Introduction: Developing The International Encyclopedia of Geography

The International Encyclopedia of Geography: People, the Earth, Environment, and Technology represents one of the most ambitious and far-reaching publication projects in the recent history of the fields of geography and GIScience. This 15-volume work is published in both hard copy and online. Annual online updates and prospective extensions to supplemental material and resources will enhance the value of the Encyclopedia to the researcher, educator, and user alike.

This six-year project, initiated by the American Association of Geographers (AAG) and Wiley Blackwell in 2010, has engaged geographers, GIScientists, and geographic societies around the globe. Its editors and authors reflect the interdisciplinary and international nature of geography’s scholarly and research activities. The sheer scale of this undertaking, in terms of the Encyclopedia’s depth and breadth of coverage and its international scope, has not, to our knowledge, been attempted before in the contemporary discipline of geography.

Geography today is a rapidly growing, dynamic, and highly relevant field, frequently leading the way in providing groundbreaking research and meaningful applications to address our world’s major societal problems and global challenges. Building on its traditional strengths of interdisciplinarity and integrative science, infused today by vibrant critical theory and revolutionary new geographic technologies, geography increasingly occupies a central place in the university and in society more broadly. Geographical theories, concepts, methods, tools, and applications now permeate environmental science, the social sciences, the health sciences, international studies, and the humanities, generating productive new research collaborations and extending knowledge frontiers.

The scope and range of The International Encyclopedia of Geography enables a much fuller discussion of the multiple subdisciplines and perspectives of modern geography than is typically the case in such endeavors, and it permits the engagement of ideas and topics from other closely aligned fields. The format of an encyclopedia, with the alphabetical sequencing of topics across the full sweep of the discipline, generates its own insightful juxtapositions and interactions among geographical topics, and illustrates how geography’s subfields inform and enrich one another. In-depth entries of up to 10,000 words allow key topics and concepts in geography and GIScience to be analyzed and presented in ways that recognize their inherent complexity.

The goal and the significant challenge of this publication project was to create the most comprehensive and authoritative in-print and online resource available for geography, while incorporating international perspectives and encompassing the discipline as a whole, broadly and synergistically defined to include:

- human geography,
- physical geography,
- geographic information science and technologies,
- economic geography and regional development,
- nature–society interactions.

It is our hope that this reference work will become the first and foremost destination for all
INTRODUCTION: DEVELOPING THE INTERNATIONAL ENCYCLOPEDIA OF GEOGRAPHY

those needing scholarly, authoritative information about these fields for decades to come. A valuable resource for educators and students at all levels, the Encyclopedia provides an enormous body of new content and analysis for teaching geography and the geosciences, environmental studies, and the social sciences and humanities fields in schools and universities. It also will be an indispensable resource for governmental ministries and planning agencies, and for private-sector firms and society at large now benefiting from the explosive growth of new geographic technologies and geospatial information.

The International Encyclopedia of Geography is designed to address a worldwide audience, which will have a comprehensive, accurate, and regularly updated account of the field at its fingertips. The entries are designed to be accessible at the advanced undergraduate and graduate level, but will also appeal to researchers and top scholars in the field. They provide accessible introductions to basic concepts as well as sophisticated coverage of complex or contentious areas.

What also distinguishes this project from other encyclopedias is that it is truly international, and it has the institutional support of the AAG and other major geographical associations from around the world. This institutional support will ensure that the Encyclopedia will be updated on an ongoing basis and, as such, has the potential to continue as the authoritative reference work in the fields of geography and GIScience for decades to come.

Where appropriate, the Encyclopedia also incorporates perspectives and input from across the full spectrum of interdisciplinary research—from science, the social sciences, and the humanities—to bear on the topics it explains and explores. Throughout, the editors also have encouraged co-authorship by collaborative teams of human and physical geographers, or teaming with GIScientists or social theorists to provide comprehensive coverage of cross-cutting and cutting-edge topics.

In sum, our goal has been to provide a serious, comprehensive, in-depth, peer-reviewed analysis of these fields for an interdisciplinary audience of scholars, graduate students, undergraduates, professionals, and other interested researchers as well as the general public. Our ongoing plan is to regularly update and maintain The International Encyclopedia of Geography as the world’s leading reference resource for the field, one that genuinely engages the needs of the international academic and professional research communities.

Structure and process of the project

The International Encyclopedia of Geography contains approximately 4 000 000 words, and 8000 printed and bound pages in 15 volumes, including an index. Over 1000 illustrations, graphs, charts, and color photographs are included. An international editorial team was responsible for ensuring that the entries are relevant, accurate, and consistent and, in so doing, ensuring that the Encyclopedia as a whole, and each of the entries, is of high quality. At least three peer reviewers assessed each entry to check that it conforms to well-established standards of scholarly publication and fairly and adequately presents the state of the field for the subject matter.

We selected an experienced and distinguished general editorial team, which helped inform the overall project and guide the way in developing a taxonomy of topical entries that knits together the research in the major fields of geography. The core editorial team consists of myself as Editor-in-Chief and five outstanding General Editors—Noel Castree (Human–Nature Interactions), Michael F. Goodchild (GIScience and Technology), Audrey Kobayashi (Human Geography), Weidong Liu (Economic Geography, Economics, and Development), and Catherine Hall (Geography and Development).
and Regional Development), and Richard A. Marston (Physical Geography)—all of whom have broad networks of contacts in their areas of expertise, as well as editorial experience with leading journals and other publications. Our core editorial team worked closely with 34 Section Editors, each a leading subject-matter expert representing a relevant subfield. Each Section Editor reviewed and edited approximately 30 entries each.

The online version of the Encyclopedia includes advanced search functions and extensive cross-linkage to related entries and to online supporting information. Considerable care was taken in creating the taxonomy of entries and in the choice of contributors commissioned, to provide state-of-the-art analyses and discussion written in an accessible style, in keeping with the aims of a definitive reference work. In addition to the entries themselves, the Encyclopedia includes a thematic lexicon by subject, an index, and an appendix of geographical associations worldwide.

The International Encyclopedia of Geography is published in conjunction with Wiley Blackwell, with whom the AAG worked closely during all editorial and production phases of the project. Distribution of the Encyclopedia will be worldwide and explicitly reflects the international scope of the work.

Using the encyclopedia

The structure and layout of the publication is designed to enhance its value as a reference work and an educational resource.

- In the online encyclopedia, abstracts and a list of key words provide a quick reference to the content.
- Search functions enable easy location of specific topics and also enhance browsing of related topics and resources.
- The entries are substantive and most are long enough (i.e., as long as regular journal articles) to be self-contained position statements on a topic, thus being suitable for students to be given on a reading list as an article to read.
- In the print encyclopedia, entries are presented alphabetically rather than grouped by narrower geographic subdisciplines, to encourage a grasp by the reader of the interactive nature of discipline and its subfields. The SEE ALSO listing at the end of each entry provides suggestions for other closely related entries in the encyclopedia.
- A list of References documents the text of each of the entries and alerts the reader to articles that expand on specific aspects of the entry. References are readily available following each individual entry so they can be consulted immediately, rather than having to refer to a distant master list of references elsewhere. Many entries also include a Further reading list that provides focused suggestions for readers who wish to explore a topic further.
- A clear example illustrating how students or others should cite an entry from the Encyclopedia is provided on page iv.
- Unlike most other encyclopedias, each entry of The International Encyclopedia of Geography has been peer-reviewed to enhance the quality and accuracy of the content, and to ensure that citations of the Encyclopedia’s entries by students and researchers are far more authoritative and reliable than is often the case in a world of increasingly unscreened and questionable Internet sources.
- The online encyclopedia allows for enhanced searching and discoverability. Users can search by key word or phrase across the full
INTRODUCTION: DEVELOPING THE INTERNATIONAL ENCYCLOPEDIA OF GEOGRAPHY

text of this resource, or browse alphabetically or by topic to explore the breadth of coverage. Users can export citations, click through reference links, and bookmark content to share on social networks.

Building international collaboration and community

The International Encyclopedia of Geography has been not only a project, but also a process. It has involved reaching out to, listening to, and working with thousands of geographers, editors, authors, educators, and reviewers around the world. Significantly, one of my reasons for undertaking this daunting project was the opportunity to build international collaboration around a major project for geography’s future. As former AAG President Ken Foote pointed out in 2010, The International Encyclopedia of Geography not only “will be an influential work for years to come but also represents an important community-building project within the discipline, both nationally and internationally.”

As such, we have sought to engage leading members of the international geography and GIScience communities as participants, authors, and editors. I am pleased to report that this first edition of the Encyclopedia includes contributions by distinguished editors and authors from 45 different countries. We also made a special effort to involve younger editors and contributors who are working at the cutting edges of new directions in geography and GIScience, including those who can address the evolving nature and diversity of our discipline, ranging from new trends in critical theory to our revolutionary and rapidly changing new geographic technologies.

Continuous updating of the Encyclopedia will provide further ongoing opportunities for broad participation by all geography and GIScience communities in this landmark project. As we move forward with regular updates of key entries in the Encyclopedia, we encourage your suggestions on candidates for editorial roles and new authors for the project, and also indications of your own interest in being considered as a contributing author or editor. Special informational and discussion sessions will be scheduled during AAG Annual Meetings in the years ahead for those who wish to learn more about The International Encyclopedia of Geography. Thank you in advance for your ideas and suggestions, and I look forward to working together with you on this most engaging project in the future.

Acknowledgments

The International Encyclopedia of Geography is the work of many talented individuals. I would like to express special appreciation for the opportunity to work with the highly distinguished and talented General Editors and Section Editors, who are listed on the Editorial Board page. A total of 1070 authors contributed 1032 entries to the Encyclopedia. To each: thank you for your time, knowledge, care, and effort.

A very special and heartfelt thank you goes to Jennifer Cassidento, the freelance project manager working with both the AAG and Wiley, without whom an encyclopedia of this scale and complexity would have been impossible to produce in the time frame allotted to us. Jennifer was with the project from beginning to end, and her efficiency, productivity, and commitment was extraordinary.

I also would like to acknowledge the editors at Wiley Blackwell, who contributed patience and good advice throughout this project. Justin Vaughan, who first approached me regarding the Encyclopedia project, has provided continuous support and encouragement throughout the
full publication process. Others at Wiley who helped at different times with editorial aspects of the *Encyclopedia* include Tiffany Mok, Lucy Wheeler, Kerry Powell, Clive Lawson, Layla Harden, Lisa Sharp, and Tanya McMullin. Barbara Duke, who led the production aspects of the project, Nik Prowse, the freelance project manager who led the copy-editing, proofreading, and indexing processes, copy-editors Giles Flitney, Jacqueline Harvey, Alta Bridges, and Kevin Dunn, proofreaders Mary Malin, Paul Sensecall, Felicity Watts, Graeme Leonard, David Adams, Jane Hammett, and Harriet Stewart-Jones, and indexers Michelle Baker, Marian Preston, James Helling, and Milla Hills in association with Indexing Specialists (UK) Ltd also contributed in essential ways, as did marketing staff Amanda Banner, Louise Morgan, and Patrick Wright.

Last but not least I would also like to acknowledge talented AAG staff members Jenny Lunn, Miranda Lecea, Joy Adams, Michelle Ledoux, Teri Martin, and Rebecca Pendergast, who provided valuable support to the project at various stages of its development. Michael Solem, Candida Mannonzzi, and Jan Monk also worked closely with Ken Foote to research and compile the appendix of international geography associations, a valuable addition to the *Encyclopedia*.

In closing, I am deeply grateful for the efforts of all those who have participated in creating this remarkable *International Encyclopedia of Geography*, which I hope will play a role in contributing to greater understanding of our world and, perhaps even more importantly, to greater understanding in our world.

Douglas Richardson
Editor-in-Chief
drichardson@aag.org
Accessibility, employment

Qing Shen
University of Washington, USA

Accessibility — more specifically, spatial accessibility — refers to the ease with which an individual or organization can reach spatially distributed opportunities, or, similarly, the potential for an individual or organization to interact with spatially distributed activities. It is a function of the locations of opportunities (activities) and opportunity seekers (activity participants), as well as the transportation and communication connections between them. This concept is frequently linked to employment because accessibility to economic opportunities, commonly known as employment accessibility or job accessibility, is especially important for the wellbeing of working-age individuals and for the productivity and livability of cities. Employment accessibility has significant efficiency, equity, and sustainability implications.

This entry first discusses the conceptualization and measurement of employment accessibility, focusing primarily on several operational accessibility measures that are commonly applied to urban planning and policy analysis. It then illustrates the usefulness of these employment accessibility measures for understanding the spatial structure of metropolitan labor markets and for examining the employment outcomes of urban residents, especially the poor.

Conceptualization and measurement of employment accessibility

Half a century ago, Hansen (1959) defined accessibility as the potential for interaction with spatially distributed activities. He operationalized this concept as a measurement of the spatial distribution of activities about a location, adjusted for the ability and desire of people or organizations in the location to overcome the spatial separation. Formally, the accessibility at location $i$ to a particular type of activity in location $j$ is directly proportional to the size of activity in location $j$ and inversely proportional to some function of the distance between these two locations; the total accessibility is the summation of the accessibility to activity in each of the individual locations around location $i$. Hansen’s conceptualization can be generally expressed by the following simple mathematical formula:

$$A_i = \sum_j O_j f(c_{ij})$$

where $A_i$ is the accessibility for people or organizations in location $i$; $O_j$ is the size of activity in location $j$; and $f(c_{ij})$ is an impedance function capturing the effect of $c_{ij}$, the cost of spatial separation between locations $i$ and $j$. For a metropolitan region with $N$ locations, $i, j = 1, 2, \ldots, N$.

Equation 1 has been widely adopted by researchers in geography, urban planning, and other related fields. While seemingly simple, application of this general formula actually
Accessibility involves many decisions, and some can be complicated. What should be considered as relevant activities or opportunities, and hence included in the variable $O_j$? What should be used to measure $c_{ij}$, the cost of spatial separation? Should it be distance, travel time, or travel cost? What if people choose different transportation modes to reach the opportunities? Moreover, how should the impedance function $f(c_{ij})$ be specified? Should it be a power function, exponential function, or some other mathematical form? Answers to some of these questions are dependent on the application context. For example, if the researcher is interested in comparing residential locations in terms of accessibility to grocery stores, $O_j$ will be the number or total size of grocery stores in location $j$. And if the residents rely on different transportation modes that have substantially different speeds, accessibility will be more appropriately measured for each mode separately, and $c_{ij}$ will probably be most effectively represented by the travel time between locations $i$ and $j$ for each mode. The selection of an appropriate impedance function is made largely on an empirical basis; the exponential function is the most frequent choice, but other reasonable mathematical forms usually lead to broadly consistent results.

Different application contexts may not just require different specifications of the components of equation 1; they may necessitate differently formulated accessibility measures. In light of several categories of application, including evaluation of the transportation and land-use system, modeling travel behavior, modeling urban development, and characterizing urban spatial structure, Morris, Dumble, and Wigan (1979) reviewed a wide range of accessibility measures. These various measures were results of many researchers’ attempts to link accessibility with application contexts and behavioral theories. Given the critical importance of economic opportunities in people’s lives and in cities’ formation and growth, a large portion of the literature on spatial accessibility has focused on employment accessibility. The approach to conceptualizing and measuring employment accessibility is to treat it as accessibility to a particular type of activities or opportunities. Following this approach, employment accessibility can be measured by adopting an appropriate accessibility measure and specifying the activity in location $j$ ($O_j$) as the size of employment, or number of job opportunities, in location $j$, with any necessary adjustment of parameter values in the impedance function. For example, equation 1 is frequently applied to the measurement of employment accessibility, typically with $O_j$ measured by employment in each location and $f(c_{ij})$ computed using data on commuting flows and times.

Employment accessibility measured by equation 1 is effective and appropriate for many applications. For example, it can be used to compare alternative locations in terms of urban development potentials because higher employment accessibility values computed using equation 1 show greater location advantages in interacting with economic activities within a city or region. Measuring employment accessibility this way is also generally useful for evaluating a transportation and land-use system by finding out whether jobs are located within reasonable proximity, or connected by effective transportation, to workers’ homes. This simple measure of employment accessibility can even help compare cities in terms of the size and diversity of a labor pool that each draws, which provides a basis for assessing the subsequent agglomeration economies that each creates.

However, equation 1 is not suitable for the measurement of employment accessibility in some other contexts. Most notably, it is not an appropriate measure of job seekers’ potential for

Given the critical importance of economic opportunities in people’s lives and in cities’ formation and growth, a large portion of the literature on spatial accessibility has focused on employment accessibility. The approach to conceptualizing and measuring employment accessibility is to treat it as accessibility to a particular type of activities or opportunities. Following this approach, employment accessibility can be measured by adopting an appropriate accessibility measure and specifying the activity in location $j$ ($O_j$) as the size of employment, or number of job opportunities, in location $j$, with any necessary adjustment of parameter values in the impedance function. For example, equation 1 is frequently applied to the measurement of employment accessibility, typically with $O_j$ measured by employment in each location and $f(c_{ij})$ computed using data on commuting flows and times.

Employment accessibility measured by equation 1 is effective and appropriate for many applications. For example, it can be used to compare alternative locations in terms of urban development potentials because higher employment accessibility values computed using equation 1 show greater location advantages in interacting with economic activities within a city or region. Measuring employment accessibility this way is also generally useful for evaluating a transportation and land-use system by finding out whether jobs are located within reasonable proximity, or connected by effective transportation, to workers’ homes. This simple measure of employment accessibility can even help compare cities in terms of the size and diversity of a labor pool that each draws, which provides a basis for assessing the subsequent agglomeration economies that each creates.

However, equation 1 is not suitable for the measurement of employment accessibility in some other contexts. Most notably, it is not an appropriate measure of job seekers’ potential for
(or ease of) reaching job opportunities. This is because equation 1 does not consider the spatial distribution of job seekers, and therefore implicitly and incorrectly assumes that the demand for the available employment opportunities is uniformly distributed across the space (Shen 1998). For opportunities that are rival goods, such as jobs — each of which is for only one worker at any moment in time — competition should be an important consideration when measuring accessibility. So long as the spatial distribution of relevant job seekers is not uniform, equation 1 generates an inaccurate or even misleading result of employment accessibility.

Because job opportunities exist in locations with various levels of demand potential, accessibility to each set of job opportunities is determined partly by the demand potential for the particular location. Shen (1998) proposed the following formula to measure the accessibility to any type of opportunities that are rival goods, and used it to compute the employment accessibility:

\[
A_i = \sum_j \frac{O_j f(\epsilon_{ij})}{\sum_k P_k f(\epsilon_{kj})}
\]  

(2)

where \(O_j\) is the number of job opportunities in location \(j\); \(P_k\) is the number of job seekers in location \(k\); and \(f(\epsilon_{ij})\) is an impedance function capturing the effect of \(\epsilon_{ij}\), the cost of spatial separation between locations \(k\) and \(j\). For a metropolitan region with \(N\) locations, \(k = 1, 2, \ldots, N\). Other notations are the same as in equation 1.

Recognizing that different opportunity seekers may depend on different transportation modes for search and travel, yet compete for the same opportunities, Shen (1998) extended equation 2 into an employment measure for a multimodal context. For job seekers who live in location \(i\) and commute by mode \(v\), employment accessibility is computed as follows:

\[
A_i^v = \sum_j \sum_k \frac{O_j f(\epsilon_{ij}^v)}{\sum_{m} P_{km}^v f(\epsilon_{kj}^m)}
\]  

(3)

where \(P_{km}^v\) is the number of job seekers living in location \(k\) and commuting by mode \(m\); and \(f(\epsilon_{ij}^v)\) and \(f(\epsilon_{kj}^m)\) are impedance functions capturing the effect of spatial separations between, respectively, locations \(i\) and \(j\) for mode \(v\) and locations \(k\) and \(j\) for mode \(m\). For a transportation system with \(M\) modes, \(v, m = 1, 2, \ldots, M\). Other notations are the same as in equation 2.

Equation 3 enables researchers to directly compare accessibility among opportunity seekers, such as job seekers, who are differentiated by transportation modes and by residential locations. Sometimes, it is also useful to examine the average level of accessibility for all opportunity seekers who live in the same location (say, a residential neighborhood). In the context of employment accessibility, this can be accomplished by computing the weighted average value for all the subgroups of job seekers commuting by different modes:

\[
A_i^G = \sum_v \left( \frac{P_i^v}{P_i} \right) A_i^v
\]  

(4)

where \(A_i^G\) is a general measure of employment accessibility for \(P_i\), all job seekers living in location \(i\); \(P_i^v\) is the number of job seekers in location \(i\) who commute by mode \(v\) and hence, enjoy accessibility \(A_i^v\). For a transportation system with \(M\) modes, \(v = 1, 2, \ldots, M\).

Shen (1998) showed that equations 2, 3, and 4 share an important property: the expected value of measured accessibility values equals the ratio of the total number of opportunities to the total number of people seeking the opportunities. This property provides a useful benchmark for assessing the resulting accessibility values, as
Advances in information and communications technologies (ICT), especially the rapid growth and wide applications of the Internet over the past two decades, have created new spatial paths and barriers that profoundly affect people’s access to economic opportunities and social services. For example, recent American Community Survey data indicate that “work at home” has the highest growth rate among all commuting modes, implying the growing importance of ICT in employment accessibility (AASHTO 2013). This has presented a major challenge to the traditional accessibility concepts and measures, which focus exclusively on the physical space. Research efforts have been made to incorporate cyberspace into the measurement of accessibility. For example, Shen (1999) extended conventional accessibility measures by conceptualizing the total accessibility from location $i$ as the sum of three components:

$$ A_i = A_i^p + A_i^c + A_i^h $$

(5)

where $A_i^p$ is accessibility to opportunities in the physical space that require transportation for access; $A_i^c$ is accessibility to opportunities in cyberspace that require ICT for access; and $A_i^h$ is accessibility to opportunities in the hybrid space that can be accessed through either transportation or ICT.

By classifying job opportunities accordingly into the three categories, and job seekers into two groups based on whether they have the ability to reach opportunities in cyberspace, equation 5 can be specified for each group of job seekers. $A_i^c$ and $A_i^h$ can be quantified because people’s activities in cyberspace (e.g., telecommuting) are embedded in their activities in the physical space that involve time and monetary costs.

Conceptualizing and measuring employment accessibility in the age of advanced ICT is still at the exploratory stage, and there is much exciting research to be done. Additionally, innovations in transportation continue to pose new questions regarding employment accessibility. For example, future development and deployment of automated vehicles may lead to some significant changes in how we measure accessibility, probably with an increased emphasis on the effects of monetary cost and intermodal coordination.

**Employment accessibility, spatial barriers, and employment outcomes**

Accessibility measures represented by equations 1, 2, and 3 have been widely applied to urban economics, geography, planning, and other related fields. Several applications of employment accessibility measures to research on the nature and impacts of spatial barriers to employment in American metropolitan areas, briefly described below, illustrate the usefulness of these measures.

Unemployment and poverty has been an important research area for urban economists, geographers, planners, and sociologists, among others. The high-level concentration of unemployed adults—who are disproportionately African Americans—in poor neighborhoods in central cities of the United States, along with industrial restructuring and employment decentralization that have been taking place for the last five decades, has drawn many scholars’ attention to the spatial dimension of this problem. A widely shared view, commonly referred to as the “spatial mismatch hypothesis,” is that the structural and spatial transformations of urban labor markets, combined with racial discrimination in housing markets, have caused the isolation of low-skilled minorities...
ACCESSIBILITY, EMPLOYMENT

in poor central city neighborhoods which are disconnected from suburban job opportunities (Gobillon, Selod, and Zenou 2007). Many proponents of the “spatial mismatch hypothesis” believe that central cities have become a geographically disadvantaged residential location for low-skilled job seekers because relevant job opportunities for them are mostly in suburban locations, and that an effective approach to solving the problem of urban unemployment is to relocate the poor to the suburbs.

To examine spatial barriers to low-skilled job seekers’ labor participation in American cities, Shen (1998) applied the accessibility measures expressed by equations 3 and 4 to an empirical study of the Boston metropolitan area. He showed that for job seekers who travel by a given transportation mode, those living in central city neighborhoods still have relatively higher employment accessibility than those living in the suburbs. He also showed, however, that travel mode is far more important than residential location in determining employment accessibility. For job seekers who can travel by car, most residential neighborhoods allow them to have good accessibility to jobs, but for those who depend on public transportation, only a few residential neighborhoods – all located in the central city – give them good employment accessibility. For the great majority of poor central city neighborhoods, the limited location advantage is far from sufficient to offset the comparative disadvantage of public transportation on which a high percentage of their residents depend. In other words, many poor central city minorities are indeed spatially disadvantaged, but the reason for this is their relatively low transportation mobility resulting from dependence on the slower travel mode (public transit), not their residential location being too far from job opportunities.

The more recently published study by Grengs (2010) tells a similar story, but it is particularly compelling because the case is the Detroit metropolitan region where the central city has suffered the most from structural and spatial transformations of the urban economy. His results indicate clearly that even though the areas of high employment accessibility in Detroit are more decentralized and dispersed than in other metropolitan regions, the central city remains a place of relatively good accessibility – better than most of the rest of the region. In addition, public transportation is shown to be such a severely inferior substitute for owning a car in Detroit that nearly every location offers better employment accessibility by auto than the location with the best employment accessibility by public transit.

Part of the explanation for central cities’ location advantage in reaching job opportunities is that they are typically the median location of such opportunities. In addition, for transit-dependent job seekers, central cities are better served by public transportation. Yet another reason, which is perhaps the most important one, is that all types of employment opportunities are still relatively concentrated in central city areas despite decades of industrial restructuring and employment decentralization. In a typical American metropolitan region, job openings are created predominantly by turnovers from pre-existing employment; new jobs typically account for only a minor portion of job openings at a given time. Even low-end jobs are still somewhat more concentrated in central city areas than low-skilled adults seeking these jobs.

Many researchers have taken another step to examine the effects of employment accessibility on employment outcomes. Most such research efforts focused on employment status, commuting distance, or cost as the outcome measure. Studies of the relationship between
employment accessibility and commuting distance or cost have obtained highly consistent results indicating that higher accessibility generally shortens commuting and reduces its cost. However, findings on the effect of employment accessibility on residents’ employment status have been mixed. While some researchers found that higher accessibility tended to be associated with higher probability of being employed, other researchers found only a weak or no relationship between these two variables.

Hu and Giuliano (2015) used the Los Angeles metropolitan area as the case study area to examine the effects of employment accessibility on the labor market outcomes of poor job seekers. Their results show that accessibility does not explain unequal employment status of the poor. On the other hand, their results indicate a significant association between accessibility and poor job seekers’ commute distance.

These studies challenge the traditional conceptualization of spatial mismatch. They indicate clearly the critical importance of transportation, rather than residential location, in determining a low-skilled job seeker’s spatial advantage or disadvantage within a metropolitan labor market. However, their findings also suggest that spatial accessibility only plays a rather limited role in job seekers’ employment status. These new perspectives point to the need to redirect public policies from encouraging housing dispersal to improving transportation mobility — perhaps by enabling more job seekers to use cars — and to develop new strategies for tackling other major barriers to employment. Indeed, Teitz and Chapple (1998) provided a wide variety of theoretical explanations for concentrated poverty in American cities, which point to economic, racial, and cultural, as well as spatial, factors as possible key causes. A critical future step toward developing effective strategies is to gain a deeper understanding of how the various spatial and nonspatial factors individually and interactively influence the employment outcomes of urban residents, especially those living in poor neighborhoods.

**SEE ALSO:** Accessibility, in transportation planning; Industrial restructuring; Information and communications technology; Poverty; Race, work, and employment; Suburbanization; Transportation and land use; Transportation planning; Unemployment and “underclass”; Urban transit

**References**


Accessibility is a deep-seated concept in spatial planning and transportation geography as it is the main product of transportation systems and has gained wide acceptance as a policy target. The concept of accessibility helps to explain the interrelationships between human activity, land use, and transportation system. Apart from its explanatory power, accessibility also plays an important role in the ex-ante and ex-post appraisal of land-use and transport infrastructure plans. At an aggregate level, accessibility is a key driver of regional development generating positive externalities for social cohesion, environmental sustainability, and economic growth.

Notwithstanding the currency and widespread use of the notion in transport planning and research, there is still no consensus about its exact definition. Definitions include “the potential of opportunities for interaction” (Hansen 1959), “the inherent characteristic (or advantage) of a place with respect to overcoming some form of spatially operating source of friction” (Ingram 1971), “the ease with which any land-use activity can be reached from a location using a particular transport system” (Dalvi and Martin 1976), and “the freedom of individuals to decide whether or not to participate in different activities” (Burns 1979). While diverging definitions abound, a common denominator is that accessibility is associated with the potential to reach spatially dispersed activities. Conceptually, this potential has been approached from the perspective of either locations or persons. The former attributes accessibility to a location representing how well this location is connected to other locations, while the latter attributes accessibility to a person indicating how well that person can reach a set of alternative destinations.

The variety of perspectives and meanings assigned to accessibility has prompted a host of different accessibility measures, with a tendency towards increased complexity. Location-based accessibility measures usually have three elements in common: (i) a reference origin (e.g., home location or zonal centroid), (ii) the quality and/or magnitude of activities at destinations (e.g., number of jobs in a zone), and (iii) a cost of physical separation between the origin and the destinations (e.g., travel time). An early yet still intensively employed mathematical expression of location-based accessibility has been formulated by Walter G. Hansen (1959) in his seminal paper “How Accessibility Shapes Land Use.” Hansen’s formulation is inspired by the notion of population potential in social physics and assumes accessibility at a certain location to be directly proportional to the size of opportunity at some other location, and inversely proportional with the distance, travel time, or composite travel cost between them. In a general form, the total accessibility $A_i$ of zone $i$ to all other zones $j$ can be expressed as:

$$A_i = \sum_j a_j f(c_{ij})$$

where $a_j$ denotes the size of opportunity at zone $j$, $c_{ij}$ represents the travel impedance between zones $i$ and $j$, and $f$ is a function reflecting
the dampening effect of distance, time, or transport costs on travel. Since the dimensions of the quantities in this equation are generally incompatible, the resulting accessibility values are often interpreted in purely relative terms by normalizing the values over a particular range. The impedance function $f$ can take different forms, with the inverse power $(c_{ij}^{-\alpha})$ and negative exponential function $(e^{-\beta c_{ij}})$ being most frequently employed. Ideally, decay parameters $\alpha$ and $\beta$ should be estimated on the basis of observed travel behavior and differentiated by mode, trip purpose, and socio-economic attributes. In practice, they are often borrowed from locally calibrated trip generation models.

Another related and commonly applied accessibility measure is a cumulative opportunity measure, also referred to as an isochrone measure, contour measure, proximity count, or container index. This measure counts the number of opportunities (e.g., jobs, hospital beds, shops) that can be reached within a predetermined cut-off distance or time. In most studies, this fixed cutoff value is not empirically substantiated but simply equated to an arbitrary threshold value that delimits catchment areas that are deemed reasonable in the application context at hand. This type of measure can be seen as a special case of gravity measure where $a_j$ denotes the number of opportunities at zone $j$ and the shape of the impedance function $f$ is rectangular; that is, it takes the value of 1 if the zone is within the predetermined threshold, and 0 otherwise. Cumulative opportunity measures have the advantage of being more intuitive and communicable than the cost-discounted values of gravity measures, making them appealing for planners and decision-makers. In contrast to gravity measures, cumulative opportunity counts do not need to be normalized since they can be directly interpreted in comparable and absolute units.

Despite their popularity in empirical work and policy, the above measures have been criticized for at least three reasons. First, zonal location-based accessibility measures are liable to aggregation problems such as the modifiable areal unit problem, meaning that accessibility values depend on the arbitrary configuration and scale of a zoning system. They also suffer from ecological errors since they attribute the same level of accessibility to all persons residing in the same zone, although these persons may dispose of different mobility resources and therefore perceive the set of destination alternatives differently. In addition, zonal measures are insensitive to local variations in accessibility resulting from intra-zonal improvements to the transport system. Second, location-based measures adopt a single reference location (i.e., origin zone or location) as the origin of travel, thereby ignoring the effects of trip-chaining behavior and the interdependence between activities. Third, place-based measures ignore temporal fluctuations in accessibility resulting from congestion, temporal availability (e.g., opening hours) of urban opportunities, and the fact that individual time budgets to travel to and participate in discretionary activities vary substantially throughout the day, between days, and across individuals.

In order to address the above limitations, more sophisticated person-based measures with a stronger behavioral basis have been developed. Two groups of person-based measures can be distinguished: utility-based measures and space–time measures. Utility-based measures are derived from discrete choice theory and assume that a decision-maker $p$ assigns a utility to each destination $j$ in a choice set $J_p$ and then chooses the alternative that maximizes his/her utility. Under the assumption of a logit decision process, the accessibility of individual $p$ to choice set $J_p$
can be expressed as a logsum, being the denominator of a multinomial logit model (Ben-akiva and Lerman 1979):

\[ A_p = \frac{1}{\lambda} \ln \sum_{j \in J_p} e^{-\lambda u_{pj}} + K \]  

where \( \lambda \) is a travel cost sensitivity parameter and \( K \) is a constant of integration which is sometimes set equal to zero, implying that no accessibility is derived if the measured utility \( u_{pj} \) equals zero. Constant \( K \) can also be canceled out of the equation by taking the difference between the accessibility benefits of two choice situations (e.g., before and after the introduction of policy). Note that accessibility expressed by a logsum measure is now person-specific rather than attached to a location or zone. Despite its theoretical appeal, the application of the logsum measure is not standard practice in land-use and transport project appraisals.

In order to measure accessibility at the individual level, researchers have also relied on the space–time approach. While this approach goes back to space–time studies of the 1960s and 1970s (Hägerstrand 1970), it was particularly during the last two decades that significant inroads were made in terms of operationalization. This is in part due to advances in geographical information systems (GIS) and the increased availability of data on individual activity-travel behavior. Space–time measures are based on the time-geographic construct of the space–time prism, which is the set of possible paths through space and time that a person can follow, given a set of space–time constraints resulting from mandatory commitments and limited mobility resources. The spatial footprint of the prism, the potential path area, gathers all locations where discretionary activities can be performed. Its size or the opportunities contained within it represent a direct measure of individual accessibility. While measures based on the potential path area enable the comparison of accessibility among different social groups, the requirement of individual-level data and the computational burden associated with their implementation hampers their practical use. Only a few insightful illustrations of the approach with real-world data have surfaced to date (e.g., Kwan and Weber 2008; Neutens et al. 2010). Space–time approaches have been integrated with utility-based approaches in the late 1970s and early 1980s through the work of Beckmann, Golob, and Zahavi (1983) on travel probability fields, and through the space–time utility function proposed by Burns (1979) and its subsequent modifications (Ettema and Timmermans 2007; Miller 1999).

SEE ALSO: Accessibility, employment; Geographic information science; Transport geography; Transportation and land use

References


ACCESSIBILITY, IN TRANSPORTATION PLANNING


Actor-network theory

Beth J. Greenhough
*University of Oxford, UK*

Actor-network theory (ANT) is an approach to research developed within the interdisciplinary field of science and technology studies (STS). ANT is shaped by the conviction that the world as we encounter it is not some pre-given, natural entity. Instead, reality is constantly in the process of being socially and materially reconstructed through the relations and associations formed between human and nonhuman actors. ANT is noted for both its refusal of hierarchies (for these assume that the order of things is given rather than socially and materially contingent) and for its recognition of the role of nonhuman agency, incorporating everything from living beings to technological devices in its analysis of how particular assemblages of agents – actor-networks – are produced and reproduced. The term assemblage is one increasingly being used within geography to stress how the phenomena we study are not discrete social or material entities, but often hybrid collections of diverse human and nonhuman, organic and inorganic, natural and technological elements. Consider, for example, genetically modified foods, which are organic and nonhuman in form but dependent on human expertise and technology and inorganic fertilizers for their production and reproduction, and which, given their controversial status, cannot be understood outside of their social as well as material context.

Origins of ANT

ANT was developed in the mid-1980s at the Centre de Sociologie de l’Innovation of the École Nationale Supérieure des Mines de Paris, by STS scholars Bruno Latour and Michel Callon, and the visiting British sociologist John Law. Latour, Callon, and Law were critical of science’s claims to objectivity – its claims to offer a “view from no-where” – and countered this by stressing the ways in which scientific knowledge is socially constructed, shaped by the social and environmental contexts of its production. However, where earlier work in social studies of science had focused on the analysis of scientific discourse and controversies, Latour and his colleagues brought a new ethnographic approach to studying the production of scientific knowledge. In their landmark publication, *Laboratory Life*, Latour and his colleague Steve Woolgar (1979) sought to demonstrate how scientific facts are both socially and materially constructed, highlighting the role of nonhuman (technologies, materials, organisms, machines, experimental subjects) as well as human actors in producing and verifying scientific knowledge. They sought a middle ground between the strong social constructivism of their predecessors (who insisted the world is shaped by relationships between human social actors, or in other words it is about who you
**ACTOR-NETWORK THEORY**

know and who knows you) and the ontological imperatives of scientific analysis to engage with the substance of the “real” world (the conviction that the world is made up of material elements with pre-given physical properties that can be uncovered through rigorous, objective scientific analysis). Scientific views of the world, Latour and Woolgar argued, were not pre-given facts but careful achievements produced by the assemblage of a diverse range of human and nonhuman actors into networks of relations. This detailed analysis of the actor-networks necessary to produce scientific facts evolved into a broader approach – ANT – asserting the inseparability of natural and social worlds and that phenomena, knowledge, objects and agents are co-constituted through the relations between them. The world is not preformed but emerges as humans and nonhumans interact with each other.

### Key aspects of ANT

ANT consists of four key elements. Firstly, ANT challenges the division of the world into structure and agency and other associated dualisms (nature-society, human-physical, global-local), instead offering “an encouragement to think relationally, in terms of associations rather than separations” (Castree 2002, 118). It criticizes the tendency within existing social theories to analytically prioritize (assign explanatory power to) one side of these dualisms over the other. One of the benefits of ANT is that it resists asymmetry whereby one side of the binary (e.g., nature) becomes subsumed as something explained through the other (e.g., society/social constructions). So, whereas early cultural geography studied landscape as something designed and brought into being by human actors, ANT would stress the role of nonhuman actors in co-constructing that landscape.

Consequently, despite the name, the proponents of actor-network theory (ANT) have always been keen to stress that ANT is not a theory, at least not in the sense that it offers a meta-level explanation for phenomena. Instead they suggest ANT should be thought of as a method or approach to understanding how the world comes to be. Secondly, therefore, ANT is a constructivist approach. In contrast to other approaches that seek to understand essential differences between their objects of study, ANT seeks explanations that work or are useful. Rather than choosing explanations based on the conviction that they are the one true explanation, ANT’s explanations are, therefore, contingent upon the social (and ANT’s understanding of who counts as a social actor is heterogeneous, see below) and environmental contexts from which they emerge.

Thirdly, ANT understands the phenomena with which it engages not as pre-given but as always in the process of being performed. Furthermore, phenomena must be performed constantly or they risk falling apart. This precarity means that pre-given explanations or theories are of little use. Bruno Latour argues that if we wish to understand the work which goes into pulling together and sustaining particular networks, then we need to study the different actors (both human and nonhuman) involved, producing detailed descriptions of how they sustain themselves. Explanation is only possible once the actor-network is fully mapped out; the material and the semiotic (conceptual) must be thought together. Importantly, this includes a reflexive acknowledgment of the researcher’s role in shaping the actor-networks they study: researchers intervene in and shape the world through the process of studying it.

Fourthly, ANT questions the assumption that agency is located within individual (usually human) actors, capable of intent. For ANT,
agency is a property of a diverse range of both human and nonhuman actants, or more accurately a property of the associations (networks) formed between them. ANT is critical of understandings which assume that the power to effect or resist change is an exclusively human property, one located within or against social actors and systems (e.g., capitalism or patriarchy). It is not surprising perhaps that this arose from a theory with roots in STS. One of the key insights from exploring the production of scientific knowledge ethnographically (as opposed to say through an analysis of official publications and papers) is a recognition of the notorious recalcitrance of nonhuman agency to comply with experimental protocols. Through experimental failures we are made conscious of the capacity of nonhuman agency to resist what humans conceive of as their ability to (re)order the natural world. Successful networks are often hidden from sight and many actor-networks only become visible when they fall apart.

ANT in geography

Whatmore and Thorne (2000) suggest that within geography ANT had two key contributions to make. First it provided a way (a metaphor and a method) for holding onto the situatedness of practices and subjects in the face of a stretched-out, distributed, globalized world. John Law suggests ANT should be thought of topologically, replacing Euclidian space and its insistence on seeing the world as a series of discrete objects with a more relational and fluid, networked sense of space. From this topological perspective, phenomena can only be located and understood by virtue of their connections or relations to other elements in the network. Rather than exploring local-global distinctions (topography), the ANT approach would be to focus instead on the connections and intermediaries between, flattening out hierarchical understandings of space and scale. One of the impacts of globalization is to make things seem as though they are both everywhere and nowhere. We all know what Coca-Cola is, but we would be hard pressed to tell you where it is. Actor-network theory brings to the analysis of this kind of phenomenon a sense that to be both universally (known) or everywhere while coming from no-where is not a given, but a hard-won achievement, not unlike the production of scientific knowledge, one that requires a complex chain of actors and networks. ANT would ask how exactly does Coca-Cola achieve this seemingly ubiquitous brand presence and would trace out the humans and nonhumans involved. There are analogies here with analyses of production chains and commodity networks. What distinguishes ANT though, is a recognition of the fragility of these networks and the emphasis placed not just on nonhuman components (as in a commodity chain) but on nonhuman agency.

Attention to nonhuman agency constitutes a second key contribution made by ANT to geographical scholarship. Perhaps unsurprisingly, therefore, ANT has particularly appealed to geographers working on nature-society relations, and ANT has been used to bring new posthumanist perspectives on the presence of nonhuman agency in the fabric of social life. Posthumanist approaches arose from the need to critique the ways in which human geography focused on human agency as if it could be detached and dissociated from nonhuman agency, and as if the social world could be understood through a focus on human needs, emotions and responses alone. Nonhuman agency remained the concern of the natural sciences and physical geography. The result
was an asymmetrical tendency within human and physical geography to attribute explanatory power to social and natural phenomena respectively, an approach Whatmore (2002) describes as deadening geography to the world and its liveliness, as though human society is formed and functions in ways that are ignorant of and unaffected by the actions and responses of nonhuman agents. In contrast, ANT insists nature and society are co-constituted. For example, Whatmore and Thorne (2000) draw on ANT in the development of their understanding of spatial formations of wildlife exchange, and the multiple ways in which elephants are reconfigured as they are caught up in social networks as zoo exhibits, database entries, and charismatic cover photos for conservation projects. Hinchliffe (2001) equally drew attention to co-fabrication of human and nonhuman worlds in his study of the late 1980s/early 1990s British Bovine Spongiform Encephalopathy (BSE) crisis, but did so in order to critique political tendencies to seek bureaucratic and technical quick fixes to complex biosocial problems. In drawing attention to complex interweaving and interdependency of human and nonhuman lives, this work stresses the impossibility of producing useful and functioning explanations of phenomena, such as wildlife conservation and zootonic disease, while insisting on a separation between natural and social worlds.

Other work contrasts the insights of ANT with Marxist approaches to understanding the role played by capitalism in the (re)production of nature-society relations. In an edited collection entitled Remaking Reality, Bruce Braun and Noel Castree contrasted a Marxist emphasis on how nature is transformed through capitalist imperatives with insights from poststructuralism and ANT (Castree and Braun 1998). They show how Marx’s emphasis on the ways in which the world is shaped by macroscale social and economic structures (such as capitalism) might be combined with an attention to how the capitalist imperative to generate profit and surplus value is dependent on, facilitated by, and might also be resisted by assemblages of both human and nonhuman actors brought together – for example in the process of developing, producing, and consuming genetically modified crops. Maria Kaika and Erik Swyngedouw (2000) combined ANT with more traditional political ecology approaches to explore how nature (in the form of water) is accommodated within and resists the urban fabric. Humans remain unaware how dependent they are on underground water supply networks until those networks fail, until the pipe bursts or the waterborne bacteria infect a human host. Beth Greenhough (2006) drew on Callon’s ANT-inspired engagement with commodity fetishism to explore the processes of decontextualization, dissociation, and detachment necessary to facilitate the extraction and circulation of human bioinformation and biosamples (medical records and DNA) as “bodily commodities” for the life science industries. She described how human society seemingly needs to be able to distance itself from its bodily-derivatives before it is comfortable with them being commodified and used in scientific research. Arguably because human society (and humanist social science approaches) gives so little agency to nonhuman materials and objects, we are rendered very uncomfortable by the idea that a large part of our human selves is a collection of organic and inorganic, social, and technical parts. In each of these cases ANT becomes a resource for emphasizing capitalist social relations and processes (e.g., commodification) as geographically contingent and precarious achievements. It shows that at any point both the human and nonhuman elements might resist their enrolment or exercise their
agency in ways that go against the imperative to make profits.

**Critiques of ANT**

Perhaps surprisingly, some of the strongest critiques of ANT come from within ANT itself. John Law and colleagues in their reflections in the edited collection *Actor Network Theory and After* (Law and Hassard 1999), lamented the widespread and, importantly, often uncritical adoption of ANT to the point where they note, somewhat ironically, that the network topology has in itself become naturalized. Or to put it another way, practitioners of ANT have become so accustomed to the idea that everything is defined relationally they fail to interrogate this assumption through a critical unpacking of the precise nature of the particular relations which shape the phenomenon under investigation. It is not enough, Law and his colleagues cautioned, to say things are complicated and relative. Instead we must always also ask how things are continually in the process of becoming complicated, the labor it takes to hold them stable, and the potential that at any point these relations may fall apart. This later work includes a much keener sense of the absences and exclusions within particular actor-networks, as in the process of assemblage some things are brought together in ways highly visible to human senses, others are present but hard to sense and others are excluded entirely. In *Actor Network Theory and After*, Michel Callon stresses the importance of enrolment; the process by which actants (human and nonhuman agents) become incorporated within particular networks of relations. Enrolment is not all inclusive, Callon suggests, and in contrast to structuralism’s insistence on an underlying (socioeconomic) imperative (the idea that everything is ordered by some underlying structure or goal, e.g., capitalism’s aim to generate profit or surplus), not all enrolments are intentional or seemingly beneficial for human interests. Again, for example, think about the spread of viruses or the contamination of water supplies.

Outside the field of actor-network theory, critiques have largely focused on the assertion that ANT offers little ground for building political critique and collective resistance. It has been argued that some networks are much more powerful than others, and ANT has been accused of an almost nihilistic relativism that refuses to acknowledge the ongoing influence of structural forces such as capitalism in shaping social inequality. In response, proponents of ANT criticize Marxist and other structuralist approaches for being insufficiently radical in their preoccupation with the impact of human-defined economic structures on largely human concerns. They argue that interests of the nonhumans in, for example, the impact of deforestation on Amazonian ecologies, become subsumed under a concern for how this affects (usually) marginalized and disempowered human populations within these communities. Latour’s more politically orientated later work therefore emphasizes the need to bring the nonhuman into political debates, challenging the tendency to rely on scientific analysis to represent its interests.

Castree (2002) seeks a middle ground between what he refers to as the “false antithesis” of Marxism and ANT, suggesting that we might equally conceptualize capitalism or other explanatory social systems as particularly enduring actor-networks, which, while unique, share many characteristics and compositional elements, including a focus on extracting surplus value. This way, we can acknowledge that capitalist imperatives play a role in the way actor-networks are assembled, sustained and reproduced. Furthermore, thinking of the seemingly dominant macrostructures of capitalism as precarious
assemblages opens up the possibility that they could be re-made along radically different lines, a possibility Annemarie Mol in *Actor Network Theory and After* refers to as “ontological politics” (Mol 1999).

**ANT futures**

ANT is often linked to the broader turn towards poststructuralism, an approach which criticizes structuralism’s emphasis on large-scale social structures (e.g., patriarchy) as the dominant forces shaping how the world comes to be. Since its inception ANT has been adopted across a wide range of fields, including health studies, sociology, anthropology, feminist studies, and economics as well as geography. However, by the late 1990s Latour and his colleagues were becoming increasingly frustrated by the overuse of the term and felt it had lost its critical edge. The above mentioned volume *Actor-Network Theory and After* made an attempt to re-orientate ANT towards competing, multiple, and discordant elements within the network, acknowledging how drawing things together also involves taking them apart. There was also an effort to pay greater attention to the fleeting and multiple nature of many actor-networks.

Within contemporary geography ANT is perhaps best recognized through its impact on three emerging trends. First, ANT is viewed as one of the reference points for the rise of assemblage-thinking within geography (Anderson and McFarlane 2011), sharing with these approaches an insistence on the relational and performative constitution of phenomena. Second, in its attention to performance and the nonhuman, it forms a key influence on Non-Representational Theory. This is an approach that emphasizes aspects of the world which are hard to capture through processes of representation, such as mood and sensation. Third, its attention to nonhuman agency and refusal of the separation between nature and society inform what Whatmore (2006) refers to as “materialist returns” within cultural geography, specifically concerned with the development of more-than-human geographical approaches to tackling key political environmental challenges such as climate change. It is arguably through these influences (or associations), rather than any rigorous adoption of ANT’s methodologies, that ANT continues to shape geographical thought and practice, a fittingly modest legacy for a theory which consistently resists its status as such.

**SEE ALSO:** Affect; Animal geographies; Construction of nature; Marxist geography; Nature; Nonrepresentational theory; Political ecology; Posthumanism

**References**


Agent-based modeling

Dawn C. Parker and Derek T. Robinson
University of Waterloo, Canada

Agent-based models: definition and context

A definition of agent-based models of geographic systems

Agent-based modeling (ABM) is a computational simulation methodology that represents key system actors as agents who make individualized, decentralized decisions. Agent-based models represent characteristics of real-world actors, items, and processes as virtual agents, objects, and algorithms, most often in computer code. These actors can include, but are not limited to, gas particles, animals, plants, people, or more complex social entities such as households or institutions. When applied in a geographic context, ABMs also generally include a spatially explicit landscape over which agents interact. Thus ABMs can situate agents within a geographic context, concurrently representing spatial aspects of actors, their decisions, and their interactions. This enables the modeling of both direct and indirect spatial interactions, which are often central to the research questions investigated by ABMs of geographical systems.

Identifying the community of ABM users

ABM is increasingly used throughout the social, natural, and information sciences. ABM has been applied in the domain of ecology, where it has been referred to as individual-based modeling (IBM) since the 1970s. Many geographical applications represent coupled human and natural systems (CHANS). In other disciplines ABM is often used to generate solutions to specific problems. For example, in computer science, software agents are designed to accomplish a specific task (e.g., harvest data from the World Wide Web), and in operations research, n-polynomial optimization problems are solved using agents. In many other cases ABM is used for pedagogical purposes (e.g., to illustrate theory or teach computer programming).

How ABMs develop decision models

Often modelers combine theoretical frames from several traditional disciplines to develop agent decision-making, action and reaction, or interaction rules among agents and between agents and their environment. In some research and applications, agent rules, attributes, and behaviors are empirically informed from social surveys, behavioral economics experiments, ethnographic fieldwork, ecological fieldwork, spatial analysis, and other data collection and analysis methods. Often these empirical efforts strive to (i) objectively identify groups of agents within a specific agent type (e.g., profit-driven, risk-averse, or hobbyist farmers), (ii) calibrate agent attributes and population numbers, and (iii) provide empirical patterns for validating model structure and output.

ABMs as representations of complex systems

Since it is not necessary to define and solve a mathematical equilibrium or optimum for
ABMs, as required by many other modeling methods, there is a high degree of flexibility that allows researchers to construct ABMs that represent a wide range of processes and feedbacks. Therefore, ABMs are ideal tools to represent and explore complex systems. While multiple definitions of complex systems are in play, commonly recognized components include: heterogeneity and interaction among agents and their environment, learning and adaptation, aggregation (i.e., emergence), path dependence, and nonlinear dynamics. Many of these components cannot be practically or wholly represented using mathematical (e.g., ordinary differential equations) or statistical (e.g., logistic regression) models.

Modelers incorporating the concepts of complex systems into ABMs are able to make both theoretical and applied advances by constructing a range of models that extend from simple toy models (e.g., flocking behavior in birds) that lack data but offer a proof of existence to high-fidelity models that enable the simulation of real-world physics and processes (e.g., a flight simulator). Thus, ABM is used both as a tool for exploration and to make substantive gains in domain-specific and broader scientific knowledge. In the real world, many phenomena cannot be methodically explored due to lack of data variation, especially given that the real world provides only a single historical trajectory reflecting particular circumstances. In such circumstances, ABMs can be used as “computational laboratories,” to explore such phenomena “in silico.”

One strong appeal for complex systems modelers using an ABM approach is to demonstrate that the individual actions and interactions of agents can aggregate to create macroscopic patterns of interest at the system or landscape level, which can include measures of spatial pattern and connectivity or equilibrium (i.e., emergence). If, through the validation process, these emergent outcomes are found to be consistent with a sufficient number of real-world data series and/or patterns, the models are said to offer a supportable candidate explanation for its hypothetical causal mechanisms. Again, although an equilibrium may emerge in the ABM, its conditions are not imposed on the model a priori. In addition to the standard notion of equilibrium, ABMs are also used to represent nonequilibrium situations including threshold effects, punctuated equilibrium, quasi-equilibrium, and out-of-equilibrium dynamics.

Elements of ABM

Landscape

An agent-based model of a geographic system generally begins with a representation of the spatial environment over which modeled agents interact. Typically, the data models used to represent the landscape correspond to raster and vector models used in traditional geographic information systems. The majority of spatially explicit ABMs use a raster-based representation of the landscape. This may be due to the ease with which data can be assimilated during model construction, accessed by agents within the model, and compared to remotely sensed data during model evaluation, among other reasons. Vector-based applications are increasingly used but face challenges when the subdivision of a polygon or other topological process requires representation. Other representations of space have also been used (e.g., triangular irregular networks, hexagon tessellations, networks). In all cases, the landscape units (e.g., cell or polygon) can store attribute information that facilitates linkages between stationary or mobile agents and neighborhood stimuli. In land-change models, for example, agents can modify the land use of landscape units, and these units in turn may be used to drive other modeled social or
environmental processes or inform the decisions of other agents.

Collection of agents

The second element of an ABM is a defined collection of agents. While many formal “agent” definitions are seen in the literature, an agent is generally a goal-oriented entity, possessing a model of cognition that links goals and behavior, that is capable of sensing and responding to changes in its environment through autonomous action. Agents are generally characterized by both internal resources (including attributes representing knowledge, experience, behavioral parameters, and decision strategies) and external resources, such as assets or position within a social network.

Each agent possesses a decision model – a set of rules that link information about his or her environment, resources, and estimated payoffs to a decision regarding possible actions. In an ABM, decisions often depend on the state of the agent’s attributes and the physical environment. They may also depend on the state of the agent’s social environment. Specifically, decisions in ABMs are often linked to and influenced by the prior decisions of other agents. A wide variety of models are used to represent decision-making in ABMs, including rule-based and classifier systems, regression and neural network models, and optimization routines such as mathematical programs or genetic algorithms. There is no consensus on the most appropriate way of representing agent decision-making in an ABM, and current decisions are based on the complexity of the actors represented, the skill and domain specialization of the modeler(s) or project team, and the data and time available.

Interaction

The third key element of ABMs is a set of rules governing interaction among agents and between agents and their environment. Rules governing agent–agent interactions may involve a distance parameter that defines the vision of the agent. For example, the implementation of a predator–prey model requires both the predator and the prey to have defined a maximum distance for which they may observe other agents in the landscape. In other cases, interactions may be based on the evolution of a social network, a change in household dynamics based on its life cycle, or observation of the behaviors of spatial neighbors (adjacent or farther away). The result of these interactions can be a change in the flow of information, interdependencies between and impacts of agent actions, and the sequencing of model events.

Event sequencing within an ABM plays an important role. In ABMs, event-sequencing mechanisms can generally be characterized as predetermined and event-driven. Predetermined event-sequencing rules are set by the modeler, and generally operate similarly over model runs. Such rules can be used to simulate synchronous behavior, where each model agent has an opportunity to make a decision, taking the last previous action of all other agents as a starting point. Rules can also be asynchronous, where, for example, each agent has a fixed probability of being active in any time period. In asynchronous implementations of agent behaviors, model results can be dependent on the order in which agents act. To alleviate this modeling artifact, most ABM platforms provide a shuffle routine that rearranges the order of agents within the population that takes actions. Event-driven rules, in contrast, dictate action when a state variable in the system – for a single agent, another agent, or the system as a whole – reaches a certain threshold. For example, a farm may be sold when the owner retires, or a family may seek a larger residence when they have children.

ABMs are also generally constructed with mechanisms to track and record information
about model states and flows. Such information can be stored in many formats—GIS layers, text files, formal databases, and others. Software and libraries designed for building agent-based models often have such capabilities built in.

When ABMs are applied to models of geographic systems

Spatially explicit ABMs have been used by researchers in a variety of different disciplines to gain insight into space-dependent problems (e.g., geography, economics, planning, political science, ecology, anthropology). Examples of spatial applications include: evaluating the effects of different types of policies on the patterns of urban development and their subsequent ecological impacts; estimating conflict and improving route scheduling for tourists in national parks; quantifying changes in crop yields due to demographic and life cycle changes in farming households; comparing social and environmental factors affecting the extinction of communities; quantifying human responses to emergencies and assessing alternative designs to affect evacuation patterns; and defining territorial boundaries for nations within states. A large suite of recent applications are presented in Heppenstall et al. (2012).

In each of these domains of study, spatially explicit ABMs can be better designed and understood by conceptualizing the sources of spatial, temporal, and behavioral complexity in the system of study and captured in the ABM. The majority of geographic applications include one or more sources of spatial complexity. While not comprehensive, spatial complexity can be broken down into three areas: spatial autocorrelation, zonal spatial relationships, and network relationships.

- **Spatial autocorrelation**, at a general level, follows Tobler’s first law of Geography: “All things are related, but nearby things are more related than distant things.” Spatial autocorrelation refers to cases in which a spatial state is either more or less likely when in proximity to a similar spatial state. In human systems, spatial autocorrelation can occur for many reasons, including imitative behavior, diffusion processes, synergies between compatible activities, spatial competition, and negative spatial spillovers. It implies a causal spatial relationship between locations.

- **Zonal spatial relationships** occur when locations share a common characteristic or state variable value, for a noncausal reason—for example, similar levels of accessibility, political or zoning designation, soil type, climate, or topography. Often different zonal relationships are non-nested (for example, political and watershed boundaries), creating measurement and modeling challenges.

- **Spatial networks** define spatial relationships between entities that are constrained by an existing structure in the landscape (e.g., roads, pipelines, forest corridors). Unlike measurements that are based on Euclidean distance (i.e., what is commonly referred to as “as the crow flies”), spatial networks provide a realistic account of distance and include impedances that can slow the travel of processes through the network. Networks can be physical, such as transportation or hydrologic networks, or social, such as communication, relationships, and trade and commerce networks. Concepts and algorithms from network analysis are commonly applied in ABMs.

Temporal complexity shares many of the same characteristics as spatial complexity, but instead of quantifying the behavior of a process across space,
it is concerned with quantifying the process over time. Temporal autocorrelation measurements are used to quantify the dependency of future measurements of a phenomenon based on past measurements. Geographical systems are often characterized by important temporal growth and decay processes. These processes can be social, such as population growth and decline, social trends, and growth of financial investments. They can also be biophysical, such as soil degradation, carbon sequestration, and erosion. Such processes lead to temporal autocorrelation, where a given time period’s state depends on that of the last time period.

When combined with spatial autocorrelation and its various drivers, processes will affect state variables across both time and space, such as species colonization and fire spread. Such processes are often modeled using cellular automaton models, which are implicitly contained in many ABMs. Further, Hagerstrand’s time geography approach is increasingly applied to monitoring individual and group behaviors in agent-based models and provides one of many geovisualization techniques to evaluate agent behavior in an ABM.

Behavioral complexity in spatially explicit ABMs refers to the degree of complexity incorporated in representing how agents make decisions as well as the complexity of social and institutional structures represented in the model. With regard to decision-making, a range of approaches are used that span from simple heuristic “if-then-else” algorithms, to utility equations, and learning and adaptation algorithms (e.g., genetic programming). While often omitted from ABMs of geographical systems, forward-looking behavior by humans and other animals plays an important role in decision-making. Examples include food and seed storage, crop rotation, strategic behavior, such as preemptive colonization, and financial investments. To account for such behavior, models must build in expectations of the future and possible responses in order to develop models of learning and adaptation.

Behavioral actors in models are often characterized by heterogeneity and interdependencies. Not only are different types of actors present in most systems, even within a type, actors may differ in their goals, expectations, strategies, and motivations, as well as their internal and external resources. Actors are also interconnected in social, economic, and ecological networks. Collections of actors may also have their own rules of behavior.

The representations of the spatial, temporal, and behavioral complexity of a study system in an ABM are each characterized by a notion of scale, and complex processes for each can operate at multiple scales. Spatial scale is characterized by spatial resolution (e.g., minimum mapping unit) and spatial extent (i.e., total area represented), temporal scale by temporal resolution (i.e., length of the smallest unit of time represented, also known as the time step) and temporal extent (duration of time represented), and behavioral scale by the minimal decision-making actors and collectives of such actors.

Processes can be characterized by feedbacks across scales in each case – for example in a spatial system, between neighborhoods and a city, in a temporal system, between fast processes such as weather events and slower processes such as climate change, and in a behavioral system, by interactions between family members and a household. Cross-scale processes can operate both from the bottom up and from the top down, with feedbacks between scales.

The collective result of these forms and sources of complexity is that systems generally modeled using ABM are highly path dependent. A small change in an initial condition or stochastic seed at a given point in time may move the system to
AGENT-BASED MODELING

a qualitatively different outcome. Modelers need to be aware of this sensitivity, testing models using a wide variety of initial conditions and/or stochastic seeds to trace out distributions of outcome variables that may emerge from the modeled system.

Model development: elements and process

While the development of ABMs often proceeds iteratively, ideally based on scientific and stakeholder feedback, the following template provides a framework that can be usefully applied to the development of most ABMs of geographical systems.

First, the purpose of modeling should be identified. While many researchers find that the process of model construction can be useful to codify existing knowledge and gain consensus as to an appropriate representation of the real-world system, ideally model design should be framed by the clearly defined research questions of what the model is for and what questions its output can be used to answer.

Next, researchers typically identify the real-world spatial, temporal, and behavioral processes they strive to represent. ABMs are structural models, which embed hypotheses about process-based relationships and causal mechanisms. The process of identifying these relationships and mechanisms in the applied problem is critically important. Often, it can be very useful to create simple system diagrams, noting the scale of operation of each process and identifying key causal linkages via arrows and feedback loops. Having done this, researchers can identify what factors and processes they might reasonably assume are exogenous (determined from outside the model — information you have before you run the ABM) and endogenous (determined as part of the model run — information you did not have before you ran the model). It is important within this process to try to distill the model elements to the minimal possible level of complexity that you can practically model. While the first question to ask is, “How can we characterize the complexity of the real-world system?” the second is, “How little complexity can we get away with modeling?”

Having undertaken this process of identifying real-world structures and mechanisms, if building an empirically supported model, the team should identify the ideal spatial, temporal, and behavioral representation (both structure and resolution) for model parameterization, calibration, and assessment data. As with any empirical undertaking, this phase is generally followed by a practical assessment of the data that might actually be available for model construction. This is generally followed by a second round of model structuring during which the modelers simplify the model plan and goals. Empiricizing ABMs can be particularly challenging, as they require data about both actors and the actors’ environment, in contrast to some geographic models that rely on spatial data alone, or social science models that use only actor data. The fine-scale, disaggregated nature of ABMs creates further challenges.

Having made high-level decisions regarding what processes to model and what data to use, modelers must turn their attention to detailed decisions regarding model design, model rules, and system architecture. While no standard texts or templates have been available historically, experienced researchers are beginning to offer texts to guide learning and model development (see below for readings). Decisions regarding software implementation must be made. In spite of community consensus that such tools could be useful, to date few standard software packages and model libraries are available. However, two sets
of software libraries are commonly used: RePast and Netlogo. Further, the Community Modeling Library hosted by the CoMSES (Computational Modeling in the Social and Ecological Sciences) Network provides code examples of previously implemented models (http://www.openabm.org/models). Journals are also providing links to model code archives with increasing frequency. Along with a push towards code sharing, many scholars have emphasized the need for standard model documentation protocols and metadata. Various approaches have been taken to the problem. Along these lines, research groups should set forth protocols for documenting, archiving, and sharing their model data and code, with access rights within and external to the research group carefully defined before modeling begins. Often such agreement might be dictated and/or constrained by institutional and access agreements. Such points should be considered when choosing data and software options.

The value of ABMs

ABMs offer many advantages. While often seen first and foremost as models, ABMs offer value by encouraging discussion and learning in a number of ways. First, even before an ABM is coded, the design phase can act as a medium for discussion and can formalize a range of assumptions and knowledge about a study system. This process continues as assumptions are formalized in computer code. This modeling process often acts as a bridge across disciplines, facilitating communication and project developments on large interdisciplinary research teams. Second, the ABM offers a proof of concept and a pedagogical framework to demonstrate the effects of different processes in isolation or in combination. Third, because ABMs are built with a focus on representing process, behavior, and interaction, they often cause data to be analyzed and interrogated in new ways. They may also highlight new types of data that have not been collected but are necessary to understand a question being answered.

Beyond their pedagogical and conceptual contributions, ABMs are valued for their ability to provide evidence to support general and overarching advances to our scientific understanding of a broad range of human and natural systems. For example, Schelling’s segregation model demonstrates that society may segregate from households having only a small preference for neighbors that are like them (e.g., 25%). New models have represented sheepdog behavior with only two agent rules and offer approaches for corralling crowds. Others have shown that simple behaviors can solve complex problems (e.g., ant-agents searching for food and leaving pheromones can solve n-polynomial optimization problems). These are but few of the many hundreds, if not thousands, of systems to which ABM are (or could be) applied.

Limitations and challenges

In spite of the successful applications described above, the development, analysis, and communication of ABMs present challenges. These challenges are widely recognized by the modeling community, and incremental and steady progress is being made to address them.

On the model development side, standard and easy-to-use software libraries, such as those available for statistical analysis, are not yet generally available for ABMs. This creates a barrier to entry to the field, as generally some programming abilities are needed to construct and modify an ABM. However, increasing availability of software packages such as Netlogo and introductory texts create a fairly accessible entry point to programming and modeling. Even for the proficient
nombre, there is a lack of standardized libraries and modeling algorithms available. In
spite of a recognition that it may not be the most efficient use of resources at a community
level, most projects start from scratch to build their own models, rather than reusing existing
model components. Thus, model comparison, evaluation, and replication are hindered.

Due to their fine-scale design, ABMs can be computationally demanding. This can create
a particular challenge in terms of scalability, especially in cases where the ABM is intended
to provide an input into a higher-scale process model. For example, an ABM may be used to
model a local economy, with the input designed to feed into a regional model. One option to
address this challenge is to explore methods to fit metamodels, or aggregate functions that rep-
resent model behavior, to the output of ABMs, then use the metafunction as an input to the
higher scale model. For example, ABMs have been developed that replicate the smooth, regular
system dynamics seen in analytical epidemiological models and predator–prey models. Further,
conceptualizing the event-sequencing methods and the modes of interaction among actors and
processes acting across and between spatial and temporal scales remains an ongoing challenge.

As mentioned above, empirical ABMs can have high data demands, especially when they com-
bine spatial and actor level data. Both cost and confidentiality concerns can be barriers to data
access. The need to separate data used for model development from that used for validation creates
further data demands. As ABMs become more empirically informed, they have become more
site specific and have rarely been applied across multiple case-study sites, limiting their ability to
demonstrate broad scientific achievements.

Once a model is constructed and verified, model analysis can be challenging due to the
complexity of model design. While fairly simply

protocols have been developed for interpretation of alternative modeling methods, such as comparative statics and dynamics for analytical equilibrium models, and interpretation of estimated coefficients for regression models, standardized methods for understanding relationships between the exogenous and endogenous elements of ABMs are not yet developed. Given the purposeful pursuit of nonlinearity and complexity in ABM, it can be especially challenging to map, interpret, and understand model behavior over the entire potential parameter space of the model. From a scientific perspective, this leads some to question the validity of the sim-
ulation modeling effort, as the global behavior of models is not well understood. A corollary
to this challenge is that it can be challenging to communicate complex model mechanisms
and interpret outcomes for nonacademic audiences, such as local stakeholders, planners,
politicians, and policymakers. Participatory or companion modeling methods, in which models
are co-designed and analyzed, can potentially address this challenge.

SEE ALSO: Agents, ontology-based
decision–support systems; Cellular automata;
Cognition and spatial behavior; Exploratory
spatial data analysis; Geocomputation;
Geography and the study of
human–environment relations;
Microsimulation; Representation: dynamic
complex systems; Spatial analysis; Spatial
interaction; Time geography and space–time
prism; Tobler’s first law of geography

Reference

Heppenstall, A., A. Crooks, L. See, and M. Batty.
Dordrecht: Springer.
Further reading


At the end of 2008, during the world financial crisis, the CEO of IBM proposed a vision called Smart Planet (http://www.ibm.com/smarterplanet/us/en/), indicating a major direction for the application of information. In addition, the proliferation of the Internet of Things has contributed to the formation of a “central nervous system” across the world, connecting all people and things using the Internet. Currently, populated by numerous sensing devices and more than one billion users, the Internet has aggregated a vast amount of data. Although users are frequently overwhelmed by the multiplicity of information available, they lack access to the information and knowledge required for decision-making. This is a paradoxical situation widely prevalent in workplaces. IBM holds that a solution to this dilemma lies in integrating intelligence and knowledge into the ocean of data to render all data locatable, allowing all information and analytical tools that decision-makers need to be made available efficiently with the assistance of intelligent agents.

How can intelligence and knowledge be obtained? Currently, the information science community has developed theories and instruments of ontology development to enable the systematic organization of human intelligence and knowledge (Hadzic, Chang, and Wongthongtham 2009). Ontology enables easy sharing of intelligence and knowledge, and assists in rapidly locating information and data. Ontology also enables the accumulation of research results and assists researchers in identifying worthwhile research topics. However, ontology alone is insufficient; intelligent-agent-related technologies are required to combine ontology with intelligent agents to maximize benefits.

If ontology is analogous to the human brain, intelligence agents are analogous to the hands and feet. Ontology supplies the intelligence and knowledge required in the decision-making process, while intelligent agents provide the driving force to solve problems.

Intelligent spatial decision-support systems (ISDSSs) are a critical research direction in geographic information science and are desired for use in governmental and corporate decision-making. In this entry, developmental trends in spatial decision-support systems, and the development of ontologies and intelligent-agent-related technologies, are addressed. A framework for ISDSS development is also proposed.

Developmental trends in spatial decision-support systems

Spatial decision-support systems are a major item of development in geographic information science (Armstrong and Densham 1990; Sikder and Gangopadhyay 2002; Sengupta and
AGENTS, ONTOLOGY-BASED DECISION-SUPPORT SYSTEMS

Bennett 2003; Sugumaran and Sugumaran 2003; Sugumaran and DeGroote 2011). As a result of the upsurge in broadband networking in recent years, geographic information system (GIS) databases have been used to provide web services, and spatial decision-support systems have been integrated into the Internet. In addition to geographic databases, model bases and knowledge bases are required in spatial decision-support systems. Evolving cloud computing technology may provide an environment for information computing that enables complex model computing. Therefore, the establishment of model bases and knowledge bases is a key concern in current development of ISDSSs. Good results in information science have been achieved by constructing knowledge bases using ontology, and continued efforts are being made to develop large-scale knowledge bases and model bases (Sugumaran and DeGroote 2011).

The development of ontology and Semantic Web technology

Advances in geographic information science and technology gave rise to the development of GIS, software, and web services (Goodchild et al. 2005). However, most geographic information has been developed by different research institutes or government organizations with different software and data formats, resulting in sizable difficulties when integrating these GIS, especially in regard to data formats, coordinate systems, or semantic differences. In recent years, the Open Geospatial Consortium (OGC) has been committed to publishing geographic information in standardized web services such as web mapping services (WMS) (OGC 2006b), web feature services (WFS) (OGC 2005), and web coverage services (WCS) (OGC 2003). The disparity in data formats can be resolved by standardizing service-oriented architecture (SOA) web services. The European Petroleum Survey Group (EPSG) (www.epsg.org) is committed to recording a standardized coordinate system format as a reference for coordinate conversion. However, semantic integration of data cannot be achieved by standardized web services.

The semantics of data refers to the meanings or concepts underlying data. In information communication, text and graphics are used to convey meanings or concepts, in part or in entirety. In different cultural contexts, different words are used to express the same meaning or concept. As a result, geographic information generated in different contexts typically leads to semantic problems in data integration. Therefore, the effective integration of semantic issues with respect to geographic information has been one of the most popular research topics in the field of geographic information in recent years (Egenhofer 2002; Kuhn 2003).

“Semantic Web” is a term coined by Tim Berners-Lee, the inventor of the Internet (Berners-Lee 2000). The Semantic Web uses the XML format to record information and describes the semantic relationship between documents; it enables computers to understand the semantics underlying information and automatically derive and query semantic meanings associated with that information. By denoting semantic information and the relationships between linked information, a web of knowledge semantics is formed on the Internet, which is known as a Semantic Web. The application of the Semantic Web to the geospatial field has resulted in the formation of geospatial Semantic Webs (Egenhofer 2002; Fonseca and Sheth 2002; Kolas, Hebeler, and Dean 2005) and the semantic reference system (Kuhn 2003).

Egenhofer (2002) defined the geospatial Semantic Web as “a web that retrieves, analyzes,
and links geographic information semantics to achieve geospatial semantic interoperability,” and meets the following requirements.

1. The creation of spatial and temporal ontologies: ontology is an explicit specification of a conceptualization (Gruber 1993) that can be understood and derived by computers. This conceptualization includes the concept definition, the relationship between this concept and others (e.g., a hierarchical relationship), and related vocabulary derived from this concept. Ontology can be perceived as an electronic dictionary that not only maintains records of terms’ definitions, but also includes all concepts and knowledge in the field. It describes and records in detail “definitions of concepts,” “relationships between concepts,” and “related vocabulary contained in the relevant concept.” Thus, the convertibility of geospatial semantics is achieved by creating spatial and temporal knowledge ontologies, in which detailed descriptions of various concepts are provided.

2. The explicit presentation of semantics: concepts recorded in ontologies can be quickly and clearly comprehended and derived by computers or humans.

3. Searchable spatial statements: spatial statements can be queried in the ontology, and whether the results of queries satisfy the query statements can be evaluated.

Research on geospatial Semantic Web applications in recent years can be broadly classified into the following categories: (i) research concerned with achieving interchangeability of geographic semantics (Bishr 1998; Harvey et al. 1999; Egenhofer 2002; Fonseca and Sheth 2002; Kuhn 2003; Lemmens et al. 2006); (ii) research concerned with investigating geospatial ontologies to establish definitions of and relationships between various concepts, and constructing geospatial ontologies (Egenhofer and Mark 1995; Mark et al. 1999; Smith and Mark 2003); (iii) research concerned with strengthening geospatial semantic retrieval and searching, such as in the project “A spatially aware search engine for information retrieval on the internet” (SPIRIT) (Purves et al. 2007); (iv) research concerned with the application of a spatial data infrastructure (SDI) to locate appropriate GIS web services, including web services incorporating GIS data and analysis methods (Bernard et al. 2003; Lutz and Kolas 2007); and (v) research concerned with the application of geospatial ontologies that are (semi-)automatically combined with GIS web services (Di et al. 2006; Lemmens et al. 2006; Yue et al. 2007).

The development of intelligent agent technology

Agent technology is an area in information science that has rapidly developed in recent years. Agent technology enables information systems to dynamically exchange and share information through agents, and accomplish tasks and goals in collaboration. Agents are characterized by autonomy, sociability, reactivity, and proactiveness, and can actively assist users by locating useful information on the Internet to complete the tasks assigned. However, without the assistance of ontologies, the effectiveness of agents is severely limited. After the integration of agent technology with ontology technologies, agents gained access to domain knowledge and became intelligent agents (Hadzic, Chang, and Wongthongtham 2009). To fully leverage its effectiveness, an ISDSS must incorporate GIS, ontology, and agent technologies.
Framework for intelligent spatial decision-support systems

The application of geospatial Semantic Webs or ontologies to improve or solve problems pertaining to geospatial semantics regarding spatial decision-support systems has not been adequately explored in the aforementioned studies. Traditionally, spatial decision-support systems operate in stand-alone mode, in which all GIS data and analysis to be used are pretreated (including GIS file formats and semantics) and saved in a stand-alone version; although current GIS web services like WMS and WFS can be applied in spatial decision-support systems, solving the GIS file format problem, geospatial semantic interchangeability is critical because of the varying semantics used by different web services. In addition, different spatial decision-support systems established on the basis of different topics may use the same GIS data and analytic methods repeatedly. Repeated use of the same GIS data and analysis may result in imperceptibly increased costs.

Therefore, proposed here is the use of a geospatial Semantic Web and geospatial ontology to improve the spatial decision-support system (Jung, Sun, and Yuan 2013; Jung et al. 2013). Detailed information about the intelligent spatial decision support system (ISDSS) framework can be found elsewhere (Jung et al. 2013). The objectives were as follows.

1. To propose a knowledge-oriented structure based on geospatial semantics, and solve geospatial semantic problems using geospatial ontologies.
2. To improve spatial decision-support systems. Based on GIS web services, geospatial ontology automatically locates and infers reasonable web services in response to various decision-making problems. Assembling is performed to generate preliminary results, which are subsequently provided to decision-makers as a reference for further evaluation. This resolves the problem of repeatedly using the same GIS data and analysis methods.

Figure 1 depicts an ontology-based information structure that consists of four components: a portal, a geospatial ontology, SDI, and web services.

Portal

The portal serves as an interface for human–computer interaction between decision-makers and the framework, and provides a web map service and registration platform for decision-makers and system developers to input spatial decision questions, including questions regarding spatial relationships and spatial scope. Any problems raised by the decision-maker are transferred from the portal to the ontology for derivation. After the appropriate web services are located, assembling and execution are performed to generate preliminary results, which are finally presented in the portal to provide decision-makers with references for further evaluation. Service providers may register on the portal to make their services available to policymakers.

Geospatial ontologies

Geospatial ontologies are the core of the structure and are responsible for receiving question inputs from decision-makers, subsequently deriving appropriate solutions and GIS web services. Geospatial ontologies that are created by experts who differ in their knowledge of geospatial domains are expandable and can be connected to each other, enabling the sharing of concepts between different ontologies.
With reference to Kolas, Hebeler, and Dean (2005), the following geospatial ontologies were constructed.

1. **Geospatial geometry ontology**: basic geometric geospatial ontologies are constructed based on the definition of the OGC simple feature specification (OGC 2006a). This includes geometric properties like points, lines, and polygons (Figure 2).

2. **Geospatial web services ontology**: common definitions and hierarchical structures for GIS web services are provided, such as WMS, WFS, and WSDL (web service definition language) services (Figure 3).

3. **Geospatial analysis ontology**: because of the present lack of a complete, unified classification system for GIS analysis, the classification system of the ArcToolbox featured in the ArcGIS 9.3 software was referenced to construct the ontology of geospatial analysis methods. The input, output, precondition, and results (IOPR) of each GIS analysis method are completely recorded in the ontology (Figure 4).

4. **Domain ontology**: domain ontology is a domain knowledge classification structure developed by experts in the field. The global Earth observation system of systems (GEOSS) addresses nine areas, including...
disasters, health, energy, climate change, ecosystems, agriculture, biodiversity, water resources, and meteorology. The content of the domain ontology is gradually built up through collaborative efforts by experts in the field mediated with GEOSS (www.ogcnetwork.net).

5 Task ontology: Timpf (2001) emphasizes that task ontology is as crucial as domain ontology. Task ontologies focus on various tasks in a given area and provide detailed descriptions of various concepts and relationships associated with that task, thereby simplifying and organizing large-scale domain knowledge. In task ontologies using this structure, each task represents a different spatial decision problem; the concepts required for solving the task and the relationships between these concepts are defined in detail. Based on GIS analytic methods, the concepts and their relationships provide a conceptual workflow.

As shown in Figure 5, when the user selects Task A, the task ontology may derive three analytical methods (i.e., Function1, Function2, and Function3), which are connected in series by hasNext relationships. For example, the Function1 hasNext Function2, and Function2 hasNext Function3.

Spatial data infrastructure (SDI)

Knowledge derived by geospatial ontology is conceptual and is not directly connected to actual web services. For example, the three functions derived in Figure 5 belong to a conceptual workflow, where conditions satisfied by each function must have been predefined in the ontology, although not directly connected to the web service. Thus, the derived functions and their associated conditions must be mapped to appropriate web services by searching within the existing SDI. During the construction of the SDI, geospatial web services are provided by the
Figure 3  Geospatial web services ontology.

Figure 4  Geospatial analysis ontology.
data provider, and the web services are registered to the portal.

**Service chain mediator**

After the appropriate web services are located, the conceptual workflow is located within the task ontology through the assistance of the service chain mediator using agent technology. Following assembling and execution, the preliminary results are transferred back to the portal, providing a reference for further evaluation by decision-makers.

The information framework presented here employs ontology technology to effectively integrate various geographic information services offered by government, academia, and private firms. The government is responsible for providing data services across the databases of SDI, academic institutes are to provide data analysis and simulation model services, and private firms are to provide various application services to government agencies and the general public. This enables complex environmental problems faced by decision-makers to be solved through joint efforts.

**Conclusion and recommendations**

The development of a “Smart Planet” is currently a major task faced by the GIS community. Proposed here is a framework for an ISDSS that may be used to accelerate the implementation of the Smart Planet initiative. In addition to developing geographic databases, an ISDSS also requires the development of knowledge bases and model bases. In the next phase, building a knowledge database and model bases is critical. Adopting the framework introduced here, ISDSSs can be constructed to comprehensively improve the quality and efficiency of the government’s policies.

**SEE ALSO:** Geospatial Semantic Web; Location-based services; Ontology: domain applications; Open Geospatial Consortium standards; Spatial decision-support system
References


AGENTS, ONTOLOGY-BASED DECISION-SUPPORT SYSTEMS


Aging

Christine Milligan
Lancaster University, UK

At a population level, aging refers to a shift in the age distribution over time towards older ages. The increased number and proportion of older people is measured on the basis of increased life expectancy at birth and declining birth rates coupled with migration.

The definition of old age against which population statistics are measured can vary. In many high-income countries retirement age (generally still taken to be 65 years of age) is the baseline against which population aging is calculated. However, at a global level life expectancy rates vary significantly; furthermore, the concept of retirement as a baseline for calculation may have little meaning (for example, in low-income countries). As a consequence, population aging may be measured on the basis of 60 years and over (see for example the World Health Organization).

Measuring aging on the basis of those aged over 60 or 65 years covers a wide age range. As a consequence, distinctions are often made between the “young old” (those aged under 75 years), the “mid-old” (those aged 75–84) and the “oldest old” (those aged 85 years and over). The terms “third age” and “fourth age” are also commonly referred to in the literature. Retirement is usually seen to mark the beginnings of the third age. There is less agreement on what marks the beginnings of the “fourth age” but most often it is used to refer to those aged 85 and over; in most high-income countries this demarcates the stage when most individuals have lived beyond average life expectancy. While these figures and terms are somewhat arbitrary, the distinctions are useful for research and policy as the oldest old are most likely to experience the highest population levels of disability arising from chronic disease that requires long-term care and support.

As perhaps the major demographic challenge of the twenty-first century, population aging will have profound consequences and implications for all places and all aspects of human life. While, on the one hand, global aging represents a triumph of human achievement in the medical, social, and economic spheres, on the other, it presents challenges to existing systems and institutions at all spatial scales that will affect nearly all aspects of society. In the social sphere, population aging will impact public services, health, care and wellbeing, family composition and living arrangements, housing, and migration. In the economic sphere, it will impact on economic growth, savings, investment and consumption, labor markets, pensions, taxation, and intergenerational transfers. In the political arena, it will influence voting patterns and representation.

The historical and ongoing shift in the aging of the population makes it important to understand how these social and economic implications are being played out in different places for different groups of people. The sheer breadth of issues affecting, and affected by, this demographic shift has led to the study of aging being addressed across a wide range of social science and health-care disciplines – traditionally through the multidisciplinary study of (social and critical) gerontology.

Those working in this field have done so from a variety of perspectives and at a range of spatial
AGING

scales stretching from the global to the local and individual.

The spatial demography of aging

Globally, the population is undergoing a significant transformation that is unprecedented in human history. The current bulge in the middle-aged demographic, together with ongoing improvements in life expectancy, means that the population of older people is projected to rise significantly, increasing from 530.5 million in 2010 to 1.5 billion in 2050. These projections point to a much older world, and a future in which approximately one in six of the global population is expected to be 65+ by 2050. This situation will double the current proportion of those aged over 65 years. Population aging is not just a feature of high-income countries; however, considerable variations do exist between regions, countries, and even within countries. Population geographers have been concerned to examine these patterns, their causes, and the implications of this spatial unevenness, and more recently to understand the why and where of spatial trends on aging.

At a regional level, Europe and North America have the highest proportion of their populations in the 65+ age group and the lowest proportion of those aged under 19 years, whilst Africa and South America exhibit the reverse population patterns. Some variation does occur within these regions, but overall the figures point to Europe as the “hotspot” of global aging, although Japan too figures highly. Population projections, however, suggest that by 2050, six of the top 20 countries/regions with the highest proportion of those aged 65+ will be Asian, including Singapore, Japan, Hong Kong, South Korea, Taiwan, and Macau. African countries will continue to have the lowest proportion of those aged 65+. It is important to recognize that these estimates are based on the proportion of the population and not actual population numbers. So while countries such as China, India, the United States, India, and Brazil may have lower proportions of older people in their populations, the actual numbers of those aged 65+ in these countries will far exceed those of European countries. As McCracken and Phillips (2005) point out, while currently these countries account for 59% of those aged over 65 years, by 2050 this figure is projected to rise to 80%.

Currently, the oldest old constitute 7% of the global population over 65 years of age. This percentage is not evenly distributed, with more developed countries having approximately 10% of the over 65s in this age group and less developed countries having around 5%. In many countries the oldest old are the fastest growing proportion of the population. This trend is most striking in Japan where it is estimated that by 2030 nearly 24% of the population will be over 85 years of age. Globally, the National Institutes of Health in the United States have estimated that between 2005 and 2030 this age group will have increased by 151%.

As life expectancy increases, causes of death will shift from infectious and parasitic diseases to chronic and noncommunicable diseases such as heart disease, cancers, and diabetes. This situation has led to debates about whether we can expect to see a compression of morbidity, where medical advance means that the young old are likely to live relatively healthy and active lives, delaying the onset of multiple health problems and disability to old old age. Others, however, maintain we will see an expansion of morbidity with an increase in the prevalence of chronic illness and disability across the total of our older populations as life expectancy increases.

The demographic transition model (DTM) is one approach that has frequently been drawn
upon to explain global patterns of aging. First formulated in the 1940s by Frank Notestein and refined over time, the model describes temporal shifts in birth rates and death rates and links the resultant changes in the population profile to transformations in socioeconomic conditions across four stages of shifting birth and death rates:

- **high stationary**: high birth and high death rates due to high infant mortality, poor medical knowledge, famine, war, and epidemics;
- **early expanding**: population growth as birth rates remain high but death rates decline significantly due to improved sanitation and access to clean water, better nutrition, and increased medical knowledge;
- **late expanding**: population growth slows due to declining birth rates resulting from shifts in attitudes to birth control and family planning and the gradual stabilization of death rates at a low rate due to continued improvements in medical technologies and nutrition;
- **low stationary**: low and stable birth and death rates due to: continued family planning and later marriages, the improved status of women; a shift in the main causes of death from disease and epidemics to chronic diseases of affluence such as heart disease, cancers.

A further stage has also been added to the model, based upon recent shifts in some (mostly European) countries that are now experiencing both an aging population and population decline as death rates remain stable but birth rates fall below replacement rates. Some argue that this disequilibrium will be balanced by controlled in-migration. Drawing on evidence from Europe and the United States, Coleman (2006) cautions that if this continues, increased rates of in-migration will eventually lead to a displacement of the original population, both by age profile and ethnic composition, leading to a radical and permanent alteration of the national population in which the original population would hold a minority position.

The DTM was modeled on relatively slow shifts over time in European countries. In many newly developing countries the shift through the stages of the model has occurred much more quickly due to their ability to take advantage of current public health knowledge and advanced medical and other technologies.

This model is not without its critics, who cite evidence of reversals in the aging of populations in countries such as Russia, Botswana, Zambia, and Zimbabwe where life expectancy at birth has fallen below 60 years of age. In Russia this fall has been attributed to social and political upheaval and economic collapse in the post-USSR period. In the African countries it has been largely attributed to war, famine, and the HIV/AIDS pandemic. These examples illustrate that the transition is not one of inevitable progression and improvement but can be subject to reversals. Using the example of Europe, the region with the world’s oldest population structure, it is also clear that cultural and political factors – such as governments’ approaches to migration, health care, and welfare – are also at play. These issues will be particularly significant for newly developing countries where the pace of population aging may make it difficult for them to adjust their economic, health, and social care systems to cope.

Despite contestation about the DTM, as a model it has proven remarkably resilient as a means of explaining observed spatial unevenness in population aging, though it is less useful as a model for explaining observed spatial variability within countries (see Population growth).
Environment and geographical gerontology

Cutting across both disciplinary and subject contexts there is consensus within aging research that understanding the environment and the individual’s location within it is a crucial factor in understanding the aging process. Both environmental and geographical gerontologists place emphasis on the importance of understanding the ordinary everyday contexts that older people inhabit (Scheidt and Schwarz 2013). This situation is not new; rather, a sustained body of research around older people’s relationships with their social and physical environments stretches back at least 40 years to the early work of M. Powell Lawton, Stephen Golant, Graham Rowles, and others. Golant, for example, looked at future housing options across a range of different scales from segregated housing projects for older people to retirement villages – developments that have sparked ongoing debates about the “ghettoization” of old age versus the development of supported and protected living environments.

Many of these early geographical and environmental gerontologists adopted behavioral and humanistic approaches to understand the ways in which older people’s environmental experiences differed from those of younger populations. Lawton and Nahemow’s (1973) model of “environmental press” is seen as a major contribution to understanding person–environment transactions in later life; here, optimal environmental fit is seen to occur when an older person’s abilities and capacities are consistent with the demands and opportunities within that person’s environment. If the demands of the environment exceed those of the person and their abilities (for example as an older person’s health or mobility declines), a person–environment misfit occurs, triggering the need for support to enable the older person to age in place successfully.

Aging in place continues to be a key theme in environmental and geographical gerontology. This involves understanding older people’s experiences of, and engagement with, domestic and community environments, the characteristics of these environments, and how they influence an older person’s quality of life, health, and mental wellbeing. A particular concern has been to unpack the continually changing relationalities between older people and their sociophysical environments and how these act to shape the human experience and processes of aging. The assumption here is that as people age, their physical and/or cognitive abilities change or decline, resulting in shifting – and often more vulnerable – relationships with the environments they inhabit. Rowles’s seminal work “Prisoners of Space” set a marker for this field of enquiry. His detailed ethnographic account of the lived experience of older people’s lives painted a rich and detailed narrative of the ever-decreasing lifeworlds that people inhabit as they age.

The work of early geographical and environmental gerontologists has stimulated a deeper engagement with older people’s social and physical experiences of, relationship with, and exclusion from, the communities they inhabit. This scholarship has mainly focused on issues of access and mobility within these communities and on what facilitates and excludes – stretching from design and infrastructure to social and structural ageism and inclusion. This work is exemplified by the emergence of a body of conceptual, empirical, and policy-oriented work around the development of age-friendly environments. Recognizing the potential challenges of our aging societies, for example, the World Health Organization has set up a global network of age-friendly cities and communities. These are designed to foster...
worldwide mutual learning about how to create inclusive and accessible urban environments that promote healthy and active aging and a good quality of life for their older residents. Similar initiatives also exist around the creation of dementia-friendly environments.

Recognition of the decreasing lifeworlds that older people inhabit, combined with policy shifts towards aging in place, has also highlighted the importance of understanding the domestic sphere and the (sometimes changing) meaning of home for older people (see Home). As a site removed from public scrutiny, the home is seen as a place where the older person can control decisions about whom to include or exclude. Such sites can provide an important buttress to an older person's sense of self and independence, particularly for those who may feel vulnerable outside the bounds of their own private spaces. The home is seen as a site of security, familiarity, and nurture that is imbued with personal meaning and which can promote successful aging in place through what Rowles later referred to as a preconscious sense of setting. That is, temporal knowledge of the home, combined with physical attachment to it and the routines performed within it, can facilitate an older person's ability to negotiate that space without coming to harm, even as physical or cognitive abilities begin to decline. But increasing frailty can lead to a breakdown of this preconscious sense of setting as the requirement for care and support increases, bringing attendant technologies of care (such as hoists, ramps, commodes, wheelchairs, etc.). Increasing dependence on both formal and informal support to maintain activities of daily life can also result in a declining power to exclude from even those most personal and private areas of the home. This process highlights the juxtaposition of private and public space and the shifting relationships of power, independence, and autonomy that accrue for older people across a continuum of physical and cognitive abilities as a consequence of these shifting juxtapositions (see Twigg 2000; Milligan 2009).

The importance attached to the meaning of home and identity has also played a significant role in work around residential care settings for older people. Here, emphasis is placed on whether it is possible to design residential spaces that facilitate the development of a sense of place or home, and the kinds of lifestyles such spaces foster and support (e.g., Peace and Holland 2001).

Much of the ongoing work in this field is grounded in conceptual and theoretical underpinnings around identity and place attachment, environmental meaning, Therapeutic landscapes, and community engagement in later life.

Landscapes of aging

Work around population aging often characterizes the older population as homogeneous in terms of their demands for social and welfare support, health care, their levels of participation in society and the workforce, and so forth. Older people are typically portrayed as more likely to place pressure on health and welfare systems and less likely to be net contributors. Such an approach fails to recognize difference and negates the considerable contribution older people make to the social, economic, and cultural landscapes of our societies.

In recognizing the heterogeneity of aging populations it is important to note that particularly in young old age, many older people in high-income countries will be healthier and better educated, and have greater disposable income and greater expectations about how they will spend their retirement years than in the past. Early retirement may be spent taking up new challenges, finding more time for existing hobbies and activities, or taking up activities once
AGING
deemed to be no longer open to older people (e.g., continued education, long-distance travel, fashion), opening up new consumer niches and new dedicated activity clubs. There is also evidence of a growing pattern of mobility and shifting migration patterns amongst newly retired populations to new locations – such as rural or coastal areas – that are seen as conducive to their quality of life and wellbeing. For others there has been a rise of so-called snowbird migration, that is, the seasonal migration of older people from (particularly) the cooler northern states of America and Canada and parts of northern Europe to the warmer climes of places such as Florida, Mexico, and Southern Europe in order to avoid the effects of cold, dark winters. Estimates suggest that some locations (for example, Florida in the United States) can swell by hundreds of thousands of temporary older migrants during these winter periods. This phenomenon can have a major impact on the resident populations of both sending and receiving communities, but has also given rise to new communities of mobile older people who seasonally migrate between these places (see Population geography).

With increasing age, however, comes a greater likelihood of experiencing a decline in both physical and cognitive functioning, increasing the demand for health and welfare services. Policy responses vary across countries and even within countries. For many high-income countries, the early decades of the twenty-first century have been characterized by austerity measures and neoliberalizing trends manifest in sweeping welfare retrenchment and a shift towards responsibilization. Risks taken throughout the life course are seen as individual, rational choices that will impact healthy life expectancy in old age. One response has been a growing call to focus on life course approaches within geographical gerontology and an engagement with sociospatial inequalities. Hopkins and Pain (2007), for example, maintained that human geography was placing a disproportionate emphasis on specific age cohorts and the extreme ends of the chronological age spectrum, arguing for “relational geographies of age” in which the focus should be on intergenerationality and life course rather than specific age cohorts.

A more established interest that has been prevalent since at least the 1980s has focused on the impact of different health-care systems for the planning, regulation, utilization, and spatial allocation of resources for older people, particularly in relation to public, private, and community-based resources. More recently, geographers have taken a critical approach that has shifted away from distribution and allocation of services for older people at national and subnational scales to consider how health and welfare restructuring is impacting older people themselves (see Identity). Theoretically underpinned by the shadow state thesis (defined by Wolch in 1990 as a parastate apparatus largely made up of voluntary and community organizations and social enterprises that are administered outside traditional democratic politics, but which have taken on many of the collective service responsibilities previously undertaken by the public sector but which remain within the purview of state control), this work has considered how welfare retrenchment is manifest in changes in accessibility and availability of services, who provides those services, and the shifting sites of delivery from institutional to community and domestic settings. Of particular note has been a widespread shift to market regulation and increased dependence on provision through the voluntary sector and social enterprise in the local, regional, and national political economies of health and social welfare. More recently this shift has also manifested in attempts to reinvigorate active communities and voluntarism as mechanisms to support aging in place and the inclusion of older people.
AGING

The critical turn in social and health geography has seen the emergence of a corpus of work that moves away from treating older people as a population statistic or as the objects of study and has instead focused on the personal and intimate and a more qualitative, in-depth engagement with older people themselves (see Critical geography). The early work of Glenda Laws and others, for example, highlighted the importance of focusing on the ways in which the body functions as a particular and fundamental determinant of the experience of places. Understanding the relationalities that exist between the aging body and places has been an important theme running through this work. A particular concern has been to examine the place of the body in everyday meanings and constructions of aging, bringing into focus what the aging body can and cannot do within places and how places can facilitate or constrain the aging body. Theoretically, this research draws on some of the early work around “environmental press” but there has also been an engagement with Bourdieu’s concept of habitus and the notion of “body capital” (see Antoinetti and Garrett 2012). This highlights the importance of focusing on the interaction between the aging body and places and how any reduction in body capital (physiological decline) not only changes habitus but also how younger age groups see older people. Diminished body capital is also viewed as diminishing other capacities accrued by the individual over the life course, creating a mismatch between the capacity of the body and the environment within which it is located (see Bodies and embodiment).

The aging body is conceived as being not only pivotal to the social construction of later life but also to the peripheralization of older people in discrete locations that may be segregated from those used and inhabited by younger people. Images of the aging body, for example, are often depicted as frail and dependent (particularly in Western societies) and as a consequence have become identified with the home, supported care, or residential settings. Within such settings there is also a strand of work around care and bodywork that has specifically looked at how gendered as well as aging bodies affect the use and meaning of space (see Gender).

From the latter half of the twentieth century onwards, there has been a gradual shift in many high-income countries away from the provision of care for the most frail of our older population within institutional settings towards aging in place. This turn has shifted the emphasis of care and support to domestic settings, where informal carers (usually family members who are largely, but not exclusively, women) have taken on many of the routinized tasks of caring for the aging body, including much of the intimate and personal bodywork involved in care, such as washing and bathing, dressing, toileting, and feeding the older care-recipient. Undertaking these normally personal and private acts gives rise to transgressions of contemporary social taboos around care in Western society – particularly cross-sex care. While the transgression of such social taboos may be less acute in spousal care-giving, it can be particularly difficult where an adult child is providing personal care for a frail older parent of the opposite sex. As a result, relationships associated with the home can be altered and challenged by the process of caring (see Caregiving).

Given the social taboos that often mark the social boundaries of bodywork in Western society, the more detached stance of the professional carer can be important in helping to make it more manageable. So the management of the care-recipient’s body and who undertakes that management can be critical to the construction of the home as a caring space. It is a body that is subject not only to management by informal
AGING

carers, but that has also been assessed by formal care services in relation to the quantity and nature of care it should receive against some institutionally defined norm. It is important, however, that the assessment of care is not interpreted solely in terms of meeting the medically defined needs of the corporeal body; if “social death” is to be avoided, the home also needs to be understood as a place where valued aspects of the social body can also be nurtured and preserved. In other words, it is important to recognize that it is not just the physical body that is attended to but that the social and emotional needs of the aging body are also recognized and met. Indeed, a growing body of research has highlighted the adverse effects of social isolation amongst older people on both their physical health and their mental wellbeing. The design of specialist housing for older people, however, often assumes an aging body that is relatively static, so requiring only limited space. This in turn can impact on an older person’s ability and opportunities to socialize within these settings. It is only through recognizing the home and body as interrelated sites and scales of analysis, that are both fluid and constantly in process, that we can gain real insight into the complex structuring of the relations that shape experiences of care.

Social and cultural stereotypes linked to public perceptions and attitudes to the aging body will shape societal responses to aging, and where these are negative, will contribute to ageism and the ongoing search for a youthful and ageless existence that acts to reinforce the “invisibility” of older people. It is important, however, to recognize that the construction of aging identities in place is both socially and culturally ascribed, hence these highly Westernized conceptualizations of the aging body will vary and be reconstituted in different ways over both time and space.

SEE ALSO: Care work; Caregiving; Demographic and epidemiological transition; Health geography; Home; Population geography

References


Further reading


Agricultural environments

Mark Blumler
SUNY-Binghamton, USA

Other than humans, only certain ants, termites, and beetles practice agriculture; they farm underground, and have relatively little impact on ecosystems. Humans, already impacting environment especially through their use of fire, began to farm about 10,000 years ago. Since then, agriculture arguably has been the major cause of ecological and evolutionary change on the Earth. Agricultural environments are discussed here from an ecological and an evolutionary perspective.

Farming began in several regions – the Fertile Crescent, China, Mexico, the central Andes, and New Guinea – at about the same time. One or two additional regions, notably the Sahel, developed largely indigenous agriculture somewhat later. Farming enabled humans to obtain more food per unit area of land than hunting-gathering, and also allowed craft specialization and technological advances. Consequently, farmers were able to spread geographically at the expense of hunter-gatherers, or to incorporate them into their system. The focus initially, and to a large extent still today, was on fast-growing, mesic, high-yielding species. Annual or short-lived perennial seed crops characterized most of the early agriculture centers, though root crops were important in New Guinea and the Andes. The seed crops, especially grains and legumes, tended to be derived from large-seeded wild ancestors, possibly because seedlings from large seeds are easier to cultivate with a primitive technology than plants from small seeds. Over time, there was an expansion in the types of plants cultivated, and in their uses, to include textiles, ornament, and today, even biofuel.

The agricultural origin centers tend to be located in the foothills of mountainous regions, that is, they have high habitat diversity and hence also high biodiversity. In addition, there is some evidence that they have high local (alpha) diversity. The Fertile Crescent and adjacent Mediterranean, for which biodiversity patterns are best documented, has the highest herbaceous alpha-diversity on Earth. Farming was inherently an attempt to simplify these ecosystems, to favor a few staples at the expense of other species. Nonetheless, the origin regions remain highly diverse today. Many species from those regions, including some crops, are now serious overseas invaders, while the origin centers themselves have suffered comparatively little from invading alien species.

Species that underwent domestication often hit the evolutionary jackpot: for instance, wild emmer, the progenitor of most varieties of wheat, occurred naturally only in the Fertile Crescent. Today, its domesticated derivatives are grown almost everywhere that agriculture is practiced in the temperate and subtropical zones, and are even penetrating the tropics. When a species expands its range so spectacularly, inevitably, other species must undergo range contractions or even become extinct. The line in the song “America the Beautiful” about “amber waves of grain” illustrates the extent to which farming landscapes have taken over and replaced indigenous ecosystems. Global vegetation maps usually show “potential natural vegetation” rather than the actual cover; if they were to show the latter, agricultural plants...
would dominate over a major portion of the terrestrial regions (see Ellis and Ramankutty 2008, for global anthropogenic biomes).

While the domesticates benefited from their mutualistic relationship with humans, other species were able to take advantage of the newly created agro-ecosystems, despite our best efforts to control or eliminate them. These unwanted species, “weeds” and other “pests,” also benefited from agriculture at the expense of other wild species, presumably. Some weeds became so successful as to overwhelm the crop, typically as farming spread into a different environment less favorable for the domesticate, at which point if the weed was edible, farmers began to harvest and cultivate it (so-called secondary domestication). Meanwhile, the wild progenitor of a domesticate frequently would evolve weed genotypes, and complex crop-weed-wild interactions might develop (Harlan 1975).

Certain families characteristically benefited from the adoption and spread of agriculture. Proponents of the “Paleolithic Diet” argue that grains and legumes were scarcely consumed before agriculture; regardless, they certainly are very important today. The Cucurbitaceae (melons, squashes, and cucumber), mustard family, and Solanaceae (potatoes, tomatoes, peppers, and eggplants), provided vegetable crops, and the Rosaceae supplied a remarkable number of temperate fruits. Some families were pre-adapted to the combination of fertile conditions and soil disturbance that characterizes agriculture, and consequently became weeds, crops, or both: the mustard family, spinach family, buckwheat family, and so on. Others were attractive for their nutritive qualities, such as the rose family.

As farming spread, humans had to transform environments so as to resemble the origin center from which they were migrating. This process is particularly evident for spread from the Fertile Crescent. There, extreme summer drought favors dominance by annual plants, and all the early domesticates were annuals. As agriculture spread northwest into Europe into a region without seasonal drought, it became necessary to plow to remove the native, perennial vegetation. Since even the weeds were largely of Fertile Crescent origin, in effect an entire ecosystem was transplanted. Fertile Crescent agriculture also spread east to India, where wheat, chickpeas, lentils, peas, and other winter annuals are now major items of the diet. But India has a monsoon climate, with rain in summer. To grow the Near Eastern crops, Indian farmers developed the “rabe” system, from the ancient Indo-European word for mustard, which also is from the Fertile Crescent. They dry fallow in summer, plowing repeatedly to prevent weeds from utilizing the monsoon moisture; the crops are sown in fall, and mature primarily on the stored moisture.

In the Old World, farming diffused until it occupied essentially all arable land. Land that was too infertile, cold, or arid was reserved for pastoralism. Technologies such as irrigation and terraces were massively developed in order to maximize the arable land. In contrast, in the Americas population density is still relatively low; arable land is currently expanding in much of Latin America. In particular, whitewater portions of the Amazon are seeing massive penetration by peasant farmers who clear the rain forest with fire. While swidden was long practiced traditionally, it is unclear how intensively the Amazon was farmed before 1492: there are indications of dense, permanent populations along the Amazon River, before Old World diseases struck, and the widespread occurrence of terra preta, a fertile, dark-colored, anthropogenic soil, also suggests that permanent agriculture was a feature of the more fertile parts of the rain forest. In contrast, the blackwater tributaries of the Amazon are simply too infertile to have supported other than shifting, swidden agriculture.
Until the Industrial Revolution, agriculture was entirely based on renewable resources, and thus in a sense sustainable; though the degree to which accelerated soil erosion may have developed into land degradation/desertification is much debated. The extent to which early Mesopotamian civilizations may have salinized soils in the Euphrates-Tigris basin is in question; certainly much of the region is still highly productive today. Erosion of slopes at times choked valley bottoms and created wetlands that became malarial to the detriment of the local populations, but such outcomes would not have lasted forever given the natural tendency of streams to unclog their drainages. In desert regions, farmers such as the Nabataeans, Hopi, and Mixtec intentionally caused slope erosion to concentrate soil in the wadis where moisture was sufficient for crop growth. Properly maintained terraces typically produced less erosion than would have occurred naturally; though during times of war or other catastrophe, when terraces were neglected, erosion could become highly accelerated. Bali represents one of the most extreme cases of land transformation though terracing and complex water management. Almost the entire surface is given over to agricultural plants, with terraced rice paddies, and domesticated palms in un-terraced spots. Rice geneticists attempted to spread the Green Revolution to Bali, but later withdrew, admitting that the indigenous system yields at least as well as a modern system would be able to do.

The infamous Dust Bowl of the 1930s had a major influence on beliefs about soil erosion, and belief in widespread land degradation persists within the environmental movement and the United Nations. But within geography, the consensus today is that: (i) the Dust Bowl region was not permanently ruined; soils remain fertile on the whole, though today if farmed they are irrigated. (ii) There is no good evidence for desertification anywhere except on a very local scale (Thomas 1993), aside from the Aral Sea debacle resulting from a misguided Soviet irrigation scheme. (iii) Increasingly, it is recognized that when soil is lost in one place it must be gained somewhere else (though admittedly that is not always a good thing).

Farming can have major effects on off-farm ecosystems. Eutrophication of waterways due to fertilizer runoff is one of the better known and more serious examples. Construction of reservoirs for irrigation purposes can alter downstream river flow, to the detriment of many organisms. The trapping of sediment behind dams is causing a reduction in the size of deltas such as those of the Mississippi and the Nile.

Today, farming is utterly dependent upon fossil fuels. For instance, the Green Revolution, which after World War II spread modern agriculture to poor countries, entails the use of gasoline-powered machines such as tractors, artificial N fertilizer manufactured using natural gas, irrigation pumped with motors powered by fossil fuel, and herbicides and pesticides largely manufactured from hydrocarbons. As such, sustainability is inherently doubtful.

These transformations reflect, and have caused, a change from subsistence to market oriented farming, with the resulting disadvantaging of relatively infertile or otherwise marginal arable lands. For instance, much of the northeastern United States, which was given over to subsistence agriculture in the nineteenth and into the twentieth century, is now abandoned and has reverted to forest because both soils and climate are not conducive to high yields, while improving transportation technology has enabled farmers in more fertile climes to ship their produce over increasing distances and undersell local farmers. On the other hand, genetic engineering
combined with massive applications of phosphorus and other fertilizer is transforming some formerly infertile regions into arable. For instance, the Cerrado of Mato Grosso, an extremely diverse, infertile shrub savanna formerly utilized for extensive cattle grazing, is increasingly given over to soy cultivation. Legumes such as soy have a high phosphorus requirement, but once that is supplied can add the nitrogen that such soils lack.

Now, the increasing use of ever-larger machinery combined with herbicide-resistant GMO crops are causing fields to become ever larger also, as well as leading to the elimination of hedgerows and other vegetated field boundary areas. The biodiversity of these fields is extremely low. Among many other impacts is the recent decline in the abundance of monarch butterflies, which depend on a food plant, milkweed, that formerly inhabited field edges and other open areas but has been greatly reduced by spraying and plowing. Many birds and other animals that depend on weeds for food, and/or use field edges for migration corridors, are now declining rapidly toward possible extinction (Stoate 2011).

Theoretical debates over the causes of biodiversity remain unresolved. For instance, many ecologists believe that high productivity is associated with high diversity, some believe that the relationship is humped (unimodal), and still others point to the low fertility of soils in western Australia and the South African Cape, which have remarkably high plant species diversity. The best evidence is that the enormous quantities of fertilizer now being applied are causing a major loss of plant biodiversity due to competition for light. The role of disturbance also is somewhat controversial, with perhaps the consensus being that up to a point it is beneficial, but that diversity declines if disturbance becomes massive. If so, farming might originally have fostered diversity, and indeed, the evidence from traditional farming suggests that may have happened. However, the enormous scale of the current industrialized system is clearly inimical, as it reduces landscape heterogeneity, a known associate of diversity. Impacts also are likely to depend on the type of natural ecosystem within which farming is practiced. Farming fragments forests, for instance, which has the effect of favoring “edge” or ecotone species, while disfavoring species that live in the interior.

Today, there is increasing concern about diversity of the crops themselves. Globally, the human diet is becoming ever more restricted to a very few staples, though in the supermarkets of developed countries the food diversity is spectacular. In part, this reflects the conversion of agriculture in much of the underdeveloped world to cash cropping, spurred on by Western institutions such as the World Bank. The spread of Green Revolution genotypes has raised concern for the genetic diversity of crops, since much of that diversity resides in traditional varieties grown in often-marginal environments that are now subject to abandonment due to inability to compete on the international market. Attempts are being made to preserve diversity in gene banks, but these suffer from introgression, and adaptation to the environment in which the gene bank is located.

SEE ALSO: Agricultural geography; Agriculture; Biogeography; Desertification; Ethnobotany; Mediterranean-type ecosystems; Soil erosion and conservation; Soil fertility and management

References


Further reading


Agricultural geography

Guy M. Robinson
University of Adelaide, Australia

Traditional approaches

The word “agriculture” is derived from the Latin word *ager* and the Greek word *agros*, both meaning “field,” and from the Latin word *cultura*, meaning “growing” or “cultivation.” The term symbolizes the integral linkage between a human land-based activity and its accompanying modification of the natural ecosystem. Since the first domestication of wild plants began in the Fertile Crescent of the Near East around 10,000 years before present, people have created agroecosystems in which natural ecosystems have had socioeconomic processes and systems laid over them. In studying agroecosystems agricultural geographers have addressed agriculture as part of a nested hierarchy extending from an individual plant or animal, with its cultivator, tender, manager, or farmer, through to systems of crop and animal populations, fields and grazing lands to farms, villages, watersheds, regions, and beyond. Hence, agricultural geography has spanned a huge range of issues and interests, from descriptive mappings of the spatial distribution of crops and livestock to analysis of systems of management, the nature of linkages to the broader economic, cultural, social, political, and ecological systems, and activities that generally occur beyond the farm gate, such as processing, marketing, and consumption of food. In particular, for the last half century research in agricultural geography has focused on the economic, social, cultural, and political characteristics of agriculture and its linkages to both the suppliers of inputs to the agroecosystem and the processing, sale, and consumption of food products.

This broad focus reflects a shift from “traditional” agricultural geography, which was dominant until the 1960s. This was concerned with location and context, emphasizing the regional characteristics of agricultural activities, and examining relationships between key variables affecting farm-based production (e.g., physical characteristics, the local economy, trading systems, cultural practices). The regional focus in agricultural geography enabled this branch of the discipline to play a major role in the growth of regional geography in the 1920s as geography’s central paradigm. Large-scale regional delimitations were based on distribution maps of crops and livestock and the classification of agricultural systems. Mapping of land use assumed national significance in some countries. For example, the Land Utilisation Survey of Great Britain begun in the 1930s by the geographer L.D. Stamp provided rich material for detailed agricultural geographies on a county-by-county basis. Post-1945 the use of statistical methods enabled multiattribute agricultural and type-of-farm regions to be formulated.

The mapping of agricultural distributions was at the heart of much traditional agricultural geography, with pioneering work by O.E. Baker in the United States and subsequent innovations involving land classification, land-use classification, and type-of-farming regions. J.C. Weaver pioneered the creation of crop combination regions in the US Midwest, recognizing that
regional production complexes usually include a range of crops rather than a monoculture. Crop areas and livestock numbers could be converted into standard man days of production, enabling livestock production to be incorporated into classifications. At a global scale, different types of agriculture were recognized and mapped, such as the world agricultural classification devised by Derwent Whittlesey in the 1930s and widely utilized. In the 1970s J.T. Coppock utilized computer methods to map agricultural distributions and type-of-farming regions in the United Kingdom.

The quantitative revolution

The theoretical and quantitative revolution that swept through geography in the late 1960s was embraced enthusiastically by several agricultural geographers who applied structural models and economic theory in work on the economics of agricultural production, often based on sample surveys of farms. One of the models adopted, the Von Thünen model of land use based on economic rent for crops, dated back to the 1820s but was not popularized within geography until the 1960s when various applications were proposed. However, the economic basis for much agricultural geography at this time was a logical outcome of the formulation of general laws of agricultural location based on economic principles. A recurrent problem was the lack of suitable economic and social data as official data on farm costs and profitability are rarely suitable for geographers’ needs. In contrast, some historical farm and estate records are of sufficient detail for microscale studies which would not be possible using modern official data.

From the late 1960s the application of computing methods extended geographers’ ability to map agricultural distributions, allying computer mapping to research on regional changes in farm inputs, farm-size structures, farm incomes, and agricultural marketing. However, the model-based, positivistic, theory-driven work dominant in the 1960s and 1970s attracted criticism as new approaches made an impact. The chief complaint was that it was an agricultural geography devoid of real people, composed of robots obeying arcane economic laws that involved minimizing effort and cost, far removed from the realities faced by individual farmers. Noneconomic considerations, such as farmers’ motivations and decisions not based solely on profit maximization, were missing from mainstream work, but appeared increasingly as geographers investigated farmers’ decision-making, the diffusion of agricultural innovations, and the responses of individual farmers to economic stimuli. This work was part of a growing body of behavioral research that was still highly empirical and positivist but beginning to develop systematic analyses of the spatial outcomes of individual decisions and to focus on the role of cognitive and decision-making variables.

The behavioral approach emphasized the characteristics and qualities of individual farmers, with research defining, measuring, modeling, and analyzing statistically the attitudes and revealed patterns of behavior of farmers based on studying a selected sample of farmers in a given area. Emphasis continued to be placed on economic forces and quantitative measurements until the mid-1980s when policies supporting environmental actions by farmers led to different types of studies investigating farmers’ adoption of environmentally friendly measures.

Torsten Hagerstrand’s work on the spatial patterns associated with innovation diffusion within Swedish farming in the 1950s and 1960s gave rise to various studies on farmers’ adoption of innovations, including new policy measures. These focused initially on orderly, predictable,
and linear progressions in the adoption process, concentrating on the demand or adopter side of change as opposed to the supply or provider side. Subsequently, research recognized the importance of lead-user inventions, change agents, and commercial marketing organizations. Factors like the influence of economic inducements, rural services, and infrastructure may be inadequately accounted for by a focus on individual decision-making behavior. Hence other models investigating different elements in the decision process have been proposed, including information models, expert and decision support systems, and learning and knowledge transfer systems.

Since the mid-1990s the most popular behavioral approach within a system or network framework has been actor-network theory (ANT), first introduced in the 1990s. The actor network is an interaction between human “actors” and nonhuman “intermediaries.” Material objects may be active in configuring human actors and their relationships, and hence human and nonhuman actors may have equal status in a network. Alternatively, nonhuman intermediaries can allow objects in the natural environment to be connected to social actors. So it maps relationships that can be simultaneously material (between things) and semiotic (between concepts), though there is a diverse variety of approaches by geographers under the umbrella of ANT. In particular, ANT has been used to understand the globalization of agriculture whereby the agri-food industry has become global in terms of both supply and demand (Robinson and Carson 2014). Applications of ANT have shown how farmers’ participation in the new industrial-style agricultural networks molds their decisions, defining the types of knowledge and skills they and others in the network value, and helping to transform farming practices in favor of unsustainable production practices that rely on artificial inputs.

### Political economy

A major development within human geography in the 1980s profoundly affected agricultural geography. This was the focus on a more systematic analysis of the political and structural frameworks within which agriculture nests. This was labeled the political economy approach, to signal its incorporation of Marxist-related content. The nature of this approach initially owed much to Piers Blaikie’s work on the political economy of soil erosion in developing countries, emphasizing the social relations of production and their ramifications for land use (Blaikie 1985). This conceptualized the farm household as having two kinds of social relations: the local and the global, with the latter increasingly assuming greater prominence.

Political economy is often closely linked to the nature of the management of the economy by the state, and the focus of this work has often been on the role of the state at various levels to set the parameters within which both economic and social change occurs. However, from the outset the term “political economy” was used fairly loosely by human geographers, often eschewing Marxist ideas in favor of broad concerns for class and power relations, and the investment decisions of agrarian and industrial capital. In agricultural geography this brought recognition of the way in which farms had been increasingly absorbed into the larger agri-food system, which includes off-farm inputs such as suppliers of seeds, fertilizers, and machinery (termed backward linkages) and recipients of farm produce such as food wholesalers, retailers, processors, distributors, and consumers (termed forward or downstream linkages).

This consideration of the wider agri-food system moved the focus of agricultural geography away from the individual farmer onto a broader research focus, acknowledging the farm as part
of the wider agri-food chain, itself embedded in a multifaceted web of economic, social, cultural, and political dimensions all impacting on farm-based production. So there has been an emphasis on the need to understand agriculture as part of the global agri-food system.

The initial framework for this political economy approach focused on four elements: uneven development (with capital penetrating different regions at different rates), geographical and historical specificity (requiring locality studies to understand the process of uneven development), examining the family farm (recognizing that the basis of agricultural production is different from other sectors of the economy) and the role of state policy (where governments have supported the family farm but often inadvertently have encouraged the penetration of agriculture by corporate capital). The new approach prompted research on the growth of agribusiness, agricultural restructuring, and the diversity of social relations and cultural practices shaping accumulation and regulation. In the United Kingdom in particular, this research focus has tended to mean that agricultural geography has been subsumed within the broader scope of “rural geography” in recent decades. However, a more distinctive and vibrant agricultural geography has continued to flourish in the United States, France, and Germany.

Critiques and alternatives to political economy soon appeared, notably criticisms of its “excessive concentration” on the agricultural production sector, dominated by the farm enterprise, farm household, landownership, farm labor, and agricultural technology. Political economies stressed structural processes as opposed to active decision-making by farm families, overlooked noneconomic decisions by farmers regarding allocation of resources, and focused too much on macroeconomic factors. However, the attention given to the broad agri-food sector prompted further consideration of three broad theories for aiding the understanding of contemporary agricultural change: food regimes, regulation theory, and a restatement of the “agrarian question.”

Food regimes

Food regimes is a concept linking international relations of food production and consumption from the 1870s onward to forms of accumulation and regulation under capitalist systems (McMichael 2013). Three “regimes” are recognized, the first extending from the 1870s to World War I (based on worldwide grain and meat production involving exports from family farms in the New World and linked to the rise of nation-states); the second from the 1920s to the 1980s (involving production of durable food for the mass market, decolonization, consumerism, and the growth of forward and backward linkages from agriculture); and the putative third regime dates from the 1990s, emphasizing fresh, organic, and reconstituted foods linked to global restructuring, with the globalization of production and consumption, international regulation, and the growth of “green” consumers.

These regimes overlap, and the third regime is still evolving as a conflation of globalizing and sustainability trends involving a mixture of biotechnology, fast food, organic production, and consumption of locally produced food (Guthman 2004). This regime includes extensions of technology from the preceding regime, including biotechnology and the use of chemicals to simulate naturalness and to control the ripening of packaged fruit.

Criticisms of the concept refer to its lack of explanatory content, omission of human agency, overreliance on the experiences of the developed world, neglect of national regulatory processes, and its high level of abstraction. Nevertheless, the concept has been useful in highlighting the
emerging divergent trends within contemporary world agriculture as globalization and the reliance on industrial-style methods meets some resistance from consumer-led movements favoring organic production and foods with fresh, local, and sustainable connotations (Pretty 2002).

Regulation: productivism and postproductivism

In recognizing worldwide transformative change, such as moves from one food regime to another, geographers have examined the ways in which different forms of regulation help shape particular economic structures and social forms. Regulation theories have referred not only to government-imposed rules but also concurrent cultural representations, political strategies, and social conflicts. In agricultural geography perhaps the main contrasting periods of regulation discussed have been those of productivism versus postproductivism, an approximate parallel with Fordism and post-Fordism in that it implies a move to different work/production practices, different relations of production, and changes in the regulations governing production.

Productivist agriculture has characterized much of the post-1945 period in which commercial farming was linked to technological development encouraged by the state through its support of research and advisory services. State support aimed at raising farm production is exemplified in the Common Agricultural Policy (CAP) of the European Union, a major subject for analysis by geographers. Productivism has been associated with substantial restructuring of farm-based production involving the overproduction of agricultural goods in face of largely stable demand, rising farm indebtedness, downstream environmental consequences, major regional differentiation reflecting dualism between small, marginal family farms and large, heavily capitalized businesses, and new forms of family farming, including part-time and hobby farming (Robinson 2004).

This strong contribution to supporting productivist farming by some governments has provided rich material for geographical study. State intervention has focused on how government has attempted to create a favorable economic environment for agriculture; it has often offered specific financial inducements for farmers; and it has introduced specific regulations such as production quotas that directly dictate farmer behavior. Particular foci have been on farm policies in the United States and the various reforms to the European Union’s CAP. The latter has been of special interest because of its evident spatial consequences, distinctive price support measures, and a sequence of reforms starting with the so-called MacSharry reforms in 1992. Some comparisons between US and EU policy have been made, notably relating to different experiences of setting aside land from production. The expansion of the European Union has also offered new opportunities for research, especially on the land reforms in central and eastern Europe. A theme in the work on government agricultural policy has been the changes to traditional farmer–state relations, linked to macro-policy changes via world trade negotiations, such as the Uruguay Round of the General Agreement on Tariffs and Trade and the World Trade Organization.

The term “postproductivism” was first popularized in the 1990s, referring to some of the responses to the “excesses” of productivism. Notably this can mean a reversal of the trend on farms toward intensification, concentration, and specialization, with moves to extensification, deconcentration, and diversification (e.g., pluriactivity). An alternative view is that postproductivism is linked in part to the loss of the central position of agriculture in society
AGRICULTURAL GEOGRAPHY

and changing public attitudes to farmers, in which intensive agriculture is often regarded as a threat to the environment and the landscape. However, one of the responses to the negative environmental consequences of industrial farming methods has been the growth of policy measures promoting environment-friendly actions, known as agroenvironmental policies. These have been accompanied by reduced production subsidies, tighter pollution regulations, and changes in property rights, which have been part of some dismantling of protectionist state policies, creating new opportunities for farmers. Examples of the new opportunities include a rising demand for “quality” foods, including organically and locally produced food, as part of new forms of agricultural production (Pretty 1995), and the commodification of former agricultural resources: land, wildlife habitats, barns, and cottages. Key characteristics of the transition to postproductivism have included a shift from quantity to quality in food production; the growth of alternative farm enterprises as part of pluriactivity on farms; state efforts to encourage a return to more traditional, sustainable farming systems through agroenvironmental policy; the growing environmental regulation of agriculture; and the progressive withdrawal of government support for agricultural production.

Geographers have contributed new theory to help understand postproductivism, identifying new regulatory conditions and links to other “transitions,” and have tested the extent to which postulated changes have occurred. Yet, despite evidence of postproductivism, most farmers have remained firmly productivist and many general agricultural policy adjustments in the developed world have focused on increasing competitiveness and the need to produce food for mass consumption. Since the mid–2000s a new and more encompassing term has been introduced to the lexicon, “multifunctional agricultural regimes,” which acknowledges that agriculture, as well as having a food-producing role, also plays major additional economic, environmental, and sociocultural roles in rural areas (Wilson 2007). These roles include the generation and management of rural landscapes and ecological features, and the social role of supporting population, especially in peripheral areas. Therefore policy reforms have had to accommodate this multifunctional role of agriculture as in the reformed CAP of the European Union.

Agricultural geographers have tended to understand multifunctionality in agriculture as a policy-led process that describes current agricultural trends rather than as a concept. However, ongoing theoretical debates are now examining where multifunctionality sits within the productivist–postproductivist transition and how it contributes to broader rural multifunctionality. These debates incorporate both economic and policy-based views, as well as emerging holistic and cultural interpretations. The policy contribution highlights, for example, the European Union’s role in urging the maintenance of farm subsidies in the World Trade Organization’s Millennium Round, and its attempt to simultaneously decouple support from production while recoupling it to agroenvironmental and rural development. Other EU policies favoring multifunctionality include: its LEADER program, based on reconfiguring local resources, redefining the social role of agriculture, and increasing the value added to farm products; rural development regulations emphasizing farm diversification, rural tourism, and landscape conservation; and a raft of agroenvironmental policies. However, within agricultural geography issues relating to multifunctionality, the role of globalization, and postproductivism continue to be hotly debated.
AGRICULTURAL GEOGRAPHY

The agrarian question

A key concern of geographers has been to understand the survival of the family farm. The latter has been the bedrock of both agrarian capitalism in the New World and of rising productivity, especially in Western Europe since 1945. Yet there are apparent contradictions between the continuing survival of family farms and the most profitable forms of business arrangement. Family farms may be threatened by the growth of specialization and the “international farm crisis” in which farmers have been squeezed by rising costs and reduced prices for farm produce, in part because of the rising power of supermarkets, wholesalers, and processors. For example, the number of farms in the United States peaked at 6.8 million in 1935. Today the country has only just over 2 million farms, of which 87% are family-operated, but with a steadily rising average age of farm operator: 60% are aged 55 and over. Just 8.5% of farms accounted for 63% of production, highlighting the growing dichotomy between large, industrial-style farm operations and smaller family farms that are often reliant only on family labor (Hart 2004). Indeed the farm labor force worldwide has fallen substantially, especially in the developed world.

Family farmers have resisted conditions favoring subsumption into fully capitalist production by drawing on family labor, earning money from off-farm work, farm-based processing, and direct marketing, and introducing new sources of income, for example, farm-based tourism. Only 45% of American farmers claim farming as their principal occupation. This reflects the growth of farm diversification and pluriactivity by farm families involving pursuit of nonfarming activities, termed “other gainful activities,” in order to generate income. Some have theorized that agriculture differs from other economic activities because of its relations with “fractious” nature and the maintenance of nonwage labor on farms. Yet, there are examples of the large-scale penetration of agriculture by corporate enterprises, for example, industrial-scale fruit and vegetable operations in California. Elsewhere farmers have gained institutional support through contracts with major food processors such as Tyson Foods and H.J. Heinz, although fewer than 5% of American farms are “corporate.”

The reduction in the farm workforce reflects changes in the nature of farm work, with mechanization reducing the need for manual labor. Furthermore, gendered arrangements have also changed within farm families so that, across the world, there is far greater involvement of women in a range of farm work than 50 years ago. Distinct theoretical approaches by geographers to this development can be recognized, with socialist feminism stressing tensions with capitalism through greater reliance on female farm labor, and poststructural feminism emphasizing the need to study the diversity of farm women’s self-image and the relevant social and policy positions. These studies have especially highlighted the contribution of women to farm labor in developing countries, for example, women in sub-Saharan Africa produce and market over three-quarters of food grown outside the corporate sector.

Agri-culture and sustainable agriculture

The engagement with the wider agri-food network has led agricultural geography to extend into studies of food retailing and food consumption, and to recognize the growth of an alternative food economy (AFE). In part this reflects the “cultural turn” within human geography, where a revival of interest in cultural aspects of human life has supplemented the previous emphasis on the economic. In
agricultural geography it has been marked by a shift of focus from production to consumption, and a recognition of the dominance of global trends in consumption (typified by “fast food” outlets such as McDonald’s, KFC, Pizza Hut, and Burger King), but alongside a contrasting process of “relocalization” and “reconnection” in which mass consumption patterns are mediated by local specificities. The latter involves concerns for place or region of origin of food as part of a desire for authenticity, greater variety, and concerns over standards of mass production and processing practices (Maye, Holloway, and Kneafsey 2007).

Studies of consumption have highlighted the rise of the supermarkets and the homogenization of food consumption across the globe, with particular brands available in almost any supermarket outlet. The growth of industrial-style agricultural production to serve this agri-industrial system has dominated agriculture in much of the developed world, though definite countertrends are evident. For example, farmers’ responses to supermarkets’ growing domination of the market for groceries across the developed world have included greater recourse to direct marketing as part of an emerging AFE. AFEs include farmers’ markets, box schemes, pick-your-own, farm shops, mail order, and home delivery. These are closely linked to the development of local connections and identities, concerns for food quality, links to traditional methods of production, and new regulations such as the European Union’s protected designations of origin (PDOs) and patented geographical indicators (PGIs). PDOs and PDIs have linked particular foods, especially cheese and drink products, to specific places and regions, for example, champagne, Stilton cheese, and Parma ham. These are specialty food products (SFPs) associated with authenticity of geographical origin and traceability, representing a “relocalization” of the agri-food system in which the products command a market benefit if they are effectively marketed (Goodman, Du Puis, and Goodman 2012).

The growth of studies focusing on links between farming and environmental issues has increasingly been framed within the context of resource management problems and notions of sustainable development. Essentially debates have evolved from the recognition of negative environmental impacts arising from productivist agricultural policies, including sharp declines worldwide in biodiversity including seminatural habitats. There has been focus on the outbreaks of diseases affecting plants and animals and the rise of particular pests, and a wide engagement with the discourse of sustainable development, especially prospects for moves away from productivist modes of production. This has included policy–related research, including examinations of the inherent tensions between “sustainable” and “development,” as encapsulated in work on sustainable livelihoods in developing countries, and the concept of sustainable agriculture (Blay-Palmer 2010; Marsden and Morley 2014). Particular interest has focused on systems of production that reduce negative environmental impacts. These systems take many guises, including organic farming, integrated farming systems, integrated crop and fruit management, and biodynamic production (Pretty 1995).

Since the mid-2000s the concept of ecosystem services has increasingly been adopted within agroenvironmental measures, reflecting the ways people benefit directly from goods and services delivered by the environment, which can be valued and rewarded by various economic valuation techniques. This has enabled the environmental externalities of farming to be considered and supported through on-farm conservation efforts. Indeed ecosystem services have become the principal concept and tool used by ecologists, conservation biologists, and conservation
planners to communicate and to implement environmental management strategies. Payments for ecosystem services (PES), or markets for ecosystems services, have been introduced in many developed countries, generally through programs that combine the provision of measurable conservation benefits while preserving working farm-based landscapes for future generations. Programs typically involve farmers voluntarily providing particular conservation actions, which can be valued so that farmers receive payment for the services provided.

**Dual economies and green revolutions**

Work on changing world trade relations has enabled agricultural geographers to engage with the impacts of globalization on farming in developing countries. Contrasting views have appeared about these impacts, both positive (enabling production in the Global South to increase to meet demand in the North but with emerging firms in China and other developing countries taking a greater market share) and negative (local producers increasingly in thrall to Western multinationals). This work acknowledges the continued presence of a dual economy in developing countries, with a dichotomy between small-scale local producers, and initially many subsistence farmers producing for the local market, and a foreign-dominated sector producing for export (Bacon et al. 2009). These two sectors have become blurred in many countries, for example Ghana, while subsistence farming has retreated in the face of an advancing wave of commercialization. Debates have included the link between agricultural intensification and population pressure, environmental consequences, and social/cultural impacts. In the 1960s Esther Boserup linked population pressure to moves from shifting cultivation (equated with low productivity) toward permanent or semipermanent cultivation with high productivity (Boserup 1965). However, her model cannot be applied in many parts of the tropics, especially Africa, where fragile soils can restrict intensification unless supplementary inputs such as chemical fertilizers are introduced or particular local management systems applied. These various systems have been investigated by geographers in terms of nested sets of interrelationships operating at the global, national, and international levels so as to provide both opportunities and constraints for individual farmers and communities (De Koninck and Rousseau 2013).

The legacy of colonialism has also been viewed through an agricultural lens, notably the application of European farming methods to tropical environments where colonists frequently dismissed existing production systems as inferior to European ones. Europeans introduced changes in land tenure, the consolidation of small fragmented plots, and mixed farming across Africa via resettlement schemes or gradual integration into existing farming practice. This was often part of a process of commercialization and intensification which in Southeast Asia followed a process of transition from 1960 whereby those living on the land eventually declined in numbers as in the West, but with a dualistic structure maintained: small uneconomic family farms alongside large fully commercial units.

This duality has a long history in Latin America where colonial powers often created large estates (latifundia), geared to producing staples such as sugar for export, alongside very small holdings (minifundia), occupied by small family farmers and landless laborers. This pattern was also widespread across the Mediterranean. In both cases land reforms, dating to the nineteenth century in some cases, have sought to establish a more equal distribution of land. This has been a fertile topic for geographical research, extending
into new geographical areas when post-1989 reforms to tenure occurred in Eastern Europe. Ongoing land reforms have also occurred in parts of Africa (Juma 2011). Increasing interest has been shown in Chinese reforms involving large-scale land-engineering schemes aimed at raising agricultural productivity while contributing to major rural–urban population shifts and rapid urbanization.

Technology transfers from the rich Global North to the poor South are symbolized in the package usually termed the “Green Revolution,” popularized in the 1960s and providing wide scope for geographical study in work on its diffusion, impacts, economic outcomes, and contribution to alleviation of world hunger (Basu and Scholten 2013). The revolution consisted of high-yielding crop varieties (HYVs), fertilizers, pesticides, machinery, and irrigation, largely promoted by funding from the West. The staple cereals of maize, wheat, and rice have been targeted through the establishment of special American-funded plant-breeding centers in Mexico (for maize and wheat) and the International Rice Research Institute (IRRI) in the Philippines, from which the improved varieties have been distributed.

So-called “miracle rice” (strain IR8), bred in the Philippines in the 1960s, was part of the package spread throughout South and Southeast Asia. It is capable, under favorable conditions, of giving increasing yields per unit area. In effect, the Green Revolution was a series of adoptions of particular innovations, with various phases.

1 Initiation and early dissemination in areas of suitable climate: Yields rose, output and farm-based labor rose, and grain prices fall, benefiting consumers.

2 Emergence of significant problems, evident by the late 1960s, with crop pests and diseases causing widespread damage to IR8 crops. It proved difficult and costly to spread the crops into new areas. Some crops were not well adapted to local cultivation practices or cuisine, and required high levels of investment to grow.

3 Second-generation HYVs to genetic modification (GM): From the early 1980s new methods to control pests led to the gene revolution and, later, a variant widely adopted in North America – GM and biotechnology.

Some contend that GM crops, livestock, and foods offer the prospect of increased production on a large scale to dramatically increase the world’s supply of certain crops and livestock products. In 2012 there were 170 million ha under GM crops worldwide, with a growth rate of 10 million ha per annum. The main GM crops are soybean, maize, cotton, rapeseed, and potatoes. The United States currently has three-quarters of the global area devoted to GM crops, mostly herbicide-resistant, offering easier control of weeds, less tillage, and reductions in soil erosion.

The cultivation of GM crops has been extremely contentious in some countries, with ongoing debates between the scientific community and the general public as to their safety (Ferry and Gatehouse 2009). This has produced severe restrictions on their adoption in some countries, notably the United Kingdom and Australia. In addition, the argument that they represent a solution to problems of food shortage in developing countries is largely not substantiated. Most developments of GM crops have been for crops in temperate farming systems rather than for subsistence crops in the tropics. Those available in developing countries have also been associated with smallholder farmers being forced to buy new GM seed every year rather than following traditional practice and holding back seed from harvest to sow the next year’s crop. This
added cost can be a major deterrent. Two-thirds of all GM crops have been engineered for herbicide tolerance rather than for any intrinsic improvement in food quality or pest resistance.

**Land-use competition**

The use of satellite remote sensing technology has enabled a flourishing of research on changes in the world’s agricultural land. Particular foci have been monitoring increases of both arable and pasture in the developing world, primarily at the expense of forest, and land-use competition in the rural–urban fringes of major cities. The loss of tropical forest within the Amazon basin has attracted special attention. Despite the substantial impact of large-scale clearance of forest for cattle ranching, one-third of Amazonian deforestation is caused by small farmers, each clearing less than 100 ha. Yet, in some parts of Asia, the extent of cultivated land is falling, notably in China, with potential implications for grain production. As a result of pro-environmental programs such as Grain for Green, China’s largest land retirement and afforestation program, substantial areas of land under grain have been converted to forest and uncultivated land (converting about 15 million ha of low-yield farmland to forest and afforesting another 17 million ha). However, this may have only reduced overall grain production by 2–3% despite some significant local impacts (Zhao 2013).

To understand the dynamics of land-use competition, simple models were utilized in the 1960s, starting with distance–cost relationships associated with economic rent, as per the von Thünen model. The complexities of land-use competition have largely consigned the model to history, but several more recent studies have focused on the land market, acknowledging the impact of measures intended to protect prime farmland, especially in the United States and the United Kingdom. For the former, key ideologically based lay discourses have been significant. These include regarding farmland as a finite productive resource; maintaining food production in the national interest; farmland preservation as environmental protection; and farming as a connection between nature and society. Lobby groups associated with these ideas have also proved fertile ground for study, especially as concerns over rural amenity remain agrarian ideals well entrenched in American social and political culture.

A different conglomeration of land uses has occurred around urban areas in developing countries, especially in sub-Saharan Africa, where “urban” agriculture contributes to sustaining the urban populace. This production also provides environmental benefits, creates employment, represents a survival strategy for low-income urban residents, and utilizes urban wastes (Cockrall-King 2012).

Competition for agricultural land tends to be at its greatest in the rural–urban fringe, where relationships between the urban economy, the agricultural economy, agricultural policy, and planning policies are at their most critical. Interconnections between the physical environment, economic activities, and the resident population produce complex land-use assemblages of which agriculture is just one component. Indeed, speculative landholding and constraints on building development often produce contrasts between idle farmland and intensive production, for example horticulture and market gardening in close juxtaposition to unfarmed land. In some countries, green belt and farmland preservation policies restrict urban sprawl and maintain farmland within a distinctive rural–urban fringe. However, strict planning controls on development can also produce proliferation of small and fragmented
AGRICULTURAL GEOGRAPHY

holdings, poor quality farmland subject to trespass and vandalism, or low-intensity activity such as agistment, equine-related activities, and recreational hobby farming. The latter has increasingly reflected the growth of interest in small-scale food production as part of a broad set of lifestyle choices that emphasize rural living, food quality, working with animals and the land, and ideas of health and fitness associated with farm work and country life. There are also some links to “back to the land” ideals, which are manifested well beyond urban fringe locations.

Green belts, green wedges, and green ribbons have proved popular measures to conserve farmland in several European countries while combining this with other objectives, such as access to the countryside for urban populations, outdoor sport and recreation, retaining attractive landscapes, and securing nature conservation interests. In the United States farmland preservation policies have endeavored to slow the long-term loss of cropland to urban sprawl and other land uses. The greatest losses of cropland have occurred in the most populous counties through a “bow wave” of metropolitan expansion consuming farmland irrespective of its quality. The preservation policies are linked to four ideologically based lay discourses: the need to preserve a finite productive resource, the need to produce food in the national interest, environmental protection and retaining farming as a connection between nature and society.

SEE ALSO: Agricultural environments; Agriculture; Environmental issues in rural areas; Food security; Human geography; Local/global production systems; Rural geography

References

Blaikie, P.M. 1985. The Political Economy of Soil Erosion in Developing Countries. London: Longman.

Agriculture

Guy M. Robinson
University of Adelaide, Australia

The agroecosystem

Agriculture, the rearing of animals and the production of crop plants through cultivating the soil, has evolved for over 10,000 years since the first domestimations of wild plants began in the Fertile Crescent of the Near East. Sheep, pigs, goats, cattle, barley, and wheat were first domesticated in this area. There were five other independent origins of agriculture: East Asia (8400–7800 BP), producing rice, millet, pigs, chickens, and buffalo; Central America (4700 BP) and South America (4600 BP), producing potato, maize, beans, squash, llama, alpaca, and guinea pigs; North America (4500 BP), producing goosefoot and sunflower; and Africa (4000 BP), producing cattle, pigs, rice, millet, and sorghum.

Domestication spread from the Middle East into southeastern Europe, where a combination of improved cultivation methods and extensive trading networks supported first the Greek and then the Roman empires, both of which gave us the term “agriculture” to reflect the inherent link between human activity and the natural environment to provide food to sustain human societies and their domesticated animals. Essentially, human modification of the environment produces an agroecosystem that constitutes both an ecological and a socioeconomic system, operating at a number of different spatial levels from a single plot or field to broader systems of production within distinct regions linked by worldwide patterns of trade.

Unlike many elements of economic activity, the physical environment is of fundamental significance to the nature of the farming system even in developed countries, where farmers have capital at their disposal to modify certain physical characteristics of the land on which farming is based. The reliance on variable environmental conditions injects elements of risk and uncertainty into farming beyond those experienced by other areas of economic activity. Hence farming is strongly shaped by the natural ecosystems on which agricultural systems are based; the agroecosystem acknowledges the reciprocal relationship between environmental factors and agricultural activity.

In agroecosystems productivity is increased through human management and control of soil fertility, vegetation, flora, and microclimate. This control is intended to generate a greater biomass than that of natural systems in similar environments, but it can also create undesirable environmental consequences. In particular, farming alters the characteristics of soils. Agroecosystems are simpler than natural ecosystems, with a less complex structure and less diversity of plant and animal species. The long history of plant domestication has produced agricultural crops with less genetic diversity than their wild ancestors. The biomass of the large domesticated animals, such as cattle and sheep, is much greater than that of the ecologically equivalent animals supported by unmanaged terrestrial ecosystems.

Agroecosystems are more open systems than their natural counterparts, with greater numbers and larger volumes of inputs and outputs. Inputs include direct energy from human and animal
labor and fuel, plus indirect forms from seeds, fertilizers, herbicides, pesticides, machinery, and water. Under cultivation a higher proportion of available light energy reaches crops and, through crop harvesting or the consumption of crops by domestic livestock, less energy is supplied to the soil from dead and decaying organic matter and humus than is usually the case in unmanaged ecosystems in similar environments.

Agroecosystems are dynamic systems of flows of matter and energy, including water, solutes (nutrients), and solids (e.g., soil particles). Inputs include soil and nutrients derived from the weathering of underlying bedrock; energy from solar radiation; precipitation; transfers from adjacent land surfaces; and inputs from the farmer in the form of seeds, livestock, manure, fertilizer, animal feeds, and fuel energy. The major outputs are harvested crops and farm animals. Farm practices manage both inputs and outputs; for example, land drainage systems and irrigation affect water loss, while tillage, soil conservation, and crop rotation can control soil erosion. Management controls key cycles in the system, especially energy, water, and nutrients.

Agriculture has created artificial communities often dominated by a single species. By incidentally concentrating food and other key essentials for competitor plants (weeds) and animals (pests), this has led to damaging population explosions of weeds, pests, and disease vectors, which have had to be controlled. Hence, the use of synthetic herbicides, insecticides, and pesticides has become commonplace on many farms worldwide. This has contributed to the loss of biodiversity, including accidental loss of species which predate on pest species, of habitat for breeding and shelter via suppression of weeds, and of essential prey species of nontarget organisms.

Solar radiation provides the fundamental energy source to support plant and animal growth. The amount of radiation received at any location depends on latitude and albedo, which varies considerably for different surfaces. Solar energy is the driving force for recycling nutrients through the agroecosystem. Crucial aspects of this recycling are the carbon (or inorganic) and nitrogen cycles. Different farming systems and accompanying management strategies have varying effects on nutrient cycling, and so produce differential impacts on the soil base. For example, soluble nitrates are vulnerable to removal by leaching and so have to be maintained by careful management. However, nutrient deficiencies are generally easier to manage through careful husbandry than the constraints imposed by solar radiation, temperature, and rainfall. The most vital nutrients are nitrogen, phosphorus, potassium, calcium, and magnesium, which together make up 10% of a plant’s dry weight.

Sustained cultivation of the same plot of land removes nutrients from the soil, and land can be rendered infertile unless farming practices replenish nutrients via manuring, mulching, or adding artificial fertilizers. In addition, leguminous crops like beans, peas, clover, and lucerne add nitrogen to soils and hence have been widely used in crop rotations, which was first developed in East Anglia in the eighteenth century when such rotations consisted of clover, wheat, turnips, and barley. Grown in rotation, these crops enabled farmers to use a plot of land continuously without recourse to fallowing. In many developed countries, crop rotation has been replaced by adding artificial fertilizers to the soil, thereby substantially increasing overall energy consumption in farming. The overall efficiency of energy use declines as the degree of dependence on fossil fuel rises.

The dominant physical or natural resource inputs to the farming system are climate, water supply, and soils.
Climate

Average temperatures, the amount of precipitation, and their annual distribution are the greatest physical constraints on agricultural activity. More localized limitations are imposed by soil type, nutrient availability, topography, aspect, and drainage. Climate determines the broad geographical region in which any given crop can be cultivated. Despite extensions to the moisture and temperature requirements of many plants through modern plant breeding, there is still a strong degree of climatic determinism in the distribution of agricultural crops. This can be seen in agroclimatological classifications, which were first popularized in the early twentieth century in the form of zonation theory, employed by Russian geographers. This theory created a world zonal pattern based on moisture and temperature criteria translated into hydro and thermal ranges respectively. In the 1930s Glenn Trewartha produced a sevenfold world classification based on climate, vegetation, and soils. More detailed classifications followed, including those by Wladimir Köppen and Rudolph Geiger in 1936 and by Charles W. Thornthwaite in 1931; after 1945 more refinements were introduced, including that by Howard Penman and John Monteith in 1965, which focuses on evapotranspiration processes and plant-specific water consumptive use. More recently, the United Nations’ Food and Agriculture Organization’s Agro-Ecological Zones project represents the most comprehensive world classification of agroclimatic zones. This is based on land resource inventories, inventories of land utilization types and crop requirements, and land suitability evaluations, including potential maximum yield calculations and matching of constraints and requirements.

The key climatic variables for plant and animal growth are receipt of solar radiation, precipitation available for transpiration, and temperature during the growing season. Relationships between these variables are rarely linear, but optimum growth conditions can be recognized where plants give the highest yields. Generally, plants are cultivated commercially in a physical region around the optimum and well removed from the absolute limit to the plant’s growth, though economic limits to production are crucial to determining location. Provided there is sufficient water, the best determinants of crop growth are temperature and light; this effectively enables distinctions to be drawn between tropical, subtropical, temperate, and cool temperate agroclimatic regions. These are broadly related to different biochemical pathways of carbon dioxide (CO₂) fixation in photosynthesis, which, in turn, reflect basic physiological differences. Temperature affects how CO₂ is used, with distinct differences between plants that thrive in temperate environments as opposed to tropical.

For every crop, there is a temperature range within which growth and development can occur. Critical temperatures are the minimum (below which there is insufficient heat for biological activity); the optimum (where rates of metabolic processes are at their maximum); and the maximum (beyond which growth ceases and higher temperatures may cause harm). Some crops have particular temperature requirements, including certain variations between daytime and nighttime temperatures. These crop-growing habits in turn affect farmers’ phasing of planting and harvesting for annual crops, to make the most efficient use of the time and space available. Generally farmers utilize knowledge about plant-growing seasons, the period during which there is production of sufficient vegetative growth to support contiguous growth of the necessary yield-forming activities.
Microclimate modification

Various measures are employed by farmers to ensure that crops are given the best opportunity to thrive so as to produce good yields. At one extreme this can mean creating an entirely artificial environment in the form of a greenhouse, but more frequently measures to control growth under prevailing extremes of weather are applied. In temperate latitudes this often means attempting to combat damaging frosts at the start and end of the growing period of the plants concerned. These mechanisms include adapting crop selection to fit local microclimates, selection of favorable aspects (especially avoiding cold air drainage into valley bottoms) and genetic modification of crops for enhanced frost resistance.

Modifications to local climate are another common practice, such as the use of ridge and furrow or through short-term temporal adjustments involving the use of burners and smoke generators. The creation of shelters, such as temporary or permanent polytunnels, is one stage removed from greenhouse cultivation. Determining which of these various strategies is most appropriate depends on the details of crop physiology, topography, albedo, synoptic weather situation, and crop value with respect to the economic or social benefits of protection in relation to projected returns.

Longer-term protection to crops can be provided by the creation of shelter belts. Typically in mid-latitudes, these are a long-standing characteristic landscape feature. Examples include the planting of Lombardy poplars \( (Populus nigra) \) in the Netherlands and northern France and of the cypress \( (Cupressus \text{ spp.}) \) in southern France and Italy to provide shelter from the mistral. They have also been utilized on a large scale in major projects recently in China and the Sudan. Diffuse barriers are generally most effective in terms of protecting crops from damaging winds, as semiporous barriers avoid severe leeside eddying which can occur behind dense barriers. The effects of a barrier are generally experienced 15–18 times its height upwind. Hence, planting a series of cross-wind barriers may provide benefits from wind speed reduction and associated growing climate over a large area. Barriers can be used to provide direct protection to crops, especially high-value crops such as flowers and soft fruit. However, they can provide several other benefits, including the extension of growing season, promotion of early germination, reduction of water loss and exposure to wind chill and excessive evapotranspiration, avoidance of disease, prevention of soil erosion, and provision of habitats for beneficial organisms such as pollinators and the predators of pests.

Water

Water supply from precipitation is fundamental to all agricultural systems, though its management by farmers can compensate for problems in the natural supply. Water transports nutrients to and through plants, and is vital in the processes of soil weathering, leaching, and erosion. Water largely controls the input of nutrients to, and losses from, the system. Losses of water can occur through evapotranspiration, drainage to groundwater, and lateral flow as runoff and throughflow to streams. Water is stored in the soil, plant tissues, and the bodies of livestock. Worldwide, agriculture accounts for 70% of all freshwater consumed, and demand is increasing by 64 million m\(^3\) annually.

Agricultural practices disrupt the pattern of the annual water balance associated with the functioning of any natural ecosystem. The balance is the outcome of the input from precipitation minus losses from evapotranspiration. Water supplies can produce waterlogging of soils and
runoff whereas, if evapotranspiration exceeds precipitation, this produces a moisture deficit. Under deficit conditions, plants deplete the store of soil moisture until they wilt. In arid areas large soil moisture deficits can develop during the growing season. Farmers can intervene in this process by providing artificial irrigation. Some irrigation systems have a long tradition, notably in the Middle East and Spain, and can work well where seasonal water deficits are short and infrequent and occur in predictable cycles, and where access to stored water is easy and affordable.

Irrigation

Irrigation techniques vary tremendously. Surface systems, sometimes called flood irrigation, move water across the surface of farmland to infiltrate the soil, and have been the most common form of irrigation traditionally. Localized systems distribute water under low pressure through a piped network to discharge small amounts of water to each plant. Typical variants are drip irrigation, spray irrigation, and microsprinkler and bubbler systems. There are subsurface systems designed to deliver water to the plant base, while drip or trickle irrigation delivers water at the top of the root zone. This can be a highly efficient system as evaporation and runoff can be minimized. Drip systems can use advanced computing technology or high-labor intensity. In sprinkler or overhead irrigation, water is distributed across a field by overhead high-pressure sprinklers or guns. Sprinklers can be mounted on moving platforms connected to a water source via a hose. Traveling sprinkler systems are commonly used for intensively irrigated areas. A widely used form of sprinkler system is center-pivot irrigation, which leaves distinctive circular areas of irrigated land that are clearly visible in aerial photography. It usually consists of several segments of pipe joined together and supported by trusses that are mounted on wheeled towers, with sprinklers positioned along the length of the pipe. Water is fed into the system from the pivot point at the center of the arc. Hydraulic and electric motor driven systems, often with GPS, have generally replaced water-powered units. Irrigated areas increased sevenfold worldwide in the twentieth century, though the greatest contribution of irrigation to national food output occurs in countries where paddy rice is a significant crop, in conjunction with semiarid climates, notably Pakistan, China, Indonesia, Chile, Peru, India, and Mexico. Irrigation investment in Middle Eastern countries has been high in recent decades, typified initially by the opening of the Aswan High Dam in Egypt in 1969, though high costs have contributed to substantial debt in some developing countries. Worldwide there are over 3.25 million km$^2$ of agricultural land equipped with irrigation infrastructure, of which two-thirds are in Asia. Major concentrations are along the Ganges, Indus, Hai He, Huang He, Nile, Yangtze, and Mississippi–Missouri rivers.

Irrigation can dramatically improve crop productivity and enable extensions to the cultivated area in dry climate regions, but it can also produce negative outcomes associated with alterations to the natural water–salt balance, increasing the extent and risk of saline and alkaline soils. Capillary action can transport dissolved solids to the active root zone and surface areas. This has occurred on a large scale in Egypt, Iran, Iraq, Pakistan, and India, and parts of Australia and former Soviet Central Asia have also been affected, the latter associated with increased extraction of river water for growing cotton. Salinization of the soil surface is a major hazard of both open air irrigation systems and systems that draw on groundwater. Salinization is the accumulation of soluble salts of sodium, magnesium, and calcium in soil to the extent that soil
AGRICULTURE

fertility is severely reduced. Under irrigation, dissolved nutrients from the soil can be left as a residue on the soil surface unless attention is given to reducing surface evaporation or to periodic flushing to remove the salts. Salinization can also occur in coastal areas through the overexploitation of groundwater as a result of growing urbanization, industry, and agriculture. The overextraction of groundwater can lower the normal water table, leading to the intrusion of marine water. The latter was also associated with the Boxing Day 2004 earthquake and tsunami, which affected coastal zones in South Asia. Salinization affects over 322.9 million ha worldwide, one-quarter of which is in Australia and around 17% in the Near and Middle East.

Research into degraded salinized landscapes in southwestern Australia suggests that new agroforestry developments may help to restore biodiversity, desalinate soils, and improve wider ecosystem services, including water management and carbon sequestration. In the Yangtze River Basin in China, soil salinity has been reduced by leaching soluble salts out of soils with excess irrigation water. Soil salinity can also be controlled by managing the water table and flushing, in combination with tile drainage or other forms of subsurface drainage. In the Australian Riverina, salinity resulting from irrigation has been mitigated through a salt interception scheme, which pumps saline groundwater into evaporation basins, protecting approximately 50,000 ha of farmland from high water tables and salinity.

Agricultural soils

The primary agricultural management practice is cultivation of the soil, which acts as a reservoir of water, minerals, and nutrients needed for plant growth. Soils can vary considerably by virtue of changes in their structure, depth, texture, plant nutrient content, and acidity. These characteristics influence not only the types of crop that can be grown but also yields, with optimum edaphic conditions existing for particular plants. In general, most plants thrive in deep, well-drained soils (for example, potatoes, which prefer deep loams), but thin limestone soils can produce good yields of shallow-rooted cereals.

In addition to depth, other key soil variables are texture (the relative importance of particles of different sizes) and acidity (measured in the potential hydrogen (pH) scale: from 0, the most acid, to 14, the most alkaline, with 7 being neutral). The large particles of sandy soils provide light, well-drained land that is readily warmed for early spring planting. In contrast, fine clay particles retain water, are slower to warm in spring, and are heavy to cultivate. However, clay soils release potassium only slowly so they are less likely to suffer from potash deficiency. They can retain water in dry conditions but their tendency to become waterlogged has meant that underdrainage has proved particularly important in helping to improve yields. In temperate climates soil acidity is greater in areas in receipt of heavy rainfall. In such climates slightly acid soils can support good crops of potatoes and lucerne, though most crops prefer neutral or mildly acid conditions.

Geographers have produced several world classifications of soils, building on pioneering work by the nineteenth-century Russian pedologist V.V. Dokuchaev and work on soil-forming factors by Hans Jenny in the 1940s. From the 1970s there was a move away from a focus on genetic factors to definitional systems based on recognizable soil properties. New systems were developed, notably in Canada, the Netherlands, the United Kingdom, and the United States. These systems have introduced new terms to soil classification, such as “pedon” in the United States, an artificial cuboid unit
with a cross-sectional area dependent on the lateral variability of properties that define classes. The American system recognizes 12 soil orders and enables comparisons with the widely used classification developed in 1974 by the Food and Agriculture Organization of the United Nations. These various classifications highlight the delicate physical and biological balance that renders agriculture possible and restricts production in various ways, limiting it to cultivation of just 15% of the Earth’s landmass and a potential limit of around 25%, but only through huge investment in irrigation and other technological inputs.

Impacts of climate change on agriculture

Since the mid-1970s it has become clear that not only have there been worldwide climate changes throughout the past 10,000 years, which have undoubtedly affected the distribution of crops and livestock, but also that recent short-term climatic changes may be having major impacts on agriculture. This has given rise to several studies assessing the potential agricultural impacts of global climatic change. For example, the CO₂ concentration in the atmosphere is regularly monitored at Mauna Loa, Hawaii. In 2013 it was 394 ppmv, having risen from 316 ppmv in 1959, when routine recording began. This represents an average annual growth rate of around 1.4 ppmv per year, though annual increases have been greater in recent years than in the 1960s. Levels may rise to 500 ppmv by 2050. This may be beneficial for the productivity of certain plants, but increased competition from weed species may also result, and concentrations of nutritionally important minerals, including calcium, magnesium, and phosphorus, may also decrease under raised CO₂ levels. Outcomes from climate changes are complex and often difficult to predict.

Global climate scenarios are usually derived from general circulation models (GCMs), which have been used for forecasting the effects of changes to atmospheric composition or macroscale climatic properties. Historical or spatial analogs and incremental changes to the observed weather record are also used to produce climate change scenarios. Within the various scenarios are predictions that crop yields in many tropical and subtropical areas will decline for most projected increases in temperature, and that rises of more than a few degrees centigrade occurring in mid-latitude regions will also lead to falls in crop yields. In general, agriculture in developing countries appears to be more sensitive to climate changes than agriculture in developed countries. Rain-fed cropland is generally more sensitive to warming than irrigated cropland, and cropland is more sensitive than livestock. However, perhaps the most serious current predictions relate to future failures of food supplies through diminished supplies of water. Such failures are projected for the Sahelian region of Africa, South Asia, and large parts of Latin America as a consequence of shifting rain belts. It is predicted that by 2025 5 billion people may suffer from periodic deficits in water supplies, compared to 1.7 billion in 2005. A decline in the quality of water is also highly likely in parts of Asia and southern Africa. Yet some regions may benefit from global warming; for example, in Canada the cultivated area could be extended northward.

One likely consequence of global warming is rising sea level. In the Yangtze delta in China, for example, sea levels may rise 50–70 cm between 2000 and 2050. This will be of particular significance around Shanghai, where 95% of agricultural land is below the high astronomical tide level. Countermeasures are being prepared, including improvements in drainage quality and...
capacity, renewal and increase of pumping facilities, development of new crops that are tolerant of a higher groundwater table, and construction of a flood barrier.

Geographers have performed various impact studies pertaining to climate change at various scales and for specific major crop and livestock systems. This work has frequently used resource rating schemes to assign land parcels to broadly defined suitability or agroecological classes. Alternatively, there is crop yield analysis, based on interactions among crop growth factors and generating estimates of output per land unit, which can be calculated at both a farm and regional level. However, criticisms of climate change impact studies have argued that too many assumptions have been based on neoclassical economics, which assumes that land will be devoted to the best economic use, with farmers accessing the best available technology and adjusting their farming practices perfectly to suit the changing and variable climate. This approach largely ignores constraints on farmers’ choices and decisions based on noneconomic motives. However, starting with work on farmers’ responses to drought and famine in developing countries, research has shown that even in areas susceptible to extremes of weather and climate, farmers often downplay the importance of climate as a major factor in their decision-making. Indeed, there is relatively little evidence that farmers have responded to recent changes in climate by changing their farming practices or that they have much knowledge of potential future climate change. Instead they negotiate risk on an ongoing basis, as failure to do so would threaten their survival.

The Intergovernmental Panel on Climate Change (IPCC) defines adaptation as “adjustment in natural or human systems in response to actual or expected climatic stimuli or their effects, which moderates harm or exploits beneficial opportunities.” In many developing countries adaptation has tended to involve technical solutions imposed by government and overseas “experts,” without any input from local knowledge. Adaptation and coping strategies are sometimes combined because what may begin as temporary coping measures (e.g., reducing/increasing inputs) can become longer-term measures.

Key technologies that offset risk in the face of climatic variations are mechanical innovations (irrigation, conservation tillage, improved drainage) and biological science (hybrids, genetic modification). However, the scale of deviation away from so-called normal conditions may define the experience of climate change, thereby stretching the ability of technological innovation to provide “solutions.” Moreover, most adaptations to climate change take the form of incremental changes characterized by short-term and small-scale actions reducing potentially damaging loss of income or enhancing the benefits of variations in climate. In contrast, transformative change could involve combinations of technological innovation, institutional reforms, behavioral shifts, and cultural changes.

While research has focused on individual barriers to, and drivers of, climate change adaptation, few studies have considered how adaptation actually occurs. However, recent work has revealed the importance of formal institutions and communities of practice. The former are groups that follow rules and procedures created, communicated, and enforced through channels widely accepted as official, such as courts, legislatures, and bureaucracies. A government agency responsible for regulating natural resource management is one example of a formal institution in this context. Communities of practice are informal structures brought together through the social construction of knowledge. A community of practice exists.
when: members share a similar set of interests, expertise, roles, and goals; opportunities exist for members to interact with one another in both formal and informal spaces; and groups share a common practice or set of practices. Farm system groups are one example of communities of practice. They exist to trial new agricultural technologies and to disseminate agricultural research and development findings to their members. Members generally attend regular meetings and social gatherings, and receive the best available knowledge of farm improvement technologies including new crop varieties, soil moisture tests, new machinery, and animal and plant control measures.

Reducing greenhouse gas emissions

It is possible that changes within the agricultural sector may contribute to future reductions in greenhouse gas (GHG) emissions, and specifically to a reduction in emissions per unit of food (or protein) and fiber produced, including the three principal GHGs – carbon dioxide (CO$_2$), methane (CH$_4$), and nitrous oxide (N$_2$O) – all of which contribute to climate change. Positive mitigation can be associated with a range of measures, including the use of set-aside (taking arable land out of production), improved livestock feeding practices, and the growing of bioenergy crops. Agriculture accounts for 10–12% of total global anthropogenic emissions of GHGs, including 50% of CH$_4$ and about 60% of the N$_2$O emitted. However, the carbon mitigation potential of agricultural land is significant, and arable land has a significant CO$_2$ sink capacity. There are several land management options to increase soil organic carbon (SOC), including the use of animal manure, sewage sludge, cereal straw incorporation, no-till farming, agricultural extensification, natural woodland regeneration, and bioenergy crop production.

The United States Environmental Protection Agency has forecast that between 2005 and 2020 methane emissions from combined enteric fermentation and manure management will increase by 21%. The escape of methane from rice paddies is also a key component, but different management strategies from those practiced traditionally can reduce outputs. Another concern is methane emissions associated with the growth of cattle rearing, as a result of the increased demand for meat associated with growing affluence in the developing world. Changed dietary habits could be one solution, but this would go against current trends among the rising numbers of middle-class in countries such as China, India, and Brazil. Meanwhile, there is an upward trend in agricultural emissions of N$_2$O, which is linked to conversions of forest to agricultural land, increased use of nitrogenous fertilizers, and the rising numbers of farm livestock. The share of animal calories in diet is projected to increase from about 15% in 2013 to about one-third in 2050. Indeed, agricultural N$_2$O emissions are projected to increase by 35–60% by 2030 as a result of the increased production of animal manure and use of nitrogen fertilizer. However, advances in precision farming may offset some rises. This could involve matching fertilizer type and quantity to site and specific crop conditions using advanced computer and satellite technology, better timing and placement of nitrogen, and the use of controlled release fertilizers and nitrification inhibitors. Measures for reducing N$_2$O emissions from grasslands could include decreasing the size of cattle herds by increasing the productivity per animal via selective breeding, lowering the nitrogen content of urine through activities such as reducing the amount of nitrogen applied to the pasture, and restricted grazing, to decrease the number of urine and dung patches. However, the latter also requires investment in housing, appropriate slurry storage
basins and slurry application techniques, and indoor feeding.

**Intensifying agroecosystems**

The simplest way to increase agricultural output is via increases to the area of land being farmed. Over millennia this has enabled farming to be focused on the best land, in terms of climate, soil quality, and topography. Under pressure to produce more food to supply rising populations, farmers have exploited more marginal lands but at a risk of lower yields, more crop failures, and greater damage to the underlying biophysical base through erosion, seasonal flooding, drought, and biological infestations. The prevalence of these risks has promoted the adoption of innovative solutions such as terracing slopes to control erosion, irrigation in drier areas, and the development of crops better suited to particular environments.

With new productive land in ever-diminishing supply, the alternative is to intensify production on existing farmland by adding more nutrients to the soils, through fertilization, and by breeding higher-yielding varieties of crops. In order for such approaches to be successful, farmers must maintain soil fertility while avoiding problems linked to long-term intensive cultivation such as pests, diseases, soil erosion, and salinization. Livestock rearing may also be intensified through improved livestock housing, supplementary feeding, and more controlled breeding programs. Mechanization may improve the timeliness of agricultural operations while enabling farmers to work a large area of land more effectively.

The use of inorganic fertilizers in developed countries increased dramatically in the second half of the twentieth century, with average rates of use between 120 and 550 kg N ha$^{-1}$, often encouraged by fertilizer subsidies. Ideally, application rates should ensure that both crop uptake and efficiency of fertilizer use are high. However, in practice, efficiency of use varies from only 20% to 70%, meaning that fertilizer may be leached from the soil or removed from the fields via runoff and rainfall events, producing losses from 35 to 155 kg N ha$^{-1}$. This has contributed to problems such as the eutrophication of waterways and the contamination of groundwater and wells. In contrast, in many developing countries, the lack of use of inorganic fertilizers is giving rise to concerns for the future of soil fertility and agricultural sustainability.

Modifications to the growing medium, water regime, and inputs to the agroecosystem in the form of energy supplements, controls via herbicides and pesticides, and labor inputs supplemented by advances in mechanization have all contributed to substantial rises in outputs per unit area. However, biotechnical advances have offered perhaps the greatest advances in crop yields and animal production in recent decades. These advances can be seen as part of a continuum of agricultural development that has its origins in the pioneering work of Gregor Mendel in the mid-nineteenth century on plant hybridization, which, taken up over three decades later, introduced the modern science of genetics. Advances in livestock breeding had occurred in the eighteenth century in the English Midlands when Robert Bakewell implemented systematic selective breeding. His experiments with sheep contributed to numerous modern breeds, and he was the first to breed cattle primarily for beef. Others then built on his work to produce larger animals and with selectively bred qualities.

In the twentieth century, continued advances in plant breeding were at the heart of the Green Revolution, the name given to a package of high-yielding crop varieties (HYVs), fertilizers, pesticides, machinery, and irrigation applied to parts of the developing world, especially Mexico,
AGRICULTURE

India, Bangladesh, and Pakistan, with funding primarily from Western nations, where the new varieties had been developed, for example, through the pioneering work of Norman Borlaug and colleagues in the United States. The staple cereals of maize, wheat, and rice were the principal crops being improved to give higher yields in a series of advances from the 1960s to the 1990s. However, it proved difficult, and costly, to spread the crops beyond areas that were physically well suited to the improved crops, and they required high levels of investment to grow. From the early 1990s biotechnology entered a new phase, initially termed the Gene Revolution, but now better known by a variant popularized in North America: genetic modification (GM). This is the use of modern biotechnology techniques to change the genes of an organism, such as a plant or animal. In agriculture four main modifications have been dominant: providing protection against environmental threats such as pests and diseases; improving nutritional value; producing an organism that can generate new outcomes, for example drugs from animals; and improved qualities, for example, yields, hardiness, larger size for animals.

GM foods were first produced commercially in 1994, with the Flavr Savr tomato, a tomato engineered to have a longer shelf life. However, concerted ethical and safety concerns have been raised regarding GM foods, especially the health implications of human consumption of GM foods: for example, the potential for toxic reactions. These concerns have led to strict controls on the development of GM crops in some countries, notably the European Union, whereas in Canada and the United States the labeling of GM foods is voluntary.

The high-tech approach of GM, often allied to industrial-style farming, global food distribution, and multinational corporations, is in contrast to farming that is geared toward agricultural sustainability based on the application of key ecological principles stressing extensive, diversified, and conservation-oriented farming systems. Sustainable agriculture implies less specialized farming, which requires mixed systems of crops and livestock to reduce dependence on purchased fertilizers. It also requires farm-level decision-making that avoids disadvantageous impacts of production, such as the contamination of groundwater and removal of landscape features that may contribute to biodiversity. Organic farming is a term generally regarded as synonymous with sustainable agriculture, though it can take various forms (e.g., no-tillage, biodynamic agriculture, and polyculture). Organic farming systems rely on crop rotations, crop residues, animal manure, legumes, green manures, off-farm organic wastes, and aspects of biological pest control to maintain soil productivity and tilth, to supply plant nutrients, and to control insects, weeds, and other pests. Many countries employ organic certification schemes whose introduction has fostered the expansion of organic farming. However, there is an ongoing debate regarding the extent to which such sustainable systems can produce sufficient output to contribute significantly to the rising food demands of the world population, which is predicted to reach between 8 and 11 billion by 2050.

SEE ALSO: Agricultural environments; Agricultural geography; Agroclimatology; Environmental management; Irrigation; Soil fertility and management

Further reading

AGRICULTURE


Agri-food multinational enterprises

Ruth Rama
National Research Council of Spain (CSIC)

Early studies and general overview of agri-food FDI

According to the UNCTAD website, “FDI refers to an investment made to acquire lasting interest in enterprises operating outside of the economy of the investor. Further, in cases of FDI, the investor’s purpose is to gain an effective voice in the management of the enterprise. The most important characteristic of FDI, which distinguishes it from foreign portfolio investment, is that it is undertaken with the intention of exercising control over an enterprise.” The distribution of foreign direct investment (FDI) is geographically uneven and varies by industry (Senauer and Venturini 2005). However, the empirical analysis poses some difficulties: data regarding FDI and multinational enterprise (MNE) activity at the sector level is frequently patchy, restricting attempts to map these activities. The few exceptions (e.g., Gabel and Bruner 2003) contain little information regarding agri-food MNEs.

The agri-food sector comprises segments such as agriculture, food and beverage (F&B) processing, research and development (R&D), and retailing. Using UNCTAD data, Senauer and Venturini (2005) have shown that food MNEs were second only to the media in their degree of internationalization, as measured by the ratio of foreign to total assets, sales, and employment. In the 1990s the rates of internationalization of these firms also increased faster than those for MNEs operating in five other industries.

Although the agri-food sector is dominated worldwide by small and medium-sized enterprises, MNEs play an important role in these substantial industries. In 2005 the world’s 100 largest food and beverage MNEs (hereafter Top 100) accounted for around one-third of the world’s turnover in the F&B processing industry, with the 15 leaders amounting to 19% of the total (Rastoin 2008). The Top 100 produced around 50% of the patented innovations available worldwide to the F&B processing industry and auxiliary industries, agriculture included (Alfranca, Rama, and von Tunzelmann 2002). FDI and trade are related in many respects, and the agri-food sector is no exception. F&B MNEs have contributed to changing the geography of international agricultural trade by influencing local dietary habits. An econometric analysis noted that the presence of US F&B multinationals encouraged local agriculture in host countries that were agricultural exporters, such as the European Union, but stimulated increased imports of agricultural goods into deficit host countries (Barkley 2005).

By the mid-1970s, food deficits were encouraging international organizations, notably the Food and Agriculture Organization of the United Nations (FAO), to put food at the forefront of development debates. The first studies on the foreign activities of agri-food MNEs were published around 1975–1980, in the context of concerns about the possible negative effects of these companies on dietary habits, food safety, land concentration, marginalization of poor
AGRI-FOOD MULTINATIONAL ENTERPRISES

farmers, and a shift to cash crops to the detriment of staples in developing countries. Rastoin’s 1975 PhD thesis, which was based on economic and financial information for 87 very large food and beverages MNEs of different geographic origins studied over the period 1963–1972, was the starting point for AGRODATA, a database containing information on the Top 100 and their approximately 8000 affiliates, published by the Institut Agronomique Méditerranéen de Montpellier since 1976 (www.iamm.fr/iamm). Arroyo (1981) published case studies resulting from a large project carried out in 14 Latin American countries by French and Latin American academic institutions. The same year, the United Nations Centre on Transnational Corporations published an analysis of multinational agri-food companies for a larger number of developing host countries (CTC 1981). The research was produced in both the Global North and South, in different languages, and the research groups remained in fairly close contact (Arroyo 1981).

These early studies aimed mainly to analyze effects of multinational agri-food companies on developing host countries rather than to systematically investigate their geographic patterns of expansion. Nevertheless, some interpretations were provided, with institutional factors in the internationalization processes of the agri-food sector placed at the forefront of explanations. The studies proposed that agri-food MNEs were only the most dynamic part of a new international food system, which promoted the worldwide homogenization of food production, commercialization, and consumption. International organizations such as the World Bank, and some national states (notably the United States), were also proactively contributing to the construction of this new system (Arroyo 1981). This notion of food systems has been criticized since the 1990s because it allegedly fails to analyze power interfaces around the food sector (Arce and Marsden 1993). In this vein, some authors propose that the globalization of the agri-food system be treated as a “contested process” (Murdoch, Marsden, and Banks 2000).

The recent geographic expansion of FDI in this sector seems to confirm early interpretations, however, since it is largely attributable to profound political and institutional changes since the end of the 1980s. According to most recently available data, outward FDI flows between 1989–1990 and 2005–2007 doubled in agriculture, hunting, forestry, and fishing – food price rises of 2007–2008 may have stimulated such investments – and quadrupled in food, beverage, and tobacco manufacturing (Table 1). To a large extent, this is due to the liberalization of capital and trade, the formation or the enlargement of trading blocs (e.g., the European Union), and the adoption of market-led systems in erstwhile state-managed economies. A new configuration of shareholders also encouraged the rapid internationalization of F&B firms (the F&B industry is considered to be a noncyclical sector, attracting institutional investors such as pension funds). This has been accompanied by demands for relentless growth and the achievement of global goals by F&B companies, which often has contributed to weakening their embeddedness in home regions by promoting their worldwide expansion (Palpacuer and Tozanli 2008).

Most FDI still originates in the developed countries, and this is also true of the agri-food sector (Table 1). Some authors note substantial changes in the composition of outward FDI in this sector between 1989–1991 and 2005–2007 (Rama and Martínez 2013; Tozanli 2005). First, the share of developed countries in outward FDI flows fell: a moderate reduction (from 98% to 95%) for food, beverage, and tobacco processing, but a substantial one (from 91% to just 52%) for agriculture, hunting, forestry, and fishing (hereafter agriculture and related industries).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Developed</td>
<td>Developing</td>
</tr>
<tr>
<td>Agriculture and related</td>
<td>467</td>
<td>45</td>
</tr>
<tr>
<td>industries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Food, beverage,</td>
<td>12 223</td>
<td>253</td>
</tr>
<tr>
<td>and tobacco</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: UNCTAD 2009.

This reflects the dynamic expansion of flows originating in developing countries. While the search for land may have become less important for MNEs based in the Triad of Western Europe, the United States, and Japan, as predicted by Dunning (1993), it remains a crucial strategy for MNEs based in developing countries, especially those lacking sufficient natural resources.

Agriculture

The main objective of South–South agricultural investments is to obtain cheap food for the home market, bypassing large multinational trading companies based in the North (McMichael 2013). No comprehensive data on the geographic origin of agricultural capital is available. Foreign investments can be difficult to trace owing to alliances with domestic investors or operations through strategic transit countries (Borras et al. 2012; Cotula 2012; McMichael 2013). Furthermore, very different geographies of interests may be involved in a single project, since lenders, insurers, contractors, and suppliers may come from different parts of the world (Cotula 2012).

One study reports that 66% of the demand for agricultural land is by Middle Eastern investors (McMichael 2013). In Latin America, however, the presence of FDI in land deals from the Gulf States, China, South Korea, and India seems to be quite small (Borras et al. 2012). Much more important, these authors observe, is the investment of intraregional investors, whose presence is also noteworthy in African and Asian land deals (Cotula 2012). Some authors observe the presence of new global players, such as state-owned firms (UNCTAD 2011; Zoomers 2010).

Zoomers (2010) suggests that the current geographic expansion of agricultural FDI cannot be understood without taking into account the liberalization of land markets and the conversion of collective and customary land rights into formal individual rights, a policy followed by many developing countries since the 1990s. In many countries, neoliberal policies had also contributed to the dismantling of the public agencies formerly used to provide credit, technical assistance, and insurance to small farmers.

In some cases, the operations of investors based in developing countries are not limited to land investments in other developing countries, but also involve several segments of global
value chains (GVC), including processing and distribution facilities in Russia and the West (Nazareth Satyanand 2011; Pozzobon 2008). The presence of these new global players coincided with the divestment from land by many Triad-based MNEs since the early 1980s, due to the nationalization of plantations and shifts to more rewarding business lines (e.g., agricultural technology and food processing) (Oman et al. 1989).

In 2007, 56% of the inward FDI stock in agriculture and related industries was in developing economies, 37% in developed economies, and 7% in southeast Europe and the Commonwealth of Independent States (CIS) (UNCTAD 2009). Governments of host developing countries have been instrumental in facilitating such investments, especially when the land targeted by foreign investors was defined as available marginal land or was located in agricultural frontier regions (Borras et al. 2012). Borras and colleagues (2012) observe that such processes may take some time to crystallize, a position supported by other studies. In Chile, for instance, according to some authors, the military’s government privatization of national assets such as forests, and the partial rollback of land reform, amounted to the reprivatization of formerly public goods: after 1973, massive expulsions of peasants took place (Arce and Marsden 1993). In 2007 the stock of inward FDI in agriculture and related industries received by Chile, a country with relatively limited arable land availability, was nearly 2.5 times the stock received by Brazil, a very large country with a substantial agricultural frontier (UNCTAD 2009). In the tradition of economic geography, Munroe et al. (2014) advise that the play of power between actors should be considered in analyses of international land deals.

To summarize, there has been a major transformation of the geographic pattern of agricultural FDI, as a large share of the total consists currently of South–South investment. The thesis of an expansion of agricultural FDI toward developing countries (Heffernan and Constance 1994) appears to be supported by recent statistical data, but with a caveat (Rama and Martínez 2013). The principal foreign investors at present appear to be MNEs based in developing countries (rather than Triad-based companies), a change that was obviously difficult to foresee at the beginning of the 1990s.

One of the important questions posed by economic geography is whether MNEs of different origins operate differently outside their home country (Dicken 2000). The literature points to differences concerning agricultural FDI. European and US food processors and retailers have often preferred to outsource production to local suppliers of agricultural products rather than investing in land. Though their geographic linkages with foreign agriculture are difficult to quantify, they have been documented by case studies (e.g., Echánove and Steffen 2005; UNCTAD 2009). Oman et al. (1989) describe such arrangements as new forms of international investment (not simply FDI): MNEs that outsource agricultural production often also finance, organize production, and influence the technological practices of their local suppliers, though they do not own land. Some authors opine that contracting has tended to generate power asymmetries and the denationalization of agricultural production, since it allegedly shifts the locus of control from local producers to the agri-food MNEs (Gwynne 2006).

The literature illustrates a variety of situations in relation to control along the GVC (see Gwynne 2006). Faced with foreign investors and joint ventures, the small Mexican fruit growers studied by Echánove (2005) seem to have lost control of land owing to their inability to organize in the context of neoliberal policies adopted
in the 1980s (including drastically reduced public support for agricultural R&D and extension services). By contrast, a study of fruit and vegetable small producers in Almería (Spain) emphasizes how their ability to organize into cooperatives, the emergence of a cooperative local bank, and public support for agricultural education and new technology have generated a very different kind of relationship with foreign retailers, with the vertical downward integration of producers and other local firms within the GVC (Aznar-Sánchez and Galdeano-Gómez 2011).

Some authors observe that agricultural FDI or contract farming has sometimes contributed to the displacement of small local peasants in developing countries. The term “land grabbing” generally refers to large-scale, cross-border land deals carried out by MNEs or initiated by foreign governments, deals that may displace local peasant populations (Zoomers 2010). The physical extent of this phenomenon is controversial, ranging from Oxfam’s estimate of 227 million ha to the World Bank estimate of 45 million ha, with Africa hosting around 70% of such land grabs (Borras et al. 2012).

The FAO defines land grabs in terms of three aspects: (i) large-scale land acquisitions; (ii) the involvement of foreign governments in land deals; and (iii) the negative impact of such foreign investments on the food security of the host country. Borras et al. (2012) argue, however, that this definition is too restrictive. In their opinion, the scale of the capital involved should be considered, not only the scale of land acquisition, a framework that includes contract farming. Following their criteria, land grabbing would be substantial in at least 10 Latin American countries, not just in Argentina and Brazil as suggested by the FAO. Land grabs also may involve financial companies with no particular geopolitical attachments (McMichael 2013).

**Food and beverage processing**

The corporate geography of multinational food, beverage, and tobacco manufacturing is quite different from that of multinational agricultural activities. As in other manufacturing industries, FDI in food, beverage, and tobacco is mainly a North–North phenomenon. F&B MNEs have faced considerable constraints and challenges in developed countries, including an ageing population, the preference of some consumers for fresh or artisan products, and competition from retailers’ own cheap brands. These circumstances could have predicted a geographic southward shift of FDI in this industry, but the most important recipients of F&B FDI are still developed countries, accounting for 87% of inward FDI stock in both 1990 and 2007 (UNCTAD 2009). Indeed, although inward FDI stocks in developing economies also grew, inward FDI stocks in developed economies increased more quickly. As a result, the share of the North grew from 67% to 84% of total inward FDI stocks during this period.

Within the OECD (Organisation for Economic Co-operation) area, the composition of outward F&B FDI has changed with the emergence of new source countries such as Australia and New Zealand. Several case studies also note the emergence of F&B MNEs based in developing countries (Burch and Goss 2005; ECLAC 2005; Nazareth Satyanand 2011). Information provided by AGRODATA also points to the foreign activities of very large F&B companies based in the developing world. In 2012, 32% of the Top 100 were based in Western Europe, 29% in the United States, 18% in Japan, and 7% in other developed countries; 13% were based in developing countries and 1% in tax havens (Tozanli 2015).

As noted by Sim (2007, 2), “theories and explanations on the internationalization (or
expansion across national boundaries) of firms were largely based on Western TNCs.” Burch and Goss (2005) agree that most of the current literature characterizes the model of globalization as a process activated by the dynamism of large corporations based in the Triad, failing to consider MNEs based in developing countries. Even less research, they claim, has been directed toward agri-food MNEs based in such countries, notwithstanding their significant regional and even intercontinental activities. A popular theory (Dunning and Narula 1996) maintains that there is an investment development path (IDP) according to which countries gradually reach a stage of development that allows their companies to go international. Burch and Goss (2005) argue, however, that the stages of development proposed by this theory imply a degree of uniformity that is not apparent in the food and fiber sector. In developing countries, they claim, a highly sophisticated and productive food-processing sector may be established on the basis of very different ownership and organizational forms: peasant farming, plantations, cooperatives, collective ownership, and so on. The previous discussion seems to corroborate this point of view.

The empirical literature on the determinants of the location of F&B MNEs is now revised. These firms prefer to invest in countries with large internal markets, high GDP per capita, large urban populations, high levels of protection for these industries, and easy availability of cheap inputs and raw materials; membership in a trade bloc is seen as a plus (e.g., Pick and Worth 2005). Oligopolistic rivalry may also influence their locational behavior (e.g., Rama and Martínez 2013). These companies are not especially attracted by low wage rates, however (Makki, Somwaru, and Bolling 2004; Giulietti, McCorriston, and Osborne 2004).

Determinants of F&B FDI may vary by type of country. Makki, Somwaru, and Bolling (2004) find that per capita income appears to be negatively related to US FDI in food-processing and foreign affiliates sales, probably because consumers may prefer to increase their consumption of fresh and other less processed foods in very high-income countries. Within developing countries, by contrast, per capita income is positively related to US FDI in food-processing and foreign affiliates sales. Their findings support the view that food chains rooted in local and regional contexts respond to a growing demand for quality products, and may eventually counteract globalization trends in highly developed countries (Murdoch, Marsden, and Banks 2000).

Most authors believe that the F&B industry is a multidomestic rather than a global industry. In global industries (e.g., automobiles), the company establishes an international division of labor between its affiliates. In multidomestic industries, the firm is unable to establish networks facilitating worldwide combinations of inputs, intangible assets, and so on. Nevertheless, the emergence of trade blocs may contribute to changes in the agri-food sector. Regional integration may facilitate the offshoring of some low valued-added segments of the food chain or the national specialization of foreign affiliates and its corollary, intense intrafirm trade within the trade bloc (Chevassus-Lozza, Gallezot, and Galliano 2005). Agri-food MNEs adopting these formulae can spatially centralize the production of some items and display strong intrafirm trade, like corporations operating in global industries. Nevertheless, these are exceptions: FDI in the F&B processing industry displays the lowest level of geographic concentration (Senauer and Venturini 2005). The explanation offered is that proximity to the market is particularly important in consumer-oriented industries.

Arguing that we are moving toward a homogenized world, many have announced “the death of distance” or “the end of geography.” The “flat
world” thesis implies that economic activities can be performed as readily in India as the United States. At first sight, the substantial decentralization of F&B MNEs seems to confirm this point of view. However, places remain very important in agri-food production, much more so than in other sectors. There is a “geography of quality”: consumers associate the high quality of particular foodstuffs and drinks with specific locales (Mansfield 2003). This country of origin effect (as labeled in the management literature) can limit the internationalization of F&B firms via FDI, promoting exports instead. Other studies also show that physical and cultural distance matter in this sector. With the exception of a few very large companies, F&B MNEs tend to expand within their respective home region (e.g., the European Union) rather than globally (Filippaios and Rama 2008; Gardner and McGowan 2010).

F&B companies are more likely to expand to host countries sharing the same language or other cultural features with the home country (e.g., Filippaios and Rama 2011). This is not surprising because F&B are culturally bound products. This is consistent with the Uppsala school of thought, which proposes that MNEs follow a sequence, from their home base to countries with greater “psychic distance.” According to this theory, an MNE’s penetration into different markets proceeds according to their cultural similarity to its home country. F&B MNEs may also be attracted by subnational markets displaying cultural proximity, as is the case of Mexican F&B MNEs penetrating the United States market by tapping in to the demands of Mexican immigrants (ECLAC 2005).

The organization and origin of capital may also influence the corporate geography of F&B firms. Anastassopoulos and Rama (2005) propose that Japanese F&B MNEs centralize production in fewer countries than their Western counterparts because their main businesses consist of exports from foreign locations to the home country, and the production of high-tech inputs for the food industry. They thus require contacts with fewer national markets. Palpacuer and Tozanli (2008) propose that the characteristics of stakeholders also influence the geographic patterns of F&B MNEs. Companies with smaller plant size, family or farmer ownership (cooperatives), and European capital follow a regional (EU), rather than a global strategy.

**Retailing**

Wrigley (2000) published one of the first comprehensive studies on the globalization of retailing capital. Examining food retailing, he noted the importance of mergers and acquisitions (M&A) as a strategy for firms to acquire global scale. Confirming previous studies, he observed the profitable and relatively easy penetration of Carrefour (France) into Brazil in the early 1990s. The takeoff of supermarkets occurred half a decade later in countries with previously state-controlled economies, such as China (Reardon, Timmer, and Berdegué 2004), but now is moving very fast.

The literature suggests that the presence of modern retailing may hinder or stimulate F&B FDI, or fuel competition in specific markets. In the 1980s and 1990s insufficient development of modern retailing facilities limited the spread of F&B MNEs in some countries, regardless of whether they were industrialized, because F&B requires refrigeration facilities or the frequent substitution of products on retailers’ shelves for optimal consumption. Elsewhere, the development of modern retailing has fueled competition in F&B processing markets. In Brazil, international brands faced unexpected competition from cheap local brands when smaller domestic processors were able to use these new channels.
to market their products (Farina and Viegas 2005). In other cases, the internationalization of retailing goes hand in hand with that of F&B production, as shown by examples of small F&B companies that follow a retailer abroad. This exemplifies the “follow the client” strategy studied by the international business literature in other industries.

It would be an error to conclude, however, that distance does not matter for retailers. Large multinational retailers prefer to expand in the home region rather than globally (Rugman and Girod 2003). The global sourcing of both fresh and processed products seems to be increasing, however.

Research and development

The globalization of business activities often starts with exports, followed by production activities abroad, and only then foreign R&D investment. The internationalization of R&D corporate activities has become increasingly common. Firms innovate abroad to adapt their products to foreign markets, to benefit from locally bound pools of know-how or to adjust to local regulations (Hotz-Hart 2000). Cross-sectional studies show that F&B MNEs are especially likely to undertake some R&D activities abroad, probably to adapt their products to different national tastes and food safety regulations. However, the control and management of such activities is retained within the Triad (Rama and Martínez 2013).

According to one study, major European F&B MNEs tend to retain their core technology within the EU region, albeit not necessarily in the home country (Martínez and Rama 2012). This supports the view of evolutionary economists (Archibugi and Michie 1995) that the internationalization of R&D may be largely a regional process, principally involving European MNEs innovating in other European countries.

Future directions in research

More research is needed to improve our understanding of the corporate geography of agri-food MNEs. Several of the topics discussed in this entry are new or have received insufficient attention. GVCs controlled by MNEs based in the South are a new phenomenon; until recently, the literature assumed that GVCs were exclusively controlled by companies based in the North. The role of cultural proximity concerning subnational markets for ethnic food is another new phenomenon which may acquire some importance, given the current scope of migration flows.

More meta-analyses are needed to deepen our understanding of a variety of situations regarding control along GVC. The wealth of published information and analyses has yet to be fully systematized. Land grabs also deserve more conceptual work, in addition to careful analyses of their spatial dimensions and timing. The lack of methodological consensus has hindered our understanding of this phenomenon. The role of trade blocs in the emergence of agri-food MNEs that centralize and specialize production, like MNEs in so-called global industries, is another interesting avenue for research. Finally, the geographic aspects of the relationship between modern retailing and F&B FDI deserve more attention. Research on these two types of FDI and their respective geographic patterns has proceeded independently to date, but there is evidence to suggest that they may be related phenomena.

SEE ALSO: Emerging market: southern corporation; Firms; Global commodity/value chains; Globalization
References


AGRI-FOOD MULTINATIONAL ENTERPRISES


Further reading
Agroclimatology deals with the complex interaction between climate and agriculture. The significance of agroclimatology continues to increase with the exponential growth in human population and the resulting demand for food. The need to provide sustainable and robust continued sources of food and fiber has become a major challenge facing humankind. Some of the broader questions that encompass this challenge include:

• What and where can crops grow?
• What can be done to improve their productivity under limited natural resources?
• How can increased productivity be achieved with minimal impact to the climate and the environment?
• How would the vulnerabilities to agriculture crop production alter under climate change?

Other questions often posed by the agricultural community include:

• What typical planting dates would provide the most potential for highest yield?
• What are the typical risks and vulnerabilities expected in a growing season?
• What can be done to reduce their impact on crop yields?
• What yield can be expected in a typical growing season, and how does this change during an El Niño/La Niña year (when the Pacific sea surface temperatures are higher or lower than climatology and significant temperature and rainfall extremes are noted in many parts of the globe)?
• What is the expected impact on food prices and global trade?

The above questions are examples. The information sought would depend on:

• The type of crop being grown (commodity or perennial or horticultural).
• The purpose of growing these plants, that is, whether it is for food, fiber, fuel, or environmental considerations (such as fixing soil nitrogen, or for regulatory compliances).

The questions would also be specific to the geographical location as well as to what regional regulations and policies need to be considered. Answers to such questions are routinely sought by the broader community and form the crux of agroclimatology.

Background and the evolution of agroclimatology

Agroclimatology is a subgenre of climate science that combines the disciplines of agronomy and climatology. Agronomy deals with the scientific and practical understanding and application of the integrated knowledge of soils, crops, nutrients, environment, and management practices with the objective of maximizing yields sustainably. The complex interaction among factors that contribute to crop growth and sustenance is of particular interest. Climatology deals with the understanding and analysis of long-term
(decades to centuries) patterns of climate and various factors that constitute it through interactions within the biosphere (biological ecosystems including plants), lithosphere (earth and soil), hydrosphere (water and moisture), cryosphere (snow and ice), atmosphere (air), and humans.

Agroclimatology can thus be considered as a branch of science wherein climatic information is synthesized, analyzed, and applied to improve agricultural practices. Historically, belts of civilization formed around regions that were found to be climatically conducive for agricultural practices. For agriculture to sustain and thrive, the primary variables of interest include access to water or rainfall, adequate moisture and nutrient content in the soil, abundant radiation for photosynthesis, the correct temperature conducive to crop growth, and a reliable year-to-year climate pattern that allows agronomic planning. Variables relevant to agricultural variables in conjunction with information on plant response to environmental stimulus or stress constitute the core of information compiled in agroclimatic datasets.

Although many aspects of agroclimatology and agrometeorology are similar, they are considered separate fields, in part, because of their temporal scales. Agroclimatology, for example, is the study of hydrometeorological variables over an extended period (i.e., months to decades) whereas agrometeorology is concerned with variables over a shorter period (i.e., minutes to weeks). For example, agroclimatology could be used to make decisions as to what crops or horticultural plants to grow in a certain region (also known as hardiness zone maps), while agrometeorology would be used in protecting crops when a frost risk or some other immediate weather-related feature needs to be considered for agricultural sustenance.

Just like agrometeorology, agroclimatology can be applied to a broader domain or specific area, such as the agroclimatology of the Corn Belt in the Midwestern United States, or for a specific location and purpose, such as the start of a planting season in a particular field or location.

The focus of agroclimatology has evolved to understand and communicate the intricate interactions between long-term agroclimatic variables (e.g., precipitation, temperature, radiation, evaporation, soil moisture, soil temperature) that affect the agricultural system from production to distribution of food, feed, and fuel.

Indeed, this aspect of the geographical locations, regions climatology and the suitability for type of plants that can be grown is well recognized. In the simplest form, plant hardiness and heat zone maps are available that relate the cold and warm season temperatures and provide guidance on which plants can generally thrive in a given region. Additional guidance on crop and plant zones has been recognized historically through delineation of “life zones.” For example, Merriam (1898) mapped the United States to identify regions based on which crops can prosper in different combination of heat and moisture availability. Additional details related to life zones were developed by Holdridge (1967): these combined information on mean annual temperature for a region, altitude, mean precipitation, and the evapotranspiration water loss. This spatiotemporal information of climatic state and the kind of agricultural variety that can thrive has thus been a backbone of agroclimatology.

Agroclimatology encompasses a broad variety of studies from increasing agricultural productivity to how to adapt to a changing climate. Therefore, agroclimatology is an important tool that can be used for natural resource management, maintaining soil health, animal production, and as an indicator of future vulnerabilities of agricultural productivity and sustainability.
Some agroclimatology terminology

Agroclimatological information is captured in a series of variables. Over the years, technology has constantly evolved but the basic desired outcomes that form the agroclimatic information have remained the same. These typically include the meteorological and climatic variables (temperature and rainfall), derived parameters (such as growing degree days and sunshine hours), indices (such as the Palmer drought severity index, the standardized precipitation index), and patterns (first and last frost date or start and end of the rainy season).

As a first step in developing the agroclimatology of a region, it is important to document the specific climatology of that region. The primary agroclimatic variables of interest can include rainfall amounts (determines the water content in the soil), radiation available for photosynthesis, and the temperature of the environment affecting crop growth. Additionally, some terms may be different depending on the geographical region of interest. For example, in temperate regions, the crop growing season entirely depends on the frost-free dates; while in the case of tropical regions, the rainy season almost entirely determines the type of crop and its entire crop life cycle from sowing to harvesting. Maps and data showing the start and end of frost season or rainfall season are used as guidance for the crop growing season. Crops are also heavily impacted by soil variables such as the loss and availability of soil moisture and temperature due to evapotranspiration, soil-water demand, and soil-water holding capacity. Therefore, maps and information related to typical evaporation loss and water-holding capacity are also considered helpful from the agroclimatic perspective.

Agroclimatic data can be integrated into crop-cycle centric activities. The growing degree-day (GDD) is an example of an agroclimatic measure that is grounded in science and has practical utility. GDD is a measure of the amount of heat units accumulated over the crop region. Plants are known to begin their growth after a certain base temperature is reached (e.g., 50°F or 10°C) and plant growth slows at a high-temperature threshold (e.g., 86°F or 30°C). The GDD can then be estimated as the sum of daily mean temperatures that are above the base temperature required for plant growth; that is, the base temperature is subtracted from the daily average temperature to obtain the GDD per day. For corn (maize) in the United States, the base temperature is taken as 50°F (10°C). The daily mean temperature is simply taken as an average of daily maximum and minimum temperatures. For any maximum daily temperature above the high-temperature or minimum daily temperature below the low-temperature threshold, the value is replaced by the threshold value. Continuing with the above example for corn in the Midwestern United States, if the daily maximum temperature is 90°F then it is replaced as 86°F in computing the GDD averages. The sum of GDD or heat accumulation is then linked to agronomic aspects such as plant growth and pest potential. For example, a GDD of around 100 is considered adequate for plants to sprout from the ground, and 600 is considered to be suitable for tassel development. A GDD of 1400 is considered for plants at full height, and full maturity for many corn varieties is about 2700. In the United States Corn Belt, crop varieties typically need a GDD around 2000–3000 depending on the location, genetic makeup, or hybrid variety. Similarly, a threshold of GDD for hatching and feeding of leaves at around 500 is considered in the case of some plant pests. For a given location and possible planting date windows, agroclimatic curves or maps showing the GDD potential for a climatological growing season (i.e., frost-free...
period or rainy season) can be used to ascertain if a particular crop variety would or would not be appropriate (that is, whether the GDD needed for the crop cycle is expected or not). The GDD and pest information are also used in planning the use of pesticide recommendations, especially when large tracts of farmland need to be covered for effective agronomic management practices. Other such agroclimatic information could include data and maps for the number of days when the temperature is expected to exceed 95°F (35°C) and crops may incur damage, when nighttime temperatures are above 78°F; the number of days of hail occurrence, or the number of days when rainfall can exceed around three inches, or other such thresholds that can cause regional ponding and impede farm operations.

Agroclimatological parameters can also capture climate variability information such as changes in typical planting dates as a function of the El Niño/La Niña phase or the expected crop yield as a result of El Niño/La Niña. Inherent to climate variability information is that agroclimatic analysis could also include frequency or interval-based information, such as the duration of the number of consecutive periods with temperature or rainfall in excess of a certain threshold or specific to a crop cycle threshold. Plant growth is reduced if there is a lack of rain two weeks after planting; therefore, soil moisture or rainfall expectations relative to the planting date also become important agroclimatic information.

Some applications of agroclimatology

One of the main objectives of agroclimatology is to effectively assist farming practices, thereby increasing agricultural productivity and yield. By observing records of precipitation, temperature, and hours of sunlight in a region, growers have a better opportunity to predict which crops and varieties are better suited for the growing conditions of a selected region. A typical agroclimatic assessment would provide data, observations, model estimates, and maps that can be interpreted by communities or agronomic advisors and includes information on the timing and variety of crops along with associated risks and management practices.

Some additional examples of agroclimatic applications are summarized here.

Natural resource management: on a global scale, agricultural activities are one of the most water expensive enterprises. Since the type and extent of crop production can heavily influence the hydrology of an area, agroclimatology could be utilized to develop methods that would improve management of water availability. By combining climatological information with that of soil and plant science, evaporative demand can be calculated to determine specific irrigation requirements and applications throughout the year. Predicting changing hydrological patterns could be improved through urban planning by designing innovative wastewater transport and treatment facilities that can safely adapt to changes in precipitation and water runoff (particularly with fertilizer use).

Droughts and hydrological extremes: climate variability and hydrological extremes such as droughts and floods play a crucial role in determining suitable agricultural practices. Drought indices such as the Palmer drought index and the standard precipitation index are useful climatic measures to ascertain anomalies in rainfall or effective rainfall (precipitation–evaporation) over a given location. Knowledge of drought occurrence and timing aids in the understanding of the risk involved and defines corresponding management measures to cope with it. Similarly, heavy rainfall events including hail and floods can cause severe crop damage in cases of insufficient preparation. Mitigation measures, such
as irrigation plans in areas where deficits exist, drains in regions where excess rain needs to be diverted, and shelters, can prevent water logging and soil erosion.

Soil health and nutrient management: as the availability of water and soil have an interdependent relationship, agroclimatological information that provides knowledge related to both is useful when preparing land management sectors for managing soil health. For example, erosion is intensified in regions of heavy rainfall climatology. Soil moisture retention and nutrient storage capabilities are also influenced by changing climate. For the use of fertilizer, the type, amount, and timing required also vary with changes in climate and nutrient availability; agroclimatic information is useful guidance for this.

Animal production: in addition to plants, agroclimatology can also be of use in animal production, another subset of an agricultural enterprise. Different animal species have adapted themselves to have optimum growth, lactation, and reproductive functions if living under specific climatic conditions, including temperature and the availability of food resources. Beyond these parameters, animals are subjected to a variety of stresses. These environmental stresses, caused by prolonged exposure to inadequate temperatures, can hinder an animal’s ability to properly regulate their body temperature as well as control their energy, hormonal, and water balance. If their stress stems from a reduction in the quantity or quality of their food, then their growth and nutrient absorption can be adversely affected as well. These impacts collectively influence not only the animal’s health but also the health of their offspring and the production of eggs and milk. Therefore, knowing how weather, climate, and an ecosystem interact with one another is critical to the productivity and welfare of animals during animal production, and agroclimatic maps can assist in the animal facility operations.

Climate modeling studies: agroclimatic information is not only used in developing practical agronomic decisions but can also be used in climatic analysis. For example, climate model studies can have enhanced interactivity between the atmosphere, hydrology, and the land surface by incorporating rules for different regions to have a change in vegetation fraction as a function of agroclimatic inputs. This is known as dynamic vegetation feedback in the climate models; agronomic data and principles are used as the driver for these model algorithms to make them realistic.

Agroclimatology and climate change

As regional and global climates continue to change with greenhouse gas emissions, land use/cover changes, and a warming planet, food security is interlinked to future sustainability. Agroclimatic information is of interest in the understanding of impacts due to climate change (e.g., if temperature were to exceed a growth threshold such as 95°F, then it could have a deleterious impact on projected crop yields). Agroclimatic information is also important is developing adaptation/mitigation strategies related to climate change. For example, there is a tendency of a poleward shift in ecological zones of different crops, pests, and weeds due to increased temperatures in the upper latitude regions. Agroclimatic maps for current and projected climate can help prepare food and ornamental crops for the inevitable problems or opportunities that arise from such changes. Being aware of these shifts can also assist pathologists and seed companies in preparing for such demands and information needs.

Agroclimatological information thus has a bearing on the regional land use/cover in terms of agricultural crop land extent and type. It is
important to note that agricultural landscapes can alter regional climate by altering the regional albedo and temperature. Agriculture combined with irrigation has also been known to change regional moisture budgets and the atmospheric energetics, which ultimately can alter rainfall patterns. Agriculture from both animal operations as well as crop and forest production contributes to the greenhouse gas budgets. For instance, plants can absorb carbon dioxide and act as a carbon “sink,” while soils can both absorb and, when disturbed for tilling and other operations, release carbon back to the atmosphere, and so act as a “source.” Thus, agroclimatic information is also important in making decisions and synthesis related to regional climate resiliency. Ultimately, agroclimatology is a science that is best utilized in mitigating climatic risks and promoting sustainable practices for optimal yields and the longevity of various natural resources, and is intimately tied to the climate change adaptation/mitigation discussion.

Broader opportunities and challenges for agroclimatology

Agroclimatology is entering an era where it is poised to be of prime importance for global challenges, such as food security and sustainable agroecosystems in a changing climate. A number of tools and sensors are emerging with capabilities for higher spatiotemporal mapping. These include integration of information from geospatial and remote-sensed datasets and wireless sensors, mobile technology, and distributed/spatial models. Agroclimatology is expanding into a field that determines information of spatiotemporal patterns locally and geospatially in an interconnected world for agronomic decision-making. Scaling this information within the agroclimatic variables and maps is needed to provide patterns and knowledge that can aid in the decision-making from global to regional to farm scale. An example of such global linkages of agroclimatic patterns and regional and local decision-making can be seen in the role of El Niño in creating drought or crop yield changes in a certain region (e.g., South America). This reduced crop yield in South America could mean higher trade potential and crop prices for farmers in North America. As a result, deciding what crop to grow, the possible returns and risks would be the norm for future agroclimatic decision tools.

This combination of agriculture, climate, economics, management decisions, and trade is the likely trajectory of future agroclimatic applications. Teleconnections in different regions and their impact on regional agricultural practices and decisions will be an area requiring newer techniques for mapping such connections and knowledge emerging from these data and patterns. Another example of the growing need for agroclimatic information is in the study areas of crop bioengineering and genetic hybrid varietals, and integrating higher spatiotemporal resolution information that can be utilized in the development of customized seed varieties. It also follows that while climate models currently provide weather and climate information that is then linked into agronomic models and for decision tools, the next generation of coupled models will have agroclimatic outputs and products that can assist with decision-making and assessments.

Two emergent challenges with these finer resolution products include higher uncertainty in projections and developing results and an information base that would be transferable from one locale to another. This will require redesigning the agroclimatological curriculum and educational materials and tools for students, farmers, and stakeholders from the global food
security perspective. Tools and educational materials are also needed to help understand the risk, uncertainty, and value of agroclimatic products and to assist with decision-making. This will include more accessible documentation and awareness of how predictions and analyses are made and what combinations or factors are being used in making a certain decision (e.g., the likelihood that El Niño during the growing season can change the planting date or cause drought stress in some regions, and if the El Niño does not evolve as projected, then the decisions can be modified and tailored to align with changing agroclimatic patterns).

While the impact of climate on agriculture has been the defining feature of agroclimatic information, in the context of the changing world it is becoming apparent that agriculture as an enterprise also has an impact on regional climate through water use, fertilizer use, greenhouse gas emissions, and land cover albedo. Future efforts will likely have more agroclimatic products providing information on this coupled feedback with newer variables and maps to assist regional planners.

Lastly, the role of humans through management practices continues to be a challenge in developing agroclimatic information because it is still poorly understood and documented. It is likely, and expected, that in the coming years the human influence within the agroclimatic variables will be better documented and utilized in agronomic decision-making.

Conclusions

Agroclimatology is a vibrant branch of science integrating fundamental and applied scientific information from multiple disciplines under agriculture and climatology, to develop a knowledge base that is increasingly important for a world seeking to feed its growing population using sustainable means.

Figure 1 shows a summary of an agroclimate decision cycle developed by agroclimatologists focusing on the US Corn Belt (Takle et al. 2014). The outer calendar indicates the approximate time during the year when different agricultural decisions are made. The inner calendar identifies the different impacts as a result of different weather and soil conditions.

Indeed agroclimatic principles, objectives, and fundamentals will remain the same as they have been, but the future is geared toward integration of newer technologies and broadening of applications where agroclimatic information is needed and will likely make its value grow exponentially in the coming decades.

Acknowledgements

This entry benefited from discussions, reviews, and comments by Marshall Shepherd at the University of Georgia, Nicole Shebesh, Renee Oblinger, and Ken Scheeringa at Purdue University, Gene Takle at Iowa State University, Roger Elmore at the University of Nebraska at Lincoln, and Olivia Kellner at Midwest Regional Climate Center at the University of Illinois. Dallas Staley provided editorial assistance. Parts of the work benefited from USDA NIFA projects 2011-67019-20042 on drought triggers through Texas A&M University, 2011-68002-30220 on making climate information useful to usable (U2U), and 2015-67023-23109 on the role of climate variability and global trade.

SEE ALSO: Agriculture; Climatology; Food security; Hydroclimatology and hydrometeorology
Figure 1  Example of agroclimatic impacts and decisions. Takle et al., 2014. © Earth Interactions. Used with permission.

References


Further reading


Agroforestry

Krisnawati Suryanata
University of Hawai‘i at Mānoa, USA

Agroforestry refers to a natural resources management system in which farmers deliberately grow woody perennials on the same land as agricultural crops and/or animals, either in some form of spatial mixture or in sequence. Although it has been a common practice around the world for centuries, the term “agroforestry” was coined and institutionalized in the 1970s when a broader interest in indigenous technical knowledge renewed interest in traditional practices. Descriptive and anecdotal information from different parts of the world had indicated that intercropping systems have fewer pest and disease problems, improve nutrient cycling, enhance the chemical and physical properties of soil, conserve moisture, and make generally efficient use of a range of farm resources. From the standpoint of environmental stabilization, agroforestry systems may reduce erosion and provide environmental services such as watershed protection and biodiversity conservation. An influential report commissioned by the Canadian International Development Research Centre (Bene, Beall, and Côté 1977) highlighted these advantages and called for global recognition of the key role that trees play on farms. It recommended the establishment of an international organization that would support, plan, and coordinate research on agroforestry. The result was the establishment of the International Council for Research in Agroforestry (ICRAF) in 1978, followed by the inaugural publication of the journal *Agroforestry Systems*, dedicated to publishing research related to agroforestry.

Throughout the 1980s, agroforestry scholars put together an inventory of existing agroforestry practices and knowledge, in order to understand the underlying processes of time-tested practices and the applicability of scientific principles to improve them (Nair 1992). The home gardens (*pekarangan*) and mixed gardens (*kebun talun*) of Java were cited, for example, to show the diversity and complexity of the structure and function of tropical agroforestry, in which intermingled trees, bushes, climbers, and herbaceous plants in a multistory vegetation system provide excellent protection against tropical storms (Wiersum 1981). Another commonly cited agroforestry practice was the *taungya* system, introduced by British colonial administrators in nineteenth-century Burma to establish teak (*Tectona grandis*) plantations on forestlands. Under this system, landless peasants were given permission to grow food crops on forestlands for three to four years in return for performing reforestation tasks. The system later spread throughout Asia, Africa, and Latin America and remains the primary method for recruiting and mobilizing labor in reforestation schemes.

While the bulk of studies at this time focused on the ecological properties of agroforestry, a few social scientists drew attention to the fact that agroforestry practices provide critical opportunities for otherwise disenfranchised groups. For example, Rocheleau (1987) demonstrates how women mobilize agroforestry strategies to make the best use of the minimal landholdings allotted to them. Proponents of cultural survival have argued that indigenous peoples use agroforestry systems to perpetuate their livelihoods and to
AGROFORESTRY

protect their cultural identity (Clay 1988). Dove (1990) argues that the diversity and complexity of mixed gardens in Java, which incorporate a wide range of cultivars with high use-value but low exchange-value, might have provided peasant groups with the means to effectively resist the extractive propensities of the state. In contrast, the taungya system serves as a tool of the state to effectively control shifting indigenous cultivators and to establish forest plantations with minimal labor costs (Bryant 1994).

Driven by narratives of economic growth and sustainability, research at ICRAF (renamed the World’s Agroforestry Centre in 2002) during its first two decades was geared toward improving agroforestry technologies such as alley cropping, hedgerows planted on contour lines, or improved fallows. These technologies were aimed at maximizing positive interactions between soil and plants in order to push the biophysical limits of the system and achieve higher productivity while stabilizing its resource base. Armed with this promise, institutional actors in forestry, major donor agencies, and national governments joined forces to promote integrated conservation and development projects (ICDP) utilizing agroforestry practices. However, optimism was short-lived. Reviews of large upland agriculture and conservation projects in Indonesia showed that the new farming systems depended heavily on unsustainable subsidies, and many farmers reverted to old practices soon after a project ended. Elsewhere, alley cropping (a prototype of agroforestry technology developed at ICRAF) turned out to have a low adoption rate among farmers. In addition to the technical difficulties of designing the system, the elevated labor cost of managing grasses and legumes in hedgerows was an obstacle that prevented farmers from adopting alley cropping. In the end, “the alley cropping experience has done considerable harm to the reputation of agroforestry research” (Sanchez 1995, 43), leading to its abandonment. Today the World’s Agroforestry Centre continues its investigations into improving agroforestry management, expanding tree species diversity, and improving farmers’ income from commodity production. It also examines institutional and policy innovations to reward agroforestry practices that supply environmental services such as watershed protection and carbon sequestration.

Political ecologists seek to expand the discussion beyond technocratic and managerial concerns, and draw attention to agroforestry as a site of contentious struggle with overlapping property and labor claims. Many societies recognize the separation of tree tenure and land tenure (Bruce and Fortmann 1988), and tree planting brings about new rights and pressures on existing property relations. The relative permanence of trees may impose restrictions on other uses, and the shading effect of trees may displace crops underneath as well as in neighboring fields. It is therefore critical to understand the overlapping, multiple claims that often develop following tree planting and the adoption of agroforestry (Rocheleau and Edmunds 1997; Schroeder 1999; Suryanata 1999), and how tree claimants exercise their control through intercropping dynamics. Farmers may view trees and forests as “liabilities” that impede their livelihood and effective resource control, particularly when the state has criminalized their removal (Peluso 1992; Rocheleau and Ross 1995).

Contemporary agroforestry regularly includes the integration of valuable commodity trees into the farming system. By practicing agroforestry, smallholders can presumably maintain a stream of benefits from annual crops before the trees start bearing fruit. Schroeder and Suryanata (2004) examine two case studies of contemporary fruit-based agroforestry initiatives, in Indonesia and in the Gambia, both of which were hailed as bold steps toward environmental stabilization.
In both projects, the lucrative value of fruit commodities caused tree growers to exercise exclusionary practices and marginalize cultivators of the underlying crops. They illustrate that commoditization tends to extend the property rights of tree growers vis-à-vis competing resource users, such as cultivators of underlying crops, forest product collectors, and pastoralists.

From a political ecology perspective, the strengths of agroforestry practices do not lie exclusively in the ways they can enhance economic productivity or stabilize the environment; they also rest in the opportunities afforded by the sheltering of multiple claims and uses. In the context of neoliberal governance, agroforestry practices based on high-value commodity crops are attractive technologies that increase the incentive to plant trees. But they are also strategies of private accumulation that can shift the systems’ distributions of benefits and produce a range of ecological and social contradictions.

SEE ALSO: Agriculture; Indigenous technical knowledge; Livelihoods; Political ecology; Property and environment

References


Aid

Jamey Essex
University of Windsor, Canada

Aid refers to the transfer of goods, services, and other resources, usually from a relatively more developed state to a less developed one, on concessionary but often conditional terms, typically as a form of development or humanitarian assistance. While the contemporary system of global development and humanitarian aid has its roots in American efforts to support reconstruction and influence political and economic development first across war-ravaged Europe and later throughout the developing world following World War II, the interlocking set of institutions that today make up the global aid system includes much more than official state institutions and government-to-government flows of money and commodities. Private charitable foundations, intergovernmental bodies, and nongovernmental organizations (NGOs) are all deeply embedded in the geography of aid, directing how development programs and humanitarian and emergency interventions operate, devising policies and sharing best practices, and working with and sometimes in lieu of governments in new forms of public–private partnership.

While aid also can include various forms of military aid, institutionally most states manage military assistance separately from development and humanitarian assistance for political reasons. The more common use of the term “aid” usually excludes overt forms of military assistance. Practices and programs of aid are nonetheless closely tied to state strategies of national security and aid has long formed an important instrument of statecraft and foreign policy. In advancing sometimes complementary and sometimes contradictory goals focused on both modernization and geopolitical strategy, donors and recipients have used aid to alleviate short-term crises and to foster longer-term structural (and highly uneven) development of national political and economic systems. More generally, flows and practices of aid have strongly shaped the relationship between the Global North and South, promoting particular (usually but not always capitalist) forms of development and generally advancing the interests of donor states while often reproducing highly unequal terms of trade and power imbalances. A geographical understanding of the contemporary global aid system must first take into account how important national and international institutions developed in the wake of World War II and through the Cold War rivalry between the US and Soviet Union, and what role aid was meant to play in the development of the “third world.” The global aid landscape has changed considerably with the Cold War’s end, the rise of neoliberalism since the 1980s, and the proliferation of official and private aid actors. This has resulted in a complex geography of aid that today includes a greater emphasis on South–South connections and flows, the opening of aid provision and governance to new nonstate actors, and the inclusion of a wide array of social and environmental objectives in aid programs and projects. On the other hand, critics from across the political spectrum have begun to more directly interrogate aid’s purpose and effectiveness, calling into question widely held assumptions about the necessity of aid for development and the benefits for recipients.
The postwar system of aid, mutual security, and development

Following the end of World War II and the Cold War’s onset, the United States developed a number of programs to channel financial, military, and other resources to numerous countries in Europe and the rapidly decolonizing “third world” to rebuild shattered economies, to modernize those newly independent states seen as underdeveloped, and to shape political and social relations in an effort to combat the threat of communism. While this was most directly coordinated through the Marshall Plan, which injected $13 billion into the economies of American allies in Europe and Turkey from 1948 to 1951, several discrete, short-term, highly targeted programs of aid existed between the war’s end and the 1961 creation of a more permanent institutional home for foreign assistance in the US Agency for International Development (USAID). Many other industrialized states followed suit in the decades following World War II, moving from short-term aid programs housed under the larger umbrella of trade and economic cooperation (or even receiving aid themselves as part of postwar reconstruction programs) to the creation of official state agencies focused on development and humanitarian assistance to developing countries in the third world. In addition to the expansion and codification of aid programs as part of national states’ foreign, trade, and economic policies, aid became by the 1960s a major part of coordinated multilateral and intergovernmental efforts, especially through the United Nations and the Organisation for Economic Co-operation and Development (OECD), to restructure the international state system and world economy. Within the OECD, for example, Western economic powers worked to coordinate development and aid policy through the Development Assistance Committee (DAC), which remains an important forum and mechanism for shaping global aid policy today.

As concerted programs of aid became institutionalized amid the heightened tensions of the Cold War, it became impossible to easily and neatly separate the humanitarian, economic, and political foundations and goals of aid, even as these often resulted in competing mandates for aid organizations and programs. Postwar development thinking in the Western context often evolved directly from modes of thought and practice closely associated with the establishment and preservation of colonial empires in the nineteenth century and which had initially framed the “tropical” regions of the world through racially and environmentally determinist arguments and ideologies. After World War II, with rapid and often violent decolonization, such deterministic approaches faded “as ‘development’ and ‘environment’ began to inhabit the discourse of ‘the Tropics’, defining arenas of the discussable for such zones” (Power 2003, 47). These discourses conditioned the ways in which Western powers envisioned the purposes, goals, and limits of aid and shaped the resulting institutions and practices that programmed and delivered aid across the newly emergent third world.

Modernization theory codified much mainstream development thinking at this time, forming the intellectual backbone of Western aid programs and emphasizing that aid should provide the economic stimulus to propel transformative societal development in newly independent and impoverished states. Practitioners and policymakers in Western donor states understood aid as a pivotal component of a wider humanitarian project that proceeded through economic and geopolitical mechanisms, providing necessary capital, expertise, and support to states threatened by poverty, underdevelopment, and communist infiltration. In the case of aid’s historical role supporting both development and
geopolitical goals, it served (and continues to serve) to help diagnose and categorize states and peoples along axes of political allegiance, economic progress, and deservingness. In this way, postwar aid and development programs, institutions, and theories demonstrate how aid, rather than simply and benevolently sending needed money and goods to the world’s poorest from its richest, actually helped constitute the Cold War balance of power by promoting allies and denying enemies, identifying spaces of underdevelopment, disorder, and intervention, and creating new frontiers of capitalist market expansion.

While this was particularly true of American development assistance as organized and delivered through USAID, other Western donors’ aid programs often took somewhat different forms and foci based on historical relations with former colonies, the need to target scarce resources more efficiently, or unique foreign policy objectives and goals. In addition, the multilateral provision of aid, especially through the many agencies of the United Nations system, allowed for a wider range of political, economic, and social criteria to enter into aid strategies and objectives than those associated with American or Western geopolitical interests alone. For its part, the Soviet Union also provided large amounts of foreign aid to numerous developing countries. American policymakers even cited sizable Soviet aid programs in the late 1950s as a threat to US interests in the third world and an impetus to expand and rationalize US aid programs. Like American and other Western donors’ aid, Soviet assistance was tied to political and economic conditions that served donor interests while also articulating ideals of international and humanitarian cooperation that were often trumped by strategic considerations and objectives associated with the Cold War and domestic politics.

Despite the way it supported the superpowers’ Cold War ambitions, foreign aid nevertheless constituted an important part of development progress across many parts of the third world in the first three decades after World War II. The number of people living in absolute poverty fell, per capita incomes rose, food security improved, and economic growth and diversification occurred in many postcolonial states. To what extent this was directly the product of foreign aid remains contested, and contemporary experts and critics still hotly debate aid’s long-term track record as an effective tool for economic, political, and social improvement in poor countries and for poor people. What is clear is that much aid was (and still is, as discussed below) donor-centered, that is, programmed, delivered, and assessed primarily on whether it met the needs of donors rather than its intended beneficiaries. In this sense, the concerted, large-scale, permanent postwar aid programs and agencies helped define and assert Global North–South imbalances, tying postcolonial states into forms of economic dependency and political subordination that often looked decidedly like colonial and imperial relations in a new guise. Yet foreign aid also became a vital part of national development strategies and government revenue in many third world states, and actively shaped domestic social relations within these states. This forged emergent political and economic structures in those states, tying and reincorporating them into circuits of global and especially American capital while supporting geopolitical allies and, often, heavy-handed military governments. Indeed, as Kodras (1993) and Nijman (1995) have shown in their detailed examinations of the geographic flows of American aid dollars from the 1950s to the Cold War’s end, US aid in particular was largely concentrated in its allies, many of which received aid in exchange for specific security-related reasons.
(e.g., basing rights), to stabilize the domestic political or economic situation, or to support the procurement of US goods and services (known as “tied aid”). Aid also had profound impacts on the cultural and spatial organization of daily life, encouraging urbanization and industrialization while uprooting and delegitimizing traditional ways of life.

By the 1970s, however, global economic crisis and critiques of aid and development programs based on modernization theory and the geopolitical objectives of the Cold War superpowers started to produce significant shifts in the ideological underpinnings and practical management of foreign aid for Western donor governments and, increasingly, the international financial institutions (IFIs) of the Bretton Woods system. While these developments did not reach their full expression until after the Cold War ended in the early 1990s, they nonetheless pointed to the increasing role of neoliberal ideals and approaches in foreign aid, and in mainstream development thinking more broadly, and of the use of aid to more fully and directly support economic rather than geopolitical objectives. Dependency theorists in Latin America produced a sophisticated critique of modernist development and foreign aid emphasizing that the third world’s underdevelopment and peripheral status in the world economy was a direct product of exploitation by the core developed Western states, supported by transnational relations between capitalist elites in both sets of countries. This meant increasingly strong criticisms of Western aid from within the third world, accompanied by stringent domestic attacks on aid in many donor states, especially the United States. Domestic budget hawks in the United States had long criticized aid as an expensive handout, and the economic and foreign policy context of the late 1960s and early 1970s – economic slowdown, increasing domestic social program costs, stalemate in the Vietnam conflict, and American insistence that its European allies pick up more of the costs of global aid – also contributed to the push to put aid programs on a new ideological and programmatic footing.

Accordingly, official development assistance (ODA) agencies and increasingly the IFIs turned toward a concern for targeting aid at the poorest peoples and the provision of basic needs in the 1970s. As Lawson (2007, 101) explains, mainstream development practice came to emphasize a “weak” form of the basic needs approach in which aid helped to incorporate the poorest into development processes by encouraging their participation in development projects funded by aid dollars. This approach promised to make aid, and development generally, more responsive to the provision and improvement of life conditions for those traditionally excluded from development (especially including women) by targeting aid more directly at necessary infrastructure and social services such as education and health. Aid was thus to become less a blunt instrument of foreign policy and more a catalyst of social transformation and improvement. This shift also, however, strengthened aid programs’ and institutions’ technical capacities and foci at the expense of full inclusion for the poor in the political decision-making that shaped aid’s purpose and operation. In other words, the widespread shift in aid donors’ approach toward a weak basic needs emphasis tended to make development aid a technical exercise in including the poor in aid project delivery and completion without tackling head-on deep structural inequalities that persisted within developing countries and globally between states.

Though even a weak basic needs approach promised to alter significantly the aid system’s objectives and impacts in favor of the poorest people, further shifts in the ideological and technical focus of aid came quickly with the
AID

rise of neoliberalism in many donor states and
the IFIs. Looming debt crises across the devel-
oping world and the aggressive pursuit of trade
liberalization and global market integration as
a preferred development strategy put aid on
a different footing in the 1980s and into the
post-Cold War period. The neoliberalization
of aid and development meant that aid donors
and recipients now were to emphasize aid’s
ability to foster market development and global
integration and a suite of preferred neoliberal
reforms within developing states, including
privatization of state and parastatal enterprises,
deregulation of significant sectors of national
economies, and trade liberalization. Aid funds
again became a kind of leveraging tool for
imposing the objectives of powerful states onto
weaker counterparts in the developing world,
but now through the discipline of global market
forces and debt. Although this had always been
the case to some degree, the rapid shift toward
neoliberal approaches in aid and development
institutions, most importantly USAID, the IMF,
and the World Bank, emphasized much more
strongly than ever before the importance of
global economic interdependency and aid’s role
in the pursuit of market liberalization.

Aid retained geopolitical purposes in the
context of particular donors’ foreign policy
strategies as well, especially in the United
States. Increasingly, and especially after the Cold
War’s sudden end, geopolitical goals became
inextricably intertwined with neoliberal eco-
nomic reforms pushed through donor states’
bilateral aid and IFI-imposed structural adjust-
ment programs. Aid became a tangled mess of
complementary and competing projects and
programs aiming to achieve multiple objectives,
though all tended toward the transformative
neoliberalization of national economies and
societies in the Global South. This also opened
the door more widely to new actors, especially
NGOs and private charitable foundations, in
what was fast becoming a much more complex
and multiscalar international aid system. While
nonstate actors had long been important in the
programming and delivery of aid, the post-Cold
War environment allowed new geographies of
aid to flourish, involving multiple kinds of actors
working at different scales, creating new spaces
of aid policy formulation, and both challenging
and reproducing sociospatial inequalities tied to
intensive forms of neoliberal globalization and
the lingering effects of colonial and Cold War
systems. In sum, the Cold War’s end helped
create the conditions for consolidating a new
international aid regime predicated largely on
neoliberal ideals and policies, and restructured
existing aid institutions contending with new
mandates, new partners, and new contexts.

Aid in the post–Cold War era

Following the Cold War’s end, Western states
and international and intergovernmental orga-
nizations with large bilateral and multilateral
aid programs sought ways to restructure aid by
giving it new purpose and institutional footing
and improving its humanitarian and develop-
ment impacts. This meant finding ways to shed
aid’s geopolitical instrumentality and make it
more compatible with neoliberal globalization,
especially in terms of the relationship with
trade. Serious questions about aid’s ability to
deliver on its humanitarian and development
promises remain, however, and critics note that
aid remains tightly bound to donors’ national
security considerations and geopolitical maneu-
vering, and that aid’s record of supporting
economic growth, political democratization,
and social service provision remains spotty at
best, especially when coupled with (or for
many, compared to) neoliberal reforms. In

5
absolute terms, aid funding channeled through ODA programs and institutions located in the OECD-DAC countries dropped through the 1990s before increasing sharply around the turn of the millennium, in part due to massive new amounts of American aid flowing to Afghanistan and Iraq as part of the “global war on terror” following the terrorist attacks of September 11, 2001. Commitments to increase ODA funding made through the UN’s Millennium Development Goal 8 (which called on states to “develop a global partnership for development” to tackle problems of underdevelopment, debt, and an unequal international economic system) and G8 and G20 meetings following the 2007–2008 financial crisis, have likewise fostered renewed attention to the amount, type, and effectiveness of aid from OECD-DAC donors.

Despite this, only a handful of OECD donor states (primarily the Scandinavian and Benelux states) have ever reached the UN’s suggested target of 0.7% of GNI as a threshold for aid funding, budget concerns in most countries have made ODA funding a target for cuts amid lingering effects of global financial and economic crisis, and aid remains highly unpopular with the public in many traditional donor states. Even so, OECD statistics indicate that DAC donors’ assistance to developing countries increased from just over US$50.1 billion in 2003 to almost US$88.6 billion in 2012 (accounting for almost two thirds of global ODA), reaching an all-time high of US$94.4 billion in 2011. The impacts of the global economic and financial crisis, however, make the short to medium term prospect of increased aid funding from most OECD donors unlikely. In turn, many official donors have sought ways to streamline their operations and improve the effectiveness of their aid dollars, often by seeking more intensive partnerships with NGOs, the private charitable and for-profit sectors, and multilateral and intergovernmental organizations. Many donor states have also moved to rationalize internal operations within their own aid systems and programs, sometimes adopting a “whole of government” approach to aid and making stronger efforts to coordinate various state institutions working on and with aid accounts. In some instances, particularly in the United States, this has evolved alongside what is now commonly called the “3Ds approach” that makes development a pivotal component of national security strategies with defense and diplomacy. In this context, recent donor concerns with aid reform, effectiveness, and funding have sought to reduce long-standing problems with overlap, redundancy, and waste while also reproducing problems associated with donor-defined and -centered objectives for aid programs and their role as geopolitical instruments.

By contrast, two important trends in the global aid system indicate the extent to which the contemporary geographies of aid are changing, with the OECD-DAC states’ bilateral programs and the multilateral channels and programs they tend to dominate now representing just some of the many important mechanisms and forums for shaping a global aid agenda. The first trend concerns the entrance of new donor states into the aid system, especially major emerging economies from the Global South, and the concomitant development of new mechanisms of donor coordination. The second is the multiplication of channels and forms of privately funded development aid, which must be seen in light of two intertwined processes, one being the emphasis on trade and foreign direct investment across three-plus decades of neoliberalization and the other being the emergence of a new form of “philanthrocapitalism” through which private capital has sought ways to make philanthropic activities conducive to profit seeking investment.
Looking first at the emergence of new donor states and forms of donor coordination, Mawd-sley (2012) cautions that it is important that many non-DAC and “emerging donors” (such as China, India, Russia, and Saudi Arabia) are in fact not new to the global aid system as donors at all. Recent changes in aid governance complicate most attempts to categorize donors and throw into disarray older categories and assumptions about how aid works globally. There is today a greater emphasis on partnership, transparency, accountability, and coordination, including forms of South–South cooperation, than ever before and a number of principles regarding aid governance have been articulated through a series of international meetings and declarations. These include the UN Millennium Declaration (2000) and the Monterrey Consensus (2002), which established the MDGs and asserted new commitments to financing development efforts in support of these goals, as well as the Paris Declaration (2005), the Accra Agenda for Action (2008), and the Busan Agreement (2011), which have steadily plotted a set of international principles for assessing and ensuring aid effectiveness and expanding the governance of aid to include recipient states and emerging donors in more robust ways. Such high-level efforts have created a new, international space of aid policy formation and coordination that is more inclusive and open than global and DAC-centered aid systems and mechanisms had been previously. On the other hand, this policy space remains highly managerial in scope and approach and far removed from on-the-ground delivery of aid. Critics continue to pose political challenges to aid that efforts at better donor coordination and articulation of globally accepted principles of aid governance have yet to adequately answer. Emerging donors, with Chinese aid to and investment in sub-Saharan Africa often serving as a prime example, seem at times to replicate the donor-centered programs of their OECD peers even as they demonstrate alternative models and practices of economic cooperation and governance not tied to Western power and conditionalities.

More deeply, many critics also strongly question the continued necessity of aid and suggest that where aid creates bonds of dependency and undermines social, political, and economic innovation, full integration into the global capitalist economy remains the more just and effective path to development with better outcomes for poverty, economic growth, and democratization than any foreign aid program could hope to produce. In this respect, the role of private channels of aid, particularly from charitable and philanthropic foundations associated with some of the world’s largest and wealthiest transnational corporations and individuals and the longstanding quest to make aid and trade complementary, form a second key trend in contemporary changes in the global aid system. Fridell and Konings (2013) argue that such forms of “philanthrocapitalism” are an optimistic response to, and outcome of, neoliberal economics and the production and exacerbation of inequality widely associated with trade liberalization, structural adjustment, and poor terms of trade for developing states. Private flows of aid, particularly those provided by foundations managed by “icons” of capitalism at the pinnacle of the world market (such as Bill Gates or Jeffrey Sachs), promise to resolve through voluntary charitable action seemingly intractable problems that decades of official bilateral and multilateral aid have failed to correct, such as poor economic growth, inequitable access to health and education, environmental degradation, and even the provision of basic infrastructure. In doing so, aid will supposedly create more effective conditions for integrating people and places into regional, national, and global economies,
and produce new opportunities for product development and market creation, though the neoliberal foundations of economic and political relations remain unchanged and unchallenged. Contemporary philanthrocapitalism thus builds on the Washington Consensus policies that helped remake development aid by encouraging market liberalization and restructuring developing state forms and functions over the last three decades. This recasts deep structural problems of underdevelopment and marginalization as market failures that can be corrected through highly targeted aid interventions and appropriate private investment unencumbered by the political considerations and objectives of official development assistance. Neoliberal assumptions about the necessity of a global market orientation for proper development are thus coupled with a kind of celebrity culture in striking new ways through philanthrocapitalism and private aid, drawing attention to a narrow range of development problems and solutions while ignoring others and reproducing some of the major faults critics have identified in official aid channels and institutions.

The future of aid

The future of the global aid system and its role in promoting development, justice, and equity across a range of social, economic, political, and environmental considerations hinges on a series of fundamental questions about the conduct of aid in a radically interdependent world. What role should aid play in a global system defined by stark social, political, and economic inequalities, imperial relations of governance and subordination, and potentially disastrous processes of environmental degradation and climate change? Should aid work to foment liberalized trade and support strategic national security objectives? Should it operate primarily to soothe Western souls and placate the poor without fundamentally challenging existing social, political, and economic relations? Can the narrow and highly technical and managerial forms of aid currently favored by many donors, both private and official, actually promote and achieve full social inclusion and empowerment of the poor? Should aid solely be employed in emergency situations, such as responding to natural disasters, and leave long-term development to market forces and privatized forms of economic connection and political governance? What alternative traditions and practices might exist or evolve that can inform our ways of thinking about and delivering aid, and our relationship to distant others? Overton, Murray, and McGregor (2013, 122) provide one of the few recent comprehensive research agendas for critically investigating the rapidly shifting geographies of the global aid system, arguing that “geographies of aid should have an explicit political goal that provides progressive space for those who suffer due to the inherent unevenness” of this system. They suggest that core geographic concepts of space, place, scale, and networks should shape the way geographers as well as aid practitioners and critics answer the kinds of difficult questions about aid outlined above to highlight not only how geography shapes aid, but also how aid creates new geographies. The contemporary global aid system, uneven and unjust as it often is, thus offers a range of potential access points for critical geographic inquiry and intervention, which can potentially create spaces and mechanisms to better fulfill aid’s emancipatory promise.

SEE ALSO: Civil society; Dependency theory; Development; Globalization; Inequality; Modernization theory; Neoliberalism; Poverty; States and development
References


Further reading


Air transport

John T. Bowen Jr
Central Washington University, USA

At any one moment in 2012, approximately 1 million people were airborne on commercial flights around the world (estimate based on ICAO 2014 and methodology described in Bowen 2010, 2, note 2). Their movements and the torrent of packages, pallets, and parcels sent daily by air have brought distant markets, resources, people, places, and hazards near. At the same time, air transport is an important factor in global climate change, and likely to become more so in the future. And yet, even as the world’s airlines are catalysts for change at the global scale, aviation’s greatest impacts, for good and for ill, are felt locally: in hubs, in spokes, and in the places in between. Across several scales, the development of aviation has affected, sometimes profoundly, the world’s human and physical geography.

Both passenger and cargo traffic have grown robustly. Over the period 1950–2012, for instance, air passenger-kilometers (one passenger flown one kilometer) and air cargo tonne-kilometers grew at annualized rates of 8.6 and 9.0%, respectively (ICAO 2014; Bowen 2010); both these rates are more than twice that attained by gross world product during the same period. The pace of its expansion has propelled aviation deep into the spaces of everyday life across much of the planet. In the United States, as in other affluent countries and many middle-income ones, air travel is the dominant mode of passenger travel beyond about 1250 km (Bureau of Transport Statistics 2006), and a growing share of even substantially shorter trips involve flying – especially on low-cost carriers. Meanwhile, aviation’s share of global freight traffic is tiny in tonnage terms but substantial – about one-third when measured in value terms (Airbus SAS 2015). The importance of aviation, combined with this mode’s inherent spatial dynamism, has made air transportation the subject of intense and sustained scrutiny by geographers (Vowles 2006).

Air transport technology and the diffusion of aeromobility

By 2012, the world’s passenger airlines operated nearly 33 000 nonstop scheduled routes (OAG 2012), a skein of links that enveloped almost every part of the globe except Antarctica and the Southern Ocean. Across this network moved 3.0 billion people (ICAO 2014). These statistics attest to the remarkable advances made since the beginning of commercial aviation.

The diffusion of aeromobility (Lassen 2006), or broadly accessible airborne mobility, began in advanced economies, but has more recently spread to developing countries. In the United States, aviation first gained economic significance in the carriage of airmail, beginning with the world’s earliest such service operating between Washington, DC, and New York City in May 1918 (Bilstein 1983). By contrast, in Europe, the other early center of commercial aviation, fragmented national markets precluded the development of airmail, and nascent carriers focused on passenger traffic instead. On both sides of the Atlantic, early airlines were either state-owned...
or state-subsidized and protected, reflecting government eagerness to realize the commercial and strategic promise of the new technology.

Rapid advances in that technology during the 1920s and 1930s yielded remarkable improvements in the speed, capacity, and safety of commercial aircraft. The most influential of the early airliners was the Douglas DC-3. The 21-seat twin-propeller plane incorporated important innovations (e.g., flaps and slotted wings), and became the first commercial aircraft to regularly fly nonstop between New York City and Chicago. In fact, adequate range to link the two business centers had been an important design criterion mandated by the launch customer, American Airlines (Bowen 2010).

As air transportation grew in scale and importance before World War II, its networks continued to resemble those of the railroads and steamship lines they were supplanting. Networks were strongly linear in form. In 1940, a transcontinental journey from New York City to Los Angeles, for instance, typically required six stops en route (Wynne 2011). Intercontinental travel was similarly arduous, as exemplified by the 32 stops and 10 days required for air travel between London and Sydney in 1939 (Wynne 2011). Air travel was expensive, too; the round-trip airfare for New York to Los Angeles was $300 or about 50% of US per capita income at the time. In 2012, the average fare on the route was about $700, but that sum equated to just 2% of US per capita income.

Even in the mid-twentieth century, however, air transportation, though costly, had such a speed advantage over alternative modes, and technological change in aviation was so fast, that the airlines continued to capture traffic from surface modes. The first sustained nonstop transcontinental commercial flights in the United States began in 1953, and by 1956, nonstop services had begun between New York City and London (Bowen 2010). These flights involved turboprop-powered aircraft of unprecedented speed, size, safety, and – of course – range. Not coincidentally, the mid-1950s also marked a watershed as transatlantic crossings by plane first eclipsed and then overwhelmed crossings by sea, and in the same decade air travel edged out rail in the number of transcontinental trips within the United States (Hugill 1993).

The advent of jet airliners in the 1950s and wide-body jets in the 1970s carried air transportation to still greater heights. The first commercial jet service was launched by British Overseas Airways Corporation (BOAC) with a British-built jet, the De Havilland Comet. Though heralding the future, the sleek jet made its commercial debut in 1952 on a route that reflected Britain’s imperial past: from London to Johannesburg via Rome, Cairo, Khartoum, Entebbe, and Livingstone (Bowen 2010). Much larger, faster, and ultimately more commercially successful jets arrived at the end of the 1950s, beginning with the 132-seat Boeing 707. The big Boeing entered commercial service on a Pan American World Airways route linking New York and Paris, with a refueling stop in Gander, Newfoundland. The ensuing proliferation of services by the 707, the similar Douglas DC-8, and then a bevy of smaller jets produced by Boeing, Douglas, and manufacturers in Western Europe and the Soviet Union, defined the geography of the early Jet Age in the 1960s.

Jet travel produced striking time-space convergence: for instance, the air travel time on Pan Am between New York and London fell from just under 11 hours nonstop on a DC-7 in 1957 to 6 hours and 35 minutes on a 707 in 1963. And as dependence on jets spread, the new technology stretched the scale and increased the speed of everyday life, much as earlier aviation innovations had done. Firms were better able to spread their operations across
transcontinental and intercontinental markets. The semiconductor industry, for instance, took advantage of the spatial flexibility afforded by air transportation to spread its integrated circuit manufacturing system across the Pacific to Hong Kong and Singapore in the 1960s, and then other parts of Asia in the 1970s. Jet aviation was similarly catalytic in extending the “pleasure periphery” (Turner and Ash 1976), a phrase that refers to the set of places to which the affluent and increasingly the middle-class retreat for holidays. Hawaii, for instance, welcomed its first jet flight in 1959, the same year it became a US state; and both changes drew the islands closer to the mainland market (Bowen 2010). At a deeper level, jet travel spurred the imagination in the 1960s. From the Jetsons to the New York Jets to jet-inspired automobile designs, the new technology was invoked as a symbol of power, modernity, and confidence in the future.

The confidence of the Jet Age fueled the rapid development of still larger commercial aircraft. In the early 1970s, the jumbo 366-seat Boeing 747 and several other wide-body jets were introduced. Unfortunately, this addition of capacity occurred just as the Arab Oil Embargo and subsequent disruptions of global oil supplies triggered a severe economic contraction across much of the world. These circumstances challenged all of the aircraft manufacturers, but ultimately proved advantageous to the newest of the competitors: Airbus Industrie.

The European industrial consortium had been formed in the late 1960s to redress what French President Charles de Gaulle lamented as “America’s colonization of the skies” (quoted in Heppenheimer 1995, 203). Linking the resources of aerospace manufacturers in West Germany, France, the United Kingdom, and to a lesser extent Spain and the Netherlands, Airbus Industrie entered the airliner business with the 250-seat Airbus A300. Unlike the other new wide-body jets, the Airbus had just two engines, which gave it an inherent fuel efficiency advantage. The jet, whose development was heavily subsidized by the governments of the respective manufacturers, made its commercial debut in 1974 on the symbolically significant hop between Paris and London. From that beginning, Airbus has risen in just 40 years to become a well-matched rival for Boeing.

Indeed, the combination of industrial economies of scale, advanced technological human capital resources, and supportive governments has fostered substantial market share gains by Airbus. By the early twenty-first century, it was the only independent European manufacturer of aircraft with more than 100 seats. On the other side of the Atlantic, meanwhile, Boeing consolidated its position as the industry leader. By 2012, Boeing and Airbus accounted for 43% and 30%, respectively, of the 23,236 commercial jets in the world’s airline fleets (Air Transport World 2013). Their relentless competition with one another has encouraged a steady improvement in aircraft performance, which in turn has made air transportation more affordable and hastened its spread.

In 2012, Boeing and Airbus accounted for more than 80% of unfilled aircraft orders (airliners are commonly ordered several years before delivery), indicating that their dominance of the skies will continue. Many of the newly built airliners will be used in developing countries, which comprise a large and growing aviation market. For instance, Lion Air of Indonesia and AirAsia of Malaysia are among the largest customers for new narrow-body aircraft from Boeing and Airbus, respectively. The significance of these airlines as customers for new aircraft technology illustrates the diffusion of aeromobility.
Air transport, globalization, and the persistence of the short-haul linkages

By 2012, the longest regularly scheduled commercial flights extended more than 13,000 km nonstop, including, for instance, Sydney–Dallas–Fort Worth, Atlanta–Johannesburg, and Dubai–Los Angeles. These links are testament to the significance of globalization, a phenomenon for which relatively low-cost aviation has been a celebrated enabling factor. For instance, freshly cut flowers flown from Kenya to Europe and tuna caught off Atlantic Canada destined for sushi restaurants in Japan illustrate the power of aviation to lower the cost of distance and thereby enlarge markets.

The spatial scale of civic society and politics, too, is stretched by air accessibility. The Association of American Geographers (AAG), for example, has become an international organization; about a third of attendees at the AAG’s 2013 conference in Los Angeles came from outside the United States. The vast majority of those conference-goers, like their domestic counterparts, flew to Los Angeles. Politically, aviation has long been an important instrument for nation-building; that importance is most clearly expressed in so-called flag carriers. The political visibility of the air transportation system also helps to explain its prominence as a target for terrorist organizations: attacks on airports or airlines can be a means of projecting a regional conflict onto a global stage.

Aviation has also influenced population geography, particularly with respect to migration and the diffusion of disease. Air travel has permitted increased transnationalism, as migrants are better able to maintain a presence in their countries of origin (Bowen 2010). Meanwhile, aviation has accelerated the diffusion of some infectious diseases. One of the most striking examples was the 2002–2003 spread of severe acute respiratory syndrome (SARS) from Southern China to 25 countries and Taiwan. The epidemic’s geography was strongly shaped by patterns of aviation accessibility (Bowen 2010).

Yet while aviation has been important in enlarging the scale and increasing the pace of some features of everyday life, it is important to note that most air travel is short-haul. In 2012, the average distance traveled per air passenger around the world was approximately 1830 km, about the distance from Boston to Memphis or London to Helsinki. The busiest routes in the world tend to be short-haul and domestic (Table 1) illustrating the persistence of distance decay and of borders, even in an era of globalization. Only one of the 20 busiest routes in the world in 2012 was more than three hours long, and that linked the world’s two most dominant cities (London and New York). More characteristic of densely traveled routes is the hour-long sector between Rio de Janeiro and São Paulo.

Airlines and their networks

In 2012, there were approximately 725 airlines offering scheduled passenger services, up from 595 in 1998 (OAG 1998, 2012). Most airlines in both years were combination carriers, meaning they also offered cargo services — typically in the lower deck space aboard passenger airliners, though some combination carriers (e.g., Korean Air) also operated substantial fleets of freighter aircraft. Additionally, a relative handful of carriers, notably including FedEx Express and UPS Airlines, operated extensive scheduled freighter services but flew no passengers.

The small increase in the number of airlines conceals the tremendous flux in airline competition in the past several decades. For instance, fewer than half of the 595 airlines flying in 1998 were still flying in 2012. Among the largest
Table 1  The world’s top 20 city-pairs ranked by 2012 scheduled airline capacity.

<table>
<thead>
<tr>
<th>Rank (1998)</th>
<th>City-pair</th>
<th>Distance (km)</th>
<th>Type</th>
<th>Scheduled seats per week (000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (3)</td>
<td>Rio de Janeiro–São Paulo vv</td>
<td>363</td>
<td>D</td>
<td>243</td>
</tr>
<tr>
<td>2 (1)</td>
<td>Sapporo–Tokyo vv</td>
<td>816</td>
<td>D</td>
<td>241</td>
</tr>
<tr>
<td>3 (2)</td>
<td>Fukuoka–Tokyo vv</td>
<td>883</td>
<td>D</td>
<td>204</td>
</tr>
<tr>
<td>4 (27)</td>
<td>Cheju–Seoul vv</td>
<td>451</td>
<td>D</td>
<td>199</td>
</tr>
<tr>
<td>5 (4)</td>
<td>Melbourne–Sydney vv</td>
<td>707</td>
<td>D</td>
<td>199</td>
</tr>
<tr>
<td>6 (11)</td>
<td>Osaka–Tokyo vv</td>
<td>417</td>
<td>D</td>
<td>196</td>
</tr>
<tr>
<td>7 (49)</td>
<td>Beijing–Shanghai vv</td>
<td>1079</td>
<td>D</td>
<td>184</td>
</tr>
<tr>
<td>8 (8)</td>
<td>Hong Kong–Taipei vv</td>
<td>805</td>
<td>I</td>
<td>162</td>
</tr>
<tr>
<td>9 (83)</td>
<td>Delhi–Mumbai vv</td>
<td>1136</td>
<td>D</td>
<td>141</td>
</tr>
<tr>
<td>10 (20)</td>
<td>Cape Town–Johannesburg vv</td>
<td>1267</td>
<td>D</td>
<td>141</td>
</tr>
<tr>
<td>11 (46)</td>
<td>Naha (Okinawa)–Tokyo vv</td>
<td>1557</td>
<td>D</td>
<td>134</td>
</tr>
<tr>
<td>12 (5)</td>
<td>Chicago–New York vv</td>
<td>1167</td>
<td>D</td>
<td>128</td>
</tr>
<tr>
<td>13 (42)</td>
<td>Jakarta–Singapore vv</td>
<td>880</td>
<td>I</td>
<td>107</td>
</tr>
<tr>
<td>14 (21)</td>
<td>Brisbane–Sydney vv</td>
<td>750</td>
<td>D</td>
<td>107</td>
</tr>
<tr>
<td>15 (13)</td>
<td>London–New York vv</td>
<td>5544</td>
<td>I</td>
<td>105</td>
</tr>
<tr>
<td>16 (16)</td>
<td>Barcelona–Madrid vv</td>
<td>481</td>
<td>D</td>
<td>103</td>
</tr>
<tr>
<td>17 (312)</td>
<td>Hanoi–Ho Chi Minh City vv</td>
<td>1159</td>
<td>D</td>
<td>101</td>
</tr>
<tr>
<td>18 (38)</td>
<td>Kuala Lumpur–Singapore vv</td>
<td>299</td>
<td>I</td>
<td>100</td>
</tr>
<tr>
<td>19 (51)</td>
<td>Brasilia–São Paulo vv</td>
<td>856</td>
<td>D</td>
<td>100</td>
</tr>
<tr>
<td>20 (109)</td>
<td>Seoul–Tokyo vv</td>
<td>1218</td>
<td>I</td>
<td>98</td>
</tr>
</tbody>
</table>

Notes: vv, and vice versa; D, domestic; I, international.
Sources: Author’s analysis of OAG Max (OAG, 1998) and OAG Max (OAG, 2012).

Airlines in 1998 that had failed or been absorbed through merger by 2012 were Northwest Airlines and TWA in the United States, Swissair and Olympic Airways in Europe, Brazil’s former flag-carrier VARIG, and Ansett Australia. Conversely, hundreds of new entrants were formed during the same period, of which 490 survived to 2012; in other words, about two-thirds of the airlines flying in 2012 had been formed in the preceding 14 years.

Both the increased size of individual carriers and the flux in competitors are reflections of deregulation. Historically, the airline industry was closely protected by governments eager to promote aviation for the sake of economic development, political symbolism, and strategic significance. In most countries, one or a handful of state-owned airlines operated state-authorized domestic networks and charged state-approved fares and air cargo rates. Even where the airline industry was in largely private hands – as in the United States – its operations were strictly controlled. International aviation, likewise, was regulated heavily. Deregulation, which began in
earnest in the United States in the 1970s, has altered the relationship between governments and the airline industry, and the result has been freer competition and greater dynamism. The diffusion of deregulation has been uneven, but generally more developed markets are more thoroughly deregulated.

A corollary of deregulation has been privatization, and that trend too is more fully articulated in richer countries. Among the dwindling number of countries in 2012 whose domestic markets were dominated by a single, state-owned airline were: Argentina (Aerolíneas Argentinas), Botswana (Air Botswana), Ethiopia (Ethiopian Airlines), Jordan (Royal Jordanian Airlines), and Laos (Air Laos). Most large markets, especially affluent ones, are now characterized by vigorous competition among mainly privatized carriers.

Prominent among the new players in the industry are low-cost carriers (LCCs) or budget airlines. Characterized by high-frequency operations over mainly short- and medium-haul sectors, LCCs typically fly one or two aircraft types, have minimal in-flight services, and rely heavily on online marketing and ticket sales. These and other strategies have tended to give LCCs a sizeable cost advantage over incumbent, full-service airlines. As a result they have gained market share in almost every market where the regulatory environment is favorable. More fundamentally, LCCs have been instrumental in democratizing aeromobility. In 2012, the largest LCCs ranked by weekly scheduled capacity were Southwest Airlines (USA), Ryanair (Ireland), easyJet (UK), gol Airlines (Brazil), and Lion Air (Indonesia) (OAG 2012). Each of these airlines is the biggest or second-biggest in its domicile in terms of seat capacity.

Nevertheless, at a global scale the airline industry continues to be dominated by legacy full-service network carriers. Table 2 ranks the largest airline enterprises, only one of which is an LCC, by revenue. The network carriers have the advantages of age – including brand recognition and control of key airport takeoff and landing slots – and size – especially economies of scale and scope. In an era of increased global traffic, the international reach of the largest airlines is an important competitive edge. As evident from Table 2, the leading carriers are now present in almost every major market – though Southwest Airlines is a clear exception.

Most of the airline enterprises listed in Table 2 comprise multiple carriers. The Lufthansa Group, for instance, includes Lufthansa, Lufthansa Cargo, Swiss, Austrian, and several other carriers that are subsidiaries of these four. The grouping of airlines reflects the advantages of size and the synergies that can come by reducing duplicate services among airlines with nearby hubs. Yet this arrangement also permits the separate airlines to retain distinct identities, which is important in an industry still redolent with nationalism and cultural preferences (e.g., the kinds of foods served in-flight).

A similar balancing act between the benefits of size and the benefits of independence has driven the growth of the airline alliances. The three largest passenger airline alliances are the Star Alliance (26 airlines in mid-2014), SkyTeam (20), and oneworld (14). Each incorporates airlines from around the world, including several major American, European, and Asian carriers each. Through schedule-meshing at major hubs (e.g., Chicago, Frankfurt, etc., for Star Alliance) and joint marketing, the alliances attempt to create an integrated service. Unlike the enterprises listed in Table 2, however, the alliances are not based on ownership, which permits them to be much more flexible in gaining and losing members.

Alliances give their members a degree of spatial coverage no single airline can match. In 2012, for instance, the Star Alliance network comprised nonstop links connecting nearly 5600
Table 2  The world’s top 20 largest airline enterprises, 2012.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Enterprise</th>
<th>Domicile</th>
<th>Revenue (billion USD)</th>
<th>Network destinations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Asia</td>
<td>Europe</td>
</tr>
<tr>
<td>1</td>
<td>Lufthansa Group</td>
<td>Germany</td>
<td>37.2</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>United Continental Holdings</td>
<td>USA</td>
<td>37.1</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>Delta Air Lines</td>
<td>USA</td>
<td>35.1</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>Air France-KLM</td>
<td>France/Netherlands</td>
<td>31.5</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>FedEx Express</td>
<td>USA</td>
<td>24.8</td>
<td>Network not publicly available</td>
</tr>
<tr>
<td>6</td>
<td>AMR Group&lt;sup&gt;a&lt;/sup&gt;</td>
<td>USA</td>
<td>24.6</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>IAG&lt;sup&gt;b&lt;/sup&gt;</td>
<td>UK/Spain</td>
<td>24.0</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>Emirates Airline</td>
<td>UAE</td>
<td>17.2</td>
<td>29</td>
</tr>
<tr>
<td>9</td>
<td>Southwest Airlines</td>
<td>USA</td>
<td>15.8</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>Qantas Group</td>
<td>Australia</td>
<td>15.7</td>
<td>13</td>
</tr>
<tr>
<td>11</td>
<td>Air China</td>
<td>China</td>
<td>15.2</td>
<td>123</td>
</tr>
<tr>
<td>12</td>
<td>China Southern</td>
<td>China</td>
<td>14.9</td>
<td>146</td>
</tr>
<tr>
<td>13</td>
<td>ANA Group</td>
<td>Japan</td>
<td>14.2</td>
<td>66</td>
</tr>
<tr>
<td>14</td>
<td>US Airways Group</td>
<td>USA</td>
<td>13.0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>China Eastern</td>
<td>China</td>
<td>12.7</td>
<td>168</td>
</tr>
<tr>
<td>16</td>
<td>LATAM Group&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Chile/Brazil</td>
<td>12.6</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>JAL Group</td>
<td>Japan</td>
<td>11.8</td>
<td>53</td>
</tr>
<tr>
<td>18</td>
<td>Cathay Pacific</td>
<td>Hong Kong</td>
<td>11.4</td>
<td>53</td>
</tr>
<tr>
<td>19</td>
<td>SIA Group</td>
<td>Singapore</td>
<td>10.6</td>
<td>62</td>
</tr>
<tr>
<td>20</td>
<td>Air Canada</td>
<td>Canada</td>
<td>7.2</td>
<td>5</td>
</tr>
</tbody>
</table>

<sup>a</sup>AMR Group is the parent corporation for American Airlines. The data on which the table is based were produced before American Airlines’ merger with US Airways.

<sup>b</sup>IAG is the parent corporation for two major airlines, British Airways and Iberia, as well as the Spanish low-cost carrier Vueling.

<sup>c</sup>LATAM Group is the parent corporation for LAN of Chile and TAM of Brazil.

Sources: Air Transport World (2013); author’s analysis of OAG (2012).
city-pairs – or about one of every six city-pairs linked by any airline, more than five times the number of United, its largest single airline. Even more important, United – as large as it was – had a very limited presence outside North America, but through alliance partners such as Ethiopian Airlines and Singapore Airlines, the reach of the American giant extended well beyond the gateway hubs of distant regions.

Nevertheless, hubs remain vital in mediating contemporary air traffic patterns. Despite the many technological, regulatory, economic, and organizational changes that have transformed the airline industry, its operations are strongly concentrated in hubs. In the case of Ethiopian Airlines, for instance, 63% of all of its flights arrived at or departed from Addis Ababa Bole International Airport in 2012 (OAG 2012). United Airlines was less singularly focused, but 74% of its flights had as an endpoint Chicago, Washington, San Francisco, or Los Angeles (OAG 2012).

Conversely, low-cost carriers often emphasize point-to-point services rather than connections via hubs, as indicated by graph theoretic measures such as the gamma index (Taaffe, Gauthier, and O’Kelly 1996). Based on the number of vertices (i.e., cities served) and edges (i.e., nonstop routes), the gamma index compares the actual extent of network coverage to the theoretical ideal of a pure point-to-point network, which in an airline context would mean a nonstop route from each city to every other city in the network. Such an ideal would yield a gamma index of 1.00, but in reality Southwest Airlines has an index of 0.37. In other words, the degree to which its network is focused in particular cities differs markedly from United (0.04) or Ethiopian (0.06), but is even more distinct from the point-to-point ideal.

For the global air passenger network (all carriers combined) in 2012, the gamma index was 0.01, a remarkably low value that attests to continued concentration in hubs, but also several constraints on network development. These include technical limitations on aircraft range, uneven economic development, and the many remaining political and regulatory barriers that continue to thwart network expansion. As an example of the latter, in 2012 Ethiopian Airlines had only six nonstop cross-border routes to neighboring countries, despite the fact that its neighbors had a combined population of more than 90 million. These constraints are almost certain to ease in aviation’s second century. The upward potential for further air traffic growth is vast.

**Containing the future costs of air transportation**

For a growing share of the world’s population, aviation confers a remarkable degree of spatial freedom, but air transportation also carries formidable costs. The relative importance of these costs varies depending on the scale of one’s analysis. At the local scale, people who live near major hubs are most frequently concerned about aircraft noise. Those who live away from airports are more likely to be concerned with greenhouse gas (GHG) emissions and the role of aviation in global climate change. Containing these costs will be a daunting challenge.

With respect to both issues, great strides have been made in improving the performance of individual aircraft. For instance, the Boeing 787 is substantially quieter and far more fuel-efficient (i.e., producing fewer GHGs) than similar-sized aircraft of a generation ago. Yet those gains are partially offset by traffic growth (Bowen 2013). Indeed, in today’s extremely competitive market, airlines are likely to pass on efficiency improvements in the form of lower fares and air freight rates.
Some have argued that biofuels or radically different aircraft designs (e.g., blended-wing bodies) might permit aviation to grow rapidly while reducing this sector’s environmental costs. Yet both of these alternatives face great obstacles (Bowen 2013). A biofuel that can meet a significant fraction of aviation demand at a competitive cost, that is carbon neutral, and that does not detract from the global food supply is a prospect somewhere over the horizon. And technological lock-in means that there is tremendous inertia behind the current designs in aviation, inertia which stems in part from the 40-year lifetime of the typical commercial aircraft.

Europe’s experience offers shorter-term solutions. First, its emissions trading scheme (ETS), applied on some flights from 2012, compels airlines to purchase carbon credits equivalent to emissions. The European ETS has been highly controversial, but such a market-based approach – especially if applied at the global scale – could be useful in compelling airlines to adopt technology and practices that will contain the growth of emissions (Bowen 2013). Second, some major European airports have chosen to restrict growth in ways that ultimately limit externalities. The UK government, for instance, has so far refused to authorize a third runway at congested Heathrow Airport, and a ban on night operations has been imposed at Frankfurt Airport.

Yet in low- and middle-income markets where the majority of people have yet to fly, it seems doubtful that such constraints will be freely adopted. As recently as 2005, only about 1% of Indians had ever flown (Bowen 2010); by comparison, more than 80% of Americans have done so (Transportation of the United States 2013). The developing world is now catching up, as evidenced by the changes in the rankings of the busiest routes (Table 1). And about half of the world’s top 20 airline enterprises are based in emerging markets (Table 2); in the late 1990s, only one was.

More rapid economic and population growth in markets such as China and India, combined with the diffusion of deregulation, suggest that the upward trajectory of aviation – with its many benefits and costs – will continue.

SEE ALSO: Airports; Economic geography; Global environmental change: human dimensions; Network analysis; Transport geography

References

AIR TRANSPORT


OAG. 1998. OAG Max database for April 1998. (The database is a compilation of airline schedules for virtually every airline in the world; on CD-ROM.)

OAG. 2012. OAG Max database for February 2012. (The database is a compilation of airline schedules for virtually every airline in the world; on CD-ROM.)


The airport is among the most important and controversial forms of land use. One hundred years after the world’s first commercial flight, which carried a single paying passenger across Tampa Bay in 1914, about 6000 people per minute were ascending from runways somewhere across the planet. And that number continues to spiral upward with the diffusion of “aeromobility” (Adéy, Budd, and Hubbard 2007) in a growing number of countries. With rising volumes of air traffic, the airport’s cultural, economic, political, and environmental significance will continue to climb too.

The earliest centers of aviation activity were airfields and seaplane bases. The Wright brothers’ perfection of the Flyer, for instance, occurred principally at Huffman Prairie, a rough field near Dayton, Ohio. Fixed infrastructure, at sites such as College Park near Washington, DC, and Croydon near London, accompanied the advent of regular services after World War I. Yet, even as aircraft grew in size and range, a substantial share of intercontinental flights arrived at and departed from seaplane bases which permitted long take-off runs for underpowered passenger planes and minimized infrastructure costs. In the 1930s, for example, Pan Am’s famous Clippers departed from seaplane bases such as Port Washington near New York City.

By the early post-World War II period, however, aircraft had so improved in performance and the nascent airline industry had gained such stature that airports with paved runways accommodating diverse landplanes became the principal gateways to the world’s airways. As airports grew in number and importance, they expanded in their footprints too. In 1948, for instance, New York International Airport (later John F. Kennedy International Airport) opened with seven runways as long as 9500 ft (2900 m); a generation earlier, the first airport to serve the city had begun operations in Newark with a single 1600 ft (500 m) runway.

Increased dimensions have been among the factors propelling airports away from urban cores. For cities served by two major airports, the newer one (e.g., Paris Charles de Gaulle, Narita International near Tokyo, São Paulo’s Guarulhos International) is almost invariably farther from the core than the older one (Paris Orly, Tokyo International (Haneda), and Congonhas respectively). And, for large cities with a single airport, it too is often the culmination of a steady outward progression. In Singapore, for instance, the first specifically commercial airport opened at Kallang in 1937, just outside the central business district (CBD). A new airport at Paya Lebar, located 10 kilometers from the CBD, opened in 1955, and then finally Changi Airport, at the northeast extremity of the island and 18 kilometers from the CBD, opened in 1980 (Bowen 2010).

Changi Airport also illustrates “terraforming” (Fuller and Harley 2004), a feature common
AIRPORTS

to some very large airports, especially in Asia. Changi is located almost entirely on reclaimed land, and the addition of space since 2000 for a planned third runway and massive new terminal has further distorted the city-state’s coastline. Other significant airports built in part or wholly on reclaimed land include Kansai International (Osaka), Hong Kong International, Incheon International (Seoul), Beirut International, and Nice Côte d’Azur among others. Elsewhere, airports have required the leveling of hills, the redirection of streams, and other changes to landforms.

In addition to space needs, another factor pushing large airports away from urban cores is the NIMBY response to environmental externalities, especially noise. Despite substantial improvements in the noise generated by individual jetliners, the increase in the number of flights and the spread of suburbs around once-distant airports have kept this issue at the forefront of aviation concerns. In response, airports have changed – to varying degrees – the way they operate. A notable example is Frankfurt, where since 2011 there has been a complete night curfew between 11 p.m. and 5 a.m. (aircraft noise is more annoying at night than during the day). Frankfurt is the busiest airport in the world with such a ban, but it is likely that more large airports – especially in Europe – will adopt similar restrictions. In the meantime, research suggests that, at least in some metropolitan areas, the burden of airport noise falls disproportionately on less well-off neighborhoods. Such neighborhoods are often found beneath a hub airport’s main landing and takeoff corridors, where land values are lower.

And yet the major commercial airport is also a vibrant source of economic activity. Geographers, economists, regional scientists, and others have documented the myriad benefits that accompany centrality in air networks. In particular, cities with good aviation accessibility tend to have disproportionately high shares of managerial and knowledge-intensive employment. Although the relationship between such jobs and air travel is obviously two-way (managers and knowledge workers fly more), careful statistical analysis makes clear that highly accessible cities have an advantage in attracting and growing these occupational categories.

The economic benefits of airports have fueled interest in the concept of airport-centered urban development in which the airport, rather than being on the periphery of a metropolitan region, is at its core. Sociologist John Kasarda coined the term “aerotropolis” to describe this phenomenon (Kasarda and Lindsay 2011). Although aerotropolises in various stages of development have been identified or proposed in regions around the world, the concept has been most fully embraced in East Asia and the Persian/Arabian Gulf. The speed of aviation growth in these regions provides ample scope for planning urban futures around new and expanded airports. Accordingly, the most celebrated examples of the aerotropolis concept include areas surrounding the main airports in Singapore and Dubai. In older air transportation centers, conversely, rather haphazard urban development typically already envelops established airports, precluding a form of integrated land use meant to optimize the economic effects of air accessibility.

Dubai certainly illustrates the potential for well-positioned cities and regions to harness aviation for economic development. The emirate’s air accessibility has been a catalyst for tourism, finance, retail trade, logistics, and other dynamic parts of its economy. And, in turn, Dubai has experienced faster traffic growth, both for air passengers and for air cargo, than almost any other large airport in the world (see Table 1;
Table 1  Leading airports ranked by passenger traffic, 2011.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Airport</th>
<th>Passengers (millions)</th>
<th>Average annual growth 2000–2011 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hartsfield–Jackson Atlanta Int.</td>
<td>92.39</td>
<td>1.3</td>
</tr>
<tr>
<td>2</td>
<td>Beijing Capital Int.</td>
<td>78.68</td>
<td>13.8</td>
</tr>
<tr>
<td>3</td>
<td>Heathrow (London)</td>
<td>69.43</td>
<td>0.7</td>
</tr>
<tr>
<td>4</td>
<td>Chicago O’Hare Int.</td>
<td>66.70</td>
<td>−0.7</td>
</tr>
<tr>
<td>5</td>
<td>Tokyo Int. (Haneda)</td>
<td>62.58</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>Los Angeles Int.</td>
<td>61.86</td>
<td>−0.6</td>
</tr>
<tr>
<td>7</td>
<td>Paris Charles de Gaulle Int.</td>
<td>60.97</td>
<td>2.3</td>
</tr>
<tr>
<td>8</td>
<td>Dallas–Fort Worth Int.</td>
<td>57.83</td>
<td>−0.4</td>
</tr>
<tr>
<td>9</td>
<td>Frankfurt</td>
<td>56.44</td>
<td>1.2</td>
</tr>
<tr>
<td>10</td>
<td>Hong Kong Int.</td>
<td>53.33</td>
<td>4.7</td>
</tr>
<tr>
<td>11</td>
<td>Denver Int.</td>
<td>52.85</td>
<td>3.0</td>
</tr>
<tr>
<td>12</td>
<td>Soekarno–Hatta Int. (Jakarta)</td>
<td>51.53</td>
<td>16.1</td>
</tr>
<tr>
<td>13</td>
<td>Dubai Soekarno–Hatta Int.</td>
<td>50.98</td>
<td>15.7</td>
</tr>
<tr>
<td>14</td>
<td>Amsterdam Airport Schiphol</td>
<td>49.76</td>
<td>2.1</td>
</tr>
<tr>
<td>15</td>
<td>Madrid Barajas Soekarno–Hatta Int.</td>
<td>49.65</td>
<td>4.1</td>
</tr>
</tbody>
</table>

Table 2). In 2010 Dubai opened the first stages of the new multibillion dollar Al Maktoum International Airport, which is located 45 km from the city’s older airport.


Yet the fortunes of airports are not solely the product of larger social or regional forces. Key airlines play a decisive role in determining whether an airport’s stature is rising or falling. Emirates Airline and Lion Air, for instance, have been instrumental in the growth of passenger traffic at Dubai and Jakarta, respectively. That is even more obviously the case among the world’s leading air cargo hubs, where the importance of Memphis International Airport, for instance, is almost singularly attributable to the success of FedEx.

Each of these airports – whether a passenger hub, cargo hub, or both – is a portal to a global economy dependent on the space–time convergence fostered by aviation. As globalization’s gateways, large hubs – especially new ones – feature vast terminals whose architectural extravagances are meant to make an impression in a world of mobile people and capital. For example, the luxuriant gardens and brisk efficiency of Changi Airport are emblematic of
Singapore. For other cities, too, the airport is an important arena of place-making. This is an ironic outcome given the perception of aviation as a technology that has rendered geography less important by shrinking space.

A further irony lies in the contemporary reality that a major airport is simultaneously an opening through which it is possible to travel far and fast and a carefully controlled checkpoint through which movement is tightly circumscribed. Via a nearby commercial airport, one can be in virtually any large city in the world within 36 hours. Yet the speed, accessibility, reach, visibility, and fragility of the air transportation system have made it central to key problems in the early twenty-first century, including terrorism and the spread of infectious disease. As a consequence, air travelers and air cargo are subject to varying levels of surveillance, probing, and in some instances exclusion.

**SEE ALSO:** Accessibility, in transportation planning; Air transport; Cities and development; Time–space convergence; Transport geography; Urban geography

### References


---

**AIRPORTS**

**Table 2** Leading airports ranked by cargo traffic, 2011.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Airport</th>
<th>Tonnes (million)</th>
<th>Average annual growth 2000–2011 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hong Kong Int.</td>
<td>3.98</td>
<td>5.2</td>
</tr>
<tr>
<td>2</td>
<td>Memphis Int.</td>
<td>3.92</td>
<td>4.2</td>
</tr>
<tr>
<td>3</td>
<td>Pudong Int. (Shanghai)</td>
<td>3.09</td>
<td>13.1</td>
</tr>
<tr>
<td>4</td>
<td>Ted Stevens Anchorage Int.</td>
<td>2.54</td>
<td>3.2</td>
</tr>
<tr>
<td>5</td>
<td>Incheon Int. (Seoul)</td>
<td>2.54</td>
<td>2.8</td>
</tr>
<tr>
<td>6</td>
<td>Paris Charles de Gaulle Int.</td>
<td>2.30</td>
<td>3.3</td>
</tr>
<tr>
<td>7</td>
<td>Frankfurt</td>
<td>2.21</td>
<td>2.4</td>
</tr>
<tr>
<td>8</td>
<td>Dubai Int.</td>
<td>2.19</td>
<td>12.8</td>
</tr>
<tr>
<td>9</td>
<td>Louisville Int.–Standiford Field</td>
<td>2.19</td>
<td>3.4</td>
</tr>
<tr>
<td>10</td>
<td>Narita Int. (Tokyo)</td>
<td>1.95</td>
<td>0.0</td>
</tr>
<tr>
<td>11</td>
<td>Changi (Singapore)</td>
<td>1.90</td>
<td>1.0</td>
</tr>
<tr>
<td>12</td>
<td>Miami Int.</td>
<td>1.84</td>
<td>1.1</td>
</tr>
<tr>
<td>13</td>
<td>Los Angeles Int.</td>
<td>1.70</td>
<td>−1.7</td>
</tr>
<tr>
<td>14</td>
<td>Beijing Capital Int.</td>
<td>1.64</td>
<td>7.1</td>
</tr>
<tr>
<td>15</td>
<td>Taipei Taoyuan Int.</td>
<td>1.63</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Source: Airports Council International 2013.

**Further reading**

Distinguishing between native and alien species is a key criterion in nature conservation, but both the distinction itself and policies based on it are contested. Such debates are profoundly geographical, hinging on questions of space, place, scale, and environment. They are important practically, because of the impacts of invasive alien species, and also conceptually, influencing how we categorize and value components of “nature.” In turn, this shapes environmental policies, conservation priorities, and social perceptions of the nonhuman biophysical world.

Defining the terms “alien” and “native” is challenging because the terms are socially constructed, emotive, and contested (Warren 2007). Discussions of definitions dating back 150 years have yet to produce consensus or consistent use. However, as commonly understood, native species are those that evolved in an area or colonized it naturally, whereas alien species are those that have been introduced beyond their natural range by humans, intentionally or otherwise. In a vernacular sense, native species “belong” in an area, whereas alien species do not. Other common terms for aliens are “non-native,” “exotic,” and “introduced,” respectively emphasizing such species’ unnaturalness, their distant place of origin, and the human agency involved.

Trade, travel, and migration have been altering the distributions of plants and animals throughout human history. Often this has been deliberate, involving the transport of crops, livestock, quarry species, garden plants, and pets, but in many instances it has been unintentional as seeds, animals, and organisms have been unwittingly carried in soil, ballast water, or cargo holds. While this has been happening for millennia, a dramatic increase in the rate and scale of such transport accompanied the period of European colonial expansion around the globe as imperial conquests triggered “ecological imperialism” (Crosby 1986). Rates of species redistribution then rose still further with the exponential increases in global trade and travel during the late twentieth century.

Most species translocated to new environments are unproblematic and some are highly beneficial for society. Indeed, alien species are integral to socioeconomic systems worldwide through their importance in agriculture, silviculture, horticulture, medicine, hunting, and the pet trade. However, a small minority of introduced species run amok in their new habitat. By outcompeting, dominating, disrupting, preying on, hybridizing with, or spreading disease among native species they can transform entire environments with profound and often irreversible ecological impacts. Furthermore, they can severely threaten agriculture, forestry, and water management, pose risks to human health, disrupt ecosystem services and cause significant cultural impacts. Worldwide, by one estimate, the costs in losses, damage, and control expenses are equivalent to 5% of global GNP annually. Numerous case histories are detailed by Simberloff (2013), but particularly infamous examples are rabbits in
Australia, zebra mussels in US waterways, and water hyacinth in Lake Victoria. Island biotas, in particular, have proved notoriously vulnerable to bio-invasions (frequently by ship-borne rats), as have anthropogenic habitats.

Consequently, biological invasions are sometimes cited as one of today’s four great ongoing global transformations, alongside (and interacting with) climate change, enhanced biogeochemical cycles, and land use change. It is hard to predict which introductions will become “problem species,” how soon, and in what ways, but the proportion that does so is low. Nevertheless, the impacts of this small minority are sufficiently serious to have generated strong policy responses nationally and internationally; over 40 international instruments address non-native species, notably the Convention on Biological Diversity (1992) and the Syracuse Charter (2009). Such policies are commonly designed to prevent, eradicate, contain or control alien species, and to preserve or re-establish native biodiversity and functions.

Although attempts to categorize species as native or alien began in the 1830s, concern about alien species is a modern phenomenon. Indeed, until the early twentieth century, the introduction of non-native species was actively encouraged, not least by acclimatization societies, for reasons of resource provision, environmental “improvement” and/or aesthetics. Today’s concerns are commonly traced to Charles Elton’s 1958 book *The Ecology of Invasions by Animals and Plants*. The inception of the subdiscipline of invasion ecology followed 30 years later, focused on understanding, predicting, and managing bio-invasions. Today, nativeness is widely considered to be a central organizing principle in conservation, and vast expenditure and effort is invested in campaigns against aliens. Whereas Elton himself saw a place for “exotic forms,” most conservationists now disparage all non-natives. This recent growth of anti-alien sentiment represents a rapid evolution in value judgments.

There are diverse motivations for preserving native species and resisting alien species. The ecological arguments for doing so relate to the functioning of ecosystems and the preservation of biodiversity. Although the arrival of new species can enhance biodiversity locally, globally the “great reshuffling” of species decreases it, prompting fears of ecological McDonaldization and the dawn of the “Homogocene” era as biodiversity is replaced with biosimilarity (Hettinger 2001). Economic arguments relate to the huge costs of controlling biological invasions, and there are ethical and aesthetic grounds for arguing that it is good and pleasing to preserve native biota, not least the moral case that humans bear a responsibility to reduce and reverse our negative ecological impacts.

Equally, there are many economic, social, cultural, philosophical, and pragmatic reasons why different individuals and sectors of society oppose controlling some or all introduced animals and plants. Attitudes toward native and alien species, and beliefs about appropriate management, flow from deeply held worldviews and philosophical commitments; anthropocentric, ecocentric, zoocentric (animal-focused), and biocentric perspectives lead to contrasting beliefs about the rights and wrongs of natives and aliens. Moreover, perceptions evolve through time, so that the “pests” of one generation become cherished native species in the next (e.g., red squirrels in the UK) (Smout 2011). The divergent nature and groundings of these pro- and anti-motivations help to explain why the attendant discourses are so passionately contested because the issues sit at the intersection of conservation science, ethics, values, public policy, and social perception.
Controversies

Debates surrounding definitions

Although the terms “native” and “alien” are often employed as binary opposites, they actually represent the endpoints of a complex continuum, making definition both difficult and contested. Rigorous definition is undermined by their essentially relative nature, both in time and space. The terms do not describe biological characteristics but dispersal history. Consequently, no species is inherently alien or native but only with respect to a particular environment at a particular moment, and the spatial and temporal boundaries of that environment can be constructed in many different and changing ways (Warren 2007; Boonman–Berson, Turnhout and Van Tatenhove 2014). For example, to be native, how long must a species have been present and within what kind and scale of space: ecological (e.g., ecosystem) or sociopolitical (e.g., nation, conservation area)? To be alien, from how far away must it have come? If a species was introduced in pre-history by indigenous people, should it be classified as alien or native? If a once-native species which has been lost is reintroduced, is it native (because it “belongs”) or alien (because it has been transported from another locale by human agency)? Such questions do not have scientific answers, and, in practice, they are answered in widely divergent ways. Choices about spatiotemporal scale critically affect classification (and hence management policies), yet scale is socially constructed, not naturally given. The native/alien paradigm is thus “not purely dependent on objective ecological criteria, but on the kind of time–space demarcations we use to identify origin or authenticity” (Hattingh 2001, 191). Since the continuum involves sliding scales, numerous intermediate terms exist such as “locally non-native,” “naturalized,” “archaeophyte,” and “neophyte.”

The definitional lines are blurred by instances in which species introduced centuries ago have become functionally native, an integral part of the food chain, such that their removal would be ecologically damaging. Some dub these “honorary natives,” an additional, imprecise category. Opinions also diverge about the degree to which genetic data should be used in applying native or alien status, and about what constitutes human agency in the arrival of non-native species; should it, for example, include anthropogenic environmental change which facilitates migration? A final, empirical, problem in classifying species according to origin is that, for many species, scientific knowledge about their status is partial and evolving.

More broadly, the concept of nativeness is criticized for being premised on a pre-Darwinian separation of human beings from nature. It only makes sense if anthropogenic species dispersal renders nature unnatural and can only operate if humans are excluded from it. Otherwise, logically, most of humanity would be classified as invasive aliens meriting eradication, so that saving nature would mean killing ourselves. It is also criticized for fossilizing nature by implying that it is static and timeless and by specifying a fixed and historically defined list of biota which may not suit future conditions.

Such considerations have generated critical questioning of the native/alien framework and its soundness as a basis for policy and management. Its defenders, however, argue that these objections apply to only a minority of marginal cases and that in most contexts there is sufficient clarity for the distinction to be useful and applicable (Simberloff 2013).

It is important to differentiate between “alien species” and “invasive alien species” (IAS). Difficulties have arisen from the elision of these two terms, and/or from the use of the former as a synonym for the latter, not only in policy and
popular discourses but sometimes in academic writing. In this way, all alien species become associated with negative impacts, creating a dualistic “native – good, alien – bad” narrative. This is inaccurate and simplistic, and has exacerbated some of the associated sociocultural debates.

Sociocultural debates

Policies designed to control IAS rest on justifiable ecological and economic foundations, but they often clash with cultural values. Whereas conservation science promotes native biodiversity, many people welcome exotic species because of their charismatic nature and their contribution to the overall richness of landscapes, wildlife, and human wellbeing. Some non-natives have even become cultural icons in their new settings, and are woven into people’s sense of place, such as Tuscany’s cypress trees and even the much-vilified Cane Toad in Australia. The lay public is often blind to the native/alien categorization, caring little about origins but greatly about “nature.” Campaigns to eradicate alien species can therefore be fiercely resisted. When popular perceptions of “cultural belonging” collide with scientific judgments that alien species do not belong ecologically, anti-alien policies can alienate the public on whose support conservation depends. Conflicting value systems can thus create incompatible convictions about the management of species, particularly about whether the preservation of one species can justify the killing of members of another. Opponents of eradication campaigns, especially animal rights activists, point out that it is humans, not the target species, who are responsible for their being in the “wrong” place. Such fault lines can create particularly acute tensions when they separate indigenous and nonindigenous groups. Opinions diverge widely on how to balance cultural preferences with conservation objectives.

A particularly vexed issue is the criticism that anti-alien policies are racist or eco-fascist, akin to ethnic cleansing. The racist critique has been extensively articulated, supported with historical evidence of the horticultural nativism promoted in Nazi Germany and with contemporary evidence that minority ethnic groups can feel threatened and insulted by ecological pro-native rhetoric (O’Brien 2006). Invasion ecologists dismiss this accusation, arguing that the motivation for resisting alien species is not an objection to their foreign origins but the desire to avoid negative impacts and to preserve distinctive, diverse communities; although native species can become invasive, the probability of damage from non-natives is far higher. But while conservationists may be entirely innocent of racism, the parallels between arguments for preserving “genetic integrity” in nature with those for racial purity in human communities are close; in both cases, foreign influence is portrayed as contaminating and threatening, and the rhetoric employed is often militaristic and inflammatory, especially in media coverage. The widespread use of “aliens” and “invasives” as interchangeable synonyms compounds this difficulty by erroneously implying that all introduced species are inherently problematic because of their foreign origins. This is exacerbated when native status is defined using political boundaries because this directly parallels nationalism; indeed, pro-native arguments are often overtly nationalistic. In the context of the pervasive intertwining of culture with nature, the argument that an anti-foreign stance is reprehensible in human affairs but benign when applied to nature is hard to sustain; “there are compelling arguments that nativist purism is undesirable in all spheres – politically, culturally and ecologically” (Peretti 1998, 189). Some therefore criticize the vocabulary of “aliens” as prescriptive and pejorative, and call for its
replacement with more descriptive and apolitical terms. These might be the older nomenclature of “pest” and “weed” which highlights impacts rather than origins, or simple temporal labels such as “long resident” and “newly arrived.”

Natives and aliens in a globalized, warming world

The ecological and economic arguments for exercising vigilant control over the import of new species are strong, and some countries, notably New Zealand, operate strict biosecurity policies. However, in the context of a hyper-connected world, it is unlikely that biodiversity preservation policies will restrain global free trade and most ecosystems worldwide already include many introductions. Millennia of anthropogenic species mixing cannot be undone, and will inevitably continue, so aliens are here to stay. Moreover, in many cultural landscapes, removing established aliens can harm ecosystem services and have unintended consequences (ecologically and socially). These realities lead some to argue that, because “resistance is futile,” established alien species should be adopted as natives and that the ecological value of “novel” or “hybrid” ecosystems should be recognized. A similarly pragmatic, positive response is the “LTL” approach – that we should Learn to Love alien species (Davis 2009) rather than waste resources on fighting unwinnable wars. But any reclassification of aliens as “new natives” further dilutes the native/alien distinction.

It is blurred still more by the prospect that global climate change will increasingly destabilize extant biotic distributions as current assemblages of native flora and fauna struggle to adapt to rapid climate change; non-native species may enhance ecosystem resilience, thus transforming aliens into allies. Anthropogenic climate change also raises new definitional dilemmas: for instance, do climate-driven migrations constitute benign range expansions or alien invasions? Most fundamentally, the difficulties and controversies associated with both defining and applying the alien/native distinction have led some to conclude that the classification itself is sufficiently flawed, nebulous, and unhelpful that it should be abandoned. Thus Chew and Hamilton (2011, 45) contend that “nativeness is a living fossil of an outmoded phytogeography.”

Such critiques date back to the 1960s but gathered pace from the 1990s as both social and natural scientists challenged the dichotomous framing and normative prescriptions of invasion ecology (Davis 2009; Rotherham and Lambert 2011), generating sharp exchanges in the literature. Notably, in 2011 when Mark Davis and 18 fellow ecologists argued in *Nature* that we should not judge species on their origins but adopt pragmatic policies better suited to a rapidly changing world, 141 invasion scientists wrote a strong rebuttal, rejecting the criticisms as being aimed at straw men. Simberloff (2013), the author of that response, argues that abandoning the distinction and surrendering to bio-invasion is overly pessimistic regarding our ability to control invasions because there have been many notable successes, and also naively optimistic about our ability to shape “novel ecosystems” beneficially while avoiding widespread problems. A thoughtful middle way through the polarized debate is charted by Shackelford *et al.* (2013) who suggest that the longer established a species is, the less significant a criterion its origin should be.

There is increasing agreement that applying the native/alien distinction in a dichotomous, dogmatic, and universal fashion is unreasonable and unsustainable, and that, instead, a discriminating and context-sensitive approach should focus on those species which pose the greatest risks, recognizing that few aliens are invasive, that
not all invasions cause significant harm, and that invasiveness, too, is a debated term. Notwithstanding the unpredictability of species’ future behavior, this targeted, impact-focused approach is the official basis of most conservation policy and practice; pragmatic realism often trumps principled opposition to non-natives, even to the extent, in certain contexts, of conserving aliens and controlling natives. Nevertheless, a belief that native species are superior because they rightfully “belong” pervades conservation and most invasion ecologists regard aliens as inherently undesirable. The fact that a species is alien is often prominent among stated reasons for controlling it, and the precautionary principle is deployed to justify calls for eradicating all newly detected introductions. In popular conservation discourses blanket vilifications of all alien species are common, employing stigmatizing terms like “junk species,” “ecological tumors” and “biological/genetic pollution.” Yet the implicit value claim – that natural equals good – equates to the naturalistic fallacy.

The importance of differentiating between change and harm, and between description and prescription, is often highlighted. Whereas change can be scientifically quantified, harm is socially constructed and value-laden. Similarly, while understanding the influence of human agency on today’s species distributions is unquestionably important for biogeographical explanation, policy prescriptions which rest primarily on such data have been widely criticized. The value judgments cannot be read directly from the science. Context is all; a species may have positive and negative impacts at different times and in different places. It is damaging behavior not alien status per se which poses a threat, and both aliens and natives can inflict serious ecological and economic damage.

Debates concerning the conceptual coherence and practical implementation of the native/alien construct are closely related to broader discussions about nature, naturalness, and the place of our own species in the “natural” world. They also intersect strongly with debates about ecological restoration and re-wilding. The period in which the native/alien framework has attracted growing critical scrutiny has, paradoxically, seen a steady reinforcement of that framework in conservation policy, and its rise to prominence in public discourse, demonstrating a disconnect between academic discourse, policy-making, and public debate. Because the Cartesian separation of nature and culture is so deeply engrained, polarities such as natural/artificial, authentic/fake, and native/alien are likely to persist.

Native and alien species may always be the focus of debate because of the multiple, competing values attached to them. Nevertheless, it seems clear that a framework in which all alien species are proscribed and all natives are sanctified is unsupportable, and that management prescriptions based solely on a species’ geographical origins are inappropriate and often counter-productive. But the potential for introduced species to cause significant damage argues for the inclusion of data about species’ origins in management decision-making, whether or not the contentious terms “alien” and “native” are retained.

SEE ALSO: Biogeography; Biosecurity; Climate change and biogeography; Construction of nature; Ecological imperialism; Environment, nation, and “race”; Nature conservation

References


Further reading


Alliances and franchises

Eirik Vatne
Norwegian School of Economics, Norway

Definitions

The term “alliance” is commonly understood as a voluntary, cooperative arrangement among independent organizations or business firms. This can involve exchange, sharing, or co-development of production, distribution, and knowledge, and include contributions by partners of capital, technology, knowledge, and other firm-specific assets. An alliance can take the form of a dyadic relationship as seen in “strategic alliances” and “equity joint ventures,” or as multilateral relationships between many firms as observed in “networks of firms.”

The term “franchise” is understood as a relationship between two independent firms based on a contractual agreement. In this arrangement a franchisor grants a franchisee the right to operate a business venture in a specific manner, place, and timespan. In return the franchisee must pay royalties/fees to the franchisor. Franchising arrangements take two general forms, as product/brand name franchising still seen in some of the distribution systems for cars or soft drinks, or as business format franchising as often observed in restaurant or retail chains.

Terms like alliance or partnership are general, but ambiguous concepts including a wide variety of institutional forms of inter-organizational relationships. Franchise is a specific contractual-based form of business relationship. Both concepts belong to the multidisciplinary research field of inter-organizational relations (IOR).

Economic governance

The traditional economic approach to understanding the organization of economic transactions is to view firms as autonomous, atomistic entities forced to compete in a free, impersonal open marketplace. The selection of a transaction partner is only based on a price/quality measurement set by the competition in a spot market. In this world there is no need for collaborative arrangements between firms. Indeed, exchanges between firms other than “arms-length” market transactions can be seen as ways to circumvent the disciplinary impulses of competition, to create cartels and to achieve monopoly power.

In a more realistic world, free market transactions will often generate costs related to activities such as simply discovering what the price would be in the market. There could be a need to handle uncertainties and risks, a need for safeguarding or monitoring outsourced production – all actions that generate costs.

If the cost of handling a market transaction is higher than the cost of internal governance, the firm will choose to internalize the transaction and a hierarchical, bureaucratic governance form will be chosen. However, if the cost of internal governance is higher than the cost generated by a market solution, the activity will be outsourced and market governance will be chosen. The firm will either make a specific item or service internally (e.g., vertical integration), or buy it in an open spot market (e.g., vertical disintegration).
ALLIANCES AND FRANCHISES

Between these two generic forms of governance, we can identify different forms of collaborative arrangements. For a given activity, a firm can create an alliance and operate jointly with the help of partners. If this is the most cost-efficient way to organize transactions, a third governance form, the hybrid or network form of governance, is chosen.

Inter-organizational relationships

Ties in inter-organizational relationships are often enduring and of strategic importance for the firms entering into them. They are most often based on attempts to gain cost reductions through a division of labor between partners, to collectively gain scale economies or to innovate by sharing information and co-producing new knowledge. Occasionally collaborative arrangements could also be an attempt to achieve monopoly profit, but the IOR field is basically interested in the value adding and efficiency seeking arguments behind inter-organizational relationships.

Table 1 illustrates some aspects of inter-organizational arrangements along the continuum from fully integrated to pure market-based governance. As shown, the degree of integration and the institutional form chosen will vary according to the strategic importance of the resources controlled by other partners, the complexity and risk of sharing resources, the degree of embeddedness, the importance of legal rules and social norms as incentives and control mechanisms, and so on.

A strategic alliance is formed when two or several firms combine resources to develop a specific activity of strategic importance for both parts. This can include activities important for maintaining the core skills of firms, reaching specific business objectives, reducing significant risks, blocking a competitive threat, or developing future business opportunities. Strategic alliances involve activities related to the competitive core of the firm and are often extended in time. Consequently, the qualities of a partner and the selection process are important, over and above contractual safeguarding. Contract-based institutional forms in the upper part of the table are normally related to strategic alliances.

A franchise involves two partners, one dominant firm with ownership of a valuable product, trademark or knowledge, and a weaker partner paying a fee to be allowed to distribute a trademarked product or set up a business concept for a specific territory. Many franchisees can be related to one franchisor. An asymmetric balance of power normally exists in franchise arrangements, and activities taking place are often standardized and easy to control. A legal, contractual agreement therefore regulates and safeguards transactions organized as franchises.

The “system area” form of institutional governance is understood as a network of multilateral relationships with many partners involved. A mutual technological or territorial context is often an important explanatory factor for the appearance, shape, and governance of the partnership. A complex supply chain connected to a large firm is often dominated by the lead firm and inter-organizational arrangements are tight and formalized. An industrial district model comprises many smaller firms collaborating in a looser and more informal way.

In all these cases, we focus on collaboration between independent firms with the purpose of achieving some common objectives to create competitive advantages through the partnership. Flexibility in starting and terminating collaborative arrangements, serving specific purposes, is a crucial argument for alliances among firms.

The choice of organizational form of an alliance depends on the purpose of the exchange,
<table>
<thead>
<tr>
<th>Degree of integration</th>
<th>Network structure</th>
<th>Strategic importance</th>
<th>Examples of types of ties</th>
<th>Examples of institutional form of governance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully integrated</td>
<td>Internalized</td>
<td>Complicated, open-end problems with high, enduring values to the firms core skills</td>
<td>Ties with long duration, large investments conditioned by a specific transaction, high degree of cooperation, involved knowledge resources</td>
<td>Hierarchy</td>
</tr>
<tr>
<td>High</td>
<td>Tight</td>
<td>Network: stability is high and mutual engagement high</td>
<td>Flexible ties with important elements of coordination and familiarity</td>
<td>Equity agreements</td>
</tr>
<tr>
<td></td>
<td>Inter-org.</td>
<td></td>
<td>Dialog-based ties where no standard solution is present, direct contact needed</td>
<td>Joint ventures</td>
</tr>
<tr>
<td></td>
<td>arrangement</td>
<td></td>
<td>Routine-based ties, cooperation is less formal, many-sided and based on standard specifications</td>
<td>Equity investments</td>
</tr>
<tr>
<td>Low</td>
<td>Loose</td>
<td>Network: stability is low and engagement low</td>
<td>Ad hoc ties with uncomplicated dialog, no mutual insight needed, investments in the ties are low</td>
<td>Non-equity agreements</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Price determined ties based on a standardized supply</td>
<td>Contract-based:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cooperatives</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Research and development consortia</td>
</tr>
<tr>
<td>Atomistic market</td>
<td>Pure market</td>
<td>Uncomplicated problems with low strategic importance to the firms core skills</td>
<td></td>
<td>Control franchising</td>
</tr>
<tr>
<td></td>
<td>decisions</td>
<td></td>
<td></td>
<td>Technology financing</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Licensing agreement</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Non-contract-based:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cooperate accords</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>System area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lead firm supply chain</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Subcontracting networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Industrial districts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Temporal coalitions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Action sets</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pure market transactions</td>
</tr>
</tbody>
</table>

Source: Modified from Christensen et al. 1990.
the degree of interdependence, and the strategic importance of the resources controlled by each partner. The specificity of investments needed to realize the exchange and the social and territorial embeddedness of the partners will also influence this decision. Some factors are specifically based on rational, short-term “economic” behavior; others are based on a specific context opening up for entrepreneurship, more complex interaction, and expanding interdependencies encouraged by a common culture for sharing and collaboration and as a consequence, collective production of externalities.

**Explaining inter-organizational relations**

A few theoretical approaches dominate the literature explaining the relational aspects of inter-organizational arrangements.

The transaction cost approach has already been introduced. The main idea is that a transaction in a free market could generate considerable costs caused by bounded rationality, opportunism among exchange partners, small number of exchange partners bargaining, and asymmetric distribution of information. Williamson’s approach (1979) stressed that the cost of transactions will increase if high transaction-specific investments are involved with low or no value in other use, the definition and performance of the transaction are ambiguous, and if the frequency of transactions is sporadic. Based on these elements, the transaction cost model will predict governance form. If investments are non-specific and uncertainty is low, market-based governance will be chosen. If transaction-specific investments are high, combined with high frequencies of transactions and high uncertainty, internalization is preferable. In situations where uncertainty and asset-specificity are at an intermediate level, a hybrid governance form will be chosen and alliances are preferred. This approach emphasizes cost minimization as the main reason why firms should enter into inter-organizational relations.

The resource-based view of the firm views a business organization as a bundle of resources understood as a set of tangible and intangible assets semi-permanently tied to the firm. Many resources are firm-specific, such as specialized production technology or knowledge bases. Firm-specific resources are scarce, difficult to imitate, often intangible, heterogeneous, and complex, and therefore unique without direct substitutes. They could be property-based resources such as machinery, patents, and trademarks, or knowledge-based resources such as organizational and technological capabilities. A firm’s access to unique resources and capabilities are the foundation for the firm’s competitive strength. A bundle of unique resources are part of the firm’s core skills and are protected through internalization. Unique resources controlled by other firms can also be mobilized through collaboration. The resource dependence theory analyzes such interdependencies. If the firm-specific resources in an alliance are complementary and combined, a division of labor can be developed between the partners. Collaboration thereby expands the possibility of deepening the core skills of each firm. Combining complementary core skills also expands the capacity to learn and thus to develop unique cross skills as a combination of the capabilities of the partners. In short, the resource-based view emphasizes value maximization through the sharing and exploitation of valuable resources, not accessible through the open market.

A third approach to exploring inter-organizational relationships could be called the network approach with a specific focus on embeddedness and local tie formations. A network is in general form a set of several actors directly or indirectly connected to each other through ties. Ties connect pairs of actors, some directly, others
indirectly through individual ties to a mutual actor. The structure of a network and the position or centrality of a specific actor is important and determines an actor’s connectivity. In this perspective the formation of inter-organizational relations is also seen as a consequence of socially embedded actions. The value of an actor’s connections can be summed up as the actor’s *social capital* – a form of capital that gives an actor access to information and resources not available through ordinary, open market relations.

An additional focus on *embeddedness* has been important. Economic actions are considered as structurally embedded in a sectoral, social, and territorial context. Repetitive relations among people and business partners generate an embedded logic of exchange different from “arms-length” market relations. A high degree of specialization and division of labor inside business networks will make necessary a high degree of adaptability, coordination, and control. Under such circumstances, behavior will be institutionalized, formed by local and social institutions, and governed by economic as well as social incentives and control mechanisms. An *untraded, relational asset* is created and embedded in local communities and in networks of enduring collaboration. Under such conditions, relations can be habitually based on trust and implicit and open-ended contracts.

**Firm relationships and geography**

In economic geography, the relational problem was earlier seen through a focus on regional economic growth explained in models based on agglomeration economies and external economies of scale. Here firms in a local economy are connected through input–output relations and technological and pecuniary interdependencies. Studies of the spatiality of industrial systems emphasized the content and structure of linkages between actors and how growth impulses diffuse through these relations. Still, the discrete relationship between individual actors of the system was seldom addressed. The same could be said about the early “geography of the firm” literature preoccupied with locational and spatial-behavioral patterns of the multi-regional enterprise as the unit of study.

Studies of “Marshallian industrial districts” emphasized localized, *flexible specialization* as an alternative organizational form to large-scale mass production (Fordism). In this model, a comprehensive specialization and division of labor combined with interfir interaction between small, local firms, are at the core of explanation. An additional focus on the importance of localized sociocultural and collective assets embedded collaborative arrangements in a territorial context. This territorial model of concerted inter-organizational relationships was suggested to be flexible, cost-efficient, and well adapted to a turbulent business environment under constant transformation.

A transaction cost approach was a starting point for the “Californian School of External Economies.” Market and technological uncertainties combined with increasing international competition forced large corporations to restructure. The subsequent disintegration and outsourcing of activities created a complex system of exchanges and interdependencies that could increase transaction costs. Agglomeration of activities could, on the other hand, help to reduce transaction costs. In this model, routinized production tasks would relocate to low-cost locations and be governed through market or hierarchical transactions. More complex tasks with a need for intricate combinations and coordination of firm-specific resources would create a prerequisite for interfir collaboration.
and agglomeration of these activities. This simple linkage-transaction cost model was later extended to include other elements of community based external economies and spillovers important for the development of “untraded interdependencies.”

Other studies, based on Porter’s cluster concept, examine spatial clusters of economic agents. A comprehensive division of labor, local specialization, collaboration and rivalry, production of externalities, local institutions, and the influence of culture and politics are all part of a socioeconomic context that can create competitive strength for firms embedded in such environments. In this body of research, localized inter-organizational exchanges and alliances are highlighted as strategically important for firms.

The literature on “learning-regions,” “innovative milieus,” or “regional innovation systems” also emphasizes collaborative initiatives and proximity among actors as important for the formation of new knowledge. Region-specific assets such as institutionalized social conventions, habits, trust-based relations, and social control mechanisms, all embed the economic agent in a sociocultural context. This context minimizes transaction costs and facilitates diffusion of information and a division of labor among independent firms, opening up for trust-based interaction and sharing of firm-specific resources, facilitating interactive learning processes and joint development of support organizations and collaborative initiatives. Region-specific assets facilitating collaboration can be seen as relational assets – untraded, intangible, and localized assets that create an atmosphere for a trust-based, cost-efficient interaction among firms open for learning and knowledge exchanges.

In these approaches, economic geographers study relations between firms in the context of a territorially embedded network of producers – as systems of relations rather than discrete relations among few actors. Geographers’ specific contribution to the IOR field is therefore a deeper understanding of the importance of a socio-spatial context as a container of relational assets facilitating the formation of trust-based, inter-organizational relationships. Localized, relational assets can influence the flexibility, efficiency, and innovativeness of collaborative arrangements and the competitive strength of individual firms. This contribution relates first and foremost to the “system area” and “network perspective” of inter-organizational relations. Several of these aspects are further developed under other headings in this encyclopedia – see cross-references below.

A strategic alliance or franchise arrangement is formed when two or a few firms join resources to develop a specific activity of strategic importance for at least one part, normally for all. Here dyadic relationships or partnerships among a small number of firms are under investigation. Studies of these forms of inter-organizational relations are uncommon in economic geography.

In the literature on innovation and clusters, an important empirical result has been the observation that the region or local milieu is not necessarily the most important arena for accessing unique information or developing new knowledge. Pipelines to global resources are often as important as local dynamics. Access to extra-local, untraded resources is often secured through strategic alliances as a “research consortia” for firms or membership in “communities of practice” for individuals.

In the literature on the geography of transnational corporations (TNCs), strategic alliances such as “joint ventures” or “technology financing” are important arrangements used by the TNCs to access and utilize unique resources worldwide or to share risks and investments in capital intensive and scale sensitive activities. To
penetrate new markets and countries, TNCs need access to specific knowledge of local regulations, institutions, and markets or a fast access to production capacity or a domestic distribution system. A strategic alliance or a franchise arrangement with a host country firm is an attractive option compared to an internalized, costly, and risky expansion into new territories. The geography of TNCs also seems to show that certain cities with a heterogeneous resource endowment and interconnections to other hubs of resources are the preferred location for collaborative initiative.

In the literature on the geography of retail or service industries, franchising is seen as a strategy to achieve scale economies through rapid expansion of production and markets, based on standardization and replication of practices. Reasons why firms choose to join forces with a franchisee can be a general resource scarcity problem under rapid expansion, an agency problem, or a need to access relational capabilities such as the social capital of a franchisee and knowledge of local specific circumstances.

**SEE ALSO:** Corporate spatial organization and producer services; Externalization; Industrial agglomeration; Industrial districts; Internationalization; Local embeddedness; Spatial social networks; Vertical integration

**References**


**Further reading**


The alternative food movement refers to recent social movements made up of diverse activists, organizations, institutions, and enterprises aiming to create food systems that differ from industrial agriculture and the industrial food system. There is considerable diversity in what is labeled as “alternative” (Venn et al. 2006). Additionally, it is more accurate to discuss alternative food movements (AFMs) rather than a singular one. Indeed, creating “convergence in diversity” among AFMs remains important political work (Amin 2011).

Agriculture in what is now the industrialized world was continuously revolutionized in the 1800s and 1900s as production for use value gave way to production for exchange value. Machinery, synthetic fertilizers, synthetic pesticides, hybrid or GMO organisms, and fossil fuels now form the backbone of industrial production. The origins and widespread adoption of these inputs have been explained as part of the treadmill of production and substitution of capital for ecological processes that is set in motion through the competitive and accumulative dynamics of capitalism (Cochrane 1979; Goodman, Sorj, and Wilkinson 1987). Industrial agriculture is supported by vast marketing, infrastructural, knowledge, and ideological systems. The rhetoric of feeding the world, called productionism, is commonly used to justify industrial agriculture, but this ignores the true causes of hunger (Lappé et al. 1998). The industrial food system that extends beyond primary production – transportation, processing and packaging, distribution, retail, and disposal – is intensive in fossil fuel use and substantially transforms most raw food ingredients into processed food by applying vast amounts of energy and anthropogenic chemicals (postharvest pesticides, preservatives, additives for coloring and taste, etc.).

Actors within AFMs argue that there is much wrong with the industrial food system. While generally productive on a per acre basis and good at provisioning food to a minority of the world’s population (those with adequate incomes and entitlements, largely in the industrialized world), critics draw attention to the industrial food system’s considerable and widespread negative consequences or “externalities” for society, public health, food access, workers, food sovereignty, gastronomy, environment, soil, climate, and animal welfare. AFMs thereby seek to change the industrial food system or create a new one. AFMs use a wide array of strategies, with most attention focused on creating alternatives to the industrial food system rather than using the state apparatus or transnational institutions to reform it. As countermovements, AFMs have generally arisen in places where the industrialization of agriculture and the food system has the longest history or in places experiencing rapid industrialization of the food system, although AFMs vary largely due to the positionalities of the actors creating them.

Since the early days of promoting alternative agriculture, many organizations – local, national,
and transnational – have come into existence. AFMs have expanded greatly from a focus on holistic production techniques and soil health to include paying farmers and farmworkers better, expanding food access and creating more just food systems, and increasing local or national control over the means of production. One process that has led to divergence within AFMs is codification of the movement’s values into formal state regulation (Guthman 2004) as this has led to critiques of “conventionalization” and new movements that seek to address issues which codification excluded.

Geographers and allied scholars have been central to a number of scholarly debates about AFMs. These debates include the extent to which alternative food truly offers an alternative, especially to neoliberal ideology and politics (Galt, Gray, and Hurley 2014; Guthman 2004; McClintock 2014; Pudup 2008); alternative foods’ accessibility both economically and culturally (Alkon 2012; Guthman 2008, 2011); and the wellbeing of farmers and farmworkers in these systems (Allen 2004; Brown and Getz 2008; Galt 2013; Jarosz 2008). There are also debates concerning specific AFMs, which show considerable diversity.

Holt-Giménez (2011) provides a typology for making sense of the diversity of AFMs. He categorizes them as reformist, progressive, and radical, each being different responses to the current neoliberal food regime. The reformist trend aims to change the industrial food system through consumer choice and through persuasion, as manifested in voluntary corporate sustainability and accountability standards. This includes the corporate mainstream faction of fair trade and organic food. The progressive trend focuses on citizen empowerment, especially through community food systems and food justice. The radical trend, increasingly framed by the concept of food sovereignty, focuses on the democratization of the food system in favor of the poor.

Organic agriculture

The roots of organic agriculture can be traced to Westerners rediscovering the wisdom and practices of peasant and indigenous producers as the industrialization of agriculture proceeded. The biological provisioning of crop nutrients and importance of soil health was a major early focus of pioneers. Rudolf Steiner in the penultimate year of his life, 1924, gave eight lectures laying out a vision or hints of an alternative agriculture that would “heal the earth.” This was arguably Europe’s first formal organic agriculture course and gave rise to what was later called “biodynamic agriculture” and “organic agriculture” (Paull 2011, 64). Around the same time, Sir Albert Howard, another progenitor of the organic agriculture movement, worked in India where he observed traditional farming and came to favor many of their practices over those promoted by Western agricultural science (Howard 1976/1940). Lady Eve Balfour (1944) played a foundational role in the early organic agriculture movement, establishing in 1939 the Haughley Experiment (the first long-term study of organic and conventional agriculture), and promoting organic agriculture for the postwar period by founding and leading the Soil Association, an international organization to promote sustainable agriculture, in 1946. In the United States, Jerome Rodale was an early advocate for organic agriculture, starting the magazine Organic Farming and Gardening in 1942 and the Rodale Institute, which promotes research on organic farming, in 1947. Another important landmark was Rachel Carson’s Silent Spring (1994/1962) which, in addition to helping catalyze 1960s environmentalism,
spurred public interest in organic agriculture and food.

According to the International Federation of Organic Agriculture Movements (2014), organic agriculture “is a production system that sustains the health of soils, ecosystems and people. It relies on ecological processes, biodiversity and cycles adapted to local conditions, rather than the use of inputs with adverse effects. Organic Agriculture combines tradition, innovation and science to benefit the shared environment and promote fair relationships and a good quality of life for all involved.” Organic food has moved from a local system of exchange between farmers and their direct consumers to oversight through organic farming organizations and eventually to institutionalization of oversight through national and international legislation which depend upon third-party certifiers to enforce the standards (Campbell and Liepins 2001; Guthman 2004). Currently in the industrialized world, most organic food is sold through mainstream channels. In the United States in 2010, mass market retailers sold 54% of organic food while natural food retailers sold 39% and the remaining 7% was mostly through direct marketing (Organic Trade Association 2014).

There has been considerable growth in certified organic farming and food in the last few decades. In the United States in 1990, organic food sales were worth $1 billion. In 2010 they were worth $26.7 billion, accounting for 4% of US food sales (Organic Trade Association 2014). Worldwide, organic sales reached $54.9 billion in 2009, and the countries with the highest per capita consumption are Switzerland, Denmark, and Luxembourg. In 2011 there were 1.8 million certified organic producers worldwide growing on 37.2 million hectares of certified agricultural land, up from 11 million ha in 1999. One-third of this land and more than 80% of producers are in developing countries, while more than 90% of consumption is in North America and Europe (Research Institute of Organic Agriculture (FiBL) 2014), which means the majority of certified organic production in many developing countries is for export. However, it is important to recognize that millions of producers in developing countries are “organic by default” – not certified, but using production practices reliant on biological/ecological processes rather than synthetic chemicals – since they cannot afford synthetic inputs or these inputs make little sense within subsistence-oriented production systems. It is in these kinds of production systems that implementing organic practices often results in considerable improvements in yields as well as other agroecosystem benefits (Hine, Pretty, and Twarog 2008).

Fair trade

Food-related fair trade as it exists today started in 1988, when an indigenous farmer cooperative in Oaxaca, Mexico, asked the Dutch development organization Solidaridad to increase its purchasing of coffee to give the coop access to mainstream European markets in volumes that would improve the small farmers’ incomes. Solidaridad created a label, Max Havelaar, which could be placed on any brand of coffee sold in any store, certifying that the farmers had received a premium price that was a “fair return.” This was when “alternative trade” became fair trade, with the focus shifting from handicrafts to food goods and into mainstream marketing channels (Jaffee 2007). The Max Havelaar label’s success led to the establishment of national fair trade groups in western European countries. Informed by dependency theory and related critiques, fair trade included a stable price for farmers and support for farming communities’ development through a fair trade premium. Sales
ALTERNATIVE FOOD MOVEMENT

Volumes increased rapidly in Western Europe. In 1997, the nationally based fair trade organizations created a worldwide umbrella certifier, Fairtrade Labelling Organizations International (in 2004 this organization split into Fairtrade International, which sets standards and supports producers, and FLO-CERT, which certifies producers and traders). In the United States, Equal Exchange, formed by activist entrepreneurs in 1986, established relationships with producer cooperatives in Latin America, Africa, and Asia, but it was not until 1999 that formal fair trade certification reached the United States through Transfair, which also coordinated efforts in Canada and Japan (Jaffee 2007).

The 1990s saw an expansion of fair trade sales and products covered by the certification. Fair trade cacao, bananas, tea, sugar, honey, rice, and orange juice appeared, mostly in European markets. Currently, 1.35 million farmers and workers participate in fair trade as covered by Fairtrade International, and there are 1149 producer groups involved. Fair trade sales recently expanded into new countries, including Kenya, Philippines, Argentina, and Lebanon, so that fair trade products are now sold in 125 countries, even though the bulk of sales remains in Europe (Fairtrade International 2013). In 2012, Fair Trade USA withdrew its membership from Fairtrade International, largely due to a disagreement about the types of producers that should qualify for fair trade. Fair Trade USA wanted to certify plantations in crops like coffee that had historically only been certified if production was organized through farmer cooperatives. Important questions remain about who should be allowed to set the terms of what is fair (Bacon 2010) and the inclusion of plantations (Besky 2013).

Civic agriculture and sustainable community food systems

A rapidly growing part of AFMs involves civic agriculture and sustainable community food systems. Civic agriculture refers to “community-based agriculture and food production activities that not only meet consumer demands for fresh, safe, and locally produced foods but create jobs, encourage entrepreneurship, and strengthen community identity” (Lyson 2004, 2). A strongly related term is a sustainable community food system, “a collaborative effort to build more locally based, self-reliant food economies – one in which sustainable food production, processing, distribution and consumption is integrated to enhance the economic, environmental and social health of a particular place” (Feenstra 2002, 100). In today’s parlance, these concepts are often referred to as “local food,” but this term can ignore embedded relationships when focusing only on distances between production and consumption.

Within these umbrella terms, there are a number of specific initiatives, from the globally connected – like Slow Food – to nationally and regionally linked – food policy councils, farm-to-school programs, and farm organizations representing smallholders – to the very locally specific – community gardens, farmers’ markets, community supported agriculture (CSA), and corner-store conversions and community markets. These variously scaled efforts are expanding rapidly, and new innovations within them are constantly being created (Venn et al. 2006). Since many of these efforts focus on community, environment, and connectedness, they often maintain more of the holistic values of the early organic agriculture visionaries than organic certification. Yet, there remain important questions about whether these alternatives allow for farmers to maintain economic viability
and farmworkers to be adequately compensated within a broader capitalist political economy (Galt 2013).

**Food justice**

Some of the most innovative initiatives to come out of building civic agriculture and sustainable community food systems belong to the food justice movement. The food justice movement “combines an analysis of racial and economic injustice with practical support for environmentally sustainable alternatives that can provide economic empowerment and access to environmental benefits in marginalized communities” (Alkon and Agyeman 2011, 6). The movement is informed by critiques of the ways that people of color and underserved communities experience the inequalities of the food system. This often manifests as a grassroots-led transition to a more equitable and sustainable food system, often within regions that have experienced the most “demarcated devaluation” occurring from the disinvestment of private capital and state support from urban areas through practices of institutional racism, such as redlining and deed covenants (McClintock 2011).

Food justice movement activists and organizations typically use the tools of civic agriculture and sustainable community food systems combined with antiracist practices. Thus, communities and organizations in Milwaukee, Wisconsin, Detroit, Michigan, and the East Bay Area of California, have been at the forefront of many food justice innovations, creating new sustainable production systems and forms of exchange serving underserved communities, providing education, and spurring greater involvement in the food system. These efforts have created higher levels of community autonomy and empowerment for groups and individuals involved (Bradley and Galt 2014; White 2011). However, there are critiques about food justice efforts that arise from the “colonial impulse” of white, upper- and middle-class activists who want to “bring good food to others” (Guthman 2008). Others have expressed concerns about food justice work underpinning the retreat of the state (McClintock 2014) and food justice organizations’ heavy reliance on markets (Alkon 2012).

**Food sovereignty**

Food sovereignty is a rapidly growing international movement that seeks to address the root causes of inequality and hunger in the world. The main goal is “democratization of the food system in favor of the poor and underserved” (Holt-Giménez 2011, 324). As envisioned by La Vía Campesina (1996), “Food sovereignty is the right of each nation to maintain and develop its own capacity to produce its basic foods respecting cultural and productive diversity.” As a movement, it stands in opposition to neoliberalism (especially the inclusion of agriculture and food in World Trade Organization (WTO) agreements) and the ever-consolidating control of large-scale agribusiness and food industry firms. In doing so, it pursues self-determination for farm workers, peasants, and family farmers by protecting their livelihoods, worldviews, and access to productive resources.

Over the last two decades, the concept has become widespread in its use, being invoked and operationalized by various groups around the world, including Brazil’s Landless Workers’ Movement (Movimento dos Trabalhadores Sem Terra, MST) and the Basque people of northern Spain and southwestern France (Masioli and Nicholson 2010). In these circumstances, food sovereignty as a concept helps groups to give coherence to historical struggles. Yet, there
remain considerable and productive debates about the scales at which it is to be implemented and how various tensions between regionally autonomous self-provisioning systems and embedded trade might best be worked out (Gibson-Graham 2006). As an example of resocializing the economy to make it serve the needs of people and the planet, the food sovereignty movement stands at the forefront (Patel 2010). As such, food sovereignty may act as a foundation for uniting AFMs since it explicitly critiques the power relations shaping the world’s food system.

SEE ALSO: Agriculture; Consumption; Environment and consumption; Environmental certification and eco-labeling; Environmentalism; Global commodity/value chains; Local embeddedness; Localization/delocalization; Neoliberalism; Race and racism; Social justice; Social movements

References


The long-standing association between anarchism and geography can be traced across the historical landscape from the towering peaks of heightened association to the low valleys of disconnection and ambivalence. Yet if Earth writing is to be understood as “a means of dissipating … prejudices and of creating other feelings more worthy of humanity” (Kropotkin 1978/1885, 7), then it seems obvious that anarchism has much to contribute to the discipline of geography. Geographical writings from influential anarchist philosophers such as Peter Kropotkin and Élisée Reclus blossomed during the late nineteenth century, when their work contributed much to the intellectual climate of the time. Following their deaths in the early twentieth century, engagement with their work started to fade, yet the lasting impact of these visionary thinkers continues to be felt within contemporary geographical theory, influencing the ways geographers think about diverse topics from ethnicity and “race” to social organization and capital accumulation, to urban and regional planning, to environmentalism and, perhaps surprisingly, even anticipating some of the key precepts of the recent “more-than-human” turn. As realpolitik and the quantitative revolution took hold of geography during the war years of the early twentieth century, the anti-authoritarian vision of Reclus and Kropotkin seemed to be pushed beyond the bounds of what were considered to be geographical concerns. Yet, as geographers rediscovered their bearings for social justice in the early 1970s, anarchism came back into the disciplinary view and was afforded serious consideration by academics advocating for what has since become known as “radical geography.” The publication of Antipode announced a new ethic for human geography, one that refused the stochastic models, inferential statistics, and econometrics that dominated geographical proceedings at the time, subverting this trajectory with qualitative approaches that placed the lived experiences of research participants at the center of its methodological focus. Anarchism played a key role in formulating this epistemological critique, where early engagements took inspiration from Kropotkin in arguing that radical geography should adopt his anarcho-communism as its point of departure.

The publication of a special issue of Antipode on anarchism in 1978 demonstrated the ongoing influence of anarchist thought and practice on geography, as well as geography’s influence on anarchism. It was not just Kropotkin’s sociospatial contributions to human liberation that were celebrated in the issue, as Reclus also received accolades for the importance of his geographical vision for freedom. A reprinting of Kropotkin’s (1978/1885) essay “What Geography Ought To Be” was meant to further demonstrate the enduring relevance of his work, while Murray Bookchin’s (1978/1965) “Ecology and Revolutionary Thought” was also reprinted, showing how the anarchism of supposed nongeographers had a significant bearing on the radical geographical thought that was beginning to make itself known. Around the same time, the newsletter of the Union of Socialist Geographers published a themed section on anarchist
ANARCHIST GEOGRAPHY

discussions that took place at the University of Minnesota in 1976. These developments were indicative of a sense of optimism for anarchist ideas to reinvigorate a collective geographical practice that was increasingly turning its attention toward social justice. Yet, as the neoliberalism of the 1980s and 1990s began to take hold of the world’s political economic compass, anarchist engagements by geographers dwindled and were largely overshadowed by Marxist, feminist, and incipient poststructuralist critiques. Nonetheless, the decade of Reaganesque and Thatcherism did see the publication of Bookchin’s (2005/1982) *The Ecology of Freedom*, wherein he advanced an anarchist critique of nature’s domination by social hierarchy. The beginnings of some introspective reflection on geography’s colonial past and its enduring state-centricity also emerged at that time, where the foundational works of Halford Mackinder, Ellen Churchill Semple, Ellsworth Huntington, Isaiah Bowman, and Thomas Holdich were taken to task by anarchist geographers who drew on Kropotkin and Reclus in calling for the abandonment of our inherited disciplinary prejudices. The 1990s fared slightly better, where a special issue on anarchism was organized for the short-lived journal *Contemporary Issues in Geography and Education*, while *Antipode* continued to publish the work of geographers who developed new, anarchist-inspired theories related to counter-hegemonic struggle and resistance to capitalism. More recently a new generation of geographers has begun actively transgressing the frontiers of geography by situating anarchism at the center of their practices, theories, pedagogies, and methodologies, (re)mapping the possibilities of what anarchist perspectives might yet contribute to the discipline (Springer 2013). This anarchist (re)turn comes as capitalism’s house of cards begins to collapse under its own weight, where intensifying neoliberalization, deepening financial crisis, and the ensuing revolt push anarchist praxis back into widespread currency both inside and outside of the academy.

From the vantage point of the present, it is important to recognize that the reduction in direct engagements with anarchism among academic geographers since the time of Reclus and Kropotkin in no way signals the decline of anarchism as a relevant political idea, something that is now actively being rediscovered by the new batch of anarchist geographers. Instead, it speaks to one of the core tenets of anarchist praxis, centered as it is on the politics of prefiguration, where anarchism lives through the organization and creation of social relationships that strive to reflect the future society being sought. Prefigurative politics is the recognition that to plan without practice is akin to theory without empirics, history without voices, and geography without context. In other words, prefiguration actively creates a new society in the shell of the old. So, while academic geography became obsessed with the trappings of positivism, and later the class-centric economism of Marxism, the geography of anarchism simply left the academy for the greener pastures of practice: on the streets as direct action, civil disobedience, and black bloc tactics; in the communes and intentional communities of the cooperative movement; amid activists and a range of small-scale mutual aid groups, networks, and initiatives; as tenant associations, trade unions, and credit unions; online through peer-to-peer file sharing, open source software, and wikis; among neighborhoods as autonomous migrant support networks and radical social centers; and, more generally, within the here and now of everyday life. In some ways what we are witnessing today, an even deeper appreciation for anarchism than we’ve ever actually seen within the academy, is the result of a century of struggle. Reclus
and Kropotkin were not able to combine their anarchism with their geographical scholarship as they might do today, but not necessarily for lack of trying. Kropotkin was offered an endowed chair at Cambridge University, but turned it down because it came with the stipulation that he give up his political commitments. Nonetheless, the closer we move toward the present moment, the more the literature demonstrates an appreciation for praxis, where the result has been a burgeoning consideration of both sides of the theory/practice divide.

Although anarchism is frequently portrayed as a symptom of mental illness and a synonym for violence and chaos, rather than as a valid political philosophy, such sensationalism is a ploy by its detractors. While violence has informed some historical and contemporary anarchist movements, and it is difficult to deny this constituent, anarchism has no monopoly of violence; compared to other political creeds (e.g., nationalism or monarchism), anarchism is decidedly peaceful. The word “anarchy” comes from the Greek ἀναρχία, meaning “without rule,” or against all forms of “archy” or systems of rule (i.e., patriarchy, oligarchy, monarchy, hierarchy, etc.). Violence can be seen as antithetical to anarchya precisely because all violence involves a form of domination, authority, or rule over other individuals. Violence is thus a particular form of “archy,” and not anarchya at all. Similarly, anarchism refuses chaos by creating new forms of organization that break with hierarchy and embrace egalitarianism. In fact, the symbol for anarchya is meant to suggest that anarchya is the mother of order, an idea advanced by Pierre-Joseph Proudhon, the first person to identify as an anarchist. Anarchism accordingly represents an unwavering political commitment that seeks to break from hierarchical structures and unfasten the bonds that facilitate and reproduce violence. It is the notion that our shared morality should not be premised on the prejudices of life as it is currently lived through a politics of consensus that is antagonistic toward difference, but rather on a version of empathy that embraces our ultimate integrality to each other and to all that is. Such a process entails the rejection of all the interlocking systems of domination, including capitalism, imperialism, colonialism, neoliberalism, militarism, classism, racism, nationalism, ethnocentrism, sexism, Orientalism, ableism, genderism, ageism, speciesism, homophobia, transphobia, organized religion, and, of course, the state. Simon Springer (2012, 1607) has accordingly defined anarchist geographies as “kaleidoscopic spatialities that allow for multiple, nonhierarchical, and protean connections between autonomous entities, wherein solidarities, bonds, and affinities are voluntarily assembled in opposition to and free from the presence of sovereign violence, predetermined norms, and assigned categories of belonging.”

Such a holistic interpretation of anarchist geographies was first laid down by Reclus (1876–1894), whose primary contribution to the discipline was the emancipatory vision detailed in The Earth and Its Inhabitants: The Universal Geography, wherein he conceptualized a coalescence between humanity and the Earth itself. Reclus sought to eliminate all forms of domination, which were to be replaced with love and active compassion between all animals, both human and nonhuman, as a process of humanity discovering deeper emotional meaning through acknowledging itself as but one historical being in the flowering of a greater planetary consciousness. Kropotkin (2008/1902) did much to contribute to such a vision as well with his monumental Mutual Aid: A Factor of Evolution, where in partial reply to the social Darwinism of his time, he observed mutual aid as cooperation among plants, animals, and humans, including mutual forms of assistance between
ANARCHIST GEOGRAPHY

species, thereby shedding light on a grander sense of agency, foreshadowing recent theorizing within the domain of more-than-human geographies. Anarchist ideas were from the outset explicitly geographical, differing greatly from the industrial imagination of Marxists, as emphasis was placed on decentralized organization, rural life, agriculture, and local production, which allowed for self-sufficiency and removed the ostensible need for central government. Anarchists were also rooted in a view of history that has been confirmed by the anthropological record, where, prior to recorded history, human societies established themselves without formal authority in ways that rejected coercive political institutions. Although early views of anarchism have been critiqued on the basis of their naturalist assumptions, we would also do well to pause and reflect on the implicit naturalizing of hierarchical structures that suggest that hierarchies necessarily arise as societies grow, rather than analyzing the patterns through which authority is actually constructed and thinking through the innumerable anarchist alternatives that could be and are being developed.

In understanding anarchist geographies, we should begin by noting that anarchism is not about drafting sociopolitical blueprints for the future, nor does it trace a line or provide a model. Prefiguration should not be confused as pre-determination, as anarchists are more concerned with identifying social tendencies, where the focus is on the possibilities that can be realized in the here and now. Anarchism accordingly points to a strategy of breaking the chains of coercion and exploitation by encompassing everyday acts of resistance and cooperation, where examples of viable anarchist alternatives are nearly infinite. The only limit to anarchist organizing is our imagination, and the sole existing criterion is that anarchism proceeds nonhierarchically. Such horizontal organization may come in the form of child-care collectives, street parties, gardening clinics, learning networks, flashmobs, community kitchens, free skools, rooftop occupations, freecycling, radical samba, sewing workshops, coordinated monkeywrenching, spontaneous disaster relief, infoshops, volunteer fire brigades, microradio, building coalitions, collective hacking, wildcat strikes, neighborhood tool sharing, tenant associations, workplace organizing, knitting collectives, and squatting, which are all anarchism in action, each with decidedly spatial implications. So what forms of action does anarchism take? “All forms,” Kropotkin once answered:

indeed, the most varied forms, dictated by circumstances, temperament, and the means at disposal. Sometimes tragic, sometimes humorous, but always daring; sometimes collective, sometimes purely individual, this policy of action will neglect none of the means at hand, no event of public life, in order to ... awaken courage and fane the spirit of revolt. (2005/1880, 39)

Anarchist organization doesn’t seek to replace top-down state mechanisms by standing in for them; rather, it replaces them with people building what they need for themselves, free from coercion or the imposition of authority. Rather than proceeding from a centralized polity, social organization is conceived through local voluntary groupings that maintain autonomy as a decentralized system of self-governed communes of all sizes and degrees that coordinates activities and networks for all possible purposes through free federation. The coercive pyramid of the state structure is replaced with webs of free association, in which individual localities freely pursue their own political economic and sociocultural arrangements.

Anarchist geographies are actually not novel, in the sense that people have organized themselves collectively and practiced mutual aid
to satisfy their own needs throughout human history. Organization under anarchism is simply a continuation of this impulse, despite its attempted disruption by the state. As Colin Ward argued:

given a common need, a collection of people will, by trial and error, by improvisation and experiment, evolve order out of the situation – this order being more durable and more closely related to their needs than any kind of order external authority could provide. (1982/1973, 28)

There is consequently no transgeohistorical narrative to anarchism as, although it has been continuously present in human societies, mutual aid is nonetheless differentiated across space and time, taking on unique and even subtle forms according to context, needs, desires, and constraints placed on reciprocity by opposing systems such as capitalism. At certain times and in particular places mutual aid has been central to social life, while at other times the geographies of mutual aid have remained largely hidden from view, overshadowed by domination, competition, and violence. Yet, irrespective of adversarial conditions, mutual aid remains prevalent, and

the moment we stop insisting on viewing all forms of action only by their function in reproducing larger, total, forms of inequality of power, we will also be able to see that anarchist social relations and non- alienated forms of action are all around us. (Graeber 2004, 76)

It is in the spirit of seeking new forms of organization that anarchist geographies have been “reanimated” as of late (Springer et al. 2012), emphasizing a do-it-yourself (DIY) ethos of autonomy, direct action, radical democracy, and noncommodification. Arguments in favor of the radical potential of DIY culture have emphasized anarchist perspectives toward the everyday transformation of our lives, a sentiment that factors heavily in a great number of social movements, where geographers have begun thinking through how impermanent spaces may arise in response to sociopolitical action that eludes the formal structures of hierarchical control. Pickerill and Chatterton (2006) have adopted such an “autonomous geographies” approach in attempting to think through how spectacular protest and everyday life may be productively combined to enable alternatives to capitalism. Routledge’s (2003) notion of “convergence space” has similarly proven influential to anarchists insofar as it appreciates how grassroots networks and activists come together through multiscalar political action to produce a relational ethics of struggle, offering a reconvened sense of nonhierarchical organization.

The application of an explicitly anarchogeographical perspective would benefit a range of contemporary issues, each with decidedly spatial implications, from the overt uprisings of the Arab Spring and the Occupy movement, to the spectacle of street theater and Critical Mass rides, to the subversive resistance of trespassing and culture jamming, to lifestyle choices of dumpster diving and unschooling, to the mutual aid activities of community gardens and housing co-ops, to the organizing capabilities of book fairs and Indymedia. Similarly, anarchism has much to contribute to enhancing geographical theory, where it is easy to envision how new research insights and agendas might productively arise from taking an anarchist approach to themes such as sovereignty and the state; homelessness and housing; environmental justice and sustainability; industrial restructuring and labor geographies; capital accumulation and property relations; policing and critical legal geographies; informal economies and livelihoods; urban design and aesthetics; agrarian transformation and landlessness; nonrepresentational theory and more-than-human geographies; activism and social justice; geographies of debt
and economic crisis; belonging and place-based politics; participation and community planning; biopolitics and governmentality; postcolonial and postdevelopment geographies; situated knowledges and alternative epistemologies; and anti-oppressive education and critical pedagogy. Kropotkin viewed teaching geography as an exercise in intellectual emancipation insofar as it afforded a means not only to awaken people to the harmonies of nature, but also to dissipate their nationalist and racist prejudices, a promise that geography still holds, and one that may be more fully realized should anarchist geographies be given the attention and care that is required for them to blossom. Retaining Reclus’s and Kropotkin’s skepticism for and challenges to the dominant ideologies of the day has much to offer contemporary geographical scholarship and its largely unreflexive acceptance of the civilizational, legal, and capitalist discourses that converge around the state. The perpetuation of the idea that human organization necessitates the formation of states is writ large in a discipline that has derided the “territorial trap,” yet has been generally hesitant to take the critique of state-centricity in the direction of anarchism. However, unlike the limited class-centricity of Marxian geography, the promise of anarchist geographies resides in their integrality, which refuses to assign priority to any one of the multiple dominating apparatuses, because all are seen as irreducible to one another. This means that no single struggle can wait on any other, and the a priori privilege of the workers, the vanguards, or any other category over any other should be rejected on the basis of its incipient hierarchy. Anarchism is quite simply the struggle against all forms of oppression and exploitation, a protean and multivariate process that is decidedly geographical. Anarchism is happening all about us.

**SEE ALSO:** Antiracist geography; Critical geography; Feminist geography; Nonrepresentational theory; Postcolonial geographies; Radical geography; Social justice; Social movements; State, the; Subaltern

**References**


Ancient geography

Kent Mathewson
Louisiana State University, USA

Ancient geography and its study was once a staple, if never quite the foundation, in geography’s scholarly genealogy from the Renaissance through the nineteenth century. Well into the Enlightenment period, the wisdom of the ancients – as received through Greek and Roman texts and techniques – was the touchstone upon which both geographic theory and practice were assayed and accepted, rejected, or superseded. The terms “ancient geography” or “classical geography” have been implicitly understood to refer to times and places bounded – the Aegean realm from the eighth century BCE to the disintegration of the vast Roman Empire a millennium and several centuries later. Within this historical and geographical expanse, the key figure – or, in most accounts, the only actors – were “the Greeks and Romans.” The origins of this strand within the larger Occidental or Western civilizational trajectory were generally traced back to the Homeric epics, with Odysseus as a heroic precursor and Homer “geography’s founder,” as Strabo dubbed him. Prior to this time, or so the standard narrative goes, any trace of geography’s firm foundations melted into myth while both moving back in time and moving out beyond the Greek oecumene, or known lands. Going forward it might be said that collectively the Greek geographers practiced Socrates’s injunction: “know thyself,” given their generally keen knowledge of the corpus of formal geographic work as it developed from the sixth century BCE up through the Hellenic period and into the first century CE with the Roman conquest and absorption. Like much else with Greek scholarly and scientific achievements, the Roman ascendancy brought little innovation, but some amplification in advancing geography in its elaboration and practice. And of course, the actual geography of the Greco-Roman world was extended greatly through Roman agency. In contrast to vague and mythical beginnings, the terminus has often been depicted as abrupt and vividly real, coincident with the Roman “collapse.” Geographical scholarship, at least in the European precincts of the former empire, was thought to have largely atrophied until slowly revived some centuries later through rediscovered Greek and Latin texts and maps, and Islamic scholarly influences. In its rough contours, this history is not all that off-center. But it is these assumptions of centrality, particularly the larger histories that it ignores – the “Other” antiquities if you will – that makes it an incomplete and therefore faulty telling. On the other hand, over the past century or so, the study of classical or ancient geography has been radically decentered, to the point of near extinction. As such, ancient or classical geography as a field of study offers opportunities for introducing new approaches and perspectives, while preserving the record and evidence that demonstrates important aspects of contemporary geography’s groundings can still be located in classical Antiquity.
ANCIENT GEOGRAPHY

Antiquity’s antecedents

Eratosthenes (ca. 276–195 BCE) is credited with being the first to use the term “geography” (Earth description) in the third century BCE; however, major figures of the Milesian school of pre-Socratic philosophy from the sixth century BCE such as Thales (ca. 624–546), Anaximander (ca. 610–546), and Hecataeus (ca. 550–476) made major contributions to the broad body of knowledge and systematic discoveries that came to be called geography. Rather than seeing their accomplishments as ancient geography’s basement or beginnings, it may be more accurate to view this center of creativity in Asia Minor as one culmination or maturation of a deep and ongoing process in which proto-geographic thought and praxis had been in formation for tens of millennia in multiple sites and contexts. The immediate proximal sites and contexts that extend Milesian or Greek antecedents back into this history (only selectively recoverable) lie to the south in Egypt and east in Mesopotamia and perhaps beyond. Specifying precocity is problematic and periodically undercut by the shifting sands of cultural historical excavation and recovery. Nevertheless, by the fourth millennium BCE, Sumerian civilization, and laterally Egyptian civilization, was developing writing systems, complex mathematics, astronomy, cadastral mapping and instrumentation, and methods of data collection for reasons of state. The earliest known map associated with a civilization is on a Babylonian clay tablet. By the third millennium BCE geographical techniques and knowledge were passing between societies and civilizations from Egypt to the Indus Valley and perhaps beyond.

Martin Bernal, in his sharply revisionist Black Athena: The Afroasiatic Roots of Classical Civilization (1987) convincingly argues that not only much of what is attributed to Greek creativity and precedence has demonstrable provenance in these Afro-Asiatic civilizations, but these influences and diffusions began earlier than commonly thought. Bernal pushes the inceptions back to the onset of the second millennium BCE, suggesting the influences were continuous and substantial throughout the second millennium BCE and well into the first. Bernal also argues that the ancient Greeks themselves were well aware of these antecedents, which he calls the “Ancient Model,” but that nineteenth century Romantic and racist scholars, enamored with the vision of a largely self-generating classical Greek civilisation, sporadically invigorated by Indo-European barbarian invasions from the north, constructed what he calls the “Aryan Model” of Greek civilizational origins. This model, foundational to conventional Western civilizational views and studies from the early nineteenth century on, seriously distorts the actual history and geography of not only the “rise” of Greek civilization, but by extension the history of geography’s early development as an art and a science. Similarly, to limit the bounds of ancient geography to only a Greco-Roman perimeter, diminishes it tremendously. Each ancient civilization can be found to have its own expressions of geographical knowledge and practice.

While Afro-Asiatic civilizations of northeast Africa and southwest Asia may well have the best-recognized pre-Greek records of geographical art and science, a full accounting of Chinese innovations and accomplishments, along with official patronage of geographical scholarship, would likely match much of the proto-geographical activity on the other side of the Asian continent and its borderlands. Nor can one rule out some interaction, diffusion, or transfer of knowledge and techniques between the eastern Mediterranean/Near Eastern lands and East Asia. There is some evidence of this toward the end of the Greco-Roman classical age, but without major import. Joseph
Needham (1959) devotes an entire volume (3) of his monumental *Science and Civilization in China* to Chinese accomplishments in mathematics, and the sciences of the heavens and the Earth. He emphasizes parallels with the West, but also points out singular Chinese achievements. One striking difference is the multi-millennial unfolding of continuities in the Chinese civilizational trajectory, rendering the categories “classical” and “ancient” somewhat more problematic to fix or bound. The opposite is true of civilizations and their geographic arts and sciences outside of the Mediterranean/Near Eastern or East Asian domains. As with Rome, most have discernable termini, usually by collapse or conquest. The Harappan civilization of the Indus River Valley and the Maya civilization of Mesoamerica both experienced civilizational simplification if not outright cessation, usually thought to be from a concatenation of causes, variously diagnosed as environmental, epidemiological, political, economic, and so on. In florescence, both produced geographic knowledge and techniques comparable in some instances with those of the larger Old World civilizational centers. Other ancient civilizational trajectories were brought to abrupt and violent ends, most notably, the New World civilizations such as the Aztec in Mesoamerica and the Inca in South America. Each had sophisticated ethnoscientific traditions and practices that generated geographical knowledge and practical works, including remaking whole landscapes for astronomical-geographical research and politico-religious ideological articulation. Perhaps ironically, their geographical enterprises, along with much of the rest of their civilizational apparatus, were in part destroyed through effective Iberian marshaling of the rapid advances in Europe’s geographic knowledge and techniques in the late Middle Ages. And some of the impetus came from the “rediscovery” of the geographic wisdom of “antiquity.”

**Even earlier geographers**

There is an increasing recognition among scholars and writers delving into the pre-civilizational past that developing and utilizing geographical knowledge and skills are a part of what makes us human. Geographer Robert Sack’s (1997, 1) *Homo geographicus* is fundamentally a “geographical being,” who transforms the Earth, and in turn is transformed. For Sack, “being geographical is inescapable” whether we are conscious of it or not. Writers such as Bruce Chatwin (1986), with his lyrical account of the Australian Aborigines’ world manifesting song lines that aid navigation through both space and time, have opened windows onto what now appear to be universal traits expressed as geographical competencies. Geographer James Blaut and environmental psychologist David Stea (Blaut et al. 2003) have argued that mapping behavior is a cultural and cognitive universal, that it has been evident since at least the Upper Paleolithic, and that it is has been a key adaptation in human evolution akin to tool use and language acquisition, aiding humans in comprehending and acting effectively in the geographical environment. Earlier work in geography and numerous studies in anthropology and archaeology have supported this proposition. Franz Boas (1888), founding figure in anthropology, worked initially in geography, studying the cognitive mapping abilities of the “Eskimo” (Inuit) of Baffin Land. Boas found the Inuit to have accurate locational knowledge of their macro-environment, all the more impressive given what to a non-native seemed often to be featureless landscapes. The case for pre- or non-civilizational geography and geographers, much of it “ancient,” albeit outside of formal institutions and expression, would seem to be obvious. As with the civilizations apart from the Egyptian/Near Eastern/Greco-Roman core, the vast vernacular geographic lore and practices
ANCIENT GEOGRAPHY

of peoples and cultures preceding them should be included in any consideration of ancient geography.

Ancient Greek geography

While enlarging the scope of ancient geography to admit the other ancient civilization’s geographical accomplishments, along with the geographical knowledge and activity generated by non-civilizational peoples everywhere, clearly the ancient Greeks produced the largest and most advanced corpus of geographical work. Locating these origins in Homeric times may have struck cords with later Greeks’ own mythmaking – and the same for nineteenth century historians of geography, but as already mentioned, the base seems to have been laid in Asia Minor at centers such as Miletus in coastal Ionia. Here commercial circuits threading throughout the “seven seas” carried commodities, but also transmitted basic geographic information. The Milesian school of philosophers Thales, Anaximander, and Hecataeus not only contemplated the cosmos, they proposed methods for its investigation. Thales and his student Anaximander have been credited with initiating what can be called the “mathematical” tradition in Greek geography. Thales was concerned with the measurement and location of places on the Earth’s surface. He was also involved in commerce, and traveled widely in the eastern Mediterranean. From the Egyptians he learned principles of geometry that he cast as propositions, and then applied to geodetic questions. Conceptualizing the Earth as a whole, he envisioned it to be a disc floating on water, with water being the primordial substance. Thales also made contributions to astronomy. Perhaps his greatest contribution was moving Greek geography beyond explanations based in myth and demonstrating the practicality of mathematical approaches. Thales’s student Anaximander followed in the mathematical tradition, introducing the Babylonian gnomon or sundial to Greek science. The gnomon allowed for determining diurnal time (at noon the shadow is shortest) but also seasonal shifts – the noon shadow is shortest at the summer solstice and longest at the winter solstice. Anaximander also seems to be the first to have produced a mappamundi or map of the world. Greece was at the center, flanked by the three continents – Europe, Asia, and Libya (Africa) and surrounded by ocean. With this concept and its depiction, Anaximander can also be credited with initiating the cartographic or mapping science tradition in Greek geography. Hecataeus, the third Milesian geographer, pioneered the historiographic or descriptive tradition. He compiled the information that flowed into Miletus and complemented it with his own travel observations of Greece, Egypt, Asia Minor, and the Black Sea Basin. From this he wrote periodos ges (circuit of the Earth) or periêgês. It was divided into two books, one for Europe and one Asia, and provided information on coastal locations including mythology, ethnography, travel directions, and topographic features.

The center of classical geographic thinking and production moved westward from fifth century BCE on. Herodotus of Halicarnassus (ca. 484–428 BCE) is usually credited with conceiving and producing the first coherent historiographic work, hence his sobriquet “patrem historiae” (“father of history”). He was particularly concerned to chronicle the uniting of various Greek polities in the face of the Persian imperial threat. Rather than simply focusing on the empirical military and political dimensions of Greek warfare as Thucydides (ca. 460–395 BCE) did, Herodotus provided extensive geographical and ethnographic detail
of both the Greek homelands as well as the Persian Empire. Herodotus traveled widely in the eastern Mediterranean world, and drew on his own observations from Egypt, the Euxine (Black Sea) basin, and much of the Persian Empire in writing his histories. He can also be considered the progenitor of historically informed human geography and ethnography. Herodotus had a lively interest in physical geographic processes and offered explanations of coastal shore and delta formation, relations between temperature regimes and wind directions, and the seasonal Nile flooding. He was perhaps the first to employ the concept of the oikoumenê (inhabited world) and suggested lands beyond. He argued for a world island (Europe, Asia, Libya (Africa)) surrounded by ocean. He based this in part on a putative Phoenician expedition (ca. 600 BCE) that had circumnavigated Africa traveling down the east coast, rounding the Cape of Good Hope, and returning through the Pillars of Hercules (Gibraltar). Herodotus’s historico-geographical work was complemented by other fifth century Greek historians. Xenophon’s (ca. 430–354) historical works, particularly Anabasis, his personal chronicle of the heroic March of the Ten Thousand (Greek mercenaries) who fought their way from the heart of the Persian Empire back to Greece, provides ample and usually accurate geographical detail. Other Greeks served Persian rulers in various capacities, often commenting on the geography of the empire.

By the fourth century BCE, the inclusion of geographic details was a common feature of historiographic practice, and geographic questions had entered the philosophical arena in significant ways. Ephorus of Cyme (ca. 400–330 BCE) produced a 30 volume universal history. He is credited with separating out the strictly geographical data and placing it in two separate books – one on Europe and the other on Asia. Both Plato (428–348 BCE) and Aristotle (384–322 BCE) can be credited with contributing to ancient geography directly and, in the case of Aristotle, providing the philosophical-paradigmatic basis on which much of the geographic discourse and discovery were mediated for almost the next two millennia. Plato’s contributions lie in several areas. Epistemologically he was an idealist and favored the deductive approach. Ontologically he was a declensionist who saw change in nature and culture as necessarily a degeneration from an original perfect state. Plato was perceptive in reading the landscape record of Attica as one in which soils had been impoverished, forests destroyed, and hydrologic systems disrupted, but not in assuming that things might be different with different modes of human agency. Plato’s postulating the lost land mass of Atlantis (larger than Europe and Asia combined) fostered the idea of terrestrial impermanence and the notion of islands that disappeared, one of ancient geography’s ideas that has provoked geographical imaginations down to present times. It also lent support to the view that the Earth’s surface was immense and the eastern Mediterranean realm was just one small portion of it, albeit centrally important. At the same time Plato deduced that the Earth must be a sphere, as this shape was the most perfect, and at the center of the universe. Eudoxus of Cnidus (ca. 408–355 BCE), Plato’s student and colleague in his academy, developed the theory of klima or zones of climate by deducing that increasing slope on a sphere (the Earth) would put higher latitude zones at further distance from the sun.

While Aristotle (384–322) was Plato’s most accomplished student, his philosophical positions were the obverse of Plato’s. He was more materialist than idealist and advocated inductive rather than deductive reasoning and approaches. Ontologically he envisioned
ANCIENT GEOGRAPHY

teleological amelioration as the natural state of things. His contributions to geography were multiple. He amended Empedocles’s fourfold division of primordial substances – earth, water, fire, and air – to add a fifth, aether, which comprised the celestial bodies. Aristotle went beyond Plato’s deduction of the Earth’s sphericity to offer inductive evidence – the circular edge of the shadow produced in a lunar eclipse, and the ascent of stars above the horizon as one travels north. Aristotle refined the concept of latitudinal zones, but planted the erroneous notion that the torrid or equatorial zone was too hot for human habitation. Like Herodotus, Aristotle drew on personal travel experience to formulate some of his theories and observations. Aristotle’s student Alexander the Great carried out this advocacy of first-hand observation to heroic levels. In part, Alexander’s imperial impetus was fueled by Aristotle’s admonition to observe as well as expand the Greek domain to the limits of the ekumene. Commercial explorers of the time included Pytheas of Massalia (ca. 380–ca. 300 BCE) who is credited with many “firsts” in the annals of ancient geography: observing polar ice, venturing to the lands of the far north or Thule (Shetlands? Faroes? Iceland? Greenland? Svalbard?), “discovering” the Baltic Sea, explaining tides as the result of lunar attraction.

The third century BCE produced the epitome of Greek geography in the persona of Eratosthenes of Cyrene (ca. 276–ca. 194 BCE). He coined the word “geography” and laid bases for its articulations. He served for forty years as head librarian of the Library of Alexandria, seat of ancient learning in Hellenistic times. With the library’s resources at hand, he produced three major works: one on the form and nature of the Earth and its processes of changes, a second book on mathematics and measurement, and a third book on the nations, peoples, and politics of the ekumene. He produced a mappamundi with Alexandria as the prime meridian. He added mathematical boundaries to Aristotle’s fivefold zonation: torrid, temperate, frigid. The torrid zone was 48 degrees straddling the equator, and the frigid zones extended 24 degrees from each pole, thus only five degrees off current delimitations. Perhaps his greatest feat, however, was estimating the circumference of the Earth within 140 miles of its actual dimensions. The culmination of Greek geography both figuratively and literally was embodied in the work of Strabo (ca. 62 BCE–20 CE). Born in Asia Minor, he traveled widely in the eastern Mediterranean, living for a time in Rome and Alexandria. Strabo’s encyclopedic Geography comprised some 17 books, largely drawing on the work of his predecessors. He also passed judgment on their findings and observations, often coming to erroneous conclusions: the veracity of Homer over Herodotus, doubting Hanno’s and Pytheas’s Atlantic voyages, setting very narrow limits for human habitability (roughly between 55 and 12 degrees north). Unlike most of the earlier Greek geographers’ writings, the majority of Strabo’s work survived, and served as the main textual source of Greek geographic knowledge from the end of Antiquity through the Middle Ages.

Roman geography serves as something of a coda to classical geography. The most illustrious figure is Claudius Ptolemy (ca. 90–168 CE) who, although a Roman citizen living in Alexandria, wrote in Greek and was most likely ethnically Greek as well. Based in the library of Alexandria, Ptolemy produced three great works in astronomy, geography, and astrology that continued to be influential for the next millennium or more. His eight-volume Geographia was primarily a gazetteer with locational tables (latitude and longitude) but the first volume dealt with astronomical data and the final was devoted to maps of various parts of the world. From the gazetteer data he also produced a world
map that remained the standard for centuries. Geographic research and knowledge production in the Roman realm was largely additive rather than innovative. Those geographical pursuits that aided commerce and imperial expansion were attended to, but theorizing, research, or even compendious writings were not much in evidence.

A conspicuous exception is Pliny the Elder (ca. 23–79 CE). Pliny’s massive *Natural History* (in 37 books) is not specifically geographical, but much of the reportage is place based and books 3–6 offer a survey of regions and places within the ekumene. The Romans did energetically produce *periploi*, or travel guides, for both land and sea. Under Roman authorship, paradoxography, the literary genre that drew heavily on geographical reportage to describe natural oddities and monstrosities, usually to be found at the edges of the ekumene, flourished, enlisting figures such as Varro and Cicero. The fascination with *mirabilia* or things marvelous and incredible within geographical contexts was one of the key legacies and links between classical Antiquity and geography’s renovation in the late Middle Ages. At the same time the more practical and prosaic pursuits of ancient geographers, whether expressed in geographical description, mathematical methodology, or cartographic exposition, provided the foundations for scholarly geography’s rebirth in fits and starts over the millennium that followed its effective dissolution coeval with the “collapse” of the Roman Empire in the sixth century CE.

**SEE ALSO:** Orientalism/Occidentalism

---

**References**


---

**Further reading**


Affect

Ben Anderson

Durham University, UK

The term “affect” has been used by geographers and others to attune to a range of intensities and atmospheres felt through bodies: to give some examples, hostility, as racism intensifies in an encounter between racially marked bodies; the surprising force of a sexually charged glance; comfort as bodies hold together; the feeling of participation in dance; “cruel optimism” (Berlant 2011), as people are almost sustained by promises that bring them harm. While theories of affect vary, use of the term “affect” signals an attention to lived experience. Work with an interest in affect shares, then, a promise and an imperative (Anderson 2014, 9–11). The promise is of a geography that senses and discerns the subtle, elusive dynamics of everyday living and touches on how abstract sociospatial processes are felt and lived. The imperative is to understand how forms of power work through affect. For some, this discovery involves diagnosing the contemporary condition as one in which affect has become a new “object target” as forms of power emerge and change. For others, it involves (re)thinking the (geo)political as an affectively imbued scene of attachments, events, and forces.

There is no single definition of affect in geography or other disciplines, just as there is no single definition for terms such as “emotion,” “feeling,” or “mood.” Interest in affect is typically associated with the emergence of nonrepresentational theories (Thrift 2004; McCormack 2014), a body of thought that is itself plural. But work on affect owes much to a number of precursors in geography and elsewhere. In addition to humanist phenomenology, research on affect is indebted to feminist scholarship that showed how emotions matter and disrupted the marginalizing or silencing of emotional experience, traced how emotions “get into” sociospatial relations (and vice versa), and experimented with new forms for representing lived experience (e.g., Valentine 1989). Nonrepresentational theories of affect share with work on emotional geographies the insights that affects express relations, and that the rational and the emotional/affective are enfolded in a complex way. From these starting points, work on affect asks a series of questions: How are affects emergent from, and part of, encounters? How, given the contingency of sociospatial orders, is affective life mediated and organized? And how, if at all, does affective life exceed, reproduce, or otherwise relate to systems of signification or representation?

These questions are refracted through different theories of affect. Thrift (2004) identifies four theories that offer different versions of what affect is: phenomenological approaches that describe the bodily processes and states through which affects occur in everyday life; psychoanalytic theories where affects are vehicles or manifestations for the drives; naturalist accounts, often deploying the brain sciences, that treat affects as biologically rooted universals; a Spinozan–Deleuzian notion of affect as always emergent capacities to affect and be affected. Despite a nascent resurgence of interest in psychoanalysis in geography, and the continued importance of phenomenological approaches, it is Spinozan–Deleuzian theories that have been central to the interest in affect. There affect refers principally to a body’s...
or bodies’ “capacity to affect and be affected.” A body’s “charge of affect” is two-sided: that is, what a body can do and does do are a matter of intensive “capacities to affect” and “capacities to be affected.” Normally this use involves invoking an analytic distinction between affect and emotion, where “emotion” is used to refer to the ways in which affects are named, interpreted, and reflected and, by contrast, “affect” refers to the intensity of experience, a quality that provides something close to the background sense of an event or practice or space. This is not to say that affect is cleaved from emotion. In what has become a well-cited passage, Massumi describes their enfolding:

An emotion is a subjective content, the sociolinguistic fixing of the quality of an experience which is from that point onward defined as personal. Emotion is qualified intensity, the conventional, consensual point of insertion of intensity into semantically and semiotically formed progressions, into narrativizable action–reaction circuits, into function and meaning. (2002, 28)

For now, we should note that, instead of affect and emotion existing in separate unconnected levels, Massumi and others stress multiple relations between emotion as a “subjective content” and affect as bodily capacities. Emotion as qualified personal content feeds back into the emergence and organization of affect, while at the same time being the most intense expression of the capture of affect and of affect’s ongoing escape (Massumi 2002, 35).

This treatment of affect and emotion has been criticized for reproducing an objective/subjective dualism and downplaying the subjective. It has also been critiqued on the basis that it willfully or naively cleaves the nonrepresentational, in the guise of affect, from the representational, and in so doing ignores, forgets, or obscures how what a body can do is determined. Such critiques turn on the question of how exactly affect can be said to be nonrepresentational and how affects connect to other processes and forces. In conclusion, affect can be said to be nonrepresentational in three ways, each of which responds slightly differently to the imperative to think about affect politically, and each of which involves different assumptions about affect.

First, affect is nonrepresentational in the sense of being never fully accessible to conscious subjects, despite being a condition for subjectivity. Before conscious thought, affects are expressed patterns of autonomic bodily reactions. These may be shaped by past encounters and may be subject to willed intervention. This approach lends itself to an emphasis on how “affect itself” is actively shaped, with work diagnosing how new forms of power work through affective spaces.

Second, and slightly differently, affect is nonrepresentational in that it names, or orientates attention to, the generative immediacy of an always overflowing life. Affects are events: they happen; they exceed their causes. Representations always fail in relation to affect; affective life exceeds fixity and containment. Associated with the invention and experimentation of new styles of research and writing (McCormack 2014), affective life is, on this understanding, conditioned but never fully determined by existing forms of sociospatial organization. The result is a politics of witnessing and cultivating events that attunes to the hesitant, cramped formation of other ways of being and living as life happens.

Finally, but closely connected, affect is understood as nonrepresentational, in that affects are formed in the midst of encounters and are mediated in ways that are irreducible to “representational–referential” systems of signification. Emphasis is placed on how affective life is perpetually becoming organized, whether through apparatuses that take affect as an object target, through encounters that enfold elsewhere.
and elsewhens, or through affective conditions (Anderson 2014). The politics of affect that accompanies this approach is, in many respects, a well-recognized one that describes and critiques the ongoing work of mediation and affirms that contingency of seemingly settled sociospatial formations.

What the three ways in which affect might be said to be nonrepresentational all offer is an understanding of how experience is organized as it happens. They differ on a question that will remain central to work on affect in geography: How are affects formed in the midst of, and become part of, sociospatial formations?

SEE ALSO: Emotional geographies; Emotional labor; Nonrepresentational theory

References

Animal geographies

Henry Buller
University of Exeter, UK

Wild animals

Between the “human” of human geography and the materiality of physical geography’s subject matter, animals – or nonhuman animals – have been largely absent from the canon of contemporary geographical thought. Only very recently has “the animal” begun to reappear in today’s human geography. In part this was a counterbalance to the hubris of the so-called cultural turn, when representations (textual, pictorial, cinematic, etc.) were a key research focus. Animals have become the newly liberated others, now noticed by human geographers who previously took no account of nonhuman presence, autonomy, difference, and multiplicity, in form or in meaning. Human geography’s newfound posthumanism has also been an important factor here, with its metaphysical, relational, ethical, and political implications for nonhuman animals and our interactions with them. In all this, animal geography is something of an epistemological menagerie, drawing on a range of subdisciplinary trajectories and influences.

For all that, animals were very much present during the early days of an emergent “modern” nineteenth-century geography. The explorer and naturalist Alexander von Humboldt (1769–1859), often described as one of the “fathers” of modern geography, sought in his writings and in his explorations to link the spatial distribution and interconnectedness of the world’s animals (and plants) to that of its inanimate landforms in a sort of grand unified bio-geographical theory of life on Earth.

Displaying a geographer’s understanding of the interrelationship of nature and culture, Humboldt drew on the pioneering mapping techniques of the earlier Wilhelm von Zimmermann (1743–1815), who produced, albeit from secondary sources, some of the first scientific cartographies of animal distribution in the mid to late eighteenth century. A nascent “animal geographer” himself, Zimmermann, like Humboldt, saw a key relationship between species distribution and the specific characteristics of different natural environments, though his belief in an essential natural fixity was soon challenged by Buffon and later by Darwin and evolutionary theory.

This early geographical concern with the spatial distribution of animal species, defined variously as geographical zoology, zoological geography, or zoogeography, achieved its defining moment in the work of nineteenth century scientists such as Sclater, Gunther, and, perhaps most importantly, Alfred Wallace. The resultant Sclater–Wallace classification of the six major faunal realms of the Earth has proved of lasting value, and continues to define and inspire the science of biogeography today. Darwin’s own faunal mapping of mammalian distributions provided him with the means to develop evolutionary theory, though he was arguably more interested in the barriers between faunal regions and their impact on species development than in the composition of the regions themselves.

Contemporary reassessments of the work of these early explorer-geographers, keen to classify
the world into distinct biophysical regions, have been quick to identify the inherent Eurocentric epistemology that drove the “discovery” and eventual naming of these exotic animals, as Humboldt’s description of the Andean Llama as a “camel-like sheep” illustrates. Yet, ultimately, the eventual acceptance that “New World” species had a uniqueness that did not fit readily into the pre-existing north European worldview and explanation, whether cultural, biological, or theological, served an almost metaphysical role in challenging traditional models and ideas about nature’s immutability, special creation, and distinctiveness, thereby ultimately paving the way for evolutionary theory and a new understanding of the possible links between human and nonhuman. And with this came an important epistemological shift. Darwin’s evolutionary theory, with its emphasis on change and adaptation, effectively drew to a close the purely descriptive approach of classic animal geography as, later, ecology and biogeography took up the challenge of explaining the spatialities, forms, and trajectories of animal movement, responsiveness, and adaptation.

The focus of all of these early animal geographies was the “natural” distribution and location of animal species, from which grand theories of biological “life” were ultimately constructed. Of course, by the mid-nineteenth century, in many regions of the world European colonial expansion had already begun to have a major transformative effect on the faunal composition and identity of many lands. Species (and peoples) deemed incompatible with the expansionist aims of the colonizers were eradicated or pushed to the margins. New species were brought into these “new” lands, creating what DeJohn Anderson calls the “advance guard” (2004, 243) of an inexorable cultural and physical reconfiguration both of “nature” and of human–nature relations. The large-scale human-induced changes to animal distributions, and the cultural as well as ecological consequences of these changes, have been explored and assessed by a number of scholars including, notably, Virginia DeJohn Anderson, whose account of the role of European livestock in transforming not only the ecologies, landscapes, and natural fauna of North America, but also the entire semiotic and representational placing of animals within both settler and indigenous populations, is truly fascinating. In Australia similar transformative processes, often consequent on colonial expansion, have been observed and discussed by writers such as Adrian Franklin (2006) and Deborah Bird Rose (2011). “At all stages of Australian history,” writes Franklin, “the animals of Australia have found themselves at the centre of biological, moral and political puzzles” (2006, 26). They also have a more material presence against which cultural narratives might have less immediate impact relevance. “Go away! (We’re British here.),” shouts Val Plumwood at the Kakadu National Park crocodile that almost kills her (2000, 131). In other disciplines, notably anthropology, the cultural roles and placings of animals within local value systems, both symbolic and material, indigenous and imposed, have a rich history, from Geertz and Levi-Strauss to Sauer and Ingold, and these too have strongly informed the development of contemporary animal geography for which, critically, animals are so much more than merely natural objects in natural places.

Finally, and most recently, a new generation of animal geographers have found a modern relevance in some of the approaches of the classical, positivist zoogeographies of the nineteenth century. Where the early writings of Humboldt, Wallace, and others tended to conceive of animals as the passive results of particular ecological, environmental, and evolutionary conditions,
modern zoogeographers have sought to emphasize the role of animals as active subjects within their environments, responsive to the affordances and possibilities of a more intentional and agentic engagement in both natural and more humanized spaces. For example, drawing both on ecology and on human geography in his account of lizards’ and snakes’ sharing of spaces with humans, Campbell (2009, 267) shows how reptilian presence and adaptive behavior actively interact with human copresence and decision-making to create what is, in effect a more holistic zoogeography.

Farmed animals

Of course, animals and humans have performed space together for centuries within the context of domestication, whether it has been for purposes of consumption or of comfort. Although the recently emergent field of “new” animal geography has been less ready to engage with domestication, largely as a result of the explicit and obvious asymmetry of both apparent agency and moral status in the human–animal relationships associated with domestication, to which we shall return, there is nonetheless an important tradition in geography’s engagement with the animal to be acknowledged here. Whatever we may think individually of the ethical basis of animal keeping, it remains a fundamental characteristic of human societies and, in most parts of the globe, a long-standing mainstay of anthropocentric economic systems and their consequent spatialization, from the earliest Neolithic settler societies to modern intensive livestock production, with its estimated global asset value of almost $1.5 trillion.

Geography has a close affinity with agriculture, though few might admit it. Both are, to a certain degree, predicated on the operation of a relationship, itself predicated on an increasingly contested dualism, between human society and the “natural” world as a resource. Livestock agriculture is in many ways the most important, and certainly the most common, form and practice of human–animal relations, involving billions of animals and ultimately billions of human (and animal) consumers at any one moment.

By their very nature, geographies of agriculture and of livestock farming have tended, like their subject matter, to see farm animals predominantly as chattels and resources, commodified objects of value, subsidy, or edibility – a means to an end. Yet, livestock and the livestock sector offer a rich terrain for a wide range of geographical approaches that reflect, in their own way, broader shifts in the discipline’s theoretical and conceptual development. For the sake of clarity and expediency, we might identify three broad areas of interest, though these are far from being discrete and distinct.

The first is perhaps the most obvious; it concerns farm animals as units of production and sites of accumulation. Although we may not wish to label such studies specifically as “animal geography,” analysis of the livestock industry, through the lenses of political economy or policy analysis, spatial distribution, corporate structures, and rural social formations within livestock farming communities have all long been commonplace themes in economic, rural, and agricultural geographies. Such studies serve an important function for contemporary animal geography. They provide empirical detail on the significance and range of human–animal interdependencies and the manner in which these presences are so fundamental to our economic and policy structures, our practices, and their spatialization. In Watts’s (2000, 295) terms, farm animals become transformed in both economic and material terms as “sites of accumulation.”

A second contribution from the study of livestock and agriculture to an emergent animal
geography has been a more culturally informed perspective, which seeks to move away from a purely political economy approach to place the relationship of farmers and domesticated animals within a broader cultural frame. Anderson’s (1997) critical deconstruction of domestication largely sets the scene for this particular area of animal geographies, but Yarwood and Evans’s (2000) take on this is specific, namely how different breeds of livestock are “socially constructed” to inform particular spatial and cultural repertoires. Others have, since the mid-2000s, taken this broad perspective into different contexts in which farm animals and humans meet, whether they be on hobby farms, in agricultural shows, or in slaughterhouses. More recently, and drawing on both critical continental thought and the sociology of science, this particular tradition within what we might call agricultural animal geography has moved toward a greater understanding of how technology, both the material technologies of livestock agricultural infrastructure and the breeding technologies of farm animal genomics, increasingly mediates and, in some cases, replaces the relations between humans and farm animals.

Finally, there has been, since the mid-2000s, an increasing intersection between animal, agricultural, and environmental geographies. This has taken a number of forms but two in particular stand out. The first is part of a growing concern for the environmental (as well as both human and animal health and welfare) implications of modern livestock farming, while the second emerges from geographers’ increasing interest in environmental histories and, for many, the place of livestock and livestock species within them. The former has, in recent years, attained particular policy relevance, with issues ranging from water pollution from farm animal wastes to the sustainability of livestock-based agriculture at the global scale. In relation to the second intersection, Cronon’s (1991) magisterial account of the growth of Chicago as “Nature’s metropolis” reveals, along with much else, the manner in which animals (in this case bison, cattle, and pigs) and humans (here hunters, farmers, slaughtermen, and packers) first co-created the “Great West” and then, ultimately, freed its principal product, animal flesh, from the bounds of nature and geography. More recently, animal geographers, with their concern for the placing and place-making of human–animal interaction, have drawn on the environmental history tradition to contribute to often wide-ranging debates around issues such as species reintroduction (whether domesticated or “wild”) and wildlife conservation and restoration programs.

Relational animals

While these two broad areas of animal geographies have certainly made significant contributions to broader geographical scholarship and have, to a large degree, been essential precursors, animal geography as a distinct and recognizable subfield within contemporary (human) geography has nonetheless, since the mid-1990s, taken on a very different allure. Some have termed this the “new” animal geography to draw a clearer distinction between it and earlier concerns. Others refer to an entire “animal turn” within the (human) social sciences and the humanities. Whatever the nomenclature, this period has seen a veritable stampede of animals into contemporary geographical thought. Like many recent subdisciplinary themes, animal geography has emerged not out of any intentional epistemological or coherent political strategy, but rather from a disparate group of people who have sought, in different ways and for various reasons, to develop a more-than-human geography. Early collections, notably two edited books published by geographers on either side of the Atlantic (Philo and Wilbert 2000; Wolch
and Emel 1998) and two journal special issues (Society and Space in 1995 and Society & Animals in 1998) draw the initial contours of the emergent subdiscipline and tell the story of its nascent flourishing. Since that time, the subdiscipline of animal geography has grown enormously through published work, conferences, journals, and book series. It now earns its own entries in dictionaries and encyclopedias of geography as well as its own Wikipedia site and Association of American Geographers specialty group.

So what is new about the new animal geography? Arguing for an acknowledgment of the historic and global significance of human–animal relations for both human and nonhuman spatial practice, Philo and Wilbert (2000, 4) assert that “a ‘new’ animal geography has emerged to explore the dimensions of space and place which cannot but sit at the heart of these relations.” For these early pioneers of the contemporary subdiscipline, animal geography was about understanding how animals – as figures of nature – and their “spaces” were differentially defined through science, through social practice, through culture and religion, and so on, whether it be as pets, as edible, as useful, as clean, as sentient, as repulsive, as metaphors, as representations, or as others.

As the definitive sites of animal ordering, zoos were also among the first “proper” animal spaces to interest animal geographers. In a seminal article from 1995 the geographer Kay Anderson wrote of zoos “as an institution that inscribes various human strategies for domesticating, mythologizing and aestheticizing the animal universe” (Anderson 1995, 276). Zoos are about establishing and reaffirming human-made, culturally inscribed spaces, boundaries, and categories in which we place nonhumans. Being predominantly located in cities, zoos are also celebrations of that boundary-making and the human–animal encounters that boundaries facilitate but also obstruct. For Anderson, zoos are about the differential human construction of the otherness of animals. Recognizing how animal otherness has been constructed through different narratives (some metaphorical) and actions (in many cases actions of spatial ordering and regulation) has been the first step of the new animal geography.

Writing a couple of years before Philo and Wilbert, Wolch and Emel (1998) sought to explain the new animal “moment” in geography and wider social theory in slightly more critical terms. For them, echoing John Berger’s famous 1970 essay “Why Look at Animals,” animals had become invisible to the social sciences and to geography, despite the fact that they were “indispensable to the structure of human affairs and so tied up with our visions of progress and the good life” (Wolch and Emel 1998, xi). Strongly influenced by the preceding cultural turn in human geography, and by writings on feminism and race, this new animal geography focused largely on the various forms and dynamics of the cultural representation and construction of animals and their spaces. Its overarching project was, in Wolch and Emel’s terms, to “bring the animals back in” – into our accounting, our explanation, our science, our ethics, our politics – a methodological as much as an ontological or epistemological challenge. Of course, for many humans, animals are already closely experienced as integral components of shared lives, spaces, and behaviors, whether as animal carers, farmers, or zookeepers. Hence, to Wolch and Emel’s reasons for the emergence of a new animal geography, Urbanik adds a further one: “the increasing acceptance of humans’ emotional connections with animals” (2012, 7).

At one level, this new animal turn extends the range of human geography to include the multitude of ways in which the presence of the nonhuman can be brought into the geographical
ANIMAL GEOGRAPHIES

understanding and explanation of predominately human worlds. It fosters an awareness of how human–animal relations, all of them variously material, semiotic, affective, and ethically engaging, are co-constitutive of, and thereby potentially redefine, spaces and spatial practices that are no longer only human. Here, then, emerges a more inclusive animal geography in which animals are themselves actors in contexts and in “beastly spaces” that are no longer necessarily entirely “natural” (as in devoid of human presence and influence) or exclusively “human.” Animals become acknowledged agents within human narratives of place, from the home to the zoo and the “wild.” Thus, in her challenge to the traditional anthropocentric understanding of the city, Jennifer Wolch offers the notion of “Zoopolis” as a “city re-enchanted by the animal kin-dom” (1988, 135). Cities now become the focal points for a new “transspecies urban theory” (Hovorka 2008) which weaves animals and humans together into spaces, networks, and topologies of heterogeneous copresence and agency. For many, this is the new animal geography: an accounting of, and attentiveness to, the multiple interrelated ways in which both nonhumans and humans perform space together. It is, nonetheless, grounded in an essentially humanist ontology of difference.

At another level, however, the animal turn goes a lot further in that it offers the potential for a more troubling and foundational critique, one that allows for other stories, other possibilities, and other ethics. “The Great Divides of animal/human, nature/culture, organic/technical, and wild/domestic,” argues Haraway (2008, 15) “flatten into mundane differences – the kinds that have consequences and demand respect and response – rather than rising to sublime and final ends.”

For some recent scholars of contemporary animal geography, the otherness of the nonhuman, anchored in such dualisms, matters far less. Indeed, the animal walks and crawls across these divides, defying the abyss which seeks to separate them while demanding some new form of affinity. Drawing on a more radical set of approaches coming from posthumanism, from new ideas about self, subjectivity, and identity, from the methodologies of actor-network theory, ethology, and multispecies ethnography as well as from a sense of the different forms of (bio)politics that may result, animal geography has moved toward a more relational, symmetrical, and ultimately generous form of understanding that acknowledges the animal not as animal, that quintessential other, not as a singular (possibly subjective) agentic force, but as partner in creative and meaningful intra-action. “All real living is meeting,” writes Barad, “And each meeting matters” (2007, 353). This has certainly given rise to a particularly fecund period for animal geography. Inspired by writers like Donna Haraway and Jacques Derrida, and faced with a largely unexplored “kingdom” of nonhuman relational potential (though some may note a preponderance of the more familiar mammals), the subdiscipline has moved comfortably into the realms of both critical social science and critical social theory, and offers what is for many a genuinely new way not only of seeing and accounting and of doing politics, but also of being with.

Animal geography, while it has not become mainstream, now has a growing presence in geographical scholarship. It is heterogeneous and wide-ranging, ethically relevant and methodologically innovative, generic and intimate (Buller 2014). It has opened doors to other stories and other worlds, and has brought particular insights, and examples, to biopolitics and the recent relational and materialist trends
within contemporary geography. The onset of the Anthropocene – more as a scientific paradigm than a geological epoch – and the end of a particular conception of nature has perhaps ironically prompted a new academic interest in the role and place of, particularly, “wild” animals and animal conservation, introducing the possibility for a lively intra- and interdisciplinary rapprochement with biogeography and the life sciences (Lorimer 2010).

Sometimes certainly, animal geography continues to act as a qualifier or foil to more abjectly humanist accounts rather than as a starting point in its own right. Other times, there is a notable epistemological tension between seeing the animal as a theoretical or conceptual object rather than as a living, material being with whom we share an innate vitality.

Given the profundity and range of its (potential) ontological and epistemological challenge – and what is for some also a political and ethical contestation – to the mainstream humanist tradition within both (human) geography and the human social sciences, one might wish it had succeeded in achieving a greater impact than it has. Certainly, this profundity and more-than-human range provide animal geography with a particular pertinence and critical power across a range of domains in human geography. Yet its attendant ontological complexity, as well as its ethical plurality and, some might say, relativism, has also been an element of division and weakness, leading, for some outside the subdiscipline, to a too ready dismissal.

For its immediate future, there are perhaps three key challenges. The first is the challenge not to use animals simply to talk about humans – the question “What is an animal?” should not only be autobiographical. The second is the ethical challenge: How do animal geographers position themselves with respect to the manifest anthropocentrism and, for many, obvious speciesism (and violence) that endures in virtually all human–animal encounters? The third is methodological: How do we account for non-humans in ways that are meaningful not just for us but also for them? A final challenge, perhaps the hardest of all, is that of convincing human geography colleagues that theirs is, by definition, only a partial understanding of the world.

**SEE ALSO:** Agricultural geography; Ancient geography; Biogeography; Cultural geography; Cultures of nature; Environmentalism; Socio–nature

**References**


ANIMAL GEOGRAPHIES


Further reading


The Annales School embraces a wide group of scholars who have introduced important innovations in their studies of the past. It is named after the scholarly journal *Annales d'Histoire Economique et Sociale* founded in 1929 by historians Lucien Febvre (1878–1956) and Marc Bloch (1886–1944), who were then based in Strasbourg. Drawing on anthropology, geography, psychology, and sociology, they promoted a break with the study of monarchs, diplomatic events, and political regimes favored by most historians at the Sorbonne in Paris. Both Febvre and Bloch had been taught geography by Paul Vidal de la Blache (1845–1918), and were responsive to work by sociologist Émile Durkheim (1858–1917) and articles in *L'Année Sociologique*. They and their entourage pioneered study of long-term historical structures over individual events and of common people rather than elites. Yet, they tended to reject class-based approaches of Marxist scholars. Unlike the Paris-based focus of the “Sorbonniste” historians, studies in regional or provincial history were welcomed in the *Annales*. From 1933 it was edited in Paris after Febvre’s move to the capital; he was followed there by Bloch in 1936. During the German occupation of France, the journal still appeared but under different titles. As a Jew, Bloch was deprived of his university post but he continued to write under a pseudonym. Febvre removed him from the editorial board to comply with censorship regulations. As an active member of the Resistance, Bloch was shot by the Gestapo on 16 June 1944.

After the war, Febvre continued to foster the *Annaliste* approach, renaming the journal *Annales, Economies, Sociétés, Civilisations* (1946–1994). He mentored Fernand Braudel (1902–1985) whose ideas on different notions of time (deep geographical settings, evolving socioeconomic and human structures, and specific moments) characterized a new phase. Braudel’s publications on the Mediterranean world, capitalism and material life, and other themes raised the reputation of the Annalistes, especially in Italy, Poland, and Latin America. From the Maison des Sciences de l’Homme in Paris, Braudel promoted the notion of “total history” encapsulating adjacent areas of scholarship including human geography. He also orchestrated networks of like-minded scholars in other countries. After combining the study of structures with that of collective ideas in the 1960s, younger Annalistes, such as Emmanuel Le Roy Ladurie (b. 1929) favored work on *mentalités* and experimented with quantification. As it grew larger, the Annales School became less cohesive and distinctive. Members of its next generation, such as Roger Chartier (b. 1945), distanced themselves from *mentalités*, preferring to embrace cultural, linguistic, and literary approaches. They continued to subscribe to the founding aim of Febvre and Bloch: to demonstrate methodological innovation. In 1994, the journal was renamed *Annales: Histoire, Sciences Sociales* but its influence had waned. An English-language edition was initiated in 2013 to attempt to capture lost interest.

Historical geographers have drawn inspiration from the interdisciplinarity and voluminous
writings of several generations of Annaliste historians. It should be remembered that an earlier “Annales school” developed from the Annales de Géographie that Vidal de la Blache founded in 1891. This journal sought to diffuse his “new geography” that highlighted regional synthesis, environmental context, and cartographic representation, while acknowledging the role of the past in shaping current human activities. Modified forms of the Vidalian approach remained important in France and in countries influenced by French scholarship until the middle of the twentieth century.

**SEE ALSO:** Historical geography; Human geography; Regional geography

**Further reading**


The Antarctic is a large continent with a surface of 14 million km², situated around the South Pole and almost entirely covered with ice. With an average ice thickness of 2200 m it represents 90% of terrestrial ice and is equivalent to 58 m of sea level. The Antarctic continent is the world's largest ice sheet. It is entirely surrounded by oceans and isolated by large and strong oceanic currents. The sea ice around Antarctica almost disappears every summer but reaches around 18 million km² during winter. The continent is generally divided into three parts: the Peninsula, which is the northernmost part of the continent closest to South America, the East Antarctica, and West Antarctica, separated by the Transantarctic Mountains. East Antarctica is the largest one. In some places, the ice thickness exceeds 4 km. Most of the ice lies on a bedrock above sea level. In several places, high basal temperature and bedrock geometry induce subglacial lakes. The most famous is the Vostok subglacial lake, located nearly 4000 m below the surface of the ice, leading at the surface to a very flat area 250 km long and 50 km wide. The western part of the Antarctic ice sheet is smaller, representing the equivalent of “only” 3.3 m of sea level, and the bedrock mostly lies at around sea level.

The Antarctic ice sheet is the coldest, highest, driest, and windiest continent of the Earth. The surface temperature decreases from the coast toward the interior from −15 to −60°C. A temperature of −93.2°C has even been measured in the East Antarctic Plateau, the world record. This cold and dense air in the interior rushes downslope and can induce strong and persistent katabatic winds. For instance, in Terre Adélie (Adélie Land), the average katabatic wind intensity is close to 40 km/h, and the maxima can reach three to four times this value. These katabatic winds carve the surface at centimetric and metric scale, and they also produce snow drift and contribute to erosion through sublimation phenomena.

Because of the extremely low temperature, the snow accumulation rate in Antarctica is very low: less than a few centimeters per year in the interior and a few tens of centimeters near the coast. As the continent is huge, this represents approximately 2200 Gt each year, equivalent to 6 mm of sea level. The snow sinks, turns into ice, and flows down very slowly toward the coast where it calves into icebergs. Most of the flow process occurs by deformation although in some places, where ice temperature is close to melting point, basal sliding occurs. Ice velocity is extremely low in the center of the ice caps, less than 1 m/year, whereas in some places near the coast it can reach over 1000 m/year. The spatial pattern of ice flow is probably not regular: large ice streams that drain ice can be detected several hundred kilometers inland. Most of the ice is evacuated by several large outlet glaciers. When they reach the coast, they terminate either in floating ice shelves, or in grounded or floating glaciers. If the ice sheet is in its steady state (in equilibrium), ice calving might be equivalent to 6 mm of sea level each year. Because of the very slow ice velocity in most of the Antarctic ice sheet, it has become, as has the Greenland ice sheet, the repository of polar records. For
instance, from the European Project for Ice Coring in Antarctica (EPICA) at Dome C (also called Dome Concordia) it has been possible to reconstruct over 800,000 years of climate history.

The shape and the volume of the ice sheet are determined by the equilibrium between snow fall and ice flow. Given the magnitude of the input (snowfall) and output (iceberg calving), even a slight imbalance between those two components can result in a significant sea level change, which is why the mass balance of the Antarctic ice sheet has been surveyed for several decades. Both snow accumulation and ice velocity depend on climate. Precipitation and evaporation rates are clearly linked to climate variations; therefore the dynamic response to a climate change could occur over tens of thousands of years. For example, because ice is a very good thermal insulator, the temperature and thus the velocity of the central part of the Antarctic ice sheet have still not stabilized since the Holocene warming that occurred 15,000 years ago. The current state of the Antarctic results from the climate of the region tens of millennia ago. There are many uncertainties in the imbalance mechanisms so it is still very difficult to model them or even to determine the trend (positive or negative preeminence). The West Antarctic ice sheet is particularly difficult because its bedrock is under the sea surface, so that the ice flows through a complex network of ice streams. To understand the stability of the ice sheet is also a challenge.

There is considerable uncertainty over the quantity of snow fall and of ice calving or melting, the margin of error is between 20 and 30%, and consequently a direct estimation of the mass balance is quite difficult. To measure the mass balance of Antarctica or to understand, model, and predict the ice sheet’s evolution, we need to know the climatological and dynamic processes that control them. We also need observations at a global scale to enhance the models. The sheer size of the continent, its relative inaccessibility, and the weather conditions make in situ measurements difficult and many observations are missing. A few scientific stations, mostly near the coast, and a few traverses during expeditions provide sparse data. On the other hand, observations from space, particularly radar observations, which are free from atmospheric conditions and illumination, provide surface and volume information with good space and time sampling. Some data can be directly linked to a relevant parameter: for example radar altimetry allows us to estimate the topography, radar interferometry helps us to estimate the surface ice flow, a scatterometer gives estimates of katabatic wind direction, and we use gravimetry to deduce the mass balance change, or optical images to derive either surface topography or surface velocity. Other data are indirectly linked to relevant parameters: for example brightness temperature from a radiometer and backscattering from radar altimetry are sensitive to ice grain size but not to accumulation rate. A long-term analysis of the observations derived from these sensors can be used to detect temporal change in geophysical parameters.

Among these sensors, altimetry and gravimetry are efficient methods to measure volume and mass balance directly at the global scale. For the last two decades, polar regions have been surveyed by radar altimeters (the European Remote Sensing (ERS) satellites ERS-1 and -2, Envisat, and Cryosat-2) and a laser altimeter (Icesat). These sensors estimate volume changes; therefore we need assumptions on snow density to derive mass changes. Moreover, they suffer from poor sampling and significant error budgets near the coast where most of the variations occur. The second remote sensing method, gravimetry, directly measures the ice sheet mass change. However, it also has limitations. In particular, it is
sensitive to glacial isostatic adjustment, due to the isostatic response after the Holocene warming.

In coastal regions of the Antarctic ice sheet, where altimetry fails due to the steeper slopes and where its spatial coverage is poor, the often complex pattern of elevation changes can be depicted using comparisons of digital elevation models (DEMs) which are digital representations of the surface topography. One source for such DEMs has been the SPIRIT (SPOT5 stereoscopic survey of Polar Ice Reference Images and Topography) project, an International Polar Year (2007–2009) initiative by the French Space Agency (Centre national d'études spatiales – CNES) that provided numerous HRS-SPOT5 (High-Resolution Stereoscopic-Satellite Pour l’Observation de la Terre 5) images over polar regions. SPOT5 captures two views of the same scene from slightly different angles that can be used to build a DEM. In the Antarctic Peninsula, well known for its complex geography and for being one of the fastest-warming places of the Earth, the comparison of multitemporal DEMs was used to assess the response of outlet glaciers to the loss of their frontal ice shelf, in particular the catastrophic breakup of the Larsen B ice shelf in March 2002. Following this event, glaciers experienced a strong acceleration and thinned by as much as tens of meters per year. This relatively small (3000 km$^2$) region upstream of the former Larsen B ice shelf has lost 10 Gt per year since 2002. New stereo images will be acquired in 2014 by the SPOT5 stereo sensor during the SPIRIT2 project and will provide a second snapshot of the coastal regions of Antarctica before the demise of the SPOT5 satellite at the end of 2014.

The analysis of all published results shows a very great dispersal in terms of mass balance. The loss is estimated to be between 0 and 150 Gt per year, leading to a potential sea level rise of between 0 and 0.4 mm/year. East Antarctica is almost in balance, with volume change probably linked with meteorological forcing fluctuations, while West Antarctica shows a worrying loss (Remy and Parouty 2009).

Most studies indicate a clear acceleration over the past decade by approximately 10 Gt/year$^2$. For instance, the altimetry analysis suggests a near balance for the 1990–2000 period and a loss of $170 \pm 4$ Gt/year for the last decade (Zwally and Giovinetto 2011). This recent acceleration in the Antarctic ice sheet mass loss has been attributed to the dynamic response of the ice sheets to recent warming, with most of the ice sheet mass loss resulting from accelerating coastal glacier flow in particular in West Antarctica (Flament and Remy 2012; Shepherd et al. 2012). One of the possible reasons is the weakening and breakup of the floating ice tongue or ice shelf that buttresses the ice stream. Because the ice shelves are in contact with the sea, changes in ocean circulation and warming of seawater may trigger basal melting and further breakup of the shelves, causing coastal ice flow to speed up.

At the end of 2013, we had three years of observations from Cryosat-2, and the altimeter AltiKa on Saral, a joint French-Indian mission, was launched on exactly the same orbit as the previous European Space Agency (ESA) satellites ERS-1, ERS-2, and Envisat. The observations from both sensors show a large and unprecedented mass loss in West Antarctica of all the glaciers flowing toward the Amundsen Sea, reaching a total of 150 Gt/year. In the near future, the Antarctic ice sheet, in particular its western part, should be closely and carefully surveyed, especially the outlet glacier. New satellites providing altimetry data such as Sentinel-3 (from ESA, to be launched in 2015) or Icesat-2 (from NASA, to be launched in 2016) will continue to maintain the continuity in the observations.
ANTARCTICA

SEE ALSO: Climate change and land ice; Glaciers; Holocene; Ice sheets; Sea level rise

References


Anthropocene and planetary boundaries

Noel Castree
University of Manchester, UK

The Anthropocene concept: definition and origin

“The Anthropocene” denotes an epoch in which the collective impact of human activities is sufficient to significantly alter the conditions of life on planet Earth. The term is said to have been coined by American freshwater ecologist Eugene Stoermer in the mid-1980s. However, it was the Dutch Nobel Prize winning atmospheric chemist Paul J. Crutzen of the Max Planck Institute for Chemistry who did most to popularize the term among Earth and environmental scientists. In 2000, he and Stoermer published a short article in the Global Change Newsletter (GCN), the “house publication” of the International Geosphere-Biosphere Programme (IGBP). The IGBP was one of several multicountry research programs designed to study contemporary environmental change on a global scale by combining the expertise of multiple academic subjects, from climatology to oceanography to biogeography. In their article Crutzen and Stoermer (2000) suggested that the intentional activities of people, such as commercial farming, traveling by car, and consuming imported foods, had unintentionally altered the “boundary conditions” defining the Holocene. The Holocene is the period of Earth surface history beginning around 11,700 years ago when the most recent “ice age” ended. Research suggests that the various climatic zones and biomes emergent during the early Holocene remained fairly stable from thereon, as did average global atmospheric temperature. The most recent centuries of the Holocene have been those where the number of *Homo sapiens* grew significantly and when large-scale migrations led to their dispersal away from centers of origin. Crutzen and Stoermer proposed that the cumulative environmental impact of people living in so-called advanced societies is such that the Holocene might be at an end – if not at the time of writing, then soon. Given how slow geological change normally is, at least when measured against an average human lifespan, the suggestion that people are now equivalent to a “natural force” (such as a large meteorite hitting Earth) was a startling one. It amounted to saying that anthropogenic climate change, already a serious scientific and policy concern by 2000, was only the tip of the proverbial iceberg.

In 2002 Crutzen presented the Anthropocene hypothesis to the wider scientific community in a short article published in the world’s leading science periodical Nature (Crutzen 2002). Counter-intuitively titled “Geology of Mankind,” the article pointed out that if current Western standards of living become globalized then people worldwide will take the earth’s biophysical systems into terra incognita. This may, he argued, necessitate technical interventions of unprecedented scale and scope, such as “internationally accepted … geo-engineering projects … to ‘optimize’ climate” (Crutzen 2000, 23). Similarly, Crutzen speculated, it may require substantial measures to engender “appropriate human behavior at all scales” (2000, 23) so as to
avoid harmful unintended impacts on fresh water resources, nitrogen cycles, oceanic currents, and much else.

The Anthropocene concept: current visibility and precursors

Since Crutzen’s Nature article, “the Anthropocene” has attracted increasing attention across a range of Earth and environmental science fields, including some branches of physical geography. The attention was by no means immediate. Even in 2006 Crutzen’s neologism rarely featured in peer-reviewed scientific research papers. However, for various reasons to be explained below, it is now a fairly familiar term in the semantic landscape of several science subjects. Indeed, a new science journal, published by Elsevier, bears the name (see www.journals.elsevier.com/anthropocene/), as does the new interdisciplinary Anthropocene Review (published by SAGE). The Anthropocene even enjoys a degree of public and political visibility, albeit currently modest. For instance, The Economist magazine devoted a whole section of a late May 2011 issue to the subject, while two months earlier National Geographic magazine ran an article by American science writer Elizabeth Kolbert entitled “Enter the Anthropocene – Age of Man.” The year ended with The New York Times publishing an op-ed on the subject (Marris et al. 2011). Meanwhile, science popularizer and environmentalist Mark Lynas used the idea of “planetary boundaries” to structure his 2011 best-seller The God Species: How the Planet Can Survive the Age of Humans. This idea, as will be explained below, is closely related to the Anthropocene concept and had already been highlighted in an April 2010 issue of Scientific American entitled “Managing Earth’s Future.” Two years later, scientists involved in the IGBP and other global environmental change research programs organized the “Planet Under Pressure” conference just prior to the United Nations Rio+20 Earth Summit. This conference, intended to capture the attention of politicians, opened with a time-lapse video of recent Earth surface change entitled “Welcome to the Anthropocene” (available online by searching on the title).

Though “Anthropocene” remains a novel term, its meaning is not entirely original. As British geologist Jan Zalasiewicz and colleagues noted, “The term Anthropocene is the latest iteration of a concept to signal the impact of collective human activity on biological, physical and chemical processes at and around the Earth’s surface” (Zalasiewicz et al. 2011, 1037). For instance, eight years before Crutzen and Stöermer’s article in the Global Change Newsletter, New York Times environmental correspondent Andrew Revkin coined the term “Anthrocene” to designate “a geological epoch of our own making” (Revkin 1992). Over a century earlier Italian geologist Antonio Stoppani (1873) had presciently talked of the “anthropozoic era.” Then, in the 1920s, the Russian geochemist and naturalist Vladimir Vernadsky was among the first to propose the idea of the “noosphere.” For him this was the realm of human imagination and desire, a constitutively social (even virtual) domain irreducible to humans’ biological capacities but with world-changing potential realizable through technology. Vernadsky had already argued that the biosphere was a force of nature akin to the movements of tectonic plates. Relatedly, his noosphere concept suggested that Homo sapiens might rise above other species to exert a global influence. In the 1940s his ideas reached a wider English-speaking science readership (see Vernadsky 1945).

These and other predecessor terms did not catch on in the scientific community (or wider
world) in the way that Anthropocene seems to have done since 2000. What explains the difference? It no doubt helps that Crutzen is a highly visible and esteemed scientist in the broader earth and environmental science community. What’s more, he has been adept at developing relationships with notable members of this community who have a professional stake in examining the Anthropocene hypothesis more closely. Among them are the aforementioned Zalasiewicz and the Australian Earth system scientist Will Steffen. In addition, the fact that anthropogenic climate change is believed by most climate scientists to be real and significant is an important background consideration. If humans can accidentally alter the world’s climate then it is not so far-fetched to take seriously the suggestion that we are taking other biophysical systems away from their Holocene norms. Relatedly, environmental measurement and monitoring systems are far more numerous and granular than a generation ago, in part because of the impetus provided by the IGBP and related research programs. This means that large claims about global environmental change are more readily testable than heretofore and easier to accept as working hypotheses that can be dis/confirmed in due course. Finally, and going back a little further in time, more than one generation of scientists, politicians, business people, and members of the public are accustomed to ideas as intellectually and practically challenging as the Anthropocene. Even prior to anthropogenic climate change (once known more pointedly as “global warming”) becoming a key term in our societal lexicon, concepts like “natural limits to growth,” global “over-population,” and “nuclear winter” were emanating from (parts of) the scientific community as far back as the mid-1960s. These concepts were taken seriously in the wider society, even if they fell out of use for various reasons. For older members of several nonacademic communities (political, commercial, etc.) the idea of the Anthropocene has thus not come out of the blue, but is only the latest – albeit most far-reaching – term used to describe human–environment relations on a global scale.

A new geological epoch?

At the time of writing, discussion of the Anthropocene remains largely confined to academic circles, especially those that Crutzen and his coauthors move in. A minority of social scientists and humanities scholars are also beginning to consider the Anthropocene, including a small number of human geographers whose arguments will be considered toward the end of this entry. Only time will tell if the Anthropocene idea becomes truly prominent outside the academic world. In the earth and environmental sciences, the debates have focused on issues of measurement and inception. The questions have been “What evidential markers can tell us whether or not the Anthropocene has begun and the Holocene ended?” and “When, exactly, did the Anthropocene start?” These questions have arisen because of Crutzen’s success in enrolling other scientists in testing his Anthropocene hypothesis. For instance, in late 2007 he co-published a paper with the abovementioned Will Steffen and American environmental historian John McNeill. Its dramatic and quizzical title was “The Anthropocene: Are Humans Now Overwhelming the Great Forces of Nature?” (Steffen, Crutzen, and McNeill 2007). The paper appeared in the peer review, multidisciplinary science journal Ambio. More detailed in evidential and analytical terms than the earlier GCN and Nature articles, it was also more confident that the Holocene is a thing of the past.
However, none of the three authors was (or now is) a geologist. By contrast, Leicester University’s Jan Zalasiewicz was, in 2007, chair of the Stratigraphy Commission of The Geological Society (located in London). He noticed Crutzen and others’ use of the Anthropocene idea. He proposed to the other 20 commission members that this idea should and could be tested using formal geological criteria for the identification of an epoch. The result was a coauthored article that appeared in GSA Today, the house periodical of the Geological Society of America (Zalasiewicz et al. 2008). Entitled “Are We Now Living in the Anthropocene?” it detailed the measures necessary to establish if and when the Holocene had ended. To quote from it at some length, “Earth has endured changes sufficient to leave a global stratigraphic signature distinct from that of the Holocene or previous Pleistocene inter-glacial phases, encompassing novel biotic, sedimentary, and geochemical change. These changes, though likely only in their initial phases, are sufficiently distinct and robustly established for suggestions of a Holocene–Anthropocene boundary in the recent historical past to be geologically reasonable. The boundary may be defined either via Global Stratigraphic Section and Point (‘golden spike’) locations or by adopting a numerical date” (Zalasiewicz 2008, 4). As a result of this paper and subsequent discussions among the academic networks of commission members, the International Commission on Stratigraphy – which is ultimately responsible for identifying geological epochs – established an Anthropocene Working Group and made Zalasiewicz its chairman. At the same time Zalasiewicz and his Leicester colleague Mark Williams joined with Crutzen and Steffen to present the question of formally establishing the Anthropocene’s epochal status to non–geologists (Zalasiewicz et al. 2010a). A set of 2011 papers in the Philosophical Transactions of the Royal Society A were intended to do the same (Ellis 2011; Steffen et al. 2011a; Zalasiewicz et al. 2011). As a consequence of these interventions, and various conference presentations by their authors, the broader environmental science community has been drawn into a geological discussion of epochal markers normally confined to earth science and normally focused on the deep past.

Which lithostratigraphic, biostratigraphic, or chemostratigraphic indicators, if any, are considered sufficiently suggestive of a phase shift in Earth surface conditions to serve agreed criteria for the Holocene’s eclipse? There are several candidates, from lake sediments to greenhouse gas concentrations to artificial isotopes produced by nuclear weapons detonations. Clearly, because these markers all pertain to both recent and ongoing anthropogenic environmental change, establishing the Anthropocene’s epochal status in a way satisfactory to geologists requires environmental scientists from a range of disciplines (including physical geography) to supply robust evidence of changes that might, thousands of years later, be considered clear stratigraphic markers by future geologists. For instance, the British Society for Geomorphology has established a fixed-term working group to advise the society on whether and how contemporary landforms might comprise stratigraphic markers (see Brown et al. 2012). The complications involved in identifying and interpreting markers of any kind are manifold (Rull 2013). First, which markers – one or several – should be taken as definitive of the Holocene’s termination? Second, given that we may only be in the very early years of the Anthropocene, it is important that any one marker be resolved to geologically precise temporal scales, like a decade or a year. This is technically challenging in many cases. Third, because the ecological impacts of human activity have been and remain diachronous,
significant environmental signatures evident in one part of the world (e.g., Western Europe) may not be replicated elsewhere until the last few years or next few decades. Finally, it is entirely possible that future environmental markers reflective of present-day human activities will prove to be more compelling indicators of the Anthropocene’s onset.

By virtue of some or all of these reasons, several environmental scientists doubt that the Holocene’s end can (yet) be measured in stratigraphic terms. For instance, in a paper about anthropogenic soils as potential stratigraphic markers Gale and Hoare (2012) conclude that “there are serious difficulties in using stratigraphic methods to define the base of the Anthropocene. In part this is associated with the worldwide diachroneity of human impact and the difficulty of establishing a single chronological datum for the epoch. Although several wide-ranging event markers exist (including bomb-produced isotopes), these fail to coincide with the worldwide initiation of [environment altering] human activity” (p. 1493). Given the several measurement issues itemized above, and given that the International Commission on Stratigraphy is very exacting about what constitutes a new geological time period, there is little chance that the Anthropocene will be declared a new epoch by geologists any time soon.

Beyond geology: current debates in the environmental sciences

This has not deterred several researchers from searching for chronometric indicators of the Anthropocene’s inception, regardless of their stratigraphic in/significance. In their 2007 Ambio paper, Steffen et al. divided the Anthropocene into three periods, beginning with “The Industrial Era” (1800–1945), followed by “The Great Acceleration” (1946–2015) and possibly followed by a period of “Earth System Stewardship” (2016–) in which humans at last take collective responsibility for their huge global environmental footprint (see also Steffen et al. 2011a). However, without discounting the chronometric significance of 1800, American paleoclimatologist Bill Ruddiman and colleagues have suggested that a global anthropogenic environmental signal can be recorded hundreds of years earlier (see the debate by Ruddiman, Crucifix and Oldfield 2011, in the pages of The Holocene). They propose that even as early as the Middle Ages (and perhaps centuries before that) collective human endeavor was able to unintentionally alter climatic conditions (see Ruddiman 2013, for a review of the evidence in favor of “the early Anthropocene hypothesis”). As with the discussion of stratigraphic markers it is unlikely that consensus will be reached on a single year or decade when the Anthropocene can be said to have begun. Instead, environmental scientists may need to accept that different worldwide anthropogenic signals of varied magnitude appeared at different times over the last six centuries. In this context, the “beginning” of the Anthropocene might be said to have been cumulative, unsynchronized, and strung out rather than punctual (see Lewin and Macklin 2013).

In light of these debates about how to measure and date a crossing of the Holocene–Anthropocene boundary, it may be thought that Crutzen and Stoermer’s hypothesis has reached an impasse a decade and a half after it was put forward. However, this is far from true. As paleoecologist Valenti Rull notes, “it is not necessary to formally define the Anthropocene as an epoch to accept that human activities have significantly changed earth system processes [and forms] during the last [few] … centuries” (Rull 2013, 4). Geologists like Zalasiewicz and Mark Williams, and environmental scientists like
Crutzen and Steffen, recognize this too. As they wrote in the journal *Environmental Science and Technology*, “quite how and when the Anthropocene is formalized is of secondary importance; its real importance is in being a means to integrate a wide range of environmental indicators [and] to consider them within the context of the whole of [the] Earth [surface]” (Zalasiewicz et al. 2010b, 6008). Consequently, it is possible for the Anthropocene to remain an informal concept in geology even as it becomes a normal part of the vocabulary of many environmental scientists. Its facility for the latter is that it denotes biophysical changes that are profound relative to the human, rather than geological, past.

One notable example of this is the recent published work of Johan Rockström and colleagues—including Crutzen and Steffen. In a 2009 paper in *Nature* (Rockström et al. 2009a) and a much longer essay published in the journal *Ecology and Society* (Rockström et al. 2009b), a large number of environmental scientists from across the disciplines have advanced the concept of “planetary boundaries.” Aside from Crutzen and Steffen, both high-profile scientists, researchers include noted ecological economist Robert Constanza and the prominent climate scientist James Hansen. Rockström et al. argue that the environmental conditions of the Holocene are preferable to those of an unknown and potentially inhospitable Anthropocene future. They identify nine global environment components constitutive of the Earth system. These pertain to climate, ocean acidity, chemical balances, atmospheric aerosols, biodiversity, land-use types, freshwater, nitrogen and phosphorus cycles, and stratospheric ozone density. For seven of these they specify a quantitative boundary, the crossing of which might take the system beyond Holocene norms (for the remaining two they are unable, as yet, to quantify the boundary). Together, the nine boundaries comprise what Rockström et al. call “a safe operating space for humanity” (2009a, 472). Modern humans, they argue, have already transgressed several of these. Recognizing that “Determining a safe distance involves normative judgements of how societies choose to deal with risk and uncertainty,” they nonetheless commend their “new approach to defining biophysical preconditions for human development” (Rockström 2009a, 472, 474).

Rockström et al.’s planetary boundaries concept is a particular version of the Anthropocene idea in all but name. We might therefore call it a “collateral concept.” It signifies environmental changes of the same scope and scale as the Anthropocene idea, but in a less politically neutral and more overtly normative way. This is not to suggest that Rockström, Crutzen, Steffen, Constanza, Hansen, and others are distorting science to serve political ends, that is, to recommend urgent action to stay within a “safe operating space.” It is simply to say that, in their view, current evidence of human impact suggests that future environmental change may pose a serious threat to human development in many parts of the world. In part, this is because the concept of multiple planetary boundaries suggests the possibility of “coupled effects,” where crossing one threshold then triggers serious biophysical changes to some or all of the other eight components of the Earth system.

Unlike the Anthropocene concept, the planetary boundaries idea has not—for obvious reasons—preoccupied geologists. It can be taken seriously without any need to satisfy formal geological criteria for the Holocene’s end. Yet, given its novelty, it has not yet received much serious scientific scrutiny outside the large network of authors represented in the *Nature and Ecology and Society* papers. It has, however, been noticed by the United Nations High Level Panel on Global Sustainability and by leading nongovernmental organizations like Oxfam and the World Wildlife.
Fund. In large part this is because it featured prominently at the aforementioned Planet Under Pressure conference in London. However, members of the American environmental think tank, the Breakthrough Institute, caution against its uncritical acceptance (Nordhaus, Shellenberger, and Blomqvist 2012). Though not practicing scientists, their evidence-based report arrived at two conclusions. The first is that six of the supposed planetary boundaries cannot be said to be global but only regional or local; the second is that there is no compelling evidence that transgressing these six “non-threshold boundaries” would necessarily diminish human welfare. Nordhaus, Shellenberger, and Blomqvist point to the choices and ingenuity humans possess to both adapt to and change their surrounding environments. They argue that people can and must “trade-off” between alternative goals and courses of action. In this light, they insist, “attempts to depoliticize [trade-off decisions] … with reference to scientific authority is dangerous, as it precludes democratic resolution of … [public] debates, and limits, rather than expands, the range of available choices and opportunities” (Nordhaus, Shellenberger, and Blomqvist 2012, 37).

The Anthropocene and planetary boundaries as strategic tools for policy relevant science

In scientific terms, it is unclear how the Anthropocene concept and the related planetary boundaries idea will fare in the years ahead. As we have seen, geologists remain open-minded yet uncertain about the former, while many environmental scientists struggle to agree on an inception date. Meanwhile the planetary boundaries idea is rather too new to have attracted sustained scientific scrutiny as yet. What is clear is that both are intended by their originators to be a scientific means to capture the worldwide attention of politicians, business people, third sector organizations, and the public. As Steffen and colleagues wrote in 2011, “The concept of the Anthropocene, as it becomes more well-known … could well drive a similar reaction to that which Darwin elicited [a hundred and fifty years ago]. There is one very significant difference, however … Darwin’s insights into our origins provoked outrage, anger, and disbelief but did not threaten the material existence of society of the time. The ultimate drivers of the Anthropocene, on the other hand, if they continue unabated … may well threaten … the future existence of Homo sapiens” (Steffen et al. 2011a, 862). Given that these authors have been directly involved in proposing it, the same can be said of the planetary boundaries idea too. Both are intended to provide “an independent measure … of the scale and tempo of human-caused change” (Steffen et al. 2011b, 756–757). They deliberately beg large non-scientific questions about what humans want for the future and how they might practically go about realizing their goals in a world of rapid biophysical change. Science can help address these questions, but only up to a point. Science is thus inserted into vital discussions about what some regard as the ultimate question, namely “How should we live?”

Cynics might argue that the multidisciplinary science networks currently advancing the Anthropocene and planetary boundaries ideas are politically canny in their hunt for future science funding. Science is expensive to pursue and highly competitive between science subjects. It could be argued that the two concepts are
eye-catching vehicles to persuade governments to throw yet more money at global environmental change science in its various forms, keeping the likes of Crutzen, Steffen, and Rockström gainfully employed for years to come. However, a different interpretation is equally, if not more, plausible. Scientists are not only scientists but also citizens, parents, members of local communities, and so on. This means they possess the same wider hopes for, and anxieties about, our future that other people possess. In many cases scientific insight can speak directly to these hopes and anxieties, without entirely dictating their content. This may explain a recent position piece published in the journal *BioScience* by – yet again – Paul Crutzen, writing with a combination of both environmental and social scientists. Entitled “Planetary Opportunities: A Social Contract for Global Change Science to Contribute to a Sustainable Future,” it argues that societies now need “useable research” from the science community that is “solutions orientated” (DeFries *et al.* 2012). The paper is framed explicitly against the “reality” of the Anthropocene and the nine Holocene “planetary boundaries” defined by Rockström and colleagues. It presents a vision of global change science at the service of societies worldwide, akin to what science policy analyst Michael Gibbons and coauthors famously called “Mode 2 research” (Gibbons *et al.* 1994). Such research, these writers argued, does not satisfy merely academic curiosity but is, instead, geared to real world “contexts of application” right from the start. Understood in this light, the ideas advanced by Crutzen, Steffen, Rockström, and allies are important not because they can ultimately be “tested” empirically but because they can frame the future conduct of societally relevant environmental science. The desire for relevance in parts of the environmental science community is currently very high. For instance, James Hansen has been a visible and vocal advocate of strong measures to mitigate future atmospheric warming. Meanwhile, British climate scientists Kevin Anderson and Alice Bows (2012) argue that too many scientists have been far too timid in emphasizing to politicians the severity of forthcoming environmental change.

The contribution of social science and the humanities

How have social scientists and humanities scholars interested in environmental change responded to the idea of the Anthropocene and its collateral concept of planetary boundaries? Some applied social scientists have used them as a (further) opportunity to argue for a new global governance architecture that can build on the not entirely successful efforts of the United Nations to ensure coordinated and decisive transnational environmental management. Notable here is the work of Frank Biermann and colleagues on the so-called Earth System Governance Project which began in early 2009. The project involves some 2000 social scientists worldwide. It is an outgrowth of the International Human Dimensions Programme on Global Environmental Change (IHDP), itself a product of the global environmental change research programs established nearly three decades ago (such as the IGBP). The project, which runs until 2018, aims to study and shape “the interrelated and increasingly integrated system of formal and informal rules, rule-making systems, and actor-networks at all levels of human society (from local to global) that are set up to steer societies toward preventing, mitigating, and adapting to global and local environmental change and, in particular, Earth system transformation, within the normative context of sustainable development” (Biermann *et al.* 2009; see also Biermann *et al.* 2010). According to this vision, social science
arguably becomes the analytical and practical twin of Crutzen et al.’s environmental science. It seeks to create, monitor, and suggest amendments to a new regulatory apparatus sufficient to govern “appropriate human behavior at all scales,” as per Crutzen’s 2002 *Nature* article.

For those persuaded of the scale and urgency of the environmental challenges encapsulated in the Anthropocene and planetary boundaries concepts, this kind of social science endeavor is both necessary and highly valuable. However, some social scientists have expressed concern. For instance, in a recent paper in the journal *Environmental Science & Policy* (Palsson et al. 2013) a mix of anthropologists, historians, and sociologists offer two criticisms of the style of “relevant” Anthropocenic social science advocated by Biermann et al. First, they worry that seemingly “objective” scientific concepts can be internalized unthinkingly by social scientists and then translated into action via policy makers. For instance, they point out that we should “view systematic boundaries as suitable constructs for a mechanistic worldview created by [scientists] … for analytical purposes and projected onto reality, rather than an intrinsic property of the observed world” (Palsson et al. 2013, 7). Second, they suggest that a rush to identify managerial “solutions” to environmental problems identified by environmental scientists may overlook the need to explore the full range of possible societal causes of these problems, and the various ways they may be perceived and felt by diverse real-world actors. All solutions are, after all, relative to how one understands the underlying reasons generative of perceived problems. For instance, a Marxist might say that profit-hungry capitalist corporations are largely responsible for unchecked global environmental change; a neo-liberal economist might say that “missing markets” are the problem; meanwhile, certain feminists might point to the deleterious effects of “masculine” reason which treats the world as a set of objects and relations to be dissected, manipulated, and exploited. In each case, the “social engineering” required to respond to the challenges of the Anthropocene and remain within a “safe operating space” would be very different than those likely to emerge from the Earth System Governance Project. And in each case we might be disposed to ask different questions about how ordinary people want significant environmental change to be tackled.

Underpinning Palsson et al.’s arguments is a clear recognition that the social sciences and humanities produce value-laden concepts, proposals, and evidence. They cannot pretend to exist “above the fray” and need to be explicit about the normative judgments written into both their analyses and proposals for alternative courses of human action. This is for the obvious reason that all values are debatable and require reasoned discussion in order to be accepted as the proper basis for social conduct and environmental husbandry. Palsson et al. are among a number of scholars calling both for more diverse and for more radical conceptions of how to respond to the challenges of our Anthropocenic times. This presents a real opportunity for social scientists and humanists to demonstrate their societal relevance, but in ways that are not narrowly instrumental or unthinkingly reproductive of the political-economic and cultural status quo. Several science policy makers are keen to maximize this opportunity. For instance, the European Union RESCUE foresight project (Responding to Environmental and Social Challenges for our Unstable Earth) created a working group focused on “revolutionary education and capacity building.” Relatedly, in key American science funding bodies (notably the National Science Foundation) “transformative research” is now seemingly de rigueur, and is conceived
ANTHROPOCENE AND PLANETARY BOUNDARIES

in wider terms than simply “transformative technologies” (like nanotechnology).

Aside from its several applied subject areas (like social work), the social sciences have long been productive of radical thought (often in a highly utopian register). This is equally true of the humanities, as they have typically been unencumbered by concerns about “policy relevance” or “stakeholder needs.” This does not render them mere ivory tower pursuits, but ensures they have the freedom to stretch the boundaries of what is thinkable and considered to be possible. Moreover, whereas social science is not terribly visible outside universities, humanistic endeavors certainly are among sections of the wider society (think of literature and poetry, for instance). During the period when the Anthropocene and planetary boundaries concepts have entered scientific discourse the so-called environmental humanities have risen to prominence within (and indeed beyond) the academy. In the broadest sense this involves non-academics and has proceeded in three connected ways. First, many professional writers in the West have been publishing works of autobiography, “faction,” literary fiction, and poetry that focus on a “rediscovery” of the nonhuman and an exploration of the role that “nature” can and should play in our everyday lives. In Britain the writings of Robert Macfarlane (2007) about the relationship of people and land are emblematic of a wider fascination with the affective significance of the material world. Second, in a range of humanities fields – most notably literary criticism – newer (like Macfarlane’s) and older (like Thoreau’s) attempts to rethink human-environment relations have produced a new body of humanistic analysis called ecocriticism (see Garrard 2013). This seeks to shine light on the content, tropes, and aims used by nature writers, photographers, and filmmakers in their attempts to have people reconsider the ways they (de)value the nonhuman. Finally, the environmental humanities have, *sui generis*, been productive of radical new conceptions of how to apprehend and register the significance of the nonhuman. Literary critic Tim Morton’s work is a good example (e.g., Morton 2010), attempting as it does to entirely eliminate the nature-society dichotomy that has for so long structured how most Westerners conceive of reality.

Geographers and the Anthropocene

How have these wider currents of thinking with and beyond environmental science influenced the discipline of geography and vice versa? Thus far, few geographers have either been involved in originating the two concepts being discussed here (University of Arizona’s Diana Liverman and University of Maryland’s Erle Ellis are exceptions) or in responding to them/fleshing out their possible significance. Given their encompassing meaning and implications, one might expect the Anthropocene and planetary boundaries ideas to engender renewed consideration of how physical and human geographers can better work together to study and improve the world. This has not happened yet but is very likely to happen in the near future. In the meantime, it is human geographers of various stripes who have so far paid the ideas most attention (or rather the first idea, since planetary boundaries have scarcely attracted geographers’ attention as yet). Since human geography has both social scientific and humanistic elements, it is no surprise that the geographers in question have offered diverse reflections on what the Anthropocene signifies for geography and/or society (see, for example, the 2014 “forum” organized by Johnson and Morehouse in the journal *Progress in Human Geography*).

In the first place, some have used the Anthropocene as an occasion to pronounce the “death”
of nature, a category that has long vouchsafed physical geography’s identity and which remains a touchstone for public pronouncements about the need to mitigate or adapt to global environmental change. According to Jamie Lorimer (2012), the worldwide ontological mixing of human forces with nonhumans definitive of the Anthropocene has profound intellectual and normative implications. For him it is creating “emergent geographies” that are spatially and temporally varied, often surprising; the implication is that we can no longer talk about a singular, asocial nature to justify various management, conservation, remediation, preservation, or restoration measures (be they large or small). He commends an “experimental ethos” (Lorimer 2012, 600) that is open-minded and reflective, challenging us to make considered, revisable decisions about how we and nonhumans might live together in a thoroughly syncretic world in whose vitality we have a big role in determining. Though Lorimer does not come out and say it, his approach to responding to the Anthropocene is almost the opposite of Biermann et al.’s vision of rule-bound institutions for “earth system governance.” His arguments are offered more in the spirit of different places and countries sharing experience and practice in the face of geographically variegated forms of future biophysical change (see also Lorimer and Driessen 2013).

Lorimer’s arguments are a challenge for environmentally minded human geographers to embrace what Morton has called “ecology without nature” (Morton 2007). They build on over a decade of research into so-called more than human geography pioneered by the likes of British geographer Sarah Whatmore (2002). In a similar spirit, but with a much keener eye on new ideas that key societal decision-makers might take seriously, Karen O’Brien (2013) has challenged geographers like Lorimer to produce “transformative research” of the kind commended by Palsson et al. – indeed she was a member of the RESCUE foresight project mentioned above (see O’Brien et al. 2013). In her view, human (and, by implication, physical) geographers need to unpick the deep and unexamined assumptions they hold about ontology, epistemology, method, and practice. Only then, in her view, can they offer nonacademic stakeholders the eye-opening ideas needed to inspire societies to travel down genuinely new roads as they face environmentally dynamic and possibly threatening futures.

In this light, ideas about the Anthropocene advanced by some more philosophically minded human geographers may appear either too radical or else too abstract to be useful as part of the “axial revolution” O’Brien wishes to see within and without geography. For instance, Kathryn Yusoff (2013) regards the Anthropocene as an occasion to discuss not the ascendancy of humanity over other life forms but the very meaning of the “anthropos.” Yet her arguments are complex, subtle, and rarefied: it is hard to imagine them being understood and acted on by many of her fellow geographers, let alone people outside the academy. By contrast, political geographer Simon Dalby (2013) points to the emergent discourses outside academia that may soon be using the “fact” of the Anthropocene to justify questionable national security and surveillance policies. Pointing to leading governmental, quasi-government, and nongovernmental actors, like O’Brien he asks what alternative, motivational concepts, and aims might be offered to them by geographers and others in the name of a more social democratic and just Anthropocenic future.

For still other human geographers, the likes of Lorimer, O’Brien, and Dalby might be regarded as too fixated on engaging (and trying to change the thinking of) mainstream societal actors, like national government ministers and
departments. Gibson-Graham and Roelvink (2009) argue that human geographers and other social researchers should both study, and actively engage with, everyday communities who want to create truly alternative ways of living. For them the onset of the Anthropocene can and should give such communities further impetus to inspire researchers and various social actors to develop real options outside mainstream thinking. For a Marxist geographer like Jason Moore (2015), however, Gibson-Graham’s arguments may appear to risk ignoring the enduring power of capitalist political economy to limit such options, even as it tries to “green” itself in order to avoid dangerous environmental change. Yet for Gibson-Graham a geographically differentiated grassroots politics lacking worldwide coordination is at least as viable a way to address the challenges of the Anthropocene as the sort of necessary, but current unviable, global anticapitalist politics that Marxists call for.

Conclusion

In sum, the Anthropocene and its collateral term planetary boundaries are arresting terms designed to characterize the human impact on the nonhuman world. It is too soon to say if they will significantly influence the terms of public debate about humanity’s future on Earth. However, they are gaining a degree of traction in Anglophone academia, and beginning to frame discussions of human–environment relations in geography. Given their enormous implications, the key question about the two terms looking ahead is: who will speak in their name, what claims will they make, whose claims will prevail, and what practical measures will follow from those claims? Like other interested analysts, geographers stand to be participants in the future discursive contests that will determine the exact answer to this important question. Though scientists can be credited with inventing and disseminating the two terms discussed in this article, it would be unfortunate if their ideas dominate future debates. By combining scientific, social scientific, and humanistic ideas about Earth’s future in a single disciplinary space, professional geographers and their students may together present the mixture of data and ideas needed to think creatively about a world to come (see Hulme 2008, for a similar argument in relation to anthropogenic climate change).

SEE ALSO: Climate change, concept of; Earth system science; Environmental futures; Environmental science and society; Holocene; Hybridity; Nature; Wicked problems

References

ANTHROPOCENE AND PLANETARY BOUNDARIES


Anthropogeography refers to a mode of systematic analysis, adopted in the late nineteenth and early twentieth centuries, of the geographical distribution of societies, the relationship between migration and the physical environment, and (that for which it is best remembered) the influence of environments upon societies. Although it can be placed on a continuum of a long history of studies of environmental determinism dating back to Greco-Roman times, the term comes from the title of a two-volume work, *Anthropogeographie* (1882 and 1891), written by the German geographer Friedrich Ratzel (1844–1904). Ratzel is best known for being the founder of geopolitics although his influence in shaping late nineteenth century human and physical geography (and anthropology) is increasingly being recognized. Anthropogeography was to influence the work of other prominent early twentieth century geographers, including, in the United States, the student of Ratzel, Ellen Churchill Semple, and Isaiah Bowman, and in the United Kingdom, the political geographer, Halford Mackinder, and in Europe, the German geographer, Karl Haushofer. It was, however, the taking up of Ratzel’s concept of living space (*Lebensraum*) by the Nazis in their justification for territorial expansion for which his work is perhaps still best remembered.

Ratzel’s practical and intellectual engagement in the imperialist politics of expansion is best understood in the political and intellectual context in which he worked at universities in Munich and Leipzig in the late nineteenth century of European imperialism. This political context has also defined the memory of Ratzel’s contributions to geography, serving to undermine that it was also the period in which the divides between natural science, social sciences, and the humanities were being carved out and in which geography as an academic discipline was institutionalized as the study of the Earth’s surface (Lossau 2009). Ratzel’s role in shaping these divides was seminal, involving going beyond the focus on geomorphology that occupied his contemporaries to also examine the life on the Earth’s surface. His work was to lead to the merging of human and physical geography, albeit in a way that was marked by causal explanations, natural laws, and one in which knowledge was derived from the ground (*Boden*) (views which his French contemporary, Emile Durkheim, was to critique), heavily influenced by the rise of social Darwinism, or survival of the strongest societies, and environmental determinism.

Ratzel’s book *Politische Geographie* (1897) is considered the culmination of his voluminous work and that in which he explicitly addressed geopolitics, the analysis of the relationships between geography and the practice of statecraft. As Lossau (2009) has pointed out, there are two competing geographical imaginaries in Ratzel’s work, of containment and expansionism. The first is that of natural lands (*Länder*), or the territory of a nation state which results from the interactions between a specific group of people (*Volk*) who are bonded not through kinship but through their relations to their occupied territory (*Boden*). This is the spatial
imaginary of regional geography that dominated the study of human geography in the early and mid-twentieth century. The second imaginary, central to geopolitics, is based on the view of the state as the biogeographical, or organic, unit that corresponds to its living space (*Lebensraum*). The state, as organism, was subject to laws of territorial growth and development (“the law of growing areas”) and would thus “naturally” seek to expand as its population increased leading to a constant struggle for the conquest of space among states, creating a cycle of progression and defeat, and what can be recognized as a process of uneven development (Natter 2005). Although there is disagreement among scholars as to the extent to which Ratzel’s work had a direct impact on the expansionist policies of the Third Reich, it was complicit in the national socialist ideology of *Lebensraum*. While Ratzel did not speak to the annexation of European countries or the racial superiority of Germans, the political implications of his work on geopolitics and on the practice of war were obvious and he actively developed them through his membership of the Pan-German League in which he advocated the overseas expansion of Germany and the development of its navy (Lossau 2009).

The dual imaginaries in Ratzel’s work complicate a straightforward narrative of environmental determinism; the belief that the environment determines social life has to be understood in relation to the dynamism of *Boden* and *Lebensraum*. While the concept of *Lebensraum* is strongly determinist, tied to the belief that societies are subject to natural laws determined by the land, that of *Boden* speaks more to the physical environment as a causal influence on a *Volk*. The latter, it has been argued, is the basis of the regional paradigm of modern geography and is open to less determinist understandings. However, it is clear that for Ratzel both concepts rely on the environment as a causal force with *Lebensraum* possessing the normative power to provide “an apparently scientific justification for political aims” (Lossau 2009, 146). The debate as to whether this taints his status as a founding figure of modern human geography can only continue. Certainly, his work in geopolitics laid the pathway for subsequent geopolitical theories and, arguably, of the interest in statecraft in critical geopolitics. That his work also falls into the tradition of biogeography (the study of the distribution of species and ecosystems in geographic space and through geological time), with continuities between the earlier work of von Humboldt, and that to come of Vidal de la Blache on *genre de vie*, and opening up questions of the relations between societies and the environment, is notable.

Most recently, the popular but controversial work of the geographer Jared Diamond has re-opened this debate. Although working a century apart, there are marked similarities between the intellectual trajectories of Ratzel and Diamond: both shared an education in the natural sciences (Ratzel in zoology and chemistry and Diamond in physiology and biophysics); both spent time travelling extensively, studying how societies evolved socially in relation to their environment, as a journalist in the case of Ratzel and as a popular science writer in the case of Diamond; and, finally, both turned to the professional designation of geographer. In both their bodies of work it is their understanding of culture that has been critiqued, both viewing it not in the realm of ideas but explained through environmental imperatives.

**SEE ALSO:** Biogeography; Environmental determinism; Geopolitics; Political geography; Regional geography
References


Further reading

By definition, the term *geomorphology* involves studying changes upon the surface of the Earth. Scientists who study these changes, then, are known as geomorphologists. They examine the changes by understanding how geomorphic agents such as wind, water, ice, gravity, plants, animals, and people sculpt the geographic landscape we interact with each day. To better understand all interactions among the various geomorphic agents shaping the landscape, geomorphologists often break the field of geomorphology down into various subfields, each of which describes a particular geomorphic agent. While the myriad of subfields can be confusing, and irritating to some, breaking geomorphology into these various categories clearly illustrates that many agents play a role in shaping the surface of the Earth.

Biogeomorphology is one such subfield of geomorphology that breaks from those traditional agents of change and examines transformative processes upon the Earth’s surface through the unique lens of life itself. It follows then that zoogeomorphology studies how animals exert an impact on the landscape, and much work has been done in this realm since its introduction by Butler (1992). Arguably, however, humans are also animals and the impacts they have rendered on the surface of the planet supersedes any past landscape transformation, denudation or deposition, by other species in the animal kingdom (McNeill 2000). Therefore, the anthropogenic element, or anthropogeomorphology, is yet another subfield within the discipline of geomorphology that is increasing in appearance within scientific literature. Despite the profound role anthropogeomorphic processes have within the scope of geomorphic processes, shaping the landscape more than any other natural process, it is a topic that has received scant attention and nowhere near the amount of research deserved. This lack of attention may stem from the difficulty researchers have in connecting what takes places within the “natural” realm of geomorphology with actions rendered by anthropogeomorphic processes.

Geomorphology and anthropogeomorphology alike are best understood knowing that the changes rendered upon the surface of the Earth come about from a collective body of geomorphic agents. These agents can remove material through the process of erosion, or they can add material through depositional processes. For example, heavy winds may pick up dust from a desert surface yet at some point the wind will slow down and deposit that dust at a downwind location, perhaps into a stream where that material will get carried into a body of water such as a lake or ocean where it settles to the bottom. Very rarely does one geomorphic agent act alone in sculpting a particular landscape. Instead, a variety of forces are at play, both eroding and depositing materials in a constant ongoing process. Anthropogeomorphologists study these agents as well as connecting the direct and indirect activities of humans as a geomorphic agent. Humans act as a direct geomorphic agent when they themselves are directly involved in the processes of erosion and deposition; activities such as...
excavation, mining, canal and dike construction, road building, and cratering due to warfare are all examples of anthropogeomorphic landforms. Indirectly, however, humans are also intimately related to the more traditional “natural” geomorphic agents such as wind and water. It is these indirect modifications, associated with the technology we have created to modify our surroundings, intentionally or not, that have the most profound impact of all on the landscape. For example, when humans clear a forest for agriculture, the accelerated rates of water erosion are indirectly related to the change in land cover created by human economic activities. For the anthropogeomorphologist, finding changes on the Earth’s surface are not difficult and indeed, what may prove more challenging is finding what geomorphic agents today are not related to some type of human activity.

Humans have always excelled at modifying their environmental surroundings. From the dawn of humankind until present, humans owe their overall success on this planet to their innate ability to somehow bend nature to their own will. Researchers in the archaeological community have even put forward theories that both the human and Neanderthal species at one point coexisted together before Homo sapiens pushed them out of existence (McNeill 2000). With a larger body size, Neanderthals required more calories than humans and therefore required larger amounts of food than humans. In times of environmental stress such as drought, humans were able to outcompete the Neanderthal by consuming what foods were available, literally depleting the land of its resources before moving on, and thus starving Neanderthal out of existence, in what would only prove to be the beginning of humans as the ultimate agent of landscape modification.

Eventually, being able to use the tool of fire, humans began to shape environments to produce both the wildlife and forage resources they desired (Mann 2005). What impact this had on sedimentation rates and the like is difficult to determine, yet it can be assumed that even in their beginnings, humans were playing a role as a geomorphic agent. Research also shows that it was this seminomadic lifestyle with modification of the surrounding environment that first led to the beginnings of the agricultural revolution (Diamond 2005). Although the numbers are debated and difficult to determine, researchers have placed the population of humans prior to the onset of the agricultural revolution at just around 1 million worldwide. With the onset of the agricultural revolution approximately 8000 years ago, the population of humans began a steady increase and with that increase in population came the evident abilities of humans to sculpt the Earth’s surface. Sediment records from thousands of years ago show that when humans cleared land protected by forest and grassland for more agricultural production to support their ever burgeoning and expanding populations, they accelerated erosion on hill slopes and increased rates of deposition in the valleys below (Montgomery 2007). In some cases, the accelerated erosion on soils formerly held together by forest cover was amply recognized and terracing was implemented as a means to prevent erosion. In many areas of the world, terraces remain in agricultural production and are an anthropogeomorphic landmark that testifies not only to what humans are capable of as direct geomorphic agent, but also that soil conservation measures are a viable practice when performed properly. Terracing, however, reflects human modification to the landscape as growing population is pushed out of prime agricultural valley bottom lands and into the submarginal soils of the upland. Terraces are also very labor intensive to maintain and are an exception rather than the rule when it comes
to past civilizations deforesting uplands in order to clear land for agricultural production. Grain harvest records, as well as sediment cores for a variety of depositional environments, attest to the number of great civilizations that have fallen due to intensive agriculture practiced on formerly forested hill sides. Montgomery (2007) provides a detailed perspective on how Western civilizations, including the Greeks and the Romans, were crippled by the massive amounts of soil erosion rendered by intensive and unsustainable agricultural practices.

Accelerated soil erosion due to intensive agriculture and forestry activities are not relegated to ancient civilizations and continue to shape the modern landscape. If anything, the widespread implementation of industrial scale farming and forestry operations has led to the removal of billions of tons of top soil around the world. Although this is a global phenomenon, nowhere is this more apparent than in tropical environments where humans have tapped one of the last agricultural frontiers to meet the food demands of a global population pushing over 6.8 billion. Formally forested hillsides have been put into agricultural production in environments subject to intensive rainfalls and soils weakly held together. Not only does this increase the sediment loads on streams, it also weakens the slopes and sets the scene for catastrophe. This was demonstrated in 1998 when Hurricane Mitch stalled out over the highlands of Central America and triggered landslides that killed nearly 20,000 people. Here, anthropogeomorphic processes were not directly responsible for the deaths, but the indirect and unintentional use of technology to engage in intensive forestry and agricultural practices led not only to massive amounts of material transported in slope failure, but in catastrophe as well. Hurricane Mitch was not unique either in that slope failures like this, as well as siltification of reservoirs and river channels around the world due to intensive agriculture, are only increasing.

Erosion from economic activities such as forestry and agriculture may be profound, but nowhere near as visible as what humans have done to reshape the Earth’s surface drainage networks. From the beginnings of the agricultural revolution until present, humans have transformed the way in how water moves on the surface of the Earth. This has been accomplished through dam construction, water diversion projects, levee creation, dredging operations, reservoir creation, drainage of wetlands, and shoreline modifications. To view the Earth from space would reveal a vast array of hydrologic modifications, many of which were completed over the course of the twentieth century. The construction of dams has allowed for the creation of reservoir style lakes in river valleys around the world. Some of these lakes, such as Lake Mead in the United States and the Three Gorges Dam reservoir in China, cover hundreds of square miles. Although profound, anthropogeomorphic creations like these may pale in comparison to the indirect ramifications of large scale canal projects. The Panama Canal connects the Pacific and Atlantic Oceans while the Suez Canal connects the Mediterranean with the Indian Ocean. If someone wanted to, they could sail up the Saint Lawrence River into the Great Lakes and then connect to the Mississippi via a series of canals and waterways locks through the Chicago River. The Chicago River, one of the largest overlooked engineering projects of the twentieth century, used to flow into Lake Michigan, but now has reversed its flow into the Mississippi. Diversions like these have ramifications well beyond changes in water levels; new corridors open up for transit of plants and animals whose abilities to act as geomorphic agents are profound and only beginning to be understood (Butler 1995).
Human hydrologic engineering has led to the creation of inland seas as well as the disappearance of others. The Salton Sea, formally the Salton Sink, in southern California was created when the Colorado River flooded in 1906, changing its course into a series of canals that were dug by some enterprising farmers attempting to divert waters from the Colorado to irrigate their crops in the fertile but dry soils of the Mojave Desert. Although shrinking in size due to increased urban demand for water in cities such as San Diego, the Salton Sea remains as an example of what direct anthropogeomorphic activities are capable of producing, intentionally or not. Another inland sea that is shrinking in yet another arid landscape is the Aral Sea that borders what is now Uzbekistan and Kazakhstan. Unlike the Salton Sea, however, the Aral Sea was not created due to anthropogenic mishaps and up until the 1960s was an inland sea the size of Belgium. The sea is located in arid landscape fed by rivers carrying snow melt off the Himalaya Mountains. Prior to the diversion projects that led to its demise, the sea was so large it buffered the temperature extremes of the otherwise harsh conditions of interior southern Asia. This benevolent climate attracted the attention of what was then the Soviet Union and the waters were diverted in a massive undertaking to grow cotton. Cotton production did ensue, but not without unforeseen consequences; the sea is now one tenth of its original size, the former fish population is decimated, the climate is no longer moderated by the sea, and toxic chemicals blow off the former surface – poisoning the remaining shell of a populace surrounding the Aral “Sea.”

Beyond hydrologic engineering projects, another anthropogeomorphic process quite visible from space would be landforms associated with excavation activities. Like agriculture and hydrologic engineering, anthropogenic landforms rendered by excavation extend back several millennia. Pits and mounds associated with our quest to obtain mineral resources from the Earth literally cover the planet, getting bigger and bigger with each technological advance. Interestingly enough, it was our quest to harness geologic resources that truly unlocked the human potential to become the most significant geomorphic agent of them all, and that was when we harnessed the power of fossil fuels.

Fossil fuels in the form of coal, oil, and natural gas are really nothing more than stored up energy from the sun. They represent energy taken in by plants for millions upon millions of years, concentrated in layers that humans harvest to unleash that energy how we see fit. By unlocking this power, we removed ourselves and our animals as the primary source of power for labor and were able to develop technology capable of acting as geomorphic agents never before thinkable. In addition, unleashing the power of fossil fuels also generated yet another surge in population in the form of the industrial revolution. In the agricultural revolution, population did grow, but at a linear rate. With the industrial revolution, human population began to grow exponentially. This exponential rise in population, combined with technology capable of moving more Earth than ever before, has left a significant mark on the landscape in the form of excavation activities (McNeill 2000). The footprints left in our quest for mineral and energy resources are everywhere. From the mountains of coal waste on the plains of Belgium, to open pit diamond mines of South Africa, to the widespread desolation of the tar sand mining operations of Northern Alberta province in Canada, the direct anthropogeomorphic excavation agent of change is readily apparent.

Yet another anthropogenic excavation agent, albeit not readily visible from space but very much widespread, are the billions of craters on the Earth’s surface rendered by explosive
munitions in time of war. So distinctive is the disturbance to the landscape surface associated with explosive munitions, that the geomorphic process is termed bombturbation (Hupy and Schaetzl 2006). Bombturbation is common and widespread across all parts of the globe, usually occurs in association with warfare, and the areas affected are spatially concentrated, for example, on major battlefields. For instance, on some World War I battlefields, over 50 million craters have been produced across a several hundred hectare area in a matter of a few months. Bombturbation is usually a cratering phenomenon with the explosion leaving behind a pit that is variously excavated of soil and underlying parent material with an accompanying rim of debris nearby. The event could then be considered both a denudative and depositional geomorphic process, although the displacement of materials is much more concentrated in the crated area, therefore making bombturbation mainly a denudative geomorphic agent (Hupy and Koehler 2012).

Quantifying the amount of material displaced by exploding munitions is, in large part, nearly impossible to tabulate; during World War II alone, 1.4 million tons of bombs were dropped on Europe and 557,000 tons of bombs were dropped on Germany by American heavy bombers (Hastings 2012). This number pales in comparison with the 14 million tons of bombs dropped over Indochina in the 8 years of US involvement in the Vietnam War (Westing and Pfeiffer 1972). Estimates regarding the number of munitions deployed in warfare are often conservatively based as it is difficult to keep accurate records of the number of weapons deployed in warfare. However, these conservative estimates easily place the amount of soil displaced by explosive munitions over the course of the twentieth century into billions of cubic meters. In Vietnam alone, Westing and Pfeiffer (1972) estimated, using air force bombing records and average 500 bomb crater sizes, the amount of topsoil and subsoil displaced to exceed $3 \times 10^9 \text{ m}^3$. Thus, the scope and magnitude of bombturbation is so immense, and the degree to which it can impact soils is so catastrophic, that it justifies singling it out as a major, singular anthropogeomorphic agent. Beyond bombturbative activities, warfare in itself represents a moment in history when many anthropogeomorphic agents are magnified in intensity. Because waging warfare is so costly in the form of resources, many of the anthropogeomorphic activities previously discussed in this chapter are being practiced at an increased level. The need for resources in warfare may also create conditions where, in the war effort, a nation might lift restrictions meant to reduce a particular negative geomorphic impact, such as in mining and agriculture activities. Warfare, one of the more unique forms of anthropogeomorphic disturbance is one such agent that has not received attention until recently. More research is needed in this field to truly appreciate the direct and indirect geomorphic implications of war.

In a world that contains 6.8 billion people, there are but a scant few places on the Earth’s surface that remain untouched by anthropogeomorphic forces. Few can argue that direct anthropogeomorphic activities have not proven themselves as a significant geomorphic agent. Modern geospatial technology in the form of satellite imagery, geographic information systems (GIS), and digital elevation models (DEMs) have made it possible to quantify just how profound this anthropogeomorphic footprint on the landscape really is. Additionally, with more geographers, environmental historians, archaeologists, geomorphologists, and other related scholars exploring our past, the scientific community is gaining insight into the role humans have served as a direct geomorphic agent in history. Humans
have proven themselves ample geomorphic agents by moving, via mining operations alone, more material over the course of the twentieth century than all other natural geomorphic agents combined (McNeill 2000). While mining may be one of the more obvious forms of an anthropogeomorphic agent that has received ample attention from the geomorphic community, other anthropogeomorphic agents lack considerable attention. What remains to be determined is how much anthropogeomorphic forces play an indirect role as an agent of landscape change.

Indirect geomorphic processes will prove challenging to the anthropogeomorphologist in future years. Linking what many would consider to be a natural geomorphic event or an “act of God” to anthropogenic activities proves difficult to establish causality, or even how human activities may have magnified an event. However daunting the task of linking anthropogenic activities to the larger geomorphic realm, significant progress has been made thanks to GIS technology and an increasing awareness of how significant humans are in shaping the surface of the Earth. One research frontier where much remains to be done is in the indirect consequences of climate change as an anthropogeomorphic agent. With time, research in this relatively young subfield of geomorphology will unlock many of the questions anthropogeomorphologists are currently asking.

SEE ALSO: Geographic information science;
Geographic information system;
Geomorphology: history

References

Antiracist geography

Sharlene Mollett
University of Toronto Scarborough, Canada

Antiracism is a political discourse and intellectual commitment to the challenge of racism, essential for a cogent movement that seeks equality as a principle component of any society. It is informed by anti-essentialist understandings of race, which unsettle constructions of race as a naturalized social hierarchy of biologically separate human populations. Moreover, antiracism, whether practiced by civil society or as part of a particular state mandate, is a contestation of the authoritarianism of the state, and its governance and distribution of the benefits to a purported citizenry; it seeks to shape a morality that questions, and targets for change, the ways in which the state and its elites “other” its populations.

As well, antiracist scholarship attends to how processes of racialization, which assign people to particular racial categories, have material consequences including the inequitable distribution of material resources, spatial meanings, representations, and various forms of power. This popular movement and intellectual narrative, although seemingly concentrated in the United Kingdom and North America, has a global reach.

With beginnings in abolitionism, anticolonialism, and civil rights mobilizations in the United States, Great Britain, and South Africa, antiracist movements are also rooted in the post-World War II effort that ultimately linked multiple racial projects from concentration camps and transfer stations at Auschwitz and Terezin to slavery and segregation in the United States, and racial apartheid in South Africa. Indeed, antiracist geographers agree that there is space for antiracist dialogue across national boundaries. For instance, in Latin America, geographers have demonstrated how critical racial analysis is helpful for understanding colonial and postcolonial relationships – both past and present – and are in conversation with Latin American postcolonial and decolonial scholars working with similar frameworks around colonialism.

Antiracist scholars, however, are divided as to whether antiracism is overly influenced by the racial histories and experiences of the United States, making the field vulnerable to calls of imperialism, due to the fact that many of the conceptual and contextual insights for struggle are shaped by the US context. For instance, Bourdieu and Wacquant’s (1999) piece “On the Cunning of Imperialist Reason,” problematizes US generalizations of the black−white binary that presume “universal” understandings of race and racism and the ways these meanings travel and are inappropriately applied to other contexts. They argue that “cultural imperialism rests on the power to universalize particularisms linked to a singular historical tradition by causing them to be misrecognized as such” (1999, 4, cited in Bonnett 2006). But geographer Alistair Bonnett challenges this claim. Simply because much of the knowledge produced by antiracists builds on US experiences does not foreclose creative applications of this knowledge (and the advance of independently produced ideas) in other contexts. Furthermore, for Bonnett, antiracist projects and imperial projects cannot be mutually exclusive. He insists that critical scholars must “give sufficient attention to the close relationship that can exist between ‘resistance’
and hegemony” (Bonnett 2006, 1091) or in this case antiracism and imperialism. He explains this convergence with the example of how US theorizing is doused in the palatable lexicon of multiculturalism and diversity in such a way that it reifies and naturalizes difference, which serves the state and its racialized governance, even when the state’s goal is seemingly to eradicate racial inequality. The naturalization of race in embodied difference can embolden more conservative positions, particularly in the ways in which multiculturalism can be co-opted in racist ways through neoliberal projects.

Scholarly engagement with the concept of whiteness has become a useful tool to denaturalize race. The increased attention and debate given to the concept of racialization, namely, the “process by which racialized groups are identified, given stereotypical characteristics, and coerced into specific living conditions, often involving social/spatial segregation and always constituting racialized places,” has been a key way in which whiteness is made visible in geography (Kobayashi and Peake 2000, 393). The spatial processes of whiteness are operationalized through banal and everyday practices and not simply aberrant violent events motivated by hate. This unveiling and centering of the “normative, ordinary power to enjoy social privilege by controlling dominant values and institutions” (Kobayashi and Peake 2000, 393) through whiteness leaves little doubt among critical scholars that “racist expressions are normal to our culture, manifest not only in extreme epithets but in insinuations and suggestions, in reasoning and representations, in short, in the micro-expressions of daily life” (Goldberg 1997, 21, cited in Kobayashi and Peake 2000, 393).

These revelations have been taken up by geographers through a number of projects that include such topics as environmental racism, food justice and security, gentrification, land conflicts, international development, migration, agricultural development, music, and labor rights. Through this work, antiracist geography critically examines how, in the words of Bonilla-Silva, “racial domination generates a grammar that helps reproduce racial order as just the way things are. Racial grammar is a distillate of racial ideology and, home of white supremacy” (Bonilla-Silva 2012, 174, emphasis in original). Antiracist geography looks to counter such universalism by making visible how whiteness distorts intellectual production within the discipline of geography. Given the white majority among geographers (and the active peripheralization of critical racial studies in geography, until more recently), the motivation for change moves at a snail’s pace; however, among antiracist geographers there is a growing urgency to highlight not just whiteness as a social condition of privilege but white supremacy hidden in the walls of institutions, the practices of scholars, and what is taught as knowledge. Such antiracist theorizing is relevant to the much needed decolonizing of the discipline of geography.

Another important discussion shaping antiracist intellectual discourse and movements is the feminist antiracist tradition, particularly marked by the Canadian experience, where indigenous women have critiqued Canada as a white settler society and where an active and contemporaneous colonial process unfolds. In a similar way, antiracist black feminist geographers cogently remind us that colonial histories and histories of enslavement set the terrain for contemporary racism, woven as they are into the white settler colonial project. While such conceptualizations unite feminist antiracist projects, tensions exist as well. Canadian indigenous scholars insist “that something is deeply wrong with the manner in which, in our own lands, antiracism does not begin with, and reflect, the
totality of Native Peoples’ lived experience, that is with the genocide that established and maintains all of the settler states within the Americas” (Lawrence and Dua 2005, 121).

These claims not only have implications for the intersectional nature of antiracist feminist traditions everywhere, but also give reason for taking a closer look at how various kinds of oppressions intersect, with particular attention paid to the ongoing colonial project. As critics assert in the Canadian context, rather than contesting the ongoing white settler colonization of indigenous Canadians, antiracism scholarship furthers the colonial agendas of the present and requires increased interrogation of the question of land as a site of struggle. Furthermore, decolonization struggles must be seen as foundational, rather than as one factor in a larger antiracist struggle. Yet, it is also argued that while increased attention to indigenous knowledges and the experiences of indigenous peoples in relation to white settler societies have become more visible in geographic scholarship, some question whether these studies are uniquely placed to reinforce antiblack sentiment in geography. Increasingly, however, antiracist geography encourages studies to explore multiple and intersecting sources of oppression that are shaped by the logics of white supremacy as a way to inform our intellectual and more plural antiracist knowledges.

Overall, such debates are productive. Like all social movements, the fate of the future of antiracism theorizing and practice will be shaped not only by the spaces outside the movement, but by the attendant struggles within.

SEE ALSO: Apartheid; Critical geography; Environmental (in)justice; Feminist geography; Ghetto; Human rights; Imperialism; Indigeneity; Necrogeography; Postcolonial geographies; Power; Race and racism; Residential segregations; Violence

References


Apartheid

Maano Ramutsindela
University of Cape Town, South Africa

Apartheid manifests perceptions, views, and habits of mind that crystallized into an ideology of the Afrikaners that was subsequently pursued through the modern South African state. The main constituent parts of apartheid are the idea, policy, and measures that worked together over time to shape the South African society. Each of these parts is important for understanding apartheid as an ideology and program of action, and its residuals in a democratic South Africa. As an idea, apartheid is underpinned by and also expresses a racist ideology. The racist ideology encapsulated in apartheid is not distinctively South African but “reflected and grew out of already existing notions of human difference” (Dubow 1992, 210). It was influenced by German missiology; a local secular intelligentsia highly connected to the international scholarship of racial thought; and the segregationist practices of the American South. The tenets of apartheid are consistent with those of racial segregation in Europe and the United States. It is for this reason that apartheid in South Africa is understood as a perfection and domestication of the same racism that underpinned colonialism. On the ideological front, the division between apartheid and colonialism is superficial hence apartheid is sometimes referred to as colonialism of a special kind. Unlike British or French colonialism that was driven from distant metropolitan centers, apartheid was colonialism from within as both the colonizer and the colonized lived in the same country. This condition both sharpened the racist ideology of apartheid and provided it with a rationale distinctive from existing forms and content of racial segregation (Dubow 1992). Apartheid ideologues saw geographical proximity of the races and the potential racial relations that might emerge as an obstacle toward the ideals of racial purity founded on biological theories of racial superiority. The black majority and the business acumen of Indian immigrants also threatened white political and economic power, respectively. Under these conditions, apartheid hardened into a permanent solution to race relations at a national scale. While the threat to white domination was generally coded as the “native question,” Afrikaner nationalists conceptualized apartheid as an “answer” to that question. It became a unique formula for the regulation of human relations and a blueprint for the future of the country and its peoples. Historians differ on the date on which the term apartheid was first used by Afrikaners but there is general agreement that it entered into the body politic in the 1940s.

Apartheid is a dynamic idea which relied heavily on a bundle of rationales that were mobilized at appropriate moments in its development. Its core pillars remained unchanged. These include the inequality of races in which whites appear as inherently superior to their black counterparts; the trusteeship of whites over blacks; a complete and permanent separation of races in the cultural, economic, social, political, biological, and territorial spheres of contact. In its South African context, the apartheid idea “embodies numerous unwritten social codes and customs [and] … its fundamental core is an attitude of mind peculiar to the Afrikaner” (Rhodie and Venter 1959, 19 – 20).
APARTHEID

Although there were disagreements among Afrikaners about the scope and pace of segregation and the measures required to achieve its goals, the South African white’s views on color informed the apartheid policy, whose objectives were threefold. First, the policies were designed in order to convey conceptualizations of race and beliefs rather than to assert these openly. Second, the policies were meant to protect, preserve, and develop the heritage of the Afrikaner in all spheres of life. Third, they sought to defend racial superiority by systematically reversing and also undermining the development of black people.

In so doing, apartheid policies sought to construct and present black inferiority as a “natural order” in the racial hierarchy they projected. This objective explains why it was necessary for the apartheid state to design an inferior education for the black population. Biological theories of racial superiority found expression in policies against mixed marriages while the need for separate territorial units was articulated through policies on land, urban planning, and the Bantustan. A major step in the geographical separation of races was the passage of a century-old Natives Land Act of 1913 through which land for occupation, use, and purchase was set aside on a racial basis. The geographical separation of races in the form of Bantustans was one of the most contentious and difficult step to realize.

The implementation of apartheid policies led to the development of two main bodies of work – one feeding into and guiding apartheid measures and the other exposing human suffering and indignity. Most Afrikaner institutions were responsible for producing the bulk of pro-apartheid literature in which theology played an important part. Liberation theology developed as a counterweight to apartheid missiology. A strong body of anti-apartheid literature drew its theoretical strength from Marxist schools of thought, Black Consciousness, and postcolonial literature, and fed into anti-apartheid political thought nationally and internationally. It played an important role in connecting South African experiences with those in other countries and as the basis for the international condemnation of apartheid as a crime against humanity. The focus on the net effects of apartheid policies led to the redefinition of apartheid as a system of racial segregation, oppression, and exploitation rather than merely an idealism of Afrikaner nationalists. Thus, the notion of apartheid can be viewed from the vantage point of Afrikaner idealism and from the perspective of the victims and the broader anti-apartheid movement.

Academically, the implementation of apartheid led to the development of two useful terminologies, namely, the apartheid state and the apartheid city that embody the South African experience. The apartheid state as a state formed through and driven by a particular version of an overt racist ideology with clearly formulated goals and a program of action under a white minority government is distinctively South African. Similar ideologically driven states operate under entirely different conditions with their own objectives and programs of action. The phenomenon of the apartheid city in South Africa has its segregationist foundation in British colonialism (Christopher 1994).

The end of formal apartheid is associated with the collapse of the apartheid state in the early 1990s; the loss of political power by the National Party that ruled the country between 1948 and 1993; and the formal abolishment of apartheid laws that gave legislative power to apartheid policies. These developments gave rise to the notion of post-apartheid, which is commonly used in literature to signify a period after the end of formal apartheid and to refer to the post-1994 democratic state. The morpheme “post” in post-apartheid is silent about the apartheid idea and the conceptualizations of race it embodies.
There is evidence that a section of Afrikaners still cherish the ideals of apartheid and that racism in democratic South Africa continues in various forms. The existence of an exclusively Afrikaner state, the *volkstaat* (own homeland), in post-1994 South Africa is an expression of the same ideals of separate territorial units for the different races on which apartheid was based.

Although South African geographers emulated the intellectual trends in the Western world, the apartheid experience gave them an avenue through which they could raise new sets of research questions. As a community of scholars, geographers in South Africa were divided along the same racial lines as the rest of society. They used their professional skills to either support or oppose the apartheid project; sometimes standing on the fence or changing positions. Much geographical work on apartheid has been on apartheid policies and measures with scant attention to its ideological foundations and its connectedness to world systems of domination and segregation. A strand of critical human geography emerged in the 1970s and 1980s to expose the hardships under apartheid, while also trying to sensitize the discipline about anti-apartheid struggles. This scholarship dissipated with the dawn of democracy in 1994. We noted that land was central to the apartheid project. While geographers have researched land issues, they have yet to show a sustained engagement with questions of land. Land issues remain critical to the economic, social, political, and environmental geography of democratic South Africa.

As a universal signifier of political and racial domination, apartheid has gone full circle. It is increasingly being used to analyze the political, civil, economic, social, and cultural strangulation in other countries, including those from which it derived its motivations. Its original meaning has been expanded to include multiple forms of discrimination and segregationist practices in the academe, health, and so on. Such practices are used to describe other states as apartheid. It is also currently used to denote inequalities in national economies and in the international political-economic system. It is from those inequalities that notions such as “global apartheid” and “apartheid economy” are derived. The multiple uses and meanings of the term apartheid serve as a global repository of the South African experience but also have the potential to create much confusion on the core elements of the term. The potency of these meanings lies in the common understanding of apartheid as an evil system and on the emotions the term evokes. Future research on apartheid will need to disentangle the mutating ideology of racism to understand various forms and texture of apartheid it generates. This is a necessary condition both for a proper exposition of segregation at various levels and contexts and elimination of racist practices. Developments in post-1994 South Africa are important fodder for postcolonial theory.

**SEE ALSO:** Antiracist geography; Colonialism, decolonization, and neocolonialism; Difference; Race and racism; Underclass theories; Whiteness

**References**


Further reading


Applied geomorphology is a branch of science within the broader discipline of geomorphology that focuses on geomorphological landforms and processes of societal concern. Specifically, it involves the application of geomorphology to answering research questions of geomorphic significance to society and finding solutions to a variety of human–environmental problems influenced by geomorphic forms and processes. Applied geomorphologists provide theory, data, and analysis to answer questions such as: What is the risk that a landslide will occur on this hillslope? Why did my house flood when the insurance-rate maps place my home outside of the flood zone? What is the most suitable location for a new bridge crossing? How much sand and gravel can be mined from this deposit? How can we reduce the impact of coastal storm surge and prevent beach erosion? And how much water should we release from a dam to scour pools and restore habitat for native fish?

Problems addressed by applied geomorphologists include landforms and processes representative of the major research areas within geomorphology including, but not limited to, coastal, fluvial, karst, arid, tectonic, hillslope, mountain, glacial, and tundra environments. The theory, science, methods, and techniques from these fields are built on a foundation of basic process knowledge which has a direct relation to solving problems in the field of applied geomorphology. Although there is not a standard text on the practice of...
weathering of stone buildings and monuments, and recommendations to prevent further decay and restore architectural integrity.

Geomorphologists are well suited to contribute solutions to these problems because they are trained to read and explain the physical landscape by identifying landforms, describing geomorphic processes, and applying site-based knowledge to interpret how the geomorphic landscape is functioning. They can identify what elements of a landscape are changing, how they are changing, and what is causing them to change. They can answer questions about whether certain changes are normal – that is, the result of natural process dynamics – or are a response to human impacts. Geomorphologists are able to translate this information to explain the historical and contemporary geomorphic context and predict future event scenarios or patterns of change that may concern societal decisions or actions.

**Geomorphic maps**

An important application of geomorphology is the creation of place-based geomorphic maps as dynamic templates representing physical landscape changes. Theory and basic science inform the interpretations of landforms, while site-specific investigations of local complexity and the interdependence of multiple factors (e.g., historical contingency, feedbacks, thresholds, lags, multiple causality, nonlinearity, human impacts, etc.) inform the process dynamics influencing the direction, magnitude, and rate of geomorphic changes. Useful geomorphic maps and unit descriptions will: (i) represent contemporary and historic landform positions; (ii) show change over time relative to different environmental and human conditions influencing the geomorphic landscape formation; and (iii) discuss the various causes of change and factors affecting landscape stability or instability (Smith, Paron, and Griffiths 2011).

These map elements are important for illustrating the spatial relationships of various geomorphic forms and explaining the direction, magnitude, rate, and causes of geomorphic change over time. For example, a river channel may be mapped with varying channel positions to provide an indicator of bank stability. In other examples, the placement of different hatch marks or symbols will be used to show changes associated with different landforms and processes, including coastal shoreline erosion, eolian sand dune migration, fluvial accretion of contaminated sediments, or glacial ice retreat. Such maps characterize the geomorphic forms and processes, enabling users to identify areas of the landscape that are more or less stable across varying temporal and spatial scales.

Historically, most geomorphic mapping was accomplished in the field and supplemented with historic aerial photographs, plan table surveys, and topographic maps. Methods for mapping and modeling landforms and physical processes have improved significantly over the past few decades and have fundamentally influenced the spatial extent, resolution, scale, and cartography of geomorphic representation used for applied geomorphology (Smith, Paron, and Griffiths 2011). Technological advancements in topographic mapping including high-resolution differential GPS (Global Positioning System) surveys, total-station surveys, LiDAR surveys, and digital orthophoto imagery have fundamentally altered how and what geomorphologists are now capable of mapping. Incorporating these data within a geographic information system (GIS) has improved the processing and accuracy of digital terrain models (DTMs) and digital elevation models (DEMs) used in geomorphic mapping, providing faster and more efficient
services to users, including planners, engineers, consultants, and managers.

Not only are geomorphic maps important for interpreting physical forms, processes, and patterns of landscape evolution, but they also provide important proxies for locating resources of economic value, identifying soil types, mapping contaminant distributions in the soil and sediment (both in transport and storage), and examining underlying physical controls on biological process and pattern interactions. For example, in the United States, geomorphic and hydrologic data derived from DTMs and aerial imagery is used to improve the National Wetlands Inventory’s wetland maps through the creation of a hydrogeomorphic (HGM) wetland classification system. The Natural Resource Conservation Service (NRCS) uses the HGM scheme to develop wetland functional assessments which are required under the NRCS Wetland Protection Policy for making decisions relative to exemption, mitigation, or restoration requirements pertaining to wetland environments and development actions.

**Applied geomorphology contributions to problem-solving**

Applied geomorphologists will often begin their research with two fundamental pathways of inquiry. One is concerned with examining how geomorphic processes and landforms affect society, and the other with examining how human actions affect geomorphic processes and landforms. These are often linked, in the sense that one will have direct impacts on the other, and thus it is critical to consider them together. The following sections are organized to show the contributions of applied geomorphology to problems and services involving hazards, land use, natural resources, environmental management, and climate change, using generalized examples from different fields of geomorphology and considering the integrated perspectives of geomorphic impacts on society and society’s impact on geomorphic forms and processes. The use of geomorphic maps, models, and predictive tools within applied geomorphology are discussed in relation to the different major areas of contribution.

**Hazards**

Geomorphologists’ skills in interpreting landforms and processes give them the ability to identify natural hazards associated with different terrains that may threaten the safety and health of society. Applications of geomorphic maps, models, and risk assessments are critical to hazard prevention and represent one of the primary tools used in the diminution and prediction of hazards, particularly those associated with floods (on rivers and coasts) and mass wasting (landslides, debris flows, avalanches, coastal cliff erosion, etc.) (Smith, Paron, and Griffiths 2011). In the aftermath of a hazardous event, geomorphologists will conduct postevent surveys to quantify the changes to the land surface. Knowledge gained from postevent surveys improves our understanding of the threats posed by different hazards and can help society to prepare for future events.

The use of LiDAR technology has revolutionized the DTMs used in delineation for identifying flood-prone zones associated with rivers and coasts. A new generation of LiDAR-derived DTM flood maps are being produced by the United States Federal Emergency Management Agency (FEMA) through its flood hazard mapping program which is part of FEMA’s Risk Mapping, Assessment and Planning (MAP) division. In many locations, these updated flood hazard maps increase the
Land area mapped within floodplains and coastal environments and, as a result, improve the forecasting and predicting of flood risks for extant communities and help prevent future risks in undeveloped locations.

Predictive landslide mapping and modeling has been the responsibility of geomorphologists for decades, and will continue to be important as society continues to develop mountainous terrains, coastal shoreline cliffs, and other areas where the potential for mass wasting is greatest. GIS spatial modeling capabilities have made developing regional landslide risk assessment much more efficient. Overlays of DTMs, slopes, drainage networks, geology, soils, and vegetation are analyzed for spatial relationships of risk potential and then extrapolated to predict locations of greatest landslide vulnerability. In many instances these models are calibrated with conditions from historic landslide events, providing a more accurate representation of hazard risk assessment. Similar GIS-based spatial analysis applications can be used for predicting avalanches, mudslides, subsidence, gully formation, and soil erosion.

An area of consistent relevance of applied geomorphology to hazards involves pollutant and contaminant mapping. Historically, there was a great interest in mapping heavy metal contaminants associated with nuclear testing. While this is still very important today, most contaminant mapping focuses on contaminants produced from mining activities associated with the extraction of precious minerals, removal of mountain-top coal, and drilling by hydraulic fracturing.

**Land-use planning**

Information from geomorphic maps and hazard assessments provide important information for land-use planning. These data can be used as a component of suitability models for determining the best locations for different types of land use in agricultural, urban, and rural settings. Sustainable development of urban areas requires detailed knowledge of local and regional geomorphology, particularly in cities undergoing rapid urbanization. Geomorphic maps and associated hazard risk assessment models, particularly those that illustrate land surfaces by gradients of hazard vulnerability or morphologic instability, are useful for making decisions on the placement of transportation networks, neighborhoods, socioeconomic centers, retail markets, places of recreation, conservation easements, and energy, water, and sanitation infrastructure.

Also of concern is how the placement of different types of urban land use and their associated impacts affect geomorphic processes and cause landform change. For example, the development of a neighborhood with a high level of impervious surface may lead to increased flash flooding and channel and bank erosion. These negative effects may increase the frequency and magnitude of downstream hazards. The ability of a geomorphologist to predict these physical changes and system responses can assist with the prevention and reduction of hazards initiated by poor planning practices.

**Natural resources**

Field surveys and geomorphic maps assist the exploration and discovery of economically valuable weathered mineral resources including clay, silt, sand, and gravel. Geomorphologists can provide information on the sources, volume, and composition of available materials. Such applications are particularly common in floodplain valleys and older river terraces that store significant volumes of alluvial sediments. Geomorphologists can also advise on sustainable extraction practices and provide input for restoring or rehabilitating
land surfaces post-mining. For example, former sand and gravel pits can be restored to create wetlands used in mitigation banking.

Geomorphology has a long history of application to natural resources through globally extensive monitoring programs for water and sediment discharge. On the River Nile in Egypt, flood events and river levels have been recorded as far back as 3100 BCE. In the United States, the United States Geological Survey manages a nationwide network of stream gauges that collect data on surface and groundwater resources. Applied geomorphologists use these records to quantify patterns of stream flow and sediment discharge, and analyze how they are changing relative to natural and human impacts, such as land cover changes, dam operations, and global climate change.

Environmental and ecosystem management

Geomorphology occupies a variety of roles in environmental management, and its application in problem-solving depends on the specific nature of the management goals. Some examples of environmental management problems that integrate geomorphology include (i) site-based classifications: using geomorphic maps to provide information on physical landforms or processes used in the characterizations of landscape and riverine settings and habitats; (ii) erosion control: monitoring and controlling sources of soil and sediment erosion for a variety of different land uses and environments; (iii) engineering: geomorphic form design and maintenance, common in coastal and river applications; (iv) ecosystem restoration: using geomorphology as a soft engineering tool to restore degraded physical habitats or modify the environment for ecological benefit; and (v) conservation: conserving geomorphic forms and processes of ecological or societal significance. In addition, geomorphic conditions represent one of the many variables included in environmental impact assessments used to evaluate development projects, and can have a significant impact at the intersection of land-use planning and environmental management, particularly where wetlands are concerned.

Environmental management of coasts and rivers requires a foundational base of geomorphic knowledge predicated on understanding current processes, interpreting past conditions, and predicting future scenarios of geomorphic change. In coastal settings, geomorphologists assist with solutions for reducing shoreline erosion, replenishing beach sediment, and restoring and conserving depositional nearshore and dune landforms. Coastal engineering involves the placement and maintenance of groins, jetties, and sea walls to control shoreline sediment and erosion dynamics. In riverine environments, geomorphologists address problems related to channel and bank stability, impacts from dams (commissioning, removal, and operation) and other in-channel control structures, catchment erosion and sedimentation, and the conservation and restoration of fluvial regimes, forms, and processes. River engineering involves channel form design and maintenance for working rivers where channel straightening, dredging, diversions, and dikes are common.

Climate change

Geomorphic responses to climate change, including potential increases in the magnitude and frequency of hazardous events, detrimental effects on natural resources (particularly water and soil), and land surface changes associated with changing temperature and precipitation regimes and rising sea levels, are of great concern.
to society. A geomorphologist’s ability to model land surface changes and to use geomorphic indicators to predict environmental changes across diverse settings is especially valuable for guiding decisions and actions that will minimize the negative effects of global environmental changes.

Careers for the applied geomorphologist

Applied geomorphologists most commonly provide services as independent contractors or academic scientists, or as members on science advisory panels. Rarely will specific job advertisements seek an applied geomorphologist, yet the niche exists across a wide array of career opportunities. Careers that use applied geomorphology occur within government research and regulatory agencies, planning and developer corporations, water supply districts and river basin authorities, nonprofit conservation and land trust organizations, and private engineering and environmental consulting firms. Individuals can receive training in applied geomorphology through geography, geology, environmental, or earth sciences programs. Many applied geomorphologists practice under the title of hydrologist, geologist, or environmental scientist, with the exception of academia where the title persists.

Graf (1996) offered the opinion that geomorphologists have the responsibility to use their knowledge of basic research to the benefit of social issues concerning environmental resource management and public policy. Goudie (2002) argued that applied geomorphology is one of the more attractive attributes of the discipline for recruiting new practitioners and for bringing the relevance of geomorphology into public perception. Collectively, these assertions provide assurance that, as geomorphology becomes better recognized for its applications to human–environmental problem-solving, the role of the applied geomorphologist will continue to gain relevance and career opportunities will continue to expand.

SEE ALSO: Anthropogeomorphology; Environmental assessment techniques; Environmental management; Environmental restoration; Environmental risk analysis; Geomorphic hazards; Geomorphological mapping and geospatial technology; Land systems science

References

McGregor, Duncan M., and Donald A. Thompson. 1995. Geomorphology and Land Management in a

Aquifers

Sheryl Luzzadder-Beach
University of Texas at Austin, USA

Aquifers are saturated, water-bearing geologic formations or sediment matrices found underground in the phreatic zone (at atmospheric pressure). In order for a water-bearing deposit to be designated as an aquifer, the water yield is expected to be “economically sufficient” (Sterrett and Hanna 2007). The top of the water table defines the division between the phreatic zone and the overlying vadose zone (unsaturated, below atmospheric pressure). Water at or below the water table in an aquifer is referred to as groundwater (see Groundwater). Water above the water table in the vadose zone is commonly referred to as soil water (see Soil water). Aquifers are composed of a wide variety of geologic units, and are created by numerous geomorphic and geologic processes over time. Groundwater enters an aquifer through percolation, and groundwater flow in aquifers is defined in terms of economic quantities, that is, whether or not it is useful in human terms. Groundwater may be recharged through the hydrologic cycle (see Hydrologic cycle), but withdrawing groundwater by pumping at a rate unsustainable by recharge is considered water mining. Some aquifers contain water accrued during past wetter climate regimes and are no longer recharged; this is called fossil water.

For a geologic unit to receive, store, and transmit water, it has to have certain physical properties related to its matrix, or assemblage of pore spaces and geologic material. These properties determine the volume of the voids between the rock particles or material, which is known as porosity, and the connectedness of those voids, known as permeability. These two dimensions combine to define an aquifer’s hydraulic conductivity or $K$ (Fetter 2001). Transmissivity takes this concept one step further into measuring how much water can travel horizontally through a given saturated thickness of aquifer, using a hydraulic gradient constant of 1. The measure looks like this:

$$T = bK = m^2/\text{day}$$

Hydraulic gradient measures the difference in hydraulic head in an aquifer segment along a slope length, $dh/dl$ (difference in head/difference in length). Adjusting then, for a specific hydraulic gradient and slope, applied to an area of aquifer cross-section ($A$) allows discharge ($Q$) to be calculated, in an equation first derived through laboratory experiments by Henry Darcy in the mid-nineteenth century (Fetter 2001). This equation is known as Darcy’s law (see Hydrologic flow models). A negative sign accounts for diminishing slope. One form of Darcy’s law is therefore expressed as:

$$Q = -KA\left(\frac{dh}{dl}\right)$$

A rock material can have a high porosity, but if those pore spaces are not connected it will not hold or transmit water, and therefore it will have a low hydraulic conductivity. If a rock material is impermeable or transmits water at noneconomic (low) rates, it is called a confining layer. This
AQUIFERS

is also referred to as an aquiclude (Sterrett and Hanna 2007). A related condition is an aquitard, which is more permeable than a confining layer or aquiclude, and may exchange or “leak” water to an adjoining layer. It may also store some groundwater, but has insufficient transmissivity to serve a well (Sterrett and Hanna 2007).

Porosity and permeability, then, relate to the particle size and sorting of the matrix and the voids, and void connectivity within the matrix. Many different geomorphic and geologic processes produce geologic units. Some are produced by igneous rock formation processes, and others are produced by secondary deposition and eventual burial, through sedimentary processes. Others still are produced by dissolution processes, such as the dissolution by naturally acidic rainwater infiltrating the limestone, and by groundwater throughflow, in a process known as karst (see Karst processes and landforms).

One of the most porous igneous geologic units with a high potential for porosity and permeability is lava rock (see Volcanic processes and landforms). During volcanic eruptions, molten lava (magma) may flow out of a rift zone or a volcanic mountain at the surface of the Earth, and the molten lava is full of gases. These gases form spaces shaped like bubbles and tubes as the lava flow cools to a more viscous or plastic state, and the trapped gas bubbles continue to trace their way through the slowly plasticizing material to a surface to escape. These pathways then harden into bubble- and tube-shaped voids in the cooled and hardened lava rock, creating spaces to hold and transmit water (Sterrett and Hanna 2007). The walls around and between these bubbles and tube spaces can become weakened and fracture, thereby creating more connectedness or permeability in the lava rock deposit. Some lava tubes can be quite large, where liquid molten lava flows through a tunnel-shaped void or space surrounded by more hardened lava on the relatively cooler exterior of the flow. These tubes can eventually empty of flowing lava, leaving behind cave-like spaces that can hold and transmit water. Other, denser igneous rock matrices such as basalt, quartz, or granite may not be as water-bearing if the crystals are small or if there is little connected pore space between crystals. However, any rock may be fractured or have horizontal joints or vertical columns (as in columnar basalt) between flows, and this will be a zone where water can accumulate and flow if these spaces are connected.

Consolidated sedimentary rock deposits such as sandstone, limestone, dolomite, and shale can serve as aquifers. Water moves either through the permeable matrix of the rock material itself (e.g., between sand grains in sandstone) or through cracks and joints between sedimentary layers (e.g., sandstone, limestone, shale).

Well-sorted, unconsolidated materials may also serve as infiltration, percolation, or recharge zones and aquifers when buried. Conversely, unsorted materials can serve as aquitards or aquicludes, slowing or stopping the flow of groundwater. Several geomorphic processes can transport, sort, and deposit these materials, and then subsequently bury the deposits to become aquifers (a few examples are provided here; for more detail please see Sterrett and Hanna 2007). For example, glaciers (see Glacial depositional processes and landforms) convey a high volume of Earth materials of variable sizes and shapes over great distances. These can be indiscriminately deposited in unsorted piles (moraines) or plains (till plains) as glaciers melt in place or melt back. Glacial till and moraines are unconsolidated, unsorted, and have a low hydraulic conductivity.

Materials can also be moved and sorted fluvially by waters melting and flowing out of the glaciers, either in streams running through melting tunnels underneath glaciers, or by
streams washing out in front of glaciers. The former process deposits a sorted stream bed, building up in the ice tunnel and as sediments drop from the melting roof of the ice cave. After the glacier melts back to reveal these features, known on the landscape as eskers, they resemble raised railroad beds in shape. When eskers are buried by glacial till, they will transmit water underground as aquifers, confined by the till. Sandy outwash may also be deposited by meltwaters in front of a melting glacier, and this too will potentially form an aquifer if buried.

Fluvial processes (e.g., Earth surface change by moving water) also remove, transport, sort, and deposit materials in an orderly way on the landscape, as can be observed in the sorted particle sizes on a sand bar or cobbled riverbed. When these deposits become subsequently buried by other geomorphic depositional processes, they too may also serve as aquifers (see Fluvial depositional processes and landforms). Besides buried riverbeds, alluvial fans make good aquifers as well. Alluvial fans are fan-shaped deposits of well-sorted grain sizes from sand to boulders, occurring at the base of steep streams as they enter flatter valley bottoms, lose energy due to the change in slope, and spread or fan out. As energy is lost, the largest particles are dropped out first, and so on, until a well-sorted stack forms. Water flowing from the stream then infiltrates into this feature and can flow underground in significant quantities; many municipal and agricultural wells are located on alluvial fans (Figures 1 and 2).

Eolian processes may also transport and deposit silt and sand-sized particles in organized sheets, dunes, and hills; once buried and confined or semiconfined by other geomorphic processes these unconsolidated materials may also serve as aquifers due to their well-sorted and abundant pore spaces between particles.

Sustainability of aquifers is threatened by climate change altering recharge rates, and by withdrawals by humans at rates faster than recharge. A prime example of a threatened aquifer is the Ogallala Aquifer, which underlies eight states in the High Plains portion of the United States (Little 2009). This shallow, stream-deposited aquifer stretches from Texas to South Dakota, and its depth to the top of the water table varies from near the surface to 150 m deep (Kromm 2015). As agricultural water use began booming in the 1940s and 1950s, some parts of the aquifer are now overdrawn by 30 m (Kromm 2015), and it is estimated it would take 6000 years to recharge it in the current rainfall regime (Little 2009). White and Kromm (1995) note that more effective groundwater management in the Ogallala Aquifer zone has slowed the depletion rate, but the rate remains at just under a meter per year and contamination issues remain a concern (Kromm 2015). Aquifers in other agricultural regions such as the San Joaquin Valley of California face similar overdraft threats and irreversible damage from ground subsidence, and salination from excessive irrigation and evaporation cycles in the vadose zone (Schoups et al. 2005). Subsidence occurs as the pore spaces within rapidly depleting aquifers are mined of their groundwater and collapse. Some parts of the San Joaquin Valley experienced around 9 m of ground subsidence between 1925 and 1977 (Botzan et al. 1999). Once pore spaces collapse, an aquifer can no longer be recharged, so programs in water banking and artificial recharge through interbasin water transfers may help protect aquifers (Galloway and Burbey 2011).

SEE ALSO: Fluvial depositional processes and landforms; Glacial depositional processes and landforms; Groundwater; Hydrologic cycle; Hydrologic flow models; Infiltration; Karst
Figure 1  Link to Google Maps Image of Chico, CA, and the alluvial fan to its east. https://www.google.com/maps/@39.7579072,-121.8409199,16127m/data=!3m1!1e3. Reproduced from Google Maps © 2015 Google.

Figure 2  Link to Google Maps Image of irrigation patterns on alluvial fans along the east side of Surprise Valley, at Cedarville, CA. https://www.google.com/maps/@41.5085594,-120.1620009,6614m/data=!3m1!1e3. Reproduced from Google Maps © 2015 Google.
processes and landforms; Soil water; Volcanic processes and landforms

References


Further reading


Most geographical research analyzes contemporary landscapes and recent spatial processes and trends, based on direct observation in the field, the manipulation of current demographic and statistical data, and the examination of recent topographic maps and aerial photographs. However, researchers investigating past or long-term spatial processes and the reconstruction of past landscapes or regional geographies have depended heavily on archival and document research. Much of the methodology for conducting this type of research was developed during the last half of the twentieth century, as British and North American historical geographers focused on recreating the past geographies of selected places or regions within Europe and the Americas.

Archival and document research is fundamental to historical geographers, but there are only a few comprehensive and critical surveys of these sources. Two such surveys by geographers pertain to British sources (Baker, Hamshere, and Langton 1970; Morgan 1979). They address such historical materials as tax records, inquisitions, land grants, manorial court records, enclosure awards and maps, tithe maps and apportionments, population censuses, agricultural and industrial censuses, probate inventories, parish registers, directories, newspapers, and transportation records.

In terms of North America, there are several articles discussing sources for the colonial period (Merrens 1963; Merrens and Ernst 1978). These articles highlight such primary sources as tax lists, customs statistics, merchant records, family letters, newspapers, estate inventories, land surveys, local court minutes, travel accounts, ecclesiastical reports, and maps. More comprehensive studies discussing both the colonial period and nineteenth- and early twentieth-century sources have taken the form of bibliographies and bibliographic essays (Grim 1982; Grim 1983; Conzen, Rumney, and Wynn 1993; and Grim, Rumney, and McIlwraith 2001). Although these compilations have identified a wide variety of primary sources that are potentially useful, historical geographers have used some types of documents, both textual and graphic, more successfully than others.

The quantity and quality of primary source materials that are available for any particular historical geographical problem will depend on the time period in which the information was recorded, the record-keeping practices of the government bodies and individuals involved, and the archival practices of the successive holders of the documents. While it is impossible to enumerate and evaluate all the potential documents and repositories that may be useful for historical geographers, a selection of basic primary sources that researchers have successfully employed to recreate past geographies of North America is discussed here.

Qualitative landscape descriptions

Exploration and travel literature, whether in manuscript or printed format, often contain
narrative descriptions of regional landscapes or primary transportation routes. Such narratives provide a basic source for a qualitative understanding of past landscapes and geographic regions. For example, explorers’ reports and travelers’ diaries have been useful in documenting the growth of geographical knowledge or the initial perception of frontier regions. Landscape descriptions recorded in these accounts are useful for reconstructing physical landscapes, settlement patterns, economic activity, and environmental perception. Observations concerning modes of transportation and routes traveled provide an understanding of generalized patterns of movement and circulation at any given time. A significant subset of travel literature includes the reports prepared by government-sponsored exploring expeditions during the nineteenth century. Although these sources have their own inherent biases – military reconnaissance and economic exploitation, for instance – they have the potential for presenting a more balanced representation than private accounts, which may suffer from exaggeration, untruth, or literary license (Grim 1982, 87–91).

Population data

Population counts and demographic statistics, whether collected at national, state, or local jurisdictions of government, can be found in a variety of documents including general population censuses, tax rolls, and militia lists. The most familiar and probably most useful source of population information for the United States is the federal government’s decennial census. Every 10 years, beginning in 1790, the government has enumerated the population for the purpose of determining representation in Congress. The information gathering in the first census was brief (name of head of household and number of free and non-free members), but subsequent enumerations collected much more. For example, the 1850 census started listing names of family members as well as age, sex, occupation, and birthplace. The manuscript census schedules on which the field data were recorded have proven indispensable for microstudies. Currently, schedules from 1790 to 1940 are available on microfilm and open to general research (www.archives.gov/research/census). US federal census data is also available online through the Historical Census Browser hosted by the University of Virginia Geospatial and Statistical Data Center (http://mapserver.lib.virginia.edu) and United States Census Data presented by the Inter-university Consortium for Political and Social Research at the University of Michigan (www.icpsr.umich.edu/icpsrweb/ICPSR/themes/census).

The original census data is collected at very local geographical units known as an enumeration districts, but is aggregated and published according to increasingly larger geographic units such as township, ward, city, county, and state. Also of interest to geographers are the statistical atlases that accompanied the censuses from 1870 to 1920; those for 1870, 1880, and 1890 are available on the Internet (http://memory.loc.gov/ammem/gmdhtml/setlhome.html).

Demographic statistics for North America’s colonial period are not as readily available or as systematic and comprehensive in coverage as data in the federal censuses. Various levels of colonial administration made population counts and compiled tax and militia lists, but with no regularity or uniformity (Grim 1982, 95–100).

Property records

Information relating to real and movable property can be found, primarily at local
administrative levels, in such records as land patents, deeds, wills, and estate inventories. However, in order to locate pertinent land records, a distinction must be made between the public-land states – those in which the federal government was the original owner of the land – and those without public lands. These latter included the original 13 states plus Maine, Vermont, West Virginia, Kentucky, Tennessee, and Texas, in which the original ownership was vested in the colonial or state government. Records from the public-land states, including surveys and initial transactions, are housed either with the National Archives or the Bureau of Land Management (successor to the original General Land Office; see Buissere 1990, 89–109, 311–315). Records of original surveys and transactions in those states without public lands are in the custody of the respective state land offices or states archives. For all states, subsequent land sales or transfers were handled locally, usually at the county level, and normally have been maintained in a county clerk’s office. They have suffered the ravages of fire and war, and in recent years the older portions of many of these records have been centralized in state archives.

The process of surveying and disposing of the public domain in the United States has generated a wealth of records that are useful to historical geographers. Land survey records, which consist of field notes and survey plats, have a primary legal value for establishing the boundaries of a parcel of land as well as a secondary value for reconstructing vegetation patterns from witness trees and related comments recorded in the notes. Land-disposal records – patents or deeds, land-entry papers, and tract books document the sale and resale of the land and are useful for reconstructing the landownership process (Grim 1982, 101–105).

Business records

The US Census Bureau’s census of manufactures is a major source of nineteenth-century industrial statistics, but other privately created documents such as merchant and plantation accounts or company archives can provide an understanding of the economic activity of individual enterprises. For example, strong collections of business and industrial archives have been gathered together at the John D. Rockefeller, Jr Library at Colonial Williamsburg (http://research.history.org/library) and the Hagley Museum and Library near Wilmington, Delaware (http://www.hagley.org/library). Newspaper advertisements and business directories are also useful for studying a wider range of economic activity. Newspapers have been used widely to study the development of political and public opinion, but they offer great potential also for geographical studies and are widely accessible on microfilm. Advertisements, specialized newspapers, and circulation statistics provide useful spatial information about commerce, manufacturing, and agriculture. City residential and business directories provide a useful complement to the manuscript census schedules for studies of population mobility and social or economic characteristics (Grim 1983, 270–273).

Graphic sources

A wide range of graphic materials can be useful for historical geographical research. Maps (at all scales), photographs (still and aerial), and landscape sketches provide both a spatial and a visual perspective for many historical studies. Although often used casually as illustrative elements of a study, many such documents, when critically appraised for their respective biases, can stand alone as primary sources. The essays edited
by David Buisseret (1990) demonstrate the possibilities, focusing on such diverse topics as exploration and military maps, property and topographic surveys, landscape views, urban maps and panoramic views, fire insurance maps, and aerial imagery.

With the introduction of large flatbed scanners and data-compression software, it is possible to scan large maps at very high resolutions and make them readily available on the Internet. The Geography and Map Division of the Library of Congress has been a pioneer in applying these technologies to cartographic collections. This digital map collection (www.loc.gov/maps) includes maps from the sixteenth century to the present. The site provides search capabilities within the entire collection or with broad thematic categories, such as discovery and exploration, transportation and communication (railroads), cities and towns (panoramic and Sanborn maps), and military battles and campaigns (American Revolution, Civil War, and World War II). Another early pioneer in providing online access to large number of eighteenth- and nineteenth-century maps is the private collector David Rumsey (www.davidrumsey.com). During the past decade other map collections including the New York Public Library (www.nypl.org/locations/divisions/map-division), Harvard Map Collection (http://hcl.harvard.edu/libraries/maps/digitalmaps), Osher Map Library, University of Southern Maine (www.oshermaps.org), and Boston Public Library (maps.bpl.org) have provided digital access to their collections, while federate sites such as Old Maps Online (www.oldmapsonline.org) are providing access to multiple collections at one digital location.

For historical geographers reconstructing past physical landscapes, the absence of a systematic program of topographic mapping in the United States prior to the end of the nineteenth century has been a major research obstacle. Consequently, they have had to search for limited availability of topographic surveys of small, scattered localities. One of the best sources is the topographic surveys that were conducted by military engineers in preparing after-battle maps for the major battles in the America’s Revolutionary and Civil Wars. In the 1880s, when the US Geological Survey (USGS) was established, a systematic and comprehensive topographic mapping program began. Aerial photography was introduced in the 1930s as a major innovation in gathering geographic information (Buisseret 1990, 238–259, 283–309).

One of the most valuable research tools for historical urban studies is fire insurance maps, of which the D.A. Sanborn Company of New York was the major publisher. Prepared primarily for underwriters seeking to determine the risk of insuring properties, these maps provide block-by-block inventories of individual buildings in approximately 12,000 cities and towns in the United States and Canada from the late nineteenth century to the middle of the twentieth century. They may effectively be used in conjunction with city directories and manuscript census schedules (Grim 1982, 43–45; Buisseret 1990, 213–237).

The panoramic view is another type of urban map useful for determining the internal spatial structure of cities. Drawn in perspective from an imagined elevation of 600 to 900 m, these commercially published bird’s eye views depict a town’s street pattern as well as individual buildings. The drawings provide reasonably accurate portrayals of the major towns and cities in the United States and Canada from the middle of the nineteenth century through the first quarter of the twentieth century. However, these views should be evaluated for their selectivity and exaggeration because city fathers commonly commissioned the views to promote a city’s
business, commercial, or industrial attributes (Grim 1982, 45–47; Buisseret 1990, 143–163).

Archival and document repositories

Researchers studying past geographical landscapes and spatial processes can find pertinent research materials within a wide variety of institutional settings. Official government records are generally maintained by archival repositories at the national, regional (state or province), and local levels while manuscript collections, printed and manuscript maps, aerial photographs, and related graphic materials can be located at historical societies, privately endowed libraries and manuscript repositories, and large public and university libraries. Many of these institutions have prepared general summary guides to their holdings or detailed listings of specific collections (Grim 1982, 74–85; Grim 1983, 252–258). While it is impossible to describe the most important repositories and their pertinent records, it is worth mentioning the two largest repositories in the United States – the US National Archives and Records Administration and the Library of Congress.

The National Archives and Records Administration (NARA), the official repository for the noncurrent records of the various agencies of the US federal government, consists of three facilities in the Washington DC metropolitan area, 17 regional archives and federal records centers, and 13 presidential libraries (www.archives.gov). The holdings of the Cartographic and Architectural Section located in College Park, Maryland, include over 15 million maps, charts, aerial photographs, architectural drawings, patents, and ships plans, comprising one of the largest such accumulations in the world. While these graphics sources are of primary interest to historical geographers, there are many other important resources for recreating past geographies as highlighted in a 1971 symposium addressed directly to historical geographers (Ehrenberg 1975). For example, numerous records pertain to the establishment and survey of boundaries, the regulation and monitoring of commerce and trade, the founding and planning of cities and towns, the utilization and preservation of natural resources, and the financing and construction of transportation systems.

The Library of Congress, also located in Washington DC, serves as the national library. It is also the largest library in the world with more than 158 million items on approximately 838 miles of bookshelves. The collections include more than 36 million books and other print materials, 3.5 million recordings, 13.7 million photographs, 5.5 million maps, 6.7 million pieces of sheet music, and 69 million manuscripts (www.loc.gov). In the United Kingdom, the UK National Archives (www.nationalarchives.gov.uk) and the British Library (www.bl.uk) serve similar functions.

Increasingly, archival repositories, manuscript collections, and libraries are digitizing their collections and making them available on the Internet. Several websites are bringing together resources from a large number of collections, including the Digital Public Library of America (http://dp.la) and Internet Archive (https://archive.org/index.php). There are also a number of websites that focus on aggregating materials that have a thematic orientation or relate to a specific genre, such as travel literature (American Journeys: /www.americanjourneys.org/), newspapers (Library of Congress: http://chroniclingamerica.loc.gov/), or nineteenth-century weekly and monthly journals (Making of America: http://digital.library.cornell.edu/m/moa/). Another useful website is World History Unpacking Evidence at George Mason University’s Roy Rosenzweig
ARCHIVAL AND DOCUMENT RESEARCH

Center for New Media (http://chnm.gmu.edu/worldhistorysources/whmunpacking.html), which provides guides to eight types of primary sources (music, images, objects, maps, newspapers, travel narratives, official documents, and personal accounts) with links to selected digital collections.

SEE ALSO: Cartography: history; Environmental history; Exploration; Historical geography; Historical settlement; Transportation history; Travel geographies

References


Areal differentiation
(or chorology)

Ron Johnston
University of Bristol, UK

A definition of geography as the study of areal differentiation of “the spatial distribution of physical and human phenomena as they relate to one another” (Gregory 2009, 35) is generally associated with Carl Sauer and his classic 1928 monograph *The Morphology of Landscape* (Sauer 1976/1928), although he indicated that his definition was synonymous with that of chorology, which has a long pedigree.

Chorology

Chorology was one component of geography as practiced by the ancient Greeks. Strabo, for example, defined a geographer as “the person who attempts to describe the parts of the earth.” His approach differed significantly from Ptolemy’s, for whom mapping was the core of geography’s *raison d’être*. Those two components, chorology and mathematical geography respectively, became institutionalized in later centuries as special geography (the description of particular areas) and general geography respectively, and these were the precursors of what in the twentieth century became known as regional and systematic geography. The term “chorology” is rarely used now to depict the description and interpretation of the spatially varying characteristics of the Earth’s surface.

From chorology to areal differentiation

In the early twentieth century, the chorological approach was promoted by both Sauer and Richard Hartshorne. Sauer offered three separate identifiers of geography’s content: (i) the study of the Earth’s physical processes; (ii) the study of “life forms as subject to their physical environment”; and (iii) the study of areal or habit differentiation – between each pair of which there was little relationship (Sauer 1976/1928, 316). Geography, for him, was one of the “great fields of knowledge,” concerned with “great categories of phenomena.” “Geography assumes the responsibility for the study of areas because there exists a common curiosity about that subject,” a disciplinary specialism that no other subject had pre-empted and that would exist as one of the “great fields of knowledge” even if geography had not been invented as a discipline to specialize in its study.

Sauer argued that geography as areal differentiation was more than an investigation of what is where: “the phenomena that make up an area are not simply assorted but are associated, or interdependent” and geography portrays those associations, what he termed the “areal connection of the phenomena and their order” (Sauer 1976/1928, 318). Thus, whereas the “facts of geography are place facts; their association gives rise to the concept of landscape”: areas differ in their landscapes, because of differences in the interrelationships among phenomena across separate areas. The study of such differences meant that geography was a scientific field rather than simply a provider of gazetteers of information. Landscapes comprise two main elements: their
AREAL DIFFERENTIATION (OR CHOROLOGY)

sites, or their physical nature; and the cultural expressions within them, the “impress of the works of man [sic] on the area.” Sauer’s career focused on that impress, on decomposing the morphology of cultural landscapes and accounting for their separate characteristics – over both space and time.

The codification and promotion of geography as the study of areal differentiation is generally associated with Richard Hartshorne’s (1939) monograph *The Nature of Geography*. The two scholars’ approaches had many similarities, but Hartshorne rejected Sauer’s cultural landscape concept (he termed it “sloppy”). His extensive, though partial, archaeology of writings about geography’s nature, particularly by nineteenth-century German scholars, led him to a broad definition (derived from Richthofen) that “geography studies the differences of phenomena causally related in different parts of the earth surface” (Hartshorne 1939, 92). As such, the discipline was composed of two main parts: systematic geography, which is the “non-explanatory description” of the distributions of individual phenomena (or chorography); and regional geography (or chorology), which explores the relationships between the phenomena that define and account for the specific features of particular areas.

Hartshorne presented geography as the discipline providing accounts for differences between areas resulting from the spatially varying interactions between humans and the physical environment. That definition was widely accepted for several decades encompassing World War II, even by those who paid relatively little heed to Hartshorne’s excavation of the subject’s/discipline’s history (for critiques see the essays in Entrikin and Brunn 1989). In the United Kingdom, for example, Wooldridge and East’s (1958, 28) *The Spirit and Purpose of Geography* defined its essential purpose as addressing “how and why does one part of the earth’s surface differ from another.”

Hartshorne revised his definition of areal differentiation in his *Perspective on the Nature of Geography* (1959), a response to critiques of his earlier book. He accepted that “the science of areal differentiation” was inadequate as a definition of geography – “either meaningless or misleading” – not least because most sciences address differences in some form or another. The underlying conception was correct, but it needed a more nuanced definition; he offered that “geography is concerned to provide accurate, orderly, and rational description and interpretation of the variable character of the earth surface” (Hartshorne 1959, 21), with several of the book’s chapters unpacking that statement because, to him, it was not “self-explanatory.” Areal differentiation was a useful shorthand statement of that more detailed definition. The largely descriptive role identified in his initial definition – geography “studies the differences of phenomena causally related” – was extended by the pair of goals “description and interpretation” (emphasis added); geography sought to understand the why as well as the what of spatial differences, thus bringing Hartshorne’s definition closer to Wooldridge and East’s.

From areal differentiation to regional geography

Hartshorne’s approach became incorporated into the established practice of regional geography, much of whose work in the 1940s–1960s focused on the definition of regions, clearly bounded areas that differed in their phenomena complexes. That “containerization” of the world, at a variety of scales from the global through the national to the local, became a major focus for many regional geographers,
who to their critics were overly concerned with “drawing boundaries that do not exist around areas that do not matter” (Kimble 1951, 159) rather than portraying and accounting for interregional differences (with many regions having at best only indistinct boundaries as one landscape type merges into another).

The region had become some leading geographers’ leitmotif, however, as illustrated in American Geography: Inventory and Prospect, where Preston James declared that there “is just one kind of geography. That kind of geography is regional geography,” whose key instrument is the map, the technology whereby areal connections can be identified and areal differentiation portrayed:

Persons who undertake to carry on geographic studies must specialize in order to develop competence in a portion of the field. Nevertheless, whether they specialize on the physical, biotic or cultural aspects of geography, the analysis of the meaning of likenesses and differences among places involves the use of certain common concepts and methods: basic to the whole field is the regional concept; fundamental to the effective study of geographic phenomena is the method of precise cartographic analysis. (James 1954, 7)

To perform their academic work, he argued that geographers must define regions, the matrix within which areal differentiation of the Earth’s surface is organized.

Hartshorne recognized the relative futility of this task, however. He argued that most regions identified by geographers were mental constructions only and concluded that “it is not possible to define sections of the earth’s surface as regions that form units in reality, [so] that we cannot correctly consider them as concrete individual objects” (1939, 281). Livingstone’s (1992, 309) summary of his position was that “regions just did not exist as all-embracing units.” The region was merely a convenient mental construction – often with boundaries on maps that were difficult (if not impossible) to locate on the ground – within which to organize and present material. Livingstone therefore saw regional geography as an “incoherent enterprise”; regions were a useful means to convey the nature of areal differentiation across an area, but their identification was not a valuable end in itself.

Hartshorne clarified this difference in his Perspective, concluding that:

The great majority of geographers have recognized that the world does not consist of a mosaic of distinct regions, and that we cannot hope to classify the more complex integrations of areal phenomena in a single objective system of regions. Nevertheless, they recognize that in order to analyze complex integrations in terms of a limited range of areal variations it is necessary to divide large areas into smaller parts. The purpose in dividing the area is to secure areal sections, or “regions,” such that within each region the elements of the segment of integration under study will demonstrate nearly constant interrelations and the maximum degree of interconnections among places. (1959, 129)

Regions provided a way of summarizing areal differentiation in the interrelations between (physical and cultural) phenomena in a matrix of areas that maximized internal homogeneity and external heterogeneity. But, whereas the region was a useful tool for uncovering spatial variations in geography’s subject matter, that tool was the means, not the end, which is to understand and explain why such variations exist. Thus he concluded that “The ‘regional concept’ and the ‘regional method’ must not be confused with what we commonly call ‘regional geography’ … The regional concept is applicable, and the regional method is used in fact at every level of geographic study” (Hartshorne 1959, 129).
The demise of regional geography – but not of the regional concept

Most regional geographers followed Hartshorne’s general strictures – at least implicitly – but some went further. For the latter group regional geography remained the discipline’s ultimate goal; systematic studies were very much a means to that end and the true nature of geography was demonstrated by its integration of material into an appreciation of how interactions of their subject matter produced unique, if not singular, places – areas of the Earth with their specific characteristics revealed by the regional method.

By the 1960s, however, this approach was under considerable attack, mainly from geographers who wished to recreate their discipline as a law-making spatial science and promoted scientific studies of particular phenomena as the way forward. In this “new geography” many used the concept of the region to describe a spatial matrix (such as urban hinterlands and labor markets, and in the nascent multidisciplinary regional science) within which society was organized, but there were reactions to that approach defending “traditional” regional geography (on which see Johnston and Sidaway 2015). As late as the 1980s, for example, a president of the Association of American Geographers proclaimed that regional description was “The highest form of the geographer’s art” (Hart 1982, 1). Resonating with some of Sauer’s argument, he claimed that:

Society has allocated responsibility for the study of areas to geography: this responsibility is the justification for our existence as a scholarly discipline. Most people are inherently curious, and they want to know more about the world in which they live. … [To satisfy that curiosity] … The highest form of the geographer’s art is producing good regional geography – evocative descriptions that facilitate an understanding and an appreciation of places, areas and regions.

Hart’s argument was that providing such descriptions should reunite the discipline, which had been fragmenting since the so-called “new geography established by the quantitative revolution.” The core concept in that reunification should be the “region,” required “in order to understand why we need the diverse and variegated systematic subfields of geography” and also to meet the popular demand for satisfying “human curiosity about how much of what is where and why it’s there, about the where and why of places and people, about the land and how people have used and abused it. … we jolly well had better give them what they want” (Hart 1982, 19).

But few geographers followed Hart’s argument, and geography as an academic discipline has very largely distanced itself from the practice of regional geography. There are now virtually no university courses or books comparable to the regional texts of the immediate post-World War II decades, and – except in some introductory undergraduate courses for students with little or no previous exposure to geography (notably in the United States) – there are very few academic geographers whose core concern is to provide descriptive accounts of areal differentiation. In the third quarter of the twentieth century this was because of the focus on spatial science, on the discovery of and explanation for spatial order, addressing general patterns rather than the characteristics of specific places. The term “region” continued to be fairly widely used, but the focus was on functional organization rather than on regional mosaics.

Critiques of spatial science have since stimulated the emergence of a renewed interest in areal differentiation, albeit with different foci, such as areal variations in levels of economic development, producing uneven geographies of economic, social, and political inequality at a variety of spatial scales, but without calling
on either the regional concept or the regional method in drawing attention to and accounting for those variations. Similarly, the particularities of specific places – how they have been created, how they are currently experienced and appear – have been explored. Most recently Murphy (2013) promoted the need for geographers to produce “grand regional narratives” for nonacademic audiences as an important disciplinary role in advancing popular appreciation of the contemporary world, but the responses to his provocative piece have indicated little support.

Conclusion

The concept of areal differentiation continues – if often only implicitly – to underpin geography’s claimed raison d’être within the academic division of labor of providing accounts of differences between places: what is where, and why? Many one-sentence definitions of the discipline focus on this question in some way: a recent geography dictionary, for example, defined geography as “The spatial and temporal analysis of human and natural systems and their interrelation across scales” and human geography as “The study of the interrelationships between people, place, and environment and how these vary spatially and temporally across and between locations” (Castree, Kitchin, and Rogers 2013, 180, 223). But within such broad definitions the myriad different intradisciplinary specialisms mean that most attention is given to the parts and very little to the whole. Some overarching descriptive accounts of an area are still produced (e.g., Haggett 2013) but they are seen as neither the rationale for the discipline nor the grand synthesis of an academic career. In general, that project is left to those who promote other geographical imaginations, such as popular geographical magazines.

SEE ALSO: Human geography; Landscape; Regional definition and classification; Regional geography; Uneven regional development

References


AREAL DIFFERENTIATION (OR CHOROLOGY)


Argentina: GÆA Sociedad Argentina de Estudios Geográficos (Argentine Society of Geographical Studies)

Founded: 1922
Location of headquarters: Buenos Aires
Website: www.gaea.org.ar
Membership: 500 (as of December 31, 2013)
President: Darío Cesar Sanchez
Contact: dario.cesar.sanchez@speedy.com.ar

Description and purpose

GÆA is one of the first nonprofit scientific and educational associations of geography in Argentina. Its aim has been, from its origin, to contribute to the production of geographical knowledge to benefit science and Argentine society.

GÆA has improved research on the vast Argentine territory, encouraging scientific exploration and expeditions, and has contributed to the academic institutionalization of geographical studies. Nowadays, GÆA shares its interests in the theory, methods, and practice of geography with other emblematic public and private institutions such as the IGN Instituto Geográfico Nacional, the Academia Nacional de Geografía, universities and teaching training institutes, and so on.

Journals or major publication series

*Boletín de GÆA Sociedad Argentina de Estudios Geográficos.* http://gaea.org.ar/BoletinGAEA.pdf

Current activities or projects

Representing a discipline with strong national and international perspectives, the GÆA outreach and research network consists of more than 7000 members in Argentina and across America and Europe. The GÆA supports and recognizes the work of professors, researchers, and professionals in geography through its many academic and scientific relationships, especially in geographical education; through research grants and awards; the annual scientific meeting devoted to geographical research and geographical education, known as Geography Week “Semana de Geografía” fosters scientific exchange and interaction; publications; and other collaborative activities, such as the library; and many conferences, workshops, and seminars about the principal geographical themes of Argentina.

Brief history

The name of the GÆA Argentine Society of Geographical Studies is composed of the Latin expression Gæa, meaning earth, followed by the subtitle “Argentine Society of Geographical Studies.” The founders intended to give it a collective, national, and scientific significance. Its emblem includes part of the peerage emblem of Felix de Azara (a Spaniard who is used as an example of the Spanish Age of Enlightenment) known as the first geographer of the Rio de la Plata. It includes the foundation year, 1922, with a symbolic value “lasting in time.”
ARGENTINA

GÆA’s principal annual contribution is the Geography Week “Semana de Geografía”. The first one was held in 1936, when these type of events were still rare. They are held annually in different cities in the country and the 75th such Semana took place in 2014. It offers a platform for the dissemination of the latest research and scholarship by high level geography teachers, as well as training opportunities, and outreach to the wider public and the media.

The members, among them remarkable foreign and Argentine scientists, have always worked with enthusiasm and perseverance. GÆA is devoted to the creation of geographic knowledge. Today members of the scientific, professional, and teaching world collaborate in this society. Their interests range from the physical/natural subdisciplines to the interdisciplinary perspectives of geography and related social sciences.

Submitted by Susana Sassone
Arid climates and desertification

Thomas A. Wikle
Oklahoma State University, USA

Arid climates are often depicted as hot, sandy deserts or as barren and unproductive landscapes with little to no surface water. Although the Earth’s arid regions include hot deserts such as North Africa’s Sahara, arid climates are extremely diverse in terms of temperature, patterns of precipitation, and physical appearance. For example, some arid regions are never hot, such as the vast Antarctica ice sheet and only a small percentage are covered with sand. Rainfall alone may not explain aridity since hot locations may receive higher amounts of rainfall than other places and still be classified as arid. As a result, finding a precise definition for what it means to be “arid” can be complicated.

An element common to all arid climates is low precipitation. Unfortunately, it can be difficult to generalize rainfall in arid regions because precipitation events are infrequent and may be moderate in areas that experience winter storms or intense in locations influenced by tropical air masses. Short-duration precipitation events in arid regions are often separated by long rainless periods. In some areas falling rain may not contribute to available moisture if high temperatures cause it to evaporate before reaching the ground. An axiom applicable to arid climates is that the lower the mean annual precipitation falling in an area, the greater the variability.

As a result of evaporative losses, locations with higher temperatures may receive more precipitation than cooler locations and still be classified as arid. In lieu of using precipitation alone for measuring aridity, climatologists compare average precipitation to average potential evapotranspiration—a measure of the demand for moisture that includes both surface evaporation and plant transpiration. The index p/PET is commonly used to represent aridity where p = precipitation and PET = potential evapotranspiration as determined by atmospheric humidity, solar radiation, and wind. PET is higher during the summer, on days that are less cloudy, and during windy conditions.

Factors contributing to aridity

Covering 6.1 billion hectares or about 47% of the world’s land area, arid and semiarid climates are found on every continent. The vast area surrounding the South Pole has been described as the world’s most arid region because its cold air holds relatively little moisture. Outside of ice sheets, the most extensive arid regions are located within 25° of latitude on each side of the equator and within central and western parts of continents. Within these regions high surface pressure serves as the principal factor influencing aridity. Heated by intense sunlight, air overlying equatorial regions rises within Hadley Cells, traveling hundreds of miles in the upper atmosphere before descending in the subtropics. The dry, subsiding air warms adiabatically, creating stable atmospheric conditions and mostly cloudless skies. The Sahara, the Arabian peninsula, and portions of Central Asia and Australia owe their aridity to persistent high pressure.
ARID CLIMATES AND DESERTIFICATION

Other factors may contribute to arid climates, including mountainous regions that remove moisture as air travels inland. This happens when precipitation falling on the windward side of mountain ranges creates a dry condition called a “rain shadow” on the opposite (leeward) side. Asia’s Takla Makan Desert is located within a rain shadow created by the mountains of Tien Shan and Kunlun Shan, making it more barren than the Gobi Desert. Likewise the aridity of South America’s Patagonia Desert is strongly influenced by its location on the leeward side of the Andes. Locations may also owe their aridity to being far from sources of moisture. Arid and semi-arid regions within interior regions of North America, Asia, and Australia are influenced by continentality, referring to their distance from sources of moisture-laden winds such as large lakes, inland seas, and oceans. Ironically, ocean currents and upwelling can also be responsible for arid conditions in coastal areas where cool ocean surface temperatures contribute to atmospheric stability. Aridity can also be influenced by land cover where lighter-colored rock or sand reflects greater amounts of incoming solar radiation compared to darker vegetation that absorbs energy, contributing to rain-producing convection. Finally, variation in rainfall within arid climates can be influenced by activities far from desert regions. For example, a weather pattern associated with differences in atmospheric pressure between the Indian and South Pacific oceans called the Southern Oscillation is responsible for periods of rainfall significantly higher or lower than average.

Classifications of arid climates

Modern techniques for classifying climates follow a system developed by the ancient Greeks that divided the Earth into broad zones based on temperature. In the early 1900s German climatologist Wladimir Köppen introduced a more comprehensive system used today for classifying climates into five major types: A (Moist Tropical), B (Dry), C (Humid Middle Latitude), D (Continental), and E (Cold). Each category was defined by temperature except B-Dry Climates, which was designated on the basis of aridity. Köppen added additional letters as modifiers representing parameters such as seasonal differences in precipitation or temperature. For example, within B-Dry climates the letter h was used for dry-hot climates with average temperatures above 18 °C (64 °F) and k was added for dry-cold climates with average temperatures below 18 °C. Defined as having potential evapotranspiration exceeding precipitation, Köppen’s B Climates include the world’s deserts (BW Climates) and surrounding and comparably less dry “steppe” regions (BS Climates). B-Dry regions are further subdivided as either subtropical or mid-latitude climates.

Centered on the western sides of continents at latitudes between 25° and 30°, BWh climates are true deserts. Clear skies and high sun angles produce surface temperatures averaging 30 °C (90 °F). BWh climates have among the highest surface temperatures on Earth. On September 22, 1922, a temperature of 58 °C (136 °F) was recorded at Azizia, Libya. A distinguishing feature of BWh deserts is average precipitation that is less than half of annual potential evapotranspiration. Subtropical deserts are among the Earth’s most rainless places, with some locations experiencing decades without precipitation. Rain that does fall is delivered through convective showers of short duration; an entire season’s precipitation may fall during a single rainfall event. BWh climates have an average rainfall less than 30 cm (12 in.) with significant year-to-year fluctuations. Driven by daytime heating, winds of more than 100 kph, such as the dry passat (trade) winds of the Sahara, carry dust and
sand hundreds of miles. BWh deserts include the Great Australian, Sahara, and Arabian. The Sahara alone covers more than 9 million km², an area roughly the size of Europe.

Located in the interior of continents, BWk climates are mid-latitude deserts dominated by subtropical high pressure. A major difference between mid-latitude and subtropical deserts is seasonality. With moderately warm summers and long, cold winters, BWk areas receive the majority of their precipitation during the summer months. For example, in the US Southwest air masses originating over the Gulf of Mexico bring moisture during a “monsoon season” in July and August. Despite summer precipitation, mid-latitude deserts receive annual precipitation that is less than half of potential evapotranspiration. The pattern of precipitation within BWk climates is also influenced by their location far from oceans and position on the leeward side of high mountains that remove moisture from prevailing winds. The aridity of North America’s Great Basin Desert is the result of both its continental location and a rain shadow effect created by the Sierra Nevada and Cascade mountains. The Great Basin experiences cold winters with a significant amount of winter precipitation falling as snow. Dominated by anticyclonic conditions, winter temperatures in BWk climates remain below freezing for about half the year with seasonal extremes contributing to an annual temperature range that may exceed 28 °C (50 °F). The altitude of BWk areas can also influence temperatures. For example, while the Chihuahuan and Sonoran deserts are located at about the same latitude, the Sonoran is hotter because of its lower elevation.

**BS climates**

Covering about 14% of the Earth’s land surface, steppe (BS) climates serve as transition areas between humid and dry climates. “Steppe” is a Russian word meaning “treeless plain.” Within steppe climates annual potential evapotranspiration is more than half but less than 100% of total annual precipitation. Steppe grasslands receive precipitation from mid-latitude cyclones or shifts in the Intertropical Convergence Zone.

Found on the margins of BWh climates, subtropical steppe (BSh) is characterized by slightly greater annual precipitation (10 to 50 cm) and moderate mean temperatures (20 °C). As in the case of subtropical deserts, subtropical steppes have minimal cloud cover, contributing to nighttime cooling, especially during the winter months. Large BSh areas are found south of the Sahara, on the border of the Great Australian Desert, and in portions of Pakistan, India, southern Iraq, and Afghanistan.

Mid-latitude steppe (BSk) climates are similar to tropical steppe but receive greater amounts of rainfall and have larger temperature ranges that may extend from −40 °C (−40 °F) in winter months to 40 °C (104 °F) in the summer. The largest BSk region, the Eurasian Steppe, covers an area stretching 8000 km from Eastern Europe to China, reaching across the Ukraine and Central Asia. Within North America, mid-latitude steppe climates include portions of western Wyoming and southeastern Washington, along with parts of Idaho, Oregon, Utah, and Nevada. With an average rainfall of 81 cm, BSk climates are generally found on the edge of mid-latitude BWk areas. North American steppe lands are located to the west of clockwise advection that moves moist air away from the Bermuda–Azores high. In contrast, dry conditions within the mid-latitude steppes of Asia are the result of continentality. The dominant vegetation of mid-latitude steppe is grass that may grow to over a meter in height. Shorter grasses are found where less rainfall is received, while trees may be located near sources of water. Because of their
ARID CLIMATES AND DESERTIFICATION

Rich soil many of the world’s steppe lands have been converted to pasture or farmland.

Coastal deserts

Ironically, some of the world’s driest climates are located in coastal areas with aridity influenced by ocean water rather than atmospheric circulation or surface topography. For example, upwelling from cool ocean water off the coast of Chile and Peru affects winds as they move inland. The cooler air has a diminished capacity to hold moisture and becomes stable as it passes over the warmer land. Without convective uplift, moisture bypasses the coastal Atacama Desert as it travels eastward, depositing precipitation on the western slopes of the Andes. This process makes the Atacama the Earth’s driest coastal desert, with no more than two or three rain events taking place each century. Along with serving as a barrier to precipitation, cool ocean surfaces also contribute to high relative humidity, fog, and low stratus clouds that move inland for a few kilometers. Flowing northward, the cold Benguela Current off Africa’s southwestern coastline has a similar affect by inhibiting precipitation over the Namib Desert as winds push cool water originating near the Antarctic continent eastward. Like the Atacama, the Namib has a significant number of foggy days (200) per year.

Although aridity is most often represented in the context of large regions, arid conditions can also be observed in geographic areas of a few hundred square meters. Microclimates are small areas where tiny differences in relief, aspect, or altitude can influence solar radiation, temperature, precipitation, and/or wind speed. As a result, considerable differences in evapotranspiration may be noticeable across short distances. For example, a shaded area may have a rate of evapotranspiration half of what would be experienced in an exposed location. Variation in surface reflectivity associated with vegetation, soil type, or surface material may also alter the heat balance of small geographic areas.

Mapping aridity

Limited climatological data within some countries has contributed to difficulties in defining precise boundaries for arid and semiarid regions. To facilitate the mapping of dry conditions, the United Nations Environment Programme (UNEP) developed an index classifying arid lands into four zones: hyperarid, arid, semiarid, and dry subhumid (Table 1). Covering just 7.5% of the Earth’s land area and nearly devoid of vegetation, hyperarid areas have annual rainfall that rarely exceeds 20 cm per year. In contrast, areas classified as arid receive summer rainfall that is not enough to support agriculture without irrigation. Areas designated as semiarid support native grasses, shrubs, small trees, and sedentary agriculture and pastoralism. Although precipitation is seasonal, dry subhumid areas facilitate some rain-fed agriculture.

Deserts and climate change

Arid regions are noted to be among the world’s most responsive ecosystems to global climate change. As noted by Lioubimtseva (2004), the climates of deserts and semideserts that exist today are known to have changed across both spatial and temporal scales. For example, during the Mid-Holocene period (9000–4000 years BP) the “Green Sahara” was covered by woodlands, scrubs, and grasslands that supported giraffe and other large animals presently found much farther south. Although paleodata are useful for characterizing climate change over long time intervals
Table 1  Dryland categories.

<table>
<thead>
<tr>
<th>Category</th>
<th>p/PET (UNEP 1992)</th>
<th>Average annual precipitation (mm)</th>
<th>% of world’s land area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperarid</td>
<td>0.05</td>
<td>200</td>
<td>7.5</td>
</tr>
<tr>
<td>Arid</td>
<td>0.05–0.20</td>
<td>200 (winter)</td>
<td>12.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400 (summer)</td>
<td></td>
</tr>
<tr>
<td>Semiarid</td>
<td>0.20–20.5</td>
<td>200–500 (winter)</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400–600 (summer)</td>
<td></td>
</tr>
<tr>
<td>Dry subhumid</td>
<td>0.5–0.65</td>
<td>500–700 (winter)</td>
<td>9.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600–800 (summer)</td>
<td></td>
</tr>
</tbody>
</table>


such as centuries, they are less useful for capturing shorter-term fluctuations in temperature or precipitation. As a result, the task of interpreting recent climatic change can be challenging.

Under current scenarios some arid and semi-arid regions are likely to become hotter and most will not become significantly wetter (IPCC 1996; 2007; 2013). Regions facing reduced precipitation include the American Southwest, the southern Andes, western Australia, and the Sahara. Even small temperature changes may affect aridity. For example, a temperature increase of 2 °C without an increase in precipitation would increase potential evapotranspiration by 0.2 to 2 mm per day. Many scientists believe that anthropocentric emissions of greenhouse gases are partly responsible for increasing temperatures in all deserts, and studies of global climate change suggest that desert temperatures are increasing faster than the global average. Within the last 25 years of the twentieth century, desert temperatures increased at an average rate of 0.2 °C per decade in the Atacama. Among hot deserts the rate of change has been even greater. The Kalahari and Great Victoria increased at an average rate of 0.8 °C per decade. In comparison, the average global temperature change for the same time period was 0.45 °C per decade (UNEP 2006). In addition to modifying the temperature regime, climate change can disturb sensitive ecosystems. For example, scientists have determined that increasing levels of atmospheric carbon dioxide may lead to an expansion in invasive plant species within arid regions.

**Desertification**

Desertification is sometimes represented as a process involving the movement or expansion of deserts. In fact, the term is misleading since deserts have no direct impact on desertification. Desertification is not a new phenomenon but, rather, a process dating back thousands of years and taking place in diverse locations ranging from river valleys of the Middle East to loess plateaus in central China and from lands surrounding West Africa’s ancient city of Timbuktu to deforested highlands in Nepal. The term was made popular by Aubreville’s 1949 book *Climats, forêts et désertification de l’Afrique tropicale*, where desertification was described as the transformation of productive agricultural land into desert through human-driven soil erosion. Aubreville’s definition didn’t describe desert lands as expanding outward, but rather as human-induced degradation caused by tree cutting, poor cultivation practices, and the indiscriminate use of fire.
ARID CLIMATES AND DESERTIFICATION

Although there is no consensus regarding its specific causes, most definitions treat desertification as a transformation of semiarid regions into arid land as a result of human-caused degradation. The United Nations has defined desertification as “land degradation in arid, semiarid and sub-humid areas resulting from various factors including climatic variation and human activities” (UNCED 1992). Land degradation associated with desertification is often represented in terms of reduced biological productivity through the loss of vegetative cover or as a result of soil compaction, decreased soil infiltration, and water or wind erosion (Schreiber 2006). In addition to increasing surface runoff, the removal of vegetative cover eliminates the binding action of plant roots that hold soil together. Subsequent erosion removes the organic content needed for the growth of soil organisms.

Desertification may also threaten biodiversity where land-use practices have allowed native plants to be replaced by invaders (Aronson and Shmida 1992). Dryland ecosystems are generally more susceptible to desertification in places where they have low resistance to stressors such as drought, fire, or human abuse. Given the difficulties in achieving consensus on its causes, some scientists suggest that desertification should be viewed simply as a nonreversible process associated with human activity (Prince 2002).

The first large-scale research on desertification was organized by the United Nations Educational Scientific and Cultural Organization (UNESCO) in the 1950s. During the 1960s a severe drought contributing to famine in areas south of the Sahara focused media attention on problems associated with the world’s semiarid lands. Subsequently the UN General Assembly sponsored an international convention in Nairobi, Kenya, in 1977 to discuss methods to mitigate desertification. In 1994 the UN Convention to Combat Desertification (UNCCD) was formed to foster cooperation in efforts to improve soil productivity and lessen the impacts of drought. That same year the UN General Assembly designated June 17 as World Day to Combat Desertification to raise awareness about the risks of drought and land degradation.

Desertification is estimated to be responsible for the declining productivity on more than 12 million hectares of land each year affecting more than a billion persons (UNCSD 1995). As noted by UNCCD, desertification impacts “the world’s poorest, most marginalized and politically weak citizens” (UNCCD 2006). As a result of its impact on populations, desertification is considered a significant threat to human welfare, especially in areas where local communities focus on a narrow range of economic activities such as farming and grazing.

Most scientists suggest that desertification should be viewed as a continuum of change extending from slight to severe. Rather than being treated as a strictly physical process, desertification is often framed in terms of overlapping economic, political, social, and natural factors. For example, government policies may favor unsustainable uses of land that contribute to degradation such as promoting sedentary farming over nomadic herding in areas more suitable for grazing. Coupled with drought, desertification may lead to prolonged famine, threatening food security in countries that have difficulty sustaining agricultural losses. In some cases population pressure or conflict may force people to occupy ecologically fragile areas, contributing to land degradation. Policies to mitigate desertification are often unsuccessful. For example, during the early 1980s a government-sponsored resettlement program aimed at moving people out of marginal lands in Ethiopia led to even greater levels of degradation. Land tenure policies may aggravate the impacts of drought and desertification when people who don’t own
land are reluctant to implement soil conservation practices. The impacts of desertification have been felt in locations far removed from where degradation is taking place. For example, airborne particles from dust storms in the Gobi Desert have degraded the air quality in Chinese cities located hundreds of miles away.

A common misconception is that droughts are the principal cause of desertification. Under most circumstances, droughts will not increase the likelihood of desertification unless the carrying capacity of nonirrigated lands has been exceeded. However, when it is combined with the abuse of land, droughts can amplify human hardship and suffering. Overgrazing leading to a loss of topsoil as a result of wind or water erosion is often cited as a cause of desertification. Plants in semiarid areas are often consumed by wild animals that follow patchy rainfall. This pattern has been emulated by pastoral humans who move their herds with the shifting availability of rainfall and forage. In some areas fencing has contributed to overgrazing by preventing the movement of wild and domestic animals. Soils are damaged through compaction from animal hooves that reduce permeability and increase susceptibility to wind or water erosion. In some locations water wells have enabled livestock to remain in areas with limited surface water, leading to the damage of surrounding vegetation and soil. In addition to soil loss, overgrazing may lead to a decline in palatable grass species or to the replacement of perennials with annual species less resistant to soil erosion. The suppression of wildfire has also contributed to changes in the composition of vegetation. Woody vegetation has also been lost through firewood collection, especially in areas near human settlements. The reduction in woody plant cover leads to accelerated wind and water erosion and may contribute to the destabilization of dunes.

Poor agricultural practices have also been cited as a contributing factor to desertification. For example, monocropping can lead to greater soil exposure and increased rates of erosion. Likewise, overcultivation may result in losses to soil productivity where exposed topsoil becomes susceptible to water erosion and gullying. When soils become dry they are more susceptible to wind erosion. In some locations poor irrigation practices may result in salinization – the collection of salt on soil surfaces. Within some areas increasing demand for food, driven by expanding population, has encouraged the shortening of fallow periods, resulting in the depletion of soil nutrients. Finally, the cultivation of marginal lands has contributed to desertification in locations where farming has replaced nomadic pastoralism. Short-grass steppe lands are particularly susceptible to erosion after being plowed, as nutrients anchored by grass and soil are removed by blowing winds.

Since the 1970s land degradation and drought have focused media attention on desertification within Africa’s Sahel region. Forming a 200 km wide transition region between the Sahara to the north and humid lands to the south, the Sahel has been inhabited by herders and farmers for thousands of years. A unique characteristic of the Sahel is alternating wet and dry periods that last several decades. For example, the region experienced a wet phase during the late 1880s that led to years of higher than average agricultural productivity followed by an extremely dry period beginning in 1900. Rainfall increased again in the 1930s until a sharp decline in the 1970s and 1980s that culminated in severe famine and the deaths of thousands of people and livestock.

Although a few parts of North Africa and Mauritania are threatened by migrating sand dunes, there is a false impression that the Sahara is encroaching into the Sahel with moving sands that swallow farmland. Within much of the Sahel
a natural belt of brush and trees protects agricultural lands from desert encroachment. The Sahel’s vulnerability to desertification is due to its strong dependence on rainwater for supporting livestock and sedentary farming. Contributing to problems in the Sahel are cultural practices that promote livestock as having value beyond serving as a food source.

Estimating impacts of desertification

Desertification can be difficult to measure because the world’s arid regions were created by natural processes over long periods or time and have expanded or grown smaller independently of human impacts. In many cases, borders of desertified areas are difficult to define precisely because vegetation exhibits a gradual transition from arid to more humid conditions, with xeric plants interspersed with other forms of vegetation across broad zones. Although there is no accepted threshold for determining when land is considered “desertified,” the United Nations Conference on Desertification has created an index for identifying the severity of desertification according to declining agricultural productivity. A 10–24% drop in productivity is considered moderate, 25–50% is high, and greater than 50% is considered very high.

As noted by Helldén and Tottrup (2008), scientists disagree about the geographic magnitude of desertification. Therefore, a significant need exists for better information on the spatial distribution and severity of desertification through the collection of baseline data and monitoring programs. In recent years methods for tracking changes in the productivity of vegetation over large areas have evolved including satellite remote sensing utilizing the Normalized Difference Vegetative Index (NDVI). However, while NDVI values are useful for estimating the condition of vegetation, they are often less helpful for comparing specific locations across time periods (Burgan and Hartford 1993).

Areas impacted by desertification

Desertification is found on every continent except Antarctica. Within Africa the largest areas impacted by desertification are the rain-supported grazing and agricultural lands of the Sahel and portions of North African plains. Within semiarid regions of sub-Saharan West Africa wind erosion tends to be a more significant factor, while fluvial erosion has been more severe in Maghreb countries such as Morocco, Algeria, and Tunisia. Overgrazing has also impacted arid regions of Asia such as rangelands in Central Asia and the Middle East. The productivity of rangelands in much of the Middle East and western portions of Pakistan is among the lowest found within the world’s semiarid lands. Fluvial erosion on cultivated lands has also been a problem in Pakistan, India, and highland areas of China. Growth in human and livestock populations have increased grazing pressures on some lands while cultivation has encroached into pastoral lands. At the same time, improvements in transportation have made formerly remote areas more accessible to grazing. Much of Australia has also suffered from the impacts of desertification. About one-quarter of the continent’s arid regions are not occupied because of unpalatable spinifex grasses or their distance from population centers. The most severe desertification can be found in the saltbrush and bluebrush areas of South Australia and New South Wales, mostly due to overgrazing.

Desertification has also impacted North America where overgrazing has severely altered native grass cover. The history of climate change in desert and semiarid regions of the American West and Southwest has been studied extensively
using indicators of past vegetation and climate that include the analysis of sediment and pollen samples. In some areas of North America desertification has been accelerated by overgrazing that has removed perennial grasses, leaving soils easily eroded by running water. Affecting more than 400,000 km², North America’s “Dust Bowl” created huge dust storms during the 1930s that reached the eastern United States, destroying crops, killing animals, and removing enormous quantities of topsoil. Areas of South America have likewise been impacted by desertification including the semiarid pampas of Argentina. Severe erosion attributed to deforestation, overgrazing, and the invasion of salt-tolerant plants has affected parts of coastal Chile and Brazil’s state of Minas Gerais.

**Mitigating desertification**

Techniques for mitigating desertification include farmer-managed natural revegetation and agroforestry – the planting of trees in and around crops. In some areas local initiatives to build wind barriers or to close areas to grazing and cultivation have been successful in combating desertification. Such methods have helped reduce evapotranspiration and improve crop microclimates. In China’s Tengger Desert researchers have attempted to stop drifting sand using a grid network of straw fences that reduce wind velocity. Another method has been the promotion of soil conservation through intercropping using perennials as anchors against water and wind erosion. Indigenous knowledge has also been important in fighting desertification. For example, on the steppe lands of Syria known as the “Badia,” Bedouin communities that maintain herds of small ruminants, camels, and equines cooperate with representatives of the International Fund for Agricultural Development in the rehabilitation of rangelands. Bedouin leaders have participated in the development of management plans for defining grazing limits established by evaluating seasonal rainfall patterns and the condition of vegetation. As a result of this effort some areas have been temporarily placed off-limits to grazing while others have been replanted or reseeded using forage plants acclimatized to local circumstances. An initiative organized by the African Union and supported by many regional and international organizations called the “Green Wall” targets desertification within the Sahel through a campaign to plant trees.

**SEE ALSO:** Biodiversity; Climatology; Desertification

**References**


ARID CLIMATES AND DESERTIFICATION


Further reading


Geographers have never seemed more open to, and interested in, art as they have been in the first decade or so of the new millennium. Yet, geography’s relationship with artists and artistic practices is arguably as old as the discipline itself. Indeed, to inspect the major phases of geographical knowledge-making is to find an important place for artists, artworks, and arts practices. Moreover, at key phases of geography’s development, the discipline’s relationship with art becomes a point of critical formulation, offering a site of epistemological evolutions that come to shape disciplinary directions and hone analytic foci (Hawkins 2014). Examples include: artists onboard ship during the long period of “capes and bays” geography, the evolving place of artists and writers in the extension and critique of the chorographic traditions of regional geographies, the role of art in the evolution of humanistic geography, the place of creative practices within qualitative GIS, and within questions of the body and affect raised by nonrepresentational theorists, together with the allied debates around aesthetics and politics. To overlook this rich history is not just to pass over an important influence on geography’s development; it is to miss out on the clear analytic value that an appreciation of the diverse roles art has played offers for exploring geography’s current phase of artistic engagement.

This relationship is not one-sided. Art theorists, artists, curators, and arts organizations have also invested in an evolving relationship with the discipline of geography, its substantive concerns and its knowledge-making practices. If art, with its concerns with site and situation, has come to feel especially geographical of late, this is not a new set of preoccupations. We can think, for example, of the intersection of art and urban theory throughout the twentieth century, as well as the broader spatial turn in the arts and humanities that found perhaps its clearest form over the last two decades (Rendell 2006). Artists have also long produced artifacts and worked with practices, and knowledge (modes of mapping, exploration, preoccupations with place, and so on), that we might more normally think of as geographical.

There is therefore much common ground between geography and art, shared conceptual and methodological territory whose occupations were not, until recently, explored terribly closely. As geography’s relationship with the humanities continues to evolve, there is a growing concern to explore the form and kind of this geography–art relationship. There is interest in mapping the multiple intersections of these expanded disciplinary and practice-based fields, around both substantive topics such as landscape or the body, as well as around common practices such as mapping, and the geographies of different mediums, whether paintings, photographs, sculpture, performance, or community-based work (Dear et al. 2012). An ongoing issue has been to attend to the different forms of these art–geography intersections, with distinctions being made between dialogues, which bring geographic ideas and thinking to bear on artworks and vice versa, and doings, wherein geographers and artists collaborate, or work

Harriet Hawkins
Royal Holloway, University of London, UK

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0663
separately, to bring artistic practices together with geographical practices and procedures of knowledge-making.

The broad spectrum of research that links geography and art displays a diversity of substantive foci (landscape, urban space, materialities, nonhuman geographies, environmental matters), analytic approaches (iconography, sensory ethnography, ethnography), and methodological questions. From amidst this diversity emerge three enduring areas of concern: the geographies of artworks and art worlds – namely the production, consumption, and circulation of art; the role of artistic practices in “making” worlds; and the role of artistic practices in geographical knowledge-making practices and methods (Hawkins 2014). These three dimensions are common not only to geography and art relationships, but also to creative geographies more generally. As such, they are applicable concerns with respect to creative writing, music, film, and craft practices (Daniels et al. 2012). This is not to collapse important distinctions of method, medium, and politics, but to enable tensions to be drawn out between these different modes.

Geographies of art worlds and art-making

Artistic production has long been recognized as a situated activity with complex and manifold geographies. Discussions of such geographies take account of the role of the art world, as well as the nested sites of studio, home, community, city, nation, and so on (Bain 2013). Geographers are very aware that the geographies of art are not just those of the initial site of the work’s production and consumption, and those places and locations that feature within the works, but are also concerned with the circulation of artworks, the various contexts through which they move, and the sites at which they are consumed. The sites and spaces and spatial distributions of artistic production, have been explored via a range of methodologies and to a variety of ends. These include economically and politically orientated studies of the networks, clusters, and scenes of the art world; detailed studies of the geohistories and geopolitics of the international diffusion of artistic materials, styles, and techniques; the sociologies and histories of art school or colonies of artists, as well as ethnographic studies of the embodied, material practices of studio production (Kaufmann 2004; Sjoholm 2012; Bain 2013).

Geographic research on art worlds has often taken the spatialities of artistic production as an analytic concern, whether such spatialities be those of local agglomeration – clusters – or the intersecting local and global networks of transnational production, circulation, and cultural exchange. Questions of creative clustering have developed out of art historical interest in arts colonies or schools, as well as from the economic imperatives of the creative cluster. Here we find artistic outputs shaped by social relations, common conceptual questions, and shared inspirational sources. For more economically inclined scholars these relations translate into “traded” and “untraded” interdependencies. If the former consist of connections formed from sharing, exchanging, or trading (for money, goods or services) painting materials, framing services, and so on, the latter might include the production of and benefit from the formal and informal social exchanges that constitute a shared creative and intellectual atmosphere. Informing these ideas have been discussions around “creative scenes,” whether these be historical studies of the evolution of particular locations or economic work on creative cities and city branding (Rycroft 2011). Running in parallel with such enquiries is research that deploys postcolonial
scholarship and ideas of transnationalism to query the mobilities of people, objects, and ideas in the production of different art forms from painting to performance work (Rogers 2014).

Accompanying work on art scenes is a growing body of geographic scholarship concerned with the sites, spaces, practices, and technologies of artistic projects, whether these be situated in the field, the studio, the street, the community, or the landscape. Across these sites attention is paid to the intersections of the material and immaterial dimensions of creative production to questions of embodied practice and to the role of technologies. Discussions of the studio, for example, explore the role of studio spaces as sites of inspiration, for thinking and dreaming, for archiving, for identity formation (Sjoholm 2012). In discussions of art-making in the landscape, the immediacy of “en plein-air” experience in shaping the artworks produced has been a key analytic dimension. In other words, how the embodied experience of being in the landscape informed the works, wherein what is being “pictured” becomes a more-than-visual account of what was sensed. In such a context, the role of technologies – from the camera to the paint-box – in mediating and shaping these relations is an important concern.

In considering artworks made on the street or in the community geographers have embraced the idea of these works as technologies of connection. Geographers interrogate these works for the form and kind of connections made between community and artwork, but also the relations within the community that are catalyzed by the artwork. Such an approach takes seriously the social “work” that these art projects do at the sites of their creation. Such “work” might relate to border and security concerns, to questions of urban super-diversity, rural community disenfranchisement, or the alienation brought about by urban gentrification (Hawkins 2014). It is clear from thinking about this work that the geographies of creative production cannot be set apart from the geographies that the work creates within the “frame,” nor from the geographies of its consumption.

Geographical scholarship on galleries as a site of artistic consumption has not perhaps been as voluminous as might be expected. Indeed a general trend in geographical scholarship has followed artworks as they moved from “studio to situation,” being more concerned with the artworks produced and consumption beyond studio and gallery sites than with reflecting critically on gallery spaces. If the geographies of galleries have been largely neglected, questions of audiencing have gained attention in the context of art produced and consumed in homes, on streets, in communities and in the landscape. While such works might involve a gallery-based period of display, often after their initial phase of production and consumption, such later forms of consumption are generally overlooked by geographers more intent on studying the transformative effects of these works on the spaces and peoples they are produced in and by. An appreciation of the potential of artworks to “make” worlds demands an analytic sensitivity to the intersections of the geographies of the works’ production and consumption – whether these be on the street, in the studio, or in the gallery – with the geographies within their “frames.”

Artistic practices and the “making” of worlds

That artworks do not so much “picture” worlds as “make” worlds is a central premise for geographers studying art. Looking across the last few centuries of geographical relations with art is to witness the evolution of geography’s appreciation
of the “work” art does in the world. In short, there is a shift from artworks understood as “picturing” worlds, with the latter verb understood to be an act of mimetic representation, to an appreciation of the force and potential of the making of worlds that happens in that picturing (but also of course the “sounding” and “performing”) of the world. This attention to the productive “doings” of art, its makings and shapings of worlds, subjects, and knowledge, has become one of the distinguishing features of geographical scholarship on art.

Picturing places: geographical imaginations

The value of art for providing “packets” of information about a place is how expeditionary art from the eighteenth and nineteenth centuries has often been discussed. In the midst of this “Age of Exploration,” artists commonly accompanied great voyages – such as those of Captain James Cook – in the service of science and empire. Indeed, the casting of aesthetic practices in relation to military and commercial interests, and as part of the production of enlightenment science, earned this era of artistic practice the moniker “the winter of the imagination.” An important figure in the context of exploration art is Alexander Von Humboldt, the eighteenth-century polymath for whom artistic, and aesthetic practices more generally, were a mode of apprehension equal to, and indeed a building block of and catalyst for, scientific knowledge. The influence of Humboldt’s geo-aesthetics can be traced into the eighteenth- and nineteenth-century formation of branches of geography as different as geomorphology and cultural geography. In geography’s twentieth century evolution, Humboldt’s scholarship is a touch-stone for a humanistic critiques of a scientized geography, and more recently it backstops both human geographers’ embrace of aesthetics and geopolitics, and physical geographers’ recognition of the role of enchantment and the arts in critical approaches to their methods and scholarship (Hawkins 2014).

The richly varied visual cultures of exploration produced during the sea-going voyages of the eighteenth and nineteenth centuries – from large oil paintings to informal sketches in diaries to scientifically accurate botanic drawings – were valorized for their situated geographies of production. Being produced on the spot such visual cultures provided privileged forms of evidence. They were valued over the written word for their production of “more perfect” accounts of places and species than words could create. Crucially such pictures of place were formative building blocks in the geographical imaginations of the European Enlightenment era. They provided not only scientific source material, but also pictured “exotic” places and others for those European populations who could never hope to travel there. Sent home ahead of explorers, they were hung in galleries, stored and displayed in scientific institutions (including the Royal Geographical Society), and reproduced in newspapers, their circulation creating globally encompassing geographical imaginaries (Driver and Martins 2005).

Such geographical imaginations are, of course, not neutral picturings of place. Rather, the imaging of places far distant was the result of a complex set of aesthetic geographies, often overlaying the terrains of non-Western places with the pictorial conventions of the Western world. William Hodge’s (1744–1797) images of Tahiti, for example, picture the island through a Western optic, a lens that owes much to the aesthetics of classical Italianate painting. This optic shaped everything from lighting to compositional layout, the types of plant species included, and the mode of picturing the inhabitants of
places. The complex geographies of these imagined, pictured worlds, importantly, did not just take form on the surfaces of the painted canvas, but also came to find form in the materiality of landscapes, the textures of social relations, and the topographies of Western imaginaries that were created as a result of these paintings.

Landscape art: from iconography to embodied experience

The force of painted aesthetics in shaping landscapes and lives is foregrounded in geographical scholarship on landscape painting. Scholars such as Stephen Daniels (1993) focus on the “Marxist duplicity of the landscape,” developing an interdisciplinary approach that begins from the Marxist cultural studies scholarship of Raymond Williams and John Berger. The empirical focus of this scholarship included the works of great English and American painters of the eighteenth and early nineteenth century. The surfaces of these canvases were examined as “veils,” aesthetic screens drawn across the landscape. Exploring both the symbolic codes and the affective and emotional qualities of the landscape, its beauty, sublimity, and order, such aesthetic conventions were understood to constitute a covering over and a turning away from the uncomfortable “truths” of the landscape. In many cases of the English picturesque – in paintings, for example, by John Constable, and landscape gardens by Humphrey Repton and Capability Brown – this turning away involved the picturing or recreation of a rural idyll, an aesthetic ideology that “hid” from view the “dark side of the landscape,” masking the squalid living conditions of the rural poor, their existence at the edge of starvation, and the widespread unrest that existed in the countryside at the time. Such ideological picturings of place produced and reproduced, indeed legitimized, particular types of power regimens, especially when, in addition to being “pictured,” they came to shape the materialities of the landscape. Migrating from the surface of the painted canvas to the earth, soil, stone, and flesh of rural landscapes and communities, these aesthetics re-aligned roadways and coppiced woods, and ensured houses were left to tumble down in picturesque ruin. Such aesthetic migrations were not just restricted to the national environs of their creation, but were rendered mobile by practices of colonization. The global circulation of these aesthetic principles was one way in which colonial power was made manifest in the landscapes of plantations, and in the lives and the bodies of their workers.

Nuggets of experience: more than “presenting”

Geographical perspectives on the ideology of aesthetics, and the transformative work such aesthetics did on subjects and urban and rural landscapes, dominated geographical scholarship on art for much of the late twentieth century – and rightly so (Pinder 2005). A parallel understanding running throughout geography’s engagement with arts in the twentieth century has been ideas of arts as offering geographers access to experience. Yi Fu Tuan (1973) and others writing of geography’s relationship with literature and art have been rather disparaging of this relationship, suggesting that geographers were scanning artistic and literary sources for “nuggets of experience.” In moves described as akin to stamp collecting, geographers become merely intellectual middlemen, seeking subjective accounts of place and experience in arts as sources, and sorting them for collation as data within the chorographic impulses of the regional geographies of the time.

As geography’s engagements with art and its own theoretical interests evolved, a rather different sense of the value of art in relation
to subjective experience has begun to emerge. In the light of a growing interest on the part of feminists, nonrepresentational theorists, and others in questions of the body, the senses, affect, and emotion, arts practices became re-valued. Within the context of these concerns painting, sound and installation art, dance, urban walking practices, video practices, among other arts practices, have become situated as sites in which geographical sensibilities can become extended and attuned to our sensing and experiencing of space and place. These studies often have a two-fold imperative: a remapping of bodily senses, challenging the privileged position of vision by asserting the value of other senses – for example, hearing, taste, and touch – in knowing place and self, and by exploring the complex intersection of these external senses with the internal senses. The second imperative which perhaps marks out geographic studies from other approaches, is a concern with how in knowing space, place, and self, we are not just concerned with the bodily senses, but also with their intersections with questions of habit, memory, and prior knowledge of these and other places (Hawkins 2014).

These ideas of the body have also reconfigured the set of analytic modes and questions geographers ask of and with art. If geographical studies of historical landscape paintings would commonly approach these paintings as a set of symbols to be decoded, then recent reconfigurations attend to art as productive of sensations and experiences. For some geographers this resulted in a revisiting of ideas of landscape in relation to abstract practices. In these works artists engage their embodied experiences of landscape as the subject of their work, creating canvases that are less concerned to picture than to create evocative blocks of sensation and experience that audience members feel in their bodies as well as explore with their minds.

In such a context artistic practices are the result of encounters with the landscape and environment, but also create encounters with landscape and environment for their audiences. From the recognition that arts practices enabled geographers and others to explore everyday embodied experiences of being in and moving through the world, recent geographical research has begun to examine the potential of artworks to open us out to experiences of the world and move us beyond the capacities of our human senses. This is to understand art as potentially creating ways of knowing the world beyond what is possible for our human cognition and our sense organs alone to develop. Art, as a result, holds the potential to create encounters that might open us out to experiences of airy and earthy matters, nonhumans, the organic, and the inorganic in ways that are important for shaping our future environmental imaginaries.

Artistic practices and the production of geographic knowledge

The place of artistic practices in producing geographical knowledge has evolved since artists accompanied explorers on board ships producing rich visual cultures of distant places. From this foundation based on the valorization of mimesis, it is possible to explore the alternative value that has been found in creative methods for geographers, which have become variously understood as creative, experimental, or artful geographies. Such work sees geographers across the discipline coming to embrace various creative practices – film-making, visual art, creative writing – both as the means through which research can proceed and by which it can be communicated and presented.

Primary among the justifications for practicing such creative geographies is to understand them
as a response to the discipline’s orientation toward embodied and practice-based doings. Such orientations demand the means by which to engage, research, and re-present the sensory experiences, affective atmospheres, and flows of life. Geographers’ development of artistic practices, either alone or in collaboration, is understood as a means to enable such messy, fleshy, and experiential modes of engagement to be explored and developed. Furthermore, creative writing practices offer a means by which to begin to write about, to grasp, such experiences and convey them to others.

For other geographers engagement with the artful proffers the means to grasp the messy, unfinished, and contingent – in everything from spatial imaginaries to knowledge-making practices – that a more scientifically inclined geography might orient us away from. Artful engagements might include critical reflections on the spaces and sites of the archive and other knowledge-making institutions; they might reflect a concern with how we know and write about places and spaces; or they might be a response to an epistemological turn within geography that seeks to do away with categories, binaries, and fixities, in favor of multiplicities, hybridities, and becomings. Rightly, or wrongly, geographers across the discipline, whether historical, concerned with nonhuman relations in the Anthropocene, or with modes of contemporary urban living, have found value in artistic practices and creative methods as a means to confront the epistemological challenges they encounter at their research sites.

A further reasoning behind geographical engagements with creative methods has been in terms of the contemporary extension of the long appreciated possibilities such methods offer for enrolling audiences other than academics. Long recognized to form affective means of communication, scholars are now recognizing the value of arts practices as proffering a means by which to constitute new and engaged publics for their work. Thus we see geographers creating plays, developing arts workshops, and configuring creative activities based on the recognized “social” work that aesthetic practices have been understood to do. As well as being driven by intellectual agendas, such practices should also be situated in the context of the contemporary political agendas and climates of higher education wherein concerns with moving research beyond the academy, and thinking through politics in a new way loom large. Where perhaps there is work still to be done is around the seeming absence of a more critical radical politics from these creative geographies (Marston and De Leeuw 2013).

GeoHumanities: looking to the future

Geography has had something of a return toward art and arts practices, as both empirical objects of study and as the methods by which research can proceed, and through which it can be disseminated. Furthermore, art practices and theories have never seemed more geographically engaged. The result has been a ground-swell of practice and attention to what has come to be called the GeoHumanities, with these intersections becoming increasingly institutionalized in interdisciplinary masters courses, in practice-based doctoral work undertaken within geography departments, as well as in the evolving collaborative research practices of academics and artists (Dear et al. 2012).

If these productive intersections are to continue to fulfill their not inconsiderable promise there are a number of perhaps self-explanatory, and also perhaps not so self-explanatory issues to consider. The first of these relates to an ongoing
need to be clear and respectful about the knowledge and practices being deployed and developed in these intersections. Geographers may like the idea of being artists, writers and film-makers, but enthusiasm should not outweigh a concern for the skillful nature of these practices. Equally, there should perhaps be a shared sense of such intersections not only being based in comings together that are predicated on the in common and the collapse of differences. One of the delights of finding similar interests – in practices of mapping, of exploration, or in ideas and terms such as region, topography, or vision – has to be also to commit to probing the differences between these understandings. This is not to create rifts within newly formed relations, but to ensure their intellectual integrity, and also to aid in the mutual development and evolution of such ideas.

Finally, and perhaps most importantly, in these comings together we should not lose sight of the importance of attending carefully to the geographies of creative production and consumption, not least because such perspectives enable a close attention to the kinds of “work” that artistic objects and experiences do in the world. Geographers need to attend to the changing geographies of art over the twentieth century, to remain attuned to how the imaginaries and materialities of places shaped by canvases denote but one important sort of work art does in the world. This variety needs to be appreciated alongside the refiguring of artistic regimens around eventful and relational aesthetics, wherein, aesthetic practices create opportunities for social encounters in which individual subjects and communities can become transformed.

There is a need to attend closely to the type of encounter with the world that is being created by art forms, to acknowledge how they affect us as academics, as makers, as geographers, but also what impact they might have on their audiences. This is to become attuned to the political possibilities, but also the responsibility associated with the art that geographers study, the creative practices deployed within their research methods, and the artworks that they might be involved in creating.

**SEE ALSO:** Cultural geography; Cultural turn; GeoHumanities; Humanistic geography; Imaginative geographies; Interdisciplinarity and geography; Landscape iconography and perception; Nature, art, and aesthetics; New cultural geography; Nonrepresentational theory; Phenomenology; Representation; Situationists/situationist geography; Visuality

**References**


Artificial neural networks in geospatial analysis

Sucharita Gopal
Boston University, USA

Background

Artificial neural networks (ANN) are computational models inspired by and designed to simulate biological nervous systems that are capable of performing specific information-processing tasks such as data classification and pattern recognition. ANN seeks to replicate the massively parallel nature of a biological neural network. A neural network is a system composed of many simple processing nodes whose function is determined by network structure and connection strengths. Similar to biological neural networks, AAN can acquire knowledge through a learning process. Interneuron connection strengths known as synaptic weights are used to store the knowledge and make it available for use. Each artificial neuron receives one or more input signals and sums them to generate an output. Usually the sums of each node are weighted, and the values are passed through a nonlinear function known as an activation function or transfer function. If a weighted sum exceeds some predetermined threshold value, then an excitatory output (1) is produced. Otherwise, an inhibitory output (−1) is produced. The transfer function can be a sigmoid function or other nonlinear, piecewise linear or step function. Researchers in a variety of disciplines including engineering, psychology, mathematics, and physics have contributed to the explosive growth in the field that has continued to this day. Two standard texts in the field are Bishop (1995) and Ripley (1996).

Neural networks can be used to address a wide variety of real-world problems. They have the ability to learn from experience in order to improve their performance and dynamically adapt themselves to changes in the environment. In addition, they are able to deal with fuzzy or incomplete information and noisy data, and can be very effective, especially in situations where it is not possible to define the rules or steps that lead to the solution of a problem. Hence they are fault tolerant. In addition, the ANN information-processing model is inherently parallel. Today ANNs are used in a variety of disciplines including engineering, finance, artificial perception, and control and simulation.

The wide use of ANN in geospatial science stems from their roles in spatial data processing and analysis. Satellites orbiting and imaging the Earth produce massive amounts of geospatial data, on the order of tera- to peta-bytes. ANN, programmed on parallel neurally inspired hardware architectures, can analyze and classify this vast amount of data quickly and draw meaningful insights via mapping or modeling. ANN’s generalization capability in dealing with classification across multiple spatial scales and resolutions is a significant advantage. Moreover, many physical processes modeled in the geospatial sciences require accurate knowledge of process dynamics and such knowledge is often unknown. In this case, ANN can be used in function approximation.

The learning that occurs in ANN is not affected by the integration of multisource data. The learning process is robust and fault tolerant.
Newer learning paradigms such as *semisupervised learning* or *self-supervised learning* can deal with incomplete data, overcoming the difficulties and expenses involved in gathering training labels for geospatial sensor data. Recent studies have attempted to make ANN more spatially explicit, either by introducing fundamental spatial principles such as spatial autocorrelation directly into the neural network structure or during post-processing and labeling using spatial neighborhood relationships. When using an ensemble approach, ANN can assist in characterizing the spatial heterogeneity of the Earth’s surface and spatial uncertainty in labeling.

**Types of ANN**

There are several types of ANN based on the learning paradigm, architecture, and function.

**Learning paradigm: supervised, unsupervised, and semisupervised learning**

In supervised learning, the ANN is supplied with a sequence of both input data and desired (target) output data; the network is thus told precisely by a “teacher” what should be emitted as output. During the learning phase, the teacher can “instruct” the network about how well it performs (“reinforcement learning”) or what the correct behavior would have been (“fully supervised learning”). The ANN learns the association between input and output classes until some criterion of successful learning is met. A commonly used metric is the mean squared error, which tries to minimize the average squared error between the network’s output and the target value over all the input-output training data pairs. The well-known backpropagation algorithm for training tries to minimize this cost using a gradient descent function.

In the ANN methodology, the sample data for supervised training is often subdivided into *training*, *validation*, and *testing* sets. The distinctions between these subsets are crucial. Ripley (1996, 354) defines the following: “Training data is a set of examples used for learning that is to fit the parameters [weights] of the classifier. Validation data is a set of examples used to tune the parameters of a classifier, for example to choose the number of hidden units in a neural network, while testing data is a set of examples used only to assess the performance [generalization] of a fully specified classifier.”

In unsupervised learning, the training scheme only consists of input data. The ANN discovers some of the properties of the dataset and learns to reflect these properties in its output. The issue of what exactly these properties are, that the network can learn to recognize, depends on the particular network model and learning method. Unsupervised ANNs can discover the underlying structure of the data, as well as encode, compress, and transform data values. This type of learning presents a biologically more plausible model of learning. The self-organizing map (SOM), a well-known unsupervised ANN, learns to produce a low-dimensional discretized representation of the input space using a neighborhood function to preserve its topological properties.

Real-world applications of supervised ANN are popular but face a number of practical challenges, such as scalability and limited availability of labeled training data. In spatial mapping, obtaining labeled datasets is expensive, and sometimes impossible. A recent solution is the development of a new learning paradigm called *semisupervised learning* (SSL). In SSL, the user strategically selects the data to be manually labeled, and then lets the ANN iteratively retrain itself on its own output using the remaining unlabeled data. A variation of SSL is the Self-Supervised ARTMAP, which uses only a
subset of input features during training. This is a great advantage in GIS and remote sensing, as new information about Earth’s features is constantly being acquired.

ANN network architecture

Figure 1a shows a typical structure of an ANN network consisting of processing units (neurons) arranged in two layers, and links (synapses) between the processing units in the different layers. There are input nodes and output nodes representing inputs and outputs in a supervised ANN. In addition, there are hidden units that can vary depending on the nature of the problem. Experimentation is often required to determine the best number of hidden units. Too many hidden units may lead the network to overfit the training data, thus reducing generalization accuracy. On the other hand, too few hidden units may prevent the network from being able to learn the required function. Each connection between nodes has a weight associated with it that is adjusted during learning. In addition, an activation function converts the processing unit’s weighted input to its output activation.

ANN network function: feedforward and recurrent networks

ANNs are also differentiated based on whether there are cycles or loops in the network. In the feedforward network (Figure 1a), information signals move in one direction, from the input nodes, through the hidden nodes (if any), to the output nodes. There are no cycles or loops in the network. Hence these are called acyclic graphs. An example is a multilayer feedforward neural network. In contrast, recurrent networks (Figure 1b) are cyclic graphs since they contain cycles or loops. Some recurrent networks (e.g., Adaptive Resonance Theory – ART) can be extremely complicated.

Three ANN models

Multilayer perceptron using backpropagation

A popular ANN classifier is the feedforward multilayer perceptron (MLP) architecture. An MLP is composed of layers of processing units in a directed graph that are linked through weighted connections (Figure 1a). The first and
last layers consist of the input variables (e.g., spectral bands) and output classes. The intermediate layers, called hidden layers, provide the internal representation of neural pathways. An MLP ANN maps sets of input data onto a set of appropriate outputs. Except for the input nodes, each processing unit has a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation to train the network.

Learning occurs in the MLP by modifying connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. This process is continued until the weights in the network have been adjusted such that the network output has converged, to an acceptable level, with the desired output. In the next phase, the fully trained network is given new data, and the processing and flow of information through the now-activated network should lead to the assignment of the input data to an output class.

Self-organizing map

The self-organizing map (SOM) is a popular tool for mapping and clustering high-dimensional data (Figure 1b). In a SOM, neurons or units are arranged in a rectangular or hexagonal group of units of predefined dimensions ($m \times n$ rows and columns). The number of units is usually small relative to the dimensionality of the input data. The SOM algorithm computes an ordered mapping, a kind of projection that forces each of the input data records to map onto a defined grid during the iterative learning process. The goal of the SOM is to preserve the similarities between samples such that similar input data records map on to the same or neighboring units in the grid, while dissimilar data records are mapped on to non-neighboring units. Thus SOM can incorporate spatial neighborhood and spatial autocorrelation effects that are commonly encountered in GIS and spatial analysis. Over the last decade, SOM has been increasingly used in a supervised fashion.

At the start of training, the grid units are initialized via a random sample of $p$ observations from the input data. Each grid unit is characterized by a codebook vector that describes the typical pattern of that unit. The aim of the SOM algorithm is to update codebook vectors so that the input data are best described by the small number of grid units. During training, the distance between phenomena in the presented dataset and the codebook vectors for each of the units is determined by a distance measure. The unit whose codebook vector is closest or most similar to the presented sample is the winning unit or node. The winning nodes’ codebook vector is then updated so that the winning unit is made more similar to the presented sample.

Self-supervised ARTMAP

Self-Supervised ARTMAP is based on Fuzzy ARTMAP (Carpenter 2013). Fuzzy ARTMAP is based on ART, proposed by Stephen Grossberg in 1976. Fuzzy ARTMAP’s internal control mechanisms create stable recognition categories of optimal size by maximizing code compression while minimizing predictive error during online learning. Fuzzy ARTMAP incorporates fuzzy logic in its ART modules. ART overcomes the stability-plasticity dilemma; namely, how a neural network can learn quickly about new objects and events (plasticity) without overwriting or forgetting previously learned memories (stability).

In the first phase of self-supervised learning, the ARTMAP system is trained with a limited set of labeled data (input features and output classes), similar to learning in a traditional supervised ANN. In the next phase, the system
ARTIFICIAL NEURAL NETWORKS IN GEOSPATIAL ANALYSIS

continues to learn using an expanded set of input features with no labels or output classes. This “unsupervised learning” enables the system to build on its existing knowledge with new information, resulting in improved accuracy, compared to that of the initial trained system, without worsening its performance. This is demonstrated by Carpenter (2013) on a remote sensing (Landsat) database of land cover related to the greater Boston region.

Review of ANN applications in geospatial science

ANN models have been used in the geospatial sciences since the 1990s. A literature review was conducted using Google Scholar to identify existing publications of relevance with key terms “neural networks and GIS,” “spatial analysis,” “spatial interaction,” or “neural networks and remote sensing.” The search was done with a five-year interval starting in 1990. A majority of publications were peer-reviewed journal articles, conference reports, and technical papers. Four key findings are: first, the number of papers using neural networks in remote sensing was much larger, likely because remote sensing is multidisciplinary, including fields such as engineering, mathematics, and physics. The number of inputs can vary up to hundreds, given the data capture capabilities of NASA’s Earth Observing System Data and Information System (EOSDIS). Classification is typically performed using supervised learning methods. In many studies, the performance of ANN has been shown to be superior to that of traditional classifiers such as Maximum Likelihood (Carpenter et al. 1997).

The measurement and characterization of changes happening on the Earth’s surface at a variety of spatial and temporal scales has long been of interest to geospatial scientists. In this context, ANN has helped to identify patterns of change through time as well as to distinguish abrupt (deforestation or forest fires) and continuous (coastal erosion) changes. This application involves mostly supervised neural networks or hybrid approaches, where classification is done at two time periods and the differences are then classified and mapped.

The main objective of clustering is to group large amounts of data into meaningful categories last category, optimization, is less prevalent in the remote sensing field. Fourth, function approximation and estimation applications are popular in climatology and hydrology and in modeling physical processes.

Common applications of ANN

Let us consider the most common applications of ANN geospatial science – classification, change detection, clustering, function approximation, and forecasting or prediction.

The goal of classification is to assign an input pattern (like spectral bands or sociodemographics) represented by an input feature vector to one of several output classes. The best known application is land use/land cover classification in remote sensing and GIS involving multisensor, multidate, multisource, and multiscale data. The number of inputs can vary up to hundreds, given the data capture capabilities of NASA’s Earth Observing System Data and Information System (EOSDIS). Classification is typically performed using supervised learning methods. In many studies, the performance of ANN has been shown to be superior to that of traditional classifiers such as Maximum Likelihood (Carpenter et al. 1997).

The measurement and characterization of changes happening on the Earth’s surface at a variety of spatial and temporal scales has long been of interest to geospatial scientists. In this context, ANN has helped to identify patterns of change through time as well as to distinguish abrupt (deforestation or forest fires) and continuous (coastal erosion) changes. This application involves mostly supervised neural networks or hybrid approaches, where classification is done at two time periods and the differences are then classified and mapped.

The main objective of clustering is to group large amounts of data into meaningful categories
that can be disjoint, overlapping, or organized in some hierarchical fashion. A category or cluster is characterized by maximum similarity among its members and minimum similarity between its members and the ones belonging to other classes. Unsupervised ANNs are often used in this context, focusing on exploratory data analysis and visualization of clusters extracted from massive data with many dimensions.

Suppose an unknown function $\Phi(x)$ (subject to noise) has generated a set of $n$ input-output pairs $\{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$. The task of function approximation is to find an estimate of $\Phi^*$ of the unknown function $\Phi$. ANN can approximate functions with arbitrarily high degrees of nonlinearity with a sufficient number of nodes and layers. These advantages have led to the recent use of ANN estimation algorithms for geophysical parameter retrievals. Typically, supervised ANNs are used in this context to solve forward and inverse remote sensing problems.

The goal of optimization is to find a solution that satisfies a set of constraints such that an objective function is maximized or minimized. The most famous problem relevant in GIS is the Traveling Salesman Problem, an NP-complete problem where a traveling salesman must visit $N$ number of cities with the following constraints: start and end the trip at the same location, visit each city only once during the trip, and minimize distance traveled during a trip. The order of visits in the trip does not matter for finding the solution. The ANN used in this context is the Hopfield model. Related modern applications include vehicle routing (such as FedEx, UPS) where the goal is to minimize the number of vehicles involved in delivery, total travel time, and total delivery time.

Given a set of $n$ samples $\{(y(t_1),y(t_2),\ldots,y(t_n))\}$ in a time series, $t_1, t_2, \ldots, t_n$, the task is to predict the value $y(t_{n+1})$ at some future time $t_{n+1}$. Two important issues must be addressed in this application: the frequency with which data should be sampled, and the number of data points which should be used in the input representation. These issues are settled empirically in most application contexts, but results from work in complex dynamic systems have employed heuristics. Typical examples of this supervised ANN approach are market predictions (price of crude oil or gold), climatological (modeling ENSO (El Niño Southern Oscillation) events using large-scale climatological parameters), and network traffic forecasting (spatial interaction in terms of flows between origin and destination pairs).

Present technology trends and the future of ANN

During the last decade, there have been changes in both ANN hardware and software. The new processors used in real-time ANN consist of electronic components that can be connected by circuits that are designed to mimic biological synapses. They are known as “neuromorphic” processors since they are based on large groups of neuron-like elements. The development of “deep learning” algorithms that can be embedded in these neuromorphic chips has led to substantial growth in real-time applications. These developments will vastly impact geospatial sciences in the next decade.

SEE ALSO: Artificial neural networks; Data structure, raster; Data structure, vector; Geographic data mining; Quantitative methodologies

References


Artificial neural networks

A-Xing Zhu
University of Wisconsin, USA

Artificial neural networks ANN are motivated by the functioning of biological neural networks. An ANN solves a problem in a stepwise fashion by first developing a memory associating a large number of input patterns with a set of resulting outputs through training on examples, and then by applying this association to produce an output when presented with a novel input pattern. A detailed account of ANN types and their applications is beyond the scope of this entry. Readers are referred to the following review articles and books for detailed discussion on ANNs: Masters (1993) and Graupe (2013). A number of studies on the application of ANNs in geography also are available (e.g., Clair and Ehrman 1998; Maier and Dandy 1996; Pachepsky, Timlin, and Varallyay 1996; Zhu 2000; Pu et al. 2008; Blackwell and Chen 2009). The following subsections provide a background on the structure, training, operation, and application of multilayer feedforward networks, the type of ANN most often used in geographic applications.

Network structure

A multiplayer feedforward network is made of many processing elements (neurons). These neurons are usually arranged in layers: an input layer, an output layer, and one or more layers in between called hidden layers (Figure 1). The neurons in one layer are connected to the neurons in the next layer with different strengths of connection, which are referred to as weights. Links support information flow (signals) with different levels of strength, as controlled by the weight. The neurons in the input layer are used only to receive external inputs, but the neurons on the hidden and the output layers have information processing capabilities (Figure 2). The processing neurons first perform a weighted sum of the inputs from different neurons in the previous layer and transfer this sum into the range [0,1] through a transfer function of sigmoidal shape (see Figure 2).

The structure (the numbers of layers and the number of neurons in each layer) of a multiplayer feedforward network is problem specific. Masters (1993, 85–88) reports that, for most problems, a three-layer model should be sufficient unless the problem domain is highly discontinuous. The number of input neurons should be equal to the number of inputs and the number of output neurons should be equal to the number of attributes that define the output. For example, if one was to use an ANN to predict soil types from environmental variables (such as elevation, slope gradient, geology, temperature), then the inputs should be the values of these environmental variables and the outputs should be the soil types. The number of hidden neurons depends on the complexity of the problem (Masters 1993, 176–177); simple problems require fewer hidden neurons. The exact number of hidden
Network training

Network training (learning) is used to determine a set of weights that will produce the best possible input/output mapping. Most network training employs a supervised approach in which the network is presented with a set of input patterns and a set of corresponding desired outputs (together referred to as training data). The training process starts by initializing all weights to small nonzero values. Then, training samples are presented to the network one at a time to produce corresponding results. A measure of error between the network outputs and the desired outputs is computed and weights are updated to reduce error. Many iterations or epochs (from presenting training samples, to measuring error and updating weights) may be required before a network reaches an acceptable level of accuracy.

The magnitude of error is determined by the combination of weights for a given network. There exists an error surface that can be described as a function of weights. The objective of network training is to find a set of weights that will minimize the error function. A number of network learning algorithms can be used to minimize network error functions (Graupe 2013; Askarzadeh and Rezazadeh 2013; Nur, Radzi1, and Ibrahim 2014).

One common method, the conjugate gradient method, improves upon the conventional back propagation method by using adaptive approaches to the determination of learning rate ($\mu$) and momentum ($\eta$), which control the convergence and the speed of network training (Masters 1993, 105–111). The learning rate determines the rate at which the weights should be modified during each epoch. The momentum controls the direction to search for a minimum. The conjugate gradient method first employs the Polak-Ribiere algorithm to determine the best direction to search for the minimum. It then uses a directional minimizing process to determine the location of the minimum.

The error surface of weights for a neural network can contain a large number of local minima since a large number of weight permutations...
Figure 3  Operation of a three-layer feedforward neural network.

will produce similar input/output mappings. These local minima can make network training even more complicated. Avoiding false (local) minima consists of two steps. The first is to avoid initiating weights in the vicinity of minima and the second is to determine if a found minimum is local. If it is, attempts should be made to escape from this local minimum. Masters (1993) describes two techniques, simulated annealing and genetic algorithm, for initially avoiding and then escaping local minima. Interested readers are referred to this book for further reading on these topics.

Network operation

The operation of a feedforward neural network is rather simple. The inputs from the neurons in the input layer are respectively fed to each hidden neuron in the hidden layer through the weighted sum approach. Each hidden neuron then transfers the sum into a range $[0,1]$ as the output of this hidden neuron. This output is then used as the input to each of the output neurons on the output layer. The best way to show how an artificial neural network operates is to work through an example as shown in Figure 3.
This section focuses on the critical issues that must be considered when applying an ANN to a problem domain. The first issue is the determination of network structure, and in particular, the number of hidden neurons in the hidden layer. This is because the number of input neurons and the number of output neurons are determined by the problem, and the size of the network depends solely on the number of hidden neurons. Learning capacities increase with increasing numbers of hidden neurons. At the same time, however, larger networks require more training samples and they are much more readily over-fitted, at which time the trained network describes random error instead of the underlying relationship. On the other hand, if the size is too small (too few hidden neurons), the network is too simple and an under-fitting situation is created in which the trained network cannot model the underlying complicated relationships. Although an optimal network structure is difficult to achieve, methods can be employed to minimize the risk of over-fitting and under-fitting (Zhu 2000).

A second issue is based on the quality of the samples used for training and validation. High-quality samples are representative of the population. Typically, two sets of samples are used for network training: the training samples, which are used to develop the association between input patterns and the output patterns, and the validation samples, which are used to examine the accuracy of the trained network. A biased sample set, particularly biased training samples, would lead to a biased network. The sample sets should be sufficient in size, otherwise, the network becomes undertrained. Prescribing a number of samples sufficient for a particular case is a challenging task. Masters (1993) has suggested that, given that the sample set is representative of the population, the minimum number of training samples for a neural network with \( m \) inputs and \( n \) outputs should be \( 2^* (m+1)^* n \), and a reasonable number of training samples should be \( 4^* (m+1)^* n \). Bartlett (1998) suggested that for good generalization performance the number of training examples should grow at least linearly with the number of adjustable parameters in the network. Recently, improved methods for sample size determination have been explored (Nuchitprasittichai and Cremaschi 2013).

The third issue concerns the applicability and portability of trained neural networks. Trained neural networks, very much like regression models, are only applicable in the domain in which they are trained, and they cannot be generalized to other domains. For example, a soil-mapping neural network trained over an area with granite parent materials will not perform well over areas with limestone parent materials because the underlying relationships are different. In geographic applications, it is extremely rare to port a neural network trained in one area to another area unless the user is sure that the underlying relationships between the two areas are the same and the training samples used are also representative of the area the network is to be ported to.

A fourth issue is the interpretability of the relationships discovered by the trained network. Unlike regression models and decision trees, the relationship, or mapping, between the input patterns and the output patterns extracted by neural networks is in the form of a weight configuration. These weights have no physical meaning in the application domain. They are simply used to match the input patterns with the expected output patterns during training.
Summary

Artificial neural networks first develop a mapping between input patterns and output patterns through training. They then apply the mapping to predict an output pattern for a given input pattern. Neural networks are capable of learning nonlinear relationships between a set of input variables and a set of output variables. However, the learned relationship is in the form of a network weight configuration that is not meaningful in the application domain. Neural network training often requires a large number of samples and any learned relationship is not transferable to areas (domains) in which the training samples are not representative. Given these properties, ANN is effective for applications in which the provision of samples is not expensive and the problem domain is fixed.

Despite these advantages, applications of ANN in geography have limitations. First, the provision of a large set of representative samples is often expensive. Second, the heterogeneity of geographic phenomena (over space and time) makes the application domain not as stable as one would hope. Third, a major goal for many geographic analyses is to understand the relationships between the inputs and the outputs. Neural networks cannot provide those relationships in a meaningful form.

SEE ALSO: Big data; Cloud computing; Fuzzy classification and reasoning; Geographic data mining; Geographic information science; Geographic information system; Machine learning; Parallel computing

References

ARTIFICIAL NEURAL NETWORKS

Further reading


Atmospheric aerosols

Shiyuan Zhong
Michigan State University, USA

Rahul Zaveri
Pacific Northwest National Laboratory, USA

Definition

Aerosols are suspensions of tiny solid and liquid particles in a gas (Hinds 1999). Atmospheric aerosols generally refer to the mixture of particles and the suspending gas, which is air. The particulate portion of an atmospheric aerosol is referred to as particulate matter or PM, but in common practice the name aerosol is used to refer to particles only and often the terms “aerosol” and “particulate matter” are used interchangeably. Examples of common atmospheric aerosols are dust, smoke, and sea spray. Aerosols vary in shape, size, and chemical composition, and they have both natural and anthropogenic origins. Despite their tiny size, atmospheric aerosols have a profound impact on climate, visibility, biogeochemical cycling, and human health.

Size distribution

Atmospheric aerosols span a wide range of sizes from a few nanometers (nm) in diameter, the size of the smallest viruses, to around 100 μm (micrometers) in diameter, about the width of a human hair (Willeke and Whitby 1975; McMurry et al. 2000). Aerosol particles are usually considered as belonging to three size modes (Seinfeld and Pandis 2006). Smaller particles with diameters less than approximately 0.1 μm are known as Aitken mode particles; however, some describe the smallest group of particles, of 0.01 μm or below, as nucleation mode. These particles are formed by a nucleation process in which gaseous molecules of low volatility condense to form solid, semisolid, or liquid matter (Hinds 1999). Aitken mode particles in the atmosphere quickly grow through processes of coagulation or gas-to-particle conversion to form new aerosol particles that reside in what is known as the accumulation mode of approximately 0.1–1 μm in diameter. Particles greater than approximately 1 μm in diameter are referred to as coarse mode and they are typically generated through physical or mechanical processes such as erosion of the Earth’s surface or sea spray, rather than through the nucleation, condensation, and other chemical processes. Collectively, all aerosol particles with diameters smaller than 2.5 μm (PM2.5), which consist of Aitken mode and accumulation mode particles, are known as fine aerosols or fine particles. Particles with diameters between 2.5 and 10 μm are correspondingly known as coarse particles.

Sources, sinks, and composition

Aerosols may be either natural or anthropogenic in origin (Seinfeld and Pandis 2006; Figure 1). Aerosols that are directly emitted into the atmosphere are called primary aerosols while those that are formed via gas-to-particle conversion processes in the atmosphere are referred to
as secondary aerosols. Nearly 90% of the aerosol mass, which is dominated by coarse aerosols, comes from natural sources. In contrast, the aerosol number concentration is dominated by fine particles. The largest natural source for aerosols is soil dust, generated as strong winds whip small pieces of mineral dust from deserts or other arid areas on Earth into the atmosphere. Because two-thirds of the Earth’s surfaces are covered by oceans, sea salt is the second largest aerosol contributor. Sea-salt particles can enter into the atmosphere by bubble bursting through the sea surface microlayer or wave crashing and they consist mainly of sodium chloride and small amounts of other salts and dissolved organic matter present in seawater. Both sea salt and dust, the two most abundant aerosols in the atmosphere, consist of mainly coarse particles, although they are also present in fine mode. Volcanic eruption, which injects huge columns of ash, sulfur dioxide, and other gases into the atmosphere, is also an important natural aerosol source. Major volcanic eruptions may inject millions of tons of sulfur dioxide high up into the stratosphere, yielding sulfate aerosols upon oxidation. Once there, these aerosols may stay in the stratosphere for about two years and they partially reflect sunlight back to space, thereby reducing the amount of solar energy reaching the lower atmosphere and the Earth’s
Aerosols can also be produced naturally through biological processes – pollens, spores, and plant waxes are examples of biological aerosols (Despres et al. 2012). The fine aerosols in the atmosphere originate from both natural and manmade sources. Primary smoke particles from tropical deforestation constitute a large fraction of manmade aerosols in the atmosphere. Biomass burning, a common practice in agriculture and land management, also introduces primary smoke aerosols into the atmosphere. Smoke aerosols are dominated by black carbon and heavy organic materials. Automobiles, airplanes, ships, smelters, and power plants that burn oil, coal, and compressed natural gas are major sources of pollution aerosols in the atmosphere and are composed of sulfate, nitrate, ammonium, organics, and soot or black carbon in varying amounts (Bond et al. 2007). Sulfate and nitrate aerosols are secondary aerosols that are typically formed from oxidation of sulfur dioxide and nitrogen oxide gases to sulfuric and nitric acids, respectively, which condense on pre-existing aerosols and react with gaseous ammonia (largely emitted from agriculture and livestock operations) to form their corresponding ammonium salts in the particle phase (George and Abbatt 2010). A significant fraction of fine aerosols is composed of secondary organic aerosols formed from oxidation of volatile organic gases that are emitted from natural sources such as vegetation and forests as well as vehicular exhaust (Andreae and Crutzen 1997). Sulfuric acid can also nucleate to form new particles, some of which may be lost due to coagulation with larger particles, while the rest may grow to larger sizes via condensation of various secondary trace gases. Oceanic phytoplankton emit trace gases such as dimethyl sulfide, which upon oxidation can lead to the formation of secondary sulfate and methanesulfonic acid aerosols in the marine atmosphere. Secondary aerosol compounds can also be produced via chemical reactions inside clouds, followed by evaporation of the cloud droplets.

Aerosols in the atmosphere have lifetimes ranging from a few minutes to several days depending on their sizes and atmospheric conditions (Seinfeld and Pandis 2006). Newly formed nanoparticles and Aitken mode aerosols have a relatively transient existence in the atmosphere, with lifetimes ranging from minutes to hours, because they transform rapidly into coarser particles via coagulation and condensation processes. The accumulation mode particles may remain in the atmosphere for several days because they are too small to settle out rapidly from the air under gravity and too large to be subject to rapid Brownian motion; the slow random motion reduces coagulation rate, thereby inhibiting further growth and removal. Because of their long lifetime, these particles can be transported by winds to thousands of kilometers downwind from their origin. In contrast, the heavier coarse mode aerosols tend to settle out relatively rapidly from the atmosphere, with lifetimes ranging from hours to days. Aerosols injected into the upper atmosphere (above the boundary layer) can remain suspended in the air for several weeks to months. Eventually, however, all aerosols are removed from the atmosphere by either settling down under gravity, referred to as dry deposition, or through scavenging by cloud droplets and subsequent rain out, referred to as “wet deposition.”

Impact

Atmospheric aerosols have significant impact on people and environment. These tiny particles
ATMOSPHERIC AEROSOLS

play an important role in the Earth’s climate system, affect visibility, provide sites for chemical reactions to take place in the atmosphere, and, when in high concentrations as found in urban environments, are a respiratory health hazard.

Aerosols have a direct effect on the Earth’s climate by altering the planetary radiation balance through scattering, reflection, and absorption of solar radiation. The sulfate aerosols ejected to stratosphere by major volcanic eruption reflect sunlight and thereby reduce solar energy reaching the lower atmosphere and the surface of the Earth (Charlson et al. 1992). Aerosols can also affect climate indirectly by influencing cloud formation and lifetime. Aerosol particles may serve as condensation nuclei, causing cloud droplets to increase in number but decrease in size, and thus becoming more reflective to solar radiation (Hess, Koepke, and Schult 1998). There is also a so-called semi-direct effect of aerosols in which some aerosols, particularly black carbon, absorb solar radiation and heat the atmosphere, resulting in a decrease of cloud cover. Black carbon aerosols, when deposited on snow, ice, and other bright surfaces, reduce surface reflectivity and produce a warming effect. The combined direct, indirect, and semi-direct climate effects of aerosols are estimated to be comparable in magnitude, but opposite in direction to those of increasing concentration of greenhouse gases in the atmosphere (Boucher et al. 2013). However, large uncertainties still exist in the estimates of the indirect radiative effects of atmospheric aerosols.

The scattering properties of aerosol particles may cause visibility degradation. Particles in the accumulation mode typically reside in the size spectrum similar to the visible range of the solar spectrum (300–400 nm) and according to Mie’s scattering theory, these particles can effectively scatter sunlight, preventing light from reaching objects in the distance and thereby impairing visibility.

Aerosols, especially those with urban origins, influence atmospheric chemical processes by providing a surface for heterogeneous chemical reactions to occur (Andreae and Crutzen 1997). The most well-known reactions are those related to stratospheric ozone destruction. Aerosols injected into the stratosphere over the polar region help to form polar stratospheric clouds and the large cloud droplets provide surfaces for chemical reactions that lead to the formation of large amount of reactive chlorine responsible for ozone destruction (Portmann et al. 1996).

High aerosol concentrations are often found in major cities and areas surrounding and downwind of the cities. The primary concern of high aerosol concentrations in these urban environments is their impacts on human health (Schindler et al. 2009). Studies have linked short-term increases in particle pollution to increased emergency room visits or hospitalization for patients suffering from respiratory diseases and long-term exposure to increased risk of death associated with lung cancer and cardiovascular diseases (Samoli et al. 2008; Stanek et al. 2011). Fine particles are found to be more harmful to human health than large particles because fine particles have longer residence time in the atmosphere and are able to penetrate into the deepest part of lungs (Downs et al. 2007).

SEE ALSO: Climatology; Precipitation; Urban climatology

References

ATMOSPHERIC AEROSOLS


Further reading


Atmospheric/general circulation

John A. Knox and Pamela N. Knox
University of Georgia, USA

The general circulation of the atmosphere (also known as the planetary circulation) is, in its most basic sense, the large-scale flow of the atmosphere across the entire planet. It is described by averages of horizontal and vertical motion of air over time and space as well as temporal and spatial deviations from the average conditions. These deviations include both large-scale waves linked to land–ocean differences and orography, and also seasonal and multiseasonal variations around the mean conditions.

The primary driver of the general circulation of the atmosphere is the differential heating of the Earth by the sun. Figure 1 depicts how solar radiation near the equator is direct and nearly perpendicular to the Earth's surface, strongly heating the tropics by direct radiation. By comparison, the same amount of incoming solar radiation is spread across a much larger area near the poles where the radiation strikes the Earth more obliquely. The solar intensity and length of day varies with season as the Earth's pole tilts towards or away from the sun.

The three-cell model

The classical description of the general circulation of the atmosphere breaks the atmosphere down into a three-celled circulation in each hemisphere (see Figure 2). The cell closest to the equator is called the Hadley cell after George Hadley, an English lawyer and amateur meteorologist, who in 1735 provided one of the earliest descriptions linking the surface trade winds, which blow generally from east to west between the equator and about 30° north and south, and a vertical convection cell driven by strong solar heating near the equator (Persson 2006). In the classical Hadley cell, warm rising air at the equator spreads out at high levels in the atmosphere and sinks at approximately 30°, returning to the equator as a surface return flow which flows from east to west due to the effects of the rotating Earth. The band of rising vertical convection near the equator, usually associated with deep clouds and heavy rainfall, is known as the Intertropical Convergence Zone (ITCZ). The band of clear skies and high pressure associated with sinking air at 30° is known colloquially as the “horse latitudes” because of the calm conditions which sometimes necessitated sailors throwing their steeds overboard to lighten their ships when caught in the calm conditions. Subtropical lands beneath the sinking portion of the Hadley cell are often arid deserts due to the lack of clouds and heating from the sinking air. Since this circulation is caused by the direct heating of the sun, it is known as a “direct circulation.”

Another major cell in the classical circulation model is the polar cell. It is also a direct circulation caused by sinking air over the frigid poles, which spreads out at the surface in the form of polar easterlies and rises at mid-latitudes, returning to the poles at high altitudes.
The third cell in the classical circulation model is known as the Ferrel cell, proposed by William Ferrel (1856). The Ferrel cell is an indirect circulation (warmer air sinking, colder air rising) which fills the gap between the Hadley cell in the tropics and the polar cell in the polar regions. In the Ferrel cell, air sinks at roughly $30^\circ$ in concert with the sinking branch of the Hadley cell, flows at the surface poleward and from west to east (the “mid-latitude westerlies”), rises at the border with the polar cell, and joins the upward flow of the polar cell in the vicinity of $45–60^\circ$ latitude, returning air aloft from pole towards the equator.

Jet streams, or regions of strong wind blowing generally from west to east, occur above the boundaries between the three cells of the classical circulation model. The subtropical jet marks the border between the Hadley cell and Ferrel cell and often serves as a conduit for tropical moisture to enter the mid-latitudes. According to classical theory, the polar jet stream is located roughly between the Ferrel cell and the polar cell, although in reality it is usually located above the area of maximum surface temperature gradient and marks the transition from warm tropical air to cold polar air. The ring of strong upper-level winds around the polar cell is sometimes called the “polar vortex,” an area which delineates the area of coldest surface air near the pole.

Because of the tilt of the Earth’s rotation axis relative to the sun and the heating changes associated with sun angle, the cells move north and south with the seasons, leading to temporal changes in the position of the ITCZ and the subtropical high-pressure regions over the year. These seasonal changes contribute to changing climate regimes, particularly at locations near the cell borders. For example, in a Köppen climate classified as Aw (tropical wet and dry), the ITCZ will lie over the region for part of the year, bringing rain and cloudiness to the area for a time. When the sun angle shifts due to the different orientation of the Earth’s axis later in the year, the ITCZ moves away, leading to a complementary dry season.
Advances to the three-cell model

While many textbooks still leave the impression that Hadley’s initial 1735 work is correct, this is not the case. For example, as Persson (2006) notes, Hadley’s theory explained only the direction, not the magnitude, of the trade winds and is flawed as an explanation for the general circulation because of the decided lack of conservation of angular momentum, among other reasons.

Furthermore, air parcels themselves do not obey the three-cell model. The Ferrel cell arises from calculations done with respect to a fixed location (Eulerian), but a single-cell equator-to-pole circulation is found using flow-following (Lagrangian) calculations and/or calculations using vertical coordinates that follow the motion of air parcels more closely than altitude, such as potential temperature. The best way in which to understand the Ferrel cell, then, is that it is a circulation driven by mid-latitude eddies (mid-latitude cyclones and the even larger planetary waves) in which the forcing on the atmosphere by these eddies (known as the Stokes drift) causes cooler air on its northern flank to rise and warmer air on its southern flank to sink (Holton 2004, 324). But, again, the Ferrel cell’s existence is reference-frame and coordinate dependent.

The most notable, relatively recent advance upon the early theory of the general circulation is that of Held and Hou (1980), who – in contrast to Hadley – studied an angular momentum-conserving frictionless model of the atmosphere. Their results for this idealized model connect the poleward extent of the Hadley cell with both the equator-to-pole temperature gradient and the planetary rotation rate. Held and Hou’s results are consistent with observations of both Earth’s and other planets’ atmospheres, and continue to be highly influential on research on the atmospheric general circulation of Earth and other planets.

Zonally asymmetric circulations and monsoons

The classical understanding of the equator-to-pole general circulation is also inadequate for understanding Earth because land and oceans are not equally distributed across the planet. Oceans, because of their high heat capacity, warm and cool slowly under solar radiation compared to the higher temperature variability of land surfaces. This difference in surface heating causes zonal asymmetries in the flow of air around the Earth and allows enhanced transfer of energy from the equator to the poles, warming polar regions and cooling the tropics by heat transport in poleward ocean currents like the Gulf Stream at lower latitudes, and by air currents at higher latitudes (Seager et al. 2006).

Because of the presence of continents that provide differential heating, the surface circulation patterns are broken up into discrete cells related to land–ocean position. This is most noticeable in the descending branch of the Hadley circulation and is seen in the persistent centers of subtropical high pressure which are observed over the oceans and land surfaces. The oceanic circulation patterns, including the gyres of surface circulation, are related to the winds which occur in the presence of these high-pressure cells.

One of the most notable current large-scale circulation patterns caused by land–ocean differences in heating is the monsoon. The monsoon is a seasonally reversing wind circulation caused by regional differences in temperature between land and water. Monsoons occur where there is a large land mass in the mid-latitudes next to a tropical ocean. In summer, the land mass warms up relative to the ocean and creates rising air,
Interannual variability and oscillations

Over time, climatologists have observed that the monsoons are not identical each year, and that in some years the monsoon has failed to appear, bringing drought and famine to areas that are expecting the monsoon rains to replenish their soils. This year-to-year variability is caused by a number of factors, including aerosol loading as well as changes in the ocean surface temperatures over time, that shift the patterns of high and low pressure which drive the global winds and their associated weather systems.

The most well-known of these patterns of interannual variability is the El Niño Southern Oscillation (ENSO), which is related to a quasi-periodic shift of atmospheric pressure which is centered between the eastern and western Pacific Ocean but which extends its influence across the globe. In one phase of ENSO, El Niño, a reduction in the speed of the trade winds allows warm water to pile up in the eastern Pacific Ocean, causing an increase in thunderstorm activity above it which pushes higher-level winds away from their neutral position. In the opposite phase, La Niña, the trade winds are stronger than usual, pushing the pool of warm surface ocean water to the central and western Pacific, shifting the area of convection with it and shifting the upper-level winds in a different direction. This oscillation of winds, which in the Pacific is called the Walker circulation, is related through long-distance interactions (“teleconnections”; Figure 4a, b) to observed changes in weather patterns across many parts of the Earth, including the strength and timing of the Indian monsoon in summer (Philander, Holton, and Dmowska 1989).

Another zonally asymmetric feature of the tropical atmosphere is the Madden–Julian Oscillation, or MJO (Madden and Julian 1972). It is also known as the 30–60-day oscillation,
and is triggered by tropical thunderstorms. From there, a slowly propagating pattern moves east at about 4–8 m s⁻¹ from the Indian Ocean to the Pacific. The MJO interacts with monsoons and tropical cyclones, and has received increasing attention in recent years for its possible downstream effects on North America.

Tropical convection also triggers a pattern of interannual variability in the stratosphere, known as the quasi-biennial oscillation (QBO;
Baldwin et al. 2001). The QBO consists of alternating descending westerly and easterly winds in the lower tropical stratosphere with a period of approximately 27 months. These wind patterns are caused by the breaking or absorption of small-scale gravity waves that are caused by the thunderstorms near the surface. The QBO can indirectly impact tropospheric weather by modulating planetary wave propagation and breaking in the stratosphere, which can then affect the troposphere.

In addition to the modes of tropical variability described above, there are also a number of oscillations which are centered on the extra-tropical region (see State Climate Office of North Carolina (2016) for explanations and current conditions for many of them). These extra-tropical oscillations include the Arctic Oscillation (AO), the North Atlantic Oscillation (NAO), the Atlantic Multidecadal Oscillation (AMO), and the Pacific Decadal Oscillation (PDO).

The AO, also referred to as the Northern Hemisphere Annular Mode, is defined by winds in the lower troposphere blowing counterclockwise around the Arctic at the approximate latitude of Edinburgh, Scotland. In the positive phase of the AO, these winds are fast and trap cold air near the pole. The negative phase of the AO is characterized by weaker westerlies, a more undulating polar jet stream, and incursions of cold air and storminess in the mid-latitudes. These phases alternate interannually, but can also change within a single winter. The NAO captures the same phenomenon as the AO, but is calculated using differences in sea level pressure between Iceland and the Gibraltar/Azores region (see Figure 5).

The AMO, discovered in 1994, captures an oscillation of a little less than a century in ocean temperatures in the North Atlantic. It may have an effect on tropical cyclone frequency and intensity in this basin. The PDO, a similar sea
surface temperature oscillation but at somewhat higher frequency in the North Pacific, has been more widely researched. Its phases, similar to ENSO, are statistically correlated with temperature and precipitation anomalies, as well as impacts on fisheries, across the Pacific basin and also across adjacent continents.

**Climate change and the general circulation**

On the longest timescales, the cumulative effects of changes in orbital parameters or large-scale changes in the location of land masses and mountain ranges due to movements caused by plate tectonics have caused significant changes in the pattern of the general circulation on Earth. For about the last 2 million years, with the continents in close to their modern configuration, small changes in sunlight due to oscillations in the eccentricity, tilt, and precession of the Earth's rotational axis have changed the yearly distribution of sunlight across the Earth, leading to the rise and fall of ice ages as snow accumulated or evaporated at higher latitudes on the continents. Collectively, these changes are known as Milankovitch cycles (Imbrie and Imbrie 1986). On longer timescales, the changing position of the continents or the opening or closing of oceanic passages changed the flow of heat through the ocean or the vertical convection cells caused by land–sea temperature differences, further altering the Earth's general circulation.

The advent of anthropogenic climate change has generated much interest recently with regard to its effects on the general circulation. Some researchers (e.g., Francis and Vavrus 2012) have hypothesized that fundamental features of the general circulation could be altered by a weakened north-south temperature gradient – which is the result of more pronounced warming at high latitudes, for example the Arctic regions, than at mid-latitudes. For example, a weakened polar jet stream could lead to more persistent and potentially more extreme weather patterns in the mid-latitudes. These “Arctic amplification” linkages to the general circulation are the subject of active research.

While much of the theory of general circulation in the atmosphere is well established, there is still much ongoing research. Please see Further reading for summaries of unsolved questions at both introductory and advanced levels.

**SEE ALSO:** Climatic modes and teleconnections; Climatology; Climatology: history; Glaciations; Monsoons; Oceanic circulation

**References**


Further reading


Augmented reality

Nick Hedley
Simon Fraser University, Canada

Augmented reality (AR) is a real-world environment that is augmented with synthetic (virtual) objects or information. AR is a specific subtype of mixed reality (MR). Mixed reality is a continuum of possible combinations of reality and “virtuality,” anchored by the extremes of pure virtuality and pure reality at each end of this spectrum.

Context: the power and shortcomings of virtual environments

Virtual environments (VE) are synthetic, spatial (usually three-dimensional (3-D)) environments constructed from 3-D modeling, spatial data (not just 3-D data), and other digital content. They are typically seen from first-person perspectives, using one or more types of display (desktop LED/LCD screens, head-mounted displays (HMDs) such as Oculus’s Rift). Users navigate and control virtual environments using various transducers and input/output devices. Combinations of displays and control devices allow a range of sensory information experiences to occur. Examples of VEs include HMD-based immersive CAVE (cave automatic virtual environment) environments supporting single or multiple users, or semi-immersive desk environments (with a “driver” and multiple viewers). Interaction modalities and interaction design together determine the way VEs can be controlled and explored, and strongly influence the “information experiences” that result.

Throughout the 1990s, VEs were explored as tools for training, communication, collaboration, and education, delivering unique and powerful experiences and abilities (i.e., fly anywhere, move under water, through outer space, inside molecular structures into the future, to manipulate phenomena and time, and to see abstract or normally invisible phenomena). Research demonstrated that VEs may allow users to form useful mental representations of large real-world spaces that they can lead to distorted spatial perception, and that previous user knowledge and experience and interface interactivity were perhaps more significant predictors of performance than the “immersiveness” of VEs.

As VEs and their enabling technology became more accessible during this period, geographers became interested in the potential of their capabilities to deliver new forms of geovisual analysis. The geovisualization community, in particular, explored geovisual VEs and methods (see Dykes, MacEachren, and Kraak 2005), and engaged the concepts of immersion, interactivity, information intensity, and intelligence of objects (MacEachren et al. 1999; Slocum et al. 2001).

One of the lessons learned from the 1960s–1990s era of VE research, however, was that in many respects VEs were too much of an abstraction from reality to be a realistic part of it. Immersive VEs often delivered dramatic and engaging visual experiences, but they suffered from limitations of expense, accessibility, and
AUGMENTED REALITY

Mixed reality: the foundation of augmented reality

In order to understand augmented reality and its significance for geographic research, one must understand the relationship between reality, virtual environments, and mixed reality. Augmented reality (AR) is a subclass of mixed reality (MR). Milgram and Kishino (1994) define MR as “technologies that involve the merging of real and virtual worlds.” Their virtuality–reality continuum construct is well known in the human–computer interaction (HCI) community, placing all interface types on a spectrum between completely real (reality) to completely virtual (virtuality) (see Figure 1).

Between these two extremes lies a middle ground of mixed reality interface technologies and environments that combine elements of reality and virtuality. Within this middle ground, MR interface environments differ by proportion of real and virtual content used. This leads to two distinct subcategories of MR: augmented virtuality (AV); and augmented reality (AR). Augmented virtuality environments are those that are predominantly virtual, enhanced by information from the real world. By contrast, augmented reality (AR) interface environments are predominantly real interface environments, augmented by virtual digital information (see Azuma 1997 and Tamura, Yamamoto, and Katayama 2001).

Defining augmented reality

AR interfaces combine real environments with virtual objects—allowing access to both real and virtual environments simultaneously. This is achieved by superimposing digital virtual objects on views of the real world (Milgram and Kishino 1994), made possible by tracking and registration (alignment) of virtual objects with real objects or geographic spaces, and display technologies that allow the user to see virtual objects in geographic space.

Azuma (1997) defined AR interfaces as those that: superimpose virtual information on the real world (i.e., they combine virtual and physical objects in the same interaction space); are interactive in real time; and are spatial—meaning the virtual objects are registered and interactive in 3-D space. As a broader range of AR techniques have emerged in the early twenty-first century, many AR researchers have adopted a broader mixed reality world view, where augmentation does not require 3-D virtual objects (Azuma’s specification), but only that virtual objects are registered in 3-D physical space in relation...
to other physical objects. This stance helps to accommodate a multiplicity of mixed reality and AR interface permutations, but also helps reinforce the difference between user interfaces (UI) that augment the 3-D space of reality versus the “annotated vision” provided by heads-up display (HUD)-type UIs.

**Types of AR**

There are various permutations of AR. Some are considered “annotated vision” interfaces, in which real views are enhanced through real-time annotation on the display surface, rather than registered to the 3-D space being viewed. Other more advanced AR systems include novel applications where users “fly into” immersive virtual worlds and participants can leave their real surroundings behind, and join others to collaborate in shared virtual spaces (see Billinghurst, Kato, and Poupyrev 2001). Two popular subtypes of AR are tangible AR (TAR) and mobile AR (MAR).

Tangible AR interfaces enable tangible user interface (TUI) control and manipulation of virtual digital content. Examples include attaching virtual 3-D objects to physical cards (see Figures 2 and 3), physical pages of books (such as the well-known Magic Book AR project), and turntables (see Figure 4).

The resulting effect is simple yet dramatic. As the user moves the card, the virtual object stays anchored to the card, and moves as if attached to it. This effect is made more powerful due to direct physical touch and control. What were previously mouse or button-actuated and metaphor-mediated activities, such as zoom, pan, and rotate in WIMP (windows, icons, menus, pointer) interfaces, are now achieved by moving the augmented objects in one’s hands — no different than if one were inspecting a coffee cup (to inspect, rotate, hold closer or further away, zoom, or pan). One of the most sophisticated examples of tangible AR in the early 2000s was the exhibit (designed by the author) for Boston Museum of Science/Lucasfilm’s *Star Wars: Where Science Meets Imagination* (see Figure 4).

![Figure 2](image-url) Examples of tangible AR, where 3-D virtual objects are attached to fiducial markers printed on physical objects (cards). Left image shows abstract 3-D object used for cognitive AR experiments. Right image from collaborative geovisualization research by Shelton and Hedley. Hedley *et al.* 2001, with permission of the author (left); Shelton and Hedley 2004, reproduced by permission of Old City Publishing (right).
Inspection of virtual 3-D content using tangible AR interfaces is very similar to everyday manipulation of real objects, and is therefore a quite intuitive interface requiring little specialized training. A further trait of tangible AR interfaces is their strong proprioceptive cues and feedback – something Hedley (2001), Shelton and Hedley (2004), Woolard et al. (2003), and others have studied empirically, in the context of geovisualization, and geographic learning.

Mobile augmented reality (MAR) is a rapidly maturing interface technology that has the potential to extend our ability to combine virtual and real worlds (Azuma et al. 2001). Notable examples include Columbia University’s Touring Machine (Feiner et al. 1997), the Naval Research Laboratory (NRL) Battlefield Augmented Reality System (Julier et al. 2000), wearable MAR for tourists visiting Greek archaeological sites (Vlahakis et al. 2002), and
real-time 3-D augmented gaming in real-world spaces (Piekarski and Thomas 2002).

**How AR works**

AR interfaces are made possible by three main ingredients: tracking, registration, and display technologies.

Tracking is important for 3-D applications involving user interaction with 3-D virtual spaces. Tracking is used to provide information about object and user location and orientation in real 3-D space (including the position/orientation of hands/head/limbs and various types of control/feedback devices) to accurately correspond to their position in virtual spaces. Accurate, low-latency tracking is critical to delivering high-performance virtual environment experiences, where the position, geometry, and actions of users and objects in the real world have one-to-one correspondences with their equivalents in virtual 3-D spaces. Different types of tracking are used to support a wide range of 3-D applications and their needs. Common forms of tracking include magnetic tracking, motion capture (“MoCap”), and eye tracking. In AR, tracking is used to determine positions of objects in real-world space, so that digital virtual objects can be registered to them. This can be achieved in several ways. It can be done using fiducial markers and computer vision software – where unique patterns can be recognized by algorithmic analysis of video feeds. When a marker’s unique pattern is recognized, its orientation and position relative to the user’s (camera’s) viewpoint can be calculated. This information allows the AR software to place and render virtual digital objects at the marker’s location and alignment. This is a common tracking method for tangible augmented reality, where black-and-white markers are printed on cards to enable 3-D augmentation of the cards that can be physically touched, held, rotated, and inspected (see Hedley et al. 2002; Shelton and Hedley 2002; 2004).

This form of tracking is vulnerable to variable light conditions. Low light conditions can make it very difficult for AR software to detect visual markers, and high brightness conditions can “wash out” video feeds, thereby undermining an ability to track objects in space and provide registration needed for rendering 3-D objects in views of the real world. Advances beyond single unique marker tracking include “natural feature tracking,” where software can be trained to recognize unique objects (such as pages of a magazine) or geographic locations (coordinates), based on unique visual patterns. A challenge in natural feature tracking is the persistence of those unique visual patterns over time. If they change, the software’s ability to recognize them as a basis for spatial/locational tracking is undermined.

A unique adaptation of tangible and regular AR interfaces uses “augmented mirrors” to achieve hands-free tangible AR without the need for an HMD. In this configuration, large displays are adapted to deliver digital AR reflections of the people and scenes in front of them. This method has been used to study AR geographic pedagogy and media production (Woolard et al. 2003). One of the most developed examples of this technique is the tangible AR interface (designed by the author) for Boston Museum of Science/Lucasfilm’s Star Wars: Where Science Meets Imagination exhibit (Figure 5).

In geographic spaces, tracking for registration of AR objects increasingly uses global positioning system (GPS) information. Hedley’s (2008) and Mower’s (2009) early geographic mobile AR work explored how to combine GPS tracking with tablets for geographic AR, before spatially enabled tablets (such as iPads) became widely available. Nowadays, many (if not most) mobile
devices (smartphones, tablets, cameras) are location aware, using embedded GPS sensors and accelerometers, enabling tracking of position and orientation in three dimensions, without having to have a device’s camera “see” an entire visual marker. Location-aware devices, fused with widespread spatial (meta)data, have led to the arrival of numerous AR “apps.” Yelp and other social media applications have experimented with AR. However, more research is needed to improve the design (and rationale) for social media augmentation of real views.

The final ingredient necessary to view real space augmented with virtual objects registered and rendered to features tracked in geographic space are displays. Most tangible AR interfaces use HMDs in order to leave users’ hands free for physical manipulation of virtual content attached to physical objects (such as cards). A number of early mobile AR applications used wearable computers (computing hardware is carried in a backpack or belt (see Touring Machine, Feiner et al. 1997; see TinMith, Piekarski and Thomas 2002), aiming to leave users’ hands free for gestural interaction. Recent advances in quality, affordability, and accessibility of HMDs such as the Oculus Rift and Samsung’s Gear VR (Figure 6) have stimulated renewed interest in HMD devices. Whereas previous technologies were highly specialized, expensive, and rare, a significant difference now is that these displays are being designed to operate seamlessly with open source game engine platforms, such as Unity3D. This fact (plus trends in Kickstarter funded technology innovation) is not just relevant to new display technology; it also defines a new model of societal use and interoperability between mainstream and homegrown VR and AR.

A more recent development has been the emergence of systems which, rather than engineering dedicated niche HMDs, are designed to transform smartphones into VR displays. Samsung’s Gear VR system transforms smartphones into location-aware, orientation-aware HMDs (Figure 6). Another notable example is Google Cardboard, which transforms a smartphone into a basic VR headset with an origami-like cardboard enclosure and a software development kit (Figure 5).

With advances in camera, GPS, accelerometer, and display technologies embedded in mobile devices, an emerging vector of AR research has been the use of tablets and smartphones as AR displays. Jun Rekimoto (1997) was one
of the first to suggest it. Instead of wearing a cumbersome/expensive HMD, the user can hold up a smartphone/tablet as if taking a photograph and is able to see an augmented view of the world through the “window” of a mobile device running AR applications. This has been taken forward considerably by researchers in HCI and spatial interface research (see section on research progress, below).

Research trends and progress

While AR can be traced back to Sutherland’s groundbreaking work using half-silvered optics in the 1960s, a key pulse of AR research and development took place in the HCI and computer science communities (see Azuma 1997; Azuma et al. 2001). Much of the early work in AR concentrated on improving AR registration and tracking (Feiner et al. 1993). AR developed rapidly during the 1990s. In the late 1990s in particular, advances in displays (such as the commercial availability of new HMDs such as the Sony Glasstron™), computer vision, and tracking solutions led to more reliable platforms. Not unlike the maturation of GIS tools, this stability provided a basis on which to try more sophisticated applied research.

MAR has been implemented for various spatial applications so far. Columbia University’s Touring Machine (Feiner et al. 1997) was a self-contained backpack-wearable computing system that included differential GPS, compass, inclinometer, mobile computer, and see-through HMD. The user sees world-stabilized information (labels attached to locations of buildings) about an urban environment. This system might constitute “augmented vision” in that it annotates views of the real world, as opposed to placing 3-D virtual objects in the view of the real world.

Some might consider Dutton’s 1978 American Graph Fleeting (a cylindrical device designed to display “holographic” 3-D maps) an early experiment in 3-D geographic visualization interfaces. Unlike tangible AR, however, there was little to no direct free-form user interaction possible. In the geographic community there was some work

Figure 6  Examples of HMD. The Oculus Rift being used to view 3-D geovisualization (left). Samsung Gear VR (right). Image © Nick Hedley/SIRL 2015 (left); Samsung, public domain (right).
AUGMENTED REALITY

done to implement “annotated vision” interfaces in the early 2000s.

Over time, MAR researchers have found that some applications favor the hands-free interaction that HMDs provide (such as Piekarski and Thomas’s TinMith research (2002)), while others have started to explore applications that may favor embedded MAR capabilities in handheld computing devices, such as tablet PCs.

Exploratory work by Radburn (2006), Priestnall and Polmear (2006; 2007), Hedley (2008), and Mower (2009) investigated how we might use handheld devices to enable mobile augmented geographic visualization and interaction in everyday spaces.

The enabling technologies and research of the 1990s and early AR research in the 2000s led to new devices and applications that integrate location awareness, video cameras, and sufficient computational power to run the tracking, computer vision, and rendering elements of many AR software libraries. This in turn led to the emergence of publicly available AR authoring software and tools, such as ARToolkit, Junaio, and Metaio.

During this transitional period, there has been a growing interest in spatial uses of AR. While it is still true that many applications of AR simply use the novelty of seeing virtual 3-D objects attached to cards and other real-world objects, there has been a steady increase in geographic uses of AR, including empirical assessment of tangible AR for geovisualization (Hedley 2001), collaborative tangible AR for geovisual interfaces (Hedley et al. 2001; 2002), dynamic tangible AR map spaces (Cheok et al. 2002), cartographic information (Schmalstieg and Reitmayr 2007), building information overlays (such as Kamat and El-Tawil’s 2007 earthquake building damage assessment tool), underground infrastructure viewing using AR (Schall et al. 2009), and broadly geographic tasks (Dünser et al. 2012).

The recent interface research of Veas et al. (2013) implemented static spatial information overlays. However, while technologically advanced, the spatial information (contour maps overlaid on real views) was not correct (contour lines draped irregularly on landscapes as textures, rather than spatially rigorous alignment to a datum) – perhaps due to approaching the problem only from a computer science perspective, rather than from a geovisualization/GIScience/geoscience perspective.

Geographic researchers in Canada recently combined spatial analysis, spatial gaming, and MAR to deliver a suite of environmental situated spatial gaming applications (see Harrap et al. 2012).

Spatially focused MAR geoscience visual analytics systems that can perform serious analyses and simulations are a major next goal for geographic MAR research. Two of the first examples of this were recently completed at the Spatial Interface Research Lab, at Simon Fraser University. A Touch of Rain enables real-time 3-D particle fluid simulations to be run in urban environments (Hedley and Lonergan 2012; Lonergan and Hedley 2014; Figure 7). More recently, researchers have succeeded in implementing the first example of using mobile augmented reality to run a 3-D physics-based tsunami simulation in situ at a coastal field site in British Columbia (the system is called Tsunamiulator) (Lonergan and Hedley 2015). Significantly, both of these recent examples do more than just deliver MAR interfaces; they actually enable on-the-fly switching between virtual and MAR space – introducing the term flexible mixed reality (FMR). These recent advances suggest that we are perhaps as close as we have ever been to realizing “augmented GIS” using situated mobile augmented reality, and importantly, linking virtual spaces of analysis and simulation with real spaces of sense-making and interpretation.
Figure 7  Running real-time 3-D physics-enabled particle simulations using a flexible mixed reality interface. Lonergan and Hedley 2014, Figure 2, with permission of the author.
AUGMENTED REALITY

Theoretical perspectives and opportunities for new sociotechnical research

With a whole-discipline audience in mind, it is important to note that augmented and mixed reality interfaces are far more than interface technologies. The real power of emerging geovisual interface technologies is not just their ability to deliver new ways to represent, explore, and understand complex geographic phenomena. It is their ability to be used by citizens, absorbed into social practice, to create new places of engagement, and new spaces of representation. In particular, they may enable us to link a multiplicity of social and cultural relationships with the same geographic spaces – allowing us to better understand the many forms of human encodings of landscape, and the way we (do and could) link these meanings to social and virtual inscriptions. In short, they define new spaces and forms of geographic representation, mediation, experience, and knowledge exchange. There is much to be done to develop new theory about augmented and mixed realities as sociotechnical hybrids.

In geography to date, interactions with geographic information and enabling tools have been engaged by research communities of GIScience, geographic visualization, geovisualization (Fisher and Unwin 2002; Dykes, MacEachren, and Kraak 2005), and more recently, geovisual analytics (Andrienko et al. 2009). However, society’s use of spatial information in “virtual” spaces (such as the Internet, social networks, massively multiplayer online games, shared virtual environments) has proliferated in a way that influences – and has become intertwined with – real spaces, social practice, behavior, and environmental perception (e.g., persistent virtual environments, AR-capable location-based apps). These trends have resulted in new questions about where reality as we know it ends, and where virtuality begins. A burgeoning field of discourse has emerged on the notion of “cyberspace” (Gibson 1984). Cyberspaces have been defined and debated in geographic discourse (see Benedikt 1991; Batty 1993; Graham 1998; Dodge and Kitchin 2001; Kitchin and Dodge 2011). Much of this discourse has focused on conceptualizations of space and place modified by the emergence of information communication technologies (ICT), virtual spaces, and new forms of representation and data in a socially networked society. Kitchin and Dodge (2011) remind us that “space is not simply a container in which things happen; rather, spaces are subtly evolving layers of context and practices that fold together people and things and actively shape social relations.” More research is needed to understand the amorphous linkages and entanglements that occur between real and virtual spaces, and how they are appropriated and mobilized in society – something that Batty (1993) seemed to envision, and implicit in “metaverses,” introduced by Neal Stephenson in 1992. Stephenson’s Snow Crash (1992) helped us begin to consider that cyberspace is not just a separate place to be viewed from without or within but a multiplicity of potential places defined by those who use and share it. Moreover – and crucially – metaverses coexist with traditional geographic spaces, enabling humans to step seamlessly between them. With a groundswell of social networking culture we now regularly “tap into,” use, share, and transact social and professional actions through many metaverses on a daily basis.

Major opportunities exist to study situated virtuality, and mobile augmented reality embedded in human social and cultural contexts, as sociotechnical hybrids. In doing so, we may consider the multiplicity of ways in which
spatial knowledge and awareness are represented and transacted – contributing to wider understandings of emerging geovisual information environments as sociotechnical hybrids, rather than just as technological gadgetry.

Theories of sociotechnical hybridization, cyberspace, cyborgs, code/space, and metaverses might inform new research into: the nature of emerging transactions between humans in real and virtual spaces; the potential (and implications) of mobile augmented reality to deliver new forms of situated geovisual analysis of human activities; the transduction of geovisual information between real and virtual space; new forms of narratives in geographic space; the identification of the geo-information and geo-experiential transactions that can occur between reality and virtuality; the exploration of the synchronicity of these transactions, and the modification of time; the emergence of forms of presence, tele-presence and multipresence; the investigation of the relationship between AR geovisual information experiences and recoding geographic space.

**Summary**

This entry provides an introductory overview to AR, its technical foundations, and its conceptual context. A cross-section of selected nongeographic and geographic AR research (including TAR and MAR) has been presented to provide the reader with a sense of key research to date. Recent examples by geographic researchers demonstrate the emergence of AR geovisualization, geovisual analysis, and mixed reality situated simulation as new capabilities for our discipline. This research establishes a foundation on which to develop new ways to link geographic spaces of fieldwork, interpretation, and communication with virtual spaces of analysis, simulation, and visualization. New empirical research will be needed to evaluate the impact of mixed reality interfaces on geographic tasks and knowledge construction. New geographic AR capabilities also open up new opportunities to extend existing theories of cyberspace and metaverses, and develop new theory for geographic mixed realities as sociotechnical hybrids.

**SEE ALSO:** Geography education: digital and online trends; Virtual geographic environments; Virtual reality

**References**


AUGMENTED REALITY


Australia: Geographical Society of New South Wales (GSNSW)

Founded: 1927  
Location of headquarters: Sydney  
Website: www.geogsoc.org.au  
Membership: 150 (as of March 17, 2013)  
President: Gordon Waitt  
Contact: gwaitt@uow.edu.au

Description and purpose

The Geographical Society of New South Wales is a professional society whose members are geography academics, teachers, practicing geographers from both the public and private spheres, and other interested members of the public. The society's mission is to advance geography in New South Wales and throughout Australia. It is dedicated to the promotion, support and, when necessary, the defense of geographical research, scholarship, and education. It encourages geographical submissions to appropriate authorities on the management of environmental and social issues in Australia. The society aspires to an expansion of geographical literacy among the public of New South Wales.

Journals or major publication series

Australian Geographer. http://www.tandfonline.com/toc/cage20/current

Current activities or projects

The society is responsible for Australia’s oldest academic international geography journal, Australian Geographer. The society also holds an annual prize night at which it recognizes excellence in the governance, teaching, and the study of geography through awarding prizes. It organizes the annual Honours Conference for final year undergraduate students to present findings from their research and awards prizes for the best presentations. It also hosts regular events for postgraduate students engaged in geographical research. The society organizes a program of study tours and runs the Travellers Club where members regularly give presentations of their travels.

Brief history

The first meeting of the Geographical Society of New South Wales was held on August 10, 1927, with Associate Professor Thomas Griffith Taylor elected as its inaugural president. In 1928, the society launched the first issue of its journal, Australian Geographer. Since the 1960s, the society has sponsored and led more than eighty study tours to many overseas countries. Participants in these study tours encouraged the society to establish its Travelers Club in 1988. To recognize the achievements of its members, it established the Macdonald Holmes Memorial Medal in 1977. The society has also awarded its Fellowship biennially since 1969. Prizes to university undergraduates have been awarded since 1943. As part of its outreach, the society
has sent delegates to the National Committee for Geography at the National Academy of Science in Canberra since 1995. It has been represented on the Geographical Names Board of New South Wales since its establishment in 1966. Since 1997, the GSNSW has hosted an annual conference where final year undergraduate honors students present the findings of their research. The society sponsored its first conference expressly for post-graduate students in 2007 and continues to hold regular events for early career researchers.

Submitted by Garth Lean
Description and purpose

To represent geography and those who contribute to it in Australia. The aims of the institute are: to promote the study and practice of geography in Australia; to advance the study of Australian geography internationally; to hold and sponsor meetings and conferences to present, discuss, and disseminate geographical studies and research; to publish the results of geographical research in a journal; to represent the interests of professional geographers in Australia; and to cooperate with other organizations with kindred purposes.

Journals or major publication series


Current activities or projects

The IAG holds an annual conference, moving between cities, with a joint conference with the New Zealand Geographical Society every 4 years (www.iag.org.au/conferences-events). It promotes research through the publication of refereed journals, *Geographical Research* and *GEOView* (which publishes undergraduate research). The IAG provides substantial financial support for postgraduate student members to attend the International Geographical Union Congress. It supports nine study groups with additional groups under consideration (www.iag.org.au/study-groups). Finally, the IAG presents awards to recognize the achievements of geographers in the discipline, the institute, and fostering international links for Australian geography (www.iag.org.au/about/awards-of-the-iag).

Brief history

The IAG was formed in 1958 with its first conference held in 1960. Prior to this time, geographers had been meeting annually within the Australian and New Zealand Association for the Advancement of Science (ANZAAS). The lack of a national organization for Australian geography prevented affiliation with the International Geographical Union, and the existing state-based geographical societies covered only Queensland, New South Wales, and South Australia.

By 1963 the three objectives for the establishment of the IAG had been fulfilled. There was a specialist conference for geographers separate from ANZAAS, Australia had been accepted as a member of the International Geographical Union (in 1960), and there was a national geographical journal called *Australian Geographical Studies*. This journal title was changed to *Geographical Research* in 2005.

Throughout its history, the IAG has been the principal body representing geographers and promoting the study and application of geography in Australia, and for advancing the study of Australian geography internationally.

Submitted by Phil McManus
Description and purpose

The main purpose of the ÖGG is to support the dissemination of geographical research and knowledge throughout society and the scientific community. To accomplish this aim, ÖGG publishes the (SCSI-referenced) journal, *Mitteilungen der Österreichischen Geographischen Gesellschaft* (Communications of the Austrian Geographical Society), organizes lectures, symposia, and field trips in Austria and Europe, and honors the scientific work of young scientists in all fields of geographical research by awarding several prizes. ÖGG’s headquarters are located in Vienna, while branches exist in several other Austrian cities. Thematically, ÖGG is composed of working groups devoted to specific research fields of geography and cartography. Membership categories are honorary, ordinary, and student members with or without subscription to *Mitteilungen der Österreichischen Geographischen Gesellschaft*. The general assembly, which convenes annually, elects the executive board, the president, and other officers.

Journals or major publication series


Current activities or projects

A poster exhibition on the transformation of the Austrian Alps on the 25th anniversary of the International Alpine Convention is shown at several locations in Austria and Germany.


A symposium was held in Vienna on December 3–4, 2014 with the title “10 Years of EU Eastern Enlargement – The Geographical Balance of a Courageous Step.”

Brief history

Founded in 1856, the ÖGG is one of the oldest geographical societies worldwide and experienced its heyday in the period between its foundation and World War I. Founded as the Imperial–Royal Geographical Society in Vienna, it had the support and recognition of the public, the scientific community, and political elites. Well-funded by the state and private promoters, the ÖGG conducted several exceptional international expeditions, for example, to the eastern Sudan, to Congo, and to the North Pole in 1872–1873, which led to the discovery of the Franz Josef Land archipelago.
AUSTRIA

The collapse of the Austrian-Hungarian Empire after World War I and the end of the age of European discoveries led to a decline of ÖGG. In the 1920s and 1930s, ÖGG faced adverse conditions including financial cutbacks and a decline in membership. During World War II, ÖGG was incorporated into Nazi-affiliated organizations related to teaching, which led to further decline.

After World War II, when academic institutions expanded and new universities were founded in Austria, the restored ÖGG was again able to expand its activities and to increase its membership in Austria and abroad. New branches were established and ÖGG succeeded in becoming a recognized institution in the academic landscape. Current challenges are the support of students as they embark on a scientific career as well as introducing academic research into public debates demonstrating that geography is relevant for society at large.

Submitted by Robert Musil
Autobiography

David Butz
Brock University, Canada

Autobiography is generically defined as a factual account of a person's life, recounted by that person using the first-person singular voice. However, authors and their audiences routinely code as autobiographical a wide range of communicative acts that exceed this narrow understanding, thereby signaling the inadequacy of equating autobiography with the formal characteristics of a particular type of first-person communicative artifact: “an autobiography.” Rather, what unites the variety of ways in which the designation is employed, and what distinguishes autobiography from other forms of communication, is that its subject and object of signification are understood to be the same, or to overlap significantly. This characteristic – or attribution – gives autobiography distinct epistemological status: to produce or receive communicative acts as autobiographical is to assert and open to challenge knowledge claims based on privileged and often exclusive insider access to the circumstances, experiences, and feelings being communicated and analyzed. Autobiography, then, is a specific epistemological orientation toward communicative acts on the part of their producers or by those who receive them; it has less in common with biography (despite shared focus on “life-writing”), than with in-depth interviews, oral testimony, ethnography, storytelling, self-directed photography, autoethnography, and everyday conversation, all of which to some degree claim authority and veracity on the basis of close correspondence between who is communicating and what is being communicated.

The assertion that autobiographers are articulating themselves from a privileged position of access is simultaneously autobiography’s key feature, its strongest attraction, and its most complicating characteristic, for those who produce autobiographical communication and who view autobiography as a potential source of data. As an often intensely and self-consciously subjective mode of address, autobiography is understood to enable the communication of personal experiences, situated understandings, subjugated perspectives, self-reflexive analyses, emotion, and affect, all of which are gaining legitimacy as aspects of geographical knowledge and expression. At the same time, the seemingly straightforward simultaneity of representer and represented in autobiographical communication can mask difficult epistemological questions about individuals’ capacity to access, analyze, and articulate their own experiences; the relationship between individual lives as told autobiographically and the social contexts through which they are constituted; the relationship between autobiographical authors and their audiences, including the capacity of audiences to understand and analyze autobiographical communications; and the power relations inherent in efforts to articulate social reality via self-narration. Purcell (2009, 235) calls these issues “critical destabilizations” that lead scholars involved with autobiographical expression to “accept that they confront a complex and mutable set of interrelationships between self and other, individual and society, and author and audience.”

Geographers have acknowledged these critical destabilizations to varying degrees, and in
various ways, as they have used – or avoided – autobiography as a means of communication and analytical resource. Pamela Moss remarked in 2001 that “autobiographical approaches are not widely accepted as a standard method of geographical research” (p. 190). The ambivalence toward autobiography Moss describes may be traced to two countervailing tendencies within geography: a lingering positivism that dismissed autobiography for lack of scientific objectivity, and a poststructuralist cultural turn that viewed suspiciously the imbrication of autobiographical accounts in larger discourses of power and subjugation. Since then, the productive potential of autobiography and associated practices of authoethnography and storytelling have gained greater recognition, in part because autobiographical practices have matured in response to post-structural critiques, but also because developments in theory have reduced geographers’ inclination to view the individual and particular solely as expressions of larger ideologies and power structures.

**Autobiography as academic practice**

**Humanistic**

Prior to the poststructural turn that transformed much social analysis in the 1980s, autobiographical accounts were typically represented as authentic expressions of a relatively autonomous humanist subject. This traditional approach conforms most closely to autobiography’s generic definition, and deals with the critical destabilizations of self-narration largely by ignoring them. Humanistic approaches may be subdivided into two subcategories. One of these, the so-called great person approach (Purcell 2009), pursues the goal of comprehensively recalling an important or influential person’s life, usually with the implication that it is worth understanding on its own terms, and also – in the case of pre-eminent geographers – because of what an important insider’s self-narration may reveal about the development of the discipline. This style of autobiography experienced a minor resurgence in the early 2000s, and continues occasionally to be employed in geography, typically as an instrument of disciplinary identity, continuity, or self-questioning. One of the purposes of this style of autobiography is to “set the record straight”; therefore, it claims a degree of facticity, if not outright objectivity, that subsequent approaches to autobiographical expression avoid.

The same epistemological move is evident in many of the autobiographical gestures that routinely and unselfconsciously pepper works that are not otherwise coded as autobiographical. As Moss (2001) notes, geographers frequently insert personal examples and anecdotes, mini-confessions, arrival stories, vignettes from the field, and first-person asides, to lend their writing warmth and personality, and to bolster its legitimacy as knowledge, often through the disguised or explicit assertion: “I was there, I should know.” Often – as in much ethnography – such empiricist claims are at odds with a work’s overall epistemology. In other increasingly common instances self-referential interjections are treated more reflexively.

The second subcategory of humanist autobiography is phenomenological self-narration. Here the objective is less to establish an individual’s importance within a given social context, than to articulate the nuances of everyday human experience. In this case an autobiographical subject’s ordinariness may be more important than his or her pre-eminence. If the great person approach flirts with assertions of objectivity, the phenomenological approach emphasizes autobiography’s intense subjectivity on the grounds that the most important aspects of human experience, including place experience,
must be accessed and articulated subjectively. Phenomenological autobiographers narrate their own lives because only their own subjective experience is sufficiently accessible to them to describe with the required “thickness” of detail.

**Reflexive**

As part of a feminist and poststructural critique of objectivity that became influential in the mid-1980s, geographers and others began to question the conventions by which their disciplines constructed and represented knowledge. A so-called crisis of representation ensued in which conventional scholarship was criticized for denying the partiality and positionality of its knowledge claims, thereby reproducing the flawed assumption that academic knowledge is unrelated to the subjectivity that produces it. One solution that was proposed was termed critical reflexivity: that is, explicit, systematic self-scrutiny on the part of researchers regarding their subject position vis-à-vis their objects of research, with the goal of understanding and turning to analytical advantage the inevitably situated nature of the research process and the knowledge it yields. Critical reflexivity inevitably involved a turn to more autobiographical forms of communication. However, unlike the humanistic styles of self-narration outlined above, the goal of reflexive autobiography is not to legitimize the facticity of knowledge or to focus analysis exclusively on researchers’ own lives, but rather to destabilize academic authority by emphasizing the socially and politically constituted nature of knowledge claims, and to gain legitimacy in a different register by demonstrating attentiveness to epistemological critiques of objectivity and impartiality.

Various styles of reflexive autobiographical writing have recently been grouped under the umbrella term “autoethnography” (Butz and Besio 2009), because the objective of self-narration is to focus introspectively on the self in order to illuminate some exterior worldly phenomenon: hence an ethnographic objective. Moreover, to the extent that critical reflexivity is predicated on blurring the lines between knowing self and known other, the subject/object of reflexive autobiography is often plural, and in this sense ethnographic rather than conventionally autobiographical.

Butz and Besio (2009) enumerate five main categories of autoethnographic practice, most of which predate the crisis of representation, which together give some sense of the range of reflexive autobiographical communication (see Table 1). Although these practices vary considerably in terms of their objectives and emerge from a range of speaking positions, they all involve a reflexive transition, whereby authors move along a continuum between agent and object of signification. This reflexive ontological move helps to break down the conventional distinction between knower and known that was the focus of political critique associated with the crisis of representation.

The first three autoethnographic categories involve academic researchers narrating individual or collective self-experience in order, respectively, to understand some social or cultural phenomenon that is informed by but exceeds their experience, to analyze the epistemological and political implications of their fieldwork relationships and the resulting knowledge, and to gain some methodological or interpretive benefit from insiderness. In these cases the transition is from agent to object of signification. The second of these, so-called narrative ethnography, is by far the most prominent in geography, and it accounts for much of the discipline’s modest recent uptake in autobiographical expression. The last two forms of autoethnography involve members of historically subordinated groups
AUTobiography

Table 1  A continuum of autoethnographic practices.

<table>
<thead>
<tr>
<th>Academic researcher</th>
<th>Research subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent of signification</td>
<td>Object of signification</td>
</tr>
<tr>
<td>Reflexive journey</td>
<td>Reflexive journey</td>
</tr>
<tr>
<td>Personal experience narrative</td>
<td>“Insider” research</td>
</tr>
<tr>
<td></td>
<td>“Native” ethnography</td>
</tr>
<tr>
<td></td>
<td>“Subaltern” autoethnography</td>
</tr>
</tbody>
</table>

Source: reproduced from Butz and Besio 2009, 1665, with permission of John Wiley & Sons.

producing collective self-representations that attempt to intervene in dominant accounts of themselves and their worlds, thus claiming status as agents of signification. In the case of indigenous ethnography this occurs from within or at the margins of academic circuits of knowledge, whereas subaltern autoethnography is a nonacademic assertion of self-identity and self-determination intended for metropolitan audiences. Insider research, native ethnography, and subaltern ethnography have much in common with what Purcell (2009) calls insurgent autobiography, the objective of which is to destabilize dominant assumptions and discourses, often from the perspective of the marginalized or oppressed. In geography, insurgent autobiographies have been written, for example, from the perspectives of women, visible minorities, graduate students, and nontenured professors. Critical destabilizations of the relationships between self and other, individual and society, and author and audience, are acknowledged and employed productively in each of these forms of autoethnographic practice, although not without occasionally lapsing into simplistic claims of identity, a tendency that is most apparent in some insider research and indigenous ethnography.

Generative

Drawing primarily from feminist, postcolonial, and antiracist theorizing, and in keeping with the assertion that the personal is political, geographers have frequently used individual self-narratives in hopes of generating social, political, and intellectual change. Feminist geographers, in particular, have insisted on the importance of narrating and analyzing individual women’s experiences as a way to challenge the totalizing discourses of a masculinist discipline and society. Other marginalized subjectivities within geography have engaged in similar autobiographical projects. In these cases, personal storytelling is employed as an epistemological resource to generate knowledge that contests the partial and ideological nature of geography’s dominant theories, assumptions, findings, and disciplinary practices, and also to build political solidarity within and among groups that have been marginalized in authorized circuits of knowledge. Some of this personal narrative work also aims to be generative in the additional
sense of storying into being new subjectivities and realities. This is to an extent implicit in the notion of forging oppositional political solidarity, but for authors like J.K. Gibson-Graham (2006), autobiographical writing has the potential to be something more, a “performative ontological politics,” which Cameron (2012, 580) describes as “a practice that involves writing about, engaging with, performing, and taking seriously the alternatives they wish to see in the world as an act of conscious, political, and creative (re)production.” It includes narrating oneself in ways that erode subjectification to dominant discourses while resubjectifying the self to alternative emergent possibilities: that is, modeling in the self the future that one desires. Gibson-Graham has applied this approach to their own self-narratives, and has guided research participants to do the same. In this and other examples of generative autobiography another of autobiography’s critical destabilizations may be discerned: between an already existing narrating subject and an emergent subject of narration. A similar destabilization is characteristic of subaltern autoethnography, which requires a certain transculturally oriented reconstitution of self.

Affective

Recently a small number of geographers have employed autobiographical expression with the goal of either building or representing affect. In terms of the former, for example, the idea that personal stories can emotionally move both storytellers and listeners in ways that create openings for the construction of communities of solidarity based around affectual connections is an important part of Gibson-Graham’s performative ontological politics. The emphasis is as much on the embodied practices of telling and listening – sharing personal narratives – as on the specifics of content. Personal stories are understood as more than exercises in critical reflexivity that speak to audiences in a rational or intellectual key: they also build affect by appealing to the embodied, placed, felt, and intersubjective attributes of narrating and listening subjects. This form of autobiographical address is more often oral and face-to-face than mediated by text, and may involve drama, music, or other performative modes. Other work in this vein is less concerned with building collective identity through the affective power of personal narrative than with disrupting existing lines of affiliation and communities of identity that are formed around axes of exclusion and oppression (Pratt 2009; Cameron 2012). In both cases the objective is to produce transformative ethical relations, although this outcome is never assured. The key critical destabilizations of autobiographical acts aimed at building affect pertain to the relationships between self and other, author and audience.

Another emerging strand of affective autobiography finds inspiration in nonrepresentational and (post)phenomenological currents, and strives to articulate (rather than construct) affect. Like phenomenological autobiography in the humanist tradition, this work is interested in the fine details of human experience and feeling, but with greater emphasis on embodied affect, especially as it relates to place, terrain, landscape, and movement. The objective is to articulate affect in ways that keep it affective, forestalling the rush to rational critique. The resulting texts are often highly lyrical, and like personal experience narrative autoethnographies have been criticized as difficult to assess, generalize from, intervene in, or relate to a world beyond themselves. Practitioners of this style of affective autobiography respond that such criticisms miss the point, which is not to present conventional
arguments or offer sources of generalization, but rather to evoke fleeting experiences, or hit readers with a flash of recognition that exposes new thinking possibilities. This critique and its response suggest that geographers need to develop new methodologies of reception that enable them to be affected by self-narratives operating in an evocative register. The imperative for better-developed methodologies of reception applies to other forms of autobiographical address as well, as many geographers continue to find first-person self-narration too personal, subjective, and self-indulgent to be useful.

Autobiography as data

In addition to its utility as a productive form of academic practice, autobiographical communication is often also understood as a useful form of data, notwithstanding what some perceive to be its overly anecdotal and solipsistic character. The distinction between autobiography as practice and as data is heuristically useful, but also somewhat artificial, for two reasons. First, autobiography is primarily a knowledge claim based on the attribution of close correspondence between who is communicating and what is being communicated. That claim can be made either by selves who produce communicative acts, or by selves who receive them as autobiographical data. To use autobiography as data is also to define certain categories of expression as autobiographical; this too is autobiographical practice in its epistemological sense. Second, various forms of reflexive autobiography – especially personal experience narrative and narrative ethnography – produce self-narrations for authors’ explicitly analytical purposes, thus blurring the line between representing self and analyzing self-representations.

Geographers have employed autobiography as data in six main overlapping and sometimes contradictory ways. First, autobiographical accounts have been useful for accessing perspectives and experiences that are otherwise difficult to obtain. The lives of Canadian pioneer women, for example, received scant attention in official contemporary documentation, but are accessible through autobiographies, journals, and correspondence. These materials may be useful for establishing the “facts” of life on the frontier, as well as for analyzing the discourses associated with, for example, settlement and colonial femininity. Second, so-called great person autobiographies provide useful material for unpacking disciplinary histories, especially as they relate to the relationship between personal biographies and intellectual currents, and several have been written with this purpose in mind. Insurgent autobiographies by marginalized subjects within geography have similar value as data, both to construct and deconstruct geography’s history. Third, geographers have found autobiographical material valuable for its emotional, experiential, and subjective characteristics. Like other types of qualitative data, autobiographical accounts offer insights into people’s opinions, values, experiences, behaviors, and feelings. Most qualitative data has an autobiographical dimension, which often accounts for its compelling depth. Fourth, geographers are turning increasingly to autobiographical communication for what it can teach them about affect, and especially the affective nature of geographical experience. A difficulty here is that often the more effective an autobiographical act is at evoking affect, the more difficult it is to treat analytically as data. Fifth, autobiography has become an important tool and object of analysis.
for scholars who produce self-narrations for critical reflexive purposes. By analyzing their own imbrication in a social context, a field of power, or a set of research relationships, scholars hope to gain a more rigorous grasp on their research object, as well as on the epistemological characteristics of knowledge they generate. Sixth, some geographers working in transcultural contexts of structured inequality are finding utility in examining the subaltern autoethnographies of those they are researching, both for what they can learn about their research subjects’ strategic self-definitions, and as a resource of critical reflexivity. The latter is aided by recognizing that researchers are an important intended audience for subaltern autoethnography.

**Epistemological tensions**

The various practices and uses of autobiographical expression are cross-cut with tensions as to whether self-narrative material should be received as a source of self-evident facticity, as a resource for identifying and deconstructing cultural ideologies and broader discourses of power, or as something in-between: “small stories” (Lorimer 2003) that express local instances of everyday lived experience in some of their contingent particularity, mundaneness, corporeality, and affectivity. These tensions express theoretical and epistemological differences regarding the nature of human experience, the relationships between subjectivity and environment, and the epistemological characteristics of human self-expression, and reflect some of human geography’s broader social-theoretical divisions. Although these differences are unlikely to diminish as autobiography continues to gain legitimacy in geography, perspectives that treat the autobiographical as either reflexive or post-representational articulations of the self are likely to gain influence in relation to empiricist or humanist approaches. Geographers therefore need to develop better ways to receive, evaluate, and incorporate evocative and affective self-expression into geographical thought; that is, to create more effective autobiographies of reception.

**SEE ALSO:** Affect; Emotional geographies; Ethnography; Feminist methodologies; Humanistic geography; Identity; Nonrepresentational theory; Oral history and narrative; Positionality; Qualitative data; Representation; Subjectivity

**References**


Further reading


Avalanche meteorology

Jordy Hendrikx
Montana State University, USA

Snow avalanches are a significant natural hazard in the mountainous regions of the world, destroying property, disrupting transportation networks, and causing physical hazard. Global estimates of avalanche fatalities range from 200–500 per year, with a high level of inter-annual variability. According to the Panel on Snow Avalanches (1990) annual mortality rates in the United States due to snow avalanches exceed the average mortality rate due to earthquakes as well as all other forms of slope failure combined. Furthermore, avalanche fatalities are rising in the United States. Since the 1950s, when the five-year moving average was less than five fatalities per year, the average has increased to close to thirty avalanche fatalities per year (Logan and Witmer 2012). In addition to these fatalities, avalanches cause widespread disruption and damage to infrastructure and transportation corridors.

To mitigate avalanche hazard, daily avalanche forecasts are produced for many mountainous areas around the world. Avalanche forecasting is conducted at a range of scales from a given slope (micro) to several mountain ranges (synoptic). Avalanche forecasting is a probabilistic enterprise with a number of factors which need to be considered, all with varying weights that change over time and space. The factors used to forecast snow instability are commonly placed into three categories which, in order of decreasing entropy, are: Class III: meteorological factors; Class II: snowpack factors; and Class I: instability factors (McClung and Schauer 2006). Class I factors are direct measures or observations of instability such as recent avalanche activity and provide the most direct evidence of instability. Class II factors provide further evidence of instability such as snowpack structure or total snowpack depth, but are much less direct. While the Class II factor information is relevant to snowpack instability, it is of less importance than Class I factors. Finally, Class III factors provide indirect measures of current and or future snow instability. Furthermore, they are commonly not recorded in the start zone of avalanches so are considered proxy measurements that provide an index measure of instability. As the spatial area of the forecast increases, the overall accuracy and resolution generally decreases. For example, when avalanche forecasting for a mountain range, there is increased reliance on meteorological data (i.e., Class III factors) to forecast mountain weather and its role in producing avalanches. This is where the intersection of avalanche and weather forecasting occurs and has been termed avalanche meteorology.

Meteorologists work with, or sometimes as, avalanche forecasters, providing detailed, area-specific weather information for mountainous areas. Avalanche forecasters integrate this information with their knowledge of snowpack and terrain to create a more comprehensive avalanche hazard forecast that describes the instability of the snowpack and likelihood of avalanche occurrence. In the absence of detailed snowpack data, avalanche potential can be estimated based on weather information alone; however, the quality of the forecasts are not as complete or accurate under these circumstances. Similar to weather forecasting, avalanche
forecasting combines observations with scientific understanding and expert judgment. The avalanche forecasting problem is compounded by the number of weather variables involved, the complexity of the terrain, the spatial and temporal variability of the snowpack, and the lack of continuous detailed on-site weather and snowpack observations. This complexity results in a high degree of uncertainty and results in the avalanche forecaster having a central role in the creation of the forecast. This requires forecasters to develop their skills through obtaining years of experience to make accurate decisions in the light of high uncertainty about the likelihood of avalanche occurrence. This diverse and dynamic combination of factors also means that current computer forecasts of avalanche conditions have limited use without expert interpretation.

The process of avalanche forecasting is dependent on the combination of the weather preceding the current conditions, the current and future conditions, and to a much lesser extent the avalanche climate. When we consider the current or forecasted conditions (or Class III factors) we often focus on a subset of variables that are well understood to relate to instability. These factors commonly include, but are not limited to, depth and water equivalence of new snow, wind speed and direction, air temperature, solar radiation, and snow surface conditions.

Depth and water equivalence of new snow is one of the more important variables as it represents an additional weight on the snowpack. As the depth of snowfall and thereby snow water equivalence increases, so too does the load on a buried weak layer. Instability and the likelihood of avalanche activity usually increases with increased snowfall. This factor is often accompanied by wind, which through preferential loading and scour in mountainous terrain can lead to areas of increased and decreased load on the snowpack. Depending on the surface conditions of the snowpack, the wind can cause increased loading even in the absence of concurrent precipitation. Wind drift is often estimated as the cube of the wind speed, so even small increases of wind speed above the required threshold for drifting can result in substantial increases in loading for areas that receive preferential deposition (e.g., wind sheltered or lee slopes and features). Both the mean value and trend of air temperatures during a storm is important. The air temperature controls the temperature and thereby, in combination with the wind speed and humidity, the density of the snow that is deposited. Fluctuating air temperatures can result in low-density cold snow being overlaid by higher-density warmer snow leading to buried weaknesses within the storm snow. Solar radiation can change the physical properties of the surface layers of the snowpack and can affect snow stability. This is especially the case in spring on solar facing aspects where the snow undergoes rapid changes due to the balance of incoming and outgoing radiation. Some of these factors have been used to develop and train statistical models that predict avalanche days for a specific operation (e.g., road or rail corridor) with some degree of success. However, these resulting models are only of limited value when applied to another location and also struggle to achieve overall prediction rates exceeding 85% (Hendrikx, Murphy, and Onslow 2014).

While these Class III factors clearly control the avalanche activity, the nature, magnitude, and impact of these factors are commonly functions of the avalanche climate and specific location. Using a simple threshold for one parameter (e.g., depth of new snow) will not necessarily translate across avalanche climates or locations. So while an avalanche climate alone will be of limited value for avalanche forecasting, due to the highly dynamic temporal variability of the
snowpack which responds to short-term weather parameters as discussed above, an avalanche climate can describe the typical weather patterns (or climatology) and resulting conditions of the snowpack, which result in a characteristic snowpack. Given that the characteristic avalanche behavior varies between avalanche climates, this approach provides a useful framework for discussing avalanche forecasting and meteorology in general terms and aids a forecaster in understanding the critical Class III factors.

In a study from the United States, Mock and Birkeland (2000) identified three primary avalanche climates: maritime (or coastal), continental, and intermountain. These terms are used globally with the intermountain term often being replaced by transitional. These avalanche climates are differentiated by their respective average snowpack depth, average winter time temperatures, and frequency and average density of new snow. When considered over climatology time frames (i.e., 30 years), locations will often be characterized into a typical avalanche climate, but may experience years in which conditions are more continental, intermountain, or coastal.

Continental climates have colder average winter time temperatures, more clear skies, and less snowfall, all of which are very conducive to the formation of depth hoar and other persistent weak layers. Typical examples of this type of avalanche climate include the Colorado and eastern British Columbia Rocky Mountains in North America, the Brooks Range in Alaska, and the Pamirs of Asia. As with other avalanche climates, most avalanches in a continental avalanche climate are caused by snow loading from direct snow precipitation or snow drift; however, the major avalanche cycles are often associated with buried persistent weak layers. Due to these weak layers, large and destructive avalanches in continental areas may result from only small snowfalls or wind events, sometimes many days or weeks after new precipitation. Mock and Birkeland (2000), using data from 48 locations in the United States, identified an average winter temperature of less than \(-7\, ^\circ C\), December snowpack temperature gradients of greater than 10\, ^\circ C, as defining criteria for continental avalanche climates.

In contrast, maritime avalanche climates have warmer temperatures, cloudier skies, and more snowfall and potential for rain, resulting in the formation of fewer weak layers. Typical examples of this type of avalanche climate include the Cascade Range in the United States, Coast Mountains of British Columbia, Canada, and the mountains of western New Zealand and Norway. Avalanches in maritime avalanche climates tend to be the result of large snowfalls and they often involve only the new snowfall, therefore relying predominantly on daily to multiday precipitation variables. As a result of this, most of the avalanche activity in this climate occurs within 24 hours of the new snowfall. Mock and Birkeland (2000) identified an average winter temperature warmer than \(-3.5\, ^\circ C\), snow water equivalence of greater than 100 cm, and more than 8 cm of rain as defining criteria for maritime avalanche climates.

Intermountain or transitional areas have intermediate conditions between these two zones. However, intermountain areas are ones with much variability as during some winters the avalanche climate will have characteristics that can be predominately continental or coastal. Coastal conditions might spread inland or continental conditions might extend toward the coast. Mock and Birkeland (2000) identified an average winter temperature warmer than \(-7\, ^\circ C\) and snowfall greater than 560 cm as the defining criteria for intermountain maritime avalanche climates when compared to continental avalanche climates.
AVALANCHE METEOROLOGY

Since avalanche forecasting techniques vary depending on the climate type, understanding and being able to predict these seasonal shifts in climate types clearly has important implications for enhancing avalanche forecasting and ensuring that the most pertinent weather parameters (i.e., Class III data) are being observed and focused on by the avalanche forecaster. While lower entropy factors like direct observations of avalanche activity provide more direct information related to snow stability, avalanche forecasting using an understanding of avalanche meteorology principles can be of critical importance when forecasting for large and or remote regions with spare snow and avalanche observations.

SEE ALSO: Environmental hazards; Mountain climatology; Natural hazards and disasters; Snow; Snow cover; Snow and ice avalanches

References


Azerbaijan: Coğrafiya İнституту Azərbaycan Milli Elmlər Akademiyası (Institute of Geography Azerbaijan National Academy of Sciences (IG ANAS))

Founded: 1945
Location of headquarters: Baku
Website: www.igaz.az
Membership: 2000 (as of December 31, 2013)
President: Ramiz Mammadov
Contact: ramiz.mamedov@geo.ab.az

Description and purpose

The Institute of Geography Azerbaijan National Academy of Sciences was organized in May 1, 1945. The main scientific work focuses on: the development and environment of Azerbaijan including study of its natural resources and their ecogeographical variation; problems of production, social infrastructure, and economic and geographical problems of territorial organization of population; constructive and regional problems of the republic; desertification processes; and factors causing the changing levels of the Caspian Sea and the study of its ecological conditions.

Now employing 210 people, the institute has 13 research departments: landscape and landscape planning; climatology and agroclimatology; hydrology of land and water resources; economical and political geography; geography of the population of Azerbaijan and social development; history of geographical thought; geomorphology and natural risk; paleogeography; geography of land resources of Azerbaijan; ecogeography; problems centered around the Caspian Sea; cartography and geographical information; and the Pirgulu station for geographical research.

The institute also coordinates several councils, such as the Scientific Council; Dissertation Council; Doctoral Council; Council of Young Scientists; Coordination Council on “Geography”; Coordination board of Caspian Sea’s problems; and the Council for the Scientific Library.

Current activities or projects

Activities and projects include: ecochemical properties of the oil polluted landscapes of the Absheron peninsula and methods of remediation; information about the industry and environment of the Caspian Sea; climatic changes and ecosystem of the Caspian Sea (model research); Landscape planning in the southern Caucasus; monitoring of oil pollution with utilization of ground observations with a multisensory, multiplatform approach; interdisciplinary analysis of the ecosystem of the Caspian Sea (MACE). IG ANAS also hosted an international conference, Estimation and Rational Utilization of Natural Resource Potential of Geosystems under Conditions of Global Changes, in Baku in 2013. The Institute has a doctoral program enrolling 35 students.

Brief history

The Azerbaijan branch of the Academy of Sciences of the USSR was established in 1937, based on the 1925 report of the Central Committee of Azerbaijan entitled On the Study of Natural
AZERBAIJAN

Resources of Republic and Region. This effort laid the foundation of the Azerbaijan geographical school which began systematic geographical research in Azerbaijan. On May 1, 1945, the Institute of Geography of the Academy of Sciences of the Azerbaijan SSR was organized. In July 1945 the departments of physical and economical geography and cartography began operation. One of the main themes of research was the geography of mountainous countries. At first, none of the staff had scientific degrees. However, by 1947, eleven of the twenty staff held scientific degrees and three were enrolled in the graduate program. By 1950, the institute consisted of six departments. As of 1959 the Centre of Problems of the Caspian Sea has been part of the Institute of Geography.

Submitted by Ramiz Mammadov
Bangladesh: Bangladesh Geographical Society (BGS)

Founded: 1955
Location of headquarters: Dhaka
Website: http://bgs-bd.com
Membership: 500 (as of December 31, 2012)
President: Shahnaz Huq-Hussain
Contact: shuqhussain@gmail.com,
       mmrahman2000bd@yahoo.com

Description and purpose

The BGS was established to enhance and disseminate geographical knowledge and to promote education and research through its journals, occasional monographs, newsletter, and other publications; local and international seminars, workshops, annual meetings; and public displays and exhibitions. It has published a number of research monographs, conference proceedings, and anthologies. The society is supported through membership subscriptions and the sale of its journals and publications. The society receives an annual grant from Dhaka University and occasional grants from the Ministry of Science and Technology and other sources. The society was first represented at the IGU meeting in Canada in 1972 and became a member of the IGU in the same year.

Besides publication, the BGS plays professional and academic roles in developing the discipline of geography and environment in Bangladesh and applying geographical knowledge and research in education, business, and government. Its members play an important role in national development.

Journals or major publication series

Bhugol o Poribesh (journal, in Bengali)
Bhugol Porikroma (newsletter, in Bengali)

Current activities or projects

Being the oldest professional geographical society in the country, the BGS encourages students by giving the Nafis Ahmad Memorial Award to the most meritorious MSc student of geography at Dhaka University every year in honor of the founding president, the late Professor Nafis Ahmad. Another award – the Rizvi Memorial Award – is given every year to the best student of the BSc (Honors) course in geography at Dhaka University. Since its inception, the society has had exchange programs with various research and professional organizations in other countries.

Brief history

Established as the East Pakistan Geographical Society in 1955 and renamed the Bangladesh Geographical Society (BGS) after the war of liberation in 1971, it is the oldest professional body of geographers in the country. The then head of the Department of Geography of the University of Dhaka, Professor Nafis Ahmad,
was the founding president of BGS. The society was inaugurated by the eminent British geographer, Sir Dudley Stamp. The Department of Geography (renamed the Department of Geography and Environment in 1996) at Dhaka University offered space for the BGS. The society has four categories of membership — donor member, honorary life member, life member, and general member. There are institutional members at home and abroad. BGS maintains a reference library that is used by students, researchers, and members.

Submitted by Shahnaz Huq-Hussain
Bangladesh: Bangladesh National Geographical Association (BNGA)

Founded: 1972
Location of headquarters: Dhaka
Website: www.juniv.edu/geograph/bnga/
Base flow

Douglas A. Howard
Georgetown University, USA

Base flow is the groundwater contribution to stream flow. This is accomplished as groundwater intercepts a stream channel and initiates groundwater discharge into the stream from points of higher to lower elevation in the water table. Streams that receive large portions of their flow from groundwater base flow tend to have relatively low variability and are more reliable water sources. Precipitation infiltration recharges groundwater reservoirs and as the water table rises, groundwater discharges into nearby streams will also increase. For base flow streams (perennial or intermittent) the amount of groundwater discharge is directly proportional to the hydraulic gradient toward the stream.

Langbein and others (1947, 6) defined base flow (also referred to as base runoff) as “sustained or fair-weather flow” that “may be considered as composed largely of ground-water effluent.” Although the modern definition of base flow does not explicitly include anthropogenic influence on the stream, here we will adopt the definition referring to naturally occurring streams.

Stream flow water is derived from components of the hydrologic cycle, primarily runoff and base flow and, in lesser amounts, throughflow, interflow, and direct precipitation. Base flow, originating from the saturated zone within the groundwater table, hydraulically connects groundwater to stream flow and contributes about 40% to the total annual stream flow volume. Base flow is a direct contribution from the saturated zone within the groundwater table, while throughflow and interflow provide highly variable water volume contributions from the unsaturated zone (vadose zone). Given that no part of the hydrologic cycle exists in isolation, base flow is a significant contributing factor. As such, the interaction between groundwater and surface water flow regimes is a topic of extensive research to understand the exchange of both naturally and anthropogenically introduced contaminants and pollutants in our water resources.

The contribution of groundwater to stream flow varies based on climate, hydrogeological responses to seasonal regional weather conditions, and topographic and hydrogeologic constraints, as well as anthropogenic activities. Over long dry periods, between water-input events such as rain or snowmelt, aquifers contribute the bulk of water to streams where the groundwater table is elevated higher than the stream channel. The water table establishes a hydraulic gradient, generally following the slope of the topography, which causes groundwater to flow. The direction of groundwater flow is from points of relatively higher to relatively lower elevation (hydraulic gradient). Hydrogeological constraints to base flow velocity and discharge through the aquifer and into the stream channel are highly dependent on the hydraulic conductivity of the rocks and soils through which the water flows. Therefore, base flow and aquifer residency will be different for fractured rock versus glacial till aquifers.

Persistent base flow is most evident in perennial streams where water continually flows from the groundwater table to the stream channel. It is a major contributor to stream flow especially when considered over a time span that includes...
BASE FLOW

dry periods of little to no precipitation. In upstream reaches, a river gains water from base flow (when the water table exists above normal stream stage level) and is referred to as a gaining or effluent stream. Gaining or effluent streams have an over-discharge increase downstream. Otherwise, if the water table is below the stream level, the stream will discharge part of its flow over the reach as a losing (influent) stream causing a groundwater mound below the streambed. Ephemeral streams that only flow in response to input events are considered losing streams. Perennial streams may also have reaches that experience influent conditions. For example, groundwater levels may be lowered by substantial groundwater pumping, reducing base flow locally to a level of becoming a losing stream (lowering the groundwater table and decreasing the hydraulic gradient) while reaches upstream and downstream may continue to be effluent.

Hydrographs are used to show the overall flow conditions of a stream as measured by a stream gauge at a given time or duration and location along the stream. The hydrograph provides a graphical representation of stream discharge versus time to show storm events and seasonal and long-duration flow conditions. Unfortunately, stream gauges and derived plots are unable to separate the direct contributions from runoff or base flow and lesser contributions from direct precipitation into the stream, throughflow, and interflow. Nevertheless, because the latter can be considered negligible contributions to the hydrograph, runoff and base flow can be estimated through various hydrograph separation methods.

Hydrograph separation methods serve to answer the question of what is the proportion of runoff to base flow contributions to the stream hydrograph. As would be expected, these contributions will differ between storm events and seasonal or long-duration hydrographs. The basis for hydrograph separation is due to the premise that surface and groundwater are hydraulically connected, so when physical (volume) and chemical changes occur in one the other will be affected.

During storm events, base flow hydrograph contributions are less than runoff but greater than direct precipitation, throughflow, and interflow. In a seasonal hydrograph, periods of precipitation show jagged rising peaks while dry season periods of the year are represented by smooth curves of discharge recession when the stream is primarily sustained by base flow. This recession in stream flow is a function of water table decline and base flow contribution. It is apparent then that base flow is the most significant contribution during a seasonal flow hydrograph period.

Although hydrograph separation is based primarily on graphical methods, base flow analysis and recession analysis utilize the base flow recession equation. Using this equation and multiple seasonal hydrographs enables the prediction of stream flow (discharge) at a given time during stream recession. The base flow recession equation is $Q = Q_0e^{-at}$, where $Q$ is the discharge at some time, $t$, after the recession of an input event, $Q_0$ is the discharge at the beginning of the recession, $e$ is the exponential function, $a$ is the base flow recession constant, and $t$ is the time interval duration (Watson and Burnett 1995, 145). Solving for the recession constant, $a$, enables prediction of the estimated stream flow discharge at a given time during recession.

Nonetheless, base flow analysis, separation, and recession analysis all attempt to identify groundwater contributions to stream flow by analysis of stream hydrographs (Hall 1968; Tallaksen 1995). The use of chemical and isotopic tracers (e.g. $\text{SO}_4^{2-}$) can also be used to identify the various water sources of stream flow. These geochemical methods are based on water quality and can help to determine
the long-term groundwater storage component from the rainfall runoff component.

**SEE ALSO:** Groundwater; Hydrologic cycle; Streams, gaining and losing

**References**


**Further reading**


Behavioral geography

Neil Argent
University of New England, Australia

Behavioral geography is a sub-branch of human geographical enquiry that became influential during the 1960s and 1970s. It emerged essentially as a reaction to the “quantitative turn” associated with the spatial sciences paradigm of the 1950s and 1960s. Fundamentally, behavioral geography is concerned with producing understandings of how and why people perceive the geographical environment surrounding them in the way they do, and how these perceptions affect their spatial behavior. Behavioral geography, which was largely responsible for introducing behavioralism to human geography, is perhaps best thought of as an approach rather than as a separate subdiscipline, given the breadth of the philosophical perspectives, research foci, methodologies, and epistemologies that were broadly attributable to it.

Behavioral geography: its origins, intellectual development, and main themes

In numerous respects, behavioral geography developed as a reaction against the then dominant paradigm in the geography of spatial science and its accompanying positivist epistemology, ontology, and quantitative methodology. For some human geographers, the spatial science school and the accompanying quantitative revolution had taken human geography into uncomfortable philosophical and methodological territory. Dissatisfaction with the direction in which quantitative spatial science was taking human geography centered on two main concerns. First, many felt disenchanted with its highly abstract and deterministic models of spatial behavior, conceptually underpinned as they were by principles borrowed from Newtonian physics and orthodox neoclassical economics. Gravity models applied to predict human migration flows were regarded by behavioralist geographers as stripping all human agency from what is usually a deeply complex decision-making process and form of spatial behavior. Second, to the extent that such models did recognize human subjectivity at all, this approach was based on the abstract notion, drawn from neoclassical economics, of rational economic man. Inherent in this postulate is the belief that individual decision-making is governed solely by the individual’s desire to pursue his/her own self-interest and to maximize his/her own economic welfare. Critics argued that the importation of the principle of individual welfare maximization from neoclassical economics provided a one-dimensional explanation for actual spatial decision-making and behavior. Unsurprisingly, then, these models could only provide, at best, partial explanations of real-world phenomena.

In a more positive sense, behavioral geographers responded by introducing the epistemology of behavioralism to human geography. Behavioralism is a body of thought that holds that humans are actively thinking beings who use their cognitive powers and abilities to gain, organize, and deploy knowledge to inform their overt behavior (Gold 1980). Behavioralism stands in direct opposition to behaviorism, a
BEHAVIORAL GEOGRAPHY

branch of psychological thought that holds that human behavior is best explained by examining stimulus–response relationships within the brain. Behavioralism became influential in human geography as individuals reacted against the growing dominance of behaviorism in other social science disciplines, such as economics, architecture, sociology, anthropology, and, of course, psychology. One particularly influential epistemological position in the development of behavioral geography was transactional constructivism, a philosophy imported from gestalt psychology (Walmsley and Lewis 1993). Put briefly, in its applications to human geographical research, transactional constructivism holds that humans interact with the environment in order to pursue their own needs and wants, but do so by drawing on past experiences, together with the relevant historical and current social contexts.

In their reactions against the quantitative turn of the 1950s and 1960s, behavioral geographers were concerned to reinstall the human actor as an active, sentient agent to the center of thinking about spatial decision-making, not as an abstract postulate behaving in an a priori purely self-interested rational fashion, but as a fallible social being operating in conditions of imperfect knowledge, bounded rationality, and, at times, conflicting goals. The behavioral school, then, sought to reverse the direction of explanation and theory building from the ideal aggregate spatial pattern produced by notional rational economic actors to the individual social agent. Previous research had already revealed that “economic agents” (e.g., workers, entrepreneurs) rarely made locational decisions based purely on the notion of least effort and highest economic return. This was not to imply that people were always bound to act irrationally but that individuals tended to operate on the basis of imperfect information and “bounded rationality.”

Therefore, an important aspect of the behavioral approach to human geography is the recognition that substantial variability exists in the manner in which individuals perceive and produce mental images; and that person-to-person differences in, inter alia, age, sex, ethnicity, socioeconomic status, physical capacity, and psychological makeup affect spatial decision-making processes, regardless of differences in spatial cognition. A crucial advance produced by the behavioral approach, then, was an appreciation that people respond to the space around them not as it is, in an objectively measured sense, but as they perceive it. A fundamental goal of behavioral geography, then, was to seek to understand how and why people perceive environments the way that they do, and how these perceptions affect actual spatial behavior; however, the philosophical and methodological pathways taken to achieve this goal varied quite substantially within the overall behavioral geography approach. This variance in approach became quite a point of contention and controversy between behavioralists, and also between the various other branches of human geography more broadly (Walmsley and Lewis 1993).

An important element of the behavioral approach to human geography was its explicit attempt to relate observable patterns across space (i.e., macro- or mesoscale phenomena) to processes (i.e., microscale phenomena). That environmental knowledge and spatial behavior are stratified was perhaps first recognized by Kirk in his model of the decision-making environment. Drawing on gestalt psychology, Kirk (1952) distinguished between an objective environment, containing “phenomena” or “facts,” and a behavioral environment – including a perceptual realm – provided by each human agent in his/her own contemplation of “the world of physical facts” (see Figure 1). In other words, Kirk saw the coexistence of two logically separate but
BEHAVIORAL GEOGRAPHY

Figure 1  Kirk’s schema of the behavioral environment and the human decision-maker.

tightly interrelated realms: that of the objective, physical environment, and that of the sentient human agent in active contemplation of this objective world, the former's perceptions of this world influenced in a potential multitude of ways by “a highly sensitive filter of cultured values” (Gold 1980, 36). Although Kirk’s model was later heavily criticized, the basic idea came to influence the subsequent development of behavioral geography profoundly.

In the earliest phases of the behavioralist approach to geography, the insights from Simon’s research – such as “bounded rationality” – were corroborated and extended by geographers working in the fields of natural hazards and agriculture. Kates’s (1962) research on settlement location on floodplains supported the finding that humans make locational decisions on the basis of bounded rationality, in which human agents are aware of only a subset of the entire range of factors that might impinge on the optimal location. Relatedly, Kates (1962) argued that his respondents were largely “satisficers” in that their spatial decision-making reflected a compromise between a range of sociocultural, emotional, strategic, and economic factors, and therefore fell short of the goal of economic optimization. Within a behavioral geography approach, relations between the human individual and her/his surrounding environment are seen as reflexive, with human behavior regarded not simply as bounded, or controlled, by physical environment settings but also as a factor in the modification of that environment. This recognition of human agency as a factor in environmental change sets in train a reflexive chain of action and decision-making in which each environmental modification triggers further changes in the human interpretation of that environment and, hence, the possibility of yet further human intervention.

Wolpert’s (1964) field research with Swedish farmers further supported Kates’s findings concerning the “satisficing” behavior of economic agents, and marked the beginning of his contribution to behavioral geography research, which included notions such as “place utility,” which would become influential in studies of human migration. A related and influential development in this regard was Pred’s (1967) “behavioral matrix” (Figure 2). This concept – and its accompanying diagram – contrasted the “ideal,” self-interested decision-making of the rational economic man of neoclassical economic theory (homo economicus) with that of the everyday human being, with the quantity and quality of information forming one axis in the matrix, and the human decision-making unit’s capacity to process this information forming the other. Although Pred’s matrix reinforced the notion that human locational decision-making rarely, if ever, conforms to the so-called ideal (located in the far bottom right-hand corner of the matrix), he also intended the matrix to be dynamic, revealing how decision-making and spatial behavior changed in response to trial and error, and hopefully evolving toward, if not necessarily achieving, the rational ideal. Pred’s matrix
also incorporated one of the key principles of the behavioral approach: that human spatial decision-making and behavior and the physical and sociocultural milieu in which these occur are mutually influential in a reflexive fashion.

This recognition led to a significant research effort into the cognitive structures and processes by which individuals make sense of real and imagined spaces and places. This became known as the cognitive mapping school. Although many associate cognitive mapping research with the popular notion of mental maps (see Gould and White 1974), the field was much broader than this generalization allowed. Early research efforts homed in on attempts to break into what was widely regarded as the black box of the mind–environment interface so as to better understand how individuals perceive their surroundings, and how such perceptions influenced actual behavior, if at all (Downs 1970). In the words of Thomas Saarinen (1979, 465), an influential figure in behavioral geography’s intellectual development, “The fundamental argument of behavioral geography is that: (i) environmental images exist; (ii) these images can be identified accurately …; and (iii) there is a strong relationship between environmental images and actual behavior.” Downs’s (1970) own schemata drew on Kirk’s dichotomous model but reversed the analytical focus by explicitly considering the dialectical links between an individual’s passive perception of the world via her/his “perceptual
BEHAVIORAL GEOGRAPHY

receptors” (essentially the perceiver’s five senses), and his/her belief system and sociocultural context (value system) and the preconceived or stored image of the real world that the individual already possesses in her/his mind. This cognitive process was then seen to lead to a decision on the most satisfactory (not necessarily the optimal) form of spatial behavior (e.g., to move or stay). Importantly, Downs’s schema contained a feedback loop which enabled the perceptual image of reality stored within the individual’s mind to be updated in the light of the post-decision-making experience. This loop allows for the fact that the process of environmental learning, together with subsequent behavioral responses, is dynamic and iterative.

Downs’s schema helped advance cognitive mapping research conceptually but was still regarded as overly general and lacking in the capacity to generate testable hypotheses. Subsequent attempts sought to correct this flaw, with mixed results (see Kitchin (1996) for a review). Gold’s (1980) schema, later adapted by Golledge and Stimson (1987), incorporated Kirk’s (1952) model of the physical and behavioral environments but sought to explicitly disentangle the perceptual and cognitive processes located within the perceiver’s mind and to explore these in greater detail than heretofore (see Figure 3).

An ambitious attempt to integrate insights from environmental psychology into cognitive mapping schema was proposed by Kitchin (1996; see Figure 4). Incorporating the principles of behavioral geography outlined earlier, Kitchin’s (1996) schema was also informed by five key theoretical propositions: (i) Golledge’s (1993) “cognitive counterparts” theory; (ii) the partial hierarchical structure of spatial knowledge theory, where knowledge of places and spaces is organized superordinately, depending on the individual decision-maker’s use of that information; (iii) Golledge’s (1978) “anchor point” theory, where information about new and unfamiliar places tends to be linked with established spatial knowledge in a hierarchical way, largely as a mnemonic device; (iv) “dual coding,” where cognitive maps are felt to be stored both as images and in propositional form; and (v) that humans use heuristic frameworks to both store and access, broadly termed, environmental knowledge. As Kitchin (1996, 72) put it, “Individuals construct schemata of everyday situations and use them when they encounter those situations. These schemata influence what we expect to see, what we look for and how we respond.” The complexity of Kitchin’s (1996) own schema is apparent in its diagrammatical representation (see Figure 4), attempting as it does to capture the multidimensional and multiscale (temporal as well as spatial) workings of the human mind in its perception of the real world.

True to the original aims of the cognitive behavioralist strain of behavioral geography (also referred to as “analytical behavioralism” by Walmsley and Lewis (1993)), cognitive mapping research focused on three main areas. First, a number of researchers were concerned with the ways in which different groups in society (e.g., children, nonmainstream ethnic groups, the disabled) constructed and organized their own mental maps. The idea here was that a person’s positionality, together with his/her sheer physical capacities, could affect his/her perceptions of, decision-making in, and interactions with, different spatial settings. The second area focused on people’s understandings of urban environments – an increasingly prominent field given the increasing pace of urbanization during the latter half of the twentieth century. This topic was largely pioneered by Lynch’s (1960) well-known work on urban legibility. Lynch’s research subjects produced simplified sketch
maps of select US cities, highlighting, inter alia, largely egocentric landmarks and greatly simplified routeways. Accordingly, Lynch argued that urban residents organized the mental images they held of their lived environment into five main categories: (i) **paths** – the routeways people move along; (ii) **edges** – the built and natural boundaries that demarcate urban regions; (iii) **districts** – large tracts of the city which have an identifiable character; (iv) **nodes** – chiefly major transport hubs, such as railway or bus stations; and (v) **landmarks** – fixed strategic points in the urban landscape which help orient residents in their travels (Lynch 1960).

In contrast to the broadly positivist, cognitive behavioralist approach discussed earlier, a more humanistic strain of behavioral geography also emerged in reaction against the quantitative...
Figure 4  Kitchin’s (1996) integrative conceptual mapping schema (reproduced from SAGE Publications).
BEHAVIORAL GEOGRAPHY

turn. Early humanistic geographical research was strongly influenced by the notion of geosophy introduced to the broader discipline developed by Lowenthal (1961). Lowenthal stressed the role that personal imagination plays in selectively filtering environmental stimuli, thereby pioneering research into landscape perception, a field that would develop into a substantial subdisciplinary area during subsequent decades. Other important and related humanistic research in this vein included Tuan’s, Saarinen’s, and Relph’s work on phenomenology, landscape perception, and place attachment respectively, and Buttimer’s research on values and beliefs in environmental appreciation. Importantly, Tuan (1974) developed the notion of topophilia in human geography, a research theme that greatly advanced the philosophical development of humanistic and, later, cultural geography. Overall, this humanistic approach was very concerned with revealing the deeply personal links between the individual and his/her surrounding environment (both physical and behavioral) as a lived totality rather than the examination of the two-way relationships between people and their environments.

With its fundamental concerns with the affective bonds that people form with environments and environmental features (broadly termed), the humanistic and cognitive-behavioralist approaches eventually parted ways.

Another important contribution to behavioral geography – and to human geography more broadly – was Hägerstrand’s research on migration and the related concept of time–space geography. As with others reacting to the sterile mathematical models of, inter alia, locational behavior, Hägerstrand began from the perspective that an understanding of human spatial interactions must begin with the individual human decision-making unit rather than with the large aggregated populations central to, for instance, gravity models. As with the first passes at cognitive-behavioral geography which recognized that an individual’s perception and cognition of her/his environmental surroundings was conditioned to at least some extent by a range of sociocultural and physical factors internal to that person (e.g., gender, socioeconomic status, ethnicity, physical ability), so Hägerstrand’s time–space geography explicitly accounted for the differential capability, capacity, and authority constraints on individuals’ everyday activities. For example, the diurnal activity spaces of a middle-aged employed male will be very different to that of a single mother with three pre-school-aged children, as would their individual experiences of the physical and behavioral landscapes that they inhabit (Hägerstrand 1970).

As already alluded to in this entry, behavioral geography played an important role in broadening – and deepening – human geography perspectives on human–environment relationships, and in forcing geographers to think more seriously about philosophical and epistemological issues within the discipline. However, this broad approach was not without its detractors.

Critiques and criticisms

At the heart of behavioral geography, at least in its initial phase, was the attempt to reclaim the agency and subjectivity of the individual in the analysis of human spatial behavior. The philosophical and methodological avenues taken by researchers in pursuing this aim varied quite substantially. Eventually, the divergent paths taken by those reacting against the quantitative revolution created schisms within behavioral geography. Those of a more humanistic bent felt that some cognitive behavioralist research, in its quest to understand the links between spatial perception, cognition, and actual behavior, had
actually slipped, philosophically, into behaviorism. Other criticisms homed in on the ongoing positivist nature of much behavioral geography research. Hägerstrand, in spite of his antipathy toward positivist spatial science and his avowedly humanistic stance, had himself earlier employed Monte Carlo random number statistical modeling methods to predict the spatial diffusion of innovations and human migration behavior. Such reductionist approaches to explanation, arguably, denied human agency as much as the spatial science and neoclassical economic schools of thought that behavioral geography was supposed to render redundant. Relatedly, criticisms were raised regarding the reliability and capacity of the various cognitive mapping techniques employed to access and supposedly capture an individual's spatial cognition, as process and/or image (Walmsley and Lewis 1993).

Perhaps more troubling were the stinging critiques and criticisms that came from the broadly labeled “radical” school of human geography. These attacks on behavioral geography were both philosophical and political in tone and content. This “radical” school, inspired chiefly by Marxian political economy perspectives, was also reacting against the mechanistic and reductionist style of reasoning inherent in spatial science, and sought to make human geography more socially and politically relevant to the times. It is important to bear in mind that the late 1960s and early 1970s saw the rise across many Western nations of second-wave feminism, student riots and uprisings, the US civil rights movement, a growing environmental consciousness, and an emergent concern over the inequality (and inequity) between the so-called first and third worlds. In these troubled contexts, behavioral geography’s concerns with recovering the locational decision-making processes and patterns of urban residents were perhaps easy targets for the epithet of “bourgeois” geography. 

Similarly, normative criticisms were made of the early behavioral geography research on natural hazards. In this case, the reliance on microscale psychosocial phenomena such as “bounded rationality” in explaining agrarian communities’ location in flood-prone territory was attacked for overlooking the structural forces at play both within and outside the host society that effectively dictate where the poor can, and cannot, live. Behavioralist accounts of urban residential location came under attack from the same quarter for their perceived voluntarism. That is, behavioral geographers were accused of limiting their analytical focus to demand-side factors (e.g., relative proximity of housing to work locations, shopping, potentially harmonious social groupings) and, therefore, neglecting supply-side influences (e.g., the role of capitalist housing markets, state and nonstate institutions). As an explanation for intraurban residential differentiation and uneven urban development, such research was regarded as potentially tautological and symptomatic of behavioral geography’s want of explanatory depth. Cumulatively, these attacks on the behavioral approach to human geography saw it fall from favor within the broader discipline after the late 1980s, with few explicitly labeling their work as “behavioralist” thereafter.

Behavioral geography now and into the future

In spite of behavioral geography’s relegation to the margins of mainstream human geographical thought and practice during the 1980s, its influence on at least some subdisciplinary fields has endured. The reaction against Marxian structural determinism during the 1990s is probably most strongly associated with the emergence and growing popularity of rival political theories
BEHAVIORAL GEOGRAPHY

such as postmodernism and poststructuralism; however, this shifting tide of academic opinion also created the space for a reappraisal of the value of a behavioral geographical perspective to a host of contemporary social and economic issues of public policy concern. For instance, in the context of the growing popularity of neoliberalist-inspired “self-responsibility” principles across a swathe of policy fields, from welfare to superannuation and pensions, some economic geographers have called for a renewal of a behavioral approach in the subdiscipline. This parallels the revived fortunes of behavioral economics since the early 1990s. A behaviorist perspective on contemporary unemployment, social welfare, and retirement policies is seen as necessary to counter the atomistic ontology of the neoclassical economic principles central to such “reforms,” and to provide much needed perspectives on the often information-poor contexts and “bounded rationality” in which many of the newly responsibilized citizens are forced to make, and act out, potentially life-changing investment decisions. Similarly, behavioral insights are being brought to bear on the critique of liberal, paternalistic public policies and strategies which attempt to inculcate changes in human decision-making across a vast range of fields, from sustainable household energy use to public bathroom etiquette. In select cases, then, cognitive behavioralism’s continued relevance can be seen in the renewal of earlier research agendas (as in economic geography), whereas in other subdisciplinary fields it was never really diminished. A further indication of the ongoing relevance of behavioral geography is its key role in the development of the progressively influential field of evolutionary economic geography. Cognitive mapping theory and methodology also continue to be influential in different parts of the discipline – witness the recent interest in children’s geographies across many subdisciplines – and in human geography’s ongoing interrelationships with other disciplines.

In summary, the advent of a behavioral approach to human geography facilitated a greater engagement with philosophical and epistemological issues, helped forge productive interactions and relationships with cognate disciplines, and laid the conceptual and methodological groundwork for human geographers to engage with contemporary social, environmental, and political issues of public policy relevance, given the increasing importance placed on individual behavior.

SEE ALSO: Cognition and spatial behavior; Dispersal, diffusion, and migration; Human geography; Humanistic geography; Landscape; Marxist geography; Phenomenology; Poststructuralism/poststructural geographies; Radical geography; Time geography and space–time prism

References


Belgium: Société Royale
Belge de Géographie
(SRBG) (Royal Belgian Society of Geography)

Founded: 1876
Location of headquarters: Brussels
Website: www.srbg.be
Membership: 125 (as of 2014)
President: Christian Vandermotten
Contact: cvdmotte@ulb.ac.be

Description and purpose

The society promotes the dissemination of geographical knowledge through the publication of an international scientific journal. It aims also to popularize geography among the general public. The society represents Belgian geography at EUGEO, the association of the European geographical societies.

Journals or major publication series

*Belgeo*: The SRBG published a journal from 1876 to 1961 when the title was changed to *Revue belge de géographie*, emphasizing its scientific character, and published under this title until 1999. In 2000, the journal merged with the *Bulletin de la Société belge d’études géographiques*, another Belgian geographical society, now dissolved, taking on the name *Belgeo*. In 2012, *Belgeo* became a free open access e-journal, published in collaboration with the Belgian National Committee of Geography. *Belgeo* normally publishes papers in English and in French and less often in Dutch. www.belgeo.revues.org

*Hommes et paysages* (Men and landscapes), a series of booklets presenting popularized topical geographical field trips through Belgium. Founded in 1986, it has published 37 titles to date.

Current activities or projects

The main activities of the society are the publication of *Belgeo* and *Hommes et paysages*. The society also organizes occasional field trips in Belgium. The library and the map collection of the society are partly located at the French-speaking Free University of Brussels (Université Libre de Bruxelles) and partly at the library of the Belgian Royal Academy for Science, Letters, and Fine Arts.

Brief history

The society was founded just a few days before King Leopold II opened the Brussels geographical conference, which aimed to promote colonization in Africa and to improve knowledge of the world. At that time the society was strongly supported by entrepreneurs, colonial companies, and the king himself. Most of the more than one thousand members were businessmen, military officers, lawyers, and so forth. Some were naturalists or professors, but very few were true academic geographers: in Belgium, geography has only been taught as a full discipline at university from 1929 onward. The young society
supported, for instance, Henry Morton Stanley’s expeditions to Congo and Adrien de Gerlache’s polar expedition to Antarctica in 1896. Following the development of academic geography after World War II, the society increasingly became a pure scientific society closely associated with academic research, especially at the Free University of Brussels. In 1997, the society was a founding member of EUGEO, the association of the European geographical societies.

Submitted by Christian Vandermotten
Belgium: Société Géographique de Liège (SGLg) (Geographical Society of Liège)

Founded: 1927  
Location of headquarters: Liège  
Website: www.sglg.be  
Membership: 13 (as of December 1, 2013)  
President: Christophe Breuer  
Contact: president@sglg.be

Description and purpose

The Société Géographique de Liège (SGLg) was founded to encourage research in geographical sciences, to promote diffusion of progress in applied and theoretical geography, and to enhance geographic education. It contributes to the formation of a network of scholars, researchers, practitioners, and students mainly in Belgium.

The society offers a discussion forum on the link between societal challenges and progress in geographical sciences. Through the publication of a scientific journal and the organization of conferences, SGLg contributes to the advancement of innovations and high-level international research in geography and related topics.

Journals or major publication series

BSGLg. www.bsglg.be

Current activities or projects

The SGLg creates a network of geographers in Belgium through four axes. First, the society organizes conferences for scholars, practitioners, and students. Lecture topics underline the contribution of geography to societal debates and knowledge progress. Speakers are local and international researchers or practitioners. Second, it publishes an international scientific peer-reviewed journal dedicated to all geographical topics. Publication of BSGLg is supported by the National Fund for Scientific Research (FNRS). Third, it organizes thematic field trips in Belgium or neighboring countries. Fourth, the society maintains contact with members through a website and a newsletter.

Brief history

The society was created in 1927 by Joseph Halkin under the name of Cercle des géographes liégeois (Circle of Liège’s geographers). It was set up by the geography department of the University of Liège (Belgium). The aim of the society was initially to promote the research of Liège’s scholars and to assist networking between geographers. In 1965, the name was changed to the current name. It extended the scope of its activities to the publication of a peer-reviewed scientific journal. This major change coincided with an increase in the number of members and the transformation from a local association to a scientific society articulating local and international research.
Between 1965 and 2013, 61 issues of the *Bulletin de la Société Géographique de Liège* have been published. In 2010 the journal has been renamed *BSGLg* and is now available both in hard and electronic copy.

Furthermore, the society also organizes conferences and field trips. These activities link Belgian geographers with the international scientific community as well as with local initiatives.

The association is facing new challenges and opportunities coming from changes in international scientific communities, exchange of scientific information, increasing specialization, and ambiguity of geographical fields of research.

Submitted by Michel Erpicum
By all accounts, the eponymous Berkeley School of Cultural Geography (hereafter “the Berkeley School”) began when Carl Ortwin Sauer moved to Berkeley’s Department of Geography at the University of California in 1923 from the University of Michigan. Prior to that, Sauer had undertaken graduate work in geology and geography, first at Northwestern University and then the University of Chicago. As Entrikin notes, Sauer’s graduate training “provided him with a model of scientific inquiry that encouraged field observation, historical explanations in terms of origins and a skeptical attitude toward abstract theoretical or philosophical debates as well as toward disciplinary boundaries” (Entrikin 1984, 389). Accordingly, Sauer also drew heavily from German intellectual traditions, especially from the works of Herder and Hettner, which later provided the bedrock for his subsequent ideas on the evolution of cultural landscapes.

While the intellectual development of an academic movement can rarely be attributed to a single person, in the case of the Berkeley School Sauer’s central role is clearly indisputable. Indeed, any discussion of the Berkeley School is inseparable from Sauer himself as a scholar and teacher who had inspired great affection and loyalty from his students. Among the key flag bearers of the Berkeley School are some of Sauer’s well-known students including Fred Kniffen, Wilbur Zelinsky, James Parsons, and David Sopher. Over the course of his career, Sauer had supervised close to forty PhD dissertations in geography ranging from topics such as the cultural geography of the prairies in southwest Louisiana, sea nomads in Southeast Asia, to the ecology of the Serengeti in East Africa.

It is also worth mentioning that Sauer’s childhood growing up in a German Methodist farming community in Midwestern Missouri among the rural setting of prairie plains had invariably shaped his later academic outlook and personal predisposition, in particular, his love for small rural places and archaic cultures. The Berkeley School also differed fundamentally from Midwestern American geography, however, with the latter being focused more on the highly practical aspects of geographical education. But more importantly, the Berkeley School that Sauer led sought to challenge the paradigm of environmental determinism that was dominant in the early twentieth century.

Sauer (1941) in particular offered what he considered to be the “three point underpinning” of geography with the emphasis on the foundational importance of studying the history of geography; that “American geography cannot dissociate itself from the great fields of physical geography”; and that “the human geographer should be well based on the sister discipline of anthropology.” Indeed, Sauer’s admiration for the discipline of anthropology and its importance for geographical inquiry is readily apparent in his proclamation that “Methodologically, anthropology is the most advanced of the social sciences, and one of its best developed methods is that of geographic distribution” (Sauer 1941, 6). Also noteworthy is the “instrumentalist character of Sauer’s epistemology” and his “pragmatist and neo-Kantian intellectual heritage” that could be traced to his

Not surprisingly, an intellectual hallmark of the Berkeley School is its strong advocacy for an interdisciplinary perspective that combines insights from anthropology, history, and natural science as well as a general disdain for positivistic social science. For Sauer, the key theme of human–environment interaction was best captured in his 1925 landmark publication *The Morphology of Landscape*, which sought to put the study of culture right at the core of geography’s intellectual project. In particular, the term “landscape” was proposed by Sauer “to denote the unit concept of geography, to characterize the peculiarly geographic association of facts” (with “area” and “region” used in equivalent terms) (Sauer 1963/1925, 300). Perhaps the most often quoted programmatic statement associated with Sauer has to be his broad invocation of cultural landscape as an entity that “is fashioned from a natural landscape by a culture group” where “Culture is the agent, the natural area is the medium. The cultural landscape the result” (Sauer 1963/1925, 343).

Also germane to the concern of the Berkeley School are ideas relating to cultural ecology – primarily understood in terms of the cultural history of human inhabitation and adaptation of the natural environment that are shaped by prevailing cultural norms and conditions, as well as the diffusion of such culture traits in the formation of cultural regions. As noted earlier, in keeping with Sauer’s upbringing, another salient feature in the works of the Berkeley School is its strong predilection for studying the rural landscapes and associated material artifacts and cultural relics such as barns, fences, animals, plows, houses, and so on. Such fascination with human-induced landscape changes was carried through detailed empirical case studies not only in the United States but also in more exotic foreign settings such as Latin America (the main focus of Sauer’s lifework). Indeed as Price and Lewis (1993, 11) point out: “Given the international orientation of the Berkeley School, ‘culture’ was a shorthand for those foreign people in whose lands geographers muddied their boots.”

To the extent that culture seems to be so loosely defined by Sauer and his colleagues to signify the diverse cultural patterns of (archaic) human communities, such a position also opens itself up to some of the most vehement critiques by scholars such as James Duncan who took the Berkeley School to task for allegedly adopting a “superorganic” notion of culture. In his view, Sauer had erroneously reified culture as a larger-than-life force dictating human behaviors and cultural traits, thereby giving short shrift to the complex inner workings of cultural processes and downplaying the agency of individuals.

In essence, the critics charged that Sauer in his earnest endeavor to provide a corrective to the prevailing paradigm of environmental determinism had replaced it with a form of cultural determinism. Following suit from Duncan’s critique, which essentially challenged the intellectual core of the Berkeley School, a second wave of attacks came from a cohort of young (mainly British) geographers such as Peter Jackson and Denis Cosgrove who questioned the social relevance of the Sauerian-inspired cultural geography, which they saw as being contented with the “static mapping” of esoteric cultural artifacts while urban crisis and social upheavals loomed large in American and British societies in the late 1970s. By the 1980s and 1990s, the sustained critiques of the Berkeley School were joined by a chorus of critical voices from Marxist and postmodernist geographers.
Notwithstanding these critiques, a more sympathetic assessment of the Berkeley School may be found, for example, in Price and Lewis’s (1993) defense of the Sauerian tradition. As they countered, much of the criticisms out there were “off-the-mark” and the term “Berkeley School” is never actually defined by those who attempted to discredit it (Price and Lewis 1993, 6). In addition, Entrikin (1984, 390) also points out that it is not entirely accurate to paint Sauer as quixotic scholar and “a reluctant participant in philosophical debate who entered the arena only when forced to do so.” Such a misinterpretation occludes Sauer and by extension the Berkeley School’s methodological contribution to the systematic study of social science as cultural history.

Nevertheless, by the late twentieth century, the Berkeley School had been ascribed the rather ignominious status as “traditional cultural geography” that is to be overtaken by the innovative waves of “new cultural geography” approaches. As many commentators have noted, however, the Sauerian tradition of cultural geography has left an indelible legacy that continues to have varying impacts on contemporary intellectual developments in the subfield. Most significantly, the enduring themes of cultural ecology, ethics of human–environment interaction and responsible environmental stewardship that Sauer had championed are all the more pertinent today than ever with the resurgence of global environmental concern.

SEE ALSO: Landscape; New cultural geography

References


The etymology of “big data” can be traced to the mid-1990s, when the term was first used to refer to the handling and analysis of massive datasets. It is only since 2008, however, that the term has gained traction, becoming a business and industry buzzword. Like many rapidly emerging concepts, big data have been variously defined, but most commentators agree that they differ from what might be termed “small data” with respect to their traits of volume, velocity, and variety. Traditionally, data have been produced in tightly controlled ways using sampling techniques that limit their scope, temporality, and size. Even very large datasets, such as national censuses, have been restricted to generally 30–40 questions, and are carried out once every 10 years in most countries. Advances in computing hardware and software and in networking have, however, enabled a much wider scope for producing, processing, analyzing, and storing massive amounts of diverse data on a continuous basis. Moreover, big data generation strives to be exhaustive, capturing entire populations or systems (n = all); fine-grained in resolution and uniquely indexical in identification; relational in nature, containing common fields that enable the conjoining of different datasets; and flexible, holding the traits of extensionality (new fields can be easily added) and scalability (can expand in size rapidly). Big data thus consist of huge volumes of diverse, fine-grained, interlocking data produced on a dynamic basis. For example, in 2012, Walmart was generating more than 2.5 petabytes (250 bytes) of data relating to more than 1 million customer transactions every hour, and Facebook was processing 2.5 billion pieces of content (links, comments, etc.), 2.7 billion “Like” actions, and 300 million photo uploads per day. Such big data, its proponents argue, enable new forms of knowledge that produce disruptive innovations with respect to how business is conducted and governance enacted. Given that much big data are georeferenced, they enable new kinds of geographic analysis and insights.

Sources of big data

Big data are produced in three broad ways: through directed, automated, and volunteered systems. Directed systems are controlled by a human operator and include closed-circuit television, spatial video, and LiDAR (light detection and ranging) scans. Automated systems automatically capture data as an inherent function of the technology and include the recording of retail purchases at the point of sale; transactions and interactions across digital networks (e.g., sending emails, Internet banking); the use of digital devices such as mobile phones that record and communicate the history of their own utilization; clickstream data that record navigation through a website or app; measurements from sensors embedded into objects or environments; the scanning of machine-readable objects such as transponders and barcodes; and machine-to-machine interactions across the Internet. Volunteered systems rely on users to gift data through uploads and interactions and include engaging in social
BIG DATA

media (e.g., posting comments, observations, photos to social networking sites such as Facebook) and the crowdsourcing of data wherein users generate data and then contribute them to a common platform (e.g., uploading traces supplied by global positioning systems (GPS) to OpenStreetMap).

Analyzing big data

Given their volume, variety, and velocity, big data present significant analytical challenges to traditional methods, which have been designed to extract insights from scarce and static data. The solution has been the development of a new suite of data analytics that are rooted in research around artificial intelligence and expert systems, and new forms of data visualization and visual analytics, both of which rely on high-powered computing. Data analytics seek to produce machine learning that iteratively evolves an understanding of datasets using computer algorithms, automatically recognizing complex patterns and constructing models that explain and predict such patterns and optimize outcomes. Moreover, since different approaches have their strengths and weaknesses, depending on the type of problem and data, an ensemble approach can be employed that builds multiple solutions using a variety of techniques to model and predict the same phenomena. As such, it becomes possible to apply hundreds of different algorithms to a dataset to ensure that the most illuminating insights are produced. Given the enormous volumes and velocity of big data, visualization and mapping have proven a popular way for both making sense of data and communicating that sense. Visualization methods seek to reveal the structure, pattern, and trends of variables and their interconnections. Tens of thousands of data points can be plotted to reveal a structure that is otherwise hidden (e.g., mapping trends across millions of tweets to see how they vary across people and places) or the real-time dynamics of a phenomenon can be monitored using graphic and spatial interfaces (e.g., the flow of traffic across a city).

Pros and cons of big data

There is good reason for the hype surrounding big data. Big data offer the possibility of shifting from data-scarce to data-rich studies of all aspects of the world from narrow to exhaustive samples; static snapshots to dynamic vistas; coarse aggregations to high resolutions; relatively simple models to complex, sophisticated simulations and predictions. Furthermore, big data consist of both qualitative and quantitative data, most of which are spatially and temporally referenced. Big data provide greater breadth, depth, scale, and timeliness, and are inherently longitudinal in nature. They enable researchers to gain greater insights into various systems. For businesses and government, such data hold the promise of increased productivity, competitiveness, efficiency, effectiveness, utility, sustainability, and securitization, and the potential to better manage organizations, leverage value and produce capital, govern people, and create better places.

Big data are not without negative issues, however. For example, most big data are generated by private corporations such as mobile phone operators, app developers, social media providers, financial institutions, retail chains, and surveillance and security firms, none of which are under any obligation to freely share the data they generate. As such, access to such data is at present limited. There are also concerns as to how clean (error- and gap-free), objective (bias-free), and consistent (few discrepancies) the data are, and as to their veracity and the extent to which they accurately (precision) and
faithfully (fidelity, reliability) represent what they are meant to. Further, big data raise a number of ethical questions concerning the extent to which they facilitate dataveillance (surveillance through data records), infringe on privacy and other human rights, enable social sorting (provide differential access to services), pose security concerns with regards to identity theft, and enable control creep wherein data generated for one purpose are used for another.

Geography and big data

Geographers have long engaged with massive datasets and nascent big data such as remote-sensing imagery and meteorological records, seeking to map and model environmental and climate change. More recently they have pioneered the analysis of volunteered geographic information (see Sui, Elwood, and Goodchild 2013), such as the mapping of georeferenced and locative social media data (e.g., data from Twitter and Foursquare) (Shelton et al. 2014), and begun to model large-scale urban data, such as the constant flow of passengers through a transport system (Batty 2014). They have also been at the forefront of debates concerning the sociospatial implications of big data technologies to urban systems and everyday life. Nonetheless, big data do pose a major challenge to the discipline, namely that, given their volume and velocity, analyzing such data requires a fundamentally different skill set to analyzing traditional forms of data, but as yet few geographers possess such skills.

SEE ALSO: Data quality standards; Metadata; Qualitative data; Quantitative methodologies

References


Further reading

Biodiversity

Thomas W. Gillespie  
University of California, Los Angeles, USA

Biodiversity can be defined as the variation of life forms (genetic, species, taxa) within a given ecosystem, region, or the entire Earth. Terrestrial biodiversity tends to be highest near the equator and generally decreases toward the poles. However, the distribution of biodiversity is complex and is based on a number of unique geographic, environmental, and anthropogenic factors over different spatial and temporal scales. The simplest measure of biodiversity is species richness or the number of species per unit area (i.e., trees per hectare, reptiles per square kilometer). The term “diversity” is more complex; technically it refers to a combination of species richness and weighted abundance or evenness data and is generally quantified as an index (e.g., Simpson index, Shannon index). These indices are used to define alpha diversity, which is the species diversity in one area, community, or ecosystem at the local scale. Beta diversity refers to the amount of turnover in species composition from one site to another or identifies taxa unique to each area, community, or ecosystem. Gamma diversity is the overall level of diversity of a region, continent, or island and this provides the best general overview of patterns of biodiversity at a global spatial scale. Yet measures of biodiversity are fundamentally multidimensional and not easily reduced to a single number. It is clear that the Earth is undergoing an accelerated rate of native ecosystem conversion and degradation as a result of human activity, and there is increased interest in measuring, modeling, and monitoring biodiversity.

This entry has three primary objectives. First, it examines the history of the study of biodiversity. Second, it examines biodiversity gradients associated with latitude, elevation, islands, and empirical patterns of biodiversity in terrestrial and marine environments. Finally, it examines threats to biodiversity and techniques that are being used to protect biodiversity into the next century.

The study of biodiversity

Humans have undertaken classifications of life for thousands of years. The Greeks and the Chinese have the longest historic records of classifications of organisms and their uses. Carl Linnaeus laid the foundations for the modern biological naming scheme by developing binomial nomenclature and greatly improved our understanding of biodiversity. The Linnaean classification systems of life began to describe species in Latin, and documented each species with voucher specimens in a standard and repeatable fashion. This coincided with broad interest in natural history collections in Europe and expeditions around the globe by naturalists that collected and described thousands of species a decade. With the publication of Charles Darwin’s *On the Origin of Species* in 1865, people began to understand the relationships between species and the mechanisms that drive evolution. Alfred Russel Wallace’s definition of zoogeographic regions based on comparative species distribution data was the first attempt to identify global patterns of biodiversity, and clearly identified Wallace as...
the founder of biogeography, a field that is today
to long-term
concerns of the fate of biodiversity.
Between the nineteenth and twentieth cen-
turies there was a rapid increase in collections of
species and estimates of biodiversity. Estimates
of the total biodiversity on Earth have ranged
widely. Recently, Mora et al. (2011) showed that
the higher taxonomic classification of species
such as phylum, class, order, family, and genus
follows a consistent and predictable pattern from
which the total number of species in a taxonomic
group can be estimated. They predict 8.7 million
eukaryotic species (any organism whose cells
contain a nucleus enclosed within membranes)
globally, of which 2.2 million are marine species.
This suggests that, in spite of 250 years of taxo-
nomic classification and over 1.2 million species
already cataloged in a central database, some 86%
of existing species on Earth and 91% of species
in the ocean still await description.
A rapid evolution of genetic techniques and
geographical information systems (GIS) databases
has changed the way biodiversity is understood.
The Tree of Life project has been developed
which shows the evolutionary relationships
between all living organisms and estimates of
their relationships over time. Currently, entire
genome sequencing for a diversity of taxa is
being undertaken at a host of international
institutions. Indeed, some groups, such as fungi,
are now being identified with genetic tools so
rapidly that there is no time to apply scientific
names to them in the classic Linnaean sense.
During this time, there have also been significant
improvements in the geographic analyses of
biodiversity in the twenty-first century. The
locations of species and species ranges have been
included in a number of GIS databases. These
databases contain millions of species records,
with the spatial resolution ranging from an
exact location on Earth to 1 or 2 degree grid
cells. This has revolutionized the way biodi-
versity is studied and understood at multiple
spatial scales (i.e., site, landscape, regional, and
global).

Spatial patterns of biodiversity

Latitudinal gradients

Species distributional patterns associated with
climatic regions have been known since Aris-
totle. As one goes from the poles toward the
equator, species richness generally increases.
However, there are exceptions to this pattern in
the forms of Mediterranean ecosystems, which
have higher diversity than many other ecosys-
tems at lower latitudes (i.e., deserts). Latitudinal
gradients of biodiversity are extremely dynamic
over geologic time scales and have fluctuated
from the poles to the equator over time. During
the Pleistocene (2 588 000 to 11 700 years ago),
vegetation types and their associated levels of
diversity have generally decreased in extent
during glacial maxima and expanded toward the
poles during interglacial periods.

The individual mechanisms behind latitudinal
patterns of biodiversity have been long debated,
but it is clear that the observed patterns are a
result of a combination of factors over different
spatial scales. Gradients of temperature and
precipitation along with area, environmental
stability, and geographic uniqueness explain a
majority of these biodiversity patterns for many
taxa. In the extreme Northern and Southern
hemispheres cold temperature and low rainfall
results in lower floristic and faunal diversity
and a number of widespread species. It should
be remembered that these ecosystems in high
latitude are relatively recent. Indeed, in places
like Canada, tundra, boreal forest, and grassland
ecosystems are relatively new to their current
locations because most of Canada was covered
by ice sheets and glaciers 18,000 years ago. At the middle latitudes or temperate regions, warm temperatures and high rainfall result in high-diversity forest areas and dry summers result in unique Mediterranean ecosystems. However, it is the tropics that clearly have the highest biodiversity in the world at almost all taxonomic levels from plants and insects to mammals. Indeed, biodiversity has flourished in the tropics for over 250 million years. Alfred Wallace was one of the first to suggest that tropical landscapes were so diverse because of the stability of the tropics. The lack of extreme seasonal variation in the tropics has led to intense competition for the resources for life and this has resulted in the evolution of species with unique ecological niches not found in higher latitudes. The tropics also cover a large geographic area of the globe, and relationships between area and biodiversity are well established.

There has been extensive research into the patterns of species and taxonomic groups since the 1980s examining latitudinal gradients in biodiversity and suggesting theories for the pattern. An extensive study by Hillebrand (2004) found that latitudinal gradients in species richness occur in terrestrial, marine, and freshwater ecosystems in both the Southern and the Northern hemispheres. The gradient appears most pronounced for taxa with more species (i.e., beetles, hummingbirds), for larger organisms in marine and terrestrial environments, and at the regional spatial scale. There are a number of exceptions for specific taxonomic groups and geographic regions such as shorebirds, penguins, freshwater zooplankton, and pinnipeds. However, it is clear that biodiversity generally increases toward the equator, and a combination of factors such as climate, area, environmental stability, and geographic uniqueness explains a majority of these biodiversity patterns for most taxa.

Elevational gradients

Elevational or altitudinal gradients in biodiversity have been known for years in mountainous regions. As one goes from sea level to the top of mountain ranges or volcanoes, there is a general decrease in levels of diversity across taxa. However, elevational gradients are complex and range from microtopography gradients in regions that were once glaciated or in low-lying regions like Florida to mountain ranges like the Rockies, the Himalayas, and the Andes. Alexander von Humboldt was one of the first to synthesize this information by collecting plants and comparing the distribution of species on mountain ranges in the Andes and temperate regions (Figure 1). He noticed that temperature drops with increasing elevation, and that vegetation communities and forest structure decrease along elevational gradients. His essays on the geography of plants were based on the then novel idea of studying the distribution of life as affected by varying physical geographic factors. These included the effects of temperature and precipitation on vegetation on mountains, the impact of the locations and elevations of the mountains (Massenerhebung effect), and interactions of the species and associated forest structural characteristics. With the increasing resolution of species distributional data from elevational gradients, it has become clear that vegetation types on mountain ranges do not occur as linear cross-sections and are significantly impacted by orographic processes and precipitation which result in a wetter and a drier side of mountain ranges and volcanoes. This impacts the location of different vegetation types and diversity on either side. In general, species diversity decreases with elevation; however, a number of studies have identified a mid-elevation bulge in diversity. This mid-elevation bulge in diversity is generally explained by overlapping ranges between lowland and montane species, by increases in
precipitation due to orographic processes, and by a decrease in area associated with increasing elevation. The tops of isolated mountains and mountain ranges almost always have fewer species than lowland areas; at the same time, these mountaintops act as sky islands and contain a high proportion of endemic species and taxa from higher latitudes.

A number of environmental factors have been hypothesized for determining relationships between species richness and ecosystem structure found along elevational gradients. Area, climate, and unique niches have all been identified as associated with levels of biodiversity similar to those found for latitudinal gradients. The area of a mountain range is associated with the number of species, and it has been proposed that high-elevation regions have lower species richness than middle-elevation regions because of the smaller area extent at high elevations. There is a well-known decrease in temperature with elevation (lapse rate) of about 3.5°F for an increase in 1000 feet around the globe. This results in colder temperatures on the tops of mountains and lower diversity. Rainfall has been correlated with species richness up to 4000 mm and mountain ranges can receive the highest rainfall at mid-elevations as a result of orographic precipitation. The heterogeneity of mountain ranges in the form of slope, aspect, and elevation has resulted in a number of niches that are not available at higher or lower elevations. Indeed, cloud forests in the tropics provide a myriad of niches for trees, epiphytes, and other species.

Although it is not possible to identify a single mechanism associated with the biodiversity of all taxa, a combination of factors mold current patterns of elevational diversity.

**Figure 1** Alexander von Humboldt's 1805 vegetation map of tropical vegetation gradients.
Islands

It has long been known that islands have lower diversity than continents, and island biodiversity has been of wide interest to biogeographers. Biogeographic metrics, such as area and isolation of islands, have long been associated with species richness at various scales. The first widely used models in island biogeography used area to predict species richness: \( S = cA^z \), equating to \( \log S = \log c + Z \log A \) when logged on both sides, where \( S = \) number of species/diversity, \( A = \) Area, and \( c \) and \( Z \) are constants (MacArthur and Wilson 1967). In this model, area serves as a correlate of other interlinked factors, such as habitat diversity or elevation, and species diversity is envisaged as the result of an equilibrium between colonization and extinction, which are influenced by island area and isolation (MacArthur and Wilson 1967). Recently, Whittaker, Triantis, and Ladle (2008) proposed a model called the general dynamic model of oceanic island biogeography, by adding the age of an island, defined as time since island emergence, as another important predictive variable for native species richness and single-island endemics on oceanic islands.

Compared to continents, islands have significantly higher endemism or unique biodiversity per unit area. This is most pronounced on island continents like Australia, large oceanic islands (which have never been connected to the mainland) such as Hawai‘i, which has 90% plant endemism, and continental fragments that broke off from continents millions of years ago which saved relict species from a previous time period such as Madagascar and New Caledonia.

Terrestrial environments

Theories of the latitudinal and elevation gradients of biodiversity have significantly advanced over the last century, but there have also been advances in the empirical locations and analyses of terrestrial biodiversity. The most accurate ways to collect biogeographic data on biodiversity are intensive field surveys or inventories of species in the field. High-resolution maps of species are available in the United Kingdom where inventories of plants and birds have been undertaken for over a decade at a 10 x 10 km resolution. Plant and animal distribution data are also available at a 50 x 50 km resolution in Europe, Australia, the United States, Canada, and South Africa. These inventories require skilled individuals and a significant amount of time in the field, and can be extremely expensive. Even in relatively well-studied areas, different field data sources can lead to dissimilar or biased maps of species distributions and diversity, and in areas such as the tropics species occurrence and distribution data are relatively coarse and not well collected. However, there has been an increase in high-resolution data and maps for a number of regions and biogeographers are continuing to research ways to map species distributions and diversity. There are a number of large spatial scale geospatial databases that have been used to measure, model, and interpret patterns of biodiversity at a global spatial scale. This has largely been done by mapping the location of a species occurrence within 1 degree grid cells. These biodiversity studies have generally focused on well-known groups such as vascular plants, reptiles, amphibians, birds, and mammals.

Vascular plants, which first evolved around 360 million years ago, now cover 70% of the Earth’s land surface and provide the structural foundation for a diversity of taxa. Native plant species richness is highest near the tropics, with exceptional levels of diversity in northern South America, followed by Southeast Asia and tropical mountain ranges (Figure 2). Other regions of high diversity occur between 30° and 40° latitude including Mediterranean ecosystems, subtropical
forests, and the Himalayas. Indeed the Mediterranean basin contains 11 700 endemic plant species. Plant diversity is clearly lower in hot and cold deserts and in low-lying areas near the poles.

Reptiles arose between 310 million and 320 million years ago and reached their peak in the Mesozoic era. Reptiles consist of approximately 10 000 species, 90% of which are lizards, snakes, or worm lizards. Lizard diversity and endemism are highest in Australia and in arid regions near the equator, and less diverse toward the poles where low temperatures limit their distribution. Snake diversity is very high in tropical forests along the equator, where they have radiated to take advantage of available niches afforded by complex forest structures. Reptiles such as turtles inhabit a wide range of freshwater and saltwater ecosystems, with the highest levels of diversity in Southeast Asia, although some tortoises have dispersed to oceanic islands.

Amphibians are composed of approximately 7000 species and, as their name suggests, live at the interface between terrestrial and aquatic habitats and require freshwater for some part of their life cycle. Most amphibians occur in the tropics and generally have small ranges and poor dispersal ability. The highest diversity of amphibians is in South America and Southeast Asia, and they are notably absent from most oceanic islands (Figure 3).

Currently, there are over 9000 bird species that are found on every landmass in the world. The highest diversity of birds occurs in South America, which was an island in the tropics for millions of years before connecting to Central America (Figure 4). High diversity in South America is also attributed to the extent of tropical forest and elevation gradients in the Andes. Other areas of high bird diversity are the East African rift zone and mountains in Southeast Asia.

There are over 5400 land and marine mammal species that occupy most of the Earth’s habitats. Land mammals have the highest levels of species richness in the Andes, Afromontane regions in

**Figure 2** Global patterns of native vascular plant species richness. (Kreft and Jetz 2007; reproduced by permission of National Academy of Sciences © 2007).

**Figure 3** Global patterns of amphibian species richness. (Jenkins, Pimm, and Joppa 2013; reproduced by permission of National Academy of Sciences © 2013).

**Figure 4** Global patterns of bird species richness. (Jenkins, Pimm, and Joppa 2013; reproduced by permission of National Academy of Sciences © 2013).
Africa, and mountains of southwestern China, peninsular Malaysia, and Borneo (Figure 5). Schipper et al. (2008) compiled data from 1700 experts and mapped 5487 mammals. They found that global macroecological patterns are very different for land and marine species but suggest common mechanisms driving diversity and endemism across systems. They suggest that diversity is similarly driven by energy availability and topographic complexity. Overall, the species richness patterns for land mammals are similar to those found for birds and amphibians; however, species richness does not always overlap with threatened species richness or rare species richness (Figure 6).

Marine environments

It is clear that all terrestrial biodiversity evolved from marine environments. Indeed, many of the oldest life forms are still in existence in marine ecosystems. Patterns of diversity are similar to latitudinal and elevational gradients on land, with diversity decreasing with depth and temperature. The isolation of deep cold regions of the Earth has given rise to some of the most unique ecosystems based on thermal vents. The highest diversity clearly occurs in coral reef ecosystems. Indeed, 25% of all marine biodiversity can be found in coral reef ecosystems that cover an area the size of British Columbia. Coral reefs are a stark contrast to highly productive temperate and polar waters because of the low levels of nutrients. However, the structural complexity of corals has resulted in the highest biodiversity in marine environments. Tittensor et al. (2010) examined global patterns and predictors of species richness across 13 major species groups ranging from zooplankton to marine mammals, and found two major patterns of marine biodiversity (Figure 7). First, coastal species showed maximum diversity in the Western Pacific. Second, oceanic groups consistently peaked across broad mid-latitudinal bands in all oceans. Sea surface temperature was the only environmental predictor highly related to diversity across all 13 taxa.

Biodiversity threats

Gradients of biodiversity have been shaped by millions of years of evolution and environmental changes; however, increases in human populations have significantly changed the levels and patterns of biodiversity. The increase in human densities and distribution in the last 12 000 years has altered levels of biodiversity for megafauna (>100 kg) across all continents and continues to impact smaller taxa. This has resulted in global, regional, and local extinctions of species at a higher rate than extinction events in the past. Indeed, scientists currently define this increase in human population and associated land-cover conversions as a geologic time period, the Anthropocene, which has been in existence for at least 12,000 years before the present. During this time humans have had significant impacts on biodiversity patterns and processes at a global spatial scale.
Figure 6  Total species richness, rare species richness, and threatened species richness for birds, mammals, and amphibians. (Orme et al. 2005; reprinted by permission from Macmillan Publishers Ltd © 2005).

Figure 7  Global species richness and hotspots of marine biodiversity: (a) all taxa (unnormalized); (b) all taxa; (c) coastal taxa; (d) oceanic taxa. (Tittensor et al. 2010; reprinted by permission from Macmillan Publishers Ltd © 2010).

Threat by taxa

These threats are not evenly distributed across geographic regions or taxonomic groups. In particular, some groups, such as plants, freshwater fish, amphibians, and mammals, have been disproportionately impacted by human activities (Figure 8). The Sampled Red List Index for Plants randomly selected a sample of 7000 out of the world’s 380,000 known plant species, and discovered that that one in five plants are
threatened with extinction, and around a third of plants are so poorly known that we do not know if they are threatened or not. Habitat loss due to the conversion of natural habitats for agriculture and livestock grazing is the biggest threat to plants’ survival. Other taxonomic groups that require freshwater, such as freshwater fish and amphibians, have also experienced high rates of global, regional, and local extinction as a result of habitat conversion, pollution, competition from non-native species, and disease. Mammals, our closest relatives, also have a disproportional degree of threat, with one in five species potentially threatened at a global spatial scale. The mammals that are most threatened are generally carnivores that have large body sizes and small geographic ranges, and require large areas of contiguous habitat. These species are significantly impacted by the destruction and fragmentation of habitat, hunting, and human air, land, and water pollution. Although all taxonomic groups are important to preserve, the threats to plants, freshwater fish, amphibians, and mammals are currently critical issues to address in biodiversity research.

**Figure 8** Proportion of taxa threatened, from several major taxonomic groups which are listed as critically endangered (red), endangered (orange), or vulnerable (yellow) on the 2007 IUCN Red List of Threatened Species.

**Threat by regions**

Threats to biodiversity are not evenly distributed across the globe, and there has been an increasing interest in prioritizing regions and landscapes that have a high conservation priority at a global spatial scale. Early research identified biodiversity hotspots or regions around the world with exceptional concentrations of diversity and endemic species that are experiencing an exceptional loss of habitat (>70%) due to human activity (Figure 9). The goal of biodiversity hotspots is to identify regions around the world where conservation priorities should be focused. The rationale is that the funding and resources available for conservation are inadequate to protect all threatened and endangered species in the world and, therefore, it would be more effective to concentrate conservation efforts in areas with the highest levels of biodiversity and endangerment that are experiencing the greatest amount of habitat loss. A total of 34 biodiversity hotspots have been identified which cover only 2.3% of the Earth’s land surface but contain over 50% of the world’s plant species and 42% of all terrestrial vertebrate species. Millions of dollars have been raised to help protect and preserve biodiversity in biodiversity hotspots since 2000.

**Conservation priorities by species**

At a global scale, the International Union for the Conservation of Nature (IUCN), founded in 1948, has led the way in conserving biodiversity by developing the Red List of Threatened Species. Since 1964, the IUCN has been ranking species conservation status as “Extinct,” “Extinct in the Wild,” “Critically Endangered,” “Endangered,” “Vulnerable,” “Near Threatened,” “Least Concern,” and “Data Deficient.” The goal of the IUCN Red List is to set standard criteria to evaluate the extinction risk of species and subspecies, and to educate natural resources
managers, policymakers, and the general public on the urgency of conservation in order to reduce the extinction of species.

Policy and law

There have been a number of global, national, and local conventions and laws established to protect the Earth’s biodiversity. The Convention on Biological Diversity has been signed by 193 nations and hopes to protect at least 17% of the planet’s biodiversity-rich areas in protected areas by 2020. The Convention on Biological Diversity is a multilateral treaty that seeks the conservation of biodiversity, sustainable use of its components, and fair and equitable sharing of benefits arising from biodiversity resources. The convention was opened for signature at the Earth Summit in Rio de Janeiro in 1992. In 2010 the Nagoya Protocol was adopted which provides a transparent legal framework for the effective implementation of the fair and equitable sharing of benefits arising out of the utilization of genetic resources, thereby contributing to the conservation and sustainable use of biodiversity.

Regionally there are a number of nations that have enacted policy to ensure that species do not go extinct. In the United States, the Endangered Species Act of 1973 tries to ensure that no species goes extinct by listing species that are threatened with extinction and by developing survival and recovery plans for species and critical habitats. The ultimate goal is to remove species from the Federal Endangered Species list. As of April 2014, there were 2149 species listed by the US Fish and Wildlife Service. However, there are recovery management plans for most of these species and a number of successful delistings of species such as the bald eagle (Haliaeetus leucocephalus).
leucocephalus), the gray wolf (Canis lupus), and the peregrine falcon (Falco peregrinus). These international and national laws are some of the best tools for protecting biodiversity.

**Protected areas**

Protected areas are widely regarded as one of the most successful measures implemented for the conservation of biodiversity, drawing on traditional and community-based approaches, governance regimes, scientific and traditional knowledge, and contemporary practices of governments and conservation agencies. The IUCN and United Nations Environment Programme (UNEP) and World Conservation Monitoring Centre (WCMC) world database of protected areas identify the extent of protected areas at a global spatial scale. This database contains IUCN categories of protected areas that were assigned by governments for their national territory. Such integrated baseline data are important for determining global conservation, research, and funding priorities. In countries with strong and well-funded natural resource management institutions, such as the United States National Park Service, there are standardized management practices to preserve species and landscape dynamic protocols related to land-cover change. In the tropics, where some of the largest decreed protected areas exist, natural resource agencies have been developing trained individuals to manage landscapes and protected areas, although resources in the tropics are still limited. It is clear that protected areas are an important component to preserve biodiversity for future generations to enjoy.

**Conservation geography**

Humans have had significant impacts on biodiversity patterns and processes at a global spatial scale. However, there are a number of methods that scientists and geographers have and will continue to employ to analyze and preserve biodiversity with the goal of reaching zero extinction. Inventorying and monitoring of species in the field continue to increase and scientists have developed new methods to identify species genetics, presence, and density, such as remote cameras, and microphones to study soundscapes, as an indicator of ecosystem health. Biodiversity and geographic data are also being displayed at higher resolutions online to help inform the public and preserve biodiversity. The Map of Life, for instance, contains data on terrestrial vertebrate and fish species for 150 million point-occurrence records from the Global Biodiversity Information Facility (an intergovernmental warehouse of digitized species data), expert range maps from the IUCN, and regional presence/absence checklists from the World Wildlife Fund, as well as citizen scientists’ databases such as eBird, which contains empirical location and density data on bird species in or near real time (Gewin 2002). Monitoring all biodiversity is not possible, so there has been an increasing interest in species distribution models and remote sensing techniques that can quantify patterns of biodiversity. There has been a rapid evolution of airborne and space-borne satellites, sensors, and techniques to measure, monitor, and manage terrestrial and marine species. There have been significant advances in baseline vegetation mapping and land-cover classifications by combining field data, multisensors, and classification techniques. High-resolution spatial (<0.5 m) and spectral (220 bands) sensors provide important near real-time data on environmental issues that are region- or site-specific, and can be used by natural resource managers to support conservation efforts. There has also been an increase in time-series remote sensing datasets on anthropogenic impacts, such as water use,
Biodiversity

fire, and land surface temperature, that have been used to assess the status of biodiversity. These types of geospatial data inform management practices, policy, and the general public on the status of biodiversity across the globe, and identify gaps where future research and management are needed.

See Also: Disturbance in biogeography; Ecogeography/macroecology (range and body size); Ecological footprint; Environmental policy; Geography of evolution; Mountain biogeography; Nature conservation; Niche theory and models

References


Further Reading

Biogeography: history

Brett R. Riddle
University of Nevada, Las Vegas, USA

Emergence of modern biogeography

Biogeography is the science that seeks to understand spatial patterns of biological diversity and how those patterns have changed through time, at geographic scales ranging from the entire planet to small islands and local habitats (Lomolino et al. 2010). Traditionally biogeographers have dealt with the diversity and distributions of organisms (species, genera, families, etc.), but today they study the geography of diversity across levels of biological organization ranging from genes to communities and ecosystems, often within the rapidly emerging subdisciplines of phylogeography and macroecology. The emerging subdiscipline of conservation biogeography seeks to develop a template and predictive framework for understanding the consequences of human activities—including habitat fragmentation and loss, invasive species, and climate change—on spatial patterns of biological diversity.

While observations regarding geographic variation in the diversity of life can be traced back through the Greek philosophers and explorers, the roots of a modern discipline of biogeography can be traced to observations derived from the expansion of knowledge about global patterns of biological diversity during the eighteenth-century age of exploration. However, few if any scientists would have called themselves biogeographers until late in the twentieth century, with the term “biogeography” probably first used in the late nineteenth century (Ebach and Goujet 2006; Schickhoff, Blumler, and Millington 2014).

Carolus Linnaeus (1707–1778) proposed in 1781 that all species originated on the slopes of a “Paradisical Mountain” and subsequently spread, without evolution, to stations on newly formed continents similar in habitat to respective elevations on the mountain. In 1761 Georges-Louis Leclerc, Comte de Buffon (1707–1788), proposed an alternative scenario with a rudimentary evolutionary component. He envisioned species originating in northwestern Europe and migrating southward as climates cooled. Importantly, he observed that different regions with the same abiotic conditions (e.g., the New World vs Old World tropics) contain different species. Because species dispersing southward from the north would have had to cross inhospitable country to get to their current locations, they would have evolved along the way, providing an explanation for why the tropical biotas of the New and Old Worlds shared few species. While wrong in details, the notion that “environmentally similar but isolated regions have distinct assemblages of mammals and birds” (Lomolino et al. 2010, 20) became the foundational principle of biogeography (Nelson 1978) and is known as Buffon’s Law.

Throughout much of history biogeography has been divided into two largely independent themes that have come down to modern times as historical biogeography and ecological biogeography (Nelson 1978), although Ebach and Goujet (2006) suggest instead that the division is better thought of as regionalization versus dispersal pathways, and that these both incorporate
BIOGEOGRAPHY: HISTORY

historical and ecological components. The historical versus ecological themes are generally traced to an essay in 1820 by Augustin P. de Candolle (1778–1841), who – in the tradition of Buffon’s Law – recognized historical factors that determined the regionalization of biotas (“habitats”), and wrote about ecological factors such as climate and terrain changes that determined the composition of what today we would call local habitats (“stations”) (Nelson 1978). Through much of the twentieth century, historical biogeography was closely associated with other historical disciplines (including evolution, paleontology, and systematics), whereas ecological biogeography had stronger connections to population, community, and ecosystem ecology. The historical versus ecological dichotomy in biogeography is increasingly being replaced by a more integrative biogeography that draws from – and often informs – disciplines in biology including ecology, evolution, and molecular biology; and others in earth sciences including geomorphology, sedimentology, paleontology, and climatology.

Persistent themes and shifting approaches

Four persistent themes in biogeography (Lomolino et al. 2010) can be traced to eighteenth- or nineteenth-century naturalists or geologists, often emerging in association with revolutions in technology and commerce (e.g., those that motivated the age of exploration), or revolutionary thinking in science, for example the hierarchical system of biological classification developed by Linnaeus; ideas on the mutability of species originating with Charles Darwin (1809–1882) and Alfred Russel Wallace (1823–1913), which required an “old” Earth, as provided by geologists such as James Hutton (1726–1797) and Charles Lyell (1797–1875); and the elucidation of an ice age by Jean Louis Rodolphe Agassiz (1807–1873), now known to consist of a series of glacial and interglacial cycles in relatively recent Earth history. The most influential naturalist to advance a modern-looking science of biogeography in the late nineteenth century was Wallace, often considered the father of biogeography. Many of the biogeographic principles that he modernized with sound evolutionary, ecological, climatic, and geological underpinnings are still being investigated today (see Lomolino et al. 2010, box 2.1).

Theme 1: The delineation and classification of terrestrial and marine biogeographic regions

By the mid-nineteenth century, data on the distribution and diversity of life were sufficient to motivate several seminal attempts to ascertain “the most natural primary divisions of the Earth’s surface, taking the amount of similarity or dissimilarity of organized life solely as our guide … supposing them to have been the result of distinct creations” (Sclater 1858, 130). Philip Lutely Sclater (1829–1913) used the distribution of avian taxa to derive essentially the six terrestrial biogeographic regions that were later generalized by Wallace in 1876, incorporating data on mammals and other groups of vertebrates and invertebrates. Sclater also attempted, less successfully, a regionalization of the oceans in 1897—newer approaches have been more widely adopted (Riddle et al. 2011). Intracontinental regions have also long been recognized (e.g., in 1852 Wallace recognized four distinct regions in the South American Amazon bounded by large rivers); recent approaches combine distributional with landform and ecosystem data into units called ecoregions.
Theme 2: reconstructing the history of diversification, distribution, and extinction of lineages and biotas, often in the context of events in Earth history

Application of a grand idea in geology, that land masses “drift” across the face of the Earth over time, to biogeographical questions might have coincided with the publication of *Die Entstehung der Kontinente und Ozeane* (*The Origin of Continents and Oceans*) by Alfred Wegener (1880–1930) in 1915, but because he got a number of details and mechanisms wrong, the revolutionary theory of continental drift was not generally accepted by geologists until it was resurrected as the model of plate tectonics in the 1960s, at which time it rapidly became foundational for students of diversification and distributional dynamics.

Prior to acceptance of the plate tectonic model of continental drift, historical biogeographers would often account for faunal and floral evidence of historical affinities in the distribution of terrestrial organisms between continents and large oceanic islands by proposing geologically implausible land connections spanning ocean basins (e.g., past continents or land bridges between the Gondwanan land masses of South America and Africa across the Atlantic Ocean, and between Africa, Madagascar, and India across the Indian Ocean) – the sorts of long-distance dispersal events envisioned earlier by Buffon and Linnaeus were a dominant theme shared by late nineteenth-century naturalists including Lyell, Darwin, and Wallace. At the same time as evidence was growing in geology for the plate tectonic model of a dynamic Earth, the biogeographer Léon Croizat (1894–1982) was amassing distributional evidence for animals and plants that he would use in his approach called panbiogeography to argue that “earth and life evolve together” (Croizat 1962, 605). Subsequent biogeographers have established that a weakness with panbiogeography was Croizat’s rejection of the formal method of phylogenetic systematics (i.e., cladistics) developed by Willi Hennig (1913–1976) for evaluating relationships of taxa – indeed, both Darwin and Wallace recognized the need for a “natural” system of taxonomy that grouped taxa on the basis of evolutionary relationships, and Hennig’s cladistics provided the method for doing so. In 1966 Lars Brundin (1907–1993) did blend Hennig’s cladistics with Croizat’s panbiogeography into an approach he and Hennig called phylogenetic biogeography, while denouncing the earlier practitioners of dispersalist biogeography with a scathing condemnation of one of its most visible and latest practitioners, Philip Darlington (1904–1983). During the 1970s, historical biogeography was guided by a cladistic-based cadre, including Gareth Nelson, Norman Platnick, and Donn Rosen (1929–1986), into the methodologically rigorous approach called vicariance biogeography, wherein taxon cladograms from codistributed taxa were used to find general relationships between a set of geographic areas, which in turn were utilized to postulate historical relationships resulting from the splitting (vicariance) of those areas by geological events. It is easy to see how this approach was motivated by the acceptance of the plate tectonic model of continental drift. More recent methods have sought to reincorporate dispersal, along with vicariance, into analytically rigorous approaches; some have shifted the emphasis away from the analysis of area histories into a more direct analysis of taxon range evolution (reviewed in Lomolino et al. 2010). Phylogeography has blended phylogenetic and population genetic approaches to develop a biogeography within species and across closely related species. The entire discipline has been transformed by, and continues to include, innovations in the use of molecular genetics to
BIOGEOGRAPHY: HISTORY

reconstruct phylogenetic and population genetic patterns.

Theme 3: Explaining differences in the numbers and types of species between geographic areas, and biotic turnover along gradients of area, isolation, latitude, elevation, and depth

A very pervasive pattern in nature, the species–area relationship, was recognized as soon as eighteenth-century voyages began to visit islands of different sizes, and became a backbone of biodiversity knowledge with a mathematical underpinning early in the twentieth century. A theory with profound influence entered biogeography in the 1960s when Robert H. MacArthur (1930–1972) and Edward O. Wilson used the species–area relationship, along with the species–isolation relationship and evidence of a dynamic turnover of species on islands, to motivate development of the equilibrium theory of island biogeography. This theory was elegant in its simplicity, predicting the species richness of islands (or isolated “habitat islands”) as a dynamic process countering immigration and extinction, predicted by island area and island distance from a larger adjacent source pool of immigrants. Nonrandom patterns of species richness across continents and oceans, as well as islands, are of great interest when biogeography is utilized to generate data for biodiversity conservation issues (e.g., recognition of biodiversity hotspots).

Biogeographers have long recognized that tropical rainforests spanning the equator are particularly rich in species, and that a latitudinal gradient in decreasing species richness stretches from the equator to the poles – a pattern replicated in many terrestrial and marine taxa. Recent attempts at a general explanation include both ecological and evolutionary components (summarized in Lomolino et al. 2010). Species turnover along gradients of elevation were of great interest to Alexander von Humboldt (1769–1859) early and C. Hart Merriam (1855–1942) later in the nineteenth century.

Theme 4: Examining geographic variation in characteristics of individuals and populations within and among closely related species, including trends in morphology, physiology, behavior, genetics, and demography

Along with the aforementioned subdisciplines of phylogeography (the study of genetic architecture within and between closely related populations) and macroecology (a statistical approach to identifying patterns and processes underlying the assembly and structure of biotas), a series of ecogeographic rules have been developed that represent attempts to explain changes in morphological features across geographic gradients and include Allen’s (limbs becoming shorter in colder climates), Bergmann’s (body mass increasing with latitude or colder climates), and Jordan’s (number of vertebrae in marine fish increasing in cooler waters). Predictable morphological size changes in island populations have led to the island rule (gigantism and dwarfism in island populations). Areography attempts to elucidate the structure of geographic ranges, including variation in their sizes, shapes, and overlap.

SEE ALSO: Biodiversity; Dispersal, diffusion, and migration; Ecogeography/macroecology (range and body size); Geography of evolution; Island biogeography; Paleoecology; Phylogeography and landscape genetics; Plate tectonics in biogeography
References


The modern field of biogeography studies the distribution and dynamics of nonhuman species and ecosystems in geographic space and through geological time. It tends to focus on physical geographical variables (such as latitude, altitude, climate, size, and isolation of an area) combined with additional biological factors (such as the location of other organisms and species communities). The subdiscipline of conservation biogeography examines human impacts on natural systems and applies the principles of biogeography to inform efforts toward biodiversity conservation. In this mode of biogeography, analysis tends toward the quantitative and is proudly “scientific” in epistemology.

There also exists a variety of alternative approaches to investigating the geographies of the “bio” that, although neither referred to nor self-identifying as biogeography, can be considered in these terms because of their comparable commitments to mapping and explaining the dynamics and distribution of nonhuman life. These alternative approaches, which are reviewed in this entry, are generally associated with human and environmental geographers, but cannot simply be understood as attempts to integrate scientific modes of biogeography with the insights of social science, though such work does exist.

Biogeography has long accommodated an interdisciplinary strand. Indeed, Alfred Russell Wallace, hailed by many as the founding figure in biogeography, paid as much attention to anthropological patterns as to the distribution of faunal and floral diversity in his classic *The Malay Archipelago*. However, the alternative approaches reviewed here have emerged out of attempts to mobilize new conceptual and methodological resources to address the core concerns of biogeographic inquiry. They have sought to expand the scope of these inquiries, to understand environmental science as a social practice and to bring politics into considerations of diversity and distribution. Sometimes these alternatives have been developed in ignorance of established work in biogeography. At other times they respond to its perceived deficiencies. Occasionally they result from conversations with its key authors.

They have emerged for three (sometimes overlapping) reasons. First, human geographers have rejected the binary distinctions between human and nonhuman life and between natural and social spaces that underpin much traditional biogeographic study. A widespread realization of the scale of human impacts on the diversity and distribution of life, notably conceptualized in the controversial logics of the “Anthropocene” (Steffen *et al.* 2011), questions the adequacy of a field that, more often than not, positions modern humans as alien to rather than inextricably implicated within contemporary ecologies. Second, the notion of the Anthropocene exemplifies a widespread and multipronged critique of the adequacy of the forms of biological knowledge that underpin much traditional biogeography. This critique includes challenges to prevalent ecological models founded on notions of equilibrium and balance on internal (coherence) and external

---

*The International Encyclopedia of Geography.*
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0559
(relevance) grounds as well as questioning the status of biological science as a singular, objective route to truth outside of politics and circuits of economic value. These concerns regarding the provenance and politics of scientific knowledge necessitate new ways of engaging publics in biogeographical decision-making. Third, diverse work in animal and (to a lesser extent) plant geographies has drawn attention to the political and ethical status of those organisms that are the subjects of biogeography – but frequently only feature as populations of species. It identifies the multiple and conflicting ways of categorizing and valuing the nonhuman world, exploring both the cultural politics of relating to and governing nonhumans. This work has also flagged the need to understand the lived experiences of nonhumans to develop biogeographies of plants and animals themselves.

This entry introduces the alternative approaches to biogeography that have been developed in the face of these challenges. It is organized into five sections that explore: (i) expanding the spaces and places of biogeographic inquiry; (ii) knowledge politics; (iii) biopolitics; (iv) biocapital; and (v) animal geographies. It evaluates their intersections with traditional biogeography, their contributions and limitations, and identifies likely future directions.

Expanding the spaces and places of biogeographic inquiry

The challenges listed above have forced a widening (or rethinking) of the “where” and the “why” in both traditional and alternative biogeographies. Traditional biogeographers have begun to look in earnest at life outside of the familiar islands and reserves in which so much historical research has focused, recognizing that ecologists need to be able to say more than feral, weedy, and non-native about the wildlife of the Anthropocene. Increasing attention is thus being given to “novel ecosystems” (Hobbs, Higgs, and Hall 2013) and “anthromes” (Ellis and Ramankutty 2008); these terms describe the unprecedented, anthropogenic ecologies that increasingly characterize the planet and in which the majority of life resides. Archetypal novel ecosystems include agricultural landscapes, cities, and island ecologies subject to large-scale colonization by non-native species. These are understood as hybrid ecologies with unprecedented, networked biogeographies of species provenance in which the vectors of globalization have scrambled evolved distributions. The likely outcomes for biodiversity of such ecological globalization are the subject of much debate. Meanwhile the impending and actual dynamics of ecologies adapting to accelerated climate change in inhabited landscapes create new, nonlinear temporalities of biogeographical change. Conservation biogeography informed by this thinking imagines new territories and forms of connectivity for mobile species as well as ambitious plans to unleash ecological processes to facilitate adaptation. Nature here is social, contingent, and on the move. The past is a guide but not a mirror.

Geologists’ diagnoses of the end of nature have been anticipated for some time by work in environmental geography, environmental history, and the social studies of science (Demeritt 1994). Two of the most influential critics of the modern idea of a nature removed from society are Donna Haraway (2008) and Bruno Latour (2004). Their respective “cyborg” and “hybrid” ontologies of dynamic and mixed forms challenge the coherence and politics of appeals to a singular nature. This work informed Sarah Whatmore’s (2002) hybrid geographies in which nature is understood as the purified outcome of various forms of modern science, rather than a transcendent reality. There is a growing body of work...
examining the implications of these approaches, characterized by multiple immanent natures akin to those found in discussions of novel ecosystems. For example, geographers have concerned themselves with the political ecologies of agriculture and biotechnology – including lab animals, livestock, and genetically modified organisms. Other research has examined the geographies of biosecurity – focusing on the mobile, sometimes pathogenic organisms that unsettle popular geographies of cosmopolitanism and globalization. A further strand of work explores the wildlife and people that flourish in urban, postindustrial, and other marginal habitats far from the purview of traditional conservation. The catalog of the places and spaces relevant to biogeographical inquiry continues to proliferate.

New frontiers for biogeography have also emerged as a result of engagements with developments in the human sciences, biophilosophy, and animal studies. For example, postgenomic research on the human microbiome is encouraging a radical reappraisal of the biological coherence and identity of the human. This work presents the human as a super-organism; a more-than-human ecology in which cells or genetic material with a distinct “human” signature are in the minority. For Donna Haraway (2008), the human, with all of its formerly special properties, should be rethought as a “companion species.” The human thus becomes a multi-species achievement vulnerable to and dependent upon the biological exchange of materials. Dissolving the coherence of the human (and, by implication, nonhuman) subject has uncertain implications for a traditional biogeography that remains, for the most part, organized around the biological species (as genetic material or morphological expression) as the primary units of analysis; and indeed reflects the wider tension between relational and essentialist ontologies (and resulting epistemologies) that currently permeates the academy.

Knowledge politics

The spatial and ontological challenges to the categories nature and human have important epistemological and political implications for biogeography. If natures are multiple, nonlinear, and discordant then what knowledge counts and through what processes might this be decided? These anxieties are compounded by the popular proliferation of knowledge controversies concerning the entanglement of the biological sciences in a range of environmental and health issues. There is thus a degree of ambivalence about the place of science among alternative approaches to biogeography, somewhat at odds with the traditional discipline. This ambivalence is infused with an understanding that certain sciences have been complicit in enabling the environmental crises associated with the Anthropocene. The resulting distrust is balanced by recognition of the roles played by scientists in both diagnosing environmental problems and delivering solutions.

The various challenges to the status of scientific knowledge production unsettle the popular (if frequently caricatured) figure of the objective biologist in possession of a singular and universal route to truth, operating in a domain outside and in advance of politics. They have informed an established literature critical of this figure and that is in search of alternative ways of bringing the biological sciences into democracy (Latour 2004). For example, geographers and others concerned with the practices of the biological sciences have mapped the difference space and place makes to its current and historic conduct (Kohler 2002; Livingstone 2003). They have sought to locate bioscience and other forms of natural
knowledge, focusing on a range of discrete places in which they are conducted. Analysis has focused on archetypal scientific locations including laboratories, hospitals, museums, and field stations as well as more vernacular places like homes, gardens, and the Internet. Furthermore, ethnographic research with scientists has documented how science involves experience, engaging in processes of learning to be affected. Here scientists and other experts calibrate their bodies and instruments to attune to relevant forms and processes. This embodied, affective epistemology has informed a range of investigations of biological science in action. Drawing on work in science studies, geographers of science have thus recognized that singular, universal forms of knowledge are a geographic achievement, involving regular and power-laden knowledge practices through which some privileged actors come to speak for others at a distance, including the world at large.

Others argue that the diagnosis of the Anthropocene (and its associated controversies) demonstrates that the experiments of modern science have escaped their laboratories and have taken over the world. Laboratory science can no longer be understood as a bounded, private, and inconsequential affair. Analysis suggests that we are all now implicated in the conduct of experiments like climatic change and “invasion ecologies,” have valuable knowledge about their drivers and consequences, and ought to be engaged in their research and governance. There has thus been a growing interest among geographers in developing concepts and methods for deliberating with multiple forms of biological expertise. This work builds on long-standing concerns among political ecologists with the colonial roots and exploitative nature of forms of biogeographical knowledge that come to inform conservation. It is most manifest in conservation controversies in the Global South characterized by stark differences in cosmology as well as political and economic power. Collaborative methods take this critique further by recognizing amateurs as experts and by developing constructive methods for working with various forms of natural knowledge (Ellis and Waterton 2005).

The collaborative ethos has driven progressive efforts to shift the point of public engagement in scientific controversies “upstream” and has involved a movement from participative to deliberative and lately to collaborative methods of democracy (Whatmore 2009). In the latter, affected and interested public comprising multiple forms of expertise come together at the advent of a research project or political decision and remain involved throughout its duration and implementation. In their most radical forms these deliberative approaches to democratic biogeography seek experimental, open-ended encounters with social and ecological difference. They are less concerned with testing hypothesis and more interested in allowing the research questions, accepted evidence, and forms of explanation to emerge during the conduct of the research. In part this ethos emerges from the ontological shifts that were outlined in the previous section. Cast off from the certainties of a fixed nature, geographers are learning to live with a nonlinear and interconnected world.

Biopolitics

Further strands of work by geographers have focused more explicitly on the human shaping of the “bio,” attending to the practices and politics through which certain groups come to speak for, benefit from, and ultimately shape the world. Some have turned to and developed the writings of Michel Foucault on biopolitics. Foucault identified two political strategies original to modern forms of government. The first, which he termed “governmentality,” described the rise
of powerful knowledge practices that construct standardized models of normal, rational healthy citizens, and inform technologies that discipline individual adherence to these subjectivities. Social scientists have reworked this concept to describe forms of “environmentality” associated with powerful efforts to create environmental citizens in the interests of conservation.

The second political strategy Foucault identified is the emergence of modern forms of “biopower” where the concern shifts from the behavior of individuals to the management of life at the scale of the (often unruly and unpredictable) population. Foucault highlights how modern “biopolitics” involves productive and destructive processes through which life is made to live or left to die. The concept of biopolitics is now commonplace in the social sciences and informs critical analysis of the deployment of biological knowledge to manage populations to secure human and environmental health.

Foucault was resolutely human in the foci of his analyses of biopower and notoriously ambivalent about animals and the environment as political problems. Post-Foucauldian scholars have developed the concept of biopower to identify and analyze the multitude of modes of nonhuman biopolitics that characterize late modern governance. This work has been influenced by Giorgio Agamben’s writings on the “anthropological machine”: the categorical procedure through which lines are drawn between human, political life (bios), and bare, animal life (zoe). Critics have argued that this approach is rather too totalizing, anthropocentric, and deathly. It presents biopolitics as the control over life and neglects both the generative dimensions of securing life and the ability of life to do otherwise.

Recent scholarship offers a range of livelier and more affirmative approaches to biopolitics that are willing to afford some power to the “bio.” One key source is Donna Haraway (2008), who presents biopolitics as processes of “living with”; modes of companionship figured as unequal, power-laden but nonetheless contingent relations. Her approach culminates in an appeal for a “cosmopolitics” – a concept she takes from Isabelle Stengers – premised on the flourishing of multispecies difference. Haraway’s cosmopolitics resonates with work in geography by Steve Hinchliffe and Sarah Whatmore and their co-researchers on the biopolitics of biosecurity and urban conservation. For Hinchliffe a cosmopolitics for living with aggregate nonhuman populations involves anticipating, nurturing, and managing events that emerge from the circulation of human and nonhuman actors in diverse spatial formations (or topologies). This cosmopolitics is not about rendering the present eternal, but involves a “careful political ecology” that is open to the immanence of nonhuman life.

These writings are part of a growing body of empirical work mapping different modes of nonhuman biopolitics, including agriculture, forestry, fishing, biosecurity, animal welfare, hygiene, hunting, and pet-keeping. Each has its own aims, privileged knowledge practices, and desired norms and subjects. Cutting across these modes, this work identifies a range of common and significant biopolitical practices concerned with understanding and intervening into the character, distribution, and dynamics of nonhuman populations. These include knowledge practices for identifying, classifying, counting, surveying, mapping, and calculating. Databases and models are key here, as are the assemblages of technologies through which governance happens. These knowledge practices inform practical management actions like culling, fencing, translocating, vaccinating, breeding, and planting.

This work traces the performative power of different modes of biological – and biogeographic – knowledge. It examines how systematic efforts to
secure the future of a valued life (both human and nonhuman) at the scale of the population come to shape different worlds, cutting up the flux of life to perform particular ideas of what life should be saved (and where it should live). Donna Haraway (2008) terms this process “ontological choreography.” Annemarie Mol (2002) explores the “ontological politics” that results from tensions between different ways of choreographing life. Drawing on this work Hinchliffe (2007) proposes an understanding of life as multiple, comprising incommensurable ontologies of individuals, species, genes, ecosystems, and processes. Politics emerges at the interface of modes of governance associated with securing versions of each.

Biocapital

In his writings on governmentality and biopolitics Foucault was especially concerned with the rise of neoliberalism. The ascendance of this mode of political economy in the last forty years has troubled a further strand of geography concerned with the growing neoliberalization of forms of environmental and social governance. Often described by the label political ecology, this work offers further alternatives to biogeography. In particular it attends to the processes of commodification through which common and public forms of life are privatized and reconstituted as “bio” or “lively” capital. This work intersects with the poststructuralist work on biopolitics reviewed above but is conceptually more aligned with neo-Marxist critiques of contemporary “biocapitalism.”

The neoliberalization of nature and associated forms of biocapital take diverse forms, but are united by at least six common properties: (i) a central focus on privatization, or the enclosure and assignation of property values to resources that were previously commonly held; (ii) a belief in self-regulating markets as the best and most efficient means of allocating value and tackling problems; (iii) which requires the commodification of everything; (iv) a resistance to what is seen as state interference, except to defend property rights; (v) resulting in both deregulation and reregulation, including the rolling back of the state, shifting governance to supranational scales, voluntary agreements, and empowering NGOs; and (vi) a focus on empowering citizens as consumers to make choices through new markets.

The critique of “neoliberal natures” has been developed in work on the political ecology of nature conservation (and elsewhere), which shows how appeals by biogeographers and others to a pure “nature” tends to naturalize certain unequal forms of valuation. Concern here focused initially on the creation of protected areas and the privatization, commodification, and iniquitous exploitation of the land and wildlife they contained (Buscher et al. 2012). This work has expanded to explore forms of value created through the circulation of spectacular imagery and the commodification of encounters with wildlife. These images and interactions are understood to fetishize certain charismatic species and places, with problematic consequences for the scope, culture, and politics of conservation.

A further strand of work has focused on the commodification of the generative, reproductive potentials of biological life, critically examining the financialization of nature. Focal products for securing the environment include carbon offsets, wetland and other forms of biodiversity banking/prospecting, and the increasingly prevalent frameworks for valuing “ecosystem services” and “natural capital.” Importantly, the fungibility performed by these commodifying practices has significant implications for the resulting geographies of valued life; the
location of a particular species and/or habitat may derive more from the confluence of development projects, associated land deals, and translocations than from ecological histories. Understanding the diversity and distribution of neoliberalized natures thus often requires paying significant attention to political economy as well as the contingencies of evolutionary processes.

Animal geographies

Further alternative approaches to biogeography are offered by animal (and plant) geographies. Since the late 1990s these subdisciplines have been the subject of a great deal of interest, connecting to the wider “animal turn” that has taken place across parts of the science, humanities, and social sciences. Here cultural and environmental geographers have examined individual and aggregations of animals as geographical subjects (Urbanik 2012). They have sought to bring the animal back into geography and in so doing have pluralized and politicized the species- and ecosystem-based approaches of traditional biogeography.

The most developed strand of this work connects with the approaches reviewed above and seeks to attend to “animal spaces” (Philo and Wilbert 2000). These are the various real and imagined geographies that humans create of and for animals. This work includes analyses of the cultural politics, biopolitics, and forms of biocapital caught up in the ordering of animal lives. It also explores the ontological politics associated with different ways of classifying and governing the “bio.” For some, this represents a political effort to address a perceived “speciesism” in the formulation of human geography. The focus here is on the rights and welfare of animals inhabiting some of the most oppressive animal spaces. For example, work has focused on the spaces of agriculture, hunting, and laboratory testing as well as more affirmative relations of domesticity. There has been limited engagement here with the political domain of conservation, which most centrally concerns orthodox biogeography. Although such concerns for animal suffering are more evident in other parts of critical animal studies, these approaches offer an important alternative form of biogeography.

A further strand of this work attends to the “beastly places” (Philo and Wilbert 2000) of animals themselves, conceiving of animals (and to a lesser degree plants) as lively bodies active in the making and sensing of place. Here various researchers have developed approaches to biogeography that are more aligned with those practiced by physical and natural scientists. Conjoining methods from ethnography and ethology, authors have sought to mobilize the expertise of field scientists as well as a wider range of people engaged in a diverse array of interspecies spatial practices – including herding, fishing, bird-watching, surveying, and dog-walking. This approach proposes the existence of comprehensible sensibilities of proximal nonhumans that can be understood across species differences. Working with field scientists and other experts, using their methods and technologies, geographers are beginning to map animals’ geographies – recognizing the varied experiences, mobilities, and topologies that characterize the nonhuman world.

To date these animal geographies have tended to focus on affirmative relations with species that are “big-like-us”; large animals are the central concern of much popular conservation and most amenable to these modes of multispecies ethnography. Field studies often concern novel ecosystems rich with interspecies interactions and adaptations. In keeping with the wider diversification in the taxonomy of animal
studies, there is a growing movement to expand the scope and character of documented inquiries and develop a more systematic comparative approach to this mode of biogeography.

Future prospects

The alternative approaches to biogeography reviewed here encompass a wide range of spaces, concepts, and methods. This is still an emergent field and is perhaps better understood in less bounded terms as an ongoing series of interdisciplinary conversations; but with the diagnosis and popularization of the Anthropocene, these approaches are likely to experience significant interest in the coming years. Likely future developments include a growing interest in a range of spaces that have been neglected by biogeography to date, including, for example, water (especially oceans), the microbiome, and synthetic biology. Methodologically there is much to be gained from the interdisciplinary collaborative approaches between different types of biogeographers that are beginning to emerge. These would harness the potential of new technologies for sensing, visualizing, and modeling the “bio” while facilitating the upstream engagement of affected publics in the formulation of and deliberation over research. Finally, the future prospects for conceptual innovation are rich. This field has benefited greatly from geographers’ ongoing enthusiasms for social theory, most specifically the growing familiarity with and contributions to work in biophilosophy and animal studies.

SEE ALSO: Animal geographies; Anthropocene and planetary boundaries; Biogeography: history; Biopolitics; Biosecurity; Conservation and capitalism; Hybridity; Nature; Political ecology; Socio-nature

References


Further reading


Biogeomorphology

Martin A. Coombes
University of Oxford, UK

Biogeomorphology as an overarching framework

Biogeomorphology is a relatively young area of earth system science concerned with the interactions and feedbacks between living organisms (ecology) and the physical landscape and the processes that shape it (geomorphology). It is very much an overarching framework of environmental study that encompasses aspects of a whole variety of more traditional fields, including physical geography, ecology, hydrology, biogeography, materials science, and engineering. In its broadest sense, biogeomorphology considers two types of interactions: (i) the contributions of organisms to the formation of landforms and the evolution of landscapes over time, and (ii) the influences of landforms and landform-forming processes on the distribution, development, and interaction of organisms (Figure 1). Under this broad framework, various terms have been used to describe more specific areas of study. “Phytogeomorphology” is the study of the influence of topography on plant growth and distribution (Howard and Mitchell 1985), whereas “zoogeomorphology” is specifically concerned with the geomorphic impacts of animals (Butler 1995). Much less studied by geomorphologists, but no less important, are the influences of micro-organisms in Earth surface systems, the study of which has been called “microbial geomorphology” (Viles 2012).

As a term, biogeomorphology is sometimes used synonymously with “ecogeomorphology” (Butler and Hupp 2013), although this has more often been applied to fluvial and wetland environments, typically focusing on plant-geomorphology interactions. Ecohydrology, hydroecology, and ecohydromorphology are related terms sometimes used to describe biogeomorphological research in fluvial environments, although not exclusively. Finally, “geoecology” is also used to describe ecological interactions with the physical Earth, often over much longer timescales and larger space scales. Biogeomorphology ultimately recognizes that landforms, landscapes, and the life forms that inhabit them are linked at various scales and in complex ways. Biogeomorphologists seek to understand this complexity so that they can explain and predict past and future landscape change.

The development of modern biogeomorphology

Biogeomorphology was first formalized in the late 1980s with the suggestion that geomorphologists had largely ignored the roles of organisms in Earth surface processes (Viles 1988). Some important work on the topic certainly existed before this, most notably Charles Darwin’s classic studies on earthworms and soil-forming processes, but the idea that landscapes were largely “abiotic” (i.e., nonliving) had prevailed in most geomorphological research. In contrast, increased recognition of the abundance and
diversity of life in virtually all environments on Earth has led to an appreciation of the often intimate associations between organisms and their physical habitat. Accordingly, study of how living (biotic) and nonliving (abiotic) aspects of the environment interact is now recognized as highly worthwhile.

As a developing field, biogeomorphology had received some early criticism given that organisms are unequally distributed in space and their individual impacts are often small and, therefore, seemingly insignificant for landscape evolution. Such issues of “scale” have remained central to the development of biogeomorphology as a scientific discipline (Dietrich and Perron 2006). Biogeomorphologists continue to grapple with these issues but organic processes are at the very least considered to be locally significant, at the landform scale, and the collective influences of populations and whole communities that form ecosystems may be of regional and even global significance. For example, although microorganisms are individually small they are globally ubiquitous and can dominate entire ecosystems where other forms of life cannot survive. This includes deep ocean sediments, exposed rock surfaces, and dry desert soils. Organic processes have further driven periods of significant environmental change in the geological past, most notably the “great oxygenation event” about 2.4 billion years ago that gave rise to major global-scale changes in climate and geomorphic process regimes.

More recently, there have been sustained efforts to integrate ideas and concepts from ecology more fully with biogeomorphological theory. This is entirely necessary for a subject that straddles the boundary between the physical and life sciences. An important development here has been the emergence of the parallel ecological concept of “physical ecosystem engineering”—the idea that some organisms create and modify physical habitat resources for other species (Cuddington et al. 2007). This is proving important for biogeomorphology because it extends and highlights the relevance of biogeomorphological research for contemporary environmental issues such as ecosystem functioning and biodiversity conservation (see entry “Applied biogeomorphology”). Increasing collaboration between ecologists and geomorphologists offers exciting opportunities for biogeomorphology (and ecosystem engineering) through the generation of new larger-scale and longer-term datasets of populations and communities that can be used in biogeomorphological modeling. Efforts to incorporate broader geomorphological ideas and concepts, such as disturbance regimes, complexity, thresholds, nonlinearity, and self-organization, are also contributing to the theoretical development of...
the field. Links between life and geomorphology over successional and much longer evolutionary timescales are also beginning to be explored. As such, the state of biogeomorphology as a subdiscipline of geomorphology is healthy and productive, and rapidly developing.

Biogeomorphological research

Over the last 25 years biogeomorphology has received sustained research interest. Special issues of key geomorphology journals have been dedicated to the topic, including collections of papers spanning the entire subject and those focused more specifically on vegetation, ecogeomorphology, and zoogeomorphology (Viles and Naylor 2002; Butler and Sawyer 2012; Butler and Hupp 2013; Stoffel, Rice, and Turowski, 2013). This body of work not only demonstrates the incredible breadth of biogeomorphological research, but the vigorousness and increasing relevance of biogeomorphology as an integrated academic and applied field.

Central to biogeomorphology is the recognition that micro-organisms, plants, and animals influence exogenic geomorphological processes (weathering, erosion, transport, and deposition) and thereby contribute to the development of landforms in many ways (Naylor, Viles, and Carter 2002). A broad range of examples exists from different environments, including forested hillslopes, coastal wetlands, desert dune fields, glaciated mountains, the ocean floor, river floodplains, and building stone in towns and cities, to name just some of them (Figure 2). Animals that actively build landforms (called “bioconstructions”) include beavers (i.e., dams), termites (i.e., termitaria), and gophers (i.e., mima mounds). In coastal environments, bioconstructions include coral reefs and worm reefs (Figure 2a), which are particularly important for the biodiversity they support. Organisms also create landforms through the removal of rock and sediment, via processes of “bioweathering” and “bioerosion.” Micro-organisms (including bacteria, algae, and fungi) coat most rock surfaces forming biofilms, and are directly involved in the sculpting of rock through chemical and mechanical means (Figure 2b). These processes, alongside weathering of rock under soils and vegetation, can be very efficient on limestone terrain, where “biokarst” is a term sometimes used to describe a distinctive suite of biogenic landforms. In rivers and streams plants are involved in passive landform development, such as the formation of tufa barrages and log jams, and on land vegetation is actively involved in the weathering of bedrock, through the destructive action of roots, and erosion by the displacement of soil during tree-throw.

Importantly, the activities of biota alter the efficacy of co-occurring geomorphological processes. In this way some seemingly small impacts can have wider-reaching geomorphological (and ecological) consequences (Table 1). By building dams, for example, beavers alter river flows, sediment dynamics, and channel morphology. They also create new habitat (ponds and meadows) and alter in-stream conditions for other species. On land, a whole host of organisms influence the development and movement of soil at scales beyond their immediate impact. For example, populations of digging and burrowing animals such as earthworms, badgers, ants, and termites mix, move, and mound soil (called “bioturbation”). By loosening and breaking up the surface, these and other organisms make sediment available for subsequent entrainment and transport by wind and water, contributing to erosion at the hillslope scale (Figure 2c). In arid and semi-arid scrublands and deserts, disturbance of the soil by animals affects vegetation patterns and ecosystem health by redistributing nutrients.
and moisture. Animals such as crayfish and fishes also move and mix sediment in rivers as well as on the ocean floor, just as burrowing and digging animals do on land. Stream vegetation (“macrophytes”) slows down water and can encourage sediment deposition; dead woody material is also very important in river systems, as it affects flow velocity, channel shape, and flood regimes. On rock, microbial bioweathering and bioerosion create simple soils and release vital nutrients for epilithic plants, which is a particularly important process in primary succession (Figure 2d). They also alter rock properties such as roughness and strength, thereby affecting resistance to subsequent breakdown as well as the types and abundance of other organisms that can grow there. At the coast, for example, microscopic cyanobacteria and algae actively bore into rock alongside a whole host of animals including piddocks and urchins. These organisms alter surface topography and the availability of physical habitat niches, such as holes, crevices, and pools, which are important for other species (Figure 2e). The influence of rock type on the activity of these bioeroders (which are typically more common in softer and calcareous rocks) is an example of geological control on the occurrence and rates of biogeomorphic processes. This recognition – that the strength of association between organisms, landforms, and landscapes will vary across space – is an important one, largely dictated by climatic as well as lithological and edaphic factors.

A different type of biogeomorphological interaction is “bioprotection” or “biostabilization” – when organisms actively or passively reduce weathering and erosion. Coral reefs, saltmarshes, and mangroves provide excellent examples of this, as they buffer storm waves. In this way, these features are not only examples of bioconstruction but they have additional consequences for erosion, the morphological evolution of coastlines, and the protection of people that live along the coast. Vegetation can also have a protective role in the terrestrial and built environment. This includes some lichens that shield rock and stone surfaces from deteriorative agents (rain, salts, thermal fluctuations, and pollutants), cushion plants that stabilize rocky debris on mountainous slopes (Figure 2f), and trees,
Table 1 Some examples of geomorphological and ecological impacts of biogeomorphological processes.

<table>
<thead>
<tr>
<th>Type of interaction</th>
<th>Example</th>
<th>Geomorphological impacts</th>
<th>Ecological impacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioerosion</td>
<td>Rock-boring of intertidal rock by micro-organisms</td>
<td>Active erosion; altered surface topography; weakened rock; rock pool formation</td>
<td>Physical habitat heterogeneity; altered colonization by macro-organisms</td>
</tr>
<tr>
<td>Bioturbation</td>
<td>Burrowing in soil by fossorial mammals</td>
<td>Soil mixing, aeration, and mounding; direct downslope transport; enhanced erosion by wind and water</td>
<td>Litter retention in pits; redistribution of nutrients and moisture; altered vegetation</td>
</tr>
<tr>
<td>Bioconstruction</td>
<td>Calcification by coral polyps</td>
<td>Construction of reef landforms; wave buffering; reduced shoreline erosion</td>
<td>Construction of reef habitat; biodiversity maintenance</td>
</tr>
<tr>
<td>Bioprotection</td>
<td>Vegetation growth on sand dunes</td>
<td>Increased surface stability; reduced near-surface wind velocity; sedimentation; dune stabilization and growth</td>
<td>Stable colonization surface for other plants; provision of refuge and food for animals</td>
</tr>
</tbody>
</table>

which can be particularly good stabilizers of soil and slopes via the mechanical reinforcement of roots (Figure 2g). In drylands, where sediment is particularly susceptible to wind deflation due to a lack of cohesive moisture, vegetation acts as a stabilizing agent by binding the surface (reducing “erodibility”) and limiting the erosive power of wind and water (reducing “erosivity”). These influences can drive the spatial and temporal dynamics of entire landscapes, such as the mobility of vegetated dune fields in response to short- and long-term environmental change (Figure 2h). Even micro-organisms contribute to the stability of desert soils, forming biological soil crusts (“biocrusts”) that are more resistant to erosion than the loose soil underneath. Microbial biocrusts and thicker “biorinds” also form through the weathering and geochemical transformation of rock, particularly in harsh environments like deserts.

Ecologically, trophic interactions (between different levels of a food chain) often have important geomorphic consequences. A simple case would be the removal of stabilizing grasses by grazing cattle, which exposes agricultural soils to agents of erosion, or grazing mollusks on rocky shores that feed on microbial biofilms and actively scrape away layers of rock in the process (another form of bioerosion). More complicated examples occur where the actions of one or a few species cause a cascade of impacts throughout the landscape, sometimes called a “trophic cascade” by ecologists. A famous example of this is the reintroduction of gray wolves (Canis lupus) to Yellowstone National Park, USA, in the mid-1990s after being largely eradicated in the 1920s. After reintroduction, the wolves altered the numbers and feeding behavior of elk, enabling trees such as aspen and willow to reestablish where they had been previously excluded by grazing. As more vegetation established, soil erosion from
valleysides was reduced and floodplains and river channels were stabilized, and their morphology altered. This new vegetated landscape provided habitat for a range of other species, including rabbits, bears, and beavers, which themselves have zoogeomorphic impacts. At the community level of enquiry then, where interactions between species become especially relevant for geomorphology, the value of ecological understanding for biogeomorphology is clear. Equally, knowledge of the ways in which the physical environment responds to organisms and ecological processes (i.e., biogeomorphology) has very clear implications for biodiversity and ecosystem function, by constraining physical disturbance regimes and habitat heterogeneity that ultimately feed-back to ecology.

Importantly, rather than occurring in isolation, these types of interactions are often multiple and involve complex feedbacks (Figure 1). In any location many examples of biogeomorphological processes are likely to be found, even if one may be dominant at a particular point in time. On a hillslope, for example, some species are involved in sediment production and transport, some in trapping and consolidating that sediment, whereas others will have a relatively benign geomorphic role. One major future challenge for biogeomorphology, therefore, is to fully understand the whole spectrum of co-occurring influences of species that make up the unique flora and fauna in any particular location. This requires studies to look beyond the one-way geomorphological impacts of single or a few species, to investigate biogeomorphological processes, feedbacks, and their outcomes at larger spatial scales and over longer temporal scales than have typically been studied so far. This is difficult, both conceptually and practically, as it requires collaboration between geomorphologists and ecologists who may ultimately be interested in different things occurring at different scales. Nevertheless, significant progress has already been made here and, as new methods and techniques are established, this will no doubt prove as fruitful as it is necessary for the development of biogeomorphology as a field of research.

**Applied biogeomorphology**

One aspect of biogeomorphology that has rapidly evolved in the last ten years or so is what might be called “applied biogeomorphology”. Just as its parent discipline of geomorphology has gained much from its application to “real-world” problems, biogeomorphological research is also proving a valuable source of understanding and guidance for practitioners and policymakers. Managing the environment in the face of a rapidly changing world dominated by people necessarily requires holistic thinking; biogeomorphology offers one possible framework for this. The portfolio of applied biogeomorphological research is rapidly expanding and covers a diverse range of different problems across many different environments and scales. The restoration and maintenance of vegetated river channels and floodplains, for example, is widely seen as the key to healthy, functional, and resilient fluvial landscapes. Here, understanding of the complex interactions between hydrology, erosion, sediment transport, vegetation, and stream fauna is increasingly applied in efforts to alleviate flood risk, reduce erosion, and conserve biodiversity.

Other examples of applied biogeomorphological research include the reintroduction of native burrowing mammals to improve the health of degraded dryland soils, using tree rings to reconstruct and date past environmental change (“dendrogeomorphology”), examining the roles of microbes and lower plants in the deterioration of built cultural heritage, predicting the response of vegetated landscapes to...
changes in the concentration of atmospheric carbon dioxide, carbon sequestration via enhanced microbial weathering of rock, the response of coral reefs to ocean acidification, and sustainable farming practices to manage soil erosion. The utility of biogeomorphology extends even to other planets, where Earth analogues are proving useful in the search for evidence of life on Mars.

Back on Earth, the use of biological structures and processes to manage environmental problems is sometimes called “ecological engineering” (distinct from ecosystem engineering as previously described). In practice, many examples of this involve facilitating or enhancing the bioprotective roles of some species and biogeomorphic systems. In low-lying coastal areas of Europe, for example, where it is increasingly impractical to use expensive “hard” defenses to protect coasts from rising sea levels, allowing vegetated wetlands to form is becoming a favored option. Other examples include afforestation to manage soil erosion and building artificial log jams to slow down flood flows in rivers. An important aspect of this kind of approach to engineering is that it has the capacity to support biodiversity conservation at the same time, by creating new habitat. Indeed, applied biogeomorphology in general sits very well within the current paradigm of valuation and provision of “ecosystem services” – those goods and services that humans gain from healthy, functional, and therefore resilient ecosystems, many of which are underpinned by biogeomorphological interactions.

With all this in mind, biogeomorphology shows great promise for contributing to the understanding and management of major environmental problems that currently dominate earth system science. This includes climate change, pollution, land degradation, and biodiversity loss. As a holistic field of study, biogeomorphology remains rich in potential and opportunity for cross-cutting, integrative, and pioneering environmental research of real applied value.

SEE ALSO: Dendrogeomorphology; Earth system science; Ecosystem services; Environmental management; Interdisciplinarity and geography; Soil biology and organisms; Weathering processes and landforms; Zoogeomorphology

References


Geosciences have experienced a large boom through the development and application of absolute dating techniques. In the 1950s, when radiocarbon dating was first implemented, absolute dating techniques favored a paradigm change in our understanding of Earth surface processes and their dynamics. All of a sudden, geomorphic features and processes could be dated exactly, such that the frequency and magnitude of processes could be studied and quantified in more detail and with a temporal perspective. This led to a better understanding of the relationship between processes and climate, environmental changes and human evolution.

Biological techniques have since been used extensively to gather valuable information about past geomorphic activity (Schneuwly-Bollschweiler, Stoffel, and Rudolf-Miklau 2012). These techniques have in common that they are based on the autotrophic conditions of organisms – such as trees, lichens, or coral – which force them to capture energy to live in a specific environment. These organisms thereby form direct witnesses of environmental changes that are taking place in their biotope and will thus record any change in environmental conditions or geomorphic process activity. As a consequence, they can be used for the absolute dating of features and processes based on the growth rate or growth anomaly of the recording organism. This entry briefly summarizes the current state of knowledge with respect to biological dating procedures and focuses on the application of these techniques in the analysis of geomorphic processes. It reviews the scientific basis of each proxy and discusses its utility and applicability.

Dating geomorphic processes with tree rings

Dendrochronology is the science dealing with the analysis and dating of growth patterns in trees and focuses on species living in temperate zones where rings are formed annually. Abrupt changes in precipitation and temperature regimes drive the meristematic activity of plants in these environments, and thus lead to the production of bands of cells during some parts of the year and no activity during others. The tree-ring patterns visible on cross-sections therefore result from the alternation of growing seasons (typically in spring and summer) and dormant seasons (normally in autumn and winter).

Tree growth and ring formation are influenced by internal (e.g., genetics, age) and external (e.g., moisture, temperature, light) factors. This fact has been used since the 1930s to attribute changes in ring widths or wood density to changing climatic conditions (e.g., precipitation, temperature). In addition, trees have been shown to record a large array of other environmental factors, such as the impact and mechanical disturbance induced by geomorphic process activity (e.g., rockfalls, debris flows, floods, snow avalanches, or erosion processes, just to name a few). Material transported by these processes can...
injure trees (Figure 1), tilt their stems, break their crowns or branches, bury stem bases, and/or expose tree roots. Mechanical disturbances are typically recorded in the tree-ring series, therefore allowing the dating of past geomorphic activity with annual and sometimes even seasonal precision. The approach, called dendrogeomorphology, represents one of the most valuable and precise natural archives for the reconstruction and understanding of geomorphic processes and usually covers the last several hundred years (Stoffel et al. 2010).

The methodological basis of dendrogeomorphology has been described by the process–event–response concept (Shroder 1980). Shroder defines the process as any geomorphic agent, such as a debris flow, snow avalanche, or rockfall, whereas the event is a specific disturbance that a geomorphic process can impact on a tree, such as an injury. The response represents the tree’s reaction to the disturbance; this is the evidence which can be used to date the occurrence of past events (see Figure 1). The following paragraphs describe the main responses of trees to geomorphic processes and form a summary of Stoffel et al. (2010).

One of the most common mechanical disturbances in trees is stem wounds. Impacting rocks and debris can partially remove the bark and destroy the meristematic cambium tissue. Trees react upon this disturbance by compartmentalizing the wounded surface to minimize rot and decay. Tree-ring research uses scars and related features to date the occurrence of past geomorphic events with up to seasonal resolution. After cambial damage, some conifers will form tangential rows of traumatic resin duct (TRD). In broadleaved trees, cambial injuries will typically result in anatomical changes such as reduced vessel sizes.

Tilted trees can be found at places where geomorphic events exert a sudden, unidirectional pressure on the stem (e.g., deposition of material by a debris flow), but also where there has been destabilization of the root plate system by landslides or soil erosion. Trees will respond to tilting with the formation of reaction wood, which can be dated macroscopically through the occurrence of eccentric growth rings in the stem.
Sudden growth suppressions can be seen in trees with crown decapitation or in partially buried trees. In the first case, the crown or branch losses will result in a sudden decrease of photosynthetic activity, and consequently in the growth rate. In the case of stem burial, the sediment deposited around the tree will prevent water and/or nutrient supply to the roots, which in turn will lead to a decrease in growth rates. Exposed roots – mainly reflecting sudden or continuous erosion processes – are characterized by noticeable changes in wood anatomy, and can thus be used as well to infer past geomorphic changes.

Geomorphic activity can also eliminate trees in areas affected by intense processes. Survivor trees tend to benefit from less competition and more light, nutrients, and water availability, so that they will show increased growth rates, yet not always immediately after an event. In addition, new plants may germinate on bare surfaces, so that the minimum age of newly germinating plants can be used to estimate the age of new landforms.

Pioneering studies in dendrogeomorphology date back to the 1960s, and focused mainly on soil erosion and floods. The use of tree-ring records has since been expanded considerably to include a large range of processes and to contribute to a better understanding of spatiotemporal patterns of mass movement activity. Stoffel et al. (2010) provide a state-of-the-art overview with illustrative examples showing how tree rings have helped to extend chronologies of processes and to determine climatic and environmental controls of process activity.

**Lichenometry**

Lichenometry is a biological dating technique using lichen growth to assess the age of exposed rock surfaces, based on a presumed specific rate of increase in the radial size of lichen over time. Measuring the diameter of the largest lichen of a species on a rock surface can thus be used to determine the amount of time a rock has been exposed. The method has been applied typically to estimate the minimum age of monuments, but also for the absolute dating of geomorphic deposits. Lichenometry has been used most frequently to date glacial deposits in tundra environments, but has also focused on the analysis of lake level changes, paleofloods, rockfalls, and snow cover persistence.

The approach is based on the principle that lichen growth – an association between an algae and a fungus – represents a site- and species-dependent process (Innes 1985). General growth rates at a specific site are determined using lichen sizes on objects of known age which are then compared to sizes of lichen growing on objects of unknown age. Lichenometry therefore is a very useful and efficient technique (i) when it comes to the analysis of processes beyond the upper limit of forests, and/or (ii) to complement dendrogeomorphic studies suffering from a limited number of old, disturbed trees.

Control points of the reference chronology typically include anthropogenic surfaces such as monuments, buildings, or cemetery headstones. Natural surfaces – dated with tree rings or historical sources – have been used in the past as well (Figure 2a). Because its ecotone covers different geographical environments, and as it can live under severe climatic conditions, *Rhizocarpon* is generally considered the most valuable species in lichenometry (Figure 2b). This species is characterized by its yellow–greenish color, a short thallus with angular areoles, and the easily recognizable circular shape. Traditional measurements of lichens were carried out with a millimetric feeler, which has been replaced more recently by photographic techniques. As a result of asynchrony between lichen sizes within the same geomorphic feature, Beschel (1950)
Figure 2  (a) The age of the different generations of debris-flow deposits can be dated by means of tree-ring analysis. (b) *Rhizocarpon geographicum* growing on a gneissic boulder transported by a flash flood in the Spanish Central System. (c) Coral structures from the Great Barrier Reef (Australia) are good recorders of past changes of marine temperatures, precipitation, and/or paleofloods.

postulated in his pioneering paper that only the largest, first-colonizing specimens should be measured on a specific geomorphic feature. In his work, he explored glacier extension and recent retreat in the European Alps by measuring the size of *Rhizocarpon* sp.
Different procedures have since been suggested to measure lichens in the field, such as the average value of largest individuals growing on a specific surface. More recently, a somewhat different statistical procedure has been suggested where the size distribution, including generalized extreme value (GEV) distribution, is assessed using a Bayesian hierarchical value model (Schneuwly-Bollscheiwer, Stoffel, and Rudolf-Miklau 2012).

Coral dating

Corals are marine invertebrates belonging to the class of Anthozoa of the phylum Cnidaria. They typically live in compact colonies of many identical individual polyps. Corals include the important reef builders that inhabit tropical oceans; these reef builders secrete calcium carbonate to form a hard skeleton (Figure 2c). The corals inhabiting shallow subtropical seas are autotrophic individuals which are adding seasonal layers composed primarily of hard calcium carbonate shells. Similar to trees, corals are very sensitive recorders of environmental changes and their growth will depend chiefly on nutrient availability and on small changes in temperature, rainfall, and water clarity. As a consequence, coral growth can be used as a valuable paleoenvironmental recorder of temperature change in tropical and subtropical oceans, changes in atmospheric circulation, pollution, and the occurrence of large floods. Changes in band density are thereby used to identify and date growth-limiting factors and to interpret changes in coral environments (Le Tissier et al. 1994). Sampling is somewhat more complex than in dendrochronology and lichenometry as it involves the drilling of skeleton calcium cores. Sample preparation consists in the cutting of thick slices from cores and their embedding in resin. The relative dating procedure of corals consists in a visual counting of individual growth bands, whereas the absolute age of corals can be obtained if density variations of annual bands are combined with chemical parameters. Changes in the \(^{18}\text{O}/^{16}\text{O}\) ratio have been used to provide a record of temperature and precipitation changes during the growing season of corals, so as to determine large-scale climatic patterns related to the El Niño Southern Oscillation (ENSO; Rimbu et al. 2003). Large amounts of sediments have been demonstrated to significantly alter the light absorption of certain algae so that the nutrient availability for corals is affected, which in turn has a negative impact on the development of coral structures, which thus allows the detection and precise dating of past, intense paleofloods (Deschamps et al. 2012).

Challenges, applicability, and limitations of biological dating techniques

This contribution has shed light on the potential of biological techniques for the absolute dating of geomorphic and/or climatic events over periods of several centuries and for different physiogeographical environments. The widespread applicability of biological techniques and the high sensitivity of the recorder organisms to environmental changes have improved our understanding of various environmental processes and changes, with previously unequalled spatial and temporal resolution. Table 1 and Figure 3 summarize the main characteristics of the three biological proxies presented here and highlight the main potential and some limitations of these different approaches.

Over the course of the past few decades, the breadth of application of these techniques as well as a multitude of methodological improvements
### Table 1  Main characteristics of the most frequently used biological dating techniques.

<table>
<thead>
<tr>
<th></th>
<th>Tree rings</th>
<th>Lichens</th>
<th>Corals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dating precision</strong></td>
<td>Annual and seasonal</td>
<td>Up to annual</td>
<td>Annual</td>
</tr>
<tr>
<td><strong>Temporal representation</strong></td>
<td>Optimum range: 0–400 years (max. 10 000 years)</td>
<td>Optimum range: 0–500 years (max. 10 000 years)</td>
<td>Up to 10 000 years</td>
</tr>
<tr>
<td><strong>Spatial representation</strong></td>
<td>Temperate climates</td>
<td>Temperate and alpine climates</td>
<td>Shallow and warm oceans</td>
</tr>
<tr>
<td><strong>Applicability in</strong></td>
<td>Earth surface processes, climate change, pollution, and archeology</td>
<td>Earth surface processes, climatic change, pollution, and archeology</td>
<td>Climate change, Earth surface processes, but mainly erosion and floods (indirectly)</td>
</tr>
<tr>
<td><strong>Main limitations</strong></td>
<td>Limited knowledge of growth conditions which operated in the past. Mostly limited to trees where growth is driven by an annual cycle of growing and dormancy.</td>
<td>Limited knowledge of growth conditions which operated in the past. High sensitivity of lichens to environmental conditions (e.g., pollution) may hamper the construction of reliable site-dependent reference chronologies. Unknown timing of colonization. Missing consensus on key methodologies within the research community.</td>
<td>Limited knowledge of growth conditions which operated in the past. Complex sampling may render the approach less reproducible and less comparable. Expensive analytical procedures.</td>
</tr>
<tr>
<td><strong>Main advantages</strong></td>
<td>Good reproducibility and temporal reliability</td>
<td>Ubiquitous presence of lichen in various geographic contexts. Inexpensive methods.</td>
<td>Paleorecords of the state of the oceans.</td>
</tr>
</tbody>
</table>

have clearly expanded the fields and areas for which biological dating can be used. It can thus be expected that more and more precise data will be gathered soon on various geomorphic processes and in areas which have been studied only poorly so far. The main challenge of present-day applications of biological dating techniques may thus reside in a better understanding of the detailed response of different species to environmental changes and in adding complementary analytical techniques (such as trace elements or stable isotopes) to infer the climatic characteristics of these environmental changes.

**SEE ALSO:** Biogeomorphology; Dendrogeomorphology; Geomorphic hazards; Hillslopes; Mass movement processes and landforms; Mountain geomorphology; Radiometric dating/techniques; Weather, extreme; Weathering processes and landforms
Figure 3  Map illustrating the most representative regions where corals, trees, and lichens can be used as proxies to date environmental processes.

References


“Biome” is a unit of biological classification. It refers to large areas with similar vegetation, primarily defined by its physiognomy (particularly its structure, as in herbaceous, shrub, or tree, its leaf type, and its spacing of the largest plants) and to a lesser extent by its associated animals (although the addition of animals was meant to differentiate the biome concept from the vegetation classification of “formation”; see Phytosociology). In size it is the most extensive and in concept the most general within the planetary biosphere (although some systems place it lower than an ecorezone, which would be a biome in other systems). Examples, individually covered elsewhere in this encyclopedia, include extensive areas of broadleaf evergreen trees (tropical rainforest) and herbaceous graminoids and forbs (grasslands). However, the specification of individual biomes, their numbers, and their borders, are a matter of choices made, as with any classification system. Most work on biomes is limited to the terrestrial realm, but the classification of aquatic and marine biomes is ongoing (see Ocean biogeography and Stream ecosystems), although some are being defined as ecoregions (e.g., www.worldwildlife.org/biomes).

The difference in the identification and number of biomes is a matter of splitting or lumping. The most highly differentiated system was that of Whittaker (1975), who listed 36 biome types, including aquatic biomes. The more conservative are in the range of seven terrestrial biomes.

The minimal number of biomes is likely the six defined by the University of California Museum of Paleontology (www.ucmp.berkeley.edu/exhibits/biomes/index.php): Freshwater, Marine, Desert, Forest, Grassland, and Tundra. A number of hierarchical systems have been developed that subdivide biomes into ecoregions or ecosystems. The World Wildlife Fund for Nature (WWF) system is a useful representation of the biome concept (Figure 1; Olson et al. 2001). Its 14 major terrestrial divisions are as follows.

- Tropical and subtropical moist broadleaf forests
- Tropical and subtropical dry broadleaf forests
- Tropical and subtropical coniferous forests
- Temperate broadleaf and mixed forests
- Temperate coniferous forests
- Boreal forests/taiga
- Tropical and subtropical grasslands, savannas, and shrublands
- Temperate grasslands, savannas, and shrublands
- Flooded grasslands and savannas
- Montane grasslands and shrublands
- Tundra
- Mediterranean forests, woodlands, and scrub or sclerophyll forests
- Deserts and xeric shrublands
- Mangrove

The boundaries between biomes are also uncertain because the change from one type of vegetation to another is often a transition in dominance, not an abrupt switch. These transitions, or ecotones, are important biogeographical features and are discussed elsewhere in this encyclopedia.
Biomes are often used as a basis for inferring relations or processes in global biogeography. For example, global patterns of net primary productivity are reported for biomes based on samples within one particular definition; in more informative instances, the range is also given.

Overall, the foremost use of the biome concept is in teaching the basic relations of climate and biogeography. The common approach is to describe the physiognomic types as adaptations to maximize net primary productivity given the climate. For example, the boreal forest has needle-leaf evergreen trees because the evergreen condition maximized productivity in the short growing season while the needle leaf can survive the winters; conversely, tropical rainforests have evergreen trees given that photosynthesis continues year-round, but are broadleaved because such leaves are more efficient at capturing energy while not constrained to survive a season of low temperatures (or water); and between, mid-latitude forests are broadleaved and deciduous where the growing season is long enough that the time and energy spent in producing new leaves in the spring is compensated for by having the more efficient broad leaves for the rest of the season, from which reserves are withdrawn and which are dropped before being killed by
cold. Similar narratives work for a gradient of precipitation. This presentation is usually complemented by noting that the biome limit in the direction of greater resource (equator-ward for the boreal and mid-latitude forests) is more due to competition than to a milder climate (i.e., more resources for productivity). The low-resource abiotic/high-resource biotic limit for the biomes narrative is being modified to emphasize multispecies interactions in all cases, and to recognize the role of microbes.

A recent addition to the biome concept is the microbiome (see Microbial biogeography). While the ubiquity of microbes has long been recognized, the ability to differentiate them through new genomic technology has allowed new classifications and differentiation of this part of the biosphere. While the spatial extents seem small, the tiny size, large numbers, and high diversity (estimates have the number of microbial cells in the human body outnumbering the number of human cells 10:1 and the number of microbial genes outnumbering human genes 100:1; http://commonfund.nih.gov/hmp/overview) mean that microbes could have a biogeography as rich and diverse as all the macrobes on the planet.

Yet another addition to the biome concept is the anthropogenic biome. Ellis and Ramankutty (2008) identified 18 anthropogenic biomes ranging in extent and intensity of use from remote rangelands to urban areas, with only 25% of the terrestrial surface and 10% of terrestrial net primary productivity in nonanthropogenic biomes of forest, sparse tree, and barren land. From these nonanthropogenic biomes to urban biomes, they described gradients in land use and land cover, net primary productivity, carbon emissions, reactive nitrogen, and native and introduced biodiversity. One conclusion is the need to more closely examine human-dominated landscapes (see Urban biogeography and Agricultural environments).

The biome concept is being reinvigorated beyond a teaching tool by the developments in anthropic and microbial biogeography. As these new ideas become common in textbooks in biogeography, the biome concept will be strengthened for the field.

SEE ALSO: Agricultural environments; Biodiversity; Boreal forest ecosystems; Desert ecosystems; Ecoregions; Microbial biogeography; Phytosociology; Polar region ecosystems; Stream ecosystems; Temperate forest ecosystems; Treeline ecotones; Tropical savanna ecosystems; Urban biogeography

References


Website

http://earthobservatory.nasa.gov/Experiments/Biome/
Biopolitics

Claudio Minca

Wageningen University, Netherlands

The notion of biopolitics and the “biopolitical turn” are at the core of many contemporary debates in the social sciences and the humanities, including human geography. Incorporating the concepts of both life and politics, biopolitics is seen to underlie discussions from the global mobility of migrants and refugees to the “war on terror,” from questions concerning biosecurity to questions of public health and the welfare state, from the pervasive tendency to establish new states of exception and emergency regimes, to the battle over new forms of population management. Crucially, biopolitics also facilitates broader debates over the “politics of life” concerning family values, abortion, and euthanasia, as well as reconceptualizations of the human body associated with the emergence of new biotechnologies and biomedical interventions.

Among political scientists, there exist positions supporting a positivist interpretation of these intersections of medicine, life sciences, technoscience, and politics. These readings, fundamentally based on the belief in the possibility of understanding political behavior via biology, are, however, contrasted by a growing body of critical literature that has questioned the basic assumptions and also the effects of considering society and politics in biological terms. In this sense, the critical, poststructuralist literature often proposes “negative” interpretations of biopolitics, viewing it as a modern form of sovereign power to be resisted for its violence and penetration into people’s lives. At the same time, some within this trend of critical thinking advance different forms of “affirmative biopolitics,” where the biopolitical is recognized as an alternative domain of empowerment and a different kind of politics to be embraced.

Despite often being characterized by distinctive disciplinary approaches to the politicization of life, works informed by critical readings of biopolitics share some common key stances across the board. First, they acknowledge the need to reflect on the role of science in forging “the social” from the nineteenth century onward, while at the same time they dismiss the assumption that culture and society can be explained merely in biological terms. They also recognize the crucial relevance of biopolitics in determining contemporary political subjectivities, a reason why they find the positivist approach to (bio)politics untenable and fundamentally deterministic. Second, most of them identify Michel Foucault as the initiator of the contemporary academic interest in biopolitics and of the putative biopolitical turn in the social sciences and the humanities. They also recognize the impact of most recent reinterpretations of biopolitics on the part of Italian philosopher Giorgio Agamben, although sometimes in rather critical terms. Third, they tend to remark on the role played by the Bush doctrine in determining new biopolitical interventions based on biometrics and biosecurity in reaction to the terrorist attacks of 9/11. Finally, they often take into consideration new conceptualizations of the body and of what “the human” is in relation to the emergence of biotechnologies. In this context, geography has largely incorporated biopolitics as a critical form of analytics and has accordingly...
entirely rejected the positivist interpretations proposed by those political scientists relying on the life science model of investigation of the real.

**Bio-geopolitical roots**

A crucial aspect of the definition of biopolitics is that concerning the origins of the term and how these may have affected its present use. Diverse genealogical reconstructions are offered by different fields and in relation to the various national academic contexts. However, a degree of consensus exists that these origins must be traced back to manifestations of biopower inaugurated by the secularized modern state from the nineteenth century onward. Eugenics, for example, is identified as an important and widely diffused form of implementation of such biopower and governance during the past century. This was an approach to life and politics that would tragically feed into the genocidal thanatopolitics exercised by the most biopolitical of all regimes realized, Hitler’s Third Reich.

It has been suggested that the term “biopolitics” was coined by political scientist Rudolf Kjellén (who, perhaps not incidentally, also coined the term “geopolitics”) in the second decade of the twentieth century, in line with the Lebenphilosophie tradition, the life philosophy that has had such an important influence on the German culture of modernity (see Lemke 2010). Other reconstructions, while acknowledging this first use of the term, however, maintain that the origin of “biopolitical thinking” may be traced back to the positivist attempts on the part of nineteenth-century French life scientists to identify prescriptive interpretations of social and political behavior in order to implement policies capable of improving the quality of the body politic. Contemporary Italian political philosophers (see, among others, Esposito 2008), for example, suggest that the deterministic view at the origin of the positivistic approach to biopolitics was already outlined in an embryonic form in the concept of biocratie (biocracy) adopted by August Comte in the 1850s. Biocracy refers to the form of government required to realize a society where a purely biological understanding of life would coincide with social, cultural, and political life. The term “biocracy” reappeared at the very beginning of the 1900s in the writings of hygienist Edward Toulouse, where pedagogy, psychiatry, and mental hygiene were assigned a fundamental role in forging and improving the social body. In the same positivist lineage, biopolitics was also explicitly theorized by Marley Roberts in his 1938 *Bio-politics: An Essay in the Physiology, Pathology and Politics of the Social and Somatic Organism*, an attempt to propose pathology as a key political technology in identifying and governing “normality.”

From the 1960s onward, a growing stream of positivist research on biopolitics in English-speaking academia consolidated around two broadly defined research trajectories, social Darwinism and ethology, both largely based on a purely technical and biological reading of the political (see Esposito 2008). Biopolitics was presented once again, especially using neurology, pharmacology, and molecular biology, with the objective of explaining human nature and political behavior. A parallel trend of biopolitics was registered in France in about the same period. Despite the French roots of positivism and social Darwinism, this trend appeared as more inclined to incorporate humanistic pedagogic interpretations of human nature in the face of dominant capitalism in shaping the social and exploiting the human body. Two exemplary cases of this approach are considered to be Aaron Starobinski’s *La biopolitique: Essai d’interprétation de l’histoire de l’humanité et des civilisations* (Biopolitics: An Essay on the History of
Humanity and Civilization) and Edgar Morin’s *Introduction a la politique dell’homme* (Introduction to a Politics of Man). In the meanwhile, positivist research lines have continued to be developed in the French- and English-speaking traditions, but they have been fundamentally ignored by the work of human geographers and by most debates on biopolitics in the social sciences and the humanities.

Italian philosophers Giorgio Agamben and Giorgio Cavalletti also identify a presumed geographical origin of the term “biopolitics,” implicitly recognized in a trajectory originating from Kjellén and the geographer Frederic Ratzel. The Ratzelian concept of population as a geographical spatial calculative political technology is indeed seen by Cavalletti as typical of the modern organic state, which sees the “geo” and the “bio” as co-implicated in its attempt to politicize living/vital spatialities. This interpretive line is also endorsed by sociologist Thomas Lemke, who claims that the concept of biopolitics emerges from Kjellén’s organicist idea of the state as a “living form” (*Lebensform*). Furthermore, philosopher Roberto Esposito suggests that this process of the naturalization of politics in Kjellén’s historical-cultural framework must be seen in parallel to the influential work of Baron Jakob von Uexküll, who in 1920 put forth the idea that the German state’s vital demands and potential pathology should be compared to the body’s anatomy and physiology.

Kjellén’s belief in the state as a living organism had an impact on many social scientists and life scientists of his time. One of them was the geographer Karl Haushofer, whose elaboration of the same approach suited what were clearly racist applications, in line with the spatial ideology of the Nazi *Lebensraum* (living/vital space). Haushofer played a prominent role in shaping the *Geopolitik* agenda and in founding the popular journal *Zeitschrift für Geopolitik* in 1924, where a 1933 article stated that the “natural science” of the state involves both biopolitics (the historical development) and geopolitics (the spatial development) of the state. This organicist state would soon become a purely racial state under Hitler’s rule, guided as it was by the objective of forging a new German Man together with a biologically homogeneous national community. Hitler’s biocracy was the first state to be mainly organized around the task of eliminating pathological human presences from the state’s organic body in space. The biopolitical and the geographical as understood by the Nazi radical ideologues thus united as one and the same, fused in the millennial horizon of the Third Reich. How these lines of positivist research on biopolitics intersected and influenced each other and how their legacy persists in the work of positivist social scientists today remains to be investigated in detail.

As already spelled out earlier, Foucault is commonly identified as the founding father of present-day critical debates on biopolitics. According to Campbell and Sitze (2013), mainstream anglophone critical social theory, in attempting to develop a theory capable of exploring both “life” and “politics,” has largely relied on Foucault’s pathbreaking work in the 1970s. Starting from his first writing of “bio-politics” in the 1976 essay “Right of Death and Power over Life” (Foucault 2013), Foucault’s analysis of the effects of the new “politics of life” in modern society is essential in reflecting on the biopolitical in geography and on the related biopolitical spatial regimes centered around population government. For Foucault, in fact, a whole system of surveillance, hierarchies, inspections, bookkeeping, and reports was established at the end of the seventeenth century and during the course of the eighteenth century. He identified the system as the emerging of a new technology of power, a “biopolitics.”
In the humanities and the social sciences today, many acknowledge Agamben’s writings (1998) to be among the most significant contributions to a reformulation of Foucauldian biopolitics, and to have brought biopolitics to the frontline of philosophical discussions. While Agamben identifies the enactments of sovereign power with the production of biopolitical regime and of “bare life” in modern politics, a number of further developments in biopolitical thought have taken the debate to different shores. This array of interventions, well beyond the field of philosophy, engage directly or indirectly with the Foucauldian take on the biopolitical. One example is the rather controversial appropriation of his work by Michael Hardt and Antonio Negri in *Empire* (2000), where they envisage a new stage of capitalism – a biocapitalism characterized by the coming together of economy and politics, (bio)creativity, and the body. Other equally important interventions critically reread Foucault (and often also Agamben) in light of recent technological and political developments, proposing new interpretations of contemporary biopower. These include, among many others, the work of Bruno Latour on the social and political role of life sciences, Jacques Rancière on policy and biopolitics, Donna Haraway on the postmodern body, Paul Rabinow on the new forms of anthropos, Nikolas Rose (2006) on “the politics of life itself,” and Achille Mbembe on the production of necropolitics. The most recent translations of Roberto Esposito’s essays on the biopolitical paradigm of immunity have also been incorporated in international debates preoccupied with new forms of governance of life. The result of these and many other similar interventions is what some identify as a biopolitical turn in diverse fields including, but not limited to, geography, political science, anthropology, and sociology.

**Geography and the biopolitical**

After decades of influence of Foucault’s work combined with the other scholarly contributions mentioned earlier, human geography “turned” biopolitical, especially starting from the 2000s. On the one hand, the discipline has seen profound theoretical work done in relation to the politics of life. Agamben’s work has been subjected to in-depth geographical readings, while present-day “vital geographies,” ethics, and the “value of life” and death have become topics of prime concern. Reflections are also made, for example, with regards to the ethico-aesthetics of life and bioethics in Félix Guattari, and with reference to Jacques Derrida’s speculations on “the beast and the sovereign,” while in other cases bare life is analyzed according to gendered biopolitics and its spatial projections. On the other hand, in an effort to understand how past and current geographical practices may imply a biopolitical dimension, and what the implications of these practices may be, geographers mostly incorporate biopolitics as a form of analytics in examining manifestations of biopower and their spatial representations and practices (for a recent review of geographers’ work on biopolitics, see Minca 2015).

For some, geography’s biopolitical turn has opened the door to historical investigations on the relationship between the “bio” and the “geo.” Two routes have been pursued in this respect: one focused on the Nazi bio-geopolitical technologies, largely influenced by Agamben’s work; the other, perhaps much more significant in terms of its impact on the discipline, focused on the biopolitics of colonialism and its post-colonial consequences. Several empirical studies have detailed postcolonial legacies and some of their biopolitical effects in relation to the “colonial present,” including the consequences of the new technologically driven ways of
waging war and, for instance, the biopolitics of “food provisioning” in colonial and postcolonial contexts.

Other major contemporary topics, especially in political geography, have also been examined in a biopolitical framework. Key areas of investigation are the new geographies of biosecurity, the “war on terror,” as well as the interventions in government of people’s (im)mobilities. Both conceptual and empirical work thus deal with the interrelated proliferation of spaces of exception, carceral geographies, new forms of bordering and border management, and relevant ideas of hospitality and/or encampment. The questions of sovereign exception and bare life are often evoked in these studies, while the specific issues include migration control, detention, airport biometrics, and the management of refugees, asylum-seekers, and stateless individuals.

In this vein, camp studies in geography have investigated sites ranging from Guantanamo to Auschwitz, to Khmer Rouge Cambodia, to the present-day camps in Lebanon and Palestine. Many of these studies draw directly or indirectly, and sometimes rather critically, on Agamben’s speculations on the camp as a biopolitical space of exception. Meanwhile, vast and rich material available on domicile citizenship, biological citizenship, and even postcitizenship has emerged recently in geographical fora. These redefinitions of the principles of citizenship and of the right to mobility have been part of the broader interdisciplinary efforts to understand the biopolitical regimes imposed on the growing masses of people fleeing situations of danger and emergency.

Other closely related research clusters around population governance and the shaping of the body politic. For one, the changing welfare state and the neoliberal state’s biopolitical operations have prompted timely scholarly interventions. Emphasis is placed on the following aspects: the state’s medical and family “care and management”; affect in biopolitics linked to the effects of neoliberal forms of biocapitalism and of global interventions against real or presumed terrorist threats; the calculative rationalities of state institutions that traditionally underpinned the constitution of specific “population regimes”; the workings of governmentality according to the new readings of Foucault; the management of emergency and its accompanying spatial strategies and conceptualizations of the individual body. Research on the effects of neoliberalism has furthermore interrogated “laboring bodies” and liberal “care” as expressed by biological and financial citizenship. In a similar context, global health and biosecurity have gained much attention, with the biopolitical understandings of public health, disease, and epidemic management playing an important role. These are complemented by work related to the state mentioned earlier. An equally crucial theme for geographers’ biopolitical analyses concerns the spaces of humanitarian intervention in relation to broader geopolitical scenarios and imaginations.

Finally, the even broader “politics of life” has entered debates in geography, largely based on reinterpretations of the Foucauldian/Agambenian legacy while incorporating the work of other scholars. This approach covers the realms of the more-than-human geographies and animal geographies, as well as biopolitical readings of major environmental issues, thus inaugurating a brave new field of investigation for nature geographies. Some work in geography has, for example, offered a biopolitical perspective on past and contemporary policies of “rewilding” nature. Animal subjectivity is analyzed in relation to livestock breeding. In conservation biology, biopolitics also inspires research and reflection on environmental risk and safety, and the realization of “biopolitical environmental citizens.” The new ecologies of global capitalism
bringing up questions of resilience, sustainable development, and climate change have also been studied through biopolitical lenses. Discussions expand into considering the Anthropocene and necropolitics. All these represent the rich heterogeneity of interventions enlivening today’s key geographical fora. In light of rapid biotechnological change and of the expansion of geopolitical interventionism on a global scale, the biopolitical turn seems to have a growing and enduring impact on the discipline.

SEE ALSO: Biopolitics; Geopolitics; Political geography

References


Further reading

The concept of biopower is no longer new to geography and other fields in the humanities and social sciences. The precise intellectual origins of this term are difficult to identify and the ideas that it encapsulate have not remained static. However, it has been the work of the poststructural social theorist Michel Foucault that has spurred the adoption and popularization of this concept as an important tool for the analysis of a variety of social spaces, right from colonial policies to present-day slaughterhouses. Foucault’s (2008, 138) elaboration of biopower as the power “to foster life or disallow it to the point of death” has formed the basis of a substantial corpus of empirical and theoretical scholarship that has used, interrogated, and modified this concept for the examination of the complexities of flows of power in contemporary society. The significance of Foucault’s work on biopower for its subsequent interpretations and applications in the social sciences and humanities thus makes familiarity with his elucidation of the concept essential.

The power over life

Foucault, at the most basic level, was concerned with the subtle and diverse ways in which power functions in society and penetrates all layers of social interaction. His signature contribution was to theorize forms and mechanisms of power that are not repressive or obviously negative and harmful. Foucault develops these ideas on power through the examination of historical texts on the nature of state rule, judicial systems, sexuality, town planning, and economic theory in Europe, offering a schema that distinguishes between sovereign and non-sovereign forms of power.

Foucault introduces the term “biopower” in the first volume of History of Sexuality (2008) where he describes changes in mechanisms of state rule in eighteenth-century Western Europe and develops an expanded understanding of the forms and functioning of power. Traditionally, power is associated with sovereignty – of a monarch, a political group, a class, a caste, a gender, maybe even a species – and therefore, with repression and dominion. This kind of sovereign power is focused on “deduction and death,” and is founded on a “right of seizure: of things, time, bodies, and ultimately life itself” (Foucault 2008, 89, 136).

Until around the end of the sixteenth century, Foucault observes, state rule was largely based on the monarch’s right to take the lives of subjects – or the corollary – to let them live. The principal objective of rule was to protect the interests of the sovereign. Such rule used mainly negative or deductive mechanisms of power such as rules, taxes, and punishments, and was often enacted by force. In this period, it was not uncommon to see public displays of violent punishment, especially directed at those who threatened the sovereign.

In the seventeenth and eighteenth centuries, Foucault notes that texts on state rule indicate a shift in modes of rule. Instead of rule based on the threat of death and violence, this period sees a transition to nonrepressive
mechanisms of rule that were focused on regulating and fostering life. Foucault uses the umbrella term “biopower” – the “power over life” (2008, 139) – to refer to these. In his telling, biopower is a positive form of power that is aimed at “generating forces, making them grow, and ordering them, rather than … impeding them, making them submit, or destroying them” (Foucault 2008, 136). To Foucault, biopower does not replace sovereign power; rather sovereignty, discipline, and biopower are a triad of distinct but copresent forms of power that imbue social relations in the contemporary world.

Biopower, as theorized by Foucault, is exercised along two principal axes: an anatamopolitical axis that targets individuals through mechanisms of discipline to make them well-functioning and suitable members of society, and a biopolitical axis that targets the characteristics of populations and collectives in efforts to enhance overall welfare and security. These two axes come together as biopower which seeks to regulate populations and reduce the randomness inherent in life processes in order to achieve a hypothetical optimal stability. Biopower is thus simultaneously individualizing (through its disciplinary mechanisms) as well as totalizing (through its biopolitical mechanisms).

Foucault further writes that biopower is associated with pastoral rationalities and practices of care and flourishing which are aimed at managing life and life processes at the level of the population or some other kind of grouping. Biopolitical mechanisms try to shape life so as to suit certain ends, and when life is not considered valuable or suitable, it is merely “let die.” The exercise of biopower involves decentralization (Gordon 1991). Biopower is not wielded by just one authority, but operates through and at multiple layers of society, at microscopic levels and through unexpected actors, including families, nonprofit organizations, hospitals, private companies, and the academia. Biopower is dispersed; it is wielded at many levels, through many rationalities, and through many creative techniques; this complicates the task of identifying individuals (or specific social groups) as sources or sites of power, or as heroes or villains.

Foucault develops these observations on biopower in the History of Sexuality (2008), lectures at the Collège de France (2009), and various interviews. While he does not use the exact term “biopower” very often, his works on governmentality and pastoral power are recognized as building on and taking forward the conceptualization of biopower and non-sovereign modalities of power more broadly (Dean 2010). In this extensive corpus, which has been taken forward by scholars from a range of disciplines, it is possible to identify four themes that are central to the understanding of biopower: population, governmentality, the welfare–violence nexus, and subjectification.

**Population**

The idea of population lies at the heart of Foucauldian theorization of biopower. Biopower functions to enhance the wellbeing and fostering of the subject-objects of power at the scale of the population or collective. In other words, biopower is exercised in the name of the wellbeing of all, in the name of the “health, prosperity and happiness of the population” (Dean 2010, 27). Biopolitical interventions might work on and through individuals, but it is the population that is the main target – the main subject-object – of biopolitical power. In the context of biopower, the population is not merely a collection of individuals; it is an entity in and of itself, which has meaning, value, and significance beyond what can be attributed to the
individuals that constitute it. While Foucault’s initial use of the term “population” was tied to territory and the nation-state, it has since been expanded by social theorists Paul Rabinow and Nikolas Rose (2006) with the use of the term “biosocial collectivity” to signify human groups that are tied together through links such as gender, race, or other forms of shared identity. Biosocial collectivity has also been used by geographers to refer to groupings that include nonhuman life, whether nonhuman animals or biodiversity as a whole (Holloway et al. 2009; Srinivasan 2014).

Governmentality

Governmentality refers to a way of thinking and acting that is critical to the exercise of biopower. Governmentality incorporates the belief that the regulation and management of the population and its various traits and processes “is not only necessary but also possible” (Dean 2010, 44). In his work on governmentality and security, Foucault shows that in contrast to sovereign power mechanisms that lay down totalitarian laws or prohibitions and attempt to suppress or forbid undesirable activities, biopower works alongside and makes use of existing biological and socioeconomic rhythms and patterns in the population that is the subject-object of power (Foucault 2009). Here, management is achieved by permitting certain levels of unwanted phenomena in a population but keeping them below what would adversely impact the population as a whole. It is this “ordering” of life forces (2008, 136) that Foucault refers to as governmentality. Governmentality enhances the efficiency and subtlety of power by using the very subject-object of power and its rhythms as aids for its functioning.

The work of other scholars, such as Dean (2010), has contributed to the elaboration of the meaning and scope of the idea of governmentality. It is now widely acknowledged that governmentality can be seen in a range of social domains, whether colonialism, international development, or environmentalism (Agrawal 2005; Li 2007). This literature explains that while government is a calculated, goal-directed activity, the norms and ends of regulation are taken for granted; the main concern of government is the means, the how of regulation, which is understood as a matter for technical investigation rather than normative reflection. A certain utopianism marks governmentality in that any failure to achieve the objectives of government does not lead to the disruption of the goals, mentalities, and techniques of government, but instead often leads to calls to reinforce and strengthen the same governmental interventions (Foucault 1977).

Governmental processes are also enmeshed with the production of particular truths, norms, and knowledges which enable power to be exercised without enforcement and threat. In particular, the production of statistics and the circumscription of the population which is the object of intervention have been observed as being key to the exercise of governmental power (Agrawal 2005; Foucault 2009; Holloway et al. 2009). Governmentality thus goes alongside the emergence and deployment of “truths” and “norms” – and the production of formal knowledge bodies which are indispensable for the efficient exercise of power (Gordon 1991). Most importantly, the governmentality literature suggests that any analysis of power must necessarily consider “how we govern … [and the] techniques and other means employed” in the process of government (Dean 2010, 18, 27). This attention to the specific technologies and rationalities of power is a particularly important and distinctive feature of the analytical toolkit offered by biopower.
Welfare and violence

Another vital characteristic of biopower is the intertwining of discourses and practices of harm and care, or as Gordon (1991, 12) puts it, in the exercise of biopower, “welfare is conjoined to exploitation.” Foucault explains that biopower’s focus on the good life, care, and wellbeing does not mean that violence and killing are removed from the equation. What alters is the justification. In the context of state rule in eighteenth-century Western Europe, Foucault observes that while the violent rule typical of sovereign power was validated in the name of the sovereign, biopolitical technologies were explained as necessary for the wellbeing of the population, as on “behalf of the existence of everyone … in the name of life necessity” (2008, 136). In other words, the entanglement of welfare and violence in biopower can be understood in terms of trade-offs between individuals and populations: biopower intervenes harmfully on and governs individuals in the name of universal wellbeing. As Dean (2010, 170) points out, an example of this entanglement can be seen in China’s one-child policy, which targets “imprudent parents and their potential offspring.”

In doing this, biopower treats individuals as not only expendable and of lesser importance but also entities to be managed and intervened upon so that they contribute to collective development and wellbeing (Srinivasan 2014). This means that under biopolitical regimes, the ethical and political significance of individuals is reduced; individuals instead become “the instrument, relay, or condition for obtaining something at the level of the population” (Foucault 2009, 42). In fact, biopower often functions so that those who “resist the regulation of the population” are excluded and subject to techniques of repression such as “exile, death and punishment” (Foucault 2009, 44).

Subjectification

The notion of subjectification or self-governance is key to Foucauldian scholarship on biopolitical and governmental power. Subjectification underlies the functioning of biopolitical power, that is, it is the motor of biopower. Subjectification refers to the process by which individual entities internalize various truth discourses about individual and population/collective wellbeing, and self-govern, that is, work upon themselves, in accordance with these discourses. For instance, in History of Sexuality, Foucault describes how norms on sexual behaviors work so as to encourage a particular form of stable family life. These stable families in turn were necessary for the smooth functioning of society as a whole. Biopower therefore works through the internalization of norms, by inculcating subjectivities (or elements of subjectivities) rather than through the external imposition of rules that dictate dos and don’ts (Rabinow and Rose 2006).

The impacts of self-governance are not always positive for the individual and could even be detrimental to their wellbeing. However, subjectification, that is, the internalization of certain norms, has the effect that the outcomes and consequences of self-governance for the individual are not explicitly evaluated. Subjectification thus enhances the efficiency of the operation of power by making acceptable and incontestable even negative impacts of self-governance and by reducing the need for externally imposed interventions. Whereas sovereign power is exercised through force and imposition, biopolitical power is exercised by means of norms and discourses of care and flourishing that make individuals self-governing subjects. Therefore, techniques of biopower, being underpinned by subjectification, even if not strictly harmless in their impacts, are more subtle and less likely to invite resistance in comparison with techniques.
of sovereign power that depend upon external force and domination.

**Biopower: applications and debates**

Biopower and the related concept of governmentality have been used in geography and the wider social sciences to examine a range of topics and theoretical concerns. This includes analyses of neoliberal policies and programs, contemporary geopolitics, food security, urban ecologies, ethical consumption, developments in biomedicine, food production, colonialism, development practice, and environmental regulation, to name a few domains in which the biopolitical frame has been usefully applied to understand the complex manners in which human life is and has been managed in the present and recent past.

The concept of biopower was originally developed for the analysis of power in human relationships, and the rich vein of scholarship that has taken forward its theorization has also focused principally on the human domain. However, recent years have seen increasing application of the biopolitical schema for more-than-human inquiry; that is, the examination of human–environment and human–animal relations. This literature takes as its starting point the focus of the biopolitical framework on the ambiguities of efforts to foster and regulate both human and nonhuman life (Agrawal 2005; Demeritt 2001), and demonstrates its relevance to a variety of more-than-human domains, from spaces of care such as animal welfare (Srinivasan 2013) and biodiversity conservation (Srinivasan 2014; Chrulew 2011) to spaces of exploitation such as livestock agriculture (Holloway et al. 2009). In much of this literature, the intertwining of harm and care in human interactions with nonhuman life-forms appears as a central issue for critical reflection.

While this body of work has offered crucial insights regarding the play of biopower in human–animal/environment interactions, questions have been raised about the limits of the biopolitical framework when it comes to understanding and analyzing more-than-human spaces. This has particularly been in respect to the processes of normalization, subject-formation, and self-regulation that are understood as driving biopower in intrahuman relations, but that are less plausible in human–nonhuman relations (Demeritt 2001). In addressing this problem, geographers have put forward and elaborated on the concepts of relational (Holloway et al. 2009) and agential subjectification (Srinivasan 2013, 2014), wherein subjectification is seen in those humans who act for, or on behalf of, animals. The writings have therefore been productive in theoretically developing biopower as an analytical tool by highlighting the different sites and relational manners in which subjectification can take place.

The concept of biopower has sparked much theoretical debate over the years. One significant line of argument has been around the twin flavors of harm and care running through Foucault’s work on biopower which have led various scholars to interpret biopolitics as either affirmative (Rabinow and Rose 2006) or harmful (Agamben 1998). For example, Rabinow and Rose (2006) argue that when biopolitics is scaled down to the molecular and genetic levels, its manifestations as self-management through gene therapy or prenatal testing do not have the same pernicious implications as biopolitics in the form of eugenics does. By contrast, Giorgio Agamben (1998) associates biopolitics with racism, oppression, and exclusionary violence. Others, such as Roberto Esposito (2008), present a view of biopolitics as double-edged and marked by
an immunitary logic in that various activities directed at fostering particular forms of life can also cause various kinds of harm to the same or other forms of life. Yet others have argued that the harmful aspects of biopower are embedded in its future-orientated elements in that current wellbeing and interests are sacrificed for the sake of the future (Hannah 2011).

The rich and wide-ranging corpus of scholarship on biopower in geography and the broader humanities and social sciences speaks to the value and relevance of this concept as an analytical tool. There is a multiplicity of empirical domains within which this concept has been deployed and of manners in which it has been theorized and reworked. However, a key theme that cuts across these diverse applications relates to the importance of questioning what is considered to be normal–natural–right and of attending to the subtle and creative flows of power in unexpected and surprising social spaces. It is this insight carried by the framework of biopower that has and will continue to stimulate social science and geographical engagement with it.

SEE ALSO: Biopolitics; Discourse; Environmental discourse; Environmentality and green governmentality; Poststructuralism/poststructural geographies

References


Bioprospecting and biopiracy

Benjamin D. Neimark
Lancaster University, UK

History of natural products drug discovery

Bioprospecting is defined as the collection, research, and commercialization of biodiversity for new medicines and other useful natural products (perfumes, cosmetics, agro-chemicals, and functional foods). Although generally thought of as a modern practice, the discovery of natural products is part of a millennia-long quest for medicines. Evidence of this can be observed in ancient texts, including the Chinese Materia Medica, the Vedas, or the Brahmamic of Hinduism, and many influential Arabic medical works (Sneader 2005).

In Europe, imperial and colonial projects of economic botany helped to establish global botanical gardens where collections of valuable herbaria and details of their medical uses were housed and used for new discoveries. The reported use by Indians in Quito, Peru, of cinchona bark (Cinchona officinalis L.) as a medicinal decoction for shivering and cold spells led European scientists in the early 1700s to adopt the use of quinine as a treatment for malaria. Another important discovery from this period included the ipecacuanha root (Cephaelis ipecacuanha), which was used to treat amoebic dysentery.

Centuries later, drug discovery was a major focus of the US government when seeking treatment for wounded soldiers during World War II. Government research immediately began to collaborate with a number of large pharmaceutical companies, such as Abbott, Pfizer, Merck, and Squibb, for the commercial production of penicillin. The large-scale research cooperation opened a gateway for the US National Institutes of Health (NIH) to promote joint public/private funding for enhanced treatment from natural products.

Bioprospecting and regulation

Contemporary bioprospecting was introduced to the conservation and development community at the 1992 Earth Summit at Rio as a model of sustainable development intervention. Proponents held that the discovery of natural drug products could provide the incentive, and more importantly the finance, for the conservation of biodiversity hotspots or ecosystems most at risk of extinction. Two of the largest multi-collaborative bioprospecting projects that evolved out of this period and debuted in 1991 were the bilateral contract between Costa Rica’s National Institute of Biodiversity (INBio) and the large pharmaceutical firm Merck and Co., and the US federally funded International Cooperative Biodiversity Groups (ICBG) program (Neimark and Tilghman 2014, 247–248).

Posed as “win–win,” these large projects were structured in a way to enable efficient high-quality research while allowing for the return of commercial benefits to both source-country laboratories and local communities. On
the research side, bioprospecting projects were meant to facilitate scientific and technology transfer between global pharmaceutical firms and source-country public and private laboratories. On the conservation and development end, international environmental organizations built relationships with rural resource users and traditional healers in attempts to deliver sustainable development in areas where the natural resources and traditional knowledge were sourced.

Yet, in contrast to previous research on natural products, which generally used targeted searches for specific plants believed to have medicinal qualities, these new projects boasted a much broader geographic scope of plant collection. In fact, bioprospectors were now spanning the globe to ensure the delivery of bioprospecting samples (plants, insects, and micro- and marine organisms) in bulk quantities.

One reason for the increased scale of collecting was the advanced technology now accessible to the industry. The adoption of user-friendly robotics and high-throughput screeners provided researchers with opportunities to run hundreds of thousands of samples per day against a particular disease target (e.g., cancer or HIV). Following this, computerized databases allowed for immediate analysis of the laborious and multistep process of elucidation and isolation. To take full advantage of these technologies, however, it was essential to collect in quantities unforeseen in the past. Major pharmaceutical firms were lured in with the reassurance that for meager investment and relatively little risk they could amass desired libraries of samples to test without leaving their laboratories.

Nevertheless, the emergence of bioprospecting brought about a host of ethical concerns surrounding the exploitation of source-country participants’ knowledge and worries regarding the ecological impacts of collection. In an attempt to placate concerns, a regulatory framework was set up to provide a structure for return benefits back to the countries where the material and knowledge was collected.

Open for signature at the first Rio Earth Summit in 1992 and ratified a year later, the Convention of Biological Diversity (CBD) was for many a significant step forward, calling for an equitable exchange of benefits in return for the facilitation of access to the country’s biodiversity. Articles 1, 8(j), and 15 of the CBD were especially important in this regard. The CBD stated that signatory countries now owned all biodiversity within its territorial boundaries, with the understanding that it had to make all attempts to facilitate access for research and/or commercial interests. This reversed the earlier Stockholm agreement in 1972, which stated that biodiversity was the natural patrimony of all humankind, and thus, a global resource (Svarstad 2005). There was some trepidation by scientists that countries might become too overprotective of their resources, but most agree that the CBD was a major step forward in providing signatory countries the incentive to conserve resources through the commercialization of their biodiversity.

In order for the CBD to work in practice, bioprospectors and source-country governments had to agree on Access and Benefit-Sharing agreements (ABS)—a requirement that was finally ratified in 2010 by the Conference of Parties at the Nagoya Protocol. Although ABSs usually differed in structure, many shared the same mutually agreed upon terms, including (i) outlining how benefits were to be distributed, (ii) confirmation that details of informed consent were granted prior to collection, and (iii) goals for biodiversity conservation. Monetary and nonmonetary benefits from bioprospecting were usually delivered in the form of milestone or royalty payments, access or licensing fees, and technology transfer of equipment, materials, and/or trainings (Ten Kate and Laird 1999).
Due to the lengthy time necessary for discovery, small-scale sustainable development projects were constructed at the local level.

Biopiracy and resistance

Right from the beginning, however, the optimism surrounding bioprospecting was quickly met with well-organized resistance. Many held that the practice was essentially biopiracy, or the systematic theft of traditional knowledge and nature. First coined by the Canadian activist group Rural Advancement Foundation International (RAFI), biopiracy was used by the Indian scholar-activist Vandana Shiva (1997) in her book of the same name. The thrust of the critique coalesced around whether rural inhabitants received fair compensation for the natural resources and associated traditional knowledge used in the drug discovery research. The resources and knowledge used by locals for medicinal purposes had been discovered, it was argued, through countless years of “trial and error.” Many of the more vocal critiques claimed that given the history of exploitation of the global south leading back to precolonial times, any form of research and/or commercialization of biodiversity under the banner of bioprospecting was biopiracy (ETC Group 2004). Resistance grew to the point that the Maya–ICBG project operating in Chiapas, Mexico, was eventually discontinued (Rosenthal 2006), followed by other projects operating in Peru, Nigeria, and Indonesia, that were encountering problems due to disagreements from participating indigenous groups. These events sent shockwaves across the bioprospecting community, ultimately leading some scientists to question the ethics of the practice and their participation within it.

Although opposition to bioprospecting manifested from larger anti-globalization and indigenous rights movements, many of the critiques can also be traced back to the 1970s–1980s surrounding the collecting and commercializing of agricultural crops for the newly emerging biotechnology industry (Neimark and Tilghman 2014, 278–279). The galvanizing issue was the World Trade Organizations’ agreements on trade related aspects of intellectual property rights (TRIPS) and the contentious debate over intellectual property rights (IPRs). The TRIPS agreement was meant to provide a way of facilitating patents on discoveries of nature for life science and agribusiness multinationals and universal IPR standards that are enforceable through trade protocols. The logic was that patents would incentivize research and discovery within a globalized marketplace, thereby opening the door for ownership of novel discoveries. TRIPS essentially codified the application of western IPR law in developing countries. Yet, many argued that this trade agreement weighed heavily in favor of multinationals who had the financial resources and the political power to guarantee patents without any recognition of the original knowledge leading to the discovery, running counter to the spirit of the CBD and, in particular, the protection of indigenous people’s intellectual property.

A second major critique of bioprospecting concerned the fairness of the ABS agreements themselves. Many began to question whether these agreements could even be equitable, given that they were negotiated between high-powered lawyers representing large pharmaceutical companies and rural resource users with very little experience in negotiating such deals. One landmark case included the ABS agreements surrounding the Hoodia gordonii – a small cactus that grows in desert regions of southern Africa used by the San people to stave off hunger. The first license agreement was made through a scientific agreement with the South African-based...
Council for Scientific and Industrial Research (CSIR) and a joint-venture between Pfizer and a UK-based Phytopharm. Once in place, the agreement was challenged because it was signed without the consent of the San. A subsequent ABS, which included the San, was also contested on the basis that the group did not have proper legal representation at the signing (Wynberg 2010). The rights to commercialize the plant were subsequently sold to the giant multinational Unilever for the development of a functional food using the plant, but the company eventually dropped efforts due to health and sourcing concerns.

Another noteworthy case of drug discovery surrounds the commercialization of the rosy periwinkle (*Catharanthus roseus*). Commonly perceived as the quintessential case study of global biopiracy, the story of the periwinkle is quite a bit more complicated. The periwinkle is a pantropical plant that is generally agreed upon by botanists to be indigenous to Madagascar. Research on the periwinkle by Eli Lilly conducted in the late 1950s focused on the use of the plant as oral insulin for the treatment of diabetes. This work followed published and anecdotal accounts of its use as a bush tea to regulate sugar levels in Caribbean and Filipino communities. After extensive research, the periwinkle was serendipitously found to have a number of tiny vinca alkaloids, which were useful as an anti-cancer treatment of non-Hodgkin lymphomas and childhood leukemia. The original source material for the research, however, was collected from multiple sources, including from India and Jamaica, not Madagascar (Irving Johnson, personal communication). Lilly did, eventually, source periwinkle from plantations in southern Madagascar, but the country did not receive any royalties from the sale of the drugs, which some estimates put close to 400 million before the expiry of the patent. The periwinkle case leads one to question what claims Madagascar can make to benefits when the original source material or medicinal knowledge leading to the discovery was not collected on the island.

Many issues surrounding the landmark case of hoodia and the rosy periwinkle can be attributed to the complex political, biophysical, and geographical challenges of bioprospecting. First, much of the biodiversity desired for drug discovery are collected in tropical and subtropical ecosystems that are very far removed from the large-scale advanced research facilities in the global north. The biological/technical resource imbalance raises a number of proprietary issues concerning the collected material and knowledge. For example, one major issue surrounding the practice is that it is extremely difficult to trace collected material housed in botanical repositories and laboratories years after it was collected. As Bronwyn Parry (2004) notes, the increasing ability to “re-mine” this material through subcontracted rental agreements to third parties raises questions about the accountability and transparency of the original scientific agreements.

Second, the range of collected natural resources used for commercialization sometimes spans political borders and/or different communities’ customary boundaries and, as such, may not be covered by benefit-sharing protocols or afforded protection of intellectual property rights. Furthermore, many benefit-sharing laws are complex and outdated and source countries have not been able to keep up with appropriate regulations to deal with the emergent technologies of bioprospecting and biotechnology. Yet, maybe the most damaging effect of not having specific bioprospecting laws is the inability of source-country governments to hold bioprospectors accountable for the misappropriation of benefits. In the end, even with laws in place, keeping bioprospectors accountable is
extremely challenging for source countries due to the largely secretive nature of drug discovery and the lack of transparency of the research agreements.

Finally, laws governing biodiversity programs and associated benefit-sharing agreements are sometimes complex and difficult to understand. The potential for problems thus exists between parties who understand the terms of the agreements and those who do not. These misunderstandings may extend into mistrust among government institutions and may develop into over-strict policies causing unnecessary restrictions even for host-country scientists to access material for basic research. Many questions also remain as to the real value of biogenetic resources and traditional knowledge, which remains difficult to calculate in economic terms and, as such, there are gross miscalculations and expectations as to the benefits that can be gained from bioprospecting.

Due to these complex issues coupled with the onset of new technologies, major pharmaceutical firms have, on the whole, for some time now been moving away from bioprospecting and shutting down their natural products divisions. Instead, the industry has made a strategic shift toward the use of libraries of computer derived generated compounds—a process called combinatorial chemistry or “combichem.” The advantage of combichem is that it enables molecules to be “tailored” to fit the desired target. This method promises to shorten the time and lessen the financial burden in bringing home the blockbuster drug. Many arguments in favor of synthetic drug development notwithstanding, the pharmaceutical industry’s output under combinatorial chemistry has not lived up to expectations, just as the industry has not lived up to expectations. This lack of output has caused some in the industry to rethink the important role that natural products may still play in drug discovery.

In fact, to write off bioprospecting would be to misunderstand the practice. Private biotechnology firms and laboratories who continue to collect natural products have diversified their approach. Beyond just researching drug discovery, bioprospectors are putting their energy and resources into an array of natural products, including industrial biofuels, agrochemicals, functional foods, cosmeceuticals, and nutraceuticals. This also parallels shifts to collect natural products from nontraditional sites, including unexplored extreme environments (called extremophiles), or plant, animal, or microorganisms that thrive in extreme biophysical or geochemical environments, such as deep-ocean thermal vents or alkaline and saline pools.

Another major trend in the industry comprises developing technologies of synthetic biology and biomimicry, and the emerging work on re-engineering life forms using models based on nature. Recent discoveries of the first synthetic genome mapping may actually revive some programs which use natural compounds as a model precursor for new drug discoveries. As synthetic biology and biomimicry research expand, however, a new wave of social resistance seems to be growing as well. Although many of the environments, practices, technologies, and laws have transformed the practice of bioprospecting over the years, resistance remains strong and the schism remains between practice and critics’ calls for more distributive and procedural justice.

**SEE ALSO:** Biodiversity; Biotechnology; Environment and law; Environmental policy; Ethnobotany; Green capitalism; Indigenous technical knowledge; Natural resources; Nature conservation; Neoliberalism and the environment; Political ecology
References


Further reading


Biosecurity, broadly meaning making life safe, has been framed by a series of events and colored by its application within various domains. The term rose to prominence at the turn of the century following a period of three decades in which a raft of newly emerging infectious diseases (EIDs) that crossed from nonhuman animals to people (so called zoonotic diseases) had come to the attention of Western states. They included AIDS (acquired immune deficiency syndrome), avian and swine influenzas, Marburg, West Nile, Ebola, and others. As a term it gained prominence and took on a militarized set of meanings as part of a post-9/11 zeal to make the Global North as secure as possible. Following the “paradigmatic” SARS (severe acute respiratory syndrome) events of 2003 wherein global health systems seemingly contained a possible zoonosis, biosecurity became a matter of ensuring preparedness and organization in the face of emerging infectious diseases. Meanwhile, biosecurity also started to apply to human-made threats, including bioterrorist activities and the inadvertent production of hazards within research laboratories. Last and by no means least, various animal diseases and food scares meant that biosecurity often became associated with threats to farming and food security.

With these and other events in mind, we can highlight four domains of biosecurity: emerging infectious disease; bioterrorism; the cutting edge of life sciences; and food safety (Collier and Lakoff 2008). Each is largely focused upon microbial threats in a mobile world of rapid environmental and sociopolitical change, with the deliberate manipulation and/or dispersal of mutable microbes, the inadvertent release of laboratory-based organisms, and the vulnerabilities of a mass food production system as core concerns. Fears provoked include epidemics and pandemics, epizootics, crop diseases, food-borne illnesses, threats to vital systems, and the “perfect storms” that are generated through intersystem failures that may include promiscuous and exuberant microbial life as one of a number of triggers. A fifth domain in which biosecurity has been invoked would include the more macrobiotic realm of invasive plants and animals, a domain that together with microbial diseases of trees draw in broader fears over national landscapes and ecological security.

A geography of biosecurity may well approximate to the following: a concern with public health, terrorism, and laboratory science dominates in the United States, while Europeans have been particularly concerned with food-sector vulnerability to disease, and Australians have been most active in the area of preventing the incursion of invasive species (see Alien and native species). Within Asia and Africa, concern has often focused on the threats to livelihoods posed by the diseases themselves as well as what some might see as the disproportionate responses of Western-based interests in reducing the emergence of those diseases (especially when other more mundane diseases remain a major threat). There are also important sectoral differences in emphasis, with national and international public health bodies like the World Health Organization (WHO) advocating pandemic preparedness and capacity building, while veterinary
and animal health organizations advance disease prevention and eradication of threats (often as a means to conform to international trading rules) and development and food organizations have vacillated between a concern to modernize production systems and the safeguarding of economies and livelihoods.

Despite this seeming variety, there is frequently a broad sense that biosecurity expresses a (human) will to exert power over life, a power that is exercised through increased border activity, disease surveillance, and interventions in living processes (Braun 2007). As we will see, this sense of human domination over nonhuman life may be a part of the problem. In order to explore why, we can focus on four key elements of biosecurity and in doing so tease out the spatial tensions that make biosecurity always far from being a straightforward exercise in control.

First, the generic issue of security, second, the relationship of security and liberalism or neoliberalism, third and fourth a pathogen-based and an anticipatory approach to danger.

First, the term “security” refers to a wide range of activities that together can instil a sense of safety and assurance as well as prevention of harm (see Security). The term may usefully be distinguished from defense. If the latter is about stopping a known enemy, security often indicates an altogether broader set of practices and states of feeling or affect. Security, in this sense, is not simply a matter of containment or shoring up existing (nation-state) boundaries (see Borders, boundaries, and borderlands). Rather, security is more constructive, and may involve building new borders within and beyond the nation-state, acting to prevent and pre-empt danger, and shaping the environment through constant activity. Security involves regulating or modulating known dangers (or risks, see Environmental risk analysis) as well as preparing for unknown though perhaps inevitable events, and possibly actively intervening in or even pre-empting dangers (like, for example, manipulating a virus in the laboratory in order to produce a highly infective strain for which, possibly, vaccines or other pharmaceuticals can be developed ahead of a viral evolution in the wild). If defense was largely a state-led exercise in preventing attack, security is more intrusive, drawing in a wide range of actors and institutions. In its call for continuous vigilance it inhabits everyday activities, affects (see Affect), and processes and even colonizes the future.

Second, securing life only really makes sense when we understand how life itself comes to be defined in liberal economies that are founded on economic expansion, exchange, and circulation. It is in these kinds of society that life comes to be defined as a set of calculable traits, which are amenable to some limited manipulation. The French philosopher Michel Foucault identified the rise of liberalism and security with the rise of biopolitics, wherein, to be brief, governance by population (more accurately depicted as a risk pool that enables calculative inscription) is continually challenged and enabled by the very relations and circulations that make it “thrive” (Foucault 2004). A population is made up of and is continually affected by the expanding space of circulations within which it exists. Indeed, circulation becomes the paradigmatic space for biopolitics, and its regulation a major preoccupation. Briefly put, expanding economic activity opens up new risks and dangers (through for example increased trade introducing new kinds of vectors for disease), and therefore its continuance depends on technologies of sorting, categorizing, and, importantly, allowing some things and people to circulate while detaining others. As an example, plague stones, dating from the sixteenth to the nineteenth century in rural England, are a material mark of the technologies of circulation. These stones, with a saucer shaped
hollow filled with vinegar, were placed at the edge of a village and allowed coinage to circulate even when a settlement was under quarantine as a result of plague, cholera, or other infectious disease. This filtering of movement at borders is similar in kind to what happens at airports, seaports, farm entrances, abattoirs, and so on, as circulation is encouraged under conditions of observation, surveillance, and regulation. Indeed, most borders and associated technologies are not there to stop circulation – they encourage certain kinds of movement albeit under altered and policed circumstances. They are infrastructures that foster circulation in ways that are deemed to be economically and politically favorable.

This logic of selective movement is underpinned by what many authors see as a contemporary reinvestment in liberal or neoliberal approaches to governance (see Neoliberalism). This is the attempt to shift the burden of security and responsibility (as well as cost) of keeping life safe onto nonstate actors, and/or the state-sponsored encouragement of new markets in security. In the biosecurity arena, this is more often than not interpreted as both a privatization of the state armature (Cooper 2008) and a move to private responsibility for risk reduction and cost sharing of both risk prevention and post-event amelioration. This broadening of security activity coincides, Cooper notes, with attempts to increase exceptional state powers in conditions of emergency. Moreover, the whole exercise is underpinned by claims of radical uncertainty and emergence which are used, she argues, to justify increased vigilance on the part of citizens and to defend emergency powers, as well as to sanction the diversion of public funds into market-led security measures. This is a contested area – for some the extension of vigilance to broad sections of society at the same time as a reinvestment in emergency powers and a marketization of security is a peculiarly neoliberal settlement wherein sovereignty reemerges as a key source of power. For others, conceding that life cannot be controlled from a central site suggests a possibility for other kinds of organization and alternative forms of politics. This is an empirical as well as theoretical issue. Indeed, the exact interplay of emergence, uncertainty, borderings, surveillance, detention, and the freedom to move and to expand is handled in a variety of ways which require further analysis. How these play out, and how freedom and security are balanced or otherwise, is a key question for contemporary governance (Lentzos and Rose 2009).

If freedom and security already start to suggest biosecurity’s spatial tensions, these are overlain in some sectors by a pathogen-based and an anticipatory approach to infectious disease and risk. Here, biosecurity equates to a desire to pursue not just the freedom to move but also a freedom from disease. As such, the spaces of circulation that Foucault emphasized are in tension with territorial expressions (like zoning) and inside/outside dichotomies. The result is a tendency to understand biosecurity as a matter of constructing and then protecting a system of spaces wherein pathogens are to be kept out. So, for example, biosecurity in many states is concerned with the incursion of infectious diseases and their impact on people, domestic and wild animals, and plants. This geography of incursion is often played out across nation-state territories or more latterly (and in poorer or more uneven states) disease-free zones, such that a trading body or zone can submit the necessary paperwork to the World Trade Organization (WTO) via the World Organization for Animal Health (OIE – Office International des Epizooties) in order to verify their disease-free status, or conversely use WTO sanitary and phytosanitary agreements to refuse trade in food stuffs where there is significant evidence of disease risk. The end result is a spatial segregation
BIOSECURITY

between the virtuous (often the Global North) and the unruly (often the Global South with the exception of the capital- and pharmaceutically intensive disease-free zones).

This drive for pathogen freedom in the name of circulation and trade is amplified by an anticipatory ethos that exists within public health, animal health, regulatory, scientific research, and food retail organizations. Here a need to protect organizational reputation drives efforts to reduce disease risk by adopting disease-free practices. This precaution is amplified by an anticipation of retrospection or an approach to a future which may well involve being called to account for the actions that were or were not taken to reduce disease risk. Anticipation of known and unknown threats exerts pressure on producers, retailers, and regulators, who either assert their sanitary agency over the living processes they organize or specify contracts and legal responsibilities such that, should the worst happen, guilt cannot be linked to the actions of the institution or organization. The result of this pathogen-focused anticipatory landscape is a territorial interpretation of biosecurity with a networked performance of distributed accountability.

A problem with this sanitary and territorialized version of biosecurity is that life is seemingly made safe while disabling the very processes that may make life possible or even flourish. A problem may be that too much attention is focused on halting the transmission of pathogens and not enough is placed on the life conditions of those living behind the wall (Hinchliffe et al. 2013). The neurosis and bare life that exist under so-called conditions of hypersecurity may in fact produce ill health themselves, and/or an overconfidence that allows for massive amplification of risk (in a similar vein to at-risk landscapes sheltering behind always imperfect flood barriers).

As this last point suggests, the four elements of securitization, liberalism, pathogen-based and anticipatory approaches to infectious disease and risk can combine in rather uncertain ways. The resulting mix of logics and tensions makes biosecurity an always ongoing compromise between economic surplus, surplus life, and the requirement to make life safe. In the food sector for example, biosecurity takes its lead from retailers and food regulation services, which tend to accentuate consumer safety through a model of hygiene and uniform public health. The result is a contractual-based and vertically integrated supply chain that (at least in theory) performs risk aversion through disease-free technique (see Vertical integration). Yet, it may not be the lack of or need for more human control over life that is the problem. It is the very affirmation of agency and the assumption of abundant capacities of deliberate human action, coupled with biosocial issues like infectious disease, which are the cause for concern. Instead of trying to separate human lives from pathogens, it may be more beneficial to learn to recognize the limits of human agency and to adapt living to what are always dense entanglements of human, nonhuman, and microbiological beings.

In the biosecurity literature, at least as it applies to public health and animal and plant life, this shift of attention to adequate responses has tended to take the form of a counternarrative whereby various ways of living with threats are opened up for debate. So instead of policing life in terms of borders between health and disease, an alternative proposes the question of what would it mean to partially abandon strict bordering and invest instead in living pathological lives. One source for this proposition came from accounts of the history of foot and mouth disease (FMD). In the twentieth century, FMD became a disease to be eradicated in the name of global trade, rather than something that previously
farmers and their animals had learned to tolerate (Woods 2004). The cost of this purification of stock became apparent in the United Kingdom in 2001–2002 when the disease re-emerged in an ill-prepared country and herd. Separately, disease ecologists have also argued that the inevitability of disease re-emergence suggests the need for a program of developing resilience rather than eradication (Waage and Mumford 2008) (see Urban resilience). In anthropology and geography, works on establishing “common cause” (Farmer 2004) or more recent attempts to think through “common sense” (Hinchliffe 2015), as well as work with lay understandings of disease and health (Enticott 2008), can open up a new space for making life safe. This involves a shift in the object of concern from pathogens to pathogenicity or health, with Farmer’s radical and clinically informed work on health and justice acting as a rallying point. Rather than eradicating pathogens, the aims should be to build health such that pathogenicity is reduced. “Living with” in this sense becomes a matter of working alongside others, and crucially with those who appreciate the complexities of health and illness, the entanglements of pathological life, and, in doing so shifts the debate from securing territories (be they national states, food premises, or bodies) towards building better health-care systems, reducing inequality and vulnerability, and stemming environmental and socioecological degradation. Some of this is about redirecting investment to basic health care, some of it is about recognizing and valuing the skills that exist within food and farming systems. The latter may risk a return to the logic and politics of which Cooper warns, vesting new powers and responsibilities in nonstate and sometimes reluctant actors as a means to displace accountability and costs, all under the auspices of an ontology (emergence) that has its own noninnocent history. While these risks are evident, the alternative proposition of working alongside people and microbes is also made in the face of a tendency to simplify life and living in the name of disease risk as well as to render those who know most about the complexities and nuances of human, animal, and plant health, its multidimensional characteristics, redundant in highly capitalized but deskilled food and health landscapes. Recent work, for example, highlights the need to arrest a tendency to oversimplify the food chain and downgrade agricultural, research, and surveillance roles within the food, farming, and health sectors. Indeed, the sources for making life safe may well be found in the maintenance of the craftwork that is involved in patching together healthy lives. Safe life depends upon a transdisciplinary assemblage of expertise, from farmers to slaughterhouse workers, from vets to virologists, from local authority health officers to wildlife volunteers and professionals. These people, who work with bodies and environments on a daily basis, and through their craft and nonhuman alignments know the complexities of health (as always more than the absence of disease), need not be imagined as exemplary neoliberal subjects, nor as integrated and so simplified by a coherent health infrastructure. They are instead, the spokespeople, the human—nonhuman collectives that make safe life possible.

SEE ALSO: Biopolitics; Biopower; Governmentality

References

and A. Lakoff, 7–32. New York: Columbia University Press/SSRC.


Further reading

Biotechnology
Phil McManus
University of Sydney, Australia

Definition and history

Biotechnology was defined in the 1992 Convention on Biological Diversity as “any technological application that uses biological systems, living organisms, or derivatives thereof, to make or modify products or processes for specific use” (FAO 2002).

This definition potentially covers many techniques now used frequently in agriculture, food production, and health care, including animal husbandry, selective plant breeding for the development of commercially valued traits, vegetative propagation (i.e., cloning) of plants, and the production of antibiotics. Yet, a conventional understanding of the term restricts the meaning of biotechnology to the manipulation – or use – of genetic material from living organisms. Current activity covered by this definition builds on work undertaken in the 1950s when the structure of DNA was identified in 1953, and research in the 1960s that identified the proteins responsible for cutting DNA. By the 1970s it was possible to manipulate and transfer genetic material between species and by 1983 it was possible to create genetically modified (GM) crops (Bridge, McManus, and Marsden 2003; Stone 2010). DNA thus became recombinant DNA (rDNA) and opened up possibilities for new activity in a number of fields. The first commercially grown GM crop was tobacco, grown in China, in 1988 (Stone 2010). Since that time, GM crops of cotton, maize, canola, and soybean, amongst other plants, have been grown in a number of countries, mainly because these plants are commercially valuable and can be modified to tolerate herbicides (particularly glyphosate-based weed killers) and to resist insects (Stone 2010). In 2013, 175.2 million hectares of biotech crops were grown globally, up from 170 million hectares in 2012, and continuing the phenomenal increases in take-up rates since commercialization of GM crops began in 1996 in western countries (James 2013).

While many commentators refer to the history of human intervention and manipulation of natural systems as being of at least ten thousand years duration, the rise of modern biotechnology is generally agreed to have occurred in the mid-1970s. This is partly related to technical breakthroughs in science, as just summarized, but was necessarily accompanied by changes in the socioeconomic domains, notably the decision by the US Supreme Court in 1980 that life forms are patentable. Since that time, biotechnology has developed rapidly in many different fields. Mosier and Ladisch (2009, 27) highlight its impact – “new biotechnology has replaced, to some extent, traditional plant and animal breeding methods with tools that enable cloning and the directed insertion of genes from one species or genetic donor into another.” In summary, the technology has come so far already that only resistance by organized and highly funded groups, such as the thoroughbred breeding and racing industry, has prevented an even greater spread of biotechnology applications.

The history of biotechnology is often traced to the beginning of agriculture in about 10,000 BCE.
and the development of brewing techniques by the Sumerians and Babylonians, the Egyptians making unleavened bread, and copper smelting in Spain by the Romans over 2000 years ago (Bridge, McManus, and Marsden 2003). Yet, biotechnology is also aligned with the notion of “progress” as “universal, a historical claim that genomic technology and transgenic crops represent ‘progress for humanity’” (Bridge, McManus, and Marsden 2003, 165). In the past this may have been sufficient justification for the development of biotechnology but changing understanding of the relationships between humans and other parts of nature has led to the questioning of conventional notions of progress. This requires the generation of new forms of justification of biotechnology to secure funding, regulatory approval, and public/consumer support for what is produced by biotechnologists.

Justification for the development and application of biotechnology varies over time, depending on the location, and is influenced by the sector in which it is proposed for application. The most common justifications relate to population, particularly at the global scale. Borém and Fritsche-Neto (2014, 1) emphasize the “continuous growth” in global population, the rural to urban migration, and the requirement that “all current food production systems must either double in productivity until 2050 or risk failing to meet the growing demand for food.” They raise the specter of Malthusian mass starvation but offer the hope of scientists who “seek more efficient food production methods,” of which “cultivar development from plant genetic breeding is considered one of the most important and has been responsible for more than half of the increases in crop yields over the last century” (Borém and Fritsche-Neto 2014, 1). This justification is rejected by a number of geographers and rural sociologists, including Dibden, Gibbs, and Cocklin (2013, 61) who argue that “food security is evoked primarily with reference to developing countries” and “is supported by other emotive and guilt-provoking terms such as ‘world hunger’ which are likely to resonate with ethically minded consumers and citizens.” They argue that food security “provides a new discursive pivot for the debate about agricultural biotechnology, enabling issues surrounding GMOs (genetically modified organisms) to be transposed to a different spatial arena (the Third World) and a higher moral plane” (Dibden, Gibbs, and Cocklin 2013, 61).

Procedures and applications

As many commentators have recognized, discussing biotechnology means engaging with the diversity of procedures that fall under this label, and the uses of these procedures. This diversity extends to specific areas within biotechnology, such as the use of recombinant DNA, or rDNA, which includes different methodologies that vary in their risks and benefits. It also varies between fields such as agricultural biotechnology and microbiological technology, particularly in the type of organisms that are modified. While this may seem to be fragmentation, the engineering of genes brings together many academic disciplines. Geography is not directly involved in the production of biotechnology or genetic engineering but has provided important contributions to the understanding of this technology in terms of its origins, application, and impacts in local areas and at wider social scales. The engineering of genes requires an understanding of four basic concepts, which McHughen (2008, 37) identified as the foundation of genetics. These are:

1. all organisms are made of cells and cell products;
2 each cell in an organism contains the same set of genes;
3 the genome contains all the genetic information necessary make an entire organism;
4 all organisms share the same genetic language.

McHughen (2008, 37) identified the last of these concepts as being “the key concept” for genetic engineering because it “allows a gene from any one organism or species to be read and understood when transferred to another.” While this understanding enabled biotechnologists to progress rapidly with the development of new products and life forms, opposition to biotechnology has also been very high. The concerns about biotechnology date to at least the 1970s, when scientists began experimenting with animal cells and animal viruses. Importantly, this concern came from among the scientists themselves, rather than a homogenous science community supporting the development of biotechnology, with prominent scientists publishing letters in *Science* and in *Nature* that, in effect, advocated a moratorium on biotechnological development until issues such as biohazards of recombinant DNA had been evaluated (Barnum 1998). The second Alisomar Conference on Recombinant DNA Molecules in February 1975 produced safety guidelines for biotechnology, but getting agreement on these guidelines was difficult. The guidelines were published in *Science* on June 6, 1975, thereby fueling a debate that continued for about four years (Barnum 1998), and arguably longer, as evidenced by the ongoing controversies around biotechnology and newer technologies such as nanotechnology, and by the accompanying increase in the production and sale of organic, GMO-free produce.

David (2008) identifies nanotechnology and biotechnology as being ultra-controversies, that is, the absence of discrete issues but the “bundling” of many issues into one controversy and highly polarized political positions in relation to the debates. McHughen (2008) highlights the importance of retaining distinctions such as red biotechnology (medical) and green biotechnology for agriculture (see later) and argues that the position of opponents to biotechnology is inconsistent. In this view, similar processes cannot be advocated in some situations but not others – “condemning one use of a ‘forbidden’ technology must condemn all uses” (McHughen 2008, 40). He compares the use of biotechnology to create new food products (which some people oppose, but he argues for the life-saving aspects of food) with the creation of medical products from rDNA that are not available in a natural form.

The issue of genetic engineering is, however, more complicated because not only are individuals able to argue about the benefits and concerns of different types of biotechnology in different contexts (for example, human health benefits of biomedical research as opposed to increasing agricultural productivity), but there are also many different perspectives based on numerous variables. Degeling, Irvine, and Kerridge (2014) investigated various faith-based perspectives on biotechnology and identified similarities between them around animal respect and animal pain, but noted differences in that some religions gave a strong mandate to biotechnology research if it resulted in reducing human pain. The importance of faith-based perceptions is an area of interest to some geographers who are concerned about the discursive construction, performative practices, and social implications of biotechnology.

Geographers have been involved in understanding perceptions relating to the introduction of new technology, including advocating caution about comparing different technologies, without thinking critically about their basis, use, and
possible implications (MacNaghten, Kearns, and Wynne 2005). Geographers have also provided the scientific and socioeconomic context for many different technologies, including biotechnology (MacNaghten 2008). This is possible because these technologies engage with many of the subjects of geography, including ideas of naturalness, environmental change, and how people adapt to their surroundings to meet their needs. Geographers have also been interested in the role of sciences in changing the world, but recognizing that this change begins with imagining and not with the “application” of pure science (MacNaghten 2008). Social and ethical dimensions of science are not limited to application but arise with “the aims and purposes underlying the production of scientific knowledge” (MacNaghten 2008, 109).

This recognition was part of the challenge facing Bridge, McManus, and Marsden (2003), who attempted to move geographical perspectives on biotechnology away from simply mapping the location of biotechnology installations. Even so, it is worth considering the importance of venture capital and the concentration of biotechnology in particular locations. The initial success in developing human insulin in the late 1970s, being the first recombinant biotechnology product approved by the US Food and Drug Administration in 1978, was dependent on the availability of venture capital (Mosier and Ladisch 2009). The risks of failure were high but the potential financial rewards were also very high. This was particularly the case in pharmaceuticals (red biotechnology), where it was relatively easy to estimate the market value of a particular product and to calculate the financial risks in developing this product, although being able to estimate the value of biotechnology companies offering the promise of future success based solely on intangible assets such as intellectual property is challenging.

It usually takes many years of investing in an expensive industry before financial returns are forthcoming, with considerable risk that there will be no returns. Schneider (2009) estimated that the average timespan for red biotechnology to progress from initial development to the market is 10–12 years. The red biotechnology sector attracted more venture capital than did other applications of biotechnology. Today, the red biotechnology sector is far larger than other biotechnology industries and is concentrated in the USA and Europe. Green biotechnology is second, while some commentators see the potential of white (industrial) biotechnology to grow rapidly as the third largest application of biotechnology development, based on a shorter (3–5 year) development phase and on the fact that there are less regulatory requirements than in other more established biotechnology sectors (Schneider 2009).

In contrast to the concentration of medical biotechnology in the developed countries, the location of biotechnology crops is diverse, with 27 countries in 2013 planting such crops, but only eight of these countries are in the developed world, with 19 being developing countries. The top five countries for biotechnology crops, in terms of area planted, are the USA, Brazil, Argentina, India, and Canada (James 2013). While the USA is by far the country with the biggest land area devoted to biotech crops, the total land area devoted to these crops in the 19 developing countries exceeds that of the eight developed countries, a scenario that emerged in 2012 largely due to the rate of uptake in Brazil (James 2013). Importantly for the future growth of biotechnology crops, biotech cotton is grown in China and three African countries (South Africa, Sudan, and Burkino Faso) grow biotechnology cotton, while maize and soybeans are also grown in South Africa (James 2013).
Importantly, however, the locations of the companies involved in pharmaceutical and agricultural biotechnology is skewed towards the developed world, with seven of the top ten biotechnology companies by market capitalization being based in the USA. This crude statistic also masks uneven distribution within the USA, with three of the five largest biotechnology centers being located in California (Los Angeles/Orange County, San Francisco Bay Area, and San Diego) (Statista 2014).

Not only are the locations of biotechnology industries and application varied, there is also significant concern about the costs and benefits of bioscience. These are also unevenly distributed, as biotechnological application exists within multiple systems and flows of capital, knowledge, and power. There are particular concerns about biopiracy (the alleged theft of genetic material, often from biodiversity rich developing countries by scientific and pharmaceutical organizations based in wealthier countries) for the creation of profit. Additional concerns about biotechnology identified by Mellon (2008) include global resentment of American technical hegemony, corporations controlling the food system, the industrialization of agriculture, the blurring of university–corporate boundaries, and the opportunity cost of investing in biotechnology research, which she believes results in other technologies being underfunded.

Bridge, McManus, and Marsden (2003, 166–167) observed that “a defining characteristic of contemporary biotechnology is that its applications are widespread and ubiquitous.” These authors meant both in its geographical distribution and its application in different economic sectors. Bridge, McManus, and Marsden (2003) claimed that the diversity of applications fell into three main economic arenas in which biotechnology was then becoming significant; namely "plant/crop breeding, animal husbandry, and medical research.” In 2005, the 12th European Biotechnology Congress used four colors to identify biotechnological applications. These were white (industrial), red (pharmaceutical), green (food and feed), and blue (environment). DaSilva (2004) developed a ten-color schema that included red (health, medical, diagnostics), blue (aqua, coastal and marine) and dark (bioterrorism, biowarfare, biocrimes, anticrop warfare). This schema varies from the six-color classification (https://www.elsevier.com/life-sciences/biotechnology), which has red (medicine), white (industrial), green (agriculture), blue (marine and aquatic processes), black (research related to bioterrorism), and grey (environmental, including both biodiversity maintenance and contaminants removal). While there is sometimes slippage and overlap between these color classifications, within these classifications the application of biotechnology to plants and crops includes the genetic manipulation of seeds, tissues or plant cells to make them more desirable. This is achieved by modifying traits such as the size, growth rate, and resistance to herbicides and pests for particular plants. In plants intended for human consumption, other characteristics such as color, texture, and taste may also be genetically manipulated. This has resulted “in the spread of genetically modified (GM) crops, which has occurred with little controversy in some areas and with fierce controversy elsewhere” (Stone 2010, 381). Geographers are interested in the mapping and spread of such crops, and in the variations in responses to the introduction of GM crops across time and space.

In terms of industry size, the most important areas for the application of biotechnology are in the pharmaceutical (red) and green (agriculture and food) sectors. The remainder of this entry looks at these two sectors and identifies the key developments, major issues, and possible future directions for biotechnology development and
application, including potential implications. It is important to note that in practice there are overlaps between these sectors, and other sectors of biotechnology development, as demonstrated through the development of transgenic pigs, which have been developed for both food supply and biomedical research because of the similarities between pig and human anatomy and physiology.

Pharmaceuticals (red biotechnology)

The development of pharmaceuticals, especially cures to life-threatening diseases, is a major component of biotechnology development. The so-called red biotechnology has been at the forefront of biotechnology investment and research development, sometimes leading to the commercial development of new products because the market demand for life-saving drugs is high. One link between red and green biotechnology is the use of certain breeds of animals, whether it be for food production or for testing the effects of particular diseases and drugs.

The creation of chimeras (animals with both human and other cells, usually primate) for medical testing is controversial. Degeling, Irvine, and Kerridge (2014) note the contrast between responses to these chimeras and the relative lack of controversy around transplanting of pig heart valves into humans. The concerns can be summarized as being the welfare of the participants, the integrity of natural processes, and the wider social implications of this research. While this research may provide knowledge and possibly cures on long-term degenerative diseases, the above concerns have slowed the rate of biotechnological development in the western world. The regulation of biotechnology to address these sorts of issues, whether they be based in scientific research or be founded on public perceptions, is important in the rate and manner of biotechnological development in medical and pharmaceutical research.

One chimera that has developed rapidly in recent years is the transgenic pig. The pig has many advantages over other animals for use in biomedical research, including early sexual maturity, favorable short generation time, multiple offspring, is similar to humans in being omnivorous, has a digestive rate and efficiency similar to that of humans, and the feasibility of efficient and precise genetic modification (Wolf et al. 2014). The rapid increase, and the projected increase, in human medical conditions such as various forms of diabetes (often related to lifestyles in developed countries, but expanding rapidly into the wealthier developing countries) is a concern for governments and health authorities. The funding of research to address this issue includes the creation of research animals, such as transgenic pigs. Technological development means that it is now possible to create transgenic pigs to address specific issues within diabetes research and to target other degenerative diseases that impact on human quality of life. The long-term goal of much of this research is to develop xenotransplantation, that is, the surgical transfer of cells, tissues, or whole organs from an organism of one species to an organism of a different species, with the intention of regeneration of the recipient. In the case of diabetes biomedical research, the development of an entire pancreas in a transgenic pig which could then be harvested and transplanted into a human is within the bounds of possibility. This raises ethical concerns about the integrity of human life and about the ability to create new forms of life, and about the transfer of infectious diseases from one species to another. At present interspecies transfer of disease involves ingestion, physical contact or a vector of some kind (for example, wind), but the potential to accelerate interspecies disease transfer is a concern for some
critics of medical biotechnology, especially if it becomes widespread and beyond the confines of the most secure laboratories.

Agriculture and food (green biotechnology)

In the area of agriculture, the application of biotechnology can be subdivided into animal and crop industries. The development of genetically engineered animals (otherwise known as transgenic animals), where a particular gene is directly transferred into an animal (unlike traditional breeding where the transfer has been indirect, or cloning which seeks to reproduce the exact genes of a highly desired animal), is a huge business. Perhaps the most famous example of genetically engineering an animal is in the field of medical rather than agricultural research, where Harvard University scientists developed OncoMouse®️, which was a mouse that had a gene which made it susceptible to cancer, and therefore ideal for use in medical research. This was controversial but has been overtaken by developments in other areas of research and commercialization of biotechnological modification of animals.

In 1992 the biotechnology company GenPharm produced a bull, initially by injecting cells with a human gene coding for lactoferrin (which is important in preventing bacterial growth) into the embryo of a cross between two Dutch breeds of cattle. The creation of Herman the Bull was extremely controversial because he was the first transgenic bovine but, even more importantly, he contained human DNA. He sired offspring that could produce milk containing lactoferrin, which is normally absent in cow’s milk but is present in the milk of women and in formula, but it was not commercially viable at the time. Following the controversial experiment in The Netherlands, the offspring of Herman the Bull were slaughtered, while Herman was euthanized at the age of thirteen (suffering from arthritis, unrelated to the genetic engineering) and is now an exhibit at Naturalis, the National Museum of Natural History in Leiden.

The genetic engineering of animals can be undertaken for multiple reasons, including increased productivity. It is possible, however, to enhance perceived attributes, such as the ability to grow faster, produce more meat, and so on, using conventional breeding techniques, or by the application of feed supplements, and so on. Where genetic engineering is unique, and hence potentially lucrative, is the ability to target particular issues, such as lactoferrin in milk in the example of Herman the Bull. Another example of such targeting is the development by researchers at the University of Guelph of the Enviropig, which is a swine that has been genetically modified by the inclusion of an E. Coli bacteria and mouse DNA to a Yorkshire pig embryo so that the pig’s urine and feces contain about 65% less phosphorus than that of other pigs. This characteristic is a boon for the genetic engineering industry, as it enables the industry to showcase the environmental benefits of reduced phosphorus entering water bodies and generating eutrophication, which leads to algal blooms and catastrophic loss of fish and other marine life. The Enviropig research program was shut down in 2012 when funding was cut, meaning that it did not become the first genetically modified food animal to be approved in the world. While the Canadian government had approved the development of the pig, no government in the world had approved the inclusion of genetically modified animals for human consumption (Pollack 2012). Despite closing down the program, and slaughtering the small herd of pigs, the semen from the pigs has been frozen and potentially can be used in the
future by a company willing to commercially
develop the product (Pollack 2012). Critics call
the pig “Frankenswine” and argue that:

the common disconnect between science and
reality is represented perfectly by the ridiculous,
and yet threateningly real GM Enviropig™
project. Enviropig™ is the grotesque realization
of early scientific aspirations and laboratory
accidents. Born of scientific curiosity, hubris
and a complete misunderstanding of the real
world, a GM pig with less phosphorous in its
feces is being proposed as a solution to water
pollution caused by run-off from factory farms.
(Sharratt 2010)

The concern among critics is that the develop-
ment of the genetically modified pig would not
have reduced the total environmental pollution
but would have enabled hog farmers in Canada
(and the midwestern United States) to increase
their stocking rates because the limiting factor
of environmental restrictions on the permitted
level of phosphorus pollution would have been
alleviated. This, in turn, would have enabled
the perpetuation of intensive factory farming,
at higher stocking rates than previously. Envi-
ropig is an excellent example of the challenges
of commercialization of biotechnology, even
though it is scientifically demonstrated to work,
of the controversies around public acceptance of
biotechnology (particularly in animal-based food
for human consumption), of the importance of
regulation in promoting or stifling biotechnol-
y, of the arguments used by biotechnology
advocates to support their activities, and of
the links between issues such as biotechnology
development and related issues such as ani-
mal welfare. This last point is consistent with
David’s (2008) labelling of biotechnology as an
ultra-issue, where there is a bundling of con-
cerns rather than a focus on an individual issue.
For opponents of biotechnology, this so-called
“bundling” is logical, as in the case of Enviropig
where the connections between genetically
modified swine, increased stocking rates, animal
welfare, alternative solutions to environmental
management, and the risk to human health are
seen as parts of a single issue.

The termination of the Enviropig research
program certainly does not mean the end of
transgenic pigs, even allowing for the freezing
of sperm for later use. According to Bou et al.
(2014, 679), “transgenic pig holds great promise
in the fields of agriculture and biomedicine.
Currently, at least 65 kinds of transgenic pigs
were derived for research in porcine meat qual-
ity, growth and disease control, human disease
model, xenotransplantation, pharmaceutics
and the fundamental biology.” Importantly,
the “normal” pig becomes the “wild type” of
genetic material in this study, highlighting the
perspective of these scientists towards the control
of nature.

The other sector for the agricultural appli-
cation of biotechnology is in the development
of genetically modified crops. Biotechnology’s
major contribution to the crop industry, to date,
has been the ability to produce glyphosate-
resistant (GR) crops. As Green (2012, 1323)
obsved, this “was a scientific breakthrough that
helped to revolutionize weed management and
provided much of the impetus to restructure the
seed business.” Glyphosate works by blocking an
enzyme that is required in the process of pho-
tosynthesis. Using biotechnology, scientists have
created plants that “produce an alternate enzyme
that is not affected by glyphosate” (Thieman and
Palladino 2004, 144), in other words, a transgenic
crop that allows farmer to spray weeds without
the danger of killing the economically desirable
plants, which was a problem with some previous
herbicides. Glyphosate was “a relatively inex-
pensive and effective herbicide with an excellent
environmental profile” (Green 2012, 1324).
The timing of this scientific breakthrough was also important because weeds were becoming resistant to many herbicides that were in use at the time. Green (2012, 1323) adds that the timing also involved an increase in farm size while the number of farm workers was decreasing and “weed management was becoming too complicated, time consuming and costly for the new agricultural systems.” The use of biotechnology to create GR crops “made weed management easy, efficient, economical and environmentally compatible – exactly what growers wanted” (Green 2012, 1323).

Some crops were particularly suited to biotechnology, with soybeans being a prime example. By 2005, 87% of the total US soybean acreage comprised genetically engineered soybeans (McHughen 2008). Another crop that is suited to the application of biotechnology is cotton, which in the same year saw genetically engineering cotton account for 79% of the US cotton acreage (McHughen 2008). Thieman and Palladino (2004, 144) in advocating biotechnology observe how the social and environmental relations around agriculture have changed as a result of biotechnology – “before the advent of resistant crops, US cotton farmers spent $300 million per year on harsh chemicals to spray their fields (This total does not include the human factor of laboring in the sun to weed between the cotton plants, which is now unnecessary!).”

These examples highlight the justifications for biotechnology. They also demonstrate the role of biotechnology within a political economy of agriculture, which includes the ongoing replacement of human labor by machinery or other technological approaches (such as the application of biotechnology to remove the need for weeding). While corn, soybeans, and cotton were early crops in which genetic engineering was applied (all became commercially available in the late 1990s), more recent applications of glyphosate resistance include alfalfa (2007) and sugar beet (2008) (Green 2012).

The application of biotechnology to various crops has influenced the construction of productivist and post-productivist agricultural landscapes (Wilson, 2007). The application of biotechnology enables proponents to increase agricultural production whilst claiming environmentally beneficial outcomes. In this sense, Marsden (2008, 192) argued that genetic modification technologies are globally becoming an integral part of a changing agroindustrial paradigm, by “providing palliatives to the technical problems of large-scale capital-intensive producers, such that they can continue to effectively supply the concentrated downstream sectors with relatively cheap and bulk agricultural inputs.” The application of biotechnology enables the use of herbicides and pesticides, thus increasing outputs while potentially reducing labor costs and, thereby, making some agricultural systems more “productivist” than previously was the case.

Biotechnology adoption is, however, uneven “both spatially, and in terms of its application in different parts of the food chain” (Marsden 2008, 193), and productivist, post-productivist, and super-productivist models can coexist (Wilson 2007). This places farmers in a difficult position regarding the adoption of genetically modified organisms (GM or GMO). They may be “caught between the promptings of industry to adopt new technology, protest groups opposed to GM, and consumer wariness in relation to GM foods” (McDonagh 2014, 5). The pressures vary between locations and specific situations. Recognizing that biotechnology is both a discursive and physical insertion into existing agricultural and food production/consumption systems also means that “very different conceptions of sustainable rural futures are possible” (McDonagh 2014, 4).
The production/consumption relations surrounding the application of biotechnology in agriculture extend beyond herbicide resistance and labor-saving practices. In summary, the benefits include B1 crops (crops inserted with a gene from a South American tree frog to resist fungal growth and insect damage), the development of soy cooking oils with reduced saturated fat (which is important in promoting healthy eating and addressing growing concerns about obesity), and the ability to delay fruit and vegetables from ripening to enable shipping and longer freshness – the now defunct Flavr Savr™ Tomato that was introduced in 1994 being a prime example of application in this area. This application was designed to ensure that the tomatoes reached markets out of season, where supply was low and they could be sold at a premium. Biotechnology, therefore, influences the form of food consumption in locations that may be far from the initial biotechnological application in agricultural settings.

Biotechnology futures

The future of biotechnology is open to speculation. Diola and Fritsche-Neto (2014, 155) believe that “the development of genetic maps with markers of easy attainment and of large-scale, highly reproducible, co-dominant and specific to linkage groups, is highly desirable for their application in plant breeding.” These genetic maps form the basis for genetic engineering experiments, which are increasingly moving beyond agriculture to feed a starving world, or pharmaceuticals to save human lives. Two areas of development are particularly important to note. First, the creation of transgenic animals for nonhuman consumption (i.e., leisure and entertainment, human companionship, and so on) is now underway with the commercial release of the GloFish. This is the first transgenic animal available to the public; it has resulted from the insertion of protein from jellyfish that enables the zebra fish (a native of rivers in India and Bangladesh, but sold as a pet in developed countries for many years) to develop enhanced fluorescence. The GloFish were first available in Taiwan, followed soon afterwards by the USA (except in California). The importance of regulation is highlighted by the fact that these small tropical fish are not for human consumption, and are therefore not threatening to the human food supply.

The second area of biotechnology development that indicates future possibilities is the genetic engineering of marmosets (a small monkey, originally from South America) that resulted in a transgene being passed to the offspring of the modified animals (Cyanoski 2009). This development has significant implications for medical research, in that, to date, the usual animal used for experimentation when research is at an advanced stage is the Rhesus macaque (a taller monkey, originally from Asia). Scientists in Kawasaki injected viral vectors with green fluorescent protein (GFP) into 91 marmoset embryos, eighty of which were to surrogate mothers resulting in five offspring, all “of which expressed the glowing transgene in some features at some point during development” (Cyanoski 2009, 492). The next step was to use artificial insemination to produce glowing second-generation marmosets. This action is controversial for a number of reasons, including the creation of transgenic animals that are produced specifically for scientific experiments (similar to transgenic mice) but particularly because primates are close to humans both in their phylogenetic composition and on the sociozoological scale, where nonhuman primates and companion animals are positioned at the top of this scale (Olsson and Sandøe 2010).
The phylogenetic proximity between humans and marmosets is advantageous for medical research, but raises issues about the potential for human genetic engineering.

**Conclusion**

Human geographers are interested in biotechnology for many reasons as indicated here, although as Greenhough (2011) noted, there has been little crossover between geographies of health and geographies of bioscience. Foremost among the reasons is the way in which the “human” in human geography is understood. As Nash (2005) observed, we increasingly understand what it means to be human in ways that are informed by genetics and biotechnology. This includes understanding the history of not just an individual, but whole races and the processes of migration through the mapping of the human genome diversity project (Nash 2005). It also includes future-oriented implications, such as the establishment of biobanks or genetic database projects that can map out genetics and make it available for the application of biotechnology to overcome diseases and weaknesses, or to develop “improved” characteristics in individuals.

Biotechnology has advanced rapidly in recent years, both in scientific development and in commercial application. There is a well-established industry that is promoting the further development of biotechnology, including the important issue of regulation. The scientific development of biotechnology has advanced well beyond the point at which regulation is the only means that distinguishes between the possible and the desirable, based on benefits and risks. Any weakening of regulation will enhance the commercial viability of a number of scientifically developed projects that are currently not commercially viable, but this is not the same as protecting societies, the environment, and animal welfare. The future of biotechnology is intricately bound up in the regulation of these processes, their outputs, and implications.

**SEE ALSO:** Agricultural environments; Nature; Socio-nature

**References**


Further reading

The importance of the body for human geography may seem at first obvious. Bodies are everywhere in scholarship, popular culture, and everyday lived space. The term “body” is in the title or keywords of growing numbers of articles and books in geography and in the social sciences more generally, particularly since the 1990s (Longhurst and Johnston 2014). The meanings ascribed to the body, bodies, and embodiment vary significantly from one subfield to another and they have changed over time, affecting how geographic research is conducted, the questions posed, and the conclusions asserted. The shifts in understanding of bodies have not of course occurred in a straightforward, chronological march toward greater understanding. Rather, geographers have combined different ways of understanding the body in recent decades, in many cases developing complementarities between longer-standing approaches to bodies with contemporary social concerns and theoretical orientations.

Feminist geography, sexuality studies, queer geographies, and critical race scholarship have been especially important contributors to understanding differences between bodies (Nast and Pile 1998). These subfields have explored the power/knowledge relations associated with diverse embodiments and the theoretical implications of attending to, from, and with difference as it articulates bodies. They challenge the historically masculinist unmarked norms of the discipline and oppose the denial and distancing of the body, and the body/mind dualism, in classical geographic research. This work explores bodies as arenas through which social hierarchies are produced, reproduced, and contested and it takes seriously the scale of the body as a window onto power relations.

Early human geography assumed a classical Western philosophical notion of the human body, implicitly conceptualizing the human body as the common-sensical equivalent of an individual, universalizable person’s physical existence. In recent decades, scholars have increasingly challenged the assumption that bodies can be viewed as taken-for-granted sites of experience, and put forward analyses that question the ethico-political status of particular bodies, the limits and boundaries of individual and collective bodies, the dynamics of diverse embodied relations, and the relations of power that connect and divide raced, gendered, sexed, (dis)abled, and human/nonhuman bodies.

Human geography is deeply engaged in exploring how, why, where, and for whom specific individual and collective bodies and embodiments matter. The extraordinary interest in “the body” as a basis for a wide range of theoretical inquiries has affected most of the discipline’s subfields. Bodies figure as critical entry points for feminist research, sexuality studies, and critical race scholarship, social and cultural geographies, economic and political subfields, as well as themes in the margins of political ecology and population geography. Bodies, embodiment, and biopolitics now figure in research on topics as wide-ranging as care labor, migration, water governance, elderly people’s care, medical tourism, art, geopolitics, and warfare.
Feminist and critical geographers have contributed to conceptualizing and re-conceptualizing bodies and embodiment in virtually every theoretical register. For heuristic purposes, their interventions can be organized along the following five lines, all of which overlap and feed into one another in various ways: (i) phenomenological, living, sensing bodies, inspired by humanistic geography; (ii) laboring bodies, engaged from Marxist and socialist-feminist standpoints; (iii) disciplined and governed bodies, approached through poststructural insights; (iv) bodies of difference, cross-cut by gender, race, sexuality, and dis/ability, and informed by feminist, critical race, queer, and postcolonial theories; and (v) vital bodies, new materialist bodies, more-than-human bodies, and affective bodies, linked to a recuperation of bodies as alive with potential.

**Phenomenological, humanistic bodies**

In the 1970s, humanistic geographers paid renewed attention to bodies. They explored, for example, the ways in which bodies and places make each other and the meanings of an embodied sense of place. Their view of the body moved away from conceptions of bodies as units of human–environment relations. Rather, the bodies of humanistic geography were – and indeed are – sensorial, emotional, and perceptive. For phenomenological philosophers, the body was the fabric into which geographic processes were woven, and it was the primary instrument of perception and therefore comprehension. Humanistic geographers invoked the “body-subject” in the 1970s as an alternative to behavioralist views of bodies as objects. They argued that in order to develop more comprehensive spatial knowledge, geographers needed to examine spaces as felt and perceived first and foremost through the immediate experience of the body from which abstract ideas develop. The phenomenological humanists understood humans to think only after our bodily experiences and they proposed that therefore all knowledge is embodied and should be recognized as such.

From this perspective, it is what the researcher’s body can touch, smell, taste, hear, and see that are the elements of human experience from which thought and analysis can follow. This lens on the body retains the longstanding binary of mind/body, and simply turns it on its head. Now, rather than privileging the observing mind as the origin point of knowledge production, scholars were encouraged to prioritize the sensing body. Attending to the bodies’ experiences in and with a place or a landscape could open up fresh ways of knowing those places that were not reducible to behavior. Phenomenological bodies were valorized for their capacities to sensorially interact with places. Embodied experiences and place-making were understood to take shape in tandem with one another.

Early humanistic geography placed people as feeling subjects at the center of analysis and it understood embodied experiences as central to the human experience of space and place; however, power was not a major concern for these scholars. Their sensing body tended to be one that was abstractable without reference to its location in political economy, discursive fields, or body politics of any kind. The differences across experiences of embodiment were not of particular interest to phenomenologists. Contemporary scholars of embodiment, however, have re-invoked phenomenology alongside feminist theories of performativity as important elements for the construction of a renewed, practice-oriented humanism (Simonsen 2013).

**Laboring bodies**
Marxist conceptions of the body have been especially important for thinking about the politics of labor albeit that capitalist processes tend to be understood as largely disembodied. From this perspective, labor is a form of mediation, a metabolic exchange, between the human body and its environment; as commodities are produced, the labor power of bodies is appropriated and surplus value is accumulated. The laboring body is exhausted by its exploitation in the labor process and its alienation from the means of production. A vast, rich literature examines laboring bodies, and the ways in which they are regulated, as keys to understanding the ways in which geographies of capitalist production and accumulation are socially and spatially organized (Harvey 1998) and how different forms of capitalism, including merchant, industrial, and post-Fordist forms, rely on different forms of embodied labor exploitation.

The body is also a key scale of analysis for geographers who study gentrification and its links to homelessness. Historical materialist work along these lines shows that homelessness takes shape through the interlinked scales of the body, public–private distinctions, the city, and the global economy. Homeless bodies are rendered undeserving poor whose lack of property ownership is both their fault and cause for their eviction from public space. As “urban renewal” programs displace low-income populations, they marginalize the bodies of low-income people. Bodies are evicted from public space both literally and representationally. Homelessness makes it difficult to care for one’s own or one’s family’s bodies. When the challenges mount of washing, feeding, and relieving one’s body in practice, homeless people face an uphill battle for social acceptance in the public sphere.

Feminist geographers added to critical political economic research with attention to gendered body spaces as structured by socially constructed gender roles within families, households, workplaces, cities, and suburbs. They first illustrated these gendered sociospatial divisions along the lines of male versus female bodies, although the body itself was not a major focus of investigation. In research on work places, they also showed how “public” and “private” spheres rely upon and produce one another through bodies. They argued that the marginalization of women’s bodies in public spaces and formal politics, and the coding of personal and private spheres as feminized, are central to the persistence of male dominance in politics and the public sphere and that bodies interact in gendered patterns that reproduce male dominance in the workplace (McDowell 2009). They found that gender divisions of paid and unpaid labor, and the devaluation of women’s work relative to men’s in both domestic and wage-earning spheres, shapes everything from global labor markets to immigration patterns and local economies.

Disciplined and governed bodies

Poststructuralist, particularly Foucauldian, scholarship has also influenced geographic understandings of bodies. Foucault placed the body and biopolitics front and center in his work. He saw the body as subjected to meticulous forms of state discipline and documented surveillance as deployed first and foremost on the body. Through objectification, classification, and enumeration, bodies are made into subjects and populations. Medical and punitive knowledge systems and technologies slot bodies into categories and types as technologies of rule. He shows how premodern public executions made a spectacle of the criminal body’s suffering and death, while the modern state exercises its power through modes of power that are interiorized by
subjects directed at their own bodies. The modern state shifts away from the body and corporeal punishment towards a preference for indirect control over bodies through governmentality.

Foucault also put forward the concept of biopower as a way of understanding how human populations are regulated and imagined. He saw biopower as integral to the growth of industrial capitalism in that capitalism relied upon the organized insertion of bodies into production relations. The ideal laboring body was produced in order to maximize its exploitation for the sake of production and to be docile for the sake of labor control. In this view, the forces of the body are managed for the sake of efficiency and productivity. He saw bodies as reflective of history and the history of capitalism as destructive of the body.

Poststructural research examines the continually produced, contested, and destroyed sets of ideas and practices that make and unmake bodies. It does not seek an originary experience in embodiment. Rather, it asks how bodies are politically produced as normal/other or civilized/barbaric and how specific ways of knowing the body (e.g., medicalized, laboring, criminalized) can shed light on the machinations of power in a particular place and time. It decenters the universal human subject and thereby encourages attention to the politics of representing, managing, and knowing bodies. Prior to poststructural attention to the body, and often in conversation with it, feminists with critical race sensibilities have produced extensive scholarship about bodies (Federici 2004).

Bodies of difference

For feminist geography, the body is at the crux of analysis. It is alternately a site of struggle, a text that is read and can be re-scripted, and the basis of experience and knowledge. When feminists introduced gender to human geography, their work demonstrated the importance of gendered and sexed bodies in the making of all sociospatial relations (Nast and Pile 1998). In early feminist geography, the body was a relatively straightforward second-wave Western feminist vision: biological sex differences existed in male and female bodies, while gendered inequalities were socially constructed and tied to sexed bodies in tandem with norms of masculinity and femininity.

Research informed by feminist theory has also explored the body as a starting point for analysis. In addition to seeing bodies as socialized and socially constructed, studies have explored the materiality of embodied experiences, grounded in flesh and blood. They have analyzed the difference that it makes to live in and from pregnant bodies, “fat bodies,” disabled bodies, and bodies undergoing weight loss (Moss and Dyck 2003; Longhurst and Johnston 2014). This research challenges the universality of Cartesian body spaces and allowed for thinking through the body to understand the making of other geographies.

Tracing the specific ways that feminist geographers have approached bodies is therefore illustrative of multiple itineraries provoked by feminist interventions in human geography more generally. Research in human geography increasingly blends strands of these theoretical schools of thought and postcolonial and antiracist work has made key contributions to geographies of bodies. It reveals the deep histories and contemporary persistence of extreme and everyday forms of violence directed at black bodies (McKittrick 2000) and charts representational and epistemic violence tied to the discursive racialization of bodies. This work illustrates black bodies’ disproportionate subjection to slavery, imprisonment, and poverty.
Colonial narratives represented black bodies as more closely affiliated with nature in distinction to white bodies’ association with knowledge and the capacity to transcend the body through reason. Critical race scholars have analyzed “whiteness” as it attaches to, and is deployed via, privileged bodies. Recent antiracist research has turned a critical eye on the power of whiteness as it is produced through embodied relations. This work examines and exposes the power inequalities tied to emplaced racial hierarchies as they emerge in old and new forms. It shows how the unmarked norms of whiteness shape structures, policies, and knowledge in racialized terms that attach to bodies. Racialized power relations emit from myths about bodies as well as through bodily comportment, intersecting with multiple sociospatial hierarchies of belonging, entitlement, and power that also take embodied forms.

Much postcolonial critique can be read as warnings about how not to read or write the body. It invites geographers to think about aspects of bodies that are untranslatable because of their difference and their submersion in official historical transcripts. The subaltern subject is both discursively produced as devalued and othered, as well as a point from which to rethink Western assumptions about being, meaning, and voice. Recent theoretical research has returned to an examination of the materiality of phenotype in terms of its inescapable embodiment. Here the body is both a productive force and produced by its interactions with other bodies, embedded in histories both conscious and unconscious. Power relations surface in and on bodies and simultaneously shape the emergence and itineraries of bodies themselves. Bodies are contested terrain and malleable to some degree and people can recast aspects of the ways their bodies are read. Studies of mixed race women have highlighted the indeterminacy of bodies and the power of performativity in making embodied meanings and identities.

Sexuality is also embodied and queer theorists have made important contributions to understanding bodies and space. Scholars of sexuality and space have unpacked the heteronormative geographies at the base of much research and they have provided alternative theorizations of space and place that take seriously the standpoints of gay, lesbian, bisexual, and transgendered people. For example, they have shown that the domestic space discussed in second-wave liberal feminism was not only a patriarchal space, but also a heteronormative one underpinned by prescriptive heterosexuality. They have shown how lesbian, gay, bisexual, and transgender (LGBT) “public” and “private” spaces differ from heteronormative usages of these spaces and categories, both in terms of homophobic exclusions from various “publics” and reworkings of “private” not only as domestic but also in conjunction with sexual identities. They have studied the specificities of LGBT spatial practices to show how sexuality, for example, affects migration aspirations and embodied experiences of borders. For scholars of sexuality and space, bodies are crucial for thinking through not just power and difference, but also desire, affiliation, and identity. The transgender body, when understood as transformed in order for the person to change what s/he felt was the “wrong” body so s/he can live in a body that aligns with his/her sexed/gendered self, requires thinking about embodiment as irreducible to the dualisms of nature–culture, bios–technos, and sex–gender.

Every one of geography’s foundational concepts has been reconsidered with attention to difference. Different bodies – gendered, raced, and sexed – encounter and make places unequally through practices of spatial inclusion, exclusion, integration, management, control, travel, habitation, and claims-making, practices...
that are directed at bodies and that are encountered and challenged through bodies. Space itself is produced through embodied social relations of difference that take specific forms over time, whether researchers are tracing racialized residential segregation, women's bodies in the street or at night, HIV/AIDS activism, access to health care, embodied consequences of economic crises, or the highly unequal effects of war. The injustice of counting some bodies, and therefore some lives, more than others has galvanized tremendous oppositional activism. “Black Lives Matter” has emerged as a key rallying phrase for activists organizing in opposition to recent deaths of unarmed African American men at the hands of police. These are struggles, like the struggles of indigenous women in Canada to bring attention to the many murdered and disappeared women from First Nations communities, and like the struggles of feminist activists in Ciudad Juárez to make visible the femicide in their midst, to reassert the value of these different bodies as human lives.

Post-poststructural bodies

There are new theoretical languages to talk about bodies and embodiment. Renewed materialisms have reminded scholars that bodies – as matter – “talk” through actions including breathing, eating, reproducing, bleeding, ailing, and dying. In geography, such dimensions of bodily expression and experience have tended to be overlooked in research originating from disembodied observation and analysis. But when bodies re-enter the conversation the terms of dialogue shift. Some bodies’ pleasures, pain, illness, injury, and death are heard more than others. Antiracist and postcolonial geography have added the narratives of previously unrecorded voices and explored the ways that recuperating these bodies reconfigures the ways researchers understand historical agents and the dynamics of political economy.

Rather than starting with individual bodies as a focus, some important research attends to the relationships that arise when multiple bodies congregate together in a place. Studies have asked what results from many bodies coming together (Saldhana 2008) tracing the aggregate characteristics of groups of bodies to begin to explore how human bodies stick together not only in crowds, but also in neighborhoods, regions, cities, and nations. This work is decidedly relational in its approach to bodies, concentrating on how bodies affect each other as collective forces, energies, directions. The goal is less to understand the interiority or limits of individual bodies and more to examine bodies as always influenced by the other bodies they encounter. Bodies that come together in alliance for a cause may generate collective political power enriched through physical contact.

The “affective turn” in social and cultural geography has also prompted fresh questions about how we understand what influences bodies and what bodies can do. In particular, it asks what affects – rather than what causes, determines, structures, or produces – bodies, thereby enabling attention to the social reverberations of, for example, fear of violence or anxiety about economic futures. Bodies are affected not only by their material surroundings, the health care that is available to them, the food they consume, or the protections they lack or enjoy, but also (and in specifically socially located ways) by the tenor of public discourse, for example, of state-circulated scare tactics or widespread anxiety about economic decline.

Defining bodies in terms of affect makes capacities of bodies emergent rather than innate. It shifts attention from what a body is to what a body can do and what a body can become, thus
emphasizing the understanding of individual bodies as processes, always open, contingent, vital, and filled with potential. In this work, a body can be anything ranging from an animal to a mind, an idea, a linguistic corpus, or a social collectivity.

Geography has also returned to humanistic and phenomenological philosophical roots with some new twists. Building on these traditions, posthumanist philosophy is opening up a host of exciting questions for thinking about how bodies figure in human geography. First, it allows for the possibility that individual bodies may be sites where multiple selves reside or as complex compositions of multiple other bodies. Second, it asks where bodily boundaries might be drawn beyond the limit of the skin, and how bodies can be understood relationally, as lives are lived. Posthuman geographers are working to think beyond the body as the core, the interior, or the scale “closest in” of human being, and seeing it instead as a force in dynamic tension with nature and more-than-human bodies. In this growing collection of work, the human body is just one among many assemblages that have power to make and remake the planet, and our bodies are compositions of shifting interconnections, unbound and unhinged (Simonsen 2013).

SEE ALSO: Antiracist geography; Difference; Feminist geography; Habitus; Marxist geography; Phenomenology; Poststructuralism/poststructural geographies; Psychoanalysis/psychoanalytic geography; Queer geographies

References


Further reading


Borders, boundaries, and borderlands

David Newman
Ben-Gurion University of the Negev, Israel

Geographers have traditionally viewed boundaries as lying at the very heart of their discipline. Since geography is concerned with the study of areal and spatial differentiation, the existence of territorial boundaries is taken as normative in the sense that the compartmentalization of social, economic, and cultural spaces assumes the presence of lines that separate these spaces from each other. The geographic literature in general, and the political geographic literature in particular, is replete with the study of boundaries as a category, building on numerous boundary case studies. While the bulk of this literature has focused on the international dimension of boundaries, the existence and functions of administrative, municipal, planning, and other forms of localized boundaries have also been studied. Nonetheless, it is the international boundary that has traditionally been seen as the most distinct of geographic demarcators, separating the sovereign state from its neighbor and, as such, determining the nature of the political and economic development on either side of the boundary (Newman 2006).

Political geographers have traditionally differentiated between the concepts of borders, boundaries, and frontiers. The border is a political concept, which identifies the territorial limits of the state and beyond where movement is limited to those with the necessary permits and documents. The boundary is a looser term, which signifies the territorial margins of the state, reflecting other social and ethnic characteristics of the population on either side. Recent literature has tended to downplay any significant difference in the use of these two concepts. The frontier is a zone, rather than single line, and signifies the area on each side of a border or boundary within which human activity is impacted by the presence of the border. This impact may change significantly depending on whether the border is “open” or “closed,” or whether it is a border between friendly countries or those at war with each other. Frontiers, also known in the recent literature as “borderlands” or “transition zones,” are those areas straddling both sides of the boundary where peoples from both sides can interact as part of the border opening process, and where, over long periods of time, ethnic and political hybridity may emerge.

Prior to the 1960s, political geographers focused on a descriptive analysis and categorization of international borders and the processes through which such borders have been delimited and demarcated. It was common for scholars to typologize the different types of border, based on the way in which they had been demarcated, during a period of major reterritorialization that took place, particularly in Europe, after each of the world wars (Jones 1959; Rankin and Schofield 2004). The classic border typology that appeared in many of the introductory political geography texts was that of Richard Hartshorne. Borrowing terms from fluvial geomorphology, he classified boundaries as antecedent, subsequent, and superimposed, the last describing the many colonial boundaries that had been drawn up by European powers in their division.
of territories in Africa, Asia, and the Middle East (Hartshorne 1936).

As with political geography and geopolitics in general, the period immediately following World War II lacked any significant border analysis, notable exceptions being Pound’s work on natural boundary ideologies (Pounds 1951) or Julian Minghi’s seminal overview of global boundaries and their categorizations in the *Annals of the Association of American Geographers* in the early 1960s (Minghi 1963).

The renaissance of border studies was spurred on by the globalization narratives of the late 1980s and 1990s (Ohmae 1990; Newman and Paasi 1998; Paasi 1998), which posited a “borderless” world as a result of structural cross-border processes such as the emergence of cyberspace and the global flow of capital, along with the historical and political contingencies of the fall of the East–West divide, along with the collapse of the Berlin Wall, the expansion of the European Union, and the removal of intra-European borders between the member states. While recognizing the impact of such processes on the function and significance of contemporary borders, scholars rejected the notion of a borderless, de-compartmentalized world, and began to meet across the disciplinary boundaries to examine the border phenomenon in greater depth and to provide a counter narrative to the idealized borderless world scenario. This collaboration included a deeper understanding of border management, power relations at the border, vertical, cultural, and social borders, invisible and perceived borders, and the spaces around the border that created borderlands, zones of transition, and spaces of hybridity.

Emerging initially as a counter narrative to the globalization theses of a “borderless world,” border studies have become transformed into a distinct subdiscipline within the world of political geography in its own right (Brunet-Jailly 2005; Kolossov 2005; Paasi 2005), drawing together scholars from diverse disciplines beyond the limited scope of geographers and political scientists. This has become evident in a growing number of research networks, conferences and workshops, and publications. These include ABS (Association of Borderland Studies), IBRU (International Boundaries Research Unit), BRIT (Border Regions in Transition Network), ABORNE (African Borders Research Network), all of which hold annual conferences and workshops. In addition to an exponential growth of border-related papers that have been published in both political geography and geopolitics, the *Journal of Borderland Studies* has emerged as an important outlet for such research in recent years, moving from a case-study oriented approach to a higher number of theoretical and conceptual contributions. Major research funding projects have, during the past decade, focused on border related topics, recent examples including the European funded research consortia, such as the FP5 EUBorder Conf and the FP7 EUROBORDERscapes, the Canadian funded Borders in Globalization (BIG) project, and the Finnish Academy funded RELATE Centre of Excellence for the study of Bordering, Identities, and Transnationalization.

A disproportionate amount of recent research has focused on case studies, especially those which describe the processes through which previously closed borders have opened and across which residents of one side undertake their journey of discovery to meet the other side. The subject of cross-border regions has been the focus of intensive discussion, especially throughout the expanding spatial sphere of the European Union. The process through which borders have opened and become more porous has been accompanied by a growing interest in the reclosing of borders, which has accompanied the securitization discourses of the post-9/11
era and a return to a renewed focus on border control, management, and surveillance.

There has not been the creation of a single border theory as such, bringing together different types and scales of border, but there has been a cross-disciplinary discussion (itself an interesting example of the ways in which borders are negotiated and crossed within the academic world and through which both scholars and practitioners with different experiences meet each other in cross-disciplinary transition zones) of a common glossary of terms of relevance to all types of border scholars (Newman 2006, 2011). These, as will be discussed in the remainder of this entry, bring together borders as a functional and dynamic process, boundary producing practices discussed within critical geopolitics in the 1990s, and the process of bordering as aptly coined by one of the leading border theorists, Henk van Houtum, from the University of Nijmegen, rather than simply as a physical and unchanging spatial outcome of a political or social process (van Houtum 2005). The common use of terms, albeit in different contexts, such as demarcation and delimitation of borders, the management and control of borders, power relations, the ways in which borders are represented, and border zones as spaces as contrasted with borders as lines, have been found to constitute common terms of interest to the growing diversity of border scholars.

Themes in contemporary border studies

Opening and closing of borders: parallel discourse

As the nature of the territory–state discourse has changed in recent years, so too has the role and function of boundaries. The end of the nation–state brought with it a parallel argument relating to the disappearance of boundaries. The impact of economic globalization, the dissemination of information and knowledge through cyberspace, and the firing of ballistic missiles over long distances paying scant regard to the physical existence or location of borders have greatly reduced the significance of boundaries in their traditional function of constituting barriers to the movement of people, weapons, goods, or ideas. The opening of borders was contingent upon the structural changes of globalization in such areas as the free flow of global capital, the cross–border cyberspaces that enabled information and communication to bypass the traditional border barrier controls, along with the political and historical changes, such as the fall of the iron curtain on the one hand and the removal of internal borders within the ever-expanding European Union on the other.

Notwithstanding, a brief glance at the map of the world shows that, despite the discourse of new world orders, the basic territorial compartmentalization of the globe remains strongly based on the existing pattern of sovereign states. Reterritorialization (as contrasted with the problematic notion of deterritorialization) continues to impact the world political map, even if the functions and significances of the borders – not least the disappearance of many of the physical and visible elements of the border fence – are constantly changing. The opening of borders and the easing of cross-border physical movement in some places have brought about new suprastate and intrastate geopolitical interactions, in many cases ignoring the state altogether. But while some boundaries were opened up to movement and became more permeable, many countries have created new fences of separation in an attempt to strengthen their own senses of identity and control mechanisms, not least in areas of continued ethnic or national conflict. In short, the processes affecting borders and boundaries in the contemporary world are geographically
differentiated, opening in some places, closing in others, contingent upon the constant dynamics of political and geopolitical change.

Nowhere is this process more apparent today than in parts of the Middle East where, in addition to unresolved problems of territorial demarcation for an independent Palestinian State, much of the region has been thrown into turmoil through the emergence of the Islamic State, which controls territories beyond the boundaries that emerged almost a hundred years ago following the dissolution of the Ottoman Empire, the Sykes–Picot arrangements, and the imposition of boundaries by the European powers.

Following the events of 9/11 and the emergence of global terror and violence, there has been a move toward the resealing and reclosing of borders as a means through which states prevent “alien” and “illegal” elements from crossing into the territories and the space of the homeland. This securitization discourse is itself a counter-narrative of the open border and has been, to a limited extent, shunned by critical scholars who see the border opening process as an idealized position that reflects the “good” as contrasted with the closed and the controlled as the “bad.” Notwithstanding, the number of closed and sealed borders as reflected in the large number of new fences and walls that have been constructed by states during the past decade as a means of preventing terrorism and violence from crossing the border is of major significance. This has become even more marked in the immediate aftermath of the mass flows of refugees and migrants escaping the ravages of war and famine in the Middle East (especially Syria) and Africa, and the inability of the European Union to deal with the humanitarian problem that has emerged. New fences have been constructed, but unlike those of previous eras, they are proving to be largely ineffective in providing barriers to the mass flow of people seeking to cross the borders at all costs. In one respect, new borders are more sophisticated in terms of the highly sophisticated surveillance and control techniques, well beyond the limited capabilities of a border guard, at a specific location along the border (such as in the US or Israel and its neighbors), while in other areas (such as the outer frontiers of the EU along the Schengen line), governments are unable to deal with the sudden mass movement of refugees, again reflecting the contrasting dynamics of contemporary border management.

The use of the securitization narrative has also been manipulated by those states that desire to prevent the continued transboundary movement of illegal immigrants from poorer countries seeking employment and life improvement opportunities. While in the past states could deal with the cross-border movement of migrants, remote from the public eye and the mass media, in whatever way they chose, such is no longer the case. Human rights groups, border ethics, and a wider focus on multiculturalism have prevented states from undertaking the sort of anti-immigrant actions that were common in the past for as long as the reason for crossing borders was to seek life improvement opportunities. The use of the securitization discourse by departments of homeland security, appealing to the base instincts of the public by highlighting the physical threat from terrorism and violence if the border remains uncontrolled, has proven to be a powerful means to enforce greater surveillance and more stringent controls over those who desire to cross borders from one side to the other.

The contemporary study of borders and boundaries reflects ongoing political practice of states and lobby groups, negotiating between two parallel border narratives: that of the borderless world, which brings about the opening and
eventual removal of borders, as contrasted with that of securitization, which brings about the closing and rescaling of borders (Newman 2015). This is not a binary distinction as each has its backers, a powerful economic lobby promoting unrestricted movement of people, goods, and capital, and a powerful securitization lobby promoting a return to stringent border control and management techniques. Both are considered important for the state as they seek to promote economic transactions and flows and to reduce the security threats (real or perceived) from the emergence of global violence and fundamentalist ideologies. There is not therefore a return to the previous notions of closed boundaries protecting the homeland population from the neighboring state, but an attempt to reshape the function of borders as barriers to threats that are themselves part of the way in which globalization has impacted and enabled the dissemination of knowledge – including knowledge that assists terrorism and violence – across borders.

In the North American example, NAFTA (Northern American Free Trade Agreement) and related economic agencies promote global trade and commerce and endeavor to create porous borders for the flow of people, labor, goods, and capital, as contrasted with the role of the US Department of Homeland Security, which endeavors to retain as tight control of borders as possible. In the eyes of Homeland Security, workers crossing the border are perceived as potential terrorist threats unless and until proven otherwise and this perception often enables governments to control the flow of labor migration under the guise of securitization – a discourse more acceptable to public opinion for justifying the stringent and, in some cases, less than ethical means used by border guards and security agents to prevent people from crossing the border from the “out” territory to the “in” territory.

Border diversity and scale

Borders may have common functions but they are diverse in nature. This point is as relevant to the spatial scale of those borders that impact our daily lives as it is to the non-geographic societal borders, the vertical constructs, through which society is compartmentalized and ordered. Equally, borders do not have to be visible to be effective. The perceived borders, the lines in the imagination, which prevent us from crossing from one place or one group into another, do not require the physical dimensions of fences and walls in order for them to impact upon our daily life patterns.

The decreasing significance of territorial borders between some states has served to highlight the increased importance of spatial borders at other spatial scales of analysis, which may affect our lives in a more powerful way than do the lines that separate one state in the international system from its neighbor. The analysis of scale in border studies is of major significance. We reside within urban, municipal, and other functional spaces, which impact our daily life patterns to a much greater extent than does the occasional crossing of the border between states. The provision of public services, the payment of property taxes, the registration of our children to schools or to health providers are but a few of the major life activities that are contingent upon the arbitrary demarcation of local borders by planners, bureaucrats, and municipal officials. Once demarcated, they can be difficult to change even if the dynamics of life change the nature of the urban segregation and stratification process much more speedily than the relatively slow, albeit sudden, changing of state borders. The establishment of municipal boundary commissions, with the object of resynchronizing urban and municipal boundaries in line with residential change and metropolitan expansion, is accompanied by much political intervention and spatial
engineering and gerrymandering. The determination of electoral districts or constituencies by redrawing borders is perhaps the most blatant of all boundary demarcation exercises, reflecting spatial change on the one hand and engineering political change on the other.

Most of us live within a highly fragmented set of spatial and social compartments. Society is too complex for all of the local borders to correspond with each other. Some borders, such as those that contain the municipalities to which people pay taxes or the school districts within which they are allowed to register their children for education, are more apparent than others. People are aware of many of the borders that determine their lives only on the few occasions when they are in need of them – a police or a fire district – and when these do not correspond with the other local borders with which they are familiar, they become unable to function with the same efficiency and certainty that they do in other spheres. Thus local borders are only of importance when the life function they demarcate is necessary; however, the fact that they play no significance in people’s daily lives for most of the time does not mean that they do not exist or that they do not play a role in the way in which society is spatially compartmentalized or socially ordered.

While such borders are mostly invisible to the eye and are not reflected in walls, fences, or security guards, their impact is of major significance. Most of the world’s population have never crossed an international border and do not possess the documentation to do so. Such is not only the case of the less developed world. When, in the 2000s, as a result of increased fear within North America from the flow of terror across the country’s borders, US citizens were obliged to have passports instead of relying solely on driving licenses to cross the border to Mexico or to Canada, it emerged that no more than 15% of the country’s citizens had ever taken out a passport or moved beyond the lines separating the United States from any other country in the international system. The most significant border crossing for much of the US population is that which separates the internal states from each other within the federal system of states, which may have different rates of local taxation, laws relating to liquor, or highway speed limits. These borders are geographical in nature – they significantly impact daily life patterns, but are vastly different in the ways that they are managed and controlled as contrasted with the borders that separate the United States from its neighboring countries.

The study of scale differences is also of importance in terms of the borders that separate one individual from the other – the self-spaces that we create, or are created for us, around the body, and in our relations with our neighbors. This entry does not address this scale of border analysis, other than to remark that a deeper understanding of personal spaces also requires another layer of understanding how such borders are created (demarcated) and what mechanisms are used (border control and management) to defend such personal spaces from the intrusion of others, while allowing loved or trusted ones to enter the space of the self.

A second form of border diversity is concerned with the vertical and societal borders, not all of which are spatial or geographical although their dispersion throughout space may have major geographical implications. Many socioeconomic or cultural groups cross borders in terms of a single fixed territorial location and their boundaries as such are movable or to be found in multiple locations at one and the same time. The emergence of a cross-disciplinary interest in borders has highlighted the common functional concerns with the way that borders are demarcated and controlled, no less in the way that people are compartmentalized within social or cultural categories than they are within
geographical districts. The hegemony with which geographers traditionally understood the concept of borders as being an essential spatial concept has been replaced by an understanding that borders create order in a complex world not only between geographic spaces, but also between cultural, social, religious, and economic groups, all of which contribute to the notion of societal compartmentalization. Each has its own border, its own processes of demarcation and delimitation, means of effectively controlling and managing the border, and its crossing points. Crossing from one territory into another may be relatively easy with the problems encountered by people attempting to cross from one cultural or religious group into another and the ways in which they negotiate, through documents, customs, or even linguistic skills (the ultimate sealed border for one who does not speak the requisite language) to move beyond the border.

The division of society into cultural, religious, and economic groups, to name but a few of the more major categories, is about borders and compartments. Notions of demarcation, delimitation, crossing borders, and border management are as much part of the discourse of sociologists, political scientists, anthropologists, and economists as they are of geographers. The functional terminologies discussed in this review are common to all forms and types of border, be they geographical or cultural, visible or invisible, local or global. The crossing of disciplinary borders creates a cross-border community of ideas, which is at the core of the recent renaissance of border studies within the social sciences.

The third form of border diversity is concerned with the tangibility or invisibility of boundaries. Borders can be perceived as much as they are visible or tangible. Within urban environments, specific neighborhoods and their assumed cultural characteristics are as much about image as they are about reality. Residents do not cross beyond certain self-perceived limits, a major transportation artery, or a major road junction, assuming the people on the “other” side to be different – in terms of racial characteristics, income levels, or ethnic diversity. The imagined borders of the urban landscape are as much about the creation of borders and self-imposed prophecies as they are about the ways in which they may reflect already existing realities and difference. In this context, Stephen Graham has discussed the role of urban dividing lines and their scalar connection to wider spaces and processes (Graham 2010). The perception of borders can be much harder to change than the physical removal of boundaries. With this knowledge many groups who are in positions of power (see next section), and are able to determine the constant formation and reformation of urban landscapes for their own political or economic ends, are able to manipulate the ways in which space remains segregated, or undergoes new processes of segregation, through the demarcation of municipal boundaries and/or the manipulation of the housing market within both the public and private spheres.

Power relations at the border

A major theme within all areas of border studies concerns the nature of power relations (Newman 2003). Power relations are as relevant in terms of those who undertake the processes of demarcation and delimitation as they are of those (often the same power elites) who determine the ways in which borders are managed and controlled. The decision to “open” or “close” borders is a political decision reflecting power interests as is the decision to relocate the control of borders away from the physical location of the border – in the middle of airport terminals in a foreign country or behind the faceless desk of a visa bureaucrat half way around the
BORDERS, BOUNDARIES, AND BORDERLANDS

world. Power thus determines the changing geographical location of the border and the border guard, aided and abetted by technology that intensifies border surveillance and management rather than contribute to the weakening of the barrier functions of borders. This management is in sharp contrast to the role of technology that was highlighted in the globalization debates of the 1980s and 1990s and that assumed a direct correlation between technology and the opening of borders as part of a uni-directional process.

The very existence of borders is essentially a function of political power. All borders are politically and socially constructed. They may result from military conquest, cultural hegemony, or from bilateral agreements, depending on the extent to which power is asymmetrical and borders are either imposed by one powerful side or agreed upon by two or more sides. The demarcation of many state boundaries has been superimposed as a result of victory in times of war. Their delimitation has not taken account of the requirements (legitimate or otherwise) of the vanquished side, not least in situations where warfare has focused around contested claims for valuable natural resources within the borderland or the inclusion of ethnic groups as part of the national entity. The construction of borders may signify identity, but is more significant in the determination of citizenship. This situation also raises many critical questions concerning the ethics of border construction, superimposition, and management that have been addressed by scholars in recent years (Williams 2003; van Houtum and Boedeltje 2009).

There is no such thing as a “natural boundary” in the sense that the physical topography determines the course and the location of the boundary. But where it is convenient for the power elites to make use of natural features and where it does not conflict with their political objectives, states will use them. The use of natural features would have been common in pre-technological historical periods of reduced mobility and access from one region to another. As such, the initial location of boundaries affected the development and consolidation of state territories over time. But where states have undergone processes of reterritorialization, especially through the past century, the natural features are only used by power hegemonies to determine the course of the boundary where it serves other political purposes. In many cases, natural features, such as the location of mineral deposits or the flow of riparian waters, are precisely the factors determining their inclusion by the power elites within their own territories. In the past, such factors were often the catalysts for territorial conflict.

In situations where ethnic majorities and minorities are split across boundaries and constitute spatial minorities in neighboring states, the formal determination of citizenship is contingent upon boundaries, even where this conflicts with critical issues of national identity. The superimposition of borders upon ethnic landscapes is a function of asymmetrical power relations. The long straight geometric lines so common throughout Africa and, until recently, in much of the Middle East are the most blatant example of the use of power to superimpose borders upon a landscape. This demarcation may have resulted in the division of homogeneous ethnic territories between neighboring states, reducing relatively large ethnic groups to minority status in neighboring political entities. Probably the most significant twentieth-century example of such division resulting from a process of re-bordering is that of the Kurds. The post-World War I victorious powers, seeking to re-compartmentalize the territories that had been under the control of the Ottoman Empire, and according to the realpolitik interests of the period, drew lines in the heart of the Kurdish territory, fragmenting the region between Iran, Iraq, and Turkey, with smaller
residuals in both the Soviet Union and Syria and leaving them as ethnic minorities within those neighboring states. One hundred years after the demarcation of Middle Eastern boundaries along the principles of the Sykes–Picot agreement, with the upsurge in internal discord, the Arab “spring,” and the rise in the power of the Islamic State, the artificially constructed borders of this region are falling away. The contemporary turbulent events in the Middle East have resulted, once again, in attempts to redraw borders and enable the Kurds to establish their own independent political territories.

The same can be said for Germans during the period of the division of Germany or Palestinians following the partition of Palestine in 1948–1949. In both cases a single ethnic group was split beyond the boundaries imposed as a result of political conflict. While their cultural identity remains unchanged, their separate citizenship is determined by the border regime. Instead of constituting a majority ethnic group as part of a “nation” state, they remain ethnic (and in some cases persecuted) minorities in neighboring states, where the borders are rigidly controlled in an attempt to prevent cross-border alliances against the existing state regimes.

Control is exercised as an absolute concept throughout state territory as far as the border. This claim contrasts with normative geographical theory in many areas of economic and social development, which emphasizes distance decay as a major explanatory factor for geographical and spatial differentiation. Notions of sovereignty do not, on paper, comply with the concept of core and periphery, where the process is stronger in one place than in another. The border is the ultimate delimiter in that it separates absolute control from zero control across a relatively small geographic space. It is this function of borders that has been challenged more than any other by globalization and that may yet serve to change the essential power relations at and beyond the border.

Border superimposition has also resulted in the opposite phenomenon, namely, the inclusion of multiple ethnic groups within a single political territory, resulting in competition for power and control and, in many cases in Africa and more recently the Middle East, civil strife and warfare, ethnic expulsion, and genocide. For a continent that has had to come to terms with the violent legacy of border superimposition in regions where territorial behavior for thousands of years was not based around fixed territories with lines that could not be crossed, contemporary westernized notions of borderless worlds are as much a superimposition of political hegemonies as was the imposition of the borders in the first place. The major difference between the two periods is that it was the Western powers, external to the region, who determined the territorial map one hundred years ago, while today it is the local power elites who determine the ways in which territorial power is to be distributed.

The same powers that determine the criteria for border demarcation also determine the ways in which borders are to be managed, especially regarding who and what can cross the border from one side to the other and under what conditions. The necessary documentation in the form of passports, visas, and other forms of identity documents and permits is a powerful form of border control that enables state agencies to permit or prevent people or goods (lacking the necessary customs fees) from freely crossing in a world of fixed territories. As border smuggling has become more sophisticated and as people cross national and other borders in cyberspace without physically moving from one place to another, so too the forms of surveillance and control have become more advanced and technologically sophisticated, as reflected in a major growth in the field of research on
border surveillance, often funded by respective departments of homeland security who seek ways in which to make borders more difficult to cross and even hermetically sealed in some places. The location of the border control point has moved partly away from the border itself to places where control can be exercised at a distance—such as the checking of passports in airports many thousands of miles away from the country to where entry is sought or to an office in Washington that collects surveillance data of movement across the US–Mexico border within seconds and relays these data back to the local control centers.

As borders were initially transformed from closed (sealed) to open and porous (or vice versa) it was initially assumed that this transformation reflected a transfer of power from one elite to another, as governments with different perspectives on global geopolitics wrested power one from the other. But in retrospect, it is the same power elites who, having determined what was in their political and economic interests at a time of border construction, are able to adapt with the changing global realities and to modify their policies vis-à-vis border management as a world of flows, movement, and networks serves their interests in the same way as did a previous world of closed compartments. It is essentially another way in which power is institutionalized within existing bureaucracies and governmental elites enabling power to be maintained even in the face of volatile global and regional transformation.

Borderlands and cross-border zones

Borders have become associated with the notion of the borderland, the functional space within which development is impacted by the existence of the border or the boundary in close proximity. A borderland is defined as a region in close proximity to a political border, within which development diverges from what would have been expected under normal planning models. Such spaces are not symmetrical on each side of the border and functions may extend more on one side of the border than the other, contingent upon both the nature of the border itself, and the government policies and investment within the border zone. In cases of cross-border tensions and animosities, where borders remain largely closed to the movement of people and goods from one side to the other, such policies can vary from extensive investment and development as a means of ensuring a bolstered presence along the border or a lack of resources in a region that is considered potentially at risk of conflict and potential destruction.

The literature has used different terminologies to describe cross-border spaces, including political frontiers, borderlands, cross-border regions, and border zones. The common point about all of these terms is that they focus on an area or space as contrasted with a single line. The extent of contact or meeting across the border line determines the extent to which the border functions as a barrier to movement or as a place where meeting takes place and, in some cases, where cultural and social hybridity is created. Cross-border spaces are particularly effective in regions where a process of border opening is taking place and where the initial cross-border contacts can take place at grassroots levels of mundane daily activities such as food provision, tourism, and commerce.

As borders in many parts of the world have become increasingly porous, the borderland has become transformed into a zone of transition where peoples and cultures, previously separated from each other, come into contact and create a place of meeting. Cultural transition between two distinct peoples or cultures may take place. In some cases there may even be cultural hybridity as the previous distance decay between state
cores and peripheries is replaced with a transition from one culture to another. The European Union in particular has created transboundary or cross-border regions within the borderlands as a means of easing the border opening process and creating places of meeting and familiarity where there had previously been places of separation and ignorance of what takes place on the “other” side. These policies have eased the eventual border opening or erasure as new countries become members of the EU, enabling free movement from one state territory to another.

Cyberspace and satellite communications have contributed a great deal to the reduction of cross-border animosities, as they depict the cross-border “other” as encountering the same daily life activities and problems. Such depictions may contrast with the governmentally constructed images of the cross-border “other” as constituting a threat from which the national self has to be protected through the construction of walls and fences and through stringent processes of border management. Cyberspace and globalization have contributed to a rethinking of the border phenomenon as borders that cannot be crossed on the ground have effectively been crossed on the internet and have removed the sense of threat and fear that traditionally emanated from the other side of the border.

Equally, the crossing of borders in cyberspace, has also brought about situations in which threat can be disseminated across boundaries that were previously closed. This is especially true of the dissemination of national ideologies, strengthening the role of diaspora groups and their attachment to their former “homelands,” even for those who are already second and third generations removed from their ancestral homelands. Such is also the case with the spread of global terror during the past fifteen years. The cross-border dissemination of messages, slogans, and information encouraging violence and anti-state behavior is enabled precisely through the use and manipulation of globalization technologies as state boundaries are no longer effective in preventing such flows of information from taking place. Thus, local populations are encouraged to take part in such activities (such as British citizens who were involved in the 7/7 bombings in London), rather than focusing on migrants who have physically crossed the border and have traditionally been perceived as the instigators of internal instability.

The existence of border regions in which the border can be crossed as part of normal life also highlights the changing understanding of cross-border difference. Much recent research in the field of border studies has focused on the changing mechanics and dynamics of the border crossing process. A misconception concerning the opening and crossing of borders was that as borders open, so difference is gradually erased and disappears. But the numerous narratives that accompany cross-border experiences strengthen, rather than diminish, the notion of difference on two sides of a border – in terms of the price of goods, the nature of the cuisine, the politeness (or rudeness) of the people, the language spoken, and so on. But the opening of the border enables such difference to be enjoyed and celebrated rather than constituting the underlying reason for threat. It strengthens the notion of the border as a line of cultural division or separation, but one that exists within a multicultural world where the previous binaries of “self” and “other” are to be shared and enjoyed, rather than feared out of a lack of knowledge and invisibility put in place by a sealed and impenetrable border. Difference that has existed across the border for hundreds of years will not disappear overnight as borders become easier to cross, but as borders become more porous and people cross them on a regular basis, difference no longer necessarily constitutes threat.
The celebration of difference within cross-border regions has also added a new dimension to the growing interest in border tourism. Both closed and open borders are a magnet for interested tourists. They are as eager to see and be photographed at the DMZ (demilitarized zone) in Korea or the points of transit between Israel and the West Bank as they are to experience different cultures and customs on each side of the border. Checkpoint Charlie in Berlin has become a major attraction for tourists, as have the few remaining border relics left by governments at former borders and around which new informal tourism commerce has emerged. There is a great potential for further border tourism and governments are eager to cash in on this new border phenomenon – at places of conflict as at places of peace.

Border representations and images

Much of the recent literature on geopolitics in general, and border studies in particular, has discussed the ways in which borders are represented in a variety of images, ranging from traditional sources such as cartography and paintings to additional media such as film, literature, caricatures, graffiti on walls, on the web, in Microsoft and other cyber products, and even poetry (Dell’Agnese and Amilhat-Szary 2015).

Images and representations of borders often reflect grassroots understandings of the impact and significance of borders. Maps and cartography were used in the past to create hegemonic views of the territorial configurations of state territories and the notion of the “correct” border in the eyes of the political elites and in the process of territorial socialization, especially in situations of territorial conflict between countries. Many caricatures of borders reflect the construction of walls and fences, rather than their removal, indicating the visual impact that such borders have on the public at large. The inability to move or communicate across borders is a theme often highlighted in films, as too the difficulties encountered when trying to cross closed or sealed borders. Smuggling across borders, illegal migration, or the crossing of the border from a dangerous to a safe space (including that of fugitives) are other common themes replayed in different geographical and political situations. Literature often focuses on the nature of difference expected to be encountered on either side of a closed border or, alternately, the discovery that perceived differences are not as threatening as originally thought when border crossings do take place. This topic offers much scope for future research and understanding of the significance of borders as they impact daily lives of both individuals and nations.

A Border Poem

Paul Muldoon (2001)

Boundary Commission

Your remember that village where the border ran down the middle of the street with the butcher and the baker in different states?

Today he remarked how a shower of rain had stopped so cleanly across Golightly’s Lane, it might have been a shower of glass that had toppled over.

He stood there, for ages, to wonder which side, if any, he should be on.

See Also: Frontiers; Nation-state; State, the

References

BORDERS, BOUNDARIES, AND BORDERLANDS


Further reading


Boreal forest ecosystems

E.A. Johnson
Y.E. Martin
University of Calgary, Canada

The boreal forest is circumpolar (Figure 1) with two continuous parts, one in North America and the other in Eurasia. The two parts have similar tree genera and physiognomy. A few genera of both gymnosperms and angiosperms are common throughout the boreal (Table 1). Only a few tree species are strictly boreal and none are circumpolar. In species richness Asia has the most species, Europe the least, and North America is intermediate. The conifer physiognomy gives the boreal its distinctive characteristics. The orthotropic, monopodial trunks and plagiotropic branches help the trees shed snow because of their shape and flexible branches. The display of the needles along the tree trunk and around the branches is conducive to light capture at lower sun angles and the evergreen needles allow longer growing seasons and nutrient conservation.

The boreal flora is believed to be descended from mid- and late tertiary mountain forests. The flora and vegetation were established by the late Pliocene. The Pleistocene saw at least eight glacial–interglacial periods. In North America, only two possible boreal areas were not repeatedly covered by continental glaciation. However, there are several cryptic glacial refugia proposed based on genomic information. Genomics has advanced our understanding of possible centers in which populations survived glaciaation and spread. In both North America and Europe, populations closer to the furthest advance of the glaciers have been the most successful in the present boreal forest (e.g., Provan and Bennett 2008). Consequently, plants and animals had to migrate south as the ice advanced and then reinvaded during interglacials. Most of the boreal species survived quite a distance south of where they exist today.

One of the important findings of paleoecology has been how easily and often ecosystems can be reassembled. This is particularly true of the boreal ecosystems which have been repeatedly removed and reassembled from various surviving areas. The conclusion of Davis in 1979 still holds: interglacials were not long enough for plants and animals to have reached the extent of their ranges in which they could have survived, ecosystems throughout interglacials were constantly being invaded by new species, and modern species evolved in the biotic and abiotic instabilities of the Pleistocene. Therefore the boreal forest we see today is a melange of histories and conditions.

Until the middle of the last century the boreal was known best around its southern edges. Access further north was by water in summer and dog team in winter. There were and still are large areas without roads although planes, particularly small float planes, have made access better. The southern boreal was the site of optimistic settlement for agriculture from the end of the 1800s into the 1920s (Johnson and Miyanishi 2012). This settlement caused serious forest fragmentation and forest clearance caused wildfires which spread into nonagricultural lands. Thus, the southern edge of the boreal, although often
seen as natural, has an imprint of past land use. Scientific studies beyond observation and plant and animal collections only began in the 1930s, and from the 1950s to the 1990s the scientific research was primarily devoted to forestry and community ecology. In the 1980s, global climate model studies showed that higher northern latitudes would experience the most climate
Table 1  Boreal forest dominant trees: examples of “Rassendreise” (circles of species), series of allopatric (replacing) “species,” which perhaps are in many cases nothing but slightly differentiated geographical races.

<table>
<thead>
<tr>
<th>Smoky Mountains</th>
<th>North America</th>
<th>Rockies Mountains</th>
<th>East Asia</th>
<th>Europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abies fraseri</td>
<td>A. balsamea</td>
<td>A. lasiocarpa</td>
<td>A. sibirica</td>
<td>A. alba</td>
</tr>
<tr>
<td>Picea rubens</td>
<td>P. glauca</td>
<td>P. engelmannii</td>
<td>P. obovata</td>
<td>P. abies (excelsa)</td>
</tr>
<tr>
<td>Pinus spp.</td>
<td>P. banksiana</td>
<td>P. contorta</td>
<td>P. sylvestris</td>
<td>P. sylvestris and spp.</td>
</tr>
<tr>
<td></td>
<td>Larix laricina</td>
<td>L. occidentalis</td>
<td>P. pumila and spp.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L. sibirica</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L. dahurica</td>
<td></td>
</tr>
<tr>
<td>Populus tremuloides</td>
<td>P. tremuloides</td>
<td>P. tremuloides</td>
<td>P. davidiana</td>
<td>P. tremula</td>
</tr>
<tr>
<td>Betula papyrifera var.</td>
<td>B. papyrifera</td>
<td>B. papyrifera var.</td>
<td>B. mandshurica</td>
<td>B. pendula</td>
</tr>
<tr>
<td>Sorbus americana</td>
<td>S. americana</td>
<td>S. scopulina</td>
<td>S. japonica</td>
<td>S. aucuparia</td>
</tr>
<tr>
<td></td>
<td>S. decora</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

warming. Further, the boreal forest accounts for one-third of the global carbon. This has resulted in a large number of carbon studies by individual researchers and large programs, for example, NASA-BOREAS.

The boreal has been aptly called a snow forest, but the zonation of vegetation is not related to winter hardiness. The boreal’s northern border is tundra while its southern borders can be semidesert, steppes, grassland, and conifer-deciduous forests. In North America the southern boundary of the boreal in the west is determined by the modal position of the arctic air mass in winter (Pielke and Vidale 1995). Support for this idea can be found in a physical model of heat and moisture limitation on reproduction of different vegetation types (e.g., Arris and Eagleson 1994). In eastern North America the boundary between the boreal and the deciduous forest is a transition zone of conifers and northern hardwoods. The northern boundary of the boreal has been attributed to the summer position of the arctic front but also to the interaction with the vegetated terrain (Bonan, Chapin, and Thompson 1995).

The boreal terrain in North America is largely glacial and mostly flat with incomplete drainage, slow moving waters, and numerous lakes and peatlands. The glacial deposits cover metamorphic and igneous rock on the Precambrian Shield. The southern edge of the Shield is covered by sedimentary rocks. The interior of Alaska has an assortment of terranes of metamorphic and sedimentary bedrock. The interior of Alaska was not completely glaciated but was affected by periglacial processes. These glacial processes explain most of the landscape forms and hydrological properties. Substrate can be divided roughly into till sediment deposed directly by the glacial ice, glaciofluvial, glaciolacustrine, and diamicton (poorly sorted and stratified material of glacial origin of a mixture of processes). Hydrological budgets usually depend on topographically defined watersheds which have
BOREAL FOREST ECOSYSTEMS

developed in landscapes that have not had long enough to come into approximate geomorphic and tectonic equilibrium. Also, the development of extensive peatlands in the last 3000 or 4000 years further disrupted the drainage systems. Surface and groundwater interact, creating complicated patterns so that groundwater does not always follow the surface topography. Buttle (2006) suggests three ways that the landscape template affects stream flow: typology defines the relative importance of vertical and lateral flow of water, topography defines the hydraulic gradient in a watershed, and topology defines the drainage network connectivity.

Overall, the water budget of boreal landscapes depends on the seasonal distribution of precipitation and evapotranspiration, the storage capacity of landscape units, permafrost, and the length of the season the water is not frozen. The boreal landscape is a series of hillslopes (uplands) and valleys in which some of these valleys will not have permanent streams. Many valleys and hillslopes have very little gradient and often are covered by peatlands (see below). The depth of the bedrock, glacial material, and peat will determine storage capacity and the spatial arrangement of the hydrological system. The runoff is further determined by a fill and spill system (Spence and Woo 2003) in which surface and subsurface water flows to the next downslope hydrological unit only after its storage capacity is filled. One can now see the importance of the landscape and seasonal distribution of precipitation and evapotranspiration. Table 2 gives a hierarchical arrangement of controlling factors showing how precipitation and potential evapotranspiration influence other factors. In years when \( P > PET \) the fill and spill connects runoff between hillslopes, valleys, and peatlands. But in times when \( P < PET \), hillslopes, valleys, and peatlands will be disconnected and runoff will be controlled by local gradients and storage. Permafrost reduces runoff by reducing the storage capacity, and in areas of incomplete coverage of permafrost the spatial differences in heat budgets play an important part in the distribution of permafrost in peatlands (e.g., peat plateaus) and hillslopes (fine textured soils on north facing slopes).

The spatial distribution of vegetation composition can be best understood by using environmental gradients once the time since disturbance has been removed. The moisture-nutrient gradients have proven to be the most important, followed by heat budgets. The latter is particularly true for those locations with upland permafrost. Moisture-nutrient gradients are most important because all landscapes consist of hillslopes which connect ridgelines and valleys even in the relatively low relief of most boreal regions. Consequently, since water runs downhill, the flow lines on hillslopes create, all else being equal, moisture gradients from drier ridges to wetter valleys. Nutrients are also carried in solution down these slopes. Figure 2 gives an example of the upland gradients in the south-central boreal forests of North America.

Peatlands are one of the distinctive features on both uplands and lowlands in the boreal. They come in many forms and the classification is sometimes confused because of different terms for the same feature. The principal environmental gradients are determined by the source of the input water and its mineral composition. Ombrotrophic peatlands, for example, bogs, are isolated from the mineral-rich ground and surface water. They get their mineral elements from precipitation. At the other extreme are minerotrophic peatlands, for example, fens and marshes, which are connected to mineral-rich ground and surface waters. Ombrotrophic peatlands are usually dominated by sphagnum and fens by feather moss (Amblystegiaceae family).
Table 2  Hierarchical classification to generalize the dominant controls on water cycling and indices to define effective hydrologic response units.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Range of factor</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate</td>
<td>Dry, arid to subhumid (P &lt; PET)</td>
<td>Wet, humid (P &gt; PET)</td>
</tr>
<tr>
<td></td>
<td>• R poorly correlated with P</td>
<td>• R closely correlated with P</td>
</tr>
<tr>
<td></td>
<td>• storage or uptake dominates</td>
<td>• runoff dominates tendency for lateral flow</td>
</tr>
<tr>
<td></td>
<td>• tendency for vertical flow</td>
<td></td>
</tr>
<tr>
<td>Bedrock geology</td>
<td>Permeable bedrock</td>
<td>Impermeable bedrock</td>
</tr>
<tr>
<td></td>
<td>• intermediate to regional flow systems</td>
<td>• characterized by local to intermediate flow systems</td>
</tr>
<tr>
<td></td>
<td>• lack of topographic control on direction of local flow</td>
<td>• topographic control on direction of local flow</td>
</tr>
<tr>
<td></td>
<td>• vertical flow dominates in surface substrate</td>
<td>• lateral flow dominates in surface substrate</td>
</tr>
<tr>
<td></td>
<td>Bedrock slope perpendicular to land surface</td>
<td>Bedrock slope parallel to land surface</td>
</tr>
<tr>
<td></td>
<td>• complex watershed boundaries</td>
<td>• simple watershed boundaries</td>
</tr>
<tr>
<td></td>
<td>• regional aquifer definition needed to determine flow direction</td>
<td></td>
</tr>
<tr>
<td>Surficial geology</td>
<td>Deep substrates</td>
<td>Shallow substrates</td>
</tr>
<tr>
<td></td>
<td>• intermediate to regional flow</td>
<td>• local flow most probable (but see bedrock geology)</td>
</tr>
<tr>
<td></td>
<td>Coarse texture</td>
<td>Finer texture</td>
</tr>
<tr>
<td></td>
<td>• vertical flow</td>
<td>• lateral flow</td>
</tr>
<tr>
<td></td>
<td>• deeper subsurface flow</td>
<td>• depression storage and/or surface and shallow subsurface flow</td>
</tr>
<tr>
<td></td>
<td>Spatially heterogeneous deposits</td>
<td>Spatially homogeneous deposits</td>
</tr>
<tr>
<td></td>
<td>• complex groundwater flow systems</td>
<td>• simple groundwater flow systems</td>
</tr>
<tr>
<td></td>
<td>• groundwater flow modeling important</td>
<td>• surface flow modeling important</td>
</tr>
</tbody>
</table>

(Continued opposite)
### Table 2  (Continued)

<table>
<thead>
<tr>
<th>Factor</th>
<th>Range of factor</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil type and depth</td>
<td>Upland mineral soils</td>
<td>Lowland organic soils</td>
</tr>
<tr>
<td></td>
<td>• subsurface flow dominates</td>
<td>• return flow and surface overland flow pathways dominate</td>
</tr>
<tr>
<td></td>
<td>• slow flow generation (matrix flow)</td>
<td>• quick flow generation (return flow saturation overland flow)</td>
</tr>
<tr>
<td>Storage</td>
<td>• deeper soils with large water storage potential</td>
<td></td>
</tr>
<tr>
<td>Transpiration</td>
<td>• deep roots access stored water</td>
<td>• shallower roots limit access to stored water</td>
</tr>
<tr>
<td></td>
<td>• ( P \approx AET ) during dry periods</td>
<td>• AET &lt; PET during dry periods</td>
</tr>
<tr>
<td>Topography and drainage</td>
<td>Gentle slopes</td>
<td>Steep slopes</td>
</tr>
<tr>
<td>network</td>
<td>• disorganized, inefficient drainage network</td>
<td>• organized, efficient drainage network</td>
</tr>
<tr>
<td></td>
<td>• large groundwater recharge</td>
<td>• small groundwater recharge</td>
</tr>
<tr>
<td></td>
<td>• small, variable runoff yield</td>
<td>• large, uniform runoff yield</td>
</tr>
</tbody>
</table>

\( P = \) precipitation  
\( AET = \) actual evapotranspiration  
\( PET = \) potential evapotranspiration.

Source: Devito et al. (2005). Reproduced by permission of John Wiley & Sons, Ltd.

Swamps are forested usually by black spruce (\textit{Picea mariana}) in North America. Paludification is common in the boreal. Paludified peatlands have increased the water table and allow peat to expand onto the uplands. Continuous and discontinuous permafrost produce a series of distinctive peat features: pulsas, peat plateaus, and polygonal peat plateaus. In North America permafrost peatland features seem to be related to the annual –1°C isotherm, suggesting a heat budget reason (Vitt, Halsey, and Zoltai 1994).

Hugh Raup pointed out in his 1940 \textit{Botanical Reviews} articles of the boreal that the idea of succession developed in temperate grasslands and deciduous forests did not work well in the boreal forest. It seemed, when using the traditional idea of succession, that most of the boreal forests were still early successional. The missing ingredient was an understanding of natural disturbances, mainly insect outbreaks and wildfires caused by lightning. Traditional upland boreal succession considered disturbance as either rare or a result of humans. When disturbance did
Figure 2  General gradient of moisture and nutrients on glaciofluvial and till hillslopes. Only the dominant trees are shown on the hillslope and steepness reflects the allometric equation of the specific substrate slopes. Source: Bridge and Johnson (2000). Reproduced with permission from John Wiley & Sons.
BOREAL FOREST ECOSYSTEMS

occur, succession resulted first in the invasion of herbs and shrubs, then fast-growing, shade intolerant trees (e.g., aspen (*Populus tremuloides*), jack pine (*Pinus banksiana*), and black spruce), followed by more slow-growing, shade tolerant trees such as white spruce (*Picea glauca*). The last stage was believed to be self-replacing. This is somewhat of a cartoon version of succession but gives the essence. Early studies of disturbance gave a descriptive chronology of the stages of recovery. Gradually, ecologists like Raup began to question and determine the disturbance frequency (or its inverse the return times). This led to a very important change in viewpoint, allowing disturbance to be seen not as an accident, but as a process that accounted for the survivorship distribution of the forested landscape. The disturbances were the result of both physical and biotic processes. Consequently, the connection of the disturbance process to the ecological processes became necessary. With this new viewpoint the changes in the boreal forest could be understood in a more in-depth manner. This new viewpoint also revealed more clearly the unrealistic assumptions made in the principal method used in determining traditional succession – the chronosequence (Johnson and Miyanishi 2008).

Incorporating disturbance requires that more detail of how a specific disturbance and ecological processes are connected to each other. One way of showing this is to start with wildfire and population dynamics of trees. Boreal fires are large size and high intensity (>4000 kWm\(^{-1}\) of fire front). Fires of these high intensities are crown fires, that is, fires burning on the ground and in the tree canopy. Wildfires are the result of forest structure and weather/climate. The forest structure creates a large amount of fine canopy fuels accessible by moderate flame height in surface fires. This is a result of canopies that do not overlap, with needles and leaves that are relatively small and thus are borne on small branches along the length of the branch and most of the trunk. The weather/climate role is in creating hot dry weather as a result of persistent positive mid-tropospheric anomaly (blocking high-pressure systems) caused by upstream negative anomalies usually over an ocean. Wildfires kill most of the trees by consuming the needles, leaves, and small branches, killing the trunk cambium and/or collapsing xylem. Thus, the fire creates an open environment with no surviving canopy trees.

Fire also consumes the forest floor organic F and H layers by smoldering combustion, leaving the mineral soil or a thin H layer exposed. This exposed surface has a good water budget for germinating seeds and is relatively free of competition from herbs and shrubs. Dispersal of seeds onto this surface depends on the season of the fire and how mature the seeds are in the cones at the time of the fire. In the case of serotinous cones, for example, on jack pine and black spruce, season is not important because the trees contain approximately 5 years of viable seeds in their closed cones. In non-serotinous, non-sprouting trees, for example, white spruce, the later the season of the fire the more likely the seed will survive and disperse because the seeds are already viable and the cones are still wet enough to reduce the heat flux into the cone from the fire (Michaletz et al. 2013). After germination or sprouting, as in the case of aspen, trees are either fast-growing and shade intolerant species or slow-growing and more shade tolerant species. Both of these tree types establish at the same time immediately after the fire, but their different growth rates give the impression of different ages by their different heights. This was the basis of much of the early observations of succession. Once trees are large enough to compete with each other the mortality rate increases, particularly among the smaller and
more shade-tolerant individuals. Wildfires may, of course, occur during this time, which will kill all or most of the trees. The fire frequency distribution has an average frequency of approximately 150 to 200 years, much less than the maximum age boreal trees can survive. The result is that the cohort that established immediately after the fire has a low probably of reaching old age before another fire. The cohorts that established after the immediate post-fire cohort have high mortality rates, very slow growth rates, and, excluding large openings (see below) in the canopy, have little chance of reaching the canopy and producing cones before the next fire.

Spruce budworm (*Choristoneura fumiferana*, Clem) is the most widespread endemic defoliating insect on spruce and fir in the North American boreal forest. They have widespread outbreaks every 25 to 40 years and are as important as wildfire in the eastern North American boreal forest. This disturbance is interesting in that it shows not only interaction between host and nonhost trees on the forest composition but also interactions between the budworm and its natural predators (Cook, Nealis, and Regniere 2007). Mature white spruce and balsam fir (*Abies balsamifera*) are preferred hosts of spruce budworm but black spruce can also be seriously affected. The outbreaks can be slow and diffuse or occur rapidly and synchronously over large areas. The twentieth century outbreaks have been large and synchronous compared to the nineteenth-century outbreaks. Suggested explanations for this have been natural changes in fire frequency and changes in land use, particularly in the eastern and southern boreal forest. The composition of the forest at the time of the outbreak plays an important role in future dynamics. To understand this we must remember that each species has different individual growth rates and that each species’ age distribution consists of different age (not size) classes which are (birth) cohorts. Cohorts, even for the same species, have different recruitment and mortality schedules, particularly for cohorts that developed in the open as opposed to those that developed under canopies. Shade tolerant species are able to grow slowly under canopies and, when a large enough opening occurs, can grow into the canopy. However, the opening must be large enough for adequate light and heat budget. The latter is quite important in the boreal forest where canopy openings can be “cold holes” and fill up with snow that melts slowly in the spring. Also, the individual must not be too old or have been damaged (e.g., by browsing) in which case they may not be able to change their growth rate to grow into the canopy. Canopies dominated by balsam fir are the most susceptible to beetles. Understory balsam fir cohorts often have seedlings that vary in age from the current year up to decades old. When the beetles kill the canopy trees some of these seedlings will grow to form a new canopy. This can occur even with partially killed canopies. Thus, balsam fir can regenerate its canopy from a “seedling” bank in its understory. Although black spruce is not as susceptible to spruce budworm as balsam fir, it can still suffer considerable damage. Its dynamics after spruce budworm-caused canopy removal involves both understory seedlings as in balsam fir but also layering. Both black spruce and balsam fir stands can also be susceptible to wildfire. Pine stands, however, are not susceptible to spruce budworm.

Industrial forestry occurs mostly in the southern parts of the boreal forest on productive sites. Starting in the 1950s in North America, sustainable forestry involved change from saw timber to both saw timber and pulp wood. In the 1980s with understanding of forest dynamics, the role of natural disturbances, wildlife ecology, biogeochemistry, and other ecosystem processes, forestry adopted a strategy of using natural
disturbances as a model to be mimicked in forest operations (Burton et al. 2003).

We now come to the elephant in the room. What is going to happen in the boreal forest if increased warming by CO₂ and other greenhouse gases continues to rise; or even more moderately, what happens in roughly 100 years if the present level of 400 ppm CO₂ stays the same? Just saying it will get warmer and wetter or drier does not tell us how the boreal ecosystems will respond.

First, we know from the past that the boreal forest has changed and rearranged its composition at the scale of hundreds to thousands of years. We also know that the landscape has been glaciated and re-invaded by organisms many times and that even in the last 12 000 years the postglacial environment has changed due to erosion and weathering in which organisms have played a significant role. Also, in the last several thousand years, green and brown mosses have created extensive peatlands, storing massive amounts of carbon and changing the hydrology. Finally, in the thousand or so years since deglaciation the megafauna that once inhabited the boreal forest and had major effects on the vegetation went extinct.

Second, we know from satellites that the boreal spring growing season has lengthened from 1982 to 2011 by one day per decade and that the amplitude of the annual CO₂ cycle increased significantly (Graven et al. 2013).

Third, the lengthening of the summer warm season will lead to the southern edge of the boreal forest being replaced by more southern ecosystems, for example, grasslands and deciduous forests. These forests often have lower carbon sequestration than the present boreal forest.

Fourth, large areas in the northern part of the boreal forest next to the tundra are developing a shrub-tundra instead of woodland or forest.

Fifth, it is not possible at present to evaluate the genetic changes that will occur to allow adaptation of animals and plants to the rapid rate of climate change. Most species’ predicted range shifts are based on the assumption that plants dispersing into new areas are already adapted (i.e., they follow the climate) and that previous local environmental adaptations will not hinder their adaptation to the new area. What we still lack is an understanding of the pattern of genetic variation and the importance of strong local natural selection and gene flow which may or may not allow plants to grow or survive rapid climate change. Davis and Shaw (2001) pointed out that plants at the southern edge of the boreal forest will generally not find pre-adapted populations further south which can move north.

These and many other effects will continue to emerge as a result of global warming. One thing is certain: in many of our lifetimes we will observe the rearrangement of the boreal forest yet again but with the forcings much faster than has happened in the last million years. Unlike ecosystems further south, the boreal forest does not have a lot of agriculture, industrial and social infrastructure, or populations to be disrupted. However, the human populations of the boreal forest are largely indigenous and have in the last few centuries been badly affected by western civilization.

SEE ALSO: Biodiversity; Climate change and biogeography; Polar region ecosystems; Temperate forest ecosystems

References


Further reading


Branch plant economy

Nicholas A. Phelps
University College London, UK

The branch plant or branch factory syndrome became fully apparent after World War II as the United Kingdom attempted to steer “new” mobile industry to old industrial regions, though those terms were not to enter the economic geography, regional economics, and regional studies literature until some time later. The managers of the industrial estates developed to attract new industry to the assisted areas were clearly aware of some of the frailties of the new industrial ensembles being created. Far from embodying new industrial districts of the sort memorably described by the economist Alfred Marshall and that had been associated with the rise of the United Kingdom as an industrial nation, they were seen in a large number of studies to have contributed to the assembly of random assortments of enterprises and industries and, as a result, the further economic decline of old industrial regions.

The branch plant/branch factory syndrome became academic theoretical shorthand for what in reality is a set of constantly shifting characteristics of individual productive investments, deserving of constant empirical reappraisal by scholars and practitioners. These characteristics, in turn, are related to evolving wider intra- and interfirm divisions of labor in and between national economies. The term “branch plant economy” grew out of a specific time and specific places, and an associated intellectual and academic theoretical context in which anxieties over the “deindustrialization” of national economies were very apparent. As such, we might question its value as a potentially rather static simplification of the reality and variety of regional economic development prospects – a simplification that is overly imbued with a sense of the territoriality of the production of specifically manufactured items in urban and regional economies experiencing long-term decline. However, the core concerns associated with its use also have been apparent in developing countries (notably in Latin America) concerned with dependent development. It is a concept that also appears to be salient to the organization of the production of services. Some of the insights associated with it may yet have their day even in the BRIC (Brazil, Russia, India, and China) countries, given their seemingly inexorable rise in economic terms. In this entry it is argued that, despite its very particular geohistorical origins, the branch plant economy concept retains a relevance for some localities and industries today and will continue to do so in the future.

If the branch plant economy is an (economic) geographical concept with significant question marks hanging over it in some respects, one purpose of this entry is to explore the continued salience of the term and, perhaps more importantly, associated insights and policy concerns. This entry begins by outlining the origins of the term and its key features as enumerated in many empirical studies and theoretical interpretations. It then explores its relevance outside the geohistorical context in which it arose. It is argued that it can still resonate with the experiences of people and institutions in many urban and regional economies hosting the branches of domestic companies and the subsidiaries of,
and contract manufacturers for, multinational companies. Following the example of literature on processes of agglomeration or clustering, the branch plant economy can usefully be updated and re-evaluated in light of changing circumstances. Here some of the key challenges involve questions of the time period over which productive investments contribute to urban and regional development, and the changing balance between intra- and interfirm divisions of labor and their ramifications for territorial economic development. The challenge theoretically – as Massey (1984) laid out some time ago – is to analyze territorial economic development in terms of the contributions of both place-bound and place-transcending economic processes.

The branch plant economy

The earliest insights regarding the branch plant economy emerged very early within the history of postwar “regional policy” in the United Kingdom. The UK government had experimented as early as the 1930s with regional policy in the form of major new industrial estates in which to house new industry in old industrial regions – for example, Trafford Park (Manchester), Team Valley (Gateshead), and Treforest (South Wales). Regional policy became more concerted after World War II, with the broad thrust of the influential 1940 Royal Commission on the Distribution of Industry (“Barlow Report”), enshrined in a later Act of Parliament. New towns and new industrial estates were developed in an attempt to deconcentrate and decentralize population and economic activity from London. Restrictions on factory development in London were combined with generous incentives for “mobile” or “footloose” industry to relocate to older industrial regions (Phelps 2009). While public policy focused on the location of manufacturing industry and its potential contribution to the United Kingdom’s regional problem, the same approach was taken somewhat later with respect to service sector “factories” – so-called back office functions. Indeed, the national government itself pioneered office decentralization to cities and regions with weak job creation, leaving a legacy in which a reliance of some local economies on public sector jobs was argued to crowd out the private sector. In the southeast of England, the Location of Offices Bureau played a significant role in steering new and expanded office development out of central London. These back offices were the service sector equivalents of branch plants, but they were only to become a more significant element of the branch plant economies in the United Kingdom, and to some extent overseas, by the 1990s, as discussed below.

These various branch facilities were sought partly because of the major direct employment contributions they made to local economies. In the case of labor-intensive manufacturing and service operations, such branches might employ several thousand people at one site. However, partly as a result of the restrictions placed on and incentives available to such “mobile” or “footloose” factories, much productive investment was in effect artificially split away from the main operations of primarily UK companies – hence the name “branch factory” or “plant.” After this time, as overseas (primarily, though not exclusively, United States) investors began to expand in greater numbers into the UK economy, the reference of the term would be extended from branch factories of multiplant domestic companies to encapsulate the overseas subsidiaries of multinational companies. This brought an appreciation of the problems of remote or external control of investment, resulting in the growth potential of branch factories being effectively attributed to decisions made...
by parent companies headquartered elsewhere, often overseas (Firn 1975).

Managers of postwar industrial estates rapidly had become aware that those footloose industries that had been attracted were diverse, to the point of having no commonalities and few local linkages with each other. The logic of regional policy was to diversify urban and regional economies in light of the perceived weaknesses of overspecialization. Yet, far from promoting the creation of new Marshallian industrial districts, these industrial estates were proving to be diverse collections of industries that were unrelated both to each other and to the existing industrial structure of the urban and regional economies in which they were situated (Phelps 2009).

Empirical studies of the branch plant economy

By the end of the 1970s, the largely empirical literature could be summarized as having centered on several propositions or suspicions regarding the potential shortcomings of the branch plant economies.

- The weak linkages, both forward and especially backward, of branch plants into the local economy. While some linkage “adjustment” was thought likely to take place as branch factories adapted to their new host economies, local backward linkages remained limited due to intermediate inputs being sourced from elsewhere within the parent company (Phelps 1993).

- Branch plants provided mainly routine, low-paid jobs. Branch factories were regarded as engaged in “screwdriver” assembly, in which mass production principles involving the minute division of labor were exploited to the extreme in order to manufacture standardized products at scale. The employment opportunities required minimal skills or even were associated with the active deskilling of workforces, leading to lower levels of remuneration. Part of the question regarding pay also centered on the fuller incorporation of women into the labor markets of older industrial region economies, for the first time on a massive scale (Massey 1984).

- Branch factories were a source of unstable employment. The dependence of branch factories on incentives triggered the idea that they were only marginal investments from the outset, rarely profitable on their own terms, and more than likely providing short-term additional production capacity. Some investors gained a reputation for being serial relocators farming available incentives, leading to the suspicion of branch factories being grant “snatchers” rather than the desired “stickers” likely to produce longer-term employment stability.

- By definition, branch factories were shorn of certain corporate functions, thereby contributing to a gradual denudation of the occupational profiles of their host regional economies. Branch factories certainly lacked conception functions such as research and development (R&D), but also other corporate functions or white-collar occupations such as management, sales, and marketing (Firn 1975). This created a “truncated” occupational structure in older industrial host region economies, with head office functions concentrated elsewhere in national urban systems.

- Branch plants were considered typically to be involved with the production of mature/standardized products, using old process technology and techniques. The common assumption was that they were manufacturing products at the end of the life cycle, using tried-and-tested process
technology requiring little in the way of supporting innovative activity. The product cycle (Vernon 1966) was a key theory used to explain the decentralization of production within metropolitan regions, from Rust Belt to Sun Belt states within the United States, and from London and the southeast of England to the older industrial regions in the United Kingdom. The growth of external control, via the acquisition of indigenous enterprise in some urban and regional economies, represented one route through which rationalization and technological asset-stripping further denuded the innovative capacity of older industrial regions.

Much of this empirical evidence was based on a quite limited and specific set of older industrial region contexts, notably Wales, Scotland, and the northeast and northwest of England in the United Kingdom, and the cities of the industrial Midwest of the United States. In the United Kingdom, these regional economies diversified their industrial complexion considerably over the twentieth century, yet continued to face economic decline notwithstanding their successes in attracting mobile investment, trumpeted by development agencies.

In fact, the evidence relating to each of these five characteristics of the branch plant economy tended to be quite mixed, of course, partly for reasons discussed below. Thus, notwithstanding these generalizations about the branch factory syndrome, most empirical studies comparing the performance and characteristics of branch factories and indigenous companies indicate that the former perform better. While branch plants generally continue to display weaker backward linkages with their local host economies than companies indigenous to the localities concerned, they also were found to have recorded higher rates of product and process innovation than indigenous firms, and were quite often found to offer better employment conditions and pay.

Thus, in certain respects, the branch plant economy thesis presents an apparent paradox. This paradox doubtless partly reflects an implicit tendency to compare branch factories to some idealized or romanticized version of a locally rooted and responsible indigenous business: a misplaced comparison. Yet, neither can it be assumed – as has been the case in much academic and policy literature – that mobile or foreign investment and the multinational enterprises (MNEs) that generate it inevitably contribute strongly to territorial economic development. Multilocal and multinational enterprises and their branch factories bring scarce capital and resources to a region, but it is not inevitable that these will spill over to other companies in the host economy.

As noted above, discussion of the branch plant economy in the United Kingdom is closely bound up with the objectives and results of regional policy. Designed to ameliorate one regional problem (diagnosed as a pattern of localized overspecialization in one or a very limited set of industries), this policy contributed substantially to a new regional problem. As an inquiry by the Regional Studies Association succinctly described: “The spatial-functional separation of activities within large corporations has led to a new territorial division of labour, apparent at both inter- and intranational scales. In the UK context, regional uneven development based on industrial specialization by sectors (e.g. coal, steel, textiles) has been supplanted to a large degree by regional differentiation based upon specialization by corporate functions (e.g. management, R&D, component manufacture assembly-line production)” (Regional Studies Association 1983, 46–47, original emphasis).
If this was a pattern that had emerged very clearly within the United Kingdom, in other advanced nations experiencing deindustrialization, and in some developing countries by the end of the 1970s, the question was how to explain it. It is to the theoretical interpretation of these empirical characteristics of the branch plant economy that this entry now turns.

Theory and the branch plant economy

By the end of the 1970s the concept of the branch plant economy existed in the Anglo-American literature as a collection of empirical studies, informed notably by behavioral theories of the firm and very loosely by the rapidly consolidating theory of the vertically integrated, hierarchically organized modern MNE (e.g., Buckley and Casson 1976) as it had emerged to that point. By the early 1980s, the term had gained currency (see, for example, Watts 1981), with many of these empirical insights being drawn together by analytical syntheses within the political economy tradition.

In the United Kingdom, these included the work of academics involved with the Conference of Socialist Economists, and drew heavily upon French regulation theory (e.g., Dunford, Geddes, and Perrons 1981). Elsewhere, these features of the branch plant economy proved a focal point for scholarship relating them to the broader political economy of relations of dependency in the international economy. In Europe, German development economists Fröbel, Heinrichs, and Kreye (1981) emphasized in *The New International Division of Labour* how highly selective forms of industrialization in developing countries were linked to processes of deindustrialization in Europe through a new imperialism orchestrated by MNEs, not states.

The idea of dependent development had gained an audience among scholars commenting on the developing world. Here very similar empirical findings were emerging with respect to the dependent development associated with the enclave forms of industrialization brought by foreign participation in national economies. Much of this literature developed in the Latin American setting (Cardoso and Faletto 1969), where there were long-standing concerns over foreign participation in industry. While enclave forms of production in a range of agricultural, extractive, and manufacturing sectors raised the production possibilities of home nations, they appeared to contribute little to stimulating the domestic economies of host nations. The various linkages of industry enclaves were primarily internationally oriented: while labor was a local input, the bulk of other inputs to production was imported.

There were some notable strengths to this literature. One was the consideration of “fiscal linkages” (Weisskoff and Wolff 1977); this was absent from discussions of the branch plant economy (and has tended to remain absent from recent discussion of processes of industrial clustering in economic geography). Some of this literature might also be regarded as proto-evolutionary economic geography in its approach to the examination of enclave forms of industrialization (Weisskoff and Wolff 1977). Finally, Cardoso and Faletto (1969) underscored how enclave forms of production in Latin America had their roots not just in the economic organization of production but also in the social and political bases of nations. Indeed, in this respect there was a measure of cross-fertilization, as some of the Latin American literature, and its associated concepts of dependency and the enclave, was brought into the Anglo-American literature.

As these ideas began to be applied more broadly, however, they also were exposed to important lines of critique from those concerned primarily with the developing world.
BRANCH PLANT ECONOMY

(Lall 1975). Some aspects of the critique offered by Lall, concerning the difficulty of establishing an adequate definition of dependence and its causes – notably in terms of the extent and nature of the role played by foreign enterprise in host economies – would be underlined by the rapid industrialization emerging in the newly industrializing, export-oriented, “tiger” economies of East and Southeast Asia.

Early attempts to theorize uneven development, linking understanding of regional problems to broader processes in capitalism, suffered from a failure to acknowledge how local institutions can exert agency to effect positive change, with a tendency to extrapolate from one or two possibly unusual trends of the time. These approaches were later adapted to discuss the emergence of semiperipheries in the international economy (e.g., Hong Kong and Singapore), with a greater concern for a variety of possible outcomes associated with changes in the division of labor coming to the fore with the development of global commodity chain (GCC) and global production network (GPN) approaches.

Perhaps the most notable of these theoretical contributions was Doreen Massey’s Spatial Divisions of Labour, which synthesized many of these theoretical elements with empirical findings in the United Kingdom. Massey’s treatment contrasted a “locationally concentrated” spatial division of labor associated with the United Kingdom’s industrial past, in which localities contained the entire division of labor associated with producing a commodity, with the new spatial divisions of labor of the branch plant economies that these regions had become. This spatial division of labor was the result of branch plants being part of “cloning” or “part-process” spatial structures associated with multiplant and multinational enterprises. Importantly, Massey’s analysis demonstrated that the incorporation of these older industrial regions into these new spatial divisions of labor presented a far more difficult and intractable new regional problem than that of the industrial specialization associated with the locationally concentrated spatial structures produced through the Industrial Revolution.

More than simply elaborating the structural underpinnings of local and regional economic development as derived from an analysis of changing corporate organization and strategy during the twentieth century, Massey sought to integrate corporate strategy and division of labor with the changing class and gender composition of labor markets in different localities. Famously, her treatment conceptualized rounds of productive investment as overlain on one another over time, with each round of investment representing a particular element in the spatial division of labor. She was signaling the emergence of a relational geographical approach, in which a sense of the peculiarities of places and place-bound processes was retained, but integrated with an appreciation of how localities are penetrated and woven together by social relations extending across national and international space. Her work also sparked a renewed interest in the economic fortunes of places, in the guise of “locality studies,” in which the commentators struggled to tease out structural change from locally contingent and locally causal sources of economic change.

Branch factories and the evolving division of labor

The branch factory syndrome proved something of a lightning rod for academic research in economic geography and urban and regional economics in the 1970s and early 1980s. It resonated with the broader political, policy, and academic climate of the time, which was skeptical of, if not
openly hostile to, the activities of multinational corporations. Much has changed since, however, not least the political, policy, and academic commentary on the value of foreign direct investment (FDI). This raises significant questions regarding the ability of this literature to address contemporary circumstances.

There have been, and remain, possibilities for injecting a degree of dynamism and contingency into the branch plant economy literature, revolving significantly around the evolving divisions of labor in society. Intra-corporate divisions of labor continue to be refined, in ways that require greater incorporation of insights from international business rather than simply geography or urban and regional economics. Intercorporate divisions of labor have evolved significantly in recent decades, producing paradigmatic approaches in the guise of GCCs (Gereffi 1999), global value chains (GVCs) (UNCTAD 2013), and GPNs (Henderson et al. 2002). These often have been rather quick to jettison any focus on arguably the key actor in such chains or networks – the firm. Thus an important opportunity remains to integrate insights from the branch plant economy literature with those from the GCC, GVC, and GPN approaches.

**Post-Fordism, factory upgrading, and the branch factory syndrome**

The earliest revisions of the branch plant economy came with a desire to consider the implications of industrial restructuring facing Western economies with the demise of Fordism. “Post-Fordist” patterns of industrial restructuring, variously presented as flexible specialization, flexible production, or flexible accumulation, involved the fragmentation of consumption, the vertical and horizontal disintegration of production, and the increased salience of time-based competition (Schoenberger 1997). These were considered to offer possibilities for resynthesizing the division of labor, for bringing back home investment that had offshored to low-cost locations, and for upgrading branch factories in the older industrial regions of the United Kingdom and the United States. This was inspired by observations in the so-called Third Italy, where an apparent decline of the “mass collective worker” had been supplanted by smaller-scale, craft-based, high-value, export-oriented production in “new industrial districts.”

While subject to considerable critique at the time (Gertler 1988), this broad-based “flexibility thesis” nevertheless presented grounds to revisit the concept of the branch plant economy. One current of thought held that many of the positive features of this brave new world of post-Fordist production flexibility, and its associated potential for the development of workforce skills, were unlikely to take root in older industrial regions. Others, however, sought to explore the implications of flexible production for branch plant economies in terms of upgrading possibilities for branch factories, previously shorn of particular functions and parts of the occupational hierarchy, into “performance” plants, which were becoming regionally embedded through the development of greater local linkages and their acquisition of significant nonmanufacturing functions such as R&D and sales (Phelps 1993).

Despite a measure of optimism at the time, born of the theoretical possibilities for such upgrading, empirical evidence for such upgrading in older industrial regions stubbornly refused to emerge consistently or strongly. What these academics had been eager to observe in branch plant economies might simply have been cosmetic: the effect of the loss of the worst branch factories from regional economies in the late 1970s and early 1980s, leaving a stock of factories that were among the most successful, technologically sophisticated, functionally integrated
and, possibly, locally embedded (Phelps 1993). Indeed, factories continued to become, if anything, more disembedded in many of these older industrial region contexts, as the full effects of corporate restructuring and regional integration agreements were realized. Patterns of input and final product sourcing became markedly more simplified and internationalized, as major corporations sought to fashion international interfirm divisions of labor with preferred suppliers and clients.

Examples of individual factories upgrading have been sparse across those older industrial regions in which a branch plant economy had taken root. Elsewhere, in the newly industrializing world, the prospects for economic development from the branches and subsidiaries of multiplant firms and MNEs are more mixed, and include examples where local economic development outcomes have been more positive to date. At least this is the tenor of much of the recent literature on GVCs, GCCs, and GPNs, which have focused primarily on interfirm divisions of labor.

### GCCs, GVCs, and GPNs and interfirm divisions of labor

The branch plant economy was intimately associated with theories of the large multiplant, multinational company as these had developed by the 1970s and 1980s. The theory developed by this time responded to the growth of these corporations and their typical organizational forms, as they had emerged during much of the twentieth century from a very limited set of origin countries, notably the United States. It presented a picture of the MNE as a vertically integrated entity, with an ever-growing internal division of labor. The essence of the MNE was conceptualized as a desire to exploit its competitive advantage across national borders through internalization. At that time, the transaction costs of dealing at arm’s length with third-party contract suppliers were considered significantly higher than those involved with direct overseas production (often in the form of what would come to be considered branch factories). Hierarchical ownership and internalization was thus the preferred means of coordinating production across national borders, with a view to minimizing the transaction costs associated with serving international markets (Buckley and Casson 1976).

With the onset of greater international economic integration, reflecting the creation, widening, and deepening of regional economic integration agreements and subsequently increased physical and virtual accessibility (underwritten by very rapid developments in transportation and information and communications technologies (ICTs)), this portrait of the MNE had become increasingly outdated. The United Nations Conference on Trade and Development (UNCTAD) first began to highlight these developments, as embodying the emergence of complex integration strategies of MNEs that contrasted with the stand-alone and simple integration strategies evident for much of the twentieth century (UNCTAD 1994). Since this time, MNEs have further outsourced final assembly and intermediate input production, seeking out a wider range of “offshore” locations from which to source inputs and/or finished products. Indeed, some MNEs (notably retail corporations) have been “born” global, operating without their own factories, but sourcing finished products from contract manufacturers to serve final markets. The MNE has thus become something of an arm’s-length coordinator of a network of operations, both its own and, increasingly, those of third-party contract producers. Indeed, some MNEs now barely coordinate such supply chains or networks, beyond a full package supplier or contract manager — often a major MNE itself — that assumes responsibility for (and
the transaction costs associated with coordinating the bulk of such networks. Thus, rather than a vertically integrated hierarchy concerned with internalization across borders, the generalized picture of the MNE that had emerged at the start of the twenty-first century resembled the imperial trading companies of the first global economy of the eighteenth to the late nineteenth century.

These developments have called forth influential theoretical approaches that seek to address, primarily, the greater elaboration of intercorporate divisions of labor. GCC and GVC analysis, emphasizing the governance structures of the production chains orchestrated by MNEs, distinguishes between producer- and buyer-driven chains. Focusing on the emergence of new clusters of production derived significantly from the expansion of retail MNEs and retail brands overseas (e.g., Gereffi 1999), these approaches are attuned to the often limited prospects for developing countries (Gibbon and Ponte 2005; Humphrey and Schmitz 2002). They also have emerged to strongly inform international policy agendas (UNCTAD 2013) and development practice. A GPN approach has emerged in geography in parallel to GCC/GVC analysis, with similar concerns but focusing less on firms than on networks of relations among both firms and a variety of other agents (Henderson et al. 2002).

The increased outsourcing of intermediate inputs may raise the prospects for the greater elaboration of the local social division of labor and linkages locally surrounding branch factories in many settings and many industries; processes of vertical disintegration of production centered on branch plants may result in the relevant intermediate inputs being supplied locally by suppliers. Indeed, empirical studies within the GCC, GVC, and GPN perspectives have drawn attention to the emergence of significant new clusters of economic activity centered on FDI. Such studies, implicitly and sometimes explicitly, are framed in contrast to the earlier literature on the branch factory syndrome, enclaves, and dependent development.

While these developments have affected manufacturing and service MNEs alike, service MNEs have emerged and internationalized remarkably quickly through these twin processes of outsourcing and offshoring. The emergence of brand name producers without factories has exposed localities to the vagaries of MNEs that – given the very low sunk costs associated with overseas production – are able to, and often do, switch production locations. Production switching is not new, though doubtless it is a greater threat when factories are operated by third-party subcontractors. Experiences across developing countries indicate that the incorporation of localities into GCCs, GVCs, and GPNs has not produced uniformly positive economic development effects. Notably, the governance structures associated with these chains and networks of activities can preclude the development of vital knowledge of consumer tastes among indigenous businesses (Humphrey and Schmitz 2002). In short, the branch factory syndrome and enclave forms of production still have considerable relevance for manufacturing activities coordinated by some retail MNEs. This is particularly the case on the African continent, where features of the branch plant economy in sectors such as clothing and textiles have often been a product of the specific provisions of trade agreements (Gibbon and Ponte 2005).

The rapid growth and internationalization of service MNEs also is evident with respect to those intermediate service inputs previously supplied in-house as part of vertically integrated manufacturing and service MNEs, but which now are outsourced. The rapid growth of employment in telemediated sales and back-office “call centers,” in localities within developed and developing countries,
BRANCH PLANT ECONOMY

has raised the same concerns registered in the branch plant economy literature: low wages and unstable employment. Some indication of the volatility of the jobs and economic development contributions made by such footloose service branch factory investments is signaled by the manner in which outsourced intermediate service inputs also rapidly have been moved offshore – to the point that the two terms are often conflated. One local economy’s loss can be another’s gain, as major new business process outsourcing (BPO) hubs have emerged in developing countries such as India and the Philippines. In India, some of the activities performed have local economic development potential (in terms of the value added) matching those in developed country settings (Dossani and Kenney 2007). The same may not be said of the presence of these activities in the Philippines to date. The build-up of BPO factories in the Philippines extends to second-tier cities where the replication of the branch factory syndrome is evident, despite some evidence of strategies of engagement between local governments and the private sector which may lead to qualitatively better local economic development outcomes in the longer term (Kleibert 2014).

Their orientation to interfirm divisions of labor means that GCC, GVC, and GPN perspectives are well attuned to the present drift of corporate organizational changes. As valuable as this emphasis is, elaborated interfirm divisions of labor nevertheless coevolve with shifting intrafirm divisions of labor, to which this entry now turns.

Intra-corporate divisions of labor: the “fine slicing” of value chains

Notwithstanding the greater development of interfirm divisions of labor, a hallmark of the MNE as it developed through much of the twentieth century has been the continued evolution of the intrafirm divisions of labor. The hierarchical organization of MNEs now coexists with, and complements, elaborated and geographically extended interfirm networks.

Indeed, in many respects the elaboration of interfirm divisions of labor described above – especially those associated with the outsourcing of intermediate service inputs – derived in the first instance from an increasing greater division of labor within MNEs, producing activities that have come to be defined as “noncore.” Yet, within those core manufacturing and service activities not yet subject to outsourcing, the division of labor within MNEs has continued to expand. International business scholarship has maintained a focus on the firm, seeking to inject a measure of evolutionary thinking into an understanding of the MNE as a heterarchy, rather than just a hierarchy, composed of a diverse group of local subsidiaries connected with one another and the parent organization. A major focus of international business research thus has been the evolving roles and capabilities of MNE subsidiaries. One implication is that heterarchical parent company organizational forms lessen the intractability of the regional problem raised by Massey (1984). Certainly, the variety of evolving roles and capabilities possessed by individual branch or subsidiary operations of MNEs have important implications for territorial economic development, but these have barely been explored by economic geographers.

There is a mutually reinforcing relationship between the more substantive and immediate nature of international economic integration, driven by a host of regulatory changes and information and communications technologies, and the internal organization of MNEs: “the more profound sense of uncertainty now said to be facing corporations is arguably leading
to greater experimentation within corporations and their subsidiaries, but also as corporations and their subsidiaries seek to experiment with the institutions that make up their external environments … ” (Cantwell, Dunning, and Lundan 2010, 571). One important aspect of such experimentation within the MNE has been the “fine slicing” of value chains (Rugman, Verbeke, and Yuan 2010), a process in which increasing division of labor results in the establishment of ever-more-specialized remits within functions (such as administration and R&D) commonly treated as indivisible. Strategies of fine slicing exploit the specialized capabilities of individual branch factories or subsidiaries in the context of local sources of technology and competitive advantage, as well as parent company, home country-derived ownership advantages overseas. This includes greater specialization within individual functions such as R&D, for example, with regional R&D centers not merely adapting products to local markets but also assuming more specialized roles for the parent organization as a whole – in some instances globally linked across time zones. These developments potentially offer economic development gains to host economies, but they also imply problems of considerable complexity for public policy seeking to secure such economic development. Notable in this respect is the difficulty of “picking winners” deserving of local institutional support in aid of upgrading.

In short, “processes of globalization … have allowed these firms to leverage greater competitive advantage from the differences between locations. This has resulted in MNEs generating heterogeneity in their own internationally networked operations. It has also meant that countries … have also learnt the benefits of generating differentiated sources of comparative advantage” (Papanastassiou and Pearce 2009, 1). An evolutionary perspective on the MNE and its internal divisions of labor (as these manifest themselves at the level of individual subsidiaries) has taken the international business literature towards a greater analysis of firm–host economy institutional interactions, including at the subnational scale. It is here that the international business literature comes closest to examining the “strategic coupling” emphasized in the GPN approach, where aspects of the older branch plant economy literature still have much to add to contemporary theoretical approaches.

The branch plant economy today

While the branch plant economy, or branch factory syndrome, became apparent in the economic decline across much of the United Kingdom and the deindustrialization of some parts of the United States, in an increasingly internationally integrated economy it retains great relevance to a wider set of countries and regions, including some that are growing in the aggregate and others that have shown signs of growth.

History and the branch plant economy

The emergence of the branch plant economy was produced by policy designed to obviate the cyclical expansion and decline of urban and regional economies that had specialized in a limited number of industries. Yet, ironically, the notion of the branch plant economy has shown a certain cyclicality of its own. Many of the same policy concerns surrounding the purpose and efficacy of the incentivizing and attraction or recruitment of inward investment, first raised in connection with the term “branch plant economy,” remain highly salient in fast-growing economies today.

This is apparent in the attraction of old and new industries alike. Extractive industries continue to be associated with enclave forms of economic
BRANCH PLANT ECONOMY

development. These have been subject to the long-term fluctuations in commodity prices made famous in long wave theory. However, new industries such as semiconductor device manufacturers, sought after as valuable industries around which to transform the skill bases of branch plant economies such as Wales, Scotland, and the northeast of England, also have been subject to intense cyclicality in international prices – for items that in many respects are commodity inputs to a wide array of modern products. It is little surprise, then, that such cyclicality has forced these national and regional economies to pay a high price for their temporary participation in high-technology industries.

Away from those industries, old and new, that are subject to commodity-like medium- to long-term price movements, branch plant economies may be subject to a degree of short-term cyclicality. There are a number of quite closely interrelated aspects to this short-term cyclicality of investment into certain city and regional economies. As time to market has become an important aspect of competition in recent decades (Schoenberger 1997), productive investments as a whole are subject to greater generalized uncertainty than in the past. Some of this uncertainty seems likely to affect even rapidly industrializing economies at some point, as in China, where there is intense competition for FDI despite the relatively strong bargaining position held by host urban and regional economies, given China’s huge domestic market.

For example, a series of cities across China find their economies significantly linked to the fortunes of a single MNE, Taiwanese Foxconn, whose own fortunes are linked to the market success of the branded products it chiefly manufactures under contract. The economic fortunes of China’s economy as a whole, as well as those of these cities, are linked staggeringly closely to this one company. Foxconn employs over 1 million people in China, in factory compounds ranging in size from 20,000 to 400,000 employees (Ngai and Chan 2012). This monopoly contract manufacturer contributes considerably to local economic development across China – at least in terms of direct and indirect job creation, but it is worth remembering that the branch plant syndrome in the United Kingdom also was closely associated with subsidiaries of US companies with similarly seemingly unassailable positions. Moreover, the Chinese cities competing vigorously to attract Foxconn factories offer generous incentives packages and “customized spaces” familiar in the West, and even bespoke company town extensions, to the point where many of the public policy dilemmas surrounding spatial competition for capital in the West are likely to be reproduced.

Singer (the sewing machine manufacturer) opened its Clydebank factory in Scotland as one of the first US investments in the United Kingdom, operating there for nearly a century. Such longevity, or anything remotely approaching it, is well and truly a thing of the past, however. Some urban and regional economies are populated by large numbers of factories associated with mature or standardized products – whose duration there is potentially limited to the span of one stage of these products’ life cycles. Empirical evidence suggests that it is actually quite hard for urban and regional economies to progress to developing new products. Instead, branch plants, and by extension their host local economies, are subjected to repeated competition as they seek to be the place to produce the next product allocated from a parent company. As such, one key empirical characteristic of the branch plant economy – the stability of employment – remains an important question for academics and policymakers. Moreover, the “financialization” of the production of many goods and services has further reduced the time
necessary to amortize productive investments, or the spatial fixity of capital. Companies have themselves sought to reduce the risks associated with long-term exposure to particular production locations. Moreover, local and national governments may inadvertently compound this because the incentives and customized spaces they offer to investors offset a firm’s sunk costs.

Ownership and the branch plant economy

The rapid regional and international economic integration since the 1980s has prompted the thought that ownership may be of little relevance to the question of how to effect local and national economic development. The international economy orchestrated by MNEs has become markedly more polycentric, with a growing array of origins and destinations of FDI and all that this implies for both the characteristics of the associated subsidiary or branch operations and the opportunities for developing countries (UNCTAD 1999). Though the 1999 UNCTAD report warned against this, this increasing polycentricity has been reflected, until very recently, in an over-optimistic assessment of the regional benefits of MNEs and their FDI. After all, the branch plant economy literature was imbued with the characteristics of FDI of Western origin, derived particularly from American MNEs as the dominant agents of international economic integration at the time. These were, in the main, vertically integrated Fordist manufacturing and service companies, enjoying at their height monopoly shares of national markets. Yet the United States was beginning to be overtaken by new sources of FDI from Japan, Germany, South Korea, and Taiwan, and more recently the BRIC countries. While Japanese MNEs have been slower to internationalize R&D and innovative activity, anecdotal evidence suggests they take a longer-term view of the development of subsidiaries and surrounding supply bases. The Korean chaebol (meaning a large family-owned business conglomerate), on the other hand, expanded rather more rashly overseas, triggering a series of high-profile tournaments for their investments but also, by the same token, high-profile failed and aborted ventures. Thus, persistent differences in the character and processes of FDI, associated with different home country origins of FDI, may produce rather different outcomes for territorial economic development than implied in the original branch plant economy discussion.

To set against this, the geographical origin of FDI and technological leadership in different industries has tended to shift more rapidly in the past decades than during almost all of the twentieth century. This creates an additional potential source of the sorts of instability in employment facing urban and regional economies discussed above. It also embodies a new, more complex geography to the branch plant economy, bringing into existence a large and almost totally unexplored research agenda centered on reappraising the term. As Iammarino and McCann (2012, 246) note, “... the intra-firm connection and inter-firm relationships which constitute the essence of multinationals represent simultaneously both complementary and inter-dependent ways of allocating and organizing resources and assets across national borders, thereby contributing at the same time to the growth and change of both the firm and the external world.”

This has made it increasingly difficult to disentangle endogenous from exogenous forms of urban and regional economic development, as the branch plant economy and dependency literatures sought to do. These changes in the origins and destinations of FDI, undermining ambitious generalizations regarding the impacts of MNEs and their FDI, vindicate the essentially empirical character of the branch plant economy.
literature; the benefits of MNEs and FDI remain to be demonstrated empirically, even at the level of the individual companies and investments involved.

The modernized dependency of the offshore enclave economy

Noted earlier in this entry was the cross-fertilization between scholarship on the branch plant economy and that on dependency and enclave forms of industrialization. The term “enclave” has continued to have a currency in recent years in the area and development studies and international relations literature, in a way that the term “branch plant” has not. To this extent it can be suggested that both sets of literature will continue to have a relevance to economic geography. The term “enclave economy” arose largely in connection with developing country economies dependent on natural resource exploitation and consequently on the resource-seeking FDI of MNEs. The notion of the enclave has been heavily criticized as an adequate depiction of the potential for territorial economic development – even that centered on those extractive industries most commonly associated with the time-limited urbanization and industrialization in enclaves. Nevertheless, its main features remain apparent in extractive industries and are a source of some concern.

The concept of the enclave economy has been somewhat overlooked, precisely because of its association with extractive industries and their geographical specificities. But it is remarkable that there has been so little elaboration of the term “enclave” in connection with secondary and tertiary economic activities, notwithstanding what Weisskoff and Wolff (1977) described as the potential for modernized *dependencia*. Just as for the formation of branch plant economies in the United Kingdom, economic development policy undoubtedly has played a role in the widespread emergence of modern forms of enclave industrialization. Special economic zones of various types are used extensively as vehicles for industrialization in East Asia and elsewhere. Territorial concessions appear to have re-emerged as part of developmental states’ extra-territorial management of the offshore movement of both domestic and foreign MNEs. Meanwhile, the overwhelming drift towards a liberalization of national law and policy regulating FDI, coupled with a drift towards the upholding of MNEs’ rights (when compared to their obligations) in international arenas arbitrating on legal disputes, has created asymmetry in relations between governments and MNEs. It has also manifested itself in the widespread promotion of economic enclaves of various sorts offering special concessions to MNEs.

Such modern-day enclaves, associated with dependent development, can be observed in an increasingly wide range of industries. Thus there is ample evidence of a growing “offshore” economy in the form of exclusive tourist resorts, export processing zones (EPZs), offshore finance centers, and secure BPO operations, shaped by a variety of concessionary agreements actively authored by nation-states and even city and regional governments (Palan 1998). In light of these developments, in different industries with potentially different local economic development implications flowing from potentially qualitatively different forms of enclave, it is strange that the academic literature has not produced the sorts of classifications or taxonomies of enclaves associated with the literature on agglomeration and clustering.

Questions remain over the longer-term economic sustainability of urban and regional economies centered on such enclaves of economic activity. Indeed, the critical questions are
very much those raised by the branch plant economy literature – whether the economic benefits really accrue primarily to MNEs and their national home economies rather than to local host economies. It is conceivable that powerful new processes of cumulative causation and urbanization may be set in train around industrialization in these enclave zones. However, it seems fair to argue that, for the majority of such cases, such narrow industrialization will fail to produce broader-based urbanization and economic development.

Conclusion

As a term with analytical value, the “branch plant economy” could be regarded as one produced from particular vantage points at a particular time and place. Yet, as this entry has tried to highlight, the branch plant economy and the branch factory syndrome remain likely to resonate with academic and policy audiences in other times and places, including today. A term that emerged in the deindustrializing United Kingdom and the United States during the 1960s and 1970s retains some salience to regions and nations less favored by the sorts of interfirm divisions of labor that now characterize GCCs, GVCs, and GPNs. Indeed, studying the effects of FDI in new host locations, including successfully industrialized developmental states, may provide empirical ammunition with which to selectively reapply, develop, and modify insights from the branch plant economy literature.

The term is certainly in need of recasting and reanalyzing. The intra- and interfirm divisions of labor to which the phenomenon is connected have continued to evolve in ways that require further conceptual analysis of the consequences for urban and regional economies. The question arises as to whether there are new spatial divisions of labor that can now be identified, alongside the cloning and part-process formats noted some time ago by Massey (1984), and whether an understanding of the evolution of MNE subsidiary roles and capabilities can shed light on the potential for sustainable territorial economic development to be fashioned from erstwhile tendencies towards the branch factory syndrome. Furthermore, the question is whether qualitatively different types of enclave can be elaborated (as has been done for industrial agglomerations and clusters over the past several decades) in order to explain the realistic potential to fashion clusters from erstwhile enclaves.

Any such analysis of the changing character of branch plant economies will have to draw on interdisciplinary insights from development studies, from political economy perspectives on industrial restructuring within GCCs, GVCs, and GPNs, and from international business and management perspectives on the parent–subsidiary relationship and subsidiary roles. The challenge is to adequately link the macropolitical economy perspectives offered by the likes of GCC, GVC, and GPN approaches with empirical findings, via relatively parsimonious meso-level concepts or theories. It is in this way that the prospects for territorial economic development can be understood as a recursive combination of the unique place-bound properties of cities and regions and the structures and processes common to global capitalism.

SEE ALSO: Dependency theory; Global commodity/value chains

References


Brands and branding

Andy Pike
Newcastle University, UK

Caribbean cuisine, Latin American coffee, Danish furniture, Thai silks, Catalan design, Broadway theater, Hunter Valley wines, London insurance … Where branded goods and services commodities are from and the places with which they are associated are integral to their meaning and value. Raising such issues encourages reflection upon how we understand and explain the geographies of economy. Internationalization, even globalization, is complicating the picture. Brand owners, managers, marketers, buyers, and authorities are grappling with questions of origin(s), provenance, and authenticity. Yet, longstanding research, fixated with the “country of origin” effect on consumer behavior, has failed to develop ways of thinking about the geographies of brands and branding that extend beyond this national frame.

In the transition from a producer- to a consumer-dominated economy, society, culture, ecology, and polity, the brands and branding of goods and services commodities have proliferated in dramatic fashion. This rapid ascendancy has led some to claim the emergence of a “brand society” wherein brands are “the most ubiquitous and pervasive cultural form” and are “rapidly becoming one of the most powerful of the phenomena transforming the way we manage organizations and live our lives” (Kornberger 2010, xi, xii, 23).

Addressing the relatively neglected geographies of brands and branding, this entry defines what is meant by brands and branding, explains how they are geographical, introduces the idea of geographical associations, and emphasizes the roles of actors – producers, circulators, consumers, and regulators – in spatial circuits of meaning and value. Origination explains the geographical associations constructed by actors related in spatial circuits in their attempts to cohere and stabilize meaning and value in goods and services brands and their branding in particular spatial and temporal market contexts. Examples of origination are discussed, from branded clothing and telemediated services.

Answering the call to “get behind the veil, the fetishism of the market and the commodity, in order to tell the full story of social reproduction” (Harvey 1990, 422), origination offers a means of lifting the “mystical veils” (Greenberg 2008, 31) woven around branded goods and services commodities by actors’ increasingly sophisticated activities. Their strategies, techniques, and practices seek carefully to create, manage, rework, and sometimes obscure where goods are made and/or from where services are delivered, and the economic, social, political, cultural, and ecological conditions where and under which they are organized.

Defining the brand and branding

The rapid growth, evolving sophistication, and widespread use of the term “brand” has been accompanied by an outpouring of definitions from academic, business, consulting, and practitioner accounts. Such efforts have served to
BRANDS AND BRANDING

fragment rather than integrate understanding and explanation. No single or generally accepted model exists of the tangible (e.g., design, function, quality) and, of growing significance, the intangible (e.g., feel, look, style) attributes and characteristics of brands and their relative importance and relationships. An influential conception defines the brand as the characteristic kind or variety of a particular good or service (de Chernatony 2010). A more refined conceptualization utilizes the idea of “brand equity,” defined as the “set of assets (and liabilities) linked to a brand’s name and symbol that adds to (or subtracts from) the value provided by a product or service to a firm and/or to a firm’s customers” (Aaker 1996, 7). Brand equity is a function of the connected tangible and intangible assets of brand loyalty, awareness, perceived quality, associations, and other proprietary resources that together generate meaning and value through their creation, articulation, and enhancement by brand and branding actors (Table 1). The elements include associations (e.g., with particular people, periods, and places), identities (e.g., image, look, style), origins (e.g., where it is designed, made, connected with, or perceived to come from), qualities (e.g., feel, form, function), and values (e.g., efficiency, reliability, reputation). Actors comprise those involved in the production, circulation, consumption, and regulation of brands and branding (Table 2). Taking a social and spatial approach to brand equity addresses the critique of Aaker’s (1996) “managerial” approach with its focus upon corporate brand ownership and control by producers rather than consumers, its relative blindness to the social construction and consumption of brands, and its weakly developed understanding of the relations between meaning and value (Kornberger 2010, 35).

If the brand is the object, branding is the process of adding value to goods and services commodities by providing meaning. Branding ensures that goods and services are no longer defined only by their material basis and functionality but also by “symbolic powers and associations” (Kornberger 2010: 13). Branding suffers from the same proliferation of definitions and conceptualizations that complicate the task of defining the brand. Branding is the “manufacture of meaning” (Jackson, Russell, and Ward 2011, 59). It is what actors do meaningfully to articulate, enhance, and represent the facets and cues of the assets and liabilities in brands in ways that create value in specific market times and spaces. Branding attempts to engender consumer trust and goodwill through constructing distinctive associations in the brand – such as

Table 1  Brand equity.

<table>
<thead>
<tr>
<th>Brand loyalty</th>
<th>Brand awareness</th>
<th>Perceived quality</th>
<th>Brand associations</th>
<th>Other proprietary brand assets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced marketing costs</td>
<td>Anchor to which other associations can be attached</td>
<td>Reason-to-buy</td>
<td>Help process and retrieve information</td>
<td>Competitive advantage</td>
</tr>
<tr>
<td>Trade leverage</td>
<td>Familiarity-liking</td>
<td>Differentiate and position</td>
<td>Reason-to-buy</td>
<td></td>
</tr>
<tr>
<td>Attracting new customers</td>
<td>Signal of substance and commitment</td>
<td>Price</td>
<td>Create positive attitude and feelings</td>
<td></td>
</tr>
<tr>
<td>Time to respond to competitive threats</td>
<td>Brand to be considered</td>
<td>Channel member interest</td>
<td>Extensions</td>
<td></td>
</tr>
</tbody>
</table>

Source: Adapted from Aaker (1996, 9) with kind permission of New York Free Press.
authenticity, quality, and style — that directly and positively influence purchasing decisions (de Chernatony 2010). Relative to commodified or more generic goods and services, the object of the brand and the process of branding are utilized by actors seeking product/image and price differentiation.

The geographical in brands and branding

Despite their rapid emergence and reach, the geographies of brands and branding are under-researched and lack conceptualization and theorization, analytical and methodological approaches, and a cumulative and substantive body of empirical research (Pike 2015). Economic geography, for example, has “consistently undervalued brands as an area of study” and “many of our theories and accounts stop abruptly as the products leave the factory gates” (Power and Hauge 2008, 123, 139).

Brands and branding are geographical in at least three interrelated ways. First, when conceptualized as an identifiable kind of good or service commodity, the brand is constituted of characteristics imbued to varying degrees and in differing ways with spatial connections and connotations. As Dominic Power and Atle Hauge (2008, 138) argue, brands have an “inherent spatiality.” Facets of brand equity — such as loyalty, awareness, perceived quality, attributes, and associations — inextricably intertwine with spatially inflected considerations of who makes the good or delivers the service and from where, as well as their identities, histories, and sociospatial connotations. Industrialization and mass production in the late nineteenth century underpinned the commercial value and meaning of brands: “Through industrialization the production of many household items, such as soap, moved from local production to centralized factories. As the distance between buyer and supplier widened the communication of origin and quality became more important” (Lindemann 2010, 3). Branding is similarly entwined in geographical associations and contexts as actors attempt to articulate the attributes and characteristics in brands in meaningful and valuable ways. Such actors deploy the meaning-making of branding to identify, articulate, and represent signs and symbols that are inescapably associated with the spatial contexts and connotations of particular brands.

Second, brands and branding are geographical because they (re)produce geographical differentiation across space and time. Modifying Michael Watts’ (2005, 527) argument for the more thoroughly branded world of contemporary “cognitive-cultural capitalism” (Scott 2007, 1466): “The life of the [branded] commodity typically involves movement through space and time, during which it adds values and meanings of various forms. [Branded] Commodities are therefore pre-eminently geographical objects.”
Some see “global” brands and branding as placeless vehicles of globalization and homogenization that cross borders as a “global fluid” and are “super-territorial and super-organic, floating free” (Urry 2003, 60, 68). Qualifying this view are more spatially sensitive interpretations of heterogeneity, diversity, and variety in how brands and branding (re)produce geographical differentiation in a spiky and sticky world (Pike 2015). In these conceptions, branded commodities travel and communicate differing meanings and values across space and time. Particular brands and their branding find geographically differentiated kinds and degrees of commercial, social, cultural, ecological, and political meaning and value. In seeking to shape and respond to the particularities of different geographical and temporal market settings, branding practices may similarly be spatially attenuated and heterogeneous.

Third, brands and branding are geographical because of the close interrelationships between their inescapable and spatially differentiated geographies and “uneven geographical development” (Harvey 1990, 432). Accumulation underpins the role of brands within marketing strategies. The actors involved in brands and branding are compelled by the rationales of accumulation, competition, differentiation, and innovation to search for, create, exploit, and (re)produce economic and social disparities and inequalities over space and time. Brand owners are therefore compelled to construct, define, segment, and exploit profitable parts of goods and services commodity markets in particular spatial and temporal settings.

Brand actors invest substantial time, effort, and resources grappling with social and geographical differences, and specifically how these can be used and perpetuated to create and realize meaning and value in spatial circuits. Spatial manifestations of economic and social differences and inequalities fuel market construction and segmentation because “[w]ide disparities between rich and poor … bring into being more luxurious types of goods than would otherwise exist” (Molotch 2002, 682). Social stratification and hierarchy underpin the relentless search by elite groups for distinction through consumption, propelling the continuous refinement and construction of new desires and wants beyond needs. Brand and branding’s differentiation imperative compels actors actively to seek out, perpetuate, and (re)produce such inequalities, fostering social polarization.

Geographical associations in brands and branding

Having defined and established the geographical basis of brands and branding, the next step is better to understand how such geographies are conceived of and articulated. Geographical associations refer to the characteristic elements – material, symbolic, discursive, visual – of the identifiable branded commodity and branding process that connect and/or connote particular “geographical imaginaries” (Jackson 2002, 3). Geographical associations do not just materially tie brands and branding in fixed relationships to certain spaces and places, but are of different kinds, varying in their extents and natures over space and time (Pike 2015). Material geographical associations include specific spatial connections to authentic and traditional methods and particular places of brand production. Symbolic geographical associations imbue and insinuate spatial referents in brand logos as proprietary markers circulated to draw attention from potential consumers. Discursive geographical associations attempt to align brands with aspirational and desirable spaces and places through stories and narratives.
Visual geographical associations utilize “origin images … recalling to consumers a rich set of associations” to surround and imbue brands and infuse branding concepts and messages (Thakor and Kohli 1996, 33). Aural associations signify geographical connotations through meaningful music, songs, poetry, language, slang, accents, and dialects.

Actors deploy these geographical associations in brands and branding. In terms of economy, the characteristics invoked include quality, tradition, and reputation. Actors work these and other attributes into their brand and branding practices and elements through design, name, and labeling. Actors involved in the financial services brand Scottish Widows, for example, deliberately articulate the attributes of frugality, integrity, prudence, and trustworthiness that are geographically associated with the national territory of Scotland. These values, symbolized in the brand name and its advertising, are material in the functional financial products they sell to consumers.

Brand and branding actors exercise much care and selectivity in constructing geographical associations in branded goods and services commodities for certain market times and spaces. Guided by particular interpretations and judgments about particular market contexts, actors seek to capture and amplify certain attributes and qualities of branded commodities while discarding and hiding others. At one end, they play up certain desirable and valued meanings such as the heritage, quality, and reputation connoted by particular places. At the other, actors mask meaningful but less commercially valuable or even damaging elements. Diversity and variety in the different types, degrees, and characters of geographical associations afford brand and branding actors a rich and pliable palette. Brands and their branding are not simply insert content afresh, however, with the exception of entirely new market entrants.

Geographical associations encompass tensions and accommodations between relational and territorial, bounded and unbounded, fluid and fixed, territorializing and deterritorializing tendencies (Pike 2015). Advancing beyond the constraints of the nationally focused approach to “country of origin” dominant in the discipline of marketing (e.g., Bilkey and Nes 1982) to engage other geographical scales, geographical associations in brands and branding can be framed territorially at different spatial levels (Table 3). Actors try to construct geographical associations to delineated, even jurisdictional, spatial entities in establishing, representing, and regulating the connections and connotations of particular branded goods and services commodities. While it can be important, the national is not the only territorial scale at which geographical associations are framed and articulated by the actors involved. Geographical associations in brands and branding are scalar and territorial as well as relational and networked. Unbounded, fluid, and deterritorializing spaces and places, stretching through circuits and networks in and beyond clearly defined and delineated territories and spatial scales, also are evident in the geographical associations of brands and branding. Combining territorial and relational understandings, actors utilize brands and branding to create differentiated meaning and value. They borrow and attempt to transcend, hybridize, and mix the geographical associations of the territories and networks of specific spaces and particular places. Nebahat Tokatli (2013), for example, demonstrates how the modernization of the luxury fashion brand Gucci drew upon and articulated geographical associations based upon the look, style, and sensibility of the City of Los Angeles, California. This representation elided a set of disparate spatial connections.
BRANDS AND BRANDING

Table 3  Scales of geographical associations in brands and branding.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supranational</td>
<td>European, Latin American</td>
</tr>
<tr>
<td>National</td>
<td>Brazilian, Japanese</td>
</tr>
<tr>
<td>Subnational administrative</td>
<td>Bavarian, Californian</td>
</tr>
<tr>
<td>“National”</td>
<td>Catalan, Scottish</td>
</tr>
<tr>
<td>Pan-regional</td>
<td>Northern, Southern</td>
</tr>
<tr>
<td>Regional</td>
<td>North Eastern, South Western</td>
</tr>
<tr>
<td>Subregional or local</td>
<td>Bay Area, Downtown</td>
</tr>
<tr>
<td>Urban</td>
<td>Milanese, Parisian</td>
</tr>
<tr>
<td>Neighborhood</td>
<td>Upper East Side, Knightsbridge</td>
</tr>
<tr>
<td>Street</td>
<td>Saville Row, Madison Avenue</td>
</tr>
</tbody>
</table>

Source: Adapted from Pike (2015) with kind permission from Wiley.

and connotations: the American designer Tom Ford was from Austin, Texas, and worked in Paris, France; the brand is headquartered in Florence, Italy; and the Gucci branded clothing is manufactured in Italy and, increasingly, internationally.

Brands and branding in spatial circuits of meaning and value

The ways in which differentiated meaning and value are created, circulated, and valorized in spatial circuits can be explained through the agency of interrelated actors – producers, circulators, consumers, and regulators (Table 2) – seeking selectively to construct, cohere, and stabilize geographical associations in commodity brands and their branding in commercially valuable and meaningful ways, in particular spatial and temporal market settings. Brands and branding are integral to the dialectic between spaces and circuits of value and meaning as well as representations of the “economic” in markets. Together with branding, “the creation and promotion of brands are … pivotal in realising the surplus value embodied in commodities and so in helping assure the smooth flow of value and expansion of capital” (Hudson 2005, 76). Together, the object of the brand and the process of branding facilitate actors in their efforts to create and fix the geographical associations constituting brand meaning and value in spatial circuits. The brand provides a temporarily stable frame that is required because “particular objects need to maintain a degree of stability of meaning in order that they can perform as commodities and so enable markets to be (re)produced” (Hudson 2008, 430).

Brands and branding provide an object and a process for the efforts of actors to construct and stabilize what branded commodities mean and are worth in particular geographically situated market contexts and time periods. Meaning and value are inextricably related, and not simply reducible to the brand’s geographical associations. Any fixity and stability achieved by actors are always unstable and makeshift accomplishments. The restless and disruptive rationales of accumulation, competition, differentiation, innovation, changing consumer tastes, and so forth, constantly threaten to recast and rupture the meaning and value in a brand and its branding in specific market times and spaces. Acknowledging meaning and value creation beyond the point of production, Hugh Willmott (2010, 526)
argues that “there is often greater potential for generating surplus from intangibles like branding – even allowing for the vulnerability of brand equity to displacement by innovation, shifts in fashion or risk of its destruction by consumer activists.” Logics embedded in cognitive-cultural capitalism risk undermining the coherence and stability of brands and branding as devices for continued value flow and capital expansion in spatial circuits. Meaning and value are not simply fixed or given by brand owners, stamped or marked on the finished goods and services commodities as they leave the factory or are communicated from the office. Brands are economic, social, cultural, ecological, and political, making them inherently fluid, unstable, “socially negotiated” (Power and Hauge 2008, 130) and “fragile things” (Kornberger 2010, 53). Some are actively contested over space and time through market rivalry and consumer activist dissent. The evolving nature of space and place further ensure that the fixing, cohering, and stabilizing efforts of brand actors are only ever partial and temporary. While brand meaning and value can sometimes be enduring in particular times and spaces, the work of brand and branding actors is always unfinished and ongoing.

### Origination

Origination conceptualizes the ways in which actors – producers, circulators, consumers, and regulators interrelated in spatial circuits – attempt to construct geographical associations for goods and services commodities. These geographical associations are selected to connote, suggest, and/or appeal to particular spatial references as part of efforts to create, cohere, and stabilize meaning and value in specific brands and their branding in the specific settings of certain market times and spaces. At particular moments in spatial circuits, actors attempt to originate branded commodities, deploying strategies, frameworks, techniques, and practices of branding to articulate and communicate meaningful and valuable geographical associations. This produces, circulates, consumes, and regulates “geographical imaginaries” (Jackson 2002, 3) in diverse and varied ways. Such actors both constitute and are compelled by rationales of accumulation, competition, differentiation, and innovation that continually disrupt the spatial and temporal fixes of geographical associations, risking the collapse of the commercial coherence and competitiveness of their branded goods and services commodities. Given the differing interests of the actors involved, origination is a highly selective process. Particular geographical associations that provide branded commodities with certain kinds of meaning and value in specific market settings are highlighted and emphasized, while those that threaten to undermine meaning and value are suppressed.

Longstanding research in the discipline of marketing has demonstrated how product–country image and origin are often decisive in consumer decision-making. Such cues underpin the longstanding “country of origin” or “made in …” effect, evident in consumer views of the geographically differentiated capabilities and historical reputations of countries for particular goods and services (Bilkey and Nes 1982). While extended to encompass “country-of-origin of brand” (Phau and Prendergast 2000, 159), this work has remained focused on the national level (Pike 2015). Yet, this national framing has been disrupted by recent changes. First, for increasingly complex manufactures with multiple component parts and subsystems, it is no longer straightforward to assign a single and definitive national geographical origin to the finished branded good provided for sale in a particular market. Second, services have
risen in importance, facilitated by advances in information and communication technologies. Their relational nature and geographical reach have stretched economic activities across national contexts and complicated any assignment of their “country of origin.” Moreover, the growing significance of intangibles, for example design and styling, is blurring the distinction between branded manufacturing and services. Third, the “national” identifiers central to “country of origin” have been played down, reworked or even eradicated, as actors seek to obscure historical associations or dissociate brands from specific national territories in the context of globalization and efforts to reach beyond particular national markets. In the wake of these changes, a belated and growing recognition has emerged in marketing that origin is more than just “country of origin.”

Consequently, the origin category for the provenance of brands and branding has been opened up, extended and articulated by actors to reflect more closely – as well as in some cases obscure – which activities are undertaken, by whom, and where. Driven by brand owners’ differentiation strategies and regulatory standards, new forms of communication and labeling formats are emerging. “Designed by Apple in California. Assembled in China” (Apple) and “Designed by Kawada in Japan. Made in Thailand” (Nanoblocks) are two from an array of such originations (Pike 2015). These practices are grappling with the notion of multiple and connected – even “global” – origins for the particular production and service activities constitutive of specific brands. New markers of origin(s) are being used by actors to reference and represent the places of different functions, including assembly, design, delivery, engineering, component sourcing, and manufacture.

Origin has now fragmented and splintered. Analytically, branding actors in spatial circuits can be identified as working with at least 10 different categories of origination:

1. where the brand was originally born,
2. where the brand was innovated or thought up,
3. where the brand is designed and developed,
4. where the brand is tested and refined,
5. where the brand is headquartered (or indirectly where the brand owner is headquartered),
6. where the branded good or service is delivered or provided from (physical, virtual),
7. where the brand is made (manufactured and/or assembled),
8. where the brand is sold (wholesale or retail channels),
9. where the branded good or service is serviced or supported from (e.g., after-sales services, maintenance),
10. where the brand is recycled and/or disassembled.

Further, the order of importance of such forms of origination has shifted too. In the context of complex and internationalized global value chains, “there remain product categories where it pays marketers to ensure that superior expertise is still widely associated with a particular place” and “the power of brand trust is expected to override any doubts customers might have as a result of products being sourced from multiple countries. Consumers are more concerned about the country or place of design and quality oversight than the country or place of manufacture” (Quelch and Jocz 2012, 44).

Significantly for origination, the multiplicity and complexity in origin(s) and its geographies afford heightened degrees of flexibility for actors in (re)working the geographical associations of brands and branding. Depending upon specific connotations in particular geographical and temporal markets, a singular origin or plural origins can be less easily or obviously discernible and
the object of promotion or obscuration by the actors involved. The rise of online services in the digital era has further reinforced the flexibility of origins in virtual space. Situating origin(s) can still include the national scale, but this is only one spatial level amongst other territorial constructs that also need to be situated in tension and accommodation with understandings of relational networks.

**Origination in branded clothing and telemediated services**

The clothing industry has a long history of internationalization, and a geographically expansive division of labor marked by the enduring importance of brands and branding as well as provenance as manifestations and signs of meaning and value. Contemporary examples demonstrate how actors are originating their brands in particular ways. US-based American Apparel, for example, foregrounds “Made in Downtown LA” as integral to its business ethos and brand, and uses this origination to mark its products and retail outlets. The actors involved differentiate the brand in competition as “Sweatshop-Free” and vertically integrated, in contrast to the low-cost, vertically disintegrated and international subcontracted business models prevalent in the clothing industry. The brand is presented as an American-based “Industrial Revolution” with its key research and development, marketing, and manufacturing activities located in downtown Los Angeles. This narrative is articulated in the brand’s circulation and consumption in its retail outlets and on its website, which invites potential consumers to “Explore our Factory.” The actors involved in this branding have originated the good and service commodity in a nationally situated brand name (“American Apparel”) and articulated a “Made in …” claim that is located in a specific territory within a particular city and state – the downtown area of Los Angeles, California. Meaning and value are sought from geographical association with LA as a center of innovation, style, and buzz in global fashion circles. Origination illuminates the particular kinds and forms of geographical associations central to the brand’s identity and its branding activities.

Branding actors deploy geographical associations also in efforts to create and fix meaning and value in branded services. In telemediated services, the origination of where actual tasks are undertaken and services delivered from has become more complex. These emergent sociospatial divisions of labor therefore have begun to influence how branding actors originate their services, inflecting the spatial connections and references articulated in trying to construct meaning and value in market settings. The standardization and routinization of back-office service functions led to the development of contact or call centers, decentralized to economically lagging places with pools of abundant lower-wage, often women, workers. Competition and cost pressures encouraged outsourcing and the international offshoring of such contact centers, facilitated by advances in information and communication technologies. However, problems with quality, customer disquiet, and competition have led some UK service providers to return their contact center services to the company’s perceived home country. Such changes have influenced the articulation and circulation of the origination of the branded services. The bank NatWest, for example, used a recent advertising campaign to bolster its brand’s reputation for “Helpful Banking” by asking: “We have award-winning 24/7 UK call centres. Does your bank?” (emphasis added). Its branding actors explicitly seek to construct their services as higher quality and accessible because of their
BRANDS AND BRANDING

origination in “UK call centres.” This emphasis upon the United Kingdom is central to the meaning and value of the brand’s proposition vis-à-vis its competitors, who may still be offering services outsourced to external – and by implication lower quality – providers operating from outside the United Kingdom. NatWest is explicitly promoting the origination of its service delivery, connecting with consumers anxious to access services from the specific territory of the United Kingdom. This origination is being deployed, even though the telemadated means used to deliver actual services means they could be provided from virtually anywhere, and certainly beyond the United Kingdom.

Conclusions

Understanding how the meaning and value of branded goods and services commodities are entwined with where they are from, and the places with which they are associated, helps explain their geographies. Engaging the dramatic rise, importance, and pervasiveness of brands and branding in contemporary economy, society, culture, ecology, and polity, this entry seeks to focus more attention upon their relatively neglected geographies. It has defined brand and branding, explained why they are geographical, introduced geographical associations, and emphasized the agency of actors in spatial circuits of meaning and value. Origination is vital to understanding the geographical associations constructed by actors related in spatial circuits – producers, circulators, consumers, and regulators – in their efforts to fix meaning and value in spatial and temporal market settings. Connecting with geographical political and cultural economy, origination provides a means to address critical questions about how, why, where, and by whom goods and services brands are associated with specific and particular spaces and places.

SEE ALSO: Consumption; Cultural economy; Global commodity/value chains; Global factory; Global production networks; Industrial upgrading; Local/global production systems; Manufacturing industry; Outsourcing

References


Brazil: Associação dos Geógrafos Brasileiros (AGB) (Brazilian Geographers Association)

Founded: 1934
Location of headquarters: São Paulo
Website: www.agb.org.br
Membership: 6500 (as of December 31, 2013)
Chairman: Renato Emerson dos Santos
Contact: nacional@agb.org.br

Description and purpose

The structure of AGB is comprised of a national executive board and dozens of local branches distributed throughout the country (with elections every two years).

Its objectives are to: promote the development of geography by researching and disseminating geographic issues; encourage the study and teaching of geography, as well as propose measures for its improvement; participate in exchanges and collaborate with other Brazilian and international entities engaged in geographical research or subjects of reciprocal interest; examine decisions of the public or private sector that affect and involve geographical science, geographers and geographical education, and research institutions, and voice its opinion on the matter; convene geographers, geography teachers, students, and other interested parties for the defense and prestige of the profession; organize meetings, conferences, exhibitions, symposiums, courses and debates, and professional exchanges; and represent the thoughts of its members before government authorities and professional, cultural, and technical associations.

Journals or major publication series

- Anais dos Encontros Nacionais de Geógrafos e Congressos Brasileiros de Geógrafos (Annals of the National Geographers Meetings and Brazilian Geographers Congresses)

Current activities or projects

The Association sponsors the biannual National Geographers Meeting, and every ten years, the Brazilian Geographers Congress, with the most recent congress being held in Vitória, Espírito Santo in 2014. It also sponsors “Speak up Teacher,” a meeting held every four years to foster and disseminate articles related to the teaching of geography in school.

Active throughout Brazil, working groups linked to local branches address, on an ongoing basis, social issues, such as urban, rural, and environmental policies, among other topics.

Brief history

Founded in 1934, the association is a civil nonprofit organization. Over time, it has become a truly national organization with members throughout the country.

Until the mid-1970s, AGB was characterized as a researchers’ association. However, at the end of that decade, spurred by the growth of the Brazilian student movement, it changed its organizational perspective, which was reflected in the reformulation of its statutes. Since then, it has expanded as an association and is more closely tied to the fight for human rights and to
the political and democratic discussions taking place in Brazilian society.

AGB’s institutional history is linked to the history of geography and Brazilian geographical thought; there is a close relationship between geographical knowledge and Brazilian society. The local branches are responsible for the publication of several scientific journals, such as the *Paulista Geography Bulletin*, the oldest in circulation, the *Gaucho Geography Bulletin*, the *Fluminense Geography Bulletin*, and others. They are also responsible for organizing regional geography meetings.

Submitted by Nelson Rego
Brazil: Sociedade Brasileira de Geografia (Brazilian Society of Geography)

Founded: 1945
Location of headquarters: Rio de Janeiro
Website: www.socbrasileiradegeografia.com.br/inicio.html
Membership: 600 (as of June 2, 2015)
President: William Paulo Maciel
Contact: sbgrj@socbrasileiradegeografia.com.br

Description and purpose

The Sociedade Brasileira de Geografia has the following goals: the study, discussion, research and scientific explorations of geography in its various branches, principles, relationships, discoveries, developments, and applications; and specialty study and knowledge of the facts and documents concerning the geography of Brazil.

Journals or major publication series


Current activities or projects

The organization does not undertake large projects or continuous activities. Its main activity is the maintenance of the journal, Revista Eletrônica da Sociedade Brasileira de Geografia, and the presentation of occasional short-term courses.

Brief history

The society was founded on February 25, 1883, as the Geographical Society of Rio de Janeiro. In 1945 it became the Brazilian Society of Geography. The creation of the society was inspired by the French model, characterized by the dissemination of scientific knowledge through the exchange of publications, participation in conferences, and exchange of correspondence with similar foreign entities, mainly European. However, its main goal was the organization of its national territory.

Submitted by Jan Monk
Buffers refer to spaces between objects or areas or within a specified distance of an object or place. They appear in a variety of contexts, including vector applications in geographic information systems (GIS), in which they are used to demarcate spaces within particular distances of features. It is frequently observed that many phenomena are limited by distance, and a buffer provides a simple way of mapping out this area. Plants and animals within a certain distance of a road may be affected by noise and runoff, and a buffer can identify those areas.

In political geography, urban planning, and conservation, buffers may be areas used to separate incompatible land uses or geographic entities.

In vector GIS, a buffer is a space defined by a user-specified distance from a point, line, or area feature. It is a common application widely implemented within GIS software, though the exact details and capabilities will vary from one software package to another. Depending on the software, it may be displayed as a graphic outline around the original feature or used to generate a new polygon dataset. A buffer around a single point will simply be a circle; those around lines and areas follow the shapes of these features, though they will be somewhat generalized. Multiple buffers may be generated at different distances from a feature, and some software may allow variable buffer distances to different features to be automatically generated based on an attribute field (or column in the attribute table). More important roads may therefore get a wider buffer than smaller roads. Overlapping buffers can generally be combined into one shape, and buffering inwards from an area’s boundary may also be possible, although it is not supported on all software. These operations are similar to dilation and erosion in mathematical morphology and image processing, with dilation generalizing a shape like a buffer to make it appear thicker and erosion doing the reverse; the latter might be approximated visually by inward buffers.

Figure 1 shows buffers of 500, 1000, and 1500 ft around the boundaries of George Washington Carver National Monument in Missouri. At farther distances the buffers become more generalized. Figure 2 shows buffers of the same size around the Monument’s Visitor Center. As this building is represented in GIS as a dimensionless point the buffers are concentric rings around the point. For clarity the buffers are shown on top of the Monument boundary. Figure 3 shows 500-, 1000-, and 1500-ft buffers around the Monument’s entrance road. At farther distances the buffer again becomes more generalized.

The GIS user must decide the distance a buffer will extend. Appropriate standards may exist for a particular application; for example, the US Federal Transit Administration uses $\frac{1}{4}$ mile (1250 ft, about 381 m) as an indicator for access to bus stops or public transit stations. In the field of road ecology, studies have been done to identify the relevant distance from roads at which particular

---

Joe Weber  
*University of Alabama, USA*

Buffers refer to spaces between objects or areas or within a specified distance of an object or place. They appear in a variety of contexts, including vector applications in geographic information systems (GIS), in which they are used to demarcate spaces within particular distances of features. It is frequently observed that many phenomena are limited by distance, and a buffer provides a simple way of mapping out this area. Plants and animals within a certain distance of a road may be affected by noise and runoff, and a buffer can identify those areas.

In political geography, urban planning, and conservation, buffers may be areas used to separate incompatible land uses or geographic entities.

In vector GIS, a buffer is a space defined by a user-specified distance from a point, line, or area feature. It is a common application widely implemented within GIS software, though the exact details and capabilities will vary from one software package to another. Depending on the software, it may be displayed as a graphic outline around the original feature or used to generate a new polygon dataset. A buffer around a single point will simply be a circle; those around lines and areas follow the shapes of these features, though they will be somewhat generalized. Multiple buffers may be generated at different distances from a feature, and some software may allow variable buffer distances to different features to be automatically generated based on an attribute field (or column in the attribute table). More important roads may therefore get a wider buffer than smaller roads. Overlapping buffers can generally be combined into one shape, and buffering inwards from an area’s boundary may also be possible, although it is not supported on all software. These operations are similar to dilation and erosion in mathematical morphology and image processing, with dilation generalizing a shape like a buffer to make it appear thicker and erosion doing the reverse; the latter might be approximated visually by inward buffers.

Figure 1 shows buffers of 500, 1000, and 1500 ft around the boundaries of George Washington Carver National Monument in Missouri. At farther distances the buffers become more generalized. Figure 2 shows buffers of the same size around the Monument’s Visitor Center. As this building is represented in GIS as a dimensionless point the buffers are concentric rings around the point. For clarity the buffers are shown on top of the Monument boundary. Figure 3 shows 500-, 1000-, and 1500-ft buffers around the Monument’s entrance road. At farther distances the buffer again becomes more generalized.

The GIS user must decide the distance a buffer will extend. Appropriate standards may exist for a particular application; for example, the US Federal Transit Administration uses $\frac{1}{4}$ mile (1250 ft, about 381 m) as an indicator for access to bus stops or public transit stations. In the field of road ecology, studies have been done to identify the relevant distance from roads at which particular
BUFFERS

Figure 1  Buffers around boundary of George Washington Carver National Monument, Carthage, Missouri. 500 feet = 152 m. Data from National Park Service and U.S. Census Bureau.

plants and animals will be affected by different highway phenomena.

In addition to displaying an area, buffers can also be used to select out other geographic features during GIS analysis operations such as “select-by-location” or “map overlay” operations. This type of operation includes selecting features by distance as one of the standard spatial relation types. In this case, vector objects within a user-specified distance of one or more features will be selected for further analysis or manipulation. This selection process works in a very similar way to buffers except that instead of generating a new object on the map the selection buffer is not seen.

Some important limitations exist. Buffers are usually based on straight-line or Euclidean distance in two dimensions only, and will therefore only approximate true distance in rough terrain. They also do not take into account the Earth’s curvature and so will provide accurate results only over small areas, though an inappropriate projection may also distort the distance value. GIS software may allow geodesic or spherical distances to be calculated and, if so, these are more accurate over larger areas. As with most
GIS applications, buffers create an absolute or Boolean space in which distance defines an area with sharp boundaries and no transition. This is represented by a typical GIS polygon in which a location is defined as either inside or outside, with no possibility of ambiguity or uncertainty. In the real world, there may be a gradual transition, as with a culture area or neighborhood unit. An additional difficulty with buffers is that they do not allow for uncertain or probabilistic distances. Geographic features may not have a fixed existence that can be precisely measured or verified. A probabilistic or fuzzy representation would be more appropriate in many applications.

A number of other GIS operations are similar to buffering. Thiessen or Voronoi polygons divide an area based on distance to the nearest seed point. These also use straight line distance but unlike buffers they partition an entire area into separate polygons that never overlap. They are also based only on vector points. A related application is visibility analysis or viewshed mapping, which identifies areas visible from one or more observation points based on a terrain surface such as a raster digital elevation model (DEM). The result is two polygon areas: those

Figure 2  Buffers around Visitor Center of George Washington Carver National Monument, Carthage, Missouri. 500 feet = 152 m. Data from National Park Service and U.S. Census Bureau.
BUFFERS

visible from a point and those not visible (or invisible) from that point. As with buffers, there is a Boolean border with no transition. They commonly have very irregular shapes with many discontinuous sections and so look very different from buffers. However, they could be used in a planning process to screen areas. This visibility does not take into account buildings, vegetation, haze, and often the curvature of the Earth. Observations points are vector points, though research has been done to accommodate linear or area-based viewsheds.

A different form of buffers can be generated using network distance, such as measured through a street network. These network buffers may be called service areas or service networks. Network analysis functions in GIS allow computation of network distance areas that show all streets within a user-specified distance of one or more origin points. Not all transport routes will permit the same speeds or costs, leading to uneven movement opportunities. These service areas therefore tend to be irregular, depending on the topology of the street network and the particular impedance function used (network distance, driving time, etc.) and how this relates to the links in the network. They may also have holes inside them in areas with few streets.

Figure 3 Buffers around entrance road of George Washington Carver National Monument, Carthage, Missouri. 500 feet = 152 m. Data from National Park Service and U.S. Census Bureau.
An application of buffering also appears in network analysis routing computations when the GIS software must connect an origin to a destination. These would be input by the user and must be within a certain distance or search tolerance of a network link. This tolerance will have a default value but will likely be adjustable by the user. The tolerance is effectively a buffer; locations within the buffer are reachable while those beyond it are not. This is especially important when the user manually places a point on a link with no width, or when geocoded addresses with a location offset from the street centerline are used.

Distance can be measured in raster GIS as well, again using straight line distance with cell resolution (size) to approximate an area within a distance of a specified cell or group of cells. Euclidean distance will be measured between cell centroids, so accuracy for a distance measurement is highly dependent on resolution. Cells within a particular distance of a cell or group of cells could be selected using a reclass function and would approximate a vector buffer. A spread function or cost distance can be used to take terrain, network, or other impedances into account by assigning each cell a friction factor. Accumulated impedance will be measured outwards from an origin cell or group of cells, approximating a network buffer. These will be displayed as a new raster surface. As with network analysis, the results will likely be irregular areas. Least-cost paths through the accumulated impedance surfaces can also be found, though these will more likely represent movement over different types of terrain or land cover than vehicle speeds through a network.

Buffers also appear in other geographical contexts, though in slightly different form. In political geography, a buffer refers to an area around or between jurisdictions that serves to separate them. In this case, a buffer exists between geographic entities, a different usage than that in the field of GIS. Demilitarized zones between hostile countries such as North and South Korea or the neutral zone that divides the island of Cyprus are examples; an independent buffer state that exists between two hostile countries, such as Mongolia between China and Russia, is another. The exact width, area, or shape of the buffer may be less important than its population, power, or ability to isolate jurisdictions.

In urban geography and planning, buffers can refer to areas both between and around particular land uses. This is particularly the case with Euclidean zoning (derived from the city of Euclid, Ohio). Incompatible land uses, such as industrial and residential, may be separated by a buffer zone or separator, whether of another land use (a use buffer) or open land (a landscaped buffer). As zoning codes commonly make use of a hierarchy of land uses, incompatibilities are defined as land uses from different levels that are adjacent or near each other. Such buffers could be based on adjacency and containment (no incompatible land uses in or touching a residential area) or by distance (none within 1000 ft of a home). Use buffers are written into the zoning code, while landscaping buffers are essentially a design feature of a property, such as vegetation to screen unattractive facilities and to provide a noise barrier. Buffers or buffer zones may also be legally enacted to limit protest activities or ensure access to public facilities such as abortion clinics.

Conservation planning also makes use of buffers as additional lands to protect a nature reserve, natural features, or water. Conservation buffers may be landscaped areas to protect a riparian corridor, vegetation planted around a farm to protect soil, a recreational greenway along a stream, a greenbelt around a city, a windbreak, a vegetative filter to trap pollutants running into streams, or similar kinds of features.
BUFFERS

large and small. Buffers may have different or multiple purposes, and may be developed into a network. GIS can be very useful for identifying potential buffer areas and their characteristics; however, this will likely be based on a different methodology than vector buffering based on distance. This notion of buffer fits within International Union for Conservation of Nature (IUCN) Protected Areas Categories V and VI, but unlike larger reserves and parks they will likely have no legal status except that under local zoning codes or land ownership laws. There is some potential for demilitarized zones to serve as nature refuges, so again multiple uses of the buffer concept may inadvertently coincide.

SEE ALSO: Fuzzy classification and reasoning; Geographic information system; Network analysis; Thiessen polygons

Further reading


The built environment can be interpreted as the product of regional cultures and of economic and technological forces, and as the legacy of successive phases of historical development. It can also be seen as a significant influence on individual and social behaviors, and is widely recognized as simultaneously both outcome and shaper of political, cultural, and socioeconomic forces. There are many subdisciplinary approaches, each with a different perspective and emphasis. Table 1 outlines the principal approaches discussed here.

The emphasis on these various interpretations has differed a great deal from one disciplinary and subdisciplinary perspective to another. The traditional view of the built environment in geography and anthropology was exemplified by the “Berkeley School” of cultural geography and its chief exponent, Carl Sauer (1889–1975), and by the work of architectural theorist Amos Rapoport (1929–). Sauer and his followers were interested in understanding the material expressions of culture on regional landscapes. This led to the concept of cultural landscapes, a tangible outcome of the complex interactions between a human group – with its own distinctive cultural practices, preferences, values, and aspirations – and the natural environment of a region (Sauer 1964). The focus of such work, for several decades of the mid-to late-twentieth century, was the character of vernacular housing and of regional building types such as log cabins, churches, temples, farm buildings, and tobacco barns. Vernacular architecture was seen as a product of the needs, means, materials, traditions, culture, and natural order of local populations. The principal critique of such an approach is that it implicitly treats regional cultures as superorganic: overarching and independent forces in their own right.

A second traditional perspective on the built environment, morphogenesis, emphasizes its spatial framework as the outcome of broader historic influences. Morphogenesis refers to the processes that create and reshape the physical fabric of urban form. Over time, the morphology of settlements changes, not only as new fabric is added but also as existing fabric is modified (Whitehand 1992). Basic forms, consisting of house, plot, and street types of a given period, become hybridized as new buildings replace old, plots are amalgamated or subdivided, and street layouts are modified. Each successive phase of settlement growth is subject to the influence of different social, economic, and cultural forces, while the growth of every settlement is a twin process of outward extension and internal reorganization. Each phase adds new fabric in the form of accretions and replacements. Morphological patterns are based on two fundamental elements: the size and shape of plots of land, and the layout of streets. Both vary according to historical period, economics, and sociocultural ideals. For example, where there is a shortage of building land, or where as many buildings as possible have to be accommodated along a given frontage (as on a waterfront or around a market square), small, deep plots tend to result.
Elsewhere, the size and form of the plot tend to be determined by the predominant house type.

In disciplines such as architecture, history, political science, and sociology, the built environment has often been treated as fossilized social history: an expression of the economic and political power exerted at different times by various individuals, social groups, and governments; and a reflection of the prevailing spirit, or zeitgeist, of the time. Architectural history, in particular, has provided a large body of knowledge that is mostly descriptive and contextual, with a focus on the aesthetic response, in built form, to the broad sweep of history. Thus, for example, Renaissance and Baroque architecture reflects the literary and scientific humanism that displaced the reactionary medieval mysticism of previous centuries. It gave expression to reason, rationality, and idealism, striving for a sense of infinite space and for imposing grandeur. Built form of this sort was only possible on any scale where centralized, autocratic power was combined with significant economic resources, and so the result was often reduced to an extravagant symbolization of wealth, power, and destiny.

The initial aesthetic response to the radical changes of the industrial revolution of the nineteenth century was reactionary. In the face of turbulence and change, architects and their clients opted for the reassurance of neoclassical styles. In the second half of the nineteenth century, architects were still struggling to find an appropriate response to industrialization and to the challenges and opportunities presented by new technologies. Neoclassicism gave way to eclectic revivals and mutations of various historic styles before the Arts and Crafts movement, based on a romantic idealization of preindustrial crafts, emerged in England. The Arts and Crafts movement was at its height in Europe between

---

**Table 1** Subdisciplinary approaches to built environments.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Perspective</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley School</td>
<td>Material expressions of culture and landscape</td>
<td>Treats culture as superorganic</td>
</tr>
<tr>
<td>Morphogenesis</td>
<td>Phases of settlement growth and change</td>
<td>Focus on house type and size and shape of plots</td>
</tr>
<tr>
<td>Architectural history</td>
<td>Built form as a reflection of zeitgeist</td>
<td>Descriptive; focus on “high” architecture</td>
</tr>
<tr>
<td>Environment and behavior</td>
<td>Built environment as independent variable</td>
<td>Tends to underplay social processes</td>
</tr>
<tr>
<td>Space syntax and symbolism</td>
<td>Social meanings of the built environment</td>
<td>Tends to underplay sociohistorical context</td>
</tr>
<tr>
<td>Structures of building provision</td>
<td>Political economy of design, construction, and growth</td>
<td>Tends to underplay urban spatial structure</td>
</tr>
<tr>
<td>Real estate development</td>
<td>Property investment and urban design</td>
<td>Emphasis on newer and larger developments</td>
</tr>
<tr>
<td>Sociospatial dialectic</td>
<td>Reciprocal and recursive relationships between society and built environment</td>
<td>A more holistic approach</td>
</tr>
</tbody>
</table>
1880 and 1910; in the United States (where it was known as the Craftsman style) it peaked between 1910 and 1925 before giving way to another reactionary design impulse in the form of the Beaux Arts and the City Beautiful movement.

The first genuinely original aesthetic response to the industrial era was German Modernism: the Deutsche Werkstätten, the Dresdner Werkstätten, and the Debschitz School in Munich in the 1890s and early 1900s. These led to the establishment of the better-known Staatliches Bauhaus in 1919. As artists and designers grappled with the implications of the rush of technical sensations of the era — electric power, telecommunication, internal combustion engines, aviation, radio, photography, and cinematography — and the dislocations of economic change and modern warfare, the Bauhaus was at the heart of a broad movement of Modernism. Modern architecture gave particular expression to the technological and scientific underpinnings of the industrial era. Among the most influential practitioners in this context was Le Corbusier (1887–1965), who advocated apartment buildings as rationally designed “machines for living.”

The social-democratic radicalism of early twentieth century Modernism was lost by mid-century as Modernism became the principal language of high design, symbolic of progress and prosperity, and adopted in the form of an International Style by corporations worldwide. In the late twentieth century a new and rather eclectic aesthetic response, postmodernism, emerged as architects sought to reflect the globalized and consumption-oriented societies of the period. Meanwhile, Modern design has been widely reasserted but in a commodified form that revises the strict form-follows-function of Modernist architecture, producing pseudomodern buildings with basically regular, Modernist proportions embellished with functionless ornamentations.

This sort of perspective on the built environment reflects the distinction famously made by architectural historian Nikolaus Pevsner (1902–1983) between architecture and “mere building.” Yet of course there is much more “mere building” than there is “architecture.” In parallel with the aesthetic responses to changing societies have been changes in building technologies that have made for new kinds of buildings while making architecture the art of the possible. Iron girders, for example, made two- and three-story “fireproof” factories possible; steel framing was a precondition for the development of skyscrapers; while the low cost and ease of balloon framing with wooden studs helped to transform the suburban landscape of North America.

In many subdisciplines, the built environment has automatically been assigned the role of an independent variable, “explaining” everything from people’s perceptual acuity to their social networks and propensity to indulge in deviant behavior. A common suggestion is that the design and configuration of buildings and spaces sometimes create microenvironments that discourage “normal” patterns of social interaction and encourage deviant behavior of various kinds. A considerable amount of evidence has been accumulated in support of this idea. The inhibiting effects of high-rise and deck-access apartment dwellings on social interaction and child development, for example, have been documented in a number of different studies; and from these it is a short step to studies that point to the correlation between certain aspects of urban design and the incidence of particular aspects of deviancy. The best-known link between the built environment and deviant behavior relies on the concept of “defensible space” (Newman 1972). Much of the petty crime, vandalism, mugging, and burglary in modern housing developments, it is believed, is
related to an attenuation of community life and a withdrawal of local social controls caused by the inability of residents to identify with, or exert any control over, the space beyond their own front door. Defensible space is now an essential component in the praxis of urban design, endorsed by many law enforcement agencies as part of an international movement known as Crime Prevention Through Environmental Design (CPTED). Such an approach, as with all studies that treat the built environment as an independent variable, runs the risk of privileging built form over social process and falling into the trap of environmental determinism (built-form stimulus → human/social/cultural response).

There is, nevertheless, a good deal of evidence of the ways in which the “syntax” of built form relates to patterns of social encounters and relationships (Hillier and Hanson 1984). The built environment can be interpreted more subtly as a text: as a socially constructed phenomenon that expresses a distinctive mix of ideas and practices, of often oppositional social groups and political relationships. The built environment is heavily endowed with social meaning, but this meaning is rarely simple, straightforward, or unidimensional. There is an important distinction to be made, for example, between the intended meaning of architecture and the perceived meaning of the built environment as seen by others. David Harvey’s study (1979) of the Sacré-Coeur in Paris has demonstrated how the intended symbolism of the building – a reaffirmation of Monarchism in the wake of the Paris Commune – was for many years seen as a provocation to civil society, and is still interpreted by the predominantly republican population of Paris as a provocative rather than a unifying symbol.

Another critical point is that the social meaning of the built environment is not static. Thus powerful symbols and motifs from earlier periods are often borrowed in order to legitimize a new social order, as in the adoption of a selection of motifs from the classical revival in Europe by Jefferson and the founding fathers responsible for commissioning public and ceremonial architecture in Washington, DC. Buildings can legitimize a particular ideology or power system by providing a physical focus to which sentiments can be attached. The ideology of civics and the power of local government, for example, are symbolized in civic architecture. In similar fashion, metropolitan élites frequently use high-profile buildings and extensive regeneration projects as part of their brand identity in a competitive world economy.

On the other hand, the built environment can carry an unintentional symbolism. Quite ordinary settings can be powerfully symbolic because they are understood as being a particular kind of place. For example, as Donald Meinig (1924–) has observed, the stereotypical New England town is widely taken to represent not just a certain type of regional architecture but also a certain kind of community with distinctive values: family-centered, God-fearing, morally conscious, industrious, thrifty, and democratic. The commonplace landscape of California suburbia, meanwhile, is symbolic of a particular lifestyle for middle-class, nuclear families: individualistic, private, informal, and recreation- and consumption-oriented.

The formal literature on architectural symbolism distinguishes three facets of symbolic meaning: syntactical meaning, or the meaning that an element of the built environment acquires by virtue of its location in relation to other elements; pragmatic meaning, or the meaning that is understood in relation to the intentions of the architect, client, or social group at the time of building; and semantic meaning, or the meaning it acquires – intentionally or not – because of the idea or ideology that it represents. The semantic meaning of the built
environment can be an important component of people’s daily surroundings: powerful but stealthy backdrops that can naturalize and reinforce dominant political and economic structures as if they were simply given and inevitable. Laden with layers of semantic meaning, the everyday built environment amounts to a moral landscape that reaffirms dominant values.

These symbolic attributes are a reminder that the built environment is socially constructed as well as physically constructed. The production and consumption of the built environment can be seen not simply as a matter of supply and demand but also as a function of time- and place-specific social relations that involve a variety of key actors and institutions, including landowners, investors, financiers, developers, builders, design professionals, business and community leaders, government agencies, homeowner associations, and consumers. At the same time, it is clear that these relations among key agents and institutions need to be understood in terms of their linkages with the broader sweep of economic, social, cultural, and political change. These sets of relations represent “structures of building provision.”

Landowners stand at the beginning of the chain of events involved in urban development, and they influence morphogenesis directly through the size and spatial pattern of parcels of land that are sold to developers and through the conditions that they may impose on the nature of subsequent development. Speculators are an important category of landowners — especially the bigger players who not only rely on an ability to anticipate trends but also hope to influence or engineer change for their own benefit. They are an important component of the overall structure of building provision in big cities: “growth-machine” coalitions of local real estate, finance, and construction interests. This rentier class of landowners, developers, realtors, commercial bankers, and construction companies (along with auxiliary players in utility companies, engineering and technical subcontractors, retailers, chambers of commerce, lawyers, title insurance and trust companies, federal, state, and local agencies, transportation and utility companies, and local media) has a common interest in securing the preconditions for growth. This involves general support for the ideology of growth and consumption as well as direct tactical involvement in local politics around local government land-use regulation, and development policy. Many political leaders become members of growth-machine coalitions because local governments rely heavily on real estate taxes to fund infrastructure and essential services such as schools, police, and fire protection.

Within local structures of building provision, demand and supply are mediated by a whole spectrum of “exchange professionals” such as mortgage financiers, realtors, surveyors, market analysts, advertising agencies, appraisers, and property managers. Most important in terms of the physical attributes of the built environment are design professionals: architects, planners, urban designers, and landscape architects. As key arbiters of style, design professionals are in a powerful position to stimulate consumption merely by generating and/or endorsing changes in the nuances of building styles. Meanwhile, by virtue of the prestige and mystique socially accorded to creativity, design professionals add exchange value to the built environment through their decisions about design. As a result, design professionals have been charged with a “silent complicity” with the agendas of the politically and economically powerful in local growth machines. More generally, design professionals draw on their particular skills to translate social and cultural values into material form: they are amplifiers and rectifiers of intellectual, social, and cultural trends; at once both products
and carriers of the flux of ideas and power relationships in society. But because of the dominance of aesthetics in their professional ideology, most design professionals tend to see their work as abstracted from the people and places that it affects. The result is often a design determinism in which people are conceived of as beholders, not willed agents/actors.

The physical production of the built environment is an important component of the economic geography of many countries. In this context, the development industry has followed the trends of other producer and service industries, with mergers and acquisitions, vertical and horizontal integration, product diversification, the deployment of new technologies, just-in-time delivery, and niche marketing, resulting in a much greater market dominance of big, publicly traded companies with complex and sophisticated operations. In the United States there are tens of thousands of home builders, most of them tiny, but the biggest builders have been getting bigger, and taking up an increasing share of the market. In 1986, the market share of the 100 largest builders (in terms of new home sales) stood at 24%. In 2012, it was more than 40%, with the top 10 builders capturing more than 25% on their own.

Meanwhile, the progressive liberalization of international financial markets since the 1980s has constituted a global driver in big-city real estate construction. It has also led to the financialization of real estate – in the form, for example, of real estate investment trusts (REITs) and pension funds with large investments in the built environment. This has introduced far-reaching institutional changes in financial markets, increasing the role of financial actors and financial institutions in creating real estate credit and in operating domestic and international real estate development processes. At the same time, the organizational structure of real estate development has become more complex and more decentralized.

In particular, the design and construction of the built environment are increasingly influenced by corporate globalization. Enabled by digital and telecommunications technologies, by advanced international business services, and by the emergence of clients with transnational operations and a cosmopolitan sensibility, the portfolio of many architecture and construction firms has an international component. The scope of operations of many of the largest firms is now truly global, and some firms are transnational corporations in their own right: huge architecture and engineering (A&E) firms with multiple international offices covering several continents. Globalization has facilitated the spread of “traveling ideas” about real estate design and development that have contributed to homogenizing trends in built form. Property-led urban regeneration schemes are arguably the principal reason for the homogenizing appearance of many cities. The success of property-led redevelopment of the likes of Bilbao, Barcelona, London’s Docklands, and La Défense in Paris has led it to become the most influential of all traveling ideas, resulting in the serial reproduction of “designscapes”: ensembles of office buildings, retail space, condominium towers, cultural amenities, renovated spaces, landscaping, and street furniture that are different in the details of styling and finish but generically similar in concept and execution.

These trends in the production of the built environment are paralleled by trends in consumption. Global cultural shifts have come to place a premium on consumer experience, celebrity, and spectacle, and “place” has become increasingly commodified. In this context, developers understand that design – especially by “star” architects – can add significantly to exchange value. Policymakers have also
recognized the demand for quality in the built environment. A key part of providing a good business climate, for many cities, is the promotion of urban design and iconic architecture. Under the pressure of increased economic competitiveness, decision-makers increasingly look to flagship architecture to combine an imagery of economic regeneration and civic pride. While for centuries the quality of the built environment was an outcome of economic growth of cities, nowadays it is a prerequisite for the economic development of cities; and urban design has acquired a new role as an instrument of economic development.

From a structuralist perspective, the built environment is seen as part of the socioeconomic superstructure stemming from the dominant mode of production. As such it reflects the Zeitgeist of the prevailing system and serves, like other components of the superstructure, as one of the means through which the necessary conditions for the continuation of the system are reproduced. Investment in the built environment is seen as an important secondary circuit of capital and a potential component of the “spatial fix” of market crises.

David Harvey (1979) characterized the restless formation and reformation of the built environment as part of capitalism’s perpetual struggle to create a social and physical landscape “in its own image” and requisite to its own needs at a particular point in time. He nevertheless recognized the danger of thinking in terms of simple causal relationships, stressing the need for a flexible approach that allows urban development to exhibit a variety of forms within the dominant mode of production. As Edward Soja pointed out (1980), the spatial imprint of capitalism is not a smooth and automatic process in which the “needs” of capital are stamped, without resistance or constraint, onto the landscape. Rather, there is a “sociospatial dialectic” of significant reciprocal and recursive relationships among individuals, the built environment, and society. The built environment is thus, in Meinig’s words, “mold and mirror” of economy, culture, and society; it is both the product of, and the mediator between, social relations, and it is central to human geographies that are simultaneously contingent and conditioning, outcome and medium, product and premise.

SEE ALSO: Berkeley School; Cultural geography; Environmental determinism; Modernity; Postmodernity; Social constructionism; Urban geography

References


Further reading


Bulgaria: Natsionalen Institut po Geofizika, Geodezia i Geografia, Bulgarska Akademia na Naukite (National Institute of Geophysics, Geodesy and Geography, Bulgarian Academy of Sciences (NIGGG-BAS))

Founded: 2010
Location of headquarters: Sofia
Website: www.niggg.bas.bg/en/
Business improvement districts

Kevin Ward

University of Manchester, UK

Business improvement districts are public–private partnerships in which property and/or business owners in a defined geographic area elect to make a collective contribution to the maintenance, development, and marketing/promotion of their commercial district. This is done in the form of a self-agreed and self-imposed levy, or “tax,” which is a percentage of the annual business rates. Collected by local government in some countries and by central government in others, this tax is then given over to a business improvement district management committee consisting of paid officials and elected representatives of local businesses and, in some cases, local residents. It decides how to spend its budget, which in a city like New York ranges from US$500 000 for a business improvement district in Harlem to over $10 million for a business improvement district in midtown Manhattan.

The business improvement district (BID) program began in a small west Toronto shopping area. In the 1960s, local retailers in the Bloor West Village realized that voluntary contributions would be insufficient for them to deliver yearly revitalization strategies. After a series of meetings between local and provincial officials and local business representatives to discuss the establishment of a self-imposed levy, the city of Toronto passed By-law no. 170-70. The Ontario Municipal Board then approved the Bloor Jane Runnymede Business Improvement Area, the world’s first business improvement district. The city is now home to 77 business improvement districts and more than 300 exist across Canada.

The first US business improvement district was established in New Orleans in 1975. Since then the number of business improvement districts across the country has grown, as states have introduced enabling legislation. By the end of the twentieth century, just 25 years after they were first established in US cities and towns, there were over 400 in 42 states, with 64% in just five: California, New York, New Jersey, North Carolina, and Wisconsin (Mitchell 1999). According to Hochleutner (2003, 374), business improvement districts have changed “the way America governs its shopping districts, commercial areas, and downtowns.”

In the past 15 years, a number of other countries have introduced BID programs. These include Australia, Germany, Ireland, Japan, South Africa, the Netherlands, and the United Kingdom. This internationalization has increased the number of BID programs at play and the differences between them: for example, between the services they provide, the level of contribution by businesses, and the way in which the contribution is calculated, which varies, sometimes subtly and sometimes dramatically, from BID program to BID program. Likewise, there is a significant variety over space in the types of business that participate in its management, those who are levied, and those who are eligible to vote. In the United Kingdom, for instance, the contribution is compulsory for all business property occupiers within the area once the business improvement district is established, setting it apart from predecessors such as town center management schemes. In the United States, it is compulsory.
BUSINESS IMPROVEMENT DISTRICTS

for all property owners in a designated zone, although in some states churches and schools are exempt from the levy. In other countries, such as Australia, the contribution is not compulsory. The name “business improvement district,” furthermore, is not always used – so in Canada it is business improvement areas, in the Netherlands “bedrijven investeringszones” (business investment zones), and in Germany, neighborhood improvement districts. So while it is possible to write of a single BID program, there is actually a variety of similar, and contrasting, BID programs in existence and circulation.

Nevertheless, these differences should not mask a number of similarities in the different BID programs. There is significant overlap between what business improvement districts do in different cities around the world. Mallett (1994, 277) has argued that the US BID program was “concerned with cleanliness and aesthetics of public space, security, achieving the best mix of activities, transportation access, and portraying an image of the whole to potential consumers.” That is, US business improvement districts “understand that simple things – such as keeping sidewalks clean and safe – matter enormously to the urban quality of life” (MacDonald 1996). This is also the case for other countries’ BID programs. In some countries, BID programs augment services provided by local government. In others, their work occurs in the space left by the redrawing of government boundaries. So, in the United States, Mallett (1994, 284) has argued that “[B]usiness improvement districts are a response to the failure of local government to adequately maintain and manage spaces of the post-industrial city.”

More specifically, what business improvement districts do can be grouped into three types of activity. First are those that target physical infrastructure. Examples include capital improvements (e.g., lighting, street furniture, shrubbery, etc.), economic development (e.g., offering incentives to businesses to relocate or stay in the BID area), and maintenance (e.g., collecting rubbish, removing litter and graffiti, etc.). The second type of activity is those focused on promotional infrastructure. Examples include consumer marketing (e.g., organizing and advertising events, producing and distributing maps and newsletters) and policy advocacy (e.g., lobbying government and other stakeholders, liaising with other business improvement districts, etc.). The third type of activity is surveillance infrastructure: for example, enhancement of public space regulation (e.g., regulating traffic flow, discouraging on-sidewalk selling, moving along the homeless, etc.) and security (e.g., ambassadors, security guards, and the use of CCTV cameras).

It is possible to understand the presence of business improvement districts in a growing number of cities around the world in a number of ways. For some, they reflect a bottom-up uprising by capital, as local businesses voluntarily involve themselves in decisions over how to govern and use public space. Here, work seeks to evaluate the impact of business improvement districts against a number of criteria or “key performance indicators.” Concerns center on efficiency and the value the BID program adds in terms of economic development (Mitchell 1999). For others, the BID program reflects the growing role assigned to capital and its representatives in the governance of cities, as governments are restructured. This work raises concerns over the “right to the city” and to whom business improvement districts are accountable. It argues that the BID program is one of a number of examples that reflect the growing interweaving of market-based and state-based regulatory arrangements, the combined effects of which have profound implications for notions of spatial and social justice (Ward 2006).
SEE ALSO: Deindustrialization; Governance and development; Public space; Urban political ecology

References


Further reading

Canada: Canadian Association of Geographers (CAG)

Founded: 1951
Location of headquarters: Montreal
Website: www.cag-acg.ca
Membership: 1023 (as of December 31, 2013)
President: Jean Andrey
Contact: valerie.shoffey@cag-acg.ca

Description and purpose

The Canadian Association of Geographers/Association Canadienne des Géographes is a national, bilingual (English and French), not-for-profit organization. It is the only national organization in Canada representing practicing geographers across the entire spectrum of the discipline. The association produces a peer-reviewed academic journal and an annual institutional directory. It also organizes academic conferences and regularly engages in communications with members and nonmembers about geographical issues. It plays a leadership role in connecting learned societies of relevance to geographers and in promoting geographic education in Canada.

Journals or major publication series


Current activities or projects

The association coordinates several conferences each year, one national and others in the regional divisions. Approximately 1000 delegates attend these meetings each year. The association facilitates other forms of communication among the Canadian geographic community including websites, a listserv with a reach to more than 2500 people, social media, and two newsletters. It also has responsibility for the Canadian National Committee for the International Geographical Union, has membership in both the Humanities and Social Science Federation of Canada and the Canadian Federation of Earth Sciences, and works with the Royal Canadian Geographical Society to strengthen geographic education in Canada.

Brief history

The concept for a Canadian association of geographers emerged, in part, out of the perceived value of geography in politics and international relations in the period during and following World War II and also out of a need to expand our understanding of and research on Canada’s north. The association was formed in 1951 when practicing geographers, both academic and nonacademic, were invited to attend an inaugural meeting at McGill University in Montreal.

Since 1951, the CAG has met annually as a national organization and regionally in smaller groups to achieve its mission which includes: dissemination of geographic research; promotion
of geographic education at all levels; recognition of geographic excellence; increased cooperation with other national and international geographic organizations; increased participation in national interdisciplinary organizations; and improved service to the membership.

Submitted by Jean Andrey
Canada: Royal Canadian Geographical Society/Société géographique royale du Canada

Founded: 1929
Location of headquarters: Ottawa
Website: www.rcgs.org
Carceral geographies

Dominique Moran
University of Birmingham, UK

Carceral geographies are an emergent and vibrant field of geographical research concerned with the spaces and practices of detention and imprisonment, informed by much longer-standing academic engagements with incarceration, namely criminology, prison sociology, and critical legal studies, and in particular with these disciplines’ increasing concern for an understanding of carceral space. Carceral geographies sit at a nexus of various interrelated developments in geographical research: the immense influence of the engagement between Michel Foucault and questions of space, place, and geography; the prominence within contemporary critical human geography of the ideas of Giorgio Agamben about bare life and spaces of exception; the growing prominence of the work of Loïc Wacquant on hyperincarceration and the punitive turn in the United States and Western Europe; and the integration of these perspectives into human geography by scholars such as Ruth Wilson Gilmore and Jamie Peck. In parallel, the recent spatial turn within criminology and prison sociology draws on human geographical understandings of space and spatiality as multiplicitous and heterogeneous, lived and experienced. As a result, an increasingly interdisciplinary approach has emerged within the social sciences that has opened a space for the coalescence of work under the heading of carceral geographies. This coalescence occurs at a critical moment in contemporary penal practice in the Anglo-American context with the expansion of workfare and prisonfare policies, the criminalization of immigration, and the expansion of the carceral estate. Of particular note is that carceral geographies’ focus on the spaces and practices of confinement extends both to “mainstream” incarceration, that is, of individuals detained by the prevailing legal system, and to migrant detention, where irregular migrants and “refused” asylum seekers are detained, ostensibly pending decisions on admittance or repatriation. Taking an overview of the scholarship produced by geographers concerned with incarceration enables the notion of the “carceral” to be brought more clearly into view. Dynamically open to transdisciplinarity, carceral geographies are informed by and extend theoretical developments in human geography, but also, critically, interface with contemporary debates such as the ongoing discussion of the aim and purpose of imprisonment.

Three main areas of interest characterize the field, broadly conceived as the nature of carceral spaces and experiences within them, the spatial geographies of carceral systems, and the relationship between the carceral and an increasingly punitive state. Taking these in turn, it is clear that scholarship in these three areas offers new perspectives on imprisonment, and that ongoing research enables novel understandings of the experiences of carceral spaces to develop. First, in work on the nature and experience of (highly racialized, classed, and gendered) carceral spaces, theorizations of incarceration informed by Foucault are debated and contested. Human geographers have shown that rather than being rendered “docile,” prisoner resistance to omni-disciplinary control can be expressed
CARCERAL GEOGRAPHIES

by reclaiming culturally defined prison space, with prisoners exerting agency to make and remake spaces, both material and imagined. In some contexts prison space is also personalized, through display of objects and belongings, and it has been suggested that this spatial modification reflects the construction of the meaning of prison spaces. Geographers have explored individual and collective prisoner agency and identified ways in which prisoners consciously negotiate tactical spatial maneuvers to find solitude in crowded prison spaces. This scholarship highlights the nature of prisons as institutions in which those in control construct and exercise their power and those incarcerated tactically create their own spaces within those places and making them temporarily their own as they occupy and move through them.

Second, spatial geographies of incarceration have been inspired by concern for the impact of places of incarceration on the communities that host or surround them. Geographers have engaged with: spatial interdependencies between prisons and cities; the impacts of prison development in persistently poor rural places in the United States, questioning prison siting as a means of encouraging economic development; as well as studying the effects of geographies of punishment on experiences of incarceration. Considering forms of confinement that take place outside the prison “proper,” they have also explored the use of electronic monitoring as an extension of mainstream carceral environments, suggesting that confinement can be independent of physical restriction and arguing that forms of punishment that are not explicitly prison-based can be just as constraining, in a different sense, as traditional incarceration. A major contribution of this body of work is in its suggestion that the carceral is something more than merely the spaces in which individuals are confined; rather, the carceral is a social and psychological construction relevant both within and outside physical spaces of incarceration.

This understanding of the nature of the carceral informs the third strand of research, into the relationship between these spaces and practices of imprisonment and a punitive state, often referred to as the prison industrial complex. It has been suggested that “transcarceral” spaces form beyond prison walls and constitute reconfineinent. Others see the carceral as inscribed on the bodies of prisoners who feel stigmatized after their release from confinement. Geographers have discussed the relationship between prisons and the metropolis in the context of hyperincarceration in the aftermath of what Wacquant (2009) describes as a swing from the social to the penal management of poverty, particularly in the United States, with a punitive revamping of public policy tackling urban marginality through containment and establishing a single carceral continuum between the ghetto and the prison. Very recently, US scholars have drawn attention to specific facets of the relationship between the carceral and the state, with analyses of the political controversy over the counting of prisoners in the US Census – where prisoners are “counted” as residents of the prison where they are held rather than as residents of the places from which they have come, with the result that prison populations are used for political purposes (such as the allocation of federal funds) from which disenfranchised prisoners are generally excluded. Similarly, geographers have drawn attention to the use of US prisoner labor, lacking protection through workers’ rights or health and safety inspections, to conduct hazardous waste-processing activity in ways that both subsidize the costs of imprisonment and undercut private recycling companies.

Carceral geographies also contribute to theory building within human geography with recent work critiquing the under-theorization...
of mobility and power by advancing notions of forced, coerced, disciplined, or governmental mobility, drawing on research into the enforced movement of prisoners and migrant detainees. Significant contributions have also been made to literature on feminist geography and embodiment, and within historical and cultural geography, where there is a growing interest in examining conceptions of crime, regimes of punishment, and their corresponding spaces of corrections and confinement, incorporating a historical-spatial focus into the study of current and former correctional institutions and their wider social contexts. Carceral geographers have made significant contributions to antiracist geographies and have also attended to the experience of transgender individuals’ experience of incarceration.

**SEE ALSO:** Biopower; Borders, boundaries, and borderlands; Crime; Human rights; Migration: international; Power; Refugees; Security; Social justice; Space of exception; Surveillance

**Reference**


**Further reading**


Care work

Shirlena Huang  
*National University of Singapore*

Care work refers to any face-to-face, often hands-on, service that contributes to the physical, emotional, psychological, and cognitive maintenance and development of the individuals receiving the care. While it can be unpaid (especially when carried out by family members or friends), care work (or caring labor) more usually refers to occupations in which those providing the service, whether in homes or institutions, are paid employees but are nevertheless motivated (at least in part) by concern about the recipient’s welfare. Because care is socially constructed around imaginings of “good care” as a “natural” attribute of women, care work is epitomized in the feminized occupations of nursing and child, elder, or therapeutic care. Geographical discourses on care work(ers) have been bound up with debates around issues of difference and discrimination, as well as rights and responsibilities, and have most recently been examined within frameworks of ethics and social justice, across various spatial scales, from the body to locale-specific sites, to the transnational.

Gerontologists and child development scholars were among the earliest to study care, focusing primarily on the recipients of care. Feminists’ foray into the field of care research moved the focus to the nature of the work, those who performed care work, and its ties to gender and social inequality. Indeed, much of the research on the geographies of care and care work has been built on and strongly influenced by feminist and gendered analyses, which have drawn attention to the associations of care work with women, as well as to the power relations and gendered inequalities at the heart of its conceptualization and organization. While the initial focus was placed on unpaid care within the family (primarily of children and the elderly) and how its naturalization as women’s work affected women’s lives, more recent academic work has centered on the neoliberal shifts that have moved care to the market and commoditized care into the family; the cultural and economic undervaluation of care labor; the structural positioning of care workers (particularly along gender, ethnic, nationality, and skills dimensions) within local and global labor markets; and the transnationalization of care work and its implications for perpetuating structural inequalities as the demand for care workers in the Global North threatens the availability of care in countries of the Global South.

The feminization of care work has its origins in the industrial model of a male breadwinner whose family and home are cared for by a wife. Despite the rise of two-income households as women's opportunities for work in the paid labor force expanded, the family ideal of male breadwinner–female caregiver has persisted. With the thinning out of the welfare state, affluent societies have turned increasingly to commoditized care. As female citizens become unwilling to take on paid care work, given the unattractive pay for work that is demanding, often dirty, and sometimes demeaning (Huang, Thang, and Toyota 2012), the care vacuum in the home created by women's labor force participation has transitioned from one traditionally filled by local (often minority) women from lower
CARE WORK

socioeconomic classes to a sector dominated by migrant women from less well-off economies. The transnational transfer of caring labor from women in the Global North to migrant women from the Global South, who in turn depend on their fellow countrywomen to fill the care gap they have left behind, has created what has been described as “global care chains” for care workers (Hochschild 2000; Yeates 2009).

Feminists contend that, when commoditized, care work is undervalued first and foremost because “care enters the market as historically gendered and domesticated practices” (Green and Lawson 2011: 650). Work done by women – in particular that of mothering – is often deemed “natural” and to be done out of love and obligation; thus care work not only pays less than other jobs, but also pays less than other women’s jobs of similar skill levels that do not involve care (England, Budig, and Folbre 2002). Paula England and her colleagues (2002) forward three other reasons why care work suffers from a relative wage penalty. First, the low pay is supposedly made up for by the intrinsic fulfillment of the job, as it is assumed that those who undertake care work gain intrinsic satisfaction from altruism and a sense of “duty to care,” rather than money. Second, much of it is labor-intensive and hence not amenable to productivity-enhancing capital substitution. Third, people who need care are usually least able to afford it; because the social return is greater than the private return for care, the state will step in to provide this public good at a cheaper cost when the market undersupplies it. The discourses of devaluation mark all levels of care work; thus, even highly educated nannies from developed countries are relatively poorly paid, as are credentialed and skilled (migrant) nurses (Cox 2013; England and Henry 2013).

Those performing care work are subject not only to the ideological construction of the tasks as feminized, but also to cultural and class stereotypes by institutional structures, industry organizations, recruiters, employers, and care recipients. For migrant workers, gender commonly intersects with class, race/ethnicity, and nationality/migrancy as co-constitutive factors to construct a hierarchy of workers in the care labor market based on stereotypes which categorize individuals, groups, and classes as more/less appropriate care workers. For example, Filipinas are often placed in the top stratum as care workers in the global labor market because of their perceived intelligence and “natural” capacity for warmth and patience. Studies of migrant nurses reveal systematic patterns of discrimination as a result of structural inequalities and employers’ overt and covert practices (including pay differentials despite similar qualifications, and being assigned work or positions that do not match their skill sets, which results in deskilling and poorer promotion prospects) because of racist attitudes in their workplaces and host societies (e.g., Batnitzky and McDowell 2011; Huang, Thang, and Toyota 2012). Immigration policies often exacerbate these labor market inequalities by restricting migrants to particular categories of jobs, often for a limited time period, despite migrants’ skills, educational histories, and previous class positions in their home countries. In combination, these mechanisms redefine the class position of migrant care workers as they move across transnational space.

While feminist scholarship has challenged the association of care work with feminization and the reproductive sector, it has not gone far enough to challenge the care/economy dichotomy. Simply seeking a re-evaluation of the worth of care work does little to disrupt the binary; instead, it entrenches the categorical associations of women and care with domesticity (Green and Lawson 2011). Future work needs to relocate care squarely within the economy rather than as subordinate to it. Indeed, it is
now generally recognized that care work, rather than simply being “natural” for women, is a form of emotional labor that requires workers to produce managed performances that display specified emotions (verbally and nonverbally) to influence the emotional state of care recipients. Hence, while an affective component is seen as critical to good care, commoditized care means that care work can be performed without real care and, as such, is purely waged work. More recently, feminist geographers working on the transnational mobilities of care work have drawn on the ethic of care to contribute to postcolonial debates about the injustice of the Global North’s extraction of care workers (akin to resource extraction) from the Global South and threatening the availability of care in countries there. By drawing attention to the interdependence that shapes the lives of people and nations – and the amount of care work that undergirds societies – a care ethics perspective works to problematize and counter neoliberal notions of autonomy and its individualizing processes (Cox 2013; Raghuram, Madge, and Noxolo 2009). In pushing for considerations of collective wellbeing, a feminist ethics of care also highlights the need for equitable wages for paid care work.

Scholarship on care and care work in geography has grown rapidly after incipient efforts in the early 2000s. One clear indication of this healthy and evolving state of the research on the care sector is the burgeoning number of themed issues on care in geography(-related) journals in the last few years, including “Postcoloniality, Responsibility and Care” in Geoforum (2009); “The Ethics of Care” in Ethics, Policy & Environment (2010); “Care of the Body” in Social & Cultural Geography (2011); “Transnational Mobilities for Care” in Global Networks (2012); and “Gendered Spaces of Commoditised Care” in Social & Cultural Geography (2013). Geographers have been able to bring a unique perspective to the study of care work, particularly in its commoditized form. By seeing place as intersubjectively constructed, geographers trouble the problematic nature of the public/private boundary and provide insights into how the meanings of care and care work shift across different spatial contexts and scales, thereby challenging the identity of the spaces and the nature, extent, and form of caring relationships that take place within those spaces. Developing an ethics of care will require more research on care work in non-Western contexts, and from the perspective of care-sending countries. We also need a better understanding of how care – and hence care work – as a concept and practice is mediated as it moves across transnational spaces and is reshaped by social, cultural, economic, and political mechanisms. There is also much scope to further explore the role of men in care work. Our current understandings of what constitutes care and care work as essentially feminized limit how we understand what men do as constituting care; as feminist scholars have indicated, focusing only on women in care work reinforces that it is women’s work.

SEE ALSO: Caregiving; Domestic workers; Emotional labor; Gender, work, and employment; Migrant labor

References


Caregiving

Andrew Power  
*University of Southampton, UK*

Caregiving refers to a range of interpersonal practices and emotions involved in supporting people. In contrast to other concepts such as help or assistance, caregiving is more deeply inscribed with a set of moral values that inform the quality and integrity of care that takes place within caregiving relationships. At the same time, it is a contested term, given its historical legacies with former practices that were anything but “caring.” It is thus a complex and dynamic concept, its meaning often depending on the trajectories of different actors involved across time and space.

Caregiving practices form a core part of the work carried out by a range of professionals including social worker, nurse, home help worker, occupational therapist, and personal assistant. In many cases, though, the terminology and values associated with care are receding from professional codes of practice. Previous experiences of “bad” care by health and social care users, particularly people with physical disabilities and mental health difficulties, have led to a disfavoring of the term. The future for caregiving is thus discussed further later in this entry.

Caregiving is also closely associated with the care work within the family. Caregivers typically provide a range of emotional and instrumental supports that enable people who may need support to continue to live in the community. Caregiving has been associated with the dual tasks of caring about and caring for a person, although these are not necessarily mutually inclusive. In the former case, caring about a person refers to being concerned for the wellbeing of another. It typically involves offering moral support and guidance and showing respect for the dignity of the person. In the latter case, caring for a person typically involves physical acts of support and personal assistance and/or actively supporting a person’s decision-making.

Caregiving for someone can involve assistance with instrumental activities of daily living (IADLs) – for example, transportation, housekeeping, meal preparation, shopping, home maintenance, and medication management – as well as activities of daily living (ADLs) – for example, bathing, toileting, eating, and personal hygiene. They may also be involved in formal service coordination and management (e.g., navigating formal health systems, linking individuals to services, and coordinating multiple services from different providers for persons with complex needs). This recognition of the distinct care roles by certain family members (over and above those deemed typical for family life) gave rise to the concept of “informal caregiver,” a distinct role to describe individuals who provide ongoing care and assistance, without pay, for family members and friends in need of support due to physical, cognitive, or mental conditions.

Women have been primarily involved in these care roles, to a certain extent because of the long-standing patriarchal assumptions of the nurturing and altruistic roles of women within families. The concept thus initially began to attract attention from feminist geographers from the 1980s in response to the ways in which women were arguably confronted with a social and moral duty or obligation to provide care,
referred to as “compulsory altruism” (Land and Rose 1985), which often led women to become involved in different “caring cycles” as various family members moved through the life course such as young children, disabled or sick relatives, and elderly parents.

This early work – and the coining of the term “informal caregiver” within the literature – also coincided with the broader context of social care policy in the 1980s, in particular the erosion of state-run institutions and care facilities. This period saw the emergence of community care, which was welcomed by many proponents of social care reform, as it was envisaged that care recipients would take up more active and valued roles in their neighborhoods. For policymakers though, it was envisioned that community care would mean care by the community rather than in the community. In other words, in many neoliberal countries that had led the way with deinstitutionalization, such as the United States, Canada, and the United Kingdom, the state arguably failed to provide the range of formal services to make independent living in the community a reality. Indeed, the development of “service dependent ghettos” and rising homelessness followed the first wave of deinstitutionalization in the 1980s in North America, leading to “landscapes of despair” within the urban environment (Dear and Wolch 1987). In this context, the locus of care shifted into a community-based asylum without walls for many people.

Given the unfolding processes of deinstitutionalization and the rise in community care, different care “contracts” began to be carved out in terms of how caregivers were being recognized across and within various countries. In response to the political debates concerning family caregivers at the time, Twigg (1989) developed a threefold typology that categorized carers variously as: resources, where family care was taken for granted and assumed; coworkers, where carers were seen as having co-responsibility alongside care professionals; and co-clients, where informal carers were seen as requiring support alongside the care recipient. In each case, Twigg argues, there are distinct differences in how carers are valued and supported, and in how their potential vulnerabilities are recognized. In practice, she acknowledges that no one model prevails, but rather each is found to varying degrees in the ways different government parties, professionals, service providers, as well as gatekeepers of public assistance, value family caregivers.

This ambiguity is evident at the comparative national scale. Work by Power (2010), for example, found distinct differences between the United Kingdom, the United States, and Ireland, in terms of how carers were categorized. In the United States there is less general acceptance of the need to support the caregiver. In comparison, in the United Kingdom there is more emphasis on caring for the carers, which includes an assessment of needs for the carer. Meanwhile, despite there being less literature on caregiving from the Global South, Kamundia (2012) notes that in Kenya mutual caring and support in the community are the norm, and it is often felt that there is no need for persons with mental health issues or disabilities to live physically and financially apart from their families. Community-based mental health-care services in Kenya are thus limited.

Given the gendered norms associated with nurturing and caring, it is primarily women who have taken up the mantle of providing care and assistance to those relocating from residential institutions to the community, as well as for the next generation of families whose children are transitioning into adult services. Despite the strong gendered lines according to which caregiving is drawn, however, geographers such as Aitken have also examined the awkward spaces of fathering within family life. Aitken (2012)
CAREGIVING unpacks the daily emotional practices that are negotiated, contested, and resisted between parents in different spaces, exploring the complex identity politics around “househusbands” and “Mr. Moms.” Rather than seeing the greater presence of male caregivers as leading to a mutual or shared care identity between men and women, he found that even the most active fathers tend to see their role as “helping out” their partners rather than taking the main responsibility for child care themselves. Others have found that men who are caring for people with complex needs are more likely to draw boundaries around the extent of intimate care they will or will not undertake. Also, male spouse carers are often more likely to accept domestic help, as it is often perceived as substituting for domestic labor previously provided by their wives.

Geographers have chronicled the complex and varied spatial aspects of caregiving: in the home, in the community, in public spaces, and within the complex array of in-between care-related places such as shelters, daycare centers, and group homes. More recently, geographers have also examined new geographies of care that are unfolding in shared spaces in the community such as libraries, museums, and cafes (Power 2013). Across all these spaces, geographers have been attuned to the embodied dimensions of care.

The home has received much attention from geographers, given the complex and evolving sociospatial relations between individuals and their homes. Milligan (2000) has examined the changing meaning of the home for family caregivers, particularly for those caring for frail older people. In this context, the home has now firmly become the preferred site of care. This has given rise to new care technologies, formal care workers, and other professionals increasingly entering the private sphere, thus blurring the boundaries between the public/institutional space and the private home space; it may be described as an institutionalization of the home. Other work has emphasized the careful controlling of noise and routines in the home that may arise from caregiving duties within the domestic sphere. These factors raise delicate moral and ethical questions for families involved in caregiving work.

Other geographic work examines how caregiving continues across public space in terms of the management of both the tasks and perceptions by members of the public. One example of this work focuses on the complex and ongoing interactions between parents of learning-disabled children and members of the public. Ryan (2005) points to the ongoing spatial coping strategies and interventions used by mothers to appease negative reactions to children who may not be able to conform to appropriate ways of behaving and using space. In this case, caregiving becomes more than physical supportive acts and involves complex relational practices and coping strategies. Carers are also often caught in the middle between promoting the independence of young adults with learning disabilities and being overprotective, which can manifest itself in the ongoing management of their spatial practices.

Caregiving has also been examined through the lens of social geographers who have scrutinized the spaces of care and welfare within the city (e.g., shelters, drop-in centers) and have examined the impacts of social care reform in urban settings. These supportive approaches to how vulnerable groups are managed on the ground are arguably downplayed by many mainstream (primarily US) accounts of urban injustice by geographers who have become largely fixated on the punitive accounts of injustice in the city – particularly within a context where the residual neoliberal welfare state rules. Indeed, injustice must coexist with and depend on these more supportive currents within urban space.
In this sense, caregiving at a very fundamental level thus allows the geographies of capitalist production to continue.

More recently, with the rising acceptance of independent living, the spaces of caregiving have increasingly become fragmented, and “place-less,” as people expect to live meaningful lives in the community. Geographers have examined the blurring of boundaries between social inclusion and social exclusion for persons with disabilities within the context of daycare center closures and greater expectations for former service users to take up valued lives in the community and to occupy positions in the open labor market, becoming effectively coerced into inclusion (Hall 2005). Thus, to feel part of the wider community a person with a disability may at times experience greater feelings of social exclusion while simultaneously being constrained from collective and interdependent forms of support (Hall 2005).

More generally, despite the increasing recognition of the level and types of caregiving across the family and community, the term has more recently lost favor, as stated earlier, particularly among the care professions such as social workers, personal assistants, occupational therapists, and clinicians. This is arguably a reaction to the criticism of the term by many groups of people who may require support, such as disabled adults (primarily those with physical disabilities) and persons with mental health issues, who prefer terms such as “help,” “assistance,” and “support.” This rejection of the term is largely a reaction to the way people were previously “cared for” by the state, in a way that was anything but caring and that often forced them into separate and institutional lives. In this sense, care became synonymous with dependency, being helpless, and being a subject of pity.

For this reason, the terms “care” and “caregiving” have become less visible within social-care policy discourse. Their place has been taken by an evolving lexicon of choice and control prioritizing autonomous decision-making and control over one’s own support. This shift has been led by a strong independent living movement in the United States, Canada, Sweden, and the United Kingdom, and has become a central policy agenda, referred to as personalization or self-directed support. What is unclear in this context of self-management and autonomy is how personalization and caregiving can cohere.

Paradoxically, despite the term losing favor among certain client groups such as disabled people, within contemporary social policy there has been a renewed focus on the privatization of caregiving and the increasing reliance on the family. With the rise of austerity politics within many neoliberal countries, which seek to reduce budget deficits in an attempt to allow markets to adjust during adverse economic conditions, policymakers have been increasingly looking toward individuals and families within the community to become more involved in local caregiving and volunteering work. Again, geographers have been centrally involved in debates in this remaking of the welfare state. This has manifested itself on the ground by restricted eligibility criteria for social care users and regional commissioning bodies (e.g., local authorities) providing support only for those in “critical” and “substantial” need. As a result, more individuals have to rely on the care and support of relatives, friends and neighbors, and local voluntary associations.

The neglect of caregiving within social policy has led some scholars like Marion Barnes to make renewed claims for an “ethic of care” to be reinstated within welfare and social policy (Barnes 2012). She draws on the work of Jean Tronto and other moral theorists to articulate six core principles of care: attentiveness, responsibility, competence, responsiveness, trust, and respect. Without an ethic of care at the core
of support policy, she argues, those in need of support can become susceptible to a more commodified, transaction-like, and potentially abusive relationship with support professionals. It is clear, therefore, that caregiving speaks to broader issues of social justice in the way we value and encourage human flourishing.

Geographers who have contributed to this debate over the contested understanding of caregiving include Sophie Bowlby, who has argued against seeing the term as inferring dependence but rather recognizing the interdependence inherent in all personal relationships. Bowlby’s work has questioned the assumption that the experiences of a caregiver are often presented as a dyadic relationship with a care receiver, and that this is a one-way relationship. This work has emphasized the interdependent and reciprocal nature of caregiving relationships. Bowlby proposes the concept of “caringscapes” to illustrate the multiple, interconnected care relationships across space, to help minimize understanding of caregiving as solely a dyadic relationship between carer and cared for (Bowlby et al. 2010).

In terms of future directions in theory and methodology, it is clear that growing numbers of older people will place greater demands on informal caregiving. On the supply side, delayed marriages, declining fertility rates, and evolving family structures (from extended intergenerational family units or traditional nuclear families to single-parent households or people living alone) mean that there are fewer family members to provide informal care. How families, voluntary organizations, the private sector, and the state will reconfigure care arrangements and renegotiate the care contract will be forced to evolve in new ways. As support budgets devolve down to the level of individuals and the privatization of care continues, it is also likely that, for professional caregivers, particularly personal assistants, caring will become an important issue, as many will likely see more precarious working conditions, zero-volume contracts, and rising transport costs exacerbated by having to call on separate clients across the community. Given this continually evolving landscape of care, and the resulting spaces and places which its different actors must access, occupy, shape, and use, geographers are well placed for examining how these issues will continue to play out.

SEE ALSO: Aging; Care work; Disability; Emotional labor; Gender, work, and employment; Health and wellbeing; Social justice; Voluntarism and the voluntary sector

References

CAREGIVING


Cartesian coordinate systems

Michael N. DeMers
New Mexico State University, USA

Cartesian coordinate systems are planar (flat, two dimensional) coordinate systems that define the location of any given point by a unique set of $x$- and $y$-coordinate pairs. These pairs can be placed on a one-dimensional line as in a number line or they can be placed on a two-dimensional (2-D) (Figure 1) or even a three-dimensional (3-D) plane. The plane may be real, representing, for example, the coordinates of a carpenter’s framing square used to prepare a piece of wood for cutting, or it may represent something more abstract such as geometric spaces in mathematics. A prominent geographic example of a 2-D geometric space is that produced by map projection, the cartographic process of converting spherical latitude and longitude coordinates of the Earth’s graticule into a flat representation.

Once the Earth’s geographic coordinates are projected it is often more important to locate objects on that projected map surface than it is on the globe itself. This is because it is more common to make measurements on small portions of the Earth than on the Earth as a whole. One advantage of using Cartesian coordinates is that calculating distances is much simpler – employing the distance equation rather than spherical geometry. The distance equation is simply a deconstruction of the Pythagorean theorem in which the hypotenuse of the right triangle is the distance in question and the values of $x$ and $y$ are represented by the differences between any two $x$-coordinates (longitudinal distances) and any two $y$-coordinates (latitudinal distances).

The Universal Transverse Mercator coordinate system

There are several classic cartographic examples of Cartesian coordinate systems. The first, the Universal Transverse Mercator (UTM) system was developed by the United States Army Corps of Engineers during the 1940s and currently employs the WGS84 (World Geodetic System 1984) reference ellipsoid. Although there are many other Cartesian coordinate systems in use around the world, the UTM is among the more universally accepted forms for cartography and geographic information systems (GIS) work.

The UTM system divides the Earth from 80°S and 84°N latitude into 60 numbered zones, each 6° of longitude wide (Figure 2). The beginning zone, zone 1, ranges from 180° to 174°W longitude and continues eastward to zone 60 where it ranges from 174°E to 180° longitude. When referencing a UTM coordinate system the band numbers are necessary because each of these 60 zones uses its own transverse Mercator projection designed to reduce the amount of distortion throughout the north–south oriented zone. The Transverse Mercator projection, unlike the Mercator projection, has its projection cylinder rotated 90° (west to east) and touches the Earth at each of the 60 central meridian longitudes. The zones are kept narrow and the scale factor of the central meridian is reduced to 0.9996, thus keeping the distortion amount for each zone to approximately 1 part in 1000.
Locating positions on the Earth using the UTM system begins by identifying the number of the zone in which the measurement is to be considered. To omit the need for negative numbers, the system measures distances in the north (called northings) and in the east (called eastings). For simplicity, all measurements of northings and eastings are in meters.

The starting point (origin) of each UTM zone is the intersection of the zone’s central meridian with the equator. To avoid dealing with negative numbers within any zone, the central meridian of each zone is set at 500 000 m East, written 500000Em. This is called the zone’s “false easting.” Thus, an object located at say 8743 m west of the central meridian would be \((500000 - 8743 = 491257)\) written 491257Em. Likewise, a reading 75 281 m east of the central meridian would be \((500000 + 75281 = 475281)\) written 475281Em. UTM eastings range from about 167 000 m to 833 000 m at the equator with the range narrowing towards the poles due to converging meridians. The use of the false origin results in easting values that are limited to six digits (Carnes 2007).

In the Northern Hemisphere northings are measured northward in meters from zero at the equator. The maximum “northing” value is about 9 300 000 m at latitude 84° N, the north end of the UTM zones. In the Southern Hemisphere northings decrease southward from 10 000 000 m at the Earth’s equator to about 1 000 000 m at 80° S latitude. In the UTM system northings one does not reference the lettered latitudinal zones but meters to a maximum of seven digits.

Because of the extreme distortion at the poles, a special coordinate system called the Universal Polar Stereographic (UPS) coordinate system, in conjunction with the UTM system, is used to locate coordinate positions on the Earth. Like the UTM coordinate system, the UPS system uses a metric-based grid on surfaces developed from conformal projections. Unlike the UTM system, it uses a stereographic projection. The UPS coordinates begin where the UTM system ends. In the north it begins at 84° N and continues to the pole at 90° N, while in the south it begins at 80° S and continues until it reaches the South Pole at 90° S. To make sure there are no gaps between the UTM and the UPS systems the UPS system also contains a 30-minute latitudinal overlap between the two systems.

At the polar portions of the Earth where all meridians converge, north and south lines also begin to converge. This results in confusion of exact north and south directions varying up to
Figure 2: Numbering system of the Universal Transverse Mercator coordinate system. Modified from http://en.wikipedia.org/wiki/File:Utm-zones.jpg.
CARTESIAN COORDINATE SYSTEMS

180° as the meridians approach the poles. To simplify things, the UPS grid north is set arbitrarily as the prime meridian.

Military Grid Reference System

A derivative of both the UTM and the UPS systems, the Military Grid Reference System (MGRS) is employed by NATO (North Atlantic Treaty Organization) military organizations for locating places on the Earth. Because it encompasses elements of both the UTM system and the UPS system, the MGRS covers the entire Earth. The significant difference is in the labeling method.

The MGRS grid reference is a point reference system that has three parts: a grid zone designator (GZD), a 100,000 m square identifier, and a numerical location in easting and northing meters. The grid zone designation includes one of the 60 numbered UTM zones, each of which is intersected by the latitudinal bands that have a lettering scheme that determines their relative latitudinal position (National Geospatial Intelligence n.d. (a)).

Each MGRS zone is divided into 20 of these latitudinal bands each 8° in latitude. For convenience, these bands are given corresponding letters starting with C at 80°S latitude and ending in X. The system omits the letters “I” and “O” because of their similar appearance to the numbers one and zero. To ensure that all land in the Northern Hemisphere is contained within the system, zone X is extended 4° to a total of 12°, thus explaining why the top and bottom of the zones are not at corresponding north and south latitudes. It is notable that zones A and B and zones Y and Z do not exist within the UTM system. That is because they are reserved for the Universal Polar Stereographic projection system employed for the north and south poles (National Geospatial Intelligence Agency n.d. (b)).

The MGRS then uses an alphanumeric system for designating the UTM or UPS coordinates. The MGRS designation comes in two parts; the first is the grid zone designation (first three characters) and the second is the 100,000 m square designation). Using the following example, 16RWC8071741305, you would get the following for the grid zone designation.

- “16” indicates UTM zone 16 (single-digit zones employ a zero before the digit).
- For polar regions (those outside the UTM coverage) the characters are omitted.
- “R” represents the 8° latitudinal zone with that letter designation (12° for zone X)
- If the area is outside the UTM, zone A is used to represent the Western Hemisphere near the South Pole and zone B to represent the Eastern Hemisphere. In the south, Y and Z are used.

For the 100,000 m square designation each UTM zone is divided into 100,000 meter squares resulting in UTM coordinates that will always be multiples of the 100,000 m. The designations are based on a column and row lettering scheme where the columns are A–Z, omitting I and O (as in the UTM system) and the rows are next ranging from A to V, omitting I and O. So for UTM zone 1 (near the equator) the columns would be A–H, zone 2 would be J–R (omitting O), zone 3 would have S–Z. The letters begin over again with A at zone 4 and continue on around the globe to the equator (Figure 3).

For row letter designations there are two alternative lettering schemes, MGRS-New (aka AA scheme) and MGRS-Old (aka AL scheme). In the MGRS-New scheme, based on the WGS84 and other modern datums, the letter for the first row just north of the equator is A in odd-numbered zones and F in even-numbered
zones. In the MGRS-Old scheme based on older datums, the row letters are shifted 10 steps alphabetically. Thus the letter for the first row would be L for odd-numbered zones and R for the even-numbered zones. The systems are unique in that, with only the exception of zone X, an identical position reported on MGRS-Old and one on MGRS-New would not be confused. So for the previous example value of 16RWC8071741305 the letter W indicates the 100,000 m column W and the 100,000 m row C.

The remaining portion of the designation, the numeric locator, indicates the position within a 100,000 m square and is designated as $n + n$ digits, where $n$ is 1, 2, 3, 4, or 5. So, for example, if one is using $n + n = 5 + 5$ then the first five digits indicate the easting in meters measured from the westernmost edge of the square and
CARTESIAN COORDINATE SYSTEMS

the last five digits would represent the northing measured in meters from the southernmost edge of the square. A $5 + 5$ designation is the most accurate, representing a resolution of 1 m. Thus a $4 + 4$ represents 10 m resolution squares, $3 + 3$ represents 100 m squares, $2 + 2$ represents 1 km squares, and finally $1 + 1$ indicates the coarsest resolution, at 10 km. The NATO standard for specifying coordinates is $4 + 4$. The previous example of 16RWC8071741305 is a $5 + 5$, or 1 m resolution, indicating an easting of 80 717 m and a northing of 41 305 m for 100 000 m square WC in zone 16R.

There are some specific rules regarding this designation that should be remembered. First, if you are in a specific grid zone and are communicating about that grid zone then specifying it is not necessary. The numeric specification is sufficient. If an area being specified is small, but it is located in the overlap of multiple 100 000 m square zones then the entire grid reference is used. Another rule is that all coordinates are read with the easting first and the northing second, or “read right up.” Finally, the current convention regarding an abbreviation of the MGRS to lower resolution or when converting UTM coordinates is not to round but rather to truncate the values.

The MGRS convention is different for the polar regions of the globe. For the southern portion, that region south of 80°S latitude, the Universal Polar Stereographic (south) projection is used rather than the UTM projection. The west semicircle creates a grid zone designated zone A while the east semicircle forms a separate grid zone designated B (Figure 4). Regions north of 84°N latitude employ the Universal Polar Stereographic (north) projection. The west semicircle is designated grid zone Y while the east is designated grid zone Z (Figure 5).

In polar regions, the lettering scheme for the Military Grid is also somewhat different. While the row letters remain the same in that they range from A to Z, omitting I and O, the columns use a smaller alphabet. Columns are lettered from A to Z but omit I, O, D, E, M, N, V, and W. They are arranged such that the rightmost column in grid zones A and Y employs column letter Z. The next column in grid zones B or Z begins with the column letter A. This system was devised to avoid confusion by guaranteeing that no UPS square would be adjacent to a UTM square that possessed the same identification.

State Plane Coordinate system

Within the United States there is a highly accurate Cartesian coordinate system, or rather a set of 124 zones, each with its own coordinate system, that together constitute the State Plane Coordinate (SPC) system. Each US state contains one or more state plane zones that typically follow county boundaries. There are a total of 125 zones – 110 in the conterminous United States, 10 in Alaska, and five in Hawaii. Puerto Rico and the US Virgin Islands each have one. The SPC is popular because it uses a simple coordinate system based on plane surveying techniques and is not linked to the geographic system of latitude and longitude. To accomplish this, the SPC ignores the curvature of the Earth thus making calculations of distance and area easy to perform. In addition, because it relies on plane surveying, the SPC is extremely accurate within each of the zones (errors less than 1 in 10 000). Because the accuracy does not extend beyond the zones, it is not as useful for regional or national mapping as UTM. Instead it tends to be used for mapping within state governments.

Of the 124 zones, most are based on either a transverse Mercator or a Lambert conformal conic map projection, the choice essentially determined by the shape and orientation of
the state and the zones. The orientation of the state (e.g., north-south versus east-west) typically employ zones that share that orientation. State SPC zones that are longer in the vertical (north-south direction) than in the horizontal (east-west direction) use the Transverse Mercator projection because of its property of maintaining accuracy in the north-south direction. Alternatively, state SPC zones that are oriented in the horizontal (east-west) direction typically employ the Lambert conformal conic projection because it tends to preserve accuracy in the east-west axis. The Alaskan panhandle whose orientation is neither vertical nor horizontal, but diagonal, employs an Oblique Mercator projection designed to minimize the combined error in both the vertical and the horizontal.

Developed in 1933 by the US Coast and Geodetic Survey under contract by the North Carolina Department of Transportation, its creation was to convert latitude and longitude coordinates into a user-friendly Cartesian coordinate system. At the time the system was based on the North American Datum of 1927 (NAD 27) and later converted to the more accurate NAD83. It is currently the
most widely used such system used for local and regional surveying and mapping in the United States (Doyle 2004). The system has been revised several times and its digital equivalent has resulted in considerable savings in GIS processing speeds. Because of the rapid increases in geographic database sizes, this characteristic has allowed the system to continue to be used in GIS despite improvements in computational efficiencies in the computers themselves.

Despite its continued use for regional mapping applications, the SPC system is not without its problems. The primary problem is that each zone uses a unique coordinate system making cross-zone work difficult. Examples would include regional studies that cross state lines or operations on metropolitan zones that include multiple zones. To perform such analyses requires that each zone’s coordinate system be converted to that of another – a procedure that is burdensome and adds to the possibility of introducing error. Today’s modern GIS software is capable of making such conversions with relative computational ease and with limited loss of accuracy.

Figure 5 Northern polar region designations for the Military Grid. Source: Modified from http://en.wikipedia.org/wiki/File:MGRSgridNorthPole.png (public domain).
United States National Grid

A Cartesian coordinate system designed at least in part to provide for improved interoperability for location-based services and other GIS operations throughout the United States is the United States National Grid (USNG) (FGDC 2001; Cavell 2005). Sharing many elements of other national grid systems, it was developed by the Federal Geographic Data Committee as part of their mandate to improve interoperability in GIS and to allow its easy integration with other national grid systems. It also very closely resembles the Military Grid Reference System (MGRS). In fact, when the WGS84 or NAD83 datum is used, the coordinates between the two are identical.

Another of the interoperability characteristics of the USNG is that its coordinates can easily be transferred to US Geological Survey topographic and similarly gridded maps using northings and eastings. These measurements are in meters making between-location distances easy to calculate. Numerical digits designating these distances vary in precision from kilometers (two digits for each coordinate) to meters (with five digits for each coordinate). The USNG uses a common system based on the USMG that permits the use of physical landmarks that allow users unfamiliar with the local area to use them.

Ordnance Survey National Grid reference system

While there are many other national grids, perhaps the best non-US system known is the Ordnance Survey National Grid reference system. As the name implies, the United Kingdom’s National Grid was developed by the Ordnance Survey as the British Grid System in 1919 and adopted for use on military maps. By 1927 the system had been replaced by the “Modified British System” which was well adapted for use in military operations.

Today’s UK National Grid breaks Great Britain into progressively smaller squares identified by letters and numbers, as seen in other grid systems. The largest of these squares is 500 km on a side. These squares are identified first by letters and then by numbers. The letters – the first of a two-letter system – range from A to Z, omitting I. However, Great Britain is only covered by four of the squares – those designated H, N, S, and T. Each 500 km square is further divided into 25 100 km squares. Letters, again ranging from A to Z, and omitting I, identify these smaller squares. This is the second letter of the two-letter system (Figure 6; Ordnance Survey n.d. (a)).

The smaller squares are themselves divided into smaller (10 km) grids, each of which is numbered starting at 0 in the southwest corner and moving east and north. A map reader can then identify a 10 km grid square by stating the two letters first and then the easting and northing grid numbers (e.g., SE 4 8). Finer resolution designations are made by estimating eastings and northings in 1/10-grid intervals, allowing the map reader to quote a location to an accuracy of 100 m on the ground.

The Ordnance Survey’s National Grid is based on the Transverse Mercator map projection. It covers all of Great Britain (Ordnance Survey n.d. (b)). Northern Ireland and the Republic of Ireland have their own National Grid system. Latitude-longitude coordinates for the true origin are located at 49°N latitude and 2°W longitude. Its false origin is just southwest of the Isles of Scilly, which was selected so that all the coordinates were positive (east of the origin at 00). As a result, 400 km are added to all eastings and 100 km subtracted from all northings. This
conforms to methods applied in other grid systems as well.

**Cartesian grid transformations**

Cartesian grids come in many forms and represent multiple projections. Because of this it is often necessary to alter the Euclidean geometry of such systems to account for differences in scale, adjustments of control points, rotation, and translation. Cases where such transformation come into play include rubber sheeting to allow co-locating identical objects digitized from two sets of unrectified imagery; projecting maps; digitizing maps either from a digitizer or using heads-up methods from a monitor; or printing and/or displaying a map using printer, plotter, or digital monitor coordinate spaces.

Although there are special modifications, the primary mechanism for making these transformations is the affine transformation. This is defined as “A geometric transformation that scales, rotates, skews, and/or translates images or coordinates between any two Euclidean spaces. It is commonly used in GIS to transform maps between coordinate systems. In an affine transformation, parallel lines remain parallel, the midpoint of a line segment remains a midpoint, and all points on a straight line remain on a straight line” (Wade and Sommer 2006). Affine transformation can translate (move), rotate, skew, scale, and skew coordinates (Figure 7). The affine transformation function is

\[
x' = Ax + By + C \quad \text{and} \quad y' = Dx + Ey + F
\]

where \(x\) and \(y\) are the input coordinates and \(x'\) and \(y'\) are the coordinates after transformation. The letters \(A, B, C, D, E,\) and \(F\) are determined by the control points of the source and destination control points and scale, skew, rotate, and translate the layer coordinates.

Coordinate distortions occur through input errors, map registration errors, lack of geodetic control, and many other causes. The process of rubber sheeting uses the basic transformational methods of scale, rotation, translation, and skew to adjust your data for improved accuracy. A major difference between rubber sheeting and transformation is that the distance features move
depends on their proximity to other objects. The process involves locking down locations that are not going to be moved and displacing those that need to be moved. Most often this is done using interpolation using temporary triangulated irregular networks.

One frequent transformation of Cartesian coordinates involves the conversion between map coordinates representing real Earth locations and those representing the map in the display space of the display device. One major difference that must be accounted for is that in normal Cartesian coordinate systems the origin is in the bottom left. In computer monitors, and other devices as well, the origin is on the upper right. As such, one additional transformational process that is required for this is called mirroring. The second major difference is that screen coordinates are recorded in pixels. The challenge then is to find a method to transform each point within the world rectangle to a point in the pixel rectangle and vice versa. Both of these transformations can be performed using affine transformation equations.

SEE ALSO: Map projections and coordinate systems

References


CARTESIAN COORDINATE SYSTEMS


Further reading

Maps can help to understand the world around us. They do so in an effective and efficient way, because they can summarize, clarify, explain, and emphasize aspects of our environment. Maps have many functions. They support navigation and decision-making, they tell stories of the past or assist in planning the future. Maps offer insight into spatial patterns and relations among the mapped phenomena, and do this well because they symbolize and abstract the reality represented.

Early definitions of a map describe it as a conventional image, mostly to scale and on a plane, of concrete or abstract phenomena that can be located in space. In this definition, conventional refers to agreements like “forest is green, and water is blue.” The International Cartographic Association provided the following definition: “A representation or abstraction of geographic reality. A tool for presenting geographic information in a way that is visual, digital or tactile” (Board 1990). The definition used here is based on the description in the book “Nature of Maps” by Robinson and Petchenik (1976), who defined a map as “a graphic representation of the milieu.”

Maps appear in many different media. Most common are maps on paper or on-screen. The first can offer a broad overview, but are more difficult to update. On-screen maps are often interactive and can be easily updated, but are limited by size. Many map types exist. They are usually divided into topographic and thematic maps. Topographic or geographic maps portray the landscape as accurately as possible with the map scale in mind. They cover, for instance, land use, hydrography, transportation, settlements, administrative boundaries, relief, and geographic names (Figure 1a). Thematic maps show the distribution and characteristics of particular phenomenon supported by (selected) topography as a base map (Figure 1b).

Maps are designed following the cartographic visualization process. This process is guided by the saying “How do I say what to whom” (Koeman 1969 in Kraak 1998). “How” refers to cartographic methods and techniques, such as generalization, and the choice of a map type. “I” represents the map-maker. In the past this was the sole domain of the professional cartographer. With the rise of geographical information systems other professionals became involved in the map-making process. Today, developments of the Internet, web 2.0, and mobile devices also allow the nonprofessional, with the help of tools like Open Street Map and mash-up technology, to create maps about virtually anything. These developments have resulted in a tremendous increase in the amount of maps produced and used.

How we “Say” things in a map refers to the symbology and semantics that represent the data. “What” refers to the geographic data and its characteristics. “Whom” refers to the potential map user, the context of its use, and its purpose. In the past any map created by a cartographer was supposed to be good enough for its intended use. Today, this supply driven approach has changed into a more user centered, demand driven map-making process. This has extended the original “How do I say what to whom?”
Figure 1 Basic map types: (a) topographic or geographic maps (source Google Maps); (b) thematic map. Both maps represent the Estonian island of Saaremaa. The topographic map emphasizes the structure of the landscape and the thematic map shows the administrative division of the island.

with “And is it effective?” In other words, cartography has become very much interested in how its products perform. This usability includes the effectiveness, efficiency, and satisfaction of the map products (Nielsen 1994).

Maps are created for many purposes, including navigation, education, urban planning, and weather forecasts. In these applications they are used for presentation and exploration (MacEachren 1994a). In their presentation role maps are designed to inform users about spatial patterns and relations. This reflects the traditional cartographic approach. The cartographer starts with a known dataset and has to select an appropriate visualization technique. The emphasis is on map design, resulting in high quality graphics presenting facts. The objective of maps designed for presentation is to communicate the information they represent. However, communication will never be perfect, even when one applies all cartographic rules. This is because information is lost, or perhaps even gained, during the communication process. The first happens because the cartographer can decide to ignore some information and/or because the map reader might not understand everything depicted. The second happens when the cartographer explains the original information and/or when the map reader combines the map data with his/her own prior knowledge.

Nowadays data are available from a variety of sources. However, one is not always fully acquainted with the data. The data has to be explored to find out if it is indeed useful to assist in the problem-solving at hand. The exploration process is often a search for patterns and trends. As a result one might end up with different alternative views that might provide a hypothesis. The emphasis is on enabling “discoveries.” An exploratory mapping environment is usually on-screen and preferably highly interactive. The maps are directly linked to databases that can be queried during the process. The graphic displays are not necessarily only maps. The exploratory cartographic visualization process is strongly information technology driven and, as such, closely linked to visualization in other disciplines. Over the last two decades progress
has been stimulated by developments in other fields such as scientific visualization and information visualization. The first of these deals with the visualization of 3-D phenomena with an emphasis on realistic renderings. The second is defined as the communication of abstract data by the use of interactive visual interfaces. This has resulted in a situation where maps are used in interactive environments to stimulate and improve explanation, insight, discovery, and decision-making.

In both presentation and exploratory mode, maps may be used to answer all kinds of geographical questions. Questions are asked because the map reader has certain tasks in mind in relation to the data mapped. Examples of these tasks are to identify the name of the province with the highest population, locate the province with the lowest population, or to make comparisons. Comparisons may be spatial (local or regional variations), based on multiple attributes (observe differences or similarities between the spatial distribution of possibly population and land elevation). Temporal comparison is applied to detect change, such as the population development over the years. To execute these tasks an extensive and flexible functionality should be available that allows selections to be made at will, based on criteria related to the attribute, locational, and temporal component(s) of the data.

However, in all circumstances, the maps should be well designed to play their role. The next section discusses how to prepare the data for an effective design.

Cartographic data analysis

When creating a map a set of constraints that will influence its design has to be considered. Among these are the purpose of the map (which questions have to be answered by the map), user characteristics (age, experience, knowledge of mapped area or phenomena), the use environment (in the field, or online) and the data characteristics. Each of these factors will influence the choice of the scale, a map projection, the required accuracy, and the map interface, and even the content. However, the data characteristics have most impact on the choice of the symbology, because qualitative data cannot be expressed in the same way as quantitative data. This section discusses how to select the appropriate symbology based on the data characteristics.

Cartographic data analysis starts with the determination of the data’s measurement level to get an idea of its character. For instance, the variable “population” in the table in Figure 5 is quantitative in nature. To allow the map user to experience these amounts the symbols to be used should give the map user the impression of different amounts. This can be realized by selecting the correct visual variable. In Figure 4a a point symbol that varies in size has been applied. The visual effect of this selection works well because a change in size is readily perceived as a change in amount.

Based on their measurement level, spatiotemporal data can be grouped into four categories: nominal, ordinal, interval, and ratio. Today, spatiotemporal data can also be derived from nontraditional sources such as photographs, video, and sound clips as well as texts. Nominal data are defined by qualitatively different characteristics. Examples include land-use categories such as road, hamlet, and border, as in Figure 2a, or the different municipalities of Saaremaa as in Figure 1b. There is no rank order, nor are arithmetic operations possible on nominal data. Ordinal data can be ranked in terms of relative quality, such as high and low, as in Figure 2b. Exact differences cannot be expressed and arithmetic is not possible. An example is a road classification such as highway, major road, and
**CARTOGRAPHIC DESIGN**

secondary road. Interval data are quantitative data. Here differences in amounts are known; however, an absolute zero is not. Distance is the defining characteristic and arithmetic manipulations possible are addition and subtraction. Examples are temperature expressed in Celsius, as in Figure 2c. The ratio scale is also quantitative but is defined by proportions. Differences in amounts are known as well as an absolute zero point. All arithmetic manipulations are possible. Examples are the number of inhabitants of provinces (Figure 4a), or the percentage of native speakers. Figure 2e shows the relation between the measurement levels, their perceptual properties and the visual variables.

The next step is to convert the measured character of the data into symbols. This has to be done so that perceptual properties of the symbology allow the map reader to readily understand the character of the data (e.g., the measurement level). In the sample maps in Figure 8a the varying size of point symbols has be used to depict the

![Figure 2](image)

**Figure 2** Measurement levels and perceptual properties: (a) nominal, (b) ordinal, (c) interval, (d) ratio, (e) the relation between the measurement level, perceptual properties, and visual variables.
number of inhabitants. Besides point, lines and area symbols are used to represent the data. Text is often used as an additional option to transfer information. All symbols can be varied using six basic visual variables: size, value, grain/texture, color, orientation, and shape (Bertin 1967). Figure 2a links the perceptual properties of the different measurement levels to these variations in symbolization. For instance, it shows that color is useful to express differentiation but ineffective to express proportions or distance (quantities) or order. Size on the other hand is not suitable to express differentiation but works well to express order, distance, and proportion. The notion of color seems straightforward, and yet it is complex. Color as used by Bertin is also known as color hue (the dominant wavelength), while his variable value is also known as color value (white added to the hue). Additionally there is color saturation (black added to the hue).

Figure 3 presents how, in the context of a map, these variations work for point, line, and area symbols, as well as text. The far right of the figure displays how these variations and their perceptual properties can be linked to the measurement levels of the data, being proportional, distance, order, and differentiation.

If the symbols to which a visual variable has been applied are spontaneously perceived as

![Figure 3](image-url)

**Figure 3** The visual variables: left as applied to point, line, and area symbols and text; right in relation to their perceptual properties.
different the variable is differentiating. Figure 3 shows that shape, orientation, and color are differentiating. Figure 1b shows an example: color differentiates the individual municipalities of Saaremaa. A visual variable is ordered if spontaneously all symbols to which it has been applied can be placed in an unambiguous order. Examples are from dark to light, small to big, or from coarse to fine. In Figure 3 it can be seen that size, value, and texture are ordered while color, orientation, and shape are not. An example can be found in Figure 10a, displaying the population density of Estonia’s provinces. If spontaneously all symbols to which a visual variable has been applied can be placed in an unambiguous order, and an estimate of the distance between the symbols within the range can be made, it enables distance perception. As Figure 3 shows, this is true for size, value, and texture, but not for shape, orientation, and color. Examples are both Figure 4a and Figure 10a. A visual variable is proportional if the differences

Figure 4 The use (a) and misuse (b and c) of the visual variables: (a) size used to express the number of inhabitants per province; (b) color used to express the number of inhabitants per province; (c) shape used to express the number of inhabitants per province.
between the symbols to which it has been applied can be expressed in distinct amounts. Figure 3 shows this is only possible with size. Again, Figure 4a is an example.

At the far right of Figure 3 it can be seen that qualitative data (the nominal measurement level) link to shape, orientation, and color. Quantitative data (the ordinal, interval, and ratio measurement levels) link to size, value, and texture. It is important to note that ratio data can be split into absolute and relative data. Absolute quantities are observed, measured, or counted and relative quantities are calculated: densities, ratios, percentages, and averages. Size should be used for absolute ratio data (Figure 7a) and value for relative ratio data (Figure 10a). Often it is seen that value is also used for absolute ratio data (Figure 10c), but when the geographical units to which it is applied are not all of the same size the impression of the map will be wrong.

What if these guidelines are not applied? Figure 4 shows one correct map and two examples of how not to apply the visual variables. The population data to represent are quantitative data, which should be represented by symbols changing in size as is correctly done in Figure 4a. The question “Which province has most inhabitants?” can be immediately answered. Figure 4b and 4c illustrate that color and shape are not suitable. Getting an answer to the question is not easy, the information is somehow embedded in the map, and if the legend is studied well an answer can be found, but not efficiently (i.e., in a short time and correctly).

There is some debate whether the six variables as defined in 1967 by the French cartographer Bertin can cover all variations possible (MacEachren 1994b; Tyner 2010). This debate gains additional impetus from the fact that on-screen maps offer options not possible on paper, for which the six variables in Figure 4 were originally developed. Examples of new variables are transparency and focus. Transparency can be used to show multiple symbols/layers on top of each other to show direct relation between the topics (Figure 8d). Focus refers to the clarity of the symbol. Giving a symbol a fuzzy appearance can for instance be used to display uncertainty. Additionally symbols can blink or be highlighted to attract attention (Kraak and Ormeling 2011).
CARTOGRAPHIC DESIGN

It has to be realized that application of the above guidelines does not always result in clear maps. Especially when multiple topics are included in a map, they may become complex. This led Bertin to distinguish between “maps to see” and “maps to read.” Maps to see will give insight in the data in a single instant (Figure 4a). Maps to read have to be studied (Figure 8d). A map to see can often only be created based on a single variable (or component), and places constraints on how many variations can be applied on the symbols.

How to execute a cartographic data analysis? In the next paragraph, examples are given based on the data shown in Figure 5, which has a base map of the Estonian provinces and a table with some population statistics data from 2012.

Figure 6a represents a workflow to execute cartographic data analysis in this case. Figure 6b and 6c are elaborated examples for quantitative and

![Figure 6](image)

Figure 6 Cartographic data analysis in practice: (a) the principle; (b) an example with quantitative data; (c) an example with qualitative data.
CARTOGRAPHIC DESIGN

The first step is to establish what the dataset is about. In Figure 6b it is the population of Estonia’s provinces and in Figure 6c it is the regions of Estonia. Bertin refers to this as the invariant, which will appear as the title of the map. In Figure 6b the measurement level of the variable “population” has to be established. These variables, or subtopics, are defined by Bertin as components. It can be seen that “population” is made up of absolute numbers (quantitative data), which makes it a ratio level. In the flow diagram’s box “measurement level,” two other columns can be observed: length and range. For qualitative data, length represents the number of categories; for quantitative data, range represents the interval between the smallest and highest value. Both values will decide if the data have to be grouped or classified before mapping. In the example, the range is 8482–55 2927. The flow diagram also has a geographic component, which will be implicitly available in all maps. It refers to the number of geographic units for which one has data. In the examples the length of the geographic component is 15.

In the next box of the figure the measurement level is linked to a visual variable with the required perceptual properties. In this example the component “population” needs a visual variable with absolute proportional properties. This makes size the only suitable variable. In the resulting map, shown right of the flow diagram size is applied on point symbols. In Figure 10c the same process is repeated for a quantitative variable “region.” The analysis of this variable “region” follows the same path. The measurement level is nominal (qualitative). It needs a visual variable that can differentiate between the regions. Three options are available, orientation, shape, and color. The last variable has been selected and has been applied to the regions, as can be seen in the map in Figure 6c.

The outcome of the analysis, for example, measurement level, the number of components one intends to include, the need to use point, line or area symbols directs the choice of the thematic map type. This choice is also related to the elementary nature of the data, as being continuous or discrete. Continuous data (or “fields”) extend over space without interruptions and values are simply observations at sample points. Examples are heights or temperature. Discrete data are distinguishable for individual entities, such as administrative units like provinces or individual objects such as towns. For a detailed discussion on different thematic map types see Slocum et al. (2008).

Design

The cartographic data analysis suggests which visual variable to use for which component and what symbol set and map type. However, it does not tell us anything about the actual graphic design, which dictates the look of the map. The map in Figure 6b is very basic, not attractive, and as a result inefficient. All lines have the same weight, all polygons have same fill color. In Figure 7b, some colors have been used but not necessarily according to any plan. To create efficient and effective maps, how the map should look must be considered. The map layout, the choice of line width, the selection of colors and fonts, the hierarchical organization of the map content, must all be considered to create an attractive map. This is the actual map design process. Despite all information to visualize it is important to keep the map as simples as possible, and avoid what Tufte (1983) called “chart junk.” Examples are abundant colors, overuse of shading, fancy lines, and exotic fonts.

What is a good design? If one asked several professional cartographers to finish the
map in Figure 6b it is likely that ten excellent but different results would be given. This is because there are no fixed “rules” or “restrictions.” A background in white or dark blue, the proportional symbols as diamonds, squares or circles, fonts in Arial or Times, all can be acceptable. One of the first design actions would be to visually organize the data such that the most important information is observed first. This introduces the design principles of figure-ground, which makes the reader spontaneously follow a sequence from important to less important. Appropriate figure-ground can be achieved by applying visual hierarchy as illustrated in Figure 7. In this figure the left map has no hierarchy applied and in the right map the information has been ordered. Below, the base layer with Estonian provinces, followed by the text layer with names of the provinces. The next layer has the proportional point symbols representing the number of inhabitants, followed by map title and legend. At the top of the hierarchy is the highlighted symbol of Tartumaa.

Figure 8 gives four examples of different but correct designs. The map titles and legend have been left out on purpose. Figure 8a is the color version of Figure 7a. The figure ground principle is applied by using a light blue for the Estonia, dark blue for the proportional point symbols to make them stand out from the base map. Figure 8b has different colors, gray for the base map and orange for the proportional point symbols, which basic shape is now a square. Figure 8c is similar to Figure 8a, but with different colors and shading added to the proportional point symbols which creates a stronger visual hierarchy. The map in Figure 8d follows yet another approach. The base map is a choropleth map representing the provinces’ population density (relative ratio data). The visual variable value is used in a scale from light blue for low densities to dark blue for high
densities. On top the proportional point symbols representing the number of inhabitants (absolute ratio data). For the selection of a proper color scale see www.colorbrewer2.org. The point symbols have been made slightly transparent and allow the map layer below to be seen. This is especially useful for the province with the highest value where the symbol is covering the province almost completely.

Applying visual hierarchy alone is not enough. A few other design principles have to be adhered to as well. It is obvious that all map content has to be presented in a clear and unambiguous manner. The map should have a balanced design and have contrast to be able to distinguish the information content. A well-designed map is not overcrowded with symbols. This is especially applicable for on-screen maps, as they are presented on a relatively small space. These maps can be “empty,” as access to other information is available via a mouse click and can be acquired at will. With today’s options to bring together information from many different sources, it is often tempting to also show all these layers in a single map, which may result in a cluttered and incomprehensible map (Figure 9).

In Figure 10a a choropleth map displays the population density of Estonian provinces in four
classes based on the data in the table in Figure 5b. Several classification methods exist to decide on the distribution of the data over those four classes. In this example a nested means classification has been applied where the average of all observations (here 24) is first defined, followed by the calculation of the average of the range below and above the overall average, resulting the classes as show in Figure 10a. Each class has been given a color defined via colorbrewer2.org. Next to a legend the map has a title in the upper right of the map, and a scale bar in the lower left. A north arrow is usually added, but often omitted when the map is north oriented. A grid with longitude and latitude could also be added. The location of this so-called marginal information depends on the space around the map. The title is preferably positioned at the upper left, for example, where one normally starts to read a document; however, in this example space does not allow this due to the shape of the mapped area. Space limitations are often the reason why legends are missing when the map is presented on-screen. The scale of the map is always relevant, and as scale indicator a scale bar is preferred because it will scale with zooming. Reference to the year of production is often needed to determine how up-to-date the map is and information on the producer can reveal something about the credibility of the map.

In Figure 10b a wrong visual variable is applied. Not value, as correctly done in Figure 10a, but color instead. The effect is that an overview
Figure 10  Good and bad and wrong designs: (a) a well-designed choropleth map; (b) choropleth map with wrong visual variable applied; (c) choropleth map with wrong data; (d) choropleth map with overdone design.

of low and high densities is no longer visible. Figure 10c show a choropleth map design, but based on absolute ratio (the number of inhabitants). This is a common error that results in a wrong impression of the mapped phenomena. An example illustrates this: when two districts, one of which is small and the other very large, have the same population, they will get the same color value in the map, but the large district will attract most visual attention, giving it a dominance over the small, but population-wise equal, area. Compare this in Figure 10c and Figure 8a. Figure 10d is an example that contains “chart junk.” Shading has been applied to all individual map items, resulting in a strange stepped effect in the map. The map is also visually unbalanced because the marginal information is distributed more or less randomly.

Online maps have many advantages when it comes to dissemination of the information and the (visual) exploration of the data. Currently, several solutions exist that allow one to publish their own data online, and interactively share it with others. Examples are CartoDB (Figure 11) and GeoCommons. The user has some simple choices to display the maps and, although the design options are limited, they generally result in acceptable maps.
For the more advanced user exploratory environments, such as offered by the Estonian bureau of statistics (Figure 12), are available. These mapping environments have a much steeper learning curve. From a design perspective default settings are provided, but during the exploration many options are offered which do not necessarily end in well-designed maps. However, if one has an exploratory objective in mind these tools can be very helpful to find and understand spatial patterns and relations. 
CARTOGRAPHIC DESIGN

SEE ALSO: Choropleth map; Color theory; Scale; Visual variables; Visualization; Web-mapping services

References


Further reading


Websites


Cartographic modeling

C. Dana Tomlin
University of Pennsylvania, USA

Cartographic modeling is a set of methodological concepts, conventions, and capabilities intended to generalize and standardize the use of geographic information systems (GIS) for the interpretation of geospatial data. It is an application-independent methodology that attempts to address a wide variety of interpretive tasks in a clear, consistent, and comprehensive manner. To do so, it establishes a system of elementary components that can be combined with relative ease and great flexibility. These include components relating to data, data processing, and data processing control.

Data

The cartographic modeling methodology does not prescribe particular data formats, but it does assume that data are organized according to a certain general construct. This is a hierarchical construct at the base of which lies the “cartographic model.” In its most common form, a cartographic model is simply a body of data pertaining to a designated geographical locality or “study area.” All of the geospatial data representing a given study area are associated with “layers,” each of which can be envisioned as a map of a single site characteristic that varies over that area. Like a traditional map, a layer is a planimetric depiction of the area that characterizes every location within it. Unlike most traditional maps, however, a layer records only one characteristic for each of its locations. Thus, the variety of topographic and demographic characteristics that might well appear together on a traditional map would instead be presented by way of a separate layer for each.

Every layer is represented by a “title” in the form of a distinguishing string of text, and it is stored with “metadata” that indicate things about the layer itself: its source, publication date, orientation, resolution, the time period to which it refers, and so on. What a layer conveys about its study area, however, is represented as a set of one or more “zones.”

A zone is the set of all locations for which a particular geographical condition is recorded. On a layer entitled SoilType, for example, each zone might represent a particular type of soil. A layer of SoilDepthToBedrock, on the other hand, would probably include zones representing distances beneath the surface of the Earth. Zones are represented by text string “labels” such as PaxtonFineSandyLoam or FreetownMuck. These can be assigned arbitrarily as long as each is unique among those within its layer. More importantly, every zone is also associated with a numerical “value” that distinguishes the zone from others within the same layer. Like labels, values can be arbitrarily assigned; unlike labels, they can also be computed.

The use of numerical values (as opposed to colors or graphical symbols) to represent geographical conditions is one of the things that most distinguishes cartographic modeling from its nondigital antecedents. It is also one of the things that enable this methodology to take advantage of the precision, flexibility, and processing power of modern computation. Zonal values can be
CARTOGRAPHIC MODELING

recorded either as integers or real numbers, and they can relate to nominal, ordinal, interval, or ratio scales of measurement that are either linear or cyclical. A special “null” value can also be used to represent the absence of meaningful data.

The final component of data within a cartographic model is the set of (what have already been casually referred to but now must be more formally recognized as) “locations.” A location is that particular portion of a study area (or a cartographic model representing such an area, or a layer within such a model, or a zone within such a layer) that can be uniquely defined by a pair of geographical “coordinates.” A coordinate is a number indicating position along one of the two perpendicular dimensions. To the extent that this number is of finite precision, it actually represents an interval or range of positions along that dimension. Thus, each location can be envisioned as a finite, 2-D portion of a study area, layer, or zone. If the coordinates that define locations refer to perpendicular and similarly calibrated dimensions, then the area uniquely associated with each location will be square, and the overall pattern of locations will be a rectilinear grid of perpendicular and equally spaced “columns” and “rows.” Alternatively, a location can be regarded as a sample point at the precise position defined by its two coordinates, and locations can thus be redefined by resampling.

Though the manner in which locations are encoded need not affect the organization of a cartographic model, it can have a very significant effect on the manner in which they are processed. For that reason, it is useful to distinguish at least between “raster” and “vector” formats.

In raster format, locations are usually called “grid cells” or “pixels,” and each is often regarded as a discrete geometric entity. In vector format, however, the degree to which individual locations are recognized as such depends on whether the zones containing them are encoded as points, lines, or polygons. While a vector-encoded point feature may well be fully defined by a set of explicitly encoded locations, line and polygon features are not. Most of the locations along a line or within a polygon are never explicitly recorded; instead, they are only implicitly referenced by a (typically much smaller) set of explicitly encoded vertices.

The practical implications of this are apparent: encoding locations in a vector format is generally more efficient and flexible than doing so in a raster format. The conceptual implications are less apparent but perhaps even more important. Whereas vector encoding casts the world as a set of distinct geometric objects or “features,” raster encoding presents a world of conditions that vary over spatially continuous “fields.” The former records locations associated with each of set of conditions, while the latter records conditions associated with each of set of locations. While it is not at all difficult to reconcile these perspectives in terms of the way in which data are stored, the two are decidedly different in terms of the way in which data are processed.

Data processing

If the purpose of a cartographic model is to store a body of data, then the purpose of cartographic modeling is to interpret those data: to convert recorded facts of potential utility into recorded facts of actual utility and, thereby, transform data into information. This is generally a matter of making implicit qualities explicit by analyzing and/or synthesizing the existing layers of a cartographic model in order to generate new layers.

Along with data preparation and presentation, data interpretation is one of the major functions of any modern GIS. What distinguishes cartographic modeling is the manner in which it organizes such interpretive capabilities.
Fundamental to this organization is use of the layer as the primary unit by which data are not only stored but also processed. Cartographic modeling capabilities are presented as a small but highly integrated set of “operations,” each of which performs a relatively primitive task, but all of which do so by taking one or more existing layers as input and generating a single new layer as output. Since the output of any one operation can, therefore, potentially be used as input to any other, only a small set of these elementary operations is required to perform an open-ended variety of arbitrarily complex tasks. Just as primitive algebraic operations (such as addition, subtraction, multiplication, or division) can be combined to form much sophisticated mathematical functions, primitive cartographic modeling operations (such as the assignment of zonal values, the superimposition of layers, the measurement of size, the determination of distance and direction, the calculation of travel costs, the characterization of shapes, the computation of slope and aspect, the delineation of viewsheds, and the simulation of movement) can also be combined to form more sophisticated “procedures.”

Cartographic modeling operations are deliberately defined in terms that relate more to the structure than to the substance of the data to which they apply. Each is defined by describing its effect on a single typical location, understanding that all locations are being processed in the same manner and (conceptually at least) at the same time. This use of locations (rather than points, lines, or polygons) as elemental units of space results in a style of processing oriented more toward fields than features. It also promotes a location-centric or worm’s eye perspective on geospatial computation that is subtly yet significantly different from the map-wide or bird’s eye perspective more commonly employed. Consider, for example, an operation that “surrounds buildings with rings of increasing distance.” The same operation could be also described as one that “measures each location’s distance from the nearest building.” While those two descriptions are equivalent, the latter reflects a point of view that is oriented more towards the locations for which new values are about to be computed than the locations from which those values are about to be generated. For all of its simplicity, it is this worm’s eye perspective that establishes the conceptual foundation for a highly efficient and effective suite of cartographic modeling operations. The suite includes operations of three major types: “local,” “zonal,” and “focal.”

Local operations compute new values on a location-by-location basis. Each generates a new layer on which every location’s value is computed from the value(s) of that same location on one or more existing layers. While some of the local operations provide for the explicit assignment of new values, most call for their calculation by way of familiar mathematical functions. Local-Classification, for example, is an operation that might be used to assign user-specified scores to different types of soil. The resulting layer of scored soils might then be combined with layers representing other site suitability factors by using an operation called LocalMean.

Zonal operations compute new values on a zone-by-zone basis. Each generates a new layer on which every location’s value is computed from the value(s) of all locations on one existing layer that share its zonal value on another existing layer. Many of the zonal operations are directly analogous to local counterparts; they call for the explicit assignment of new values or their calculation by way of familiar functions. Zonal-Classification, and ZonalMean, for example, are operations that might be used to summarize site suitability scores within each of a set of towns.

Several of the zonal operations go beyond this kind of zone-wide summary to characterize each location’s existing value in relation to those of
the others within its zone. *ZonalRanking* is an operation that might be used to sort site suitability scores within each of a set of towns, for example, while *ZonalPercentile* might be used to set each location to a value indicating how much of its town has a lower score.

Focal operations compute new values on a neighborhood-by-neighborhood basis. A “neighborhood” in this context is the set of all locations that are situated at designated distances and/or directions from a given location, which is then referred to as the “focus” of that neighborhood. Those distances and directions may either be the same for all such foci or vary from one to another. Each of the focal operations generates a new layer on which every location’s value is computed from the values of neighboring locations on an existing layer, and there are focal equivalents to many of the local and zonal operations that summarize existing values. *FocalMean*, for example, is an operation that could be used to set each location to the mean of its nearby site suitability scores.

There are also focal counterparts to those zonal operations that characterize each location in terms of the relationship between its own existing value and those of the other locations being summarized. *FocalPercentile*, for example, might be used to set each location to a value indicating how much of its neighborhood has a lower suitability score.

A number of additional focal operations consider not only the value of each neighboring location but also (or instead) its distance and/or direction from the neighborhood focus. *FocalGradient*, for example, is an operation that might be used to calculate the rate at which site suitability scores change over space.

Focal operations vary as well according to the nature of the neighborhoods to which they are applied. “Local” neighborhoods include only those locations that lie adjacent to each neighborhood focus, while “lateral” neighborhoods extend that reach to encompass all neighboring locations lying at specified distances and/or directions. “Visible” neighborhoods limit what might otherwise be regarded as radial neighbors to those that lie within sight of each neighborhood focus, given an intervening landscape of visual obstructions. And “accessible” neighborhoods are defined in terms of distances that are measured not in units like meters or miles but in units like minutes, dollars, ergs, or levels of environmental impact. These are units that accumulate as a consequence of implied motion from each neighborhood focus and do so at rates that can vary according to conditions impeding that motion.

Given an ability to specify sequences of local, zonal, and/or focal operations that generate new layers from existing layers, it becomes possible to draw an important distinction between “actual” layers that are stored as data and “virtual” layers that are stored as procedures to be applied to actual data. In order to include such virtual layers in a cartographic model, the procedures involved must be recorded in a format that is not only legible to humans but also executable by machine—a language for data processing control.

**Data processing control**

The notation used to represent cartographic modeling data and data processing activity is not intended to replace the programming language employed in any given GIS. Rather, it is intended to relate to such languages in a manner that transcends their minor differences in order to focus on fundamental commonalities. It is a pseudocode that employs “statements” to represent individual operations and sequences of statements called “programs” to represent procedures. Statements are given in an imperative form.
much like that of conventional algebra but with greater use of words than symbolic characters. The formal syntax of this “map algebra” is more fully described elsewhere but is illustrated below with a typical example. Here, a new layer entitled *Summits* is generated from an existing layer entitled *Elevation* by applying a procedure that:

1. smooths a 3-D surface of topographic elevations,
2. calculates local deviations between the original and smoothed surfaces,
3. relates those deviations to a scale from 0 to 100,
4. identifies areas of high positive deviation as hills,
5. distinguishes each hill from the others,
6. determines the maximum elevation of each hill, and
7. isolates the location(s) situated at each hill’s maximum elevation.

Note that this notation refers to layers by their titles. Note, too, that this particular program refers to its input layer by way of a title (*Elevation*) that could either identify a specific layer that is currently available or instead represent a generic layer that has yet to be provided. A program that refers only to specific input layers can be regarded as a (virtual) layer itself. A program that refers only to generic input layers, however, acts more like an operation than a layer. This is an important distinction that effectively broadens the definition of a cartographic model to include what amount to nouns and verbs of up to four major types:

- standard operations stored as GIS-based functions,
- special operations stored as programs with generic inputs,
- actual layers stored as data, and
- virtual layers stored as programs with specific inputs.

**Development**

The cartographic modeling paradigm can be traced to manual techniques that were employed by landscape architect Warren Manning in the early 1900s and popularized by landscape architect Ian McHarg in the 1960s. Its digital antecedents start with SYMAP (Fisher 1966) and extend in a direct lineage to GRID (Sinton and Steinitz 1969), IMGRID (Sinton 1977), and MAP (Tomlin 1980). The term “cartographic modeling” was introduced in the late 1970s and the methodology was refined throughout the 1980s (Tomlin 1983, 1990).

Since then, cartographic modeling capabilities have been widely adopted and adapted. They remain primarily associated, however, with data that are encoded in raster format. Much of the reason for this relates to the simplicity and generality of that format – one in which all manner of geographical phenomena are represented by the same type of geospatial unit. When working with vector-encoded data, units defined
CARTOGRAPHIC MODELING

by points, lines, and polygons present a much greater challenge, particularly with respect to focal operations. This is a worthwhile challenge, however, and a challenge that can only be met by recognizing that the world outside is one of both fields and features.

SEE ALSO: Data structure, raster; Data structure, vector; Geodesign; Geographic information system; Map algebra; Quantitative methodologies; Spatial analysis

References


Further reading


As a rubric, History of Cartography has two distinct connotations: (i) the chronicle of technical and intellectual developments surrounding mapmaking and map use and (ii) the interdisciplinary endeavor that invokes historical context in studying both the map as a visual object and the societal impacts of cartographic technology and mapping activity. (It should not be confused with historical cartography, the mapping of past events by historians and historical geographers.) A third and equally pervasive connotation is the comprehensive History of Cartography initiated by J.B. Harley (1932–1991) and David Woodward (1942–2004) and published by the University of Chicago Press. Because the several volumes of the History reflect major divisions in map history, the series provides a convenient framework for much of this entry, which also identifies important English-language synopses and appropriately emphasizes developments since 1900, including dynamic maps and geospatial technology. This entry also examines academic activities, including conceptual frameworks, key individuals, and the institutional apparatus of map collections, exhibitions, conferences, and journals.

Encompassing geospatial technology, this entry demands a definition of map with two levels: the narrow, traditional sense of a graphic framework whereby symbols representing geographic features facilitate navigation, planning, and general understanding (Harley and Woodward 1987, xv–xvii) and the broader, more recent sense of topologically structured geographic information intended to support spatial analysis, graphic display, or both. The word cartography merits a similar disambiguation insofar as it encompasses not only the traditional focus on mapmaking but also the broader context of map use and mapping as a governmental, commercial, and intellectual endeavor.

Frameworks and chronologies

Systematic study of salient trends in map history involves the overlay of three factors: a classification scheme that recognizes the diverse ways in which maps are produced and used; a chronological schema akin to the eras or stages of development common to broad-brush historical narratives; and distinctive themes such as the impacts of technological innovations like printing and photography and political revolutions like the rise of the nation state. Science, technology, and state formation have precipitated distinctive changes in cartographic symbols and formats as well as in the social relations surrounding maps and mapping.

Modes of mapping practice

How maps are made and used, and to what ends, is arguably the most revealing basis for sorting cartographic artifacts into categories for historical analysis. Competing categorizations include map scale, which is both the lynchpin of cartographic theory and a key characteristic of any map. Even so, categories based on scale are too coarse and internally varied to sustain a broad range of meaningful interpretations and
conclusions. Similarly limited are categorizations based upon whether the mapmaker is a governmental or a private-sector entity, whether the map is circulated on paper or electronically, or whether replicated images were copied by hand, engraved on a copperplate or lithographic stone, or transferred photographically. Although the reproduction method leaves distinctive marks on the artifact, these differences in appearance are less socially and intellectually significant than differences among cartographic modes such as property mapping, marine charting, and geodetic surveying.

The concept of cartographic modes was introduced by Matthew Edney (1993) as a framework for studying the development of European cartography between 1500 and 1850, which includes the Age of Discovery and the rise of capitalism. His schema initially consisted of nine modes of cartographic practice, defined as “sets of cultural, social, and technological relations which define cartographic practices and which determine the character of cartographic information” (Edney 1993, 54). Although these modes overlap and interact, each has a typical range of map scales as well as distinctive forms, innovations, and user communities. Celestial mapping, with ties to astrology and astronomy, looks heavenward and has the smallest scales. Small-scale maps are also the most characteristic product of geographical mapping, which summarizes spatial knowledge for the entire world and its regional subdivisions. Geodetic mapping, which measures the size and shape of the planet, is a continental or global endeavor most conveniently depicted (if at all) on small-scale maps; even so, its most important products are the highly precise triangulation networks that provide a geometric framework for the large- or medium-scale maps generated by topographic mapping, which focuses on the representation of places and the Earth’s surface, and by urban mapping, a historically important endeavor focused on cities. Measurement is no less basic to property mapping, which promotes land ownership with large-scale representations, and boundary mapping, which delineates national territories and their provinces on medium- or small-scale maps. Marine charting, which produces small-scale maps of the seas and medium-scale maps of harbors, interfaces with topographic mapping along the coastline. Thematic mapping, which encompasses the broadest range of scales, typically exploits topographic maps and marine charts as spatial frameworks for plotting natural or social phenomena.

As Edney (1993, 59–64) observed, these modes represent a convergence of mapping practice around 1500 followed by a marked divergence before 1800. By the sixteenth century the collective emergence of mercantilism, printing, the commodification of land, and the modern territorial state had led to four distinct modes – charting, geodesy, small-scale chorography (place mapping), and large-scale topography – which by the early eighteenth century had amalgamated into a relatively amorphous “mathematical cosmography” inspired by the Scientific Enlightenment and exemplified by systematic and detailed national surveys like the mapping of France by the Cassini family. Fragmentation into the aforementioned modes by 1800 reflects the increased political, economic, and intellectual complexity of European society.

Edney’s schema also included four institutional endeavors that cut across multiple modes. Military cartography, for instance, employs geographical maps for strategic planning, topographic maps for tactical operations, and marine charts for naval activities. At local as well as provincial and national scales, administrative cartography intersects property mapping, thematic mapping, topographic mapping, and urban mapping. The map trade, which became map publishing after printing and engraving technologies vastly expanded the
cartographic marketplace in the sixteenth century, encompasses the commercial distribution of maps of all types, and map collecting by institutions and individuals brings together maps useful for military intelligence, general reference, and scholarly research. And in the twentieth century in particular, antiquarian dealers have touted rare and merely old maps as aesthetically attractive investments.

Technological advances in the twentieth century required two additional modes of cartographic practice: overhead imaging, whereby aircraft and satellite platforms and diverse imaging technologies vastly increased cartographic coverage and content, and dynamic cartography, whereby interactive and animated displays offer comparatively complete and often engaging displays of temporal or complex phenomena. In addition, academic cartography emerged as a distinct institutional endeavor focused on the systematic, scholarly study of not only the design and production of maps but also the cultural connotations and societal impacts of mapping activities.

**Historical approaches**

The scholarly study of map history has roots in the eighteenth century, but systematic treatments were largely lacking until Manuel Francisco de Barros e Sousa Santarém (1791–1855), the Portuguese scholar-diplomat believed to have coined the term cartography to describe the study of old maps, published his *Facsimile-Atlas* of early maps in 1841. Another milestone was *l’Atlas des monuments de la géographie*, published in 1879 by the French civil engineer Edmé-François Jomard (1777–1862), who assembled a collection of facsimiles to emphasize cartography’s close ties to progress in science and discovery. In turn, Jomard’s atlas was the source of several reproductions in the *Facsimile-Atlas to the Early History of Cartography, with Reproductions of the Most Important Maps Printed in the XV and XVI Centuries*, published in 1889 in Swedish and English by Adolf Eric Nordenskiöld (1832–1901), a Finnish natural scientist, arctic explorer, and map collector. Although many of Nordenskiöld’s images were necessarily reduced in size, the *Facsimile-Atlas* (1889) was a useful general reference for librarians, historians, and map collectors. Systematic cataloging of old maps fostered problematic notions of inevitable, ever increasing cartographic progress propelled by advances in mathematics and natural science, measuring techniques, and graphic reproduction.

Facsimile atlases were followed by narrative histories, such as Lloyd A. Brown’s *The Story of Maps*, published in 1949. A former map curator at the University of Michigan’s William L. Clements Library, Brown traced cartography’s evolution as a modern science from the ancient geographies of Strabo and Alexander the Great to the International Map of the World, a heroic multinational endeavor frustrated by two world wars, limited cooperation, and lack of a clear-cut need. Military strength and world peace, he argued, depended on reliable maps, and fully complete and reliable maps were impossible without peace and prosperity. Like most narrative histories, Brown’s *Story* put maps, geometric accuracy, and cartographic coverage at center stage and paid little attention to the complexity of maps as social and cultural texts.

Erwin Raisz offered a more schematic treatment in his 1938 textbook *General Cartography*. Raisz was a Harvard faculty member as well as an accomplished illustrator, widely recognized for his insightfully detailed physiographic diagrams. His calligraphic dexterity is readily apparent in elegantly hand-lettered timelines for Antiquity (300 BCE–300 CE) (Raisz 1938, 15), the Middle Ages (400–1400) (1938, 25), the
CARTOGRAPHY: HISTORY

Renaissance (1470–1700) (pp. 36–37), Modern Maps (1700–1930) (pp. 52–53), and the United States (1750–1930) (pp. 60–61). Most events are tagged with the name of a specific mapmaker; his Renaissance timeline includes separate columns for England and France; its Modern Maps counterpart assigns noteworthy developments to Italy–Spain–Latin America, France, Netherlands–Germany–Austria, Great Britain, and “Others,” whereas his United States timeline is divided equally between “Surveys, Official Maps” and “Private Cartography.” This emphasis on persons, national traditions, and technology is typical of the “internal histories” that arose in the late nineteenth century, when mapmakers began to enrich their manuals and textbooks with historical narratives of steady progress in overcoming diverse challenges.

J.B. Harley and David Woodward questioned the reliability of internal histories, which not only emphasize scientific progress, famous firsts, and intriguing artifacts, but also tend to ignore the influence on mapmakers and mapping institutions of intellectual and political movements. Harley was a historical geographer with Marxist leanings, whose careful examination of late eighteenth-century English county maps and detailed maps produced during the American Revolution aroused a critical appreciation of old maps as historical sources with limitations beyond the obvious consequences of imprecise instruments and sloppy fieldwork (Edney 2005b). Woodward was a historian of cartography whose research on map engraving and printing inspired a wider interest in how maps were circulated and used (Edney 2005a). In spring 1977 he persuaded Harley to forego plans for a four-volume history of mapping in North America and collaborate on a general history of cartography focused on the societal context within which maps were produced and consumed (Woodward 2001, 23–24). They planned to cover the period from prehistory through the end of World War I in four books, divided chronologically at 1470, 1660, and 1800 and totaling no more than a million words. Their hope of having all volumes in print by 1992, the 500th anniversary of Columbus’s widely celebrated voyage to North America, greatly underestimated the richness of map history as an academic endeavor.

A “big book” project

Their million-word estimate proved similarly naive insofar as the History of Cartography, when completed in the early 2020s, will be at least seven times as large, and take its place alongside monumental “big book” reference works like the Dictionary of American Regional English (1985–2012) and the Dictionary of Scientific Biography (1970–1990). From the beginning Harley and Woodward recognized that their project would require numerous collaborators, grant support, and a committed and competent publisher (Woodward 2001). The University of Chicago Press, with which Woodward had worked, filled the publishing role, and “the History Project” (as it came to be known) has drawn substantial support from the National Endowment for the Humanities, the National Science Foundation, and the University of Wisconsin–Madison, which added Woodward to its geography faculty in 1980.

Volume One, titled Cartography in Prehistoric, Ancient, and Medieval Europe and the Mediterranean, appeared in 1987. Its 21 chapters are largely interpretative narratives focused on a region, a time period, a type of map use, or perhaps all three, as exemplified by the title of chapter 12: “Maps in the Service of the State: Roman Cartography to the End of the Augustan Era.” Topics covered in various chapters include the measurement and calculation methods of the
Babylonians and Greeks, Ptolemy’s *Geography*, maps for urban planning and civil engineering in the early Roman Empire, Woodward’s recognition that medieval *mappaemundi* were symbolic representations of history rather than geometric representations of the planet, the development and use of *portolan* sailing charts in the fourteenth and fifteenth centuries, and an increased number and variety of maps overall.

In their preface, Harley and Woodward proclaimed a fundamentally new definition of *maps* as “graphic representations that facilitate a spatial understanding of things, concepts, conditions, processes, or events in the human world” (Harley and Woodward 1987, xvi), and in chapter one, “The Map and the Development of the History of Cartography,” Harley appealed for recognition of map history as a distinct scholarly pursuit devoted to questioning the map’s role in sociopolitical discourse. Eight other scholars contributed most of the essays, and the editors concluded with a call to fill gaps in the historical record, enhance understanding of the map’s emergence as a means of human communication with diverse geometric frameworks, and widen appreciation of the social context, including the map’s use as an intellectual weapon. In their view, a comprehensive reference work that got the facts right could stimulate critical scholarship as well as enlighten curious readers. To this end, Volume One set a standard for the thorough fact checking, careful editing, and accessible typographic design that permeates the series.

As their plans evolved, Harley and Woodward added a fifth volume to encompass the twentieth century. The series swelled to six volumes after they discovered numerous scholars who could collectively cover mapping by indigenous and non-Western societies. Though not part of the original four-volume plan, these topics became the focus of Volume Two, published as three separate books: *Cartography in the Traditional Islamic and South Asian Societies* (1992), *Cartography in the Traditional East and Southeast Asian Societies* (1994), and *Cartography in the Traditional African, American, Arctic, Australian, and Pacific Societies* (1998). Harley died in 1991, but Woodward continued the series, often with the support of associate or assistant editors.

When Volume Three, *Cartography in the European Renaissance*, was published in 2007, its 2334 pages, bound as a set of two large books, dwarfed Volume One’s 654 double-column, 21.59 × 27.94 cm (8.5 × 11 inch) pages, which had established the series’ internal design. Aware of the challenge posed by the explosion of mapping activity after 1650 and sharing the publisher’s concern about production costs – Volume Three overran its original budget of a million words by 30% – Woodward adopted an encyclopedic format for the remaining three volumes. This shift from a small number of relatively long narrative essays authored by comparatively few scholars to a much larger number of encyclopedia entries also acknowledged cartography’s increased complexity, allowed the participation of many more contributors, and afforded a workable strategy for controlling each volume’s length. Consistent with the series’ role as a reference work, the encyclopedic strategy and Woodward’s plan to share editorial responsibilities also promised to reduce production time.

CARTOGRAPHY: HISTORY

2021. Entry-term lists for these encyclopedic volumes reflect the systematic polling of broadly constituted boards of international advisors as well as hierarchically integrated conceptual clusters (HICCs) based on Edney’s modes of mapping practice and institutional endeavors, which were reworked into multiple levels so that important topics would not be overlooked.

Each volume’s HICCs reflect key developments in its respective era. For example, Volume Four details the impact on cartography of the flowering of science, technology, and rational philosophy in the eighteenth century, when mapping became an important instrument of scientific inquiry, especially in the natural sciences, and maps based on observation and measurement had become markedly less decorative than their Renaissance counterparts. In addition, the map was now a powerful device for organizing knowledge in a revealing, readily accessible way as well as a commonplace tool of governmental administration, including planning and maintaining public civil infrastructure. Military cartography faced a similar need for systematic mapping after the open battlefield replaced the prolonged siege of fortified cities as the primary strategy for confronting hostile rivals. No less important to Enlightenment cartography was an increased public consumption of maps fostered by an expanding middle class and related improvements in literacy and education. By 1800 maps were a distinct part of Europe’s visual culture.

The nineteenth century witnessed a strengthening of cartography’s relationships with measurement, mathematics, and statistics, readily apparent in the diverse and complex thematic maps that accompanied advances in the physical, biological, and social sciences and in medicine and public health, all of which used maps for exploratory analysis, hypothesis testing, and serendipitous discovery. Geodetic networks grounded in improved technologies for measuring angles and distances provided a framework for the mapping and demarcation of political boundaries and for the detailed topographic and hydrographic maps important to economic development and military defense. Mapping also supported the growth and densification of railway networks, particularly in Europe and North America, as well as mineral exploration and nascent efforts to inventory natural resources. By mid-century sparse networks of atmospheric observations had provided an empirical basis for understanding the development of storms and demonstrated the practicality of forecasting weather; by century’s end map-intensive operational meteorology had become an essential service of central governments. International cooperation in meteorology, geology, and geography fostered an appreciation of standardized measurements that culminated in the International Meridian Conference, which established the Greenwich meridian as the global standard for longitude in 1884 and provided a stable framework for the worldwide system of time zones, which continued to evolve in the following century.

Mapping’s close relationships to society and technology are particularly apparent in the emergence of a mass cartography promoted by lithography, photography, and industrialized map production, exemplified in turn by inexpensive reference atlases for the home, textbooks with abundant maps, and school atlases. Increased consumption of maps by government, business, and the public accompanied the growth of geographic and general literacy, which also benefited from improved access to public education and municipal libraries. By 1900, a fitful increase of maps in newspapers reflected a growing thirst for spatial information as well as advances in printing and engraving.
Twentieth-century developments

Technology and map use combined with state formation and global forces to make the twentieth century a distinctive and coherent period of map history. Overhead imaging emerged early in the century when the integration of photography and propeller-driven airplanes provided systematic aerial surveys for topographic mapping and related administrative and military endeavors. Satellite surveillance and the electronic capture and processing of overhead data, which followed in the 1960s, were part of a far-reaching electronic transition that included geographic information systems, global positioning systems, enhanced methods for measuring distances and angles and for reconciling measurement errors, and the computational power required for fly-by animation, LiDAR imaging, network adjustment in geodesy, advanced spatial statistics, and the numerical simulation of environmental processes (Foresman 1998).

Technologies that combined transport and mapping fostered new kinds of national territory. Maps were used to partition and control navigable airspace, extend terrestrial boundaries seaward, and divide the seabed on the continental shelf and beyond. The cartographic coastline provided an anchor for exclusive economic zones 200 nautical miles wide, which encroached on the high seas, created heretofore improbable neighbors like Japan and the United States, and precipitated litigation before the World Court when coastal neighbors could not agree on a common boundary. In addition, submarine mapping and space exploration produced remarkably detailed maps of the ocean floor, the Moon, Mars, and asteroids a few kilometers wide, which in turn revealed prominent features like seamounts (sub-marine volcanoes) and impact craters amenable to international naming conventions. As with the electronic gazetteers developed to support topographic mapping, computerized inventories of feature names provided a framework for applying new toponyms and resolving disputes.

Exploration and mapping of outer space and the seafloor continued the longstanding association of cartography and warfare, a synergy further underscored by weapons systems like the cruise missile and the unmanned aerial vehicle, or drone, which made air strikes both more precise and more likely, thereby undermining diplomacy while reducing (but hardly eliminating) civilian casualties. Military spending was a driving force, directly or indirectly, behind many developments in geospatial technology, which trickled down to civilian applications and reproduced a multitude of prohibitive cartographies (Monmonier 2010). Developments rooted in military research include GPS, digital terrain models, and high-interaction display software like Google Earth. These origins spawned debates, still ongoing, over whether geospatial technology can ever be value-neutral (e.g., Harley 1991).

Particularly noteworthy is GPS, originally devised to guide cruise missiles but used widely after 2000 for vehicle navigation and traffic-flow monitoring, land survey and environmental data collection, and tracking sex offenders, criminal defendants, Alzheimer’s patients, children, pets, and packages. An essential component of almost all wireless telephones, GPS supports instantaneous map displays of the user’s immediate surroundings and can compile extensive location histories, invisible to the user but archived indefinitely in the interest of national defense. Although the marked expansion of satellite surveillance after 1990 epitomizes the diverse impacts of mapping on society during the twentieth century, earlier cartographic activities with important influences on individuals and society at large include highway mapping, weather prediction, hypothetical environments proposed by planners and landscape architects, and video
entertainment for both mass audiences and individual users. Video games, which are inherently spatial if not blatantly cartographic, enhanced players’ hand-eye coordination and raised expectations among a new generation of map users.

The twentieth century also witnessed a markedly increased use of maps as tools of public administration, for diverse activities such as land-use zoning, hazard modeling and mitigation, real-property assessment, community policing, and the creation and adjustment of special districts for historic preservation, additional municipal services, and school attendance. Particularly significant is the redistricting software that raised partisan gerrymandering to new levels, first by promoting bizarre congressional districts in the remap following the 1990 Census and then by using shape geometry to craft equally effective but less visually inflammatory gerrymanders that undermined collaborative decision-making by giving one party a distinct advantage in the vast majority of districts. The inherent ambiguity of geospatial technology became apparent when interactive software that encouraged citizens to create alternative redistricting plans reinforced the questionable notion that single-member districts and first-past-the-post voting are essential for representative government (Monmonier 2001).

No less emblematic of twentieth-century cartography is the paradox of globalized practices and customized content. Global standardization reflected increased communication within distinct mapping communities such as national mapping agencies, commercial mapmakers, and academic cartography. Textbook authors codified symbolization strategies, and successful software developers, most notably Esri, introduced new terminology and a standardized aesthetics readily implemented by a swelling cohort of eager users no longer hamstrung by the tedious, error-prone tasks of pen-and-ink cartography and manual digitizing. New mapping communities arose along a spectrum stretching from the necessary uniformity of aeronautical charts and to the purposeful pursuit of novel advertising and tourist maps. In addition map projection software allowed a wider range of custom-tailored geometric frameworks, and a web cartography that empowered map users to be their own mapmakers precipitated widespread customization of geographic scope, map scale, and content. Because millions of these use-specific maps were printed on inexpensive ink-jet and laser printers, the predicted demise of the paper map was delayed if not circumvented.

Map history as an institutional endeavor

As an intellectual enterprise, the history of cartography is particularly indebted to the Russian historian Leo Bagrow (1881–1957), who built upon the facsimile atlases of Santarém, Jomard, and Nordenskiöld by establishing the journal *Imago Mundi* in Berlin in 1935. Its premier issue included several short notices, a bibliography of recent literature, and twelve articles – nine in German and three in English. Its subtitle, printed in both languages (“Jahrbuch der alten Kartographie/Yearbook of Old Cartography”), underscored a focus on pre-nineteenth-century maps. Bagrow had written on map history while working as a hydrographic surveyor in the Russian navy. In 1918, he emigrated to Germany and in 1945 he moved again, to Sweden, where he continued to edit the journal until his death in 1957 (Skelton 1959). Despite multiple publication gaps related to geopolitical turmoil and editorial turnover – no issues were published for 1936, 1938, 1940–1946, 1957–1958, 1961, 1973–1974 – an increase from one to two issues per year in 2004 underscored *Imago Mundi*’s role
J.B. Harley and Michael Blakemore highlighted *Imago Mundi*'s crucial role in a content analysis of its first 30 volumes. Most of the articles published between 1935 and 1978 examined readily available printed maps “published in relatively large numbers, on paper or similar transportable media, yet which were sufficiently old to be of historical importance” (Blakemore and Harley 1980). Less than 5% of the articles examined mapmaking after 1800, and 78% of the mapmakers were Europeans. In concluding that prior scholarship collectively underrepresented important activity after the “Age of Discovery,” they identified three overarching themes: a Darwinian paradigm that emphasized steady, progressive improvement, typically by hypothesizing distinct stages of development; an “old is beautiful” paradigm that underscored an antiquarian fascination with Renaissance maps, their decorative flourishes, and heroic mapmakers; and a nationalist paradigm that reflected convenient access to maps produced in the authors’ homelands. They concluded that the history of cartography was a fragmented discipline lacking coherence and balance.

Matthew Edney (2014), who tabulated both pages and articles, extended the analysis through 2010. In general, empirical studies dominated scholarship on map history throughout *Imago Mundi*'s first 75 years. Edney identified three primary approaches to map history, which he labeled traditional, internal, and sociocultural. Traditional approaches, including biographical and bibliographical essays and content analyses, were dominant until the early 1960s. Internal studies, which focused on map forms and their related technologies and procedures, were then preeminent in the journal until around 1985 when traditional approaches enjoyed a brief resurgence before losing ground to sociocultural studies after 1990. Geographical mapping has been the dominant mode throughout the period, accounting for 65% of all articles followed by marine charting at 15%. Although Renaissance cartography was the dominant chronological focus throughout the period, followed by Enlightenment cartography, “modern topics” (essentially the eighteenth and nineteenth centuries) emerged in the 1980s as a third theme. Overall, the twentieth century has fared no better than non-Western cartography, which accounts for only 10% of *Imago Mundi*'s content.

The sociocultural turn is synonymous with the rise of a “critical history of cartography,” a rubric too readily conflated with Marxist, structuralist, poststructuralist, postmodern, and other avowedly “theoretical” academic ideologies that sometimes read like conspiracy theories. Nonetheless, the sociocultural history of cartography is fundamentally critical in its focus on questioning the map’s creation and use in contexts that are intrinsically social, political, or cultural. This contextualized probing is inherently interdisciplinary as demonstrated by the active participation of art historians, literary theorists, social and political historians, and other humanists and social scientists at academic conferences on map history and in the pages of *Imago Mundi*.

As an intellectual discourse, the history of cartography relies not only on a refereed international journal but also on the synergy of face-to-face discussion at meetings like the International Conference on the History of Cartography (ICHHC), initiated at London in 1964 and held every two years since 1967. *Imago Mundi*, Inc., a registered charity in Britain, which appoints the journal’s editor and editorial board, oversees ICHCC. In 2011 a group of scholars associated with the charity formed the International Society for the History of the Map, which
CARTOGRAPHY: HISTORY

holds its own symposia and whose dues-paying members elect officers and are eligible for a reduced-rate subscription to the journal. The Commission on the History of Cartography, affiliated with the International Cartographic Association, also promotes map history. The Commission holds its own biennial symposia, a year out of sync with the ICHC and focusing on the nineteenth and twentieth centuries.

A third organization, the International Map Collectors Society serves map collectors, map dealers, and cartophiles in general by sponsoring an annual map fair in London and publishing the quarterly IMCoS Journal to which academic map historians often contribute. The symbiotic relationship between cartophiles and historians of cartography accounted for short-lived popular periodicals such as The Map Collector's Circle and The Map Collector, published in Britain 1963–1975 and 1977–1996, respectively, and Mercator's World, published in the United States between 1996 and 2003. The Portolan, published by the Washington Map Society since 1984, is a prominent example of a map history newsletter published by a state or local map society.

The intellectual apparatus of map history also includes specialized bibliographies of maps or mapmakers, such as Robert Karrow’s masterful 876-page Mapmakers of the Sixteenth Century and Their Maps: Bio-bibliographies of the Cartographers of Abraham Ortelius, 1570. Published in 1993 by the Newberry Library, Mapmakers exemplifies map history’s close ties to map collections at national, university, and endowed research libraries like the Newberry. Institutional map collections regularly sponsor themed exhibitions, and bibliographies and cartobibliographies advertise their holdings to historians and other empirical scholars.

Map history’s scholarly infrastructure also includes online map collections and directories. Most institutional map collections have a presence on the Internet, and many websites provide carefully researched descriptions as well as zoomable map viewers, which allow detailed inspection of symbols, labels, and text. Particularly prominent is the David Rumsey Map Collection (www.davidrumsey.com), a trove of high-resolution images scanned from the private holdings of a successful real estate investor. In 2009 Rumsey announced his plan to donate his entire collection of maps and digital images to Stanford University. Other valuable online resources are the Map History Gateway (www.maphistory.info), maintained by Tony Campbell, former map librarian at the British Library, and the International Directory of Researchers in Map History (www.maphistorydirectory.org), hosted by the Osher Map Library at the University of Southern Maine on behalf of Imago Mundi, Inc., which published nine editions of a printed directory between 1975 and 1998.

In sum, both cartography and map history have changed markedly during the twentieth century when overhead imaging and dynamic cartography emerged as distinct modes of mapping practice and both mapmaking and map use became more standardized globally, more customized to the preferences of individuals, and more crucial to science and government. In this vein the societal impact of mapping became more readily apparent not only in enhanced surveillance by the state but also in formal resistance through counter mapping. And by century’s end map history had moved beyond the antiquarian tastes of map collectors to become an interdisciplinary scholarly endeavor with strong ties to geography, librarianship, the humanities, and the history of science. The History of Cartography series exemplifies this intellectual
maturation in its concern for the societal context within which maps were produced and consumed.

SEE ALSO: Cultural turn; Ethics in GIScience; Exploration; Geodesy; GIS: history; Governmentality; Interdisciplinarity and geography; Scale; Surveillance; Technology; Toponymy; Visualization; War

References


Further reading


and Map Use in 1900.” *Historical Geography*, 28: 157–178.


Case study approach

Susan W. Hardwick
University of Oregon, USA

The case study approach has been a popular and accepted research strategy in the social sciences for almost a century. Case study research is particularly useful for geographical research because it is a holistic, nuanced, and integrated approach. The case study approach is defined as the study of a single instance or a small number of instances of a particular phenomenon in order to explore the relationships and contextual influences on that phenomenon (Baxter 2010).

Sample research topics based on a case study approach that have been explored recently by geographers include studies of the incorporation experiences of immigrants in a particular place or context and research on race and whiteness; development; biotic change; participatory planning; and the impacts of climate change on a particular place.

Since focusing on a case study makes it possible for scholars to compile a detailed analysis about a selected place, process, or group, it is useful for doing both theoretical and empirical work. However, during the past three decades or so, and due to the important influences of feminist geographers (and other scholars in the field who are interested in developing or testing theory), research based on a case study approach has been increasingly popular among theoretical and qualitative researchers in the field.

There are three main types of case studies – exploratory, descriptive, and explanatory – although no exclusivity exists between each of these categories (Schell 1992). Some of the best case studies, in fact, are both exploratory and descriptive or descriptive and explanatory (such as William Whyte’s frequently cited 1940s study of street corner society). Exploratory case studies are often a beginning point, or a preliminary analysis, to be used to help develop a plan for conducting a more in-depth research project later on. A descriptive case study, on the other hand, is more often based on a preconceived theoretical and/or empirical framework that may need to be explored or developed in some depth prior to launching a larger project. In comparison, explanatory case studies are most often used for cause-and-effect geographical investigations.

Robert Stake identifies three other kinds of overarching case study categories useful for geographers (1995). These include intrinsic (when the researcher already has an interest in the case before it is studied); instrumental (when the case study is used to study and analyze more than what is obvious to the observer); and collective (when a small group of different case studies are analyzed and compared). Both human geographers and physical geographers have long favored the use of the collective case study approach (often referred to as a multisited analysis), because the study of people and processes in several different places provides the kind of detailed comparative data needed to test theory or for theory building and also for transferability to other places.

It is important to note that literally all types of case studies are multiperspective approaches. In other words, a case study researcher not only considers the voice and perspective of the actors in particular places but also observes and analyzes the context and the relationships and interactions between and among them (Tellis 1997).
CASE STUDY APPROACH

**Origin and evolution of the case study approach**

During the twentieth century, two widely respected academic programs – the Sociology Department at the University of Chicago and the Harvard Business School – developed, refined, and disseminated the use of the case study approach for scholarly work (at Chicago) and for classroom teaching (at Harvard) to scholars and educators in the United States and, ultimately, in many other parts of the world. Many urban and social geographers and scholars in other related fields in the social sciences have long respected and built upon the work of urban sociologists from the Chicago School. The intensive case study-based research they conducted and published on immigrant neighborhoods in Chicago is credited with pioneering the use of the case study approach for urban studies and, later, for a wide variety of other research specializations in human geography and beyond.

The work of Chicago sociologists such as Robert Park, Ernest Burgess, and Louis Wirth popularized the use of the case study approach for urban research, and their widely read publications also encouraged other researchers to emulate their use of a mix of quantitative and qualitative data sources. Most scholars credit the widespread adoption of this Chicago School approach as a positive contribution to the importance of what today is referred to as mixed-methods research. Others, however, are less enthusiastic about this overly positive view of the methodological contributions of urban scholars at Chicago. They argue that the widespread dissemination of their blended methodological approach all too often has resulted in a tendency for scholars to exaggerate the differences between the value of quantitative (“objective”) and qualitative (“subjective”) approaches, and that this division continues to devalue the importance of qualitative work today (see Pratt 1992).

In the 1950s, the follow up contributions of the founding faculty at the Harvard Business School added both impetus and importance to the value, integrity, and validity of the case study approach. During the development of this now prestigious program, early business faculty were surprised to discover a complete lack of appropriate textbooks or other materials needed for classes to be offered as a part of their innovative program. As a result, a core group of faculty thereafter decided to help fill this curricular gap by collecting a group of “real world” case studies of business practices and business practitioners to use as classroom texts. They began this process by conducting personal in-depth interviews in the field and then compiled detailed interview transcripts of their findings.

These interview-based “reports” then were used to develop “cases” that were structured around particular learning objectives and refined in the classroom prior to publication. These case studies were supplemented with a variety of other classroom materials that contained both quantitative and qualitative data. Although the case studies at Harvard were not designed for scholarly work, they nonetheless greatly furthered the importance of the use of a mixed-methods case study approach for guiding student inquiry and also helped validate the use of the case study approach as a legitimate academic tool.

**Rigor, validity, and objectivity of the case study approach**

As with other kinds of research and teaching frameworks, it is vitally important to construct a careful plan to guide and defend all case study-based projects. Otherwise, the internal and external reliability of the outcome of the
study is compromised. One way to enhance the validity of case study findings is to use multiple approaches and multiple data sources. To achieve this goal, geographers frequently depend on “triangulating” case study data sources and research methods by compiling and analyzing data, for example, from such sources as interviews, participant observation, census data, and survey questionnaires. This approach not only helps corroborate the findings of a project, but also adds rigor and validity to the outcome of the study.

A recent example of a geographic study based on a triangulated, multisited, case study approach is a comparative study of the migration flows, incorporation experiences, and spatial patterns of US-born residents of Toronto and Canadian-born residents in Atlanta, Georgia (see Hardwick and Smith 2012). The validity of this comparison of two urban case studies was enhanced by triangulating a suite of methods and data sources including the use of interviews, census data, and other government documents and participant observation. This multisited case study project of cross-border immigrants in two different urban areas (like the countless other research projects based on the case study approach that appear regularly in geography journals, book chapters, and conference papers each year) provides evidence of the ongoing popularity, value, and usefulness of case study-based research in the field in recent years.

**Research design, theory, and methods for case study research**

Case study approaches are useful for geographers who favor using both an inductive and a deductive research design. For clarity, a deductive case study-based research approach carefully develops the research questions and chooses the most relevant theoretical framework to use prior to launching the project’s data collection phase. This approach is quite different from the work of inductive scholars who allow their conclusions and theories to grow out of the discoveries that the data bring to the fore during the project. Geographers favor the use of either or both of these approaches – with positivist, quantitative scholars in the field usually preferring inductive approaches and geographers interested in the use of qualitative methods, grounded theory, and/or other kinds of theoretical frameworks more often favoring research based on deductive approaches.

That said, however, it is important to note that qualitative research in practice is rarely a purely deductive or purely inductive approach. Instead, it often is more of a cycle in the sense that theory stated up front as hypotheses or more loosely simply as possibilities (deductively) to study issues in the “real world” to gather information is then used to generate new concepts (theory) to explain what is observed (inductively) (Baxter 2010). The use of grounded theory is one helpful way to move through this cycle to create and form a detailed and in-depth understanding of the findings from a case study.

Regardless of the often vast chasm dividing these various diverse ways of working and thinking, it is essential for all case study researchers to first develop a well-thought-out research design before launching a project. Likewise, a carefully constructed list of the most appropriate research methods must be developed in advance of launching the project. As mentioned above, one of the many strengths of centering a research project on the use of one or more case studies, especially as compared to other approaches, is that data can (and should) be collected from multiple sources. This triangulation of data-collection strategies provides essential “multiple sources of evidence.” This approach
CASE STUDY APPROACH

not only helps confirm the validity and rigor of research findings, but also can be based on an overlapping set of quantitative and qualitative methods.

Most sources agree on the basic methods/data sources for a case study analysis. These include:

- documents,
- archival records,
- direct observation,
- participant observation,
- physical artifacts.

Data can be collected from one or more of these sources of information during the research process although case study research is likely to be taken more seriously by other scholars if more than one of them is used. Thus, case study scholars should use as many sources as are relevant to the goals of the study. The key is to use one or more of these methods for case study research to collect enough information to (i) document and analyze patterns revealed by the data; and (ii) be able to look for patterns or processes that give meaning to the case study as a whole.

In sum, Yin (1994) suggested three principles to guide the data collection process for case study research.

1. Use multiple sources.
2. Create a case study database.

Strengths and weaknesses of the case study approach

As discussed in this entry, the case study approach has remained an often criticized but viable research strategy since the early twentieth century when the Chicago School conducted case studies of local neighborhoods in the city. Despite this strong foundation, however, in most of the major social science textbooks, the case study approach was not discussed as a central or very important research approach until the mid- to late 1980s when qualitative research began to gain favor.

Because of the lack of support for the case study approach until recent decades, and despite the body of early work by the Chicago School that was based on the case study approach, many scholars continue to view case study-based research as exploratory at best and overly descriptive, narrow, and pedestrian at worst. This negative perception gained momentum among geographers during the discipline’s quantitative revolution in the 1960s when the testing of statistical data and the development of static generalizable research outcomes were favored. As a result, by the early 1970s, quantitative methods and statistical data resulting from survey research and government documents had become the norm with case study research used and discussed much less often. As a whole, the dominant perception at the time was that case studies completely lacked objectivity and rigor as a research approach (Rowley 2002, 16).

However, as discussed in earlier sections of this entry, it is possible for the case study approach to serve as a useful and reliable research method when executed with care. It also has a number of distinctive strengths. First, case studies are the most useful and effective way to gather and analyze detailed information about a particular place, group, or process. Second, case studies are more useful than other approaches when questions prefaced by “how” and “why” are being asked about a contemporary set of events over which the investigator has little or no control (Yin 1994, 9).

Third, another important argument in support of case study research is its usefulness for generating theory or for theory testing. This is one of
the most important strengths of this approach for qualitative scholars as compared to other research approaches because conducting research based on case studies allows theoretical propositions either to be stated before entering the field or, for scholars who prefer using ethnographic approaches or grounded theory, for theory generating later on.

Finally, since all case studies are multiperspectival, the voices of the powerless and often overlooked actors in particular places or groups are heard and considered along with the powerful people involved in the analysis.

Despite these many strengths, it is also important to note that the case study approach does not use sampling as a research strategy. However, with the careful development of a sound research design, strategic planning of the most appropriate research methods to use, and triangulation of data sources and methods, the rigor and validity of a study accomplished using the case study approach can be improved.

In sum, use of the case study approach has been criticized as lightweight, overly descriptive, and insufficiently rigorous since the quantitative revolution in the 1960s. Of greatest concern has been the lack of generalizability of research findings from this approach. However, in recent decades, the increased use of case study research by feminist geographers and other qualitative, theoretical scholars in the field has called attention to the increasing importance, significance, and usefulness of this approach for theory testing, theory building, and overall more nuanced outcomes.

**Conclusions**

As this entry has made clear, the use of a case study approach is an ideal choice when a detailed holistic investigation of a particular place, group, or process is needed. Case studies are especially useful for bringing out the viewpoints and voices of all participants in a study via a triangulation of multiple sources of data. Some case study projects only study one group or one site. Others are multisited analyses that involve analyzing more than one case for comparative purposes and greater transferability or findings.

There are many different kinds of case studies – but more than one category may be used at the same time if they are designed to achieve a common goal. Regardless of which category defines the type or types of case study framing a project, all of the different kinds of case studies are grounded in a multiperspective approach from beginning to end. This means that the researcher in this kind of project is interested in more than just an analysis of the voices of project participants. The interaction between and among actors is also important, as is the context of the place or group being studied. Along with the contextual interpretation of data from this kind of study, perhaps even more importantly, is the consistent way that case study projects call attention to the impacts, perceptions, and voices of the powerless as well as the powerful. Thus, ultimately, despite all the criticisms of the case study approach over the years, this multiperspectival quality ensures that case study analyses continues to be a valuable and timely research tool.

**SEE ALSO:** Archival and document research; Chicago School; Mixed-method approaches; Qualitative data

**References**

CASE STUDY APPROACH


Further reading


Cellular automata

Andrew Crooks
George Mason University, USA

Over the last few decades our understanding and modeling of geographical systems have changed from looking primarily at aggregate patterns to examining individuals, and from assuming equilibrium to looking for far-from equilibrium through the lens of complexity science. One way of exploring how aggregate patterns emerge from the bottom up (i.e., individuals) is through the use of cellular automata (CA) models. A CA model is an arrangement of individual automata, or a software representation of a self-contained entity, that is situated within a regular tessellated space, such as the rectangular grid shown in Figure 1. An individual cell represents a discrete spatial confine of the automata, and an individual automaton is understood as neighboring some other automata. Each automaton processes inputs such as the characteristics or “state” information (e.g., whether a cell is alive or dead) from its neighbors and itself, and then decides whether to change its own state (e.g., from dead or alive), in light of instructions programmed within itself. It is important to note that in traditional CA models, all cell states are transformed by the same set of rules. The neighborhoods may take on a number of different configurations: agglomerations of adjacent cells are defined by their distance from an individual automaton. Two standard neighborhood configurations are illustrated in Figure 1: the nine cell “Moore” neighborhood (which includes the cell in question and its eight surrounding neighbors) and the five cell “von Neumann” neighborhood (which includes the cell in question and its four surrounding neighbors) along with a variation of one of these, the extended $5 \times 5$ Moore neighborhood.

An important feature of CA is that the automaton’s location does not move; the automaton can only change its state. This is different from an agent-based model where mobility is often key (see Crooks and Heppenstall 2012 and Agent-based modeling). The position of the cells and their neighborhood relations remain fixed over time. Information is explicitly exchanged, or “spread” only through neighborhoods (i.e., a local diffusion process). This notion of neighborhoods fits well with one of the tenets of geography, that of “everything is related to everything else, but near things are more related to each other” (Tobler 1970). Additionally, traditional CA models only have one attribute; for example, a cell could be occupied or unoccupied, but the cell cannot contain multiple attributes such as building type or date built. However, over time, this assumption has been relaxed to create more expressive models (as is discussed further later).

To illustrate how CA models work, Conway’s Game of Life is used. This is a simple model that operates over a rectangular lattice with square cells. A cell has only one attribute and one of two states, either dead or alive. The rules of the model are remarkably simple: each cell checks the states of itself and its nine surrounding cells (i.e., Moore neighborhood) and, depending on the states of these cells, the cell will do one of three things. If the current cell state is alive and two cells in its neighborhood have a state that is alive, then the state of the cell remains in the
alive state, otherwise it dies. However, if the cell is currently dead, it will become alive if the state of three neighboring cells is alive. Lastly, the cell will die if it has fewer than two neighbors with a cell state of alive or if more than three neighbors are alive. Figure 2 shows these rules in action.

Specifically, Figure 2 shows how dead cells (white) become alive (black) depending on the number of alive cells around a specific cell in question. Some patterns are stationary or unchanging over time, such as the first four patterns on the left of each image where the same cells stay alive, while others oscillate, for example the three patterns on the right of each time slice. Others translate across the space, such as the “glider” in the bottom of each image, so named because the collection of live cells appears to slowly move across the grid. Such a model demonstrates that from simple, well-defined rules, more complex patterns can emerge and persist through space and time. In essence, each cell processes information from external sources and reacts to these sources based on rules. This notion will be revisited in the applications section of the entry. The key message is that depending on the transition rules and the current state of the cell, different complex patterns can emerge.

CA models have become a standard way of modeling of complex systems because they offer a way to model the idea of emergence, that of complex patterns emerging from simple and well-defined lower level interactions via well-specified rules. The attractiveness of using CA models relates to why there is a need to model more generally. Specifically, the fact that the world is infinitely complex and trying to model all of its behavior is impossible. Modeling allows researchers to unravel this complexity and examine key elements of a system that are of most interest, to test different ideas and theory, and to help understand and predict past and future events. CA models allow us to do just this within the environment of a computer. Once the model has been created, it can be run to see how the system evolves through the interaction of cells.

**Background**

The history of CA dates back to the beginning of computer history when Alan Turing demonstrated the idea that computers could “reproduce” themselves, but it was not until the 1950s and 1960s that the field became alive, so to speak. This earlier work relates to that of John von Neumann and the theory of self-reproducing automata, which described the conceptual principles of a machine that was able to self-replicate (similar to the Game of Life example above). Such work laid the foundation of CA models today. Interested readers wishing to know more about the foundations of CA are referred to Benenson and Torrens (2004) and Cioffi-Revilla (2014).

While the discussion above has highlighted how 2-D CA models work, CA models can also be developed in other dimensions, such as one and three dimensions. Initially, many CA were developed on a 1-D space. For a 1-D model, cells are arranged along a line where the right-hand side is joined to the left-hand side to avoid edge effects. Wolfram (2002) notes that for 1-D CA
models, only 32 different rules can be used with respect to cells changing their state. This is because there are only a finite number of dead and alive states for a cell and its two neighbors. For example, Figure 3 shows an implementation of Wolfram’s “rule 30” (Wilensky, 2002), where a specific pattern emerges from an initial seed cell in the middle of the image.

This rule states that a cell becomes alive (white in Figure 3) or dies (black) depending on the cells to the right and left of it. More specifically, the cell state at \( t+1 \) can be written as:

<table>
<thead>
<tr>
<th>Time (( t ))</th>
<th>Cell state</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t = 0 )</td>
<td>AAA AAD ADA ADD</td>
</tr>
<tr>
<td>( t + 1 )</td>
<td>D D D A</td>
</tr>
<tr>
<td>( t = 0 )</td>
<td>DAA DAD DDA DDD</td>
</tr>
<tr>
<td>( t + 1 )</td>
<td>A A A D</td>
</tr>
</tbody>
</table>

where \( A \) is alive and \( D \) is dead. For example if the current cell is dead (\( D \)), and both its left and right neighbors are alive (\( A \)), the cell dies in the next iteration.

Such simple rules also form fractal patterns, in the sense that the patterns repeat themselves across many scales due to their growth rules reflecting successive applications of local neighborhood principles and, as such, make them attractive to geographers as self-similarity is seen in many geographical settings such as cities (see Batty and Longley 1994, and Fractal analysis). Moreover, 1-D CA models do not just generate interesting patterns but have been used to explore many social science questions, including the emergence of residential segregation or how new political actors and empires emerge over time (Axelrod 1995). However, to apply CA to these systems it must be assumed that each cell is an agent, and that all agents employ the same rules. Such models also assume that agents will not change their behavior and that their behavior is based on what has occurred in the previous iteration, which some authors question as being unrealistic of human systems (Miller and Page 2007).

Moving to a 2-D view could be considered more suitable to geographical questions given their spatiality. Specifically, how objects and spaces relate to one another can be explored, as CA allow for diffusion and interaction (Batty 2005). However, Wolfram (2002) notes that by adding additional dimensions there is no remarkable difference in the occurrence of complex phenomena. Several things change, however, by adding an extra dimension. First is the increased number of possible rule sets and

Figure 2  Example of cells changing state from dead (white) to alive (black) over time depending on the states of their neighboring cells.
the ability to have different neighborhoods (e.g., instead of just right and left in 1-D CA, allows the von Neumann and Moore neighborhoods). While the vast majority of 2-D CA models operate on a square lattice, CA models have been developed on hexagonal and triangular lattices (Cioffi-Revilla 2014). However, such delineation of space has not gained popularity within the geographical community. On the other hand, there are a growing number of models that use a Voronoi tessellation of space (Benenson and Torrens 2004). Such spatial partitions have been developed due to criticisms that geographical phenomena, such as urban growth, do not always fit the regular geometric partition of space of a square lattice. Unlike in the CA models introduced above, the form of the neighborhood and the number of neighbors varies between automata in the CA. For example, neighborhoods using Voronoi tessellations can be based on shared edges as shown in Figure 4: cell 258 has neighbors 261, 264, 266 and 274, while cell 266 only has one neighbor, that of 258.

CA models have also been developed in 3-D and are sometimes referred to as voxel automata. In such models, cells are organized into identical cubes or voxels, and the Moore and von Neumann neighborhoods are extended in the third dimension. The predominant area of application has been in computer graphics, such as the modeling of clouds in flight simulators. However, there are a number of 3-D CA models developed to explore geographical problems, such as the infiltration of water to that of river deposits (Karssenberg and De Jong 2005).

Applications of CA

CA models did not begin to feature much in the geographical literature until the 1990s. That is not to say that CA models were not being developed before that time. Some of the earliest models were developed in the 1960s and 1970s. For example, Chapin and Weiss (1968) developed a cell space model to explore urban growth in Greensboro, North Carolina. While not adopting the term CA, their model used a grid to represent the settlement, and the attractiveness of a cell to become urban was based on adjacency relationships. With a CA-like model that generated a movie of growth in the Detroit area, Tobler (1970) showed how land-use could be defined using a neighborhood and how land-use could change over time. Such models, along with others, showed how one could model socioeconomic processes at a fine scale.
Figure 4  Voronoi tessellations of space where each polygon has a different number of neighbors based on a shared edge.

The rapid growth in applications relates to two broad advances within the sciences more generally. Specifically, the evolution of complexity science and the digital revolution, when computing capacity and graphics became more accessible to researchers combined with the release of remotely sensed imaging (Batty 2005). The CA model regular lattice structure is analogous to the raster data structure from many remotely sensed images and allows a one-to-one mapping between cells within the CA model and remote sensing. An example is the DINAMICA model (Soares-Filho, Cerqueira, and Pennachin 2002), which was designed to simulate landscape dynamics in the Amazonian tropical forests: its cell transition rules are based on spatial data pertaining to soil types, vegetation, altitude, slope, distance to rivers or distance to roads; the
CELLULAR AUTOMATA

model uses this information to explore land-use and land-cover change.

Many consider the incorporation of spatial data into CA models of geographical systems important because this integration provided the ability to have cells that are related to actual geographic locations. This is of crucial importance with regard to modeling geographical systems, as everything within a city, region or country is connected to a place. Furthermore, it allowed modelers to think about how objects (i.e., cells) and their aggregations interact and change in space and time. For geographical information systems (GIS) users, it provided the ability to model the emergence of phenomena through individual interactions of features within a GIS over time and space. Moreover, some would consider this linkage highly appealing in the sense that while GIS provides us with the ability to monitor the world (i.e., land cover), it provides no mechanism to discover new decision-making frameworks, such as those that can be explored through a CA model by applying rules to cell states. Moreover, the use of spatial data within CA models also allows researchers to validate and calibrate their models. Several validation and calibration techniques exist; readers are referred to Pontius et al. (2008) for a detailed discussion of these techniques, as it is beyond the scope of this entry (see also Validity and verification).

The flexibility offered by CA has resulted in this technique being used across the spectrum of geography: from that of exploring the physical world to that of exploring the human world and anywhere in between. For example, representative applications of where CA models have been developed to explore social science issues include the emergence of political actors, segregation, social interaction, and traffic dynamics. In the physical arena, CA models have been developed to explore geomorphological processes, groundwater flow and the contamination of soils, earthquakes to that of the spread of diseases, and the spread of fires. Perhaps the largest area of applications within the geographical community is that of land-use change. CA models have been developed to explore deforestation due to pressure from farming to that of exploring the re-growth of forests (Soares-Filho, Cerqueira, and Pennachin 2002) to that of urban growth. The popularity of CA in this area is due to the fact that to truly understand and explore land-use change it is necessary to capture both temporal and spatial dynamics of such change, which is a movement away from past generations of models, which focused on equilibrium outcomes rather than on their dynamics and which assumed that land-use change processes were well behaved (Benenson and Torrens 2004).

With respect to geographical modeling, CA models represent a significant improvement on previous generations of simulation models: they are spatially implicit while contemporary models generally conceive of space in relative terms (e.g., spatial interaction and gravity models, econometric models, and location-allocation models). CA additionally treat individual entities or objects at a microspatial level and as temporally dynamic, embodying processes that do not imply any equilibrium end state, and reflect simulations based on local rules which generate development, one cell at a time, over many time periods, all of which capture notions from complexity science (Benenson and Torrens 2004). With respect to defining transition rules, multiple statistical and nonstatistical methods have been developed. For example, Clarke and Gaydos (1998) used a Monte Carlo method, while logistic regressions are favored by others such as Verburg et al. (2002), and nonstatistical methods for defining transition rules include artificial neural networks or particle swarm optimization (readers are referred to Liu and Feng (2012) for further discussion).
However, as with all modeling approaches, CA models have their weaknesses. One is their reliance on neighborhoods for the propagation of information, which inhibits action-at-a-distance from taking place and makes them unable to handle explicit spatial interaction between cells. Additionally, real-world neighborhoods come in many different shapes and sizes that may not map well onto the CA tessellation. Also, when modeling physical applications, care must be taken that the basic object being modeled is consistent with the physical processes involved. Another weakness of CA models, specifically in relation to geographical systems, is the inability of automata cells to move within the lattice in which they reside. It has been argued, for example that one can only go so far with representing urban development as fixed locations (Batty 2005). There is a need for mobile objects that can move between cells, for example, when dealing with mobile objects such as pedestrians, migrating households, or relocating firms. However, this limitation has been overcome using the cells just as locations in which the agents can move around and, more importantly, can communicate with other agents. For example, the FACS (free agents in a cellular space) models by Portugali (2000) show how important it is to consider mobility within the process of segregation.

As a result of these weaknesses, CA models applied to geographical systems rarely invoke the strict limits of CA outlined above. Researchers have adapted the formalism of CA to meet their simulation needs. For example, Clark and Gaydos (1998) relaxed the cell neighborhood assumption entirely, in that cells are chosen for development across the entire space without any constraints imposed by local neighbors. O’Sullivan and Torrens (2000) discuss adding several cell states to account for continuous change in land-use and demographic makeup, instead of purely discrete states (e.g., dead or alive). Another weakness of CA models is their homogenous state transition rules, which are applied to all cells in a way that limits modeling processes that vary over space and by individual, such as human behavior. While this is not an issue for many physical systems, it is an issue for systems where humans are involved.

Future research directions

In response to the various challenges faced in using CA, one area that is currently being investigated is the linkage of CA with agent-based models (ABM) (Heppenstall et al. 2012). In practice, CA and ABM have often been applied separately to explore a wide variety of urban phenomena, but they are increasingly joined. For example, CA models are commonly applied to represent possible land-used changes, while ABM are often applied to crowd dynamics and traffic simulation. However, models are increasingly being developed using a combination of CA and ABM techniques to overcome some of the weaknesses identified above and the distinction between them is increasingly blurred. For example, cells may be granted some degree of agency in their state descriptions and are simply reinterpreted as artificial agents and/or agents are imposed on top of a CA layer, while simulated agents are interpreted as responding to average cell conditions and are able to move location. This combination of CA and ABM results in the continuum between the two approaches narrowing and, as such, these models are better described as agent-based models operating in cellular space (CS) models rather than pure CA models (Batty 2005). By combining the two approaches, the entire range of action and interaction effects relevant to geographical systems can be accessed, for example, action-at-a-distance. Additionally, by granting
CA models some degree of behavior, decisions can be taken account of; for example, in a CA model, it might be assumed that a land cell will be converted from rural to urban if its surrounding cells are all urban. However, ABM allows the behavior of individual cells to be considered. For example, if the land cell owner does not desire to sell the land cell to a developer, the land will not be developed. This individual behavior is often missing in CA models of urban growth.

One approach that uses this combination is that of “geosimulation” (Benenson and Torrens 2004), which stresses the importance of studying space when seeking to understand our environment. These involve problems where nonfixed objects (e.g., people or cars) are located on fixed objects (features such as houses or roads) allowing for the simultaneous processing of immobile and mobile objects, at the same time directly accounting for spatial relationships between various types of urban and nonurban objects, revealing the emergence of spatial patterns at higher levels of urban hierarchy.

Summary

This entry has introduced CA models, presented its history with respect to geographical systems, discussed its application domains, and highlighted future directions of research. Readers wishing to explore the topic further are referred to Benenson and Torrens (2004), Batty (2005), and Wolfram (2002), whose books give further reviews of CA principles, models and their applications.

SEE ALSO: Agent-based modeling; Complexity in biogeography; Fractal analysis; Land-use/cover change and climate; Location-allocation models; Representation: dynamic complex systems; Validity and verification

References


Census geography

Wardlow Friesen
University of Auckland, New Zealand

The term “census geography” is used to define broad issues of the implementation, analysis, and application of a census and related issues of its political, social, and economic implications. Comprehensive censuses are largely a phenomenon of the eighteenth century onward, but the history of census-taking goes much further back, probably to as long ago as 6000 years. The earliest censuses were taken for the purposes of military conscription, labor recruitment, and taxation, and were carried out in various early civilizations including China, Egypt, Babylonia, and Rome. A more recent census was the Domesday Book of 1086 in England and Wales, which involved the enumeration of each landowner according to the value of the land and livestock they held, for taxation purposes.

The “modern census” is said to have started with the 1749 census of Sweden, although a few earlier censuses had some modern attributes. The ideal modern census is complete, accurate, and simultaneous. Completeness involves an enumeration of the whole population. It also implies the collection of information on a range of variables for individuals and usually households. By the end of the nineteenth century most of the countries of Europe, as well as the countries of significant European settlement, had implemented modern censuses. Through the nineteenth and twentieth centuries some colonial powers initiated censuses in their colonies, and by the end of the twentieth century most nation-states had a census. International organizations are also involved in census-taking, with the United Nations recommending a standardized range of core variables; in many countries significant logistical input is provided by the United Nations Fund for Population Activities (UNFPA). In recent years, a number of countries, especially in Europe, have stopped taking full enumeration censuses in favor of other types of data collection, including censuses based on a sample of the population and population registration systems.

The functions of a modern census are rarely related to the earlier purposes of conscription and taxation, but rather are undertaken to promote the wellbeing of national populations. In some democracies based on universal suffrage and “one person one vote,” including the United States, Canada, and New Zealand, the census has a constitutional role in facilitating the drawing of equitable electoral boundaries. In virtually all countries, census functions include the provision of information for an understanding of the social, cultural, and economic characteristics of the population, as well as to facilitate the planning of systems of health, education, labor force, transport, and other services.

With the focus of contemporary censuses on the wellbeing of the population, demographic and socioeconomic analyses are central. Much demographic analysis of fertility, mortality, and migration is dependent on census data. In countries that have comprehensive birth and death registration systems, vital rates are partly based on compulsory registrations, but are still dependent on census data to establish base populations for the calculation of rates. In many less developed countries such registration systems
are absent or inaccurate, so determining rates of fertility and mortality are largely dependent on census analysis. A census also provides data on many contingent variables, which can be analyzed in conjunction with data on vital rates. The analysis of international migration, for example, is dependent on census data as well as other data such as arrival and departure data and immigration visa information. Besides detailed empirical studies, censuses tend to be the main source of data on migration outcomes, through the analysis of birthplace and ethnicity data in conjunction with other variables related to economic activity, occupation, income, housing tenure, and many other variables.

The determination of patterns and characteristics of internal migration is largely dependent on census data, except in those countries that have comprehensive population movement registers. Census questions about residence at a previous point in time provide at least a snapshot of internal movement. If a census does not ask a previous residence question, the residual method of migration analysis can be used, whereby changes in the age–sex structure of an area (combined with birth data and life tables) between censuses can be used to infer the age–sex characteristics of net migration.

Demographic information produced in censuses is particularly important for planning through population projections at national, regional, and local levels. These usually involve the incorporation of estimated undercounted populations and the use of assumptions based on past trends and forecasts of economic and other conditions in the future.

Population geographers and others are involved in a great range of socioeconomic analyses based on census data on educational achievement, occupation, income, housing tenure, ethnicity, gender, and migrant status, among other variables. Methods of analysis include descriptive cross-tabulation, correlation analysis, geospatial analysis, and the creation of multivariate indicators of wellbeing or deprivation.

Resources and services provided by governments are often allocated on the basis of population distribution and characteristics determined from a census. In some cases, like the United States, allocation has a legislative basis, while in others, this mechanism is less formal. Thus, if a population undercount occurs, resources may not be equitably distributed. Although most censuses aspire to 100% coverage of the population, there is no evidence that any census has ever achieved this goal, despite penalties for nonparticipation. Impediments to full enumeration include the suspicion of some, including undocumented migrants, that census data may be passed on to other government agencies; the rise of libertarian or anarchist movements averse to government in general; the difficulty of locating highly mobile or homeless people; and the reluctance of people to fill out a sometimes lengthy form.

Post-enumeration surveys in countries with effective census systems have typically shown undercounts in the range of 2–4% of the population; however, they also show that the undercount is age, gender, and ethnically selective, so undercounting among young males in minority ethnic populations may be three or four times that of other groups. In general, minorities in large cities are undercounted, as are indigenous groups. In the United States this has been a political issue after every census, since funding for education, health, and other services is dependent on census counts. In addition, the electoral representation of minorities is negatively impacted in many countries.

The late twentieth and early twenty-first centuries have seen the convergence of two arguably contradictory directions in census geography: the move away from comprehensive traditional
full enumeration censuses by some governments despite the development of new technologies that have much greater potential for improved and cost-effective census-taking and analysis.

In the latter part of the twentieth century, new technologies have allowed much more comprehensive processing and analysis of census data than had been possible earlier. Computerization has permitted a much greater range of variables to be collected and analyzed, and geographic information systems (GIS) have aided enumeration procedures, as well as spatial analysis and display. Most censuses through the twentieth century involved door-to-door enumeration, often involving at least two visits by enumerators to each household to ensure compliance, but in some countries e-censuses involving online enumeration are increasingly being adopted, usually resulting in significantly reduced costs.

Despite these new potentials, some authors have claimed that it is the twilight of the traditional census because of the high cost (especially of enumeration), privacy issues, declining response rates, and the need to obtain more timely and regular data, especially when some censuses are conducted only every 10 years (Coleman 2012). The primary alternative has been the analysis of databases based on population registration, sometimes combined with other administrative data. These have been in use for several decades in the countries of Scandinavia, which have adapted existing registration systems, and in Japan. These countries also have populations that are accustomed to registration and which largely accept this level of monitoring. Most of the other countries that have moved away from the traditional census are in Europe, with only a few elsewhere.

Kukutai, Thompson, and McMillan (2015) counter the twilight argument by pointing out that only 12 out of 241 countries (including territories) did not have a census of any sort in the 2010 round of censuses (this number includes censuses based on registration). Further, the great majority of the world’s countries have taken a full enumeration census in recent years, including some that had never taken a census before. The reasons for this include greater international support for census-taking and analysis, the reality that most countries do not have comprehensive systems of registration or extensive administrative data, a sense that the use of such systems would be even more intrusive than a confidential census, and the realization by politicians and planners that the collection of a comprehensive range of census variables is critical to future planning. Another complicating factor is that some variables are not collectable or are not usually collected in registration systems. These include ethnic self-identification, languages spoken, household and family composition, and variables relating to services such as mode of transport to work or school.

Even if the traditional census does not disappear, there is the possibility that its extent and quality may be compromised in some cases, with Canada being an example. At short notice, and apparently for political reasons, the compulsory long form of the 2011 Canadian census was abandoned in favor of a short form with a limited number of questions, supplemented by a voluntary National Household Survey (NHS). The response rate to the NHS was only 69%, causing academics, planners, and others dependent on census data to suggest that the resulting data are of limited use because of the skewed nature of the sample (Walton-Roberts et al. 2014).

These changes in census methodologies and their resulting data outcomes are a challenge not only for population geographers and others in terms of data analysis, but also for many in relation to their role as advocates for the ongoing collection of census and other data to allow for
CENSUS GEOGRAPHY

robust academic research and effective planning decisions.

SEE ALSO: Fertility; Migration: internal; Migration: international; Mortality; Population geography; Quantitative methodologies

References


Central business district

Peter O. Muller  
*University of Miami, USA*

The downtown central business district (CBD) of a city consists of its core concentration of largely nonresidential activities that is endowed with superior accessibility to the surrounding urban area it serves. Prior to the last third of the twentieth century, that reach was dominant throughout the metropolis. But over the past half-century, that tributary area has contracted to mainly encompass the central city as spatial restructuring has turned the American city (especially) inside out, forging a new polycentric urban form anchored to a suburban constellation of new metropolitan-level centers as well as the downsized CBD.

The origins of the central business district can be traced back at least as far as the agora that formed the nucleus of the ancient Greek city, which in turn evolved from the central activity cluster that probably marked urban settlements since they first appeared more than 5000 years ago. Nonetheless, formal CBDs did not emerge until the mid-nineteenth century, when innovations from the Industrial Revolution triggered the modern age of urbanization by unleashing a massive migration from rural areas into the newly forming manufacturing cities. Swiftly, the loosely organized networks of small preindustrial cities mutated into a national-scale urban system, connected by an expanding web of increasingly efficient railroads that linked raw material sources, producers, and markets. Internally, the transformation of the larger nodes of the new urban system was just as dramatic, particularly the spatial reorganization of its economic activities.

A key component of that restructuring was the expansion and differentiation of urban employment, enhancing the role of centrality in the nascent industrial city that facilitated the rapid emergence of the CBD after 1850 (Ward 1971: 85–102). The transportation-cost savings of a central location were most advantageous not to the operators of factories, but to the producers of commercial and service activities that accompanied the rise of the industrial urban economy. The latter outbid all other land uses in the competition to locate at the most accessible node where the city’s trolley (and, later, subway and commuter rail) lines converged, sorting themselves into increasingly specialized retail, warehouse, and financial districts. Administrative activities also grew alongside them, and by 1900 had evolved into modern office complexes as improved elevator technology enabled the proliferation of skyscrapers that added significantly to the available floor space for business when the CBD’s density approached its maximum.

During the first half of the twentieth century, CBDs dominated the metropolitan areas they centered. (The term “central business district” was coined in the 1920s by core-city civic leaders to assert the primacy of downtown as new highways began to spawn commercial centers in the suburbs.) But, by 1960, the steadily strengthening forces of decentralization, propelled by the automobile and new superhighway networks, were shaping the rise of an outer suburban city that surrounded an aging central city still dependent on its declining public transit systems. In short order, the CBD’s exclusive hold on region-wide centrality disappeared as accessibility became a
ubiquitous good – now almost any location on the metropolitan freeway network could easily access the entire urban area. By the 1970s, the metropolis was both turning inside out and spatially reorganizing as new metropolitan-level centers – edge cities – mushroomed in the suburban ring at major freeway intersections, and the monocentric urban form of the past gave way to a new polycentric metropolitan structure that has been maturing ever since.

The consequences of these changes for the CBD over the last half-century have been substantial. From the late 1950s into the 1980s, downtowns stagnated as the central cities entered a new (and still continuing) era of economic and population decline commensurate with their reduced role in the restructuring metropolis. But the core city’s civic leaders did not give up on the CBD, and by the 1990s a turnaround was underway. An important driver was the “back to the city” movement of young professionals, which emphasized downtown-area residential opportunities, especially the gentrification of CBD-bordering neighborhoods. At the same time, the livability of the city center was being enhanced in numerous ways, both to entertain its new residents and to attract visitors from the rest of the metropolis and beyond. The layout of the CBD now began to change as well, with the long-standing office–financial–hotel–retail complex (increasingly marked by reconstructed buildings) forming the core, surrounded by a revitalizing frame containing myriad parking facilities, new entertainment and convention districts, and expanding medical and government clusters (Hartshorn 1992: 337).

By the 2010s, this core-frame pattern had become established in several major central cities (leading the real estate and revitalization booms at mid-decade were the CBDs of Denver, Seattle, and Chicago). Yet still unclear was whether these successes amounted to a true renaissance of the city center or merely a temporary reversal in the long-term economic and demographic decline of the United States’ old metropolitan cores. Taking no chances, many CBDs vigorously continue their quest to reinvent themselves – in fact, these days the very term “central business district” has mostly been replaced by “downtown.” This movement has fascinated researchers across the broad spectrum of urban studies, and a sizable body of scholarly literature has accumulated since 2000 to document the rebirth of downtowns. Apart from significant work on such specific topics as gentrification and homelessness, however, the spatial generalization and theorization of downtown redevelopment remains a research frontier in its formative stage.

The most widely cited and comprehensive treatment of the subject continues to be Larry Ford’s (2003) America’s New Downtowns: Revitalization or Reinvention?, a classic study of landscape change, interpretation, and conceptualization based on meticulous fieldwork. Concluding that the old CBD has evolved into something more substantial than a “postmodern downtown of consumption and display” (the superficial label used by many researchers), it provides the foundation for a new urban geography of the city center, the cornerstone of which is Ford’s (2003, 308–310) composite model of downtown spatial organization. Incorporating the key features that mark post-1990 downtown expansion (new attractions to draw in people; restoration of local historic districts; residential growth; and transport innovations), the model consists of seven major elements: (i) the spatial form of downtown, especially its degree of linearity; (ii) theme districts (new entertainment facilities, cultural centers, sports venues); (iii) enhancement of physical amenities; (iv) new commercial-building “footprints” as walls and gates; (v) competition between private and public spaces; (vi) new residential and recreational
spaces to fill vacated commercial properties; and (vii) making downtown a family-oriented event center.

All of these elements continue to be vital for monitoring and understanding the processes that are transforming the heart of the old metropolitan core. During the decade since the composite model was first proposed, it became apparent that the forces of globalization constitute yet another element in the reshaping of downtown. An overriding impact of globalization on the metropolis has been the increasing fragmentation of people and activities within central cities and suburban rings. Thus transnational migration intensifies the diversity (and internal spatial differentiation) of the metropolitan social mosaic, while economic activities increasingly attuned to global networks localize in state-of-the-art business centers – especially in and around suburban edge cities. As this process unfolds, the implications for big-city CBDs are enormous. To date, most have demonstrated not only resiliency but also a commitment to compete for their share of the action, nurturing continuity when possible but always open to new opportunities to reinvent themselves.

SEE ALSO: Accessibility, in transportation planning; Edge city; Gentrification; Homelessness; Industrial agglomeration; Infrastructure; Intrametropolitan location; Polycentricity; Suburbanization; Urban geography; Urban redevelopment; Urban transit

References


Further reading


Central place theory

Peter W. Daniels  
*University of Birmingham, UK*

One of the notable features of patterns of settlement or (for example) shopping districts within cities is a sense of regularity in their spacing and size. Smaller groups of shops or freestanding settlements are more numerous than larger shopping centers or settlements. Moreover, the larger entities are generally spaced farther apart than the smaller ones. During the early twentieth century such attributes in the spatial organization of places did not go unnoticed by a German geographer who was intrigued by settlement patterns he had observed in southern Germany. Walter Christaller (1966/1933) developed a theory concerning the size and distribution of what he termed “central places” (settlements). Central place theory seeks to account for how different types of settlements (differentiated by size and function) are located in geographical space.

To appreciate the value of central place theory, it is useful to consider the functions performed by settlements that can range in size from a very small group of homes with few people to many thousands of homes and millions of people alongside, for example, numerous business, administrative, and cultural activities and buildings. Settlements exist not only to support people’s needs for goods and services but also to provide goods and services for those nearby; each settlement has a market area and is centrally located within it; the reach of the settlement is determined by the number and different kinds of goods and services that it provides. Settlements supplying the most goods and services are described as higher-order centers, while those supplying the least are the lowest-order centers. It follows that, because the lowest-order centers possess a very limited range of functions, they will tend to be used by those who are nearest. Some of the functions located in the larger places will be more specialized and used less frequently, but these will command a much more extensive geographical market.

The purpose of central place theory is to provide a universal set of rules for explaining the location, hierarchy, role, and function of settlements in geographical space (Hsu 2012). It can be used to understand historical and contemporary patterns of settlement distribution. The starting point is a set of simplifying assumptions that can be used whenever and wherever the theory is put to the test. For the purpose of his theory, Christaller assumed that the pattern of settlements develops on a flat, limitless, and homogeneous surface (an isotropic plain); that populations are uniformly distributed across this surface; that resources are evenly distributed; that consumers of goods and services have the same income levels and purchasing power, and behave in the same way as consumers; that perfect competition exists between sellers who all seek to maximize their profits; that only one mode of transport exists which is equally easy to use in all directions; that the cost of transport uniformly increases with distance covered; and that the pull power of any one supplier of a good or service steadily declines as distance to a user increases.

Using these simplifying assumptions, Christaller introduced two concepts to translate their effects into settlement arranging. First, the concept of
Central Place Theory

Range is the maximum distance users of goods and services are prepared to travel to reach them. This distance varies according to the type and frequency of use of different types of goods and services but a point is reached at which the cost outweighs the need or benefit. The second concept, threshold, is the minimum number of users (or market) for a good or service that is required to make it viable. With the range and threshold varying according to the type, size, and specialization of a good or service, it becomes apparent that every good and service cannot exist everywhere across the isotropic plain. Consequently, central place theory posits that a hierarchical system of centers will emerge, with those at the base of the hierarchy closest together and offering a narrow range of frequently used functions (such as newsagents, village stores, and post offices) and centers at the top of the hierarchy providing a larger range of goods and services (as well as all the frequently used functions) that are reached from longer distances and less frequently, and are therefore dependent on a larger market (population). Examples include fashion stores, jewelers, department stores, photographic equipment suppliers, and car dealerships.

It remains to consider how these assumptions and key concepts shape the distribution of settlements. Christaller started with the assumption that each settlement offering the same order of goods and services will have the same circular market area delimited by the range and threshold. From an examination of the overlap of market areas for settlements at different levels in the hierarchy, Christaller concluded that the most efficient pattern for serving markets with minimum overlap would take the form of a triangular/hexagonal lattice. Three different settlement layouts were identified: the marketing ($k = 3$) principle, the transport/traffic ($k = 4$) principle, and the administrative ($k = 7$) principle (Figure 1). Take, for example, the $k = 4$ principle which involves the minimization of the length of roads connecting central places at levels of the hierarchy of settlements. The highest order center (1) has a market area that includes half of the market area of each of the six neighboring lower-order centers ($0.5 \times 6 = 3$). These lower-order centers are located on the sides of the hexagons around the higher-order settlements and this creates the most efficient transport network.

There are other ways in which the principles of central place theory can be used to account for the pattern of settlements. Rather than working downward from the highest-order central places, August Lösch (1954) started with a system of self-sufficient farms (lowest order) that were distributed regularly in a triangular-hexagonal pattern. From this smallest scale of economic activity, he worked upward, using mathematics to derive a number of central place systems that also included the three systems identified by Christaller. This analysis addressed Lösch’s concern that the three options derived by Christaller paid too much attention to location and not enough to how the pattern of centers would maximize the benefits to consumers rather than to the distributors of goods and services. To Lösch, his method created a distribution of settlements as an ideal consumer landscape where the requirement to travel for any good or service was minimized and profits were held level.

![Figure 1](image.png)
The assumptions used in central place theory for explaining patterns of settlement are actually difficult to replicate in the real world. Few landscapes are uniformly flat; physical features such as lakes, estuaries, and mountains distort transport routes and costs. Also, a population’s socioeconomic attributes are unlikely to be homogeneous; some households will have one or more cars, others will only have access to public transport; some households will have higher incomes than others. Finally, households can display behavior that does not necessarily conform to the principle that they will travel to the nearest appropriate center for a good or service. This tendency will also be influenced by real or perceived differences between centers in the quality, cost, or even experience of shopping. The role of location and cost of travel is therefore compromised by numerous other factors. Set against these limitations, the increasingly ubiquitous use of information and communications technology for online retailing, banking, or arranging home service deliveries (ranging from groceries to vehicle tire fitting) has dramatically changed the extent to which the assumptions of central place theory can be fulfilled.

This is not to suggest that central place theory is not useful. This was recognized very early on by the Nazis; they hired Christaller to develop a plan for repopulating the depopulated areas of Germany and Poland using the principles of central place theory (Barnes and Minca 2013). Many geographers, especially during the period 1950–1980, set out to validate the theory by examining real-world patterns of settlements (Berry 1967) or the distribution of retail centers within cities (Johnston 1966). Its principles have also been widely utilized for planning the location of public services such as hospitals or commercial shopping centers. The relative absence of central place theory in recent research does not reflect its demise; rather, it is probably the case that, since the 1990s onward, its tenets have been implicitly, rather than explicitly, utilized in activities such as planning for the location of new shopping centers, office parks, new towns, or new transport infrastructure. Some have recently suggested that it is time for the re-emergence of central place theory (Mulligan, Partridge, and Carruthers 2012).

**SEE ALSO:** Location-based services; Spatiality; Urban geography

**References**


Change detection, in the geospatial context, is broadly defined as the identification and characterization of the effects of natural and anthropogenic processes in a landscape. Examples of landscape change include conversion of agricultural land to residential areas, and shifting forest species composition due to climate change. Change detection is important in a wide range of applications, such as urban planning, wildlife habitat monitoring, and carbon flux estimation. It is difficult to imagine a natural resource or environment-focused application that does not depend, in some way, on knowledge of the changing composition of Earth’s landscapes.

Change detection can be accomplished with many types of geospatial data, including field measurements, land-cover/use maps, and various layers such as road networks, cadastral data, and elevation models. However, the most commonly used data type in change detection is imagery acquired from aerial, satellite, and in situ sensors. Advantages of imagery for change detection include the many sensor options available in: (i) spatial resolution (i.e., pixel size), from submeter to multiple kilometers, allowing for studies at different spatial scales of interest; (ii) temporal resolution, allowing for flexible revisit times based on the needs of the application; (iii) spectral resolution, with existing and planned sensors offering sensor bands in optical wavelengths suitable for varied terrestrial and marine applications; and (iv) radiometric resolution, with newer sensors such as Landsat 8 able to distinguish fine gradations in reflectance.

Landscape change results from a variety of activities and processes. A distinction is often made between changes in land cover, that is, the landscape element that is apparent to the viewer or sensor (e.g., forest or water) and land use, or human activities that affect the appearance of a landscape (e.g., agriculture or silviculture). However, change can occur in many forms. The causes of some changes may be readily apparent, such as the construction of a lumber mill in a formerly forested area. Other changes may have multiple or uncertain causes, such as a forest fire: Was the fire the result of a lightning strike or a poorly tended campfire? Since it may be difficult to determine whether a change is natural or anthropogenic, the characteristics of the change, as measured using the input data, are of primary importance.

An important determination affecting the accuracy of both the detection and the characterization of landscape changes, regardless of their causes, is whether they are discrete or continuous. A discrete change is one that results in a new land-cover label for an area. For example, in the construction of a housing development in a former crop field, the old label of “agriculture” or “row crop” could be replaced with “low density urban” or “house,” depending on the scale and thematic detail of the resulting map. In contrast, a continuous change is
one in which the thematic label assigned to an area is constant, but change nonetheless occurs. For example, as a forest ages the crowns of its trees may grow together, resulting in an increase in crown closure and a decrease in visibility of the understory or ground vegetation. The label “forest” would remain, but maps showing crown cover percentage as a thematic variable would require updating. Complicating factors in discriminating discrete from continuous change include the lack of temporally sufficient data to identify a continuous change, and, given enough time, a continuous change may cause a discrete change. For example, vegetation in a drying wetland may gradually be replaced by upland species, resulting in a discrete land-cover label change from “wetland” to “upland.” Annual imagery may be suitable for identifying and characterizing such a continuous change, but images acquired several years apart may be useful only for mapping the discrete change.

In a properly designed change detection study, the change identification and characterization goals are specified such that available data and methods used can support those goals. In a research environment often characterized by limited funding and short duration projects, one rarely has the ideal dataset to conduct a given change detection study. Thus, comprehensive planning must be undertaken to ensure that maximum utility is extracted from the resources that are available. Land-cover label definitions must be clearly specified so that ambiguity about what constitutes a change is minimized. The types of changes to be mapped, for example discrete versus continuous, must be described. The change-mapping goals must be appropriate for the quality and spatial, temporal, spectral, and radiometric resolutions of the input data. Finally, the methods used must be appropriate for the changes to be mapped.

Change detection and visualization methods

Many change detection techniques have been developed and studied since the mid-1970s (Singh 1989; Coppin et al., 2004; Lu et al., 2004; Jensen 2004). These techniques have been generally divided into two groups: preclassification change detection and postclassification change detection.

Within the preclassification change detection (pre-CCD) group, techniques are generally pixel-based, that is, operating on one image pixel at a time. These techniques analyze multitemporal images on a pixel-by-pixel basis and do not take into consideration spatial context of the surrounding pixel. Examples of pre-CCD pixel-based techniques include: (i) multidate composite image, (ii) image algebra, and (iii) transformation band.

Multidate composite image change detection combines images from two or more dates with the purpose of creating one multilayer, multidate image, and extracts any change information from the composite image. A common example of detecting changes with this technique is through the use of variation in pixel values for the different dates to transform the input image into change and no-change pixel groups.

Image algebra techniques use arithmetic operations on the spectral values, along with a threshold value to determine changed areas. Common examples of techniques within this category include: image differencing, image regression analysis, image ratioing, vegetation index differencing (e.g., Normalized Difference Vegetation Index, a measure of vegetative land cover), and change vector analysis. The majority of these techniques are easy to perform and understand. The most common pre-CCD technique used within this category is image differencing in which one date of imagery is
subtracted from another to produce a “difference image” that highlights pixels for which the spectral values have changed. The main difficulties with pre-CCD are that these techniques do not provide from–to class label information and they require the selection of a suitable pixel value threshold to detect changed areas. For example, in the image-ratioing change detection technique the ratio of the pixel values is first computed and then a threshold is applied to the ratio to discriminate change from no change.

Transformation band change detection refers to the use or manipulation of color or multispectral band transformations. The following are examples of techniques used within this category: principal component analysis (PCA), Kauth–Thomas transformation, Gramm–Schmidt (GS) process, color invariants, and chi-square transformations. The most commonly used techniques in this category are PCA and K–T because of their relative simplicity and wide availability in commercial remote sensing software. The main advantages of this pre-CCD category are the reduction of data redundancy between image spectral bands and the maximization of image information shown in derived components. These techniques suffer from limitations similar to those of image algebra methods.

Pre-CCD techniques are particularly useful for identifying changes rapidly and for determining the magnitude of changes through comparison of image pixel values over multiple dates. The main challenges with pre-CCD techniques are that the change detection results do not describe changes in a particular class (e.g., urban or forest) and the computed pixel differences can be affected by a range of factors including sensor calibration, atmospheric conditions, illumination and view angle, and groundwater and soil conditions.

Postclassification change detection (post-CCD) techniques can be grouped into two categories: pixel-based and object-based image analysis (OBIA). In general, these techniques use thematic classifications of the input images produced using a classification technique (e.g., supervised or unsupervised classification). Once the images are classified, the class labels for each pixel or object are compared to determine whether a change has happened.

Pixel-based post-CCD techniques are most often used with coarse to moderate spatial resolution imagery. Class labels for these relatively large pixels are compared to determine from–to change. For high spatial resolution imagery, and increasingly also for moderate resolution imagery, OBIA methods are more commonly used. In OBIA change detection (OBCD; Chen et al., 2012; Hussain et al., 2013), the first step is to create meaningful image objects through a process known as image segmentation, in which spectral, contextual, shape, texture, and ancillary information is used to group spectrally homogeneous neighboring pixels into groups called objects (Blaschke 2010). These objects can then be classified in accordance with an appropriate system of class labels. The labels are compared over multiple dates to determine change.

The main advantage of post-CCD is the provision of class-specific from–to information for each change. Disadvantages include a lack of information about the spectral magnitude of changes and dependence on the accuracy of the input classifications. The accuracy of the change map can only be as good as the input classifications. Errors in any classification will lead to errors in the change map.

The visualization, or presentation, of change results is an important component of any change detection study. The following methods are generally used to represent the change detection results: animations, difference or change maps, and change matrices. The animation technique is a qualitative tool that helps to display a set of images in sequence. Difference or change
maps allow the visualization of from–to changes for different years. For example, areas could be highlighted to represent changes from one class to another.

A change matrix provides a tabular representation of from–to change information for classes of interest. The matrix rows represent one classification date and the columns another. The cells of the matrix then show either the number of pixels or area that have changed (off-diagonal elements) or have remained the same (diagonal elements) between the two dates. For example, in a postclassification comparison analysis using a satellite image from 1990 and another image from 2000 for the same area, and using five land-cover classes, the matrix of change will contain the following information: the total area for every class in year 1990 and 2000, and a detailed distribution of how much of the original area for each class has either changed into another class or stayed the same. This matrix of change technique reveals the reduction or increase of specific land cover classes in a quantitative and straightforward way.

Validation and complications

Change detection analyses require proper data selection and postacquisition processing steps to minimize errors in the resulting change map. Data acquired by the same sensor are assumed to have similar scale and spatial resolution. Thus preprocessing steps are essential for accurate registration and appropriate temporal scales between the analysis dates. Geometric processing, atmospheric corrections, and radiometric normalization are conducted postacquisition and prior to comparing multitemporal images. These procedures are used to correct horizontal registration, adjust for atmospheric effects and cloud cover, and normalize pixel values across all time intervals in the analysis.

Geometric processing, in which the geographic coordinates of a feature on one image are matched to the same coordinates in an overlapping image, is accomplished using a reference image or ground control points, resampling method, and elevation data. This process minimizes positional errors by adjusting image pixels to the same location based on the reference image or ground control points. Elevation data are integrated to reduce displacement from the sensor passing over variable terrain and causing changes in the scale of ground features closer to or further from the sensor. Root mean square error (RMSE) provides a useful metric for comparing positional accuracy between images. This metric describes the amount of horizontal offset between several sample locations in the reference data and the geometrically corrected image. Historical data sources may also require a geographic transformation to a common projection, coordinate system, and datum prior to producing a change map. The transformation provides an updated reference for the geographic position of overlapping images and the features on them such that the same coordinates and measurement units are applied to the analysis.

The view angle, or the angle from the sensor to a geographic position within the sensor’s field of view, should be considered when selecting images for comparison. View angle is directly related to areas of occlusion from building or tree lean and shadows from varying solar azimuth and zenith angles. Optimally, images used for change detection should have the same, or very similar, view angles.

Cloud cover, atmospheric moisture, and haze should be assessed prior to the change analysis and addressed by removing or replacing clouded and shadowed pixels. Atmospheric corrections are necessary for reducing attenuation from
haze or humidity. Radiometric normalizing per scene will provide similar spectral characteristics between multiple dates of imagery, provided the acquisition dates and phenology are similar. This process adjusts individual pixel values for the images by reducing noise from atmospheric conditions or sensor errors.

Vegetation, agricultural, and urban phenology are important considerations when selecting image dates. Anniversary dates minimize the impact of vegetation phenology where different development stages modify image pixel values. Depending on the change detection method used, such variations may result in errors within the change map. Phenology must be considered when determining the temporal scale for assessing change; short temporal scales might underestimate change in general, while longer temporal scales might underrepresent change for plants that regenerate over shorter life cycles. Agricultural growth stages, the cycle of field preparation, planting, crop growth, and harvesting should correspond to the dates of analysis such that change is not falsely detected using images acquired during the growing season and compared to images acquired before the growing season or postharvest. Urban greenness changes, such as the transition from low-density to high-density development, could be underestimated over longer temporal scales where the change occurs in a dispersed pattern rather than as a continuous increase along the urban fringe. Weather events preceding acquisition must also be considered, as increased precipitation or drought conditions will impact water/wetland area, soil moisture, and vegetation greenness. Such weather-caused changes may be confused with class label change.

Spatial resolution and radiometric resolution become concerns when change detection involves data from multiple sensors. Change maps derived from sensors with different spatial resolution limit the analysis to the coarsest resolution such that change detected at a finer scale will be aggregated into a larger unit (Walsh, Butler, and Malanson 1998). Radiometric depth frequently varies between sensors and must also be adjusted so that the bit depth matches between datasets. Images acquired with the same spatial resolution but different radiometric resolution may lack sufficient comparability for identifying subtle differences between features. A forest stand surrounded by shrubs might appear discrete in images with high radiometric resolution while the same area in an image with lower radiometric resolution might appear as more transitional between the vegetation types. These concerns should be addressed using the appropriate image-processing tools prior to analysis and be considered when assessing the accuracy of the change map or the resulting classifications used for postclassification change detection.

Assessing the accuracy of remotely sensed data has proven challenging as image resolution and classification procedures have advanced in recent decades. These challenges have been well documented and discussed in the literature (Congalton and Green 2011; Olofsson et al. 2012). Accuracy assessment and quality control and quality assurance procedures specific to the intended use of the change map should be followed.

Accuracy assessments based on reference data from photo interpretation or ground-based observations provide error metrics for determining the level of uncertainty associated with a change map. However, accuracy assessments for individual classifications are insufficient for assessing the accuracy of the change map and should only be considered for the accuracy of their corresponding classification. Misclassified features within one classification may be difficult to classify in images acquired at other dates, and result in compounding error in the change map.
CHANGE DETECTION

These errors may also extend to the reference data or mislabeled features in each classification. These uncertainties are inherited in the change map; however, registration errors, mixed sensors, and different spatial resolutions should also be considered when assessing the accuracy of the change map. Small horizontal offsets between dates that produce underrepresentation or over-representation of change require assessing the accuracy of the change map independent of the individual accuracy metrics for each date. Photo interpretation or repeat field sampling should be conducted to assess the accuracy of the change map.

Classification schemes also present challenges when comparing multitemporal results. A feature classified using different values or labels at one date compared to a subsequent date can result in false changes. Further aggregation of classes or conflicting descriptions of classes such as an area classified as agriculture at one date and pasture at another date might be identified as change when the only change is in the class label scheme or definitions of the classes. Classification schemes should be reviewed and redefined to a common scheme prior to performing analyses.

Applications

Change detection techniques can be applicable to a variety of fields. The following are examples of applications of change detection.

The use of remote sensing to examine deforestation is long-standing. A study by Hansen et al. (2013) used multitemporal images for mapping global forest change. A group of researchers led by University of Maryland scientists created the first global forest change map from 2000 through 2012. They used 654,178 growing season multidate Landsat 7 satellite images to show a global loss of 888,000 square miles of forest and a gain of 309,000 square miles of forest. Also, they discovered that the deforestation rate in the Brazilian Amazon has been reduced by half since the mid-2000s, as opposed to Indonesia where the deforestation rate has doubled in the same period. In another study, conducted by Guild, Cohen, and Kauffman (2004), pre and post change detection techniques were used for identifying deforestation and land-cover conversion. They used two images (1984, 1992) from the Landsat Thematic Mapper (TM) sensor to identify deforestation and cattle pasture formation in Rondônia, Brazil, employing two pre-CCD techniques and one post-CCD technique to identify changes in the forest and other land-cover areas. The two pre-CCD techniques were used to enhance differences between the forest and cleared and regrowth areas and to identify any overall. The postclassification technique helped them to quantify specific land-cover class changes. Over eight years, they found that the forest areas decreased by 11%, having been primarily converted to pasture, and that 7% was lost because of hydroelectric dam flooding.

Another common use is tracking the extent and condition of water bodies. A dramatic example of an environmental change detection application is tracking the rapid loss of the Aral Sea. This water body was once the fourth largest freshwater lake in the world. However, between 1973 and 2000 approximately 60% of the lake disappeared due to a large water transfer project which diverted fresh water to provide irrigation for cotton and food crops. Thus, a comparison of multiple dates of remotely sensed imagery could be used to identify the changes that occurred from 1973 to 2000. For a more quantitative analysis, a post-CCD approach could be used with satellite images to identify the conversion of specific classes including water and its conversion to other land-cover classes.
Given the increasing rates of urbanization in many parts of the world, studying such increases and their implications for urban planning issues such as infrastructure development, storm water flow management, and zoning is of great importance. Yuan et al. (2005) demonstrated a method to map landscape change in a major metropolitan area over 16 years using Landsat imagery. Their results indicated that urban cover had increased from 24% to 33% of the area, with some areas exhibiting a near-total conversion of classes such as agriculture to urban. Such information is of importance to planning and policy groups at all levels of government.

Change detection is also useful for monitoring hazards. For example, pre-CCD techniques and multiple images can be used to identify changes in the landscape due to drought conditions. Washington-Allen et al. (1998) provide an example in which a 15-year time series of Landsat images from 1972 through 1987 was used to evaluate the effects of drought on vegetation covers in the Bolivian Altiplano. They identified the different changes in vegetation cover due to drought using the image-differencing pre-CCD technique. Visualization change maps were used to show how all vegetation types were impacted by drought. The results indicated that nearly 90% of the vegetation types had not changed between 1972 and 1987. Furthermore, they found that the wet meadow vegetation type was the most resistant to drought conditions, which is a key resource for the agropastoral communities in the Bolivian Altiplano. This is a practical use of change detection techniques to discover information regarding vegetation species that are more resistant to drought conditions, which may be of use to local communities.

Geomorphologists use multitemporal imagery for identifying landscape change as a result of flowing water. Water flowing from an upland area in a watershed or catchment becomes channelized as it moves downhill and forms stream channels. The meandering or winding patterns formed by streams can be measured and monitored as the surrounding landscape is modified at varying time intervals by removing or depositing sediment and vegetation. These dynamic processes result in changes to the channel width, increasing or decreasing meanders, and new landforms such as gravel bars within the stream channel. Multitemporal data can be used to detect these changes over short time intervals in dynamic streams where loose sediments and varying flows cause semi-annual and annual changes, and for longer time intervals where the channel migrates across the valley floor and impacts natural and human areas. Images provide this view for individual stream processes and the network of streams as water moves over the landscape.

Geomorphologic change detection using imagery is more broadly applied to identify changes on the landscape and the resulting impacts on natural and anthropogenic environments. Multitemporal images have been used for viewing lateral changes such as channel widening from erosion that threaten human land uses, debris jams that might create temporary dams and flood infrastructure and agricultural land, and areas where sediment deposits have straightened the channel and present flooding hazards. Erosion, flooding, and deposition of sediments that alter the stream course reduce arable lands, threaten urban infrastructure, and change property boundaries. As a result, stream channels have been straightened and armored for moving water away from an area more quickly (e.g., reducing sediment deposits and flood risks) and for minimizing loss of land from erosion or shifting channels.

Finally, satellite sensors have provided data on human rights violations and humanitarian relief efforts. Civilian satellite imagery can be acquired...
on temporal scales that facilitate monitoring human rights violations. Darfur (Western Sudan) provides an example of using satellite imagery to map military conflict and human rights abuses. Sudanese militias moved into the Darfur region in 2003 and began attacking residents and destroying their villages. Individual homes were classified in an image acquired preconflict and compared to a classification based a postconflict image (Sulik and Edwards 2010). Sri Lanka provides another example of human rights violations that were detected using high-resolution satellite imagery. Satellite imagery spanning four days (May 6–10, 2009) were acquired following reports of mortar attacks on civilians. Analysts identified the position of the mortar attacks and the location where the mortars were fired from, thereby linking the attacks to military forces. These images were also used for identifying an increase in gravesites from casualties in an area identified as safe for civilians.

Summary

Change detection is an important and widely used technique across a range of disciplines. The benefits of change detection include flexible methods that can be customized to a particular application, quantitative estimates of change in land-cover/use categories of interest, and a synoptic view of a changing landscape that may be impossible to obtain in other ways. Care must be taken to ensure that the change detection methods used are appropriate for the study goals and input data.

SEE ALSO: Hyperspectral remote sensing; Land change science; Map projections and coordinate systems; Microwave remote sensing; Optical remote sensing

References


Chicago School

Winifred Curran  
DePaul University, USA

Based at the University of Chicago from 1914 through the 1940s, and using the city of Chicago as its laboratory, the Chicago School sought to use deep ethnography in the form of detailed neighborhood case studies to develop larger, universal theories about the city’s growth and organization. Lacking an urban tradition of its own, geography looked to the theory and methods of the Chicago School of urban sociology to build the subdiscipline of urban geography. This theory was shaped by human ecology, the assumption that the city functioned as an organism, and thus, that this landscape’s organization was shaped by competition over space in the struggle for existence. It was highly influential in the development of an urban geography, especially among those who attempted to use quantitative methods to understand the organization of urban space.

The pre-eminent figure in the Chicago School was Robert Park. Park constructed his human ecology perspective of urban sociology on accepting his job at the University of Chicago in 1914, where he served until retiring in 1933. He was heavily influenced by the philosophy of Georg Simmel and his Kantian separation of social form from content, as well as by the concept of social space: space as a mental category necessary for the study of social forms. The urban was central to this conception. For Simmel, the experience of the city was inherently unsettling, but one to which people could adapt. This concept directly fueled the Chicago School’s focus on the process of the organization, disorganization, and reorganization of urban space.

While Park was well schooled in human geography, notably under Alfred Hettner, he was at pains to distinguish it from his human ecology. He assigned to geographers the role of describing the facts of location, but saw human ecology in the much more expansive role of discovering the underlying principles used to interpret facts; geography was idiographic, ecology and sociology nomothetic (Entrikin 1980).

Chicago proved an ideal laboratory for Park and the other members of the Chicago School, notably Ernest Burgess, R.D. McKenzie, and Louis Wirth, all of whom contributed to the definitive text announcing the arrival of the Chicago School, The City (1925), as well as Harvey Zorbaugh, author of the equally influential The Gold Coast and the Slum (1929). They observed and wrote during the first period of urban life in Chicago, which had morphed from a small lakefront- and river-based prairie settlement into a major industrial city just beginning to garner national attention between 1880 and 1920. Park and Burgess came to Chicago before World War I; their personal transformations and that of the city were interconnected; as an ecologist would say, they were symbiotic. The theory went on to become foundational, the yardstick against which other American cities would be measured, such as in E. Franklin Frazier’s (1937) import of the theory and methodology of the Chicago School to his study of Harlem, in which he argued that a segregated racial or cultural group would develop the same patterns of zones in the larger city. Dear (2002) attributes
the Chicago School’s success to the model’s beguiling simplicity, the construction of a substantial body of literature, and the fact that the model was seen to work in many different cities over a long period of time.

Even within this relatively small group of researchers, there was some disagreement as to the limits of ecology as a metaphor for urban changes. Park argued that reality was constructed and that an individual’s actions stemmed from the exercise of free will within the constraints of the larger social whole, but were never determined by it. Zorbaugh, on the other hand, argued that “the city is a natural phenomenon and has a natural history”; it proves “curiously resistant to the fiat of man” (1926, 188, 187).

Regardless, there are a number of ecological processes that are central to the Chicago School’s conception of the city: centralization, concentration within natural areas within the city, competition for position, segregation over natural areas, invasion, and succession, all operating within a city that was assumed to be a unified whole.

Burgess’s (1967/1925) concentric zone model is the most visual expression of the Chicago School’s theory (see Figure 1). It presents a series of zones emanating from the central business district organized according to levels of cultural assimilation and economic and social status. Zorbaugh (1926, 190) summarized the model, which he argued applied to all American cities:

To begin with, they segregate into broad zones as they expand radially from the center — a “loop,” or central business district, a zone of transition between business and resident; an invasion by business and light manufacturing, involving physical deterioration and social disorganization; a zone of working men’s homes, cut through by rooming house districts along focal lines of transportation; a zone of apartments and “restricted” districts of single family dwellings; and, farther out, beyond city limits, a commuters’ zone of suburban areas. Ideally, this gross segregation may be represented by a series of concentric circles, and such tends to be the actual fact where there are no complicating geographical factors.

The expansion of the city involved each wave of urban immigrants in the inner zones invading the next outer zone in a process of succession, a term Burgess borrowed from plant ecology. This process accounted for both physical growth and levels of social organization within the city. Burgess describes this process as analogous to a process of metabolism in order to explain the ways in which individuals were incorporated into a city’s life. The model was predicated on continued high levels of immigration.

The resulting “differentiation into natural economic and cultural groupings gives form and character to the city. For segregation offers the group, and thereby the individuals who compose
the group, a place and a role in the total organization of city life” (Burgess 1967/1925, 56). There would be a similar process of disorganization, reorganization, and differentiation in the division of labor. Social organization broke down in areas where mobility was greatest, leading to a zone of deterioration, “areas of demoralization, of promiscuity, and of vice” (Burgess 1967/1925, 59). The disorganization fueled by the influx of lower-status immigrants created negative environmental consequences and thus generated the move outward of wealthier residents. Burgess found that variations in land values offered the best single measure of mobility, and thus of all the changes taking place in the expansion of the city. While Burgess argued that some level of disorganization was normal as it was necessary for reorganization, an oversimplified view of the model had the effect of presenting the suburbs as models of stability and implied that urban pathologies were inescapable (Warf 1990).

Each new group would cycle through this experience as part of the process of assimilation, which was considered “apparently progressive and irreversible” (Park 1950, quoted in Jackson 1984, 171), and part of the “moral order” (Park 1926, quoted in Harvey 1973, 131). As such, the Chicago School completely ignored the effects of racism and other structural constraints. This is not to say that the Chicago School was insensitive to the issue of race. Park’s own view was that race was not biological. Heavily influenced by working as an assistant to Booker T. Washington at the Tuskegee Institute (1907–1914), where the focus was on education and economic advancement, Park developed his theory of the race relations cycle, in which contact was followed by conflict, then accommodation, and finally assimilation in a pattern that was repeated seemingly everywhere. The assumption that assimilation was inevitable had the effect of naturalizing the continued segregation, most notably of African Americans, in the most marginalized neighborhoods.

Similarly, class conflicts were not adequately addressed. Harvey (1973, 132–133) comments that Engels had recognized concentric zoning in the city but had attributed it to economic class rather than culture, as the Chicago School did, arguing that it is “a pity that contemporary geography has looked to Park and Burgess rather than to Engels for their inspiration.” Instead, planners and geographers reinforced, through policies like zoning, the pattern of the model by assuming it to be true, rather than questioning how it could be changed.

While the Chicago School’s ethnographic method might come to be considered humanistic (Jackson 1984), positivist spatial scientists were attracted to the Burgess model because, Sibley (1995) argues, it was one of the few spatial theories with universal claims. Robert Park, Ernest Burgess, and the other members of the Chicago School saw their sociology as purely scientific. Burgess saw the sociologist as a “natural scientist” (Sibley 1995, 42). Sociology was to be a pure science, removed from emotion, and strictly apolitical. According to Carey (1975, quoted in Sibley 1995, 42), “an overriding concern of the time was to distinguish sociology from socialism.” As such, Sibley (1995, 46) argues, this “objective,” “scientific” (quotes mine) knowledge gained legitimacy precisely because the political climate was amenable.

The assumptions of the Chicago School heavily influenced social area analysis, factorial ecology, and other methods that attempted to make sense of the city through quantitative analysis. Spatial scientists sought to use quantitative data on housing, demographics, and socioeconomic status to understand and explain the pattern of residential differentiation. Books like Berry and Kasarda’s Contemporary Urban
CHICAGO SCHOOL

Ecology (1977) used ecological analysis of spatial distribution and processes. They go back to a Chicago School conception of community, understanding human ecology as the study of how populations adapt in environments which are constantly changing and yet restrictive. While the Chicago School model remained the foundation for further research, much of this work found disparities between the actual and expected patterns. This finding led to a more thorough questioning of urban ecological assumptions and the development of more behavioral approaches as an alternative.

The scientific detachment of the Chicago School and its followers contrasted with the other “Chicago School” doing similar work at the time, that of the University of Chicago’s School of Social Service Administration (Sibley 1995). Much of this research was carried out by women such as Edith Abbott, Sophonisba Breckinridge, and Helen Rankin Jeter, who conducted in-depth studies of the Chicago housing market, with a special emphasis on landlordism and the effects of discriminatory practices and a focus on policy implications. While Burgess recognized this research as “outstanding,” he differentiated this “practical” social work from the supposedly more serious pursuit of “theoretical” sociology. Park was more vehement, declaring that these women reformers “had done more damage to the city of Chicago than corrupt politicians or gangsters and he discouraged students from taking their courses” (Sibley 1995, 41). Modern-day urban geography has continued this gendered knowledge. Any text on urban geography and planning will include the Burgess model, but not the research of the School of Social Service Administration.

Beyond these structural critiques, there were, even early on, questions about the accuracy of the Burgess model and the Chicago School theories. Although often lumped in with the Chicago School, Homer Hoyt argued as early as 1947 that the city’s residential patterns more closely resembled sectors than concentric rings. Similarly, the multiple nuclei model of Harris and Ullman contradicted a fundamental assumption of the Burgess model. Others critiqued the model for trying to universalize the experience of the American city, for the Burgess model fails to explain the growth of many European cities, for example.

The focus on social pathology related to another fundamental understanding of the Chicago School, that of urbanism as a fundamentally distinctive way of life. Wirth (1938) argued that urban dwellers had weak primary contacts as a result of the city’s size, density, and heterogeneity, and that superficial contact was a way of immunizing themselves against the sensory assault. Wirth further argued that the only way for individuals to become effective was to organize into groups, and he sought to build an urban sociology that would make sense of the complexity of urbanism by building on this understanding. A large body of work contested this understanding of urban life, giving evidence of extensive community and nontraditional family ties, and indeed arguing against the very notion of urbanism as a separate way of life.

One positive legacy of the Chicago School is the in-depth case studies of “natural communities” that fueled the theory. In this way, communities that may not otherwise have received sustained academic attention were documented in detail, through surveys, participant observation, and interviews. So, for example, Chicago School studies serve as a resource for understanding aspects of gay life in Chicago and the geographic range of the community at the time. Burgess and Park took an interest in the urban expression of sexuality and directed graduate students to conduct field work among Chicago’s homosexual communities. Aspects of
this work were later published by Burgess under the categorization of social pathology. Glimpses of gay life can also be found in Zorbaugh’s *The Gold Coast and the Slum* (1929).

Whatever usefulness the Chicago School may have served in explaining the twentieth-century American industrial city, many argue that the process of gentrification has undone the model and renders it no longer appropriate, with upper-income professionals increasingly attracted to the central city, displacing the working class and ethnic minorities to locations farther and farther from the urban core. Even in Chicago, what had been the zone of transition has been transformed into luxury condominiums. Disused train tracks have become Millennium Park, one of the most popular attractions in the city. And yet, as Warf (1990) comments, this process exhibits precisely the characteristics of residential competition on which the theories of the Chicago School were built.

The theories of the Chicago School maintain a hold on the urban geographic imagination, with various attempts to rescue the theory from its own shortcomings. Warf (1990, 91) tries to recover some elements of usefulness from the Chicago School, arguing that it provides some of the best work on neighborhood change and ethnicity, which should be blended with an understanding of the division of labor to produce a new social ecology that underscores the “contextual nature of social action.” Even Soja (1989), in his exploration of a postmodern Los Angeles School of urbanism, recognized an underlying logic consistent with the Burgess model, with radial growth from a central core. Indeed, elements of the model are still visible in certain sections of Chicago to this day. The Loop remains the central business district, with much of the manufacturing that remains in the city concentrated in what was the industrial zone. The north and west sides of the city are bordered by suburbs (Evanston and Oak Park) that are much wealthier than the urban neighborhoods inside the city limits.

Perhaps one of the most enduring legacies of the Chicago School is the idea of a comprehensive social theory based on the work of individuals researching the same locale. This notion was revisited by Michael Dear (2002) in his call for a new school, the Los Angeles School, to replace the Chicago School, given the evolution of the city in a globalized world. Dear finds evidence of the transition from the Chicago School to the Los Angeles School in Louis Wirth’s final chapter of *The City* (Park et al. 1967). Wirth isolates two fundamental features of urban evolution: that the city lies at the center of, and provides the organizational logic for, a complex regional hinterland, and that the development of satellite cities can exert a determining influence on the direction of future growth. Dear argues that this development foreshadows the postmodern city, of which Los Angeles is the most iconic, in which it is not the center that organizes the hinterlands, but the hinterlands that determine what remains of the center. While the idea of the Los Angeles School never exerted the influence of the Chicago School, it does illustrate the longevity of the original theory.

More recent understandings of the Chicago School have tended to emphasize not the antiquated vocabulary or overly simplistic models, but rather the focus on attempts to reorganize social groupings and identities after modern urban life has disrupted the traditional institutions of family and locale. This attempt to redeem the fragmentation of urban life, to see the potential in the chaos, drives many urbanists and may explain, in part, the rediscovery of the urban work of theorists like Walter Benjamin, whose dialectical urbanism recognized the shortcomings of the capitalist city while celebrating
its potential to enhance individual potential and collective freedom.

Despite the many and varied critiques of the Chicago School, we can perhaps appreciate this focus on the city as the object of study, the depth of their methods, and agree with Park that what is best about cities is their diversity:

In the freedom of the city, every individual, no matter how eccentric, finds somewhere an environment in which [s]he can expand and bring what is peculiar in his [her] nature to some sort of expression. A smaller community sometimes tolerates eccentricity, but the city often rewards it. Certainly one of the attractions of the city is that somewhere every type of individual – the criminal and beggar, as well as the [wo]man of genius – may find congenial company and the vice or the talent which was suppressed in the more intimate circle of the family or in the narrow limits of the small community discovers here a moral climate in which it flourishes.

(Park 1952, quoted in Jackson 1984, 174)

SEE ALSO: Central business district; Ethnography; Gentrification; Human ecology; Los Angeles School

References


Today nearly half of the planet’s 7 billion people are aged 25 years or younger, and a quarter are thought to be younger than 14 years. An estimated 9 out of 10 of these young people live in developing countries. Given such demographics, it is surprising that the study of children and how they inhabit, experience, adapt to, and change their environments has only relatively recently received attention from geographers. In part this is because the newly emerging studies (geographies of children) are set against a wider demographic revolution in the developed world, associated with smaller family sizes, shrinking populations, and a dramatic aging of the population. The United Nations Charter on the Rights of the Child defines children as those aged under 18 years. From postwar research into children’s understanding of maps, and map-making, to contemporary longitudinal cross-cultural and interdisciplinary research, the study of children and their environments has blossomed (Holloway 2014). This trend has been supported by the recent creation of active national research groups including the British Royal Geographical Society’s Geographies of Children, Youth and Families Research Group and dedicated journals including Children’s Geographies and Children, Youth and Environments.

Thirty years ago, developmental psychologist Bronfenbrenner (1979) defined children’s environments in social-ecological terms. Bronfenbrenner depicted children’s environments as a complex series of nested interactive systems, in which the child is both embedded in a dynamic world of everyday micro-level interactions and influenced by indirect but equally significant macro-level processes, including economic, political, cultural, economic, and physical change. These wider institutional interactions can include, for example, conditions of parental employment, media or education policy, tribal and kinship structures, and government policies about urban development, natural hazard mitigation, or climate change.

Universally, children today face four intersecting macro-environmental issues: dangerous environmental change, growing social inequality, unprecedented global youth unemployment, and weakening local democracies as communities struggle to hold global corporate power to account (Hayward 2012). Since the 1987 Brundtland Commission on sustainable development, policy advisers, urban planners, and political leaders have been urged to consider the needs of “future generations” at all levels of environmental decision-making, from macro-policy to micro-level planning. Children’s geography also encourages attention to the future generations who already live among us in this policy
process: the world’s children. By 2050, 70% of the world’s children youth will live in cities, yet children do not experience these urbanizing environments in uniform ways. For many children, urban communities afford an unparalleled quality of life, facilitating new opportunities for education, social interaction, and employment. For children in families with a high income, resource-intensive, urban consumer lifestyles may be an expression of identity; for others, however, cities may afford few choices. Some children experience little control over their environment, being “locked in” to intensive energy use in everyday living, with few opportunities to choose more sustainable alternative living environments, or displaced from place as trafficked or forced migrants.

Cities are also precarious environments for the world’s poorest young people. Half of all people under the age of 25 live in situations in which they and their families have access to just 9% of global income, and in this context child labor may be the means of survival (Ortiz and Cummins 2011, 21). UNICEF (2012) estimates that one city-dwelling child in three lives in conditions that lack security of tenure; lacks access to adequate health services, housing, education, or diet; and faces complex risks of exploitation, discrimination, and violence. Recent research is aiding greater understanding of children’s environmental experiences in the Global South and North as developed economies also struggle with recession and urban inequality, which disproportionately affects children and young migrants (Wells 2009).

Understanding children’s diverse environmental experiences

Understanding children’s diverse experiences is the subject of growing interest, culminating in the emergence of the field of children’s geographies. Early studies of children and the environment in Europe and North America often depicted children as passive, vulnerable dependents who needed to be protected from the impact of modernity and guided in their development. Rousseau and Wordsworth, for example, wrote on the importance of children’s play in the natural world as a means of countering the pressures of modernity. Elements of this Romantic perspective still resonate within contemporary debates about whether children are suffering from a “nature deficit disorder,” as a result of rapid urbanization and associated loss of access to green space and opportunities for unstructured free play (Louv 2005) and the need for biophilia (identification with the natural world) and environmental place-based education (Sobel 1996) to nurture values of environmental stewardship and a positive sense of self and a place-based identity.

Beyond efforts to bring urbanized, and often middle-class, European children into closer contact with nature, there are contrasting debates concerning the appropriateness of children’s presence – and more especially their behavior – in a range of outdoor and public spaces. At times, this debate has ignited as a moral panic, for example, over whether children are becoming more risk averse in the outdoors and turning into a generation of “cotton wool kids” over whom overly protective “helicopter parents” constantly hover to ensure their “protection” from imagined harm (Valentine 2004). Children denied the opportunity to engage in wild play, instead, are assumed to spend their leisure hours indoors, preferring screen entertainment to face-to-face interaction. In turn, this fuels worry about a new generation of isolated, lonely, and anxious children and teens who experience anomie, are losing the ability to relate to others, or are loitering in ways that threaten public order.
Public attitudes to children and their technological environments are no less fraught with contradictions and anxieties. On the one hand, the new technology of smartphones is reducing the costs and barriers to entering the virtual world, in turn enabling new opportunities and greater connection with others. Yet the digital divide continues to exist between wealthy communities and children in communities without affordable and efficient connections. Moreover, the issues of surveillance and exploitation of children through the Internet, including loopholes in popular apps such as Farmville and Facebook, mean that the Internet is an increasingly problematic environment for children as a potential platform for social interaction and social control.

In the Global North, children and young people are often seen as perpetrators of the worst excesses of consumerism in their quest for self-expression, peer affiliation, and cultural participation; yet, on the other hand, they are often held up as our “best chance” for a sustainable future; “Trojan horses” of change who care deeply about the environment and are committed to living within planetary means (Collins and Hitchings 2012). At the same time, neoliberal economies and corporations are increasingly fulfilling the role of the entities that are not only “feeding, clothing, entertaining, socialising, but also moralising and ‘caring for’ children” to build a lifetime of brand loyalty and a habit of spending” (Schor 2011, 207), thus limiting opportunities for children to practice forms of social life and interaction that are fundamental to childhood and to the development of autonomous citizens.

When children do leave the screens of their smartphones or computer games consoles to explore the physical world, their unstructured play in urban environments is often depicted as a threat to the prevailing social order (Vanderbeck and Johnson 2000). Again new surveillance technology has evolved to document children’s movements in public space from personal GPS tracking to closed circuit television, regulations, and signs or mosquito-sound alarms placed in supermarket entrances or other places where children are likely to “loiter.” Other attempts to control children include England’s wide-ranging antisocial behavior laws that can be used to restrict minor activities like ball sports to the wider prevention of youth congregating in public spaces and the searching of children if they are suspected of participation in potentially criminal acts.

Against these contradictory experiences, many researchers recognize that children can no longer be understood simply as dependents or victims but may be regarded as agents of change who have varying influence in their environment. In this context new research methods aim to give voice to the diversity of children’s experiences. Movements like “guerrilla geographers” (Geography Collective n.d.) and other critical place-based pedagogies have drawn inspiration from Paulo Freire’s liberation learning, to promote child-centered education and to help children to understand their environment, reflect critically, and act with others to implement strategies for desired change in the places they value (Gruenewald 2003).

**Child-centered, longitudinal, and interdisciplinary research**

What has the plethora of children’s contradictory experiences meant for the ways in which research is conducted? While communities experience social injustice differently, what drives much contemporary interdisciplinary study of children and their environment is an implicit or explicit recognition that children have a significant stake
in the outcome of planning decisions because they will experience the effects of these decisions over the course of their lifetime, and should therefore also have the opportunity to be heard and to influence decision-making.

Longitudinal research methods direct attention to important questions about how children engage with their environment and the way they are exposed to environmental burdens and opportunities over time. Many children in mid- to high-income communities, for example, currently have unprecedented access to world travel and quality local outdoor environments, and live in homes that are well built, heated, and digitally networked. At the same time, one in three children are disproportionately exposed to a range of risks, including domestic violence and alcohol and drug abuse, which compound macro-level socioenvironmental threats of war, civil disruption, climate change, or economic hardship (UNICEF 2012) and enormous numbers are subject to forced migration (Dona and Veale 2011). The experience of children in the globalizing South or in low-income and ethnic minority communities can be quite sharply divided. Many children in these communities are inadequately protected from, for example, toxic waste pollution; their daily lives are disproportionately impacted by clearance for private development; or they are exposed to excessive risk as child laborers. Poor children are 10 times more likely to bear the brunt of severe climate change, while women and girls are 14 times more likely to die in a disaster (Plan International 2011). Many children, especially girls of low-income ethnic minority communities, are “doubly exposed” to both physical and social-economic risks in their environment, including risks of severe storms, urban heat waves, and flood events associated with dangerous environmental change. These physical risks are compounded by the lack of quality education, information, and health security and by secure or insecure housing and exploitative work environments (Leichenko and O’Brien 2008).

Understanding of the experiences of children in their environments has been enriched by development studies and research into citizenship, globalization, mobility, development, feminism, and sexuality. Researchers in these fields have supported a growing interest in how children are socialized into power relationships and “governed” in space and time. In turn, there is also interest in youth agency, or the ability of children to imagine and effect change (Hayward 2012). In a classic lecture, Ruth Lister (2007) argued that children do not have to “become” citizens (i.e., to assume the full expectations of adult responsibility and entitlements). She argues that children already “belong” as citizens who participate in and make claims of their communities in the present. In this light, geographers have also turned their focus toward understanding the diverse ways that children’s public space is regulated and increasingly commercialized, forcing children, for example, to play in restricted ways and often in highly consumerist contexts such as shopping malls. By contrast, there are fewer spaces where children can create the conditions of play for themselves and thereby give their own meanings to those spaces.

Finding ways to better understand children’s diverse environmental experiences received an impetus with the formative work of critical children’s geographers like Valerie Walkerdine and Roger Hart. The latter argued that children’s voices are rarely given significant space or serious contemplation by adults (Hart 2008). While adult conferences might adorn their walls with children’s artwork, they rarely listen to or consider the issues that really trouble children. Hart adapted the urban planner Sherry Arnstein’s (1969) ladder of participation for children
to illustrate how children are frequently incorporated into environmental decision-making processes in ways that amount to little more than co-option or placation. He concluded that it is rare that children are regarded as partners in urban planning and even rarer that they are able to lead a decision-making process.

In response to such concerns, critical geographers and political ecologists interested in children’s issues have sought a shift in focus, widening their search for ways to bring the voices of children to the center of the research process. Research methods which do not directly involve children (such as large-scale observation) have been criticized for working on, rather than with, children. Instead, more inclusive and participatory methods have been developed, often in conjunction with longitudinal life course or cohort studies. These approaches aim to involve children as active coproducers of knowledge, raising their awareness of their own knowledge of their environments, as well as developing a deeper understanding of children’s perspectives through research methods suited to their capabilities. In particular, visual and verbal research techniques, such as drawing, photography, stories, and song can be effective ways of eliciting the experiences, opinions, and perspectives of children.

The child-centered turn in research has also begun to emphasize children’s agency in their environment in the context of enhancing children’s “resilience” (the ability to cope or recover from stress or disaster) (Tanner and Seballos 2012) and the child as a political actor in processes such as forced migration (Dona and Veale 2011). While the recognition of children’s ability to take action is welcome, there is a danger that too great an emphasis on children’s individual ability to adapt or cope to new environmental stress is undesirable. Children can accommodate significant change, but critics argue that there are limits to how much imposed socioeconomic or physical environmental change young citizens should have to adapt to before it becomes unreasonable, unsustainable, and unjust (Ungar 2012).

While much research in children and environments is driven by the Global North, new voices are highlighting the way children experience diverse local environments, as street children in mutual solidarity, as child soldiers, or as child labor. As more studies reflect the complex issues of child health and children’s exposure to natural or human-made disaster (including economic crisis), new questions arise about how children experience temporal (intergenerational) justice and spatial (intragenerational) fairness (Katz 2004; Wikenden and Kembhavi-Tam 2014). Work by authors like Wells and Katz reminds us that an emphasis on agency cannot condone or excuse unrealistic or unjust expectations, for example, that children should atone for, or carry the costs of, environmental degradation or social suffering caused by the action (or inaction) of previous generations. It may take a village to raise a child but in the twenty-first century, as the village has become the megacity, our boundaries and ethics of care for children are challenged to expand. Studies of children and their environment are highlighting the complex ways that actions in one place and community impact on the life chances of children separated by space and time (Mills and Kraftl 2014). This complexity and the enormity of the issues confronting new generations challenge social researchers to ensure that the outcomes of their work can support children in their diverse and rapidly changing environments.

**SEE ALSO:** Children and youth; Citizenship; Social movements; Social resilience and environmental hazards
References


Valentine, Gill. 2004. _Public Space and the Culture of Childhood_. Aldershot, UK: Ashgate.


**Further reading**

Children and youth are relatively new subjects and foci of study within geography. Despite their recent entry into the discipline, the growth in academic work related to this younger age group has been exponential, for two key reasons. One is that more geographers have turned their attention to working with younger people and, in this way, have contributed a great deal to understanding of the roles of space, place, and spatialities in young people’s lives. They have critically examined the concept of age, the role of structures in young people’s lives, and the complexities and contradictions around children’s and young people’s rights (Holt 2011). Second, many geographers have been determined to make children and young people more visible within the discipline as part of an intellectual and political project. Their work has demonstrated that younger people, their identities, their spatialities, their agency, and their lived experiences, illustrate different kinds of geographies from those of adults and force us to rethink our conceptual frameworks. Young people’s geographies are being examined in ways that not only “add them in” to geography as a discipline but also form a range of conceptual challenges to taken-for-granted notions and approaches within human geography. Just as young people themselves challenge adultist sociospatial constructions, geographical research with young people provides interventions in adult-centric normative geographies.

The invisibility and neglect of children and young people within the discipline of geography has been one of the key academic and political imperatives that have driven the development, expansion, and intellectual project of geographies of children and youth. With the aim of making children and young people evident within the wider discipline, the subdiscipline effectively broadened the purview of geography as a new group of social actors has been included, to challenge assumptions and presumptions about key concepts such as mobility, politics, and urban life. Just as feminist geographers challenged the masculinism of the discipline, children’s and young people’s geographers confront the power relations of adultism, recognizing that children and young people are significant members of human society and actors in the physical and social environment who should be taken seriously.

Geographers working with children and young people have had to transform existing methodological approaches and to develop new methods, tools, and techniques to capture children’s and young people’s articulations of their lives. This exercise has led to remarkable innovation and invention in order to rework adult–child power relations, and to make research fun and meaningful for participants. Children and young people’s geographers have used modern technologies such as smartphones, Global Positioning System (GPS), texting, Facebook, and other tools to develop research methods that are relevant and connect effectively with a particular age group. To challenge the power inequalities of the adult–child relation, children and young people’s geographers have adapted more equitable and engaging methodologies, such as participatory and action research as...
CHILDREN AND YOUTH

well as “go-alongs” or walking interviews. To engage younger people and make the research more appropriate, they use photo elicitation, play activities, drawing, and dynamic subject-centered cartography (Kesby 2007). Additionally, geographers (and other social scientists) have faced considerable challenges in terms of the ethical issues and tensions at play in doing research with younger people. Ethics is particularly problematic in relation to protocols that fix children as vulnerable, incompetent, or unknowing. Geographers have been influential in establishing that children and young people are competent social and spatial actors who have the right to participate in research that affects their lives, those of their families, and of their communities (Beazley et al. 2009).

Early work in the 1970s and 1980s challenged geography’s way of focusing on urban planning and experiences of place without considering children and youth. Geographers critiqued developmental psychologists’ conceptualizations of children as immature and dependent on adults by demonstrating children’s spatial and mapping competencies. In the late 1990s and 2000s geographers, influenced by debates in sociology, engaged with the new social studies of childhood where emphasis was placed on recognizing childhood (and latterly youth) as a social construction that is temporally and spatially specific. Additionally, children were conceived as agents and actors in their own right, in contradiction to the perception of their being “adults in waiting”; however, it is very clear that there are substantive differences between a child of 5, a teenager of 15, and a young adult of 25. A call for an expansion of focus away from just children was made in the late 1990s (Skelton and Valentine 1998). Drawing from youth studies and cultural studies, geographers examined the complexities of young people’s lives through spatial and scalar lenses and effectively launched young people as a legitimate and necessary focus, distinctive from and yet overlapping with the lives of children. In the context of development and geopolitics, Katz (2004) illustrated the ways in which children in Sudan and the United States experienced complex interactions with their physical environment; were active within the production of social and cultural reproduction; and were heavily affected by global economic restructuring. More recently, children’s and young people’s geographers have engaged with re-examinations of social reproduction and the role of the family. Theoretically and conceptually, children’s and young people’s geographers have drawn on and developed approaches that range from poststructural, postmodern, and nonrepresentational, to life course and transitions, through to a focus on embodiment, emotions, and affect. They have demonstrated the ways in which these theoretical orientations have relevance in young people’s lives and that attention on younger people enriches and enlivens related debates in the wider discipline.

Defining children and youth

Children and youth have much in common, yet at the same time they are quite different from each other. Geographies of children have tended to be pre-eminent and there is a wealth of published material that captures the spatial and social complexities of childhood and children’s lives in both the majority and minority worlds (otherwise identified as the Global South, or developing world, and the Global North, or developed world). Children’s geographers have rejected the notion of a singular, universal childhood tied to age-standardized durations; they work with a concept of childhood as socially and politically constructed and culturally and historically contingent. They argue that
childhood is neither biologically nor developmentally determined but, rather, is constituted through being in the world. Theoretically and politically, geographers have contributed to and expanded the new social studies approach to childhood through their exploration and identification of the diverse production and experiences of childhood in different parts of society and geographies across the world. These approaches have demonstrated that children are actors in and creators of their own childhoods, that they create their own social worlds because they are active and competent members of society. Consequently children and childhood deserve to be studied in their own right and as legitimate social entities. Once this approach became more mainstream within geography, space was created to expand from childhood and children to include young people and families.

Young people, especially teenagers, are viewed as having an in-between status, sharing characteristics of being both child and adult at the same time as not really being either. This ambiguity is seen as an unsettling position; young people do not quite fit into one age group or the other and consequently are often highly scrutinized and regulated. Nevertheless childhood and youth do converge in quite important if sometimes confusing ways. The 1989 United Nations Convention on the Rights of the Child (CRC) defines childhood as 0–18 years; hence teenagers are included in the convention. The CRC has had important implications for the protection, provision, and participation of children and youth within their families, communities, and nations. The United Nations also defines “youth” as 15–24 years; hence in this definition children are those under 15 rather than 18. The International Labour Organization has a focus on child labor, and attempts to limit and control the amount of exploitative and dangerous work to which children aged 5–17 are exposed. The legal systems within countries specify very different ages at which children and youth can legally get married, leave school, drive a car, have sex, vote, or be held criminally responsible for their actions. The chronological age definition is therefore messily split up according to international and national definitions, posing the question of how valuable categorization by age actually is. This critique allows a degree of intellectual freedom to resist categorizing younger people into discrete and universalized types, and instead focuses on age as a relational construct. This approach has further stimulated examination of the intersectionality of age with other social identities and sites of marginalization and exclusion such as class, (dis)ability, ethnicity, gender, race, and sexuality.

Much of the initial work with young people was developed in the minority world and academically engaged with cultural and youth studies. Youth studies tends, importantly, to be oriented toward policy, and legal and cultural analysis. Research with young people in the majority world is frequently related to development issues, much of it located in Africa, but we still know relatively little about young people’s lives in the Americas and Asia. This situation is quickly changing as scholars from outside and within these regions engage with geographies of young people and children in the majority world and offer critiques and new perspectives.

A predominant trope around young people is that they are either unnoticed or neglected as a group or they experience heightened attention because of their antisocial behavior, their disaffection with politics and/or society, their lawlessness, and their delinquency. The binary around children means that they are described as either angels or devils, inherently good or evil, connected with Apollonian and Dionysian discourses. For young people the same binaries exist, but they are also bound up with more
complex notions of being at risk, troublesome, out of place, or dysfunctional, and with expectations that they should be cured, contained, or coerced to become the “right” kind of citizens. For geographers working with young people there is a determined rejection of the stereotypical categorization of many young people in favor of analyses relating to place, space, and spatiality and the ways in which young people interact with, and act on, their social, cultural, political, and physical environments. Geographical work with and about young people has taken more time to develop than children’s geographies but it is growing in confidence, conceptual depth, and visibility.

The ways in which children and youth are defined and identified within geography are flexible and dynamic. The two are often linked, and linking has a conceptual and analytical validity at times, but in many instances it is important to separate out the 7-year-old from the 14-year-old. The importance of these distinctions is apparent in the varying focus on children and/or youth within major topics in geography. For instance, play might be more associated with children, and politics with young people. Increasingly, though, the broad range of empirically grounded and conceptually critical geographical work disrupts these associations; young people play in urban space through parkour, and children, like those involved in the Arab Spring uprisings in 2011, have a role in political actions.

Play

Play is often seen as quintessentially childlike and a necessary part of childhood; it is something that children expend a great deal of energy on and enjoy. All over the world children play. Play has been identified as an important factor in the enhancement of children’s intellectual abilities, socialization, and wellbeing. These are spatially and temporally contingent; not all play in all places and at all times is beneficial. Moreover, there is a plethora of adult-created forces and power relations that encroach on and delimit the possibilities of space and time for play. Such forces include reductions in safe and attractive spaces for play through modernization, urbanization, and the agroindustrialization of rural areas; increases in automobility, which moves children around but also threatens their safety as pedestrians or cyclists; a stretching of children’s everyday places where they have to spend time (the school on the other side of the city, the after-school and other organized clubs, the tuition centers, the sports training areas); and an intensification of learning and examination pressures. Time–space compression and the forceful speeding up of life that adults experience in a rapidly globalizing and developing world also affect children. They have little time or space for play, whether outside or inside, alone or with others, imaginative or organized, screen-time play, or quiet daydreaming. Despite these restrictions on children’s time and space, conceptually, play is still considered to be an extremely important part of a child’s life and essential for their social and physical development. Children are almost all experts at play and are ever inventive. Wherever they are – unless expressly forbidden (and even when they are) – children will start to play: they run around; they race against each other; they play with their voices and echoing sounds; they jump across or onto lines in the pavement; they hide from and find each other or their adult companions; they climb walls and spin around in revolving doors; they become fascinated by a small detail – a chip in a glass window, a snail in the garden, a loose brick in a wall, pamphlets in the doctor’s waiting room. Play is a form of agency children enact and create, and they can
demonstrate extraordinary competency in terms of design, construction, and contextual meanings for the spaces they play in and the things they play with. Children also enact inventive mimicry of adult worlds and through play learn to make sense of the world around them, including adult worlds and activities (Katz 2004).

The conceptualization of play by geographers recognizes that play is complex and socially and culturally constructed in sometimes problematic ways. Play is dynamic within contexts of time and space. While children and youth are decidedly agents of their own play, playing, play spaces, and appropriate play, in particular, are usually determined and designed by adults. Playgrounds are specifically designed around adult notions of what children *should* do rather than what they *want* to do. Children and youth often work to disrupt and upset these adult-defined notions of play, and conceptually play has been defined as permitting a revolutionary imagination or as lying at the intersection between being and becoming. Even in everyday spaces such as school grounds, children territorialize playgrounds through intricate practices of control, domination, and resistance that are bound up with social relations such as gender, ethnicity, (dis)ability, and age. Consequently play and the spaces/places of play create distinctive and contested sociospatial practices and material realities of childhood linked with aspects of identity and performativity. This is also true for youth if digital and screen-based play, as well as play/movement in public spaces such as skateboarding or rollerblading, is considered.

**Mobilities**

Exploration and movement throughout their environments is considered important for younger people, and is an area of research where geographers have made important contributions. Children’s investigations and engagement with spaces and places beyond their immediate home environments are seen as foundational in their social and personal development. Through such travel, children learn to exercise autonomy and independent decision-making. It also allows them to cultivate relationships beyond their immediate families and households. These elements of spatial mobility can enhance their cultural capital and lead to forms of social mobility. At the same time, such movements through spaces, especially the public space of the street, are considered dangerous and a duty of care is required to protect children and, where possible, transport them through such spaces to reduce risk. For children in poorer communities and countries, arduous, lengthy, and sometimes dangerous journeys have to be taken as they walk to school, contribute to household or economic labor, or fetch and carry necessary requirements, such as water, for their families. The actual pathways of movement themselves can be risky and dangerous. Whether it is the lack of pavements/sidewalks in new home developments in Australia, or crossing streams swollen with monsoon rains in India, the risk of injury or death for children remains. At a different scale of mobility, migrant children and youth who transition to new geographies and cultures can find the new environments difficult and distressing, especially for refugees and asylum-seekers.

A major preoccupation of children’s geographers and child mobilities in the minority world relates to the journey to and from school. In several European countries the school run can contribute up to 40% of urban traffic. Recent initiatives to get children to walk to school and to encourage traffic safety have increased the number of children walking to school, but very few of them do so unaccompanied – children’s autonomous journeys to school have decreased
CHILDREN AND YOUTH

Figure 1 In Kunming, China, children walk home from primary/elementary school: through walking, children are able to navigate their city and have independent leisure time with their peers. (Source: author’s photograph.)

as significant in young people’s lives as they have been in the past or as they are in the majority world. In the latter the street remains a space of vibrant interaction, a locale of play, but also an ambiguous site of danger, risk, and support, especially for street children (Beazley 2003).

Politics

Questions of politics and agency form another key component of geographical work with children and young people. A special issue on “Political Geographies of Children and Young People” (Philo and Smith 2003) was one of the first collections to focus on younger people and politics. Its own political project was to work against the disempowering of young people and toward a political geography that would recognize them as political actors in everyday life and beyond. Young people were shown to be very much political actors in their everyday lives.

A decade later, in “Children and Young People’s Politics in Everyday Life” (Kallio and Häkli 2013) geographers evaluated the ways in which a focus on young people’s politics had connected and critically engaged with the subdisciplines of political geography and geopolitics. The papers demonstrate the ways in which geographers take on the role of making visible young people’s actual and potential political agency, subjectivities, competencies, participation, and capabilities for decision-making. They also illustrate the diversity of ways in which children and young people are “being political” and that there is a great deal to learn about such political processes.

A focus on young people and politics provides the opportunity to critically interrogate established theories, concepts, and methodologies within political geography.

Importantly, a focus on young people and children’s politics carries a specific political
imperative of exploring social justice issues in order to expose processes of marginalization and the exclusion of young people. Borders are a key concept in political and geopolitical geographies. Children’s and young people’s geographers have provided insights into young people’s border-making, resistance, and revolutionary potential from both majority and minority world contexts. They show that young people engage with bounded events, places, and institutions and move them toward an emancipatory state of *sin frontera* (without borders) (Aitken and Plows 2010). Other work on young people’s politics is asking significant conceptual questions about what constitutes politics, and political agency in particular. It recognizes the role younger people play in politics both with a big P and a small p. The former has been examined through formal involvement in political institutions such as local youth councils in the United Kingdom or the United Nations. Small p politics is recognized as being located locally and communally and as involving children of all ages in practices of care and informal citizenship.

Since the formalization of participation as a right of children and young people by the CRC, political agency and participation have been written about as inherently positive. Some geographers, however, have begun to question what kind of agency is politically and socially acceptable and the power relations behind definitions of acceptability: What about agency used in negative ways (to promote racism or sexism, for example)? What about the participation of young people in increasing militarized activities and organizations such as uniformed groups, enforced national service, or “fun” days in military museums? When is participation in politics exploitative, brutalizing, and damaging (child soldiering, being token participants, for example) (Skelton 2013)?

**Urban identities**

“Urban studies,” a term used to encompass urban geography, sociology, economics, planning, and policy, has a mixed history in relation to young people’s geographies. Urban geography has, with some exceptions, exhibited a generic disconnection from young urban actors. In the context of development studies, urban analysis of the situation of children and young people in the world’s poorest cities has been evident for some time, and a good range of edited collections and monographs about youth and the city exists. There is currently a rejuvenation and transformation of theoretical, conceptual, and methodological approaches in urban geography, and this has created space for previously neglected urban actors, notably young people, to be part of that challenge and change. Such changes have included the revitalization of comparative methodologies, and youth geographers (alongside anthropologists) have compared key urban experiences of young people in Latin America, Africa, and Asia, focusing on spatial concepts such as home, neighborhood, and street. Young people constitute the largest demographic groups in many of the world’s cities and play significant roles in the making, changing, negotiating, and navigating of urban spaces. Increasingly, geographers are writing young people into the sophisticated approaches used to capture aspects of the intense complexities of contemporary cities (Skelton and Gough 2013).

Geographers have shown that young people remain ignored or actively excluded from decisions that affect them; most city authorities are not connected with youth in terms of politics, planning, and policy. Geographical work has revealed that young people are social and spatial agents who dwell in the city in ways that differ from those of adults and that they play significant roles in the diversity of urban life. Focusing on
the urban dimensions of young people’s lives allows different questions to be asked about the relationality of cities and the social identity of age. In a spatial context, young people are particularly interesting because of the ways in which they occupy the liminal and interstitial places of the city. These spaces are used or occupied, sometimes by choice, to avoid observation as they indulge in activities such as skateboarding, parkour, unsupervised time with their peers, or illicit and illegal activities; and at other times by force in order to avoid the authorities, find a safe space to sleep, or live out socially rejected identities.

Geographical work on young people in cities has covered a wide international geography and provided a range of analysis. For example, geographers (along with sociologists) have examined the possibility of urban, social, and cultural mobility, and the ways in which place identities serve to support young people in their environments but also hinder them from traveling to study or work in other cities. Recent work in Western, African, and Asian cities has explored the use of mobile technologies for navigating cities and building social and economic connections that can facilitate their possibilities of income generation and social connectivity across urban space. A key feature of examining the cultural and visual elements of the city links to the complex role of graffiti and creative urban expression, much of which is produced by young people. Cities are spaces of conviviality but also sites of intense inequalities. Geographers have explored the many complexities of urban processes and practices in relation to the communality and tensions between people based on gender, ethnicity, (dis)ability, race, sexuality, age, and socioeconomic status. Youth and children in urban spaces have been investigated to critique their apparent antisocial/street-living behavior and confrontations with authority in order to consider possibilities and strategies of more harmonious dwelling but also to encourage a greater respect for young people’s needs and rights as urban dwellers.

Future directions

Geographies of children and young people are continuing to expand and, in doing so, they challenge the established disciplinary approaches, theoretically and conceptually, within geography and the broader social sciences. Working with children and youth forces new theoretical questions to be asked and conceptual tools to be developed, because it is not possible to fit younger people into existing analyses. They demand more creative academic inventiveness. This is happening and will continue to be a work in progress. In order to work with children and youth, adult researchers have had to rethink methodological approaches, as described. The innovation of methodological research tools is a major contribution from the subdiscipline to human geography as a whole.

SEE ALSO: Citizenship; Critical geography; Difference; Disability; Ethnicity; Feminist geography; Identity; Life course; Political geography; Social capital; Social geography; Urban geography

References


Further reading


China: Chung-kuo ti-li hsüeh-hui (The Geographical Society of China (GSC))

Founded: 1909
Location of headquarters: Beijing
Website: www.gsc.org.cn
Membership: 9720 (as of December 31, 2013)
President: Liu Yanhua
Contact: gsc@igsnrr.ac.cn

Description and purpose

The GSC aims at promoting scientific research and education in geography, encouraging the full participation of its members and geographers, popularizing geography, training professionals, and making contributions to the society. The society has over 9000 members from research institutes, colleges and universities, secondary schools, government agencies, and other organizations.

To accomplish its goal, the GSC established 29 commissions, 6 working committees, 7 regional divisions, 3 research/working groups, 31 local societies, and 13 journals. The society organizes academic conferences, scientific activities and awards, and educational programs to disseminate geographic knowledge. The society also works regularly with government and its international associates to promote the application of geography science.

Journals or major publication series

*Journal of Geographical Sciences*. www.geogsci.com

Current activities or projects

The GSC annually hosts international and national academic conferences and scientific events as well as publish public academic journals (http://www.gsc.org.cn/xh/xhjj//bjcb.htm) to promote academic exchange; awards outstanding
CHINA

scholars, young researchers, and educators (http://www.gsc.org.cn/xh/xhjj//bzjl.htm); support geographers in international geography affairs; undertake the functions commissioned by government; assist and work consistently with the International Geographical Union (IGU), International Association of Geodesy (IAG), and the Imagery Product Archive (IPA).


**Brief history**

The GSC is one of the first academic societies established in China. Its predecessor was the Chinese Geographical Society which was formed by Zhang Xiangwen in Tianjin in 1909. In 1934, Zhu Kezhen, a renowned scientist, founded the Geographical Society of China in Nanjing. After the formation of the People’s Republic of China, the two organizations were merged into the current Geographical Society of China in 1950. Since then, the GSC has grown rapidly, joined the IGU, IPA, and IAG and gradually built a strong relationship with geographical organizations in France, Japan, Korea, and the United States of America.

The GSC has been working to unite all geographers and those interested in geography to improve the development and application of geographical science and technology in China. This has involved popularizing geographic knowledge for the public and improving the geography education and training conditions, working on government projects, communicating and collaborating with 31 local geography societies in China, and cooperating with other academic disciplines. The GSC council is elected democratically and a standing council is in charge of the GSC’s routine operations. The GSC has been working for over a century to build the foundation needed for scholars, researchers, students and its members to engage in academic exchange and pursue academic freedom in geographical research.

Submitted by Zhang Guoyou
China: Hong Kong Geographical Association (HKGA)

Founded: 1969  
Location of headquarters: Hong Kong  
Website: www.hkga.org  
Membership: 150  
Chairperson: Xu Jiang  
Contact: jiangxu@cuhk.edu.hk

Description and purpose

HKGA was founded with the aims of promoting interest in, stimulating teaching of, and cultivating research about geography. The association actively promotes geographical education and awareness through a wide range of activities including the Hong Kong Geography Day, field trips, seminars, workshops, orientations, and subject-related competitions. It serves as a platform bridging secondary and tertiary geography education, offering professional support and advice for both teachers and students. The association has been working closely with government departments on curriculum development and an assessment scheme.

HKGA also publishes papers, newsletters, and journals.

Journals or major publication series

Asian Geographer. www.hkga.org/site/resources/asian-geographer  
Hong Kong Geographer. www.hkga.org/site/resources/hong-kong-geographer

Current activities or projects

HKGA has initiated many activities to promote geography education in Hong Kong and to connect the local community with the wider geographical world. The Hong Kong Geography Day is hosted biennially in collaboration with local universities to foster academic exchange and promote geography education. The Hong Kong Geography Olympiad is an annual open competition for all secondary school students. Latest development in the discipline is constantly shared in its Annual General Meeting and Seminar. HKGA also organizes field trips to improve understanding of local and regional geographical issues. Recently, it has held a series of workshops and discussions on Hong Kong’s new geography curriculum.

Brief history

In 1969, a group of local teachers gathered and established HKGA to promote geography education. In 1981, the association initiated the first Hong Kong Geography Day for the development, popularization, and application of geography. To articulate quality education and professional studies in the subject, HKGA has published the Hong Kong Geographer and the Asian Geographer since 1982 and 1991 respectively.

HKGA has members from local universities, secondary schools, NGOs, and other organizations. It is led by an executive council which consists of the officers and honorary representatives from local education institutes. HKGA is an advisory management committee member of the Hong Kong Teachers’ Centre. In past years, it has been working on curriculum development and institutional changes, especially for the

The association has also been a dedicated player in the international arena. It is a member of the International Geographical Union, and has actively participated in events such as the International Geographical Congress and the International Geography Olympiad. The association itself has also hosted many workshops and activities in collaboration with local and overseas institutes.

Submitted by Xu Jiang
China: Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences

Founded: 1940
Location of headquarters: Beijing
Website: http://english.igsnrr.cas.cn
Membership: 596 (as of December 2012)
President: Quansheng Ge
Contact: geqs@igsnrr.ac.cn

Description and purpose

To serve as the leading, strategic scientific organization focused on leading and supporting the regional sustainable development of resources and environment in China, solving major natural resource and environmental problems related to long-term national development, and improving our scientific capabilities for independent innovation and sustainable development.

Journals or major publication series

Acta Geographica Sinica. www.geog.com.cn
Journal of Natural Resource. www.jnr.ac.cn

Current activities or projects

The institute focuses on physical geography and global change, human geography and regional development, natural resources and environment, geographical information system and surface simulation, terrestrial water cycle and water resources, ecosystem network observation and modeling, Chinese agricultural policy, and other related topics. It currently supports four key laboratories and three field observation stations. It supports cooperative research with corresponding academic institutions in more than fifty countries and regions. It contributes to international scientific programs through twelve branch organizations and nine joint research centers.

Brief history

For the past 70 years, our institute has been leading geographical research in China in the fields of rational use of natural resources, ecological and environmental protection, comprehensive land consolidation and management, regional sustainable development, and resources and environmental information systems. Many of these research projects have been of considerable significance, such as the integrated research on regional differentiation of China’s natural environment, the comprehensive management and development of land in medium and low yield fields in Huang-Huai-Hai Plain, the study on the uplift of the Qinghai-Tibet Plateau and its effects on natural environment and human activities, the compilation of The National Physical Atlas of China, the theoretical research and practical application of regional spatial exploitation, and setting up of a China ecosystem network. Since 1978, these outstanding scientific outputs have won 248 science and technology awards, of which 43 being at national level.

Submitted by Yi Liu
China: Royal Geographical Society – Hong Kong (RGS-HK)

Founded: 1995
Location of headquarters: Hong Kong
Website: http://www.rgshk.org.hk/public/general.php?pageId=87
Membership: 2000 (as of 2014)
Director: Rupert McCowan
Contact: director@rgshk.org.hk

Description and purpose

The Royal Geographical Society – Hong Kong is the Hong Kong branch of the Royal Geographical Society (with IBG). It provides a forum where members can regularly meet and listen to leading local and international speakers from the world of geography and related sciences, exploration, travel, research, the environment, and conservation.

The RGS-HK accomplishes its aims through holding some 65 lectures each year for its 2000 members and the public. It also runs a schools outreach program, serves as a contact point for geographical academics interested in the region, and funds geographical research.

Current activities or projects

The RGS-HK is actively engaged in advancing geography and related sciences in the region. The RGS-HK hosts some 65 lectures a year, provides annual scholarships and awards, publishes an annual review, and holds dinner events including an annual dinner and a gala dinner. It also runs a schools outreach program, through which it hosts educational events at schools throughout Hong Kong. During 2013, the Schools Outreach Programme reached over 5000 students with speakers ranging from distinguished academic geographers to explorers such as Sir Ranulph Fiennes.

Brief history

Previous speakers who have spoken at the RGS-HK include the polar explorers Sir Ranulph Fiennes and Robert Swan, mountaineers Sir Chris Bonington and Doug Scott, primate expert Dame Jane Goodall, the botanist Professor David Bellamy, leading environmentalist Sir Crispin Tickell, former space shuttle pilot Dr James van Hoften, moon walker Commander Dave Scott, Hong Kong explorer Wong How Man, round-the-world yachtsmen Sir Robin Knox-Johnston and Sir Chay Blythe, yachtswoman Tracy Edwards, and authors Simon Winchester, William Dalrymple, Paul French, Jan Morris, and Mark Tully.

Since its inception, the RGS-HK has also strived to assist geographical research in the greater China region. In addition to serving as a point of contact for geographical academia interested in the region, the RGS-HK awards multiple scholarships yearly. These include a post-graduate award and several scholarships for undergraduate and master’s degree research.

The RGS-HK’s board meets three or four times a year to review future speakers and decide policy matters. Elections to the board are held once a year at its annual general meeting. The RGS-HK’s chairman is James Riley. The branch is run by a director, Rupert McCowan, and two administrators, Wendy Poon (events and schools outreach administrator) and Frances Chim (membership administrator).

Submitted by Rupert McCowan
One of the most common quantitative thematic map types is the choropleth map, in which magnitudes of area-based statistics are portrayed by fill colors, shades, or patterns as they occur within the boundaries of the unit area (Figure 1). The word choropleth is taken from the Greek words chorós (place or area) and plethein (to fill). The earliest choropleth maps can be traced back to the early 1800s. Choropleth maps became progressively more common throughout the nineteenth and twentieth centuries, first with the increasing prevalence of census and other statistical data, then with the advent of computer mapping, geographic information systems (GIS), and other software that automated most of the cartographic procedures that were previously done using time-consuming manual methods. Web delivery of choropleth maps further accentuated their widespread use and appeal; this trend continues today. Some of these maps also have interactive and/or animated capabilities not available in printed static maps.

Choropleth maps are also popular because the rationale behind a sensibly designed choropleth map is typically well understood by most map-readers. They are able to see the geographical patterns in the mapped statistical values (also called attributes) and are able to compare different areas on the map. When two or more choropleth maps are involved (e.g., unemployment and education rates for a particular region or US population density in 1950 and 2000), the map-reader can compare areas between the maps. When the maps are small in size and have similar structure and context in order to facilitate comparisons between the maps, they are called small multiples (Tufte 2001). The choropleth map can also be portrayed in 3-D, where the enumeration units’ heights are scaled to the data values (Figure 2). In many interactive choropleth maps, the individual attribute values can often be seen by a simple click or mouse-over in the unit area. However, for the reader who wants to find only individual values apart from the spatial distribution, it would be more efficient to consult a table.

**Attribute data**

Effective portrayal of statistical data is of prime concern in thematic mapping. Although choropleth maps can technically be created for any area-related numeric value, cartographers emphasize that some form of derived data (e.g., rates or ratios) should be used in the mapping process. The enumeration unit areas tend to vary in most choropleth mapping situations and the use of derived data provides standardization throughout the map. Ratios can involve areas (e.g., people/mile²) or other ratios independent of area (e.g., per capita income or birth rates).

The attribute data are always tied to some form of enumeration unit, such as countries, provinces, states, counties, census tracts, regions, or other divisions. In Figure 1, the unit is the county. Since all maps are generalizations of reality, it is important to remember that a choropleth map will portray the attribute value uniformly across the enumeration unit. This
implied homogeneity is a common criticism of choropleth maps, especially when using larger enumeration units such as countries. A choropleth map can become less generalized by using smaller enumeration units, although the appropriateness of using smaller units is highly dependent on the map purpose, level of acceptable generalization, map scale, and often, most importantly, data availability. If these levels of generalization are not acceptable, one alternative is to use another thematic map type. For example, in dasymetric mapping, zones of a more uniform statistical value are created that may not necessarily follow enumeration unit boundaries.

Data classification

Attribute values are often grouped into classes in order to simplify the pattern of distribution for the map-reader (Figure 1). This type of generalization is called data classification, a process that is very common in choropleth mapping. Cartographic convention has been to use anywhere from four to seven classes as a possible starting point for choropleth maps, but the number of observations, consideration of possible data outliers, the color choices, the map medium, and the purpose of the map can all influence the number of classes. A greater number of classes reduces each class range and, thus, lessens the amount of generalization in the map, but the increased complexity may make interpreting the map more difficult. If a choropleth map does not undergo data classification, it is considered to be an unclassed choropleth map, in which there is a unique color or symbol for each observation in the attribute data (discussed below).

There are a number of different data classification methods that may be employed when classifying data for choropleth maps. Natural break methods, either by visual inspection of the

---

**Figure 1** Simple choropleth map using a sequential color scheme adapted from ColorBrewer 2.0, the Jenks natural breaks classification method, and a legend that includes four classes and an outlier class in which the outliers are labeled. See text for discussion. Source: Dent, Torguson, and Hodler 2009. Reproduced with permission of McGraw-Hill.

**Figure 2** A 3-D grayscale data visualization of choropleth map data. In this figure it is the percentage of population age 65 or older by county for the year 2000. These visualizations are also called prism maps or block diagrams. Source: Dent, Torguson, and Hodler 2009. Reproduced with permission of McGraw-Hill.
data or by the use of a computer algorithm such as Jenks Optimal, group data so that all values within a class are as alike as possible (maximizing class homogeneity), and class breaks occur where there are larger gaps in the data. Natural break classification methods produce visual results that a map-reader will typically expect – that enumeration units with like symbolization will have greater similarity in attribute value than enumeration units with different symbolizations. Note that this method typically produces varying class intervals and numbers of observations in each class.

Equal interval and equal frequency (also called equal step and quantiles, respectively) are two of the more traditional classification methods used in choropleth mapping. Equal interval classification groups the data so that class ranges are all the same. This method is often used for multiple choropleth maps that use the same legend, as well as animated choropleth maps, where enumeration units may change classes and their subsequent symbolization throughout the animation. Some cartographers also feel that the equivalent class ranges make the map legend easier to read. Equal frequency classification, which generates an equal number of observations in each class, is useful for statistical tests between the classes that require the same number of observations, as well as for comparing quantile classes between different choropleth maps. Note that equal interval and equal frequency methods typically fare poorly when quantitative assessment measures for class homogeneity are applied (Dent, Torguson, and Hodler 2009).

Other classification methods used in choropleth mapping include arithmetic and geometric classification, which are sometimes used for data with skewed distributions. For example, class ranges will become progressively larger to accommodate attribute data that are increasing or even increasing at an accelerating rate. When the statistical data have a normal distribution, many demographers and other social scientists employ the mean and standard deviations to determine the class breaks. In this case, class breaks fall at standard deviations above and below the mean value in the attribute data. Still other cartographers use defined breaks (user-defined) that match a special purpose. The classification breaks may fall along critical values, such as tax brackets, a poverty line, or other break point that is relevant to the map (Kimerling et al. 2012).

An unclassed choropleth map results when each observation has its own unique area symbol (Tobler 1973). Such a map eliminates the generalization caused by data classification, as each class “range” with a single observation achieves class homogeneity. Unclassed choropleth mapping began receiving academic attention in the 1970s but, until recently, relatively few unclassed maps were created compared to classed choropleth maps. Unclassed choropleth maps are typically more difficult to create than conventional classed maps; many cartographers feel that classification is a necessary part of the map generalization process and leaving the map unclassified gives it an unstructured look. Nonetheless, map-readers can detect major trends in the distribution patterns (Figure 3) and, with the influences of scientific and cartographic visualization, unclassed choropleth maps are both more numerous and, perhaps, less controversial than a few decades ago (Dent, Torguson, and Hodler 2009).

Symbolization

The symbolization used in choropleth maps involves the use of one or more visual variables to portray areas of higher and lower attribute values for each enumeration unit. In black and white (or grayscale) mapping, patterns or shades within each unit create a light-to-dark
Figure 3  Purple America is an unclassed choropleth map in which the percentage of those voting Republican or Democrat for each county or county equivalent is represented by a unique mixture of red and blue. This legend illustrates the use of a bipolar or divergent color scheme (see text for discussion). Note that visualizations such as this may or may not have an associated legend. The point is simply to get a feel for the overall data distribution without getting bogged down with numbers. Source: Robert J. Vanderbei. Reproduced by permission. http://creativecommons.org/licenses/by-sa/4.0/legalcode.

progression suggesting lower and higher attribute values. For pattern fills, units with the highest attribute values will have the most dense texture or pattern. With solid shade fills, the highest values (or percentage black) will have darker tones. Solid shade fills are more common today than pattern fills. Black and white choropleth mapping is still common where higher printing costs associated with color maps are an issue; it also works well for readers with color deficiencies.

Choropleth maps today are more likely than ever to be produced in color, largely because they are often produced as virtual, on-screen maps. For printed color maps, the cost disparity between color and black and white printing, while still significant, is much less than it was a few decades ago. Color also affords the cartographer many more design options, but those options make design multifaceted, as there are over 16 million colors to choose from on many platforms today. The designer has to choose wisely in order to make an effective choropleth map. In color mapping, darker hue, hue values, and/or more saturated colors indicate higher
attribute values and vice versa. The resulting sequence is often referred to as a color ramp. Sometimes the color ramp values are along one hue (e.g., light red to dark red) and sometimes multiple hues are used (e.g., light yellow to medium green to dark blue).

While most GIS and mapping software today provide quite a number of options for various color ramps that visually suggest “more” and “less” in the data, web applications such as ColorBrewer 2.0 (Brewer 2014) are commonly used to assist cartographers in making better color choices for their choropleth maps. ColorBrewer 2.0 is a guide that allows the user to select the number of classes, the type of color ramp, and the specific color scheme. A map shows what the result looks like, and will also indicate if the map will work well for individuals with color deficiencies, if it will photocopy well, and if it is appropriate for liquid crystal display (LCD) and/or print formats. Additionally, it will export the color values to a number of formats, including those appropriate for GIS, artistic drawing programs, and web programming platforms.

Sometimes statistical attribute data do not simply progress from a lowest to highest value. Occasionally, data diverge in two directions from a central value or a central class range. A bipolar map results when a divergent color ramp is applied to this type of data (Figure 3). A few examples include percentage increase/decrease, profit/loss, or above/below average (Dent, Torguson, and Hodler 2009). If the data for a bipolar map are distributed normally, the mean and standard deviation classification scheme is sometimes employed (see discussion above).

Bivariate and multivariate choropleth maps show two or more related variables on the map through symbol variations in color and/or pattern fills. For example, median age and per capita income can be compared by incorporating a 2-D color ramp to create a bivariate map (Figure 4). Although research in multivariate choropleth mapping with three or more variables has occurred (Carr, White, and MacEachren 2005), cartographers must employ quite a bit of ingenuity to create these multivariate maps, perhaps by using scripting, since they are not directly supported in most GIS and mapping software. Some cartographers will find it easier to simply juxtapose multiple choropleth maps side by side for comparison.

Bivariate and multivariate data can also be mapped by using the choropleth map as a base, and then superimposing a dot-density layer, a proportional symbol layer, or other thematic map type on top of the choropleth map. While this approach is easily done in most GIS and mapping software, some map designers caution that it is easy to visually overload such a composite map with too much data (Dent, Torguson, and Hodler 2009).

**Legends and other design issues**

The legend on most choropleth maps serves as a referent or visual anchor between the map symbols and the legend. This is a supporting role, so the legend is typically lower on a visual hierarchy (it is typically smaller than the map, for example). In Figure 1, four class ranges represent most of the counties; a fifth class has the outliers labeled. On some choropleth map legends, the class ranges are depicted in a noncontinuous fashion. That is, they represent only the minimum and maximum values for each class. Noncontinuous reporting is useful for narrowing the class ranges so the reader more accurately interprets the data in each class. The continuous style (depicted in Figure 1) has no gaps or overlaps in the class ranges and is a more traditional presentation of a choropleth map legend. The continuous style is also often applied to animated or sequential map series, in
which only one legend is used and data values can drop in and out of the different class ranges throughout the series or animation.

Legend titles are typically reflective of the attribute data in the choropleth map and will be consistent with the title or figure caption. For example, if the data is population density or poverty percentages, the legend title will be something like “People per square mile” or “Percentage” as opposed to “Legend,” “POP10_DEN,” or “%.”

Another important aspect of choropleth map design is the choice of map projection. All maps contain projection distortion, especially at smaller scales. Since the distribution of data is being compared throughout the map or between a set of maps, most cartographers tend to favor equal area (or equivalence) projections for thematic maps, including choropleth maps. For example, Figure 4 was created using an Albers equal area projection. Although other map projection properties (conformal, distance, direction) or minimum error/compromise projection (balancing distortion properties) are also used in choropleth mapping, the equivalence property is typically the most important.
As with most maps, choropleth maps have very subjective aspects to their layout and design. For example, some choropleth maps also include general reference features such as cities, roads, rivers, and the like. Some designers like to keep choropleth maps simple, arguing that the inclusion of these features causes clutter and, if taken to an extreme, may take away from the main point of the map. Others feel that selective use of reference features can provide important contextual information and enhance the value of the choropleth map. For example, a few large cities portrayed on a choropleth map may help with the interpretation of patterns such as population densities or crime rates (Figure 5).

Another example of subjectivity in choropleth map design can be found in what is included in a choropleth map layout with regard to the area surrounding the map body. Traditional

**Figure 5** The addition of some reference information can give additional context to the choropleth map. Source: Dent, Torguson, and Hodler 2009. Reproduced with permission of McGraw-Hill.
cartographers often insist that all maps must have a certain set of basic map elements, such as a scale bar, north arrow, title, legend, and so on. Designers favoring a data visualization approach to cartography often eschew most, and sometimes all, of these basic elements to allow the viewer to see the data in an as unfiltered manner as possible. Most current textbooks on thematic cartography tend to take a nuanced view in that what is required on a map may change according to the map purpose, the intended audience, the scale of the map, and the final output for the map.

The future of choropleth mapping: mashups and more technology

As noted already, web delivery of choropleth and other thematic map types is on the increase. One increasingly popular method of displaying choropleth maps is to export the map layout to a format such as KML (keyhole markup language), in order to display the map in popular web applications such as Google Earth, Google Maps, or other readily available web mapping platforms. This approach is often referred to as

Figure 6  Choropleth map mashup with Google Maps. Average gasoline prices are depicted for each county. Used by permission. Source: GasBuddy.com, http://www.gasbuddy.com/gb_gastemperaturemap.aspx.
a mashup, as the choropleth map is displayed on top of the existing map, satellite, or aerial photography provided by the web application in a single view (Figure 6). The advantage to this method is that the map-reader then has all of the benefits of using the platform’s capabilities, such as zooming, panning, and so on, as well as a familiar interface. With most choropleth map mashups, the user is also able to click on individual enumeration units to see the name, individual value, and other attributes.

At the time of writing, one possible disadvantage to this technique is that the base map on which the choropleth map is displayed uses the web Mercator projection. The conformal properties of this projection ensure that roads that intersect at 90° angles in the real world also have right angles on the map. While a necessity for traffic-routing and navigation purposes, the Mercator projection has one of the highest amounts of distortion for mid-to-high latitude small-scale maps. This distortion is most noticeable on world maps (Greenland appears as large as Africa) and, as such, is typically not recommended by most designers for choropleth or other thematic mapping activities.

Technological advances are happening at an increasing rate. While predicting technological change is difficult, support for choropleth maps in GIS software and mapping programs will, nonetheless, continue to be nearly universal, as is the case at present. Increasingly powerful computers, faster Internet speeds, cloud computing and data storage, a wider range of platforms for map display, and futuristic human–map interfaces will all likely play a role in future choropleth map developments.

SEE ALSO: Cartographic design; Color theory; Dasymetric mapping; Map projections and coordinate systems; Mapping mashups; Visualization

References


Further reading


Cities and development

Garth A. Myers  
Trinity College, USA

Questions about the relationships between cities and development have echoed through the last 50 years of both urban geography and development geography. Some take the relationship as a given: cities are development, cities bring development, development leads to more urbanization, and the cycle repeats. Others take nearly the opposite view: cities are parasites that suck (under-)development out of rural people, and cities destroy human development. The realities in the world fall somewhere in between, or combine some elements of each side’s claim. Cities do manifest development, but always at a cost to other places and people, and always unevenly. It is perhaps one of humanity’s greatest current and future challenges to find the ways to ensure that urbanization and both economic and social development lead to the betterment of the existence of all – or at least a great majority – of the world’s ever-increasing urban population.

It is useful to note that there are no secure agreements across the world on how to define either cities or development. Both official and scholarly definitions of what a city is vary considerably across the world, despite general agreement on the basic idea that a city is “a relatively large, dense and permanent human settlement” (Beall and Fox 2009, 7). Beyond that general understanding, though, countries use different population parameters for delineating the threshold of cityhood, to say nothing of which areas are, or are not, included in a metropolitan area. Development, too, is understood very differently around the globe beyond a generally shared basic notion of progress. Great variation exists in the relative weight given to economic versus social indicators in measurements of development (for example, a typical measure of the former is the gross domestic product (GDP) per capita of a city, while the latter might use life expectancy or adult literacy rates as proxy measures), and data vary tremendously in availability and quality across the world’s cities for making comparisons.

Another fundamental problem lies with the fact that for a long time the study of cities and the study of development existed as essentially separate realms of geography, divided by the region of the world in which scholars conducted research. Urban geography essentially belonged to cities of what is variously termed the First World, the Developed World, the More Developed Countries, or the Global North. By contrast, scholars studying cities in what is called the Third World, Developing World, Less Developed Countries, or the Global South were presumed to be studying development, since cities in the Global South were not quite developed as proper (Western) cities (Robinson 2006). To some extent, that sort of dichotomization of previous scholarship may be overdrawn, and in any case it has changed considerably in the last decade, during which a mounting critique has challenged universal understandings of urbanism that emanate from European and North American cities. This critique has energized a reconsideration of the need to construct global hierarchies of cities, the criteria for positioning on such hierarchies, and the desirability of policymaking aimed at boosting a city’s rankings on them (Myers 2014). The critique has also
coincided with a prolonged global economic crisis which has unsettled common understandings of or policy prescriptions for development (a palpable manifestation of this confounding state of things came to light, for example, when Portugal, nominally a country of the Global North, sought aid from Angola, its former colony in the Global South, amidst its severe financial crisis in 2011). While urban geography scholarship continues to be divided by world regions and by ideological economistic categorizations into Developed/Developing World, First/Third World, or Global North/South, there is now a substantial countertrend in the literature, matching the doubts which now surround the mainstream developed-world models for development in the developing world.

Many scholars of Global South urbanism now demand greater attention for the urban areas in their midst in the development of theories about cities, potentially to the point of upending the privileged place of Global North cities in these theories (Roy 2009). But there is also a growing arena of research exploring common ground and reciprocal or even reverse idea flows between Global North and Global South urban contexts, South–South comparativism, and growing networks of learning and policymaking across the Global South. There is an increasing amount of exploration of interests held in common between those working on Global North and Global South cities, leaving behind biases and assumptions of privilege in an effort to focus on emergent patterns and processes specific to different cities or city-regions.

As the divides have begun to break down, what is emerging is not necessarily a new universality – though that claim has its proponents, whether in theories of planetary urbanization or of global city-regions – but a greater appreciation for how varied the processes of urbanization are in different parts of the world, and even within one country’s cities. Instead of force-feeding cities into one lens on their development, it increasingly makes sense to speak of themes held somewhat in common but with comparative variation in urban development worldwide. The examination of seven of these themes below (urbanization itself; economic versus social development; poverty and inequality; housing; urban governance and planning; urban services and infrastructure; and urban environments) briefly highlights the commonalities and the differentiations across global urban geography. If we start from the very basic questions of what scholars study when they study both cities and development at once, then it is not difficult to center in on these themes. Since urban geography is quite vast, it is impossible to encompass the entirety of relevant work within these themes (let alone to account for other genuinely valid themes, such as crime and violence, megacities, or the future of urban design), but they definitely provide a broad map of work at the intersection of cities with development.

**Urbanization itself**

During the last decade, the proportion of people living in cities passed 50% of the world’s population. During the nineteenth and early twentieth centuries, Europe, North America, and selected countries in Latin America and Asia rapidly urbanized. Generally speaking, for much of the last half-century, though, the predominant theme has been the rapid population increase of cities in Asia, Latin America, and, increasingly, Africa. It follows that we have had several decades now of contrasting rates of growth for urban areas in the Global North and South. Many Asian and African countries, in particular, have experienced a half-century of rapid urbanization coterminous with a substantial slowing – or
even reversal – of urbanward population flows in much of Europe and North America.

What is more, the character and geographic patterns of land use for metropolitan growth have varied considerably. In some parts of the world, urban growth corridors, extended metropolitan regions, or global city-regions are highlighted, but even within these conceptualizations substantial variation exists. Where an earlier generation of urban geographers generated human-ecological models of urban land use out of development processes in cities like Chicago and applied them universally, it is more common now for scholars to seek less universalizing models given the great variability and fluctuation of the last few decades in urbanization. To wit, many of the world’s most rapidly urbanizing countries of the past half-century have been in sub-Saharan Africa, and that region’s rapid urbanization rates provide urban geographers with many questions which provoke caution for grand theorization. In some countries (e.g., Botswana or Ghana), rapid rates of urbanization appear to correspond well with high rates of economic growth, but in other countries with rapidly growing cities this urban growth has occurred during a period of economic stagnation or even amidst an economic downturn. Substantial debates have surrounded the actual data on urbanization, such that some scholars make persuasive claims for a region-wide slowing of urbanization, with even a few reverse migration trends, wherein there is no one single correspondence with overall development trends (Potts 2009).

Economic development versus social development

This debate over the relationship between urbanization and development carries over into a debate over how to characterize development. The role of cities in fostering economic development is a strong theme of longstanding centrality in urban geography research, but a parallel theme – exploring connections between cities and sociocultural development – has risen to the fore in recent decades. So on one hand, urban geographers have shown the deep connections which exist between urban growth and economic prosperity. Relatedly, as cities grow, their significance to sociocultural development in their respective societies expands. On the other hand, urban geography research has long debated the socioeconomic inequalities attendant with urbanization.

Indeed, urban geographers with deeply varied ideological commitments have focused their research concerns on the bifurcated processes and outcomes of urban development, spatially and socially. As United Nations Secretary-General Ban Ki-moon put it in his preface to the UN Habitat’s (2013, iii) *State of the World’s Cities* report, “cities can be prime driving forces of development and innovation. Yet the prosperity generated by cities has not been equally shared, and a sizeable proportion of the urban population remains without access to the benefits that cities produce.” This general understanding then leaves room for considerable discussion of urban poverty, inequality, and injustice within the realm of “cities-and-development” studies.

Poverty and inequality

Following from the concerns with the social down side to urbanization processes, research has identified the rise of inequality in many cities, in both the Global North and the Global South. The growth of the urban middle class has run parallel to persistent poverty. Gini coefficients of income inequality are fairly consistently higher
for urban areas in comparison to rural areas in most countries. Even though urban incomes are broadly twice as high as rural incomes across the world, levels of urban poverty are often shockingly high, even in the world’s wealthiest and most urban societies. Hartford, Connecticut, which a 2012 Brookings Institution study identified as the richest metropolitan area on the planet, contains at its core the city of Hartford itself, ranked as the fourth-poorest city in the United States, with a regionally high Gini coefficient of income inequality (Chen and Bacon 2013). The nearby metropolitan area of Fairfield-Bridgeport, Connecticut, is simultaneously the heart of the region’s Gold Coast bedroom communities of millionaires and billionaires, and home to the worst Gini coefficient of income inequality of any US city. Similarly, cities such as Johannesburg, South Africa, Windhoek, Namibia, and Lusaka, Zambia, contain among the world’s highest levels of income inequality, and yet each contributes a far higher proportion to its nation’s GDP than its population would warrant on a per capita basis (UN Habitat 2014). The fragmentations and polarizations attendant with urbanization in the twenty-first century demand attention from both policymakers and urban activists or advocates for social justice (Mitlin and Satterthwaite 2012). The inequality tangibly evident in urban housing and housing rights provides a visual and visceral reminder of this significant challenge for contemporary cities around the world.

Housing

As the UN Habitat (2013) report indicates, making the prosperity that cities generate available for the benefit of a larger proportion of the urban populace is now a major global urban development policy focus. The struggle to make housing and residential development more equitable has become a central feature of these policies. Housing has been a crucial research theme of urban geography scholarship across the globe for more than 50 years. Global North urban scholarship on housing development has long concentrated on the contrasting housing markets of many cities, made more intense with globalization. One major contemporary theme concerns the transformation of public housing under various forms of governance. Another focus has concerned the housing dynamics of an increasing population inversion whereby, in the United States for instance, suburbs (especially inner-ring suburbs) are becoming increasingly diverse while declining economically as city centers slowly begin to rebound economically through increasing demand for upper-income residences (Ehrenhalt 2012). These US patterns parallel patterns in some European cities, but there are also marked distinctions in the transformations of inner-city and suburban housing in London, Paris, and Rome, for instance, such as in the quite varied impact of changes in peri-urban public housing projects (discussed in the Parisian context as banlieue suburbs) or rather different dimensions of residential squatting.

The contradictory expansion of both gated communities and informal settlements (often appearing as or described as slums) is a major concern for scholars of Latin American, African, and South- and Southeast Asian cities. Migration dynamics provide for both expanded transnational housing development strategies and dramatic changes in household dynamics. The gentrification of poor and working-class areas in cities, long debated in Global North contexts, has now increasingly become a theme in Global South scholarship, with the notable twist that many researchers challenge the applicability of the debate’s Northern terms within Southern cities. Still more Global South scholarship has
highlighted the contradictions in the progress made in reducing slum housing conditions while overall slum populations continue to grow and elite housing expands apace (Huchzermeyer 2011). In some notable cases in particular such as Brazil, claims are made for approaching goals set in the Millennium Development Goals for slum eradication, and yet favela (informal settlement) populations have increased. This debate on how to measure or gauge progress in slum eradication in turn generates broader debates about urban policy in terms of governance and planning.

**Urban governance and planning**

Neoliberalism, generally speaking, entails support for free trade and free markets, and limitations on the role of governments in the economy. In urban contexts, neoliberalism is manifested in municipal governance that is heavily reliant upon private sector involvement and public–private partnerships. Urban geographers have made a broad distinction between rollback and rollout neoliberalism – where the former entails elimination of urban governmental units or agencies as part of a privatization of urban management, and the latter entails the creation of new governmental units or agencies whose purpose is to facilitate capitalism (Peck and Tickell 2007). During the 1980s and 1990s, urban policies and development policies around the world came under what came to be called the Washington Consensus, the tacit agreement among the major development donor states (the Organisation for Economic Co-operation and Development (OECD) member countries, mainly) that neoliberal strategies were essentially the only path for development. During the 2000s, and particularly following the prolonged global financial crisis that began in 2007, the Washington Consensus began to crumble, leading to substantial reconsideration of neoliberalism in the governance of many countries and cities around the world (Robinson 2011).

Debates concerning neoliberal governance have transcended geographic boundaries. The expansion of neoliberal urban governance tactics and strategies into the cities of the Global South from the 1980s onward, alongside the spread of professional urban planning practice heavily influenced by the West, has shaped much of the discussion of the roles of states in urban development. Scholars have engaged in increasing debate on what the most appropriate governance arrangements and policy approaches can be for making cities prosperous and inclusive at the same time as they forge rights to the city for all (Brenner, Marcuse, and Mayer 2012). Proponents of neoliberal urban governance are rather rare in scholarly research despite the ubiquitous turn earlier on toward various forms of neoliberal planning in urban practice and in the policy realm – and the continuing significance of neoliberal thinking in OECD development agencies and the international financial institutions (the World Bank and the International Monetary Fund).

The increasing global significance of BRICS countries (Brazil, Russia, India, China, and South Africa) as both sites of and investors in urban development has brought increasing attention to their governance and planning approaches, particularly in terms of their potential relevance, or lack thereof, in cities of the developing world. In many BRICS cities, strict adherence to neoliberal principles is rare. In Brazil, and in many Latin American cities more broadly, municipal and national governments have actually played an increasing, and often populist or progressive, role in urban management in the last decade or so. China’s unprecedented urbanization has similarly involved a substantial role for local, regional, and national state
institutions in urban governance, planning, and management.

This flows inevitably into questions surrounding the relationships between democracy and urban development. Neoliberalism’s advocacy for increasing private sector involvement in urban governance is inconsistent in regards to democracy, with highly uneven degrees of access to citizenship or to rights to the city. Cities around the world have been, to some extent, set aflame by rebellions often based around exactly this disconnect (Harvey 2012). The United Nations Sustainable Cities Program and UN Habitat as a whole brought forward an agenda for urban planning and management in the late 1990s, strongly influenced by what was then the World Bank’s Cities Alliance, advocating “Environmental Planning and Management” (EPM) in cities around the world as a means of fostering inclusiveness in planning. Yet the UN itself on numerous occasions championed fundamentally undemocratic planning practices and processes as supposed best practices of this EPM model. The reasons given for lauding these specific examples often emerged in their efficiency, particularly in service delivery, at the direct expense of popular participation and input (Pieterse 2008).

Urban services and infrastructure

The most tangible governance and planning questions in cities concern the provision of municipal services and infrastructure – for water, solid waste, electricity, transportation, and education, among other factors. The drive for the privatization of service provision, under both rollout and rollback neoliberalism, has stalled in the past decade. From Detroit and New Orleans, to Cape Town and Shanghai, and on to London or Paris, cities are in search of the most effective and efficient balance between public, private, and popular sectors for service and infrastructure provision. No one single approach has proven most effective, despite the claims of various partisans. For every case study of privatization successes in urban service provision, one can find counterexamples of failed engagements of the private sector; similarly, there is quite a range of outcomes for state-led service provision. In some cities, public–private or public–private–popular sector partnerships have improved water supply, solid waste management, or public transportation, but scholars have likewise documented shortcomings for many of these varied partnerships.

The economic rise of China in the past two decades has further transformed discussions about urban service provision in the world’s cities, particularly in terms of infrastructure construction. Across Africa, Latin America, and Asia, Chinese engineering firms have been active in developing sanitation, airport, seaport, bus, subway, highway, and light rail systems for rapidly growing cities around the Global South. Other Global South – and in some cases, Global North – development partners have expanded their urban infrastructure investments abroad as a result. This then further muddles the policy debate questions surrounding neoliberalism in urban governance. It also often draws attention to the increasingly profound environmental impacts of urbanization.

Urban environments

Urban development has profound environmental consequences. Air pollution, water pollution, and land degradation are among the most immediate of these consequences. The more rapidly that societies urbanize, the greater the potential for negative externalities in the environment. The above discussion of urban services and
infrastructure has suggested connections to the environment which can now be made more explicit: the lack of waste and sanitation services, or lack of urban management of air, water, and land, can have and has had direct negative consequences in many cities for the health and wellbeing of humans and other beings in the urban setting. From the more quotidian concerns with these externalities, recent scholarship has turned to analysis of global climate change and its implications for cities. Environmental disasters, such as the 2005 and 2011 tsunami events which devastated many eastern Indian Ocean Rim communities and then eastern Japanese cities, respectively, or the recent devastating earthquakes in Chile, New Zealand, and China, have remained a crucial focus of research in urban geography and urban development studies, but often with the added component of the climate change dynamics potentially linked with these mega-events.

Urban geographers have brought vital attention to the politics of urban environments in urban development processes. Urban political ecology has been influential in drawing a political focus on urban metabolisms into the frame in a manner which extends past concern for mere nature and environment (Heynen, Kaika, and Swyngedouw 2006). At the same time, urban political ecology has been challenged over questions of the universality of its language and applicability of its Global North-derived concepts in cities of the Global South (Loftus 2012). This critique extends to questioning the literature of sustainable urban development, which likewise often fails to account for Global South urban dynamics. The Sustainable Urban Development Reader (Wheeler and Beatley 2009), for instance, contains very few readings on Global South cities, and none on African cases. Regardless of these shortcomings, urban environments are sure to be central to research at the interface of cities and development for the foreseeable future.

SEE ALSO: Development; Economic geography; Environment and urbanization; Global cities; Rural/urban divide; Urban ecology; Urban geography; Urban planning: human dynamics; Urban political ecology; World cities

References

Myers, Garth. 2014. “From Expected to Unexpected Comparisons: Changing the Flows of Ideas about


Citizenship

Lynn A. Staeheli
Durham University, UK

Citizenship is a complex, contested, and even confusing concept. It is most commonly defined as a legal standing that connotes membership in a national polity, the guarantee of rights, and consent to obligations associated with membership. This definition is insufficient, however, as citizenship is most often acquired passively by virtue of being born in the territory of a country (jus soli) or to a relative who is a citizen (jus sanguis); typically, only those who go through a process of naturalization or who renounce citizenship take a conscious choice to accept the status, rights, and obligations of citizens. Furthermore, the concept is freighted with normative expectations that may be hidden by defining it simply as a legal status.

Legal and substantive citizenship

Many philosophers distinguish between the legal or procedural aspects of citizenship and its substantive aspects. Discussions of the legal aspects include the mechanisms by which the formal standing of citizenship is conferred, the specific rights and protections accorded to citizens, and the obligations or responsibilities that citizens accept. In liberal political theory, the legal aspects of citizenship are generally only relevant to individuals. Individuals, rather than social groups or places, are generally the holders of rights and responsibilities, although the means by which citizenship is extended may involve an element of group struggle, such as in women’s suffrage movements in the nineteenth and twentieth centuries in the Global North. While the legal standing as citizen has always been exclusionary in some way, the narrative of the history of citizenship has been one of increasingly greater inclusivity, as citizenship has been extended from a few landed elites under a feudal system, to property-owning males, to most males (or at least in the Global North, to white males), to women, to immigrants, and so on. In this way, the legal standing of citizenship is often described as universal, or open to all, even if not everyone living in a country holds the legal citizenship of that country.

While the legal standing of citizens is important, liberal and republican theories of citizenship draw attention to different elements of it. From the perspective of liberal theory, citizenship entails a set of protections from a state (or sovereign) through law and legal standing. Such protections often include freedoms of speech and assembly, freedom from unreasonable restraints of liberty, and the rights to fair and equal treatment by the state. From the perspective of republican theory, by contrast, participation in governing is crucial, and legal supports – and requirements – for participation are highlighted. These supports might include trial by jury, systems of deliberation and voting, and other forms of participation. While these two perspectives are sometimes put in opposition, they can be seen as reinforcing one another, as protection of basic freedoms from the state may be necessary for citizens to participate in self-governance. Both perspectives, however, recognize the centrality of legal standing and legal systems to citizenship.
CITIZENSHIP

Even in focusing solely on the legal aspects of citizenship, its spatiality is evident in at least four ways. First, in many countries with a federal structure, individuals may simultaneously be citizens of multiple jurisdictions, ranging from a municipality to provinces, cantons, or states, to the national government. To some extent, then, the rights and responsibilities held by individual citizens are qualified by the geography of the countries in which they live. Second, countries that have been partitioned – often through war and subsequent “peace processes” – sometimes embed complex and contested citizenship identities associated with internal territorial division. In the case of Bosnia-Herzegovina, for instance, the country is divided into two “entities,” the Republika Srpska and the Federation of Bosnia and Herzegovina. Leaders of the entities disagree whether national or entity-level citizenship should be primary, and the persistence of these divisions reflects the continuing claims of Serbia and Croatia for territory within the country. The unsettled questions of where or of what one is a citizen are literally mapped in the geographies of citizenship in these contexts. Third, and since at least the middle of the twentieth century, with the Universal Declaration of Human Rights in 1948 and the subsequent development of supranational and international rights regimes, the ways in which citizenship is affected by legal structures at scales other than that of the nation-state have been a focus of many academic texts that posit the possibilities for, and potential of, a global citizenship regime with a legal apparatus – such as the International Court of Justice – to support it. Finally, at a time when many people live in countries other than the ones in which they have legal citizenship and when dual citizenship has greater acceptance, the spatiality of the legal aspects of citizenship becomes ever more complicated.

Citizenship, however, has always implied more than legal standing, and some scholars focus on what are described as the substantive aspects of citizenship. From this perspective, citizenship is better understood as a public standing (Shklar 1991) that is recognized by conferring a legal status labeled as “citizen”; the extension of citizenship to larger portions of the populace reflects a changing view of social groups (e.g., men without property, women, racialized minorities) as being worthy of citizenship, as being able to act as autonomous citizens, and as capable of participating in self-rule, historically at the level of the city. Scholars making this argument recognize the legal aspects of citizenship but argue that the legal standing of citizenship is predicated on social recognition of the political subjectivity and personhood of individuals.

In addition to a more inclusive view of who should be recognized as a citizen, there has also been a fluctuation in what citizenship entails. T.H. Marshall (1950) may be the most prominent exponent of this view. In his essay “Citizenship and Social Class,” Marshall argued that the rights associated with citizenship had expanded in European countries from those associated with civil rights (such as those protecting individual freedoms) to include political rights (such as voting, and the right to stand for office) to social rights (such as the right to education and social welfare support). Significantly, Marshall’s formulation emphasized the responsibilities of the state to its citizens. The recognition of citizenship as predicated on public standing and as encompassing a broader range of rights to support equality are both discussed in terms of the substantive aspects of citizenship.

Attention to the substantive aspects of citizenship considerably broadens its scope to include structural relationships and the positioning of individuals and social groups with respect to the economy, society, and politics, often in
contradictory ways. Poverty, for instance, is not simply a lack of money but also affects the ability to meet obligations such as payment of taxes. If the obligation not to be a burden on others is also seen as a responsibility of citizenship, then reliance on social welfare support – as a social right of citizenship – may also be seen as the failure to be or to act as a responsible citizen, and can be used to justify the withdrawal of social rights. Similarly, behavior and deportment that violate social norms may be used to justify exclusion from at least some benefits of citizenship, or to support claims that groups or individuals are unworthy of citizenship. Thus, debates over who should be recognized as citizens and what kinds of rights should be extended to citizens are complicated by the positioning of individuals and social groups with respect to the economy, society, and laws that extend beyond the formal definitions, the laws that condition citizenship and who will be seen as a citizen include those related to education, welfare, family structure, and more. In this way, citizenship is structured through both law and the relationships of daily life and experience (Staeheli et al. 2013).

The substantive aspects of citizenship have a complex spatiality that extends beyond that noted earlier. The complex legal structures identified earlier are, of course, important. Beyond these, however, citizenship takes on different configurations and meanings around the world, reflecting differing legal structures, economic conditions, political settings, and social relationships and histories. The rights of citizenship, for instance, differ in accordance with the development of the welfare state in different countries, as well as with determinations of who should be able to access them and the ability of the state to meet its own obligations. For instance, housing is a right of citizenship conferred through the South African constitution. The explicit mention of housing as a right is unusual in a national constitution, but it reflects the country’s vision of itself as a modern cosmopolitan country that is committed to human rights in their entirety. The inability of the country to even begin to meet that commitment, however, reflects the poverty and inequality that are enduring legacies of apartheid. It also means that the “right” to housing is rhetorical, and perhaps normative, rather than something that can be accessed in practice.

The geography – or spatiality – of citizenship is, thus, a reflection of historical legacies, social and economic cleavages and relationships, and geopolitical positionings. Sometimes these differences can be dismissed as a mismatch between rhetoric and reality, as in South Africa. Sometimes, differences between the citizenship regimes of countries reflect norms related to what it means to be capable of assuming the responsibilities of citizenship. This point is perhaps most clear with respect to children and young people; countries differ in the ways that children and young people are viewed in terms of their abilities to assume the responsibilities of citizenship and the rights that they are believed to be able to wield effectively. More controversially, differences in citizenship may reflect histories of colonialism, political “development,” “culture,” and geopolitical relationships. Put bluntly, people in some countries are discursively constructed as being incapable of acting as citizens, either because their states have “deskilled” them (e.g., this was the argument about Eastern and Central European countries during postsocialist transitions) or because of Orientalist discourses that construct citizenship as fundamentally modern/Western and therefore somehow incompatible with some cultures or regions (e.g., as is claimed about Middle Eastern countries; see Isin 2002).
CITIZENSHIP

Who can be a citizen?

As noted, legal standing as a citizen is usually conferred through birth, and is sometimes described as a birthright. Yet recognition and the ability to act as a citizen are more complicated than this assumption might imply, and the complication is not simply with respect to the substantive aspects of citizenship. One of the key issues is the perceived ability of individuals to govern themselves and to act autonomously and responsibly. One’s standing as a citizen, therefore, is evaluated in terms of personal characteristics and deportment, which are then compared with normative expectations of how citizens should behave. Failure to meet those expectations rarely results in the formal withdrawal of legal citizenship, but one’s standing as a citizen may still be called into question. In many countries, for instance, convicted criminals often lose the right to vote, in addition to the rights to movement and liberty. More controversially, “workfare” regimes sanction citizens who fail to be self-sufficient (in other words, they rely on the welfare state to meet basic needs), often through the withdrawal of welfare and some of the support that has come to be seen as constituting the social rights of citizenship, as previously noted. Thus, the characteristics and behaviors of individuals can be critical to their being recognized as citizens and to their ability to access rights in a given context.

Perhaps the most important, normative characteristic of citizens in democracies is the ability to act autonomously and rationally so that they can participate in self-governance and self-rule. In this way, citizenship is often linked to democracy, even if the nature of democracy is contested. Indeed, for some theorists, the concept of citizenship is not meaningful in the absence of democracy, and they would describe members of an undemocratic state to be subjects, rather than citizens. Citizens are required to participate in debate, to deliberate, to make decisions as to the public good, and to meet obligations. These presuppose the ability to reason, the ability to govern oneself (not just participate in the governance of others), and a degree of independence so that decisions and opinions are not coerced or subject to undue influence. Any presumed deficiency with respect to these (e.g., as with children who may be seen as irrational and not independent) becomes a reason to withhold citizenship from a social group.

The other characteristic of citizens is that they are recognized as somehow “belonging” to the entity with which citizenship is associated. In practical terms, this is often assumed to be the nation-state, and recognition is conferred through the legal status of citizen. It should be noted that the earliest philosophers did not make that assumption, however. Instead, the association between citizenship and the nation-state arose with the solidification of the Westphalian system, and has been reinforced by the legal recognition that states have the ability to determine who can be a member. Much of the research on citizenship has, therefore, focused on what Bosniak (2006) describes as the two “domains” of citizenship: boundary issues related to who belongs or who is a citizen, and issues related to the ways citizens are to be treated once they are recognized as citizens.

Aligning citizenship with the nation-state may be pragmatic and legally recognized, but the alignment is tested by the large numbers of people who live outside the state in which they hold citizenship, by the growth of dual citizenship, by the fragmentation (or at least layering) of citizenship within federal states, and by people who are not recognized by any state (Bauböck 1994). Of these, stateless people may face the greatest challenges. People become stateless and therefore lack recognition as citizens.
in many ways. Some national groups lack a state and may not be recognized by the states in which they were born or live (e.g., Palestinians in the Occupied Territories or in Lebanon); new national boundaries and laws may result in the loss of citizenship (e.g., the Izbrisani in Slovenia); children born to refugees may be unable to be registered in the countries their parents hold citizenship (e.g., babies born to Syrian refugees in Turkey or Lebanon). Without formal recognition and acceptance anywhere, such people face additional barriers in gaining the legal standing of citizenship from any state, which in turn complicates their ability to access rights and to combat prejudicial treatment against them.

Making citizenship and citizens

When citizenship is understood as a status or standing that can change, attention is directed to the political, economic, and social structures that create or enable those changes, to broad processes of citizenship formation (Marston and Mitchell 2004), as well as to efforts to encourage behaviors and qualities in individual citizens who can act rationally and autonomously in order to support democracy and the public good. Each of these terms — “rational,” “autonomous,” “democracy,” “public good” — is contested and malleable, so there is a politics and agonism surrounding the development of citizenship and citizens (Mouffe 2005).

The structural context of citizenship formation has been described by Sandel (1996) in terms of the political economy of citizenship. This includes the kinds of economic opportunities in a society, the distribution of wealth and well-being, as well as political and social structures that enable the development of a public sphere in which individuals can participate, unencumbered by gender, race, religion, age, or other social differences. In different times and places, these structures are configured in ways that make it possible — and sometimes impossible — for people individually and for members of different social groups to act as citizens. Feminist philosophers, among others, have argued that, while formal standing as a citizen has been gradually extended, inequalities are endemic, and they limit the ability of groups to participate as citizens. Feminists note the formation of “subaltern counter publics” in which the qualities of citizens (or, in their terms, members of the public) are developed in specific ways (Fraser 1990).

Autonomy and rationality might be seen as ideal characteristics of citizens, but at a practical level the kind of rationality required under particular circumstances rarely occurs without concerted effort to foster it; indeed, political and cultural economies militate against the development of citizens who can participate as equals. Furthermore, as economies, social relations, and governing strategies change, new qualities of citizens may be desired. As such, considerable effort is expended in fostering citizens who are capable of acting in ways that seem rational and appropriate given a particular context. As noted, however, there is a politics to the ways that citizens and citizenly characteristics are imagined, encouraged, and trained.

One of the primary sites in which citizens are trained is in public education systems supported by nation-states. Such training has typically focused on imparting the skills necessary to support the state and to build loyalty to and identification with the nation; as noted, the desired qualities of citizens reflect the broader political and cultural economies in which citizens are embedded, so specific elements of the training may differ. Consider two examples. As the United States was consolidated as a state and nation in the late 1700s, public education was intended to create a new kind of citizen who embodied a particular political subjectivity.
CITIZENSHIP

These subjects were to be American (rather than European), hardworking, capable of supporting the economy of the new country, law-abiding, able to make rational judgments in many arenas, and adaptable to new circumstances without guidance from the state. Thus, the public education system emphasized reading, mathematics, animal husbandry, and farming more than it emphasized history, science, or moral philosophy. By contrast, the postapartheid education system in South Africa was designed to train citizens with a slightly different set of skills and characteristics that would allow the country to assume a new place in the world. As in the early American system, history was downplayed but, in this case, it was because attention to past injustice and redress was believed to limit the ability of the country to move forward in a unified fashion. Instead, self-sufficiency, commitments to human rights and justice, and healthy behaviors were emphasized; for a country in the midst of an HIV/AIDS epidemic, the latter was important if citizens were to avoid being a burden on the state or other citizens. In this way, South Africa was little different from other countries emerging from conflict, where agreement on a common narrative for the country and identity can be difficult. The result was a curriculum that seemed to overlook or to forget apartheid and instead emphasized the ways citizens would be treated going forward. Looking at these two cases, it is apparent that the ways in which citizenship is taught and the desired characteristics of citizens may differ, but the belief in the role of education in creating citizens and solidifying the state is held in common.

Yet, learning about citizenship does not occur solely in classrooms. Critical education scholars use the term “pedagogy” to refer to a broad set of practices and ideologies that are used to reproduce social orders, such as that implied in a citizenship regime; a regime in this sense is a routinized set of practices that reinforce dominant power relationships. Contemporary pedagogical practices often discursively locate citizenship in civil society and local communities, rather than in the state. The practices and qualities of “good citizens” include civility, volunteerism, and active engagements to meet the needs of oneself, one’s family, and fellow citizens in the community. These are generally discussed and promoted as “active citizenship.”

This form of citizenship has been prioritized in the current context of neoliberalism. A component of neoliberal governing strategies, contemporary pedagogies of citizenship promote the ideology that citizens are disempowered by state incursion into arenas and activities that are more appropriately the realm of citizens. Engagement in civil society is proposed as the means of empowering citizens so that they can make their needs known and, crucially, take responsibility for meeting the needs of themselves and their communities; Jessop (2002) describes these moves as combining neoliberalism with neocommuunitarianism. While the governing strategy and political economic context may be neoliberal, the form of citizenship that is promoted is better described as a blend of liberal and republican citizenship, with the state imagined as protecting basic freedoms and individuals assuming responsibility for what had, in Marshall’s theory, been seen as the social rights of citizenship. Whether the resultant form of citizenship is described as liberal, neoliberal, neocommunitarian, or republican, there is a shared view of a basic quality of citizens as self-governing, autonomous, participating in their communities in order to reduce demands on the state, and reliant on the state only for the protection of basic freedoms rather than for social rights.

While much of the academic discussion of contemporary citizenship regimes has focused on the Global North and on countries that have
long been recognized as democracies, there is considerable effort devoted to developing citizenship in the Global South and countries where democracy is only now being established or consolidated. Such efforts are often funded and supported by international or supranational organizations, such as UNESCO or the European Commission, or by government agencies from the North, such as the United States Agency for International Development (USAID) or the Swedish International Development Agency (SIDA). While sharing some elements of the pedagogy of citizenship that circulates in the North, such as the commitment to self-sufficiency, self-governance, and active participation in civil society, the pedagogy in the South and in newly emerging democracies and citizenship regimes also typically includes attention to rule of law, nonviolence, cross-cultural communication, human rights, and entrepreneurship. Many concepts, however, need to be translated to suit both the political structure of countries and the goals of international organizations. In such circumstances, manifestations of the complex spatiality of citizenship are evident: international organizations may promote a form of citizenship that is legally recognized by the state, takes its inspiration from cosmopolitanism and human rights, and is enacted or performed within civil society and communities.

In trying to make good citizens, educators and practitioners do not rely directly on more theoretical or philosophical treatises, so the kind of citizenship they promote does not map neatly onto divisions between legal and substantive aspects or between liberal and republican theories. Instead, and drawing on the pedagogy of citizenship articulated by Osler and Starkey (2005), they often work with a model that conceptualizes citizenship as involving three elements: status, feeling, and practice. Status involves the legal standing of citizenship, as well as the rights and responsibilities that citizenship entails. Feeling reflects individuals’ affective sense of belonging to a political community. The practice of citizenship refers to collective activities in which individuals engage. These elements of citizenship are seen as mutually reinforcing, which means that one might be highlighted or emphasized in promoting citizenship when another element is seen as lacking. For instance, a feeling of belonging might be used to encourage participation in civic affairs, creating a more active citizenry. When that feeling of belonging is either muted or suppressed, people can be reminded that their legal status means that they belong to, or have a stake in, the community and, in turn, that they have responsibilities to be engaged in collective activities to improve the community. Finally, when noncitizens participate in those collective activities, they are likely to be recognized as worthy members of the political community and this recognition may be used to leverage changes in their legal status. Importantly, practice is often associated with localities and involvement in local communities, thus offering individuals affective recognition as citizens, even if legal status may be denied or be impossible. Thus, partial and hybrid citizenships may be formed, even in the absence of legal status.

Critique and alternative formulations of citizenship

There is not a singular, unified account of citizenship. Like all political categories and concepts, the very essence of the term is contested and fragmented. While some of the main lines of debate have been outlined here, there are emerging fault lines in citizenship debates and alternative formulations of the concept.

One of the most significant debates is whether – or perhaps how – citizenship can ever
CITIZENSHIP

be a unifying concept or identity that lends itself to the formation of an inclusive political community or political entity. When citizenship is associated with the nation-state, its very production relies on processes of inclusion and exclusion: people are included as citizens of a particular state with certain expectations of how they should be treated or they are excluded from that nation’s citizenry. These are the boundary issues mentioned previously. But inclusion is not simply based on legal status; questions of deportment, rationality, autonomy, and so forth serve to include and exclude members of social groups as citizens. The characteristics of citizens and the efforts to promote those characteristics, it is thus argued, play an important role in disciplining individuals so that they can be recognized as belonging to the political community.

Many scholars argue that the boundaries of citizenship rely on assumptions about who can perform as rational, autonomous political subjects as they engage in debate and complete their duties and obligations. Because there is a presumption of equality among or between citizens, such political agents are described as disembodied and unencumbered when they operate in the public sphere. Feminists, queer theorists, and cultural critics, to name a few, highlight the ways in which the private sphere and the embodied realities of life condition people’s abilities, their views, and the ways they are perceived – or recognized – in public, and how such recognition in turn affects their standing as citizens. Important strands of argumentation from these perspectives include analyses of the ways in which certain subjects are constructed as incapable of acting as citizens, but also of the ways in which citizenship represents a moral claim and political resource in their struggles for inclusion and justice. Thus, ideals about citizens are used to exclude subjects from citizenship, but those same ideals can be used to demonstrate the injustice that comes from the exclusion of particular groups. For example, women’s suffrage movements in the Global North used the figure of Liberty, who is always depicted as a woman, in their campaigns to argue that women had the requisite characteristics of citizens. Yet the moral argument about the injustice of excluding women did not extend to racialized minorities, thus demonstrating the partial or, at best, incremental potential that citizenship offers to drives for inclusion and justice.

Other critiques imagine alternative sites or locations for citizenship, often by attempting to de-link citizenship from the nation-state; doing so, however, necessitates a conceptualization in which legal status is less important than citizenship’s substantive aspects. Some of these critiques draw from long-standing ideas about the importance of civil society as a site in which citizenship is enacted and from which the state can be held to account. As noted, enacting citizenship in civil society lends itself to action at the local level and to civic or republican forms of citizenship. Radical theories of citizenship, however, also emphasize local action and practices. Such theories argue that citizenship should not be granted by the state but, instead, be based on inhabitance. Inhabitance draws on the lived experience that comes from living and participating in the city and that generates an affective attachment and feeling of belonging. Radical citizenship theorists argue that inhabitance generates a better understanding of the needs of others, and, without being constrained or defined by the state, gives rise to politics in which concerns for social justice will be dominant. In this view, the state is not irrelevant but, rather, blocks or channels the democratic impulses that citizens would otherwise enact. Such views may be particularly compelling in postcolonial contexts in which the state and its system of governance have often been imposed externally, rather than
emerging through democratic iterations within a country.

Yet another strand of critique argues that the nation-state is becoming irrelevant to citizenship owing to forces of economic globalization, to transnational migration, and to the development of human rights regimes at the international level. These critiques often lead to arguments for global citizenship and highlight its potential for creating more just citizenship regimes, even as inequality between and within nations seems to be increasing. Many scholars also argue that attention to human rights and global citizenship can provide a counterweight to nationalist tendencies and policies enacted by states as they attempt to fortify their boundaries, often excluding people from citizenship outright or marginalizing them through the withdrawal of social rights.

Implicit in many of these critiques are questions about agency and whether citizenship – the standing or status – affects the actions of individuals as citizens. Historically, agency did not figure prominently in descriptions of citizenship, other than through questions of autonomy, rationality, and performance of duties and obligations. More recently, and often attributed to neoliberal citizenship, emphasis has been placed on the development of active citizens who will be engaged in their communities through volunteerism and other similar activities. Some analysts are concerned, however, that volunteerism and engagement are channeled into activities that allow state withdrawal from supporting the social rights of citizenship; these scholars also argue that active citizenship may mute claims on and against the state as communal responsibility becomes internalized. In its place, such scholars call for activist citizenship, in which the state is directly challenged by citizens in what Isin (2008) has called “acts of citizenship.” Such acts break with the everyday practices of citizens and of the state, and potentially call forth or enable new relationships and new political practices, including acts of insurgency. Such insurgent citizenship (Holston 2009) represents a destabilization of existing citizenship regimes, often through urban-based struggles for housing or equality. Insurgent citizens struggle for more radical forms of democracy, reconfiguring the rights and protections of citizenship in ways that pose fundamental challenges to existing state structures and to the political economy of citizenship more broadly. Insurgency is never complete, however, but is an ongoing struggle. In this way, it reflects the continually contested nature of citizenship itself.

SEE ALSO: Civil society; Democracy; Nation-state; State, the

References


CITIZENSHIP


Further reading


“City logistics” is a recent term in the scientific literature as well as a new practice. It has been mostly developed in Europe since the 1990s and is now applied worldwide. City logistics represents more than the mere transportation of goods within urban areas. It can be defined as any service provision contributing to efficiently managing the movements of goods in cities and providing innovative responses to customer demands. The objective of city logistics is to make deliveries to urban residents and businesses possible at the highest economic, social and environmental standards.

City logistics includes physical operations such as order preparation and packaging, transportation and deliveries (including home deliveries), short-term storage of goods, management of drop-off/pick-up boxes for parcels, return of goods, waste management including the management of recycled goods, and the management of empty pallets and packages. City logistics makes use of sophisticated information and communication technologies to increase coordination, efficiency, and adequate enforcement of urban delivery operations.

Defined as such, city logistics does not apply well to the way goods are currently delivered and picked up in cities worldwide. A significant portion of these activities remains inefficient, with numerous and poorly coordinated freight trips and many unclean delivery vehicles (Dablanc 2007). To decrease cost, increase efficiency, and promote environmental sustainability, there has been a recent surge in innovative practices in urban freight and city logistics. These operations involve various stakeholders, such as trucking companies, logistics providers, real estate developers, and major retailers, as well as start-up companies. New logistics services have emerged in cities around the world, especially in areas with high levels of activity, such as city centers.

Although not yet significant in terms of volume and environmental benefits, these practices provide new directions for more sustainable city logistics activities. Organizations and projects in city logistics have been described in books (Melo and Macharis 2011; Gonzalez, Semet, and Routhier 2013). Giuliano et al. (2013) compare city logistics issues for the United States and Europe. Wolpert and Reuter (2012) provide a review of city logistics in the scientific literature. An International Conference Series on City Logistics (Taniguchi and Thompson 2016) provides an opportunity to share academic results in this field every other year. The US Transportation Research Board and the European Commission organized a joint symposium to discuss city logistics research issues (TRB 2013).
chains, from the supply of office products to commercial activities to the delivery of food to restaurants. These urban supply chains are the result of logistics decisions that are, in turn, based on the demands of the production and distribution sectors, themselves dependent on the behavior of economic agents such as households and firms. Each activity (commercial, service, industrial, administrative) taking place in an urban environment can be associated with a specific freight profile. Freight vehicles are very diverse: trucks and vans of all sizes and weights, motorbikes and bicycles, pedestrian push-carts, and, more rarely, rail and waterborne transport. Urban delivery systems unique to one city can also develop, such as the famous “dabbawallas” in Mumbai, India: 200 000 lunch boxes made at home are delivered each day to businessmen at their workplace through a collection/sorting/delivery system using bicycles, trains, and pedestrian modes of transportation. Cities can have individual logistics specificities. For example, Chicago has been preoccupied with maintaining its prominence as a rail freight hub for North America, and is thus concerned about rail freight movements between the numerous rail terminals located within the city. Los Angeles is primarily concerned with air pollution and targets drayage associated with the ports of Long Beach and Los Angeles. Shanghai has become the largest cargo port in the world and its logistics added value is estimated at 13% of its GDP. Activities from the three ports of the bay of Tokyo add much truck traffic to Tokyo’s streets. In Mexico City, 42% of the working population works in microcompanies, of which half are home-based or street-based, generating very specific patterns of deliveries. The wholesale market in Mexico City (Central de Abastos) generates 52 000 truck trips every day.

A sustained effort in urban freight data collection and modeling has yet to be accomplished in many cities around the world. Although a lot of progress has been made in the last decade, survey methods remain heterogeneous, making it difficult to compare results from one city to another. In France, the FRETURB model (Laboratoire Aménagement, Economie, Transports, Lyon) has adopted a methodology well suited for urban economics. It identifies detailed freight patterns of urban establishments from extensive — but expensive — “urban goods movements surveys.” The last surveys were made in Paris and Bordeaux in 2011–2015. These surveys revealed a key ratio for French cities: there are about as many urban deliveries a week as there are jobs in a metropolitan area (Ambrosini, Patier, and Routhier 2010).

Urban freight is more polluting than long-distance freight transportation because of its modal profile; the average age of the vehicles and the high number of short trips and stops. Impact surveys carried out in European cities (BESTUFS 2006), show that urban freight generates between 20 and 60% (according to the pollutants considered) of local transportation-based pollution. Nitrogen oxides (NOx) and particulate matter (PM) are pollutants for which urban freight has a particularly important responsibility. Both pollutants are very harmful to the health of urban residents, and in European cities their concentration tends to stagnate or even increase, contrary to many other urban air pollutants. In Mexico City, the average age of trucks and vans is twelve years. The renewal of urban freight vehicles is generally slow (slower than for nonurban road freight traffic), because most trucking companies operating in urban areas are very small, competition among operators is acute, and profit margins are low: many operators save money by buying used trucks or vans and using them as long as possible.

Greenhouse gas emissions and noise pollution are also among the most severe environmental impacts of freight in cities. Freight represents
about one-quarter of transportation-related CO₂ emissions in European cities.

A survey of 1650 Mexico City truck drivers (Lozano Cuevas 2006) identified the challenges met by urban freight operations that are actually common to operators in large cities in many countries: congestion, lack of loading and unloading space, complex legislation, risk of theft, and lack of safety are among the drivers’ greatest concerns. Another important urban freight issue is road safety related to truck traffic. Trucks have a low share of accidents in cities but the accidents involving them are more serious. The reconciliation of truck traffic with rapidly increasing bicycle use has been a recent cause of concern in Paris and London following much-publicized fatal collisions. Recent marketing trends such as same hour delivery for residents in selected cities (such as New York) put delivery personnel at risk. Drivers, often on bicycles or motorbikes, use whatever driving and parking tactics they find that can make them meet the promised delivery times.

Emerging city logistics services

City logistics services respond to new demands from urban businesses and consumers. Consumer behavior has changed rapidly in recent years. In many countries, online shopping has become a mainstream consumer activity, especially in large cities. This generates a demand for new logistics and freight services, among them home deliveries or deliveries at the office or in drop-off/pick-up depots. Demand for instant deliveries is growing. Companies also have a growing demand for new logistics services. Surveys have shown that many inner-city retailers are interested in, and willing to pay for, new services such as the rental of storage space, the provision of dedicated areas for receiving deliveries out of the stores’ premises, and specialized services for the pick-up of pallets and empty boxes and packages. City administrations have also expressed growing demands for a more efficient urban freight and city logistics global organization.

In response to these demands, entrepreneurial initiatives in city logistics have emerged, first in Asia and Europe, then in the United States and other regions. Companies such as Shurgard provide urban storage space for storekeepers. Star’s Service, a transportation company with 1500 truck drivers, has been highly successful by specializing in home deliveries of food products to Paris households. In Germany, DHL/Deutsche Post has installed thousands of “PackStations,” or automated lockers, in cities’ public spaces so that e-commerce consignments can be delivered to customers at any time. In Japanese cities, takkyubins (or takuhaibins; such as Yamato and Sagawa) have responded quickly to the evolution of consumer behavior and deliver everything from general parcels to frozen goods to large products, such as luggage or golf equipment. The size of their networks, the diversity of their services, their high level of logistics efficiency, and the extensive use of home deliveries by Japanese families make takkyubins stand out as a model, not yet applied anywhere else in the world. In Paris and London, cargo-cycle companies provide last-mile parcel delivery services to businesses and residents in urban locations. The use of cargo cycles brings environmental benefits. It has been estimated that the Paris-based freight company La Petite Reine, using 40 electrically assisted freight bikes, saves 200 tons of CO₂ per year compared to the previous organization based on diesel vehicles. However, the use of cargo cycles necessitates the use of small terminals located in inner city locations to prepare delivery tours. Since 2003, the municipality of Paris has carried out a program to implement “urban logistics spaces,” small trans-shipment facilities located
CITY LOGISTICS

in the densest areas of Paris, most of them in underground municipally owned car parking facilities. These logistics spaces are allocated at a low rent to the logistics operator offering the greenest mode of delivery. Cities in the United States have recently experienced a surge in sophisticated city logistics services. Amazon provides goods pick-up lockers to urban residents in 7-Eleven convenience stores. It is testing grocery deliveries by appointment to residents of Seattle, San Francisco, and Los Angeles, and the company has generated much attention with its futuristic drone parcel delivery idea. In major US cities, instant deliveries are provided by services such as Uber Rush, Postmates, or Instacart. FedEx and UPS routinely use electric delivery vehicles to operate in urban areas or university campuses across the country.

Public policies for city logistics

Involvement in urban freight and city logistics for the public sector can be manifold and at various levels. Cities are in charge of the urban road network, which is heavily impacted by commercial vehicle traffic. Cities are also interested in local economic development and, therefore, have to make sure that the provision of goods and logistics services is adequate. Cities control the use of land, including the land necessary for logistics activities (warehouses and terminals), and are in charge of local traffic and parking regulations, including all regulations that relate to delivery vehicles. Finally, cities are concerned with environmental and social issues, which are associated with urban freight activities.

Despite these many mandates and potential objectives, actual local public policies regarding freight and logistics are quite modest. Most cities still view truck traffic as something that should be banned or at least strictly regulated, and few of them consider freight activities as a service they should help organize. Local policies are generally very parochial and can be conflicting. In the Lyon metropolitan area (France) there are as many as 30 different municipal rules on truck access regulations based on weight and size, forcing truck drivers to decide which rules they will comply with, and which ones they will disregard. Major events such as the organization of the Olympic Games provide opportunities to organize a change in urban freight operations citywide. Barcelona and London, today, are cities that have a clear logistics strategy. In many other cities, major urban development projects, such as the implementation of a tramway or a subway system, can also play that role.

The tool that cities, when confronted with freight activities, often use first is truck access restriction. These restrictions are based on various criteria such as time windows, weight, and size (length, surface). The most famous truck ban in Europe is the London lorry ban, in place since 1975. Heavy goods vehicles over 18 tons cannot circulate at night and weekends. All trucks in Seoul have been banned since 1979 from the central areas during working hours. Truck access rules based on weight or size policies have several drawbacks. They tend to promote small-capacity vehicles (vans, light trucks), which increases total congestion and reduces the efficiency of urban freight. Regulating truck access requires enforcement and control, meaning a sufficient and well-trained staff. The most recent trends in access restrictions are environmental standards and road pricing. Both can be combined, as is the case in the London congestion charge. Urban tolls are not specific to commercial vehicles, while many environmental zones or “low emission zones” are. In London since 2008, a low emission zone prohibits access to trucks older than the Euro IV emission standard (trucks manufactured since 2001). Large vans have also
been targeted. The area where the rule applies is surrounded by the M25 motorway and covers about 600 square miles. New enforcement technologies have increased the efficiency of truck access policies. Automatic control systems, such as automatic number plate recognition cameras, mobile enforcement, vehicle positioning and on-board equipment, have been introduced in many cities in Asia and Europe. This technology comes at a cost and generates controversy over privacy issues. The reduction of noise caused by deliveries and the promotion of night (or early morning/late evening) deliveries are also an emerging urban freight policy target. According to a survey carried out in New York City, businesses most likely to change to off-peak deliveries are shippers doing their own deliveries (own account transportation), and receivers open during extended hours, such as restaurants (Holguin-Veras 2008). In the Netherlands, in 25 pilot cities, the national government provides financial help for operators investing in silent delivery equipment, which includes vehicles and handling equipment generating noise emissions below 65 dB, for night deliveries at supermarkets. Tests have shown that companies delivering at night save 30% in delivery costs and 25% in diesel consumption.

Freight operators need good access to dedicated loading and unloading areas, be they public or private, on-street or off-street. The lack of sufficient delivery spaces transfers delivery operations onto traffic lanes or sidewalks, and leads to congestion and traffic accidents. In busy urban areas and city centers, adequate on-street loading and unloading bays must be identified and their use must be better controlled. Time sharing is a good way to improve street parking capacity for trucks and vans. In Barcelona, the municipality has created an innovative organization on some of its main boulevards by devoting the two lateral lanes to traffic during peak hours, deliveries during off-peak hours, and residential parking during the night. The city is also renowned for its dedicated mobility motor squad, consisting of 300 agents circulating with a motorbike and controlling all on-street parking activities, including loading/unloading zones. This has prevented illegal long-term parking and made these zones always available to delivery truck drivers.

Consultation processes in urban freight provide collaborative opportunities between private companies and local administrations that otherwise are not willing to work together. In September 2013, the city of Paris and the most important carriers’ and shippers’ associations signed a charter for sustainable city logistics. In the United Kingdom, freight quality partnerships have been developed in many cities. Consultation processes can lead to initiatives, such as certification processes, that identify the best urban freight operators. Certification confers privileges on an operator, such as extended delivery hours or the use of designated loading/unloading facilities. It may also provide operators with a competitive advantage when bidding for contracts. An example of such an initiative is the FORS (freight operator recognition scheme) in London. The FORS certifies operators that comply with a list of efficiency, safety, and environmental impact criteria at bronze, silver, and gold levels. It gives companies access to data, benchmark information, and training programs for their drivers.

Local policies for city logistics also include the control of logistics land use through planning provisions. A pattern of “logistics sprawl” (Dablanc and Ross 2012) has been identified in several North American and European metropolitan areas in which warehouses are being sited in increasingly ex-urban locations. Decentralized logistics facilities tend to generate more truck miles, as delivery trucks need to cover much longer distances to reach urban
CITY LOGISTICS

destinations. This adds to regional congestion. A regional overview of logistics land uses is therefore needed. Another potential strategy is the development of logistics parks (sometimes called freight villages), partly through public investment. Some notable examples include Güterverkehrszentrum (GVZ) at the outskirts of large German cities, or the Raritan Center in New Jersey. These logistics parks commonly include services for all companies located on the site, such as surveillance, catering, fueling and cleaning stations for trucks, overnight truck parking, and night accommodation for drivers. Former industrial areas (brownfields) can be used when developing freight villages. However, necessary remediation efforts (e.g., for the removal of polluted soil) should not be too expensive, as programs for logistics spaces cannot tolerate high land costs.

Many cities impose the building of off-street delivery areas in new commercial or industrial developments. The Tokyo off-street parking ordinance of 2002 compels all department stores, offices, and warehouses to provide for loading/unloading facilities when they have a floor area of more than 2000 m². Cities should also be careful in avoiding inadequate regulations in their building and planning codes. Access to underground parking, for example, whether in public parks or in private buildings, can be very difficult for commercial vehicles, even small ones, because of height and size limits. The Chicago Downtown Freight Study (January 2008) mentions this as an important problem for truck drivers, and the municipality has worked on a freight plan including recommendations for building provisions.

Some municipalities accept or even promote the development of multi-activity buildings, with logistics on street levels and other activities on the upper levels. Some also welcome multistory warehouses. The logistics developer Prologis has built seven-story logistics terminals in downtown Tokyo. The acceptance of such projects is increasingly conditioned to environmental criteria and to a careful attention to the integration of buildings into the existing urban landscape. The development of “urban consolidation centers” has often been identified as an interesting strategy to reduce urban freight impacts. These facilities provide a service of bundled and coordinated deliveries, often requiring public subsidies. Up to 200 such terminals existed in European cities in the 1990s and early 2000s. However, due to operating costs, most of them closed down when municipalities could no longer subsidize them. Today, a few urban consolidation centers still operate, mostly in medium-size cities: Bristol in the United Kingdom, Modena, Padua, and other Italian cities, La Rochelle in France, and Yokohama in Japan. These consolidation centers are expensive and the allocation of operating costs is difficult.

Conclusion

Despite traffic difficulties and costs associated with activities in urban areas, freight operators manage to serve the vast majority of their urban customers. In this way, today’s urban logistics are highly efficient. However, this often comes at an environmental cost. Goods movements represent 10–15% of all transportation in an urban area (in vehicle miles), but one-third or even more of the local pollutants related to urban transportation. Patterns such as “logistics sprawl” add more vehicle miles and congestion to the metropolitan and regional road network. To these many challenges, city logistics has recently emerged as an innovative response. It intends to meet urban businesses, and customers’ demands for new services while responding to cities’ concerns over the environmental impacts of current urban
deliveries. However, the pace of actual introduction of city logistics in urban areas is slow and varies greatly from one city to another. To characterize this diversity, Dablanc and Rodrigue (in press) have proposed four illustrative categories of city logistics cities: (i) land-efficient urban logistics cities (such as Tokyo) that integrate freight facilities in very dense urban settings; (ii) cities as mega distribution centers (large US and European gateways) acting both as distribution facilities for local markets as well as regional hubs for the grouping and redistribution of goods to regional and national markets; (iii) dual urban logistics system cities (megalopolises of developing countries) in which a modern freight and logistics sector coexists with an informal system of pick-ups and deliveries for home-based artisans or street vendors; and (iv) smart logistics cities (the historical centers of some European cities, some Asian cities) with an emphasis on cleaner, more silent operations, consolidated deliveries, a high level of home delivery services, and high-tech enforcement systems. This last category still represents a marginal share of total freight activities in metropolitan areas, but receives a lot of media and decision-makers’ attention.

SEE ALSO: City logistics; Climate change, concept of; Corporations and e-commerce; Environment and urbanization; Environmental policy; Global commodity/value chains; Logistics; Public policy; Sustainable transport; Transport geography; Transportation planning

References


CITY LOGISTICS

Civil society

Andrew D. Davies
University of Liverpool, UK

The Enlightenment conception of civil society

The concept of civil society is generally used to describe an understanding of politics that has its roots in European Enlightenment thought, where it is the sociopolitical domain that exists between the household or family and the state. In particular, Hegel’s work on the links between civil society and the state has become central to many understandings of civil society. As he set it out most clearly in *Philosophy of Right* (1967/1821), Hegel, saw civil society as a liberal political formation, grounded in modernity and operating through *Gesellschaft*, or societal and social practices – an arena of relatively formal political behavior. This distinct zone was separate and untainted by the more affective and potentially emotional sociality of *Gemeinschaft*, or community practices, which have the potential to be more populist and are driven by issues concerning familial or communal identities (see also Tönnies 1955). This liberal conception creates a clear zone where there is a distinction between the community, civil society, and the state, and, as a result, civil society is central to processes like democratic deliberation. While civil society is a zone where individuals can come together to mobilize and make claims on the state, it is also far from being a zone of equality – power relations between individuals and groups determine their ability to change their circumstances. Thus, from the outset, the functioning of civil society is contingent on the ability of individuals and organizations to mobilize and act with or against each other in order to convince the state of the merits of their case. The state therefore becomes the top of a relatively linear hierarchy, in which individuals sit at the bottom. This hierarchical understanding crucially reifies the state as the core arbiter of political activity. Organizing in civil society, therefore, occurs in a buffer zone, wherein the state arbitrates from “above,” and people mobilize from “below.” Given this framing of the state, this conception of civil society lacks the potential for radical political change: the core dynamic of politics is the interaction between civil society and the state, and thus any radical reform that sees the state as something to be defeated or removed (e.g., an anarchist politics) does not fit into this political framework.

While Hegel himself was critical of this liberal version of civil society as a bourgeois construct grounded in an overly rational modernity, the understanding of civil society as a buffer zone between the state and the family or individual remains the dominant framing of the term. The liberal Enlightenment conception of civil society was contested by a Marxist reading, particularly that developed by Gramsci (1971/1929–1935). While keeping the hierarchical architecture of the Hegelian reading of civil society, Gramsci argued that the links between the formal political society of the state and civil society were actually closely interwoven, with bourgeois and capitalist interests often shaping the two domains for their own benefit. Gramsci argued that, instead of a liberal zone of potential freedom and...
democratic deliberation, civil society was where hegemonic relations were made in order to maintain class differences. These two readings of civil society – as a space of liberal negotiation and deliberation or as a domain where hegemonic power plays out – remain the two pre-eminent interpretations of the term.

Today civil society is still broadly conceived as the space where negotiation, resistance, domination, and other political practices take place and are worked through, although it also incorporates a diverse range of nonstate actors such as social movements and nongovernmental organizations (NGOs) and is increasingly influenced by social media and the Internet. Despite being theorized as outside the state, civil society is closely linked to, and indeed exists only because of, the state. Unsurprisingly, given the Hegelian origins of the term, the state and civil society are often linked in a dialectical relationship. As a result, through its claim to represent those political actions that sit beyond formal, state-based politics while at the same time being closely implicated in formal decision-making, civil society has become a key term for geographers, especially development and political geographers, although the term still assumes a relatively linear system – with the state at the top; individuals, families, and communities at the bottom; and civil society operating in the middle ground. Crucially, this framing also conceives of civil society as a relatively orderly and structured part of the political system. Both of these principles – civil society fitting into a hierarchical scalar order and as an orderly space of politics – have been critiqued (as we shall see). In addition, there remains considerable conceptual vagueness about civil society, and, as a result, the utility of the term can often be called into question; put simply, is “civil society” such a broad umbrella term that it generalizes and obscures the range and depth of political activity in the world? Going further, given Gramsci’s treatment of civil society as a zone of hegemony, does civil society actually have the potential for change that it is supposed to offer? Can we actually rely on civil society as a buffer zone to change society for the better, or can we at best hope to change the odd marginal decision while powerful stakeholders maintain their hegemony? These questions lie at the heart of many contemporary debates about civil society.

Contemporary forms of civil society

With the rise of NGOs (and later international NGOs) as institutions in the 1970s, civil society became a crucial topic of contemporary political concern. This rise shifted the concept from a topic of debate for political philosophers to a key object of study across the social sciences. This is often seen as a part of an emerging neoliberal economic consensus, with state powers being rolled back and diminished as market-led approaches became dominant. As a result, civil society organizations, particularly those in the “third sector,” came to be seen as providers of services that the state was now unable or unwilling to provide. Over the next two decades, and with the decline of state socialism globally (including a perception that civil society was important in precipitating this decline in Eastern Europe), there was increased conjecture about the role of civil society. As a result, from the early 1990s onward there was a huge range of writing about civil society, particularly the potential benefits and problems associated with it. In particular, once globalization became a key academic and policy buzzword, the potential for a new “global” form of civil society sat quite comfortably among discourses that stressed the increasing economic and cultural integration of the world. At their most optimistic, ideas about global civil society stressed the potential
for new forms of peaceful integration at a global scale. These works also attempted to expand the concept of civil society as it was used by Enlightenment scholars and to map out the new spaces for engagement that were emerging as the world globalized. For example, Kaldor’s (2003) global form of civil society encapsulates three overlapping contemporary areas: an activist sphere, based on self-organization and a rejection or contestation of state or market power; a neoliberal sphere, based on third sector charities and NGOs taking on increased responsibility as the state was rolled back by market forces; and a postmodern sphere, an anti-universalist space, where the difference between individuals and groups is recognized and accepted. This reading of a new global civil society was an attempt to categorize and understand the increasingly complex political, economic, and social structures that were (and are) emerging as a consequence of globalization.

Although it moved beyond the liberal-democratic framework of early conceptions, the broad notion of a global civil society still regards civil society as a key arena for redressing imbalances in global power structures, and holds that the increasingly globalized economy will respond to and can be regulated by the activities of civil society as well as those of the state. Campaigns against and boycotts of transnational corporations over, for example, perceived unfair labor practices became one of the most visible signifiers of this turn. More problematically, such conceptions of globalism often create the ideal of civil society as a normative project, where certain universal standards are seen to be applicable to many groups across the world. This marks a shift from using civil society as an explanatory tool for understanding how societies negotiate their political differences into a more prescriptive term where what and which social practices take place are of equal importance. In turn, this normative reading has meant that civil society can arguably be seen as a new vehicle for the enforcement of neocolonial and neo-imperial practices where those who are judged to be more valid civil society actors (most often those based in the Global North) are able to tell others what they regard as acceptable standards of behavior and to set goals or standards for those others to aspire to.

The idea of a liberal global civil society is further complicated by examining how the political practices of civil society play out in reality. The play of power between interest groups in civil society is often far less certain than the optimistic readings of advocates of globalist thinking. Authors have increasingly argued, following Gramsci’s conception, that civil society in practice often fails to create the leverage necessary to create more just alternatives to existing political structures and practices (see Bénit-Gbaffou 2012 for a geographical perspective), and in many cases has served simply to reinforce existing hegemonic frameworks that prioritize the powerful at the expense of the powerless. Indeed, to some, civil society is a meaningless concept in a world where state power has been amalgamated into the more complex framings of the neoliberal global order. James Ferguson (2006, 91), for example, argues from an anthropological perspective that treating civil society uncritically simply serves to obscure, and at worst legitimate, a “profoundly antidemocratic transnational politics” that works, through civil society, on the globally poor and disenfranchised in both the Global North and South. This argument harks back to Gramsci’s earlier critiques of civil society as a tool for maintaining capitalist hegemony; in particular, Ferguson critiques the inherent hierarchy present within liberal civil society and argues for a topological reading of the power relations that structure global development.

This topological reading, rather than seeing power as rooted in nation-states, places power in the hands of those who create and perpetuate the
CIVIL SOCIETY

unequal relations of power at work within civil society. This means complicating the conceptual framework so that, instead of creating a linear hierarchy between the individual or family, civil society, and the state, we should include market structures within our understanding of the global order, but also recognize that power operates through and across these various spheres in contingent and dynamic ways. This vision develops globalist perspectives and creates a reading of power that goes beyond seeing civil society as a mere buffer zone, and instead as complicit in the various political projects that seek to maintain or alter existing social structures. This is particularly attractive to geographers, who have long argued for relational understandings of spatial processes, particularly John Allen’s (2011) work on topological space which sees power relations as more nuanced than being blunt practices of domination and resistance. These geographical understandings of power emphasize the limitations of seeing civil society as a rational space of deliberation which the state, market, and individual somehow sit outside. Indeed, as Craig Jeffrey (2011, 148) has recently pointed out in relation to work on young people and politics, continuing to read civil society as clearly divided between Gesellschaft and Gemeinschaft occludes relations where identities and political formations can shift over time and space as they come into contact with other actors and organizations. Thus, the idea of civil society as a separate zone of rational activity is inherently problematic because it is not clear where civil society begins and ends, but also because it fails to account for how groups move across or straddle these boundaries; for example, according to UK law, a registered charity cannot take part in political activity, yet many often have implicit political purposes (e.g., many registered charities were involved in the 2005 Make Poverty History campaign against the G8, which had clear political aims). This is a problem of how UK law defines politics, but also an example of how civil society’s limits are often unclear and permeable. This lack of clear conceptual boundaries, together with the continuing desire to see civil society as a distinct political sphere, lies at the heart of postcolonial critiques of the concept, as we shall see.

These conceptual difficulties can be seen by examining one area where civil society has undoubtedly expanded over the past few decades, as global forms of interaction facilitated by the Internet have hugely increased, particularly driven by Web 2.0 practices such as blogging, social networking, and wiki creation. This growth has also necessarily raised questions about the potential for change allowed by civil society activity. On the one hand, movements like the Zapatistas in Mexico offer relatively hopeful tales of transnational organization facilitated by the Internet. There is obviously the potential within these movements to rework and transform social relations as they occur across space, and to share knowledge, organizational practices, and strategies across a transnational civil society (see Jazeel (2010) for key geographical issues here). However, the tensions raised by the Internet also raise questions about what it means to be truly active within civil society. Olivier Marchart (2011) has argued that “minimal” politics represents the lowest threshold of what can be considered an actual form of political activity. The Internet’s ability to spread knowledge means that a new range of interaction takes place, such as the organizing of international meetings, writing blogs, and more mundane activities such as signing an e-petition or posting news of a topical issue on websites like Twitter. Internet civil society has become an important topic in academic studies, but it is a further development away from the rigid, state-led structures provided by the original liberal Enlightenment
conception. This creates a whole new range of activities that can be identified as broadly belonging to civil society, which in turn also expand the boundaries of what count as attempts to mobilize in civil society. For example, is it now possible to be political without leaving one’s desk or home, and what new forms of politics are created as a result? In addition, thinking about what counts as minimal raises questions about how effective and useful these new forms of civil society action are. This is not to say that the mundane cannot be politically effective in the transnational spaces created by the Internet. Far from it: these actions could potentially act as a useful rejoinder to accounts that see only certain forms of engagement (demonstrating, voting, lobbying, etc.) as truly political. But we do need to question how (and whether) these new forms have the potential to reshape what we regard as political behavior, removing an emphasis on supposedly grander forms of mobilizing and allowing for more action from what would traditionally be seen as “below.” There is also a need for caution here, as the Internet’s potential for change is often overstated, and not everyone has access to this new online civil society, and this means recognizing that these online domains are fraught with power relations and access issues just like any other part of civil society (see, e.g., the debate about blogging and public geography in Kitchin et al. 2013). Indeed, it is from critiques of civil society as an elite space that one of the major conceptual challenges to the term has emerged since the beginning of the twenty-first century.

Moving from “civil” to “political” society?

Despite the changes in civil society driven by globalization and the Internet, postcolonial critics have continued to question the elitist and universalizing tendencies inherent in the term. One of the most visible of these critiques comes from the work of Partha Chatterjee (2004, 2011). Given the large numbers of people around the world who occupy various informal or illegal statuses, Chatterjee has argued that large swaths of the global population do not have formal citizenship and are therefore excluded from civil society. If civil society is regarded as a domain that belongs to the relatively privileged, Chatterjee argues, popular politics in much of the Global South occupies a differential zone, what he calls “political society.” Political society stems from the particular forms of governance that many states now employ. Instead of being treated as individual, rights-bearing citizens, in the Foucauldian language of governmentality, people are grouped according to their characteristics (e.g., as informal settlers or as belonging to a certain communal group), thus allowing the state to target these populations according to their criteria, and therefore to fulfill the state’s perceived obligation to “improve” these populations’ conditions. However, not all groups can be targeted because of the limitations of the “rolled back” neoliberal state. As a result, it is those groups that can make themselves visible but also argue that they are the most worthy of aid that are able to claim resources from the state. Thus, groups that are able to mobilize effectively and present their case in a way that the state can understand, and that the state sees as worthwhile (e.g., fitting with the latest policy target) are more likely to receive the benefits of limited state funds. Here, then, is not a formal, rational, Hegelian civil society. Instead, political society occupies a “normatively nebulous zone” (Chatterjee 2004, 50), where practices like clientalism, brokerage, and other more or less formal/legal activities take place as those excluded from formal civil society try
to get what they can from the state. Again, here the boundaries between *Gesellschaft* and *Gemeinschaft* become indistinct as groups move strategically between the categories according to their need. As a result, the concept of political society explicitly argues for a more diverse set of political practices than is usually allowed under the term “civil society,” which Chatterjee, following Marx and Gramsci to an extent, frames as a bourgeois zone. Political society then, is a space where a series of mundane, possibly minimal, political activities – such as visiting the office of a government official on a daily basis – rather than the clearly “political” gestures of formal civil society, are practiced. It is also avowedly not a radical space of transformation, but a space where communities and individuals try to get by as best they can by explicitly engaging with and utilizing the state’s facilities.

Chatterjee, with his background as a scholar in the Subaltern Studies Collective in the 1980s and 1990s, sees the intellectual trajectory of political society as fundamentally a rejection of the idea that there can be a global and universal civil society. Civil society as defined by “the West” is fundamentally normative, where certain democratic and moral choices are viewed as the ideal for everyone. Instead, Chatterjee argues that it is the incomplete and problematic application of European thinking across the Global South, as a result of colonialism, that lies at the heart of the problem, as a result of which civil society does not adequately cover the range and scope of political activities available to the majority of the world’s population. The challenge is to think through how European modernity has created a normative set of rules that in turn have created the sense that civil society (and any other category that emerged from European thought) is a form of hegemonic domination. Thus, the task is to interrogate how these universal norms are taken up, interpreted, and appropriated according to specific historical, social, cultural, economic, and political circumstances. Political society then can be read as a project that partially continues the subaltern studies ideal, which in turn demands that the legacies and unequal power relations caused by colonial rule in any discussion of civil society and that we still see at work today are taken seriously in any discussion of civil society. The key geographical question then is interrogating how these power relations play out in reality, and how the difference between political society and civil society functions for those who are involved in contesting and negotiating with the powerful on a daily basis. Chatterjee leaves this relatively unexplored in his own writing, but it has been taken up by numerous authors since (Gudavarthy 2012), and indeed it takes us back toward the ideas about contingency and specificity raised by the earlier discussion of topology.

Political society has brought renewed vigor to discussions of the limits of civil society, and the concept has become one of the most widely debated interventions on civil society in recent years (Gudavarthy 2012). However, political society has been discussed overwhelmingly in an Indian context, and as a result the question remains as to how far political society functions in other areas of the world. In *The Politics of the Governed* (2004), Chatterjee argues that similar processes must be occurring in other areas where people are governed and treated as populations. This opens up interesting questions about the limits to civil society and how these are spatially uneven, with different states/spaces/places experiencing differing forms of governmental activity. As Harrison (2012) has pointed out, however, the danger of political society is that it replaces one conceptual black box (civil society) with two (civil and political society). In arguing that political society exists as a realm distinct from civil society, and leaving civil society as a
relative ideal occupied by the elite, Chatterjee has little to say about either his conception of civil society or the relationship between civil and political society. Political society shares many characteristics with civil society and, in its direct contact with state functions through processes like corruption, vote rigging, and clientalism, it is seemingly less democratic and more entangled with the state than civil society. However, this again highlights the dynamic nature of the interaction between the state, its citizens, and wider market forces. Chatterjee’s political society is not an alternative to civil society; it is, instead, a recognition that civil society is a fundamentally limited concept that fails to grasp the messy, and often morally dubious, nature of politics as it is practiced by those who sit outside the traditional liberal political framework which underpins it.

Conclusions

Civil society remains a central concept to geographical understandings of politics, democracy, and social organization. However, it is also a concept that should be approached with a degree of care. Despite the seeming logic of compartmentalizing extra-state, extra-communal, and extra-market political practices into a distinct zone, there are continuing issues as to how this framework, rooted as it is in Enlightenment values, creates normative hierarchies and in many cases fails to deliver on its supposed promise for democratic deliberation. With transnational and global interconnections making ever more visible the complexity of social relations across the globe, the conception of civil society as a distinct arena remains problematic. Optimistic readings would still view civil society as performing a useful service in helping us to understand non-state political activity. For left-leaning liberals, civil society can be seen as still possessing the potential for change and resistance to the increasing corporate and capitalist hegemony of the world. Chatterjee’s more pessimistic reading has led to the more troubled zone of political society, where the majority who cannot gain access to civil society are left to compete for the scraps that neoliberal governmentality offers them. What is clear is that geographical understandings of politics and political organization are still informed and shaped by these competing conceptions of civil society as a space of political potential, and debates about their utility are likely to continue.

SEE ALSO: Citizenship; Community; Democracy; Governance and development; Governmentality; Neoliberalism; Political geography; Postcolonial geographies; Social movements; State, the; Subaltern; Topological relations

References

CIVIL SOCIETY


Class

Marianna Pavlovskaya
Hunter College and CUNY Graduate Center, USA

Class is one of the most important, widely used, and complicated concepts in human geography and the social sciences. Its meaning varies according to particular social theories and human geographical traditions, and to related conceptualizations of the economy. Different geographers, therefore, use and understand class in a number of ways. It is important to understand these distinctions because they affect our understanding of how class works and have different implications for policy, politics, and geographic research. Class as a concept is foundational to economic geography; it permeates research on local and global issues, the gendering of work and socioeconomic space, geographies of race and racialized economies, postcolonial struggles, and labor migrations.

Class as a group (neoclassical economics)

The concept of class entered Anglo-American economic and social geography after World War II. The Great Depression and the war made such a profound impact on the economic activities of the time that they required special analytical and theoretical understanding. The works of the prominent German sociologist Max Weber on social stratification and social class, published in the first decades of the twentieth century and translated into English in the 1940s, became influential in modern American sociology and neoclassical economics. Influenced by Karl Marx, Weber observed social groups marked by dramatic differences in wealth and social standing. Rich industrialists and impoverished immigrant laborers, middle classes and state bureaucrats, large masses of destitute homeless—all were emerging in a society being remade by the seemingly unruly forces of capitalism. According to Weber’s theory of social stratification, these groups possess drastically different fortunes which are shaped by differential access to economic opportunities provided by the market (called economic class), social status, and access to political power. Weber saw economic class as a major factor in the stratification of social classes. This factor, in contrast to more vague notions of status and power, can be relatively easily measured in terms of income.

At the same time, the economy was increasingly explained in terms of neoclassical economic theory and as an entity governed by its own objective internal laws. These laws, which economists aimed to discover, guided the invisible hand of the market in sorting people into groups based on their economic performance. This concept of class as a group of people with particular characteristics of income has dominated social sciences for decades and is still widely used today in reference to the upper class, middle class, and lower class, for example. In geography, this concept led to prolific research on social and residential differentiation that sought to identify homogeneous neighborhoods using social area analysis techniques and, with the advent of computation, so-called factorial ecologies. Measures of income were often combined with those of educational attainment,
household size, and race/ethnicity. Using census data and surveys, geographies of class were calculated and mapped for research into housing markets, access to schools and other services, and even, to a degree, the creation of foundational urban models (concentric, sectoral, and multinuclear). These models explained spatial and socioeconomic differentiation as a result of a tradeoff between land values and commuting costs contingent on advances in transportation. Mobility between classes (and therefore urban neighborhoods) depended on a household’s ability to raise higher income through participation in the labor market (usually as a function of education or entrepreneurial acumen). Class mobility had a spatial component, most prominently expressed in the relocation of upper and then middle classes to the suburbs (where money can buy more land and a better house and absorb commuting costs) and, in recent decades, back to the rapidly gentrifying inner city. Such census-based empirical analyses have received another boost in the current era of mapping, which allows a combination of advances in geographic information science, geocomputation, and big data analysis. Explanation, however, now draws on theories of class rooted not only in neoclassical economics but in Marxism, feminism, poststructuralism, and their many combinations.

Class as structural relation (Marxism)

Although Marx developed his theory of class much earlier than Weber, his legacy took hold in Anglo-American geographical theory considerably later. Marxian theory was introduced into geography most prominently by David Harvey, Doreen Massey, Dick Peet, Bill Bunge, and others in the 1970s and 1980s, as part of the critique of the post-World War II quantitative revolution and neoclassical economics. Class is a central concern of Marxism but its meaning differs from that of neoclassical economics. Marx viewed class as a structural relation that informs the whole of human history but is hidden from direct empirical observation. Classes are distinguished by their position in relation to ownership of the means of production instead of as groups of individuals with similar socioeconomic characteristics. The source of wealth lies in exploitation, which results from the appropriation of surplus from those who produce it but who do not own the means of production. Under capitalism, the means of production includes factories and other modern ways of manufacturing material and nonmaterial commodities, and the two main classes are capitalists, who own the means of production, and workers, whom they employ for wages and whose surplus they appropriate and use for capital accumulation. Tied into an exploitative relationship, classes are in tension and their struggle defines the workings of the capitalist economy and development more generally. Accordingly, Marxist scholars reject market-based causes as valid explanations for the making of economic geographies. Instead they advance explanations tied to the logics of capital accumulation and the dynamics of class exploitation. Class mobility, therefore, depends on the ability to gain control over the means of production and can be achieved by joining the capitalist class or by working toward social change which would transfer this control to workers themselves.

Marxist geographers have theorized space as integral to the process of capital accumulation and class relations. Thus, David Harvey (2001) advanced a theory of spatial fix as a means of temporarily resolving— or of always delaying—the crisis of accumulation. He argued that capitalism must search for new geographic markets, into which to invest surplus capital
in order to avoid losing it. In this way, new territories and resources enter the circuits of capitalist accumulation, a process that has become increasingly globalized. Similarly, surplus capital can be invested into (re)building infrastructure within capitalist space once an obsolete urban fabric becomes a new investment frontier (through, for example, suburbanization, urban redevelopment, and gentrification). Both geographical expansion and the remaking of the built environment offer only a temporary spatial fix; capital never rests securely and continues to produce new capitalist space and, by extension, class relations. Neil Smith (2008) refuted the tendency toward equilibrium espoused in neoclassical economics and argued that uneven spatial development is a necessary condition for creating new frontiers for capital. The production of the built environment, and its destruction, channel investment into real estate, including suburbanization and, most recently, gentrification, during the recurring crises of capital accumulation. Thus, Marxist geographers mainly focus on the production side of capital accumulation and examine geographies of the working class primarily as part of this process. Gordon’s (1978) explanation of urban form, for example, is that it is an attempt by capital to retain its control by spatially dispersing the working class. In Marxist analyses of class politics in the Fordist era, the working class was largely assumed to consist of unionized male and white industrial workers, an assumption challenged by feminist developments in understandings of class.

**Class as social reproduction (feminism)**

Since the 1960s, Marxian theory has been taken up, critiqued, and radically transformed by feminist scholars. While they allied with the idea of class exploitation as the cause of economic and social inequality, feminists have challenged the omission in Marxist analyses of (working-class) women in waged work. They have also challenged the conceptualization of social reproduction – the space of women’s unpaid work in patriarchal societies – as secondary to the sphere of production in which the exploitation of the (male) working class takes place. Feminist scholars have argued that unpaid women’s work in social reproduction is as important for capital accumulation as waged manual work and that this unpaid work has been the source of the specific exploitation of women by capital and by their household members. Early theoretical debates focused on the relationship between two structures of exploitation: capitalism as the exploitation of workers and patriarchy as the exploitation of women. In the course of theorizing this relationship, feminist scholars have developed a gendered critique of capitalism that links class exploitation to other dimensions of experience such as gender and extends it beyond the workplace: capitalists exploited working women as well as working men, while capitalists and working-class men also exploited women in the household. Feminist scholars, therefore, placed or positioned social reproduction alongside production, profoundly changing the meaning of class.

Feminist geographers have specifically focused on the relationship between class, gender, and space. They have shown that the gender gap in wages, for example, is also related to differences in commuting time between men and women and that commuting time also differs for women with different class and racial backgrounds (McLafferty and Preston 1991); that the gendered nature of the workforce plays a major role in shaping regional economies (McDowell and Massey 1984); that access to work outside the home is mediated by unequal
domestic responsibilities (Hanson and Pratt 1995); that, within the workplace, women have systematically been undervalued and underpaid (McDowell and Massey 1984; Wright 2006); and that women are generally more likely than men to engage in the informal economy. More recently, feminist scholars have also problematized the conflation of women’s unpaid domestic work with emotional support and care (Atkinson, Lawson, and Wiles 2011) and argued for the recognition of women’s agency as a means of empowerment.

Another important contribution of the feminist rethinking of class is that it has enabled scholars to see class experiences as embodied and class subjects as empowered actors (Mitchell, Katz, and Marston 2004). The embodied class subject stands in stark contrast to neoclassical economics, which regards labor as an input and class as an income category. It also contrasts with Marxist theory which highlights the power of capital over workers and rarely recognizes the agency of workers in the production of capitalist landscapes. By creating an embodied and empowered class subject, feminism has also opened class analysis to further theorization through the lens of black, postcolonial, and poststructuralist theory.

**Racial economy and class**

Antiracist, black, and postcolonial scholars have drawn on Marxism to understand the political economy of racial exploitation while also complicating the Marxian concept of class. They argue that colonialism and slavery have been instrumental in enabling large-scale capital accumulation in Europe and the United States. Structural racism, in other words, has been a necessary condition for the rise of capitalism. The exploitation of colonies supplied the capital needed for expanded capital accumulation during industrialization, and secured a food supply for impoverished and overexploited European industrial workers and peasants. Not only did slavery make plantation economies prosper in the southern parts of the United States, Latin America, and the Caribbean, but slave labor in the American North supported most middle-class households, urban economies, and city governments in crucial ways. Moreover, modern capitalism continues to thrive because of continuing racial exploitation. Examples include concentrated poverty in urban ghettos in the United States, where people of color have no prospects of work but experience the ever-present threat of incarceration; continuing institutional racism at the workplace and in the housing market; and the exploitation of numerous immigrant workers, many of whom are undocumented, and who have fled economic devastation, war, or gang violence in their home countries (Wright 2006; Gilmore 2007; McKittrick 2011). In short, capitalism moves surplus from black and brown bodies to capitalist owners and to larger segments of white waged workers, who often fail to recognize their white privilege and exclude people of color from their solidarity struggles for wages and benefits. Capitalism, of which a racial economy is an integral part, can be thought of as a thoroughly racialized class relation at scales from local to global.

**Intersectionality and class as a process**

The constitution of class by the social relations of gender and race can be illuminated using the concept of intersectionality. Articulated in the 1980s by black feminist scholar Kimberlé Crenshaw to explain black women experience
as defined by both race and gender, it has gained in popularity in the social sciences. Intersectionality draws on poststructural concerns with decentered identity, that is, when the identity of an individual or a group is simultaneously shaped by class, gender, race, sexuality, nationality, and other relevant power relations. As these intersecting registers of social experience can be understood only in their interaction, intersectionality provides an entry point into the analysis (Valentine 2007).

This thinking is clearly related to earlier feminist theorizing of class and gender, as well as race and gender, as inseparable or mutually constituting relations. But an explicit reworking of class and gender through the lens of poststructuralism occurred in feminist geography in the works of Julie Graham and Katherine Gison who wrote together under the pen name of J.K. Gibson-Graham (1996; 2006). Drawing on a Marxist anti-essentialist concept of class, J.K. Gibson-Graham do not see class as a group or structural relation but as a process that leads to the production and appropriation of surplus. Class processes occur when any kind of work takes place – as a waged relation within a factory or as unpaid domestic work within the household – because class exploitation takes place when those who produce the surplus do not control its distribution, and in a patriarchal household men typically appropriate the surplus produced by women through cleaning, cooking, and caring. Patriarchy, therefore, generates a class process, albeit one that is different from waged employment in a factory.

Another crucial consideration of class as a process is that people can participate in more than one class process, and can be exploited as well as exploit others (e.g., a waged factory worker who appropriates his family’s surplus at home). Capitalism and patriarchy can interact with other exploitative but different class processes such as modern slavery (i.e., forced labor, human trafficking for sex work, and the enslavement of ethnic and religious groups). The bottom line is that, while capitalism is dominant discursively and in terms of traditional measures of wealth, modern societies are complex and embrace many types of class processes including those with the potential for progressive class politics.

In sum, poststructural class analysis seeks to identify a range of class processes within capitalism and to explain how they are mutually constituted with social relations of gender, race, sexuality, nationalism, and so on. In this respect, J.K. Gibson-Graham’s approach and intersectionality go hand in hand but Gibson-Graham explicitly use Marxist concepts of class and exploitation as a point of entry while intersectionality may or may not be concerned with class. Gibson-Graham’s “diverse economies” approach also aims to advance transformative class politics and eliminate class exploitation.

Class in the neoliberal era and changing labor geographies

The era of post-Fordism or late capitalism has been associated with the rise of neoliberalism, which promotes privatization, globalization, and the deregulation of markets at all scales. These developments have profoundly changed the nature of economies and posed new challenges to class theories and analyses. Unions representing working-class, white-collar, and professional workers have long, and with some success, advocated for secure employment that provides financial stability for working families and communities. Yet one of the most striking developments is the shift away from permanent full-time employment, with a living wage and benefits as a golden standard of the modern economy, and its replacement by
temporary, part-time, low-paying jobs with no benefits (Standing 2011). In many cases, even the remaining full-time jobs no longer pay a living wage. Those who work multiple jobs but still remain in poverty join the growing ranks of the “working poor” while the relatively secure middle class continues to shrink.

Highly paid professional occupations in banking, law, information technology, and higher education are also moving rapidly toward temporary contracts that no longer provide job security. Temporary labor agencies now supply highly skilled professionals in addition to low-skilled service workers. Information technology specialists, graphic design artists, and lawyers get hired for a project and are fired when the project is terminated. In many North American universities, adjunct professors now teach the majority of courses at a fraction of the salary of a full-time professor. While professional workers usually have more resources than low-paid service workers to weather periods without work, the point is that job security and wage levels across sectors and occupations have declined. The trend amplifies as corporations push for the further deregulation of employment and justify reductions in wages and benefits by global competition. The decentralized global production networks that rely on multilayered subcontracting have fragmented employment to a high degree.

As well as the nature of work, both the composition and the geography of workforces have also changed. Today’s working class is no longer white, male, and unionized but largely female, nonwhite, and fragmented. It includes Sun Belt workers in the United States, immigrant and minority workers in the Global North, and outsourced industrial and service workers in the Global South (Wright 2006). And, while notions of the working class have expanded to embrace the greater numbers and diverse kinds of working people, the power of class politics in the last several decades has declined because it has been centered on a male industrial working class.

Is class still relevant?

This question raises the possibility that class may not be relevant in the neoliberal world, where both work and workers are fragmented along lines of gender, race, nationality, occupation, and other dimensions. Neoliberalism, to some, has made it impossible for working people to connect and articulate collective demands on corporations and capitalism. With traditional class politics linked to unionism in decline, no other leverage is in sight. Some see the subjection of working people to corporate power as almost complete. The only resolution to unfettered domination and exploitation may be the inevitable, large-scale class war that may replicate forms of militancy in armed struggles against colonial regimes (see Harvey 2014).

Yet others think that, although traditional working-class politics is declining, neoliberal restructuring has created not only new sites of exploitation but new class subjects as well. As disenfranchised workplaces extend to household economies, agricultural fields, urban neighborhoods, biotechnology labs, university campuses, and global production networks, new class subjects come into being. Domestic workers, unemployed college graduates, service and construction workers, immigrants, teachers and university professors, lawyers, and information technology specialists, peasants and farmers, and subcontracted garment workers across the globe can now all fill the ranks of the working class. These diverse class subjects have different levels of education; work in manual, service, and professional occupations; and can be of any racial group or gender identity. But they all...
share the condition of the increasingly precarious employment and income that, therefore, makes vital new class alliances and transformations that can potentially bring these new workers security and control over the economy. Social theory struggles to understand the nature of impending class transformations, with one frequently invoked scenario being the rise of the “precariat” – economically and culturally precarious and politically ambiguous proletarian masses of all walks of life (Standing 2011; Munck 2013; also see Harvey 2014). At the same time, the diversification of class subjects may provide an opportunity to broaden progressive class politics to build alliances across places, scales, and cultures and to cultivate imaginations of the shared future. The major challenge is how to build these alliances and make them work. In short, class analysis and politics have not lost its relevance; class remains a central concern.

Economic crisis, the problematic of class, and new class struggles

Class relations have indeed gained new attention during the protracted and continuing global economic crisis triggered by the financial collapse of 2008 in the wake of the pervasive foreclosure crisis in the United States. The Occupy Wall Street movement has somewhat waned since then, but it achieved astonishing success in bringing the problematic of class into the center of public debate, albeit a problematic framed more in liberal than in Marxian terms. Liberal discourse usually tolerates economic inequality under capitalism as a price for economic efficiency and growing overall prosperity, but the extreme concentration of wealth becomes counterproductive to capitalism as it interferes with growth. This reframing of class inequality as an impediment to economic development has rung bells across the political spectrum, and produced a feeling of disillusionment with capitalism as the principal model of development. At the same time, it has created room for debate about new economic policies and economies that are different from capitalism and that could lead to more sustainable social, economic, and ecological futures. It is vital, therefore, to turn to theories that begin to point to the ways in which progressive class transformations can occur.

Primarily concerned with the analytics of class, J.K. Gibson-Graham grounded their vision of transformative class politics in feminist and poststructuralist theory. Feminist theory helped to expand the scope of class processes into households and communities and to advance the possibility of decentralized but widely spread class transformations. These transformations are theoretically possible when the economy is read in poststructuralist terms as constituted by multiple economic processes, of which capitalism is only one. Some already existing noncapitalist practices, such as cooperative enterprises or egalitarian households, may act as the seeds of progressive class transformations, leading to the development of community economies. The latter are economic practices at all geographic scales, from household to international, that are guided by ethics of cooperation, care, and mutual support, and in which the participants collectively create and appropriate a surplus. On the one hand, seeing the economy as diverse instead of homogeneous and capitalist fragments of capitalist space and diminishes the power of capitalism over societal practices more generally. On the other hand, it creates and multiplies opportunities for progressive class transformations. Once in the realm of public imagination, the already existing alternatives to capitalism, such as economies of cooperation and solidarity, may stimulate creative engagement and the creation of new progressive class processes.
While J.K. Gibson-Graham have most prominently advocated for the possibility of transformation here and now (1996; 2006; Gibson-Graham, Cameron, and Healy 2013), other conceptions of class also work to advance the politics of possibility. Thus, Andrew Herod’s (1997) insistence that workers, in addition to corporations, actively shape economic landscapes can be pushed further to argue that workers can exercise their agency to produce noncapitalist landscapes by, for example, reorganizing social production based on cooperative ownership of enterprises, land, and housing.

Intersectionality also helps in understanding the nature of class subjectivities as embodied, gendered, and racialized. This proliferation, instead of being seen as a barrier to solidarity, can be regarded as multiplying opportunities for both resistance and progressive class transformations. Thus, households can work to eliminate domestic exploitation; transnational immigrant households can make economic demands of national governments based on the impact of their remittances; fishermen can set up community-supported fisheries instead of working for corporate boats; indigenous people may work to rebuild their communal economies even in neoliberal contexts; and domestic workers can exercise their solidarity by using the Internet to overcome dispersed workplaces. Working people can join collectively owned enterprises at greater scales, which would help to increase worker security, pool resources around collective housing, and use the financial services of credit unions instead of private banks. New community economies can also form around the commons that ensure collective control over land, oceans, housing, and urban space (Gibson-Graham 1996; Pavlovskaya 2013; Gibson-Graham, Cameron, and Healy 2013; Roelvink, St Martin, and Gibson-Graham 2015).

While the politics of resistance retains its importance in the face of the continued consolidation of neoliberal control over the economy and politics, there is also a need to cultivate the capacity of diverse working classes to enact progressive class transformations. This capacity is already evident in the growing international movement of solidarity and social economies, economic democracy, and other creative innovations coming together at the World Social Forum (Miller 2005). These movements seek to build alliances between economic actors who pursue ethical goals (e.g., social justice, workplace democracy, cooperative ownership, and environmental sustainability) instead of profit maximization and competition. By forming noncapitalist production networks with each other, they have the economic power to grow and to nurture new solutions and imaginations.

Despite the fragmentation of class politics by neoliberal policies around the globe, the relevance of class as a dimension of social experience and category of analysis has not diminished. The concept of class has been evolving to account for the fragmented nature of work and the embodied nature of class processes. The experiences of diverse – in terms of gender, race, ethnicity, gender, occupation, economic sector, nationality, and so on – labor forces are in need of new theoretical insights that can articulate opportunities for progressive class politics here and now, in multiple economic contexts, and across geographic scales.

Their differences notwithstanding, notions of class in different human geographic traditions are not mutually exclusive; they may all usefully address specific aspects of class problematics. Census-based socioeconomic variables are helpful for constructing the changing geographies of class differentiation, while Marxist concepts of class as social relations help to focus critique on growing capitalist class exploitation and
consolidation of economic power as causes of this differentiation. Made ever more powerful by analytics of gender and race, class analysis is now able to account for complex social experiences and to be more in tune with the rich fabric of daily life while poststructuralist interventions open class to progressive transformations here and now.

SEE ALSO: Feminist geography; Gender; Gender, work, and employment; Globalization; Intersectionality; Labor geography; Marxist geography; Poststructuralism/poststructural geographies; Race, work, and employment; Radical geography

References


CLASS


Further reading


Climate adaptation/mitigation

Gordon A. McBean
Western University, Canada

As human activities have evolved and expanded, they have had, and continue to have, impacts on the climate system through emissions of greenhouse gases (GHGs) and aerosols and changes to Earth surface characteristics which affect the climate; conversely, there also continue to be impacts on humans. The economic disaster losses from weather- and climate-related disasters have generally increased over the past decades, ranging from a few US$ billion in the 1980s to above $200 billion in the period since the turn of the century (IPCC 2012). Economic, including insured, disaster losses are higher in developed countries while fatality rates and economic losses expressed as a proportion of gross domestic product (GDP) are higher in developing countries.

The objective of the United Nations Framework Convention on Climate Change (henceforth Climate Convention) is (Article 2): “…stabilization of GHG concentrations in the atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system. Such a level should be achieved within a time frame sufficient to allow ecosystems to adapt naturally to climate change, to ensure that food production is not threatened, and to enable economic development to proceed in a sustainable manner.” The Intergovernmental Panel on Climate Change (IPCC 2014) has concluded that for crop yields there have already been more negative than positive impacts of a changing climate.

The Climate Convention addresses the two fundamental policy responses to the risks associated with climate change: mitigation and adaptation. Mitigation, which means actions to reduce GHG concentrations in the atmosphere, generally through reductions in emissions to the atmosphere or the enhancement of uptake of GHGs from the atmosphere, has been the central focus of most national and international policies concerning climate change. Climate change adaptation is defined as adjustment in natural or human systems in response to actual or expected climatic stimuli or their effects, which moderates harm or exploits beneficial opportunities. It should be noted that in the disaster risk management field the word mitigation means structural and nonstructural measures undertaken to limit the adverse impact of natural hazards, environmental degradation, and technological hazards, which makes it closer to adaptation in terms of actions.

Projected climate changes

To guide actions there is need for enhanced information on future climate change impacts at a scale necessary for decision-making (NRC 2010). Over the next few decades, the range of Intergovernmental Panel on Climate Change (IPCC 2013) scenarios does not produce discernibly different climate change outcomes, with projected warming of about 0.2°C per decade. After 2040, the projected temperatures diverge so that by 2100, for the range of scenarios and including the uncertainties, the warming is between 0.3 and 5.5°C. For most of the scenarios, warming continues for centuries to follow. Note that these are global mean
changes, and temperature changes over land at higher northern latitudes will be at least twice as great. Sea level rise is a critical parameter for the billions of people who live on or near the coasts. For the same scenarios, the IPCC (2013) projects sea level rise in the range of between 0.4 and 0.76 m by 2100. With the uncertainties, the sea level rise could be close to a meter by 2100.

One important issue is very much the possibility of abrupt changes in the climate system. As noted in the National Research Council (NRC) report *Abrupt Impacts of Climate Change: Anticipating Surprises* (NRC 2013), there is need to pay increasing attention to the possibility that at least “some changes will be abrupt, perhaps crossing a threshold or ‘tipping point’ to change so quickly that there will be little time to react.” It is noted that the rapid decline of Arctic sea ice over the past decade due to warmer polar temperatures was not really projected and there is evidence of very rapid changes (over a few decades or less) in the past with major impacts on ecosystems. Since the rates of warming over the past few decades and projected for most of the rest of this century are “probably as fast as any warming event in the past 65 million years,” there are increased risks of species extinctions. The destabilization of the West Antarctic Ice Sheet, which would result in raising sea level by tens of meters and more, with huge impacts, is still an abrupt change of unknown probability.

**Response strategies**

This entry is about mitigation and adaptation as responses to a changing climate, recognizing that adaptation is also necessary for a naturally varying climate and, as such, has been underway to some extent since the beginnings of life on this planet. Actions for climate adaptation, with adequate support, usually result in benefits for the short and medium term and locally to globally. Because of the global nature of the climate system, the mixing of GHGs, and the delayed response of the climate system, mitigation benefits are delayed and are hence long term; further successful mitigation requires global agreement or, at least, agreement of all significant emitters.

**Mitigation – emissions reductions**

Although overall GHG concentrations are rising, now at a greater rate than in the 1990s and reaching 400 ppm in 2013, there are in place effective policies and economy-wide packages of policy instruments for reducing GHG emissions in different sectors and many countries. Substantial technical and economic potential exists for the mitigation of global GHG emissions over the coming decades. An illustrative way of looking at changes of emissions in the future is (IPCC 1996):

\[
\Delta CO_2 \text{ emissions} = \Delta (\text{pop}) \times (\Delta (\text{GDP/pop}) \\
\times (\Delta (\text{energy/GDP}) \times \Delta (\text{CO}_2/\text{energy}))
\]

Here, pop is population and GDP is gross domestic product (wealth by many definitions). Hence, changes in future emissions will depend on growth in population, per capita wealth, energy intensity (how much energy is needed to produce the wealth), and carbon intensity (what fraction of the energy is derived from fossil or carbon-based energy sources). The main focus needs to be on how to reduce energy intensity and the carbon intensity of the energy being generated. There are important human choices through changes in lifestyle and behavior and management practices.

Under the Climate Convention, countries, depending on their capacities, are required to mitigate climate change through limiting GHG emissions and protecting and enhancing
sinks and reservoirs, and are required to update national inventories of GHG emissions and removals. Mitigation can be applied to different sectors, such as energy supply and demand, transport, buildings, industry, agriculture, forestry and waste management, or across the economy. Some techniques for carbon dioxide removal include: land-use management to protect or enhance land carbon sinks; use of biomass for carbon sequestration as well as a carbon neutral energy source; enhancement of natural weathering processes to remove CO₂ from the atmosphere; and the direct capture of CO₂ from ambient air.

The Kyoto Protocol of the Climate Convention had an overall target for developed countries of 5% emissions reductions compared to 1990 levels over the five-year period 2008–2012. Under the Climate Convention and its Kyoto Protocol there are mechanisms for countries to work together in reducing emissions in effective ways. These include: joint implementation of projects that assist one industrialized country to meet its targets by participating in projects with another country; international emissions trading (IET) enabling the buying and selling of GHG emission allowances between industrialized countries; and the clean development mechanism (CDM), which allows public or private entities to invest in GHG-mitigating activities in developing countries and earn abatement credits. Mitigation can also be addressed through protecting and enhancing sinks and reservoirs. Examples of the latter are: reducing emissions from deforestation and forest degradation (REDD) and land use, land-use change, and forestry (LULUCF), whereby the rate of buildup of CO₂ in the atmosphere can be reduced by taking advantage of the fact that carbon can accumulate in vegetation and soils in terrestrial ecosystems.

Under the Climate Convention, Annex I parties (43 parties) are required to submit a GHG inventory according to a common reporting format. Over the period 1990–2014, total aggregate GHG emissions including LULUCF for all Annex I parties decreased by 22.6% (UNFCCC 2014). The United States never ratified the Kyoto Protocol, and Canada, which did ratify it in 2004, withdrew in 2011. Although many countries achieved Kyoto Protocol targets, others did not and non-Annex I countries did not have targets. Conditional pledges have also been made through the Copenhagen Accord. The future of these processes is not clear at this time.

Changing climatic events

The IPCC (2014) has concluded that changes in climate have in recent decades caused impacts on natural and human systems across the continents and oceans. Further, the impacts of climate-related extremes events have shown the significant vulnerability and exposure of some ecosystems and many human systems. People living in poverty have especially negative outcomes. The IPCC (2012) concluded that economic losses from weather- and climate-related disasters have increased, noting there is large spatial and interannual variability. The economic losses associated with weather, climate, and geophysical events are higher in developed countries but the fatality rates and economic losses expressed as a proportion of GDP are higher in developing countries. Their assessment is that the increasing exposure of people and economic assets has been the major cause of long-term increases in economic losses from weather- and climate-related disasters. As the climate changes, these pressures will increase (Burton 2008). It is important to note that strategic approaches to mitigation can have adaptation benefits and vice versa. There is at least some inherent capacity to cope with variable weather and climate in every nation and community but the changing climate threatens to exceed this “coping range.” As the climate...
changes, there will be changes in the frequency, intensity, spatial extent, duration, and timing of extreme weather and climate events; these may come together in unprecedented ways (IPCC 2012).

Based on climate model projections (IPCC 2012, 2013) for the rest of the twenty-first century the frequency and magnitude of warm daily temperature extremes will increase while cold extremes will decrease. A 1-in-20-year hottest day is likely to become a 1-in-2-year event by the end of the twenty-first century in most regions. Although there are other factors, the increases in hot days will add to impacts on human health (IPCC 2014). Health impacts will also occur due to storms, air pollution resulting from more hot days, and other factors.

Although changes in total precipitation will vary, extreme precipitation events over most of the mid-latitude land masses and over wet tropical regions will very likely become more intense and more frequent by the end of this century. The monsoon regions are likely to expand in space and time, with monsoon winds likely to weaken but the monsoon precipitation likely to intensify due to the increase in atmospheric moisture. Heavy rainfalls associated with tropical cyclones are likely to increase with continued warming. Increases in heavy precipitation will result in a 1-in-20-year annual maximum daily precipitation amount likely becoming a 1-in-5 to 1-in-15-year event by the end of the twenty-first century. These changes in extremes are already having their impacts and additional actions are needed (ICLR 2012). Another issue identified by IPCC (2012) is the combination of heat waves, glacial retreat, and/or permafrost degradation leading to high mountain phenomena such as slope instabilities, movements of mass, and glacial lake outburst floods.

With regard to the global frequency of tropical cyclones, the numbers will either decrease or remain essentially unchanged, but there will be a likely increase in both global mean tropical cyclone maximum wind speed and rainfall rates. The frequency of the most intense storms will increase substantially in some basins. These projected precipitation and temperature changes imply possible changes in floods. With mean sea level rise, these higher winds will lead to even greater risks of coastal storm surges, which are a major cause of fatalities and damage in coastal communities, such as with Hurricane Sandy in 2013, and particularly in tropical small island states.

The possibility of abrupt climate changes and abrupt climate impacts also needs to be considered in adaptation policies (NRC 2013). The National Research Council recommends that action be taken to improve society’s ability to anticipate abrupt climate changes and impacts through the development of an Abrupt Change Event Warning System (ACEWS). This could allow for the prediction and possible mitigation of such changes before their societal impacts become severe.

Adaptation strategies

The impacts of climate hazards on a community or ecosystem depend not only on the climate threat but very much on their exposure and vulnerability, which are the key determinants of disaster risk and of impacts when risk is realized (IPCC 2012). The occurrence of certain weather or climate events may affect community vulnerability to future extreme events by modifying resilience, reducing coping capacity, and increasing sensitivity to the stresses they impose, thus reducing the capacity to adapt to these stresses. Climate adaptation policy is a course of action chosen by public authorities to mandate or facilitate adjustments to practices,
processes, or structures, aimed at gaining benefits and reducing current and future impacts of a changing climate while also linking it closely to disaster risk management. The central goal of adaptation policy must be to reduce exposure and vulnerability. Adaptive capacity refers to the ability to adjust practices, processes, or structures to moderate or offset potential negative impacts associated with climate change (Burton 2008). A community with a greater capacity to adapt is less vulnerable to negative impacts associated with climate change. The core of countries’ capacity to meet the challenges of observed and projected trends in exposure, vulnerability, and weather and climate extremes will depend on national systems (IPCC 2012). The IPCC (2014) lays out adaptation principals for managing future risks and building resilience. These adaptations will be place and context specific and can be enhanced through individuals’ and governments’ complementary actions. Noting that climate-resistant pathways need to be consistent with sustainable-development trajectories, additional research may be needed on the socioeconomic costs of extreme events and the savings and benefits of adaptation strategies to influence decision-makers.

With the broad implications of climate change, the involvement of many government organizations, including industry, economic development, and trade, is important so that the resulting development of the climate change adaptation strategy can be “mainstreamed” into all public- (local, subnational, national, and international development) and private-sector policies and practices (IPCC 2012). Climate change adaptation actions will need to address the human impacts and disaster losses in sectors with the closer links to climate, such as water, agriculture and food security, forestry, health, and tourism. In these sectors there are also potential benefits from a changing climate, such as longer and warmer growing seasons enabling the growing of some crops. With the increased risk of heavy precipitation events and possible high wind conditions, there are risks to infrastructure that can be addressed through building codes and municipal planning. Early warning systems are important for reducing impacts on humans.

The scientific basis for concern about a changing climate is strong, and humans will have to adapt to a changing climate (McBean 2015). Adaptation challenges which are already very large will be much greater if stronger actions on mitigation are not taken (IPCC 2014). Some ecosystems will not be able to adapt and there is also the risks of abrupt changes. Research questions include how to develop and implement optimum adaptation and mitigation strategies, including the factors that will influence public and private decision-makers to undertake them. There is a need for all societies to more aggressively address these issues.

SEE ALSO: Climate change policy; Climate and societal impacts; Geoengineering/climate intervention; Global climate change; Global climate models; Intergovernmental Panel on Climate Change (IPCC)

References


Climate and societal impacts

Colin Polsky
*Florida Atlantic University, USA*

Climate shapes the behavior of all living organisms. Variations or changes in climate should therefore affect human and ecological processes. Such effects are broadly known as “climate impacts.” This concept is intuitive and widely shared among biophysical and social scientists, and the general public. However, from a research perspective, definitively classifying a given change in social or environmental systems as a climate impact is not as straightforward as it might seem. For example, paleoclimate reconstructions suggest a significant warming, and then cooling, in Greenland ≈ 1000–800 years BP. These climate variations appear to correspond with the timing of the colonization, and then abandonment, of the region by the Norse (D’Andrea et al. 2011). It is therefore tempting to infer that the climate produced specific impacts on humans by making habitation tolerable, or not tolerable. Similar claims of climate controls on humans and ecosystems have been made for many other cases. In the Norse case, the warming (and subsequent cooling) presumably affected the population by significantly enhancing (and then degrading) the local habitats for livestock and agriculture.

Yet an observed correlation between a change in climate and a change in human or environmental systems does not necessarily imply causation. For example, variations at a given point in time in the abundance of wildlife or fish, or in political-economic conditions, might be unlinked to climate, and yet contribute to the observed changes thought to be the product of climate. Climate should therefore be analyzed in the context of other contemporaneous processes. In general, any statement about climate impacts needs to be evaluated by testing alternative, non-climate explanations for the human–environment changes hypothesized to be linked to climate (see Wisner et al. 2004). Despite the challenges in classifying whether or not specific outcomes in human and environmental systems are the product of climate variations, there is a compelling and growing need to do so. The promise of anthropogenic climate changes in the coming decades due to atmospheric loading of greenhouse gases means that understanding the specific conditions under which a given climate change will negatively (or positively) impact people and ecosystems is a pressing research and policy need. This encyclopedia entry begins by providing a brief overview of recent climate impacts scholarship and ends with a review of some observed climate impacts trends and future research needs.

The roots of climate impacts scholarship

Climate impacts research is not new. This field has examined both how people affect climate and how climate affects people and ecosystems. The human role in modifying climate is not limited to the case of modern atmospheric accumulations of greenhouse gases. Other human–environment
CLIMATE AND SOCIETAL IMPACTS

interactions (such as when deforestation modifies local radiation and water budgets) have long been known to shape local climates (see Marsh 1965/1864). The direct impacts of heat waves, drought, storms, and so forth, on human and environmental systems have also been a focus of study for generations.

By contrast, the more systematic study of changes in meteorological parameters over space and time is a relatively new, and growing, science. The development of climatology as a discipline – and by extension, the allied field of climate impacts research – had to wait until enough meteorological variables had been measured for a sufficiently long time, and in a sufficiently large number of locations, that comparisons of trends and anomalies had sufficient statistical power to make reliable predictions and inferences about the trends and underlying physical processes. Thus despite the long history of interest in climate impacts, systematic studies linking environmental and social outcomes to climate – as opposed to links with non-climate contemporaneous trends or events such as wars or earthquakes – begin to appear only in the twentieth century when meteorological instrumentation networks had generated sufficient data for synoptic study.

Unfortunately, some of the early scholarly climate impacts studies (e.g., Semple 1911) were used to support discriminatory policies or practices by, for example, claiming that human characteristics or achievements were strongly or even necessarily linked with climate. Typically the inference was that higher temperatures retarded human, social, and cultural progress. This logic suggested that tropical societies were limited by climate in their potential to develop. These arguments supported racial discrimination. Such explicit “environmental determinism,” which was generally grounded in opportunistic case studies, is now absent from mainstream research, even if some contemporary scholarship continues to attribute a definitive and independent role to climate impacts on economic development (e.g., Nordhaus 2006).

Much recent climate impacts scholarship draws on a comprehensive conceptual-analytical framework that emerged in the second half of the twentieth century. The “environmental risks and hazards” approach examines how social and biogeophysical processes interact to produce specific adverse environmental and social outcomes. This field has largely examined environmental phenomena not framed in climate change terms, such as earthquakes, floods, landslides, and even nuclear waste or other technologically related hazards. To the extent that climate or meteorological phenomena were examined (e.g., tropical storms, droughts), the motivation was grounded in the phenomena on their own merits, that is, without respect to climate change.

Yet following the growing concern in recent decades of scientists and policymakers about global warming, this risks-hazards framework has been adopted for the systematic study of climate dynamics and the associated potential implications for terrestrial, freshwater, marine, and human outcomes in an anthropogenically warmed world. In this framework, three dominant human–environment processes are highlighted for explaining why groups of people exhibit varying impacts from environmental, including climate, processes: land use, emergency response, and adjustment control.

Land use refers to activities that individuals and groups of people undertake to support the provision of their basic needs and luxury wants. At a high level of abstraction, agriculture and urban settlement are examples of land use. In the “climate impacts” context, the land-use system is important to understand because these activities can introduce social dependencies on climate. Farming by definition introduces social needs
for specific levels and timings of precipitation and solar insolation for production to reach economic levels. Understanding land use is also important in the climate impacts context because land use can increase (or decrease) social and ecological exposures to climate variations. For instance, urbanized landscapes tend to enhance nighttime temperatures and convective storm development by modifying local energy budgets due to the replacement of vegetation with impervious surfaces.

Second, the emergency response process is the means by which people or goods are rescued from immediate harm following exposure to an environmental hazard, and then provided short-term relief to mitigate the most pressing negative impacts. The plight of many elderly New Orleans residents exposed to the flooding associated with Hurricane Katrina illustrates how this process can succeed, or fail. Finally, the process of long-term adjustment control describes the human decisions, often but not always regarding land-use planning, that are, or could be, implemented with the explicit aim of reducing the incidence of future adverse climate impacts. Relevant to this domain of climate impacts research is the (at times paradoxical) role of government policies (e.g., Burby 2006) and the expected recovery time for a city following exposure to a major climate/weather event (e.g., Kates et al. 2006).

Recent climate impacts scholarship and some notable climate impacts

The field of climate impacts research was arguably codified for a global audience in the 1980s. In 1985, the landmark SCOPE (Scientific Committee on Problems of the Environment) report Climate Impact Assessment: Studies of the Interaction of Climate and Society was published. Then in 1988, the World Meteorological Organization joined with the United Nations to create the Intergovernmental Panel on Climate Change (IPCC; see www.ipcc.ch). The IPCC is charged with reviewing the most recent scientific literature on how climate affects human–environment systems. The IPCC makes no policy prescriptions but instead cleaves to evaluating the scientific research on global climate processes and associated drivers, impacts, and feedbacks. The first IPCC report presented a tentative conclusion about the presence and causes of contemporary and future climate conditions across the globe, with more speculations than definitive statements about climate impacts. By contrast, the most recent fifth IPCC report, finalized in late 2014, argues unambiguously that climate change is underway, that it is produced by human loading of atmospheric greenhouse gases, and that, consequently, a suite of impacts spanning terrestrial, freshwater, marine, and human domains can now be observed and should be expected to continue for the coming decades – and for longer duration if significant emissions reductions are not achieved soon.

Given this evolution of increasingly confident (and high-profile) statements about global climate change from IPCC, the volume of information about specific trends in climate impacts on human and environmental systems, and associated policy and other response options, has grown significantly in recent decades. Periodic accounts are found in multiple scholarly journals, such as the Bulletin of the American Meteorological Society, and Annals of the Association of American Geographers. The US federal government has also produced a series of accounts of trends in climate, climate impacts, and associated response options, in the form of the recurring National Climate Assessment, published by the US Global Change Research Program.
Climate and Societal Impacts

Select findings from this rapidly growing collection of conclusions include the observations that climate has already contributed to an 80% decrease in Caribbean coral cover in recent decades and to increasing wildfire extent across the western United States. Significant changes in biological and ecological properties, such as individual fitness, population dynamics, distribution and abundance of species, and ecosystem structure and function, have also been observed in recent decades across the United States. Recent warm and dry conditions in the Colorado Rocky Mountains have resulted in increased insect outbreaks. Glaciers and ice sheets continue to shrink in many locations worldwide, with direct effects on runoff and downstream water resources. A host of social and economic impacts on human health and infrastructure is also being observed, and acted upon, by governments (local, state, federal), the private sector, and community groups.

Future research needs

This growing catalog of observed and projected impacts of climate change on human and ecological systems suggests that the research template has been set for the coming decades. Specifically, future climate impacts research needs to center on adding to this database of climate impacts across space and time. Yet there is at least one additional research frontier in the climate impacts domain: adaptive capacity. Beginning with the third IPCC report, scholars have increasingly turned to arguing that even though scientific understanding of climate dynamics (i.e., the processes that generate exposures to climate variations or changes) and of short-term socio-ecological responses to the exposures (i.e., the sensitivities to the exposures) is rapidly improving, much less well understood, by contrast, are the longer-term capacities of these systems to adapt such that the exposures and/or sensitivities are eventually diminished.

The term “vulnerability” has been adopted to refer to this three-dimensional concept: not only the exposures and associated short-term sensitivities, but also the longer-term adaptive capacities that dictate how well the system can learn from specific events by reducing future negative impacts from repeated climate exposures (Turner et al. 2003; Figure 1). Accordingly, the climate impacts literature has experienced a renewed interest in the adaptive capacity process (what the earlier literature termed “long-term adjustment control”), resulting in a considerable number of publications in recent years.

However, the insights from the recent scholarship on climate-related adaptive capacities (as well as on climate exposures and sensitivities), as valuable as they are considered individually, may be difficult to generalize. The wide variety of processes involved in socioecological adaptation means that finding a common set of metrics meaningful to all systems has proven elusive to date. Systematizing the study of climate impacts, adaptive capacity, and, by extension, vulnerability, will require translating a growing library of results of varying natures into comparable qualitative statements that permit decision-makers of all kinds to plan for the future based on our understanding of the past. The analytical means for advancing on this front exist (Polsky, Neff, and Yarnal 2007), but has yet to be applied systematically to the climate change impacts domain. A generalization of the factors that lead to greater, or lesser, climate impacts is as needed now as ever in the United States. The recent tragedy of Hurricane Sandy is just one of what appears to be an increasing incidence of weather events the impacts of which are measured in the tens of billions of dollars. Hence the growing call among urban planners, politicians, private
CLIMATE AND SOCIETAL IMPACTS

Figure 1  Climate and societal impacts. Source: Turner et al. (2003). Reproduced by permission of National Academy of Sciences © 2003, National Academy of Sciences, USA.

Businesses and citizen groups, and academics to transition human–environment interactions to less vulnerable/more sustainable modes of development.

SEE ALSO: Climate change adaptation and social transformation; Climate change and biogeography; Climate change communication; Climate change, concept of; Climate change and health; Climate change and land ice; Climate change and permafrost; Climate change policy; Climate change and sea ice; Global climate change; Ice sheets; Intergovernmental Panel on Climate Change (IPCC); Vulnerability; Water and climate change

References


Climate change adaptation and social transformation

Karen O’Brien
University of Oslo, Norway

Climate change is among the most complex and challenging problems facing humanity. The impacts are being observed around the globe through increases in temperature; changes in the amount, timing, and intensity of precipitation; warming of the oceans; reduced sea ice in the Arctic; shrinking of glaciers almost worldwide; sea level rise; and changes in extreme weather and climate variability. These changes have widespread yet differentiated consequences for human society, with implications for health, livelihoods, settlements, and security. They also influence the number, geographical range, migration patterns, and seasonal activities of individual species, and the composition and distribution of ecosystems. Continued emissions of greenhouse gases will cause further warming and long-lasting changes in the climate system, increasing the likelihood of severe, widespread, and irreversible impacts for people and ecosystems (IPCC 2014).

Reducing the risks of climate change requires human responses at all levels of society. Both mitigation and adaptation have been the focus of climate change policies in international negotiations and through regional, national, and local strategies and actions. Mitigation refers to the reduction of concentrations of greenhouse gases in the atmosphere in order to limit future climate change, while adaptation refers to the process of adjusting to actual or expected climate and its effects. These are seen as complementary responses, as there is no doubt that adaptation to observed and near-term climate changes is necessary regardless of mitigation, but it is also clear that substantial mitigation is a prerequisite for successful long-term adaptation. Yet mitigation and adaptation alone are not considered sufficient for realizing an equitable and sustainable world. There is increasing recognition of the links between climate change and other environmental and social issues, including poverty and inequality, hence the growing attention to the need for broader and deeper social transformations in response to climate change.

Climate change impacts and vulnerability

Humans have always influenced the climate, for example through land-cover changes that alter the surface reflectivity, or albedo. However, the rate, magnitude, and scale of human influence on the climate have been increasing since the onset of industrialization, through the extensive use of fossil fuel energy sources (coal, oil, and gas), land-use and land-cover changes such as deforestation, and increases in livestock production. The consequences of these activities are global and systemic. The latest assessment report of the IPCC (2014) documents numerous changes in the climate system. For example, the globally averaged temperature is estimated to have increased by 0.85°C in the period from 1889 to 2012. Each of the three decades following 1980 has been successively warmer than any preceding decade since 1850, and in the Northern Hemisphere the most recent 30-year period from 1983 to 2012 was likely the warmest in the last
1400 years. The IPCC (2014) concludes with 95% certainty that human influence has been the dominant cause of an observed warming since the 1950s.

Although human influence on the climate system is clear (see IPCC 2014), relatively little has been done globally to reduce emissions of carbon dioxide (CO$_2$), methane (CH$_4$), nitrous oxide (N$_2$O), and other greenhouse gases. In fact, CO$_2$ emissions have increased significantly over past decades and are likely to continue increasing under most business-as-usual scenarios. Unless greenhouse gas emissions are drastically reduced, the IPCC expects global average surface temperature increases to exceed 1.5°C by the end of the twenty-first century, relative to the period from 1850 to 1900. Some scenarios show a globally averaged warming of more than 3°C by the end of the century. Sea levels are expected to increase under all scenarios by 0.25 cm to 0.88 cm in the coming century, with further increases for centuries to come. Projections of future global and regional climate change are highly dependent on global emission trajectories, which are usually linked to scenarios of economic development, technological change, and population growth.

The consequences of climate change have been observed in all parts of the world, with the strongest impacts visible in natural systems. Climate change affects the global water cycle, leading to increasing contrasts between wet and dry regions and seasons (although with regional exceptions). Agriculture and food production are affected by increased variability and changes in temperature, precipitation, soil moisture, pests, and diseases. The warming of oceans influences ocean and atmospheric circulation patterns, as well as fisheries and coral reefs. Sea level rise contributes to greater storm surges, coastal flooding, and salinization of soils. As a result of the cumulative effects of past, current, and future emissions, and the thermal inertia of the warming oceans, most of these changes will persist for many centuries, even after emissions of CO$_2$ are dramatically reduced. Increases in carbon dioxide are contributing not only to climate change, but also to ocean acidification, which has implications for marine and coastal ecosystems and the human livelihoods associated with them.

The impacts of climate change on human systems are particularly visible through changes in climate variability and extreme events, which tend to reveal the social dimensions of risk, including the differential impacts within and across societies (Blaikie et al. 2003). Droughts, floods, wildfires, or cyclones can influence ecosystems and food production, damage infrastructure, affect human health, and contribute to injuries and deaths, loss of livelihoods, and population displacement.

Both increased variability and long-term climate change have significant implications for human security; in addition to decreasing food and water security, they can indirectly increase the risk of violent conflicts, serve as a “push” factor for migration, and challenge the territorial integrity of many low-lying and small island states (IPCC 2014). With more than half of the world’s population living in cities, the impacts of climate change on urban populations are of particular concern. Many of the world’s largest cities are located in coastal areas and are thus vulnerable to sea level rise, cyclones, and storm surges. The effects of heat stress are likely to be felt by urban populations, particularly the elderly and poor. Although populations or households that lack access to infrastructure and services, or live in marginal areas, are often the first to experience the impacts of climate change, the impacts on urban infrastructure (transport, waste management, water, electricity, buildings) can make large populations vulnerable to climate change and extreme events.
Adaptation to climate change

There has been a surge of attention in recent years to adaptation research, policy, and practices as a means of reducing the risks of climate change (or, in some cases, as a means of taking advantage of new opportunities). This growing interest in adaptation is tied to the recognition that climate change is already happening and affecting species, ecosystems, economic sectors, livelihoods, and human security in most parts of the world. Although strategies and actions to reduce greenhouse gas emissions can significantly influence the rate and magnitude of future climate change, society will still have to adapt to impacts over the next decades. This is because some heat-trapping gases, such as CO₂, can remain in the atmosphere for many decades, as it takes time for accumulated emissions to be reabsorbed. In addition, about 90% of the accumulated heat has been taken up by the oceans, and this heat will be released slowly because of the thermal inertia accumulated. Although there have been intensive international efforts to promote mitigation of climate change through agreements such as the 1992 UN Framework Convention on Climate Change (UNFCCC), the 1997 Kyoto Protocol, the 2014 Lima Accord, and the 2015 Paris Agreement, these political commitments alone are not considered sufficient to limit global temperatures increases to below 2°C, and many experts are suggesting that average temperatures may rise by 2.7°C or more by the end of the twenty-first century if business-as-usual scenarios persist. Adaptation is now considered an increasingly necessary part of planning and development processes, as warmer temperatures, sea level rise, ocean acidification, and other impacts have implications for sustainable development, poverty reduction, and other societal goals.

Adaptation research has expanded significantly since the mid-1990s, and includes theoretical papers that discuss definitions, typologies, and approaches, as well as case studies from diverse communities, sectors, and world regions (Smit and Wandel 2006; Schipper and Burton 2008). It has also become an important part of climate change policy debates, adding to discussions about the reduction of greenhouse gas emissions through mitigation strategies. Questions are raised in these discussions regarding the transfer of technology, the means of increasing adaptive capacity, mechanisms for financing adaptation, and the governance of adaptation processes. More recently, issues of responsibility, including compensation for loss and damages, have become an important part of adaptation debates.

Adaptation requires coordinated and complementary responses across all levels, spanning a wide group of stakeholders, including citizens, policymakers, nongovernmental organizations, and the private sector. The governance of adaptation across different interests and groups is not always clear, particularly in relation to questions of funding and finance for low-income households, communities, or nations. There have been discussions about mainstreaming adaptation into all sectors of society (Ford and Berrang-Ford 2011), including into overseas development aid (ODA) (Tanner and Horn-Phathanothai 2014). Approaches to adaptation are closely linked to understandings of the causes of vulnerability. A focus on the biophysical drivers of vulnerability usually draws attention to adaptations that minimize the impacts associated with different scenarios of climate change. Such adaptations are often sectoral, focusing, for example, on water, agriculture, energy, health, or buildings. There is a tendency for such adaptations to emphasize technical measures, which may include changes to standards and regulations, better technologies, new types of seeds and cropping techniques, infrastructure and engineering projects, as well
CLIMATE CHANGE ADAPTATION AND SOCIAL TRANSFORMATION

as early warning systems and other “climate services.” They may also include changes in behaviors and management practices, such as drinking more water during heat waves or planting crops earlier.

While adaptation projects and programs that address a specific climate risk are important (not least because they can be more easily implemented, monitored, and evaluated according to various success criteria), there is a risk that these technical adaptations will result in piecemeal solutions that address specific impacts without considering systemic factors and the underlying causes of vulnerability, or the wider consequences of adaptation, that is, the implications for other sectors, systems, or groups. There is also a risk that they will ignore or be insensitive to cultural aspects, such as traditional practices, cultural values and preferences, identities, and sense of place. Cases where adaptation in one sector or by one group exacerbates or contributes to new vulnerabilities for others, both in the present and future, are often referred to as “maladaptation.”

Social approaches to adaptation tend to focus on the underlying drivers of vulnerability, whether by improving social services and safety nets, increasing education and awareness, developing new economic instruments, providing insurance, or by identifying other policies and measures that specifically target the most vulnerable groups. Such approaches recognize that vulnerability and risk are not merely the result of the hazard but, rather, are also an outcome of historical conditions and social structures that expose people to risks or limit their capacity to respond. These conditions often include social relations linked to gender, class, or ethnicity, which are related to historical, cultural, and political contexts. From this perspective, adaptation is about changing social relations and addressing historical injustices and contextual factors that surface as “climate change vulnerability.”

However, adaptation policies and measures that focus on a broader social context can be difficult to distinguish from general development projects and processes, and may not be uniquely recognized by some as climate change policies. Alternative approaches to adaptation that address systemic causes of vulnerability are less common, and few have challenged dominant development paradigms that are associated with increasing risk and vulnerability (Bassett and Fogelman 2013).

Resilience approaches to adaptation focus on complexity, feedbacks, linkages, flexibility, and adaptive capacity of social–ecological systems. As Nelson, Adger, and Brown (2007, 412) argue, “Whereas much of the adaptation literature is focused on reducing vulnerabilities of specific groups to identified risks, a resilience approach is concerned with developing sources of resilience in order to create robustness to uncertainty and to maintain the flexibility necessary.” They show how a resilience perspective on adaptation can contribute to better understandings of both incremental and transformative systems change. Indeed, as the observed and projected impacts of climate change become more visible and better understood, it is increasingly recognized that small adjustments are, in some cases, unlikely to be sufficient responses to climate change. More recently, the concept of transformational adaptation has attracted attention within the research community. Transformational adaptation goes beyond incremental adjustments or approaches, and may include changes in form or structure through novel, large-scale actions. Examples include relocating coastal populations in response to sea level rise or shifting to new types of agriculture. Transformational adaptation may be pursued in anticipation of or in response to observed or expected impacts; it may involve coordinated or uncoordinated actions; and it may be deliberate or inadvertent. Kates, Travis, and Wilbanks (2012) describe three types of
transformational adaptations: those adopted at a larger scale or intensity; those that are novel to a region or system; and those that transform places or involve a shift in location. Transformational adaptation may be more needed for some places and systems than for others.

Although there are diverse approaches to adaptation, there has been relatively little discussion of what constitutes “successful adaptation” to climate change, especially in cases where climate change is only one among many changes that are being experienced (Moser and Boykoff 2013). Successful adaptation is a relative and subjective concept, and has spatial, social, and temporal limitations. One location or group’s adaptation may increase the vulnerability of others, in relation, for example, to flood or fire management, irrigation schemes, migration, or coastal protection. Current adaptation measures may also increase the vulnerability of future generations, particularly if they do not take into account climate change mitigation, distributional effects, and other aspects of sustainability. What appears to be successful adaptation at one point in time may be considered inadequate in the future, as climate change is characterized by nonlinear and systemic changes that include tipping points and other ecological surprises. It is recognized that there are limits to adaptation or cases where something of value will be permanently lost (Adger et al. 2009). This raises important ethical issues regarding who decides about the policies that influence not only the rate and amount of climate change, but also the types of adaptation responses that are acceptable.

Adaptation to climate change is not a neutral, apolitical process. From a critical geographic perspective, it can be considered problematic, particularly when it involves adapting to changes that are associated with existing political, social, and economic relations, perpetuated by some interests, and normalized by media, habits, and social practices. Across the emerging field of climate change adaptation, discussions and debates rarely address the full complexity of the challenge, including whose interests and ways of doing things are strengthened by promoting adaptation to climate change impacts, particularly through technical measures. When it is delinked from climate change mitigation politics and policies or from discussions about past, current, and future development pathways, adaptation can be considered an alibi for business as usual. Indeed, adaptation could be interpreted as the active promotion of passive responses to climate change, diverting attention from other possibilities, including responses that minimize risk and vulnerability by challenging existing systems, structures, and power relations. Rather than simply a technical problem that can be addressed with more expertise, know-how, and innovation, climate change can be considered as an adaptive challenge that is fundamentally linked to mindsets, particularly individual and collective beliefs and assumptions (O’Brien and Selboe 2015). Viewed from this perspective, adaptation to climate change is about integrating the personal and political dimensions of change to promote larger social transformations toward sustainability.

From adaptation to transformation

As described, adaptation is an important strategy for reducing vulnerability to climate impacts, yet rarely does it involve questioning the systems, structures, and interests that perpetuate climate change risks and vulnerability. There are, however, more progressive interpretations of adaptation that do not involve merely complying with or adjusting to changes that are already occurring or are expected to occur. For example, Pelling (2011) distinguishes between adaptation as resilience, adaptation as transition, and adaptation
as transformation. While the first two types are seen as preserving or reforming existing systems, adaptation as transformation offers the potential to fundamentally change the underlying causes of risk and vulnerability. Distinct from the transformational adaptations made in response to the impacts of climate change, adaptation can thus be thought of as a potential mechanism for progressive and transformational change that shifts the balance of political or cultural power in society, overturning established systems and imposing new regimes (Pelling 2011).

Policies and decisions associated with transformation extend beyond the status quo, and often challenge traditional ways of thinking about things, doing things, and planning for the future. As with adaptation, transformation is not a neutral, unbiased response to climate change; it reflects the beliefs, values, worldviews, and interests of different actors and institutions, and there is no doubt that power and politics influence discourses on transformation. Transformation is a concept with wide appeal, and it can be interpreted and used to promote diverse interests and agendas, ranging from transformations to a green economy to transformations through geoengineering. Thus, although transformation is widely talked about, there are many partial, fragmented, and even contradictory understandings of what needs to change and how such changes come about.

Transformation encompasses many different approaches, interests, focal points, goals, and objectives, and these can sometimes threaten traditions, interests, power structures, and social hierarchies. This in turn can trigger resistance, opposition, oppression, and fear. Because transformations often upset the status quo, it can be difficult to see or agree on their benefits, and it is not surprising that conservative responses tend to dominate. As Pelling (2011) points out, history is replete with examples of transformation processes being captured by vested interests or new elites. Importantly, “both the poor and powerful are aware of the costs of change and prefer the known even if it is a generator of risk” (Pelling 2011, 97). Thus, even as climate change and other environmental and social changes radically transform the world, there is a tendency to avoid deliberate transformations to sustainability for fear of change.

Transformation is nonetheless considered by many to be a necessary means of responding to environmental challenges. However, it is important to recognize that there are many theories of social change based on different understandings of the role of human agency and its relation to social and political structures, and different conceptualizations of human–environment relationships. A materialist understanding of social change considers physical matter as the fundamental and only reality, and thus tends to view change as causal manifestations or results of physical properties and processes. Consciousness, or subjective experience, is seen as a product of brain interactions, and free will and intentionality thus have no real role in social change. Interpretivist and constructivist understandings of social change, in contrast, recognize the significance of ideas, meaning, and perceptions, including how power and interests forge or perpetuate relationships that hinder or facilitate change. Alternative nondual understandings of nature–society interactions have fostered new approaches to engaging with social change (see, e.g., Castree 2005). It is important, however, to recognize that dominant discourses and paradigms still tend to define both problems and solutions, “naturalizing” the types of transformation that are considered possible and desirable.

For example, often the transformation of human behavior is promoted as the most important climate change response. Pointing to the individual, numerous efforts have been
made to understand, influence, manipulate and “nudge” behaviors toward sustainable practices. A focus on “attitude, behavior and choice” has been criticized by Shove (2010) for ignoring the underlying systems of provision and the extent to which behavioral options and possibilities are structured by institutions and governments and social practices. A body of research on transformations to sustainability focuses on social-technical transitions, drawing attention to the need for changes in energy, transport, agriculture, and other systems. This research emphasizes the importance of learning processes, adaptive management, innovation, and experimentation, yet also points to the limits to rapid transformative change because of deep structures and the lock-in of technologies. Such approaches tend to be consistent with ecological modernization, which seldom challenges dominant political structures and economic paradigms centered on competition and growth.

Many “transformative” responses to climate change ignore the role of politics and power in perpetuating business as usual. Swyngedouw (2010) is critical of the nonpolitical and non-partisan nature of climate change responses, particularly its implicit acceptance of capitalism and a market economy as the only possible organizational structure for the social and economic order. This, he argues, comes at the expense of a “politics of the possible” and a naming of alternative socioenvironmental futures that introduce difference, conflict, and struggle. The repoliticization of climate change is thus seen as an important means of envisioning, enacting, and realizing alternative futures. Yet repoliticization is difficult when dominant paradigms downplay the role of human and political agency in change processes. The structural changes or changes in activity considered necessary to reduce greenhouse gas emissions, for example, include improvements in reductions in greenhouse gas emissions intensity, more efficient use of energy, behavior and lifestyle changes, and improved carbon sinks (afforestation, bioenergy and carbon capture and storage) (IPCC 2014). While these may be important, the scenarios pay little attention to individual and collective agency, social movements, and the potential for radical social change. One challenge for successful climate change responses involves integrating recent research from the natural sciences, social sciences, and humanities to develop new narratives of both adaptation and transformation. Integrated approaches are more likely to view successful adaptation to climate change as a transformation to sustainability rather than as a response to the inevitable impacts of a changing climate.

Alternative futures that are equitable and sustainable represent key challenges for this century. Such futures are unlikely to come about without the questioning of many of the assumptions that are explicit and implicit in current development pathways and practices; without challenging many of the beliefs, interests, and loyalties that maintain power relations that support the status quo; and without changes to science itself. There are tensions between adapting to climate change and transforming the very systems and structures that contribute to risk and vulnerability. While both can be considered necessary, a critical understanding of the relationship between adaptation and transformation is needed to respond to the complex challenges of climate change. All transformations associated with climate change responses are likely to involve a combination of technological innovations, institutional reforms, behavioral shifts, and cultural changes. However, successful responses to climate change are likely to be those that question the paradigms and patterns of thought that normalize climate change and its impacts, and instead recognize the role of humans in collectively shaping the future.
through conscious social transformations. These are likely to involve changes in both science and society, and to depend on individuals and groups who are willing to critically engage with adaptation through social transformation.

SEE ALSO: Climate adaptation/mitigation; Climate and societal impacts; Global climate change; Global environmental change: human dimensions; Social movements; Sustainability science; Vulnerability

References


Further reading


The Intergovernmental Panel on Climate Change report that the “warming of the climate system is unequivocal” (IPCC 2013, 4). Extreme weather events such as heat waves and heavy precipitation are likely to have become more frequent in many regions, while it is very likely that we are seeing fewer “cold wave” events at the global scale (IPCC 2013). While there is considerable uncertainty in climate change models, which has implications for policy, analyses suggest that these trends will continue for a considerable time into the future.

Health geographers have researched diverse issues concerning the links between climate change and human health, as reviewed, for example, by Curtis and Oven (2012). The theoretical frameworks informing this research draw on diverse approaches to health geography and are also situated in a wider interdisciplinary literature on risk and resilience. Key themes include: the physical and psychological health impacts of environmental conditions associated with climate change; the processes of mitigation, adaptation, and resilience to climate change and extreme weather events; and the socially and geographically unequal pattern of these impacts that give rise to research on environmental justice.

Changes in prevailing average temperatures and in the frequency of extreme weather events are expected to impact physical and mental health. Our understanding of these impacts is situated in the literature which views environmental risks to health as dependent both on exposure to hazard and on characteristics that influence vulnerability to the hazard.

Changes in the pattern of death and illness associated with extreme weather-related events such as heat waves and floods are expected to become more common as the climate changes. Heat waves (and associated deterioration of air quality) are associated with excess mortality and morbidity due to physical health conditions such as cardiovascular disease, stroke, and respiratory illness, especially among older people and others who may be at particular risk due to existing health problems. Changes in the prevailing climate also have the potential to impact the prevalence of mortality and morbidity due to increases in vector-borne infections and food poisoning, as well as food shortages. Although excess mortality associated with cold weather is expected to decline in future as colder periods become less frequent, it is likely that they will continue to be significant, partly because the population may in future be less well adapted to cold weather (Oven et al. 2012; Curtis and Oven 2012).
CLIMATE CHANGE AND HEALTH

Mental health may also be impaired by post-traumatic stress and losses arising from extreme weather events. This reflects the ways in which attachment to place is disrupted and undermined by the loss of places, people, and livelihoods that are important to wellbeing in psychological as well as physical terms.

Mitigation, adaptation, and resilience

Individual and collective responses to the threats to human health and wellbeing presented by climate change can be considered in terms of mitigation, adaptation, and resilience. Fundamentally, mitigation would involve action to reduce the rate and severity of climate change, notably through reduction of carbon emissions due to human activities in agriculture and industry. From a health geography perspective, it is particularly interesting to consider that the medical and pharmaceutical sector is in itself a significant source of carbon emissions. The sector is looking to become more sustainable, and there is a growing debate about how this can be achieved by changes to health-care processes and to built infrastructure such as hospitals and clinics.

Strategies to improve adaptation and resilience accept that mitigation measures, even if they are effective in slowing the rate of climate change, are unlikely to prevent or reverse it in the foreseeable future. Therefore, to protect health in the face of continuing climate change, we must also change health-related behaviors and strategies to become more resilient. Health geography concerned with climate change therefore connects with the wider geographical agenda on the social and ecological dimensions of resilience. It is important to understand the social processes influencing human knowledge, perceptions, and behaviors, at the individual and collective levels, which may contribute to vulnerability or resilience (Wolf et al. 2010). There is also a growing literature in fields including geography that demonstrates how individual and local community knowledge and action may contribute to adaptation and resilience to climate change in ways that protect health and wellbeing (e.g., Pelling 2011).

Geographical dimensions of environmental health and social justice

Much of the literature on health impacts of climate change demonstrates socially and geographically unequal effects. It also shows that the individuals, communities, and places that are most severely impacted are often those that have relatively limited resources, or are societies, mainly in the developing world, that are dependent on resources that are sensitive to climate change. Furthermore, the societies being impacted most significantly may not be contributing most heavily to the human activity thought to be responsible for accelerating climate change. This work highlights the links between this field of health geography and other geographical research concerned with sustainability and environmental justice.

Geographical discussion of climate change and health also reflects the broader interest among health geographers and other public health specialists in debates concerning global health, and conceptual frameworks offered by political ecology concerning the complex and emergent relationships between political, social, and environmental processes at different scales that influence inequalities in human experiences of climate change affecting human health (King 2010).

In order to address the challenges to human health presented by climate change, action will need to be taken at the international and national scales, as well as the local level. The future agenda
for health geographers will need to be framed by these theories informing understanding of the global health agenda, as well as through fundamental geographical inquiry into how local variations in populations and places relate to health risks associated with climate change.

**SEE ALSO:** Environment and health; Mental health geographies; Migration: international; Vulnerability

**References**


Climate change and land ice

Andrés Rivera
Centro de Estudios Científicos, Chile
Universidad de Chile

Claudio Bravo
Giulia Buob
Centro de Estudios Científicos, Chile

Glaciers provide excellent records of climate change. Samples of particulates and gases can be extracted from ice cores, from which the past composition of the atmosphere can be reconstructed. Ice cores also record temperature and moisture, and all of this information is stored in dateable, consecutive layers. The position of end moraines, past and present, can be used to discern past glacier positions, and these deposits can be dated with radiogenic isotopes to reveal the timing of past shifts in climate. The response of a glacier to climate change may be fairly immediate for alpine cirque glaciers, but a response time is usually involved, ranging from decades for valley glaciers to thousands of years for ice sheets. Long-narrow glaciers will have a longer response time than will compact ice caps of the same area. A negative mass balance may cause an immediate response in the position of the glacier snout, but a positive regime may not be detected at the snout for many years. Weekly, monthly, and seasonal variations in glaciers are especially sensitive to ablation rates and meltwater lubrication. Glacier response over years and decades is tied to mass balance, which reflects a range of climate conditions, especially as influenced by ocean temperatures and position of the jet stream. Glacier response also depends on the distribution of glacier area with elevation. As climate warms, the elevation of the snow line will progressively experience upstream migration, shifting precipitation from solid to liquid on low-altitude glaciers. In this context, a small mountain glacier with most of its area at lower elevations will experience retreat and thinning due to rain. However, a valley glacier with an important accumulation area at higher elevations will experience stability or even thickening if the maximum snowfall is also migrating upstream. Thus, glaciers with bigger and higher elevation accumulation areas may advance during periods of climate warming, whilst small glaciers without high-altitude areas will more likely tend to disappear if ongoing climate trends persist over time.

Scientists have incomplete geographic coverage for monitoring long-term glacier changes, especially for cold-ice glaciers in the high latitudes. The Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) provided enough evidence of human-induced climate change, especially in connection to the generation of greenhouse gases such as CO₂, which has increased 40% since 1750 (IPCC 2013). According to the same IPCC report, worldwide glaciers continued to shrink in the past decade, as revealed by time series of changes in glacier length, area, volume, and mass. In a global context of glacier area reduction, an outline of glacier changes taking place in the extra-tropical Andes of Chile and Argentina since 1985 is presented.

The selected glaciers are located from the northern arid zone down to the humid and cold
southern tip of South America (Figure 1), and their areal changes have been mapped between 1985/1986 and the most recent cloud-free available satellite image (Table 1).

Methods

Variations in glacial areas have been obtained through a comparison of several satellite images, acquired for all the selected glaciers. The glacier outlines have been obtained by analyzing false color composite scenes produced from Landsat Multispectral Scanner (MSS), Thematic Mapper (TM), Enhanced Thematic Mapper (ETM), and Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) bands. The outlines were manually digitalized using ArcGIS commercial software.

The error of the glacial area calculation was estimated by multiplying the pixel size, or image resolution, by the perimeter of the digitized polygon. The average relative error for each zone varies between ±0.2% and ±12%.

Finally, the temperature reanalysis data from NCEP-NCAR (National Centers for Environmental Prediction-National Center for Atmospheric Research) is compared with the observed glacier area changes. NCEP-NCAR reanalysis data corresponds to monthly mean values at the pressure level close to the altitude of each glacier area (Table 1). The annual mean and anomalies estimations have been calculated with respect to the 1977–2006 mean.

General overview of Chilean glaciers, climate, and climate changes

The water cadastre (Dirección General de Aguas) of the public works ministry (DGA 2013) of Chile is the governmental institution in charge of updating the Chilean glacier inventory. Recent figures they have for the total number of Chilean glaciers is 24 114 with a total area of 23 460 km² (2013). This inventory includes any ice body bigger than 0.01 km², including rock glaciers, glaciarets, debris-covered ice, and clean ice. These glaciers are distributed from 18°S down to 55°S and experience a contrast of climatic conditions, from areas with only a few millimeters of precipitation per year along the Atacama Desert in northern Chile, to more than 3000 mm along the islands facing the Pacific in the southern part of the country (Quintana and Aceituno 2012). Glaciers located in northern Chile, at high altitude, are characterized by cold ice, little summer melting, high sublimation, and very low ice velocities. Located in a dry environment, these glaciers are more sensitive to precipitation changes, with sublimation being the main process defining ablation; therefore, these glaciers are particularly sensitive to changes in atmospheric humidity which affect the ratio between sublimation and melting at the ice surface. Conversely, glaciers located in southern Chile, under wet and cold conditions, are more sensitive to temperature changes because ablation is mostly dominated by melting. These characteristics of the ablation process define somewhat the different sensitivity of glaciers located at the same latitudes in the Andes in response to orographic effects.

Changes in temperature have a double influence on glacier mass balance: first, they affect the amount of energy available for the melting process; and second, the air temperature determines whether precipitation falls as snow or rain, defining the amount of accumulation during the hydrological year. However, the effect of changes in air temperature depends on the 0°C isotherm position, since an increase in air temperature will not affect the mass balance of a glacier until the freezing line reaches the elevation of that glacier.
Figure 1  Location map of selected glaciers.
For instance, in northern Chile, where glaciers are far above the 0°C isotherm, an increase or decrease in temperature probably would have a scarce effect on the mass balance; by comparison, a change in precipitation is likely to strongly affect the net mass balance of such glaciers.

**Recent climate change in Chile**

Several studies indicate that air temperatures in Chile have warmed in recent decades. However, the spatial pattern of these temperature changes exhibits a considerable variability. In situ radiosonde and satellite information from 1979 onwards indicates a marked cooling (−0.20°C/decade) along the coast of north-central Chile between 17°S and 37°S, and a marked warming (0.25°C/decade) inland and over the Andean slope (Falvey and Garreaud 2009). In southern Chile (38°S to 48°S), temperature trends over land are weak and do not show a clear identifiable spatial pattern (Falvey and Garreaud 2009). By comparison, in the Chilean Lake District (38°–42°S) the surface temperature trend exhibits a cooling in the second half of the twentieth century; nevertheless, in the middle to upper troposphere, radiosonde data from Puerto Montt show a warming trend between 0.019°C a⁻¹ and 0.031°C a⁻¹ (Bown and Rivera 2007). This warming is demonstrated by the increase of the 0°C isotherm elevation in central Chile, both in summer (200 ± 6 m) and in winter (122 ± 8 m) (Carrasco, Casassa, and Quintana 2005). All these observed data suggest that mid-troposphere warming is the main cause of glacier retreat along the Chilean Andes.

---

**Table 1**  Glacier areas.

<table>
<thead>
<tr>
<th>Glacier areas (number of studied glaciers per area)</th>
<th>Region</th>
<th>Glacier type</th>
<th>Area changes Initial (km²)</th>
<th>Area changes Final (km²)</th>
<th>Mean altitude (m asl)</th>
<th>Studied period</th>
<th>Total % of area change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maranceles and Tronquitos (7)</td>
<td>Northern Chile</td>
<td>Mountain</td>
<td>15.9</td>
<td>9.4</td>
<td>5263</td>
<td>1985–2013</td>
<td>−41</td>
</tr>
<tr>
<td>Volcán Marmolejo and San José (5)</td>
<td>Central Chile</td>
<td>Ice-capped volcanoes</td>
<td>22.3</td>
<td>21.1</td>
<td>4653</td>
<td>1986–2013</td>
<td>−5</td>
</tr>
<tr>
<td>Sierra Velluda (8)</td>
<td>Chilean Lake District</td>
<td>Mountain</td>
<td>19.0</td>
<td>11.4</td>
<td>2666</td>
<td>1985–2014</td>
<td>−40</td>
</tr>
<tr>
<td>Volcán Tolhuaca (3)</td>
<td>Chilean Lake District</td>
<td>Ice-capped volcano</td>
<td>3.1</td>
<td>1.5</td>
<td>2424</td>
<td>1985–2014</td>
<td>−50</td>
</tr>
<tr>
<td>Volcán Michinmahuida (4)</td>
<td>Northern Patagonia</td>
<td>Ice-capped volcano</td>
<td>39.7</td>
<td>36.2</td>
<td>1758</td>
<td>1985–2013</td>
<td>−9</td>
</tr>
<tr>
<td>Los Moscos and Narváez Norte (3)</td>
<td>Central Patagonia</td>
<td>Valley</td>
<td>21.5</td>
<td>18.9</td>
<td>1483</td>
<td>1985–2014</td>
<td>−12</td>
</tr>
<tr>
<td>Torres del Paine (5)</td>
<td>Southern Patagonia</td>
<td>Freshwater calving</td>
<td>423.1</td>
<td>380.4</td>
<td>1284</td>
<td>1986–2014</td>
<td>−11</td>
</tr>
<tr>
<td>Cordillera Darwin (6)</td>
<td>Tierra del Fuego</td>
<td>Tidewater calving</td>
<td>129.9</td>
<td>118.6</td>
<td>1044</td>
<td>1986–2014</td>
<td>−10</td>
</tr>
</tbody>
</table>
In Chile, precipitation trend changes are clearer south of 30°S. For instance, Quintana and Aceituno (2012) detected a negative trend in annual precipitation between 30°S and 43°S from the beginning of the twentieth century. Between 30°S and 35°S annual rainfall showed a persistent negative trend from the beginning of the twentieth century until the mid-1970s, followed by a significant increase in the 1980s; since then, the rainfall regime has exhibited a marked decadal variability in the absence of a well-defined trend (Quintana and Aceituno 2012). Between 37°S and 43°S a downward trend in annual precipitation is detected from 1950. This trend is, however, interrupted by a positive trend in the mid-1970s, registered by all the stations taken into account, except between 40°S and 43°S. In general, annual precipitation rates between 37°S and 47°S have reduced since 1950. This negative trend intensified in the last part of the twentieth century (Quintana and Aceituno 2012). According to Garreaud et al. (2013) a tendency for weaker westerlies has been detected at mid-latitudes (approx. 45°S); meanwhile, around 60°S, an increase in westerlies prevails. This tendency in the westerlies accounts for a decrease of 300–800 mm/decade over north-central Patagonia and an increase of approximately 200 mm/decade further south (50°S and beyond).

Independently from the changes in precipitation, the warming detected over the Patagonian ice field has caused a decrease in the amount of solid precipitation.

Recent decadal glaciers changes

Semiarid northern Chile

The Maranceles (28°25′S/69°40′W) and Tronquitos (28°32′S/69°43′W) glaciers are located in the dry Andes of northern Chile, at very high altitude from 4800 to almost 5800 m asl. These mountain glaciers have suffered important areal reductions, with a total loss of 41% since 1985.

Mediterranean central Chile

The Marmolejo (33°44′S/69°55′W) and San José (33°49′S/69°54′W) glaciers are contributing meltwater to the Maipo basin located in the central zone of Chile near Santiago, Chile’s capital. These glaciers are located on the western and southern flanks of active volcanoes that have not experienced any eruptive events since the beginning of the nineteenth century. The Marmolejo and San José glaciers have an extended altitudinal range (between 3500 and almost 6000 m asl), and have suffered some of the smallest changes within the studied glaciers, with areal losses of 5% since 1986.

Chilean Lake District

The glaciers in Sierra Velluda (37°28′S/71°25′W) and at Tolhuaca volcano (38°19′S/71°39′W) in the Lake District, have smaller altitudinal ranges than those in central Chile, ranging between 2000 and 3250 m asl. This is due to a generally lower altitude Cordillera, interrupted by few volcanoes rising above 3000 m asl. Sierra Velluda is partially covered by several mountain glaciers, all of them experiencing strong area shrinkage, with a total area loss of 40% since 1985 (Figure 2). The active Tolhuaca volcano is also partially covered by mountain glaciers, showing the strongest areal reduction of this study, with areal losses of 50% since 1985.

Northern Patagonia

Active ice-capped Volcán Michinmahuida (42°48′S/72°27′W) has experienced a total areal loss of 9% since 1985. This area shrinkage
Figure 2  Sierra Velluda glacier changes (Chile).
is similar to the changes observed by Rivera and Bown (2013) in other ice-capped active volcanoes of the region.

Central Patagonia, Los Moscos

In central Patagonia, the Los Moscos mountain complex (48°25’S/72°25’W) is situated at the border between Chile and Argentina. This mountain complex is covered by several valley glaciers that have experienced retreat since the first aerial photographs from 1944/1945. A total area loss of 12% has been observed since 1985. On the Chilean side, the Mosco 1 and 2 glaciers (Figure 3) were merged in 1985; however, at present, due to frontal retreat they are completely separated. On the Argentinean side, the Narváez Norte glacier was calving into a lake in 1985, but at present the glacier front is noncalving (Figure 3).

Southern Patagonia, Campo de Hielo Sur

Campo de Hielo Sur (Southern Patagonian Ice Field) is a large glacier complex spanning from 48.3°S to 51.6°S, considered the biggest temperate ice field in the Southern Hemisphere, with almost 12 500 km² of ice (DGA 2013). In the southern end of this ice field (50–51°S), at the Torres del Paine National Park of Chile and at the Los Glaciares National Park of Argentina, five glaciers located in the eastern side have been selected for this study: Grey, Olvidado, Dickson, Cubo, and Frías (Figures 4 and 5). Their historical (since the Little Ice Age) elevation, frontal, and areal changes were studied in detail by Rivera and Casassa (2004), and here all these glacier areal changes have been updated up to 2014.

The selected glaciers in this region have experienced a total area loss of 11% (42.7 km²) since 1986, a higher percentage compared to the changes obtained by Rivera and Casassa (2004) between 1945 and 2000 (8%). All of these glaciers are at present calving into freshwater lakes; however, at the beginning of the series, in 1986, the Dickson glacier had two arms, a freshwater calving front to the south and a noncalving arm to the north, which were merged together at that time with the Frías and Cubo glaciers. In 1986 only small proglacial lagoons were visible at the common debris-covered front between Frías, Cubo, and Dickson’s northern arm. Now, all are clearly separated with both arms of the Dickson glacier calving into an expanded Dickson lake, where the Cubo glacier is also calving (Figure 4). The Frías glacier is completely separated from the two other glaciers, having a debris-covered lower tongue that is partially calving into several lagoons that are progressively increasing their size. This glacier experienced shrinkage at a rate of 0.3 km²/year between 2000 and 2014, almost three times the rate recorded between 1945 and 2000, of 0.11 km²/year (Rivera and Casassa 2004).

Dickson glacier’s southern arm has been calving into the homonymous lake since the Little Ice Age (Figure 4). In total, the Dickson glacier reduced its area at a rate of 0.35 km²/year between 2000 and 2014, an accelerated rate when compared to the 0.15 km²/year observed between 1945 and 1975, 0.3 km²/year between 1975 and 1986, and 0.17 km²/year between 1986 and 2000 (Rivera and Casassa 2004).

Located a few kilometers further south, the Grey glacier is calving into its homonymous lake (Figure 5). The lower tongue of Grey has two calving fronts separated by an island where trim lines are clearly visible, showing the retreat and thinning since the Little Ice Age. Of all the selected glaciers in this study, the Grey glacier has experienced the largest area shrinkage during the observation period (a total loss of 19.2 km², or 7% of its 1986 area). Areal reduction rates reached 0.71 km²/year between 2000
and 2014, again accelerated when compared to the 0.15 km²/year observed between 1945 and 1975, 0.27 km²/year between 1975 and 1986, and 0.46 km²/year between 1986 and 2000 (Rivera and Casassa 2004).

A couple of kilometers north from Grey is the Olvidado glacier (Figure 5), which in 1986 was a valley glacier without any adjoining water body. At present, as a result of retreat, a small lake has formed next to the frontal position into which the glacier now calves. This glacier has lost almost 16% (3.6 km²) of its 1986 area, and has retreated at a rate of 62 m/year between 2000 and 2014, more than twice the rates observed between 1945 and 1975 (25 m/year) and 1975 and 2000 (29 m/year) (Rivera and Casassa 2004).

Tierra del Fuego, Cordillera Darwin

The Cordillera Darwin ice field is located in the main range of Isla de Tierra del Fuego in southern Chile (55°S). Six tidewater calving glaciers have been selected for this study (Figure 6), all of them located at the end of the Parry fjord. Most of these glaciers have been relatively stable since 1986; however, the biggest glacier (Darwin, 54°42'W/69°29'W, 41.5 km² in 2014) suffered a strong frontal retreat (almost...
Figure 4 Dickson (Torres del Paine National Park, Chile), Cubo, and Frías (Los Glaciares National Park, Argentina) glaciers. Pre-1985 data from Rivera and Casassa 2004. Reproduced from Regents of the University of Colorado.
Figure 5  Grey and Olvidado glaciers (Torres del Paine National Park, Chile). Pre-1985 data from Rivera and Casassa 2004. Reproduced from Regents of the University of Colorado.
3 km, 1985–2014). This particular frontal retreat is more likely related to the bathymetry of the fjord and is not necessarily fully connected with climatic changes. In total, the six glaciers studied here have lost 10% of their area since 1986.

Related climate change and glacier areal losses

When comparing all the glaciers studied in each zone (Figure 7), it became clear that there is a general trend of areal loss, and that the rate of loss has accelerated in recent years, especially in northern Chile. However, the largest individual percentage glacier area loss was observed in the Lake District, where the reduction trend does not show a significant variation since 1989. Overall, glacier area change rates (Table 1) show that in some zones, glaciers may disappear by the end of the twenty-first century if ongoing deglaciation continues without alteration. Conversely, glaciers in Patagonia will likely last for many more centuries, even in the worst scenario of climate change that will be discussed in the following section.

In the Chilean Andes, high-altitude climatic stations that are in close proximity to glaciers are sparse. One of the few stations with such characteristics is El Yeso (33°41’S/70°06’W, 2475 m asl) (Figure 1). Data from this station shows a clear warming trend since the 1960s (Figure 8), a trend that is also observed in Santiago (Quinta Normal at 527 m asl). However, this warming trend is highly dependent on the climate shift of the 1970s associated with the El Niño Southern
Oscillation phenomenon (Falvey and Garreaud 2009). This long-term warming trend, at high altitude in the Andes where the glaciers are located, is forcing an upward migration of the regional snowline, effectively increasing ablation and reducing snow accumulation on glaciers (Carrasco, Casassa, and Quintana 2005).

Figure 9 shows a comparison between glacier areal losses for each glacier considered in this study and the temperature anomalies obtained from NCEP-NCAR reanalysis at the approximate altitude of the glaciers (Table 1). All these time series of temperature anomalies show that the climate shift of the mid-1970s, described above and below, exhibits contrasting behaviors.

The northern zone shows a clear trend of higher temperatures in the past 15 years, which could be triggering the strong upward migration of the snowline, reducing snow on the glaciers, which would explain the area shrinkage of the Maranceles and Tronquitos glaciers. The central zone shows very small temperature increases after 1970, as observed in El Yeso (Figure 8); consequently, a small glacier retreat is understandable for glaciers on the Marmolejo and San José volcanoes, illustrating the variety of glacier behaviors in this region. Further south, where the Tolhuaca volcano and Sierra Velluda glaciers are experiencing very high areal losses, there is no temperature trend in NCEP-NCAR reanalysis data (Figure 9). Since little volcanic activity has been observed in this zone, the glacier shrinkage must be related to precipitation reductions, as detected by Quintana and Aceituno (2012). Near Michillinahuida volcano, air temperatures have shown a moderate increase since the mid-1970s, and the areal reductions must be explained by precipitation changes and volcanic activity (Rivera and Bown 2013). In the rest of Patagonia, near the Northern and Southern Patagonian Ice Fields, the air temperature trends are clearly positive, especially in the past 10 years, this trend being the more likely triggering factor in the generalized negative areal changes.

**Climate projections**

For the twenty-first century, the IPCC (2013) indicates an increase in warming and a decrease in precipitation throughout Chile. The north of Chile (from 18°S to 30°S) appears to be the most affected zone in CMIP5 (Coupled Model Intercomparison Project Phase 5) models’ mean composite for each scenario, with temperature anomalies in 2100 (with respect to 1986–2005 mean climate) from 1.2°C (rcp26) to 5.3°C (rcp85) (Figure 10). The projected temperature change for the central zone is estimated to range between 0.8°C and 3.4°C. Precipitation appears more consistent in all future scenarios, with reductions ranging from 5 to 10%, between 30°S and 43°S, without significant changes further south.

In all these future scenarios of climate changes, glacier retreats will continue, with glaciers especially affected in northern Chile, where
small and low-altitude glaciers will very likely disappear by the end of this century. In central Chile, the glaciers will be increasingly affected, as long as end-of-summer snow lines (equilibrium line altitudes) keep rising (Carrasco, Quintana, and Casassa 2005). In the Lake District, the ongoing deglaciation will accelerate if the ongoing reduction in precipitation is not reversed. In Patagonia, temperature trends will certainly increase ablation, inducing high thinning and negative mass balances, as observed in the past decade.

Conclusions

The 41 selected glaciers included in this study, located along the Andes, including a wide variety of glacier types, are showing area reductions and negative frontal changes. This trend is not uniform among the selected glaciers, with the strongest changes taking place in northern Chile and in the Lake District. In central Chile, glaciers are also retreating but at lower rates than in northern Chile. In Patagonia, the changes are important in terms of total areas, but due to the huge amount of ice in this southern zone, areal loss percentages are much smaller. In this zone, there are also some anomalies, especially among calving glaciers, some of them exhibiting an enhanced retreat compared to noncalving glaciers, and in two cases there are advances. The changes detected in this work are representative of the area reductions taking place in other glaciers along the Chilean Andes, with some of them showing much stronger area shrinkages (some in connection with volcanic activity) and some exhibiting smaller area reductions (for example on the partially debris-covered glaciers). The main driving factor explaining these generally negative changes is certainly the observed atmospheric temperature changes, especially at high altitude in the Andes. However,
precipitation variability also plays an important role. All the scenarios of climate change for the rest of this century are indicating an increment in temperatures for almost the entire country, and a reduction of precipitation along the Andes between 30°S and 43°S. In these scenarios, glaciers will experience further reductions, with the smaller and lower-altitude glaciers probably disappearing by the end of the century. In central Chile and Patagonia, the ongoing changes will be enhanced, but the great number of large glaciers, the altitudinal ranges of these glaciers, and the amount of accumulation will certainly preclude their disappearance in the long term.

Acknowledgments

The research reported here has been supported by FONDECYT 1130381 and Centro de Estudios Científicos (CECs). CECs is funded by the Basal fund of CONICYT among other grants. Ryan Wilson improved the text with his comments and suggestions. María de Los
Ángeles Gacitúa helped with figures. Richard Marston also contributed to improve the text. CODELCO Andina Division and Anglo American Chile mining companies supported this research.

SEE ALSO: Cryosphere: remote sensing; Glacier changes; Glaciers; Intergovernmental Panel on Climate Change (IPCC)

References


Climate change and sea ice

Josefino C. Comiso
NASA Goddard Space Flight Center, USA

One of the most contentious issues that are confronting society in recent years has been the influence of humans on the Earth’s climate system. The warming impact of the increase of greenhouse gases in the atmosphere was studied in the late nineteenth century by Svante Arrhenius and his results indicated that a doubling of CO$_2$ would cause an increase of about 4°C in global surface temperature. It was estimated at that time that it would take more than 3000 years for this doubling to occur. Since 1880 when accurate surface temperature measurements around the globe became available increases in surface temperature have indeed been observed. It became apparent, however, that the rate of increase in CO$_2$ in the atmosphere as observed from the 1950s at Manoa Loa, Hawaii, by Charles Keeling has been a lot faster than was initially anticipated. Current measurements indicate that the fraction of CO$_2$ in the atmosphere has increased by more than 30% since the 1880s and is already much higher (400 ppm) than the peak (290 ppm) of observed natural fluctuations of CO$_2$ (and also methane) over the last 850,000 years as inferred from ice cores. The impacts of such high rates of increase and the ability of humans to control it have been the subjects of many studies (IPCC 2014).

Globally, climate change is already manifested in many observational studies and confirmed by modeling studies to be caused primarily by increases in anthropogenic greenhouse gases. The confidence has steadily increased as more in-depth studies were made leading to the conclusion by the Intergovermental Panel on Climate Change (IPCC) 2014 that climate change as influenced by humans is unequivocal. The increased confidence compared to previous assessments has been made possible by significant advances in our knowledge of the climate system that have been strongly aided by satellite technology and the advent of high performing computer systems. Satellite sensors enabled the collection of much needed global datasets while high performance computers greatly enhanced the capability and accuracy of numerical models.

The impact of global warming to human society can be profound if not devastating. For example, warming is expected to cause an intensification of the global hydrological cycle since a warmer atmosphere holds more water. Such intensification may be occurring already since in recent years we have observed a higher frequency of extreme events exemplified by the occurrences of super hurricanes like Katrina that almost wiped out the city of New Orleans in 2005, Sandy that devastated the US east coast in 2011 and Haiyan that almost destroyed the city of Tacloban, Leyte in the Philippines in 2013. Persistent rain has caused intense flooding and landslides in England, the Philippines, China, and other regions from 2009 to 2014 while largescale droughts have been occurring in Australia and California since 2012. Moreover, higher sea surface temperature and glacial melt have caused an acceleration in the rise of sea level that in turn caused the disappearance or abandonment of some inhabited islands. Also, higher temperatures have led to persistent heat waves that have caused the death of thousands of people and have
CLIMATE CHANGE AND SEA ICE

had serious direct and indirect effects on human health, biodiversity, and the environment. There are currently still gaps in our knowledge about the linkages of the various variables and more scientific research is needed to establish the strength of the correlations. But in the meantime it is apparent that serious mitigation measures are needed to prevent even more serious impacts and possibly irreversible consequences of climate change.

Observed changes in our climate system

Surface temperature measurements from meteorological stations, buoys, and other sources made using modern instruments since the 1880s have been very useful in quantifying recent changes in the climate system. Such global surface temperature data have been compiled and quality checked to take into account errors or biases such as those associated with expanding urban areas. The result of linear regression analysis of yearly averages of global temperatures from 1880 to 2013, using data described by Hansen et al. (2010), reveals a trend in surface temperature of about 0.0650 ± 0.003 °C per decade. Such trend in temperature is considered unprecedented over the last 1000 years (Mann 2012) and the impact could be more serious than expected because the warming is not uniformly distributed spatially and could be much stronger in some areas than in others. Such spatial variability in the trends is not unexpected and could vary depending on the time period evaluated as depicted in Figure 1. Three periods were selected: (a) 1880 to 2013 for optimal record length using modern data; (b) 1964 to 2013 for an intermediate record with a more robust dataset; and (c) 1981 to 2013 to be able to make use of satellite records (Comiso 2003) that provides greater spatial coverage than the pre-satellite data. The long-term trend map presented in Figure 1a shows the most uniform distribution with a consistent warming evident almost everywhere except for small areas in the Antarctic and south of Greenland. The trend map for the intermediate period, as depicted in Figure 1b, shows stronger positive trends (than those in Figure 1a) in the Northern Hemisphere and especially in the Arctic region but more cooling in parts of the Southern Ocean and the Pacific Ocean. This period is also when the global data are enhanced with more meteorological stations and other data sources and therefore provide improved spatial coverage. The trend map during the satellite era as presented in Figure 1c represents a further improvement in global coverage including high latitude regions and shows an even more intense warming in most parts of the globe especially in the Arctic. During the same period, there also appears to be more significant cooling in many areas like the eastern Pacific Ocean region and parts of the Antarctic region. Such cooling is relatively mild compared to the warming in most other regions. The unexpectedly large areas of negative trend appear counter-intuitive but are consistent with those derived from other datasets (e.g., Reynolds et al. 2002). Analysis of the data from 1981 to 2013 yielded a global trend of 0.17 ± 0.02 °C per decade which is 2.6 times that of the 1880 to 2013 period indicating an apparent acceleration in the warming (Comiso and Hall 2014).

It is also apparent that warming is amplified in the Arctic region with the trend for the region >64°N over the period 1880 to 2013 being 0.153 ± 0.011 °C per decade while that for the period from 1981 to 2013 period is 0.60 ± 0.07 °C per decade using the Hansen et al. (2010) data. Using AVHRR satellite data, the trend for the same region from 1981 to 2013 is 0.67 ± 0.06 °C per decade which is a bit higher because of improved coverage especially in the mid-Arctic region where the trends are
Figure 1  Global surface temperature trend maps for the periods (a) January 1880 to December 2013; (b) January 1964 to December 2013; and (c) August 1981 to December 2013 using data as described by Hansen et al. (2010). The trend map in Figure 1(c) makes use of AVHRR data as described in Comiso (2010) for polar regions generally higher than 55° North or South.
CLIMATE CHANGE AND SEA ICE

higher than average. The warming is much more modest in the Antarctic region as indicated in Figure 1 likely in part because of large differences in the geographical distribution of land and ocean areas and different atmospheric circulation patterns in the two hemispheres. The warming trend in the Antarctic (>64°S) is estimated to be $0.037 \pm 0.014$ °C per decade using the 1881 to 2013 data while the trend for the period from 1981 to 2013 as derived from satellite AVHRR data is about $0.072 \pm 0.038$ °C per decade. The aforementioned warming is expected to continue in the foreseeable future unless the emission of greenhouse gases into the atmosphere is controlled. Since CO$_2$ has a residence time in the atmosphere of a few centuries, it will actually take a long time (after a reduced emission of greenhouse gases) for the surface temperature to stabilize and reverse its course.

The generally warming trends are clearly manifested in multiple and independent observations in the atmosphere, ocean, land areas, and the cryosphere. The number of warm days and nights has increased on a global scale while the frequency of heat waves has increased in large parts of Europe, Asia, and Australia. Ocean salinity measurements have also indicated that the contrast between wet regions and dry regions over the ocean has increased in the last 60 years suggesting changes in precipitation patterns. Increases in the frequency, intensity, and amount of heavy precipitation have also been observed in land areas while the number of extremely intense tropical cyclone activities has been going up. The characteristics of the ocean which covers more than 70% of the Earth’s surface and where about 60% of the net increase in energy is stored are also changing significantly. The heat content of the ocean up to 700 m in depth have been observed to be increasing with the upper 75 m layer warming by about 0.11 °C per decade (IPCC 2014).

Overall, warming signals are most apparent in the polar regions and in particular in the various components of the cryosphere. This is mainly on account of the amplification of climate signals in the polar regions associated with ice-albedo feedback (Holland and Bitz 2003). A large fraction of solar radiation is usually reflected back to outer space because of the high reflectivity of snow and ice that covers the region. But as the snow and ice cover retreats, more heat from the sun gets absorbed by the exposed surface that in turn cause the surface to be warmer causing further retreat of the snow and ice cover. An acceleration in the loss of ice mass in land has been observed with the rate of loss in the Greenland ice sheet increasing from 34 Gt per year from 1992 to 2001 to about 215 Gt per year over the period 2002 to 2011 while the rate of loss in Antarctic ice sheet changed from 30 Gt per year over the period 1992 to 2001 to 147 Gt per year for the period 2002 to 2011. Meanwhile, the rate of loss of ice mass in the glaciers increased from 226 Gt per year over the period 1971 to 2009 to 275 Gt per year for the period 1993 to 2009 (IPCC 2014). The snow cover in the Northern Hemisphere has also been declining in spring from 1967 to the present at the rate of 2.17% per decade (Comiso and Hall 2014) and at a much faster rate of about 12% per decade during the month of June. The permafrost temperature has been increasing by about 2 °C to 3 °C in parts of Alaska and Russia while the discontinuous permafrost has been advancing to the north. The most dramatic change, however, especially as observed during the satellite era, is the large decline in the perennial ice cover or the ice that survives the summer melt in the Arctic as discussed in the following section. Sea ice is a key component of the climate system and drastic changes in the ice cover could profoundly change the state of the climate.
The polar regions have been referred to as the heat sink and changes in the sea ice cover could have a profound effect on the global heat balance and hence the weather and the climate (Fletcher and Kelley 1978). In winter, sea ice keeps ocean heat from being released to the atmosphere while in summer it limits the amount of heat from the sun that is absorbed by the ocean. Sea ice also redistributes salt in the ocean in that it releases low salinity water where it melts in the spring and summer while it enhances salinity through brine rejection where it forms in autumn and winter. The melt of sea ice in spring and summer has also been associated with the occurrences of ice edge blooms and high productivity near the ice edge regions because the layer of low density meltwater becomes an ideal platform for photosynthesis since the layer stays at the surface where sunlight is abundant. Also, the growth of ice in coastal polynyas and other regions causes the production of high salinity and high density cold water that sinks to the bottom and becomes part of the global ocean thermohaline circulation. Furthermore, the marginal sea ice region is regarded as the site of strong ocean–atmosphere interactions and is usually visited by intense storm activities, known as polar lows, and cold-air outbreaks.

Satellite passive microwave sensors have provided much of what we now know as the seasonal and interannual variability in the global sea ice cover. These sensors are able to provide day or night almost all weather coverage for the sea ice cover in both hemispheres on a daily basis. The resolution of the sensors is relatively coarse (around 12 to 25 km) but algorithms have been developed to estimate the concentration of ice within the field-of-view of the sensor. Other parameters can also be quantified through the use of the multichannel data. The sea ice concentration data have been used to assess the spatial distribution of the sea ice cover and to locate areas of divergence, polynya formations, and other features of the ice pack (e.g., occurrence of an Odden, or ice tongue, in the Greenland Sea). The geophysical parameters that have been used to quantify changes in the sea ice cover are the sea ice extent and sea ice area. Sea ice extent has been defined as the summation of the area of all data elements with concentration of at least 15% ice concentration while sea ice area is the summation of the product of the ice concentration and corresponding area of each data element. The ice extent provides the means to assess how the ice covered ocean has been changing, while ice area has been used to estimate the volume and hence the ice mass, knowing the average thickness.

The sea ice cover is very seasonal with the Arctic ice extent changing typically from $6 \times 10^6$ km$^2$ in summer to $16 \times 10^6$ km$^2$ in winter while in the Antarctic the corresponding change in sea ice cover is from 3 to $19 \times 10^6$ km$^2$. With more than 34 years of satellite data it is now possible to assess not just the interannual changes but also the trends in the sea ice cover. Because of the large seasonality, it is useful to use monthly anomalies as shown in Figure 2 for trend studies. The monthly anomalies were derived by subtracting the climatological average (average of data from 1978 to 2013 for each month) from each monthly ice extent data. Figure 2a and 2b show the monthly anomalies in the Northern and Southern Hemisphere, respectively, and it is apparent that the trends are not only different but also have opposite signs. The trend lines indicate that the sea ice extent in the Arctic has been declining at the rate of $-480 000 \pm 21 000$ km$^2$ (or $-3.79 \pm 0.16\%$) per decade, while the Antarctic ice extent is expanding at the rate of $174 720$ km$^2$/decade (or $1.77 \pm 0.18\%$/decade).
Figure 2  Ice extent monthly anomalies (in black) and yearly anomalies of ice minimum (in red) for the period 1979 to 2013 in the (a) Northern Hemisphere; (b) Southern Hemisphere; and (c) combined Northern and Southern Hemispheres. Trend lines from regression analysis of the data are also shown.
For comparison, the trend in ice area (not shown) is $-527,000 \pm 20,000 \text{ km}^2$ (or $-4.5 \pm 0.2\%$) per decade for the Northern Hemisphere while that for the Southern Hemisphere (not shown) is $207,841 \text{ km}^2/\text{decade}$ (or $2.2 \pm 0.1\%$ per decade). The errors indicated are statistical error (one standard deviation) of the slopes of the trend line and do not include systematic errors that may be associated with other factors like differences in instrument calibration, resolution and incident angle between the different satellite passive microwave sensors used in this study. The impact of the latter on the trend, however, is expected to be minor because the techniques used to derive sea ice concentration accounts for such error sources (Comiso 2010).

The asymmetry in the trends for the two hemispheres is intuitively surprising but not totally unexpected because of the large regional variations in the warming as indicated earlier (see Figure 1). The big difference in the geographical distribution of land and the ice cover in the two hemispheres may also contribute to the difference in the interannual changes in sea ice distribution. The significance of the magnitude and sign of the trends in the two regions is also different. In the Arctic, the sea ice cover is primarily in the Arctic Ocean basin where the surface temperature is extremely cold and is surrounded by land that restricts expansion of the ice cover beyond the land boundary. On the other hand, the coldest region in the Southern Hemisphere is a snow covered continental area called Antarctica. The sea ice cover that surrounds Antarctica is in relatively warmer water and the extent is basically controlled by sea surface temperature. The summer ice cover in the region is also dispersed and relatively small compared to that of the Arctic and the location changes from one year to another. The impact of ice-albedo feedback in the two regions is therefore different. In particular, the observed retreat of summer ice in the Arctic causes a significant warming of the mixed layer of the Arctic Ocean while a retreat in the summer ice cover in the Antarctic is likely not as impacting because of large interannual variability in the location of summer ice and because of the more dynamic ocean environment that minimizes the change in mixed layer temperature associated with less summer ice.

It is apparent that in the Arctic, there is large monthly but small interannual variability from 1978 to 1996, followed by a relatively modest variability but consistent decline up to 2007 which was then followed by a much larger monthly and interannual variability. The large interannual variability in the recent period is associated with the large decline in the perennial ice cover which is the ice that survives the summer. The perennial ice extent has been derived from summer minimum extent (Comiso 2010) and is shown in a red dashed line in Figure 2. In the 1980s, the Arctic basin region was almost all covered by ice even at the end of summer melt because of the abundance of multiyear ice (or ice that survives at least two summers) which is the thick component and mainstay of the Arctic ice cover. In recent years, however, the perennial ice cover, represented by the ice cover during ice extent minimum in summer, has been reduced to less than half of what it was in the 1980s with a yearly trend of $-11.4\%$ per decade. The percentage decline in the multiyear ice extent is even larger which is about $13.5\%$ per decade (Comiso 2012; Parkinson and Comiso 2013) and in addition, this ice type has been observed to be thinning (Kwok and Untersteiner 2011). The loss of this thick component of the sea ice cover would make the Arctic ice at the end of the summer vulnerable and the high rate of decline suggests that the Arctic basin might become totally ice free in a few decades unless a sustained cooling in the region occurs and the sea ice cover recovers its previous thickness. The latter will not
happen if the warming in the region as indicated earlier continues. The occurrence of an ice free summer will be unprecedented since, historically, a significant summer ice cover has been observed in the Arctic for at least the last 1450 years. A disappearance would have profound effects on the climate, ecology, and environment not just of the region but also globally.

It should also be noted that the magnitude of the trend of the ice cover in the Arctic is more than two times that in the Antarctic. Global sea ice cover can be studied by simply combining the extents and areas of sea ice in the Northern and Southern Hemispheres for each month. To estimate the trend, monthly anomalies of the combined Arctic and Antarctic monthly sea ice extents are presented in Figure 2c. The results show a relatively uniform extent for total sea ice cover with a net reduction over the 34-year period and a trend of only $-0.13 \pm 0.125\%$ per decade. It is interesting that the trend of the global sea ice cover when compared with the trend of the Arctic ice cover is analogous to the trend in global temperatures when compared to that of the Arctic.

The trend of the sea ice cover in the Antarctic is relatively modest with the magnitude less than half the trend in the Arctic. Some studies indicate that the trend is related to more ice production owing to the decrease of ocean salinity in the region as a result of the melting of ice shelves. Others point out that a warming in the ocean would result in more evaporation that causes more solid precipitation over the region leading to colder surface temperature that enables higher rate in the growth of sea ice. A key observation is the increase in the dynamics of sea ice in the region and enhanced formation of coastal polynyas that serves as ice factories. Through modeling studies, it has been postulated that the increased dynamics might have been caused primarily by the occurrence of the ozone hole in recent years. The ozone hole apparently causes a deepening of the lows in the west Antarctic region which in turn leads to stronger southerly winds and higher ice production along the coastal region of the Ross ice shelf leading to the increase in the extent of the sea ice cover in the Ross Sea. Actually, the ice cover in the Bellingshausen and Amundsen Seas has been declining at a high rate but such decline is overwhelmed by the increase in the ice cover in the Ross Sea. There are also studies that indicate that the average extent of sea ice cover in the 1950s, as observed from whaling and other data during the period, was actually higher than the average extent during the more recent years as observed from satellite data. This is further supported by observations that the surface temperature in the region has increased significantly since the 1950s.

Ice–atmosphere–ocean interactions and modeling studies

Changes in the global climate and those in the polar regions are strongly influenced by processes that are occurring in the atmosphere and the ocean. In the Arctic, atmospheric circulation is believed to be basically controlled by the arctic oscillation (AO) sometimes referred to as the Northern Hemisphere Annular Mode (NAM). A similar phenomenon called the Southern Hemisphere Annular Mode (SAM) occurs in the Antarctic region. The characteristics of the AO have been studied using AO indices that represent the difference of the average sea level pressure of two Arctic regions. Previous studies indicated that the indices are highly correlated to the extents of the sea ice cover in the 1980s and 1990s. However, the correlation fell apart and became weak in the 2000s (Overland and Wang 2005), which was a time period when drastic changes in the sea ice cover were occurring.
This led some to hypothesize that the phase of the AO and the circulation patterns in the Arctic atmosphere has changed.

A very prominent sea ice feature in the Greenland Sea, called Odden, was reported in the early 1900s and was observed almost every year using satellite data in the 1980s and 1990s but the feature was basically missing in the 2000s (Comiso 2010). The bay surrounded by the Odden called Nordbukta has been regarded as one of only four regions of the world where deep ocean convection is observed. Moreover, the temperature of the mixed layer of the Arctic Ocean has also been observed to be warming in part because of a rapidly declining summer sea ice cover. The demise of the Odden and the warming of the Arctic Ocean are likely influenced by the increases in surface temperature in the Greenland Sea. In the Antarctic, a large open water region, called polynya, in the middle of the ice pack was observed using satellite data in 1974, 1975, and 1976, and caused significant cooling up to 3000 m depth in the region. Since then the polynya has not reappeared likely because of changes in climate of the region as may be associated with SAM. Although the Antarctic sea ice extent has generally been increasing, there are some regions, like the Bellingshausen/Amundsen Seas near the Antarctic peninsula where sea ice is declining at a relatively rapid rate comparable to those in the Arctic. The peninsula has also been regarded as an anomalous region since it has been experiencing warming at a rate higher than most regions of the world. Ice shelf melting has also been occurring at a relatively fast rate in the region and also adjacent regions in recent years causing significant calving. These phenomena illustrate apparent inconsistencies that may not be unexpected because of the existence of offsetting fluctuations, seesaw patterns, and teleconnections in the climate system.

Studies of the drivers of the changes in climate reveal that the largest contributor of the radiative forcing since the industrial revolution is the increase in atmospheric CO₂ followed by the increase in CH₄. Modeling studies indicate that the most dominant cause of the observed warming since the twentieth century is anthropogenic greenhouse gases (IPC 2014). In particular, the changes in the sea ice cover, in the ice sheet, glaciers, snow, and permafrost as well as the heat content of the ocean is caused primarily by human influence. This means that the future climate would depend primarily on future greenhouse gas emissions. A control of these emissions is apparently imperative to avoid dangerous if not irreversible impacts of climate change. A modeling study by Washington et al. (2009) reveals that emission reduction of about 70% is needed by 2100 to prevent half of the expected changes in temperature and precipitation from occurring.

Discussion and conclusions

Multiple indicators of climate change are observed in the atmosphere, land, ocean, and ice from instrumental and paleoclimatic sources. Global circulation models have been able to simulate many of these observed changes and to distinguish human–induced from natural changes making it possible for the IPCC to conclude that anthropogenic climate change is unequivocal. Ocean salinity measurements indicate that the contrast between wet regions and dry regions over the oceans has increased over the past 60 years while the heat content of the ocean up to 700 m depth has increased significantly. Among the strongest signals of warming are those from the polar regions where climate change is expected to be amplified and even slight increases in temperature have caused significant changes in the timing of the melt season, retreats in the snow cover, loss of
mass in ice sheet and glaciers, and thawing of the permafrost. But among the most dramatic change that has been observed on a global scale has been that of the sea ice cover in the Arctic.

The annual mean extent of the Arctic ice cover has declined at the rate of 4% per decade while the perennial and multiyear ice extents have declined at 11.5 and 13.4% per decade, respectively, during the 1979 to 2013 period. In addition, a significant thinning of multiyear ice in the Arctic basin has been observed while the length of the melt season has increased. The multiyear ice cover is the mainstay of the Arctic ice and the trends suggest that the Arctic basin could become ice free in the summer in the foreseeable future. The impact of an ice free summer in the Arctic can be profound and would include a change in the climate, ecology, and environment of the region and changes in ocean circulation. Strong mitigation measures through reduced greenhouse gas emissions are needed to stabilize surface temperature and reduce warming thereby minimizing potentially devastating if not irreversible impacts of climate change.

SEE ALSO: Glaciations; Global climate change; Global climate models; Lake ice

References


Further reading

Climate change and biogeography

David Goldblum  
*University of Calgary, Canada*

Christine Carrier  
*The Morton Arboretum, USA*

Climate change and biogeography: observations

Biogeography is the study of the distributions of species and ecosystems through time and space examined at global, regional, or local scales. Modern research in this discipline combines theories and data from biology, geology, ecology, and physical geography. Climate change, including rising temperatures, modified precipitation regimes, and altered chemical concentrations in the atmosphere and oceans are being investigated as a driver of biogeographical changes on our planet.

The Earth has witnessed a long period of variable climate over the past 2.5 million years which included multiple glacial–interglacial cycles. During the glacial periods, temperatures at the regional level may have changed as slowly as 1°C per millennium, or as rapidly as several degrees over a few decades, and included changes in precipitation and CO₂ concentrations. Numerous types of paleoecological records can give precise and detailed records about past plant species distributions and their responses to postglacial warming events.

Through research done by palynologists, fossil pollen patterns have shown that many taxa found in tundra, boreal forests, and temperate forests migrated to lower latitudes during glacial periods – spending the long glacial periods in refugia (region in which a species spends an unfavorable period). However, populations from a subset of species may have persisted in low densities in more poleward locations in cryptic refugia. As temperatures increased when global conditions transitioned from glacial to the warmer interglacial periods, vegetation undertook a poleward migration. Pollen evidence from both North America and Eurasia over the last 18 000 years, when the last glacial period began to ameliorate, shows tree species generally spreading poleward from the margins of their glacial refugia.

Given the high density of fossil pollen sites in eastern North America palynologists have created detailed species range maps covering the past 18 000 years for dozens of common tree genera (and species in some cases). From these maps it is possible to determine the speed of migration, the timing of postglacial migration events, and migration routes followed. In North America most boreal tree taxa (e.g., *Picea, Pinus*) spent the glacial period in proximity to the southern limit of the Laurentide ice sheets. The tree taxa currently associated with the present-day temperate deciduous forests (e.g., *Acer, Fagus*) were found south of the boreal taxa; and those species currently found in today’s warm mixed forests (e.g., *Liquidambar, Carya*) of the southern United States were found in the extreme southern United States (Florida). It is clear from pollen studies that tree taxa responded individually to postglacial warming, thus vegetation community assemblages formed over the
past 18 000 years as taxa moved north were often different from contemporary ecosystems, in some cases comprised of combinations of species not currently found in North America. These novel combinations of species have been termed “no-analog” communities. Similar patterns have been recorded in Europe, with vegetation shifted to lower elevations and southern refugia in the Iberian, Italian, and Balkan peninsulas.

Based on pollen range maps, rates of post-glacial poleward tree migration ranged from 100 to 400 m/year, in large part depending upon seed dispersal vector. However, based on recent chloroplast DNA surveys it is possible that some taxa maintained cryptic refugia near the ice sheets (and therefore near modern range limits). In those cases the estimated migration rate might be significantly lower than those rates inferred from pollen evidence. In examples from both North America and Eurasia, routes of postglacial migration differed by taxa, with some largely migrating up coastal plains, others migrating along river valleys, and some moving along mountain chains.

With these documented biological responses to the postglacial increase in global temperatures over the past 18 000 years, ecologists have speculated that recent and future anthropogenic warming will lead to similar ecosystem responses mirroring patterns deduced from pollen evidence — although future temperature changes are modeled to be substantially more rapid than those from the past.

Since global surface temperatures from 1906–2005 warmed by 0.74 °C, and the Northern Hemisphere has seen an earlier spring as the length of the frost-free season has increased (IPCC 2007), field studies over the past several decades have been undertaken to examine biological responses to recent climatic changes. A meta-analysis by Parmesan (2006) found that every continent, ocean, and taxonomic group has documented evidence of direct recent climate change impacts. Although general observations can be made at the species level, most research is conducted at the local scale and fewer at the regional level. There has been a geographical bias in this research with the majority of studies conducted in North America, Europe, and Russia. While there has been limited research in Africa and Asia, more information is available from South America, South Africa, and Japan. The global impact of climate change is generally acquired from independent yet similar surveys including plants, animals, insects, and aquatic life, which are compared using statistical meta-analysis to identify recurring patterns. For example, in a study of 1598 species, 59% displayed alterations to their distributions, phenologies (timing of biological events), or both in the past 20–140 years, and these changes followed the expected patterns caused by regional climate change (Root et al. 2003).

**Phenology**

Recent changes in the timing of biological events (phenology) have been well documented in light of warming climates over the past century. A meta-analysis of over 500 plant species found 78% of all springtime phenological records advancing, while only 3% were delayed. The changes include advances in the leaf out dates of deciduous species, earlier flowering, egg laying, migration, and breeding dates. Meta-analyses of phenological changes across the Northern Hemisphere find that biological spring events are occurring 1.0–5.1 days earlier per decade. Timing of fall events (e.g., leaf fall) has changed much less, with events occurring 1.3 days later per decade. Several studies have found little or no change in phenology. In some cases these studies were undertaken in areas experiencing stable temperatures over the past century, but
a few studies found no change in phenology despite warming temperatures. While documented phenological changes are most common in spring, advances in the timing of early spring events are more pronounced than late spring events. Advances in spring timing may be occurring faster in coastal areas compared to inland regions. Organisms living at higher elevations are advancing in the spring more rapidly compared to lower elevations; however, little difference has been detected in magnitude of phenological change across latitudinal gradients. Temporally, rates of phenological changes over the past few decades (since the 1970s and 1980s) are greater than in previous decades.

Migration

A meta-analysis (Chen et al. 2011) of elevational range changes found species shifting upslope at a median rate of 11.0 m per decade in Europe, North America, and Southeast Asia. In the Ural Mountains, during the last century, treeline advanced approximately 20–60 m upslope, corresponding with a doubling of winter precipitation and summer warming of 0.9 °C during the twentieth century. A study from the European Alps documented the expansion of vascular alpine plants toward the summit of 0–4 m per decade. There was an increase in species richness at 70% of the summits in the Alps that may be the result of global warming trends since the nineteenth century.

In a thirty-year period from 1974–2004, the Sierra de Guadarrama mountain range in central Spain has had a mean annual temperature rise of 1.3 °C, which has increased the lower elevational limits of 16 species of butterflies an average of 212 m. A study of Costa Rican cloud forests suggests that anthropogenic warming has elevated the cloud heights reducing moisture from cloud contact. Many plants and animals are dependent on these moisture inputs and with its decline, changes have occurred in both distribution and abundance of these species.

A meta-analysis of numerous animal taxa from Europe, North America, and South America found a median rate of poleward migration of 16.9 km/decade (Chen et al. 2011). In a study of 166 global latitudinal treeline sites, 52% experienced poleward advancement since 1900 (only 1% exhibited recession); with the probable cause linked to local winter warming. Saplings of tree species preferring higher temperatures have, over the past century, begun to germinate 50–300 km northward in northern Sweden.

A survey of more than 100 bird ranges in the United Kingdom found that from 1968 to 1991 the bird’s northern range margins shifted an average of 18.9 km. In 35 years, the sachem skipper butterfly (Atalopedes campestris) has expanded its northern boundary hundreds of kilometers from California into Washington State (Crozier 2003). The change in distribution corresponds with an increase of 2–4 °C over a 50-year period, and the butterfly’s persistence in these new areas is possible due to the warmer winters at the northern edge of their margins. A study of British dragonflies and damselflies found that all 37 nonmigrating species exhibited northward range shifts, attributable to a warming climate, from 1960 to 1995; the mean northward shift was 74 km.

While examples of poleward limits shifting to higher latitudes are common, there is much less evidence that low-latitude limits are changing due to anthropogenic climate change, particularly for woody species. Finally, it must be noted that globally elevational and latitudinal range shifts are variable, about one quarter of species analyzed moved downslope or toward lower latitudes.

General patterns

The impact of climate change on terrestrial ecosystems may encompass considerable
geographic variability. Projections of anthropogenic climate change generally suggest that warming will be greatest at higher latitudes. However, analyses of global data sets of insects, birds, frogs, lizards, and turtles suggest that future species extinction risk is greatest in tropical ecosystems. Tropical species typically have narrow thermal tolerances compared to higher latitude insects and vertebrates which have broad thermal tolerance and often are living in environments below their physiological optima, so warming may enhance their fitness.

Marine environments are also being altered as a result of a changing climate. Rising levels of atmospheric CO₂ absorbed by the oceans increases acidification which impedes marine animals from producing calcareous skeletal structures, and may directly impact the food web and species distributions. Water temperature increases of 0.79 °C between 1931 and 1996 in Monterey Bay, California led to increased abundance of invertebrate species from southern areas and decreased abundance of species from northern areas. In the North Sea, where water temperatures have risen 0.6 °C from 1962 to 2001 there have been detectable impacts to the distribution of bottom dwelling fish due to climate change. Roughly two thirds of the species exhibited a northward shift in mean latitude, and further temperature changes may influence community interactions and effect commercial fisheries in the region.

Rising sea temperatures can cause bleaching of coral reefs and frequent or severe occurrences may impact their survival at large geographic scales. Fossil records indicate that reef-building corals were once present at higher latitudes during warmer periods, and then retreated equatorward as temperatures cooled. Following this trend, a northern range expansion into the northern Gulf of Mexico and along the Florida Peninsula has been documented for two species of reef-building coral, coinciding with climatic warming and increased sea-surface temperatures. However, there is no clear evidence that the world’s coral reefs can adapt to moderate future warming. Coral reef communities may experience a number of changes; increased mortality may reduce the age structure, more tolerant coral species may replace those more sensitive to environmental stress, and coral species may be replaced completely by other groups of organisms as coral communities are unable to keep pace with projected warming.

An earlier breakup of sea ice on western Hudson Bay in Canada has been correlated with rising spring air temperatures from 1950–1990. Polar bears living at their southern range boundary in this area have demonstrated a decline in their mean body weight, as well as their population numbers. These animals hunt from the ice shelves, and as the lengthening of ice-free periods on Hudson Bay causes their feeding grounds to disappear, they are forced to live off their reserves. If these conditions persist, it is possible that the bears will move into areas of human habitation, threatening human life and property.

Multiple factors may alter species distributions in polar regions. Warming temperatures at these higher latitudes may raise soil and permafrost temperatures as seen in Alaska in the late 1980s to early 1990s, which has the potential to dramatically alter or even destroy the ecosystem as a whole. Evidence seen in northern Canada and Alaska as well as parts of Russia show that shrubs have been expanding into areas of tundra, coinciding with three decades of rising temperatures. Much of the warming has occurred during the winter months and has created a positive feedback loop, as shrubs trap and hold snow to insulate the soil. An increase in microbial activity and plant-available nitrogen allows for increased shrub establishment and growth, which may
impact animals such as caribou, which prefer tundra species.

Although accessibility in Antarctica still limits the locations and organisms selected for investigation, the establishment of research stations in the late 1950s and satellite imagery have greatly increased information available from this important area. Trends found in 244 marine glaciers on the Antarctic Peninsula show that 87% have retreated southward toward the pole and the pattern is comparable with warming temperatures in the region and it appears that the most significant changes are taking place along the coast. Penguin species in the Antarctic have been impacted due to changes in climate conditions. Adélie penguins are found throughout the Antarctic and have appeared and disappeared throughout the paleocological record in response to changing ice cover. Regional warming has reduced the extent of sea ice, thus limiting the wintering habitats for individuals on the Antarctic Peninsula. In contrast, Gentoo penguins generally establish in ice-free areas and are able to expand their breeding ranges as the changes in climate reduce the extent of sea ice and provide more suitable habitats.

On the western side of the Antarctic Peninsula, there has been a decline in krill stocks, which depend on the sea ice to breed successfully. These shrimp-like crustaceans serve as a primary food source for many species in the area and feed on phytoplankton, a community that is also changing as the diatoms are being replaced by less mineral-rich cryptophytes as their regional habitat changes. These changes may alter the food web and impact the species that are able to inhabit the area. Anticipated results of climate change in Antarctica will most likely encourage the establishment of new taxa to the region and increase their chances for survival by providing more favorable environmental conditions; however, they may also provide direct competition to local species already living at the edge of their physiological limits.

Finally, considerable uncertainty surrounds biological responses to climate change at the community level. As numerous studies have shown there is a common directional pattern in plant and animal taxa response to recent changes. However, given the complex interactions (e.g., predator–prey, plant–pollinator, etc.) that occur between species, it might be difficult to predict future ecosystem trajectories.

**Biogeography and climate change: models**

Given the numerous and growing documented cases of species range shifts over the past several decades biogeographers and ecologists have applied a long-standing interest in understanding the distribution of species toward predicting how species’ ranges might shift under future climate change. Looking to the next century, modeled climate change scenarios predict that global mean surface temperatures may increase 1.4°C to 5.8°C, a more rapid rate than the gradual warming at the end of the last glacial period, which took place over a period of 5000 years (IPCC 2007). Precipitation contrasts will likely continue to increase between wet and dry season and among regions. While global average temperatures are modeled to increase, the average values mask considerable geographic variation. Generally, temperature increases are most likely to be greatest in high northern latitudes, particularly in the winter months. Although tropical regions are modeled to see smaller increases in air temperature, the tropics are most likely to experience novel climates (lacking modern analogs). Since most ecological models are based on modern observations, predicting ecological responses to these no-analog climates might be challenging.
In recent decades biogeographers have used quantitative models to predict vegetation patterns under modeled future climate change conditions. The models fall into two broad classes: vegetation-type models (dynamic global vegetation models) and species-specific models. The vegetation-type models operate at the global scale and generalize vegetation into plant functional types or biome type, not focusing on single species or their individual range shifts. The species-specific models focus on individual species. The species-specific models are best divided into two broad categories: correlational (empirical) models and mechanistic (processed-based) models. Recently, several researchers have advocated a hybrid (combined) approach which incorporates the strengths of the correlational and mechanistic models.

**Dynamic global vegetation models**

Static global biome models were initially developed to model global vegetation distribution patterns under constant climate conditions based on eco-physiological constraints and resources limitations. The dynamic global vegetation models (DGVMs) are a natural extension of these in that they can be used to predict the dominance of plant life forms (or plant functional types) at large spatial scales under changing climate conditions over decades to centuries. These complex models typically incorporate vegetation dynamics (e.g., nutrient cycling, competition, disturbance), soil-vegetation-atmosphere transfer, biogeochemistry (e.g., canopy physiology, soil processes), land management change, and soil conditions. In part, DGVMs are designed to track vegetation changes driven by climate variability. DGVMs have proven to be useful when modeling global scale vegetation response to modeled anthropogenic climate change because DGVM output can be coupled with general circulation model (GCM) predictions. In contrast to correlational models, DGVMs do not take into account the behavior or responses of individual species, and generally don't perform well at regional scales; rather they are designed to provide predictions at the global scale, necessarily sacrificing detail. An advantage of DGVMs is they can be applied where anthropogenic environmental changes may have already altered species observed ranges or in cases where a species may be absent from a previous portion of its range due to human activity – both scenarios cause problems for correlational models.

**Correlational models**

This category of model is known by several names including: species distribution models (SDMs), empirical models, bioclimatic niche modeling, niche envelope modeling, and habitat suitability models. Correlational models identify statistically significant relationships between species occurrences and a number of environmental variables thought to predict suitability of the habitat for the species (e.g., temperature, precipitation, solar radiation, soil nutrients, and topography). Provided the species’ geographical distribution is correlated with the environmental predictor variables then the model can be employed to make spatial predictions based on maps of the predictor variables (including future climate maps). The generated maps can identify areas of potential species occurrence based on the suitability of those locations. Moving beyond single species modeling, “stacked” SDMs can combine modeled distributions for multiple species in a geographic space to create community-level predictions. SDMs require little understanding of the mechanistic (ecological or physiological) links between the organisms and their environment,
which can be advantageous if the analysis involves numerous species and/or large geographic areas. However, predictions from these models might be limited in their biological realism and transferability to novel environments. Studies have found that dispersal capacity and response to disturbance affected predictive strength of a correlational mode.

There are a number of important assumptions of these models: (i) species location data is representative of the species’ true distribution which is problematic if spatial occurrence data is incomplete or limited, (ii) species are in equilibrium with the environmental factors used in the model, and (iii) the correct environmental predictor variables are included in the model. Furthermore, other assumptions of correlational models include that the processes controlling range limits are fixed in space and time – this is not likely to be the case under future climate conditions, and that genetic and/or phenotypic plasticity of species being modeled is low. Finally, correlational models are typically defined as static models in that they do not incorporate the ability of a species to migrate to the predicted geographic space, which can be remedied by including realistic rates of species migration.

Mechanistic models

Process-based or mechanistic models focus on the role that functional traits such as physiological, behavioral, and morphological limitations and life history characteristics such as development, growth, and reproduction have on species distributions. These factors, which are not included in correlational models, can influence a species distribution and abundance, potentially constraining a species’ geographic distribution to a subset of locations that might otherwise be deemed suitable by correlational models. Mechanistic models require a large number of parameters to be estimated which may be of limited availability particularly at high spatiotemporal resolution.

Hybrid models

These models are a middle ground between the high data demands and complexity of a full mechanistic model and less data intensive mechanistic models. Hybrid models combine the static spatial predictions from correlational models with some process-based factors (e.g., dispersal, phenology, population dynamics, trophic interactions, and community dynamics). The inclusion of these factors will likely constrain predicted future species ranges compared to correlational models, and might allow for predictions of species distributions under changing climatic conditions.

Problems and limitations – DGVMs

While climate has long been recognized to be a major determinant in vegetation distributions it is possible that under future (anthropogenic) climate change conditions vegetation might not respond as it has in the past. There are numerous factors that lead to this conclusion. In the future, habitats will be more fragmented than they were in the past potentially limiting species dispersal, dispersal vector models are not fully developed, elevated atmospheric CO₂ levels may have unknown impacts on plant physiology, disturbance regimes may differ from current conditions, and biotic interactions may be altered.

As is often the case with climate or ecological models, the generalizations necessary to model global phenomena can create problems. For example, DGVMs represent competition poorly as they focus on identifying the dominant plant functional type, not accommodating species coexistence well and poorly representing
CLIMATE CHANGE AND BIOGEOGRAPHY

ecosystem heterogeneity. Additionally, based on evidence of vegetation response to Holocene warming it is clear that seed dispersal and migration of species is a critical response to environmental changes. Yet, DGVMs might not represent plant migration processes well, such that replacement of an extinct plant functional type by a new plant functional type would likely not occur immediately as modeled by DGVMs; rather a lag of several decades to 150 years might occur due to dispersal limitations or competitive interactions.

Problems and limitations – correlational models

The strength of correlational models is almost entirely dependent on the availability of correlational data. When using correlational models to predict vegetation response to climate change there is little regard for the dynamics of potential shifts from one species’ dominance or vegetation type to another. As described above for DGVMs a correlational model does not account for lags in vegetation change. Additionally, correlational models focus on the factors most highly correlated with species distributions, but it is possible that critical climate factors, which are rare or not directly measured, may not be included in the predictive model undermining the predictive power of the model. Interspecific interactions (e.g., competition) which might lead to a species being absent from geographic space within its climatic envelope complicates correlational associations and can lead to inaccuracies in predicted distributions under future climate conditions. While perhaps desirable to combine single species SDM predictions to model community-level patterns (S-SDMs), stacking results often leads to an over prediction of species richness. To resolve this, researchers have combined SDM output with macro-ecological variables and ecological assembly rules to constrain species richness predictions. Finally, novel climate conditions might develop yielding future climate conditions for which there is no current analog and therefore for which we lack correlative data.

Problems and limitations – mechanistic models

Mechanistic models require more time, effort, resources, and data than correlational models as they often require specific information about a species’ traits. As above, any geographic variation in a species’ physiological limits proves a challenge for mechanistic models. Similar to the issues described above, the parameterization of processes is based on correlative models and the causality of the observed correlations is not assured.

Quantitative models have been used to project likely impacts of climate change on species distributions and biomes since the early 1990s. DGVMs, correlational, mechanistic and hybrid models differ in both the geographic and taxonomic scales at which they’ve been applied. At global scales climate dominates distributions, whereas at finer scales factors such as topography and geology affect microclimate, nutrient availability, and water availability, which in turn influence species distributions.

Local and regional scales

Correlational, mechanistic, and hybrid models are most commonly used for modeling species shifts at scales ranging from hillslopes to continental scale and are useful for conservation planning in light of anthropogenic climate change. The primary concern with modeling over small scales, however, is the requirement of fine resolution climate and species data which is not readily available.
Myriad studies have been published modeling climate change impacts on single taxa (including terrestrial animals, plants, and aquatic organisms). Other studies are based on overlapping geographic distributions (stacking) of multiple species to assess community-level patterns (species diversity and forest type changes) likely to occur under future climate conditions. However, simply assessing species richness patterns by overlapping ranges of individual species may lead to an over prediction of species richness since no constraints are placed on species coexistence that might naturally arise through resource competition.

A DGVM analysis (Bachelet et al. 2001) of potential changes to vegetation in North America showed a general northward shift in biomes. Generally coniferous forests expanded slightly, temperate deciduous forests shifted northward, being replaced by southeast mixed forest or savanna. Tropical forests appeared as a new vegetation type in Louisiana where they replaced the northward shifting southeast mixed forest. Shrubland was lost in the Great Plains, being replaced by savanna. The greatest reduction was for the tundra biome, decreasing by more than 80% by the end of the twenty-first century.

Global scale

DGVMs have been typically reserved for continental to global scale modeling efforts where biome level patterns are of primary interest given that output is fairly coarse. Scholze et al. (2006) modeled changes in land-cover type (forest vs. nonforest) at the global scale. Depending on the level of temperature increases over the coming century there is a high risk of forest loss in Eurasia, eastern China, Canada, Central America, and Amazonia. The models suggest forest expansions into the Arctic and semiarid savannas around the world.

SEE ALSO: Biodiversity; Biogeography: history; Boreal forest ecosystems; Climate change, concept of; Disturbance in biogeography; Ecosystem services; Intergovernmental Panel on Climate Change (IPCC); Paleoclimatology; Paleoecology; Treeline ecotones

References


Further reading


Climate change and permafrost

Nadine Salzmann
University of Fribourg, Switzerland

Isabelle Gärtner-Roer
University of Zurich, Switzerland

Definition, occurrence, and basic processes of permafrost

Definition

Permafrost is a thermal phenomenon and is defined as ground (soil or rock and included ice or organic material) that remains at or below 0°C for at least two consecutive years. According to this definition, the occurrence of permafrost does not necessarily require the existence of ice.

Permafrost consists of an active layer, which is the surface layer of soil that thaws each summer after the snow cover has melted and refreezes each winter (Figure 1). The active layer thickness refers to the annual maximum thaw depth at the end of the summer. It ranges from less than a few decameters in continuous permafrost along the arctic coast, to a few meters in discontinuous permafrost (e.g., Southern Siberia), and several meters in mountain ranges such as the European Alps or the Qinghai-Tibetan Plateau.

The main permafrost body is bounded at the top by the permafrost table and at the bottom by the permafrost base (Figure 1), as a result of the balance between heat conduction from the surface (MAGST; mean annual ground surface temperature) and thus the atmosphere, and the ground thermal heat flux from the Earth’s interior. The position is determined by the vertical temperature profile of the ground and only varies over very long time periods, depending on the permafrost thickness and the ice content (Figure 1). Permafrost temperatures at depths reflect variability of climate conditions as heat diffuses slowly through the ground.

The depth of zero annual amplitude (ZAA) is where the permafrost temperature shows no seasonal variation, that is, below the depth of ZAA, long-term changes in average climate conditions are reflected. The depth of ZAA varies from a few meters in discontinuous permafrost to 20 m or more in continuous permafrost or in bedrock.

Occurrence

The occurrence of permafrost is determined primarily by the regional climatic and the site-specific surface and subsurface properties and in general is to be expected where the MAAT (mean annual air temperature) is below 0°C. About 20% (or 25 × 10⁶ km²) of the global land surface is characterized by permafrost. These areas are located mainly at high latitudes and/or at high altitudes (Figure 2). Regional permafrost occurrence is classified into three zones, based on the fraction of land area underlain by permafrost: (i) the continuous permafrost zone, where 90–100% of the area contains permafrost; (ii) the discontinuous zone, with 50–90%; and (iii) the sporadic zone, with 10–50%. Isolated occurrence is given in regions where less than 10% of the area is underlain by permafrost.
CLIMATE CHANGE AND PERMAFROST

Figure 1  Schematic view of the ground thermal regime and important terms (left) and general response times (right).

Figure 2  Worldwide occurrence of permafrost. Reproduced from GTN-P, http://ipa.arcticportal.org.
As an example of permafrost occurrence, this entry showcases the Swiss Alps. In Switzerland, about 5% of the land surface is characterized by mountain permafrost, and the permafrost bodies reach depths of a few decameters in debris (Von der Mühll and Haeberli 1990) and up to more than 1 km in the bedrock of the highest summits. Its position varies only over very long time periods, depending on the permafrost thickness and the ice content.

Low temperature conditions, as well as the occurrence of ground ice, are the most effective drivers for mechanical weathering, geomorphic dynamics, and related landform evolution (Swift et al. 2014). Pingos are the most prominent landforms in high latitudes (e.g., in North America, Svalbard), developing and persisting only in permafrost landscapes. They are characterized as ice-cored hills with a height of about 60 m and a lifetime of up to 1000 years. Closed-system pingos (hydrostatic processes) occur in lowland settings within continuous permafrost, while open-system pingos (hydraulic processes) are more common in valley bottoms and footslopes in both discontinuous and continuous permafrost. They are characterized as peaty frost mounds consisting of alternating layers of segregated ice and peat or mineral soil material. In high mountain regions, rock glaciers (Figure 3) are among the most prominent landforms and indicators for the occurrence of permafrost. They represent climatological boundary conditions with a MAAT below $-1$ to $-2 \, ^\circ C$, and annual precipitations of $<2500 \, \text{mm}$ (Swift et al. 2014). Rock glaciers commonly occur at the foot of free rock faces (talus rock glaciers) or below moraines (debris rock glaciers), and form tongue- or lobe-shaped bodies with lengths of several hundred meters. Active landforms contain ice and move at rates of $>0.1 \, \text{m year}^{-1}$, whereas inactive ones still contain ice but do not move any more. Relict rock glaciers indicate former permafrost conditions.

Key processes

The present state of permafrost occurrence (as well as its future development) is influenced by climatic conditions. Permafrost is a thermal system with a relatively slow response to climatic changes (compared, for instance, with glaciers). The response time of permafrost depends on the thermal conductivity, the ice content, and the thickness of the frozen ground. In fact, the thickness and some marginal occurrences of alpine permafrost most probably still reflect maximum Holocene cooling during the Little Ice Age, which culminated in the nineteenth century. Temperature profiles observed in boreholes have proven permafrost thicknesses in excess of values expected from present-day mean annual ground temperature (MAGT).

In general, the thermal boundary conditions of the uppermost layers of the lithosphere are determined by atmospheric energy fluxes and the heat flow from the Earth’s interior. The geothermal heat flux is relatively low and quite constant in space and time. In the context of permafrost, its variations caused by mountain topography can be neglected. As a consequence, the ground thermal regime and, in particular, its potential changes are determined mainly by the atmospheric net heat exchange at the surface, which is controlled by the local topography, by the surface and subsurface conditions, and by the ground properties.

The surface energy balance, equation 1, describes the energy exchange of heat and moisture between the atmosphere and the ground surface. According to the conservation principle,
the sum of all fluxes must be balanced for any timescale and every location.

\[ Q^* + Q_H + Q_{LE} + Q_G + Q_M = 0 \] (1)

where \( Q^* \) (net radiation) = \( K \downarrow + K \uparrow + L \downarrow + L \uparrow \) and where \( K \downarrow \) = incoming shortwave radiation, \( K \uparrow \) = reflected shortwave radiation, \( L \downarrow \) = incoming longwave radiation, \( L \uparrow \) = outgoing longwave radiation, \( Q_H \) = sensible heat flux, \( Q_{LE} \) = latent heat flux, \( Q_G \) = ground heat flux, and \( Q_M \) = latent heat of fusion.

From equation 1 it is evident that solar radiation is the most important energy source of the surface (if geothermal heat is neglected; see above). Therefore, the surface temperature at a specific location depends significantly on the amount of incoming shortwave radiation, which varies on an annual cycle with latitude and topography. Furthermore, the partitioning of the energy balance components also varies significantly due to site-specific factors, such as ground properties, snow cover regime, and so on (see below).

In general, the propagation of the ground surface temperature into the ground depends on the ground thermal properties for heat transfer.
The response of permafrost to atmospheric changes is furthermore dependent on the ice and liquid water content of the ground, which can be spatially highly variable, particularly in mountain topography. Ice and water in the ground determine the rate of temperature change by the effect of latent heat, that is, ice-rich ground slows down and damps temperature change because of heat absorbed by partial melting of interstitial ice. Ice content, however, is difficult or impossible to measure directly, and quantification must typically be based on geophysical methods such as electrical resistivity tomography or refraction seismic tomography (Hauck 2013).

Finally, glacier retreat may lead to permafrost aggradation in regions where the current MAAT still favors the development of frozen ground.

**Observed changes**

In general, permafrost is sensitive to changes in temperature and other controls, such as snow cover, ice content, and vegetation (Voigt et al. 2010). First of all, permafrost conditions are strongly dependent on atmospheric changes. In the following, a short overview of observed changes of major atmospheric parameters (air temperature and precipitation, including snow) is thus provided before observed changes in permafrost temperatures are further discussed.

**Observed atmospheric changes**

For air temperature, the past decades show a clear positive trend at the global scale, with an average increase of about 0.85°C for the period 1880–2012 (Figure 4). The warming is spatially heterogeneous and regional variations diverge in magnitudes of between about 0.2 and 2.5°C. Also note that in particular in cold regions,
Figure 4  (a) Observed global mean combined land and ocean surface temperature anomalies, from 1850 to 2012 from three datasets. Top panel: annual mean values. Bottom panel: decadal mean values including the estimate of uncertainty for one dataset (black). Anomalies are relative to the mean of 1961–1990. (b) Map of the observed surface temperature change from 1901 to 2012 derived from temperature trends determined by linear regression from one dataset (orange line in panel a). Reproduced from IPCC (2013).
meaning high altitudes and high latitudes, data coverage is usually sparse.

For precipitation there is no clear overall global trend observed. There are regions that show an increase while others show a decrease of annual precipitation (Figure 5).

For permafrost, more important than the annual precipitation rate is the annual distribution of precipitation in relation to air temperature, that is, in particular, whether precipitation reaches the ground in the form of rain or snow. As outlined above, the duration of the snow cover extent is of major importance. In general, there is a decrease in spring snow cover extent observed (Figure 6), which typically increases the exposure of ground to strong radiation during springtime.

**Observed changes in permafrost**

In general, climate change indications are given by alterations of the permafrost in its spatial extent, thickness, and temperature, as well as by landform changes. However, due to the lack of long-term data series and the large heterogeneity of the different sites (continuous, discontinuous permafrost), the observation and monitoring of the response of permafrost to climate change can be rather complex. Nevertheless, recent warming in the polar and mountainous regions has resulted in documented generally warmer permafrost, deeper active layers, and the development of taliks. While warming of permafrost is distinct in northern Europe and North America, trends are less clear for mountain regions, for instance for the European Alps, because of large variations in snow conditions and complex

---

**Figure 5** Maps of observed precipitation change from 1901 to 2010 and from 1951 to 2010. Reproduced from IPCC (2013).

**Figure 6** Extent of Northern Hemisphere March–April (spring) average snow cover. Reproduced from IPCC (2013).
heat exchange mechanisms in warm permafrost close to 0°C. For most other mountain ranges worldwide, including large mountain ranges such as the Himalayas, systematic monitoring and observation of permafrost conditions do not exist.

Changing active layer thicknesses represent the direct response to annual climate conditions and show interannual variations between +20% in Svalbard (Janssonhaugen) and +100% in the Alps (Schilthorn), depending on site-specific conditions (Voigt et al. 2010). On Svalbard, significant near-surface warming was reported as the result of a strong temperature anomaly during winter and spring 2005/2006, resulting in a 1.8°C higher mean ground temperature at the permafrost table as compared to the mean of the six previous years. A similar anomaly occurred in summer 2003 in the Alps; a relatively direct response was recognized in the acceleration of rock glaciers (Roer, Kääb, and Dikau 2005) and destabilization of rock walls (Gruber, Hoelzle, and Haeberli 2004).

Permafrost temperatures have risen over the past several decades in Alaska; coastal sites show continuous warming since the 1980s and this warming trend has propagated south toward the Brooks Range, with noticeable warming in the upper 20 m of permafrost since 2008 (see Figure 7; Romanovsky et al. 2011). Permafrost in the Alaskan interior warmed in the 1980s and 1990s, but has generally stabilized during the past 10 years. Northern Russia and northwest Canada show increases in permafrost temperature similar in magnitude to those in Alaska during the past 30–35 years. Recent permafrost temperature records from three boreholes in Sweden, Norway, and Svalbard suggest rapid warming, with rates of 0.04–0.07°C year⁻¹ at the permafrost table. In the Swiss Alps, there is evidence for warming at rates of about 0.01°C year⁻¹ at a depth of 48 m in bedrock (Harris et al. 2009).

**Landform changes**

Changes in permafrost extent, thickness, and temperature are used as first diagnostics of climate change. However, since these changes are not visible and not easily detectable, topographic or landform transformations are often the first indicators for ongoing changes. The increase in temperatures leads to warming and thawing permafrost, thus initiating the degradation of single landforms or entire landscapes, visible for example in the expansion of thermokarst lakes in the high latitudes of Russia. The occurrence of thermokarst mirrors the decay of permafrost in soils and sediments and features sink holes, slumps, and lakes caused by the disturbance of the thermal equilibrium brought about by
CLIMATE CHANGE AND PERMAFROST

geomorphic, vegetation, or climatic changes, or a combination thereof (Swift et al. 2014). Another geomorphological indicator of climate change in permafrost regions is the increased erosion of arctic coastlines, as a complex of the ocean–land surface system. The smaller, shorter, and thinner distribution of sea ice allows for the longer activity of waves at the shore. Also, in high mountain systems, the warming ice, as well as the occurrence of unfrozen water, leads to changes in the thermo-hydrological system and related processes. This is reflected by the increasing number of accelerated and destabilized rock glaciers (Figure 8), active layer detachment slides, and rockfalls. During recent decades it has been observed that several landforms are moving at increased rates, probably related to warming permafrost temperatures. The monitoring of seasonal and interannual changes in rock glacier

Figure 8  Collapsing tongue and development of deep cracks of rock glacier Grueo1 (Valais, Switzerland) between 1975 and 2001. The cracks started to develop on the orographic right side, while later (between 1993 and 2001) the landslide-like failure extended over the entire tongue. Between 1975 and 2001 the rock glacier advanced about 60 m ($\approx$2.3 m year$^{-1}$). (See also Roer 2007, Kääb et al. 2007.) Orthoimages of 1975, 1987, and 1993 © Swiss Federal Office of Topography (Swisstopo). Orthoimage of 2001 © RTG 437, Department of Geography, University of Bonn.
CLIMATE CHANGE AND PERMAFROST

dynamics indicates regionally synchronous patterns, reflecting response to varying ground surface temperatures. In addition, destabilized rock glaciers are also described; they show very high deformation rates and often display crevasse-like features at the surface indicative of shear mechanisms similar to those known for landslides.

Monitoring of permafrost

International permafrost data such as borehole temperatures and active layer thicknesses are made available in the Global Terrestrial Network of Permafrost (GTN-P), which aims to manage a global network of permafrost observatories for detecting, monitoring, and predicting climate change. The basis is a database with open access permafrost data from all over the world. Beside this superordinate network, some thematic as well as national networks exist (see Figure 9). A well-established national network is PERMOS (Permafrost Monitoring Switzerland; www.permos.ch), which was formally implemented in 2007, after a 6-year pilot phase. The aim is the systematic, long-term documentation and investigation of permafrost in the Swiss Alps, as compiled at different sites with varying topographic settings (mountaintops, crests, talus slopes, rock glaciers) based on three different types of observations: (i) ground temperatures measured in boreholes as well as near the ground surface, (ii) changes in subsurface ice and water contents, and (iii) velocities of creeping permafrost landforms. A good example of a more thematic network on permafrost is the

Figure 9 Permafrost borehole temperatures at depths of around 10 m, from the Alps (Switzerland), Norway, and Svalbard (Janssonhaugen). Voigt et al. 2010. Reproduced by permission of the authors.
Circumpolar Active Layer Monitoring (CALM) program, which coordinates measurements of active layer thickness at 260 sites, measurement of 168 of which has been ongoing since the 1990s.

Impacts of permafrost thawing for society

Thawing of permafrost can lead to increasing active layer depths, the development of taliks, and generally a decrease in the spatial extent of permafrost areas. There are several second-order impacts associated with permafrost thawing, some with adverse impacts for human societies and/or positive feedback mechanisms.

In the following some selected examples of impacts related to permafrost thawing are outlined.

Instabilities related to permafrost thawing

Loose ground, debris slopes, or steep rock walls are stable when pores and cracks are filled with permanent ice (permafrost conditions), and infrastructure constructions are stable when anchored in permafrost areas. When permafrost thaws, steep slopes can destabilize and infrastructure such as roads, railways, and houses in arctic or high-altitude areas can lose their stability.

In relatively warm permafrost, air temperature anomalies or heavy precipitation events can provoke rockfalls and landslides. Among the most prominent anomalies was the extraordinarily hot summer of 2003 in central Europe, which caused an enlarged active layer thickness or permafrost thaw in the Alps, contributing to increased rockfall activities (Gruber, Hoelzle, and Haeberli 2004). Also, short high-temperature anomalies or precipitation events, it is suggested, are able to trigger large slope failures, as observations from different mountain regions of the world have shown. Increasing permafrost temperatures also pose engineering challenges of infrastructure stability in arctic regions or, for instance, for the railway on the Tibet Plateau.

Changing landscape features

Degradation of permafrost changes geomorphologic features and can lead to the formation of thermokarst terrain, expansion of thaw lakes, or erosion of arctic coastlines. These impacts that mainly occur in arctic regions can affect construction work and infrastructure stability.

Feedbacks (carbon cycle)

The two examples above describe impacts at the local or regional scale, but permafrost thaw in the arctic regions can have a significant global impact, and is among the greatest fears regarding global climate change. Frozen organic soils, including the subsea areas on the shallow shelves of the Arctic Ocean, some relics from the last glaciation, contain considerable quantities of carbon (Figure 10). In fact, these areas hold more than twice the amount of carbon dioxide (CO$_2$) currently present in the atmosphere. Permafrost thawing exposes previously frozen carbon to microbial degradation and releases radiatively active gases, such as carbon dioxide (CO$_2$) and methane (CH$_4$) or nitrous oxide (N$_2$O). These gases are very strong radioactive forcing gases and create positive feedbacks, which amplify significantly global warming.

Modeling (future) permafrost conditions

There is a range of models (from simple to complex) available to simulate permafrost. Each model approach has its own advantages and
disadvantages, and the objective of a study determines which modeling approach is most appropriate.

Assessments for large areas are typically modeled by using **empirical-statistical models**. These models provide probabilities for permafrost occurrence as a function of topo-climatic parameters and statistical relation, for example between potential direct solar radiation and MAAT.

**Process-based models** focus on a detailed understanding of the energy fluxes between the atmosphere and the permafrost ground. They explicitly parameterize the energy balance at the surface and require a correspondingly large amount of measured or computed data. These models also consider subsurface characteristics and surface layers such as seasonal snow. Due to their large computational demand and the extensive input of observational data, these models are restricted to modeling only relatively small regions or specific sites. However, only with process-based models can ongoing processes related to climatic changes be studied.

Several thermal modeling approaches and studies have been conducted and developed in Norway for regions with gentle topography (Westermann et al. 2013), and the use of climate model output for permafrost modeling in alpine environments is among the most promising techniques for scenario generation (Salzmann et al. 2007). For the Swiss Alps, a recently finished large project (funded by the Swiss National...
Science Foundation) on The Evolution of Mountain Permafrost in Switzerland (TEMPS) made a particular effort in assessing future permafrost evolution. By combining various methods, including scenarios from regional climate models and a one-dimensional coupled heat and mass transfer model (COUP), it was shown that air temperature change has the strongest impact on permafrost in the Alps. Also, there is an overall degradation of permafrost observed, which will very likely continue. Furthermore, the changing snow cover regime could be associated with the high variance found in permafrost time series, and ice content is responsible for the rate of permafrost thawing.

SEE ALSO: Climate change, concept of; Global environmental change: human dimensions; Patterned ground; Periglacial processes and landforms; Snow cover changes

References


Climate change is widely recognized to be one of the major challenges facing humanity in the twenty-first century. The Intergovernmental Panel on Climate Change (IPCC) has provided increasingly clear signals that human actions are contributing to growing concentrations of greenhouse gases in the atmosphere which, if unchecked, may lead to an increase in global average temperatures of more than 2°C above preindustrial levels by 2030 and a 5–6°C rise in the longer term (IPCC 2013). The IPCC describes the consequences of even the lower increase as “dangerous” and warns of rapidly rising sea levels, greater climatic variability (e.g., increased drought occurrences and severity), more frequent extreme weather events, and acidification of the world’s oceans caused by the accelerated uptake of atmospheric carbon dioxide.

Such cautions have transformed climate change from a relatively narrow scientific concern into a high-profile political issue. However, governing climate change is intensely challenging for a number of reasons. First, limitations in current understandings of natural and human-induced effects on the Earth’s climate system create significant problems for assessing the costs and benefits of different courses of action. Second, greenhouse gas emissions are integral to virtually all human activities, including energy production; transport and mobility; land use; planning; building design; waste and, ultimately, individual behavior. As such, climate change is a classically “wicked” environmental problem, the management of which requires high levels of scientific knowledge, major economic restructuring, and strong coordination across tiers of governance, industry, finance, and civil society.

Such complexities and uncertainties have prompted serious disagreements over methods for addressing climate change and how to apportion responsibilities for reducing greenhouse gas emissions. Within United Nations discussions, many countries from the Global South that have contributed relatively little to current atmospheric greenhouse gas concentrations argue that developed countries in North America and Europe should reduce their emissions first and that their development should not be impeded by efforts to decarbonize the world economy (Falkner, Stephan, and Vogler 2010). However, many developed countries have been unwilling or unable to show strong climate leadership. Inequality in the distribution of the effects of climate change has been another source of dispute. Although climate change is projected to produce global impacts, some regions stand to be more severely affected than others. Examples of disparities include: the magnified dangers of sea-level rise for low-lying territories in Bangladesh and major coastal cities; faster-than-average warming in polar regions caused by losses in albedo (created by snow and ice melt, reduced reflection of sunlight, and increased heat absorption by the ground); and desertification in semiarid areas in sub-Saharan Africa caused by altered
CLIMATE CHANGE POLICY

weather patterns. Such physical differences are compounded by differences in vulnerability to climate stresses. Low-income areas in the Global South with high reliance on agriculture, seasonal water availability, or fisheries, for example, face particular difficulties and even the prospect of climate-related conflicts and climate-induced migration.

Climate change thus raises serious issues about fairness and responsibility between nations, communities, and generations that are further aggravated by ongoing rapid industrialization and spiralling emissions in countries like Brazil, India, and China. Geographical perspectives are therefore crucial both to understanding climate change itself and to developing strategies to combat and adapt to climate change. Climate policy must contend not only with the grand project of decarbonizing the global economy in aggregate but also with how to do this while addressing rather than reinforcing inequalities and enabling rather than impeding economic development. It must also grapple with how to help societies and communities adapt to a more unstable climate.

The entry begins its review of these debates by examining the main types of climate policy and general problems within the idea of governing climate change. It then discusses climate policy developments at the global and national levels, and the emergence of transnational climate-governance networks comprised of public and private sector actors as a major (and, some argue, leading) force in climate governance and critical scholarship on carbon markets as a means of reducing emissions. The entry concludes by examining emerging debates on adaptation and loss and damages policy.

Defining climate policy

Climate policy traditionally divides into adaptation and mitigation measures, where adaptation encompasses initiatives to help societies adjust to the effects of climate change and mitigation comprises policies to counteract the causes of climate change by reducing emissions. Adaptation policy can be further divided into autonomous initiatives (e.g., households protecting their dwellings from local flooding) or government initiatives (e.g., funding and constructing flood defenses for areas threatened by sea-level rise), while mitigation policy can be either direct (e.g., measures to reduce the carbon intensity of an industrial process or vehicle fuels) or indirect (e.g., policies to promote renewable energy that reduce emissions by reducing demand for fossil fuel energy generation).

Adaptation policy is sometimes argued to be geographically distinctive from mitigation policy in the sense that although adaptation can be instigated at any spatial level (e.g., international, national, and regional assistance programs), physical adaptation activities generally take place locally to ameliorate local climate-related stresses. In contrast, mitigation policy has more of a global dimension because greenhouse gases diffuse rapidly over the atmosphere. As such, mitigation actions anywhere in the world theoretically lessen overall global impacts. Conversely, mitigation is also local because emissions sources themselves come from local point sources (e.g., individual power stations, factories, vehicles, and people). The implications of these distinctions are discussed further later in the entry.

A further subcategory of climate policy concerns compensation for loss and damages caused by climate-related impacts. Loss and damages first gained official status within the United Nations Framework Convention on Climate Change following the 16th Conference of Parties in Cancún, Mexico, in 2010 and broadly differentiates between loss and damages on the basis of whether reparation or restoration for negative impacts is impossible (e.g., loss of freshwater
resources) or possible (e.g., damages to buildings from storm events) (Surminski and Eldridge 2013). Although commonly associated with adaptation policy, loss and damages can also be incorporated into mitigation policy, for example, where governments provide transitional assistance to industries affected by a national carbon tax if this creates competitive disadvantages compared with companies operating in countries that have not implemented equivalent carbon pricing measures.

Beyond these basic categories, the character and boundaries of climate policy become hazier. Although the physical processes associated with climate change are essentially environmental, tackling the causes of increased atmospheric greenhouse gas concentrations is an economy-wide issue, not just an environmental policy problem. Mitigation and adaptation efforts are thus unlikely to succeed unless climate issues are integrated into all major policy areas and there is clear policy programming to ensure mitigation and adaptation operate cohesively. Viewing climate change solely as an environmental issue also makes it susceptible to being marginalized in government or to attack by groups whose interests are negatively affected by mitigation or adaptation policies (Newell and Paterson 2010). Policy integration also poses major challenges for compartmentalized models of government, where each department has its own priorities against which climate concerns must compete and where departmental boundaries limit information flows and impede coordinated action.

Some authors go further and question the fundamental goals of climate mitigation policy. Hulme (2008) criticizes what he sees as the intellectual hubris behind the idea that humans can or should attempt to stabilize the Earth’s enigmatic and dynamic climate system. Beyond the practical difficulties of defining the idea of “a normal climate” and mechanisms that might achieve climate stability, Hulme suggests that the mindset of climate stabilization raises deeper ethical issues about who has the right to determine desired climate outcomes and how costs and benefits should be distributed, on what authority, and in whose interests they should act. Among the questions he asks is: on what grounds is it defensible to stabilize the climate to lessen economic losses in developed countries if this sacrifices development and poverty reduction in the world’s poorest nations?

Despite the undoubted value of such critiques and the need for greater reflection on the balance between mitigation and adaptation, mitigation has tended to occupy a dominant position in political and social science debates on climate policy. One explanation for this is that it aligns with the general orthodoxy that public policy can essentially solve or manage policy problems within acceptable limits. This outlook has a particular tradition in environmental policy, where considerable successes have been achieved in controlling ozone-depleting substances, sulphur dioxide, and other pollutants, and could also be regarded as more politically acceptable than resigning humanity to potentially devastating changes in the climate system. The growing emphasis being placed on adaptation, loss and damages, and nonstate forms of climate governance nevertheless suggests that climate change is severely testing deeply held beliefs in the problem-solving abilities of nation-states and public policy. The entry now examines some of the main developments in scholarship on international, national, and transnational modes of climate policy and carbon markets.

**International climate policy**

Anthropogenic impacts on the Earth’s climate system are a global problem in three important
respects: (i) greenhouse gas emissions anywhere in the world contribute to accumulating stocks of greenhouse gases in the atmosphere and, by extension, to climate impacts across the world; (ii) all nations stand to be affected by climate change, though in different ways and with differing levels of severity; and (iii) no one country can tackle climate change single-handedly. Many commentators have concluded from this that effective mitigation can only come about through global cooperation and interstate dialogue under the auspices of the United Nations.

Reflecting this viewpoint, the United Nations Framework Convention on Climate Change (UNFCCC) has worked since the 1992 Rio Earth Summit to establish a global climate regime and targets, timetables, and mechanisms for reducing greenhouse gas emissions and enabling adaptation. This global deal approach built upon an established model that enjoyed considerable success with other transboundary environmental problems. The 1985 Vienna Convention and 1979 Geneva Convention have been particularly hailed as creating effective legal frameworks for subsequent protocols limiting or prohibiting long-range transboundary air pollution and ozone-depleting substances (Falkner, Stephan, and Vogler 2010). It also aligned well with a state-centered view of geopolitics that assumes nation-states possess the capacity and willingness to work cooperatively through the United Nations to combat commons threats.

The international climate negotiations have, however, exposed serious shortcomings in state-hierarchical approaches to climate diplomacy. Although the 1992 Framework Convention provided the political and legal foundations for the 1997 Kyoto Protocol, negotiations on a successor treaty were hampered by the refusal of the United States to ratify Kyoto and ongoing disputes over how the burden of reducing emissions should be shared between developed and developing countries, with particular attention focusing on rapidly emerging economies such as China and India. Two main reasons can be proposed for these difficulties: (i) the UN’s inability to force agreement between countries; and (ii) a lack of willingness by governments to compromise national interests (whether related to economic competitiveness and development or to the provision of financial and technological assistance to mitigate and adapt to climate change) (Falkner, Stephan, and Vogler 2010).

Most major developed countries nevertheless agreed legally binding emissions reduction targets in the Kyoto Protocol (the main exceptions being the United States and, until 2007, Australia, though Canada withdrew from its Kyoto commitments in 2011). Additionally, the Kyoto agreement incorporated the Clean Development Mechanism and Joint Implementation, two instruments designed to leverage investment in overseas emissions-reduction projects to ease the costs of meeting emissions targets whilst promoting sustainable development in host countries. However, the 2009 Copenhagen Accord only managed to secure loose commitments on targets and funding and fueled fears that the UN process could collapse as a result of disagreements on how to apportion responsibilities for reducing emissions and the provision of funding.

Work by subsequent Conferences of the Parties to retrieve this situation led in December 2015 to the adoption of the long-awaited and hard-won successor to the Kyoto Protocol, the Paris Agreement. Among the key undertakings made in the agreement are: (i) to hold the increase in global average temperatures to well below 2°C above preindustrial levels and to pursue efforts to limit this to 1.5°C; (ii) to achieve a net balance between anthropogenic greenhouse gas emissions and removals between 2050 and 2100; (iii) pledges by over 180 countries to cut or curb their emissions, although these take the form of Intended
Nationally Defined Contributions (INDCs) and are not legally binding; (iv) the creation of a five-yearly review mechanism to enable countries to increase their INDCs to close the gap between current commitments and the emissions reductions needed to remain below the 2°C target; and (v) to accelerate finance flows alongside a mechanism for addressing financial losses faced by vulnerable countries from climate impacts.

Although the Paris Agreement appears to represent a significant step forward in the intensity and timescale of action being taken to avoid dangerous climate change, and in the levels of cooperation between countries in the Global North and Global South, its real value will only be known as governments develop hard-and-fast policies to reduce emissions, support adaptation, and provide for loss and damages.

**National and transnational climate policy**

National climate politics has been a somewhat underexplored topic within geographical research on climate policy, particularly mitigation policy. Early research on this topic was instead dominated by environmental economists analyzing the use of carbon trading and carbon taxes to reduce emissions (Helm 2005). One chief attraction of this research was that it directly addressed the question of how to mitigate climate change without harming prosperity or competitiveness, and so aligned strongly with traditional state economic interests. However, the rather normative approach of environmental economics has been criticized for treating greenhouse gas emissions as a purely technical question of correcting market distortions and for ignoring broader social and political debates on how societies should both mitigate and adapt to climate change (see discussion by Hulme (2008) above).

The political dimensions of national climate policy have been addressed more directly by comparative research by political scientists examining how national characteristics such as per capita wealth, education, dependency on fossil fuels, press freedoms, and the strength of domestic environmental groups affect the strength or weakness of national mitigation policies. Harrison and Sundstrom (2007) conclude from studies of the European Union, the United States, Russia, Canada, Australia, and China that compliance costs were a consistent influence, but they also acknowledge that the impact of many factors was mixed in both direction and extent. For instance, public opinion has been important in persuading politicians to introduce mitigation policies in the European Union but has been less influential in the United States, Canada, and Russia, while shifts in public opinion and electoral competition in Australia have both driven and constrained mitigation policy, in turn shaping the emphasis on and nature of adaptation policy.

Structural explanations for why some countries adopt more ambitious mitigation policies nevertheless still fail to capture the multi-issue, multi-actor and partisan nature of political debates on climate policy in many countries. Bailey and Compston (2012) attempt to counteract this by investigating the tactics used by industry groups, electorates, the media, and national legislatures to oppose new climate policies and political strategies that governments can utilize to counteract such opposition. Examples of these strategies include: (i) taking action early during a new government to allow time for political disputes and concerns about new mitigation and adaptation initiatives to fade before the next election; (ii) offering policy concessions to influential groups such as major industries and electorates to dissuade them from opposing climate measures; (iii) using persuasion...
CLIMATE CHANGE POLICY

techniques, such as stressing the moral imperatives for mitigation and adaptation actions and emphasizing climate policy as an opportunity to develop competitive advantages in low-carbon industrial sectors; and (iv) employing manoeuvres to strengthen climate policy relative to more powerful government departments, such as finance and industry. Compston and Bailey, however, stress that the political, economic, and social conditions shaping climate policy in different countries are often highly specific and that such frameworks only provide a figurative toolkit for how governments might deal with nationally distinctive barriers to climate policy. Detailed studies of national mitigation policies and politics nevertheless remain relatively sparse and have emanated mainly from political science rather than geography. A noteworthy example of such research is Carter and Jacobs’s (2014) investigation of factors aiding the creation of the United Kingdom’s Climate Change Act between 2006 and 2010. In particular, they stress the importance of interparty competition for climate leadership and policy entrepreneurship by individual politicians who opt to make issues like climate change part of their “narrative identity” and political legacy.

The question still remains why greater geographical scholarship has not been directed towards analyzing national climate policy, particularly given the difficulties geographical differences present for identifying general factors affecting mitigation and adaptation efforts. A major explanation for this trend is a more general shift in geographical thinking away from conceptualizing places and spatial scales as rigid and independent phenomena (e.g., the international and national “layers”) and towards modes of enquiry that emphasize the fluidity of economic, social, and political interactions being brought about by globalization and economic liberalization. This has led to increased questioning of the idea that nation-states are pre-eminent political actors and growing interest in multi-actor, multilevel, and transnational forms of environmental governance. Geographers have been particularly energetic in exploring how climate leadership is increasingly being exerted by regional and city administrations, multilateral institutions (e.g., the World Bank and World Trade Organization), businesses, nongovernment organizations, and individuals through their everyday conduct and opinions.

This trend is further reflected in the growing diversity of mechanisms through which climate change is being governed. Rather than mitigation policy being dominated by state regulations and financial incentives, research shows that nonstate initiatives such as corporate sustainability reporting (e.g., the Carbon Disclosure Project), industry-led certification schemes (e.g., the Forestry Stewardship Council), and grassroots initiatives are rapidly becoming mainstream forms of climate governance. Geographers have also analyzed the growth of transnational climate governance networks led by non-nation-state actors and involving cooperation between multiple actors from the public and private sectors. One prominent example is the International Council for Local Environmental Initiatives (ICLEI), an association of over 1000 cities and towns in 86 countries working to promote local action for global sustainability (www.iclei.org/). Another is the Transition Network, founded in 2007 to encourage and support communities to develop grassroots initiatives to reduce emissions and build local resilience (www.transitionnetwork.org/). By 2013, over 1000 Transition initiatives had been established in the United Kingdom, mainland Europe, Australasia, and North America.

Geographical thinking about the spatialities of climate governance thus raises important questions about the changing nature of climate
policy. Pivotal among these are the implications of transnational and nonstate forms of climate governance for the role of the state. To what extent are state functions being “bypassed” or are states being left behind as climate-governance actors because of the slow progress made by UNFCCC negotiations and state-led climate policies, or are states simply changing how they govern, with greater focus on creating regulatory frameworks to support and steer more dynamic and less encumbered nonstate and transnational initiatives? Most analysts agree that “governing beyond the state [does not mean] governing without the state” and that many – though by no means all – nonstate governance arrangements (particularly within the adaptation sphere) involve state intervention in establishing ground rules and enabling and monitoring mechanisms for nonstate-led initiatives (Bulkeley and Jordan 2012, 564). At the same time, geographical analysis has demonstrated that state authority over the governance of socioenvironmental problems is being replaced by more polycentric and autonomous modes of governance.

Further questions concern the effectiveness and legitimacy of public–private climate networks. Although most commentators agree that interstate negotiations and state actions have made limited progress in establishing effective climate governance regimes, will greater nonstate actor leadership yield more effective outcomes, and what mechanisms exist to ensure their accountability to those affected by nonstate initiatives? There are few definitive answers to these questions, not least because of the number and diversity of nonstate initiatives; however, the fragmentation of governing authority within nonstate climate governance networks creates obvious challenges for coordination and coherence. Additionally, transnational climate-governance activities cannot be assumed to place reducing emissions or promoting adaptive capacity above the private interests of the actor groups involved. The nonstate network approach thus relies on climate goals coalescing with other motivations. For example, the World Bank is a major actor in climate governance through its involvement in the Climate Investment Funds and its influence over development policy in many developing countries. Some of its investments may therefore work to reduce emissions or finance adaptation while others, particularly those linked to economic growth, may undermine these goals. Although tensions also exist between state climate and economic strategies, corporate motivations for becoming involved in climate governance depend on them genuinely acknowledging the impact of their activities on the social and physical environment and their willingness and capacity to dedicate resources to promoting collective mitigation and adaptation actions (Newell and Paterson 2010).

**Carbon markets**

Carbon markets have become a core element of both state-led and private responses to climate change in recent years. The essential logic underpinning carbon markets is that markets are the main driving force for innovation and productivity within economies; therefore, if markets can be designed that internalize the social and environmental costs of carbon emissions, in theory they can provide a mechanism for achieving effective and cost-effective reductions in emissions. Three main types of carbon market can be distinguished.

“**Cap and trade**”

“Cap and trade” is where a governing body sets an overall emissions cap, distributes emissions allowances to affected industries (typically
large power generators and energy-intensive manufacturers), and allows companies to trade allowances in order to help achieve emissions targets. The purpose of trading in this instance is to give emitters with lower emissions-abatement costs a financial incentive to sell their surplus allowances to those with higher abatement costs so as to produce net overall welfare gains at reduced cost compared with other policy approaches. The European Union Emissions Trading System (EU ETS) is the world’s largest cap-and-trade scheme. In 2014, the scheme covered all 28 EU member states plus Iceland, Liechtenstein, and Norway, and around 45% of the European Union’s total greenhouse gas emissions. From 2013 onwards, the scheme was expanded to incorporate aviation to and from participating countries.

**Government-led project schemes**

Such schemes involve governments or organizations investing in emissions-reduction projects elsewhere in the world to generate emissions credits for sale on international markets. The most prominent example of this type of scheme is the Clean Development Mechanism (CDM), which allows developed nations that ratified the Kyoto Protocol to buy credits generated from projects in developing countries to meet a proportion of their UNFCCC targets. In addition to enabling countries to source emissions cuts at lower cost compared with decarbonizing their domestic energy or transport systems, the CDM aims to promote investment, technology transfer, and sustainable development in host countries. Typical investments permitted by the CDM include wind and solar energy, methane capture from landfill gas, and the destruction of potent synthetic greenhouse gases.

Most project-based schemes utilize a *baseline-and-credit* methodology, where the number of credits issued represents the difference between measured emissions from activities and estimates of business-as-usual emissions if the project had not been implemented. Because calculations are based around a hypothetical comparison, baseline-and-credit systems need robust oversight by the CDM Executive Board and Designated National Authorities to determine the types of project permitted, approve projects, and verify the number of credits generated.

**Voluntary carbon offsets (VCOs)**

VCOs utilize a methodology similar to that of government-led project schemes but provide mechanisms for private individuals, companies, and governments to purchase carbon credits to offset their greenhouse gas emissions. Credits are again generated from remote abatement projects funded by credit sales, for example, where an individual buys carbon offsets from a tree-planting initiative to compensate for emissions from personal air travel. However, because VCOs are run by private organizations (e.g., Future Forests, ClimateCare, and the CarbonNeutral Company) rather than governing bodies, VCOs are generally subject to less strict controls, meaning that a broader range of projects is utilized and projects are not always screened as rigorously.

The philosophy and practice of carbon markets have provoked intense academic debate. Most environmental economists see them as valuable mitigation mechanisms because of their cost-effectiveness and flexibility in reducing emissions. Geographical research has, however, highlighted two main sets of difficulties: (i) the political and practical problems of translating economic theories on carbon pricing into effective policy; and (ii) a tendency for carbon markets to enable the commodification and expropriation of aspects of nature in ways that sustain uneven
CLIMATE CHANGE POLICY

and crisis-prone development patterns at the expense of justice-based approaches to climate governance. These arguments are now examined in turn.

A crucial challenge for carbon markets is how to translate different greenhouse gases into common units so that they can be traded. Greenhouse gases do not possess the intrinsic value or tangible qualities of normal commodities, so creating markets requires estimates of the carbon savings gained, for example, from replacing wood-burning stoves with gas stoves in India, and methodologies for converting gases with different climate-forcing potentials into carbon dioxide equivalents so as to create a tradable commodity. Such proxy commodification is inherently susceptible to pressures from political and economic actors with vested interests in the profitability of carbon markets.

Other research stresses the political difficulties of governing carbon markets. Most economists agree that cap and trade depends on setting strict emissions caps to create strong incentives for businesses to innovate and reduce emissions. However, the European Union has struggled to agree targets for its emissions trading scheme because some member states have pressed for less challenging targets to protect major industries and shield consumers from higher energy prices (Bailey, Gouldson, and Newell 2011). The CDM has also encountered difficulties in: ensuring rigorous rules for monitoring, reporting, and verifying emissions reductions; resisting pressure to relax rules on whether projects are judged to be additional to business as usual; and structuring carbon markets so that they direct finance and technology to countries according to their development needs. Surveys of the CDM indicate that investment has been instead concentrated towards rapidly industrializing countries, particularly China, where more lucrative projects exist, while Africa has benefited least from CDM finance despite having more pressing development needs (Newell and Paterson 2010).

Problems with the EU ETS and the CDM also underscore the geographical challenges facing carbon markets. In particular, they reveal tensions between a general regulatory logic that sees carbon markets as effective and efficient mechanisms and the territorial logic of governments seeking to safeguard consumers and strategic industries. The governance problems encountered by the CDM, meanwhile, illustrate the difficulties of reconciling market imperatives for profitable investments with promoting sustainable development and a just apportionment of costs and benefits between countries that have industrialized by burning fossil fuels and those in greatest need and most vulnerable to climate change.

Marxist critiques of carbon markets express similar concerns about their potential to reinforce unequal power relations and inequitable development, but draw broader connections between carbon markets and the history of global capitalism. Central to this approach is the idea that capitalism must continually find new modes and sites of wealth accumulation in order to avert periodic crises in an economic system that is structurally organized around perpetual growth as it encounters resource limits. Harvey (2006) argues that capitalist systems feed value production through accumulation by dispossession – the commodification and privatization of new resources at the expense of pre-existing rights irrespective of whether they were previously regarded as common heritage or property. The commodification of greenhouse gas emissions is seen as a key instance of such dispossession wherein carbon markets assign property rights to, and essentially privatize, aspects of the atmosphere and livelihoods to fuel capital speculation. Böhm, Misoczky, and Moog (2012) describe how the use of rice husk in a CDM biomass power generation project in Thailand denied
local peasants a traditional source of fertilizers and forced them to buy chemical fertilizers from transnational corporations. Newell and Paterson (2010) similarly portray carbon markets as effecting a redistribution of wealth and access to resources under conditions of unequal exchange, where markets create and sell cheap credits to the Global North obtained from the Global South. Such projects often have limited regard for delivering benefits to host regions and mainly enable high-emitting countries to avoid forgoing carbon-intensive lifestyles.

Bailey, Gouldson, and Newell (2011), meanwhile, examine connections between carbon markets and ecological modernization, the influential school of thought which suggests that climate change can be addressed through technological and market innovation without sacrificing societies’ ideological commitment to market capitalism and economic growth. The innate compatibility of capitalism and environmentalism proposed by ecological modernization provides one reason for the political acceptance of carbon markets by mainstream political and business actors. However, the concern again is that the preference for carbon markets within such interpretations of ecological modernization effectively locks out alternative – and perhaps more effective and equitable – ways of mitigating climate change. In particular, ecological modernization’s emphasis on efficiency pays limited attention to distributional equity and the possible need to limit demand for energy and emissions-producing activities. This in turn raises questions about whether ecological modernization thinking enables political and business elites to recognize and respond to structural problems within carbon markets. In essence, does ecological modernization involve genuine reflection on how to reduce the environmental impacts of capitalism or does it simply reinforce “business-as-usual” modes of thinking?

Policies on adaptation and loss and damages

So far, the discussion has focused chiefly on mitigation policy or issues common to both mitigation and adaptation. However, geographers have also been heavily involved in studying the ways in which human adjustments to climate risks operate across spatial and societal scales. An obvious foundational question for adaptation research concerns visions of adaptation. Pelling (2011) argues that climate adaptation should not be limited to interpretations of resilience that emphasize preserving pre-existing states and practices (e.g., the use of new seed varieties to sustain existing agricultural practices). Although resilience in its broader interpretation incorporates multiple and dynamic visions of change in response to shocks and stresses, Pelling also emphasizes the role of transitional adaptations that promote incremental changes to prevailing economic, political, cultural, and social relations, over time, to produce more radical change. Transformational adaptation, in turn, stresses the possibility of utilizing climate risks consciously to reform or replace dominant societal objectives and political-cultural regimes.

Recognizing that adaptation policy is equally about responding to and creating new realities, and not just about preserving the present, raises further questions about the processes used to evaluate the effectiveness, efficiency, equity, and legitimacy of adaptation policy. Adger, Arnell, and Tompkins (2005) argue that these are again geographical issues because adaptation operates at different spatial and societal scales and involves a broad landscape of actors, including individuals, corporations, civil society, local, regional, and national governments, and international agencies. Of particular importance is the potential for operational adaptation decisions to be influenced and constrained by higher-level institutions.
involved in developing adaptation frameworks, but at the same time, for local interpretation of higher-level decisions to have a significant bearing on adaptation outcomes. Equally important are the perspectives held by different actors on the goals of adaptation policy (if these are even defined) and, thus, what effective, efficient, equitable, and legitimate adaptation means and entails. Recognizing and responding to such interconnections can be critical in avoiding maladaptation, that is, impacts that adversely affect the vulnerability of other regions, sectors, or social groups (Barnett and O’Neill 2010).

Recognition that mitigation and adaptation efforts cannot avoid all the impacts of climate change has also prompted growing attention to loss and damages as a means of addressing climate impacts. The UNFCCC responded to this emerging agenda in 2010 by initiating a work program to assess current knowledge of climate-related loss and damages, mechanisms for addressing loss and damage, and the UNFCCC’s role in enhancing the implementation of loss and damage policies. Approaches to address loss and damage from extreme events and slow-onset processes can be divided into four categories: risk reduction; risk retention (e.g., social safety nets and contingency funds); risk transfer (e.g., insurance); and approaches for managing loss and damage from slow-onset processes.

While loss and damages approaches seek to promote the principle of distributive justice within adaptation policy by ensuring that regions and communities’ ability to adapt to climate change is not constrained by resource limitations, implementing the concept can be problematic. First, severe difficulties lie in attributing individual weather events to climate change and even more so in pinpointing culprits from whom to seek compensation. Second, the unpredictability and magnitude of certain climate risks mean that insurance contracts and claims for some climate events may become exorbitantly expensive. As such, the willingness of governments and other actors to contribute is likely to depend on the establishment of multilateral and bilateral “no-liability” financing mechanisms, while further challenges lie in ensuring local communities have a genuine say in international, national, and local planning and decision-making processes and in establishing governance arrangements for the identification and management of climate loss and damages.

Conclusion

Climate policy and governance have become increasingly important themes in recent geographical research on human–environment relations. This interest reflects both the growth of climate change as a policy concern and its intersections with core geographical interests in environmental and economic futures, development and equity, and the changing character and spatialities of governance. The debates reviewed reveal a multitude of uncertainties and disputes about human influences on the climate system, the feasibility and ethics of “governing” climate, different strategies for mitigating and adapting to climate change and their coexistence with development and justice priorities, and even definitions of climate policy itself. Geographical perspectives play a valuable role in probing the complexities lurking within the idea of climate policy, while the epochal and uneven effects of human-induced climate change provide a compelling case for continued geographical investigations of the challenges of developing effective and equitable climate policy.

SEE ALSO: Climate adaptation/mitigation; Climate change adaptation and social
transformation; Climate policy; Climate and societal impacts; Commodification of nature; Ecological modernization; Environmental (in)justice; Environmental policy; Environmentality and green governmentality; Neoliberalism and the environment

References


Climate change, concept of

Mike Hulme
King’s College London, UK

In 1966, the World Meteorological Organization (WMO) published a technical report on climatic change in which the statistical properties of different meteorological time-series data were systematically compared. Climatic change was proposed as the term to embrace all forms of climatic inconstancy on timescales longer than 10 years, irrespective of cause. Although climates were understood as changeable on all timescales, most scientific attention was paid to past changes which had occurred on multi-century to multi-millennial timescales. The favored term was “climatic change,” climatic being used as an adjective to describe this particular type of change (as opposed to, say, political or economic change). Thus the first academic journal dedicated solely to the study of climate change was launched in 1977 with the title *Climatic Change*.

From the 1970s onwards, understandings of climatic change began to change. There was a growing appreciation that a wide range of human activities, from energy use to food production, had the potential to alter the physical functioning of an interconnected global system. “Climate change” – a noun – now became “an issue” rather than the technical description of changing weather it had been for the WMO in 1966. It took its place in public life alongside other “issues” like global poverty, human rights, or water pollution. With the establishment of two new international institutions – the Intergovernmental Panel on Climate Change (IPCC) in 1988 and the UN Framework Convention on Climate Change (UNFCCC) in 1992 – climate change supplanted climatic change as the dominant linguistic term. Climate change became a phenomenon caused by complex chains of human actions (and natural processes), but also an agent which could influence far-reaching sets of material and imaginative phenomena. As this “new” idea of climate change traveled around the world, it became inscribed with multiple and complex political and cultural meanings. In this sense, the idea of climate change has never fully escaped older cultural readings of climate and its changes which have been retained in many non-Western cultures.

Historical and cultural perspectives

The idea of climate change – whether of natural or human origin – has a genealogy that can be traced back to Greek civilization. Not only were changes in climate discerned by Greek scholars, but these early observers were also able to trace causation to human actions in the world. For example, in the third century BCE, Aristotle’s student Theophrastus observed and documented local changes in climate induced by human agency: the clearing of forests around Philippi in Greece warmed the climate, while the draining of marshes cooled the climate around Thessaly. Non-Western cultures also had accounts of climate change, often captured in linguistic expressions for climate which combined both descriptive and causative elements of change. Thus the Inuktitut word *sila* or the Marshellese phrase *oktak in mejatoto* are both expressions which capture the enveloping climate of a place.
CLIMATE CHANGE, CONCEPT OF

and a causative account of cosmological stability or change.

Western Enlightenment discourses about climate change from the seventeenth century onwards frequently turned to the effects of deforestation. For example, the historian Edward Gibbon could see the beneficial warming effects of tree-clearing, both in changes of climate through time and in differences in climate caused by geography. In the late eighteenth century, Gibbon claimed that the “improvement” in European climate since classical times was due to the clearing of “immense woods” and he believed that contemporary forests in Canada subjected that land to a climate as fierce as that of ancient Germany. Empirical support for such understanding seemed to emerge from the land practices of colonizing Europeans in America. Within half a century of large swathes of forest being cleared along the eastern seaboard, observers were able to remark that winters had become less harsh and summers cooler. Climate had not just changed, but had been changed.

The idea of climate change also occupied the imagination of European colonists in the tropics and subtropics. Here, the destruction of forests was believed to exacerbate the droughts that many settlers in the eighteenth century found endemic in subtropical regimes. Climate change was thus not only a phenomenon that humans induced in the physical world; the resulting changes in climate could also challenge the economic wellbeing of a colony and the health of its exogenous inhabitants. Climate change functioned both as an index of change, but also a putative cause of wider environmental and social change.

In western cultures, the temporal horizons over which climates were believed to change were massively extended in the nineteenth century through the combined work of geologists and physicists. Huge swings in global climate were implied by massive and ancient glaciations, the traces of which were newly diagnosed in the landscape by scientists such as Louis Agassiz in the 1830s. But alongside these emerging radical ideas of global climatic instability occurring over previously unimagined timescales, some observers were still grappling with the extent to which human activities could alter contemporary regional climates. The prominent Austrian geographer Eduard Brückner maintained that statistical evidence could be found for contemporary changes in regional climates. In the 1880s he demonstrated that average temperature and precipitation for areas of central Europe and Russia when measured over successive 35-year periods differed substantially, claiming that such changes in climate would have implications for rivers, lakes, agriculture, and human migration.

Despite the work of Brückner and a few others at the time, the dominant Western view during the first decades of the twentieth century was that climates were basically constant on timescales that mattered to human planning and action. The British climate historian Hubert Lamb was thus able to remark in 1959, “not so very long ago … climate was widely considered as something static, except on geological time scale[s], and authoritative works on the climates of various regions were written without allusion to the possibility of change” (Lamb 1959, 299).

The globalization of climate change

Into the 1960s, the dominant approach to understanding climate – and hence climate change – remained through the comparative analysis of meteorological statistics. The WMO’s technical report on climatic change published in 1966 (WMO 1966), systematically compared the statistical properties of different meteorological time series data. Climatic change was defined
as “all forms of climatic inconstancy, regardless of their statistical nature (or physical causes),” although inconstancies over less than a decade in length were to be regarded as climatic variations. Thus climatic change encompassed climate periodicities (regular and irregular), fluctuations, oscillations, vacillations, discontinuities, and trends. With few exceptions, most scientific attention and public interest at the time was paid to (natural) worldwide changes which had occurred in the distant past on multi-century to multi-millennial timescales. Climatic change that originated through human activities and which occurred on the timescales of human generations was confined to local and regional scales.

The preconditions for a new understanding of global climate change began to emerge in the 1960s, as evidenced in a joint UNESCO/WMO symposium on changes in climate held in Paris in 1963. First, a few scholars such as Hubert Lamb in the United Kingdom and Emmanuel Le Roy Ladurie in France began publishing accounts of historical climate change based on long climatic reconstructions derived from documentary and environmental evidence. Lamb in particular helped to popularize the notion of a Medieval Warm Epoch (in the late Middle Ages) and a Little Ice Age (in the early Modern period) – both of which were imagined to be hemispheric, if not global, in reach (Lamb 1966).

Second, new developments in scientific monitoring of the planet, boosted by the International Geophysical Year in 1957/1958 and by new satellite technologies, prompted a new conception of climate as an interconnected system of atmosphere, ocean, biosphere, and cryosphere. Local and regional understandings of climate largely gave way to global understandings, facilitated by early computer simulation models of the newly imagined “Earth system.” This opened the way for numerical experimentation using these models through which the global climatic effects of volcanoes, sunspots, carbon dioxide and, later, nuclear explosions could be safely simulated.

Third, this idea of a human-induced change in global climate found sympathy in the broader currents of intellectual thought of the 1960s and 1970s. The emergence of a new environmentalism focused attention on the planetary scale effects of human activities on the physical world. One of the first associations of anthropogenic climate change with notions of danger was in a 1963 conference of scientists convened by the Conservation Foundation of New York which warned of a “potentially dangerous atmospheric increase of carbon dioxide.” And the first governmental and international assessments of the prospects of climate change were conducted during this period. In the United States for example, the President’s Scientific Advisory Committee in 1965 published a report on “Restoring the Quality of our Environment,” which included a specific section on “the climatic effects of pollution.”

The possibilities of changes to the “global climate system” – driven by human activities – therefore began to be articulated. The first use in a scientific journal of the term “global warming” occurred in June 1971 (Russell and Landsberg 1971) and a major focus of the WMO’s First World Climate Conference in 1979 was on climate change. During the “greenhouse summer of 1988” the idea of anthropogenic climate change – commonly depicted as the enhanced greenhouse effect – penetrated deeply into popular culture in the West, although more superficially or not at all in other parts of the world. The growing political resonance of climate change was partly explained by the dissolution of the Soviet Union between 1989 and 1991. Fears of Cold War destruction were displaced by those associated with climate change, prompting the observation at the time from cultural theorist Andrew Ross that, “apocalyptic
fears about widespread droughts and melting ice caps have displaced the nuclear threat as the dominant feared meteorological disaster” (Ross 1991, 8).

The evolving scientific definition of climate change can be discerned through successive assessment reports of the IPCC. In its First Assessment Report in 1990, no very precise definition of climate change was offered, although it was stated that the “climate change we are addressing in this report is that which may occur over the next century as a result of human activities” (IPCC 1990, xxxvi). In fact the report discussed both natural and human changes in climate. A few years later, the Second Assessment Report was more explicit. Climate change meant “climate fluctuations of a global nature … and which includes the effects due to human actions … and those due to natural causes” (IPCC 1996, 56). By 2001 the IPCC had settled on a definition which also prevailed for its Fourth and Fifth Assessment Reports: “Climate change refers to statistically significant variation in either the mean state of climate or its variability, persisting for an extended period (typically decades or longer). Climate change may be due to natural internal processes or external forcings, or to persistent anthropogenic changes in the composition of the atmosphere or in land use” (IPCC 2001, 788).

While the IPCC understood climate change to embrace both natural and human causes, for the UNFCCC – signed in 1992 – climate change was to mean “a change of climate which is attributed directly or indirectly to human activity that alters the composition of the atmosphere and which is in addition to natural climate variability observed over comparable time periods.” The contrasting definitions between these scientific and political usages have resulted in considerable strategic ambiguity in the public meaning of climate change.

Meanings, imperatives, and language

Since the 1960s, climate change has moved from being a technical description of a physical phenomenon (climatic change) to becoming a political (i.e., a contested) issue. Thus, each December since 1995 has witnessed a two week meeting of the Conference of the Parties (COPs) to the UNFCCC, where politicians, diplomats, business leaders, and civil society organizations have gathered to negotiate political responses to climate change. Further evidence of this shift includes how some government ministries have been reconstituted around climate change. For example, since 2008 the United Kingdom has had a Department of Energy and Climate Change, while in 2010 the European Union appointed its first Commissioner for Climate Change Action.

At the most generic level climate change now acts as a synecdoche for the collective impacts of humanity on the physical process and sustainability of the Earth. This reading of climate change has helped give rise to the idea of the Anthropocene, a new geological era which marks the global impact of human activities on the Earth’s ecosystems. But as the idea of climate change has traveled around the world through scientific, political, and civic networks it has become inscribed with complex meanings each carrying different local resonances. It is not possible to view climate change “from nowhere.”

Contrasting accounts of climate change offer different interpretations of cause, significance, and responsibility, even while anchoring the idea of climate change in the changing weather attributes of the climate system. Thus Ban Ki-Moon, the UN Secretary-General describes climate change as “the defining challenge of our age … the time to act is now,” while the economist Lord Nicholas Stern describes it as “the greatest example of market failure the world
has seen.” Contrasting meanings can be found in the work of ethicists and anthropologists. So theologian Michael Northcott sees climate change as “the Earth’s judgement on the global market Empire and the heedless consumption that it fosters” and cultural anthropologist Jerry Jacka, writing of the Porgeran tribe in Papua New Guinea, explains climate change as “due to societal breakdown between [the Porgerans] and the rituals oriented toward powerful spirits that control the cosmos.”

The proliferation of linguistic terms used to express the idea of climate change lends further variety to these multiple meanings. With climate change (noun) having replaced climatic (adjective) change it became necessary to describe what sort of climate change the world was facing. Climate change is therefore described as natural, anthropogenic, and/or human-induced. It is portrayed as gradual or rapid, smooth, or abrupt (see Table 1). For example, the term “climate catastrophe” first appeared in the context of anthropogenic climate change in the German language in the cultural magazine Der Spiegel in April 1986. The phrase “catastrophic climate change” (Klimakatastrophe) continues to be used to deliberate effect in certain discourses. And other linguistic entrepreneurs have sought to find more persuasive formulations than the dominant expression “climate change.” So “global warming,” “weather weirding,” “the climate crisis,” “global heating,” and “climate disruption” have all been used as alternative, more vivid, descriptions of the idea of anthropogenic climate change.

**The future of climate change**

There are few issues in the world today that can match the salience and cultural reach of climate change. Books dealing with climate change now appear at the rate of more than one a day in the English language alone, compared to less than one a week a generation ago. They cover topics such as climate change and architecture, diet, football fans, gender, theology, visual art, time, forests, justice, democracy, disasters, trust, law, aviation, migration, and capitalism. Not only is the reality of climate change contested, but so too are its causes, consequences, and meanings. Climate change today therefore needs to be understood as an emergent phenomenon that is simultaneously a physical transformation of the climate system and an evolving cultural symbol. The insights of anthropologists, psychologists, sociologists, literary critics, historians, theologians, and philosophers are needed to do justice to the idea of climate change, as much as those coming from natural and physical scientists.

All of human life is now lived out not just in the presence of a physically changing climate, but...
CLIMATE CHANGE, CONCEPT OF

but in the new discursive and cultural spaces that have been created by the idea of climate change. Most human practices and disputes can now be expressed through the language and symbolism of climate change. Thus photography, music, cartoons, literature, theater, poetry, dance, religious practice, architecture, educational curricula, personal identity, politics, and so on, use climate change as a medium of expression. And disputes about flood management, landscape aesthetics, child procreation and child rearing, trade tariffs, development aid, industrial patents, social justice, taxation, even democracy itself, are formulated in the language and argumentative spaces of climate change. As suggested by the eco-critic Greg Garrard, we “feel there might not be any narrative whose meaning we cannot re-evaluate in relation to climate change” (Garrard 2013, 183). Climate change has become a new condition through which human life now takes shape.

SEE ALSO: Climate policy; Global climate change; Global environmental change: human dimensions; Globalization; Intergovernmental Panel on Climate Change (IPCC)

References


Further reading


Climate change communication

Susanne C. Moser  
*Susanne Moser Research & Consulting, USA  
Stanford University, USA*

Communicating climate change – as a special case of science and risk communication, and as an opportunity for civic engagement with global environmental change – is among the greatest communication challenges in contemporary society. The social-scientific interest in climate change communication has risen over the past decade in response to the growing understanding in the problem and the lagging societal response (Moser 2010, 2016). Over this period, understanding of climate change communication has grown significantly, although it remains difficult because of the nature, scope, scale, and speed of the problem, the enormous political and financial stakes involved, and the changing societal context that shapes people’s receptivity of the issue (Whitmarsh, O’Neill, and Lorenzoni 2011; Moser and Dilling 2007).

For some, communicating climate change is a matter of science education and as such a desirable goal in itself; for others, it is a means to increase scientific literacy so as to increase people’s capacity to participate in an informed manner in the decisions of a democratic society. For yet others, communicating climate change serves advocacy, that is, informing and mobilizing society for or against certain courses of climate action. The resulting politicization of climate change science has also created considerable interest in how to effectively counter disinformation campaigns and overcome societal and political polarization on the issue (McCright and Dunlap 2011).

Geographers’ interest in climate change communication

Given both physical and human geographers’ long-standing involvement in the study of global (climate) change, communication is a central concern for them. As humans modify the climate, environment and life support systems of the planet, and thus cause a fundamental shift in human–environment interactions, including in the availability of natural resources and the occurrence of climatic hazards, climate change communication becomes a foundational skill and tool for geographers. Geographers are interested in conveying and explaining observed and projected changes in the environment – a task made more difficult by the fact that the causes and early impacts of climate change are difficult to see, slow to emerge, and challenging for lay audiences to connect systemically. This makes climate change communicators part of the “early warning” system of a society and thus part of the human response to climate change.

In addition, for many geographers, effectively communicating the causes, risks, and response options to climate change is a matter of practical necessity and ethical obligation. Not only as educators and researchers in multidisciplinary teams, but as expert informants of policy debates or management responses, as consultants, and as individuals participating in civil society, geographers are grappling with how best to communicate the issue to different audiences.
CLIMATE CHANGE COMMUNICATION

With the expertise inherent in the field of geography, in earth science and human-environment interactions, its unique interest in linking phenomena across space and scale, and its ability to understand change in the richness of place-based and regional contexts, geographers are well positioned to help society understand climate change, facilitate meaningful interpretation of the science, and foster an informed debate about possible responses.

Focal areas for climate change communication

The practice and study of climate change communication have centered on different aspects of climate change. Early on, interest was primarily in articulating the science of climate change, that is, establishing its existence, providing multiple lines of evidence for its reality, explaining its causes, particularly the growing understanding of its human causation and how scientists have come to know and become confident in this conclusion, as well as illuminating the complex challenge of global climate modeling and future projections of climate change. Much of the communication challenge with regard to climate science has been in simplifying complex science for audiences with varying levels of science literacy. This involves identifying useful mental models that help individuals understand the fundamental aspects of the problem, and to help audiences sort out (scientifically unsubstantiated, but nevertheless frequently repeated) counterclaims asserting that climate change is not real, not human caused, and not significant or certain enough to warrant action.

A second major focus of climate change communication is the emerging and future impacts of climate change, and the ecological and societal vulnerabilities in different regions and sectors to these consequences. For communicators, there have been both opportunities and challenges in that focus. Climate change impacts as the concrete manifestations that “bring climate change home” (Slocum 2004) offer communicators an opportunity to make the abstract topic of concern visible, tangible and meaningful to audiences (Moser 2014). Years ago, communicating impacts typically involved pointing to the future, to places far from where the majority of humanity lived, and to other species – such as polar bears on melting ice floes or low-lying islands in the Pacific ocean being inundated by the rising sea. These science- and necessity-driven ways of “distancing” climate change in space and time matched the human propensity for psychological distancing (Spence, Poortinga, and Pidgeon 2012), that is, the common desire to keep threats at arm’s length as a way to manage the psychological impact of taking the magnitude and profundity of global climate change seriously. More recently, the early consequences of climate change have become evident in every region of the world and thus offer opportunities to link the observed with the scientifically understood and with future projections. In turn, this challenges communicators not only to find ways to communicate what is (and is not yet) understood in scientifically credible ways, but also to help people process their affective responses (e.g., worries, fears, guilt, denial, hopelessness) to these emerging and projected impacts (Swim et al. 2011).

The third major focus in climate change communication (as in the climate change science field more generally) is on response options. With regard to communicating climate change mitigation, that is, efforts to reduce the causes of anthropogenic climate change, one major challenge has been to persuade society of the human causation of contemporary climate change, and thus of humanity’s responsibility for minimizing future climate change. Concurrent
is the challenge of conveying just how substantial the reductions in emissions from energy consumption, transportation, land use, and industrial activities need to be if society wishes to avoid significant disruption of life support systems, economies and livelihoods, human safety and wellbeing, and place identity. A third set of challenges in communicating mitigation lies in translating complex policy mechanisms and associated responsibilities and roles for different actors at different scales of governance into understandable approaches and clear actions.

Communicating the complementary response of adaptation, that is, the diverse set of strategies and options for preparing for and minimizing the risks of negative impacts from climate change and for taking advantage of possible positive consequences, is a more recent challenge. Many view this as a growing opportunity to “make climate change real,” to take actions that produce immediate and tangible results (as opposed to the delayed climate-related benefits of emission reductions), and to address local challenges that diverse interests may be able to rally around. However, the concept of adaptation is still unfamiliar to people, and to some it signals passivity, defeat, and competition to mitigation. Research on whether or not these assumptions and perceptions are borne out in reality and on how best to communicate adaptation is only beginning to emerge (Moser 2014).

A range of related but more specific issues have emerged, such as the deliberate modification of the climate system through various geo-engineering approaches, or carbon capture and storage, that communication researchers are also beginning to examine. In all of these instances, common communication concerns revolve around public perceptions and understanding, linguistic preferences, mental models, beliefs and attitudes toward responses, the willingness to fiscally or politically support a particular course of action, and what roles different actors may (or should) play in any of them.

**Messaging, language, and imagery**

While the categorical areas of climate change communication directly mirror the concerns of climate change researchers more generally, it is far from obvious exactly what, how much, and in what words and imagery the substance of climate change should be communicated. Information needs, levels of understanding, cognitive processes, cultural values, and the context in which information is being communicated are among the key influences on what information different audiences take up, accept, understand, and perceive as persuasive, how they interpret and retain it, and whether or not and how it influences their behavior (Swim et al. 2011; Hulme 2009; Moser and Dilling 2007).

Particular attention in this context has been paid to mental models, that is, the conceptual, symbolic, often intuitive ways in which we represent real-world processes in our minds. Mental models are simplified explanations, constructed and reinforced by language, images, symbols, sounds, colors, and tone of voice. They predispose us to think about problems and associated solutions in certain ways, for example, the “greenhouse effect” offers a mental model of global warming that likens the effect of heat-trapping gases in the atmosphere to the panes of a greenhouse, which allow sunlight to come through, but heat to be retained inside the greenhouse (Figure 1).

Considerable applied research thus has also gone into research on messaging, that is, what and how best to communicate climate change, including:

- how to avoid the conceptual confusion of weather and climate;
The greenhouse effect

Some solar radiation is reflected by the Earth and the atmosphere.

Some of the infrared radiation passes through the atmosphere. Some is absorbed and re-emitted in all directions by greenhouse gas molecules. The effect of this is to warm the Earth’s surface and the lower atmosphere.

Most radiation is absorbed by the Earth’s surface and warms it.

Infrared radiation is emitted by the Earth’s surface.

Figure 1 The greenhouse effect has long served as a mental model to explain how certain natural and human-generated gases trap heat closer to the Earth’s surface and warm up the planet. Reproduced from US Environmental Protection Agency, 2012. © US Govt/EPA.

- word associations and preferences such as climate change vs. global warming, or adaptation vs. preparedness;
- the benefits and drawbacks of different mental models such as carbon pollution vs. greenhouse effect; and
- resonance of different frames, such as economic vs. national security vs. moral reasons for acting on climate change (Whitmarsh, O’Neill, and Lorenzoni 2011; Moser and Dilling 2007).

More recently, communication researchers have also become interested in the use of imagery, especially iconic imagery and graphics, in conveying climate change messages (O’Neill and Smith 2014) (Figure 2).

Communicators and trusted messengers

Historically, scientists in academia and other research institutions were the primary communicators of climate change — a logical result of scientists discovering and working to understand this global, but at first “invisible,” problem. With growing understanding of the far-reaching implications of climate change, however, spokespersons...
Figure 2  Research into the use and impact of iconic imagery on issue salience and people’s personal sense that they are able to do something about climate change (self-efficacy) has revealed that images of climate change impacts (such as extensive flooding – photo in (a)) can increase salience but undermine self-efficacy. By contrast, images that help envision a clean energy future (photo in (b)) promote self-efficacy (O’Neil and Smith 2014).

from non-governmental organizations, politicians, government agency representatives, and influentials from other sectors of society (e.g., faith and business communities, entertainment) have become common messengers in the public arena. Communication researchers have examined the various roles they play with regard to science translation, advocacy, agenda setting, education, moral suasion, and political influence. A key factor in the amount of influence is the degree to which the messenger is perceived as a trusted source of information (and judgment) among a given audience (Moser and Dilling 2007).

Communication channels and the role of traditional and new/social media

Another focus in climate communication research is the pathway through which different audiences hear and learn about climate change and through which they communicate about it. The strongest focus has been on mediated forms of communication. In particular, the important roles of the media as communicators, translators, and explainers of emerging science, as shapers of public and policy agendas, as influentials on public understanding of science and the scientific consensus on climate change, but also as amplifiers or attenuators of influential voices in the public debate are key research foci (for ongoing tracking and analysis of climate change media coverage see: http://sciencepolicy.colorado.edu/media_coverage/). A critical understanding of these roles of the media requires a consideration of the political economy of the media industry, journalistic norms, editorial and professional influences on what, how often, and how climate change is being communicated, and of the influence of technology on social interactions, cultural expectations, and changing civic behavior (Boykoff 2011). This has also led to a growing interest in the role and potential/limits of non-traditional media (new and social media, internet, Web 2.0).

An important, if quantitatively smaller, body of work on the topic of communication channels focuses on direct interpersonal, face-to-face, or virtual communication. As the oldest form of communication, the research and practical focus on face-to-face, direct communication and dialog were somewhat neglected in the face of the rapidly rising technology-based, mediated forms of communication. A range of disciplines (from neuroscience to psychology and beyond) have contributed to a better understanding of the observed politicization and polarization in public opinion on climate change, that is, the ideologically and values-driven divergence in public understanding and opinion about climate change, and the resulting antagonistic exchange among divergent factions of society. Based on these insights, dialogic approaches to communication of climate change (use of active, two-way conversation as opposed to one-way information delivery) are being increasingly studied and explored in practice.

Understanding audiences

Deliberate or strategic communication, that is, communication aimed at achieving an intended outcome (such as a certain level of climate literacy, or public support for a certain climate policy), aligns the contents to be communicated with the communication channel(s) and messenger(s), as well as with frames, mental models, and language that are meaningful and resonant with a particular audience. This requires communicators to move away from thinking of a generalized “public” and instead clearly define and understand particular publics so that climate communication can be tailored accordingly.
In fact, considerable research has gone into understanding different audiences. The longest-standing strand of research in this regard has focused on tracking public perceptions of and attitudes toward climate change, as well as public understanding of climate change (science), including identifying dominant beliefs and key gaps in understanding (e.g., Gallup, Pew, and Stanford opinion polls). Because different messengers have come under attack over time, public opinion surveys also commonly now track attitudes toward scientists and other prominent communicators (e.g., on trustworthiness). Opinion surveys are also used to track public preferences among response options. Mostly, these have focused on mitigation such as different energy sources or particular policy mechanisms, for example, energy efficiency standards, a carbon tax, a cap-and-trade system, and so on, but increasingly they also ask about adaptation and geoengineering.

Communication researchers are interested in better understanding different audiences’ belief systems, worldviews, and values that shape how and to what extent climate change information is taken up and how it is interpreted (e.g., cultural cognition of risk, www.culturalcognition.net). Another audience-focused research interest is gaining traction among climate communication researchers as the limits of an exclusively cognitive approach to engaging publics are becoming more apparent: affect and emotional responses to climate change play an important role in people’s responses to this global problem; worry, fears, anger, grief, guilt, hope, and other feelings affect people’s valuation of the seriousness and urgency of climate change, their assessment of the need for and possibilities of personal and societal response, and their level of cognitive, behavioral, civic, and political engagement (Swim et al. 2011).

The cumulative insights on differences among publics have led to different audience segmentation studies. For example, early approaches focused on demographic variables (e.g., age, gender, ethnicity), finding, for example, that women across cultures appear to be more concerned about climate change than men and wish to see a more urgent and comprehensive response. Others have differentiated audiences mostly along political ideological lines, finding that more liberal, left-leaning audiences tend to be more convinced about the science of climate change, show greater concern, and feel greater urgency to take strong action on climate change than their more conservative, right-leaning counterparts. They also have tracked the growing polarization in climate change attitudes and beliefs, particularly in the US. A more recent variant of audience segmentation research distinguishes audiences by prevalent cultural worldviews and belief systems (e.g., Leiserowitz and colleagues’ ongoing survey since 2008 on “Global Warming’s Six Americas,” which is based on cultural theory; Leiserowitz, Maibach, and Roser-Renouf 2008; for regular updates see: http://environment.yale.edu/climate/).

In summary, different audiences share different convictions about the reality, human causation, urgency, and severity of climate change; they trust different messengers, and preferentially use different media channels; and they share different concerns, emotional responses, and preferences for responses (e.g., level of government involvement, regulation vs. free-market, voluntary or individual actions). The notable polarization over the past 20 years of climate change communication, resulting in a self-reinforcing process of opinion hardening, without necessarily increasing climate science literacy, has resulted in a very challenging communication environment. This has led to a growing interest among both researchers and communicators in the
CLIMATE CHANGE COMMUNICATION

possibilities of reducing that polarization, bridging differences and finding common ground among divided societal factions and elevating climate literacy to reduce the gap between the ever-stronger scientific consensus on the reality, causation, and urgency of climate change and lay audiences’ and policymakers’ understanding and responses to climate change.

SEE ALSO: Affect; Behavioral geography; Climate adaptation/mitigation; Climate change, concept of; Climate literacy; Climate and societal impacts; Environment and the media; Environmental citizenship; Environmentalism; Visualization

References


Climate literacy

Lesley-Ann L. Dupigny-Giroux
University of Vermont, USA

Climate literacy refers to both a body of knowledge and a way of understanding and acting in light of this knowledge. It is a distinct subset of science literacy, or the knowledge, skills, and attitudes needed to apply inquiry or problem-based approaches to new situations and decision-making. Climate literacy involves a deep appreciation of the complexity and interconnectedness of the climate system over space and time; the role that humans exert in modifying and interacting with the climate system; the ability to “act accordingly” having understood the above; and the recognition of bias or the change in behavior due to insights gained about an issue or concept (Dupigny-Giroux 2008, 2010). The Essential Principles of Climate Science Literacy (US Global Change Research Program 2009), which were developed through a collaborative process among a number of science agencies, scientists, educators, nongovernmental agencies, and US governmental agencies including NOAA (National Oceanic and Atmospheric Administration), AAAS (American Association for the Advancement of Science), and the National Science Foundation (NSF), state that “People who are climate science literate know that climate science can inform our decisions that improve quality of life. They have a basic understanding of the climate system, including the natural and human-caused factors that affect it. Climate science literate individuals understand how climate observations and records as well as computer modeling contribute to scientific knowledge about climate. They are aware of the fundamental relationship between climate and human life and the many ways in which climate has always played a role in human health. They have the ability to assess the validity of scientific arguments about climate and to use that information to support their decisions.”

One of the keys to unlocking climate literacy is understanding the distinction between weather, climate, the climate system, climate variability, and climate change. The interchangeable usage of these terms in the public lexicon not only leads to confusion and bias, but also erodes the fundamental underpinnings of science literacy in general. Weather refers to the “short-term (minutes to days) variations in the atmosphere” of such variables as “temperature, humidity, precipitation, cloudiness, visibility, and wind” (American Meteorological Society 2000). Climate, on the other hand, refers to “The slowly varying aspects of the atmosphere–hydrosphere–land surface system. It is typically characterized in terms of suitable averages of the climate system over periods of a month or more, taking into consideration the variability in time of these averaged quantities” (American Meteorological Society 2000).

The climate system refers to the processes occurring in the Earth’s spheres, as well as the interaction between them that determines a region’s climate. These spheres are the atmosphere, hydrosphere (lakes, rivers, oceans), cryosphere (frozen ground, glaciers, ice, and other frozen parts of the hydrosphere), biosphere (vegetation on land and marine biota in the oceans), and the lithosphere (crustal matter). The interconnectedness of the climate system suggests
CLIMATE LITERACY

that a systems-based approach be applied to understanding and quantifying climate processes.

It is important to appreciate the nuances of the related concepts of climate variability and climate change. Climate variability encompasses “variations in the mean state and other statistics (such as standard deviations, the occurrence of extremes, etc.) of the climate at all spatial and temporal scales beyond that of individual weather events. Variability may be due to natural internal processes within the climate system (internal variability), or to variations in natural or anthropogenic external forcing (external variability)” (IPCC 2012, 557–558).

Climate change refers to: “A change in the state of the climate that can be identified (e.g., by using statistical tests) by changes in the mean and/or the variability of its properties and that persists for an extended period, typically decades or longer. Climate change may be due to natural internal processes or external forcings, or to persistent anthropogenic changes in the composition of the atmosphere or in land use” (IPCC 2012, 557).

These definitions highlight the timescales involved, the tight coupling between the atmosphere and the other spheres on Earth, and the role of both anthropogenic and natural drivers. What is missing from these definitions is the awareness that uncertainty exists in observations of changes in the climate system, certain drivers of climate change (e.g., aerosol and cloud forcing), recent changes in the climate system, and projections of global and regional climate change (IPCC 2012). While the presence of uncertainty makes it challenging to be “responsive to policymakers’ needs for expert judgment at a particular time, given the information currently available, even if those judgments involve a considerable degree of subjectivity” (Moss and Schneider 2000), uncertainty should not be misconstrued as unreliability or the complete lack of scientific climate knowledge. The 2013 American Meteorological Society’s (AMS) policy statement “Climate Science is Core to Science Education” states that “uncertainty is a natural component of the all scientific endeavor” and “provides a sound example for the broader instruction of the scientific method.” An incomplete appreciation of uncertainty, feedbacks, and scale in the climate system leads to weather and climate becoming either interchangeable concepts or synonymous with only one process (e.g., climate change) or variable (e.g., temperature and precipitation) (Dupigny-Giroux 2010).

As with any discipline or language, climate understanding hinges upon terminology. Core concepts or processes such as teleconnections, electromagnetic radiation, and the hydrologic cycle are not always well understood by nonpractitioners. After learning the rubrics of the climate language, undergraduate students report having a “more cohesive comprehension” and being less likely to question the legitimacy of science in general due to a lack of understanding. Disconnects in terminology and physical processes often exist between the public and climate spheres. For example, Choi et al. (2010) outlined the ways in which the term “air pollution” (soot, aerosols, particulate matter) has become confused with and is now interchangeable with greenhouse gases, and ultimately global warming. Another disconnect lies in statistical climate averages and the usage of the term “normal.” The World Meteorological Organization (WMO) mandates that climate variables be averaged over a consecutive period of 30 years, producing 30-year statistical averages that are also called climate normals. A misconception that only 30 years of data are needed to define climate, or that climate is only relevant to the last 30 years, may have arisen from the layperson’s use of the term “normal” – meaning “the expected value” or “according with, constituting, or not deviating
from a norm, rule, or principle” (Merriam-Webster 2012). Related concepts such as the probability of an event’s occurrence, the mean or average of observations over time, and the fact that these averages change with the length of the observational record and length of time being averaged, are key to quantifying the characteristics and changes of a region’s climate. Finally, the importance of cultural norms and linguistic choices is often overlooked in the translation of science and mathematical concepts or results from one language to another.

Achieving climate literacy can only occur in the context of interdisciplinarity, ranging from the mathematical sciences, through the natural and social sciences, and into the cognitive sciences and psychology. The communication of climate understanding and knowledge must hinge on the lessons derived from cognitive and neuroscience literature on how humans learn and assimilate new information. Social constructivist theory reveals that new knowledge is socially constructed from previous cultural and epistemological experiences and knowledge acquisition, especially at the pre-high school levels. Varying degrees of climate illiteracy emerge when this prior knowledge is riddled with misconceptions about the climate system. Climate misconceptions are held by both students and lifelong learners alike. Many of these misconceptions are related to global climate change, while others point to a fundamental flaw in science literacy in general. Hendriques (2000) identified possible sources of misconceptions in textbook diagrams, verbal explanations, personal observations, and the stories recounted to young children. In summarizing 41 of the most commonly held misconceptions among middle and high school students, Choi et al. (2010) focused on the role of textbooks in contributing to the initial formation of climate misconceptions around types of radiation, the nature of selective absorption, greenhouse gas emission, stratospheric ozone depletion, and the distinction between particulate air pollution and gases.

Perhaps the most pervasive of all climate misconceptions is the notion that stratospheric ozone depletion (which is important for short wavelength ultraviolet radiation) causes tropospheric atmospheric warming (which is a function of the absorption and emission of long-wave radiation). Not only is this misconception particularly resistant to instructional remediation (Chi 2005), but it has been reported worldwide and has its roots at the middle school level. The multiple inaccuracies convolved in this misconception are exacerbated by students’ inability to think or visualize in three dimensions (which affects their understanding of where in the vertical profile of the atmosphere the processes occur) and their flawed notion that stratospheric ozone is a single species while tropospheric ozone is composed of other gases such as carbon dioxide, methane, nitrous oxide, and oxygen.

Dupigny-Giroux (2010) highlighted six challenges to achieving climate literacy. These included the language of the climate sciences (or scientists speaking in code; see also Somerville and Hassol 2011); the persistence of misconceptions; the timing of curricular interventions; the importance of learning styles; the importance of formal educators; and the role of life experience. In the formal academic setting, achieving climate literacy (and science literacy in general) is intricately linked to student preparation in terms of academic content and lifelong learning skills. Undergraduate students report being better able to grasp atmospheric and climate concepts when they had taken either earth science or physics at the pre-baccalaureate level. Similarly, lack of familiarity with the scientific method and scientific habits of the mind (e.g., observation, quantitative analysis, and deductive reasoning) also poses barriers for those who do not consider
themselves as being predisposed to science. Achieving climate understanding will involve targeting both the analytical and the experiential processing systems of the human brain (Center for Research on Environmental Decisions 2009). Climate literacy demands the ability to view connections over time and space, and in the language of Bloom’s taxonomy, to move beyond simple knowledge or remembering to higher orders of applying, analyzing, evaluating, and creating new information. Climate understanding relies on process-based approaches and the ability to synthesize different elements of the land surface. Finally, achieving climate literacy calls for a paradigm shift away from a student-centric world view on the part of students in which they “forget the big picture,” to a broader temporal and spatial view of the climate system around them. This is facilitated by helping students shift from a “fixed intelligence mindset” in which they tend to avoid challenging material, to a “growth intelligence mindset” (Dweck 2006) where they have acquired the habits of the mind needed to embrace and persist in overcoming their personal barriers to climate literacy.

SEE ALSO: Climate change, concept of; Climatology; Geography education: primary and secondary; Geography education, workforce trends, twenty-first-century skills, and geographical capabilities; Geography in higher education

References


Climate policy

Paul A.T. Higgins
American Meteorological Society, USA

Climate policy is decision-making that relates to the climate system. This can include efforts to increase understanding of weather and climate events or to manage risks and realize opportunities associated with current weather patterns, climate variability and climate change (whether natural or caused by humans). Increased knowledge and understanding of the climate system results primarily from scientific observations and research. Weather and climate services help apply that knowledge and understanding for societal benefit (American Meteorological Society Policy Program 2012).

Weather and climate observations reveal dangers from severe weather, create a long-term record for assessing climate variability and change, and provide a rigorous basis for the development, testing, and validation of the models used for forecasts, projections, and predictions. Weather and climate observations provide information on temperature, precipitation, humidity, cloud cover, and other atmospheric conditions. Observations also record physical conditions at the Earth’s surface (e.g., coastal inundation, the status of water resources, timing of lake and river freezing and thawing, etc.) and biological characteristics (species ranges and the timing of seasonal events such as bud burst, flowering, leaf drop, and migration). These observations come from surface (terrestrial, oceanic, and cryospheric), airborne, and satellite-based instruments.

Weather and climate science consists of basic and applied research (analysis and experiments). It may be conducted in the laboratory, the field, or in computer models and is intended to expand our knowledge and understanding of the characteristics and functioning of the climate system. The information that results helps identify and characterize risks and opportunities associated with the climate system, which can enhance commerce and help society minimize (or avoid altogether) weather- and climate-related dangers. Scientific research on the climate system is conducted in academic institutions, government agencies, and the private sector (for-profit and not-for-profit organizations).

Weather and climate services help us to apply knowledge and understanding to benefit society. For example, services can help improve public health and safety, expand economic opportunities, protect environmental resources, and promote national security. Weather and climate services include weather forecasts and warnings, flood and drought prediction and monitoring, natural hazard preparedness and response, public health monitoring, disease prevention and control, assessment and management of fire risk, and decision support for water resources, agriculture, transportation, and other key economic sectors.

For example, weather and climate forecasts support agricultural decision-making such as which crops to plant, when to plant them, and how best to irrigate, fertilize, or control for pests. Weather forecasts and climate projections can also help identify when social challenges or unrest may occur because of crop failures and reduced yields.
Policy choices with respect to the climate system can include decisions about how much to invest in (and how best to conduct) observations, science, and services or specify building codes, land-use patterns, disaster insurance requirements and subsidies, and efforts to monitor, prepare for, respond to, and recover from disasters. These choices influence vulnerability and resilience to weather and climate events.

Government agencies at all levels (national, regional, and local) fund scientific research, determine the balance of investments among disciplines and between basic research and applied objectives, establish and maintain observations, provide weather and climate services, and create weather- and climate-related regulations.

The National Meteorological and Hydrological Services (NMHSs) such as the National Weather Service in the United States provide data, forecasts, and warnings. For-profit companies and humanitarian institutions also provide weather and climate services. These services inform and support routine activities for people and businesses and help protect life and property from extreme weather events.

Climate change as a policy issue

As a policy topic, climate change boils down to four overarching issues: (i) climate is changing, (ii) people are contributing to climate change, (iii) the societal consequences of climate change are highly uncertain but include the potential for serious impacts, and (iv) there are numerous policy options for climate change risk management, most of which are well characterized (i.e., have known strengths and weaknesses) (Higgins, 2014b, from which portions of this section are adapted. © American Meteorological Society. Used with permission.)

Climate is changing. The scientific conclusion that climate is changing is overwhelming because there are many separate lines of evidence that all agree and that have been verified by many different experts. Think of it this way: if you feel heat, smell smoke, hear a fire alarm, and see flames then you have independent confirmation from four senses that there is a fire. The evidence is conclusive. The same is true for climate change. The evidence that climate is changing comes from more than a dozen independent measurements, including: (i) temperature increases in the air measured over land and the oceans using thermometers, (ii) temperature increases in the air measured by satellites, (iii) warmer ocean temperatures (i.e., greater ocean heat content), (iv) melting glaciers throughout the world (the vast majority), and (v) species moving where they live and shifting the timing of their key life events (e.g., migration, reproduction, and periods of activity). These, and other, independent lines of evidence demonstrate that climate is changing.

People are contributing to climate change. Multiple independent lines of scientific evidence demonstrate this as well. Basic math and a growing chemical signature of carbon from fossil fuels demonstrate that people are causing carbon dioxide concentrations to increase in the atmosphere. The warming influence of greenhouse gases is clear based on laboratory experiments, evidence from past changes in climate due to greenhouse gases, and the role of greenhouse gases on other planets (e.g., that Venus is much hotter than Mercury despite being further from the sun).

Additional lines of evidence to link human activities with climate change relate to the patterns of change underway. Think of it like a whodunit where the suspects each have a unique fingerprint – a pattern of climate change they cause. The changes in climate we have witnessed over the past several decades match the patterns expected from greenhouse gases well and do not match the characteristics we would expect from the usual suspects: the sun, volcanoes, aerosols,
land-use patterns, or natural variability. That, in addition to what we know about greenhouse gases, is conclusive evidence that humans are contributing to climate change.

The societal consequences of climate change in the decades ahead are hard to predict because exactly how climate will change and how capable human society will be at absorbing climate impacts are issues characterized by deep uncertainty. This deep uncertainty will almost certainly remain for the foreseeable future.

Notably, different experts reach different conclusions about the seriousness of climate risks. Some experts think the consequences of climate change over the next several decades are likely to be relatively small – perhaps a few percent of GDP. They tend to foresee some combination of stabilizing climate feedbacks, relatively low sensitivity of physical systems, biological resources, and social institutes to climate changes, and a large capacity for society to deal with climate impacts (due, in part, from humanity’s considerable scientific and technological capabilities).

Other experts see climate change as a very serious risk to society. The rationale for this view may include that: (i) the expected changes over the next several decades are unprecedented over the past 10,000 years (i.e., since the start of human civilization); (ii) relatively small changes in climate in the past have, at times, had large societal consequences locally or regionally; and (iii) the physical characteristics of our planet, the biological resources on which society relies, and the social systems that we have developed are often heavily adapted to existing conditions. This increases the potential for changes in climate to be disruptive.

Even in the absence of deep uncertainty over climate change’s specific consequences, policy responses necessarily integrate both objective information about the climate system and our relationship with it, and subjective value judgments (e.g., whether we are more averse to the risks of changes in climate or the policy responses, the ways we assess issues of fairness among nations and peoples, and the consideration we give to cultural heritage or nonhuman species). This creates a complex and often contentious risk management challenge.

Policies relating to climate change can be thought to fall into four broad categories (Figure 1): (i) mitigation – reducing greenhouse gas emissions, (ii) adaptation – increasing society’s capacity to cope with changes in climate, (iii) geoengineering or climate engineering – deliberately manipulating the Earth system to counteract at least some of the impacts of greenhouse gas emissions, and (iv) knowledge base expansion – learning and understanding more about the climate system, which can help reveal risks and opportunities and support risk management efforts.

Reducing emissions is a little like disease prevention (e.g., exercise, eat well, and don’t smoke). Adaptation is like managing illness (e.g., take medicine to cope with symptoms and alleviate problems). Geoengineering is a little like organ transplantation – best avoided but potentially better than the alternative even if you happen to be the first (or only) patient.

Each category of response consists of a family of possible options (described in greater detail below). In some cases the boundaries between the categories become fuzzy (e.g., efforts to reduce emissions might influence adaptive capacity in some cases, and vice versa). Indeed, some consider geoengineering approaches as falling under either mitigation or adaptation. Furthermore, none of the proactive risk management options are mutually exclusive – we could simultaneously enact policies intended to mitigate, adapt, and geoengineer in a range of combinations. Comprehensive climate change risk management almost certainly includes a
Climate change risk management consists of three proactive risk management strategies (mitigation, adaptation, and geoengineering) and efforts to expand the knowledge base with respect to the climate system through research, observations, technology development, and scientific assessments (Higgins 2014a).

### Mitigation

By reducing emissions, mitigation decreases society’s future contributions to greenhouse gas concentrations in the atmosphere. This translates into smaller future changes in climate, which increases the chance that societal impacts will remain manageable. However, the climate has already changed and will continue to warm due to past emissions, which makes some climate impacts unavoidable. Mitigation does little to help with changes in climate that have occurred or are already entrained.

Mitigation could also create risks. Some approaches could reduce access to energy or cause excessive energy price increases. Adverse secondary consequences are also possible. For example, policies to promote biofuel production could lead to inefficient uses of land, water, or agricultural crops. Biofuel use could also reduce air or water quality.

However, mitigation might also confer benefits unrelated to climate change (often called co-benefits). For example, reducing emissions of greenhouse gases is expected to reduce more traditional forms of pollution associated with coal-fired power plants (e.g., mercury emissions, smog, and acid rain), which would benefit public health. Similarly, mitigation would likely lead to a reduction in oil consumption. This would help reduce impacts associated with oil drilling, transport, and use, and lessen foreign oil dependence, which could improve the environment, the economy, and national security.

Approaches to reducing emissions fall into several broad categories. These include: (i) regulation; (ii) research, development, and deployment of new technologies; (iii) conservation; (iv) efforts to increase public awareness; (v) positive incentives to encourage choices that lower emissions; and (vi) adding a price to greenhouse gas emissions, which creates an overarching incentive to reduce emissions.

Regulations often specify what activities are permitted and the manner in which they may be conducted. This can include (among others) specifying fuel efficiency standards for vehicles, determining land-use practices, establishing mandates to use specific emission control technologies, making harmful practices illegal, establishing renewable energy requirements, or enacting building codes and construction practices that help reduce energy consumption (e.g., requiring the use of energy-efficient appliances and minimum amounts of insulation).

Research, development, and deployment can help create or improve next-generation technologies, including those that might reduce emissions. Conservation of energy or biological resources (e.g., forests and wetlands) can help reduce emissions associated with energy use or
land-use practices. Public awareness campaigns can help ensure that the public understands the implications of their choices and encourage individuals to adopt practices that reduce emissions. Similarly, positive incentives such as tax breaks or subsidies can help shape consumer preferences toward products and choices that result in lower emissions.

Adding a price to greenhouse gas emissions is a particularly noteworthy policy option because it would be expected to have a broad-reaching impact on emissions, has received a great deal of attention from the research community, and has been a focus of policy discussions since climate change emerged as a public issue.

Three economic principles suggest that adding a price to greenhouse gas emissions might be a beneficial way to manage climate change risks. The first economic principle relating to emission pricing is that having less of something (greenhouse gas emissions, in this case) almost certainly requires an increase in the price of those activities that cause it. This is because a price increase for emitting activities causes emissions reductions by encouraging efficiency (a reduction in emissions for a given amount of the activity) and also frugality (more sparing engagement in the activity) (Daly 2007). Critically, increasing the efficiency of an activity without a corresponding increase in the price makes engaging in the activity cheaper, which encourages more of the activity. As a result, efficiency gains without an increase in the price of an activity may not lead to emissions reductions.

The second economic principle is that incorporating the costs associated with climate change into the price emitters pay for their emissions (i.e., through an additional price on emissions) would be expected to increase overall economic wellbeing. This is because maximum economic benefits are realized when individual decision-makers (the entity choosing to emit, in this case) pay all costs and receive all benefits associated with the activity.

Currently, the societal consequences of climate change are distributed across the entire population, including future generations. As a result, potentially significant economic costs associated with greenhouse gases (i.e., the societal costs of climate damage) are not paid for by those who emit. Instead, the people who endure the consequences of climate change pay those costs. This constitutes an economically harmful subsidy that emitters receive from the broader society. Incorporating the costs of climate damage into the price paid by emitters would reduce that subsidy and therefore bring net economic benefits.

Note, however, that greenhouse gas emissions result, in part, from six separate market failures (Higgins 2010). These include: (i) that the cost of climate damages associated with emissions are not included in the price paid by the emitter (i.e., a negative externality is unaccounted for, as described above); (ii) split incentives, in which the narrow interests of a decision-maker are maximized when creating much higher costs for someone else (e.g., a landlord’s incentive to minimize capital investment expenses even when doing so ensures that their tenants’ excess energy expenses will be greater than the landlord’s savings on capital equipment); (iii) imperfect information, in which decision-makers do not know or understand their options and the implications of their choices; (iv) monopoly power, which limits consumer choices for low-emission alternatives; (v) long-lived (fixed or immobile) factors of production, which locks in less efficient technologies because the existing capital stock makes emitters less responsive to market signals; and (vi) the absence of a market for climate stability (a public good) because the private sector cannot provide and price public goods.

Adding a price to emissions addresses the first market failure (the externality) but is insufficient
or ineffective at addressing the other market failures. This means that including a price on emissions that accounts for climate damages associated with emitting would likely be insufficient for fully addressing all the relevant market failures that contribute to human caused climate change (i.e., greenhouse gas emissions).

Furthermore, we cannot quantify precisely the cost of climate damage associated with emitting greenhouse gases because the consequences of climate change are characterized by deep uncertainty, as described above. This means we cannot know the economically optimal price to add to emissions. In practice, adding a price to greenhouse gas emissions will either be too high or too low to maximize economic benefits.

A price that is too high (i.e., exceeds the cost of climate damage associated with emissions) would sacrifice some economic wellbeing relative to a lower price because energy prices would be too high. A price that is too low (i.e., fails to account for the costs of climate damage) would sacrifice some economic wellbeing relative to a higher price because too much climate damage would occur. Risk aversion to climate change implies erring on the side of a price on emissions that is more likely to be too high than too low, whereas risk aversion to price increases for energy implies erring on the side of a price that may result in excessive climate damage.

The third economic principle related to emissions pricing is that market mechanisms are generally the most economically efficient way to reduce emissions. This means that a price-based approach can be expected to result in the greatest amount of emissions reduction for the least cost or, equivalently, the most emissions reduction for a given cost. An important caveat to this basic conclusion is that it applies when the externality constitutes the dominant market failure because the additional market failures described above are not entirely responsive to the emissions price approach and therefore may be more effectively addressed through regulation or another policy option.

Despite these basic economic principles, adding a price on emissions may be insufficient or undesirable. As noted above, the price on emissions will necessarily be too high or too low because we cannot know what the actual damage to the climate system will be from a given amount of the emissions, and a price-based approach cannot address the full range of market failures.

Furthermore, adding a price to greenhouse gas emissions could have significant distributional consequences (i.e., there would be winners and losers) even while overall economic benefits would be expected to increase. Depending on the specific details of the policy design, those distributional consequences could be particularly significant for heavy emitters or low-income families. In addition, losses are likely narrowly distributed whereas benefits are broadly distributed. This creates both legitimate questions of fairness and political challenges for climate policy (described below).

Some policy options for mitigation also require allocating scarce resources toward emissions reduction efforts (e.g., investing in low-emission technologies). In the event that the consequences of climate change turn out to be less harmful than expected, investments in mitigation could constitute an inefficient use of limited resources. Note, however, that this does not apply to approaches that reduce market failures (e.g., adding a corrective price on emissions). Although emission pricing at first seems to require the use of scarce resources because energy prices may rise, it is actually reducing a hidden subsidy (i.e., reducing societal investments in climate damage) as previously described.

Notably, societal values other than maximizing economic efficiency also matter, perhaps
more than economic efficiency, in some cases. For example, fairness, the role people play in determining the Earth’s characteristics and functioning, and how to weigh the impacts on cultural heritage or other species are all questions outside of economic efficiency.

In general, there are two market-based approaches for adding a price on emissions (Higgins 2010). Policymakers can set a limit on the amount of emissions (a cap or limit on the quantity of emissions) and allow emitters to buy and sell permits to emit. This approach (often called cap-and-trade) leaves it to the market to determine the price of emitting. Alternatively, policymakers can determine the price that emitters must pay when they emit (a fee or a corrective tax). This approach leaves it to the market to determine the quantity of emissions. Notably, these two approaches have much in common because emission prices and quantities are linked. Both are market mechanisms for addressing climate change.

Hybrid approaches that combine elements of both approaches are also possible. For example, cap-and-trade can include a price ceiling (an upper limit on prices at which additional permits are always sold) or a price floor (a minimum price on emissions at which permits are always purchased). Similarly, fee-based hybrids can include automatic increases in the price if emissions quantities exceed an upper limit, thereby ensuring a minimum level of climate protection.

Policy deliberations for these approaches must determine the initial price (or quantity) of a pricing mechanism and the rate that it changes over time (Higgins 2010). Higher prices (or lower quantities) translate into larger, faster emission reductions, but may trigger larger price increases for energy and transportation.

Emission fees or permits can be collected at the oil well, coal mine, or point of entry for imports (often called “upstream”), closer to where the actual emissions occur, such as the tail pipe or power plant (downstream), or in between (e.g., petroleum refineries). Upstream implementation helps ensure comprehensive coverage of emissions, generally reduces the administrative burden placed on regulators and emitters, and minimizes transaction costs. Downstream implementation more closely aligns the point of regulation with the point where emissions occur. However, where the fee is collected likely does not affect who pays the fee, because market forces generally determine the relative burden on producers and consumers.

Revenues generated from adding a price on emissions could be used in a wide range of ways. For example, revenues can be used to lower existing taxes, returned in a lump-sum payment to citizens, used to promote research and development of low-emission technologies, or assist those most heavily hit by the fee, among other options. How these revenues are used can reduce or exacerbate distributional consequences of emission pricing. For example, a tax shift – one that applies the revenue generated from emission pricing to lowering existing taxes – or the lump-sum return of revenues to citizens on an equal per capita basis could increase the progressivity of the approach. Similarly, the disproportionate impact on heavy emitters can be softened by providing a small number of permits freely (under cap-and-trade) or by directing some of the revenue generated by a fee to hard-hit sectors.

Emissions pricing can include offsets or credits for emissions reductions that occur elsewhere (e.g., carbon capture and sequestration, forestry projects, and international mitigation efforts) and the banking or borrowing of permits in the case of cap-and-trade. Offsets can encourage emission reductions and reduce the costs of achieving a given level of climate protection. However, offsets also pose challenges because seemingly legitimate reductions may not last over
time. Borrowing and banking can also help even out price fluctuations in a cap-and-trade system.

Adaptation

Adaptation involves planning for climate impacts, building resilience to those impacts, and improving society’s capacity to respond and recover. This can help reduce damage and disruptions associated with climate change. Adaptation might also help with existing threats due to current weather patterns (e.g., routine and severe weather events) or from other natural and human-induced disasters unrelated to climate changes. This represents an important potential co-benefit of efforts to build adaptive capacity to climate change.

Adaptation also has limits or downsides. It is possible that some climate impacts will be too severe to manage through adaptation. Efforts to promote adaptive capacity could also prove maladaptive (counterproductive) due to uncertainties over future climate projections and the expected impacts of climate change on physical systems, biological resources, and social institutions.

Adaptation policy can include regulation to decrease vulnerability (e.g., through land-use planning and building codes), response planning, disaster recovery, impact assessment for critical systems and resources (e.g., water, health, biological systems, agriculture, and infrastructure), observations and monitoring, and efforts to minimize compounding stresses such as traditional air pollution, habitat loss and degradation, invasive species, and nitrogen deposition.

Implementing adaptation policies successfully may require detailed consideration of location-specific factors because climate change impacts will vary geographically and depend on the uneven distribution of societal resources and institutions. As a result, centralized policy responses may be somewhat more limited for adaptation than for mitigation or geoengineering.

Nevertheless, centralized regulations have the potential to promote adaptive capacity by altering land-use patterns on a wide scale (e.g., floodplain development, management of coastal zones, and insurance practices) in ways that increasingly account for future climate change impacts. Similarly, centralized approaches to disaster relief efforts, the establishment and design of wildlife reserves, and management of water and agricultural resources could all help account for vulnerabilities anticipated by climate change.

Centralized adaptation policies can also potentially promote decentralized efforts by creating broadly useful sources of: (i) scientific information about climate change impacts and vulnerabilities; (ii) information about the advantages and disadvantages of particular adaptation measures and the specific conditions under which different options work best; (iii) support for (or incentives to encourage) local and or regional-level adaptation planning and implementation, including the provision of technical expertise and/or financial resources; and (iv) monitoring and reporting on the effectiveness of adaptation efforts.

Geoengineering

Geoengineering refers to deliberate, often global-scale, manipulations of the climate system (American Meteorological Society 2009). In general, the goal of geoengineering would be to counteract the effect of human greenhouse gas emissions or some of their consequences. Geoengineering could help to lower greenhouse gas concentrations in the atmosphere, counteract the physical impact of increasing greenhouse gas concentrations, address specific climate change impacts, or offer desperation strategies in the event that abrupt, catastrophic, or otherwise
unacceptable climate change impacts become evident.

Geoengineering could also create new sources of risk because attempts to engineer the Earth system on a large scale could lead to unintended and adverse consequences. Notably, the complexity of the Earth system (which couples numerous physical and biological systems and processes) and society’s relationship to the Earth system (which involves further coupling with social institutions) makes it challenging for scientific research to fully identify and quantify the potential consequences associated with geoengineering. As a result, the possible impacts from geoengineering could inadvertently compound the dangers associated with climate change.

Even to the extent that potential consequences of geoengineering can be well characterized, those consequences would almost certainly differ among countries and individuals. This raises complex legal, ethical, diplomatic, political, and national security concerns. The use of geoengineering as a desperation strategy could also distract from mitigation and adaptation, either of which may have a higher probability of contributing positively to risk management efforts.

Nevertheless, two categories of geoengineering are most prevalent within scientific and policy discussions: solar radiation management and carbon removal and sequestration. These, and other approaches, likely wouldn’t address all possible impacts associated with greenhouse gas emissions. For example, solar radiation management will not reduce the amount of carbon dioxide in the air or the ocean, and would therefore have no impact on ocean acidification or other direct effects of carbon dioxide enrichment on biological systems.

Policy options for geoengineering generally fall into five categories. We could conduct research and analysis in order to develop or vet options. We could study the impacts and likely unintended consequences. We could create punitive measures to discourage unilateral (or what might be considered reckless) attempts to geoengineer. We could create policies that promote cooperation and transparency or help ensure that governance issues would be addressed. Of course, policies could also seek to implement geoengineering approaches. To date, policymakers have rarely considered geoengineering explicitly.

Expanding the knowledge base

Policies can also be designed to expand the knowledge base relating to the climate system (described above) or to reveal information relating to the management of risks associated with climate change.

Research, observations, scientific assessments, and technology development can help reveal risks and opportunities associated with the climate system and support decision-making with respect to climate change risk management. Expanding the knowledge base allows policymakers to understand, select, and refine specific risk management strategies, and to thereby increase the effectiveness of risk management efforts. Knowledge base expansion can, in some cases, also reveal entirely new opportunities for protecting the climate system or reducing the risks of climate change impacts. As a result, policies to expand the knowledge base can underpin and support the proactive risk management strategies described above (mitigation, adaptation, and geoengineering).

Climate system research spans numerous disciplines and subdisciplines, including those within the atmospheric sciences, oceanography, hydrology, biology, cryology, and paleoclimatology, among others. Determining the societal consequences of climate variability and change depend on understanding how human systems depend on and will respond to potential impacts on
physical systems, biological resources, and social institutions. That also requires information from disciplines in the social sciences, including (but not limited to) economics, sociology, history, and political science (Steinbuck and Higgins 2013).

The political landscape

There are several political obstacles to climate change risk management in general and to pricing greenhouse gas emissions in particular. Climate change can be characterized as a “wicked problem” (Rayner 2006). This means that climate change, as a public and policy issue, is characterized by: (i) contradictory certitudes (i.e., different people believe – as fact – different things that are actually incompatible), (ii) having redistributive implications for entrenched interests, (iii) being related to deeper problems (e.g., the scale of human activities relative to the Earth system’s capacity to withstand disturbance), (iv) having relatively little room for trial-and-error learning, and (v) tending to be incompletely solvable (i.e., we must live with climate change in some sense).

As a result, policy deliberations (and public debates) about climate science are often at odds with the assessments of the relevant subject-matter experts (Leiserowitz et al. 2014). This is because the complexity of the issue and the underlying science increases the potential for nonexperts to be unaware of expert assessments and more prone to believing rhetorical arguments that seem convincing, even when those arguments do not withstand the scrutiny of subject-matter experts. This contributes to political polarization with respect to climate change. Furthermore, those who would tend to be hurt by climate change risk management (particularly through emission pricing) often know that policies could harm them, care about a relatively small number of issues other than climate policy, and tend to be politically powerful and well organized. In contrast, the winners from emission pricing often do not realize they will benefit (i.e., because they take the climate system for granted and do not recognize the risk emissions pose for climate). Winners also tend to care about a wide range of other issues more than climate change. Because benefits are broadly distributed, the constituency for climate protection is a relatively disorganized group that is politically weak. These differences between the winners and losers create significant political obstacles to enacting policy solutions.

To some degree, these challenges are exacerbated at the international level because the contribution to (and risk exposure from) climate change are unequal among nations. For example, many of the impacts of climate change are expected to be most severe for countries and peoples who have contributed minimally to the atmospheric stock of greenhouse gases (e.g., low-emitting small island states, which face limited options for adaptation, and developing countries, which may be highly vulnerable to climate change and possess limited adaptive capacity). This separates the sense of urgency to respond to climate change from the capability of doing so. Of course, these differences also contribute to the complex ethical dimensions of climate policy.

The global nature of climate change also creates future challenges for climate policy. There is a genuine need for a coordinated global effort to reduce emissions because atmospheric greenhouse gas concentrations are well mixed (i.e., they are based on emissions anywhere in the world). This makes unilateral action more difficult and potentially less effective. It also contributes to powerful rhetorical arguments against mitigation efforts. Why should nation A begin to reduce its emissions when nation B has no similar plan to do so?
There are policy solutions to these challenges. For example, unilateral action by one nation could be conditional on other countries making similar efforts, or could include incentives for cooperation such as border tax adjustments to account for those who do not incorporate climate damage into prices paid by emitters (which constitutes a subsidy).

Voluntary emission targets and actions taken by individual nations may also encourage similar efforts by other nations and thereby create a positive feedback cycle for emissions reductions. This is a central component of the current international approach to reducing greenhouse gas emissions, as established by the Copenhagen Accord in 2009 and further implemented through the Paris Agreement in 2015.

Note: This entry is based on and adapted from Higgins (2014a). © American Meteorological Society. Used with permission.

SEE ALSO: Climate change communication; Climate literacy; Climate and societal impacts; Environmental governance; Geoengineering/climate intervention; Social vulnerability and environmental hazards

References

Climatic modes and teleconnections

Andrew M. Carleton
Pennsylvania State University, USA

Climate variations, climatic modes, and teleconnections

Temporal departures from the long-term means of climate variables such as temperature, precipitation, humidity, cloud cover, and winds, tend to be organized spatially and to be internally consistent for a given climate region (e.g., warmer than normal, drier than normal, reduced cloudiness, etc.). Climatic modes are recurring climate anomaly patterns on regional to hemispheric and even global scales accompanying intraseasonal, interannual, and multiyear variations of the atmospheric circulation. Intraseasonally (i.e., 30–90 days), these modes often are linked to deep convection pulses over the tropical Indian Ocean that spread both eastward and poleward to influence the weather and climate of extratropical latitudes, and which help modulate the interannual climate modes. Prominent examples of the latter include the opposite-signed climate anomalies in western versus eastern North America in the colder months, and between Greenland/Iceland and western Europe. On hemispheric and larger scales, important interannual climatic modes include the opposing anomalies of temperature between middle and high latitudes of the Northern Hemisphere, and of precipitation between the western and eastern tropical Pacific Ocean.

The spatial coherence of climate anomaly patterns (i.e., alternating positive and negative departures) on regional to global scales, expresses their association with atmospheric circulation as a teleconnection, or long-distance (tele-) repetitive association (connection) of climate variations. For example, the European heat wave of summer 2003 was accompanied by persisting cool anomalies further eastward in western and central Russia. A teleconnection is primarily the statistical manifestation of a climatic mode but also implies a physical explanation for such modes. Thus, in the foregoing examples, the tropical Madden–Julian Oscillation (MJO) comprises the dominant source of intraseasonal climate variability for the globe; the Pacific–North America (PNA) pattern accompanies the opposition in climate anomalies between western and eastern North America; the North Atlantic Oscillation (NAO) expresses the Greenland–Europe anomaly couplet; the Arctic Oscillation (AO) projects onto opposite-signed temperature anomalies in middle versus higher latitudes; and the El Niño Southern Oscillation (ENSO) explains tropical precipitation anomalies between the western and eastern Pacific, and both temperature and precipitation anomalies for the extratropics of both hemispheres (Table 1). On decadal and multidecadal timescales, climatic modes and teleconnections express the dominant roles of temperature, salinity, and wind stress in the ocean gyres; the Pacific Decadal Oscillation (PDO – or “ENSO-like” pattern – has a period of about

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0494
### CLIMATIC MODES AND TELECONNECTIONS

**Table 1** Major teleconnections and their key attributes.

<table>
<thead>
<tr>
<th>Name (acronym)</th>
<th>Spatial scale</th>
<th>Timescale</th>
<th>Index defining variable(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Madden–Julian Oscillation (MJO)</td>
<td>Hemispheric/Global</td>
<td>Subseasonal</td>
<td>OLR, surface winds</td>
</tr>
<tr>
<td>Quasi-Biennial Oscillation (QBO)</td>
<td>Hemispheric/Global</td>
<td>Biennial</td>
<td>Tropical stratospheric winds</td>
</tr>
<tr>
<td>El Niño Southern Oscillation (ENSO)</td>
<td>Global</td>
<td>Interannual</td>
<td>SLP, SST, OLR</td>
</tr>
<tr>
<td>Arctic Oscillation (AO)</td>
<td>Hemispheric</td>
<td>Monthly → interannual</td>
<td>SLP, temperature</td>
</tr>
<tr>
<td>Antarctic Oscillation (AAO)</td>
<td>Hemispheric</td>
<td>Monthly → interannual</td>
<td>SLP, geopotential height</td>
</tr>
<tr>
<td>Semi-Annual Oscillation (SAO)</td>
<td>Subhemispheric</td>
<td>Interseasonal</td>
<td>SLP, height, winds.</td>
</tr>
<tr>
<td>Trans-Polar Index (TPI, Southern Hemis.)</td>
<td>Subhemispheric</td>
<td>Monthly → interannual</td>
<td>SLP, height</td>
</tr>
<tr>
<td>North Atlantic Oscillation (NAO)</td>
<td>Regional</td>
<td>Monthly → interannual</td>
<td>SLP, height</td>
</tr>
<tr>
<td>North Pacific Oscillation (NPO)</td>
<td>Regional</td>
<td>Monthly → interannual</td>
<td>SLP, height</td>
</tr>
<tr>
<td>Pacific–North American Oscillation (PNA)</td>
<td>Regional</td>
<td>Monthly → interannual</td>
<td>Height, SLP</td>
</tr>
<tr>
<td>Indian Ocean Dipole (IOD)</td>
<td>Regional/Subhemispheric</td>
<td>Seasonal → interannual</td>
<td>SST, OLR, sea surface height</td>
</tr>
<tr>
<td>Meridional Index (M1, Southern Hemis.)</td>
<td>Regional</td>
<td>Monthly → interannual</td>
<td>SLP</td>
</tr>
<tr>
<td>Antarctic Circumpolar Wave (ACW)</td>
<td>Subhemispheric</td>
<td>8 years</td>
<td>SLP</td>
</tr>
<tr>
<td>Pacific Decadal Oscillation (PDO)</td>
<td>Subhemispheric</td>
<td>40–60 years</td>
<td>SST</td>
</tr>
<tr>
<td>Atlantic Multidecadal Oscillation (AMO)</td>
<td>Subhemispheric</td>
<td>70 years</td>
<td>SST</td>
</tr>
</tbody>
</table>

40–60 years (i.e., each phase is 20–30 years), while the Atlantic Multidecadal Oscillation (AMO) has a periodicity of around 70 years.

Any teleconnection comprises two extreme phases marked by a reversal in sign of the sea level pressure (SLP) or tropospheric height anomalies defining the pattern in the key region. These phases are typically denoted as positive and negative. For example, the AO extreme phases comprise the “high index” (colder high latitudes, warmer middle latitudes) and “low index” (warmer high latitudes, colder middle latitudes), whereas the ENSO extreme phases are named La Niña (cold event or intensified...
normal pattern) and El Niño (warm event or dry western Pacific, wet central and eastern Pacific).

Teleconnection patterns can be classified into two broad types: zonally symmetric–annular–and zonally varying, or meridional. In both the Northern and Southern Hemisphere extratropics, the dominant modes are annular, comprising the AO and the Antarctic Oscillation (AAO), both of which are expressed in the stratosphere as well as troposphere. Meridional patterns involve longitude and amplitude anomalies of the tropospheric long (i.e., stationary or Rossby) waves and may be embedded within – and highly correlated to – the annular modes; prime examples being the NAO in relation to the AO and the Southern Hemisphere’s Trans-Polar Index (TPI) and AAO (Table 1). Thus, although usually defined as discrete (i.e., independent) phenomena, a teleconnection can correlate significantly with others; for example, the PNA, NAO, and AO in the North America sector during certain winters, and their modulation by the MJO, thereby increasing their area of influence. Winter climate anomalies in western North America often are of the same sign as those in the Labrador–Greenland area, while those in eastern North America can have the same sign as Europe.

The reduced land–sea contrasts in Southern Hemisphere middle latitudes versus the Northern Hemisphere impart a lower amplitude to the stationary waves. Thus, teleconnection characteristics can be quite different between the Northern and Southern hemispheres (e.g., the Pacific–South America (PSA) pattern versus the PNA), with shorter duration times for intraseasonal “blocking” episodes. Conversely, the semi-annual (interseasonal) oscillation – SAO – of middle and subpolar latitudes is much stronger in the Southern Hemisphere than in the Northern Hemisphere, where the annual cycle dominates.

The persistence of climatic modes and teleconnections on timescales ranging from weeks to months and even across seasons and years, results primarily from positive feedbacks – reinforcing nonlinear interactions – between the atmosphere and Earth surface climate variables; notably the sea surface temperature (SST), soil moisture and vegetation activity, snow–cover latitude extent and depth, and sea ice conditions (extent, concentration). The Earth surface climate anomalies and their atmospheric interactions often are evident as departures in large-scale satellite-retrieved variables of the outgoing longwave radiation (OLR), and Normalized Difference Vegetation Index (NDVI). Indeed, fluctuations in these and other variables frequently are used to monitor teleconnections and their accompanying climate anomalies. Much of the pioneering work on teleconnections – particularly that by Jerome Namias (1969) and J. Bjerknes (1969) – used the association with anomalies in surface conditions to help predict circulation climate on monthly to seasonal timescales, or “medium- and long-range weather forecasting.” The interactions are thermal (involve the radiation and energy budgets), hydrologic (involve the moisture budget), and dynamical (involve the atmospheric circulation). Anomalies in the number, position, and amplitude of the tropospheric troughs and ridges accompanying teleconnections, and their associated jet streams and synoptic circulation systems of high and low pressure, impart a juxtaposition of positive and negative surface climate anomalies over space.

History of teleconnections research
and teleconnections in climate history

Teleconnection studies have a long history. The term dates back to Ångström (1935) but studies of climatic fluctuation patterns include H.H.
Hildebransson’s (1897) discovery of inverse SLP variations between southeast Australia and southern South America (Barry and Carleton 2001), Sir Gilbert Walker’s pioneering efforts nearly 100 years ago to link Indian monsoon rainfall with SLP variations over the tropical Pacific Ocean (the “Southern Oscillation”) and elsewhere, and J. Bjerknes’ work in the 1960s linking tropical Pacific sea surface temperature (SST) variations to fluctuations in the extratropical westerlies. During the 1980s and 1990s, an explosion of research identified a large number of teleconnection patterns either accompanying extratropical climate anomalies or modifying the influence of tropical teleconnections. In the 2000s came the realization that improvements in both weather and climate prediction occur with explicit consideration of the MJO phase. Additional advances included numerical modeling that attempted both to reproduce the dominant teleconnections and to clarify the physical processes linking the atmospheric circulation fluctuations with the Earth surface anomalies. Much recent work assesses the long-term stability of teleconnection–climate associations given the contemporary climate changes related to trace gas increases (“global warming”) or other human activities (e.g., tropical deforestation, irrigated agriculture), including proposed climate geoengineering solutions.

In terms of climate history, the Medieval Climate Anomaly (MCA) and subsequent Little Ice Age (LIA) are partly explained by teleconnections: during the MCA, an extended La Niña-like pattern with some resemblance also to positive (+) NAO, especially in the Europe sector; during LIA, an extended El Niño-like pattern and one resembling the meridional (−) NAO. Spatial patterns of Northern Hemisphere temperature trends in the last quarter of the twentieth century also suggest a strong association with teleconnections, only comprising an El Niño pattern in the Pacific and + NAO. However, trends since about 2000 indicate more the global warming signal, especially in higher northern latitudes, superimposed on a greater frequency of the low index (−) AO pattern.

### Teleconnection indices

A number of statistical methods can extract the dominant spatial pattern(s) of variation in, for example, SLP, geopotential height, or OLR, on regional to subcontinental and larger scales, although these have favored empirical orthogonal function (EOF) analysis. More recently, self-organizing map and artificial neural networks have clarified and extended the impact of different synoptic circulation regimes in teleconnection phases. The role of a given teleconnection in climate anomalies often is quantified using correlation or composite (map average) approaches for the extreme phases.

Once a teleconnection pattern has been confirmed as robust (e.g., from time series analysis of eigenvalues), it is useful to depict it as an index that can be updated monthly or seasonally for climate monitoring purposes. Indeed, in areas lacking long-term mapped atmospheric data – the Southern Hemisphere oceans in particular – teleconnection indices have been crucial for assessing climate linkages to large-scale SST and sea ice variations. Although both qualitative and quantitative indices exist (e.g., for ENSO), the majority of teleconnection indices involve measurements of climate variables that express the phenomenon. Because most teleconnections involve anomalies in SLP or geopotential height of the major circulation “centers of action,” their related indices typically utilize monthly or seasonal values or their differences from the long-term (climatological) normals.
For example, the NAO index (NAOI) traditionally comprises the SLP anomaly difference between Portugal or the Azores (North Atlantic High) and Reykjavik (Icelandic Low). From the geostrophic wind equation, the NAOI thus captures variations in the strength of the westerlies over the North Atlantic. These winds are stronger during +NAO, and weaker for −NAO, with concomitant variations in temperature and precipitation anomalies for western and northwestern Europe (i.e., milder and wetter in +NAO, colder and drier in −NAO). Somewhat analogously, the PNA pattern was originally defined by Wallace and Gutzler (1981) as simultaneous variations in 500 hPa heights at four grid points representing the waves over North America and the northeast Pacific, although it can also be defined using three grid points. Both the AO and AAO are defined using zonally averaged SLP anomaly differences between middle and high latitudes (Table 1). The tropical Quasi-Biennial Oscillation (QBO) is defined primarily by the lower-stratospheric winds; either in their easterly or westerly phases. Regional-scale teleconnection indices can depict variations in meridional winds; for example, the M1 index captures fluctuations in the trough in the Tasman Sea, whereby a weakened trough (i.e., anomalous ridge) often accompanies El Niño.

Recent improvements to teleconnection indices include, for example, consideration of spatial variations in location of the centers of action that may not be captured by station SLP observations; a monthly NAOI that considers the grid points of lowest and highest pressure in the Icelandic Low and Azores High, respectively. Other teleconnection indices may be refined using additional climate variables involved in the pattern. Thus, the traditional ENSO index is the SLP anomaly difference between Tahiti (i.e., South Pacific High) and Darwin (northern Australia monsoon low), or Southern Oscillation Index (SOI). Because the ENSO is primarily a tropical ocean–atmosphere interaction mode, fluctuations in tropical Pacific SST and OLR improve upon the SOI; notably in the Multivariate ENSO Index, MEI (Figure 1).

Associations between teleconnections may be either constructive (i.e., reinforcing or amplifying) or destructive (damping), with concomitant impacts on the climate signal. Thus, a positive or warm phase of the PDO tends to enhance the frequency and/or intensity of El Niño events, while a negative (cool) phase enhances La Niñas. These relationships can be evident in the frequency and area burned by wildfires in the western United States, for example. Finally, analysis of longer-term Pacific SST data can refine the subregion of greatest sensitivity to ENSO, consideration of which improves ENSO forecasting (e.g., the Niño 3.4 vs Niño 1 + 2 regions).

**Physical processes and mechanisms**

Because the tropics experience an annual surplus of solar and net radiation, they are the Earth’s heat reservoir; energy and moisture move poleward on average, typically along relatively narrow corridors in the atmosphere that are influenced by synoptic phenomena on weather timescales and by climate teleconnections. Extratropical to tropical atmospheric interactions may also occur, such as East Asian cold-air surges that help trigger enhanced convection in the western tropical Pacific, and which may, in turn, propagate out of the tropics as moisture plumes or “atmospheric rivers.” Despite their persistence in a given phase, teleconnections can shift quite abruptly to the opposing phase, resulting in a similarly abrupt reversal of climate anomalies in region(s) influenced by the teleconnection. This behavior manifests the “almost intransitive”
character of Earth’s climate system, whereby large fluctuations or changes may occur with only minimal changes in boundary conditions.

The dominant global-scale teleconnection pattern – ENSO – arises from strong ocean–atmosphere interactions in the tropical Pacific that are transmitted to the extratropics by standing waves and the subtropical jet stream (STJ). During neutral and also cold (i.e., La Niña) events, negative anomalies of SST (−SSTA) occur in the eastern and central tropical Pacific, positive anomalies in the western (“warm pool”) area. However, during El Niño events the SST anomaly pattern is reversed and features an area of +SSTAs near the International Date Line, extending eastward and along the west coast of South America (Figure 2). The basin-wide pattern of tropical Pacific SSTAs evident in Figure 2 combines two “flavors” of El Niño: that which develops primarily near coastal Peru (or traditional El Niño, 1st EOF), and that which develops in the central tropical Pacific, or Modoki El Niño. This 2nd EOF features −SSTAs both to the east and west of the +SSTAs, and occurs in summer and winter. Its extratropical climate impacts can also differ from the traditional El Niño.

Over +SSTAs and upper-ocean warmth (i.e., depressed thermocline) in the tropics, atmospheric convection is enhanced (negative OLR anomalies) due to increased evaporation and instability of the air, and SLP tends to be low. Conversely, over −SSTAs and upper-ocean coolness (thermocline near the surface), convection is suppressed – air sinks and cloud dissipates (positive OLR anomalies) – in the more stable atmosphere, with corresponding higher SLP. Precipitation anomalies, such as those retrieved from the tropical rainfall measuring mission (TRMM) satellite, accompany the OLR anomalies: negative OLR with higher daily rain rates; positive OLR with low or zero rain rates. Thus, when the SOI or MEI (Figure 1) is either close to the long-term normal or indicates La Niña, most deep convection is located in the western tropical Pacific – an intensified North Australian summer monsoon – with subsiding air in the eastern tropical Pacific. At these times, the Walker (west–east) Circulation is enhanced relative to normal. During both traditional and Modoki El Niño events, subsiding air in the western tropical Pacific weakens the Australian monsoon, whereas the +SSTAs in the central and eastern tropical Pacific during traditional

Figure 1  Time series of the MEI from 1950 through February 5, 2016. Positive values (red) indicate El Niño events, negative values (blue) indicate La Niña events. Source: http://www.esrl.noaa.gov/psd/enso/mei/.
El Niño promote increased convection and precipitation (Figure 3).

SSTAs in the central North Pacific tend to be of opposite sign to those in the central and eastern tropical Pacific (i.e., negative in El Niño, positive in La Niña: Figure 2), although they may have the same sign in Modoki events. As shown by Namias, −SSTAs in the North Pacific help anchor a strong trough there because of the reduced surface insolation receipt (more cloud cover) and greater wind stress that brings up colder water from depth. A tropospheric ridge tends to be enhanced downstream (i.e., in western North America), both because of vorticity advection and by the warmer water along the coast (weakened California current), which increases the tropospheric thickness. Thus, the extratropical SST-OLR anomalies of the teleconnection are opposite those in the tropics (i.e., −SSTA with −OLR, +SSTA with +OLR). Moreover, the enhanced latitude gradients of SST and air temperature in traditional El Niño events intensify the “thermal wind,” helping produce a stronger STJ that advects moisture into the southwestern, southern, and southeastern United States. Accordingly, more cloudiness and precipitation usually accompany El Niño along the Gulf...
Coast and in southeastern states (Figure 3). When an El Niño is accompanied by $+\text{PNA}$, drier and warmer than normal conditions occur in the northwest United States and southwest Canada, with colder and wetter than normal conditions in the northeast. The atmospheric associations may be intensified in winter by snow cover–circulation positive feedbacks. Over more extensive snow cover, the greater net radiation deficit refrigerates the air relative to a no-snow situation, reducing the tropospheric thickness, and thereby helping “fix” a long-wave trough longitudinally that encourages further snowfall along and near its equatorward margin. Thus, the $+\text{PNA}$ typically is marked by positive snow extent anomalies in the eastern United States, but $-\text{PNA}$ has positive snow extent anomalies in the western United States, helping promote persistence in the two patterns. Somewhat analogously, more extensive snow cover in Eurasia may be involved in the onset and amplification of ENSO.

Certain teleconnection patterns – the PDO in particular – are not yet explainable physically, although the positive (warm) and negative (cool) phases are accompanied by a suite of climate anomalies along the west coast of North America. The PDO probably represents a “slow
teleconnection” with the upper ocean that involves the North Pacific Gyre, somewhat similar to that proposed for the Southern Hemisphere’s Antarctic Circumpolar Wave, although the ACW is a more mobile pattern. Accordingly, the PDO appears to influence the occurrence and character of El Niño and La Niña events. For example, the frequency and intensity of El Niño events increased starting around 1976/1977 with the transition from cool to warm PDO; before that, El Niños and La Niñas occurred with about equal frequency (Figure 1). It is interesting that the late twentieth-century warm PDO was also marked by the longest El Niño on record (1991–1995) and by the very strong 1982/1983 and 1997 events; the last marking a peak in globally averaged temperature the following year and the start of the so-called “climate warming pause.” This most recent period has seen increased oceanic uptake of heat, perhaps due to the stronger surface wind stress promoting greater vertical mixing within the water column that has accompanied more frequent La Niña events, and/or the AMO’s interaction with the global ocean “conveyor belt.” The slowed atmospheric heating may also be the result of increases in atmospheric particulates resulting from the rapid industrialization of China and India.

Climate impacts

Temperature anomalies

Globally, El Niño events tend to accompany positive (i.e., warmer) anomalies of surface temperature while La Niña tends to accompany negative (cooler) anomalies. However, considerable regionality occurs for both phases. For example, the US Gulf Coast and southeastern United States typically experience cooler conditions in El Niño winters owing to greater cloud amount and precipitation (Figure 4a). Conversely, this region typically sees warmer conditions during La Niña events, due to reduced cloudiness (Figure 4b).

Colder winters in eastern North America are more likely with an intensified or larger West Pacific Ocean “warm pool” (i.e., in La Niña or neutral ENSO and/or cool PDO phases) because these favor the +PNA pattern, with the timing of cold-air outbreaks influenced by the MJO (greater convection over the tropical Indian Ocean).

The opposition in sign of SLP anomalies between the Arctic and middle northern latitudes expressing the AO has associated surface temperature anomalies as follows: colder (warmer) in the Arctic (middle latitudes) for the high index phase, and warmer (colder) in the Arctic (middle latitudes) for the low index phase. Accordingly, winter sea ice conditions tend to be milder – less ice – for low index than high index AO, and trends in the AO (e.g., from lower to higher index values between about 1968–1997) explain a proportion of the wintertime trends in surface temperature over the extratropical Northern Hemisphere.

Regional anomalies of temperature are favored by teleconnection modes comprising meridional atmospheric circulation (e.g., +PNA, −NAO/AO). They tend to be expressed subseasonally as increased blocking activity. Blocking refers to the high amplitude wave – and therefore very slow moving, or even retrograde – phase of the intraseasonal “index cycle.” At its most extreme, the meridional phase resolves into an Omega Block comprising a blocking high and positive temperature anomalies in higher latitudes (e.g., the Greenland–Scandinavia region), flanked by “cutoff” lows in lower–middle latitudes (e.g., Southeast United States, western Mediterranean). Because this pattern represents a reversal of the normal latitudinal pressure/height
Figure 4  United States climate division composite surface temperature anomalies (°F) for (a) nine El Niño events and (b) eight La Niña events, with respect to the 1971–2000 climatology. Note: 1.8°F = 1.0°C.  Source: Reproduced from NOAA.
CLIMATIC MODES AND TELECONNECTIONS

and temperature/thickness gradients, the thermal wind also is reversed; hence, the westerlies slow – and may become easterly – in the region of the blocking high, and the Polar Front Jet (PFJ) splits (“split flow”) upstream of the block. An Omega Block is more likely to accompany negative NAO/AO; it results in milder than normal conditions in Labrador–Greenland but colder conditions in western Europe and the eastern United States (e.g., the winter of 2009–2010). The NAO is not restricted to wintertime and can be responsible for extreme climate anomalies in other seasons. For example, the highly anomalous surface melt that occurred over 97% of the Greenland ice sheet in mid-July 2012 was accompanied by a strong blocking ridge, with cutoff lows and negative temperature anomalies over Labrador and the British Isles (Figure 5a), during a −NAO for the month (Figure 5b).

Droughts and floods

Because the ENSO “cycle” involves relocations of the centers of rising and sinking air and their associated Walker circulations, El Niño events favor drying – even drought – accompanying the South Asian and Australian summer monsoons, and in the Brazilian Nordeste, but an increased incidence of flooding rains in Peru, Ecuador, and eastern Africa. The opposite is generally true for these regions during La Niña. In the extratropics during boreal winter, El Niño brings increased precipitation – sometimes flooding – to California, the Gulf Coast, and southeastern United States (Figure 3). The accompanying split-flow regime along the US West Coast can induce atmospheric rivers of extreme moisture (cloud, precipitation) that result in flooding and mudslides in California. In the Northwest US states and southwest Canada, the weakened PFJ and strengthened ridge typical of El Niño reduce precipitation in that region.

Increased precipitation typically characterizes El Niño events in parts of East Asia. ENSO-related precipitation anomalies are sensitive to the strength of the El Niño, the type of event (traditional vs Modoki), as well as to MJO phase. The US Midwest extreme drought of summer 1988 resulted from a tropical–extratropical teleconnection as the El Niño of 1987 transitioned to a La Niña, establishing a strong ridge over central North America that was intensified by land surface–atmosphere interactions (e.g., dry soil, lowered evapotranspiration rates).

Wildfire

Wildland fires occur when an ignition source (e.g., lightning, human activities) causes combustion of dry fuels in grassland or forest eco-types, and windy, low humidity conditions prompt a rapid expansion of the burned area. Because fine fuels – from dry soils and vegetation – are a necessary prerequisite for wildfires, there are strong associations of fire frequency, intensity, and burned area with climatic modes and teleconnections. Thus, El Niño events tend to accompany increased bush-fire activity in eastern and southeastern Australia, and more severe forest fire conditions in Canada’s Yukon, northern Prairie Provinces, and central Quebec. Lag relationships also are evident: in the southwestern United States, the positive anomalies of precipitation associated with El Niño increase vegetation growth that then dries out in the subsequent La Niña or neutral ENSO episode, providing fuel for large fires. A PDO influence is superimposed, whereby wildfire area burn or intensity may be increased in constructive phases (i.e., El Niño during warm PDO, La Niña in cold PDO). Subseasonally, the MJO modulates wildfire variations of the western United States, partly via its influence on dry lightning. Of course, the encroachment of human settlements
Figure 5  (a) Average 500 hPa heights (m) for the four–day period of the Greenland ice sheet major surface melt event in mid-July 2012; (b) monthly average 500 hPa height anomalies (m) for July 2012. Source: Reproduced from NOAA.
into wildland areas increases the risk of fire hazard, as likely also does global warming.

**Tropical cyclones**

Intense rotating synoptic-scale vortices originating in tropical latitudes from strong sea–air interactions (i.e., warm-cored cyclones) also vary in their frequency and tracks of movement due to teleconnections. The primary response is to the SST anomaly pattern, whereby tropical cyclogenesis increases over +SSTA and decreases over −SSTA. The climatic relationship is most evident in association with ENSO: more frequent and/or intense tropical cyclones develop in the tropical eastern Pacific in El Niño than in La Niña, with the reverse being generally so for the western Pacific. During the Australian monsoon (October–April), tropical cyclone activity in the southwest Pacific and eastern tropical Indian Ocean shows a close association with the monsoon trough, which tends to be more (less) active in La Niña (El Niño) years. Subseasonally, however, there is strong modulation by the MJO on both ENSO extreme phases in most tropical cyclone regions. In the Caribbean Sea, Gulf of Mexico, and western Atlantic, the relationship to ENSO is more indirect. There, hurricane activity decreases, on average, in El Niño years relative to La Niña and also neutral years, because the +SSTAs in the eastern tropical Pacific enhance the latitude gradient of temperature and, thereby, a stronger STJ. The associated greater vertical wind shear in El Niño years is unfavorable to intensification of tropical cyclones. Fluctuations in longitude of the North Atlantic High – partly a function of the NAO – influence the steering of hurricanes and tropical cyclones; into the Gulf of Mexico for westward displacement, versus along the US Atlantic coast with eastward displacement. On longer (interdecadal) timescales, increased versus decreased hurricane activity in the North Atlantic shows some association with the AMO.

**Extratropical synoptic systems and extreme weather events**

Because teleconnections represent low-frequency variations of the atmospheric circulation, the associated higher-frequency (synoptic-scale) circulation regimes can be quite different for the extreme phases as well as between events of the same sign (e.g., weak versus moderate versus strong El Niño). Thus, “nor’easter” winter storms bringing heavy snowfall along the US East Coast are more likely during El Niño, +PNA, and −NAO/AO because of the phasing of the STJ and PFJ that accompany a deep trough in those longitudes. Conversely, intense mid-latitude cyclones bringing heavy precipitation to Britain and western Europe are favored by +NAO/AO patterns. Accordingly, decadal and longer-period variations in the prevalence of NAO extreme phases are accompanied by quite different synoptic regimes and their associated regional surface climate anomalies (e.g., the 1960s and 1970s versus the 1980s and 1990s). In the Tasman Sea, atmospheric blocking – more frequent in El Niño than La Niña years – favors winter–time persistent cold-air outbreaks over New Zealand that often contain intense meso-scale cyclones (“polar lows”) bringing snowfall to the South Island.

The frequency and intensity of tornadoes (especially the more violent vortices) in the central and Midwest United States show some association with ENSO; somewhat greater in La Niña than in El Niño events, because of the stronger PFJ – and more cold air aloft – in the former, and the fact that higher SSTs in the Gulf of Mexico are maintained over winter by the lack of cloud and precipitation (cf. El Niño). Recent work suggests that the springtime
transition between a La Niña and El Niño sees strong increases in tornado activity in the central United States. Superimposed on these and other synoptic–teleconnection associations as to the timing of events is the role of MJO.

**Teleconnections and global warming**

Contemporary climate changes, such as the reduced Arctic sea ice extent and thickness, may be influencing teleconnections and the frequency and intensity of synoptic and subsynoptic weather events over middle latitudes, via “Arctic Amplification.” This positive feedback involves greater absorption of solar radiation by the Arctic Ocean in summer (due to the reduced surface albedo) and stronger upward fluxes of latent and sensible heat in winter (more open water or thinner ice), which decrease the latitude–temperature–thickness gradient, slow the zonal winds, and increase the amplitude of Rossby waves and incidence of blocking; all expressed as more frequent – NAO/AO patterns.

Finally, given their important roles in climate prediction and the development of climate change scenarios, it is pertinent to consider how well global climate models (GCMs) capture the circulation variability comprising teleconnections and climatic modes. As might be expected when simulating ENSO, the performance has improved substantially with better coupled (i.e., ocean–atmosphere interaction) models, although there is some dependence on El Niño type – traditional versus Modoki – and the model convective scheme. For teleconnections of extratropical origin, particularly the AO, GCMs generally confirm their intrinsically atmospheric character; that is, coupled models tend not to significantly improve the simulation of these teleconnection patterns relative to atmosphere GCMs.

**SEE ALSO:** Atmospheric/general circulation; Climate change and sea ice; Climate and societal impacts; Climatology; Climatology: history; Global climate change; Global climate models; Hydroclimatology and hydrometeorology; Monsoons; Oceans and climate; Quantitative methodologies; Temperature; Weather, extreme

**References**


**Further reading**


Climatology: history

Cary J. Mock  
University of South Carolina, USA

Early developments

Climatology as an academic and professional discipline has increased in popularity in the last few decades. This is also clearly evident in popular culture such as television, newspapers, and blogs and forums on the Internet. It is in contrast to what Carleton (1999) noted for much of the twentieth century, when climatology was viewed more as an ugly duckling of statistics and applied services. Much of the increased visibility of climatology can be traced in several historical themes that have impacted the development of the discipline. This entry summarizes themes of climate data and technological advances, methods in climatology, interdisciplinary linkages, scale aspects, and societal relevance. It focuses more on the geographic aspects and contributions, but also notes linkages with other disciplines (notably meteorology) and concepts that extend back to before the twentieth century. Important events on geographical climatology are noted in Figure 1.

Daily weather observations, both instrumental and verbal, have provided a framework for climatological research since the ancient Greeks. Very early weather and climate descriptions also include those written by Muslim scholars in the Middle Ages which were used in climate classification. However, instrumental observations did not begin until after the Renaissance, and widespread observational data networks in North America mostly began in the early nineteenth century. Thus, climatological research in the nineteenth century based on these data was mostly basic description, which enabled the construction of the first large-scale climate maps and climate classifications (e.g., by Koeppen). Some climate and weather processes were debated, such as the law of storms and potential relationships between climate and health. During this era, both meteorological and geographical climatology were not official academic disciplines, and thus were not distinguished from one another.

Modern developments to the mid-twentieth century

By the end of the nineteenth century, expanding observational networks of weather data began to be more evident at the global scale, while regional-scale data networks became denser. The first geography (and meteorology) academic departments in the United States were established at the University of Chicago and at the Massachusetts Institute of Technology in 1903 and 1928 respectively. Physical geography and climate studies from the early- to the mid-twentieth century assessed the relationship of the Earth and its inhabitants, probably because of the popularity of environmental determinism at this time, partly through the work of Ellsworth Huntington (Skaggs 2004). This included mostly descriptive climatology in the form of climate as “averaged weather,” as well as some climate classification. From the 1910s to the 1940s, climatology started to move beyond observation
toward the analysis of processes and synthesis, as more detailed meteorological observations became apparent from aircraft and the upper atmosphere. In particular, aviation-related observations increased sharply in the United States during the 1920s and 1930s. Thus research incorporated new meteorological theoretical advances such as the Bjerknes mid-latitude cyclone model and Rossby’s model of atmospheric circulation. During much of the “regional era” of the 1920s–1940s, research advances in the processes in climatology were made mostly in the physical sciences (Skaggs 2004). These advances subsequently led to the development of synoptic and dynamic meteorology by World War II.

Skaggs (2004) regarded the “modern era” of geographical climatology as beginning after 1945. Prior to the modern era, the notion of climate in relation to the environmental determinism paradigm that prevailed in geography earlier in the twentieth century, and to the decline of the regional geography paradigm, had a huge impact on society. The research perspective on climate classification continued, but influences from meteorological advances were increasingly introduced into academic geography. By around 1950, climatology in some university geography departments was taught by meteorologically trained scholars employing concepts derived from meteorological technology and developed prior to and during World War II. During this time, physical geography specialty areas, including climatology, started to subdivide into more specialized fields (definition of fields vary up to today); for example, climatology could incorporate the fields of micro-,
CLIMATOLOGY: HISTORY

regional, and synoptic climatology. The quantitative revolution in geography was apparent during the mid-twentieth century as well, with the increase of the application of statistical and hypothesis testing approaches in physical geography. In separate activities outside American academic geography in the 1950s, applied meteorology and climatology and climate data collection and archiving increased in popularity with the advent of the State Climatologist Program; the establishment of what eventually became the National Climatic Data Center in Asheville, North Carolina; the establishment of the National Hurricane Center at Miami, Florida; and the advent of the era of satellite meteorology.

Developments of recent decades

The 1960s into the 1980s saw further developments and ideas from other fields influencing mainstream geographical climatology, including chaos theory. Data management and monitoring accelerated with the creation of the National Oceanic and Atmospheric Administration (NOAA) in 1970 and the National Aeronautics and Space Administration (NASA) shortly after. The National Climate Program Act in 1977 renewed governmental state climatological public services, and regional climate centers also contributed to data management activities. Geographical climatology continued with its traditional positivist, deductive approach, with different climatological subfields still largely based on observational data, statistics, and hypothesis testing but becoming a bit more diverse in the spatial scale of study. Many of the subfields in climatology (e.g., synoptic climatology) developed within academic geography, aided by the development of faster computation, with applications utilizing objective as well as subjective methods. Some studies adopted new perspectives to examine climate processes (e.g., the air-mass climatology approach). Urban and energy balance climatology, and arctic and alpine climatology became prominent at some academic geography centers (Brazel et al. 1991). The development of Thornthwaite’s concept of potential evapotranspiration merged hydrological data with climatic data and linked the processes between climate and the hydrosphere.

By the mid-1980s geographically trained climatologists were analyzing various components of the climate system in greater detail, and much research was integrating climate processes at various time (from meteorological subdaily to paleoclimatic) and spatial (from microscale to global) scales. At the close of the twentieth century there were still distinctive characteristics of how geographical climatologists conducted research compared to meteorologists; for example, they placed more emphasis on classification and less on forecasting (Carleton 1999). However, given the need for a strong holistic approach in analyzing and linking climate processes at different spatial and temporal scales, geographical climatologists began collaborating and borrowing more ideas from meteorology as well as other Earth, physical, and natural science disciplines. The mapping and analyses of numerous climate variables from NOAA’s 20th Century Reanalysis project have enabled much more detailed studies of climate variability and change. Quiring (2007) noted that a shift occurred around 1998, with climatologists publishing less in geography journals and more in nongeography journals, perhaps for greater interdisciplinary visibility, to reach more of a mainstream climate audience, or as a result of the culture of funding agencies or the growth of more journals. From 2000 onward, while positivism and empirical generalization remained popular in climatology, the development and increased popularity of numerical modeling as a result of faster computers also focused more
CLIMATOLOGY: HISTORY

attention on processes. Up to the present, not many modelers trained in dynamic climatology are in academic geography departments, but some are in geography departments that include atmospheric science programs. However, an increasing number of geographical climatologists are utilizing data from global climate model simulations, including for downscaling studies to link larger-scale processes with smaller-scale ones. Climatologists started to use new innovative applications with climatology from geographical information science and spatial statistics developed in the last few decades. The need to calibrate and verify climate models with observational data continued to be a common theme in geographical climatology, covering a range of time scales from the interannual to the paleoclimatological, and some geographical climatologists continued to analyze the data quality of weather and climate observations.

Study of the human dimension in climatology dates back to the ancient Greeks; thus the idea of humans modifying the climate and the physical environment is not new. The idea of humans causing increased global warming can be traced back, for example, to the work of George Perkins Marsh in the mid-nineteenth century. However, with the advent of the Intergovernmental Panel on Climate Change (IPCC) in the 1980s, climate change and detection studies from the regional to the global scales became increasingly popular and gained in importance, including among geographical climatologists. This activity has led to increased interdisciplinary research between climatologists and social scientists and humanities scholars on climate impacts, adaptation, and policy (Kates 2011). The result has been something of a renewal of environmental determinism concerning the role of climate on premodern societies, since the turn of the century (Diaz and Trouet 2014), prompting more activity in high-resolution paleoclimatology. The increased popularity of extreme climate events, such as years of unprecedented extreme droughts and heat waves, in recent times and potentially in the future, have also led to more case study approaches and interdisciplinary climate research. Geography, with its unique interdisciplinary tradition and linkages to the natural, physical, and social sciences as well as the humanities, will likely continue to play a prominent role in continuing climatology’s increased visibility well into the twenty-first century.

SEE ALSO: Climate and societal impacts; Climatology; Global climate change; Global climate models

References


Further reading


Climatology

Julie A. Winkler
Michigan State University, USA

Climatology is a key component of contemporary physical geography. In this contribution to The International Encyclopedia of Geography, the evolution of climatology as a field of inquiry is summarized. In addition, the data sources available to climatologists, along with their limitations, are described, and the wide range of statistical methods and numerical models used in climatological research is highlighted. The various subfields of climatology are grouped into two categories, those that represent an analytical perspective and those that are thematic in focus, and are briefly described. Particular attention is paid to recent developments in applied climatology. Finally, the growing popularity of the terminology “climate science” and the consequent implications for the meaning of “climatology” are noted.

Evolution of climatology as a field of inquiry

Climatology is often described as either “the study of climate” or “the science of climate.” These broad definitions, however, mask the evolution through time in the perceptions of what constitutes “climate,” and the impact of these changes on the foci, objectives, and approaches of climatology. Danish science historian Matthias Heymann argues that climate “cannot be reduced to a clear, neat, and unequivocal definition;” but, rather, “its interpretation is closely linked not only to a state of scientific knowledge, but also to the broader cultural contexts of its time” (Heymann 2010, 82). Changes in the perception and interpretation of climate have not only influenced the nature of climatology as a field of inquiry, but also the positioning of climatology among, and within, other disciplines, including geography.

Climatology before 1900

The term climate originates from klima, a Greek astronomical term for the length of the longest day at different latitudes. In the third century BCE, Greek philosopher Eratosthenes described five latitudinal zones (two frigid zones, two temperate zones, and a tropical zone), distinguished by similar sun angles and day lengths, which he referred to as climates. Early climatological thought was strongly influenced by the philosophy of weather and climate espoused in Meteorologica, authored by Aristotle around 340 BCE. Another influential treatise was On Airs, Waters and Places, published around 400 BCE by Hippocrates, who attributed common diseases to locations, seasons, winds, and air.

The scientific era of climatology begins in the 1600s, when, as succinctly described by American mathematician H. Howard Frisinger, the study of climate moved “beyond the natural philosopher into the hands of the natural scientist” (Frisinger 1966, 444). Scientific breakthroughs that furthered an understanding of the Earth’s climate included French mathematician Blaise Pascal’s observations in the seventeenth century of the relationship between atmospheric pressure and elevation, and the
development in the mid-eighteenth century of a conceptual model of the general circulation by British amateur meteorologist George Hadley. The invention, beginning in the late sixteenth century, of key meteorological instruments (e.g., thermometer, barometer, hygrometer, anemometer) led to systematic measurements of the atmosphere and also to the sharing of observations. Observational networks date to the mid-1600s and were commonplace by the late 1800s. According to geographer David Miller, these measurements provided climatology with numerical data for analysis before most other environmental sciences, and the term climatology became synonymous with the “ordering of any variable displaying an annual regime” (Miller 2005, 284). Science historian Heymann (2010) argues that with these new measurements, the intellectual focus of climatology shifted to climate’s changeability (i.e., the temporal dimension of climate), with less emphasis on the spatial (i.e., geographical) dimension. Notwithstanding the strides made in providing a foundation for the physical explanation of climate, climate determinism also gained popularity during this period. The underlying belief of climate determinism is that the physical environment, particularly the climate, has a controlling influence on the development of human societies and cultures.

Substantial shifts in climatological thought occurred during the nineteenth century. Alexander Humboldt is often credited for refocusing climatology on the spatial, rather than temporal, variations in climate and, early in the 1800s, Humboldt prepared the first isoline map of the spatial distribution of annual temperature. The introduction in the late 1800s of long-term averages as a reference for comparing typical conditions between locations is attributed to Austrian meteorologist Julius von Hann. Climate classification also was used to describe the spatial variations of climate, and the well-known classification by German geographer Waldimir Köppen was first published in 1884. In the nineteenth century, climatology was primarily a subdiscipline of geography and meteorology and, to a lesser extent, of geology. Climate determinism remained popular among some groups throughout the nineteenth century.

The modern era of climatology

The modern era of climatology begins in the late nineteenth and early twentieth centuries. Early in this period, Norwegian meteorologist Vilhelm Bjerknes outlined the primitive equations (i.e., equations for the conservation of mass, momentum, and energy) for predicting large-scale atmospheric motion, and British mathematician Lewis Fry Richardson provided initial numerical techniques for solving these equations. Bjerknes and his colleagues, including his son Jacob Bjerkness and Halvor Solberg, also were responsible for the first conceptual models of extratropical cyclones, air masses, and frontal systems, providing a platform for relating the climate at a location to the frequency of different types of weather systems. Atmospheric teleconnections, defined as the relationships between distant atmospheric and sea-surface temperature anomalies, were first introduced by British physicist Gilbert Walker in the late nineteenth century, and by the second half of the twentieth century were widely used to explain interannual climate variability.

Energy and water fluxes between the atmosphere and Earth surface were another focus of twentieth century climatological research. The initial conceptualization of the energy budget is attributed to the work of Swedish physicist Anders Ångström in the 1920s, although numerical models of the energy balance were not developed until the late 1960s. In 1955,
geographers Charles Thornthwaite and John Mather published their seminal monograph *The Water Balance*, outlining methods to estimate water surpluses, runoff, and recharge. Improvements in observations also contributed to advances in climatology. Regular balloon soundings of the upper atmosphere began in the 1930s and, from the 1970s onward, climatology benefited from the greater spatial and temporal coverage of satellite observations of the atmosphere and land surface.

In the second half of the twentieth century, concerns about anthropogenic contributions to climate change spawned an enhanced interest in climatology. In the 1850s British physicist John Tyndall discovered that atmospheric molecules such as carbon dioxide and water vapor absorbed thermal energy, and in the first half of the twentieth century global temperature was theoretically linked with changing carbon dioxide concentrations by Swedish physicist Svante Arrhenius and British engineer Guy Callendar. Yet, concern regarding the impact on climate of increasing greenhouse gases did not become widespread until after Carl Keeling initiated in 1958 systematic measurements of atmospheric carbon dioxide at the Mauna Loa Observatory in Hawai’i. Soon after, anthropogenic-induced land cover change was recognized as an additional forcing of local, regional, and global climate. With this heightened interest in the anthropogenic influences on climate came a greater interest in the temporal, rather than spatial, variations of climate. Although initial attempts to estimate trends in global surface temperature date to the early work by Köppen in 1881, published time series of globally averaged temperatures were not available until the 1970s and 1980s. These include, for example, the temperature time series produced by scientists at the University of East Anglia Climatic Research Unit in the United Kingdom, and at the US National Aeronautics and Space Administration Goddard Institute for Space Studies.

Climate models became essential tools for studying climate. Meteorologist Norman Phillips is credited with developing, in 1956, the first general circulation model of the atmosphere, and in the late 1960s and early 1970s climatologist Syukuro Manabe and oceanographer Kirk Bryan constructed and refined a coupled atmosphere and ocean model. By the early twenty-first century, global models were widely used to project future climate conditions, and archives of model simulations were established. Interest in paleoclimatic approaches to studying climate also increased during the modern era, in line with an evolving focus of climatology on the temporal variations of climate. As early as 1837, Swiss geologist Louis Agassiz proposed that the Earth had experienced past ice ages, and in 1941 Serbian astronomer Milutin Milankovitch related long-term climate change to cyclic variations in Earth motion (orbital eccentricity, obliquity, and precession). Beginning in the 1800s, proxy measures of climates, such as the growth bands in trees and deposits in lake beds, were recognized as a means to extend the climatological record backwards in time. Ice cores were first drilled in the 1950s, and, around the same time, analyses of deep-water corals suggested variations of ocean temperatures over glacial time scales.

**Philosophical paradigms of climatology**

The philosophical paradigms of climatology, particularly as practiced within geography, have shifted considerably over time. In the early twentieth century, climatology was heavily influenced by environmental (climatic) determinism, which at that time was a central theory of the discipline of geography. After environmental determinism fell out of favor in the 1920s, climatology as practiced within geography returned to a focus
on the spatial differences in climate that was popular during the previous century, whereas the focus of climatology practiced within the discipline of meteorology was on the time-averaged components, or synthesis, of weather. From the mid-twentieth century until recently, philosophical paradigms were not extensively debated by climatologists, whatever the discipline with which they most closely identified. For the most part, climatologists comfortably worked within a positivist framework and employed the scientific method. American geographer Richard Skaggs contends that the lack of a widely accepted orthodoxy during this period, and the diversity of subject matter and methods, contributed to the rapid advancement of climatology from the 1950s onward (Skaggs 2004). But beginning with the early twenty-first century, climatologists became more cognizant of the paradigms that guide their work. Climatologists Gavin Schmidt and Steven Sherwood, for example, recently commented on the influence of climate models on climatological thought, pointing out that these models have led to the dominance of “the paradigm of understanding emergent properties of the complex system via the bottom-up agglomeration and interaction of small scale processes” (Schmidt and Sherwood 2015, 165). Interest in the coupling of climate and human systems also has grown and British geographer Mike Hulme has argued that geography should “reclaim climate from the natural sciences,” where it is defined in physical terms only, and, instead, “treat it unambiguously as a manifestation of both Nature and Culture” (Hulme 2008, 6).

Data sources for climatological analyses

A large number of climate variables are used to describe the many facets of climate. Observations of these variables are essential for analyzing the spatial and temporal variability of climate. Climate variables can be measured directly or they can be estimated using remote sensing techniques. Worldwide routine in situ measurements include surface air pressure, maximum and minimum temperature, precipitation, humidity, wind direction and speed, and visibility. A global network of balloon soundings provides upper-air measurements of pressure, temperature, humidity, and wind speed and direction for multiple levels in the upper atmosphere. Satellite observations of the atmosphere began nearly fifty years ago. Observations are obtained from both Earth-orbiting satellites and geostationary satellites. Retrieved variables include, among others, surface land temperature, sea surface temperature, soil moisture, snow and ice cover, and precipitation estimates.

Although climatologists often collect atmospheric observations as part of field campaigns and maintain specialized observational networks for specific applications, the extensive use of historical climate data archives is a hallmark of climatology. Most often, these archives are maintained by national organizations, such as the US National Oceanic and Atmospheric Administration’s National Centers for Environmental Information (formerly the National Climatic Data Center). Working with these archived observations can be challenging. Most observational networks were initially designed for short-range weather prediction, where the focus is on the accuracy and precision of measurements. While these qualities are also a consideration for climate monitoring, the temporal and spatial consistency of observations is as, or even more, important. Heterogeneities introduced into the historical climate record by changes in instrumentation, observation protocols such as time of observation, and station location complicate the use and interpretation of climate observations, particularly the
interpretation of temporal trends. Consequently, identifying and adjusting for biases in time series of climate observations has been, and continues to be, an important aspect of climatology.

In addition to archives of “raw” observations, gridded fields of historical climate observations also are available. Gridded fields can be obtained by: (i) averaging anomalies or climatological values for stations located within a specified grid box (e.g., the Global Historical Climatology Network (GHCN) global gridded temperature and precipitation products); (ii) spatial interpolation of anomaly and climatological fields based on distance between observing stations, such as the global gridded monthly time series and climatological values of terrestrial air temperature and precipitation developed by the University of Delaware; or (iii) spatial interpolation that, in addition to distance, considers elevation (e.g., the WorldClim and Daymet datasets) or elevation and slope (e.g., the PRISM (parameter-elevation relationships on independent slopes model) gridded fields of temperature and precipitation). As noted by PRISM developer Christopher Daly, an important consideration when using gridded observations is that the fine resolution of many of these datasets can give an appearance of realism that is often not consistent with the spatial resolution of the initial observations (Daly 2006).

Reanalysis fields, first introduced in the late 1990s, are another popular data archive for climate research. These gridded fields are a “blend” of observations and model output. Very generally, for a particular time, the value of a parameter at a location is initially obtained from the short-term forecast of an operational weather forecast model and then modified by the surrounding current observations of the atmosphere. The spatial coverage of reanalyses ranges from regional (e.g., the North American regional reanalysis) to global (e.g., ERA-40 from the European Centre of Medium Range Forecasting), and the horizontal resolution varies from fine (e.g., about 38 km for the Climate Forecast System reanalysis) to coarse (e.g., 2.5° latitude × 2.5° longitude for the NCEP/NCAR reanalysis). Reanalysis fields are affected by biases associated with the operational weather forecast model and from the number and type of available climate observations; thus, they can deviate from observations especially in data-poor regions.

Archives of simulations from global and regional climate models are another important resource for climatological analysis. These simulations are useful for numerous applications, particularly for assessments of climate change impacts and adaptation. Simulations of historical and future climates obtained from global climate models developed by more than twenty modeling groups from different countries are available from the Coupled Model Intercomparison Project, phase 5 (CMIP5). Archives of finer-resolution simulations from regional climate models are also available. For example, the ENSEMBLES project distributes climate simulations for Europe for the period 1950–2100, that were obtained from regional climate models driven by reanalysis fields and multiple global climate models. For North America, multiple regional climate model simulations for two time slices (1960–1990 and 2040–2070) are available as part of the North American Regional Climate Change Assessment Program (NARCCAP).

Research methods in climatology

Climatologists employ a plethora of research methods, ranging from the simple to the complex, that are constrained only by the researcher’s innovation. Statistical techniques and numerical modeling are, in particular, widely employed in climatological research and applications.
Statistical techniques

Statistical techniques are used for the extraction, quality control, and synthesis of climate information, and for establishing associations between variables. The calculation of long-term averages and the use of composite mapping are some of the first, albeit relatively simple, usages of statistical methods in climatology. For example, climate standard normals of temperature and precipitation, introduced by the World Meteorological Organization in the 1930s, are defined as 30-year averages, updated every decade, for individual stations. Composite maps, on the other hand, display the ensemble average of spatial fields of climate variables, and have frequently been used to identify the typical circulation patterns associated with a particular weather phenomenon, such as severe weather, or the typical temperature and precipitation patterns under different atmospheric teleconnections. Other early introductions of statistical methods to climatology include the use of parametric probability distributions to estimate the magnitude and return frequency of climate extremes. A classic example, dating to the 1960s, is the *Rainfall Frequency Atlas of the United States* authored by hydrologist David Hershfield, which provided extreme precipitation estimates for hydrologic design. Time series analysis also has been a cornerstone of climatological analysis. Linear and nonlinear regression methods have been, and continue to be, used to estimate temporal trends in climate variables, and techniques such as harmonic, spectral, and wavelet analysis are helpful for detecting interannual and interdecadal climate variability.

Climate classification also employs statistical techniques. The goal of any classification is to minimize intraclass variation while maximizing interclass variation. Classification has been used within climatology to summarize the spatial variations of climate and to identify frequently-occurring atmospheric circulation patterns. Early classifications were developed subjectively, such as the aforementioned Köppen classification scheme, which grouped locations based on average temperature and the amount and seasonality of precipitation, and the well-known catalog of circulation types for the British Isles developed by climatologist Hubert Lamb in the early 1970s. Over the past several decades, subjective classification methods have, for the most part, been replaced by computer-assisted approaches based on multivariate statistics. In particular, cluster analysis is frequently used to identify climate regions. Statistical approaches for identifying circulation types include correlation analysis, empirical orthogonal functions (i.e., principal components analysis), and self-organizing maps. A classic example is the use of empirical orthogonal functions by meteorologists John Wallace and David Gutzler in the early 1980s to identify wintertime teleconnection patterns for North America.

Statistical techniques are also used in climatology to downscale simulations from global and regional climate models to finer spatial and temporal resolutions, and to adjust for biases in the model simulations. An extensive literature exists that evaluates the efficacy of a range of techniques for this purpose, including regression procedures, canonical correlation analysis, artificial neural networks, support vector machine algorithms, and weather generators (Winkler et al. 2011 provides a review). The resulting empirically-downscaled and bias-corrected climate projections are frequently used in climate change assessments. Other uses of statistical techniques within climatology include the development of transfer functions to estimate climate parameters from proxy records of climate-dependent phenomena, such as the

---

**Climatology**

Statistical techniques are used for the extraction, quality control, and synthesis of climate information, and for establishing associations between variables. The calculation of long-term averages and the use of composite mapping are some of the first, albeit relatively simple, usages of statistical methods in climatology. For example, climate standard normals of temperature and precipitation, introduced by the World Meteorological Organization in the 1930s, are defined as 30-year averages, updated every decade, for individual stations. Composite maps, on the other hand, display the ensemble average of spatial fields of climate variables, and have frequently been used to identify the typical circulation patterns associated with a particular weather phenomenon, such as severe weather, or the typical temperature and precipitation patterns under different atmospheric teleconnections. Other early introductions of statistical methods to climatology include the use of parametric probability distributions to estimate the magnitude and return frequency of climate extremes. A classic example, dating to the 1960s, is the *Rainfall Frequency Atlas of the United States* authored by hydrologist David Hershfield, which provided extreme precipitation estimates for hydrologic design. Time series analysis also has been a cornerstone of climatological analysis. Linear and nonlinear regression methods have been, and continue to be, used to estimate temporal trends in climate variables, and techniques such as harmonic, spectral, and wavelet analysis are helpful for detecting interannual and interdecadal climate variability.

Climate classification also employs statistical techniques. The goal of any classification is to minimize intraclass variation while maximizing interclass variation. Classification has been used within climatology to summarize the spatial variations of climate and to identify frequently-occurring atmospheric circulation patterns. Early classifications were developed subjectively, such as the aforementioned Köppen classification scheme, which grouped locations based on average temperature and the amount and seasonality of precipitation, and the well-known catalog of circulation types for the British Isles developed by climatologist Hubert Lamb in the early 1970s. Over the past several decades, subjective classification methods have, for the most part, been replaced by computer-assisted approaches based on multivariate statistics. In particular, cluster analysis is frequently used to identify climate regions. Statistical approaches for identifying circulation types include correlation analysis, empirical orthogonal functions (i.e., principal components analysis), and self-organizing maps. A classic example is the use of empirical orthogonal functions by meteorologists John Wallace and David Gutzler in the early 1980s to identify wintertime teleconnection patterns for North America.

Statistical techniques are also used in climatology to downscale simulations from global and regional climate models to finer spatial and temporal resolutions, and to adjust for biases in the model simulations. An extensive literature exists that evaluates the efficacy of a range of techniques for this purpose, including regression procedures, canonical correlation analysis, artificial neural networks, support vector machine algorithms, and weather generators (Winkler et al. 2011 provides a review). The resulting empirically-downscaled and bias-corrected climate projections are frequently used in climate change assessments. Other uses of statistical techniques within climatology include the development of transfer functions to estimate climate parameters from proxy records of climate-dependent phenomena, such as the
growth rings of a tree, quality control of observational data including tests for heterogeneities, and the spatial and temporal interpolation of atmospheric observations.

**Numerical modeling**

Numerical models are physically-based models developed using the principles of conservation, the first law of thermodynamics, and the laws of motion. Models are an important tool for improving the scientific understanding of the processes and internal dynamics of the climate system, evaluating responses of the system to perturbations, projecting future climate conditions, understanding the processes contributing to paleoclimates, and translating climate processes into useful information for applications and policy. Climatologists employ a variety of numerical models that differ in terms of their dimensionality and resolution, and by the number and types of the components of the climate system that are included in the model. The choice of model depends on the questions to be addressed or the applications for which the simulations are intended.

Examples of 1-D models include energy balance models, which are typically used to investigate latitudinal variations in surface temperature as a function of the Earth's energy balance, and radiative-convective models, which are single column models focusing on vertical variations but ignoring horizontal variations. Early 3-D climate models were atmosphere-only general circulation models. Although a coupled atmosphere–ocean model was first introduced in the 1960s, it was not until the late 1980s that atmosphere–ocean general circulation models (AOGCMs) were widely used. AOGCMs simulate the dynamics of the climate system, rather than only the atmosphere, and incorporate feedbacks between the atmosphere, ocean, and land and sea ice, allowing for more realistic modeling of interannual and longer-term variability of the coupled system. Earth system models (ESMs) expand on AOGCMs to include additional components of the climate system (e.g., the carbon cycle) and represent the current state-of-the-art in 3-D climate modeling. ESM simulations are routinely run for periods of a century or longer. In addition to ESMs, Earth system models of intermediate complexity (EMICs), are available. These models have more idealized representations of the climate system components, and are particularly useful for long (i.e., millennial) model integrations. As climate models became more complex, the acronym “GCM” evolved from its original meaning of “general circulation model” to the more general term “global climate model” that broadly refers to 3-D models of the global climate system. In addition to GCMs, limited-area 3-D climate models are also available. Referred to as regional climate models (RCMs), as they only examine a portion of the Earth's surface, RCMs typically have a higher spatial resolution than GCMs and can simulate sub-GCM-scale processes and distributions of climate variables. Lateral boundary conditions to drive a RCM are obtained from reanalysis fields or from GCM simulations.

The immense effort expended by climatologists to develop, evaluate, and apply climate models cannot be overstated. Through the efforts of climatologists and others, these models continue to increase in skill and scope. Furthermore, extensive effort has gone into producing the inputs needed for the initial and boundary conditions of climate models and the long-term, spatially-extensive datasets used for diagnostic evaluations. Maintenance of archives of multimodel simulations is another hugely cost and labor intensive effort, but archives such as
CLIMATOLOGY

CMIP5 are essential for assessing model differences and uncertainty. In addition, climatologists have written extensively on the challenges of interpreting multimodel ensembles, which in effect are “ensembles of opportunity” with interdependent rather than independent members (Knutti 2010). Much effort also has been expended in developing procedures for downscaling climate simulations to scales appropriate for applications and for accounting for biases in simulations. A glimpse of the potential use of model simulations for climate assessments and decision-making was provided by geographer Linda Mearns and her colleagues who summarized the diverse applications of the RCM simulations from the NARCCAP archive (Mearns, Lettenmaier, and McGinnis 2015). Climate models are also a primary tool for evaluating the influence of land use/land cover change on local, regional and global climate. Thus, it is not difficult to understand why Schmidt and Sherwood (2015) refer to numerical simulation as the “new pillar of inquiry” in climatology.

Subfields of climatology

Climatology is routinely divided into subfields. Many different subfields have been proposed, representing differences in scale (e.g., microclimatology), techniques (e.g., statistical climatology), and thematic focus (e.g., building climatology), although there is usually considerable overlap between an individual subfield and other subfields. The subfields also differ considerably in the number of adherents, and the popularity of different subfields has waxed and waned with time. For this entry, subfields within climatology are grouped by whether they represent a different perspective on climatological research or whether they are primarily distinguished by a thematic focus.

Subfields as analytical perspectives

Several subfields of climatology reflect different perspectives to climatological inquiry; namely, physical climatology, dynamic climatology, synoptic climatology, and paleoclimatology. Physical climatology is primarily concerned with the interactions between the Earth’s surface and the atmosphere, including the spatial and temporal distributions of energy, moisture, and momentum exchanges. One focus area of current significance is the role of surface cover types on climate. Physical climatologists are interested in energy and mass exchanges from the micro to the global scales. Physical geography is characterized by the extensive use of numerical models, ranging from simple to complex, although other analytical procedures including statistical techniques are also employed.

Dynamic climatology and synoptic climatology are closely related. In general, dynamic climatology refers to the climatology of atmospheric dynamics and thermodynamics. Dynamic climatologists often, but not exclusively, focus on atmospheric circulation at the global scale and are concerned with climate variability at longer (i.e., interannual) time scales. Synoptic climatology, on the other hand, is concerned with the relationship between the atmospheric circulation and local or regional climate. Synoptic climatology takes the viewpoint that climates differ because their component weather types and the frequencies of these weather types differ, which in turn are affected by atmospheric circulation at a range of scales from planetary to regional. Historically, dynamic climatology has emphasized the use of numerical modeling, whereas empirical methods, particularly classification methods, were widely used in synoptic climatology. This distinction has blurred, however, with statistical techniques (e.g., empirical orthogonal functions) frequently used in dynamic climatology and with numerical models (e.g.,
GCMs, RCMs) now an essential tool in synoptic climatology.

The historical perspective on climatology is provided by paleoclimatology. Here the focus is on reconstructing past climates (usually before the start of the instrumental record) and linking these proxy records to atmospheric circulation. Climate reconstructions make use of pollen analysis, tree-ring dendrochronologies, lake and river sediments, stratigraphic variations in the chemical and dust content of cores obtained from ice caps, and other proxy measures. Statistical techniques are used to develop the transfer functions between the proxy measures and other time series of climate variables (e.g., temperature and precipitation), and climate models are used to simulate past climates in order to understand the mechanisms for past climate changes and to extrapolate between the local and global scales of paleoclimates.

Thematic subfields

Innumerable thematic subfields of climatology can be defined. Of these, urban climatology, hydroclimatology, bioclimatology, and agricultural climatology are particularly active areas of climatological inquiry.

Urban climatology originated in the late 1800s and initially was descriptive in nature, primarily focusing on observational studies and the characterization of the urban heat island (i.e., the elevated temperatures of urban areas compared to surrounding rural environments). With the development of numerical models, the focus of urban climatology shifted to energy and momentum exchanges in the urban environment at all spatial scales, although micro- and mesoscale environments such as urban canyons have been of particular interest. Other areas of active research in urban climatology include air quality, urban forestry, biogeochemical cycles, water movement and storage, and the potential impacts of climate change on the urban environment.

Hydroclimatology, often described as the study of the influence of climate on hydrologic events, emphasizes the interface between the atmosphere and terrestrial water. All aspects of the hydrologic cycle, including precipitation variability, floods and droughts, snowfall and snow cover, river discharge, and groundwater recharge, are investigated, using a range of methods from descriptive summaries to numerical simulations.

Bioclimatology focuses on the interactions between the atmosphere and living organisms. Interest in the influence of climate on biota dates to Hippocrates; thus, bioclimatology is one of the oldest subfields of climatology. Bioclimatology is a diverse subfield, with research ranging from changes in plant development, including trends and interannual variability in the timing of the spring emergence of plants, to the influence of extreme heat events on human mortality. The analytical approaches used in bioclimatology are as diverse as the topics studied. Although agricultural climatology can be considered a part of bioclimatology, it is often singled out as a separate subfield because of the extensive work in this area and its economic significance. Research within agricultural climatology ranges from climate-crop interactions at the field and subfield scales to climate influences on global food security.

The “new” applied climatology

Applied climatology is often considered to be a subfield of climatology. In his classic textbook General Climatology, which was published in 1960, geographer Howard Critchfield claimed that applied climatology is one of three major subdivisions of climatology, along with physical climatology and regional climatology. Since
then, few textbooks and other references have been published without specific mention of applied climatology. However, this distinction as a subfield masks that applied climatology, or at least modern-day applied climatology, draws on, and incorporates, any of the analytical perspectives of climatology and, moreover, can be situated in any of the thematic subfields of climatology. In other words, applied climatology crosses all the subfields (analytical and thematic) of climatology and even extends into the social and policy sciences.

British geographers John Thornes and Samuel Randalls argue that, in spite of considerable interest in applied climatology in the mid-twentieth century, and even calls by prominent climatologists such as Werner Terjung and Stanley Changnon for climatology to address real-world problems, it was not until the late 1990s, with rising concern about climate change, that applied climatology became an interdisciplinary endeavor involving both natural and social scientists (Thornes and Randalls 2014). Prior to this, climate had primarily been treated as a hazard or risk, and applied climatology involved mostly natural scientists. Geographer Marc Tadaki and his colleagues explain that the aim of applied climatology in the 1960s and 1970s was “not to explain social life, but the far more pragmatic one of optimizing economic returns and biophysical indicators” (Tadaki, Salmond, and Le Heron 2014, 397). They contrast this with the current status of applied climatology, arguing that “climate applications have never been more embedded into human organizations than at present” (Tadaki, Salmond, and Le Heron 2014, 399).

Much of the early work in applied climatology could also be thought of as applied in potential but not necessarily in practice, with the choice of problem often selected by the climatologist and findings simply handed off to potential users. Recognition of the limitations of this top-down approach is leading to greater calls for the coproduction of knowledge by climatologists, stakeholders, and representatives from other relevant disciplines, who work together to outline concerns and goals, jointly identify options for mitigation and adaptation, and then evaluate the potential co-benefits and negative externalities of different options. This strategy is much more time consuming than top-down approaches but offers an exciting new direction for applied climatology and increased potential for more informed decision-making.

The future of climatology

Over the last few decades, the study of climate has expanded dramatically. Geographer Andrew Carleton contends that “climatology has become a subject of active research in disciplines that formerly either eschewed it (e.g., meteorology and atmospheric sciences) or had little need for such studies (e.g., geology)” (Carleton 1999, 714). Moreover, climatologist Hervé Le Treut and his colleagues note that, in addition to greater interest in climatology, the scope of climatology also has expanded and is “now far more wide-ranging and physically comprehensive than was the case only a few decades ago” (Le Treut et al. 2007, 98).

Much of this expansion can be attributed to increased interest in, and concern with, natural and anthropogenic climate change. Environmental scientists Michael Grieneisen and Minghua Zhang recently queried the Web of Science database to identify publications related to climate change during the period 1997–2009 and found that the number of publications had increased exponentially with over 100,000 publications reporting climate change-related research during this period (Grieneisen
They point out that this total is equivalent to the research output of major scientific disciplines. Climate is now studied across a range of disciplines, although Grieneisen and Zhang found that, with the exception of economics, the social sciences remain underrepresented compared to the natural sciences. Furthermore, Andy Reisinger argues that, in spite of cross-disciplinary interest, climate-related research has been slow to integrate the insights and contributions of different disciplines within an individual research program but, instead, remains primarily “multidisciplinary” rather than fully “interdisciplinary” in character (Reisenger 2011).

Even though the future for the study of climate is bright, the future of the term “climatology” is in question. More and more, the study of climate is referred to as “climate science” rather than “climatology,” with the meaning of “climatology” instead constrained to the description, often using statistical methods, of the spatial and temporal characteristics of climate. This change mirrors nomenclature changes occurring in meteorology, where “meteorology” is being replaced as the umbrella term for the study of the atmosphere with “atmospheric science.” Propo-

nents of this nomenclature change argue that the “climate science” terminology is more encompassing and better portrays the increased use of numerical models, more frequent multidisciplinary participation, and greater recognition of the human dimension in the study of climate. Further evolution in the meaning and usage of “climate science” and “climatology” can be expected in the future.

SEE ALSO: Atmospheric/general circulation; Climate change, concept of; Climate and societal impacts; Climatology: history; Dendroclimatology; Earth system science; Earth’s energy balance; Global climate change; Global climate models; Hydroclimatology and hydrometeorology; Paleoclimatology; Temperature; Urban climatology; Water budget

References


Further reading


Cloud computing

Ramanathan Sugumaran
John Deere Corporation, USA
University of Iowa, USA
Marc P. Armstrong
University of Iowa, USA

Background and rationale

Geographic information continues to proliferate in both quantity and type, thus presenting significant impediments to its effective management and analysis. Cloud computing is being used to meet such challenges by providing access to capacious storage and high-performance computing resources. In its most general form, cloud computing is an evolving construct that refers to the provision of distributed, configurable computer services, such as applications, compute cycles, and storage, over a network. A key advantage of the cloud model is its inherent flexibility, or what is sometimes termed “elasticity”; users and organizations are easily able to adapt to rapid and uncertain changes in demand, thereby reducing the capital expenditures that are required if infrequent demand peaks must be satisfied. Commonly used sobriquets of this adaptability feature are “pay-as-you-go,” “on-demand,” and “utility” computing. This represents a rather dramatic organizational shift, from an approach involving substantial capital expenditures, which have often taken the form of single-user, powerful personal computers individually configured with software (requiring routine upgrades), to one that incurs operational expenses based on lightweight, networked clients (Garrison, Kim, and Wakefield 2012). Other benefits of cloud computing accrue as a consequence of increased ease and reliability of software and hardware maintenance, thus reducing support costs and risks to organizations.

Cloud services are provided using three primary deployment modes that vary in terms of their accessibility and levels of security. Major commercial vendors, such as Amazon and Microsoft, currently provide open-access public cloud services. In other cases, institutions may opt to develop private cloud resources to support their particular computing objectives. Such services are typically developed by large organizations and are hidden behind a firewall. In yet other cases, hybrid services are provided. For example, public services, such as geographic information system (GIS) applications, may be used with proprietary data held in a secure repository.

Cloud services are often viewed as a set of layers in a stack. The layer closest to the end user is often termed “software as a service” (SaaS) and is generally manifested as managed, network-enabled applications. A good example is Google Apps, and GIS software and services vendors are also rapidly embracing this model. Beneath that layer the “platform as a service” (PaaS) refers to configurable foundational software components such as databases and the middleware that handles flows of information among applications. At the lowest level, “infrastructure as a service” (IaaS) provides the physical
CLOUD COMPUTING

computing resources that are configured by the user to meet variable needs. This layer is often viewed as being centralized, but in reality may be highly distributed as location independence is an important objective of the cloud model. Authoritative discussions about cloud computing concepts, including definitions, architectural issues, and service models, are provided by Armbrust et al. (2010), Garrison, Kim, and Wakefield (2012), and Liu et al. (2011). Research articles focused specifically on emerging technical issues related to cloud computing are published in IEEE Transactions on Cloud Computing.

Geospatial applications

Many geospatial computing tasks are data and computation intensive (Armstrong 2000). Cloud environments are an appealing way to achieve high performance in the form of scalable, massive parallelism, at low cost. One major cloud services vendor, for example, enables buyers to purchase cycles in compute-optimized, memory-optimized, and graphics processing unit (GPU) (massive parallelism) instances. Though researchers are in the early stages of reporting on advances in the use of clouds for geospatial applications, the results show promise (Yang, Xu, and Nebert 2013).

Spatial data reduction and transformation

Like most sensor technologies, LiDAR instruments continue to increase in spatial, spectral, and temporal resolution, thus driving up data volumes (in this case, point clouds). Hegeman et al. (2014) demonstrate how a cloud environment (Amazon EC2) is used to extract topographic information from raw point clouds. Tang and Feng (2014) develop a private cloud and exploit GPU-based parallelism to process LiDAR point clouds; their implementation scales well across the limits of their evaluation process.

Remote sensing

Remote sensing requires significant computer power to process high-resolution multispectral imagery, though such processing is rarely time-critical. Nevertheless, several researchers have developed cloud-based strategies for unsupervised image classification. Sugumaran et al. (2015) provide an overview of remote sensing applications in cloud computing environments. Wang et al. (2013) implemented a maximum likelihood classifier that uses MPI (message passing interface) and MapReduce. When compared to a single node, their approach boosts performance considerably.

These two domains represent only a partial perspective on possible geospatial applications. Other areas that can benefit from low-cost on-demand parallelism include spatial optimization and various forms of spatial statistical analysis, particularly those employing Bayesian methods.

Current cloud computing challenges

Though the advantages of cloud computing are widely touted, it is not a panacea. In some cases, particularly those involving invariant demand for resources, seemingly attractive economic advantages fail to materialize. General cloud computing challenges are emphasized in Yang et al. (2011). There are two common problems with important geospatial implications worth mentioning here.

Network latency

Though computer network bandwidth continues to increase in both wired and wireless
forms, latency can occur if significant traffic volumes are encountered, or when applications are data-intensive, requiring substantial input/output. In such cases, preservation of data locality may become important.

Security

Some applications, health applications of GIS being a prominent example, require fastidious attention to security details, as any lapse can be catastrophic. Cloud computing, by definition, requires storage off-site (and therefore out of local control), with data often being transported across commercial networks. As a consequence of this openness, institutional data policies may prohibit the use of cloud services. Such concerns might be misplaced, however, since the security of many enterprise systems, including those located on university campuses, is highly variable, as can be seen in the accounts of data breaches that appear regularly in the press.

Summary

Cloud computing is emerging as an important computing model that is supported by rapid advances in network infrastructure and bandwidth, as well as innovative software architectures. The approach provides flexible, configurable computer resources to end users. GIS applications are increasingly taking advantage of the opportunities afforded by cloud environments and it is likely that the approach will become commonplace over the course of the next decade.

SEE ALSO: CyberGIS; Geocomputation; Service-oriented architecture

References


Further reading


Clouds

Lin H. Chambers
NASA Langley Research Center, USA

Clouds are familiar to all of us from our youngest days. They can bring gloom or bright, fanciful shapes; they can bring gentle, life-giving rains or flood-inducing downpours, dancing snowflakes or driving blizzards. Anyone who flies regularly has seen how quickly one can go from the gray, rainy underside of clouds to the sunny, fantastical landscape of cloud tops. Despite this familiarity, many people know very little about clouds.

What are clouds?

Clouds are collections of liquid water droplets or ice crystals suspended for a time in Earth’s atmosphere. Clouds form when conditions are cold and moist enough to condense or freeze water vapor, and disappear when conditions are warm and dry enough to evaporate water or sublimate ice. In terms of moisture, this condition is referred to as saturation – the point at which the atmosphere holds as much water vapor as it can in a gaseous form; beyond the saturation point, moisture will begin to condense out of the atmosphere (dew on the ground comes from this source). But the formation of clouds is most often mediated by tiny particles in the atmosphere called cloud condensation nuclei (CCN) through a process called heterogeneous nucleation. This is the process whereby gas molecules of H₂O (water vapor) first stick to a condensation nucleus – a small particle floating in the atmosphere such as dust, smoke, sea salt, even small bits of biological matter kicked up by the wind – then clump together to form water droplets if enough water vapor is available to do so. Without these nuclei, clouds would only form in the coldest parts of the atmosphere, below −40 °C/F, where homogenous nucleation (water vapor directly with water vapor) can occur. This is because very high supersaturation levels (humidity above 100%) are required to overcome the surface tension of tiny, pure water droplets and allow more molecules to adhere and grow the droplet; but the presence of a foreign substance within the water enables the droplet to grow at much lower supersaturation levels. Ice nucleation – allowing the formation of ice crystals – requires very specific nuclei that facilitate water molecules arranging themselves in a crystalline structure to create solid ice. More details on these processes can be found, for example, in Rogers and Yau (1989).

How are clouds classified?

While clouds are infinitely variable, certain patterns recur and are meaningful. Luke Howard (1772–1864) created a useful and enduring classification of clouds first published in 1803 (Howard 2011/1865/1803; Hamblyn 2001). His classification used Latin terms to identify three primary classes of clouds:

- cumulus – heap, pile, mound: describes puffy, convective clouds;
- stratus – spread, laid out: describes flat, featureless cloud layers;
- cirrus – curl, fringe: describes wispy, high ice clouds.

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0584
### Clouds

#### High-level clouds:
- **Cirrus (Ci)**, **Cirrocumulus (Cc)**, **Cirrostratus (Cs)**

- **Cirrus** (Ci): in the form of filaments, strands or hooks, not progressively invading the sky.
- **Dense Cirrus** (Ci): in patches or entangled sheets, which usually do not increase, and seem to be the remains of the upper parts of a Cumulonimbus.
- **Dense Cirrus** (Ci): often in the form of an oval, being the remains of the upper parts of a Cumulonimbus.
- **Cirrus** (Ci): in the form of filaments, strands or hooks, progressively invading the sky and generally becoming denser. Cirrus and/or Cirrostratus progressively invading the sky, generally growing denser, but does not reach 45° above the horizon.

#### Mid-level clouds:
- **Altostratus (As)**, **Alto cumulus (Ac)**, **Nimbostratus (Ns)**

- **Altostratus** (As): the greater part of which is semitransparent; through which the sun/moon may be weakly visible, as through ground glass.
- **Thick Altostratus**, greater part sufficiently dense to hide sun (or moon), or Nimbostratus.
- **Alto cumulus**, the greater part of which is semitransparent; the various elements of the cloud change only slowly and are all at a single level.
- **Alto cumulus undulatus** (lens-shaped); most of which is semitransparent; occurs at one or more levels and continually changing in appearance.
- **Semi-transparent Altostratus** in bands, or Alto cumulus as one or more layers (semitransparent or opaque), progressively invading the sky.

#### Low-level clouds:
- **Cumulus (Cu)**, **Stratus (St)**, **Stratocumulus (Sc)**, **Cumulonimbus (Cb)**

- **Cumulus** (Cu): of fair weather, little vertical development and seemingly flattened.
- **Cumulus of moderate extent** with protrusions in the form of cones or towers, may be associated with Cumulus, Stratocumulus with bases at the same level.
- **Cumulonimbus**, the summits of which, at least partially, lack sharp outlines but are neither clearly fibril or in the form of an anvil.
- **Stratocumulus formed by the spreading out of Cumulus.** Cumulus may also be present.
- **Stratocumulus not resulting from the spreading or flattening of Cumulus**.

#### More weather phenomena:
- **Mammatus** are the drooping underbelts of heavy, rain-saturated clouds.
- **Tornadoes** are extremely destructive whirlwinds extending from the base of a thunderstorm.
- **Wall cloud**: Lowering from rain-free base of thunderstorm indicates area of strongest updraft. Tornado may form.
- **A shelf cloud forms in association with a cold front from a squall line or thunderstorm.**
- **Waves generated by a nearby thunderstorm ripple through the base of Stratocumulus.**

---

Figure 1 continued
These three primary cloud classes could then be combined to describe further variations in clouds, such as stratocumulus: a vertically restricted layer of extensive cumulus, and so forth. The cloud chart in Figure 1 provides one example of a modern classification based on this system. Many of these cloud types are associated with the occurrence or approach of a particular type of weather system, as noted in the figure. Thus, familiarity with cloud classification...
CLOUDS

is useful for anyone who spends substantial amounts of time outdoors, or whose work is in some way impacted by weather events.

How do clouds form?

There are a number of ways that the atmosphere can reach the saturation conditions to form clouds; these involve either a drop in temperature or an increase in the water content of an air mass, or both. This can occur in several ways, as follows.

- By wind forcing air upward over surface topography such as a mountain range. As air rises, it cools. This mechanism is called orographic lifting and explains how mountains can create a rain shadow on the downwind side of a range by causing most water to condense and precipitate out of the atmosphere on the upwind side of the range.
- By frontal lifting, in which air in a warm air mass is forced upward over an equally invisible, but denser, cold air mass. As the warm air rises, it cools, and water vapor may condense to form clouds. Thus, the arrival of a warm front often signals precipitation.
- By the simple process of heat rising, or convection. As sunlight warms a surface during the day, it in turn heats the air above it, which becomes less dense and begins to rise. If sufficient heating occurs, for example, on a summer day, this can cause air to rise and cool enough to reach saturation and form clouds. Often under these conditions a collection of cumulus clouds will form with their bases all at essentially the same level – the point at which the air has risen and cooled to the dew point temperature where water begins to condense. If heating continues, stronger convection will occur in the atmosphere, resulting eventually in towering cumulonimbus storm clouds.
- By radiative cooling, which occurs after the sun has set and is no longer warming the Earth’s surface. The air now cools due to emission of thermal or infrared radiation (heat) to space, and may drop below the dew point temperature. Fog often forms by this mechanism.
- By advective cooling in which air flows over a cooler surface and cools to its dew point. This may lead to coastal fog in regions where the land and water temperatures differ.
- By mixing some source of moisture into a cold, dry atmosphere.

A common question is “Why don’t clouds fall from the sky?” The answer, of course, is that they do – they are subject to gravity like everything else. However, because individual cloud particles are tiny, they fall very slowly against the resistance of the air: approximately 0.6 m/min. for a typical 10-micron diameter cloud particle (smaller than the width of a human hair). This allows even gently rising air currents (convection) to keep a cloud at a near constant altitude for quite some time. It also creates the distinctive appearance of cirrus clouds, in which falling ice particles drift downward through upper atmosphere winds into shapes that look like horses’ tails. And of course, individual droplets or crystals form and evaporate at different times in different parts of the cloud due to environmental conditions, including entrainment, the process by which drier air from outside the cloud gets mixed into the cloudy air. For example, droplets may evaporate at the edge of a cumulus cloud due to heating by the sun, while new droplets condense at the cloud base as convection lifts air upward to the dew point.

Once water or ice particles form, they can begin to clump together and grow into larger particles, and eventually some may become large enough to fall from the sky as rain, snow, sleet,
or hail. Raindrops, which are approximately 20,000 times larger than cloud droplets, fall about 400 m/min.; large hail falls even faster. Such precipitation provides a second mechanism by which clouds disappear or lose mass. Rogers and Yau (1989) provide more details.

How do clouds affect weather?

The connection between clouds and precipitation is clear and obvious to humans everywhere. While clouds may or may not bring precipitation, the absence of clouds is a near-perfect predictor of dry weather (although occasionally a few droplets may be blown by the wind from a nearby cloud to fall in an area of blue sky) – note the more cloud-free regions in Figure 2 correspond generally to drier areas of the surface. In effect, clouds are a visible massing of water in the atmosphere, which sooner or later will fall to the ground somewhere. In addition, clouds modulate the surface temperature of the Earth: cloudy or overcast days are generally cooler than clear days within the same season, while cloudy nights are warmer than clear nights.

Where precipitation falls is governed largely by the large-scale circulation of the atmosphere, which is in turn controlled by the movements of the sun and the rotation of the Earth. Several large-scale features resulting from these governing forces can be observed in Earth’s cloud patterns. The *intertropical convergence zone* (ITCZ) is a persistent band of cloudiness near the equator that results from convection near the subsolar point. It moves seasonally with the sun. In the

---

**Figure 2** Average global cloud fraction from 6 years (2007–2012) of CALIPSO satellite data obtained from http://mynasadata.larc.nasa.gov. Blue areas have less cloudiness than white areas. Source: MY NASA DATA.
CLOUDS

mid-latitudes, Coriolis forces create moving air masses and generate comma-shaped fronts that are associated with clouds in various ways depending on their relative temperatures. Satellite observations have revealed that the bulk of Earth’s precipitation falls on the oceans. Salby (1996) provides more details on these processes.

How do clouds affect climate?

One of the least appreciated roles that clouds play is in their modulation of the Earth’s energy budget. While human experience leads us to appreciate the cooling shade of a passing cloud on a hot sunny day, few people understand the global importance of clouds in controlling Earth’s climate. Clouds reflect sunlight – thus cooling the Earth – and also modulate the heat emitted from the Earth system, depending on their altitude (and thus their temperature). This combination of factors means that different cloud types can have different effects.

- High, thin ice clouds (cirrus, cirrostratus, etc.) reflect some sunlight but also reduce the amount of heat emitted to space because of their extremely cold cloud top temperatures (−40 °C/F and below) at high altitudes above 5 km. Taken together, this results in a local warming effect from such clouds.

- In contrast, low, thick water clouds (cumulus, stratus, etc.) reflect lots of sunlight but also emit large amounts of heat because of their relatively warm cloud top temperatures (just a little cooler than the nearby ground surface). Taken together, this results in a local cooling effect from low clouds.

Satellite measurements of the entire planet were necessary in order to understand the global impact of clouds (Ramanathan et al. 1989). Measurements beginning in the mid-1980s show that low clouds dominate in terms of the radiative effect of clouds, due in large part to the existence of persistent marine stratus cloud decks on the west side of land masses (such as San Francisco’s or the United Kingdom’s famous fogs). On balance, clouds have a huge impact on the Earth’s energy budget: were it not for their presence, the average global temperature on Earth would be 20 °C warmer.

The competing effects of different cloud types also explain why clouds remain one of the most important and challenging variables in the prediction of future climate. A successful model must be able not only to predict whether clouds will increase or decrease in a changing climate, but also how each type of cloud (high or low) will increase or decrease, as well as whether the clouds will thicken or thin, last for longer or shorter periods of time, and so forth (for more information, see Randall 2012).

How do humans affect clouds?

Clouds, driven by energy from the sun and the resulting water cycle, transport vast amounts of water and energy around the planet, so it might seem unlikely that humans could affect them. However, since the scale of each cloud particle is tiny, human activities can and do affect clouds. For example, adding condensation nuclei to the atmosphere (e.g., through the burning of renewable or fossil fuels) provides more locations for water to aggregate and results in clouds with more and smaller particles. These small particles grow more slowly and thus suppress precipitation, with the result that such clouds last longer than natural clouds. They also are brighter – another way of saying that they reflect more sunlight – because they contain many more small particles rather than fewer large particles (an example of this effect resulting
from ship traffic impacting low marine stratus clouds over the ocean can be seen online at http://earthobservatory.nasa.gov/IOTD/view.php?id=80203).

Airplanes can also affect clouds. By adding hot water vapor to the atmosphere, a product of fuel combustion within airplane engines, airplanes can “seed” the formation of a special type of mixing cloud, called a contrail – short for condensation trail. This occurs when conditions (temperature and humidity) in the atmosphere where the plane is flying are right: that is, it must be cold enough – typically −40°C/F and below – and moist enough that the addition of the plane’s exhaust brings the atmosphere to or beyond the saturation point (Schumann 1996). Because of the extremely cold temperature, this condensed water immediately freezes into ice crystals, which will remain until they sublimate. In areas with extensive air traffic, observation of the sky can thus provide clues to the moisture content of the upper atmosphere: if the atmosphere is too dry, the moisture in the exhaust will simply humidify the air slightly, but remain invisible vapor.

With only a little ambient water vapor present, condensation can occur but sublimation will proceed quickly as the water returns to a vapor state in the atmosphere. This leads to short-lived contrails – a short tail that follows the plane but quickly dissipates. When more ambient water vapor is present, the additional vapor and energy from the passage of the plane can generate a persistent contrail that may remain in the sky long after the airplane has passed. When the atmosphere is nearly saturated (perhaps due to emissions from other airplanes in high-traffic areas), the plane’s passage can create a persistent spreading contrail, which grows by condensing and freezing water vapor from the nearby atmosphere. Eventually such contrails can grow into extensive cirrus sheets, nearly indistinguishable from natural cirrus clouds.

On the other hand, airplanes can also cause existing clouds to dissipate. This can occur when heat from a plane’s exhaust causes cloud particles to evaporate, a phenomenon referred to as a “distrail,” short for dissipation trail (an example of this effect can be seen online at www.spc.ncep.noaa.gov/cooling/distrail/). Since clouds affect the Earth’s energy budget, these human impacts on clouds also mean that humans impact the energy budget and, like high and low clouds, some effects are toward warming and others toward cooling.

How do scientists characterize clouds?

Beyond simple visual observation and classification, scientists now use a variety of techniques to quantitatively measure and characterize clouds.

- The cloud fraction is a measure of the amount of sky covered by cloud. Its meaning is slightly different for an observer at a point on the ground who is affected by perspective and foreshortening, compared to a satellite remote sensing instrument which obtains a nearly two-dimensional view.
- The cloud height can be measured in a variety of ways, depending on point of view. From the ground a ceilometer – a device that sends a light or laser beam upwards, and then detects its reflection from the cloud – can be used to measure cloud base (referred to by pilots as the ceiling). From space, passive remote sensors can infer an effective cloud top height by measuring the temperature of the cloud. Active remote sensors (radar or LiDAR) can determine the cloud top height by ranging methods.
- The effective temperature of a cloud, typically measured by satellite remote sensing, provides a measure of the amount of heat
Clouds

released to space by the cloud top (and is also used to estimate cloud height as above).

- The liquid (or ice) water content of a cloud, as its name suggests, is a measure of the total amount of water contained in the cloud, or in a segment of the cloud. This is often measured by satellite remote sensing, which allows scientists to measure the total water content in a vertical column of the atmosphere.

- Cloud phase describes whether the particles in the cloud are predominantly liquid droplets, ice crystals, or a mixture of the two.

- The cloud particle size is a measure of how big the water droplets or ice crystals in the cloud are. Typically particles will be distributed across a range of sizes and one can characterize the particles with a variety of statistical measures such as mean or effective size and various moments of the size distribution. Different size particles interact differently with light and heat of different wavelengths traveling through the cloud.

- Optical depth is a measure of the visual opacity of a cloud – how opaque it is – as well as characteristic of the mean free path of light photons within a cloud. Mean free path is, statistically, the average distance a photon (or light ray) travels within the cloud before being scattered, absorbed, or reflected. Light transmission through the cloud falls off as \( \exp(-\text{optical depth}) \). Experience in heavy fog can illustrate just how quickly visibility can be reduced.

How can individuals observe clouds?

Clouds and citizen science

As a uniformly accessible but highly changeable part of the Earth system, clouds have inspired a variety of forms of citizen science and other types of observation projects. These include the following.

- The Global Observations to Benefit the Environment (GLOBE) Program, a global student program of environmental observation. Its cloud observation protocol – one of more than 50 observation protocols – has been popular since the program began in the mid-1990s. Thus GLOBE has a large amount of cloud observation data in its database. With the launch of CloudSat, a satellite carrying the cloud profiling radar instrument, in the mid-2000s, a subset of GLOBE participants have been involved in more in-depth cloud observations, which are used to help validate the CloudSat retrievals.

- The Students’ Cloud Observations On-Line (S’COOL) project, a similar, but more focused, project to obtain ground truth observations of clouds and sky conditions at the time that a NASA satellite carrying the Clouds and the Earth’s Radiant Energy System (CERES) instrument passes overhead. Since 1997, students around the world have provided tens of thousands of observations to this project. Participants are provided with matching satellite cloud retrieval data (where applicable) to allow them to pursue further scientific inquiry around clouds and remote sensing.

- The SatCam iPhone app developed over the past few years to enable mobile device users to provide photographic validation of clouds for the Moderate Resolution Imaging Spectroradiometer (MODIS), an imager on the Terra and Aqua satellites. This app uses geolocation features on the iPhone to precisely locate an observer in the satellite image, which it returns to the user after a photo is submitted.

- The Cloud Appreciation Society was established in the United Kingdom as a place for
cloud lovers to unite. Apart from its website, it has also created a CloudSpotter iPhone app, which allows users to submit and classify photos of clouds, then receive expert confirmation on their cloud classification.

SEE ALSO: Atmospheric/general circulation; Climatology; Precipitation; Weather, extreme

References

Cluster detection

David C. Wheeler
Virginia Commonwealth University, USA

Background

Investigating a spatial pattern of events or event counts to find a geographic area with statistically significantly increased likelihood of event occurrence is known as spatial cluster detection. A spatial cluster, therefore, may be defined as a statistically unusual collection of events in a geographic area when compared with the pattern expected under a particular event-generating process (Besag and Newell 1991). Cluster detection is used frequently in many disciplines where a spatial analysis is conducted, including archeology, biology, criminology, health services, and epidemiology, among others. Spatial cluster analysis has been used in archeology to determine if there are clusters in ancient grave sites of different time periods, and also to identify clusters of individuals in grave sites according to a common characteristic, such as extracted wisdom teeth. In biology, cluster detection may be used to determine if there is a cluster of a certain disease among a species of trees, or to determine if there is a cluster of a particular species of tree in a forest. In criminology, cluster detection is used to identify hotspots of crime within a city. Cluster detection has been used in health services to demarcate catchment areas for hospitals. In spatial epidemiology, cluster detection is used to identify areas of significantly elevated or decreased risk of disease incidence or mortality, potentially to reveal etiologic clues.

A major motivation for performing cluster detection in spatial epidemiology is that many diseases have risk factors that are distributed unevenly in the environment. Examples for cancers include lung cancer and radon, bladder cancer and arsenic, and leukemia and benzene. The uneven distribution of known or unknown risk factors can lead to spatial patterning in disease risk. When risk factors are unknown, studying spatiotemporal patterns in disease events may reveal clues about disease etiology. In fact, there is a long history of research analyzing geographic patterns in disease incidence and mortality with the objective of discovering environmental determinants of disease. Examples of risk factors revealed by analytic epidemiologic studies that followed upon observations of geographic patterns of cancer include exposure to asbestos from shipyards as a risk factor for lung cancer among men along the southeastern United States seaboard and chronic use of snuff as a risk factor for oral and pharyngeal cancer among women in the southern United States.

Due to the relationship between disease risk and known risk factors that may exhibit a spatial pattern, it is crucially important to adjust for known covariates when conducting cluster detection analysis. Failing to account for known risk factors may lead to detection of an uninteresting cluster, one that would be expected to occur if the risk factors were considered. For example, age is often positively associated with cancer risk and could lead to a detected cluster of disease in an area with a relatively large elderly population. To focus this discussion of cluster detection, the following review of methods is
limited to a few probabilistic models that allow for adjustment for covariates. This review is not intended to be comprehensive.

Methods

Contemporary cluster detection methods assess the statistical significance of the elevated risk in some area to evaluate the area as a cluster. Of interest is the question of how likely the area of elevated risk is under a certain null hypothesis. A typical null hypothesis is constant risk, where all areas are assumed to have the same risk after adjusting for pertinent, known risk factors. The null risk level may be taken to be the overall risk in the entire study area. Under the null hypothesis, the disease risk does not vary within the study area. An area with an elevated risk well over the null risk may be detected as a cluster if the statistical test has good power in the area, where power is the probability of the test to reject the null hypothesis when the null hypothesis is false. The methods to detect clusters operationalize the null hypothesis through probability statements of a particular quantity of interest to summarize how unusual the observed risk is in some area. The methods develop a distribution of the quantity of interest under the null model and then compare the quantity of interest for the observed data to the distribution under the null model. In this way, cluster detection is conducted in a statistical hypothesis-testing framework by comparing the observed risk measure to the expected risk measure or statistic under the null hypothesis. Rejecting the null hypothesis in some area by finding an unlikely risk measure compared with the expected value under the null leads to cluster detection. For example, a county in the United States with eight diagnosed cases of esophageal cancer in one year when only two cases were expected based on the national rate of esophageal cancer incidence may be determined to be a cluster in a nationwide study. In this example, cancer incidence in the local area, the county, is compared with the national incidence rate in a relative risk (null value of 1) to determine what is unusual assuming that the national rate applies equally to all areas.

The method one uses for cluster detection depends on the type of data under study. Different methods exist for individual-level data and aggregate data. Individual-level data are defined by the geographic coordinates for some location for each subject in the study. The type of location is typically the residence, but it could also be a place of employment. In either case, the geographic coordinates are determined from an address-matching process of reported addresses. Individual-level data also include an outcome variable value, which is often a binary disease status indicator (disease present or absent). Covariate values for individuals are typically available for case-control and cohort studies used in cluster detection and commonly include age, gender, and race. Lifestyle factors, environmental exposures, and socioeconomic status variables such as income and education may also be available for individuals. In contrast to individual-level data, aggregate data are counts of the outcome variable and measures of the covariates within defined geographic areas. The geographic unit is often an administrative boundary, such as a census tract or county in the United States, or a postal code, such as a ZIP Code in the United States. Importantly, aggregate data also include a measure of the at-risk population that is used to calculate the disease rate or proportion in each area. It is crucial to consider the size of the at-risk population when conducting cluster detection because areas with larger populations will be expected to have more counts of the outcome variable under the null hypothesis. For example, an area with 10 new
cases of childhood asthma and 1000 residents will be less likely to be classified as unusual than would an area with 10 new cases and 100 residents. In case-control studies, the controls are a sample from the at-risk population and are used as a contrast with the cases in terms of spatial location in a cluster detection analysis. Advantages to using individual-level data over aggregate data include more precise spatial information and more specific information on risk factors, particularly for levels of environmental exposures. With aggregate data, it would be necessary to use a mean environmental exposure in each area instead of individual-based exposures.

One method for cluster detection with individual data is the generalized additive model (GAM), which provides a flexible approach to modeling spatial residual variation in risk after considering known risk factors. This type of model is a natural fit for a case-control study with individual-level spatial information. Consider \( i = 1, \ldots, n \) subjects located within a region \( A \), each with known residential location \( s_i \) at a particular time \( t \) and a binary label \( Y_i \) for disease status. Defining the spatial risk function as \( r(s) \), equal to the probability that a person at location \( s \) will be a case, the general spatial log-odds function is \( l(s) = \log[r(s)/(1 - r(s))] \). We can model the log-odds of disease through an adjusted generalized additive model as:

\[
\log \left[ \frac{P(Y_i = 1)}{P(Y_i = 0)} \right] = \alpha + X_i \beta + l_i(s_i) \tag{1}
\]

which adjusts for covariates \( X_i \) parametrically with coefficients \( \beta \) and an intercept \( \alpha \). This model could include known environmental exposures as covariates. The model includes a nonparametric term for the spatial log-odds using residential locations at lag time \( t \). The spatial log-odds is a bivariate function of spatial coordinates and models residual spatial variation in risk. The residential locations at only one time are used in this model specification. In a typical study, the time would be the time of diagnosis of disease or time of study enrollment. If spatial information for other times is available, through residential histories for example, then the spatial risk component at time lags can be evaluated. The residual odds ratio surface may be calculated through the spatial log-odds and the mean of the spatial log-odds by \( \exp[l(s) - \bar{l}] \). The spatial component estimated by the GAM can be visualized by mapping the log-odds or odds ratios, or by mapping risk using a logarithmic base 2 scale where each unit increase corresponds to a doubling in risk. Two convenient forms of the spatial log-odds function are a locally weighted scatterplot smoother (LOESS) and a thin plate regression spline (TPRS) smoother (Wood 2006). These smoothing functions contain a smoothing parameter than can be estimated by minimizing the Akaike Information Criterion (AIC).

The null hypothesis of primary interest in this type of analysis is that the risk is constant over space, that is, \( r(s) = r \). An approach to evaluate the significance of the spatial component in the GAM for both types of the smoothing function is to use a Monte Carlo permutation test. This test has conditions based on the number of cases and the locations of subjects at one time point, performs randomization of the case labels among the locations, fits the model to the randomized data, and calculates the change in deviance with the spatial component in the model. Typically, 999 randomizations are used to build the permutation distribution for the test. Random labeling of subjects as cases or controls is equivalent to the null hypothesis of constant risk over space. Thus, the random labeling process is akin to repeatedly generating data from the null hypothesis. The \( p \)-value for the overall spatial term is calculated by dividing
CLUSTER DETECTION
the rank for the observed data of the change in
deviance among the permutation distribution
by the number of randomizations. Statistical
significance of the local log-odds or odds ratio
can also be evaluated with the Monte Carlo
permutation procedure. The local log-odds or
odds ratio, either calculated at each data point
or predicted for each cell of a grid overlaid
on the study area, is regarded as statistically
significant if it is outside the 2.5% and 97.5%
ranked values from the local permutation dis-
tribution. These areas of statistically significant
risk are the detected clusters. When multiple
locations for subjects are available over a range
of time, this randomization procedure may be
used to determine the temporality of the spatial
component that is most associated with risk of

A method for cluster detection with aggregate
data is to use Bayesian hierarchical modeling.
This can be done for outcome counts assum-
ing either a binomial or Poisson distribution.
Cluster detection with Bayesian hierarchical
models may be performed using either the
model residuals or the exceedence probabilities
for a quantity of interest, such as a relative risk,
compared with some threshold value (e.g., 1 for
relative risk). Focusing on a model for Poisson
distributed counts, one can denote the counts
in each area as
\[ Y_i \sim \text{Poisson}(E_i \theta_i) \]
where \( E_i \) represents the expected number of cases and \( \theta_i \) is the relative risk in area \( i \). The expected
counts are based on an overall risk in the study
area. One can specify a log-linear model for
the relative risk that includes area-specific risk
factors \( X_i \) as:

\[
\log[\theta_i] = \alpha + X_i \beta + v_i
\]

where \( \alpha \) is the intercept and \( v_i \) denotes spatially
unstructured random effects. The model param-
eters can be estimated with Markov chain Monte
Carlo (MCMC) simulation in software such as
WinBUGS. The Bayesian residuals for this model
are calculated as
\[
\hat{r}_i = (y_i - e_i \hat{\theta}_i) / \sqrt{e_i \hat{\theta}_i},
\]
where \( \hat{\theta}_i \) is the mean estimate of the relative risk from
the MCMC samples. Mapping the probability
that the residual exceeds some threshold value,
say 2 or 3, can identify hotspots of elevated,
unexplained risk. Another approach is to define
an exceedence probability as the probability that
the relative risk exceeds a threshold level \( c \). The
exceedence probability is estimated from the
MCMC samples as:

\[
\text{Pr}(\theta_i > c) = \frac{\sum_{g=1}^{G} I(\theta^g_i > c)}{G}
\]

where \( I() \) is an indicator function that returns
1 if the expression in the function is true and
0 otherwise, \( \theta^g_i \) is the relative risk from the
\( g \)th MCMC sample, and \( G \) is the number of
MCMC samples. This exceedence probability
is the proportion of the MCMC sample values
of the relative risk in the \( i \)th area that exceed
the threshold value, where the threshold is
frequently set to 1. Hotspots may be identified
by selecting the exceedence probabilities that are
greater than a conventional level, such as 0.95
or 0.99.

A method that can be used for cluster detec-
tion with either individual-level or aggregate
data is the local spatial scan, which scans over the
study area with circles of varying sizes to find the
most unusual area of elevated risk. The method
compares the null hypothesis of constant risk
to alternative hypotheses where the disease rate
inside a scanning window is greater than outside
the window. The scanning window size is set
to contain no more than a certain proportion
of the total population in the study area, for
example 50%. A test is calculated based on the
maximum likelihood ratio statistic across all
circles as:
where $Y_j$ and $E_j$ represent the observed and expected number of cases inside circle $j$, $Y_+$ is the total number of observed cases, $Y_+ - Y_j$ is the number of cases outside the circle, and $I()$ is an indicator function taking on the value of 1 if the observed number of cases exceeds the expected number in the circle. The sampling distribution of the test $L$ under the null hypothesis is constructed using a large number (e.g., 999) of Monte Carlo randomizations of the case labels, and then the observed $L$ is compared with the null distribution to calculate a p-value associated with the window with the maximum likelihood ratio statistic. The most likely cluster can be identified using the software SaTScan (www.satscan.org), and then mapped using other statistical or geographic information system software. It is not possible to directly adjust for covariates with the local spatial scan method implemented in SaTScan, but one can perform analyses stratified by a categorical variable, where a spatial scan test would be calculated for each level of the covariate.

There are several issues that are problematic for cluster detection. One of the important issues for cluster detection in spatial epidemiology is residential mobility. Most cluster detection analyses assume that pertinent unknown exposures occur in and around the home, but use only one reported address per individual, which is typically the address where subjects were diagnosed with disease. This makes the assumption that subjects do not move, or that all environmental exposures relevant for disease occur at only the most recent location. This is a major assumption that becomes increasingly wrong as disease latency increases, and as population mobility increases. The power of cluster detection tests to identify a spatial signal in risk will suffer as the assumption becomes more tenuous. For other applications of cluster detection, such as crime hotspot analysis, population mobility is much less of a concern. The power to detect a cluster generally varies from test to test, and also depends on the magnitude of the risk in the potential cluster, the shape of the elevated risk area, and the location of the elevated risk area within the study area (Waller, Hill, and Rudd 2006). The spatial scale of the cluster detection analysis is an issue with aggregate data, as a small area of significantly elevated risk at one spatial scale (e.g., census block group) may not be detected when using a larger spatial unit (e.g., county). This is related to the well-known modifiable areal unit problem. Of course the results of a cluster detection analysis will depend on the address–matching completeness and spatial accuracy of the geocoded addresses. Using many addresses matched to a centroid of a large geographic unit may result in a relatively uninformative cluster if relevant exposures are thought to occur at the neighborhood level.

**SEE ALSO:** Environment and health; Geocoding; Health geography; Public health: human dynamics; Quantitative methodologies; Spatial analysis; Spatial epidemiology; Spatiotemporal analysis

**References**


CLUSTER DETECTION


Further reading

Coastal depositional processes and landforms

Nancy L. Jackson
New Jersey Institute of Technology, USA

Coastal depositional environments are generally located on low gradient passive margins where there is sediment of a size that can be moved by waves or wind. The combined effects of the processes of winds, waves, tides, currents, and sea level give rise to a number of emergent depositional landforms including barrier islands and spits, and associated beaches, dunes, and marshes, as well as sedimentary features at the shoreline such as salients and tombolos (Figures 1 and 2). The spatial and temporal scales of processes and landform adjustments range from broad-scale evolution of barrier islands in response to long-term sea level and sediment supply variation, to small-scale adjustment of beaches and dunes in response to episodic storm and non-storm conditions. Human processes also have a direct influence on the evolution and adjustment of coastal depositional landforms by changing the size, shape and location of these features via beach nourishment, dune building, and inlet dredging and filling activities.

Barrier islands and spits

Barrier islands are wave-built, emergent, shore-parallel elongated sedimentary accumulations of unconsolidated sand and/or gravel that generally form chains separated by tidal inlets. They comprise about 5–10% of the world’s coast and are generally associated with passive margin shorelines. The longest barrier island chains are located on the US Atlantic and Gulf coasts but barrier islands can be found throughout the world, including South America, India, and Europe.

Barrier islands are found in micro- (<2 m) and meso- (2–4 m) tidal environments but not in macro- (>4 m) tidal environments, where the lack of wave energy concentration at a single elevation prohibits their formation. The planform morphology of barrier islands is a function of wave height and tidal range. Barrier islands in wave-dominated environments are generally long, narrow, and straight with a pronounced flood tide delta, few tidal inlets, and prominent overwash deposits. Barrier islands in mixed-energy environments (wave and tide processes) tend to be short and wide with large, well-developed ebb tidal deltas and numerous tidal inlets.

Barrier islands are important as a buffer from storms to the mainland and protection for the ecologically productive back-barrier lagoon or bay environment. Barrier islands are defined separate from other barrier formations because they comprise a system of distinct interactive sedimentary subenvironments. The subaerial portion of the barrier includes a beach and dune system and back-barrier shoreline and is underlain by a subaqueous platform fronting the shore face. The barrier is bounded by tidal inlets and associated ebb and flood tidal deltas on their ends, and is backed by a low-energy lagoon or estuarine system. Marshes are present on the sheltered reaches of the back-barrier and mainland shorelines of these basins (Oretel 1985). Although barrier islands are commonly associated with oceanic environments, the
existence of barriers on estuarine shorelines has sparked recent debate on whether these estuarine landforms can be classified as barrier islands (Otvos 2012).

Formed during the Holocene transgression, there are four explanations for the genesis and development of barrier islands (Davis and Fitzgerald 2004). The first explanation involves the presence of nearshore sand barriers that subsequently migrate inland with sea level rise to form emergent barrier islands. The second explanation is that barrier islands begin from
the development of sand spits due to longshore sediment transport from an existing headland. Breaching of the spit during storm activity and development of an inlet establishes the barrier island. The third explanation is that barrier island formation is the result of sea level rise and subsidence of mainland shorelines drowning low elevation areas adjacent to beach and dune ridges on the mainland. The low elevation areas backing the beach and dune ridges form lagoon systems with marsh environments. The fourth explanation is that barrier island development involves a combination of all three explanations.

Barrier islands can migrate based on available sediment supply and sea level variation. Transgressive migration occurs where sediment supply is limited or relative sea level rise is high. Transgressive barriers are narrow, low in elevation and characterized by numerous washover deposits. These deposits form when waves and surge overtop or breach the dune ridge during storms and result in the landward transport of sediment from the ocean side toward the back-barrier environment. Regression occurs where sediment supply is ample or the rate of sea level rise is low. A prominent feature on regressive barriers is multiple beach or dune ridges, which can be vegetated and indicate the position of former shoreline positions.

Barrier spits are emergent, elongated, narrow, and composed of sand and/or gravel but differ from barrier islands in that they are attached to the mainland or an island. They develop on the downdrift end of littoral cells and, thus, evolve by longshore transport of sediment from the mainland. The shape and orientation of barrier spits are a function of wave angles and local water depths. The distal end of a spit terminates in open water and is often the location of multiple beach ridges that develop in response to an alternating increase and decrease in longshore sediment transport. Transport of sediment into deep water leads to the growth of a spit platform at the distal end. Wave refraction leads to a curvature at the distal end and transport of sediment bayward. Barrier spits that prograde completely across entrances to bays or lagoons are referred to as baymouth barriers (Otvos 2012). Barrier spits may breach during storms forming an inlet, separating the landform from the mainland, to form a barrier island.

Beaches

Beaches are located on exposed oceanic shorelines as well as on back-barrier shorelines or within estuarine or lagoon systems. Beaches are sloping unconsolidated deposits of sand and/or gravel that form at the land/water intercept where there is sufficient wave energy to rework available sediment. The beach profile extends from the position where the subaerial backshore intersects the upland or dune toe to the subaqueous nearshore zone where one or more bars may be present.

Research on the morphodynamics of beaches examines the mutual interaction of morphology and hydrodynamics. Chief among these interactions is between the type of wave breaking (an indication of the degree of reflection or dissipation of wave energy) and morphology as a function of the textural properties of the sediment and slope of the bed. Wright and Short (1984), using field data from a number of beaches in Australia, classified beach systems in high-energy microtidal environments as one of six possible states based on the surf similarity parameter ($\varepsilon$) (Guza and Inman 1975):

$$\varepsilon = \frac{(a_b \omega^2)}{(g \tan^2 \beta)}$$

where $a_b$ = breaker amplitude, $\omega$ = incident wave radian frequency ($2\pi/T$ where $T$ = period),
COASTAL DEPOSITIONAL PROCESSES AND LANDFORMS

$g =$ acceleration of gravity, and $\beta =$ the gradient of the beach and surf zone. The two end members of the beach classification spectrum are reflective and dissipative beaches. Reflective beaches, $(\varepsilon < 2.5)$ are analogous to a swell or summer profile and characterized by a steep foreshore slope, the lack of an offshore bar or surf zone, and coarse sediment. The backshore is narrow and there is a pronounced step at the base of the foreshore. Waves surge up the steep foreshore and may reach 1 m in height. The beach foreshore is generally planar but cusps may form under higher wave energy. Dissipative beaches $(\varepsilon > 20)$ are analogous to the classic storm or winter profile. They are characterized by fine sediment and spilling waves (>2 m) that dissipate energy over a wide low gradient surf zone. There is large subaqueous sediment storage in the surf zone and generally at least one to three linear shore-parallel bars present. Between the reflective and dissipative limits are intermediate beaches that are distinguished by both reflective and dissipative conditions and the presence of bars and rip currents in the surf zone.

Over time, beaches will exhibit a modal state based on wave conditions and sediment characteristics. Using a modified dimensionless fall velocity parameter $(\Omega = H_{b}/W_{s} T)$, where $H_{b}$ is the wave height at breaking, $W_{s}$ is the sediment fall velocity, and $T$ is the wave period, Wright and Short (1984) showed that beaches were modally reflective when $\Omega < 1$ and modally dissipative when $\Omega > 6$. Masselink and Short (1993) extended the work of Wright and Short (1984) to include meso- and macro-tidal beaches within the morphodynamic classification system. They introduced a parameter called the relative tide range $(RTR = MSR/H_{b})$ where $MSR$ is the mean spring tide range and $H_{b}$ is the wave height at breaking. The RTR parameter reflects the importance of swash, surf zone, and shoaling wave processes on the beach profile over the tidal cycle. As the RTR increases, modally reflective beaches transition from low tide terrace beaches with rips to low tide terrace beaches without rips and modally dissipative beaches transition from nonbarred dissipative to ultra-dissipative beaches.

Coastal dunes

Coastal dunes are aeolian depositional landforms formed under winds capable of entraining available sediment on the beach and where there is an obstacle to reduce winds and allow deposition to occur. The obstacle is often vegetation or wrack accumulation on the backshore but on highly developed coasts, where vegetation or wrack is intentionally eliminated by human action, a sand fence with artificial plantings may be used to trap sediment. Vegetation traps sediment in transport creating shadow dunes. Growth of vegetation after burial by sediment deposition creates conditions for additional sediment deposition and growth of the dune form. Vegetation also may emerge landward of the toe of the dune and create new areas for sediment deposition.

Several models have been developed to predict aeolian sediment transport potential. Application of these models to coastal environments has yielded higher rates of transport compared to actual transport rates observed on beaches. One of the most frequently used equations to quantify the rate of transport $(q)$ is based on the work of Bagnold (1941) and relates the rate of transport to sediment size on the backshore and the shear stress imparted by the wind to the sand surface:

\[
q = C(\rho_{a}/g)(d/D)^{0.5}u_{*}^{3}
\]

where $q$ is the rate of sediment transport, $C$ is an empirical constant ranging from 1.5 to 2.8, $\rho_{a}$ is air density, $g$ is acceleration due to gravity, $d$ is grain diameter, $D$ is a reference grain diameter of 0.25, and $u_{*}$ is shear velocity.
The underlying assumptions of many aeolian transport equations, including the above, are that the wind is unidirectional, fully turbulent, uniform, and steady, that the sediment is uniform in size and dry, and that the bed is planar, flat, and unobstructed. Often one or more of these conditions are violated in coastal environments. For example, the presence of salt crusts, vegetation, and wrack accumulations on the backshore, surface sediment moisture, topographic variation across the backshore, and changing fetch distance as a function of beach width, can reduce sediment entrainment and transport potential for dune formation. Field studies of aeolian sediment transport on beaches have refined sediment transport models over the past two decades, providing insight to turbulence and unsteadiness of the wind, fetch distances required to achieve full transport potential, and spatial and temporal variability of transport rates due to the presence of surface sediment moisture, lag deposits, and microtopographic variability across the backshore (Davidson-Arnott 2010).

The interaction and sediment exchange between beach and dune are critical to understanding the size, shape, and persistence of dunes on the backshore. The likelihood for dune formation and growth has been explained using a morphodynamic approach over short time scales and a sediment budget approach over longer time scales. The morphodynamic state of a beach, driven by nearshore waves and currents, influences sediment characteristics, bed slope, and backshore width. These attributes, in turn, influence the potential rates of sediment transport for dune formation. The dune provides a source of sediment to the beach and nearshore system during storms and the net result is an exchange of sediment between the beach and dune system. Dunes will form on both reflective and dissipative beaches but dunes will be smaller and lower in height on reflective beaches due to the narrow backshore widths and steep slopes (Short and Hesp 1982). The sediment budget approach links sediment volume in the beach and dune and between beach and dune to the size of a dune with increasing dune size occurring under a slightly negative beach budget (Psuty 1988).

Saltmarshes

Saltmarshes and mangroves are found in estuaries and lagoons at the margins of rivers and on the landward side of barriers islands and spits. Saltmarshes are found in the mid- and upper latitudes and their location in the estuary or lagoon is a function of wave energy, tidal flow, sediment availability, and recent sea level rise history. In tropical areas the marsh will be vegetated by mangroves. The low wave energies in estuaries and lagoons promote growth of salt-tolerant vegetation that is capable of dissipating flow velocity, focusing fine sediment deposition, and increasing the surface elevation of the marsh. Sediment may be derived from the drainage basin, from the ocean via transport through the tidal inlets or via storm overwash of the barrier. Biomass production, influenced by subsurface processes such as groundwater flow and oxygen availability, is also an important contributor to marsh development (Townend et al. 2011). Tidal creeks within the marsh provide the transport network for the movement of sediment and organic matter. Saltmarshes occupy the upper intertidal zone and there is a zonation of vegetation based on surface elevation, and frequency and duration of inundation (hydroperiod). Less salt-tolerant species (i.e., *Spartina patens*) are found at the mid- and upper elevations of saltmarshes where tidal inundation occurs less frequently and over short durations (i.e., spring high tide). More salt-tolerant species (i.e., *Spartina alternafora*) occupy the lower elevations.
where tidal inundation occurs daily and for longer durations.

The elevation of a saltmarsh is a function of tidal range but the long-term stability of the system is a function of the marsh surface elevation relative to eustatic sea level rise or local subsidence. With an increase in sea level or local subsidence and no increase in surface elevation the marsh system will translate inland or drown in place if there is insufficient accommodation space for migration. Most influential in the loss of saltmarsh systems are the cumulative effects of human action. Diking to allow conversion for agricultural production, filling to provide a substrate for human development, and restricting the inland migration with human structures have reduced both the extent and function of these systems.

Salients, tombolos, and cuspate forelands

Salients are triangular, accretionary shoreline features that can form in the shadow zone of an offshore island or reef system. Shore protection structures such as detached breakwaters also result in the development of shoreline salients with the ratio of the length of the offshore obstacle and the distance of the obstacle to the initial shoreline influencing the morphology of the salient. Wave refraction and diffraction around the obstacle result in lower energy conditions and sediment deposition within the lee of the structure or obstacle. When the apex of a salient attaches to the offshore feature it is called a tombolo. A cuspate foreland is a large-scale, triangular, accretionary shoreline feature and may occur in isolation or in a series. Examples of these features include Dungeness, on the south coast of United Kingdom, and Cape Hatteras and Cape Canaveral on the southeast coast of United States. Several explanations for the formation of cuspate forelands have emerged, including changes in wave magnitude and direction, reworking of deltaic deposits, and wave refraction around offshore shoals (Davis and Fitzgerald 2004).

SEE ALSO: Coastal zones; Coasts

References


Coastal erosion processes and landforms

Alan S. Trenhaile
University of Windsor, Canada

Most types of coast experience occasional erosional events that can lead, for example, to the formation of scarps on beaches and fore-dunes, and low muddy cliffs along the seaward margins of salt marshes. Erosional features on primarily depositional landforms are generally quite ephemeral and quickly masked by renewed deposition. Conversely, erosion is irreversible and is the dominant mode of development of rock coasts. Rock coasts are found in all types of environment, in tropical regions where they are often cut into raised coral limestones, in cooler, frequently storm wave-dominated middle latitudes, and in high latitudes where they may have been eroded by glaciers in the past and are subjected to the effects of frost and shore ice today. Bare, rocky surfaces are generally dominant, but there are often small, stony pocket beaches between rocky headlands and sandy beaches over shore platform foundations. Rocky coasts can be an important source of sediment for adjacent coastal areas depending on: rock resistance and the strength of the erosional agents, which determines rates of erosion and sediment production; the grain size of the sediment, which reflects rock grain size, joint density, and weathering efficacy; and whether the potential sediment sinks are wave-dominated, which promotes deposition of coarse-grained sand and gravels in beaches, or tide-dominated which promotes deposition of fine-grained clays and silts in tidal flats, salt marshes and mangrove swamps, and deltas.

Erosional processes

Rock coasts experience a very wide range of erosional agents whose efficacy varies with the chemical and physical characteristics of the rock and the wave regime, climate, and other aspects of the morphogenic environment. Although the terms are poorly defined, the distinction has been made between “hard” and “soft” rock coasts. Soft rock coasts are generally considered to consist of fairly nonresistant materials such as cohesive clay, marl, shale, and highly weathered rock with low resistance to mechanical wave erosion, and cliff recession rates of more than 10 meters per century. Hard rock coasts (limestone, gneiss, granite, etc.) are much more resistant to wave erosion, with cliff recession rates generally ranging from almost nothing up to a few meters per century.

Mechanical wave erosion, including abrasion by sand and gravel and hydraulic quarrying by water hammer (impact), shock pressures from breaking waves, and air compression in rock discontinuities, is usually the dominant mechanism in the storm wave environments of the mid-latitudes and in other areas where there is strong wave action. Quarrying processes depend on alternations of air and water, and abrasion is most effective in shallow water where wave-generated bottom currents are strongest. As standing, breaking, and broken waves also generate the highest pressures at, or slightly above, the mean water surface, mechanical wave erosion must be most effective at or close to this
level, although wave-generated bottom currents are also effective on cohesive clay and other soft rock coasts. Only a few workers have measured wave transformation and energy dissipation on rocky foreshores. Recent studies have shown that infragravity energy is present on shore platforms, although it is unclear whether it plays any role in their erosion and evolutionary development.

Because of abundant moisture and saline solutions, coasts provide nearly optimum conditions for many weathering mechanisms, including chemical and salt weathering, frost, and wetting and drying. Most weathering, with the exception of frost which is important in cool coastal regions, is generally considered to be most effective in warm to hot environments. Weathering may be the dominant mechanism in such areas today, but by operating along joints it also facilitates hydraulic quarrying in temperate, wave-dominated environments.

Microflora and faunal borers and burrowers erode rock surfaces whereas vermetid gastropods, coralline algae, and other organisms produce hard crusts that protect the underlying rock. Bioerosion, bioconstruction, and bioprotection play important roles in the formation of deep notches, organic protrusions (corniche), and narrow shore platforms (trottoir, surf ledges) on coral, aeolianite, and other calcareous substrates in tropical and Mediterranean environments.

Bioerosion also occurs in cooler regions where its main role may be to promote limestone solution and karren formation through faunal respiration and carbon dioxide production in pools, which reduces the pH of the water at night when it cannot be extracted by algae.

Rockfalls, landslides, and other types of mass movement result from the undercutting and oversteepening of coastal slopes by waves and other marine processes. The type, size, and frequency of these slope failures are determined by rock structure, lithology, and other geological characteristics, and by the absolute and relative intensity of the subaerial and marine processes. Rock falls typically occur in well-fractured rocks, translational slides in seaward-dipping rocks, alternations of permeable and impermeable strata, and massive rocks overlying incompetent materials. Rotational slumps occur in thick, fairly homogeneous deposits of clay, shale, or marl. Human activities promote slope movements in some places because of the addition of water from septic systems, irrigation, and runoff disruption, and the depletion of protective beach material by the construction of coastal obstructions and the damming of rivers.

Rates of erosion

Much of the data on cliff recession rates, which range from virtually nothing up to 100 meters per year, is of questionable quality. Although there are erosional data for many parts of the world, it is impossible at present to identify any reliable relationships with rock type or wave and tidal conditions. Cliff retreat rates have been measured or estimated in a variety of ways, using sequential terrestrial and aerial photography, old maps, repeated surveys, cliff-top stakes or metal pins driven into the cliff face, morphological evidence, erosion of ancient and modern anthropogenic structures, and dated inscriptions. Unfortunately, whereas cliff recession is strongly episodic, with long intervals between large slope failures, the most precise techniques have been used usually only over short periods of time. Sequential air photographs and satellite images can be used to determine recession rates over half a century or more, although there may be too little change to hard rock coasts over this period, and too much change between consecutive images on soft rock coasts. Sequential airborne and terrestrial LiDAR is increasingly used to record cliff recession and other elements of rock
Coastal erosion processes and landforms

Coast evolution and, by scanning from the beach or shore platform, it can record the detachment of small rock fragments ranging from only a few centimeters in size up to large falls and slides. Microdrones may also provide an additional source of high-resolution airborne data in the future.

Micro-erosion meters have been used by many workers to measure rates of surface downwearing (erosion in the vertical plane) by weathering and abrasion in the intertidal and supratidal zones. These instruments cannot be used to record the dislodgement of larger rock fragments, however, or to measure backwearing (erosion in the horizontal plane) resulting from the retreat of rock scarps. Additionally, the responsible processes must be inferred from the erosional data. Mean rates of downwearing are usually from essentially 0 up to a few millimeters per year, with occasional rates of a few centimeters per year corresponding to the removal of thin slivers of weathered rock (Porter et al. 2010).

Climate change

The most important effect of climate change is likely to be from the rising sea level, which will allow waves to break closer to the coast and expend more energy at the shore, and possibly in some areas from increased storminess. Although all rock coasts will be affected, the greatest increases in erosion rates, and particularly in rates of cliff recession, will be on soft rock coasts, which will impact human settlements and infrastructure. Higher precipitation, more extreme rainfall events, coral reef mortality, and decreasing sea ice will also promote cliff failures. Rising sea level, changes in wave regimes, and other climate changes will trigger local adjustments to wave refraction patterns, longshore transport paths, and the production of sediment, causing rapid changes to beaches on and adjacent to rocky coasts. Many other climate-induced changes will only become apparent over longer periods, including the modification of weathering processes and efficacies, and the longitudinal and attitudinal migration of marine organisms with corresponding changes to their bioerosional and bioprotectional activities.

Geology

The geology of the substrate, including the structure, lithology, and mineralogy of the rock, usually controls the general morphology of the coast. Geological factors help to determine the intensity and efficacy of the erosional mechanisms; the amount, mobility, and size of the debris at the foot of the cliff; the degree of surface irregularity in the nearshore zone, which influences rates of wave attenuation and, in the intertidal zone, the degree of ponding and abrasional efficacy; and whether a coast is entirely contemporary or inherited in part from interglacial stages when the sea was at a similar level to today’s.

Rock hardness is frequently measured in the field with a Schmidt rock test hammer. Shore platform gradient and elevation and the height of the cliff foot (the cliff/platform junction) generally increase with rock hardness or resistance. Nevertheless, erosion usually occurs along joints, bedding planes, and other structural weaknesses, and the resistance of a rock, which depends on the processes to which it is being subjected, may have little to do with its physical strength or hardness. Rock resistance is also affected by the strike and dip of alternating beds of sedimentary rock, which determine rates of wave attenuation, and whether the weaker strata are sheltered from incoming waves behind more resistant, upstanding beds.
Landforms

Coastal scenery is the product of a unique combination of factors including the gradient, height, and other morphological characteristics of the hinterland; wave environment and tidal range; geological structure and lithology; and climate and sea level history. Although these variables provide infinite variety to rock coasts, their general morphology may reflect the regional geotectonic setting. Coasts are generally oriented along convergent plate boundaries where oceanic plates are subducted beneath continental plates. These coasts are characterized by narrow continental shelves; high, steep hinterlands, often with staircases of elevated marine terraces; and an abundance of beach sediment from steep stream courses. A much greater variety of rocks can be exposed along plate-embedded or trailing-edge coasts (passive margins) which may run at high angles to the structural grain. This produces indented coasts with bays and headlands; hilly, plateau, or low hinterlands; and wide continental shelves. Plate tectonics can therefore provide a partial explanation for differences in coastal characteristics, but rock coasts with high cliffs are widely distributed along both convergent and trailing or plate-embedded coasts.

Coastal plan shape

Variations in rock lithology and structure, often associated with folds, faults, and joints, create coasts consisting of alternating headlands and embayments. Whereas differences in the type of rock along a coast are usually fairly obvious, variations in rock structure can be quite subtle, involving changes in joint density and orientation, bed thickness, and the strike and dip of the rocks. Bays can also develop in more homogeneous rocks because of Holocene submergence and the more rapid retreat of the lower cliffs around the mouths of streams than of the higher cliffs on the adjacent interfluvies. Differences in the type and amount of beach material along a coast, and consequently in the degree of abrasion or protection afforded to the cliff and shore platform, may also produce crenulated coasts without marked variations in the resistance of the rocks to erosion.

Caves, arches, and related forms

Small coves, narrow inlets, blowholes, caves, arches, and stacks can form in rocks with closely spaced weaknesses, which could include a joint or fault system, or a particularly weak stratum, although their best development is in rocks that are strong enough to support high, near-vertical slopes and the roofs of caves, tunnels, and arches. The sea can also occupy and erode subterranean karstic systems and the mouths of lava tubes in igneous rocks, producing long tunnels, caves, and inlets.

Cliffs

Marine erosion undercuts coastal slopes creating steep cliff faces, whereas subaerial mechanisms — including weathering; runoff; and creep, slides, and other mass movements — produce more gentle, convex slopes. Consequently, steep and gentle cliffs predominate where, respectively, marine and subaerial erosion are dominant. Cliffs that are steep near the foot and sloping near the top develop where both process suites are effective. Steep cliffs are predominant in exposed sites in the stormy mid-latitudes and more gently sloping, vegetation-covered cliffs in sheltered, often coral-protected, humid tropical regions, although the collapse of deep notches commonly produces steep cliffs in tropical limestones. Geological factors are at least as important as climate and wave regime in determining the relative efficacy of marine and
subaerial processes. For example, weak rocks in the upper part of cliffs promote weathering and runoff gullying and the formation of sloping surfaces, whereas strong rocks inhibit weathering and allow steep, wave-cut slopes to extend to higher elevations, often to the top of the cliff. Rock structure is also important. Very steep cliffs tend to develop in horizontally or vertically bedded rocks and sloping cliffs along the joint or bedding planes of, respectively, landward- or seaward-dipping rocks.

Cliffs often have composite profiles composed of two or more major slope elements. These profiles may simply reflect elevational differences in rock resistance as, for example, when weaker rocks in the middle portion of a cliff face have a lower slope gradient than stronger rocks in the lower and upper parts of the cliff. Other types of high, composite cliff in resistant rocks are the result of reactivated marine erosion during interglacial periods of high sea level, and cliff abandonment and frost and other cold-climate weathering and slope reduction during periods of low glacial sea level. They include hog’s back (beveled, slope over wall) cliffs which have a steep, wave-cut face at the bottom and a convex or straight seaward-facing slope above, and multistoried cliffs consisting of two or more steep faces separated by more gentle slopes. Composite cliffs have been reported most frequently from northwestern Europe, and on other ancient, resistant massifs and many volcanic islands in the Southern Hemisphere. Resistant basaltic and other islands often have plunging cliffs which stand in deep water without any shore platforms or beaches at their foot. These cliffs experienced little erosion as sea level rose during the Holocene because the unbroken waves were reflected from the cliff face, and the lack of sediment prevented abrasion and reduction in the depth of the water at the cliff foot.

Shore platforms

The term “shore platform” is used in preference to older genetic alternatives, including wave-cut or abrasion platform, to refer to horizontal to gently sloping, intertidal rock surfaces. Shore platforms usually terminate landward in a sea cliff, although they may extend under sand or pebble beaches or colluvial deposits where the hinterland is low.

Much of the traditional debate was concerned with the relative contributions of wave and weathering processes to the formation of predominantly sloping shore platforms around the North Atlantic and subhorizontal platforms, terminating abruptly seaward in a low tide cliff, in Australasia. It is generally accepted today that mechanical wave erosion and weathering play important and often supportive roles in platform development. The contributions of these process suites vary through time, with changes in such factors as climate, sea level, and platform morphology, and in space with changes in elevation and distance from the cliff foot. Surface topography plays an important role in determining the efficacy of wave quarrying, which is highest where waves impact seaward-facing scarps in horizontally bedded rocks or the upstanding, more resistant beds in steeply dipping strata. Without abrasive material, mechanical wave erosion is much less effective on shore platforms with fairly smooth, even surfaces, especially if they have very low gradients and low tide cliffs that cause the waves to break at their seaward edge (Stephenson and Kirk 2000). Subhorizontal shore platforms are often dominated by weathering today. These platforms in much of the Southern Hemisphere and in areas affected by glacio-isostatic movements of the land may have been cut by waves during a period of higher sea level in the Holocene, and subsequently leveled and lowered to their present elevations by weathering.
COASTAL EROSION PROCESSES AND LANDFORMS

Traditional explanations for the occurrence of sloping and subhorizontal platforms were based on differences in climate and wave regime, and consequently in the efficacy of mechanical wave erosion and weathering. There is strong evidence, however, to suggest that regional mean platform gradient is determined by the tidal range, although gradients may increase within regions with rock hardness and possibly with the size and mobility of any beach deposits. Tidal range influences mechanical wave erosion by determining the degree to which wave energy is expended at particular elevations within the intertidal zone; and weathering and biological erosion and protection by controlling the frequency and duration of the wetting and drying cycles.

Inheritance

Cohesive clay and other soft rock coasts erode fairly easily and are contemporary features related to modern morphogenic conditions. Conversely, shore platforms and other elements of hard rock coasts may have been inherited in part from interglacial stages when sea level and climate were similar to today. Hard rock coasts may also retain vestiges of former conditions during the Quaternary, such as scree slopes and raised beaches that developed when morphogenic conditions were quite different from today.

The occurrence of wide terraces on tectonically uplifted coasts shows that shore platforms can develop during single interglacial stages, and presumably therefore over similar periods of time in the Holocene. Although dating is difficult, it has been opined that some shore platforms in very resistant rocks are too wide to have developed in the few thousand years since the sea reached its present level. These platforms are often backed by ancient composite cliffs covered in places by glacial or periglacial deposits, and by elevated erosional ledges and raised beaches. There is convincing evidence that some shore platforms were formed incrementally in the Holocene and over one or more interglacial stages. They include dated platforms in southern Australia, and in northwestern Spain where they extend under ancient cemented beach deposits and dated alluvial and colluvial slope deposits from the early and much of the middle portion of the last glacial stage (Blanco Chao et al. 2003). Cosmogenic beryllium-10 dating suggests that the seaward portions of shore platforms in Korea are up to about 150,000 years old, and that the modern platform is cutting into its interglacial predecessor (Choi et al. 2012). Modeling also suggests that inheritance has been important in the development of shore platforms on tectonically stable coasts.

Modeling

It is difficult to determine the long-term development of rock coasts because they change very slowly and usually lack any datable sediment. Many workers have modeled these coasts but it is difficult to represent the myriad interacting geological and morphogenic factors that play a role in their development. Nevertheless, models have been used to study the development on hard rock coasts of shore platforms, erosional terraces, and submarine shelves on stable, uplifting, and subsiding landmasses, with sea level either constant or changing over periods ranging from a century or so up to a million years or more. Erosion is more rapid on cohesive clay and other soft rock coasts, and the threat to human life and property has generated a need for predictive erosional models with constant and rising lake and sea levels, usually operating over much shorter timescales than those for hard rock coasts. Less
modeling has been conducted on the plan shape of rock coasts by incorporating the effects of wave refraction and integrating erosional changes in the vertical and horizontal planes.

**SEE ALSO:** Coastal depositional processes and landforms; Coastal zones; Coasts; Climate change, concept of; Sea level rise

**References**


**Further reading**


Coastal zones

Randolph A. McBride
Christopher T. Seminack
George Mason University, USA

Definition

The coast is a dynamic environment because it involves the triple intersection of land, ocean, and air; thus it experiences terrestrial, oceanic, and atmospheric processes concurrently. Major sea level changes have caused the active coastal environment to migrate tens to hundreds of kilometers horizontally over geologic time. The most recent changes in sea level occurred during the Quaternary period (from 2.6 million years ago to the present) in response to the growth and decay of continental ice sheets (i.e., glaciations and deglaciations) driven by Milankovitch cycles. Consequently, global (eustatic) sea levels fluctuated up to 135 m vertically during intervals of climate cooling (i.e., glacial or sea level lowstands) and warming (i.e., interglacials or sea level highstands) during the Quaternary period, resulting in the coast migrating back and forth across the entire continental shelf repeatedly. In the light of the dynamic and migratory nature of the coast during the Quaternary, the specific spatial boundaries of the coastal zone are delineated in Figure 1 and follow the definitions of Inman and Nordstrom (1971), Inman and Brush (1973), and Masselink, Hughes, and Knight (2011).

Specifically, the upper and lower boundaries of the coastal zone correspond to the elevational range over which coastal processes have operated during the Quaternary period, and include the coastal plain (Quaternary deposits only), shoreface, continental shelf, and the waters that cover the shelf and shoreface, as well as waters filling coastal water bodies, such as estuaries, bays, lagoons, sounds, harbors, and inlets (Inman and Brush 1973; Masselink, Hughes, and Knight 2011). As such, the landward limit of the coastal zone includes the oldest Quaternary coastal plain deposits laid down during sea level highstands at or several meters above present-day levels, whereas the seaward limit extends out to the continental shelf break, typically lying between the 100 m and 200 m isobaths with an average depth of 135 m.

Primary types of coastal zones worldwide

Globally, the geology and morphology of continental margins are determined by the following factors: (i) plate tectonics; (ii) wind, waves, and wave-generated currents; (iii) tidal range and intensity of tide-generated currents; (iv) sediment supply to the coast and longshore sediment flux; (v) storm impacts (tropical and extratropical); and (vi) coastal climate. Based on a combination of these factors, Inman (1994) identified five primary types of coastal zones worldwide: (i) collision coasts; (ii) trailing-edge coasts; (iii) marginal sea coasts; (iv) cryogenic coasts; and (v) biogenic coasts.

Collision coasts face trenches (subduction zones) offshore and are predominantly erosional features. These coasts are tectonically active and are typically characterized by narrow continental
Coastal zones are characterized by three main regions: Quaternary coastal plain, shoreface, and continental shelf. The coastal plain is a low, net progradational plain with its seaward edge at the mean low-tide line along an oceanic shore, and its strata are either horizontal or gently dipping toward the ocean. The shoreface extends from the mean low-tide line to the depth of average wave base (depth varies depending on coastal setting but typically averages from 5 m to 10 m water depth), while the continental shelf occurs from average wave base to the shelf break. Source: Modified from Christian and Mazzilli (2007).

Collision or leading-edge coasts are common around the perimeter of the Pacific Ocean (i.e., Ring of Fire), where trenches and volcanoes dominate (e.g., Peru–Chile Trench, Cascadia Trench, Aleutian Trench, Kuril–Kamchatka Trench, Japan Trench, Mariana Trench, Tonga Trench), as well as in the eastern Indian Ocean (Java Trench), as documented by Inman and Nordstrom (1971). Trailing-edge coasts face mid-oceanic ridges (divergent boundaries) and are dominated by depositional features. These coasts are tectonically quiescent and are typically characterized by wide continental shelves; moderate wave energy; depositional features such as barrier islands, spits, and deltas; long rivers with large drainage basins and robust deltas (e.g., Amazon, Niger); and well-developed coastal plains. According to Inman and Nordstrom (1971), trailing-edge coasts are common around the perimeters of the Atlantic and Arctic oceans where few to no trenches are present, as well as around most of the perimeter of the Indian Ocean.

Marginal sea coasts face the backside (non-open ocean side) of volcanic island arcs (e.g., Aleutians, Kuril Islands, Japan, Philippines, New Zealand, and the Greater and Lesser Antilles) and are dominated by depositional features. Marginal sea coasts are fronted by shallow water bodies with more limited fetch than an open ocean, as observed in the Bering Sea, Sea of Okhotsk, Sea of Japan, East China Sea, South China Sea, Coral Sea, Tasman Sea, Gulf of Mexico, and Caribbean Sea. Marginal sea coasts are found along the...
entire eastern mainland shoreline of Asia (e.g., Russia, China, North Korea, South Korea, Vietnam, Cambodia, and Thailand), around most of the perimeter of the Gulf of Mexico, and the western Caribbean Sea shoreline of Central America (Inman and Nordstrom 1971). Marginal sea coasts can be similar to trailing-edge coasts in appearance. Their tectonically stable setting and general protection from open ocean waves and tidal processes afford fluvial processes the opportunity to dominate marine processes, thus enabling large fluvial deltas to develop (e.g., Yangtze, Yellow, Mekong, and Mississippi). Marginal sea coasts are variable but may be characterized by wide continental shelves and coastal plains (e.g., northern Gulf of Mexico coast).

The primary cryogenic coasts are found around the perimeter of the Arctic Ocean, specifically the northern shorelines of Alaska, Canada, Greenland, and Russia. Based on the tectonic coastal classification of Inman and Nordstrom (1971), these Arctic coasts are categorized as trailing-edge coasts because they face a mid-oceanic ridge (spreading center) located on the seafloor of the Arctic Ocean. These Arctic coasts are tectonically aseismic and are characterized by wide continental shelves, broad coastal plains built of fluvial and glacial deposits, low-profile depositional features (e.g., barrier islands, spits, and deltas), microtidal conditions, and extensive permafrost, including subsea permafrost. These Arctic coastlines experience two primary seasons: eight to nine months of winter when the sea ice expands to cover most open water and the coast is frozen solid, in contrast to a short summer when the sea ice retreats and open water returns, allowing normal coastal processes to operate ephemerally (i.e., wind-generated waves, longshore sediment transport). In winter, coastal processes are dominated by ice-push phenomena and other cryogenic processes (see Washburn 1979; Yershov 1998).

Biogenic coasts (coral reefs and associated carbonate environments) are found in tropical and subtropical latitudes, particularly between 30°N and 30°S. They are best developed across the Indian Ocean, and along the western margins of the Atlantic and Pacific Oceans, although biogenic coasts are generally excluded from the eastern margins of the latter two oceans by the upwelling of cold water (Woodroffe 2003, ch. 5). Globally, the most prolific reef development area is known as the Indo-Pacific province, in contrast to the less prolific area known as the Atlantic province. Coral reef growth and evolution are controlled by a complex interaction among plate tectonics, sea level changes, environmental conditions (i.e., water temperature, light, wave activity, water turbidity, salinity, nutrient levels, etc.), and biological activity, as discussed by Woodroffe (2003, ch. 5) and Masselink, Hughes, and Knight (2011, ch. 10).

History of definitions

The term “coastal zone” has been used extensively in scientific literature and, more recently, in legislation. Popular in the 1800s as a general expression, “coastal zone” lacked a scientific definition. Perhaps the first documented scientific definition of the coastal zone was based on Stephen (1934), who specified that the coastal zone extends seaward from the littoral zone to the neighborhood of the 40 m bathymetric line. “Coastal zone” is now an important term used in environmental legislation, as well as in the physical and biological sciences.

The US government has played a significant role in developing a political definition of the coastal zone through the passage of extensive legislation beginning in the late 1960s. During this time, the US government passed groundbreaking environmental legislation that set the
stage for present-day international policy. All environmental legislation in the United States ultimately stems from the National Environmental Policy Act (NEPA) of 1969. This piece of legislation forced the US government to review the environmental impact of all government and government-funded actions. NEPA represents the first step in creating a political definition of a coastal zone.

The Coastal Zone Management Act (CZMA) of 1972 provides federal funding for a state to meet a balance between development and the environment (Vernberg and Vernberg 2001). This act is the only piece of US legislation that addresses the definition of a coastal zone; the CZMA permits each state to develop its own definition because of each state’s unique coastal setting. However, the CZMA does supply a blanket guideline of a coastal zone as follows:

The term “coastal zone” means the coastal waters (including the lands therein and thereunder) and the adjacent shorelands (including the waters therein and thereunder), strongly influenced by each other and in proximity to the shorelines of the several coastal states, and includes islands, transitional and intertidal areas, salt marshes, wetlands, and beaches. The zone extends, in Great Lakes waters, to the international boundary between the United States and Canada and, in other areas, seaward to the outer limit of State title and ownership under the Submerged Lands Act (43 U.S.C. 1301 et seq.), the Act of March 2, 1917, (48 U.S.C. 749), the Covenant to Establish a Commonwealth of the Northern Mariana Islands in Political Union with the United States of America, as approved by the Act of March 24, 1976 (48 U.S.C. 1801 et seq.), or section 1 of the Act of November 20, 1963 (48 U.S.C. 1705), as applicable. The zone extends inland from the shorelines only to the extent necessary to control shorelands, the uses of which have a direct and significant impact on the coastal waters, and to control those geographical areas which are likely to be affected by or vulnerable to sea level rise. Excluded from the coastal zone are lands the use of which is by law subject solely to the discretion of or which is held in trust by the Federal Government, its officers or agents. (Coastal Zone Management Act 1972, 4)

The CZMA contains special programs within its legislation. The National Estuarine Research Reserve program created a system of natural laboratories in order to study ecosystems (Vernberg and Vernberg 2001). Another program created by the CZMA is the Special Area Management Plan. This program was created to address coastal environmental issues which affect multiple jurisdictions (e.g., federal, state, and local governments).

Many US legislative works involve the definition of a coastal zone or its subhabitats. The US Army Corps of Engineers regulate activities in navigable waters under the Rivers and Harbors Act of 1899. They also spearhead construction projects along the coast for hard structures (i.e., jetties, seawalls, etc.) and soft structures (i.e., beach replenishment; Vernberg and Vernberg 2001). The Clean Water Act of 1972 regulates discharges into water bodies, both terrestrial and coastal (Vernberg and Vernberg 2001). The Clean Air Act of 1972 aims to “Preserve, protect, and enhance the air quality in national parks, national wilderness areas, national monuments, national seashores, and other areas of special national or regional natural, recreational, scenic, or historic value” (section 160). This act primarily affects development along the coast (i.e., oil and gas industry; Vernberg and Vernberg 2001). The National Flood Insurance Act offers subsidized insurance for coastal flooding damage to property owners. Federal funding to recover from natural disasters is provided by the Federal Emergency Management Agency (FEMA). The agency will reimburse (up to 75% in most cases) public entities such as publicly renourished beaches, water and sewer systems, roads, bridges, recreational
<table>
<thead>
<tr>
<th>Organization/author(s)</th>
<th>Definition of coastal zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stephen (1934)</td>
<td>The coastal zone extends from the littoral zone to the neighborhood of the 40 m (bathymetric) line.</td>
</tr>
<tr>
<td>Inman and Nordstrom (1971)</td>
<td>“The coastal zone is defined in terms of the large-scale features of coasts, including first-order features, such as the coastal plain, the continental shelf, and the waters that cover the shelf; and second-order features, such as large bays, estuaries, lagoons, coastal dune fields, river estuaries, and deltas.”</td>
</tr>
<tr>
<td>Ramsar Convention (1971)</td>
<td>“Areas of marsh, fen, peatland, or water, whether natural or artificial, permanent or temporary, with water that is static or flowing, fresh, brackish or salt, including areas of marine water the depth of which at low tide does not exceed six meters. [This] may incorporate adjacent riparian and coastal zones, islands or bodies of marine water deeper than six meters at low tide lying within the wetland.”*</td>
</tr>
<tr>
<td>Inman and Brush (1973)</td>
<td>“The coastal zone may be defined in terms of the large-scale tectonic and erosional-depositional features which have lengths along the coastline of the order of 1000 kilometers and widths extending from the coastal plain out into the water of an order of magnitude less. The coastal zone is composed of the coastal plain, the continental shelf, and the waters that cover the shelf; it also includes other major features, such as large bays, estuaries, lagoons, coastal dune fields, river estuaries, and deltas.”</td>
</tr>
<tr>
<td>Perillo (1995)</td>
<td>“An estuary is a semi-enclosed coastal body of water that extends [landward] to the effective limit of tidal influence, within which sea water entering from one or more free connections with the open sea, or any other saline coastal body of water, is significantly diluted with fresh water derived from land drainage, and can sustain euryhaline biological species from either part or the whole of their life cycle.”</td>
</tr>
<tr>
<td>Food and Agriculture Organization of the United Nations (1995; 1998)</td>
<td>“All of the coastal and upland areas, the uses of which can affect coastal waters and the resources therein, and extends seaward to include that part of the coastal ocean that can affect the land of the coastal zone. This may also include the Exclusive Economic Zone.”</td>
</tr>
<tr>
<td>The World Conservation Union (IUCN) and National Oceanic and Atmospheric Administration (Sherman and Duda 1999)</td>
<td>“Regions of ocean space encompassing coastal areas from river basins and estuaries to the seaward boundary of continental shelves and the outer margins of the major current systems.”</td>
</tr>
</tbody>
</table>

*(Continued opposite)*
**COASTAL ZONES**

**Table 1  Continued**

<table>
<thead>
<tr>
<th>Organization/author(s)</th>
<th>Definition of coastal zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vernberg and Vernberg (2001)</td>
<td>“In general, the coastal zone encompasses both the neighboring uplands and adjacent salt waters that are mutually influenced by the interactive complex of various ecological processes (natural and human-influenced) occurring in each region. Various problems are associated with establishing both the landward and seaward boundary lines. Scientifically, variation in such features as habitat, soil type, and land formation complicate the establishment of a boundary line of the coastal zone. Politically, a legislative and regulatory definition of the coastal zone (which is not necessarily based on sound environmental grounds) has led to conflict among different levels of governmental and regulatory agencies.”</td>
</tr>
<tr>
<td>Millennium Ecosystem Assessment (2003)</td>
<td>“Interface between ocean and land, extending seawards to about the middle of the continental shelf and inland to include all areas strongly influenced by the proximity to the ocean (between 50 m below mean sea level and 50 m above the high tide level or extending landward to a distance 100 km from the shore.”</td>
</tr>
<tr>
<td>UNESCO (2003a)</td>
<td>“The coastal stream and river runoff areas for lagoons, bays, and estuaries.”</td>
</tr>
<tr>
<td>The Coastal Ocean Observations Module of the Global Ocean Observing System (UNESCO 2003b)</td>
<td>“The head of the tidal waters to the outer limits of the Exclusive Economic Zone (EEZ).”</td>
</tr>
<tr>
<td>International Geosphere–Biosphere Programme (2005)</td>
<td>International waters and their drainage basins, including habitats comprising marine, coastal and freshwater areas, and surface waters as well as groundwater.</td>
</tr>
<tr>
<td>Global International Waters Assessment (2006)</td>
<td>Does not outline a specific definition of the coastal zone; rather it introduces an adaptable approach, stating that the geographical structure has to be flexible and based on natural, political, and institutional realities. Existing geographical and program structure should be used where appropriate. Habitats include coastal wetlands, estuaries and deltas, mangrove, coastal reef and seagrass beds.</td>
</tr>
<tr>
<td>Masselink, Hughes, and Knight (2011)</td>
<td>“The coastal zone extends from the landward edge of the Quaternary coastal plain to the continental shelf break and includes the Quaternary coastal plain, shoreface, and continental shelf.”</td>
</tr>
</tbody>
</table>

* The Ramsar Convention utilized a definition of “wetlands” rather than “coastal zone.”


facilities, and buildings (Vernberg and Vernberg 2001). Significant marine areas of historical, educational, research, ecological, recreational, or aesthetic value are protected under the National Marine Sanctuaries Act (Vernberg and Vernberg 2001). The National Coastal Barrier Resources Act put in place legislation to reduce the development of undeveloped barrier islands in hopes of reducing storm-related damage to people and property, as well as preserving coastal
COASTAL ZONES

habitats (Vernberg and Vernberg 2001). Finally, the Endangered Species Act of 1973 is one of the most powerful pieces of US legislation. The goal of this act is to conserve endangered and threatened species. Many of the species listed under this law reside in the coastal zone (Vernberg and Vernberg 2001).

Political definitions of the coastal zone extend to the international stage as well. Table 1 provides brief summaries of coastal zone definitions from various international bodies and conferences, as well as authors of scientific articles.

Conclusion

The future viability of coastal zones worldwide will depend on numerous factors and impacts. The most important factors and impacts include global climate change, rising sea levels, human population growth and disturbance in coastal areas, pollution, utilization of coastal resources (e.g., groundwater, aggregate, and energy extraction, e.g., oil and gas), and the implementation of sustainable coastal management (SCM) practices (see Vernberg and Vernberg 2001, ch. 9; Masselink, Hughes, and Knight 2011, ch. 12).

SEE ALSO: Climate change adaptation and social transformation; Climate change and land ice; Coastal depositional processes and landforms; Coastal erosion processes and landforms; Coasts; Cryosphere: remote sensing; Cryosphere studies: history; Ecosystem services; Environmental management; Eolian erosional processes and landforms; Geomorphic systems; Global climate change; Holocene; Ice sheets; Intergovernmental Panel on Climate Change (IPCC); Land-use/cover change and climate; Late Cenozoic polar glaciations; Oceans and climate; Oceans and seas: human geography; Oceans and seas: physical geography; Periglacial processes and landforms; Permafrost: definition and extent; Quaternary geomorphology and landscapes; Quaternary glaciations; Sea level rise; Weather, extreme

References


The coast is where the land meets the sea. Coasts attract people because of their beauty, abundant supply of natural resources, and their special appeal as a place to live, harvest resources, and recreate. Nearly half the global population and associated development is already concentrated at the coast, and this is likely to intensify in the future. Consequently, coasts are crucial for human wellbeing but they are subject to mounting pressures. Unsustainable practices are widespread and worsening in many localities in this narrow strip. Coasts are also impacted by distant activities, such as poor farming practices far inland or pollution at sea. Coastal ecosystems are now among the most threatened on the planet, jeopardizing the livelihoods, wellbeing, and future prospects of coastal communities, nations, and humanity. Therefore, one of the defining challenges for the twenty-first century is to overcome the failures of sustainability, both distant and local, that are concentrated at the coast. Increasing attention is being focused on how to manage the coast to promote sustainable coastal development. Coastal communities have managed coastal resources for millennia. Contemporary coastal management efforts involving coastal communities and their governing authorities developed rapidly from the early 1970s and are now underway in most coastal localities. There are notable success stories but they tend to be small-scale and coastal sustainability writ large is elusive despite dedicated efforts and a wide range of progressive coastal management practices. To compound matters, global environmental change is making many coasts less suitable for human habitation. Climate change, in particular, is compounding coastal hazard risks especially in areas already exposed to the impacts of coastal erosion and flooding. Future prospects are dismal unless the root causes and drivers of unsustainable practices are addressed and effective measures institutionalized that enable coastal communities to build adaptive capacity and resilience. Transformative change is needed to govern the coast sustainably. Such systemic change challenges the status quo and is likely to be resisted. Issues to address are how sustainability failures can be transformed into successes and what coastal management “best practices” need to be fostered to bring about sustainable coastal development. A starting point for addressing these issues is to understand the distinctive characteristics of the coast.

Defining the coast

The coast can be defined as the interface between land and sea. It is a relatively narrow transitional zone characterized by direct interactions between purely terrestrial and marine environments with indirect influences extending far inland and out to sea. For example, activities in remote catchments, such as damming of rivers or farming practices that cause soil erosion, nutrient enrichment, or other polluting activities, cause environmental impacts that negatively affect the health, productivity, and diversity of coastal waterways like estuaries and associated ecosystems. Activities in the marine
realm, such as an oil spill, can have significant impacts on coastal ecosystems and the people whose livelihoods depend on them.

Land–sea interactions involve multiple biogeochemical processes and gradients and a myriad of human–nature and human–human interactions, which vary greatly in geographic and temporal scale and duration (Crossland et al. 2005; Figure 1). Consequently, coasts are highly interconnected, diverse, and dynamic environments. These intensive land–sea interactions give rise to a variety of distinctive coastal ecosystems and human communities that are interdependent and constitute a coupled coastal socioecological system made up of the coastline or seashore, coastlands, and coastal waters.

The coastline is the area between the low- and high-water marks but extends land- and seawards depending on seasonal influences and longer-term changes in coastal biogeochemical processes. Coastlands are areas above the high-water mark extending inland that have an influence on or are influenced by the sea. The inland extent depends on the reach of these direct and indirect land–sea interactions. Rivers are a pivotal connector between inland areas and the marine realm and channel the influence and impacts of catchment activities that, for example, compound and abstract water, pollute rivers, and add or trap sediments. Coastal waters extend from the seashore into the ocean as far as terrestrial processes and activities have influence and impact. River discharge, for example, can have a powerful influence far out to sea. Land-based sources account for most marine pollution and include persistent, bioaccumulative, and toxic substances that can be magnified up the marine food chain and pose a severe threat to human health and marine ecosystems. Marine activities can negatively impact coastal ecosystems through, for example, pollution from passing ships and plastic debris that is carried via ocean currents far from their source to distant coasts.

The coastal socioecological system can thus be thought of as a meeting place – of land and sea, of people and cultures, and a plethora of activities from fishing to coastal development, recreation, farming, mining, and nature conservation (Figure 2). It is also a place where different societal institutions and administrative jurisdictions intersect. For example, local
COASTS

authorities typically use land-use planning provisions to regulate development and the use of coastal resources from the high-water mark landwards – typically under a regime of private property rights. In the marine realm, from the high-water mark seawards up to the boundary of the exclusive economic zone, by contrast, the allocation and exercise of user rights and responsibilities are shaped by national (e.g., fisheries legislation) and international (e.g., the UN Convention on the Law of the Sea) provisions that regulate user practices under a regime of state ownership and/or stewardship in the public interest. The terrestrial and marine legislative and property rights regimes are bifurcated by the high-water mark. They need to be aligned, coordinated, and integrated because the coastal socioecological system needs to be managed holistically to realize the system-wide benefits that coastal ecosystems provide to coastal communities, nations, and the global community at large.

Over time various definitions of the coastal zone have been used, including fixed and variable distance (and elevation/depth) options; definitions according to particular coastal issues, interests, activities, or uses; and definitions that combine elements of the foregoing. No one boundary definition for the coastal zone is suitable for all purposes because of the variable relevance and extent of direct and indirect land–sea interactions. Defining appropriate boundaries for the coastal zone therefore needs to be guided by the purpose at hand. For the Millennium Ecosystem Assessment, Agardy et al. (2005) defined the boundaries of the coastal zone to include land-based influences up to a maximum of 100 km from the seashore or 50 m elevation (whichever is closest to the sea), and the seaward extent as 50 m depth in the marine environment. On this basis, the coast comprises less than 5% of the terrestrial surface area of the Earth. Exploring further the “geography of the coast” reveals more about the distinctive characteristics of the coast and the particular challenges and opportunities it presents for coastal management and governance.

The geography of the coast

Physical geography of the coast

Coasts are characterized by intensive, often high–energy interactions, significant nutrient and material fluxes, and incessant change – from short-term changes driven by winds, waves, currents, and tides to changes across seasons, decades, and long-term climatological and biogeochemical changes (McLusky and Wolanski 2012). Coasts encompass an incredible range of features and habitats from cold, arid coasts like those fringing Antarctica to subtropical islands like those in the Pacific and a panoply of ecosystems – including estuaries, marshes, salt ponds, coastal aquifers, and lagoons; mangrove forests; intertidal habitats, deltas, beaches, and dunes; rocky shores; coral reefs and atolls; seagrass beds or meadows; kelp forests; other benthic communities, like rock and shell reefs, mudflats, coastal seamounts, and rises; near-shore islands; and semi-enclosed seas (Figure 3). Coastal ecosystems are among the most biologically productive systems on Earth (Agardy et al. 2005).

The dynamism, diversity, and productivity of coastal ecosystems stem from complex interacting biogeochemical processes and fluxes. These coastal ecosystems are resilient to change but they are finite and have thresholds beyond which degradation and even collapse occurs. These ecosystems perform a range of ecological functions, such as the role played by mangroves and estuaries in providing a safe and food–rich habitat for the growth and development of juvenile fish, crab, and prawn species. Coastal ecosystems and
the ecological functions they perform yield a wide range of ecosystem services that benefit people, such as subsistence, recreational, and commercial fishing. The physical and human geographies of the coastal socioecological system are thus interwoven.

**Human geography of the coast**

For millennia, people have flocked to the coast because of its beauty, diversity, productivity, and cultural and spiritual significance. Albeit a narrow strip comprising a fraction of the Earth’s habitable surface, about 40% of the world’s population live within 100 km of the seashore (Agardy et al. 2005; Martínez et al. 2007; Small and Nicholls 2003). Most people are concentrated in densely populated rural areas, and small and medium-sized cities, chiefly in the vicinity of estuaries, mangroves, and coral reefs. Average coastal population density is three times greater than the global average. More than 50% of coastal nations have 80–100% of their total population in this narrow strip. Wealth and wellbeing is concentrated at the coast with per capita income typically four times higher and lower infant...
mortality and higher life-expectancy than inland areas.

Many people live in coastal cities and megacities (Pelling and Blackburn 2013): 16 of the world’s 23 megacities with populations over 10 million people occur here. Large-scale urbanization is causing widespread transformation of coastal ecosystems. There is especially rapid urban sprawl, fragmentation, and agglomeration in less wealthy countries. Coastal cities play a crucial role in the search for sustainability pathways because they are focal points of consumption and social reproduction and are pivotal centers of global trade, innovation, and geopolitical leadership (Figure 4).

The coast is the locus of global economic and human development with immense aesthetic, cultural, spiritual, and educational value and an estimated 61% of global total gross national product sourced within 100 km of the seashore (Agardy et al. 2005; Patterson and Glavovic 2008). Coastal ecosystems provide a wealth
of provisioning services (e.g., fish for food, raw materials for construction and boat building, fresh water, and medicinal resources), regulating services (e.g., carbon sequestration and storage, local climate and air quality, moderation of extreme events, wastewater treatment, erosion prevention and soil fertility maintenance, pollination and biological control), habitat or supporting services (e.g., habitat for coastal species and maintenance of genetic diversity), and cultural services (e.g., recreation and mental and physical health, tourism, aesthetic appreciation and inspiration for culture, art, design, spiritual reflection, and sense of place). Coasts contribute proportionately more ecosystem services than most other systems, including those that cover vastly larger areas. A preliminary estimate by Costanza et al. (1997) indicated that the total value of coastal ecosystem services is about 43% of the total value of global ecosystem services. Estuaries and marshes are among the most fertile coastal environments and provide the widest range of ecosystem services. For example, if the natural capital of the Mississippi river delta were treated as an economic asset, it would provide annual benefits of US$13–47 billion and have a minimum asset value of US$330 billion–1.3 trillion (Batker et al. 2010).

Sustaining these benefits is already imperiled and the prognosis for the future is dismal (Agardy et al. 2005; Moser, Williams, and Boesch 2012). In recent decades, there has been a rapid increase in coastal population, urban sprawl, and development intensification, especially in parts of Africa and central Asia, driven by, among other things, migration to the coast, high fertility rates, and, in some regions, high visitation by tourists and those seeking recreation (Figure 5). Coasts are consequently undergoing more rapid and pervasive change due to human pressures and impacts than at any other time in human history despite long-standing pressures in some localities. These changes include direct physical (e.g., infilling of wetlands for urban development) and biological (e.g., overexploited fisheries) transformations. In addition, there are indirect changes due to, among other things, changes in water flow in rivers and sediment patterns that lead to altered deposition and erosion dynamics at the coast which are exacerbated by chronic degradation and chemical transformations due to land-based (e.g., eutrophication, and bioaccumulated persistent toxins) and marine-based pollution (e.g., plastic disposal) as well as the proliferation and spread of invasive non-native species.

The coastal population and associated drivers of change are not uniformly distributed. People are concentrated in close proximity to the most productive coastal ecosystems (Agardy et al. 2005) – over 70% of the coastal population lives within 50 km of estuaries. In tropical coastal regions, many communities are located within reach of mangroves and coral reefs (58% of the world’s coral reefs are within 50 km of urban centers of 100,000 people or more and 64% of mangrove forests and 62% of major estuaries are near such centers).

Furthermore, access to and distribution of ecosystem service benefits, and bearing the consequences of degraded or destroyed coastal ecosystems are also differentially distributed with some groups and communities faring much worse than others largely because of their sociopolitical and economic marginalization, for example, small-scale fishers and traditional communities. Gross inequity is common at the coast with marked economic disparities within and between rural and urban coastal areas; and poverty and inequity is pronounced and increasingly concentrated in urban coastal areas (Brown et al. 2008; Pelling and Blackburn 2013). Brown et al. (2008) estimate that there are over 250 million coastal poor, with 80% living in just 15 countries – 27% in India, 13% in Indonesia,
and 9% in Bangladesh. The coastal poor receive a disproportionately small share of benefits from coastal ecosystems and have had minimal impact on overall changes in coastal ecosystem services. However, in some locations, unsustainable practices by those with limited livelihood alternatives have caused major coastal degradation. There is a complex interrelationship between poverty, ecosystem health and service provision, and livelihood sustainability. Coastal poverty and inequity can fuel conflict and catalyze a vicious spiral of ecologically and socially unsustainable practices that increase vulnerability and risk and undermine already fragile livelihoods and future prospects, especially in the face of climate change.

Climate change is compounding and exacerbating coastal hazard risks in localities already prone to inundation, coastal erosion, and saltwater intrusion into groundwater, especially in areas with vulnerable populations (Glavovic et al. 2015; Nicholls et al. 2007; Wong et al. 2014). Coastal systems are particularly sensitive to increases in sea level, ocean temperature, and ocean acidity. Low-lying coastal areas will be exposed to erosion, flooding, and submergence due to inevitable sea level rise, albeit variable in different localities and regions. Coral reefs are

Figure 5  Qingdao, a major seaport, naval base, and industrial hub in Shandong Province, China. Photos by Bruce Glavovic.
the most vulnerable coastal-marine ecosystem to increasing ocean temperature and acidity and have little prospect for adaptation with inevitable coral bleaching and other changes, including altered ranges of coastal-marine species. Climate change will cause other coastal changes but they are difficult to distinguish from human-related drivers of unsustainable practices (e.g., transformation of coastal habitats and ecosystems through land-use changes, development expansion and intensification, resource exploitation, and pollution). Climate change will exacerbate socioeconomic and infrastructure impacts due to:

- unsustainable practices such as loss of coastal property and land;
- loss of and damage to assets, including safe housing, public services, and infrastructure;
- negative impacts on coast-dependent livelihoods and human health due to loss of coastal resources, food insecurity, exposure to extreme events, heat stress, vector- and water-borne diseases, and diminished access to potable water;
- increasing costs of insurance (and rising un-insurability) and protective works and diversion of resources for adaptation;
• increasing mortality
• unforeseen socioeconomic disruptions.

Social institutions, cultures, and ways of life will also be adversely affected by increasing uncertainty, dynamism, and unprecedented events that disrupt community routines and lifestyles (Figure 6). Extreme events may become more frequent and unpredictable. Displacement and migration of vulnerable people may become more commonplace. Some coast-dependent cultures and ways of life (e.g., those dependent on coral reefs) may be imperiled. There are likely to be adverse impacts on cultural identity, national prestige, and sovereignty (e.g., as a result of submergence of low-lying small island states like Kiribati). Sociopolitical instability, unrest, and conflict may escalate in localities adversely impacted by climate change. Climate change will, however, also open up new opportunities for some groups and communities. Prospects for realizing these opportunities and adapting to change vary significantly within and between developed and developing coastal nations, with low-lying regions (e.g., Mississippi delta), developing nations (e.g., Bangladesh and Vietnam), and small island states (e.g., Kiribati) expected to experience very high impacts and related loss and damage as well as adaptation costs.

It is unequivocal that coastal hazard risk is escalating due to rapid coastal population growth, urbanization and development intensification, and climate change (Wong et al. 2014). By 2100, without effective adaptation, hundreds of millions of people will be impacted by coastal flooding and displaced due to land loss, mostly in east, southeast, and south Asia. Many coastal cities face escalating flood risk due to population growth and urbanization compounded by subsidence and climate change. According to Hallegatte et al. (2013), average global flood losses in the 136 largest coastal cities were about US$6 billion per year in 2005 and are expected to increase to US$52 billion per year by 2050 based on projected socioeconomic changes alone. Taking into account subsidence and climate change, existing protective measures will need to be upgraded to avoid unacceptable losses of US$1 trillion or more per year.

Many coastal communities, nations, and regions are vulnerable to climate variability and shocks that will compound the impact of unsustainable practices. Understanding and addressing this predicament is, however, difficult because of the complex interrelationships between coastal community wellbeing, resource and urban development trajectories, sustainability, and climate and global environmental change. Many vulnerable coastal communities will experience additional stress, existing problems will intensify, new problems will emerge, and risk will escalate – especially for those who are poor, and socially and politically marginalized, and thus least likely to benefit from new opportunities created by a changing climate. Climate change therefore poses a formidable challenge, especially for those already adversely impacted by inequitable and environmentally unsustainable coastal development. It raises vexing questions about how to mobilize, integrate, and apply scientific and local knowledge in practical ways to promote coastal resilience and sustainability (Figure 7).

Coastal challenges and opportunities in the Anthropocene

Post-industrial human activities have caused environmental impacts that may transgress critical planetary boundaries and threaten human wellbeing and survival, leading Nobel laureate Paul Crutzen (2002) among others to say that we now live in the Age of the Anthropocene. The coast is the frontline of the sustainability crisis
arising from the dominant influence humans now have on the Earth’s biogeochemical and climatological processes (Glavovic 2013a).

Humanity faces a “quadruple coastal squeeze” in the Anthropocene: (i) population growth, resource exploitation, urban sprawl, and development intensification; (ii) coastal ecosystem degradation, conversion and loss; (iii) rising CO$_2$, climate change, and alteration of coastal-marine biogeochemistry and ecosystems; and (iv) ecosystem thresholds and tipping points being transgressed, resulting in rapid and irreversible changes in the coastal socioecological system and societal responses. Consequently, we face a dilemma at the coast because as the coastal population grows there is increasing dependence on coastal ecosystem services, but unsustainable practices are pernicious, pervasive, and intensifying, compounded by global environmental change and climate change in particular, accelerating coastal hazard risk and making the coast less suitable for human habitation (Agardy et al. 2005; Moser, Williams, and Boesch 2012; Glavovic 2013a; Glavovic et al. 2015).

There are no ready-made “fixes” for the challenges facing coastal communities in the Anthropocene. Communities will need to chart their own pathways in the face of the complexity of the coast, including inevitable change, systemic uncertainty, and surprise; nonlinear and emergent interactions with intricate feedback loops; and real socioecological limits and barriers. Opportunities to chart pathways that foster adaptive capacity, resilience, and sustainability
are constrained by prevailing institutional structures and processes that are ill-equipped to forge timely solutions and which incentivize unsustainable coastal development. Business as usual is therefore untenable. Incremental stepwise improvements may alleviate some impacts and reduce some risk, but there is a compelling and urgent need to comprehend and address the underlying drivers and root causes of unsustainable coastal development, climate change complacency, institutional inertia, and maladaptive policies and practices. Ultimately, transformational change is necessary to chart resilient and sustainable pathways at the coast.

Science does not offer a panacea, but it is foundational for understanding the nature of the coastal sustainability crisis. The coastal science–policy–practice interface needs to be bridged and research on priority issues carried out through co-design and coproduction of knowledge involving scientists and the policymakers, practitioners, and local coastal stakeholders and community members who play vital roles in real-world coastal management (Bremer and Glavovic 2013). Inevitably, potentially viable solutions will be contested because there are multiple legitimate views shaped by divergent values, perceptions, and interests about how to address coastal problems. Moreover, the capacity or “power” to chart new pathways is not spread evenly within and between coastal communities and nations. Coastal communities need to be enabled to reflect critically on the challenges and opportunities they face, learn from experience (while recognizing the limits of relying on past practices to confront novel challenges), and envision, investigate, and explore new pathways to reverse localized and distant unsustainable practices, reduce risk, adapt to change, and foster resilience. Innovations in coastal management offer the best prospect for charting such pathways.

From coastal zone management to coastal governance for sustainability

Five broad approaches to coastal management can be distinguished: traditional community-based coastal management; ad hoc sector-based coastal management; coastal zone management; integrated coastal management (ICM); and coastal governance.

Traditional community-based coastal management

For many centuries, coastal communities and cultures developed and implemented management regimes to benefit from and sustain a variety of coastal and marine resources and fisheries in particular. Some Pacific Island cultures, for example, used measures such as limited entry, closed seasons, closed areas, size limits, and even, on occasion, gear restrictions enforced through cultural taboos, incentives, and sanctions, to manage these resources. Many of these practices have been eroded by globalization and Westernization with the spread of cash economies, export markets, new technologies, and other socioeconomic, geopolitical, and cultural influences (Johannes 1978). However, there has been a resurgence of interest in and application of traditional practices over the last three to four decades in Oceania (Johannes 2002), and different forms of devolved co-management, involving traditional communities, governing authorities, scientists, and nongovernmental organizations, have been adopted around the world (Figure 8).

Ad hoc sector-based coastal management

Up to the early 1970s, in developed countries in particular, coastal management was typically carried out in an ad hoc and compartmentalized
manner with provisions tailored to address sector-based interests and issues, for example, coastal development including housing and public infrastructure, nature conservation, fishing, farming, and recreation. Little consideration was given to the impact of ad hoc sector-based decisions on the coastal socioecological system as a whole. Public engagement was limited, environmental concerns were given minimal attention, and management efforts were largely reactive. The shortcomings of this approach became increasingly apparent with growing environmental consciousness.

Coastal zone management

The sea and seashore have long been considered a public asset and have enjoyed special legal status since pre-Roman times as an asset that should not be owned but enjoyed by citizens or held as public property vested in the state as custodian on behalf of the public. Such provisions were, however, insufficient to ensure effective management of the coastal zone as a whole. One of the earliest nationwide comprehensive efforts to manage the coastal zone was instituted with the passage of the US Coastal Zone Management Act in 1972. This initiative recognized the compelling case for dedicated government attention to be focused on the coast through formulation of state-wide integrated coastal zone management plans that make explicit the need to bridge different spheres of government responsibility, enable integration across sectors and activities, and involve the public in planning and decision-making. This enabling legislation, which includes technical and financial supporting measures and a federal consistency provision that requires federal agencies to comply with approved state plans, incentivized states to develop and implement plans that support federal objectives and standards for coastal zone management.

Focused attention on environmental concerns and proactive efforts to manage the coastal zone better grew rapidly from the early 1970s, and by the time the UN Conference on Environment and Development (UNCED) took place in Rio de Janeiro in 1992, many nations had embarked on initiatives to institutionalize dedicated coastal zone management regimes. One of the difficulties encountered was to define the boundaries of the coastal zone in a way that enabled effective management of the coastal socioecological system as a whole, taking into account both direct and indirect drivers of change. As explained above, defining a single static set of boundaries for the coastal zone is problematic because of the variable extent of local and distant influences on the coastal socioecological system. The need to shift the focus from “the zone” to “integration” came to the fore at UNCED and the notion of ICM gained traction.

ICM

A dedicated chapter on ocean and coastal areas, with an emphasis on integration as the primary
means to foster sustainable coastal development, was included in Agenda 21 (ch. 17) – the action plan produced by UNCED. This emphasis was reinforced in Johannesburg at the 2002 World Summit on Sustainable Development and in many other international and regional initiatives (UNEP Regional Seas Programme, and EU initiatives on ICM and Marine Spatial Planning to complement framework directives on water and fisheries, among others) as well as national and more localized ICM practices (e.g., the passage of South Africa’s ICM Act in 2008).

Defining ICM

According to GESAMP (1996), ICM can be defined as a continuous and dynamic process that enables government, coastal communities, scientists, and coastal managers to work together to formulate and implement an integrated plan that sustains coastal ecosystems and reconciles sector-based, private, and public interests in the coast. The goal is to improve the quality of life of coastal communities who depend on healthy, productive, and diverse coastal ecosystems. Achieving this goal requires an integrated management regime that coastal communities consider legitimate, equitable, transparent, and adaptable in the face of change. In short, the goal of ICM is to promote sustainable coastal development.

Why ICM?

First, the coast is an interconnected socioecological system that presents distinctive challenges and opportunities compared to either the marine or terrestrial realm per se. Management practices cannot simply be transferred from the marine or terrestrial realm and applied at the coast. ICM needs to be founded on tailored and hybrid approaches that take into account the intersection of land and sea influences, processes, interests, activities, and governance regimes. Bridging these realms requires integration across a number of domains (see Table 1).

Second, the coast is a public asset that requires dedicated and integrated management to maintain and secure it as a public heritage, for example, by providing public access and right of way over private property to the seashore and sea and reconciling public and private as well as intergenerational interests.

Third, ad hoc sector- and interest-based management efforts result in a “tyranny of small decisions” that might be individually rational but fail to realize system-wide benefits and undermine the healthy functioning of the coastal socioecological system. The whole is greater than the sum of the parts. ICM is necessary to address the crucial issue of interconnected scales of interaction at the coast that extend from the local to global level.

By way of illustration, constructing a causeway that impedes river flow to the sea might be compelling to minimize road construction costs. But estuarine functions (e.g., as a nursery for fisheries) may then be disrupted, diminishing ecosystem services (e.g., lower fish catches) with dire consequences for those whose livelihoods are dependent on such services. Choices made that result in incremental degradation or irreversible impacts might yield short-term benefits for some stakeholders but cause long-term costs that are borne by others and the wider community. The environmental impacts (including ecological, social, cultural, and economic impacts), social costs and benefits, and risks associated with alternative sector- and interest-based decisions need to be considered in the context of the coastal socioecological system as a whole. Hence the ICM imperative.
<table>
<thead>
<tr>
<th>Table 1</th>
<th>Dimensions of integration.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographic/spatial integration</td>
<td>Connections extend across the land–sea interface, landwards and seawards, and up and down the coast. Water is a major connector at the coast. Geographic interconnections are not only physical but include important cultural, social, economic, trade, geopolitical, and institutional linkages that can reach worldwide.</td>
</tr>
<tr>
<td>Temporal integration</td>
<td>Short-term decisions can lead to impacts with cumulative (a seemingly trivial choice assumes greater import in the face of multiple such choices over time) and synergistic (the impact of two interacting factors that produces an impact greater than the sum of their individual impacts) effects with significant long-term consequences and hence the need to take into account the interaction between short- and long-term consequences and intergenerational interests and equity.</td>
</tr>
<tr>
<td>Sectoral integration</td>
<td>Integration between sectoral activities is vital to ensure that decisions made by those involved in coastal property development, public infrastructure, agriculture, conservation, recreation, and so on take into account sector-specific impacts on the wider socioecological system.</td>
</tr>
<tr>
<td>Political, administrative, and institutional integration</td>
<td>Jurisdictional boundaries typically define the geographic extent of administrative and institutional responsibility. Such boundaries seldom coincide with coastal ecosystem boundaries, let alone align with the coastal socioecological system, and hence the need to integrate across jurisdictional boundaries to ensure coherent decision-making. In addition, “vertical” integration is vital to ensure that spheres of government from local to national and even international levels adopt approaches consistent with ICM principles and best practices. There also needs to be coordination and integration between relevant laws, policies, and practices that impinge on the coast and in so doing enable integrated action by coastal stakeholders, including those in government, civil society, the private sector, and research community.</td>
</tr>
<tr>
<td>Disciplinary integration</td>
<td>Understanding the nature and dynamics of the coast and associated management challenges and opportunities necessitates research, education, and learning approaches that draw on disciplinary insights, but, where appropriate, integrate such insights to understand the coast holistically. The complexity, uncertainty, and dynamism that characterizes the coastal socioecological system reinforces the need to draw from and integrate a wide range of disciplinary perspectives, recognize the limits of our understanding, and explore ways to align scientific understanding with local, traditional, and tacit knowledge.</td>
</tr>
<tr>
<td>Science–policy–practice integration</td>
<td>Disciplinary, interdisciplinary, and transdisciplinary knowledge and learning are necessary but not sufficient. Such understanding needs to be framed by policy and practice imperatives. Addressing the challenges of the coast requires more effective bridging across the science–policy–practice interface – an undertaking that necessitates co-design and coproduction of applied research and social learning processes that enable community-wide awareness and application of context-relevant best practices.</td>
</tr>
</tbody>
</table>
What does ICM involve?

ICM encourages those involved in discrete coastal activities to think about the system-wide and long-term consequences of their decisions and actions. ICM is usually based on establishing institutional mechanisms that enable cooperation, coordination, and integration between spheres of government (vertical integration) and different actors, activities, sectors (horizontal integration), and other dimensions of integration (see Table 1). ICM promotes social learning and integrates scientific understanding with local, traditional, and tacit knowledge. ICM requires strategic planning that is applied in locally relevant and locally informed ways and adapted over time based on deliberate lesson-learning through monitoring, evaluation, and program adjustment. ICM involves a range of activities, including:

- engaging coastal stakeholders in formulating and implementing a plan to foster sustainable coastal development;
- maintaining ecosystem health, diversity, and productivity;
- sustaining the benefits of ecosystem services;
- enabling sustainable livelihood opportunities;
- resolving conflicting interests;
- facilitating awareness of and understanding about the coast;
- stimulating cooperation, coordination, and ultimately integration.

Resolving conflicting interests is central to ICM. Direct conflict can occur between and within coastal resource user groups especially when there is intense demand for limited space or resources, and/or where activities are incompatible. For example, shore angling, bathing, surfing, off-road vehicles, and dune mining are not feasible in the same locality without interests conflicting. Indirect conflict also occurs as in the example of the road causeway in which a road engineering decision can unintentionally undermine the interests of fishers and other stakeholders.

Deciding how to allocate and utilize the increasingly scarce resources of compromised coastal ecosystems presents a profound management challenge. It needs to be determined how tradeoffs should be made and conflicting interests resolved given the synergies and interconnections between coastal ecosystems, associated services, livelihoods, and the inequitable distribution of entitlements, benefits, costs, and risks that frame such choices.

ICM draws upon and seeks to integrate a range of disciplinary, sector-, and profession-based processes and practices to make such choices, including:

- land-use planning, development approval processes, and monitoring and evaluation protocols;
- strategic planning;
- provisions for housing and service delivery;
- infrastructure planning;
- transport planning;
- economic development planning;
- community development planning;
- emergency management;
- tourism and recreation;
- nature conservation;
- catchment management planning;
- integrated pollution control and waste management.

ICM does not replace these processes and practices but seeks to integrate them through a series of steps and successive cycles of policy and program formulation and implementation that include (i) identification and assessment of coastal issues, problems, and opportunities; (ii) negotiation and formulation of ICM policies and a course of action or program; (iii) formal
adoption by governing authorities of the ICM program or plan and provision of funding; (iv) implementation of agreed policies, provisions, and actions; and (v) ongoing monitoring, review, lesson-learning, and revision (GESAMP 1996). ICM is thus framed as an ongoing and iterative process of negotiated program- and plan-making, lesson-learning, and adaptation (Cicin-Sain and Knecht 1998; Kay and Alder 2005).

ICM in practice

ICM practices have varied significantly around the globe. The European Union (EU), for example, has outlined a general framework for ICM but it is merely advisory and non-binding on member states, and, unlike the US Coastal Zone Management Act, does not provide administrative, financial, and technical incentives for member states to adopt and implement ICM. ICM in developing countries has been mainly donor-funded and project-based. Billions of dollars have been invested in ICM by bilateral and multilateral agencies since the 1990s, aside from in-kind, financial, and nonfinancial contributions by international environmental and humanitarian organizations, national governments, and community-based and nongovernmental organizations. Upscaling and sustaining ICM initiatives has, however, proved challenging in practice despite significant progress made in places as diverse as the Philippines, Vietnam, Bangladesh, South Africa, Ecuador, the United States, and the EU (Kremer and Pinckney 2012; Krishnamurthy et al. 2008). Many ICM initiatives have been small-scale and limited in scope and influence and/or they have not been sustained over sufficiently long periods of time to secure systemic institutional change and sustainable outcomes. ICM is increasingly reframed as coastal governance, wherein divergent goals, understanding, and interests are negotiated in political interactions between coastal stakeholders (Figure 9).

Coastal governance

What is coastal governance?

Governance refers to the interactions of key actors from the state (government), civil society, and private sector, together with the scientific community and media, in seeking to address societal problems through power sharing, social coordination, and collective action. Coastal governance involves interactions between these actors to reconcile sectoral, private and public, and short- and long-term interests in pursuit of coastal resilience and sustainability. These interactions take place through a range of institutions that include both formal (e.g., law and government administration) and informal (e.g., sociocultural norms and “rules”) structures and processes to make social choices that shape community prospects at the coast.

Governance institutions invariably have two features that coexist in dynamic tension. On the one hand, they are authoritative, constraining, disciplining, and stabilizing; and, on the other hand, they are generative, enabling, empowering, and innovative. The art of coastal governance lies in mobilizing these dynamic and countervailing features in ways that foster individual, group, and community behaviors and interactions that are socially equitable, adaptive, resilient, and sustainable. Coastal governance institutions thus convey norms and confer rights, roles, routines, responsibilities, and the resultant culture of coastal management. The outcome of governance interactions and the extent to which the aspirations of different groups and communities are realized is dependent among other things on the power and entitlements
Managing coastal activities: scenes from Guanabara Bay between Rio de Janeiro and Niteroi, Brazil. Diverse and productive coastal ecosystems, especially mangroves, have been heavily impacted by urbanization, deforestation, and water pollution. Photos by Bruce Glavovic.

There has been a tendency to move away from reliance on centralized government authority through top-down prescription, regulation, and control toward more localized, bottom-up, and networked governance interactions. However, both top-down and bottom-up governance modalities are important. Government plays a crucial role in establishing the policy and legislative framework, enforcement, and providing enabling resources. Establishing credible and legitimate governance arrangements depend on the level of trust that communities have in their governing authorities. Civil society thus plays a crucial role in legitimizing government, building social capital, and realizing community aspirations. The private sector and market interactions play a crucial role in entrepreneurship, economic development, and enabling alternative coastal livelihoods.

The remarkable dynamism and diversity of the coast precludes a static “one-size-fits-all” approach to governance. Governance institutions,
processes, and practices need to be tailored according to the distinctive features of different coastal cultures, communities, and jurisdictions. Many coastal issues transcend the scale at which prevailing governance arrangements in coastal communities and even nations operate, for example, land-use planning provisions that stop at the high-water mark or bisect water catchments, or the peril of sea level rise facing low-lying coastal localities and even nations. Particular attention therefore needs to be focused on developing scale-appropriate governance arrangements and enabling integration across scales. Localized issues may need to be addressed through devolved, decentralized, and bottom-up provisions. However, such provisions invariably require higher-level enabling legislation, policies, and resourcing. Cross-scale governance interactions are crucial for addressing issues that transcend the local scale, for example, to address the distant indirect influences of activities in a remote catchment. Institutionalizing “bridging” arrangements may require novel initiatives such as creating forums for dialogue between coastal stakeholders and those who live far inland and who might not even think of themselves as playing a role in coastal governance.

In recent years, attention has been focused on the role of coastal governance in building adaptive capacity in the face of climate change. Any long-term plans or decisions (e.g., infrastructure provision) need to anticipate escalating risk, cope with a wider range of changing conditions, and deal with systemic uncertainty, turbulence, and surprise. This prognosis has significant implications for coastal governance and underscores the need to move beyond traditional linear, science-based decision-making processes that seek to predict future change and resultant impacts, and then prescribe solutions. Attention needs to be focused on how coastal governance can enable communities to deal with this deepening uncertainty, reduce risk, and enable adaptation in the Anthropocene by adopting “no or low regrets” strategies (that are beneficial independent of climate change); adopting redundancy provisions, and flexible, and reversible, policies and practices; maintaining safety margins in significant public investments; and keeping open options to pursue adaptive pathways in the face of inevitable but uncertain change (Moser, Williams, and Boesch 2012; Glavovic et al. 2015). Future prospects at the coast will be shaped by the extent to which governance practices address the root causes and drivers of unsustainable practices, reduce climate risk, and overcome the institutional inertia and maladaptive practices that prevail at the coast.

**Toward best practice coastal governance in the Anthropocene**

Coastal governance practices need to take into account the distinctive features of the coast and draw on lessons from coastal management experience.

**Distinctive features of the coast**

The following distinctive features of the coast need to inform coastal governance endeavors.

The first feature is that of interconnected systems. The coast is defined by powerful and dynamic linkages and direct and indirect interactions between the terrestrial and marine realms, and between coastal ecosystems, functions, services, and people’s livelihoods. Consequently, a single, static definition of the boundaries of the coastal zone is not appropriate for all governance purposes. The geographic extent of the coast needs to be defined in the light of these interconnections and influences relevant for different governance endeavors, including enabling
institutional innovations that resolve local and distant sustainability failures concentrated at the coast.

The second feature is the immense value of coastal ecosystems. These interconnections give rise to a system that is highly dynamic, diverse, and productive, founded on ecosystems that provide services that underpin coastal livelihoods and benefit coastal communities, nations, and humanity. The value of coastal ecosystem services far exceeds that of services derived from terrestrial and marine ecosystems alone. The coast is foundational for enabling and sustaining global human development because it is the primary habitat of humanity, with proportionally more people and economic development concentrated here than elsewhere, and this proportion is growing rapidly. The coast is also the locus for innovation in the Anthropocene.

Third is pervasive and intensifying unsustainable practices. Coastal ecosystems are among the most threatened on Earth, resulting in irreversible loss of ecosystem services that endangers coastal livelihoods, and human wellbeing and development prospects. The coastal sustainability crisis is thus more than an ecological crisis. Poverty and inequity are commonplace and coastal livelihoods are imperiled necessitating that the root causes and drivers of unsustainable practices be confronted.

The fourth feature is the concentration of complexity, change, and uncertainty at the coast. Addressing this crisis is challenging because of the complexity of the coastal socioecological system and the powerful and intricate linkages between natural and human-induced changes in inland, coastal, and marine ecosystems. Our understanding of this complexity is growing but remains far from complete. Uncertainty prevails about the precise nature of locality-specific change in the Anthropocene. Moreover, many coastal communities are likely to face challenges that are unprecedented in recent human history (e.g., rapid melting of arctic ice presents novel challenges for arctic communities and cultures). Coastal governance endeavors must be reframed to deal with the deep complexity, dynamism, uncertainty, and turbulence that characterize the coast in the Anthropocene.

Fifth, the coastal socioecological system is resilient but finite and subject to escalating vulnerability and coastal hazard risks. Accelerating global change, and climate change in particular, will have its most immediate and long-lasting impacts on ecosystems and human activities in coastal and near-shore marine areas due to the very strong linkages between catchments, the coast, and marine systems. The coast, and its constituent ecosystems and human communities, is resilient, but prevailing practices have transgressed critical system thresholds and tipping points and coastal hazard risks are increasing rapidly. Those already vulnerable to the adverse impacts of unsustainable practices, and especially those also exposed to flooding and coastal erosion, will have to contend with increasing sea level, ocean temperatures, and acidity, along with other climate change impacts. Coastal communities are therefore on the frontline of the sustainability crisis in the Anthropocene.

Finally, business as usual is untenable and transformative change is imperative. Traditional community-based coastal management efforts have not been able to withstand the impacts of globalization and Westernization. Over five decades of dedicated and widespread coastal management efforts have not stemmed unsustainable practices globally. Business as usual is fatally flawed for addressing the foregoing features of the coast and is indefensible given the reality of and prospects in the Anthropocene. A paradigm shift in coastal governance thinking and practice is necessary.
Coastal management has evolved rapidly in recent decades and builds on much longer standing efforts. There is growing recognition of the distinctive features of the coast and best practices in coastal governance (Burbridge, Glavovic, and Olsen 2012; Cicin-Sain and Knecht 1998; Glavovic et al. 2008; Glavovic 2013b; Kay and Alder 2005; Stojanovic, Ballinger, and Lalwani 2004). Despite this recognition by coastal scholars, and even rhetoric to this effect in coastal management provisions in diverse coastal settings from the international to local level, prevailing practices tend to prioritize private, short-term interests and economic considerations over longer-term concerns about public safety and the ecological, cultural, and social dimensions of coastal sustainability. Consequently, unsustainable practices prevail.

Best practices for transformative coastal governance

A transformative praxis of coastal governance for sustainability needs to be framed by and take account of the distinctive features of the coast, founded on the following best practices.

People-centered and empowering

Geared toward the needs and wellbeing of current and future generations, transformative coastal governance resolves conflicting interests and empowers coastal communities so that vulnerabilities are reduced and adaptive capacity, resilience, and sustainability fostered. It is framed as a capability building process for all governance actors. Particular attention is focused on sustaining coast-dependent livelihoods and distinctive coastal cultures and lifestyles. A people-centered approach recognizes that coastal communities are reliant on diverse, healthy, and productive coastal ecosystems. It secures the coast as a public heritage.

Ecosystem-based

Coastal governance best practice enables alignment of otherwise incongruent ecosystem and administrative jurisdictional boundaries through ecosystem-based approaches. Coastal ecosystems are managed to sustain ecological functions and ecosystem services. Biophysical limits, assimilative capacities, and ecosystem and resource constraints are understood and respected and ecosystem functions and services secured by, among other things, establishing marine protected areas, sustainable use of renewable coastal resources, avoiding irreversible impacts from nonrenewable resource use, integrated waste and pollution control, and taking into account the powerful physical processes (e.g., currents, tides, and waves) and fluxes that take place in the coastal zone.

Participatory and collaborative

No single coastal stakeholder or actor from government, civil society, or the private sector can secure sustainability outcomes on their own. Inclusive, meaningful opportunities for stakeholder and public engagement in governance processes and decision-making is key to building strategic and practical partnerships, sharing information, developing common understanding and a shared vision for the future, pooling resources and capabilities, and institutionalizing and adapting governance provisions and practices. Authentic participatory and collaborative processes resolve conflicting interests and enable effective partnerships that empower coastal communities and endure over the long term to facilitate coastal sustainability.
Holistic and integrative

Best practice coastal governance is holistic, recognizing the interconnectedness of the coastal socioecological system and the need to take into account powerful, dynamic, cumulative, and synergistic interactions across multiple scales. Integration is enabled iteratively and progressively. It does not replace the need for sector- and interest-based management but invokes institutional mechanisms that enable cooperation, coordination, and, ultimately, scale-relevant vertical and horizontal integration. Complementary bottom-up and top-down measures engage key actors from government, civil society, the private sector, science, and the media in a locally relevant and locally informed manner, recognizing cross-scalar interactions. Account is taken of the timescales relevant to ecosystem functioning and sustainability as well as the shorter timescales relevant to business, political, and project cycles. Attention is focused on meeting the needs of the current generation without foreclosing options for future generations to meet their own needs.

Visionary, strategic, and contingent

Business as usual is unsustainable. Visionary leadership and authentic community engagement enable the paradigm shift in thinking and practice required to chart new pathways in the Anthropocene. A strategic focus enables progress on pre-eminent community concerns and issues, taking into account institutional constraints and capabilities in agenda setting (i.e., be realistic and focus on a few issues that can be done well rather than being overambitious and do many things poorly). Coast-dependent activities and livelihoods (e.g., fishing, ports, and harbors) are prioritized. Coastal governance endeavors are tailored to account for locality specific ecological, cultural, and livelihood diversity at the coast.

Risk averse and precautionary

Risk avoidance and mitigation measures are proactively adopted to reduce vulnerability and build resilience in the face of escalating coastal hazard risks instead of relying on ad hoc postevent response and recovery measures. Particular attention is focused on reducing exposure and vulnerability to coastal hazards, especially those compounded by climate change. If the consequences of a decision are uncertain and potentially dangerous, preventive actions are undertaken even if there is scientific uncertainty about whether or not detrimental impacts will occur and, in the absence of scientific evidence, the burden of proof falls on those advocating the policy or action that may cause harm.

Deliberative and reflexive

Deliberation is a noncoercive process of information sharing, reflection, dialogue, and negotiation over societal values, preferences, and opinions. It is foundational for coastal governance in the Anthropocene. Albeit contested, time-consuming, and resource intensive, deliberative governance builds the community capabilities necessary to understand and address coastal challenges better. Reflexivity refers to critical self-reflection and self-correction which are crucial qualities to develop in the face of the adversity, uncertainty, surprise, and conflict that prevail at the coast. Deliberative and reflexive modalities of governance enable intentional community reflection on current circumstances as well as future prospects, the conscious charting of pathways that are attuned to changing circumstances over time, and the
local and distant drivers of sustainability failure that are concentrated at the coast (Figure 10).

In the long run, these coastal governance best practices need to be mainstreamed into day-to-day planning and decision-making. Hence, it is imperative to reframe and underpin livelihood choices and business practices, and technological development and application, with institutional innovations that facilitate transformation for coastal sustainability. Support and commitment by key governance actors from the highest level down are necessary to build and institutionalize a governance culture of innovation, compliance, and enforcement that fosters sustainable coastal development in the Anthropocene.

Charting pathways to coastal sustainability in the twenty-first century

Unsustainable practices are pervasive, many coastal livelihoods are already precarious, and coastal risk is escalating. Achieving the requisite paradigm shift in thinking and practice, however, poses a fundamental challenge to business as usual and is likely to be strongly resisted by those who benefit most from maintaining the status quo. It remains to be seen if coastal communities will chart new pathways toward sustainability or if seemingly unstoppable trajectories of cumulative demographic, development, and environmental change will be foisted on them. The social benefits of anticipatory strategies that chart new pathways are likely to far outweigh the costs of imposed change. Far-sighted leadership and governance innovations will be needed to transform myopic practices that undermine adaptive capacity, resilience, and sustainability.

Coastal communities need to reconcile the contending governance imperatives of stability and predictability on the one hand, which is prone to rigidity and path-dependent maladaptation, and, on the other hand, flexibility and nimbleness to anticipate and respond to changing circumstances. Maintaining the status quo is unsustainable in the long run, but radical change in the short term could be disruptive, socially unacceptable, and paradoxically impose an unfair burden on those most vulnerable to the adverse impacts of prevailing practices. “Windows of opportunity” to trigger transformative governance innovation can open up, for example, after coastal disasters, but waiting for such opportunities may be futile because experience shows that recovery efforts often re-entrench pre-event vulnerabilities.

Despite mounting evidence about the need to chart a new course, coastal communities find it difficult to build consensus on the need to depart from business as usual – what course ought to be charted and how best to implement it. Locally relevant anticipatory strategies need to be tailored for changing conditions and possible future scenarios. Choices in the short term can be geared toward “no or low regret” outcomes while retaining options for future generations. The relevance and feasibility of local adaptation pathways are likely to be increased when triggers for change have social impacts salient to local people and when pathways resonate with coastal stakeholders.

Transformative change challenges prevailing business and technology and underpinning societal values, institutions, and practices. Profound societal change has occurred in the past, for example, the abolition of slavery, but it invariably involves protracted struggle to overcome formidable obstacles. Deliberate exploration of alternative coastal sustainability pathways is urgently needed. New markets need to be created and alternative governance strategies mobilized. Safe arenas need to be created for communities to experiment with
alternative ways to implement the coastal governance best practices outlined above. In so doing, coastal communities can build shared understanding about the limits of prevailing practices and the opportunities created by alternative pathways. Institutional support, including government support, will be needed to facilitate the diffusion and mainstreaming of emerging innovations. Ultimately, coastal governance practices that build adaptive capacity, resilience, and sustainability need to become progressively more people-centered and empowering; ecosystem-based; participatory and collaborative; holistic and integrative; visionary, strategic, and contingent; risk averse and precautionary; and deliberative and reflexive.

**SEE ALSO:** Climate change adaptation and social transformation; Coastal depositional processes and landforms; Coastal erosion processes and landforms; Coastal zones; Coasts; Environmental governance; Global environmental change: human dimensions;
References


Glavovic, B.C., M. Kelly, R. Kay, and A. Travers, eds. 2015. Climate Change and the Coast: Building Resilient Communities. Boca Raton, FL: CRC Press.


Cognition and spatial behavior

Daniel R. Montello
University of California, Santa Barbara, USA

A core goal for geographers is to describe, predict, and explain human activity on the Earth. The concept of spatial behavior highlights the geographer's focus on the spatial and temporal aspects of this activity – where people go to carry out particular activities at particular times, and how they travel there. Because the smallest cohesive unit of human spatial activity may be one trip carried out by one person at one time, some geographers study human activity at the level of single trips by individual people rather than groups of trips or the activity of groups of people. In contrast, most studies by human geographers traditionally take a group approach, analyzing activity at the level of neighborhoods, cities, institutions, cultures, and so on. In the 1960s and 1970s, what was then the new individualistic approach to understanding human geography came to be known as behavioral geography.

Historically, this turn toward understanding human activity at the disaggregate level that was at the heart of behavioral geography led relatively directly to the further idea that human activity should be understood in terms of human beliefs and reasoning. From the inception of behavioral geography, scholars realized that individual people are not passive agents in the world but active decision-makers who reason about behavioral choices. This reasoning, it was assumed, would not be based upon actual properties of the world and the individuals, but upon people's beliefs about properties of the world and themselves. That is, behavior is based on the subjective or perceived world. The subjective world consists of beliefs about spatial properties, such as location, distance, direction, connectivity, and containment. It also consists of beliefs about nonspatial properties, including temporal properties and thematic properties, such as what objects and events are in the world, what their positive or negative hedonic qualities are (i.e., their potential as resource or hazard), what preferences and abilities a person has, and so on.

Human beliefs and reasoning are part of the study of cognition, and behavioral geographers focusing on cognition are sometimes said to be doing cognitive geography. (Other behavioral work is not specifically cognitive but does focus on the behavior of individuals.) Cognitive geographers participate in a multi- and interdisciplinary scholarly endeavor, in terms of theories, concepts, and methods. Besides connecting with the work of other geographers, such as those in economic geography, human–environment relations, and cartography and geographic information science, they connect directly with the work of psychologists, including those in cognitive, environmental, perceptual, social, and developmental subfields. They also connect with the work of other cognitive scientists, including those in computer science, philosophy, linguistics, and neuroscience, and other social-behavioral scientists and scholars, including those in sociology, economics, anthropology, animal behavior, and planning and architecture.
Several concepts are basic to understanding cognition and spatial behavior. **Sensation** is the first response of the nervous system to stimulation from the world. Sensory receptors transduce patterned world energy, such as light or chemical, into patterned nervous system energy, which is electrochemical. A widespread misconception is that humans have but five sensory modalities; depending on how one slices it, the correct number is at least eight or nine, and includes vision, hearing, smelling, tasting, pressure and texture sensing, temperature sensing, kinesthesia (limb position and movement), and vestibular sensing (linear and angular acceleration). Pain is sometimes considered a sense. To (sighted) humans, vision is probably most important for sensing the spatiality of the world, but hearing, kinesthetic sensing, and vestibular sensing all play important roles. Senses like smell and texture and temperature sensing are important for place or feature identification, but play a limited role in humans for sensing spatiality. One can detect the direction of the wind or the sun on one’s face, which sometimes provides a useful cue to heading or travel direction.

The organization of sensory input into meaningful impressions of oneself and of the world, as influenced by prior beliefs, is **perception**. In commonsense experiential terms, the environment as perceived is three-dimensional with oneself in the center, and detachable objects and events against a stable background. The perceived world has varied sensory qualities, such as colors, tones, and so on; presents information redundantly (e.g., different depth cues typically coincide); is incompletely apprehended from a point of view and point in time; demonstrates constancies; and tends to be perceived in terms of meaningful and familiar objects, events, and settings. **Constancies** are the phenomena wherein entities appear to stay the same even as the sensory information they offer changes radically. A long-standing puzzle for perception theorists has been to explain constancies in the face of our locational perspective from a single place and time. How does a swinging door appear to remain rectangular even as its outline on the retina changes shape?

Perception is considered to be part of the larger topic of **cognition**, which is about knowing and knowledge (believing and beliefs), and also includes thinking, learning, memory, concepts, imagery, language, and reasoning. The focusing or directing of cognition is **attention**. Attention is controlled internally, as when you intentionally retrieve a certain belief from memory or try to listen to a particular speaker, or externally, as when a loud sound or colorful building captures your attention (the degree to which stimuli attract attention is called **salience**). Some types of cognitive processing require explicit attentional resources – they are said to have **capacity limitation** and are readily subject to **interference** by distracting tasks. Attempting to find your way in an unfamiliar environment is an example. But at the same time, other cognitive tasks require little or no attentional resources. Following your typical route to work each day is an example. Walking around without running into walls is another.

Cognition clearly influences and is influenced by **emotion** (affect), as when certain mental images lead to emotional states or certain emotional states motivate one to focus attention on certain aspects of the world. Although human emotions can be complex and nuanced, at their core is the central component of **evaluation** or hedonic tone – a positive or negative response to beliefs about the world. The results of coupling beliefs with affective states, possibly leading to behavioral intentions, are known as **attitudes**.

As we described earlier, geographers are interested in cognition primarily because of its links to behavior. **Behavior** is potentially observable,
goal-directed body movement; it is sometimes called action or activity. Behavior is goal-directed in that its purpose is to achieve some end, such as getting to a place, obtaining a resource, or avoiding a hazard. It is not an internal mental state, such as thoughts or moods, although it is profoundly interrelated to them, as we have noted. In fact, the relationship of behavior to cognition (and emotion) provides a large and complex set of issues for ongoing research and theoretical debate in several fields. That said, behavior is clearly influenced by cognition, if not entirely determined by it. Sometimes overlooked is that cognition is also influenced by behavior, as when people travel about in order to gather information about their surroundings.

Just as behavior is not equivalent to internal mental states, it is not equivalent to neurological states of the brain or the rest of the nervous system. The relationship of the mind and brain is the profound and very old mind–body question in philosophy. Briefly, this asks, what is the nature of mind (experience, awareness, soul, spirit), the nature of body (brain, body, physical world), and their relationship? There are many highly developed philosophical views on this question over the centuries. Many behavioral and cognitive scientists believe the mind emerges from the action of a brain in a body, living in a physical and sociocultural world. Thus, the mind requires a brain but is not reducible just to the brain. Still, without committing to the strong reductionism that equates mind and brain, it is evident that we learn more about the mind (and behavior) by better understanding the brain. In the past two or three decades, there has been a great increase in research on the neuroscience of cognition, much of it fueled by advances in brain imaging technologies, especially fMRI (functional magnetic resonance imaging). Such research attempts to relate patterns of activity in various regions of the brain to particular cognitive states, such as certain reasoning styles or content. Cognitive neuroscience has made substantial contributions to our understanding of the cognition of space and place over the last couple decades, and such studies are just now beginning to appear in geographic literature. We can expect to see the influence of neuroscience grow in the discipline of geography, as it is doing in nearly every other field that studies human mind and behavior.

**Empirical methods for studying cognition and spatial behavior**

As in other problem domains within geography, researchers who study cognition and spatial behavior need methods to observe and measure their phenomena of interest. The major approach to such methods in the study of cognition and behavior has been observing and recording the behaviors of research participants, individually or in groups. Other useful methods include examining secondary archives, including data mining of the Internet and social media. Physical traces can support inferences about where behaviors occurred and what people believed about something; these physical traces may have been intentionally or unintentionally created. Computational modeling, including robotics and other forms of artificial intelligence, continue to shed light on cognition and spatial behavior. During the late twentieth century, neuroscience methods became available for studying brain activity during ongoing thought in conscious, nonclinical respondents. Finally, a truly geographic approach to studying cognition and behavior should involve thorough analysis of the environment, whether natural or built, and any informational material involved in cognition and behavior (such as cartographic maps, photos, texts, etc.).
Recognizing this methodological diversity, it can still be noted that behavioral methods are the dominant way geographers study cognition and spatial behavior. These can be based on verbal or nonverbal behaviors, and include where people travel and along which routes, where and at what they look or point, what they draw, and what they say or write. Some behavioral geographers focus on the observed behaviors themselves as their phenomena of interest. For instance, some geographers record daily commuting trips without considering what people believe about their trips or why they think they make them. Other geographers with more cognitive interests typically make inferences about thoughts and beliefs from the observed behaviors. Either way, one can broadly distinguish behavioral methods involving the nonmanipulated observation of ongoing behavior from those that ask people to explicitly express their beliefs. The latter are called explicit reports: whether written, drawn, or spoken, these expressions are behaviors. Explicit reports are used to study people’s beliefs about themselves or others, about places or events, about objects or activities; any of these beliefs can include beliefs about spatial properties or attributes. When providing data via explicit reports, research participants know they are supplying information about their beliefs to a researcher, and these beliefs must be consciously accessible to the participants. Explicit reports can be further distinguished as tests, which generate responses that are evaluated for accuracy, or surveys (polls, interviews), responses to which cannot readily be judged in terms of accuracy or whose degree of accuracy is not of central interest to researchers (although truthfulness is). That is, surveys generally assess opinions, preferences, or personal experiences rather than knowledge. Of course, the explicit reports used to study spatial cognition are more often tests than surveys, but other cognitive geographers use surveys frequently.

Sketch mapping

A widely used method to obtain data for the study of cognition and spatial behavior is having research participants sketch maps (or construct physical models, etc.) of places or regions. Sketch mapping is one of the most straightforward ways to find out what people believe about spatial layouts at any scale, and perhaps hundreds of studies and thousands of informal demonstrations have collected sketch maps. But as with other open-ended methods, the ease and simplicity of collecting sketch maps is not matched by easy and simple coding and analysis. In the end, the questions you want answered by the sketch maps should determine how you code and analyze them – there is no omnirelevant approach. Types or specific instances of features such as paths or landmarks can be counted. Spatial properties such as distance or direction can be measured. The orientation of the sketches can be coded, as can their drawing style, or the presence or absence of particular verbal labels. The challenge of unambiguously identifying map features and making comparisons across individuals or groups can be made easier if the sketching task starts with more structure than just a blank sheet of paper or blank computer screen. Cardinal directions, distance scales, road networks, or mountains and water bodies can be present at the start. A list of features to be located on the map can be provided, allowing the researcher to focus on the location of placed features rather than whether they are included at all.

Scaling

A diverse set of explicit-report techniques used to study spatial cognition is known as scaling.
When scaling, research participants directly express their beliefs about quantitative properties, meaning properties that are not just classified but rated or estimated at a metric level of measurement – interval or ratio. Scaling derives from two methodological traditions within the history of research psychology: psychophysics and psychometrics. Psychophysical scaling originated during the nineteenth century. It requires participants to estimate quantities of a property that researchers relate to values of objectively measured quantities. For instance, participants can estimate distances, directions, or sizes. A specific example would be asking someone to estimate the distance between two cities in miles. Psychometric scaling, in contrast, originated during the early twentieth century. It requires participants to estimate quantities of a property that cannot be compared directly to objectively measurable quantities. For instance, participants can quantitatively express attitudes, abilities, preferences, or moods. A specific example would be asking someone to rate how much they like different cities on a scale from 1 to 10. Both psychophysics and psychometrics demonstrate persuasively that subjective mental states in humans can be scientifically studied.

Values generated using scaling techniques can be statistically interpreted as individual variables describing individual entities, such as the area of a city, or relational variables describing pairs of entities, such as the distance between two cities. A sophisticated way to analyze scaling data, called multidimensional scaling (MDS), however, provides a way to analyze and interpret values concurrently across an entire set of entities of interest. To understand this technique intuitively, imagine using a ruler to measure distances on a map between pairs of points representing cities in order to create a traditional mileage chart: a matrix of distances between pairs of the cities. MDS effectively reverses this procedure (computationally) to create a configuration of points given only the matrix of pairwise distances between cities in a given set. In fact, MDS algorithms can generate metric configurations from nonmetric input, such as ranks of distances, instead of metric distances, and other spatial properties such as directions can provide a basis for MDS. Importantly, when the separations between points represent dissimilarities between the entities rather than literal distances, so-called semantic spaces can be created that use space metaphorically to represent meaning with any types of entities as a spatial configuration.

Cognitive maps and mapping

The concept most central to the study of cognition and spatial behavior is the cognitive map. Downs and Stea (1973, 9), in the introductory chapter to their influential edited collection, defined cognitive mapping as “a process composed of a series of psychological transformations by which an individual acquires, codes, stores, recalls, and decodes information about the relative locations and attributes of phenomena in his [or her] everyday spatial environment”; the stored information is the cognitive map. Cognitive maps are fundamentally idiosyncratic to individuals but are partially shared among groups of people. The notion that mental representations of the environment guide behavior may be found in the academic literature at least as long ago as the early twentieth century, but the specific term “cognitive map” is attributed to the animal behaviorist Tolman (1948) to explain the behavior of his rats in a tabletop maze.

The concept of the cognitive map is a metaphor, suggesting that mental representations of the environment are like cartographic maps in the mind. Many other metaphors for this
Cognition and Spatial Behavior

Concept are possible and have been suggested by various scholars, such as imaginary map, mental model, or cognitive collage. But like any metaphor, the source domain of cartographic maps may not capture the target domain of environmental knowledge perfectly; research literature has discussed this issue at length. Both cartographic and cognitive maps are representations that contain spatial and nonspatial information, both are selective, both distort properties of the world, both encode from different spatial perspectives, both represent features on a continuum of abstractness (from relatively iconic to relatively arbitrary), and both serve various functions beyond just guiding navigation. The metaphor can be quite misleading, however. The cognitive map is not a unitary and uniform representation, contrary to the idea of a “mental picture” in the mind. It is a collection of pieces (separate representations) that are not continuously integrated with a constant or continuously varying scale. The cognitive map derives from multiple sources and is mentally represented in various formats and perspectives. The pieces that make up the cognitive map are not mutually coordinated, at least not completely, and they may express spatial beliefs that do not follow a Euclidean metric or any metric geometry at all. While cartographic maps can certainly stimulate emotional responses in map users, cognitive maps themselves incorporate emotionality, such as attitudes of fondness or fear toward places.

In the end, aspects of evolution and learning shape beliefs and decision-making to generally support adaptive behavior, leading people to be spatially oriented, travel efficiently, make rewarding choices and avoid harmful ones, and communicate effectively. In this regard, the idea that internal mental states and processes mediate observed spatial behavior leads us to see geographic beliefs and reasoning as functioning to organize, direct, and enrich experience. But early in the scholarly history of this domain, behavioral researchers also recognized that the subjective world may deviate greatly from the objective world and that beliefs can be mal-adaptive. Spatial knowledge includes beliefs that are in error relative to objective reality. These errors have both systematic (consistent) and random components. Researchers study errors as a way to understand several aspects of spatial knowledge, including its content, its resolution, how it is structured and processed in the mind, and how it relates to emotion and behavior.

Landmarks

An important component of people's knowledge of the environment is the landmark, which in general terms is a feature or object in the environment that is relatively distinct and can be noticed and remembered. People use landmarks to recognize places and orient themselves, and to communicate this with other people. In more sophisticated terms, landmarks can express a symbolic meaning for a place, serve as cues for actions, or function as reference points around which place knowledge more broadly is mentally organized. Although we usually think of landmarks as providing a key to location, in many cases, it is our knowledge of a landmark's location that provides the context to disambiguate its identity as a landmark in the first place. And even though landmarks are often caricaturized as prominent point-like features (the Eiffel Tower!), they are often extended features that are line-like or areal, especially at particular spatial scales. Even acknowledging how common it is that people orient themselves by visually (or via other senses) recognizing the environment, it is important to realize they often do it by recognizing entire scenes within the visual field, perhaps without being consciously aware they are doing it. Such “landmarks” are not useful as
part of verbal route directions the way landmarks as distinct and discretely labeled features are.

**Regions**

Another important structural component of spatial knowledge of the environment at any scale is the areal (two-dimensional) concept of *regions*. Long studied by geographers and used by them to organize their understanding of the Earth’s surface, regions are central to the spatial thinking of laypeople as well. Regions break a continuous Earth surface into discrete pieces, essentially spatial categories. They are bounded, and the boundaries vary in their precision or vagueness, and in their permeability to matter, energy, and information. Furthermore, regions are often organized hierarchically, with regions at different levels of status (such as size or power) connected to each other in relations of containment. This allows hierarchical reasoning, wherein spatial relations between places are inferred from relations between the regions to which they belong, rather than stored directly (Stevens and Coupe 1978). More subtle models of hierarchical reasoning suggest that locations are inferred by weighting combinations of coarse regional membership and precise metric location.

There are various types of regions in geography. *Cognitive regions* (also called perceptual regions) are informal regions in the mind, socially/culturally shared to varying degrees. A diverse set of behavioral phenomena has been offered as evidence for regional organization and hierarchical reasoning, including systematic errors in direction or distance judgments, particular patterns of response latencies to answer questions about spatial relationships, systematic ordering of sequences of recalled places or features, and more. Friedman and Brown (2000) collected estimates of cities’ latitudes to show that people organize their knowledge of spatial relations at the continental scale in terms of regions (they termed them *cognitive plates*) that compress spatial relations within the regions but exaggerate them between regions. Furthermore, the separate regions can be mentally shifted independently of each other, producing distinct patterns of distorted knowledge (see Figure 1).

**Environmental spatial learning and development**

Any process of systematic change over time may be called *development* or *evolution*. Processes of change are of interest across the discipline of geography. They occur at any temporal and spatial scale and for entities at many levels of analysis, including individuals, cultures, institutions, ecosystems, species, and landforms. Cognitive researchers in geography are most interested in *ontogenesis* (ontogeny), the development of cognition from an individual’s conception to his/her death (also called child or lifespan development), and *microgenesis* (microgeny), the development of cognition over shorter time periods from initial exposure to new information or new situations to later states of familiarity.

An important distinction for those studying ontogenesis and microgenesis is that between *learning* and *maturation*. Learning refers to relatively permanent change in the content, structure, or processing of knowledge due to specific cognitive experiences such as those involving new information or semiotic representations, not physical experiences such as those involving illness, injury, or exercise. Maturation is the unfolding of innate change processes over time, which do not require specific environmental experiences (but usually require general experiences, such as adequate nutrition). Of course, while learning occurs as part of both developmental processes, maturation typically
Figure 1  Estimated latitudes of cities in North America, Europe, and North Africa, ordered left to right from cities judged most northerly to cities judged most southerly, and plotted against actual latitudes. The pattern reveals subjective latitude compression within regions, exaggeration between regions, and north- or southward shiftings of cities in certain regions. Source: Friedman and Brown (2000), Fig. 2. Copyright American Psychological Association. Reprinted with permission.

does not take place within the limited time periods of microgenesis, which often occur within as little as seconds or minutes. Nonetheless, the nature of an individual’s learning is strongly linked to his or her maturational development, as when older children can learn more complex spatial layouts than younger children can in the same exploration period.

Commonly, theorists studying change of any type in any system contrast stage and continuous models of development. Stage models propose that change occurs in qualitatively distinct episodes or periods, each stage having a coherent theme of interrelated events or abilities, with relatively abrupt transitions between stages, and an invariant sequencing of stages (although the timing of transition may vary). Continuous models oppose this idea, suggesting that change occurs relatively continuously rather than in discrete stages, and is more accurately described
as quantitative rather than qualitative in nature. Although the contrast between stage and continuous models is intellectually fruitful, it is typically ambiguous to choose between them definitively, in part because processes that appear discontinuous at one scale of analysis appear continuous at another.

Scholars study a host of specific aspects of cognitive change, but central to those studying geographic spatial cognition is the issue of how children and adults organize their understanding of location on the Earth’s surface, both in thought and in communication. This is the topic of reference systems (sometimes called frames of reference). As geographers know well, all locational information is relative; reference systems are the ways that locational statements are defined or anchored. These include the precise and quantitative systems familiar to surveyors and geographic information scientists, known as coordinate systems. Cognitive geographers are usually more interested in the approximate and qualitative systems mostly employed by laypeople. A variety of typologies of cognitive and linguistic reference systems have been proposed over the years. Such typologies usually distinguish egocentric systems based on one’s body (“front-back,” “left-right”) from allocentric systems based on something external to the body. The latter is sometimes further distinguished as being based on nearby features or landmarks that define location only very locally (“near the tree,” “at the gas station”) versus distant or global features that define location over large areas (“head toward the ocean,” “go north”); those based on global features are also known as absolute reference systems. The topic of reference systems is a rich one. Individuals of different ages clearly tend to use different reference systems, as do people differing in their level of familiarity with places. Different cultural or linguistic groups use reference systems somewhat differently; even people of the same cultural group apparently use them differently as a function of their residential environment (for instance, whether they live in flat or mountainous terrain). And a single person can use multiple reference systems in thought and language, either as a result of subtle changes in context or even with the same chain of reasoning or expression.

Ontogenesis of environmental cognition

A great deal of research for over half a century has looked at the spatial cognitive development of infants and children. Early in their life, infants develop various aspects of spatial perception, including the perception of shape, depth, size constancy, and perspective changes. Important to perceptual and cognitive changes are changes in motor behaviors, whether lifting one’s head, crawling, or walking. As infants become toddlers and young children (variously up to age 6 or so), they are typically able to travel much further in their environment and acquire richer, more extensive cognitive maps. Other research has looked at the acquisition of spatial concepts at different levels of geometric sophistication, including metric and topological properties. As young children become middle children (up to adolescence) and teenagers, they further develop their ability to plan and choose routes, and conduct organized spatial searches. Their ability to give and interpret verbal directions improves, and their understanding and use of cartographic maps increases in accuracy and complexity.

During the mid-twentieth century, the theories of the Swiss psychologist Jean Piaget and his colleagues were widely explored to explain cognitive development in children, including spatial and environmental cognition. His stage theory
of cognitive ontogenesis saw development as a biologically evolved process of adaptation to the complex, uncertain environments that children encounter. For Piaget, knowing involves actively selecting, interpreting, and constructing. Action upon the world and its effects on the surroundings is critical to knowledge development. The nature and modification of knowledge structures (schemas) can be characterized in terms of four major stages, typically taking place during certain characteristic age spans: sensorimotor (ages 0–2 years), preoperational (2–7), concrete operational (7–11), and formal operational (11+). The theory is rich and rather complicated, but in brief, changes through these stages can be summarized as cognition going from simple to complex, prelogical to logical, concrete to abstract, and perceptual/action-based to conceptual/reasoning-based. Besides his elaborate theoretical framework, Piaget’s work also provided a series of influential empirical tasks and concepts for later scholars, including object permanence, spatial egocentrism and perspective-taking, conservation, and nonmetric spatial thought.

Substantial problems have been identified with Piaget’s large body of work, including doubts about the timing of his stages, his tendency to underestimate the abilities of infants and young children, and his tendency to overestimate the abilities of teenagers and adults. Various theoretical alternatives have cast doubt upon the existence of coherent general stages of cognitive development and criticized Piaget’s relative ignoring of the roles of language, culture, and social interaction, including formal and informal schooling. Some of these alternative theories include information-processing theory, situated cognition, linguistic relativity, and conceptual nativism (see Newcombe and Huttenlocher 2000).

Microgenesis of environmental cognition

By analogy to stage theories of ontogenesis, an influential framework for understanding the microgenesis of environmental spatial knowledge arose during the 1960s and 1970s. This “dominant framework” was most eloquently expressed by Siegel and White, as described by Ishikawa and Montello (2006). It proposed that spatial knowledge in a new environment, such as when a person moves to a new city, develops over time in an invariant sequence of three stages. The first is landmark knowledge, which is not explicitly spatial at all, but merely implies recognition of distinctive features or objects. The second stage is route knowledge, which refers to locomotor routines connecting sequences of landmarks. Knowledge at this stage, at least initially, is nonmetric and organized only egocentrically or with reference to local landmarks. The third and most mature level of knowledge is survey knowledge (or configurational knowledge). This is two-dimensional layout knowledge, simultaneously representing spatial interrelations of landmarks and routes. It contains metrically scaled distances and directions, and integrates stored landmarks and routes into a unified, coordinated representation organized according to a global allocentric reference system. Taking inspiration from Tolman’s (1948) rats, the behavioral sign of survey knowledge is the ability to create novel shortcuts and detours — in short, to navigate creatively. Although presentations of this framework have never specified how much time passed between stages, they implied that the time periods would be substantial, perhaps in the order of months or years.

Some researchers have questioned this dominant stage theory of spatial microgenesis. Ishikawa and Montello (2006) describe an alternative framework, published earlier by Montello, that proposes that development is relatively continuous and quantitative, not stage-like. That is,
spatial knowledge acquisition starts immediately upon arrival at a new place, and the extent, accuracy, and completeness of knowledge continue to grow indefinitely (the exact time-course and ultimate limit of knowledge development were recognized as important research questions). This spatial knowledge includes some approximate metric knowledge right away; there is no period of pure nonmetric landmark or route knowledge. This framework does include the relatively abrupt step that separately learned knowledge of routes and regions acquired during unitary travel episodes can relatively suddenly be integrated into more complex, hierarchically organized structures in what might be considered a form of spatial insight.

Ishikawa and Montello (2006) presented a longitudinal study designed to compare these two frameworks for microgenesis. Participants rode individually with the researcher on automobile trips through a local neighborhood they were not previously familiar with, taking one ride per week for 10 weeks. Their distance and directional knowledge about the relationships between landmarks on two test routes was tested each week, and participants drew sketch maps of the routes every other week. Unexpectedly, the results, shown in Figure 2, indicated that different participants showed

![Figure 2](image_url)
very different patterns of knowledge acquisition over time, which neither framework implies. Contrary to the dominant framework, some participants acquired accurate and surprisingly precise metric knowledge after their first trip, and continued to demonstrate very good spatial knowledge throughout the 10 weeks. Contrary to both frameworks, other participants acquired little or no metric knowledge after their first trip, and even 10 weeks of visiting the site resulted in little or no metric spatial knowledge. Only a subset of participants showed marked improvement over 10 weeks, and this improvement appeared more continuous than stage-like.

Navigation and orientation

All human geographers are interested in people’s activities in space and place, and perhaps the most central example of such activity is travel between places, whether temporary, as in a shopping trip, or more permanent, as in migration (residential relocation). Perhaps the most salient aspect of travel for cognitive geographers is that people so often do it in a coordinated and efficient manner. Such coordinated and goal-directed movement of oneself through the environment is navigation. Navigation includes not just the specialized activity of professionals on ships and airplanes, but something almost everyone does many times every day, when they walk to class, drive to work, and so on. Because the literal navigation referred to here – actual body movement over the Earth’s surface – provides such a concrete and universal case of problem-solving within a complex situation, laypeople and scholars alike often speak of other problem-solving contexts metaphorically as navigation. For example, we may navigate through a math problem, a difficult text, a website, or an emotionally vexing relationship crisis.

One can distinguish two components of navigation: locomotion and wayfinding. Locomotion is the coordination of the body to the local surrounds during movement, serving to avoid obstacles and barriers, move toward perceptible landmarks, and so on. It involves processing information about the surrounding environment that is directly accessible to sensory and motor systems, and thus does not typically require internal (memory) or external (cartographic maps, etc.) representations of the environment. Locomotion takes place via various modes, whether strictly body-based or involving technologies, like automobiles or airplanes. People locomote at various speeds, facing various headings, and following various courses. Active locomotion is often distinguished from passive, usually referring to whether the movement is self-directed or not. Navigation also involves wayfinding, the goal-directed planning and decision-making part of travel. The goal of navigation is getting to a destination efficiently. Wayfinding is the set of processes by which a person strives to achieve this in the common situation where the goal is not perceptible from the person’s current location. That is, wayfinding is coordinated to the distal environment not directly accessible to sensory and motor systems. Thus, wayfinding does require either an internal or an external representation of the environment, although it may be very schematic or incomplete. Typical wayfinding tasks include choosing routes, scheduling trips, and maintaining orientation to the environment beyond the immediate surrounds.

Maintaining geographic orientation – a sense or knowledge of where you or where your destination is on the Earth’s surface – involves some combination of knowledge about location, distance, and direction. Of course, a significant wayfinding problem, unfortunately common for some individuals, is that people are not always oriented. You are oriented when you think you
know where you are or where to go, and you are correct. If you are not oriented, you may be either disoriented or misoriented. Disorientation is subjective, occurring when you think you don’t know where you are or which way to go, or aren’t sure; it is what people usually mean by saying they are lost. Misorientation is objective, occurring when you are actually not where you think you are or are actually going the wrong way. Each state often happens without the other, especially being misoriented without being disoriented. This is unfortunate, as people do not attempt to reorient if they are misoriented without being disoriented.

There are two types of processes people use to update their orientation as they move about: landmark-based updating and dead reckoning. Landmark-based updating, also known as pilotage, position fixing, or taking a fix, is based on recognizing external features or objects. Sometimes a visible (or audible, etc.) landmark is available at a person’s destination location, and the person can simply locomote toward it, a process called beacon-following. This is actually fairly rare as a case of landmark-based updating. Much more commonly, people pilot by recognizing visible landmarks in the surrounds, which in turn allows them to identify their location and heading within their cognitive or cartographic map, which then allows them to relate their current position to the locations of distal features and places. In contrast to landmark-based updating, dead reckoning, also known as path integration or inertial navigation, does not involve recognizing external features. Instead, given knowledge of one’s initial location, information about the direction and speed of one’s travel can be used to infer one’s final destination, after travel. Without specialized technology, people can sense movement direction and speed with the help of idiothetic (internal) or allothetic (external) signals. An important example of idiothetic information is the information one gets about linear and angular acceleration from vestibular sensing. An important example of allothetic information is the information one gets from optic flow, the movement of textural elements through the visual field, without recognizing specific features. Of course, dead reckoning is useful in completely unfamiliar areas, but it has some important limitations. It requires that you know the location where you started for full orientation on the Earth’s surface. And it suffers from the accumulation of error over time, which must be corrected by position fixing.

When cartographic maps are used to maintain orientation during ongoing travel, the navigator must coordinate his or her current location and heading on the map with his or her location and heading in the surrounding environment. Examples are “you-are-here” (YAH) maps or road maps used while riding as a passenger in a car. Such navigation maps exhibit orientation specificity: They are perceived and interpreted best in a single orientation. That orientation is generally “forward-up,” with the navigator’s forward heading in the surrounds represented as up on the map (assuming a vertically held map, facing the navigator). Any other relationship between the map’s orientation and the navigator’s heading tends strongly to produce misalignment effects (usually called “alignment effects” in the literature). This is the extra time, error, effort, and/or displeasure that results when using misaligned navigation maps. Travelers typically rotate detached navigation maps to keep them forward-up aligned as they change headings; digital navigation systems usually have the option of automatically reorienting in this way. Fixed YAH maps cannot be rotated in this way, and it is surprisingly common to find such misaligned maps mounted in public places. Various cognitive strategies have been identified for reasoning effectively with misaligned navigation.
maps, but the map user must recognize these strategies are called for, they must know how to use them correctly, and they must have the cognitive ability to do so. Interestingly, while the automatic forward-up realignment of maps in digital navigation systems helps most people stay oriented, it may not support the acquisition of cognitive maps of an area as well as a fixed orientation, such as north-up, does.

Distance and direction knowledge

Being oriented often requires knowing something about distances and/or directions in the environment, and effective spatial planning and decision-making more broadly typically require some knowledge of distances and directions. From a cognitive perspective, the concept of an environment implies a space that is so large and otherwise obscured by features that one can directly apprehend its spatial layout only by locomoting around and mentally integrating separate sensorimotor experiences over time. It cannot be viewed entirely from a single point, as a cartographic map allows. Thus, beliefs about distances in directly experienced environments are beliefs about distances along traveled routes. Research has suggested a variety of types of information that can provide a basis for beliefs about environmental distances, or at least that can influence judgments of distances if they do not entirely determine them. Evidence indicates that the presence of environmental features such as path intersections and segments, turns, barriers, and landmarks will typically increase subjective distances. The cognitive mechanism for this is not entirely clear. People might just mentally equate distance traveled with the number of features noticed and remembered. It appears more likely that they use prominent features to subjectively segment routes into pieces, which in turn leads to longer subjective distances on more segmented routes because of category or psychophysical scaling effects in estimation.

The presence of environmental features can also affect one’s sense of travel time, which in turn can influence one’s sense of traveled distance. Geographers often treat travel time as distance itself, insofar as they consider distance abstractly to mean any measure of the cost of overcoming the separation between places. In many cases, people undoubtedly use travel time as the relevant cost to consider when they spatially plan. This does not mean that people cognitively equate time and distance, however, only that in some situations they use time instead of physical distance to make decisions. Limiting cases show that people do cognitively distinguish time and distance, as when they realize they have not gone far when sitting in a traffic jam, or they know they have gone far while sitting for a couple of hours on a plane. But it is still quite likely, especially when people have poor access to information about their movement speed, that travel time influences judgments of traveled distance. Evidence does not suggest this relationship is very robust, however, whatever intuition suggests. And basic psychological research indicates that the relationship of events (such as the presence of features) to subjective time is not simple. Events can expand estimates of past time intervals when the events are recalled retrospectively as part of judging the interval. But events can shrink estimates of time intervals if they are the focus of attention prospectively, during travel – they can distract one from focusing on time or distance. Finally, physical effort has also been proposed to influence judgments of traveled distances, at least over relatively short extents, but demonstrations of the influence of effort are also not robust and invite severe
criticisms about the possible role of participant expectations on observed judgments (known as demand characteristics).

Knowledge of directions in the environment is also important to spatial planning, and especially central to being oriented. Given that localizing entities on the Earth’s surface is mostly a two-dimensional problem, the focus of cognitive research has almost entirely been on azimuths rather than elevations or slopes. Many studies ask research participants to point to features (with their hand or a tool) from their current location and heading, or from an imagined location and heading (“point to the grocery store as if you were at the courthouse steps, facing west”); these are often called judgments of relative direction. When exploring theoretical questions about the resolution or vagueness of spatial knowledge, researchers are most interested in random or unsystematic errors in direction estimates. It is generally correct to calculate variable errors in this case. At other times, researchers want to explore theoretical questions about bias or distortion in spatial knowledge; they are most interested in systematic tendencies to estimate directions either clockwise or counterclockwise of the correct direction. It is generally correct to calculate constant errors in this case. At still other times, researchers want to explore theoretical questions about how people differ from each other in their abilities to estimate directions; they are interested here in analyses that combine the two types of error, calculating absolute errors to do so. Properly calculating absolute errors is just a matter of taking the absolute value of the difference between estimates and correct directions. Properly calculating variable and constant errors is more involved and requires using circular (directional) statistics.

Cognition of cartographic maps and other geographic information displays

People experience the Earth and acquire geographic information, including spatial, temporal, and thematic information, from directly interacting with the world via sensorimotor systems or from interacting with some type of symbolic medium. The nature of people’s experience and the information they acquire varies as a function of the nature of this interaction. If one directly interacts with the world, they can be stationary or locomoting. If locomoting, they can be crawling, walking, or running. Their locomotion may be mechanically aided, with a bicycle, car, boat, or plane. If one interacts indirectly with the world via symbolic media, they may do so with static pictorial representations, such as graphs, maps, drawings, or photos. Or they may do so with dynamic pictorial representations, such as animations or movies. They may experience the world through natural language like English or Mandarin Chinese, whether spoken, written, sung, or signed, or formal language like mathematics, symbolic logic, or a computer programming language. Virtual reality systems may be more like dynamic pictorial representations or more like direct experience, depending on the nature of the system. Of course, people commonly experience the world through more than one of these, either simultaneously or sequentially over time. Ongoing research addresses how information from multiple sources is combined, or even whether it is.

These different ways of interacting with the world present the world and its properties differently, and likely influence the beliefs one acquires about the world. They involve different sensory and motor systems; for instance, some involve body locomotion and some do not.
COGNITION AND SPATIAL BEHAVIOR

Some present static information about the world and some present dynamic; furthermore, some present their information statically and some present it dynamically (for example, a map with arrows statically presents dynamic information). Some ways provide nearly simultaneous access to the world and others provide it sequentially, perhaps over long time periods. Indirect interaction requires the interpretation of symbol systems. The semiotic abstractness of symbols varies, ranging from very iconic (resembling what they represent) to very arbitrary (not at all resembling what they represent). Some symbols require scale translation for interpreting spatial or temporal scale, and some are more flexible in showing different scales. Different ways give different viewing perspectives on the world, ranging from horizontal to oblique to vertical. They vary in the precision with which they depict spatial properties, as well as the detail they provide.

Map symbols represent spatial, temporal, and thematic entities and properties of the Earth’s surface. That is, they express meaning (semantics) by referring (corresponding) to a portion of the Earth’s surface and the events and features found there. Cognitive map research asks questions such as how information is perceived from maps, how it is interpreted and stored in memory, how it is used to reason and solve problems, how it is used to guide behavior, and so on. An ongoing academic debate explores the degree to which map skills (interpreting, using, making maps) have an innate basis, and how maps should be incorporated into early education. But there is no question that the sophistication of different map skills develops over childhood, and in fact, that many adults struggle with various map skills. The meaning of largely arbitrary symbols, such as contour lines, can be obtuse to children and adults; iconic symbols, such as green for vegetation or blue for water, can readily mislead children and adults to over-interpretation.

Understanding spatial scale and translations between scales confuses many people, as does the proper interpretation of symbol generalizations and perspective transformations (e.g., overhead to terrain-level perspectives). The appropriate interpretation of size, distance, and direction as depicted on various projections may even vex quite a few professional geographers.

Cognitive aspects of geographic information science

Almost from its beginning in the 1980s, basic research on geographic information science (GIScience) included concerns with human beliefs, communication, reasoning, and decision-making about and with geographic and environmental information. A distinct area of study within this focus on cognitive GIScience looks at cognitive and computational geo-ontologies. The traditional philosophical study of ontology concerns the ultimate nature of what exists in reality. During the 1980s and 1990s, it became widely recognized that geographic databases and information systems are essentially computational models of reality, and at the same time, mental and linguistic representations are conceptual models of reality. Cognitive GIScientists address the possibility and desirability of increasing the congruence of these computational and cognitive ontologies.

Another central cognitive research program within GIScience continued the tradition of research on map perception and cognition, but no longer dealt only with traditional flat and static pictorial maps. Digital technologies allowed a host of new forms of geographic information displays, often referred to as geovisualizations (notwithstanding the unintended limitation this implies to the visual modality). Images need not be flat but could incorporate stereopsis (three-dimensional vision); they could include
nested images and textual annotations revealed only with interaction; they could include dynamic animations; and they need not be only visual, but could incorporate sound, touch, even smell and taste! Critically, the traditional passive consumption of maps designed by others and inflexible in appearance could be replaced by interactive user control, including variable or theme selection, slider bars, brushes, zooming, panning, and more. Spatial displays of nonspatial information that resemble familiar spaces like landscapes were dubbed spatializations. This fit in with the explicit recognition that the appearance of computer interfaces functioned metaphorically, by suggesting particular familiar domains for users. Ongoing cognitive research looks at various digital interface-design issues, whether maps on the Internet, on cell phones, on eyeglass screens, or as part of location-based services.

Natural language and space

Another symbolic medium for mentally and externally representing spatial properties is natural language. The relationship of language to thought is an old question for philosophers and social scientists. It is relevant to geographers insofar as it suggests something about cultural (linguistic) variation in spatial thought, which in turn has important implications, for instance, for the design of geographic information systems (GIS) outside a monolingual context.

Linguistic expressions often include spatial content. The spatial properties of the characters, objects, places, and events that make up narratives are often described, whether size, shape, or location; changes to spatial properties, especially location, often figure centrally in the story. Other common examples of linguistic expressions containing spatiality include instruction manuals, road signs, and giving verbal route directions. GIScientists are quite interested in entering, storing, processing, and outputting verbal geographic information in systems like general-purpose GIS, navigation systems, digital libraries, tourist systems, and more.

A variety of intriguing issues concern how language expresses spatiality. Spatial terms sometimes involve a spatial scale for their interpretation, and this scale is often provided implicitly by context. We can be “near the mailbox” or “near Mumbai.” “Near” likely refers to very different extents in these two cases. Language expresses mostly nonmetric or very imprecise metric information, such as in terms like “near” and “right” (although it can express precise information). Does this reflect something fundamental about the imprecision and nonmetric nature of spatial thought, or does it simply mean that language avoids encoding our more quantitative thoughts because communication typically doesn’t need it, perhaps because metric precision is so often perceptually available in the surrounding environment? Spatiality is expressed in nearly all grammatical classes, including nouns (“top”), verbs (“approach”), adjectives (“far”), and adverbs (“nearby”). Prepositions are an especially important and interesting case, as most deal with spatial relations, and yet they are difficult to translate – even native speakers often struggle with them. We get “on a bus” but “in a car”; we refer to “the house on the lake” as well as “the boat on the lake.” Linguistic scholars study the expression of spatial relations in prepositions in different languages and the nature of constraints on preposition use. For example, it has been shown that larger, more stable objects usually serve as reference objects for figural objects; we say the “book is on the table” not the “table is under the book.”

Probably the greatest amount of cognitive research on spatiality in language of interest to geographers has been on verbal route directions (navigational instructions). In a prototypical
COGNITION AND SPATIAL BEHAVIOR

route direction exchange between two people, the direction-giver (person G) has a series of cognitive and social tasks to accomplish as part of providing directions to the asker (person A). G must identify the identities and locations of A's start and destination. In many cases, the start is the location where the direction-giving exchange is occurring, but even then, G must become cognizant of where the two of them are, and how it relates to other locations in a cognitive map that includes the start, destination, and intervening route. (In rare situations, such as at an information help desk, a person may have “canned” directions in their mind that can be expressed without accessing a cognitive map.) After G has determined the spatial relationship of the start and destination from the cognitive map, he or she must plan a route for A; this may be presented in whole or piece by piece. G must select the information advisable to communicate to A, including which turns, street names, landmarks, and so on. G must monitor A for ongoing comprehension, repeating and/or revising instructions as necessary. To complete the interaction, G and A must achieve consensus that the directions have been understood and make sense. In live situations, gestures are critical, although they are absent from navigation systems (arrows have been described as “graphical gestures”).

Cognitive geographers have carried out descriptive analyses of how people actually give and interpret route directions. Does G choose the shortest, simplest, safest, or most aesthetic route? G must judge A's ability to handle routes with particular characteristics. How many landmarks does G include, and which landmarks are those? How much metric information about distances and directions is included, if any? How do people use gestures when giving directions? Does G show sensitivity to aspects of routes that are potentially more ambiguous? What does G mean by the common expression “you can’t miss it” and what makes G say this?

Researchers have also carried out prescriptive analyses aimed at determining how directions can optimally be generated and interpreted. Researchers have suggested many ideas as to what constitutes “good” or “best” directions, but more empirical evaluation is needed. Every question that can be asked about what people say or write when giving directions can be turned into a question about whether they should say or write those things – whether they help or hinder the traveler’s thoughts, emotions, and behaviors. Substantial research suggests that directions work better when they contain explicit reference to landmarks, especially salient landmarks at critical decision points along routes, but also along routes for course maintenance, and off or beyond routes for error correction. How much metric information should be given? How valuable are corrective or overshoot statements, considering that they require extra time and effort? How much redundancy is good? Efforts to automate direction-giving, as in digital navigation systems, struggle with two considerable complexities. First is the substantial difference among individuals and groups of individuals as to what is optimum. A prominent example of such differences that has received considerable empirical support is the distinction between route thinkers and survey thinkers. As in the theories of spatial microgenesis discussed earlier in this entry, route thinkers reason about the environment in terms of linear sequences of places and simple turns, such as right or left; their thought is one-dimensional and largely nonmetric. Survey thinkers reason about the environment in terms of a two-dimensional layout of places that supports spatial inferences directly between places, even if the thinker has not previously traveled directly between those places; their thought requires more metric distance and direction.
knowledge. As well as this and many other potentially relevant individual and group differences, a second complexity in determining optimal route directions is the large difference between particular places and routes as to what directions will work best there. This depends on a variety of factors, such as spatial scale, street geometry, street signage, architectural style, and topography.

SEE ALSO: Behavioral geography; Cognitive geoengineering; Geographic information science; Ontology: theoretical perspectives; Qualitative spatial and temporal representation and reasoning; Routing and navigation; Spatial concepts; Spatial thinking, cognition, and learning; Time geography and space–time prism; Visualization

References


Further reading


COGNITION AND SPATIAL BEHAVIOR


Cognitive geoengineering

Martin Raubal
ETH Zurich, Switzerland

Cognitive engineering

Donald Norman (1986) invented the term “cognitive engineering” in an effort to integrate cognitive and computer science approaches to the design and construction of machines. More broadly, cognitive engineering is referred to as an interdisciplinary approach to the analysis, modeling, and design of engineered systems or workplaces where humans and technologies operate together to achieve system goals (Lee and Kirlik 2013).

According to Norman, cognitive engineering is a type of applied cognitive science. When people interact with different everyday things, such as telephones or doors, one often notices a discrepancy between psychological user variables and physical system variables. The psychological user variables include goals, intentions, concepts, and spatial and cognitive abilities. The physical system variables describe the state of the system. During task performance users must interpret the physical system variables in the context of their psychological goals and translate their psychological intentions into physical actions upon the system. The goal of cognitive engineering is to bridge the so-called gulf of execution and evaluation (Figure 1), which results from the differences between user and system states in terms of form and content. This gulf can be bridged from two sides: either the system designer moves the system closer to the user by finding better matches to his/her psychological needs, or the user bridges the gap by approximating the description of his/her goals and intentions to the system’s language. In his account of cognitive engineering as a new discipline, Norman focused on computer design in general and on the design of user interfaces in particular. A major viewpoint in the analysis was that different users may require different interfaces, even when performing the same tasks and working with the same system. Norman therefore advocated a user-centered system design, which starts with the user’s needs regarding a particular problem.

Cognitively engineered geoservices

The goal of cognitive geoengineering is to design geographic information systems and services (such as location-based services) based on the principles of spatial cognition, human communication, and reasoning (Raubal, Mark, and Frank 2013). It is an extension of the cognitive engineering approach into the domain of geographic information. Cognitive geoengineering is an interdisciplinary endeavor, involving the disciplines of geographic information science, cognitive science, computer science, and engineering. A special focus lies in human–computer interaction based on the integration and processing of the spatial and temporal aspects of phenomena. The field of cognitive geoengineering is motivated by the belief that useful and usable solutions to people’s geospatial problems can only be found by considering the cognitive abilities and strategies people bring to the spatiotemporal problem-solving process.
Geoservices are unique in the way they use data, which are related to locations in space and time, and the potential ways of processing the data with respect to spatial locations. This leads to increased complexity during reasoning with and analysis of data. Resolving the discrepancy between psychological user variables and physical system variables in the geodomain goes beyond the user interface level, because space and time are integral elements of reasoning processes, such as when people need to agree on spatiotemporal concepts or perform various kinds of spatiotemporal reasoning. People use different conceptualizations of space and time depending on the context, and these conceptualizations need to be matched between users and systems. For example, in transportation, the semantics of the term “road” can be depicted as an open way of travel whereas, in a wildlife habitat, “road” may represent a potentially dangerous border for various species. Semantic heterogeneity, that is, two contexts that lead to different interpretations of the same information (Wache et al. 2001), can lead to erroneous decisions when geoservices are used. Results of a spatiotemporal analysis can be correct for the data model of the dataset, but may not meet the user’s expectations and can therefore lead to wrong outcomes. An example using topographic data for noise abatement planning (Lutz, Riedemann, and Probst 2003) is shown in Figure 2. When determining the noise effect of roads on residential areas, those roads touching or crossing these areas must be identified. A user may have the concepts of roads and residential areas as illustrated on the left in Figure 2, whereas the system uses the representations as shown on the right. If the user is not aware of the system designer’s conceptualization she might assume that roads overlap residential areas and will thus use the dataset as input for an intersect operation to find roads crossing residential areas. Based on the system’s representation though, the user will not find any roads by doing so.

Spatial reasoning and decision-making in a spatiotemporal context include characteristics which must be accounted for during human–computer interaction. Spatial reasoning is about topology, distance, orientation, and shape with regard to objects and configurations of objects in space and time. Instead of doing exact calculations, people apply qualitative methods of spatial reasoning (Cohn and Hazarika 2001) that rely on magnitudes and relative values. The differences between wayfinding instructions as given by humans and those given by machines serve as an example in this respect. When people perceive space through different channels, they arrive at various kinds of information, which are usually qualitative in nature. Well-designed cognitively engineered geoservices will take these qualitative aspects into account in order to move closer to users.

Although also applicable to and useful for the design of static desktop environments, cognitive geoengineering has recently gained importance due to the ongoing evolution of
personalized (mobile) geoservices. Nowadays, different providers offer such services for geospatial problem-solving and by that means sell geographic information to many individual users in small quantities. Location-based services (LBS), that is, information services that are sensitive to the location of a (mobile) user, are prominent examples (Raper et al. 2007). They support users of transportation systems, inform them about the locations of nearby points of interest and friends, or help people find a new apartment or house. In the optimal case, such services are personalized, that is, they emphasize individual persons or personal details. From the perspective of cognitive geoengineering this translates into the customization and adaptation of geoservices to their users, who vary in cognitive styles, abilities, and preferences. Cognitive research helps to understand how individuals and groups of people differ in their cognition of geographic information, and its results can be used to improve the usability and efficiency of geoservices (Montello and Raubal 2012). Based on work done in the areas of psychology, spatial cognition, and user interface design, one can distinguish between three categories of cognitive user parameters: generic, group, and individual (Raubal 2009). Designing cognitively engineered geoservices also means accounting for the consequences of age and gender differences, or the effects of different background and culture. Further explanatory variables for differences in spatial cognition as part of a comprehensive (but non-exhaustive) list are genetic constitution, physiology and anatomy, sex, education, expertise, socioeconomic status, family membership, and residential environment (Montello 1995).

Examples of cognitively engineered geoservices can be found in various domains, such as navigation, tourism (mobile guides), emergency response, geosocial networking, or location-based health. Specific techniques and methods that follow such an approach include the following:

- Implementing natural-language query interfaces to allow for a more natural human–computer interaction.
Cognitive Geoengineering

- Spatializing user interfaces to account for people’s spatial experiences.
- Designing geoservices that consider user preferences for the performance of spatiotemporal tasks.
- Applying cognitive design principles to cartographic maps and geographic visualizations to facilitate people’s understanding and knowledge acquisition.
- Personalizing wayfinding services that offer customized directions based on landmarks.
- Designing geoservices for disabled people, such as for the visually impaired.

Future research directions

The further development of cognitive geoengineering requires future research along various avenues. As mobile geoservices will continue to gain importance it will be necessary to address questions regarding mobile decision-making: How do people operate in dynamic and often complex situations? What kinds of information do they need? How can such information be communicated effectively? Studies within different domains will shed light on how people actually make decisions while on the move and how these decisions are impacted by technology and its current limitations. Researchers must not only investigate and specify spatiotemporal constraints for different environments but also consider people’s interaction with mobile devices, and perceptual, cognitive, and social processes. Bridging the gap between psychological user variables and physical system variables requires further insights into the formalization of human conceptual representations. Such representations need to be integrated into computer systems so that system views will come closer to user views. This will enhance the likelihood of facilitating human–computer interaction and delivering cognitively adequate answers to the users’ spatiotemporal problems. Novel mobile sensors, as can be found nowadays in mobile phones and personal digital assistants, will help in acquiring real-time values for context parameters, such as velocity or visibility. This data feed from sensors to mobile device will support decision-making processes. Important questions relate to filtering context data for a particular user and making this filtering process adaptive to the user’s spatial and cognitive capabilities. Emerging technologies such as mobile eye-trackers can be employed for novel interaction methods based on gaze. For example, a user’s gaze history can be utilized to enhance human–computer interaction on small display maps (Giannopoulos, Kiefer, and Raubal 2012). Finally, an important research topic concerns spatial learning: it has been argued that the use of mobile navigation services may result in users losing concentration (Parush, Ahuvia, and Erev 2007). A key aim must therefore be to ascertain how the functionality of such services can be adapted so that users achieve immediate and longer-term objectives.

See also: Cognition and spatial behavior; Geographic information science; Geographic information system; Information technology and mobility; Interoperability of representations; Location-based services; Spatial concepts; Spatial context; Technology

References

Collaborative resource management

Thomas Gunton
Simon Fraser University, Canada

Collaborative resource management is the delegation of responsibility for resource planning to a group of stakeholders who engage in a face-to-face dialogue to reach consensus agreement on planning and management decisions (Gunton and Day 2003). Although it can be structured in many ways, it normally commences by government appointment of a group of stakeholders to a planning committee that is given responsibility for a specified task such as preparing a land and resource management plan for a specific geographical area. The stakeholders typically develop their own terms of reference and operating principles consistent with the general direction provided from government. The process is managed by a facilitator and supported by professional experts who work under the direction of stakeholders to undertake joint fact finding and analysis to support stakeholder deliberations. The plans generated by collaborative processes are, in most instances, submitted as recommendations to be ratified by governments who retain their legislative and fiduciary obligation for final decision-making.

Collaborative resource management is emerging as the preferred model of planning in many jurisdictions in North America, Europe, and Australia by agencies such as the US Forest Service, the US Environmental Protection Agency, and many state and local planning agencies (Gunton and Day 2003; Sabatier et al. 2005). The most comprehensive application of collaborative planning to date has been in British Columbia, Canada, where collaborative planning was successfully used commencing in 1992 to develop land and resource management plans for the entire provincial land base (Gunton and Day 2003; Morton, Gunton, and Day 2012).

The collaborative model arose in response to growing opposition to the scientific model of resource management, which has been the dominant paradigm for much of the twentieth century (Yaffee and Wondolleck 2003; Susskind, Van der Wansom, and Ciccarelli 2003). The scientific paradigm relies on experts working for independent government agencies such as the US Forest Service to make decisions based on scientific principles such as sustained yield. Opposition to the scientific model is rooted in fundamental disagreements over values and goals of resource management, such as preservation versus development, that have no right or wrong answers and are beyond the capacity of the scientific management paradigm to resolve. The scientific paradigm is useful for assessing impacts of different management options but is unable to assess the validity of competing goals and values or to resolve differences among stakeholders.

Resource managers have recognized the limits of the scientific paradigm and the importance of engaging the public in resource management processes for some time. Many tools for engaging the public have been developed and tested including: notifying the public, seeking public comments, holding public meetings and open houses, appointing public advisory committees, funding stakeholders to prepare alternative plans, and using formal public hearing processes such as quasi-judicial hearings (Beierle and Cayford...
These varieties of public engagement can be divided into three basic categories: information sharing, consultation, and collaboration (Gunton, Rutherford, and Dickinson 2010). Information sharing fulfills resource managers’ statutory obligation to inform stakeholders about resource management without formally seeking stakeholder input. Consultation formally seeks stakeholder input without any obligation to incorporate the input into decision-making. Collaboration fully engages stakeholders in decision-making by an interactive process of dialogue among stakeholders to seek consensus agreement.

Advocates of collaboration conclude that the collaborative model is superior to other models of resource planning for the following reasons (Gunton and Day 2003; Innes and Booher 2010; Yaffee and Wondolleck 2003; Susskind, Van der Wansem, and Ciccarelli 2003). First, collaborative planning is more likely to produce a better plan in the public interest because it meets the diverse interests and objectives of all stakeholders by relying on consensus agreement and development of more creative options developed through interactive dialogue. Second, collaborative planning is more likely to lead to effective implementation of plans because stakeholders actively support implementation of a plan that they developed and that benefits all stakeholders. Third, collaborative planning generates additional “social capital” benefits, such as improved knowledge and skills of participants and improved stakeholder relationships that help reduce conflict and improve public decision-making.

Collaborative planning also faces challenges (Gunton and Day 2003; Frame, Gunton, and Day 2004). Collaboration requires those in control of resource management to give up their power by delegating management responsibility to other stakeholders. In many instances, resource managers may resist this change and therefore prevent a collaborative process from being created. If a collaborative process is allowed to occur, its success is contingent on participation of well-organized stakeholders representing the spectrum of society’s interests. In many cases this broad spectrum of capable and willing stakeholders may not exist. A third challenge is that the consensus rule for decision-making is difficult to achieve and may result in vague or second best solutions in order to reach agreement. The likelihood of achieving consensus is further reduced the more challenging the planning problem is and the larger the differences in values of stakeholders. Collaborative management may increase the time and cost of planning by engaging stakeholders in a complex, lengthy process.

Empirical evaluations of collaborative resource management generally confirm that the benefits outweigh the challenges (Cullen et al. 2010; Frame, Gunton, and Day 2004; McKinney and Field 2008; Morton, Gunton, and Day 2012; Scholz and Stiftel 2010; Susskind, Van der Wansem, and Ciccarelli 2003; Susskind, Camacho, and Schenk 2012; Wieble and Sabatier 2009). Evaluation studies show that consensus plans are achieved and successfully implemented in the majority of applications, collaborative planning generates social capital benefits including improved stakeholder relations, reduced conflict, and improvement of societal skills and knowledge, and collaboratively developed plans are also more likely to meet the public interest because they meet the interests of all stakeholders, are approved by democratically accountable governments, and are based on sound scientific information developed by participation of experts in joint fact finding. Overall, stakeholders who have been engaged in collaborative resource management rate the collaborative models as superior to other models of planning.
The evaluation research on collaborative resource management emphasizes that realizing these benefits is contingent on using the following best practice design and management principles (Gunton and Day 2003; Margerum 2011). First, all interests need to be represented in the planning process. To ensure that the number of participants is manageable, organizations with similar interests can be presented by one party at the main planning table. A second subsidiary planning table can be used by the common interest organizations to ensure that their delegate at the main table adequately represents their interests. Collaborative planning processes also need clear objectives, clear delineation of roles and accountability, adequate support staff, and training and financial support for stakeholders. It is particularly important for stakeholders to be supported by experts to provide sound scientific information for stakeholder deliberations through joint fact finding. Collaborative planning processes require sufficient time to reach decisions, sometimes over four years for complex management problems. Attempts to expedite the process may undermine its success. Collaborative planning requires a good facilitator and consultation through more traditional mechanisms such as open houses and public meetings with the general public, who may not be represented at the stakeholder table. Governments also need to retain their final statutory decision-making authority over planning proposals recommended by collaborative processes to provide a final review and ensure democratic accountability.

Collaborative resource management is an innovative model that shows considerable promise. Although evaluation of performance of collaborative planning is still in its early stages, the evidence to date confirms its many benefits relative to other models of resource management based on the more traditional scientific paradigm. Given the increasing popularity of collaborative planning, continued evaluation and development of best practice guidelines will be important to ascertain its strengths and weaknesses.

**SEE ALSO:** Citizenship; Civil society; Community-based natural resource management; Environmental governance

**References**


Collaborative Resource Management


Description and purpose

ACOGE is a private, nonprofit organization with four membership types: patrons (former chairmen of the association or individuals who have contributed one-time sizeable donation to the Geographer’s Home project), regular members (individual geographers or scholars from related fields keenly interested in geography), student members (undergraduates in geography), and institutional members. The promotion of geography as a science as well as a professional career is central to ACOGE’s undertakings. One of its practical purposes is to provide a friendly setting for intellectual interaction where geographers can introduce their research and discuss ideas concerning developments in geographical theory and practice.

Journals or major publication series

*Semestre geográfico*
*E-Boletín-ACOGE. http://www.acoge.net/e-Boletín.html*

Current activities or projects

The association’s principal activities concentrate on organizing the Colombian Congress of Geography, a biannual meeting that has been held twenty times during the past four decades (in one occasion, in 1977, this meeting was held in conjunction with the Conference of Latin Americanist Geographers (CLAG), an USA-based international organization). Since 2012, all activities of the association have been concentrated in a five-year plan designed to commemorate its semi-centennial, which takes place in 2017. Several initiatives are underway including the Geographer’s Home project; research and writing about the history of modern geography in Colombia; a comprehensive geographic treatise of the country and other books; and the Convention on Geographical Education, among others activities. But the main project is the organization of the first international convention of Spanish and Hispanic-American geography, Congreso Panhispánico de Geografía, which will be held in mid-2017 in Tunja, the city where ACOGE was founded. This project has earned the endorsement of the Spanish, Mexican, and other geographical communities, aiming at bringing interested scholars together to explore and set forth strategies to further develop geography in the Spanish-speaking world.

Brief history

Three scholars represent the phase of pre-modern geography in Colombia: Francisco José de Caldas (1768–1816), Agustin Codazzi (1793–1859), and Francisco Javier Vergara y Velasco (1860–1914). Caldas, a self-educated geographer of the cosmographic school, is generally regarded as a scholar par excellence.
COLOMBIA

and the “father” of Colombian geography. He became personally acquainted with Humboldt in Quito in 1801. The land surveys of Codazzi in the mid-nineteenth century confirmed a trend which led to the founding of the Colombian Geographical Society (1903) and the Geographic Institute Agustin Codazzi (IGAC) (1935) as government-sponsored agencies.

Academic geography at university level, however, was a late development. Since the mid-1960s, the Colombian Pedagogic and Technological University (UPTC) collaborated with the IGAC to achieve this goal. This process started in Tunja in 1967 when the UPTC hosted a small gathering of would-be geographers, conveyed there by Dieter Brunnschweiler, a Fulbright visiting scholar from Michigan State University, and Hector F. Rucinque, his Colombian liaison. The outcome of that meeting was the creation of the Association of Colombian Geographers (ACOGE).

For over four decades, ACOGE has provided leadership to build a geographic profession and to promote geography as a modern scientific discipline. Graduate seminars were taught in the 1970s by American geographers led by C.W. Minkel, a Michigan State University professor. Minkel helped students of those seminars to continue their graduate training in the USA; one of them, Hector F. Rucinque, already holding a Master’s degree from Wisconsin-Madison, would become the first Colombian ever to earn a PhD in geography (1977).

The UPTC and IGAC founded a highly qualified Master’s program in Bogotá (1983), staffed with six geographers holding doctoral degrees. Later on, but not under Rucinque’s leadership, a doctoral program was established within the UPTC-IGAC project. In general, the UPTC-IGAC formed the geographers that have built a thriving geographical career in this country, through subsequent individual projects started in the 1990s in Bogotá, Pasto, Popayán, Montería, and Cali. There are schools of geography in three private colleges as well.

Submitted by Hector F. Rucinque
Colonialism, decolonization, and modern imperialism

In her account of imperialism and the rise of totalitarianism in Europe, Hannah Arendt (1951) saw the three decades from 1884 to 1914 as the period in which the bourgeoisie became “politically emancipated.” It was an era in which the bourgeois ruling class, hitherto contained within the nation-state – “it had left all political decisions to the state” (Arendt 1951, 3) – realized that the nation-state itself proved unfit for the further growth of the capitalist economy. Expansion is everything, said the great nineteenth century imperialist Cecil Rhodes. It was precisely expansion as a permanent and all-encompassing aim of politics that, says Arendt, “is the central idea of imperialism” (Arendt 1951, 5). Imperialism, or rather modern imperialism, was born at the moment when capitalist production and an expansionary capitalism system ran up against national limits to its growth. As Arendt put it, “the process of the never ending accumulation of power necessary for the protection of never ending accumulation of capital determined the ‘progressive’ ideology ... that foreshadowed the rise of imperialism” (Arendt 1951, 23).

The original sin of primitive accumulation required by capitalism demanded other sins, and imperialism bore its name. Unlike true imperial structures, where the imperial power integrates conquered territories, the late nineteenth century imperium involved a sort of trusteeship for conquered peoples in which national metropolitan and colonial administrations were separate though bound together by relations of control and power. This classical age of empire, as it is often called, in which one-third of the globe was colonized by the competitive drives among expansionary nation-states of Europe (and Japan), stood awkwardly in relation to nationalism. On the one hand colonization reflected new nationalist aspirations of imperial powers reflected in new forms of class alliance (what Arendt called an alliance between capital and the mob); on the other hand when the nation-state appeared as an imperial conqueror it “aroused national consciousness and desire for sovereignty among the conquered people” (Arendt 1951, 7).

Politically, an empire is a geographically extensive group of states and peoples (ethnic groups) united and ruled either by a monarch or by a political oligarchy. In geopolitical terms, empire has denoted very different political and institutional forms of territorial rule: at the strong end is the extensive British Empire of the nineteenth century and at the weaker end the Holy Roman Empire, which began in the eighth century but at the time of its final dissolution in 1806 was more a city-state than a territorial empire. Somewhere in between is the Spanish empire of New Spain of the sixteenth century. These imperial forms reflect different expressions of globalization that vary in their extensivity of reach, their intensivity of interaction, their velocity of exchange and communications, and
the potential impact of geographical integration. Imperialism and empires, whether of sixteenth or nineteenth century provenance, express transcontinental flows and networks of activity, interaction, and exercise of power. Arendt is, of course, describing a distinctively modern form of empire, namely the creation and maintenance of an unequal economic, cultural, political, and territorial relationship between states and their dispersed colonies based on domination and subordination. Colonialism refers to the establishment and maintenance of rule for an extended period by a sovereign power over a subordinate and alien people that is separate from the ruling power. The colony is a territory under the immediate political control of a metropolitan-state and its emissary, so to speak, the colonial state. Colonialism is one form of, and also a consequence of, imperialism: it is defined by the means by which disparate parts of the world are subordinated to the typically national interests, drives, and dictates of a separate and distant imperial center (often in competition with other imperial centers over markets, resources, and territory). For colonies in antiquity, city-states would often found their own colonies. Some colonies were historically countries, while others were territories without definite statehood from their inception. In modern forms of empire, the metropolitan-state is the state that possesses the colony and over which it extends its sovereign power.

Colonialism

Modern colonialism in its various norms and forms extended over the period from the fifteenth to the twentieth centuries. In the New World, which had been subjected to Spanish, French, Portuguese, and Dutch colonial rule in the First Age of Colonialism, the first wave of decolonization occurred in the eighteenth century. In this regard, the so-called Classical Age of Imperialism in the last quarter of the nineteenth century was short, the first decolonizations of the second wave being achieved after the end of World War II. The two cycles of imperialism both concluded with a limited phase of decolonization followed by the rapid collapse of empires and an irresistible push to political independence. What neocolonialism as an analytical category offered was a critical sense of how this process occurred and whether it was in any sense complete.

These phases of colonization and decolonization constituted not only the making of the “West and the Rest” (Hall 1992) but also of a distinctively new and different modern world system. The new world economy that arose in the sixteenth century and thereafter differed from earlier empire systems because it was not a single political unit. Empires depended upon a system of government which, through commercial monopolies combined with the use of force, directed the flow of economic goods from the periphery to the center. Empires maintained specific political boundaries, within which they maintained control through an extensive bureaucracy and a standing army. Only the techniques of modern capitalism enabled the modern world economy, unlike earlier attempts, to extend beyond the political boundaries of any one empire. The definition of a modern world system by Immanuel Wallerstein (1977) is intended to combine the developments of the different societies since the sixteenth century in different regions into one collective development. The main characteristic of Wallerstein’s definition is the development of a global division of labor, including the existence of independent political units (in this case, states) at the same time. It is divided into core, semi-periphery, and periphery and is ruled by the capitalist mode of production.
and in normative terms a free market. There is no political center, compared to global empires like the Roman Empire; instead, the capitalist world system is identified with the global market economy.

If colonialism is an enduring relationship of domination it entails both forms of dispossession (the sovereignty of a colonized people and their indigenous political structures is compromised, typically violently, and in ways which may involve enslavement, forced removal, and subjugation) and the exercise of power by a colonizing minority that is not only convinced of its own racial and cultural superiority but also deploys forms of rule that simultaneously include coercion, persuasion, collaboration, and compromise through colonial state apparatuses. The term colonization refers to the raft of strategies and projects – from exploration, mapping, to violent conquest, pacification, occupation, state building, forms of rule, settlement, market building, mobilization of labor – entailed by typically quite different colonial enterprises in different regions in various historical epochs. Colonization, in sum, entails the constructions of cores and peripheries within a global totality, what Franz Fanon described as a world cut in two and a colonial world divided into compartments (Fanon 1963).

A useful way to think about the variety of forms of modern colonialism is by understanding the conditions of possibility for the existence and reproduction of a colony (Bernstein 1990): namely, the colonies have to be acquired territorially (often as part of a competitive inter-imperial rivalry), they have to be financed (there is always the question of cost, whether through standing armies or administrators), they have to be governed (“stabilized” through forms of direct, indirect, or autocratic rule), and they have to be made profitable (sources of accumulation and resource acquisition, and forms of labor mobilization). These conditions can be met, of course, in a variety of ways – a fact that highlights the considerable diversity of forms of colonial political economy (even within contiguous regional or continental territory). The British–ruled settler economy of Kenya in 1930, for example, looks rather different from King Leopold’s Congo in 1900.

The differing structures of colonial political economy can be mapped by using the tripartite structure developed by Samir Amin (1978) in relation to African colonies. He distinguished settler, peasant trade, and concessionary colonies. In colonies in which European settlers were central to the political economy (Kenya, for example), colonial states typically ruled directly in order to protect white interests while African cultivators were displaced and dispossessed (typically shunted into native reserves). Labor was mobilized through a combination of wage, smallholder, and tenant (squatter) labor forms, and landholding inequality along racial lines was its hallmark. The peasant colonies (Nigeria, for example) were distinguished by indirect rule (local African elites were converted into colonial bureaucrats) in which African cultivators (peasants) were effectively left intact and presided over by their own indigenous elites (so-called customary rule and law). African and European traders and intermediaries used taxes and the cash nexus to acquire the resources (cotton, cocoa, groundnuts) they needed for European markets and industries. Labor was for the most part drawn from the household. The concessionary or mine economy (South Africa) was dominated by the extractive sector in which foreign capital and migrant and coercive labor systems were distinguishing features. White and capitalist industrial interests were combined in a racialized form of accumulation, called apartheid in the case of South Africa, but the broad elements of the political economy can be
found in other colonial mine economies such as the Belgian Congo. Each of these differing colonial political economies possessed differing sorts of political dynamics and trajectories, and each left rather different postcolonial legacies.

Analyses of the reasons behind colonialism in its modern forms is, of course, a source of considerable intellectual debate and contention. The vast literature on theories of imperialism dating back to the nineteenth century (Karl Marx, J.H. Hobson, Lenin, Luxemburg, Weber, Schumpeter, and so on; see Brewer 1990 for a review) and continuing up to contemporary accounts of the post-9/11 American imperium and the war on terror, all address the complex social and political forces, all operating within a global political economy, that are in play in various colonial and imperial projects (Harvey 2005). To take, for example, the case of the late nineteenth century scramble for Africa, a raft of “driving forces” of vastly different theoretical persuasions has been identified by scholars: commercial competition and inter-imperialist rivalry, the impact of global economic recession, the demands of the so-called “second industrial revolution,” the limits of informal empire, intraregional class antagonisms, social atavism, the expanding role of finance capital (“over accumulation”), the civilizing mission, the spillover effects of the South African mineral boom, the drive of expansionary capitalists such as Cecil Rhodes, and the “political emancipation of the bourgeoisies.” All have been invoked as key to the rise of the late-nineteenth-century African partition embodied in the Treaty of Berlin in 1884 (Hobsbawm 1989). What is clear is that there are multiple imperial or colonial trajectories in which the relations between colonizer and colonized — between metropole and periphery — were complex, unstable, and to a certain degree open. That is to say, colonialism — and the operations of colonial states — was a dialectical engagement: colonial powers were in some cases weak and unable to obtain what they wanted; in others the very fact of colonization gave birth to anti-colonial nationalism; for some countries the trajectory was relatively short (in world time) punctuated by serial crises; and in others trajectories were long and relatively stable. The key themes of the quite different colonial projects have been well articulated by Jane Burbank and Fred Cooper (2010) in their book *Empires*, which describes differing colonial “trajectories.” They focus on the politics of difference, imperial intermediaries, and local compradors, the intersections among and between empires and colonies, imperial imaginaries, and repertoires of power and contention.

The danger is then to see colonialism as monolithic, in reducing the experience to vapid generalization (for example, seeing Indians or Kenyans as simply colonial subjects or as national or proto-nationalist actors). An alternative approach pursued by the so-called Subaltern School (Guha and Spivak 1988) sees colonialism as a hegemonic project fragmented as colonial rule attached itself to local idioms of power. From this experience characterized by hybrid forms of identity, of blurred boundaries, and contradictory practices, the process of colonization must necessarily look more complex than imperial domination and peripheral subordination. The colony is a particular sort of space in which, to quote Edward Said, the struggles are not only about soldiers and canons but “about forms, images and imaginings” (Said 1993, 7). Place and identities — class, ethnic, racial, gender — were all inscribed and contested in the colonial crucible, producing unstable forms of hybridity, intermixing, and dialectical self-definition. All colonial projects had their forms of “orientalism” (i.e., cultural discourses), their own cultural technologies of rule, and their
own forms of discourse, practice, and investigative modalities (Cohn 1996). These imaginative geographies (Gregory 1994) of the colonial world have produced a vast and often contested scholarship – much operating under the sign of postcolonial theory – that attempts to understand not only the complex ways in which culture is generative under conditions of colonial oppression, but also how other colonial histories can be told that depart from the totalizing vision associated with the inexorable march of capital and the advancing frontiers of western domination. Under its most robust analyses (Goswami 2004), the very nature of the colonial project takes on not simply a contingent form but exposes the contradictory ways in which colonial state space is produced and reproduced.

Decolonization

Decolonization refers to the process, often long, tortuous, and violent, by which colonies achieve their national aspirations for political independence from the colonial metropolitan power. Decolonization can be understood as the period of later colonialism but implicit in the notion of neocolonialism is the idea that decolonization was incomplete or perhaps aborted to perpetuate a form of metropolitan or imperial hegemony. The first challenge to the first wave of colonization came in 1776, as British North American colonies declared independence and subsequently the French were overthrown in the Haitian Revolution (1791–1804). While Britain maintained its Caribbean and Canadian colonies, the Napoleonic upheavals in Europe so weakened Spain and Portugal that European settlers from Mexico to Chile expelled their imperial masters. By 1825 the Spanish and Portuguese empires were dead. In the subsequent 115 years up to World War II, decolonization was limited to Cuba in 1898 and two groups of British colonies: the white settler colonies (Canada, Australia, New Zealand, and South Africa) granted internal autonomy and finally full sovereignty in 1931 (although full constitutional powers came later), and Egypt and Iraq after World War I. World War II marked the death knell for European colonization: India’s separation from the British, Indonesia’s from the Dutch, the remaining Arab mandated territories and Indo–China from the French. The independence of Ghana in 1957 marked an avalanche of liberations in Africa, although the process was not complete until 1990 (Namibia). Between 1945 and 1989, over one hundred new independent states were created.

Decolonization is a process marked by the achievement of political independence but the duration, depth, and character of decolonization movements vary substantially. In some African colonies, colonization was barely accomplished and resistance movements of varying degrees of organization and institutionalization attended the entire colonial project. In other cases, an organized anti-colonial and nationalist movement came late, accompanied by a rapid and hastily assembled set of political negotiations in which it is clear that the metropolitan power wished to hand over the reins of power with utmost expedience (Nigeria). In others it took a war of liberation, a bloody armed struggle by leftist guerillas or nationalist agitators pitted against white settlers or intransigent colonial states (as in Laos, Vietnam, and Zimbabwe).

One of the problems with analyzing decolonization, as Fred Cooper (1997, 6) notes is that the story “lends itself to be read backwards and to privilege the process of ending colonial rule over anything else that was happening in those years.” It should also be said that any account of decolonization – or for that matter neocolonialism – presumes an account, or a theory,
of colonialism itself: top-down interpretations take colonial projects at face value whereas the nationalist account denies any reality to the goal of modernization that the colonial state purported to bring. In general, decolonization is seen as either: (i) self-government as an outcome of negotiated preparation and vision from above by a colonial state apparatus; or (ii) as a nationalist triumph from below in which power is wrested (violently or otherwise) from recalcitrant colonizers (Burbank and Cooper 2010). In practice, decolonization was an enormously complex process involving something of each, shaped both by the peculiarities of colonialism itself and by the particular setting in world time in which the nationalist drive began.

There are two forms of decolonization that rest on what one might call nationalist triumph. The first is built upon social mobilization, in which a patchwork of anti-colonial resistances and movements (many of which are synonymous with colonial conquest itself) are sown together into a unified nationalist movement by a western-educated elite (Malaysia, Ghana, or Aden). Mobilization occurs across a wide and eclectic range of organizations – trade unions, professional groups, ethnic associations – bringing them into political parties and propelled by a leadership focused on racism, on liberation, and on the sense of national identity of the colony given its own history and culture. The second is revolutionary – Franz Fanon was its most powerful and articulate spokesman – in which the vanguard is not western-educated elites or indeed workers, but the peasants and lumpenproletariat. It rested upon violence and rejection of any semblance of neocolonialism. Decolonization rejected bourgeois nationalism (of the first sort); rather, as Fanon put it, “the last shall be first and the first last. Decolonization is the putting into practice of this sentence” (Fanon 1963, 30).

Both views depict nationalism as subsuming all other struggles and, hence, obscure and miss much history; both posit a true cause, as Cooper (1997, 7) puts it, in which there is little truck with opposition. Mamdani’s (1995) enormously influential book on Africa makes the important point that decolonization posed the possibility of breaking with traditional European colonial indirect rule (what he called “decentralized despotism”), in which African custom granted enormous powers to local systems of traditional (and therefore cultural) authority, to develop instead a sort of civic nationalism, in which cultural politics did not play a key role. Most African states continued the colonial model in which African colonial subjects were granted a racial equality and citizenship rights, but in which “indigenes” were simultaneously a sort of bonus. In much historiography, the nationalist road to self-government tends to take for granted the depth and appeal of a national identity. It is precisely the shallowness of such nationalisms in the postcolonial period that reveals how limited the simple nationalist account of decolonization itself is. In practice, decolonization occurred in the context of all manner of contradictions and tensions between the national question and other social questions.

There is another narrative of decolonization, which has a singular vision but from the side of the colonial state. It was the colonial bureaucracy, long before nationalist parties arose, that shaped self-government on a calculus of interest and power derived from an older conception of colonial rule (New Zealand and Canada) as a stepping stone to independence. In this view Africa, by 1947, had already been set on the road to decolonization in spite of the fact that the Colonial Offices typically saw early African leaders as schoolboys or demagogues (Cooper 1997).
Another version of the dirigiste theory is rendered through the cold calculation of money and cost. It was the decision-making rationale of accountants estimating costs and gains against the backdrop of weakened metropolitan economies after World War II that sealed the fate of the colonies.

Neocolonialism

A critical understanding of decolonization and the political forces that are assumed all point to the knotty problems surrounding political independence and neocolonialism. Kwame Nkrumah, the anti-colonial leader and first president of Ghana, defined the neocolonial condition as: “modern attempts to perpetuate colonialism while at the same time talking about ‘freedom’” (Nkrumah 1965, 41). In his formulation it was the “last stage” of imperialism because it emerged in the context of the Cold War (in which the third world became the site of proxy anti-communist struggles) and the deepening militancy of ex-colonial territories. This was no longer “naked colonialism” but rather more invisible modalities – economic, ideological, political, and cultural – in which colonial exploitation was perpetuated. For Nkrumah, control by imperial powers over nominally independent states was achieved through new forms of corporate and, especially, financial forms of capital, by a psychological dependency among third world elites, by the effects of what he called “limited wars,” and by the capitulation of African, Latin American, and Asian leadership to the hegemonic forces of the former colonial states.

While Nkrumah first gave voice to the concept of neocolonialism as an analytical device, the substance of the term was an integral part of African anti-imperialist theorizing. Leopold Senghor, Alioune Diop, and Albert Memmi all articulated similar sorts of ideas in the 1950s and 1960s. It was Franz Fanon, in particular in his searing account of Algeria and what he called “the pitfalls of nationalist consciousness,” who laid out a profound map of the forms and norms of the neocolonial condition. The Algerian war of liberation played an absolutely indispensable role in the formation of neocolonialism as a category of thought, and French intellectuals – most obviously Jean Paul Sartre – also contributed to the theorization of, in his case, French neocolonial rule and the deployment of violence (the assassination of Patrice Lumumba in particular) as its particular instrumentality.

The term neocolonialism has fallen out of fashion since the 1970s. It is quite true, however, that the broad thrust of critical development work from the late 1960s onwards – dependency theory emerging from Latin America, the calls for a new international economic order, unequal exchange theory, and the French Marxisant modes of production debate, and world systems theory (Wallerstein 1977) to take a handful of the most prominent trends in development discourse – all spoke of relations of exploitation between former colonial states and the advanced capitalist core. Indeed, any theory of imperialism, almost by definition, presumes the exercise of powerful forms of dependency among first and third world states, and to this extent the analytic core of neocolonialism identified by Nkrumah has been central to any critical account of global political economy.

More recently in the so-called anti-globalization protests – the movement of movements – the critique of corporate power, of the expanded role of finance capital in the impoverishment of the Global South, and the imperialist role of multilateral development institutions such as the IMF and World Bank, is entirely consistent in substance with Nkrumah’s account. The
revivication of neocolonialism is especially clear in the World Social Forum. First convened in January 2001 in Porto Alegre, Brazil, the World Social Forum (WSF) is an annual meeting held by members of the so-called anti-globalization movements – sometimes dubbed the “movement of movements” – to provide a setting in which global and national campaigns can be coordinated, shared, and refined. It is not an organization or a united front but “an open meeting place for reflective thinking, democratic debate … by groups and movements of civil society that are opposed to neoliberalism and to domination of the world by any capital or any form of imperialism” (http://www.worldsocialforum.info/). The WSF has grown substantially from its first meeting in Brazil.

Neocolonialism more than anything was a key marker of a certain sort of 1960s third world nationalism. Neocolonialism was a by-product of its largely African and Marxist origins, of the Bandung moment, and of the contradictions of decolonization as it unfolded in the wake of World War II and in the heart of the Cold War. For a while it was central to thinking of the theories of imperialism within a Marxist frame but it fell out of intellectual fashion. The so-called neoliberal counterrevolution and the devastating consequences of structural adjustment and economic reforms on a large part of the Global South (and Africa in particular) have given neocolonialism a shot in the arm, seen in the various forms of anti- or alternative globalization movements. There are also close affinities between neocolonialism in the Nkrumah sense and postcolonial theory, but the latter always distanced itself from the sort of determinism and historical telos that accompanied so much of Leninist and orthodox Marxist accounts of empire.

SEE ALSO: Geopolitics; Globalization; Nationalism and geography; Nation-state

References


Further reading


Color theory

Amy L. Griffin
UNSW Canberra, Australia

Color theory includes several sets of ideas about how colors can or should relate to each other. Mirroring the cartographic design tradition more broadly, these ideas draw from both scientific and artistic investigations of color. Today, the use of color in maps is commonplace, a result of the increasing ease and economy with which map production technologies can create it. As a type of map symbol that easily draws the attention of map-readers, its use by map-makers can greatly affect the legibility of the map as well as map-readers’ reactions to and engagement with the map. Although it is easier than ever to include color on maps, this technological ease can be counterproductive when there is a lack of understanding among map-makers of the effects of their color symbol choices on how map-readers interpret the map.

Color serves many roles on maps, ranging from the decorative to the functional, though in practice it is often difficult to separate these roles neatly. Furthermore, some cartographic treatments of aesthetics have been normative rather than discriminating, such as Arthur Robinson’s (1952) contention in The Look of Maps that the art in maps has a single purpose: to heighten the efficiency of the map’s communication. Yet, part of the “art” in cartography lies in the map-maker’s capacity to develop his or her own expressive style for effectively conveying information (i.e., a house style), and color plays an outsized role in this process. The challenges of choosing colors that are both aesthetically pleasing and effective for communication of spatial relationships are heightened by the many uses to which maps may be put and the tasks they may be asked to support. Some of the major functions of color in map design include its use to differentiate different map features or feature types; provide emphasis to particular map features; indicate all map features of a particular type; make an association with a real-world feature or another map element; signify a change in trend or condition; and capture the map-reader’s attention and improve his or her memory of the information contained within the map.

Human perception of color

What is colloquially known as “color” is produced by the interaction of the human visual system and electromagnetic energy. Electromagnetic energy can be scientifically described using several objectively measurable spectral dimensions. These spectral dimensions include hue, which refers to the wavelength of the energy; lightness, the quantity of energy; and saturation, the range of wavelengths of light, also sometimes referred to as chroma (see Figure 1). However, human perception of color cannot be perfectly summarized with these three spectral dimensions, as light with the same spectral characteristics can be experienced differently by different people and under different conditions (e.g., levels of light).

Fundamentally, the effective use of color in maps relies on a solid understanding of the human visual system and its particularities. Human color vision relies on a combination of
COLOR THEORY

Figure 1 Differences in the spectral reflectance characteristics of electromagnetic energy that correspond to different spectral dimensions of color, including approximate red, green, and blue (RGB) color values for each color swatch. Spectral simulations were generated using Nan Schaller’s Spectrum Applet (http://www.cs.rit.edu/~ncs/color/a_spectr.html).

a light source that emits light with wavelengths in the visible portion of the electromagnetic spectrum (ca. 400–700 nm); the eye and its sensory cells; and the part of the brain devoted to visual processing, the visual cortex. These sensory cells, two types of neurons known as rods and cones, respond differently depending on both the quantity and wavelength of light.
Rods respond to very low levels of stimulus, as little as one photon of light. Cone cells, on the other hand, require much brighter light to fire. This is the reason that it is not possible to see in color in very dim light. Cone cells come in three variants, short (S), medium (M), and long (L), which differ in the wavelengths of light to which they are most sensitive, though all of them respond to some degree to a range of wavelengths. Individually, they do not produce perception of the color of wavelength of light to which they are most sensitive. Rather, the perception of color is derived from the brain comparing the relative strengths of the signal from the different kinds of cones. So, for example, strong absorption by the S cones but weak absorption by the M and L cones will result in perception of the color blue. Strong levels of absorption across all three cone types, on the other hand, results in the perception of the color white. This account of color vision is known as the trichromatic theory.

Understanding that color perception is a function of the relative levels of absorption of multiple cone types leads naturally to diagnosis of the cause of color vision impairment: defective function or the absence from the eye of one or more types of cone. In other words, if the “accurate” perception of a color requires a comparison of the relative strength of signal between all three cone types, then a diminished or missing signal for one of the three cone types will lead to the map-reader perceiving a color “inaccurately.” Color vision impairment is a reasonably common disability. Though its prevalence is geographically variable, up to 8% of males and about 1% of females experience some form of the condition. It can be caused by a variety of factors, some congenital, others acquired as a result of one of several chronic illnesses, accidents, exposure to industrial chemicals, or physiological changes that the eye experiences as it ages. Therefore, it is important for map-makers to be aware of how color vision impairments may change the ways in which map-readers can see colors in their maps, and to choose colors carefully so that their maps are legible for their end users. Color confusion diagrams, which identify color hues that are likely to be confused with each other, can be used to avoid color hue pairs that may be problematic for map-readers with particular types of color vision impairments (see Figure 2). Researchers have developed a variety of software tools, such as Color Oracle or Vischeck, to assist map-makers with identifying map symbols that are likely to be confused by map-readers with different kinds of color vision impairments.

It is not just the types of functional sensory cells that a person has that influence whether he or she can see colors accurately. The spatial arrangement of cones on the retina is also important for color vision. The highest densities of cones are found in the fovea, the area directly behind the lens. While cones are present across the retina, they are greatly reduced in number outside the fovea. Rods, on the other hand, are found in most areas of the retina except the fovea. The implications of these spatial distributions for color vision are that (i) it is best in the fovea (the center of your vision where your eyes are focused), and (ii) the ability to differentiate between different colors for objects of a given size decreases as the objects move further into peripheral vision. So, if it is important that map-readers can discriminate between different colors, and the map features are small, the colors need to be more distinctly different than if the map features are large. Otherwise, the map-reader will need to move his or her eyes much more to foveate (focus on) each map feature to differentiate it from the others. In many instances, particularly for areas, because map features are tied to a geographical extent and location, it is simply not possible to make the symbols larger at the periphery than at
Figure 2  Example of confusing and accommodating renditions of qualitative color schemes, where the confusing scheme falls along a color confusion line, while the accommodating scheme crosses color confusion lines (top). The image pair at the bottom simulates what a person with deuteranopia (red–green color vision impairment) would see when looking at the confusing and accommodating schemes. The color vision impairment simulation was created using Color Oracle. Adapted from Olson and Brewer (1997, 131).
The background that surrounds a color can influence how the color is perceived. The central figures in both examples at the top are spectrally identical: a square of 30% black. However, the square on the left is perceived as being darker than the square on the right due to the tendency of a color to induce its opposite hue. Hence, the white background induces its opposite and makes the central figure appear darker. The same effect can be observed for the spectral dimensions of saturation (middle) and hue (bottom).

**Figure 3** The background that surrounds a color can influence how the color is perceived. The central figures in both examples at the top are spectrally identical: a square of 30% black. However, the square on the left is perceived as being darker than the square on the right due to the tendency of a color to induce its opposite hue. Hence, the white background induces its opposite and makes the central figure appear darker. The same effect can be observed for the spectral dimensions of saturation (middle) and hue (bottom).

A second account of color vision explains what happens to the neural signals after they leave the photoreceptors (i.e., the cones and rods) and are integrated in the opponent neurons. This theory explains the experience of a phenomenon known as simultaneous contrast, in which the surround of a color affects how it is perceived (see Figure 3). The opponent process model explains color vision as comprising the action of opposing color pairs: red/green, blue/yellow, and black/white. When one member of each of these pairs is activated, the other is inhibited, preventing, for example, the perception of a yellowish blue or a reddish green. Simultaneous contrast effects can be found in all three spectral dimensions of color: hue, lightness, and saturation. Because map symbols are rarely viewed in isolation, map-makers need to account for the effects of simultaneous contrast to reduce the likelihood that different map symbols will be confused with each other. The ColorBrewer tool (http://colorbrewer2.org) can be used to evaluate the effects of simultaneous contrast for the color schemes presented there (see Figure 4). While it is probably not possible to “control for” this effect in maps (i.e., adjust each symbol to account for its background), it is possible to select color sets in which all elements are discriminable across the range of surrounding symbols that appear in the map. The difficulty, however, of achieving this increases with increasingly small visual differences between symbols. Therefore, to ensure that all symbols are discriminable in a choropleth map, a good rule of thumb is to use no more than six or seven different classes.

**Methods for describing color: color models**

The complexities of the human experience of color notwithstanding, scientists have been
searching for ways to systematize and organize the description of color for centuries. Color models, as these organizing systems are called, are abstract systems that can be used to describe a color precisely. The complete set of colors that a given color model can produce is known as a color space, or gamut. Different color models have arisen because different methods of producing color create different gamuts – different subsets of the set of all possible colors that human vision can perceive. The CIE color model represents the complete set of colors that a standard observer can perceive under particular illumination conditions, called a standard illuminant.

One early color model with which most people will be familiar is the hue circle, which was first described by the physicist Sir Isaac Newton. The idea for organizing colors as a circle of hues had its origins in his observations during optical experiments with prisms that white light was composed of a mixture of lights of different wavelengths. A modern color model founded on similar principles of adding together light produced at particular wavelengths in different proportions, whose increments typically range in value from 0 to 255, is the RGB (red, green, blue) color model employed in many electronic devices, such as computer monitors, televisions, and scanners (see Figure 5). Because the specific hues and saturations of red, green, and blue light produced by the particular phosphors or dyes used by a device manufacturer may vary, the appearance of a color defined by a given location in the RGB color space may differ from one device to the next, giving rise to the need for color matching when colors need to be specified and seen precisely.

Another early color model that anticipated a modern color space was the hue circle described by Goethe in 1809. While Newton’s hue circle was developed by his observations of the
COLOR THEORY

properties of light, Goethe’s was derived from observations of the eye’s experiences of color, that is, human perception. In particular, Goethe was interested in the perception of the color of pigments on paper and in paintings, rather than of light emitted from a light source. Thus, his focus was the perception of reflected rather than emitted light. His model also presaged the concept of opposing colors, by developing a symmetrical hue circle in which each hue opposed another. There are two prominent modern color models that share some similarity with Goethe’s hue circle: the Munsell color space and the CMYK (cyan, magenta, yellow, black) color space. Both color spaces are based on how we perceive color from pigments rather than from emitted light.

The Munsell color space is based on perceptually uniform and independent dimensions of color: hue, lightness, and chroma (saturation). In particular, the Munsell scheme was the first to identify perceptually equal steps in these different dimensions. Of particular note is the fact that human visual perception of the change in color dimensions is not linear: a point that is perceived to be halfway on the continuum between white and black would actually have a luminance of 20% (on a scale of 0%, black, to 100%, pure white). However, because Munsell’s system is perceptually uniform, a value of 5 in his space is one that is perceived to be in the middle of white and black (his lightness scale ranges from 0 to 10). Inspection of his space reveals that human perception of color is not symmetric across the three dimensions. That is, humans can perceive more steps of some hues at a particular lightness or chroma level than at others. A major drawback of Munsell’s system in the context of making maps is that the Munsell color gamut is not necessarily the same as the gamut of the map production system that is used to produce a map. For example, while there are translators that produce approximate RGB values for a particular Munsell color, the RGB gamut is restricted (as demonstrated in Figure 5), so not all Munsell colors can be produced in the RGB color space.

The CMYK color space, like RGB, is a production-oriented color space that is also sometimes called process color, and is used in offset printing. This color model produces different colors by mixing together different amounts of pigments (inks) using the three hues and black. Because the inks are typically applied to a white background, it is called subtractive, as the color that is perceived is based on the wavelengths of light that are not absorbed by the pigments that are laid down on the paper (i.e., those that are “subtracted” from white). Depending on the particular printing device (e.g., inkjet printer, color laser printer, or offset printer), the color gamut may be slightly different. When maps are printed, particularly in large volumes on offset printers, the color gamut may be expanded by also adding one or more spot colors (a non-CYMK offset ink) to the printing process.

While the RGB and CMYK color spaces are most commonly used for specifying colors when map-makers create maps for digital and printed maps, respectively, the Munsell color space most closely matches the way that cartographers have conceptualized how color can be employed as a visual variable or as a building block for making meaning with maps. In his seminal description of visual variables, Jacques Bertin (1983) identified hue, which he called “color,” and lightness, which he called “value,” as fundamental visual variables. While Bertin discussed saturation, he merged it with hue, and did not treat it as a separate dimension or variable. Of particular note is the distinction that Bertin made about the use to which “color” and “value” could be put when constructing a map. His syntax (set of rules for matching visual variables to
data characteristics) identified “value” as being appropriate for ordinal information with the basic logic of “darker means more,” and “color” as appropriate for nominal information. Other cartographers have since described saturation as a third color variable, but its use as a visual variable in its own right is uncommon.

Color schemes for maps: logical relations between colors

Reference maps and thematic maps commonly use slightly different strategies for developing color schemes. While reference maps often use hue to represent differences in kind or qualitative differences (e.g., blue lines for water features such as rivers and brown lines for contours), the assignment of particular hues to particular feature types often draws on conventional associations that have been built up over centuries of mapping. So, for example, while the Amazon River may often appear brown when one looks at it owing to the suspended sediment load it contains, it is usually represented using a blue line or area on maps, depending on the map’s scale. Similar conventions have been developed for other types of features, though these may vary from one country to the next. For example, while most countries’ topographic maps represent vegetation using a green symbol, Finnish topographic maps use orange and white symbols for vegetation. This decision is driven perhaps by a desire to save ink, given that forests cover about 75% of Finland’s land mass. Lightness is sometimes used in reference maps to give an indication of the relative size of features. For example, minor tributaries of major rivers may not only vary in the width of the line used to represent them, but also be represented using a lighter shade of blue.

A typical symbolization challenge in reference maps, which commonly depict many different types of map features and many layers of information, is to place particular map features at the desired place in the map’s visual hierarchy. The use of a high-contrast color, such as a saturated red or yellow, often moves the feature up the visual hierarchy. This is one reason that major roads may be represented using a red or yellow line rather than a thick black line (see Figure 6).

By contrast, fewer conventional color associations exist for thematic maps, though there are a few well-known exceptions, such as maps of voting in elections, where a political party in a given country is often associated with a particular hue (e.g., blue for Democrats and red for Republicans in the United States). A map of unemployment rates, for example, could use any hue and communicate the spatial distribution effectively. What is particularly important with thematic maps is that the logical relationships in the data are matched to logical relationships in the colors used, and that this logic is based on what is understood about how map-readers perceive color. The ColorBrewer tool summarizes a body of research undertaken by Cynthia Brewer and colleagues on how to match the logical relationships in data with the logical relationships in a color scheme. It presents three types of color schemes that are commonly used in thematic maps: qualitative, sequential, and diverging (see Figure 7).

Qualitative schemes are those whose symbols vary in hue, while holding lightness relatively constant. This type of scheme might typically be used to map nominal data such as land-use types or majority race/ethnicity within an area. Within qualitative color schemes it is also possible to associate pairings or relationships between classes by using similar colors. So, for example, if a land-use map included coniferous forest and deciduous forest, both feature types might be represented with different and differentiable green hues. A particular challenge with designing
qualitative schemes is choosing colors that will not be confusing for readers with impaired color vision, as qualitative schemes by definition avoid the use of lightness, which may help such readers correctly infer ordinal or quantitative relationships in the data even when hues are confusing.

Sequential schemes are best used to represent data with ordinal or quantitative characteristics, where a goal of the map is to indicate how much of something is present in a given location. The classical sequential scheme is one that uses only lightness, with a gradient from white to black, but often sequential schemes use one hue in combination with lightness, for example, using a gradient from light orange to dark orange. Because a simple gradient, linearly interpolating between 0% of a hue and 100% of a hue, does not produce perceptually equal lightness steps, as illustrated by Munsell’s work, it is useful to begin by specifying a color scheme in the Munsell system, and then translating it to the production specification that is needed for the production environment the map will be made within, noting that the mismatch between the gamuts of the two specification systems may limit which colors within the Munsell space can be used in

**Figure 6** The use of color contrast to move a map feature up the visual hierarchy. Lightening and increasing the saturation of the blue used to represent the river features increases the contrast and lifts them above other linear features, such as roads or contours.
Figure 7  Examples of maps using (a) qualitative, (b) sequential, and (c) diverging color schemes. Sources: City of Ithaca, NY; UN Population Division, UNEP.
practical application. Because most hues are not highly saturated when they are very light or very dark, the saturation within a sequential scheme is often not constant, with mid-range lightness values having a higher saturation level than the end points (see Figure 8).

Many visualization software packages that can be used to create maps include what is known as a spectral scheme as a default symbolization option. While there is a logical ordering of hues in a spectral scheme (i.e., they are ordered by decreasing wavelength of light), there is not a logical ordering of lightness in the symbol set. This can prove problematic in maps of quantitative information, as it can lead readers to incorrect inferences about the nature of the spatial distribution of the mapped quantity. Researchers who work with other types of scientific visualization have also noted this problem, yet, unfortunately, it persists as a default symbol set in many applications. Despite the problems of spectral schemes, it is possible to use a subset of the spectral order to increase the information-carrying capacity of sequential schemes. A sequential hue–lightness transition that begins with light yellow and progresses either through orange to red or through green to blue can provide map-makers with a larger number of clearly differentiable classes, while preserving the “darker means more” logic of schemes based solely on lightness.

A final type of commonly employed color scheme is the diverging scheme. This scheme comprises two sequential schemes that diverge in opposite directions from a critical value, such as zero, a mean, or the eligibility threshold for a government program. The central value in a diverging scheme is light (e.g., white, light gray, or yellow), while the end points of the scheme are dark. These schemes may or may not be symmetrical (i.e., use the same number of classes on each side of the critical value).

Figure 8  Comparison of color schemes derived from the Munsell space that do (left) or do not (right) employ saturation transitions.
COLOR THEORY

Figure 9 Examples of bivariate color schemes: (clockwise from top left) qualitative/sequential, sequential/sequential, diverging/diverging, diverging/sequential.

These different types of color schemes may be combined when mapping multiple variables, to represent relationships between variables. Useful combinations might include qualitative/sequential schemes, sequential/sequential schemes, diverging/sequential schemes, and diverging/diverging schemes (see Figure 9). Great care must be taken to ensure that the visual logic of individual variables is appropriately represented in the final symbol set, and this may require fine-tuning of saturation levels to ensure that all symbols are clearly differentiable.

Color can also be employed persuasively, in order to amplify the memory of a map’s overt or less overt message. For example, colors can be chosen to deliberately emphasize particular elements on the map. High contrast color choices (large areas of black, red, etc.) draw the reader’s attention to particular areas of the map, often at the expense of others, and moreover create emotional associations with the map’s content.

While red symbolizes joy or good fortune in Chinese culture, in many Western cultures it is associated with danger. When cultural meanings of color are combined with high levels of contrast, they are particularly likely to evoke a high level of arousal and a strong emotional response. There is a reason that both saturated reds and black feature strongly in propaganda maps: it’s hard to feel threatened by a pastel country (see Figure 10).

SEE ALSO: Cartographic design; Choropleth map; User-centered design; Visual variables; Visualization; Web-mapping services

References


Further reading


COLOR THEORY


Figure 10 “The war is the national industry of Prussia”: example of the use of high-contrast color to create an impression of danger in a propaganda map. Source: Bibliothèque Nationale et Universitaire de Strasbourg; http://www.bl.uk/collection-items/propaganda-map-war-national-industry-prussia#.
COLOR THEORY


Commodification of nature

Ryan Gunderson  
*Miami University, USA*

The commodification of nature refers to the process of incorporating biophysical entities and/or information about them into economic systems for the purpose of exchanging the good or service for a profit. The notion is most commonly employed in political ecology, human geography, environmental sociology, and environmental and sociocultural anthropology. When the term “commodification of nature” is used to denote a frame for investigation or a process mediating human–nature relations, the work is typically in the Marxist tradition, whether classical, revisionist, or simply borrowing conceptual categories from the Marxian catalog. There is a direct effort in this work to denaturalize capitalism’s historically unique relationship with nature based on generalized and specialized commodity production. Emphasizing the historical character of nature commodification shows that the biophysical world has existed and can exist free from the commodity form. While all human societies have employed labor to work up nature to produce things for use and survival, nature commodification literature takes great pains to show that the current mode of production, which appropriates nature primarily to exchange for a profit, is a relatively recent social formation and, because nature commodification is historical, it is also mutable. This historical and critical approach is a unifying theme in nature commodification literature that distinguishes it from ahistorical and uncritical work on economics–nature relations. Nature commodification literature is typically directly or indirectly critical of the further expansion of the market into natural domains on environmental and/or social grounds.

The most recent integration of literature on the commodification of nature is over a decade old (Castree 2003), though it has excellent organization and typologies that are employed here in revised and updated form. This entry first highlights the theoretical origins of this area of inquiry and then outlines the basic elements and structures that comprise nature commodification. It then explores three research emphases in nature commodification literature: (i) the way commodification processes function and adapt in different domains of nature; (ii) the socioecological consequences of, and barriers to, nature’s commodification; and, more recently, (iii) the limitations of neoclassical environmental economic valuation models and their applications. It goes on to argue that there should be a scholarly effort to conceptually and theoretically link future case studies in order to achieve some level of generalizability and connect the area of inquiry.

Classical theoretical background: Marx, Polanyi, and three modern classics

Although the theoretical underpinnings of nature commodification research vary, many can be traced back to two classical frameworks put forth by Karl Marx and Karl Polanyi and/or expansions...
of these two classical frameworks, especially elaborations offered by Neil Smith, Allan Schnaiberg, and James O’Connor. Marxist-oriented social scientific research exploring the causes and outcomes of nature commodification go back as far as Marx himself. As Foster (2000) has shown, Marx has numerous applications for studying the relationships between capital and nature. Most importantly, Foster retraced a formerly underappreciated development in Marx’s analysis of agriculture. Marx incorporated insights from Justus von Liebig, a German chemist, to theorize capitalism’s role in soil fertility loss in his investigations of the town–country divide. Following England’s removal of peasants from the soil through the enclosure of common lands, rural populations were forced to flock to cities for work, leading to a lack of human waste for fertilizer and long-distance trade of food and fiber to urban centers. Marx argued that this movement, combined with the early intensification of agricultural methods, created a “metabolic rift,” or a disruption of energy and material flows, which ultimately undermined soil fertility in the long term at one end and caused an accumulation of waste in cities at the other (e.g., Marx 1977, Chapter 15; 1981, Chapter 47). Additionally, as Ricoveri (2013) has detailed, Marx’s (1977, Part 8) analysis of the early land enclosures – not only a prerequisite for the development of capitalism, but also one of the earliest movements to commodify nature on a massive scale – has contemporary applications for understanding the commodification of commonly owned environments in the Global South. Harvey (2004) too emphasizes the continuation of what Marx called “primitive accumulation” as a persistent feature of global capitalism today, and describes many of the processes detailed below as “accumulation by dispossession.” More generally, Marx’s theories of how capitalism operates and the nature of the commodity form provide the conceptual building blocks for contemporary work on the commodification of nature, whether or not this is made explicit.

A second important classical influence on contemporary research on the commodification of nature is the work of Hungarian economist Karl Polanyi, specifically his notion of “fictitious commodities.” In The Great Transformation (1944), Polanyi described “land” (which denotes commodified nature) as a make-believe commodity, meaning that the natural environment, a prerequisite for production, is treated as if it were produced for the market through labor, even though it is not (though this deserves some nuance today due to biotechnologies, see “Commodification of internal nature,” below). Land exists within and outside the production process. Polanyi (1944, 178) asserted that liberal capitalism could only develop and function because the fiction of nature as a commodity, “perhaps the weirdest of all undertakings of our ancestors,” is sustained. His argument has been influential in nature commodification literature because it is helpful for conceptualizing the utopian nature of treating nature as if it were a commodity, whether to criticize an environmental economist who attaches exchange values to biodiversity, to understand why a logging company assumes that a forest acts – or should be made to act – in accordance with market forces, or to explain why nature presents “barriers” to commodification in the first place. Further, scholars studying the commodification of nature are confirming Polanyi’s (1944, 73) assertion that allowing “the market mechanism to be sole director of the fate of human beings and their natural environment … would result in the demolition of society. … Nature would be reduced to its elements … landscapes defiled, rivers polluted … the power to produce food and raw materials destroyed.” Additionally,
Polanyi has influenced some attempts in the environmental social sciences to formulate alternatives to wholly commodified human–nature relations. Following Polanyi’s assertion that capitalism “dismembeds” the economy from society and nature (i.e., social relations and nature are subjected to economic activity), some environmental social scientists have argued that economic systems must be re-embedded in both domains for economic activity to meet human needs and operate in accordance with natural laws.

As noted in the introduction, exploring the processes and consequences of nature commodification is common in human geography and environmental sociology. Three “modern classics” from these subfields are worth noting as they have stimulated a great deal of increased inquiry into nature commodification in the last few decades (though Mann and Dickinson 1978, Kloppenburg 1988, and Cronon 1991, explored below, could certainly be placed under this “modern classics” label as well). Perhaps the most influential work on nature commodification in political ecology and human geography is Smith’s *Uneven Development* (1984). In attempting to break down the notion of nature as detached from society, Smith argued human labor, especially as organized under capitalism, clearly “defies” this dualism. Economic production is, at bottom, the “production of nature.” What makes capitalism’s production of nature unique is its global scale and “exchange value relation” to nature (see below). For Smith, capitalism is a system that reaches across the globe to commodify nature to the extent that the distinction between “first” and “second” nature becomes problematic. Likewise, Schnaiberg’s *The Environment* (1980) is the foundation of a great deal of (neo-)Marxist work on the relationship between commodity production and nature in environmental sociology. His “treadmill of production” theory maintains that the market’s further encroachment into nature is inherently unsustainable because capitalism is growth dependent, requiring more material and energy throughput. This means the overuse of resources and unchecked pollution are inevitable features of capitalist societies. A third notable recent classic is James O’Connor’s (1998) second contradiction of capitalism thesis, first developed in the late 1980s, which has inspired a good deal of research on nature commodification in both political ecology and environmental sociology.

Pulling together Polanyi and Marx, O’Connor has maintained that capital may undercut the natural prerequisites (“land” in Polanyi and a “condition of production” in Marx) necessary for further production because noncommodified environments have no immediate economic value to individual capitals. The ideas of Smith, Schnaiberg, and O’Connor are explored in more detail below after an outline of the constellation of features that characterize nature commodification.

### Aspects of nature commodification

Along with highlighting the theoretical underpinnings of nature commodification research, it is helpful to outline the basic elements and structures of nature commodification before exploring contemporary work and themes. Castree (2003) has presented a helpful typology for analytically distinguishing six “aspects” of commodification commonly used or assumed, though rarely together, in work on nature commodification: (i) privatization, (ii) alienability, (iii) individuation, (iv) abstraction, (v) valuation, and (vi) displacement. In combination, these six aspects are meant to describe the chief characteristics of nature commodification in capitalist societies.
COMMODIFICATION OF NATURE

1 **Privatization** refers to a legal title being assigned to a private entity (usually an individual or corporation) for exclusive, or nearly exclusive, use. Private ownership and use can be clearly contrasted with public, collective, or common ownership and use. Because commodity production depends on the private ownership and operation of goods and resources, privatization of a natural entity or class of natural entities is a precondition of nature commodification.

2 Much of the work on nature commodification has emphasized the role of privatization in the commodification process due to neoliberal environmental policies.

3 As commodities are produced to be sold, they must be **alienable**. This means that for a natural entity to be commodified, it must be able to be separated from its sellers to go to new owners.

4 **Individuation** is the process of partitioning the world with material and/or legal boundaries in a way that makes possible commodity production and distribution. In relation to nature commodification, this means separating a class of biophysical things from ecological contexts.

5 **Abstraction** is a process whereby distinct, individualized natural entities are grouped into categories based on similarities (functional abstraction) and/or a distinct, individualized natural entity in one place is used as if it belonged to the same category in another place (spatial abstraction).

6 Another unique aspect of nature commodification involves the leveling and reduction of qualitative differences between individualized natural entities and between categories of natural entities to quantity, specifically in the form of money (**valuation**). In theory, this makes qualitatively distinct objects commensurable for the market. All ways of valuing objects are subsumed and/or replaced by price values.

6 The last aspect of nature commodification identified by Castree (2003) is **displacement**. Displacement refers to the way in which the commodity form masks the social and material relations that brought the commodity into being (termed “fetishism” by Marx (1977, Chapter 1)), though work on nature commodification focuses on what happens to the environment behind the commodity (i.e., the ecological effects at the site of production that are hidden by the commodity).

An analytically distinct and especially important aspect of commodification, subsumed under “valuation” in Castree (2003), can be teased out when examining the (vii) **purpose** or **logic** of market-oriented production regimes. This is also stressed by Smith (1984) and Schnaiberg (1980). Marx (1977, Chapter 4) outlined this purpose in a straightforward way in his “general formula” for capital: $M \rightarrow C \rightarrow M'$ (money-commodities-more money). As elaborated in the second volume of *Capital*, Figure 1 represents the basic processes of capitalist production and exchange.

Money–capital (M) is used to buy labor power (LP) and the means of production (MP) (machinery, tools, inputs, etc.), which are...

![Figure 1](image-url)
commodities (C) in the form of productive capital. LP and MP are utilized in the production process (P) to create a commodity of greater value, or “commodity-capital” (C'). C' must be sold in the marketplace for M plus a profit (Δ). Money-capital is reinvested, reproducing the circuit. Capitalism is unique in that the purpose of economic production is to amass private wealth for further investment and that it universalizes this form of production. This is the essential distinction Marx placed between commodities (as the term is used here) and the products of labor in pre-capitalist societies. The commodity is a good produced with the sole aim of exchanging it for a profit in the marketplace. Without question, economic products in pre-capitalist societies were often exchanged, but, unlike today, the general purpose for production in pre-capitalist societies was to create goods for personal use or for a local community. Exchange was “subordinate,” as Polanyi put it. Today, with some exceptions, the reason goods are produced is to accumulate capital for the owners of the means of production for further investment. Capitalism is concerned with accumulating capital “for the sake of accumulation” and “production for the sake of production” (Marx 1977, 742), not with meeting human, animal, or ecological needs. In Marxist terminology, the purpose for production in pre-capitalist societies was to produce use values, while capitalist societies are structured around the production of exchange values. This distinction is central for understanding what is meant by the commodification of nature. When nature is incorporated into economic systems in order to be sold on the market for a profit, it has been commodified. Although these basic elements and structures of nature commodification are rarely discussed in their entirety in studies on nature commodification, they provide conceptual groundwork for understanding how commodification processes work in the varying “domains” of nature.

Domains of nature commodification

Thus far, this entry has explained the theoretical origins of nature commodification research (Marx, Polanyi, and contemporary elaborations) and the basic elements and structures that comprise nature commodification (privatization, alienability, individuation, abstraction, valuation, displacement, and the logic of capitalist production). This section will explain the way commodification processes function and adapt in different domains of nature. Although all commodities have origins in the biophysical world and/or the human body, it is important to analytically distinguish specific categories of natural entities because (i) certain entities have particular environmental and/or social consequences that deserve special attention and (ii) focusing on the material makeup of these specific entities, common in nature commodification literature, casts a stronger light on the material processes of nature commodification (Castree 2003). There are four broad and often overlapping domains of nature focused upon in work on the commodification of nature: external nature (i.e., the nonhuman, biophysical environment), internal nature, the human body, and information about nature (Castree 2003). This typology is retained here, though a fifth domain is added: the commodification of animals, an area of inquiry that has increased since Benton’s (1993) work. In the following subsections, each domain of nature is delimited and defined (external, internal, animals, the human body, and informational) along with examples from principle works or recent case studies to show how commodification processes function and adapt in these different domains.
Commodification of external nature

One domain of nature examined in nature commodification literature is “external” nature. External nature here simply refers to what is commonly called the biophysical world, the natural environment, or nonhuman nature. The commodification of external nature refers to the extension of the commodity form into formerly noncommodified natural arenas. Even if not explicitly stated, every modern study of extractive and agricultural industries engages in an analysis of an “exchange value relation” to natural entities, excluding studies of commonly held or publically owned natural resources. Yet studies of agricultural and extractive industries that draw on the commodification of nature as a frame for analysis usually stress the historical and socioeconomic forces that led to and shape the particular industry. Cronon’s (1991) now famous account of the ecological roots and consequences of Chicago’s industrial growth in the nineteenth century is representative of this approach.

*Nature’s Metropolis* shows in great detail that the expansion of a metropolitan economy (Chicago’s) played a central role in linking city and country and shaping the landscape of the frontier. Made possible by the railroad, new markets opened up for grain (Chapter 3), timber (Chapter 4), and meat (Chapter 5). These markets radically transformed the ecosystems of the Great West: a few farm plants replaced the dozens that formerly made up the prairie ecosystem, northern forests were clear-cut, and the bison of the Great Plains were almost exterminated to be replaced by pasture and, later, early feedlots. In short, the commodification of nature “had transformed prairies into wheat, forests into lumber, livestock into meat,” a harbinger for “many of the environmental problems we face today: large-sale deforestation, threats of species extinction, unsustainable exploitation of natural resources, widespread destruction of habitat” (Cronon 1991, 149, xiii).

Employing similar terminology to that used by Smith (1984), the theoretical and conceptual significance of Cronon’s work is that capitalism, via the commodification of nature, problematizes the separation of a pristine “first nature” and the artificial “second nature” created by humans. The expansion of capitalist markets transforms first nature in its own image, thereby “merging” first and second nature. “What had once been a diverse prairie landscape produced fewer and fewer species. Although the local ecological conditions of first nature continued to influence which species grew where, the economic imperatives of second nature – distance from the city, cost of transportation, supply and demand, price – played an ever more important role in determining the shape of the landscape. As the human inhabitants of Chicago’s hinterland responded to the siren song of its markets, they simplified local ecosystems in the direction of monocultures. … The merging of first and second nature was thus a shift from local ecosystem to regional hinterland and global economy” (Cronon 1991, 266, 267).

While it would be brash to level a critique of Cronon’s masterful work in this short summary, it is important to note that other scholars argue that *Nature’s Metropolis* fails to explain the underlying mechanism for Chicago’s commodification of “first nature,” as if trade itself were such a mechanism (see Walker 1994, special issue of *Antipode*). Thus, for future researchers drawing from Cronon (1991), it is helpful to supplement this drawback with some of the conceptual insights summarized above, especially the Marxian account of capital accumulation.

Commodification of internal nature

A second domain of nature explored in nature commodification literature is “internal” nature.
Castree (2003) uses the term the commodification of “internal” nature to refer to the process of strategically altering natural entities in a way that overcomes their “resistance” to complete commodification in order to become “pure” or “real” commodities. Thus, “internal” refers to a smoother “internalization” of nature into the market system through biophysical modifications via technological means (genetic modification, growth hormones, hybridization, etc.). The function of the commodification of internal nature is to change the biophysical makeup of a given natural entity for more efficient incorporation into the market to overcome natural barriers and increase productivity (see “Physical and ecological barriers to nature commodification,” below). A helpful way to conceptualize the different processes of, or rather, opportunities presented in the commodification of external nature described above and the commodification of internal nature is to distinguish the “formal” and “real” subsumption of nature. Analogizing Marx’s (1977, Appendix) distinction between real and formal subsumption of labor, Boyd, Prudham, and Schurman (2001) argue that, in contrast to the formal subsumption of nature – where extractive, nonbiologically based industries cannot manipulate the given resource to increase productivity (e.g., the diamond industry) – the real subsumption of nature, the transformation of nature into a productive force itself, is becoming the norm in biologically based industries (e.g., plant and animal cultivation) through biotechnologies.

Today, there are many examples of how the very nature of nature is altered for the market through biotechnologies. Perhaps the most famous text that provided groundwork for examining the internal commodification of nature is Kloppenburg’s First the Seed (1988). Addressing the theoretical side of the Marxist “agrarian question” (how does/will capital deal with and transform agriculture), Kloppenburg (1988) focused on how enduring pre-capitalist social relations are broken down by “binding” farmers to inputs – particularly, to commodified, hybridized seeds. He argued that the seed had “resisted” commodification for a long time, but through privatization laws, the barring of public agricultural scientists from providing cheap or free new seed varieties, and hybridization by private research labs, the seed has fully “submitted” to the commodity form. Castree (2001) has depicted Kloppenburg’s argument within Marx’s basic model of capital accumulation (see Figure 2).

![Figure 2](image_url)

**Figure 2** Seed firms, farmers, and capitalist accumulation. Source: reproduced with permission from Castree 2001. Malden: Blackwell Publishers, p. 200.
Hybrid seeds are produced by seed companies with the help of scientific research into nature (SRN) and sold to farmers as an MP. These hybrid seeds are sterile and, thus, must be purchased each growing season with part of the given farmer’s profits (Δ*).

In relation to nature commodification research, Kloppenburg’s classic shows how science and capital interact to fundamentally transform the makeup of natural entities, as well as global social relations, for the accumulation of private wealth for reinvestment. More examples of internal commodification can be found in the genetic engineering of foods and industrial animal agribusiness “husbandry” practices, described shortly. However, as Castree (2003) notes, technical attempts to overcome “incomplete commodification” through internalization can produce risks (see “Physical and ecological barriers to nature commodification,” below).

**Commodification of animals**

The commodification of animals is a third domain of nature studied in nature commodification literature. Although animal commodification falls under the categories of “external” and “internal” nature commodification, there are two immediate reasons to distinguish animal commodification from the former two categories. First, the unique physical and social features of animals compared to the rest of nature, not to mention the differences between animal species, demand analytical division, especially when examining the unique “barriers” animals pose to full commodification. Second, the commodification of sentient beings presents unique moral concerns. Ted Benton’s *Natural Relations* (1993) provided much of the neo-Marxist groundwork for this area. In Benton’s typology of human–animal relationships, he specifies animals as sources of profit as a distinct category. He justifies this distinction by arguing that the requirements of capital can dramatically transform the form and organization of human relations with animals. Factory farmed animals are designated as the clearest representative of this category. The lives of food animals have been fundamentally altered by the demands of capital. By drawing on a revised early Marx, Benton argues that factory farmed animals are “alienated” beings. The social and technological transformation of animal agriculture has led to a contradiction: intensive animal agriculture is attempting to maximize output through an array of technical means (and animals are treated accordingly) yet, to be “productive,” there are constraints posed by the social, organic, and psychological requirements of farm animals that are undercut in the process of instrumentalizing animals (animals become ill, cannot reproduce, etc.). Benton predicts that attempts to overcome these barriers through antibiotics, growth hormones, and other technical means will lead to unintended consequences (see “Physical and ecological barriers to nature commodification,” below).

In addition to factory farmed animals, Benton (1993) argues that there are animals commodified as a “direct consequence” of capital, namely those animals that are subjected to safety experiments for cosmetics, new foods, and medicine. Of course, the commodification of animals precedes factory farming and advanced animal experimentation. For example, Van Sittert (2005) has retraced a 100-year history (1850–1950) of wild animal commodification in Cape Colony/Province (now part of South Africa, hereafter the Cape). Inverting the usual colonial narrative of game conservation in southern Africa (i.e., urban elite sportsmen saved select wild animals from the market in state reserves), Van Sittert shows, in the Cape, it was actually a rural gentry who used conservation as a step toward the privatization and commodification...
of various game species. Public game reserves were uncommon. In fact, in this colonial period, some of the only wild animals deemed “public property” were “vermin” (mostly jackals and hyraxes) who were slaughtered in large numbers for a bounty with proof of slaughter. Game stocks were enclosed by the rural gentry for incorporation into their market (as “trophies” and meat for consumers and spectacles for zoos). The ostrich and eland were privatized, domesticated, and commodified as agricultural animals during this period. Van Sittert (2005) shows how studying the commodification of nature can challenge taken-for-granted narratives.

Commodification of the human body

A fourth domain of nature in nature commodification literature is the human body. While human bodies as physical things are not empirically detachable from social influence, it is helpful to theoretically separate the somatic from other realms of the human, as well as from other natural entities, because of the body’s unique makeup (Castree 2003). Further, it is clear that the increased “fungibility” and alterability of the human body through technological innovation has increased the opportunities for its commodification. A great deal of work on the causes, consequences, nuances, and domains of human body commodification has occurred in sociocultural anthropology, a line of inquiry spanning back to early studies of slavery and prostitution (Sharp 2000). Drawing from Marcel Mauss and Marx, sociocultural anthropologists explore the “social lives” of bodies and body parts as commodified things. Recent work has emphasized the commodification of reproductive capabilities, organ transplant markets, “elective” surgeries (cosmetic and gender reassignment), genetic material, and the fusion of human, animal, and technological parts (as “cyborgs”). They have asserted that the Cartesian mind–body dualism (in which the mind encompasses the extent of the self) in concert with scientific reductionism has helped legitimize the commodification of the body and its parts.

In a similar vein, Dickens (2000) argues that the human genome project (and its implications in capitalist societies) is another good reason for linking the social and natural sciences. Dickens discusses the commodification of the human body on two levels. The first entails the genetic implications of the basic processes of capitalism. He makes clear that the commodification of the body shapes intergenerational genetic makeup. For instance, social stratification—currently based on the commodification of labor power for amassing private wealth for elite groups—predisposes certain populations to various diseases and developmental outcomes even before birth due to poor prenatal environments. In other words, Dickens argues that the commodification of the body alters the very makeup of human genetics even without biotechnologies. The second level is based on the potential and various routes biotechnologies have for altering human beings in capital’s “own image.” Like sociocultural anthropologists, he stresses the revolutionary role genetic modification could play in body commodification. More specifically, upper classes may soon have the ability to manufacture ideal offspring through commercial means, further magnifying existing inequalities. Both Dickens (2000) and Sharp (2000) demonstrate that the human body shares a similar fate to that faced by the rest of nature under capital.

Commodification of informational nature

The fifth domain of nature explored in nature commodification research is information about nature (“informational” nature). Informational nature, while clearly “materially different” from
COMMODIFICATION OF NATURE

The other natures outlined above, is not only still tangible and physical, but can be, and is, commodified (Castree 2003). Further, privatized informational nature, whether a genetic database or indigenous knowledge about the medicinal and agricultural utility of plants, is often easier to alienate, individuate, and abstract than biophysical entities. The commodification of genetics is perhaps the quintessential example of informational nature commodification. Directly related to the discussion of human body commodification above, Everett’s (2003) argument is that the genetic privacy movement is a direct reaction to the commodification, or potential commodification, of one’s genes for employers, pharmaceutical companies, and insurance companies. However, she clearly shows this countermovement is inadvertently perpetuating genetic commodification by utilizing the same logic as biotechnology companies. Genetic privacy supporters argue that individuals should be able to sell their own genetic information; they support private property laws if applied to one’s own body, and adopt a similar genetic determinist framework (i.e., one’s self is reducible to one’s genes). In short, property discourse becomes the only discourse available for supporters and critics, illuminating the completeness of our “exchange value relation” to nature and the body.

Another emphasis in the work on the commodification of informational nature involves the selling of local, traditional, communal, or indigenous knowledge about natural entities to Western pharmaceutical and agricultural corporations. “Bioprospecting” involves pharmaceutical or agricultural corporations actively searching for local knowledge about, and genetic samples of, natural entities, usually plants, in order to patent and use this local knowledge and genetic material for a profit. Because multinational corporations profit from seizing local knowledge and genetic material, usually from developing countries, without offering financial compensation, the term “biopiracy” directly calls into question the legality and morality of this process, perhaps most famously and forcefully by Shiva (1997). One of the most publicized examples is the patenting of genetic material taken from the seed of the Indian neem tree as a pesticide property by WR Grace & Company. Although the neem tree had been used in India as an insecticide safe for humans for years (as well as for medicinal purposes for thousands of years), neither India nor Indian farmers were compensated, raising a wealth of moral, legal, and political questions related to the commodification of informational nature (Marden 1999).

The socioecological effects of, and barriers to, nature commodification

Along with exploring the way commodification processes function and adapt in different domains of nature (external, internal, animals, the human body, and informational), explaining the socioecological consequences of, and barriers to, nature’s commodification is a research emphasis in nature commodification literature. The explanatory thrust of contemporary work on the commodification of nature can be roughly divided into two broad and interrelated arguments. Both arguments counter apolitical explanations for environmental degradation as well as dominant market-oriented and technical approaches to normative environmental research. First, the process and/or outcomes of nature commodification are ecologically harmful because (i) commodified entities have material impacts on noncommodified environments that are systematically linked to and/or near the former and (ii) nature commodification brings more natural entities into an inherently unsustainable economic system. Second, there are natural
“barriers” to “real” commodification. More specifically, (iii) internal and external obstacles arise during the commodification process that can create unintended socioecological side effects (risks) if “overcome” through technological means that can potentially further complicate the commodification process. Castree (2003) has helpfully labeled (i) “commodification effects” and (iii) “incomplete commodification.” Here, (ii) is included with (i) under “commodification effects.” (iii) is labeled “incomplete commodification effects,” and broken down into physical/ecological, social, and economic barriers to commodification. First, this section summarizes the two subtypes of commodification effects (impacts on noncommodified environments and the expansion of the “treadmill of production”). Then it summarizes the three subtypes of incomplete commodification effects, or the barriers to nature commodification (physical/ecological, social, and economic).

Commodification effects

The first subtype of commodification effects asserts that the production and consumption of commodities have negative ecological impacts on proximate or related noncommodified environments. The basic claim is that spaces of the natural environment that are noncommodified, because they are not immediately economically valuable for individual capitals, are harmed by the production of other commodities. For instance, vehicle production and consumption pollutes the atmosphere and, for an example more relevant in this case, commodifying a single tree species in the Amazon harms surrounding forest and fauna (Castree 2003). O’Connor (1998) has explained this process well. Because degrading the natural environment does not immediately negatively impact a given firm’s balance sheet (in fact, it can increase profits), individual capitals do not have any short-term, instrumentally rational motivation to sustain production conditions (for the unintended economic dilemmas created by this process, see “Economic barriers to nature commodification,” below).

The second subtype of commodification effects is the claim that nature commodification is ecologically harmful because it brings more natural entities into an inherently unsustainable economic system. Capitalism is a mode of production that operates and reproduces itself largely through intensification and geographical expansion. It is growth dependent: it must continuously accumulate capital on an ever-increasing and ever-expanding scale. A number of (neo-)Marxist environmental scholars have argued that the inbuilt expansionary mechanisms of capitalism make it fundamentally unsustainable. The most representative theory that expresses this view is Schnaiberg’s (1980) notion of a “treadmill of production” (ToP). The expansion of production in capitalist economies creates a production cycle (“treadmill”) that inevitably increases “withdrawals” from the environment (natural resource extraction) and “additions” into the environment (pollution) (thus, ToP theory also includes many of the claims of the first commodification effects subtype above). Businesses must reinvest surplus profits in production to create more commodities to be sold for a profit. ToP theorists argue that this treadmill is the system at the core of ecological degradation. In relation to the commodification of nature, the encroachment of the market into natural spaces and increase in material and energy throughput is not only inherently unsustainable, but possibly unstoppable within the constraints of capital. Unlike a great deal of works on nature commodification that rely on the case study method, studies drawing from ToP theory frequently provide national- and international-level data that show the ecological
Incomplete commodification effects: barriers to nature commodification

Here, “incomplete commodification effects” refers to (i) the immediate barriers to nature commodification and the consequences of these barriers as well as (ii) the side effects (risks) created when attempting to overcome these barriers that can further complicate commodification processes. Barriers to nature commodification sit on both sides and in between the society–nature interface. Thus, it is helpful to analytically separate three barriers to commodification: physical and ecological barriers, social barriers, and economic barriers.

1 Physical and ecological barriers to nature commodification. A first barrier to nature commodification results from the nature of nature. Physical and ecological barriers to nature commodification refer to material obstacles natural entities exhibit in the process of commodification as well as the socioecological risks produced by attempting to overcome these barriers through technological means. Bakker’s (2003) examination of the difficulties encountered in the privatization of water in England and Wales is an excellent example of physico-ecological barriers to nature commodification. For Bakker, water is the quintessential “uncooperative commodity.” Arising from neoliberal criticisms of the welfare state’s failure to efficiently manage public services, water was reassigned the status of a tradable good, as opposed to a public one. On the one hand, the socially created and defined (“second-order”) scarcity of water presented an opportunity for capital as a justification for the privatization and commodification of water. On the other hand, Bakker (2003, 28–29) asserts that the “materiality” (biophysical properties) of water erects barriers to commodification: it is (i) a “flow resource” and (ii) heavy. Flow resources are “not easily bound” (Bakker 2003, 32). This makes it difficult to create property rights boundaries (see “individuation” above), and “negative externalities” (undesired environmental harms) can easily flow from user to user. The heaviness of water, although cheap to store, makes its transport expensive “relative to value per unit volume, requiring large-scale capital investments in infrastructure networks which acts as an effective barrier to market entry” (Bakker 2003, 33). In short, the materiality of water presents obstacles to commodification.

Many examples of physical and ecological barriers to nature commodification can be found in modern agricultural practices. In Mann and Dickinson’s (1978) classic article, they argue that there is a “lag” in agricultural production between “labor time” and “production time” due to biophysical processes (e.g., photosynthesis, gestation) that pose natural barriers to capital accumulation. Thus, capital is unable to completely subsume agricultural spheres, a thesis that continues to influence a great deal of critical work on agriculture. Yet capital attempts to overcome these natural barriers through various technological means (see “Commodification of internal nature,” above). In this case, overcoming natural barriers often means an attempt to accelerate natural processes. These attempts can bring about manufactured socioecological risks. For example, as Benton (1993) predicted (see above), there are consequences of submitting animal bodies...
commodification of nature to the imperatives of capital, such as BSE (bovine spongiform encephalopathy) and avian flu (Castree 2003). These risks, what one could call incomplete commodification side effects, present further barriers to the commodification process.

2 Social barriers to nature commodification. A second barrier to nature commodification results from social movements. Social barriers to nature commodification refer to social movements that organize explicitly to oppose the further encroachment of markets into nature or social movements whose interests overlap with this goal. Although all environmentalist groups oppose environmental degradation, not all oppose capitalist expansion and growth. Similarly, although all social justice movements support human flourishing, not all link social injustice to environmental degradation. Despite this, there has been a more clear and active overlap between traditional leftist politics and environmentalism since the 1980s. Red-green political coalitions and permanent parties, environmental justice groups, Third World movements against the privatization of commonly held resources, and farmer opposition to buying seeds are all movements that attempt to strengthen the barrier between nature and the commodity form. Social barriers to nature commodification present unique obstacles as they are the only impediment to commodification with the potential of formulating and implementing alternative economics–nature arrangements.

While physico-ecological and economic barriers interrupt, divert, and restructure the ways nature is commodified, social barriers can potentially improve society–nature relations. For example, following the neoliberal structural adjustments required by the IMF and World Bank, a number of indebted countries in the Global South privatized and commodified water services, especially throughout the 1990s. However, there was a sharp and extended backlash from social justice movements, especially in Latin America. In concert with the difficulty of making large profits in poor countries through tap water commodification, these movements have pushed many multinational water firms out. Yet capital also challenges and overcomes social barriers to nature commodification. For instance, bottled water, a more profitable and portable route to water commodification, has increased steadily despite resistance from social movements at the sites of production and consumption (Jaffee and Newman 2013).

3 Economic barriers to nature commodification. A third barrier to nature commodification is economic. Usually, nature commodification literature focuses on how economic forces are antithetical to ecosystem health. However, some have gone further and claimed that capitalism may undermine its ability to accumulate capital due to the antithesis between capital and nature. This view is typified by O’Connor’s (1998) second contradiction of capitalism thesis. For O’Connor, the environment is a condition of production that does not operate in accordance with the “laws” of the market and value. Because the natural environment is a “fictitious commodity” (see above), it is neglected (degraded) by individual capitals. Over time, this impairs the environment’s productive powers for future capital accumulation. In Marxist terminology, the reproductive demands of the conditions of production are not met, creating economic crises. While Marx (1981) theorized a demand-side crisis
COMMODIFICATION OF NATURE

of overaccumulation/underconsumption, O’Connor’s “second” contradiction of capitalism leads to a cost-side crisis, which can come about in two ways. First, individual capitals maintain or increase profit margins through means that damage or diminish the natural conditions necessary for production. Second, the social movements described above can put pressure on capital to sustain or repair natural environments. Although individual capitals maximize profits by degrading the natural environment in the short term, this negatively affects capital as a whole and can undercut long-term productivity of given industries. Thus, according to O’Connor, only the state can effectively regulate access to and use of natural resources, but O’Connor maintains there is no reason to believe it necessarily will under current conditions. Indeed, in mainstream economic theory and international environmental policy, markets are often put forth as a solution instead.

“Proxy” nature commodification: criticisms of neoclassical environmental economics

Together with exploring the way commodification processes function and adapt in different domains of nature and explaining the socio-ecological consequences of, and barriers to, nature’s commodification, critiquing neoclassical environmental economics is a third research emphasis in nature commodification literature. Unlike the commodification processes described above, “proxy” commodification, as Castree (2003) calls it, refers to the process of placing price values on noncommodified or not-yet-commodified natural entities. This section (i) briefly summarizes the logic behind proxy nature commodification; (ii) explains how proxy nature commodification is used to further real nature commodification; (iii) provides two critical examples of the latter (emissions trading and wetland offsets); and (iv) abridges four general criticisms of neoclassical environmental economics leveled by (neo-)Marxists and institutional ecological economists.

1 Neoclassical environmental economic valuation schemes assume environmental degradation is a result of market failure and, thus, are designed to make up for “missing markets” (i.e., when an “unmet demand” exists without a market). The claim is that when property rights are nonexistent or vague, natural resources are overused. According to neoclassical environmental economic theory, without markets for noncommodified environmental goods and services people fail to recognize that environmental goods and services have “value,” which is synonymous with “positive prices” in neoclassical environmental economic theory (Adaman and Özkaynak 2002). Similarly, pollution occurs because the “social costs” of economic processes are not represented in the “private costs.” This means that environmental harms (conceptualized as “negative externalities”) are not accounted for either in the market price of the product or in the product’s production costs. This occurs when neither producers nor consumers are liable for the costs of environmental harms. Despite the fundamental differences between the neoclassical environmental economic and (neo-)Marxist positions detailed below, both camps agree that noncommodified environments are harmed because they are not considered economically valuable. For (neo-)Marxists, this is only one feature of the environmental
COMMODIFICATION OF NATURE

harms caused by commodification (see the first subtype of “commodification effects” above) and these destructive circumstances further solidify the need for a qualitatively different social formation. However, for neoclassical environmental economists, the solution to this problem is to internalize these negative externalities. Thus, in opposition to (neo-)Marxists, who argue that further nature commodification will increase environmental degradation, neoclassical environmental economists argue that further nature commodification via economic valuation models is the solution to environmental degradation.

2 Foster (2002) explains how neoclassical environmental economic valuation schemes operate in three stages. The first two stages depict what is meant by proxy commodification while the third stage depicts the use of proxy commodification models for the real commodification of nature. The first stage is to analytically split up the non-commodified environments into particular categories of goods and services (see individuation above). Second, price values based on manufactured supply and demand curves are assigned to these goods and services. Demand curves are often created through willingness to pay (WTP) and willingness to accept (WTA) methods. WTP refers to the highest amount a consumer is willing to pay for a good or service (e.g., clean air) or to prevent a negative externality (e.g., algal blooms), a method used when this amount is unknown. WTA refers to the amount a consumer is willing accept for a good or service lost (e.g., compensation for a flooded homeland due to a dam project) when this amount is unknown. Third, new markets are created based on these valuation schemes or, if markets already exist, policies, regulations, and new mechanisms are implemented to alter existing markets. This third step should make clear that proxy commodification is not simply an abstract way of monetarily depicting nature’s value for economic productivity. It is also used as a platform for furthering real commodification. There are several ways to put valuation schemes into practice (Adaman and Özkaynak 2002; Foster 2002), the results of which often appear under the heading “payments for ecosystem services” (PES). The first step is introducing clearly defined property rights of environmental goods and services when none exist or they are not explicit (see privatization above). This will ensure that, in theory, compensations for use and/or pollution of natural resources are appropriately allocated. Second, regulations are imposed. The simple option is to impose taxes for environmental impacts and subsidies for environmental improvements. More abstract and complex options, which are preferred by neoclassical environmental economists (Foster 2002), involve the creation of new markets with incentive systems via state intervention and policy. Although there are a number of proxy commodification schemes put into practice (e.g., tradable harvest rights in the fisheries industry, wildlife enhancement schemes, biodiversity credit schemes, and debt for nature swaps), the Kyoto Protocol’s emissions trading scheme and the creation of wetland mitigation banks are used as examples here.

3 The idea of emissions trading, born in the United States (who did not ratify the treaty), is to make carbon sources economically scarce through imposed emission limits and, subsequently, assigning tradable legal rights to emit. The prices created through bargaining would, in theory, represent how
much society “valued” the resource. The United Nations distributed rights to emit to the Northern countries that ratified the treaty. These rights were then usually distributed to leading industrial polluters. These polluters could sell unused permits to heavier polluters. Alternatively, those who wanted to avoid caps were given two more options: build or invest in new carbon dumps (offsets) or, invest in business ventures and infrastructure abroad, usually in developing countries, if the results can be shown to emit less than if they had not invested (Lohmann 2005). Usually framed as a “first step” to climate change mitigation, the Protocol has not only failed to reduce total emissions but has other social and ecological implications. Lohmann (2005) has argued that the emissions trading scheme has successfully redirected intellectual and financial resources from innovations and social changes that have the potential to actually reduce emissions. Additionally, alternative social futures that lie outside “national allowances” and the like are effectively marginalized. Indeed, Lohmann details how environmental criticisms of the Protocol were scorned as a “do-nothing” stance. Lastly, the treaty is thoroughly technocratic and corporatist, which excludes input and solutions from nonprofessional and noncorporate groups.

Wetland offsets via mitigation banking is another (in)famous example of applied proxy nature commodification (Robertson’s work is especially notable here, e.g., 2004). Due to failures to enforce the compensatory mitigation provision in the Clean Water Act (i.e., if a firm fills a wetland it must restore or set aside a set acreage of wetland in compensation), President George H.W. Bush’s administration devised a “No Net Loss to Wetlands” policy goal whereby developers who fill a wetland for development can buy “wetland mitigation credits” from owners of a wetland who agree not to develop in the future or companies who restore degraded wetlands (“wetland mitigation banks”) (Robertson 2004; Smith 2007). The intention is for wetland acreage in the United States, at a minimum, to remain constant. For example, a developer who fills a wetland to build apartments can buy wetland credits from an off-site mitigation bank instead of conserving or restoring wetlands on the same parcel of land his/her company is developing. However, the attempt to make commensurate the value lost at the development site and the value gained at the mitigation banking site creates problems. The value gained at the banking site is often conceptualized as “bundles” of ecological functions and measured through algorithms known as “rapid assessment methods,” while at the development site acreage is often used as a proxy measure of these bundles of ecological functions. A contradiction emerges when mitigation banks begin to “bank” on the different minute and opportunistically conceptualized ecological functions between the development site and the mitigation banking site in order to take advantage of their mitigation bank’s unique ecological characteristics. This search for additional fabricated markets clashes with the mitigation permit market’s need for quantitative abstraction (Robertson 2004).

In addition to the specific criticisms of emissions trading and wetland banking, (neo-)Marxist and institutional ecological economic literature levels a number of criticisms at economic valuation schemes and their applications (e.g., O’Neill 2001; Foster 2002; Smith 2007; Kosoy and Corbera...
The first two of the following four lines of criticism are in response to the valuation schemes of neoclassical environmental economics. The second two are in response to their application.

Neoclassical economic valuation models reduce human beings to the utility maximizing, instrumentally rational, fully informed, and egoistic *homo economicus* of liberalism. This model of humanity is distilled in neoclassical economic theory and is presupposed in their environmental economic valuation schemes. Economic man, as a one-sided and empirically imprecise depiction of humanity, implicitly omits other forms of relationships real humans have with nature that stand outside the limits of *homo economicus*.

Nature has no other value than exchange value in neoclassical environmental economic models. Or, more specifically, any value nature has can be quantified and expressed in monetary terms. This assumption carries two problems. First, qualitatively distinct natural entities may not be comparable in quantitative ways, or, they are incommensurable to begin with. Second, this assumption excludes the possibility that natural entities have intrinsic worth irrespective of human thought and action. Even if one assumes only humans can assign value to objects, neoclassical environmental economic models are incapable of accurately taking account of noneconomic values. However, this does not stop WTP technicians from asking “consumers” how much a view of a landscape – a “cultural service” – is “worth” to them in monetary terms, for example. That is, neoclassical environmental economic models assume all values can be reduced to, and successfully translated into, dollars and cents. (Neo-)Marxist nature commodification literature firmly disavows this assumption, as do other schools of thought, such as institutional ecological economics (Adaman and Özkaynak 2002). This is not only seen in left-leaning scholarly critiques of greened neoclassicism, but also in the refusals of some study participants to price environmental goods and services in accordance with WTP and WTA surveys, interpreting the questions as corrupt and an answer as a betrayal of personal values (O’Neill 2001).

Another line of criticism asserts that some of these valuation schemes, when put into practice, may reproduce or intensify global inequality and poverty. For instance, a Third World farmer who is compensated for preserving a forest will not likely experience any long-term increase in life quality, an opportunity that could have come about if that forest had been developed. The belief that there is voluntary cooperation among those who provide “carbon sequestration services” in the Third World may hide the fact that participation exists due to structural poverty. Similarly, it is not clear how WTP and WTA methods should practically deal with wealth stratification among its “consumers.” The question is whether demand curve calculations should be hierarchized by magnifying the preferences of those with more disposable wealth and minimizing or excluding those with little or no disposable wealth. On the other end of the social pyramid, mitigation permit banks are making a good deal of money and there is already a speculative “environmental derivatives” market. On top of this, there is concern about how property rights are implemented, contested, and negotiated. Those in power assign property rights and, often, negotiations are structured by extant power inequalities.

When put into practice, economic valuation schemes only serve to further the socioecological consequences associated with nature commodification described above. Many acknowledge there may be short-term environmental gains, but, in the long term, capitalism is the problem, not the
solution, or “the [proposed market] solution is the problem,” as the subtitle of O’Neill’s (2001) essay puts it. This is the fundamental critique of neoclassical environmental economics and PES schemes put forth in nature commodification literature. In particular, PES schemes are criticized for severing natural entities from their ecological contexts as “functions” (see individuation and abstraction above). This undermines the wealth of data that depict the interdependent, relational, and interactive nature of nature.

In summary, Marxian-rooted nature commodification literature and institutional ecological economists level a series of criticisms that undercut the conceptual and empirical core of the market-based environmentalism common today: (i) “rational economic man,” the agent presupposed in nearly all neoclassical economic models, is not an accurate characterization of real human beings and the diverse relations they have with nature; (ii) the totality of nature’s value cannot be expressed in monetary terms; (iii) applications of valuation schemes may reproduce social inequality; and (iv) increasing the scope of capital’s penetration of nature will only serve to perpetuate the structural causes of environmental problems.

Future directions

Before discussing future directions work on nature commodification can take, it is important to clarify what is already present in current studies that is valuable and should be sustained. First, studies of nature’s commodification should continue to stress the historical and contingent character of nature commodification. As stated in the introduction, past social formations have traded economic products derived from the biophysical world, but capitalism is unique in that the primary purpose of “producing nature,” as Smith puts it, is to amass private wealth, not to meet human and ecological needs. Historicizing and denaturalizing these relations is a defining feature of nature commodification literature that ought to be preserved. Second, studies of nature’s commodification should continue employing critical perspectives. While there is a diverse range of lenses available for exploring the commodification of nature, from classical Marxism to feminist political ecology, these perspectives are unified by a normative stance that the further penetration of nature by capital is problematic and the relationship between economic and natural forces ought to be restructured. Castree (2003) has argued that researchers should make their normative critiques explicit. Third, studies should continue to study the commodification of nature in its diverse domains, geographical areas, and historical periods. One lesson gained from studies of nature’s commodification is that the natural entity in question alters the way commodification proceeds and operates. Thus, case studies from a wide range of natures, areas, and times are necessary for advancing the area of inquiry. Fourth, studies should continue elaborating the effects nature commodification has on the natural entity in question, surrounding ecosystems, social relations, and the commodification process itself. Relatedly, to further the generalizability of one’s study, it may be helpful to conceptually align the results to critical and general explanations for environmental degradation, such as Schnaiberg’s treadmill of production theory or O’Connor’s second contradiction of capitalism thesis. Fifth, studies of nature’s commodification should continue to illuminate barriers that arise in the commodification process as well as resulting manufactured risks when industries attempt to overcome these barriers, if applicable. Lastly, scholars should continue critically evaluating neoclassical environmental
economic valuation schemes and their applications. Neoclassical valuation schemes have a great deal of influence in regional, national, and international environmental policy today, limiting the scope of alternative solutions.

Despite the strengths of nature commodification literature, there are openings for new directions and improvements. Most glaringly, there should be a scholarly effort to conceptually and theoretically link future case studies in order to achieve some level of generalizability as well as to connect the field. There are two openings that could help meet this goal. The first is to draw from a similar stock of conceptual categories. It seems difficult to envision something that looks like a general theory of nature commodification due to the diverse ways the commodification process adapts to the given natural entities and vice versa. Thus, as stated above, case studies will continue to be an important feature of nature commodification literature in order to understand and explain these nuances. Because abstraction and generalization are made more difficult, nature commodification scholars should be more attentive about employing similar concepts in order to approximate generalized explanations and predictions. Castree’s (2003) integrative typologies, which were retained here with some modifications, seem especially applicable. This is not a call to limit the perspectives utilized to theorize nature commodification, but a proposed means to help unify the area of study. A second way to help unify the area of inquiry is to increase communication and collaboration between disciplines. Because the topic/problem sits in-between fields, the diversity in disciplinary backgrounds studying the commodification of nature appears necessary. Further, one of the reasons nature commodification literature is innovative and exciting seems to emerge from this amalgamation of traditions. However, this diversity also carries difficulties, namely an abundance of sometimes unfamiliar perspectives and their accompanying jargon. Increasing explicit collaboration and communication across fields (interdisciplinarity) could benefit future studies.

**SEE ALSO:** Bioprospecting and biopiracy; Biotechnology; Environment and resources, political economy of; Environmental degradation; Environmental valuation; Nature; Neoliberalism and the environment; Political ecology; Production of nature

**References**


COMMODIFICATION OF NATURE


Community

Briavel Holcomb
Rutgers University, USA

The word community, derived from the Latin communitas, has variable meanings but all imply relationships, whether among people, plants, or animals. Until relatively recently, and still to a large extent today, community had spatial implications. A common usage refers to the relationships among people in a particular place and indicates a sense of belonging both to place and to the social group(s) within it. Attention to the notion of community followed from Ferdinand Tonnies’s 1887 groundbreaking distinction between gemeinschaft (communal society) and gesellschaft (associational society). The former references traditional, personal, face-to-face relationships typical of rural life at that time. The latter identifies less personal relationships following from individuals interacting with businesses and bureaucracies. Gemeinschaft was romanticized as providing deeper senses of belonging and mutual aid, while industrialization weakened traditional bonds of family, religion, and kinship. The writings of Karl Marx center on conception of community. For Marx, community is a democratic form of association which replaces the state. Humans are communal animals and can only achieve their highest form of existence in a community. The individual develops in association with others but Marx saw traditional village community as stifling, enclosed, and not conducive to the true freedom and individualism to which all should aspire. While the cities of the Industrial Revolution provided the antithesis of community, they could eventually, in Marx’s view, provide the possibility of evolving to an environment in which the individual contributes according to his ability and receives according to his need. That perfect community is democratic and free. It could be, but is not necessarily, place-based. For Marx, democracy is of the nature of community. The ideal communist society is one in which the freedom of the individual is obtained through their free association with others in which all productive activity is a matter of individual or collective choice.

The famous “Chicago School” in sociology also offered a distinctive notion of community. In the first half of the twentieth century, this School focused on ecological ideas of urban form and stressed the importance of place and location in ideas of community formation, while reinforcing the notion that the modernization of urban life was reducing traditional values of community belonging. Louis Wirth’s classic article on urbanism, published in 1938, argued that people living in cities depended on more daily interactions than their rural predecessors, but that those interactions were impersonal, superficial, and transitory. In a heterogeneous society, people belonged to more groups, but their needs and preferences were subsumed into those of the majority. People join groups in order to pool resources which are directed to serve the average constituent. Communities, then, are rational or functional rather than emotional. Another member of the Chicago School, Robert Park, applied ideas of ecology developed for plant and animal communities to observations of urban life. The notions of competition, invasion, dominance, and segregation which were current in ecology were used to study human behavior and Park...
COMMUNITY

emphasized the importance of communication in the formation of community. Park saw the city as a series of “cells” in the urban organism with spatial contiguity of neighborhoods, immigrant groups, and functional districts providing bases for community. Nonetheless, although he stressed place and locality, Park also was open to the possibilities of nonspatially determined forms of human association. He noted that various urban institutions from schools, to churches, to street gangs, aided by communications including newspapers, could create communities of interest not reliant on spatial contiguity. The place-based idea of community so redolent in the Chicago School has remained a strong theme in both geographical theory and praxis, particularly in urban areas. However, the Chicago sociologists assumed the urban societies were characterized by impersonality, social disorganization, alienation, and anomie. Urban places were essentially anti-community in the conventional sense. Marx’s optimistic predictions made in nineteenth-century Manchester about the (r)evolution of industrial cities were not evident in twentieth-century Chicago.

While ideas of community historically have not been exclusively place-based or spatial, the second half of the twentieth century saw growing recognition of the existence of aspatial or at least not place-based communities. An early contribution was Melvin Webber’s 1963 “Order in Diversity: Community without Propinquity” in which he argued that the sociologists’ lament that community was dead/dying was short-sighted, and that people were continuing to form communities based on common interests, professional identities, and other nonspatial associations. This focus on time–space distantiation was a relatively fresh idea in 1963, but was a precursor to subsequent thought which has, of course, “exploded” with the Internet. Webber likewise noted that traditional distinctions between rural and urban were diminishing as people of all occupations could participate in national urban life.

Around the same period, Herbert Gans and others argued that the Chicago School was too pessimistic in its view that industrial cities destroyed the real sense of community which had characterized rural village life. Gans’s 1962 classic study of the West End of Boston found that the cohesive social networks and sense of identity associated with rural life were replicated in “urban villages” often based on either ethnicity (Italians in Boston’s case) or class (working class in London or Chicago). Some reinforced this view, pointing to the strong ties formed by shared experience, kinship, friendships, and attachment to place. But others pointed out that working-class neighborhoods were frequently characterized by economic insecurity, social stress, conflict, and crime. High density, sometimes combined with ethnic diversity, was not conducive to mutual support and community enhancement. Some considered Gans a romanticist. But few found a sense of community in suburbia. Mid-twentieth-century critics, such as Lynds, Mumford, and Whyte (*The Organization Man*), saw suburbia as a place of nuclear families enjoying privacy and individuality rather than strong community ties. Nevertheless, Gans’s (1967) study of suburban Levittown found evidence of communities in which residents chose to participate in relatively loose association for purposes ranging from parent–teacher associations to country clubs, gardening, and the like. It seemed that urban versus suburban communities were similar in purpose but different in execution.

Whether or not community was dead, it was resurrected by activists and government in the 1960s and 1970s. Faced with “the urban crisis,” civil disturbances, and cries for a war
on poverty, various federal programs based on
the presumption of community were initiated,
and in many cases have survived well into
the twenty-first century. The Community
Development Block Grant (CDBG) program
was enacted in 1974 under President Gerald
Ford with bipartisan support. Cities and states
received federal grants and were required to
submit allocation reports showing where and
how the money was spent. The amount of
the grant was determined by such measures
as poverty rate, overcrowding, and amount of
deteriorated housing, and a rural “set aside” sent
a third of monies to rural areas. The idea behind
CDBG was that it would be a “bottom-up”
approach with local people setting priorities
and being involved in deciding where and how
monies would be spent. It was also a reaction to
the Johnson administration’s “war on poverty”
which had spawned a large federal bureaucracy
inappropriate for Republican administrations.
However, perhaps ironically, proponents of
CDBG included such organizations as ACORN
(Association of Community Organizations for
Reform Now), begun in 1970, which grew to
the largest community organization for low and
moderate income people before disbanding in
2010. ACORN and other activist community
organizations of the period were inspired by
the ideas of community organizer Saul Alinsky
(who had studied with Robert Park at Chicago).
Alinsky’s book Rules for Radicals (1971) was based
on his experience organizing with the Industrial
Areas Foundation. He recommended working
with the “have nots,” the poorest segment of
society, and stressed the need to communicate
in organizing. Alinsky was both idealistic and
pragmatic. Among his rules were: Make the enemy
live up to its own book of rules. Ridicule is man’s
most potent weapon. Never go outside the expertise of
your people. A good tactic is one your people enjoy.
ACORN survived until 2010 but other less
radical alternatives, such as the Industrial Area
Foundation, continue today to carry on similar
community organizing.
CDBG has, of course, had its critics, who
argued that its priorities and execution were
flawed. Critics noted that some projects on
which funds were spent were frivolous and did
not target the most needy, and the amount
of money allocated had declined (from US$2.7
billion in 1975 to $1.7 billion in 2014). Although
the program was conceived with an antipoverty
agenda, much of the money now is widely
distributed to middle and even upper income
places, while that going to lower income com-
munities may be spent on infrastructure which
makes places more attractive to middle income
residents. CDBG monies have paid for tennis
courts, recreation centers, and senior citizen
centers in well-off communities, sugarcane mills
in Hawaii, and, reportedly, psychic readings for
a mayor in Texas. Yet CDBG remains a viable
program as local people receive federal funds
for locally determined needs. In the United
States, the Community Development Society,
an organization formed several decades ago with
a mostly US-based membership, is devoted to
helping professionals engage with community
members to improve the sustainability and well-
being of communities. Among their principles
is that members should disengage from support
of any effort that is likely to adversely affect the
disadvantaged members of a community. The
society holds annual conferences for practitioners
of community development and publishes Com-
munity Development quarterly. The articles include
both theoretical and case-based contributions,
focused primarily on North America.
A considerably more international perspective
is that of the Community Development Jour-
nal, a quarterly that claims to be the leading
international journal in the field and seeks
to publish articles that challenge conventional
wisdom, suggest innovations, and relate to issues of social justice, diversity, and environmental sustainability. Published by Oxford University Press, a recent (July 2013) issue included articles about communities in Columbia’s mining region, fishing versus oil production in Ghana, and the role of anthropology in New Guinea’s extractive industries. While providing provocative reading, the journal complicates the task of defining community development and challenges the conventional wisdom that community is local by showing the many ways communities are linked to outside, sometimes distant, forces. The adage “think globally, act locally” (a saying attributed to various people) might be thought of as a theme of the journal.

Cities throughout the world have recently seen the growth of “specialized” communities often catering to particular interest groups such as the elderly or the affluent. Ironically, perhaps, the first retirement community established was Youngstown, Florida, and was followed by hundreds of others in warmer states like Arizona, California, and other parts of the world. Such communities may provide a continuum of housing from independent living to assisted living or even nursing homes. They typically cater to middle income residents and have services which build community among them, such as shared club houses and sports facilities. Not surprisingly, they have both advocates and critics of segregating older people into physically discrete locations, most of which ban younger residents and children, except as visitors. But proponents praise the social capital built by shared spaces and activities.

Gated communities, now proliferating in many countries, range in size from small developments to larger villages. They typically have limited and guarded entrances and may be surrounded by fences. It is assumed (possibly wrongly) that this reduces crime, but it also increases the number of rules and regulations residents must obey, many to do with the appearance of property. Gated communities are popular, with over 11 million residents in the United States. They offer shared amenities and, presumably, an enhanced sense of community. They are, however, criticized as exclusionary and catering to the affluent. In some Latin American countries, Saudi Arabia, India, and South Africa, gated communities are desired for their assumed increased security as well as their prestige. Critics of gated communities are many, but so are their proponents. In possible contrast, “intentional communities” are part of a movement which has its roots in utopianism and blossomed in the 1960s in the United States, but continues to be a lifestyle of choice for a small but significant segment of the population. Intentional communities are those in which residents live together with a common purpose, often working cooperatively to create a lifestyle that is typically eco-friendly, and based on shared values. They include rural villages, residential land trusts, and urban housing co-ops, as well as communities for particular groups such as Christians or gays. The movement publishes a magazine, Communities, which gives advice on establishing and maintaining intentional communities.

In a perceptive book Lyn Lofland (1973) discussed how in preindustrial societies most people knew most of the people with whom they came in contact on a daily basis. Strangers were few and far between; however, in the industrial city, we know very few people we see in the course of a day with whom we are acquainted. Most people we see are strangers. Thus, while we may feel we live in a “community,” we have evolved ways of interacting with unknown people by assuming commonalities with them. Subsequent critics have argued that in fact current city dwellers no longer assume or rely on community, but instead go “bowling alone,” as Putnam (2001) warned. Putnam
described the decline of participation in civic and other social groups during the period after about 1970 in the United States. He notes that voter turnout and membership in voluntary associations, including churches, all declined. Americans and other “Westerners” started to spend more time on individual leisure pursuits (movies, watching—not playing—sports) and less on eating meals together, playing on teams, and other communal activities. Americans gave less to charities proportionately, and while the rate of volunteering grew from 1970 to 1990, most of the increase was due to people over sixty years old. Trust in strangers decreased and hitchhiking died. While community thrived among some specialized groups (e.g., evangelical Christians, self-help groups), overall people felt much less connection to community, and what Putnam (and others) call social capital declined. Putnam attributes this decline in community to such things as suburban sprawl and the replacement of local shops with regional malls, the increase in commuting to work alone by car, and the increase in TV viewing. But, according to Putnam, if we can rebuild social capital we will be more prosperous, healthier, and happier. We might do so by involving young people in meaningful civic projects, making workplaces more family-friendly, reducing commuting times, using mixed-use zoning and pedestrian-friendly streets. While Putnam’s arguments were not wholly new, his book was influential in calling attention to what may be a decline in feelings of place-based community commitment in some parts of the world in this century. Other advanced industrialized and suburbanizing countries (e.g., United Kingdom, Japan, Australia) experienced similar trends. In China the rapid growth of cities caused by rural in-migration has disrupted traditional rural communes, but new forms of urban communities have evolved. Putnam’s work generated considerable response, many people pointing out that while local neighborhood participation may have declined, membership in larger groups has grown. DeFilippis (2001) is among Putnam’s critics, pointing out that social capital without financial capital has limited utility in community development.

Technology, at first the telephone, now the Internet, has greatly facilitated the emergence and maintenance of a great diversity of communities, some place-based but many essentially aspatial. The idea of geographically dispersed community was central to Anderson’s (1983) *Imagined Communities* in which he describes ways in which people of a nation feel a sense of belonging to a group which never meets in person, but which is assumed to have certain interests in common. While such communities have territorial limits, they are imaginary in the sense that the members of a nation may have little in common and few mutual interests. Ironically, perhaps, while Putnam and many others attribute a major reason for the decline of community to media (such as TV), Anderson attributed the imagined communities of nations of early modern Europe to the ability of entrepreneurs to print books in the vernacular rather than Latin. Early European states were thus able to find and distribute ideas of nation in their own language, leading to a decline in the belief of divine rights and the beginning of modernity.

Ideas about community have been revolutionized by the Internet. Today, it is probable (at least in the United States) that citizens feel more connected to online, virtual communities than to their place-based community. An early discussion of some implications of this was Rheingold’s (1994) *The Virtual Community: Homesteading on the Electronic Frontier*. Rheingold’s involvement followed from an online discussion group (the WELL—Whole Earth ’Lectronic Link) in 1985. The group had grown to over 8000 members by 1993 and while some
met in the “real world” (San Francisco, naturally!), most communicated online. By 1994 Rheingold noted that the Internet enabled many communities, and allowed people to find a date, sell a lawnmower, publish a book, and conduct a meeting. He was largely optimistic about the community-building potential of the Internet, noting that it encourages the sharing of interaction, interests, and passions. He argued such interaction could challenge the monopoly of mass media and invigorate democracy. But he also presciently warned of the danger of corporate monopolies in cyberspace and the role of government in monitoring web content. He understood the challenges to privacy that the then new technology provoked, and which remain today.

Since Rheingold’s initial book on virtual communities, much scholarly work has been devoted to understanding these new iterations of community, with both praise and criticism of their effects. While critics point out the loss of in-person socializing (the many hours spent in front of person-avoiding screens and the stultifying tedium of much content), others note the ability to connect with others with similar interests (or problems) who are geographically distant, the ability of invalids and other “shut ins” to make meaningful contact, and even the emotional benefits to being part of a network of people coping with similar adversities such as disease or disability. Similarly, while in-person attendance at local community meetings may have declined, civic engagement among especially the young has been encouraged by online media. Online chat rooms provide possibilities for interaction with others with similar interests and concerns. Emotional support can be accessed quickly and cheaply, though not always reliably. Renninger and Shumer (2002) praise the opportunities for building diverse communities – from young girls to math teachers and other learning groups – which online communication facilitates.

While there are thousands of online communities, virtual worlds are probably the most interactive. Second Life, a virtual world run by Linden Labs, has millions of members, although most are inactive and regular “residents” number about a million. Nevertheless, Second Life is a “place” which emphasizes community and enforces community standards (such as forbidding intolerance, harassment, and “assault” – you may not shoot or shove virtually when the resident is in a safe area). Although teenagers may join Second Life, much of its content is inappropriate for children and a rating system is used to designate “adult” areas of the virtual world. As its website claims, in Second Life there is always someone to talk to, dance with, and learn from. You can meet people from all over the world without leaving your home. A children’s equivalent is the game Minecraft in which players manipulate digital landscapes to build desired terrains either individually or in teams. Minecraft has been downloaded 20 million times and is quite addictive.

Changes in meanings of the word “community” from place/neighborhood to interest group, often facilitated by online communication, are evidenced by data from Google’s Ngram viewer. The single word “community” appeared in books infrequently until the twentieth century, but increased rapidly after 1950, with double peaks in the 1970s and 1990s. (In Chinese the word “community” peaked in 1920 and 1970. In French and German there were marked increases after 1940.) However, the term “Urban Community” in English had a major peak in the early 1970s (correlating with the passage of US federal legislation related to community), while the term “gay community” rose rapidly in the 1980s to the end of the century (and probably continues to the present, though the Ngram
viewer does not cover the period after 2000). Google Scholar shows over a million entries under “gay community” while there are 1.5 million for “women’s community,” 3.2 million for “black community” (and 2 million for “African American community”) and 3 million for “white community.” Obviously, none of these “communities” are constituted of people who know each other, though they may (or may not) have interests in common. Clearly, the meaning of community has migrated from primarily place-based association, or geographical community, toward that of mutual interest or characteristic, while still often maintaining its original territorial connotation. As Collins (2010) argued, this shift from territory to interest group has important political implications as communities of various kinds now constitute sites of political engagement and contestation.

The spatial or geographical meaning of community continues to be prevalent in the fields of community mental health and community policing. The former are services which substituted for care provided in mental hospitals, many of which were closed in the mid-twentieth century in a movement to deinstitutionalize mental patients. Community mental health centers were established to provide outpatient psychiatric treatment to lessen the social exclusion of the mentally ill. Though it has its critics and is habitually underfunded, it remains the model for mental health provision in both the United States and United Kingdom, and assumes that people are better off staying “in the community” where social relations can be maintained. Similarly, the notion of community policing, which emerged in the late 1970s, encourages police to develop good relationships with citizens and key stakeholders in neighborhoods, to establish neighborhood substations, conduct foot patrols, and attend community meetings.

The largest online “community” is probably Facebook, with over a billion members. While Facebook publishes community standards (relevant to terrorist threats, nudity, hate speech, and the like), it also has “communities” within it devoted to cooking or celebrities or whatever generates interest. Facebook nurtures off-line communities as it provides a way for friends and neighbors to keep up with each others’ lives. It has been shown to enhance social capital (among college students) and to build political support for candidates. For geographers, the location-based social networking site Foursquare is of interest since it uses global positioning technology to locate members, enhance commercial enterprises, and enable the meeting of friends in real time/space. The rapid growth of online communities and the geographical implications of location-based networking is a fertile field for further research (Erickson 2010).

If the meaning of community is stretched to include online communication between people who are gathering in a virtual space for political change, one can argue that the various Facebook/Twitter “revolutions” are bringing people together both online and in the material world to create sometimes radical social change. The first such occasion was in Moldova in 2009 when the Communist Party was overthrown for a coalition government. Although public demonstrations had been planned prior to the election, the role of social media (especially Twitter) in effecting the mass mobilization which happened in the streets of Chisinau was partly responsible for the political movement. The following year (2010), Iranian “dissidents” hoping to foster change used Facebook to rally support for the candidacy of an opposition leader, Mir-Hossein Mousavi. The Facebook page was administered from outside Iran, due to government censorship, but nevertheless was able to organize defiant nightly “Allahu Akbar” chants from rooftops in Tehran.
as well as various street protests. Ultimately, this has not led to the change desired by government opponents. It could be said that the Jasmine Revolution (in Tunisia) was enabled by the community-building potential of online social networking.

In conclusion, as the notion of community has expanded from place-based localities to international connectivity, the geography of community still matters, and Smith (1999) urges us to consider the implications of community for ethics and morality. Geographers are among those who contribute to the development and strengthening of real-world communities. In New York State, the Syracuse Community Geography group provides geographical assistance to help build a more just society, and geographers in many parts of the world are offering their expertise and energy to encourage ecologically and socially beneficial communities.

SEE ALSO: Behavioral geography; Chicago School; Emotional geographies; Friendship, geographies of; Gated communities; Geovisualization of social media; Home; Neighborhood; New Urbanism; Social geography

References


Further reading

Community-based natural resource management

Chasca Twyman
University of Sheffield, UK

Community-based natural resource management (CBNRM) encompasses both a broad approach to the management of natural resources as well as a defined project format or program that is implemented through specific policies and practices. At the heart of CBNRM is community control over decision-making about the local environment. CBNRM emerged in the 1980s as both approach and practical project closely aligned with developments in participatory approaches and farming systems research of that time. Its reach was significant with its principles underpinning a range of projects around the globe, such as community forestry projects in Nepal and Bolivia, wildlife conservation in Namibia and Zimbabwe, and marine conservation in Madagascar and the Caribbean. It was seen as the panacea of community conservation and heralded as a transformative form of local development and environmental management. However, CBNRM has come under fire in recent decades for departing from its central ethos and even contradicting some of its core principles of social justice and socioecological sustainability.

This entry explores CBNRM with a particular focus on the Global South. The history of CBNRM provides an important account of how a revolutionary approach evolved and changed over time. The typology of CBNRM explains its current diversity based around resource type, operational mode and level of devolved control. CBNRM was appropriated by different parties in different regions, and arrives today at a fundamental crisis of purpose. The entry examines these contemporary crises focusing on the politics of control, the role of conservation science, the marketization of the environment, and the everyday challenges of living in poverty before examining the opportunities for CBNRM in the future.

First, however, it is important to review what we mean by some of the key terms in CBNRM. It is easy to use the term community without adequate precision and definition. It needs to be recognized that not all communities are alike and membership of communities is often negotiated, contested, highly heterogeneous, and exceedingly political. For the purposes of CBNRM we tend to assume a neat village easily identifiable on a map, but in reality communities involved in CBNRM can be extremely diverse, disparate, and even ephemeral. They may encompass subcommunities of self-selected participants within urban areas, they may include (or exclude) mobile pastoralists or hunters, and communities may struggle over definitions of who counts as a resident when the stakes involved are high. For an approach or project to be community-based it assumes a commonality in purpose among a group and a willingness and ability to work together toward a common goal. Ostrom (1990) provided influential work on how common resources (i.e., those owned or held in trust communally, not individually) could be managed by a set of rules that defined key parameters within a community that would lead to successful outcomes. Inherent within these notions of community and...
community-based are issues of power and representation that are explored later in this entry. Lastly, natural resources within CBNRM refer to specific resources (water, marine resources, land, wildlife, forests, and so on) but also to the environmental services they provide (mitigation of soil erosion, enhancement of range quality or landscape aesthetics, sequestration of carbon, and so on). Projects are often defined around specific resources, though more general everyday forms of CBNRM often encompass a more holistic recognition of the environment.

The emergence of CBNRM

History suggests that in the past, rules and regulations regarding access to natural resources were organized and precise, though rarely written down, and were enforced through traditional institutional structures and moral codes of conduct within societies. Thus everyday forms of CBNRM, rooted in these histories and traditions, have been structuring the day-to-day management of natural resource in most societies around the globe for centuries. CBNRM was a phrase coined to capture the formalization of an approach to natural resource management particularly through policy and prescribed practices. CBNRM initiatives were greatly influenced by theories around common pool resource management heralded by Elinor Ostrom. Ostrom (1990) produced seminal scholarship around the role of public choice in decision-making about the production of public goods and services and on how human interactions with ecosystems could maintain long-term sustainable resource yields. Her guiding principles for common pool resource management are still widely used today.

CBNRM initiatives came into vogue in the 1980s, following a wider move within development discourse away from prescriptive top-down policies and practices toward grassroots community development and local management of natural resources. These approaches aimed to involve and integrate communities in management processes within a structured policy framework (i.e., not wholly top-down or bottom-up), with the ultimate aim of achieving social justice alongside environmental sustainability. Much of the initial influential research and practice in this area was in Africa (particularly for wildlife and rangelands) and Nepal (noted for its social forestry).

One well-known and influential project was the Community Areas Management Programme for Indigenous Resources, better known as CAMPFIRE, in Zimbabwe (Taylor 2009). This pioneering program emerged in the late 1980s and aimed to return land and resource rights to poor communal farmers to promote environmental and economic welfare. The CAMPFIRE association was formed to represent the interests of district councils who had been granted authority to take the leading role in the coordination of CAMPFIRE activities and the representation of CAMPFIRE within national, regional, and international fora. Its success spread and in 1992 Zimbabwean ecologists helped to establish a similar scheme in the Luangwa Valley of Zambia called Administrative Management Design (ADMAKE). In Zimbabwe, the CAMPFIRE program was implemented widely. By 2003, the population of elephants on communal lands, under CAMPFIRE management, had doubled, even though human population had also doubled. Other wildlife numbers were reported to have increased by 50%. The greatly increased revenues meant that land that would have been settled or cultivated was protected at the local level for wildlife. The success has not been without difficulty. The politics of decentralization have been complex with accusations of elite capture, missed beneficiaries, and
COMMUNITY-BASED NATURAL RESOURCE MANAGEMENT

Corruption. However, it is still heralded as one of the pioneering CBNRM programs to date.

Another notable example of CBNRM is community or social forestry in Nepal. Community forestry is a branch of forestry that deals with the communal management of forests for generating income from timber and nontimber forest products while also regulating and ensuring sustainable ecosystem services. The evolution of community forestry in Nepal dates back to the late 1970s and was first devised as an attempt to improve the management of forest resources and address environmental issues that were of great concern given the country’s failing centralized forest policy. In recognition of these problems, in 1976, the government introduced innovative and farsighted legislation whereby the Forest Department could hand over forest land to the panchayat (the local administrative unit, usually comprising several villages). A completely new level of land tenure was thus introduced, aimed at encouraging public participation in forestry activities by giving ownership of the resource, as well as responsibility for its management, to local communities. Further decentralization was promoted in 1982 to provide more local control, and the 1993 New Forest Act allowed handover of the national forest to the adjoining forest users for accountable management. The legislation of CBNRM in this case provides a powerful structure through which control over natural resources can be handed over to village councils for fully devolved management. However, though the policies are far-reaching on paper, in reality elite capture of benefits and the inability of the poorest (often the most dependent on forest resources) to have a voice in the decision-making processes means that the principles of social justice have not always been upheld in CBNRM in Nepal.

Although this entry focuses on CBNRM in the Global South, the movement has not been confined solely to this region. CBNRM projects can be found all over the world, from North America and Europe to Australia and Russia. Not all go under the name of CBNRM and various terms such as co-management and participatory management may be used alongside other terms referred to in this entry. Given the great variation in CBNRM projects it is useful to think about how they might be categorized to better understand how these differences might influence the success, or not, of different projects.

A typology of CBNRM

CBNRM is diverse and manifests itself differently across the globe, so it is useful to think about a typology for categorizing CBNRM. Taking a resource or sector-based approach is one of the most common ways of thinking about CBNRM. Community-based conservation or community-based wildlife management has one of the strongest identities in this category. These projects may revolve around single species (e.g., pheasants in Nepal) or may focus on collections of resources such as habitat preservation incorporating a range of flora and fauna conservation. Community-based conservation may be driven by a strong conservation science approach with the community element added in, or it may be more socially driven to enable the achievement of the dual goals of livelihood enhancement alongside conservation. Both approaches are not without difficulty, and sometimes controversy, and outcomes can be highly variable. For example, in many countries ownership over wildlife is vested in the state and the state may grant rights to hunt or cull wildlife through licensing schemes or permits; however, governments often retain control over the revenues generated. Community-based wildlife management schemes often involve the devolution of control and decision-making about wildlife within a designated area, though the real
extent of this devolution is sometimes debatable. Good examples in this category include the CAMPFIRE program in Zimbabwe (Taylor 2009) and Rema-Laenga Wildlife Sanctuary in Bangladesh (Chowdhury et al. 2013).

Community-based management of pasture and rangelands is an often under-recognized form of CBNRM despite, for example, arid and semiarid grasslands occupying 40% of the world's land area with 100–200 million people deriving a living from pastoralism. Commonly, pastoralists have been highly mobile, exploiting the variability of the rangelands over space and time. These grasslands may be managed through formal or informal rules governing who has access and use rights. When external pressures such as population growth, agricultural expansion, climate change, and policy shifts come into play, then these rules governing rights are often disrupted and the outcomes are complex for all stakeholders and for the rangeland ecosystem itself. Good examples in this category include community-based grassland management in western China (Banks et al. 2003) and Du Toit, Kock, and Deutsch's (2010) comprehensive work on Wild Rangelands.

Community-based forestry management is another well-known resource sector approach to CBNRM. It is estimated that most of the world's forests are publically owned (by the state) with the rest held by municipalities, communities, and individuals. Communities in many countries retain access and use rights over these public forest resources. The case of Nepal showed that while success can be well recognized, the exclusion of some households can be highly problematic. A further challenge and opportunity is coming from the United Nations' scheme to reduce emissions from deforestation and degradation (REDD+). Under this scheme, community forest management can be an opportunity for communities depending on how it is negotiated. Community involvement in REDD+ schemes is reshaping how states, the private sector, and communities interact to manage forest resources. A good example from this category can be found in Cronkleton, Barton Bray, and Medina (2011).

Community-based marine management encapsulates a wide range of different resources such as fish (through fisheries and fishing), plants (aquaculture), and reef and coastal environments (including mangroves and coral reefs). Many of these resources may be located within marine protected areas (MPAs) and increasingly it is being recognized that the poor performance overall of MPAs can be traced to a failure to effectively include local communities in the design and implementation of conservation practices. Recent efforts to incorporate community-based management systems into this sector have struggled as conservation of marine resources is often constrained by a lack of adequate policy frameworks, insufficient customary management in the face of economic or social-ecological changes, incompatibility with MPA design and policy, the risk of elite capture, or a lack of coordination and integration beyond the local level. A good example is a coastal ecosystem management project in the Gulf of Mannar in India (Samarakoon 2011).

Community-based water management is perhaps one of the least prominent but most widespread forms of CBNRM. Rural water supply programs have striven to build and strengthen community management of local water resources either at the water point scale or local catchment scale. As with other forms of CBNRM, there are complex issues governing rights of use and access over water. More recently some countries have begun to privatize the management of water supply which further changes the complex governance of this resource. Good contrasting examples of the successful and problematic
COMMUNITY-BASED NATURAL RESOURCE MANAGEMENT

outcomes of community-based water supply are Crow’s (2012) research in Kenya and Mandara, Butijn, and Niehof’s (2013) work in rural Tanzania.

Another way of categorizing CBNRM is to think about an operational distinction. In this categorization, distinction is made between “everyday” general CBNRM institutions and practices, and the more project-based “focused” CBNRM activity. “Everyday” CBNRM refers to what is happening anyway, that is, the everyday ways in which people and communities organize themselves to use and manage the natural resources around them aside from “focused” implemented schemes and activities. These “focused” activities are often initiatives funded by governments, donors, and other conservation and development agencies. These types of projects often receive the most funding and prominence but actually involve far fewer people overall than the everyday CBNRM that structures most people’s rural (and sometimes urban) lives. External CBNRM initiatives that address the landscape as a whole are not common, yet many communities are managing multiple resources, often through different initiatives and structures. Communities may have quite complex institutional configurations with village development committees, water point committees, forestry groups, and wildlife committees, and so on. Thus recognizing the difference, and interactions, between everyday practices and implemented projects is important.

Last, CBNRM can be categorized according to the power that is devolved from the state to communities. CBNRM is fundamentally based on the devolution of responsibilities, rights, and authority from central government to local communities and the bodies or committees they designate for management. The transition from centralized natural resource management to CBNRM can be measured by the level of local control over socioeconomic benefits and revenue flows from natural resource management. Globally, the term CBNRM is applied to a wide range of situations along a transitional axis from full state control toward full community control, where local people make management decisions and benefit fully from the resources. In centralized states, CBNRM is often used to describe situations where local people are mobilized as labor for government programs under state control, particularly in situations where the resource has high monetary value. Toward the middle of the CBNRM transitional axis, decision-making authority remains with central government, but certain services and administrative activities are decentralized from central government to regional and district level government and co-management provides some benefits to local people. At its most advanced, CBNRM refers to community control over resources, implemented with technical and conflict resolution support from national government agencies and district level administration. CBNRM at that point is integrated into the overall land use and income generation and livelihood strategies used by communities. Under fully devolved CBNRM, rural communities are able to evaluate local ecological conditions and decide how to manage the harvest of their fish, wildlife, and forests with technical advice (not control) from government agencies.

Table 1 summarizes the CBNRM typology discussed above, grouping the approaches according to resource/sector-based, operation distinctions, and power relationships, but does not draw comparisons between them.

CBNRM in crisis?

While CBNRM has been on the ascendancy for several decades and plays a leading role in conservation strategies worldwide, Dressler et al. (2010) argue that CBNRM is currently experiencing a
Table 1  A typology for CBNRM

<table>
<thead>
<tr>
<th>Resource/sector-based approach</th>
<th>Operational distinctions</th>
<th>Power relationships</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservation/wildlife</td>
<td>Everyday activities</td>
<td>Centralized (state holds all power, even if communities involved)</td>
</tr>
<tr>
<td>Pasture and rangeland</td>
<td>Formal focused projects</td>
<td>Devolved/decentralized (some power handed to communities but may be limited in extent)</td>
</tr>
<tr>
<td>Forestry</td>
<td></td>
<td>Community-owned (fully devolved control to communities)</td>
</tr>
<tr>
<td>Marine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Crisis of identity and purpose. These discontents can be grouped as follows: the dominant role of conservation science, the controversial politics of control, the marketization of the environment and the everyday challenges of living in poverty.

**Conservation science** occupies a powerful role in community-based conservation and CBNRM. There is a current resurgence of the “protectionist paradigm” in conservation that aims to separate conservation and development and this is having impacts on the success of CBNRM projects. This backlash upholds that an excessive emphasis on development can lead to a de-emphasis of conservation goals to the extent that they are no longer seriously addressed. The surge in protectionist conservation has sidelined some forms of CBNRM and in some instances ensured that preservation has taken precedence over indigenous people’s demands and rights over natural resources. This move “back to barriers” (see Hutton, Adams, and Murombedzi 2005) is worrying as it is antithetical to the core values of CBNRM, that of shared conservation and development outcomes. Protectionists have also worked to align devolved conservation with market-oriented solutions, a further complexity that may cloud the successful outcomes of CBNRM projects.

The politics of control is central to any contemporary analysis of CBNRM. While community-based natural resource management frameworks have at times been seen as solely conservation projects (or vice versa), they have rarely been critically evaluated in terms of development theory, which acknowledges the power and positionality of the different stakeholders. There is a critical need to examine these programs in terms of local understandings and opinions of CBNRM initiatives and local relationships with the environment. The extent to which CBNRM programs have been shaped by local priorities or government agendas will reflect the power relationships involved and the balance between conservation and development objectives within the programs. Central to the ethos of community-based natural resource management is the “participation” of local people and their “empowerment” through the development process. Yet we have seen that there is great variety in the real amount of devolved control to communities and even in so-called successful projects, the poorest are sometimes still excluded from the benefits of such schemes.

CBNRM has been used by some communities to levy new political control and openings and is heralded as a success in securing true empowerment to manage the natural resource base. However, in most cases the politics of who is in control, who trusts whom, and who has the power to make decisions and enforce actions
is highly variable and hotly contested. The differential access to resources inside local communities is often mirrored in community-based management approaches. People who use fewer resources might not depend less on them, but simply lack the entitlement for access. A broadening of the involvement of local communities via established leadership structures will not necessarily lead to the integration of everybody’s interests. Situations may occur where only particular members of a community attend meetings, while others avoid them as they perceive they have nothing to gain from participation. At a higher level, the complex administrative and policy structures have been hybridized with wider neoliberal restructuring to make the relationships between local communities, the state, and nonstate actors mired in complexity. For example, Botswana was praised for its CBNRM approach in the 1990s and the devolution of control and facilitation of joint-management activities. However, the government is now considering re-centralizing the management of natural resources away from local communities in key areas, arguing that wildlife is a state resource and the whole country should benefit. This goes against the very principles of CBNRM that suggest local communities (in close proximity to the natural resources) have a vested interest in managing them.

One of the ways that CBNRM has been molded in the twenty-first century is through the *market-based solutions* to the perceived environmental problems. The view is that if communities can see a monetary return on their investment in natural resource management, then more and better conservation outcomes are likely. Therefore joint-ventures and co-management regimes have proliferated (with the sharing of safari and tourism revenues) and more recently through global financial mechanisms, such as REDD+, for the payment of ecosystem services, such as carbon sequestration. A key driver here is that strengthening the role of the market in conservation actually moves conservation from a concern to *protect nature from the market* to a concern to *achieve conservation through the market*. This neoliberal approach is not necessarily compatible with achieving the goals of community controlled conservation and development in all cases. In the process of integrating market-based approaches with CBNRM the tendency for devolved conservation to become hybridized and bureaucratized has grown considerably. In some cases, where CBNRM is aligned with market-based solutions, it has strengthened the hegemonic position of the state rather than weakened it (e.g., Botswana).

The individuals and communities involved in CBNRM projects are facing the *everyday challenges of living in poverty*. They have to secure livelihoods, negotiate benefits, and participate in projects to manage their natural resources. Dressler *et al.* (2010) argue that CBNRM has been so reconfigured and standardized over the years that it is often misaligned with local realities. The bureaucracy of CBNRM, led by the state and NGOs, has rendered livelihood problems as concrete, with pre-assigned solutions that involve greater market-integration and intensification. Those charged with community-based design and practice have begun simplifying community problems in order to offer clear solutions already aligned with expected outcomes. Thus the process has become a technical exercise rather than an engagement in empowerment through conservation. As such, projects often become divorced from the everyday challenges of living in poverty in many of these environments.

**The future of CBNRM**

Community-based natural resource management programs were based on the premise that local
COMMUNITY-BASED NATURAL RESOURCE MANAGEMENT

populations have a greater interest in the sustainable use of natural resources around them than that of more centralized or distant government or private management institutions. These local communities were credited with having a greater understanding of, as well as vested interest in, their local environment and were thus seen as more able to effectively manage natural resources through local or traditional practices. This move in global and local development discourse was part of a wider reassessment of the goals of conservation by international bodies such as the UN (for example, the Convention to Combat Desertification and the Biodiversity Convention both advocate community-based approaches), national governments north and south (for example, indicated by the number of countries signed up to the sustainable development goals of Agenda 21 and the Millennium Development Goals), and NGOs and “community-based” organizations across the world. Several decades into CBNRM policies, there is now increasing recognition that natural resource management must be effectively linked with issues of equitable access to natural resources, the promotion of sustainable livelihoods, and the alleviation of poverty through participatory and empowering processes of development.

CBNRM arose as a powerful ideal with promise to become a near universal strategy for actors to use for rendering otherwise complex problems into manageable solutions. In the process of moving from small-scale diverse grassroots projects, where funds were low but perspective was clear, it has been scaled up as a global prepackaged solution to local problems. CBNRM’s universality may be leading to its demise. Its bureaucratic intricacies and convoluted layers of governance are making being poor more complicated. Ultimately, CBNRM needs to re-privilege social and environmental justice (such as individual and communal rights) over neoliberal logic and return to its core principles.

Adams (2008) sets out a series of points to pave the way for CBNRM to return to its core principles of community control over decision-making about the local environment to ensure social justice and socioecological sustainability of the local natural resource base. He argues that conservation-based poverty reduction should provide the focus with “pro-poor conservation” providing conservation-based natural resource management strategies that are used as a tool for poverty reduction and social justice. These strategies of sustainable use of natural resources should allow optimization of economic return and/or impacts on poorest people, not just in terms of accrued revenues. There should also be a focus on livelihood-linked conservation such that conservation can and should contribute directly to local livelihoods and generate positive benefits for local communities. Although conservation should still be the overall objective, programs should be designed to generate maximum benefits for poor people. This might include resource-sharing (or revenue-sharing) in protected areas or their “buffer zones” (e.g., wildlife tourism). Poverty neutral conservation is another principle that would ensure the net impact of conservation on the poor is neutral. Conservation can have negative impacts on the poor and should compensate fully for local opportunity costs of protected areas or laws that constrain resource use. Poverty can be a constraint on conservation success and therefore poverty needs to be addressed in order to deliver on conservation objectives. Addressing the poverty of neighbors and others who have the power to disrupt conservation programs (e.g., those who might poach or encroach on parks) can also ensure long-term sustainable success of many activities.

Ultimately, for CBNRM to be effective, policy and practice must facilitate devolved community
control over natural resources and protect the interests of the very poor to ensure social justice is rooted within socioecological sustainability. CBNRM has much still to offer to communities and to the world as both approach and practical project.

SEE ALSO: Collaborative resource management; Conservation and capitalism; Environmental issues in rural areas; Indigenous knowledge; Livelihoods; Natural resources and human conflict; Neoliberalism and the environment; Participatory development; Political ecology

References


Further reading

Community/continuum in biogeography

Christopher R. Hakkenberg
Dennis D. Tarasi
Robert K. Peet
University of North Carolina at Chapel Hill, USA

An ecological community refers to the association or assemblage of living organisms that coexist within a given space–time context. The dual concept of “community/continuum” refers to two complementary approaches to characterizing communities: as a community-unit or as a continuum of compositional variation. The community-unit hypothesis describes ecological communities as distinct and repeatable entities across the landscape. By contrast, the continuum hypothesis represents communities as idiosyncratic assemblages, conceptually represented as a region within a multidimensional continuum of independently distributed species that track environmental gradients and reflect chance events of dispersal and disturbance.

The community-unit concept came to prominence at the turn of the twentieth century through the concurrent work of continental European phytosociologists such as Josias Braun-Blanquet and North American and British plant ecologists such as Frederic Clements and Arthur Tansley. Clements argued that a biotic community is analogous to a “superorganism,” consisting of interdependent parts that together constitute a coherent whole. Envisioned as such, communities were asserted to express a significant degree of cohesion by means of positive interactions as well as similar adaptions to competition, disturbance, and environmental change. In the Clementsian view, the developmental development of communities through time is best characterized by the deterministic replacement of individuals (and species), eventually resulting in a stable “climax” community.

The Clementsian community-unit concept was challenged by Henry Gleason and colleagues, who observed that while biotic interactions affect species presence, the abiotic environment ultimately assumes the most significant role in constraining species distributions. In Gleason’s individualistic approach to species assemblages, the composition of communities is the product of overlapping yet independent species distributions, rather than discrete repeatable community units. In addition to environmental factors, the composition of any one stand is constrained by stochastic disturbance events and historical contingencies affecting dispersal and recruitment from the local species pool. While Gleason’s individualistic hypothesis was initially met with resistance, by the mid-twentieth century empirical evidence increasingly supported his view.

Skepticism regarding the concept of the Clementsian superorganism and the discreteness of the community-unit hypothesis reached a turning point in the 1950s with the simultaneous, independent development of the community-continuum hypothesis by ecologists John Curtis and Robert Whittaker. The community-continuum hypothesis describes variation in floristic composition and structure as the product of species’ individualistic responses to continuous environmental gradients. In this view, communities can be characterized as regions within multidimensional compositional space where spatiotemporal variation in community

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg1043
composition is driven by site-level environmental variables as well as biotic interactions such as competition, pathogens, and herbivory. While Curtis and Whittaker described the principle of community continuity in space and time using a Gleasonian framework (as individual species, rather than entire communities, replacing each other over time), it was their focus on community-level assemblages that distinguished their community-continuum concept from Gleason’s species-scale individualistic hypothesis.

John Curtis and his students proposed a continuum approach and developed quantitative methods to identify continuous gradients of compositional variation between forest stands, independent of environment or history (Curtis 1959). This use of numerical similarity to order vegetation samples is referred to as indirect gradient analysis, or ordination, and remains in wide use today. Recognizing that abiotic factors can strongly influence establishment and dominance, Curtis applied known environmental values such as pH and soil texture to describe and differentiate compositionally dissimilar groups. This post hoc inclusion of abiotic drivers to describe differences among communities provided an initial framework for Curtis and colleagues to examine compositional variation in vegetation as well as other groups of co-occurring organisms ranging from birds to lichens.

Concurrently and independently, Whittaker advanced the continuum hypothesis through his innovation of direct gradient analysis, an analytical method to describe and explain spatial patterns in community composition (Whittaker 1967). Whittaker contextualized compositional variation in relation to measureable aspects of the physical setting using either direct environmental gradients or complex gradients such as elevation and topographic position. Gradients were graphically depicted as orthogonal axes where one could visually distinguish distinct regions, and upon which community mosaics could be overlaid. Viewed in ordination space, the $n$-dimensional environmental gradients described by Curtis and Whittaker define species’ fundamental niche space and the stage on which biotic forces such as competition constrain species to their realized niche space. While Curtis and Whittaker represented species distributions as unimodal response curves to multidimensional environmental factors, the gradient analysis approach has since been broadened through the application of a broad array of ordination techniques, including those which incorporate polymodal response curves (see Austin 2013).

Whittaker recognized the need to categorize community types in a standard framework for community classification. Although Gleason correctly noted that no two communities are exactly alike, modern community classification efforts are predicated on the idea that one can designate community types while still recognizing compositional variation and the idiosyncrasy of individual stands. From this dualistic perspective, the community-unit and continuum concepts are concurrent and compatible models of great utility in describing complex biological phenomena, much like the morphological species concept is useful despite sometimes large amounts of intraspecific genetic variation. From this perspective, community units exist as nominal entities such that the classification of vegetation into discrete types is feasible, provided the arbitrary character of their delineation is recognized.

Terrestrial community types are most commonly, though not exclusively, characterized by dominant or diagnostic flora. Variation among communities can be quantified by the degree of compositional similarity/dissimilarity between them, and used as the basis of a synthetic community classification system whose typological structure is determined by statistical relationships and expert opinion. Communities can
be classified as belonging to a single, exclusive “crisp” type, or assigned to a set of “fuzzy” types ranked according to relative membership probability. These classification efforts recognize that species associations are empirically observable and that broadly defined community types exist in replicate over landscapes, while still acknowledging that individual assemblages are fluid and dynamic in composition. Though some disagreement remains between national schools regarding the best methods for organizing community types into a coherent classification structure, recent efforts have aimed to globally standardize classification protocols for consistent standards across geographic regions (see De Cáceres et al. 2015 for a review).

Although most ecologists today recognize the utility of the Gleasonian concept of individualistic species assemblages and the continuum model underlying community classification, debate continues regarding the relative merits of these models for use in different contexts. Part of this debate may be attributed to different methodological approaches or scales of reference, with no one perspective inherently correct. Where the community unit is an abstraction based on geographical space, the community continuum reflects an abstract environmental space in the context of biological–environmental relationships (Austin 2013). For land managers and policymakers, envisioning communities as convex volumes in environmental hyperspace may not be as useful as characterizing them as discrete units upon which management plans can be effectively implemented. On the other hand, recent trends in landscape ecology emphasize communities as continuously intergrading over larger spatial extents. Species distribution models increasingly reflect abiotic proxies and likelihoods for species presence as a wall-to-wall raster grid, rather than a patch mosaic of discrete community polygons (see Evans and Cushman 2009).

Land cover maps based on remotely sensed imagery provide another arena to highlight the distinction between the two approaches. While remote sensors detect energy from the Earth’s surface and convert these measurements to a continuous range of integer values, the land cover maps produced from these pixel arrays are the result of the classification of continuous data that partially reflect the subjectivity inherent in the choice of sensor design and classification algorithm. For example, the US National Gap Analysis Program (GAP) uses remotely sensed imagery to produce wall-to-wall land cover maps that depict vegetation as discrete polygons representing plant community types. Ultimately, both the community-unit and the individualistic-continuum perspectives possess their own distinct advantages in describing the nature of ecological communities. With climate change and other global environmental trends demanding disparate responses from scientists, policymakers, and land managers, the practical imperative to classify community types into discrete classes will remain, even as projections for how biotic assemblages will respond to global change will be conducted within a framework of continuous compositional variation.

SEE ALSO: Biodiversity; Biogeography; Climate change and biogeography; Landscape ecology; Nature conservation; Niche theory and models; Phytosociology

References

COMMUNITY/CONTINUUM IN BIOGEOGRAPHY


Comparative advantage

Godfrey Yeung
National University of Singapore

Comparative advantage is defined as the capacity of an agent, a region, or a country to produce a specific commodity or provide a service at a lower opportunity cost (the value that one has to forgo) over another agent, region, or country. This concept is based on the relative cost ratio for one agent to produce a commodity compared to the cost ratio of another agent. It is one of the fundamental concepts in international economics and is normally used to support free trade and pertinent policy agendas in the Washington Consensus. The original concept of comparative advantage and the neoclassical international trade model are explained before a brief overview on the propositions outlined by heterodox economists and neo-Marxist social scientists.

Ricardian comparative advantage

The concept of comparative advantage was first mentioned in Adam Smith’s seminal book (1904/1776) *An Inquiry into the Nature and Causes of the Wealth of Nations* and this brilliant idea is elucidated in David Ricardo’s (1973/1817) classic *On the Principles of Political Economy and Taxation*. Under the scenario of a barter economy with a fixed exchange rate of 1:1, he illustrated the concepts of absolute and comparative advantages with the opportunity costs of producing wine and cloth in Portugal and England. In Portugal, the cost of producing one unit (gallon) of wine was 80 labor hours and one unit (yard) of cloth took 90 labor hours (Table 1). In England, the corresponding opportunity costs were 120 and 100 labor hours, respectively. Portugal clearly had the absolute cost advantage over England (i.e., it used fewer labor hours) to produce both commodities: their cost ratios are 80:120 (0.66; Portugal utilizes 66% of England’s labor hours to produce the same output) in wine and 90:100 (0.9) in cloth. Naturally, this suggested Portugal should produce both types of commodities and England should just buy (import) them from Portugal.

Instead of focusing on the cost ratios or absolute cost advantages between Portugal and England alone, Ricardo (1817) argued that the cost ratios of producing both commodities within each country needed to be considered. Portugal then had a comparative advantage in wine production while England had a comparative advantage in cloth production. The opportunity costs of Portugal to produce one yard of cloth was equivalent to 1.125 gallons of wine, that is, $1 + [(90 - 80)/80]$, while the opportunity costs of England to produce one gallon of wine was 1.2 yards of cloth, that is, $1 + [(120 - 100)/100]$. Therefore, both countries should specialize in producing the commodity that gave them a comparative advantage and trade with each other. In other words, the best course of action was for Portugal to produce wine instead of utilizing 90 labor hours to produce one yard of cloth because it could get 0.125 yards more cloth by trading with England (at the 1:1 barter exchange rate), that is, better to use 90 labor hours to produce 1.125 gallons of wine and exchange it for 1.125 yards of cloth from England (so the 90 labor
COMPARATIVE ADVANTAGE

Table 1  Opportunity costs of producing wine and cloth in Portugal and England.

<table>
<thead>
<tr>
<th></th>
<th>Labor hours required to produce</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 gallon wine</td>
<td>1 yard cloth</td>
</tr>
<tr>
<td>Portugal</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>England</td>
<td>120</td>
<td>100</td>
</tr>
</tbody>
</table>


hours yielded one yard of cloth directly and another 0.125 yards of cloth indirectly). By the same token, it was better for England to produce cloth rather than utilize 120 labor hours to produce one gallon of wine because a further 0.2 gallons of wine are available by trading with Portugal, that is, the 120 labor hours to produce 1.2 yards of cloth could have been exchanged for 1.2 gallons of wine from Portugal. Rather than each country trying to produce all the commodities they need, Ricardo further argued that countries can gain from trade if they produce commodities for which they have the comparative advantage and then trade them for commodities from other countries.

The neoclassical Heckscher–Ohlin–Samuelson trade model

In the original two-country and two-commodity model, Ricardo showed that even if one country had the absolute advantage in producing both commodities, both countries could benefit from trade if they specialized in producing the commodity where they had comparative advantage, that is, the opportunity costs of producing that commodity would be lower than those of their trading partner. In Ricardo’s example, both England and Portugal would have gained from trading, even though Portugal had a cost advantage for producing both wine and cloth. In the Ricardian version of comparative advantage, differences in the technologies employed by labor in different countries determine costs and comparative advantages.

However, Ricardo’s extraordinarily influential conceptualization of comparative advantage is based on three highly restrictive assumptions. First, his illustration is based on a two-country and two-commodity barter economy with a fixed exchange rate of 1:1; no currency and exchange rates are considered (see Dunford et al. 2014 for the importance of wage costs and exchange rates for international trade). Second, labor, one of the factors of production (labor, capital, and land are the three main factors of production), is immobile. Third, there are no transaction costs (costs incurred in the collection of information and the processing of exchanges, including transportation) in international trade.

The Ricardian conceptualization of comparative advantage was further developed by Eli Heckscher and Bertil Ohlin, and later enhanced by Paul Samuelson into the canonical neoclassical Heckscher–Ohlin–Samuelson (H–O–S) model of international trade. To address concerns about the restrictive assumptions of Ricardian comparative advantage, Ohlin (1967/1933) relaxed the single factor assumption in the two-country, two-commodity, and two-factor (labor and capital) model and argued that comparative advantage, geographical specialization, and trade are driven by regional factor endowments in his book Interregional and International Trade. According to the Heckscher–Ohlin theorem (H–O theorem), countries export products that make intensive use of production factors (labor and capital) and natural resources that are relatively abundant and relatively inexpensive. In other words, a capital-abundant country exports capital-intensive goods to
labor-abundant countries while importing labor-intensive commodities in return.

Neoclassical economists therefore argue that the international division of labor (specialization of production on a national basis) is a natural consequence of the spatial differentiation of labor productivity (Ricardian comparative advantage) and (Heckscher–Ohlin) factor endowment. As the production factors (labor and capital) are mobile within countries but immobile between countries (as stipulated in the H–O theorem), and markets are perfectly competitive, factor prices will be equalized across countries as a result of international trade, according to Samuelson’s factor price equalization theorem (see also Barro, Mankiw, and Sala–i–Martin 1995; Mankiw, Romer, and Weil 1992). Therefore, the neoclassical theory of international trade hypothesizes that incomes across countries will converge and development will not be uneven. In other words, free trade promotes economic growth and reduces regional disparity across space.

It is unquestionable that the H–O–S model of international trade is a brilliant extension of Ricardian comparative advantage. However, it is unable to reconcile the uneven economic growth observed between developed and developing countries since World War II. The H–O–S model still suffers from a set of, although much less restrictive, unrealistic assumptions: the production function has constant returns to scale (i.e., no accounting for technological advancement), production factors are immobile across space (countries), and markets are perfectly competitive.

To address the lack of solid empirical evidence for the convergence of income across countries hypothesized by the H–O–S theorem, international trade economists, notably Paul Krugman, relax the constant returns to scale assumption and argue that the divergence of income between developed and developing countries is due to the increasing returns to scale of investment among the former in the new trade theory.

The new international division of labor and the neo-Marxists’ interpretation

Instead of the “win-win” scenario portrayed by neoclassical international trade theory in general and W.W. Rostow’s type of modernization theory in particular, heterodox economists and neo-Marxist social scientists argue that there is limited empirical evidence to support the H–O–S theorem. As there is increasing divergence of income across countries, especially between developed and developing countries since World War II, heterodox economists argue for the protection of infant industries in developing countries. Newly established industries in developing countries are unable to compete with the established giants in developed countries, hence, they need to be sheltered by import tariffs and quotas until they become competitive internationally. Hans Singer and Raúl Prebisch have similar policy recommendations based on their Singer–Prebisch thesis: developing countries are under unfavorable and deteriorating terms of trade (the relative price of exports compared to imports) and suffer from trade deficits due to the low export prices of raw materials and minerals relative to the higher import prices of machinery and manufactured goods.

The increasing dominance of transnational corporations (TNCs) in foreign direct investment (FDI) in developing countries, an important phenomenon of the new international division of labor due to the relocation of labor-intensive manufacturing activities from developed countries to Asian developing countries since the 1970s, fuels further debates about the effects of international trade on development. In his book, *Kicking Away the Ladder: Development*
Comparative Advantage

Strategy in Historical Perspective, Ha-Joon Chang (2002) highlighted the historical paradox that advanced economies (like the United States and the United Kingdom) used interventionist and protectionist economic policies to accumulate capital during their early stages of economic development yet bar developing countries from following similar pathways today.

Neo-Marxist social scientists go further by highlighting the exploratory nature of free trade and the subsequent dependent relationships between developed and (newly independent, especially former colonies) developing countries. Dependency theorists, notably Andre Gunder Frank, argued that the existing model of capitalist development promoted the “development of underdevelopment” in developing countries. He illustrated his argument by referring to the postcolonial (under)development experience in Latin America in his book Capitalism and Underdevelopment in Latin America (1967). By tracing the history of development in sixteenth-century Western Europe and America, in The Modern World System (1974) Immanuel Wallerstein articulated similar arguments in his world systems theory. Despite the fictitious nature of capital and the endless accumulation of capital by competing agents (see below), Wallerstein argues that the world economy is sustained by a dominant capitalist core (developed countries), which extracts surplus value (resources and capital) from the periphery (developing countries, especially former colonies, and to a lesser extent, semiperipheral countries at the developmental stage between the core and periphery) through a complex network of economic exchanges (including unfavorable terms of trade) and their control over the economic and political institutions. In an edited volume, The World System: Five Hundred Years or Five Thousand (1993), Andre Gunder Frank and Barry Gills further contributed to the debate by pointing out that the global and economic power of a nation is largely determined by its historical circumstances and geography.

Instead of lack of integration into the international trade, as normally argued by neoliberal trade economists, dependency theorists argue that developing countries are perpetually locked into a world system of core–periphery dependency where the core (developed) countries dominate the advanced technologies and global political and economic institutions while the developing countries are locked into unfavorable terms of trade and remain exploited.

Neo-Marxist geographers have typically highlighted the fictitious nature of capital in the debates. In his seminal work The Limits to Capital (1982, 2006), David Harvey pointed out that the relocation of manufacturing sectors from developed to developing countries is a spatial fix of TNCs due to the overaccumulation of fictitious capital and the intrinsic crisis nature of capitalism rather than the results of FDI and free trade.

Harvey’s geography of capitalism developed from the theoretical foundation of Karl Marx’s classical work on “fictitious capital” about the speculative quality of credit money that results in excessive accumulation. In Capital (1906/1867), Marx argued that capital has a dual role. As productive capital, it lubricates circulation and exchange by allowing the purchase of commodities in production and the exchange of transformed commodities back into money before being reinvested in production (through the financial system) and thus maintains the circulation of capital and the operation of capitalism. As commercial capital, it is an expression of value. Under the guidance of “miserly thinking” (see Doel 2009) where “scarcity” (the world is unable to fulfill insatiable consumption demands) and market prices control allocation, capitalists maintain the operation of neoliberal
capitalism by accumulating the surplus value produced by the working classes through primitive accumulation, such as privatization and financialization (Harvey 1982, 2006). (According to Karl Marx, surplus value is the difference between the value of money as capital used to purchase commodities and the value contained in the commodities transformed by labor.)

Should an investment involve some form of credit, financial capital remains a form of speculative holding (i.e., the value of commodities is “fictitious” and does not materialize until they are sold) in the circulation of capital and could be devalued. Financial institutions then act as intermediaries to facilitate the market for fictitious capital in the circulation of surplus value by offering interest payment to capital owners and charging interest premiums to capital borrowers. Once the expansion of production capacity goes beyond market demand, however, capitalists compete for limited opportunities to deploy their capital profitably, for example too much machinery and materials competing for too little labor power results in a reduction in surplus value. Subsequently, financial institutions may not be able to pay interest to capital owners due to the diminished surplus value generated from production, and hence the devaluation of the fictitious capital. The owners of capital thus withdraw it from circulation by liquidating their fictitious capital holdings and hoarding them as commercial capital, and this overaccumulation of capital leads to its further devaluation. Marxists believe this withdrawal of productive capital from the system happens from time to time, and the economic system is thus inevitably prone to cyclical and systematic crises as state intervention is unable to correct the intrinsically destructive force of the excessive and unsustainable accumulation of capital and the devaluation of surplus value through the circulation of capital (Harvey 1982, 2003, 2006).

Comparative advantage versus competitive advantage

Instead of comparative advantage, official policy papers prepared for governmental and nongovernmental institutes these days normally use the term “competitive advantage.” Competitive advantage was coined by Harvard Business School guru Michael Porter. In his seminal work Competitive Advantage: Creating and Sustaining Superior Performance (1985), he argued that countries have a competitive advantages either by producing commodities at lower costs or at different specifications or forms where consumers are willing to pay higher prices. In a major departure from the utilization of existing comparative advantages and factor endowments outlined by the H–O–S theorem, he emphasized productivity growth should be the focus of national strategies as countries cannot depend on nonsustainable low labor costs and the exploitation of natural resources alone.

Interestingly, Project Socrates was commissioned by the Reagan administration in the United States to establish the foundation for a technology-based competitive strategy development system that could be used to guide international trade policy while the Washington Consensus was germinating in the 1980s.

SEE ALSO: Corporations and global trade; Dependency theory; International division of labor; Trade, FDI, and industrial development

References


Competitiveness

Luis Rubalcaba  
*University of Alcalá, Spain*  
*World Bank, USA*

Competitiveness is about the ability and effort required to grow and to improve the living standards of citizens, organizations, social communities, cities, regions, and countries within the international global arena. In economic terms, it is usually defined as the performance of firms, regions, and countries in the global economy.

The concept of competitiveness

The term “competitiveness” is a rather vague, fuzzy word that can mean different things. Synonyms explaining what it means to be competitive are numerous: productive, good, best, cheap, efficient, effective, and so on. “Competitive” is defined by Oxford Dictionaries online as “relating to or characterized by competition” (synonyms given are ruthless, merciful, aggressive, fierce), “having or displaying a strong desire to be more successful than others” (synonyms are ambitious, driven, vying, combative, contentious, aggressive; insistent, driving, pushing, zealous, keen), and “as good as or better than others of a comparable nature.” Therefore, competitiveness is about success and improvement when judged against others. In economic terms, this is not necessarily the case, with concepts ranging from the competitive capabilities of workers, the market position of companies, the ability of a region, a country, or a group of countries to face global competition and to export more than they import, to the ability to steadily increase their participation in global supply, support, and raising income levels. All these alternatives assume the continued growth of productivity and hence the incorporation of technical progress in production processes.

The notion of competitiveness usually involves taking a position in relation to two factors. Firstly, it is a comparative term that requires comparisons to be made between countries, regions, companies, or individuals. Second, it is a term that presupposes a capacity for countries and other entities to modify their position in such a comparative context. The diversity of meanings has not prevented competitiveness from being used as a tool for analysis in economic geography, since it is basically about how organizations and countries are acting in new, global, interrelated spaces (regions, cities, etc.).

The construction of economic activity requires, at a given time, that distinctive features are present. These are the differentiators relative to the current state of things. These differentiators may be based on price, quality, reputation, or any other attributes that help to create comparative advantage. The success of economic development involves competing against the previous situation and with or against others. In its etymological sense, “competitiveness” is the result of a tension within a competitive jurisdiction; it is the action of “competing with.”

The term “competing with” involves three aspects. First, there must be a plurality of agents (at least two). Second, there must be a common aspiration (something shared, an objective). Third, some effort must be exercised (to make an attempt, a strategy, and a deliberate engagement).
COMPETITIVENESS

Following from this, competitiveness implies an ontological dimension (e.g., what a company, a region, or a country is in relation to others, or what a country is able to do) and a behavioral dimension (what a company, region, or country does and how it affects others). To understand this distinction, an analogy can be drawn with human behavior. What defines a person’s social dimension is “being with,” or what it means to “be with and to others,” rather than specific behaviors “with or against others.” A person is a social being by nature but can suffer antisocial deviations or enjoy high levels of sociability. The ontological significance of competitiveness, first of all, implies an improvement of an agent’s position in an environment in which others try to do the same. The individual who improves significantly against himself/herself, regardless of the impacts on others, is the competitive one. Competitiveness is more than an action; it also requires the ability to compete. In this sense, economic competitiveness at an international level must be defined as the economic fitness of two or more countries or regions to improve their economic welfare. They compete, for instance, to achieve higher levels of employment, income, or quality of life.

According to this concept, competitiveness is about the ability and effort required to improve the living standards of citizens, organizations, social communities, and countries, not only or even primarily to increase market shares abroad (this would be a behavioral meaning of competitiveness, “compete against”). In economic terms, it is usually defined as the performance of firms, regions, and countries in the global economy. To the extent that international trade is part of the economy of a country or a region, competitiveness concerns international trade, among other things.

If the ontological dimension of competitiveness is not taken into account, policy actions would be understood in the context of conflict behaviors and zero-sum games in which one country or region, for example, wins and another one loses. Collaboration with the competitors should be avoided in this approach. However, the world’s economies are not mostly zero-sum games and everybody can win from international trade; this makes collaboration highly positive for the competitiveness of everybody.

Definitions of competitiveness

There are many definitions of competitiveness. BusinessDictionary.com defines it as the “Ability of a firm or a nation to offer products and services that meet the quality standards of the local and world markets at prices that are competitive and provide adequate returns on the resources employed or consumed in producing them.” This definition clearly relates to business success. In relation to regions and countries, productivity appears to be the key determinant of competitiveness. According to Michael Porter (2004, 50) competitiveness is defined by the productivity with which a nation utilizes its human, capital, and natural resources; the starting point must be a nation’s underlying sources of prosperity. Productivity is at the center of a country’s standard of living and the way to measure competitiveness. For Porter, what matters most is not exports per se or whether firms are domestic or foreign-owned but the nature and productivity of the business activities taking place in a particular country. This also applies to individual regions, cities, and productive regional clusters.

There are two institutions in Switzerland that publish the key international reports on competitiveness, including definitions. In the World Competitiveness Report (WCR) Sala-i-Martín et al. define competitiveness
as the set of institutions, policies, and factors that determine the level of productivity of a country. The level of productivity, in turn, sets the level of prosperity that can be reached by an economy … a more competitive economy is one that is likely to grow faster over time. (2014, 4; emphasis original)

For Sala-i-Martín et al. there are two components of competitiveness: static and dynamic. The former refers to the ability to sustain a high level of income while the latter is related to growth potential. A rather different approach is employed in the IMD World Competitiveness Yearbook (ICY), where Bris and Caballero (2015, 2) state that competitiveness is “the ability of a country to facilitate an environment in which enterprises can generate sustainable value.” Thus, the WCR focuses on factors behind productivity growth while the ICY mainly focuses on factors behind value creation, mainly at firm level.

Based on the previously defined ontological dimension of competitiveness, a more general definition is provided that is not only about the productivity capacity or value creation in countries. In this instance competitiveness can be defined as the capacity (ability and effort) of people, organizations, cities, regions, and countries to create growth in living standards within the global arena. In economic terms, at country level, this growth can mainly be measured by productivity and value creation but welfare gains are also included. Under this overall umbrella, there are different types of competitiveness: trade competitiveness (the most used connotation in international economics), business competitiveness (at firm level), industrial competitiveness (at sector level), regional competitiveness (highly relevant for studies of cities and regions), social competitiveness (performance in terms of social indicators such as poverty and shared prosperity), and global competitiveness (the holistic approach followed in the WCR and ICY).

Each of these specific concepts requires specific indicators and measures, such as the regional profiles compiled using the EU Regional Competitiveness Index (Annoni and Dijkra 2013) or the sectoral approach for service industries based on comparative measures in productivity, internationalization, innovation, regulations, and competition (Rubalcaba 2007).

The social context

The notion of competitiveness has important social as well as economic implications. The ongoing wave of globalization largely explains the interest in this topic since the 1990s, beginning with the international expansion of trade in the 1960s. Large global flows enabled by a revolution in information and communications technology (ICT) emphasized the deepening economic relations between peoples, markets, cities, and countries. The first great wave of modern globalization between 1870 and 1914 was linked to the political and industrial strength of the colonial powers and it was relatively easy for politicians to lead the process. The European Union emerged after World War II with the objective of political peace and reconciliation, in which the economy served as a cornerstone. In today’s globalization, it is no longer so evident who leads the processes; economic power has often proved stronger than political power.

In this new context, to be competitive has more than ever become a priority in a more interconnected world. The welfare of workers, employers, shareholders, and citizens is increasingly linked to their relative position in the hierarchy. Workers therefore seek to improve their educational levels, language abilities, and technical skills in order to be able to compete for better jobs and a higher degree of employability. Employers must be alert to increasing
COMPETITIVENESS

competition in a world offering a higher level of opportunities but also with certain vulnerabilities. Shareholders and investors are increasingly dependent on the evolution of the major exchanges in the leading global financial centers (such as London, New York, Sydney, Shanghai), where they must compete to identify and acquire the best information. Citizens struggle to improve their living standards and compete for the best services in schools, housing, utilities, and all kinds of other public and private goods and services. Cities and regions are looking for ways to become more competitive in order to attract trade, investment, and economic activity that will enhance their economic and social welfare. Being competitive has become an unavoidable action for achieving improvements in welfare.

Competitiveness has reinforced the value of dynamism and flexibility. Not only is the competitiveness of a company defined by its income and profit, but it must also be able to continuously adapt to new requirements and to keep rethinking its goals and strategies. A sector is competitive when new firms generate new value, creating comparative advantage. A city or a region is competitive when it achieves sustainable levels of local development. A country is competitive when it is able to attract income and wealth over and above that arising from short-term economic development. In essence, competitiveness is an evolutionary concept.

The academic debate

Krugman (1994; 1996) argued that there are dangers associated with the careless application of the term “competitiveness”; its use should be restricted to well-defined cases. At best, it should be used synonymously with “productivity.” This is based on the assumption that, on the one hand, competitiveness mainly refers to international trade and, on the other hand, the economics of countries does not work in the same way as for businesses or firms. At a business level, there are zero-sum games, whereas this is not the case at international level where there are plenty of win–win transactions. Krugman later wrote on his New York Times blog (2011):

It’s OK to talk about competitiveness when you’re specifically asking whether a country’s exports and import-competing industries have low enough costs to sell stuff in competition with rivals in other countries; measures of relative costs and prices are, in fact, commonly – and unobjectionably – referred to as competitiveness indicators. But the idea that broader economic performance is about being better than other countries at something or other – that a country is like a corporation – is just wrong.

This position is useful in that it alerts us to a mistake often made by some policymakers when they blame external conditions in order to cover up internal problems, for example, when they set up protectionist measures instead of making domestic reforms to improve competition and productivity. However, Thurow (1994) and Cohen (1994), among others, did not concur with Krugman’s view of the role of external trade in domestic economy, while the IMD World Competitiveness Center (2014) notes that “Some scholars claim that nations themselves do not compete, rather their enterprises do. Nations compete because world markets are open.” To what extent do nations compete or not compete? These questions underpin disagreements between scholars’ views on competitiveness.

The aforementioned ontological dimension seems to support the Krugman side of the debate, since it is not restricted to international trade and closer to the win–win games view proposed by neoclassical theories. However, there is also room for behaviors such as zero-sum games. Following previous findings (Rubalcaba 2002),
what is right and wrong depends mainly on three factors: the geographical units and their sizes, the performance of these geographical units, and the dynamics of economic systems. The more that small and micro(geographical) spaces are considered, the more likely it is that a zero-sum game may exist. On the other hand, the bigger the units (large countries, large regions), the more likely it is that nonzero games exist. However, size is only one factor; it is also the case that not all geographical units have the same institutional and productive endowments. Nor do they have the same settings and capacity to change specialization patterns, mobilize resources to shift to more competitive sectors, and adapt to the advantages of international trade. Geographies evolve over time in such a way that the balances between win–win and zero-sum games will depend on system dynamics and evolutionary efficiency. Therefore, the scale, the flexibility, the quality, and the dynamics of each geographical space matter.

The drivers of competitiveness

The drivers of competitiveness will vary, depending on which concept of competitiveness is adopted. If it just involves international trade, division of labor in international markets, and market share, then factors related to business trade performance, such as costs, prices, and product differentiation will be very important. This is also true for some institutional and macroeconomic factors affecting trade, such as exchange rates, inflation rates, and the business climate for trading and investment. The relative importance of each driver can be classified and rated, depending on the theoretical model. Some examples include the classical absolute advantage theory (Adam Smith) which is based on intrasectoral advantages; the neoclassical thinking model (David Ricardo) which is based on relative comparative and intersectorial advantages; and the Heckscher–Ohlin–Samuelson (O–H–S) model which is based on general equilibrium and different factor endowments. If trade competitiveness is not substantially determined by price or cost factors, which is very often the case in both goods and services markets, it is referred to as the Kaldor (1978) paradox, whereby the market shares of sectors abroad are driven by factors not related to prices and costs; due to quality and differentiation, industries and countries may have more exports and better market shares in the context of higher prices and costs.

When competitiveness is approached from a more macro and holistic angle, all the factors that influence productivity, value creation and growth play a role. Productivity and growth are determined, in particular, by institutions, policies, markets, agents, and geography at the macro, meso, and micro levels. Michael Porter (1998/1990) proposed a “diamond” framework to help to understand what determines competitiveness at various local, regional, and country levels. The diamond comprises four pillars: (i) factor conditions (relative factor endowments, new factor dynamics, and intra- and interfactor dynamics); (ii) demand conditions, referring to the composition of domestic demand, size and patterns of demand growth, internationalization of domestic and demand dynamics; (iii) support-related sectors, for which a nation may have a competitive advantage in supplying sectors or related services around a particular industry; (iv) strategies, structure, and rivalries of companies, referring to the role of corporate competitiveness strategies (strategy and structure of domestic firms, goals and objectives, etc.). Within this diamond framework there is also room for the effects of chance and for public intervention.
COMPETITIVENESS

Recently, a more comprehensive approach has been proposed. The World Economic Forum (WEF) uses 12 pillars of competitiveness: institutions, infrastructure, macroeconomic environment, health and primary education, higher education and training, goods market efficiency, labor market efficiency, financial market development, technological readiness, market size, business sophistication, and innovation. Using these factors in combination, countries are placed along a continuum ranging from factor-driven economies to economies driven by efficiency and innovation. Innovation is the ultimate stage of development and is closely related to both economic and institutional factors. Technology change (Fagerberg 1996) and national systems of innovation (Lundvall 1992; Nelson 1993) are also highly relevant in this context. The World Competitiveness Center utilizes a similar set of factors but in a different way. Four pillars are proposed: (i) economic performance; (ii) business efficiency; (iii) government efficiency; and (iv) infrastructure. Four other dimensions are then analyzed for these pillars: attractiveness versus aggressiveness, proximity versus globalization, assets versus processes, and individual risk-taking versus social cohesiveness. Such approaches to global competitiveness rely on a comprehensive multidimensional view, with each having the potential to highlight important implications for policymaking.

Competitiveness policies

Whether competitiveness is understood in a narrow sense as a position on the international trade market (with other countries or regions) or in its traditional sense as improvement and superiority in key variables of economic growth and welfare (employment, income, etc.), identifying and utilizing appropriate policies is an important task. First, policies designed to improve the competitive position of a country or region relate primarily to foreign trade, although with obvious interrelationships to other policies. Second, specific competitiveness policy and economic policy in general largely overlap: policies that improve economic and welfare performance do establish the full range of possible policies, namely stabilization, structural, macro- and microeconomic policies. A competitiveness policy is, in this sense, a comparative economic policy in that the purpose is to analyze the factors that influence the relative positions of countries or regions. However, since the relative differences between countries and regions are related to structural factors and long-term policies, competitiveness policy tends to privilege the structural features that are linked to factors of production. Short-term stabilization policies thus remain at the level of macroeconomic policy formulation for a country as a whole, even though this may influence national or regional competitiveness.

From the perspective of regional or urban economic policy, six types of policy can shape competitiveness and competitive wellbeing.

- **Trade policy**, understood as directed toward the growth of international trade in the region and attracting investment; this is a policy of actual trade competitiveness.
- **Productive factors policy**, comprising natural resources and infrastructure, labor market and human capital, physical capital and financial capital. All affect competitiveness in general, and trade competitiveness in particular. The human capital factor is a central element in this set of policies.
- **Market and industrial policies** relate to the environment and markets in which industries are operating, and include state intervention in the markets for goods and services (regulation, competition policy, business climate)
and industrial policy in the traditional sense (clusters policy and industrial development).

- **Innovation policy** as a competitiveness element: here, incentives are used to encourage innovation in goods and services provided by the market, especially if market or system failures are detected. Innovation policy may be incorporated into technology and research and development policy as well as programs directly aimed at innovation and adoption and absorption of technology and knowledge, including labor market and employment policies (active and passive), social benefits (social security, pensions, disability), environmental policy, and income redistribution;

- **Urban planning policy**: the organization of space in cities and regions is a key factor rendering the city or region attractive to trade and investment. Also fundamental is support for the functional needs of the markets, such as logistics or the production of high-quality residential environments for workers. This is especially challenging when devising policies aimed at enhancing the availability of skilled professionals in knowledge-intensive services.

### Concluding remarks

Competitiveness is important, but studies of it are still at an early stage of development. Most date from the 1990s onward, even though Adam Smith and others mooted theories of competitiveness as far back as the early eighteenth century. A future research agenda may include further development of conceptual frameworks; addressing measurement issues; the new factors emerging as competitiveness drivers; the role of new global value chains and policy design. Competitiveness is a complicated multidimensional phenomenon; a multi- and pluridisciplinary approach will be important for further research. Geographers, economists, sociologists, urban planners, marketing and business researchers, and practitioners should work together more to deepen understanding of a process that ultimately has an impact on the economic and social wellbeing of us all.

**SEE ALSO:** Comparative advantage; Globalization; Industrial location theory; International division of labor; Internationalization

### References


Further reading

Complexity in biogeography

Daehyun Kim
University of Kentucky, USA

Complexity is generally defined as a property of a system that results from the interactions between many basic components according to fundamental rules, with a potential to generate unexpected order (Levin 1998). In this broad definition, how such system components interact with one another is not stated explicitly due to differing views among geographers on the relationships between nonlinearity and complexity. Malanson (1999) considered that multiple variables per se do not constitute complexity; complexity exists when these variables are dynamically intertwined over space and time in a nonlinear way. This perspective is in accordance with the notion that feedbacks that lead to nonlinearity and, eventually (if not always or inevitably), to self-organized criticality characterize nonequilibrium systems, which are the building blocks of the science of complexity (Holland 1995).

However, Phillips proposed that “not all nonlinear dynamics are complex; not all complexity is nonlinear” (Phillips 2003, 4). To him, nonlinear dynamics and complexity have their own conceptual territory, although they share a commonality, termed complex nonlinear dynamics (CND), which represents the complexity that arises from intrinsic nonlinearities within a system. Put another way, CND simultaneously implies a subset of nonlinear system behaviors that lead to complex behavior and a subset of complexity that arises from nonlinearities. This entry will not evaluate which of these two perspectives is more relevant to the contemporary literature of biogeography; in either case, nonlinearity is an important part of the conceptual background of the complexity theory.

The complexity theory resulted from recognizing the inability of the traditional reductionist approach to explain how higher-level patterns of physical and biological systems across space and time emerge from and control localized interactive processes occurring at lower levels. Simply put, in biogeography, the detailed study of individual trees with regard to their physiology and anatomy (i.e., reductionism) provides limited insight into how they compete with and/or facilitate one another to develop a population. The systems in which such feedbacks occur between higher and lower levels are regarded as typical examples of complex adaptive systems (CAS) (Holland 1995).

The limitations of reductionism arise from several related properties of CAS, including nonlinearity and emergence. Nonlinearity is one of the cores in the behavior of CAS, and it makes a whole larger (or smaller) than the sum of its parts. A CAS cannot be investigated fully by viewing each of these parts in isolation because they continuously interact within positive or negative feedback loops. Salt marshes provide a biogeomorphic example: pioneer plants begin to colonize the uppermost portion of tidal flats, and they increase sedimentation rates by reducing the energy of tidal flows and by trapping suspended fine particles during submergence. As a result, the surface elevation of the flat increases (i.e., decreases in waterlogging frequency and...
duration). Moreover, the early-successional plants gradually ameliorate the salinity levels of sediments. This, in turn, facilitates the growth of plants and augments the height and density of vegetation cover, further enhancing the sedimentation process. This positive feedback mechanism explains how the marsh surface constantly readjusts itself toward equilibrium as sea level changes. In such a CAS, nonlinear dynamics may lead to chaotic behavior such that minor variations in the initial conditions across the marsh surface are magnified over time. In the example above, salt marsh plants are considered ecosystem engineers that modify their surrounding physical conditions and modulate the availability of resources to other species and/or themselves.

Another important concept in CAS, emergence (or emergent properties), is a feature or order that generates spontaneously without any definable rule during the process of self-organization. Although emergence is an unexpected property, it still constitutes coherent and integrated structures in CAS. In the biogeographical literature, emergence has often been associated with spatially divergent self-organization in the evolution of landscapes, or spatial differentiation of landscapes from a homogeneous to a heterogeneous pattern (Phillips 1999). This divergence can be characterized by entropy, which is an appropriate measure of the organization as in the following model:

$$\bar{\delta}(t) = A e^{\lambda t}$$

where $\delta(t)$ is a perturbation event at time $t$, $A$ is a constant (or initial variation) at $t = 0$, and the arrows represent vector quantities. $\lambda$ is a Lyapunov exponent. The number of Lyapunov exponents in a system indicates the number of system components. Therefore, when there is at least one positive Lyapunov exponent (i.e., $\lambda > 0$), the associated system should experience an exponential increase of variation, that is, unstable and divergent landscape evolution. Overall, such a situation is defined as deterministic chaos. The concept of Lyapunov exponents is also related to the Kolmogorov (K-) entropy, which is calculated by summing up all positive Lyapunov exponents. A positive K-entropy implies a magnification of the initial spatial variation of, for instance, soil nutrients over time. This nonlinear chaotic behavior of soil resources may in turn induce a differentiation of vegetation cover into diverse spatial units.

In salt marshes, environmental controls, such as elevation and/or salinity, greatly influence early-successional vegetation patterning and dynamics over space. Over time, as biological interactions between plant species proceed and organic/inorganic matter accumulates, patterns of vegetation dynamics become stochastic and complex on the increasingly mature marsh surface. Therefore, researchers have often proposed that distinct plant species composition may be found even at locations having the same surface elevation. Such an unexpected compositional variability produced by internal marsh processes (i.e., self-organization) can be considered an emergent property. These discussions imply that vegetation patterns and dynamics in salt marshes cannot be fully understood or explained by elevation only; thus, the traditional elevation-centered perspective may be an oversimplification of the complexity of tidal marshes. Similar ideas of augmented complexity and stochasticity in vegetation patterning and dynamics through time have been reported in other contexts such as old-field succession, post-fire succession in boreal forest, and Mid-Holocene mangrove succession.

Within the framework of complexity theory, Stallins (2006) has attempted to integrate ecological and geomorphic components to conceptualize four unifying themes in biogeomorphology: (i) multiple causality; (ii) the
function of ecosystem engineers; (iii) the expression of ecological topology; and (iv) ecological memory. Multiple causality recognizes the importance of reciprocal interactions between form and process. For example, disturbance by fire, via its impact on vegetation, can influence surface erosion and slope stability (i.e., process affects form). Conversely, topography, through its variations in slope angle and aspect, determines the rate and pathway of fire spread (i.e., form affects process). Stallins emphasized the need to investigate how these form–process interactions play out across time and space, and to what extent this multiple causality becomes self-organizing. Adopting the concept of domains of scale, he demarcated different stability domains in which distinct form–process feedback mechanisms are observed. The regions bridging these domains are characterized as transient states. Overall, the stability domains and transient regions are combined to comprise ecological topology. Within this framework of ecological topology, a shallow lake can flip over time between nutrient-rich and nutrient-poor states. It should be noted that this switch is often mediated by ecosystem engineers that directly or indirectly modify flows of matter and energy, and even disturbance regimes. Depending on the type of ecosystem engineers, a system may dynamically switch between several different stability domains or remain persistent with a certain domain. Finally, ecological memory suggests that disturbance becomes encoded and reinforced in both vegetation and topography across a landscape. In barrier island dune systems, for example, the frequency and spatial extent of overwash events driven by storm surges are dynamically encoded in plant species abundances and dune morphologies. These plants, as ecosystem engineers, actively construct and reinforce their topographic niches in light of the historic disturbance regimes. A specific vegetation–landform complex can potentially perpetuate through this developmentally intertwined feedback.

SEE ALSO: Biogeography; Biogeomorphology; Spatial organization and structure

References


Confidentiality

Matt Duckham  
*RMIT University, Australia*

Confidentiality is concerned with protecting the privacy of information about an individual. Information privacy itself is defined as the “claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others” (Westin 1967, 7). Thus, the topic of confidentiality covers more specifically what types of information have privacy implications, and how that information can be protected from invasions of privacy.

Information privacy

Information privacy is one of several types of privacy. Other types of privacy include bodily privacy (e.g., freedom from physically invasive medical procedures) and territorial privacy (e.g., freedom from intrusion into homes). Privacy in its broadest sense is recognized as a human right in the United Nations Universal Declaration of Human Rights.

A lack of information privacy is associated with a range of negative effects on individuals and society. Specifically, a lack of information privacy can lead to misuse of personal information about an individual in several ways.

- **Unsolicited marketing**: Personal information can be used by unscrupulous companies for unsolicited marketing, such as sending “spam” (inappropriate and indiscriminate) mail and emails to individuals. The inconvenience caused by these activities has direct and negative economic and personal consequences for the individuals, and for society more broadly.

- **Identity crime**: Personal information about an individual can be used by a hostile agent to assume another individual’s identity (termed “identity fraud” or “identity theft”). In 2013, identity crime affected more than 10 million people in the United States of America alone, and it can lead to significant financial and personal distress.

- **Reduction in personal safety**: Detailed information about an individual’s circumstances (e.g., home address, medical history, place of work) can potentially be used by a hostile agent to endanger that individual’s personal safety.

- **Intrusive inferences**: Personal information can be used to make unfair or unwarranted inferences about other areas of an individual’s life. Such inferences can be particularly damaging when made by companies or government institutions. For example, personal medical coverage might be adversely affected by information about a person’s web searches for “cancer” and related terms, or his or her visits to an oncology ward.

- **Inhibiting technological advances**: A perceived or actual lack of privacy can in turn inhibit the development or uptake of potentially beneficial technological advances. For example, if privacy concerns had prevented the acceptance of Internet applications, such as email, social networking, and cloud computing, society would at the same time have forgone the substantial benefits associated with those technologies.
CONFIDENTIALITY

As a result of these negative impacts of breaching information privacy, mechanisms for protecting information privacy, and so ensuring confidentiality, are important in order to avoid or at least mitigate these negative effects.

Privacy protection

In general, there are two distinct approaches to information privacy protection: those that aim to proscribe unfair use of private information, and those that aim to prevent the release of private information in the first place.

There exists around the world a wide range of regulations for dealing with private information. For example, more than 80 countries around the world have comprehensive laws that explicitly recognize information privacy rights (Greenleaf 2012). While privacy regulations vary in their details, in general they all adhere to five fundamental principles of fair information practices (FIPs, also termed “information privacy principles,” IPPs). These five principles can be traced back to a 1974 US Department of Health, Education, and Welfare report (on “Records, Computers, and the Rights of Citizens”).

1 Notice and transparency: Individuals must be made aware of when personal information about them is collected, by whom, and for what purpose.
2 Consent and use limitation: An individual’s consent is required in order to collect personal information about them. Personal information can only be collected for specified purposes, and the subsequent use of that information is limited to those purposes.
3 Access and participation: Individuals have the right to access personal data that refers to them. In case that stored data contains any inaccuracies, individuals may also require that errors be corrected.
4 Integrity and security: Collectors of personal information must make reasonable efforts to ensure data is accurate and up to date. They must also protect against unauthorized access, disclosure, or use.
5 Enforcement and accountability: The collectors of personal information must be accountable for any failures to comply with principles 1–4.

There are, however, two main drawbacks associated with information privacy regulations. First, striking a balance between allowing innovation and precisely defining fair use is difficult or impossible to achieve in all cases. Regulations that are too specific may inadvertently stifle unanticipated and potentially beneficial technological innovations. Regulations that are too unspecific risk exploitation of loopholes, resulting in inadequate privacy protection. Second, irrespective of the balance struck, privacy regulations can only hope to deter information privacy infringements, and penalize those who remain undeterred. However, once a privacy infringement has occurred, it is typically difficult or impossible to undo. Once private information has been publically disclosed, for example, it can never be undisclosed.

As a result, a second approach to information privacy protection aims to protect the personal information itself, preventing its release in the first place. One approach is to use technologies and standards that support privacy policies specifying the acceptable use of information collected or released. For example, P3P (the Platform for Privacy Preferences Project, Cranor 2002) enables websites to declare in a machine-readable format what information is collected by the service provider about visitors to that site, and how that information is used. Like privacy regulations, privacy policies have their drawbacks.
P3P, for example, has been widely criticized for its high levels of complexity and lack of enforcement mechanisms (EPIC 2000), and is today only rarely used in practice.

Other, complementary technologies for preventing the release of personal information center on the use of cryptography to ensure information can be transmitted securely. The PGP encryption system, for example, can be used to ensure only trusted entities are able to decrypt encrypted information. It is important to note that despite the name (PGP, pretty good privacy), information privacy and information security are distinct concepts. Security is concerned with protecting information from unauthorized access; privacy is concerned with the control of personal information. It is possible to secure information that is not in any way private (e.g., encryption of a nonpersonal data file using PGP); and it is possible to preserve privacy without using any security (e.g., reject a website’s P3P policy, and not transmit any information to that service). Although they are distinct concepts, privacy and security are clearly complementary. Confidentiality is best supported by a system that both is secure and respects information privacy principles.

Location privacy

Location privacy is a special type of information privacy with particular relevance to geography and geographic information systems (GIS). Location privacy concerns the right of an individual to control information about his or her location. Location information is potentially highly sensitive. All of the potentially negative effects of a lack of privacy discussed above (unsolicited marketing, identity crime, threats to personal safety, intrusive inferences, and inhibiting technological advances) can also accrue from a lack of location privacy specifically. However, a lack of location privacy is associated with particular risks in three areas. First, location-based spam is unsolicited marketing based on a person’s current location or past patterns of movement. Shoppers, for example, could be bombarded with adverts and special offers for nearby retailers on their mobile phones if they are unable to protect their location information. Second, knowing someone’s current location or past patterns of movement can present particular risks to personal safety, for example through stalking. Third, a person’s location strongly influences the activities they are likely to be engaged in. Intrusive government agencies, organizations, or individuals can potentially use a person’s location to infer information about that person’s state of health (e.g., located at a specific hospital ward or clinic), political views (e.g., located at a political rally or demonstration), sexual orientation, religion, profession, hobbies, interests, and so forth.

In addition to presenting particular privacy risks, location information also presents particular challenges when applying IPPs. The IPP of notice and transparency, for example, is increasingly difficult to maintain because there is an ever-wider range of circumstances in which an individual’s location information may be captured as an incidental, secondary by-product of another function. A cell phone network, for example, must maintain information about the locations of mobile phones (in terms of the cell in which that phone is located) whenever a phone is switched on, in order to route calls to that phone. Making a credit or debit card purchase at a shop or ATM (automatic teller machine) implicitly locates the owner of that card (a feature widely used by banks to combat card fraud). Speed and safety cameras on roads track the movements of personal vehicles, and implicitly locate the registered owner of the
CONFIDENTIALITY

vehicle. Public transport smartcards (travel passes on an integrated circuit card) allow the movements of individuals to be tracked around the transport network.

Even in cases where location information is clearly understood to be the focus of a function, such as when using a personal navigation system or “what’s nearby” location-based service, notice and transparency may be impractical to maintain. Because location is likely to be continually tracked for those services (e.g., in order for a navigation service to provide turn-by-turn routing instructions), it is usually impractical to notify users every time their location is communicated to the service provider (possibly every few seconds or even more frequently). Alternatively, when users are notified about privacy issues at periodic intervals (such as reminders every month while they continue to use a service) they may forget and inadvertently reveal private information at specific, sensitive locations or times. Finally, in addition to common location-aware technologies, such as personal navigation systems, mobile phones, and smartcards, new technologies are continually emerging for identifying individuals at locations. For example, gait recognition of individuals using CCTV camera surveillance is continually improving, with recent advances making its widespread use a realistic possibility in the near future.

Location privacy protection

Given the potential sensitivity of location information, it is perhaps surprising that specific capabilities for location privacy protection have lagged behind the increasing use of location-aware technologies.

One problem facing legal avenues for location privacy protection is ambiguity over whether location information should even be regarded as “personal” information. Relatively few laws explicitly address protection of information about location. Instead, privacy protection usually focuses on “personal information” more broadly, usually defined as information “about an individual whose identity is apparent, or can reasonably be ascertained, from the information” (Australian Government 1988).

For many types of personal information, such as medical or banking records, privacy protection can be achieved relatively simply, by deleting any information about the identity of the person to whom that record refers. The process of removing an individual’s identity from information about that individual is termed “anonymization.” A special case of anonymization replaces an individual’s actual identity (e.g., their name) with a system-generated persistent identity (e.g., a student number), termed “pseudonymization.”

However, for location information, anonymization and pseudonymization alone may not be enough to guarantee a person’s privacy. Location information is special in that identity can frequently be inferred from location, especially from a history of locations over time. A person’s precise location recorded, say, over the course of a few days would often be enough to identify that person. For example, patterns of movements between a home, a workplace, a child’s school, and other locations over the course of a week would easily allow a resourceful hostile agent to identify a person, cross-referencing locations with other publically available data sources, like maps and telephone listings. Several studies have demonstrated the surprisingly small amount of anonymized location information that is needed to infer the identity of the individual tracked at those locations (e.g., Krumm 2007; de Montjoye et al. 2013).

Turning to privacy policies, technologies for specifically protecting location information
have been considered. For example, the IETF (Internet Engineering Task Force, an international open organization of networking experts) GeoPriv working group has developed PIDF-LO (Presence Information Data Format–Location Object). The proposed standard sets out mechanisms by which users can automatically annotate location information with privacy policies that specify acceptable uses of that location information. By allowing users to specify acceptable usage, PIDF-LO represents an example of a user-oriented privacy policy (in contrast to P3P, a service-oriented approach to privacy policies). Users may specify features such as how long location information may be retained, whether that information may be retransmitted, and usage notification requirements. Interestingly, PIDF-LO also provides explicitly spatial mechanisms to protect location privacy by restricting the level of precision of location information. The information that I am located “in Melbourne,” for example, is both less precise and more private, than revealing my (precise) latitude and longitude as $-37.7999253, 144.9610477$. The process of deliberately reducing the precision or accuracy of a person’s location information is termed “obfuscation” (Duckham and Kulik 2005). In order to provide differing levels of privacy, the level of precision may also be adapted depending on the number of other people in an individual’s vicinity. For example, to achieve the same level of privacy, a larger obfuscated region might be desired by a person when hiking on an isolated mountain than when walking around a busy city-center shopping mall.

Census data

While the growth in use of location-aware technologies, like smartphones and personal navigation systems, has brought location-privacy issues into sharp focus, confidentiality and privacy issues in geography are not new. Indeed, geographers have a long history of working with sensitive information with a range of privacy implications in the form of census datasets. Census data contains a wealth of detailed personal information, typically including ages, income, education, health, and religion, as well as home locations of citizens. Census data, then, provides a helpful illustration of the combination of many of the approaches to confidentiality discussed above.

First, when collecting or using census data information privacy principles are scrupulously adhered to, to ensure personal information is protected. For example, strict controls are placed on the capture and storage of census data to ensure information is secure. Those who deliberately or inadvertently disclose an individual’s private data from the census can incur severe penalties. Census data is also frequently anonymized, removing from the data the identities of the individuals to whom specific census records refer. Further, census data is usually only made publically available in an aggregated form, with records from many individuals combined together across defined geographical areas called census tracts. In some cases, random errors with known distributions may also be added to smaller units. Deliberately aggregating data and introducing errors to census data effectively reduces the precision and accuracy of data to a level that makes it much harder for a hostile agent to infer information about specific individuals, much as obfuscation aims to do for an individual’s location information.

The future of confidentiality

Sun Microsystems Chief Executive Officer Scott McNealy famously exclaimed: “You have zero privacy anyway. Get over it!” (Sprenger 1999). In a digital age, with exponentially increasing
volumes of information captured and communicated, attempting to safeguard information privacy can appear a futile endeavor. Indeed, personal information about us is not simply harvested by agencies who value this information, from search engines to state-sponsored surveillance; many individuals freely disseminate their own personal information through a range of channels, from retail store cards to social networks.

In this context, it is hard for anyone to predict what forms of confidentiality and levels of privacy protection will be common in the future. If the apparent trajectory of increasing disclosure continues, it is conceivable that societies may ultimately arrive at a new, stable equilibrium, such as the “transparent society” envisioned by Brin (1999), where everyone knows what everyone else is doing. It seems, however, equally plausible that changing attitudes to confidentiality apparent in digital societies may be a temporary phenomenon, yet to be checked as the negative impacts of a lack of privacy, discussed above, become more common and more widely understood.

SEE ALSO: Privacy, personal privacy

References


Further reading

Conservation and capitalism

Dan Brockington
University of Sheffield, UK

Capitalism has transformed the material world: think of all the mines sunk, land cleared, fish harvested, or air polluted in the name of making money. Capitalism also dominates our symbolic lives: think of the television, radio, and billboard adverts that daily sell various commodities to us, like mobile phones and automobiles. Conservation is normally seen as the antithesis of, and sometimes as an antidote to, all of this. Conservation is manifest in the form of parks and protected areas, maps of endangered species habitats and rare ecosystems, and exhortations to rethink our engagement with the environmentally destructive system that is capitalism.

Historically, this basic incompatibility has intensified as both spheres of influence have expanded in geographical terms. In its starkest form, and using now outdated language, “nature” is commonly seen as retreating before human despoliation. For many conservationists it is traditional to define the solution as the opposite to development, industry, and the use of natural resources. Some of the defining historical moments of the movement have been integrally linked to the protecting of landscapes from human use and depredation. Examples include the early establishment of protected areas in Europe and Africa (and the continued fight to keep them pristine, as manifest in Africa’s Serengeti recently), the failure to protect Hetchy in the United States, the success in protecting Australia’s Jervis Bay, and the fights over the Franklin River in Tasmania.

However, despite these iconic moments in conservation’s history, and ongoing disputes that bring human interests into conflict with particular forms of nature, the relationship between conservation and capitalism is changing. Many varieties of conservation and corporate capitalism now work closely together. This is not merely a case of opposing interests that have become adept at satisfying each other’s demands. It is more intimate than that. Nowadays it is hard to distinguish between the ways in which conservation reshapes the world using capitalism as its instrument from capitalism doing the same through conservation. The presence of this entry in this encyclopedia is testimony to the importance of this trend. Its purpose is to explore the ways in which conservation and capitalism are intertwining, and the consequences of the fact both practically and for research purposes.

The entry will first examine certain writings about conservation and neoliberalism, in which much current research on the interactions between conservation and capitalism is grounded. It will then explore manifestations of this relationship in important conservation practices, first with respect to the work of conservation NGOs, and second with respect to the marketization of conservation policies, especially the proliferation of new commodities associated with ecosystem services. Finally the entry explores the sorts of politics that are associated with capitalist conservation, and particularly the problems associated with postdemocracy and environmental change.

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg1084
CONSERVATION AND CAPITALISM

Neoliberal conservation

Writings on neoliberal conservation have an interesting history. They stem from encounters with geographer Noel Castree’s seminal reviews of writings on neoliberal nature and a group of researchers who were exploring the impacts of market-based conservation policy (Castree 2007a; 2007b). Castree’s writings provided the framework to explain what was happening within conservation policy, to understand how these developments were interconnected, and to communicate them to a broader academic and activist community. He argued that neoliberalism is one particular form that capitalism has taken in recent decades. It is characterized by the following processes.

- **Privatization**: the assignment of clear private property rights to social or environmental phenomena that were previously state-owned, unowned, or communally owned. New owners of hitherto unprivatized phenomena can potentially come from anywhere across the globe.

- **Marketization**: the assignment of prices to phenomena that were previously shielded from market exchange or for various reasons unpriced. These prices are set by markets that are potentially global in scale, which is why neoliberalism is often equated with geographically unbounded free trade.

- **Deregulation**: the rollback of state interference in numerous areas of social and environmental life so that (i) state regulation is light touch and (ii) more and more actors become self-governing within centrally prescribed frameworks and rules.

- **Reregulation**: the deployment of state policies to facilitate privatization and marketization of ever-wider spheres of social and environmental life.

- **Market proxies in the residual public sector**: the state-led attempt to run remaining public services along private sector lines as efficient and competitive businesses.

- **The construction of flanking mechanisms in civil society**: the state-led encouragement of civil society groups (charities, NGOs, communities, etc.) to provide services that interventionist states did, or could potentially, provide for citizens. These civil society groups are also seen as being able to offer compensatory mechanisms that can tackle any problems citizens suffer as a result of the previous five things listed.

Castree argued that neoliberalism is consistent with attempts both to exploit and protect the physical environment, noting the growing use of neoliberal measures in the field of environmental policy worldwide. This is what is meant by the phrase “neoliberalization of nature.” For instance, individual transferable quotas issued to fisherman have become a favored way of protecting fish stocks, a form of state-led reregulation of human behavior that has privatization and marketization of environmental assets at its core.

There followed a series of landmark publications and international meetings which have explored different aspects of what neoliberal conservation on land and at sea entails (Igoe and Brockington 2007; Brockington, Duffy, and Igoe 2008; Brockington and Duffy 2011; Büscher et al. 2012). A fascinating commentary and appraisal of the epistemic community formed thus has been, appropriately enough, recently published by Castree and Henderson (2014).

It is notable how interdisciplinary this work is. It brings together geographers, environmental anthropologists, development studies, media studies, and politics and international relations. Yet, despite this disciplinary diversity, through these activities of meeting, writing, and sharing research ideas, methods, and successes, an
Conservation and Capitalism

An epistemic community has been forged which has come to tackle a series of similar problems in method, conceptual frameworks, epistemologies, and by contests over the same. Critical analysis of neoliberal conservation, and explorations of the interactions between conservation and capitalism, are a recognized subdiscipline.

What are the main tenets of writings on neoliberal conservation? At their core there is a suspicion of the role of markets in conservation policy. Markets come in many forms, but in this case the reference is to ones where commodities are exchanged for money between buyers and sellers. Opponents of neoliberalism observe that giving over things that people value to market control can result in the commodification of ecological and social relationships. Indeed, it can require impossible feats of quantification as attempts are made to reduce these relationships somehow to commensurable (that is, comparable or equivalent) forms. More than that, neoliberalism is associated with greater levels of inequality, with the withdrawal of social and state support for the disadvantaged and with forms of political decision-making that favor the rich and powerful. As Warren Buffet, one of the world’s richest people, said repeatedly, there is a class war, and it is his class, the rich, that is waging it, and they are winning.

Opposing neoliberal economics and politics is difficult. It has yet to result in a particularly satisfying alternative. In the academy, opposition entails documenting and exploring what these policies actually entail for nature and for society as they are implemented. Researchers explore the tensions and contradictions that develop around them, and the ways in which market-led governance can change levels of inequality, political forms, scientific discourse, and so on.

Igoe and Brockington’s 2007 paper in Conservation and Society provides a useful starting point for this literature. These authors observed that there had been a growing literature concerned with the neoliberalization of nature, and flourishing in particular among radical geographers. They also observed that these critiques did not mention changes in conservation policy. Yet, they argued, the key processes that were part of the neoliberalization of nature could also be observed within new conservation initiatives, particularly in Africa where their own empirical research has been situated.

These processes included all those summarized by Castree: deregulation, reregulation, privatization, marketization, and so on. These are visible in the new forms of conservation estate that is set aside outside protected areas to make space for new corporate ventures, including hunting and photography safaris in alliance with the rural owners of the lands. They include a proliferation of conservation NGOs working with companies and local groups. They include new markets for new forms of tourist activity such as green hunting (in which trophy animals are merely darted, not shot dead; the more daring versions are reputed to include elephant paintball), as well as markets for carbon in village forest reserves and other ecosystem services. But this deregulation, privatization, and marketization, all of which hand domains traditionally controlled by the state to other actors, is also accompanied by reregulation and new areas of state activity. Traditional protected areas have continued to expand, and indeed expand most during neoliberal times. And new forms of conservation estate remain governed and overseen by state actors, even where they are oriented toward catering to free-spending international tourists.

These developments are also accompanied by further changes which allow alliances of capital and conservation to become more profitable. There are new forms of territorialization which, following Vandergeest and Peluso (1995), refer
to “the demarcation of territories within states for the purposes of controlling people and resources.” Protected areas are one such form of territorialization. Another development is the rise of new attempts to manage and use resources more effectively. This is done by altering the ways rules control who can use resources, when, and how, or, in other words, how resource use is governed. Under neoliberal policies, governance is meant to be achieved by the market, meaning that who gets to use resources, and the ways in which they are to be used, are determined by who is willing to pay for their use. But to usher in and to make markets work, new networks of actors are required that include nongovernmental organizations (NGOs), companies, philanthropists, in various alliances with states. So powerful are these alliances that they can be described as a form of privatized, if sometimes spatially restricted, sovereignty.

In 2012 Büscher et al. (2012) wrote an assessment and critique of neoliberal conservation. Their contribution, they argued, was partly necessary because critiques of the neoliberalization of nature were still not recognizing the importance of conservation. The emphasis in writings about neoliberal nature was all about how nature and natural resources were used in the transformation of neoliberalization. The demands of capitalism were the main drivers of change, and conservation was somehow regarded as separate from all of that. Indeed, in Harvey’s work they found a tendency to treat environmentalism as a bastion against capitalism (Harvey 2005). Protest against capitalism, which Harvey lauded, was often environmental protest, and the complicity of NGOs in the neoliberal project that he critiqued did not include the conservation NGOs. Yet Büscher and colleagues (2012) had repeatedly shown that conservation NGOs have been at the forefront of bringing in new forms of neoliberal conservation, sometimes to the detriment of local groups.

Büscher and colleagues argued that this thinking missed the point that capitalism can expand even if it is not using and exploiting resources. Capitalism can expand through finding profitable ways to conserve resources. Thus they argued that “neoliberal conservation shifts the focus from how nature is used in and through the expansion of capitalism, to how nature is conserved in and through the expansion of capitalism” (Büscher et al. 2012, 5).

They suggested that there were three main arenas of activity within neoliberal conservation that required attention from academics and researchers. They first noted a basic contradiction in the thinking that promoted much neoliberal conservation. Current ecological crises are driven by economic growth and the hunger of market-based capitalism for profit, yet neoliberal conservation policies see markets as the solution to these problems. “Economic growth,” Büscher and his colleagues argued, “becomes, paradoxically, the prerequisite for healthy ecosystems, whose conservation in turn becomes the basis for further economic growth” (2012, 14).

And yet, Büscher and his colleagues observed that the potential contradictions and pitfalls of this policy are concealed by the ideology and the win–win rhetoric that inevitably accompany it. The objections and difficulties that arise from this project become reduced to technical problems. When considering problems associated with community forestry, for example, the problems include how one measures growth in carbon and how one pays for it. More fundamental questions of who decides which bits of the forest should be conserved, what forms of use and users are excluded, and whether poorer people (pastoralists and charcoal burners) suffer as a result are simply not part of the discussion. Another way of saying this is that what are actually intensely
political projects involving contests over how resources should be used, and by whom, are portrayed as apolitical issues, with the conflict and tensions driven from them.

Büscher and his colleagues (2012) go on to observe that neoliberal conservation achieves this smooth and happy facade through the clever use of image and spectacle (see Environment and celebrity). This is the second arena of activity they mention in their article. Spectacle in this instance refers to images that mediate relationships between people. Examples include dramatic images of tsunamis or earthquakes, tragic pictures of starving children or drowned toddlers, or heartwarming images of animals seeking adoption. These images mediate relationships between people because of the way that viewers act on them, and because of the attitudes to the places and issues portrayed that they cultivate. What we do and how we think as a result of these images affects the way we relate to other people.

Spectacle is integral to the establishment of an appearance of harmony and productive peace between the previously opposing spheres of the economy and the environment. Spectacle can be staged in public and experienced in person. It can occur in major showpiece events and global conservation meetings, when the new alliances between conservation and capital are celebrated. It encourages audiences to welcome and accept the new developments. More often spectacles are not experienced directly but are mediated, through websites, television broadcasts, and Internet videos which help forge the idea that, for example, carbon produced in the North can be interchangeable with carbon sequestered in forests in the South.

It is through such spectacular work that nature can come to generate value and profit for capital as conservation organizations raise finance and support for their ventures in the world’s metropoles. Elizabeth Garland’s (2008) work shows how images of wild nature are produced and cultivated by conservation NGOs and then turned into money by presenting them to audiences in fundraising events and campaigns. They can also be turned into social and political capital, changing the mindsets and decisions of key politicians and corporate leaders (Garland 2008).

The final arena of activity Buscher et al. (2012) observed was the disciplining of dissent. Mainstream neoliberal conservation is characterized by a broad (if somewhat superficial) consensus. It has been the unfortunate experience of many of those opposing its core tenets that objections are not often countenanced. Anyone speaking out against the smooth and comfortable visions of neoliberal conservation can expect a sharp reaction. Maintaining the happy vision is essential for the financial models and political influence of important conservation organizations. They take disagreement seriously.

At the same time as the criticisms of neoliberal environmentalism have consolidated, the criticisms themselves have come under closer scrutiny. On the one hand, there is concern that the word “neoliberal” is used too liberally. It is often too loosely defined, or it refers only to a general set of activities. A second pertains to the unbalanced agenda of neoliberalism’s critics. Their outcries against power have yet to produce a viable and well-defined alternative to neoliberal conservation. They focus too much on critique and not enough on what could feasibly replace neoliberal policies. For instance, the practical implications of their objections should be considered. For example, should an objection to the alliance of extractive industries with conservation be extended to an objection to the presence of metal in society more generally? Or perhaps we should object to the pulping mills and learn to live without paper?
Conservation and Capitalism

The challenge becomes not to resist any alliance with, for example, the extractive industries, but rather to forge alliances that are inclusive and egalitarian, even while maintaining a principled stance against unsustainable production practices, which are all too often “greenwashed,” that is, represented as being environmentally friendly by the corporations that engage in them.

The Marketization of Conservation Policies

At the core of neoliberal thinking on environmental and conservation issues is the idea that inefficiencies and failures in the allocation of resources can be solved by more efficient and effective markets. Or, in other words, if conservation problems cannot raise the resources to bring about the required solutions, then markets need to be created to allocate those resources according to conservation priorities. The key problem, argue the protagonists, is that many conservation values do not have a financial value that can be exchanged or realized through markets. It is impossible for consumers or citizens to express their choices financially, and therefore difficult for conservation programs to raise the revenues they require. Conservation’s problem is that it sells itself cheaply; nature’s problem is that it gives away its bounties free. If they could be costed and paid for, there would be less degradation.

For example, forest cover on hilltops is thought to provide enduring, regular water supplies to rivers draining hill catchments, and so to urban and industrial users downstream. The cost of that service however, is not recognized or paid for. If it were to be included in water bills, this would create and generate funds for conservation policies. More generally, this issue is part of the debate about payment for ecosystem (or environmental) services (PES), which is proposed for a broad range of benefits including carbon sequestration, pollination, amenity values (see Ecosystem services). Sven Wunder’s (2007, 50) widely cited definition insists that PES requires:

1. a voluntary transaction in which
2. a well defined environmental service (or a land use likely to secure that service) (3) is “bought” by a (minimum of one) buyer (4) from a (minimum of one) provider (5) if and only if the provider continuously secures the provision of the service (conditionality).

The idea that conservationists should realize the value of the goods and services that they are conserving has become extremely popular. The mantra is that nature has to be worth something in order to be valued by decision-makers, and that it would be much easier to convince politicians of the value of conservation if we could show how much conservation is worth. This logic is driving the work of organizations like the Economics of Ecosystems and Biodiversity (TEEB), which is collating and synthesizing numerous reviews to determine what value to place on different ecosystem services.

The move is interesting in a number of ways. First, it has been met with enthusiasm by the conservation community, particularly the research community. This is clearly visible in official conservation conferences where selling ecosystem services is widely promoted. Second, while providing valuations of goods for which no market yet exists is an uncertain science, this does not dampen the enthusiasm of economists, who have many ways of tackling the problem. The results of these analyses, however, can vary unpredictably.

Beyond the theoretical problems of valuation there is also a series of difficulties in implementing related policies of resource use, which derive from the difficulty of creating the new commodities of ecosystem services that could be traded on markets and of establishing the new markets themselves. It is not easy to delineate or measure...
carbon (for example), or to fix it permanently in trees in ways that can be paid for in advance. There is a lack of case studies, but those that are available underline the fact that the creation of markets requires more regulation and state governance, notwithstanding their stated ideologies. This does not mean that PES are all problematic and likely to fail, just that it requires particular conditions in order for them to work – such as clearly defined property rights or effective commodification of the service in question. Another crucial element is a culture of recognizing the legitimacy of payments. If people are not used to paying for forest cover to protect water supplies, or even paying for water, it will be difficult to persuade them to start doing so.

PES are more of a prospect and mood than an active and effective policy; the idea is more vaunted than it has been put into practice. Nonetheless, the philosophy and implications of them are already alarming critics. Nicolás Kosoy and Esteve Corbera (2010) argue that the commodities on which PES hinge are characterized by three simplifications which conceal important aspects of the services the PES schemes are trying to promote. First, ecologically, the commodities cannot capture all that is valuable in nature, only that which the markets try to value. Carbon forestry, for example, is not necessarily the same thing as preserving a biodiverse or watershed-preserving forest. One could promote a forest that maximizes carbon storage, but it may be full of fast-growing exotic species that are bad hosts of other biodiversity.

Second, reducing nature to commodities will simplify the social relations that pertain to natural resources, ecosystems, and environments. What one group might value, for example wildlife for hunting, another might detest as dangerous pests (in this example). Setting up a market for hunting wildlife requires healthy populations of those animals. Once again political disputes over how landscapes ought to be used become reduced to technical problems of how to set up markets that govern already determined uses. Moreover, markets cannot easily capture spiritual or aesthetic values.

Finally markets conceal the power inequalities that structure and shape them. Allowing those with money to purchase ecosystem services depends on a just distribution of money in the first place. And, for all the talk about efficient and free markets, in practice markets are structured by political pressures in a myriad ways even as their rhetoric denies that influence.

Redford and Adams (2009) list further issues which, in part, derive from the nature of conservation as a movement. Specifically, they argue that conservation is prone to fads, that is, wholeheartedly chasing the latest idea, sometimes before the evidence exists to show that it is robust. Furthermore Redford and Adams note that ecosystem services can include natural processes which are harmful to people (flood, fire); harmful to nature (when provided by alien species); and prone to being radically altered by climate change, which is reconstituting ecosystems under our feet. They welcome the possibility that the scheme offers, but call for more skepticism in testing and developing it. The difficulty, then, is not with PES as an idea or theory, but rather with the marketization of nature as a movement.

Conservation NGOs

One of the hallmarks of neoliberalism is its distrust of government. One of the characteristics of the rolling out of neoliberal policies is the growth in NGOs that have arisen to address the demand for services left by declining states. Conservation is part and parcel of this trend,
and for this reason conservation NGOs deserve special consideration.

The influence of conservation NGOs predates neoliberalism and derives from their concentration of research and policymaking skills. NGOs and NGO-based researchers have been at the forefront of numerous debates in conservation, whether these be concerned with identifying conservation priority areas, maximizing return on conservation investment, or the discussion of new neoliberal policies. Conservation NGOs do not speak with one voice here – far from it. Many cannot speak with one voice in themselves. They are large, diverse, and eclectic organizations. But the point is that they are a vital source of ideas and experimenters in new policies, particularly in poorer parts of the world where government policies are weaker because they have fewer resources and less capacity.

Yet, despite their diversity, there are some common trends among NGOs, particularly the more powerful organizations. Conservation NGOs have been the subject of several stinging critiques which have accused them of becoming too close to government and corporate power and of not doing enough to promote conservation interests where this risked upsetting the powerful. Mac Chapin’s (2004) critique is perhaps the most severe. It was published in 2004 in the popular magazine *Worldwatch*, complete with choice cartoons that compared conservation NGOs to imperial conquistadors. Chapin alleged that large conservation NGOs, in their pursuit of corporate funding and in trying to negotiate with powerful governments, had become identical to the capitalist and industrial interests they were meant to be opposing. This led to tragic and farcical situations where local environmental protestors and indigenous groups might find themselves opposing alliances of mainstream conservation NGOs, corporate power, and governments. Surely, Chapin, insisted, any conservation NGO worth the name would be found on the other side of the barriers.

Chapin’s polemic highlighted some stark contradictions within the conservation movement, but it is probably fair to say that the most egregious errors are not typical. It is rare to fund conservation organizations allying with mining companies against the interests of indigenous forest dwellers. A far more common alliance is NGOs embroiled in new policies that neoliberal conservation entail, for they often serve to help implement new forms of commodification and market governance, such as carbon credits, or new safari packages on community-owned lands. They are, in some ways, at the vanguard of neoliberal conservation, serving as entrepreneurs and restless innovators. In this respect they behave very like the capitalists described by Karl Marx who are endlessly and energetically seeking new opportunities to create profit.

In some respects, it must also be noted that the changes NGOs bring are welcome – the money provided by new commodities can be a valuable source of income to poor people’s livelihoods, as can the changes in governance. However, in other respects NGOs can be part of a deeper problem. Jesse Ribot’s studies of governance and devolution in the forestry sector have found that many potentially empowering incidences of devolution have been stymied and incomplete. They could have brought welcome change to impoverished places, but they have not done so because power has been devolved to institutions and entities, including NGOs, which are not downwardly accountable (Ribot 2004). NGOs are not elected by local groups. They cannot be dismissed. They are funded by outside agencies, donors, or rich foreigners and nationals. This means that, while they should exist to serve their local demographics, they may fail to do so and may be more closely oriented to serving their
funders. In this sense NGOs obstruct participatory local governance of natural resources. Their skills and competence may make them good at working with local people and with natural resource management. But their mandate is self-derived and their authority ultimately arbitrary.

At the same time, from the point of view of researchers, working with conservation NGOs can be rich and rewarding. For these are intelligent, reflective communities, passionate about their calling and thoroughly engaged in the practices researchers are trying to understand. (The same is true of the conservation filmmaking community.) The communities are also immensely diverse, for the nature and purpose of conservation organizations differs, and within larger organizations there will be many different views at work. Thus, while it is important to understand the nature and role of conservation NGOs, it is also important to realize that much understanding about them is coproduced with the NGOs themselves. There are interesting revolving doors between communities of natural scientists, who work on conservation issues, and conservation NGOs. Careers can switch between the two. These reciprocal relations are yet to develop properly between social scientists who study conservation and conservation NGOs.

The politics of capitalist conservation

One of the most interesting things about the close relationship being forged between conservation and contemporary capitalism is the politics of the decision-making and the sort of democracies in which this relationship has come to thrive. It is useful here to turn to writings on postdemocracy, particularly as described by Colin Crouch (2004). Postdemocracy, for Crouch, describes situations where the ruling elites have learned how to manipulate and control the voting public, and in which the voting public appears not to mind the development that much. Crouch notes that enthusiasm and participation in democracy tend to be uneven, with particular historical moments when it is strong followed by less activity. When voter turnout and participation in political decision-making decline, government becomes dominated by elite-level lobbies which are much less accountable to publics.

Crouch’s objection is to the form that liberal democracies take, particularly in unequal societies. Liberal democracies are founded on the belief that any form of political expression (within some limits) should be allowed and that everyone should be free to politically support the interests that matter to them. This becomes an issue in unequal societies when liberal democracies inevitably become dominated by the interests of the most powerful, who have the most money and resources to devote to supporting their interests. In postdemocratic situations, without a vigilant public to restrict the actions of politicians, liberal democracies will not be healthy democracies. The key issue for Crouch in this state of affairs is that it produces politics where decision-making can be dominated by elites, and that must inevitably tend toward unequal societies. Such elites tend not to be egalitarian or to benefit from egalitarian policies, yet they are able to prosecute their interests far more effectively than other groups in the politics of elite-dominated lobbies. The public interest can become subsumed beneath the pressures from well-funded, and often corporate-funded, interest groups.

It should be clear why this state of affairs matters for analyses of conservation and capitalism. The decision-making and politics that determine conservation policy directly, and issues of resource use and exploitation (and so,
indirectly, conservation issues) will be determined in elite-level forums. Conservationists who wish to prosecute their policies and interests successfully have to become adept at walking in these powerful corridors and lobbies. Viewed thus, the celebrity-filled, industry-sponsored events of the International Conservation Caucus in the United States are simply the necessary response of the conservation movement to the political dictates of the age (Corson 2010). So too are the business councils and corporate sponsorship programs that large organizations have pursued. As Crouch (2004) observes, organizations hoping to make a difference in postdemocratic politics have no alternative but to engage in elite-lobbying.

Nonetheless the common accusation is that the process of becoming so close to power has resulted in conservation organizations losing their critical edge (as can be seen in Chapin’s (2004) critique). Corporate capitalism is not afraid of conservation NGOs any more. Conservation organizations pose little threat to corporate brands or public opinion. More than that, there are times when corporate interests and conservation activists seem so closely aligned as to be indistinguishable. CEOs of major conservation groups sit on the same boards of major industries; they work closely together to build brand and income. A postdemocratic politics appears, of necessity, to be about consensus, and without the oppositional agonism that will make contentious decision-making more transparent.

It is in the realm of image and spectacle that so much postpolitical and postdemocratic conservation debates become visible. The lack of agonistic politics, the melding of elite leaders and rulers, and the coincidence of corporate and conservation agendas all become particularly clearly visible, and are actively enforced and maintained in the advertising campaigns, media events, and staged interactions that characterize protest and environmental debate in media-saturated societies. One of the reasons why Büscher and his colleagues (2012) identified spectacle as being so important for neoliberal conservation is because of its role in postdemocratic politics.

Summary

The study of conservation and capitalism provides important insights into the way conservation and environmental issues function in capitalist societies, and how each is being incorporated into the other. They help us to understand the sorts of relations between society and nature that are being constructed, and the hybrid socio-natures that result. This is illuminating both for what it tells us about contemporary societies and political economy and for what it tells us about nature. This is an exciting time to be involved in the social science of conservation, and in particular the relationships between conservation and capitalism. The field has expanded enormously, and there is now a strong theoretical framework guiding current inquiry and tying it to more general concerns about neoliberal environmental policies and critical geographical inquiry in the field.

SEE ALSO: Ecosystem services; Environment and celebrity; Environment and development; Environmental policy; Environmental valuation; Tourism

References


Constructing 3-D GIS

Keiji Yano
Ritsumeikan University, Japan

We live in a three-dimensional (3-D) world and throughout human history the 3-D world has been represented through pictures and photographs presented in bird’s-eye view. Until now, geographers and GIS scientists have basically sought to graphically express their understanding of 3-D spatial aspects in reality through two-dimensional (2-D) maps. With a 2-D map it is possible to turn the surface of the round Earth into a planar graphic form via map projection and using contour lines to represent the various elevations. This allows us to imagine the 3-D world and create 3-D models. Using 3-D GIS, it is possible to view the entire 3-D space from various viewpoints.

Many types of 3-D GIS data structures have been developed and many related software applications have been created. In addition, methods for data acquisition in real life designed for these data structures have also been developed. This entry provides an overview of 3-D GIS data structure and 3-D GIS data acquisition, while introducing the latest technologies and several 3-D city models.

3-D GIS data structure

There are two kinds of 3-D spatial object representations: surface-based and volume-based. A surface-based representation is where an object is represented by surface primitives, whereas a volume-based one is where the interior of the object is described using solid information (Abdul-Rahman and Pilouk 2008). Surface-based representations include grid models, shape models, facet models, and boundary representations (b-rep); volume-based representations include 3-D arrays, octree, constructive solid geometry (CSG), and 3-D TIN (or TEN). Some of these representations are common in computer-aided design (CAD) systems but not in GIS.

Here, grid models and facet models (TIN) are outlined as surface-based models, and 3-D arrays, octrees, and 3-D TINs are outlined as volume-based models defined according to Abdul-Rahman and Pilouk (2008).

Grid model

A grid is a method for surface representation in GIS and digital terrain modeling (DTM). It is a structure that specifies the height value at regular locations and it is simple to generate. The topology of grid points can easily be determined, since each grid point is relative to other points. The grid structure is considered as an array structure in computer programming. Each array element represents the x and y locations of the grid, and the value of the element is z, which indicates the height. The positions of the grid points are easily defined, and their layout may be regular or irregular (Figure 1). The terrain surface model can be derived from this structure. However, this structure has disadvantages in the case where surfaces have multi heights such as vertical walls and overhangs.
CONSTRUCTING 3-D GIS

Figure 1  Grid model.

Figure 2  Facet model (TIN).

Facet model (TIN)

A facet model describes an object’s surface using planar surface cells that can be of different shapes and sizes. One of the most popular facet models using triangle facets is known as a triangle irregular network (TIN). A TIN is a vector-based representation of the physical land surface or sea bottom made up of irregularly distributed nodes and lines with 3-D coordinates (x, y, and z) that are arranged in a network of nonoverlapping triangles (Figure 2). TINs are often derived from the elevation data of a rasterized digital elevation model (DEM). An advantage of using a TIN over a raster DEM in mapping and analysis is that the points of a TIN are distributed variably based on an algorithm that determines which points are most necessary for an accurate representation of the terrain. Data input in this case is, therefore, flexible and requires fewer points to be stored than in a raster DEM with regularly distributed points. However, a TIN may be less suited than a raster DEM for certain kinds of GIS applications, such as analysis of a surface’s slope and aspect.

3-D array and octree

3-D array is the simplest structure in the 3-D domain, being easy to understand and implement. In the 3-D array (Figure 3), the size of the array elements is equal and each occupies the same amount of computer space. However, as a result, a 3-D array model requires large computing power to render. A much better way of representing 3-D objects is by varying the size of the voxel, known as the octree technique. Octree is one of the hierarchical data structures.
CONSTRUCTING 3-D GIS

3-D GIS data acquisition

3-D GIS data acquisition and reconstruction is the generation of 3-D or spatiotemporal models from sensor data. The techniques and theories, generally speaking, work with most or all sensor types, including optical, acoustic, laser scanning, radar, thermal, and seismic sensors. Acquisition can occur from a multitude of methods, including 2-D images and acquired sensor data.

Acquisition from 2-D images

Photogrammetry or stereo photogrammetry is the technology used for creating 3-D data \((x, y, z)\) of object features using stereo image pairs. Stereo photogrammetry is broadly divided into two types: aerial photogrammetry from an airplane and close-range photogrammetry on land. In the former case, serial images are taken using one camera with the same parts overlaid. In the latter case, the image of an object is taken by moving one or two cameras, so that parts of the captured images overlap with each other. Both methods can be used to acquire topologically structured 3-D data from 2-D aerial stereo images.

In both Figure 5a and Figure 5b, point P within the ground coordinate is recorded on the projection planes as p1 and p2 using cameras with projection centers O1 and O2. However, there is no difference in data processing between aerial photogrammetry and close-range photogrammetry without the distance between the camera and the object, and the parallel photographing with high accuracy. In the case of aerial photogrammetry, parallel photographing is easy, but in the case of close-range photogrammetry it is rather difficult to achieve (Susaki and Hatayama 2013).
Acquisition from acquired sensor data
LiDAR

Currently, laser scanners are fast and powerful enough to generate airborne-scans of large areas. Laser measurement using LiDAR (light detection and ranging, laser imaging detection and ranging) is one of the optical remote sensing methods where laser is used to obtain high-density sampling points to generate highly accurate 3-D data. Using GIS software, ArcGIS, it is possible to manage, handle, visualize, and share a cluster of many points in a 3-D space. LiDAR is used for airborne laser mapping instead of conventional photogrammetric methods for its cost–benefit performance (Figure 6). The main hardware of the LiDAR system consists of on-board equipment (attached to an airplane, helicopter, or car), tripod stand, laser scanner system, GPS (global positioning system), and INS (inertial navigation system). The INS system is used to determine the role and gauge the pitch and direction of the LiDAR system.

It is also possible to obtain semi-automatic building extractions from LiDAR data and high-resolution images. This approach allows modeling without the need to physically move towards the location or object in question. From airborne LiDAR data, a digital surface
model (DSM) can be generated and objects higher than the ground are then automatically detected using the DSM. Based on general data on buildings, information pertaining to geometric characteristics – such as size, height, and shape – is then used to separate the buildings from other objects. For instance, using ArcGIS in LiDAR Analyst it is possible to automate the extraction of 3-D terrain data, and other information such as buildings, trees, and forest areas from the LiDAR data obtained through aerial laser survey.

Around the year 2000 there were some semi-automatic methods available for creating 3-D GIS data from LiDAR data, such as Virtual London by UCL (Evans, Hudson-Smith, and Batty 2005) and Virtual Kyoto using MapCube by Tokyo CadCenter Corporation (Yano, Nakaya, Isoda 2007).

Virtual London has been utilized as a platform for communication, and Virtual Kyoto as 4-D GIS with a time dimension has provided a virtual time-space into which the digital contents in the historical city of Kyoto can be put.

MapCube, consists of prismatic 3-D building block models based on building footprints and heights of an entire city. The 3-D models are created by extruding building footprint data with airborne laser-profiler data that obtains height values at the interval of 2.5 meters within an error margin of ±15 centimeters (Figure 7).

In 2005 the extremely popular Google Earth was launched. It is a free web application that provides 3-D city models by adding aerial images to 3-D terrain models and shows land features and landmarks including 3-D buildings. The 3-D buildings are modeled using free SketchUp. These types of 3-D building models have, until now, been created using expensive software applications such as 3-DMAX and Maya, and the compatibility of the data format of these 3-D models is a problem. In general, however, common formats including VRML are used to maintain the compatibility of 3-D building model data.

Recent cutting-edge technologies

From around the year 2000, aerial 3-D mapping using LiDAR has become common and the accuracy has been greatly improved. In addition to that, many people have also opted to build 3-D models by integrating LiDAR data with oblique photographs. The following are two of the latest examples of how 3-D data are actually collected.

**High-resolution LiDAR data** Conventionally, it was difficult to create accurate DEMs of forest areas because trees prevent the laser from reaching the land surface. Today, LiDAR data collected from a helicopter are 3-D data containing over 20 points per square meter. Through high-resolution LiDAR data, it is not only possible to create detailed DEMs and contours but also 3-D models with even the height of trees represented. As a result, it is now also possible to estimate the volume of trees in forests.

For instance, based on the 20-point/m² LiDAR data obtained for forests in Kyoto prefecture, a 25-cm DEM was made, from which maps showing detailed land surface shape and tree heights can be created (Figure 8). The performance of LiDAR equipment is likely to further improve in the future and we can expect to obtain more detailed LiDAR data. Meanwhile, development of methods to recognize ground surface, buildings, trees, and other objects, as well as to clean out noise from images is an urgent matter. It is also important to develop software for rapid display of detailed groups of points and for converting such data into vectors.
Extremely realistic 3-D city models using oblique photography 3-D city models up to now have been 2.5-dimensional models created by measuring the height of buildings. However, detailed full-3-D models have become possible through using aerial oblique photography. Specifically, this method involves taking pictures at direct and slanted angles at the same time to obtain an enormous amount of overlapping photography (multilayer) images using an oblique

Figure 7 MapCube. Source: CadCenter Corporation.
camera to construct extremely realistic 3-D city models. This technology can be used to collect 3-D data not only from an airplane but also from a low-flying helicopter or using a vehicle MMS (mobile mapping system) from land (Figure 9).

Figure 10 shows the comparison between a conventional 2.5-D model and a detailed full 3-D model of the Yokohama seaside district. In the case of the full-3-D model, oblique photography is applied to construct a super-realistic 3-D city model without the need for texture mapping of photographic images into shape models. In particular, this makes it possible to show the detailed shape of the sides of buildings and to apply texture mapping.

Future of virtual city models

The needs for 3-D city models are growing and expanding rapidly in various fields, including geography, urban planning, design, architecture, environmental visualization, and many more. The development of methods for creating 3-D city models is an important topic in academic and commercial environments (Lin and Batty 2009). Presently, there are various terms used to denote 3-D city models, such as “virtual city”, “cybertown”, “cybercity”, and “digital city” (Dodge et al. 1998) and there are a large number of 3-D city models across the world, ranging from CAD models through various 3-D GIS to VRML web content.
Figure 9  New technologies of 3-D data acquisition using oblique photography. Source: Pasco Corporation.
In the case of Google Earth, for example, virtual city models can be easily explored using walkthrough and flythrough functions. On the other hand, ESRI CityEngine allows users in GIS and CAD to easily generate 3-D cities from existing 2-D GIS data. However, while ESRI CityEngine can promote virtual city models under the GIS environment, creating 3-D city buildings is still costly and time consuming. In response to this issue, Google has provided the toolkit, SketchUp and BuildingMakers, necessary for creating 3-D building models via the web. By encouraging the public to freely create 3-D buildings, based on volunteered GIS, Google has set off an apparently very successful trend in the creation of 3-D building models.

3-D GIS data acquisition technologies are expected to become further developed in the future, and detailed virtual city models may be built based on the improved technologies. Virtual city models may become easy to display and use via the web or a portable device. With the anticipation of further development in 3-D GIS data structure and data acquisition, it is necessary to consider how the technologies can be utilized in the future.

**SEE ALSO:** Geographic information science; Geographic information system; Modeling uncertainty in digital elevation models; Representation: 3-D; Virtual reality

**References**


CONSTRUCTING 3-D GIS


Further reading


Foster, Shaun, and David Halbstein. 2014. Integrating 3-D Modeling, Photogrammetry and Design. Berlin: Springer.


The “construction of nature” is the idea that nature, rather than or in addition to being a given in the world and an object of inquiry by humans, is a human or social construction, the outcome of human activities or imagination, often called social or produced nature. This idea, which is commonplace in social constructivist and postmodern literature, has been understood in diverse and contradictory ways due to ambiguities about which nature we are dealing with (the nonhuman, biological, or material world; the whole reality), and whether we are dealing with nature itself or with humans’ representations of nature. This entry attempts to disentangle the confusion that results from this multiplicity of “natures” and proposes a reassessment of the idea in this light. It shows that, depending on how nature is defined, the “construction of nature” is either a trivial idea (if dealing with representations of nature) or implies metaphysical paradoxes (if dealing with nature itself). It argues that (i) radical deconstruction should be pursued but should target the multiple representations of nature produced by humans; and that (ii) defining nature itself as a unique, nonsocial, nonconstructed reality is one of the conditions for successful deconstruction and, more broadly, for the production of relevant knowledge about the biophysical and social worlds and their interactions.

Collingwood (1945) presented a history of the idea of nature that identified three key periods, each corresponding to a particular way of seeing the world and a particular conception of nature. First is the “Greek view of nature,” according to which nature encompasses all things found in the material world and has a mind which is the source of its regularities. Animals would participate intellectually as well as materially in the ordering of the world, making it a live and intelligent organism whose mind and body are made of the same substance.

Second is the “renaissance view of nature,” which appeared during the sixteenth or seventeenth century. The natural world is not viewed anymore as an autonomous organism with intelligence. It is a machine to be commanded from outside, through divine laws created and imposed upon it. Matter and mind are made of different substances, which raised the question of the direction of the causal relation between the two.

Berkeley’s (1710) answer to that question is that the material world is produced by a universal spirit (God), which reveals itself to human minds through an ordered language that we call nature. Nature is thus God’s construction as revealed to humans. This conception contributed to the rise of radical idealist views that consider that nature and the reality itself can only exist as outcomes of the mind. Kant (1838/1781), by making a triple distinction between mind, things, and nature, gave an answer conducive to
less radical idealist views. He showed that “the proper object of scientific knowledge is not mind or things in themselves but nature,” which is “mere phenomenon, a world of things as they appear to us, scientifically knowable because their ways of appearing are perfectly regular and predictable” (Collingwood 1945, 119). The “thing” in itself that constitutes the substance of the world would be something other than nature. It would be inaccessible by experience, and philosophy may be more relevant than science to produce knowledge about it.

These two answers regarding the causal relation between mind and nature show that three different things are typically called nature: (i) the world out there as it is holistically, independent of what we know about it (Nature); (ii) the subset of this world that is intelligible to humans (Kant’s nature); and (iii) the representations of Nature or of Kant’s nature produced by humans (the so-called social or produced “natures”). The frequent conflation of these three concepts may be the main cause of heat and confusion surrounding academic debates about nature. Note that Kant’s nature is the most ambiguous one because even if it is real (there are intelligible phenomena out there), it is defined from an anthropocentric perspective, because only humans can tell what is intelligible to them.

The third conception, the “modern view of nature,” developed during the nineteenth century. Change was increasingly seen as being ubiquitous, leading to the notions of progress in the social sciences and evolution in the natural sciences. The machine metaphor was abandoned because a machine is something finished. It does not change unless it is redesigned, broken, or wears out. The material world was considered again as carrying within itself its own design but this design was no longer seen as something static and perfect. It was considered dynamic, improving, which led to blurring frontiers between long-established categories that describe it. Pairs like structure and function, matter and movement, time and space, were seen as single things, consistent with recent discoveries in physics. It was recognized that the way we see the world depends on when and how we observe it. This conception strengthened claims that certain knowledge about nature was impossible and provided an additional basis on which to argue for the “social construction of nature.”

The construction of nature

In spite of this modern view that considers that nature is dynamic, the view that nature is a static world in balance prevailed until the second half of the twentieth century among scientists, and still dominates today among policymakers and in collective imaginaries. The idea that nature is a construction arose as a reaction against this view. It gained more momentum during the 1960s and 1970s when it was realized, within emancipatory intellectual and social movements like feminism, that invoking an objective and immutable nature was used by dominant groups to legitimate control over minds and bodies and block social changes. Foucault (1982), outlining the connections between power and knowledge, provided philosophical and sociopolitical arguments for contesting this domination. At the same time, social scientists studying laboratories from an anthropological standpoint identified and described the processes that led to the “social construction” of scientific facts (Latour and Woolgar 1979). Inspired by these “postmodern” and “social constructivist” ideas, scholars in the social sciences and humanities emphasized the ever-changing character of the natural world, its hybridization with the social world, the inseparability of sciences and politics, and the impossibility of knowing nature, or of
CONSTRUCTION OF NATURE

knowing it with certainty, depending on the level of radicalism adopted. The idea of nature was considered irrelevant (Latour 2004) or had to be reinvented (Cronon 1996), or multiple truths about nature had to be accepted. Increasingly, nature itself, and not simply the idea of nature, was considered an outcome of human activities. This encouraged emancipatory claims that nature should be managed on behalf of social choices.

The critique of the social construction of nature

Consistently, with its rejection of an objective nature, the “construction of nature” idea led to the disputing or rejecting of the possibility of producing objective knowledge in general. Statements were generally ambiguous, though. They can be represented on a line of increasing radicalism that goes from the trivial (knowledge about the reality is partially socially constructed) to the radical (knowledge about the reality is entirely socially constructed) and the counterintuitive or paradoxical (the reality itself is socially constructed, and multiple). The most radical views triggered a counterattack from biophysical scientists, leading to the so-called science war (Gross and Levitt 1997; Lingua Franca 2000).

Some social scientists also criticized radical or ambiguous constructivist views. Gingras (2000) and Bourdieu (2001) argued that the use of counterintuitive, paradoxical statements was a strategy aimed at creating a hype conducive to mediatizing scholarly work in a context of great demand for radical new ideas that could turn the establishment upside down, while ambiguity enabled the maintaining of a link with common sense in order to retract if the approach was deemed too adventurous. Bourdieu (2001), in order to escape from that hype while making advancement in the sociology of science, proposed a reflexive method of inquiry where science is considered a collective, human activity framed within a given history and social domain, but is also recognized for specificities that make objective knowledge possible. He argues that scientists’ collective agreement with social rules such as control by peers, barriers to professional entry, or acceptance of testing through conducting experiments, are conditions for the possibility of producing objective knowledge about nature. He further investigates social phenomena and processes that have the potential to bias the production of knowledge, such as belonging to a particular social class and the appropriation of a habitus. He recognizes the subjective character of knowledge but emphasizes the need to redress social biases, and identifies the mechanisms that science, as a social institution, has already put in place to do so.

The “construction of nature” today

A recent synthesis (Castree 2014) helps summarize the advancements made possible by the “construction of nature” literature, in spite of and beyond the epistemological and ontological confusion and paradoxes created by the most radical or provocative views.

First, we read in this literature that “nature” is not a given, waiting to be analyzed, experienced or interacted with.” It is, instead, “a way of categorizing and labelling the world” (Castree 2014, 34). The nature that is at stake here is Kant’s nature. It is not Nature, that is, it is not the reality or the “thing” reality is made of. It is, rather, a structure that is superimposed upon Nature in order to make sense of it. This structure can be real, in the sense that there can be something real out there, a subset of Nature, that matches with it, but it is not necessarily real, because it is defined by humans and they
can be wrong. In spite of this, it constitutes the grid that humans, and more particularly scientists, use to render the world intelligible. It is a representation that is being used to produce representations. This is why it absolutely must be questioned. Behind that grid, and whatever it is that actually exists that the grid matches with or not, there is a broader reality, which we called Nature. It is precisely because Nature exists out there beyond the grid that an essentialist view of Nature needs to be retained, as a justification to questioning the grid. The mistake of mainstream positivist science and social constructivism alike was to adopt the essentialist view of nature for the Kantian nature (the grid), instead of adopting it for Nature. This may result from what Bhaskar (1979) called the “epistemic fallacy,” a frequent mistake made by both social and natural scientists and which consists in confusing natural laws, which are invisible generating mechanisms or tendencies, with patterns of events, which are visible, intelligible phenomena provoked by natural laws but to which these laws are not reducible. Once this point is clarified, we can say that Nature is in fact a given, waiting to be analyzed, but this analysis is usually done using a template (Kant’s nature) that is often also called Nature, is frequently confused with it, and is used to render the world and its people intelligible; that is, the conflation of Nature with Kant’s nature is used to exert illegitimate control over people, which brings us to the second point.

The second key statement found in the “construction of nature” literature is that reference to nature gives authority to govern people and ideas. It is so because nature is perceived as being nonsocial, an “ontological given” (Castree 2014, 98). It is nonsocial indeed, if one refers to the essential, non-Kantian Nature. But this essential Nature, although it exists (unless the existence of reality is denied), is unknown, or uncertainly known, and unknowable entirely, so humans cannot claim full authority on behalf of knowledge about it in the conduct of their affairs. Moreover, the nature about which true knowledge claims are made is in fact the Kantian nature, that is, the subset of Nature that is rendered intelligible to humans by its regularities. Accepting authority claims derived from knowledge about Kantian nature alone risks reducing humans and the world they live in to much less than what they actually are, through reification. Through this process, the whole domain that is located in between Nature and Kantian nature, which may be accessible to some extent through philosophy and collective experience, would be denied existence, or considered as if it did not matter.

The third key proposition of the “construction of nature” literature is the argument that representations are “performative” (Castree 2014). That is, the representations of nature and the world, and the grids used to create these representations, determine the way we shape the world materially, through reification. In the end, the things in the world out there reflect the ideas we have about them, even if these ideas did not reflect them in the first place. In other words, representations are “constructs that can themselves construct” (Castree 2014, 139). The world progressively becomes the outcome of the mind but indirectly, as a consequence of humans using their increasingly comprehensive but partially false knowledge to transform it as they wish. The great influence on politics of atomistic views of the social world developed by economists is probably the most terrible, tremendous, and striking case of such a reification process taking place. Humans are redesigning their environment, but also themselves, in that process. We are just starting to question the consequences this can have on us and our environment. We are discovering that limits are set by the way the world actually works; that is, they are set by Nature, not by the Kantian nature that is used as a template
to create human designs. This is why, again, the essentialist view of Nature needs to be retained, as a remedy to the reification of Kantian nature.

Fourth, it is a great challenge to escape the performative character of social constructions because “the question of ‘what do we know, believe, feel and do?’ is umbilically connected to the question ‘who or what is seeking to shape our attitudes, values, emotions and actions?’” (Castree 2014, 98). The production of knowledge and meanings is inseparable from identity, politics, and power, and scientists are not free from these influences. The knowledge they produce about nature is determined to some extent by who they are, which institutions they belong to, and what their social or political agendas are. The more they hold power, or depend on powerful institutions, the more their political agendas and those of the institutions they belong to pervade their work and bias their results, which justifies questioning the knowledge they produce. This is particularly true when scientific work is embedded in unequal social relations like postcolonial north–south relationships, and when the topic and issues at stake are heavily mediatized, as in the case of environmental conservation.

These findings justify pursuing the “construction of nature” agenda. They further show that science may not suffice to understand the relationship of humans with their world, at least for as long as it limits its inquiry to Kantian nature and/or confuses it with Nature. Some questions cannot be answered using the dominant scientific mode but need to be agreed upon in order to collectively confront future changes. Knowledge with no or highly uncertain “scientific” ground might be necessary to fill knowledge gaps in between Kantian nature and Nature. This might explain why myths, which are produced by a larger collective than the scientific community and are socially constructed to a greater extent than scientific knowledge, always played important and positive roles in shaping relationships between humans and nature. They might still be necessary today, sometimes against scientific knowledge, if science fails to account for this space in between Nature and Kantian nature.

A “new” definition of nature?

Today, there is agreement that the “construction of nature” idea enabled progress in the understanding of the natural and social world we live in, in spite of the metaphysical paradoxes and epistemic battles generated by radical or provocative views. How, then, to keep on advancing in social-constructivist research about nature while conserving an essentialist view of Nature as object of scientific inquiry? We suggested that finding a solution relies on clarifying what we mean when we use the term “nature.” For this reason, we conclude by proposing a definition of nature, as well as of a term that is often used interchangeably with it: “the environment.”

We suggest using the term “Nature” for the world out there, as it has been created, by god or by itself, or as it has always been. Nature in the Kantian sense does not fit this definition and should be called otherwise, in order to avoid confusion. If we refer to the work of Bhaskar (1979), it may be that Kantian nature is the “domain of the actual,” where are found patterns of events and regularities generated by Nature, while Nature in the sense we propose here is the “domain of the real,” where are found the laws that generate these patterns. Nature, then, looks like it is another term for reality. But it emphasizes the facts that reality works through its own rules, and that these rules are part of the reality itself. They are the natural laws that organize reality and generate regularities. They have always been there and will always remain, although one
CONSTRUCTION OF NATURE

could discuss whether they are immutable or changing. We believe natural laws are immutable but, as they never act alone and combine in infinite and unexpected ways, patterns of events change. Hence humans cannot change natural laws but they can play with them, combining them in various ways to generate particular patterns of events, that is, particular regularities that enable increasing knowledge and control of the world. Humans can change the domain of the actual (the things that happen) but not the domain of the real (the generative mechanisms of the things that happen). They cannot change nature, but they can change the things that nature creates, by playing with the laws of nature.

Hence Nature is the essence of all things and cannot be socially constructed, by definition. Only representations of it can be socially constructed. The most influential of these representations today is a system of intelligible processes that humans produce through the activity they call science. This system matches to some extent with things or phenomena that exist, because there is indeed an intelligible “world” out there. But this “world” is only a subset of the world. It is defined by a series of features which are mere human representations, some of them true, some of them false. This is why it must not be called Nature.

Another definition of nature, less anchored in philosophy and more operational in everyday life, might also be necessary. In common language, Nature, by opposition to culture, is often defined as the sum of things that are not humans, or that are not the outcome of human agency. We believe that these definitions contribute to the confusion, and we propose, instead, to define nature as aspects that are not the outcome of human agency, together with the generative mechanisms of these outcomes, within things which are, as a whole, the outcome of both natural and cultural processes (of both natural causes and human choices). As humans play with the rules of nature, the things they produce are always hybrids of nature and culture. They are the outcomes of natural necessities as well as human choices and designs. Skyscrapers carry nature in the material used to build them and culture through the design of their architects. Architects need to know the rules of nature to choose the right materials, but nature alone could never build the skyline of large cities. The world inhabited by humans is coproduced by nature and culture. But the outcome of this coproduction is not nature, if we retain the definition above. How should we call it then? It is, simply, our environment. Maybe the second greatest cause of confusion surrounding the “construction of nature” idea, after the conflation of Nature, Kantian nature, and “natures,” is the misguided conflation of the terms “nature” and “environment.”

SEE ALSO: Cultures of nature; Environmental discourse; Environmentalism; Environmentality and green governmentality; Nature; Nature conservation; Production of nature; Social constructionism; Socio-nature

References


Consumption

Juliana Mansvelt
Massey University, New Zealand

Consumption is the set of social relationships, discourses, and practices that focus on the sale, acquisition, use, and disposal of commodities. While a commodity can be understood as a good or service that is sold in a market, consumption itself is much broader than a momentary act of purchase. Consumption incorporates a range of practices that are both material and symbolic, and that may extend from spaces of market exchange to sites of wasting. Consumption has an important role to play in material survival and wellbeing and is important in the ways things are valued and devalued economically, socially, and symbolically. Consumption of commodities is necessary to the continuation of the capitalist mode of production, but also has a significant role in the uneven social reproduction of society, in the production of meaningful social relationships (or socialities), and in the construction of real and imagined people and place identities. As a set of social relations, discourses, and practices, consumption plays a significant role in the formation of built environments and material culture, and is integral to the maintenance and expression of everyday life. A central concern of much writing on consumption has been to understand the connections between production and consumption and between economy and culture.

As a sphere of social processes involving purchase, use, disposal, and wasting of commodities, consumption occurs in place, but also has an important part to play in how spaces are created at a range of scales from the body to the global. Bodies, homes, workplaces and play spaces, cities and rural areas, and nation-states all provide a context for consumption, and are made meaningful through consumption practice as spaces of meaning, experience, and imagination. Consumption also intersects with age, gender, sexuality, race, and other dimensions of subjectivity to influence identity formation. Consumption creates and reflects a range of morally inflected social relations between people, and between people and things, across the life course, over time, and in place, for example, reflecting geographies of love, care, hate, and grief. Geographical research on consumption is currently characterized by a wide range of theoretical approaches and empirical studies, which examine the spatialities, subjectivities, and socialities of consumption in both contemporary and historical contexts.

The emergence of geographic research on consumption – disciplinary contours

While consumption has been an integral part of the functioning of societies and the uneven manifestation of social relations and spatial structures throughout history, it was not until the last few decades that substantial numbers of geographers turned their attention to its role and expression. Most prior geographical consumption research focused on the mapping, modeling, and description of consumer behavior, commodity provision, and the collective consumption of goods and services. A relative explosion of interest in the manifestation, role, and conceptualization of consumption in
CONSUMPTION

the 1980s and 1990s appeared to have emerged from both disciplinary and societal change which became most visible from the 1970s.

The 1970s were also a period in which the continued viability of Keynesian politics and welfare states as a means of securing the economic wellbeing of citizens in many capitalist countries was being questioned. Governments subscribing to neoliberal ideologies sought to roll back the operations of the state and to open state-managed or provided services in areas such as health, transport, education, and housing to market forces. These societal changes emphasized the individual as a “citizen-consumer” of goods and services. Citizen consumers were viewed as rational economic agents, who if given the right information would make efficient market choices. At the same time, political economic changes associated with globalization re-shaped geographies of production and commodity networks, influencing consumption practices. The deregulation of economies, reductions in trade and tariff barriers, and opening up of state economies to financial, information, people, and commodity flows, combined with developments in information and communication technologies, improvements in transportation, logistics, and innovation networks, and the development of flexible systems of production, provided new possibilities for where, what, and how commodities were produced, moved, and sold. These changes also facilitated trade, media and cultural flows, and the exposure of more individuals to a wider range of consumption goods, services, and consumption spaces whether they live in circumstances of lack or plenty. The development of advertising, branding and marketing firms in the earlier part of the twentieth century, and of a range of retail spaces (from markets to malls and the Internet) have also encouraged consumers to reflect on the representational nature of commodities and their inferred importance in mediating everyday relationships and practices. Social changes have been significant too; new forms of consumption practice, access and exposure to an increasingly commodified world of goods and services, and the growing visibility of both real and virtual consumption spaces have provided consumers with a diverse range of ways in which bodies, identities, social relationships, culture, experience and meaning can be constructed and negotiated. However, not all such social, political and economic changes surrounding consumption have been positive. Access to consumer goods and services is still marked by stark inequalities both within and between nations, and the wider ecological, environmental, and social effects of current levels of consumption and resource use (including that contributing to climate change) are of concern to many.

As consumption seemed to occupy a more visible place in social and spatial relations, human geographers, in common with other social science scholars, increasingly turned their attention to the manifestation and operation of consumption. Disciplinary changes were also relevant to the intensification of interest in consumption; the rise of postpositivist conceptualizations of society and space – including humanist, Marxian, and feminist geographies – prompted geographers to view the role and function of consumption with a more critical lens, shifting the emphasis from description and modeling toward explanations of the functions of consumption in capitalist societies. A number of geographers were keen to address what they believed were the productionist and masculinist underpinnings of theories of consumption based on Marxian political economy, aiming to study consumption as a meaningful sphere of social processes in its own right. The development of postcolonial and poststructuralist theories challenged many of the assumptions implicit in earlier foundational theories of consumption.
In emphasizing diversity and difference, these perspectives supported a more reflective and situated analysis of consumption, its historical emergence, and its role in configuring social and spatial relationships. Research on consumption continues within both positivist and critical human geography, although the explicit identification of researchers as geographers of consumption seems less evident than in the early 1990s to 2000s, perhaps as a consequence of the increased recognition of the legitimacy of consumption research, and of the growing complexity and diversity of geographical work on consumption. Research on geographies of consumption is now more evident as an integral part of research in other subdisciplinary fields. Geographers less often assert the value of and need for consumption research, but continue to contribute to understanding of the ways in which consumption impacts on commodity networks (including those concerned with waste and sustainability), how it is entwined in moral and ethical geographies, how it involves the intersection of human and nonhuman relations in place, and to the role of consumption practice in reproducing a politics of embodiment, mobility, and identity.

Consumer spaces

Much of the early postpositivist work on consumption by human geographers was concerned with how consumption was produced in place and emphasized the more spectacular spaces of consumption such as malls, theme parks, and fairs. This writing was primarily informed by Marxian political economy, which understood these sites as representational spaces designed to encourage consumption and the realization of surplus value. While this research was critiqued for its productionist and masculine focus, research on shopping malls, fairs, and theme parks and other consumer spaces continues to reveal the significance of these spaces as part of contemporary consumer culture. Many of the early studies drew on traditions of reading landscape within cultural geography, highlighting how imagined geographies and the created environment of retail space had a significant role in encouraging practices of consumption (Goss 1999). In focusing on the production and manifestation of these spaces, however, such studies tended to emphasize the production of space for the consumer, rather than by the consumer, an aspect subsequently addressed by ethnographic research which has highlighted the practice, experience, and embodiment of consuming. Work on spaces of consumption now encompasses a diverse range of first- and second-hand retail spaces (malls, stores, high streets, markets, and the Internet) but also examines how sites of work and leisure such as homes and gardens, work, and tourist spaces are shaped in and across different contexts. The bulk of research focuses on urban areas although geographers have highlighted the impacts of the commodification of rural places and nature. Rather than endeavoring to understand these sites of consumption as discrete entities, much geographical research examines how these spaces are embedded in and connected to social, political, cultural, and economic relations across a variety of scales.

Identity, embodiment, and consumer practice

From the 1990s, geographers using ethnographic methods have studied the practices, experiences, identity, and embodiment of people as consumers, providing opportunity to consider the mundane but significant work that comprises consumption. This acts as a foil to beliefs that consumption is a frivolous, wasteful, and always
CONSUMPTION

pleasurable realm of social life. Initially focusing on shopping, geographers have turned their attention to appropriation and possession, and the use, reuse, and disposal of commodities in order to understand their significance in the constitution of people and place relationships and identities. With regard to acquisition of commodities, a number of geographers have argued that the purchase, display, and use of commodities, while linked to cultural imperatives to construct consumer identities and lifestyles, is more often a means of expressing sociality – that is, of establishing and maintaining relationships with significant others. Research on consumption practices (such as shopping, drinking, and eating) has consequently revealed the complex ways in which consumption practice and commodities are managed and negotiated in everyday life. Geographers have demonstrated that discourses of race, age, class, gender, and sexuality are both a reflection of and an instigator of forms and practices of consumption. In recognizing that consumer identities are heterogeneous and ephemeral, geographers have revealed how these alter and are framed differently in response to life course changes and transitions. The role of consumption in shaping parental and familial relations and in relation to both heterosexual and homosexual identities has also been a focus. While work on meanings and experiences of consumption practices have predominated, more recently the embodied, sensorial, performed, and emotional constitution of consumption practice has been given greater consideration.

Commodities and commodity networks

A third area of geographical investigation relates to commodities. Most of this work has centered upon commodities that have a lasting material form, rather than services, although there have been some studies of consumption of financial and tourist services. Commodity studies of food have been prevalent, but music, consumer durables, drugs and alcohol, and apparel/fashion have also featured. The role and function of commodities as objects of material culture has been emphasized and insights from material geographies and actor network perspectives have informed geographers’ interpretations of the consumption of things.

Geographers have been interested in the ways in which things in commodified and noncommodified moments and spaces enable insight into how places and social relations (such as families, homes, friendship, communities) are constructed. For example, it is possible to understand housing-related consumption as a materialization of the imagined construction of home (Jacobs and Smith 2009). Geographers have also considered how the material properties of commodities imply certain competencies, revealing the ways commodities influence and accommodate human practice. This research has demonstrated how the properties of nonhuman things can be important in the construction of visceral and emotional experiences of activities such as cooking, listening to music, and DIY (Do It Yourself). Accordingly, commodities can be understood not simply as repositories of value but as artifacts of memorialization, identity, embodiment, and emotion. Geographers have also begun to unpack the ways in which mundane consumption practices help shape wider political and economic geographies, such as how the materiality of household disposal is connected with municipal systems of recycling and disposal (Bulkeley and Gregson 2009) and how the acquisition of air conditioning can become a symbol of nationhood and state citizenship (Hitchings and Lee 2008).

The second tradition of consumption research centered on the commodity emphasizes the
social and spatial lives of commodities through commodity chains from production to disposal, often following the work of anthropologist Arjun Appadurai (1986). Geographers have been concerned to understand not just commodity trajectories and how things move in and out of the commodity form as they are produced, acquired, used, disposed of, and wasted, but also the spaces through which these processes occur, and how the social relations, discourses, and practices surrounding commodities and commodity flows are themselves produced across different places and scales. Much of this research has focused on following agricultural and food commodities (Cook 2006), but geographers have also been interested in addressing the productionist focus of much commodity chain analysis to consider how the consumption of commodities, consumer practice and consumer anxieties reverberate through commodity networks (Jackson 2006). As part of research on the shaping and movement of commodities, research on commodity cultures surrounding commodities such as food and music has provided insights into the ways in which cultural and economic practices are shaped transnationally. The connection between processes and cycles of production and consumption and their complexity in a globalizing world has also been illuminated by work on wasting, which seeks to look beyond purchase and acquisition of commodities to understand how the materiality of waste and disposal practices influence a range of other production and consumption activities (Gregson et al. 2010).

**Ethical geographies and sustainable futures**

Through geographical research on consumer spaces, practices, and commodities, geographers have highlighted the ways in which ethical relations are shaped with near or distant others and the kind of moral politics created. In common with work on moral economies, this research has recognized that consumption is not a neutral sphere of social relations, but rather comprises an arena for the exercise and expression of power. Consumers, for example, are encouraged to care through purchase of fair trade, slow, “green,” and cruelty-free products, and “Cool Aid” consumerism (Andrews et al. 2011). A great deal of work on ethical consumption has explored the formal shaping of ethicality across commodity chains and networks with an emphasis on how consumers are enrolled and engaged in caring at a distance. In studying consumers, and particularly household practices of consumption, geographers have shown how sustainable practices may be constrained or enabled by structural, political, and economic contexts. Hence geographers have revealed that behavioral change of individual consumers is not in itself sufficient for sustainable consumption to be achieved. They have demonstrated the desirability for green consumerism, for example, to be aligned with other consumption imperatives, actions, and preferences. While recognizing that all consumption practice is political and underpinned by forms of moralities, research on “ethical” forms of consumption such as farmers’ markets and food security schemes has highlighted the need to interrogate the unintended (and sometimes negative) effects of such practices on others.

Geographical research on consumption has not only revealed the material and representational structuring of places for consumption, but has provided insights into the place of consumption in contemporary and historical social life, emphasizing the complex ways in which consumption is shaped and made meaningful in and across space. It has highlighted not just the ways in which places are made meaningful for consumers, but how consumption is a meaningful
CONSUMPTION

social practice in place. Through studying consumption spaces, consumption practice, commodity trajectories, and material attributes, geographers have provided important insights into operations of power and the moral geographies that underpin them. While the bulk of consumption research is based in Anglo-American contexts more research from other nations (especially in the Global South) is emerging.

Attention to the uneven political and moral constitution and expression of consumption, and to concerns about resource use, sustainable futures, and the environmental impacts of consumption including deforestation, food insecurity, declining biodiversity, pollution, and climate change, provide further possibilities for the extension of consumption research and for developing greater intersections between human and physical geographies. In terms of reflecting on the role of consumption in securing alternative futures, there is still much work to do.

SEE ALSO: Actor-network theory; Commodification of nature; Cultural politics; Environment and consumption; Everyday geographies; Global commodity/value chains; Identity; Inequality; Social geography; Social reproduction

References


Further reading


In its strictest sense, the term “containerization” applies to goods of all kinds that are placed in large steel containers which can be transported by many different modes. It is a technology that developed in the 1960s for ships, but which has greatly facilitated the movement of goods between different transport modes. Its impact has been very extensive: transforming shipping and seaport operations and locations; changing relationships between port cities; reshaping the organizational interactions in port city regions; and facilitating the globalization of international trade and industry.

Prior to containerization, freight had to be loaded or unloaded separately between the dockside and the ships’ holds in a manner that had evolved little since Antiquity. It was a system that required a lot of labor and handling was a lengthy process. Cargo ships would spend on average three weeks in port unloading and reloading. Shipping lines generate revenues only when the ships are at sea moving between ports, and port calls are a cost factor. The number of revenue-generating voyages was limited prior to the adoption of containerization.

Because of their standard size, containers can be lifted between ship and shore by specialized gantry cranes thereby obviating the need for human resources, and can also be moved around port terminals and stacked by other mobile equipment. They can be placed on trucks, rail, or barges for onward delivery with speed and efficiency. Containerization thus represents a revolution in mechanized cargo handling, facilitating both scale economies and improvements in speed of handling.

While there had been growth in container traffic in American waters ever since the first ship owned by Malcolm McLean sailed from New York to New Orleans in 1956, the first international sailing did not take place until 1966. The economies of containerization were recognized more widely thereafter, beginning a massive conversion of general cargo into containers. This conversion of existing traffic has been one of the principal factors of growth in container traffic.
shipping, but it has been greatly supplemented by the growth of trade around the world from the 1970s to the present. Today less than 10% of nonbulk cargo is handled by noncontainerized ships. Recently a new factor has become evident with the conversion to containers of certain kinds of bulk cargoes. These include lumber and timber as well as grains.

The growth in containerized trade globally has been spectacular since the 1980s, exceeding gross domestic product and population growth by large margins. If measured by the number of containers handled in ports around the world, traffic has grown from 36 million TEUs in 1980 to 601 million in 2012 (see Figure 1). The trajectory of containerization has been marked by significant regional shifts. Today, East, South, and North Asia are the most important markets, Europe is second, and North America third. This is reflected in the dominant trade lanes, with intra-Asia routes accounting for most traffic today, Asia-Europe second, Asia-North America third, and the transatlantic fourth.

Container shipping has been a factor in the process of globalization of industry and commerce. It has provided a safe and inexpensive way of shipping components and finished goods around the world. The size of the box means that it can accommodate large quantities of goods. For example, it is estimated that for a pair of running shoes sold in the United States for over $100 the transport cost from China is 25 cents. Of course, globalization has been enabled by other factors such as trade facilitation features, but data cited by the United Nations Conference on Trade and Development (UNCTAD) (2013, 23) suggest that between 1970 and 1990 bilateral trade facilitation accounted for 45% of global growth, that membership in the General Agreement on Tariffs and Trade (GATT; the precursor of the World Trade Organization) accounted for global growth of 285%, but that the container as the driver of world economic growth was responsible for growth of 790%. Evidently there is an interdependent relationship between containerization and globalization: the global spread of industrialization and the growth in consumption has provided the shipping industry with greater demand and new markets that it can uniquely serve.

**Figure 1** Global container traffic 1980–2011.

**Containerization and the shipping industry**

The growth in world trade precipitated a significant increase in the size and capacity of container ships. Scale economies are significant, especially where voyage lengths are great. For long trade lanes, such as North Asia-Western Europe, costs per container slot of a 2000 TEU ship were double those of a vessel with twice the capacity. These economies were made possible by improvements in engineering and ship design.

The first specially built container ships appeared in the late 1960s. These were ships with a capacity of 1000 TEUs. The progression in the 1970s was to double the capacity and increase
the speed. By the early 1980s the capacity had doubled again to in excess of 4000 TEUs. By this time vessels reached the maximum size to be able to traverse the Panama Canal, hence they are referred to as Panamax. This represented a plateau that was not exceeded until the late 1980s, when the first post-Panamax ships were deployed. By 1996 the largest ships reached 7000 TEUs and within 10 years had grown to 14,500 TEUs. These megaships entered the market just in time to benefit from the boom preceding the premarket collapse of 2007–2008, and several companies placed orders for vessels of this capacity. The addition of significant new capacity coincided with a global reduction in world trade, and resulted in serious financial difficulties for the carriers. Nevertheless, in 2011 Maersk (the largest carrier) placed an order for 18 new super post-Panamax ships of 18,000 TEUs. The first of these ships began entering into service in 2013.

Thus, in the space of just less than 50 years, from the first international sailing of a container ship in the 1960s to the present, container vessel capacity has increased 18-fold. In no other transport system has such growth in size taken place. The world fleet of fully cellular container vessels has risen spectacularly, from nothing in 1960 to 206,577,000 deadweight tons in 2013.

Container shipping has required significant capital investments to build vessels that are increasingly large. Thirteen of the top 20 carriers in 1979 had disappeared by 2013. Many of the older companies who had played a major role in the conversion to containerization were among those that disappeared, while many newer companies have entered. Thirteen of the carriers in the top 20 in 2013 are Asian, many of which are either subsidiaries of major industrial and banking conglomerates, such as the Japanese MOL (Mitsui OSK Lines), NYK (Nippon Yusen Kabushiki Kaisha), and K-Line (Kawasaki Kisen Kaisha) or the Korean Hyundai and Hanjin, or largely state-owned, such as COSCO (China Ocean Shipping Company) and APL (American President Lines). However, family firms continue to play a big role in the industry. The top four carriers, three of which are European, are largely under family ownership, and there are four other family firms in the top 20. The family firms have been among the most dynamic of the entire industry, introducing new vessel types, developing new networks, and entering new markets. Their organizational structures are simpler than those of the other corporate firms, and are thus better able to respond quickly to changes (Slack and Fremont 2009).

The demand for containers and the growth in vessel size necessitated important spatial and operational restructuring of container shipping. In seeking to serve global markets, the carriers have designed a range of network configurations, including end-to-end, line bundling, trans-shipment, and relay (Ducruet and Notteboom 2012). Trans-shipment has become an important feature of the industry, accounting for nearly 30% of all container traffic. It is especially important in the Mediterranean and Caribbean seas, involving ports such as Gioia Tauro, Italy, and Kingston, Jamaica.

The global nature of the market has led many carriers to combine in strategic alliances, enabling them to pool resources and mount combined services. This has been the practice of most of the Asian carriers. On lesser trades, where demand may be limited, even the big European carriers may offer joint services with regional shipping lines or may buy space on ships operated by other lines, a practice called slot-sharing.

A major challenge to shipping has been the increase in the cost of fuel (bunker). Between 2005 and 2012 the price of bunker nearly
CONTAINERIZATION

tripled, from $138 per ton to $639 per ton. At these prices fuel costs represent over half of operating costs. This has necessitated action by the container shipping lines. They have added a surcharge to their customers and they have reduced the speed of their ships from 20–23 knots to 15–18 knots in order to conserve fuel, a practice called slow steaming. Orders for ships with greater fuel efficiency have been placed as well. Slow steaming means that delivery times are lengthening, requiring adjustments along supply chains. However, the speeds are now below the optimum efficiency for many older vessels, and may produce more emissions and require more maintenance as a result. The new shipbuilding orders are for ships with engines with greater fuel efficiency that can operate at lower speeds. This is a feature of the new orders by Maersk for 18,000 TEU ships whose twin engines have been totally redesigned to reduce fuel consumption and emissions.

Containerization, intermodalism, and supply chains

The box is eminently designed to be transferred from dockside to other modes of transport. There are no global data on the modal split, but it can be safely said that most containers in the world are moved by truck to and from dockside. In part this is due to the spatial location of producers and consumers. In much of the world, such as Japan, China, South America, Australia, and Africa, markets are located in coastal zones and can be served most effectively by trucking. The opportunities for transfers to other modes are limited therefore. In the United States, however, many large markets are located hundreds of kilometers from ports, and thus rail is an important factor in inland distribution (Plant 2002). This has led to an innovation, where the contents of maritime ISO containers are transloaded into larger capacity domestic containers 53 ft in length in the port cities, which are then transported by rail to inland hubs from where they are delivered by truck to local customers.

Europe is an intermediate case. Many markets are coastal and others are at distances within the competitive range of trucking. Rail transport is limited to certain corridors, from Hamburg to Central Europe for example, but as the rail industry regulations are being reformed, more and more services are being developed (Debrie and Gouvernal 2006). A further inter-modal development in Europe is the use of rivers and canals, especially along the Rhine corridor.

Container shipping has reduced unit transport costs by sea, but on land distribution costs are frequently higher. Land connections and market access have become an important focus for reducing costs and improving efficiency along supply chains. Two elements are at work: first, attempts to relieve growing congestion in ports; and second, managing the supply chains. Logistics plays a major role in both areas. Inland ports have been established that evacuate containers directly from the port terminals by shuttle services, and where preliminary sorting, storage, and onward shipments to distribution centers are made (Rodrigue et al. 2010). The distribution centers are located close to major markets. It is there that containers are stripped or filled, as the case may be, and their contents are sorted, labeled, and assembled. These port-focused networks are managed by third-party logistics providers, or major shippers themselves. This logistics-based integration of the maritime industry with the inland distribution systems is giving rise to a new form of spatial phenomenon, port regionalization (Notteboom and Rodrigue 2005).
Containerization and local and regional impacts

The local impacts of containerization have been profound and not always beneficial. The victims of containerization have been the dock workers (Levinson 2006). In the past, gangs of longshoremen toiled physically carrying cargoes to dockside, where slings and ships’ hoists lowered them into the holds and where the goods then had to be stacked. They were a significant labor force that had established their homes in proximity to the ports, forming distinctive communities. Furthermore, the sailors spent weeks in port and there developed businesses, such as bars, rooming houses, and brothels in “sailor towns.” All this changed with containers. Modern gantry cranes can load and unload boxes at rates of 30 lifts per hour per crane. If 20 of these boxes are 40-footers, the hourly rate is equivalent to 50 TEUs. If the contents of each container weighs 8 tons, one crane with one driver can handle 400 tons of goods per hour (in many Asian ports these rates are higher). This volume of cargo would have taken many gangs many days to handle. Thus, the old system of labor in ports had to change, with painful consequences for the workers, their families, and communities.

At the same time, containerization required ports to expand to new sites. These facilities were often far removed from the old port areas, divorcing further the port from the old port city communities (Hoyle 2000). This has been captured in models of port morphology that reveal the spatial separation of terminals and the increasing dimensions of new sites. These intraport models are being integrated with interport models of port system development as well as with inland networks (Notteboom and Rodrigue 2005).

The relocation of container terminals from the old port sites has afforded many port cities opportunities to redevelop urban sites that were redundant. These redevelopment schemes, such as London’s docklands, or San Francisco’s Fisherman’s Wharf, have been transformative in attracting commercial institutions, gentrifiers, and tourists, but few have taken into consideration the nearby communities that suffered the dislocations and hardships of the transformative changes of containerization.

Future trends

Today containerization is a well-established global intermodal transport system and has now achieved a state of maturity. A challenge for the future will be an old problem – how to match supply with demand. The demand is cyclical, as market conditions fluctuate, but ships have a lifespan in excess of 20 years. In addition, in developing vessel capacity to meet the demand shipping lines face problems, since the provision of new ships is a lengthy process, in which the design, the securing of financing, the placing of the order, the construction, and the sea tests extend over several years. Alternately, they may seek to lease ships. The Baltic Exchange in London is where this takes place, and its index is a widely held indicator of global economic activity. In periods of high demand for vessels the rates are very high, and conversely in low demand periods the index is low. Thus, when high market growth occurred, such as the period between 2005 and 2007, the Baltic Exchange index was very high, and many ship owners decided to place very large orders for new vessels, only to find that on delivery the market had collapsed, leaving them with extra capacity on their hands and facing greatly reduced freight rates.

For the last 20 years or more experts in the industry and in academia have forecast that vessel size increases would cease, only to see size passing 5000 TEUs, then 10 000 TEUs, and then 15 000
CONTAINERIZATION

TEUs. In 2013 Maersk’s new 18 000 TEU vessels entered into service. With this history it would be foolish to predict no further increases in size. While each increase in size limits the number of ports that can accommodate the vessels because of the increasing dimensions, there are always ports willing to invest in dredging deeper channels and expanding terminals. The ultimate limits may be established by the Suez Canal (24 m) and the Straits of Malacca (25 m).

SEE ALSO: Containerization; Corporations and global trade; Global commodity/value chains; Logistics; Ports; Transport geography; Transport networks

References


The intertwined concepts of center and periphery, denoting the relative positions of given entities in space, are widely used not only in the social sciences, but also in urban design, physics, and the history of science. In the case of physics and the study of the atom, we know that electrons are tiny particles that randomly wander around a central nucleus, wherein are located the protons and neutrons, the particles that contain the bulk of the mass of the atom.

In the context of the social sciences, the core refers to regions that have consolidated their hegemony in terms of leadership in science and technology and economic power, that is, the highly industrialized countries in Europe and North America and some Asian countries.

Even before the end of World War II, Polish economists such as Kalecki (1993) and Rosenstein-Rodan (1943) were concerned about the future of the countries of central Europe; in particular, with how to overcome the effects of late capitalism that impacted on them and put them in a peripheral position in relation to the Western European countries. Their question, essentially, was: What is the future of this periphery?

For liberal or neoclassical economists, underdevelopment was only a delayed stage in the onward march of capitalism. In 1958 W.W. Rostow presented his theory of the stages of economic growth as a universal evolutionary process through which every society necessarily has to pass. There are five stages of growth: the initial stage of a traditional society; a transitional stage; the stage of takeoff or “boot”; a phase of “maturity”; and finally the age of mass consumption. This route, which is the path of capitalism, was founded on the premise of progress, rooted in Western culture, and on technical progress, which was seen as neutral and as an expression of rationality (Rostow 1958).

A different theory of development had been offered by Raúl Prebisch 10 years earlier, in 1949. In his introduction to the first economic survey of Latin America, The Economic Development of Latin America and Its Principal Problems, he elaborated the thesis of “unequal exchange” that structured center–periphery relations (Prebisch 1950/1949). The appearance of this publication was also the occasion of the launch of the executive secretariat of the Economic Commission for Latin America.

In his 1964 book on the dynamics of Latin American development, Prebisch analyzed the problem of external bottlenecks that the region was subjected to, and indicated the congenital weakness of the periphery because of which it remained unable to retain the fruits of its technical progress. These regions imported manufactures, the unit values of which increased rapidly, while the unit values of their primary exports rose slowly. Attempts were made to overcome this imbalance via industrialization through import substitution policies. But his interpretation did not neglect politics, because he realized that economic development could not accelerate without the transformation of the social structure. And he could see that democracy would be threatened without an articulated process of income distribution.
For many Latin American developmentalists, a sociocentric concept of development would have to mean, at the very least: (i) recovering a critical and long-term vision; (ii) establishing as a fundamental axis a role of the state which assumes responsibility for contributing to the definition of a national strategy in the medium and long term with a view to achieving greater democracy and overcoming poverty; and (iii) focusing on citizenship, that is, the creation of participatory institutions that would facilitate decentralization, regionalization, local initiative, and a profound cultural change.

During the second half of the twentieth century, a change in the developmentalists’ understanding of the concept of development seems to have occurred. The transformations that had been taking place within capitalism itself, and their perverse social effects, in turn produced a partnership that was assuming new forms of articulation and action. Among these were the movements of peasants, slum dwellers, ethnic communities, environmentalists, gender activists, youth, consumers, human rights activists, and, more specifically, those social segments that had traditionally been marginalized, even including middle-class sectors. All of these groups began to draw up more comprehensive demands featuring new nontraditional social actors. Among these new demands were ones relating to solidarity, participation, and partnership, and to forms of development that were not limited to economic growth and that, until then, had never been visualized as processes of development. But, at the same time, during the extended periods in which nonrepresentative political regimes in many peripheral countries implemented structural adjustment policies demanded by the International Monetary Fund, strong bonds of a shared perception of predatory self-interest between large segments of the upper and middle classes, who aspired to achieve North American levels of consumption, had also emerged and became consolidated. This process was further reinforced by the political alienation of the rest of the population.

A theme that has a central place in the work of Celso Furtado (1983) is underdevelopment, a theoretical challenge on which he worked all his life. For him underdevelopment is an autonomous historical process, in the sense that it is not a stage through which economies have to pass in order to achieve a higher level of development. Although Furtado recognized this autonomy as necessary for the completion and maturation of the phenomenon, he also understood that an underdeveloped economy should not be considered in isolation from the international division of labor in which it is embedded. He also recognized that underdevelopment is essentially a phenomenon of cultural and political domination.

He believed that an explanation of the process of capital accumulation should lie at the very base of the theory of economic development. Choices and economic decisions are policies. However, investments and allocations decided by economists are regarded as something unambiguous. This supposed “axiological neutrality” of the procedures of positive economics was not accepted by Celso Furtado, and he posed arguments and questions of a different order: “What are the relationships that exist between social stratification, systems of domination and the changes that occur in a society as a result of accumulation? How can economic development be integrated into the process of social change and how does it relate to the decision making systems and structures of power?” (1976, 25).

To analyze the state of the periphery in the period since the mid-1980s, which has witnessed a technological revolution and a strong and concomitant process of globalization, it is essential to establish a link between technology
and development theories, mainly because the periphery has tended to remain a passive recipient of innovation and not a focus of creativity (Costa Lima 2008). It is from this perspective that Furtado (1998, 26) propounded the concept of a “technological imperative,” which holds that the trend of globalization is such that markets will impose their logic and rhythm on the world economic space.

In the 30 years after World War II the history of capitalism has been characterized by a marked geographic concentration of industrial activities, alongside a more egalitarian distribution of income, in core countries. These trends are the result of the struggles in these countries of organized workers which led to a rise in real wages, as well as of workers’ demands that their governments adopt policies to protect domestic markets and the fact that these countries did not suffer colonization. The dynamism of the capitalist economy thus grew out of the interaction of two factors: technical innovation, which is reflected in increased productivity and achieved mainly by reducing the demand for labor, and the expansion of the market with higher wages.

It is important to note the inherent logic of capital: that the development of a society is not independent of its structure, and, furthermore, that increases in the efficiency of the system, which are generally identified as the principal indicator of development by conservative economists, are not a sufficient condition for satisfying the basic necessities of the population. It may even be the case that the introduction of more sophisticated techniques leads to a degradation of living conditions of the bulk of the population. In other words, the technical advances may benefit only certain social groups (Furtado 2000). It is, therefore, important to consider the quality of the development, to ensure that the efforts made to innovate and promote technological development benefit the largest possible number of people rather than generating or reinforcing a structure of privileges.

There exists a rich corpus of theory on development, including, among others, the writings of heterodox authors like François Perroux (1969), Gunnar Myrdal (1957), and Albert Hirschman (1958), who criticized the concept of stages and spoke instead of asymmetries, dualism, and dependence. There are also Marxist authors who rejected developmentalism and dualism altogether, preferring the concepts of imperialism and the colonial past which, they believed, impeded the process of autonomous national development in the periphery. It is in this tradition that Andre Gunder Frank (1966) presented the thesis of the “development of underdevelopment,” which was further explored by Samir Amin (1967) and Arghiri Emmanuel (1969), who dealt with the mechanisms of global accumulation, from which one can escape only by rupture, or a clear break with the prevailing realities created by capitalism. World-systems theory, which is mainly attributed to Immanuel Wallerstein (1974), but was also enriched by the contributions of Amin and Frank, is rooted in classical sociology, Marxist economic policy, and the thinking of dependency theorists.

The works of Emmanuel and Amin explain the extent of the leakage suffered by the peripheral countries as a result of deteriorating terms of trade, that is, disparities are perpetuated by the unequal exchange that prevails between the countries at the periphery and those at the center. For Amin, the transfer of the surplus to the center, in various forms, is the central issue, and these mechanisms of primitive accumulation feed capitalist expansion (Costa Lima 2011, 146).

According to Chase-Dunn (2007, 1060), the modern world system can be understood structurally as a stratified system: first, composed of central societies that are economically, culturally, and militarily dominant and in competition
with each other; and, second, by peripheral and semiperipheral regions. Some dependent regions were, of course, successful in improving their relative positions within the broad center–periphery hierarchy, but most had to remain in their peripheral and semiperipheral positions. This structural approach allows us to analyze the cyclical profile of social change and the long-term pattern of development in a historical and comparative perspective.

In the early twenty-first century, the center–periphery structure has continued to remain intact in global comparative terms. The center, with its large international capital, has not only gained a global specialization, but has also undergone radical technical transformations in production. The role of the dependent sphere as an exporter of capital continues through the permanent mechanism of debt leakage, and this situation leads to political, social, and economic instability, with predatory effects on the environment.

The drastic changes brought about by globalization and financialization have rapidly transformed the social reality and interpretive categories. The consolidation of financial hegemony, mediated by stock exchanges or by mutual and pension funds, redefines a new geoeconomic orientation under the control of the United States. In this new context, Latin America has completely abandoned its project of development, whether autonomous or dependent, and is now working under conditions of financial bankruptcy that require external control by national banking systems, through privatizations, and also by the states and their instruments of intervention. As Fiori (2001, 82) eloquently affirms: “financial capital diluted and softened to the maximum possible extent the variable frontiers of its economic territories, moving from country to country and from one world region to the other, without creating any permanent attachment, much less any kind of ‘civilising’ project for the periphery of the system.” This turns the countries of the periphery into hostages to the logic of international movements of capital and subject to their moods and the crises created by them.

The contemporary theoretical and empirical discussion arising out of globalization processes—which focuses on changes that are taking place in space, technologies, production processes, the maintenance of social inequalities, regional dynamics, territories, the phenomena of metropolitanization, the environment, and the role of the state—began to require an inter-, multi-, and transdisciplinary approach to these new realities and phenomena (Costa Lima 2013). How can the fact that economic activities tend to concentrate in a finite number of well-defined points in space be explained? How can the new hierarchies and centralities of development be explained?

In geography, the first studies that sought to identify an urban structure organized around a center/periphery dichotomy were more descriptive than properly paradigmatic or theoretical. In general, the empirical findings about urban structure and the distribution and/or segregation of the population in certain areas within the city could be observed in the studies of Georg Simmel and Max Weber, who largely influenced the research developed by the Chicago School. In 1925 Ernest Burgess conducted a pioneering study, which proposed the construction of an analytical model that included the expansion of the city around a central point, and identified several concentric circles as one moved toward the periphery which spatially delimited the various forms of sociofunctional clustering of a city.

From a functionalist perspective, segregation, dispersion, and urban inequalities were examined in analyses focused exclusively on the individual, based on the notion that his/her housing decisions would be exclusively guided
by tastes, preferences, and social networks. These were examined, therefore, in a reductionist manner that made the social production of space a product of individual rationality, leaving aside economic and structural variables.

Differentiated and critical approaches came into vogue which sought to understand the areas and urban spaces that were peripheral, distant, precarious, and devalued in the property market. These also focused on the conflicts that occurred, on the one hand, because of the violence of the status quo and marginalization, and, on the other, because of demands for the allocation of urban space from the poorer sections of the population whose access to alternative housing was foreclosed by the market or the state, thus creating irregular settlements, built without order or outside the law, on protected or inappropriate land. At the other end was a center, well-equipped, valued, and with a large number of highly skilled jobs and services. It is in this context that the confrontation takes place – between the unused and closed properties awaiting an increase in market value and the possibility of low-cost housing and renovation projects that, in the process of execution, expel the poorest sections of the local population. It is here, therefore, that the theories created in the countries of the North and the South are clearly differentiated from one another, with regard to their application to advanced industrial regions or to the regions of low development. The regional disparities are not just international phenomena but are also national in scope.

Efforts are underway to further develop theories on the geographic concentration of economic activity. Cities are perceived as places where material and immaterial exchanges happen, where networks of various forms and contents settle in continuous, endogenous, and exogenous flows. They are places where innovation happens, producing information and knowledge. But these cities also have their networks of hierarchies, structured in a center–periphery manner.

In recent years, a growing number of economists have turned their attention to this phenomenon and, in a broader sense, to their impacts in relation to regional development. Since the early 1990s the new economic geography (NEG) has provided several theoretical studies of regional and urban structures. As stated by Fujita, Krugman, and Mori (1999), cities have increased their importance as the basic units of international economic systems.

Since the late 1990s, income inequality, as measured by the Gini coefficient, has increased in almost half of the developing countries, but, according to the data presented by Joseph Stiglitz (2012) in his latest book, inequality has also increased significantly in the United States, the richest country in the world. The Nobel laureate speaks about the growing inequality of income and wealth in the United States since the 1980s. In the Preface, he notes that “there are moments in history when people around the world seem to stand up to say that something is wrong.” His data are conclusive on the effects of economic policies that generated inequality of income and wealth in the United States after 1980, namely that 25 years ago the richest 1% of Americans owned 33% of national wealth, and that in 2007 that figure had risen to 40%. In 1976 the richest 1% received 9% of national income compared with nearly 25% today. The real income of a typical male full-time worker has stagnated for more than one-third of a century.

As Dunford and Yeung (2011) point out, global economic inequalities, in general, have been increasing over 200 years. According to them, the reason for this is that very few economies in the pursuit of economic and social modernization had managed to achieve sustained growth. More recently, the rise of Asia has been an important step in the reconfiguration of
global development, precisely because a number of large emerging market economies have rejected the prescriptions of the Washington Consensus. One of the reasons for this was, thanks to increased global integration, a sharp increase in the global workforce. A second and more important reason was that, in Asia, as a result of currency devaluations, strong growths in exports were achieved in some countries. These emerging countries, which had earlier been debtors, were transformed into creditors while rich core countries became debtors.

Dunford and Yeung’s (2011) central argument is that, in all likelihood, the group of emerging economies, especially China, will sustain its recent advantages of growth albeit in the context of slower global growth. In the Western world, a set of strategies designed to restore sustained growth has proved insufficient, and the recent economic crisis will have an even bigger and more stubborn negative impact in the developed world because these countries do not have an alternative model of growth in reserve. China, on the other hand, has the potential to sustain relatively high growth rates for many years. To do so, however, it will need to bring about profound changes in its development model. If it manages to do so, and if the other major economies in the BRIC group, India, Russia, and Brazil, also continue to grow, there is a strong possibility that these emerging countries will lead global growth. If this happens, then, within 25 years, Asia will account for 66% of global GDP. The global configuration created by the Industrial Revolution, colonialism, and imperialism would then be reversed. Inequalities of income per capita will still continue for a longer period, but the center–periphery relationship will have undergone substantive changes.

The global economic crisis that began in 2008 has given rise to perplexities and changes in our understanding of global phenomena. Economic inequality in all its forms has forced as many as 1.57 billion people to live in a state of “multidimensional poverty” (PNUD 2013). In any event, after the big changes taking place in emerging countries, led by China, new concepts are still worth designing with a view to developing new perspectives on the long-established center–periphery structures which are becoming ever more complex. Given this context, the question is to what extent the realities subsumed within the current dynamics of North–South and center–periphery relations are likely to change, along with their interpretation.

SEE ALSO: Dependency theory; Development; Developmentalism; Industrialization; Power and development

References


Corporate environmental responsibility

Abigail Efua Hilson
Royal Holloway, University of London, UK

Corporate environmental responsibility (CER) is a voluntary activity by corporations to assess and minimize their ecological footprints in the course of conducting business. These activities fall outside the legal requirements of organizations to reduce the levels of pollution, emissions, and environmental degradation associated with their business activities.

Organizations were previously perceived as closed systems, whose links to their environment and society were of little interest. As a result, organizations existed for the sole purpose of making a profitable return for their owners or shareholders. In recent years, recurring corporate scandals over loss of livelihoods, destruction of the natural environment, and the collapse of financial markets have inspired a rethink of how organizations can and should operate. Specifically, how their existence affects society and the environment is now being formally considered. Today, organizations are considered porous or open systems. The activities of organizations transcend the periphery of their immediate boundaries and affect the general society and natural environment.

Corporate activities have resulted in water pollution, waste generation, natural resource use and exploitation, harmful emissions, and accelerated levels of energy consumption. In response to the realization that corporate activities can be far reaching, organizations have been called on to consider their social and environmental impacts through corporate social and environmental responsibility programs. There is some regulation of corporate environmental impacts by governments; however, for the most part, corporate environmental responsibility is voluntary. The application of corporate environmental responsibility across organizations worldwide varies widely as it is a voluntary exercise. For example, in 2010, in the course of its oil drilling activities in the Gulf of Mexico, BP plc caused a catastrophic accident dubbed the Deepwater Horizon oil spill. This spill resulted in the death of 11 people and unprecedented destruction to marine and wildlife. It is widely held that BP took a lackluster approach to its environmental obligations and that this had caused the accident. Environmental groups attribute the spill to the absence of adequate regulation of the ecological impacts of industries by government. Corporate environmental responsibility is a phenomenon that has gained traction in recent years, especially in the face of climate change. This entry reviews the concept of corporate environmental responsibility by outlining its origins and how it is currently being practiced, and offers a brief discussion of its anticipated future.

Evolution of the concept of corporate environmental responsibility

In order to understand how corporate environmental responsibility has evolved, it is instructive to trace its antecedents. Corporate environmental responsibility is a subset of the broader term or activity corporate social responsibility (CSR), hence it evolved in tandem with the concept
CORPORATE ENVIRONMENTAL RESPONSIBILITY

and practice of CSR. Until recent years, CSR was the umbrella term for all social and environmental initiatives undertaken by organizations. Over time, it became necessary to distinguish between environmental and social responsibility as it was recognized that business activities have direct effects on the environment. Environmental advocates now demand that businesses be accountable to the environment in its own right as a stakeholder.

This demand that business take notice of the environment has been met with intense debate. Critics charge that operating with an environmental conscience runs counter to the very purpose of business: making a profit. Milton Friedman (1962) rejected the suggestion that businesses have any wider social and environmental responsibilities, and contended that the market should be the final arbiter of what they do. To focus on anything else, like the wider impacts of the business on the environment or society, would not only be inefficient and but also represent a gross misappropriation of shareholders’ investment.

At the time Friedman publicized his argument, the corporate social and environmental responsibility agenda was only beginning to take shape. Andrews (1973) describes how the issues of corporate responsibility had been gestating in the years prior to Friedman’s infamous declaration. The academic discourse prevailing at the time was that business was accountable to three stakeholders; namely, its investors, its employees, and its consumers. The emphasis here was on demonstrating a social conscience and not necessarily an environmental conscience. Moura-Leite and Padgett (2011) provide a useful account of the evolution of corporate social responsibility, drawing on the seminal works of Bowen (1953) and Drucker (1954). Bowen (1953, xi) is said to have spawned the modern period of CSR by asking the question “What responsibilities to society are businessmen reasonably being expected to assume?” Bowen’s text sought to spark debate by underscoring the importance of the several hundred largest businesses at the time to “centres of power and decision making,” furthermore highlighting how “the actions of these firms touched the lives of citizens at many points” (1954, 6, 4). The latter, explain Moura-Leite and Padgett, helped to advance the importance of including “public responsibility as one of the eight key areas in which business objectives should be set” (2011, 530). Drucker famously asserted that managers must “consider whether the action is likely to promote the public good, to advance the basic beliefs of our society, to contribute to its stability, strength, and harmony” (1954, 388). This foundational work provided an impetus for scholars to develop working definitions of CSR that aligned closely with the business ethics of the day.

Friedman’s (1962) critique of CSR was issued in the face of an emerging environmental movement that expanded the scope of corporate responsibilities. Rachel Carson’s book Silent Spring, written in 1963, captured the public conscience and helped to spark a radical overhaul of environmental regulations, as well as facilitating a change in the general attitude toward industrial pollution worldwide. Pressure was put on regulators by both the general public and environmentalists to restrain the environmental destruction wrought by businesses.

In response, many in the business world seized on the idea of corporate environmental responsibility as a way to earn their social license to operate. Critics, however, complained that there was no definition of the concept of social and environmental responsibility, making it easier for businesses to define the concept in line with their own purposes. CER was and continues to be an increasingly perplexing problem for corporate strategists to find an appropriate posture for what
may be called corporate citizenship, environmental responsibility, or social responsibility. Among the things that make it difficult are two prevalent myths: that the activities relating to environmental or social responsibility are in fundamental conflict with the interests of the profit-seeking investor; and that the motivation for such activities lies only in a sense of noblesse oblige on the part of the group formulating the strategy and, again, is not in the equity holder’s interest.

The criticisms against social and environmental responsibility continued into the late 1990s. The general sentiment among those against the CSR movement was that business executives were against the concept. One scholar asserted that “executive interest in corporate social responsibility is waning” when new standards for pollution control and safety came into effect (Shanklin 1976, 75). Toward the mid-1970s, however, a much broader body of analysis – one more positive in outlook – began to take shape. A large share of this literature popularized the merits of CSR (e.g., Gavin and Maynard 1975; Carroll 1974; 1978; 1991; Carroll and Beiler 1975). A series of complementary studies argued how, in addition to their profit-making motives, large businesses were beginning to show a willingness to address environmental and social concerns more proactively. It was at this time that organizations had begun to alter their structures and to appoint corporate social responsibility officers. These officers were responsible for managing community development and environmental concerns. The discussion at this point was showing how organizations were embracing mounting societal and environmental demands and how this new approach to business was in the best interest of organizations.

The conceptual work of Archie B. Carroll (1974; 1978; 1991) helped transform business attitudes toward community development and environmental management. In 1991 Carroll developed the “pyramid of corporate social responsibility,” a hierarchical model of the various business responsibilities, starting with economic responsibilities (the foundation of all other business activities), legal responsibilities, ethical responsibilities, and philanthropic responsibilities at the apex. Carroll’s pioneering model, which even today is still portrayed as landmark, spawned a new genre of CSR scholarship. The insights it offered continue to serve as the foundation of the case for CSR being practiced by organizations around the world.

Gradually, many organizations have begun to embrace the call to respond to the environmental agenda. Incidents such as the Exxon Valdez oil spill in 1989, the Bhopal disaster in 1984, and the more recent Fukushima nuclear disaster in 2011 have continued to raise awareness on the importance of environmental responsibility. Most of these disasters have resulted in huge sums of money being spent on litigation. Curiously, ethical businesses have been rewarded by increased consumer loyalty. Today’s consumers are prepared to shell out more for sustainable products. Buzz words such as “biodegradable,” “organic,” “carbon offsetting,” “FSC certified,” “eco-friendly,” and “renewable” attract consumer attention. People are “going green” – making environmentally friendly choices in the products they buy. Companies are therefore expected to do the same. Today’s organizations are not expected to exist solely for the motive of profit maximization but are also required to align with the social, environmental, and regulatory structures within the countries they invest in. Investments in CER programs are now perceived as a requirement for companies to
get the necessary buy-in from the communities within which they operate. That is part of what is termed the “social license to operate.”

Another condition that has contributed to the widespread investment in CER programs is the role that governments have played in encouraging businesses to minimize their ecological footprints over and above the requirements of the law. The activities of firms in the extractive industry, for instance, result in significant environmental impacts. One way to fend off regulation from governments is through CER reporting suggesting that governmental intervention is unnecessary, because business can be trusted to manage the environment responsibly. Extractive industry firms have been quick to highlight their numerous contributions to society through comprehensive sustainability reports. In the 1980s they were heralded as trailblazers for environmental and social responsibility programs in the developing world. In line with idea of the “triple bottom line” corporate reporting put forward by Elkington in 1994, these firms published and continue to publish extensive sustainability reports on their social and environmental efforts. Elkington (1994) urged organizations to consider their social, environmental, and financial performance and not just to focus on the traditional measure of financial performance. Elkington’s suggestions spawned a new wave of corporate reporting. The annual financial reports typically released by public companies were now accompanied by social and environmental reports.

Companies outside the extractive industries soon followed suit. Social and environmental reports, typically dubbed “sustainability reports,” are now published for stakeholders to see how companies are making a contribution to society and taking care of the natural environment. Several theories have been advanced to explain why companies publish such reports. It is argued that, fundamentally, companies publish these reports for the sake of legitimacy. Irrespective of the reasons given for corporate disclosure of environmental activities, the increased presence of these voluntary reports perhaps points to the fact that a lot of companies have embraced the call to take responsibility for their society and the environment seriously.

Visible negative effects of corporate activities have also increased the need for corporations to operate sustainably as the natural resource base of the Earth is finite. The past century has witnessed several corporate disasters, which have led to this call for action. The Exxon Valdez oil spill alone affected 1300 miles of shoreline. The carcasses of more than 35,000 birds and 1000 sea otters were found after the spill, but, since most carcasses sink, these are considered to be a small fraction of the actual death toll. The best estimates are 250,000 seabirds, 2800 sea otters, 300 harbor seals, 250 bald eagles, up to 22 killer whales, and billions of salmon and herring eggs.

Typical steps in corporate environmental responsibility planning

The concept of CER continues to be applied differently across industries and between organizations. As CER is voluntary, organizations are free to evaluate their internal and external processes in order to identify the perceived impacts on the environment. Generally, organizations take three steps to address their CER. Figure 1 depicts this dynamic process.

Step 1

The initial step is to identify the environmental impacts of all business activities. This includes reviewing the company’s supply chain, internal processes (administration and manufacturing,
Step 1

where applicable), and its service or product delivery outlets to identify the ecological impacts of such activities. Such assessment often also involves an extensive stakeholder analysis.

Step 2

The next step is to evaluate the extent of the damage to the natural environment. Within this step, organizations prioritize the environmental impacts they identify in step 1. At this stage, stakeholders are grouped in order of significance. As organizations perceive their primary stakeholders as the most important stakeholders, issues that affect these stakeholder groups are given priority. Cost–benefit analysis is often performed at this stage to determine how best to address these ecological issues. Organizations also consider the prevailing standards in their industries and align their choices with those of the industry. Table 1 shows a selection of international sustainability frameworks, guidelines, and “best practices” typically adhered to by organizations. The standards guide organizations in reporting their environmental and social efforts.

Step 3

After prioritization of the environmental impacts, companies take steps to mitigate their ecological impacts. Typically, responsibility plans include reducing energy consumption, recycling to reduce landfill disposal, reclaiming mined lands, reducing hazardous waste, reducing water usage, replanting trees, and using sustainable wood and other products.

It is now commonplace to find environmental sustainability links on corporate websites. Most organizations have a corporate environmental policy. Such policy documents are composed of environmental codes, details of environmental principles or standards that the organization subscribes to, and environmental standards and codes specific to that organization.

Future directions in CER

Despite gaining significant traction over the years, there is still considerable room for improvement in the area of corporate environmental responsibility. While companies operating in the developing world have managed to deflect negative attention by projecting positive images of “community development” and “environmental management” in localities affected by their operations, extensive corporate lobbying in the developed world has slowed down environmental regulation. There is widespread belief that corporations will continue to deflect attention through innovative reporting; extensive public engagement; the assembling of dynamic corporate relations departments; and the skillful manipulation of the local media. An
emphasize on outcomes has led to organizations adopting a narrow view when making environmental decisions, the effects of which can be detrimental when it comes to a disempowered, marginalized stakeholder such as the environment. Companies are using their CER programs as a tool for enhancing their legitimacy. This is especially true in areas where activism for the environment is strong. The accounting literature has discussed, extensively, the positive role that voluntary sustainability disclosure plays in an organization’s efforts at gaining or enhancing legitimacy. In some instances, it is argued that organizations with strong environmental and social programs use sustainability reports to claim superiority in the area of CSR or to counter a perceived legitimacy threat. CER was used to deflect criticisms and to regain legitimacy following the Exxon Valdez oil spill in 1989 and the BP oil spill in 2010. Academic scholars note that there were increased self-laudatory environmental disclosures, focusing heavily on preventive measures and emergency response, which, as demonstrated, were made purely in response to heightened public pressure.

If these trends continue, it is anticipated that CER will, for the most part, continue to be a facade, a mere tool that companies use in order to gain or enhance their legitimacy, to the detriment of society and, more importantly, future generations.

SEE ALSO: Conservation and capitalism; Ecological footprint; Environmental citizenship; Environmental (in)justice; Social capital; Sustainable development

References


### Further reading


Corporate financialization

Manuel B. Aalbers
KU Leuven/University of Leuven, Belgium

The rise of financialization

The concept of “financialization” has rapidly become popular in social science. On April 22, 2011, there were 1950 and 4680 hits, respectively, for “financialisation” and “financialization” on Google Scholar (Engelen 2012). Almost 40% of those were added between the beginning of 2009 and April 2011. Three years later, on April 22, 2014, there were, respectively, 5940 and 12600 hits for the UK and US spelling of the concept. This means that 64% was added in those 3 years – a true explosion of the concept of financialization. What is it that makes this concept apparently so attractive to academics, and what do they mean when speaking of “financialization”?

It would be hard to deny that the global economic crisis in 2007–2008 and its persistence in the following years explain part of the popularity of the use of the term “financialization.” The economic crisis is often framed as a financial crisis caused by unscrupulous financial practices in both the global financial command and control centers (London, New York, and so forth) and the daily life of consumer banks and their customers. There is a feeling that both the economy at large and daily life have become more financialized: that is, finance is thought to play a bigger part in both the Economy with a capital “E” and in the many economies with a minor “e.” The burgeoning literature on financialization tries to answer the who, what, how, why, when, and where questions of the presumed financialization of the E/economies. One might assume geographers would have prioritized the “where” question, but they have rarely studied the where in isolation of the other questions. Furthermore, political economists of different stripes and different disciplinary backgrounds have also included the where question in many of their analyses.

The literature on financialization is truly multidisciplinary, in the sense that most contributors to the financialization debate appear to rely on literature from different disciplines. The authors of the 19 most cited (i.e., at least 100 citations in Google Scholar) publications on financialization have backgrounds in economics, sociology, political science, cultural studies and arts, history, and geography, and most of these papers reach their high citation scores by receiving cross-disciplinary citations. Whatever may divide these approaches, there appears to be a shared conviction that mainstream, neoclassical theories provide little fertile ground to understand the contemporary financialized economy. Most of the economists active in financialization debates rely on Keynesian, Marxist, or, more generally speaking, heterodox and political economics. Many of the noneconomist protagonists rely on either some form of multidisciplinary political economy or on some combination of post-structuralist and cultural-economy accounts. Many of the financialization protagonists suggest or explicitly argue that a great deal of work within their discipline or subdiscipline for too long has either ignored finance or presented an outdated view on the role of the financial sector in contemporary capitalism.
CORPORATE FINANCIALIZATION

Financialization also has been criticized, either because the concept is considered imprecise, vague, and chaotic or because the presented evidence supporting the financialization claim is disputed. To some extent, the critics are right: financialization can be a very loosely defined concept that covers many processes, structures, practices, and outcomes at different scales and in different time frames. Furthermore, sometimes financialization is the *explanandum* (the phenomenon to be explained), sometimes the *explanans* (the thing that explains), and at other times it is not even clear which of the two it is. In that sense, financialization is not that different from other concepts whose academic (and media) popularity rose quickly and which are simultaneously criticized for being imprecise and vague – globalization and neoliberalism are cases in point.

The popularity of each of these concepts lies, at least in part, in their imprecision: that is, in their ability to transcend different lines of argument, originating from different disciplines, and taking place at different scales. It is the inability of existing perspectives, concepts, and data to deal with the complex realities of contemporary societies that explains an important part of the popularity of such imprecise concepts. Moreover, these concepts become popular so rapidly exactly because in the real world it may be hard to tell the *explanandum* from the *explanans*. Part of the intellectual journey of the use of concepts is that they problematize existing conceptualizations and understandings of what caused what. While this may initially create more confusion, it also reflects an, often implicit, acknowledgment that we do not live in a closed system in which causations are linear, one-dimensional, and single-scalar. The literature on financialization thus is part of a larger attempt to understand the nonlinear, multidimensional, multiscalar complexity of contemporary societies/economies. This entry, by adding a little to the complexity, seeks to shed light on the different elements of the financialization literature and their interrelations.

**Defining financialization**

The financialization literature is commonly divided into three different conceptualizations: financialization as a regime of accumulation, financialization as the rise of shareholder value, and the financialization of daily life. This division has become problematic: a great deal of the literature makes connections between these strands or moves outside the arguments presented within them. Thus the following 10 themes are proposed as encompassing contemporary scholarship on financialization:

1. financialization as a historically recurring process that signals the autumn of hegemonic powers
2. financialization of banking: that is, the rise of nonbank financial institutions
3. financialization of the economy in narrow terms: that is, the financial sector becoming increasingly dominant in economic terms
4. financialization of nonfinancial firms: that is, traditionally nonfinancial firms becoming dominated by financial narratives, practices, and measurements
5. financialization *within* nonfinancial firms: that is, traditionally nonfinancial firms increasingly partaking in practices that have been the domain of the financial sector
6. financialization of the workplace: that is, employees and their labor practices increasingly shaped by financial narratives, practices, and measurements
7 financialization of the (semi-) public sector: that is, government, public authorities, education, health care, social housing, and a range of other sectors becoming dominated by financial narratives, practices, and measurements

8 financialization of public policy: that is, the financial industry’s concerns becoming increasingly privileged in the policy domain

9 financialization of households: that is, financial motives, rationales, and measures becoming increasingly dominant, both in the way individuals and households are being evaluated and approached, and in how they come to make decisions in life

10 financialization of the discourse: that is, finance becoming increasingly dominant as a narrative and metaphor, as a language to see/view/measure/assess/evaluate all things economic and noneconomic.

Here financialization is not separated out as a regime of accumulation, although it shares a focus with the first and third elements. A regime of accumulation is more than the sum of different elements: understanding it means focusing on both macro and material aspects as well as meso, micro, and discursive aspects. The seventh and eighth elements separate two quite different processes underlying the financialization of the state (which also includes the first, third, and tenth elements). Indeed, it is impossible to think of financialization as an accumulation regime without considering the role of the state in its different constitutive elements. The ninth and tenth elements are both about the “financialization of daily life,” but it is important to realize that the financialization of the discourse plays an important part in all the other elements distinguished here. A definition of financialization encompassing these dimensions would be: the increasing dominance of financial actors, markets, practices, measurements, and narratives, at various scales, resulting in a structural transformation of economies, firms (including financial institutions), states, and households.

This entry on “corporate financialization” focuses on the first eight elements (for the financialization of households, see inter alia Martin 2002; Aalbers 2008; Langley 2008). While not ignoring financialized discourses, these will not be discussed separately from the other elements. Of course, all these elements are related, which is why they are not discussed as different conceptualizations of financialization (although that would arguably result in a less vague and more precise definition) but rather as different dimensions of a complex phenomenon that can only be understood – whether as explanans, explanandum, or purely as discourse – by a strong awareness of their interdependence. For analytical as well as practical reasons, however, it may be necessary to study some elements in relative isolation; holistic understandings do not always fit together very well with empirical research projects. In what follows, these elements will be discussed one by one.

Financialization as a historically recurring process

The “-ation” part of financialization suggests that it is not a state or end result but an action, something that is made. Many financialization scholars situate the beginning of financialization in the 1970s with the rise of neoliberalism, the industrial crisis in the West, the breakdown of the Bretton Woods system, and other developments. Others have pointed at financial deregulation and the associated changes on Wall Street and the City of London in the 1980s, including technological developments and the rising influence of
CORPORATE FINANCIALIZATION

pension funds and other institutional investors. The decline of communism and the fall of the USSR at the end of that decade are also mentioned as contributing factors, in part because they discredited noncapitalist alternatives and underwrote how neoliberal and financial discourses became hegemonic. Structural changes in welfare states are also seen as crucial, although it is not always clear to what extent welfare state changes drive financialization and to what extent financialization drives welfare state changes. More generally speaking, financialization is part of and key to structural transformations of advanced capitalist economies. According to some scholars, we have been here before, and financialization thus should be understood as a recurrent phase in capitalist development.

When studying the historical trajectory of financialization, the work of Giovanni Arrighi (1994) has been very influential. Building on seminal contributions by Braudel and Wallerstein, Arrighi argues that hegemonic capitalist powers in the autumn of their hegemony can be characterized by a phase of financial expansion. Earlier hegemonic capitalist powers – Genoa, Holland, England – had already started their decline when their economies became financialized internally but also financially hegemonic externally. Arrighi sees the financial expansionary phase as a response to overaccumulation. Capital is switched to the financial sector to avert a crisis, but the real economy and hegemonic power nevertheless decline. Yet, financial power typically remains with the declining hegemon while the next hegemonic power is in its initial stages. Typically, the old hegemon finances the new one, not because they want to hasten their own decline, but because they see this as the best way to keep on realizing return on investment. The “autumn hegemon” produces a rentier class that comes to dominate the real economy but continues to produce profits that can extract financial rent. Typically this financialized stage of the declining hegemonic power benefits fewer citizens than in the prefinancialized state. This tends to undermine middle-class consent, and social polarization and inequality increase.

The historical parallels appear clear and partly justify the focus of financialization scholars on the United States, which can be considered to be in the autumn of its hegemony, having entered the financial expansionary phase and the associated decline in middle-class consent and rise in income inequality and social polarization. From this perspective, other OECD (Organisation for Economic Co-operation and Development) countries can be seen as under American hegemony, with the United Kingdom as the United States’s closest ally. That does not explain, however, how and why the City of London in some ways has become more important than – or at least as important as – the main financial center of the “autumn hegemon,” Wall Street. It also does not explain how US debt is, in fact, increasingly financed and backed by foreign, particularly Chinese, capital. This could be a new development in the historical trajectory of declining hegemons, which in an era of globalized finance no longer need their own excess capital to enable financial expansion. It could be equally a sign of financial power and of the real decline of the hegemon. Arrighi’s analysis suggests neither the one nor the other: the fall of world hegemonic powers is at least as slow as their rise, and the seemingly paradoxical combination of increasing financial power and decreasing political-economic power more generally is just another fundamental contradiction of capitalism. For our purposes, it is important to keep in mind that the different elements of financialization may often be more advanced and visible in the United States than elsewhere but that financialization is increasingly visible in
other Western and non-Western countries. This does not necessarily mean that financialization is as advanced elsewhere, but that the trend in different places and at different scales often goes in the same direction.

Financialization of banking

It appears counterintuitive to separate banking from finance. But finance is not just the business of banks. First, the role of banks has shifted from more or less passive intermediaries towards active financial actors. Second, traditional banking has become less important vis-à-vis other financial actors and activities with the explosion of nonbanking financial institutions, ranging from pension funds and mortgage companies (so-called “nonbank lenders”) to private equity and hedge funds. As finance has moved beyond its traditional intermediating functions, it has come to be seen as a growth industry in its own right.

In the popular discourse, banks make money through the difference between interest charged on loans and interest given on savings, but the real moneymakers for most financial institutions, including banks, are leveraging and charging fees. This may have happened to varying degrees in different countries, but no Western country on either side of the Atlantic and ever fewer non-Western countries have escaped the trend towards transaction- and leverage-based banking business models – some speak of a financial services revolution (Moran 1990). For many banks, issuing loans is primarily interesting because they can be repackaged into new financial products for which fees can be charged. Furthermore, once the loans are repackaged and sold, the money can be reinvested in other financial products. Thanks to the financial leveraging powers of banks, this pumping around of money, mostly between financial institutions, could continue unprecedented for some years. The crisis that started in 2007–2008 slowed down the leveraging and debt machine, but has not stopped it.

Not only do nonfinancial firms rely less on banks for their finances, but banks also invest less in the so-called real economy and increasingly put their money in financial assets: “During the 2000s, lending for finance, real estate and household purposes replaced ‘productive’ lending as the driving force in the loan portfolio of banks” (Lapavitas and Powell 2013, 371). This appears to be an international trend, although there is quite some variation in who receives most of the loans (e.g., real estate firms, homeowners, financial intermediaries). All in all, we can speak of a debt explosion, not just in volume but also in the geographical scope of debtor–creditor relations.

Financialization of the economy in narrow terms

Financialization is, among other things, a pattern of accumulation in which profit-making occurs increasingly through financial channels rather than through trade and commodity production (Krippner 2011). Due to the slowing down of the overall growth rate and the stagnation of the real economy, capitalism has become increasingly dependent on the growth of finance to enlarge money capital (Sweezy 1995). Therefore the capital accumulation process becomes financialized, focused on the growth of finance to benefit actors within financial markets, such as investors, rather than benefitting the real economy. Furthermore, some commentators argue that financial investment is replacing physical investment.

To illustrate this argument, different authors cite different statistics to show that a whole range of financial markets have grown rapidly since the 1970s. The market for derivatives, in particular,
CORPORATE FINANCIALIZATION

has virtually exploded between 1990, when
the market was almost too small to measure,
and 2006, when the number of outstanding
contracts added up to $370 trillion, as the Bank
for International Settlement has demonstrated
(BIS 2008). Krippner (2011) demonstrates that
finance has become the dominant source of
profits since the 1990s, a trend that may be par-
ticularly pronounced in the United States but can
be witnessed in most OECD countries. For the
27 member states of the European Union (i.e.,
before the accession of Croatia), EUROSTAT
has calculated that the FIRE (finance, insurance,
and real estate) sectors together contributed to
29% of gross domestic product (GDP). Even in
Germany, which is said to have a less financialized
economy, the FIRE sectors contributed more to
GDP than industry (respectively, 30% and 20%).
Financial assets held by institutional investors
as a percentage of GDP grew rapidly in all
OECD countries and now represent more than
200% in countries like the United States and the
United Kingdom and around 100% in countries
like Germany and France, increasing threefold
(United States) to tenfold (France) between 1980
and 2001 (Deutschmann 2011). By contrast, the
wage share of national income has fallen across
the board, although less so in countries with
strong labor unions (Epstein and Jayadev 2005).

Financialization of nonfinancial firms

Nonfinancial firms have always been dependent
on credit, but the rules and logics of Wall Street
are increasingly becoming the rules and logics
outside Wall Street. The corporate narrative has
also become financialized. The idea of share-
holder value has become dominant in how firms
“ought” to be run, and senior managers have
become responsive to such demands: “Managers
were no longer considered as skilled professionals
but as agents of shareholder value maximization”
(Deutschmann 2011, 358). Financial numbers
had to be framed to make them appear promis-
ing. Many senior managers became busier with
communicating positive stories to convince
credit rating agencies, market watchers, and
stockholders than with innovation or production
gains (Froud et al. 2006). The increasing finan-
cialization of nonfinancial firms has been noted
for a wide range of industries: for example, the
car industry, the so-called new economy, real
estate, retailing, computing, pharmaceuticals,
insurance, aviation, and infrastructure.

An important driver of the financialization
of nonfinancial firms has been the ownership
of publicly traded firms. In the 1950s, US
households held approximately 90% of corpo-
rate stocks. Fifty years later their share was just
42%, whereas the share of institutional investors,
including pension funds, had increased to 46%
(Crotty 2005). Furthermore, since the 1980s
nonfinancial firms are increasingly led by CEOs
with a financial or legal background (Fligstein
1990). The ideology or myth of shareholder
value is prioritized in leveraged buyouts, stock
repurchases, mergers, and acquisitions over
long-term profitability or firm survival. Many
financialized firms seem able to prop up their
stock prices or impress the rating agencies for
some time, but the effective return on capital
rarely goes up structurally and appears more
vulnerable to both conjunctural and structural
shifts in the industry.

One reason that it is important to consider the
financialization of nonfinancial firms, beyond
the creation of shareholder value, is that many
companies are not publicly listed and traded, but
are still financializing. Financialization changes
the way money is made in many industries: there
generally is a narrow focus on outsourc-
ing and short-term profits at the expense of
integrated development, long-term investment,
and nonfinancial innovation. As a result, nonfinancial firms have increased financial flows to the financial sector through interest payments, dividends payouts, and share buybacks (Lazonick and O’Sullivan 2000; Crotty 2005). The financial market’s and financialized management’s response to unsuccessful cases of financialization is typically more, not less, financialization.

Financialization within nonfinancial firms

If profits are the bottom line, firm management may be expected to engage in activities that generate the highest profits. As the profit rates in the financial industry for some time were higher than in most of the so-called real economy, some nonfinancial firms became mixed nonfinancial/financial firms. Derivatives, in particular, proved hard to resist for many formerly nonfinancial firms. As a result, nonfinancial corporations increasingly derive profits from financial activities and own a greater proportion of financial relative to nonfinancial assets (Krippner 2011; Lapavitsas and Powell 2013).

Optimists say that if a nonfinancial firm realizes high profits through investments in, say, derivatives, this results not in “crowding out” real investment but in additional funds becoming available to investments in the nonfinancial parts of the firm. Critics, however, argue that this overlooks the fact that once a firm is investing in financial assets, it will most likely use profits to expand such activities – that is, if such financial investments create higher profits than nonfinancial investments, they will be tempted to shift more money towards investments that deliver higher profits. The bigger problem of the optimistic argument is that financial investments tend to be quite volatile and may jeopardize the survival of the firm, or at least its nonfinancial activities.

An important consequence of this element of financialization is that statistics of the “real economy” versus the “financial sector” become blurred. Measuring financialization as the increasing dominance of the financial sector in GDP statistics and financial firms’ profits misses an important dimension of financialization. There clearly is more room for studies that investigate how traditionally nonfinancial firms increasingly partake of practices that used to be the domain of the financial sector. It is crucial that such studies include the practices of publicly traded companies as well as nonlisted firms. Research methodologies focusing explicitly on individual firms, such as developed by Julie Froud and colleagues (2006), are important in this respect.

Financialization of the workplace

Financialization does not only change the senior management of the firm through the ideology of shareholder value, but also impacts employees throughout the firm, from CEOs to partners and senior and junior employees. Financialization can be an employee control strategy. Faulconbridge and Muzio (2009) and Alvehus and Spicer (2012) demonstrate this in law and accountancy firms, respectively, including effects on recruitment, human resource management, management by objectives, career systems, and client control. Financialization aims to transform the working lives of employees into an investment activity in its own right, using billable hours as both a measure of profitability and investment in future higher pay, and possibly entry to the firm’s partnership. Working lives are transformed in this financialized system of controlling and steering “human capital” and come to be defined in monetary terms and discussed in terms of investment, trade, speculation, and leverage.
Employees, in return, play the system, not only by increasing the number of billable hours but also by offloading work downstream (Alvehus and Spicer 2012). Financialization puts pressure on workers’ wages, the amount of hours they work, and the rights they have or can exercise (Lazonick and O’Sullivan 2000).

One does not have to be a skeptic to see parallels here to the ways academic work has become not only commodified but also financialized, with publications in indexed journals replacing billable hours as financialized time and academics investing their time to leverage them for tenure, promotion, and grants. Although most academics are still interested in knowledge per se, it would be hard to dismiss the evidence pointing to a gaming of the system not unlike that in law and accountancy firms. The transformation of the working lives of people in law, accountancy, academia, and elsewhere is uneven but rarely entirely voluntary, as the principle of “up or out” (you need to achieve a certain rank within a fixed period of time) can apply in these fields, forcing employees to play along to some extent.

Financialization of the (semi–) public sector

States and semipublic industries are increasingly dependent on financial markets and are also evaluated in similar ways to firms. Rating agencies provide scorecards for governments, not only national governments but also local ones. States are not only evaluated like companies; with the popularity of New Public Management, both public and semipublic institutions also became managed more akin to private firms than at any time in the past. One important consequence is a redefinition of citizens into consumers and a further redefinition of consumers as financial assets or cash cows (Allen and Pryke 2013).

The spread of New Public Management and of the domination of financial narratives, practices, and measurements is not limited to government institutions, but is also apparent in the working of other public authorities as well as semipublic and commodified sectors such as education, health care, and social housing. Both Engelen, Fernandez, and Hendrikse (2014) and Beverungen, Hoedemaekers, and Veldman (2012), for example, argue that universities have become increasingly financialized. Academic management is controlled by, and controlling, employees though financial metrics, measurements, and increasingly also narratives. This is not only visible in university annual reports but also in the expectations it has of its employees.

The financialization of universities and other semipublic sectors is not limited to a combination of the above-mentioned elements of financialization – that is, of nonfinancial firms and of the workplace – but it is also an example of financialization within traditionally nonfinancial companies, as universities become involved in deriving profits from investments in derivatives. The consequences are mostly negative: “less professional autonomy, more administrative chores, more overhead, more standardization, higher throughput and less academic exchange” (Engelen, Fernandez, and Hendrikse 2014, 16). Indeed, “by extending its leverage and balance sheet, [it] is in danger of strangulation by debt, risking the funding streams to the activities for which it was established: teaching and research” (Engelen, Fernandez, and Hendrikse 2014, 16).

Financialization of public policy

In contemporary capitalism, we can witness a “great risk shift,” in which households can rely less on public institutions for their long-term security and become increasingly dependent
CORPORATE FINANCIALIZATION

on private firms (Hacker 2008): in particular, financial institutions. The state actively promote this movement away from the state and into financial markets. In fact, a combination of commodification/privatization and de/re-regulation has created such markets, sometimes out of thin air. The state is no bystander in the financialization of the economy, firms, households, and of the state itself. It has actively promoted financialization, although rarely in a linear and one-dimensional way – state institutions at different scales and with different responsibilities have often acted in diametrically opposing ways.

The global financial crisis that has been dragging on since 2007 made this painstakingly clear. In this crisis, “the financial industry has managed to externalize its own problem and to transform it into a problem of the state” (Deutschmann 2011, 384). States became “victims of the transformation they helped to bring about, and have been forced to bail out their debt-encumbered banks and financial systems … the state role of ‘risk absorber’ is expanded for the private market sector rather than for the citizenry” (Christopherson, Martin, and Pollard 2013, 352). The public management of the financial crisis can be seen as an example of “socialism for the rich and capitalism for the poor,” in which the state protects the interests of the financial elite at the expense of most of its citizens (Aalbers 2013).

Lobbying plays a significant role in this financialization. A study by Corporate Europe Observatory, together with the Austrian Federal Chamber of Labour and the Austrian Trade Union Federation, suggests that the financial industry spends at least 120 million euros a year on lobbying the European Committee through more than 700 organizations. They are estimated to outspend trade unions and civil society by a factor of more than 30. Furthermore, of the 17 European Union (EU) official advisory councils that the researchers investigated, 15 were dominated by the financial industry. Although it is hard to measure the success of 600 million euros of “investment” in 5 years’ time, it would be hard to imagine that this would not bring the industry at least 600 million euros in beneficial regulation. In other words, finance-to-EU lobbying is likely to have provided the industry with much more than an additional 600 million euros in savings. Indeed, in a financialized political environment it will be difficult (but not necessarily impossible) to get anything done that runs counter to the expected benefits of the most powerful group.

Geography and corporate financialization

Geographers have repeatedly stressed that financialization is an inherently spatial phenomenon that should be much more central to economic geographic analysis. Local, national, and macroregional institutions act as filters of how financialization plays out and is perceived. Often, financialization is not much limited by existing institutions, but these institutions are mobilized and transformed to enable financialization. This is why the structural transformation of the state is intrinsically linked to the process of financialization: the state filters financialization processes while financialization at the same time furthers the transformation of the state itself. Human geographers have contributed to the idea that there are not only varieties of capitalism (VoC) but also varieties of financialized capitalism (VoFC) that do not entirely flow from the expectations of the VoC framework of liberal market economies versus coordinated market economies. The embeddedness of national political economies in global capital markets is
not limited to liberal market economies, as one would expect based on a reading of the VoC literature. Small, open economies like those of the Netherlands, Belgium, and Ireland appear both globalized and financialized (Engelen 2012).

A singular focus on national economies, however, stifles a full understanding of financialization, not because the national scale is irrelevant but because it is only one of the many relevant scales, including but not limited to the global, macroregional, and metropolitan scales. The financialized economy is perhaps not concentrated in global cities (although the financial industry is); yet, these cities are command and control centers for both the financial industry and the globalized, financialized economy more generally. Furthermore, national statistics can both over- and underestimate levels of globalization and financialization. Thus the Netherlands appears to be one of the major investors in many countries, but such statistics largely reflect the attractiveness of the Netherlands as a tax shelter rather than real or Dutch globalized investment: money flows through rather than from the Netherlands.

Financialization, globalization, and neoliberalism are interdependent. Offshoring, for example, whether financial or nonfinancial in nature, may be motivated by financialization but its effect is economic globalization. Furthermore, both globalization and financialization are often promoted and furthered through a neoliberal agenda, sometimes through false pretenses of leveling the playing field while in fact redrawing the field in favor of corporate and financial elites, and their shareholders. Duménil and Lévy (2004) suggest financialization causes neoliberalization, but it is hard to disentangle the causal relationships for two concepts that are both so widely defined and in fact part and parcel of each other.

Although the literature on financialization has exploded, many avenues for research remain underexplored. For financialization to be taken seriously, not only by academics but also by policymakers and the public at large, financialization needs to be measured more rigorously in different countries and in internationally comparative ways. It is important to focus on all dimensions of financialization. It is justified for studies to isolate one aspect from the others to make empirical investigation possible in the first place, but this should not take away from the inherent complexity of financialization, nor should it be limited to the aspects of financialization that can be researched relatively easily through readily available statistics. Qualitative and discursive analyses, including corporate case studies, are equally important. Furthermore, geographers and other social scientists should not ignore the study of spaces of financial exclusion, expropriation, and exploitation. Finally, the financialization of the state is also an important avenue for future research. Research into the financialization of public and semipublic institutions is still in its infancy, while the interlinkages between finance and power have also been too much assumed rather than put to empirical scrutiny.

SEE ALSO: Corporations and the nation-state; Economic geography; Financial geography; Firms; Globalization; Labor geographies and the corporation; Mainstream and shadow banking; Neoliberalism; Regulation; World cities

References

Corporate identity

Nick Clifton  
Cardiff Metropolitan University, UK

Stefan Gärtner  
Institute for Work and Technology in Gelsenkirchen, Germany

Corporate identity – meaning and concept

The term corporate identity (CI) is most simplistically a combination of the words identity and corporation. That identity is closely related to geography or space has been described often in academic debates (Lynch 1960). That corporations are strongly linked to a location and that corporations are always part of the local economy and, therefore, important determining factors for regional development is obvious, but the linkage between the concept of “corporate identity” and geography is less clear. To understand why it is worth applying the concept to place and geography it is necessary to examine the concept of CI.

CI is the strategic concept of how to position a corporation. In doing so, it is important to look for the defining traits of identity and integrate them into a congruent action concept. As shown in Figure 1, CI consists of corporate behavior (e.g., values, norms), corporate design (e.g., logos, corporate colors, architecture) and the corporate communication strategy (e.g., public relations, information). Taken together, this constructs the corporate image as the set of means and narratives by which a corporation is known, and through which people tell stories about it.

CI has been a topic of interest since the period of mass industrialization of the early twentieth century. New forms of corporations – like publicly listed commercial companies – were established at this time and brand identities started to become relevant. Since the 1960s and 1970s CI has become an issue for nearly all corporations, becoming ever more complex since the 1980s, as all corporate activities have come into sharper focus and increasingly strategic. Various features of identity, philosophy, culture, behavior, communication, and appearance have been brought together into one strategic approach. This increasing significance of CI has gone hand in hand with the dissolution of common social values. With globalization, identities have become increasingly hybrid, liquid, or overlapping.

Even if corporate identity is linked to an overall image of a corporation in the minds of customers, investors, and employees (or of a city or region, in the minds of its residents), meaning that communication is important, CI is more than communication. It should be based on realistic positions and visions that must be transformed into reality (corporate behavior) before a communication strategy (corporate communication and design) can be applied. Identity, in this context, is to be seen from an individual as well as from a collective point of view.

Bringing together corporate identity and place is relevant because cities, regions or national states, and corporations benefit from each other’s identity and image in a reflexive process. In recent years, many stakeholders studying the
influence of regions on corporations, and vice versa, have turned their attention to culture and identity. Country image identifiers and product nationalities are growing in importance as competition and standardization in globalizing markets reduce uniqueness and stimulate the demand for authenticity and provenance. Relatedly there has been growing discussion regarding what some authors term the “experience economy” – shorthand for the many intangible qualities surrounding consumption decisions. As such there is an iterative movement back and forth between how people act (both as consumers and as social citizens) and the representation of a place.

Cities and regions compete to attract capital and creative people by using certain images. Due to a high level of communication, interaction, and mutual reference, common regional identities emerge, which then can differ from those of other cities and regions. For cities with a long tradition of being recognized as a place of knowledge and knowledge-intensive services (e.g., London, Paris, or Hong Kong) it is easier to distinguish themselves via a corresponding image. This is much more difficult for cities and regions which are affected by structural change or are too peripheral to be well known.

**Place identity, or “Who are we?”**

Since the 1980s, the “world city hypothesis” (Friedmann and Goetz 1982) has addressed the extraordinary importance of certain cities, those that are highly integrated and create a common hegemonic space. Proponents of this thesis claim that transnational corporations compete for power and control of global economic space and its urban nodes. World cities remain important but, nowadays, smaller cities and regions also are searching for their own specific identities and unique selling points.

It is not possible to create an image out of thin air but, generally, it is vital that an image is rooted and seen to be authentic within the region. It is possible to create radical departures from tradition but they must be consistent and actually lived. Although lighthouse projects have been widely criticized, cities like Bilbao with its Guggenheim Museum, or Dubai – where it seems possible to construct everything from mega buildings to ski parks and whole islands – have succeeded in creating an identity and image with an impact both internally and externally. Especially for cities in Western pluralistic societies, where the government is based on participation and a focus of quality of life is important, it is not enough to create an image without considering the actual living and working conditions and the involvement of civil society. Today, quality of life, sustainability or new urbanism seem to be keywords in this regard. This value change is caused by different aspects, including an increasing awareness of ecological problems and a growing hedonistic orientation, partly expressed in an increasing demand for
authentic emotional products. This could be connected with a backlash against increasing, disorienting globalization, to be compensated by collective identifications through regional, religious or ethnic affiliations. Expressive aesthetic manifestations, indicated by growing numbers of museums and accentuation through architecture, are also increasing importance. However, this not only symbolizes an aesthetic turn in public spaces but also stands for the need to profile one’s own place in the face of global competition.

Dealing with places’ identity means trying to answer the question of “Who are we?” There is no general answer to this question: cities and regions are subjective and intangible, and, unlike corporations, are always contradictory and paradoxical. A helpful starting point, however, could be trying to identify the relevant subject and target groups (who belongs to “us”? ) and what is the relevant space or area for each topic.

Who belongs to us?

In regional science it is increasingly recognized that regional success does not only depend on companies; a triple helix network of business, science, and politics is central (Etzkowitz and Leydesdorff 2000). More recently, taking into account how new forms of knowledge sharing, decision-making, and participation have become powerful, civil society also has come to be considered a success factor – resulting in a quadruple helix (Carayannis and Campbell 2012). Yet, while the innovative societal power of regional bottom-up approaches cannot be ignored, civil society is not always active in contributing to a viable city life. Therefore, the analytical concept for exploring existing regional identity must be distinguished from the political agenda of making civil society part of the city image (which might be considered wishful thinking). Political agendas, urban life, protest culture, and participation have to be fostered, and political spaces and places have to be provided for these processes.

Social movements and civil society initiatives often have local origins, but some spill over between countries (e.g., from Tunisia to Egypt) and across organizations (e.g., the Spanish Occupy movement was inspired by the Arab Spring movement). This new spatial form also is a chance for cities to enhance their global connectivity. And, at the same time, this raises the issue up of how much political space and place should be given to which groups. Different target groups require different aspects and the question arises as to what extent values and norms like sense of responsibility, (family) tradition, perseverance, and pragmatism play a role in a process of identification for a region or city.

As cities always have multiple identities, the right balance instead of the dichotomous choice has to be found. London, for example, symbolizes the capital of finance for a certain target group but the capital of punk rock for others – both valid identities – for corporations, and for other actors, and thus worth communicating.

What is the relevant space?

Regional networks, social capital, environmental factors, and cultural issues and identities can all have locational potential, capable of shaping a region’s or city’s strategic position, and should be a starting point for an independent and distinctive strategy, whereas the reference areas typically go beyond strict administrative boundaries and locate themselves only partly on a local level.

The concept of space has been totally rethought and detached from geography by some researchers in recent years. In relational understandings of space, space is constructed
through processes of cognition, imagination, and retrospection. Space only exists by anthropological acting, that is, people have to interact in order to create space. Many contemporary researchers dealing with questions about place and space agree that the bordering container space has to be redesigned and regions have to be understood as looser subspaces or fields, relevant to specific target groups (economic, intellectual, artistic, etc.). Each such space has its own logic and development dynamic, but also interacts with the others. Understanding a place’s identity thus requires multilevel perspectives. Consider, for example, the three perspectives illustrated here.

First, a global scale perspective means being aware that proximal spaces within a city matter, while recognizing that it also is possible for spaces to develop and connect globally. Knowledge spaces can be developed internationally because regional knowledge networks sometimes include international companies or their affiliates. Economic development should take into account the most important supra-regional spatial contexts, connected to regional fields of knowledge and supporting the local and international integration. Indeed, globalization encourages thinking beyond space as a single region or national state, both in academic circles and the media. For example, the *Times Magazine* highlighted “Nylonkong” as a new space, consisting of the three cities of New York, London, and Hong Kong. “Linked by a shared economic culture (...), connected by long-haul jets and fibre-optic cable, and spaced neatly around the globe, the three cities have (by accident – nobody planned this) created a financial network that has been able to lubricate the global economy (...)” (Elliot 2008, 31). Nylonkong illustrates the idea of relational space: it cannot exist without traders, compatible business cultures and identities, and interconnecting transactions.

Second, a regional-scale perspective means developing awareness that the spaces where compact knowledge potential lies, knowledge spillovers occur, identities are developed, and images are shaped need not be limited to a particular administrative region. It is not enough, as was done with respect to defining some metropolitan areas, to stretch the boundaries. It is less the size of an area that ultimately matters but more its mix of spaces; having the right balance of areas for each specific activity.

For world cities or global urban agglomerations with a good reputation and a long-standing tradition, the surrounding space is not so important to their identity. It is different, however, for cities and regions with “less reputation” – whether peripheral or urban, prosperous or structurally weak, whether world-market- or locally oriented. Successful positioning within international competition requires different strategies, which will depend on the type of city. Cities lying within the hinterland of attractive metropolitan areas, for example, characterized by a narrow economic structural link with the metropolis, face three development options.

1 Occupying a complementary position in relation to the metropolis. This entails a risk of high dependency, as well as uncertainties stemming from possible competition with other small towns in the vicinity of this metropolis.

2 Sharpening the city’s profile to compete with the metropolis, promoting the advantages of smallness while in reach of the metropolis (lower land costs, better living conditions, etc.). The danger here is that the city’s individual profile is not distinct and thus would not stand out.

3 Positioning the city as a “dormitory town.” This can be an attractive economic alternative, extending the function of the city from
a pure “sleep function” towards meeting demand for leisure activities and recreation.

These three perspectives are ideal types that cannot be implemented as urban policy on a one-to-one basis. The strategy of a city (from corporate behavior over design to communication) and its positioning with respect to city types will vary greatly.

Third, from a local scale perspective it is necessary to be aware that certain districts affect different identities that depend only partly on the general orientation of a city. For certain districts, different cities can have more in common and carry out more transactions – due to similarities – than do single districts within each city. For example, some economic sectors compete between the districts of the City of London and Wall Street (New York) rather than between London and New York; or, to take another example in Germany, between Friedrichshain-Kreuzberg (Berlin) and St. Pauli-Karolinenviertel (Hamburg). The creative economy is dependent on fast changing trends, so its actors seek to be close to the source of new trends, namely the subcultures that typically arise in certain metropolitan districts.

With increasing segregation, neighborhoods have become increasingly homogenous, each accommodating citizens seeking the same kind of living conditions, activities that fit with the cultural infrastructure, and specific consumption bundles. As stereotypes like rich and poor or working class districts have become apparent, personal identities are more closely linked to those of the district. Districts can have a bad image, of course, such as old-industrial urban districts with historical layouts but bad housing conditions, or post-World War II districts with a large amount of social housing, concrete tower blocks, and more or less monoculture structures. For a strategy of regeneration, it is not enough to campaign for a better image; it is necessary to change the conditions and break the monoculture. This means attracting new users to the space: residents, enterprises, shops, restaurants, and so on. Economically less successful areas and deprived neighborhoods also offer opportunities for new ideas and unconventional, innovative business practices. Urban planners widely believe that deprived urban places with vacant buildings and tolerance for “otherness” can be significant hotspots for the creative economy (Florida 2002). Thus, cities have set up local business development agencies in deprived districts to exploit this potential. Transforming the identity of such districts, converting formerly poorly reputed districts into, creative, viable, and young urban quarters, can benefit the city’s image also. On a more local level, neighborhoods position themselves by focusing on certain lifestyles and targeting specific groups. Changing identities means not only applying the quadruple helix but also developing a broader understanding of the relevant actors; this should involve such actors as social workers, churches, sport associations, and so on.

Reconnecting: place identity shaping corporate identity

The interrelation of places’ entities includes corporations, competences, values, norms, brands and products, profile identities, and images. This is a complex process and it is nearly impossible to detect a general mechanism of action. Here, the discussion is restricted to the reflexive interrelations between companies, brands, products, and regions. Even if identities are neither static nor homogeneous, cities and regions possess some kind of shared ideas and values about how companies operate, and the kinds of products produced and their quality. On this basis, specific
outside perceptions can help establish brands. In the following this is illustrated first for branding and brands; how corporate identities and reputations, and those of regional networks, are based on a regional or city identity, as well as vice versa, are then examined.

Products, brands, and places

Nations and regions have images, or “brands,” that stakeholders often attempt to manage or at least influence. In turn, many products have an association, sometimes positive and sometimes negative, with a region or place. These associations interact and shape each other. Country-of-origin (COO) research has tended to focus upon how geographical associations may assist the marketing of certain products, with the role of symbolic value and identity being important.

The role of such symbolic meanings in the consumption of goods and services was semi-nally described by Veblen (1949). This is now more acute than ever, given the need to differentiate one’s product from the many others in the marketplace with similar functionality or technical specification. This requires reference to immaterial qualities or symbolic value. The race for innovation that has defined the post-Fordist economy has given us high-quality products. Increasingly, however, their functionality is taken for granted; consumer choices are determined by the symbolic value of the product or the company behind it. The strength of a brand thus stems from a combination of how it performs and what it stands for.

Successful global brands typically originate from a place with its own brand image, with the product often strongly linked with that image – whether it is an Italian sports car, associated with the qualities of style, speed, and innovative design, a chic and classy French perfume, or Japanese consumer electronics that associate implicitly with the merits of high-tech expertise and affordable quality. Goods produced within any given region are imbued with an associated symbolic value – whether weak or strong, positive or negative (Anholt 2003) – associations that typically are activated implicitly in consumers’ minds by the mere presence of country-of-origin (COO) information. The basic idea that a product’s COO can consistently influence consumers’ judgment of it is simplistic; a more nuanced understanding of the concept is required in the age of globalized supply chains, complex corporate ownership structures, and the international division of labor. This requires explicitly recognizing that a single product is designed, assembled, manufactured, and so on increasingly across in different places. Indeed, there is evidence that COO might often be incorrectly identified by potential consumers.

Firms and specific territories

The French term “terroir,” without a direct English translation, essentially refers to being from a defined geographical area or terrain, and also the customs and idioms inherent in that particular location. It is about the physical environment as well as the constructed and the cultural one. There is growing realization that terroir does not apply solely to physical products; different nations (and the subregions thereof) possess their own “intangible” assets or specializations – places that in many ways are directly in competition with each other for highly skilled labor, inward investment projects, for tourists and visitors, and so on. A common identity can be developed if corporate and regional cultures coincide. With identity and the relevant kinds of knowledge varying widely, depending on industry and place, agglomerations
emerge, with geographical specialization. Urban spaces and metropolises are of great importance for sophisticated knowledge-based services, with cities and districts attracting a creative culture, knowledge distribution, education, and so on, gaining in importance in recent years. This is not simply due to the availability of skilled labor, important trade partners and customers, or random meetings and the presence of qualified employees. As financial centers, cities like New York, London, and Tokyo have so-called surplus-meaning. Transnational companies locate in these locations for reasons connected to identity and image, triggering cumulative effects in the sense of a self-fulfilling prophecy. Location-based sources of industrial advantage (clusters, innovations systems, creative cities, etc.) are well rehearsed in the economic geography literature, but beyond providing the “industrial systems” that deliver innovative products, successful places also must offer spaces and channels for the negotiation and communication of immaterial value (Jansson and Power 2010). It follows that the attributes needed for the production of successful functional products, and for realizing their successful immaterial or symbolic aspects (or indeed brands), are place-based. These platforms or “brand channels” are different from the 1-D place-branding initiatives associated with the boosting of places of poor repute; they are a complex set of interactions enabling the transmission of what can be termed “concurrent place myths” – myths because these are powerful and lasting perceptions that go beyond straightforward brands; concurrent because large, diverse places are likely to possess multiple myths co-existing simultaneously. The key point, here, is that the interaction between product and place should firmly be placed on the policy agenda.

Over time, regional identities and mentality can foster a specific demand, shaping the attributes of the regional products (e.g., long-lasting, good quality). For regions, this basis in the “real” economy may ultimately be less important than perception and identity – they can define the prestige of a region and sometimes become an important location factor for companies (“a good address”).

This is obvious for world cities but also true for such peripheral regions as the City of Solingen in Germany, well known for its knives. The image and brand “made in Solingen” used by its famous large companies helps create the regional identity. Solingen’s historical significance and expertise in producing tools, blades and cutlery strengthen this field of competences. There is a high number of small and medium-sized highly specialized manufactures, closely cooperating with business in the design sector, and offering the possibility of innovating new products in very small and special series – a positive feature and unique selling point. This example illustrates how a firm’s image or reputation is connected to the competences and identity of a region. It be should be recognized, however, that past industrial strengths can hamper a region’s shift into new spheres of economic activity; working practices, skills, cultures, and mentalities contribute to inertia and lock-in. Consider the case of Sheffield, UK, a city once also famous (and to a degree still today) for producing steel goods, cutlery, and so on; it has encountered a difficult transition into a new economic climate, with steel manufacturing largely now absent.

While small companies cannot foster a regional identity or create a worldwide image on their own, a network of regional companies can make this possible. Networking, a reflexive phenomenon, influences cultural change and regional identities. Networks emerge through collective and evolutionary processes, connecting
network members with different cultural backgrounds, company size, world of production, position in the value chain, and so on. Again, the important question is how networks establish an identity and for which target groups. Networks or clusters and city’s regions’ CI influence each other in a reflexive way: well-known examples worldwide include Silicon Valley, the Third Italy in the Emilia Romagna, or the ICT cluster in Bangalore. Here, not only the identity of brands and corporations but also the cluster’s or network’s identity is based on the place identity, and vice versa.

In summary, identity and image are determined by the real entities of places (behaviors, economies). To find their own identity, places must focus on structures and conditions, on different spatial layers, and on different thematic activities as discussed. To minimize the risk of anonymity, places should identify consistent patterns and shared internal attributes. Although place and corporate identity are different concepts, they are fundamentally interrelated. A corporation will not have a single identity or image but these are manageable, whereas places have many internal and external identities and perceptions – or myths – capable of reaching different internal and external groups. Thus, reflexive interrelations between companies, brands, products, and places exist and influence each other, meaning that ultimately the processes establishing corporate identity are inseparable from those shaping place identity.

SEE ALSO: Brands and branding; Competitiveness; Creative class; Cultural capital; Globalization; Glocalization; Human capital; Identity; Local development; Networks, social capital, and development; New Urbanism; Social capital; World cities

References


Further reading


Corporate retailing

Shuguang Wang
Ryerson University, Canada

Retailing is the business of selling goods directly to end users. As such, it is the essential link between production and household consumption. From the political economy point of view, all consumer goods have surplus values locked up in them, and the surplus values are not realized until the consumer goods are purchased by consumers through various distribution channels (Blomley 1996). The accumulation of capital is achieved through “repeated acts of exchange” between consumers and retailers (Ducatel and Blomley 1990, 218).

In a typical capitalist market economy, the retail sector consists of diverse types of retailers. Some of them are corporate retailers, which operate multiple stores in the form of a chain and which have a complex organizational structure. Others are independent retailers with a single store location. In terms of ownership, corporate retailers can be further differentiated as publicly traded companies and privately owned chains. Relative to independent retailers, corporate retailers are small in number, but they command the largest share of the consumer market. Table 1 lists the top 30 corporate retailers in the world. They are based in seven advanced economies: the United States (12), France (5), Germany (5), the United Kingdom (2), the Netherlands (2), Australia (2), and Japan (2).

Since the mid-1980s, a series of revolutionary changes have taken place in the retail industry worldwide amid the process of globalization of the world economy. The new trends and characteristics are embodied mainly by corporate retailers, and have led to the retheorization of the orthodox retail geography and the emergence of a new retail geography with a major paradigm shift. As the retail planning process becomes more complex, so too does the role of retail geographers.

Characteristics of the contemporary capitalist retail economy and recent trends

Significant restructuring has taken place in the retail industry

In the contemporary capitalist retail economy, retail chains continue to be the most important form of retail concentration. Without exception, all corporate retailers are chain operators, as the retail chain provides a way of introducing scale economies while avoiding the restrictions of market size. Retail chains often account for 70% or more of the total retail sales in a metropolitan market. The 1990s and 2000s witnessed a significant restructuring and reshuffling of retail chains, leading to concentration of retail capital in the hands of a few “super leagues” (Marsden and Wrigley 1996). Some sought to acquire, or to merge with, others to consolidate resources, and rose to the status of global corporations. For example, in 1991 the French firm Carrefour took over two other domestic hypermarket chains, Euromarche and Montlaur, and in 1999 it merged with Promodes to create the largest European food retailing group and the second largest worldwide (at the time).
### Table 1  The top 30 corporate retailers in the world (by 2012 retail revenue).

<table>
<thead>
<tr>
<th>Company</th>
<th>Home country</th>
<th>Retail revenue ($ million)</th>
<th>Dominant format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walmart Stores, Inc.</td>
<td>USA</td>
<td>469,162</td>
<td>Hypermarket/supercenter/superstore</td>
</tr>
<tr>
<td>Tesco PLC</td>
<td>UK</td>
<td>101,269</td>
<td>Hypermarket/supercenter/superstore</td>
</tr>
<tr>
<td>Costco Wholesale Corporation</td>
<td>USA</td>
<td>99,137</td>
<td>Cash &amp; carry/warehouse club</td>
</tr>
<tr>
<td>Carrefour S.A.</td>
<td>France</td>
<td>98,757</td>
<td>Hypermarket/supercenter/superstore</td>
</tr>
<tr>
<td>The Kroger Co.</td>
<td>USA</td>
<td>96,751</td>
<td>Supermarket</td>
</tr>
<tr>
<td>Schwarz Unternehmens Treuhand KG</td>
<td>Germany</td>
<td>87,236</td>
<td>Discount store</td>
</tr>
<tr>
<td>Metro AG</td>
<td>Germany</td>
<td>85,832</td>
<td>Cash &amp; carry/warehouse club</td>
</tr>
<tr>
<td>The Home Depot, Inc.</td>
<td>USA</td>
<td>74,754</td>
<td>Home improvement</td>
</tr>
<tr>
<td>Aldi Einkauf GmbH &amp; Co. oHG</td>
<td>Germany</td>
<td>73,035</td>
<td>Discount store</td>
</tr>
<tr>
<td>Target Corporation</td>
<td>USA</td>
<td>71,960</td>
<td>Discount department store</td>
</tr>
<tr>
<td>Walgreen Co.</td>
<td>USA</td>
<td>71,633</td>
<td>Drug store/pharmacy</td>
</tr>
<tr>
<td>CVS Caremark Corp.</td>
<td>USA</td>
<td>63,654</td>
<td>Drug store/pharmacy</td>
</tr>
<tr>
<td>Aeon Co., Ltd.</td>
<td>Japan</td>
<td>63,100</td>
<td>Hypermarket/supercenter/superstore</td>
</tr>
<tr>
<td>Groupe Auchan SA</td>
<td>France</td>
<td>59,041</td>
<td>Hypermarket/supercenter/superstore</td>
</tr>
<tr>
<td>Woolworths Limited</td>
<td>Australia</td>
<td>58,602</td>
<td>Supermarket</td>
</tr>
<tr>
<td>Amazon.com, Inc.</td>
<td>USA</td>
<td>58,570</td>
<td>Nonstore</td>
</tr>
<tr>
<td>Seven &amp; i Holdings Co., Ltd.</td>
<td>Japan</td>
<td>58,329</td>
<td>Hypermarket/supercenter/superstore</td>
</tr>
<tr>
<td>Edeka Zentrale AG &amp; Co. KG</td>
<td>Germany</td>
<td>55,944</td>
<td>Supermarket</td>
</tr>
<tr>
<td>Wesfarmers Limited</td>
<td>Australia</td>
<td>54,231</td>
<td>Supermarket</td>
</tr>
<tr>
<td>Casino Guichard-Perrachon S.A.</td>
<td>France</td>
<td>53,375</td>
<td>Hypermarket/supercenter/superstore</td>
</tr>
<tr>
<td>Lowe’s Companies, Inc.</td>
<td>USA</td>
<td>50,521</td>
<td>Home improvement</td>
</tr>
<tr>
<td>Rewe Combine</td>
<td>Germany</td>
<td>48,984</td>
<td>Supermarket</td>
</tr>
<tr>
<td>Best Buy Co., Inc.</td>
<td>USA</td>
<td>45,085</td>
<td>Electronics specialty</td>
</tr>
<tr>
<td>Centres Distributeurs E. Leclerc</td>
<td>France</td>
<td>44,807</td>
<td>Hypermarket/supercenter/superstore</td>
</tr>
<tr>
<td>Safeway Inc.</td>
<td>USA</td>
<td>43,322</td>
<td>Supermarket</td>
</tr>
<tr>
<td>Koninklijke Ahold N.V.</td>
<td>The Netherlands</td>
<td>42,236</td>
<td>Supermarket</td>
</tr>
<tr>
<td>Sears Holdings Corp.</td>
<td>USA</td>
<td>39,854</td>
<td>Department store</td>
</tr>
<tr>
<td>J Sainsbury plc</td>
<td>UK</td>
<td>36,840</td>
<td>Supermarket</td>
</tr>
<tr>
<td>ITM D</td>
<td>France</td>
<td>50,286</td>
<td>Supermarket</td>
</tr>
<tr>
<td>The IKEA Group (INGKA Holding B.V.)</td>
<td>The Netherlands</td>
<td>36,111</td>
<td>Other specialty</td>
</tr>
</tbody>
</table>

Source: Deloitte (2014).
In 2002 the American electronics retailer, Best Buy, expanded by acquiring a Canadian retail chain, Future Shop. Large corporate retailers are better able to invest heavily in information technology and centralized distribution systems, the use of which enables them to more effectively perform their competitive functions of reducing the turnover time of commodities (Hughes 1996). Meanwhile, many other chains were not as fortunate and subsequently went under. Eaton’s, a Canadian department store chain with a history of more than 100 years, was one such victim.

**Big-box stores have emerged as new leading retail formats**

The rapid progress of suburbanization in the late 1950s and the 1960s required that cities be planned following the principle of hierarchical organization. Accordingly, a hierarchy of shopping centers was incorporated in this form of city planning, consisting of neighborhood, community, and regional shopping centers. Each type of shopping center had a clearly prescribed tenant mix, trade area size, and even physical form (Casazza 1985). In the 1970s a newer and larger type of shopping center — the super-regional shopping center — began to be built. These centers often combine entertainment and recreation with shopping under one roof, becoming new palaces of consumption. Typical examples are the West Edmonton Mall in Canada (493,000 m²) and the Mall of America in the United States (466,000 m²). As suburbanization continued, the metropolitan city became multicentric, and a network of retail nodes evolved to serve the expansive city. Each regional and super-regional shopping center became a node, spatially distributed in a hierarchical fashion analogous to central places in Christaller’s Central Place Model.

By the late 1980s, retailing had become an important and dynamic economic sector in many developed nations. As well, major technological innovations took place in the distribution system. Highly computerized goods-tracking and inventory control systems enabled direct communications with, and direct shipping from, manufacturers (Hughes and Seneca 1997). In the process, it has led to the emergence of a new retail format, known as “big-box” stores. At first, such stores were freestanding outside shopping centers; but eventually several big-box stores began to cluster together at one location in the form of a planned plaza, commonly called a “power center” (also known as a retail park). In less than 10 years, this new format has been adopted by a variety of retail businesses. These new retail spaces have been created and manipulated by the innovative retailers and commercial real estate developers to induce consumption. The big-box stores and power centers have become leading nodes in the Western retail system, widely blamed as the cause of the “graying” of regional and super-regional shopping centers and the demise of department stores in North American cities.

**Retailer–supplier relationship has tipped toward retailers**

For many years, the producers of consumer goods effectively dictated the brands and types of products being sold in the retail market, as well as their price. Corporate retailers are no longer passive receivers of consumer goods supplied to them by producers. Since the late 1980s, retailing has been shifting from being the sales agent for manufacturing and agriculture to being the production agent for consumers, and the balance of the retailer–supplier relationship has tipped toward the large retail chains. While still mediating between consumers and producers, they are no longer neutral mediators. Instead,
they induce consumers to select goods with larger profit margins. They erode manufacturers’ share of the surplus value by influencing patterns of consumption in their own favor and by using their bargaining power to lock manufacturers into retailer-led supply chains (Hughes 1996, 99). They also aggressively develop own-label products as a competitive strategy. With their significant purchasing and bargaining power, they manage to exploit negative working-capital cycles by negotiating for extended credit payment periods, reducing circulation costs and accelerating the accumulation of capital (Marsden and Wrigley 1996). Pressure is often placed on suppliers to develop just-in-time production programs and factory-to-warehouse distribution systems to match retailers’ demands.

Internationalization of retailing has intensified

As their home markets became saturated, large retail chains were motivated to explore foreign markets, using a variety of entry paths ranging from licensing and franchising to joint ventures and wholly owned subsidiaries. Almost every successful large retail chain has been seeking possibilities of expansion into other countries. Initially, they chose markets that were closer in terms of physical and psychic distance, in order to minimize cost and the degree of uncertainty about sourcing and operation. In recent years, the major international retailers have increasingly turned their attentions to emerging markets in Asia and Latin America. For example, Walmart became an international company in 1991. By 2012 it had more than 5000 stores in 26 countries. Carrefour, which went international in the 1970s, intensified its overseas operations since the late 1980s, and now operate 9500 stores in 32 countries. For many multinational retailers, revenue from overseas operations has become an increasingly important part of their business success, or even survival.

Distribution channels have diversified

In the past, goods were distributed to consumers in a linear fashion: manufacturers sold large quantities of products to wholesalers, which in turn sold the products to retailers (see Figure 1). Since the 1990s the distribution system has become much more complex. In many cases, either wholesalers or retailers are bypassed. For example, some producers sell products directly to end users through self-operated factory outlets. Certain wholesalers (such as Costco and Sam’s Club) also sell goods directly to households.

In the early 1990s a new agent of retailing and form of retail space came into being: online shopping and web stores. Online shopping is one form of electronic commerce, whereby consumers purchase goods (or services) from a seller over the Internet, without having to visit a brick-and-mortar store. Shoppers must have access to a computer connected to the Internet and a method of electronic payment – typically a credit card. Once a payment is received (usually through a third-party e-commerce business that facilitates online money transfer between consumers and retailers, such as Paypal), the goods can be shipped to a prescribed address via post or courier services, or picked up by the consumer from a nearby store. This type of electronic commerce is used for business-to-consumer (B2C) transactions and mail orders, and led to the creation of Amazon in 1995 and eBay in 1996. The 2000s saw an increasing amount of retailing using online websites, electronic payment, and delivery of merchandise to consumers via a courier service. Initially, online shoppers were deterred by fraud and privacy concerns, such as the risk of cyber theft of credit card numbers and identities. In recent years, such
risks have been greatly reduced by the use of Secure Sockets Layer (SSL) encryption and firewalls, and “e-tailers” improving their refund policies. As a result, more and more consumers are encouraged to shop online for convenience and saving, as web stores are open all year round with no after-business hours and no restrictions of location or distance.

Most people tend to think that web stores are purely virtual space, but “e-tailers” must set up physical facilities such as warehouses and regional distribution centers by either constructing new buildings or leasing existing spaces, in order to reduce delivery time and cost. These facilities require heavy monetary investment. Not all merchandises have the same level of suitability for sale online. Yet, “e-tailers” have been expanding their offerings and taking market shares away from the brick-and-mortar retailers.

In North America, Amazon continues to be on the march, successfully moving into merchandise that Walmart and other brick-and-mortar retailers traditionally have sold. In China, Alibaba’s online portals handled 1.1 trillion yuan (US$170 billion) in sales in 2012, more than eBay and Amazon.com combined (Economist 2013). More retailers responded to the new shopping pattern by offering both multichannel retailing (selling in-store and online in parallel) and cross-channel retailing (order online but pick up in-store).

The increased popularity of online shopping is leading to another revolution in the retail industry. In the new world order of retail, shoppers have become more tech savvy, often using smartphones to shop, order, and pay. Or, they go to a store to look at and feel a product, and then buy it online from Amazon or Alibaba, turning brick-and-mortar stores into showrooms.
The new geography of retailing: a paradigm shift

Retailing is a subject of study by two schools of scholars and students, business management and geography, but they approach the subject from different perspectives. The former concentrate on consumer behavior and business organizations, including logistics, merchandising, marketing, in-store display, and customer service; whereas the latter have traditionally focused on the selection and determination of business locations and the delimitation of trade areas in localized markets.

Retail geography, which originated as marketing geography but lately has been called business geography by some, became a separate field of study in the mid-1950s in the United States. William Applebaum is widely regarded as the chief architect of this field of study. He emphasized that marketing geography should be viewed essentially as an applied, rather than as a purely academic, subject (Applebaum 1954). He also considered that the best place to develop the field of marketing geography was in business itself. Marketing geography was further developed and advanced by prominent geographers such as Brain Berry and David Huff. In its first 40 years, retail geography was primarily concerned with the identification (i.e., areal expression) of the demand for various goods and services and with the spatial arrangements for supplying them through an efficient distribution network (Davis 1976). In general, this orthodox retail geography, as it is called by some contemporary retail geographers, was founded on a predictive and instrumentalist epistemology. Most of its ontological presumptions are linked to the neoclassical economic view of the world, such as central place theory, the gravity model, and distance decay function, which conceive space as a neutral container, at most affecting transportation costs. It focused primarily on the retailer rather than on the supply chain as a whole, and on the geography of stores while neglecting the geography of such important trends as the centralization of retail distribution operations. The traditional maxim of “location, location, location” as a study focus was well reflected in orthodox retail geography (Jones and Simmons 1993).

The early 1990s witnessed a major paradigm shift in the study of retail geography: orthodox retail geography was retheorized by Wrigley, Lowe, and a few other European economic geographers, and a new geography of retailing was advocated to reflect the important changes in the global economy highlighted in the preceding section. The new geography of retailing has three important characteristics in comparison with the orthodox retail geography (Lowe and Wrigley 1996).

First, it takes a political economy approach, seeing retail capital as a component of a larger system of production and consumption (Blomley 1996). This extends the scope of study to include production spheres in the system of circulation activities, particularly the production–commerce interface (the changing relations between retailers and suppliers). Aided by advanced technologies, retailers have developed logistically efficient stock control systems and centrally controlled warehouse-to-store distribution networks. These systems permit shorter and more predictable lead times, with important implications for the configuration of new retail spaces as well as the reconfiguration of existing retail spaces.

Second, it concerns the geography of retail restructuring and the grounding of global flows of retail capital (i.e., sinking of capital into physical assets in overseas markets) and its geographical expressions (i.e., spatial outcomes) in the form of retail facilities of different formats. Innovative retailers, teamed up with developers, create differentiated spaces of retailing in
the same or different markets, organized to induce consumption (Ducatel and Blomley 1990).

Third, it calls for much more serious treatment of market regulation because the regulatory state is an important force, influencing both corporate strategies and geographical market structures. Lowe and Wrigley (1996) argue that orthodox retail geography was remarkably silent about regulation and the complex and contradictory relations of retail capital with the regulatory state. With the exception of a discussion concerning the constraining influence of land-use planning regulations, the transformation of retail capital appeared to take place in a world devoid of macroregulatory environment shaping competition between firms and the governance of investment. In this sense, regulations include both public interest and private interest interventions, with the former concerning relations between retailers and consumers, and the latter concerning relationships between retailers and suppliers (including producers).

The retail planning process and the role of retail geographers

The success of a retail corporation in business development depends on a well-defined planning process, where retail geographers can play an important role. As illustrated in Figure 2, the retail planning process starts with the retail company setting up its development goals and objectives on the basis of the company’s market position and resources (see Box 1 in Figure 2). The goal of resource-rich and competitive companies is usually to grow or expand. There are two general types of growth strategies: nonspatial and spatial. The former refers to driving sales from existing stores, the latter to expanding the network by adding more stores at more locations.

For companies with mediocre performance, the goal is often to maintain the status quo and improve the performance of existing stores. Less competitive and weaker companies are most likely to choose to reduce scales of operation by closing unprofitable stores or selling off the entire operation.

If a company decides to expand, the next step is to select a target market (Box 2). A market can range in size from a country or region to a city or town. Alternative strategies include expanding in an existing market or entering new markets. A suitable market is selected based not only on extensive analysis of market conditions but also on a careful examination of the retail environment, particularly when a new market is being considered. Some countries are defined as having developed markets, with well-established trading rules and regulations, and others as emerging markets, which converge toward the developed markets in advanced economies. While they may not have reached the level of market efficiency with strict standards in accounting and securities regulation to be on a par with advanced economies, emerging markets are nonetheless sought after by multinational retailers because of the prospect of high returns. They experience faster economic growth, although investments are also often accompanied by high risks due to political instability, domestic infrastructure problems, and currency volatility. Within most countries, there exist considerable or even significant variations between subnational markets in terms of both market condition and retail environment. Some regions are more mature and more affluent markets than others. Different provinces/states and even municipalities often have different regulations and bylaws that directly or indirectly affect retailing operation and expansion.

Market conditions are assessed for both the demand and supply of consumer goods (Box 3).
Typically, demand is estimated on the basis of population size, population composition, and disposable income, which determine demand quantity and purchasing powers. In some cases, the level of automobile ownership is also considered because it affects consumer mobility and therefore shopping behaviors. Population projection is an important consideration in market condition analysis as well, as it facilitates decision-making about the scale of retail development necessary to meet future demand while minimizing investment risks and potential sunk costs.

On the supply side, market condition assessment includes both retail structure analysis and competition analysis. This helps to gauge the level of market saturation, the strengths and weaknesses of potential competitors, and the market space left for new entrants. For example, Walmart introduced its Sam’s Club to Canada in 2006. In the next three years, it opened six club stores in southern Ontario, where Costco had already established a strong presence with a large base of loyal customers. As the result of an underestimation of the competitiveness of Costco in Canada in its planning process, the six
The retail environment includes regulations, local attitudes (toward certain retailers and formats), and infrastructures (Box 4). Different countries, but also provinces and municipalities, can have different regulations, such as tax laws, anticompetition laws, land-use bylaws, and foreign investment policies. Within the capitalist market economy, governments rarely own or operate retail stores, except for a few selected consumer goods, such as the nonmerit goods of alcoholic beverages and cigarettes; in Ontario, Canada, for example, all liquor stores are owned and operated by the Liquor Control Board of Ontario, a provincial government enterprise. While retailers enjoy a high degree of freedom in business decision-making, governments do intervene in the retail industry with regulatory measures, often in the form of public policies, to mitigate negative impacts. In turn, these policies, which are supposed to reflect a society’s values, impose limits on retailers’ freedom to deal with competitors and to conduct business with suppliers and consumers, affecting retail operations and the overall market structure.

Governments also regulate the distribution system and the relationships between retailers and suppliers and between retailers and consumers. Dawson (1980) has generalized five types of public intervention that commonly exist in Western capitalist economies: (i) location restriction through land-use planning and zoning bylaws to minimize spatial externalities; (ii) price control to protect consumers from inadequate advertising and predatory selling practices, and to generate revenues for governments; (iii) optimization of business structure through fair competition laws to prevent monopoly and encourage innovation; (iv) promotion of business efficiency by controlling new entries into a market to avoid excess retail capacities and waste of economic resources; and (v) protection of consumer wellbeing and safety through licensing/inspections and compulsory labeling. Dawson also pointed out that few policies are directed toward retailing but numerous indirect policies aimed at other sectors impinge on retailing. These various types of public intervention are often imposed by different levels of government and implemented at different spatial scales (Jones and Simmons 1993).

In Western democracies, the public attitude toward a particular retailer is also an important consideration in retail planning. In the United States, anti-Walmart sentiments in some communities have derailed the retailer’s plan to enter a number of local markets. The opposition is not just against the big-box format; it has also opposed the way Walmart conducts business, including its anti-union corporate culture, low wages, and heavy-handed treatment of suppliers. While Walmart now operates 4800 outlets across the United States, it has been unable to enter New York City – the United States’ largest metropolitan market. In 2012 Walmart planned to open its first store in Brooklyn, but met strong oppositions from many members of the local community and unions (Greenhouse and Clifford 2013). Despite having spent nearly $10 million in news media to project a positive image of the retailer to New York City’s citizens, Walmart was forced to abandon its plan. This signifies that retailers must explore local temperaments in the planning process before it goes too far; elected municipal officials cannot afford to ignore the concerns of their constituents. Although retailers can appeal government decisions, the process can be lengthy and costly.

Once a market is chosen, the retailer must decide what format of retailing is most appropriate and profitable there (Box 5). The choice of which retailing format depends on market conditions. Formats include the department
CORPORATE RETAILING

store, supermarket, supercenter, membership club, warehouse retail store, and factory outlet. Some retailers operate in more than one format. Walmart entered Canada in 1994, choosing the discount department store format. This proved less profitable than expected, so it later introduced both the supercenter and Sam’s Club to Canada. Its supercenters are flourishing, despite the failure of Sam’s Clubs. When Walmart entered China in the mid-1990s, it introduced and experimented simultaneously with three formats: the supercenter, Sam’s Club, and the neighborhood supermarket. Again, the supercenters are the most successful, but Sam’s Clubs grew very slowly: Walmart has opened 362 supercenters in China, but only 10 Sam’s Clubs; neighborhood supermarkets were a failure, with only two stores.

Consideration of retail formats and selection of market entry paths often go hand in hand (Box 6). Alternative paths include organic (or internal) growth, merger, acquisition, joint ventures, and franchising. Each entry path has its merits and limitations, as detailed by Birkin, Clarke, and Clarke (2002).

Organic growth refers to the process of opening new stores in an existing market or a new market. From a real estate point of view, this can be achieved through the construction of new buildings on retailer-owned land or in leased properties. Birkin, Clarke, and Clarke (2002) differentiate between two types of organic growth: hierarchical diffusion and contagious diffusion. The former refers to the opening of new stores, first in the largest cities, and then working systematically down the national urban hierarchy. The latter refers to the diffusion of retail outlets from a single city to adjacent cities in the same region or province. While organic growth gives the retailer more control over the pace of store development and operation, its spatial growth can be slow as a result of the lengthy process of store location selection and the difficulty of acquiring properties, particularly in foreign markets.

Merger and acquisition can produce fast network expansion and rapid geographic growth in markets where the retailer is currently un- or underrepresented. Through merger and acquisition, the retailer can also eliminate competition from the target market. However, buying a large existing chain requires a large up-front investment. Achieving organizational and cultural fit in the stage of implementation and integration can also be very costly.

Joint venture is mostly used when a retailer enters a foreign country, where it usually forms a new business entity with a local partner, either by choice or as dictated by the host country government. Before China’s admission to the World Trade Organization (WTO), most foreign retailers entered by forming joint ventures with a local partner, as required by the Chinese government. The retailer in question can benefit from the local partner’s native knowledge of market conditions and the local retail environment. The local partner can also be instrumental in searching for business sites and obtaining government approval. Depending on the nature of ownership, the investing retailer may not be able to implement its business plan smoothly without friction with its local partner, especially in equal-ownership or minority-ownership joint ventures. Thus joint ventures often result in a takeover or buyout of the local partner, turning the joint venture into a wholly owned subsidiary, as has sometimes been the case in China.

There are two main forms of franchising: direct franchising and master franchising agreements (Quinn 1998). The former links the franchisor and the franchisee directly. In the latter, the franchisor grants the master franchisee the right to subfranchise to others within an exclusive
territory. This method of growth has been especially popular in recent years for international growth, where the franchisor is less familiar with the local market. Intermediary organizations are often in a stronger position to handle cultural and language issues. Franchising can also produce fast store network expansion with less capital investment and therefore lower risks. However, the level of profit is also low, and the franchisor’s control over the operation of the franchisees can be weak.

For the purpose of organic growth (i.e., new store development), detailed location analysis is necessary (Box 7), usually after approval has been obtained from the relevant government authorities: nation-states in the case of foreign markets, and local governments in domestic markets. Location analysis includes site selection, market condition analysis at the store level, search for suitable and affordable real estate, pro forma analysis, review of zoning bylaws, trade area analysis, and projection of sales potential. Retail chains often also conduct location-allocation analysis, which is concerned with the selection of multiple store locations simultaneously to serve a spatially dispersed population (i.e., to allocate a given spatial distribution of demand to a specific number of outlets). Location-allocation modeling seeks to determine how many stores are needed in the target market, how many can be supported, and where they should be located.

Spatial expansion of a store network must be supported by a network of distribution centers. For retailers operating large chains, the deployment and construction of distribution centers is also an important part of retail planning (Box 8). A distribution center operated by a corporate retailer is a specialized warehouse stocked with goods to be redistributed to the retail outlets in the same regional market. It is a vital part of the entire order fulfillment process, enabling a single location to stock a vast number of products.

After the business is opened, store performance is continually monitored (Box 9). Depending on ongoing performance, corporate goals/objectives are revisited and adjusted. Again, strong performers and profitable retailers are most likely to choose to grow and expand; mediocre performers may decide to maintain the status quo and focus on improving performance of existing stores; and poor performing retailers are forced to contract (Box 1).

As illustrated in this section, the retail planning process is complex, requiring knowledge from multiple disciplines and involving the participation of experts with different professional backgrounds. Retail geographers with training in spatial analysis are equipped with techniques and tools to play an important role in market condition analysis and location analyses, as highlighted in Boxes 3 and 7 in Figure 2. Analytical methods and tools also have been automated with the use of geographical information systems (GIS). Thus Esri Business Analyst now offers a comprehensive suite of tools including the calibrated Huff Model and various location-allocation models that meet different criteria in response to different constraints. These tasks reflect the typical concerns of orthodox retail geography, but the conceptualization of them is now informed and improved by political economic theorizations in the new geography of retailing.

SEE ALSO: Corporate spatial organization and producer services; Corporations and e-commerce; Economic geography

References


Corporate spatial organization and producer services

James R. Faulconbridge
Lancaster University, UK

Jonathan V. Beaverstock
University of Bristol, UK

Sarah Hall
University of Nottingham, UK

Geographical research on producer services has been dominated by questions about the role of services in processes of corporate globalization. Producer services such as accountancy and law facilitate the globalization of manufacturers, and producer service firms themselves have actively globalized over the past 40 years, both to serve existing home-country clients overseas and to penetrate new markets. This has led to significant interest in how internationalization theory can be applied to producer services. Recognition of the growing centrality of producer services in the economy dovetailed with the growth of postindustrial economies in Western Europe and North America, with research highlighting how by the late 1980s and early 1990s urban economies were increasingly founded upon producer service activity, in particular in relation to finance (Sassen 1991).

Studies can be broadly grouped into two clusters. First, research sought to describe and explain the spatial organization of producer service firms within countries. In this work, particular attention was paid to the urban geographies of producer service firms, with observations being made that in Western Europe and North America it was not unusual to find in excess of 90% of producer service employment in cities (Bryson, Daniels, and Warf 2004). This metropolitan bias inspired a genre of work on the causes and effects of uneven geographies of producer services, not only across the urban–rural divide but also between cities as spatial divisions of city-based labor emerged within countries. Second, attention also turned to the geographical dimensions of the drivers, processes, and outcomes of producer service globalization. By focusing upon the geographies of the overseas activities of producer service firms in sectors including accountancy, advertising, engineering, and law, it was shown that a range of motivations for globalization were leading to concentrations of global producer service firms in selected world cities (Sassen 1991), with a range of organizational strategies being deployed to establish offices and service clients.

Urban fix

The rise of the producer service corporation has a distinctive geography that is characterized by the clustering of firms within a relatively small number of cities. This urban concentration of producer services counteracts claims that the growing importance of information and communication technology renders service sector activities footloose and geographically mobile as part of what O’Brien (1992) termed “the end of geography.” For authors such as O’Brien, services would be rendered highly mobile since the knowledge and information that is central to their operation could, theoretically, be circulated
globally, thereby removing locational constraints. However, research in economic geography and cognate social sciences from the 1980s onwards has demonstrated that, far from being evenly distributed globally, producer services are one of the most, if not the most, geographically concentrated sectors of economic activity. It has been demonstrated that a small number of leading cities, globally and at the national level, act as nodes within networks, housing significant clusters of a range of producer services. Saskia Sassen (1991) most clearly articulated this urban geography in her work on what she terms the “global city.” For Sassen, London, New York, and Tokyo form a powerful triad in the global economy, housing the majority of headquarters and commercial activity of producer services—a triad that also controls flows of knowledge, capital, and people that drive the contemporary global service economy.

In order to understand why this urban geography typifies producer service firms, it is necessary to focus on the nature of knowledge used within firms in this sector, reflecting its importance as the “raw material” of producer services. In particular, it has been demonstrated that the bespoke, specialist knowledge that is developed at the boundary between the producer service firm and their client requires close interaction, typically involving face-to-face contact and hence geographical proximity. Moreover, in addition to this client–provider relationship, product development within producer services benefits from cluster-based traded (i.e., formal transactions between actors in a cluster) and untraded interdependencies (understood as the informal benefits of being co-located, such as being able to take advantage of a specialist, highly skilled labor market or more informal interaction through sociocultural activities in clusters).

Understanding of the distinctive urban nature of producer services is very well illustrated by financial services where research has consistently demonstrated the dominance of a small number of cities in the international financial system. London and New York are key examples of such international financial centers (IFCs). Thrift (1994) identifies four elements that explain this urban hierarchy. First, the nature of financial services firms themselves serves to encourage them to co-locate in IFCs as they compete with each other for labor and clients but also simultaneously cooperate in sharing intelligence about financial markets and taking advantage of the highly skilled labor markets that characterize IFCs. Second, IFCs incorporate diverse and deep markets for financial services that allow firms to enter and exit relatively easily and learn about market pricing information. Third, the culture of IFCs highlights their role as key places of information processing and knowledge generation. As such, they serve as clusters of highly skilled individuals who co-locate to interact with colleagues and competitors in both informal and formal settings—activities that are central to the operation of the international financial system in terms of learning about new products, client demands, and subtle market pricing signals. Fourth, IFCs generate important external economies of scale due to the depth and liquidity of financial and producer services within them. These include the ability to share fixed costs such as infrastructure related to market settlement and drawing on the associated clustering of related producer services, notably law and accountancy, which have become more important to financial services as a result of the rise of financial products such as securitization from the 2000s onwards (Faulconbridge and Hall 2014).

London and New York have been widely identified as the leading IFCs. Similar patterns of urban concentration in a small number of cities can be seen for other producer services. For example, leading law firms are overwhelmingly
headquartered in the United States or United Kingdom, with those based in the latter leading corporate international expansion into Asian markets. This international expansion is being facilitated through the opening of offices in new markets (rather than servicing markets from a distance relying on ICT) because legal regulations are country specific and hence lawyers can only offer advice in the country in which they are registered to do so. Moreover, the information upon which legal transactions and products are based is often confidential and sensitive. As a result, it is best exchanged through trust-based relationships which are predominately facilitated through face-to-face interaction, thereby requiring an office presence in markets to be served.

Whilst the dominance of a small number of global cities is well documented across a range of sectors, more recently, attention has shifted to new features of the centrality of cities for the locational choices of producer service firms. Two strands of research stand out as particularly significant. First, the relationship between leading global cities and the clusters of producer services in smaller urban settings has been documented, with particular attention to countries such as the United Kingdom which are heavily dominated by one global city. This is valuable because it demonstrates how clusters of producer services in smaller cities such as Birmingham, Manchester, and Leeds, in the case of the United Kingdom, interface directly with the global economy by developing their own connections and functionally specific clusters.

Second, the changing geographies at the global scale of the clustering of producer services in cities have been explored. At one level, this has shown the growing importance of Asian cities as places for the clustering of producer services. In part this reflects the stagnation of the Japanese economy and the associated decline of Tokyo as a global city, as identified by Sassen (1991), when compared to New York and London. However, it is also being driven to a very considerable extent by the growing power and reach of China in the global economy and the desire for producer service firms from Europe and North America to open offices there in order to service a rapidly expanding market. While the geography of concentration in China in terms of city location is still unfolding, established clusters in Singapore and Hong Kong are facing increasing competition from emerging clusters in Shanghai and Beijing (Lai 2012).

A further important trend in terms of the city location of producer services is the relationship between global cities and the offshore activities of producer service firms, particularly in terms of finance. Here, building on earlier work on offshore financial centers, it is shown how global cities rely upon a network of offshore activities for their own economic growth as well as, particularly in the case of London, acting as a quasi-offshore center themselves as producer service firms use the dense highly skilled labor markets to devise new products and forms of service delivery associated with the growth of offshoring.

Rise of the producer service multinational corporation

At the root of the corporate spatial organization of the producer sector in the world economy has been the emergence of the producer service multinational corporation (MNC), especially in the latter quarter of the twentieth century. The rapid growth of the producer service MNC has effectively been the “driver” (Bryson, Daniels, and Warf 2004, 76) for the internationalization of producer services and responsible for their intensive locational preference for the urban fix of a metropolitan area. Across different producer service sectors, from logistics and distribution to
banking, finance, insurance and real estate, and business services, MNCs dominate the accumulation of global revenue, size of market share, and ultimately, levels of profitability in their specific sector, leaving small and medium-sized firms in their wake.

The growth of the producer service MNC can be explained and documented in two different ways: first, by focusing on the main (theoretical) reasons postulated for the growth of producer service MNCs – which are both external and internal to the firm; second, by examining the way in which producer service MNCs have grown and internationalized their investments outside their home country (via new office expansion or mergers and acquisition activity, for example) through waves of foreign direct investment (FDI) in host countries aided by regulatory factors, such as the General Agreement on Trade in Services (GATS).

**Externalization and internalization**

There is a very long tradition of literature on the producer service MNC benefiting from the externalization of functions by other firms. Over the last 25 years or so many firms, both MNCs and small and medium-sized enterprises (SMEs) in all sectors of the economy (primary, manufacturing, public administration), have externalized previously in-house functions to specialist producer service firms. Externalization has occurred not only for specialist business services such as accounting, advertising and marketing, consulting, human resources, legal services and real estate, but also more recently in activities such as logistics and distribution (reflected, for example, in the rise of DHL and Federal Express) and education (see Bryson, Daniels, and Warf 2004). Such services, once internalized in MNCs, are very likely to be externalized and thus create a derived demand for specialist producer services that can be called upon using quite flexible contractual arrangements for the supply of specific services and bespoke expertise. It is often the case that the global producer service MNCs maintain repeat business arrangements with a mix of clients from the private and public sectors provided, of course, that they deliver reliably and incorporate state-of-the-art knowledge and expertise.

Another aspect of externalization has been the rapid growth of producer service business-to-business demand for other producer service inputs from a wide array of buyers and co-suppliers who require both specialist products and bespoke advisory services. The genesis of 24-hour global financial markets and global financial centers, in Europe, North America, and Asia, and more recently in mainland China and Brazil, Russia, and India, has fostered unprecedented demand for producer service inputs, particularly in co-location and interdependency with other producer services. The proliferation of business-to-business (B2B) producer service inputs has acted as a major catalyst for the geographical expansion and coverage of producer service MNCs throughout the world, particularly in the “advanced” producer and business services. For example, during the late 1980s and 1990s, the major US law firms entered markets in Asia and Europe in response to demand from their US corporate clients (mainly wholesale banks, insurance companies, advertising agencies) operating out of cities such as London, Frankfurt, Singapore, and Hong Kong. The ICT revolution and introduction of fiber-optic technology and infrastructure were important vectors which aided the internationalization of producer services throughout that period, but at the turn of the new millennium, the rise of the Internet age and e-commerce (mainly B2B) has certainly provided new technologies, practices, and processes that producer service MNCs use to market and sell their products and advisory
services to external buyers in more innovative and efficient ways.

In parallel with the rise of the producer service MNC, in response to new demand from external buyers, a series of internal processes which are a cornerstone of a firm’s strategy for entering and penetrating new markets have developed: the ability to deliver (sell) a rich knowledge base and often bespoke expertise to clients (buyers) at the point of demand. In essence, producer service firms have entered into international production through the establishment of branch office networks directly in the marketplace, either via the growth of wholly owned subsidiaries or through mergers and acquisitions with local host firms. Since producer service MNCs are offering knowledge, expertise, and intelligence that is mainly embodied in their professional workforce of accountants, brokers, advertising creatives, consultants, head hunters, lawyers, and wholesale bankers, the most efficient mechanism for delivery is through a physical entity that ensures close co-location between the buyer and seller. Producer service MNCs have therefore invested in extensive international office networks outside of their home country to sell to buyers directly in the market. This has the added benefit of enabling them to protect and enhance their corporate reputation and trust relationships with buyers and, importantly, co-suppliers.

There are two other internal factors driving the rapid office expansion, both in terms of the numerical expansion of offices worldwide and in terms of an increase in their actual size (e.g., number of professionals employed): sector-specific concentration, and product diversification and innovation (see Bryson, Daniels, and Warf 2004). Throughout many producer service sectors, a small group of very large MNCs (which may range from 4 to 10 firms in any one sector) dominate global revenue, market share, and profitability, followed by a larger group of other (smaller) MNCs, often with regional coverage, and then a constellation of literally thousands of SMEs and sole traders. This concentration emerged from the late 1970s/early 1980s as many of the successful firms merged to create global producer service firms. Such concentration is very well advanced in sectors such as accounting, but also in advertising, consulting, legal services, and logistics, amongst others. Often firms that have merged with their competitors have done so to increase the reach of their office networks into selected geographical world regions rather than to fulfill rationalization or restructuring objectives typical for firms in other sectors of the economy.

Foreign direct investment and globalization

The rise of the producer service MNC has resulted in rapid growth in FDI. For example, between 1990 and 2002, the estimated world outward flow of FDI in finance and business services increased from US$447,937 million to US$3,305,192 million (an increase of 638%), and much of it has financed the establishment of wholly owned offices and merger and acquisition (see UNCTAD 2004). Such FDI decision-making by producer service MNCs has reaffirmed their structural characteristics (concentration and diversification) and consolidated their locational preferences for world cities. The global regulatory environment has encouraged this growth in FDI, especially the ongoing efforts to liberalize international trade and investment in services, embodied in the GATS, introduced in the early 2000s by the World Trade Organization.

Organizational spaces

As a result of the internationalization and city-based geographies of producer services,
a number of distinctive forms of organization have developed. These are designed to exploit the local-global geographies of producer service firms, in particular because of the importance of the production and leverage of knowledge as part of the process of delivering advice to clients. Two distinctive modes of spatial organization have been noted.

First is the project team. In many cases, the delivery of producer service work to the client relates to a specific problem or task as part of a time-bounded project. For instance, lawyers may be called on for advice in relation to a merger or acquisition, whilst consultants may be used for advice about the introduction of a new computer system. Such work requires specific types of expertise that only certain employees of the producer service firm might possess. As a result, a distinctive means of service production is the temporary project team (see Grabher 2001) that is established for a specific client project, and then disbanded upon completion of the advisory task. Of particular interest, in relation to corporate spatial organization, is the geography of the project team.

In line with earlier discussions of the interrelated processes of internationalization and localization in cities, project teams have local-global geographies that inflect in important ways the nature of the advice provided to clients. Locally, project teams are often assembled not only by including workers based in the office of the producer service firm in question, but also by exploiting particular advantages of the city-based geographies of firms discussed earlier, by bringing in expertise from other co-located firms. Such city-based project ecologies are especially important for creative producer service firms such as advertising (see Faulconbridge et al. 2011).

Project teams can also have global geographies. Grabher (2001) highlights how the global office networks of firms, including the social networks of individuals within them, allow project teams to be staffed with workers not just from the office managing the project. Individuals may be selected to be part of a project team based on their expertise and how they can contribute to the success of a project rather than on where they are located.

Secondly, and related to project teams, another important spatial characteristic of the organization of producer services is labor mobility. Partly as a result of the local-global geographies of project teams, but also related to wider strategies of talent management within firms, portfolios of labor mobility (Millar and Salt 2008) now exist within producer services. At one level these portfolios exist around interfirm mobility. At the city level there is extensive churn of individuals between competing producer services that is facilitated by the urban ecologies mentioned earlier. Globally, there are also significant flows in elite labor markets; individuals and their careers are defined by their mobility between key sites of producer service work. This mobility is facilitated by executive search firms, which are producer services that specialize in helping their clients to identify mobile talent that can be recruited to fulfill a range of specialist and senior management roles (Beaverstock, Faulconbridge, and Hall 2015).

At another level, these portfolios are associated with intrafirm mobility, often as part of the internal labor markets of a firm. A key feature of these portfolios is the different temporal duration of the multiple forms of mobility that exist. Mobility ranges from the short term, such as business travel which may last a few days or weeks, through to longer-term rotations or secondments which last a few months, and expatriation, which can be for several years (Millar and Salt 2008). Underlying all of this mobility of workers is the need for firms to
CORPORATE SPATIAL ORGANIZATION AND PRODUCER SERVICES

manage their most fundamental asset, the human resources that can most effectively produce and leverage the knowledge required to deliver the best advice to clients.

The importance of temporary project teams and portfolios of labor mobility has two further implications. One important effect arises from the management practice of those running producer service firms. Such practice requires an acute awareness of the spatialities of the work of firms, meaning both their embeddedness within a city and also the global relational networks that make work possible. For instance, those managing firms must become part of the professional communities made up of competing firms, regulators, and industry commentators in the cities where they work, this being mediated through professional associations as well as more informal social gatherings in the bars and restaurants. Partners and CEOs must also undertake frequent business travel associated with the management of global producer service firms. This not only allows the work of project teams to be monitored and controlled, but also enables corporate cultures and strategic planning to be overseen on a global basis. In particular, this means carefully handling institutional and cultural differences that affect work in different office settings around the world, and ensuring that a balance exists between global integration and local responsiveness as far as work practices are concerned.

A second important implication of the local-global geographies of producer service organization relates to the impacts on the practice of clients. In some cases, particularly when a producer service firm is not global, clients are primarily looking for a locally tailored service which is sensitive to place specificities. For example, advertisers may be expected to provide advice about engaging with clients in a particular country or region (Faulconbridge et al. 2011), and in such cases, producer service firms help ensure the local appropriateness of their clients’ practices. When global producer service firms are involved, advice is more likely to be sought on the adoption of global practices. Here, the role of the producer service firm is to help create and/or provide advice about globally recognized approaches to a particular issue. For instance, global law firms are hired to offer advice about globally recognized bankruptcy practices that correspond with local legal requirements and the expectations of international institutions such as the International Monetary Fund. Global architecture firms are used by clients seeking building designs that are globally recognized, iconic world city styles, not local vernacular architectural styles. In both cases, the firms are likely to be part of the production of these global standards, both through their work for other clients, which helps to define global practices, and through collaborations with other producer services. In all cases, however, firms must also be able to advise on the local application and implications of global standards. As such, local embeddedness and global integration are not either/or options; both are necessary to be effective in delivering the most innovative and profitable work.

Case studies

Accounting

At year–end 2012 the “Big Four,” PricewaterhouseCoopers (PwC), Deloitte, Ernst & Young (E&Y), and KPMG (in rank order) had accumulated global revenues of US$100.2 billion, were present in over 150 countries each, had upward of 700 offices each (except KPMG, which had 650+), and employed a staggering 539,724 professional staff (partners and fee earners) (Accountancy Age 2013). These global networks, along with other international
accounting organizational forms such as associations or alliances of independent member firms, have penetrated new markets and internationalized their functions by supplying clients with a range of services through corporate offices located in national capitals, regional cities, and other important metropolitan areas. In effect, the internationalization of accounting is a metropolitan-based activity where firms supply services such as: audit and assurance; corporate finance; consulting and advisory; insolvency and business recovery; and other specialist activities such as executive search, technology, corporate social responsibility, and change management.

The internationalization of accounting firms actually started during the nineteenth century, but the phase of opening up new offices outside of firms’ home countries only really gathered pace from 1945. The period between 1945 and the end of the 1980s saw the growth of firms’ office networks in Asia, Europe, North and South America, and Africa and the Middle East. Between 1975 and 1985, the leading 20 international accounting firms expanded their office networks by more than 122%, from 2026 to 4622 (Daniels, Leysh, and Thrift 1988). Importantly, this period also marked the beginning of the concentration of the sector with the establishment of the “Big Eight” through various mergers and acquisitions between UK, European, and US multinational firms. The lineages of today’s “Big Four,” PwC, Deloitte, E&Y, and KPMG, are essentially the outcome of mergers and acquisitions activity between the “Big Eight” in the 1990s, for example the creation of PwC following the merging of Price Waterhouse and Coopers & Lybrand in 1998.

The corporate office is the firm-client interface for firms in different worldwide locations, from the “Big Four” to small and medium-sized enterprises. Usually, it is wholly owned or leased by the accounting SME or global network such as the “Big Four,” or by an association or alliance of independent member firms. In 2012, the leading global accounting associations and alliances were: GGI (388 member firms in 101 countries); Praxity Global Alliance (60 in 109); The Tag Alliance (259 in 94); and Leading Edge Alliance (190 in 102) (Accountancy Age 2013). Global accounting firms (networks, association, and alliances) require a physical office in a city because clients and co-suppliers (such as lawyers or corporate bankers) expect accounting professionals to be co-located. Like many other advanced producer services, global accounting firms adopt extensive international mobility programs to ensure that professional staff circulate within and between international office networks.

Executive search

Executive search producer service firms originated in the United States in the 1960s and 1970s. They were spinoffs from management consultancy firms when consultants providing the service began setting up dedicated companies which specialized in hiring executives in senior management positions, such as chief executives and chief finance officers. Faulconbridge, Hall, and Beaverstock (2008) document how, from its US origins, executive search quickly internationalized, first through London and then more widely in Western Europe, followed by Russia, South America, and Asia in the 1990s and 2000s. Key firms driving the internationalization of the industry include Korn Ferry, Heidrick & Struggles, Russell Reynolds, and Spencer Stuart. In 2014, these four firms had between them nearly 100 worldwide offices, led by over 300 partners (plus many thousands of junior consultants). Offices are located in major world cities, reflecting the important role of other
producer services as clients. In common with other producer services, internationalization was inspired by a desire to follow existing clients who were themselves internationalizing, to tap into new markets but also, and particularly important for this industry, to respond to internationalization in labor markets as elite labor became increasingly mobile (Faulconbridge et al. 2009).

Two things are noteworthy about executive search firms. First, globalization has been achieved using a variety of organizational forms. This includes wholly owned offices, best friend networks where global firms collaborate with local firms to serve a particular market, and hybrid forms where multiple local firms adopt a global brand but remain independent entities (Beaverstock, Faulconbridge, and Hall 2015). Second, executive search illustrates the way in which global spaces of organization have worldwide effects, in this case on elite labor markets. Through internationalization, executive search firms have promoted a particular model of elite labor recruitment that involves using search firms but also assessing candidates using particular models of talent (Faulconbridge et al. 2009). As a result, executive search firms now define the “rules of the game” in elite labor markets. In addition, the internationalization of executives stands out as a distinctive example because of the way firms have sought to create a new industry through internationalization. Since executive search firms did not exist in the form they do today until the 1960s, their internationalization has spread a new model of elite labor recruitment and created a whole new producer service industry. This is significant because it reveals the way that producer service internationalization responds to existing demand for services, but also actively constructs demand to render firms profitable (Beaverstock, Faulconbridge, and Hall 2015).

Executive education

Geographers interested in producer services have not given much attention to executive education, even though education is increasingly understood as a service sector. It is, for example, included in tables showing the contribution of exports by sector to national economies. This oversight has recently been addressed on the basis that educational services, and particularly education offered to professionals after completion of their first or undergraduate degree – executive education – deserve more attention as part of ongoing work on the globalization of producer services. Indeed, executive education is central to the knowledge networks that other producer services rely upon and that shape their city-based geography.

Executive education replicates the forms of internationalization and preference for city-based offices typical of many producer service providers. For example, the size and country of ownership of firms operating in the sector are dominated by a small number of large firms predominately headquartered in the United States, including the Apollo Group, Pearson Education, and the Washington Post Company. Beneath these is a long tail of small, specialist boutiques (Hall 2015). Moreover, rather than being passive agents in education markets, and in common with other relatively “new” producer services such as executive search, executive education firms play an important role in “making” education markets by educating clients about the nature and value of the services being offered. This serves legitimate executive education and hence stimulates demand (Hall and Appleyard 2011). Finally, the highly variegated institutional contexts that give rise to distinctive national and regional education landscapes mean that educational service firms demonstrate high levels of territorial embeddedness. These institutional contexts are shaped and reproduced through
the combination of formal regulations with more informal expectations about the nature of education and styles of learning in different countries (see Hall and Appleyard 2011). Moreover, this variegation is complicated in the case of educational service firms because they have diversified beyond their initial service offering. For example, several leading business education firms, notably the Apollo Group, have entered the executive education market having previously specialized in compulsory schooling and private universities. As a result, this sector demonstrates a diverse range of organizational architectures in which the management of international offices is typically highly localized and organizationally specific.

SEE ALSO: Economic geography; Elite labor; Financial geography; Firms; Globalization; Industrial agglomeration; World cities

References


Corporate world and work: Indian women in the IT sector

Saraswati Raju
Jawaharlal Nehru University, India

The chaotic global reshuffling and flows of capital, information, goods, ideas, and human beings across subnational and transnational borders, which is unprecedented in terms of scale and the number of actors involved, is increasingly seen as a frictionless “space of flows” that is overtaking a more heterogeneous “space of places.” Such discourses imply the erosion of local differences; geographical constraints on social and cultural outcomes are seen not only as weakening but also as destabilizing the physicality of places. Places are seen as irrelevant, expunged in what some see as “the end of geography” (Graham and Marvin 1996; Graham 1998).

Undoubtedly, the traditional conceptualization of spaces as self-enclosed preconstituted autonomous containers has become highly problematic. This entry proposes, however, that places and geographical locations continue to matter – not in the sense of Cartesian space but as sites of continuous negotiation and renegotiation. This is evident from an examination of the sociospatially localized manifestations of gendered identities in the world of work, and particularly in information technology (IT) – the presumed epitome of frictionless global flows.

A topic as broad as corporations and work is beyond any single encyclopedic entry. This entry focuses on technological innovations that have resulted in the restructuring and reorganization of work epitomized by IT, IT-enabled services, and business process outsourcing (BPO). These services have facilitated time-space distanitation and redefined the boundaries of production and consumption. While they have expanded the limits of what it means to be an individual worker, there still exist regionally specific sociocultural constructs within which labor dynamics play out.

Confining the discussion to Asia in general and India in particular, the entry is organized as follows. It first discusses the outsourcing and offshoring of labor, then deals with the segmentation of global and local labor forces, before turning to the question of how nation, the work, and domestic spheres are entangled with the gendering of Indian IT labor.

Outsourcing and offshoring labor

With the collapse of the communist bloc, the entire planet now seems drawn into capitalism, with few exceptions. Some of this reflects the age-old spirit of competition in the “pursuit of profit and individual satisfaction,” but information and communication technologies provide new means of realizing global production networks (Castells 1999, 2). Growing interdependencies, mutually beneficial even as they are exploitative, make this brand of capitalism simultaneously “very old and fundamentally new” (Castells 1999, 2). Outsourcing and offshoring production, to reduce corporate taxes and avoid such mandated benefits as social security, safety measures, managerial and labor cost, is symptomatic of such globalization.
Rarely used until the 1970s, the term “outsourcing” refers to the contracting out of economic activities and labor to a third party. It started appearing in dictionaries from the mid-1990s, and has become a buzzword in contemporary globalization discourses. Contracting and outsourcing, while related, are not synonymous. Whereas contracting out suggests a passive exchange of services between corporate bodies, outsourcing is “what one company actively does in relation to another, and that means that the two managements are active agents” (Gupta 2010, 19). In short, managerial discourse dominates outsourcing. Core concerns in such a discourse, Gupta observes, are lower labor costs and the speculative shift of capital from distinctly productive to overtly speculative investment. Debates surrounding outsourcing became murkier with the spread of neoliberal globalization and the offshoring of outsourced activities. Such offshore outsourcing affects a relatively small proportion of low-profile jobs, but their relocation from developed North to developing South has generated complex normatively loaded debates and anxieties.

In this increasingly integrated global economy the IT and IT-enabled industries have received particular attention. The associated shift from Fordist mass production to more flexible accumulation has been marked by “complex transnational production networks [spread across the globe] rather than through vertically integrated multinational companies” (Upadhya and Vasavi 2008, 12), facilitated by the IT sectors. These networks may constitute a very small part of the global labor force, but are presented as “new generation ventures,” dispersing to hitherto developing countries such as India. A key aspect of this has been the emergence of virtual workspaces: call centers are such sites, one of the most rapidly expanding areas of work globally. (Incidentally, this is coupled with a heightened sense of the possible loss of Northern jobs, particularly in the United States.)

The corporatization of IT lies in its very nature. Abraham (2008) characterizes this as (i) improved telecommunication technology with concomitant cost-effectiveness; (ii) standardized software platforms that homogenize skill requirements across organizations (and geographical locations); (iii) availability of a desirable labor pool for multinational and domestic companies that is fluent in English, Spanish, French, and German in parts of the developing world; and, crucially, (iv) the substantive wage differentials between developing and developed countries (particularly the United States). Offshored labor can also take advantage of the time differences: day wages in India are far lower than (overtime) night wages in the United States, the classical argument of time–space compression. Taken together, these features make offshoring a profitable venture for corporate IT giants.

IT work is essentially based on manipulating symbols in computer systems, dubbed the “textualization” of work (Upadhya and Vasavi 2008, 12), rather than on material objects. As a result, “production and services have become dematerialised, disembodied, and divided amongst workers located in geographically distant site[s].” In these technologically mediated organizational structures, the workers are expected, in terms of training, skills, and behavioral norms, to operate in a contextually isolated manner, valorizing a singular model of “global corporate culture” (Upadhya 2008, 105).

At a more generalized scale, those benefiting from the development of the information technology enjoy material and symbolic privileges: the workforce is urban, largely young, with a high level of education and professional skills. India’s high-end airports, gated communities, and mushrooming high-technology parks in rapidly growing cities reflect elite lifestyles that
at times are better than those of their counterparts in developed countries. The gradual move up from routine coding and maintenance jobs requiring no special problem-solving skills to specialized positions in the value chain places workers at the upper end of the global economy (Radhakrishnan 2011b). Even if this is just the symbolic language of code, it has provided Indian workers with relatively better options, higher wages, and the opportunity to join the elite global network of IT workers, itself seen as an advantage (Aneesh 2006, quoted in Radhakrishnan 2011b). The resultant overlapping of an educated urban workforce with the IT professional sector has given rise to a new vocabulary of social differentiation, to a “new” middle class: new largely in terms of lifestyles marked by burgeoning conspicuous consumption (Fernandes 2000).

IT workspaces also are “new”: open, nonhierarchical, and informal work settings, in direct contrast to earlier traditionally framed or bound hierarchical and formal places (Harvey 1989). They represent a merging of diverse management styles, developed in the United States and drawing “American values of egalitarianism, teamwork and individual initiatives … into a single dominant model – an ideal type ‘global’ corporate culture that every company, worker and manager must work towards in order to compete in the global market” (Upadhya 2008, 105). Interlinking local with global and compressing time, these transnational spaces also imitate the physical ambience of the US IT culture. “The interior office environments of most large IT firms in India are designed to look like their Silicon Valley counterparts, or more accurately, to look placeless, as if they could be ‘anywhere’” (Radhakrishnan 2011a, 61; emphasis added). Radhakrishnan cites an Indian IT professional calling this feeling of being “anywhere” a “space wrap.”

In these workplaces, a worker is an individuated entity, rewarded for his/her initiatives and managed primarily through indirect and subjective techniques of cultural anchoring (Kunda 1992, quoted in Upadhya 2008, 106).

If the industrial era consisted, in terms of the labour process, of taking a population of peasants and craftsmen and bringing them into socialized conditions of labour, the information age is exactly the reversal. It is the de-socialization of labour and the increasing flexibility and individualization of labour performance. (Castells 1999, 8)

The individuated worker also is invested with socialization processes associated with becoming “global”:

… a cosmopolitan sensibility that connotes status and refinement outside the workplace, as well as added bonuses with it. Such a strategic approach to culture that begins in the IT workplace, stemming from the capitalist imperative to streamline differences for the sake of efficiency, becomes a way of understanding culture more generally, as corresponding notions of “India” take shape alongside competing notions of the “global.”

(Radhakrishnan 2011a, 55)

**Labor market segmentation**

Such processes are not without contradictions, however. Even as the global technical worker is being created by virtue of working across transnational borders, conflicting cultural images are being produced that require a segmented workforce serving capitalist interests. According to Upadhya (2008, 103), such segmentation of labor is not straightforward but

a dialectic process in which pre-existing cultural communities, gendered identities or racially marked groups are transformed into labour
forces that perform particular roles in the production process, which in turn marks these social identities with the stamp of capital. For instance, the type of work that is performed or the working conditions have implications for the reconstitution or reinforcement of masculinities or femininities … The interactions between capital and cultural identity become more complex as globalization gathers pace and both “cultures” and identities get increasingly commoditized, becoming … capital itself.

While India takes pride in its claim to be the world’s emerging power in the technically advanced IT sector, its workers are often projected as cyber coolies, placed at the lower end of the commodity chain. Such representations persist even as developing economies are increasingly internationalized and integrated into the global system. For instance, Japanese transplants in Malaysia, producing mature goods in a competitive international environment, continue to occupy relatively low-value activities in mass assembly while engaging in limited product design or process innovation. This segmentation parallels piece-rate production from the “global assembly line” export-oriented garment subcontractors in Sri Lanka and/or Bangladesh (Upadhya 2008). As Vo (2007, 198), puts it:

The globalisation processes of uneven development, interdependence between equals and those who are less equal alongside interactions of conflicting and common interest, within which state institutions are embedded, do not simply sustain a single and homogeneous pattern of firm behaviours.

In contrast, by creating segmented labor markets, firms hail and appropriate regionally specific cultural discourses and differences, to be used as a “strategy of control” (Upadhya 2008, 102). This can be seen in the engagement of child labor (especially girls) in the corporate cultivation of Bt cotton in Andhra Pradesh in south India. Here, the notion of purity/impurity associated with menstruation is used to justify employing pre-pubertal girls to pick (pristine) white cotton in the fields — quite conveniently disguising the “beneficial” aspect of employing child labor. Similarly, the overarching construct of Indian women’s primary location being the private sphere of domesticity conveniently legitimizes the progressive informalization of women’s paid labor and their concentration into home-based work — a cheap labor supply beyond the reach of protective labor legislation (Raju 2013).

**Gendered geographies of IT labor: the case of India**

Under these circumstances, a key geographical question is: How far do national contexts influence the international transfer of multinational corporations’ management practices? Radhakrishnan (2011a, 55), outlining the cultural terrain at work in the corporate world, defines its global culture as characterized by (i) a sense of placelessness and the insulation of work sites; (ii) standardized communication style and skills, which she calls low context, as clues are verbal rather than situational or relational; and (iii) nonhierarchical organizational/managerial styles. Together, these indicate a potentially empowering environment at work, which, it is assumed, will filter through into personal lives. That is, the ensuing cultural flows between these two spheres — the global culture at the workplace and the local in home spaces — would be “more fluid than they appear at first.” However, contrary to such expectations, research shows a much complex situation for multinational corporation (MNC) operations in developing and transitional economies, whereby they have to adopt a “very high level of flexibility,” and in some cases compromise and confront conflicting institutional
constructs (Vo 2007, 199). One such, often invisible, divide, is profoundly gendered, whereby women are encouraged to negotiate the merging of the two identities – being a woman and working in an MNC. The overwhelming constructs, however, continue to lock them up in “Indian middle-class cultural norms of respectability and domestic femininity” (Vo 2007, 203–204).

The IT industry invokes the rationality of merit as the marker for upward professional mobility, but the overall lower percentage of women professionals in the upper echelons does not reflect such a claim. IT women consistently explain their absence from these echelons, however, in the “language of personal choice,” ignoring “the ways in which its articulation in the everyday lives of IT professionals explicitly overlooks the specific conditions of its production in the history of India” (Radhakrishnan 2011a, 91). The improved access to employment opportunities and financial independence, seen in progressive educational attainment in accordance with the contemporary corporate world in India, should presumably enable women to exercise “their reasoned agency” (Sen 1999, xii). Does this happen, however? Does paid employment in the corporate sector necessarily increase women’s freedom and agency, especially under conditions of globalization? For the information and communication technology (ICT) and BPO sectors, the answer has to be “only superficially.” Despite the glamour and an invoked sense of articulate modernity (Tara and Ilavarasan 2011), women in IT and BPO sectors continue to operate within the paradigm of thus far and no further, the no further limit being (re)constituted by persistent gendered constructs that continue to encode women’s primary place within domesticity, even as the confining vocabulary may undergo some cosmetic changes. For example, although the line between long and erratic hours of IT-sector jobs and women’s responsibilities toward the home and family gets blurred, questions regarding priorities arise – women rarely privilege career aspirations over family life.

Krishna (2001, 412) argues that contemporary labor geographies should be placed beyond the limits of the nation-state instead of being bounded by a self-spatialization that “names itself as a destiny, a genius, a culture, a civilization, or homeland – a contiguous and identifiably discrete and separate entity.” Such a position is problematic, however, when viewed from the contextually spatialized world of women’s labor. Krishna advances a “meta-myth” (Bradley 2000) of globalization that is highly contentious in a country like India, with its irresistible “ironies and resistances” of local making (Appadurai 1996, 29). Indeed, such makings seem to have become hegemonic in India’s contemporary corporate culture where, despite women’s participation in an elite IT professional work culture, corporate norms and expectations are woven into a distinctive and self-conscious “Indian” idiom of “respectable femininity” that spills over into their everyday lives (Radhakrishnan 2011b, 195).

Kagitcibasi’s (2005) concept of the “autonomous-relational” self is helpful in understanding this. He questions the oft-posed separation of autonomy and/or agency, and their intersections with relatedness or heteronomy. For him, the separation between autonomy and/or agency and relatedness has a root in the Euro-American cultural context, with its ideological background of individualism. Autonomy and heteronomy are thus two ends of the same spectrum. An individual can have an interdependent but also independent self, in a “dialectic mutuality” or “coexistence of opposites.” These selves are embedded within societally encoded gendered constructs, which are internalized during socializing processes (D’Mello 2006, 139). In the Indian context,
there is very little scope for mutually dialectical exchanges in the constitution of women’s identities. Instead, the independent self (the articulation of which could have been drawn from education and paid work status) annihilates itself under self-efficacy and self-definition, notwithstanding the other means through which a relational self is overwhelmingly present, in the making of the self.

From her empirical study of professionals in the field of information technology in Mumbai, D’Mello seems to reiterate Kagitcibasi’s observations:

The autonomous-relational self … is operationalized among IT workers, in ways that are more dichotomous than dialectically mutual. The breadwinner ideology predominated as a central aspect of masculine identity constructs, while the relational self occupied center stage in feminine identity constructs. Individual responses as well as the coping means used by women reinforce the view that the relational self was the predominant pathway [for women] … Negotiating for power and equality in the family system [was found to be] a challenging task, threatening disintegration of family relationships … many women [in her study] compromised their own career aspirations. (D’Mello 2006, 152)

Similarly, Radhakrishnan (2009) argues eloquently that the larger consciousness surrounding women’s place in urban India, now being constructed in a family, largely coincides with the emergence of India’s new middle class. Working women, particularly those in professional IT sectors, signal the arrival of a “global” nation (Radhakrishnan 2009, 197). Notwithstanding this global image, the nationalist (work) culture continues to draw on an idealized construct of (Hindu) women based on a powerful dichotomy between the inner and the outer, material and spiritual, and home and market, which is inevitably gendered. Contrasting with its earlier emphasis on the spiritual, according to Radhakrishnan, the contemporary construction of middle-class women hinges on the notion of family (and the location of women within it). Using Bourdieu’s notion of symbolic capital and his conception of gender and the family, she proposes the concept of “respectable femininity”: the essentialized framing of women within the domestic sphere, a positioning within the family’s perceived “normalcy” “legitimated through middle-class status” (Radhakrishnan 2009, 197). Placed within this framework, the symbolically authoritative dominance of the middle class allows the conception of “family first” to bear on the national consciousness, opening “up space for grappling with the embeddedness of gender” (Radhakrishnan 2009, 197). This shapes the labor market outcomes, even for those in new generation employment such as IT, in terms of upward mobility within the profession, as women internalize familial priorities over their professional aspirations beyond certain thresholds, despite individual negotiations and interlinkages with the global economy. IT women thus enact a highly competent “professional” femininity that continues to remain “markedly Indian” (Radhakrishnan 2009, 200–201, 209).

In summing up, a few priorities for future research will be identified. This “markedly Indian” manifestation of gendered constructs is intercepted, in a nuanced manner, by localized differences in the overarching patriarchal structures that prevail across India. The IT sectors are predominantly located in the southern Indian cities of Bangalore and Hyderabad—social spaces that have been relatively more gender-egalitarian than elsewhere, especially in comparison to northern Indian cities and states (Raju 2011). Other hubs are emerging, such as Faridabad, closer to Delhi and located in a highly patriarchal Northern social space. It will be interesting to
see how the microenvironment of these spaces works to further constrain and/or encourage professional women as they negotiate/contest well-entrenched, albeit localized, gendered structures. More generally, what kinds of strategies do women adopt in distinct subnational regional contexts in this regard?

The IT industries do provide emancipatory spaces to women, with the potential also to enhance women’s bargaining position in domestic spaces. However, there is very little geographical research examining the discursive interfaces between public and private spheres, in terms of the work–life balance, strategies and negotiations, domestic struggles in daily lives, and so on.

The heterogeneity of women workers also needs more attention. Paul’s (2013) research in the city of Kolkata, based on a purposive sample of about 360 women working in the IT and IT-enabled services contrasted with women workers in traditional sectors such as teaching, underlines the heterogeneity of women workers marked by their age, education, and socioeconomic background, and their differentiated responses to various professional and familial challenges. Paul’s study also interrogates how questions of emerging new places of consumption such as shopping malls, and women’s spatial interaction with them, intersect with the sectors they work in. Few professional women have been able to break through the glass ceiling in their careers. The conditions propelling them to the positions they occupy, the replicability of those conditions, as well as the institutional, social, and structural changes required, are further unexplored areas with exciting possibilities for research, by geographers in general and in the Indian context in particular.

SEE ALSO: Corporate identity; Cultures of work; Gender, work, and employment; Glocalization; Labor geographies and the corporation; Labor market

References


Corporations and global trade

Martijn J. Burger
Erasmus University Rotterdam, Netherlands

One of the characteristic features of contemporary globalization is the historically unprecedented increase in foreign trade. Between 1970 and 2013, worldwide trade rose by almost 1000% in real terms. This outpaced growth in worldwide real gross domestic product (GDP), which only quadrupled in the same period (Figure 1). In the economic geography and international economics literatures, this remarkable growth in global trade has been attributed to several social and economic developments. These include increased economic wellbeing, declining transportation costs with advances in shipping and information and communication technology, trade liberalization in terms of a reduction in tariff and nontariff barriers to trade, and the increased outsourcing of production processes across borders and the related rise of global value chains. Several studies also have shown that trade has been an important source of growth for many countries over the past decades (Badinger 2005; Bradford, Grieco, and Hufbauer 2006).

Nonetheless, there are signs that corporate internationalization remains restricted to the “happy few” (see Mayer and Ottaviano 2008) in that firm participation in global trade is surprisingly uncommon. Data obtained from the Organisation of Economic Co-operation and Development (OECD) Eurostat Trade by Enterprise Characteristics (TEC) shows that only 4.5% of US firms export abroad. A similar share of firms from the European Union (EU) member countries engages in intra-EU trade, while only 2.7% of EU firms export to countries outside the EU. The bulk of trade is driven by only a handful of large corporations, exporting many products to many countries. The top 50 US exporters by export value account for about 30% of US export value and the top 50 EU corporations account for about 40% of intra-EU export value and 30% of extra-EU export value. Likewise, the top 1000 exporting firms account for approximately two-thirds of trade in both the United States and European Union. At the same time, firms’ propensity to export varies by sector. For the United States and Europe, firms in the final goods sector generally import and export more frequently than firms in the intermediate goods sector (Bernard et al. 2007); goods generally are more frequently exported than services (Burger et al. 2014).

Conventionally, trade theories seek to explain the volume of trade between countries in terms of either differences in relative productivity across different industries and in the availability of specific types of labor and physical capital goods or by increasing returns to scale, specialization, and consumer demand for variety (Helpman and Krugman 1985). As Helpman (1999) notes, these theories are reasonably able to explain the architecture of the global trade network, but depend on the oversimplified assumption of the representative firm. In these theories, a model firm is used to explain the trading behavior of all firms within an economy, an approach that hides the differences between firms that also demand theorization.

Recent work on trade, however, increasingly recognizes the fact that global trade is
concentrated within a few firms, and that the architecture of the global trade network is mainly shaped by *firms’ participation in trade* (the decision to trade, the number of products traded, and the number of countries with which firms trade) rather than *the average amount of trade per firm*. Further, it is becoming commonly accepted that firm heterogeneity is associated with firm participation in trade, which in turn shapes the pattern of international trade flows. In particular, reducing barriers to trade (e.g., through trade liberalization or decrease in transportation costs) alters the organization of firms through adjustments they make in the range of products offered in foreign markets and their decisions about whether to serve foreign markets through trade or overseas facilities (i.e., FDI) or to outsource foreign production (Bernard *et al.* 2007).

Accordingly, since the early 2000s attention has gradually shifted from countries and industries to firms and products as the unit of analysis. This shift has been accompanied by increased availability of firm-level trade data as well as product-level trade statistics in Eurostat and the United Nations COMTRADE databases. Within the empirical literature on corporations and global trade, two topics are commonly addressed: (i) trade participation and firm characteristics and (ii) geographical scope of trade.

Studies examining firm participation in exports have found that exporting firms are generally larger and more successful. Focusing on inter-firm differences within industries, Bernard *et al.* (2007) find that, compared to nonexporting firms, exporting firms are larger in employment and sales, have a higher capital stock to total employment ratio, a higher nonproduction workers to total employment ratio, higher value added per worker, and pay higher wages. Although most of the literature focuses on exports, similar results are found for importing versus nonimporting firms (Bernard *et al.* 2007).

Following Melitz (2003), the direction of causality is not simply from export performance to firm characteristics; more productive firms are better able to overcome the barriers to foreign trade in terms of fixed export costs (i.e., costs that exporting firms incur irrespective of the amount they export). Thus the finding that exporting firms are more capital-intensive and skill-intensive can be related to the fact that exporting forces firms to make use of more advanced technologies in order to cover these costs (Burstein and Vogel 2010; Sampson 2014). Exporting firms also tend to produce higher quality products, typically requiring more skilled workers (Verhoogen 2008), and also higher marketing costs, requiring a more intensive use of nonproduction employment (Arkolakis 2010). Melitz also shows that globalization increases aggregate productivity: lowered barriers to trade through trade liberalization or reduced transportation costs change the landscape of firms. High-productivity firms tend to survive and expand whereas low-productivity firms contract or exit. This shakeout improves average productivity within industries.
The geography of trade

Apart from the fact that most firms do not trade, the majority of exporting firms export only to one country, yet the majority of trade is concentrated in a few corporations exporting many products to many countries. The fixed export costs firms incur when entering a certain country, for example, related to cultural differences or compliance with regulation, vary across export destinations, as do demand for their exports (key to ensuring that profits cover the fixed export costs).

As a result, countries differ in terms of their popularity as export destinations. Eaton, Kortum, and Kramarz (2011) find that: (i) firms are more likely to export to larger markets, (ii) firms exporting to less frequently targeted countries are more likely to export to more frequently targeted countries, and (iii) firms exporting to less frequently targeted countries also sell more in their home country. According to Helpman, Melitz, and Rubinstein (2008), the variation in entry costs across countries also explains why there is no trade at all between many countries. Low demand levels and high fixed export costs can make exports infeasible even for the productive firms that engage in exporting.

SEE ALSO: Free trade zone; Trade, FDI, and industrial development; Trade and regional development

References


Further reading


Corporations or large business organizations are a major focus of interest in economic geography. While they have a longer history stretching back to the nineteenth century in some cases, multinational corporations (MNCs) became recognized as major economic actors in the 1960s and 1970s as the international economy became increasingly integrated. More recently, much of this research has been conducted under the rubric of globalization, with MNCs becoming emblematic of the increased mobility of capital across space, which enables them to shift resources and operations between locations in response to changing economic conditions. Informed also by the neoliberal nostrums of competitiveness and entrepreneurialism, these strategies encourage localities and regions to compete for investment on the basis of cost and flexibility, sparking fears of a “race to the bottom” as wages and social entitlements are ratcheted downward.

The local state also emerged as a distinct object of analysis in the 1970s, defined by one prominent British-based researcher as “both the local presence of the national state agencies, such as the police and courts, and those functions, such as housing and education, that are defined as local responsibilities” (Cockburn 1977, 363). The concept of the local state has subsequently become broader and more diffuse, operating through a range of networks and partnerships involving private interests and voluntary sector agencies alongside traditional government bureaucracies, as signaled by the adoption of the term “local governance.” The geographical definition of the local state can also be somewhat elastic, sometimes encompassing regional government and agencies, suggesting that it might be more accurately defined as the subnational state.

Like their national counterparts, local states play an important role in coordinating and mediating social relations, such as those between capital and labor, but also across other axes of difference such as race, gender, and age. Local states are neither neutral arbiters between the often competing claims of different groups nor the exclusive instruments of any particular group. Rather, as suggested by the strategic-relational approach developed by Jessop (1990), local states incorporate an unevenly structured set of institutional capacities or powers that are more or less open to different groups, depending on the strategies that they adopt toward it. Thus, local and national states have become increasingly skewed and “tilted” toward business and economic interests since the 1980s through the adoption of neoliberal strategies of competitiveness, while the interests of labor and lower-income residents have often been marginalized. At the same time, however, local states remain important sites of contestation and struggle as particular initiatives can spark local opposition and protest.

The nexus of relationships between corporations and local states has been an area of great interest in geography since the 1980s as local states have adopted “entrepreneurial”
strategies based on place promotion and the attraction of external investment. This reflects local states’ roles in mediating and coordinating social relations, particularly in terms of forming partnerships with local business interests and seeking to mold and package local assets such as labor according to the needs of external corporations. Here, local states are key players within the development of local labor control regimes (LLCRs), defined by Jonas (1996, 329) as "relatively stable institutional environment[s] for capital accumulation, albeit circumscribed by struggle and the periodic crises to which capitalism is prone." LLCRs typically involve cooperation between local state institutions and employers to regulate local labor markets and manage wage and work expectations, fostering reciprocities between the realms of production, reproduction, and consumption.

Over time, research on the corporations–local states nexus has broadened away considerably from the original concern with local economic development to cover a wider range of state practices, corporate strategies, and forms of social and political mobilization and struggle. This is reflected in the remainder of this entry, which is organized into three main sections covering corporations, local state strategies, and subnational economic development in the Global North; work on the intersections between labor, corporations, and local state practices in the Global South; and debates on social movements, resistance, and corporations that help to link issues and concepts in the Global North and South.

**Corporations, local states, and economic development**

The effects of economic restructuring and deindustrialization in the 1970s and 1980s created severe unemployment and poverty in many industrial regions of Western Europe and North America. One of the main ways in which local states responded to deindustrialization was by developing new strategies to promote economic development and attract external investment. This is part of a profound shift in the purpose and ideology of local state intervention, famously encapsulated by Harvey (1989) as a shift from urban managerialism to entrepreneurialism. Managerial approaches, prevalent in the 1960s and 1970s, viewed the role of the local state as supporting collective consumption and social reproduction (Cockburn 1977). Urban entrepreneurialism, by contrast, saw localities focus on the need to generate growth and employment. This can be seen as supporting the development of new LLCRs emphasizing the need for workers to be flexible and responsive to the needs of investors.

Urban entrepreneurialism is associated with the development of new forms of local governance based on partnerships between local authorities and business. In many cities in the industrial Rust Belt of the United States, such as Pittsburgh, Detroit, and Boston, local governments and local business interests involved in the exchange of land and property such as developers, construction companies, and realtors came together to promote and redevelop the city through the attraction of external investment. In a highly influential formulation, these interests were described as “urban growth machines,” supported by their shared interest in the expansion of the local economy as the ultimate source of tax revenues and profits (Logan and Molotch 1987). Urban entrepreneurialism is closely associated with intense interurban competition as cities and localities have become “hostile brothers” (Peck and Tickell 1994), engaged in often similar strategies designed to attract corporations and new facilities to their areas.
In the United States, many cities created public–private partnerships to leverage private sector investment and to compete for limited redevelopment funding from state and federal governments. Another key instrument is the establishment of special purpose, business-oriented agencies to regenerate an area. For example, urban development corporations were set up in certain run-down inner city areas of England and Wales in the 1980s. Other measures designed to attract investment by external corporations include the designation of special economic areas or zones offering tax and other financial incentives and simplified planning procedures. A good example of this type of measure are the enterprise zones and special purpose districts established in the United States by local government and the enterprise zones created in the United Kingdom by national government, first in the 1980s and again since 2010. In addition, tax increment financing has been widely utilized in the United States to fund the redevelopment of blighted districts, allowing local states working in partnership with private developers to raise finance by selling bonds which are subsequently paid off by the additional property taxes generated by redevelopment.

Urban entrepreneurialism is also associated with the creation of a new urban landscape, based on the regeneration of former industrial areas (e.g., Baltimore’s Harborplace, London’s Docklands, and Bilbao’s Abandoibarra area). These strategies have created new jobs in sectors such as tourism, retail, culture, and leisure, although many of these jobs tend to be low-paid and part-time, compared to the often better-paid full-time manufacturing employment lost through deindustrialization. In many cases, the benefits of regeneration have failed to reach low-income groups in disadvantaged inner-city neighborhoods or outlying public schemes or projects.

The second strand of literature covered in this section is research on the role of subnational governments and development agencies in attaching and “embedding” inward or foreign direct investment (FDI), focusing mainly on the regional rather than the local scale. Attracting FDI has been a key component of local and regional economic development policy since the 1960s, particularly in old industrial regions suffering from the contraction of traditional industries. The experience of high levels of unemployment has supported the introduction of new forms of labor control involving highly selective modes of recruitment by incoming investors, while also requiring substantial state support for the process of labor market adjustment.

Early studies of FDI in developed economies, however, suggested that inward investment could lead to relations of dependency as local economies became increasingly subservient to the needs of large MNCs. In the 1970s, the term “branch plant economy” was coined to reflect how some of the older industrial regions of Western Europe and North America were being repositioned within the wider space economy of capitalism. The concept is defined by functional truncation due to a lack of higher-status activities such as strategic planning and research and development, limited linkages between MNC plants and the local economy, with many components and materials being imported, and a high degree of external ownership and control.

The negative branch plant stereotype became increasingly questioned in the 1980s and 1990s, however. Compared to the low-skill assembly line activities of previous investments, the new wave of inward investment into the United Kingdom and United States from Japanese and German companies was geared toward the creation of locally integrated industrial complexes that were more firmly tied to their host
CORPORATIONS AND LOCAL STATES

regions. This more optimistic scenario hinges around a number of organizational changes taking place in MNCs themselves in response to globalization, characterized as a shift away from centralized bureaucratic hierarchies to flatter and more decentralized structures, meaning that more decision-making powers and higher-level operations are devolved to local branch management. In these circumstances, it is argued that investment can become strongly embedded in host regions. For instance, the Japanese car manufacturer Nissan set up a manufacturing plant in the northeast of England in 1986 which has expanded to employ over 5000 workers directly and to support a significant regional supply chain. At the same time, regions can use the global knowledge networks of MNCs to upgrade and improve their competitiveness. For example, Japanese branch plant investments in the Great Lakes region of the United States during the 1980s were seen as instrumental in rejuvenating the regional economy, through the transfer of “best practice” management and production techniques to local firms.

The evidence to support such claims is mixed, however, with research finding little evidence of increasing embeddedness in peripheral regions of the United Kingdom (Phelps et al. 2003). At the same time, case study analysis has pointed to the increasing volatility of inward investment and the enduring significance of global market trends and corporate decision-making processes. These long-standing concerns have been reinforced by some high-profile plant closures in peripheral regions of the United Kingdom in recent years, together with the relocation of many facilities to the lower-cost regions of Central and Eastern Europe. The electronics sector has proved particularly vulnerable, with both the German firm Siemens (1100 jobs) and the Japanese firm Fujitsu (600 jobs) closing semiconductor plants in the northeast of England in 1998, while Scotland’s “Silicon Glen” lost around 40% of its jobs between 1998 and 2006. Companies that have relocated from the United Kingdom’s regions to take advantage of lower costs in the Czech Republic include Japanese firm Matsushita (from Wales, costing 1400 jobs) and US firms Compaq (Scotland, 700 jobs) and Black and Decker (northeast England, 600 jobs).

The role of local and regional states in attracting and managing inward investment from corporations can be understood in terms of the concept of “strategic coupling,” which is closely associated with the global production network (GPN) approach in economic geography. Strategic coupling occurs through the activities of local and regional state institutions in shaping and molding regional assets to fit the strategic needs of lead firms in GPNs. Such activities include providing financial support; preparing sites and assisting with infrastructure and services; coordinating the supply and training of labor; developing local supply chains; working to generate technology transfer; and offering a host of “aftercare” services to existing investors. In some cases, this can result in “corporate capture,” whereby MNCs are able to appropriate regional institutional capacities to serve their interests at the expense of other actors such as indigenous firms and workers. This often reflects MNCs’ capacities to engage in regulatory arbitrage, effectively playing off localities against one another to achieve the most favorable deal in terms of financial and institutional support.

Strategic coupling should be seen in dynamic terms, focusing attention on processes of decoupling and recoupling (MacKinnon 2012). Decoupling involves rupturing the relationship between a region and a GPN through plant closure and disinvestment. Recoupling, by contrast, is related to the attraction of repeat investment,
recombining existing plants and regional assets with a new round of investment. Local states will often support local plant management in their efforts to attract repeat investment through a range of direct and indirect services; decoupling typically involves local states undertaking an ameliorative role often involving the establishment of task forces to deal with the effects of closure, particularly in terms of providing support to redundant workers.

The concept of strategic coupling reflects the adoption of a largely positive perspective on FDI and regional development, stressing the scope for cooperation between local states and lead firms in GPNs. As has been emphasized, this is consistent with the strategies of many local states to attract and retain investment. In certain circumstances, however, local states can also play a redistributive role, redirecting resources toward particular groups and areas. This is particularly apparent in the mining sector, where the unprecedented commodity boom of recent years has generated windfall profits for corporations. This has fostered a renewed, if highly uneven, resource nationalism and “resource regionalism” in the form of increased political pressures for revenues to be used to benefit local communities. In the state of Western Australia, for instance, increasing royalty rates levied on the major mining corporations have been distributed to the mining regions of the state through the Royalties for Regions fund. Decoupling and plant closures can also spark contestation and resistance from labor and local residents, exerting pressures on local states. In many cases, local states remain distant from local protest campaigns, but there are examples of local (and national) state personnel supporting these campaigns, such as that opposing the proposed closure of the Nestlé-owned plant of St Menet in Marseille.

Local labor control regimes, corporations, and uneven development

Work on strategic coupling usefully draws attention to the multifaceted and contested nature of relations between corporations and local states. An understanding of such relations can be extended through thinking about some of these relations historically and in the contexts of the Global South. Such a focus can shed light on important aspects of the uneven histories and geographies of cooperation between corporations and local states. Examining the shifting relations between corporations and particular LLCRs offers a useful way of tracing different relations between capital and labor in diverse spatial and temporal contexts.

Colonial labor geographies were shaped by particular, often racialized, relations between corporations, imperial state formation, and labor organizing. This can be exemplified by shipping companies such as Elder Dempster, which held “the monopoly of carrying freight, specie and mail between the UK and the west coast of Africa from about 1910 until the 1930s” (Sherwood 1994, 130). Elder Dempster was just one of a number of shipping companies which, together with local state actors in port cities such as Cardiff and Liverpool, and the National Union of Seamen, practiced unequal and racialized labor practices. The company was notorious for the “gross discrimination” it practiced against seamen of colour “with the connivance of the unions and the government” (Sherwood 1994, 130). Corporations such as Elder Dempster, then, shaped LLCRs which were in part concerned with regulating flows of mobile labor between port cities in West Africa and the United Kingdom. These exclusionary practices were in turn contested by organizations such as the Colonial Seamen’s Association.
In the postcolonial period such exclusionary and racialized geographies of labor were challenged, albeit with uneven consequences. Independence had important effects on the geographies of corporations, not least the collapse of shipping lines in the postindependence context. Further, many postcolonial nations adopted economic policies such as import substitution models directed at “national and subregional markets” (Pérez Sáinz 1999, 116–117). The Cold War context of postcolonial state formation meant that LLCRs could be shaped in important ways by attempts to circumvent left-wing political interventions at the level of the local state. The AFL-CIO, for example, began in the 1950s to fund segments of the Guyanese labor movement that were hostile to the moderate left-wing political leader Cheddi Jagan for fear that he might regain power in Guyana and “turn British Guiana into another Cuba” (Herod 2001, 186). These strategies used new local housing projects “to sway the local working class’s political loyalties” (Herod 2001, 186).

Links between anticommunist US foreign policy and right-wing military dictatorships in Latin America were significant in shaping the emergence of neoliberalism in the 1970s and 1980s, most famously in relation to the neoliberal economic project of Augusto Pinochet in Chile after the ousting of Salvador Allende in September 1973. Anticommunism also shaped many LLCRs in pernicious ways during the 1970s and 1980s in contexts such as Guatemala. Thus in the 1970s and 1980s workers in the Coca-Cola plant in Guatemala City, who were organized through the Coca-Cola Bottling Company Workers’ Union (STEGAC), were at the forefront of struggles to maintain and reproduce a trade union movement in the face of severe repression by successive military dictatorships. These struggles took place in a context where the local state was deeply enmeshed in relations between the military and capital, linkages which facilitated the extreme and violent repression of STEGAC. While the immediate postwar period had seen a massive upsurge of trade unionism in Guatemala, the US-sponsored coup against Jacobo Árbenz in 1954 ushered in decades of repression which drove the labor movement in Guatemala underground. Workers in the Coca-Cola franchise bottling plant in Guatemala City were central in resisting this repression, especially during the 1970s and 1980s:

[STEGAC] suffered more violent attacks in the 1970s than any other single union. Because of this, by 1978 the union had gained potent support from international labor and human rights groups that allowed it to survive in the late 1970s and to persist under military rule in the early 1980s. (Levenson-Estrada 1994, 176).

Such dense linkages between a transnational corporation and particular local political contexts emphasize the need to think about the diverse ways in which corporations and local states intersect. Understandings of processes of neoliberalization, however, have often been narrated in a rather undifferentiated way. Thus Sangameswaran argues there is an important need for ongoing processes to “be analyzed not only with reference to neoliberalism but also other processes at the local, state and national level” (2009, 230). An attention to local state strategies can be a significant contribution to such an approach to neoliberalization. The rejection of economic policies such as import substitution has been integral to economic strategies such as India’s New Economic Policy and embrace of neoliberalism since 1991. This had very tangible effects such as allowing companies like Coca-Cola to operate in the country. Neoliberalization has also been produced through the strategic use of spaces such
as “export processing zones” (EPZs) or “special economic zones” (SEZs).

These zones are configured to directly expand “exports with the introduction of special enclaves subject to duty-free imports and extensive tax concessions. The intention of these quasi-autonomous SEZs is to attract foreign direct investment to set up manufacturing and other enterprises for the international market” (Oza 2010, 243). These sites have been seen as significant for the ways in which they bypass local government. There is evidence that panchayats (local assemblies) and urban local bodies are marginalized from decision-making processes in the siting of SEZs. The incentives enjoyed by corporations locating in SEZs can also have further negative consequences on local states, not least in terms of lost tax revenue.

The uses of SEZs and EPZs have had particular effects in relation to different sectors of the economy, such as the garment sector. They also depend on particular gender divisions of labor. Wills (2005, 1) notes that the “women workers upon whom the garment industry depends for its wealth are largely invisible, increasingly distanced from the major brand-name retailers in the industry by complicated chains of subcontracted production.” Yet these processes can intersect with local labor market conditions where labor is not entirely without some agency.

In the context of free trade zones in Honduras and Costa Rica, Pérez Sáinz argues that managerial hiring strategies intersected with the “contact networks of the workers themselves,” observing that “cases have been identified in which firms have implemented relocation strategies with the aim of securing local labor markets.” Thus “one Costa Rican firm set up its second plant in a rural area in the Central Valley, claiming that in its original location, a Free Trade Zone, the hiring of labor is competitive and sometimes disloyally so” (Pérez Sáinz 1999, 121).

While SEZs and EPZs are spaces where corporations bypass engagement with local states, strategies of neoliberalization have also entailed much more direct linkages between the local state and corporations. This partly reflects the broader reorientation of the local state toward business interests in recent decades. A case in point is the privatization of key utilities in India as liberalization intensified after 1991. Thus in 1992 the Maharashtra State Electricity Board signed a memorandum of understanding with Enron, setting in motion the Dabhol Power project (run by the Dabhol Power Corporation (DPC)), the largest corporate-led commercial venture in Indian history, which also involved Bechtel and General Electric (all US corporations) as venture partners. This was part of a shift from a broadly Keynesian to a liberalized model of electricity supply; until 1991, the Indian power industry had been nationalized. Priya Sangameswaran has observed similar dynamics in the water sector in Maharashtra. She notes that the emphasis of discourses of self-sufficiency in urban local bodies is indirectly providing a push for private sector participation, and compromising the process of decentralization that is also simultaneously under way. Further, as in neo-liberal projects elsewhere, there is an emphasis on changing subjectivities and this has resulted in symbolic changes in the water sector. (Sangameswaran 2009, 228)

Such processes of neoliberalization have been contested. Thus the proposed site for the DPC power plants was the district of Ratnagiri in the state of Maharashtra, but opposition to “the project was triggered by local environmental concerns” generating significant resistance to this major neoliberal project even before it gained official clearance (Ahmed 2012, 1065). This resistance was partly mobilized in relation to local state actors. Thus Ahmed (2012, 1066) notes that activism “in the local communities
CORPORATIONS AND LOCAL STATES

forced a public hearing to be held at the Maharashtra State Secretariat in Mumbai on the 8 and 9 September 1994.” This raises significant questions about the relations between corporations and the local state and articulations of resistance, to be explored in the next section.

Corporations, militant particularism, and progressive localism

Waquar Ahmed’s account of the resistance to the Dabhol Power project argues that the anti-Enron protests started as a “form of environmental and livelihood militant particularism” (2012, 1067). He notes, however, that in this “struggle against a mighty corporation, local communities jumped scales by accessing spaces beyond the local to get publicity for their concerns” (2012, 1067). Through such strategies the campaign sought to connect the “particular struggle of the farmers and fisher-folks to a general struggle against the nexus of exploitative and manipulative global corporation and repressive state” (Ahmed 2012, 1073). The Bharatiya Janta Party (BJP) proved successful, however, in co-opting and colonizing such opposition. Ahmed details how an alliance of Shiv Sena with the BJP, the state-level and the national-level right-wing Hindu fundamentalist parties respectively, directed anger at a Congress state government that was seen as corrupt and supportive of Enron.

In framing the struggle against Enron in these terms, Ahmed draws on Raymond Williams’s account of the relations between militant place-based politics and wider claims for socialism. Williams’s account of “militant particularisms” was developed by David Harvey in his account of the “struggle to create a public and political campaign over the future of Rover’s car plant at Cowley” in east Oxford (Harvey and Swyngedouw 1994, 25). In this context Harvey and Swyngedouw drew attention to a campaign to keep a General Motors plant open in Van Nuys, California, in 1988 which combined “shop-floor struggle” and “wider community movements in this campaign” (Harvey and Swyngedouw 1994, 24). They argued that the way the Van Nuys campaign brought together “unlikely partners” such as “civil rights campaigners, local business people who would suffer from the closing down of the car plant, the women’s movement and local politicians” could serve as a potential model for opposition to the closure of the Cowley plant.

For Harvey, tensions were raised that made difficult a straightforward defense of the Cowley plant, because of its problematic environmental impact and the fact that these were “shit jobs.” Harvey and Swyngedouw’s account of the Cowley plant dramatizes the consequence of unfettered power of mobile capital on communities which they construct as much more limited and settled. Critics of this position, such as Noel Castree and Jeremy Anderson, argue that Harvey and Swyngedouw’s position overemphasizes the unfettered power of corporations in relation to local workers (Anderson 2009; Castree 2000). Yet such debates have frequently rather ignored the agency of the local state as well as the workers.

Indeed, while the local state has often allied itself directly with transnational corporations against local forms of resistance by labor and social movements, there are important cases of local state actors playing significant roles in relation to such resistance. Thus, when workers in Uddingston near Glasgow occupied a factory of the US multinational Caterpillar, protesting plans to close the plant, elements of the local state offered different forms of support to the occupation. The occupation, which lasted for 103 days in 1987, was a high-profile political event in Scotland, with local authorities playing important roles in supporting the occupation.
Glasgow District Council “donated £1,000 directly to the campaign while Councillor Lawrence McGarry of Strathclyde Regional Council’s influential economic development committee promised £10,000 towards the funding of an independent feasibility study of the plant” (Woolfson and Foster 1988, 91–92). There was also strategic use of the local state’s role in welfare provision to support workers involved in the occupation. Motherwell District Council, where there was a district unemployment rate of 18.63% in the month of the closure announcement, spearheaded

the implementation by local authorities of Section 25 of the Housing Benefit Act, which allowed local authority housing departments to give rent rebates to tenants who were Caterpillar workers, as they had also done during the miners’ strike [in 1984–1985]. Those who had mortgages and qualified under the Act, generously interpreted by Strathclyde region, also received rebates. (Woolfson and Foster 1988, 91–92)

The support given to the Caterpillar occupation demonstrates some of the ways in which key local state actors can facilitate and offer important forms of support to militant place-based resistance to events such as plant closures. There also are important examples of the local state adopting a more proactive role, seeking to contest some of the underlying dynamics which have led to such processes of closure. Thus in the mid-1980s, in a context of significant unemployment especially in sectors such as manufacturing, the Greater London Council (GLC) under Ken Livingstone and the Greater London Enterprise Board (GLEB) sought to develop a strategy which they termed “Restructuring for Labour,” which “threatened to curb the power of some of London’s employers” (Massey 2007, 83). Doreen Massey, who was closely linked with this strategy, has argued that it offered a dynamic response to the ways in which “capitalism itself was changing, away from the cost sensitive mass production of Fordism towards smaller-batch and higher quality production.” It also was aimed at producing “room for manoeuvre for improving the conditions of inner-city labor while still remaining competitive – in other words a less degrading strategy for working-class jobs than that on offer from the Thatcher government” (Massey 2007, 83). For critics such as Peter Nolan and Kathy O’Donnell, however, the project overemphasized the agency of local management in London and downplayed the power of large multinationals. This strategy was also produced in a context of entrenched political conflict between particular local state projects and the Thatcherite national state, which eventually resulted in the abolition of the GLC.

These political engagements have long shaped Massey’s influential theoretical interventions on the terms on which places are understood and mobilized politically. In particular, her arguments about a “politics of place beyond place” allow local state politics to be conceptualized in ways which dynamically relate them to wider, unevenly shaped, flows and connections. These interventions are important for locating the local state as part of a strategic engagement with the terms on which corporations act both within and between places. It also opens up possibilities for understanding the ways in which corporations might be challenged as part of post-neoliberal social relations.

There is now a renewed focus on the role that local states can play in fostering alternative development trajectories, in contexts defined by “austerity.” This can involve fostering a distinctly critical attitude to the terms of engagement of corporations with place. It can also emphasize how place-based movements and progressive factions of the local state have the potential to combine in productive ways. A significant
example is recent movements for the remunicipalization of electricity and water provision in major German cities such as Berlin and Hamburg. Beveridge, Hüesker, and Naumann (2014, 66) note that, after 12 years of water privatization in Berlin, a campaign involving local state actors, social movements, and left-wing political parties led to the German energy utility producer RWE agreeing in principle to sell its 24.9% share of the Berlin Water Company back to the city of Berlin in May 2012. The following year, after initial threats of legal action against RWE, the second private partner, French utility Veolia, also indicated that they would be willing to sell their 24.9% share.

They argue that these “decisions mark a remarkable turnaround,” creating the conditions for what was at first presented by various Berlin governments as impossible, then simply unrealistic and irrational – the remunicipalization of Berliner Wasserbetriebe (BWB). This also emphasizes how in a Europe marked by political strategies of austerity there remain possibilities for local state actors to offer different political responses to the postcrisis conjuncture.

Conclusions

This entry has engaged with the multifaceted relations between corporations and local states. It has delineated some of the key ways in which these relations have been theorized by geographers and other social scientists, and discussed the different terms on which these relations are constructed in different contexts. Further, attention has been drawn to the ways in which these relations can be configured in different ways and actively challenged, even if it would appear that local states have been reconfigured and reoriented toward business interests and corporations since the mid-1980s. The entry has also sought to engage with the mutating/changing terms on which these relations have been produced, particularly in relation to the evolution of neoliberal projects over time and across space.

In “austere” times it is tempting to see the possibilities for “local states” to challenge corporations as severely limited. In many cases this is clearly the case. However, it would be wrong to subscribe to a straightforward teleology of increasing neoliberalization. Austerity is itself an uneven project, with geographically uneven consequences, partly in relation to differentiated local contexts, often having the most adverse effects on disadvantaged localities which lack the resources and connections to navigate this latest form of crisis. At the same time a withdrawal of local state support and investment is not the only possible strategy within a political conjuncture defined by “economic crisis,” as indicated by the remunicipalization of services in German cities. This suggests that there are possibilities for forms of “progressive localism” that seek to and can involve the local state in challenging uneven power relations and, in certain contexts, the power of corporations, and in so doing generate outward-looking strategies that build connections with other places, through, for instance, various forms of public–public partnership.

SEE ALSO: Branch plant economy; Export processing zones; Global production networks; International division of labor; Labor geographies and the corporation; Local development; Local state; Corporatization of race: an American case study

References


Corporations and the law

Joshua E. Barkan
University of Georgia, USA

Much of the geographic research on corporations has examined the ways companies organize the production and distribution of goods across the spaces and territories of capitalist economies. Legal rulings and regulations are integral to this process, yet the geographic study of corporations and the law remains in its infancy. Nevertheless, the law provides an important avenue for research on corporate geography, particularly under conditions of economic globalization. Because laws and legal systems are almost always produced, implemented, and enforced within the boundaries of specific political jurisdictions, the attempt to shape and regulate socioeconomic relations that are increasingly transnational is fraught with tensions and contradictions. For this reason, laws play an important role in constituting the uneven landscape of contemporary corporate capitalism.

Although contemporary interest in corporate geography focuses almost exclusively on business firms, corporations were legal entities long before they were used for capitalist enterprise. In Western law, legal definitions of corporations date back to the ancient fellowships, leagues, societies, colleges, and guilds of classical Greece and Rome. These associations were created for a wide variety of social purposes including the carrying out of religious rituals, the organization of social and occupational classes, and the undertaking of public duties and administration (Radin 1910). Moreover, the religious conception of corporations as collectives of believers persisted after the rise of Christendom, with a variety of Christian sects organized in corporate forms. The Roman Catholic Church itself was conceptualized as a corporation under the Pauline term corpus Christi, which united believers into a unified body with Christ as its head. Ernst Kantorowicz (1957) has described how this religious image of the church was transformed during the twelfth and thirteenth centuries into the corpus mysticum. This new conception incorporated the sociological and administrative capacities of the church into a single body that encompassed “Christian society as composed of all the faithful, past, future, present, actual and potential” (Kantorowicz 1957, 195). Most importantly, it was this image of the corporation that provided a template for early modern political and legal theorists to conceptualize secular states as corporate political bodies led by secular rulers and monarchs rather than Christ.

The treatment of associations as corporate bodies meant that these collective institutions were treated as persons within the scope of the law. Although there are continuous debates concerning the nature of corporate personhood as either real (and therefore equivalent to natural persons) or fictitious (and therefore created by a legal authority), the organic metaphor established a collective subject of the law whose existence transcended the life or death of any individual member. What thus unites Roman collegium, monastic orders, early modern states, and even modern business corporations is that any individual member of these corporate bodies can leave these associations without threatening the legal standing of the association as a whole. In this sense, the primary importance of the
CORPORATIONS AND THE LAW

corporation is to provide a consistent legal designation for a group across time that enables them to pursue and accomplish some legally defined end.

If the temporal aspects of the corporate legal form were central to its development, so too were its geographic dimensions. Yet the relationship between the corporate legal form and space or territory was far less consistent than the relationship between corporations and time. Advocates of the natural entity theory of corporate personhood argued that the corporate nature of associations preceded the establishment of either the state or the law. In this sense, associations might have organic relationships to land that were only retroactively recognized by states through legal designations of property and territory. For instance, Otto von Gierke, the nineteenth-century political theorist and major proponent of the natural entity of corporate personhood, focused on the long history of German fellowships, arguing that they emerged out of the collective life of German communities during the medieval period and therefore antedated both the emergence of modern nation-states and the incorporation of Roman law into German legal codes. Many of these fellowships, such as corporate towns and “city commonwealths,” were geographically circumscribed and connected with specific places. In this sense, land could be an organic component of the corporate entity, but the jurisdictional dimensions of these corporations were secondary to the broader social relations that had produced the associations (see Gierke 1990).

In contrast, artificial entity theories, including the grant theory that shaped the initial development of Anglo-American corporate law, argued that corporations were creations of sovereign authority. Sovereigns issued grants to groups of individuals to encourage them to undertake actions that were deemed beneficial for the state and beyond the scope or capacity that any individual could accomplish. Corporations were created for a wide variety of purposes, such as philanthropic organizations, educational and religious institutions, and towns, as well as institutions for banking, transportation, and trade. Because corporations were viewed as creations of law, their legal standing – including their abilities to hold property, to sue and be sued, and to enter into contracts – was confined to locations in which the authorizing sovereign’s legal system was recognized. Sir Edward Coke famously codified this definition as part of the “essence” of incorporation in the 1612 Case of Sutton’s Hospital, arguing that corporations were always “of a place, for without a place no incorporation can be made.” Geographic elements were designated in the charters and “letters patent” that created corporate entities, and one of the primary purposes of incorporation was that it allowed for members to hold property collectively, including but not limited to land, under the common name and seal of the corporation.

For corporate institutions such as towns or hospitals, which were located in specific places, such geographic designations were fairly straightforward grants of property to corporate entities. Corporations whose enterprises spanned multiple distinct jurisdictions, however, could not rely on their chartered grants alone. This was especially the case with the European imperial companies of the seventeenth, eighteenth, and nineteenth centuries, whose participation in the project of European colonialism brought them into legal relationships of conflict and compromise both with representatives from multiple European states – including their home and foreign governments, along with other competing corporate entities – and with local rulers in the areas that European powers were seeking to colonize. For this reason, the development of international law, which coincided with
the growth of the imperial corporations, gave special consideration to the legal regulation of chartered corporations operating outside of the states in which they were chartered and, often, in international spaces such as the seas that were not under the *dominium* of any other recognized European state. For instance, Hugo Grotius, generally considered as the founder of modern international law, articulated the concept of the freedom of the seas as part of his defense of the Dutch East India Company’s appropriation of trade routes in Asia. In terms of the legal rights of corporations on land, Grotius argued that corporations had rights to appropriate territory that was not being used by its owners. Although foreign sovereigns had a right to police persons, including corporate persons, that entered their jurisdiction, they had no right to prevent them from occupying lands that were considered unused or “waste” (see Tuck 1999). In this manner Grotius’s legal arguments helped facilitate the colonization of indigenous lands by Dutch chartered corporations.

The treatment of foreign corporations in international law also shaped the development of corporate law and the corporate economy in the United States. Much of the initial British colonization of North America was carried out by chartered corporations and many of the colonies were governed under corporate charters. After the US War of Independence, the power to create corporations was reserved for state governments as corporations shifted from quasi-public entities of government to private entities created by contracts between states and shareholders. States chartered corporations to carry out much of the work of local development, including the construction of roads, bridges, and railways, the administration of poor relief, the organization of intellectual, educational, and religious institution (such as learned societies, universities, and churches), the government of cities and corporate towns, and the organization of business and manufacturing. Because individual states were considered independent sovereigns in their capacity of chartering corporations, US federal jurists turned to principles developed in international law to help regulate corporations that engaged in business outside the jurisdiction in which they were chartered. As with international law more generally, the regulation of foreign corporations in US courts through the nineteenth century facilitated the expansion of the corporate economy, with states generally recognizing foreign corporations outside of their home jurisdictions except in those rare cases where local state legislatures explicitly prohibited them (see Henderson 1918).

By the turn of the twentieth century, however, local forms of corporate regulation grounded in the issuing of charters were wholly insufficient for regulating a dramatically expanding economy. Liberal general incorporation laws by states such as New Jersey and Delaware provided easy access to the privileges of the corporate legal form, complete with limiting the liability of investors and authorizing holding companies. Moreover, political concerns over the growth of cartels, monopolies, and trusts in key industries such as sugar, oil, meatpacking, tobacco, and, most importantly, the railroads gave rise to new forms of corporate regulation. Beginning in the 1880s, the US federal government undertook a variety of regulatory steps to control corporate consolidation, including the regulation of railroad rates through the Interstate Commerce Commission in 1887; the creation of administrative agencies to gather economic information and make policy, such as the US Industrial Commission in 1898, the US Bureau of Corporations, and the Department of Commerce and Labor in 1903, and the Federal Trade Commission in 1914; the institution of antitrust laws via the Sherman Anti-Trust Act of 1890; and, finally, a series of
regulations on banking and finance as part of the New Deal, including the Glass–Steagall Act of 1933 and the creation of the Securities and Exchange Commission in 1934. Unlike earlier forms of corporate regulation that controlled corporations directly, through the mechanism of the state-issued charter, federal regulation focused on controlling the effects of corporate consolidation in the market and, primarily, in terms of prices. Forms of consolidation that hurt competition and artificially raised prices for consumers were targeted for regulation, although some limited monopolies were permitted in areas like public utilities and railways, where unrestrained competition might jeopardize the industry in its entirety. Key components of this regulatory compromise, which Martin Sklar (1988) has called the “corporate reconstruction of American capitalism” or more simply “corporate liberalism” and that geographers most commonly refer to, with a slightly different emphasis and historical timeline, as “Fordism,” held sway in US economic regulation up to the deregulatory movement associated with neoliberalism beginning in the 1970s.

What is critical about this story from a geographic perspective is that, in the context of the US economy, the federal government supplanted the localized form of early corporate regulation with a national project designed to promote the growth of corporate capitalism while limiting its anticompetitive tendencies. In other words, corporate liberalism depended on the national jurisdiction of federal law and administration to match the increasing geographic scope of the corporate economy. Yet, in terms of the international or global economy, there was never a globally coordinated system of law or regulation that could even provisionally stem the crisis tendencies of global capitalism. Instead, the last 100 years have witnessed the rollout of a variety of provisional measures and regional attempts to address specific aspects of corporate regulation. From the grand projects of economic regulation associated with international institutions – such as the League of Nations’ attempts to unify private international law, including the legal definition of corporations, beginning in the 1920s, or, later, the New International Economic Order’s struggles at the United Nations during the 1970s and 1980s to produce legally binding corporate codes of conduct that would protect the economic sovereignty of “third world” states—to the long history of bilateral and multilateral treaties, agreements, and compacts designed to regulate trade and industry, the modern international order has never had a standardized approach to corporate regulation. Like all international economic activity, corporate-led globalization transpires in an economic space governed by the uneven relations between international actors and institutions, including states, corporations, and global, transnational, and regional agencies. With respect to the legal definitions of corporations, however, fundamental issues such as their nationality or domicile and their responsibilities to conform to regulations issued by sending and receiving states, as well as international law more generally, continue to be debated by scholars and litigated in courts. Likewise, the investment decisions of multinational corporations are shaped in important ways by the uneven nature of the global regulatory landscape. Corporations regularly engage in regulatory arbitrage to find more profitable environments for their activities. Finally, in light of the challenges of global legal regulation, the international community has come to focus on “soft law” approaches to regulation that emphasize voluntary norms over legally binding codes, such as the guidelines concerning corporate social responsibility articulated in the United Nations Global Compact.
Given the wide range of contemporary issues involving the intersections of corporations, law, and geography, a comprehensive account of current themes, much less their different treatment in disciplines such as law, geography, and political science, is beyond the scope of this entry. Moreover, because the regulatory landscape associated with economic globalization is dynamic, detailed empirical accounts only capture snapshots of a constantly shifting terrain. Nevertheless, there are some persistent topics involving corporations and the law that are particularly central to the concerns of human geography. In terms of global issues, these include the transnational regulation of corporations and corporate activities in domains such as taxation, finance, labor, and the environment, broader questions about corporate social responsibility, and the responsibilities of corporations for the protection of human rights (see, generally, Crane et al. 2008; on taxation see Avi-Yonah 2000). These issues intersect with more geographically localizable concerns, such as the relations between corporations and indigenous communities (a particularly important topic in discussions about the human rights obligations of corporations) or the role of corporations in structuring and governing various places of production, from the increasingly ubiquitous export processing zones to the enclaves associated with diverse types of resource extraction (see Sawyer 2004). In addition, there are a host of topics that concern political questions about corporations, ranging from the legal basis of various forms of privatized government under both imperialism and neoliberalism to the extensive use of private corporations in military and police actions (Freeman and Minnow 2009; Mbembe 2001; Singer 2003).

Noticeably, these topics are either treated by legal scholars, on the one hand, or by geographers and other critical social scientists, on the other, but rarely in consort or conversation with one another. Thus much of the ethnographic study of the politics of capitalist production happens with little consideration of the legal dynamics that are essential to understanding issues like the mobility of capital or the structure of industrial relations. Likewise, legal discussion about attempts to harmonize corporate regulations of labor, the environment, or taxation often transpire without careful consideration of the geographic dimensions of these problems, operating not just between distinct national jurisdictions but with scalar dynamics as well. Future research on corporations and the law from an explicitly geographic perspective could address both of these gaps, while also giving much greater attention to the domain of legal politics and the politics of law in structuring social and spatial relations of capitalist economies. Research is also needed on the treatment of corporations, law, and economy in diverse legal traditions. Although there is a great deal of scholarship on the development of the corporation in European systems of common law and civil law, there is little comparative or integrative scholarship that examines other corporate entities in non-Western legal systems or the ways that non-Western corporate entities developed through colonial encounters of European and non-European legal cultures.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Corporate environmental responsibility; Corporate identity; Corporate world and work: Indian women in the IT sector; Corporations and local states; Corporations and the Anthropocene; Corporations and the nation–state; Export processing zones; Firm migration; Human rights; Labor geographies and the corporation; Legal geography; Sovereignty; States and development
CORPORATIONS AND THE LAW

References


Kantorowicz, Ernst. 1957. The King’s Two Bodies: A Study in Medieval Political Theology. Princeton: Princeton University Press.


Corporations and the Anthropocene

Martin Perry
Independent research contractor, New Zealand

The Anthropocene is a controversial term used to distinguish the period of geologic time over which human populations are having the biggest influence over the planet’s environment in terms of atmospheric, geologic, hydrologic, biospheric, and other Earth system processes. The term has been proposed for formal adoption by the International Union of Geological Sciences (IUGS), the professional organization in charge of defining the Earth’s time scale. The IUGS has convened a group to advise it whether to officially declare that the world is currently living in the Anthropocene. Meanwhile, the Earth remains officially in the Holocene, the geological epoch starting from the end of the last glacial period around 11 700 years ago.

A potential start date of the Anthropocene is when humans had their first irreversible environmental impact affecting a natural ecosystem. This would date the beginning of the Anthropocene to about 50 000 years ago in the Pleistocene epoch and prior to the commencement of the Holocene. This was when humans contributed to the mass extinctions on most continents of megafauna such as mammoths and giant kangaroos. While significant in changing ecosystem dynamics, this impact may be considered modest in terms of its capacity to exert a geological impact. The next potential start date brings the Anthropocene forward by about 42 000 years to the time when humans began to practice various forms of agriculture. With agriculture came extensive land clearing achieved mainly through fire, which had atmospheric impacts as well as causing soil erosion of some land cleared of vegetation. Such change is significant but is also a product of natural processes as lightning produces fires too, so making the early Anthropocene hard to separate from the post ice age epoch.

The industrial revolution and its evolution into the atomic age are the prime candidates for identifying the onset of the Anthropocene. This is the start date favored by Paul Crutzen, the Nobel Prize-winning scientist who first caused substantial interest in the onset of the Anthropocene. The term itself is traced to Andrew Revkin and his 1992 book, Global Warming: Understanding the Forecast, in which he discussed how human contributions to climate change had created the “Anthropocene.” Crutzen suggests the term Anthropocene originated when he gave a conference presentation in 2000: “Where someone said something about the Holocene … I suddenly thought that this was wrong. The world has changed too much. So I said: ‘No, we are in the Anthropocene.’ I just made up the word [constructed by combining ‘anthro’ (human) and ‘-cene’ (epoch)] on the spur of the moment. Everyone was shocked. But it seems to have stuck” (Pearce 2007, 21). Further justifying the use of the term, Crutzen (2002) went on to say that: “The Anthropocene could be said to have started in the late eighteenth century, when analyses of air trapped in polar ice showed the beginning of growing global concentrations of carbon dioxide and methane.” Beyond the contribution to changing the mix of atmospheric gases, the first nuclear weapon test on July 16, 1945, is viewed as a further definitive
CORPORATIONS AND THE ANTHROPOCENE

step into the Anthropocene with the release of unique isotopes across the globe.

Critics of the concept argue the distinguishing feature of the present century may not be the widespread existence of global environmental pressures – 2 billion years ago cyanobacteria oxygenated the atmosphere and powerfully disrupted life on Earth – but the greater consciousness of the impacts of human activity on the environment. Epochs revealed by the Earth’s geologic record encompass millions of years whereas the existence of the Anthropocene is influenced by the perception of what people’s impact on the environment might be as well as what has resulted to date. The environmental journalist Fred Pearce (2007), for example, argues for the Anthropocene by referring to the possibility of critical thresholds being crossed that will result in abrupt global environmental change. Human-induced global warming has potential to create lasting environmental change in this way but it remains possible that variations in solar eradiation or the onset of massive volcanic and tectonic activity might ultimately define the post-Pleistocene period.

Regardless of whether the Anthropocene is accepted officially, the term is being used to reflect the widespread agreement that human populations are having impacts on the environment at the scale of the planet as a whole. Considering this and the evidence of a significant acceleration in human impacts since the industrial revolution, there are two main reasons to single out corporations as disproportionately responsible for changing the planet’s environment. First, economic power is concentrated in large corporations who are responsible, directly and indirectly through the businesses that serve corporations as suppliers and subcontractors, for the largest share of environmental impacts. Second, corporations exert influence across the political and economic environment to perpetuate a system that enables their continued dominance and precludes the emergence of more sustainable forms of economic activity that might lessen further human impacts on the environment.

The economic and technological power of corporations

Saleem Ali, in his commentary on the history of human relationships with the Earth’s resources, records how a sign by the entrance to the National Mining Museum, Leadville, USA reads: “More minerals have been consumed since the start of World War II than in all of the rest of human history” (Ali 2009, 8). He cites evidence from the USA Geological Survey that from 1975 to 1995 total consumption of materials increased by 67% in the United States. Arguably of more significance for the possible onset of the Anthropocene, the amounts of material used by weight had moved from 50% renewable in 1900 to only 8% renewable in 1995. The Bingham Canyon mine in Utah, USA, is one visible symbol of this consumption boom: the world’s single largest human excavation. Opened in 1906, the mine now measures around a mile deep and 2.5 miles across. The Grand Canyon is of similar depth and up to 18 miles wide but, as Ali observes, it took 6 million years for the Colorado River to carve the Grand Canyon; it took 100 years for humans to excavate the Bingham Canyon mine.

As well as the growing scale of material consumption, the last century has seen a growing diversity in the range of materials consumed. At the beginning of the twentieth century, around 20 elements of the periodic table were exploited for commercial product manufacturing. At the end of the century, manufacturing drew from all 92 of the naturally occurring elements and incorporated some of them in entirely novel
CORPORATIONS AND THE ANTHROPOCENE

materials such as various types of plastic and carbon composites created through intricate production processes involving hundreds of material inputs. Such materials have proved to be particularly persistent in the environment as they do not decay as readily as natural materials that can convert back to their elemental origins. A consequence is seen in the way that oceans have become repositories of discarded plastics that accumulate in concentrations created by the oceans currents, such as the north Pacific subtropical gyre between Hawaii and the California coast. In this part of the Pacific it is estimated that the volume of human-made materials (principally plastic fragments) accounts for six times that of naturally occurring oceanic zooplankton.

With the demand for energy to fuel economic growth has come an even more rapid growth in carbon emissions, increasing from 1500 million tons annually in the 1950s to around 7000 million in the 2010s (Marland, Boden, and Andres 2008). This is reflected in the increased concentration of carbon dioxide in the atmosphere: almost 400 ppm from approximately 300 ppm prior to the onset of industrialization, with this increased concentration generally accepted to be the cause of an increase in global temperatures of about 0.74°C over the past century (NASA 2011). As well as changing the Earth’s biosphere, rising sea levels, and shrinking ice coverage, global warming is linked to changes in the alkalinity of the ocean as a result of acidification by anthropogenic carbon emissions (Steffen 2011).

There is no precise accounting of the sources of these environmental changes. The responsibility of corporations is inferred from the coincidence of the increasing concentration of economic power in a comparatively small number of big businesses that operate with global reach alongside the deteriorating state of the Earth’s environment. This global reach is exercised through both the ownership of business assets around the world and their control of global supply chains by means of which notionally independent businesses are configured to serve a corporate customer. Moreover, the concentration of economic power is particularly high in sectors that have the greatest environmental impact, including mining, energy, chemicals, transport, and food.

The political power of corporations

More than their direct impact on the environment through resource extraction and pollution, corporations are accused of having so much power that they are able to “capture the state” (Korten 1995) and “rule the world” (Monbiot 2000). This accusation alludes to the way that corporations have more influence over patterns of development than it appears many national governments do. This influence has been linked to what has been called the “transnational power of withdrawal” (Beck 1997). By encouraging a perception that their operations are highly mobile, corporations are said to gain negotiating power over nation-states. The outcome is not necessarily reflected in governments simply consenting to the demands of corporations, who may decide that a better strategy is to allow governments to regulate their activity and then find ways of mitigating the impact of the regulation by identifying loopholes and means of compliance that do damage to their core strategies. An information asymmetry between the regulator and corporations, created by the corporation’s ability to control operational and financial data measuring various dimensions of its performance, its ability to draw on its worldwide resources, and its global mobility, can make it hard for governments to achieve their intentions through laws and other regulations. Moreover, the extent to which people and organizations have come to depend upon the
benefits created by corporations makes it hard to envisage alternative futures without them.

Geographers such as Peter Dicken (e.g., see Dicken 1997, 2011) argue that corporations are extremely diverse in their size and characteristics and that this should caution against treating them as a monolithic group yielding enormous power. Most large transnational corporations operate in only a small number of countries outside their region of origin and even those with extensive international operations tend to retain a home base. They tend to be national firms with overseas operations rather than truly placeless institutions. Having a home base means that corporations do not evade all forms of national regulation simply by locating activities overseas: the US Foreign Corrupt Practices Act 1977 and Alien Tort Claims Act 1789 are examples of national law extending to the overseas operations of transnational companies. Similarly the power of corporations needs to be judged against their need to adapt to shifts in technology, organizational opportunities, and markets. To survive, even big corporations must be learning institutions, and economic geographers see this as a key reason why most transnational corporations retain a national identity. Learning is something that both embeds corporations within their home country and drives them to seek integration within the social, economic, and political fabric of that place. Corporations may actively seek to shape the institutional and regulatory environment of their home country and may use their global reach to evade national regulation, as where they relocate their tax domicile away from the places where activities are actually conducted, but they are also to some degree carriers of the changing social and political environment of their home country. This is reflected, for example, in the tendency for transnational corporations to vary in their commitment to environmental and social responsibility according to the emphasis placed on these issues in their home country.

Geographers also point out that the power and influence of corporations is not simply a product of corporations extending their global reach: intensifying competition between nation-states also is part of the context. States compete with each other to enhance their international trading positions and to capture the economic benefits of international trade and investment from competitor nations. This is reflected, Peter Dicken argues, in governments deregulating aspects of their economies partly targeting sectors such as financial services and telecommunications where prospects for enhancing their share of global industries is thought to exist. Cooperation between states in the formation of regional trading blocs is another manifestation of inter-state competition. Trading blocs vary in their administrative form and objectives but providing an attractive environment for transnational corporations is typically part of the agenda. Where individual countries come together in a trade bloc, the greater ease of cross border transactions enables some corporations to grow and use their scale to expand globally.

The weakness of inter-governmental institutions for managing the global environment has been exposed as corporations have grown in importance. As discussed by Michael Blowfield and Alan Murray (2008), a governance vacuum at the global level is evident in the absence of entities with clear responsibility or accountability for issues such as human rights, poverty, or the management of the global commons. The entities that do exist, such as the World Trade Organization, are viewed as overly influenced by the interests of the wealthiest nations, or as limited to promulgating principles of good practice and promoting dialog but with limited means of enforcing change. In 1999, for example, the United Nations announced its Global Compact
to bring companies together with UN agencies, labor organizations, and civil society to promote responsible corporate practices and encourage compliance to UN development and environment goals. The Compact has succeeded in gaining affiliation from over 7000 businesses including most of the world’s largest corporations but questions remain over whether it is succeeding in changing business behavior (Rasche 2013).

Corporations and green growth

Partly encouraged by the UN Global Compact, statements of commitment to corporate social responsibility and sustainability are now commonplace among big businesses. Claims of a commitment to “zero waste,” “100% renewable energy,” “zero toxics,” “zero deforestation,” and “100% sustainable sourcing” are among the bold promises being made by major multinational companies such as Walmart, Nestlé, Nike, McDonald’s, and Coca-Cola. Yet it is not easy to disentangle the significance of the steps being taken from the rhetoric of corporate responsibility. A general judgment is that the steps being taken are more about helping big businesses to continue to grow than about promoting environmental protection (Dauvergne and Lister 2013).

Beyond the individual initiatives of corporations, there is a possibility that future economic growth can be “green.” The Organisation for Economic Cooperation and Development (OECD 2011) explains the concept of green growth as fostering economic growth and development, while ensuring that natural assets continue to provide the environmental services on which life depends. Exemplary activities seen as contributing to green growth include renewable energy, smart electricity grid technologies, biofuels, and improved recycling processes. As well as such technological innovation, the OECD recognizes that there is a requirement for environmental externalities and market failures to be more fully incorporated within economic decision making, to incentivize companies to adopt green technologies. The green growth agenda relies on the assumption that business will substantially reduce its impact on the environment once an appropriate price has to be paid for the use of the environment. Experience with carbon credits and emissions trading indicates that corporations accept this approach, but that such market-based approaches to controlling environmental challenges are hard to design in ways that remain effective (De Freitas and Perry 2012; Perry 2015).

Transnational civil society, in the form of transnational social movements, global mass media, and local grass roots groups is seen by some to be a potential counterweight to the power of corporate globalization. In this context, civil society refers to forms of social organization leading to collective action that exists above the scale of the individual but below that of the state. One attempt to quantify the scale of transnational civil society identified around 20,000 transnational nongovernmental organizations of which 27% were concerned with human rights issues and 14% with the environment, with women’s issues accounting for the next largest grouping (Willetts 1998). Although lacking the financial resources of the corporations that are the main target of much transnational activism they have proved effective where their campaigning has threatened the brand value and reputation of corporations. This is possible because much of the dominance of corporations relies on carefully constructed marketing and image generation, particularly in the case of corporations that directly serve end-consumers as compared with corporations that have a largely business clientele. Especially in areas bearing a potential environmental or
public health impact such as the chemical and food industry, the public have become more open to claims of corporate wrongdoing.

Campaign groups can succeed where environmental issues are turned into a competition for public attention rather than a reasoned discussion of available options. A paradigmatic example of the challenges that a publicly exposed corporation may face from transnational campaign groups is the 1995 controversy between Royal Dutch Shell and Greenpeace over the disposal of the Brent Spar oil-storage tank. Greenpeace succeeded in painting Shell as an environmental villain in the minds of consumers and many elected government officials, even though Shell had selected the best environmental disposal option and Greenpeace’s estimates of the volume of toxic waste involved were wrong. This case illustrates how even operating within the law may not provide corporations with necessary legitimacy to maintain public support. Yet it also illustrates the weakness of transnational campaign groups that depend on mobilizing public opinion against corporations. When public opinion is influenced more by perception than a rational assessment of available evidence it becomes challenging to effect concerted change based on systematic evidence of the relative importance of different environmental challenges to humankind’s future. Anti-corporate campaigning can be successful in individual cases, but has yet to demonstrate capacity to bring systematic change on a scale matching the concerns informing the Anthropocene.

**SEE ALSO:** Biopolitics; Climate and societal impacts; Corporate environmental responsibility; Corporations and global trade; Earth system science; Ecological footprint; Environment and waste; Environmental futures; Environmental history; Environmental movements and protest; Environmental science and society; Holocene; Natural resources; Natural resources and human conflict; Nature; Resource extraction; Sustainable development

**References**


Further reading


Corporations and the nation-state

Henry Wai-chung Yeung
National University of Singapore

As two discrete concepts, corporations and the nation-state have been well studied in economic and political geography. Putting them together necessitates a significant reorientation in both subfields, however, toward a new geographical understanding of uneven capitalist global economy. According to Sheppard (2011), this new understanding can be described as geographical political economy. Grounded in the subfield of economic geography, this entry attempts to trace the growing importance of corporations or firms, the nation-state, and their relationships (firm-state) in contemporary geographical research since the late 1960s.

In retrospect, corporations did not receive substantive analytical attention in economic geography until the emergence of the geography of enterprise approach in the late 1960s and the early 1970s. The nation-state and its manifold relationships with corporations came to the forefront of geographical interests only in the late 1980s. In his Roepke Lecture at the 1989 annual meeting of the Association of American Geographers, Brian Berry (1989, 2) declared that “I believe that nation-states are primary agents shaping the economic geography, drawn together by other global actors – the multinationals that link the global economy.” This was his “realist” moment, reflecting then, and still now, a dominant intellectual paradigm in international relations and international political economy that places an overriding emphasis on the competitive and rational power politics of the nation-state in shaping the global economy. To him, corporations, including multinationals, are economic actors embedded in territorially bounded cultures or worldviews predetermined by the nation-state. These embedded economic actors then interact with nation-states to produce the comparative geography of the global economy.

By the mid-1990s, this firm–state interactive nexus had been firmly established as one of the central problematics of economic geography (the others including “geographical industrialization” and “industrial districts and clusters”). Described by Peter Dicken (1994, 101) in his 1993 Roepke Lecture as “exciting and challenging times in economic geography,” geographical analyses of these firm–state interactions along the global–local power nexus provided a very fruitful way forward for understanding economic globalization and its highly uneven consequences. These research agendas considerably influenced the dynamic development of the subfield in the subsequent two decades.

This entry has three principal sections. It first revisits the role of corporations in the emerging geographical political economy. It then examines the role of the nation-state as a regulatory and promoter of corporations in distinctive capitalist political economies. The final section focuses on the territorial embeddedness of foreign firms and a relatively new phenomenon in the complex firm-state nexus – the globalization of national capital in the form of sovereign wealth funds. The entry ends with a brief note on future research directions for studying the geographical political economy of these evolving firm–state relations.
Corporations in geographical political economy

While Weberian location theory dominated industrial and economic geography in the 1950s and the 1960s, little analytical attention was given to corporations because the primary objective was to identify the least-cost location of particular industrial activity. Having inherited many assumptions from neoclassical economics, such as perfect competition and information, profit maximization, and economic rationality, this Weberian geography approached corporations as a locational “black box” responding to competitive initiatives in accordance with the law of diminishing returns (i.e., production costs and market incentives) and producing economic outcomes in space.

By the late 1960s, there was a growing discontent with this Weberian theory on industrial and economic geography. The neglect of the organization of corporations was particularly frustrating. As noted in Storper and Walker’s (1989, 125) *The Capitalist Imperative*, “Neoclassical economics and Weberian geography beg the question of organization by treating the market as the mediator of all economic transactions, the plant as a production function, the firm as single plant, the industry as made up of representative firms producing a single product, and the region as a blank slate on which firms individually pick out the best spots to locate.”

This major dissatisfaction with location theory prompted a new behavioral approach after the late 1960s in the form of the geography of enterprise or “corporate geography.” The object of investigation in the geography of enterprise was neither an industry nor a region consisting of a large number of anonymous corporations. It was rather the single firm or a small number of representative firms or plants. During the 1970s, research in the geography of enterprise was primarily preoccupied with the motivations, locational decisions, and adjustments of individual corporations. Five interrelated fields of geographical inquiry characterize this behavioral approach (Hamilton 1974, 6).

1. The traditional problem of selecting the initial location of the industrial firm and its manufacturing plant.
2. The spatial forms and implications of the growth and changing organization of the corporation.
3. The corporation’s decision-making procedures employed in, and motivations for, choosing any kind of locational behavior.
4. The spatial adjustment by corporations to their existing or chosen environments or in changing the size or use or relative importance of existing plants or facilities within their control.
5. The comparative spatial behavior of corporations of different organization and scale, of industrial organizations in contrasting economic and social systems, and of corporations in diverse cultural environments.

This behavioral approach did not go far enough, however. Much of the research remained dominated by neoclassical approaches. While continuous transformation was occurring within the geography of enterprise approach, the late 1970s and the 1980s also witnessed an upsurge of a radical Marxist interpretation of geographical industrialization eloquently summarized in Storper and Walker (1989). But corporations and their relations with the nation-state continued to receive insufficient analytical attention.

During the 1980s, corporate geographers became much more aware of the weaknesses of the earlier corporate growth models and branch plant literature, in particular the separation of the
corporations from its structural and institutional environment. Growing analytical emphasis was paid to the spatial organization of corporations, as shaped by the interaction between the structural and institutional context of the corporation and its strategic predisposition.

The corporate growth approach in the geography of enterprise also was inadequate in its conceptual accounts of the structural context in which corporations were (re)produced. Although environmental factors were explicitly analyzed, these were conceptualized mainly in terms of interfirm competition and consumer markets. There was a serious lack of attention to inherent and asymmetrical power relations in intra- and interfirm corporate networks. Given any specific environmental context, theory gave far too much autonomous power to decision-makers in these corporations, an unintended consequence of its conceptual foundation in the behavioral approach. Spatial outcomes were merely expressed as a result of corporate strategies. Corporate actors were conceptualized as making decisions as if there were no institutional impediments and power struggles with other actors, such as the nation-state and nonstate organizations.

The role of the nation-state in shaping the territorial outcomes of corporate growth was mostly neglected. In his reappraisal of corporate location theory, Berry (1989, 3) argued that “[w]hile much is being written about multinational decision making, much less has been codified about the role of the nation-state in the new locational calculus.” Even when nation-states were mentioned in the branch plant literature, the intimate relationships and mutual coexistence between corporations and nation-states were only mentioned in passing, with little in-depth theorization of these complex firm–state relationships.

By the turn of the 1990s, leading scholars in the geography of enterprise were making a much stronger plea that “[w]e are perhaps approaching the end of the beginning; we are certainly not at the beginning of the end” (Dicken 1990, 243) and that “there is a real and continuing need for research into the ‘geography of enterprise’ and not just at the margins of economic geography” (Dicken and Thrift 1992, 288). Most importantly, the theoretical and empirical frontiers of research in this branch of corporate geography were redirected toward the firm–state nexus and the territorial embeddedness of business organizations. Both of these research foci were explicitly structural and institutional, approaching the sort of geographical political economy characterized in Sheppard (2011). The aim was to elucidate the complex ways and processes in which corporations interact with nation-states and become (re)produced in localized territories. As detailed in the next two sections, this shifting nature of conceptualization and empirical investigation presented an exciting challenge to corporate geographers who were increasingly interested in an emerging form of geographical political economy.

The nation–state as a champion of “national champions”

Since the early 1990s, the firm–state nexus has emerged as a leading frontier in geographical political economy. The nation–state is no longer an “environment” or “context” in which corporations thrive/wither and the global economy prospers/declines; it is rather the key architect of economic globalization and changing global economic governance. Despite the dire predictions popularized by such “ultra-globalists” as Kenichi Ohmae and Thomas Friedman, the globalization of economic activity by (trans)national
CORPORATIONS AND THE NATION-STATE

corporations has not inevitably led to the end of the nation-state and geography and the rise of a flat world. Much to the contrary, economic globalization has accentuated the dialectical relationships between nation-states and (trans)national corporations (Yeung 1998, 2016). In a recent Roepke Lecture, Harvard political economist Dani Rodrik (2013) argued strongly for the continued vitality of the nation-state in global economic governance (for a critique, see Agnew 2013). To him, geographical heterogeneity and institutional diversity among nation-states are the keys to a “sane globalization” that can “accept the right of individual countries to safeguard their domestic institutional choices” (p. 17).

In short, institutional differences among nation-states can make a huge difference to the successful formation and viability of domestic economic actors such as corporations and industries. But what exactly are these institutional differences? It is useful to introduce a simple typology of nation-states in today’s global economy. In Coe, Kelly and Yeung’s (2013, 100–108) Economic Geography: A Contemporary Introduction, the nation-state has been broadly categorized into four types with varying degrees of institutional capacity, political control, and policy orientations.

1 Neoliberal states: state institutions seek to keep a distance from corporations and industries. The main role of the state is to establish market rules via legislation and to enforce these rules through their regulatory mechanisms. Some contemporary examples are the United States, the United Kingdom, and New Zealand.

2 Welfare states: in these states, labor unions and state institutions play a more direct role in corporate governance and firm behavior. Corporations do not necessarily have a free rein in labor management. Instead, there are stringent labor laws and other state-sanctioned welfare provisions that shape the business and investment behavior of corporations. Well-known examples are France, Germany, and Scandinavian countries.

3 Developmental states: the state is relatively autonomous from the influence of interest groups, powerful businesses, and the population at large. This institutional autonomy – often achieved through authoritarian political control – enables the state to develop and pursue interventionist policies favoring economic development. To achieve their economic development objectives, these states exercise economic control through developing elite, state-sponsored economic agencies and strategic industrial policies. Many East Asian newly industrialized economies tend to experience such a state form (e.g., Japan, South Korea, Taiwan, and Singapore; see Yeung 2014, 2016).

4 Authoritarian states: these states combine a highly centralized political system with an increasingly open economic system. Many are former socialist states that have sought to liberalize their economies while maintaining strict political control. Their economies tend to be dominated by a mixture of state-owned enterprises (SOEs) and corporations (domestic and foreign). The state exercises its control over domestic economies through owning stakes in SOEs and strictly regulating corporations and industries. Authoritarian states are mostly found in central and eastern Europe (e.g., former Soviet republics), Russia, and Asia (e.g., China and Vietnam).

This great diversity of nation-states points to their highly variegated institutional relationships with corporations – domestic and foreign. In
CORPORATIONS AND THE NATION-STATE

most of these state types, the state performs a *regulatory* role by defining and enforcing the prevailing “rules of the game,” such as property rights, market regulation, and governance laws. Generally, corporations have a significant degree of freedom to make investment, production, and distributive decisions; their economic actions and fortunes are subject to market, not state, discipline. Depending on the specific nature of their economic strategies and policies, however, some of these nation-states also serve as a *promoter* of corporations. Directly involving themselves in steering the growth and development of domestic corporations, these states have become “developmental” in the sense that national economic development serves almost as their *raison d’être*. In this form, firm–state relations are particularly intertwined and mutually constitutive. In the name of certain national development objectives, the state favors a highly selective group of national firms commonly known as “national champions.”

A developmental state can deploy a wide range of policy instruments and/or even political coercion to favor particular corporations and/or industrial sectors (details in Yeung 2016). In the East Asian examples of Japan, South Korea, and Taiwan between the 1960s and the 1990s, the developmental states explicitly engaged in certain protective trade policies to nurture their chosen “national champions” and “infant” industries (e.g., Nippon Steel and POSCO in steel making, Toyota and Hyundai in automobiles, and Samsung and TSMC in electronics). These industry-specific trade policies were combined with strategic industrial policies and preferential financial tools to incentivize these corporations to enter new industries prioritized by the state for national development purposes. Relatively autonomous from pre-existing business and industry interests, these states’ ministries of industry developed comprehensive five-year development plans for nurturing new industries, initially through import-substitution programs, later through export-oriented industrialization.

In Japan and South Korea, the respective finance ministries or central banks also offered “policy loans” to distort deliberately the financial market system in credit allocation so that domestic corporations could “afford” to invest in these highly capital-intensive industries. Following state-led instrumental policies, these loans were priced so cheaply or even offered as grants to chosen corporations to help them overcome the initial, often huge, barriers to entry in new industries. In some extreme cases (e.g., South Korea’s President Park Chung Hee during the late 1960s and the 1970s) the state leader personally supervised the progress of these “national champions,” embodying and implementing national industrial policy. In lieu of a market-based credit allocation system, top state officials made allocation decisions based on a complex web of personal favors and crony interests.

While much of the enormous literature on these East Asian developmental states has been produced by political scientists, economic sociologists, development scholars, and East Asian area specialists since the late 1980s and throughout the 1990s (reviewed in Yeung 2014, 75–79, 2016), it is important to note that leading economic geographers, including Brian Berry (1989) and Peter Dicken (1994), also paid serious attention to these developmental state–firm relations. According to Berry (1989, 6), success of East Asian “development states” “centered on the ability of the central government to forge bonds of trust with the business community, the intelligentsia, and the working class. Linked to this is the idea that politics is the arena in which the vital decisions about the long-term economic health of the nation are to be made and the strong belief in meritocracy through competitive
examination and in the improbability of the human condition through communal self-help.”

Dicken (1994, 112) couched these developmental states in the context of the emergence of the “competition state,” which takes on some of the characteristics of corporations as they strive to develop strategies to create national competitive advantage. As such, the nation-state is much more than a contributor to the “environment” of corporations. Indeed, he argued that “it is not unreasonable to regard many of the current politico-economic tensions in the global economy as being a reflection of a clash between competition states occupying different positions within the market-rational/plan-rational space” (p. 113). While the developmental state belongs to the plan-rational political-economic system, the regulatory state epitomizes the market-rational variant. Geographical political economy can therefore be much better understood if the unequal power relationships and complex interactions between these state types and their domestic corporations are duly taken into account. This geographical political economy can be extended to an analysis of the state-led globalization of domestic capital.

Territorial embeddedness and the globalization of national capital

Insofar as the developmental state can serve as a champion of “national champions” (i.e., domestic corporations), the nation-state can also become a highly involved actor in shaping economic globalization in two ways. First, the nation-state can influence or shape the business and investment behavior of foreign transnational corporations (TNCs) coming into their national economic spaces. Generally speaking, the nation-state prefers to embed these TNCs territorially in order to maximize the positive spillover of their investment for local and regional economies. Second, and in somewhat more limited cases, the nation-state can become a global investor in its own right, actively steering the globalization of its own national capital. This two-way involvement of the nation-state in globalizing corporations has attracted very significant research in economic geography since the mid-1990s.

In particular, Dicken (1994, 117–121) has rigorously argued that the dynamics of firm–state bargaining relationships are the critical key to understanding and mitigating intense global–local tensions in today’s global space-economy. These bargaining relationships need careful geographical analyses because of their inherent complexity and evolutionary tendencies. As corporations from different countries of origin governed by variegated nation-states, TNCs embody different home state imperatives and are bearers of different business “recipes” or cultures. Even in the same broad industrial sector, a South Korean chaebol or conglomerate such as Samsung Electronics clearly operates rather differently from its key competitors – American high-tech electronics giants such as Apple (smartphones) or Intel (semiconductors). While these TNCs are primarily interested in new markets, technologies, and resources or lower costs of production, they organize their global production networks by searching and scanning different locations throughout the world (Coe and Yeung 2015; Yeung and Coe 2015).

From the perspective of the nation-state “hosting” these TNCs, nevertheless, the benefits of their presence must outweigh the costs over time. Revisiting the earlier discussion of the East Asian developmental state, at one time foreign corporations were not favored in industrial sectors designated for domestic “national champions.” The potential cost of foreign corporations dominating these sectors was perceived
CORPORATIONS AND THE NATION-STATE

as far greater than the low skill employment and limited technological linkages they might bring. But in many of today’s developing countries, endowed with much weaker state capacity and political control, nurturing and relying on emerging “national champions” is not an option for kick-starting their domestic industrialization. Despite the potential cost of crowding out domestic corporations, TNCs have become sought after through the provision of huge investment incentives and the removal of regulatory barriers. This discursive movement away from the “sovereignty at bay” view of TNCs by developing country states represents a significant shift toward a more neoliberal form of economic development. In more advanced industrialized nation-states, such pro-TNC stance also is increasingly evident with greater recognition of such benefits as enhanced employment, market access, and technologies or knowhow.

As foreign corporations and host nation-states enter into ever more complex bargaining and cooperative relationships, economic geographers have become particularly concerned with the territorial embeddedness of these corporations in specific localities and regional economies. When foreign corporations are more willing to localize their operations and engage with domestic corporations in these regional economies, they are deemed as territorially embedded. Crucially, the extent of this embeddedness can often be substantially enhanced through both policy initiatives of the host state at national and regional scales and the availability and quality of pre-existing industrial and/or market linkages. The territorial embeddedness of TNCs tends to be higher in well-established high-tech industrial clusters (e.g., Silicon Valley and Baden-Württemberg) and/or global cities (e.g., London and New York). Conversely, foreign corporations are less likely to be embedded in exports enclaves in developing countries or low-tech regions in industrialized economies.

With hindsight, this geographical specificity in firm–state relations is highly important both in theory and in practice. At the theoretical level, territorial embeddedness represents a key concept in “relational” economic geography, dating back to Dicken and Thrift’s (1992) early work (see Yeung 2005). It provides a useful conceptual apparatus to analyze how corporations, domestic and foreign, are grounded in territorial ensembles coproduced with the nation-state (Dicken and Malmberg 2001; Dicken 2003). It also paves the way for explaining the geographical political economy of regional development through global production networks (Coe et al. 2004; Yeung 2009, 2016; Coe and Yeung 2015).

In practice, territorial embeddedness presents a crucial opportunity for policy interventions by the host nation-state seeking to capture value and gains through the presence of foreign corporations. A variety of empirical studies have illustrated these dynamic firm–state relations in the context of continuous bargaining and policy initiatives. In China’s socialist variety of capitalism (Peck and Zhang 2013), for example, this bargaining relationship between the authoritarian state and foreign corporations is evolving rapidly. In studying China’s automobile industry, Liu and Dicken (2006) point to the critical role of state interventions in “obliging” foreign automobile corporations to embed themselves in different regional markets in China through developing joint venture arrangements with state-owned enterprises. In the North African and Eastern European apparel industry, Smith et al. (2014) and Smith (2015) have shown the continual role of host state policies in enabling domestic industrial upgrading through the presence of foreign corporations and their cross-border production networks.
CORPORATIONS AND THE NATION-STATE

On the other hand, the globalization of national capital is a relatively more recent phenomenon, associated in particular with the rise of sovereign wealth funds (SWFs). Since the mid-2000s, it has begun to attract growing analytical attention from economic geographers (Clark et al. 2010; Clark, Dixon, and Monk 2013; Yeung 2011; Haberly 2014). In theory, SWFs represent merely a common organizational form of strategic investment and asset management corporations. But due to their wholly state ownership, with liabilities limited to these sovereign owners, SWFs are viewed as embodying and exercising more than market principles or market-based logic of portfolio management and financial venturing. Some market critics have argued that the investment objectives of SWFs are almost always intertwined with state objectives that range from domestic wealth preservation and economic needs to geopolitical interests and national security. If these SWFs wake up from their dormancy to become active or even strategic investment corporations in the global economy, this mixing of financial with political imperatives can threaten and potentially destabilize the global political economy. These dual roles of SWFs are often viewed through the lens of what is critically known as a “double bottom-line” mentality.

In globalizing national capital, the emergence of SWFs brings about several important reconfigurations of firm–state relationships that merit further discussion. First, the sheer size and scale of these SWFs confers enormous bargaining power on their investors, that is, state owners, when it comes to negotiating with potential investment target corporations. As the primary investment holding vehicles for nation-states and, in some cases, regional states, SWFs can manage investible funds and investment-class assets ranging from a few hundred million to a few hundred billion US dollars. The Sovereign Wealth Funds Institute estimates that the combined value of the world’s 74 largest SWFs exceeded US$6.3 trillion as of April 2014 (http://www.swfinstitute.org/fund-rankings, accessed on May 20, 2014). The top ten SWFs account for some 76% of this mammoth value, indicating an extremely high level of financial concentration in these top players. From a systemic risk perspective, these top SWFs may well be beyond “too big and too interconnected to fail,” since no individual nation-state or international organization, such as the International Monetary Fund, can mobilize sufficient financial resources to save a large but failed SWF.

Second, SWFs are quite diverse in terms of their historical, sectoral, and geographical origins. While some small public sector investors such as Texas Permanent School Fund (US$30 billion) can be traced back to 1854, the largest SWFs are much more recent. Among the top ten SWFs, Kuwait Investment Authority (1953) and Singapore’s Temasek Holdings (1974) are the oldest. The role of oil and gas wealth in constituting SWFs is also significant, accounting for four of the top ten and 59% of the total value. In turn, this sectoral specificity points to the unique national origins of oil and gas-related SWFs – from nation-states well endowed with these natural resources. Among the top ten SWFs related to oil and gas, only Norway is not from the Middle East. Meanwhile, a significant number of SWFs from China, Hong Kong, and Singapore are noncommodity in their origin, and represent national wealth accumulated from long periods of export earnings and successive budget surpluses.

Third, and not surprisingly, these SWFs have different foreign investment motives and horizons. Some SWFs are more passive in their approach to investing in foreign corporations. Others deliberately take a low profile in their foreign forays, investing in smaller stakes (typically below 1% of equity) and avoiding
direct intervention in corporate management. The above “double bottom-line” mentality should not necessarily be construed as the automatic basis for bad SWF behavior or investment practices. Comparing Singapore’s two SWFs (GIC and Temasek Holdings) to other nonstate global investors, for example, there is no guarantee that privately owned hedge and mutual funds will be more ethical or considerate of “national feelings” when making their investment decisions. Because of their direct links to Singapore’s developmental state and their deep integration with its national regime of capital accumulation, these SWFs are more likely to consider political risks carefully and act less aggressively in order to reduce potential negative reputational effects of their investments on national interests. On financial grounds, these two SWFs are likely to seek longer term returns to investment, providing steady capital and liquidity for host corporations’ assets (Yeung 2011).

This relatively passive approach to firm–state relationships contrasts quite sharply with some Middle Eastern SWFs, which not only have acquired majority stakes in iconic global corporations, but also have taken up directorships proportionate to their equity holdings. Coined “White Knights from the Gulf” by Haberly (2014), these SWFs have taken over some global leading firms in major industrial sectors, restructuring them into formidable industrial corporations in line with their state-led domestic industrialization strategy. For example, the Abu Dhabi Investment Authority (ADIA) spent several billion US dollars to acquire the US corporation Advanced Micro Devices’ loss-making semiconductor manufacturing facilities and Singapore’s Chartered Semiconductor, in order to establish GlobalFoundries in 2009 – a new corporate entity created to spearhead the development of semiconductor industry in Abu Dhabi. In the same year, Qatar Investment Authority acquired a 17% stake in Germany’s Volkswagen for US$10 billion, giving it a seat on Volkswagen’s supervisory board and a firm foot in one of the world’s most competitive automobile manufacturers (Clark et al. 2010, 2288; Haberly 2014, 17–19). The intimate role of SWFs in bringing together firm–state relations is clearly evident here. They serve as tools of engagement for a nation-state to gain access to global productive capacity unavailable in its domestic economy.

Coda

Taken together with the earlier discussions of the changing role of the nation-state in global political economy and the territorial embeddedness of foreign corporations, this globalization of national capital presents an exciting challenge to future economic-geographical research because of its multiple juxtapositions of different economic actors (nation-states and corporations), industrial sectors (finance and nonfinancial industries), and governance systems and jurisdictions (state-centric and transnational). As a relatively new and yet potent force in shaping global political economy, SWFs and their global reach are significant. Studying them geographically can help unravel the mutual dependency and co-evolution of different economic actors such as corporations (private and public) and nation-states. Such an analytical approach can advance economic geography well beyond the neoclassical and behavioral heydays of location theory and the geography of enterprise, bringing together the two major subfields of economic and political geography. This would effectively address the challenge outlined at the beginning of this entry – a reorientation toward new understandings of the uneven capitalist global economy.

Future geographical research into SWFs also can help transcend the sectoral boundaries and
industrial specificities in many existing studies of corporations. In financial geographies, these nation-state economic instruments are substantial and important subjects in their own right. But their economic and noneconomic linkages far exceed the financial sector itself and/or even most global financial networks. SWFs cut across different global networks, from production, innovation, or financial to social or even environmental. The underlying political power of this national capital and its market-making capacity has the potential to transform massively current geographies of the global economy.

The governance of the global economy remains an elusive concept and practice, partly because of excessive state-centrism in the social sciences (Agnew 2013; Yeung 2014, 2016). As explained in this entry, corporations and nation-states are the most significant actors in shaping today’s capitalist global economy. But their significance should not preclude or overshadow other actors, geographical loci, and cross-border flows and networks constituting the broader governance of this global economy. As argued cogently in Clark et al. (2010, 2274–2275), “overly territorialized models of capitalism that categorize countries differently according to static institutional criteria can easily obfuscate the deep interlinkages, interdependencies, and commonalities between different political-economic spaces.” Future directions in geographical political economy must eschew a simplistic form of universal capitalism and its underlying methodological nationalism; it should instead embrace a geographically variegated understanding of how individual sovereign forms of national capitalisms – neoliberal, welfare, developmental, authoritarian, and so on – become mediated through transnational flows: not just of people, products, and ideas, but more importantly of corporations and, ultimately, nation-states.

**SEE ALSO:** Branch plant economy; Corporations and local states; Economic development zones; Economic geography; Firms; Global production networks; Globalization and rural areas; Institutions and development; Internationalization; Local embeddedness; Nation-state; Political economy and regional development

**References**


Corporations and corporate power

Corporations are legal entities or juridical persons. Corporations, as an organizational form, represent a group of people/shareholders/owners who are authorized to act as a single legal entity or person. In other words, corporations are afforded many of the legal rights of individuals, particularly in matters of property, contract, and torts. There are provisions in corporate law that allow for a variety of corporate forms, but in lay and scholarly discourse the term is most often used to refer to for-profit firms organized generally around the pooling of investment capital, the issue of transferable shares, and the vestment of oversight in a board of directors taken to be distinct from the management of the firm’s day-to-day operations. Ensuring limited liability for members or shareholders of corporations is fundamental to this organizational form. Common usage most often refers to publicly traded corporations, the shares of which are bought and sold on regulated public exchanges. There are, however, relatively large corporations that remain privately held, owned either by a small number of individuals or by a larger firm. Both privately held and publicly traded corporations come into existence and are regulated by state authority. They must be chartered by or registered with particular governments, a process that varies across jurisdictions but which is generally far more rigorous and more heavily regulated for publicly traded firms.

Charters for substantial joint stock companies were initially issued in the mercantilist period by regents, ostensibly for public purposes, but they contributed to the development of capitalism in the seventeenth and eighteenth centuries in their Western European core. European expansions were facilitated by corporate ventures like the Hudson Bay Company and the East India Company, whose efforts to control vast swaths of the planet received state approval and support in an attempt to consolidate imperial rule. The rise of joint stock firms also fueled a number of the earliest speculative bubbles, notably the South Sea Company in England and the Mississippi Company in France in the early eighteenth century. Both resulted in financial crises, engendering substantial anticorporate sentiment in their respective countries (Garber 1990). In the case of the English South Sea Company bubble, parliament responded by instituting a ban on charters for joint stock companies which remained in place until 1824 (MacKay 2003). In the United States, corporate law first took its inspiration from the English, and was initially characterized by ambivalence and a general sense that charters ought to be issued for public purposes, but it evolved independently over the nineteenth century as individual states, in competition with one another for business, adopted increasingly permissive laws (Maier 1993). Beginning in the 1880s, on the basis of what at least one historian has characterized as a “deliberate misuse of facts” (Graham 1938, 384), the Fourteenth Amendment to the US
Constitution has consistently been interpreted by courts to apply not only to natural persons (e.g., freed slaves) but also to corporations. In the 1890s and early 1900s Congress passed legislation aimed at curbing monopoly (e.g., the Sherman and Clayton antitrust acts), and additional regulation was passed into law during the Depression era. In spite of these limited efforts by the state to check corporate power, corporations have continued to play an increasingly central role in American life. In 2014 the Supreme Court extended the legal rights of corporations to political speech in *Citizens United v. FEC* in 2010 and to religious freedom in *Sebelius v. Hobby Lobby Stores*. Broad legal recognition and rights, coupled with increasingly sophisticated organization and communications, have converged to make possible corporate bodies with millions of employees worldwide, annual earnings in tens of billions of dollars, market capitalizations approaching one trillion dollars, with one-third of all international trade now between branches of the same corporation. Corporate holdings have become so Byzantine that it is often impossible to trace the chain of their ownership (Ellwood 2006). According to a study conducted in 2000 by the Global Policy Forum (Anderson and Cavanagh 2000), 51 of the largest 100 economies in the world were organized by corporations and not countries. General Motors was bigger than Denmark, and Walmart than Poland or Greece.

Transnational corporations (TNC) are firms with the power to coordinate and control business operations in two or more countries, even if they do not own them (Dicken 2007, 106). The chief international draws for TNCs are market size and their desire to access global labor and knowledge pools. As TNCs expand horizontally, across nations, they compete through market adaptation by modifying products to suit the requirements of local markets. The transnational corporate model emerged particularly during the 1930s, on account of political-economic factors incentivizing them to decentralize in response to national market differences. Corporations become transnational by developing or setting up new facilities or by acquiring or merging with local firms – the more common methods of entry to gain access to new products and geographical markets (Dicken 2007). As corporations grow, they have greater power to wrest concessions from regional and national governments with promises of job creation and increased national incomes.

Some proponents of economic globalization argue that TNCs are the new ambassadors of democracy. They insist that free markets and political freedom are bound together, with the introduction of the first leading to the second (Friedman 2000; Boockmann and Dreher 2003). They also insist that governments are inefficient and that bloated bureaucracies waste taxpayers’ money, prescribing privatization as a remedy. Even prisons and parks are being privatized as governments reduce public expenditure to meet market demands for balanced budgets. In the United States alone the total expenditure for prisons, jails, and other correctional facilities in 2007 was $74 billion (Ellwood 2006; Kyckelhahn 2010); hence corporations see a tremendous business opportunity here and continue to lobby and increase their share in the prison industrial complex.

Decreasing rates of profits on account of inter- and intracapitalist competition propel surviving corporations toward growing into monopolies and oligopolies. The most profound consequences of big business corporate capitalism are sociopolitical: monopolies and oligopolies are perceived to play a strategic role in capitalism’s tendency to endanger the institutions on which it depends (Marx 1906, 836–837). Capital, through its creative success and manifestation in
the form of corporate power, leads to its own destruction, undermining the very institutions on which it depends, for example the state or supportive political and civil organizations (Schumpeter 1950, 61). Public perceptions of corporate overreach in domestic politics and foreign affairs, as well as glaring examples of corporate irresponsibility with regard to consumer safety, ecological malfeasance, harsh labor discipline, and the socialization of risk have made corporations the subject of broad and intense criticism. Shifts in the spatial division of labor have intensified since the 1980s, driven by core capitalist investment in and relocation of semiautomated labor to developing countries, bringing the role of corporations front and center in debates about development and international relations. Although critics have argued that very large and rich corporations have essentially captured electoral systems in the United States, United Kingdom, and elsewhere (Ferguson 1995; Wolin 2008; Peet 2007), there is mounting evidence that large, publicly traded firms also exert considerable influence in global governance and development projects, for example, in negotiations over “free trade” agreements that codify rent seeking by firms from wealthy nations in poorer ones (Baker 2015), and in the increasing influence of a for-profit development–security assemblage embedded within the operations and policies of international institutions (Nagaraj 2015).

Multilateral organizations: trading and evading conflicts

Multilateral organizations are transnational institutions with the mandate to regulate generalized behavioral norms between three or more nations. They can be formal, with rules and procedures, or informal, representing coalitions of interests, like the United Nations or the European Union. They can be international in focus, fostering integration between countries such as the South Asian Association of Regional Cooperation or the World Trade Organization (WTO), or supranational, with power vested to govern certain aspects of member states. Multilateral organizations can be strategic (like NATO), political (e.g., UN voting blocks), sociocultural (UNESCO), or economic (the World Bank or WTO). Here, the focus is on multilateral organizations with direct or indirect relationships with economies and corporations.

Multilateral forms arose in the modern era to mediate matters of sovereignty and territoriality where uni- and bilateral solutions were inadequate, for example, the establishment of an international maritime order during the early era of European colonialism in the Americas and elsewhere. In the interval between the Napoleonic and Crimean wars, the five leading powers of the Concert of Europe – Austria, Great Britain, Prussia, Russia, and France – forged a multilateral convention aimed at preserving peace and fending off revolutionary threats to dynastic power (Ruggie 1993). By the mid-nineteenth century, the concert had effectively disintegrated as sociopolitical transformations dented the domestic legitimacy of participant regimes and reconfigured the European balance of power. Great Britain, the pre-eminent industrial capitalist state, moved to liberalize its own trade policy, later pursuing strategies that promoted bilateral liberalization. Increasingly, European bilateral trade pacts came to include most-favored nation provisions that guaranteed parity to each party, with concessions granted to third-party states in other bilateral agreements. Regardless of intent, in the latter third of the nineteenth century these engagements tended toward a multilateral form that coalesced into a liberal trade regime based
on free trade and the gold standard (Ruggie 1993, 20–21).

Whereas multilateral forms were only rarely embodied in formal organizations prior to 1900, multilateralism thenceforth became much more formal in the wake of war and shifts in global power. The ad hoc international gold standard, of which the British had been custodians, ruptured. During World War I, Britain’s payment imbalances afforded the United States an opportunity to transition from debtor to creditor. Having cemented its role as international financial, industrial, and military power, the United States was poised to take the lead in organizing “multipurpose, universal membership” international organizations: first the League of Nations and then the United Nations (Ruggie 1993, 22). The emergence of the United States as a dominant power after the two world wars has been critical to the character of the major international institutions that have evolved since. Their multilateral and formal status was not a given, but was contingent on the domestic and foreign policy positions of postwar America.

At the close of World War II, responding to a war-weary public ambivalence toward foreign entanglement and a need to engage Moscow, the Roosevelt administration moved toward a model of multilateral security collaboration in the form of a universal membership body, the United Nations. The United States also took on a critical role in the new international monetary order negotiated at the Bretton Woods Conference in 1944, based on fixed but adjustable currency exchange rates pegged to a gold-convertible dollar (Eckes 2014). The Bretton Woods Conference laid the groundwork for the International Monetary Fund (IMF) and the International Bank for Reconstruction and Development (IBRD, subsequently the World Bank). Later, the United States would call on UN member states to charter an International Trade Organization (ITO) with the aim of negotiating reciprocal trade liberalization. While the US Congress refused to ratify this charter, elements of the ITO proposal were incorporated into the General Agreements on Tariffs and Trade (GATT) (Goldstein 1993). In the late 1940s European anxiety over the tightening sphere of Soviet influence spurred collective security engagement for Western Europe and anglophone America in the form of the 1949 North Atlantic Treaty. With the Soviet Bloc Warsaw Pact, NATO, the United Nations, and the new international trade and monetary order (the IMF, IBRD, and GATT), the postwar international institutional framework coalesced into a world order characterized in the West by American hegemony, Cold War security anxieties, and more or less Keynesian, managed capitalism, driving reconstruction in war-wrecked Europe and Japan.

The strongly institutional forms of multilateralism that emerged after World War II are novel in their organization and scope. Under the security and economic regimes put in place in the wake of the war, the “first world” saw rapid economic growth and impressive industrial output, though Cold War tensions continued. By 1971, the costs of US engagement in Vietnam, which slowed productivity and export growth, and inflation prompted the 1971 Nixon shock in which the dollar was rendered inconvertible to gold. This was followed by the reconfiguration of the international monetary order from a gold-backed system mediated by convertible dollars and pegged currencies, to one of floating fiat currencies priced via currency exchange markets. Characterized by successive oil shocks, and high and rising inflation alongside high and rising unemployment, and slow growth, the stagflation of the 1970s proved amenable to an ideological shift among anglophone political elites away from postwar Keynesianism and the regulatory state (Peet et al. 2003, 117). Beginning in the
Reagan–Thatcher era, international institutions like the IMF, World Bank, and WTO (formerly GATT) were reconfigured along monetarist lines and refocused on minimizing state intervention in markets. These preferences were reaffirmed in the wake of the 1991 collapse of the Soviet Union and the relative prosperity in the West during the 1990s. These events and shifting intellectual currents ushered in the period of market triumphalism which informs contemporary international economic institutional policy and has legitimated a political economic and ideological tendency called neoliberalism, which is oriented toward rapid liberalization, fiscal austerity, minimal state regulation of markets and business activity, and privatization of state-owned enterprises (Peet et al. 2003; Klein 2007).

Summarizing, contemporary multilateral organizations are products of the historical circumstances of both world wars, postwar American power, and successive geopolitical and financial crises. Their historical development has been marked by an at least nominal emphasis on equality between member states in both security and trade agreements and by formal organization. Their rise to global prominence has paralleled the development of corporate personhood and the centrality of business corporations to daily life in much of the world. The rest of this entry will examine the overt and subtle interplay of multilateral institutions and business corporations.

**Corporate–multilateral interplay**

The relationship between business corporations and the state, while at times conflictual, is generally characterized by mutual necessity. On the one hand, the interests of corporations differ from those of the state, the former organized around the generation of profits for shareholders and the latter around material welfare and the social cohesion of its citizens. On the other hand, states need corporations, and corporations need the state. States need firms to produce goods and provide employment for their people. Business corporations need the state for both physical and social infrastructure, from roads and bridges to the enforcement of property rights and the production of an educated workforce. Conflicts arise, however, on the basis of disparate interests and objectives, the negotiation of which shapes much of the politics of the capitalist core and beyond.

Competitive pressures inherent in capitalism ineluctably drive firms in search of new markets and lower costs. To these ends, business corporations have often extended their operations across national borders in the pursuit of new markets, to exploit lower prevailing wages in developing countries, to secure access to raw materials, and to invest (Harvey 2005). From the perspective of the state, international expansion by domestic firms is a means to secure or maintain market position and is thus in the national interest. Public opinion is often split, however, largely owing to the perception that foreign investment by domestic firms translates to higher domestic unemployment and trade imbalance. Developing countries often feel pressure to attract inward foreign direct investment, competing to secure it via all manner of inducements from preferential taxation arrangements to relaxed regulations. The assumption is that foreign investment will create jobs, generate revenue, and stimulate growth (Ahmed 2011). Inward foreign direct investment can create problems in both developing and developed countries, which often derive from real or perceived threats to domestic producers, firms, and workers, or from nationalist sentiment aroused by economic activity that is seen as challenging national sovereignty and territoriality. As the logic of the world economy, in multiple forms, has transcended the scale of
Corporations and Multilateral Organizations

The nation-state, multilateral organizations have become more relevant than ever (Knox, Agnew, and McCarthy 2014).

To an extent, it is important to focus on the United States given its position as the world’s largest economy (backed by the world’s most powerful military) and its hegemonic position. In this context, the nexus between the US government, TNCs, and multilateral organizations assumes tremendous significance (see Figure 1).

In general, foreign direct investment has risen sharply around the world since 1990. In the case of the United States, between 1990 and 2010, outward investment increased about ninefold and foreign investment into the United States trebled, with inward FDI outpaced in dollar terms by US business engagement abroad. These changes reflect increased market liberalization, and shifts in the US domestic market and the global spatial division of labor (particularly in manufacturing). Increased business abroad strengthens motives to engage with global governance institutions like the WTO. American business corporations have evolved not only along lines of efficiency and organizational improvement but also in political management. Their massive influence over domestic politics within the capitalist core, particularly in the United States, is the vector sum of interinterest wrangling, but the general tendency has been the prying open of foreign markets and elimination of foreign capital controls and financial regulation. Business corporations manage the policy preferences of global governance not only through domestic politics, but also through expert opinion expressed via trade association and industry experts and intellectual apparatuses such as think tanks (Peet 2007).

For an example of the influence of corporations over international development work, one need look no further than the US Millennium Challenge Corporation (MCC), an aid agency formed in 2004 under the George W. Bush administration. The MCC provides aid to developing countries through development contracts. Candidate countries are evaluated on 17 criteria before they are offered such a contract and aid can begin. The evaluation of each of these criteria is conducted not by MCC but by a third party, such as UNESCO, IMF, and the World Bank (none of which is surprising). However, the “Trade Policy” criterion is evaluated by the Heritage Foundation (MCC 2015), a big business-friendly think tank whose funding remains largely opaque but has been associated with corporate interests, from ExxonMobil to Koch Industries (Robbins 2015). The MCC is not itself strictly multilateral as it works through bilateral development agreements, but its preferences are generally aligned with and diffuse into the broader international aid and development community with which it regularly interacts (IMF, World Bank, etc.). This exemplifies a direct flow of business corporate influence from corporations and their wealthiest owners and managers into a think tank with immense power.

Figure 1 Geography of corporate and multilateral power.
over the trade policies of developing countries desirous of US development monies and other aid. In this way, facilitated by the global governance institutions, the will of American business becomes the reality in developing countries.

Trade agreements are negotiations between states, acting in what their representatives take to be their national interests. In the process, insofar as that interest is often equated to the designs of the largest and best-connected domestic corporations, their will is often transmitted into global governance and multilateral endeavors. For example, the Trans-Pacific Pact has drawn fierce criticism for its rent-friendly intellectual property provisions that broaden and lengthen protections for intellectual property holders over and above the typical anxieties among labor groups over job losses. Correspondence obtained through Freedom of Information Act requests reveals a close working relationship between US trade representative staff, business corporation leadership, and paid lobbyists, almost certainly representing a level of access and influence unparalleled by any other segment of US society (MCC 2015). At the level of bureaucracy, domestic and international trade and governance personnel, corporate management, and policy think-tank funders and intellectuals very often overlap, engendering strong social ties often referred to in the US context as the economy–society–polity revolving door phenomenon (Vidal, Draca and Fons-Rosen 2012).

Such relations between powerful states and their corporations affect developing economies in different ways, engendering contentious debate over motives and outcomes. Claims that liberalization and foreign direct investment represent an unmitigated good, promising growth, technology transfer, and broad improvements in living standards, are challenged in the geographic literature by numerous case studies demonstrating a much more complex reality. Developing countries that enter into loan arrangements and other aid contracts with multilateral organizations do not do so on equal terms. When the IMF or World Bank extends credit to countries, the money comes with conditionalities that, since 1980, have centered on broad liberalization: lowering tariffs, eliminating subsidies for domestic industry, privatizing state-owned enterprises, eliminating capital controls, and other deregulation (Ahmed 2014). These reconfigurations of domestic policy and international trade relations can have deleterious effects for domestic producers and firms, while at the same time affording considerable opportunities for profit-taking by foreign firms.

Consider, as an example, the Indian power industry. In the early 1990s India received structural adjustment loans from the World Bank and IMF to cope with a balance of payments crisis. As part of structural adjustment, the Indian system of power generation was effectively privatized, with significant protections for the American firms (particularly Enron), which were granted power generation contracts that included pegged utility rates and guaranteed sales (Ahmed 2010). When the terms of the deal precipitated the collapse of Indian electricity distributors, US energy executives exerted pressure on the George W. Bush administration to threaten sanctions to enforce its terms. This episode makes clear the strong relationships between states, business corporations, and global governance institutions. American firms worked with the latter to secure strong market positions in the Indian power industry at very favorable terms, an opportunity afforded them by the coercive power of conditionality in international lending, the dominant role of the United States in the administration of global governance institutions, the promulgation of neoliberal politico-economic ideals, and the political influence of American corporations.
Contrary to the rhetoric of advocates for neoliberal reforms, the outcomes in cases like this look much less like a rising market tide that lifts all boats than a form of state-supported piracy.

Another linkage between business corporations and multilateral institutions is an emerging for-profit international development–security assemblage. The blurring of development and security has accelerated since 2000, in part owing to the prosecution of the “global war on terror,” such that by 2008 a World Bank president called for “bringing security and development together” (Nagaraj 2015, 606). Very large international development contracts are increasingly awarded to very large for-profit corporations, mostly originating in the anglophone West, and are created through successive rounds of merger and acquisition (Roberts 2014). As these firms have grown, they have broadened their scope, often evolving from engineering or agricultural consultancies into one-stop shops for international development contracting. In 2011 a single such firm, Chemonics, received enough USAID (United States Agency for International Development) funds to rank third among nations in terms of the dollar value of aid received. Recently, some of these firms have taken on roles as providers of security, accompanying American and British troops into Afghanistan, Iraq, and Pakistan. As the geopolitical meanings of development and security have blurred and converged, so too have the agendas of development corporations, whose bottom lines remain plump in times of war and peace. Corporations whose business has been to develop infrastructure, facilitate commerce, and foster civil society are now equipped to provide policing and private security services as well as logistical support to military forces deployed around the world.

Conclusion

Business corporations exert considerable influence on national politics, global governance, and international projects. The rise of corporate personhood in the capitalist core has enabled very large-scale accumulations of capital and investment in correspondingly large projects. Broad legal recognition and rights, coupled with immortality and limited liability for shareholders, enables business corporations to wield economic and political power that is readily convertible into influence over the ideology, policy, and operations of multilateral organizations. Corporations enjoy unparalleled access to and influence over international policymakers and bureaucracies, influence that is informed by the channeling of profits to corporate owners and managers rather than by national interests or politico-economic ideals (not even that of a free market for corporations), and that often exhibits undemocratic and antisocial tendencies. Some obvious ethical questions for future research emerge around the readiness of business corporations to embed themselves as easily in military invasions and occupations as in nominally humanitarian projects, and the degree to which corporate power emanating from rich countries underwrites arguably predatory global financial and development activity under the auspices of multilateral institutions – often at the expense of very vulnerable people.

SEE ALSO: Corporate identity; Corporations and global trade; Corporations and local states; Corporations and the nation-state; Emerging market: southern corporation; Globalization; Neoliberalism
References


CORPORATIONS AND MULTILATERAL ORGANIZATIONS


Corporations and e-commerce

Barney Warf
University of Kansas, USA

Essentially, e-commerce consists of commercial applications of the Internet. These include a variety of activities, including business-to-business (B2B) transactions and those linking businesses to their customers (B2C). B2B e-commerce constitutes the bulk of such activity and takes a variety of forms, including electronic data interchange (e.g., inventory data, digital invoices and contracts, purchase orders, and product updates), Internet recruiting, and advertising. (Related activities such as back office and call center relocations are not discussed here.) B2C e-commerce includes electronic retail shopping (e-tailing), Internet banking and gambling, distance learning, and entertainment, where the digital convergence of the Internet, telephony, and broadband cable has allowed companies to bundle these services.

This entry approaches e-commerce in two ways. First, it offers an overview of the varieties of types of e-commerce, including its B2B and B2C forms. It also includes the Internet’s effects on the tourism industry, higher education, and gambling. Second, it focuses on the geographies of e-commerce, noting that because it is embedded within widely varying social, legal, and cultural contexts, it takes a variety of forms in different places.

The varieties of e-commerce

E-commerce has decisively restructured how firms do business. The rising popularity of e-commerce forces firms to change their business models, giving rise to new products and processes and annihilating older ones in a ceaseless process of creative destruction. For the most part, these activities are restricted to large commercial actors, although the Internet also creates opportunities for small and medium-sized establishments (SMEs) to reach out to national and global markets (Grandon and Pearson 2004), a process that allegedly “levels the playing field” and obviates the advantages of size.

One important dimension is electronic data interchange (EDI) systems, which are widely used in B2B contacts. Common uses of EDI include low cost advertising; online product catalogs; the sharing of sales and inventory data; submissions of purchase orders, contracts, invoices, payments, and delivery schedules; product updates; and labor recruitment. E-commerce reduces delays and marketing and delivery costs, improves supply chain management, and has led to a greater emphasis on connectivity, ideas, creativity, speed, and customer service. For example, in insurance e-commerce has streamlined underwriting policies and centralized database management systems. E-commerce also has improved the reliability of product delivery, such as using radio frequency identification devices (RFID) on trucks and boxes of goods. B2B e-commerce includes electronic
advertising, which is relatively cheap, although its effectiveness varies considerably among national cultures. Internet advertising has proven to be difficult, in part because the Internet reaches numerous specialized niche markets rather than mass audiences.

B2C e-commerce, also known as “e-tailing” or electronic retailing, reveals the growing commercialization of the Internet: in 1993, 2% of all websites were commercial (i.e., “dot com”); by 2012, 70% were so categorized. Shopping on the Internet requires only access to cyberspace, a credit card (or a service such as PayPal), and a parcel delivery service. Such a mode allows effortless price comparisons and intensifies competition among suppliers. In 2012, global e-tailing sales were estimated to exceed $700 billion. The proportion of sales conducted over the Internet varies markedly by economic sector. E-shoppers tend to be relatively young, have above-average incomes, and to be relatively well educated. Web-based banking has experienced slow growth, even though it is considerably cheaper for banks than tellers or automatic teller machines. However, Internet-based payments of bills, mortgages, and insurance have grown rapidly. Of course, the use of e-tailing depends heavily on consumers’ confidence that private information placed online will remain secure. There are real concerns: 0.25% of Internet credit card transactions are fraudulent, compared to 0.08% of non-Internet transactions.

In the United States, e-commerce retail sales exceeded $224 billion in 2012. E-tailing comprised the vast majority of travel reservations ($8 billion/year), 62% of all banking transactions, 17% of computer sales, and 11% of book sales. Despite predictions that “click and order” shopping would eliminate “brick and mortar” stores, e-tailing has been slow to catch on, comprising only 4.6% of total US retail sales, perhaps because it lacks the emotional content of shopping. The most successful example perhaps is Amazon.com, started by Seattle entrepreneur Jeff Bezos, which now is responsible for 60% of all books sold online. Other examples include online auctions (e.g., eBay) and downloading of music files, which has provoked a firestorm of opposition from music companies concerned about infringement of their intellectual property rights and declining over-the-counter music sales. Indeed, e-books now comprise 10% of all book sales, and music downloads exceed purchases of CDs, sending sales of the latter crashing in a downward spiral of creative destruction. iTunes alone comprises 10% of all US music sales. Streaming on-demand downloads of movies, through companies such as Netflix, have increased their share as well. Internet-based sales of stocks (e.g., E*Trade) now make up 15% of all US trades. E-advertising forms only 1% of total revenues in the United States and is overwhelmingly focused on computer and software firms.

The Internet also has had important effects on tourism, the largest industry in the world by employment. Traditionally its most important role in the tourism sector has been travel and hotel reservation services. However, the Internet has enormously increased the ease of obtaining information about potential destinations, and tourism companies and institutions have changed their promotion and distribution channels accordingly (Buhalis and O’Connor 2005). Internet-based tourism enables companies to organize their activities in new ways, including better communications with clients and suppliers and easier agreements with business partners. For customers, the Internet facilitates price transparency, provides better visual and graphic information, and generates time savings.

Universities also have become involved in e-commerce, many investing heavily in web-based distance learning courses. Such programs
are designed to attract nonlocal and nontraditional students, many of whom may not be able to take lecture-based courses in the traditional manner, also reflecting the mounting financial constraints and declining public subsidies that many institutions face. Many see distance learning as a means of attracting additional students, and tuition, at relatively low marginal costs. The largest example of web-based teaching is the University of Phoenix, based in Arizona but with 400,000 students located around the world, making it the world’s the largest university. Some universities offer free online courses (MOOCs, or massively open online courses) to thousands of students at a time, giving birth to corporations such as Coursera.org. Distance learning has provoked fears that it accelerates the corporatization of academia and the domination of the profit motive; others question whether the chat rooms that form an important part of its delivery system are an effective substitute for face-to-face teaching and classroom learning. Others suggest that distance learning programs may be better suited to professional programs in business or engineering than in the liberal arts.

More morally ambiguous is the growing role of Internet-based gambling systems, which include a variety of betting services, especially concerning sports events, and even online slot machines in which gamblers may use their credit cards (Wilson 2003). (Some complain that online gambling does not adequately substitute for the heady experience of a gaudy casino in Las Vegas or Atlantic City.) Because the geography of legal gambling is highly uneven, the existence of such systems also challenges the laws of communities in which gambling is illegal. Offshore gambling centers have grown quickly, particularly in the Caribbean since Antigua licensed its first Internet casino in 1994. In 2011, an estimated 1000 online casinos, mostly in the Caribbean, attracted roughly 12 million users. More generally, e-commerce has provoked worries about images of child sexual abuse, tax evasion, and sales of illegal goods.

In most of the developing world (excepting China), e-commerce has grown modestly, hobbled by low disposable incomes, modest Internet penetration rates, lack of secure online transactions, and relatively infrequent use of credit and debit cards. Many SMEs in the developing world lack the human capital, computers, and support services to take advantage of Internet-based trading, financing, and product delivery. Nonetheless, some SMEs, such as in Africa, have used the Internet to carve out niche markets for their products (Moodley and Morris 2004) or to promote tourism (Wynne et al. 2001).

Geographies of e-commerce

Geographical scholarship on e-commerce has taken a variety of forms (see Brunn and Leinbach 2001; Malecki and Moriset 2008 for overviews). E-commerce lies at the intersection of literatures concerned with the spatiality of consumption, geographies of the Internet, transportation analyses, and recent perspectives on cities such as learning regions. For Janelle (2001), e-commerce is a means by which the Internet facilitates the globalization of regional and national economies. A major theme of such works is its potential to reduce transportation costs for both consumers and producers, although this assertion has been challenged (Mokhtarian 2004; Weltevreden and Rotem-Mindali 2009). Ren and Kwan (2009) studied the impacts of e-commerce on different types of consumers using travel diaries, noting that those who lived in remote regions with few retail opportunities were more likely to engage in e-commerce. Clearly e-commerce has changed the meaning of accessibility, reducing search costs, and enhancing competition among
CORPORATIONS AND E-COMMERCE

retailers. Another theme is the ability it offers to SMEs to expand their market reach. Some explore whether e-commerce may emancipate consumers from the “tyranny of the region,” that is, traditional retail landscapes, although Couclelis (2004) is dubious.

Because it will always be an imperfect substitute for brick-and-mortar retailing, e-commerce will never lead to entirely friction-free consumption. Therefore, there is no “one size fits all” model of e-commerce. Rather, its forms and impacts are highly dependent on the contextual specifics of individual regions. As Gefen and Heart (2009) point out, e-commerce is intimately wrapped up with cultural norms of trust, reputation, familial contacts, tacit knowledge, and other related dimensions. E-commerce is also differentially enabled and constrained by national legal systems, including intellectual property rights and the degree of security of online transactions, all of which conspire to give it a very real regional geography.

Despite the common view that cyberspace is placeless, e-commerce in many respects conforms to the locational logic of nonvirtual companies. For example, although the United States is one vast national market with relatively few internal cultural and political barriers, e-commerce is nonetheless unevenly distributed among its cities and regions. Zook (2005) offers a detailed portrait of the regional Internet economy of the San Francisco Bay Area, a churning hub of dot.com entrepreneurship in the 1990s and 2000s, with companies such as Yahoo!, Google, America Online, Ebay, and Mosaic (which later gave birth to Netscape and Hotmail). This analysis goes far to debunk the myth of the “spaceless” Internet economy, pointing instead to the reality of regional clusters in which agglomeration economies, including close linkages to venture capitalists, still figure prominently. As the largest single concentration of Internet-based firms in the world, the San Francisco–Silicon Valley complex operates, ironically, very much through the circulation of tacit information in narrowly defined social channels. Similarly, German e-commerce firms tend to consolidate spatially in distinct clusters, particularly in Berlin and Baden-Wurttemberg. Germany’s success in this regard reflects its long tradition of adopting new technologies, its history of mail order and catalog houses, and the politics of retail regulation (Aoyama and Schwarz 2004). In Italy, e-commerce supplements rather than substitutes for the highly successful model of industrial districts such as in Emilia-Romagna.

It is important to emphasize that e-commerce cannot be understood independently of its local and national cultural contexts; it is socially and spatially embedded. Telecommunications are an imperfect substitute for face-to-face contact, including eye contact and body language, which are essential to the transmission of tacit (as opposed to standardized) knowledge. Without the exchange of tacit knowledge, which is often irregular in nature, not printed, and changes quickly, it is difficult to establish trust. The creation of trust is a problematic process that is often time-consuming. For this reason, many businesses and consumers approach e-commerce with caution, which helps to explain its relatively slow rate of diffusion in cultures in which trust is impaired by prevailing cultural norms. For example, in economies traditionally based on cash and face-to-face interactions, such as in much of East Asia, Internet-based financial transactions diffused relatively slowly.

Conclusion

E-commerce is profoundly changing how and where corporations do business, accelerating product cycles, enhancing competition,
improving efficiency, and reducing uncertainty. As the Internet has steadily expanded in size and scope, it has become increasingly essential for ever larger numbers of firms to reduce costs, reach markets and clients, and manage their supply chains. Geographically, this process has enhanced the locational flexibility of many firms. It would be unwise, however, to assume that the growth of e-commerce unfolds in the same way everywhere. The implementation of e-commerce is profoundly shaped by national legal systems, labor markets, corporate cultures, and the views and practices of consumers, all of which exhibit significant spatial variations. In short, geography matters: no “one size fits all” model can explain e-commerce in diverse economic, political, and cultural contexts.

SEE ALSO: Consumption; Corporate spatial organization and producer services; Corporations and global trade; Economic geography; Industrial linkage; Internet and global capitalism; Transaction cost

References


Corporatization of race: an American case study

Bobby M. Wilson
University of Alabama, USA

Many who have sought to disentangle the complexities of advanced capitalist societies have accepted the Marxist idea of the centrality of production in political economy. Corporations are the key operational units of capitalist economies. Each corporation operates in geographical space, with multiple locations and connections to suppliers and customers. Yet the structural imperative of the corporate world is not only production expansion but also consumer spending. In order to realize the profits undergirding long-term viability, corporations need everyone’s needs. Marx recognized the important role of workers as consumers. Historically, labor situated its politics at the intersection of production and consumption, building worker solidarity while also mobilizing workers’ purchasing power in its struggle with capital. Yet anglophone Marxist scholars have failed to grasp that transcritical moment when production and consumption intersect, making the distinction between the two a difficult dialectical synthesis. Just as much surplus value can be made in retailing as in production or manufacturing. Indeed, the world’s largest corporation, Walmart, is a retailer which subcontracts manufacturing. More generally, the offshoring of production, together with the decline of public sector jobs, brought about the disappearance of the black corporate working class. Yet, as US corporations closed domestic factories, they gradually shifted emphasis to distribution and consumption.

Combining case history and analysis, this entry examines how blackness and whiteness underpins the operation of corporations in the United States. Race-connected practices are not self-determined but are constituted by their corporate relations with production, distribution, and consumption. As the nation’s largest retailer, Walmart carries a meaning for consumers that extends well beyond the price of goods and services (Weston 2009). Since the 1880s, the court has ruled that corporations are in fact people entitled to free speech rights guaranteed by the Constitution. Race can be a barrier to capital’s expansion of commodity exchange, but it can also be used to facilitate such an expansion to build empires. In order for capital to govern production, distribution, and consumption, it also regulates race-connected practices to stay within reach of the black domestic market. This entry focuses first on how corporations in the United States were able to extract surplus under the rules governing the racial doctrine of “separate but equal.” It then examines how corporations culturally manipulated the black consumer to extract billions of dollars in profits from the black domestic market, thereby underdeveloping the black community (Marable 2000).

Separate but equal corporate regime

Race is just one of many barriers that capital seeks to overcome. Race-connected practices are diverse, fluid, and varied rather than fixed and monolithic. Race asserts itself in unexpected ways across time and space. Race, both in the
past and now, is a product of this tension. It cannot be seen merely as a reflection of racial ideologies. It must be situated within a framework that understands and critiques capitalism as an ideology that sustains and reproduces the capital relations of production.

While race can be a barrier that capital seeks to overcome, it can also be used to underpin capital through the production of differences marked by and constructed through race. Capital’s earliest renderings of the production of difference took the form of black slavery and the seizure of indigenous Indian land. The transatlantic slave trade became tantamount to the Western world’s first multinational corporation, which was responsible for the exportation of almost 12 million Africans to the so-called New World where they were sold as commodities on the slave market. From its origins as a manufacturer of soap, Unilever became a global corporation on the basis of profits from its African-based enterprises (Marable 2000). The seizure of indigenous land in America forced the removal of approximately 70,000 Native Americans to reservations beginning in the 1830s. Land taken from black families in the twentieth century became corporate country clubs, and the US South emerged as a leading sporting destination for corporate elites. Many of these gaming areas now include landing strips capable of handling corporate jets.

The racial doctrine of separate but equal regulated US race-connected practices. This doctrine lasted from the time of the Plessy (1892) decision to the 1954 Brown decision. Status was fixated in whiteness but also in the market. It meant that capitalists could separate themselves from blacks but had to remain within market reach. Race facilitated commodity exchange and consumption while also serving as a barrier to such an exchange. Blacks purchased racial pride and respect by consuming certain commodities. The corporate world thus drew on black consumption while also providing the sociospatial structure needed for the continuous authentication of whiteness.

The only decade when evidence shows that the corporation was out of the reach of the black consumer market was 1919 to 1929. It was the most extensive period of racial segregation but also a golden period for the growth of black businesses. It was a time in which virtually every black neighborhood or town claimed a number of independent black entrepreneurs who were providing goods and services to an exclusively black consumer market (Marable 2000, 146). For the white corporate world this lack of consumer demand was not sustainable, given the lessons learned from the Great Depression. Totally separate economies undermine capital’s need to continuously expand the sphere of commodity exchange and consumption and to avoid another depression.

White corporations eventually gained control of the black domestic market. Following the end of slavery, the US Supreme Court never affirmed a common law right for businesses that were open to the public to exclude blacks completely on the ground of race. Nor did the court deny property rights on the basis of race – for capital, access to market, and property ownership must be protected at any cost, even if it meant violating the separate but equal doctrine. For most of the twentieth century, the corporation charmed consumers with whiteness to create value and expand commodity exchange and consumption, bestowing exchange value on whiteness. This drove the accumulation of capital. All consumers were deemed white. Henry Ford manufactured his Model-T for the white masses who wanted not only the utility value of the commodity but also the fixed social relations encoded in it.

The 1893 World’s Columbian Exposition in Chicago, designed to advance the causes of
American nationalism, imperialism, and consumerism, exhibited commodities that ridiculed and denigrated blacks for the sake of white consumption. A prominent Midwestern flour-milling firm, R.T. Davis Milling Company, persuaded Nancy Green, a 57-year-old former slave, to become a living advertisement at the fair for the company’s Aunt Jemima Pancake Mix. Green played the role of a stereotypical plantation mammy, winning the milling firm an award at the fair. Aunt Jemima thus became a corporate brand that gained national acceptance as an early icon of America’s emerging culture of mass production and consumption. In this sense Green became what Baudrillard (1983, quoted in Mullins 1999, 43) terms a “simulacrum”: “she materialized a black caricature that never existed in reality, yet that caricature assumed the status of authenticity in the white imagination because of pervasive belief in it,” marginalizing the status of black women who historically performed domestic labor (Mullins 1999, 46). In Michel Foucault’s terms, this racialized representation is a privileged discourse that exercised, in this case, “inordinate influence on the definition, circulation, and naturalization of racial subjectivities and white superiority” (quoted in Rabinow 1984, 108; see 108–113). Race is thus mutually constituted through the formulation of a commodity exchange circuit, a totality with the distinct moments of production, distribution, and consumption.

Following the 1954 Brown decision, consumer demand for a wider range of products made greater choices possible — choices that would have been illegal or impossible under the separate but equal doctrine. As encoding commodities purely in whiteness became increasingly a liability, the corporate world, beginning in the 1960s, moved increasingly toward a more flexible regime of accumulation without the status fixated in whiteness. Separate was no longer considered equal, setting the stage for the Civil Rights Movement and eventually the creation of a more diverse market. The one size of whiteness no longer fit all, ushering in a more flexible regime of accumulation that could better exploit the many dimensions of pluralism. The corporation promoted the production of differences to expand markets. Differences pay. Consumption coincides with the many dimensions of identity: race, ethnicity, gender, sexuality, religion, and language, among others. Geodemographic marketing accelerated this kind of differentiation, which was missed from Marx’s Capital.

Pluralist-led market regime

A political and social push for an identity that is not fixed just in whiteness can be traced back to the 1920s, with the “New Negro Renaissance.” This was not only a period of incredible economic boom but also of cultural ferment. What was held to be “authentically” black began to change, as the popular musical form ragtime gave way to jazz in 1917, when some white New Orleans musicians, calling themselves the Original Dixieland Jazz Band, began to produce their encoded version of the music. Their whiteness paved the way for the widespread acceptance of jazz which entered the mainstream in white skin, even though black musicians had been playing in New Orleans since the turn of the twentieth century (Wynter 2002, 48). The race that gave blacks the everyday “blues” could now make money playing and singing the blues. It was mainly through white artists that much of “black folk culture, black inventiveness, black talent gave the twenties its distinctive image as the ‘Jazz Age’ and dictated the character of mainstream American popular music and culture for years to come” (Wynter 2002, 48). The one-size-fits-all of whiteness meant that black artists could not
CORPORATIZATION OF RACE: AN AMERICAN CASE STUDY

speak for themselves. The “King of Swing,” Benny Goodman, had no problem owning up to the black roots of swing, but many in the white community continued to have problems with this, fixating the dance form in whiteness (Wynter 2002, 48, 69).

By the 1940s, black jazz artists began to challenge this one-size-fits-all racial paradigm in cultural political economy. John Coltrane, Miles Davis, and other jazz artists were encoding a musical style that informed a new black self-consciousness in opposition to whiteness. These black artists rebelled against attempts to “whiten” jazz “to bleach away its African roots,” rejecting values that politically and culturally subjugated blacks (Kofsky 1998, 57). They were posited as the basis of collective identity, after having initially been rejected as a valid form of classification. Having defined the terms of domination in whiteness, “white” colonizers found their rule challenged by people asserting a counteridentity grounded in a celebration of blackness.

Richard King (1996) has observed how the struggles of the 1960s and 1970s were defining moments, driven by race and culture. The 1960s freed people to experiment with a diversity of lifestyles and identities. This cultural turn captured the different identities of a divided America, constructing and deconstructing the identities of blacks, but also women, men, homosexuals, Hispanics, Asians, and others. The movement was emblematic of a counteridentity grounded in blackness. Black ghetto formation during the first half of the twentieth century produced a new black ethos that contributed to an unprecedented degree of black nationalism, especially in Northern urban areas. Whereas the Civil Rights Movement initially sought to deal with power that resided mainly in the ideological state apparatus or in representative institutions like the state and church, the “Black Power” movement sought to address the power that resided in blacks themselves. It located its politics of resistance in the sphere of immanence, on an ontological rather than simply an economic or institutional basis. In doing so, it provided the basis for a new identity in the black community that was encoded in blackness, not whiteness. Resistance against whiteness, the highest justification for this new identity, took the place of political ideological differences and made the hitherto fixed borders more fluid.

At a very early stage of the Civil Rights Movement, with his nonviolent direct action campaigns and proposed march on Washington, A. Philip Randolph attempted to build black support for the Socialist Party of America based on class identity. Randolph kept the movement all-black, however. As Kenneth O’Reilly (1995, 130) put it, “The greatest antidiscrimination event in the first half of the twentieth century began by adopting discrimination as its own policy,” undermining the possibility of a class alliance between the Civil Rights and labor movements. This undermined the availability of class, one of the most powerful alternative means to overcome the vicissitudes of the “primal” identities acquired through family, school, and the market. Class identity was no longer suitable for the management and regulation of an increasingly postmodern society (Aronowitz 1992). But the new black identity also was unwilling to subject itself to a privatized existence in which political identity is at best sporadic. Instead of class, new social and cultural formations, including nationality, race, gender, and sexuality, now provided the basis for group and individual identities (Aronowitz 1992, 97). The managerial use of race was more acceptable than class for those of the disappeared working class, who cannot consume enough to be of interest to corporate America and its encoders. The emergent black consumer is the US black middle class,
leaving out of the picture the majority of blacks, including those of the disappeared working class.

Stokely Carmichael, later known as Kwame Ture, was one of the most prominent, firebrand young (student) leaders of this new identity formation. When Carmichael forcefully sounded the bugle for “Black Power” as a call to reclaim black heritage, it struck a deep chord among the young generation of blacks, many of whom had grown impatient with the “slow” pace of the Civil Rights Movement. When Carmichael forcefully sounded the bugle for “Black Power” as a call to reclaim black heritage, it struck a deep chord among the young generation of blacks, many of whom had grown impatient with the “slow” pace of the Civil Rights Movement. Huey Newton and Bobby Seale, inspired by the revolutionary theories of Frantz Fanon’s *Wretched of the Earth*, organized the Black Panther Party, which introduced a new understanding of black identity and its place in the United States. Socially isolated, economically excluded, and visibly frustrated by the continued absence of material progress, black youth not only had an intellectual fascination with the vision of black power but actually embraced it as their rallying cry, as a social and economic cause worth fighting for. The dissemination of black power throughout the black community signaled a clearly defined appropriation of a group identity, often referred to as “black America.” The emergence of this new paradigm of power provided for more diversity in the realm of culture and employment, countering the one-size-fits-all paradigm of whiteness that dominated the separate but equal regime.

To represent the strength and dignity of blacks, Carmichael countered the white rooster used to symbolize white supremacy in the state of Alabama with the symbol of the black panther. In his critique of the one-size-fits-all paradigm, Carmichael rejected the notion of the universal intellectual who represented all oppressed groups. Black power challenged the conventional identity of blacks with a new identity that held out the possibility of a black politics. It challenged the fixed binary logic of modernity, sweeping the black other into the forefront of politics. A new era had dawned for black America that the corporate world could not ignore, given its need to underpin the continual expansion of commodity exchange and consumption in an increasingly consumer-oriented, diverse society. The Black Power movement opened the possibility for a biopolitical practice to counter white supremacy. Capital transformed and now valorized the black body for consumption.

In seeking to dismantle the Jim Crow system in the South, the Civil Rights Movement was a reform movement that led to an increase in federal transfer payments to the black underclass and in affirmative action programs for admitting blacks to educational and employment programs. Since the unemployed made for lousy or poor consumers, the corporate world strongly supported the extension of these benefits into the realm of consumption. The Civil Rights Movement also set the stage for providing political and economic incentives for maintaining an individual’s sense of racial and cultural identity. But the desire for freedom as expressed in the biblical idiom of collective liberation, in the psychopolitical quest for a transformed sense of self, and in the political emphasis on participatory freedom (i.e., genuine self-determination encompassing both individual and collective dimensions), suggested that the movement had more in mind than the goals entailed by liberal reform (King 1996, 201–202). Martin Luther King retained some skepticism about black power but eventually realized the importance of building a greater sense of racial pride and potency, a new identity among blacks.

In short, the black community became more autonomous, speaking for itself not only as a producer but also as a consumer. Beginning as early as the late 1920s, “Don’t Buy Where You Can’t Work” campaigns were an important protest
strategy that spread from Chicago to corporations across America. In the 1930s the Harlem Housewives League mobilized women to boycott local chain stores for employment discrimination, arguing that blacks should be hired in proportion to the amount of money they spend in the stores. Through television, black activists began to speak in their own voice directly to audiences without the filter of white journalists (Bodroghkozy 2012). It was a black renaissance with a distinct economic bent. National liberation movements, women’s movements, the Civil Rights Movement, and other antiracist struggles are all inheritors of this cultural turn directed at undermining the separate but equal regime’s one-size-fits-all logic of whiteness.

The notion of beauty, so prominent in consumer culture, broadened beyond the classic Western one to take into account the standards of other cultures. The focus on bodies, the subject of a biopolitical power, was not to whiten skin or to straighten hair but to affirm blackness, to bring about a radical transformation of subjectivity to participate in a project of liberation, or a cultural rebound. The black body is also “beautiful.” James Brown’s “Say it Loud, I’m Black and I’m Proud” and Curtis Mayfield’s “People Get Ready” epitomized the cultural rebound that was taking place in the black community. Eager to accentuate and promote what they considered to be a positive racial identity, the black power generation took many of their cues from Africa to express black pride and culture. Dashikis became popular attire for many blacks during the 1960s, and “Afro” and cornrowing were the “natural” looks in hairstyling. Influenced by the existentialist movement, Richard Wright had already predicted this cultural rebound in the 1940s, publishing works such as Native Son (1944), Black Boy (1945), and The Outsider (1953).

For the sake of capital, black slaves were commodities to be bought and sold in the marketplace. By the time of the Civil Rights Movement, the corporate world, in its effort to expand commodity exchange and consumption, was becoming less dismissive of the black body. Capital had become interested not only in the commodification of the labor component of the black body but also in its cultural component. Under a market-led pluralist regime, capital could increase profit through the production of difference marked by race, not just abstract labor.

Recommodification of the black body

These many dimensions of cultural diversity quickly became encoded into the commodity form. For the corporate world to continue to capture and control the black consumer market, it would have to change how it approached blacks as consumers and possessors of exchange value. The one size of whiteness no longer fit all: status became more fixated on market pluralism. The black community came to embrace this recommodification of the black body as a sign of black progress. This is quite unlike slavery, when black bodies also were commodities to be bought and sold. Yet, the function of both recommodifications is the same: to underpin capital, in different ways.

Following the ghetto explosion of 1965, large corporations established departments of community affairs to combat the growing anticorporate influence in the black community. By 1967, corporate interests, led by the Ford Foundation, were deeply involved in financing most major civil rights organizations, a form of recompense to black America for its support of capital. Carmichael thus did not have the last word on black power. To stay within market reach of the black community, corporate market makers re-encoded, as black capitalism, the movement...
for “Black Power,” which had become the revolutionary cry of the black community. Weems (1998, 79) concluded that, “while many in the 1960s associated African Americans with the word ‘revolution,’ it was the corporate market marketers, the producers of media encoded texts, who had actually conducted a successful coup.” The producers of media-encoded text made it easier to penetrate the black community, simplifying its planning and programming for the sake of capital. As Earl Ofari (1970, 85) explained:

[With] the new alignment of the black politician, the black businessman (old black elite) has formed a new vested-interest power base in the black communities. Working hand-in-hand with corporate enterprise, the neo-black elites rapidly transformed the once expressive (and potentially revolutionary) cry of the black masses for black control of the black communities into capitalist control of the black communities. This only reaffirms the American ethic that everything is expendable, including principles. Everything from nail polish to Black Nationalism can be, and is, hustled as long as there is a prospect for private profit.

A call went out from within the black community for the desegregation also of the money market. Jesse Jackson urged black businessmen to move from civil rights to silver rights. Many black liberals and veterans of the Civil Rights Movement went along with this neoliberal shift. It was reinforced by the 1969 appearance of two magazines in the black community, Essence and Black Enterprise. With the aid of white corporate advertising, the new subjectivity that blacks were embracing recoded the commodification of blackness as a sign of black progress, not as a sign of whiteness. These magazines, along with their famed predecessor, Ebony, were visible symbols of black pride as well as of black America’s growing economic and popular cultural base. In its early issues, Essence carried a full-page self-promotional ad with a penetrating message that read: “Subscribe to Blackness.” In a small way, this ad bore testimony to black America’s defiant emergence from centuries of slavery, suppression, and segregation.

In the 1940s and 1950s the publisher of Ebony, John H. Johnson of Johnson Publishing Company, encouraged corporate America to reach out to the “black market” in the same way that it appealed to the white market; by 1964 he made the case for separate but equal racial retailing, almost in the same way that W.E.B. Du Bois had for a separate black economy over 30 years earlier. According to Johnson (1964, 119, 120):

The Negro market … is not a special market within the white market; it is, on the contrary, a general market defined, precisely, by its exclusion from the white market … Negro consumers tend to think, buy, and act in significantly different ways than white consumers … What Negro consumers want now is recognition of their humanity and an industry wide respect for the Negro image … [They] are growing not less self-conscious but more self-conscious … more critical … of media and products and services that deny them acceptable images … The basic factor here is the migrating mood of Negro consumers who have a new appreciation of themselves as non-whites … a newfound appreciation of blackness … It will be increasingly difficult to influence Negroes with ego ideals and ego models that are exclusively white.

For Johnson, the black body would no longer be fixated in whiteness. In the past, the black market was separate but not equal; now it had become separate but equal. Capital encoded in their commodities different identities and perspectives that would better capture and control the growing black consumer market, creating one of the greatest obstacles to a present-day black capitalist strategy.

The Congress of Racial Equality (CORE) organized a drive to convince 14 major corporations to make ads “a more realistic reflection of
the composition of the American population” by using the black body as a model in their advertisements. CORE argued that the positive use of blacks in advertisements would not adversely affect corporate sales. Corporations could still sell and realize surplus value (Advertising Age, September 9, 1963; December 9, 1968). Jesse Jackson’s Operation PUSH (People United to Save Humanity) and the SCLC’s (Southern Christian Leadership Conference) Operation Breadbasket also pressured corporate America, threatening a black consumer boycott, the most dreadful thing for capital to imagine. Martin Luther King also acknowledged the success of the consumer boycott strategy in Birmingham:

> It was not the marching alone that brought about integration of public facilities in 1963. The downtown business establishments suffered for weeks under our almost unbelievably effective boycott. The significant percentage of their sales that vanished, the 98 percent of their Negro customers who stayed home, educated them forcefully to the dignity of the Negro as a consumer. (quoted in Washington 1986, 602; emphasis added)

The corporate world needed little persuasion, however, knowing full well that it had to sell to the black segment of the market. It could mean the difference between making and not making a profit.

This compulsion was evident, at least in terms of buying advertising space in the black print media. Full-page ads in very early issues of Essence and Black Enterprise show that cigarette manufacturers, regardless of their growing concerns about the connections being made between smoking and cancer, were among those who led the way in showing a keen interest in the black consumer. The number of full-page ads increased noticeably in Essence within less than two years of its initial publication. Essence carried merely five full-page ads in its January 1971 issue, but within few months this more than doubled to a dozen. Significantly, the magazine’s corporate client base expanded to include some of the major corporations of the time, including Sears, Levi’s, Pepsi, and producers of Martex, Kleenex, Tampax, Wohl Shoes, and Elan (perfume). Even TWA, Pan Am, Earhart (luggage), and Modern Living by Nancy Heine-man (household furniture) began to buy ads that targeted black consumers without offending them. The gains that Black Enterprise posted in terms of securing corporate advertisements were no less impressive. By 1974, within five years of its inception, its advertisers included such corporate heavyweights as Merrill Lynch, New York Life, Metropolitan Life, IBM, AT&T, Amoco, Exxon, Gulf, GM, Ford, Chrysler, Xerox, Delta, Rouse Company, Puma, Polaroid, GTE, and CBS. In one 1974 advertisement about increasing diversity, marketers at CBS noted: men and women of all races desired. This said it all about the emerging new corporate paradigm.

Through advertising in the black media, white corporate interest was able to solidify its control of the black consumer market, realizing more surplus value. Leading companies bought product advertisements in the black print media whose circulation was predominantly concentrated among black households. In the mid-1960s the late H. Taylor Fitzhugh pioneered the concept of black consumer marketing as an executive at Pepsi-Cola whose ads also appeared in Essence (Graves 1997). Having spent several million dollars in black-oriented advertising, Pepsi-Cola saw its annual profits increase from $157.6 million to $250 million between 1960 and 1964. In its fortieth anniversary issue, Ebony contained 162 pages of advertising worth $6.2 million, offering a vision of a bland consumer culture in multiple shades.
These magazines were not simply icons and outlets for black pride and consciousness. Profit-oriented enterprises themselves, they clearly understood the need for capital to sell and realize surplus value. It is, therefore, no surprise that the black print media served as a confluence not only of black power and pride but also of power and pride in the expansion of the commodity form for the sake of capital. As the popularity and circulation of these magazines grew among black consumers, they provided a genuine alternative and hence competition to white magazines of their genre.

With the black print media acting as a powerful and most visible conduit, black households had effectively become the target of a separate but equal marketing campaign on the part of corporate marketers and producers of media-encoded text. The black print media, represented by *Black Enterprise*, *Essence*, *Jet*, and a host of other newcomers, clearly served a purpose in flexing their muscles to demonstrate black consumer power. The specter of black power was finally hitting the world of marketing research because capital has to sell. Corporate marketers of women’s beauty and personal care products, in particular, sought to sell both “beauty” and “status” to black women, but this time in its black form (Weems 1998, 91). The advertising campaigns for cosmetics and beauty products were particularly racially targeted. The black community had evolved into a potent segment of America’s consumer market.

Black-owned ad agencies used their professional expertise to target black consumers for various corporations and their products. Several of the major black-owned advertising agencies were established in the 1970s, one as early as 1956 but most in the 1980s. In 1992 they spent $785 million just to reach the black market segment alone.

Target marketing, instead of one size fits all, had emerged as the corporate strategy aimed at a particular demographic and lifestyle category at a time when lifestyles were becoming increasingly the basis of social distinction. While market segments are delineated by geography, age, and sex, others contain racial and ethnic components, differentiating them from the mainstream market defined by its dominant white American client base. Marketing segmentation rejects the production of blacks as a replica of whiteness; this leads to blacks, as well as other racial and ethnic identities, becoming a market niche encoded by corporate advertising. It amounts to the commodification of blackness as a cultural or aesthetic category. No longer viewed as dark-skinned white people, blacks had their own identity or culture which could be marketed just like any other commodity. Consumer choice is thereby diverted into a mechanism for differentiating on the base of various dimensions of diversity, underwriting the production, distribution, and consumption of commodities.

This has the effect of expanding markets across multiple borders: the more differentiated or segmented the market, the greater the number of target markets from which to realize surplus value. Culture and media industries accelerate segmentation and identity formation through creating specific delimited consumer markets. Something is provided for all, so none can escape the invisible hand of the market: distinctions are emphasized and extended in these postmodern times.

The postmodern cultural turn produced a brave new world with a greater emphasis on choice and product differentiation; on marketing, packaging, and design; on the targeting of consumers by lifestyle, taste, and culture rather than just by categories of social class. Identities are reflected in lifestyles that are closely associated with brands and the products they label.
This enables consumers to express individual sovereignty through identities, constructed by acquiring and displaying goods that convey information about them and their position within a constellation of social networks. Rather than seeing identity as fixated on whiteness, the one-size-fits-all strategy, identities are now considered plural and dynamic, hybrid and fluid. Consumers carry, as Mort (1989, 169) noted, “a bewildering range of different and at times conflicting identities around with them in their heads at the same time.” A flexible regime of accumulation reconstitutes the production and distribution of commodities to exploit this hybridity and fluidity in society. By recommodifying the black body to enhance consumption, a racialized and more diversified underpinning of the corporation has taken hold.

A flexible regime of accumulation

A main flaw of postmodern analysis is that it tends to ignore the social relations of production or the class structure of capitalism. Market makers, encoders of text, and consumers as decoders of text are not free of structural constraints. The postmodern cultural turn suggests an open and dynamic egalitarian culture, with constant cultural flows and mixtures. Exploiting this openness and mixture in the constitution of production, and the diversification of consumption and distribution, requires a regime of accumulation more flexible than the Fordist regime.

Whereas through the first half of the twentieth century Fordism was producing conformity, by the 1970s a flexible specialization regime was offering greater authenticity, individuality, difference, and rebellion as a market strategy. Corporations could now maximize profits through the social production of difference, marked also by race. Featherstone (1994, 404) notes: “Since the 1960s there has been a more general informalization and elaboration of previously restricted codes of behavior.” If the trauma of World War II meant that the corporate world faced a society mentally wedded to mutual cooperation and togetherness, 20 years later the Civil Rights Movement and the Vietnam War set the stage for a more fragmented society which the corporation was able to exploit in different ways.

Left to its own devices, the corporate world never abandons a profitable regime. The separate but equal regime had been profitable, reproducing the racial hierarchy and making money in the process. It was basically a producer-driven commodity chain whose primary objective was the production and consumption of whiteness. Through the mass production of standardized goods, it counted on adequate consumer demand but had little need to listen to the market. This regime defined the market as massive and white, paying little or no attention to differences within it. Feedback from consumption to production did allow some changes in the market to translate to changes in productive engineering, but this communication circuit was restricted, slow, and lacked flexibility.

The mass market became increasingly saturated as a result of rapid rises in productivity after 1945. Consumer markets began to stagnate, posing a noticeable constraint to economic growth in the industrialized world. The corporate world expanded mass production and consumption, increasing the competition for the consumer dollar. Mass marketing became highly competitive. But additional advertising and promotion yielded diminishing returns. The market for consumer goods was increasingly governed by replacement demand turnover and obsolescence rather than by the opening up of new markets, which had been possible throughout the 1950s and early 1960s. Having reached its maximum potential, mass marketing intensified efforts to open up new markets. But the fixity and
standardization of the separate but equal regime began to pose obstacles to the further progression of capital. A more mobile and flexible regime had become necessary, which was not fixated on whiteness or on a particular identity or status relation.

The separate but equal regime lacked the flexibility to exploit the growing demand for more differentiated products or hybridity in the marketplace. Thus, capital had to engage in the social production of difference. Wendell Smith (1956, 7) had already warned of the growing danger of product competition with mass production and mass marketing. He noted that the theory of perfect competition, the basis of neoclassical economic theory, assumes homogeneity, that one size fits all, but diversity or difference had become the rule. He proposed exploiting “market segmentations,” arguing that, in place of a mass market, goods would “find their markets of maximum potential as a result of recognition of difference in the requirements of market segments.” Concluding that “their core markets have already been developed,” Smith (1956, 7) advised companies that “attention to smaller or fringe market segments, which may have small potentials individually but are of crucial importance in the aggregate,” could be profitable. Capital could maximize its profits not just by rendering labor abstract but also through the social production of difference.

Consumption has become symbolic of social distinctions. Mass-produced products lack status or lifestyle differences, the sign value that targets a particular socioeconomic group to buy. Decades before Baudrillard’s (1983) cultural analysis of the commodity sign, the automobile industry had already started to encode its products in market segments, manipulating the sign value of its commodities to communicate messages about the status or lifestyle of the consumer. Growth required selling cars whose appearance and features become obsolete on an annual basis. It was better to divide up the mass market according to distinctive needs and desires. Instituting model and color diversity, the automobile industry placed its products in different geographical and social contexts to denote a specific rather than a generic product. Production became a bundle of changing characteristics designed, produced, and marketed to appeal to different market segments of the population.

With geodemographic analysis, the corporation is able to predict variations in consumption based on a host of demographic factors such as race and where a person lives. Mass culture becomes micro: smaller fringe markets matter. Commodities are branded on the basis of where they are made. Corporations are increasingly emphasizing localized production and consumption.

The cultural and racial discontent of the 1960s and 1970s made universal or mass marketing increasingly impractical, weakening the strategy of one size fits all. In the corporate world, the fixed relations that characterized the separate but equal racialized corporate regime became increasingly impractical and illegitimate. Corporations become less receptive to the notion of social purification.

It is not coincidental that the newly constituted regime corresponded to the historical era of discontent in the 1960s and 1970s. If it was not for the Black Power movement, the student revolts, the women’s movement, and other antiracist struggles, the corporate world may have been content to maintain the one-size-fits-all corporate strategy, happy to avoid the trouble of having to reconstitute production, distribution, and consumption. Instead, in the new social and cultural formations, capital saw the opportunity to create more value and to expand commodity exchange and consumption by recoding and decoding race, bringing the strange, the other,
into the space of the ordinary. According to Cohen:

social and cultural groups such as African Americans, women, youth, and senior citizens began to assert themselves in a way that came to be called “identity politics,” where people’s affiliation with a particular community defined their cultural consciousness and motivated their collective political action. Marketers, who had they not been in search of ways to avoid saturating the mass market, might have despaired at the splintering of the mass in the 1960s, soon seized the new opportunities for selling it offered. (2003, 308; emphasis added)

Needing to sell to realize value, corporations seized on the opportunity offered by the racial discontent of the 1960s and 1970s to do just that.

Conclusion

Consumption is diverse, fluid, and varied rather than fixed and monolithic. The tension between race and corporate capital’s need to expand commodity exchange and consumption asserts itself in unexpected ways across time and space. Race, then and now, is a product of this tension. It cannot be seen merely as a reflection of racial ideologies. It must be situated within a framework that understands and critiques capitalism as an ideology that sustains and reproduces capital relations of production.

From the consumption of whiteness, the one-size-fits-all paradigm, the corporate world has shifted to a market-led pluralism that recognizes different consumer market segments. This postmodern paradigm shift reached new heights as black cultural elements, once ignored in the marketplace, were widely marketed by corporations. As a result, blackness is no longer represented as the opposite of whiteness but as a commodity to be consumed by all. The classical black sound of rap, the new rhythm of consciousness and resistance, and its hip-hop cultural form are commodified to be bought and sold on the world market. This regime of accumulation is less fixated on whiteness and conformity, on maintaining fixed status relations. The consumerism of postmodernity, held to be incommensurate with the fixed relations of modernity, accommodates differences to expand markets. Contemporary producers of encoded texts welcome the youth-led and hip-hop cultural revolution as an ally in their own struggle to manage and regulate society for the sake of capital. Rap and hip-hop are prime examples of the commodification of race for the sake of capital, having become extraordinarily commercially successful worldwide. Rap is globally produced and consumed more than ever, and in diverse markets. The underlying motivation for positing race and culture as the bases of collective identity was social, but not far behind was the neoliberal need to express race and culture as a commodity form to realize surplus value.

The black print media reproduced neoliberalism in the black community. Yet it often blames black and brown populations for their economic circumstances and for failing to fight against structural factors that keep them economically subordinate. The black corporate world has “earned the confidence of the white corporate hierarchy and the capitalist state by keeping alive the bogus illusion of a black capitalism” (Marable 2000, 158). Yet over four-fifths of all black-owned US firms are small businesses, struggling against massive odds to not fall back into the ranks of the proletariat. Further, the black community lacks an effective voice to articulate the growing black demand for economic opportunities.

More research is needed on the ways in which the corporation reworks cultural elements to expand commodity exchange and consumption,
and the implications thereof. Has this reworking of culture elements produced a culture war over meaning and lifestyle? Can we put a neoliberal face on this culture war? Many condone this reworking of culture because it sells. Hip-hop reproduces neoliberalism by emphasizing narratives that laud violence, crime, and drug use in the black community. To expand commodity exchange and consumption, the corporate world welcomes all, even those from the ‘hood. It manages and regulates this life of diversity opportunistically to make money. Social networking sites such as MySpace and Facebook are invaluable in identifying niche markets of potential consumers for capital.

To what extent has this cultural manipulation of the consumer by capital contributed to the underdevelopment of the black community? The 1968 Report on Civil Disorder by the Kerner Commission documented the pent-up desire for consumer goods that television, movies, and advertising held up. While black consumer spending has increased, few dollars circulate in the black community. Black consumers are foot soldiers for corporate interest, siphoning away dollars from the black economy (Weems 2005, 253). For every $8 to $10 spent in the black community only $1 remains. An informal and predatory community, consisting of illegal practices such as drug trafficking and prostitution, develops as the poor person’s way to cut himself or herself in on the consumer society and to survive in a community that’s leaking a substantial amount of capital.

SEE ALSO: Brands and branding; Geodemographic profiling; Marxist geography; Postmodernity; Corporatization of race: an American case study; Race and racism; Whiteness

References

CORPORATIZATION OF RACE: AN AMERICAN CASE STUDY


Further reading


Many who have sought to disentangle the complexities of advanced capitalist societies have accepted the Marxist idea of the centrality of production in political economy. Corporations are the key operational units of capitalist economies. Each corporation operates in geographical space, with multiple locations and connections to suppliers and customers. Yet the structural imperative of the corporate world is not only production expansion but also consumer spending. In order to realize the profits undergirding long-term viability, corporations need everyone’s needs. Marx recognized the important role of workers as consumers. Historically, labor situated its politics at the intersection of production and consumption, building worker solidarity while also mobilizing workers’ purchasing power in its struggle with capital. Yet anglophone Marxist scholars have failed to grasp that transcritical moment when production and consumption intersect, making the distinction between the two a difficult dialectical synthesis. Just as much surplus value can be made in retailing as in production or manufacturing. Indeed, the world’s largest corporation, Walmart, is a retailer which subcontracts manufacturing. More generally, the offshoring of production, together with the decline of public sector jobs, brought about the disappearance of the black corporate working class. Yet, as US corporations closed domestic factories, they gradually shifted emphasis to distribution and consumption.

Combining case history and analysis, this entry examines how blackness and whiteness underpins the operation of corporations in the United States. Race-connected practices are not self-determined but are constituted by their corporate relations with production, distribution, and consumption. As the nation’s largest retailer, Walmart carries a meaning for consumers that extends well beyond the price of goods and services (Weston 2009). Since the 1880s, the court has ruled that corporations are in fact people entitled to free speech rights guaranteed by the Constitution. Race can be a barrier to capital’s expansion of commodity exchange, but it can also be used to facilitate such an expansion to build empires. In order for capital to govern production, distribution, and consumption, it also regulates race-connected practices to stay within reach of the black domestic market. This entry focuses first on how corporations in the United States were able to extract surplus under the rules governing the racial doctrine of “separate but equal.” It then examines how corporations culturally manipulated the black consumer to extract billions of dollars in profits from the black domestic market, thereby underdeveloping the black community (Marable 2000).

Separate but equal corporate regime

Race is just one of many barriers that capital seeks to overcome. Race-connected practices are diverse, fluid, and varied rather than fixed and monolithic. Race asserts itself in unexpected ways across time and space. Race, both in the
past and now, is a product of this tension. It cannot be seen merely as a reflection of racial ideologies. It must be situated within a framework that understands and critiques capitalism as an ideology that sustains and reproduces the capital relations of production.

While race can be a barrier that capital seeks to overcome, it can also be used to underpin capital through the production of differences marked by and constructed through race. Capital’s earliest renderings of the production of difference took the form of black slavery and the seizure of indigenous Indian land. The transatlantic slave trade became tantamount to the Western world’s first multinational corporation, which was responsible for the exportation of almost 12 million Africans to the so-called New World where they were sold as commodities on the slave market. From its origins as a manufacturer of soap, Unilever became a global corporation on the basis of profits from its African-based enterprises (Marable 2000). The seizure of indigenous land in America forced the removal of approximately 70,000 Native Americans to reservations beginning in the 1830s. Land taken from black families in the twentieth century became corporate country clubs, and the US South emerged as a leading sporting destination for corporate elites. Many of these gaming areas now include landing strips capable of handling corporate jets.

The racial doctrine of separate but equal regulated US race-connected practices. This doctrine lasted from the time of the Plessy (1892) decision to the 1954 Brown decision. Status was fixated in whiteness but also in the market. It meant that capitalists could separate themselves from blacks but had to remain within market reach. Race facilitated commodity exchange and consumption while also serving as a barrier to such an exchange. Blacks purchased racial pride and respect by consuming certain commodities. The corporate world thus drew on black consumption while also providing the sociospatial structure needed for the continuous authentication of whiteness.

The only decade when evidence shows that the corporation was out of the reach of the black consumer market was 1919 to 1929. It was the most extensive period of racial segregation but also a golden period for the growth of black businesses. It was a time in which virtually every black neighborhood or town claimed a number of independent black entrepreneurs who were providing goods and services to an exclusively black consumer market (Marable 2000, 146). For the white corporate world this lack of consumer demand was not sustainable, given the lessons learned from the Great Depression. Totally separate economies undermine capital’s need to continuously expand the sphere of commodity exchange and consumption and to avoid another depression.

White corporations eventually gained control of the black domestic market. Following the end of slavery, the US Supreme Court never affirmed a common law right for businesses that were open to the public to exclude blacks completely on the ground of race. Nor did the court deny property rights on the basis of race – for capital, access to market, and property ownership must be protected at any cost, even if it meant violating the separate but equal doctrine. For most of the twentieth century, the corporation charmed consumers with whiteness to create value and expand commodity exchange and consumption, bestowing exchange value on whiteness. This drove the accumulation of capital. All consumers were deemed white. Henry Ford manufactured his Model-T for the white masses who wanted not only the utility value of the commodity but also the fixed social relations encoded in it.

The 1893 World’s Columbian Exposition in Chicago, designed to advance the causes of
American nationalism, imperialism, and consumerism, exhibited commodities that ridiculed and denigrated blacks for the sake of white consumption. A prominent Midwestern flour-milling firm, R.T. Davis Milling Company, persuaded Nancy Green, a 57-year-old former slave, to become a living advertisement at the fair for the company’s Aunt Jemima Pancake Mix. Green played the role of a stereotypical plantation mammy, winning the milling firm an award at the fair. Aunt Jemima thus became a corporate brand that gained national acceptance as an early icon of America’s emerging culture of mass production and consumption. In this sense Green became what Baudrillard (1983, quoted in Mullins 1999, 43) terms a “simulacrum”: “she materialized a black caricature that never existed in reality, yet that caricature assumed the status of authenticity in the white imagination because of pervasive belief in it,” marginalizing the status of black women who historically performed domestic labor (Mullins 1999, 46). In Michel Foucault’s terms, this racialized representation is a privileged discourse that exercised, in this case, “inordinate influence on the definition, circulation, and naturalization of racial subjectivities and white superiority” (quoted in Rabinow 1984, 108; see 108–113). Race is thus mutually constituted through the formulation of a commodity exchange circuit, a totality with the distinct moments of production, distribution, and consumption.

Following the 1954 Brown decision, consumer demand for a wider range of products made greater choices possible — choices that would have been illegal or impossible under the separate but equal doctrine. As encoding commodities purely in whiteness became increasingly a liability, the corporate world, beginning in the 1960s, moved increasingly toward a more flexible regime of accumulation without the status fixated in whiteness. Separate was no longer considered equal, setting the stage for the Civil Rights Movement and eventually the creation of a more diverse market. The one size of whiteness no longer fit all, ushering in a more flexible regime of accumulation that could better exploit the many dimensions of pluralism. The corporation promoted the production of differences to expand markets. Differences pay. Consumption coincides with the many dimensions of identity: race, ethnicity, gender, sexuality, religion, and language, among others. Geodemographic marketing accelerated this kind of differentiation, which was missed from Marx’s Capital.

Pluralist-led market regime

A political and social push for an identity that is not fixed just in whiteness can be traced back to the 1920s, with the “New Negro Renaissance.” This was not only a period of incredible economic boom but also of cultural ferment. What was held to be “authentically” black began to change, as the popular musical form ragtime gave way to jazz in 1917, when some white New Orleans musicians, calling themselves the Original Dixieland Jazz Band, began to produce their encoded version of the music. Their whiteness paved the way for the widespread acceptance of jazz which entered the mainstream in white skin, even though black musicians had been playing in New Orleans since the turn of the twentieth century (Wynter 2002, 48). The race that gave blacks the everyday “blues” could now make money playing and singing the blues. It was mainly through white artists that much of “black folk culture, black inventiveness, black talent gave the twenties its distinctive image as the ‘Jazz Age’ and dictated the character of mainstream American popular music and culture for years to come” (Wynter 2002, 48). The one-size-fits-all of whiteness meant that black artists could not
CORPORATIZATION OF RACE: AN AMERICAN CASE STUDY

speak for themselves. The “King of Swing,” Benny Goodman, had no problem owning up to the black roots of swing, but many in the white community continued to have problems with this, fixating the dance form in whiteness (Wynter 2002, 48, 69).

By the 1940s, black jazz artists began to challenge this one-size-fits-all racial paradigm in cultural political economy. John Coltrane, Miles Davis, and other jazz artists were encoding a musical style that informed a new black self-consciousness in opposition to whiteness. These black artists rebelled against attempts to “whiten” jazz “to bleach away its African roots,” rejecting values that politically and culturally subjugated blacks (Kofsky 1998, 57). They were posited as the basis of collective identity, after having initially been rejected as a valid form of classification. Having defined the terms of domination in whiteness, “white” colonizers found their rule challenged by people asserting a counteridentity grounded in a celebration of blackness.

Richard King (1996) has observed how the struggles of the 1960s and 1970s were defining moments, driven by race and culture. The 1960s freed people to experiment with a diversity of lifestyles and identities. This cultural turn captured the different identities of a divided America, constructing and deconstructing the identities of blacks, but also women, men, homosexuals, Hispanics, Asians, and others. The movement was emblematic of a counteridentity grounded in blackness. Black ghetto formation during the first half of the twentieth century produced a new black ethos that contributed to an unprecedented degree of black nationalism, especially in Northern urban areas. Whereas the Civil Rights Movement initially sought to deal with power that resided mainly in the ideological state apparatus or in representative institutions like the state and church, the “Black Power” movement sought to address the power that resided in blacks themselves. It located its politics of resistance in the sphere of immanence, on an ontological rather than simply an economic or institutional basis. In doing so, it provided the basis for a new identity in the black community that was encoded in blackness, not whiteness. Resistance against whiteness, the highest justification for this new identity, took the place of political ideological differences and made the hitherto fixed borders more fluid.

At a very early stage of the Civil Rights Movement, with his nonviolent direct action campaigns and proposed march on Washington, A. Philip Randolph attempted to build black support for the Socialist Party of America based on class identity. Randolph kept the movement all-black, however. As Kenneth O’Reilly (1995, 130) put it, “The greatest antidiscrimination event in the first half of the twentieth century began by adopting discrimination as its own policy,” undermining the possibility of a class alliance between the Civil Rights and labor movements. This undermined the availability of class, one of the most powerful alternative means to overcome the vicissitudes of the “primal” identities acquired through family, school, and the market. Class identity was no longer suitable for the management and regulation of an increasingly postmodern society (Aronowitz 1992). But the new black identity also was unwilling to subject itself to a privatized existence in which political identity is at best sporadic. Instead of class, new social and cultural formations, including nationality, race, gender, and sexuality, now provided the basis for group and individual identities (Aronowitz 1992, 97). The managerial use of race was more acceptable than class for those of the disappeared working class, who cannot consume enough to be of interest to corporate America and its encoders. The emergent black consumer is the US black middle class,
leaving out of the picture the majority of blacks, including those of the disappeared working class.

Stokely Carmichael, later known as Kwame Ture, was one of the most prominent, firebrand young (student) leaders of this new identity formation. When Carmichael forcefully sounded the bugle for “Black Power” as a call to reclaim black heritage, it struck a deep chord among the young generation of blacks, many of whom had grown impatient with the “slow” pace of the Civil Rights Movement. Huey Newton and Bobby Seale, inspired by the revolutionary theories of Frantz Fanon’s *Wretched of the Earth*, organized the Black Panther Party, which introduced a new understanding of black identity and its place in the United States. Socially isolated, economically excluded, and visibly frustrated by the continued absence of material progress, black youth not only had an intellectual fascination with the vision of black power but actually embraced it as their rallying cry, as a social and economic cause worth fighting for. The dissemination of black power throughout the black community signaled a clearly defined appropriation of a group identity, often referred to as “black America.” The emergence of this new paradigm of power provided for more diversity in the realm of culture and employment, countering the one-size-fits-all paradigm of whiteness that dominated the separate but equal regime.

To represent the strength and dignity of blacks, Carmichael countered the white rooster used to symbolize white supremacy in the state of Alabama with the symbol of the black panther. In his critique of the one-size-fits-all paradigm, Carmichael rejected the notion of the universal intellectual who represented all oppressed groups. Black power challenged the conventional identity of blacks with a new identity that held out the possibility of a black politics. It challenged the fixed binary logic of modernity, sweeping the black other into the forefront of politics. A new era had dawned for black America that the corporate world could not ignore, given its need to underpin the continual expansion of commodity exchange and consumption in an increasingly consumer-oriented, diverse society. The Black Power movement opened the possibility for a biopolitical practice to counter white supremacy. Capital transformed and now valorized the black body for consumption.

In seeking to dismantle the Jim Crow system in the South, the Civil Rights Movement was a reform movement that led to an increase in federal transfer payments to the black underclass and in affirmative action programs for admitting blacks to educational and employment programs. Since the unemployed made for lousy or poor consumers, the corporate world strongly supported the extension of these benefits into the realm of consumption. The Civil Rights Movement also set the stage for providing political and economic incentives for maintaining an individual’s sense of racial and cultural identity. But the desire for freedom as expressed in the biblical idiom of collective liberation, in the psychopolitical quest for a transformed sense of self, and in the political emphasis on participatory freedom (i.e., genuine self-determination encompassing both individual and collective dimensions), suggested that the movement had more in mind than the goals entailed by liberal reform (King 1996, 201–202). Martin Luther King retained some skepticism about black power but eventually realized the importance of building a greater sense of racial pride and potency, a new identity among blacks.

In short, the black community became more autonomous, speaking for itself not only as a producer but also as a consumer. Beginning as early as the late 1920s, “Don’t Buy Where You Can’t Work” campaigns were an important protest
strategy that spread from Chicago to corporations across America. In the 1930s the Harlem Housewives League mobilized women to boycott local chain stores for employment discrimination, arguing that blacks should be hired in proportion to the amount of money they spend in the stores. Through television, black activists began to speak in their own voice directly to audiences without the filter of white journalists (Bodroghkozy 2012). It was a black renaissance with a distinct economic bent. National liberation movements, women’s movements, the Civil Rights Movement, and other antiracist struggles are all inheritors of this cultural turn directed at undermining the separate but equal regime’s one-size-fits-all logic of whiteness.

The notion of beauty, so prominent in consumer culture, broadened beyond the classic Western one to take into account the standards of other cultures. The focus on bodies, the subject of a biopolitical power, was not to whiten skin or to straighten hair but to affirm blackness, to bring about a radical transformation of subjectivity to participate in a project of liberation, or a cultural rebound. The black body is also “beautiful.” James Brown’s “Say it Loud, I’m Black and I’m Proud” and Curtis Mayfield’s “People Get Ready” epitomized the cultural rebound that was taking place in the black community. Eager to accentuate and promote what they considered to be a positive racial identity, the black power generation took many of their cues from Africa to express black pride and culture. Dashikis became popular attire for many blacks during the 1960s, and “Afro” and cornrowing were the “natural” looks in hairstyling. Influenced by the existentialist movement, Richard Wright had already predicted this cultural rebound in the 1940s, publishing works such as Native Son (1944), Black Boy (1945), and The Outsider (1953).

For the sake of capital, black slaves were commodities to be bought and sold in the marketplace. By the time of the Civil Rights Movement, the corporate world, in its effort to expand commodity exchange and consumption, was becoming less dismissive of the black body. Capital had become interested not only in the commodification of the labor component of the black body but also in its cultural component. Under a market-led pluralist regime, capital could increase profit through the production of difference marked by race, not just abstract labor.

Recommodification of the black body

These many dimensions of cultural diversity quickly became encoded into the commodity form. For the corporate world to continue to capture and control the black consumer market, it would have to change how it approached blacks as consumers and possessors of exchange value. The one size of whiteness no longer fit all: status became more fixated on market pluralism. The black community came to embrace this recommodification of the black body as a sign of black progress. This is quite unlike slavery, when black bodies also were commodities to be bought and sold. Yet, the function of both recommodifications is the same: to underpin capital, in different ways.

Following the ghetto explosion of 1965, large corporations established departments of community affairs to combat the growing anticorporate influence in the black community. By 1967, corporate interests, led by the Ford Foundation, were deeply involved in financing most major civil rights organizations, a form of recompense to black America for its support of capital. Carmichael thus did not have the last word on black power. To stay within market reach of the black community, corporate market makers re-encoded, as black capitalism, the movement
for “Black Power,” which had become the revolutionary cry of the black community. Weems (1998, 79) concluded that, “while many in the 1960s associated African Americans with the word ‘revolution,’ it was the corporate market marketers, the producers of media encoded texts, who had actually conducted a successful coup.” The producers of media-encoded text made it easier to penetrate the black community, simplifying its planning and programming for the sake of capital. As Earl Ofari (1970, 85) explained:

[With] the new alignment of the black politician, the black businessman (old black elite) has formed a new vested-interest power base in the black communities. Working hand-in-hand with corporate enterprise, the neo-black elites rapidly transformed the once expressive (and potentially revolutionary) cry of the black masses for black control of the black communities into capitalist control of the black communities. This only reaffirms the American ethic that everything is expendable, including principles. Everything from nail polish to Black Nationalism can be, and is, hustled as long as there is a prospect for private profit.

A call went out from within the black community for the desegregation also of the money market. Jesse Jackson urged black businessmen to move from civil rights to silver rights. Many black liberals and veterans of the Civil Rights Movement went along with this neoliberal shift. It was reinforced by the 1969 appearance of two magazines in the black community, Essence and Black Enterprise. With the aid of white corporate advertising, the new subjectivity that blacks were embracing recoded the commodification of blackness as a sign of black progress, not as a sign of whiteness. These magazines, along with their famed predecessor, Ebony, were visible symbols of black pride as well as of black America’s growing economic and popular cultural base. In its early issues, Essence carried a full-page self-promotional ad with a penetrating message that read: “Subscribe to Blackness.” In a small way, this ad bore testimony to black America’s defiant emergence from centuries of slavery, suppression, and segregation.

In the 1940s and 1950s the publisher of Ebony, John H. Johnson of Johnson Publishing Company, encouraged corporate America to reach out to the “black market” in the same way that it appealed to the white market; by 1964 he made the case for separate but equal racial retailing, almost in the same way that W.E.B. Du Bois had for a separate black economy over 30 years earlier. According to Johnson (1964, 119, 120):

The Negro market … is not a special market within the white market; it is, on the contrary, a general market defined, precisely, by its exclusion from the white market … Negro consumers tend to think, buy, and act in significantly different ways than white consumers … What Negro consumers want now is recognition of their humanity and an industry wide respect for the Negro image … [They] are growing not less self-conscious but more self-conscious … more critical … of media and products and services that deny them acceptable images … The basic factor here is the migrating mood of Negro consumers who have a new appreciation of themselves as non-whites … a newfound appreciation of blackness … It will be increasingly difficult to influence Negroes with ego ideals and ego models that are exclusively white.

For Johnson, the black body would no longer be fixated in whiteness. In the past, the black market was separate but not equal; now it had become separate but equal. Capital encoded in their commodities different identities and perspectives that would better capture and control the growing black consumer market, creating one of the greatest obstacles to a present-day black capitalist strategy.

The Congress of Racial Equality (CORE) organized a drive to convince 14 major corporations to make ads “a more realistic reflection of
the composition of the American population” by using the black body as a model in their advertisements. CORE argued that the positive use of blacks in advertisements would not adversely affect corporate sales. Corporations could still sell and realize surplus value (Advertising Age, September 9, 1963; December 9, 1968). Jesse Jackson’s Operation PUSH (People United to Save Humanity) and the SCLC’s (Southern Christian Leadership Conference) Operation Breadbasket also pressured corporate America, threatening a black consumer boycott, the most dreadful thing for capital to imagine. Martin Luther King also acknowledged the success of the consumer boycott strategy in Birmingham:

It was not the marching alone that brought about integration of public facilities in 1963. The downtown business establishments suffered for weeks under our almost unbelievably effective boycott. The significant percentage of their sales that vanished, the 98 percent of their Negro customers who stayed home, educated them forcefully to the dignity of the Negro as a consumer. (quoted in Washington 1986, 602; emphasis added)

The corporate world needed little persuasion, however, knowing full well that it had to sell to the black segment of the market. It could mean the difference between making and not making a profit.

This compulsion was evident, at least in terms of buying advertising space in the black print media. Full-page ads in very early issues of Essence and Black Enterprise show that cigarette manufacturers, regardless of their growing concerns about the connections being made between smoking and cancer, were among those who led the way in showing a keen interest in the black consumer. The number of full-page ads increased noticeably in Essence within less than two years of its initial publication. Essence carried merely five full-page ads in its January 1971 issue, but within few months this more than doubled to a dozen. Significantly, the magazine’s corporate client base expanded to include some of the major corporations of the time, including Sears, Levi’s, Pepsi, and producers of Martex, Kleenex, Tampax, Wohl Shoes, and Elan (perfume). Even TWA, Pan Am, Earhart (luggage), and Modern Living by Nancy Heine-man (household furniture) began to buy ads that targeted black consumers without offending them. The gains that Black Enterprise posted in terms of securing corporate advertisements were no less impressive. By 1974, within five years of its inception, its advertisers included such corporate heavyweights as Merrill Lynch, New York Life, Metropolitan Life, IBM, AT&T, Amoco, Exxon, Gulf, GM, Ford, Chrysler, Xerox, Delta, Rouse Company, Puma, Polaroid, GTE, and CBS. In one 1974 advertisement about increasing diversity, marketers at CBS noted: men and women of all races desired. This said it all about the emerging new corporate paradigm.

Through advertising in the black media, white corporate interest was able to solidify its control of the black consumer market, realizing more surplus value. Leading companies bought product advertisements in the black print media whose circulation was predominantly concentrated among black households. In the mid-1960s the late H. Taylor Fitzhugh pioneered the concept of black consumer marketing as an executive at Pepsi-Cola whose ads also appeared in Essence (Graves 1997). Having spent several million dollars in black-oriented advertising, Pepsi-Cola saw its annual profits increase from $157.6 million to $250 million between 1960 and 1964. In its fortieth anniversary issue, Ebony contained 162 pages of advertising worth $6.2 million, offering a vision of a bland consumer culture in multiple shades.
These magazines were not simply icons and outlets for black pride and consciousness. Profit-oriented enterprises themselves, they clearly understood the need for capital to sell and realize surplus value. It is, therefore, no surprise that the black print media served as a confluence not only of black power and pride but also of power and pride in the expansion of the commodity form for the sake of capital. As the popularity and circulation of these magazines grew among black consumers, they provided a genuine alternative and hence competition to white magazines of their genre.

With the black print media acting as a powerful and most visible conduit, black households had effectively become the target of a separate but equal marketing campaign on the part of corporate marketers and producers of media-encoded text. The black print media, represented by *Black Enterprise*, *Essence*, *Jet*, and a host of other newcomers, clearly served a purpose in flexing their muscles to demonstrate black consumer power. The specter of black power was finally hitting the world of marketing research because capital has to sell. Corporate marketers of women’s beauty and personal care products, in particular, sought to sell both “beauty” and “status” to black women, but this time in its black form (Weems 1998, 91). The advertising campaigns for cosmetics and beauty products were particularly racially targeted. The black community had evolved into a potent segment of America’s consumer market.

Black-owned ad agencies used their professional expertise to target black consumers for various corporations and their products. Several of the major black-owned advertising agencies were established in the 1970s, one as early as 1956 but most in the 1980s. In 1992 they spent $785 million just to reach the black market segment alone.

Target marketing, instead of one size fits all, had emerged as the corporate strategy aimed at a particular demographic and lifestyle category at a time when lifestyles were becoming increasingly the basis of social distinction. While market segments are delineated by geography, age, and sex, others contain racial and ethnic components, differentiating them from the mainstream market defined by its dominant white American client base. Marketing segmentation rejects the production of blacks as a replica of whiteness; this leads to blacks, as well as other racial and ethnic identities, becoming a market niche encoded by corporate advertising. It amounts to the commodification of blackness as a cultural or aesthetic category. No longer viewed as dark-skinned white people, blacks had their own identity or culture which could be marketed just like any other commodity. Consumer choice is thereby diverted into a mechanism for differentiating on the base of various dimensions of diversity, underwriting the production, distribution, and consumption of commodities.

This has the effect of expanding markets across multiple borders: the more differentiated or segmented the market, the greater the number of target markets from which to realize surplus value. Culture and media industries accelerate segmentation and identity formation through creating specific delimited consumer markets. Something is provided for all, so none can escape the invisible hand of the market: distinctions are emphasized and extended in these postmodern times.

The postmodern cultural turn produced a brave new world with a greater emphasis on choice and product differentiation; on marketing, packaging, and design; on the targeting of consumers by lifestyle, taste, and culture rather than just by categories of social class. Identities are reflected in lifestyles that are closely associated with brands and the products they label.
This enables consumers to express individual sovereignty through identities, constructed by acquiring and displaying goods that convey information about them and their position within a constellation of social networks. Rather than seeing identity as fixated on whiteness, the one-size-fits-all strategy, identities are now considered plural and dynamic, hybrid and fluid. Consumers carry, as Mort (1989, 169) noted, “a bewildering range of different and at times conflicting identities around with them in their heads at the same time.” A flexible regime of accumulation reconstitutes the production and distribution of commodities to exploit this hybridity and fluidity in society. By recommodifying the black body to enhance consumption, a racialized and more diversified underpinning of the corporation has taken hold.

A flexible regime of accumulation

A main flaw of postmodern analysis is that it tends to ignore the social relations of production or the class structure of capitalism. Market makers, encoders of text, and consumers as decoders of text are not free of structural constraints. The postmodern cultural turn suggests an open and dynamic egalitarian culture, with constant cultural flows and mixtures. Exploiting this openness and mixture in the constitution of production, and the diversification of consumption and distribution, requires a regime of accumulation more flexible than the Fordist regime.

Whereas through the first half of the twentieth century Fordism was producing conformity, by the 1970s a flexible specialization regime was offering greater authenticity, individuality, difference, and rebellion as a market strategy. Corporations could now maximize profits through the social production of difference, marked also by race. Featherstone (1994, 404) notes: “Since the 1960s there has been a more general informalization and elaboration of previously restricted codes of behavior.” If the trauma of World War II meant that the corporate world faced a society mentally wedded to mutual cooperation and togetherness, 20 years later the Civil Rights Movement and the Vietnam War set the stage for a more fragmented society which the corporation was able to exploit in different ways.

Left to its own devices, the corporate world never abandons a profitable regime. The separate but equal regime had been profitable, reproducing the racial hierarchy and making money in the process. It was basically a producer-driven commodity chain whose primary objective was the production and consumption of whiteness. Through the mass production of standardized goods, it counted on adequate consumer demand but had little need to listen to the market. This regime defined the market as massive and white, paying little or no attention to differences within it. Feedback from consumption to production did allow some changes in the market to translate to changes in productive engineering, but this communication circuit was restricted, slow, and lacked flexibility.

The mass market became increasingly saturated as a result of rapid rises in productivity after 1945. Consumer markets began to stagnate, posing a noticeable constraint to economic growth in the industrialized world. The corporate world expanded mass production and consumption, increasing the competition for the consumer dollar. Mass marketing became highly competitive. But additional advertising and promotion yielded diminishing returns. The market for consumer goods was increasingly governed by replacement demand turnover and obsolescence rather than by the opening up of new markets, which had been possible throughout the 1950s and early 1960s. Having reached its maximum potential, mass marketing intensified efforts to open up new markets. But the fixity and
standardization of the separate but equal regime began to pose obstacles to the further progression of capital. A more mobile and flexible regime had become necessary, which was not fixated on whiteness or on a particular identity or status relation.

The separate but equal regime lacked the flexibility to exploit the growing demand for more differentiated products or hybridity in the marketplace. Thus, capital had to engage in the social production of difference. Wendell Smith (1956, 7) had already warned of the growing danger of product competition with mass production and mass marketing. He noted that the theory of perfect competition, the basis of neoclassical economic theory, assumes homogeneity, that one size fits all, but diversity or difference had become the rule. He proposed exploiting “market segmentations,” arguing that, in place of a mass market, goods would “find their markets of maximum potential as a result of recognition of difference in the requirements of market segments.” Concluding that “their core markets have already been developed,” Smith (1956, 7) advised companies that “attention to smaller or fringe market segments, which may have small potentials individually but are of crucial importance in the aggregate,” could be profitable. Capital could maximize its profits not just by rendering labor abstract but also through the social production of difference.

Consumption has become symbolic of social distinctions. Mass-produced products lack status or lifestyle differences, the sign value that targets a particular socioeconomic group to buy. Decades before Baudrillard’s (1983) cultural analysis of the commodity sign, the automobile industry had already started to encode its products in market segments, manipulating the sign value of its commodities to communicate messages about the status or lifestyle of the consumer. Growth required selling cars whose appearance and features become obsolete on an annual basis. It was better to divide up the mass market according to distinctive needs and desires. Instituting model and color diversity, the automobile industry placed its products in different geographical and social contexts to denote a specific rather than a generic product. Production became a bundle of changing characteristics designed, produced, and marketed to appeal to different market segments of the population.

With geodemographic analysis, the corporation is able to predict variations in consumption based on a host of demographic factors such as race and where a person lives. Mass culture becomes micro: smaller fringe markets matter. Commodities are branded on the basis of where they are made. Corporations are increasingly emphasizing localized production and consumption.

The cultural and racial discontent of the 1960s and 1970s made universal or mass marketing increasingly impractical, weakening the strategy of one size fits all. In the corporate world, the fixed relations that characterized the separate but equal racialized corporate regime became increasingly impractical and illegitimate. Corporations become less receptive to the notion of social purification.

It is not coincidental that the newly constituted regime corresponded to the historical era of discontent in the 1960s and 1970s. If it was not for the Black Power movement, the student revolts, the women’s movement, and other antiracist struggles, the corporate world may have been content to maintain the one-size-fits-all corporate strategy, happy to avoid the trouble of having to reconstitute production, distribution, and consumption. Instead, in the new social and cultural formations, capital saw the opportunity to create more value and to expand commodity exchange and consumption by recoding and decoding race, bringing the strange, the other,
into the space of the ordinary. According to Cohen:

social and cultural groups such as African Americans, women, youth, and senior citizens began to assert themselves in a way that came to be called “identity politics,” where people’s affiliation with a particular community defined their cultural consciousness and motivated their collective political action. Marketers, who had they not been in search of ways to avoid saturating the mass market, might have despaired at the splintering of the mass in the 1960s, soon seized the new opportunities for selling it offered. (2003, 308; emphasis added)

Needing to sell to realize value, corporations seized on the opportunity offered by the racial discontent of the 1960s and 1970s to do just that.

Conclusion

Consumption is diverse, fluid, and varied rather than fixed and monolithic. The tension between race and corporate capital’s need to expand commodity exchange and consumption asserts itself in unexpected ways across time and space. Race, then and now, is a product of this tension. It cannot be seen merely as a reflection of racial ideologies. It must be situated within a framework that understands and critiques capitalism as an ideology that sustains and reproduces capital relations of production.

From the consumption of whiteness, the one-size-fits-all paradigm, the corporate world has shifted to a market-led pluralism that recognizes different consumer market segments. This postmodern paradigm shift reached new heights as black cultural elements, once ignored in the marketplace, were widely marketed by corporations. As a result, blackness is no longer represented as the opposite of whiteness but as a commodity to be consumed by all. The classical black sound of rap, the new rhythm of consciousness and resistance, and its hip-hop cultural form are commodified to be bought and sold on the world market. This regime of accumulation is less fixated on whiteness and conformity, on maintaining fixed status relations. The consumerism of postmodernity, held to be incommensurate with the fixed relations of modernity, accommodates differences to expand markets. Contemporary producers of encoded texts welcome the youth-led and hip-hop cultural revolution as an ally in their own struggle to manage and regulate society for the sake of capital. Rap and hip-hop are prime examples of the commodification of race for the sake of capital, having become extraordinarily commercially successful worldwide. Rap is globally produced and consumed more than ever, and in diverse markets. The underlying motivation for positing race and culture as the bases of collective identity was social, but not far behind was the neoliberal need to express race and culture as a commodity form to realize surplus value.

The black print media reproduced neoliberalism in the black community. Yet it often blames black and brown populations for their economic circumstances and for failing to fight against structural factors that keep them economically subordinate. The black corporate world has “earned the confidence of the white corporate hierarchy and the capitalist state by keeping alive the bogus illusion of a black capitalism” (Marable 2000, 158). Yet over four-fifths of all black-owned US firms are small businesses, struggling against massive odds to not fall back into the ranks of the proletariat. Further, the black community lacks an effective voice to articulate the growing black demand for economic opportunities.

More research is needed on the ways in which the corporation reworks cultural elements to expand commodity exchange and consumption,
and the implications thereof. Has this reworking of culture elements produced a culture war over meaning and lifestyle? Can we put a neoliberal face on this culture war? Many condone this reworking of culture because it sells. Hip-hop reproduces neoliberalism by emphasizing narratives that laud violence, crime, and drug use in the black community. To expand commodity exchange and consumption, the corporate world welcomes all, even those from the 'hood. It manages and regulates this life of diversity opportunistically to make money. Social networking sites such as MySpace and Facebook are invaluable in identifying niche markets of potential consumers for capital.

To what extent has this cultural manipulation of the consumer by capital contributed to the underdevelopment of the black community? The 1968 Report on Civil Disorder by the Kerner Commission documented the pent-up desire for consumer goods that television, movies, and advertising held up. While black consumer spending has increased, few dollars circulate in the black community. Black consumers are foot soldiers for corporate interest, siphoning away dollars from the black economy (Weems 2005, 253). For every $8 to $10 spent in the black community only $1 remains. An informal and predatory community, consisting of illegal practices such as drug trafficking and prostitution, develops as the poor person’s way to cut himself or herself in on the consumer society and to survive in a community that’s leaking a substantial amount of capital.

SEE ALSO: Brands and branding; Geodemographic profiling; Marxist geography; Postmodernity; Corporatization of race: an American case study; Race and racism; Whiteness

References


**Further reading**


Cosmography

Arild Holt-Jensen

University of Bergen, Norway

Cosmography literally means the description of the universe; today, this would automatically be associated with astronomy, which as a discipline has very little to do with geography as it is taught in schools and universities. On the other hand, geography literally means description of the world, understood as and limited to our planet Terra. However, geography is a scientific discipline and, as such, not confined to description; geographers explain phenomena on Earth. Geology would have been an appropriate disciplinary name, but that was adopted earlier on for another subject with explanatory ambitions for only some of the phenomena on the Earth. Disciplines change their content, methods, and investigation themes over time. Geography was used as a term at the Museum in Alexandria about 300 BCE, but there are only feeble links from then to the discipline of today.

Even though cosmography seems far from geographic interest today, it is highly relevant for the historic understanding of the discipline. One simple reason is, of course, that understanding of the cosmos, or universe, was far more limited, say 200 years ago, than it is today. Cosmographic studies focused on the Earth and the closest celestial bodies. A second reason was that geography was not an institutionalized discipline; writers and teachers that have been listed in this entry as geographers could write and teach about what they found most interesting, they were not confined by any defined disciplinary matrix or paradigm. Many of them should be regarded as cosmographers rather than geographers.

Eratosthenes (276–194 BCE), who was chief librarian at the Museum in Alexandria, succeeded in calculating the circumference of the Earth with remarkable precision. Of equal importance was his development of systems of coordinates for the world, that is, latitude and longitude, which he used to locate places and to measure distances, thus making it possible to draw the first passably accurate maps. This work was termed geography but it is quite clear that its basis was a new understanding of cosmos, particularly that the Earth was a globe. From this early start the cosmographic approach was integral to the geographic task of mapping.

Until the end of the nineteenth century, voyages of discovery and the mapping of formerly unknown lands were closely associated with geography. Wayne K. Davies (1972, 11), for instance, maintained that geography enjoyed its strongest relative position among the sciences during the so-called “golden age” of exploration from the fifteenth to the nineteenth centuries. This was not due to the academic status of the subject during this period but to the work of a number of people who were actively involved with the mapping and description of the new lands being discovered. To the extent that they were working scientifically they could best be described as cosmographers.

Bernhard Varenius (1622–1650), whose *Geographia Generalis* was published in Amsterdam in 1650, was one of the first scholars to present a scientific understanding of the nature of geography. But his book includes much material that today would be treated as mathematical geography or as astronomy. It is divided into three parts:
COSMOGRAPHY

1. the absolute or terrestrial section, which describes the shape and size of the Earth and the physical geography of continents, seas, and the atmosphere;
2. the relative or cosmic section, which treats the relations between the Earth and other heavenly bodies, especially the sun and its influence on world climate;
3. the comparative section, which discusses the location of different places in relation to each other and the principles of navigation.

Varenius intended that the Geographia Generalis should be followed by what he called, in his Foreword to the book, “special” geography, in which descriptions of particular places should be based upon:

1. “celestial conditions,” including climates and climatic zones;
2. terrestrial conditions with descriptions of relief, vegetation, and animal life;
3. human conditions, including trade, settlement, and forms of government in each country (Lange 1961).

It was Alexander von Humboldt (1769–1859) and Carl Ritter (1779–1859) who laid the scientific foundations for geography as a branch of knowledge. These two men had many views in common and were united in their criticism of the casual and unsystematic treatment of geographical data by their predecessors. Humboldt was the last of the great polymaths. Penck (1928, 31) regarded him as a cosmographer rather than a geographer. He mastered a number of disciplines and put all his energy (and fortune) into travel and research in order to understand the whole complex system of the universe. In his great work Kosmos, with its subtitle Sketch of a Physical Description of the World, he attempted to assemble all the contemporary knowledge of the material world. Kosmos was published (1845–1862) in five volumes at the end of his life, while the results from his long and important journeys in Latin America (1799–1804) had been published in 30 volumes in Paris (1805–1834).

Cosmography was termed in a more limited way by Schmithüsen (1976, 10), as he saw cosmos as the Earth and not the original definition as the universe. But he defined it as an umbrella discipline including not only geography and cartography but also natural sciences, such as biology, geology, and geophysics, and social sciences, such as anthropology, which only achieved their independent academic standing towards the end of the nineteenth century. Exploration, and all these other fields of cosmographic activity, were also regarded as being part of geography by the general public because they were carried out, to a large extent, under the auspices of the geographical societies, and they became, in political terms, quite important for the establishment of geography as a discipline in universities and high schools.

Geography developed as an academic discipline partly on the basis of a cosmographic philosophy that was developed to give coherence to the different activities of the geographical societies. Gradually theoretical studies made an increasing contribution to the advancement of a specific geographical methodology. The chief emphasis remained, however, on geography as a science of synthesis, a science linking humanity and environment and creating a bridge between the social and natural sciences. The cosmographic tradition is still at the roots of the contemporary discussions.

SEE ALSO: Ancient geography; Ethnophysiography; Geodesy; Geography and the study of human–environment relations; Historical geography
References


Further reading

Discourses of cosmopolitanism have recently surfaced in relation to the accelerated pace of globalization of capital flows, material production, cultural expressions, and consumption. On the one hand, cosmopolitanism is often cast as the humanist counterpart of globalization in line with visions of global democracy and world citizenship. On the other hand, given that cosmopolitanism’s normative orientation distinguishes it from the condition of globalization, more critical accounts make clear that, as the world becomes more globally connected, there are no automatic or immediately obvious pathways toward cosmopolitanism.

As an ideal, a desirable goal, and a dream, the notion of the cosmopolitan speaks of a spirit of openness as opposed to closure, and emphasizes the extension of bonds of inclusivity between people, societies, institutions, and organizations based on foregrounding the interests of humanity. Cosmopolitanism espouses a “universalism plus difference,” a culture of openness and acceptance, underpinned by the values of “inclusive heterogeneity, tolerance of difference and otherness, equitable (re)distribution of resources and privileges, recognition of others’ freedoms, (comm)unity in diversity, unqualified practice of fairness, kindness, and generosity” (Dharwadker 2001, 7). Cosmopolitans thus need to have the ability to function productively wherever they are, accepting of others by minimizing concerns over nationality, race, or class, and are connected to the Earth but not to a particular place (Robbins 1998). This inclusivity becomes an indication of a tolerance of all peoples as not similar to self, but rather as having a recognizable, expected, and accepted difference. Jazeel (2011, 89–93) goes one step further in advocating the “unlearning” of cosmopolitanism in order to be able to imagine a future of living together “not just with difference but … with radical, heretofore unimaginable, difference” within particular places which are “relationally intertwined, dynamic and turbulently moving on.”

Scholars have noted cosmopolitanism’s inherent politics and its elitism. Critics argue that discourses of cosmopolitanism merely articulate the ideology of the neoliberal managerial class (about living in a “global village”) while at the same time complicitly reinforcing the hegemony of this class. Robbins (1998, 4) decries its presumed detachment asemasculating, suggesting that

the cosmopolitan is held to be incapable of participating in the making of history, doomed to the mere aesthetic spectatorship that he or she is also held secretly to prefer, and that cosmopolitan identification with the human race serves as a thin, abstract, undesirable antithesis to a red-blooded, politically engaged nationalism.

Here, cosmopolitanism is reduced to a matter of “consumption” or “aesthetics” best embodied in the emergent culture of the transnational capitalist class whose globe-trotting lifestyles, connoisseur tastes, and disembedded social networks present a revolt against the nation-state. Elitist and exclusionary, such a version of cosmopolitanism is devoid of social or political transformative power. Harvey (2000, 557)
further points to the “banality” of a cosmopolitan vision in the absence of a recognition of geographical differences, and argues that a meaningful cosmopolitanism must go beyond some passive contemplation of global citizenship: “Cosmopolitanism bereft of geographical specificity remains abstracted and alienated reason … Geography uninspired by any cosmopolitan vision is either mere heterotopic description or a passive tool of power for dominating the weak.”

Cosmopolitanism must hence imply the presence of a geography, of places which are different from others; would-be cosmopolitans must thus learn to navigate a nonhomogeneous landscape. In this light, Jazeel (2011, 92) cautions against Eurocentric visions of universal humanity that may lurk beneath cosmopolitan thought, and advocates “opening out some of the concept’s more restrictive geographical imagination” to thinking of “place as inherently relational, … constituted by the coming together of narratives, trajectories and border crossings from elsewhere.”

Expectedly, scholarly geographical attention to rooting cosmopolitanism in place began with cities. In contrast to discourses on multiculturalism, integration, and social cohesion that usually frame discussions of diverse urban societies in terms of fixed ethnic identities, cosmopolitan logics have been invoked as a means to “avoid pitfalls of essentialism or some kind of zero-sum, all-or-nothing understanding of identity issues within a nation-state framework” (Vertovec and Cohen 2002, 3). Cosmopolitanism has been inserted back into discussions on urban diversity and emergent civil spaces, particularly in the context of the rising presence of transnational migrants and the increasing proliferation of contact zones that renders the contemporary globalizing city as a meeting place with the other. For example, Filipina domestic workers’ subversive “occupation” of the Central District of Hong Kong for weekend gatherings alerts us to the ways in which a city’s public space can be symbolically transformed into sites where alternative meanings about one’s identity are rewritten and reinscribed onto the city’s landscape. In a variety of urban contexts, studies of social interaction processes among residents make clear that demographic assortment alone is no indicator of cosmopolitan hospitality, and that geographies of encounter in the city are structured by sociospatial inequalities and insecurities.

Other scholars have begun looking for cosmopolitan beginnings in unlikely places: instead of concentrating our focus on the public face of the city as the quintessential reflection of cosmopolitanism’s topography, as is often the case in projects of urban boosterism, bringing into view the inner workings of more privatized, vernacular places characterized by relations of care and intimacy may help us in unlearning the normalized contours of cosmopolitanism. For example, Datta (2011, 3) proposes moving “beyond the public sphere into more private and affective spaces” of the city to search for a fragile “cosmopolitan neighbourliness” amid a Delhi squatter settlement in the context of communal violence.

Notwithstanding suspicions that cosmopolitan ideals are all too easily pressed into service to ideologically endorse neoliberal capitalism, geographers attentive to the difference that space makes have also argued that to ascribe the values of cosmopolitanism to only those who have the benefit of dwelling in a “global” space is tantamount to denying those living outside it a chance to express the same kind of human relationality. In this light, geographical scholarship has turned to notions of actually existing cosmopolitanism or cosmopolitanism from below to reflect on dimensions of social life that is grounded in more open, unfixed, provisional, and ethically conscious notions of the other. For example, Kothari’s
COSMOPOLITANISM

(2008) ethnographic work with Bangladeshi and Senegalese street traders in Barcelona associates non-elite cosmopolitanism with ordinary people such as street peddlers who are artful political subjects in their use of translocal social networks to acquire resources and respect, while other scholars have focused on non-elite forms of cosmopolitan modernity articulated among slum dwellers. By exploring changes in consumption patterns, possibilities for cultural learning, the development of new sensibilities, and the negotiation of cultural differences, scholars identify emancipatory hope for migrant domestic workers as potential working-class cosmopolitans despite the retrogressive contours of transnational domestic work. Through treating cross-cultural exchanges unfolding in different kinds of spaces from the flexibly open-minded to the severely constrained, the true breadth of cosmopolitanism can be more fully appreciated.

SEE ALSO: Difference; Global cities; Multiculturalism

References


The creative cities concept emerged at the start of the twenty-first century as a new way to imagine urban planning and economic development. Rather than envision cities as centers of manufacturing and corporate offices, as had been the model in the twentieth century, advocates of creative cities emphasized human capital and suggested that, to thrive, urban economies must focus on leisure, entertainment, tourism, innovation, entrepreneurship, the arts, and, in general, human creativity. The two major articulations of creative cities were Landry’s (2000) notion of the creative city and Florida’s (2002; 2005; 2008) concept of creative class. Landry (2000, xiii, xv) maintained that “Cities have one crucial resource – their people. … [and] At the heart of creativity are creative people and organizations who have particular attributes: when these come together in one area they establish a creative milieu.” Florida (2005, 35) argued that cities are competing globally to attract such creative individuals: “The Creative Class is moving away from traditional corporate communities,” and is moving to “Creative Centers … [which have] high concentrations of creative economic outcomes, in the form of innovation and high-tech industry growth.”

Advocates proposed that urban policies should be designed to generate creative cities and argued that a vibrant cultural scene is essential to revitalizing both neighborhoods and economic development. To encourage the emergence of a creative city, for example, urban planners could revise existing zoning laws to allow live–work studios in former manufacturing buildings; create multiuse buildings that can aid new business start-ups; provide nontraditional office space for technology entrepreneurs; redesign transportation to reduce car use and suburban commuting and, instead, encourage the use of public transit, bicycling, and walking; subsidize city-wide broadband and wireless Internet access; give tax breaks to encourage environmentally sustainable urban design and buildings; and host myriad festivals, concerts, and other events. Typically cited examples of creative cities in the United States include San Francisco, Boulder, and Austin, where investment in the annual South by Southwest (SXSW) celebration of music, arts, and technology has branded the city a creative urban center. Calgary, Toronto, Amsterdam, Nagoya, Bilbao, and Sydney are also often cited as creative cities. The rapid acceptance of the concept was exemplified in 2004 when UNESCO established its Creative Cities Network (UCCN) in which member cities work “together towards a common mission for cultural diversity and sustainable urban development ... ‘Creative hubs’ that promote socio-economic and cultural development in both the developed and the developing world through creative industries” (UNESCO 2009–2014).

As making cities more creative has quickly become an unquestioned goal of international urban policy, critics maintain that they produce exclusive enclaves for urban elites. Peck (2005, 742, 766) argues that “the creative-cities script has found, constituted and enrolled a widened civic audience for projects of new-age urban economic development ... ‘Creative hubs’ that promote socio-economic and cultural development in both the developed and the developing world through creative industries” (UNESCO 2009–2014).
privileged actors,” in which “elite-focused creativity strategies leave only supporting roles for the two-thirds of the population languishing in the working and service classes.” McLean (2014, 673), in turn, proposes that creative city policy “initiatives put a positive spin on precarious creative industry work while simultaneously gentrifying neighborhoods, a dynamic that displaces low income residents — especially seniors, people living with disabilities, and the racialized poor.” The use of public monies to fabricate creative cities typically means reducing spending elsewhere and, therefore, creative cities may exacerbate urban inequalities.

SEE ALSO: Creative class; Creative class culture; Gentrification; Global cities; Knowledge-based economy; Local development; Regional development policies; Urban elites; Urban geography; Urban politics; World cities

References


Further reading


Creative class culture

Ulrike Gerhard
Heidelberg University, Germany

Creative class culture refers to the circle of understandings, beliefs, and values of a distinctive group of urban dwellers, the so-called creative class, described by Richard Florida (2002). It is characterized by a modern urban lifestyle, expressed by the usage of and attraction to sidewalk cafes, bistros, small shops, and galleries. Thus, creative class culture can be assigned to gentrified inner-city neighborhoods with an older housing stock that has been redeveloped by a more affluent group of inhabitants or developers, displacing former residents and transforming the urban landscape and infrastructure. While creative class culture is strongly related to cultural capital assets, its specific consumption patterns show a slight bias to street level culture that disfavors customized mass products but instead looks for more authentic items and places (e.g., small, independent cafes instead of Starbucks coffee shops, whole food stores instead of big supermarket chains). Often, these preferences adhere to a higher socioeconomic status with a certain willingness to pay higher prices for more quality or individuality. Life and leisure plays an important part in creative class culture: “live the life” (Florida 2002, 166) is the attitude that includes intense, high-quality, multidimensional experiences and favors active recreation over passive pastimes (e.g., running or rock-climbing rather than watching television).

Despite these characteristics, creative class culture is a rather broad description of an urban lifestyle that can be found throughout the globalized world, characterizing a revaluation of the urban as a progressive ethos (e.g., democratic, participatory, modern) that is not new to urban theory (e.g., the Greek Polis, Louis Wirth’s Urbanism as a Way of Life). Cities function as nodes for the global space of flows that also entail flows of people. More recently, some of the most observed socioeconomic groups have been described under different headings: e.g., Yuppies (Young Urban Professionals), Bobos (bourgeois and bohemian), or DINKS (Double Income No Kids). The creative class concept, though, captures best a formal institutionalized class development infecting many cities across the world: the struggle of cities to compete for investments, affluent habitants, and global attention. Creative class culture thus emphasizes the economic value of a certain group of people, by neglecting that (i) not all members of that group are necessarily affluent, and (ii) a vast majority of inhabitants do not qualify as being “creative.” Thus, creative class culture can spur urban uneven development. Furthermore, it is a very dynamic concept originally characterized as individualistic, progressive, and assigned to childless households, one can now detect a maturing within the lifecycle – more established, family-oriented households, often with working mothers, increasingly adhere to that group as well. In urban geography and social theory, the term is therefore very useful to analyze the social and cultural impact of urban developments in the transforming knowledge society.
CREATIVE CLASS CULTURE

SEE ALSO: Creative class; Cultural capital; Gentrification; Urban redevelopment; Urban uneven development

Reference


Further reading

The last several decades have witnessed a major shift in capitalism. This transformation is marked by the rise of a “creative class,” a term coined by Florida (2002) to refer to workers who generate economic value through their creativity. This class encompasses knowledge, professional, and technical workers, as well as workers in the arts and cultural industries. These workers draw on intellectual rather than physical capacities to create meaningful new forms. This entry discusses the rise of a creative class and what it means for the nature of work. Criticisms of the notion of a creative class will also be reviewed.

According to Florida (2002), the transition to a knowledge-based or creative economy is tied to the crisis in an earlier phase of capitalist development known as Fordism. Fordism connotes an economic system that came into being in the post-war period. This system was based on the mass production of standardized commodities. A key problem with Fordism is that it afforded minimal possibilities for intuition, initiative, or creativity.

Fordism fell into crisis in the 1970s, when advanced industrialized countries faced increased competition from overseas producers. The costs of this system also became apparent in the low quality levels characteristic of mass production systems. These challenges gave way to a new competitive strategy based on the production of smaller batches of niche market products differentiated on the basis of their quality, symbolic value, or design. This transition has fueled the demand for new skill sets, and heightened the importance of creativity. Florida argues that workers in a variety of jobs, including those in manufacturing and services, are increasingly valued for their creative capacities.

This is especially so in creative class occupations. Florida (2002, 68) defines class as a social grouping or identity based on economic function. This grouping gives rise to a unique set of identities, desires, and consumption patterns. The creative class can be divided into two segments. The first segment is a super-creative core composed of scientists, engineers, professors, artists, writers, designers, architects, and entertainers. These workers engage fully in creative endeavors, producing new products, theories, artworks, or designs. They not only solve problems but also identify new challenges and needs. The second component of the creative class is made up of creative professionals working in financial services, law, health, and business management. According to Florida, these employees draw on complex bodies of knowledge and use their judgment to solve problems. These professions necessitate a high level of education. Workers may occasionally develop new products or methods but it is not central to their job (Florida 2002, 69).

The rise of a creative class is also related to new patterns of work. It is associated with greater opportunities for autonomy and initiative. Florida equates it with open concept workspaces, flexible hours, and a relaxed dress code. Creative class workers switch jobs frequently, creating horizontal labor markets. They value individuality, diversity, and meritocracy.
CREATIVE CLASS

Florida distinguishes the creative class from the working and service classes. The working class consists of those who work in production, transportation, repair, and construction, while the service class encompasses lower-wage occupations in health care, food service, office work, and caring. The creative class makes up approximately 30% of the workforce and is growing rapidly; the service class constitutes 43% of the workforce and is also expanding; and the working class accounts for roughly 25% of all employment but is shrinking (Florida 2002, 74). Florida maps this new class structure onto US cities, arguing that some cities have a higher concentration of creative class occupations (i.e., Washington, Boston, Austin, and San Francisco), and that these centers are more likely to experience growth. By comparison, major tourist centers such as Las Vegas and Miami have high concentrations of the service class. Other cities such as Grand Rapids, Memphis, and Buffalo are confined to being working-class enclaves.

While different class concentrations characterize different cities, Florida (2002) sees a relationship between the rise of the creative class and the service class. The creative class works long hours and has high incomes. As a result, they have a strong demand for restaurants, as well as for cleaning and child-care services. This fuels the growth of low-wage service employment, leading to polarization in some cities.

There are many parallels between Florida’s theory and other accounts of economic transformation. In the 1960s Bell (1973) described a major transition to a postindustrial economy, a move which he equated with the demise of the working class. A variety of more recent theories also chart a shift in advanced industrialized economies. Scott (2007, 1466), for example, cites the rise of a “cognitive-cultural economy” which emphasizes intellectual and affective labor. He includes high-technology industries, the service sector, neoartisanal manufacturing and cultural industries in this economy.

Autonomist Marxists such as Lazzarato also cite the rise of what they call “immaterial labor.” Immaterial labor refers to labor that produces the informational and cultural content of a good (Lazzarato 1996). It includes manual labor that has an increasing intellectual component, but also work that involves setting fashion trends, aesthetics, and consumer tastes. While there are many similarities between Florida’s notion of a creative class and immaterial labor, there are important differences. For Lazzarato, the rise of immaterial labor is accompanied by new techniques for managing communication and creativity, and for creating and controlling worker subjectivities. This new system threatens to be even more totalitarian than the old Fordist system because capitalists increasingly seek to incorporate the worker’s personality within the production of value (Lazzarato 1996). Immaterial labor is also highly precarious. Workers are often self-employed, working freelance or on contract. Work extends into leisure time, eroding the distinction between life and work. While Lazzarato does acknowledge some possibilities for cross-class mobilizing around precarious work, a much more negative picture is developed of the rise of creative/immaterial labor (see also Gill and Pratt 2008; Ross 2008).

There have been other criticisms of the notion of a creative class. Markusen (2006), for example, argues that the creative class represents a fuzzy concept – a category that unites disparate things under one umbrella while separating groups that may be similar. The notion of the creative class is based on occupational groupings, but some of these groups include uncreative as well as creative jobs. Business and financial occupations, for example, include claims adjusters, while
the category of engineer also includes drafting technicians. Some of these jobs may be creative but so too could many jobs designated uncreative by Florida – such as tailors. Markusen (2006, 1924) argues that it is problematic to label some workers as creative and others as not (see also Morgan and Ren 2012).

A further problem is that Florida tends to conflate creativity with high levels of education where there is no direct link between the two (Markusen 2006). The assumption of a common class position also disguises differences in training, career trajectory, and economic capital between different creative occupations. These differences have implications for the locational preferences and political values of creative workers. This suggests that Florida is not talking about class as defined by a common class interest, outlook, or behavior (Markusen 2006).

Others question the politics of Florida’s theory, which positions the creative class as the source of economic growth in a city region. As Peck (2005, 757) points out, this relegates the working and service classes to a passive role in economic transformation, and leads to policies focused on the interests of the creative class. This is problematic because this elite class is already well positioned to succeed in the new economy (Shearmur 2006–2007, 8).

Florida (2002) argues that, to attract the creative class, cities need to enhance the quality of place characteristics of an urban region, making investments in bike paths, streetscapes, and bohemian neighborhoods. This agenda provides funding for the arts, and for events that promote values associated with creativity, such as openness, diversity, and tolerance. Florida contends that attracting the creative class to an urban region will benefit everyone. However, as noted earlier, there is evidence that the presence of a creative class is associated with increased inequality (Donegan and Lowe 2008).

Cities around the world from Amsterdam to Singapore have been reorienting their economic development agendas in accordance with Richard Florida’s creative class thesis. Investments have been made in creative industries and in lifestyle amenities designed to attract the creative class. Creativity policies fit nicely with existing neoliberal agendas (Peck 2012). These policies support the desire of cities to attract foreign talent and become “global cities.” In Singapore, for example, creative city agendas are associated with a liberalized approach to homosexuality. However, the rhetoric of tolerance toward gays and lesbians stops short of genuine enfranchisement (Oswin 2012, 1636). Similarly, while talented “creative” workers are encouraged to join the “national family” and to bring their families to Singapore, other foreign workers, performing more “unskilled” work, have no rights to citizenship. Like gays and lesbians, these workers continue to be positioned as “out of place” in the creative/global city (Oswin 2012). This foregrounds the way in which a creative class politics is associated with a narrow, hierarchical, and polarizing vision of the city.

Richard Florida’s notion of the creative class highlights significant changes in occupational structure in advanced capitalism. It describes the rise of new forms of work, which rely on the intellectual and creative capacities of workers. However, critics suggest that there are many problems with the way he characterizes these changes and, in particular, with the notion that creativity is located within a specific class. This leads to the privileging of particular groups of workers in the new economy, which has contributed to increased inequality.

**SEE ALSO:** Creative class culture; Knowledge-based economy
References


Further reading

Crime

Robert Haining
University of Cambridge, UK

Crime is any action that constitutes an offense as defined by the criminal law of the country, and that is punishable by law. A crime involves an offense, one or more offenders, and one or more victims (who may be individuals, or may be communities). To study crime is to study one facet of societal risk (of becoming an offender as well as being a victim) and how that risk might be moderated through different forms of individual and social action. The study of crime overlaps many areas of the social sciences with different disciplines tending to focus on different questions, emphasizing different sets of factors (e.g., social-institutional structural, as opposed to environmental-situational, explanations). Geography has a particular interest in understanding why offenses take place where they do, the sorts of neighborhoods that produce, or are perceived to produce, disproportionate numbers of offenders, and geographies of victimization. Geography, in common with other disciplines, is also interested in the policing response, and since police forces are often organized on a territorial basis their response may involve geographical targeting, which is of particular interest. These questions of crime and policing are examined at many different scales as well as involving national and international (cross-cultural) comparisons. And geography, again in common with other social sciences, also has an interest in the wider consequences of crime. The study of the geography of crime and criminality, as well as contemporary responses to control crime and criminality, can be traced back to the formation of modern industrialized and urbanized societies in the nineteenth century. This field of study has been variously referred to as the geography of crime by geographers, environmental criminology, and, most recently, sociospatial criminology by criminologists (Bottoms 2012).

A common thread in these different disciplinary endeavors is to understand the role of place and space in explaining these geographies, where place concerns the attributes of specified areas at defined scales (for example, blocks and street segments, neighborhoods, communities, cities, countries), and space refers to the spatial relationships between places (concerned with such properties as distance, nearness, and proximity). Understanding may focus on the here and now of an observed pattern for some area and for some defined period of time (perhaps considered typical or representative), or it may focus on repetitive patterns for an area over short time periods (daily, weekly, seasonally) or on longer periods of time in relation to crime pattern trends. The last of these (for example, the seemingly inexorable rise in acquisitive (e.g., theft and burglary) and violent crime post-World War II in increasingly affluent Western societies) have at different times given fresh impetus to the field, challenging researchers and requiring policymakers, urban planners, and the police to develop new responses. Crime profiles shift over time as do public attitudes and people’s fears about crime (not necessarily well correlated with recorded crime levels) and these trigger shifts in research emphasis. These longer term shifts in crime, criminality, and public attitudes and explanations for them, usually involving aspects
of societal and technological change, have given rise to contrasting political interpretations and even paradigm shifts in the field of criminology as, for example, in the United States and Western Europe in the 1970s.

Viewing criminal behavior as a rational choice rather than a purely pathological act and placing rational choice theory (RCT) at the meta-theoretical heart of Anglo-American criminology has had a profound impact on research into crime since the 1970s, providing an important framework for organizing ideas about who commits crime and where. The offender is viewed as the self-interested individual who seeks rewards and who balances those rewards against the risks that must be faced. RCT is perhaps most relevant in the case of acquisitive crime where motivated offenders may seize the opportunity for personal financial gain in those circumstances where they perceive the balance of risk to reward to be in their favor. In the case of expressive crime (e.g., some forms of homicide, sexual assault) “reward” may be couched in terms of pleasure or release from pain. In both cases risk is defined in terms of the probability of being caught and punished, and the scale of that punishment. RCT is criticized for overestimating its role in criminal activity as well as overestimating human powers to access and then process complex information, so that “full rationality” is usually substituted with the concept of “bounded rationality”; however, the inclusion of decision-making theory into predatory criminal behavior provides an essential ingredient for an important theory in criminology – routine activity theory (RAT). Routine activity is defined as any recurrent or prevalent activity that provides for basic population and individual needs and includes, for example, activity associated with work, obtaining food, shelter, schooling – anything in fact that is part of everyday life.

RAT was empirically grounded in the rising crime rate (in terms of both violent and property crime) in the United States between 1947 and 1974 (Cohen and Felson 1979). These post–World War II years were characterized by rising affluence when it might have been expected that crime rates would fall, or at least not rise, as more people were able to satisfy their material wants. In fact, crime rates were to go on rising both in the United States and in many parts of Europe into the 1990s. RAT argued that predatory crimes, which involve direct contact between the victim and the offender, require the convergence in space and time of the likely or motivated offender and a suitable target (the victim) in the absence of guardianship – someone (e.g., security guard, police patrol) or something (e.g., locks and fences) – that would deter the act by significantly increasing the motivated offender’s risk of being caught or by simply acting as a discouragement. Cohen and Felson (1979) argued that even if only one of these elements was missing, a crime would not occur. It was RAT’s contention that changes in routine activity could influence crime rates by affecting this convergence and that it was not necessary to speculate about a rising tide of individuals willing to commit crimes. They argued that in the United States since 1945 there had been a massive shift in routine activities away from the home associated with more people enjoying a better life and that these shifts had increased the likelihood of criminogenic interactions.

RAT combines the idea of the rational, motivated offender, drawing on rational choice theory, while also recognizing the importance of the environment in the study of human relations in space and time drawing on ideas from human (or social) ecology. While opportunity, defined in terms of target visibility, attractiveness, and accessibility, is a crucial element in RAT, it also argues that this cannot be abstracted from
the lives of ordinary people and how they interact with the social and built environments. Human ecology, the study of the relationship between humans and their environment, had underpinned the work of the Chicago School of Sociology in the 1920s and 1930s. Early forms of social disorganization theory (SDT) argued that neighborhood delinquent behavior was not just a product of individual and family characteristics such as age and race, but was also influenced by neighborhood characteristics such as high levels of unemployment, population turnover, and single-parent households. But in RAT, the perspective provided by human ecology was extended to embrace the everyday lives and environments both of offenders and of their victims.

Routine activities provide offenders with suitable targets (opportunities). Daily routine activities, whether associated with work or with leisure, determine the location of property and people and their vulnerability to crime. But RAT embraces other aspects of “everyday routine.” Routine production activities affect what consumables are available and how easy they are to steal since goods of high value but low weight and small size make for ideal targets. While community organization may provide the circumstances under which crime thrives, technological advance is also important both through its creation of desirable objects to acquire and through its development of security and surveillance devices that determine how well people can safeguard themselves.

There is another important empirical observation. The geographical distribution of offenses tends to concentrate in certain places. Crime pattern theory (CPT) extends the underlying ideas behind RAT to suggest why this might be so (Brantingham and Brantingham 1993). CPT argues that offenders will selectively choose areas to target that are cognitively known to them because they lie within their “activity space” and will not venture into areas that are unknown to them – and areas are familiar or not as a consequence of the offenders’ own routine activities. Although certain areas in a city may offer very attractive targets (houses in wealthy neighborhoods might reasonably be expected to contain many high value goods), if would-be offenders are unfamiliar with these areas they might feel unsure about the level of risk – of being seen and caught – they are facing. This argument is supported by the empirical observation that neighborhood crime rates are often highest and most persistent in poorer neighborhoods with a large number of resident offenders. Offenders target poorer areas close to where they live because of familiarity, which enables them to make a more informed assessment of the balance of risk and reward. A further element is often added to this calculation: the principle of least effort. When adapted to offender behavior this principle argues that, other things being equal, an offender will not expend any more effort than is necessary to achieve the desired outcome, and since travel involves effort, if offenders believe they can meet their needs traveling a shorter rather than a longer distance then that is what they will do. This combination of elements provides an explanation of why better-off neighborhoods close to neighborhoods with many offenders (typically poorer neighborhoods) tend to experience higher crime rates than those of similar neighborhoods that are not close to high offender areas.

Crime pattern theory helps understand long and short duration crime hotspots. The study of crime hotspots has been advanced by police geocoding of crime sites and the use of modern computer mapping systems, together with processing algorithms to detect and highlight spatial concentrations of cases. When crime data are analyzed at fine spatial scales it has been observed that crime incidents tend to concentrate in a
small number of street segments, block faces, and other such “micro-places,” and often in proximity to bars, schools, and transit stops. This level of concentration tends to be stable over extended periods of time; moreover, there is also evidence of a high degree of stability in terms of which specific “micro-places” in any area are responsible for this concentration (Sherman, Gartin, and Buerger 1989).

Two further theories support current understandings of crime geographies: social disorganization theory (SDT) and situational action theory (SAT). SDT is associated with the recognition that there is a neighborhood effect in addition to an individual and family effect in generating delinquency. At the heart of SDT is the idea that poor neighborhoods that lack close social relationships among residents produce disproportionate numbers of young male offenders, a social group responsible for a large proportion of crime. There is strong overlap here with the concept of social capital, that close networks among residents bring benefits to individuals and groups, and that the absence of such networks creates various social problems. More recently, it has been argued that the central focus should be not on the lack of social relationships (middle-class neighborhoods often lack these too) but rather on the inability of certain, usually poor, neighborhoods to achieve what the community wishes for and that trust is absent in community relationships. This is referred to as collective efficacy, which acts as a mediating variable between neighborhood social structure and levels of crime and disorder (Bottoms 2012).

SAT also addresses the role of the social environment in crime causation (Wikström et al. 2010). Social disorganization theory tends to overestimate the neighborhood of residence of the offender rather than recognizing that people move around and are exposed to many social-environmental influences; routine activity theory is criticized for not explaining how convergence in space and time results in a crime taking place thus focusing only on the necessary, not the sufficient, conditions for crime to occur. SAT argues that since crimes are breaches of moral rules of conduct, to understand the occurrence of crime we need to explain why people break moral rules. The theory argues that the convergence in space and time between a person’s criminal propensity (their morality and ability to exercise self-control) and the exposure setting (the moral rules of the environment in which a person spends time and their enforcement) initiates a perception-choice process whose outcome may be an act of crime. This approach requires us to understand people’s activity fields: the various environmental settings people are exposed to in any given time period.

As this review has demonstrated, a rich literature has evolved in criminology examining the role of place and space in understanding offense and offender geographies. The links between criminology and sociology are deep as evidenced by the contents of the *Oxford Handbook of Criminology* (see in particular the contribution by Bottoms 2012). On the other hand, the links between this academic literature and geographical research were perhaps closest in the 1980s with the work of David Evans and David Herbert (1989). Be that as it may, the work of sociospatial criminologists in specifying the criminogenic attributes of neighborhoods, their deployment of rational choice concepts and routine activity theory to account for crime events and where they happen, and the place-specific signals that lead to (or do not provide sufficient discouragement to) criminal acts have provided fertile ground for more recent geographical work into women’s fear of crime and the victimization of racial minorities. The
attention of social geographers has been largely
drawn toward studies of different populations,
declared by race or gender or both, who are
at high risk either of becoming victims or of
becoming offenders. In the United States there
has been a long tradition of research into the
relationship between crime and race rooted
not only in the violence historically directed at
African Americans but also in the disproporti-
ately large number of African Americans (but
also native Americans and people of Hispanic
origin) held in incarceration relative to their
presence in the population. Although much of
this work is sociological in disciplinary origin
the contribution of the geographer Harold
Rose is notable. Rose’s study of lethal violence,
especially among young black males, is illus-
trative of this work (Rose and McClain 1990).
Feminist geographical research on crime, again
part of a wider thrust in sociology, has focused
on women’s fear of crime, linked to the effects
of poor urban design, as well as to women’s
experiences of crime and the underreporting of
some offenses against women, notably domestic
violence (Pain 2000). The latter research exposes
the contradiction for some women of the home
as a place of supposed safety and security but in
reality a place of abuse.

To date much of the work by geographers has
linked more closely with the concept of social
justice rather than risk. Social justice is con-
cerned with issues of what is or is not fair about
society and the extent to which everybody in
society enjoys an equal chance to succeed in life
and to realize their potential. When life chances
are not distributed equally (in this case, certain
groups are subject to higher rates of victimization
or are at greater risk of becoming criminals) a
“just” society is one where its members should
feel an obligation to address the inequality. The
concept of social justice has its critics: it lacks
objectivity and can quickly become politicized
around issues associated with individual freedom
and the role of the individual; however, in the
case of crime the term social justice brings to
the forefront ideas about what, at the societal
level, is the right thing to do to minimize
the risk of becoming a victim of crime or an
offender. In that sense the fruitful link between
these two modes of human understanding is in
society’s response to crime and how risk can be
reduced in ways that do not lead to, for example,
social exclusion or the stigmatizing of particular
subgroups such as certain racial and ethnic
groups, or young males. With risk reduction in
mind, we now turn from theories addressing
the risk of crime and the uneven geography of
that risk to examining social responses aimed at
mediating those risks through urban design and
policing.

At the center of many debates about the role of
urban design are two rather contrasting visions
of how to effect crime prevention through
environmental design. In both rational choice
theory and routine activity theory a central
feature in the decision to commit an offense is
the offender’s perception of the risk of being
catched. Natural surveillance increases would-be
offenders’ perceptions that any criminal act
will be observed, thereby increasing the risk of
being caught. Architectural design that promotes
natural surveillance means urban design that
improves visibility (through the placement of
physical objects and good lighting) and promotes
social interaction (by encouraging a diversity
of uses as well as through the creation of sites
where people will naturally congregate such as
parks). This vision of crime prevention is
described in the urban planner Jane Jacob’s work
on Greenwich Village, New York. An alternative
vision is presented in the work of architect and
planner, Oscar Newman. He argued that while
natural surveillance is important in curbing
crime it is not sufficient and that emphasis also
needs to be placed on social control (the idea that residents of an area should have a strong sense of ownership of and hence responsibility for their neighborhood) and security through such things as limited access (so that escape routes for would-be offenders are also limited). Just as Jacob’s ideas on urban design stressed influencing the offender’s perception of risk, so Newman’s ideas stressed imparting the perception of neighborhood security and solidarity in influencing offender behavior. Perhaps the most extreme expression of Newman’s vision is the gated community, but there are other examples including “alleygating,” which involves cutting off passageways between or through clusters of homes. This literature has a particular resonance for feminist and prefeminist research, which has emphasized how post-1945 urban design has led to architectural forms that have fueled women’s fear of crime in public spaces (Ceccato 2012).

Policing is the second area through which society seeks to reduce the risk of crime. A feature in developed Western societies has been the decline since the 1990s in levels of high volume crimes such as burglary and car crime and also some forms of violent crime (e.g., homicide) thereby reversing the post-World War II trend in such offenses. If such crimes are currently declining, however, others are increasing. Terrorism, people trafficking and human slavery, money laundering, and drug dealing, crimes often linked to globalization, are increasing. Hate crimes, particularly racially motivated hate crimes, are increasing (and significantly underreported) as are crimes associated with local disorder and linked to social fragmentation. Some forms of crime associated with the internet (e.g., fraud, sexual grooming) also appear to be on the increase. The range of these crimes and police responses are too large to overview here; hence, only two aspects will be considered: the role of crime mapping to support police activity and the implementation of neighborhood-focused reassurance policing. Geographers have provided direct inputs into crime hot spot detection and modeling through the use of local spatial statistics and GIS as well as into assessments of targeted programs (Craglia, Haining, and Wiles 2000; Li et al. 2013).

Crime mapping has enabled many police forces to add value to their routinely recorded crime and service-call data by helping to identify and empirically quantify high-risk locations (crime hotspots) where additional resources, such as more police patrols or situational crime prevention (SCP) measures, might be targeted. SCP, theoretically grounded in rational choice theory and routine activity theory, is concerned with the identification and implementation of strategies that seek to reduce the opportunities for offenders to commit crimes and to increase the risks they face if they do. Typical SCP strategies include: target hardening (putting in place security measures that increase the effort would-be criminals need to expend); improving surveillance; and deflecting potential offenders. They focus on improving the safety of buildings and streets by making physical changes, by involving individuals and communities in preventing crime, and through inter-agency collaboration. An interesting extension of the crime mapping agenda is geographic profiling, the geographical equivalent of psychological profiling, which is underpinned theoretically by crime pattern theory. But the reductions that have taken place in some crimes have not necessarily been matched by the public’s sense of safety. One response has been the introduction of reassurance policing, which is based on a detailed understanding of places and their crime and disorder problems and what needs to be done to address residents’ fears. Certain crimes as well as forms of antisocial behavior act as signals that
a community is under threat. Neighborhood reassurance policing will often focus on reducing such “signal” crimes and in some cases the response is geographically targeted, as in the case of Neighborhood Watch schemes and in the United Kingdom the implementation of No Cold Calling Zones to reduce doorstep cold calling often associated with distraction burglary.

The social scientist’s interest in crime extends to the broader social and economic consequences of crime. A few of these broader areas of interest include the damaging impacts of people’s experiences or fears of crime on health, at both the individual and population levels, on social relations and social exclusion, and on quality of life. High levels of crime also have economic consequences for house prices and for local business investment. Second, as research and scholarship have become more global there has been an increase in research into crime outside the developed countries of the West, raising questions about the universality of theories about crime and criminality built from Western experience and Western datasets. This trend has also given added impetus to research into crime outside the developed countries of the West, raising questions about the universality of theories about crime and criminality built from Western experience and Western datasets. This trend has also given added impetus to research into, for example, the impact of corruption and violent crime on economic development and the effects of political-economic transition on crime rates and crime profiles in, for example, China (post-1979) and the former Soviet Union and its satellite states in Eastern Europe. These areas are of relevance in economic, political, and social geography but to date geography’s representation in these fields is modest.

SEE ALSO: Behavioral geography; Built environments; City logistics; Cluster detection; Cognition and spatial behavior; Feminist geography; Local statistics and place-based analysis; Neighborhood; Security; Social geography; Spatial analysis; Spatial context; Urban geography

References


Further reading

“Critical geography” is a term used to describe a diverse set of perspectives that are committed to practicing forms of emancipatory thinking and action. This approach is directed both within and beyond the discipline of geography, encouraging progressive social change while also challenging the processes and institutions through which geographical knowledge is produced. Under the circumstances, attempting to have an international or all-inclusive view of the field of critical geography poses a challenge since the term covers an array of Marxist, feminist, anti-imperialist, anarchist, antiracist, and poststructural approaches (Peake and Sheppard 2014). “Not only is critical geography everywhere,” writes Nick Blomley (2006, 90), but “it is diverse and inchoate.” Consequently an attempt to narrate an overarching definition operates in contradiction with the very objectives of the production of critical geographical knowledge where the desire to challenge dominant forms of power has involved a deliberate move away from high theory and the privileging of abstract textual issues (Harvey 2006). This diversity is partly a reflection of the proximity of this concept to other affiliated terms, most notably radical geography. Where radical geographies emerged in the 1970s, heavily influenced by Marxist thought, critical geography was a slightly later intellectual and political development. Like radical geography, it shares a desire to challenge inequality and oppression, but it is less wedded to structural interpretations of this process and consequently less focused on political economy. Where radical perspectives have sought to illuminate the underlying class conflict shaping geographical phenomena, critical geographies have interrogated questions of representation, culture, and language. The entry will return to this distinction later, after spending some time considering the evolution of this field of study within and beyond the discipline of geography.

Historical trajectory of geography

It is tempting to trace the emergence of critical perspectives in human geography to the widespread incorporation of social theory into geographical approaches in the late 1960s and early 1970s. But to do so would be to overlook the long tradition of critical dissent within the study of geographical systems and process. For example, Gerry Kearns’s (2009) account of the influence of Halford Mackinder’s imperialist perspectives of late nineteenth- and early twentieth-century Britain demonstrates the presence of contemporary thinkers questioning dominant imperial geopolitics. For example, Petr Kropotkin, a Russian émigré to the United Kingdom, studied the struggle of communities in eastern Siberia over limited resources, where he witnessed both forms of competition but also – crucially – practices of cooperation. The challenge for thinkers such as Kropotkin was to confront geographical perspectives that relied on principles of environmental determinism,
an approach based on a paradigm of social Darwinism where imperial incursions were justified on account of the supposed primitivism of native populations. Perhaps the most powerful element of such determinism was its claim to scientific neutrality, where exploitative and violent geopolitics were represented as necessary responses to adverse climate, topography, or other environmental conditions. In constructing a critical response, Kropotkin did not reject social Darwinism but, rather, accused such imperialist writers of misrepresenting evolutionary biology, where species success is not simply a product of struggle but also of cooperation and mutual aid.

This example sets out some of the established tenets of critical geography that have flourished over the latter part of the twentieth century. Perhaps the most prominent characteristic of this perspective is a challenge to the imagined objectivity and neutrality of science while exposing and confronting the power relations that underpin scientific “objectivity.” Certainly, much of the Marxist, feminist, and humanistic approaches that emerged in Anglo-American geography of the early 1970s confronted the underlying positivism of the spatial science practiced in the 1950s and 1960s. One of the precepts of being a part of science was to have rational and objective viewpoints – the “enlightened” way of seeing the world as orderly, neatly organized, and causally generalized. In a quest for an essentially rational order, the production of “truth” meant the use of methods that would lead to overarching theories. It was not surprising, therefore, that objectively produced knowledge, which could stand scientific rigor, scrutiny, and validation, took precedence over other forms of knowledge. Geography came to be seen as nomothetic (a science of general or universal law) rather than ideographic (place-specific or unique work), leading to increasing emphasis on spatial scientific approaches. Perhaps the most tangible evidence of this intellectual prioritization can be found in the quantitative revolution, which had marked an intellectual shift, particularly in Anglo-American geography, from the mid-1950s. This phase was characterized by theorizing and the adoption of statistical techniques of description and empirical verification in a positivistic manner. This approach, often termed spatial science, held that space itself was an external characteristic of the world with its own discoverable laws.

The challenges to spatial science were both plural and mutable. Early critical approaches were directed not toward creating new political or social theory but to addressing the absence of certain topics from geographical scholarship. In consequence, the early 1970s saw scholars drawing into the disciplinary concerns of geography issues as diverse as civil rights claims in the United States, rising global environmental concerns, the conflict in Southeast Asia, and Cold War conflict. Following this widening of thematic concerns, attempts began to consider how the observed inequality and oppression, operating at multiple geographical scales, could be adequately theorized by geographers. Initial works, first by David Harvey but later including James Blaut, Richard Peet, Neil Smith, and others, drew on Marxist approaches to connect geographical inequality to the operation of the capitalist economic system. In a move that has become a characteristic of critical approaches, their works drew attention to the dual practice of oppression (economic or social marginalization through unequal economic power relations) and practices of representation (the way the world is narrated and symbolized in policy and media environments). Critical geography in these terms was as much about trying to challenge and disrupt capitalist representations of the world as it was about building an overarching spatial theory of capitalism.
Interest in challenging the production of knowledge and unsettling powerful representations of the world were traits that extended beyond Marxist geography. At a similar time, a range of scholars, illustrated by but not limited to the Women and Geography Study Group in the United Kingdom, explored ways in which feminist political perspectives could both challenge conventional geographical perspectives and deepen Marxist critical approaches. The late 1980s witnessed debates regarding the merits of drawing on feminist and Marxist concepts to see if gender and class politics can be intersected, in particular around practices of patriarchy (Longhurst 2002). Such initial studies were largely simple descriptions and a mapping of the differential geographies of men and women in terms of their disparate access to services, employment, and facilities while also addressing women’s position and status within the discipline. Later, scholars started offering newer and more complex interpretations of place and their intersections with the existential specificities of women’s lives, as well as new insights into women, gender, and gender relations. And yet, these developments were essentially taking place in an Anglo-American context. Elsewhere in the world, such as in India, collective thinking and open debates about such intellectual concerns were largely absent, while the intellectual frameworks within which these issues were discussed continued to remain firmly on empirical terrain.

The situation changed somewhat with the inclusion of gender. Although the initial themes studied in South Asia appear rather archaic when placed against issues that have been explored in the Anglo-American world, the very entry of gendered subjects was to push the limits of geographical analyses to critical thinking in opening up a terrain of investigation into asymmetrical power structures between men and women, their differentiated spatial relationships and practices, challenging the established androcentric tenor of the discipline. One of the preoccupations of critical thinking has also been decentering gender as the primary category of analysis. For example, the presumed linear developmental trajectory of education, access to gainful employment, and women’s empowered status is being critically looked at to situate gendered subordination in complex and multilayered structural environs marked by caste, class, religion, and ethnic differences, mediated through spatial locations. India offers a case where spatial and spatially embedded social structures and norms have significant holds on gendered relations, notwithstanding the argument that gender relations cannot be conceptualized as fixed and static (Raju 2002; 2011).

In tandem with the emergence of feminist critical perspectives, scholars sought to draw on burgeoning postcolonial scholarship to consider the ways in which geographers could offer critical accounts of both enduring forms of imperial practice and the mechanisms through which non-Western voices were silenced in the geographical discipline. Often the “third world” was seen as providing empirical “case-studies from … exotic others for the consumption of powerful knowers in the centre” who are from the Anglo-American worlds (Simonsen 2004, 526). As with feminist approaches, postcolonial critical geography is thus focused as much at the sites, bodies, and language through which geographical knowledge is produced as it is on specific colonial histories or relationships. In this way, this wide body of postcolonial scholarship has sought to illuminate the inequalities between colony and metropole while simultaneously exploring how processes of knowledge exchange, representation, and claims to expertise allowed such inequities and oppressive practices to be masked, reproduced, and, on many occasions, celebrated.
CRITICAL GEOGRAPHY

The diversity of intellectual perspectives did not end in the 1980s. The years since have seen a constant process of extending, challenging, fragmenting, and integrating different theoretical approaches in order to sharpen the tools of critique and social change. Perhaps the distinctive approach of the 1990s was the more wide-ranging incorporation into critical geography of poststructural approaches, in particular practices of deconstruction inspired by the work of Michel Foucault and Jacques Derrida. This move has oriented attention to the relationship between the production of space and its representation and, in particular in the work of Foucault, tracing the historical genealogies through which social categories gain meaning and significance. Language is no longer regarded as a neutral medium of communication but it is seen as the means through which particular hierarchies, classifications, and forms of inequality are normalized and reproduced. Such a focus on words turns inward on the discipline itself, where scholars such as Sparke (2005) have sought to deconstruct the etymology of geography to understand the practice as geo-graphy, quite literally, “earth-writing.” In these terms, geography is less a distinct academic discipline than a discourse through which particular imaginations of the world are stabilized and established as truth and which also provides the tools for resistance and subversion.

Critical geography and geopolitics

The turn to deconstruction as a form of critique fostered diverse reactions and responses within and beyond geography. The emergence of a field of critical geopolitics provides a lens to explore such contestation. Critical geopolitics, just like critical geography, is a diverse field which some see less as a coherent body of work and more as a style of scholarly analysis, or simply as a never-met aspiration. But the term has been applied to the body of work emerging in the late 1980s and early 1990s that sought to decode powerful accounts of international affairs, and in particular the foreign policy accounts of major Western powers. Starting with the policy prescriptions enacted toward the end of the Cold War, and proceeding through similar texts confronting the 1991 Operation Desert Storm in Iraq, through the interventions in the fragmentation of Yugoslavia, and on to the US-led “war on terror,” this work has sought to illuminate the ways in which political agendas are legitimized through specific representations of global space. Ó Tuathail and Agnew (1992, 192), two key proponents of early critical geopolitics, suggest that the purpose of this critical perspective is to rethink geopolitics as a “discursive exercise” by which “intellectuals of statecraft spatialize international politics in such a way as to represent it as a ‘world’ characterized by particular places, peoples and dramas.”

The reason for pausing on critical geopolitics is to investigate the response to this style of discursive technique, a set of reactions that provide a telling microcosm for broader concerns about critical geography. In particular, debate has focused on what is “critical” about critical geopolitics. Where initially the critical credentials of this scholarship were signaled by the disruption of powerful narratives about the world, increasingly feminist and postcolonial scholars have confronted the ability of such accounts to offer alternative visions of the world. One concern is that a rather straightforward toolkit for critical deconstruction can be established that neither encourages transformative politics nor gives voice to those who suffer the consequences of violent foreign policy incursions. In addition, feminist scholars of geopolitics have argued that these initial works
within critical geopolitics had little consideration for their own embodiment (mainly the works of men) or situated nature (in elite Anglo-American universities). In response to these concerns, work adopting a feminist geopolitical perspective has sought to reorient the focus of geopolitical analysis away from formal state pronouncements to everyday sites through which geopolitical ideas circulate and are challenged. Consequently these approaches foreground alternative loci of security, whether in civil society, the community, or the family.

Where feminist and postcolonial scholars have critiqued the transformative potential of critical geopolitics, more recently scholars have begun to challenge the emphasis placed on representation. Rather than simply analyzing and deconstructing policy texts, these interventions have drawn attention to the ways in which geopolitical discourse is performed through bodies, materials, and institutions. Such work could be framed as a more-than-representational turn in critical geopolitics, or as a rematerialization of critical geopolitical inquiry that moves beyond texts and images to consider the wider economic and social contexts of their existence. But there is a problem here. Scholars in critical geopolitics have always been interested in the wider material context within which texts and other forms of representation are produced; look at the fine-grained historical analysis included in the works of scholars such as Ó Tuathail or the biographical positioning of the life and work of noted US geopolitical Isaiah Bowman in Neil Smith’s (2003) account of early twentieth-century US geography. In addition, discourse is never simply textual or merely representational, it is always also a performance that can be traced through its production (as scholars have provided behind-the-scenes accounts of journalistic and policy practices), its performance in particular arenas, and its various receptions by different audiences. These are processes as material, mutable, and dynamic as any other.

**Critical and radical geography**

The vexed question of the relevance of critical geography necessitates revisiting the distinction between radical and critical geography. Considering the wide array of perspectives outlined earlier, it is no surprise that some scholars have sometimes become dissatisfied with the “critical” label as being too diffuse and inclusive, claiming instead that their work and political practice is “radical,” reflecting the longer lineage of Marxist geography conducted under this latter label. Supporting this position, Purcell (2003) critiques the coherence of critical geography, arguing that “islands of practice” have emerged, independent of each other and working on parallel forms of political oppression. Within this narrative, scholars interested in capitalism were often intellectually and politically isolated from those engaged in studies of patriarchy, racism, or heteronormativity. Similarly, there existed disconnects between scholars exploring global processes and those involved in local-scale processes.

Drawing a line between radical and critical continues to be a fraught practice (not least as many would – and do – self-identify with both). Geographers have presented radical geography as a diverse field that is interested in undertaking scholarly work that has explicit relevance for people’s everyday lives. At the same time, much of the work that is gathered under the label of “radical geography” has remained wedded to structuralism, thereby seeking to illuminate and challenge the frameworks of power that reproduce inequalities (broadly understood) and legitimize oppression. But there is also a question about the location and privileges of knowledge production, where radical scholars have claimed
CRITICAL GEOGRAPHY

that the “radical” elements of critical geography have been lost as feminist, Marxist, and post-colonial practices have been incorporated into the disciplinary structures of Anglo-American human geography. But, in many ways, critical geography has also attempted to retain a sense of distance from traditional disciplinary institutions and practices. One example would be the establishment of the International Critical Geography Group (ICG), a coalition of scholars from across the world (the 2005 ICG Conference in Mexico had 35 participants) which challenges nationally aligned geographical societies while thinking through transnational practices of critical geography. Similarly, the journal ACME, established in 2002, defines itself as an international e-journal for critical geographies and is organized around a fully open access format, subverting the restrictive financing model of more mainstream geographical journals.

Critical geography and relevance

A controversial issue in critical geography has been social relevance. Some argue that social relevance can be acquired by establishing a bridge between academia and (political) activism. A few geographers examining gender are also trying to bridge the gap between research and praxis by getting involved in activities of civil societies and action research, and by positioning themselves between academia and policy implementers. And, yet, what constitutes such activism and should it be pursued? Whether critical issues should be attended to from within academia rather than by joining direct action outside remains contested and unresolved. A number of critical geographers have also sought to move away from such bifurcation between the academic and activist categories, arguing for a more nuanced reading of the locations and identities for critical geography. There can be a third space between academia and activism, a space of continual flux and movement wherein there can be activism within academia and academic endeavors within activist settings (Blomley 2008).

In the South Asian context, the question of geography’s relevance has to be placed in a wider frame. There the rigid definition of what constitutes geography and what does not, which of the strands are appropriate and which are not appropriate, has to be continually challenged, reworked, and contextually grounded. This challenge has to happen without giving up (despite interdisciplinary exchanges and discourses) one’s training as a geographer because of the disciplinary edge geographers have in decoding the spatial embeddedness of social relations and social formations, and providing a nuanced analytical framework of interdependencies between the environmental bases of spatial attributes and the social and historical processes leading to a theory of social causation. Such integration could open up the tremendous possibilities for context-specific critical geographies. Recent research shows that geographers are well equipped to address issues of rising fundamentalism and ethnic reasservations as reactionary outcomes of the hegemonic appropriation of resources and opportunities. Scholars have related these concerns to questions of distributive justice and the role of civil society, distorted spatial relations in a globalizing world, and marginality issues in the context of international capital flows.

In sum, a grounded recognition of how contextualized understanding of the processes of knowledge production has to position critical geography in multiple ways, dispossessing a desire for a monolithic construct of the content of critical geography. Critical geography is now an established branch in much of Western academe and critical geographers occupy places
of eminence. Critical geography has ushered in not only normative changes in human geography, but also, in the process, transformative effects on the discipline itself, for critical geographers are using a wide range of critical social theories to address substantive issues related to race, caste, class, colonial power, and inequality.

SEE ALSO: Feminist geography; Marxist geography; Postcolonial geographies; Radical geography

References


Critical GIScience

Nadine Schuurman
Simon Fraser University, Canada

In 1999, Schuurman published a monograph entitled Critical GIScience: Theorizing an Emerging Discipline. This work outlined the emerging phenomenon of human geographers engaging with the computationally based subdiscipline of geographic information systems (GIS). The monograph examined the human developers of geographic information science (GIScience), as well as the technology, through the lens of critical human geography (Schuurman 1999). Though this was the initial use of the term, “critical GIScience” has broad antecedents in GIScience, human geography, and social theory. The use of the term “critical” in human geography and social theory more generally is a means of marking the work as an inquiry into a range of perspectives that include gender, social construction, and science and technology studies (STS), as well as Marxism.

The word critical is derived from the Greek word *kritikos*, which focuses on questioning and analyzing phenomena, especially those that might otherwise be taken for granted. Over the past 50 years, a number of social theorists have identified their work as “critical,” with the implication that they are looking at social and practical implications of a given event, practice, or movement, rather than taking its existence at face value.

The critical movement in human geography was definitely a chief influence in the development of critical GIScience (Leszczynski 2009). Since the 1950s, there have been fault lines in the discipline between those committed to quantitative analysis and those who pursue more human and social axes of inquiry. GIS, as they emerged from the mid-1960s to the 1990s, were squarely lumped in with quantitative geography. In the mid-1990s, two parallel developments altered this division. First, the National Center for Geographic Information and Analysis (NCGIA) was formed and started to integrate historical and participatory aspects of GIS into its agenda (Goodchild and Mark 1993). Second, human geographers, influenced by STS and critical theory, began to question the agenda of GIS (Sheppard 1993).

By the 1990s, STS and critical human geography converged into a series of critiques of GIS. Perhaps the earliest (and most viral) of these was by the late Neil Smith when he linked the first Gulf War directly to GIS technologies. In 1993, a meeting was held in Friday Harbor, Washington, to examine some of the suspicions being cast on GIS. This meeting was attended by a number of GIS scholars, as well as by human geographers, and resulted in a special issue of the *Cartography and Geographic Information Society Journal* in 1995.

It was in this special issue that accusations of gender bias, Cartesian perspectivalism, military links, and issues of race emerged as a collective set of criticisms. They were informed in large part by the emerging literature on STS, as well as by emerging feminist and critical geographies. Their goal was to illustrate the axiom that the direction of science is not separate from its surrounding culture; indeed, they are inseparable. More specifically, the concerns and direction of science are closely related to the important issues of the day. Bruno Latour famously describes this inter-
twinement of scientific and societal priorities by illustrating how scientific discourse was shaped during the development of smallpox vaccine.

For the most part, influential STS scholarship has focused on scientific inquiries in physics and biology. However, STS-informed critiques of GIScience had perhaps a greater impact in the discipline of geography – perhaps because geography is a small enough discipline that ripples in the pond are noticed.

If the journal special issue was a ripple, the next disturbance was a tsunami. In 1995, close on the heels of these early criticisms, a new book entitled *Ground Truth*, edited by John Pickles, was released (Pickles 1995). This compilation addressed various perceived flaws as well as the positive potential of GIS. For the most part, however, its tone was somewhat derogatory, and it included entries that did not display a sophisticated understanding of the underlying technology nor its representational challenges. The edited book, and its contribution to critical GIScience, is well documented in Schuurman (2000). A positive outcome of this challenge to the technically oriented and philosophically detached GIScience research community was to force consideration of some of the issues that had been raised in STS, such as theoretical underpinnings to semantic heterogeneity (Harvey and Chrisman 1998), the potential role of algorithmic choice in scientific outcomes, and the absurdity of implicating GIS in the woes of the world.

This abbreviated history of critical GIS is elaborated on elsewhere (Schuurman 2000). However, it is important in setting the context for the development of a cadre of scholars that took the issues associated by critical GIS very seriously. For many, this became their career path. It is these scholars who, today, are beginning to explore mapping-related artifacts of georeferenced Internet-based data and services collectively referred to as Web 2.0, as well as neogeography.

Web 2.0 is often synonymous with social media. Neogeography, coined by Andrew Turner in 2006, is a spatial iteration of Web 2.0 in which the old professionalism associated with map-making disappears as the practice becomes widely accessible to untrained users. This profound democratization of mapping and geovisualization affects conceptualization and expectations of an emerging GIScience as all online mapping venues are expected to include interactions (Sui and Goodchild 2011). Indeed, exchange of information with attendant ability to interact, exchange data, and download “shareware” has become routine. In this, GIS follows all social media. It is this enactment of GIScience as media that is the basis for volunteered geographic information (VGI), location-based services (LBS), spatial big data, and attendant privacy concerns.

Leszczynski recently alerted us to the dangers of distinguishing “neogeography” from plain old GIS and thus exempting its many dimensions from the scrutiny of critical GIScience (Leszczynski 2014). Never has this counsel been more apt as critical GIScience scholars begin to grapple with the implications of unimaginably large collections of geographically referenced data on the web. These include critical examinations of VGI, LBS, and big data, all of which are the basis for an emerging *new critical GIScience* that encompasses the many implications of Web 2.0 and neogeography.

VGI was coined (like the appellation GIScience) by the pre- eminent GIS scholar Michael Goodchild (Goodchild 2007). He recognized very early that mobile devices and persuasive computing were the basis for users to contribute widely and frequently to “citizen science” efforts. Indeed, VGI is the basis for populating spatial databases of many stripes. He
used the example of the forest fires that raged across southern California in 1997. Previously, government officials were charged with warning potentially affected homeowners when they should evacuate. Residents, for the first time, were using Flickr *en masse* to report on the fires’ extent. Homeowners could go onto Flickr and see “where” the fire was. Geotagged photos were far superior reporters of the geographic coverage of the fires than any government official, who could, after all, not be everywhere at the same time (Goodchild 2007). During a subsequent fire, volunteers used geotagged information to create online maps of fire extent that received hundreds of thousands of hits (Goodchild and Glennon 2010). Web-based geotagged information, if reported by enough people, was everywhere at once.

VGI is both a boon and challenge to governments around the world. After all, anyone can report data through geovisualization and maps. Urban planners are now faced with multiple possible neighborhood futures and must seek ways to acknowledge points of view that may have been unsolicited. Legal concerns shadow VGI as well. Liability, ownership, and reliability often remain unknowable (Haklay 2010). The proliferation of VGI also destabilizes spatial data infrastructures that were painstakingly erected and controlled by national governments (Elwood and Leszczynski 2011).

The once fanciful idea that the world is populated by a pervasive army of potential reporters on a wide range of social and scientific phenomena is now a reality. But trenchant critics of GIS have pointed out that VGI is not in fact reported by all, or even a cross-section, of the population. Instead, it remains the domain of mostly male, prosperous first-world citizens – and only a slim portion of those (Lesczynski and Elwood 2015; Haklay 2013). This reminds us that all technology (and data) is socially embedded. And the data that they report on reflect this. There are many classifications of bars and pubs reported in some cities by VGI contributors but none for child care (Stephens 2013). This raises a question central to the *new critical GIS*: Whose data is reported and for what purpose?

This single question is the basis for emerging scholarship that links critical GIS to the geoweb. For instance, VGI is the basis for populating the brave new world of independent maps and software owned by “the people.” But when the “V” in VGI is absent, the data falls into the domain of LBS. Contributors to LBS (e.g., most of us sporting smartphones) are also unwitting data volunteers for corporations keen to sweep up any crumbs of personal information that could result in more persuasive advertising and greater consumption. Any example of VGI as a positive scientific contribution (e.g., ubiquitous temperature reporting in the interests of monitoring global climate change) can be matched with a more sinister one linked to LBS, such as “stalker” apps that report users’ locations to potential predators, often without their knowledge or consent (Leszczynski and Elwood 2015).

When VGI does occasionally sport a dark side, there are almost always LBS involved. For example, locationally aware mobile devices reported girls’ locations back to a central server, which distributed spatial data on proximate girls to subscribers of the service Where the ladies at, mapping the locations of concentrations of women (based on public Foursquare check-ins) as “hotspots” or densities. While this particular LBS has been discontinued, the potential for ethically suspect use of locational data is enormous. Critical GIS scholars are alerting us to ethical dimensions of LBS (Wilson 2012). Perhaps the most striking is the uptake of LBS by corporations largely enhanced through mass participation in social media.
But it is our devices themselves – unrelated to social media – that perhaps offer the greatest technological advances with attendant increases in privacy violations. Since the introduction of Apple iOS 8, iPhones and iPads have been using iBeacon, a Bluetooth-enabled service that allows retailers to determine exactly where potential customers are located (McFarland 2014). Using triangulation, iBeacon can pinpoint in which aisle of a grocery store someone is standing or which painting they are in front of in a museum (a more benign example). In a retail environment, a coupon could be sent to the customer as they view pizzas or big screen TVs in front of them. In a museum or other public setting, a person could sign up for a guided tour of the facility. Proponents will argue that these “intrusions” are agreed upon by the customer through a consent process. However, when money is involved, consent can be a dubious concept.

These small spatial incursions are multiplying. Since the Mac OS called Yosemite was released in 2014, just opening the search function called Spotlight triggers the OS to send a precise location to Apple (Soltani 2014). Tweets, Facebook entries, Instagram photos, and other forms of self-publication on social media contribute to an emerging profile of each of us. It is the linking together of these various bits of personal information (much of it spatial) that we all seed on the web that forms the behemoth that we now refer to as big data.

Big data is, in effect, all the little data crumbs that each of us leave as we cruise the web. Increasingly corporations have harvested these crumbs and developed individual profiles of each of us in the interests of promoting consumption (Leszczynski 2014). The ability to cross-reference these data from various sources to build detailed profiles of individuals as they move through their lives both horizontally and vertically requires significant algorithmic and processing power. Corporations around the globe have stepped up to the task.

While these incursions into our most private lives may seem minor when considered individually, they are significant in the aggregate. The New York Times started to take notice in 2012 that companies were using marketing information that they both collect and purchase (which often is scraped from social media and linked to email addresses) to do “predictive analytics.” Data is not just about what we buy and where, but about how we live. Insurance companies are compiling rates and restrictions based on our lifestyle choices. Whether you are a deer hunter, an avid runner, or a slothful pizza consumer, all of these attributes are linked to life and other insurance applications. These infiltrations are becoming more pervasive as they are linked to scraps we leave on social media. Facebook is at the forefront as it harvests postings of personal activities and combines these with data they purchase from other data scrapers to create custom ads for individual users. Google takes the stakes higher as it created educational apps aimed at college students and then mined their email addresses to link to other activities. That almost all of these data acquisition forays contain valuable spatial data seems self-evident. But when Nordstrom put up signs divulging that they were tracking customers as they roamed department stores, customers were nonplussed and decidedly unhappy.

Perhaps this is because most big data profiling is conducted stealthily and in North America people have chosen to be relatively complacent (in contrast to Europe, which supports greater consumer vigilance). As scholars, we have also been complacent, having paid relatively little attention to spatial aspects of big data. This is changing as scholars from many disciplines wake up to the social implications of pervasive and ubiquitous data collection. Our attention is also stretched
thinly as data collection is not confined to corporations but eagerly pursued by governments. Governments are increasingly the most thorough gatherers of Internet data about individuals. The Snowden leaks in the United States are perhaps only a glimmer of the future in which data (including geographical data) on individuals are assembled en masse. In Canada, in January of 2014, the government was found to have collected travel pattern data on individuals using Wi-Fi at major airports. Sentinel to both examples is not just the collection of data but its use to follow individuals as they move around the globe – far outside the jurisdiction of the governments tracking them. And big data transactions are (or should be) the purview of new critical GIScience scholars and citizens concerned with democratic concepts of liberty.

Digital divide

Some of the potential problems associated with neogeography are first-world problems. As Graham (2012) enumerates, high-speed Internet access is very unevenly distributed across the planet (Graham 2012). Indeed, variable web access reinforces existing divides between the privileged and underprivileged across all the familiar categories, including education, income, and social capital, not just in low- and middle-income countries but in Europe and North America. Just fixing the problem of income inequality will not necessarily close the digital divide. Digital inequality in access to computing equipment and in skills to operate is only a symptom of existing inequities.

Long ago, Haraway reported on the danger of allowing a patriarchal computing world to emerge (Haraway 1991). Rather than chastise men for their role in the evolving “cyborg,” she urged women to become more active in its architecture. Only by learning to program, and engaging with the waking beast, could women and feminist objectives be woven into the construction of the web. The contemporary artifact is the geoweb and its character.

A number of critical GIS scholars have reported on ways that gender relations are reinforced rather than reforged as the geoweb emerges (Leszczynski 2011; Leszczynski and Elwood 2015; Stephens 2013). However, there are many openings for women to shape and be shaped by emerging technologies (Schuurman 2002). Perhaps the largest barrier to this aspect of the digital divide is our own epistemological programming. Pavlovskaya (2009) urges us to use the opportunities of the geoweb to forge new epistemologies and, as a result, allow new ontological categories to emerge in GIS. Cope and Elwood (2009) literally open the door for this development in their book Qualitative GIS, in which they provide many road maps for these categories to emerge from qualitative data that was historically not considered suitable for GIS.

If traditional GIScience and the geoweb have reinforced a quantitative masculinity, they also open the door for new modes of expression, including an acknowledgment of the role of emotion in human expression – whether digital or in person (Kwan and Guoxiang 2008). LBS are freely utilized in this contemporary expansion of expression as men and women broadcast their preferences (along with location) in what Wilson (2012) brands conspicuous mobility. Ironically, the openings that women seek in virtual worlds are frequently those that make them physically vulnerable as well as corporate prey as data is harvested on their every preference and move (Leszczynski and Elwood 2015).
Arguably the antecedents to critical GIS began in the 1980s as STS gained a foothold in the social sciences. A highly distilled account of the messages of STS might be: to understand science, examine its social context. This lesson, like so many from STS, is especially pertinent as critical GIScience remakes itself in the wake of Web 2.0 and the geoweb. Dawn Wright (2012) expressed the value of critical GIS in this context articulately: as GIS becomes inseparable from the web of devices and data that surround our earth, there is an increasingly urgent need for critical evaluation of their effects on geographical scale, proximity, society, and surveillance (Wright 2012).

In our delight over the wonder and possibilities of a newly configured geoweb, we can lose track of the fact that corporations evince a parallel delight in ubiquitous spatial data (Leszczynski 2014). Governments likewise have jumped on the data collection bandwagon with egregious incursions into personal lives in the name of security (Lyon 2014).

As forecast long ago, we have become digital personae (Poster 1996), lightly sketched holograms made of data bits scavenged without our permission but willingly donated to corporate profiles as we virtually cavort through the playgrounds of social media. The cookies we leave on the web, along with our email addresses and other personal identifiers, are being woven into our profiles. It is these implications of the Brave New World of VGI, LBS, and other aspects of web mapping and spatial analysis that form the basis for an emerging new critical GIScience over the coming decade(s).

SEE ALSO: Critical geography; Critical spatial thinking; Feminist geography; Geographic information science

References


Critical spatial thinking

Diana S. Sinton
Cornell University, USA

Background and overview

Critical spatial thinking incorporates spatial concepts and geographic principles to guide and inform reasoning. A critical spatial thinker wisely and appropriately uses geographic information to understand problems, derive solutions, and communicate effectively about the process and its outcomes. Thus, he or she is following the regular practices of critical thinking within a context or situation involving geographic information, whether he or she is the producer or consumer of such information. In its simplest form, critical spatial thinking blends habits and processes of both critical thinking and spatial thinking.

The practice of critical thinking itself involves careful and deliberate consideration of the factors and variables that contribute to a situation or problem, and a subsequent analysis or synthesis that leads to a well-reasoned conclusion. As Facione (1990) stated, “the ideal critical thinker is habitually inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-minded in evaluation, honest in facing personal biases, prudent in making judgments, willing to reconsider, clear about issues, orderly in complex matters, diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are as precise as the subject and the circumstances of inquiry permit.”

Critical thinking is not a skill that necessarily occurs developmentally, or develops automatically. As an approach to knowledge acquisition and reasoning about the world, it can become a habit, and thus easier and more automatic over time and through persistent effort (Thomson 2008). Habituated actions that build upon constructivist cycles of knowledge production, application, interpretation, assessment, and re-production may be driven by an intellectual commitment to regularly use these thinking skills to guide behavior.

Spatial thinking is an ability to visualize and interpret location, position, distance, direction, relationships, movement, and change through space and over time (Sinton et al. 2013). This broad and inclusive definition is designed to cover subgroups of spatial thinking, such as the skills and abilities often associated with spatial cognition. A location or position that is being visualized and interpreted can be something in geographic space, such as a bend in a river; or something in the space of indoor human activity, such as the corner where two walls meet a ceiling; or something in abstract or digital space, such as the top of a stack of blocks being projected on a computer screen. Although the scales and environments vary, in each of these situations our cognitive spatial thinking processes rely on our capacity to visually comprehend and discern factors, such as the location or position of the phenomena and the distances, directions, angles, and orientations of how all things fit together in their respective contexts. Factors such as these determine how the where of the world is perceived and appreciated. Combining critical thinking with spatial thinking brings together the tenets of these two cognitive approaches to bear on tasks of
reasoning and decision-making. Scale, location, distance, and spatial dependence are among the factors that will affect how someone interprets a situation in a critical and informed manner.

Scale

Scale is of central importance to spatial thinking. At a personal or human scale, spatial thinking and cognition are often of interest to psychologists as they assess the ways in which individuals and groups of people are able to mentally rotate 3-D objects, or visualize objects from folded papers. But at an environmental or landscape scale, spatial thinking is consistent with geographical thinking, as the patterns and processes of the natural and social world are considered.

During the process of spatial thinking, scale interacts with location in terms of measuring and describing space. For example, the location of a point somewhere on the Earth’s surface can be identified by many different spatial reference or coordinate systems, all referring to the same place. Computers rely on these reference systems to store locations relative to the surface of the Earth, but the coordinates are also relative to a scale. For example, on a map of a particular scale, the Eiffel Tower in Paris, France, can be depicted as a single point. A computer may store the coordinate location of this zero-dimensional point with as many as 8–10 decimal places, specifying a location down to the atomic level. Yet on the ground, the actual structure has additional dimensions of height, length, and width that are not associated with the zero-dimensional single point. Without having previously recorded the 3-D data, the computer can only reference the single point location.

Thus, critical spatial thinkers need to become aware of how scale affects measurements, how scale can limit graphic representations, and how computers are capable of storing and displaying location data. Computers are capable of recording and storing coordinate data with a high degree of precision, but this is often out of alignment with both tools for measuring geographic locations and the very nature of a location in geographic space. Appreciating these characteristics of location, and how its measurements aid in assessing and evaluating evidence, is part of the practice of critical thinking. Becoming aware of the ways in which this knowledge affects the correct and informed usage of geographic information systems (GIS) is one way to develop expertise as a spatially literate geographic information scientist (GIScientist) (Sinton and Schultz 2008).

Scale has different connotations within geographical and spatial thinking, and all of these meanings affect how an individual reasons with and through space. Cartographic map scale is the mathematical relationship between units of distance on the map and the same units of distance at the place being mapped. Scale also references the resolution of geographic data in its raster and vector digital form. Lastly, scale is used to describe the extent of a study area. In the process of thinking critically about space, all of these meanings would be understood and applied appropriately and correctly in context. With this knowledge in place, an individual is ready to understand processes and patterns as they cross scales, such as interpreting hierarchies, networks, and fractals.

Location

Location is a nuanced idea that can be both relative and absolute. Any object or structure, such as a single park bench, can have its location described by coordinates and placed on a map of the Earth, based on those coordinates. Critical spatial thinking then involves knowing how that particular locale is situated within countless
other physical, environmental, and social frames of reference. How a bench in an urban center differs from a bench along a beach path, or how the same one in a city could be a different place at 11:00 a.m. versus 11:00 p.m., means appreciating the spatial and temporal context of a location. No geographic location exists within a vacuum; the setting of a location provides extensive other information to critically understand a place.

Appreciating the broader context of location is essential for critical spatial thinking. Location becomes an explanatory variable in a model. Physical, environmental, and social data are all interrelated. A location can be understood by what is taking place upwind, upstream, or uphill of it, and how, in turn, it is affecting the downwind, downstream, or downhill locations. This perspective enables scaling of the pattern and process, those ideas of hierarchies, networks, and fractals, mentioned earlier.

Location varies in its descriptions and measurements as a function of scale. Information about a location or place is collected at and appropriate for a particular scale or level of aggregation. For example, data collected about the city of New York only as a whole, with one set of values covering all of its five boroughs, cannot later be divided to provide details only about Manhattan or the Bronx. However, the data from the individual boroughs can be combined upwards, or to a coarser, larger scale. So this scaling property is unidirectional.

Furthermore, data that were collected about residents living within the same single zip code area cannot readily be compared to data aggregated at a different areal unit, such as voting precincts or school district boundaries. Because the data are not uniformly distributed throughout these areas, they cannot be divided by their areal values and it be assumed the proportions remain consistent. These situations invoke the modifiable areal unit problem (MAUP) and issues of ecological fallacy. Data and modeled relationships are ideally suited only for the scale at which they were originally collected. This truth is a challenging one under which to operate consistently, and it is one that is often violated by individuals not thinking critically about geographic data.

Distance

Like location, distance has both relative and absolute properties, and is a complex notion associated with any space comprised of more than one object (Gatrell 1983). Distances between locations within Euclidean or Cartesian space can be measured with standardized units, and by applying known conversion formulas such distances can be converted across reference systems. But distance can also be subjective and relative when considered in the context to other variables, such as topography or currents. Running a shorter distance uphill and into the wind can be a more formidable task than a running a greater measured distance downhill and with the wind from behind. Distance and rates of travel also have a temporal component. For example, during rush hour it will take more time to travel the same distance as it would another time of the day.

Being savvy about distance informs critical spatial thinking by providing a means to evaluate variables and their expected relationships. This means matching or aligning approaches to measuring distance with particular situations is the best strategy for making the wisest decisions. For example, knowing when to consider absolute or relative measures of distance, or selecting between Euclidian (as-the-crow-flies) and network routing methods, is desirable or possible for modeling movement or flow.
CRITICAL SPATIAL THINKING

Spatial dependence

Distance also underlies the fundamental principle of spatial organization, that everything is related to everything else, but near things are more related than far things (Tobler 1970). This “first law of geography” drives spatial autocorrelation, the likelihood of similar characteristics and values being present in close phenomena. It also affects how geographic patterns can be interpreted through statistics, because the samples of phenomena are not independent from one another, a result of the underlying variables that contribute to their location or placement. Critical spatial thinkers anticipate these spatially dependent patterns and are in the habit of expecting them to occur. That helps them build comparisons, envision patterns, and extrapolate to analogs, other practices of spatial thinkers. Moreover, the habit of expecting spatial dependence facilitates the identification of outliers and anomalies when they do arise. Identifying and investigating these would, in turn, be a practice of critical thinkers as they assess and evaluate interrelated variables.

Error and uncertainty in critical spatial thinking

There are no spatial data that are 100% perfect or absolutely complete for all situations, over all scales and time frames. Gaps, inconsistencies, and unknowns are inherent in the knowledge of this field. Recognizing the existence of uncertainty and being able to address and mitigate it are hallmarks of effective critical spatial thinking. First, important distinctions exist between uncertainty and error, which affect critical spatial thinking in different ways. Error is an inaccuracy known objectively; it is the measured deviation of a value from truth. Error as it is thus defined has its groundings in mathematics, logic, and the law, among other disciplines. Uncertainty, on the other hand, is a facet of our imperfect and inexact knowledge of the world. It is a measure of the difference between the data and the meaning attached to the data by the current user (Zhang and Goodchild 2002). Uncertainty is a result of error, ambiguity, or the lack of information. Errors can be deviations from objective standards, and identifying, quantifying, or documenting them may be possible. Thus, it may be possible to disclose and share the source, amount, and extent of errors. In contrast, uncertainty is more challenging to characterize and easily address, especially in any automated manner.

Both error and uncertainty are likely to be associated with geographic data and spatial analyses, and the existence of uncertainty within spatial analyses is inevitable. Because the complexity of the world cannot be modeled perfectly, and all natural and social phenomena vary over space, all digital data used within GIS-based analyses, for example, are subject to these multiple types of uncertainty. Ways in which the certainty of geographic data can vary include their accuracy, precision, completeness, consistency, lineage, currency, credibility, subjectivity, and interrelatedness (Buttenfield 1993). Through categorizations like these that help people appreciate how diverse and extensive uncertainty can be, particular types can be recognized, better understood, and then addressed.

This is the key connection between uncertainty and critical spatial thinking. Eliminating uncertainty may be an impossible task but acknowledging and understanding it are essential, especially in terms of its broader impacts. For example, products derived from GIS-based analyses are often used in decision-making by people who do not have knowledge or understanding of how the data were collected, generated, or processed. They are unaware of the many ways that uncertainty can affect the nature of the final
product. This undermines and compromises the likelihood of their making optimal choices.

Critical spatial thinking and geographic information systems

Establishing explicit connections between critical spatial thinking and the use of GIS is a goal that is both desirable and elusive. The digital platform allows users to overlay and analyze datasets that were derived at different scales and levels of aggregation, with different projections, and years apart in provenance, among many other distinguishing features. The software can provide “answers” to queries and analyses that violate basic premises of physical, social, and statistical understanding. Posing those queries and executing those analyses requires no knowledge other than pressing buttons in a particular sequence. Critical spatial thinking is neither a prerequisite nor an automatic outcome, but it is a practice that can be modeled and demonstrated effectively and efficiently with the use of a GIS. Understanding problems is as important a recognized outcome as attempting to solve them. Many of the issues and situations addressed with a GIS do not have one single correct solution but may have explanations and better choices that can be informed by critical spatial thinking and abductive reasoning.

Critical spatial thinking requires that both producers and consumers of geographic information be aware of how their information is shared and how they increasingly rely on data that are shared by others. Incorporating a reflexive component into one’s critical spatial thinking habits is one step toward becoming a spatially informed citizen (Gryl and Jekel 2012). The “geographic sense of self” is evolving very rapidly (Downs 2014), and successful critical spatial thinkers will keep aligned with and adapt to new digital technologies and applications while still following the long-recognized tenets of critical thinking and reasoning.

Critical spatial thinking and map interpretation

Another stage at which critical spatial thinking is relevant within decision-making is anticipating how others are going to perceive geographic data and its representations. There can be a false sense of credibility applied to computer-generated graphics of geographic information, via traditional maps or other graphic representations. Though many more people are now making simple maps online of point data, it is still a very uncommon practice for someone in the public sphere to use a GIS to generate a map. Thus, most people are unfamiliar with the process and do not appreciate how simple it may be to manipulate and distort the information and its depiction.

Maps convey multiple messages, whether the author and designer intended so or not. This can be attributed to the map projection that was used, a decision that results in distorted shapes, sizes, distances, or angles and bearings of the phenomena being mapped. Other map elements that can yield particular messages to map-readers include the title, the colors used, and the graphics in the map’s marginalia. One of the most common and effective ways to manipulate a map’s message is to modify its data classification scheme. This is particularly easy to do when quantitative information is being displayed via a GIS. Owing in part to the notion of inherently trusting a map, many map-readers are not adequately attentive to a map’s legend and discerning of its values. Using the GIS, the map-maker can readily modify the range of numerical values to have very unequal “bins” of values, emphasizing particular values,
Critical spatial thinking

locations, or patterns. When using a single color, it has become a convention to have light hues represent lower values (concentrations, densities, etc.) of an attribute and darker hues represent higher values. By reversing these, the map-maker also risks confusing map-readers who only take a cursory, brief interpretive glance at the image. Choosing color schemes that are mismatched with a particular type of data, or a range of data values, is a sure way to obfuscate.

Effective map interpretation habits of a critical spatial thinker include careful review of a map’s title, its legend, its scale, its projection, its data sources, its design and color schemes, and any other information that conveys messages. This becomes a deep reading of a map, rather than a blanket acceptance of the medium’s message. Such a reading of a map is the same practice that a critical thinker applies to other media: images, graphics, tables, figures, and text. Critical thinking means evaluating arguments and suppositions, and a map itself is included among these. This is linked with the skills of twenty-first century media literacy, habits developed for careful consideration of all of the forms of media used to communicate information. Often less visually evident but perhaps no less significant are the messages and intentions behind the medium itself. Comprehensive map interpretation extends to evaluating the intentions behind the map, the agenda of the cartographer or the objectives of the organization producing the data. This may invoke the guiding principles of critical GIS when digital information is being considered.

Critical map interpretation is one specific form of critical image interpretation, a foundational practice of both media literacy and visual literacy. Understanding, producing, and using images to communicate requires knowledge about representations themselves as well as a cultural perspective to be able to anticipate how others will perceive and interpret the information. This sense of perception distinguishes between whether one person her- or himself may be a clever and effective spatial thinker and how well that person may be able to teach critical spatial thinking skills to others, or help others make decisions informed by critical spatial thinking. People with natural capacities and abilities for spatial thinking are not necessarily effective instructors of these skills. The expert-to-novice barrier can be significant.

Critical spatial thinking in decision-making

The capacity to make, and help others make, informed decisions involving geographic data and situations is one of the most important outcomes of critical spatial thinking. The decision can be trivial and mundane, such as where to park one’s car to ensure continued shading, or of a significant, global nature, such as how to design and plan for protected landscape and seascape areas. The magnitude and degree of influence may vary but all such decisions still involve their own respective issues of location, direction, scale, and spatial dependence, among other spatial variables.

Critical spatial thinking may take place at different stages of decision-making processes. As mentioned earlier, a user of geospatial digital data has opportunities to think critically about data throughout its life cycle: how it was collected, at what scale or level of aggregation, how its location is recorded and with what accuracy and precision, how are its attributes defined, and so on. By being a savvy critical thinker, someone can be both proactive in taking steps to minimize error and uncertainty at the beginning of the process and reactive to error and uncertainty later.
CRITICAL SPATIAL THINKING

on. Meanwhile, certain disciplinary fields, such as planning and supply chain management, may demand critical spatial thinking as a continuous facet of a workflow, rather than an episodic one.

Multiple forms of and approaches to reasoning are engaged during a critical spatial thinking instance. In a deductive framework, there are certain theories, principles, and truths that inform hypotheses about geographic processes. That water flows downhill, and that the elevation value of one locale is likely to be similar to its immediate neighbors, informs our understanding of how to model hydrologic processes. Likewise, that a set of watersheds can be nested is also applied knowledge involving scale, distance, and spatial dependence. Expecting certain processes to occur or scale differently on the basis of their relative position within nested watersheds is an example of critical spatial thinking informed by deductive reasoning.

At the same time, initial observations – rather than predictive theories – also inform understanding about spatial processes. Inductive reasoning draws from observations and measurements to ascertain how phenomena might have been affected by their position in the landscape and their subsequent interactions with variables at different scales, for instance.

Drawing conclusions from observed patterns is a type of informal inferential reasoning that is common with geographic information, especially that which is displayed via a GIS. GIS facilitates the symbolization and concurrent display of co-located datasets that may be diverse and completely unrelated to one another. In the absence of critical spatial thinking, merely displaying spatially coincident datasets concurrently can suggest to viewers that a relationship exists between the datasets. Hence, the use of some geographic data during decision-making processes may unintentionally contribute to misguided informal inferential reasoning. Critical spatial thinking reduces the risk, but cannot eliminate it.

Geographically based decisions frequently involve information derived at many spatial scales, from diverse sources, and often in an incomplete state. Thus, another approach to reasoning in the geographic realm is often an abductive one. Abductive reasoning relies on prior knowledge to qualify observations when drawing a conclusion or making a decision, and may mean making a “best guess” or taking an intellectual leap. Often associated with design processes, abductive reasoning can contribute important insights to creative problem-solving. For these reasons, and its accommodation of uncertain data, abductive reasoning fits naturally with critical spatial thinking. Making wise inferences from geographic situations typically requires this more-flexible but informed approach to understanding the world.

Critical spatial thinking in everyday life

Today’s geo-enabled society relies on location to seek and share information, and this has created another benefit to critical spatial thinking. Cell phones and other digital devices are capable of receiving and transmitting location coordinates, whether the device’s owner is aware of it or not. This information may be deliberately used by someone to help locate him- or herself on a digital map if lost, or to find goods, services, or people that are located nearby. Such technologies have spawned microtargeted advertising possibilities, such as a coffee shop being able to text a coffee coupon as a person is approaching the store. Overall, the use of global navigation satellite systems (GNSS) such as the global positioning system (GPS) have enabled and
empowered many people to travel with much greater confidence and competence.

However, there are downsides to the ubiquitous use of location information. Critical spatial thinkers keep themselves in control of how and with whom they share and leverage such data, to be able to benefit and protect themselves concurrently (Gryll and Jekel 2012). For example, cellular devices can have their location services turned on or off, based on the circumstances and applications. The sharing of location information should be done with awareness of the digital geographic footprint that may result. Nefarious individuals are able to create a comprehensive profile of a person based on when and where they travel throughout the week. Critical spatial thinkers need not become paranoid of sharing or using location information, but they can do so with forethought and intention.

One basic reason why using location to share, access, and distribute information has become widespread is that a “map” is a familiar template for organizing any type of information. Graphic displays of information are effective whether the data reside in geographic space with coordinates from the Earth’s surface, or in abstract space on the back of an envelope. Basic gestalt principles are invoked to group data or differentiate between what is near or far. This geography of our physical and intellectual spaces relies on our inherent capacity for spatial thinking as a way that our understanding of knowledge and the phenomena of the natural and social world are organized. This links our production of mental, internal representations and maps with the custom of generating and using external representations as well. Sketching a simple diagram, graph, or flowchart to communicate with others is a prime example. When these acts become familiar practices, the act of critical spatial thinking becomes a more regular habit.

Critical spatial thinking examples from the world around us

In March 2014, a Malaysian jet airliner disappeared from radar during a flight from Kuala Lumpur to Beijing. To try to locate the missing plane, hundreds of people from many different countries have interpreted vast quantities of digital audio files, satellite imagery, and other types of recorded information. Through a careful and critical process, the researchers attempted to separate the signal from the noise, enabling them to narrow the search area. From the very beginning, this task relied on critical spatial thinking to estimate flight ranges, triangulate radar pings, and design underwater search expeditions across steep, deep, and irregular bathymetric terrain.

Public health and disease monitoring are other places where critical spatial thinking is evident. Tracking the spread of contagions and infected individuals requires careful attention to both social and geographical networks of travel and communication. Being able to interpret patterns of distances, associations, and clusters contributed to quarantine protocols for the 2014 outbreak of Ebola.

Analyzing political dynamics is another activity that involves critical spatial thinking. For example, there are ecological and social principles that operate at different scales, from the global to the local. Certain political institutional forms of governance may be consistent across nations, but there is also truth in “all politics being local.” These ideas of global versus local analyses have informed the development of new spatial statistical techniques over the last few decades, producing computational methods to measure both global and local patterns of spatial autocorrelation, for instance. Global measures indicate the presence of spatial clustering or dispersion patterns somewhere across the whole
Critical Spatial Thinking

study area, and local measures can pinpoint these locales. Application of these analytical methods help political scientists understand the past, present, and possible future patterns of countries in political turmoil, such as Syria, Argentina, and the Ukraine.

Similarly, economists have argued that globalization has created a flat world in which geographical distinctions mean less and less for economics and commerce. A counterpoint argument, that place will always matter, has been put forth by geographers who maintain that human and environment interactions are paramount, whether they are historical ones that shaped society and culture or future ones that will inform resource use and trade. These are both valid points and their conclusions are not mutually exclusive. Scale and context make the difference; thinking critically and spatially means that both perspectives can be appreciated.

See Also: Cognition and spatial behavior; Critical GIScience; Decision analysis; Geography education, workforce trends, twenty-first-century skills, and geographical capabilities; Spatial concepts; Spatial thinking, cognition, and learning; Spatiotemporal analysis; Uncertainty; Visualizing uncertainty

References

Facione, Peter. 1990. Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assess-
ment and Instruction. Millbrae: California Academic Press.

Further Reading

Croatia: Hrvatsko geografsko društvo (HGD) (Croatian Geographical Society)

Founded: 1897
Location of headquarters: Zagreb
Website: www.hagede.hr
Membership: 300 (as of 2014)
President: Danijel Orešić
Contact: doresic@geog.pmf.hr

Description and purpose

The Croatian Geographical Society was founded to encourage the dissemination of geographic research and knowledge, to cooperate with other similar societies in Croatia and abroad, and to generally popularize geography. Improving geographic education and professionalization is among its major goals especially after World War II. Its members are mainly teachers, researchers, and students as well as alumni employed in business, government, and nonprofit organizations in the Republic of Croatia.

The society accomplishes its goals through its journals, one scientific and one professional/popular; occasional monographs/special editions; national congresses and other scientific meetings; lifelong learning programs; national and international school geography competitions; web geography education; and popularization. In order to achieve many of its goals, the society cooperates with the Ministry of Science, Education and Sports, government educational agencies, university geographical institutions, and regional geographic societies.

Journals or major publication series

Hrvatski geografski glasnik/Croatian Geographical Bulletin http://hrcack.srce.hr/hrvatski-geografski-glasnik

Current activities or projects

The society publishes Hrvatski Geografski Glasnik (since 1929, currently 2 issues annually) and Geografski Horizont (since 1955, currently 2 issues annually). It also organizes the Croatian Geographical Congress (every 4 years, sixth congress in 2015), the Geographic Monday Public Lectures (a dozen annually, since 1947), the Winter Seminar for geography teachers (annually since 1955), the Summer Seminar for geography teachers (annually since 1951), the geography teachers workshops, the national school geography competition (annually since 1994) with participation in the IGEO – International Geography Olympiad (participation since 2009), the web programs e-school for young scientists, and the educative portal www.geografija.hr.

Brief history

The Croatian Geographical Society was founded in Zagreb in 1897 by a group of academics initiated by the first geography professor in Croatia, Petar Matković, and lead by geographer Hinko Hranilović and biologist Alfonz Heinz. In 1899 the society joined the Croatian Natural Science Society as its first independent section and was very active until World War I. Professor Artur Gavazzi revived the society activities and in 1929 launched its scientific publication, the Hrvatski
geografski glasnik (Croatian Geographic Bulletin). After World War II the society was registered under the new political dispensation in 1947 and led by Professor Josip Roglić. The new statute defined as its mayor goals the cooperation with education authorities, including concern for teachers and their lifelong education, improvement of geography teaching practice, and so forth. To accomplish them many now activities where created, such as Geographic Monday Public Lectures (a dozen annually, since 1947), the Summer Seminars (annually since 1951), and the Winter Seminars (annually since 1955) for geography teachers. Since 1955 the society publishes a professional publication named Geografski Horizont, mainly serving geography teachers. Its active speleological section was created in 1956. The Croatian Geographic Society continued its scientific activity in the postwar period when it organized the First Congress of Geographers of Yugoslavia in 1949 and building on this first congress, organized the Seventh Congress of Geographers of Yugoslavia in 1964.

As Croatia gained its independency, the Croatian Geographic Society organized the First Croatian Geographical Congress in 1995 and continues to do so every 4 years along with organizing other scientific meetings. The society has remained the umbrella organization for all its geographers, though independent regional geographic societies are also active. The society furthered its cooperation with the Ministry of Science, Education and Sports and with educational agencies and worked on geography curriculum questions, state high school graduation exams, and organized many geography teachers workshops, school geography competition on national and international levels, and so on. The future mission of the Croatian Geographic Society is to remain on the path of advancement of geography education and of validation of geography as a modern science which, in dealing with functional organization of geospatial systems, can provide many answers to questions of our sustainable development.

Submitted by Danijel Orešić
Cross-border, transnational, and interregional cooperation

James W. Scott
University of Eastern Finland

For several decades, subnational processes of cooperation across state borders have enjoyed considerable research attention within social and economic geography. Interest in these areas has been spurred by political and economic changes of global dimensions that have unfolded since the end of World War II and that have accelerated since the final demise in 1991 of the Cold War order. An inclusive but yet concise definition of cross-border cooperation is that of a political project carried out at subnational levels by public, private, and civil society actors with the express goal of extracting benefit from joint initiatives in various fields. Generally speaking, there are three distinct but often interconnected levels: cooperation within local and contiguous cross-border spaces; cooperation between regions beyond local border contexts; and transnational networks that can involve cities or groups of cities.

The transnational is the most flexible and least spatially fixed of these cooperation contexts; it basically involves the project-oriented networking of actors (e.g., cities, nongovernmental organization (NGOs), economic groups). Conceptualized as an example of “subnational paradiplomacy,” transnational cooperation networks have largely been a focus of international relations and scholars of globalization theory. Research in this area was especially plentiful during the 1990s when putative “postsoverign” transformations of nation-states were understood to produce a new global patchwork of social and political relations between substate actors (Aldecoa and Keating 1999). Examples of transnational cooperation actors include Pacific Rim cities such as Vancouver, Canada, and have sought to intensify relations with Chinese counterparts, and regions such as Quebec which have developed working relationships with other francophone regions around the world. More recently, the concept of transnational cooperation has been appropriated by the European Union within its regional policies; in this case it refers to flexible international networks between national, regional, and local actors in dealing with specific development issues (see Figure 1). The interregional level of cooperation is more specifically European as it emerged out of EU attempts to decentralize, denationalize, and “Europeanize” structural policies. The European Union’s Interreg program – dedicated to cross-border cooperation – provides a platform for development projects between regions within member states.

Finally, the local level of cooperation, which will be referred to here generally as cross-border cooperation (CBC), is the most intensive as it generally involves multiple networks of public, private, and civil society actors that work in close geographical proximity. For historical reasons, it was in continental Europe’s economic core that cross-border cooperation, as it is understood today, began to take root. Here, the confluence of local interests, spurred by the momentum of European economic recovery and prospects of integration, promoted the development of new forums for dialogue and conflict resolution.
The inauguration of a trinationally focused Regio Basiliensis, for example, was driven by a need for the Swiss city of Basle to compete with Zürich and to seek alternatives to the parochialism of national politics, dependent as it was (and still is) not only on its immediate French and German neighbors, but on the greater European context as well (Briner 1986).

While all three cooperation contexts are significant, it is the latter level, that of cooperation between contiguous border areas, that has received most research attention within human geography. One reason for this has been the subject itself – border cities have been a long-standing subject of geographical inquiry, and the investigation of the functional relationships within cross-border urban areas presaged considerations of direct political cooperation. Larry Herzog’s (1990) study of the San Diego–Tijuana urban region on the US–Mexican border and Claude Raffestin’s (1974) analysis of the Geneva functional region between Switzerland and France are important examples of such research. Another major reason for the development of cross-border cooperation as a research field in its own right can be attributed to attempts at creating zones of free trade and continental integration in North America and Europe. These processes brought into focus many of the issues and problems involved in the elimination of barriers to social, economic, and political interaction. European integration in particular was instrumental in enhancing the status and visibility of local CBC.

Social geography deals with cross-border cooperation in all its forms as it has been associated with state–society paradigms that suggest that new forms of politically relevant action increasingly take place beyond the seemingly inflexible territoriality of the state. Similarly, research has been driven by a concern with the consequences of state–society transformations for economic, political, social, and cultural life. The putative significance of CBC is that it provides new political spaces within which to address locally issues that include social affairs, economic development, minority rights, cross-border employment and trade, the environment. Furthermore, through such territorially flexible forms of political and economic interaction – both institutional and informal – it has been suggested that greater cost-effectiveness in public investment can be achieved, economic complementarities exploited, the scope for strategic planning widened, and environmental problems more directly and effectively addressed. While such cooperation initiatives have proliferated in Europe and North America, they have also developed in Asia and elsewhere, lending credence to the notion that CBC is a global phenomenon.

Comparative analysis of local CBC processes has emphasized their contextual nature and, in general, research has indicated that CBC outcomes are subject to (geo)political agendas as well as cultural and structural factors that can enable as well as constrain cross-border cooperation (Scott 1999). In other words, basic differences between regional situations with respect to international integration logics, the conditioning role of states, local cooperation contexts, and different modes of supporting cross-border cooperation explain much of the variation in CBC experiences.

As Figures 1 and 2 indicate, the European Union has developed a comprehensive policy of supporting cross-border interaction at the three different levels discussed earlier. These maps suggest a highly networked European Union that is more complex than a political community of sovereign states. With regard to transnational networks in Europe, numerous programs and initiatives have been launched with the specific goal of opening new spatial
Figure 1  Supported transnational cooperation areas in the European Union, 2007–2013. Non-EU cooperation areas are indicative only and are subject to modification. This map includes Intellectual Property from European National Mapping and Cadastral Agencies and is licensed on behalf of these by EuroGeographics © EuroGeographics.
Figure 2 Cross-border cooperation program areas in Europe, 2007–2013. This map includes Intellectual Property from European National Mapping and Cadastral Agencies and is licensed on behalf of these by EuroGeographics © EuroGeographics.
CROSS-BORDER, TRANSNATIONAL, AND INTERREGIONAL COOPERATION

perspectives for cooperation between cities and regions in various areas of economic development and regional policy. These are generally targeted at specific development concerns that are at once national and regional but that serve to link very different regional spaces within the European Union. Providing incentives for the creation of new communities of interest within and between nation-states was intended not only to instill a sense of European identity, but also to diffuse innovations in economic promotion, job creation schemes, and revitalization strategies, among other areas.

In the case of the US–Mexican border, local and regional cross-border trade, as well as the development of export-oriented industrial complexes (maquiladoras), helped fuel the growth of binational urban economies and development problems out of which developed the cross-border cooperation institutions now in place. One indicator of the intensity of interaction is the development of binational coalitions and advocacy groups that have striven to maintain and/or improve conditions for bilateral trade investment but also environmental sustainability. Similarly, a plethora of business development initiatives (including innovative schemes for industrial incubators) have materialized in US–Mexican border cities during the 1980s and 1990s (Peña 2007). In contrast to the European situation, the North American Free Trade Area (NAFTA) project has pursued a logic of limited integration that recognizes, for example, the functional interdependence of border cities and regions but that does not envisage a “borderless” North America nor a comprehensive policymaking process at the supranational level – at least not within the foreseeable future. The reasons for this reflect basic differences between European and North American approaches to interstate integration. Perhaps most importantly, while sovereignty issues appear to have lost much of their divisiveness in Europe, they continue to block attempts at broader political and sociocultural cooperation in the North American context. Fears of losing local and regional decision-making abilities as a result of the adoption of supranational regulations regarding commerce, trade, and environmental protection remain one of the largest sources of resistance to further integration within the Americas. Furthermore, fears of undocumented immigration and drugs trafficking, as well as terrorist threats, have made the US–Mexican border a defensive, heavily guarded demarcation line. Since the September 11 attacks of 2001, CBC along both the US–Mexican and the US–Canadian border has, unfortunately, lost considerable impetus as well as research attention.

Development of the field

The research field of CBC has expanded rapidly since the 1980s and the focus has shifted since the 1980s from empirical research on transnational urban networks and their cooperation mechanisms to the study of local and regional forms of policy-relevant cross-border interaction. In addition, since the 1990s CBC research has expanded to other areas of the world. African, Asian, and Latin American examples give evidence of interest in local and regional forms of cross-border planning coordination, development, and rapprochement. The African Union Border Programme (AUBP), for example, has attempted to provide a policy instrument that encourages CBC and in this way transcend barriers created by colonial borders (Asiwaju 2005). However, while border studies is genuinely international, the principal focus of CBC research – at least in human geography – has shifted to Europe since the 1990s. For this reason, this research has been conditioned by
CROSS-BORDER, TRANSNATIONAL, AND INTERREGIONAL COOPERATION

the experiences of European integration and of EU policy contexts. There is thus, admittedly, a strong Eurocentric character to CBC research which might limit its international comparability.

As the research field has developed and expanded, so have the theoretical approaches that characterize it. This is in part a result of the development of associated research fields, border studies in particular. The focus on local cross-border cooperation processes has become more differentiated thanks to tensions between, but also to an interlinking of, structural, normative-functional, and contextual understandings of CBC. To begin with, a structural perspective on CBC assumes that changes in territorial politics are a result of changes in underlying economic relationships within the world-system. CBC is thus a logical outcome of globalization pressures on states, regions, and local communities. With advanced capitalism and neoliberal ideologies weakening the role of the state and its redistributive policies, CBC follows the path of entrepreneurial local development already broken by cities and regions more generally. Normative and functionally oriented perspectives generally operate from the assumption that, while structural conditions are the principal driver of political change, permeable borders emerge from a rationality that is conditioned by a need to manage sociopolitical problems. CBC is thus inevitable and its relative performance is of central importance. The significance of transnational regionalism is determined, in other words, by governance and addressing in a politically meaningful way local and regional concerns that transcend traditional national problem-solving capacities. Building on the conceptual foundations of subnational paradiplomacy, the normative-functional view offers a specific focus on cross-border policy integration as a form of multilevel governance (Perkmann 2007). Prospects for transboundary region building are thus defined by the outcomes of a gradual and complex process of institutional innovation and capacity building at national, state, and local levels (Scott 2000).

CBC can indeed be interpreted as a form of place-making or region building in which borders are used as strategic resources. However, one major weakness in understanding CBC as a new border-spanning form of governance is a frequent neglect of sociopolitical and historical contexts. CBC experience indicates that induced and institutionally thick Euroregions cannot by themselves lead to a transcending of borders; historical experience and memories of past conflicts continue to affect cooperation outcomes, as the examples of the German–Czech and German–Polish border regions testify (Balogh 2014; Mirwaldt 2010). Partly as a response to the limitations of the governance approach, borders scholars, especially since the turn of the millennium, have elaborated constructivist understandings of CBC as a contested regional development project. Indeed, the normative political language of integration often contrasts with local realities. Rather than represent a new governance ideal, cross-border cooperation often reflects competing territorial logics at the supranational (e.g., EU), national, regional, and local levels and conflicting attitudes toward more open borders (Popescu 2008).

Links to border studies

Research on cross-border cooperation has been closely linked to the study of borders themselves. If the initial phase of de-bordering euphoria that had already begun in the 1980s tended to de-emphasize the actual sociopolitical significance of borders, CBC research – with its close ties to border studies and political geography,
has since developed a more contextually sensitive understanding of the nature of borders themselves. In common understanding, borders result from sociopolitical processes of ordering. Borders, however, also reflect the cultural and symbolic construction of societies at a more general level and as such play an important role in framing and regulating social relations, as well as setting conditions for local and regional development. As a result, borders have been used as explicit symbols of European integration, political community, shared values, and hence identity by very different actors (Perkmann 2007; Popescu 2008). CBC has thus been related to not only the functions but also the symbolism of state borders as a spatial organization of difference.

Similarly, local CBC has been closely linked to the concept of borderland. In traditional academic debate, the term “borderlands” has generally referred to a space of transition between (national) societies (Rumley and Minghi 1991). The processes that contribute to borderland formation operate at different levels and involve a dialectic relationship between local societies and territorial spaces defined by borders. On this view, processes of CBC are not merely a result of more porous borders but are themselves active creators of borderlands and thus of transitions, both physical and cognitive, between different political spaces. Borderlands are formed through processes of cross-border regionalization at different levels and in different realms of agency: cross-border cooperation and political projects of place-making, as well as everyday economic, social, family, and cultural practices of transnationalism. Unlike classical notions of frontier as a regional situation that precedes the consolidation of state territoriability within formal borders, a borderland can also describe an area that closely reflects the physical, political, and social impacts of state borders.

CBC and European integration

CBC research is also linked to the analysis of European integration and cohesion. CBC, understood in terms of cross-border local and regional development, forms part of a rescaling of state–local relationships and the political geography of the state that has been set in motion by EU policies. As mentioned earlier, CBC began as a subnational political project in Europe in the 1950s. Since 1989 it has burgeoned as a result of both local initiatives and orchestrated networking strategies supported by EU programs. According to O’Dowd (2002), CBC has developed in Europe as a function of shifting state formations and changing border regimes. O’Dowd has also indicated that, as part of integration and enlargement logics, European borders have been being reframed in terms of their (often conflicting) significance as barriers, bridges, resources, and symbols of identity, and how these reconfigurations relate to the project of European integration and enlargement.

Here as well, the borderlands concept has served to conceptualize the sociospatial changes generated by European integration. This is most clearly the case with Euroregions, which were pioneered and developed as locally based cooperation initiatives in Dutch–German border regions as early as the 1960s (Perkmann 2007). The Euregio, formally established as the first Euroregion in 1965, was the product of postwar rapprochement between Dutch and German municipalities. Since then, Euroregions have become a part of complex regional policy networks, particularly in west European contexts. Indeed, the Dutch–German Euregio, a Euro-region with its own local council and close ties to German and Dutch state agencies, has served as a model of sorts for the development of border region associations. Having thus graduated from local and/or grassroots beginnings to the level of
a more or less coherent policy area in the late 1980s, transboundary cooperation and border regions development are now largely determined by attempts to promote the goal of political and economic cohesion.

The European Union has also had an important impact on the nature of cross-border relations in Eastern and Central Europe (Zhurzhenko 2010; Scott 2006). In preparing Central and East European countries for membership, the European Union adopted a strategy based on institutionalized CBC and aimed at a gradual lessening of the barrier function of national borders. These policies have also been aimed at integrating previously divided border regions in order to build a more cohesive European space. Nevertheless, Gabriel Popescu (2011) has argued that this normative political language of Europeanization (e.g., as a process of de-bordering regional development) often contrasts with local realities where CBC reflects competing territorial logics at the EU, national, regional, and local levels, and conflicting attitudes toward more open borders.

Current major emphases

Contemporary research on CBC focuses on cross-border cooperation as a form of contingent territorial politics which reflects different and often conflicting uses of borders. On the one hand, CBC has been assessed as a regional development project that utilizes borders as resources, primarily through creating synergetic working relationships between public and private actors. On the other hand, however, CBC has also been investigated with a view to border management and securitization policies that openly challenge cooperation agendas.

Exhaustive appraisals of the results of cross-border cooperation in Europe are difficult due to the vast number of border region initiatives either completed or in realization. However, even if the promotion of a sense of cross-border regionality through common institutions has received ample support, in practice institutionalization patterns have been uneven – both in terms of governance capacities and in their performance of actual cooperation. In well-organized border regions (e.g., the Dutch–German Euroregions), public sector and NGO cooperation has been productive in many areas, especially in relation to environmental protection, local services, and cultural activities. In less successful cases, such as the Hungarian–Slovakian border, cross-border projects have often merely served to enhance local budgets without stimulating true cooperation. Generally speaking, it has also been very difficult to stimulate private sector investments as part of cross-border regional development projects. Rarely has CBC produced rapid results in terms of economic growth and regional development. It is a process that can only produce long-term results in these areas. On the other hand, CBC is of considerable symbolic significance as a de-bordering and trust-building exercise, and it has also contributed to the participation of local governments in EU policy implementation processes.

In a more critical vein, contemporary research on CBC has drawn attention to problems of reconciling border security prerogatives with the aim of cross-border cooperation. Since the terrorist attacks of 2001 in the United States, and of 2004 and 2005 in Spain and the United Kingdom, respectively, it has become difficult to discuss CBC without considering border management issues. Contemporary border studies, for example, focus both on border management (as control and confirmation) and border crossing (as cooperation, contestation, and transcendence) as parallel and simultaneous processes. The crossing and control of borders
CROSS-BORDER, TRANSNATIONAL, AND INTERREGIONAL COOPERATION

compete with each other for hegemony: open and more flexible borders are vital for economic reasons, while tighter and more closed borders are seen as important security measures. It is a delicate balance which, in recent years, has in favor of the securitization proponents because of the emphasis on issues such as personal and physical safety against cross-border threats. The securitization of borders (i.e., the simultaneous erection of administrative and physical obstacles to control migrations) is an attempt not to close space and territories but to filter and sort transnational flows according to their degree of desirability (Jones 2012).

Security issues have seriously curtailed attempts at local CBC in the North American context. However, we are also witnessing a strong dose of political realism creeping into European discussion of borders and cross-border cooperation. Hardening external borders and an increasing EU emphasis on the territoriality of political community are challenging the spirit and raison d’être of CBC. Borders as expressed by visa regimes, citizenship, residence rights, and new forms of physical control of the European Union’s internal and external frontiers gives evidence of the creation of new categories of sociocultural distinction – and thus of contested and partly dividing borders. Furthermore, while the European Union expresses a desire to avoid recreating political divisions, new visa regimes and other restrictions of cross-border interaction threaten to exacerbate development gaps between the European Union and its neighbors.

Future perspectives

Transcending boundaries – at least rhetorically – remains a leitmotif of a postnational understanding of political agency. And yet, more than two decades after the end of the Cold War, borders themselves often remain an obstacle to greater political and social interaction even if the defensive character of many international borders has diminished. What CBC research has often neglected in the past is the changing nature of the power of borders. If anything has become clear in comparing different attempts at cross-border region building, it is that these forms of cooperation are inherently a process of sociopolitical construction and cultural negotiation. Borders not only have different meanings for different actors but are a manifestation of power relations within and between societies at different scales. Contextual and sociocultural issues will therefore most likely dominate future CBC research agendas. Along with a need to further explore the possibilities for creating postnational political spaces, ethical aspects of borders require greater research attention. Finally, a major challenge will lie in attenuating the dominant eurocentrism of CBC research through more inclusive international comparisons.

SEE ALSO: Borders, boundaries, and borderlands; National and regional integration; Political geography; Regional definition and classification; Regional economic integration; Regional geography; Regional political movements

References


Further reading


Cross-cultural research

Tracey Skelton
National University of Singapore

Human geographers have a particular focus on people, space, place, and spatiality. Some geographers, not all of them self-defined as cultural geographers, conduct research projects with people that focus on their cultural identities, practices, and meanings. These identities and practices are located in places and spaces, but they are fluid and dynamic, not necessarily bounded and fixed within those locales. Many geographers conduct research in locations different from those of their own cultural milieu and across spatial boundaries. This is defined as conducting cross-cultural research.

Since the 1990s, books on research methodologies in human geography have usually included a chapter that engages with cross-cultural research. Geographers have produced academic papers that either name cross-cultural research as their explicit theme or imply it through terms such as difference and diversity, insider/outsider positionalities, ethical approaches toward “others,” cultures of knowledge production, crossing boundaries. So there is apparently more cross-cultural geographical research taking place than is declared as such. A definitive definition of cross-cultural research can, therefore, be difficult to pin down. Nevertheless, as the body of work on cross-cultural research expands, albeit more slowly in the past decade than in the 1990s and early 2000s, it has become possible to filter out key elements of what geographers consider to be cross-cultural research (more details below). This expansion also allows us insight into the complexities and considerations of this kind of research.

In the early 1980s, observations were made and concerns raised about human geography continuing to pursue positivist research methods, attempting to emulate the physical sciences in research design and the use of particular methodological approaches, notably quantitative methods. Human geographers argued that unless human geography’s dissidents challenged the discipline’s foundational assumptions and normativities there could not be a re-examination and change of research methodologies (Hodge and Lester 2006). Such changes were necessary if the complex, messy, and dynamic geographies of people’s everyday lives were to be central in human geography research. Post-1980 challenges within geography began to examine and change processes and procedures of research. Key elements of the challenges and pushes for change were found in feminist geography, debates about ethics in geography, arguments over quantitative versus qualitative methodologies, and research in cross-cultural contexts. So from the early 1980s onwards, cross-cultural researchers have been working to challenge normative and positivist research methodologies and practices and move toward more inclusive, participatory, holistic, and decolonized approaches. Within this collective, feminist geographers and geographers working with indigenous people have been very influential in pushing the boundaries of inclusive, open, and equitable research adopting reflexivity and coproducing knowledge. Using culture
as a means to designate the value accorded to difference, meaning, and particularity has been a tactic of cross-cultural geographers to disturb and question notions of objectivity that have been embedded in geography’s interpretations of fieldwork or quantitative methods (Barnett 1998).

While there are complex issues at play in geography as to the meaning and theoretical and conceptual value of the term “culture,” there are geographers committed to practicing and working through the complex and often difficult processes of cross-cultural research. Such research, at a general level, is designed to challenge the power relations embedded in research projects, tackle the difficulties posed by crossing cultures, and work to create research which is more than the value of its publishing status. Most cross-cultural geographers aim to produce warranted and rigorous scholarship but also something that has value and meaning for the people they work with. Cross-cultural researchers face many dilemmas and one of them is the issue of representation and what outputs are produced by the research. Cross-cultural research and its advocates ask important questions of the academy and about the production of academic work, such as what is the value of the research and for whom.

This entry addresses both the processes and practices of cross-cultural research, specifically the methodological doing of this type of research. It also examines the range of dilemmas researchers and their co-researchers/participants have to work through in order to achieve effective, equitable, and empowering research. First, this entry considers what cross-cultural research is and some of the theoretical and political challenges involved. Second, it examines the more specific and practical aspects of research approaches and the methods used. Third, the entry explores the politics, tensions, and rewards of representation and research outputs.

What is cross-cultural research?

An academic journal, called Cross-Cultural Research, publishes articles that describe cross-cultural and comparative studies in all human sciences (geography is not listed). It supports and encourages interdisciplinary comparative research with the objective of establishing scientifically derived generalizations about human behavior. This is not the type of cross-cultural research conducted by geographers and indeed could be one of the reasons why many geographers who conduct cross-cultural research do not call themselves cross-cultural researchers. For them their research is about understanding and interpreting details and specifics of a particular group or set of cultural practices, not establishing generalizations. The cross-cultural element is more about a way of doing research than establishing comparative generalizations on human behavior. Geographers’ cross-cultural research also has to consider epistemological and theoretical challenges, such as the definition of culture.

Don Mitchell’s 1995 publication stated that there is no such thing as culture in order to call for a reconceptualization of culture within “new cultural geography.” He argued that cultural geography reified culture and accorded it an ontological and explanatory status that it did not warrant. For Mitchell, the most important aspect of culture for geographers was the material development of the idea of culture and how that idea of culture was used to explain material differences, social order, and relations of power. He also argued that culture is used to “order, control and define ‘others’ in the name of power or profit” (Mitchell 1995, 104). His argument is complex, compelling, and controversial and this entry cannot do it full justice. However, it highlights one of the theoretical complexities that have to be considered as part of the context of cross-cultural research. If there is no such
thing as culture, then what are we crossing in our research?

If we consider the situation of indigenous people in many countries of the world then we can see that defining them as a cultural group was part of the mechanics of their marginalization, exclusion, and denigration in relation to wider non-indigenous societies. Parceling groups of people up into distinct cultures then proves that cultures exist, the concept of culture is made real and we believe in those differences between people (of course this holds echoes with the negative power relations of racial classification and differentiation). Historically such cultural categorizations have been dangerous and established power relations of “us” and “them/others.” Researchers became classified as insiders or outsiders; most of them “entering a culture” in order to study and, in the past, classify and categorize it. However, at the same time, for indigenous groups and other people who define themselves as having a cultural identity, the notion of distinctive cultures can be used in resistive ways as it has political potential to challenge negative power relations and cultural subordination. For Welsh-speaking people the assertion that their language was an essential part of their culture allowed them to bid for funding from the UK government and the European Union in order to “save” their linguistic cultural identity. In the context of Australian Aboriginal people and Torres Strait islanders, assertions of tribal cultural ancestry and traditional practices have accorded them some rights in relation to land access, caretaking, and consultation on development and planning initiatives. These rights have been hard won and there are many battles still being fought; not only the implementation of these legal rights, but also the designation of a cultural identity and its link to particular spaces and places have been very important to their struggles.

Hence, on the one hand we have an academic argument that there is no such thing as culture, and it is a strong, theoretically and politically important intervention. On the other hand we have considerable evidence that in practice people believe that they belong to a culture and that this often has a geographically defined spatiality. Cross-cultural researchers work with this latter formation of “culture,” but there are caveats. Most cross-cultural researchers do not see cultures as fixed, unchanging, and spatially bound. They do not consider their co-researchers or project participants as “others” but as knowledge producers. They recognize that they are culturally laden and that, for many, their Western culture means they construct their worlds, their identities, meanings, and practices in different ways to the people they work with. Cross-cultural research is designed to be reflective and the researcher’s positionality vis-à-vis the research participants is interrogated and made explicit. Although the value of some aspects of positionality has been challenged by arguments to attend more to how personality affects field research and the production of knowledge (Moser 2008). Cross-cultural researchers place substantial emphasis on field-based research that involves respectful attention, listening and engaging with co-researchers, and on choosing or devising methods that allow such research engagements to take place.

Cross-cultural researchers are therefore aware that they are working with a problematic (intellectually and politically) concept and that culture does quite literally mean different things to different people. Yet, it remains a concept to be worked with and through and certainly has a valence for many people in the world, often those most marginalized economically, politically, socially, and additionally culturally categorized in powerful but negative ways. A key political project for cross-cultural researchers linked with
the notion of cultural politics is to challenge the traditional modes of engagement as practiced in the past by geographers and anthropologists. Such approaches have been termed “colonial research,” which was the dominant mode of research in cross-cultural research in geography, and subsequent approaches that resist these highly problematic and questionable research practices are known as postcolonial and decolonizing research (Howitt and Stevens 2005).

Colonial research approaches to cross-cultural fieldwork embody power relations with “others” that reinforce the domination and exploitation of such groups. It tends to dismiss researched subjects’ own cultural and traditional knowledge and practices and uses intrusive top-down research methods (note that the present tense is used here because there are many examples of this kind of field research still being conducted). Postcolonial approaches act in reaction against and rejection of the above negative practices and work toward the self-determination and welfare of the research participants. This is done through the use of methodologies that enable the production of research findings that value the particular cultural community/ies involved in the research. Their rights, knowledge, perspectives, concerns, and aspirations are recognized, determined as valid, and, where possible, supported by the research. Decolonizing research pushes the politics of openness and equality further as the actual research processes and practices are designed to interrogate and deconstruct cross-cultural “discourses, asymmetrical power relationships, representations, and political, economic, and social structures through which colonialism and neo-colonialism are constructed and maintained” (Howitt and Stevens 2005, 32). Decolonizing researchers (who are often heavily influenced by feminist geography critiques of research practice) will consider their research participants to be co-researchers as all cross-cultural research is in effect a coproduction of knowledge (see Maclean 2015 for an excellent example of decolonized coproduction of knowledge in Australia). Most geographers working cross-culturally pursue one of these three research approaches with, increasingly, the last two or a combination of both in the ascendency.

Practicing cross-cultural research

Human geographers’ writings and debates about practicing cross-cultural research demonstrate a wide range of themes to be considered. Many of these are also shared with other research practices and methods, but they are often more visible and demand more care and attention in cross-cultural research. On an everyday basis the researcher and the cultural group they are working with may have different conceptualizations of reciprocity, personal space, time, socially acceptable behavior and speech, gift giving, hospitality, privacy, bodily comportment, and so forth. These are frustrating but intriguing and part of the co-learning that takes place in open, inclusive, and decolonizing research. Researchers often describe the ways in which they had to learn and present a different way of being, adapt their research processes in situ to make sure they worked effectively and inclusively, and adopt new ways of engaging with people. For example, first and subsequent meeting protocols, what is said and done when and how, can be very intricate in different cultures.

Ethical practices are relevant to all research with human participants. Many argue that they are particularly important in cross-cultural contexts, especially if trying to practice decolonized research. Ethics should be integrated at all stages of the research: planning, design, implementation, and production. Ethics are culturally constructed. Ethical protocols for cross-cultural
Cross-cultural research are more culturally aligned with the researcher than with the project participants. What is defined as ethical practice in one culture and place may not be so ethical or appropriate in another. Often an immediate problematic appears in relation to written participant information sheets and the collection of signed consent forms. This happens when a researcher is working with illiterate people or with people for whom signing a form is an anathema of human interaction or has been a tool of deceit in the past as people signed away land or rights without understanding the consequences. Here the meaning and sentiment of ethical practice comes to the fore and other culturally appropriate mechanisms of informed consent should be used. A key ethical principle in Western research protocols is privacy and confidentiality. Interviews should be conducted in private, away from family, the household, or the community. In some cultures, particularly those based on collective principles, such behavior is considered rude or suspicious. Such “strange” behavior can serve to reinforce notions of a researcher being potentially dangerous, unpredictable, and self-centered in relation to their own research project rather than working collaboratively with their participants. Cross-cultural researchers therefore have to tread carefully (sometimes literally in the context of sacred spaces) to learn from those they work with as to what is ethically acceptable for them and their community. Knowing and understanding as much as possible about the cultural, ethical, and appropriate ways to behave, communicate, and interact with different communities is extremely important. A key ethical principle increasingly demanded by groups that find themselves the “object” of research is how and in what way the project will benefit them and their communities. What is the value in it for them? Who benefits most and least once the project is completed? In decolonized, ethically practiced research the community needs to be met and cross-cultural geographers have to work to build ethical and equitable relationships with them. This is particularly true in working with indigenous people who have frequently been exploited, denigrated, and damaged by Western-style colonial research projects (Louis 2007).

Discussions of cross-cultural research are linked with a presumed binary of researcher status, that is, insider/outsider. In fact, researchers report that they are often in some slippery, interstitial space between the two; the term “betweenness” is often used as researchers differentially connect or disconnect with members of the community. Such relationships also vary across time – first encounters can often mean and feel very different from final partings, and possibly future returns; many cross-cultural researchers participate in longitudinal research and the insider/outside binary becomes meaningless. However, being perceived as an outsider can lead to very rich research as participants provide more information to ensure the researcher really understands the community well and there is no presumption of prior knowledge. Additionally, in communities where gossip is a key aspect of social control, especially of women, then an outsider (female) researcher may be tested, trusted, and, subsequently, given extremely valuable information not usually shared with an “insider” researcher (Skelton 2001). On the other hand, “insider” researchers may be seen as more empathetic, trustworthy, and having the best interests of the community in mind throughout the research.

Cross-cultural researchers discuss a second binary that relates closely to the identity, positionality, and personality of the researcher. This is the concept and practice of same-ness/difference. This is more about the interrelationships that are established between research participants/co-researchers and the researcher/s than about the association or...
CROSS-CULTURAL RESEARCH

disassociation between these two parties. It links to cross-cultural methodology where researchers consciously acknowledge the differences between themselves and co-researchers, consider the consequences of it, and incorporate it into their research practice. Interpreters and/or locally based researchers may be employed to establish some linkage of sameness. Researchers might emphasize their “difference” in order to connect with participants as someone who genuinely wants to learn and understand and work through and beyond the differences. The last approach is often used as a technique when working with elites and people in authority. More usual though is the work that researchers do to find a basis of sameness and commonality with the people they work with. There might be a connectivity found that transcends difference that allows the co-creation of an affinity or alliance which smooths cultural and research exchanges. Feminist geographers have reported using gender, motherhood, and sexuality as points of connection. These elements of sameness may counterbalance differences that cannot be connected to so easily or effectively such as “race,” religious affiliation, or age. However, sameness and difference are not fixed: as part of culture they are fluid and dynamic. Different researchers will find different points of connection and over time differences can be overcome as the actors in the research project spend time and communicate together. What researchers and co-researchers involved in decolonized and feminist influenced approaches often report is that there are reciprocal exchanges taking place in the search for commonalities that allow for shared spaces of similarity from which meanings and understandings can emerge.

An important facet of cross-cultural research practice is interpretation and translation. All forms of cross-cultural research require some degree of interpretation in order to be able to research and talk across differences. Indeed, feminist geographers have discussed the need for interpretation within the academy and across worlds, questioning whether it is possible to talk in meaningful ways about research and publishing in different parts of the world based on their actual cross-cultural research experiences (Dyck, Lynam, and Anderson 1995; Raju 2002; Staeheli and Nagar 2002). In terms of the practices of research, often written tools of research have to be translated; for example, questionnaires have to maintain the same meaning in different cultural contexts. Carefully translated research questions are essential in the early stages of the research, ideally in situ so there can be a discussion about the most appropriate terms that capture the essence of what the research is about. A project on the meaning of “wealth” for a community will provide completely different data if that is translated as “money.” Interpreters and translators (hereafter just called interpreters) hold interesting and complex positions within a research project and can be both insider and outsider. They may differ from the community because of education and social class but have an affinity through family connections. For geographers working cross-culturally there are examples where interpreters have been indispensable to the research project and have enhanced all manner of data collection. Effective processes of translation/interpretation create new spaces of exchange and learning where all research actors develop a cross-cultural relationship around the subject area of the research. New meanings can be created as “in-between” forms of understanding. However, there are also incidents of problems and major clashes with interpreters who will not take the time to work through the finer nuances of linguistic meaning, summarize questions rather than ask them in full, lead participants into particular answers because they “know” what the researcher is
“looking for,” and convey their own prejudices into the research, such as refusing to interpret between a researcher and lower class/different caste/migrant people (Twyman, Morrison, and Sporton 1999). Cross-cultural researchers have to accept that there will always be areas of incomprehensibility, knowledges kept hidden and not shared, and elements that can never be translated.

Politics, tensions, and rewards of representation and research production

Some aspects of politics in cross-cultural research have been discussed above. This final section considers the political power of questioning established forms of authority in the production of knowledge combined with a discussion of the power and politics of representation.

If cross-cultural researchers adopt the decolonizing approach then rich, new, and constructive knowledges will emerge that may counter perceived prejudices, stereotypes, and existing power–knowledge relations. It is important to examine the power relations, inequalities, and injustices at play in communities and to explore the ways in which cross-cultural research can challenge these and bring about change. Such approaches are influenced by feminism and postcolonialism. There must always be questions about the impacts of the research project on those involved, what the research is for, and what will be the value for the research participants. Knowledge production for the sake of knowledge production is no longer considered a valid approach and nor is the presumption that it is the researcher/academic that produces the knowledge. All research is coproducive and cross-cultural research in particular has to acknowledge this in order to empower and fairly collaborate with their “co-researchers.”

Methodologies have to be used which integrate co-researchers’ voices and knowledge. Renee Louis (2007, 131) makes a very explicit call for the use of indigenous methodologies which state that research in indigenous communities is conducted respectfully, from an indigenous point of view, and contributes in some way to the betterment of the community. Such political and methodological approaches within cross-cultural research can establish conversations, connections, and knowledge sharing that is equitable and reciprocal and works to not reinforce Western/northern/colonial authority. Through such culturally and politically aware approaches, spaces and discourses of progressive politics can be established. Such methodological approaches and the ensuing research production (reports, papers, presentations, datasets, advocacy, testimonials) can challenge academic authority and make space for cross-cultural and spatial differences to establish alternative interpretations of practices and power.

In academic research and production there are always tensions. Cross-cultural researchers have wrangled with difficult issues of representation, in particular misrepresentation that emerges from misunderstanding and mistranslation. Careful thinking through of how research findings are produced and reported is essential because much is at stake when a group of people or a place is researched, spoken, and written about. Researchers are part of the production of people and spaces and can be responsible for misrepresentation. It is important that they do not appropriate coproduced knowledge as their own. Continual questioning about what is produced, for whom it is produced, and how the knowledge can be best represented is required. In order to follow through the ethical research practice of “do no harm,” researchers have to avoid and actively counter the reinforcement and perpetuation of certain cultural stereotypes.
CROSS-CULTURAL RESEARCH

(misrepresentations) – this must be done at all stages of the research from interim reports to final publications.

Geographers working with indigenous people provide some of the best examples of this kind of political sensitivity, avoidance of misrepresentation, and recognition of the coproduction of knowledge. Kirsten Maclean and the Bara Yarralji Bubu Incorporation conducted a collaborative research project that enabled an indigenous group in Australia to articulate their knowledge and expertise of water governance on their traditional country. Government engagement with indigenous people is often not very successful mostly because of culturally inappropriate approaches combined with negative stereotypes about the indigenous groups. Through sympathetically and scientifically “crossing boundaries” all researchers in the project demonstrated their learned and practiced knowledge and produced a research report. The co-authors call this a “boundary object” because the report can translate indigenous people’s knowledge, values, and management interests in a form that maintains their cultural integrity and is easily accessed and interpreted by non-indigenous planners and scientists (Maclean 2015, 144).

Future directions

It appears to becoming less common to identify geographers’ work specifically as “cross-cultural,” and there are indeed questions being raised about the value of such a term. In a highly mobile and globalized world it is questionable if there are still distinct cultures to cross into. Cross-cultural researchers would argue not, because no cultures are fixed, completely distinct, or bounded. However, these are important questions to raise and to keep on raising. Cross-cultural researchers have been grappling with the complexities, tensions, and frustrations of working across cultures in ways that have changed for the better: decolonial rather than colonial, reflective, based on principles of coproduction, equitable, rigorous, and effective for researchers and participants. Geographers crossing boundaries are producing “boundary objects” that are transparent and produce legitimate information. All stakeholders (those within or without a research project) can access and understand the information. Researchers acting as “boundary agents” help to communicate the knowledge such objects carry, guarantee their validity, and ensure the recognition of coproduction.

In 1999 Linda Tuhiwai Smith published Decolonizing Methodologies: Research and Indigenous People. While she was focusing on indigenous people her work is highly relevant for all groups of people who have experienced colonization, injustice, and marginalization on the basis of their culture. Her book is a critique of previous colonial research practices and a presentation of indigenous methodologies. She stated: “The spaces within the [dominant] research domain through which indigenous research can operate are small spaces on shifting ground. Negotiating and transforming institutional practices and research frameworks is as significant as the carrying out of the actual research programmes” (Tuhiwai Smith 1999, 140). More than 15 years on many cross-cultural researchers and geographers crossing boundaries are still working on such transformations, challenging institutional practices, reorienting research approaches, and coproducing knowledge that is meaningful for the co-researchers and the communities – step by respectful step.

SEE ALSO: Cultural geography; Culture; Difference; Ethics in geography fieldwork; Ethnicity; Everyday geographies; Feminist
geography; Feminist methodologies; Fieldwork in human geography; Friendship, geographies of; Geographies of exclusion; Identity; Indigeneity; Indigenous knowledge; Inequality; Multiculturalism; Positionality; Postcolonial geographies; Power; Reflexivity; Representation; Social justice

References


The frozen realm of the cryosphere consists of snow, sea ice, mountain glaciers, ice sheets, and perennially and seasonally frozen ground. From the polar regions to the tropical mountains, cryospheric elements are found in any location where there are prolonged freezing temperatures and frozen water. The cryosphere plays a fundamental role in Earth’s climate as it responds to changes in temperature and precipitation. Through its influence on Earth’s energy budget, the cryosphere itself drives changes in climate from local to hemispheric spatial scales and from weekly to millennial time scales. Due to the vast and remote expanses covered with snow and ice, the direct measurements needed to characterize cryosphere changes are not attainable. Satellite and airborne remote sensing platforms are, therefore, indispensable tools for mapping, characterizing, and monitoring the cryosphere in all parts of the globe.

Satellite remote sensing of the cryosphere dates back to the early 1960s when US photographic reconnaissance satellites (such as CORONA) used high-definition photography to image regions around the globe, including the Greenland and Antarctic ice sheets, permafrost and sea ice regions in the Arctic. Declassified in 1995, data from the 1960–1972 CORONA satellite missions has been used for cryosphere studies involving change detection (Bindschadler and Vornberger 1998). Unclassified, near–real-time orbital imagery became available starting in 1966 with Nimbus-2, allowing researchers to map snow cover and sea ice extent by examining spatial differences in visible band reflectance. That year also began the start of one of the longest remote sensing time series: snow cover extent. The National Oceanic and Atmospheric Administration (NOAA) used manual interpretation of visible imagery to produce weekly charts of snow extent and produced these weekly charts until 1999. It subsequently turned to an automated daily mapping system that used multiple sensors including passive microwave data, which can map the surface even in darkness and beneath clouds. Active microwave remote sensing of snow and ice began with the 1978 launch of SEASAT-SAR, the first synthetic aperture radar (SAR) instrument. Passive microwave remote sensing of snow began in 1978 with data from the scanning multichannel microwave radiometer (SMMR) on board the Nimbus-7 satellite. The earliest sea ice mapping began with the Nimbus-5 electrically scanning microwave radiometer (ESMR) from 1972 to 1976. Daily polar sea ice mapping with space-borne passive microwave began in 1978 and continues today. These satellite measurements of snow and sea ice are the longest cryospheric environmental data records.

### Electromagnetic radiation and definition of key terms

Before proceeding with an overview of remote sensing of the cryosphere, it is helpful to cover some of the basic, relevant remote sensing terminology.
Remote sensing measures the interaction of energy with materials. There are three basic ways that energy can interact with matter: absorption, emission, and scattering. Absorption and scattering represent attenuation of energy by either conversion to thermal energy (such as with absorption of sunlight) or redirection of refracted/reflected energy (such as with scattering of sunlight or microwave radiation). Terms that are commonly encountered in remote sensing of the cryosphere include:

- **Albedo**: the albedo of a material is the amount of energy reflected relative to the energy incident upon it. Albedo can be a spectral quantity (spanning a narrow range of wavelengths) or a broadband quantity (spanning a wide range of wavelengths, usually across the solar spectrum from about 0.3 μm to 5.0 μm). Albedo and reflectance are not the same quantity. Reflectance, or, more specifically, the bidirectional reflectance, is what is measured by space-borne instruments. Reflectance depends on measurement geometry whereas albedo is independent of viewing geometry.

- **Brightness temperature**: in the microwave portion of the spectrum, satellite-observed brightness temperature ($T_B$) is linearly related to the emitted radiance of a material. Through a simplified relationship, $T_B$ can be defined as the product of the spectral emissivity of the material ($\varepsilon_\lambda$) and its thermodynamic temperature ($T$): $T_B = \varepsilon_\lambda T$.

- **Dielectric constant**: in the thermal and microwave portions of the spectrum, the dielectric constant is used to describe how energy interacts with snow. The dielectric constant, $\varepsilon$, is sometimes called the “relative permittivity.” Permittivity is a measure of a material’s ability to transmit an electric field and $\varepsilon$ is the ratio of the permittivity of the material to that of a vacuum. At 0°C, the dielectric constants of snow constituents are: air $\varepsilon_a = 1.0$, ice $\varepsilon_i = 3.2$, liquid water $\varepsilon_w = 81$.

- **Emissivity**: the ability of a material to emit energy. Materials that are good emitters are poor reflectors and vice versa. Emissivity is a function of the dielectric properties of the material and the shape and geometry of their arrangement. Dry snow has a very low emissivity whereas wet snow has high emissivity.

- **Radiometric resolution**: the quantization of brightness levels expressed as $2^n$-bits where $n$ is an integer (e.g., 8-bit resolution = $2^8 = 256$ brightness levels).

- **Spatial resolution**: the number of spectral bands and their spectral width.

- **Spectral resolution**: the minimum size of an object that can be detected and/or distinguished from another object in an image.

- **Temporal resolution**: the frequency with which remotely sensed data are acquired over an area.

In the following sections, the remote sensing capabilities are described for snow, ice sheets and glaciers, sea ice, and permafrost and frozen ground.

### Remote sensing of snow

Snowpack quantities that can be derived from remote sensing include snow-covered area (SCA), albedo, effective grain size/specific surface area, depth, snow water equivalent (SWE), and melt-thaw status. For each of these properties there are limitations, sometimes significant, in terms of measurement accuracy, spatial/temporal resolution, detectability, and length of data record.
Snow covered area

The unique spectral properties of snow in the visible and near-infrared wavelengths (Figure 1) make it readily distinguishable from other land cover types. Thus, SCA can be readily mapped using the difference between visible and near-infrared snow reflectance.

The normalized difference snow index (NDSI, equation 1) is the basis for the binary (snow/no snow) SCA product from the moderate resolution imaging spectroradiometer (MODIS).

\[
\text{NDSI} = \frac{\text{band 4} - \text{band 6}}{\text{band 4} + \text{band 6}} \quad (1)
\]

where MODIS band 4 corresponds to a wavelength of 0.55 μm and MODIS band 6 corresponds to a wavelength of 1.6 μm. Remote sensing of snow in vegetated areas is challenging and the MODIS snow product accounts for vegetation by adjusting the snow/no snow threshold using a vegetation index. This approach works well for a sparse-to-moderately dense canopy, although there is no algorithm that can map snow under dense forest canopy.

Fractional SCA (fSCA), is the proportion of snow cover in a pixel and is another widely used metric for snow-covered area. Highly accurate estimates of fSCA are possible using a spectral mixture analysis method that decomposes the pixel reflectance into the individual reflectance values of its components (e.g., snow, vegetation, soil/rock). Compared with binary snow mapping methods, fractional snow mapping works very well at times when snow cover is patchy, such as during the melt season.

Snow albedo

Snow albedo can be computed from multispectral measurements of spectral reflectance that are converted to spectrally integrated broadband albedo using regression coefficients derived from measurements. The spectral albedo of clean snow is controlled by grain size (Figure 1), with
coarse-grained snow having a lower albedo than fine-grained snow in the near-infrared wavelengths. Thus, grain size (also referred to as the optically-equivalent grain size or the specific surface area) can be derived from multispectral and hyperspectral remotely sensed data in the near-infrared wavelengths. Snow grain size can be used as input to a radiative transfer model from which snow albedo can be computed, assuming the snow is clean. However, for snow containing dust or soot particles, the visible albedo will be decreased by as much as half.

Snow depth

To date, accurate remote sensing of snow depth has only been accomplished with airborne and ground-based instruments (Figure 2). LiDAR remote sensing of snow depth uses a differential approach in which snow-free elevations are subtracted from snow-covered elevations (Deems, Painter, and Finnegan 2013). Positional accuracy is critical, especially in mountainous terrain where slight shifts in the x-y position can yield large snow depth discrepancies, especially on steep slopes. Thus, kinematic GPS is required to provide ground control points for horizontal and vertical accuracy and to orthorectify the resulting elevation surfaces. For example, NASA’s Airborne Snow Observatory (ASO) uses LiDAR and a hyperspectral sensor to retrieve snow depth, fractional snow covered area, snow albedo, and dust content as part of an
operational approach to improve water resources management in the California Sierra Nevada.

Ground-based LiDAR, also known as terrestrial laser scanning (TLS) uses 360° scanning laser instruments with distance ranges up to 4 km to compute watershed-scale snow depths with a spatial resolution and vertical accuracy of about 5 m and 10 cm, respectively. For maximum efficacy, TLS requires permanent reflectors at multiple locations in the study area and minimal obstruction by vegetation and topography.

Snowpack depth can also be inferred from frequency modulated continuous wave (FMCW) radar, but such measurements require ground-based or helicopter deployments and measurements need to be calibrated with field observations. This active microwave technique holds promise for mapping and characterizing snowpack stratigraphy, as the radar returns respond to grain size and density of individual snow layers.

**Snow water equivalent and melt status**

The natural microwave emissivity from the snow-free ground is attenuated by snow cover and this serves as the basis for passive microwave estimation of snow water equivalent. Passive microwave radiometers have been used to estimate SWE for nearly four decades. However, this measurement has relied on a simplistic empirical relationship between microwave brightness temperature at 19 GHz and 37 GHz with adjustments for the effects of forest cover, differences in snow density, and the presence of tundra lakes below the snowpack. The passive microwave signal reaches an asymptote at about 150 mm of SWE. Passive microwave measurements require a very large footprint because the signal is relatively weak. Determinations of SWE using passive microwave techniques are limited to dry snow because when liquid water is present in the snowpack, the snow itself becomes a strong microwave emitter and it is no longer possible to distinguish snow from snow-free areas. However, this snow property also means that passive microwave can be used to identify the onset of snowmelt.

Passive microwave snow observations have been a valuable addition to model-derived SWE. GlobSnow, from the European Space Agency, is a modeled SWE product that uses passive microwave and ground-based weather station data for the period 1979–2013. SnowDAS, from the NOAA National Weather Service’s National Operational Hydrologic Remote Sensing Center, is a gridded daily snow product (including SWE) derived from a data assimilation model that incorporates passive microwave and optical satellite data, airborne gamma surveys, and ground-based data.

In airborne campaigns, dual frequency Ku-band (12–18 GHz) radar has been moderately successful in measuring SWE (Yueh et al. 2009). As SWE increases, so does radar backscatter due to the larger path length and greater number of scattering elements in the snow volume. Both grain size and SWE affect radar backscatter but these can be solved independently with two Ku-band frequencies. However, the relationship between other snow properties and radar backscatter is complicated. For dry snow, factors that affect radar backscatter include snow grain size, snow density, and snowpack stratigraphy (layering, especially ice lenses and depth hoar). As with passive microwave, wet snow significantly modifies the radar backscatter signal. In limited cases, radar has been used to retrieve the liquid water content of the snowpack.

**Remote sensing of ice sheets and glaciers**

Glaciers and ice sheets store most of the fresh water on our planet. The ice sheets of Greenland and Antarctica are extensive glacial ice masses,
CRYOSPHERE: REMOTE SENSING

exceeding 50,000 km². Ice caps are smaller than ice sheets and consist of multiple, coalesced glaciers, such as Iceland’s Vatnajökull ice cap. Other types of glaciers include mountain glaciers (such as the Arolla glacier, in Switzerland), piedmont glaciers (such as the Malaspina glacier, in Alaska), ice fields (such as the Juneau ice field, in Alaska), and others. Key questions that can be answered using remote sensing include: How thick are the glaciers and ice sheets? What are their surface properties? Are they increasing or decreasing in mass? How fast are the glaciers moving?

Ice sheet surface properties

As with snow albedo, the albedo of the glaciers and ice sheets is critical for estimating surface energy balance. On both Greenland and Antarctica, albedo has been measured and monitored using multispectral sensors such as MODIS. Low contrast between ice and clouds makes cloud detection a major challenge over the bright, large expanses of Greenland and Antarctica. Atmospheric correction of cloud-free imagery can also be a challenge at high latitudes due to low sun angles limiting the applicability of atmospheric correction models. At the high elevation portions of the Greenland ice sheet, the dry snow zone, the broadband albedo typically exceeds 0.85. Albedo decreases to about 0.70 in the wet snow zone and in the bare ice zone exposed firn and bare ice have albedo values of about 0.40. Along the Greenland ice sheet margins, dark algae growing on the snow/ice surface during the summer season can lead to very low albedo – about 0.10–0.20.

Melt extent on the Greenland ice sheet has been mapped using gridded brightness temperatures from passive microwave sensors such as SMMR, the special sensor microwave/imager (SSM/I), and the special sensor microwave imager/sounder (SSMIS). Horizontally polarized 37 GHz brightness temperatures are used to discriminate nonmelting from melting snow in each 25-km gridded image (Figure 3).

Glacier zones such as the dry snow, wet snow, percolation, and bare ice zones have been mapped using albedo and surface roughness. Surface roughness is an expression of surface–atmosphere interactions as winds redistribute snowfall and carve patterns into the surface creating dunes, sastrugi, and wind crusts at scales from millimeters to several kilometers. Roughness is also affected by subglacial topography, creating large-scale (1–100 km) surface features, and by glacier flow, expressed in 1–100 m scale features such as crevasses. Roughness increases radar backscatter and SAR sensors with spatial resolutions of about 3–100 m (depending on scan mode) can detect fine scale surface features. Multi-angular visible reflectance from the multi-angle imaging spectroradiometer (MISR) has also been effective at mapping roughness at scales from 0.1 to 10 m.

Ice thickness

To quantify ice sheet and glacier thickness, scientists have developed airborne coherent radar depth sounding systems that transmit a radar pulse through the ice and measure its return time and characteristics. The pulse delay time has been empirically related to ice thickness to an accuracy of about 2%. Longer than usual radar wavelengths, radio frequencies at 50–500 MHz can penetrate several kilometers of ice and can also provide information on internal layers needed for ice flow modeling. Cryospheric scientists have also used ground-based and airborne radar depth sounding to map the ice–bedrock interface, identifying subglacial channels and preglacial topographic features.
Glacier velocity

Using interferometric synthetic aperture radar (InSAR), studies have been carried out to map ice surface velocity on ice streams and outlet glaciers. After removing the effects of topography, it is possible to compute the phase difference between two SAR images to map surface motion along the line of sight of the radar:

\[
\phi = \phi_1 - \phi_2 = \frac{4\pi}{\lambda} (R_1 - R_2)
\]

where, \(\phi_1, \phi_2\) are the radar phase of each of the two images, \(R_1\) and \(R_2\) are the range from the radar instrument to the ground element and \(\lambda\) is the wavelength. Ice motion must be relatively slow to prevent phase decorrelation between images. Ice velocities for the Greenland ice sheet have been mapped using InSAR data from RADARSAT-1 and the advanced land observation satellite (ALOS) and TerraSAR-X (Joughin et al. 2010). Combining ice velocity with ice thickness gives an estimate of the mass flux of water from ice sheet to ocean. A second method to determine ice surface motion involves cross-correlating glacier features that move with the ice (such as crevasses) and that

---

**Figure 3** Greenland melt from passive microwave brightness temperature. Reproduced from National Snow and Ice Data Center.
are visible in images from one acquisition time to another. This technique requires adequate spatial resolution and feature contrast for effective image-to-image cross-correlation. Landsat thematic mapper (TM) has been used for this purpose, measuring ice velocity on Antarctic ice streams and mountain glaciers.

Surface elevation and mass balance

Radar altimetry and laser altimetry measure the surface elevation of glaciers and ice sheets using the known altitude of the spacecraft and a range measurement from the instrument to the surface. Elevation changes, when combined with density estimates can be translated into changes in ice mass balance. Radar altimeters are nadir-pointing sensors that send a radar pulse to the surface below and measure its round-trip travel time and the returned waveform. Using the satellite altitude and applying corrections for water vapor and ionization, the time estimate is converted to an altimetric surface height. The footprint of space-borne radar altimeters is about 2 km, although this increases over rough surfaces due to increased travel time for the return waveform. Early radar altimeters such as SEASAT-ALT were not designed for ice altimetry but CryoSat-2, launched in 2010, is specific to polar ice surfaces with a high inclination orbit and is augmented by InSAR capabilities to map elevations of sloping and rough ice sheet surfaces. On board NASA’s ice, cloud, elevation satellite (ICESat) the geosciences laser altimeter system (GLAS) was a space-borne near-infrared (1.04 μm) laser altimeter that measured ice sheet elevation. With it 70-m footprint and 170-m spacing, GLAS measured elevations of ice sheet and ice shelf features with detail not previously available from satellite. Due to laser failures, the ICESat was decommissioned in 2010. In its stead, Operation IceBridge, a multi-instrument airborne science program is providing extraordinary 3-D perspectives on the polar regions in more detail than ever before.

One of the most innovative approaches to remote sensing of ice sheets is the gravity recovery and climate experiment (GRACE). Since 2002, GRACE has measured monthly changes in Earth’s gravitational field to estimate regional changes in mass over the ice sheets and large glacier regions (Luthke et al. 2006).

Glacier extent, terminus position, and snowline

With their steep elevation gradients, alpine glaciers are especially sensitive to changes in climate and have also been significant contributors to sea level rise in recent decades. As such, satellite remote sensing has been an important means of tracking glacier characteristics that can be used to estimate changes in glacier mass balance or used as a proxy for climate change. Glaciologists have developed automated approaches to map glacier extent and terminus position using multispectral imagery from sensors such as Landsat TM, enhanced thematic mapper (ETM+) and the advanced space-borne thermal emission and reflection radiometer (ASTER). In the near-infrared wavelengths such as 1.6 μm (ASTER band 3 or TM band 5), ice is very dark compared with non-ice features, including most clouds. In the visible wavelengths ice is very bright. Simple ratios using near-infrared-to-visible spectral combinations or the NDSI (equation 1) are effective for glacier mapping using either TM or ASTER. Combined with a digital elevation model, changes in glacier extent can also be used to estimate changes in glacier mass balance. Using satellite-derived glacier outlines, the global land ice measurements from space (GLIMS) program and the Randolph glacier inventory have mapped hundreds of
thousands of glaciers globally, providing new capabilities for monitoring glacier change. Debris cover on glaciers obscures the contrast between ice and surrounding bedrock and moraines, confounding glacier extent estimates. Limited research has shown that InSAR can map the motion of debris-covered glaciers though to date, there has not been a robust method for their regular mapping and monitoring.

Remote sensing of sea ice

Sea ice forms on the sea surface when the water is sufficiently cold to freeze. The complex atmospheric and ocean circulation processes that characterize the polar regions control its seasonal development and motion.

Sea ice area, extent, and concentration

Sea ice area is the total area of sea ice greater than about 15% concentration in a pixel. Extent is the geographic extent demarcated by the ice margin. Concentration is the fraction of sea ice in each remote sensing grid cell. Optical sensors can readily distinguish open water from sea ice but cloud cover and the dark of polar winter inhibit their views. Passive microwave remote sensing has an advantage of being able to monitor sea ice during periods of darkness and cloudiness. The brightness temperature of open ocean is low compared with that of sea ice, so microwave instruments can map sea ice concentration and extent with high accuracy. Although the passive microwave spatial resolution is coarse (the footprint size is roughly 28 × 37 km for SSM/I and 14 × 8 km for NASA’s advanced microwave scanning radiometer-Earth observing system, AMSR–E), the concentration of sea ice in each grid cell can be determined using a linear mixture analysis approach. This requires an estimate of brightness temperature for sea ice and open water and assumes that the overall pixel brightness temperature is a linear mixture of the brightness temperatures and amounts of each component in the pixel. The accuracy of this method depends on the ice types and other components present, how well they can be distinguished, and their complex spatial arrangement.

Regular, operational monitoring of sea ice started with SMMR in 1979 followed by SSM/I in 1987. Data from these instruments have been processed to create a time series of daily sea ice concentration and extent at a grid resolution of 25 km (Figure 4). This time series continues to the present with the special sensor microwave imager/sounder (SSMIS) sensor and has been augmented with 12.5 km resolution data from the AMSR–E sensor for the period 2002–2011.

Sea ice thickness

Since 2010, CryoSat–2 has been successfully mapping and characterizing sea ice conditions, including sea ice freeboard (the thickness of sea ice above the ocean surface) and sea ice thickness. Cryosat–2 uses an algorithm that determines the optimal fit between the measured waveform and a modeled waveform for a given sea ice freeboard (Laxon et al. 2013). The algorithm requires ancillary data such as the depth of the snow layer on the sea ice and a measurement of sea ice concentration, as ice thickness retrievals are only performed for grid cells having greater than 70% ice concentration. Assuming hydrostatic balance, ice thickness is then computed based on the snow depth and retrieved freeboard.

The GLAS sensor also derived estimates of sea ice freeboard and sea ice thickness. These laser altimeter measurements are limited to cloud-free periods during the 15 periods when the lasers were functioning from February 2003
Figure 4  Sea ice concentrations shown for the Arctic, September 2012. This was the lowest sea ice extent on record. The magenta line shows the historical average extent (1981–2002). Reproduced from National Snow and Ice Data Center.

to October 2008. Since the end of GLAS on the ICESat mission, Operation IceBridge has measured sea ice freeboard and snow depth using a combination of airborne laser altimeters and a snow radar, respectively. Below the ice, declassified upward-looking sonar data from United States and British Naval submarines have provided measurements of ice thickness along their Arctic traverses. US submarine data for the period 1958–2000 combined with the GLAS measurements, create a 42-year time series showing accelerating declines in sea ice thickness over much of the Arctic.

Sea ice type

When first forming, sea ice can be smooth and as thin as a few centimeters; over time, it grows...
thicker and more extensive with characteristic rough pressure ridges and cracks (known as “leads”). “First year” sea ice forms during the polar winter and melts in summer, but when it persists beyond one melt season it is called “multiyear” sea ice and can be 2–3 m thick.

Radars such as SEASAT-SAR, ERS1,2, and JERS-1 demonstrated the immense value of SAR for mapping sea ice characteristics. Classification of multiyear versus first year sea ice works with SAR because multiyear sea ice has more air bubbles (since the bubbles’ briny contents have drained out over time), so it has higher backscatter than first year ice. RADARSAT-1,2, operated by the Canadian Space Agency, have been used operationally to differentiate ice types and for daily ice charts. Such information has been critical for navigation through pack ice in polar regions. SAR’s improved spatial resolution over passive microwave allows mapping of leads and polynyas. Intermittently since 1978, scatterometers such as SEASAT-A, ESCAT, NSCAT, and SeaWinds have measured radar backscatter and have been used to characterize sea ice surface roughness. Scatterometer spatial resolution is comparable to that of passive microwave (e.g., SeaWinds spatial resolution is 25 km) due to the need to average radar backscatter over large areas to achieve adequate response. CryoSat-2 derives sea ice surface roughness from pulse-broadened waveform, which assumes a Gaussian distribution of roughness elements within the radar footprint.

**Sea ice motion**

Sea ice motion is computed by automated tracking of identifiable features from one image to another. This approach uses repeat temporal coverage from satellite data and has been extensively used with optical data (e.g., advanced very high resolution radiometer, AVHRR), passive microwave (e.g., SMMR and SSM/I), and SAR (e.g., RADARSAT-1,2). Using tie-points, image correlation algorithms record the pixel-by-pixel displacement of sea ice features. Sea ice motion tracking is performed on daily to weekly time scales and provides estimates of ice export in the Arctic and for sea ice conditions essential for polar navigation.

**Remote sensing of permafrost and frozen ground**

Soil, sediment, or rock that has remained at a temperature of 0°C or colder for more than two years is defined as permafrost or perennially frozen ground. Permafrost is typically overlain by an active layer, which freezes in the winter and thaws each summer. Seasonally frozen ground is that which freezes during the cold season and thaws during the warm season and it may exist in the absence of permafrost. Key properties include permafrost spatial extent, soil freeze–thaw status, active layer thickness, and thermokarst detection. Remotely sensing permafrost properties is challenging because most sensors only respond to the top few millimeters of the soil surface. Therefore, most methods measure permafrost-related variables or combine remote sensing measurements with models.

**Permafrost/frozen ground extent**

Because freezing is the hallmark of permafrost and seasonally frozen ground, land surface temperature (LST) is a key measurement for sensing the extent and freeze–thaw state of these soils. Satellite measurements of LST have been made using thermal band data from AVHRR and MODIS with spatial resolution of about 1 km. Land surface temperature is also derived from passive microwave brightness temperatures, such
CRYOSPHERE: REMOTE SENSING

as from SSM/I and AMSR-E, at spatial resolutions of 25 and 12.5 km, respectively. Although LST may not be an exact predictor of soil or rock temperatures at tens of centimeters below the surface, these data are valuable input to models that predict the spatial and temporal distributions of permafrost and seasonally frozen ground.

Soil freeze–thaw status

There is a fundamental difference in the dielectric properties of frozen and thawed soils. The bulk dielectric constant of soil is governed by that of water, as the liquid water dielectric constant is an order of magnitude greater than that of ice and soil minerals. Thus, passive and active microwave sensors respond to the freeze–thaw status of soils. Radar backscatter from a thawed soil increases by about 2–9 dB depending on variations in moisture content and surface properties. The 37 GHz vertical polarization passive microwave data have been used to map soil freeze–thaw status on a daily basis from 1979–2012 using SMMR-SSM/I-SSMIS and AMSR-E data (Figure 5). Because of their coarse spatial resolution, passive microwave brightness temperatures are influenced by subpixel tundra lakes, so soil freeze–thaw estimates need to be adjusted based on total lake area. Snow cover is another challenge for passive microwave measurements of frozen ground, as snow obscures the soil and some frozen soil detection algorithms use the same brightness temperature difference (with 37 and 19 GHz frequency) as used for snow mapping. When snow cover is present, an energy balance model must be used to ascertain the thermal state of the soil.

The ESA’s soil moisture and ocean salinity (SMOS) uses an L-band passive microwave radiometer to monitor seasonally frozen ground. The long wavelength of this instrument (21 cm) can provide information on soil freeze–thaw for depths up to 35 cm (depending on soil type and vegetation cover) though at a coarse spatial

![Figure 5](image-url)  Daily soil freeze–thaw status at 25 km from AMSR-E for April 10, 2010; a composite of morning (a.m.) and afternoon (p.m.) satellite overpasses. FR = frozen, NF = nonfrozen, TR = transitional (a.m. frozen, p.m. thawed), INV-TR = inverse transitional (a.m. thawed, p.m. frozen). Kim et al. 2014. Reproduced from the National Snow and Ice Data Center DOI:10.5067/MEASURES/CRYOSPHERE/nsidc-0477.003.
resolution of 35 km. NASA’s soil moisture active passive (SMAP) mission uses an L-band SAR to monitor soil freeze–thaw state. With a temporal resolution of 2–3 days and spatial resolution of 3 km, the SAR instrument on SMAP can capture spatial patterns not evident with passive microwave, although not necessarily the heterogeneous landscape processes that affect seasonally frozen ground.

Active layer thickness

Active layer thickness (ALT) is the maximum depth of seasonally thawed ground that overlies permafrost. Surface-based ground penetrating radar is effective for mapping ALT though only over small areas. InSAR provides an indirect method of estimating ALT by measuring the subsidence that accompanies soil thaw during the warm season. Promising new approaches with greater soil penetration depth and less influence of vegetation include the airborne P-band SAR (e.g., AirMOSS) and kilohertz frequency data (airborne electromagnetic method, AEM) can retrieve active layer thickness.

Surface expressions of permafrost and thermokarst

In regions where permafrost is thawing, the loss of ground ice is a spatially heterogeneous process that leads to localized frost heaving, surface deformation, and subsidence known as thermokarst. Thermokarst landscapes are typically hummocky and include numerous small lakes that form in melt depressions, mounds that remain when ice-rich polygonal ground thaws, and thermo-erosional gullies that form when massive ground ice melts along a slope. These features are readily observed in fine resolution optical satellite imagery with spatial resolutions of 0.5–4 m (e.g., IKONOS, QuickBird, GeoEye-1, WorldView-1,2,3, and Pléiades commercial satellites). Repeat measurements with active sensors such as airborne LiDAR and space-borne InSAR map surface deformation due to frost heave and subsidence.

SEE ALSO: Glaciers; Hyperspectral remote sensing; Ice sheets; LiDAR; Microwave remote sensing; Optical remote sensing; Radar remote sensing; Sea ice, ice drift, and oceanic circulation; Snow; Soils of cold and permafrost-affected landscapes

References


Beginnings

The cryosphere consists of all terrestrial forms of snow and ice—snow cover, floating ice, glaciers, ice sheets, frozen ground, and permafrost. It is a critical element of the climate system because of its high reflectivity, its insulating effects on the land and ocean, and its storage of water on short and long timescales.

Recognition of the past existence of more extensive Alpine glaciers during ice age conditions came in 1837 with the work of Louis Agassiz in Switzerland. However, the first glacier sketch map dates from 1601 for the Vernagtferner in the Austrian Tyrol. The earliest research on glaciers was carried out in Iceland by S. Pálsson in the 1790s, in studies that were only translated into English in 2004. He made a sketch map of the Vatnajökull glaciers.

Most scientific study of the different components of the cryosphere—snow cover, glaciers, ice sheets, freshwater ice, and sea ice—began around 1890–1900, whereas the study of permafrost and ground ice started in the Soviet Union in the 1920s but did not begin until World War II in the West.

The first definition of the cryosphere and its components was made by A.B. Dobrowolski (1923) in a massive treatise written in Polish, but the term was not widely adopted until the 1990s. Most writings treated single components until the publication of *The Global Cryosphere: Past, Present and Future* by Barry and Gan (2011) and *The Cryosphere* by Marshall (2011).

**Land components**

**Land ice**

Data on glacier fluctuations, from 1881 onward, were published in 1895, and the annual records have been maintained in the series “Fluctuations of Glaciers” published at five-year intervals by the World Glacier Monitoring Service (WGMS) in Zurich. Mass balance data were first collected on the Rhône glacier in 1874 and then continuously on Claridenfirn in Switzerland from 1914–1915. Other programs operated in the 1920s to the 1940s, but continuous annual measurements of glacier-wide mass balance began on the Storglaciären, Sweden, in 1945–1946 and expanded to other mountain ranges only in the 1950s. The WGMS currently lists annual data on terminus locations for 1967 glaciers and on mass balance for 266 glaciers, but most are small and not globally representative. There is now a nearly complete global glacier inventory (the Randolph Glacier Inventory), which contains 170,000 glacier outlines and a global ice area of 737,000 square kilometers. However, details about many of the glaciers remain incomplete. Consequently, the global area and volume of ice remain uncertain.

Theoretical work on glacier flow and snow mechanics began in Britain and the Alps in the late 1930s and resumed after World War II. Extensive fieldwork on glaciers in the USSR began during the International Geophysical
year (IGY; July 1957 to December 1958), and continued through the early 1990s. In Canada and Alaska such work mainly began during the International Hydrological Decade (1965–1974).

High-resolution remote sensing, which began with Landsat in 1972, has greatly expanded the information available on world glacier distribution and characteristics. The Global Land Ice Measurement from Space (GLIMS) program uses the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) instrument, aboard the Terra satellite, and the Landsat Enhanced Thematic Mapper Plus (ETM+). The GLIMS database has 116,000 glacier outlines. Interferometric synthetic aperture radar (InSAR) mapping has been used to determine motion on selected glaciers since the 1990s.

Ice sheets

Expeditions to the Greenland ice sheet were undertaken in the 1880s and 1890s, but the first substantial work was performed in 1930–1931 by the British Arctic Air Route Expedition of Lauge Koch and Gino Watkins at Watkins’s ice cap station (2440 m) and Alfred Wegener’s Greenland Expedition to Eismitte (3100 m). In 1946–1948 P.-E. Victor led the Expedition Glaciologique Internationale au Groenland (EGIG), which established a survey line across the ice sheet. Summit station (3216 m) has operated continuously since 1989.

Antarctic scientific expeditions began in 1899 with the first overwintering on the continent, but some of the earliest glaciological research was conducted by the Norwegian–British–Swedish expedition of 1946–1947 in Queen Maud Land. During the IGY 12 nations established over 60 stations in Antarctica including South Pole and Vostok on the East Antarctic plateau.

Ice core research began in Greenland during the IGY when 300–400 m cores were extracted and analyzed by C. Langway Jr. Cores to bedrock were drilled at Thule, Greenland, in 1966 and at Byrd station, Antarctica, in 1968. W.C. Dansgaard laid the foundations of paleoclimatic research using ice core data in 1969. The longest climate record to date was recovered from the European Project for Ice Coring in Antarctica (EPICA). The 3,139 m of core so far analyzed spans 740 ka and covers eight glacial-interglacial cycles.

Ice thickness surveys using airborne radio-echo sounding began in Antarctica in the 1960s. During the 1990s and 2000s ice sheet altitudes were mapped by satellite radar and lidar surveys, and melt extent on Greenland has been mapped from passive microwave data. Since 2002, volume change estimates have been derived from the Gravity Recovery and Climate Experiment (GRACE) satellite.

Snow cover

Snow surveys began in California in 1906 and soon expanded to most western US states. These were generally made at monthly intervals in late winter and spring as the basis for streamflow estimates. Snow depths are recorded daily at weather stations but the data are not exchanged internationally.

Hemispheric analysis of snow cover began in the United States in October 1966 from the polar orbiting Very High Resolution Radiometer (VHRR). Initially weekly, maps became daily in February 1997. These were complemented by maps from the 500 m resolution Moderate Resolution Imaging Spectroradiometer (MODIS) on Terra (2000) and Aqua (2002). Snow water equivalent has been determined from passive microwave remote sensing for the northern hemisphere since 1979, although forest cover and melt layers limit the accuracy of the products.
Freshwater ice

There are records of ice on Lake Constance dating from 875 CE and on Lake Suwa, Japan, from 1443 CE. Observations of river ice breakup started in Finland on the river Tornionjok in 1693, and on lakes Kallavesi and Näsijärvi in 1833 and 1836. Ice conditions on the river Danube were studied in 1848 following two years of flooding and in the late nineteenth century on the St. Lawrence River in Canada. Canada maintained an extensive network of around 240 lake observing sites from the 1940s to the 1990s, but almost all of these have since been terminated. In 1995 the Canadian Ice Service began a program of monitoring ice extent on small lakes using high-resolution satellite imagery. The amount of ice on each lake is determined weekly by visual inspection of Advanced Very High Resolution Radiometer (AVHRR) and RADARSAT images. The program had expanded to 118 lakes by 1998.

Trends in lake ice formation and breakup were studied in Finland in the 1940s and 1980s, and subsequently for other regions across the northern hemisphere.

Permafrost and frozen ground

The existence of frozen ground was first noted in 1577 in Baffin Island and then again in 1642 in Siberia. The earliest scientific recognition of frozen ground is attributed to Karl von Baer in 1838. The first maps of permafrost in Russia were published in 1872 and 1899. M.I. Sumgin published the first book on perennially frozen soils in Russian in 1927. The first English text was not published until 1947 by S.W. Muller. Permafrost research in China started in the 1950s. A circum-Arctic map of permafrost distribution and ground ice content was prepared by the International Permafrost Association and published in 1997 by J. Brown and others.

Monitoring of the active layer thickness in the circum-Arctic began in 1991.

Marine components

Sea ice

There is an index of East Greenland ice off the coasts of Iceland from 1150 CE, although some authorities consider it to be unreliable before 1600. Ice edge positions were first documented in the Nordic seas in the 1550s and in the North Atlantic in the 1890s.

In the Antarctic there are whaling ship records from the 1930s to 1950s. Regular aircraft reconnaissance of Arctic ice off Eurasia began in the USSR in 1933 and continued until 1992. Summer ice patrols began in the Canadian Arctic in 1957 and continued through the 2000s. Satellite mapping of ice extent and concentration began in the mid-1960s, with all-weather global passive microwave data available from 1979 and high-resolution radar data from 1991. Ice draft in the Arctic was first measured in 1958 using submarine upward-looking sonar. Sonar data continued to be collected from 1971 through the 2000s, but from 2003 to 2009 the ICES at laser altimeter enabled the direct measurement of sea ice freeboard. In 2010 CryoSat 2 was launched with interferometric radar for the same task.

William Scoresby Jr., an English whaling captain, published a notable book on ice and ocean conditions in the Greenland Sea in 1820. F. Nansen performed the epic drift across the eastern Arctic Ocean in the Fram during 1893–1896. The first book on the physics of sea ice was written by F. Malmgren in 1927 based on observations made during the Norwegian ship Maud's north polar expedition, 1918–1925.
CRYOSPHERE STUDIES: HISTORY

Ice shelves and icebergs

Ice shelves consist of grounded and floating land ice. They exist mostly around the Antarctic ice sheet. The eastern part of the Filchner–Ronne shelf was first explored by W. Filchner in 1911–1912. The western part was photographed from the air in 1947. Bases were set up on the Ross ice shelf from 1929 to the IGY. Major scientific work began in the 1960s and 1970s.

Iceberg calving and ice shelf breakup have been monitored by satellites from the 1970s. Icebergs in the northwestern North Atlantic were documented annually from 1912, after the Titanic disaster.

Organization of cryospheric science

The first organization concerned with the cryosphere was the International Glacier Commission, established in 1894, which documented glacier changes, initially in the Alps. This became the Commission on Glaciers and Snow of the International Association of Scientific Hydrology (IASH) in 1939, which in 1948 was renamed the International Commission on Snow and Ice (ICSI). During the 1990s ICSI members increasingly felt constrained by the restricted scope of their activities under their parent organization, the International Association of Hydrological Sciences (IAHS). Between 2001 and 2004, Gerry Jones, the president of ICSI, decided to push for the establishment of an association of cryospheric sciences as part of the International Union of Geophysics and Geodesics (IUGG). A committee (Roger Barry, Jan Hagen, and Wilfred Haberli) documented the case for the new association and in 2007 the International Association of Cryospheric Sciences (IACS) was approved at Perugia, Italy, as the eighth association of the IUGG. It held a successful first meeting in Davos (DACA-13) in July 2013.

The International Glaciological Society (IGS) was founded as the Association for the Study of Snow and Ice in 1936. It became the British Glaciological Society in 1945 and changed its name and scope in 1971. The International Permafrost Association (IPA) was founded in 1983, and in 1989 it became an affiliate of the International Union of Geological Sciences. In 2000 the World Climate Research Programme (WCRP) established the Climate and Cryosphere (CliC) project to coordinate international research efforts in this field.

SEE ALSO: Climatology; Glaciers; Ice sheets; Permafrost: definition and extent; Sea ice, ice drift, and oceanic circulation; Snow cover

References


Further reading


on Snow and Ice (ICSI) to the International Association of Cryospheric Sciences (IACS).” Annals of Glaciology, 48: 1–5.


Cuba: Instituto de Geografía Tropical (Institute of Tropical Geography)

Founded: 1962
Location of headquarters: Havana
Website: www.geotech.cu
Cultural capital

Louise C. Johnson
Deakin University, Australia

The elaboration of a term like “capital” beyond its financial confines has occurred through the work of economists, labor market analysts, and sociologists. Unpacking or extending the term into the various forms that capital can take occurred in the second part of the twentieth century as scholars acknowledged the limitations of neoclassical and Marxist economics. The result has been discussions of “human capital,” the ability of people to create value through their labor and related actions; “natural capital,” or environmental resource endowments; and various forms of capital relating to human values and social positions, via “symbolic capital” and “cultural capital.” The latter term has been particularly important to sociologists, educationists, economists, and geographers as they have theorized the rising importance of consumption over production as well as culture in the constitution of identities and contemporary economies. Fundamental to this reorientation was the work of the French sociologist Pierre Bourdieu who, in his classic 1984 text, Distinction, put status and the quest for social distinction at the center of his analysis of the French class system. This entry considers the nature of cultural capital and related terms developed by Bourdieu before focusing on the main ways in which geographers have utilized the term to analyze the cultural industries and related transformations of economic and social space.

Definitions

The French sociologist Pierre Bourdieu derived and used the notion of cultural capital to analyze and explain the high failure rate of working-class children in the French school system. It was later extended to detail and socially situate the tastes of the bourgeoisie and the ways in which particular cultural forms acquired value. He argued that “culture” has the effect of reinforcing and legitimizing middle-class power at the expense of lower-class taste, status, and power (Bourdieu 1984).

In contrast to a social model that prioritized disembodied class relations or an economic one that privileged rational behavior, Bourdieu admitted human values, taste, and the quest for status into his sociological analysis. Cultural capital thereby describes the possession of knowledge, accomplishments, and formal and informal qualifications – all of which are embodied by individuals and used by them to negotiate their social position. A person’s socialization via family upbringing, formative peer groups, and formal education endows them with stocks of generalized cultural capital that are expressed as habitual tendencies and generative predispositions, forming a habitus. Acquired primarily through formal education, cultural capital is subsequently expressed through consumption practices, which tend to reproduce
socioeconomically based status distinctions. The status of a given type of cultural capital is thereby set by the values, interests, and tastes of the dominant social class factions. In Bourdieu’s analysis, although it does not necessarily match the distribution of economic and social capital, cultural capital tends to reinforce the unequal class order of late twentieth-century France (Bourdieu 1986).

Bourdieu argues that when consumers accumulate new forms of cultural capital they are making an investment in particular constellations of skills, knowledge, and aptitudes. The status value of these stocks of cultural capital lies in their conversion rate – into additional forms of cultural capital but also into social capital (advantageous social connections), economic capital (via occupational opportunities), and symbolic capital (social recognition and respect). The conversion rate is related to the social field in which it is deployed and to the underlying relations of socioeconomic advantage and disadvantage that structure the distribution of social and economic resources among social fields.

If cultural capital is held (or embodied) by individuals as a consequence of their family background, education, and placement in the class system, the cultural object – or objectified cultural capital – intersects with the schema to attain its value through its position in a field of cultural production (Johnson 2009). Within each artistic field, for Bourdieu, there are objective qualities which govern success, and these include price, awards, or grants. These fields have aesthetic, symbolic, and historical value. Each field is the consequence of forces and struggles between key arbiters of taste; institutionalized interests; those who gaze on, value, and purchase the artistic products; and the political context in which transactions occur. As a result, changes in literary or artistic possibilities result from alterations in the power relations, which constitute the positions and dispositions by those involved in defining art and its value: bureaucrats, patrons, critics, producers, arts managers, viewers, and consumers (Bourdieu 1996). It is the intersection of embodied and objectified output with artistic institutions that thereby gives any art object or activity its value. In his detailed analysis of particular art forms Bourdieu thereby linked artistic objects and activity to class, power, politics, and taste.

Uses of cultural capital

Bourdieu’s notion of cultural capital – as a set of acquired markers of social distinction as well as a means of unpacking the value of cultural objects – arose from his sociological investigation of the role of the French school system in reproducing class relations. It is therefore not surprising that much of the work utilizing this concept and comparable methodologies has occurred in education as researchers have charted the ways in which class is either reproduced or modified through schooling. There has also been a use of the concept in literary studies, to unpack the establishment of a literary canon or national identity, and in sociology, with numerous studies of the quantity and quality of cultural capital held by particular social groups.

If such work has built directly on that of Pierre Bourdieu, the notion of cultural capital has been usefully extended by those who have focused on the apparent shift toward creativity in the economic system. Notions such as the cultural industries, the creative class, and the creative city have thereby emerged (see Hesmondhalgh 2002). Key to this conceptual shift, in addition to those prefigured by postmodernism and post-Fordism, was the work of Richard Florida.

In an enormously influential book on The Rise of the Creative Class (2002) Florida argues...
that the various economic, technological, and social transformations that have made the media and communications industries vitally important have also engendered other changes. Beginning with the geography of regional growth across the United States, Florida isolates a number of common characteristics which for him explain the emergence of cities like Washington, DC, Boston, Austin, Chicago, and San Francisco as centers of rapid population growth. He argues that their growth did not arise from firms moving into these areas but rather from creative people seeking out and making these urban environments their own. These cities in turn become centers of innovation and high-technology industry. What they have in common, Florida argues, is “technology, talent and tolerance” (2003, 10). Technology involves the mobilization of capital and innovation to create new things, processes, and industries by those with talent. These talented individuals, in turn, Florida argues, are attracted to places that have a social milieu that fosters and supports their creativity and associated work and lifestyles (Florida 2002; 2003).

Florida’s analysis has proliferated across the globe and has inspired major policy shifts toward the arts and the cultural industries, and political tolerance in a range of places. While it is criticized for being tautological (i.e., if you have a broad definition of a growth class you will necessarily get growth regions), somewhat instrumental, empirically flawed, and overly optimistic, Florida’s work has catapulted the notion of the cultural industries and of cultural capital (a city focused on the development of its creative class and culture) to the top of many urban political agendas. The work that has impacted the most on geography has engaged with the role of the creative arts in the establishment of class taste and power within the city as well as in the growth of the cultural and creative industries which offer new jobs and vehicles for urban and regional regeneration.

**Uses in geography**

Few geographers use Bourdieu’s framework without significant modification, as a result of their application in other disciplines and the very different questions posed and because the world is a different place from 1970s France, when his research was done. Its use can be linked to the revitalization of cultural geography in the 1980s. Thus, the “new cultural geography” is concerned with how landscapes and dominant meanings are produced through the interplay of social relations and texts. Now a vast subfield covering everything from urban images and agricultural landscapes to marginalized social groups, the new cultural geography has focused on cultural and symbolic capital.

In a vital early study of loft living in New York, Sharon Zukin (1982) wrote of the many costs and problems associated with the “symbolic economy,” an intertwining of cultural symbols with entrepreneurial capital. For her the symbolic economy in action was corrupt, as the association with entrepreneurial capital devalued the artistic product, compromised creative people, and imposed exploitative social relations on all those involved. But, in her analysis, Zukin (1995) connected the creation of cultural symbols to urban transformations, materializing the arts into a transformative political urban economy.

In his study of gentrification in Toronto, David Ley (2003) invokes the notion of cultural capital more explicitly as a vital component in changing the value and class character of particular parts of the city. He charts how previously run-down areas with few economic resources have been reclaimed by social groups (artists, academics, intellectuals) with large amounts of cultural
CULTURAL CAPITAL

capital but limited economic capital. In this process, there is a shift as the areas are injected with labor and value to become sites of both cultural and economic capital, in other words, gentrified. It is a powerful argument and demonstration of the power of Bourdieu’s concept in understanding the changing nature of the western city.

If cultural capital has been used to understand the process of gentrification, the notion has also been connected to emerging economic sectors. Allen Scott (2000) has analyzed the importance and specific character of the “cultural products industries.” In addition there have been geographical studies of particular cultural industries, such as the film industry, fashion design and retailing, and the manufacture of surf boards, as industries that basically combine rich concentrations of cultural capital with artisanal production methods to ensure the transformation of cultural to economic capital.

The role of the cultural industries and concentrations of cultural capital in urban and regional regeneration formed the basis of an entire European Union-sponsored program of naming cities as and elevating them to centers of culture or “cultural capitals” (see Landry and Bianchini 1995). The first of these – Glasgow – best illustrates the challenges and benefits of such a designation as this declining manufacturing city was reimagined and rebuilt around its pre-existing cultural assets but also a well-funded and supported program of art, physical regeneration, and programs to grow its cultural industries (Johnson 2009). Directed by the subsequent theorizing by Florida, Landry, and others, this program has been replicated across Europe, North America, and Australasia so that the mobilization of cultural capital has become synonymous with urban regeneration, social animation, and reimagining.

If these examples have taken notions of cultural capital and either applied them directly or merged them into studies of the cultural industries and cities, there have also been geographers who have reformulated their theoretical premises. In their analysis of emerging forms of cultural capital in Europe, Prieur and Savage argue that the types and social relations underpinning its theorization in Distinction need to be updated because of the decline in “high-brow” culture and the emergence of what they argue is a cosmopolitan cultural capital within a European field of a well-educated professional and managerial class. They further argue that, as popular cultural objects become aestheticized and as elite objects become popular, the objective form of cultural capital is supplanted by the embodied form, with the upper classes now more wide-ranging and discerning in their cultural practices (Prieur and Savage 2013).

Cultural capital, therefore, has been used directly and indirectly by a number of geographers to analyze existing phenomena such as gentrification, to understand emerging economic sectors (the cultural or creative industries), and to form the basis of new policy agendas for urban, regional, and social regeneration. In this exercise they have both used the concepts developed by Bourdieu and modified them in the light of Bourdieu and modified them in the light of contemporary economies or their analytical purpose.

SEE ALSO: Art; Cities and development; Creative cities; Creative class; Cultural geography; Cultural turn; Gentrification; New cultural geography; Rust Belt cities; Urban renewal

References


Cultural diffusion

Victor Roger Savage
National University of Singapore

The first use of “cultural diffusion” goes back to the American anthropologist Alfred Kroeber’s (1944) influential paper on “stimulus diffusion” or transcultural diffusion. Since then, the concept of diffusion has been used across many disciplines. For geographers, the concept is used to map the travel of both tangible and intangible culture over space. The Swedish geographer Torsten Hagerstrand (1952) has contributed most to an understanding of spatial diffusion processes. Many of the conceptual paradigms of diffusion come from the Swedish school of spatial diffusion. Four of these diffusion processes are often cited in the literature: expansion diffusion (spatial expansion of cultural ideas/artifacts from one of multiple sources); relocation diffusion (implant of a culture in a different place and time); hierarchical diffusion (movement of culture from larger entities to smaller places); and contagious diffusion (how ideas move rapidly through societies/communities through personal communication). The Marxist geographer Jim Blaut (1993) is critical about conceptualizations of diffusion, arguing that on a world scale “diffusionist ideas” are often biased, Eurocentric, and based on single belief systems.

All cultural diffusion needs to be understood as a product of three intertwining factors: (i) culture must have an origin(s), or cultural hearth(s); (ii) it must be transported in time or across space and society; and (iii) it must find expression in a new time, place, and community. Cultural diffusion is as old as the first Homo sapiens seeking effective ways of ensuring survival in varying ecosystems.

Much has been written about cultural diffusion between civilizations and cultural hearths around the world. Academic debates continue to persist about the transference of scientific knowledge, the idea of democracy, and the beginnings of capitalism especially between East and West. For example, while Fernand Braudel, Immanuel Wallerstein, and Eric Wolf state that capitalism originated in Europe and spread to the rest of the world, Janet Abu-Lughod, Jim Blaut, and Andre Gunner Frank posit that capitalism originated in the East and spread westward.

Cultural diffusion has been most widespread when there are institutional forms of cultural transference and indoctrination. Three institutional forms of cultural diffusion may be identified: colonialism, religion, and education.

Throughout history, developed states and civilizations often engaged in conquest and colonization processes, which in turn have given rise to forced cultural diffusion. For example, for over 1000 years China ruled over what is now Vietnam, which it converted into one of its provinces, thereby leading to the diffusion of Chinese culture to the Vietnamese. Edward Said (1994) offers an interpretation of how Western colonialism was not only a colonization of land but an ideological enterprise of the mental subjugation of indigenous peoples. Benedict Anderson (1991) identified three specific ways of achieving this colonial enterprise: the use of mapping to define the colonial state; the development of museums to provide the historical narrative of the colonial state; and the introduction of population census to survey the
CULTURAL DIFFUSION

extent and character of colonial subjects – all current instruments of modern states.

Religious and spiritual beliefs are pervasive across many communities around the world. These include world religions such as Christianity, Islam, Buddhism, and Hinduism, but also an increasing range of spiritual beliefs such as Wiccanism and esotericism. An explicit form of active cultural diffusion is evident in the proselytizing of religions by missionaries. The spread of religion is often accompanied by the diffusion of religious rituals and taboos such as an absence of pork consumption among Muslims or avoidance of beef by Hindus.

The adoption of public and mass education in nearly every state around the world has led to cultural diffusion in two ways. On the one hand, all state public education attempts to inculcate in students a sense of the “national history and culture.” On the other hand, education has also been responsible for the most enduring diffusion of cultural ideas, from those of science and technology, through the social sciences and humanities.

Despite Huntington’s “clash of civilizations,” which remains anchored in Cold War, East–West polemics, the world is becoming more integrated with new technologies (Huntington 1996). The Jesuit priest Teilhard de Chardin’s idea of noosphere (sphere of knowledge) as the repository of global human knowledge is now becoming a pragmatic reality with cyberspace, through blogs, activity streams, event streams, Wikipedia, Internet, Twitter news streams, knowledge, and ideas for a global community. Larry Sanger, cofounder of Wikipedia, argues that Wikipedia is a new industrial revolution, a product of collaboration and the democratization of knowledge, ideas, innovations, news, and information. Such democratization necessarily relies on the rapid cultural diffusion that is facilitated by the new technologies.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Cultural geography; Religion

References


Cultural economy

Andy C. Pratt
City University of London, UK

The cultural industries comprise publishing, film, TV, music, theater in both their recorded and their live forms. There has been a focus on the economic reproduction of these cultural forms and, usually separately, of the artists who feature in them. Less common has been a concern with the various technicians and support staff. Accordingly, a more comprehensive framework has sought to include both the objects and texts, as well as labor: the system of production and reproduction of cultural forms, or a cultural field. With such a dynamic topic there is ongoing disputation of its boundaries and value; this illustrates the relational nature of both the circulation of cultural meaning and its constitution.

There is a range of socially and politically privileged cultures and practices, and others that are merely tolerated or even censored. The role of cultural governance is also important within any institution, and specifically within state institutions. Historically, state support for culture has been legitimated through a welfare economics discourse: that culture was subject to market failure, but society regarded it as a merit good, hence it was supported by the state. Historically, culture in this sense is restricted to immaterial and material cultural heritage and the fine arts. In recent years this state-legitimated culture has been placed in tension with the mainly commercial cultural industries. This tension was the primary concern of Adorno (Horkheimer and Adorno 2002/1947), who coined the term “culture industry” as representing a disparaging negative mirror image of “real” culture. The tensions between the mass-produced and craft, and authenticity and its other, have a resilience in public debate. The old tension between the economic and the cultural has been given new force with the pincer movement of, on the one hand, austerity-driven cuts in state budgets or simply neoliberal small-state initiatives, and, on the other hand, the profitability and economic driving force of some aspects of the cultural field.

Another body of work that has used the term “cultural economy” in recent years is that of cultural, economic, and social geographies of the economy (Amin and Thrift 2004). These are a response to the “cultural turn” in geography and seek to apply the tools of cultural analysis to the object of economic action. It is interesting to note that, thus far, none of these analyses has taken the cultural economy (as outlined in the current entry) as an object of analysis.

Terminology

As noted earlier, the originating term was “culture industry,” an object defined by its negative relationship to real culture. By the 1960s the scope of what was and was not culture had been widened, and French communications scholars sought to deploy a new terminology that would reflect the field as a serious area of research. They
also wanted to point to the differences within the field, so they used the term “cultural industries” (Garnham 2005). This term proved helpful in drawing to attention the expansion of related activities under the “cultural” umbrella, both state-supported and commercial. Recognizing the social and economic importance, and registering its purchase on politics, the cultural industries became a topic of policy concern; however, such concern was separate from the traditional focus on cultural policy. In most cases the cultural industries were the responsibility of business and industry departments, while culture came under the remit of heritage. As the cultural industries climbed the political agenda, there was scope for tension with cultural policymaking.

In 1997 the UK government, initiating a re-focus of its governance of the cultural field, took the bold step of placing both concerns within one department, linked to communications. Similar strategies had been adopted by Australia, Canada, and New Zealand. It focused on the commercial cultural industries by renaming them the “creative industries,” oriented to those industries that generate value through intellectual property rights (DCMS 1998). In political terms, this turn made significant overtures to the “knowledge society.” The notion of the creative industries has often been elided with the discussion of the creative class by Florida (2002), which itself conflates with the notion of the creative city.

In order to avoid the fragmentation of a fast-developing field, academics and policymakers began to use the term “creative economy,” taking “creative” from the identification with the “creative industries,” and “economy” from a pre-seventeenth-century usage as a general management of resources. This latter inflection was important in that it sought to loosen the boundaries between for-profit and not-for-profit and formal and informal that clearly united the field in practice. Countries with a very strong public tradition of public funding for the high arts used the term “cultural and creative industries” as it allowed them to speak of the commercial and the cultural at the same time. The narrowing confines of the “creative” modifier have been a source of concern, potentially focusing in an idealist romantic version of “creativity” and genius, as opposed to the social process of cultural reproduction. Recently, the policy community has adopted the term “cultural economy,” thus avoiding the reduction to “creativity” and foregrounding the socially situated constitution of “culture.”

### Long-term change

How and why has the cultural economy come to such prominence? The concept of the separation of culture from other parts of life rose in prominence in eighteenth-century Europe, with an increasing concern with the perfection or beauty deriving from a deity. Humanism and romanticism were woven into a powerful conception of the artist as author and interpreter of the world. These sensibilities were the basis of decisions about which cultural forms were to be supported and cultivated. That they were not “economic” but “merit” goods (goods that the market would not value correctly) gave the state – if it valued these goods – a justification and obligation to support culture (Throsby 2001). Of course, culture also has an ideological role, which is critical in nation building.

There are two aspects to the transformation of economies in the Northern Hemisphere from an industrial to a postindustrial mode of organization. First, in general, people had more money to spend on consumption and so cultural consumption grew. Moreover, participation
CULTURAL ECONOMY

in higher education and extended years of compulsory education created a new consumer power: the teenager. Second, the balance of the economy has shifted to the postindustrial—or service—sector, termed the “knowledge economy.” It has been claimed that the creative, or cultural, economy is at the breaking edge of this new wave of economic change. Such a dual rationale places the cultural economy in a conceptually strategic position to lead growth.

Empirical measures of the cultural economy have since the mid-2000s showed that such anticipated transformation and growth are already a reality (UNCTAD 2010). The growth rates of employment, gross value added (GVA), and trade in the cultural economy have outstripped most other sectors; moreover, they have maintained positive growth through the recession. In distributional terms, much of this growth is concentrated in major cities (highlighting a quite different dimension to the instrumental and consumption-based creative cities agenda); indeed in many Northern cities the cultural economy is in the top five industries. While cultural employment and trade are concentrated in the Global North, rates of growth are higher in the Global South, indicating that the cultural economy is of growing international significance.

Concepts and measures

The notion of the cultural economy is a contested one, and steps to find reliable measures have been hard-won. There is a simple reason. The analytical and statistical framework that constitutes our understanding of the economy is founded (taxonomically) on an early twentieth-century snapshot. Many high-growth and transformative industries of the late twentieth and twenty-first centuries did not exist previously, or were so small as not to be considered worth a taxonomic category. The industries that had grown out of others, such as chemicals, engineering, and pharmaceuticals, were captured by incremental change and the modification of taxonomies. Others, the ones that were not “industries” (such as public cultural forms, orchestras, theaters, etc.) were not recorded or were bundled up with other activities that did not fit the taxonomies. Economically speaking, this was not of great concern, as they were merit, or nonmarket, goods that were not expected to enter into the price system. Of course, as such divisions have dissolved, state bodies have struggled to keep apart the cultural sector and the creative industries, administratively, as they were governed under different logics (indeed, in Adorno’s initial conception, the two categories are corrosive of one another and hence must be separate). So, gross activity in the economy may be measured, but any attempt to find out how many are involved in the computer games industry would draw a blank. As indicated previously, around the end of the twentieth century a number of countries began to experiment with measures for (what they believed to be) a growing cultural economy. The United Kingdom coined the term “creative industries” and included a list of 13 industries selected on the basis of pragmatism (i.e., that they could measure) rather than a rigorous attempt to delimit the field (Box 1). However imperfect it was, the DCMS (1998) “mapping document” proved to be a much replicated template.

National statisticians like good time series, and are wary of making changes to taxonomies lest the time series be broken. The emergence of a new “industry” thus presents a problem. Accordingly, a compromise was adopted: simply, the method was to go back to the basic statistical
Box 1 The DCMS’s 13 creative industries.

The DCMS introduced the notion of the *creative industries* in 1998 defined as “those industries which have their origin in individual creativity, skill and talent and which have a potential for wealth and job creation through the generation and exploitation of intellectual property.”

Taxonomies for industries and occupations at their most disaggregated and then to reconstruct an artificial new taxonomy that included all the activities of the cultural economy. Normally employment statistics on the economy show a breakdown at a one- or two-digit level. The one digit gives descriptors such as agriculture, mining, and manufacture. The two digit gives types of industries, for example car-making, chemicals. The job of taxonomic reconstruction involves going to the four- or five-digit level analyses. When the new “codes” are combined, data can be associated with them to construct a “creative economy” category. Two measures are commonly used: those of employment and of economic output. Researchers have approached the employment measures in three ways: an analysis based on occupations (Markusen et al. 2008) or one based on industry (Pratt 1997), or a combination of the two systems. Clearly, this is not an ideal state of affairs, but, until the whole international statistical taxonomic system is revised, a satisfactory historical record cannot be constructed. Measures of cultural trade are even more problematic. Trade is normally measured in weight, and trade in “invisibles” is underrecorded (UNCTAD 2010). Therefore, although much effort has been put into this task, it significantly underrecords the economic value of culture (let alone the noneconomic value).

1 Advertising
2 Architecture
3 Arts and antique markets
4 Crafts
5 Design
6 Designer fashion
7 Film, video
8 Interactive leisure software
9 Music
10 Performing arts
11 Publishing
12 Television
13 Radio

The logic for selecting particular occupation or employment categories and excluding others has been the subject of considerable debate; however, work from a conceptual foundation has provided a more comprehensive underpinning for future taxonomies and mapping exercises. In general, the difference between these approaches is that the occupational and opportunistic approaches rest on a quasi-romantic conception of “pure” culture/creativity which, like a ripple in a pond, becomes less and less creative or cultural until one reaches the rest of the economy (Throsby 2001). Such a model places artists at the core and technicians on the periphery. There is also no principle as to where to draw the line between culture or creativity and the rest of the economy. An alternative approach takes the notion of a production system that describes the set of activities and skills necessary to produce a given cultural output. Here there is the concept of a codependency of multiple
The idea
Production/crafting
a model
Reproduction/mass
production
Distribution
Exchange/performance
Archiving/reflecting

Figure 1  The creative economy ecosystem.

processes on a final product, what we normally understand by an industry (Pratt 1997). It is this latter approach that underpins the UNESCO (2009) framework for cultural statistics. This approach has advantages in that it is neutral as to whether the activity is state- or privately funded, high or low culture: it captures comprehensive processes. These approaches have yielded insightful geographical analysis that foregrounds the flows of people and ideas in the construction of cultural value (see Figure 1).

Organization

The analytical challenges of the cultural economy do not end with the collection of quantitative data; qualitative dimensions present additional problems. As noted earlier, the immaterial aspects and the novel (not previously measured) highlight the cultural economy’s outputs, with differentiations based on causal processes rather than appearance or social value systems. From a normative perspective, industries of the cultural economy appear chaotic, or at least quite different. Such criticisms are sometimes used to reinforce popular prejudices about the economic utility of culture, the management abilities of artists, and the “inefficiencies” of the state-subsidized sector. Extensive and detailed analyses of the industries of the cultural economy are relatively new; however, they indicate a variety of novel organizational modes, and disclose rationalities that are not always anticipated by normative approaches. These organizational forms show five common characteristics.

First, the organizational structure of the cultural economy has been characterized as one with a “missing middle,” meaning that mid-sized organizations are underrepresented and small and microenterprises overrepresented. This depiction creates an organizational problem, one that can be solved broadly in two ways: internalize all activities in a larger corporation, or hope for intermediaries to emerge. This echoes a trend of outsourcing that has characterized much of the economy; it has been argued that the cultural economy was a leader in such changes: post-Fordist avant la lettre. Currently the cultural economy is characterized by many freelance and self-employed people, by those in microenterprises (smaller than small), and by a small number of transnational corporations. The articulation of the two organizational forms generates both risk and opportunity.

The second is project working. Smaller enterprises that are able to mediate are formed as short-term projects. These are temporary groupings (firms) of workers and microenterprises to deliver a project (such as a film or a TV series). At the end of the project, the team disperses. The advantages of this form are that unique skill combinations can be managed at short notice. The disadvantages are that workers are in a constant state of impermanence. Such short-termism militates against training and knowledge transfer within organizations. It often
CULTURAL ECONOMY

means that learning, training, and knowledge transfer happen between or outside companies in what has been termed the wider ecosystem of the cultural economy.

The third characteristic is risk. The risk of not having a sustainable job, which is a consequence of project working, is amplified by the nature of the cultural market place. In all markets some products succeed and others fail, but the spikiness is notorious in the cultural economy; it produces a phenomenon that has been termed “winner takes all” markets. Moreover, the uncertainty as to which product will win in the market is extreme. It is a commonly cited rule that the ratio of success to failure is 1:10. The only way to hope to survive the law of averages is to produce at least 10 potentially winning products; if one is fortunate a single product out of the 10 will succeed. This may sound extreme, but it is common in other high-risk industries such as financial services, where it is termed portfolio management.

The fourth is turnover. One of the reasons that stakes are so high is that cultural value is hard to predict and changes quickly. So, even if a product is successful, it may be superseded by another in a short space of time. This places significant pressure on companies to be innovative. While a car company may introduce a new model every two years, clothing fashions change by the month. In an indication of the speeding up of turnover, fashion used to have a seasonal turnover; fast fashion seeks to have a constant renewal. The prime example is the music company, which releases many songs a week, and those that are hits may be so for only a week.

The fifth is boundary spanning. Cultural innovation happens in all parts of the economy and society; some are planned but others are not. Cultural practices take place in the formal and informal economy; the public and private; and in both for-profit and not-for-profit realms. Cultural practitioners tend to work across these boundaries in part in order to put together jobs that may constitute a sustainable strategy; but also to gain new experiences and skills. The mobility of cultural workers can be an embodied form of knowledge transfer; however, it can be sustained only within a cultural ecosystem.

Work

Work is an important theme in research on the cultural economy, and represents the move away from the idealization of culture and artists that commonly precluded them from the lens of cultural analysis. First, the organization of cultural work exhibits a complex ecosystem of firms (commonly with 1–10 workers) and freelance workers, organized around short-term projects. Second, this temporary work engagement leads to precarious employment conditions for most, and the reconception of what a career is. Third, cultural labor includes physical tasks and mental labor, but is to some extent driven by a passion for a particular cultural practice. For some commentators, this raises new issues of “immaterial labor” and to what extent the “self” is sold or is over- and self-exploited. Fourth, and related, is the interest in affective labor and the presentation of self in cultural work. As noted, this affective turn characterizes the cultural economy for many geographers.

Place/space

In the analog age, culture and expression were commonly associated with the places in which they were practiced as if the soil itself gave rise to culture. This tendency was countered by “high” or transcendent culture, which was “out of place.” While craft production and the
limited distribution of cultural goods or service took place, they were not economically significant (except in the trade of goods comprising high-value components, e.g., jewelry). With the mass reproduction of cultural goods huge returns to scale, and in many cases monopolistic market forms, were possible: music, broadcasting, newspapers, and film are all good examples. Cultural production is reliant on large markets, and therefore global cities are popular. Moreover, the association with power and taste-making means that curatorial and fashion trends are found in cities.

It was expected by many that digitization would free producers, especially cultural producers (who were the archetypical symbolic analysts), from the constraints of place, as their work could be done anywhere. The “death-of-distance” hypothesis was much discussed, but geographers discovered in practice that digital producers like to be in close proximity, even though they did not need to be, in terms of the physical transfer of goods (Pratt 2000). As has been suggested, a well-populated and diverse cultural ecosystem is valued by the cultural industries and their workers; moreover, these ecosystems are rooted (but not exclusively located) in places. Cultural industry clusters have increasingly been a topic of analysis of economic geographers concerned with networking and knowledge exchange (Bathelt and Gluckler 2011). The attraction to place is associated with the social and economic reproduction of labor (people have to live and work in a place), but also with the local cultural discussions. Significantly, a taste community is a valuable asset in minimizing risk regarding new products. Quite simply, the notion of a “scene,” which is a community of practice and a social institution, has led to the clustering of cultural producers in cities, often within one or two streets, a hypercluster. Such clusters are stereotypically urban in the sense that the agora of the city, the in-betweenness, is the material and social manifestation of the ecosystem.

The cultural economy has a distinct geography, one that is clustered in cities and in the Global North. The reason for the North/South division is not cultural; rather, it is the global organization of cultural production that is dominated (but not exclusively) by major corporations based in the North which control intellectual property rights. Thus culture has some similarities to global commodity chains, which span the globe but filter control and value added to a small number of actors. It is particularly challenging in the South that colonial and imperial trade and distribution systems provided a first-mover advantage for the North – one that is echoed by the expertise in intellectual property control. These legal issues (access to markets and protection of property) are a significant challenge today.

Governance

Since the mid-1990s, the issue of governance of the cultural economy has grown in economic as well as social and cultural significance, and the modalities of its exercise have become more complex. Previously, the field was occupied by cultural policy and the regulation of culture. The former was usually associated with the public provision of merit goods, or the conservation of heritage, a function that is now increasingly associated with tourism and with national identity. Regulation was exercised through the control of monopolies (economic and of communication) and censorship.

In the 1990s, an instrumentalization of culture took place in many jurisdictions. Culture was mobilized to fulfill social and economic ends, examples of which range from improved mental health to community cohesion to urban regeneration. Previous iterations of cultural policy were
instrumental in terms of nation-state politics and cohesion, but this new wave explicitly used culture for specific ends. The objective was not cultural excellence or even participation; these objectives were held by core cultural policy. Governments lost faith in the justification of art for art’s sake or for the public good, and instead justified it by their other contributions. In this sense, “creative city” policies sought to achieve instrumental aims (Florida 2002) with what they hoped would be politically popular means. As many examples have shown, the investment in opera houses and art galleries to attract foreign direct investment did not benefit the whole population, especially the poor.

In the 2000s debate began about how policy could address the intrinsic value of the cultural economy in either its economic or its cultural value dimension, that is, create jobs in the cultural economy and supporting businesses and/or those sustaining local cultural value systems and communities of practice. The how of this new governance continues to be a challenge for the reasons set out earlier in terms of information about the sector and an understanding of its organization and processes. The emergent concerns are threefold: how to support and sustain cultural intermediaries; how to achieve local capacity building (education, training, archiving, and institution building); and how to gain foresight and information so that practitioners and policymakers can develop and manage strategy. Alongside these new approaches, however, more traditional instrumental approaches still obtain: regulation of monopolies; use of cultural attractions to facilitate foreign direct investment in financial services; and traditional public good policies.

The Great Recession of 2008 which impacted Europe and North America particularly was used by most governments to justify further cuts in public expenditure; these hit the cultural sector, already under pressure, with full force. Thus, in terms of finance and justification, the mid-twentieth-century settlement on culture as a public good managed by welfarist state economic logic has more or less ended. In its place is a complex pattern of governance of the cultural field, a field that has grown significantly and that has, in some cases, become a major player in local economies. The challenge to understand this delicate balance of relationships that constitute cultural ecosystems will continue to be an important field for social scientific research.

**SEE ALSO:** Creative class; Creative class culture; Industrial agglomeration; Intermediaries; Knowledge-based economy; Precarious work

**References**


Further reading


Marked by several distinct theoretical and methodological positions, cultural geography is often recognized as one of the most vibrant as well as contested subfields in human geography. Indeed, it is not an overstatement to say that cultural geography, with its widespread popularity, has frequently been accorded equal footing with human geography as a whole. That cultural geography has come to represent human geography in its entirety should come as no surprise, especially in the eyes of nonacademic audiences who tend to equate the study of human geography with learning about geographically differentiated and often exotic cultural traits. Such close alignment between cultural and human geography is also evident in titles of popular textbooks that often place “human” and “cultural” geography side by side. Even cognate subfields such as social geography are sometimes seen to be subsumed within cultural geography as their research agenda and theoretical approaches become increasingly indistinguishable to the point that the two sub-disciplines are often referred to conjointly as “social-cultural” geography.

Yet it is also interesting to note that there is no single and unequivocal definition of “cultural geography,” so much so that its flagship journal (formerly *Ecumene*) was suitably renamed *Cultural Geographies* in the plural to reflect the diversity and polyvocal nature of the subfield. There are also multiple competing views on what is deemed appropriate for the study of cultural geography in terms of its scope and method of inquiry. In fact, cultural geography means quite different things to different people depending on their specific encounters and subdisciplinary convergences (Longhurst 2007). From postcolonial and ethnic studies to research on feminism and the consumption of popular cultures and new media, cultural geography has left its imprint on both theoretical and methodological debates. It is thus not an exaggeration to say that the intellectual reach of cultural geography has clearly gone beyond human geography per se and has indeed crossed over to the academic terrains of a wide range of disciplines in the social sciences and humanities.

Notwithstanding its vibrant diversity, it will be surprising to many that the subfield has recently been said to be undergoing something of an intellectual crisis, with skeptics readily pointing to the growing disenchantment with the contemporary relevance and value of cultural geography for the wider society. Indeed, contemporary cultural geography has often been criticized for being too fragmented and diffused, while for others it is derided for being narrowly microscopic and “navel-gazing” in its orientation. Whether justified or not, such critiques actually reflect the inherent ideological tensions and schisms between what might be termed “old” versus “new” cultural geographies (more on this later). Arguably, the fragmented and diffused nature of the subfield has also been seen by some practitioners as having forestalled the potential development and contribution of cultural geography at the turn of the century. Not
least, for Jane Jacobs (2006, 1), the unfulfilled promise of cultural geography has led her to question its continued existence. As she notes:

A mere decade ago cultural geography was seen as an analytical frame that could promise not only a productive, but also a necessary, reshaping of geographical scholarship. Now it seems we can’t decide if we want this subfield to be dead or alive!

If indeed cultural geography is undergoing an existential crisis and disciplinary anxiety (Nash 2002), part of the problem may be traced back to the vexed old question of what constitutes “culture” and what work culture accomplishes. As Don Mitchell (2000) has usefully pointed out, we need first to understand what culture is and what it isn’t, or if it even exists in the first place! Beyond challenging the ontological presupposition and the often taken-for-granted view of culture, we need to unpack the conceptual black box of culture to examine “how what gets called culture is itself socially constructed through all manners of contests and cooperations over the materials – places, jobs, pictures, foods, art, histories, ethnicities, sexualities, etc. – that make up our lives” (Mitchell 2000, 12).

Revisiting the question of culture

Often said to be one of the most complicated words to define in the English language, the notion of culture has been keenly debated by geographers and cultural theorists alike. From Clifford Geertz to Raymond Williams, culture has been given various intellectual interpretations from the purely idealist and phenomenological to the staunchly materialist. Taken at face value, culture is often used to represent a whole litany of ideas and behavioral traits, and may refer to many things including cultural beliefs, value systems, lifestyle practices, and even material artifacts. In more popular conceptions, culture can be used to signify “high culture” expressed through the fine arts that are consumed by the upper class, but it also has a more democratic notion that refers to the quotidian way of life of ordinary groups of people. In fact, it is not overstating the case to say that the definition of culture can be an intellectual muddle and can indeed mean all things to all people.

Nonetheless, cultural geographers have attempted to come to terms with the complex usage of the concept of culture by providing a working definition that is appropriate for the study of cultural geography. Several thematic distinctions have been invoked to describe culture as the spatial distribution of things, as a way of life, as symbolic meaning, as a form of practice, and also as the manifestation of power relations. More importantly for cultural geographers, the fundamental concern is how we think “spatially about culture” while also thinking “culturally about spaces” (Anderson et al. 2003).

Despite the myriad ways that culture has been defined and used, cultural geographers have all underscored the importance of “taking culture seriously” in geographical analysis. At the same time, some geographers have also warned against using “culture” too loosely or simply treating it as a shorthand for describing place specificities and cultural variations seen on the ground. Undeniably, the call to put culture at the forefront of intellectual or geographical enquiry has also precipitated the rise of the “cultural turn” in human geography and the social sciences and humanities more broadly. To the extent that culture is now everywhere (and yet nowhere), the cultural turn has attracted both supporters and detractors. In particular, it has been argued by many scholars that, with the permeation of the cultural agenda into practically every sphere of research and
everyday life, cultural geography as a subfield has been hollowed out and has become something of an empty signifier for *all things cultural*. To put it more bluntly, it may be said that cultural analysis and the mobilization of “culture” as the unit of analysis can no longer be the preserve of cultural geographers alone given that everyone now has a stake in culture as a *researchable* object.

If cultural geographers have no proprietary claim over the use of the concept of culture, and given its diffused and fragmented nature, the corollary to this assertion is what distinct contributions cultural geography can offer as a unique subdiscipline. More pointedly, can cultural geography as an academic enterprise stand on its own? Rather than being consumed by attempts to (re)claim the concept of culture, or even attempt to provide an unequivocal definition of culture or cultural geography, it may be more fruitful to focus on the diverse practices of doing cultural geography on the ground. After all, as Shurmer-Smith (2002, 3) argues compellingly, “culture is practiced and not owned” and we should think of it as “adverbial” – that is, “what people do, not what they have, and they keep doing different things in different ways, with different other people all of the time.”

We can perhaps make an epistemological distinction here between culture as a practical idea as opposed to its being a static category of analysis. This distinction is an important one especially when the meaning of culture is so protean and polymorphous that it makes little sense to try to pin it down in a definitive and substantive manner. To the extent that culture can be seen as a practical category or a conceptual toolkit that can be mobilized to work on a number of different geographical problems and issues, it is timely to review the different approaches and methods that have been deployed over the years. Before proceeding further, it is important to qualify at the outset that these different approaches should not be seen as radical breaks or as moving in a linear trajectory but as coevolving. It should also be noted that, while it is often difficult to neatly divide the contemporary development of cultural geography according to distinct chronological time frames, a historiographic narrative is nevertheless useful to provide some broad sketches of the evolution of the subfield since the twentieth century. And, as with any historiographic accounts, they constitute the collective memories (however contested) and self-scripted narratives that together help to shape the evolving identities of the subdiscipline.

### Cultural geography in action: paradigm shifts and turns

Contemporary cultural geography in the anglophone academic world traces its origin to Carl Sauer and the Berkeley School. Similar to the Vidalian tradition of the French regional studies, a key intellectual hallmark of the Berkeley School was its strong advocacy for an interdisciplinary perspective that combines insights from anthropology, history, and natural science, as well as a general disdain for positivistic social science. With an emphasis on cultural ecology and humans’ role in transforming the physical environment, Sauer’s Berkeley School sought to bring culture squarely into the agenda of geography’s intellectual project and also to provide a corrective to the prevailing paradigm of environmental determinism that was dominant in the early part of the twentieth century. If there is an emblematic work that represents the Berkeley School, it has to be Sauer’s landmark publication “The Morphology of Landscape,” in which he argues for the study of cultural landscape as an entity that is forged from a natural landscape by a culture group (Sauer
CULTURAL GEOGRAPHY

For Sauer, the agency of culture is often imprinted on land which acts as the natural medium; the resultant "cultural landscape" is often thought to be the end product that bears testament to the agentic forces of culture acting on a natural area.

Insofar as the Berkeley School of cultural geography shares several similarities with the French regional school (although Sauer himself was influenced more by the German intellectual tradition, especially the works of Herder and Hettner), the Sauerian-inspired approach has also been criticized for its analytical fixation on empirically oriented studies that focus on historical material artifacts and a tendency to reify the cultures of dominant groups. By the late twentieth century, however, what is deemed the traditional Berkeley School of cultural geography was roundly challenged by an emerging group of "new" cultural geographers led by geographers such as Peter Jackson, Denis Cosgrove, and James Duncan. Expressing dissatisfaction with the conservative conception of culture as representing the values of dominant groups in society, proponents of the new cultural geography were vociferous in their attack on Sauer's alleged treatment of culture as a "superorganic" entity, among many other objections. Unlike Sauer, whose main concern was to discover (in a more formal sense) the culture of particular social groups in the form of their unique customs, arts, antiquarian lifestyle, and so on, the new cultural geographers were more concerned with how culture is implicated in struggles over social differences such as class, ethnicity or "race," and gender. To the extent that the Berkeley School is often criticized for its static mapping of cultural artifacts, the new cultural geography may be described as a self-conscious response to the upwelling of cultural politics within contemporary social life in many globalized Western cities that have witnessed the surge in diverse population groups marked by class, gender, and ethnic differences.

If Carl Sauer's (1925) "The Morphology of Landscape" sought to place the role of human culture at the forefront of (cultural) geography's intellectual project, Peter Jackson's (1987) Maps of Meaning represents an attempt to revitalize cultural geography and to retheorize the field. At the core of the new cultural geography is an attempt to displace the traditional (Sauerian) "unitary view of Culture" in favor of "a plurality of cultures, each of which is time- and place-specific" (Cosgrove and Jackson 1987, 99). The new cultural geography argues for a new spatial constitution of culture which is best encapsulated in the following programmatic statement: "Culture is not a residual category, the surface variation left unaccounted for by more powerful economic analyses; it is the very medium through which social change is experienced, contested and constituted" (Cosgrove and Jackson 1987, 95).

In tandem with the retheorizing of culture as such, the idea of cultural landscape has also been reworked to recognize it as "a sophisticated cultural construction: a particular way of composing, structuring and giving meaning to an external world whose history has to be understood in relation to the material appropriation of land" (Cosgrove and Jackson 1987, 96). In a nutshell, the cultural ecology of landscape and human–environment interaction soon gave way to ideas on the contested representation and symbolic meaning of landscapes that soon became the new buzzwords in cultural geography.

It is equally important to point out that the critiques leveled at the Berkeley School are highly partial, resting on the selective targeting and oversimplified reading of certain programmatic statements while ignoring the broader remit and influence of the Berkeley School. The challenges of new cultural geography in many
ways point to the ideological clashes between Sauer and Jackson. As Olwig (2010, 177) noted: “Sauer and Jackson were like oil and water, they didn’t mix, and this, unfortunately worked to polarize cultural geography.”

To the extent that such “orthodox” narratives that pit “traditional” against “new” cultural geography have become almost the hegemonic account of the subdiscipline, we risk ignoring other significant parallel developments in the field. Since the 1970s, humanism as a philosophy had also found its way to a receptive pool of audience among human and cultural geographers who were disenchanted with the economic determinism and dehumanizing logic of the “quantitative revolution” which was steadily gaining ground during that time. For these groups of scholars, the humanistic movement is seen to both endorse as well as challenge existing paradigms in cultural geography. While the humanistic approach is critically aligned with the new cultural geography in that they both reject the superorganic conception of culture in favor of the view of culture as actively constructed by individuals, historically inflected forms of humanistic scholarship were, at the same time, also seen as compatible with the more traditional forms of cultural geography with its focus on antiquarian cultures and civilization histories and origins. In particular, humanism, broadly speaking, has provided cultural geography with the academic license to place the study of people at center stage, with an emphasis on their distinctively human agency, meanings, values, and purposes.

If humanism has laid the important foundations for a revitalized cultural geography, cognate humanist philosophies such as idealism, phenomenology, existentialism, and pragmatism have further animated cultural and humanistic geographical research, giving rise to themes such as “sense of place” (Tuan 1977) and “lifeworld” (Buttimer 1976) and at the same time injected new epistemological as well as methodological innovations through studies on landscape iconography, mental maps, environmental perception, and everyday geographies. It may be observed that a key idea driving humanistic research for many cultural geographers is how diverse human experiences and conditions are being shaped by the spaces that people inhabit, and in turn how humans transform the geographical world through various place-making endeavors.

It must, however, be noted that, despite the close association between humanist geographers such as Yi-Fu Tuan with the Berkeley tradition, humanistic work, strictly speaking, did not set out to reform cultural geography at the outset. As is often pointed out, both humanistic and cultural geography could be said to be united in their response to a common ideological adversary—quantitative spatial analysis in the 1960s and 1970s.

Nonetheless, if humanism provided the much needed intellectual booster for cultural geography on both sides of the Atlantic by the 1980s, the Marxist critical approach also served to further invigorate and “radicalize” cultural geography. As is laid bare in his “call-to-arms” manifesto “Towards a Radical Cultural Geography,” Cosgrove (1983, 1) averred that “Both Marxism and cultural geography commence at the same ontological point.” Arguing for a synthesis between Marxist theory and cultural geography, Cosgrove drew on the works of Raymond Williams and E.P. Thompson among others to push for a rapprochement between historical materialism and cultural analysis. Applying such historical materialist understanding to cultural concepts such as landscape, Cosgrove (1984, 2), in his book *Social Formation and Symbolic Landscape*, further argued that

the landscape idea does not emerge unprompted from the minds of the individuals or human
CULTURAL GEOGRAPHY

groups ... historically and theoretically it is unsatisfactory to treat the landscape way of seeing in a vacuum outside the context of a real historical world of productive human relations.

Needless to say, Cosgrove was not alone in this endeavor, given that David Harvey had earlier argued for culture to be understood as inseparable from capital production and circulation itself. While it does not reduce culture wholly to the circulation of capital, a radical cultural geography approach takes on the uneasy relationship between culture and economy by questioning how culture is spatially produced and whose interests and benefits it represents. Eschewing a totalizing “cultural explanation” for a more concretely political economic approach, Marxist cultural geographers stress that cultural analysis is best operationalized by foregrounding the unequal social structures that shape the production of meanings in social life. To put it more starkly, “culture is politics by another name” (Mitchell 2000, 294).

A whole generation of critically minded cultural geographers responded to the call to radicalize cultural geography by embarking on various projects that interrogate the structural imperatives and power relations behind the production of spectacular landscape and the urban built environment. From schools, religious institutions, ethnic enclaves, iconic skyscrapers, to shopping malls, power politics and “culture wars” are read into the production of these “landscapes of power” (although it has to be noted that few of these works are strictly Marxist in orientation). In other words, radical cultural geography went beyond a strictly Marxist mandate to engage more broadly with critical cultural studies that drew inspiration from neo-Marxist social theories such as those emerging from the Frankfurt School. Revisiting the work of Theodor Adorno and Max Horkheimer, many cultural geographers also critically examined the roles of the cultural industries and commodities in contemporary society, which are seen as bearers of cultural codes that mediate the production and consumption of everyday life. On the latter, researchers have stressed that cultural resistance is as much a part of everyday life and that people are not “passive dupes” who uncritically internalize the hegemonic meanings encoded in the landscape. A typical research strategy with such an approach is to uncover how ideologically charged landscapes are being read, decoded, and appropriated or even subverted by the masses. Searching for “slippages” and “contradictions” in the landscapes became the dominant leitmotifs that characterized the research agenda of cultural geography in the 1990s, although it should be pointed out that “geographies of resistance” rarely ever correspond neatly to “geographies of domination,” and vice versa.

If there is a concept that distinguished the new cultural geography from its predecessors, the idea of landscape as text is a strong contender. In both epistemological and methodological terms, the textual metaphor of the landscape, developed with insight from poststructural linguistic and literary theories, cannot be more different from that of the Berkeley School. Whereas landscape in the traditional sense was often seen as reflecting the cultures of social groups which could be readily apprehended through empirical observation, the “textual turn” saw landscape itself as constitutive of culture through a signifying system that helps to reproduce power relations in society. The latter suggests that different textual communities are at work to mobilize their own readings and interpretations of the landscape or text, and that struggles over different textual interpretations can have real material consequences (see Duncan 1990). Notably, the notion of landscape as text not only opened a novel perspective to landscape studies that combines Marxist cultural materialism with different ways of seeing,
but it also provided a common vocabulary for cultural geographers, literary theorists, and others to engage in critical discursive analysis of landscape representation and cultural politics.

Another observation that could be made about the new cultural geography was the shift in empirical focus from farms and villages to cities and urban places. Arguably, if traditional cultural geography has an unmistakably rural bias, then the new cultural geography from the late 1980s onward has become urbanized, with much of the research output dedicated to the interrogation of urban cultural politics. This period also coincided with the rise of entrepreneurial urban regimes in many Anglo-American cities that provided fodder for research, with topics ranging from urban place marketing to the interrogation of urban cultural politics. This period also coincided with the rise of entrepreneurial urban regimes in many Anglo-American cities that provided fodder for research, with topics ranging from urban place marketing to the contested cultural and spatial logic of postindustrial cities.

Beyond urban cultural politics, another strand of critical cultural geography work has been the focus on the “hidden” labor that goes into the production of landscape and commodities. Don Mitchell (2003), for example, in his penetrating analysis of the political economy of California’s agricultural landscape, argues that the bucolic scenery of the rolling fields and strawberry farms is sustained by the “dead labor” that is embodied in the production of commodities such as the strawberry, as well as the spatial injustices that (re)produce both the visible and the hidden landscapes around us. Fundamentally, both the real and the symbolic meanings of the landscape are seen as a function of who has the power to represent that landscape.

Since the 1980s, interdisciplinary inspiration from a wide range of thinkers such as Michel Foucault, Michel de Certeau, Bruno Latour, as well as Gilles Deleuze and Felix Guattari, have provided cultural geographical scholarship with a further “radical turn” through the postmodern challenge. Responding to the challenges of postmodernism requires cultural geographers to problematize the use of discrete and bounded ideas such as cultural hearths, landscapes, or regions and to attempt to deconstruct these taken-for-granted spatial concepts by interrogating their internal diversity and heterogeneities. Such a radical turn in postmodernizing cultural geography also extends to demanding a new attitude and sensibility toward multiple forms of otherness, including marginalized social groups and communities based on ethnic minority status, sexual deviance, or disabilities, as well as disenfranchised groups of working-class, urban poor, migrant workers and also women, elderly people, and young children. Collectively, these diverse groups of people are often absent in research programs or are seen as a lesser counterpart to the dominant white middle-class heteronormative (Western) society that frames academic discourse and analysis.

In other words, if new cultural geography sought to displace the uniform view of culture of dominant groups, the postmodern turn went further to give voices to diverse subaltern groups, including feminists, those who are disabled, ethnic minorities, postcolonial subjects, and to entrench cultural politics at the forefront of cultural geography’s research agenda. Indeed, the “politicization” of new cultural geography has practically become a quintessential feature in almost every research program spanning the period. What unites such critically informed cultural geography research seems to be the commitment to the rather formulaic conception that cultural politics is ubiquitous since nothing is politically neutral, and that the cultural landscape we inhabit must necessarily reflect and manifest the contestation of cultures even in the most mundane and unremarkable of settings such as the home, school, workplace, shopping mall, theme park, and so on.

Nonetheless, while the new cultural geography, with its postmodern critical inflections, had
CULTURAL GEOGRAPHY

become academically fashionable by the end of the twentieth century, its novelty soon wore out and before long even new cultural geography itself had fallen out of fashion and ceased to function as anything more than a convenient descriptor for demarcating what was seen as the “old” and the “new,” or as merely a marketing tagline for academic publishers. While criticism and misgivings over new cultural geography have been well documented, it is worth pointing out that a major disgruntlement lies with an overemphasis on the politics of representation. Many cultural geographers, for example, were concerned that the knowledge generated in cultural analysis is no longer capable of doing anything productive apart from just dabbling with the discursive politics of representation. To put it more bluntly, as Neil Smith (2000, 27) charged: “cultural geography, perhaps human geography in general, has downgraded the importance of fieldwork and has too often come to think of empirical research as a question of perusing texts – magazines, adverts, movies, landscapes – for representations of this or that.”

For the harshest critics, the whole enterprise of cultural geography in the 1980s and its revival was largely built on an intellectually stultifying preoccupation with the politics of representation that valorizes the symbolic and rhetorical at the expense of actual practice and expressive life.

In response to what has been termed the “death of representation,” a body of work rallying around “non-representational geographies” (Thrift 1996) has emerged to place the multisensory practices and experiences of the everyday life at the forefront of geographical analysis over its abstract representation. For this group of scholars, new cultural geography and the cultural turn is seen as overly theoretical, contemplative, and even too culturalist. Indeed, if traditional cultural geography was criticized for its super-organic conception of culture, new cultural geography is similarly seen as locked in a totalizing view of culture and its endless representation.

Nonrepresentational scholars, instead of representing the world through abstract language and theories, make a conscious and deliberate attempt to focus on the performative aspects of the multisensuous corporeal world that is lived and experienced through bodily spatial practices. Strategically speaking, focusing on the creative and performative aspects of cultural geography has also been a way to elevate the relatively low public profile of the subdiscipline as it begins to make inroads into more popular public art forms such as photography, videos, dance, and creative writing. All in all, more-than-representational scholarship may be characterized as an approach to the study of the “restless” cultural world, which is primarily concerned with meanings that are produced and enacted through mobile activities (walking, dancing, climbing, cycling, driving, etc.). In essence, movement and the mobile structuring of everyday life are seen as the core of what constitutes being purposive human agents.

Another contribution of the more-than-representation work has been to draw attention to the “mobility turn” and the “new mobilities paradigm.” While geographers have always been interested in the study of people and things on the move, new mobilities studies see mobility as more than just an empirically observable brute fact out there, (i.e., the movement from point to point) – as an embodied spatial practice that is socially produced and conveyed through a diverse array of representational strategies that endow it with social-political meanings (Cresswell 2006). In particular, the new mobilities paradigm makes the case for paying close attention to the bodily practices and performance of mobility, and makes mobile embodied practices central to how cultural geographers study and experience the world. Mobile practices such as walking,
running, and flying, and the spaces they traverse, are often inflected by complex geographies and histories, and are seen as embroiled in a whole range of processes from political-economic to cultural and ethical concerns. By highlighting the diverse variety of mobile experiences that characterize a world in motion, the new mobilities paradigm can also be credited with building connections between cultural geography and other subdisciplines such as in transport studies, as well as tourism and migration research.

Insofar as nonrepresentational theory has shone a spotlight on mobile spatial practices and the materiality of social life, Sarah Whatmore (2006) further argues for cultural geography’s “materialist recuperation” and calls on geographers to attend to the more-than-human geographies we inhabit. This entails several analytical shifts that redirect our mode of inquiry from discourse to practice, from meaning to affect, from the politics of identity to the politics of knowledge (Whatmore 2006, 603–604). But, more importantly, the more-than-human call challenges geographers to venture beyond the indifferent stuff of a world “out there” and to signal a return to the “livingness of the world” that is articulated through the corporeality and materiality of the human existence. To be sure, while this commitment to a more-than-human cultural geography has been to steer away from a purely anthropocentric position, it does not exclude humans at the same time and instead views humans/animals and technological devices as agents provocateurs in the “co-fabrication of the socio-material worlds” (Whatmore 2006, 604).

Beyond the purely anthropocentric concerns, the more-than-human approach also resonates with Wolch, Emel, and Wilbert’s (2003) call to “reanimate cultural geography” by bringing the “animal question” back to the forefront of geographical research agenda. This is not entirely a new research endeavor given that human–environment and human–animal interaction has long intellectual roots in geography from pre-Darwinian zoogeography to Carl Sauer’s pioneering work on the role of animal domestication and the evolution of cultural landscapes, to Yi-Fu Tuan’s study of the dominance and affection that underscores the power relationship in pet-keeping. Drawing inspiration from social theory and responding to new cultural geography, contemporary study of animal geographies departs from earlier approaches by radically reconceptualizing culture, nature, and animal subjectivity as well as having a strong commitment to environmental ethics and the role of animals in the moral landscape.

In tandem with the focus on embodied practices and materiality, cultural geographers have also called attention to the role of affect and emotion in mediating the spatiality and temporality of everyday life. Beseeching geographers to pay greater attention to how the geographical world is at once being constructed and experienced through human emotions of pain, anger, happiness, feelings of affection, and so on, Anderson and Smith (2001) have led the call for an “emotional turn” in geography that recognizes the complex range of human emotions and sensations (anger, grief, happiness, affection, etc.) as important and valid ways of knowing, being, and constructing the geographical world; however, the apparent lack of attention to emotions by cultural geographers needs to be qualified. After all, the geographical literature is already replete with emotionally laden concepts such as sense of place, topophilia, or place alienation. Nevertheless, it is true that, at least on the surface, the discipline of geography as it has been represented historically tended to be portrayed as a rational and emotionally detached science that eschews any unquantifiable objects of analysis such as the human emotions. In many respects, such a scientific view of the geographical academic
CULTURAL GEOGRAPHY

discipline is largely a product of the Western Enlightenment tradition, rooted in Cartesian and Stoic philosophies that are often suspicious of the ineffable nature of fickle human emotions. But if, indeed, emotions do matter in geography, the emotional turn will demand an epistemological shift to take human emotions into account in the production of geographical knowledges beyond their usual visual or textual registers.

In many ways, the focus on emotions and bodies has paid intellectual dividends by reinvigorating and widening the scope of cultural geography, which has previously been seen as being more interested in landscape interpretation and textual analysis than in real people. Arguably, this shift from the purely representational to the messy materiality of bodies, emotions, and sexual differences offers a new discursive space for feminist geographers, who have erstwhile been apprehensive about being aligned with cultural geography per se.

Clearly, this call for emotions to be taken seriously in human/cultural geography has not gone unheeded, judging by the burgeoning literature on affectual and emotional geography, although Pile (2010, 17) has rightly cautioned against the uncritical importation of “an ever-expanding shopping list of expressed emotions,” and of taking “an increasingly cognition-centred, humanistic and romantic view of expressed emotions,” without ever reflecting on why and how emotional geographies matter in the first place.

New cultural geographies, new methodologies?

The new cultural geography has opened up a fresh intellectual terrain and even brought about several paradigm shifts and turns, but several geographers have pointed out that much of its contribution seems to be theoretically driven and offers little serious consideration to the issues of methodology and fieldwork innovation beyond the narrow range of qualitative techniques of interpreting and decoding landscapes. More seriously, Thrift (2000, 3) has lamented the lack of methodological innovation and the conservative approaches that cultural geographers bring to the table: “how wedded they still are to the notion of bringing back the ‘data,’ and then re-presenting it (nicely packaged up as a few supposedly illustrative quotations), and the narrow range of sensate life they register.”

There is no doubt that, as cultural geographers move forward with a renewed cultural geography, it is crucial (for the subdiscipline) to be more explicit and articulate about methodological choices and assumptions. Furthermore, if contemporary cultural geography is often observed to be searching relentlessly for new terrains of research from mobilities to human emotions, such vigor also has to be equally matched by bolder methodological innovations. Not least, nonrepresentational currents of thought in geography have already challenged much of the conventional methodological toolkit available for geographical fieldwork – in particular, participant observations, in-depth interviews, and focus groups – which are seen to accentuate the contemplative and discursive aspects of geographical fieldwork and inquiry.

Recently, practitioners of cultural geography have lobbied for greater methodological pluralism by juxtaposing traditional qualitative research methods that rely mainly on discourses and textual representation with more experimental practices such “go-along ethnography,” walking, participant photography, and so on in order to augment the multisensory embodied experience and the expressive modality of the research subject at hand. Thrift (2000, 4–5), in particular, has called for greater openness as well as “cultural and practical generosity” in
learning from practice found in street theater, dance and music therapy, performative writing, and creative storytelling, which help to “co-produce the geographical world that we live in.” As Jacobs (2006) warns, however, such innovation and experimentation should not detract from the foundational methodological commitments of cultural geography, which in her view means attending to the relationalities between material and immaterial, small and big, mundane and exceptional, or global and local (Jacobs 2006, 22). Using the modernist skyscraper as an example, Jacobs demonstrates how cultural geographers can profit by avoiding the fixation on representational geographies and to take on board a “semiotics of materiality” that pays attention to the “relational assemblages of human and non-human that work to ‘make’ a building event” (Jacobs 2006, 2).

Also, while nonrepresentational theory tends to privilege prediscursive and noncognitive practices, it “does not seem to allow room for considering visual and textual forms of representation as practices themselves,” and appears to “offer more theoretical guidance for considering practices over representations rather than strategies for bringing them together” (Nash 2000, 662). For Nash, it is vital that proponents of nonrepresentational theory do not totally abandon the study of representation and texts, which make up the core tradition within cultural geography. As she notes, “Exploring practices, performances, texts, objects and images together rather than abandoning the knowable for the unknowable may be less theoretically ambitious than ‘nonrepresentational theory’ but it is also more politically effective” (Nash 2000, 661).

Contemporary cultural geography has certainly come a long way, and has precipitated various paradigm shifts and conceptual turns within the subfield, but much of these intellectual developments remain firmly entrenched in the anglophone academia, bypassing much of the non-English-speaking world. Such omission of non-anglophone voices is particularly jarring if we are to take seriously contemporary cultural geography’s mission to “decolonize” and “decenter” Western academic discourses. Many of the earliest work in cultural geography can actually be traced to non-anglophone traditions in Greco-Roman as well as Chinese and Muslim scholarship. Indeed, as many scholars observe, it would not be surprising to find cultural geographers (or their equivalents) as part of the intelligentsia of expansionist empires of the ancient world such as the Greek and Roman civilizations, the Tang dynasty, or the Ottoman empire.

In recent years, criticism of “faux-cosmopolitanism” in Anglo-American geography (Smith 2000, 27) has led some scholars to reflect on their disciplinary practices, and in particular on the role of language. Several geographers have begun to challenge what they perceive as linguistic hegemony, which is often evident in anglophone publications. It has been pointed out that challenging the dominance of English in everyday academic discourse and publication is an urgent task if cultural geographers (in the anglophone world) are serious about transgressing the linguistic boundaries that they have been advocating. If subverting the dominant lingua franca of academia seems to be a way to disrupt and destabilize Anglo-American geography, a wider implication here is to examine how different ways of doing and practicing (cultural) geography on a daily basis – including how we formulate research ideas and theories as well as conduct scholarly activities such as publishing and attending conferences – can all be modified to open up the discipline to a plurality of non-Anglo-American voices as well as “alternative ways of valuing forms of geographical enterprise” (Kitchin 2005, 2).
If the country reports in the flagship journal *Social & Cultural Geography* are anything to go by, the hegemonic status of Anglo-American geography (whether real or perceived) has warranted great concern among geographers. It has been pointed out that anglophone-centric geographical literature has a tendency to reinforce its own dominance through the use of English language as the de facto lingua franca of academia. Nevertheless, as Kitchin (2003, 253) notes, the country reports that appear in the journal hope to serve as both an *intellectual* and a *political* project that “seeks to disrupt and destabilize the prevalent trend towards English and Anglo-American hegemony in the international production of geographic knowledge.”

Attempts to bring social and cultural geographers from around the world into dialogue should not be conceived as an attempt to bring the work of “geographers at the margins” to the Anglo-American center, or worse, to co-opt non-Anglo-American geographers into the fold of “Anglo-American ways of knowing, interpreting and writing” (Kitchin 2003, 253).

For that purpose, the commissioned country reports on social and cultural geography in the respective regions are published in both English and the contributor’s native language. Indeed, such a practice is now widely adopted in many major geographical journals such as *Area, Annals of the Association of American Geographers, and Transactions of the Institute of British Geographers*, where article abstracts often appear in more than one language, although the main manuscript remains published solely in the English language. While such a linguistic gesture toward including “native” languages in the journal abstracts is certainly a welcome addition that gives the journals an “international” appeal, it remains to be seen how far such endeavors go in terms of bridging the gaps between (non-)anglophone academic worlds.

Part of the difficulties in engaging with non-anglophone scholars goes beyond simply the linguistic divide. As various country reports in *Social & Cultural Geography* seem to suggest, the dearth of social and cultural geography works that engage with an international (English) audience can equally be attributed to the exclusiveness of non-anglophone academic communities and their general mistrust and/or disdain of foreign influences, and to Western intellectual trends that are often perceived as threats to the integrity of local traditional scholarship. Even in France, despite the popularity of works by French theorists such as Lefebvre and Foucault in Anglo-American geography, much of French social and cultural geography literature remains rather insular and patchy. A huge intellectual gulf remains between English-language geography and French geography. Indeed, part of the reason for the deficit in cultural geography in some places hints at much broader structural conditions and the state’s role in shaping geography as an academic discipline. This role is especially pronounced in countries such as China which have experienced massive social and political transformations in recent decades. Against the changing political and ideological backdrop in China, several commentators have shown how cultural geography (alongside various subfields in human geography such as political and social geography) had been under severe state sanctions owing to their perceived Western bourgeois influences and became forbidden areas of research in the 1960s and 1970s. It was not until after the economic reforms in the late 1970s and early 1980s, with the gradual opening up of the country, that Chinese human geographers began to engage with Western geographical theories and to apply them to understanding many pressing contemporary concerns in China, ranging from
globalization, tourism development, and urban growth politics to environmental protection and sustainable development.

Clearly, there are regional differences as well as internal heterogeneity in how cultural geography takes root and develops in different parts of the world which reflect not only different institutional norms and state practices but also the underlying diverse geohistorical contexts. For instance, social and cultural geography in the fledgling nation-states of Southeast Asia after the postwar period, in the 1950s and 1960s, often focus on the politics and challenges of postcolonial nation-building and globalization (see, e.g., Bunnell, Kong, and Law 2005). While cultural geography in different places has developed and evolved in quite different ways from the Anglo-American pathways, there are also signs of convergences which are often spearheaded by significant individuals who often act as agents of change. In Japan, for instance, several globally oriented scholars initiated contact with anglophone, French, and Italian geography, and also translated key publications from English and French authors, including David Harvey, Paul Claval, Ron Johnston, Jacqueline Burgess, and John Gold. Notably, the *Jim bun Chiri* (*Japanese Journal of Human Geography*), published by the Human Geographical Society of Japan, has recently translated several Japanese articles into the English language to facilitate the cross-fertilization of ideas across linguistic divides. In a similar vein, translations of key cultural geography texts such as Anderson et al.’s (2003) *Handbook of Cultural Geography* have also found a receptive audience among Chinese scholars in contemporary China.

While it is encouraging to see such signs of the bridging of different linguistic divides, we should question if such changes can ultimately disrupt and destabilize dominant anglophone geographical scholarship, or if they merely signal the “reeling in” of native geographers into the fold of “Anglo-American ways of knowing, interpreting and writing” (Kitchin 2003, 253). Kong (2010, 601), in particular, has issued the challenge that the study of China’s cultural geography should avoid becoming just “another case study” illustrating the theoretical logics established in “Western” settings. In this sense, local knowledge production is crucial and “could lead to exciting, perhaps even revolutionary, new theoretical insights, representing a welcome opportunity to unsettle Anglo-American scholarly hegemony.”

As a theoretical rejoinder to dominant Western narratives, a fairly typical response by scholars working in the Global South has been to underscore the uniqueness and particularities of their respective local case studies as a way to speak back to dominant Anglo-American theories. By highlighting the different local contexts, it is often hoped that such nuanced local knowledge and contextual differences might translate into different ways of knowing and theorizing places. To be sure, the production of idiogetic and locally specific knowledge does not preclude the production of more generalizable nomothetic theories. Indeed, the production of close-grained empirical scholarship which has a rich tradition in much non-Anglo-American academic geography is often valued as a starting point to engage with more abstract theoretical work. Yet it would be naive to think that empirically based work can somehow be unproblematically grafted onto existing (Western) theoretical frameworks. Nor should such projects be driven by a self-gratifying need to theorize for theory’s sake.

While cultural geography in the Anglo-American context has evolved from a distinctly Western philosophical tradition, cultural geography in many non-anglophone contexts has developed from a vastly different intellectual environment and genealogy. It is critical to
CULTURAL GEOGRAPHY

Reflect on how non-anglophone cultural geographers can venture beyond merely engaging nominally with international academic literature by being active agents in the production of geographical knowledge and theories within their own intellectual lineage and tradition. For example, we might ask how humanism, which has a long-standing tradition in Asian philosophical thought, can shape contemporary debates in cultural geography. What are the alternative forms of (non-Western) modernities that ground the cultural experiences of the Global South?

There is also a danger of falling into the “Orientalism” trap by mistaking such endeavor as an attempt to recover the lost “authentic” history of non-anglophone intellectual traditions. Rather, the impetus here is to draw on local intellectual resources and scholarly imaginations in order to engage in theorization from within. Necessarily, such an intellectual project will entail going beyond the translation of Western ideas and concepts into native languages, to represent a concerted attempt to provide an alternative epistemological and even ontological framework for understanding the complexities of the geographical world that is invigorated by different positions and views from elsewhere. Arguably, if past decades have seen how “the intermingling of American and British geographers has helped to create sensitivity to the subtle differences in what is understood by culture in both places” (Olwig 2010, 178), it will be interesting to see how the increasing interaction with non-anglophone cultural geographers will further deepen and broaden intellectual dialogue.

Looking to the future

The agenda that was championed by (cultural) geographers in the 1980s is now widely accepted and has even become a hegemonic project to some extent. It is not surprising nowadays to find geographers specializing in political, economic, historical, or even transport geography who consider themselves nominally as cultural geographers. Such a paradoxical situation points ironically to the success of a renewed cultural geography enterprise and its “crisis of overproduction” (Nash 2002). But, more fundamentally, the disciplinary anxieties expressed earlier also reflect the different ways that cultural geographies have been written about, valued, learned, and practiced over past decades, both old and new, traditional and avant-garde, conservative and radical.

While the old versus new divide has often been used to frame the historiography of cultural geography, being overly fixated with such dichotomous representation can lead to a dangerously oversimplified view of the complex evolution of intellectual ideas and to promoting a falsely heroic worldview of (mostly male) academic movers and shakers associated with different schools of thought and the dramatic paradigm shifts that they have engineered. Such highly partial and even sensational accounts also tend to exaggerate the internal unity and coherence of different ideological camps and, further, to encourage antagonistic caricaturing of their respective research programs at the expense of providing a nuanced appreciation of the historical continuities of different intellectual lineages and traditions. Indeed, such historical continuities are already evident in the recent resurgence of interest in human–environment interaction and environmental history, echoing the Sauerian tradition as a result of widespread environmental concerns.

While the cultural turn has revitalized the discipline as a whole and given cultural geography a new identity by shaking off its “old nerdy image,” it has also generated tensions with other subfields and been duly accused of
eclipsing and marginalizing “social” geography (Valentine 2001). What matters here is not only that boundaries between subfields are porous and fluid, but also that developments within cultural geography have unintended spillover effects on other cognate subfields. For Valentine (2001, 170), a real practical concern about the cultural turn “is not so much with the evacuation of the social but rather with the depoliticization of social geography.”

Indeed the lack of an explicitly political agenda is also a critical issue that confronts contemporary cultural geography, with its preoccupation with meaning, identities, and representation at the expense of giving due emphasis to class inequality and even social environmental justice. While geographers have been quick to embrace the latest styles offered by intellectual fashion houses in France, many such “cutting-edge” and “avant-garde” works are jargon-ridden and have little relevance for public policy. Even when such academic works do attempt to speak to important social and economic issues of the day, they are often written in such a convoluted manner that they are difficult to translate into policy-friendly language that can be digested by policymakers and planners.

For the detractors and critics who are eager to take a potshot at cultural geography, the subfield is often an easy target for ridicule because of the increasingly inaccessible corpus of work that is riddled with complex jargon, abstract metaphors, and pompous language which produce quixotic forms of scholarship catering only to insular intellectual peer groups. Whether such accusations are justified or not, Don Mitchell (2014) has warned against what he calls the “complicated fetish” in geography whereby intellectual premium seems to be accorded to ever more complex ways of framing understandings and explanations of the world. To be sure, such critiques are not meant to express anti-intellectual sentiments or, worse, to advocate the dumbing down of geographical scholarship. On the contrary, it is precisely because the geographical world is infinitely complex that researchers and scholars need all the more to make such “hard-won knowledge” more accessible to people rather than forestalling greater understanding and action through unnecessary jargon-laden rhetoric (Mitchell 2014, 126).

As cultural geography progresses into the future, it is also imperative that the subdiscipline continues to demonstrate critical relevance in its scholarship in terms of both the production and the dissemination of knowledge. This is not, of course, to suggest that cultural geographers working in the earlier tradition have been uncritical in their work. Without denying or ignoring the historical specificity of any approach to critical geography, what counts as critical is often itself a product of its time. Any number of approaches in the longue durée of geography’s intellectual development from environmental determinism in the late twentieth century, through Sauer’s cultural ecology, the quantitative revolution of the postwar years, and the humanist and postmodernist movements in the 1980s and 1990s, can rightly be labeled as “critical” geography of its time. They all offer a critique and challenge to dominant approaches and ideologies. An important issue confronting contemporary cultural geographers is how the conditions for the realization of critical geographical praxis have changed in the context of an increasingly globalized and neoliberal (academic) world. Indeed, many cultural geographers now count themselves active members of critical geography groups and online communities, and journals have appeared such as *ACME: An International E-Journal for Critical Geographies*, but just what does being critical mean and who gets to say and define what constitutes being critical in the contemporary age? We also need
to question not just what cultural geographers are being critical of but also what are they being critical for. What higher purpose and goals does critical scholarship serve beyond being a rallying point for motley groups of self-proclaimed leftist/radical/anarchist scholars? Too often, being critical in academic writing takes the rather well-trodden, even formulaic, route of citing the fashionable critical theory of the month or celebrity authors rather than genuinely engaging with an honest critique of hegemonic ideologies, social inequality, and spatial injustices on the ground.

On the whole, the intellectual project of cultural geography is still an unfinished business, as noted by several commentators, and is best characterized as an ongoing project of critical self-reflection, ideological disagreement, and intellectual commitment. It is certainly premature to sound the death knell on cultural geography. It will remain, in the foreseeable future, as a vibrant subfield, albeit one that is confronted with the critical challenges of being innovative while staying socially relevant.

SEE ALSO: Berkeley School; Critical geography; Cultural turn; Culture; Emotional geographies; Humanistic geography; Landscape; New cultural geography; Nonrepresentational theory; Social geography

References


Heritage definitions and classifications

Heritage may be defined as anything of value inherited from the past, to be used at the present and to be passed on to future generations. This simple definition, however, belies the complexities that often attend the heritage selection, interpretation, and presentation process. In multicultural and multiethnic societies, conflicts emerge when questions about whose heritage and who decides what is heritage are asked. In multifunctional and multiuser sites populated by residents, migrants, and tourists, questions about what heritage is and who benefits from it further compound the complexities of heritage definition and utilization.

Heritage may be classified as either natural or cultural. Natural heritage includes physical environments and biological features such as unique geological landscapes, national parks, and coastlines, as well as historic gardens and archaeological sites. The habitats of threatened species of animals and plants, and natural areas of outstanding scientific or aesthetic value, are also included. Cultural heritage refers to the handiwork of humans deemed worthy of preservation and may be either tangible or intangible. Tangible cultural heritage includes buildings, monuments, artifacts, and historic places. Also known as built heritage, it ranges in size from a single monument or museum to an urban district or an entire historic city. Such built heritage is often significant to the archaeology, architecture, science, or technology of a specific culture.

Intangible heritage, also known as living heritage, refers to the practices, representations, expressions, and skill sets and knowledge of communities, which include oral traditions, performing arts, customs, and rituals, and traditional crafts and skills. As these intangibles are passed from one generation to the next, they are constantly adapted to suit changing needs, values, and technological standards of society. Although mutable, intangible heritage nevertheless provides communities with a sense of history, continuity, and identity, bonding them as a unique people.

The 1970s may be regarded as a pivotal period when heritage came to general public awareness. Two key initiatives were responsible, the first being the 1972 World Heritage Convention where proposals on nature conservation and the preservation of cultural properties were brought together in a single document that was agreed on by member states of the United Nations. The convention document was adopted by the United Nations Educational, Scientific, and Cultural Organization (UNESCO) and came to define the World Heritage List. In 2014 the list comprised 981 properties: 759 cultural and 193 natural sites, and 29 mixed properties, spread across 160 countries (the website at http://whc.unesco.org/offers information on World Heritage Sites and their classification).

Another significant milestone was the designation of particular years to celebrate heritages. For example, the European Architectural Heritage Year (1975) was highly successful in promoting public awareness in and pride among European residents of their shared urban heritage. Interest
among tourists and increased conservation practices across the continent inspired further designations such as the European Campaign for an Urban Renaissance (1981), the European Year of the Environment (1989), and subsequently the European City of Culture designations and the millennial project of European Capitals of Culture. Such campaigns not only raised heritage consciousness but also drew attention to the role of heritage as a tourism resource and unifying political tool.

**Geography of heritage**

It has been argued that a “geography of heritage” exists. The study of heritage may be approached from various geographical perspectives, for example, through an understanding of location (where heritage is sited and who it belongs to) or distribution (why heritage is found in a particular site or region and not another). Scale is another dimension in heritage studies. Heritage may be individually owned (e.g., a family heirloom) or take the form of a single site on a national scale (e.g., a monument or historic city) (Graham, Ashworth, and Tunbridge 2000). Heritage is most often conceived at the national scale because national heritage is a critical component for nation-states in consolidating a sense of belonging while differentiating themselves from other territories.

It is also possible to think of heritage at a global scale, particularly with respect to UNESCO’s World Heritage Sites. Such sites are recognized as possessing outstanding universal value from an aesthetic or scientific point of view. They are also bound by international codes on conservation and management practices as defined by UNESCO and its subsidiary organization ICOMOS (International Council on Monuments and Sites). Access to a world reserve of professional expertise consisting of architects, archaeologists, historians, geographers, engineers, and town planners is also available (see the section on “Heritage management and sustainability” later in this entry).

Geographers of different subdisciplines are interested in heritage issues for different reasons. While historical geographers are concerned with unraveling the facts and meanings of places through heritage landscapes, urban geographers focus on conservation efforts and the effects on residents, land uses, and place images. Cultural geographers have explored how heritage is linked to wider debates on representation and identity, while economic and tourism geographers have studied how heritage may be used in the planning and regeneration of towns, cities, and regions.

Finally, some geographers have argued that heritage planning cannot be divorced from the agencies that own the land on which heritage is sited and whose plans inevitably affect the geography of the land. Conservation trusts and development agencies are viewed as geographic agents of change and a worthy aspect of heritage geography research. The works of agencies like the National Trust in England, Wales, and Northern Ireland (established in 1895), Heritage Canada (1973), and the Urban Redevelopment Authority of Singapore (1964) impact not only the physical character of cities and the countryside but also the functional and social character of places. The agencies’ roles in place conservation and change management are well aligned with geography’s concern with the environment and human impacts on it.

**Commodification and the heritage industry**

Beginning with museums and heritage centers in the late nineteenth century, which first
charged visitors entrance fees to today’s historic theme parks, the heritage industry emerged as an outcome of commodification. The term “commodification” denotes the appraisal of heritage as a commodity to be manufactured, marketed, and sold to a paying consumer. While some view commodification as an offshoot of tourism, others see it as a way to stave off declines in the manufacturing sector while boosting pride in national history. Robert Hewison’s (1987) *The Heritage Industry*, for example, attributes Britain’s obsession with heritage in the 1970s and 1980s to the frustrations associated with modern life, including inner city unemployment, coalminers’ strikes, and the ugliness of modernist architecture. Moments of economic decline are often critical in triggering people’s nostalgic impulses and the emergence of a heritage sector.

There exist many forms of heritage commodification. The most common are museums, castles, palaces, historical relics, and national parks, where entrance fees are levied on visitors. Free access contemporary sites may also be regarded as examples of commodification, especially when heritage is deployed for commercial ends. These include mega shopping malls where aspects of the past are recreated (e.g., West Edmonton Mall in Edmonton, Canada, and its replica of Columbus’s Santa Maria), as well as restored waterfronts and festival marketplaces that adaptively reuse historic buildings for contemporary needs (e.g., Singapore’s Clarke Quay, which features retail, leisure, and office uses housed in previously defunct warehouses). Whether heritage is celebrated or commercially exploited is debatable.

Critics of commodification charge that the heritage industry reduces historical sites and cultural resources to trivialities appreciated only for their photogenic qualities. The primary function is usually commercial, while the historical and cultural significance recedes into the background. Supporters, on the other hand, claim that new uses for historic sites and resources ensure their continued existence and greater access to members of the public. Without contemporary uses, old buildings and architecture would have a much slimmer chance of survival. The pros and cons of commoditization are points of debate not only among researchers but among policymakers too.

While it is easy to dismiss commodification outright, the economic benefits of a heritage industry should not be underestimated. Employment brought about by tourism (both quantitative increase in numbers and qualitative changes in the types of jobs created), spin-offs in such sectors as retail, convention, and construction, infrastructural improvements that benefit local residents, and the upgrading of human capital are positive outcomes of the industry. Whether economic benefits outweigh the cultural drawbacks of heritage commodification, and what local residents and other stakeholders feel about the tradeoff, should be considered in heritage planning and development.

**Heritage tourism and authenticity**

Heritage tourism is any form of tourism where the focus is on historical resources. It involves tourists visiting sites of cultural and natural relics, and the industry’s attempt to cater to this demand. With the rush to develop historical resources for tourists, the authenticity of the heritage product and experience is often questioned.

The concept of authenticity assumes a verifiable authentic past to which the contemporary heritage product may be compared. Such a comparison can take different forms, one example being the contrast between handicrafts and trades. Tourists usually purchase handicrafts or souvenirs because of their cultural link to
CULTURAL HERITAGE

Cultural heritage refers to the collective memory of a particular place or people. Handicrafts are created specifically for the tourist market and are often mass-produced. By comparison, trades refer to methods and techniques by a workforce operating independently of the tourism industry. The authenticity of trade items are thus contrasted to the inauthenticity of tourist handicrafts (Urry 1990).

Another area of comparison revolves around preservation and conservation. While preservation entails wholesale restoration with an emphasis on historical accuracy and unchanging customs, conservation refers to adaptive and selective reuse of old buildings and practices in order to cater to contemporary needs. Under conservation, historical sites are rendered accessible and tourist-friendly, with modern amenities including toilets, souvenir shops, restaurants, and staged performances. Under preservation, minimal intervention is permitted and the outcome is usually an unchanged environment or cultural form that is traditionally authentic, even if slightly inaccessible to tourists.

How the past is presented to tourists is the result of interpretation. Often, state authorities and commercial enterprises decide on what aspects of history or culture to offer, and how the heritage product is to be communicated and marketed. Such interpretations are criticized for being highly selective and overly aligned to commercial agendas. For this reason, nongovernmental organizations and heritage interest groups have increasingly asserted their inputs. In what has been described as “histories from below,” alternative interpretations that do not always reinforce dominant political or commercial interests are also being heard (Hardy 1988). Answers to questions of whose history and what heritage are broadened beyond the scope of the elite, allowing a greater range of opinions and a more encompassing view of conservation to emerge.

Heritage, identity, and nation building

The interpretation of heritage goes beyond tourism and is often part of the political process of nation building. Nation-states foster a sense of identity and shared bonds by harking back to a common heritage. Promoting the glories of past regimes, emphasizing shared ancestry of land, language, customs, and values, as well as censoring select aspects of history that are detrimental to the nation’s image are all integral aspects of interpretation and nation building. Defining a national heritage can also help to neutralize competing claims by fractious groups or regions within the nation while differentiating it from other territories.

The search for a common heritage may also be a response to external threats. In an era of globalization, many nations face the threat of cultural homogenization and loss of identity. Without a bedrock of shared history, customs, and traditions, communities run the risk of becoming rootless and soulless. In Singapore, for example, a Committee on Heritage in 1988 defined heritage as offering a social and psychological defense to the nation. As citizens become more westernized, it is feared that many will forget their common ancestry as Asian immigrants and the customs and traditions that bind them as a people. Through conserving historic landscapes and designating national monuments, architecture, and even food and music, heritage is rendered in explicit material form that can hopefully bind citizens as a community and also as a nation (Kong and Yeoh 2002).

As nations evolve, how heritage is interpreted changes too. In countries such as Australia, Canada, New Zealand, and the United States, traditionally Western conceptions of heritage have been increasingly challenged by indigenous and migrant communities offering alternative interpretations of history, culture, and beliefs.
In newly independent countries, heritage reinterpretations also provide opportunities to forge new coalitions and a sense of belonging. For example, before apartheid ended in April 1994, heritage in South Africa had followed an elite Western interpretation consisting mainly of colonial architecture, battlefields, and archaeological sites. In the postapartheid era, new preservation laws were passed that added heritage sites and practices that were deemed important to the indigenous people. Such “new heritage” includes places of spiritual significance, sites associated with the history of slavery and the independence movement, graves, cemeteries, and apartheid-era prisons, as well as artifacts associated with tribal oral traditions (Green 2011). Such inclusions play a necessary role in forging a new sense of identity for the evolving nation.

Beyond the nation, heritage can also play a role in fostering transnational identities. In the run-up to the formation of the European Union in 1993, for example, much of the discussion on creating a pan-European identity revolved around the need for a shared heritage. Quintessential European achievements and heritage such as the Renaissance and the Enlightenment, and certain innovations in the arts, architecture, science, and technology, were identified, endorsed, and celebrated as pinnacles of European history and experience (Graham et al. 2000). For a “new Europe” to emerge, history had to be reinterpreted and a “new past” or “new heritage” devised (Ashworth and Larkham 1994).

There are many challenges in co-opting heritage for nation-building purposes. The term “heritage dissonance” describes discord whenever people have conflicting views about heritage and what best represents a nation and its people. Apart from the obvious differences across citizens and migrants, the views of citizens may also differ along the lines of age, gender, class, ethnicity, and political affiliation. State-sanctioned heritage adds another dimension on top of already complex pre-existing layers of identity, further complicating questions of what heritage is and who decides. As much as heritage has a cohesive potential, it also has a separatist or divisive tendency that must be addressed if nation-building projects are not to go awry.

Heritage management and sustainability

The importance of cultural and natural heritage necessitates long-term care and administration. Heritage management involves resource restoration and protection, as well as management of resource conflicts. Such conflicts arise when different groups or countries contest each other by asserting exclusive claims on or use of particular resources.

A central tenet in heritage planning and management is cultural resource management (CRM). CRM entails a multifaceted approach, beginning with a holistic view of resource selection, a thorough consideration of how heritage is affected when it is being developed, vigilant control of looting, and systematic archiving of resource documentation and conservation practices. In addition to taking care of heritage sites physically, it is also important to manage sites socially. This involves active management of people’s behavior, particularly in locations that are popular with tourists, and public education programs targeting locals, visitors, and resource owners.

Heritage management is also deeply concerned with the issue of integrity. Heritage integrity is defined as the intactness of a particular resource and its attributes. All the elements of a heritage must be present in order for its outstanding universal value to be expressed (Gullino and Larcher 2013). To ensure integrity, management plans must be long-term and holistically conceived.
CULTURAL HERITAGE

Local participation is crucial and entails involvement by stakeholders such as local site managers, communities, local and regional governments, heritage interest groups, and nongovernmental organizations (NGOs) (Landorf 2009). A cycle of continuous action that involves planning, implementation, monitoring, evaluation, and feedback must also be put in place. While all this is feasible in theory, the reality is that the execution of management plans is fraught with difficulties, particularly in the developing world. The lack of political will and administrative resources, as well as heavy financial costs, frustrate the implementation of plans.

For UNESCO World Heritage Sites, management is easier to enforce and monitor. Formal nomination entails commitment on the part of owners and country governments to protect heritage through a detailed plan, which includes long-term legislative, regulatory, and institutional provisions. Heritage sites also have access to grants and technical assistance from the World Heritage Fund. World Heritage Site listing may also be withdrawn if management guidelines are not adhered to. Such sites are first placed on a “World Heritage in Danger” list by the World Heritage Committee and negotiations are undertaken with local authorities on remedial strategies. Two natural heritage sites have so far been delisted – one in Oman and another in Germany – because of developments near the sites that were regarded as detrimental to their scenic qualities and their flora and fauna. Such delistings are unfavorable for countries not only because of the loss of funding opportunities but also because of the embarrassment to local authorities. One of the guidelines for World Heritage listing is advanced recognition of threats in the form of development, pollution, tourism, and agriculture. Management strategies must also be identified to alleviate these threats.

It should be clear that heritage management is about not only the physical care and protection of sites, but also the management of people's behavior and the provision of information to the public. While the former contributes to physical sustainability, the latter concerns social sustainability. Each heritage site is conceived as having both a physical and a social carrying capacity. If the physical carrying capacity is breached, heritage sites will experience wear and tear, resulting in destruction, denudation, and the loss of tangible artifacts. In the case of social carrying capacity, poor management results in the loss of local traditions, skills, and customs, particularly in sites that are overwhelmed with tourists. Resentment of tourists requires monitoring and, where needed, mitigation through schemes that benefit local residents. These include free access to heritage sites or discounted membership for cultural programs. Both dimensions of physical and social control must be acknowledged if management is to be holistic and sustainable.

The future of heritage

Although heritage consists of relics from the past, it is crucial to think about its future. Heritage evolves over time in both a quantitative and a qualitative sense. In a quantitative sense, the sum total of what is regarded as heritage changes as “new” heritage is added to the sum or when existing heritage is lost because of environmental problems. In a qualitative sense, what constitutes heritage will also change as societies, politics, and technologies evolve.

The most immediate cause of heritage change is the threat of physical destruction brought about by climate fluctuations. Over time, an increase in global temperatures resulting from the emission of greenhouse gases has led to the demise of natural and cultural heritage. Changes
in water temperature have caused coral bleaching and loss of marine biodiversity. The Great Barrier Reef, for example, experienced severe bleaching in the late 1990s and early 2000s, which led to about 5% of its reefs being stripped of 50–90% of corals (UNESCO 2007). Global temperature rise also affects snow and ice cover, depleting upland glaciers and leading to severe consequences for lowland rivers. Cultural heritage is not spared either. Increased urban floods affect historic buildings, many of which were built with materials not accustomed to excessive ground moisture. Changing soil conditions also affect archaeological sites, while rising sea levels impact historic coastal areas. How much heritage survives into the future depends on the severity of environmental problems and the effectiveness of mitigation measures.

Qualitative changes also take place and such changes refer to how heritage is defined. The World Heritage List is a good example. It is only since 1992 that new categories have been added to the list which, at that time, was made up of mainly religious, cultural, and archaeological heritage. Since then, recognition has been given to commercial industrial heritage (such as railways and factories), buildings of local architectural vernacular, rural cultural sites, and pilgrimage and migration routes. Such reappraisals take place because of changing values and what people consider as historically or culturally important. Likewise, today as the composition of society changes, new constructs of identity, authenticity, and belonging emerge. In many cities and countries, the contributions of migrant ethnic communities are acknowledged and their cultures recognized as part of a collective heritage. South Asians in London and Chinese in Toronto and Vancouver are examples of migrant communities whose heritage have been accepted as part of the historical and cultural fabric of their host societies.

Another change lies in the area of tourism. The mass tourism of the past has given way to segmented groups of travelers characterized by specific interests. What was earlier identified as heritage tourism may be more accurately recognized today as comprising niche groups such as culinary, education, or architecture tourism. Tourists are no longer satisfied with the passive observation of heritage but demand active participation and immersion too. For example, culinary tourists no longer desire only good food but also wish to visit local markets, interact with chefs, and engage in cooking and food preparation. A deeper appreciation of heritage is possible through active engagement. Tourism planners and service providers must ensure that, as new forms of heritage are made available, the exchange between guests and local hosts is mutually beneficial.

Finally, new media and communication technologies are presenting innovative ways for heritage to be archived, publicized, marketed, and consumed. In the twenty-first century, many heritage boards, trusts, and museums have turned to digital and web technologies to change the way heritage is communicated to the public. Artworks, museum pieces, and heritage sites are showcased on different web platforms for virtual visitors to enjoy. Technology has also allowed three-dimensional “up close and personal” viewing in a way that reality cannot or will not permit. The archiving of oral histories in digital format and immediacy of access have also brought intangible heritage closer to people. Feedback from the public on any aspect of heritage can also be instantaneously channeled to managers and administrators.

As debate continues over whether virtual tourism can ever substitute for actual visits to heritage sites, there remains considerable interest in heritage and increasingly novel ways of presenting it to the public are found. Sensitive
CULTURAL HERITAGE

management of heritage and sustainable ways of communicating it are essential if heritage is to live up to its definition as something precious to be protected and passed onto future generations.

SEE ALSO: Built environments; Conservation and capitalism; Cultural heritage; Environmental management; Identity; Landscape; Sustainable development; Tourism; Urban redevelopment

References


Further reading


Cultural politics

Orlando Woods
Independent scholar

Cultural politics stems from the fact that we are all – to varying degrees and in various ways – conditioned by culture. Our cultural identities automatically (and often unknowingly) create a taxonomy of understanding that deeply influences how we see ourselves (and people like us) and how we see others (or people not like us). Such taxonomies are often deeply rooted, as societies traditionally have formed, developed, and strengthened along the lines of cultural similitude and spatial segregation. Less than a century ago, it would not be uncommon to find communities that were more or less culturally homogeneous. Since then, shifts in global politics and economics, and advances in technology, have caused the mixing and dilution of cultural norms, and the creation of new cultural forms. Such shifts and advances have facilitated processes of migration and the development of global communications networks. In turn, these processes and developments have shaped the sociocultural landscapes of the modern world in ways that are unprecedented in scale and effect. Most noticeable is the fact that they have resulted in the constant and largely unavoidable blending of people, ideas, symbols, and values, and ever-greater assertions and contestations of power. The relative safety of cultural homogeneity has given way to the uncertainty of cultural heterogeneity, and with it a whole raft of questions and concerns about cultural difference have emerged.

Cultural differences create power asymmetries, which are enshrined in a number of oppositional pairings. Such pairings – power and resistance, dominance and subordination, sameness and difference, inclusion and exclusion, insider and outsider – form the lexicon of cultural politics. Each pair reveals the different standpoints and tensions of different cultural stakeholders. Such standpoints also play a central role in the ordering, classification, and use of space, and the politics contained therein. Indeed, space and place can both reflect and enforce the cultural politics of a given community or area, and have helped shape the research agendas and theoretical contributions of cultural geographers since the 1980s. Such agendas and contributions have engaged with debates in the broader social sciences to show how cultural politics not only come from within (i.e., how culture dictates the framing of self and other), but from without as well (i.e., how culture is studied, understood, and interpreted by scholars and public discourses that seek to advance specific interests, or a specific idea of “culture”). “From within” and “from without” are two approaches to understanding cultural politics.

The politics of cultural knowledge production

The new cultural geography of the 1980s sought to dismantle existing frameworks of cultural understanding, and to replace them with a new focus on cultural practices and processes, and their relations to power. Doing so has, among other things, led to a scholarly focus on understanding the power relations that are
CULTURAL POLITICS

inherent to cultural knowledge production. This has resulted in the interrogation of academic knowledge production itself, with a particular emphasis on exposing and reconciling epistemological norms and biases. Cultural knowledge production has, as a result, evolved from being the identification of ontological “fact,” to a more situated and subjective – and inherently contestable – interpretation of culture instead. Accordingly, it is now widely accepted that there can be multiple constructions and interpretations of the same cultural form; however, not all constructions and interpretations are shared and accepted in an equitable manner. As Tim Cresswell (2004, 124) puts it:

The study of culture is … closely connected with the study of power. A dominant group will seek to establish its own experience of the world, its own taken-for-granted assumptions, as the objective and valid culture of all people. Power is expressed and sustained in the reproduction of culture.

Any form of cultural production reflects a struggle between dominant powers, orders, or understandings, and those that seek to resist such dominance. Efforts have been made to bridge the gap between forms of knowledge production from above (i.e., codified, universalist, hegemonic) and from below (i.e., tacit, locally sensitive, emancipatory). Doing so has led to the acceptance and interrogation of culture as an idea (i.e., something constructed and understood by different people in different ways), rather than something that is inherent in people and place.

Such efforts have been strongly influenced by a variety of epistemological developments associated with postmodernism, all of which have led to a rethinking of the power relations associated with cultural knowledge production. Two developments in particular have helped to unravel and to understand (and, in some respects, to further) cultural politics: postcolonialism and postructuralism. Each contributes significantly to the politics of cultural knowledge production that is imposed by constructors on the constructed – a cultural politics from without.

Postcolonialism

Since the Enlightenment thinkers of the seventeenth and eighteenth centuries, cultural knowledge production has served to develop a sense of European (more recently, Western) superiority and civilization, the aim being to further justify and expand European power in the colonies. Cultural knowledge production is intimately bound with geography. Creating the geography of the world (and the discipline itself) coincided with the global spread of capitalism, imperialism, and modern scientific discourse; the mapping of the world and of nation-states in accordance with pre-existing ideas of superiority and inferiority; and the production of cultural knowledge in the colonies. Indeed, the very premise of cultural knowledge production during the colonial epoch was to understand hitherto unknown people and places in order to extend and enforce control over them. To the extent that the political apparatus of colonialism has since been dismantled, so too has postcolonial thought striven to dismantle the construction of culture as the intellectual products of a Western elite who observed, wrote about, and claimed to represent the cultures of non-Western peoples. And yet, as Audrey Kobayashi (2004, 242) rightly claims, “it would be difficult, indeed hypocritical, to avoid the fact that the discipline of geography is dominated by Northern, Western, white scholars whose lives and careers have been constructed out of the very colonial systems that produced them.” This sort of awareness has led to critical self-reflection among cultural knowledge producers, their aim being to bring about the decolonization not
just of territory and people, but of thought and knowledge as well.

Drawing in particular on the work of cultural theorists such as Edward Said, new cultural geographers have attempted to unravel the politics of knowledge production and representation. Said’s (1979) book *Orientalism* has gained notoriety for revealing how Western cultural representations served to legitimize the physical structures of colonial rule. Such representations often served to demean and dominate non-Western peoples and places through language and knowledge. As Dan Clayton (2004, 450) put it:

> the momentum of western power is now deemed to lie as much in the “cultures of imperialism and colonialism” – in language and knowledge, texts and discourses, images and representations, and the iconography of power – as in the political economy of imperial expansion and colonial incursion.

As a result, culture has been rethought in more relational terms. Culture is always produced in relation to the producer, meaning that critical interrogation of the production of culture itself can reveal unequal (or inequitable) patterns of cultural representation and acceptance. Processes of postcolonialism have, in other words, necessitated a focus on who is writing about culture, and the politics of (mis)representation that are inherent in the production of culture.

This development has had a number of tangible effects on the academy. Not only has the recognition of researcher positionality – and associated assumptions and biases – come to the fore, but the study of the subcultures of marginal groups has been elevated to a level that is almost on a par with that of dominant or “official” cultures. Greater physical intermingling of dominant and marginal cultural groups has, without doubt, contributed to such an elevation. Indeed, the formation of subaltern studies – a disciplinary collective that seeks to uncover and legitimize the cultural productions of ostensibly “hidden” (or subaltern) groups – is a clear reflection of the growing prominence of the marginal in mainstream discourses. Some scholars have even been accused of expressing a preoccupation with marginality, to the extent that they have been accused of “fetishizing” the other. Whichever way it is presented, it is clear that the processes of postcolonization have contributed to the devolution of cultural knowledge production, and a relentless scrutiny of who is producing culture and for what purpose.

**Poststructuralism**

The emergence of a new cultural geography in the 1980s marked a point of departure from structuralism. As the predominant organizing principle of much Western academic thought, structuralism classified the world according to oppositional binaries of us versus them, self versus other, center versus periphery, modernity versus tradition. Since the 1980s, however, poststructuralism has largely supplanted structuralism in academic discourses. An important contribution of poststructuralist thinking has been the effort made to overcome these binaries, which establish figurative lines of difference that set people and places apart. In a globalized world where people are increasingly interconnected and cultural symbols increasingly shared, such binaries have become as fallacious as they are redundant. There is no such thing as an all-encompassing “self,” as it is likely that most of us will share one or more facets of cultural identity with the “other.” To this point, one of the most pervasive legacies of poststructuralist thought has been the work of Michel Foucault (1977) and his rethinking of the cultural “subject.” Such rethinking has witnessed an evolution from understandings that are coherent
CULTURAL POLITICS

and bounded to those that are more fragmented and multifaceted.

The pervasive influence of Foucault’s work has caused the cultural to become a more subjective, fluid, and complex frame of understanding. The intellectual hallmark of poststructuralism – and part of Foucault’s enduring legacy in the social sciences – is social constructionism, and the associated belief that knowledge is not inherent nor predetermined by biology but constructed (or invented) by humans advancing specific interests and viewpoints. Female characteristics such as being more passive, dependent, and emotional than males are now believed to be the result of ascription by one person or group to another, rather than an intrinsic characteristic that is unique to women. This has brought about a scholarly focus on the power relations that are enmeshed within the practices of everyday life, and which contribute to the production of identities and the boundaries that demarcate difference. Geographers in particular have since focused on the “messy and inherently spatialized entanglements of domination/resistance, as always energized and traversed by the machinations and effects of power” (Sharp et al. 2000, 2). Feminists, for example, have rallied against the effects of patriarchy on society, and the creation of spaces that serve both a constraining and an enabling role for women.

The politics of cultural identity

Cultural identities are not homogeneous, nor are they coherent. Poststructuralist thought stipulates that identities are multifaceted, fluid, relationally constituted, and sometimes contradictory; they are not essential. Identity grants acceptance into some cultural groups but not others, just as it enables access to some spaces and not others. Above all else, the cultural politics of identity are concerned with diversity, and with the inclusion or exclusion of cultural minorities in the dominant sociospatial order. That is, power is not held and yielded unilaterally by one group. As much as institutional power is enforced from above, it is also complemented, challenged, and sometimes nullified by power (or resistance) from below. The gaps that arise in the difference between the yield and effect of power from above and below are what drive the politics of cultural identity – a cultural politics from within. The following subsections explore the cultural politics of the group, the cultural politics of the individual, and the cultural politics of landscape. While the first two subsections focus specifically on the operation and effect of society in space, the third focuses on the effect of space on society.

The cultural politics of the group

Social groups tend to draw on one or more facets of cultural sameness in order to bind themselves together. The power of the group lies in its ability to define itself in relation to others, and in so doing to establish and enforce ideas of sameness and difference. Importantly for geographers, it is also able to impose conditions or restrictions on the access to, and behaviors within, the spaces that it claims, owns, and occupies. Groups that draw on deeply rooted (and often immutable) cultural traditions are typically more conservative in nature, and share an often idealized understanding of cultural values and legacies. They can be oppressive in their willingness to protect and maintain the prevailing sociocultural order. As Iris Young (1990, 303) reminds us, they are typically intolerant of difference and are defined by the pursuit of “the same social wholeness and identification that underlies racism and ethnic chauvinism on the one hand and political sectarianism on the other.” Dominant groups often strive to keep minorities apart – and, by doing
so, enforce social and spatial division rather than diversity – while minority groups often seek to resist and subvert dominant groups in the bid for institutional acceptance and to express their own cultural identities. Invariably, therefore, majority and minority cultural groups often exist in a state of tension, with each trying to control, suppress, or undermine the other.

Nationalism is the most territorially specific form of cultural identification, and also one of the most potent drivers of identity politics around the world. Majority cultural groups that share common ethnic, linguistic, and/or religious traditions may form a national union that claims the sovereign right to a given territory. Given its implicit numerical dominance, such groups are often closely aligned with the state. This alignment (in the form of the nation-state) serves to strengthen the cause of nationalism while encouraging the intolerance of cultural minorities. Such intolerance typically involves the marginalization – even persecution – of minority groups, or culturally distinct migrant communities with whom a common territory or citizenship may be shared. Minority groups may resist by campaigning for greater autonomy or even liberation from the restrictive or controlling influence of the dominant group. In its most extreme form, the cultural politics of the group can lead to secessionist war or even genocide, as witnessed in the Holocaust against the Jews from 1941 until 1945, the mass slaughter of Tutsis in Rwanda in 1994, and the civil war between Sinhalese and Tamils in Sri Lanka that played out over two and a half decades from 1983 until 2009.

The cultural politics of the individual

Every individual is a cultural being, and cultural identity is an embodied phenomenon. Every body has a set of physiological characteristics that reveal the building blocks of a person’s cultural identity. Physiology makes visible a person’s sex and race, but it is the different interpretations of – and expectations associated with – these identifiers that often drive the cultural politics of the individual. As such, there are socially constructed – and socially accepted – ways of being male or female. While the physiology of the body cannot (without invasive surgery) be changed, the way the body is dressed, decorated, and adorned by the cut and color of hair, the style of clothes, the presence of tattoos, piercings or makeup, the accessorizing with jewelry – not to mention the style of speech, mannerisms, attitudes, and behaviors – gives the individual unlimited freedom to interpret and represent their cultural identity as they wish. In other words, while it is nearly impossible to escape the biological entrapments of cultural belonging, everybody has the power to use their body in a way that can enforce or challenge predetermined cultural stereotypes.

Cultural geographers have explored how space can dictate and control the movement of bodies and, specifically, how cultural norms and expectations can restrict mobility. For example, gendered dress codes in office spaces – narrow skirts and high-heeled shoes – have resulted in the inability of women to walk quickly and assertively. In some instances, the evolution and expansion of cultural identifiers has resulted in the clearer demarcation of social space. Class, for example, was once a descriptive term that referred to income, status, occupation, or consumption habits, but has since developed into a more meaningful term that can be simultaneously exploitative, judgmental, and spatially restrictive. Gated communities have developed as spaces that do not welcome – or permit the entry of – working-class bodies without due reason, while working-class council estates are spaces wherein the bodies of the affluent may
CULTURAL POLITICS

feel threatened and vulnerable to attack. While the cultural identifiers of the body can grant and restrict access to various spaces, they can also dictate behaviors in space as well. For example, the movement and freedom of Muslim women in public spaces can be controlled through dress (having to wear the burka or hijab) and prescriptions on behavior (in Saudi Arabia, for example, Muslim women are banned from driving cars; in many Western contexts, however, restrictions on the freedom to wear Muslim dress are often depicted as violations of human rights). Transgressing such conditions of movement can lead to a range of outcomes, from emancipation to social outcasting and persecution.

The cultural politics of landscape

Cultural landscapes are at once a reflection and a manifestation of the dominant cultural order. Such landscapes are most commonly associated with public spaces, and can be divisive insofar as they help to instill and enforce feelings of cultural inclusion and alterity. While public spaces typically refer to the built environment, they also include other mainstream spaces associated with the media, education, politics, and housing. As David Chidester and Edward Linenthal (1995, 17) state with regard to religious spaces—a maxim that has wider-ranging application beyond religion—the cultural landscape can reflect and establish “hierarchical power relations of domination and subordination, inclusion and exclusion, appropriation and dispossession.” The presence of cultural minorities and the associated alteration of the cultural landscape can be a cause of much resentment and resistance, and can lead to greater expressions of cultural diversity on the one hand, and be a source of bitter intercultural conflict (and the more stringent regulation and surveillance of public spaces) on the other. Accordingly, the spaces of cultural minorities—also referred to as subcultural spaces—often function as spaces of resistance to a dominant cultural order. Clearly, cultural landscapes provide some of the most tangible and visible evidence of cultural politics, and are one of the most long-standing areas of inquiry among cultural geographers.

In many Western contexts, public spaces have traditionally been associated with masculinity, heterosexuality, and able-bodiedness, and imprinted by the dominant racial, religious, and linguistic group of a territory. Neighborhoods and housing are typically segregated along the lines of class or ethnicity, whereas public spaces have traditionally—whether intentionally or not—enforced various stereotypes associated with gender and sexuality. While feminists have shown a preoccupation with overthrowing—or at least moderating—the patriarchy associated with mainstream culture, scholars of queer studies have sought to do the same with the assumptions of heterosexuality that permeate everyday cultural symbols found in advertising, film, and TV storylines, and in residential and educational spaces. For example, the Stonewall Inn (a gay bar in New York City) has been credited with marking the beginning of the modern struggle for gay and lesbian rights after its patrons defiantly resisted a raid by a homophobic police force in 1969. For other groups, private spaces are where resistance to the prevailing cultural order can be realized and fomented. House churches in communist China and Buddhist-majority Sri Lanka have been shown to enable the proliferation of Christianity in cultural contexts where public displays of religious presence and growth are discouraged via a variety of formal and informal means. Indeed, it is almost to be expected that, wherever a cultural landscape reflects the ideals of a dominant group at the expense of
minorities, it will encounter some form of resistance.

The politics of cultural representation

Finally, as part of the emancipatory agenda of the new cultural geography, debates around the different forms of cultural representation have come to the fore. In particular, a rift between the scholarly foci on texts, theories, and things (or representations of culture), and experiences, practices, and feelings (or fleeting and often more performative representations of culture), has caused a rethinking of how culture is studied and understood. Cultural representation has, in other words, expanded and become more nuanced as researchers grapple with different – and often unthinking – ways of studying, understanding, and replicating cultural symbols and property. Nonrepresentational theory attempts to provide recompense to the belief that normative representations of culture have become increasingly decoupled from experience, leaving behind a representational void that is filled by nonrepresentational forms of cultural transmission. As such, nonrepresentational cultural geography focuses primarily on the performative aspects of culture, such as music, dance, art, and other forms of sensory-affective experience. In many respects, the politics of representation come from the same root as the politics of cultural knowledge production – that is, how the cultural world is understood, interpreted, and (re)produced.

Pioneered by Nigel Thrift in the mid-1990s, nonrepresentational theory aims to celebrate the everyday workings of culture – through creativity, imagination, and play – and how such workings are used to undermine systems of power (see Thrift 2007). The argument is that academics do culture a disservice by losing the contextual nuances of culture by writing about it in academic books and journals. In a bid to alleviate this problem, nonrepresentational theory seeks to present culture not as a post hoc event, but in the moment of creation. Doing so has seen a methodological shift toward more in situ methods such as filming, and asking research participants to document their lives through photography, film, and journaling in a bid to create real-time data. Indeed, alongside the growing sensitivity toward cultural politics, there has also been growing interest among geographers in not just studying cultural politics, but also addressing and attempting to alleviate its effects as well. Participation has ranged from charitable and volunteer work to the embracement of participatory methods and, in its most persuasive form, political activism and lobbying.

Nonrepresentation is not, however, without criticism. Just as nonrepresentational forms of cultural study attempt to break away from representation, they cannot, ultimately, disconnect from representation entirely. Representational forms of communication are needed to translate the nonrepresentational into something that can be shared and understood by a wider academic community. While it has been criticized for promoting a sort of intellectual circuitousness – one obsessed with linguistic games and the decoding of meanings – nonrepresentational theory has nonetheless helped to provide another approach to the study and understanding of cultural politics. In doing so, it serves to further highlight that cultural politics is an inherent part of cultural study, and pervades every aspect of cultural production, identity, and representation.

SEE ALSO: Difference; Identity; Landscape; New cultural geography; Nonrepresentational theory; Orientalism/Occidentalism; Postcolonial geographies; Postmodernity; Poststructuralism/poststructural geographies
CULTURAL POLITICS

References


Further reading


Cultural studies

Julie Cupples
University of Edinburgh, UK

Kevin Glynn
Massey University, New Zealand

Cultural studies is a critical interdisciplinary (and sometimes antidisciplinary) field that studies cultural dynamics, forms, struggles, and transformations. Defining cultural studies is difficult as the field has always resisted disciplinary codification and closure. It draws on various strands of Marxism, feminism, semiotics, and poststructuralism to examine culture as it is lived, experienced, performed, contested, and transformed. Culture is understood as a site of struggle through which meanings, pleasures, discourses, representations, identities, and practices are fashioned and reworked. Its proponents are interested in how relations of power and oppressions of class, race, gender, and sexuality are reproduced and can be subverted, and they see everyday cultural practices as important sites for such reproduction and subversion. Cultural studies unsettled established academic disciplines because it dealt with cultural forms and practices that were not traditionally studied or taken seriously within universities, including ordinary and popular cultures of everyday life in contemporary societies such as working class tastes, youth subcultures, and women’s cultures. It focuses on the structural constraints within which these cultures come into being, but also on the forms of agency, pleasure, and resistance that enable ordinary and subordinated people to generate meanings, creative practices, and forms of struggle. Cultural studies famously refused to view contemporary cultural consumers as “cultural dopes” (Hall 2005, 67).

While there are some standalone departments and programs of cultural studies, scholars who identify with the label are widely dispersed throughout the humanities and social sciences. Cultural studies has inspired, and been inspired by, many other disciplines, including English, sociology, anthropology, geography, and education. It also interacts heavily with and informs many interdisciplinary fields, such as American studies, media studies, postcolonial studies, science and technology studies, and gender studies. Cultural studies as a field of study is now thoroughly institutionalized and is visible in universities, conferences, academic societies, and publications throughout the world, and its intellectual contributions have been fundamental and decisive. Its defiance and critique of traditional disciplinarity, and its engagement with subject matter historically disregarded by the academic establishment, have reshaped how many fields and established disciplines operate. Nevertheless, in the age of the increasingly corporatized university, cultural studies often occupies a contested and sometimes precarious position. While cultural studies work has been and continues to be highly influential throughout the world, the field’s radical interdisciplinary and, arguably, increasingly amorphous identity have rendered it vulnerable in the context of neoliberal universities’ cost-cutting, restructuring, and deepening audit cultures. Indeed,
cultural studies was hastily eliminated at the University of Birmingham in 2002 after a research assessment exercise result that fell one point below expectations.

The question of what constitutes the boundaries of cultural studies is a complex and contested one. While a spirit of theoretical experimentation and radical interdisciplinarity has long been central to cultural studies’ ethos and its practitioners have struggled against the calcification of the field into an orthodoxy, it has also been apparent from the start that “cultural studies” must be more than a synonym for any and all studies of culture. Those who edit the field’s journals, organize its conferences, and run its degree programs necessarily encounter a recurring need to articulate the boundaries of cultural studies, however provisional, expansive, and permeable they may be. Hence the present, future, and even the past of cultural studies have been matters of ongoing debate since the field’s earliest days.

Emergence of cultural studies

While there have been politicized analyses of contemporary culture throughout much of the twentieth century and the world, cultural studies as a specific field of intellectual enquiry came into being in the United Kingdom via an initial creative blending of critical pedagogy, adult education, and Marxist analysis, joined later by semiotics, structuralism, feminism, ethnography, and poststructuralism. Richard Hoggart’s *The Uses of Literacy* (1957), Raymond Williams’s *Culture and Society* (1958), and E.P. Thompson’s *The Making of the English Working Class* (1964) are often described as key texts for the emergence of British cultural studies. While quite different from each other, these works are united in taking working class culture seriously and emphasizing, albeit to varying degrees, the agency and capacities of ordinary people to make their own culture.

The naming and institutionalization of cultural studies began with the creation of the Centre for Contemporary Cultural Studies (CCCS) at the University of Birmingham in 1964. Richard Hoggart, who had worked as an adult educator before joining the university in 1962, was the center’s first director. He was succeeded in 1968 by Stuart Hall, a Jamaican-born leftist. Hall’s thinking was more theoretically inclined than Hoggart’s, and under his leadership the center engaged closely with Marxism, structuralism, and semiotics. While Althusserian and Barthesian approaches enabled those at CCCS to develop sophisticated analyses of contemporary media culture, by the end of the 1970s the center had turned decisively to Gramsci, whose theory of hegemony offered a way to rethink the functionalism increasingly associated with Althusser’s account of ideological state apparatuses (ISAs) and subjects, and thus more effectively to identify sites of hegemonic contestation and possibilities for political transformation across contemporary cultural landscapes.

Those who worked at the center described it as a theoretically productive and experimental space, full of tensions and vibrant debates. During the 1970s, CCCS scholars produced pathbreaking analyses of working-class lads, popular music subcultures, and television news and its audiences. In the mid-1970s, in the midst of second-wave feminism, a number of female graduate students, including Angela McRobbie, Chris Weedon, Charlotte Brunsdon, and Janice Winship, took the faculty to task for their lack of attention to gendered power and to the cultures of women and girls. In 1978 they published *Women Take Issue: Aspects of Women’s Subordination*, which combined theoretical analysis with case studies such as women’s magazines, teenage girls’ cultures, and the home lives of...
working-class women. This initiative paved the way for the development of feminist cultural studies that explored things like soap operas, romance novels, and the cultures of women’s and girls’ everyday lives. Hall (1992, 282) wrote that like a “thief in the night,” feminism “broke in, interrupted, made an unseemly noise, seized the time, crapped on the table of cultural studies.” Similarly, there were important interruptions and redirections of the Birmingham project’s original focus on class as an object of analysis with the attention given to race in one of the center’s most celebrated publications, *Policing the Crisis: Mugging, the State, and Law and Order* (1978), and in *The Empire Strikes Back: Race and Racism in 70s Britain* (1982).

Stuart Hall: meaning, discourse, and ideology

The turn to Gramsci at the center in the 1970s facilitated analyses of the ways in which subordinated social groups negotiate the social circulation of meanings and ideologies in contemporary culture. Among cultural studies’ primary targets were various reductionisms (economic as well as textual) that too readily presume the effects of the “culture industries” in securing the domination of the power-bloc. As Hall (2005, 71) writes, popular culture is best understood in much more contingent, dynamic, and processual terms as a site where the struggle “for and against a culture of the powerful is engaged.” Hence, although “it is partly where hegemony arises, and where it is secured,” it is at one and the same time, for Hall (and for cultural studies), a thoroughly contradictory terrain where competing perspectives, discourses, and interests – top–down as well as bottom–up – clash and compete, and where “the people” might be constituted as “a popular-democratic cultural force.”

Hall encouraged us to think of cultural productivity and agency in terms of articulation, understood as both a contingent link that can be broken and a process for putting into meaning or “language-ing” (see Grossberg 1996). Meaning is thus understood as a product of contingent linkages between different discourses, sub-discourses, and other signifying elements that may, though need not necessarily, be conjoined within larger ideological formations. The meanings that result from such breakable connections can be shifted through practices of disarticulation and rearticulation; the conjunction of previously disjoined discourses necessarily brings new meanings into social circulation. Hall’s self-styled “post-Marxist” approach to this issue thereby evades long-standing problems of reductionism and essentialism in Marxist thought concerning the relationships between economic practices and subjectivity/consciousness (or other cultural products). Hall replaces the notion of pregiven social formations (such as classes or races) and of fixed ideological ones (e.g., “dominant ideology,” revolutionary consciousness) with a more flexible and contingent set of concepts that nevertheless remain anchored in a materialist epistemology and politics.

Within this framework, when the subelements of an existing discourse are negotiated, broken up, rearranged, and articulated to subelements of other discourses, new subject positions emerge that generate fresh ways of making sense and of engendering social crises and mobilizations. Such newly rearticulated discourses and subject positions might then become part of emergent social identities and movements, which are themselves understood as the products of these ongoing negotiations and articulatory practices (rather than being understood as preexisting “classes” waiting to be activated into consciousness). Hall’s
presumption that there are no pregiven, fixed identities prior to the negotiation and cultural articulation of meanings is not a form of idealism (e.g., “everything is discourse,” “anything can be articulated with anything else”), but rather a rejection of essentialism and excessive determinism.

Hall gives the example of Jamaican Rastafarians, who came to political consciousness and to social and cultural activism by rearticulating the subelements of preexisting discourses from the text of the Bible into a thoroughly modern and wholly new set of expressive practices. As Hall writes of the Rastafarians, because the Bible was a text that “did not belong to them,” they had to rearrange and invert it to produce meanings “that fit their experience. But in turning the text upside-down they remade themselves; they positioned themselves differently as new political subjects; they reconstructed themselves as blacks in the new world: they became what they are. And, positioning themselves in that way, they learned to speak a new language” (Grossberg 1996, 143). Condensed into this highly dialectical example are key ingredients of a nonreductive, nondeterministic mode of cultural studies thinking: cultural innovation comes not from whole cloth but rather from a materially anchored and socially motivated reworking of existing sense-making resources for new purposes. In this reworking, the cultural agency of subordinated peoples is neither unconstrained nor fully subordinated to the control of centralized administrative bureaucracies or conglomerated capitalist enterprises; rather, it exists within a field of continuous and ongoing contestation over the disposition and mobilization of signifying resources: a site of semiotic struggle or a terrain characterized by conflict in terms of a politics of meaning that is fully enmeshed within and articulated to lived conditions of material existence wherein significant stakes hang in the balance.

**Internationalization of cultural studies**

Throughout the 1980s, cultural studies increasingly went global and started to become a much more diverse and variegated set of enterprises. Key practitioners moved from Britain to Australia at the onset of Thatcherism and helped generate a very significant cultural studies scene down under. Meanwhile, cultural studies took hold in the United States as an alternative to both economistic and more pessimistic modes of Marxist analysis. During the same decade, cultural studies began its extended engagements with poststructuralism, post-Marxism, and postmodernism. The first two of these, particularly in the guise of Michel Foucault and Michel de Certeau, would lead to even further complication of cultural studies’ theories of power and subjectivity. To a significant degree, the third formation, postmodernism, began giving way to the conceptual problematic of globalization as a source of theoretical energy and terrain of struggle in cultural studies from the 1990s onward.

Throughout the period of cultural studies’ internationalization, the field has multiplied its issues and perspectives, engaging with an ever-expanding range of theorists and objects of analysis. In recent decades, scholarship in cultural studies has dealt with contemporary media; popular audiences and fandom; critical pedagogies; genders and sexualities; race, ethnicity, and nations/nationalities; subjectivities, identifications, and identity politics; urban life and cultures; postcolonialism and decoloniality; cultural policy; globalization and transnational flows; science and technology studies; surveillance; human–animal relations; cultures of environmentalism; satellites and remote sensing; bodies and bioethics; new social movements; human rights and citizenships; the War on Terror and securitization; neoliberalization; and consumption.
While the Birmingham School is central to the development of cultural studies as an intellectual field, it is important to recognize the diverse sites in which cultural studies emerged and continues to proliferate. The cultural studies project has been taken up around the world in a variety of ways (see Ríos, Trigo, and Del Sarto 2004; Abbas and Erni 2005). Today, cultural studies has many regional formations that extend far beyond the United Kingdom into Asia, Oceania, the Americas, Africa, and continental Europe. While some cultural studies formations recognizably build on the intellectual foundations developed by Stuart Hall and others in the United Kingdom, it is important to observe that they also draw on other regional cultural histories, theoretical underpinnings, and intellectual trajectories. Latin American cultural studies, for example, is influenced less by Raymond Williams and Richard Hoggart than by José Martí, José Carlos Mariátegui, Angel Rama, and Darcy Ribeiro, and this work in turn influences Latin American geographers. A recognition of and engagement with such diverse genealogies has led to the indigenization, decolonization, and de-Westernization of cultural studies and to productive comparative debates across geographical difference. For instance, there is a body of literature emerging from Latin American cultural studies, termed the modernity/coloniality/decoloniality (MCD) paradigm, which draws on earlier decolonizing historical events and scholarship, and on indigenous and popular knowledges from América. As John Frow (2005) has noted, cultural studies’ intellectual, theoretical, and institutional history is highly contested and heterogeneous; there is no singular trajectory that can be neatly mapped, but there are important theoretical detours, key theorists, and debates that vary depending partly on where one practices cultural studies. While there are many ongoing arguments about the strengths and weaknesses of cultural studies, the field’s sense of radical experimentation persists and has tended to prevent its congealment into a fixed disciplinary identity.

One strain of cultural studies work that has been both influential and controversial in recent years has developed around the expansion of participatory digital media and the emergence of what Jenkins (2008) calls “convergence culture.” In the new media environment, where audiences, citizens, and fans can produce, circulate, remix, refashion, vote on, and comment about all manner of cultural events, products, and figures, spaces are opened for the participatory negotiation and rearticulation of significant political issues (Cupples and Glynn 2013). In the digital media ecology, people develop and hone skills that lend themselves to the production of new modes of participation, creativity, and collective intelligence, new forms of political and transcultural engagement, decreased dependence on official expertise, and enhanced capacities for collaborative problem-solving (Jenkins 2008).

Cultural studies and geography

In many ways and for various reasons, cultural studies set out deliberately to disrupt established modes of disciplinary thinking and practice. The intellectual debates engendered by the work of the Birmingham School had a profound effect on disciplines such as sociology, mass communication, and geography, all of which underwent a “cultural turn.” Geography, for example, was reworked in the 1980s and 1990s as practitioners drew on intellectual sources and developments taking place elsewhere, including in cultural studies. This reworking has given us a significantly expanded terrain of geographic subject matter in the form of the new cultural geography.
A seminal text that explicitly brought cultural studies into dialog with human geography was Peter Jackson’s *Maps of Meaning* (1989). This book drew on the work of the Birmingham School (and other theoretical influences) to give questions of race and racialization a distinctly spatial inflection and thus to demonstrate that both racisms and struggles against them have their own specific geographies (see also Jackson 1987). In the process, Jackson developed a critical and materialist cultural geography that moved beyond the rigidity and structuralism of classical Marxism. Without denying that ordinary people in Thatcher’s Britain operated under structural constraints, Jackson emphasized the spaces for agency and resistance they created through their uses of cultural resources available to them. Since the publication of *Maps of Meaning*, scholars such as Tim Cresswell, David Morley, Kevin Robins, Nick Couldry, Lynn Spigel, Clive Barnett, Jason Dittmer, and others have produced work explicitly positioned at the intersection of human geography and cultural studies.

Meanwhile, cultural studies has undergone something of a spatial turn, which has proven to be highly compatible with the field’s commitments to conjunctural and contextual analysis. For example, significant work has explored the micropolitics of living rooms as places where TVs and digital media are both embedded within and constitutive of (particularly gendered and generational) relations of meaning and power. An important recent collaboration at the intersection of human geography and cultural studies, the Kilburn Manifesto (http://www.lwbooks.co.uk/journals/soundings/manifesto.html), seeks to intervene in widespread processes of neoliberalization. This initiative, led by Stuart Hall, Doreen Massey, and Michael Rustin, is a series of publications that attempt to stimulate the generation of alternatives to neoliberalism by radically rethinking the current economic order in terms that are missing from many contemporary political debates. Geography’s cultural turn and cultural studies’ spatial turn have made it difficult to distinguish categorically between contemporary cultural geography and cultural studies.

**SEE ALSO:** Critical geography; Cultural politics; Cultural turn; New cultural geography; Popular culture

**References**


**Further reading**

Cultural turn

Clayton Rosati
Bowling Green State University, USA

The cultural turn describes a broad set of shifts in the social sciences and humanities, which blossomed from the 1970s to the early 2000s. This shift moved geography’s object of study from economic logics, the direct exercise of power (in geopolitics), the quantitative and positivist geography of the early 1970s, and the morphology of landscape, to questions of shared meaning, representation, and the politics of language and consent. Especially in critical geography, modes of explanation shifted from political economy to culture. Culture became an object of study not just in its traditional home of cultural geography but across the gamut of human geography, from economic geography to critical geopolitics to urban geography.

Coinciding with and building on emergent critiques of positivism, universalism, determinism, essentialism, patriarchy, heteronormativity, racism, and Eurocentrism, this shift incorporated many theories and methods, including feminism, phenomenology, poststructuralism, postcolonial studies, black studies, ethnic studies, queer studies, linguistics, symbolic interactionism, and qualitative methodologies. One might see the roots of the cultural turn developing with Italian communist leader Antonio Gramsci’s interwar critiques of the “vulgar” Marxism – particularly forms of economism, ignoring especially political struggle, intentionality, and intraclass difference – that dominated the Second International after Marx’s death. Just as importantly, the 1950s’ seemingly permanent postwar prosperity in the “first world” demanded further reconsideration of traditional Marxist critiques of impoverishment and crisis, provoking instead critical theories of social management and mental freedom. The English translation and subsequent popularity of Gramsci in the 1970s found utility in an analogous tension between determinism (“structuralism”) and theories of difference (“culturalism”) in the 1960s and 1970s, identified most commonly in E.P. Thompson’s critique of Louis Althusser (Thompson 1978) (although Thompson’s postscript to The Poverty of Theory rejects the idea that he is a culturalist, e.g., pp. 297–298). The interventions of poststructuralists like Derrida and Barthes, and critics of modernity – and Marxist complicity with capitalist modernity – like Foucault, Baudrillard, and Lyotard, struggled to push critical and radical thought beyond the limited horizons of Hegelian grand narratives of history and structuralist fixity in meaning and language.

All of these influences converged in the work by Hall, Gilroy, and others from the Centre for Contemporary Cultural Studies at the University of Birmingham, beginning in the early 1960s, which became crucial for mainstreaming critical scholarship on culture and a central resource for critical geographers. Ultimately, after the failures of the 1960s revolutions worldwide, the cultural turn also came to focus on micropolitics and a rejection of parties, grand strategies, and revolution, in general. With the rise of Thatcher, Reagan, and neoliberalism’s popularity among the Anglo-American working classes in the 1980s, the “natural majorities” and assumed alliances among social movements had to be rejected and, instead, fought for (e.g., Hall
1988). “Culture” became a way of describing the terrain of that struggle, drawing heavily on Gramsci’s theories of building class alliances and winning consent for dominant social norms in “civil society” before they take hold in the state. Here, Barthes and Derrida’s influential work on language was equally important. On the one hand, culture was that which is determined by shared “maps of meanings,” discourse, ideology, and forms of “common sense.” And, on the other hand, it became that which is constantly differentiated, local, historically and geographically specific, and made contingent through use. The cultural turn, thus, embodied a paradox and a continuation of old tensions.

In geography, the cultural turn’s impact was felt widely and drew heavily on these shifts, impacting – among others – contemporary cultural geography, including landscape studies; economic geographies of labor, regulation, innovation, agglomeration, culture industries, urban economic geography, and studies of the capitalist state; and geopolitics involving Gramscian and Foucauldian international relations, scale and activism, development, and neoliberalism; however, culture and the cultural have invited critique. Within geography, the latter part of the 1990s wrestled with the concept after its critique by Mitchell (1995) for reifying culture into a thing, sphere, or realm, making culture a causal force, and ignoring the power of culture industries in producing ideas of difference, if not difference itself. Beyond this, the sensibilities of the cultural turn’s Gramscian emphasis on meaning and representation, and its tendency toward valorizing individual autonomy, are increasingly under critique by non-Gramscian currents in affect theory, actor-network theory, and other variants of new materialism.

Contemporary cultural geography, including landscape studies – where geography’s cultural turn had its principal and most resounding impact – begins in the early twentieth century with the rejection of environmental determinism and social Darwinism, which led to the adoption of an anthropological model whereby the diffusion of particular cultures determined the morphology of vernacular landscape. This, now, “old cultural geography,” in the 1970s came to focus increasingly on the ordinary and everyday aspects of landscape as society’s “unwitting autobiography” (Lewis 1979). With the growing dominance of concepts from cultural studies, however, a “new cultural geography” developed, which rejected the structuring determination of life implied in this old school of thought. Duncan’s (1980) thorough critique of the “superorganic” – Hegelian–like cultural entities above human beings, which determine their actions and works – in cultural geography emphasized the internal differences, choices, conflicts, and struggles that were erased in this old model. The so-called new cultural geographers of the late 1980s and the 1990s engaged emergent theories of difference and linguistic understandings of use and practice (see an influential collection of papers from the 1991 Royal Geographical Society–Institute of British Geographers conference edited by Chris Philo). Landscape, as with the old cultural geography, was still a text to be read. But, with the new incorporation of semiology’s emphasis on the relationship between sign, signifier, and signified, reading landscape also involved reading social attitudes and the maps of meaning produced in interactions with not just landscape morphologies, but also previous meanings associated with them. Drawing also from cultural studies, the status of particular meanings and types of morphological forms themselves varies, understood as dominant, residual, or emergent. Landscape is, thus, seen as part of the constituent process of hegemony – life lived under dominance, not as force but as given conditions and
assumptions beyond which life is difficult to imagine. Work on landscape and urban design also eventually extended this through more Nietzschean critiques of truth into Baudrillardian notions of the hyperreal and simulacra, of signs referring only to other signs.

In economic geography, the cultural turn had several impacts, mostly in branching away from strict and deterministic understandings of capitalist accumulation and of industrial firm location. The impact of French regulationism on economic geography is significant in its engagement with transitions away from economism and versions of ahistorical capitalist logic. Here, as in cultural geography, the inclusion of factors like contingent and historically specific systems of social meaning and social agency in labor geography, capitalist accumulation, and studies of the capitalist state tempered more determined understandings of economic systems and structures. So-called cultures of production, and shared differences between firms’ production practices, were studied as means of developing competitive advantages. “Innovation” became another associated way of understanding the “cultural” or forms of difference in production practices. If culture was conceptually inserted into terms of the practice of production, cultural artifacts by corporations – small and large – also became an object of study for economic geography. With some key exceptions, these studies generally were not critical or did not engage with the content of those artifacts. Rather, as critics have noticed, they were interested mostly in the economic behavior and success of capitalist firms. Importantly, decades after Raymond Williams (1977) imploded the overbaked base–superstructure relationship, economic geographers began to engage with meaning, symbols, and differing social practice as a constitutive part of the foundation of economic and capitalist production. Ultimately, by the mid-2000s, the distinctions – more so there than in cultural geography – between the “cultural” and “economic” aspects of social life were reconceptualized and described as the cultural economy.

Critical work on urban economic geography and on the capitalist state pulled important lessons from Foucauldian and feminist understandings of biopower and discourse. Likewise, urban and economically oriented work on critical race studies and the prison industrial complex (e.g., Ruth Gilmore’s work) were central contributions to the critical side of geography’s cultural turn. Some of the most controversial work from economic geography’s cultural turn is on diverse economies and the epistemological politics of capitalism (e.g., J.K. Gibson-Graham’s work), which aims to subvert conceptualizations of the “organic totality” of global capitalism toward envisioning and enacting “new class futures.” They claim that the power, scale, and coherence of capitalism are determined by our discourses, which ascribe it those characteristics. While knowledge and systems of meaning are foundational to understandings of economic power, the question of what kinds of epistemological politics are appropriate for which kinds of capitalist subjection remains open, if hotly contested. Here, the post-1968 political question of how to avoid reproducing what is being fought is central to ongoing critiques of capitalism. But, particularly after the 2007 world financial crisis, the macro-conditions – even if produced through myriad micropractices – of the capitalist mode of production still demand to be studied. Though much work from the 1990s and 2000s focuses on critiques of neoliberalism, perhaps to this end, they have tended to sidestep the questions “What is capitalism?” (of which neoliberalism is a specific articulation) and “How do we transform it?”
CULTURAL TURN

In geopolitics, the cultural turn brought Foucauldian knowledge/power and Gramscian hegemony into the study of relations between states as relations based on more than the direct exercise of force. Toal’s (1996) Foucauldian engagement with the principles of geopolitics and Agnew’s (1998) Gramscian utilization of hegemony pushed discussions about American geopolitical power beyond notions of simple dominance. Although raw military force remains important, the development of American international power has to do with a foundation of institutional, philanthropic, corporate, and other agents of building consent for American leadership. With this came research into the internationalization of the nation-state as a mechanism for building and maintaining hegemonic leadership in geopolitics. Here as well, the Washington Consensus, which aimed at liberalization and fiscal discipline, and at neoliberalism more broadly, became the framework for understanding forms of consensus in international development and economic policy, enacted within various governmental and nongovernmental institutional pressures. Scholarly engagements with protests at the disruptive nature of these projects and resistance to development projects more broadly also incorporated lessons from Gramsci. By focusing on the construction of class alliances and social movement tactics that would help scale up a social movement’s reach and impact, these movements engaged in Gramscian “wars of position” over accepted ideas and appropriate questions to ask so as to eventually capture elements of state power and policy. Some recent work on this has questioned the effectiveness of traditional movement tactics involving large urban protest spectacle, especially on the use of media for jumping scale, due to the increased police control over permissible public spaces for protest.

Many critiques of culturalist approaches exist, some from within and others from without the broad cultural turn. Mitchell’s famous and still controversial 1995 critique from within, “There’s no such thing as culture,” challenged the very notion of culture as either a causal object or as a sphere or realm distinct from other aspects of social life. While the theories and methods of new cultural geographers rejected economic reductionism, Mitchell’s critique suggested that in doing so they reified culture in these separations. Drawing on work by James Clifford and echoed by critics like Terry Eagleton, the idea of culture became increasingly disenchanted in geography. Culture, it seemed, either as systems of meanings or as localized diverse uses, constantly deferred the questions of social power, politics, and how it was produced. Further, and more to the crux of the matter, since the 1970s, “culture” had become the most fashionable means among policymakers to blame the poor for their impoverishment, rather than looking to exogenous, economic factors. Here, in policy debates about the so-called culture of poverty/welfare/entitlements is where the seeming arcane debates about culture’s ontological status have clear political force and implications. Beyond critique, Mitchell’s work on culture, more importantly, aimed to reconstruct the object of study in cultural analysis and the bases of culture’s production within the limits, pressures, and material social power of political economy. But in Mitchell’s critique, unlike others, the production and politics of meaning remain crucial to the mission of critical geography.

By the mid-2000s, a number of new theoretical directions – drawing from environmental/ecological crisis, quantum physics, biological and neurosciences – had taken root in human geography and in cultural studies, which challenged several of the foundations of
the cultural turn. Affect studies, actor-network theory (ANT), a set of loosely confederated “new” (non-Marxian) materialisms, and a broad engagement with Deleuze and Guattari, directly disputed humancentric orientations, meaning in social exchange, and historical processes. Like an echo of earlier determinisms, affect studies and so-called nonrepresentational theories push previous understandings of biopower beyond the discourses of norms and bodies into the direct modulation of feeling and affective response, which happen before and thus determine meaning and choice. From that vantage, the political struggles around Gramscian hegemony seem defunct on ontological grounds, since meaning, language, and representation are no longer presumed to be the operative mechanism of political life. Debates commenced, subsequently, about “posthegemon” theory. Further, the insistence of ANT and theorists of “vibrant” matter on the equality of humans and nonhumans as agents or social actors, and on human consciousness as a kind of determined effect, is like the metastasization of Althusser’s antihumanism, against which Thompson wrote so pointedly. Many of these approaches minimize or outright reject systems (e.g., capitalism), history, material fixity, and other concepts that would either extend beyond locally observable phenomena or determine a world seen as radically immanent and contingent.

In the contemporary period, the tensions and paradoxes subsumed within the cultural turn have been extended and pushed to their limited conclusions. Gramsci and systems of meaning, which determine the production and use of material life, are still important in human geography but exist in contention with these recent challenges more so than they did in previous decades, during the flourishing of the cultural turn.

To the extent that they have affected radical politics, these new challenges have left them even more constrained – resistance confined often to the accidental and social meaning subordinated to the agency of inanimate objects. The extension of micropolitics and intensification of antielectoral political tendencies, however, is unreflective as a rejection of previous (pre-1968) failures, in its lack of comparative engagement with other potential strategic paths and with subsequent failures and incongruities. As such, the insurgent view of Gramsci’s irrelevance increasingly limits the question of where Gramsci and, for example, the affective may overlap or where meaning and mindfulness may intercede in or perhaps overcome the determining attempts at affective control. Again, the post-2008 global financial collapse demands a reconsideration of economic systems and history, not to mention a re-engagement with notions of a world public and global political organization. Even further, the focus on objects and the agency of matter misses many of the determining – real limits and pressures of inheritance, profit and interest, and so forth – and particular meaning-related processes that produce the objects in our social world, one of the key principles in landscape studies, for instance. The cultural turn’s critique of positivism’s or empiricism’s lack of skepticism of the world as it appears is implicit here. When it is incapable of addressing these, the new materialism of geography requires a cultural re-turn but one that further engages the materiality of meaning and struggles for consent and studies the production of systems of meaning through political economic force and the technological composition of that force.

SEE ALSO: Economic geography; Geopolitics; Landscape; New cultural geography; Nonrepresentational theory; Postcolonial geographies; Poststructuralism/poststructural geographies; Subculture of poverty
CULTURAL TURN

References


Further reading


Culture is a central concern of geography, for its perceived role in shaping spatial variations in social relations and human–environment interactions. There is also great diversity in how culture is defined and used in the production of geographical knowledge. It can be simultaneously a verb and a noun, a category and a quality.

Defining the term literally, as a verb, links culture to practices of cultivation – to grow, extend, feed, or rear – as in agriculture, horticulture, or viticulture. To culture is to steward or produce material objects and artifacts in particular places and times. This is one of the older interpretations of culture in geographical thought, and is commonly found also in cognate disciplines such as anthropology and archaeology, where the study of material culture (artifacts, crafts, objects) is common. Indeed, in the colonial period culture was pivotal to European understandings of what it was to be human, distinguishing Homo sapiens from nonhuman species. Charles Darwin (1871, 122) himself famously argued that “The highest possible stage in moral culture is when we recognize that we ought to control our thoughts.” Cultivated practices distinguished humans from other nonhuman nature. A capacity for cultivation enabled humans to secure food production in ways that are seemingly beyond “natural” laws and ecological limits. Such antecedents set in train a linguistic and ontological binary dualism between “culture” and “nature” (the latter positioned as antonymic to the former) which would later come under fire from critical geographical scholars.

Other nuances of meaning in the term also derive from the colonial era. Caught up in the Enlightenment pursuit of scientific knowledge, geographers were important public intellectuals mapping regions hitherto unknown to Europeans, observing difference and attempting to theorize extant variation. Culture, understood as practices of cultivation, was a central part of this context. The dominant thinking was that geographical variants in aspects of nonhuman nature (including the distribution of soil, landform, climate, and geological types) determined agricultural practices and differences. Nature explained observed variations in the geography and manner of human social, economic, and political development. Human culture and its geographical diversity were, in essence, thought of as manifestations of underlying physical biophysical variability. Such ideas would later be heavily critiqued within geography for their reliance on environmental determinism at the expense of human agency and contingency.

Throughout the same period, culture slipped into usage as a noun, denoting a quality, something possessed by a society that has progressed or “civilized” to the point of developing a set of shared norms, behaviors, and rationalities. This explains the related term “cultured.” Such a meaning was absorbed into European ways of thinking and acting hierarchically. Deeming themselves “cultured,” Europeans claimed technological and intellectual superiority. Such claims fused with prevailing imperial and social Darwinistic thinking, whereby humans were classified into different cultural types (e.g., “races”), and their humanity ranked (e.g., as...
CULTURE

“less human” savages, or as “children” who are in a stage of civilizing development). At its apex, such thinking linked culture as an achieved human quality with technological and imperial might, at a time when biological or organic theories of geopolitical expansion were manifest in the extreme fascism of Nazi Germany.

Related to this colonial legacy, but without its violent overtones, culture has often been used as a synonym to describe, often in a hierarchical manner, the creative expressions, and intellectual and aesthetic achievements of talented people within particular societies. People who understand and value intellectual and creative expressions are often positioned as critics, agents, and managers – arbiters of taste responsible for differentiating between what is considered “high” art (the domain of “cultured” people) and “common,” popular culture, such as pop songs and television shows. Influenced by French theorist Pierre Bourdieu, as well as the Birmingham School of Cultural Studies in the late twentieth century, such thinking was critiqued for its links to a politics of class division and tension. Nevertheless, such distinctions linger on, in for instance demarcated cultural precincts within cities (which typically feature an art gallery, museum, civic library, theater, or performing arts center) and in preferential government financial support for the “fine” arts as part of their nation-building and place-branding efforts.

Perhaps the most common definition of culture is as an identifiable way of life. In this definition, culture is understood as a set of shared values and beliefs based on language, religion, customs, or ethnicity. These cultural traits are expressed in material cultures, including crafted products, clothing, foods, buildings, and occupations, as well as in collective tendencies, political predilections, and behaviors. Such notions of culture are understood as “superorganic,” in that they suggest that culture is an independent, active force working on or through people and possessing casual power beyond that of the individual. Culture could be deployed to make generalized claims about all people within a particular geographically bounded area, termed a cultural realm or region.

In geography, culture as a way of life remained a dominant and largely uncontested meaning for over six decades from the 1920s to the 1980s. It was especially pervasive in the United States, where the influence of Carl Sauer, professor of geography at the University of California, Berkeley, was profound. Sauer sought to reinstate humans as active agents in the production of landscapes. In “The Morphology of Landscape,” Sauer argued that humans possess the unique capacity to modify landscapes (Sauer 1925). The Berkeley School of Cultural Geography, led by Sauer and his disciples, became more interested in the diffusion of culture traits (rather than in evolutionist generalizations of their origins), the identification of culture regions (in both material and nonmaterial terms), and the role of culture in conditioning human perceptions of nature and the environment.

Since the 1980s, however, the successive influences of poststructuralist, feminist, multicultural, queer, and more-than-human thinking on contemporary geography have challenged the primacy of viewing culture as a superorganic force. Drawn to contest categories and to reveal them as contingent and unstable rather than fixed, many contemporary cultural geographers resist any attempt to approach culture as a cause or to define culture as a thing, type, or even subdisciplinary field. Don Mitchell (1995) famously quipped that “There is no such thing as culture,” implying that culture is a fabricated category, a construct of ideology made to perform the work of power relations, without any real material existence.
Such arguments have meant that culture now occupies an ambivalent position in contemporary geographical thinking and research. Prevalent nowadays is an interpretation informed by a less everyday sense of the cultural: culture as knowledge production. In philosophical terms: What exists in the world? And how do we know what we know about the world? Such questions encourage researchers to identify links between knowledge production, power, practices, behaviors, and possibilities for change and transformation. Culture as a way of thinking and knowing is informed by a concern to examine how particular ideas are established, circulated, maintained, and challenged.

But even this sense is being superseded by a push from within the discipline to rematerialize culture in analysis. The previous emphasis on cerebral discourses, ideas, knowledges, and representations has been at the expense of acknowledging embodied interactions and the materiality of everyday life. It has also privileged human cognition and agency over the experiences and capacities of nonhuman others. Culture is more than mere ideas, and involves humans and nonhumans struggling for survival and meaning amid ongoing, unfolding, and emergent relationships – be they fleeting or more permanent – between animals, plants, material things, and bodies. At the vanguard, culture has almost slipped from the geographical lexicon, because of its deeply problematic association with a stable and fixed category of the human as cognitively superior and ontologically distinct.

What remains, then, is a highly variegated landscape of thinking and writing about culture across geography. How culture is adopted and used itself varies geographically within academic research – influenced by the degree of adherence to regional, national, and/or linguistic traditions. In the United States, where the legacy of Sauer remains strongest, culture as way of life is still a common understanding. In the United Kingdom, where the “cultural turn” of the 1980s was more thorough, waves of poststructuralist, relational-materialist, and posthumanist interpretations have seemingly distanced questions of culture from the superorganic, regional tradition. Elsewhere are hybrids of influences, blending global theories with endogenous factors and traditions (including language itself – an imprint of culture that has shaped, for instance, distinct French and German geographical traditions). In Singapore, China, and much of Asia culture in geographical research intersects with questions about national and religious identities and is therefore as much a topic of political geography as of cultural geography. In Canada, Australia, and New Zealand questions of culture in geography also bring to the fore their anxieties as settler-colonial nations still struggling with colonial injustices toward indigenous peoples and seeking to accommodate indigenous rights, practices, and traditions. In such circumstances, culture signifies not just a way of life or a sphere of meaning and identity, but a deeply politicized area of struggle for recognition and respect.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Cultural geography; Landscape; Nature; Nonrepresentational theory; Race and racism

References

CULTURE


Further reading


Cultures of nature

Lesley Head
University of Melbourne, Australia

Culture and nature

Geography has a long-standing indeed foundational interest in the relationships between culture and nature (and their parallel distinctions “society and nature,” “people and environment”). Study of how such relationships have varied over time and space has been a fundamental part of the discipline, and for many people it is embedded in the division of the discipline into human and physical geography. Both sides of the discipline (and the sciences generally) have approached such relationships within a humanist tradition, which assumes fundamental differences between humans and the rest of nature, whether living or nonliving. Phenomena such as imagination, conscience, ideology, worldview, and planning do not have their direct counterparts in other living organisms. So plants, animals, earth, water, and climate were approached in physical geography using the observational tools of the natural sciences.

The term “cultures of nature” refers to scholarship over the last two to three decades in which those categories and divisions became not so self-evident, largely due to human geographers taking on subject matter traditionally understood as the preserve of physical geography. The history of this research effort is often traced to Margaret Fitzsimmons’s (1989) call for human geographers to pay more attention to the idea of nature and to social natures. Various authors seen as part of the “cultural turn” in geography, influentially Kay Anderson, Sarah Whatmore, and Bruce Braun, drew attention to the implications of culturally variable understandings and perceptions of nature (e.g., Anderson 2003). In particular they argued that the clear binary or dualism between culture and nature was neither empirically nor ontologically valid. The cultural turn was also influenced by parallel trends in the sciences that systematically deconstructed the notion of humans having a special or distinct place in nature. The nature of human difference and similarity from other organisms became an open empirical question, not one to be approached with assumptions of essence.

Since then there has been a great flourishing of work across a number of disciplines, but particularly cultural geography, seeking to analyze and deconstruct the separation of culture and nature and think about it differently. Emphasis is placed on relations rather than essences; phenomena are seen to develop their characteristics in relation with a host of others. Alternative concepts advocated at different times by different authors include hybridity, actor-network theory, assemblages, and posthumanism influenced by key thinkers outside geography including Donna Haraway, Bruno Latour, John Law, and Annemarie Mol.

Components of “nature,” and its companion concepts

Nature in these debates is generally understood to encompass a wide variety of nonhumans. A
particularly strong body of work emerged under the umbrella term of animal geographies (Wolch and Emel 1998). Human interactions with many different species, including fish and birds, were studied, connected to a broad variety of topics. Many stemmed from animal welfare concerns, including in the context of farming and eating, and the consequent ethical debates. These deal with issues of subjectivity closely associated with concepts of mind and consciousness – how might we understand animals against the norm of the human subject? Hence geographic approaches drew in new work in philosophy and animal biology. The spatial context of many studies was urban – encompassing both pets and non-domesticated animals – helping to break down the parallel binary of cities as places of culture and rural areas as places of nature. Studies in protected area contexts included fraught issues such as the reintroduction of wolves in North America and northern Europe.

New approaches to organisms somewhat more different from the human emerged more slowly but have flourished in recent years. An influential work on trees (Jones and Cloke 2002) led plant approaches, which now encompass gardens, food, and invasive plants as well as more conceptual work. Fungi, microbes, and bacteria have been the subjects of recent critical study. While many animal geographic studies are framed by the similarities and commonalities between humans and other animals, work on bacteria or weeds has to start with the question of difference. Such work is no less relational, but it has forced different kinds of thinking about subjectivity and agency, which have then to be discussed without recourse to human or at least mammalian concepts such as mind and consciousness. This is more difficult but also opens up exciting new possibilities.

At the boundary between living and nonliving (a divide itself open to critique in the natural sciences) is a set of works that include among their nonhumans water, air pollution, weather, seawater, and the indifferent earth itself (e.g., Clark 2011). Clark has stimulated important discussions about what constitutes the political in inhuman nature. Now increasingly referred to as the “geological turn” in human geography, it remains to be seen whether this provides a basis for reforging links with physical geographers.

If the nature/culture binary was to be challenged, so was a set of companion concepts which assume their distinction. Foremost among these were wilderness and the wild, built out of long-standing but complex intellectual Western traditions of pristine nature. Indeed empirical challenge to the concept of pristine wilderness, founded on archaeological, indigenous, historical, and paleoecological evidence and picked up in a number of human geography studies, constituted one of the earliest contributions of cultures of nature scholarship. There are strong spatial contexts and applications in these debates, for example, when an understanding of national parks as pure nature is manifested in maps, texts, and images, thus drawing a line on the ground between natural space inside to be protected from cultural space outside. Boundaries, borders, and belonging are common themes in the literature. For example, vociferous debate over the appropriate places for native plants and animals, compared with invasive or feral species, shows how concepts related to nature (e.g., nativeness, alien, feral) have material outcomes with implications for which organisms are privileged over others in different spatial contexts.

**Cultural variability**

Broadly speaking, “cultures of nature” research has built an extensive body of work in which the culturally variable engagements with that
which is understood as nature are documented. Culture was here understood in vernacular and everyday terms, not just as “high” or elite culture. It is understood to be very dynamic and to pervade and be part of discourses, practices, and beliefs. Comparative studies of indigenous and settler engagements with land were a fundamental building block in Australia, New Zealand, Canada, and the United States of America; influenced by the rising consciousness of indigenous land rights and associated political struggles in the last several decades of the twentieth century. This work had purchase partly because of its very clear relevance to debates that were happening around world heritage sites, biodiversity protection, protected area conservation – it became clear, albeit contested, that mainstream settler land management processes were not themselves self-evident but had a cultural dimension and a number of embedded assumptions.

Indigenous environmental knowledge and modes of environmental interaction are not only interesting for their own sake, as an important dimension of cultural variability. Rather, there are also conceptual challenges attendant on taking seriously indigenous understandings, for example, how they help identify capacity deficits in existing natural resource management institutions and practices. Indigenous and other non-Western ontologies offer a range of cultural resources in the necessary reframing of relationships between humans and their environment. The Australian indigenous concept translated as “Country” provides an important example (Suchet-Pearson et al. 2013).

A flourishing body of work proceeded to analyze the gender, age, class, and religious dimensions of human interaction with the nonhuman world. Studies from the affluent West dominated much Anglo-American cultural geographical discourse with considerable geographical variation in how the issues have been dealt with. There were productive interactions with urban studies, for example, Kong’s (2000) work on children growing up in mostly apartment contexts in Singapore. As seen also in the examples of animal studies and gardens, cultures of nature research in urban contexts (often referred to as urban natures) helped challenge a further binary, that between the urban and the rural. An emerging feature of research in settler societies and those Western societies with high migration is greater attention to the role of ethnicity, and in particular nonwhite ethnicity and nature (Buijs, Elands, and Langers 2009).

In needing to maintain a very dynamic view of culture, researchers had to consider the intersections of different influences, and also consider cultural change over time. For example, Australian anthropologists discuss the ways Christianity has been absorbed into traditional Aboriginal worldviews as an outcome of the colonial experience and the implications of this for landscape attachment and understandings of country.

Methods

People make and remake culture as individuals, communities, and in institutions and academic disciplines. They do it in everyday practices and over both long and short periods of time. It follows that to understand environmental cultures or cultures of nature we need fine-grained methodologies that enable understanding of motivations, behaviors, contradictions, tradition, change, and flexibility in requisite depth. Cultural research pays a lot of attention to that which is taken for granted and understood to be common sense. It also follows that cultural research methods can be applied to the examination of how large-scale institutions, economies, and structures of meaning are created and maintained, and how different understandings of
nature are embedded in organizations charged with land and natural resource management on the ground.

In keeping with the diversity of the subject matter, a wide variety of methods are used in research on cultures of nature. Most common are qualitative methods in the ethnographic tradition where there is overlap between cultural geography and anthropological methods. In-depth interviews and participant observation are frequently undertaken with different cultural groups, for example, with indigenous groups in their own country. Many studies have undertaken media and textual analysis as a means of understanding the cultural discourses around nature and their power within society. Recent work focuses particularly on embodied practices such as walking. Quantitative approaches include surveys and photo-elicitation approaches, in which broad trends are established, for example, exploring which groups of people prefer which landscapes. Other researchers have used geographic information systems to map places where people feel most connected to nature.

Extending traditional qualitative methods more recently are various forms of embodied and performative work, which pay attention to the emotions and affective dimensions through careful examination of processes of walking, looking, listening, and touching. Angelo’s (2013) research on relationships between people and birds provides an instructive example. Working with both birdwatchers and ornithologists, in a context that included conflict over the killing of birds for taxidermy, she shows clearly how “experience makes nature, differently for different people and over time” (Angelo 2013, 367).

A controversial extension of the methodological challenges in going beyond the human–nature binary is the idea of multispecies ethnography (Kirksey and Helmreich 2010), which seeks to bring animals, plants, fungi, and microbes into the conversation as subjects, rather than as the traditional objects of ethnobiological approaches. This is prompting much discussion of the extent to which ethnography is an inescapably human-centered (anthropocentric) method.

New ontologies, yet persistent dualisms

A paradoxical contribution of cultures of nature scholarship is to challenge if not undermine the value of its two constitutive concepts. New traditions, sometimes referred to as posthumanist, argue that the binary between humans and environment, or society and nature, is ontologically impossible to sustain. Posthumanist perspectives critique “the human” as an essentialized and unified category, and further emphasize relational approaches, whereby the characteristics of phenomena are constituted in the process of their relationships with other phenomena. Important contributions in this area have come from materialist approaches that illustrate the fundamental material basis of life. A further implication of the materialist approach is that the distinction between living and nonliving matter seems to disappear, so technology and materials become part of the debate. Particularly rich contributions in this vein are seen in work on biosecurity, quarantine, and genetically modified (GM) foods, which draw in science, disease, laboratories, and public fears to the concerns of some cultures of nature research.

Castree (2014) makes a strong argument that the culture–nature dualism cannot be wished away, because it is deeply embedded in social and mental structures. It is in some ways an inescapable dilemma: “Our life condition appears to be ‘both/and’ rather than ‘either/or,’ obliging us to use the contradictory ideas of nature as ‘external’ and ‘universal’ when discussing ourselves” (Castree 2014, 29). His argument is that scholars need to keep paying attention to how
the concept and notion of nature is mobilized in many different ways in modern society. It will not disappear just because criticized within academia.

Davison (2008) illustrated the complex interdependence of dualistic and nondualistic understandings of nature within a study of urban environmentalists: “ideas of untouched nature exist in complex interdependence with non-dualistic understandings of the seamlessness of social and natural existence in the lives of many environmentalists.” The latter particularly relate to their suburban existence and everyday rhythms of

Walking and weeding in nearby bushland, growing and preserving food, sharing homes with companion animals, observing suburban wildlife, reading the moods of suburban mountains and coastlines, finding peace and privacy in suburban gardens: such everyday engagements revealed often subtle navigation of socionatural complexity. (Davison 2008, 1294)

Davison thus identifies untapped potential in these liminal understandings and experiences – not fully recognized by the environmentalists themselves. There are ongoing ontological and practical challenges in this complexity.

The “so what?” question

The point is not necessarily to resolve the tensions elucidated by Davison, but to recognize and use them. Research and scholarship going under the umbrella of cultures of nature often faces the “so what?” question. It often comes from physical geographers and natural scientists who ask how such research makes a difference on the ground. Embedded within that question is often a taken-for-granted view of the world and how things are, but the questioner is rarely likely to be sympathetic to deconstruction and critique for its own sake. It also comes from more politically attuned parts of the discipline, such as political ecology, where there is justifiable concern about insufficient attention to unequal relations of power in some connected or networked descriptions.

Critique makes an important contribution in its own right, and history already shows cultures of nature scholarship to have had a profound influence in some key environmental and land management debates. Understanding diverse perspectives provides a basis for clarifying and potentially resolving conflicts, for example, between environmentalists, miners, forestry interests, and hunters, all of whom have been studied within this tradition. This scholarship has also shown how concepts and categories of nature, and of culture, become embedded in societal structures, and how these concepts have power once they are incorporated into legislation, institutions, and public discourse. A notable example is the question of heritage, and the relationships between “natural” and “cultural” heritage used by UNESCO and many member countries. These categories of heritage have required significant revision over the last two decades, in direct response to cultures of nature research.

A further contribution is providing critical and historical analysis of what stands for common sense and normal behavior and how it came to be. Understanding environmental norms and practices helps identify both barriers to and thresholds for change in the context of contemporary sustainability challenges. For example, attention to cultural variability within and between households allows resources for change to emerge from unlikely places, including among older people who explicitly do not identify as “green” but who nevertheless have a number of sustainable practices based on frugality and not wasting.
Further, the influence of this scholarship throughout the discipline can be recognized in seemingly unrelated debates. It is scarcely possible now to discuss key questions in urban geography without explicitly considering nonhuman natures. And there is increasing recognition among scientists and policymakers that sustainability and climate change issues have significant if not easily tractable cultural dimensions.

Despite the persistent dualisms, there are good reasons for pushing forward to better conceptualize relations between humans and the rest of nature. The evidence of the Anthropocene, that humans are embedded in all Earth surface processes, has confirmed that the binary between humans and nature is both empirically and thus ontologically questionable. We have no choice but to think differently. The key contribution of cultures of nature research may yet be that it helps us all to think in such a way that the concepts, and thus the terminology, have been superseded by something better.

SEE ALSO: Alien and native species; Animal geographies; Borders, boundaries, and borderlands; Cultural turn; Hybridity; Indigenous knowledge; Nature; Posthumanism; Qualitative data

References


Cultures of work

Trevor J. Barnes
University of British Columbia, Canada

A culture of work is a distinct and stable set of shared values, norms, meanings, institutional forms, artifacts, and embodied practices that unifies a group of workers as they labor at a given site of employment. A culture of work is not separate from the labor process; it is not an extra, something that is added. Rather, it is there from the very beginning, joined with the labor process, indissolubly connected to it. It shapes how work is performed; the relation that workers have with one another, including with their bosses; how things are done and by what means; and, even more fundamentally, what counts as work in the first place.

Recognizing a culture of work is important because it demonstrates that work is more than just a bare economic transaction. Orthodox (neoclassical) economics conceives of work as only an economic calculus. In exchange for expending physical effort and/or mental energy to complete tasks set by an employer, the worker receives a wage. The wage is the price of a laborer's work. In neoclassical economics the determination of the wage rate in the labor market is the only interesting fact one needs to know about work. Once the price of labor is determined everything else follows. Such a position is contested by the idea of the culture of work, however. Rather than the wage rate in the labor market determining everything about work, it is everything outside the labor market bearing on work that determines the wage rate. This position is forcefully argued by the French sociologist Pierre Bourdieu (whose work will be discussed further). Orthodox economics begins with an economic relation, the determination of price in the marketplace that then determines all other noneconomic relationships. Bourdieu argues the reverse – that it is the noneconomic that determines the economic. As he puts it, “it is not prices that determine everything but everything that determines prices” (Bourdieu 2005, 77). The culture of work is part of the noneconomic, and gives it central importance.

One more general point: the culture of work is not singular but plural. There is not one culture of work but many. Even at the same geographical site, there can be a multiplicity of work cultures. These cultures may overlap at different points or may be separated; they may be tacit or formally articulated; they may be imposed from above or arise organically from below. However, to be a culture of work it must be shared. A culture of work never pertains to just one individual worker. It must have the capacity of being taken up, learned, transmitted, and imparted to other workers in either the same place or in other places even if they are half a world away. Exactly how that culture is taken up, learned, transmitted, and imparted varies geographically and historically, but the existence of such a capacity is essential.

This entry is divided into two main parts. The first provides a conceptual framework for understanding the culture of work. As an idea, it stems from political economy, but since the 1970s it has increasingly severed those ties, becoming more squarely associated with social and cultural theory. The second provides a set
of illustrative empirical cases drawn mostly from research in economic geography, the discipline within human geography most concerned with work and its culture. As a field it has progressively moved away from orthodox economic analysis to approaches emphasizing the social and the cultural.

Theorizing cultures of work

Workers employed in England’s “dark satanic mills” of nineteenth-century industrial capitalism were generically called “the Hands.” It is a term Charles Dickens memorably satirized in his novel *Hard Times* (1854), set in the ultimate (fictional) gritty industrial city of Coketown. Dickens’s use of the label “Hands” was symptomatic of the cruelty of industrial capitalism, which reduced the wholeness of a worker to only two bodily appendages. That, at least, was how Mr Bounderby, the antihero capitalist at the center of the novel, thought of his workers. But, Dickens suggests, and as illustrated by one of the novel’s characters, Stephen Blackpool, workers were much, much more than merely a pair of hands. Work as Dickens portrayed it was always embedded within a larger setting of cultural sentiment and meaning, that is, within a culture of work.

Karl Marx wrote at the same time as Dickens and in the same city. He too was concerned with the dire and horrific character of work and the fate of workers within English industrial capitalism. But, whereas Dickens’s criticisms were couched in terms of morality and sentiment, Marx’s were framed by a rigorous and stark political-economic logic. Marx believed that logic incontrovertibly proved the unsustainability of industrial capitalism. His analysis consisted of many elements, one of which bore specifically on work, and which he formulated as worker alienation. Alienation occurred because of a disjunction between the nature of workers as fully rounded human beings and their transformation into less than humans under capitalism. At work they were deployed merely as things, as inputs, part of the machinery of production. This was at odds with their real nature as sentient humans. Consequently, they were alienated from their product, from their work, from themselves, and from other workers. Workers lost any culture of work. They were simply grunt laborers, employed as Hands, and for nothing else. Marx’s political project was a restorative one: to eliminate worker alienation by re-establishing a culture of work, which he thought would be possible only through a working-class revolution. It was his analysis that initiated a political-economic approach to the culture of work.

Dickens and Marx, in their different ways, both pointed to the need to provide a wider discussion of work that went beyond simply a description of acts of labor and its remuneration. There was a need to situate work and workers within a larger social and cultural context. Subsequent depictions initially followed the political-economic tradition of Marx, but were later supplemented and then overshadowed by explicitly culturally inflected analyses.

Political economy

Marx and his collaborator Friedrich Engels were concerned primarily with factory workers and their work. In volume 1 of *Capital* (1867), Marx relied on reports completed by factory inspectors to document the appalling conditions under which workers labored. The shocking excesses of the descriptions made it clear that work was never simply an economic calculus. Engels saw those factory working conditions up close in 1842 when his father, part owner of a
textile mill in Manchester, sent the 22-year-old there to learn the business. Apart from allowing Engels to record working conditions, this also gave him the opportunity to document the pitiful and hazardous state of working-class urban life in Manchester. During his time off, Engels roamed the city, writing up his observations as *The Condition of the Working Class in England* (1844).

For there to be a change, Marx and Engels argued, there needed to be a working-class revolution, which would happen in part as a result of the alienation of workers at work: workers would rebel against their horrific treatment. But to mount an effective revolution, workers needed to collaborate, to act collectively, to forge a class consciousness, to become a class for itself. In turn, that required a working-class culture, a culture of work.

That was the thesis at the heart of the British historian E.P. Thompson’s now classic volume, *The Making of the English Working Class* (1963), which documented the rise of an English working-class consciousness in the period 1780–1832 which was intimately tied to the experience of work within early industrial capitalism. Perhaps the most important part of Thompson’s title was the present participle “making.” It signaled a commitment to understanding the working-class consciousness as an ongoing process that, historically and geographically, took variegated forms. How workers came to consciousness at their workplace, and the form and consequences that consciousness took, differed substantially across period and place. As Thompson (1963, 9–10) wrote:

Class-consciousness is the way in which experiences [at work] are handled in cultural terms: embodied in traditions, value-systems, ideas, and institutional forms. If the experience appears as determined, class-consciousness does not. We can see a logic in the responses of similar occupational groups undergoing similar experiences, but we cannot predicate any law. Consciousness of class arises in the same way in different times and places, but never in just the same way.

Thompson was concerned with the emergence of class consciousness, and its equivalent the culture of work, during early forms of capitalism. Those forms were in continual flux, however, often undergoing radical transformation. As Marx and Engels (2010/1848, 16) put it, “all that is solid melts into air.” Capitalism was always changing its spots, and as it did so new cultures of work appeared. The “machinofacture” that Marx, Engels, and Dickens described, in which work consisted of feeding inputs into machines, had given way by the first part of the twentieth century to a new form, Fordism, in which workers in effect became part of the machinery itself. It was a new culture of work.

The culture was clinically analyzed by Harry Braverman in his book *Labor and Monopoly Capital* (1974). The beginnings of that work culture lay in the late nineteenth-century time and motion experiments carried out by Frederick W. Taylor (1856–1915). Taylor was concerned with designing workplace strategies that would squeeze from workers the greatest amount of work from the least amount of work time. To formulate those strategies, Taylor carried out experiments on workers, varying their work conditions and timing and recording their performances. From the results he developed what he called three principles of scientific management that could be applied to all laborers, and which would maximize efficiency at work. The first was dissociating the labor process from the skills of the laborer. This was what Braverman (1974) later called “deskilling.” Workers were not employed for the manual skills they possessed but for their capacity to carry out a minimal set of repetitive simple tasks that required no skill at all. For Braverman, the deskilling of labor became ever more marked and pervasive.
as capitalism developed historically. The second was the separation of conception from execution. Workers were paid not to think (conception) but only to do (execution). Taylor found from his experiments that thinking took time, producing inefficiencies and requiring elimination. Workers were most efficient when they became part of the machinery, carrying out rote mechanical acts without thought. The third was to use “task cards” to specify exactly what was to be done, how it was to be done, when it was to be done, and how long it should take.

The new culture of work, Taylorism, was perhaps best seen in the factories of mass manufacturing production that emerged in the early twentieth century. It was exemplified by the Rouge automobile plant in Dearborn, Michigan (by 1928 the largest integrated factory in the world), built by Henry Ford, where the production method combined assembly-line techniques with the systematic deployment of principles of scientific management. The culture of work was for workers to stand in place along the assembly line and robotically carry out the same single minimal task over and over and over again. It was culture that was spoofed by Charlie Chaplin in his film *Modern Times* (1936) and (strangely) celebrated by the Marxist Mexican painter Diego Rivera in his Detroit Institute of Arts murals of workers at the Rouge (1932–1933).

Only a few years after it was published, the sociologist Michael Burawoy criticized Braverman’s work. In *Manufacturing Consent* (1979), Burawoy argued that the culture of work under Taylorism was defined less by draconian control enforced by authoritarian scientific managers (Braverman’s view) than by a consensual process in which workers agreed to their own domination. In making this argument Burawoy drew on the idea of cultural hegemony originally put forward by the Italian Marxist Antonio Gramsci (1891–1937). This was the notion that capitalists exerted social control not by strong arm coercive tactics but by winning consent from the larger population. Consent was gained by capitalists seemingly granting the masses what they wanted and meeting their interests. In reality, though, as Gramsci argued, the interests of the proletarian masses were never met. They only appeared to be met. Consent was secured under false pretenses, with the elite the only beneficiaries in the end. Burawoy took Gramsci’s idea and applied it to Taylorist workers. Under a regime of cultural hegemony, Burawoy argued, workers consented to their own oppression and deskilling. They mistakenly believed that Taylorism gave them choice and widened their work experience. That was why they consented. In reality it gave them neither, allowing capitalist employers to exploit and oppress them to an even greater degree.

Cultural hegemony was also central to Paul Willis’s classic ethnography, again written in the late 1970s: *Learning to Labour: How Working Class Kids Get Working Class Jobs* (1977). Willis explored not so much the culture of work as the culture at school that prepared working-class boys for Taylorist jobs in a factory. Conducting interviews with a dozen schoolboys in their last year and a half at a secondary modern school in Britain’s West Midlands (then at the heart of the country’s automobile industry), Willis identified two groups: “the Lads” and the “Ear ’oles.” The Lads did not pay attention at school. They were bored by it. Their main aim was “to have a Laff,” which they did by resisting school authority in various forms (Willis 1977, 14). The “Ear ’oles,” in contrast, paid attention, strove to complete their school work, and respected their teachers. Willis’s argument was that, for the Lads, the cultures of class and masculinity at the school prepared them perfectly for the culture of work in a factory into which they went when they left school. Work, like school, would also be boring and not require their attention,
but it could be endured by having a “Laff” and by circumventing and subverting authority whenever they could. This was a form of cultural hegemony because the Lads consented first at school and then at their workplace to their own domination. Of course, they did not see it in such terms, believing that their acting out provided them with freedom and agency when in reality it only met the interests of capitalists. Working-class kids got working-class jobs, and the larger system of industrial capitalism was smoothly reproduced.

Of course, it wasn’t always so smooth. Even during the golden age of Fordism, there were occasional downturns when it was difficult for working-class kids to get working-class jobs. From the late 1970s, however, there was a tectonic shift in the Global North, when massive deindustrialization and economic restructuring profoundly impacted employment, labor, and work. “Old Father Ford” didn’t immediately die, but he became very poorly, making those 1970s studies, and the masculinist assumptions that underpinned them increasingly redundant as male Taylorist workers themselves were made redundant. This change also brought about a tectonic shift in the literature on the culture of work. Up until the 1970s the culture of work, for the most part, meant men’s work. After the 1970s, however, and in step with the larger structural changes in the economy in the Global North (and for that matter in the Global South too), there was much more emphasis on examining the work of women. The old political-economic perspective, with its emphasis on the male working class, was out of sync with the new times. Instead, a different, more relevant, theoretical perspective was taken up that emphasized the cultural, drew on feminist and poststructural theory, and conceived of class as not only a narrowly economic category.

Cultural theory

Deindustrialization cut a grim swathe across Global Northern manufacturing regions from the 1970s, eliminating an enormous number of relatively high-paying, Taylorist, often unionized manufacturing jobs held by men. Such changed work practices became fodder even for popular culture. For example, the 1997 film *The Full Monty* (1997), later a Broadway and West End musical, featured a group of male Sheffield steel-workers who, when they were made redundant, took off their boilersuits and steel-capped boots and entered the traditionally female service sector occupation of exotic dancing.

While the film may have been a fantasy, it pointed to a fundamental change around work and its culture. The Global North was moving toward a postindustrial, service-based economy, turning upside down older Fordist nostrums of work. Employment now came in two main forms. There were the “McJobs,” which were low-paid and required few skills and little experience. Even in this general category there could be variation, depending on the particular culture of work in the McJob. At the higher end of the job hierarchy was employment in establishments of conspicuous consumption such as boutique cafes and designer clothing stores. At the lower end was flipping burgers in fast-food restaurants or as “associates” working the floor of big-box discount retailers. The other type of service sector employment was in the creative economy, or the new or cultural economy. Workers here were often highly paid, and frequently required skills that were formally credentialed. Work could be in existing sectors that had been reconstituted such as advertising or higher education, or in entirely new sectors like video game or web design. Both kinds of service work – the low end and the high end – required embodied interactions with customers and with other workers, calling for cultural, social, and even
emotional (soft) skills. Unlike in manufacturing, where one worker’s body was like another’s, simply the repository of brawn to be dispensed in the industrial process, in the service sector the body was often a vital part of the product sold. What kind of body made a difference, and how it was dressed, comported itself, spoke, and looked could be crucial. Out-of-the-box Taylorist management principles were still applied, especially to low-end service work. But typically there was scope, which increased as one moved up the service sector job ladder, for conception, even creativity and individuality, as well as execution.

Within this transformed setting the culture of work came to mean something different. It didn’t necessarily lead back to class, and to questions of social revolution, as it did under political economy. Instead issues of cultural identity around gender, race, sexuality, and age were central. And even when class was emphasized it was conceived differently, with a cultural inflexion.

Bourdieu’s work, especially his book *Distinction* (1984), made that clear. Bourdieu argued that economic differences in income and wealth were not the only, indeed were not even the most important, criteria in determining class membership. For Bourdieu class location was set by three kinds of capital that an individual controlled: economic (money), social (networks and contacts), and cultural (aesthetic tastes). The latter was not just an add-on but, for Bourdieu, often the most important determining influence, albeit interacting with the other two factors. Furthermore, Bourdieu argued, while economic and social capital were accumulated over time, cultural capital was learned at an early age, inherited from one’s family. Once one’s tastes were fixed, they were then difficult to change, and as a result ossified class distinctions.

The culture of aesthetic distinction channeled different class members into different jobs. For a number of service sector jobs, such as those in the hospitality sector, recognizing distinction was imperative. How one presented oneself, the clothes one wore, the words one chose, the accent in which they were spoken, and the manners one displayed all made a difference. They were the required culture of labor for that job. But it was a culture that one either had or didn’t have depending on one’s inherited cultural capital. In Yasemin Besen-Cassino’s (2014) study of youth labor in America, the possession of cultural capital mattered even for minimum-wage service jobs. Besen-Cassino researched student workers employed in hip coffee shops in an East Coast city. Although remuneration was minimal, in order to be hired and to keep their jobs students needed to possess the appropriate cultural capital: to like and to be able to judge the correct music, to be able to converse appropriately and on fitting topics, and to look right. The employees may not have had much economic capital, but they possessed cultural capital. Indeed, the economic capital they acquired was quickly turned into what Bourdieu called symbolic goods, commodities whose aesthetic qualities only further demonstrated their good taste, reinforcing their privileged class membership.

The work of the baristas that Besen-Cassino (2014, 40–42) studied was, as she noted, an exercise not only in cultural distinction but also in cultural performance. The dais on which the baristas made their cappuccinos and lattes was their stage, the customer line-up the audience, the physical layout of the cafe and the chairs and tables the scenery, their smart talk the script, their uniforms, theatrical costumes, and their accessories props. Treating cultural and social life as a performance has been discussed by scholars for at least the last half-century. It was found in philosophy in the work of J.L. Austin on “performatives” in sociology in the writings of Harold Garfinkel on “ethnomethodology,” and more recently in feminism in the
texts of Judith Butler (discussed further). The original use of the “dramaturgical” metaphor, as it has been called, was with the Canadian-born sociologist, Erving Goffman, and his book *The Presentation of Self in Everyday Life* (1959). His argument was that, in order not to embarrass ourselves and also to control what others think of us, we put on a performance. We use the stage setting and any appropriate props to conjure a positive image of ourselves, to present ourselves in everyday life as we would like to be rather than as who we really are. Afterward we return backstage, taking off the mask and reverting to our true selves.

Goffman’s dramaturgical metaphor seems especially suited to understanding service sector occupations, in which the culture of work plays such a central part. Besen-Cassino’s case study of college baristas is, of course, one such example. Another is Arlie Hochschild’s (1983) study in her book *The Managed Heart* which examined, among other occupations, female flight attendants. Hochschild introduces the idea that some workers are employed to manage emotions, both of those whom they are serving and their own. In both cases there is a performance of what she terms “emotional labor.” First, the worker tries to cultivate in the customers they are serving the “right” emotions, in the sense that they allow the customers to properly appreciate the service they are purchasing. Second, for this to happen and for it to be authentic, like all good actors, the workers must also induce within themselves the “right” emotional feelings while at the same time suppressing any “wrong” feeling that might disrupt their performance. In the case of female flight attendants, a major task is to quell any emotional fears passengers may have of flying. This is “managed” by the attendants persuading passengers that the flight is an exotic journey, luxurious even. Hence the smiles; the trolley of drinks and tasty snacks; the in-flight entertainment; the jokey, even flirtatious, chit-chat; the short skirts; the tailored uniforms; the coquettish headgear. But there is also the need for the flight attendants to manage their own hearts which, as Hochschild shows, can be difficult. Apart from boorish talk and occasional violent acts by passengers, they have to deal with their treatment by the airlines, which subject them to horrendously long and physically taxing work days; meticulous disciplining of bodies and appearance; and continual surveillance and control – moral, corporeal, and work – both inside the cabin and out.

Hochschild shows the dark underbelly of the culture of work, located within a larger force field of changing relations of power that are sometimes positive, often not. There are also connections to the French social theorist Michel Foucault and his concern with power, surveillance, control, and the body. Foucault’s ideas certainly link to earlier formulations of the culture of work put forward by Frederick Taylor. Taylorist workers are nothing if not controlled, disciplined, and surveyed. But Foucault’s interests apply perhaps even more so to the contemporary service sector and the bodies that perform it. The body is so important to Foucault because it is a site on which so many discourses are inscribed: gender, race, sexuality, disability, age, among others. The consequent cultural identities, in turn, shape what kinds of jobs people undertake, and their experiences at work.

The body as Foucault suggests can also be a site of potential resistance. This is a theme further developed in Judith Butler’s work, especially *Gender Trouble* (1990), which provides a critical feminist reworking of the idea of performance. For Butler gender is always performed. That performance takes the form of repeated iterations, which can mistakenly give the illusion that gender is fixed and natural rather than socially constructed and continually remade. Of course, there...
are immensely powerful regulative discourses that maintain the formative iterations of gender. But, and this is where there is at least the possibility of bodily resistance, one iteration may not be exactly the same as the one before, and raises the possibility of progressive change, however gradual. The idea that performances of the culture of work can produce at least a possibility of progressive change has become a strong theme in the geographical literature, and to which the entry now turns.

Cultures of work in geography

Within human geography, work as a topic has been most often discussed by economic geographers, and linked to discussions of the labor market. Very little has been written explicitly on the culture of work, however. Instead, it needs to be teased out from individual studies, which is the aim of this section.

From its very beginning, economic geography recognized that labor influenced the location of economic activities. Initially, that locational influence was conceptualized (as in, e.g., Alfred Weber’s classic location triangle) in terms of the cost of workers (i.e., wages) and not their culture. The first glimmer that the culture of work might also have a locational effect came with geographical studies of industrial districts undertaken during the 1940s. An industrial district is a tightly bound area within a city specializing in one particular type of industrial production, for example the garment district. The British economist Alfred Marshall, working at the turn of the twentieth century, was the first to recognize the phenomena of industrial districts, explaining them as a consequence of agglomeration economies. He came to realize, though, that this explanation went only so far, and augmented his account with an additional factor, “industrial atmosphere.” Subsequently, Marshall’s “atmosphere” was interpreted as the force of culture – the missing ingredient that explained why industrial production geographically clumped rather than dispersed. Specifically, the culture of work, in this case interpreted as common social ties and institutional affiliation linking similarly skilled workers, was used to explain why particular types of employees cluster in one part of the city, which then functioned as an inducing labor pool for an industrial district.

For the two decades or so after the end of World War II the topic of work mostly disappeared from the agenda of economic geography, and the focus was on labor and final output rather than on work and the production process. This began to change from the early 1970s with the emergence of political economy. David Harvey, the most prominent geographer writing within that tradition, made little of work (and even less of culture) at first. Labor for him was simply the kick-start of value – in his schematic, labor power (LP) – which then began a complex geographical circulation (Harvey’s central preoccupation). This changed in the early 1980s when the topics of work and culture became more prominent, the result partly of the Global North’s firestorm of industrial restructuring and deindustrialization that made work, or the lack of it, so prominent; and partly of the beginnings of feminist geography which started with a discussion about women and work, the so-called domestic labor debate. Those two approaches, political economy and feminism, opened up the culture of work to scrutiny.

The hollowing out of manufacturing led those in the political economy tradition to attend directly to work. This involved in part enumerating the hemorrhage of manufacturing jobs, but also describing changes in the nature of the industrial work that remained. With deindustrialization came industrial restructuring, a move from an older Fordist form of work based on
Taylorism to a new (leaner) post-Fordist form of work based on labor flexibility. Here workers learned the manual, could potentially operate at any workstation, contributed ideas to “quality circles,” could stop the assembly line when they saw flaws (the ultimate no-no under Fordism), and were paid to think as well as to do. It was a new culture of work, and which those from within the political economy tradition began to describe variously as post-Fordism, lean production, or flexible accumulation (Amin 1995).

Doreen Massey’s writings, especially Spatial Divisions of Labour (1984), were particularly important during this period. Massey applied ideas of culture directly to work. An especially potent case study was her analysis of the South Wales economy which, in the late 1970s and early 1980s, shifted away from its historical specialization of heavy industry (coal, iron, and steel) to the assembly of electronic goods. Primarily American and Japanese multinational firms were drawn to South Wales by the prospect of realizing a particular culture of work, which was to be realized not by men, the traditional employed worker in the region, but by women. The corporations wanted an educated, “greenfield,” and female labor market that was geographically trapped, and consequently would be cheap, pliable, and knowledgeable. Massey argued that the culture of South Wales had created just such a workforce. The culture of the male workers of South Wales’s heavy industries had in effect produced a pool of potential female workers who, while they had gone to school, had not generally worked outside the home because of the region’s masculinism and patriarchy. When retrenchment started in heavy industry, female partners went out to perform paid work, although they remained geographically constrained because of demands to continue domestic (household reproduction) duties. The American and Japanese multinationals were waiting for them.

The other strand pushing the culture of work in geography, and again one in which Massey was also central, was feminism. In the early 1980s feminist geographers attracted to political economy followed the larger socialist feminist literature which at the time was preoccupied with the question of female domestic labor and household reproduction. That literature identified the household as a site of (re)production. The household produced workers and future workers (children), ensuring that they were clothed, fed, and sheltered, ready to turn up to work (or school) the next day. But how should the unpaid domestic work of women, and who (re)produced the central commodity of capitalism, labor, be treated conceptually? On the surface it appeared as if the work women did in the home was different from laboring in a factory. It was different culture. Work disappeared. But the classical economistic and functional Marxist interpretation reduced domestic labor, women, and the political ends of feminism only to class position. Culture disappeared.

Partly as a result of this, from the late 1980s and early 1990s there was a break as feminists in economic geography moved away from Marxism toward poststructural theory, which emphasized the cultural. Often the focus was the body, both men’s and women’s, and the various (Foucauldian) social and cultural discursive inscriptions written on it. One of the first studies bearing on work from this new sensibility came from Doreen Massey who shifted her focus from a declining old industrial economic region, South Wales, to a brand new one, the Cambridge high-tech research triangle in England. In both regions, though, there was a distinct culture of work in operation. Massey, along with her co-author David Wield, showed in High-Tech Fantasies (1992) that the culture of work was inextricably linked to the culture at home. The seemingly disembodied discourse of science
CULTURES OF WORK

and rationality practiced by the overwhelmingly male research workers at high-tech companies in the Cambridge research triangle was possible only because of the embodied labor of the overwhelmingly female partners at home and female support staff at work. The culture of work was possible because of the culture of home.

Later works drew especially on Goffman’s dramaturgical metaphor, as well as Butler’s post-structural feminist version, to reveal cultures of work. Philip Crang’s (1994) study of a themed restaurant somewhere in southeast England disclosed that everyone, both workers and customers alike, adhered to a cultural script, dressing up, using the right prop at the right time, and saying their lines when prompted. Especially interesting, Crang observed, was when things went wrong, when people muffed their lines or didn’t follow stage directions, or when the props broke. Such moments were like when the curtain went up prematurely, revealing the backstage. It showed just how the frontstage presentation was pulled off.

A more politically pointed use of performance following Butler and focused on gender and the body was Linda McDowell’s study of merchant bankers in the City of London, Capital Culture (1997). Like other service workers, bankers must not only talk the talk but also walk the walk. They embody the product they sell: they are finance incarnate. McDowell was especially interested in how gender and sexuality were performed as the culture of work. It was done by clothing choice (leather skirt or demure suit); by language (English public school or the street market); and by acts and gestures (a head nod or a flapping arm). McDowell showed how narrowly constrained and vigorously policed such cultural performances were. Transgressive moments were necessarily subtle and fleeting: a grey suit that was just a little too grey, a hand gesture in the trading room that was just a little too enthusiastic. They might not be transgressions that would bring down financial capitalism but they could get you fired. Or they just might, following Butler, redefine an acceptable culture of work in the performance of gender and sexuality.

Another related example is the cultural performance of race at work. Beverley Mullings’s (2012) research about male and female sex-trade workers in the Caribbean showed that, just like gender, race is performed into existence. Sex-trade workers in her study used their bodies to perform that version of race expected of them by their often European (colonial) clients. Race, like gender, then, was pliable, taking on different forms. It was not natural but always a performance, always a response in this case to the larger culture of work in which it is situated.

There is one more form of performance around the culture of work, the performance of academic researchers as they try to change the culture of work that they are investigating. J.K. Gibson-Graham’s The End of Capitalism (as We Knew It) (1996) was pioneering. Gibson-Graham used their workshops, focus groups, and interviews with the female partners of Australian miners working in the Latrobe Valley, Victoria, to perform a different version of what a miner’s partner might be. They tried to create a space for feminist politics that altered not only the culture of home life, but, through the effect on the male partner, the culture of work life as well. Perhaps the end point of this trajectory is when academic research becomes literally a stage performance. Geraldine Pratt’s work on Filipina domestic workers in Vancouver, Canada, has become that. Pratt initially wrote it in standard academic prose but subsequently transformed her research material about Filipina domestic workers into a theatrical script for performance on stage (Families Apart, 2012). That production was a critical exploration of the culture of work of Filipina
nannies in Vancouver. Through influencing its audience, like Gibson-Graham’s project, it aspired to change the current culture of work, to perform a new world of work into being.

Conclusion

The culture of work is an overstuffed term. It seems as though it contains everything interesting that is left out of the standard economic account of the labor market, which is an enormous amount. The theories reviewed in this entry try to give the culture of work a more precise shape and definition. The political economy approach conceives of the culture of work as an addendum to class, which has the potential to be used strategically both to advance class interests (e.g., raising class consciousness) and also to thwart them (e.g., as cultural hegemony). The other approach, cultural and social theory, resists linking the culture of work only to class. Instead it seeks to explore the full spectrum of its elements – gender, race, sexuality, disability, age, and so on. Both approaches have their origins in part in the character of the work to which they have been applied. Crucial here, as suggested, was the profound shift during the 1970s and 1980s in the nature of work in the Global North as it moved from a manufacturing to a service-based economy. It triggered a change in the conception of the culture of work. However, the need to remember that the economy, including work, is always linked to culture is unchanging. Perhaps that is the most important role of all that the idea of the culture of work plays.

SEE ALSO: Economic geography; Emotional labor; Gender, work, and employment; Knowledge-based economy; Labor geography; Labor market; Race, work, and employment; Restructuring; Social reproduction

References

CULTURES OF WORK


Economies as self-correcting or self-reinforcing systems

At a very broad level of conceptualization, there have long been two main, and opposing, views of how economies function and develop. The first sees economies as essentially self-correcting systems that inherently and constantly tend toward equilibrium or balance. Theories based on this assumption or ontology regard equilibrium as the “natural” or normal state of affairs, and, if some disruption or event disturbs this state, various mechanisms are automatically set in motion – via the normal functioning of market processes – that act to restore equilibrium. The second view is quite different, and sees economies as systems in which market forces, rather than leading to equilibrium in a self-correcting way, tend instead to lead to disequilibrium and imbalance, and, moreover, operate to reinforce such imbalance. The former view is typically associated with conventional neoclassical and related schools of economic thought, the latter view with new growth theory and heterodox economics, and in particular with the theory of cumulative and circular causation.

These two distinctive worldviews carry equally distinct implications for how the economic landscape is perceived to develop. Under the neoclassical perspective, which assumes constant returns to scale, diminishing returns to labor and capital, the unhindered mobility of the latter, and ubiquitous knowledge, if a development gap opens up between regions or localities, as manifested by differences in per capita output, profitability, (real) incomes, and employment opportunities, then labor and capital movements are triggered such that labor will move to the higher-income, higher-productivity regions while capital will move in the opposite direction to take advantage of cheaper labor and the potential for major productivity advance. These movements act to bring regional per capita (real) incomes and productivity back in line, that is, to restore spatial “equilibrium.” Regions may differ in their particular industrial mixes, but the thrust of this view of regional development is that, over time, per capita (real) incomes should converge across space, or at least converge conditional on irreducible differences in consumer tastes and preferences, technology, or industrial structure. Empirical tests of this proposition for numerous countries have revealed, however, that in most cases the rate of regional convergence is very slow, that in some instances it is nonexistent, or that even in others divergence rather than convergence appears to be dominant.

Under the self-reinforcing or cumulative causation perspective, the fact that regional disparities in economic growth and prosperity may not be self-correcting and convergent but persist or even increase over time, is entirely consistent with how actual market economies function. Leading development theorists, such as Myrdal, Hirschman, and Nurske, were stressing this fact back in the 1950s. Thus Gunnar Myrdal, widely...
regarded as one of the key exponents of cumulative causation theory, argued that “the main idea I want to convey is that the play of the forces in the market normally tends to increase, rather than decrease, the inequalities between regions” (1957, 26). At about the same time, Albert Hirschman similarly claimed that “we may take it for granted that economic progress does not appear everywhere at the same time and that once it has appeared, powerful forces make for a spatial concentration of economic growth around the initial starting point” (1958, 183). For these theorists, uneven regional development is not some self-correcting, equilibrating process, but a self-reinforcing, cumulative one, although possibly subject to various spread effects that might limit the eventual degree of regional disparity.

It was not until the 1970s and early 1980s, however, that the idea of cumulative causation was developed further, primarily by Nicholas Kaldor (1981). Kaldor set the idea within a nonequilibrium theory of export-led manufacturing growth based on increasing returns, and, like Myrdal and Hirschman before him, used the notion to explain regional income disparities. Then, from the early 1990s onward, the notion began to attract wider attention, stimulated by a number of factors: a more general rediscovery by economists of the idea of increasing returns and its incorporation into different aspects of growth theory (Buchanan and Yong 1994), including new endogenous growth theory; the development of path dependence economics (David 1989; Arthur 1999; Setterfield 1997; 2010); and the emergence of the new economic geography (Krugman 1991, 1997), to name but three. Some have gone so far as to argue that the principle of circular and cumulative causation (CCC) provides the foundations of a new nonequilibrium economics (Berger 2009). And, for their part, economic geographers have embraced the concept, elevating it to a key principle for understanding (uneven) regional development:

any attempt to answer the questions posed above [concerning the economic development and specialization of regions] must formulate the problem by reference to a dynamic of cumulative causation whose logic is definable not in terms of some primum mobile or first cause, but in terms of its own historical momentum. (Scott 2006, 85)

All this is not to suggest that there is now a single, coherent, and generally agreed theory or model of CCC. Kaldor (1970; 1981) came close to a synthesis, which Setterfield (1997; 2002) has sought to extend and formalize; but a comprehensive CCC theory linking production, technological change, economic growth, and income distribution, let alone regional development, has yet to be advanced (Toner 1999). Furthermore, CCC theorists themselves often deploy the notion in different ways. Nevertheless, there is consensus about the basic principle of CCC, and about its key mechanisms.

**The principle and mechanisms of circular and cumulative causation**

It was Myrdal who first used the phrase “circular causation of a cumulative process.” As the phrase suggests, it refers to a process in which there is causative feedback between key aspects or features of a system, which feedback serves to reinforce the development of those same features in a cumulative manner. Thus, for example, an initial increase in some variable, X, induces changes in a second variable, Z, which result in a further increase in X, and so on. This interaction continues indefinitely, without leading the system toward a position of equilibrium. Kaldor’s model exemplifies this basic principle (Kaldor 1970; 1989). It posits a circular causal
process that links three basic mechanisms or determinants of regional (and indeed national) growth: trade-driven demand for a region’s exports; increasing returns and externalities; and technological change. To capture these links, the Kaldorian CCC model appeals to Verdoorn’s law, according to which the rate of growth of productivity depends on the rate of growth of output. The Verdoorn law is commonly understood as a dynamic analog of Adam Smith’s original dictum that represents the influence of output growth not just on the extent of specialization in the production process, but also on learning by doing, the propensity to engage in research and development, and firms’ willingness to invest in new physical capital that embodies technological improvements. Rapid growth induces dynamic increasing returns, via Verdoorn’s law, which enhance export competitiveness and hence export growth, so that the circular cumulative causation process takes the form:

\[ \text{High } X_t \rightarrow \text{High } Y_t \rightarrow \text{High } q_{t+1} \rightarrow \text{High } X_{t+1} \]

and so on, where \( X, Y, \) and \( q \) are the rates of growth of exports, output, and productivity respectively.

Given that regions, for the most part, are highly open economies, crucially dependent on the external demand for their products (or services), the model is highly relevant for understanding regional growth and development (Figure 1). The demand for a region’s products (or services) will depend on their price, and the price will depend on the productivity of the region’s industries providing those products (or services), and their prices relative to those for similar goods and services produced elsewhere (in world markets). The higher the productivity of a region’s industries, compared to the wages in those industries relative to those elsewhere, the more competitive the region’s industries will be and the higher will be the demand for its products (or services). The faster the growth of a region’s exports, the faster the growth of regional aggregate output (which includes multiplier effects on industries and services that are not exported but which supply the export sectors or are stimulated by the increases in incomes generated within the export sectors). With increases in output comes greater scope for exploiting increasing returns and externalities. Within the neoclassical system, increasing returns are generally examined at the level of the plant or firm, and via the concept of scale economies, that is, benefits to the plant or firm arising from expanding output, such as indivisibilities in factor inputs, gains from specialization, and the use of superior and more efficient techniques. While CCC theory recognizes these sources of increasing returns, it also puts particular stress on dynamic economies, especially those associated with technical “learning by doing.” Cumulative experience with a production process leads to technological change, which in its turn raises productivity. Growth itself thus stimulates technological change, which in its turn helps to stimulate productivity, and hence competitiveness, which makes a region’s products and services attractive in external markets; this then boosts exports demand and thus regional growth, and so on.

In emphasizing the role of increasing returns in his version of CCC theory, Kaldor (1989) referred explicitly to the importance of geographical agglomeration. Drawing on Alfred Marshall’s account of “industrial localization,” Kaldor argues that one of the most remarkable features of modern industrial development is the spatial concentration of economic activity, especially manufacturing. The spatial concentration of industrial activity is explained largely by the fact that agglomeration confers a variety of benefits or externalities on firms. These advantages go well beyond the economies of large-scale
CUMULATIVE CAUSATION, ENDOGENOUS GROWTH, AND REGIONAL DEVELOPMENT

Figure 1  The process of circular and cumulative regional growth: a Kaldorian-type schema.

operations, such as plant- and enterprise-scale economies. According to Kaldor, the advantages of geographical concentration arise from the local availability of skilled labor, joint production, and the possibilities for ready communication of trade, technical know-how, and management practices. Further, the close proximity associated with spatial agglomeration affords a ready identification of market opportunities for a firm’s products or related products, and for monitoring changing market trends and competitors’ behavior. In his version of CCC theory, geographical agglomeration reinforces the effect of Verdoorn’s law.

But, in addition to these externalities arising from agglomeration, an additional mechanism that is of critical importance for the operation of self-reinforcing cumulative development in a spatial (regional) context is the geographical mobility of capital and labor. As mentioned above, whereas in neoclassical growth theory, such interregional movements are argued to lead to convergent growth paths and to equalize regional incomes over time, in CCC theory they can have the opposite effect. Imagine a national economy composed, geographically, of two regions, north and south, and assume that, for some reason — possibly an accident of history, or some natural or environmental advantage — the...
south gains a lead in economic growth and begins to develop the increasing returns effects and positive externalities associated with the spatial concentration of activity. The prospect of better job opportunities, higher incomes, and greater profits in the south draws in finance capital (investment) and workers from the north (and possibly also from overseas – increasingly important in a globalized world). In effect, these inflows raise the full employment growth ceiling (FEGC) in the south above what it would otherwise be, preventing supply-side constraints from emerging, and thereby enabling further growth to take place, and a yet greater spatial concentration of activity in that region (see Figure 2). To the extent that it will be the more skilled and enterprising workers who move from the north to the south, bringing knowledge and technical expertise with them, this reinforces their positive impact on the south’s productivity, its competitiveness, and hence its growth rate. The outflow of capital and labor from the north has the opposite effect, constraining or even lowering its full employment growth ceiling (FEGC), and hence limiting or even reducing its growth rate, particularly if the workers who

Figure 2  Labor and capital mobility and cumulatively divergent regional growth (adapted from Holland 1976).
migrate from the north to the south are indeed among the more skilled sections of the north’s labor force (see Holland (1976) for a detailed exposition of this sort of model). In addition, the greater the agglomeration of firms in the south, the greater will be the knowledge and technological spillovers encouraged by close proximity, and the faster will growth be in the south compared to the north. In a spatial setting with relatively free movement of capital and labor, but localized technological spillovers, the process of circular and cumulative causation is thus essentially one of combined and uneven (divergent) regional development. As Harvey expresses it thus:

Small pre-existing geographical differences, be it in natural resources or socially constructed endowments, get magnified and consolidated rather than eroded by free market competition. The coercive laws of competition push capitalists to relocate production to more advantageous sites and the special requirements of particular forms of commodity production push capitalists into territorial specialisations … Agglomeration economies (including those achieved through urbanisation) generate a locational dynamic in which new production tends to be drawn to existing production locations … Circular and cumulative causation within the economy then ensures that capital rich regions tend to grow faster while poor regions grow [relatively] poorer. (2006, 98)

Three issues arise in connection with CCC theory. First, how does it relate to the recent rise of endogenous growth theory, given that the latter also allows for increasing returns, and thence for the possibility of divergent regional growth patterns? Second, given that in CCC theory regional growth is endogenous to its past history, how does it relate to path dependency theory, which itself has attracted considerable attention from economic geographers since the mid-2000s as a way of thinking about regional development? And, third, what are the limits to circular and cumulative causation? Since regional growth paths do not necessarily diverge indefinitely, what does this imply for the assumption of continuous self-reinforcing dynamics?

**Endogenous growth and cumulative causation**

The distinctive feature of CCC theory is that growth is viewed as endogenous to the economy, generated entirely from within by recursive causal mechanisms. As Kaldor stated:

it is impossible to assume the constancy of anything over time, such as the supply of labour or capital, the psychological preferences for commodities, the nature and number of commodities, or technical knowledge. All these things are in a continuous process of change but the forces that make for change are endogenous not exogenous to the system. The only truly exogenous factor is whatever exists at a given moment of time, as a heritage of the past. (1989, 61)

This is in stark contrast to standard neoclassical growth theory. The latter identifies the underlying rate of output growth as the sum of exogenous labor force and capital stock increase, and exogenous technical change. Indeed, there is a basic theoretical paradox in neoclassical theory, in that technological change is seen as of fundamental importance for economic growth but is not explained by the theory. Instead, it is simply assumed to occur autonomously outside the economic system, which is hardly a very realistic assumption since it ignores purposive research and development by firms. In CCC theory, technological change is stimulated by growth itself, through the increasing returns and learning effects that growth generates. Further, as numerous empirical analyses have shown, technological change (in the form of research...
and development activity, innovation, patenting, and the like) is not distributed evenly across geographic space but tends to concentrate in particular regions or urban centers, where it is fueled by an endogenous circular process by virtue of the externality and increasing returns effects generated by the dense locational proximity of innovation-oriented firms and workers.

However, recent developments in neoclassical growth theory have sought to endogenize technical change and give due recognition to the role of increasing returns in growth. In general terms there are three main variants of endogenous growth theory: those that attribute technical change (and thence growth) to investment in capital stock (which generates learning by doing); those that stress investment in human capital, via education and intentional research (which also increases spillovers of knowledge); and Schumpeterian versions that allow for purposive and profit-seeking improvements in technology by entrepreneurs (for a detailed overview of endogenous growth models and their implications for regional development, see Martin and Sunley 1998). Essentially, such models seek to endogenize technical change within the standard neoclassical production function model, and allow for increasing returns to capital and labor.

All three variants (and there are also several hybrid forms) can be applied to the analysis of regional growth and development (see Martin and Sunley 1998). Under certain assumptions, for example about the relative mobility of capital and labor, the in situ propensity for local workers to invest in education and skills, and the scope for and extent of local knowledge spillovers, endogenous growth models predict varying degrees of self-reinforcing regional divergence, akin to CCC models (see Skott and Auerbach 1995). Which outcome is predicted depends much on the extent, if any, of interregional spillovers of technology, which work to disperse growth and hence to prevent the continuing accumulation of activity and growth in certain regions at the expense of others.

One particular type of model of the economic landscape that draws on endogenous growth ideas (especially learning by doing, induced technological progress and knowledge spillovers), and which incorporates a version of increasing returns and cumulative causation type mechanisms, is the new economic geography (NEG). The new economic geography literature that emerged in the 1990s sought to formalize certain cumulative causation mechanisms that account for the uneven geographical distribution of economic activity. Depending on assumptions concerning transport costs, initial economic structures, and labor mobility, regions are shown to fall into rich (core) and poor (peripheral) categories as production concentrates where the market is larger. In Krugman’s (1991) seminal contribution – the basis of much of the subsequent development of NEG – the tendency for firms and workers to concentrate in one or more core regions is driven by the interaction of labor migration and transport costs, in the presence of increasing returns effects and external economies. Larger markets attract more firms, which in turn attract more workers. In more recent versions of the model, the agglomeration of industry occurs via cumulative processes triggered by input–output linkages between intermediate (upstream) producers and final goods (downstream) firms. With increasing returns to scale, intermediate producers will have an incentive to concentrate where there is a large final goods industry, in order to produce at a more efficient scale and to take advantage of the possibilities of inventing new combinations. This in turn will make it attractive to final goods producers to locate where there is a large intermediate industry, as the cost of intermediate goods will be cheaper there. The
process is thus a circular one, and may result in large and self-perpetuating uneven geographical configurations. In one extreme variant of the NEG model, all activity will concentrate in just one region (with a “catastrophic” emptying of other regions). In still other versions, the specific spatial equilibrium outcome depends on the strength of dispersion forces, such as market-crowding effects which encourage firms to locate in less highly agglomerated regions.

In his early exposition of NEG, Krugman (1991) briefly acknowledged the influence of Kaldor’s work on the relationship between increasing returns, cumulative causation, and regional agglomeration. And it can be argued that, in certain respects, Kaldor’s work may well have been a prologue to the NEG literature (see Bhattacharjia 2010), as captured in his statement that

The spatial aspects of competition under conditions of cumulative causation constitutes a field that has not been explored yet, but which may call for radical changes in the prevailing views concerning the effects of freedom of trade between different countries or regions. (Kaldor 1989, 34)

However, Kaldor’s dismissal of the relevance of equilibrium – which is a core feature of NEG models – no doubt explains the subsequent lack of reference to his work, and indeed to CCC theory more generally, by NEG theorists. In fact, the process of circular and cumulative causation raises key questions about whether actual regional economies can ever be in a state of equilibrium.

Cumulative causation and path-dependent regional development

With increasing returns, the productive possibilities in an economy in any given period depend crucially on the level of production achieved in the past, since, in the presence of increasing returns, technical possibilities for production today are endogenous to those past levels of production. Cumulative causation is therefore a historical or path-dependent, rather than a conventional equilibrium, process. Again, Kaldor makes this point:

Once however, we allow for increasing returns, the forces making for continuous change are endogenous … And the actual state of the economy during any one period cannot be predicted except as a result of the sequence of events in previous periods which led up to it. (Kaldor 1970, 1244)

The basic idea behind the notion of path dependence is that, once a particular path or pattern of development is initiated – be it a particular technology, industry, or geographical distribution of an industry – various processes emerge that tend to reinforce that path or pattern. Although there are different models of path dependence economics (see David 1989; Arthur 1999; Setterfield 1997; 2002; Martin and Sunley 2006), most invoke some form of increasing returns effects as the mechanism by which a self-reinforcing path or pattern of development assumes momentum.

In Arthur’s (1999) work, for example, it is assumed that the autocatalytic or self-reinforcing mechanisms that generate economic path dependence have a spatial dimension in that, choosing where to locate, firms are attracted by the presence of other firms in a region. In the spin-off version of this model, the path-dependent geographical distribution of industry occurs through a process of local firm spin-offs from existing parent firms. Arthur suggests that this type of cumulative, self-reinforcing process has characterized the geographical concentration of the US electronics and car industries. In the agglomeration economies version, if one region by chance gets off to a good start, its
attractiveness and the probability that it will be chosen by firms will be enhanced; further firms may then choose this region, with the result that it becomes yet more attractive because of the emergence of various agglomeration economies and externalities, and the concentration of firms (and perhaps also particular industries) there becomes self-reinforcing. Arthur refers to three such agglomeration economies or externalities: dynamic learning effects (learning by interaction with other proximate firms – knowledge spillovers); coordination effects (various local network effects, such as local supply chains, specialist intermediaries, and the like); and self-reinforcing expectations (when the increased prevalence of other similar firms doing similar things or making similar products enhances belief and confidence in a local developmental path).

Arthur (1999) suggests that Silicon Valley in the United States is a possible example of this sort of cumulative, self-reinforcing spatial agglomeration of a given industry, in this instance electronics. Certain key people – entrepreneurs like the Hewletts, the Varians, the Shockleys of the industry – happened to set up near Stanford University in the 1940s and 1950s, and the local labor market expertise and interfirm market they helped to create in Santa Clara County made subsequent location extremely advantageous for the thousand or so other firms that followed them. Arthur cites this as a classic example of (almost) unbounded agglomeration economies, which have fueled a cumulative, self-reinforcing, and path-dependent process of clustering. Another example is the high-tech cluster in Cambridge. The origins of this cluster are usually traced back to the establishment of a research consultancy and transfer enterprise (Cambridge Consultants) in 1969, followed soon after (at the beginning of the 1970s) by the construction of a science park just outside the city by Trinity College. These events set off a cumulative, path-dependent process whereby new high-tech firms have been created or attracted into the cluster, which in their turn have promoted the growth of a highly qualified scientific workforce. And, as the cluster has grown, this has stimulated research and innovation, increasingly closer links with scientific research within the university, the emergence of a local circuit of venture capital and business angel finance, and a general climate of entrepreneurship and enterprise formation, all of which have maintained the cluster’s growth momentum. In many respects, then, clusters display the sort of self-reinforcing, path-dependent development that typifies the process of circular and cumulative causation.

The limits to cumulative causation

But, in its basic form, there is an obvious problem or caveat in the cumulative causation model of regional development. Essentially it suggests that success breeds success and failure begets failure indefinitely. The recursive, self-reinforcing nature of cumulative causation means that once a particular pattern and path of regional development is initiated, it simply continues and cumulates as time progresses. It appears, therefore, that an initially faster-growing region will permanently experience faster growth, while its less successful partners will endure slower growth: this is the impression given in Figure 2, for example. Put another way, once the initial relative success or failure of a region is known, then, given the path-dependent nature of cumulative causation, so too is its subsequent history. By the same token, the basic CCC model predicts progressive and unabated regional divergence in per capita incomes.

Yet, from the stylized facts of regional growth, it is known that disparities in regional per capita incomes do not widen indefinitely. Some classic
CUMULATIVE CAUSATION, ENDOGENOUS GROWTH, AND REGIONAL DEVELOPMENT

studies have argued that, as a country develops, so regional disparities widen for a while as economic development agglomerates in certain core regions, but that with time they begin to narrow, as peripheral regions catch up. Both Myrdal and Kaldor recognized this possibility. Myrdal, for example, argued that various “spread” and “backwash” effects – such as induced demand from the core for the products of the periphery, or dispersion of economic activity from the core because of rising costs there – could limit the extent of uneven regional growth caused by cumulative causation-propelled spatial agglomeration. More recently, Leon-Ledesma (2002) has suggested that the diffusion of technology from core growth regions to lagging peripheral ones may allow the latter to catch up. In other words, with time “spread” or “centrifugal” effects can come into effect to limit and counter the agglomerating or “centripetal” pull of the prosperous core region. That is to say, the balance between agglomerating (centripetal) and spreading effects may shift over time in favor of the latter, thereby allowing peripheral or lagging regions to improve their relative growth performance and then to embark on a virtuous circular and cumulative developmental path of their own.

But the stylized facts also indicate that, notwithstanding the possibility of such spread effects, successful regions can lose their inherent relative growth advantage, and subsequently experience periods – often prolonged – of slow growth or relative depression. There are essentially two ways of conceptualizing such eventualities – by invoking an external shock that interrupts the cumulative causation process from without; or by recognizing that various processes or mechanisms may emerge endogenously from within the process of cumulative causation itself. As an example of the former, consider the key role that external (export-led) demand plays in the development and growth of regional and urban economies. In any round of the cumulative causation process, the dynamics of a virtuous regional growth circle depend on sufficient export demand for the region’s products or services being forthcoming, induced by the rise in efficiency brought about by the realization of scale economies and by positive externalities associated with the increased spatial agglomeration of activity in the region in the previous round. But what if the rate of growth of exports is subject to exogenous shocks over time? A major new external competitor may appear, offering far more superior forms of the products or services in question, or at much cheaper prices. A major reduction in the export demand for the region may then occur, and this in turn will disrupt the cumulative growth process, possibly causing it to break down. Economic history is replete with examples of countries, regions, and cities that were once national, even international, leaders in their respective specialisms, but then lost that leadership to the rise of other, more competitive, nations and regions elsewhere, and with it their economic dynamism. For example, since the late 1970s, the rise of cheap labor producers in the BRIC countries, especially China, has undermined the export growth of much of the Western mass production industry. The positive export-led cumulative causation dynamics previously enjoyed by many Western cities and regions has in effect been transferred to certain cities and regions in the BRIC nations, which have enjoyed extraordinary growth rates as a consequence.

Faced with an exogenous shock of this sort, a region’s firms have three main possible responses: to struggle on with declining market share and falling profits, and eventually go under; to restructure and reorganize production via technological and product upgrading (a technological fix), in order to resume competitiveness and hence export demand for their products;
or to shift some or all of their production to cheaper or more dynamic locations elsewhere (a spatial fix).

But a region’s cumulative causation dynamic may break down and be lost as a result not of an external shock but of processes and features that emerge from within the cumulative process itself. The very process of cumulative spatially agglomerated sectoral specialization, though making for cumulative and self-reinforcing growth, or positive path-dependent lock-in, for a certain period, may itself eventually give rise to structural, technological, and institutional rigidities and fixities that then usher in a period of negative path-dependent lock-in which impairs the ability of the region’s economy to realize transformational growth (Setterfield 1997; Thirlwall 2013). The increasing rigidity of the regional economy may occur for various reasons, depending on the structure and organization of that economy. According to the extent to which new types of fixed capital must be accumulated in the pursuit of economies of scale, the ability of a region to realize dynamic increasing returns may be impaired if it suffers from technological interrelatedness, both within local firms and, crucially, across local firms. Because this phenomenon makes technological change a nonmarginal process, it can become highly costly or subject to coordination failures, and this may result in the negative lock-in of the regional industry in question to a previously existing technological regime. The ability of the region to transform its stock of productive assets in a manner in keeping with the realization of increasing returns and with maintaining its export competitiveness may suffer accordingly.

For example, in a regional economy based on an industry with a detailed interfirm horizontal division of labor, with different firms specializing in different stages or aspects of production of a finished export good, the high degree of interrelatedness across firms may come to hinder the introduction of major technological improvements to production. This is particularly likely to occur if the industry in question involves large sunk costs in expensive dedicated machinery or plant. In such a regional production complex, no one firm will invest in new techniques of production unless other firms do so, with the result that none may be willing to risk making the first move. As a consequence, technological inertia may develop and eventually hold back productivity advance. The regional economy then becomes especially vulnerable to the rise of competitors that have switched to the more advanced and more efficient techniques.

A prime example of this situation is the Lancashire cotton textile sector in the United Kingdom. From the late eighteenth century onward, this region enjoyed strong self-reinforcing cumulative, path-dependent, spatially agglomerated growth, involving thousands of local firms linked in a complex interfirm horizontal division of labor in cotton textile production. By the mid-nineteenth century Lancashire had become the world’s largest exporter of cotton textiles, accounting for two-thirds of the world’s exports, and employing over 600,000 workers locally. But, toward the end of the nineteenth century and into the early decades of the twentieth century, the region came under increasing international competition from overseas competitors that had adopted new technologies, including upright automatic looms, while Lancashire’s firms still clung on to the old horizontal looms on which its original productive superiority had been founded. The high degree of interfirm division of labor and technological interrelatedness between Lancashire’s firms, together with an institutionalized self-confidence (some would say overconfidence) born of past success, became sources of negative technological lock-in and
entrepreneurial inertia, which then slowed down the region’s technological upgrading, with devastating consequences. Between the two world wars, several thousand of the region’s textile mills closed down, and the industry never recovered.

More generally, there is evidence that the economic benefits that can accrue through the self-reinforcing spatial agglomeration of industry may themselves follow a sort of life cycle over time, initially rising as the cumulative process gets underway, then reaching a maximum as the regional industry matures, and then declining thereafter (Potter and Watts (2011), find this, for example, in the case of manufacturing industries in the United Kingdom’s Birmingham region). That is to say, agglomeration economies may not increase continuously over time, or even remain unchanged once established, but rise and fall as the industries around which they developed rise and fall. The reality is that regional industrial structures change and evolve over time, and the processes of cumulative causation can be expected to ebb and flow as a consequence. Theorists have as yet given insufficient attention to the endogenous evolutionary dynamics of cumulative causation, yet this would be necessary if the model is to capture the long-run evolutionary paths of regional economies. The issue is akin to the way that the notion of path dependence is typically used in economics, regional studies, and institutional studies, whereby a given path and pattern of economic, technological, and institutional development, once established, is assumed to get locked in and to change only when subjected to an external shock of some kind. In the real world, economies, technologies, and institutions also display differing degrees of ongoing endogenously generated change and adaptation. Both path dependence theory and the CCC model need to incorporate and account for such change and adaptation (Martin 2010).

SEE ALSO: Population mobility and regional development; Regional development models; Regional inequalities; Regional unemployment and regional labor markets; Trade and regional development; Uneven regional development

References


CUMULATIVE CAUSATION, ENDOGENOUS GROWTH, AND REGIONAL DEVELOPMENT


CyberGIS – defined as GIS (geographic information science and systems) that is based on the development and use of advanced cyberinfrastructure – represents a new modality of GIS in the era of big data; it has emerged during the past several years as a vibrant interdisciplinary field (Wang 2010; Wang, Wilkins-Diehr, and Nyerges 2012; Wang et al. 2013; Wright and Wang 2011). Cyberinfrastructure (sometimes referred to as e-infrastructure and e-science) typically integrates digital environments to support collaborative research, as well as computing- and data-intensive scientific problem-solving. Specifically, cyberinfrastructure consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high-performance networks to improve research productivity and enable scientific breakthroughs that are not otherwise possible (Wang and Zhu 2008). An initial landmark in the history of cyberinfrastructure is the blue-ribbon panel report titled Revolutionizing Science and Engineering through Cyberinfrastructure that was published by the US National Science Foundation (NSF) (Atkins et al. 2003).

Major advances in cyberGIS as an interdisciplinary field are rooted in multiple research domains including advanced cyberinfrastructure, complexity science, computational and data sciences, distributed computing and systems, GIS, high-performance and parallel computing, spatial analysis and modeling, and a number of domains concerning geospatial questions and problems. Early research on geo-middleware set an important foundation in cyberGIS for integrating high-performance and distributed computing resources and services to enable collaborative and computationally intensive geographic analysis (Wang 2004).

A variety of research related to cyberGIS had been pursued before 2010 (Wang, Wilkins-Diehr, and Nyerges 2012; Yang et al. 2010). For example, Wang and Liu (2009) demonstrated a suite of technical capabilities for establishing an online science gateway to bridge cyberinfrastructure and GIS functions based on a service-oriented architecture and high-performance and distributed computing. The first holistic understanding of cyberGIS, however, was gained by the development of an integration framework for synthesizing cyberinfrastructure, GIS, and spatial analysis and modeling to solve computationally intensive and collaborative geographic problems (Wang 2010). The case study used to demonstrate the application of this cyberGIS framework focused on revealing spatial patterns of the impact of climate change on crop yields in the United States. The computation, data, and collaboration requirements of the study could not be met by conventional GIS approaches based on a monolithic architecture, sequential computing, and single user interfaces.

Several cyberinfrastructure, geospatial, and GIS research communities have been working together to advance the field of cyberGIS. A key community event was the NSF TeraGrid...
Workshop on CyberGIS that took place in conjunction with the University Consortium for Geographic Information Science Winter Meeting in February 2010. The workshop report laid out a compelling cyberGIS roadmap that underpins fundamental issues of cyberGIS for enhancing cyberinfrastructure while contributing to the advancement of the next generation GIS, which will integrate high-performance, distributed, and collaborative capabilities to support geospatial discovery and innovation (Wang, Wilkins-Diehr, and Nyerges 2012). Built on the success of the workshop, a multi-institution and multidisciplinary proposal to NSF was funded in 2010 to begin a five-year $4.43 million project on “CyberGIS Software Integration for Sustained Geospatial Innovation” with the following six major goals (Wang et al. 2013):

1. engage multidisciplinary communities through a participatory approach to evolving cyberGIS software requirements;
2. integrate and sustain a core set of composable, interoperable, manageable, and reusable cyberGIS software elements based on community-driven and open-source strategies;
3. empower high-performance and scalable cyberGIS by exploiting spatial characteristics of data and analytical operations for achieving unprecedented capabilities for geospatial scientific discoveries;
4. enhance an online geospatial problem-solving environment to allow for the contribution, sharing, and learning of cyberGIS software by numerous users, which will foster the development of cross-cutting education, outreach, and training programs with significant broad impacts;
5. deploy and test cyberGIS software by linking with national and international cyberinfrastructure to achieve scalability to significant sizes of geospatial problems, amounts of cyberinfrastructure resources, and number of users;
6. evaluate and improve the cyberGIS framework through domain-science applications and vibrant partnerships to gain better understanding of the complexity of coupled natural and human systems.

This major NSF cyberGIS initiative has established three interrelated pillars of a cutting-edge cyberGIS software environment, including the CyberGIS Gateway, CyberGIS Toolkit, and GISolve middleware.

The CyberGIS Gateway has evolved from the TeraGrid GIScience Gateway (Wang and Liu 2009) and provides an online environment for making cyberGIS capabilities accessible to a large number of users for various research and educational purposes (Wang et al. 2013). The CyberGIS Gateway serves as a platform for a suite of scalable spatial analysis and modeling applications built on GISolve middleware. These applications are broad and diverse, including, for example, assessment of climate change impacts, hydrological information analysis, social media analytics, and spatial econometrics, which are supported by the modular architecture of the CyberGIS Gateway. The computation, analysis, and modeling tasks are conducted using multiple cloud computing environments along with several leading cyberinfrastructure environments, such as the NSF extreme science and engineering discovery environment (XSEDE) and the Open Science Grid. The user interfaces of the CyberGIS Gateway are designed to accommodate specific application characteristics in a user-centric way, so that a large number of users who do not possess in-depth cyberGIS technical knowledge can access complex cyberinfrastructure and cyberGIS workflows.

As a result, the CyberGIS Gateway is able to serve as a learning environment and is also
accessible to the general public. Numerous undergraduate and graduate students have used the CyberGIS Gateway to learn advanced cyberinfrastructure, service-oriented architecture, GIS, high-performance computing, spatial analysis and modeling, and computing- and data-intensive geospatial problem-solving at universities in the United States and beyond.

The CyberGIS Toolkit is composed of loosely coupled geospatial software components, each of which can take advantage of advanced cyberinfrastructure resources (e.g., those provided by XSEDE) in a scalable way. The CyberGIS Toolkit is open source and enables computational and data scientists in various geospatial domains to directly contribute to and benefit from community-driven scalable software and tools. The approach used to build the CyberGIS Toolkit develops and leverages scalable computing strategies that are required to solve computing- and data-intensive geospatial problems, and also exploits high-end cyberinfrastructure resources, such as supercomputers integrated by XSEDE and high-throughput computing resources on the Open Science Grid. A rigorous process of software engineering and computational intensity analysis (Wang and Armstrong 2009) is applied when a software component is incorporated into the CyberGIS Toolkit, including software building, testing, packaging, performance and scalability analysis, and deployment. A selected set of software component examples of the CyberGIS Toolkit include the following.

- Parallel agent-based modeling (PABM) software is for multiscale and spatially explicit agent-based modeling (ABM) of coupled natural and human systems.
- Parallel genetic algorithm (PGA) is used to address computationally intensive spatial optimization problems; it can efficiently scale up to hundreds of thousands of processor cores available on multiple cutting-edge supercomputers.
- Parallel PySAL provides a set of scalable PySAL functions. PySAL is an open source library of spatial analysis functions written in Python. Parallel PySAL uses a multiprocessing Python library.
- pRasterBlaster is a high-performance map reprojection software originally developed by the Center of Excellence for Geospatial Information Science within the US Geological Survey (USGS).
- Terrain analysis using digital elevation models (TauDEM), created by Dr David G. Tarboton at the Utah State University, is enhanced to include a suite of high-performance digital elevation model (DEM) tools for watershed delineation and the extraction and analysis of hydrologic information from topography as represented by DEMs.

GISolve middleware bridges the CyberGIS Gateway and Toolkit by managing the complexity of interactions among heterogeneous and distributed resources and services of advanced cyberinfrastructure (e.g., XSEDE and various cloud computing environments) and providing standard interfaces for the integration of scientific applications into the cyberGIS software environment. GISolve represents the leading middleware approach to integrating advanced computational and data infrastructure with geospatial data and computing capabilities for computationally intensive and collaborative geospatial problem-solving, and is openly accessible via a suite of open service application programming interfaces (APIs) (Wang et al. 2013). These APIs are compatible with community standards such as those facilitated by the Open Geospatial Consortium, and adopt widely used web service protocols such as the representational state transfer (REST) architectural style. Consequently, GISolve supports
interoperability among different cyberGIS software environments. For example, it has been demonstrated that ArcGIS Online and the CyberGIS Gateway can interoperate for applications to take advantage of both environments while accessing advanced cyberinfrastructure based on GISolve APIs.

Previous work shows that scientific problems exhibiting high spatial complexity require holistic approaches to integrating rich spatiotemporal data, analytics, and models to form novel problem-solving approaches enabled by cyberGIS (cf. Wang 2010; Tang et al. 2011). Conventional scientific approaches to solving such complex geospatial problems, however, tend to be fragmented in space and time and are constrained by an inability to take advantage of complex, diverse, and massive geospatial data, which makes extrapolation across large and multiple spatial and temporal scales difficult or infeasible. Researchers from multiple institutions are engaged in research and development activities on the three pillars of the cyberGIS software environment. These investigations span various domains (e.g., bioenergy, emergency management, geography and spatial sciences, and geosciences) as illustrated by the examples given in the next sections.

**Complex geographic problem-solving**

Multiscale simulation of geographic dynamics often uses spatially explicit ABM approaches to study spatiotemporal complexity of coupled natural and human systems (e.g., multiscale disease spread). However, significant computational and data challenges must be resolved when dealing with massive multiscale agent–agent and agent–environment interactions, efficient exploration of alternative individual-level scenarios, and rigorous sensitivity and uncertainty analyses. In a simulation study of disease spread involving billions of individual-level agents, operating across multiple spatial and temporal scales, Shook, Wang, and Tang (2013) identified that variations in spatial and temporal scales had a strong influence in the representation of transmission processes for disease spread dynamics. Such multiscale geospatial simulations often require high-resolution spatiotemporal datasets during and after computation, while producing massive geospatial datasets for representing and examining spatiotemporal dynamics. This type of complex geographic problem-solving is difficult or impossible to pursue without cyberGIS.

Not only does cyberGIS provide a means to resolve computational and data challenges in solving such complex geographic problems, but it also represents a new paradigm for conceptualizing geographic processes based on parallel and distributed computing. Real-world geographic processes take place concurrently and in parallel with interactions across various spatiotemporal scales. But due to the dominance of sequential computing approaches in research and development of GIS, geographic process modeling has been limited by missing theoretical and methodological underpinnings for systematically representing concurrency, parallelism, and related spatiotemporal interactions. CyberGIS is built on parallel and distributed computing; thus, it is well suited to nurture such underpinnings for transforming computational and data-driven research on complex geographic problems.

**Data-intensive knowledge discovery**

Geospatial data have become ubiquitous and are increasingly important in numerous research fields (e.g., ecology, environmental science and engineering, geography and spatial sciences,
geosciences, and social sciences, to name just a few). Such data are instrumental to solving many critical scientific and societal problems (e.g., disaster management and environmental sustainability). With the increasing availability and capability of location-based sensors and devices, the explosive growth in volume and complexity of geospatial data will continue to accelerate in the foreseeable future. For example, it is estimated that the 3-D elevation program (3DEP) of the USGS for acquiring 3-D elevation data using LiDAR will result in data volumes that are of the magnitude of 7–9 petabytes, considering only the point cloud, intensity signals, and bare-earth elevation model for covering the United States. Conventional GIS approaches are ill suited to resolving such big data challenges. The realization of the full potential and benefits of 3DEP, estimated at $1.2–13 billion in new benefits annually by the national enhanced elevation assessment (NEEA) study (Snyder 2012) is dependent on the ability of cyberGIS for management, processing, visualization, and analysis of such massive geospatial data.

Extracting features from such elevation data promises to create new valuable data products (e.g., high-resolution stream network and drainage area) for integration with the USGS national map, which is a collaborative effort among USGS and other federal, state, and local partners to improve and deliver topographic information for the United States. Fine-resolution DEM data have a significant impact on the accuracy and reliability of hydrological modeling. Though tools exist for deriving topographic information from DEM data for hydrological modeling, when such tools are applied to fine-resolution DEM data, the resulting computational burden makes GIS approaches based on sequential computing unfeasible. TauDEM integrated within the cyberGIS software environment promises to resolve this data-intensive challenge.

Through rigorous computational-intensity analysis, the ability of TauDEM to take advantage of advanced cyberinfrastructure resources for data-intensive terrain data analysis is eliminating barriers to desirable computational performance. The enhanced TauDEM software can scale to thousands of processor cores and is able to analyze tens of gigabytes of DEM data within a single analysis, which was not possible on small-scale high-performance computing resources and lies far beyond the capabilities of GIS tools based on sequential computing strategies. With large data volumes of tens to hundreds of terabytes for a single basin, cyberGIS-based scalable data analytics is arguably the only viable approach that will allow effective analysis of 3DEP data to yield valuable geospatial information and enable its sharing with user communities. Meanwhile, tackling the significant computational and data challenges as posed by 3DEP will require development of a rich array of new cyberGIS capabilities and software that will, in turn, drive the frontiers of data-intensive geographic knowledge discovery.

Emergency management

Social media data (e.g., from Twitter) and volunteered geographic information are becoming increasingly useful in emergency management due to widespread use of location-aware mobile devices and near-real-time emergency status updates. Leveraging the cyberGIS software environment, in 2012 a collaborative project called Global Twitter Heartbeat analyzed a real-time stream of 10% of Twitter’s several hundred million daily tweets as they were posted. The entire process from data ingest to analysis to visualization was optimized to be completed
within a few seconds. This project demonstrated the tremendous power of the cyberGIS software environment for resolving the computational and data challenges of processing big data in an efficient and scalable manner. While no other existing tools can simultaneously manage, process, and analyze such massive, semi structured or unstructured, and streaming data in this near-real-time fashion, such cyberGIS capabilities are crucial for emergency management purposes.

Specifically, the project analyzed every tweet to determine its location and sentiment tone values, and then visualized related information in a heat map infographic that combines and displays tweet location, intensity, and tone. The project examined Hurricane Sandy, on which millions of tweets were posted, making cyberGIS a significant platform for communicating what was happening at every moment across a number of spatial scales. The live demonstration of this research outcome was debuted at the 2012 International Conference for High Performance Computing, Networking, Storage and Analysis (SC12) in Salt Lake City, UT.

Geodesign and spatial decision-making

Bioenergy is an important source of renewable energy. Optimal and sustainable management of bioenergy infrastructure, such as its supply chain, is a complex geodesign problem, which requires the collaboration and participation of different types of decision-makers who are usually not experts on cyberinfrastructure, GIS, or spatial analysis and modeling. Normally, a biomass supply chain optimization model is computationally intensive when applied to realistic scenarios across multiple spatiotemporal scales; this computational complexity makes it infeasible for decision-makers to directly interact with pertinent GIS analytics during the duration of a typical meeting. Another challenge for such decision-making processes is related to the data used as input and output for optimization modeling, and the integration of diverse geospatial data types (e.g., crop, land use, soil, transportation, and water).

A computationally intensive optimization model of a biomass supply chain – BioScope – has been integrated into the CyberGIS Gateway to demonstrate that cyberGIS can enable the computation of the model by accessing powerful cyberinfrastructure resources while providing an interactive and simple user interface for decision-makers who are not required to possess any technical knowledge of cyberGIS (Figure 1) (Hu et al. 2015). Decision-making experiments accessible to a large number of users can be conducted at the spatial scale of the contiguous United States, which is difficult or impossible to achieve without cyberGIS.

The four examples (i.e., complex geographic problem-solving, data-intensive knowledge discovery, emergency management, and geodesign and spatial decision-making) described above are not intended to form a complete taxonomy of cyberGIS-enabled research and science. The overarching emphasis is rather placed on demonstrating the role of cyberGIS in supporting important research breakthroughs.

Computational theories and methods

CyberGIS has been advanced through the synergy of critical computational and spatial thinking (Wang and Zhu 2008). For example, the theoretical construct of a spatial computational domain was established to guide the development of generic methods and efficient algorithms for multiscale spatial modeling (Shook, Wang, and Tang 2013) and data-intensive geospatial
Figure 1  CyberGIS-BioScope application user interface.

analytics (Wang 2010; Wang et al. 2013). The formalization of the spatial computational domain addresses the following important questions: How can massive shared cyberinfrastructure resources be harnessed to solve computationally intensive geographic analysis problems? What is the nature of computational intensity of geographic analysis? Why is a spatial perspective needed to address the first two questions? Such theoretical work also involves methodological and empirical investigations that reveal how the spatial characteristics of data and analytical operations affect computational requirements for cyberGIS-based geographic analysis (Wang, Wilkins-Diehr, and Nyerges 2012).

Due to the complexity of evolving cyberinfrastructure and the diversity of geographic analytical approaches, significant scientific and technological challenges must be tackled before generic computational methods can be established for cyberGIS-based geographic problem solving (Wang 2013). Building on the theoretical foundation of the spatial computational domain, a number of computational methods have been developed for solving complex geographic problems such as spatially explicit agent-based
models that simulate the movement of massive numbers of individual agents through space and time (Shook, Wang, and Tang 2013), and data-intensive viewshed analysis (Wang et al. 2013). These methodological advances enable efficient and shared use of massive cyberinfrastructure resources for scalable analysis of geospatial big data and solving computationally intensive geographic problems (Tang et al. 2011; Wang 2013).

A major empirical element of cyberGIS centers on an open and functional process for understanding the nature of computational challenges in various geographic problems and for overcoming these challenges to empower desirable research productivity and progress. This process facilitates the development of scalable computational methods and associated algorithms that are encapsulated into geospatial middleware that manages the complexity of cyberinfrastructure. Previous research on geospatial middleware has provided interoperable geographic analysis services for empowering a large number of users and diverse applications (Wang et al. 2013).

Digital geospatial ecosystems

CyberGIS fulfills an essential role in enabling computing- and data-intensive geospatial research and education across a broad swath of academic disciplines with significant societal impacts. The future realization of cyberGIS’s reach is compelling as the role and complexity of big data continue to challenge the state of the art of collaborative, interactive, and scalable knowledge discovery through the processing and visualizing of complex and massive amounts of geospatial data and performing associated analysis, simulation, and visualization.

CyberGIS-based scientific approaches have already made a significant impact in several research domains (e.g., hydrology and water resources, complex coupled natural and human systems, emergency management, and spatial econometrics). CyberGIS based on critical computational and spatial thinking is needed to combine rich and complex geospatial data, analysis, and models, and is expected to support the effective and timely resolution of many grand geospatial challenges (e.g., energy and environmental sustainability, emergency and disaster management, and health and wellness).

CyberGIS exemplifies and enhances the vision of integrated cyberinfrastructure and, thus, will continue to positively contribute to broad scientific advances. Integrative approaches are required to meet the increasing needs for combining diverse and massive digital resources and services across individuals and organizations, to serve various research and education purposes. To navigate through the extensive digital transformation across science and society, cyberGIS has an important role to play, as it bridges the divide between broad digital and geospatial realms by nurturing sustainable digital geospatial ecosystems. As argued by Wang (2013), the ecosystem metaphor for characterizing cyberGIS emphasizes the notion of systems of systems that are developed and evolved in heterogeneous computing, communication, information, and human environments.

As a recent example of geospatial innovation for cyberGIS hardware and software, in 2014 the CyberGIS Center for Advanced Digital and Spatial Studies at the University of Illinois at Urbana-Champaign received an NSF major research instrumentation grant to establish a national cyberGIS facility. This facility is based on a single tightly integrated instrument – a high-performance computing system optimized to deal with geospatial data and computation, that is equipped with:
• approximately seven petabytes of raw disk storage with high input/output bandwidth;
• solid-state drives for applications demanding high data-access performance;
• advanced graphics processing units for exploiting massive parallelism in geospatial computing;
• interactive visualization supported with a high-speed network and dynamically provisioned cloud computing resources.

CyberGIS software and tools integrate these system components to support a large number of users who are investigating scientific problems in areas as diverse as biosciences, engineering, geosciences, and social sciences.

The design of the cyberGIS facility is driven by requirements posed by these diverse scientific investigations that are critically dependent on scalable spatial data infrastructure and cyberGIS workflows involving interactive steering of simulations, and processing and visualization of large, complex, and heterogeneous geospatial datasets.

The potential of cyberGIS to contribute to broad scientific and technological advances continues to be empowered by rapid developments in digital technology in general, and advanced cyberinfrastructure in particular. It is important to recognize that cyberGIS exhibits significant complexity, as it represents digital geospatial ecosystems embedded within massive evolving systems, dynamic connections, and complex relationships. Management of such complexity while sustaining innovation is a grand challenge for cyberGIS as an interdisciplinary field.

SEE ALSO: Agent-based modeling; Big data; Cloud computing; Digital divide; Geocomputation; Mapping cyberspace; Parallel computing; Representation: dynamic complex systems; Spatial analysis; Technology

References


Czech Republic: Česká geografická společnost (ČGS) (Czech Geographical Society)

Founded: 1894
Location of headquarters: Prague
Website: www.geography.cz
Membership: 485 (as of January 1, 2014)
President: Tadeusz Siwek
Contact: tadeusz.siwek@osu.cz

Description and purpose

The main aim of the Czech Geographical Society is to promote geographical research and geography as a scientific discipline. It represents Czech geographers in their contacts with Czech political authorities and with other scientific organizations in the Czech Republic and abroad. It is a base for cooperation among all Czech geographers.

Journals or major publication series

Informace ČGS. http://geography.cz/informace-cgs/

Current activities or projects


Brief history

Czech geography, as a science, was born at the end of the nineteenth century in the Austro-Hungarian Empire when the Germanized old Czech university was divided into a German and a Czech section in 1882. Jan Palacký (1830–1908) was appointed the first Czech professor of geography. The Czech Geographical Society was established in 1894 and between 1920 and 1939 and between 1945 and 1990 it existed as the Czechoslovak Geographical Society. The best time for its activity was the period of coexistence with the Geographical Institute of the Czechoslovak Academy of Science (1962–1993). The most famous director of this institute was the geomorphologist Jaromír Demek (born 1936), who was also the president of the society from 1978 to 1981. Nowadays, Czech geography is predominantly an academic discipline and it is taught at eight universities (Prague, Brno, Ostrava, Olomouc, Plzeň, České Budějovice, Ústí nad Labem, and Liberec) where regional branches of the society have been established. The most renowned one is Charles University in Prague, which is known especially for the Albertov School of Human Geography owing to two theoretically oriented geographers Jaromír Korčák (1895–1989) and Martin Hampl (born 1941).
Soil development: conceptual and theoretical models

Ronald Amundson
University of California at Berkeley, USA

Models and theories

A model implies the creation of a verbal or mathematical representation of how something works. There is considerable inconsistency in the definitions or distinctions between model and theory. However, it is widely agreed that theories provide falsifiable predictions of the phenomenon being examined. For example, a mass balance model of soil on a hillslope that states that the soil thickness is the balance of erosion and soil production provides little in the way of a falsifiable prediction. Yet, such a model that states that rate of soil production is inversely correlated with the soil thickness, and the rates of erosion are positively correlated with soil thickness, is indeed a testable hypothesis, providing a theoretical underpinning to the model. This review of models of soil development focuses on the combined development of both theory and conceptual models in the field of pedology over the past two centuries.

The concept of soil

The science of pedology focuses on the processes that form and distribute soils on the landscape. Pedology is largely a historical and observational science, as opposed to other related branches of soil science that are mainly laboratory based and experimental. Soil is the object of study in pedology. It is sometimes said that “soil means different things to different people.” Yet, for pedology to be a science, there must exist a definition that reflects a theoretical understanding of soil. For most of human history, soil was viewed through the lens of agriculture, or as a construction material. Since about the mid-1800s, however, the definition of soil has had a sound theoretical basis. The Soil Science Society of America’s Glossary of Soil Science Terms provides two definitions of soil.

1. The unconsolidated mineral or organic material on the immediate surface of the Earth that serves as a natural medium for the growth of land plants.
2. The unconsolidated mineral or organic matter on the surface of the Earth that has been subjected to and shows effects of genetic and environmental factors of climate (including water and temperature effects), and macro- and micro-organisms, conditioned by relief, acting on parent material over a period of time. A product-soil differs from the material from which it is derived in many physical, chemical, biological, and morphological properties and characteristics.

The first definition above is purely utilitarian and has no theoretical basis. The second definition is based on a theory, one that began with the work of the Russian scientist Vasily Dokuchaev. By the time of the publication of his treatise Russian Chernozem (Dokuchaev 1883), Dokuchaev had...
SOIL DEVELOPMENT: CONCEPTUAL AND THEORETICAL MODELS

already clearly recognized that soil is a natural body – a segment of nature distinct from other components – that reflects the “a) ground, b) climate, c) vegetation, d) topography, as well as e) the duration of the formation period of the soil” (Dokuchaev 1883; Minasny, McBratney, and Salvador-Blanes 2008). Dokuchaev examined the individual effects of each of these five factors on the soils of the Russian steppe. Most importantly, he recognized that this relationship had predictive power: “it is quite obvious that any two given points emerging simultaneously from under sea level, identical in physicochemical composition and topographic situation, inhabited by identical vegetation and affected by identical climatic conditions must have identical soils” (Dokuchaev 1883). For this reason, Dokuchaev and his colleagues are credited with developing the theoretical and definitional underpinnings of our field.

State factors

Nearly 60 years after Dokuchaev’s Chernozem, Hans Jenny began his book Factors of Soil Formation: A System of Quantitative Pedology with the simple sentence: “As a science grows, its underlying concepts change, but the words remain the same” (Jenny 1941). This concise, spartan prose announced what would be a reworking of Dokuchaev’s concepts into a physical framework amenable to numerical quantification and mathematical representation. A physical chemist by training, Jenny had been hired by the University of California at Berkeley in the 1930s to teach courses on pedology and mineralogy (Amundson 2005). In the course of this effort, Jenny began to view soil as an open physical system (with boundaries chosen by the investigator) whose properties could be defined by a set of external variables (e.g., the factors of soil formation)

(Figure 1). In the case of a starting material like an igneous rock or loess, the parent material is largely isotropic at \( t = 0 \). The fluxes of mass, water, gas, and energy between the atmosphere and the soil (along with biological and physical mixing) create a variously anisotropic system over time. As Jenny knew, the vertical trends in soil properties identify or demark the soil horizons.

These external variables or factors, drawn from Dokuchaev and others, were carefully defined so that they might be quantified and represented numerically in a mathematical framework. Briefly, these are:

- climate \((cl)\): the temperature, rainfall, humidity conditions surrounding the system;
- organisms \((o)\): the potential biota that can inhabit the soil system from the immediate surroundings (e.g., the gene pool of the potential biota). Note that this definition and conception are fundamentally different to the biota existing within the system, which is a system property or dependent variable, as opposed to an independent variable. Jenny (1958) elaborated at length on the nature of the biotic factor. This paper is critical
reading for anyone interested in the state factor model;

- relief or topography (r): the initial geographical configuration of the soil system;

- parent material (p): the initial physical/chemical makeup of the soil system (or the configuration at the beginning of a chosen period of study);

- time (t): the elapsed time since the parent material stabilized or since a period of observation began.

These factors, as defined by Jenny, are commonly independent of the soil system itself. Thus, the change in the soil system (s) itself is due to the combined impacts of each factor on the soil, written in the form of a differential equation (equation 1).

\[
\frac{ds}{dt} = \left( \frac{ds}{dl} \right)_{o,r,p,t} dl + \left( \frac{ds}{do} \right)_{d,r,p,t} do + \left( \frac{ds}{dr} \right)_{d,o,p,t} dr \\
+ \left( \frac{ds}{dp} \right)_{d,o,r,t} dp + \left( \frac{ds}{dt} \right)_{d,o,r,p} + \ldots (1)
\]

Because the nature of the mathematical relations between soil properties and any of the factors (e.g., ds/dl) was only poorly understood at the time, Jenny emphasized that pedologists strive to derive these functions through the observational study of sequences of soils (climosequences, chronosequences, etc.). A sequence of sites (carefully selected such that the other variables are constant) represents a fundamental experimental design that can allow the researcher to determine the nature of the relationship between the soil property and any given factor. In a sequence, four factors are held constant, while one (the one of interest) is allowed to vary. Jenny noted that there are two different ways to hold other factors constant. First, hold the variation in the other variables to zero. For example, if the property being examined has little relation to variations in soil age, then \( ds/dt = 0 \).

The appeal of this fundamentally clear and inherently understandable approach to soils is unmistakable. For this reason, many soils textbooks have been outlined by discussing the role of each factor on soils and soil properties using “sequences” of soils as examples. Indeed, a quick search of the Web of Science using “________sequence” as the keyword reveals hundreds to thousands of papers that use and incorporate these types of “sequence” concepts.

A key component of equation 1 is the ellipses or dots that follow it. They represent additional factors that may be important but are, as yet, undiscovered or unaccounted for. Indeed, Jenny had devoted considerable conceptual development to humans as a separate factor of soil formation, but chose in 1941 to keep them within the biotic factor. In the decades following his book, some noteworthy papers were published articulating the need to make humans an explicit soil-forming factor (Bidwell and Hole 1965; Yaalon and Yaron 1966). However, an advancement of this idea was provided in Amundson and Jenny (1991). In this paper, the authors struggled with the issue of why humans require a designation as a separate soil-forming factor. Briefly, the paper focused on the uniqueness of humans, and one of their qualities that is special relative to probably all other species: culture. Culture, like genes, is malleable and can be affected by “selection” when transferred to new environments (Diamond 2012). Culture contains the tools, technologies, and ethical framework that determine what people do to soil. Amundson and Jenny thus defined the human state factor as \( o_h, c_i \), where \( o_h = \) human genotypes and \( c_i = \) cultural inheritance. These independent (of
the system and of each other) external variables, once emplaced within a soil system, become dependent variables that over time may diverge from the values emplaced in the system at \( t = 0 \).

Another concept that Jenny (1961) elaborated on is what he called the “larger system” (soils + overlying biosphere), which is what is now known as an ecosystem. As Amundson and Jenny (1997) noted, this quantitative, factorial articulation of an ecosystem did not appear to become part of the ecological sciences until possibly the 1990s, as ecosystem ecologists and biogeochemists became increasingly interested in ecosystem and watershed mass balances. As an example of the linkage of Jenny’s (1941) ideas with modern ecology, many ongoing Hawaiian Island chrono/climo gradient studies have been designed using the state factor model (Vitousek 2004).

The state factor model is “phenomenological,” and articulates the relationship of soil properties to factors. Jenny was particularly clear that the soil forming factors “are not forces, causes, or energies, not are they necessarily environment. They have but one feature in common. They are the independent variables that define the soil system.” And, later: “The solutions given in this study are formalistic, as contrasted with mechanistic. In simple words, we endeavor to determine how soil properties vary with soil-forming factors. We exhibit but little curiosity regarding the molecular mechanism of soil formation and thus avoid lengthy excursions into colloid chemistry, microbiology, and so on. Such treatment will be reserved for [a] later occasion.”

Essentially, using this guidance, soil is viewed as a “black box.” A black box, in science, is a system where the workings remain unknown and only the output and the controlling conditions or processes are understood. The factors thus set the boundary conditions for the box; they are relevant to subsequent process-based or mechanistic investigations but do not explain or articulate the actual surficial and pedogenic processes that eventually form the soil. That is why this model is referred to as a factorial model, and not a process-based model.

In the intervening years, other models have been devised, especially to highlight the role of energy in the soil system (Runge 1973). Others have conceived of models that shift the emphasis to pedogenic processes and the resulting impact on the nature of the soil profiles (Simonson 1959; Johnson and Watson-Stegner 1987).

Finally, although the “state factor” approach is generally referred to as a model, it is not mechanistic or subject to simulation like a typical model. Additionally, it provides predictions that can be tested in the field. For example, existing work has shown that soil organic matter in well-drained soils increases with rainfall and decreasing temperature (Jenny 1941), a relationship mechanistically understandable when the processes involved are considered (Jenny, Gessel, and Bingham 1949). Thus, it should be possible to predict the organic matter content of unstudied locations based on climate information. And then, the predictions can be tested by subsequent field study. Taken to its extreme, as Dokuchaev predicted, two soils with the same soil-forming factors must (or, at least, should) be identical. While ultimately the distinction between model and theory may not seem essential, it is important to recognize the intellectual foundations of our science, which in turn impacts or guides other fields.

Factors versus processes: opening the black box

Opening the “box” that is the soil system to process-based research and models essentially occurred simultaneous to other developments
of the state factor approach – and sometimes by the same people. This synergy is to be expected, as one wishes to understand why a functional relationship exists, or how a given set of processes may be distributed across the landscape. However, the study of soil processes also owes its success to the contributions from other sciences; here, some of the key contributors to our understanding of soil dynamics are reviewed.

One overarching model for soils involves mass balance estimations. This approach was perhaps first conceptually developed – in pedology – by Roy Simonson (1959). Simonson’s goal was to recognize that soils, and soil-forming processes, operate as a continuum across the surface of the Earth. To most effectively link these features and processes, a general “balance” model was needed to organize or account for changes via additions, removals, transfers, and transformations to, within, and out of the soil system. This generic framework was deliberately designed to replace or quantitatively understand soil horizon changes due to more vaguely defined processes such as “melanization,” which is really a complex combination of processes resulting in the darkening of soil by organic matter, or rubification, which in general implies soil reddening. As Simonson (1959) articulated, the model can be applied to any number of soil processes. Here, some of the major developments in these specific models that have occurred over the last several decades, focusing on physical, chemical, and biological processes, are examined.

Physical processes

Soil on hillslopes is in constant, slow motion – downslope due to slope and surficial processes, downward due to gravitationally driven processes, and upward due to some forms of pedoturbation. Biological and physical processes mix or move soil particles, and, with the aid of gravity, they move downslope, as well as downward within the profile. This view, while now part of the pedological mainstream, began to develop before the science of pedology existed, and it took many years before it became a widely adopted pedological concept. Let us start at the beginning and trace forward.

James Hutton (1726–1796), the Scottish naturalist and scholar, is viewed as the father of modern geology. And yet, his reputation for this achievement might be based on his even more startling grasp and articulation of a conceptual mass balance model of soil formation. Hutton (1788) began his contemplation of upland soils from what, today, might be considered an unusual starting premise: we inhabit a world designed by God to support life and, in such a world, fertile soils are necessary to support plants. Yet Hutton was well aware that erosion occurred following rainstorms, and simple reasoning suggested to him that without a restorative mechanism of underlying rock being converted to soil, the soil mantle would gradually disappear. This outcome was inconsistent with his view of a habitable Earth. Considering this further, Hutton recognized that although soil erosion and production were inherently slow, together they would eventually wear down and level the landscape, leading to geologic/pedological stagnation. Thus, from his reading of geological strata – and his view of the requirement of a habitable planet – Hutton recognized that geological uplift and recycling must occur to maintain a general planetary steady state. Hutton’s contemplation of the relatively slow rates of soil formation and erosion, when linked to observations of geological strata in an array of rock records, provided “no vestige of a beginning, no prospect for an end.”

Nearly 90 years later, the American geologist G.K. Gilbert participated in the early geological expeditions to the American West. In his work
SOIL DEVELOPMENT: CONCEPTUAL AND THEORETICAL MODELS

Figure 2 A depiction of a hillslope mass balance model. The losses from the soil are due to physical losses (flux out flux in) and chemical erosion. Gains are due to atmospheric deposition and, most importantly, to the conversion of rock or saprolite to mobile soil. From Owen et al. (2010). Reproduced from Elsevier.

on the Henry Mountains of Utah, Gilbert (1877, 105) wrote that “over nearly the whole of the earth’s surface there is a soil.” Yet, unconstrained by Hutton’s starting premise, Gilbert noted that “the general effect of vegetation is to retard erosion; and since the direct effect of rainfall is the acceleration of erosion, it results that its direct and indirect tendencies are in the opposite directions.” Thus, Gilbert suggested that plants are responsible for the maintenance of hillslope soils. Gilbert’s short discussion of hillslope soil processes (Gilbert 1877, 99–108) added a number of testable hypotheses to Hutton’s general hillslope mass balance model. First, Gilbert indicated that erosion is proportional to “declivity” or downward slope gradient. Second, Gilbert hypothesized that the rate at which rock is converted to mobile soil (soil production) is inversely related to soil thickness, down to some critical minimal threshold, beyond which the soil cover no longer is capable of accentuating rock degradation, and the soil mantle disappears (see Figure 2). This model is today commonly referred to as the soil production function (Figure 3). It remained for physical scientists like Carsen and Kirkby (1972) to formalize these concepts and translate them into the language of mathematics:

\[
\frac{dH}{dt} = P - E \tag{2}
\]

where \(H\) = the mobile soil thickness layer (commonly equivalent to the A horizon) (L), \(P\) = soil production (L/T) and \(E\) = erosion (L/T). Some functions that are found to commonly describe \(P\) and \(E\) are:

\[
P = P_0 e^{(-\alpha H)} \tag{3}
\]

\[
E = \partial Q_s = K \nabla^2 z \tag{4}
\]

where \(P_0\) = maximum soil production rate at \(H = 0\) (or some critical depth), \(\alpha = a\) constant \((1/L)\), \(K = a\ constant\ \((L^2/T)\), and \(\nabla^2 z\) = slope curvature \((1/L)\). Once dislodged, soil material is inherently mobile from a geologic perspective, especially on slopes. This model is applied to
convex hillslopes, where curvature is negative. When curvature is negative, slope increases with distance downslope and there is a net removal of soil. When curvature is positive, slope declines and mass is accumulated from upslope from erosion. When curvature is zero, the slope is planar, and no net loss or accumulation occurs via erosion.

In what may be one of the most important papers for pedology in the last 15 years, Heimsath et al. (1997) developed a means of using cosmogenic radionuclides to determine the rate of soil production, or the conversion of rock or saprolite to soil material (e.g., equation 3). In their paper, the authors provided the first empirical test of Gilbert’s (1877) hypothesis that soil production from bedrock should be inversely proportional to the thickness of the overlying soil mantle. This paper in effect launched the present synergy between pedology and geomorphology, and introduced a new set of tools that now stimulate both fields. Subsequent to Heimsath et al.’s (1997) physical mass balance model, Riebe, Kirchner, and Finkel (2003) and Yoo et al. (2007) provided means for quantitatively integrating chemical and physical mass balances on hillslopes. In particular, Yoo et al.’s (2007) numerical framework, when coupled with cosmogenic nuclide derived soil production rates along hillslope transects, allowed soil scientists to determine if the rate of soil erosion is related strictly to slope gradient, or to other parameters such as soil thickness.

Chemical processes in soils

Water is a pervasive solvent of rock, and a mechanism which moves both suspended and dissolved material through the soil profile. Quantifying the chemical alteration of soil is needed to understand weathering rates, the release of plant nutrients, and the role of soils in global biogeochemical cycles. By the late twentieth century, characterization of soil profiles had advanced greatly. Terms and procedures had been standardized for describing and characterizing the physical, chemical, and mineralogical composition of soils. Yet, these measures by and large were selected for their applications to agriculture, soil classification, or other practical uses of soil.

In large part, total chemical analyses of soils – so popular in the early twentieth century – had fallen out of favor during this period. A fortuitous renaissance occurred when Hans Jenny took two Berkeley geoscientists on a field trip to examine a unique groundwater in Spodosol on the northern California coast. One outcome of this trip was the inclusion of this soil as an example in a paper illustrating how total chemical analyses can be used to quantitatively examine mass and volume changes over time. This chemical mass balance approach, which was introduced to soil science by Brimhall and Dietrich (1987), or arguably reintroduced (Merrill 1906), has become one of the most important soil mass balance models in the last 30 years. Applying principles and methods used to study ore body evolution in economic geology, Brimhall and Dietrich (1987) described the use of index elements to quantify gains and losses of elements that had occurred in soils via chemical weathering. A key component of this model is the “tau” (τ) value, the fractional gain or loss of an element in a soil horizon relative to the parent material:

\[
\tau = \frac{R_s}{R_p} - 1
\]

where \(R_s\) and \(R_p\) are ratios of the concentrations of mobile to immobile elements in soil horizons or the parent material. While not explicit on the specific mechanisms involved, the approach...
provided quantitative tools for determining total elemental changes during soil development. This quantification of chemical processes in soil assisted in changing the soil system “from a black box to a grey box,” as suggested by Chadwick, Brimhall, and Hendricks (1990). Minasny, McBrantney, and Salvador-Blanes (2008) also pointed out that the “work by Brimhall and Dietrich (1987) is the cornerstone of the last 20 years of soil formation work by many people.”

This relatively simple, soil chemistry based, mass balance model might be viewed as the foundation on which more process-based modeling is presently based. In particular, White et al. (2009), Brantley and Lebedeva (2011), and Maher (2010) used total soil chemistry, mineralogy, pore water chemistry, and thermodynamic data to understand mineral weathering rates, rates and depth of clay formation, and the limiting factors controlling chemical weathering rates of soils. While these very quantitative models lie outside of the scope of this review, they are representative of the novel directions currently underway in soil modeling (Minasny, McBrantney, and Salvador-Blanes 2008).

Biological processes in soils

The importance of macroscale, biological processes on soil formation have been recognized since the development of the field, yet their quantitative study seems to have waxed and waned over time. Dokuchaev and others clearly understood the role of vegetation in the formation of the dark surface horizons of the Chernozem soils of Russia. Gilbert (1877) recognized the role of biota in stabilizing loose sediment on hillslopes, as well as in their effect on breaking up rock and converting it to loose soil. In a largely unrecognized side note to the development of hillslope geomorphology in the nineteenth century, Darwin (1881) independently arrived at the idea that sediment movement downslope by earthworms is also proportional to slope gradient. In spite of his self-acknowledged weakness in mathematics, Darwin provided data on rates of soil movement across slope contour lines. These data provided scientists the opportunity to calculate a numerical value for a slope-dependent, earthworm-mediated, transport coefficient that is implied in Gilbert’s (1877) discussion (i.e., sediment flux rate = K • (slope)).

However, an additional component of Darwin’s research focused on the effect of earthworms on soil mixing and their direct and indirect movement of large particles/clasts over time. Darwin’s observations of gravel, stones, and artifacts in soils revealed that, over time, fine material brought to the surface by worms eventually led to the downward migration and burial of the larger objects, which the worms (bioturbators) could not move (Figure 4). Thus, these larger objects tended to settle to the maximum depth of their burrowing activities. Although the broader implications of these processes to pedology (by gophers, etc.) seem obvious in hindsight, Darwin’s work remained outside of the pedological mainstream for decades.

In the past few decades, an important addition to the pedological lexicon has been the term biomantle. A biomantle is composed of material that has been mixed and homogenized by plants and burrowing organisms. It usually occupies the uppermost part of the profile (Figure 4). The term first appeared in 1975, but its growth as a key process in pedology is largely due to the efforts of Donald Johnson (1934–2013). Johnson, Domier, and Johnson (2005) suggested that Darwin’s work on earthworms and bioturbation represents a key and ignored pedological paradigm. Although Darwin did study how worms help mix and form A horizons, his focus
was not on soil formation as a phenomena, or on the soil itself.

The biomantle concept also links well with the model of hillslope soils discussed above. Geomorphologists recognize that such a “mobile soil layer” – soil lacking any residual rock or sediment structure – is commonly the soil’s A horizons, and thus the biomantle constitutes a conveyor belt of soil that slowly creeps down many hillslopes. This concept has been elaborated upon and developed in a model of soil and hillslope development called the dynamic denudation

Figure 4 The conceptual stages in the biological mixing of a parent material into a biomantle (Johnson, Domier, and Johnson 2005. Reproduced by permission of Taylor & Francis Ltd., www.tandfonline.com). The mixed, mobile layer is commonly equivalent to the A horizon in many soils.
model (Johnson 1993), which not only helps explain slope processes but also the biomantles and stone lines that commonly underlie them.

Conclusions

Dokuchaev and those that followed provided the academic community with both our central pedological theory and a scientific definition of soil. The theory has been subsequently refined and made more amenable to quantification, and has guided and stimulated research in a number of fields. The conceptualization of soil as an open physical system, in turn stimulated the development of mass balance models, many of which are now quite mathematical and amenable to computer simulations. However, it is the original, conceptual breakthroughs or illuminations that started the process.

Although it might be easy to critique the sometimes vague scope of early conceptual models and breakthroughs – because they do not immediately lend themselves to a particular quantitative application – it is important to consider how they set in motion new ways of seeing things and transforming a field. Theories are the parables of science, allowing us to see and investigate what we otherwise have not seen.

SEE ALSO: Soil mapping and maps; Soils in geomorphic research; Soils on slopes: catenas; Soils of urban and human-impacted landscapes

References


Disturbance in biogeography

Thomas T. Veblen
University of Colorado Boulder, USA

Disturbance defined

Disturbance is an inherent component of all ecosystems affecting both the local and broad-scale spatial and temporal patterns of biota across the Earth. Although definitions of ecological disturbance vary in their nuances, the definition most widely adopted by biogeographers and ecologists is: “a disturbance is any relatively discrete event in time that disrupts ecosystem, community or population structure and changes resources, substrate availability, or the physical environment” (White and Pickett 1985, 7). Understanding the roles of disturbances in different ecosystems is particularly critical to the field of biogeography because it aims to understand how natural processes, as well as human activities, determine biotic patterns across spatial scales from the local to the global and across a range of temporal scales. Although disturbance is an important concept in marine and aquatic sciences, this review will focus on terrestrial ecosystems, given their greater relevance to physical geography.

Intellectual context and societal importance

Ecosystems are shaped by natural disturbances of numerous kinds such as windstorms, fire, flooding, snow avalanches, landslides, herbivory by large mammals as well as by insects, and outbreaks of disease-causing organisms (pathogens). Anthropogenic disturbances such as logging, treading, altering fire activity, or manipulation of grazing animals are also key influences on many ecosystems. The ubiquity of disturbances in all ecosystems, and their importance in shaping ecosystem structure and function, make them a vital component of biogeographical analyses of how and why vegetation patterns vary across the globe.

Disturbances influence plant communities through their direct effects on plant biomass by killing entire plants or removing parts of plants which in turn affect competition, environmental conditions, and the availability of substrate and resources to other plants. When disturbance removes or partially removes a large plant, in many environments the resources released may be used by other plants of the same or of other species. This sets in motion changes in biotic interactions and population dynamics that may significantly alter the composition and structure of the vegetation. Thus, disturbance is a primary cause of spatial heterogeneity in ecosystems. When disturbance affects a single plant or small
DISTURBANCE IN BIOGEOGRAPHY

group of plants (such as a single tree-fall), the result is an alteration of fine-grain patchiness. Coarse-grained patchiness results from larger areas disturbed such as in the case of fire. Thus, in the context of biogeography, given its focus on patterns at all scales, disturbance is a major source of patchiness and heterogeneity in the landscape.

Repeated disturbance is a key process in maintaining biodiversity because species have adapted over evolutionary time to environmental conditions reflecting different frequencies and intensities of disturbance. Following a particular disturbance, some species invade or increase in abundance while others are eliminated or decrease in abundance. Some species are adapted to the environmental conditions that immediately follow a severe disturbance, whereas others thrive only after many decades of biotic and abiotic changes that occur in the absence of a severe disturbance. The former are sometimes termed pioneer, early seral, or colonizing species, and the latter are often called climax, late seral, or late successional species. The co-occurrence of plant species with different adaptations to different frequencies and severities of disturbance is of fundamental importance to the resiliency of plant communities to disturbance. Differences in species traits such as dispersal and competitive ability and the capacity to grow under different levels of resource availability (e.g., light, nutrients) provide mechanisms by which ecosystems recover after disturbances and contribute to overall ecosystem stability.

Appreciating the role of natural disturbances in vegetation dynamics is fundamental for understanding the roles of humans in altering natural communities and in managing resources. This general principle is inherent in land management paradigms such as ecosystem-based management, emulation of natural disturbance, and historical range of variability (Hayward et al. 2012). The idea that ecosystems should be managed in a manner consistent with their historical structure and functioning as related to the disturbances that shaped these ecosystems was proposed in the 1990s as an alternative to management paradigms guided primarily by resource extraction. In the context of land management there is a strong consensus that effective land management of forest and rangeland ecosystems depends on a thorough understanding of the history and effects of disturbances on such ecosystems. Ecological science has demonstrated the importance of natural disturbances in determining ecosystem conditions and trajectories: equilibrium conditions are rare or short-lived in most ecosystems. During the second half of the twentieth century, in ecological science there was a shift away from a previous emphasis on equilibrium (i.e., stasis) toward acceptance of continuous variation in species composition and ecological structure. This resonated in the 1990s with the practical experiences of many land managers in dealing with the effects of changes in fire activity and the consequences of species introductions and extirpations (local extinctions). Research in disturbance ecology is inherently interdisciplinary and is strongly developed in disciplines such as physical geography, ecology, forestry, and range management.

Historical perspective: from Clements’s superorganism to Odum’s strategy of ecosystem development

The modern concept of plant succession was developed largely in North America at the turn of the nineteenth and twentieth centuries, and owes much of its origin to the trend among natural scientists during the nineteenth century to shift away from static views and toward dynamic views of nature. At broad scales, this trend toward acceptance of changes in species distributions and their traits is based on the
development of ideas in evolutionary theory and their applications to explaining plant and animal distributions by Charles Darwin and Alfred Russell Wallace. Key to their viewpoints was an understanding of plant populations changing over time as opposed to creationist views of the Earth and its landscapes which stressed stasis. This shift toward an appreciation of the Earth’s surface as a dynamic phenomenon is also exemplified by the theory of the “geographical cycle” developed by the geologist and geographer William Morris Davis in 1899. Davis’s theory stressed the orderly development of landforms, starting with the uplift of a new erosion surface and culminating in the “peneplain,” a rolling landscape of low relief. The progressive development of the physical landscape was explained by analogy with the development of a human organism by describing a landscape as passing through stages of youth, maturity, and old age. This organismal metaphor, emphasizing change and evolution, was a departure from earlier views that emphasized the stasis of landscapes, and such metaphors were typical of the post-Darwinian nineteenth century. One of the pioneers in the early scientific study of plant succession was Henry C. Cowles, who used a theoretical framework similar to Davis’s geographical cycle to describe and explain changes in the species composition of dune vegetation on the shores of Lake Michigan.

It was Frederick Clements, an American botanist, who in publications during the first two decades of the twentieth century developed the first comprehensive theory of plant succession that included general mechanisms and causal explanations (Glenn-Lewin, Peet, and Veblen 1992). Clements’s theory was based on an order and logic that dominated most research on plant succession for much of the twentieth century and that continues, often unacknowledged, to influence the way land managers and the general public perceive vegetation change. Much of the modern debate about the causes and nature of vegetation change harken back to the issues of predictability, convergence, and equilibrium which were stressed in Clements’s theory of succession. Clements viewed succession as a highly predictable and orderly process in which vegetation change was, in fact, the “life history” of a plant community, thus making an analogy to an organism. He believed that, although initial starting points may differ in environmental conditions and species composition, succession would eventually result in convergence to the same type of mature or climax vegetation over an area of uniform regional climate. In his theory, the climax vegetation was believed to be a highly stable equilibrium condition unlikely to change unless there was a significant change in the regional climate. In modern ecology this viewpoint is termed an equilibrium perspective because it requires the concept of a stable or climax vegetation. Today, it is also regarded as a deterministic model because it postulates such an orderly and predictable series of vegetation changes inexorably culminating in the development of the climax vegetation. In such a deterministic model, the role of randomness is denied, and the model is assumed to always produce the same output from a given set of initial conditions.

Two key early critics of Clements’s succession theory were the American botanist Henry A. Gleason and the English ecologist Arthur G. Tansley (Glenn–Lewin, Peet, and Veblen 1992). As early as the late 1910s, Gleason argued that Clements’s view of a plant community as a highly integrated organic entity, sometimes termed a superorganism, was not consistent with field observations of the individualistic behavior of plant species. Gleason stressed the role of chance events such as availability of bare areas for colonization and unpredictable timing of the arrival of dispersal units (seeds,
DISTURBANCE IN BIOGEOGRAPHY

fruits) of a plant species with overlapping but not identical environmental tolerances. Thus, in Gleason’s view, the combined influences of chance dispersal of propagules and variability in sites over time and space will lead to independent distributions of species as opposed to static combinations of the same species. In his view, plant communities were composed of species with individualistic but overlapping distributions so that broadly similar plant communities actually were unique assemblages of species. While they might be similar in dominant species they were actually different in species composition from place to place. His view became known as the individualistic concept of the plant association.

Tansley, writing mostly in the 1930s, challenged Clements’s assumption that all vegetation in a region of similar climate would ultimately converge in species composition toward a single type of climax vegetation. He argued that other controlling factors, such as unusual geological substrate and topographic position, would result in undisturbed vegetation types quite different from that which might be allowed by climate under more favorable geological and topographic conditions. Thus, he recognized a series of stable vegetation types that might be associated with waterlogged soils induced by topographic position or nutrient-poor soils associated with particular rock types. His viewpoint became known as the polyc climax theory in contrast to Clements’s monoc climax theory. Thus, Tansley and his followers’ critiques of Clements’s view of the climax centered on the degree of convergence to a single type of climax. As discussed later in the entry, this debate has largely become moot as modern ecological studies revealed the importance of repeated disturbances, as well as climate variability and climate trends, in resetting the direction of successional change.

In the early 1950s the American ecologist Robert H. Whittaker merged the key elements of Gleason and Tansley to describe vegetation as varying spatially simply because of the continuous variability of the underlying physical environment. Whittaker’s viewpoint, known as the gradient concept of vegetation, eventually became the predominant theoretical framework for conceptualizing how vegetation varies across space and was illustrated by numerous quantitative studies of plant species distributions along environmental gradients in mountainous environments. Highly similar ideas emerged at approximately the same time in the 1950s based on quantitative studies of species distributions along environmental gradients in less topographically dramatic settings in the upper Midwest of the United States. These studies were led by J.T. Curtis of the University of Wisconsin, and the resulting conceptualization of spatial variability in vegetation was termed the continuum concept of vegetation.

While many aspects of Clements’s theory of plant succession are not supported by modern ecological science, his ideas were important in directing research on the mechanisms by which vegetation changes in response to disturbance. These mechanisms include the dispersal and establishment of plants at a disturbed site, competition and other biotic interactions, and the modification of the abiotic environment by plants in ways that may affect further vegetation change. He particularly stressed the role of reaction as the dominant mechanism resulting in successive changes in plant community composition toward some sort of climax. The aspects of Clements’s theory that are no longer strongly supported include his emphasis on long-lasting stability in the form of a climax vegetation, the high degree of convergence of initial differences toward a uniform climax, and his overall emphasis on predictability. Modern views of vegetation change recognize greater variability in successional pathways and the frequent importance
of stochastic events such as timing of dispersal events, and do not stress long-term stability. In addition, modern viewpoints stress repeated disturbances of varying types and severities that continue to alter vegetation in contrast to the role assigned by Clements to disturbance as simply an initial trigger of progressive vegetation development.

In 1969 Eugene Odum published a highly influential paper entitled “The Strategy of Ecosystem Development” which succinctly summarized numerous trends to be expected in the course of succession. Odum’s synthesis reflected a conventional view of succession as presented in textbooks of the time, and aimed to integrate earlier perspectives on succession with key concepts of the emerging field of ecosystem science. He summarized a number of successional trends in community- and ecosystem-level properties which he postulated result from the tendency of ecosystems to develop toward greater homeostasis or control of the physical environment in the sense of attaining maximum protection from disturbances. He postulated that the strategy of ecosystem development is, in fact, directional change toward greater stability or homeostatic control of the environment by the organisms dependent on that environment. He described succession as an “orderly process of community development that is reasonably directional and, therefore, predictable” (Odum 1969, 262). The mechanism of successional change that he emphasized was the modification of the physical environment in ways that drive species changes in a predictable way. Odum asserted that succession eventually culminates in a stabilized ecosystem in which maximum biomass and self-reinforcing interactions among organisms are attained for that particular physical habitat. Thus, although described in ecosystem parameters not defined until the mid-twentieth century, Odum’s strategy of ecosystem development is highly similar to Clements’s theory of succession. Although Odum’s synthesis was widely accepted at the time, reproduced in textbooks, and cited for its implications for the management of global resources, it was seriously challenged by William Drury and Ian Nisbet in a critique published in 1973. They systematically critiqued the 24 trends identified by Odum, finding that the supporting data were weak, contrary, or nonexistent. Although Drury and Nisbet’s 1973 paper was published in a relatively obscure (at least to ecologists) journal (Journal of the Arnold Arboretum) after apparent rejections from mainstream ecological journals, the questions about the Clements–Odum model raised by this paper triggered a renewal of research interest in succession in the 1970s. The late 1970s to early 1980s saw the publication of numerous empirical and review papers re-evaluating the mechanisms of succession, emergent properties associated with succession, and the relative stability of successional end points (i.e., climax). These studies formed the basis for the development of the modern nonequilibrium paradigm of vegetation dynamics in which repeated disturbances play central roles.

### Disturbance in the context of the modern nonequilibrium approach to vegetation dynamics

The modern nonequilibrium view of vegetation dynamics is a consensus paradigm articulated in numerous publications in 1980s for which Drury and Nisbet’s 1971 and 1973 papers were foundational, regardless of whether they were cited. Drury and Nisbet (1971) insightfully described Odum’s and Clements’s views of succession as developmental theories. In such a developmental theory, succession is viewed
largely as the consequence of relationships and interactions within a community, whereas external influences such as coarse-scale disturbance, climate variability, and the arrival of new species are either held constant or relegated to minor roles. In the developmental theory of succession, it is assumed that the physical site and external influences on biota remain constant for a long period during which there is a progressive development toward a type of ecosystem that maximizes stability. Thus, an inexorable trend toward greater homeostasis in Odum’s model is the deterministic driving force underlying successional trends. In contrast, Drury and Nisbet (1971) proposed that there should be no assumption of long-term site stability (i.e., the absence of disturbance or climate variability), nor should the theory depend on the existence of a stable end point. They termed their ideas a “kinetic” viewpoint of vegetation change that is now more commonly termed a dynamic and nonequilibrium view of vegetation change.

While the nonequilibrium view of vegetation dynamics emerged as the consensus view in the 1980s, it is important to recognize that other ecologists had articulated similar ideas much earlier. Remarkably, as early as 1901 Cowles described succession as “a variable approaching a variable rather than a constant” to emphasize the transience of any predicted climax vegetation (Glenn-Lewin, Peet, and Veblen 1992, 6). Likewise, in work published mostly in the 1930s to 1950s Hugh M. Raup of Harvard Forest pointed out that the time period required for succession from a pioneering community on a bare site to a climax community was so long that the assumption of environmental constancy, especially of climate, was not tenable. In the 1910s to 1930s, with a strong focus on the population mechanisms responsible for temporal change in forest structure and composition, Alexander S. Watt of Cambridge University conducted a series of insightful studies of beech forest dynamics in England that became the basis for modern studies of tree-fall gap dynamics in forests. He lucidly revealed the importance of the role of tree-fall gap-phase processes of tree population dynamics in maintaining the relative stability of species composition at a coarse scale across the entire forest stand. Raup and Watt presaged, mostly through their empirical field observations, the recognition of the lack of constancy in the determinants of vegetation attributes at any particular place in the landscape and, in particular, the importance of variability induced by disturbances of different types, severities, and frequencies.

While Clements recognized the importance of coarse-scale disturbances such as fire and glacial retreat as triggering events of succession, he and his followers focused nearly all their research on how interactions among plants and plant influences on the site (i.e., reaction) would determine the course of succession during long disturbance-free intervals. Modern ecologists also continue to examine the types and relative importance of these plant-centered mechanisms of succession, but today equal attention is devoted to quantifying disturbance regimes and developing frameworks for predicting both disturbances and vegetation responses to disturbances (Turner 2010).

The definition of disturbance by White and Pickett (1985) presented earlier has been widely adopted as a general purpose definition of proven utility in many ecological applications, but it is not meant to be an exclusionary definition. Thus, it easily applies in the common situation where a disturbance event is destructive of biomass and thus releases resources, often creating a relatively bare area or open space. Frequently, a disturbance results in the release of resources, increasing their availability to surviving or newly arriving plants such as in the case of a tree-fall and the resulting understory
response. However, this is not always the case. A disturbance such as an extensive volcanic ash deposit may result in a subsequent reduction in soil-related resources such as available plant nutrients and soil moisture. Although many disturbances, such as fires, can result in relatively discrete patches in the landscape, others, such as herbivory, may be spatially diffuse. Disturbances that occur suddenly are easily perceived, but at some spatial scale some disturbances may not be distinguishable from slow, gradual trends in environmental conditions, sometimes referred to as secular environmental changes. “Disturbance” and “perturbation” are often assumed to be synonyms, but the latter is more commonly applied in experimental situations where variables can be artificially manipulated and equilibria can be perturbed. The term “disturbance” has proven more durable in applications in field ecology and related modeling treatments.

A large amount of definitional and classificatory terminology has developed around the term “disturbance,” most of which is not particularly helpful in understanding disturbance processes or their effects on ecosystems. Sometimes disturbances are classified as endogenous as opposed to exogenous, where the former originate from processes inherent in the plant community (e.g., plant death due to senescence) as opposed to exogenous disturbances that originate independently of the plant community (e.g., volcanic eruptions or earthquakes). However, many disturbances cannot be predicted or their origins understood without considering both factors that are internal and external to a plant community, such as fire which depends on the fuel attributes of a plant community as well as external ignition sources and weather. Analogously, “autogenic” and “allogenic” are somewhat archaic terms used to distinguish between processes controlled by the actions of plants themselves and processes that result from factors independent of the plant community (e.g., drought).

More important than the terminology surrounding the concept of disturbance is a practical way to describe the attributes of disturbances affecting a particular ecosystem which is accomplished with the concept of a disturbance regime. The sum of all the disturbances affecting an ecosystem is known as its disturbance regime. Thus, a disturbance regime is a description of all the disturbances affecting an ecosystem, landscape, or even a larger region. To be useful, a disturbance regime must be explicitly linked to a real place in the landscape and to a defined time period. For example, it is necessary to identify a specific study area (e.g., perhaps a small valley or a larger landscape) and to state the time period for which the disturbance regime applies, which may range from a few years to millennia. Synthesis of descriptions of disturbance regimes for many actual ecosystems and landscapes may allow generalization for these ecosystem types (i.e., classification categories), but it cannot be assumed that superficially similar ecosystem types will have identical disturbance regimes. The key descriptors of a disturbance regime are the type of disturbance, spatial attributes, temporal characteristics, specificity, and interactive components.

An important research goal in modern ecology and biogeography has been to quantitatively describe as many of the potential descriptors of disturbance regimes (Turner 2010) as feasible. The applicability of the various descriptors will vary according to disturbance type, with discrete high-severity fire events probably being the disturbance type for which it is easiest to conceptualize the descriptors. Spatial characteristics include the size of the area disturbed, potentially its shape, and its spatial distribution relative to pertinent environmental settings (e.g., according to slope aspect, steepness, and other topographic variables). Temporal characteristics include, most
DISTURBANCE IN BIOGEOGRAPHY

importantly, frequency (number of events per period of time) which, when combined with data on the area disturbed, can yield an estimate of a rotation period or the time required to disturb an entire study area once given the estimated frequency and areas of disturbance events. Environmental heterogeneity within a large study area often results in multiple fire events (or other disturbance events) in parts of the study area before the first such event occurs elsewhere in the same study area. Thus, rotation periods must be interpreted cautiously in terms of their applicability to a particular point in a large landscape. The same caution applies to a disturbance such as forest blowdown, where turbulence induced by distance from ridges or tree instability associated with topographic differences in soil drainage complicate the computation of a rotation period for a large, topographically complex landscape. Magnitude refers to the intensity of the event (e.g., heat released by a fire) whereas severity refers to the damage done by the disturbance (e.g., size and number of trees killed by a windstorm or by a fire).

Specificity as an element in the description of a disturbance regime describes whether the disturbance is selective according to plant species, which is likely to be an important factor in cases of herbivory or pathogen attack. Specificity by the size or age of the plant may also be important in some cases such as preferential insect attack on only large, old trees or greater susceptibility to wind damage among taller trees. The residual organisms or propagules present after a disturbance event such as a severe fire or blowdown are fundamental in determining potential post-disturbance recovery trajectories. These are also called biological legacies or remnants (Turner 2010) and are particularly important in assessing the effects of human-caused disturbances (e.g., land clearance) on the future viability of patches of vegetation surrounded by intense human land use. The term “synergism” was initially used to describe the potential effects of one disturbance on another, and (as described later in the entry) disturbance interactions constitute a theme of major current research activity.

Current research frontiers

Applications in wildland management

The most common context for research conducted by biogeographers and ecologists on disturbance phenomena and their impacts on vegetation during the late twentieth century has been in the area of management of wildlands. In some cases the land-use context has been commodity production (e.g., timber, livestock), but in most cases the context has been environmental protection and conservation science. Specific research goals have often been to inform land management policies aimed at ecological restoration or mitigation of societal vulnerability to disturbances such as wildfire. Studies of disturbance ecology play a central role in modern research management approaches such as ecosystem-based resource management and emulation of natural disturbance. There is widespread agreement among scientists and managers that disturbance ecology, and in particular understanding of the history and effects of disturbance in the landscape, provides essential insights for decision-making. This consensus is based on the premise, supported by ecological theory and practice, that comprehending the causes and consequences of temporal variability in ecosystems is crucial to successful resource management (Hayward et al. 2012).

Numerous studies in disturbance ecology have been and continue to be conducted because of their direct relevance to wildland management issues such as ecological restoration and fire mitigation. Such studies are solicited by the public
and policymakers because of three key principles (Hayward et al. 2012): (i) knowledge of past natural variation provides an essential reference for evaluating impacts of modern land-use practices such as grazing, fire suppression, and logging on current ecosystem conditions and processes; (ii) past natural disturbances have played key roles in structuring contemporary ecosystems, and will continue to do so in the future; (iii) hypotheses about the drivers and mechanisms of future ecological change driven by climate change must be informed by understanding the roles of past climate-related disturbances on the landscape.

Disturbance interactions

Despite the existence of hundreds of studies of disturbance regimes, there are relatively few which explicitly consider more than a single type of disturbance in the same study area. Thus, most fire history studies consider data only on fire and most tree-fall gap dynamics studies present only data on tree-falls. In many of the study areas targeted by these studies, it is evident that multiple types of disturbances play important roles in shaping the ecosystem structures. For example, many fire history studies recognize that grazing in the same area has probably had an important influence on the vegetation and indirectly on fire through alteration of fuel quantity or fuel type, but rarely are quantitative data available on the grazing component of the disturbance regime. Analogously, there are numerous studies of hurricane impacts on tropical coastal forests in areas where land-use practices (grazing, hunting, burning, and agricultural land abandonment) are known to be important influences on forest composition and structure, yet rarely are quantitative data on land-use practices effectively integrated with temporal and spatial data on blowdowns. Such integration of quantitative data on multiple types of disturbances is one of the current research challenges in disturbance ecology.

In contrast to the past tendency to focus on a single disturbance type, numerous studies of coniferous forests in western North America recently have yielded robust quantifications of disturbance regimes of both fire and lethal insects (typically bark beetles) at landscape scales (e.g., Veblen et al. 1994). Many of these studies have been motivated by questions of whether widespread bark beetle outbreaks will influence subsequent fire frequency, extent, or severity. This emerging research frontier is examining both linked disturbances (in the sense of Simard et al. 2011) and compound disturbances (in the sense of Paine, Tegner, and Johnson 1998). Linked disturbances are disturbances that interact by altering the extent, severity, or probability of occurrence of a subsequent disturbance. Although extensive tree mortality caused by bark beetles intuitively leads to expectations of increased subsequent extent, frequency, and severity of fires, dozens of retrospective field studies and fuel-driven fire behavior simulation studies have failed to demonstrate a clear or consistent pattern of increased wildfire activity. This is currently a subject of intense research activity but, at this date, the consensus is that widespread beetle kill may somewhat alter fire behavior under particular conditions (i.e., pre-beetle kill forest structure, severity and time since beetle kill, and fire weather conditions), but it has not fundamentally altered the wildfire hazard in affected forests. Both beetle kill and wildfire activity are strongly driven by warm, dry weather conditions, and both have increased in extent in western North America in conjunction with climate warming during the late twentieth century. Compound disturbances refer to the occurrence of two disturbances at relatively short intervals so that they have a synergistic effect greater than the effect predicted from the sum
of the individual impacts of the disturbances. For example, two surface fires occurring at short intervals may prevent successful tree regeneration as a result of insufficient time for juveniles to reach a size at which they are fire-resistant. Another example could be a beetle outbreak that selectively removes seed sources of one species so that, when it is followed by a fire, there may be greater success of postfire regeneration of a non-host species such as in the case of beetle-killed conifers and quaking aspen in western North America. Although the potential for disturbances of the same type or different types, occurring in different sequences, to amplify or attenuate the effects of individual disturbances has long been recognized, there are few empirical studies that have quantified these potential effects.

Alternative stable states

A second major area of current research activity in disturbance ecology, which overlaps with the study of disturbance interactions, is the study of how positive feedbacks from disturbances may result in major long-lasting switches in vegetation types, or so-called alternative stable states. Alternative stable states in species assemblages occur when under the same broad-scale environmental conditions (especially regional climate) and a similar biota, dramatically different species assemblages can occur side by side in apparently the same environment. For example, the presence of extensive tropical savannas of sparse populations of small trees and shrubs mixed with tall grasses and juxtaposed with tall tropical forests under the same regional climate has long intrigued biogeographers. Ecologists and biogeographers have documented the role of different fire regimes in maintaining such sharp boundaries. Whereas savannas are easily burned by natural or human-set fires and therefore are subject to frequent grass-fueled fires, forests in the same environment rarely burn because of their less flammable fuel types and their microclimate which is less conducive to fire. Repeated intentional understory burning of forests often associated with exceptional drought can result in microclimatic and fuel-type feedbacks that increase flammability and result in a switch to a savanna.

Alternative stable vegetation states with abrupt boundaries resulting from different fire regimes are also documented in temperate ecosystems where tall forests are replaced by shrublands or mixtures of tall shrubs and small trees. For such vegetation switches to be maintained beyond the life span of the initial postfire occupants of the site requires feedback mechanisms that increase the probability of a subsequent fire to perpetuate the postfire species assemblages. These mechanisms in the pyrogenic plant communities have been documented to include a wide variety of attributes such as greater flammability of individual species due to their physical and chemical properties, spatial arrangements and sizes of fuels, and fire-favoring microclimatic conditions. Some of these pyrogenic traits may also be selected for by the effects of herbivory by introduced wild and domestic mammals. For example, where herbivory prevents the regeneration of obligate seeding trees but allows the resprouting of browse-resistant shrubs with plant architectures that promote the spread of surface fires into shrub canopies, herbivores contribute to the perpetuation of the more flammable shrub alternative stable state. In some ecosystems, rare, severe droughts are required to facilitate fires in forests that otherwise are not easily burned, and to initiate the switch from forest to nonforest. Thus, in the context of climate warming, altered fire regimes and the positive feedbacks from pyrophilic vegetation become the mechanisms for possible large-scale transitions from forest to shrublands or grasslands. Such transitions
often termed “tipping points” because of the accelerated rate of vegetation change in response to climate change. Much research is needed in disturbance ecology to improve our understanding of mechanisms responsible for switches to alternative stable states and their relationships with climate change.

**Invasive plants and animals**

Invasions by alien plants and animals are disturbances per se but may also have synergistic impacts through their alteration of disturbance regimes. In the example mentioned in the previous section, introduced herbivores had a fire-promoting effect through their selective influences on regenerating plant species, but introduced herbivores also often have a dampening effect on fire through removal of fine fuels. Introduced plants may have profound effects on species composition and functioning of native ecosystems by altering edaphic conditions and plant competitive interactions. They can have particularly dramatic effects on native ecosystems by altering fire regimes such as the promotion of fire by the invasive cheatgrass (*Bromus tectorum*) in semiarid environments in the Great Basin region of western North America. In the context of climate change, disturbance ecology research on the effects of climate alterations of disturbance regimes on the success of invasive species, and in turn their influences on disturbance regimes, will continue to be a research priority.

**Future research directions**

Although disturbance ecology has traditionally been dominated by field-based empirical studies conducted at scales ranging from tens to hundreds of hectares, the development of geospatial technologies and remote sensing in the final decades of the twentieth century has provided important opportunities to scale up the study of disturbance processes and effects to large landscapes and regions. Thus, datasets that can be verified through careful field data collection are now available for parameters such as wildfire severity and tree mortality at regional or subcontinental scales. Merging of approaches based on observations across a range of spatial scales is rapidly improving our understanding of the broader scope impacts of disturbances such as fire and beetle kill. In the context of climate change, simulation modeling of climate impacts on future vegetation patterns, and ecosystem function, has long been based on research results on the physiological responses of individual plants (and to a lesser extent of small experimental ecosystems) to altered climate and atmospheric gas concentrations. Currently, however, the rapidly improving capacity to observe and monitor disturbance processes such as fire and tree mortality is allowing the integration of disturbance ecology into regional and global modeling of the effects of climate change on regional and global ecosystem function and the feedbacks to the atmosphere. Research on the potential feedbacks of fire and tree mortality on the atmosphere through carbon dynamics is clearly one of the frontiers in disturbance research (Turner 2010).

Climate scientists agree that the anthropogenic release of greenhouse gases has been and will continue to be the major driver of warming temperatures on Earth for many decades. However, research is needed on the ecological consequences of climate change and to improve understanding of how ecological processes may result in feedbacks that could further accelerate future warming. For example, increased wildfire activity is recognized as a likely consequence of warming, with a potential positive feedback to the atmosphere through the release of carbon dioxide from burning vegetation. To address the
actual impacts of future wildfire activity on the atmosphere and future climate requires an interdisciplinary approach in which disturbance ecology must play a major role. Research in disturbance ecology needs to be integrated with studies of Earth surface processes and studies of land-use change in a broad earth-system science framework. Such a framework views the biosphere and atmosphere as a coupled interacting system in which fire plays a major role. The long history of research on the controls and ecological consequences of wildfires conducted in the context of disturbance ecology and land management provides a strong foundation for addressing the challenge of integrating research on wildfire activity into an earth-system framework.

Large-scale biogeographic responses to climate change, including disturbances such as fire and interactions with land-use practices, incorporate feedback effects of land-cover changes to atmospheric processes through changes in albedo, evapotranspiration, and roughness. Clearly, research in disturbance ecology on the effects of fire and tree mortality (either from insect outbreaks or from the direct effects of warming and drought) is of primary importance in understanding mechanisms that affect regional- and global-scale carbon dynamics. Such research is of vital importance in addressing the efficacy of land-use policies aimed at using carbon sequestration by vegetation to mitigate the effects of burning fossil fuels.

Multiscale research in disturbance ecology is needed on the differential susceptibility of various vegetation types to wildfire and on postfire transitions from one cover type to a different cover type (e.g., from closed canopy forests to shrublands or grasslands), which in turn may further enhance fire potential under a warmer climate. At a global scale, the area burned annually is estimated to have increased in the second half of the twentieth century, but trends are highly variable regionally. Recent increases in wildfire activity in some regions have been linked to changes in climate and land use, operating independently or in combination. However, warming does not necessarily result in increased fire activity in all biomes because decreases in moisture availability can reduce the amounts of biomass for burning in already low-biomass ecosystems. Land-use practices have been shown to have significant influences on wildfire activity at a global scale, based on strong associations of fire activity with human population densities and socioeconomic variables (e.g., gross domestic product per capita and ratio of crop cover to other land covers). However, the relationships of these coarse-scale indicators of socioeconomic variables to wildfire activity need to be more thoroughly examined at regional and local scales to identify actual mechanistic explanations of these relationships.

The development of useful regional- and global-scale dynamic models of future fire-driven vegetation patterns requires a mechanistic understanding of fire determinants such as the amount and flammability of biomass to burn, fire weather, and sources and probability of ignition. At a landscape scale, feedbacks such as time since last fire and land-use impacts on fuels have been demonstrated to be critical to future landscape heterogeneity and the associated potential for fire. Thus, research in disturbance ecology, integrated with studies of land-cover change and land surface–atmosphere interactions will continue to play a major role in the development of regional, and ultimately global, dynamic vegetation models.

SEE ALSO: Biodiversity; Biogeography: history; Biogeomorphology; Climate change and biogeography; Dendroclimatology; Dendrogeomorphology; Ecosystem; Ecosystem services
References


Further reading

Domestic spaces

Irene Cieraad
Delft University of Technology, Netherlands

The term “domestic space” may not have existed before the 1970s, but its function as a household’s living space within an enclosed structure of a house or hut existed for thousands of years, as archaeologists have lavishly demonstrated in their maps of excavated prehistoric settlements. Material remains serve as testimonial to possible functions of the spaces, as traces of a fireplace might indicate the main living space. In this context of the emergence of “domestic space” as a new concept in the 1970s, it also must be noted that the word “space” was not used before the late nineteenth century.

Although archaeologists used the concept “domestic sites,” it was the British social anthropologist Mary Douglas (1972) who introduced the concept of “domestic space” to an audience of archaeologists in 1970. She cautiously presented the anthropological and symbolic interpretation of tribal settlements and the tribal use of domestic space as a possible tool for understanding the symbolic meaning of spatial arrangements in excavated settlements. Her examples illustrated, however, that spatial arrangements and material traces as such were insufficient clues for anthropologists as they depended for their symbolic interpretation on people’s behavior and practices. In other words, she was rather pessimistic about a possible archaeological endeavor of reading the meaning of domestic space from material remains alone.

Despite Douglas’s pessimism, the concept of domestic space was more readily adopted by archaeologists than by fellow anthropologists. With the adoption of the concept of domestic space in the writings of archaeologists, the discipline witnessed a shift of focus from the function to the meaning of spatial divisions in settlements. Nowadays, anthropological studies on symbolic interpretations of tribal spatial arrangements have become a source of inspiration for archaeologists, and so too have sociological studies of behavior in public and private space, providing the availability of written sources as in the case of classical antiquity (Nevett 2010).

In the late 1970s the concept of domestic space gained momentum in opposition to the more familiar concept of public space within the discourse of academic feminists, and contributed to the conceptual framework of the newly initiated women’s studies, which thrived on male/female opposition. Domestic space was interpreted not so much as the living space of a household but as a secluded female domain in which women took care of children and the household, while men spent much time in public space earning a living and socializing with other men. Although the male/female opposition was a common symbolic opposition in tribal spatial arrangements as described by anthropologists, the new domestic/public opposition gave the male/female opposition more of the character of a male conspiracy in secluding women within the home.

In crude feminist interpretations, capitalism was to blame for the economic separation of the domains of consumption and production into separate domains of living and working. The exclusion of women from the domain of
DOMESTIC SPACES

production started within the class of better-off merchants in the seventeenth and eighteenth centuries and gave rise to a spatial segregation of the secluded female home place in opposition to the public male work place. In the nineteenth century the separation was translated into emotional contrasts of the intimate and caring relationships within the domestic domain versus the anonymous and businesslike relationships within the public domain of industry and commerce. Privacy and domesticity became the much praised and near-sacred attributes of the domestic space of the nineteenth-century home where the woman and housewife had to guard and cherish this moral sanctuary from worldly evils.

Although this concept of domestic space did not exist in the seventeenth or eighteenth centuries, it was presented by feminist scholars as part and parcel of the so-called domestication of women in the merchant’s class. Decades of historical research on the position of women have corrected not only the beginning of the domestication process, but also the parallel image of the subjugated homebound housewife. Medieval sources revealed that prosperity was a main incentive for domesticating women as it allowed merchants’ families to imitate the lifestyle of the gentry and aristocracy, whose women led more secluded domestic lives and engaged in recreational needlework (rather than waged needlework).

Also, the defining characteristic of domestic space as a nonproductive domain is corrected in historic research on the productive aspects of housework and the many skills of housewives. Even domestic space as a female-dominated domain is contested by male scholars, like John Tosh (1999), who thinks that the gender contrast is overrated and that the Victorian home was as much a man’s as a woman’s place. As such, Tosh presents the beginning of the modern ideal of domestic space as home, portrayed as a family leisure space and a refuge from the obligations of work and school.

The domestication of women in the merchant class is illustrated by the changed architecture of merchants’ houses in late seventeenth-century Amsterdam, at that time one of the biggest cities of Western Europe. The access to the house changed from an open entrance at street level into raised doorsteps leading to a majestic, but shut, front door. Previously, housewives supervised the open entrance hall while monitoring the public space of the street, but the changed architecture of the merchant’s house mirrored a changed attitude of these upper-class women, who retreated from the public eye of the street into the back zones of the house. In most European cities, however, the domestication of upper-class women and children alike became evident in the nineteenth century (Cieraad 1999).

Twentieth-century historic research revealed that the domestic space of the nineteenth-century gentleman’s house was in fact an intricately zoned hierarchy of spaces, not only the superior domestic spaces of the family with different gradients of domesticity and privacy as opposed to the inferior service spaces of the domestic staff, but also differently decorated domestic spaces for male and female family members and visitors, for children and adults. As such, the plural of “domestic spaces” seems more adequate than the singular.

The more recent use of the plural bears witness to a shift of focus in today’s scholarly literature on domestic space. Initially, the boundary between the domestic space of the house and the public space of the street was of prime importance, especially to feminist scholars. Also, in the traditional anthropological and symbolic interpretation of Mary Douglas, the threshold as the border between the inside and the outside of the house represents an important ritual boundary, which is still obeyed in the old custom of the
groom carrying the bride into the house. With the more recent shift of scholarly attention toward the internal zoning of domestic space according to use, gender, and age, not only has the threshold as the border between inside and outside lost its importance as a ritual boundary but so too have the walls of the house as the confines of the domestic domain.

The domain of present-day domestic spaces most likely extends into the domestic garden, as the popularity of the outdoor kitchen illustrates, or onto the roof, as the popularity of the roof terrace illustrates. Even seasonal influences on these outdoor spaces are countered by all-weather furniture and domestic comforts like fireplaces, heating devices, and awnings. In today’s interpretation of domestic spaces, the traditional interiority of domestic space is not defining anymore, but its domestic usage, in particular of these outdoor spaces, is. Where the garden or the roof terrace is used for food preparation, dining, relaxing, or receiving guests, it becomes one of the many domestic spaces in and around the house. More than anything, the short scholarly history of domestic space illustrates human geography’s concern “that space and place are more than just containers of social processes; they are the social processes” (Short 1999: x).

SEE ALSO: Borders, boundaries, and borderlands; Built environments; Feminist geography; Gender; Home; Privacy, personal privacy; Public space

References


Domestic workers

Rachel Silvey
University of Toronto, Canada

Domestic workers are paid to carry out household labor. They do paid child care, elder care, house cleaning, laundry, and often cooking and pet care in the homes of families that are not their own. In human geography, feminist geographers spearheaded research interest in domestic work, and the body of scholarship on the topic has burgeoned in recent decades within the discipline, as well as in neighboring fields. The now large body of research on domestic workers is a reflection of the heightened importance accorded to feminist approaches in geography in general, and an indication of substantive shifts in the global organization of domestic work.

Historically and internationally, paid domestic work was and continues to be almost entirely a women’s occupation, cross-cut by class relations, rural/urban divisions, and racial and ethnic distinctions. The numbers of domestic workers have grown with each phase of global capitalism. Geographers have focused primarily on the period after 1970, when the international migration of domestic workers began to grow considerably, and immigrant women began filling the jobs of nannie and maids in rising numbers. Many of the jobs that immigrant domestic workers fill have been newly created in recent decades as a result of expanding middle-class populations and lifestyles, and changing gendered global divisions of labor and their reverberations in the domestic sphere. Research on domestic workers has shed light on the ways in which nationality and citizenship, as well as new and long-standing racialized hierarchies, shape who does the “dirty” work and for whom.

Studies of domestic workers began to gain traction in the discipline in the 1990s when geography was also beginning to register engagements with poststructural, feminist, and postcolonial theories. They have opened up questions about the politics of methodologies and embodiment, the diversity of care migration and social reproduction, the social construction of space and scale, and chains of care work.

While most immigrant domestic workers are now paid for the labor they perform for others, many work under conditions of indentured servitude, and the majority earn incomes at the bottom of the wage scale in their respective countries of work (Romero, Preston, and Giles 2014). They are often faced with the garnishing of their wages, and high levels of debt are common. Indebtedness intensifies dependence on employers and heightens vulnerability to excessive workloads (sometimes reaching 18 hours a day), underpayment and nonpayment of wages, and a wide range of other abuses, including verbal and sexual harassment, violence, and rape (e.g., Yeoh, Huang, and Gonzalez 1999).

Domestic workers have fewer rights than most other workers in most countries. Because their work is carried out in the “private” space of the home, governments have commonly considered it beyond the reach of the law. In locations where laws are in place to protect domestic workers’ rights, such as the United States and the European Union, in practice workers are often unaware of these rights or are afraid to claim them, given their status as contract workers or their liminal citizenship status. Domestic
DOMESTIC WORKERS

workers’ rights intersect with their experiences of civic space, transnational legal abandonment, the micropolitics of households, and national and urban sociolinguistic and racialized hierarchies.

The dividing line between the roles of paid domestic worker and that of unpaid household worker and wife are not always clear-cut. Indeed, some domestic workers seek to marry citizens in their host countries as a pathway to citizenship and to escape repeated short-term labor contracts abroad. In some cases domestic workers who marry local citizens play important roles in remaking their new local communities’ identities and social fields (Faier 2009). Furthermore, domestic workers often also work in other sectors while they are employed as domestics, engaging in various supplemental formal and informal sector income-earning activities. They are frequently misrecognized as victims of human trafficking, and antitrafficking policies have criminalized domestics working without legal documentation.

Sending states play active roles in promoting and managing the overseas migration of domestic workers with an eye toward earning foreign exchange in the form of remittance income. In many parts of the world, migrant domestic workers are required to live in the homes of their employers, an arrangement that contributes to their social isolation and exclusion from the public sphere. Control of the living and working space of domestic workers is conjoined with the control of their bodies. Gender ideologies of sexuality and proper womanhood, religious piety, and place-specific norms about gendered family roles have also shaped the migration and working conditions of domestic workers. They are taught to embody a subservient, desexualized manner by labor brokers who recruit and train them, as well as by their local communities and overseas employers.

The formal and informal conditions under which domestic workers labor differ considerably across countries and remain highly dependent on the whims of individual employers. Nonetheless, the globalization of domestic work has also sparked numerous progressive, creative, and increasingly successful movements for the rights of domestic workers. Meanwhile, scholarship has challenged the representational reduction of domestic workers to either heroes or victims (Gibson, Law, and McKay 2001), and deepened its analyses of domestic workers’ emotions, desires, religions, and sexualities (Silvey 2007).

Domestic work is intimate labor, and when it is carried out by immigrant women it interweaves global, local, and embodied processes. Each of these scales of analysis is shaped by the gendered material and discursive “forces of domesticity,” which persistently position women in the cleaning and caring roles of households, nations, and the global economy. Migrant domestic workers’ contracts tend to require them to leave their children at home, and their sojourns thus usually involve long family separations. Though these patterns take highly diversified forms across specific places, women family members such as sisters or grandmothers, or lower-income women neighbors in the origin countries, pick up the residual local domestic and caretaking roles, and are usually paid low wages drawn from remittances income, which is also expected to cover the everyday needs and education costs of the families left behind. These transfers have been termed “global chains of social reproduction” (Parreñas 2015), such that the labor of migrant women, and the associated intensified labor of women left behind, produce a direct subsidy at low cost to the social reproduction of households in higher-income countries. A recent debate has emphasized the conceptual distinctions between care work and social reproductive work, and
underscored reasons for privileging the latter (Raghuram 2012).

Domestic workers who leave children in their home countries tend to practice transnational mothering while overseas, but their long absences from home are often traumatic for family relationships over the life course (Pratt 2012). Moreover, the experience of returning “home” tends to be fraught, in practical and emotional terms, for both migrant domestics and their families (Yeoh and Huang 2000). Research has only begun to examine the ways that overseas domestic workers’ experiences play out over the long term, and attention is beginning to address the experiences and needs of aging and elderly current and former migrant workers, many of whom work in home care for the elderly (Raghuram 2012).

The political horizons of the future of global domestic work look promising from some angles. In 2011 the United Nations International Labour Organization (ILO) passed the Domestic Workers Convention (C189), which covers a range of basic human and labor rights specifically for domestic workers, and, as of early 2016, 22 countries had ratified the convention. This success was the result of decades of organizing on the part of domestic workers worldwide, and alongside its passage several countries and states have passed their own new laws creating new rights for migrants (e.g., a minimum wage for overseas domestics from the Philippines; a requisite day off for domestics working in Singapore; and a Domestic Workers Bill of Rights in four US states covering overtime pay and a weekly day of rest, among other protections). Domestic workers’ rights advocates have taken courage from these legal changes, fueling the proliferation of organizations in support of domestics (including, for example, national alliances in many countries, regional centers, and the International Federation of Domestic Workers).

In addition to these political gains, research has argued that, while domestic work is work and while domestic workers deserve rights, recognition, and respect for their work, domestic workers are more than workers. They are also family members, artists, and lovers, and often in addition to their work as domestics many have also had careers as nurses, entertainers, office professionals, and teachers. They and their allies have created artistic, poetic, and theatrical performances that enrich and deepen the complexity of their experiences in a way that traditional social scientific methodologies have not been able to capture. As research continues to deepen its appreciation for the wide range of domestic workers’ experiences, geographers can continue to contribute to the exciting theoretical and empirical work tying together the geographies of their lives. Studies of domestic workers have been at the forefront of collaborative methodological innovation with activists and artists, and such work is vital for continuing to understand the relational production of not just public and private spaces, and productive and reproductive labor spheres, but also the dynamic interlinkages between the intimate and the global.

SEE ALSO: Care work; Emotional labor; Gender, work, and employment; Labor migration; Migrant division of labor; Migrant labor; Precarious work; Social reproduction

References


DOMESTIC WORKERS


Droughts or water shortages are repetitive natural hazards created by lower than expected moisture levels at a given location. Drought and water shortages occur across the globe but can only be clearly identified and addressed within a regional and impact specific context. The core concept of a drought or water shortage definition is the physical condition of below “normal” moisture levels. Such moisture levels may be, but are not limited to, average monthly precipitation, average annual stream flow, or mean seasonal soil moisture. The moisture levels represent the amount of water available at a given time in the hydrologic cycle which includes all states of water (liquid, vapor, ice) at a given time and place. The hydrologic cycle includes evaporation and transpiration from the Earth’s surface, condensation to form clouds, precipitation followed by runoff, and movement or storage of water underneath the Earth’s surface. Water shortages associated with drought are identified by tracing the paths that water takes through the hydrologic cycle, what the water is doing at various stages along each path, and how the quantity and characteristics of water are altered as they pass through the cycle.

Critical to the analysis which leads to the identification of a water shortage (or a surplus of water) is the concept of a water budget. In this approach, the net available water at any given location is represented as a function of gain from precipitation, loss by evapotranspiration, loss to or gain from soil moisture, and addition or subtraction by other physical (such as groundwater flow, stream flow) or human (such as irrigation, drinking water withdrawal) processes. Commonly, a water budget is expressed by the equation.

\[ W = P - E_v T \pm S_m \pm P_{misc} \pm H_{misc} \]  

Each variable in the equation offers an account or “budget” of water as it travels through the hydrologic cycle within a local landscape. A surplus occurs when the input of water from precipitation, soil moisture, or other processes is greater than any withdrawal of water from the system. A shortage is when all withdrawals are greater than the inputs. It should be noted that a shortage or surplus is rarely created by a single human process or a single physical process. Rather, most problems are created by a combination of physical and human processes. For instance, an agricultural water shortage may be created by both low precipitation and inefficient water delivery systems. The key in identifying a drought or water shortage is to complete a water budget analysis.
DROUGHTS AND WATER SHORTAGES

that incorporates all processes which account for
the storage and movement of water at a given
location.

In combination with the core concept of low
moisture conditions, the definition of a drought
also includes a link to the supply and demand
of an economic good, service, or other human
activity. Such goods, services, or activities are
dependent on water in one of its three states for
production, delivery, or completion. Some of the
more direct examples of these goods, services,
and activities include agricultural crops that
require water for growth or river water used for
municipal water supply. Examples of less direct or
less obvious goods, services, or activities depen-
dent on water are when livestock in one area
are impacted by drought in another area because
not enough feed can be grown due to dry
conditions and decreased supply of goods due to
the inability of barges to navigate rivers because
of low water levels in the navigation channel.

As one attempts to identify a drought or water
shortage, it is very important not to confuse a
drought or water shortage with arid conditions.
Arid conditions exist at a location or area that
consistently has very low rain fall totals and high
evaporation rates. The two most common forms
of arid conditions on the Earth’s surface are the
steppe and desert. Such an arid location is Alice
Springs in central Australia which only receives
on average 252 mm (about 10 in.) of rain a year.
Accordingly, a water shortage in Alice Springs
represents a deviation from an expected rainfall
or moisture level pattern or, in the case of Alice
Springs, when a year’s rainfall is lower than
252 mm.

General types of drought

Once a drought has been identified it can be
classified as one of four types: meteorological
drought, agricultural drought, hydrologic
drought, or socioeconomic drought (Wilhite
2000). Meteorological drought refers to precipi-
tation deficiencies alone and does not include a
link to a good, service, or activity. An example
of a meteorological drought is below aver-
age precipitation for the summer at a specific
location in a given year. Typically meteorolog-
ical drought serves as the physical framework
within which impacts upon human activities are
identified. Agricultural drought refers to preci-
pitation and/or soil moisture deficiencies that
impact agricultural production. Such impacts
are not limited to crop production alone, such
as drought in the midwestern United States
that reduces soybean production, but can also
include drought impact on livestock through
less drinking water available and less water for
production of feed crops. A classic example of
agricultural drought was the Dust Bowl which
was an extreme drought that impacted the Great
Plains region of the United States for consecutive
years in the 1930s. During this event, wheat
and corn yields plunged by 50% to as much as
75%, and millions of tons of topsoil were lost as
desiccated soils were eroded by wind (Schneider
and Londer 1984). It should be noted that
tilling, cropping, and other agricultural practices
contributed significantly to this drought event
and associated soil erosion, highlighting the need
to combine both physical or social conditions
and processes to understand the true nature of a
drought event.

Hydrologic drought represents a shortage in
surface or subsurface water supply that impacts
a good, service, or other human activity. A
common form of hydrologic drought is the dry
conditions which lead to reduced urban water
supply from rivers or reservoirs. For example,
beginning in the winter of 2005–2006 through
summer 2008, much of the southeastern United
States experienced a significant rainfall deficit
DROUGHTS AND WATER SHORTAGES

which led to an extreme drought event (Manuel 2008). Due to this drought event, the region experienced widespread restrictions on water use, dangerous drops in reservoir levels, damages to crops, and conflicts within and between states on access and use of scarce water resources. Low reservoir and lake levels forced power companies such as the Tennessee Valley Authority and Duke Power to reduce hydroelectric generation and temporarily replace it with more expensive fossil fuel generation. Water levels of Lake Lanier, GA, source of drinking water for 3 million people in Atlanta, were reported to be as low as 5 m below normal.

Socioeconomic drought represents all other forms of drought in which a water or moisture shortage impacts supply, demand, production, or occurrence of some good, service, or other human activity. Given the broad definition, this type of drought includes the widest variety of events. Examples of this drought type are poor fish catch due to high salinity from low freshwater stream flow into an estuary or reluctance of tourists to visit a resort community due to an unpleasant odor produced by dead fish as a lake lowers due to drought. Socioeconomic droughts usually have the widest impact in terms of the sectors of society involved and geographic area impacted. It has been suggested that this type of drought is linked to the collapse of civilizations, such as the Classic Maya collapse in 800–1000 CE (Gill 2000). This megadrought event contributed to the death of millions of Mayans and played a role, along with internal strife, in the collapse of the empire (Dunning, Beach, and Luzzadder-Beach 2012).

It should be noted that these four drought types are not mutually exclusive and frequently overlap, which limits their use beyond organizational or conceptual guides. For instance, it is possible to identify a drought which represents both a meteorological drought and an agricultural drought; lower than normal rainfall at a specific location that impacts maize production. It is also possible to have an event which represents a combination of more than two of the definitions. A combination of meteorological, hydrologic, and socioeconomic droughts can exist when low rainfall causes low stream flow that reduces rafting trips down a river, and the inability of barges to use the river due to too shallow a navigation channel. Such overlap between the four types underscores the complexity of drought as a natural hazard.

The study of drought

The importance and relevance of drought to society is drawn from its ubiquitous and insidious nature (Tannehill 1947). Drought is ubiquitous in that it occurs across the globe, in water “rich” and water “poor” areas. Deserts or arid areas can experience droughts as well as rainforests, as illustrated with the example of Alice Springs. The insidious nature of drought is evident in that it is a slowly developing phenomenon that does not have a discrete beginning or ending point and most people do not notice or react to drought conditions until they are in the middle of a water shortage and faced with its greatest impact. Given such characteristics, drought is considered the most complex and most difficult natural hazard to effectively address. Further, drought can impact a very wide geographic area in contrast to other natural hazards which typically have a relatively small area of impact (i.e., tornados or landslides). The Dust Bowl, referred to in terms of agricultural drought, impacted an area greater than 900,000 km² from Nebraska to Texas. Given the existence of drought over a wide variety of landscapes, the difficulty in identifying drought event onset, repeating occurrence, and the potential to impact a wide area, drought has received much attention from
DROUGHTS AND WATER SHORTAGES

academic, governmental, and nongovernmental entities. As a result, a wide variety of literature exists in the fields of meteorology, climate science, agronomy, geography, and natural hazards that pertain to drought as well as volumes of government documents concerning drought policy which attempts to plan for, respond to, and mitigate drought. Within this large body of drought research and literature four main themes or approaches can be found: (i) physical science approach to drought; (ii) social science approach to drought; (iii) holistic human–environment approach to drought; and (iv) drought policy.

The physical science approach to the study of drought focuses on the specific environmental components and mechanisms through which drought develops, unfolds, and recedes. Examples of these components and physical mechanisms include, but are not limited to, precipitation, soil moisture, stream and groundwater hydrology, plant physiology, cropping techniques, and irrigation systems. Common themes in the physical approach to the study of drought include identification of the environmental components and processes which allow for prediction and identification of the onset of drought, monitoring drought conditions, engineering and management techniques for drought prevention, and engineering and management techniques to mitigate the impacts of drought.

The social approach to understanding droughts focuses on how humans contribute to the existence of drought and the impacts of drought upon society. Some common social science topics in the research of droughts include community drought preparedness and mitigation, social adaptation to drought, economic dimensions of drought, prediction and management of societal water demand, community and individual vulnerability to drought, drought risk management, and transboundary watershed management.

The holistic human–environment approach to drought emphasizes analysis of how physical and social processes, entities, and agencies interact and combine to produce drought. In this approach to the study of drought, researchers do not follow an atomistic approach in that physical and environmental aspects of drought are studied separately from the social aspects of drought. Rather, an attempt is made to combine the two components (physical and social) together to understand drought. Examples of such analyses include the study of desertification, drought within the context of global change, and drought as a natural hazard.

The final approach to the study of drought is drought policy which focuses on how government agencies contribute to drought management, response, adaptation, and mitigation. Drought response represents action taken to diminish negative impacts once a drought has started. Drought adaptation refers to the development of a course of action through which society or individuals adapt to the occurrence of drought before, after, and during a drought event. Drought mitigation refers to government action to stop or hinder the occurrence of drought. Given this wide array of approaches to studying drought, it is obvious that the study of drought is multidisciplinary. However, the study of drought’s physical science aspects has a tendency to fall within the disciplines of meteorology, climate science, geology, geography, ecology, hydrology, plant science, agronomy, and engineering. The social, human–environment, and policy approaches to the study of drought typically fall within the disciplines of geography, economics, sociology, political science, planning, public administration, and environmental studies.

Geography offers a unique perspective to the study of drought, a perspective that focuses on the spatial components and scale of drought.
Such a perspective is needed because supply and demand of water are not in equilibrium with geographic space and time (Dzurik 2003). Within the context of drought, this equates to the demand for water being greater than what is available at a specific place and/or time. Geographic analysis of drought is a problem-solving approach that explicitly recognizes the role of spatial characteristics and interactions of entities across the Earth’s surface. Hydrologic features (watersheds and aquifers) and social structures (water management districts and varied stakeholders such as landowners, environmental activists, and so on) manifest themselves differently in different geographic locations.

In order to conceptualize the differences of drought across geographic space, geographers find it useful to employ the concept of scale. Scale is a concept that allows geographers to think of water moving across local, regional, national, and international levels through physical and human systems which operate in a variety of interrelated geographical scales from the individual to international and global politics (Swyngedouw 2004). Such a characterization illustrates the importance of scale to water resources and drought. Scale can refer to the areal extent of an entity (e.g., a county within a state) or to the entity’s area of influence (e.g., a company that operates at local, regional, national, or international scales). Droughts unfold within and across these scales with each of the component environmental systems and social entities existing at one of these scales or across multiple scales. For instance, the watershed of the Jordan River is shared by Jordan, Israel, Palestine, Syria, and Lebanon. Consequently, drought management in this region unfolds across all geographic scales: local towns, state or department administrative units, and across international boundaries. Such a situation creates a very complex geography of water resources. Analyzing the political and environmental relationships across the scales is essential to understanding and reconciling competing perspectives in drought management.

The existence of multiple-scale components in a drought issue means that water resource managers inevitably are dealing with interrelated geographic scales. Typically, interrelated geographic scales create poorly defined boundaries of stakeholder influence, which can in part lead to conflict between stakeholders, such as local versus international or regional versus national interests. In the case of the Jordan River system, a local versus international mismatch may represent stakeholders in the city of Jerusalem versus stakeholders in Syria; or stakeholders in the West Bank versus the stakeholders of Israel. Without the identification and understanding of interrelated geographic scales, managers will experience little success in addressing drought management issues.

Current and future study of drought

A review of the history of the research on drought indicates that it has consistently evolved over time through advances in associated disciplines and adaptation of new technology by stakeholders. Despite such advancements, one of the most consistent themes in the study of drought has been the struggle to craft a universal definition of drought (as evident in the overlap of the four general drought types). For years, scholars, government officials, and stakeholders have attempted but still have to develop an accepted universal definition for drought (Wilhite 2000). Such debate has led many drought experts to conclude that a universal, precise definition of drought is unattainable and only a vague, generic conceptual definition can be provided that applies to all drought occurrences. Consequently, experts indicate the only appropriate approach
DROUGHTS AND WATER SHORTAGES

to defining drought is from an operational standpoint within a regional and use-specific context. Such a regional, use-specific operational definition allows for the identification of precise drought characteristics and specific threshold values that define the onset, continuation, and termination of drought episodes, as well as their severity. Such an approach takes into account the climatology and typical hydrologic conditions at a local or regional scale (e.g., county, state) and examines how this local or regional setting impacts a specific good, service, or activity (e.g., maize production or tourism). The characteristics used to identify a drought event can be a meteorological value such as soil moisture and an agricultural variable such as the amount of crop lost to drought. The thresholds are specific values for each of these two variables which indicate onset, continuation, termination, and severity of a drought. For instance, onset of a drought event may be identified as when soil moisture falls below 20% water content and at least 30% of crop planted are lost due to a lack of soil moisture. A severe drought is then defined as soil moisture below 15% water content and loss of crop to dry conditions greater than 50%.

Outside of the fundamental question of defining drought, several other themes have recently received focused attention in the current study of drought. With the development of modern ecology and environmental sciences in the 1970s, a greater emphasis has been placed on the human–environment approach to drought analysis, stressing the recognition that neither a physical nor a social component alone can be used to fully understand drought. In this approach to studying drought, researchers attempt to combine physical and social aspects/components of drought from the beginning of analysis and then determine how they interact or coevolve to produce a drought event. A current example of a holistic human–environment approach to drought is the concept of double exposure and vulnerability to drought. Double exposure refers to community or individual exposure to both economic change and environmental change which combines to create increased vulnerability or the susceptibility to experience a negative impact due to drought (Leichenko and O’Brien 2008). For instance, a farmer may receive a low price for tomatoes harvested due to a more competitive global food market. As a result, the farmer is unable to afford installation or repair of an irrigation system on their farm. This loss of financial capital for irrigation due to a change in the global food market combines with drier conditions created by global climate change, resulting in greater farmer vulnerability to drought. Thus, double exposure to both economic and climate change creates vulnerability to drought, a vulnerability that cannot be completely understood without investigating the combination of both the physical and social aspects of economic and climate change over time.

Another current focus in drought studies is the application of new satellite data and numeric modeling technology to improve the forecast and identification of drought and drought monitoring methodologies (Hayes et al. 2005). Meteorologists have continued to develop and improve operational monthly and seasonal predictions of drought based upon empirically derived statistical relationships from satellite data and numerical ocean atmosphere coupled computer models that identify teleconnections between moisture conditions in one location and global scale atmospheric dynamics and processes such as variability in sea surface temperatures, ENSO events, jet stream variability, and development and absence of long-term wind and storm systems (Ropelewski and Folland 2000). These techniques have led to recognition that the absence of moisture advection, increased
DROUGHTS AND WATER SHORTAGES

Evapotranspiration, and increased regional atmospheric subsidence are the major physical contributors to the occurrence of drought.

A major component of the recent efforts to improve forecasting and monitoring of droughts is directed toward the development of effective drought indices (Keyantash and Dracup 2002). A drought index is a numeric or alphanumeric expression that clearly and efficiently communicates drought conditions at a given location based upon an aggregation of in situ observations, statistical comparisons, model assessment, and expert analysis. A successful drought index is useful (robust), data compatible (tractable), clear (transparent), applicable across time (extendible), and uses fundamental units or a ratio (dimensions) (Keyantash and Dracup 2002). Some of the most common drought indices used by drought experts are rainfall decile, standardized precipitation index (SPI), total water deficit (TWD), Palmer hydrologic drought index (PHDI), and the crop moisture index (CMI). The US Drought Monitor, the official monitoring product of the US National Drought Mitigation Center, displays drought conditions across the United States in the form of a drought index ranging from D0 (abnormally dry) to D5 (exceptional drought) as well as the designation of S for short-term water shortage (less than 6 months) and L for long-term water shortage (greater than 6 months) (Svoboda et al. 2002). The D0–D5 alphanumerical value is derived from six key physical indicators and expert evaluation and was designed to be easily recognized by the public as inspired by Fujita tornado scale and the Saffir-Simpson hurricane scale.

Linked to this current theme of use of new technologies and data to predict and monitor drought is the application of such techniques to determine the potential for increased drought and water shortage in a future anthropogenically warmed climate. The Intergovernmental Panel on Climate Change (IPCC 2013) states that the warming of the global climate system is unequivocal with a linear trend of 0.85°C (0.65 to 1.06°C) in combined land and ocean surface temperature for the period 1880 to 2012. Corresponding to this warming trend is an analysis of drought (using the PDSI drought index) that indicates very dry areas have more than doubled since the 1970s, particularly in the tropics and subtropics, due to surface warming (Dai, Trenberth, and Qian 2004). In the future, predictions indicate this drying trend is likely to continue in mid-latitude and subtropical dry regions while high latitudes, the equatorial Pacific Ocean, and mid-latitude wet regions will likely become wetter. The result is a future climate in which contrasts in amounts of precipitation in wet and dry regions will increase causing greater potential for drought in already dry areas alongside less probability for drought in wet areas. Further the changes in precipitation and melting of snow will alter regional hydrologic systems which in turn may cause regional water shortages or drought. Some of the key risks associated with the increase of water shortage and drought and changes in regional hydrologic systems are the breakdown of infrastructure networks and critical services associated with water supplies, food insecurity caused by the breakdown of food systems, and loss of rural livelihoods due to insufficient access to drinking and irrigation water and reduced agricultural productivity (IPCC 2014).

Conclusion

In conclusion, drought is a ubiquitous and insidious natural hazard that can potentially impact any location worldwide. Given the unique nature of drought at any given location, it is very
DROUGHTS AND WATER SHORTAGES

difficult to develop a universal definition and prescription for drought. Each drought event must be evaluated within its regional geographic and human activity context. Drought events can be classified as one of four types: meteorological drought, agricultural drought, hydrologic drought, or socioeconomic drought. Over the years the study of drought has followed four main themes or approaches: (i) physical science approach to drought; (ii) social science approach to drought; (iii) holistic human–environment approach to drought; and (iv) drought policy. Currently, drought experts are utilizing new technology to monitor and to improve forecasts and predictions of drought. These improvements will be vital to address the greater potential for drought in mid-latitude and tropical dry regions due to anthropogenic climate change. Geographers are poised to contribute to these efforts by using a unique perspective that stresses the spatial components and scale of drought.

SEE ALSO: Arid climates and desertification; Climate change, concept of; Climate and societal impacts; Hydroclimatology and hydrometeorology; Hydrologic cycle; Natural hazards and disasters; Scale; Water budget; Water resources and hydrological management

References


Oxford: Blackwell.

Wilhite, D.A. 2000. “Drought as a Natural Hazard:
Concepts and Definitions.” In Drought: A Global
Assessment, edited by D.A. Wilhite, 3–18. London:
Routledge.
Daily mobility

Sandrine Wenglenski
Université Paris-Est, France

A matter in progress

Mobility is not a straightforward concept. In its common usage, it simultaneously describes a potential for action (the capacity to move) and a real action (the actual movement). In its academic usage, it covers several dimensions (social, spatial, symbolic, etc.) and does not only refer to physical movement: its early use applied to social mobility, which relates to a change in position at a social level. Even when mobility specifically addresses geographical concerns, it considers various entities (individuals, goods, data, ideas, etc.) and fields (international, residential, daily, virtual movements, etc.). Gradually, the concept has been adopted in many studies, expressing the perception of an increasingly changing world, sometimes referred to as a “mobility turn.”

At first sight, daily mobility is more straightforward. It is commonly agreed that it relates to the day-to-day physical movements of people. However, this simple statement leaves areas of uncertainty: what are day-to-day actions? Who are the people involved? Should not virtual movements (distance communication) be taken into account in view of their possible interactions with physical movements? What about the links with social or other types of mobility? How research answers these questions depends on the environment and concerns of the day. The initial study of daily travels was prompted by the need to solve practical and technical problems related to increasing urban traffic and congestion, in which recurrent trips such as commuting to work played a key role. Accordingly, the travel survey protocols that were developed considered mobility as the outdoor physical movements of local residents pursuing their everyday life agenda, in the sense of activities carried out fairly close to individuals’ homes as opposed to long-distance journeys. These surveys still provide the basis for the statistical knowledge of daily mobility. Although the mobility of tourists, freight, and information, as well as long-distance travel, is acknowledged to be partly linked to that of residents, currently no common framework binds them all.

Yet, from the time it was first studied on a regular basis, daily mobility – as well as the physical and social environments within which we live – has changed substantially. The trends and changes involved have been a key field of observation and a means of gaining a better understanding of mobility, thanks to input from a variety of disciplines and methods. Such input has participated in the continuous questioning of the scope, implications, and meanings of daily mobility. Daily mobility appears to be linked to many dimensions of social life and social issues. Depending on the social, economic, and technological context, it both derives from ways of life and spatial forms and also influences them.

Mobility as a by-product

With the exception of travel for the sake of pleasure, traveling arises from a forthcoming
DAILY MOBILITY

activity that cannot be performed on the spot. We need to travel in order to buy food we do not have at home, even though our primary desire is not to travel but to feed ourselves. This is why daily mobility is considered to be a by-product; it is a “demand” which is derived from purposes other than itself. The study of travel was initiated by engineers responsible for forecasting transport demand so as to adjust infrastructures to growing travel needs. Consequently, thinking about mobility was largely inspired by the figure of homo economicus. Analysis focused on single trips, perceived as the derived demand of the need to achieve one primary activity – for example, working, consuming, entertaining – regardless of the number of an individual’s trips during a period of time or of the interplay of several purposes. One action has one reason. In addition, travel was regarded as a pure cost: a loss of money and time. As a result, all things being equal, it was assumed that people will attempt to minimize their travel costs and distances: the lower the cost the better.

This view was supported by the orders of magnitude observed in trip structures. It grounded the gravity model which expresses the amount and direction of flows between origins and destinations within (and between) cities as a function of the economic importance of areas (generally the number of job opportunities as a proxy for the potential of a location to provide destinations of activities) and the distance between them. The denser and nearer a zone, the greater the flows it attracts. Yet, it has hardly helped us to understand the variety of behaviors and the spectacular increase in distances traveled daily observed from the 1970s to the end of the 1990s in Western countries, and also apparent today in emerging countries. Unpredicted flows and increasing distances ran counter to the pervading theoretical conception of travel being strictly dependent on economic urban hierarchies and a willingness to minimize distance.

The fact is that all things never are equal. People are subject to various constraints and have many goals, not all of which can be achieved under the same conditions, and daily mobility is rarely the automatic consequence of the desire to perform either a single or a primary activity which would be the trip’s obvious purpose, and the nearer the better. For example, the mode of transport people use to travel to the location of their first activity of the day generally impacts how they travel to the second, and vice versa. Similarly, the characteristics of a person’s journey to work depend on the location not only of their job but also of their home. The home location choice may be one of their primary purposes in travel decision-making. Mobility encompasses many dimensions and factors and, depending on them, individuals may place uneven importance on close distance and minimizing travel costs.

Mobility as a tradeoff

Two major perspectives contributed to the development of a view that humanized and contextualized the protagonists of mobility: one analyzes the constraints that weigh on people; the other emphasizes their degree of choice.

Torsten Hägerstrand (1970) highlighted that people are limited by a wide range of circumstances. Because there are only 24 hours in each person's day, the number of distant activities (number of trips), the size of the area within which a person can carry out these activities in a day (length of trips), and the choice of their location (destination of trips) are determined by his/her characteristics and spatial and social environment. An individual’s daily mobility depends on his/her possible travel speed (relying
on his/her physical capacities, the economic budget he/she can afford to spend on a nonmuscular means of transport, the local transportation supply, and existing technologies); the spatial distribution of activities; the working or opening hours of companies, retail outlets, and public services; and the social conventions that allocate different roles and activity programs to members of the household and the community (by race, gender, or age). Hägerstrand’s work provided a more realistic picture of the physical and social experience of traveling, revealing factors and constraints that restrict mobility potentials and behaviors.

In contrast, Yacov Zahavi (Zahavi and Talvitie 1980) drew attention to the dimension of choice that lies behind mobility behaviors. Zahavi noticed that, across Northern and Southern countries of the world and at different periods, people appeared to spend the same average amount of time a day traveling. What distinguished travelers living in different geographical and technological contexts was not how much time they could spare but rather how far they could go. Counterintuitively, people who are able to use faster modes of transport do not seem to take advantage of them to spend less time traveling, but instead travel greater distances. This phenomenon challenged the traditional assumption that trip production is fundamentally determined by the minimization of travel costs. Within undoubtedly limited time and money budgets (on average one hour a day and 10% to 15% of the household income per month), people may extend their range of travel rather than reduce their budgets whenever this increases the choice of possible destinations and the chance of one of these better meeting their aspiration. All things not being equal, all individuals have a specific set of options within which they make tradeoffs and act, depending on what they value – for example, living in a larger house or a wealthier neighborhood, getting a higher-paying job, attending a better school – all of which affect their travel. This means that daily mobility is not just a pure cost derived from a single and immediate need. It can be seen as a means, the result of tradeoffs that are made in order to achieve personal goals in a constrained world.

The spatial footprint of social phenomena

This research paved the way for a great deal of work on the part of geographers, socioeconomists, sociologists, and planners that explored and deepened our understanding of the many determinants of daily mobility (Hanson and Giuliano 2004). Travel abilities and behaviors vary (Orfeuil 2004). Sociodemographic characteristics (age, gender, household composition, etc.) determine the physical autonomy of individuals and the social obligations weighing on their personal schedules. Social and economic positions (cultural and social patterns of exposure to others, monetary capacities) determine the freedom to break free of (or afford) proximity and to access diverse destinations. Spatial locations (with respect to the location of urban resources and transport provision) determine accessibility to opportunities and resources. These factors affect individuals’ potential mobility, which shapes mobility practices. Through these factors, the values and trends that pervade society at large affect the characteristics and changes of daily mobility. They influence the two-way relationship that links mobility to both lifestyles and places of living.

The sociologists of the Chicago School were among the first to analyze the changes in urban life in expanding cities, and to demonstrate how spatial mobility was involved in it. First, cultural and economic clusters in city districts
created the need to travel every day from one specialized place to another. Second, the progressive social and economic integration of new residents allowed them to move from the centers to distant, pleasant, neighborhoods from where they had to travel long distances. This meant that social reality is translated into spatial terms: social organization has effects on spatial interactions and organization. For example, upward social mobility often created spatial mobility. The increase in distances traveled daily, coupled with the constancy of travel budgets, proved to be a means of accessing ownership of single-family houses in low-density environments for car-owning households. Extensive daily mobility was the price for the achievement of residential aspirations in a wealthy economic period when car use and out-of-town housing were highly valued. In most cities worldwide, depending on the population’s level of wealth and the extent to which individuals have adopted the values of the consumer society, daily mobility sustained a way of life based on car and home ownership, and a new pattern of urban expansion leading to edgeless cities, where mass transit was nonexistent or worthless. The potential for daily mobility provided by the automobile opened up residential choice and released spatial configurations from the “tyranny of distance.”

The relationship between space and mobility patterns has stimulated an important line of research which investigates how urban form and daily travel influence each other – and whether and how to plan cities so as to reduce car use and traveled distances (Newman and Kenworthy 1989; Cervero and Ewing 2010). Not only have people who located far from city centers because of the ease of travel contributed to urban sprawl and are nowadays held responsible for growing congestion and pollution problems, but, since the beginning of what has been called the “white flight” in the United States, this move has raised concerns about segregation and access to jobs. The massive departure of the middle classes from inner areas to the outer suburbs (sometimes avoiding proximity to ethnic minorities) has had impacts on both residential specialization and the location of jobs. The inner-city neighborhoods vacated by the middle classes became impoverished and dominated by disadvantaged groups. Meanwhile, following the residents, and in view of their ability to commute by car, a great number of jobs relocated to outlying suburbs, where rents were lower. This geographic context created a “spatial mismatch” (Kain 1968), with the poorest residents from the dense neighborhoods being unable either to move closer to jobs or to travel the considerable distances needed to reach them. According to John Kain, this spatial mismatch was a cause of the higher unemployment rate experienced by the lower classes living in inner cities. Ultimately, one group’s ability to travel created a spatial context that worsened the consequences of the other group’s lower travel ability.

The evidence of social change – and permanence

Because it allows people to free themselves from the constraint of geographic closeness and affects spatial locations, social interactions, and the organization of everyday life, daily mobility has come to be blamed for weakening social ties. Easy travel is sometimes said to increase distances between people by fostering dispersed urbanization and the conversion of social distances into geographic distances, led by those being able to afford great travel distances. However, it is noticeable that one of the major spatial trends that dominated the second part of the twentieth century (urban sprawl) had moderated by the end of it. Living in core cities and being
able to benefit from urban amenities and to socialize locally came to be more and more highly valued. In many urban areas, the upper classes reconquered the central places, which became synonymous with quality of life and sustainable mobility. In city centers, mean daily traveling distances have recently decreased. This trend does not contradict Zahavi’s conjecture, since he stated that people prioritize their goals, of which minimizing travel costs may not be the most important, since shorter travel distances do not necessarily equal shorter travel times. Cycling from home to work is nowadays the prerogative of the happy few who live and work in town and are able to choose slow, environmentally friendly modes of transport. This has not curtailed the tendency for urban segregation: it occurs via other means. This suggests that daily mobility can serve various purposes and that its significant characteristics and relationships with space reflect deep social values and changes. On this ground, it has been said to be a “total social fact” (Bassand and Brulhardt 1980).

The social trends and ways of life that are embedded in daily mobility are obviously interconnected with the state of technology. In large cities, the spread of personal motorized vehicles and the development of efficient mass transit has combined with social and economic pressures to go faster and further. These changes have affected various segments of the population. Emerging countries are currently experiencing huge, albeit uneven, traffic growth. In Western countries women’s travel behaviors are in the process of catching up with men’s. Clerical staff and blue-collar workers are no longer distinguished by short home-to-work trips. As transport technologies have improved, extensive mobility has become an increasingly important feature of everyday life. According to some researchers (Bacqué and Fol 2007), not only has daily mobility increasingly become the rule, but it has also become more and more unavoidable, as it is required both by social norms and by the nature of modern life. The highly efficient transport that is accessible to part of the population has raised the norm of mobility expected of all. Nowadays, being mobile is a minimum requirement. In addition, urban sprawl, the instability of employment, and the growing pace of urban life require mobility and reactivity. One has to be mobile, and to be able to reform one’s practices in order to be able to keep up with change. Thus, daily mobility is actually less and less a choice for those who perform it, while those who have difficulties doing so are excluded from social integration – as mobility, although widespread, is still not equally accessible to all. Local travel and simultaneously benefiting from a high degree of individual accessibility is a rare privilege. In a number of cities, low-income households have now little option but to select less expensive and less accessible outlying locations from where they spend a lot of money on commuting.

It is as if social status matched travel practices, the hierarchy of which had shifted according to changing norms.

In contrast, two overall observations show constancies in mobility behaviors. First, it is not really contested that, over time, the average daily travel-time budget has remained basically unchanged: despite growing distances, there was no significant increase in the amount of time spent traveling in a day. Second, the same applies to the mean number of trips per day per person (between three and four). While long-distance communication has greatly increased, daily out-of-home occupations have neither decreased nor increased (there are still only 24 hours in everyone’s day). The good news is that the importance of face-to-face contact remains undiminished. The bad news is that it
DAILY MOBILITY

DAILY MOBILITY does away with the initial hypothesis (and hope) that virtual mobility (through information and communication technology) could provide a way of reducing physical mobility and its ecological impacts.

The future of mobility

Today, one of the significant issues that arises from daily mobility is its environmental impact. The need for mobility and the main means by which these needs are fulfilled are acknowledged to be considerably harmful in terms of the energy consumed and the pollutants emitted. In the near future, rising mobility is set to be more of an issue in developing countries. In Western countries, car ownership is already at a high level and car use has even fallen in some regions. Overall daily distances should not increase to any marked degree as the population is aging and not due to grow much. Yet, private modes of transport accounted for 9 out of 10km daily traveled in the United States in 2009 and 8 out of 10km in the United Kingdom in 2012. In France, the transport sector generates one-third of carbon dioxide emissions, and car use is responsible for half of this. But the most worrying issue is caused by the rising standards of living in large countries where car ownership used to be low. In recent years, car ownership has dramatically increased in China and India. This state of affairs is leading academics, planners, and politicians to look for ways of reducing car use or travel distances. This requires public policies that act together on the transport system, land use, and real estate markets – because where people are able to live defines how they are able to travel. The issue is challenging. Just-in-time and ultra-competitive economic models that require fast interaction, the decline in the resources available for improving public transport, and the laissez-faire attitude toward urban sprawl which leads to car dependency all stand in the way of virtuous behaviors. However, norms and attitudes are changing. Laws are gradually implementing environmental obligations, and environmentally friendly travel solutions are being developed (e.g., car sharing, electric cars, bike-share services).

Finally, daily mobility ties in with many other types of mobility (social, residential, and virtual) and aspects of social life (social norms, economic models, technological change) that affect how and where people live. This pivotal position, together with the growth of all other types of physical mobility (professional long-distance journeys, freight, tourism), and the dramatic state of flux that characterizes contemporary social organizations in an increasingly connected world, has led researchers to think of mobility as a broad concept referring to today’s changing world. The “mobility turn” scientific community does not advocate a specific vision of mobility itself or a specific way of grasping it. It is much more concerned with an overall vision of the contemporary world where actual spatial movements, the opportunity to move, and the pace of lifestyle changes increased significantly throughout the twentieth century and dramatically since the start of the twenty-first. It encourages experts and stakeholders to include this phenomenon in their analysis and policies. John Urry (2000) goes even further, taking the view that sociology should no longer be the study of structures (of societies) but of the movement (of individuals, goods, ideas, information, images, money, and so on). Although debatable, this proposal has the merit of inviting us to open up the conceptual boundaries of scientific approaches and to initiate new questions and methodologies in the area of mobility. Over time, quantitative approaches have improved. More recently, qualitative tools
have been developed in order to grasp less visible and less measurable aspects of travel behaviors such as experiences, representations, and perceptions. Thus the overall societal context and academic endeavor continue to pose questions about the definition and nature of daily mobility. The future of debate on mobility is assured.

SEE ALSO: Chicago School; Mobility gaps; Residential mobility; Social geography; Soils of mountainous landscapes; Time geography and space–time prism; Transportation and land use; Travel geographies; Urban geography

References


Dams

William H. Renwick

Miami University, USA

Definition of dams

Dams are barriers constructed with the purpose of restricting water flow. Most dams are constructed across streams, but some are built at upland sites that did not previously have flowing water, but which can impound water that falls as precipitation or that is otherwise delivered to these sites.

Summary statement of extent, purposes, and consequences

Today millions of dams exist around the world. Their impacts on human societies and on natural processes cannot be understated. They provide reliable supplies of drinking water to billions of people and irrigation water that has dramatically increased the extent and quantity of agricultural production and renewably generate 17% of the world’s electricity. These benefits come with significant costs, however. Construction of new dams displaces thousands and sometimes millions of people. Worldwide, dams impound over 6000 km$^3$ of water, or about 16% of total annual stream flow, significantly altering downstream aquatic and riparian environments. In addition to storage, an unquantified but certainly substantial portion of the world’s river flow passes through reservoirs that trap sediment, carbon, nitrogen, phosphorus, and other key components of global ecosystems. Dams have altered the distribution of plants and animals, both creating new fisheries and causing extinctions of riverine species.

Dams vary tremendously in size, with respect to the height and length of the dams themselves and also the reservoirs that they create. The highest dams in the world are hundreds of meters high, and the longest ones are tens of kilometers long. These very large structures typically impound the largest amounts of water, but are relatively few in number. Large dams – those that impound at least 0.1 km$^3$ of water – number in the thousands. Smaller dams, down to those that create small ponds of 1 ha or less, number in the tens of millions. These smaller dams account for more than 99% of all dams, but probably only about 20% of total reservoir surface area. The number of dams in the world continues to grow, especially small dams.

Because of their imposing size, importance to society, and environmental significance, dams are controversial. When they are built they displace natural ecosystems and human populations as well as inundate valuable land. Dam failures, while relatively few, are among the great disasters in history. Dams become semipermanent features of the landscape with lasting and profound impacts on water resources, infrastructure, economic activities, and the distribution of population. Opposition to construction of large dams has grown in recent decades and this, along with the existence of alternative sources of electricity and approaches to water-resource needs, has contributed to a decline in the rate of large dam construction, particularly in industrialized countries. However, we continue to build new small dams.
DAMS

by many thousands each year. In some areas older
dams – both large and small – are being removed
as sedimentation or changing needs have reduced
their value to society, and removal can reverse
some of their negative environmental impacts.

History of dam-building

Ancient dams

The earliest known dams are located in the
Middle East, associated with irrigation-based
civilizations that developed in present-day Jor-
dan, Syria, Iraq, and Egypt, and date from
the fifth–third millennia BCE (Fahlbusch 2009).
Ancient dams also survive in South Asia, China,
Japan, and Central America. Dams were an
integral part of the co-development of irri-
gated agriculture and hydraulic engineering that
supported early civilizations. While irrigation
remained the dominant function of most early
dams, they also supplied water for other pur-
poses such as domestic use, often carried in
via complex systems of canals and aqueducts.
The Roman Empire made many contributions
to hydraulic engineering technology, including
dams. Several Roman dams survive and are in use
today, including Cornalvo Dam in Spain, Subi-
aco Dam in Italy, and Lake Homs and Harbaqa
Dams in Syria.

Ancient dams were mainly earthen and/or
masonry dams that relied on their weight and
gravity to remain in place and resist the pressure
of water on the upstream side. The Romans built
such gravity dams but also were among the first to
use masonry buttresses and arches to strengthen
dams. Concrete was also used extensively in
Roman dams, as it is today.

Dams are intimately connected with civi-
lization. Dam construction inevitably requires
high levels of social organization to mobilize
labor and capital resources to build dams and
distribute the water that is made available by
them. Nearly all large dams in the world have
thus been built by governments, or at the
very least with government authorization and
coordination.

Dams and hydropower

In addition to providing water for irrigation
and human uses, dams have also been used for
millennia for generating mechanical power – for
grain milling, sawing lumber, and more recently
powering industrial facilities and generating
electricity. Mill dams were common in the
middle ages in Europe as well as in colonial
times in North America, but the industrial
revolution of the eighteenth and nineteenth
centuries as well as accompanying urbanization
spurred a dramatic increase in the number of
dams. These new dams were built for industrial
purposes (especially sawmills in North America),
as feeders for canal systems, and for urban water
supply. Examples include Bosley Reservoir and
Anglezarke Reservoirs in the United Kingdom,
built circa 1830 and circa 1850 for canal water
and drinking water supplies, respectively, and
Telos Dam in Maine, built in 1841 to support
timber harvesting.

With the nineteenth century development
of efficient water turbines as opposed to water
wheels, the use of dams for hydroelectric power
became much more feasible. The first hydro-
electric power station opened in Appleton,
Wisconsin, in 1882. Use of dams for hydropower
grew rapidly thereafter in the industrialized
world, with many dams of varying sizes con-
structed in the early twentieth century. Today
hydroelectricity is a dominant use especially
in very large reservoirs: more than half of the
world’s reservoir storage capacity is used primar-
ily for generating electricity.
Modern multiple-use dams

Because reservoirs offer diverse benefits to society, most reservoirs today are used for multiple purposes although one or two purposes are typically foremost. The most important uses for dams today are irrigation, hydroelectricity, water supply, flood control, waste containment or processing, navigation, and recreation. The opportunity for multiple uses increases the political attractiveness of undertaking construction of dams.

The middle to late twentieth century was the period of greatest dam-building around the world, as measured by volume of water storage capacity (Figure 1). Ninety percent of world capacity was built between 1940 and 1995; 50% was built between 1960 and 1980. The era of big dam-building was slightly earlier in the United States: 90% of capacity was built between 1915 and 1980; 50% was built between 1950 and 1968. Since about 1980 the pace of dam construction in the world has dropped dramatically, particularly in industrialized nations.

Large dams today

Dams are among the largest human-built structures in the world, and if one includes the reservoirs they impound they are probably the largest. For example, the Three Gorges Dam in China incorporates more than 27 million m$^3$ of concrete, and the impounded reservoir is 1045 km$^2$ in area with a volume of 39.3 km$^3$. Hoover Dam in the United States was the world’s largest concrete structure when it was built in 1935 (2.4 million m$^3$). It is 223 m high and impounds a reservoir 640 km$^2$ in area and 37.3 km$^3$ volume. The Rogun Dam in Tajikistan, at 335 m high, is among the world’s highest. The Grande Dixense in Switzerland is 285 m high. These very high dams tend to be in narrow valleys. The longest dams in the world are tens of kilometers in length. The Yacreta Dam in Paraguay is composed of an 800 m long concrete structure and a 65 km earth dike. The Martin Power Plant just east of Lake Okeechobee in Florida is supplied by cooling water from a reservoir that is contained within a 27 km-long dike, 10 m high.

Figure 1  Historic capacity growth of large reservoirs in the United States and the world (in km$^3$). Data sources: US Army Corps of Engineers (2013) and Lehner et al. (2011).

World

Dams are found throughout the world (Figure 2). Reservoirs cover >500 000 km$^2$ of Earth’s surface – an extent comparable to the total area of the world’s cities. They are also extremely common: in the conterminous United States there is an average of about one dam for every 2 km$^2$ land area. The greatest concentrations of large dams are in relatively industrialized
countries, but some of the largest dams are in less-industrialized countries. As of 2011 there were nearly 6900 large reservoirs in the world with a storage capacity of >0.1 km³ each (Lehner et al. 2011). The total storage capacity of these reservoirs is about 6200 km³, or about 16% of annual runoff from land areas of the world to the oceans. Among the largest dams in the world, by impounded water volume, are Kariba (Zambia, built 1959), Bratsk (Russia, 1964), Aswan (Egypt, 1970), Akosombo (Ghana, 1965), and Daniel Johnson (Canada, 1968). The five countries with the greatest impounded water volume in large dams are Canada (14% of world total), Russia (13%), United States (12%), Brazil (8%), and China (7%). These five countries account for more than 50% of total world reservoir capacity. In addition to these very large artificial reservoirs, many large lakes have dams at their outlets that allow water level regulation. These dams are typically of only modest height and may only regulate lake levels by a few meters, but because the lakes are very large the total storage is also large. Dams at the outlets of Lake Victoria (Uganda, 1954) and Lake Superior (Canada/USA, 1921) account for the largest storage volumes in the world and the United States, respectively.

The principal uses of the world’s large dams are irrigation (34% of dams) and hydropower (29% of dams) (Lehner et al. 2011; Table 1). About 67% of storage capacity is in hydropower reservoirs and 19% in irrigation reservoirs. Hydropower represents more capacity because more humid and often remote regions are often favorable sites for large dams. Areas that need irrigation water are dry and so less water to fill a large reservoir is available, and also land that would be lost to inundation may be more valuable. Water supply is the third most common primary use of large reservoirs, followed by flood control and recreation. Today, navigation is a relatively minor purpose of dam construction, although in the nineteenth century when canals were more central to the transport systems of industrialized countries they were proportionately more significant.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>World (GRanD)*</th>
<th>United States (NID)†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity range (million m³)</td>
<td>100–205000</td>
<td>100–37300</td>
</tr>
<tr>
<td>Number of dams</td>
<td>5284</td>
<td>1205</td>
</tr>
<tr>
<td>Purpose</td>
<td>Percent of dams</td>
<td>Percent of storage</td>
</tr>
<tr>
<td>Irrigation</td>
<td>33.7%</td>
<td>19.6%</td>
</tr>
<tr>
<td>Hydroelectricity</td>
<td>29.1%</td>
<td>65.4%</td>
</tr>
<tr>
<td>Water supply</td>
<td>16.0%</td>
<td>2.9%</td>
</tr>
<tr>
<td>Flood control</td>
<td>10.4%</td>
<td>9.0%</td>
</tr>
<tr>
<td>Recreation</td>
<td>5.5%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Other</td>
<td>3.9%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Navigation</td>
<td>1.1%</td>
<td>1.9%</td>
</tr>
<tr>
<td>Fisheries</td>
<td>0.3%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

*GRanD database (Lehner et al. 2011), excluding Owens Falls Dam at outlet of Lake Victoria.
†National Inventory of Dams (US Army Corps of Engineers 2013), excluding Soo Compensating Works at outlet of Lake Superior.
DAMS

United States

In the United States, most large dams are found in the northern plains (Upper Missouri River), west and southwest (Pacific Coast and Colorado River drainages), south-central (Texas and Oklahoma), and southeastern (Tennessee Valley, southern Piedmont) regions. These regions, though varying widely in climate and land use, were all foci of government dam-building activities in the twentieth century. In the west, need for irrigation water stimulated the 1902 Reclamation Act, which ultimately led to the construction of many dams including Hoover and Glen Canyon Dams (Arizona, 1935 and 1963, respectively) on the Colorado River, Garrison and Oahe Dams (North and South Dakota, 1963 and 1953 respectively) on the Missouri River, and Grand Coulee Dam (Washington, 1941) on the Columbia. The Tennessee Valley Authority, created in 1933 to address a variety of economic and environmental problems that plagued the region, built many large dams including Kentucky and Wolf Creek Dams (Kentucky, 1944 and 1951, respectively), and Norris Dam (Tennessee, 1936). The Army Corps of Engineers, through its authority to improve navigation and conduct flood control activities, has built many dams including Sam Rayburn Dam (Texas, 1965), Bull Shoals (Arkansas, 1951), and J. Strom Thurmond Dam (Georgia, 1954).

The Corps of Engineers maintains a National Inventory of Dams (NID; US Army Corps of Engineers 2013), which includes dams that are more than 7.6 m (25 ft) high, hold more than 800 000 m³ (650 acre-feet) of water, or are considered a significant hazard if they were to fail. In 2013 the database contained 86 461 impoundments; some of these are contained by multiple dams. The total storage capacity of these dams is 1689 km³, or about 58% of total annual runoff from the United States. Ninety-two percent of this capacity is in 1205 large dams with capacity >0.1 km³. Considering just the 99.7% of NID dams that are within the conterminous 48 states, storage is 103% of total annual runoff.

The primary uses of large dams (capacity >0.1 km³) in the United States differ from those in the rest of the world: flood control is the most common primary use, followed by hydroelectricity, water supply, and irrigation (Table 1). The emphasis on flood control probably results from the United States’ longer history of dam construction combined with a level of economic development and urbanization that places populations and valuable facilities at risk. When the uses of the 1205 large dams in the NID are compared with the relatively smaller 85 256 remaining dams, the differing uses of large and smaller dams are evident. Among the smaller dams in the United States recreation is the most common primary use, followed by “other,” flood control, and irrigation. The “other” category includes fire protection (which accounts for 13% of dams but only 1.6% of capacity), wastewater treatment, power plant cooling water, mine tailings storage, and erosion control. These are uses that tend to be widely dispersed and typically do not require large volumes of impounded water. Finally, about 3% of the dams in the NID are “off-stream,” which means that the dam is built on a level or upland area rather than across a stream. Many of these off-stream dams are for water treatment (including tailings storage), fisheries, power plant cooling, and pumped-storage hydroelectricity.

While the NID is a very comprehensive database, it includes only a fraction of the dams in the United States. We have a relatively accurate inventory of all the lakes and reservoirs in the United States in the National Hydrography Dataset (NHD) – a digital version of the detailed topographic maps produced by the US Geological Survey (n.d.). The NHD includes about 6.57 million lakes, ponds, and reservoirs, but does not
distinguish between those of natural origin and those created by dams. In the aggregate we can estimate how many of these features are likely to be of natural or human origin based on known lake-forming conditions. For example, natural lakes are common in glaciated areas, while unglaciated areas with rolling or steep topography rarely contain natural lakes. Using this knowledge we can infer that about two-thirds of the lakes/ponds/reservoirs in the United States are of human origin. In other words, there are perhaps 4 million to 4.5 million dams, most of them too small to be included in the NID. Roughly half of the ponds created by these dams are less than 0.1 ha in size (about a quarter of a football field). So while very small reservoirs less than 0.1 ha in area are extremely numerous, they do not account for a very large portion of total water surface area – only about 1%.

Environmental consequences

Flow and flow regime

Dams are designed to store water and thus alter stream flow: to prevent floods, augment low flows, or indirectly modify flow by releasing water in relation to electricity demand, irrigation needs, or other purposes. Dams also divert water from streams to off-stream uses such as irrigation that increase evapotranspiration and thus decrease total stream flow. Although each individual dam will vary in its hydrologic effects, in general the hydrologic effect of dams is to reduce the magnitude or frequency of flooding, reduce total flows, and sometimes increase low flows.

The effects of dams on flow regime are determined by the design of the dam and its reservoir and, for dams that have outlets that can be manipulated, the operation of those outlets. Most large dams have outlets that are controlled by gates or valves, and so the amount of stored water (also known as pondage) can be deliberately varied. For hydroelectric power dams, the rate outflow is usually governed by demand for electricity, and the outflow from such dams may vary accordingly by time of day as well as by season. Dams with the principal function of flood control are usually operated so as to have relatively little water stored at the beginning of the high-flow season, to make room to store runoff. Dams primarily intended for irrigation typically retain water during times when irrigation demands are lower, and release it when those demands are higher, with corresponding impacts on downstream stream flow.

In contrast, most small- and many medium-sized dams have little or no pondage, and so the amount of water that spills over the dam is mainly a function of inflows. Such dams are sometimes referred to as run-of-the-river dams. But even these store a small amount of water during times of rising inflow and release it when inflows are lower, and so exert some moderating influence on downstream flows, reducing peak discharges.

In addition to their direct impacts on flows, dams are key components of water withdrawal and consumption systems that have resulted in reduced total flow in many areas. Such effects are most dramatic in arid and semiarid regions with large irrigation water use, such as the southwestern United States, China, and central Asia. Well-known examples of such impacts include the catastrophic reduction in the size of the Aral Sea in Kazakhstan and Uzbekistan, and the elimination of perennial flows in the lower Colorado River (US and Mexico) and the Huang He (China).

Sedimentation and sediment transport

Sedimentation is probably the most significant water quality effect of dams and their reservoirs. Sediment transported by streams that flow into reservoirs is trapped in the reservoir and thus not carried on downstream. All dams trap some
sediment and large reservoirs often retain nearly all the sediment they receive. Because of this sedimentation, reservoirs eventually can fill with sediment if loads in inflowing streams are high enough. Most dams are built with anticipated reservoir life spans of at least several decades, although some are built specifically for the purpose of removing sediment from the water. In the conterminous United States today, total reservoir sedimentation is estimated at 2–3 km$^3$ per year, or about 0.1 to 0.2% of total reservoir capacity (Renwick et al. 2005). This estimate masks great local variability, however: in areas of substantial stream sediment loads it is common for reservoirs to lose a substantial portion of their total capacity in as little as a few decades. Sometimes reservoirs are dredged in order to remove accumulated sediment.

Because dams are so ubiquitous, the total amount of sediment that is removed from the world’s river systems is very large. Perhaps 50% of all sediment transported by rivers is trapped in reservoirs (Vorosmarty et al. 2003). In addition to reducing the storage capacity and usable lives of reservoirs, this sedimentation has downstream effects. Sediment loads are reduced below reservoirs causing some streams to erode their beds and banks, particularly in the area immediately downstream from the dam. The loss of sediment from streams can also have ecological consequences to downstream environments, particularly in very dynamic streams where sediment deposition and erosion are important parts of riparian ecosystems. Such problems have been a major concern in the Grand Canyon below Glen Canyon Dam, for example.

**Aquatic communities**

Dams and their reservoirs transform riparian environments by breaking relatively continuous free-flowing streams into a series of stream segments separated by deep, low-velocity impoundments and the physical barriers created by the dams themselves. Whereas a river is relatively shallow, with frequent variations in depth and width associated with stream flow variations, reservoirs may have large surface areas and large volumes of water that has only limited interaction with adjacent shallow-water and terrestrial habitats. Deep water may become depleted in oxygen, especially in eutrophic reservoirs common in populated regions where inputs of nutrients are high. Aquatic communities are thus transformed from those that require higher velocity flowing water and high oxygen content to those that are more typical of lakes (Nilsson and Berggren 2000). Because they typically have higher inflow rates in relation to water storage than in natural lakes, reservoirs are more likely to suffer from water quality problems associated with high nutrient and sediment levels.

Impoundment of flowing water in reservoirs has water quality consequences, some of which are positive. Sedimentation removes mineral and organic matter from the water, and along with it excessive nutrients may also be removed. If water is released from deeper portions of a reservoir, river reaches immediately downstream from dams may have low temperatures and/or low oxygen content.

Dams also form formidable barriers to upstream–downstream migration of fish and other aquatic organisms. Perhaps most important among these are anadromous fish such as salmon that are economically important as food sources as well as key components of some ecosystems. Salmon have been virtually eliminated from portions of the Columbia River system upstream from dams on several major Columbia River tributaries. Today salmon stocks on the Columbia are at less than 2% of their levels prior to the era of big dams and several populations are considered endangered.
The possibility that snail darters, an endangered species of fish that lives in Tennessee, would become extinct as a consequence of completing a major dam was the focus of a major controversy involving endangered species protection. The controversy, which ultimately did not prevent the dam in question from being constructed, increased public awareness of the environmental effects of dams.

**Dam failures**

Dam failures are relatively rare, but when they occur they are major disasters. Dam failures can cause great loss of life when populated areas lie downstream, which is often the case given the importance of dams in meeting human water needs. The floods that result from dam failures are large and often unpredicted, and the water rises very quickly, making it difficult for those in the path of the flood to escape. Dam failures have occurred throughout the history of dam engineering, and still occur despite the increasing sophistication of dam engineering. A review of some recent dam-related catastrophes provides some insight as to the range of causes and impacts.

- The Machhu failure in Gujarat, India, 1979, occurred when heavy rains caused flooding that exceeded the capacity of the emergency spillway (a channel intended to carry water when normal water-release structures are insufficient). The resulting flood inundated the town of Morvi downstream, killing thousands. The design criteria in use at the time the dam was built were found to have been inadequate, and the dam has since been rebuilt with a larger spillway.
- The Teton Dam, an earth-fill dam in Idaho, collapsed in 1975 as the reservoir was being filled, killing 14 people. The rock adjacent to the dam was highly permeable, and although during dam construction efforts were made to seal the rock this work was inadequate. Water seeped through the rock and undermined the dam, which then collapsed.
- A coal slurry impoundment dam failed during heavy rains in West Virginia in 1972, causing a flood on Buffalo Creek that killed 125 people. The dam had been built on unconsolidated mine waste and was not constructed to prevailing engineering standards.

These examples, among many others, show that dam failures are nearly always attributable to lack of adherence to standard engineering practices. In general, the existing knowledge base is sufficient to support dam construction and management practices that will prevent dam failures.

**Future of dam-building**

**Potential for new large dams**

The greatest potential for construction of new large dams is in less-densely populated regions where displacement of communities is less controversial. Worldwide, the greatest potential for hydropower production is in China, Russia, Brazil, and Canada. The US Energy Information Administration (2013) projects an increase in total world hydroelectric production of over 75% between 2015 and 2040; presumably most of that increase would be accomplished by building new dams although some could come from adding hydroelectric capacity to existing dams. There will also be increased demand for irrigation water as well as public water supply, and this will certainly mean construction of additional new dams.
DAMS

Where population growth has slowed, growth in water demand is also slower, and the shift from industrial to service economies has also reduced growth in water demand. Resistance to dam construction has also increased, driven by interest in conserving natural environments and concerns about displacement of populations in areas that would be inundated. Given the combination of decreasing availability of dam sites and major untapped water resources, as well as increasing public opposition to dam construction in at least some areas, it is unlikely that we will see a return to the fast pace of dam construction that was witnessed in the mid- to late twentieth century.

Dam removal

In some areas, notably the United States, there has been a significant increase in dam removal. Among the millions of dams that exist in the United States are many that have outlived their former utility and are no longer serving important purposes. Some of these have lost value because of sedimentation, while others have become unsafe as a result of structural deterioration. Removal of these dams is motivated by a desire to restore the ecological health of the rivers that they impound. While some dams in the United States were removed in the early twentieth century, the pace of dam removal has increased rapidly since the 1980s, and in recent years typically 50–70 dams have been removed each year. Over 1000 have been removed as of 2014. Most of the dams removed have been relatively small dams (2–5 m tall), but some larger dams have been removed. Among the more prominent dam removal projects is that on the Elwha River, Washington, in which two major dams have been largely removed allowing five species of Pacific salmon to return to the upper Elwha River.

SEE ALSO: Ecological footprint; Energy resources and use; Hydrologic cycle; Lakes and limnology; River basin management and development; Rivers and river basin management; Rivers and streams; Stream ecosystems; Surface water; Water and climate change; Water budget; Water: drinking; Water engineering; Water quality; Water resources and hydrological management

References

Further reading


Website

www.americanrivers.org
Dasymetric mapping

Jeremy Mennis
Temple University, USA

Definition

A dasymetric map is a type of thematic map intended to represent a statistical surface of density, that is, a measurement that varies continuously over space and which captures the rate of occurrence of some phenomenon over space. In a dasymetric map, the statistical surface is represented by partitioning the space into a set of spatial units where the unit boundaries occur at the steepest escarpments of the surface, that is, where the density transitions most abruptly from one location to the next. Thus, a dasymetric map has spatial units that maximize within-unit homogeneity while minimizing between-unit homogeneity. Ideally, the density within any given spatial unit of a dasymetric map should be approximately equal. A dasymetric map is often contrasted with a choropleth map, which is also a thematic mapping technique used to represent the spatial distribution of a quantitative variable, density or otherwise. Unlike a choropleth map, however, where the spatial unit boundaries are typically derived from some convenience of enumeration or data collection that has little or nothing to do with the actual spatial variation in the data, a dasymetric map’s spatial units are derived from, and are intended to depict, the nature of the spatial variation in the mapped variable.

Consider, for example, the map of a set of points shown in Figure 1(a). A dasymetric map of this point distribution would depict the point density as a set of spatial units. Figure 1(b) shows such a dasymetric map, where the colors of the units correspond to densities ranging from relatively sparse (blue) to highly concentrated (red).

The depiction of a statistical surface using a dasymetric map demands that the true point distribution, or at least its representation in a statistical surface, is known. Most research in dasymetric mapping, however, is actually geared towards situations in which the true point distribution or underlying statistical surface is unknown, and the objective is to develop a strategy to infer and represent such a surface given the limitations of the available data. The process of dasymetric mapping, also called dasymetric modeling, typically refers to a set of techniques used to transform data encoded in a choropleth map representation into a dasymetric map, or at least a representation that better approaches a dasymetric map as compared to the original choropleth map encoding.

For this purpose, dasymetric mapping utilizes ancillary data, a spatial dataset separate from the original variable being mapped, but related to it. By defining a functional relationship between the ancillary data and the original mapped variable, data in the original choropleth map may be redistributed within a new set of spatial units that more accurately and precisely depict the true spatial distribution of the variable. Dasymetric mapping thus spatially disaggregates data encoded in a choropleth map to produce a more precise representation of the distribution of a density variable.
Figure 1  A dasymetric map representation of point density. (a) Distribution of points. (b) Corresponding dasymetric map representing the point density using a red to blue color gradient, where red indicates higher density and blue indicates lower density.

Origins and development

The idea of dasymetric mapping can be traced back to early modern maps of population distribution such as George Julius Poulett Scrope’s 1833 world map of population density. Petrov (2012) established that the term “dasymetric” (roughly fusing the Greek words for “density” and “measure”) was coined by the Russian cartographer Benjamin Semenov-Tian-Shansky in 1911 and implemented for an ambitious project to map the population of Eastern Russia. The technique was introduced to American cartographers in a handful of articles appearing in the journal *The Geographical Review* in the first half of the twentieth century, perhaps the most notable being J.K. Wright’s (1936) article, which described the basic form of the dasymetric mapping approach in detail. With a few exceptions, dasymetric mapping remained a somewhat obscure cartographic corollary to the more established choropleth mapping approach until the early 2000s, when geospatial technologies such as geographic information systems (GIS) and environmental remote sensing provided the data and computational processing to facilitate dasymetric mapping, resulting in a rapid rise of research articles advancing dasymetric mapping techniques.

The development of these geospatial technologies, and their application to mapping, has facilitated the convergence of cartographic dasymetric mapping research with related research in spatial statistics and environmental remote sensing. In statistics, the topic of areal interpolation addresses the issue of incompatible spatial
units, that is, where one has two spatial datasets to be analyzed but they are encoded in different tessellations of spatial units. Areal interpolation refers to the transformation of data from one set of spatial units to another, typically to facilitate data integration for analysis. Dasymetric mapping can be considered a specific type of areal interpolation, where an ancillary dataset is used to facilitate the transformation of data from one set of spatial units to another, finer set of spatial units. Such an approach using an ancillary dataset is also referred to as intelligent areal interpolation (Goodchild, Anselin, and Deichmann 1993).

Researchers in environmental remote sensing have also focused on modeling population distribution using remotely sensed imagery. This requires the development of a functional model between remotely sensed pixel values and population, which is typically derived using pre-existing population data encoded in a choropleth map. Thus, the use of remotely sensed imagery to derive population estimates is inherently dasymetric in its approach.

Applications

Dasymetric mapping can be applied to any spatially distributed variable that is conceptualized as a statistical surface representing density. However, it was developed specifically for population mapping, and in practice the vast majority of dasymetric mapping applications remain population mapping. Thus, the emphasis here is on population mapping, although the principles of dasymetric mapping may theoretically be applied to mapping the density over space of any punctiform variable. Dasymetric mapping has been applied to a wide variety of areas where more spatially precise population information is required compared to what may be available through surveys. In less developed countries, where census surveys may produce rather coarse spatial estimates of population, dasymetric mapping may be used to develop higher spatial resolution population estimates. Thus, dasymetric mapping has been used to develop high-resolution population datasets at national, continental, and global scales (Bhaduri et al. 2007).

One of the most prominent applications of dasymetric mapping has been for emergency disaster planning, where precise information on population distribution is required for identifying populations at risk due to floods, earthquakes, and industrial accidents, as well as to other hazards such as pollution. Here, dasymetric mapping is used to match incompatible spatial units containing population data and data describing degree of hazard or vulnerability. Relatedly, social services provision analysis, such as for health care, has utilized dasymetric mapping to develop more precise measures of population relating to accessibility to services. Dasymetric mapping has also been used extensively to develop historical series of population data, first, in cases where historical population data are encoded at a very coarse resolution and more spatially precise population estimates are desired, and, second, in cases where census enumeration units have changed over time (Gregory and Ell 2005). Dasymetric mapping has been used to develop a time series dataset of population change encoded using a single spatial tessellation and, thus, to facilitate the analysis of population change.

Ancillary data

Dasymetric mapping employs an ancillary spatial dataset to transform data from a set of choropleth map spatial units to a map that more precisely depicts the underlying statistical surface. The ancillary data layer is often an area-class map,
where the region of interest is exhaustively tesselated into a set of spatial units taking on nominal values, as with, for example, land cover classified into residential, industrial, agricultural, and vegetated categories. Typically, the result of dasymetric mapping is not necessarily an idealized dasymetric map, where the spatial units perfectly partition the surface into maximally homogeneous regions, but rather a map partitioning the original choropleth map units into a finer tessellation of spatial units using the ancillary data. This is typically performed by overlaying the original choropleth map with the ancillary data layer to produce a dasymetric map, where the dasymetric map units are perfectly spatially nested within the original choropleth map units.

The most common ancillary dataset for use in dasymetric mapping is land cover, often derived from classified remotely sensed imagery. In its most basic form, for instance, inhabited and uninhabited land may be distinguished between, and the dasymetric mapping procedure would then apportion the population within each spatial unit of the original choropleth map only to those portions of the dasymetric map with a land cover classified as inhabited. Thus, the population density of the inhabited portions of the map would increase compared to their original encoding in the choropleth map, and the population density of those uninhabited regions would be reduced to zero.

This basic formulation can be extended depending on the attribute precision and classification scheme of the ancillary data. For example, a land cover dataset that classifies inhabited land as either high-density residential, medium-density residential, or low-density residential, may be used to apportion a greater population density to high-density residential areas compared to the other land covers. Other useful land cover classifications include distinctions between various types of forest cover, agricultural land, impervious surface cover, and distinctions between developed and undeveloped land. While such land cover classes may not directly indicate habitation, they may be used as rough proxies for population density. For example, it might be expected that agricultural regions have a lower population density than urban regions, or that higher percentages of impervious surfaces indicate developed land with relatively higher population density compared to areas with a lower percentage of impervious surface. It is precisely these differences in the relationship of population density with different types of land cover that dasymetric mapping techniques seek to identify and exploit to transform population density data from a choropleth map representation to a dasymetric map representation.

While land cover is the most prominent ancillary dataset used in dasymetric mapping, many other types of datasets have been employed. Remotely sensed data from satellite and airborne platforms have been used extensively, including the use of raw spectral reflectance values, aerial photographs or panchromatic imagery for the manual and automatic detection of residential structures, and light detection and ranging (LIDAR) for estimating building volumes. Remotely sensed data indicating nighttime lights, vegetation indices, and digital elevation models (DEMs) have also been used. Road network data have been used effectively in dasymetric mapping (Reibel and Bufalino 2005), where higher road density is associated with higher population density, and may be used to distinguish developed areas from more rural areas. Parcel-level data, including zoning and building attributes, have also been used (Maantay Maroko, and Herrmann 2007), as has address-point data which encode the locations of addressable units. Publicly available spatial data
such as bus stops, schools, and raster pixel maps have also been used effectively (Langford 2013).

Techniques

Dasymetric mapping seeks to transform population data encoded in a choropleth map, or what can be termed a set of “source zones,” into a finer resolution dasymetric map composed of a set of “target zones.” Generally, this is done by overlaying the sources zones with an ancillary spatial data layer within a GIS software package, where the target zones are composed of the resulting polygons that spatially nest within the original source zones. This might be considered a process of dis-aggregation of the original source zones into a greater number of smaller target zones. Allocation of population to the target zones is dependent on defining the functional relationship between the underlying surface of population density, which is unknown, and the ancillary data layer.

Techniques for defining the ancillary data–statistical surface relationship for dasymetric mapping can be divided into those techniques that stem from the original cartographic basis of dasymetric mapping and those, generally more recent, methods that take a more statistically oriented approach. The most basic formulation, sometimes called the “binary” method, is simply to use an area-class map with only two categories of inhabited and uninhabited land. The entire population of each original choropleth source zone is then allocated wholly to the portion of the source zone that is occupied by inhabited land.

As an illustration of ‘binary’ dasymetric mapping, consider, for example, Figure 2. The map on the left of Figure 2 shows a set of source zones with the population density values for each spatial unit annotated. The map in the middle of Figure 2 shows the ancillary data layer with inhabited and uninhabited regions annotated. The map on the right of Figure 2 shows the resulting dasymetric map, created by overlaying the source zones and ancillary zones to create a set of target zones. Note that the target zones perfectly nest within the source zones. Here, the target zone component of each source zone that is spatially coincident with the uninhabited region specified in the ancillary layer is assigned a population of zero. The target zone that occupies the remaining inhabited portion of the source zone thus has a higher population density as compared to its original host source zone, as the

![Figure 2](image)

**Figure 2** Illustration of the “binary” method of dasymetric mapping: source zones (left), ancillary layer (middle), and target zones (right).
entire portion of that source zone’s population is assigned to the inhabited target zone that lies within it. For example, consider the source zone in the northeast corner, which has a population density of 10 people/km². This source zone lies half within an inhabited region and half within an uninhabited region, resulting in the original source zone being split into two target zones in the dasymetric map. Thus, the target zone in the uninhabited region has a population density 0 people/km² whereas the target zone in the inhabited region has a population density 20 people/km² – double that of the original host source zone – because the inhabited target zone accounts for the entire population of the host source zone in only half the area.

An extension of the binary method is the multiclass method, where, as the name implies, more than two categories of an area-class map are used. Here, however, the functional relationship of population density and the area-class map categories is not restricted to simply inhabited versus uninhabited areas, but rather indicates magnitudes of expected population density for each area-class map category. For instance, we might distinguish between agricultural, suburban, and urban land cover, where population density is assumed to increase in magnitude from the former to the latter land cover category. Thus, the population of a source zone composed equally of each of those three land cover categories would ascribe the most people to the urban land cover target zone, less to the suburban zone, and even less to the agricultural zone. The allocation of population to each ancillary class can be imposed a priori by the analyst, or may be derived empirically from the data.

A central principle for both the binary and multiclass dasymetric techniques is that of the limiting variable. That is, a nominal value in an area-class map serving as ancillary data limits the population density ascribed to areas with that nominal value. Once that population density threshold is reached, population is distributed to other dasymetric zones of other area-class map values. The binary method is the simplest implementation of the limiting variable approach, as the technique limits uninhabited land to a population density of zero, and then distributes that population within any given source zone to the other, inhabited, target zones within the original host source zone.

Another important principle of dasymetric mapping is the pycnophylactic property – the idea that the dasymetric map represents the same total population for the region as the original choropleth map. More specifically, the sum of the population for all the dasymetric target zones that fall within a given source zone (recall that the target zones spatially nest perfectly within the source zones) will be equal to the total population of the source zone in the original choropleth map. In dasymetric techniques that produce population estimates outside the bounds of the original choropleth map, the pycnophylactic property can be enforced by simply scaling the estimated population density for each ancillary class, such that the dasymetric technique distributes the population of the source zone to its constituent dasymetric zones based on a proportion of its population density estimates.

A key parameterization of dasymetric mapping is the functional relationship between the nominal ancillary data map categories and population density. This may be done through the expert knowledge of the analyst (Eicher and Brewer 2001), for instance an analyst may enforce a population density limit of zero for areas covered by water. More typically, however, the functional relationship between ancillary data and population density is derived through some sort of sampling (Mennis 2003). If it is possible to identify a source zone in the original choropleth map that lies entirely within a single type of land cover,
Figure 3  (Continued overleaf)
the population density of the source zone can be used to indicate the population density of that land cover class. Alternatively, if ancillary data polygons occur wholly within a choropleth map polygon, the population density for that ancillary data class can be observed. Various types of sampling techniques have been proposed to identify candidate sets of observations to estimate population density values for different ancillary classes. Various statistical approaches have been developed to define the functional relationship between the ancillary data and population density. A simple approach, given a sample of observations associating the ancillary data classes with population density values, is to simply calculate the mean population density for each ancillary data class and scale the figures to preserve the pycnophylactic property.

Figure 3 (from Mennis 2009) shows an example of this approach developing a dasymetric map of population for Delaware County, PA, where population data from the US Bureau of the Census at the tract level (Figure 3(a)) have been disaggregated using classified remotely sensed land cover data (Figure 3(b)) to produce a dasymetric map of population density (Figure 3(c)). Figure 4 shows a close-up view of the red boxed area in Figure 3(c). It can clearly be seen that the spatial precision of the population distribution has been substantially increased, with many
portions of the original tracts being assigned zero population while other tract components have increased in population density.

Regression has also been widely used for dasymetric mapping, where, say, the percentage that a particular ancillary class occupies within a target zone serves as the explanatory variable to estimate the total population. Other approaches include the use of the expectation maximization (EM) algorithm and maximum entropy (ME) modeling to iteratively refine the estimated population density for different ancillary data classes, as well as the use of kriging to improve the estimates derived from regression.

Considerations and future research

One of the major challenges of dasymetric mapping is in assessing the accuracy of its results, which is typically measured using root mean square error (RMSE) or similar metrics. Research has demonstrated that dasymetric mapping clearly offers a more accurate representation of population density than areal weighting, which does not employ an ancillary dataset and simply allocates population from source zones to target zones assuming a homogeneous distribution of the population. However, no single statistical methodology has been shown to be consistently better than the others. And while certain types of high-resolution ancillary data, such as parcels (Maantay, Maroko, and Herrmann 2007) and address points, have been shown to produce better results than coarser-resolution ancillary data, such datasets are not nearly as accessible or widely available as remotely sensed imagery indicating land cover, for which global datasets are publicly available. This issue is particularly germane in the case of less developed parts of the world where population data are

Figure 4  A close-up view of the red boxed area in Figure 3, comparing the dasymetric map (left) with the original choropleth map (right).
DASYMETRIC MAPPING

particularly coarse or unreliable and dasymetric mapping may be most useful.

The performance of dasymetric mapping is also highly dependent on the characteristics of not only the statistical technique and the ancillary data employed, but also the character of the region to which dasymetric mapping is applied, that is, whether it is urban versus rural, the types of settlement patterns, and so on. Comparisons of different dasymetric methods have shown that the most accurate method and most useful ancillary dataset can differ from region to region, even in a single country. So, while the use of ancillary data and dasymetric mapping clearly produces more accurate population maps, there is no single approach or dataset that is considered best for all situations.

SEE ALSO: Choropleth map; Interpolation: areal

References


Further reading


Data model, event-oriented

Kathleen Stewart
University of Maryland, USA

In recent years, the development of theories that address spatiotemporal aspects of data modeling has attracted an increasing number of researchers from different disciplines. One aspect that has gained traction with researchers in the field of geographic information science is the development of theories and data models for events, that is, distinct happenings or dynamics, that involve change and are distinguished with both beginnings and endings (Galton 2006). Events have been identified as a core concept for using spatial information. This view holds that a spatiotemporal perspective is key for bridging different disciplines, and is necessary for addressing many contemporary grand challenges involving, for example, sustainability, biodiversity, climate, energy, and health (Kuhn 2012). Many of the dynamics that users are interested in representing in geographic information systems (GIS) can be understood as events – for example, the different interactions by individuals or vehicles as they move through a city (e.g., traffic jams, or stopping for traffic lights), and the kinds of dynamics that are experienced when a natural hazard such as flooding or wildfires strikes (e.g., setting up flood or fire barriers, or evacuating individuals). The interest in developing data models for events stems from the need for one or more of the following: to understand the range of events for a particular domain, that is, distinguish and classify different kinds of events that hold for a domain; to track the sequence of events over space and time; and to represent patterns and trends involving events, and predict where and when future events will occur.

In an event-oriented data model, events are given first-class status and treated explicitly in the data model. Events are spatiotemporal by their very nature (Grenon and Smith 2004; Worboys and Hornsby 2004). Events are associated with settings defined as spatiotemporal regions where events occur. This gives rise to representations of events that include space–time points, trajectories (space–time paths), and space–time narratives, or histories. Time is an integral part of events, and they may be formalized as being durative or punctual (Galton 2006). When event durations are modeled, this opens up possibilities for describing relations between events, for example, events that are spatially disjoint, overlapping, or simultaneous. Modeling events from the perspective of being durative (i.e., intervals) or punctual (i.e., instants) also invokes discussions of granularity, which is another core topic for geographic information science.

Events may be classified in a subsumption hierarchy or ontology. In this way, higher-level events (e.g., hazard events) as well as more specialized domain events (road closure events) may be classified and formalized. Methods for extending ontologies such that they capture dynamics as well as more commonplace static features are being discussed among researchers (http://vocamp.org/wiki/GeoVoCampSB2014).

Understanding the evolution of events over space–time or describing a sequence of events can be key for comprehending change in a domain; for example, for vegetation monitoring it may be useful to distinguish periods of abrupt versus gradual vegetation changes and the
events that may be driving these variations in vegetation (Kulik, Stewart Hornsby, and Bishop 2011). New details about the evolution of storm events at sea are revealed through the use of geosensor networks such as ocean observing buoys that measure barometric pressure over space–time in the ocean (Rude and Beard 2012). Different kinds of changes in pressure signal the onset or presence of storms and can be tracked as a series of events. In another study, changes experienced by different environmental regions, such as oil spills, may be similarly tracked, in this case as a series of topological events (e.g., appearance and disappearance, splitting, and merging events), and monitored by sensor networks (Sadeq, Duckham, and Worboys 2012). The evolution of such topological events gives insight into behaviors in the domain, offering opportunities for evaluation and management.

The patterns of different kinds of events can reveal important information about the dynamics experienced by a domain. Recently, researchers have applied geographic information retrieval techniques to automatically extract spatiotemporal information from a set of documents. Such extraction can result in a rich representation of spatiotemporal knowledge that is described in text, capturing what events have been reported in the documents and where and when they were reported. In this case, the motivation may be to understand the different kinds of events that occur. These events can be extracted and used in combination with associated spatial and temporal information that is also extracted from the documents to reveal more about the dynamics described within the documents. For example, spatiotemporal information from online news reports about severe storms impacting an area includes domain-related events associated with hazards (e.g., power outage or evacuation events), as well as higher-level events (e.g., more general hazard response or hazard recovery events) (Wang and Stewart 2014). Automated extraction of events from documents followed by geolocating the extracted events makes it possible to produce geographic visualizations of the extracted dynamics, highlighting the spatiotemporal progression of different kinds of events. In other research, an integrated software system called STempo offers users a set of exploratory and interactive tools for tagging or labeling events, analyzing sequences of extracted events, and interacting with visual representations of events detected by the analysis of text or media content (Hardisty et al. 2011; Stehle and Peuquet 2014). Used to extract and analyze, for example, political events from RSS (rich site summary) news feeds, this system offers capabilities to discover sequences of events that may contribute insights into political processes in a big data setting and contribute new methods that impact spatiotemporal knowledge discovery and analysis.

SEE ALSO: Geographic information science; Geotargeted alerts and warnings

References


Data model, F-objects and O-fields

Thomas J. Cova
University of Utah, USA

A data model describes how reality is to be represented in a database. More specifically, it defines the logical entities, attributes, and relationships that comprise the structure of a database. An important step in data modeling is selecting the real-world phenomena to represent. A second step is deciding how to represent these phenomena in a simplified manner to yield a database schema. This process is defined in three steps: conceptual data model design, logical data model design, and physical data model design. The most challenging decisions lie at the conceptual level because this is where the domain semantics are defined. In contemporary database design, the focus at the logical level is on representing the conceptual model as a set of related tables, and the physical level, which addresses storing and retrieving information from a digital medium, is managed by a database management system (DBMS). As data modeling is the means by which the real world is represented digitally, it lies at the heart of database theory and practice.

In a geographic data model (GDM), some of the logical entities in the conceptual model will have a spatial expression. Furthermore, entity pairs will have well-defined qualitative and quantitative spatial relationships. A common entity type in transportation applications is a road, and an example relationship between two roads might be crosses if one road intersects another. In this case crosses is a relationship class, given the many ways in which two or more roads can cross. Geographic data modeling is described using the same three steps as data modeling, but additional design decisions must be made regarding which entity types require spatial representation, as well as how each one will be represented. Raster and vector are the two primary GDM approaches, with the former relying on subdividing geographic phenomenon into nonoverlapping spatial units of equal size and shape (e.g., square, triangle, hexagon) and the latter relying on representing geographic phenomena using points, lines, and polygons.

As noted, a significant step in geographic data modeling is deciding whether geographic reality comprises discrete geographic objects or continuous fields (Couclelis 1992). In the former view, space is viewed as a container populated with geographic objects, each with identity, extent, attributes, relationships, and behavior. This view is very compatible with natural language, as geographic reality is generally described using objects. From physiographic features such as mountains, rivers, and lakes, to cultural ones such as roads, buildings, and cities, people have an inherent tendency to ascribe identity and extent to geographic features. A good example is weather phenomena, as in, “A cold front (object) is crossing the desert (object) and heading towards the mountains (object).” The alternative to an object view is a continuous, field-based view where geographic reality comprises surfaces. In this view, every location is assigned a value drawn from an attribute domain. Temperature, elevation, and land cover are three themes that are commonly viewed as fields, and it is common to view every location on the landscape as having a scalar value for each of these.
attributes. Using weather phenomena again as an example, “The temperature at every location (field) on the earth’s surface is influenced by its elevation (field) and land-cover type (field).”

It is important to recognize that the field and object views are not inherent properties of geographic phenomena. Rather, they are conceptual perspectives that can be adopted based on the application needs of the user or decision-maker. Thus, any geographic theme can be perceived through the lens of either view. For example, an object view of terrain might treat landform features as entities with identity, attributes, and spatial relationships with other features (e.g., an alluvial fan that abuts a mountain or a meandering river that bisects a valley). In contrast to a landscape populated with objects, adopting a field view of terrain involves viewing every location as having a value for each landscape attribute, as in elevation, slope, and aspect. In many cases, the view that a decision-maker adopts is revealed in casual conversation. For example, reference to a region’s highpoint or “the location of greatest annual rainfall” reveals a field view, but reference to a count of geographic features such as peaks, lakes, and canyons reveals an object-based view.

In the case of the field view, geographic phenomena are conceptualized as surfaces where every location in an infinitely dense space (in theory) is mapped to a value drawn from an attribute domain. In most cases, the attribute domain consists of a single value drawn from a nominal, ordinal, interval, ratio, or cyclic scale. A land-cover map is a common example of assigning every location a designated land-cover type (e.g., forest, grassland, impervious), and a temperature map is an example of mapping every location to a value selected from an interval scale (e.g., Fahrenheit or Celsius). In addition to nominal and scalar fields, vector fields refer to the case where every location is mapped to a vector with direction and magnitude. Vector fields are often used in modeling physical processes such as the flow of wind, water, or fire across a landscape, but they are also applied in human geographic settings including modeling migration, economic, and transportation flows. Tensor fields are a unique case where every location is mapped to a matrix that represents strain at that location, and this class of field is widely used in geologic applications such as seismology.

In an object view, geographic space is conceptualized as an infinitely dense set of locations populated by discrete objects that represent geographic features. Objects can be point, linear, or areal in nature and may be contiguous or not. Example features routinely viewed as objects include buildings, roads, rivers, subdivisions, and airports. However, real-world features are not limited to being viewed as one object type. For example, a city may be viewed as a point at one map scale but a polygon at another. Objects can also be combined to create composites, as in the case of a subdivision object that comprises a set of constituent parcel, building, and road objects, each represented differently using points, lines, and polygons. In theory, each finite geographic object still occupies an infinite number of locations because there is no limit on locational precision within its boundary. However, precision limits are introduced in representing objects and fields using raster and vector geographic data models in the context of a finite digital store.

Although the field/object dichotomy is well articulated and widely accepted, there are interesting examples that blur the line between the two views. One example is a density surface where a filter (or kernel) of a fixed size and shape is used to define a count at each location of a particular object type contained within the filter. The term “density” implies an object view, as something is being tallied per areal unit, yet the term “surface” implies a field view. Population
density surfaces are a common example where the underlying object of interest is an individual, but a filter is used to assigns a value to each location in a field based on an estimate of the number of people within the filter (e.g., circular buffer of 1 mile). Thus, a peak in this surface would be the location with the highest population density using this size and shape of filter, and a valley would be a lower-density area. A second case that blurs the field-object line is a fuzzy spatial model of a geographic feature. In a fuzzy spatial model, each location is defined by a function ranging from 0 (not in the object) to 1 (completely in the object) that defines the degree of membership of that location in the object. While the underlying feature of interest is an object (e.g., lake, mountain, valley), the result is a field view of the locations that comprise the object to varying degrees.

Although fields and objects address a key means by which geographic phenomena are conceptualized, representing them in a computer requires a finite data spatial model. While it is common to map a field view to a raster data model and an object view to a vector data model, fields are also represented using a vector data model and objects are represented using a raster data model. There are many examples of representing a field using vector data, as in a soil class map where every location has a soil class defined using polygons. Contours are another example where every location has an elevation (albeit with some interpolation) but the representation relies on the vector model. An example of using a raster to represent objects is using the fuzzy membership example above where every cell in the raster represents the degree of membership of that cell in a larger object using a 0–1 scale.

While the field/object dichotomy is widely accepted, researchers have explored hybrid views for decades. An obvious line of inquiry is to combine or link these two views. One example is objects that possess internal field-like qualities where every location within the object’s boundary is assigned a scalar value (Yuan 2001). A good example of this form of representation arises in the context of Doppler radar data for a moving thunderstorm. In this case, a storm is considered a moving object with identity, extent, and relationships with other objects (e.g., the storm is over the city). Yet pulse-Doppler radar data routinely reveal significant variation within a storm cell regarding spatial attributes such as precipitation type and rate (e.g., intense rain). This representational approach has also been applied to wildfires which are viewed as objects within identity and extent. While object-based statistics are commonly used to describe wildfires, such as percent contained, which requires a well-defined boundary, they also have important internal field-based variation in spatial attributes such as burn intensity and severity. This hybrid view has been dubbed a field-object or F-object.

Another approach to combining the field and object views is to define a field of spatial objects (Cova and Goodchild 2002). Similar to scalar, vector, and tensor fields, an object-field (or O-field) assigns to every location one or more objects, which may or may not include this location in its footprint. One example that arises in hydrologic theory is the case where every location on the landscape has a watershed defined by an uphill catchment or basin. In this case, a field of watershed objects can be defined where every location is mapped to its watershed (object). In moving to a more practical finite context, each location’s watershed can be represented as a set of contiguous raster cells or a polygon. Another interesting example is a viewshed, or the set of locations that can be seen from a given location. If every location is mapped to its viewshed, which may or may not be a contiguous region, this defines a field of viewsheds. To generalize O-fields, every location is mapped to one or
more objects that can be point, linear, or areal in nature (or a combination), and these objects may or may not contain the original field location in their extent. These mappings can be viewed as a spatial hyperlink that explicitly links a location to an object or set of objects. For example, the location of a residence could be linked to each family member’s work, school, or current location, which might be useful in designing household dispersion metrics.

One application area where O-fields are useful is in interactively exploring a large set of spatial alternatives that are too difficult to view in a single static map. This problem can arise in using spatial decision support systems (SDSS) to generate candidate solutions derived from an optimization model. Examples of spatial optimization model classes include location-allocation, corridor, and land-allocation models. The focus in an SDSS context is often identifying a single best solution, or a small set of nondominated solutions if the problem is multi-objective, to present to a decision-maker. This can place too much value on a single best solution, which may include a significant amount of model and data uncertainty. This uncertainty arises from the quality of the input data, the sensitivity of the model to changes in its parameters, or the very definition of “optimal.” For most spatial optimization models, a facility or location can be constrained to being in the solution, which can then be used to generate a field of solutions (point, line, or area), where every location has an optimal solution that includes that location (Cova and Goodchild 2002). This field can then be provided to a decision-maker for interactive exploration to view a larger number of possible solutions to a spatial planning problem. While many of these solutions may not be nondominated in the globally optimal sense, the resulting O-field allows a decision-maker to ask the question, “What is the best solution if we put the facility here?” where the definition of “here” can be controlled by the decision-maker.

A second area where O-fields and F-objects may have application is in representing new types of geographic variation. While scalar, vector, and tensor fields cover much of what we perceive in geographic reality, there are many types of variation that remain unexplored and unrepresented. Examples of this often arise in considering the universe of possible geographic features that might exist in any context. For example, consider the geographic variation that arises from visualizing all possible biological reserves, facility location patterns, watersheds, urban futures, fire-spread scenarios, or spatial interaction flows in a given study area. The results of this thought experiment yield very different types of geographic variation than typically handled by traditional fields, which excels at representing variation such as temperature and elevation. Similarly, the internal variation within geographic objects that is not often represented, such as the temperature variation across a residential rooftop throughout the diurnal cycle, or the variation in albedo across an aircraft’s fuselage, is also an example of where an F-object may have value.

Geographic dynamics also raise very interesting cases where F-objects and O-fields may have application. Consider the case where F-objects and O-fields are allowed to vary over time in location, shape, or internal variation (Goodchild, Yuan, and Cova 2007). For example, a field of biological reserves, where every location has an associated optimal reserve that includes that location, may be allowed to change in shape so as to incorporate the impact of climate change predictions on the elements used to define habitat suitability over time. This would allow a decision-maker to ask questions as to how resilient a given biological reserve might be under climate change. Resilient reserves in this
DATA MODEL, F-OBJECTS AND O-FIELDS

In summary, while fields, objects, and current hybrids are established concepts in geographic information science, there are numerous areas for further research. One example is exploring the deeper structure of fields and objects, particularly as it relates to theories that might unify the two perspectives (Goodchild, Yuan, and Cova 2007). Another example is dynamic geographic modeling, as much of the phenomena in the geographic world of interest are in motion (Galton 2004; Pultar et al. 2010). Spatial uncertainty is another area that holds opportunity for advancing our understanding of F-objects and O-fields, as models of uncertainty in this context would be hybrids of existing uncertainty models. Designing geocollaborative spatial decision support systems is another area where hybrid models may show promise in helping decision-makers visualize and compare a larger number of spatial alternatives. Multiscale geographic objects is another area that may have potential to advance theory and practice, as in the case where geographic features are viewed as different object types at different scales (e.g., a crowd of people as an object at one scale but a set of individuals at another). Finally, the intersection of each of these areas is also a potentially rich area for exploration as in uncertainty in dynamic geographic domains, or scale in spatial decision support systems.

SEE ALSO: Ontology: theoretical perspectives; Representation: complex objects; Representation: fields; Spatial database; Spatial feature classes

References


Data model, moving objects

Atsushi Nara
San Diego State University, USA

A data model for moving objects is a blueprint that describes data of interest related to movements of objects and a set of operations represented in an information or database management system. Designing an appropriate data model fosters conceptual understanding of the dynamic phenomena of moving objects, yields efficient data storage and retrieval, and further facilitates spatiotemporal analytical processes.

Various data models have been proposed to efficiently and effectively store, manage, query, and analyze moving objects data. This entry starts with the description of fundamental elements of moving objects and their abstract data types. Then it reviews data modeling approaches for representing moving objects including conventional spatiotemporal data model, constraint model, moving object data model, semantic trajectory data model, event-based data model, and models for handling big moving object data.

Fundamental elements and abstract data types

Fundamental elements of moving objects can be defined by space, time, and objects, where each element has its own properties represented by values of attributes. Space consists of locations or places, where a reference system is necessary to distinguish positions in space, such as coordinate-based referencing, division-based referencing, and linear referencing (Andrienko et al. 2013). Using a coordinate reference system, positions can be defined by a pair of coordinates (2-D) or a trio of coordinates (3-D), such as global positioning systems (GPS) points. Division-based referencing defines positions based on compartments of a geometric or semantic-based division of the space, which can describe movements based on visited places (e.g., city and state). Linear referencing refers to relative positions of a moving object along linear objects such as streets. For instance, movements can be described by a sequence of street addresses.

The time element in moving objects refers to a moment, or position in time. The Gregorian calendar and standard units of time (i.e., seconds, minutes, and hours) are used as a typical temporal referencing. Natural and human cycles are also common temporal referencing, such as the weekly/monthly/seasonal cycle and the cycles of working hours/days. In addition, a set of two successive time elements can represent the interval or period of moving objects.

Moving objects are entities that change their position over time. In the physical world, an act of moving objects or movement is a continuous phenomenon in space and time, where the spatial continuity is a 2-D or 3-D process and the temporal continuity is a 1-D and unidirectional process. Due to technical and technological limitations, moving object data can only be collected in the form of spatially and temporally discrete objects. Such moving object data may be modeled by treating movements as discrete entities in space and time. On the other hand, a continuous data model, which employs a function to represent movements in order to retrieve any
temporal snapshot location of a moving entity, may be used.

Common abstract data types for moving objects are moving points, moving lines, and moving regions, which are extended from basic vector-based spatial data types: points, lines, and polygons, respectively. The moving point abstraction describes an entity moving around in a 2-D plane or higher dimensional space, for which only the time-dependent location is of interest; examples of moving point entities are pedestrians, cars, animals or packets represented as points. The moving line abstraction is defined as an entity moving around in the plane or higher dimensional space, for which the time-dependent shape and/or the 1-D extent are of interest; examples include traffic congestions on a road network, a shoreline, or a fire front of a wildfire event represented as polylines. The moving region abstraction represents an entity moving around in the plane or higher dimensional space, for which the time-dependent shape and/or 2-D extent are relevant; examples are hurricanes, city/urban expansions, and flocks of migrating birds represented as polygons. Each abstract data type can be represented discretely or continuously (Figure 1).

Spatiotemporal data model

Moving object data consist of three fundamental elements, space, time, and object, which are generally classified as spatiotemporal data; therefore, any existing spatiotemporal data model can represent some aspects of moving objects. The historic development of spatiotemporal models, including snapshot model, space–time composite (STC) data model, simple time-stamping model, event-oriented model, three-domain model, history graph model, spatiotemporal entity-relationship (STER) model, object–relationship model, spatiotemporal object-oriented data model, spatiotemporal unified modeling language (STUML), and moving object data model, has been reviewed by Pelekis et al. (2004). Their review describes the taxonomy in detail in terms of temporal semantics, spatial semantics, spatiotemporal semantics, and query capabilities.

Almost all models can handle discrete changes, while movement and continuous change was not in the priorities in earlier approaches such as snapshot, STC, and simple time-stamping models (Pelekis et al. 2004). In addition, most of the models can capture some aspects of

Figure 1  (a) Discretely changing point, line, and region; (b) continuously changing point, line, and region. Source: Güting and Schneider 2005. Adapted with permission of Elsevier.
the notion of change, including changes in geometry, topology, and/or attribute; however, the earlier models mentioned previously do not support all types of change. More recent approaches support these two concepts (movement and continuous change) and provide functionality that is related to the evolution of spatiotemporal objects; for example, recent data models based on object-oriented and moving object approaches maintain methods and functions for defining the topological relationship and measuring the evolution of spatiotemporal objects in space and time (Pelekis et al. 2004). These two approaches permit the database representation of the complex structure of moving objects, which arise from continuous motion modeling, multiple time line representation, complex behavioral and semantic spatiotemporal queries, and big data management (Pelekis et al. 2004).

Constraint model

The basic idea of the constraint approach for modeling a continuous movement is to finitely represent the movement by means of a Boolean combination of polynomial equalities and inequalities (Macedo et al. 2008). Generally, a constraint model is defined as:

\[ \{ (x_1, \ldots, x_k) \in \mathbb{R}^k | \Phi(x_1, \ldots, x_k) \} \]

where \( k \) is the number of variables, and \( \Phi(x_1, \ldots, x_k) \) is a formula over variables \( x_1, \ldots, x_k \). A constraint model can represent trajectories by considering them as a collection of infinite points connecting a finite number of sample points. For example, with observed moving points, \( P_{(t,x,y)} = \{ (0, 2, 1), (3, 1, 2), (5, 1, 4) \} \), the 2-D movement can be described by the following linear constraint model.

A moving object data model considers time as an integral part of the spatial entities and captures both change and movement of entities represented by abstract data types: moving points, moving lines, or moving regions. A trajectory-based approach is a specific moving object data model that focuses on moving points. A trajectory is defined as the expected route of a moving object from one location to the next. As a simple abstraction, a trajectory \( (T) \) can be a model for a motion path of a moving object consisting of a sequence of sampled time-stamped locations \( (p_i, t_i) \) where \( p_i \) is a 2-D point \( (x_i, y_i) \) and \( t_i \) is the recording timestamp of \( p_i \) defined as (Pelekis and Theodoridis 2014):

\[ T = \{ < p_1, t_1 >, < p_2, t_2 >, \ldots, < p_n, t_n > \} \]

By assuming that the moving object moves with constant speed along a straight line connecting two succeeding observed points in the unconstrained Euclidean space, a simple liner

\[ \{(t, x, y) \in \mathbb{R}^{14^2} | t \geq 0 \land t \leq 3 \land x + y = 3 \land x \geq 1 \land x \leq 2 \lor t \geq 3 \land t \leq 5 \land x = 1 \land y \geq 2 \land y \leq 4 \} \]
interpolation function can find a location \( (p_t) \) at a nonrecorded timestamp \( t \) as follows:

\[
p_t = \left( x_i + \frac{t - t_i}{t_{i+1} - t_i} (x_{i+1} - x_i), \right.
\]

\[
y_i + \frac{t - t_i}{t_{i+1} - t_i} (y_{i+1} - y_i) \right)
\]

where \( x \) and \( y \) represent a coordinate of a 2-D space, and \( t_i \) and \( t_{i+1} \) are two consecutive recorded timestamps. For modeling network-constrained movements (e.g., a moving vehicle on a transportation network), segment-oriented model, edge-oriented model, or route-oriented model can represent the location of an object by segment, edge, or route respectively.

Trajectory-oriented queries are expected to be of great interest in a database that supports trajectory data models. Figure 2 shows examples of trajectory-oriented queries: (i) a timeslice query to find the locations of trajectories at a given timestamp \( (Q1) \); (ii) a spatiotemporal range query to find the portions of trajectories inside a given spatial region during a given time interval \( (Q2) \); (iii) nearest-neighbor queries to find the trajectories that lie nearest to a given point during a given time interval \( (Q3) \), nearest to a given trajectory during a given time interval, and nearest to a given point in every time instance of a given time interval \( (Q5) \); and (iv) a trajectory similarity query to find the trajectories that are the most similar to a given trajectory \( (Q6) \) (Pelekis and Theodoridis 2014).

To efficiently perform queries on moving object data, the design and construction of index structures play a significant role in a database. The main task of an index structure is to ensure fast access to target records in a database on the basis of a search key (Güting and Schneider 2005). One of the earliest techniques for indexing trajectories of moving objects is the 3-D R-tree, which is an extension of R-tree in a 3-D space (Pelekis and Theodoridis 2014). 3-D R-tree for trajectory data treats time as an extra

---

**Figure 2** Examples of trajectory-oriented queries. Source: Pelekis and Theodoridis 2014. Adapted with permission of Springer.
DATA MODEL, MOVING OBJECTS

Spatial dimension and it decomposes trajectories in 3-D segments storing segments’ minimum bounding box (MBB) in its nodes (Peleakis and Theodoridis 2014). One challenging issue with 3-D R-tree is to find the optimal number of MBBs per trajectory. Other proposed indexing techniques for moving object data include, for example, TB-tree for free space movement, and FNR-tree and MON-tree for network constrained movement.

To facilitate effective representation and data management of moving objects in a database, several spatiotemporal database management system (DBMS) prototypes have been developed. SECONDO, one of the first database system prototypes that can handle moving objects, is an extensible DBMS and provides a native implementation of the moving object model (De Almeida, Güting, and Behr 2006). HERMES is another database engine that can process, manage, and analyze moving object data (Peleakis et al. 2008).

Semantic trajectory data model

Semantically enriched trajectories can support a meaningful interpretation of the behavior of moving objects. The process of semantic enrichment involves the annotation of a whole trajectory or parts of a trajectory with behavioral characteristics. For example, trajectory segments of an individual can be contextualized with transportation modes. Similarly, a trajectory from raw GPS data can be converted to a sequence of visited places and then labeled with activities performed. Spaccapietra et al. (2008) introduced a semantic trajectory data model based on the concept of stops and moves. A stop is defined as a part of a trajectory where a moving object stays for a minimal amount of time, while a move is the movement of the moving object between stops. Figure 3 represents a trajectory design pattern, which holds object types for representing trajectories and their “begins,” “ends,” “stops,” and “moves” (Spaccapietra et al. 2008). In the model, stops are described by a group object type B.E.S. consisting of begin, end, and stop objects, while moves are described by an object type Move. Both B.E.S. and Move objects have a life cycle and a geometry. Moreover, a life cycle corresponds to a time interval and a geometry is shown as a point for B.E.S. or as a time-varying point for Move. Two relationship types From and To bearing both a topological constraint and a synchronization constraint enforce the connection between a B.E.S. and a Move in space and time. A B.E.S. and a Move can be linked to a hook spatial object type SpatialOT that represents the corresponding location, such as city, lake, road, and store.

Event-based data model

Instead of focusing on the location-based aspect of movement or quantitative descriptions of motion, an event-based model of moving objects employs the perspective of events associated with movement. Hornsby and Cole (2007) introduced an event-based approach and reveal notable patterns of events. In the model, an event sequence, $E$, is represented as a set of events:

$$E = \{ e_{i,z,v,t(1)}, e_{i,z,v,t(2)}, \ldots, e_{i,z,v,t(n)} \}$$

where $e$ is an event (e.g., departure, arrival), $i$ is the identifier of the object for which the event is associated, $z$ refers to a spatial region or zone where the event happens, $v$ is an event attribute, and $t$ is a timestamp reflecting when the event happened. Events are modeled as being instantaneous (i.e., no duration and mark a change in state of an object). To reveal distinguishing patterns from event sequences, the model uses
pairwise groupings of events or event combinations based on three key attributes: object id, event description, and location. The event combinations identify eight possible patterns of events in transits (Table 1). While P1–P4 are patterns of events arising from multiple sequences that correspond to the movements of single objects, P5–P8 represent patterns of events involving multiple objects. The combination of key attributes of events can identify three notable patterns of events such as repeating (P1), collocating (P2), and reiterating (P5), whereas P1, P3, P4, P6, and P7 can be classified as common patterns in routine transits unless further information is available. P8 is a nonrelated pattern where different objects experience different events in different zones. A repeating event pattern (P1) describes that a transit occurs when a moving object is associated with two events having the same description and occurring in the same spatial region. A collocating event pattern (P3) explains that different event combinations of the same object take place in the same spatial region.
A reiterating event pattern (P5) represents occasions where the same event combinations occur to multiple objects in the same zone within a constraining time window. This pattern captures the convergence of objects into an area or the exodus of objects out of an area such as traffic congestion and evacuation respectively.

Handling big moving object data

Wearable devices equipped with location-aware technologies are rapidly becoming a part of everyday life. These devices enable us to collect data about moving objects in large volume, high frequency, and high variety. A massive amount of data, a very fast update cycle, and a heterogeneous data structure pose new challenges of how to efficiently and effectively store and manage such big moving object data in database systems.

One solution is to utilize the high performance computing (HPC) in the form of parallel and/or distributed databases, such that data processing and computation can operate simultaneously on multiple moving objects, for example. A parallel database system is implemented on a tightly coupled multiprocessor and increases performance through parallelization of various database tasks, such as loading data, building indexes, and evaluating and executing queries. There are three architectures for passing information in parallel database systems: (i) shared memory: any processor has access to any memory module or disk unit; (ii) shared disk: any processor has access to any disk unit; and (iii) shared nothing: each processor has exclusive access to its main memory and disk units. A distributed database system is a collection of logically interrelated databases distributed over a computer network usually in a shared-nothing architecture.

An alternative approach to handle big data is to utilize MapReduce, a simplified programming framework with an associated implementation for processing and generating large datasets using a scalable distributed file system. MapReduce consists of two processes: mapping and reducing. Mapping is a process that restructures an input dataset of key-value pairs into groups of intermediate key-value pairs. Reducing is a process that takes an intermediate key and a collection of values to generate a smaller set of tuples. Tan,
Luo, and Ni (2012) developed a scalable big spatiotemporal data storage system, CloST, to support data analytics using Hadoop, which is an open source implementation of MapReduce along with the Hadoop distributed file system (HDFS). A data model used in CloST consists of three core attributes: object id, location, and time along with other attributes (Oid, Loc, Time, A₁, ..., Aₙ). CloST introduces a three-level hierarchical partitioning approach based on the core attributes to enable efficient parallel processing of all object queries with any possible spatiotemporal ranges (Figure 4).

Parallel SECONDO is another data model that combines Hadoop with a set of SECONDO databases to perform queries and analyses on massive sets of trajectory data in the cloud (Lu and Güting 2014).

Data models for moving objects are, by definition, spatiotemporal and multidimensional. Hence, the data object representation involves semantic, spatial, and temporal components. Various data models have been developed to support management, query, and analysis of moving objects. Now, ubiquitous location-aware mobile devices have been generating a deluge of
moving data that can be benefitted greatly with HPC and MapReduce to meet data processing, management, and computational challenges.

**SEE ALSO:** Data model, event-oriented; Data model, object-oriented; Indexing; Qualitative spatial and temporal representation and reasoning; Spatial database; Spatiotemporal analysis; Time geography and space–time prism; Trajectories: analysis

**References**


Levels of representation

The representation of space involves three levels of abstraction, including conceptual models, spatial data models, and spatial data structures. Conceptual models describe the conceptualized space, that is, the essential spatial components and the relationships between them. For example, a population can be conceptualized as individuals at various locations, subpopulations within administrative regions, or population density varying continuously in space. The relationships between them refer to those between locations, between regions, and within a continuous space. At the second level, the spatial data model formalizes the spatial components and their relationships in terms of basic geometry (e.g., points, lines, polygons, and grids) and topology. The spatial data structures at the third level implement these components and relationships in a computing environment.

There are various nomenclatures used to refer to the three levels, but it is widely accepted that the spatial cognition process of humans is at multiple levels and the number of levels is often three, as discussed in the literature. The conceptual models, spatial data models, and spatial data structures represent three distinct yet sequential perspectives. That is, a naive observation, a formalization of the observation, and an implementation of the formalization.

Spatial sciences are interested in the spatial distribution and temporal dynamics of phenomena, be it environmental or human behavioral. Unlike those models that directly describe environmental and behavioral processes, spatial data models are models of data. They focus on the representation of spatial forms of phenomena, such as geometry and topology.

One limitation is essential to the formulation of these data models. The computing environment is finite and discrete; a geographic space must be divided into a finite number of discrete pieces to allow for the implementation of a data model. Because of this limitation, spatial data models focus on the design of the discretization and the relationships between the discretized pieces.

Three types of information are associated with a spatial data model: location, attributes, and topology. A location is represented by a pair of coordinates. Attributes describe properties of partitioned pieces. Topology, on the other hand, represents the spatial relationships between these pieces. There are two classic data models: vector and raster. These are two fundamentally different types of data model, representing two different ways of dividing space according to the two commonly recognized dichromic views of space: object and field views.

Object and field views

Field is a fundamental concept in science, especially in mathematics, physics, and other science disciplines that are built on mathematical and physical principles. The concept of object came
DATA MODEL, OBJECT-ORIENTED

with the rise of object orientation, the current computing paradigm since its introduction in the late 1980s. When the two concepts were extended to the spatial sciences and paired together, the object-field view offered a profound framework for conceptual models of spatial phenomena. With the object view, reality is conceptualized as discrete objects littered in space (hereafter referred to as “spatial objects”). In contrast, the field view conceptualizes the reality as a continuous plane with variations at an infinite number of locations. The object versus the field view has been used to guide the conceptualization of spatial phenomena.

It is easy to conceptualize certain phenomena as discrete spatial objects and other phenomena as continuous fields. A continuous phenomenon can be divided indefinitely without changing its essential nature, while a discrete phenomenon most likely cannot be divided without altering its identity. For example, the water in a bucket may be continually divided into halves and yet remain water, but once a bucket is divided into two halves, neither half is a bucket any more. The object and field views are inclusive. A phenomenon can be represented as either a spatial object or a field depending on the spatial scale of the observation, the purpose of a study, and convention. The two conceptualizations can coexist as well.

In addition to easily recognizable spatial objects and fields, spatial regions are a third phenomenon that has been discussed in the literature. Spatial regions have dual qualities. They appear to be discrete as spatial objects, but are not quite as independent because they are always parts of a continuous field. The spatial objects, spatial regions, and fields are considered conceptual models of spatial phenomena. They are not directly tied to either of the two spatial data models, vector or raster. It is first necessary to understand the characteristics of spatial objects, spatial regions, and fields before selecting the ideal data model to formalize them.

Characteristics of spatial objects, regions, and fields

The characteristics of spatial objects, spatial regions, and fields have been actively discussed in the geographic information science literature. Spatial scale, boundary, attributes, process, and mobility have been used to characterize the three types of conceptual models as well as to distinguish between them.

Spatial scale is an essential characteristic of all spatial phenomena. From an ontological perspective, this characteristic uses the human body as the reference scale to categorize spatial phenomena into small and large objects, as well as small and large fields. The small objects include everyday tabletop objects that are smaller than the human body and can be manipulated. The large objects are larger than the human body, but can be easily conceptualized as objects, such as buildings. Spatial scale is applicable to spatial regions when they are equated with spatial objects. Fields can be small or large although they are continuous. The small fields are those that can be experienced by humans, although not completely all at once, such as a landscape. Large fields are an extension of small fields beyond the range of direct human experience, such as space beyond the Earth. These categories of objects and fields have important implications in identifying appropriate conceptual models for spatial phenomena.

Boundary is also a common characteristic of spatial phenomena, especially for those that are conceptualized as spatial objects and regions. Boundaries are as essential as the internal content to the ontological makeup of spatial objects and spatial regions. The aforementioned small and
large objects, along with spatial regions, are commonly considered as well bounded. There are three types of boundaries for spatial objects and regions: bona fide boundaries, fiat boundaries, and indeterminate boundaries. The bona fide boundaries include those that exist physically and can be precisely located in space, such as the boundary between a lake and land. The fiat boundaries include those that do not exist physically but can be precisely placed in space, such as administrative boundaries. The third type refers to indeterminate boundaries that physically exist but cannot be located precisely, for example, the boundary between “tall grass prairie” and “short grass prairie” or between “Midwest” and “Northeast.” With boundaries, spatial objects and spatial regions can be readily identified. Fields are spatially extended without a boundary or the boundary is not a concern.

Attributes are a third characteristic. From the ontological viewpoint, things are distinguished from one another by their unique attributes. These attributes can be spatial (e.g., area, adjacency) or nonspatial (e.g., magnitude, type). The distinction between spatial objects or between spatial regions is assessed by the attribute values. These include, for example, the magnitude or threshold of attribute values (e.g., climate zones), the homogeneity or dominance of attribute values (e.g., land-use and land-cover zones), or the spatial association of attribute prototypes (e.g., soil associations). Fields, on the other hand, are viewed as a mapping between attributes and an indefinite number of spatial locations. Each location is associated with attributes, and these attributes vary continuously across the space.

Process is another essential characteristic of spatial phenomena. According to ontological principles, all things change. Process represents changes to the state of spatial objects, spatial regions, and fields. Similar to attributes, processes can be spatially oriented (e.g., direction, speed) or nonspatial (e.g., magnitude, type). There are numerous examples of identifying spatial objects and spatial regions by environmental and socio-behavioral processes, such as the delineation of watershed and traffic generation zones. Similar to attributes, each location of field is associated with processes, and these processes vary continuously across the space. Because process can be represented as an attribute change, the attribute and process characteristics are often paired together in the identification of spatial objects, spatial regions, and fields.

Mobility can be considered as a special case of the process characteristic, but because mobility implies the independence of objects from locations, this characteristic stands in its own right. While moving, objects maintain their identity, properties, and behavior. Many spatial objects can be mobile or moved, such as an animal or a vehicle. When an object moves, all parts of the object move together. Spatial regions can also be mobile, but it is their attributes and processes that are passed to the adjacent regions, not their own parts. The same “pseudo” mobility applies to fields. A field does not move, nor does its parts. It is its attributes and process that move across space. A close analogy is the movement of an ocean wave. When the water molecules make a vertical motion, but have zero horizontal displacement, the attributes (the wave crest) and process (force on the shore) of a wave move forward. Another example is the movement of a “wave” in a sports stadium. By a sequence of vertical motions of the members of an audience (standing up and sitting down), a wave moves around the stadium, while its parts (the audience members) do not change their locations. The style of mobility helps distinguish between spatial objects, spatial regions, and fields.
In summary, spatial objects, spatial regions, and fields possess any number or combination of the aforementioned characteristics. The spatial objects exist in both small space and large space, have discrete boundaries, properties, and processes, and are mobile (or movable). Spatial regions have definable boundaries, attributes, processes, and can be mobile. They are extracted out of a continuous field and yet they are often conceptualized as objects. They carry dual qualifications but do not comply fully with the discrete definition of objects or the continuous definition of fields. The dual qualification of spatial regions presents challenges as well as flexibilities in their representation. Fields have large scales, no boundaries, continuous change of the attributes and processes, and pseudo-mobility.

Spatial networks are a unique form of spatial phenomenon (e.g., stream and transportation systems). A network consists of nodes and links. Unlike single-formed spatial phenomena, such as polygon-formed phenomena or line-formed phenomena, the geometric makeup of networks requires two forms, points and lines. The dual formed nature of networks adds unique complexity to its conceptualization. Of spatial object, spatial region, and field models, spatial networks most closely resemble fields. They are 1-D fields embedded in a 2-D space. Networks can be fields because they exist in small and large space; they are continuous although in a linear form, carry attributes and processes, but occur only in the network and not in the intervening spaces, and move following the pseudo-mobility principle. The formal model of fields is applicable to networks, where there is an infinite number of locations constrained to this linear field. The attributes can be any of the types defined for 2-D fields. Linear segments extracted from a network are equivalent to those spatial regions in a 2-D field.

Due to their dual form, it is easier to extract spatial regions out of networks than from their 2-D counterpart. The spatial regions in networks can be well bounded because nodes and links bound each other. Both nodes and links have their own attributes and processes. The attributes associated with links are consistent with those of 2-D fields, while the direction attribute of networks is more prominent. The attributes and processes of networks can vary on a continuous basis along a link, but it is more often the case that the attributes and processes associated with a node and a link are considered homogeneous. Given these traits, it is conceptually easier to conceptualize individual nodes or links as objects than spatial regions in 2-D fields.

Spatial objects, spatial regions, and fields in geography

Geography deals with a variety of spatial phenomena. Some of these can be easily thought of as spatial objects, some as fields, and others as spatial regions. The five characteristics discussed already, namely scale, boundary, attributes, process, and mobility, help further categorize these spatial phenomena into different types of spatial objects, spatial regions, and fields. This typology helps the selection of appropriate spatial data models for the intended geographic study. Five typical categories of spatial objects, spatial regions, and fields are identified and discussed here.

1 Individuals. Individuals exist in both small- and large-scale space, and have clear boundaries, attributes, and processes. They can be mobile or sedentary. Mobile individuals are independent of spatial locations, while the sedentary individuals are bound to their locations. The most typical examples of mobile individuals are individual animals.
or vehicles, while the typical examples of sedentary individuals include individual plants and buildings. These individuals are characterized by their unique identity, attributes, and behavior (including the mobility). Individuals, both mobile and sedentary, are normally conceptualized as objects. Both types are fundamental subjects of many geography studies.

2 Masses of individuals. This category is an extension of the category of individuals. As the constituents of the mass, the individuals are identifiable but are small in size or large in quantity, or both. These constituents can be mobile or sedentary, relative to the temporal scale of a study. The most common examples of the masses of mobile individuals include an animal population composed of mobile individual animals or a traffic flow composed of individual vehicles. Examples for the masses of sedentary individuals include vegetation composed of individual plants or an urban landscape composed of individual buildings. The collective attributes and processes across space, and the mobile or sedentary nature of the masses, mean they are conceptualized as fields. Their field-like traits are attractive for the environmental and sociobehavioral subdisciplines in geography.

3 Regions of individuals. This category is an extension of the masses of individuals or, alternatively, an extension of the individuals. These are spatial regions extracted from the continua of the masses of individuals or an aggregation of the individuals. A region can be mobile or sedentary, depending on whether the location of the region as a whole is mobile or not. A school of fish and traffic flows are examples of mobile regions of individuals. Plant biomes carved out of continuous vegetation and electoral districts extracted from the spatial distribution of a population are examples of sedentary regions. The unique identities, attributes, and processes associated with the regions, either mobile or sedentary, often mean they are conceptualized as objects and are of great importance to geographers.

4 Continuous masses. Phenomena of this type are spatially extended, continuous, and spatially varying. The attributes and processes of this type vary across space, and the mass may be dynamic or static. Temperature and population density can be examples of dynamic masses (note that population density refers to a continuous surface expressed as a rate, while the population example used in the masses of individuals refers to a collection of discrete individuals). Land and roads are examples of sedentary masses in 2-D and 1-D spaces, respectively. The continuous form and the dynamic or static nature of the attributes and processes of this type are of great interest to geographers. Both continuous sedentary mass and continuous dynamic mass are typically conceptualized as fields.

5 Regions in mass. This category is an extension of the continuous masses because the regions are extracted out of the continua. Similar to regions of individuals, the constituents of the regions in mass can be mobile or sedentary. Hurricanes and pollution plumes are examples of mobile regions. Land-use and land-cover areas extracted from the continuous land and high-accident sections of a road are examples of sedentary regions in a mass. Both mobile and sedentary regions are often conceptualized as objects. The unique attributes and processes of these regions are important for research in geography.
DATA MODEL, OBJECT-ORIENTED

The five categories discussed above represent a spectrum of spatial continuity with the most discrete type, objects, on one end and the most continuous type, fields, on the other, with both gradually transforming to spatial regions in the middle of the spectrum. The first three types begin with phenomena that are conceptualized as spatial objects and then aggregated into spatial regions and fields. The other two types begin with continuous fields and are then discretized into spatial regions. The mass of individuals as a field can always be decomposed into individuals.

Spatial phenomena in geography can be placed anywhere along this spectrum of spatial continuity, but not necessarily in distinct categories. The same phenomenon may qualify in a number of categories depending on the objective, scale, and conventional treatment of a study. For example, a population can be conceptualized as either mobile or sedentary individuals, a mass of mobile or sedentary individuals, regions of individuals, a continuous mass if presented as population density, either dynamic or sedentary, or regions in the mass of population density.

The characteristics of these spatial objects, spatial regions, and fields help select the appropriate spatial data model for the intended study. The next section describes the design principles of the two classic data models, the vector data model and the raster data model, to help assess the compatibility between conceptual models and data models.

Spatial data models

Both vector and rater data models formalize the conceptual models of spatial phenomena. They both divide the space into a finite number of discrete pieces, but by fundamentally different principles.

Vector models divide the space into spatial “features,” which are implemented in the geometric forms of points, lines, and polygons. Each feature is associated with the three types of information – location, attributes, and topology – as previously discussed. The location is recorded as the coordinates of points, turning points of lines, and outlines of polygons. Each feature has multiple attributes. The vector data model can explicitly represent topology between the features, such as which polygonal features are on either side of a line feature or which point features are at either end of a line.

Semantically, the geometric features correspond well to spatial phenomena, whether they are conceptualized or physically present. This allows the vector data models to represent these phenomena in a meaningful way, in terms of their geometric form, attributes, and topology. Vector data models are mostly advantageous for spatial phenomena that have a clear identity, are discrete, internally homogeneous, and static.

Raster models divide the space into a matrix of regular cells (or points). These cells have a predetermined and identical geometry in terms of their shape and size. The coordinates of a cell are implicitly referenced by its row and column positions. A single attribute value is associated with a cell. The adjacency between cells in the four cardinal and the four diagonal directions is explicit.

In contrast to the vector data model, the regularly shaped cells do not correspond directly to any spatial phenomena. Usually an assemblage of a number of cells does. This assemblage is not recorded in the raster data models. Consequently, the attributes of the spatial phenomenon and the topology between phenomena are not recorded either. Because the cells, not the phenomena, are the basic spatial unit in the raster data model, raster models are advantageous for phenomena that are continuous, heterogeneous, and dynamic. The explicit adjacency between cells...
DATA MODEL, OBJECT-ORIENTED

is particularly effective for supporting localized simulations that often query information about areas adjacent to a local cell.

Spatial data models are often implemented in spatial databases. The databases are developed as generic tools to manage the data and to support their usage. These tools are mostly for basic spatial operations, not for the modeling and analysis of specific environmental and behavioral processes.

There is no information in the database about what phenomena these data models should represent. The separation of the spatial data models from the phenomena that they represent allows the spatial databases to be flexible, a necessary quality to support a wide range of applications. This separation, however, makes it critical for users to choose data models appropriate for the intended study. There are almost always multiple spatial data models that can be used to represent the same phenomenon. The choice of the data model can have significant impact on the outcome of a study. An appropriate choice depends on many factors, but the first and foremost is how a phenomenon is conceptualized and how it subsequently is represented.

Object-oriented implementation of data models

Today’s geographic information system (GIS) data models are implemented in an object-orientation environment. As the current computing paradigm, object orientation is based on several basic principles that closely resemble the human cognition process. The most important are the encapsulation and composition principles. With encapsulation, an object has a unique identity, attributes, and behavior. Attribute values describe the state of the object, whereas behavior can change the state. These identities, attributes, and behavior are encapsulated within the object. Here the behavior can be equated to the processes discussed earlier. With the composition principle, objects can play different roles when organized in several different ways: as a subclass in a class hierarchy, an integral part of an aggregated entity, or a member of an association set.

Because these principles are intended to resemble the general human cognition processes, they are independent of specific domain applications. There are several hidden assumptions behind the application of object orientation. For example, it is the responsibility of the domain users of object orientation to identify their objects and that the objects should be identifiable. Further, because field-like phenomena are not explicitly considered in the object orientation, it is the responsibility of the domain users to conceptualize fields into some types of objects. These assumptions brought both conceptual advantages as well as difficulties in the design of object-oriented GIS databases.

The spatial data models implemented in the object-orientation environment bear the same geometric form as that in the pre-object-orientation era. Geometric primitives, such as points, lines, polygons, their derivatives in the vector data models, and grids of cells in the raster data models are implemented as software objects. Attributes and functions (equivalent to process) are encapsulated within the software objects and are organized into various hierarchies, aggregates, and associations. Note that these software objects should be distinguished from the spatial objects discussed earlier. The spatial objects in the earlier sections are discussed in the context of conceptual models to represent how spatial phenomena are conceptualized by humans. The software objects in this section are discussed in the context of the technical implementation of the spatial data models.
In the vector data models during the pre-object-orientation era, locations were the base of the data model. Attributes and topology were attached to the location. In a current object-orientation environment, the software objects are built on their identifiers. Location is treated as an attribute, such as the “shape” attribute in ArcGIS™. With this approach, changes in location can be easily implemented by updating the location attribute, while the identity, other spatial and nonspatial attributes, and functions of the software object remain intact. This change is one of the major improvements of object-oriented vector models over traditional ones and the change has significantly increased the utility of GIS in many applications.

The object-oriented raster data model keeps the same geometric form of cell arrays as that in the pre-object-orientation era. In the object-orientation environment, the cells may be aggregated into different levels of blocks according to the needed granularity. Each block can be treated as a software object. While technically feasible, the conceptual advantage of this implementation is inconclusive. Because object orientation has become the software industry standard since the early 1990s, the pressure to comply with the standard might have compromised the conceptual considerations in the implementation of object-oriented raster data models.

The five types of conceptual models of spatial objects, spatial regions, and fields discussed earlier can be formalized by various vector and raster data models, although to various degrees. The appropriate selection of spatial data models for the representation of spatial phenomena in a study depends on the compatibility between the characteristics of conceptual models and the design principles of data models along with their object-oriented implementations.

GIS data models for spatial objects, regions, and fields

Spatial data models for spatial objects

Spatial objects include the category of individuals discussed earlier. The vector formalization of spatial objects is rather straightforward, as these data models are designed to focus on the unique identity, discrete forms, and the encapsulated attributes and functions. These principles readily support the representation of scale, boundary, attributes, processes, and mobility of spatial objects.

The attributes of the vector data model can support many characteristics of spatial objects. These characteristics can be represented as their internal (geometric, biophysical, socioeconomic, etc.), background environment, and spatiotemporal attributes, for example. The representation of the scale and boundary can be accommodated by internal attributes, such as geometry. The representation of other attributes is straightforward. The processes of spatial objects may include the actions of individuals, the interactions between them, and their interactions with the background environment. These processes can be represented as an attribute change. Similarly, the mobility of an object can be represented by updating the spatial and temporal attributes.

For raster data models, a single cell or an assemblage of connected cells may be used to approximate the shape of a spatial object, but the representation of boundary is implicit. All cells have an attribute value and can support processes. However, the attributes and processes are associated with a cell, which normally does not correspond to a meaningful spatial object. Additional programming tools are normally needed to identify the spatial objects and support their related functions. The mobility of spatial objects can be represented by the pseudo-mobility...
approach. This is done by a sequential change of an attribute for a cell or an assemblage of cells along the path of movement.

Raster data models are not readily compatible with the spatial object conceptual model. A square cell does not resemble most spatial phenomena and the lack of identity for meaningful spatial phenomena prevents the raster models from being a good choice for formalizing spatial objects.

Spatial data models for spatial regions
Spatial regions include two categories, the regions of individuals and regions in mass. A spatial region is more likely to be conceptualized as a spatial object rather than a field. The design principles of vector data models, for the most part, are conceptually compatible with these regions, especially those taking polygonal and linear forms. The vector representation can support the scale, boundary, attributes, and processes associated with spatial regions as they would for spatial objects. However, the representation of mobility is an exception. Unlike spatial objects that are independent of their background environment, spatial regions are bound to the field. The representation of their mobility needs to follow the pseudo-mobility approach typically used for representing the mobility of fields.

Raster formalization of spatial regions faces similar challenges to that for spatial objects. This is primarily due to the inability of the raster data model to identify a meaningful spatial region. On the other hand, the raster data model has an advantage when representing the mobility of spatial regions due to its ability to represent continuous, heterogeneous, and dynamic phenomena.

The dual quality of spatial regions, that is, they are discrete spatial objects but are always parts of a continuous field, makes them partially compatible with vector and raster data models.

Spatial data models for fields
Fields include the categories of masses of individuals and continuous masses. Both vector and raster data models can be conceptually compatible with a field. The design principles of both data models can support the characteristics of a field, that is, the spatial continuity, the spatial variation of attributes and processes, and the dynamics of these attributes and processes. Six data models have been discussed specifically for the representation of fields. These are polygons, triangulated irregular networks (TINs), contours, cell grids, point grids, and irregular points. Of these, polygons, TINs, contours, and irregular points are vector data models, while cell grids and point grids are raster models.

Of the four vector models, the polygon and TIN models formalize a field as contiguous regions across a planar space. Although each polygon has explicit boundaries, there is no overlap and no holes between polygons in the space. The spatial continuity of a field can be appropriately represented. Other characteristics of a field, such as the spatial variation of attributes, processes, and their dynamics can also be supported. The TIN model is considered a special kind of polygon model.

Contours and irregular points, although in a vector form, are intended to formalize a continuously varying field. The contour lines and irregular points represent attribute values of a field at sampled locations, either along the contour lines or at certain point locations. Attribute values for the rest of the field can be estimated based on those values observed at the sampled locations. In addition to spatial continuity, these two models
DATA MODEL, OBJECT-ORIENTED

can support the representation of spatially varying attributes and processes, and their dynamics through a sampled form.

The two raster data models, cell grids and point grids, bring their advantages into full play when used to formalize fields. Their simple form, the explicit representation of attributes and their spatial variation, and the support for spatial functions allow the grid models to effectively represent the spatial continuity, spatial variation of attributes and processes, and their dynamics of fields. This is especially so for those continuous fields without explicit regions, where each cell is a part of a continuous field and not part of any spatial object or region.

Composite structures

Both raster and vector data models have been implemented in a layered structure. Raster data models are inherently limited to a single type of data (scalar, vector, or tensor) in the layer. While vector data models are confined within a planar space and limited to a single theme and geometry. For example, a school bus routing project may involve the student population in census block groups (polygons), the bus route (lines), and bus stops (points). In many GIS software packages, this project must be prepared with a minimum of three layers, a polygon, a line, and a point layer, respectively, although all belong to a single theme. Today, it is both theoretically and technically feasible to develop a “composite” structure to support the assemblage of different geometries or themes in a single layer. For the example above, all three types of geometric features can interact in a single layer for the intended study.

A multigeometry and multitheme “object field” has been proposed. Vector data of various geometries and themes are integrated with a continuous raster field with a different theme. Moreover, spatial objects, regions, and fields can be organized into hierarchies, aggregates, or associations according to the needs of the study. A composite structure requires a more sophisticated topology between different types of geometries. The nine-intersection system has provided a notable theoretical framework to support the complex topology between points, lines, and polygons. Spatial data models are a classic topic in geographic information science. Since the 1980s, the research in data models has experienced enormous advancement. When both this research and the GIS industry matured, the spatial data models implemented in proprietary software packages stabilized. Continued theoretical development in the research community and the technical development from the GIS software industry are critical for advancing this composite structure or for the development of other advanced spatial data models.

SEE ALSO: Data model, F-objects and O-fields; Data model, moving objects; Data structure, raster; Data structure, vector; Fiat and de facto objects; Geographic information science; Geographic information system; Ontology: theoretical perspectives; Representation; Representation: fields; Spatial database

Further reading

Couclelis, H. 1992. “People Manipulate Objects (but Cultivate Fields): Beyond the Raster-Vector Debate in GIS.” In Theories and Methods of Spatio-temporal Reasoning in Geographic Space, edited...
“Practically overnight, access to terabytes of geographical information, much of it in three dimensions, has changed the way people work, live and play” (NGAC 2009). As geographic data have become ubiquitous and widely used, more and more people are relying heavily on spatial information to support a wide range of applications. Geographic information has evolved from being shared exclusively via paper maps to data shared on networks, first exclusively within the domain of geographic information systems professionals, and, only recently, beyond the realm of professionals to anyone with a web browser or a smartphone. This invaluable ecosystem of digital maps, spatial data, and location-aware applications and devices is made possible through the leveraging of many different standards. Standards-setting bodies like the Open Geospatial Consortium (OGC), the World Wide Web Consortium (W3C), and the International Standards Organization (ISO) have enabled the seamless and efficient discovery, exchange, and exploitation of geographic information through the development and advancement of numerous internationally recognized geospatial technology standards.

Data quality standards enable data producers to provide quantitative and qualitative information regarding the accuracy and validity of the data they make available to potential consumers. In turn, data consumers can use this information to determine the data’s “fitness for use.” The concept of fitness for use highlights the importance of the consumer’s perspective of data quality, as it is ultimately the consumer who must determine whether any given dataset can or should be used for a specific purpose. When using any source of data that originates from other individuals or organizations, data consumers should review a standardized description of the quality of that data in order to best support defensible and sound decision-making. Data quality standards have evolved over time to help ensure that producers are able to sufficiently describe their data so that others can understand any dataset’s strengths, weaknesses, and limitations, as well as how to access and process the data.

In the field of geography, and specifically with regards to the use of geographic information, the challenges associated with documentation and appropriate interpretation and use of information on data quality have increased over time as geospatial information has become widely available and more heavily used. In the past, the primary consumers of spatial data were geospatial science and technology experts and geographic information systems (GIS) practitioners. Techniques and practices for measuring and documenting data quality evolved primarily in support of the technical requirements of geographic data producers and a relatively specialized audience of data consumers.

Specific measures of spatial data quality typically vary, based on the type of spatial information being described. There are specific parameters often used to report the accuracy of computer-classified satellite imagery, for example, which differ from those tailored particularly to enable producers to measure and
document the quality of vector-based geographic features such as roads. Sensors that produce and collect observational data with geographic parameters sometimes have their own specific measures of data quality that are useful to data consumers. A number of standards initiatives have been advanced to help document data quality in ways that are consistent and repeatable across data sources and providers.

Standards bodies including the US Federal Geographic Data Committee (FGDC), the French Institut Géographique National (IGN), Open Geospatial Consortium (OGC), the International Standards Organization (ISO), and others have developed data quality standards that enable data producers to describe geographic data quality using relatively similar constructs. The ISO standard ISO 19157:2013, “Geographic Information–Data Quality,” builds on these efforts and standardizes the components and structures of spatial data quality measures while enabling the documentation of data quality to be captured from the perspectives of both data producers and data consumers.

Information on data quality can be documented in metadata records associated with the data themselves. In the United States, the Federal Geographic Data Committee released the Content Standard for Digital Geospatial Metadata (CSDGM) in 1998, with data quality attributes based on the 1992 Department of Commerce Spatial Data Transfer Standard (SDTS). Today, many organizations across the United States and throughout North America utilize the CSDGM standard when publishing geospatial metadata and documenting quality. Many of these concepts have contributed to the development of an international metadata standard developed through the collaboration of many nations under the umbrella of the International Standards Organization (ISO). ISO 19115:2003, “Geographic Information–Metadata,” defines a standard schema that allows data producers to consistently describe data quality using a set of structured fields. ISO 19115 and other commonly used metadata schema typically include a structured methodology for documenting spatial data quality in a number of general categories including the following.

- **Lineage/source/processing steps:** describes the origin and provenance of the data, including a description of how the data were transformed or otherwise processed to produce the final product.
- **Spatial/positional accuracy:** measures how close (in three-dimensional space) locations in the dataset are to their actual locations.
- **Attribute accuracy:** an assessment of the accuracy of the identification of entities and assignment of attribute values in the dataset.
- **Semantic accuracy:** the quality with which features are described in accordance with the selected data model. Semantic accuracy is often used to characterize geospatial datasets generated from aerial photographs or satellite imagery as a way to depict the relative certainty that spatial features captured from the imagery (e.g., roads and wetlands) are categorized correctly.
- **Temporal accuracy:** measures of the accuracy of time measurements present in the dataset.
- **Completeness:** the relative degree of presence or absence of geographic features and their attributes.
- **Logical consistency:** degree of adherence to logical rules of data structure, attribution, and relationships between data elements.

The practice of using formal metadata records to document data quality has been a long-standing requirement within many government organizations that produce and manage important geospatial data assets. In the United
DATA QUALITY STANDARDS

States, Executive Order 12906 (1994) directs all federal organizations to develop metadata records for the geospatial data they collect or produce, either directly or through indirect mechanisms, including the issuance of grants to other organizations. Under Regulation 1205 (2008) of the Infrastructure for Spatial Information in the European Community (INSPIRE) Directive, metadata requirements and policies are clearly defined for government organizations in the European Union. Other nations and units of government have similar policies and requirements in place, and, as a result, metadata that adequately and accurately documents the quality of spatial data is available from many governmental organizations. Additionally, many governments have policies and laws in place that require government data producers to adhere to a set of standard practices and procedures to maximize data quality and report on certain metrics. In the United States, the Information Quality Act was passed into law in 2001, and requires the Office of Management and Budget to issue government-wide guidelines that “provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by Federal agencies” (Office of Management and Budget 2002).

These types of policy directives have ensured that, for many spatial datasets produced and maintained by governmental organizations around the world, metadata records have been made available to help users understand the relative quality of data from these providers so they can make an appropriate determination as to the fitness of the data for their specific purposes. However, it is increasingly important to be mindful of the fact that private companies are generally not required to comply with well-known data quality standards or to publish formal metadata with their data offerings. The geospatial community has grown and changed rapidly, and the private sector has taken on a primary role in the production and dissemination of spatial data. At the same time, the use of GIS to support decision-making has grown across all sectors of the economy, as has the relative degree of reliance on data from external providers in these efforts. This situation is perhaps summarized best by Goodchild (1995, 424): “GIS is its own worst enemy: by inviting people to find new uses for data, it also invites them to be irresponsible in their use.”

The long history of publishing spatial data for open access, combined with the relative ease with which tremendous volumes of geographic information can be quickly and easily assembled and analyzed using geospatial technology, brings particular challenges with regards to the appropriate use of data. To make matters more complicated, new types and sources of geographic information are becoming available at an accelerating rate, and the tools to exploit that data quickly and easily via the construction of web “mashups” and other methods are widely available and heavily used. With the rapid advancement of the “Internet of Things” and sensors of all types and sizes producing and sharing data in volumes that were, only recently, unimaginable, the reality is that much of this data will not be sufficiently characterized with respect to quality. This presents challenges for users who are faced with trying to utilize this data for decision-making without fully understanding the consequences of using data of initially unknown quality.

As the Internet has evolved beyond a one-way communication medium toward an infrastructure that is participatory and collaborative via Web 2.0 concepts and technologies, opportunities now exist for data consumers to publish and share information on the quality of data they utilize. While outside the scope of evaluating data
quality per se, perhaps the best-known example of this type of consumer-based evaluation of quality is the use of star rankings for products made available through Amazon.com and other online commerce sites. These rankings and the comments that users provide build a valuable information resource that helps subsequent consumers determine whether the product will suit their needs or interests. Examples of this type of consumer-based feedback with regards to data quality and utility are becoming more commonly incorporated into open data catalogs. The City of Asheville, NC, for example, allows users to rate and comment on datasets accessed in its open data catalog. One of the largest commercial sharing environments for geospatial information, Esri’s ArcGIS Online, also incorporates capabilities to facilitate user feedback and rankings on many of the datasets that are available through this catalog.

This feedback can be of substantial benefit to a data consumer who is evaluating the fitness of use of a given dataset for their purposes. Moreover, even when structured, quantitative data quality metrics may be available from a data producer, and this type of user feedback can be beneficial to data consumers. Research has shown that, in some instances, data producers may believe that their data has “perfect” data quality in terms of assessed accuracy, yet end-user perceptions of the data quality can vary substantially (Wang, Strong, and Guarascio 1994). Wang, Strong, and Guarascio (1994) suggest that the parameters of data quality that are important to data consumers are substantially different from the largely accuracy-based measures that data producers often report on, and include, the following.

- **Intrinsic data quality:** consists of believability, accuracy, objectivity, and reputation.
- **Contextual data quality:** consists of relevancy, timeliness, completeness, and appropriate amount of data.

- **Representation data quality:** consists of interpretability, ease of understanding, representational consistency, and concise representation.
- **Accessibility data quality:** consists of accessibility and access security.

More and more geospatial data discovery and distribution systems are incorporating increasingly sophisticated capabilities for data consumers to document and share their perceptions and measures of data quality. These community-generated measures and assessments of data quality will continue to play an important role in helping end users determine whether data are appropriate for their intended use. As technology advances, and open data initiatives expand and gain momentum around the world, the audience for geospatial data will continue to grow and consumer-developed measures and indicators of data quality will play an increasingly important role. When these are combined with measures of data quality developed by data producers and distributed via formal metadata records, end users have a full spectrum of information available to them to help inform and support decision-making using data from the widest possible variety of sources.

**SEE ALSO:** Geographic information system; Geospatial metadata; Metadata; Spatial data infrastructures; Spatial resolution; Validity and verification; Volunteered geographic information: quality assurance

**References**


Further reading


Data structure: spatial data on the web

Stefan Steiniger  
Pontificia Universidad Católica de Chile

Andrew J.S. Hunter  
University of Calgary, Canada

Spatial data on the web and in our daily life: Where do we need spatial data and what are spatial data?

Everyone who has a computer or television at home looks at or uses spatial data that have been transferred over the Internet. Weather forecasts on TV or weather maps on the Internet are based on spatial data: temperature, humidity, precipitation, and so on that are gathered and sent mostly automatically from weather observation networks. These data may arrive every 10 minutes or so at a weather office; then derived maps, forecasts, and other information are generated or updated. Similarly, spatial data are analyzed and transferred over the web when one searches for the shortest or fastest route to a vacation destination, when one gathers statistics on one’s weekly biking or running activities, or when consulting traffic and location services using Google Maps, for example. Sensing services used to monitor and analyze earthquakes or to forecast flood risk need to transfer spatial data over the web. Similarly, many public sector activities including transportation, planning, and development require large volumes of spatial data. Authorities that collect spatial data, for example, a forestry service, land registries, or a transportation agency, often send data to other agencies to plan and undertake day-to-day activities. Some uses are tolerant of long (slow) data responses, but other applications are time critical.

“Data structures for spatial data on the web” denotes specialized data formats that are used to transfer spatial data from one place to another place via communication networks and allow (almost) immediate display, manipulation, and, perhaps, further processing of the data. Note that either a person or a computer can be at the receiving end of the data transfer. In the case of a computer, data are transferred between two web servers. Here one web server “serves” data to another server, which is also known as the “processing web server”, which may generate maps, statistics, and so on. In the case of a person requesting such maps and statistics, by using their home PC, mobile phone, or a tablet, data are transferred between a server and a client. A client displays the data and may even improve data presentation. But the client does not serve the (processed) data to someone else.

The types of geographic data that are transferred over the web are similar to information that we see on daily news sites, nytimes.com, for example, or social networks such as facebook.com. that is, images and text. What makes the data “spatial data,” however, is the fact that the data contain additional information, called “geographic coordinates,” that allows us to relate the information to a place.

Spatial data are stored and transferred in two very distinct data formats: as vector and raster
data. Both of these formats are described in detail below. However, to summarize: if we transfer raster data over the web, then we transfer some type of image format (e.g. a *.jpg or a *.png file), which stores “continuous” information in a grid structure – raster data examples include satellite images, terrain elevations, or temperature surfaces for a region. One or more discrete real-world objects, such as a bridge, a building or a forest, are transferred as vector data. When using vector data, each object is described by a geometry, a set of attributes that characterizes the real-world object (see the example given later in Figure 2), and often styling information for presentation of the information.

Discussed here are important considerations when sending spatial data over the web, how data can be transferred from one location to another, and what data formats and standards exist for storing and transferring spatial data.

Important technical considerations for sending spatial data on the web

Previously, spatial data formats have often been developed for use with desktop geographic information system (GIS) software to store data in some practical and convenient way on hard disks. File format development has focused on re-use of data with desktop GIS programs. That is, developers have been concerned about the speed of data read and write operations to disk, accessibility of file content, file size, and so on. However, many initial formats were not developed to support the interoperability that is necessary when data need to be transferred from one computer to another. Hence, the requirements and constraints that are important when transferring spatial data via the Internet are examined.

(i) Probably the most important thing for the average Internet map user is that they can see the map or map data as soon as possible after sending their request to a web server. Rapid display and update of online maps depend, among other minor factors, on available network bandwidth and the amount of map data to be transferred. Reducing data size is possible by restructuring the data, by generalizing the data (see Map generalization), by using compression methods, or by sending only the data that the user really needs or wants at that very moment.

However, there are also other requirements. (ii) The format of the data should not be prone to transmission failure due to network interruptions; that is, if there is a network interruption, then the data do not need to be resent from the beginning, but can be restarted from where the interruption occurred. (iii) Data formats need to be independent of the computer platform, that is, a format can be read regardless of the operating system that the sending or receiving computer uses. This independence is also termed interoperability; it is discussed further in later sections.

Depending on the particular use of the data, it may be useful for the user to preview the data, choose the level of detail needed (which is similar to viewing maps at different cartographic scales), and select the themes and geographical areas that they are interested in – before sending a data request. This means that, on the one hand, (iv) the spatial data format should contain or link to metadata that describe the datasets spatial extent, the phenomena included in the dataset (e.g., roads or forests), the purpose of its generation, the year of creation, and so on – so that the user can decide if a dataset is of interest. On the other hand, (v) a data transfer service, which answers a user’s request for data, should allow data browsing, as well as content and spatial filtering of the data, to avoid sending out data that is irrelevant.
(vi) Data formats should ensure data consistency. This means that when compression methods are used to reduce the size of the data, then the compression method should not change the data in any way. If we look at the geometry of two geographic objects, such as a field and a forest that are adjacent to each other, then the relationship between their borders and their topology should be the same before compression and after de-compression. (vii) The final important criterion for spatial data web formats and services is the separation of content and styling. That is, a user may be interested in knowing about all the forests near a certain city for an environmental assessment; but the user may not be interested how the forest should be portrayed on a map, that is, its fill color and pattern, line style, and so on. In this case, sending map styling information is not important.

The importance of standards for sending spatial data over the Internet

When data are transferred between two computers the data server needs to understand what data are requested, and the client needs to understand the format used to send the spatial data so that the data can be stored, processed, and displayed. Syntactical interoperability deals with information exchange between systems through the use of common data formats. Several organizations, such as the World Wide Web Consortium (W3C), the International Organization for Standardization (ISO), and the Open Geospatial Consortium (OGC), have developed and published many XML-based formats, along with their data schemas, as well as standardized image formats, such as JPEG, TIFF and PNG, to support interoperability. W3C is mainly concerned with general Internet standards, whereas the OGC has a particular focus on standards for web mapping applications and spatial data exchange. In particular, the OGC has released standards to describe spatial data and data format encodings (examined later). OGC has produced numerous XML schemas for web applications and services, for instance the web feature service (WFS), data encoding languages such as the geography markup language (GML), style languages such as the styled layer descriptor (SLD) for visualization, and filter encoding (FE) specifications. However, several de facto spatial data formats are also frequently used that have not been standardized by ISO, OGC or W3C. Examples include ArcGIS's Rest API, the GeoJSON format (examined later), and the GPX format.

How to send spatial data over the Internet

Solutions for sending data: files versus services

To send spatial data over the web, one can distinguish between two options: sending files and sending data via a data service (Figure 1). Sending files is convenient if the dataset is small or the user understands the data structure. Using a data web service makes sense when the content of the data is unknown, when data sets are large and only a subset is needed, and when data should be updated. It is important to note that sending data via a service is very often a file-based process, with the files being created ad hoc for transfer.

Sending files. If data are sent by file, then they are often compressed first to reduce file size, for instance using GZIP compression. Transferring data by (compressed) file, however, does not allow the user to see and browse the data before they receive them completely. Furthermore, the send process is not responsive in that it may
Figure 1  Two general approaches to sending spatial data via the Internet.
fail during network interruptions, and users are not able to select a subset of the data. Probably even more importantly, the user needs to be sure that their software can read the data format. For instance, the user may download geometry data in Autocad’s DWG format, but few GIS are able to read this format. Here, syntactic interoperability is not ensured. Converting the data into a standardized geospatial file format, such as GML or KML, before making them available to others can resolve this interoperability problem. In comparison with the DWG format, OGC’s GML and KML format specifications are open to the public, and hence can be implemented freely by software developers. However, in practice only a few (desktop) GIS support the GML standard fully.

Using data web services. If a data web service, such as OGC’s WFS and Web Coverage Service (WCS), is used, then several of the problems previously outlined can be avoided. For instance a data service, such as a Spatial Data Infrastructure (SDI), enables browsing of the data, selection of subsets, can cope with network interruptions, and ensures interoperability. HTTP GET and DESCRIBE communication commands, such as GetCapabilities, GetMap, GetFeature, DescribeFeatureType, and so on, are used to inform the client about the offerings from the data services and to select data subsets. The use of standardized methods and data descriptions enables the client application to identify in which data format the requested data can be sent so that the client can read the data correctly. Spatial data are mostly transferred using one of four data serialization standards: URLs that contain key-value pairs, eXtensible Markup Language (XML), JavaScript Object Notation (JSON), and protocol buffers (Table 1). Some data serialization formats, such as XML and protocol buffers, provide a schema in addition to the data. The schema contains a description of the data structure and data types for the data during data transfer. This allows easier parsing and validation of the data received.

While we will discuss particular spatial data formats in more detail below, note that XML and JSON-based data formats are not designed to minimize the amount of information to be sent. Hence, standard HTTP compression capabilities are often employed (e.g., gzip) to reduce the transmission time. Furthermore, choosing a subset of the data can significantly reduce the size of the dataset that needs to be transferred. Several of the OGC’s spatial data service standards allow the selection of subsets using filter encoding standards. One can select data within a particular geographic area of interest, by geographic theme, image band of interest, and by level of detail (similar to map scale), and one can select certain geographic objects or raster images based on client-specified criteria. Standards that allow subset selection include OGC’s WFS, WCS and the web map tiling standard (WMTS).

2-D – Vector data formats and services

As outlined, vector data can be transferred as files or via a data web service. Common vector data formats include ESRI’s Shapefile format and AutoCAD’s DXF format. Both data formats are proprietary and pre-date the geospatial web. Hence, data transfer via the Internet was not likely a design objective for Shapefile and DXF formats. The Shapefile format, although fairly small in size because of its binary form, allows only one geometry type (i.e., polygons or lines or points) per file, and all features must have the same attribute set within a file – so trees cannot be mixed with buildings in the same dataset. Furthermore, one Shapefile consists of at least four files that need to be packaged together. Similar
## Table 1  
Internet encoding/serialization standards used for transferring spatial data.

<table>
<thead>
<tr>
<th>Name (specification)</th>
<th>Purpose</th>
<th>Elements and example</th>
<th>Characteristics</th>
</tr>
</thead>
</table>
| Key-value pair (IETF RFC 2616) | For data requests and data submission, i.e., HTTP GET, HTTP POST | - Form: “keyword = value”, separated by ampersands (&) appended to an URL  
- Example: http://example.com?name=Smith&initial=JS | Human readable, keyword-value pairs |
| XML (IETF RFC 3023) | For description of data, including documents, to request and send data | - Version and character encoding declaration  
- Elements declared by a start tag and end tag, e.g., `<step>` and `</step>`  
- Element attributes, e.g., `<step number = "3">`  
- Example: `<person> <name>Smith</name> <initial>J S</initial> </person>` | Human readable; can be used for hierarchical structured data; self-describing, extensible |
| JSON (IETF RFC 4627) | For sending data objects (not for documents) | - Array and record based  
- Curly brackets to indicate object start and end  
- Attribute-value pairs  
- Example: `{"name":"Smith","initial":"JS"}` | Human readable, attribute-value pairs. Simpler than XML (smaller grammar); can be used for hierarchical structured data; self-describing |
| Protocol buffers (Google) | For sending and storing structured data objects | - Using a proto schema file to describe data first, followed by serialization of one or many data objects  
- Proto schema example: `message Person { required int32 id = 1; required string name = 2; optional string initial = 3; }` | Binary: i.e., small size but not human readable; name-value pairs, can be used for hierarchically structured data; not self-describing – requires .proto files, extensible, forward and backward compatible |
problems exist for the popular DXF format. DXF does not allow user-defined attributes. This means that DXF files cannot carry information to characterize a tree as an oak tree with a height of 10.5 m and a stem diameter of 80 cm. Hence, new web-specific spatial data formats were developed to address such requirements.

Based on XML serialization, the OGC developed and standardized two spatial data formats: the geography markup language (GML) and the keyhole markup language (KML). Another existing XML-based format not adopted as a standard by OGC is OpenStreetMap XML (OSM). GeoJSON, derived from JSON, has not been formally recognized as a standard yet (in 2016), but GeoJSON is widely supported by browser-based mapping applications (e.g., Leaflet and OpenLayers). Finally, spatial data formats based on the protocol buffer file (PBF) standard are the OpenStreetMap PBF format and MapBox’s vector tile format. Table 2 presents the characteristics of several vector data formats.

GML, KML, and GeoJSON formats have adopted the OGC simple feature specification (OGC SFS) to describe the geometry of geographic objects. This specification, among other things, defines the basic geometry elements: Point, LineString, and Polygon, as well as GeometryCollections of those (see Data structure, vector). GeoJSON files and protocol buffers are compact formats in comparison with XML-based data formats. However, GeoJSON and XML formats (i.e., GML, KML or OSM) can be compressed. Some of the vector formats, in particular KML and GML, can include styling information for point symbols, labeling, line width, color, and so on, for display. Each of the formats has its advantages and disadvantages with respect to size, extensibility and flexibility. Hence, choosing a transfer format should be based on the characteristics of the particular use case. Since formats for 2-D vector data only are being discussed, readers with an interest in 3-D vector data are referred to OGCs CityGML standard and the American Society for Photogrammetry and Remote Sensing (ASPRS) organization’s LAS format.

There are two web service standards for transferring vector data: the OGC web feature service (WFS) and ISO 19142: 2010 Geographic Information – Web Feature Service. The OGC standard forms the basis of the ISO WFS standard. The standards utilize the GML standard to send vector data. Operations described by the WFS standards, such as DescribeFeatureType and GetFeature, allow a client to browse, retrieve (get) a subset of a dataset, or even a single geographic object only using their respective filter encoding specifications, and also create, delete, and update features in a dataset using a transaction service.

Figure 2 presents a vector data example (see image on the bottom left) showing how different geographic objects can be described in GeoJSON and KML formats. The dataset contains: a building with the attributes “name” and “address;” a road with the attributes “name” and “road type;” and two trees, with the attributes “tree type” and “height.” As can be seen, the geometry type of the building is Polygon, the road geometry is of type LineString and the trees are described as Points. In both formats the locations, that is, coordinates, are geographic coordinates, latitude and longitude. GeoJSON coordinates are assumed to be in the coordinate reference system CRS84, (urn:ogc:def:crs:OGC:1.3:CRS84) unless specified otherwise, whereas KML is WGS84 (EPSG:4326). GML, not shown here, expects that the coordinate reference system (CRS) will be defined explicitly as an attribute of the feature, and requires a schema (.xsd) describing the feature types included in the GML document. In the given example, no styling information
### Table 2  Geospatial data formats used for online data transfer.

<table>
<thead>
<tr>
<th>Name</th>
<th>Purpose</th>
<th>Serial</th>
<th>Key elements</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>GML</td>
<td>To describe application schemas and to transport and store geographic information</td>
<td>XML</td>
<td>- Data schema file (*.xsd) to describe data types, and a data file (.gml)</td>
<td>Human readable, (very) flexible, for raster data, vector data, and triangulated irregular networks, permits specification of application schemas (e.g., CityGML, netCDF)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- FeatureCollections with FeatureMembers with geometries, or coverages, and attributes, spatial reference system (SRS) information, display styles</td>
<td></td>
</tr>
<tr>
<td>KML</td>
<td>To display geographic data in a (Google) Earth browser</td>
<td>XML</td>
<td>Placemarks, ground overlays, paths, polygons, network links (e.g., for dynamic data), display styles</td>
<td>Human readable, fixed schema, a specialization of GML, for raster and vector data, coordinates only in WGS 84</td>
</tr>
<tr>
<td>OSM</td>
<td>To transfer OpenStreetMap data</td>
<td>XML, PBF, JSON</td>
<td>Nodes, ways, relations</td>
<td>XML: human readable, fixed schema</td>
</tr>
<tr>
<td>GeoJSON</td>
<td>To transfer geospatial data based on JSON</td>
<td>JSON</td>
<td>FeatureCollection with Features with geometries and properties</td>
<td>Human readable, not very practical for image or raster data, coordinates in WGS84, no style information</td>
</tr>
<tr>
<td>TopoJSON</td>
<td>Extension of GeoJSON to encode topology</td>
<td>JSON</td>
<td>Like GeoJSON plus additional topology object type with an “arcs” member</td>
<td>See GeoJSON; files should be smaller than GeoJSON files</td>
</tr>
<tr>
<td>MapBox VectorTile</td>
<td>To encode tiled geospatial data to be shared across clients</td>
<td>PBF</td>
<td>- Data schema defined in vector-tile .proto, data file</td>
<td>Binary; for squared tiles in spherical Mercator projection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Layers with features with geometries (described by drawing commands) and attributes</td>
<td></td>
</tr>
<tr>
<td>GeoRSS/Atom</td>
<td>To extend existing feeds with geographic information</td>
<td>XML</td>
<td>Feed data: title, author, entry, updated, etc.; with a geometry element as part of an RSS entry element</td>
<td>GeoRSS Simple coordinates are in WGS84; GeoRSS GML is a formal GML application profile that allows the spatial reference system to be specified</td>
</tr>
</tbody>
</table>
Figure 2  Example for a geographic dataset stored in GeoJSON and KML format.
(KML) is included for the display of the objects. When compared, the GeoJSON format is more compact (1788 Bytes) than KML (4343 Bytes). GML is generally more verbose (3358 Bytes plus a 2408 Byte schema).

2-D – Raster data formats and services

Examples of spatial raster data include satellite images, gridded elevation models, and temperature or wind speed snapshots of a region. Spatial raster data also include image tiles used by web mapping services to render a base map, such as Google Maps or Here.com.

Raster data and raster maps are usually transferred in an image format, such as JPG, TIFF or PNG. Supplemental information is added to each raster image indicating the real-world coordinates of the image corners, the map projection used, and perhaps information describing any image transformations applied (e.g., stretching and rotation). Spatial data web services offer the same image formats. Since JPG, TIFF and PNG have built-in data compression methods, HTTP compression is usually not applied when transferring image formats. Which type of image format is best to use depends on whether the user is looking for a ready-made map (with styling), or the actual raster dataset. In the case of a map, it makes sense to use JPG or PNG formats. JPG employs lossy data compression, whereas with raster data the TIFF format is recommended as it offers lossless LZW compression. The image formats, JPG, TIFF and PNG, do not allow separation of data and styling information. However, newer formats may support this (see for instance the OGC symbology encoding specification).

There are at least three Open Geospatial Consortium (OGC) standards for distribution of raster data using a data web service: the OGC web mapping service (see also ISO 19128), the OGC web coverage service (WCS), and the web map tiling service (WMTS). However, there are other standards that address distribution of raster data as well, such as OGC GML in JPEG 2000, OGC GML version 3, OGC NetCDF, and ISO 19123.

An OGC WMS returns complete maps of (dynamic) spatially referenced data in an image format. The WMS service may offer its own styling, or an OGC styled layer description (SLD) file can be sent with the map request. A fully conforming OGC WMS service returns PNG, GIF, JPEG, and TIFF formatted images. The bounding box of the area of interest, image dimensions, themes of interest, and the map projection can be specified in the WMS request. The service does not allow updates, so-called transactions, to the data on the server.

An OGC WCS service is not used for retrieving completely styled maps but allows detailed access to raster data of space/time-varying phenomena. Supported output formats may include GeoTIFF, GRD, NetCDF, and JP2000. However, metadata about the raster data are sent using OGC GML. The user request to the service can specify parameters such as the bounding box for the area of interest and the time – to select data for a particular date and a particular image band (e.g., near infrared) to ensure that only data of interest are received. The map projection, image dimensions, image resolutions, and interpolation methods for the rescaling of an image can be specified as well. Version 2 of the standard extends capabilities to work with multidimensional and time series raster data as commonly encountered in meteorology and climatology. Extensions that allow a WCS to be transactional, that is, to update datasets (WCS-T) and to process and filter data (WCPS), exist as well.

The introduction of a web map tiling service (WMTS) resulted from deficiencies (response time) encountered when working with a WMS
service. Often WMS services require a few seconds to respond to an image request, because the map image is generated on the fly. When there is a large number of concurrent users, for example, serving many users at the same time, these CPU-intensive services become impractical. Google Maps and later OpenStreetMap were the first to use the so-called “slippy map” approach to address this processing problem. Here, one may imagine that a big map image is rendered first and then cut into many small images that are called “tiles.” Only a few of those tiles are of interest to the user and need to be sent to the client. Individual tiles for a complete map can be precomputed over a range of map scales (different levels of detail) so that when a user zooms in and out tiles can be retrieved more rapidly. This technique was formalized in the OSGeo TMS standard and later in the OGC WMTS standard. A restriction of this service is that each tile map can have only one map projection and so-called “faces” are needed to cover the Earth’s polar regions. It is recommended that tiles be sent in JPG or PNG formats and tile sizes are 256×256 or 512×512 pixel large. The number of zoom levels (scales) available usually depends on the map projection used and the WMTS data provider. The storage structure generally takes the following form: “www.domainname.com/mapfolder/z-zoom level/tile-column/tile-row.”

Sending event data: web updates and sensors

So far, we have discussed formats best suited for data requests for a particular geographical object or an entire regional dataset. However, many real-time applications require event data with constant updates of spatial information. For example, when traveling by car the car navigation system may send information about the vehicles current position every three seconds. Another example of event data is seismometer data collected to record earthquakes. To map and analyze earthquake events over time the information from the seismometer should include the strength of the earthquake, the time stamp for the earthquake event, the geographic location of the seismometer, and perhaps a reference to calibration metrics for the instrument.

Formats that are commonly used to send event data include two XML-based formats: rich site summary (RSS) and Atom, and JSON. Similar to GeoJSON the RSS format has a spatial equivalent called GeoRSS (described in Table 2), which is also used for Atom-formatted data. Use of the GeoRSS or Atom format makes sense for updates on world news webpages. Since 2013, GeoJSON seems to have established itself as the de facto standard for transmitting location events. For instance, the US Geological Survey’s Earthquake Hazard Program makes real-time earthquake information available in GeoJSON and Atom formats. However, some webpages offer location event data in KML format.

The Open Geospatial Consortium has also defined a set of “sensor web enablement” standards for description, communication and data exchange with (environmental) sensors. One particular standard of this sensor web enablement set is the OGC Sensor Observation Service – OGC SOS. This service standard defines a GetObservationService request that returns sensor measurement data. Other, optional, commands allow the retrieval of information about the sensor itself (e.g., its location) and information about the object that is monitored. (More information on this topic is to be found in Sensor networks, wireless.)

Current developments

As with most technical fields there are many development activities in the area of spatial data
transfer techniques and standards. In particular, we see work focusing on the delivery of background maps as vector tiles instead of raster tiles. In this context, new vector data encodings are being introduced. Among those is the Topo-JSON format by M. Bostock (Table 2). This format was developed to reduce the size of the datasets that are to be transmitted by exploiting their topological relationships. For instance, a common border of two touching polygons is stored only once, rather than twice, as in a purely object-based geometric model.

A second format that may find wider application is MapBox’s VectorTile format, which is based on the protocol buffer serialization (Tables 1 and 2). Several objectives have led to the development of the format. Two of them are that vector-based tiles can be much smaller than image-based tiles for geographic regions with few map objects, such as a desert, a forest, or water environments, and that using the vector format allows the immediate application of custom map rendering styles (e.g., chosen by the user) as opposed to pre-rendered image tiles. The development and testing of vector map tiling were and are performed not only at MapBox but also by others; for instance, Apple and Google Maps use vector tiling for map display on mobile devices now (in 2016) as well.

Finally, it should also be mentioned that substantial ongoing research investigates efficient data transmission of spatial data for use in multiplayer online games. If one thinks of online games as places where people explore cities and landscapes (such as in first-person shooter games), then large volumes of terrain data need to be transferred and rendered by the game engine. Developments in this area are of importance to GIS developers because the gaming engines are used for online games but also play an increasing role in pilot and driver trainings, as well as in military and medical simulations.

**SEE ALSO:** Data structure, raster; Data structure, vector; Interoperability of representations; Map generalization; Representation and presentation; Sensor networks, wireless; Spatial data infrastructures; Topological relations; Web-mapping services

**Further reading**


The raster data structure is an approach for encoding spatial data computationally for use within a geographic information system (GIS) or remotely sensed image processing software package. Its general organizational approach is the exhaustive tessellation of space into a set of nonoverlapping spatial units, typically square grid cells. Each individual grid cell contains a value that represents a property of that grid cell location, whether it is the presence of a geographic feature or measurement of some kind. The raster data structure, along with the vector data structure, composes the primary geometric approach for representing spatial data computationally.

Raster-based GIS has its roots in the late 1960s and 1970s. Its development was driven in part by advances in computer graphics data handling and the limitations of line printers for cartographic output at the time, which operated by overplotting characters in individual cells of a rectangular grid to create shaded regions. The raster structure is also closely related to the products of digital scanners and environmental remote sensors affixed to airborne and satellite platforms. Notable early academic research included the raster data handling software packages IMGRID and MAP (Tomlin 1990) and the development of standards for raster data structures and processing (Peuquet 1979; Samet 1984). Many of these early contributions are now embedded, and have been substantially advanced, within academic, open source, and commercial GIS and remotely sensed image processing software packages, such as GRASS GIS, ERDAS IMAGINE, IDRISI, and ArcGIS, among others.

Raster data can be used to represent nearly any type of geographic phenomenon but certain types of raster data are particularly prominent. Remotely sensed digital imagery gathered from ground, ship, and satellite platforms are encoded in raster format, and these data form a large portion of GIS data generally, particularly for their use as base, or framework, data on which other datasets may be registered. Other important raster data sources are digital raster graphics (DRGs), scanned maps of US Geological Survey topographic quadrangle maps, and digital elevation models (DEMs). DEMs encode elevation data in grid form and, though they were initially created by interpolation from contour line data, most current wide-extent DEM datasets, such as the shuttle radar topography mission (SRTM) data with coverage over nearly the entire Earth, are derived from remote sensing. In addition to elevation data, many other environmentally oriented datasets, such as those related to hydrological and meteorological modeling, are in raster format, due to the spatially continuous nature of these types of environmental variables as well as the affinity of many computational simulations of environmental processes with raster data organization.

**Raster data modeling**

Data structures, including the raster data structure, are a component of spatial data modeling, which details how a geographic phenomenon in the real world is conceptualized and ultimately
represented and encoded in a computer. Data modeling proceeds through a series of stages, where each stage represents a greater degree of abstraction in the representation of a real world phenomenon (Peuquet 1984). The stages include the conceptual model, data model, data structure, and file structure (Figure 1). The file structure concerns the actual physical encoding of data on a storage medium, and is not relevant to the discussion here. The present chapter thus focuses on the conceptual model, data model, and data structure as they relate to raster modeling.

### Conceptual model

The conceptual model concerns how a geographic phenomenon is conceived of in the mind of the analyst who is creating the computational representation. Generally, two high-level conceptual models are recognized – object based and field based. An object-based conceptual model of the world is one in which the world is viewed as an empty space occupied by individual objects whose boundaries are discrete and definable. The field-based conceptual model of the world is often used for phenomena that are not thought of as discrete entities but rather which are considered to vary continuously through space. Often, such phenomena are considered qualities that vary from location to location, and thus while they have no discrete boundaries they can be measured or observed at any given location. Many environmental phenomena fit this description, such as temperature, relative humidity, and soil salinity, things which are qualities that can be measured but are not distinct entities or features in and of themselves. Such conceptual models are often materialized in language, such that while it makes perfect sense to say “Look at that building over there,” it does not make sense to say “Look at that temperature over there,” simply because temperature is conceptualized not as a discrete object but rather as a quality that exists as a continuum throughout space.

### Data model

A spatial data model is an approach to using a certain type of geometry to represent geographic phenomena computationally, where the geometric configuration extends from the conceptual model. There are two primary spatial data models – vector and raster. The vector data model uses point, line, and polygon geometric primitives to represent the spatial expression of a geographic phenomenon. The raster data model utilizes an exhaustive tessellation of space into a set of spatial units as a representational device. Each spatial unit encodes a value that represents some property or measurement at that location. Though irregular tessellations can be used, regular tessellations are the norm. Several different types of regular tessellations have been used, including tessellations of triangles, hexagons, and squares (Figure 2). Hexagons,
unlike triangles and squares, have the advantage of having their centroids equidistant to all adjacent hexagons, which is a substantial advantage when performing operations related to the measurement of distance from one spatial unit to another. Note that for a tessellation of squares, the distance from the centroid of one square to the centroid of an adjacent diagonal square (i.e., two squares that share a point) is 1.4 times the distance between the centroids of two adjacent squares that share an edge.

The tessellation of space into squares has, however, several advantages that tessellations of hexagons and triangles do not share. Tessellations of squares are similar in structure to remotely sensed imagery derived from aerial and satellite platforms, which serves as a major source of data for GIS. Square-based tessellations also have an affinity with conventional computer graphics data structures, which supports the adaptation of computer graphics processing and manipulation to the GIS environment. Such a structure is also directly compatible with the Cartesian space of planar coordinate systems used to georeference data within GIS. Squares can be recursively subdivided into smaller spatial units of the same shape and orientation, too. In other words, a square can be subdivided into a set of smaller squares. This provides substantial efficiencies in data encoding, retrieval, and analysis.

For these reasons, by far the most common approach in raster data models is to employ a tessellation of squares in a matrix, also commonly called a grid, or sometimes, a raster. Each individual square in the grid is called a pixel, formed from an amalgamation of the words “picture element,” a term commonly used in computer graphics and image processing. Pixels are also referred to as grid cells, which is the term that will be used here. Each grid cell encodes a value that captures some observation or measurement at that grid cell location. For example, a standard raster data model representation of Earth surface elevation would be a grid, where each square grid cell contains the elevation at that grid cell location. Because, unlike with vector, the property of a phenomenon is attached to a location, rather than some geometric representation of an object, the raster data model is often said to be “location based.”

The reader may notice a logical extension of the raster data model from the field-based conceptual model and the vector data model from the object-based conceptual model. Certainly, the raster data model, with its location-based orientation, has a natural affinity for the representation of phenomena that are conceptualized as varying continuously over space, as such continuous variation can be represented by the set of values encoded in the raster’s grid.
cells which are distributed regularly throughout space. And vector is a logical choice for phenomena that are conceptualized as objects, where points, lines, and polygons are used to capture the location and spatial extent of features. However, it should be emphasized that any geographic phenomenon can be represented using either the raster or vector data models.

**Data structure**

A data structure is the encoding strategy for the spatial data model, a specific design of how to store the geometry expressed in the data model within a computer. It is typically expressed as a set of diagrams and tables that illustrate the computational storage of spatial data in a database or set of files. The data structure is implemented within a programming language. Thus, while the raster data model encompasses the idea that spatial information can be stored using a tessellation of space into regularly spaced spatial units, with each unit encoding an observed value, there are many different raster data structures that specify how to actually computationally encode such a geometric configuration. The raster data model may be thought of as composing a family of specific raster data structures. The algorithms that form the analytical raster component of a GIS package are developed for the specific raster data structure implementation employed by that GIS package.

Perhaps confusingly, the terms “raster data model” and “raster data structure” are often used interchangeably. There is an ambiguous semantic boundary between a data model and a data structure, as the former implies the concept of data organization while the latter implies a more specific encoding strategy. However, even the same general encoding strategy can be implemented somewhat differently in different programming language for different data formats. Many GIS software vendors have developed their own proprietary raster data formats, all of which are based on the same idea of tessellating space into a regular grid, but which differ in their implementation details.

**Fundamental principles of raster**

Considered here is by far the most common type of raster data model, the raster grid, where the space is tessellated exhaustively into a matrix of square grid cells. The grid is oriented so that the rows express the coordinate position along the north–south axis and the columns express the coordinate position along the east–west axis. Encoding a value for each grid cell is thus based on recording a value for each row and column position. GIS packages handling raster data typically represent geographic features using a set of grids in which each grid can be considered an individual spatial data layer. Each layer represents one theme – that is, the spatial distribution of one specific type of geographic object, spatial property, observation, measurement, or characteristic. Some basic layers, for example, would be elevation (above sea level), the locations of rivers, and the locations of cities. There are many possible themes, and the specific themes involved are determined by the database creator. These multiple layers are stored in a GIS database to depict the salient features of a region.

**Resolution**

A key property of any raster is the resolution, which refers to the size of the grid cell. The resolution is given as the length that one edge of a single grid cell measures on the Earth surface. For example, a 12 km resolution grid consists of a matrix of grid cells, each of which measures 12 km along its side. Thus, the area of each grid cell in such a grid is $144\text{km}^2$ ($12 \times 12$...
The resolution of a grid is a measure of its spatial precision—the finer the resolution, the greater the precision. Given a raster of a particular extent, as the size of the grid cells gets smaller the degree of precision increases, but the number of grid cells necessary to cover the extent also increases. Every time the resolution is cut in half, the number of grid cells necessary quadruples. Consider the example shown in Figure 3 for an analog map of two polygons, labeled “1” and “2,” with an extent of 12 km. At a resolution of 2 km, 16 total grid cells are required to cover the entire area. If the resolution is doubled (made coarser) to 4 km, the number of grid cells drops to nine, and at a resolution of 6 km, only four grid cells are required. Note also how the precision, that is, the level of detail at which the raster captures the spatial variation of the boundary between the two polygons, is reduced as the resolution is made coarser.

The decision of what resolution to use resides with the analyst, if it is not predetermined by the technology used to generate the data (e.g., as with an image derived from remote sensing technology). While, ideally, it is preferable to have the finest resolution possible, there is a tradeoff in the granularity of the representation and the storage volume necessary to hold an increasing number of grid cell values. A raster with a very fine resolution over a large extent can be cumbersome in terms of the memory required for storage and also for processing. Thus, the analyst should consider the precision required for the application of the data when deciding on a resolution.

### Encoding

The most basic raster encoding scheme uses only values of one (“1”) and zero (“0”) to encode the presence and absence, respectively, of some geographic feature. Consider, for example, an analog map of a landscape presented on the left hand side of Figure 4, which shows two lakes bisected by roads, along which are located three houses. In the analog map representation, the lakes are represented as areas, the roads as lines, and the houses as points. In a raster representation, each type of feature is encoded as a separate layer—thus there is a houses grid, a roads grid, and a lakes grid. Every grid cell that contains that feature is given a value of “1,” otherwise the grid cell is given the value of “0.” Figure 4 shows grids of houses, roads, and lakes for the analog map.

A pertinent question relates to how a grid cell is assigned a “1” or a “0.” For point features, the assignment is relatively simple. If a grid cell contains a point, it is given a value of “1”; if

---

Figure 3  Impact of changing the resolution for a raster representation of a mapped area with an extent of 12 km: 2 km resolution grid, 4 km resolution grid, and 6 km resolution grid.
not, it is given a value of “0.” For line features, if a line crosses or touches a grid cell, that grid cell is given a value of “1”; if not, the cell is given a value of “0.” For polygon features, the issue of assignment may be more complex, as a polygon may only partially overlap a grid cell, or multiple polygons may overlap a single grid cell. Two conventional approaches are used for polygon assignment. The first approach is the “centroid rule,” where a grid cell may be assigned the value of the polygon that contains the grid cell centroid. The second approach is “winner-take-all,” where the polygon that occupies the greatest area within a grid cell is used for assignment. Occasionally, other criteria may be used to prioritize assignment for certain themes. For example, if a raster is being created that describes the presence of certain animal species with the aim of habitat preservation, one may want to code any grid cell that overlaps with that species distribution as containing that species, even if the species range did not occupy a majority of the grid cell area.

There is also the question of whether the value of each grid cell is assigned to the entire area of the grid cell or only to its centroid. Theoretically, if one conceptualizes the geographic phenomenon being represented by the raster as a homogeneous entity, it makes sense to consider the entire cell as occupied by that entity. For example, a raster representation of city neighborhoods indicates that each entire cell belongs to a neighborhood. If, however, the phenomenon is conceptualized as a continuously varying statistical surface, as with an elevation surface, then it makes sense that the raster is a set of samples of that surface occurring at the centroids of the grids cells. Notably, this distinction occurs only conceptually – there is no difference in the nature of computational encoding whether one considers the value occurring over the entire grid cell or only at its centroid. However, certain raster functions operate on the centroid of a grid cell, often when measuring, say, the distance between grid cells, while other functions treat the area of the grid cell.

Grids can take on a variety of values beyond simply zero and one, and can contain nominal (categorical), ordinal (ranked), and interval/ratio (continuous) data types. Most GIS packages that handle raster data, however, encode only numeric, not text, values. Thus, in the case of nominal data, numeric values are used to stand for nominal categories. For example, Figure 5 shows an analog map of four nominal category land covers and its representation in raster form.

**Figure 4** Raster representations of an analog map with point (houses), line (roads), and area (lakes) types of features.
Ordinal data and interval/ratio data can be encoded directly as numeric values in the grid.

**Data types**

Early raster GIS packages were mainly limited to integer values due to the storage and processing demands of encoding and handling rational numbers, which demand a greater amount of memory devoted to storing the value of each grid cell. While today’s computers have greater storage and processing capabilities, many GIS packages continue to distinguish between grids that are intended to hold integer values from those that are intended to hold rational values. While not enforced as a rule, integer values are often used to encode nominal and ordinal data, whereas rational values are used to hold interval/ratio data. Nominal values in raster are often used to depict features, such as roads and buildings, as well as partitions of an area such as neighborhoods, political districts, and census boundaries. Here, each grid cell holds a nominal value that stands for the identity of the feature; for example, a neighborhood can be represented by a group of adjacent grid cells that hold the same nominal value.

Phenomena that are continuous in nature, such as elevation or temperature, likely have grid cells values that do not repeat but change in a continuous manner throughout space. Thus, rational values stored as a floating point data type are used to depict this continuous variation. The choice of integer versus rational values has implications for volume of storage, as the assignment of a short integer data type (with values ranging between 0 and 255) to store the value of each grid cell is less demanding in terms of memory than the assignment of multiple bytes of storage for each grid cell, as required by rational numbers. Additionally, redundancy in the values of integer type grids can be exploited to reduce the memory required to encode the raster, as explained below.

The grid data type also has implications for the storage of attribute data. Traditional raster data structures did not include attribute data in separate tables, as is the norm with vector data, but were limited to a set of grids, each representing a single theme. Some GIS packages have linked the raster data structure to the relational data model to handle attribute data, as with the vector data model. The standard approach is for an individual grid to be associated with a table, where each row in the table represents one unique value that occurs in the grid. It is NOT the case that each grid cell is represented by a row in the attribute table. Consider, for example, the grid of land cover presented in Figure 5. There are four unique values that occur in the grid: 1, 2, 3, and 4, each representing a particular land cover. Each unique value is

**Figure 5** Raster representation of nominal data: a land cover map with residential (1), commercial (2), industrial (3), and agricultural (4) land covers. The table (right) shows an attribute table attached to the grid.
represented as a single row in the associated attribute table; thus, there are four rows in the attribute table. The table contains a field that encodes that grid value (the “value” field, which can serve as the primary key in the table) and a field that encodes the count of the number of grid cells that contain that value (the “count” field). As with any table in the relational data model, other fields may be added to include additional properties, such as the name of the land cover and the area of the land cover.

This approach for linking the raster data model to tabular data is generally limited to integer grids intended to represent nominal or ordinal data types. In a grid holding rational number data, each grid cell may have a unique value, as the grid is typically used to represent a continuously varying surface. In such a case, the attribute table would have a number of rows equal to the number of grid cells. In even a medium-size grid this results in a number of rows that would be cumbersome to manipulate in most desktop-based GIS packages – a grid with just 1000 rows and 1000 columns would have an attribute table with 1,000,000 rows.

### Raster encoding and compression

#### Full raster encoding

The most basic raster encoding is simply a matrix or 2-D array of values, where the row and column position of the value in the array indicates its position in geographic space. This is termed “full raster encoding” or “cell-by-cell encoding.” Notably, it is unnecessary to explicitly store a coordinate position, or row and column position, for a grid cell value. A header file typically accompanies the data file, and contains information on the total number of rows and columns in the grid, the grid resolution, the memory allocated to each grid cell value, and information regarding the coordinate system. The coordinate system information often also includes the extent of the grid as captured in the minimum and maximum x and y coordinates of the area covered by the grid. Also of note is that because the data are often stored in an array, the row and column positions of a grid often begin at the upper left (northwest) corner of the grid, where for instance the top (northern) row is numbered row zero and rows below (to the south) are numbered sequentially. Likewise the column to the left (west) is numbered column zero and the columns to the right (east) are numbered sequentially. This is in contrast to most planar coordinate systems which typically have an origin to the southwest of the area being mapped.

Because a value is encoded for each grid cell position, the volume of raster datasets can grow to be very large and unmanageable. Thus, there are compression algorithms that allow for the storage of raster data with reduced memory requirements. All compression algorithms seek to exploit the principle of spatial autocorrelation – for raster, the idea that grid cells nearby or adjacent to one another are more likely to have identical, or at least similar, values. Such a spatial pattern allows for encoding schemes that may be more efficient than storing the value of each individual grid cell. Compression algorithms can be divided into “lossless” and “lossy” approaches. As the name implies, lossless compression allows for the recovery of the full raster encoding without any loss of data, whereas lossy compression results in some loss of data – the original full raster encoding cannot be completely recovered from the compressed file.

#### Lossless compression

A prominently used lossless compression for encoding raster is called “run length encoding.” Run length encoding stores runs of the same value, that is, sets of adjacent grid cells that share
the same value. For example, consider the analog map of three areas, labeled areas “1,” “2,” and “3,” shown on the left of Figure 6. To the right of the map is the full raster encoding in a $6 \times 6$ cell matrix, where 36 values are required. The run length encoding file shown to the right of the full raster encoding contains three columns: val (the value in the grid cell), run (the number of adjacent grid cells that hold this value), and row (the row number for that run of values). The encoding proceeds from the top left grid cell and reads left to right, top to bottom. The file reads that the value “1” occurs in the first row six times, the value “1” occurs in the second row six times, the value “1” occurs in the third row two times, the value “2” occurs in the third row four times, and so on through the entire grid. Note that this run length encoding approach requires only 24 values to encode the entire grid, and that the original full raster encoding can be easily reconstructed from the run length encoding table.

An even greater degree of compression can be had by eliminating the need to store the row number, and storing runs of adjacent cells that extend from one row to the next, again starting at the top row of the grid and reading each row from left to right. Here, only the value and the number of adjacent grid cells need to be encoded. Such an approach is illustrated in Figure 6 on the far right, where it is recorded that the value “1” occurs in a run of 14 cells (stretching from the first row to partially through the third row), the value “2” occurs in a run of four cells, and so on. Obviously, the reconstruction of the full raster encoding using this approach requires the GIS software to know how many rows and columns are in the grid, information that is contained in the header file.

Another well-known compression strategy for raster data is the region quadtree data structure, a recursive tessellation that utilizes a variable resolution (Samet 1990). It is similar in concept to run length encoding but is extended to two dimensions. The quadtree structure recursively partitions an area into quadrants, labeled “0,” “1,” “2,” and “3.” Any quadrant that is not homogeneous, that is, not all grid cells within the quadrant share the same value, is further partitioned into quadrants recursively until the resolution of the raster is reached. Only values of homogeneous quadrants are stored, which

Figure 6  Run length encoding (RLE). The analog map on the left is encoded using full raster encoding (middle left), run length encoding by row (middle right), and run length encoding by value (far right).
Figure 7  Quadtree raster data structure. Representation of the analog map (far left) using a recursive partitioning into quadrants to three levels. The gray square grid cells represent the gray polygon shown in the map.

reduces the redundancy of data storage for areas of the raster with repeating values.

Figure 7 shows an example of a quadtree. On the far left is an analog map with one polygon shown in gray. The first partition into quadrants is shown to the right, where each quadrant is labeled. Note that quadrants “2” and “3” are homogeneous, that is, there is not variation within the quadrant, but quadrants “0” and “1” are not, that is, the polygon occupies only a portion of the quadrant. Therefore, quadrants “0” and “1” are further divided into subquadrants and labeled accordingly, where, for example, quadrant “0” is subdivided into quadrants “00,” “01,” “02,” and “03.” Each of these quadrants is subdivided again, if not homogeneous. The final quadtree structure is represented on the far right, where grid cells – of various resolutions – are colored gray if they represent the gray polygon in the map.

Figure 8 shows a diagrammatic representation of the same quadtree shown in Figure 7. Here, one can clearly see the tree structure composed of nodes in the tree, where a node occurs at the beginning and end of each line, or edge, in the tree. The node at the very top of the diagram is called the root, and every other node, each of which represents a quadrant, has the potential to be subdivided into four other nodes, called children, that is, four subquadrants. When a node is homogeneous, and thus is not split into subquadrants (does not have any children), it is referred to as a leaf node. Other nodes which have children are referred to as branching nodes. In Figure 8, the leaf nodes that represent the gray polygon shown in the map in Figure 7 are colored gray.

Ordering strategies

A variety of different ordering strategies for storing series of grid cells can be employed. These strategies are called “space-filling curves” as their purpose is to transform data in two dimensions (i.e., the rows and columns of a raster) into one dimension for storage on a disk (van Oosteroom 1999). For example, the run length encoding example shown in Figure 6 employs a row ordering strategy – first the grid cells in row 0 are encoded, then the grid cells in row 1, and so on. There are variations of this strategy whereby the rows may be read from right to left, or alternate from left-to-right and then to right-to-left as one traverses from the top of the grid to the bottom (called row-prime ordering), or proceed from column to column rather than row to row. The row ordering strategy (and its variants), however, gives preference to row adjacency, and thus has the disadvantage of having two adjacent cells that share the same column position but are in different rows separated in the 1-D storage.
To address this, certain space-filling curves have been proposed to maximize the degree to which grid cells near one another in 2-D space will also be located nearby one another in the 1-D ordering. The space filling curve depicted by the ordered labels in the quadtree data structure in Figure 8 is one such strategy and is called a Morton curve (also called a Peano curve, an N-curve, or a Z-curve). The Hilbert curve is another commonly used space filling curve. Figure 9 illustrates each of four commonly used space filling curves.

Run length encoding and quadtree data structures are appropriate for raster data with large numbers of grid cells with redundant data values, typically raster representations of nominal or ordinal data. For continuous raster data, such approaches are inefficient, and may even increase the data storage requirements. For compression of continuous raster data, lossy approaches adapted from computer graphics have been used. These approaches typically capture the spatial patterns expressed in a raster through mathematical functions that approximate the continuous variation in grid cell values over space. They often allow the analyst to choose the degree of compression (and, consequently, loss of information) employed.
Lossy compression

An approach to compression of continuous raster data gaining in prominence is the wavelet transform. Here, a mathematical function expressing a regular wave, or oscillation, is used to model the continuous variation embedded in the raster. The deviations of each pixel from the wave are recorded as a set of coefficients. These coefficients can be used to reconstruct the original raster, in which case the compression is lossless. The coefficients can also be summarized statistically by taking averages of adjacent grid cells, thus decomposing the original wave into smaller wavelets. This process may be used recursively, such that two adjacent grid cells may be averaged, and then those averages may be averaged again to attain a greater degree of decomposition of the wave. This latter approach of statistically summarizing the grid cell values allows for substantial compression of the original raster, but is lossy, if only the summarized data and not the original coefficients are retained. One primary advantage of wavelet compression is that it supports storage and retrieval of the raster data at variable resolutions, as the raster can be retrieved using different levels of decomposition of the original wave.

Raster formats

Band sequential, or “BSQ” for short, is a way of organizing multiband imagery – that is, raster data in which there are multiple values per grid cell. Such data may be thought of as multiple individual raster layers in GIS, where each raster layer is a separate “band,” but within computer graphics and image processing it is common to store multiple values at single location, for instance, in order to store the magnitudes of red, green, and blue colors at each grid cell location in a graphic image. Or, in remotely sensed image data, each grid cell may have several values collected by different sensors. Band sequential encodes all the data for the first raster band, or layer, encoded in the file first, followed by all the data for the second band, and so on. Band interleaved by line, or “BIL,” implies the file is organized line by line, such that the first row of grid cell values is encoded for the first band, then the first row of grid cell values is encoded for the second band, and so on. A third organizational approach is band interleave by pixel, or “BIP,” where the values of the first grid cell are written for all bands, then the values of the second grid cell are written for all bands, and so on.

There are numerous raster file formats available, many of which are used in, or descend from, compression formats used in computer graphics. Many of these are open (i.e., their specifications are made public), patent free, or intended for cross-platform use. The JPEG (Joint Photographic Experts Group) format, with files having a “.jpg” filename extension, is a lossy compression algorithm that employs a discrete cosine transform (DCT) to model the variation in an image as a series of mathematical functions. This format was originally developed for sharing photographs and other images on the World Wide Web and continues to be widely used for that purpose. JPEG images with high compression ratios often have a blocky appearance, due to the nature of the compression, which operates over discrete square portions of the raster. JPEG2000, with the filename extension “.jp2,” is an updated approach that uses the wavelet transform. A metadata standard for georeferencing has been defined for JPEG2000 by the Open Geospatial Consortium (OGC) such that JPEG2000 images can be registered with other raster data in a GIS package.
The TIFF (tagged image file format) format, with filenames having the “.tif” extension, is now controlled by Adobe, Inc. It is a compression format that employs a form of run length encoding and has been widely used by digital graphics professionals because it is lossless and thus preserves the original full raster encoding. It has been extended to GeoTIFF, which includes a metadata standard for georeferencing information to support the seamless integration of TIFF files into GIS packages. The GIF (graphics interchange file) format has the “.gif” extension and also is widely used for graphics on the World Wide Web. It utilizes a lossless compression but is limited in the range of values that can be stored in each grid cell as compared to some other raster formats. The PNG (portable network graphics) format, using the file extension “.png,” was developed as a replacement for GIF. It is also lossless and offers a greater range of values as compared to GIF. MrSID (multiresolution seamless image database) is another raster format developed at Los Alamos National Laboratory and owned by LizardTech, Inc. It employs the wavelet transform and offers both lossy and lossless compression options.

There are also many proprietary formats developed for specific GIS and remotely sensed image processing software packages. GRID is the native raster format for the ArcGIS GIS package developed by Environmental Systems Research Institute, Inc. (ESRI). It is a lossless compression format using run length encoding. Another run length encoding-based lossless compression format is the “.img” format used by the remotely sensed image processing package ERDAS IMAGINE. The file format used by the remote sensing/GIS package IDRISI, with the extension “.rst,” is a full raster encoding.

Outlook

The raster data structure has distinct advantages and disadvantages. Advantages include its simplicity both in encoding and in the application of analytical algorithms. Basic encoding of raster data as a 2-D array is straightforward in nearly all programming languages, and raster algorithms can be built on well-defined array operations for the development of more sophisticated modeling and analysis functions. And because raster is the native format for much remotely sensed imagery and computer graphics, there has been a natural cross-pollination of raster data handling in GIS with these closely related fields. A disadvantage of raster is its blocky appearance for cartographic products, though this depends on the raster resolution. The main challenge with raster, however, has always been the large data volumes required, which can make storing and analyzing raster data problematic. Even given the rapid advances in computer storage capacity and processing speeds, this issue continues to be a prominent one, due to the ever-increasing volume of spatial data generated from the advances in, and increasing ubiquity of, geospatial technologies.

Future research in raster data structures should address the issue of efficient data encoding, and the relationship of such encoding to support raster analysis. Such research is particularly important for extending raster data structures and analysis from two dimensions to three dimensions in space, and four dimensions when time is also included (Mennis 2013). Other research frontiers include the development of data structures for more sophisticated semantic representation, which often integrate raster, vector, and object-oriented modeling principles (Goodchild, Yuan, and Cova 2007), as well as tighter integration of raster GIS with research in
DATA STRUCTURE, RASTER

computational simulation of environmental and socioeconomic processes.

SEE ALSO: Cartographic modeling; Data structure: spatial data on the web; Data structure, vector; Map algebra; Representation: fields; Sorting spatial data

References


Further reading


Data structure, vector

F. Benjamin Zhan  
Texas State University, USA

Yan Lin  
South Dakota State University, USA

Data structures are ways in which data are organized in a computer. This entry focuses on the topic of vector data structure for organizing geographic objects in a geographic information system (GIS). A GIS based on a vector data structure has been customarily called vector GIS. Another common way of organizing data is the raster data structure, which is the topic of another entry in this encyclopedia. On a paper map, geographic features are typically represented as points, lines, and polygons. In the vector data structure, the location of a point is represented by a pair of coordinates, a line is represented by an ordered sequence of shape points connected by line segments, and a polygon is represented as a set of lines in a given sequence forming the boundary of the polygon. Network data structure for supporting network analysis is also a form of vector data structure; it is covered in a different entry.

To understand the historic evolution in the development of vector data structure, it is necessary to have some background information about the traditional manual map production process before computers were used for mapping. In that process, the coordinates of a geographic feature to be mapped were first obtained through field surveying and/or photogrammetry. The feature was then drawn on a piece of paper either manually or by a mapping instrument based on the coordinates and a predefined symbol representing the geographic feature. This was a very labor-intensive process involving not only the initial mapping of any area but also the subsequent process of keeping the map current when any changes on the ground had to be reflected on the paper map.

Many paper maps were accumulated over the years since maps were invented for communication and archiving. In the 1950s and early 1960s when people began to realize the potential of using computers to support mapping, the immediate challenge was to transfer the paper maps into the computer. To store large volumes of maps in a computer requires a “good and efficient way” to organize map data. Therefore, the search for a data structure to organize maps in a computer based on coordinates was one of the most significant research challenges in automated mapping.

Figure 1 is a map generated from modern GIS showing some basic geographic features in San Marcos in Texas, the home city of Texas State University. This is a typical reference map that shows some basic geographic features in an urban area. Although it is a computer-generated map, it still serves the purpose of illustrating what a paper map usually looks like. If you are new to the fields of GIS and mapping, and are asked to find a way to organize the geographic features shown on the map in a computer, how would you get started and conceptualize it? Moreover, what data structures would have the power and versatility to support sophisticated analysis of geographically referenced data to solve complex problems beyond mapping?
Figure 1  A typical reference map of San Marcos, Texas.
In the rest of this entry, we will first walk through some of the key developments in the evolution of vector data structures in GIS together. These developments led to the topological vector GIS data structure that is widely used in different parts of the world today. Then, we will manually construct a vector data structure using an example to understand how it actually works. Furthermore, we will illustrate the power of topological vector data structure in supporting computations in GIS and discuss some of the constraints of the existing data structures as well as the challenges related to the development of vector data structures.

Evolution: contributions from the US Census Bureau

The simplest and earliest vector data structure is the so-called “spaghetti model” – a direct line-for-line translation of the paper map using simple points, lines, and polygons. Among other problems, the spaghetti model does not represent the inherent nature of topology between features on a map. Topology refers to the relationships between geographic features represented on a map that do not change after geometric transformations such as translation, scaling, and rotation. Examples of these relationships include connectivity and adjacency. Since the beginning of the development of computer mapping and GIS technologies in the late 1950s and early 1960s, many researchers naturally began to search for a vector data structure that could adequately represent both geometry and topology.

Although Robert Dial developed a data structure related to mapping in his master’s thesis in the early 1960s (Dial 1964), and Tomlinson stated in his 1984 paper in the Operational Geographer that the Canadian GIS developed in the 1960s had functions to correct topology in the data (Tomlinson 1984), it is now widely recognized that the most influential innovations in the development of topological data structure have been through the contributions from the US Census Bureau, beginning in 1967 in New Haven, Connecticut, in the United States. Milestones in the developments at the Census Bureau include: (i) the development of the DIME (dual independent map encoding) in 1967 (Cooke and Maxfield 1967; Cooke 1998); (ii) the development of a 2-D version of DIME in the 1970s and early 1980s that helped the US Census Bureau complete its digitizing project in time to support operations related to the 1980 decennial census (White et al. 1981; Lockfeld and Smartt 1984; Cooke 1998); and (iii) the development of the nationwide TIGER (topologically integrated geographic encoding and referencing) database from 1982 to 1990 to support the 1990 decennial census (Marx 1986; Cooke 1998). Table 1 is a summary of the data structures and key individuals involved in these developments.

The US Census Bureau made a decision in the mid-1960s to use a mail-out/mail-back approach in its 1970 decennial census instead of employing enumerators to deliver questionnaires to households and collect the completed questionnaires in the field. To prepare for the mail-out/mail-back operation, the Census Bureau established a research unit called Census Use Study in New Haven. Among other responsibilities, researchers of the New Haven Census Use Study team were charged to automate the processes of geocoding each address to a census block and evaluate methods of mapping that could be used to support operations of the Census Bureau.

Donald F. Cooke and William H. Maxfield were hired by the New Haven Census Use Study
team in early 1967 while finishing their studies at Yale. One of their tasks was to conduct the mapping experiments. Naturally they started with information from the New Haven Address Coding Guide (ACG) provided by the Census Geography Division in their mapping experiments as early as February 1967. The ACG used block side (one side of a city block) as the geographic entity to record address information for census enumeration. Basic information in each record of the ACG included: street name, state code, county code, place code, ZIP code, census tract number, block number, and address range.

In their initial experiments, Cooke and Maxfield used coordinates of the two ends of each block side and the line between the two ends to map a block side, amended the coordinates to the ACG records, and generated maps using the amended ACG data. They quickly realized that the ACG was not a good geographic base file for mapping. Among other problems (Cooke and Maxfield 1967), each cross-street intersection had to be read as many as eight times. In addition, the “true” shape of a block side of a curvilinear street was lost when only the two ends of a block side were recorded.

Because of the limitations of the ACG as a geographic base file for supporting mapping, James Corbett of the Bureau’s Statistical Research Division and George Farnsworth, then working as a statistical intern at the Bureau, proposed an alternative mapping database structure (Cooke and Maxfield 1967). This data structure was initially called dual incidence matrix encoding (DIME) and renamed to dual independent mapping encoding later. DIME is theoretically grounded in graph theory. More details about the topological principles of maps can be found in a technical paper of the Census Bureau (Corbett 1979).

Technical innovations in the development of DIME, as we see it today, include: (i) the introduction of topological principles from graph theory in applied mathematics to mapping; (ii) the use of the line (street) segment as the most fundamental unit of the data structure in mapping; (iii) the first use of “node” in a (cartographic) data structure, the beginning of “assigning numbers (identifiers) to nodes,” as well as organizing the numbered nodes in DIME; and (iv) the design and implementation of the “topological editing” procedures for quality control. Cooke implemented the procedures for topological editing without directly using the concepts of incidence matrices introduced by Corbett (Cooke 1998). This innovation is also the beginning of the algorithms for “topology building” of vector data in subsequent generations of GIS in the decades following the invention of DIME. More details about “topological

<table>
<thead>
<tr>
<th>Data structure</th>
<th>Year(s)</th>
<th>Key individual(s)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIME</td>
<td>1967</td>
<td>James Corbett, Donald F. Cooke, and William H. Maxfield</td>
<td>Cooke and Maxfield (1967); Corbett (1979); Cooke (1998)</td>
</tr>
<tr>
<td>2-D DIME (precursor to TIGER)</td>
<td>1970s–1981</td>
<td>Marvin White</td>
<td>Corbett (1979); White et al. (1981); Lockfeld and Smartt (1984); Cooke (1998)</td>
</tr>
</tbody>
</table>
Despite the innovative features in DIME, it suffers several limitations. Among the limitations, it is very inefficient to retrieve and update the records because the line segments are stored in the system without any particular order. This means that all records in an entire file had to be scanned to search for a particular line segment. This is a very time-consuming and inefficient process from a database management perspective.

Because of the limitations of computing technologies at that time, the DIME system developed by the New Haven Census Use Study team was an offline and batch processing system that did not provide graphic feedbacks to operators who performed digitizing and editing activities. A transition to an online interactive environment with 2-D graphic display had to be made to overcome the limitations of the original DIME system. This transition was achieved about a decade later, in the late 1970s, by Marvin S. White, who was a researcher at the Statistical Research Division of the Census Bureau from the 1970s to the early 1980s.

Starting with the online DIME editor called ARITHMICON left by Corbett at the Census Bureau after Corbett’s retirement, White designed a 2-D version of DIME and implemented it in a microcomputer environment using Pascal (White et al. 1981; Lockfeld and Smartt 1984). The system was based on the topological principles of 2-D maps described by Corbett (1979). It used 0-cells, 1-cells, and 2-cells to organize nodes, lines, and blocks. The system employed the B-tree data structure to index the data for faster data retrieval and manipulation. It provided an interactive environment for 2-D graphic display that facilitated a much more efficient process of creating and updating DIME files (Lockfeld and Smartt 1984). Under the pressure of meeting deadlines of the digitizing project to support census operations related to the 1980 census, the Geography Division of the Census Bureau implemented the system based on White’s design. This newly implemented system helped the Geography Division complete the digitizing project on time (Cooke 1998).

The usefulness and potential of the 2-D system apparently played a significant role in convincing managers at the Census Bureau to make the decision in 1982 to develop a nationwide TIGER database to support the 1990 decennial census (Cooke 1998). In the 1980s, Robert Marx, then serving as the Chief of the Geography Division at the Census Bureau, spearheaded the development of the first nationwide TIGER database through collaborations with the US Geological Survey (USGS) (Marx 1986).

The overall architecture of TIGER is well documented (Marx 1986; US Census Bureau 2006). Theoretically, it is based on the topological principles of 2-D maps described by Corbett (1979). Although the basic units (nodes, lines, and polygons) of TIGER are still the same as the ones used in DIME, the data structure of TIGER is more sophisticated. TIGER uses lists to store the topological relationships among 0-, 1-, and 2-cells. The 0-cell and 2-cell lists can be accessed through directory files that are the entry into the TIGER database (Marx 1986). Directory files are stored in a B-tree structure to enable efficient access, fast manipulation of data, and optimal computer storage. Directory files contain one record per cell and a pointer to the corresponding record in the cell list. The list files contain more fields than directory records. For example, the 0-cell list contains the coordinates for each record, selected attributes, as well as a pointer to the first record in the 1-cell list that has this 0-cell as one of its end points. The 1-cell
lists hold feature attributes and several pointers (e.g., a pointer to the start and end 0-cells and a pointer to the left and right 2-cells). The 2-cell lists include several feature attribute fields and pointers such as pointers to 1-cells (Marx 1986).

The TIGER database and the products from the Geography Division of the Census Bureau have evolved over the years. The Census Bureau adopted the terminology used in the spatial data transfer standard (SDTS) (USGS 1992). In the 2006 edition of the TIGER database, the concepts of complete chain and GT (geometry and topology)-Polygon replaced those of line and polygon mentioned earlier in this chapter. In addition, the concepts of spatial objects, features, left- and right-side data fields, and single-layer topology are used (US Census Bureau 2006).

Spatial objects are the digital representation of real-world spatial phenomena (e.g., Texas State University and Spring Lake). Spatial objects defined in the TIGER database support geometric drawing and topological operations, which have coordinate locations, shapes, as well as topological relationships. There are three types of spatial objects in the TIGER database: zero-dimensional (e.g., node, entity point, and shape point), one-dimensional (e.g., complete chain or network chain), and two-dimensional (e.g., GT-Polygon) spatial objects. The term feature defined in the TIGER database is used to describe spatial objects that are more complex than node, complete chain, or GT-polygon. For example, a street that consists of a set of complete chains is defined as a feature in the TIGER database. The left- and right-side data fields list information about the left and right spatial objects or features of a complete chain based on the start node and end node of that chain. Therefore, the fields provide necessary information about complete chains for users to construct polygons or features.

**Evolution: development of vector data structures at other institutions**

The evolution of vector data structures extended far beyond the efforts at the US Census Bureau. Reports about research activities of the development of vector data structures by other institutions in the United States and different institutions in the rest of the world are scattered in the literature. It is beyond the scope of this entry to provide a comprehensive review. Other institutions in the United States that made significant contributions include the University of Washington, Harvard Laboratory for Computer Graphics and Spatial Analysis, USGS, Environmental Systems Research Institute (ESRI), and Intergraph Corporation.

As mentioned earlier, one of the earliest documented innovative research activities in the United States was Robert Dial’s master’s thesis research completed in 1964 at the University of Washington under the supervision of Professor Edgar M. Horwood (Dial 1964; Mark et al. 1997). The title of the thesis is “Street Address Conversion System” (SACS). SACS was developed to meet the need for geocoding street addresses in a large database. It translates street addresses to point coordinates using an efficient segment file structure. Although similar to DIME, the system developed by Dial had an important innovation which was not available in the initial implementation of DIME — the use of a grid for indexing nodes (Mark et al. 1997).

In the 1970s, the Harvard Laboratory for Computer Graphics and Spatial Analysis implemented the concept of POLYVRT (POLYgon conVeRT) (Peucker and Chrisman 1975). POLYVRT is a hierarchical vector data structure. One important contribution of POLYVRT is the introduction of the concept of “chain” that is widely used today. In addition, POLYVRT explicitly defines neighborhood relationships...
DATA STRUCTURE, VECTOR

that are important in certain applications of spatial analysis.

The digital line graph (DLG) developed by the USGS is another vector data structure. It was converted from hardcopy maps and stored in ASCII format. DLG files are available at 1:2 000 000, 1:100 000, and 1:24 000 scales. After several revisions, the DLG–3 data structure in the mid-1980s consisted of three basic elements: nodes, lines, and areas. Line is the basic element and users must start by processing line records. DLG contains one-way pointers to nodes at each end of a line and to areas on each side of a line. Unlike some other vector data structures, DLG does not explicitly store attribute information but stores unique feature codes to represent the attribute features.

In 1992, the USGS released the spatial data transfer standard (SDTS) to meet the need of supporting data transfer and exchange (USGS 1992). The SDTS facilitates Earth-referenced spatial data transfers between different computer systems, while minimizing the potential information loss. It includes standards for meeting a wide range of transfer needs such as spatial data concepts, attributes, data structures, and logical and physical file structures. Although SDTS cannot be categorized as any data structure type, it is mentioned in this entry because of its importance in facilitating transfer of vector data.

Among other data structures, ESRI introduced the arc–node data structure (coverage) and shapefiles to organize vector data in GIS. In the arc–node data structure, “arcs” are the same as “lines” and “chains” mentioned earlier in this entry. Topology is represented in the arc–node model in a way similar to those of DIME and TIGER. Like most of other vector data structures, shapefile represents spatial objects using points, lines, and polygons. However, shapefiles use a nontopological vector data structure that does not have built-in topology in the data.

A shapefile consists of three separate and distinct types of files: main files, index files, and database tables. The main file is a direct-access file that contains the shape as a list of vertices. The index file contains character length and offset information for locating an element. The database table contains attributes of the shapes. A main advantage of shapefiles is that it allows rapid cartographic representation of large geographic datasets.

Autodesk developed the design exchange format (drawing exchange format now) (DXF) in 1982. DXF is used for communicating design information between computer-aided design (CAD) and other programs. The DXF data structure is a tagged data representation that uses an integer number (called “group code”) to define each data element. Unlike some vector data structures that use features to define spatial objects, DXF format uses listing elements consisting of strings of vertices. A DXF file consists of several sections: header, classes, tables, blocks, entities, objects, thumbnailimage, and end of file. The DXF format has data redundancy because it is a nontopological data structure.

Two other vector data structures, GeoMedia and imagestation feature collection (ISFC), were introduced by Intergraph Corporation. GeoMedia has no proprietary data format. It accesses data in a variety of formats (e.g., ESRI shapefile and CAD data) using technologies such as geographic data object and data server without translation. ISFC is a MicroStation-based format that can be used to facilitate digital photogrammetry workflows.

GeoJSON is an open standard format used by many open source GIS packages. Like other vector data structures, GeoJSON uses points, lines, and polygons to represent spatial objects. Additionally, it uses multipart collections of these three types of data, including multipoint, multilinestring, and multipolygon. However,
compared with the data structures described above, GeoJSON was developed and maintained by a group of volunteers. It also differs from some other vector data formats in that it uses objects consisting of members. Each member in GeoJSON has a name represented by a string as well as a value represented by a number, string, object, or array.

A vector data structure can be implemented using the keyhole markup language (KML). KML was initially developed by Keyhole Incorporation (now part of Google). In 2008, KML became an international standard of the Open Geospatial Consortium (OGC). As of 2014, KML is one of the most widely used ways to support operations related to vector data. For example, it can be used to describe geographic annotation and visualization (e.g., polygons, place marks, and 3-D models) involving internet-based 2-D or 3-D maps.

State of the art of vector data structure by example

It helps to understand the concepts related to vector data structure if a topological data structure is manually constructed for the features shown in Figure 2 using a five-step procedure. Recall that a node is used to represent either an intersection of two or more linear features or the end of a linear feature. The direction of a chain starts at a “from node” and ends at a “to node” (Figure 3). Based on the direction of a chain, its left-side and right-side polygons can be determined. In the five-step procedure, we first define the nodes and assign an identifier (ID) to each node. Second, we give an ID to each chain connecting a pair of nodes. Third, we construct a “Node” table. Each record in the “Node” table contains all attributes associated with a node, including the coordinates representing the location of the node. Fourth, we build a “Chain” table. The essential attributes of a record in the “Chain” table include an ID, the “from node,” “to node,” left-side polygon, right-side polygon, and any other attributes associated with a chain. Fifth, we construct the “Polygon” table. The contents of the tables are shown in Figure 3.

In practice, most of the work in the five-step procedure mentioned above can be done automatically using topology-building functions available in a GIS. After digitizing the coordinates characterizing the shape of a chain into GIS from a paper map, a commercial GIS at present has the capabilities to build the three tables mentioned above and the topological information in the tables automatically. When building topology, planer enforcement requires that all features (node, chains, and polygons) are on a 2-D surface. No overlaps among chains or polygons are allowed on the same layer. Additionally, there must be an intersection at each chain crossing.
Figure 3  Basic components of a topological vector data structure for the nodes, chains, and polygons shown in Figure 2.
Power of vector data structure to support database management and analysis

The vector data structure plays a vital role of supporting database management. It lays the foundation to facilitate computational procedures that are necessary to solve complex problems involving processing geographic data in a vector form. The topic of spatial database management is covered in a different chapter. As for the computational aspect of vector GIS, it is not the intention of this entry to provide a comprehensive coverage. Instead, we only illustrate the usefulness and power of vector data structure in supporting computational procedures using some simple examples.

These computational procedures in vector GIS may be used to support spatial search, spatial analysis, and other sophisticated analysis and computational tasks. Spatial search is used to provide an answer to a query about geographic data based on spatial relations between geographic objects. These relations may include distance relations, directional relations, and topological relations. Because coordinates of features are explicitly stored in a vector data structure, the search of the locations of features can be easily accomplished. Distance calculation is a simple procedure based on the coordinates representing locations of the features in question. Additionally, a spatial search involving topology is more efficiently supported in a topological data structure than that in a nontopological data structure because the relationships among points, lines, and polygons are well defined. For example, the task of searching for countries sharing a border can be efficiently solved by searching the left and right polygons representing the countries sharing a border from the chain topology table illustrated in Figure 3.

Spatial analysis refers to the analysis of geographically referenced data using statistical and/or other quantitative methods to reveal patterns hidden in the data and gain insights about the phenomena observed. Vector data structure facilitates spatial analysis in performing analytical and computational procedures involving topological relations such as adjacency, connectivity, and containment. For example, it is often necessary to determine adjacent polygons in some tasks of spatial analysis. These tasks can be easily accomplished by “walking” along the boundaries of a polygon in question when the polygon is represented in a topologically structured vector format.

As a simple example, imagine that you have a map showing several layers of geographic features covering the United States. These layers include cities, counties, major roads, water bodies, and parks. A simple spatial search that one may perform is to “find all counties with a passing-through river whose county seats are within 200 km of a straight distance from the City of San Marcos in Texas.” This is a simple spatial search that involves a topological relation and a distance relation. In vector GIS, this search can be easily accomplished using a few steps. First, calculate the distance between the location of each county seat in Texas and the city of San Marcos, and select the counties whose county seats are within 200 km from San Marcos to form the first set of counties. Let us assume that the data are in the Cartesian coordinate system and the effect of the curvature of the Earth can be ignored for the sake of simplicity. The distance is a simple Euclidean distance. Second, determine if a county in the first set of counties has a passing-through river. In computational terms, this means that we will check if any chain forming the boundary of a county in the first set of counties intersects with a chain representing a river. Any county with its boundary intersects with a river is kept in the set of counties. Otherwise, a county is deleted from the set of
counties. The counties left in the set are the counties satisfying the search criteria.

Topological overlay is one of the most widely used operations in GIS. It is a computationally intensive operation when the overlay involves two map layers each consisting of many polygons. The overlay operation first determines the boundaries of new polygons in the output map layer from the two input map layers, and then updates the attributes of the new polygons based on the rules dictating the overlay. For example, Figure 4a is a map layer showing polygons of four different soil categories (I, II, III, and IV) in an area of interest, Figure 4b is a map layer of polygons covering the same area, and the type of planned development in each area covered by a polygon is marked as residential (R), industrial (I), commercial (C), and agricultural (A). The city wants to obtain a map showing areas that are planned for residential development and have either soil category I or III.

The topological data structure in vector GIS can be used to support this type of overlay analysis. In addition to support operations involving other spatial relations, topological data structures facilitate analyses concerning adjacency, coincidence, and connectivity. Coincident topology allows one boundary from a layer to lie on top of another boundary from another layer. In the example presented in Figure 4, because layers 1 and 2 share some common edges along boundaries (coincident geometry), coincident topology in vector data structure can facilitate the overlay analysis between these two layers. Figure 4c is the output map layer showing the resulting polygons marked in different colors. Polygons in light blue are the areas that are planned for residential development with either soil category I or III. Polygons that are not colored are areas that do not meet the conditions.

There are several computational procedures involved in this overlay. First, for each chain forming the boundary of each polygon in the first input map layer, the GIS checks the chain against every chain forming the boundaries of polygons in the second input map layer and determines if the two chains intersect with each other. If the two chains intersect with each other, then the GIS computes the intersection of two chains, and records that intersection as a node of a new chain in the topological data structure of the output map layer. This process is repeated for all possible pairs of chains of the two input map layers that intersect. Second, the GIS builds the topological data structure of the output map layers and generates new polygons in the output layers and generates new polygons in the output

![Figure 4](image)

**Figure 4** An overlay example: (a) input map layer 1: polygons showing areas with different soil categories; (b) input map layer 2: polygons illustrating areas planned for different land uses; and (c) output map layer: initial resulting polygons after the overlay showing residential areas with either soil category I or III (in light blue) and other areas.
map layer. Third, the GIS assigns an attribute to each polygon in the output map layer based on the rules dictating the overlay operation.

In actual implementations, the efficiency of the computational procedure and exact algorithm can be improved by limiting the search space of the pairs of the intersecting chains in the two input map layers to reduce the search space. These topics are beyond the scope of this entry. From a computational perspective, the “line intersection” procedure and the “topology construction” procedure are two critically important computational procedures for supporting the overlay operations in GIS. Indeed, they are two of the most fundamental and widely used computational procedures in vector GIS and play important roles in a variety of applications.

The computational power of vector GIS can also be observed in its functions to support spatial analysis. For example, spatial autocorrelation is a fundamental concept in spatial analysis. It means that the observed value of a variable at one location is correlated with the values of the same variable in nearby locations. When areas are used as the geographic units for observation, almost all existing methods for determining whether or not spatial autocorrelation exists in a given dataset use a contiguity matrix to measure the nearness of area units represented by polygons. The contiguity matrix then is used to calculate a statistic to evaluate the existence of spatial autocorrelation. The contiguity may simply be the number of boundaries shared by two adjacent polygons. When the number of polygons is large, it is a time-consuming process to search for all adjacent polygons of each polygon in the study area. Fortunately, the topological data structure in vector GIS provides a convenient and efficient way to facilitate this type of search, and supports the computational tasks related to spatial autocorrelation analysis.

For organizing, processing, and visualizing objects in 2-D space, vector data structures described in this entry serve their purposes well for most applications. It is almost a necessity to use a vector data structure in applications where the demand on positional accuracy is high. For data involving more than two dimensions, 3-D vector data structures and other data structures are needed to adequately represent and process 2-D data. The development of vector data structures to support the management and handling of higher dimensional data is a challenging task and it is discussed elsewhere.

**SEE ALSO:** Data structure, raster; Data structure: spatial data on the web; Indexing; Routing and navigation; Spatial analysis; Spatial database; Spatial feature classes

**References**


Lockfeld, Frank, and Brian Smartt. 1984. “2-D: Address Matching and Dime Maintenance.” In *Proceedings of the 22nd Annual Conference of the*
DATA STRUCTURE, VECTOR

_Urban and Regional Information Systems Association_, 194–206. Seattle, WA.


Further reading


Decision analysis

Yupo Chan
University of Arkansas at Little Rock, USA

Decision analysis is a key methodology in the field of “methods, models, and geographic information sciences,” or “GIScience and technology” for short. The term decision analysis in the context of GIScience usually refers to the siting or locating of an object, facility, or land use on the landscape. For example, locating an airport is a complex decision that is the result of a consensus between the public and the authorities, if not beyond, who must account for a variety of considerations. Central to classic decision analysis is the concept of utility, which serves to quantify the value of a decision; in this case, the value added to the community by the new airport. The public, the authorities, and other stakeholders need to agree on the fact there is value in constructing a new airport and ideally account for costs and benefits. To quantify value, a utility function is employed to convert different attributes to a single measure called “utile;” these attributes may include accessibility to the airport and environmental concerns among others. Following classic multi-attribute utility theory (MAUT), these attributes need to be collapsed into a single measure, or a “common currency of exchange,” through a utility function. The process by which this utility is established is the heart of decision analysis.

Basics

The last 50 years have witnessed significant advances in the field of MAUT, in which meticulous methodologies have been crafted to overcome many challenges, and the field is still evolving. A central challenge is the fact that attributes can be incommensurate. Combining them may very well require comparing “apples” to “oranges;” for example, how does the decision analysis process reconcile the cost of noise pollution with the benefits of economic activity that come with an airport? More than that, there are many other intangibles, such as the attitudes of the public or the attitude of the authorities, both of which are difficult to quantify.

The simplest case of decision-making is termed dominance, which allows straightforward ordering among alternatives. An example is a “less expensive and better quality” shirt in a department store, which is preferred by all shoppers, all other factors being equal. The shirt that is both less expensive and better quality simply dominates other shirts because they are more expensive and of poorer quality. Note that a nondominated alternative is not necessarily a singleton, meaning that there are usually several shirts (if not more) in a department store that are equally attractive to shoppers. In fact, there are typically a huge number of nondominated solutions (many different shirts in this example) and the challenge is to focus on the ones that are most appealing.
More complicated than dominance is a situation where tradeoffs are involved, or Pareto preference. Continuing the example of shirts, the nondominated shirts may include those that are superior in quality yet more expensive, and others that are the less expensive ones but may not last as long. Here the tradeoff is between cost and longevity as a proxy of quality. The field of multicriteria decision-making (MCDM) addresses these tradeoffs, where the two criteria under discussion are price and quality. Multicriteria optimization (MCO) is a key component of MCDM besides MAUT. Central to MCO is the concept of Pareto optimality, defined as the choice in which one or more criteria achieve a higher utility without compromising any other criteria. This precisely describes the diversity of shoppers’ decisions with different pocket books while the department stores seek to stock items that will please a cross-spectrum of shoppers.

Of particular interest in MCO is interactive multicriteria optimization (IMCO). In considering shirts, buyers do not negotiate with the shirt manufactures, but in many real-world applications of decision analysis, the process of assessing options is interactive. With the airport example, the stakeholders’ perception of value will change as they negotiate and interact with one another. There is no rigid utility function that remains unchanging in many real-world problems, especially those that are controversial. By interacting with one another, the stakeholders’ preferences change, and through negotiation these evolving preferences are often revealed. Due to its applicability to a range of problems, IMCO is an exciting body of knowledge that has drawn a lot of attention in recent years, and new concepts are still emerging.

This entry uses a case study of choosing an airport location to illustrate the IMCO process, in order to examine how decision analysis works. First, the context of the airport-location case study is introduced to illustrate generic features of decision analysis, starting with basic facility-location concepts. (See elsewhere in this encyclopedia for details on facility location.) Such concepts include medians and centers, where a median is an airport location that is closest to all the travelers on the average, and a center is one that is closest to the furthest traveler. This is followed by defining a multi-attribute utility function that captures the incommensurate attributes of accessibility to the airport and environmental concerns in a single measure. This is accomplished by an interactive process rather than by a static function because the attitudes of the stakeholders, including the public and the authorities, emerge as they work together. More formally, an IMCO procedure is carried out as it is emerging, which means that it does not depend on an explicit definition of the utility function. This implicit quantification process shows the potency of MCDM in addressing intangible values and representing the intent of this chosen case study. Decision analysis is then further shown to shed new light on some of the classic problems of facility location in GIScience and technology. This entry ends with a projection of how the field of decision analysis will evolve to serve the GIScience and technology community.

Interactive location decisions: an airport example

Two central concepts in facility location are median and center. Some facilities, such as a distribution center for goods or services, are best located at the median, so that the retailer can best serve its clientele in aggregate. In contrast, a fire station may be located at a center, so that the furthest citizens are as likely to receive a fast response. Elsewhere in this encyclopedia, the reader can find procedures to locate medians and
centers in networks and planes (in the sense of a surface). Let us define a network as a representation of a geographic area in terms of nodes (such as street intersections) and arcs (such as the streets themselves). A plane is simply a blank map on which there are no streets or intersections. For a linear-cost function of accessibility, Hakami’s extremal condition (Hakami 1964) suggests facilities tend to locate at extreme points, or at a node of a transportation network rather than a “central” location on an arc. This is sometimes referred to as the nodal-optimality property. For a simple two-node network linked by an arc, a facility will be placed at either end of the single arc. Likewise, a facility will be located at the three nodes that make up the three corners of a triangular network connected by three arcs. Locating a center will be different, with the airport placed in between the two ends (on the arc) even for a linear cost function (when \( \beta = 1 \)). In other words, the optimal airport location is invariably at either end of the arc connecting the two nodes should travel cost assume a linear function of distance, such as \( c_{ij} = d_{ij} \). It can be shown that the airport will be located somewhere other than at the end points on the arc should noise pollution be the only concern. In this case, the cost function is increasing return to scale, or more precisely with \( \beta < 1 \). An example noise cost function is \( c_{ij} = d_{ij}^{-2} \). As plotted in Figure 1, one can see the progression of \( \beta \) as it keeps decreasing from \( 1/2 \) to 0, and onto \(-2\). It suggests that noise can be mitigated at a farther distance.

The case study and solution algorithm

Consider the example of selecting a site between Cincinnati (C) and Dayton (D), Ohio, for a regional airport so that both travel cost and noise impacts are minimized. The Cincinnati and Dayton metropolitan regions have populations of approximately two million and one million, respectively. Consider the simple linear network shown in Figure 2, where \( x_1 \) is the distance from Cincinnati and \( x_2 \) is the distance from Dayton. In
this referencing scheme, an airport located at the mid-point on the arc will carry the coordinate \((x_1, x_2) = (30, 30)\), or thirty minutes from either Cincinnati or Dayton. Travel cost is simply \(2x_1 + x_2\) passenger-minutes-of-travel, and noise impact is simply \(2x_1^2 + x_2^{-2}\) passenger-decibels.

The locational decision can now be modeled by the following MCO mathematical program:

\[
\begin{align*}
\min \ & \nu(f_1(x), f_2(x)) = \nu(2x_1 + x_2, 2x_1^2 + x_2^{-2}) \\
\text{s.t.} \ & x_1 + x_2 \geq 60 \\
& x_i \geq 0 \quad (i = 1, 2)
\end{align*}
\]

While travel cost \(f_1(x)\) and noise impact \(f_2(x)\) are quantified, notice the model is set up without specifying the precise form of the utility/value function. It is simply expressed as an open function of the two attributes \(f_1(x)\) and \(f_2(x)\), \(\nu(f_1, f_2)\). In this model, the criterion functions \(f_1\) (for accessibility) and \(f_2\) (for noise impact) are not necessarily all linear. In particular, the noise function, with a \(\beta\) value of \(-2\), certainly has \(\beta << 1/2\), suggesting a strongly increasing return to spatial scale, as shown in Figure 1.

The undefined value function and the nonlinearity of noise impacts complicate the solution of this MCO problem. Let us propose a simple solution by a gradient search procedure based on a linearized value/utility function, and an interactive marginal-substitution procedure. The former will approximate a nonlinear, yet to be determined, value function by its local tangent. The latter will take care of the two criteria by evaluating \(\lambda_1\) versus \(\lambda_2\), defined as the weight placed on travel cost and noise impact respectively. As will be shown, the Frank–Wolfe (F–W) gradient search procedure may be employed for the former. To apply the F–W solution method, the following condition must exist: all \(f_i(x), i = 1, 2,\) are convex and continuously differentiable in their respective domain and the constraints form a convex and compact set (a set that is closed and bounded).

Suppose the feasible region for the mathematical program shown in equation 1 is defined as \(\kappa\) in Figure 3. This 2-D figure illustrates the computational procedure of the F–W algorithm, should it start at a chosen initial feasible solution \(x^0\). First, the tangent at this point is constructed and taken as an objective function. Together with the feasible region \(\kappa\), the solution to this linear program (LP) is obtained, yielding the interim optimum at location \(x^1\). Another tangent is now constructed at \(x^1\). Together with the first linear objective-function and the feasible region \(\kappa\), it forms another LP. For the current LP, the optimal

![Figure 3](Frank–Wolfe algorithm.)
solution is again obtained at \( x^2 \). This procedure is repeated as many times as required. It can be seen that after \( k \) iterations an optimal solution \( x^k \) to the original nonlinear mathematical program, consisting of the bimodal objective function and the linear feasible region \( \kappa \), is obtained. Notice step size determination is part of the F–W procedure, as shown by the step size \( \alpha d^0 \) in this figure for the initial step. Here \( d^0 \) is the initial gradient of the climb, and \( \alpha \) is the step size, with \( \alpha \) ranging between 0 and 1. The F–W procedure, while being a robust algorithm, can only find local optima, rather than global optimum. (There is no guarantee of a single, best solution.) Whether a local or global optimum is obtained will be determined by the choice of the initial feasible solution or the starting point, in this case \( x^0 \). It can be shown that should one start at an alternate location, \( x^{0'} \), at the other end of the feasible region \( \kappa \), a global optimum will be obtained.

Importantly, the F–W procedure can be used in an interactive manner by stakeholders to overcome a “hidden” value function form, where the explicit form of \( v() \) is unknown. Taking the gradient of the value function yields:

\[
\nabla_x v(f) = \frac{\partial v}{\partial f_1} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} \end{bmatrix} + \frac{\partial v}{\partial f_2} \begin{bmatrix} \frac{\partial f_2}{\partial x_1} \\ \frac{\partial f_2}{\partial x_2} \end{bmatrix} \tag{2}
\]

Let us approximate the nonlinear value function by its tangent at the location \( x \) where the gradient is evaluated by equation 2. Setting \( \partial v/\partial f_1 = \lambda_1 \) and \( \partial v/\partial f_2 = \lambda_2 \), with both evaluated at alternate airport locations \( x^0, x^1, x^2, \ldots x^k \), and so on yields a tangential linear function for each alternate location \( x^k \) (shown in Figure 4):

\[
\lambda_1 f_1 \big|_{x^k} + \lambda_2 f_2 \big|_{x^k}
\]

The relevant decision-makers decide interactively on the marginal rate of substitution \( \lambda_1/\lambda_2 \) to trade between accessibility and noise impacts. At the initial location \( x^0 \), halfway between Cincinnati and Dayton, for example, the decision-maker is asked about the increment of travel cost) \( f_1 \) for which s/he is willing to trade against a decrement of aircraft noise) \( f_2 \). The slope of this indifference curve is precisely \( -\lambda_1/\lambda_2 \), which is perpendicular to the gradient \( \nabla v(f_1, f_2) \). Suppose the decision-makers decide that the marginal rate of substitution is “fifty-fifty,” or \( -\lambda_1/\lambda_2 = 1 \). Without loss of generality, let us set \( \lambda_1 = 1 \), which means \( \lambda_2 = 1 \) in this example. The initial gradient at the half-way point between Cincinnati and Dayton \( x^0 = (30, 30) \) is now obtained as follows:

\[
\nabla_x v(f(x^0)) = \lambda_1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} -4x_1^{-3} \\ -2x_2^{-3} \end{bmatrix} \big|_{x^0 = (30, 30)} = \begin{bmatrix} 2\lambda_1 -0.00015\lambda_2 \\ 1\lambda_1-0.00007\lambda_2 \end{bmatrix}
\]

This gradient leads toward an equivalent LP:

\[
\min_{x \in \chi} \nabla_x v^T(f(x^0)) x = \min_{x \in \chi} (2\lambda_1 - 0.00015\lambda_2, \lambda_1 - 0.00007\lambda_2) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}
\]
Now substitute $\lambda_1 = \lambda_2 = 1$ into the above. By solving the following LP, the optimal solution $x^* = (0, 60)$ is determined:

$$\min \{2x_1 + x_2 | x_1 + x_2 \geq 60; x_i \geq 0 \} \quad (i = 1, 2)$$

Thus at this iteration, we are moving the airport toward Cincinnati from the halfway point between the two cities according to the steepest “ascent” direction $d^0 = x^* - x^0$, where $x^0 = (30, 30)$ and $x^* = (0, 60)$ or $d^0 = (-30, 30)$. In the space of the LP, this is illustrated in Figure 5.

**Algorithmic iterations**

The decision-makers (DMs), consisting of the authorities and the public, now determine the step size $\alpha$ to move along the direction $x^0 + \alpha d^0$. The DMs, assisted by tabular or graphic displays of the function $f(x^0 + \alpha d^0) = [f_1(x^0 + \alpha d^0), f_2(x^0 + \alpha d^0)]$ determine the step size $\alpha^0$ between the values of 0 and 1. One possible way to obtain the best step size $\alpha$ is to display the values for the two criterion functions $f_i(x^0 + \alpha d^0)$ for $i = 1$ and 2 as a function of $\alpha$ over the selected values of $0 \leq \alpha \leq 1$ in a tabular or graphic way. The DMs then determine a value of $\forall$ for the most preferred values of the corresponding criterion functions. In short, the following optimization problem is solved: $\min_{0 \leq \alpha \leq \nu} (f(x^0 + \alpha d^0))$. Suppose $\alpha^0 = 0.5$.

We are now at $x^1 = x^0 + \alpha^0 d^0 = (30,30) + 0.5(-30,30) = (15,45)$, and the iterations continue until the incremental “ascent” of the preference function $\nu$ is minuscule, as with most “hill-climbing” algorithms.

To fill in the details, let us now derive the example travel function for starting point $(30, 30)$, which results in $d^0 = (-30, 30)$ for the first step of the F–W procedure, moving toward Dayton with a step size $\alpha$ yet to be determined. This is shown in Figure 5, with $\lambda_1$ and $\lambda_2$ already determined to be equal to 1 each. The linear travel-cost function is now a function of the step size $\alpha$, as shown in Figure 6: $f_1(x^0 + \alpha d^0) = 2(30 + \alpha [-30]) + (30 + \alpha [30]) = 90 - 30 \alpha$. Notice the function is decreasing with step size, suggesting the closer we move the airport toward Cincinnati, the less passenger-minutes-of-travel becomes. Similarly, the nonlinear noise function is now a function of the step size $\alpha$. The parametric noise function used for determining the step size $\alpha$ is the following for the initial $(30,30)$ location: $f_2(x^0 + \alpha d^0) = 2(30 + \alpha [-30])^{-2} + (30 + \alpha [30])^{-2} = 3(30 + \alpha [-30])^{-2}$, and the function can be plotted as Figure 7. Notice the noise function is increasing with step size, in contrast to the
travel cost function, which is decreasing. Notice also that the travel-time and noise functions will change as the potential airport location changes.

Both the travel and noise functions are displayed side-by-side in Figure 8. With such a graphic display, the travelers and environmentalists are invited to reach a consensus on a common step size $\alpha$, where they are comfortable with both the travel cost versus noise levels. It can be shown that for equal weights upon the two criteria, that is, the decision-maker is indifferent between noise and travel and a fixed step size of $\alpha = 0.5$, an airport location converges. From Table 1, it can be seen that while the airport steadily moves toward Cincinnati in the first five iterations, starting at iteration 6, there is a directional reversal toward Dayton. If we make subsequent iterations, it will begin moving toward Cincinnati again (Figure 2). However, there is a limit to this westward movement — namely at a point east of the previous reversal point $(0.9375, 59.0625)^T$. This point is further from Cincinnati and closer to Dayton, or $x_1 > 0.9375$ and $x_2 < 59.0625$. As this point subsequently moves east toward Dayton again, there is once again a limit to its movement. In this case, it falls short of $(30.46875, 29.53125)$ — the previous “point of reversal.” Thus, the airport location bounces back-and-forth within a shrinking interval, eventually converging toward a final equilibrium point at $(1.587, 58.413)$, or about 1.6 minutes (or 2.6 km) outside Cincinnati.

As shown in Table 1, each computational step simply recommends another potential location for the airport, starting with the midpoint $(30, 30)$. In terms of the real-world context, this decision analysis framework can be placed in a public participation context. For example, public hearing could be held at each location and corresponding charts for travel cost and noise similar to Figure 8 be displayed in order to determine step size. Decision analysis provides the tools for determining the location of the airport as a function of noise vis-à-vis

![Figure 7](image_url)  
**Figure 7** Step size determination for noise function.

![Figure 8](image_url)  
**Figure 8** Graphical display to determine step size.

<table>
<thead>
<tr>
<th>Iteration $k$</th>
<th>Airport location $x^k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$(30, 30)$</td>
</tr>
<tr>
<td>1</td>
<td>$(15, 45)$</td>
</tr>
<tr>
<td>2</td>
<td>$(7.50, 52.50)$</td>
</tr>
<tr>
<td>3</td>
<td>$(3.750, 56.250)$</td>
</tr>
<tr>
<td>4</td>
<td>$(1.875, 58.125)$</td>
</tr>
<tr>
<td>5</td>
<td>$(0.938, 59.063)$</td>
</tr>
<tr>
<td>6</td>
<td>$(30.469, 29.5319)$ etc.</td>
</tr>
</tbody>
</table>

Source: Chan 2011, 232.
Extending the basic model to two dimensions

Beyond network-center location problems like that examined above, decision-analytic techniques are equally applicable to median and center location on a plane. As above, the locational decision involves spatial organization articulated through the cost functions as shown in Figure 1. In the following planar example, we identify the cases in which $\beta = 1$, $> 1$, and $< 1$. Through the lens of decision analysis, one can view median locational decision as a system optimization model, whereby the “common good” is being upheld. Let us build upon the above airport case study by considering a regional airport for three cities (instead of two): Cincinnati, Columbus, and Dayton, Ohio. Suppose the population in Cincinnati is two million, the population in Columbus is three million, and the population in Dayton is one million. Note that the population in Columbus is the same as the population of Cincinnati and Dayton combined. This is done on purpose to show an ambiguous case for the Hakami extremal condition—a planar median-location decision based on system optimization. It is challenging because the solution is sensitive to model formulation and computational procedures.

Now, refer to Figure 9, which illustrates various locational solutions. For a linear-cost accessibility criterion, the airport will be located at the largest city—Columbus in this case—as shown by the label LB, with LD serving as the alternate solution. Here, the notation L stands for an LP solution, B stands for base solution, and D stands for alternate solution. Please note that numerical approximation may result in imprecise locations, shown here as “wedges.” Using planar gradient search, HB denotes the base solution, and HD is the alternate solution, where H stands for employing a planar gradient search procedure. Notice that both LP and gradient search yield the same base solution at Columbus. However, the alternate solutions LD and HD are in the middle of the triangle (instead of at a corner) and the two locations—LD and HD—are disparate. When noise concerns are the focus, spatial cost becomes nonlinear with $\beta < 1/2$, and the airport is now located at HB_n, where B_n stands for the base solution for noise abatement. Obtained by gradient search, the noise-abatement HB_n solution is somewhere between the accessibility-based LD/HD solutions and the LB/HB locations (at Columbus).

Viewed in decision-analytic terms, the center location problem is a user optimization problem, where the most-distant user is being taken care of. It can be shown that the airport will be located among the three corners of the triangle representing the three cities, even for a linear cost function. Of particular interest is that an MCO airport-location solution will juxtapose the single-criterion results shown above, essentially being sited by an interpolation of the locations shown in Figure 9. This result is similar to the two-city case study. It suggests that real-world
decisions, which invariably include multiple considerations, tax classic facility-location approaches.

Returning to the two-city study, it can be seen that an effective way of performing multicriteria nonlinear programming is the interactive Frank–Wolfe approach. A key feature of the F–W algorithm is to convert a nonlinear program into a set of LPs. We extended this concept by assuming the existence of an underlying preference function, as illustrated in Figure 4, but never actually requiring this preference function to be identified explicitly. All that is necessary to do so is to construct a tangent approximation of the value function at each candidate airport location. The tangent approximation helps to weigh accessibility against noise through the tangent slope $-\lambda_1/\lambda_2$. Meanwhile, the simultaneous graphic display of the travel and noise functions in Figure 8 helps to determine the next potential site for the airport through the step size $\alpha$. The basic idea is that even if the DMs cannot specify an overall preference function, they can provide local information regarding a preference at a particular situation. In so doing, we narrow the number of nondominated solutions to one at each iteration. The iterative approach moves from an initial feasible solution toward the optimal solution by finding the direction of steepest “ascent” and the optimal step size toward that direction. Again, explicit knowledge of the overall preference function is not essential. Only local information concerning the revealed preference of the decision-makers is required and this, in turn, is sufficient to determine the relocation direction and step size.

It is apparent from the above case study that interactive programming of this sort is highly numerical in nature and there are issues not addressed in this entry. Aside from convergence issues, for example, different attribute functional forms may give rise to drastically different solutions for a preset set of weights and step size. An example is to re-write the value function in terms of $v(f_1(\mathbf{x}), f_2(\mathbf{x})) = v(-2x_1-x_2, 2x_1^2+x_2^2)$, with the criterion functions to be maximized (instead of minimized). In practice, however, such an extreme result is unlikely to occur, since the weights and step size is anything but constant. The above cited computational results are built upon the assumption of $\alpha = 0.5$ and $\lambda_1 = \lambda_2 = 1$ throughout the iterations. This simplifying assumption is made mainly for our computational convenience. In fact, the entire foundation of interactive procedures of this sort is to explore the DMs’ revealed preferences, guided by charts and tables, in various situations. Correspondingly, their reactions, as reflected by values of $\alpha$ and $\lambda$s, are expected to be different at each iteration. Convergence in this case is obtained not so much from numerical properties as it is from the decision-maker’s behavioral changes through the process of negotiation and deliberation. These shifts from one iteration to another can compensate for the effects of different numerical procedures, leading toward consistent location decisions (Chan 2011). More detailed discussions of interactive multi-objective programming are found in Seo and Sakawa (1988) and Haimes and Steuer (2002).

**Future research directions**

What are the challenges of multicriteria decision analysis in the foreseeable future? One answer lies in the view that MCDM can be seen as comprising four different subareas. First, multicriteria mathematical programming will continue to focus on advancing our ability to formulate and solve decision problems mathematically. Second, research on multicriteria discrete alternatives (including integer MCO) will focus on problems where tradeoffs do not
DECISION ANALYSIS

vary smoothly, but instead have set, discrete values. Third, multi-attribute utility theory will examine the larger issues of how decisions are made. Finally, negotiation theory deals with how individuals can or should work in concert to come to decisions. In all four areas, extending the state-of-the-art lies not only in advancing a particular set of issues, but also addressing larger questions. For example, do stakeholders have multiple, conflicting objectives that must be reconciled? Is there significant uncertainty in the decision outcomes? Is one making a single, one-time decision or a sequence of decisions over time? Answering these questions will involve working across research areas.

If the problem calls for involving multiple stakeholders and multiple competing attributes, for example, then tools that emphasize group support and value elicitation are worth exploring. Group decision-making remains an interesting problem that should be further investigated, and has traditionally been addressed using incentives, game theory, facilitation, and other methods (Abbas 2012).

Fortunately for researchers and policymakers, there are decision analysis and multicriteria decision-making software packages available. Most of the examples given below cover multi-objective decision analysis and MAUT. Ideal software should be intuitive to the user, explainable to a variety of stakeholders, and readily support iterations among the various stages of the decision analytic process.

- 1000Minds – 1000Minds Ltd.
- Crystal Ball standard, professional & premium editions – Oracle.
- DPL – Syncopation Software.
- ForeTell /Æ – DecisionPath, Inc.
- Hiview3 – Catalyze Ltd.
- RPM-Decisions – Systems Analysis Laboratory, Helsinki University of Technology.
- SMILE (structural modeling, inference, and learning engine) – University of Pittsburgh.

Recent decision-analysis techniques tend to have a number of features in common and future research will very likely continue to enhance and expand on these features. First, they should not cramp the style and involvement of decision-makers. Second, they have interactive feedback mechanisms including graphics. A visual display, possibly through the use of computer graphics, can greatly assist the decision-makers in performing MCDM analysis. (We have seen such an example in Figure 8.) Third, they have a built-in evolutionary process, allowing modification of the model as the analysis progresses; improvements should include problem structuring and elicitation features. Finally, these software packages can provide decision-makers with fast-turnaround analysis techniques.

In some cases the analytic team will have large quantities of data that can be used to formulate and support the decision-making process (Chan 2011). Some of these decision-making software packages (as well as some statistical packages) have learning algorithms that will build the joint probability distribution from the available data. If this option is available, the analysis team can supplement automated efforts by involving subject matter experts in the review of the resulting model. These experts can help find errors in the data and, just as importantly, they can supplement the model with their knowledge. Combining what is learned from data and what is learned from experts usually yields a better model and results in a higher likelihood that the effort devoted to decision analysis will be successful. The primary advantage of involving experts is to increase the information learned as a result of asking preference questions of decision-makers.
Finally, there are broader issues of power and knowledge at play. In a competitive environment, such as the relationship between the airport authorities and the public in the above example, recent research suggests that those with better knowledge about other stakeholders tend to do better than those who do not. Under this situation, it is important to assess the role of epistemic knowledge, which is defined as knowledge supporting a belief, truth or hypothesis. Epistemic utility is a measure of the value of this knowledge in supporting a hypothesis, that is, a measure of the importance of the information about the other stakeholders relative to the belief that the other stakeholders are going to adopt a certain strategy. Epistemic utility theory has so far furnished us with a number of arguments for some of the central norms governing partial beliefs (Pettigrew 2010). For example, is it legitimate to employ the notion of expected utility when the belief function with which the expected utility is calculated is not a probability function? Stirling (2002) proposed a theory of multi-agent coordinated decision-making using epistemic utility, which represents an alternate theory of multi-agent decision-making, termed satisficing games, which is designed to address the limitations that are imposed by the assumption that players are motivated exclusively by self-interest. However, it does not abandon the principles of value and performance that are vital to rational behavior, which includes both common good and self-interest. Aside from our observations reported here, there will no doubt be other evolving schools of thought that differ from both the classic-utility- and epistemic-utility-theoretic schools.

SEE ALSO: Location-allocation analysis; Location-allocation models; Multicriteria decision-making; Spatial decision-support system

References

Deep mapping and sensual immersive geographies

Trevor M. Harris
West Virginia University, USA

It is apposite that the term “geographic information system” (GIS) was coined at an early stage in the development of geographic information science (GIScience) not least because of the fundamental spatial concepts and spatial reasoning that underpin the technology. But while GIS and other labels such as spatial information system are often viewed as synonymous terms, in reality an important distinction might be drawn between them. Both represent digital spatial information systems, but the leading adjective represents an important nuanced differentiation in that the geographic serves a broader audience within the field of geography whereas emphasizing the spatial alludes to a more specific focus on spatial objects, spatial metrics, and spatial analysis. Within the field of geography this is an important distinction, and while GIS has seen dramatic diffusion and uptake within academia and almost every sector of the economy, it has nonetheless experienced resistance within parts of the discipline of geography. This is less a reluctance to embrace the technology but more a resistance to the way in which GIS represents nature and society, which appears at odds with many conceptual and theoretical advances made in human geography since the quantitative turn.

Many have questioned the epistemological and ontological construction of GIS and the perceived inability of a positivist GIS to address the needs of humanists and human geography. The important broadening of GIS from a focus on system toward one on science has not diminished concerns about the positivist and reductionist tendencies of GIS, embedded as it is within the scientific method. Critical GIS, critical cartography, and the GIS and society debates of the 1990s captured the early social theoretic critique of GIS and the concerns of human geography in particular, which saw GIS as a return to the positivist traditions of the quantitative revolution and which questioned the ways in which GIS represented nature and society. GIScience is clearly not a panacea for geography, but the primary focus of GIS on the spatial has contributed to a lack of engagement with other important aspects of geography such as place. While it is quite appropriate to contend that certain aspects of geography lie outside the bailiwick of GIScience, it is argued here that this space–place divide prematurely concedes considerable ground. Recent advances in both the technology and innovative approaches indicate that GIScience has a role to play in addressing aspects of geography previously considered beyond the purview of a spatially focused technology.

It is in the engagement of GIScience and the humanities that many of these issues have come to the forefront and also where possible advances have been made toward bridging this divide. There are good reasons why the sciences and social sciences were early adopters of GIS and have seen demonstrable success, yet the humanities have lagged these early adopters except where GIS lends itself to the mapping project or the analysis of spatially structured data such as within historical GIS. It is in the conceptual and methodological engagements between GIScience and the humanities, and in what
has been termed “the spatial humanities,” that scholars have sought to bend GIScience to the needs of humanists. In so doing, advances in the conceptualization and application of geospatial and geovisualization technologies indicate ways in which GIS might escape its reductionist spatial corset and embrace representations of human behavior, events, and place that enable GIS to contribute to a more human-centered geography. This entry examines ways in which deep mapping and immersive geographies point to alternative GIS representations of space and place.

Spatial representations – placial representations

GIS provides a powerful digital technology with sophisticated spatial data handling, spatial analysis, and mapping capabilities that brings with it a particular representation of the world. GIS necessarily represents an abstraction of the real world, for capturing reality in its entirety would involve an improbable 1:1 relationship. GIS captures, stores, and represents the world as spatial primitives, objects and fields, with the topological connectivity that relates them. Objects must have a geographical location invariably recorded according to some measured Euclidian coordinate geometry. GIS is driven by concerns to be error free and to achieve high accuracy and precision in the location and attribution of objects. Considerable effort is thus expended on issues concerning spatial data acquisition, data categorization, scale, distance and direction, datum and projections, and geometric topology. GIS seeks to be reproducible and verifiable, hallmarks of the scientific method. Sophisticated GIS mapping capabilities draw upon long-established cartographic traditions to effectively map spatial data that are difficult to describe through narrative, which is ill suited to represent the higher-order dimensionality of complex spatial data. System functionality enables sophisticated spatial analysis and spatial statistics to be performed on the data. Space thus provides a powerful integrating theme in GIS that enables multiple scaled explorations of spatial relationships by virtue of their common location and in ways that many alternative approaches cannot. Yet, despite the digital construction of an “authentic” representation of reality, GIS and maps possess an objective, authoritative, and reductionist allure and there is a tendency to separate and disembodied spatial analytical geographies from the mixed methods approaches of social, cultural, and political geographies (Sui and DeLyser 2011).

The social-theoretic critique of GIS through the GIS and society debates of the 1990s, critical GIS, and critical cartography, provides a valuable context for exploring many of the issues associated with the epistemology, ontology, assumptions, and biases of GIS in representing nature and society (Pickles 1995). As a result of these largely academic endeavors, particular types of data, representation, logic systems, and ways of knowing the world were seen to be privileged over others. There can be no question that the ability to deconstruct and reconstruct spatial entities and regions within GIS has brought about a profound technological transformation. However, the heavy reliance on available spatial data at an appropriate spatial and temporal scale and attribution, the commodification of data, the Boolean logic system of GIS, spatial determinism and the silences in the data, structural knowledge distortion, and the emphasis on mapping have all contributed to a particular representation and way of knowing the world.

Ultimately, GIS privileges reductionism and disambiguation in its organization and analysis of knowledge, and the map has assumed a seemingly unassailable sense of spatiality, surety, and finality. In many respects this disambiguation may have
contributed to the marked success of GIS as a critical decision-making support tool in many areas of society. Critics contend, however, that GIS privileges master narratives and categorical causation at the cost of contingent causation and alternative interpretations and renderings of space and particularly place. This reductionist metanarrative stands in contrast to many parts of human geography and the humanities that are heavily versed in multivocality, intertextuality, and contingency, and where place is equally important as space in the construction of human geographies. In this regard GIS is considerably less empathetic to scholarly work where knowledge is nuanced, individualized, spatially imprecise, uncertain, ambiguous, and part of a world of multiple realities. Handling ambiguity in GIS represents a challenge, though, this entry contends, not to a GIScience engaged in a meaningful reciprocity between geospatial technologies and humanist traditions.

But this raises the question about the role of place as inhabited space. Space and place, as intertwined phenomena, are central to the works of geographers (Cresswell 2004) and historians (Ethington 2007). For Tuan (1974) the individual’s relationship with the livable world, their being-in-the-world, and the connectedness between the physical fabric of landscape and human emotion lies at the heart of understanding place and the larger networks of places that is space. Tuan argues that human experience, while highly individualistic, is based on biological and cognitive traits that are common to everyone and that these innate traits allow our experiences of space and place to be shared. However, any experience of place carries with it the observer’s own reflexivity and worldview, encapsulated within social and cultural meanings that have been layered onto the landscape. While the work of Tuan, Tilley (1994), and others is compelling in its focus on the lived experience of space, place, and landscape, these ideas are difficult at best to translate into the medium of GIS and geospatial technologies (Goodchild 2011). The success of GIS as a tool for synthesizing, displaying, and analyzing geographic information is clear, but the limitations imposed by current data models and modes of representation make it difficult to develop a digital framework for humanistic and experiential aspects of place. Blending a phenomenological and experiential approach to interpreting places that are relative, mediated, and socially produced represents a real challenge to GIScience because place encompasses not only the physical but also the symbolic, the emotional, and the many meanings associated with place. As Entrikin (1991, 10) points out, “To understand place in a manner that captures its sense of totality and contextuality is to occupy a position that is between the objective role of scientific theorizing and the subjective role of empathetic understanding.” The quantitative nature of GIS favors representations of space that are absolute, whereas addressing issues of place requires a sensuous and reflexive GIS that is nuanced and sensitive to the traditions and epistemology of humanists.

Humanists place considerably greater emphasis on place than on space. The question then arises as to how it might be possible to link and explore the processes that make up place and place-making, such as locale, memory, experience, myth, ideology, symbology, and events, the grain and patina of place, to GIS. The spatial turn in the humanities confronted the perceived undue emphasis on historicism and reinserted space as an active participant in human events and behavior (Soja 1989). Ironically, the reinsertion of space in the humanities has been the result not only of the spatial turn but also of the (re)discovery of the power of the map through the application of GIS and historical GIS: social theory and positivist science together. However, the spatial analytical power of GIS does not
easily accommodate the ways in which humanists explore or interpret phenomena and place. Humanists struggle to fit their questions, data, and methods, the emphasis on narrative and storytelling, and a resistance to pre-emptive resolution, into the rigid parameters of the technology. In these instances, where knowledge is multi-valent, equivocal, contradictory, and conflictual, and where social life is contingent, implicated, or unpredictable, interfacing with the spatial precision and reductionism of GIS is problematic.

Deep mapping

Deep mapping arose out of the imaginary of the spatial humanities and represents a critical engagement between geography, GIScience, and humanist traditions (Bodenhamer, Corrigan, and Harris 2010). Deep mapping is aspirational and methodologically largely untried, but the concept pursues a GIS-humanist theme and engages a GIScience-enabled representation of place. It is too simplistic to speak in terms of a qualitative-oriented humanismandaquantitative GIS, yet the contrasting emphases on conceptual mapping rather than cartographic mapping do represent a significant methodological divergence. The intellectual origin of deep maps lies in a combination of eighteenth-century antiquarian approaches to geography, history, people, culture, and place; in the detailed local histories of historical geographers such as Hoskins (1955); in the insightful observations into deep mapping and performance drama of Pearson and Shanks (2001); in de Certeau’s (1984) spatial stories and practices of everyday life; in the early work of the Situationist International (DeBord 1955); in psychogeography (Stedman 2002); in Tuan’s Topophilia (1974); and in the thick description of William Least Heat-Moon’s PrairieErth (1999). Deep mapping challenges GIS to move beyond a view of humans as entities and data points and to include not only the material but the imaginary embodied world, together with the behavior and relationships that make up a nuanced, nonreductionist, and scaled conception of place and a habituated world. Deep mapping engages the material and the contingent world and interrelates theories of practice and agency and how people both create their material world and, in turn, are created by it. This linking of critical geographies, postmodernism, and GIScience provides a conceptualization of a recursive humanist GIS that is negotiated, nonauthoritative, and nonobjective, and which is framed not as a statement but as a conversation. Deep mapping is thus inherently unstable and continually changing in response to new data, new perspectives, and new insights (Bodenhamer, Corrigan, and Harris 2015).

Maps remain a crucial instrument in deep mapping because space contextualizes place. Traditional GIS maps, however, emphasize precision and accuracy and the mapping of material physical features, infrastructure, population, and other features traditionally found on topographic maps. The term “thin” map, rather than “shallow” map, the antonym to deep map, is used to describe these map products to avoid implications of shallowness, superficiality, or inconsequentiality for which there is abundant evidence to suggest otherwise (Harris 2015). These “thin” maps have formed the backbone of GIS and myriad applications.

Deep maps, in contrast, interlace autobiographical, art, narrative, folklore, stories, and memory with the physical form of a place. As Pearson and Shanks (2001, 64–65) suggest, deep maps “record and represent the grain and patina of a location … [through the] juxtapositions and interpenetrations of the historical and the contemporary, the political and the poetic, the factual and the fictional, the discursive and the sensual; the conflation of oral testimony,
DEEP MAPPING AND SENSUAL IMMERSIVE GEOGRAPHIES

anthology, memoir, biography, natural history and everything you might ever want to say about a place.” Deep maps speak to the contingent nature of cultural processes and the agents of change and transformation and for dismembering generalizations through rethinking and recombining (Bodenhamer, Corrigan, and Harris 2015). Deep maps are the stories, conversations, and lives lived out in a place, which are inseparable from the cultural context in which they exist. Deep maps are place-based and capable of integrating multiple voices and views based on a GIScience that is sensitive to the needs of humanist scholars. Deep maps resist reductionism and universal truths and grand narratives in favor of discursive collaboratories abounding with multiple voices, contingency, and contested meanings.

Moving from deep map concept to method is not without its challenges. There have been several attempts in GIS to explore place and latent place-rich geographies through GIS (Cooper and Gregory 2011). Perhaps the early focus on GIS has diverted attention away from other promising approaches under the GIScience umbrella even though GIS would appear to be an obvious analog for deep mapping through its layering and emphasis on space as a bounding framework. Under the umbrella of GIScience, a cornucopia of tools, techniques, and methods are now available, including geovisualization (Dykes, MacEachren, and Kraak 2005), geotagged social media and technologies, crowdsourced volunteered geographic information (Sui, Elwood, and Goodchild 2012), and Web 2.0 geospatial web mapping, mashups, and tools including neogeography (Turner 2006), capable of generating, storing, accessing, and mapping spatial data in multiple ways. Deep maps represent hybrid methodologies that fuse positivist and quantified geospatial technologies with qualitative data and nuanced analysis: a distinct move toward qualitative GIS (Cope and Elwood 2010). Deep maps thus draw heavily on space and place as an organizing framework to understand the world and yet challenges GIS as to how this might be achieved. A number of approaches have been proposed to achieve deep maps, including neogeography (Warf 2015); ghost maps of highly detailed and layered text, photographs, and maps (Ethington and Toyosawa 2015); geographical text analysis combining GIS with corpus linguistics (Gregory et al. 2015); ethnographic biographical storytelling, ethnopoetry, and tender mappings that weave the personal, memory, emotions, behaviors, and attachment to place into everyday geographies (Aitken 2015); infinite depth systems for storyboarding (Corrigan 2015); narrative generation platforms (Yuan, McIntosh, and DeLozier 2015); and coordinate frameworks for narrative recording (Martin 2015), each of which have sought to identify potential avenues for deep map development.

Embodied sensual GIS and immersive geographies

To illustrate one possible approach to deep mapping and deepening GIS into the experiential, emotional, and local associations that constitute place, an example exploring the literary work of a nineteenth-century writer, Rebecca Harding Davis, is used. Text analysis represents a continuing issue for GIS and yet it is the mainstay of the humanist tradition. In examining the writings of Davis (1861) the GIS back end is augmented by drawing on a fusion of three-dimensional (3-D) modeling and serious game engines displayed within the immersive virtual environment of a cave automatic virtual environment (CAVE). The intent here is to move beyond the reductionist two-dimensional map of the material fabric of a landscape toward an
Deep mapping and sensual immersive geographies

Experiential immersive environment that enables the emotional, the sensual, and the immaterial to be included to create a nuanced dimensional richness of place. Linking phenomenology and the experiential in such a manner in GIS is problematic for reasons detailed above, yet this study suggests ways in which it is possible to drill deeper into the geography of space and place.

Rebecca Harding (née Davis) (hereafter Davis) (1831–1910) was a leading American literary figure of the nineteenth century who lived in Wheeling, then Virginia and now West Virginia, and who wrote evocatively about the industrializing nature of the antebellum town. Davis witnessed dramatic changes in Wheeling as the town became an industrial center. While the distinction between factual and fictional sources, and between realism and the imaginary, is often blurred, Davis’s writings for the local newspaper and for fiction paint a fascinating picture of Wheeling as a place that augments the massive data repositories of newspapers, travelogs, and other archival papers available for Wheeling. The degree of fidelity to geographical reality in Davis’s writings is difficult to assess, though there is good reason to pay heed to her insights as one of the nation’s first realist writers and because of her personal knowledge and experience of Wheeling. Brosseau suggests that readers often assign the subjective element within fiction to the physical elements of a place while assuming that the interpretation of the relationships between places and people is reliable as novelists are seen as “some kind of spokesperson … They immerse us in the various attitudes, values and conflicts shared by the people of a particular region in relation to their environment” (1994, 336). Darby (1948), for example, suggests that Thomas Hardy manipulated some topographical details in his novels to facilitate his storyline and yet remained true to the underlying social relations. This is an important point that seeking insight into the human–environment relations rather than spatial description is perhaps most revealing and rewarding, and especially so in historical contexts where corroboration of these social–cultural–economic–environment relationships is so difficult to establish.

Davis’s seminal work, Life in the Iron Mills (1861), is recognized as among the most important works in nineteenth-century American Literature (Baym 2011). Davis writes in powerful evocative prose of mid-nineteenth-century Wheeling: “A cloudy day: do you know what that is in a town of iron works? … The air is thick, clammy with the breath of crowded human beings. It stifles me. I open the window and looking out, can scarcely see through the rain the grocer’s shop opposite … I can detect … the foul smells ranging loose in the air . . . . I want you to hide your disgust, take no heed to your clean clothes, and come right down with me, here, into the thickest of the fog and mud and foul effluvia. I want you to hear this story . . . . I want to make it a real thing for you” (Davis 1861, 1–2).

Yet, while her writings identify township locations, the world of emotions and place-based qualities evoked and experienced by Davis exceeds the representational dimensionality of a two-dimensional map. Davis’s world is multidimensional and sensual for her writings are awash with references to smell, texture, light, touch, and sound—all elements of a sense of place. Capturing the textual, experiential, and nuanced world of Davis’s Wheeling, to make it a “real thing” for the reader, requires moving beyond the strict spatial framework of GIS and its focus on the material townscape.

Within the past few years, a number of advances in computer science, computer hardware and software, computer graphics, virtual reality, 3-D GIS modeling, serious gaming engines, and geovisualization have made possible impressive rendered immersive representations.
that more closely emulate the world around us (Fisher and Unwin 2003). Immersion, augmented by geosensory inputs, more closely creates a sense of place and of “being there” (Harris and Baker 2006/2007). Immersion is a powerful psychophysical experience that draws a user into a seemingly real but virtually rendered environment and transforms the user from passive observer to active participant (Harris and Hodza 2011). The visual-cognitive model draws on the creative power of the mind to move seamlessly between the physical, virtual, and imaginary world and to be disengaged from the physical reality of the projection space. The effect is a powerful one in transporting users into a multidimensional and multisensory environment embedded within GIS in which users interact with the displayed scene in a highly dynamic way through multiple viewpoints.

The CAVE used in this study is a specially designed structure comprising three walls and a floor made up of 10-foot screens (Figure 1). Stereo images are rear projected and bounced via a mirror from above onto the floor display. The projected images are computationally made into one seamless image projected across the CAVE walls (Cruz-Neira et al. 1992). Users standing in the CAVE wearing 3-D glasses experience a feeling of immersion within a lifelike sensory environment. The sense of immersion is attained through a combination of stereo depth perception, peripheral vision capture, head tracking devices, and interactive handheld controllers. The effect is to make images appear holographic.

**Figure 1** Viewing the smog and iron mills of Davis’s Wheeling in the CAVE. Attribution: Christabel Devadoss, Fang Fang, Aaron Ferrari, Trevor Harris, Deborah Kirk. Photo taken by the author.
DEEP MAPPING AND SENSUAL IMMERSIVE GEOGRAPHIES

and for the user to maneuver through and around the imagery as if it were a real environment. Such an experience is difficult to describe, but the user has a profound sense of being in and experiencing the virtual world.

To situate Davis’s neighborhood, the urban fabric of Wheeling was reconstructed and rendered using 3-D modeling software such as Trimble SketchUp (www.sketchup.com) and ESRI’s rule-based City Engine (http://www.esri.com/software/cityengine). The 1884 Sanborn fire insurance maps detailed building footprints and dimensions, and the material from which buildings were constructed, as well as the number of floors and their height. Historical photographs and extant buildings were used to generate architectural styles and building appearance typical of the era. The intent here is not to be photorealistic but to capture a scene representative of the geography of Davis’s neighborhood at that time. Terrain and surface imagery was ingested and the entire scene displayed through ArcScene and a game development platform, Unity3D (www.Unity3d.com), which provides a visually stunning 3-D graphical rendering. Unity has superior graphics, interaction and navigation capabilities, animation, particle physics to emulate flowing water and smoke, and the capacity to trigger sound events.

Not only does the virtual environment provide a visually stunning immersive insight into the Wheeling described by Davis, but additional senses are also engaged to reinforce the sense of place (Figures 2 and 3). Davis describes streets adjacent to her neighborhood where she felt concern for her personal safety. A GIS map of her geography of fear was created and the 7.1 sound system of the CAVE was used to capture her emotional response to these contested places. As a user navigates through the virtual Wheeling scene the sound of a steady beating heart is heard until the GIS detects intrusion into the areas of

Figure 2  The Wheeling skyline with the persistent smog and smoke. Attribution: Christabel Devadoss, Fang Fang, Aaron Ferrari, Trevor Harris, Deborah Kirk.
concern identified by Davis’s map of fear. The sound of the beating heart increases in pace as the user moves deeper into these spaces and diminishes as the user recedes to the margins. In this way an atavistic sound is used to emulate Davis’s fight or flight reality and to replicate her innate emotional response to place. Additional soundscapes are triggered as a user navigates the scene, with a mix of ethnic voices in conversation, everyday sounds of street life, commercial activity, animals, and machinery. Even the presence of drunken groups as described by Davis can be captured with off-scene catcalls and ribald comments.

Additional geosensory inputs augment the visual and soundscape. The smells of Wheeling permeate Davis’s writings. In one instance she describes the everyday street smell of rotting detritus and of a street where a tannery and a brewery were in close proximity and comments on the resulting mix of smells that pervaded the area. Smell has a powerful effect in triggering and evoking memory and place, and smell was added to the CAVE scene to reinforce the virtual scene and soundscapes. A ScentAir (www.ScentAir.com) smell diffuser was used to create smellscapes throughout the virtual city whereby street garbage, livestock, food, cut grass, and burning wood were variously inserted into the virtual scene at appropriate locations. Tactile senses were not enabled in this particular scene, but in another study re-creating the virtual Cherokee Trail of Tears, a walking device used in the gaming industry enables users in the immersive scene to walk the scene, which, while not coming close to the experience of the Cherokee, does at least extend the experience beyond interacting from a standing or sitting stance. The addition of geosensory inputs adds to the sense of place as identified.
by Davis and moves beyond an otherwise rather sterile map. The effect of these sensory inputs is pronounced when compared with just the visual immersive experience. The graphics of Pudding Lane (Dempsey et al. 2014), a re-creation of seventeenth-century London using the Crytek serious game engine, is visually dramatic but is accompanied by incidental music which detracts markedly from experiencing the place that is graphically depicted so incredibly well. The soundscapes and smellscapes of Davis’s virtual Wheeling create additional dimensions of depth to the experience of place.

The immersive, sensual, and emotional geographies drawn from Davis’s writings articulate lifeworlds and places and an evocative image of life in an industrializing town. It is suggested here that such a phenomenological and experiential framework and representation of place, combined with the spatial analytical power of GIS, will resonate with readers as a means to investigate notions of place not previously possible through a conventional mapping of space. In the absence of other narrative writings, Davis’s work and its immersive representations become a primary source and medium for exploring and understanding mid-nineteenth-century Wheeling as place. Immersive geographies reside in the classic “uncanny valley” where the perceived distinction between reality and representation becomes more noticeable the closer one gets to representing reality (Mori, MacDorman, and Kageki 2012). There is rightful concern about striving for hyperrealism and verisimilitude in virtual modeling where the reconstructions are seen to stand in an inferior position to the original referent (Gillings 2002). At some level the human mind fills in the spaces of an obviously nonreal representation and yet becomes uncomfortable and hypercritical the closer a building depiction or human portrayal emulates the real. It is for this reason that human avatars were used with caution in this study. Although the virtual scenes stand as a subjective rendering of a bygone place, this should not detract from their value in augmenting traditional cartographic products which are arguably equally subjective constructions (Crampton 2001).

Drawing upon soundscapes and smellscapes in an immersive representation of a place opens up additional opportunities to pursue a sensual and experiential GIS for exploring place and geography. The approaches outlined here enable the representations of place embodied in text to extend beyond the physical and locational referents of space and allow us to peer into the symbolic, the emotional, and the sensuous aspects of, in this case, literary places.

**Conclusion**

Tuan (1974, 3) suggests that “Space and place are basic components of the lived world,” yet any digital representation of space and place is mediated in its creation and interpretation by our own experiences and worldviews. Capturing the vividness and complexity of place within the vector and raster representations of GIS is challenging, not least because a defining characteristic of place is the meaning gained through experience and abstract qualitative elements. The early focus on GIS as a system may have diminished the potential impact of GIScience on humanist research, though an ever-growing range of advanced geospatial technologies and approaches that augment the power of GIS to handle, manage, and analyze spatial and placial representations are increasingly available and add considerable sophistication to the armory of GIS tools and techniques available. Deep maps provide a conceptual goal for broadening the nature of the mapping process and examining the nuanced and qualitative aspects of place as
well as space, and in the process raise unexpected meanings and questions not previously pursued.

Immersive geographies move beyond the two-dimensional plane to provide an experiential and shared representation of geography. The representation is more intuitive because it replicates the three-dimensionality of our own lived world. In his argument for nonrepresentational theory, Thrift seeks to counter what he sees as an undue emphasis on mapping in human geography and emphasizes the need to move beyond representation and the practices of “witnessing” that produce “knowledge without contemplation” (Thrift 2008). Nonrepresentational theory focuses upon practices and how human actions are enacted, rather than on representing social relationships and what is produced. In the process, phenomenology, poststructuralist theory, and hybrid geographies come together (Whatmore 2002). In this regard, a fusion of hybrid geographies forces us to explore the familiar with unfamiliar technologies and to extract more from the geography than spatial associations. A nuanced, reflexive, sensual GIS, with geosensory inputs such as smellscapes and soundscapes, moves GIS beyond the visible fabric of society and reinforces the sensual nature of place and the role of emotion in understanding the cultural, social, and behavioral sensations, ideas, and feelings related to place. A critical, self-aware approach to these technologies is important to separate the layers of meaning embedded in place, and to recombine them in meaningful ways. Such nuanced representations provide open-ended experiential frameworks which, in combination with the spatial analytical approach of GIScience, enable the deep mapping of humanized space and a thick weave of multiple voices and viewpoints to destabilize our understanding of both space and place.

**SEE ALSO:** Augmented reality; Critical geography; Critical GIScience; Nonrepresentational theory; Place; Virtual reality

**References**


Deindustrialization

Peter W. Daniels
University of Birmingham, UK

Deindustrialization represents a reversal of an earlier period of structural change in societies and economies that involved a shift from agrarian pursuits to industrial activities. In this sense it is the opposite of industrialization. It has been shown that all economies experience a structural shift that involves moving from a reliance on primary sector activities, to manufacturing sector activities and finally to tertiary (or service) sector activities (Gershuny and Miles 1983).

A key characteristic of structural change is that it is persistent and long-term, and results in changes in the relative importance of the different sectors that make up an economy. This is usually measured, for example, by examining changes in the sectoral share of employment, output, or exports. This process of structural change is a phenomenon that has received much attention in economic theory. It is commonly understood as “the different arrangements of productive activity in the economy and different distributions of productive factors among various sectors of the economy, various occupations, geographic regions, types of product, etc.” (Machlup 1991, 76). Such shifts in the structure of economies may be related to the level of economic development or to changes in the composition of a production system that are necessary to sustain economic growth.

In line with the principles of structural change, after a phase of industrialization economies undergo changes caused by a contraction of their industrial capacity, notably in labor-intensive traditional manufacturing industries such as textiles, shipbuilding, or iron and steel. The process has been widely observed in the advanced economies since the 1970s and is now widely referred to as deindustrialization (Blackaby 1979). Although it was primarily a twentieth-century phenomenon, a degree of deindustrialization occurred during the eighteenth and nineteenth centuries in countries like China and India where manufacturing declined following colonization by the British or the French. The shift from hand to machine production methods was at the heart of the Industrial Revolution that took place in Great Britain between 1780 and 1840. Major advances in technology, such as the development of the steam engine, combined with the rapid exploration of new scientific ideas that led to innovations in machines, products, and modes of transport, enabled the construction of canals, railways, roads, mills, and factories. The period also witnessed the transformation of life and work accompanied by the emergence of large towns and cities. This model of industrialization was adopted later in Western Europe and North America and, depending on levels of development, by the rest of the world throughout the nineteenth and twentieth centuries. It is continuing into the twenty-first century, with industrialization continuing apace in East Asia and Latin America. Since it was the first to
DEINDUSTRIALIZATION

industrialize, Great Britain was also, arguably, the first to experience deindustrialization, followed later by the economies of Germany, France, the United States, and others.

The causes of deindustrialization are numerous but there are a number of key determinants. Some of them are internal to each of the advanced (and more recently the emerging) economies, while others, such as trade flows, are external. One of the most important determinants is the way in which improvements in the productivity of manufacturing, compared to services, has reduced the demand for workers while the converse is true in service industries. At the same time, the process of industrialization has caused household incomes to rise and the resulting increases in disposable incomes (after things like housing, food, and energy have been paid for) has fueled a propensity to increase the demand for services more than for manufactured goods. The expansion of global trade linkages between the advanced and the emerging economies is another important influence on deindustrialization; it has both direct and indirect effects on employment patterns in economically advanced countries.

The widespread occurrence of deindustrialization across the advanced economies over the last 40 years, which are themselves different with respect to the structure of economic activities, means that there are also variations in interpretations of the phenomenon. One of the most widely utilized interpretations is that deindustrialization occurred when the share of total employment or output in manufacturing is lower than the equivalent values for services. Another interpretation is that deindustrialization is evident when there is a decline in employment or output of manufacturing industries. Some care is needed when evaluating both interpretations since short-term trends may reflect short-term cyclical economic effects that do not ultimately show up as long-term trends. A third interpretation of deindustrialization arises from the importance to national economies of external trade in manufactured goods; where such exports constitute a declining share of total exports, it becomes more difficult to achieve a positive trade balance (a surplus of exports over imports) because it becomes more difficult to generate sufficient revenues to pay for the imports required for the further production of goods. This does rely on the assumption that revenues from service exports cannot make up for the loss of revenue from goods exports; in the case of the United States, the United Kingdom, Hong Kong, and Singapore, for example, this is certainly no longer the case. Finally, deindustrialization is seen as arising from the new international division of labor whereby the competitiveness of goods exports from countries like China has encouraged manufacturing to shift away from the advanced economies, leaving service industries to fill the gap.

At the start of the twenty-first century the emerging economies are undergoing deindustrialization, with a structural shift toward tertiary industries. However, there is a difference in that advances in communications and in information technology have enabled developments in the structure and organization of production activities that have blurred the previously clear distinction between manufactured goods and services. Many service activities support manufacturing or are based on material inputs and technology-intensive goods produced by the manufacturing sector. Therefore the links between manufacturing and services have become stronger and more complex, epitomized by the rise and role of producer services. In this way deindustrialization of the emerging economies will not necessarily be measured with respect to reductions in the share of employment in manufacturing or in the share
of goods exported. It may be more useful to analyze contemporary structural changes in the emerging economies as a process of quaternarization, whereby the rise of sophisticated producer services are used as inputs both before and after the production of goods (Peneder, Kaniovski, and Dachs 2003). Deindustrialization is now a less clear-cut process than it was in the advanced economies during the second half of the twentieth century.

Deindustrialization is not confined to the national scale; it has a geographical imprint that varies across time and space in regions, cities, and at the local scale (Martin and Rowthorn 1986). At a country level, an example is the United Kingdom, where by 1979 employment in agriculture and in manufacturing accounted for fewer jobs than in 1964. This trend continued through to the present. By 2010 manufacturing employed only one-quarter of the number for 1964 even though total national employment had grown from some 22.5 million to more than 30 million. In the case of agriculture the change was even more dramatic; employment in 2010 was less than 15% of the level in 1964. The losses of manufacturing and agricultural jobs were balanced by increases in services, especially financial, professional, and scientific services, as advances in technology improved employee productivity and rising average incomes increased the demand for consumer services such as tourism or retailing. Thus, service sector employment expanded by 11.9 million as manufacturing employment contracted by some 4 million, leaving a net increase of some 7 million in total employment. Deindustrialization along these lines, even if the precise details vary from one country to another, has taken place in most of the advanced market economies as well as in many emerging economies. Service industries now account for both the largest share of employment and gross domestic product (the total value of economic output) in many countries.

Whether deindustrialization is seen as a positive or as a negative consequence of structural economic change will depend on the scale of analysis. For some UK regions such as the northeast and Wales, as well as cities such as Manchester and Newcastle, deindustrialization has not involved the full replacement of shrinking manufacturing and primary industries with service employment. Many of the services are in the public sector and are not wealth-creating, market services in the way that is typical of the southeast region, particularly London. Here, deindustrialization is visible at every turn, with services providing more than 80% of employment, the lion’s share of total economic output, and has a physical impact in terms of numerous existing and new office complexes that eclipse equivalent developments in all the other cities and regions of the United Kingdom put together. This, in turn, means that the southeast region and London continue to attract a disproportionate share of new national, and particularly international, investment in services and related activities at the expense of other places. This reinforces the view that deindustrialization may not be avoidable but it occurs at the expense of increasing economic and social disparities between regions and cities within a country.

SEE ALSO: Informal sector; Local development; Restructuring; Uneven regional development

References

DEINDUSTRIALIZATION


Demand and supply for producer services

William B. Beyers
University of Washington, USA

Markets are at the heart of systems of demand and supply in capitalist economies. Demand comes from household consumers, governments, investors, and on “intermediate” account – as an input into the production and consumption of goods and services. The supply of goods and services is offered by producers across all categories of economic activity in markets. This entry first describes the general principles of supply and demand before turning to issues of supply and demand for producer services.

The neoclassical approach to supply and demand is conceptualized as a market environment in which buyers and sellers are presumed to have perfect knowledge. The quantity demand for a particular good or service is generally assumed to be an inverse function of the price offered, as depicted in Figure 1; a high price is related to relatively small quantity demand, while a low price is related to relatively large quantity demand. From the supplier’s perspective, the quantity offered in the market is presumed to be proportional to the price offered. The intersection of the demand and supply functions \(D_1\) and \(S_1\) in Figure 1 describes the equilibrium situation, in which market price and quantity supplied are defined as \(P_1\) and \(Q_1\).

Changes in demand and supply functions can alter the equilibrium market price as well as the quantity demanded. In Figure 1 a contraction in the level of demand, from \(D_1\) to \(D_2\), and a simultaneous increase in supply from \(S_1\) to \(S_2\) would result in a lower equilibrium price \((P_2)\) and also a lower level of quantity demand \((Q_2)\). The reduction in the demand function could occur due to changes in the tastes and preferences of consumers, or if the market involved is one related to production of a good or service it could be the result of technological change in the way that a particular input is utilized. Similarly, the change in supply functions could be related to market adjustments of various types, including expansion or contraction of the number of suppliers in a particular market, or changes in cost functions for particular goods or services that alter the willingness of suppliers to offer goods in particular markets.

The simple model depicted in Figure 1 is the basis for many variants. The slope of the demand and supply functions can vary, leading to situations of varying “elasticity.” If quantity demand is totally insensitive to price, then the demand function in Figure 1 is a vertical line, referred to as perfectly inelastic demand. However, if quantity demand is highly sensitive to price differences, then demand functions are considered to be highly elastic. If demands by an individual consumer are aggregated with other consumers, demand functions become more elastic. Similarly, if supply functions of individual suppliers are aggregated to total supply in a given market, supply functions become more elastic.

In the capitalist economy markets can be viewed as a system (Figure 2). Final demand is divided into consumption by households and demand for government goods and services. It also includes a link to the investment process, which is required by all productive components as well as by households (for purchases such as...
housing stock) and governments (for purchases such as infrastructure). Savings from households become a part of the source of funds for investment. Households sell their labor and receive income from work (loop 1). Factor markets exist for labor, also for resources and goods and services, as well as for capital. Inputs from factor markets power industrial production in all categories of economic activity, described in greater detail in the entry on Input–output analysis (loop 2). The output of the industrial system involves vertical linkages, defined by the technology of production in each industry. Goods and services from the system of industrial production are destined for product markets (loop 3), or become embodied as new investment. Product markets (for goods and services) deliver output to households (and government) for consumption, with payments for these goods and services flowing into product markets (loop 4). This circular system should be viewed as dynamic, with a changing mix of output over time due to changes in the tastes and preferences of consumers (and governments), technological change, and processes of innovation and invention that continually amend the spectrum of goods and services produced in economies. From the perspective of a region or a city, this system should be viewed as open, with imports and exports associated with all of the system components defined in Figure 2.

The structure of markets is also variable, as measured by the number of buyers and sellers. A schematic diagram (Figure 3) arrays these numbers from a single seller or buyer to markets with many buyers and sellers. In some cases, these markets could be highly competitive (as in a situation where there are many buyers and sellers); in other cases there could be essentially no competition, such as in a market with only one buyer and one seller. Figure 3 could also be conceptualized within a particular geographic context – ranging from a national or global market system – to a localized market environment. The explicit introduction of space into models of markets is at the heart of location theory.

Markets in capitalist economies may also fail to provide a sufficient supply of particular goods or services, as judged from a public policy perspective. This leads to government intervention to supply levels of these goods and services through taxation; examples include public education and national defense. There are significant variations among countries in the degree to which governments intervene in the supply of goods and services. For example, in the United States most health care is provided by the private sector, while in Canada most health care is provided by the public sector.

**Producer services**

The general discussion of demand and supply was based on the common neoclassical economic approach to markets, a framework that makes strong assumptions about the knowledge that producers and consumers have about elements of
the market environment. One broad component of modern economies for which the basic tenets of demand and supply differ from the neoclassical framework is producer services. These activities have expanded relatively rapidly in advanced and developing economies, as measured by employment, suggesting growing markets for the services that they provide.

Businesses may choose to produce the inputs that they require in-house or they may turn to the marketplace and buy these services from freestanding suppliers. The literature suggests different tendencies in this regard among countries. A tendency for German firms to produce more services in-house when compared with those in the United Kingdom and the United States was documented. Analyses of structural change in the United States through the lens of input–output models have found rising levels of purchase of producer services, suggesting that, over time, firms may be making more decisions to buy rather than make producer services. However, there is also evidence that the overall utilization of producer services has expanded for a variety of reasons, implying that there could be growth simultaneously in the
DEMAND AND SUPPLY FOR PRODUCER SERVICES

Figure 3 Alternative market structures.

internal production of producer services, as well as their external acquisition. Analyses of the occupational structure of nonproducer service firms find this growth takes the form of internalized work of a producer services nature. The neoclassical model described earlier implied that these decisions would be made purely on a cost basis. However, the literature on the demand for producer services finds strong noncost bases for expanding the acquisition of these services in the marketplace.

Scholars have distinguished between cost-based, quasi-cost-based, and noncost-based reasons for producer services acquisition. The transactions cost approach focuses on elements of internal costs (such as wages, administration, or overhead) versus the costs of external acquisition (including transactions costs related to external acquisition). It has been argued that external acquisition has led to supplier firms being able to take advantage of scale economies, and thereby offer their services at a lower cost than the cost of internal provision. Another thread in the literature argues that firms want to be “lean and mean,” to externalize functions not core to their markets, leading them to increase external purchases to lower internal production costs.

Quasi-cost considerations have also been argued to be a basis for the growth of the supply of producer services. Flexible specialization arguments revolve around the way in which demand for many types of producer services is unpredictable, making it necessary for firms to make choices as to whether or not to retain internal staff to provide these services. Strategies used by Danish firms were documented and shown to increase flexibility, allowing risk reduction while dealing with infrequent demand, and a desire to concentrate on core skills.

Noncost considerations have been shown to influence decisions on the external acquisition of producer services. These include the fact that some producer services are specialized and require very advanced levels of expertise for them to be supplied, and it has been argued that the development of in-house capabilities is simply not possible in these cases (Goe 1991). In addition to cost barriers to developing this type of expertise, the length of time that it would take to develop this expertise could also be a major barrier to internal provision of a producer service. Buyer/supplier dynamics have also been argued to be a major noncost consideration, such that rapid changes in the market and technological environments of clients of producer service firms lead to fluid relationships with suppliers of these services. Other factors seen as important are third-party information needs, such as independent audits, expert testimony, and government requirements for analysis, or the growing complexity of the management environment that encourages firms to seek outside experts to assist with their management needs.

Survey research has documented the relative importance of cost, quasi-cost, and noncost factors related to the demand and supply of producer services. A large-scale survey of producer service suppliers found that noncost factors dominated the reasons why their clients sought their services (Beyers and Lindahl 1996). Using a sample of 403 establishments in 11 producer
service industries, this survey used a Likert-scale to determine the importance of 16 different dimensions explaining why their clients sought their service. The importance of particular factors varied by industry but cost factors were not found to be an important determinant of producer service demand; quasi-cost factors were found to be moderately important. These findings suggest that the neoclassical framework does not offer a robust explanation for the growth in the demand for producer services.

Paralleling Beyers and Lindahl’s analysis of the perception of reasons why the clients of producer services sought their services were analyses of actions taken on the supply side by these establishments to position themselves in their marketplace. It was found that most producer service firms tried to position themselves in a “niche” market – to differentiate themselves in the eyes of the buyer from their competitors (Beyers and Lindahl 1996). This analysis was extended to incorporate aspects of Porter’s approach to the construction of a position of competitive advantage by firms (Porter 1990). His framework builds on two paradigms – market position defined through cost leadership strategies versus market position defined through differentiation from competitors. Lindahl and Beyers (1999) found clear evidence that producer service firms embraced differentiation strategies consistent with efforts to identify niche market positions. In addition, sales per employee and sales growth in the previous five-year period before the firms were surveyed were significantly higher for businesses pursuing differentiation or focused differentiation strategies than for businesses pursuing a cost or cost focus strategy. These results are consistent with work undertaken in the United Kingdom by O’Farrell, Hitchens, and Moffat (1992).

The issue of externalization was also explored by Beyers and Lindahl, building on work by Coffey and Drolet (1996). This research explored whether growth in demand for producer service firms came about because of externalization of producer service functions by their clients. About 41% of clients did not have in-house departments providing services similar to the producer service firms being interviewed, while about 36% had noncompeting in-house departments. In the remaining firms (23%) there were competing in-house departments. Regarding the tendencies for externalization by those establishments that did not have in-house departments, the majority had never produced the service provided by the producer service establishment. This result is consistent with the strong tendency, already noted, towards niching and differentiation. For the noncompeting in-house departments, the results were mixed, with some tendency for externalization, suggesting an expansion of the procurement of a producer service simultaneous with the internal production of that service. In contrast, firms with competing in-house departments were split between externalizing and not externalizing (Beyers and Lindahl 1996).

It is also useful to look at the supply side with respect to changes in services offered, changes in industries to which services are sold, and changes in the geographic locations in which services are sold (Beyers and Lindahl 1997). This analysis was framed within the notion of adaptive behavior and business success developed by Ansoff (1965). Firms that were the most adaptive in changing these supply dimensions showed higher growth rates in sales and employment than those that did not adapt.

The modes used to supply producer services have also been examined (Beyers 2000). The strong evidence for differentiation processes on the supply side, and the lack of externalization tendencies on the demand side, are related to ways that the outputs of producer
service businesses find their way to clients. The unstandardized nature of much of this work means that clients and suppliers must get together on a personal face-to-face basis to conceptualize exactly what the producer service firm will deliver. These face-to-face meetings were overwhelmingly important, either with or without written documents, in the production and delivery of producer services. Supplementing these face-to-face meetings were contacts using the telephone and electronic media such as fax and e-mail. Contrary to some expectations, face-to-face meetings had not diminished in importance, even following the inexorable rise of digital technologies.

The survey research of Beyers and Lindahl was conducted in the 1990s. It clearly supports Porter’s differentiation strategy as a foundation for the growth in supply of producer services, and it leads to the conclusion that on the demand side growth has not been the result of externalization but rather has been driven by noncost and quasi-cost considerations. Survey research was primarily conducted in the United States, the United Kingdom, and in Canada on these demand and supply issues. Survey research on this theme does not seem to have been conducted recently; current research could lead to different conclusions.

There has been much speculation about how the rise of new information technologies could revolutionize the supply and demand of producer services as a consequence of the “the death of distance” and arguments that “the earth is flat,” implying fundamental restructuring in the organization of economies. Thus far, there is no evidence to suggest that, even in the face of Internet commerce, smartphones, and other forms of digital communication, the fundamental foundations for the demand and supply of producer services have been altered. The continuing, relatively rapid growth of this segment of contemporary economies is related to the underlying forces of demand and supply, including innovations in the lines of service offerings that find a marketplace. Forecasting industrial structural change is risky, and it is certain that new innovations in producer services will emerge as people develop ideas that find their legs in the marketplace.

SEE ALSO: Central place theory; Industrial location theory; Input–output analysis; Producer services: definition and classification

References


Democracy

Clive Barnett
University of Exeter, UK

Murray Low
London School of Economics and Political Science, UK

Democracy has a straightforward meaning: “rule by the people”; however, the meaning of “the people” and of “rule” is far from straightforward. The historical geography of democracy is a story of various attempts to find answers to practical problems such as: Who belongs to the demos? How should rule be organized? What is the scope of the activities over which democratic rule should be extended? Disagreement over basic definitions, practices, and institutions is constitutive of the dynamism of actually existing democratic politics.

Questions of the scope of rule and the delimitation of the people include important geographical questions. For example, the problem of rule raises issues of the spatial reach and territorial organization of the mechanisms of popular rule. Is democratic rule synonymous with electoral practices and representative mechanisms? At what geographical scales are such activities best organized? And does democracy also extend to more participatory practices such as citizens’ juries or local participatory budgeting? Likewise, the problem of defining “the people” raises questions about how to determine membership of a polity. Should cultural criteria of shared belonging to a community prevail? Should criteria of place of residence be used? Or should a spatially expansive principle of affected interests be used to define the identity of those who should have a say in shaping decisions?

Despite the centrality of geographical questions to the practical politics of democracy, democracy has not been an explicit object of analysis in geography until very recently. Heightened interest in this topic is in part a reflection of real-world processes of democratization, and an awareness of the significance of electoral processes to the exercise of power in the real world. It is also, however, a reflection of shifts in the normative paradigms which underwrite critical human geography in particular. Until recently, there has been little explicit focus in human geography on the normative issues that are at the core of debates about the relationship between democracy and spatiality. Geographers have often avoided issues raised in fields of political philosophy on the grounds that they seem to be too “liberal.” Since the 1990s, however, there has been a shift in various subdisciplines toward investigating the ways in which normative problems central to democratic theory play out in practice. These issues are reflected, for example, in discussions of citizenship, deliberation, and participation in areas such as development geography, urban studies, planning studies, and environmental studies. The absence of democratic institutions is now recognized as a key factor in the reproduction of social injustice and inequality, and the exposure of vulnerable or marginalized groups to serious harm. At the same time, geographers have begun to investigate innovative mobilizations by social movements that reconfigure the meanings and practices of democracy. At a more abstract level, the theories of spatial ontology

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0837
DEMOCRACY

developed by geographers have opened up new understandings of “the political,” and have drawn the discipline into debates informed by a distinctive strand of contemporary political theory that focuses on the agonistic, dissensual aspects of democracy. This strand of theory supports critical evaluations of the contemporary world as increasingly “postdemocratic” and/or “postpolitical” (see Swyngedouw 2011).

In short, issues of democratic practice and problems of democratic theory are now much more evident in human geography than they once were. Broadly speaking, this increased concern with democracy is divided across two areas of inquiry: research on geographies of liberal democracy and explorations of a variety of alternative forms of democracy.

Geographies of liberal democracy

Liberal democracy is usually taken to refer to forms of institutionalized popular representation, involving periodic mass election of representatives to authoritative legislatures (and, in many cases, of executive heads of state), under conditions of freedom of speech and of association. This is a model unevenly developed in the Global North, and widely emulated and/or imposed in other parts of the world. Liberal-representative democracy is usually legitimated by appealing to the basis of politics in the preferences and liberties of private individuals recognized as equal citizens under the rule of law. There is an enormous field of literature, largely in political science, devoted to examining the conditions under which such liberal democratic institutions develop, as well as to particular aspects of these systems, including political parties, elections, voting behavior, citizenship, and the media. There is also a substantial body of liberal political theory associated with the justification of institutions of liberal democracy, a body of work that often provides critical insights into the operations of actually existing liberal democracies. Both political science and political theory assume a given spatial formation of democracy as the norm: territorialized national states, internally divided in some way or another.

Modern democratic theory has therefore developed around a particular geographical imagination. It assumes that democracy is framed around bounded national territories which are internally organized through a nested hierarchy of scales. In geography, research on liberal-representative democracy has focused on two areas: electoral geographies and geographies of democratization.

The subfield of electoral geography focuses on explaining and mapping the spatial distribution of votes. Geographical processes can make a significant difference to the outcomes of elections in terms of who wins and who loses. Geographers have demonstrated how the spatial organization of electoral systems affects how votes are translated into representative majorities in liberal democracies. Recent attention has focused on developing more sophisticated and contextually sensitive explanations for voting behavior. Research on electoral processes has broadened out to include the geographies of campaigning, party formation, and political communication. Work in electoral geography has shown, then, that the spatial organization of formal democracy has consequences for outcomes in terms of criteria of equality, fairness, and representativeness. Accordingly, recent research in electoral geography has more explicitly addressed the normative issues raised by practices of gerrymandering, political redis-
Geographers have also explored the process of democratization as part of a broader examination of the conditions for the formation and maintenance of liberal democratic polities around the world. The diffusion of liberal democracy as a global norm of governance since the 1980s has followed in the wake of the collapse of communism in Eastern Europe; political transitions away from authoritarianism in Latin America, Africa, and Asia; and the application of norms of democratic governance in the geopolitics of Western international financial policy, trade negotiations, and military engagements. Geographers have investigated the degree to which the adoption of democratic forms of governance can be accounted for by specifically geographical factors. This research contributes to debates over whether democracy can be established and sustained only after various socioeconomic and cultural prerequisites have been met, or whether it is itself the prerequisite of economic development or peace-building.

Geographers have also critically assessed the theoretical assumptions through which liberal forms of electoral democracy have been circulated as the global norm. Debates about democratization raise questions regarding the degree to which the norms of Western liberal-representative democracy can and should be practically applied in non-Western contexts and deployed as normative benchmarks of critical analysis. Rather than democracy having a single form that is traceable to a unique origin point, the mobility of democratic practices suggests that the devices through which different imperatives of democratic rule are enacted can be combined and adapted in different ways. As a consequence, the values enacted through these devices — for example, the ways in which interests are represented, conflicts resolved, participation practiced, accountability enforced — may be highly variable in different geographical contexts (Bell and Staeheli 1998).

Alternative democratic spaces

Criticisms of theories of liberal-representative democracy that assume the nation-state as the natural container of democratic politics have encouraged geographers to give increasing attention to various alternative models of democratic politics. There are two aspects to this focus: a concern with scales of democratic politics above and below that of the nation-state; and a concern with the potential for radicalizing democratic politics in established and new arenas.

The first aspect of the attention to alternative models of democratic politics is the focus on the spaces of democratic politics that exist above and below the level of the nation-state. While electoral geography tends to focus on the translation of subnational electoral outcomes into aggregate national election results, a burgeoning literature focuses on the city or the region as a privileged scene of democratic politics in its own right. The focus on these subnational scales is related to the argument that place is the key spatial modality in which identities and interests are experienced and negotiated. At the same time, in political science and political theory, the territorial framing of democracy is increasingly subjected to critical investigation by confronting democratic theory with the facts of contemporary transnationalism and globalization. Increasingly, settled assumptions about inclusion and participation in democratic politics at national scales are thrown into question by transnational social mobilizations and identifications, which point up the limitations of territorialized models of representation. Settled assumptions about the objects of democratic decision-making are, likewise, increasingly under strain with the
proliferation of nonterritorial concerns, such as global climate change, pandemic disease, and integrated transnational financial markets.

The shift of attention to the complexity of spaces and scales through which democratic politics operates is related to a second aspect of research on alternative democratic spaces in geography. It is concerned with searching out and affirming various fractures in liberal representative regimes. These explorations are committed to thinking of democracy as more than simply a procedure for legitimizing the decisions of centralized bureaucracies and holding elected representatives accountable. Critical human geographers tend to think of democracy first and foremost in terms of contestation and dissent. It is assumed that “the political” is a realm in which new identities are formed and new agendas generated, and through which the stabilized procedures, institutions, and identifications of official politics are contested and potentially transformed. Such an assumption is related to an alternative sense of “democratization,” understood not so much as the geographical diffusion of established norms of democracy but as the deepening of democratic impulses and their extension of new arenas of everyday life. In contrast to the focus of electoral geography and mainstream political science on the formal democratic procedures of elections, voting, and parties, attention is paid to the diverse practices and sites where questions of accountability, citizenship, justice, and participation are contested.

Different schools of democratic thought which inform geographers’ focus on alternative democratic spaces tend to share an emphasis on “participation” as the central normative value of democratic politics. Geographers have drawn on a broad family of egalitarian thought about the relationships between democracy, rights, and social justice, including post-Rawlsian political philosophy; post-Habermasian theories of democracy, including feminist critical elaborations of this tradition; various postcolonial liberalisms; agonistic liberalisms and poststructuralist theories of “radical democracy”; the revival of republican theories of democracy, freedom, and justice; and anarchist and “autonomist” traditions of political activism. All of these traditions share a conviction that democratic politics is about more than formal procedures for the aggregation of individualized voter preferences. Taken together, they amount to a tradition of thinking about “radical democracy.” In geography, radical democracy has most often been conceptualized with reference to a post-Marxist approach developed by the political theorists Ernesto Laclau and Chantal Mouffe (see Massey 1995). Their work emphasizes the idea of democracy as a diverse set of practices of challenging the boundaries of “the political” in two senses: by establishing new issues as counting as political, and by focusing on the active formation of subjectivities and identifications in and through political practices of agonistic contestation and collective mobilization. Another feature of this style of radical democracy is the considerable emphasis it puts on issues of space, understood as a scene of multiplicity and encounter with difference, and therefore as the very site of democratic culture. It is an understanding which emphasizes the idea that democracy is an inherently contested, evaluative, incomplete formation that is always open to further perfection, without hope of its finally culminating in a pure form.

Rather than being restricted to this tradition of post-Marxist theory, heavily inflected by poststructuralist ideas, the notion of radical democracy might actually be better thought of as referring to a variety of approaches that emphasize the inclusive, participatory, and contestatory qualities of democratic politics.
The most influential of these approaches is the family of theories of so-called “deliberative democracy,” often shaped by a critical interest in the ideas of Jürgen Habermas. Approaches that develop strategies for informed, structured participatory discussion and decision-making are now central to debates on democratic theory. These approaches propose a much more active role for citizens in all facets of decision-making, as well as, in many cases, the extension of democratic norms to a far wider array of activities. They also involve thinking about a range of activities that are not directly political, in the sense of being tied directly to decision-making processes, as nonetheless important facets of a broader democratic culture by virtue of their contribution to the quality of the public sphere.

Theories of radical democracy, whether of a Habermasian or poststructuralist inflection, often focus on forms of noninstrumental cultural practice and performance. These are understood as key mediums of democratic opinion formation or democratic subject formation. In turn, both traditions share the problem of whether the forms of influence that can be generated in the public sphere or in cultivating an agonistic democratic ethos can or should be articulated with the institutionalized exercise of power. Post-Habermasian approaches finesse this issue through concepts of weak and strong public spheres. Poststructuralist theorists who address the communicative conditions of pluralistic democracy through figures of democratic ethos are even more wary of drawing too close to sites of decision-making and rule. For them, the force of democratic politics is reserved for the disruptive energies of contestatory practices that reconstitute the dimensions of the political arena.

In practical terms, theories of radical democracy are often associated with calls for the decentralization of decision-making and political participation to subnational scales of regions and cities. This localizing impulse easily falls into the trap of assuming that subnational scales of governance are somehow more democratic by virtue of being closer to people’s everyday concerns. In contrast, the rescaling of political authority away from the nation-state is sometimes presented as being a means of depoliticizing issues. At the same time, increasing attention is given to emergent forms of transnational democracy, focusing on whether systems of globalized economic and political governance can be subordinated to democratic oversight. In work on issues of global justice and transnational democracy, one finds sustained attention to the diverse agents of justice through which democratic justice can be pursued and secured, moving beyond an exclusive focus on states as the privileged containers of democratic politics. Again, sometimes this process of rescaling is seen to exemplify the development of “postpolitical” or “postdemocratic” forms of rule. It may be more productive to think of emergent transnational aspects of democratic politics as raising two important challenges to how the geography of democratic ideals and practice are conceptualized. First, this process challenges any simple contrast between representative and participatory forms of democracy (Low 1997). Any viable form of democratic polity is likely to combine aspects of these practices in different ways. For both practical and normative reasons, representation seems an irreducible aspect of any viable, pluralistic model of democracy. Not only do representative procedures enable the “time–space distanciation” of democratic politics, but they also embody important principles of difference and nonidentity within the “demos.” Representation is also an unavoidable mechanism for the integration of so-called mute interests, such as future generations or nonhuman actors. Second,
DEMOCRACY

Transnational political processes are suggestive of a move beyond the predominant territorial framing of the spatiality of democracy. Rather than thinking simply in terms of the need to articulate subnational and national scales with global scales, discussions of these topics increasingly focus on the diverse spatialities of democracy, ones that articulate territorial and nonterritorial practices, and on scalar and nonscalar conceptualizations of space.

Conclusion

Geographers’ contributions to spatializing the analysis of democratic practices such as elections, and to imagining alternative democratic geographies, overlap with emerging concerns about how democracy can be understood in a world where the sites of politics are multiplying and crucial decisions are made in ways that cut across territorially organized political regimes. In political science and political theory, contemporary debates about democracy often hinge on geographical themes and geographical problems. In short, geography is at stake in a new way across the social sciences: understanding contemporary democracy requires geographical contributions by everyone, and democracy as an object of analysis offers considerable potential for new forms of cross-disciplinary engagement. Further engagement with these debates by geographers, however, presents a challenge for prevalent understandings of critical vocation in the discipline. The predominant understanding of critical analysis in human geography has always distanced itself from what are presented as “liberal” approaches, where liberalism is usually understood by reference to economics. It is difficult to appreciate the concerns of contemporary theories of radical democracy without acknowledging the degree to which such theories draw on key themes from traditions of political liberalism, such as pluralism, citizenship, rights, or the separation of powers. Radical democracy in this expansive sense is not best thought of as a rejection of liberalism tout court.

One of the lessons learned by diverse strands of radical political theory from the experience of twentieth-century history is that struggles for social justice can create new forms of domination and inequality. Liberalism can be thought of as a potential source of insight into the politics of pluralistic associational life. The problems of coordination, institutional design, and justification, on which liberal political theory dwells at length, are unavoidable in any normatively persuasive and empirically grounded critical theory of democracy. It is tempting to see in the energy of theories of radical democracy a promise that everything can, and should be, democratized, and that all interests can, and should be, included in democratic practices; however, an abiding aspect of liberal theories of democratic politics is a concern with the justifiable limits of collective decision-making, a concern informed by an acknowledgment of the potential harms, injustice, and violence that can be generated by the putatively democratic exercise of coercive power. Democracy, it should be remembered, is a way of doing politics, and not a substitute for politics. Thinking about the ways in which commitments such as those based on faith, or practices mediated through bureaucratic institutions, or relationships underwritten by the force of violence, can and should be democratized remains a pressing problem of empirical, normative, and theoretical analysis.

SEE ALSO: Citizenship; Civil society; Electoral geographies; Political geography; Social movements; State, the
References


Further reading


The term *demographic transition* refers to the shift in vital rates within population groups (local, regional, and national) from a pattern of high birth (fertility) and death (mortality) rates to one of low rates. The *epidemiologic(al) transition* is part of the mortality decline side of this transition, referring to the change from an infectious disease-dominated cause of death regime (e.g., diarrhea, pneumonia, and tuberculosis) to one in which chronic, degenerative ailments/"noncommunicable diseases" (e.g., cancers, cardiovascular diseases, mental disorders, and dementias) predominate. An important consequence of demographic and epidemiological transition is changing population age structure, from “youthful” to “aged” numerical dominance, that is, an *aging transition*. Various other transitions are, in turn, associated with these changes, in both cause and effect interactions, for example, nutrition, mobility, urbanization, technology, and social and economic development (Figure 1).

**Demographic transition**

In pretransition populations, high birth and death rates virtually cancel each other out and population growth is, therefore, slow. Typically, the first stage of demographic transition is a fall in death rates, while fertility remains high. As a result, population growth begins to accelerate. Later, birth rates also begin to fall. Ultimately, fertility and mortality reach relatively similar low levels, producing a new slow population growth equilibrium. Post-transition societies also have slow (or even negative) growth as the new lower birth and death rates similarly virtually balance each other.

This broad three-stage typology of population change – pretransition, transition, post-transition – is one of demography’s longest established models. The label “demographic transition” was first used in 1945 by American demographer, Frank Notestein, but the idea of transition and similar terminology can be traced back to earlier writings. Notestein is also generally credited with providing the first definitive statement of demographic transition theory, explaining the observed different patterns and levels of fertility and mortality around the world in the context of social, economic, and political modernization. His and other modernization-centered work of the time is now referred to as classic demographic transition theory.

Decades of subsequent research have, not unexpectedly, produced a more complex and nuanced understanding of fertility and mortality decline. The emphasis on modernization has been long dismissed, with detailed research at various scales showing that while modernization can stimulate birth and death rates to drop, it is not essential. Numerous examples of improving mortality and falling fertility in the absence of any significant social and economic development can be cited (e.g., Bangladesh and Sri Lanka).
**Figure 1** Demographic, epidemiological, aging, and nutrition transitions. Source: McCracken and Phillips (2009). Reproduced by permission of Elsevier.
The world of gradual birth and death rate change that Notestein and other early transition theorists wrote about has in the decades since changed to one of accelerated change, of national government involvement in population health and reproductive matters, of formal family planning programs, international assistance agencies, cheap and effective reproductive control technologies, and enormously increased life-saving public health capabilities. Explanations of the fertility decline side of demographic transition are now couched in a variety of perspectives – social, economic, cultural, ideational, and institutional (Kirk 1996).

Demographic transition first occurred in Western Europe, with death rates in various countries showing signs of decline from around the mid-eighteenth century. The exact reasons for the early improvement in health conditions have been and still are a matter of debate, though better nutrition is generally accepted as having been one key factor. Fertility subsequently started to drop, at first gradually, but then rapidly over the latter nineteenth and early twentieth centuries. A similar fertility transition occurred at much the same time in overseas European populations (e.g., North America, Australia, and New Zealand), and by the late 1920s the transition to low birth and death rates was essentially “completed.”

During the 1930s and 1940s, falling death rates and incipient fertility decline were apparent in Japan, some Latin American countries (e.g., Cuba and Argentina), and a small number of other developing nations; however, for most parts of the “developing” world the transition did not discernibly begin until after the World War II. In the immediate postwar years, public health campaigns achieved rapid life expectancy gains in many such countries, and rates of population growth accordingly accelerated. Widespread fertility transition across the developing world started in the 1960s and has continued in the decades since. Progress through the transition has varied (Table 1). Some erstwhile developing nations (e.g., China and South Korea) now have well below replacement level fertility comparable to or even lower than many developed nations. Other nations are at a more intermediate stage in the transition, with birth rates still well above replacement level. Meanwhile in a few cases (mainly in sub-Saharan Africa) fertility transition has essentially still to begin, with current total fertility rates above six births per woman.

In developing nations, the fall in death rates has generally been faster than historically occurred in Europe and other more developed regions. This difference, with the normal lag in fertility decline, has resulted in considerably more rapid mid-transition population growth than experienced in developed countries. Also, over time birth and death rate transition has started at progressively lower levels of development. Within contemporary developing countries fertility decline is usually substantially more advanced in urban than in rural areas. Similarly, fertility transition has generally progressed further in the more developed provinces/regions of countries.

Today, over one-third of the World Health Organization’s member states have below replacement level fertility (usually assumed as a total fertility rate (TFR) of 2.1 children per woman). In around 40 of these, the TFR has fallen to 1.6 or lower, this move to extremely low fertility being termed a second demographic transition by some population analysts (Lesthaeghe 2010). This new stage in demographic development is seen as reflecting the virtually full control now enjoyed over fertility and a shift in norms toward materialism and individualism leading to a move away from formal marriage and parenthood.
Table 1  Demographic and epidemiological transition indicators, early 2000s.

<table>
<thead>
<tr>
<th></th>
<th>World</th>
<th>Low income countries*</th>
<th>Lower-middle income countries*</th>
<th>Upper-middle income countries*</th>
<th>High income countries*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude birth rate (per 1000 population), 2011</td>
<td>24.3</td>
<td>35.5</td>
<td>26.9</td>
<td>15.4</td>
<td>12.5</td>
</tr>
<tr>
<td>Crude death rate (per 1000 population), 2011</td>
<td>7.9</td>
<td>9.4</td>
<td>8.0</td>
<td>7.0</td>
<td>8.4</td>
</tr>
<tr>
<td>Total fertility rate (per woman), 2011</td>
<td>2.4</td>
<td>4.0</td>
<td>2.9</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>Life expectancy at birth (years), 2011</td>
<td>70</td>
<td>60</td>
<td>66</td>
<td>74</td>
<td>80</td>
</tr>
</tbody>
</table>

Causes of death (%), 2010

<table>
<thead>
<tr>
<th>Causes of death (%)</th>
<th>World</th>
<th>Low income countries*</th>
<th>Lower-middle income countries*</th>
<th>Upper-middle income countries*</th>
<th>High income countries*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communicable, maternal, perinatal and nutritional conditions</td>
<td>24.5</td>
<td>54.3</td>
<td>33.8</td>
<td>10.2</td>
<td>6.6</td>
</tr>
<tr>
<td>Noncommunicable diseases</td>
<td>66.4</td>
<td>36.4</td>
<td>55.8</td>
<td>80.6</td>
<td>87.3</td>
</tr>
<tr>
<td>Injuries</td>
<td>9.1</td>
<td>9.4</td>
<td>10.4</td>
<td>9.1</td>
<td>6.1</td>
</tr>
<tr>
<td>Median age (years), 2011</td>
<td>29</td>
<td>20</td>
<td>25</td>
<td>33</td>
<td>40</td>
</tr>
</tbody>
</table>

*World Bank income groups.

In a number of European countries and Japan, the crude birth rate is now lower than the death rate, and natural population decrease is occurring. United Nations’ projections see this phenomenon becoming more widespread in coming decades. Whether long-term population decline becomes the norm in many countries, however, obviously cannot be foretold.

Epidemiological transition

The term epidemiologic(al) transition describes the change in patterns of health and disease as populations undergo demographic transition and was introduced by Abdel Omran in 1971 in what is now considered one of the classic papers of epidemiology and public health. The notion of such change, though, can be found in earlier demographic and international health writings.


1. That mortality is a fundamental factor in population dynamics.
2. That during the transition, pandemics of infection are gradually displaced by degenerative and “man-made” (human induced) diseases as the chief form of morbidity and primary cause of death.
DEMOGRAPHIC AND EPIDEMIOLOGICAL TRANSITION

3 That the most profound changes in health and disease patterns during the transition occur among children and young women.

4 That the shifts in health and disease patterns that characterize the transition are closely associated with the demographic and socioeconomic transitions that are part and parcel of modernization.

5 That variations in the pattern, pace, determinants and consequences of population change differentiate three basic models of epidemiological transition.

From these propositions Omran went on to build his theory around the ideas of “transition stages,” “models of transition,” and “disease determinants.”

The shift in diseases during the course of demographic transition referred to in Proposition 2 was argued to involve three major successive stages (“ages”) (Omran 1971, 737–739), as follows.

1 The age of pestilence and famine: the period before the transition when mortality is high and fluctuating and dominated by epidemics, famine, and wars.

2 The age of receding pandemics: when mortality declines progressively, and epidemic crises become less common.

3 The age of degenerative and man-made diseases: when mortality continues to decline and the shift from an infectious to degenerative disease-dominated health profile is completed.

Three basic models of epidemiological transition were identified (Omran 1971, 739–739).

1 The classical (Western) model, in which the shift in mortality patterns occurred gradually and progressively over a century or more.

Most western countries were seen as fitting this pattern.

2 The accelerated model, in which the mortality change occurred over a far shorter period of time, Japan being cited as typifying this model.

3 The contemporary (or delayed) model, characterizing the “relatively recent and yet-to-be completed transition of most developing countries,” Chile and Sri Lanka (Ceylon) being identified as illustrative examples.

Finally, three major categories of disease determinants were outlined (Omran 1971, 739–741): ecobiologic; socioeconomic, political, and cultural; medical and public health. In the classical (Western) model, primacy was attributed to ecobiologic and socioeconomic factors, with medical factors “largely inadvertent until the twentieth century, by which time pandemics of infection had already receded significantly” (Omran 1971, 741). For Japan and other accelerated model countries, the mortality reduction and disease transition were seen as stemming from “sanitary and medical advances,” along with “general social improvements,” while in the case of the delayed model imported medical technology and public health measures were deemed paramount.

In essence, Omran promulgated a generally uninterrupted progression of countries from the pretransition age of pestilence and famine through to the age of degenerative and human-made diseases. The impression conveyed was of the third stage being the apparent end-point of human epidemiological evolution. Also implicit was the notion of the transition as a universal process, an epidemiological pathway along which all countries would eventually proceed (McCracken and Phillips 2009).

Since Omran’s paper was published, the concept and theory of epidemiological transition
have received wide attention from researchers across an array of disciplines as well as by population health planners and policymakers in both more and less developed nations. As might be expected, there have been numerous critiques and extensions of Omran’s foundational statement. Since his time of writing, the global HIV/AIDS pandemic and health reversals in several of the former Soviet Union republics in the upheaval following the dissolution of the USSR have clearly shown that smooth, uninterrupted passage through the transition is not guaranteed. The third and apparent final stage outlined by Omran has also subsequently been shown not to be the end-point of epidemiological change. Life expectancy in developed (and many less developed) countries has risen to levels well beyond that envisioned by Omran. The degenerative disease domination of late stage transition still applies, but in high income nations there has been a significant delaying of deaths from these conditions to later ages. More attention has also been given to the fact that, during the transition, many countries experience a challenging dualistic phase, when there is simultaneously heavy loss of life and healthy years from both infectious and degenerative diseases – what has come to be variously termed the double burden of disease or epidemiological polarization (WHO 1999). A social pathology type health profile of violence, substance abuse, suicide, accidents, and HIV has also been recognized in seriously disadvantaged groups in some countries – for example, inner city young black males in the United States, some indigenous communities in Australia. As well, greater emphasis has been given to the morbidity side of epidemiological transition and a more complete view of health status and its determinants. This is especially true in the more aged societies, where “epidemics” of dementias have been identified. In this vein, some researchers and policymakers prefer to talk of health transition rather than what they see as the narrower idea of epidemiological transition (Caldwell 1993).

Table 1 shows the current cause of death patterns for the world as a whole and by income groups of countries (as defined by the World Bank). Globally, epidemiological transition has advanced to the stage where noncommunicable diseases account for two out of every three deaths. One-quarter of all deaths, however, are still due to “communicable, maternal, perinatal and nutritional conditions,” despite a large proportion of these being relatively easily and cheaply preventable. In low income countries, over half of all deaths are due to these conditions. The emerging double burden of disease in these countries poses major difficulties in determining respective resource allocation between old pretransition style diseases and new “modern” health problems in pressured health budgets. Lower-middle income countries also face the difficulties of the double burden and many are more advanced in the transition to a degenerative disease-dominated health regime. In upper-middle income and high income countries, “old style” (infectious disease) mortality has largely been displaced by “new style” (degenerative) causes of death. Infections, however, remain a major (and often underestimated) health problem in terms of morbidity and healthy wellbeing in many countries and it is feared that potentially uncontrollable infections may again come to the fore, as antibiotic resistance grows and many new or resurgent viral conditions emerge for which there is currently little effective remedy.

Therefore, comments on future epidemiological transition developments can, by definition, only be speculative. The past half-century has seen both the emergence of numerous serious new infectious diseases (e.g., HIV/AIDS, SARS, hepatitis C, and Middle East respiratory syndrome coronavirus – MERS-CoV) and the resurgence of some that appeared close
DEMOGRAPHIC AND EPIDEMIOLOGICAL TRANSITION

to being beaten (e.g., tuberculosis, especially drug-resistant varieties, and malaria). A global pandemic of virulent influenza is always a threat. Along with the growing problem of antimicrobial resistance amongst pathogens, these examples serve as a reminder that infectious disease is a long way from being vanquished and could produce new and even unexpected epidemiological futures (Fauci and Morens 2012). An obvious major uncertainty regarding infectious disease futures is the impact that global climate change and population concentration into large cities may prove to have.

While infection-based health challenges will almost certainly occur over coming decades, the order of magnitude (i.e., virulence and ease of transmission) required for them to return all humanity to a new infectious disease era is probably unlikely. More likely is a continuing degenerative disease-dominated world, but with evolving patterns. For example, the major decline in cardiovascular disease mortality in developed countries over the past 40 years will probably continue, with cancers coming to be the leading cause of death globally. In turn, the growing importance of dementia morbidity and mortality in developed countries in recent years will almost certainly intensify in coming decades in association with continued population aging. Dementia will also become a major health burden in today's developing countries as they undergo population aging.

Most population health specialists see human life expectancy continuing to improve over the course of the twenty-first century, although there is considerable debate about what ultimate level is biologically achievable or socially likely. A number of subnational population groups around the world have already attained life expectancy at birth levels of 90 years or more and thus give cause for optimism. A few analysts, however, argue that factors such as the growing global obesity epidemic could in fact spur the reverse, that is, average life expectancy falling and/or health status deteriorating. A major issue if longevity does continue to increase will be whether the additional years of life are years of good, bad or indifferent health.

SEE ALSO: Aging; Environment and disease; Environmental health; Fertility; Health and development; Mortality; Population geography; Public health: human dynamics; Spatial epidemiology

References


Further reading


Dendroclimatology was the first application of the newly founded science of dendrochronology in the first decade of the twentieth century, based on initial findings and later principles formulated by Andrew Ellicott Douglass of the University of Arizona, United States, when analyzing tree rings from ponderosa pine trees. Douglass noted similarities in the patterns of ring widths over the years from one tree to the next, not just in trees growing next to each other in one forest stand, but also in trees growing at a considerable distance from each other. Recognizing these similar patterns across trees and across a geographic region eventually led Douglass to establish the basic principle and technique in dendrochronology known as “cross-dating” (Douglass 1941). Cross-dating exists because all trees within a region respond to the overarching climatic processes that operate from day to day, month to month, and year to year. This response will impart a common climatic signal in all trees in a pattern of wide and narrow tree rings over the years, despite some influences from other types of environmental process that also affect tree growth, such as insect herbivory, wildfires, and air pollution. The value of cross-dating lies in its ability to provide the dendroclimatologist with a precise chronology back in time with annual resolution by accounting for tree rings that may add inaccuracy, such as locally absent rings (rings that are absent from the trunk of the tree where samples are usually taken) and false rings (rings that display intra-annual density fluctuations that can often mimic a true annual ring). Such problem rings could potentially add “noise” to the climatic “signal” desired in dendroclimatology. By accounting for these problem rings, dendrochronologists minimize the noise and maximize the climatic signal, making it possible to determine which climatic factors are most responsible for that common signal in the tree rings and capitalizing on this to eventually reconstruct climate in the past for the full length of the tree-ring chronology.

Douglass recognized the importance of pushing back the climatic record from tree rings for millennia back in time using the giant sequoia trees of the western slopes of the Sierra Nevada in California, but his efforts soon turned to the dating of prehistoric structures throughout the American Southwest. After founding the first laboratory dedicated to the tree-ring sciences in 1937, the Laboratory of Tree-Ring Research (LTRR) at the University of Arizona, Douglass assisted his colleague and former student Edmund Schulman by initiating an ambitious project that would take Schulman across the Americas and elsewhere in search for old trees to extend the record of Earth’s past climate. Schulman’s expertise lay in climatic applications and he was the first to extrapolate precipitation patterns for the American Southwest back for 800 years. Before his unexpected death in 1958, Schulman had analyzed the climatic potential of many tree species, mainly across North America and South America. Key among his discoveries was the venerable bristlecone pine trees of the
White Mountains in southeastern California and Nevada, which reached unprecedented ages, including many over 4000 years old (Schulman 1958). The current age record for one individual bristlecone pine is 5062 years. This particular bristlecone pine, originally discovered by Schulman in the 1950s, was later analyzed by the eminent dendrochronologist Thomas Harlan of the LTRR, who found several hundred additional tree rings to reveal the first verified 5000-year-old living tree.

Beginning in the 1930s and into the 1940s, dendroclimatology was also taking a foothold in Europe, led by Bruno Huber, then at Tharandt in Saxony, Germany, whom many consider the “father of European dendrochronology” (Liese 1978). By the 1960s, Huber and his colleagues would have laid the groundwork and developed improved quantitative methods for analyzing tree-ring chronologies throughout Europe that would eventually aid in understanding the climate of Europe for the past 10 000 years at annual resolution based on tree rings. Further, techniques developed by European dendrochronologists would eventually aid in the dating of literally thousands of historic structures from the Roman period into the Middle Ages and later all across Europe and beyond. Huber realized early in his career that the ring patterns in Europe were distinctly different in their variability across the years from ring patterns in the western United States, making visual comparisons of tree-ring patterns for cross-dating more challenging. Huber and his colleagues introduced the Gleichläufigkeit (percentage of agreement) statistic, the first of its kind, to help add a degree of statistical rigor in the cross-dating of tree-ring patterns.

More formal quantitative techniques for analyzing the climate signal in tree rings and for reconstructing past climate were later generated by Harold Fritts and his coworkers at the LTRR (Fritts 1976). By then, dendroclimatologists recognized the importance of sound training in botanical and biological principles to better interpret the often confounding signal contained in tree rings. These techniques first focused on generating more accurate bioclimatic models of tree growth that incorporated tree-ring data. In other words, to better understand past climate from tree rings, you must first know how trees grow by embracing a more holistic view of environmental influences on tree growth. By the 1970s, quantitative reconstructions of past climate were becoming commonplace, pushed by an emerging need to also reconstruct hydroclimatic parameters (precipitation, discharge, lake levels) from tree-ring data. By the early 1990s, statistically validated millennium-length reconstructions of regional and hemispheric temperature and precipitation patterns were becoming common, spreading from Europe and the United States to Mexico, Canada, Australia, New Zealand, Chile, Argentina, and the Middle East, later occurring also in Russia, China, Mongolia, and Southeast Asia. Recent efforts in dendroclimatology concentrate on analyzing new tree species for their climatic potential, especially in the tropics; extending our knowledge of climatically sensitive tree species into new geographic regions; evaluating the impact of nonclimatic processes on the climate signal in tree rings; establishing whether the relationship between climate and tree growth has remained stable over recent decades; and incorporating tree-ring data into larger ensembles of proxy data for reconstructing past climate. These tree-ring data and the hundreds of climate reconstructions generated from them in many regions of the world are providing valuable assessments of long-term climate patterns over the past two millennia that play a key role in understanding the probable trajectory of the Earth’s climate in the future, as shown in the 2013 report by the
Techniques

Dendroclimatology assumes that established practices and principles have been followed when identifying and collecting tree-ring samples for understanding past climate. Key among these are the Principles of Site and Tree Selection (Speer 2010; Grissino-Mayer 2014) to ensure that climatically sensitive locations were identified during the project design and field reconnaissance phases, and that trees were carefully chosen and sampled to provide the maximum climatic information while minimizing the adverse effects of noise from confounding ecosystem (both natural and human-caused) processes. Furthermore, tree-ring data analyzed are not limited to total ring widths only. Especially useful climatic information has been obtained using maximum latewood density measurements (particularly for European tree species) and stable carbon isopes ($^{13}$C/$^{12}$C ratios) (especially in Europe and the western United States). Additional physical properties used in dendroclimatology include minimum earlywood density, average ring density, earlywood and latewood widths, latewood percentage, “light” rings (i.e., faint latewood), oxygen isotope values ($^{18}$O/$^{16}$O ratios), and annual basal area increment (BAI).

Two approaches are used in dendroclimatology. The first involves the development of (mechanistic) bioclimatic models of annual tree growth by modeling growth (via annual tree-ring data) as a function of one or multiple climatic variables (usually monthly values). In this strategy, tree growth is a function of one or more climatic parameters. Basically, before tree-ring data are used to reconstruct climate in the past, we must first know to which climate variables the trees are responding, as recorded in the physical or chemical properties of tree rings. Techniques first developed by Fritts (1976) and his colleagues of the LTRR used principal components to first transform monthly climatic data (such as temperature and precipitation) and remove interdependence between the predictor climate variables, then used these transformed data in a multiple regression to determine which monthly data most influenced tree growth on an annual basis. This technique was termed response function analysis (RFA), and the models generated are especially important for (i) isolating the primary climate variable to which the trees are responding (e.g., drought or temperature); (ii) determining which months/seasons are critical for tree growth; and (iii) evaluating potential changes in tree and forest growth responses given a changing climate.

Climate variables analyzed include temperature (mean, minimum, or maximum); precipitation; drought (Palmer drought severity index, or PDSI); snowpack; streamflow; and sea-level pressure. Recent research has focused on large-scale ocean–atmosphere patterns, such as the El Niño–Southern Oscillation, Pacific Decadal Oscillation, North Atlantic Oscillation, and Atlantic Multidecadal Oscillation, among others. The months analyzed and modeled always extend back to the previous growing season because tree growth is strongly influenced by conditions set during the prior year. For example, a year of favorable climate could allow trees to fix abundant carbon in their plant tissues (e.g., tissues in the apical meristems), prior to the dormancy period, that translates to enhanced growth during the following growing season. Although such principal components-based analyses are still common, RFA as practiced today involves more the use of simple correlation functions to determine the primary variables to
DENDROCLIMATOLOGY

which the trees respond during specific months and seasons.

The second technique involves the development of (stochastic) time-series models of past climate as a function of tree growth. In this strategy, historic climate data are modeled as a function of tree-ring data (usually over the twentieth and twenty-first centuries) to develop a calibration model that predicts (or “retrodicts”) climate for the full length of the tree-ring record back in time. The dendroclimatologist must first determine the one primary climate variable to which the trees are responding using RFA before the reconstruction can be accomplished. The climate variable can represent a single monthly value (e.g., July drought as modeled using the PDSI); a range of months over the current growing season and/or months prior to the current growing season (e.g., average summer temperature); or an annualized climate variable (such as water-year precipitation, common in hydrologic studies). Most tree-ring-based reconstructions of climate involve first withholding climate data for a specific period during the twentieth/twenty-first centuries and generating predicted values of climate from tree-ring data for that period. If the predicted values from the calibration model track the actual values for that period (and can be verified statistically), then the climate–tree growth relationship is considered stable enough to use the calibration equation for reconstructing climate back in time. In most cases, however, the calibration model would use the full record of climate and tree-ring data available for the twentieth/twenty-first centuries rather than just a specific calibration period.

Main findings

The response by trees to climate varies greatly across geographic regions. For example, in the American Southwest, precipitation is the primary limiting factor and the primary climate variable to which trees respond in their annual growth. Across mainland Europe and in parts of South America, (summer) temperature is the primary climate variable to which the trees respond. In the southeastern and northwestern regions of the United States, some tree species respond primarily to temperature while other species respond primarily to precipitation. Trees that grow in the tropics and subtropics experience relatively little variation in climate from month to month and season to season, such that trees that grow in lower latitudes have limited potential for climatic analyses, but recent advances demonstrate that tropical trees do harbor climate information that can be extracted. Trees that grow in higher elevations and at higher latitudes display a greater sensitivity to temperature variations, while trees that grow in more arid conditions or at the lower elevation limits for that species display greater sensitivity to variations in precipitation. Even within a specific geographic location, trees can be found that respond to one climate variable while adjacent or nearby trees can be found to respond to a different climate variable because of differences in local topography and microclimate.

Recent research has focused on evaluating whether the relationship between climate and tree growth has remained stable over time during the historic (twentieth/twenty-first centuries) period. A major concern among dendroclimatologists (and scientists in general) centers around the issue of “divergence” (D’Arrigo et al. 2008), which describes the diverging temporal trends between temperature records (increasing in the twentieth/twenty-first centuries) and tree growth (decreasing beginning about 1960), especially for trees in the higher latitudes of the Northern Hemisphere which have been prominently used for reconstructing past temperatures. Prior to this, temperature and
DENDROCLIMATOLOGY

Tree growth showed a strong positive relationship: that is, as temperature changed, so did tree growth. The implications of a temporally unstable relationship between tree growth and climate during the twentieth/twenty-first centuries are substantial because dendroclimatology assumes that environmental processes that operate today also operated in a similar manner in the past—the so-called uniformitarianism principle. If this relationship does not hold (remain temporally stable), then tree-ring data may not be reliable indicators of past climate. Researchers have discovered that some tree-ring chronologies display unstable relationships with climate over time, but most studies have found that the tree growth–climate relationship is indeed stable over the twentieth/twenty-first centuries.

Perhaps the greatest contributions of tree-ring data to society have been the reconstructions of past climate generated for many parts of the world that extend our knowledge on seasonal and annual climate well before historical weather records were kept. Reconstructions have been developed for nearly all the climate variables analyzed, across multiple temporal scales (monthly to multidecadal), for nearly all temperate and high-latitude locations, and now extending into the tropics. For example, water-year precipitation was reconstructed for the American Southwest from 136 BCE–1992 CE and revealed several centuries-long periods of above-average and below-average rainfall, including a “megadrought” in the late sixteenth century that has since been reconstructed for other portions of North America. Reconstructed PDSI for Mesoamerica suggests a strong forcing of drought by the El Niño–Southern Oscillation and a strong influence by the Atlantic Multidecadal Oscillation on summer climate variability over Mexico. Reconstructed streamflow for the Colorado River in the western United States suggests that river discharge during the twentieth century was not representative of discharge in previous centuries, calling into question government policies that allocate water resources to various western states in the United States.

Particularly important for understanding past climate have been the reconstructions of temperature for the Northern Hemisphere generated using primarily maximum latewood density data, although ring-width data also have been used. Temperature reconstructions were first pioneered by Fritts and his colleagues (especially Valmore LaMarche) at the LTRR beginning in the 1970s, before eventually being applied in the 1980s by others in the United States (Gordon Jacoby and Edward Cook of the Tree-Ring Laboratory at Columbia University) and in Europe (specifically Dieter Eckstein, Keith Briffa, Joel Guiot, Tom Wigley, and Malcolm Hughes). These initial reconstructions pushed our knowledge on trends in past temperature back several centuries, covering the important climatic episodes known as the Little Ice Age (ca. 1400 to 1850 CE) and the Medieval Warm Period (also called various other names, including the Medieval Climatic Anomaly), which occurred from about 1000 to 1300 CE. The first well-replicated millennium-length reconstruction was developed for Fennoscandia (Briffa et al. 1990) but showed little evidence for either of these two climate excursions. Evidence from Huon pine trees growing in Tasmania, however, showed enhanced growth during the twentieth century that could be attributed to anomalously warm temperatures, but such enhanced growth was not revealed for long-lived alerce that grow in southern Chile which were used to develop the first multimillennial reconstruction of climate (Lara and Villalba 1993).

Beginning in the late 1990s, dendroclimatologists employed hundreds of temperature-sensitive
DENDROCLIMATOLOGY

tree-ring datasets across large portions of the Northern Hemisphere to reconstruct past temperature, based largely on ensemble methods, a strategy that focuses on the development of several different (though related) reconstructions. The ensemble strategy also occasionally uses different proxy sources of past climate (such as borehole temperatures, ice core data, and speleothem data), and averaging/combining the results of the reconstructed temperature. Such ensemble approaches are desirable because error variance in the final reconstruction can be reduced. Mann, Bradley, and Hughes (1998) focused on reconstructing temperature for the past 600 years and found that annual temperatures during several years in the late 1990s were the warmest since the 1400s, stoking an ensuing controversy surrounding the graphic that would later become known as the “hockey stick” (1998: Fig. 5b and later versions). Later independent analyses, using the original but updated tree-ring data as well as new tree-ring datasets, would later confirm that the latter part of the twentieth century was indeed warmer than at any time during the past 1000 years. Recent tree-ring studies have also confirmed that the rate at which temperatures have increased in the past 100 years has been unprecedented during the past 1000 years. New tree-ring studies, however, point to a general cooling trend over the past 2000 years since the Roman Warm Period, a period claimed to have experienced temperatures warmer than those experienced during the twentieth century.

SEE ALSO: Biologic dating techniques; Climate change and biogeography; Dendrogeomorphology; Environmental science; Global environmental change: human dimensions; Nature conservation; Paleoclimatology; Water and climate change

References

The significant contribution of tree rings to documenting Earth-surface processes lies in their capacity both to preserve evidence of past geomorphic activity and to provide critical information on their dating with annual or subannual resolution. In many climates, the tree-ring record may represent the most valuable and precise natural archive for the reconstruction and understanding of past and ongoing Earth-surface processes during the last several hundred years. As many geomorphic processes are significant natural hazards, understanding their distribution, timing, and controls provides valuable information that can assist in the development of mitigation of and defense against these hazards and their effects on society.

This entry provides an overview of tree-ring based reconstructions of surficial processes resulting from different geomorphic agents. It outlines the impact of Earth-surface processes on tree morphology and wood anatomy, and clarifies the approaches used for tree-ring reconstruction of past activity of Earth-surface processes. It also illustrates the breadth and diverse applications of contemporary dendrogeomorphology and underlines the growing potential to expand these studies, possibly leading to the establishment of a range of techniques and approaches that may become standard practice in the analysis of specific hazards.

What surficial processes can be analyzed with tree rings?

Dendrogeomorphology has been used in the analysis and reconstruction of a wide range of surficial processes. Some of the earliest applications of dendrochronological research to Earth-surface processes utilized roots rather than stem samples to study erosion processes. In the 1960s exposed tree roots were used as an indicator of ground surface at the time of germination of 5000-year-old *Pinus aristata* Engelm. in California. More recent work focused on areal denudation in badlands in the French Alps and central Spain to illustrate limitations and possible caveats of past approaches using growth rings of roots to reconstruct erosion rates (Stoffel et al. 2013). After a series of pioneering studies conducted on Mount Shasta (California), tree-ring records were used by several European researchers to reconstruct debris-flow activity in the Dolomites, and more recently in the Alps, where the frequency, spatial patterns (i.e., lateral spread and extent of flows), frequency–magnitude relations, and precipitation thresholds of debris flows have been analyzed (Stoffel, Tiranti, and Huggel 2014). Dendrogeomorphic analyses of landslides were realized primarily in Canada and the French Alps (Braam, Weiss, and Burrough 1987) to document spatial patterns of activity,
Dendrogeomorphology

reactivations, and rainfall patterns triggering landslides. In the first study in the subtropics, suppression–release patterns and inception of *Alnus acuminata* Kunth. trees have been used to date landslides in Argentina. The occurrence of snow avalanches has quite frequently been analyzed with tree rings since the mid-1970s. For a long time, tree-ring-based analysis of snow avalanche activity was almost exclusively used in North America (Butler 1979). Work in the European Alps started in the early 2000s, with a focus on runout distances and triggering weather conditions (Schläppy *et al.* 2014). The analysis of growth rings has also been used in paleohydrology. The first studies linking fluvial processes and botanical evidence go back to the 1960s (Sigafoos 1964). Later on, a broad array of flood signatures was defined to date palaeofloods. More recently, scarred trees have been used to estimate flood discharge and to analyze flood frequency distributions (Ballesteros-Cánovas *et al.* 2014). Surprisingly, and despite the potential of dendrogeomorphic methods, rockfall activity has only rarely been studied with tree rings. Studies analyzing the seasonality, frequency, and spatial patterns of rockfalls using tree rings were only very recently conducted, which provided a reliable basis for the analysis of rockfall risks and the accurate assessment of rockfall models (Trappmann, Corona, and Stoffel 2013).

**Tree reactions to Earth-surface processes**

Apart from the site-specific information common to many trees at any site (e.g., genetic information, age, or climatic conditions), individual trees also record the effects of mechanical disturbance caused by external processes. Trees can be injured, can suffer breakage of their crown or branches, and may have their stems tilted or partially buried or their roots exposed. Evidence of these events can be recorded in the tree-ring series of affected trees (Figure 1). The analysis of geomorphic processes through the study of growth anomalies in tree rings is called dendrogeomorphology (Alestalo 1971). Dendrogeomorphic research is normally based on the process–event–response concept where the “process” is represented by any geomorphic agent, for example, debris flows, snow avalanches, or rockfall (Stoffel *et al.* 2010). Individual geomorphic “events” that affect the tree may result in a range of growth “responses.” These events and associated responses are illustrated in the following paragraphs.

Mechanical disturbance and wounding (Figure 2a) are very common in trees affected by geomorphic processes. When impact locally destroys the vascular cambium, increment cell formation is disrupted and new cell formation ceases in the injured segment of the tree. To minimize rot and insect attacks after damage, the injured tree will compartmentalize the wound and start production of chaotic callus tissue (Figure 3a) at the edges of the injury. The extent of wound healing greatly depends on the annual increment rate, the age of the tree, and the size of the scar. After injury, tangential rows of traumatic resin ducts (or TRD; Figure 3a) are produced in the developing secondary xylem of certain conifer species. They extend tangentially and axially from the injury. When wounding occurs during the vegetation period of the tree, resin production will start a few days after the event and ducts emerge within three weeks after the disturbing event. Therefore, in analyzing cross-sections, the intraseasonal position of the first series of TRD can be used to reconstruct previous events with monthly precision. In broadleaved species, cambial damage will initiate various anatomical responses, among those a reduction in vessel sizes (Arbellay, Fonti, and Stoffel 2012).
Inclination of the stem results from the sudden pressure induced by geomorphic processes directly, by the associated deposition of material (e.g., avalanche snow, debris-flow material) or by the slow but ongoing destabilization of a tree through landslide activity or erosion. Tilted trees (Figure 2b) are common in many areas affected by geomorphic processes, and have therefore been used in many dendrogeomorphic studies to date previous events. Subsequent growth in the trunk of a tilted tree will attempt to restore its vertical position, and the reaction will be

Figure 1  Events affecting a tree and its reaction on a cross-section and in ring widths.
Figure 2  Characteristic examples of trees with disturbances visible on the stem surface: (a) open wound induced by rockfall; (b) tilted trees on a landslide surface; (c) buried tree stems standing in a debris-flow levee; (d) multiple decaptitions in a Larix decidua Mill. tree growing on an avalanche slope.

Figure 3  Reactions of trees to mechanical disturbance: (a) callus tissue (right) and tangential row of traumatic resin ducts (left) in Larix decidua Mill. after wounding; (b) severe growth suppression resulting from tree decapitation; (c) reaction wood in Picea abies (L.) Karst. after stem tilting; (d) growth release in a survivor tree following elimination of its neighbors.

most clearly visible in that segment of the tree to which the center of gravity has been moved through the inclination of the stem axis. In the tree-ring record, eccentric growth (and related reaction wood; see Figure 3c for details) will be visible after a tilting event and thus allows accurate dating of the disturbance. In coniferous trees, compression wood will be produced on the underside of the trunk. Individual rings will be considerably larger and slightly darker in appearance compared to the upslope side. The difference in color results from the much thicker and more rounded cell walls of early- and latewood tracheids. Multiple tilting events in the same stem may be recognized by changes in the amount, color, or orientation of reaction wood series in the ring series. In contrast, stem tilting in broadleaved trees leads to the formation of tension wood on the upper side facing the tilting agent. Broadleaved trees also react on
tilting with ultra-structural modifications that are visible only when studied on micro-sections. In addition to the formation of different types of reaction wood, trees may also respond with reduced growth after tilting.

Debris flows, floods, landslides or “dirty” snow avalanches may bury trees (Figure 2c) by depositing material around their stem base. Growth in these trees will normally be reduced (Figure 3b), as the supply of water and nutrients will be temporarily disrupted or at least limited. If stem burial exceeds a certain threshold, trees will die from a shortage of water and nutrient supply. According to case study results from the Italian Dolomites, *Picea abies* may tolerate a maximum burial depth of 1.6 m to 1.9 m in environments dominated by fine-grained debris flows composed of calcareous material. Occasionally, buried trees produce adventitious roots close to the new ground surface. As adventitious roots are typically formed in the first years after burial, the moment of root sprouting can be used to derive an approximate date for the sedimentation processes.

Bouncing rocks and boulders, debris in flowing water, debris flows and lahars, or the windblast of snow avalanches may cause the decapitation of trees (Figure 2d) or the removal of branches. The loss of the crown or branches is more common in bigger trees, when stems have lost their flexibility. Apex (the uppermost part or the growing tip of a plant) loss has also been observed as a result of rockfall impacts close to the ground level. In such cases, the sinusoidal propagation of shock-waves in the stem results in the breakoff of the crown. This phenomenon has been described as whiplash or a hula-hoop effect. Trees react on decapitation or branch loss with distinct radial growth suppression (Figure 3b). One or several lateral branches will form a “leader” that replaces the broken crown, resulting in the tree morphology called “candelabra” growth (see Figure 2d).

Erosional processes and the (partial) denudation of roots may generate different growth reactions, both in the stem and in the exposed roots. The type and intensity of the reaction(s) will depend on the nature of the erosive event, which may be instantaneous or progressive and gradual. If several roots are completely denuded during a sudden erosive event (e.g., debris flow, lahar, flood, or landslide), they are no longer able to fulfill their primary functions and will quickly die. The tree subsequently suffers from a shortage of water and nutrient supply, resulting in suppressed tree growth and the formation of narrow rings in the stem. In cases where only part of a root is exposed and its outer end remains in the ground, the root will continue to grow and fulfill its functions. In the exposed root, however, anatomical changes will occur and individual growth rings similar to those in the stem or branches will be formed. The localization of this change in the tree-ring series may allow determination of the moment of exposure. The continuous slow exposure of roots is usually caused by gradual processes and relatively low denudation rates, for example by overland flow; the slow opening of cracks in soils (e.g., soil creep, landslides) and in disintegrated bedrock, along rivers and streams, and beside lakes and oceans (floods, shore erosion), as well as by faulting activity and displacements in relation to earthquake activity. Provided that the roots are gradually exposed with time, it is also possible to determine the erosion rates in such cases.

Geomorphic processes can eliminate trees along channels or couloirs through uprooting and stem breakage and leave neighboring trees intact. This phenomenon can be observed with, for example, rockfalls, debris flows, lahars, extreme floods, landslides, and snow avalanches. The uninjured survivor trees have
less competition and more light, nutrients, and/or water. As a consequence, they may start to produce larger increment rings (Figure 3d). However, observations indicate that this growth release in survivor trees can be delayed, and therefore this reaction cannot always be used to date past destructive events with yearly precision. Nevertheless, the growth release in survivor trees can corroborate the dating of events that have been identified from evidence in other trees at the same site, for example from scars and tilted stems (Stoffel and Corona 2014).

Many different Earth-surface processes can eliminate surface vegetation including entire forest stands, leaving no direct dendrogeomorphic evidence. In such cases, the germination ages of trees growing on bare surfaces can be used to estimate the time of creation of new landforms and the disturbance event. This approach provides a minimum age for that surface and has frequently been used to date landforms or to assess the minimum time elapsed since the last devastating event in snow avalanche couloirs, debris-flow channels or floodplains. It involves estimating the time between the exposure of the surface and the germination of the first surviving seedling on that surface. This ecesis estimate varies according to the environment, substrate, available seed sources, and several other factors. Problems of ecesis determination have been extensively discussed in studies that attempt to date the formation of glacier moraines (McCarthy and Luckman 1993) where ecesis estimates may vary from a few years to several decades.

Field approach and sampling design

The types of damage to trees described in the preceding section may result from a wide variety of Earth-surface processes. Linkage between the damage and the causative process depends on a critical evaluation of the sampling site. In many cases, investigations are undertaken with a specific context where the process or processes involved are known, and appropriate sampling methods and sampling design may be applied. These vary considerably with the nature of the process or investigation and are best discussed in the specific context of individual processes; for example, the sampling network needed to define past earthquake activity may be very different from that required to reconstruct the history of snow avalanches and debris flows on a large debris cone.

At an individual site the choice of sampling design and selection of the trees to be sampled will depend on the purpose of the investigation and the processes or hazards being sampled. The sampled area and the number of trees sampled may be very small and specific (e.g., on a small landslide) or may involve systematic sampling of a larger area (e.g., within a large runout zone of a snow avalanche) or target individual trees damaged at the margins of an event. As a rule of thumb, processes that tend to spread considerably will likely leave larger spatial footprints and will thus be visible in a larger number of trees. In the case of landslides, a sample size of 50 trees seems sufficient; a slightly larger sample size (100 trees) appears appropriate for snow avalanches, while at least 150 trees are needed to characterize debris flows appropriately. The sample size in rockfall studies will depend on the size of the study site, stand density, and mean rock sizes. The choice of the individual trees sampled is based on an inspection of the stem and on the location of the trees in relation to the processes studied.

The tree-ring record of growth disturbances created by past events is analyzed with cross-sections, wedges, or increment cores. Cross-sections or wedges are normally taken at the
location of the growth disturbance, and provide an excellent and complete insight into the tree’s history. However, in many locations (e.g., protection forests, national parks) felling may be prohibited or is ill-advised for aesthetic or safety reasons. Most tree-ring studies thus use cores extracted with an increment borer. The nature of visible growth defects observed in the tree’s morphology will influence sampling height, sampling directions, and the minimum number of samples to be taken per tree.

In trees with visible scars, previous geomorphic events are most easily dated through the destructive sampling of trees and the preparation of cross-sections taken at the location where the injury is largest. This approach will facilitate an accurate and intraseasonal identification of the onset of callus tissue (and TRD) formation, and therefore allows a reconstruction of the impacting event with very high temporal resolution. Alternatively, wedges can be sawn from the overgrowing callus and an increment core extracted from the side opposite to the wound.

Tilted stems are usually analyzed with at least two increment cores extracted per tree, one in the direction of the tilting and the other on the opposite side of the trunk. The reaction wood will be visible on the tilted side in conifer trees (= compression wood) and on the side opposite to the tilting direction in broadleaved trees (= tension wood). Individual cores are best extracted at the location where tilting is strongest, based on an outer inspection of the tree morphology. However, in situations where multiple tilting events may vary in direction – for example in snow avalanche tracks – additional cores, transverse to the downslope direction, may be needed to develop a more complete record.

In the case of buried stem bases, the sampling of two increment cores in opposite directions (ideally from the upslope and the downslope sides) will normally allow accurate reconstruction of the event that has led to the sedimentation of material around the stem’s base. It is best to sample these trees as close to ground level as possible (ca. 20–40 cm) to extract a maximum number of tree rings and obtain maximum information. However, the influence of roots should be avoided.

The analysis of exposed roots is normally based on cross-sections because they regularly show partially or completely missing rings. The position of samples needs to be carefully noted with respect to the ground surface or the noneroded parts of the root. These samples are essential for the understanding and interpretation of continuous erosion processes and the determination of denudation rates.

The germination of trees on new landforms is best determined by counting the annual growth rings in increment cores taken from the root crown level. However, branches, obstacles, and rot may sometimes require sampling positions higher up on the stem. In these cases, an age correction factor needs to be added to compensate for the time a tree takes to grow to sampling height. A simple height–age correction can be achieved by dividing the tree height by the number of tree rings to get an average rate for the yearly apical increment.

In some cases a reference tree-ring chronology may need to be developed from a nearby forest stand to assist in the cross-dating of trees at the sampled site to verify the age assignment of rings. Adequate sample size varies between species and also depends on the strength of the common signal. In general, 20–30 trees is a useful guide for most species, to minimize the potential influence of geomorphic processes and hidden growth disturbances.
Sample analysis and interpretation

Ring widths of the disturbed samples may also be measured, and the series compared graphically or statistically with the reference chronology. These tools also allow for an assessment of cross-dating accuracy between ring-width series of individual disturbed trees and the reference chronology. Once all tree-ring series are checked, individual growth curves may be analyzed visually to identify the tree’s reactions to geomorphic processes, for example, the initiation of abrupt growth reduction or releases. In addition, the microscopic appearance of the cells may be investigated to yield further evidence of disturbance. Other features, such as callus tissue overgrowing scars or the presence of TRD formed following cambium damage, can only be identified through a visual inspection of the cores and cross-sections. All growth reactions identified in the samples are noted in order to identify disturbance events. Except for processes with limited volumes (e.g., single rocks or boulders involved in rockfalls), one growth disturbance identified in a single tree is not considered sufficient evidence to identify an event, and therefore the identification of events needs to be based on quantitative or semiquantitative thresholds (Stoffel and Corona 2014).

Conclusions

Initially tree rings were employed in geomorphic studies simply as a dating tool and other environmental information and records of damage contained within the tree were rarely exploited. However, these unique, annually resolved, tree-ring records preserve valuable archives of past Earth-surface processes on timescales of decades to centuries. As many of these processes are significant natural hazards, understanding their distribution, timing, and controls provides valuable information that can assist in the prediction and mitigation of, and defense against, these hazards and their effects on society. This entry has provided illustrations of these themes, demonstrating the application of tree rings to studies of snow avalanches, rockfalls, landslides, floods, earthquakes, wildfires, and several other processes. It illustrates the breadth and diverse applications of contemporary dendrogeomorphology and underlines the potential of expanding dendrogeomorphic research, which could lead to the establishment of a range of techniques and approaches that may become standard practice in the analysis and understanding of Earth-surface processes and related natural hazards in the future.

SEE ALSO: Biogeomorphology; Biologic dating techniques; Dendroclimatology; Disturbance in biogeography; Geomorphic hazards; Hillslopes; Mass movements in periglacial environments; Paleofloods; Riparian ecosystems

References

*Arctic and Alpine Research*, 11: 17–32.


Denmark: Det Kongelige Danske Geografiske Selskab (Royal Danish Geographical Society)

Founded: 1876
Location of headquarters: Copenhagen
Website: www.rdgs.dk
Membership: 350
Secretary General: Henrik Breuning-Madsen
Contact: rdgs@geo.ku.dk

Description and purpose

The society aims to promote knowledge of the Earth and its people and to further interest in geographical science. The society organizes six annual meetings with presentations on geographical topics; publishes a journal, monographs, and atlases; and keeps a library and collection of maps. The society’s activities particularly promote and disseminate Danish geographical research but also form links to other geographical and related societies.

Journals or major publication series


Current activities or projects

The society arranges an annual program of meetings and lectures on geographical themes. It organizes expeditions and research projects and hosted the circumnavigated Galathea 3 expedition in 2006–2007, which visited six of the seven continents. Most recently, the society participated in the interdisciplinary GeoArk initiative in northeast Greenland combining geography, biology, and archaeology.

The society publishes a range of books, an Atlas series, and an international English language journal, Geografisk Tidsskrift–Danish Journal of Geography, which has been the main forum for scholarship on the geography of Denmark and the work of Danish geographers since 1877. The society also maintains a minor museum including books, instruments, and a collection of maps.

Brief history

The Royal Danish Geographical Society was established in 1876 with the dual purpose of disseminating knowledge of the Earth and its inhabitants and promoting interest in geographical sciences.

Up until World War I, the society had an important role in initiating, supporting, and reporting Arctic research, particularly in relation to Greenland. Most of the world’s leading polar researchers were awarded the society’s Gold Medal and gave lectures in Copenhagen.

Exploration remained a leading activity in the time between the two great wars, for example, Knud Rasmussen’s fifth Thule expedition along the entire polar edge of the North American continent to trace Eskimo origins. While Danish expeditions in central Asia (Haslund-Christensen) provided much new information, domestic geomorphology also attracted increasing interest, specifically the dynamics of the Wadden Sea coastal landscape.
DENMARK

After World War II, the society maintained its interest in Arctic research with focus on landscape, soil, and archaeology (Geoark) and it hosted the circumnavigated Galathea 3 expedition in 2006–2007, which visited six of the seven continents.

Since 1876, the society has published Geografisk Tidsskrift–Danish Journal of Geography mainly in Danish until the 1990s when it changed to English. The society publishes atlases and books and arranges at least six public meetings per year.

Submitted by Henrik Breuning-Madsen
Denmark: Geoforum
Danmark (The Danish Association for Geographic Information)

Founded: 2001
Location of headquarters: Copenhagen
Website: www.geoforum.dk
Membership: 250 companies and 400 individual members (as of January 1, 2014)
President: Winn Nielsen
Contact: bf04@tmf.kk.dk

Description and purpose

The purpose of the organization is to provide a network for organizations that work with geography on any level in both the public and private sector as well as to promote the use of geography. The organization’s work is heavily based on volunteers and has a variety of groups where members meet to discuss certain themes and aspects of geography. Furthermore, there are five groups that arrange events throughout the year, including the annual three-day conference, Kortdage, where the entire industry meets to catch up with the latest news.

Journals or major publication series

Geoforum (members’ magazine in Danish).
http://geoforum.dk/udgivelser/geoforum/
Geoforum Perspektiv (journal in Danish and English).
http://geoforum.dk/udgivelser/perspektiv/

Current activities or projects

Professional events throughout the year include courses, networks, and the new Geodataprisen (geodata awards) that celebrates new and innovative products created by the industry. Also, the annual Kortdage – a three day conference in the fall – brings the entire industry together.

The board also works on several products to bring attention to the industry, ensure politics in favor of geography and geodata, and improve the subject of geography on all educational levels.

Geoforum provides two Internet portals to share good examples (in Danish): www.brugstedet.dk is the Geoforum collection of examples on the use of geographic information and www.kortfordig.dk is the Geoforum education portal oriented towards high school students aimed at recruiting them for geodata careers.

Submitted by Jesper Høi Skovdal
Density estimation

Fang Ren
University of Redlands, USA

With significant advances in spatial analytical methods and rapid development of geographic information systems (GIS), density estimation techniques have been widely applied in many disciplines that involve spatially referenced data during the past forty years or so. One important use of density estimation is to convert a set of spatially discrete events, including but not limited to disease cases, crime incidents, road accidents, and animals’ locations into a continuous field or density surface to reveal the spatial distribution and spread of the events. In addition to the presentational advantages, the resulting density may be further tested for statistical inference or aggregated to compare to other socioeconomic and environmental variables.

Density estimation, which is rooted in probability and statistics, seeks to construct a probability density function \( f \) from a sample for an unknown population. Converting discrete data values into smooth curves greatly facilitates data distribution analysis (e.g., skewness), confirmatory statistical analysis such as classification, and pattern detection. There are two groups of estimators: parametric and nonparametric. Unlike parametric estimators, nonparametric estimators are not built upon a presumed population distribution. The simplest nonparametric density estimator is the histogram, in which the probability function is calculated as the proportion of observations contained in each bin divided by the width of the bin. Sophisticated estimators include fixed bandwidth kernel estimators, adaptive (or variable) kernel estimators, the nearest neighbor method, orthogonal series estimators, and restricted maximum likelihood estimators, to name a few (Silverman 1986). Because kernel density estimation (KDE) is the most commonly used method, this entry focuses on KDE and its applications in spatial studies.

Kernel density estimation

Given \( n \) observations in a 2-D geographic space, a general formulation of bivariate kernel estimators is expressed as:

\[
\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left( \frac{d_i}{h} \right)
\]

where \( \hat{f}(x, y) \) is the estimated density at any location \((x, y)\), \( d_i \) is the distance from location \((x, y)\) to the \(i\)-th observation, \( h \) is the bandwidth or smoothing parameter that determines the amount of smoothing, and \( K \) is the weight function that usually is a symmetric probability density function. An illustration is given in Figure 1, where the shape of the “bump” is determined by the choice of the kernel \( K \). The build of the kernel makes itself of particular relevance to the distance decay principle in geography, which implies that spatial interaction declines when distance increases.

The bandwidth of the kernel can be fixed or adaptive. The purpose of using an adaptive kernel is to vary the amount of smoothing based on the local density of data. In a sparse area, a relatively large bandwidth is preferred to include enough observations to avoid under-smoothing (the right kernel in Figure 1), whereas a
DENSITY ESTIMATION

Figure 1 Kernel density estimation with adaptive bandwidth.

A smaller bandwidth is suitable for a dense area to prevent oversmoothing (the left kernel in Figure 1).

There are two ways to implement KDE depending on where the kernel is placed: one is to directly calculate density at each location by putting the kernel on that location (named the site-side estimate), while the other is to center the kernel only at each event and add up the $n$ kernel density functions (called the case-side estimate) (Shi 2010). Although the two methods are mathematically equivalent for fixed bandwidth KDE, the case-side method is more computationally efficient and better suits the scenarios where the distribution of the events varies greatly in space.

It is well known that the bandwidth has a greater effect on KDE results than the choice of the kernel. If the bandwidth is too large, data will be oversmoothed, such that variations in data distribution might be obscured, whereas an incredibly small bandwidth may lead to spurious spikes in the final density surface.

To select an appropriate bandwidth, a variety of selection methods have been proposed, some of which are simple while others are complex. Theoretically, the optimal bandwidth minimizes the discrepancy between the true density distribution and the estimated one, which is commonly measured by mean integrated square error. Unfortunately, both measures depend on the unknown population distribution. To solve this problem, various data-driven strategies have been developed, including least-squares cross-validation, biased cross-validation, “plug-in” methods, smoothed cross-validation, and local polynomial fitting. A comprehensive review of bandwidth selection strategies is available elsewhere (Jones, Morran, and Sheather 1996).

A popular “rule of thumb” of bandwidth selection has been given by Silverman (1986). It assumes the underlying density function is a normal or Gaussian distribution that is the most popular kernel:

$$h_{NR} = 1.06\hat{\sigma}n^{-1/5}$$

where $h_{NR}$ is the normal reference bandwidth, $\hat{\sigma}$ is the estimated population standard deviation, and $n$ is the number of observations. For a set of points, $\hat{\sigma}$ is the standard distance of the point distribution in the horizontal plane.

This computationally simple bandwidth only works well when the population follows a normal distribution. In other cases, however, it is often excessively large and produces over-smoothed densities. To correct for multimodal distributions, Silverman (1986) recommended a slightly different formula:

$$h_{SROT} = 0.9A n^{-1/5}$$

where $h_{SROT}$ is Silverman’s rule of thumb, $A$ is the smaller of the sample standard deviation (or standard distance for a set of points) or sample interquartile range (or interquartile range of distances from each point to the mean center) divided by 1.34, and $n$ is the number of observations. In practice, multiple experiments with different bandwidth values may be useful in determining a reasonable bandwidth for the problem at hand.
Density estimation with spatial data

Due to its strength in data visualization and exploration, density estimation has played an increasingly important role in exploratory spatial data analysis, to reveal spatial patterns of geographic phenomena and their spatial relationships. For example, it is often challenging to identify the location, scale, and orientation of the spatial clusters from a large volume of point data. The overall spatial pattern, however, will become far more noticeable when disconnected points are smoothed into a continuous field. Hence, density estimation has become an effective approach to examining the first-order variation (inconsistent intensity/density) of the underlying spatial processes.

Variants of kernel density estimation

To adapt standard KDE to support the studies that involve spatially referenced data, methodological advances have been made. Brunsdon (1995), for example, developed an adaptive kernel algorithm to determine the probability distribution of a set of point data, which can be applied to points in any dimension in Euclidean space. Diggle (1985) designed the optimal bandwidth that minimizes the mean square error for a set of spatial points, which is widely adopted in epidemiological studies. Edge effects, which often distort point pattern analysis results, have also been corrected in KDE.

It has been found that conventional KDE is not appropriate for mobile objects (e.g., migratory birds) because the points representing a mobile object’s trajectory are not independent. Integrating time geography with traditional KDE, Downs (2010) developed time-geographic density estimation (TGDE) to model the density of moving objects from tracking data. In a typical formulation of TGDE, density is estimated as a function of a geo-ellipse generated for each pair of two consecutive points on the object’s space–time path. Compared to KDE, TGDE excludes the areas that cannot be approached by the animal given its space–time path and the degree of smoothing is determined objectively by the travel velocity of the animal.

For those applications where spatial events only take place over a network (e.g., road accidents or animal roadkill), the planar kernel estimators tend to overestimate the density of the events. To address this issue, Okabe, Satoh, and Sugihara (2009) specifically formulated three kernel functions, namely the “similar” shape, equal-split, and equal-split continuous to accommodate the linear spatial configuration of the events; the latter two are recommended as they produce statistically unbiased density estimates.

Another important advance in kernel density estimation is the development of space–time KDE that estimates density distribution in both space and time. Temporal dynamics is a very important aspect of many geographic phenomena; for example, residential burglaries usually take place in daytime, implying one single static density surface cannot appropriately describe how the intensity of residential burglaries varies throughout a day. Brunsdon, Corcoran, and Higgs (2007) formulated kernel estimators as a function of geographic location (X and Y) and time (T) to estimate the probability of a crime incident at a space–time location (x, y, t).

Although a continuous field is a great way to explore spatial patterns, it does not imply statistical significance that is usually required for pattern detection. The fundamental concern of pattern analysis is to evaluate whether the clusters or areas of high risk are generated by chance or by systematic spatial processes. The Monte Carlo simulation procedure, a commonly
density estimation

used empirical inference approach in spatial
statistics, can be combined with KDE to extract
statistically significant clusters by comparing the
observed density to the simulated densities of
randomly distributed events.

Further, density estimation has been incor-
porated in popular statistical software, such as
R, S-Plus, and SAS, and geographic analysis
enabled software including ArcGIS, CrimStat III,
and STAR (Space-Time of Regional Systems).
In particular, the 3-D rendering capability of the
software provides an appealing 3-D visualization
of density distribution. Overall, the method-
ological advances and widespread availability of
new tools in both statistical and GIS software
have significantly broadened the applications of
density estimation.

Applications of density estimation in spatial
studies

In medical geography, KDE has been applied to
study spatial epidemiology, access to health-care
facilities, and service provision of health-care
facilities. Since Bithell (1990) first demonstrated
that KDE was an effective way to estimate a rela-
tive risk function, KDE has been used to examine
disease distribution and identify associated risk
factors in spatial epidemiology. To evaluate the
impact and spread of a disease, case density by
itself is often of limited use because the spatial
pattern of the disease usually coincides with
population distribution, indicating that disease
density needs to be adjusted for the density of
population at risk or a random sample from the
population (controls). In doing so, the areas of
elevated risk, which otherwise would be con-
cealed due to low population, may be disclosed.
Two measures have been suggested for assessing
relative risk: the ratio of case density over con-
trol density and the relative difference between
these two densities. In addition, KDE enables
the discovery of significant spatial variations in
relative risks of two diseases (Kelsall and Diggles
1995).

One important theme of the public health lit-
erature is concerned with social disparity of geo-
graphic access to health care. The most straight-
forward measure is the provider-to-population
ratio calculated at a predefined scale such as
metropolitan statistical areas. This measure may
be misleading because it assumes that a provider
only serves the individuals in the same area unit.
The barrier of border crossing can be overcome
by a field model, which explains why KDE has
gained broad recognition in many public health
works. Guptill (1975), for example, overlaid a
density surface of physicians with neighborhood
borders to calculate average physician density
for each neighborhood in Detroit, MI. The
density distribution of health-care providers
may be directly compared to that of popu-
lation to reveal underserved neighborhoods
or adjusted by population density to measure
relative accessibility to health care. Inequality
in geographic access to health care among dif-
ferent social groups may be further uncovered
by comparing accessibility densities of these
subgroups.

Parallel to medical geography, spatial point data
representing crimes, persons, and animals have
also been extensively analyzed with KDE and its
variants. In criminological research, KDE and
space–time KDE have become popular methods
for detecting crime-prone zones in space or
in space and time. For example, Chainey and
Ratcliffe (2005) discussed the advantages and
issues of KDE in crime hotspot analysis. In spite
of the issues faced by KDE, such as subjectiv-
ity involved in both bandwidth selection and
gray scale classification in the legend, it is a
very effective way to depict the location, size,
extent, and orientation of crime hotspots. Their
example of comparing robberies with burglaries
further suggests that the spatial structures of two geographic phenomena are different if distinctive smoothing effects are observed with the same bandwidth.

In the context of mapping people, KDE has been applied to examine geographic distributions of customers for marketing and human activity–travel patterns. Using KDE in conjunction with map algebra and 3-D geovisualization techniques, Kwan (2000) compared space–time activity density patterns of different subgroups with a travel diary dataset collected in Portland, OR. For ecological applications, KDE and time-geographic density estimation are found to be useful in defining home ranges of species (Worton 1989). Usually, a density surface is first developed with the tracking data, and then home ranges and core areas are delineated using specified contours of relative intensity (e.g., 50%).

Further, effort has been spent to develop density estimation for spatial phenomena that are restricted over a network, such as road accidents (Flahaut et al. 2003). Identifying road accident hotspots, also called black zones, is critical for developing effective strategies for the reduction of traffic accidents. Compared to the traditional approaches that focus on road segments or junctions, a density surface of accidents, or a relative risk surface if controlled on traffic volume or other factors, can better represent the dangerous spread that, in fact, should not be confined within a well-defined street network boundary.

**Summary**

In summary, the benefits that density estimation brings to the spatial studies are threefold. First, density estimation greatly supports spatial pattern analysis by smoothing discrete point data into different kinds of continuous fields, including density surfaces (e.g., crime or activity densities), risk surfaces when background is controlled (e.g., risk of a disease or relative accessibility to health care), or surfaces of difference when multiple events or groups are compared. Combined with the Monte Carlo simulation procedure, statistical significance can be tested on these resulting density or risk surfaces, with which policymakers can make informative decisions to combat crime, promote accessibility of underserved populations, or target groups for preventive health care.

Second, by depicting discrete geographic phenomena as spatially continuous variables, density estimation successfully circumvents the modifiable areal unit problem and the border-crossing issue that are raised by aggregating individual points into predefined area units. Although administrative boundaries are convenient for managing data, spatial phenomena like diseases or criminal activities do not operate accordingly. For the applications that only involve areal data, such as population counts of census tracts or disease rates of districts, different spatial smoothing techniques, including interpolation and empirical Bayes estimation, need to be applied.

Last, an important use of density or risk surface is to identify risk factors associated with the events of interest. For the problems in which risk factors are spatially continuous variables, such as air pollution or proximity to a contaminated source, the density of these variables can be directly correlated with the risk surface of the events. In other cases where the relevant factors are areal data (e.g., median household income or percentage of a particular ethnic group), the density or risk of the events may be aggregated as a summary statistic like mean or median for an in-depth analysis on relationships.
Density Estimation

See Also: Crime; Exploratory spatial data analysis; Geographic information system; Spatial epidemiology

References


Further reading


Dependency theory

William G. Moseley
Macalester College, USA

In reaction to modernization theory, a new set of development ideas began to arise in the 1950s that emphasized structure, or the global framework under which countries operated. Dependency theory was a leading example of this new type of structuralist thinking. This emphasis on structure meant that the relationships between countries were as important as, or more so than, internal policies for determining the future development of a country.

The earliest ideas about dependency are most closely associated with two economists, Hans Singer and Raul Prebisch, who published papers on the topic in 1949. They had observed that the terms of trade for the poor, tropical countries had deteriorated over time relative to those of the wealthy, industrialized nations. Their idea, known as the Singer–Prebisch thesis, was that poor countries were acquiring fewer and fewer manufactured goods in exchange for the raw materials they exported.

This idea was then elaborated upon by a number of Marxist scholars, one of the most well known being Andre Gunder Frank. Frank (1979) suggested that economies in the tropics (Africa, Latin America, Asia) were essentially “underdeveloped” during the colonial era as European countries refashioned these economies (through a combination of taxation policies and forced coercion) for their own benefit. According to Frank, elites in developing countries often colluded with the colonizers to organize primary extraction in exchange for wealth and power. In many instances, farmers in tropical countries were encouraged or forced to produce commodity crops for the European market (often at the expense of subsistence production) and/or pushed to leave their own farms to pursue wage labor on large plantations or in emerging mining sectors. These primary commodities from the tropics (the satellites) were traded for manufactured goods from Europe and other industrialized countries (the metropoles) in a global system of unequal exchange that favored Europeans. Taylor (1992) described the creation of the modern world economy as being “made by Europeans for Europeans as one great functional region centered on Europe.”

Wallerstein (1979) expanded on dependency theory through his world-systems theory, which basically gave a spatial face to dependency theory by depicting the world in terms of the core, semiperiphery, and periphery. Under this schema, the core countries represent the most developed states that are dominated by industry, financial services, and an information-based economy. The semiperiphery is an emerging group of states where high levels of cheap manufacturing increasingly take place. Finally, the periphery represents those countries whose primary role in the global economy is to provide raw materials. All three of these regions operate as a world system, with deep connections existing between the spheres.

The problem, as dependency theorists see it, is that many tropical countries will find it difficult to break out of their role of producing goods with peripheral processes, including relatively low wages, low levels of technology, and low levels of education. In other words, Rostow’s
DEPENDENCY THEORY

(1959) stages of economic growth will not occur because the least developed countries are locked into an exploitative set of relationships with countries that are more developed. This has led to the idea of an international division of labor, wherein one has unskilled laborers working on the periphery and a skilled workforce in the core countries (Baran 1957). Given this division, highly skilled workers in the periphery will be tempted to migrate to the core countries (a problem known as brain drain) because employment opportunities for them are limited in their own countries.

Dependency theory was also influential at the subnational scale via the concept of dualism. Influenced by notions of dependency, dualism refers to situations where two areas are in relationship with one another, and one area is developing at the expense of the other. This concept may be examined at a variety of different scales, including global (as is the case with dependency theory), regional (within country), urban, and rural.

Within most countries, dualisms often exist between urban and rural areas. In the United States, for example, many would argue that core areas exist where a disproportionate amount of investment and wealth accumulate. These areas, typically cities, have historical relationships with more rural zones of resource extraction that function as peripheries to the urban cores. In Nature’s Metropolis (1992), environmental historian and geographer William Cronon wrote of the relationship between Chicago and vast interior regions of the Midwest. Timber, iron ore, livestock, and grain historically came to Chicago from the Upper Midwest in exchange for manufactured items. Cronon explained that while Chicago became fabulously wealthy during the nineteenth and twentieth centuries, the regions of extraction (such as northern Minnesota, southern Illinois, and northern Wisconsin) have little to show for their exports today.

Dualism is apparent in rural areas where one may have an export-oriented commercial agricultural sector that exists alongside smaller-scale, mostly subsistence farming. The classic situation is one wherein small subsistence farmers spend the day laboring on nearby large plantations, returning to work on their own farms in the evenings and during the weekends. The linkage between the two sectors is often disadvantageous for the small farmer or laborer. For commercial farms, the advantage of laborers who moonlight as small farmers is that they need not be paid a living wage because they have their own production to cover a portion of their annual food needs. Many small farmers who end up working on larger farms may initially take on such employment in order to bridge a food production shortfall. In so doing, however, they often embark on a slippery slope of declining production as they spend critical time off the farm compromising their own ability to produce food.

A similar set of dualistic relationships exists in urban areas between the formal and nonformal employment sectors. Formal sector employers, such as large export-oriented firms, may draw on temporary laborers who are otherwise employed in the informal sector, including a wide array of undocumented full- and part-time employment in activities such as petty commerce and artisanal production. This system allows for the relatively cheap supply of labor to the formal sector, as these laborers may also rely on the informal sector for income.

Policy outcomes

At the international scale, dependency theory did lead to some real policy changes in developing
countries. Probably the most significant of these was an approach known as “important substitution industrialization,” often associated with the Argentine economist Raúl Prebisch previously mentioned. Prebisch (1959) argued that tropical countries would be forever stuck as producers of primary products (and therefore fail to develop) unless they took proactive steps to change the nature of their economies in relation to those of others. The idea behind import substitution was that manufactured goods needed to be produced at home rather than imported from the core countries (hence the name of “import substitution” for such policies). Given a lack of private capital available for industrialization, and stiff competition from producers in the core countries, the governments of peripheral states often became directly involved in the creation of such enterprises. These state-run enterprises came to be known as parastatals. Governments typically also erected tariff barriers to protect such nascent or infant industries until they could stand up to international competition.

The other policy, at least partially inspired by dependency theory, was the nationalization of some industries, often mines, that were owned and run by transnational corporations (TNCs) in tropical countries. The argument was that the foreign companies that ran these industries would not manage them in the interest of local people and development. While capital accrued from mining, for example, might normally be reinvested in industry and lead to the development of the local economy, the problem was that TNCs repatriated most of the profits. By taking over such industries (and running them as parastatals), the state could, in theory, insure that the profits from these mines were plowed back into developing other industries in the country. Good examples of countries that did this are Chile and Zambia, who nationalized their copper mines in the 1960s and 1970s.

Import substitution, and to a lesser extent nationalization, was quite popular until the late 1970s when the “third world” debt crisis struck. This crisis involved a number of peripheral countries that were on the verge of defaulting on loans owed to private commercial banks as well as public creditors. The crisis was largely brought on by government involvement with increasingly inefficient, state-run enterprises and the energy crisis of the 1970s when high oil prices were particularly challenging for oil importers.

Neoliberal and poststructural critiques of dependency theory

Dependency theory, and structuralist thinking more broadly, began to face two very different sorts of critiques in the 1980s. The first type of critique came from neoliberal economists who argued that dependency theory-inspired import substitution policies were inhibiting economic growth. These policies, they suggested, were largely responsible for the “third world” debt crisis of the late 1970s. Neoliberal reforms were initially developed and undertaken in Latin America, led by economists from the University of Chicago known as the “Chicago Boys.” These polices would then spread to other parts of the world. For example, the so-called Berg report (World Bank 1981) offered a searing critique of import substitution policies in Africa and proffered neoliberal reforms, known as structural adjustment, as the solution. This package of reforms called for reduced government, free trade, export orientation, and the privatization of state-run enterprises.

Critics of neoliberalism (largely supporters of dependency theory) suggested that this was a return to the economic policies of the colonial era (where tropical countries were encouraged
to focus on the export of raw materials). Some subsequent work in Asia (e.g., Bruton 1998) has shown that import substitution policies were critical for developing the export-oriented industries that led to the rise of the Asian Tigers (Hong Kong, Singapore, South Korea, and Taiwan). Ironically, it was the export orientation of these countries that was used as evidence in support of neoliberal economic reform (conveniently leaving out the policies that helped build these industries in the first place).

Another type of critique of dependency theory came from the poststructuralists, a group of social scientists influenced by postmodernism. This group argued that structuralist thinking (to which dependency theory is central) did not give adequate attention to “agency,” that is, the power of individuals and groups to change their circumstances through individual or collective action. This group also asserted that greater attention needed to be accorded to discourse as a form of power and influence. While the discourse on any given subject is often controlled by elite interests, and the masses have often internalized these narratives (the definition of hegemony), collective action could also lead to the production of counternarratives, or a new conventional wisdom. In many ways, the poststructuralists were not abandoning the insights of dependency theory, but they wanted to build on it to account for power of agency and discourse. Good examples of such poststructuralist thinkers include Peet and Watts (1996) who wrote about the emancipatory potential of poststructuralist thinking when applied to an older, more structuralist leaning, political ecology (a subfield of human–environment geography). The poststructuralist, feminist economic geographers J.K. Gibson-Graham also write eloquently about the limitations of dependency theory. In The End of Capitalism (As We Know It): A Feminist Critique of Political Economy (Gibson-Graham 1996), they write about how we imagine capitalism to be more powerful and ubiquitous than it is in reality. Once we begin to realize all of the noncapitalist activity around us, we can rewrite or reimagine the dominant narrative and feel empowered to act differently (or express our agency). For them, feminist scholarship is critical for recognizing the heterogeneous nature of the economy.

More recent debates, referencing ideas and intellectual offshoots of dependency theory, include claims for new regional growth in Africa, or the “Africa Rising” assertion, based on natural resource-based exports (compare Roxburgh et al. 2010 to Carmody 2011 or Moseley 2014). The question is whether this recent growth represents something new or simply a return to dependency. The notion of path-dependent growth is now also widely debated in the social sciences (Martin 2010).

SEE ALSO: Colonialism, decolonization, and neocolonialism; Development; Modernization theory; Political ecology; Poststructuralism/poststructural geographies; World-systems theory

References


Further reading


Desakota

T.G. McGee

University of British Columbia, Canada

Desakota is a neologism developed after fieldwork in the Jakarta mega-urban region in the 1980s and discussion with Indonesian colleagues. (McGee 1991). The term was coined to recognize the fact that rural areas adjacent to the built-up central areas of cities were being subjected to increasing urban expansion. This process was creating areas of mixed urban and rural activity that were increasing in territorial size and population. The term desakota combines two words from the Indonesian language, desa (village) and town (kota); it captures the ideas of rurality and urbanity that have existed in Indonesia for many centuries. These areas were growing rapidly as the result of urban spatial expansion that accelerated in the large mega-urban regions of Asia from 1970 onwards.

In the period since the 1980s, these processes of urban expansion have accelerated in Asia, particularly in the faster growing economies of the region, often leap-frogging along urban corridors beyond the immediate hinterlands of the mega-urban regions to secondary urban nodes that were developed to link new industrial areas, residential estates, and retail complexes to other parts of the mega-urban regions. This process has been facilitated by the growth of digital communication that facilitates networks which link these activities to core cities in the mega-urban region, nationally, and globally. The rapid increase in the price of land in the urban cores also encouraged decentralization to the cheaper land markets located on the edge of these cities. These developments are reshaping the urban space of the mega-urban regions for which the models of the organization of urban space that have been developed in the past for developed countries are no longer valid; they are replaced by the concept of mixed urban land-uses in a form of hybridity that reflects the collapse of space and time in which these extended urban spaces are linked by networks of flows of commodities, people, capital, and information that facilitate urban expansion. Today these desakota regions are a kind of “in between” in the “rural–urban dichotomy” that have become a prevailing form of land use. Indeed, national space is now a form of rural–urban hybridity. This does mean that “rurality” and “urbanity” as concepts do not capture reality and people’s beliefs and actions, but these are still all subject to cultural, historical, and economic variations at a global, national, and local level. In the contemporary era, these developments have been complicated by the new challenges of environmental change and deterioration and rapid but uneven economic growth in many of the developing countries of Asia (McGee 2011). These challenges will continue in the next thirty years with almost 60% of all global urban growth occurring between 2010 and 2050 in Asia, much of it located in the extended urban space of large mega-urban regions. This suggests that ideas that inform the study of the organization of urban space need to be critically evaluated in this new era of “planetary urbanization” (Brenner 2014).

SEE ALSO: Cities and development; Globalization; Megacity; Rural/urban divide


Desert ecosystems

Taly Dawn Drezner
York University, Canada

The world’s deserts have a unique assemblage of species, and are not the lifeless zones often envisioned. Nearly one third of the Earth’s land surface is identified as arid or semi-arid, and perhaps 15% of the world’s land areas are true desert. Characterized by low precipitation, high evapotranspiration rates, little cloud cover, high summer temperatures, and large annual and diurnal (day-night) temperature ranges, desert climates are also highly variable in rainfall receipt from year to year, and in seasonality of rainfall. Winter conditions can vary from mild in lower latitude deserts to higher latitude deserts that regularly see subfreezing winter temperatures and sometimes snowfall, such as the Gobi in central Asia. Only the coastal deserts associated with cold ocean currents (such as the Atacama and the Namib) deviate from this scheme. These deserts typically experience smaller temperature ranges, milder summer temperatures, and higher humidities associated with regular fog. Desert organisms have adaptations to cope with the extreme temperatures and limited moisture of desert environments. The focus of this review is on traditional “hot” deserts, excluding the polar deserts.

The pulse-reserve paradigm

Perhaps one of the most central theories on desert function is the pulse-reserve paradigm, proposed by Noy-Meir (1973), along with ideas by Westoby, Bridges, and colleagues credited within Noy-Meir (1973). Noy-Meir (1973) proposed that desert ecosystems are controlled by the unpredictable availability of rainfall and water, and thus behave in a pulse and reserve manner. This view suggests that a rain event produces an influx of the controlling resource (water) and triggers a pulse of production (Noy-Meir 1973). While some pulse-driven resources are later lost, such as to mortality of newly germinated seedlings, or by consumption through grazing of new biomass produced from the rain event, some of the pulse goes into carbon and energy reserves including seeds (Noy-Meir 1973). Noy-Meir (1973) develops these ideas further, such as relationships between the pulse and depth of water penetration, root systems, and differences across taxa, as well as impacts of soil texture, salinity, and the timing of rainfall events on the pulse. He even extends these basic ideas to higher trophic levels as available forage, water, and prey for animals will also depend on inputs of water both in the open (i.e., standing water) and in potential food (Noy-Meir 1973).

Reynolds et al. (2004) update and modify these ideas, such as consideration of soil moisture conditions prior to a rain event, soil water recharge, rain event size, and total seasonal rainfall. Noy-Meir (1973) discusses soil moisture, and Reynolds et al. (2004) observe that soil moisture is a better predictor of plant uptake and use than rainfall. They suggest that smaller rain events may only be biologically important when falling in conjunction with a larger storm event (rain falling in temporal isolation may be only minimally beneficial biologically), and that...
DESERT ECOSYSTEMS

Rainfall itself does not fall randomly over time, but rather, tends to be somewhat clustered over a season (Reynolds et al. 2004). Further, different plant functional types exhibit different responses to soil moisture and precipitation, and seasonality of precipitation (e.g., winter rain evaporates more slowly) may also be important, depending on the rainfall regime of the particular desert (Reynolds et al. 2004).

Primary production and food chains

Primary production and amount of biomass created by plants are strongly linked to rainfall in desert ecosystems. Primary production in arid environments is low in comparison with more humid ecosystems, but there is considerable variability in primary production within deserts, with primary production sometimes rivaling that of more humid temperate ecosystems (Whitford 2002). Deserts are also home to localized mesic environments and associated vegetation, such as riparian areas near perennial streams, and communities associated with springs and oases.

There has been some disagreement among researchers about desert food webs and linkages within the ecosystem. For example, some studies have suggested the desert ecosystems have relatively few links while others have observed many more (Ward 2009). Other points of disagreement include, for example, whether two species preying on each other (loops) are rare or not (Ward 2009). Generally, trophic level interactions and food chains are far more complex than often recognized. In some cases desert consumers may be subsidized by the close proximity of another community, such as in a coastal desert where food sources and prey species become available from the nearby marine ecosystem. Ward (2009) provides a succinct review of some of these issues.

Animal adaptations

Animals have a wide range of adaptations to the heat and aridity that typify desert environments, and in some deserts, also adaptations to protect them from cold winter temperatures. Adaptations to heat and aridity include rapid uptake of water and mechanisms to reduce water loss, to tolerate and/or escape high temperatures, and to reduce metabolism (Whitford 2002). The examples below hint at some of this variety.

Many insects, amphibians, and mammals burrow as temperatures may be 30°C cooler than at the surface. Many insects, spiders, and scorpions remain in burrows for much of their lives. Some species remain in their burrows until a relatively large rain event, when they emerge, eat, and mate before returning, such as amphibians who may use ephemeral ponds for breeding. Some toad species can remain in their burrows for more than two years without eating, typically with a highly reduced metabolism. Even snails are found in deserts, sometimes dormant during the hot months with increases in activity following rain events.

In southern Africa, the Cape ground squirrel (Xerus inauris) can lower its operative environmental temperature by 8°C by shading itself with its own tail and aligning itself relative to the sun’s position to maximize shading, which can double the number of hours it can spend foraging during the day. Kalahari lizard species maintain a nearly constant body temperature by alternating between sunny and shady spots over the course of the day, while some insects and lizards move around to different parts of a plant to find the most suitable shade and temperature conditions (Whitford 2002). Many insects and rodents come to the surface to seek food at night. Nocturnal activity is found in many mammals including commonly in small mammals, and a variety of insects, spiders, and scorpions. Even insects that
are not nocturnal may limit their activity during the hottest times of the day, and some large mammals (e.g., the Arabian oryx) may only seek food at night during the hot months.

Many small mammals in deserts are known for concentrating their urine to reduce water loss. Some amphibians secrete and cover themselves with wax to minimize water loss. In some cold current coastal deserts commonly characterized by fog, some insect and lizard species have adaptations to condense and capture fog water onto their backs and then to drink that water as it drains toward their mouths.

Perhaps the most famous desert residents are camels, which originated in North America 40–45 million years ago and include the dromedary camel (one hump, Middle Eastern deserts) and the Bactrian camel (two humps, central Asian deserts); their ancestors crossed the Bering Land Bridge into Asia about 3 million years ago. Rather than having a more even distribution of fat over their bodies, which would act as an insulating layer, camels have humps that are fat reserves concentrated in one place. Camels can go without water for two weeks, and can store that water initially in their gut, thus enabling them to consume large quantities at once without overhydrating. The camel has a wide variety of other adaptations, including highly concentrated urine, the ability to withstand elevated body temperatures of several degrees, and a specialized circulatory and breathing system that allows the brain to be kept at a cooler temperature than the rest of the animal, as has been found in other large desert mammals (e.g., eland, wildebeest). Camels and other arid land animals (e.g., ostriches) have long legs and necks that help with cooling.

There has been some debate about bird adaptations in deserts (Williams and Tieleman 2005). Research shows that metabolism is lower in desert birds, some species lay fewer eggs, and evaporative water loss is lower in desert birds than in other birds (Williams and Tieleman 2005). Desert-dwelling birds typically seek microhabitats and conditions that minimize their exposure to extreme temperatures which includes selecting shaded and sheltered nesting sites and foraging in the morning and evening. Some birds have glands that secrete salt, including the ostrich and roadrunner. A variety of birds pant to aid in cooling themselves and their eggs (Ward 2009).

Plant adaptations

Plants exhibit numerous adaptations to the high temperatures and lack of water that are typical of desert environments. Many species avoid the extreme conditions by remaining in the soil as seeds and only germinating during favorable periods, followed by a short life cycle and a return of seeds to the seed bank. Seed dormancy is discussed further below. Other species are drought tolerators, which can desiccate to extreme levels, such as the resurrection plant (*Selaginella lepidophylla*) in the Chihuahuan Desert.

Root adaptations include extensive root systems that maximize capture of surface water, deep roots that can tap to a water table or to soils that retain their moisture longer, and the production of rain roots within hours of a rain event to increase water uptake. Some plants (called halophytes) are adapted to soils with high salt concentrations. In some cases, these species do not establish in non-saline areas where they may have to compete with other species.

Leaf adaptations include fewer stomata (pores through which gas exchange and water loss occur), and their greater distribution on the shaded underside of leaves. Some species have waxy leaf surfaces, while others have pubescent leaves (e.g., with hairs, such as *Encelia farinosa*), or have lighter colored leaves (e.g., *Ambrosia dumosa*) to increase albedo (reflectance) and reduce leaf
**DESERT ECOSYSTEMS**

Temperatures and sometimes to reduce photosynthesis (Whitford 2002). In foggy deserts, some species have many narrow leaves that promote water interception and capture from fog. Some perennial plants are drought-deciduous, dropping their leaves during periods of reduced soil water, in some cases reclaiming nutrients locked up in the leaves prior to dropping them (resorption) (Whitford 2002), and some plants orient their leaves to minimize insolation during the hottest times of the day. Similar adaptations have been noted in cacti, such as prickly pears (*Opuntia* spp.) that orient their pads to minimize exposure during the hottest times of the day, while other cacti tilt equatorward to minimize receipt of solar radiation during the hottest times of the day. Other leaf adaptations include a reduction of leaf size to reduce water loss; cacti are an extreme example of this, where leaves have reduced to spines. While spines provide protection from animals, they also cast shade reducing the plant’s surface temperature, and can elevate nighttime stem temperatures by trapping heat and sometimes expanding the species’ range into cooler areas (Anderson 2001), and spines reduce surface area exposed to PAR (photosynthetically active radiation) reducing photosynthesis and growth (and water loss), among other benefits (Anderson 2001). Spines can even aid in the transport of stem segments to a new location for some species where a new plant can establish vegetatively.

Succulence, the storing of water in leaves, stems, and/or roots, is often associated with desert plants, along with CAM (Crassulacean acid metabolism) photosynthesis. While about 90% of the world’s flowering plant species engage in the traditional C₃ photosynthetic pathway, some species utilize the CAM or C₄ photosynthetic pathways. Unlike C₃ plants where stomata are open and gas exchange occurs during the day, in the CAM pathway, stomata are closed during the day and CO₂ uptake occurs primarily at night (Nobel 1988) when tissue temperatures are lower, thereby reducing water loss. As a result, the plant's ability to store carbon dioxide at night can limit photosynthesis (Nobel 1988), and thus the trade-off for this more efficient use of water is slow growth, typical of CAM plants such as cacti. When ample water is available, CAM plants can switch to C₃ photosynthesis (Nobel 1988). The C₄ pathway also improves water efficiency but is more common among monocots and grasses. Nobel (1988) reviews these pathways.

Many of the tall columnar cacti (such as the famous branched saguaro cactus of the American southwest) have an accordion-like external design of ridges and furrows. The cacti’s shallow and extensive roots help them quickly absorb rainwater, which they store by expanding the accordion-like folds. This process also regulates receipt of solar energy for photosynthesis with furrow exposure increasing the photosynthesizing surface area of the plant that increases with greater water in storage.

Cacti are perhaps the plants most associated with deserts, though cacti are native only to the Americas (one epiphytic species was probably carried by birds to tropical Africa) (Anderson 2001). Thus the Sahara and Middle Eastern deserts, the Australian deserts, and all the Old World deserts have no native cacti. Today, some introduced species persist and some are cultivated. The cactus family is also not limited to deserts in their natural range, though their richness (number of species) is relatively high in the North American deserts, particularly in Mexico, and in arid and semi-arid areas in South America (Anderson 2001). The cold hardy *Opuntia fragilis*, for example, reaches the boreal forests of Canada and northward, while other cacti are found to 50°S latitude in southern Argentina. Some cacti are found in the humid,
temperate regions of eastern North America and many species are native to the tropics (Anderson 2001). These species may be epiphytic in more humid areas, while cold-tolerant species such as those living in Canada experience acclimatization to cold as well as lower stature to increase coverage by snow which helps protect plants from extreme air temperatures, among other adaptations. Caution needs to be used to not mistake cacti with the more general characteristic of succulence, which has evolved independently in multiple groups and includes Aloe vera, some Euphorbia (spurges) species, which also include leafy, non-succulent members such as poinsettia (Euphorbia pulcherrima), and other species.

Seed dispersal in space and time

Most plants worldwide have adaptations for dispersing their seeds by animals (e.g., fleshy fruits) and abiotically (e.g., “parachutes” to carry seeds in the wind like dandelions). The benefits of dispersing seeds away from the parent plant are many, including colonizing new areas, reducing competition with siblings or other conspecifics (= of the same species), reducing reproduction with closely related individuals, and general competition for space and nutrients (Van Rheede van Oudtshoorn and Van Rooyen 1999). While dispersal away from a parent dominates in more mesic areas, this is not necessarily the case in arid environments. Ample open space eliminates space competition as a driver of dispersal, and establishment sites are not as widespread as in a mesic environment, often limited to favorable microsites with greater water or other key resources. Further, with a substantial portion of desert florals being annual plants (which would thus die before the next generation germinated and established), the best strategy may be to leave seeds at the mother site, which is necessarily hospitable as the mother plant has already been successful there (Van Rheede van Oudtshoorn and Van Rooyen 1999). Atelechory (absence of features on the diaspore to facilitate long distance dispersal) and antitelechory (features, structures, or mechanisms that hinder dispersal away from the parent plant) have been documented in deserts and soon the “mother-site” theory emerged. Ellner and Shmida (1981) question this theory as atelechorous and antitelechorous annuals are widespread across many desert areas, raising doubts about limited favorable sites, and further, many perennial species are atelechorous or antitelechorous, raising questions about these perceived adaptations in annuals (where the parent plant would not be a competitor in subsequent years). Antitelechorous species that are amphicarpic and geocarpic (produce some or all, respectively, of their seeds below the soil surface) are better represented in more mesic areas nearby (Ellner and Shmida 1981). The mother-site theory would also suggest continuity in species composition over time at microsites, but Ellner and Shmida (1981) did not find this to be the case. They propose that many of the atelechorous and antitelechorous adaptations observed are not so much for dispersal over space (or lack thereof), but for dispersal over time as water availability is more important than competition for space, and plants need to take advantage of periodic wet years. Mortality is often density-independent so adaptations to dispersal may not be needed (atelechory) (Ellner and Shmida 1981). Antitelechory adaptations such as burial can improve water uptake and shelter seeds from rodents and other seed predators, and some structures may anchor seeds to avoid being washed downstream during heavy rains, rather than simply acting to hinder dispersal as previously suggested (Ellner and Shmida 1981).

Seed dormancy is a common bet-hedging mechanism in deserts whereby germination
DESERT ECOSYSTEMS

occurs over multiple years; thus one devastating
year (such as early rains that trigger germination
that are followed by little to no rain resulting
in widespread mortality of seedlings) does not
leave the soil devoid of seeds, particularly crucial
for annuals that do not have living adults grow-
ing in the vicinity to replenish the seed bank
(Venable 2007). A certain fraction of the seeds
from a cohort delay germination to subsequent
years. The fraction of the seeds produced that
are dormant is variable (e.g., Venable 2007),
including a lower germination fraction with
increasing risk of reproductive failure. Typically,
the more unpredictable and arid the desert,
the higher the proportion of the flora that is
annual (e.g., Van Rheede van Oudtshoorn and
Van Rooyen 1999). Because of their dynamics,
annual plants are more prone to local extirpation
than perennials, as well as to invasion.

Invasive species

It has been suggested that deserts are not invaded
as successfully as other biomes as a smaller
portion of the desert flora is typically made up
of non-natives. Those non-native species that
are found in deserts are often unoffending to
the native biota. The relative lack of success of
invasion is thought to be related to the extrem-
ities of the climate and environment (Lambert,
D’Antonio, and Dudley 2010). In fact, more
mesic species from surrounding ecosystems that
invade the desert often establish in those more
mesic microsites, as many species, including
trees, typically do not succeed in establishing
in the desert because the environment is too
harsh (Ward 2009 and references therein). Not
surprisingly, some of the greatest threats to desert
biotas worldwide occur in more mesic areas,
such as along riparian corridors. Non-native
Tamarix (tamarisk) species negatively impact
native ecosystems in a variety of ways including
lowering water tables, decreasing faunal diversity,
and outcompeting many native species such as
cottonwood (Populus spp.) and willow (Salix
spp.) and reducing their presence along streams
(Ward 2009). Some of the rarest forest types in
North America have been invaded by Tamarix.

Invasive species, including grasses, such as the
problematic buffelgrass (Pennisetum ciliare), and
herbs such as Sahara mustard (Brassica tournefortii)
have outcompeted native species, and also altered
the native fire regime causing further damage
to the desert ecosystem (Lambert, D’Antonio,
and Dudley 2010). Buffelgrass contributes to the
production of combustible ground cover and
thus to fire, and causes native species to decline.
Many other species have become problematic in
a variety of arid and semi-arid areas worldwide,
such as members of the Opuntia cactus genus.
Today, Opuntia ficus-indica, Opuntia stricta,
and others are invasive and well established over large
areas in Australia, the southern African Karoo,
and the Middle East. Non-native and invasive
animal species have also established in deserts,
including the European starling (Sturnus vulgaris)
and cane toads (Rhinella marina) in Australia, a
highly toxic species that has caused declines in
larger native predators that eat the toads.

Convergent evolution

While some species or closely related species
have dispersed to many parts of the world, species
inhabiting the different deserts of the world are
typically not related, despite their similarity in
appearance (e.g., succulence, small light-colored
leaves) and that they may fill similar niches. The
nearest relatives of most desert plants are from
the more humid areas in the surrounding regions
(Ward 2009). Certain traits confer advantages
for survival and reproduction; thus desert species
have adapted to a similar suite of conditions, and as a result they have apparently similar features (adaptations). Environmental conditions are an important driving force in the evolution of continental species (Melville, Harmon, and Losos 2006). This similarity of form, ecology, and/or function (such as succulence or small leaves) that evolves independently in unrelated groups in response to similar environmental conditions is termed convergent evolution. For example, the succulent aloes (*Agave* spp.) are closely related to lilies but only distantly related to cacti. In areas where fog is common, narrow leaves have been observed convergently for water capture in unrelated species.

These parallels occur across both plants and animals. Ungulates are lighter in color in desert environments, likely related to camouflage and concealment. Unrelated lizards in North America and in Australia exhibit similar adaptations (e.g., water collection on their bodies), and convergence has been documented in many lizard species (e.g., Melville, Harmon, and Losos 2006). A classic example of convergence is that of desert rodents (and a marsupial) in form, who share many similarities such as bipedal hopping, tufted tails, and various aspects of coloring, as well as urine concentration which is well studied in small mammals and is convergently found in a South America marsupial “mouse” as well. However, no two species ever converge completely and there are also differences in all convergence examples as no two niches are exactly the same. In both plants and animals, differences in form and function persist.

**Keystone species**

Keystone species are those that play a disproportionately important role in the ecosystem and food web and, if they were to become extinct, would likely affect the survival and reproduction of other species to the point where some may become extinct by the loss of the keystone species. Keystone species are found across a variety of plant and animal groups. Small rodents have been identified as being important, as have arthropods, particularly ants and termites. Ants and termites may fill niches typically filled by other taxa, and have substantial impacts on the structure and function of desert systems. Termites and ants affect nutrient cycling, plant production, and a variety of soil properties. The saguaro (*Carnegiea gigantea*) and other columnar cacti found in the southwestern US and Mexican deserts are keystone species (Drezner 2014). A variety of leguminous trees have been identified as central species in deserts worldwide including the Sahara, Israeli, Australian, and North American deserts, sometimes by increasing soil fertility while in some cases simply through microclimate and microsite modification as nurse plants (see below). For example, in the Sonoran Desert, the leguminous tree *Olneya tesota* acts as a nurse to over 200 plant species and over 250 animal species are known to utilize the plant; in some cases, *Olneya tesota* nurses more species beneath its canopy than are found outside of it (Burquez and Quintana 1994).

**Nurse associations and the stress gradient hypothesis**

A nurse plant is one that creates a beneficial environment under its canopy or in its proximity, thus promoting the survival of other species near or under it. Extreme heat and cold temperature amelioration, protection from soil surface temperatures of over 70°C, and shelter from herbivores are common benefits, and some nurse species create fertile islands with increased soil nutrients, among other benefits.
Protégé species often produce fruits that are consumed, carried, and deposited by animals to shaded areas such as under potential nurse canopies, including birds that may defecate the seeds while perching. It has been widely documented that nurses are often woody species, though some disagreement has arisen about whether this may be the result of greater survivorship in such areas (i.e., a true facilitative effect), or whether protégé distributions are simply a reflection of the nonrandom seed dispersal by animals to such microsites (e.g., Sosa and Fleming 2002). It is possible that survivorship is higher in such areas due to reduced seed and seedling predation under nurse cover, though for some species, the abiotic climatic benefits are important, as reviewed in Sosa and Fleming (2002). Many species are nurse plants and many species are protégés in arid environments, although the reasons for the large number of such associations in these environments are still debated. Some species simply do not survive without a nurse (e.g., Carnegiea gigantea) while for other species, such associations are beneficial but not essential.

For the nurse, disadvantages of such associations range from none, particularly when protégé individuals are very small, to many. The protégé cactus Carnegiea gigantea has very shallow roots that are laterally extensive, thus sitting above its nurse’s main root system and capturing rainwater percolating down through the soil ahead of its nurse. Carnegiea effectively compete with their nurse for rainwater, and their towering size as adults enables them to take up and store large amounts of water, sometimes resulting in the premature death of the nurse. It has been suggested that a replacement cycle may occur where the protégé causes early mortality of its nurse thereby reducing nurse presence and thus potential nurse establishment sites for future protégé individuals, resulting in protégé population decline, which leads to an increase in nurse establishment, driving up protégé recruitment, and so on. However, this relationship would operate at relatively long time scales, and since many protégés are successful under a variety of different nurse species, any such dynamics are complicated. In some cases, the nurse may benefit from the presence of a protégé. For example, the grass Hilaria rigida nurses several cholla cactus species (Cylindropuntia spp.) (Cody 1993). These cacti have easily detaching jointed segments that are both difficult and painful to remove from skin, and many desert animals avoid these plants. Thus the protégé cactus provides its nurse grass natural protection from grazing (Cody 1993).

There are also drawbacks to the protégé species for establishing in close proximity to another plant. These include reduced sunlight and PAR which can result in slowed growth, and root competition not only with a nurse, but also with other plants (the same and different species) similarly clustered beneath the shared nurse. However, the extreme environmental conditions typifying desert habitats make any such disadvantages secondary; that is, the benefits (and sometimes requirements) conferred by a nurse may be so important that they outweigh the disadvantages to the protégé with regards to its survival and reproduction. It has been suggested that in communities where climate is less extreme, biotic factors are limiting and competition becomes the dominant interaction; conversely, places with more extreme abiotic conditions (e.g., high temperatures and aridity as found in the desert) limit species more than does competition (Bertness and Callaway 1994). The stress gradient hypothesis suggests that in more abiotically harsh environments, facilitation and positive plant interactions should be relatively common in comparison with less extreme environments, where competition would be far more important in shaping communities (Bertness and
A variety of studies have been done both debating and modifying the stress gradient hypothesis.

Deserts are diverse regions that are home to numerous endemic species that are unique to those regions. A wide variety of adaptations have evolved to deal with irregular and often unpredictable rainfall and inputs of water, as well as with the temperature and other environmental extremes. Many theories about desert function have been developed, including the influence of rainfall, and on positive interactions between species to assist in survival in such abiotically extreme environments.

**SEE ALSO:** Biomes; Desertification; Ecoregions; Plant–water interactions

**References**


Desertification is as much a powerful idea as it is a biophysical process. Indeed, it is the idea of desertification, rather than the processes of land degradation in arid regions, that has been more influential over the last two centuries. Many millions of dollars have been spent, and many people’s lives disrupted, around the world in order to combat “desertification” in places where dryland ecologies have been largely misunderstood and degradation has actually been limited. First coined by the French colonial forester Louis Lavauden in 1927, the word desertification has long been associated with notions of spreading deserts and of human culpability for deforestation, overgrazing, and land degradation. Lavauden, following an old Anglo-European tradition of thinking about arid lands and indigenous people, wrote that “desertification is uniquely the act of humans … [T]he nomad has created what we call the pseudo-desert zone” (Lavauden 1927, 267). From its very first use, then, the word desertification has been ideologically and politically laden with many preconceptions and misconceptions that led to the dominant view during the twentieth century that deserts are spreading and that desertification is primarily the result of human mismanagement.

Contemporary research in arid lands ecology, however, has reached a broad consensus that the extent and severity of desertification has been greatly exaggerated (Helldén and Tottrup 2012; Herrmann and Hutchinson 2005; Nicholson 2011; Reynolds and Stafford Smith 2002; Reynolds, Stafford Smith, and Lambin 2007; Thomas 1997). After years of what have come to be called the “desertification debates,” a great majority of scholars now agree that nonequilibrium dynamics dominate in arid lands around the world, that documented cases of severe arid lands degradation are relatively few, and that they are found in discrete locations associated with particular social, political, or economic problems. The limited number of cases of serious dryland degradation are found primarily in places under the strong influence of political economic forces such as capitalist expansion (US southwest), authoritarian state rule (China, USSR) or the developmentalist state (Egypt). That is to say, degradation in arid lands has certainly taken place in some areas for certain reasons, in particular from over-irrigation and plowing marginal soils, but the common claim of the global crisis of desertification has repeatedly been demonstrated to be false. In fact, a recent “greening” has been detected in many arid regions over the last two decades (Dardel et al. 2014; Helldén and Tottrup 2012; Nicholson 2011).
Most deserts are not continually expanding but, rather, they grow and shrink according to rainfall patterns more than any other single factor, and have over the last several millennia. Indeed, as noted by the geographer Andrew Goudie, many deserts including the Sahara, the Namib, the Kalahari, and the Gobi have been more or less where they are today since the beginning of the Tertiary, about 65 million years ago, shrinking and expanding with global climate conditions (Goudie 1986). Deserts, according to most scientists, are the arid and hyper arid regions covering about 20% of the globe that receive very little rainfall of a highly unpredictable nature (Laity 2008). There is now little debate that these are naturally occurring physical features. Another 20% to 25% of the Earth’s surface is variously categorized as semiarid and together deserts and semiarid areas are often called the “drylands.” Although usually thought of as sparsely populated, the drylands are estimated to contain approximately 38% of the world’s population (Reynolds, Stafford Smith, and Lambin 2007). Most of the debate about desertification concerns the semiarid zones and fears that they will become arid deserts, leaving many of these people without a means of livelihood or subsistence. Located primarily in the low humidity, high pressure zones of 30° north and south latitude, dryland environments are highly variable, ranging from sand seas to stony pavements to arid mountains to desert crusts and varnishes to steppelands, grasslands, shrublands, and savannahs.

The majority of arid and semiarid regions contain vegetable and animal life well-adapted to heat, aridity, and drought. In many drylands numerous plants, such as annual grasses and wildflowers, are only visible for short periods following rainfall while spending the majority of their life cycles as below ground biomass in seed banks and in bulbs or other forms. Although often invisible to the casual observer, most arid zones are rich in biodiversity (Laity 2008). The high interannual variability of rainfall in the majority of arid and semiarid regions, however, creates ecological conditions
of nonequilibrium that render many conventional development approaches (like eliminating grazing) and common baselines for defining degradation questionable and sometimes invalid. Moreover, many of the plants, a majority in places like the Mediterranean basin and related climate regions, are also well-adapted to fire and grazing. Indigenous systems of extensive pastoralism, moving livestock frequently over large areas, therefore, is often one of the best uses of land and resources in these regions (Adams 2009, 228–233). Given the nature of arid and semiarid lands and their biophysical components, it is not surprising that well-documented cases of large-scale desertification are relatively few.

Many arid lands scientists agree, in the words of meteorologist Sharon Nicholson, that the extent of desertification has never been adequately determined and in most cases “it is virtually impossible to separate the impact of drought from that of desertification” (Nicholson 2011, 433, 441). A large part of the problem is that there exist at least 100 definitions of desertification and no agreement has been reached on how to measure or monitor desertification (Nicholson 2011; Thomas 1997). This is not to say that degradation in arid and semiarid lands has not taken place or that degradation is not a potential risk. Many details of how degradation can occur in the drylands have been documented, and include salinization, crusting, erosion, nutrient depletion, and acidification (Nicholson 2011, 437–439). Such degradation has occurred but it is limited to relatively few discrete locations usually in the semiarid zones (Helldén and Tottrup 2012; Reynolds and Stafford Smith 2002). How permanent the degradation is and whether it constitutes desertification, which almost always implies irreversibility, is still being debated by scholars. Evidence is growing, however, that a great proportion of the drylands are more resilient and less fragile than previously believed. Many areas categorized earlier as desertified have since recovered with increasing rainfall (Adams 2009; Dardel et al. 2014; Nicholson 2011; Thomas and Middleton 1994).

The persistence of the global crisis narrative of desertification has been analyzed by a
DESERTIFICATION

number of scholars who have shown that it has remained influential because it can be very useful to various powerful actors (Adams 2009; Davis 2007; Swift 1996; Thomas and Middleton 1994). Despite the large and growing body of evidence that undermines it, the crisis narrative of desertification continues to be used by many because it is an efficient way for governments to raise development funds, to spur action, and to justify social policies that control “difficult” populations like nomads. Furthermore, it allows international agencies like the UN as well as multiple NGOs to raise funds and justify many of their programs at the same time that it helps scientists and other scholars generate grants and professional opportunities. The crisis narrative of desertification also persists because it has deep roots in our environmental imaginaries about deserts and arid lands and the histories of these regions.

The roots of our global concern about the crisis of desertification are usually traced to the drought in the African Sahel and the related famine which occurred in the early 1970s because it generated an unprecedented level of concern about desertification around the world that resulted in coordinated action. Some scholars have also highlighted the importance of earlier events such as the Dust Bowl in the United States and colonial activities in West Africa in the early twentieth century in shaping our thinking and policies about desertification. Governments, however, particularly Western colonial governments, used the idea of menacing spreading deserts to justify a host of programs, laws, and policies that facilitated state rule for a long time before the word desertification was first used in the early twentieth century. Many of these programs, laws, and policies carried over into the postcolonial period and several remain influential today. The remainder of this entry traces the history and evolution of Anglo-European thinking about deserts and desertification because it is the legacy of these notions about spreading deserts that directly informs contemporary mainstream global anti-desertification policy today.

Spreading deserts before the “age of desertification”

Deserts were not always considered the ruined or undeveloped wastes that they have been for the last couple of centuries. From antiquity to the age of European exploration, deserts and their edges were primarily considered by Europeans to be naturally dry, sometimes sandy features that presented difficulties for traveling and were known to produce strange and exotic people like the Ethiopians who were believed to be “scorched” by the torrid climate and its burning sun. Importantly, though, they were not conceived as ruined or deforested spaces, nor regions where crops could or should be grown. Early Christianity held the desert in Egypt more or less sacred as the site of ascetic perfection where the desert saints went to prove their piety. For a few, then, the desert represented a heaven on Earth. For others, like Marco Polo, deserts such as those in Central Asia, were strange and wonderful, but also sometimes home to threatening people and frightening spirits.

The “age of exploration,” when Europeans “discovered” the new world, and the rest of the old world, brought new knowledge of very different environments, especially the tropics and tropical islands, but also of other deserts. It is after the so-called Columbian exchange that we see Western perceptions and understandings of deserts and drylands begin to change in significant ways. The sixteenth century witnessed not only the development of early capitalism and colonialism, but also the beginnings of
an articulation of desiccation theory which forged the long-lasting link between deserts and forests. Desiccation theory included the idea that deforestation causes the climate to dry out and diminishes rainfall and also, the corollary, that reforestation restores lost rainfall and creates a generally more humid and healthful atmosphere. There are many problems and weaknesses with this theory and it only applies, if at all, to relatively small “micro-climates” in certain ecological settings.

Early scientific work on desiccation theory primarily focused on the beneficial effects of deforestation and desiccation. This late seventeenth- and early eighteenth-century work tended to highlight the benefits to health and agriculture of clearing forests in cool, humid, continental climates like eastern North America. It did not mention any dangers of deforestation creating deserts. By the mid-eighteenth century, however, this began to change and by the 1760s and 1770s connections began to be drawn by colonial officials like Pierre Poivre and Joseph Banks between deforestation, apparent desiccation, and the creation of desert-like conditions, particularly in tropical island settings such as Mauritius and St. Helena (Grove 1995). Grazing by livestock was frequently blamed along with deforestation for the desiccation of these island environments. The solutions that began to take shape to prevent the creation and spread of arid, infertile conditions were reforestation, elimination of grazing, and creation of forest reserves. These ideas, so well articulated by Pierre Poivre, spread rapidly among French, British, and other European thinkers to become dominant during the nineteenth century.

By the dawn of the nineteenth century, the presumed relations between deforestation, desiccation, and the creation of desert-like conditions were becoming well known and began to be paired with the political preoccupations of the time. Europeans’ notions of the “decline” of the Ottoman empire and its “despotism,” weakness, and deterioration had grown to be quite common. Many thinkers in Europe began to worry that Europe might decline in very similar ways and it stimulated much writing on the subject. Others began to put together these common fears of political/economic decline with fears of environmental decline via deforestation, desiccation, and “desertification.” The region we now call the Middle East, most of which was under Ottoman control during this period, was often used as a withering example of the dangers of deforestation and overgrazing since it was widely presumed that these were the reasons, largely due to Ottoman ineptitude, that the Middle East apparently had become a worthless desert. Many feared that European countries could suffer a very similar decline into despotism and desertification if care were not taken. The most common solutions were forest conservation, which entailed curtailing grazing, fire suppression, and mass reforestation overseen by an enlightened government. Very similar ideas and policies were applied to the arid lands of the United States in the nineteenth century.

These early notions of what came to be known as desertification were also influenced by the increasingly powerful ideas of agricultural improvement and enclosure of commons and waste land during the rise of economic liberalism in the eighteenth and nineteenth centuries. An important transformation took place that changed perceptions of wastes and commons as a “productive remainder” that could provide gathered foods, grazing, and provisions for the poor, to understanding these uncultivated lands as “wasted land” that needed to be enclosed and made agriculturally or otherwise valuable and productive. This transformation included not only the notion that waste land was not being used efficiently but also that it was actually being
DESERIFICATION

degraded by gathering wood or medicinal plants, overgrazing, and the like. For the agricultural improvers, drylands and deserts were the ultimate wastes and they needed the most improvement. Especially in France, the long battle of the improvers to stamp out itinerant grazing in order to enable enclosure over the course of the nineteenth century created widespread antipathy to herding and grazing which became important to the administration of arid lands during the colonial period.

Few European countries had arid lands within their borders, although several southern countries like France, Spain, and Italy had semi-arid regions. Most direct experience with arid lands, then, was encountered by European powers in their colonial territories in Africa and Asia. As a result, much European thinking about and managing arid lands was bound tightly to colonialism and its various goals for more than a century. By 1800 the idea of ruined desert wastes had become firmly rooted in the Anglo-European imaginary. The early nineteenth century was therefore a crucial time for these ideas to be “tested” in the colonial territories. Some of the earliest and most influential experiences were in British Cape Colony (South Africa), French Algeria, and British India.

The French were likely the first in putting many of these early ideas about arid lands into practice outside of Europe, in Algeria, where the colonial narrative of indigenous deforestation and desertification of the environment provided powerful fodder for the development of a wide array of laws and policies that expropriated the local Algerians of significant amounts of their lands, forests, and herds during the colonial period (Davis 2007). Within just a few years of conquering Algeria in 1830, many involved with the colonial project came to understand, incorrectly, the arid and semi-arid regions of Algeria as having been deforested and overgrazed by the indigenous Algerians with their “primitive” agricultural methods and their supposedly destructive nomadic grazing practices. Despite a few dissenting voices at the time, this flawed understanding of the Algerian environment resulted in policies that were said to be aimed at restoring the environment to its former fertility and prevent the desert from spreading. The ideals they aimed to emulate were the “rational” and productive agricultural fields and thick forests back home in France. The resulting policies included sedentarizing nomads, protecting forests from grazing and fire, promoting reforestation, and encouraging or enforcing intensive European farming methods. Many of these goals were supported with a legal transition to private property regimes from collective property regimes justified with the same story of environmental ruin by the “natives.”

From the nineteenth century into the twentieth, this process of arid lands “development” that restricted and criminalized native grazing and agriculture, promoted reforestation, and later built irrigation works for settled farming, intensified across the colonial world from the Maghreb to South Africa to the Middle East to India. Many of these policies have proven harmful to indigenous populations who have lost their livelihoods and also harmful to the environment. In South Africa, for example, afforestation undertaken in the name of desiccation theory in the late nineteenth and early twentieth centuries, resting on the flawed belief that local populations had deforested and desertified the environment has been shown by recent research to be a major contributor to landscape desiccation (lowered water tables). It has led the government to develop plans for large exotic tree removal programs and to formally regulate any new tree plantations (Showers 2010). In India, the Punjab was declared an unproductive desert waste in the late nineteenth century by
the British, the nomads were settled and large irrigation works were built by the administration to intensify farming. This resulted, however, in sizable areas of land suffering waterlogging and salinization and also in significant increases in water-associated diseases such as malaria.

France and Britain were the two colonial powers who controlled the most territory in arid and semiarid lands during the nineteenth and early twentieth centuries. With the dissolution of the Ottoman empire following World War I, arid lands administered by the French and British grew substantially with the creation of the Mandate territories in the Levant. The understandings of arid lands that developed in France and Britain while administering their imperial territories, have been, therefore, especially influential in global thinking about drylands and desertification. As a result, the dominant policies implemented in the drylands at the turn of the twentieth century consisted of sedentarizing nomads, eliminating or reducing grazing, fire suppression, and reforestation in various combinations. During the twentieth century, this Anglo-European dominance of knowledge and policy in the drylands grew and was enhanced with growing American influence.

The age of desertification: the twentieth century

It was within this long-standing Anglo-European imperial context, then, that the French colonial forester Louis Lavauden coined the term “desertification” in 1927 and pronounced that desertification was the result of the improvidence of nomads and the overgrazing of their livestock (Lavauden 1927). Such had been the thinking in the colonial territories of the Western world for a long time, and it built on an old European intellectual tradition of anti-nomadism (Adams 2009; Davis 2007).

Other indigenous livelihood activities were also frequently blamed for desertification, including local agricultural practices (especially the use of fire). The British forester Edward Stebbing increased fears of anthropogenic desertification with his alarmist articles during the drought years of the 1930s. He warned in dire terms of the southern spread of the Sahara desert, despite the evidence from the Anglo-French forestry commission to West Africa that the Sahara was not spreading to the south. The American Dust Bowl of the 1930s further heightened knowledge of and concern with what had become known as desertification and with it focused attention on the problems of soil erosion in dryland environments.

The events of the 1930s thus led to the internationalization of the idea of a “crisis of desertification” that raised the profile of environmental change in the drylands to a global level. The concern with erosion raised by the American Dust Bowl experience was taken up in many Western colonial territories in Africa and Asia with American-inspired policies to curb erosion in ways that were often ecologically inappropriate. All too frequently, the new anti-erosion policies, along with more well-established anti-desertification policies, led to socially repressive programs and the implementation of political goals in the name of environmental protection (Adams 2009). In Mandate Palestine, for instance, such thinking led the British-run forest department to report, incorrectly, that half of the habitable area of Palestine had become an artificial desert due to overgrazing and to implement grazing restrictions and reforestation programs with the help of Zionist interest groups.

Following World War II, as many former colonial territories were gaining, or about to
Figure 3  Camels grazing in southern Morocco. Without an understanding of arid lands ecology, scenes like these camels grazing an apparently bare landscape are easily misinterpreted to be cases of “overgrazing.” Camels are actually one of the few domesticated species able to survive arid conditions due to their unique ability to conserve water and eat plant species normally unpalatable to most other livestock. Source: Photo by the author.

gain, independence, the newly created United Nations formed the Arid Zone Program in 1949 in its UNESCO branch. Framed as “man’s battle against the desert,” the program grew in size and influence and by 1956 it had been renamed the Arid Lands Major Project, with permanent seats held only by Great Britain, France, and the United States (Selcer 2011). A few other countries, mostly from the Middle East, were represented on a rotating basis. The Arid Lands Project became a centerpiece for UNESCO and allowed the young organization to gain an international reputation in both science and resource management. It incorporated the vast majority of earlier Anglo-European thinking and policies on arid lands and threatening deserts and further spread these ideas and policies globally with training programs and publications such as The Future of Arid Lands, the Arid Zone Newsletter, and the Arid Zone Research Series. Although it was subsumed under the International Hydrological Program in the 1960s, the Arid Lands Project set the stage in several ways for the later development of the UN Conference on Desertification a decade later.

When a serious drought hit the African Sahel in the early 1970s, attention to desertification exploded. The dramatic famine that accompanied the drought killed hundreds of thousands of Africans and their livestock and was covered sensationaly in the Western media. This particular drought appeared all the more dramatic because it followed a couple of decades of above normal rainfall in much of Africa. The “natural disaster” of drought was commonly believed to have pushed environments and burgeoning communities already weakened by decades, if not centuries, of alleged desertification through improper land use over the edge of survival. The drought also followed many decades of colonial rule in most of the region, however, that had disrupted precolonial social and agricultural
DESERTIFICATION

systems to a great degree which led to a higher death toll. As a consequence of this widely publicized disaster, the 1977 United Nations Conference on Desertification (UNCOD) was held which generated the desertification branch of the UN Environment Programme (UNEP).

Following in many ways the example of the earlier Arid Lands Program of UNESCO, the new UNEP desertification branch wrote action plans and published the Desertification Control Bulletin and the World Atlas of Desertification, which echoed much of that earlier UNESCO work. It also produced numerous reports that were highly politicized and rarely based in field research or ecological detail (Nicholson 2011). As a result, the extent and severity of desertification globally were quickly exaggerated. Just as during the colonial period, however, the “crisis of desertification” suited the purposes of many newly independent governments in their efforts at state building and controlling both agricultural development and populations perceived as unruly. UNCOD’s perspectives and prescriptions were reaffirmed in the 1994 UN Convention to Combat Desertification (UNCCD) and further institutionalized with the creation of its subsidiary body, “the global mechanism,” designed to disperse anti-desertification funds in 1997 (Adams 2009). In 2006, the UN declared the International Year of Deserts and Desertification which entailed more publicity of the global “crisis” of desertification.

Conclusion

The hegemonic perception of deserts and desertification embodied in UNCOD/UNCCD, and many other institutions, though, is profoundly flawed. Moreover, “it seems to have triggered many problems, of understanding and of action, that manifested themselves in the ways in which desertification was conceptualized, represented and approached as an environmental, social and political issue” (Thomas and Middleton 1994, 49). According to these two geographers, and many other specialists, dryland ecosystems “appear to be well-adapted to cope with and respond to disturbance, demonstrating good recovery characteristics” (Thomas and Middleton 1994, 160). In other words, most drylands around the world are resilient, not suffering significant degradation, and deserts are not expanding. A large and growing body of scientific research has reaffirmed these conclusions over the last two decades. Some researchers have even proposed eliminating the term desertification altogether to facilitate greater understanding and more sustainable development in the drylands.

Arid lands scientists Stefanie Herrmann and Charles Hutchinson have insightfully noted recently that “we have a non-equilibrium world that is saddled with an overriding equilibrium mindset and [anti-desertification] policies that reflect it” (Herrmann and Hutchinson 2005, 38). They conclude, as have many others, that “policies that affect people on the ground may be formulated largely independent of science that is current and thus may serve to degrade rather than enhance the lives of people most affected” (Herrmann and Hutchinson 2005, 38). One of the biggest points of contention surrounding desertification, then, is between the scientific/academic community and the global policy community, especially at the levels of international agencies like the UN and many national governments.

Although some agencies and programs have begun to understand and act on problems of land degradation in the drylands as intricately connected to problems of poverty and social,
DESERTIFICATION

political, and economic issues, too many projects are still implemented to “roll back the desert” which uncritically assume desertification. Such projects usually rely on some combination of sedentarization and destocking, reforestation, and increasing irrigated and non-irrigated agriculture. These old policies, derived from the long-standing assumption that deserts are the result of overgrazing and deforestation, have been shown to increase degradation in many nonequilibrium, arid environments. Other approaches exist that can enhance dryland environments when tailored for specific environmental conditions such as terracing, water harvesting, and extensive grazing, among others (Adams 2009; Nicholson 2011). The difficulty that many policymakers have in understanding and implementing these differences is strongly related to the underlying environmental imaginary of deserts and desertification that has deep roots in the colonial period and earlier. Bridging the gap between the science of drylands and mainstream policy development in these regions is crucial for building more sustainable environments and livelihoods in the world’s magnificent drylands.

SEE ALSO: Desert ecosystems; Development; Famine; Geopolitics of the environment; Land degradation; Livelihoods

References


Deskilling consists of simplifying the tasks of individual or collective labor. This includes lessening its technical (e.g., the amount of formal training/education) and/or its conceptual (e.g., the level of conscious planning and thought to complete a task) content both within firms and/or in the domestic sphere, as the result of either an explicit strategy or an unintended effect of changes in work organization and/or labor supply. Beginning in the 1970s, deskilling became the object of an intense debate in the social sciences over the nature of the capitalist labor process, sparked by Harry Braverman’s (1974) book *Labor and Monopoly Capital*. Influenced by Marx, Braverman argued that the introduction of Taylorism and scientific management in US manufacturing in the early twentieth century allowed the separation of conception, or the planning and organizing of work, from its execution. Conception was appropriated by a new cadre of managers and supervisors, while work formerly performed by craft workers was reduced to deskill assembly tasks. This allowed a significant increase in the intensity of work and productivity, especially when combined with the new assembly-line techniques of manufacturers like Ford. Braverman argued that deskilling had become dominant not only in manufacturing, but also in newer service employment.

Braverman’s arguments were subject to criticism by both Marxian and non-Marxian scholars, who argued that control of the labor process was only one means to capital accumulation and could often lead to the development of new products and processes, which also created new skills (Burawoy 1985). Geography scholars also stressed the uneven sectoral and spatial development of the production process (Sayer and Walker 1992) while feminist (including feminist geography) researchers noted that Braverman and much of the labor process debate had ignored the critical role of gender not only within the formal capitalist workplace but also in the domestic sphere (Cockburn 1983; Hanson and Pratt 1995). Other scholars argued that firms using Japanese production methods had reskilled employees or, more generally, that most jobs require higher levels of education and training over the long term (see Block 1990). Defenders of Braverman argued that “reskilling” strategies could be consistent with work intensification and did not challenge deskilling as a dominant long-run tendency in capitalism. The debate over skill and the pervasiveness of deskilling has continued, for example, in relation to the “creative class” (see Florida 2012) and the impact of new technology. Thus Krugman (2013), citing a McKinsey Global Institute (2013) study, argues that new technology is likely to be disruptive and displacing of employment, including that requiring high levels of education.

SEE ALSO: Gender; Skill; Technology

References

DESKILLING


The nature and scope of development

One of the hallmarks of contemporary global societies is that, at whatever scale we view them in terms of “levels of development,” they are almost universally characterized by major contrasts and differences. As proof, a recent 2006 study carried out by the World Institute for Development Economics Research on behalf of the United Nations (UNU-WIDER) showed that 50% of the world’s total population owned total assets of less than US$2200 and collectively accounted for only 1% of the world’s total wealth. Indeed, over half the world’s population is currently to be found living on less than just over $2 per day. Furthermore, the same study showed that one did not have to own much in the way of resources to be part of the richest group of the population on Earth. Specifically, owning assets worth $500,000 or more, or about £350,000, or more, puts one into the richest 1% of people on the planet who control some 40% of the world’s total wealth. Such widely divergent conditions, standards of living, and associated qualities of life require explanation not only because such unfairness should not be tolerated but because such inequalities belie the promises concomitant with modernization and growth theories.

It is this central developmental issue that frames this entry, which seeks to provide an overview of the evolving nature of the cross-disciplinary and interdisciplinary field(s) of development studies and of development geographies and cognate areas of social science, as well as a brief introduction to the practice of development in the global policy arena (see also Potter et al. 2012). It may be argued that, in the past, development geography and the cognate interdisciplinary area of development studies showed signs of being treated as the Cinderella of the discipline of geography in particular, and of the social sciences in general. This can no longer be tolerated because the current pace and penetrative power of globalization and the spread of the canons of the free market or neoliberal order are making the world more and more unequal and unjust. This means that it is vital to include development–related work at the core of scientific assessments of societal change and progress in this iniquitous, unruly, runaway world (Conway and Heynen 2006; Piketty 2014).

So, what exactly is implied by the word “development” in its global-developmental context? This entry proposes that development at its most basic level is about improving the life conditions that are faced by the global majority, and specifically this means reducing existing levels of poverty and inequality at local, regional, national, and global scales. A more detailed working definition of development might be adopted that stresses development as attempts to reduce poverty and world inequalities in an effort to guide the world to a situation of betterment and improvement over time (see Kothari, 2005; Potter et al. 2008). It would seem axiomatic that development should provide for the betterment
of the poor majority in the first instance, as this will add the most to what may be referred to as the common good. The essence of development is that there is a poor world and there is a rich world, and it is implicitly assumed that it is the responsibility of the latter to assist the former. In other words, there is a pressing ethical need to equalize the highly disparate conditions that currently exist in and between the poor and rich worlds, so that the tenets of Denis Goulet’s (1971, 2006) and Des Gasper’s (2004) “development ethics” are constantly at the forefront of “thinking” and “doing” development (Desai and Potter 2014).

Development, therefore, is very much a practical subject and nations, international agencies, nongovernmental organizations (NGOs), and community-based organizations are all involved in the policy-related processes of trying to promote development on the ground. Combined together, the practice of development and the concepts that underlie such practice give rise to what is referred to as the broad arena of “development studies.” Development studies, therefore, represents theoretical, empirical, and policy-related studies concerning how development has, and should be, implemented. It can be further argued that, rather than being either inter- or multidisciplinary, development studies can make a strong claim to be cross-disciplinary in nature in that it serves to bring together a large number of fields in the study of poverty and inequality. This is represented in graphical summary form in Figure 1. The central concern of development studies may be regarded as the existence and seemingly inexorable deepening of global poverty and inequality.

Development studies has its origin in the 1940s, when development became an overt aim of Western or advanced nations, specifically those of Europe and North America. In its early stages, “breakaway” economists were strongly involved in the rise of the field. Reflecting this in the core discipline, a distinct subdiscipline within economics can be recognized that is now conventionally referred to as “development economics” (see Figure 1). Geographers, with their strong tradition of regional and area studies

![Figure 1](image-url)
represented another area of involvement and interest. In the same manner as in the case of economics, the rise of development geography as a distinct field can be recognized. Other mainstream social science disciplines, such as politics and sociology, also contributed to the rise of development studies and came to be known respectively as the sociology of development and the politics of development. Attesting to the truly cross-disciplinary character of the field, subjects such as demography, international relations, anthropology and history, and urban and regional planning should also be identified as making a distinct contribution to development studies.

Indeed, cross-disciplinary theory development, and regional empirical contextualization would be fostered within this sharing of disciplinary concern for fair and just human development. Throughout the second half of the twentieth century and into the twenty-first there were a plethora of alternative perspectives, alternative visions, and alternative approaches that challenged the modernity theorizing that was at the core of the Global North’s “classical traditional” approaches (see Figure 2). The evolving nature of these alternatives is reviewed later in the entry. By way of an introduction, competing measurements and assessments of

![Figure 2](image-url)
what development should mean to the world’s majorities are first reviewed.

Assessing and measuring development

Both researchers and policymakers have sought to assess and measure development with competing approaches that directly reflect the historically contingent and evolving conceptualizations of development as a process (Desai and Potter 2014; Potter et al. 2012). During the 1950s through to the early 1980s, development was generally measured in terms of the level of economic growth achieved by countries and, in particular, the growth of production and income. From the late 1980s through to the 1990s, changes in the way development was being envisioned were directly recognized in the promotion of wider indices of human development and change. This trend toward recognizing the multidimensional nature of development has continued from the 1990s through to the second decade of the twenty-first century, whereby wider sets of factors, reflecting more subjective and qualitative dimensions, have increasingly been employed to define development. Three general approaches have characterized these trends (see Potter et al. 2012):

1 measuring development in terms of economic growth, by means of gross domestic product (GDP) and/or gross national product (GNP) per capita;
2 measuring development in terms of human development: the Human Development Index (HDI); and
3 measuring development in terms of wider dimensions, including human rights and freedoms.

Throughout, the account in this entry is strongly based on that provided in Potter et al. (2012).
labor that accompanies low levels of physical capital (natural resources), human capital (e.g., education), and technology.

HDI and wider measures of development

Development is far wider than the growth of income alone, and many development studies practitioners and scholars view GDP/GNP measures as highly limited in their ability to capture the depth and breadth of socioeconomic change. Specifically, GDP and GNP take no account of the distribution of national wealth or of output between different groups of the population or between different areas/regions. Further, such income-based measures do not take into account the broader wellbeing of people, which includes more than goods, money, and material wellbeing. Given these limitations, and the increasing recognition of the need to account for noneconomic factors in development, new or alternative measures of development came into being, starting in the 1980s.

In 1989 the United Nations Development Programme (UNDP) promoted the Human Development Index as a wider measure of development. HDI data were published for the first time in 1990 in the inaugural Human Development Report (UNDP 1990). In the original HDI, the emphasis was placed on assessing human development as a more rounded phenomenon. There was still a measure of economic standing, but this was only one of three principal dimensions identified:

1 a long and healthy life (longevity) – originally measured by life expectancy at birth in years;
2 education and knowledge – initially measured by the adult literacy rate and the gross enrolment ratio (the combined percentage of the population in primary, secondary, and tertiary education);
3 a decent standard of living – originally measured by GDP per capita in US dollars, as outlined in the previous section.

Since 1990, the Human Development Report has been published by the UNDP every year, but in the Human Development Report 2010 the formula was changed. The three dimensions of health, education, and living standard have been translated into four indicators, and these are summed to give a single HDI. The common practice, therefore, has been for HDI to be used to divide nations into what have come to be referred to as high, middle, and low levels of human development. Recently, the classification has been extended to also include a very high human development category.

It should be stressed that the HDI is a summary and not a comprehensive measure of development. For example, in the years since its introduction, various methodological refinements and spin-offs have been made by the United Nations, including the Human Poverty Indexes 1 and 2, the Gender-Related Development Index, and the Gender Empowerment Measure. These are all variations on the basic Human Development Index. In each case, additional variables were brought in to reflect the revised index. (For a more in-depth discussion, see Potter et al. 2012.)

Measuring development in terms of wider dimensions, including human rights and freedoms

While the HDI extends our measure of development into the noneconomic realm, a number of writers have stressed the importance of self-esteem, basic freedoms, and human rights
as further components of the development equation (see, e.g., Sen 1999). Such views represent specific recognition that wider aspects of development are vital, particularly those that relate to the quality of people’s daily lives, their freedom from various inequalities, and the attainment of human rights and basic freedoms. In an attempt to account for and measure progress in relation to these dimensions of development, the UNDP, starting in 2000, began developing a set of indicators and objectives through which to guide global efforts at achieving broad-based development.

The resulting Millennium Development Goals (MDG) are designed as instruments and metrics to steer the world to enhanced levels of development (Rigg 2014). By 2013, there were eight broad goals that serve as priorities for the global development project:

1. Eradicating extreme poverty and hunger – measured by the percentage of the population living on less than $1 or $2 per day (now $1.25 and $2.50 per day);
2. Achieving universal primary education;
3. Promoting gender equality and empowering women;
4. Reducing child mortality;
5. Improving maternal welfare;
6. Combating diseases;
7. Ensuring environmental sustainability;
8. Achieving global partnerships for development.

For each of the eight MDGs there are associated targets and detailed indicators (see http://www.un.org/millenniumgoals for details). The indicators can be seen as dimensions that can be employed in order to assess the progress of nations and regions toward the goals and targets, and thereby represent measures of the wider dimensions of development.

An impression of how such indicators can be used as measures of progress in development can be gained from the National Reports covering progress with the MDGs that are available on the UNDP website. The 94-page Report for India 2009, for example, shows in considerable detail the mixed success achieved on the 12 targets that apply to it. In terms of basic human rights, an interesting approach is to chart the extent to which countries have ratified the six major human rights conventions and covenants (e.g., the United Nations Convention on the Rights of the Child, the Convention against Torture; see Potter et al. 2008). In a similar manner, the United Nations Human Development Report 2010 introduced the HDI-derived Gender Inequality Index (GII). The statistics input to the GII include the national female and male shares of parliamentary seats and educational attainment. The GII also includes female participation in the labor market. The GII thereby represents a direct effort to measure the progress made by countries in advancing the standing of women in wider political and economic developmental terms.

The mainstream economic development project

Having broadly defined what is meant by development and how it can be assessed and measured, the entry now provides an overview of the changing concepts that have directed development in what has been referred to in a multitude of terms, including the “third world,” “Global South,” “developing” or “poor worlds,” or whatever other phrase is used to denote those parts of the globe which are deemed to be in pressing need of enhanced rates and levels of development. At the outset, the account considers the multidisciplinary concept
of modernity as the focus of societal development manifest in recent mainstream (Keynesian and neoliberal) approaches to achieving development. In the sections that follow, the historical record of alternatives to these top-down economic development approaches are detailed.

**Modernization, modernity, and Keynesian capitalist development**

“Modernization,” in its earliest promotional guise as “Western” modernist thought, firmly placed third world or Global South “traditional” societies as “backward” if not also “savage” and “uncivilized.” For development and progress to take place, modernization and modernist changes similar to those in play in Europe and the “West” had to occur (Power 2003). Modernity is distinguishable from modernization in terms of the new technologies and innovations that accompany capitalism’s evolutionary transformations, from early urban industrialization to advanced capitalism, and on to neoliberal and global capitalism.

Modernity brings about new modes of transport, new sources of energy, new management and communication technologies, new media forms, and the like. These, in turn, change social relations, power relations, and institutional structures and supplant or modify previous ways and means. To Giddens modernity may be defined in terms of being:

1. associated with (1) a certain set of attitudes towards the world, the idea of the world as open to transformation, by human intervention; (2) a complex of economic institutions, especially industrial production and a market economy; (3) a certain range of political institutions, including the nation-state and mass democracy. (Giddens and Pierson 1998, 94)

It was, however, the economic constructs built on modernization’s top-down tenets that had the most lasting influences during the early development decades of the 1950s, 1960s, and 1970s. More specifically, modernity and modernization both became the main conventional economic models favored by the development establishment since the second half of the twentieth century. Between 1940 and 1980, modernization (as development) was thought to be best propelled by the Keynesian model of advanced capitalism that prevailed in the developed “first world” of the United States, the reconstructed “war-torn” economies of Europe, and a rebuilt Japan. Keynesian capitalism favored governmental intervention to enable robust growth such that both social welfare safety nets and public subsidization programs were common partners to private enterprise.

Central to this economic formulation for the post–World War II “developing world” was the work of W.W. Rostow and his influential 1960s book *The Stages of Economic Growth: A Non-Communist Manifesto*. As evidenced in the title, Rostow had interwoven two major concerns, one political and strategic, the other economic and developmental. Rostow was geopolitically concerned with the international and political contexts of the process of transition in third world nation–states. His model, accordingly, was constructed in large part to further the strategic influences of the United States. Rostow’s model proposed that all societies would be able to transit through five stages: traditional society, preconditions for takeoff, takeoff, the road to maturity, and the age of mass consumption. Crucial to his takeoff and road to maturity stages were the infusion from outside (presumably the United States or its Western allies) of two factors of modern industrial success:

1. the existence of already functioning modern technology which can be made available by technology transfer from benevolent outsiders and public–private partners; and
the existence of international aid and technical assistance (including technical education and consultants) provided by the governments of developed countries and their contracted professional/technical advisers.

Neoliberalism and market-led development

Following the global recession of 1978–1983, a new capitalist model, labeled neoliberalism, brought about a dramatic turn away from Keynesian economic thinking and political-economic practices just about everywhere in the globalizing world. Accordingly, supply-side economics solutions to economic recovery – Reaganomics and Thatcherism, for example – were offered as alternatives to combat inflation, reduce government overspending, reduce public sector workforces, and roll back wages. Neoliberalism as an ideological discourse and unchallenged model of economic efficiency, fiscal austerity, and capitalist enterprise has prevailed since the 1980s, forging the contemporary era of neoliberal capitalism and globalization.

In theory, neoliberalism is about increasing the scale, scope, and pace of the movement of goods, commodities, resources, and commercial enterprises so that cheaper resources can be accessed to maximize profits, enhance efficiency, and increase income per capita worldwide. Neoliberal strategies seek to achieve these objectives through the removal of various controls (e.g., tariffs, quotas, capital regulations) and restrictions on capital flows and investment across national boundaries. Beyond its transnational implications, neoliberal ideologies and policies also advocate for or promote the following within nation-states:

- the rule of the market – freedom for capital, goods, and services, where the market is self-regulating – allowing the “trickle-down” notion of wealth distribution; it also includes the de-unionizing of labor forces and removal of any impediments to capital mobility, such as regulations (the freedom is from the state or government);
- reducing public expenditure for social services, such as health and education, by the government;
- deregulation, to allow market forces to act as a self-regulating mechanism;
- privatization of public goods, resources, and services (ranging from water, power, and transportation to information dissemination and exchange, Internet use, and communication);
- changing valuations of public and community worth to privilege individualism and individual responsibility.

With respect to development issues in the Global South, the International Monetary Fund (IMF) and World Bank have been the stalwart “enforcers” and promoters of neoliberal policies, notably through their Structural Adjustment Programmes (SAPs), Poverty Reduction programs, and austerity and “stabilization” measures designed to reduce public spending, remove subsidies on basic foods and other commodities, privatize state-owned enterprises, reduce the government’s wage bill, and devalue or free-float domestic currencies. In addition, neoliberal initiatives promote market-led forms of public service provision (e.g., through public–private partnerships) and economic liberalization measures which remove tariffs and trade barriers and deploy tax reductions and subsidies to encourage local entrepreneurship and inward foreign investment by transnational corporations. The World Bank, IMF, and the regional development banks (e.g., Asian Development Bank, Inter-American Development Bank) often place “conditionalities” on development loans to ensure that such
policy measures are enacted and that developing countries are opened up to more international trade and foreign corporate investment.

All told, while neoliberal initiatives have fostered GDP growth in some countries, particularly those rich in natural resources such as minerals and oil, there is little or no evidence to suggest that widespread wealth distribution is on the rise, that inequality is decreasing, and/or that economic diversification (à la Rostow’s stages) is occurring on a large scale. In essence, the contemporary circumstances seem remarkably similar to those observed in the colonial and early postcolonial period where surplus extraction by Western states and corporations and South-to-North dependencies stifled widespread human development in the Global South. As Harvey effectively summarizes, the contemporary neoliberal project fails to adequately or justly address pressing questions of inequality and redistribution:

The world is broadly polarised between the continuation (as in Europe and the United States) if not a deepening of neoliberal, supply-side and monetarist remedies that emphasise austerity as the proper medicine to cure our ills and the revival of some version, usually watered down, of a Keynesian demand-side and debt-financed expansion (as in China) that ignores Keynes’s emphasis upon the redistribution of income to the lower classes as one of its key components. (2014, xi)

Alternative, radical, and people-centered theories of development

Revolutionary and radical ideas concerning societal development were commonplace during the second half of the twentieth century. As if the end of World War II signaled that dramatic, nonviolent change needed to be brokered differently, from the 1950s onward, societal upheavals, popular resistances, elite power-brokering, militarism, pacifism, colonial oppression, and postcolonial independence movements occurred throughout the world. Writings that enflamed passions, ignited revolutionary (or radical) ardor, and cried out for social justice for all — not just for the wealthy and privileged — were as much at the heart of violent revolutionary action as they were central to the development of radical approaches to development.

This section focuses attention on people’s ideas and writings on “radical” or “alternative” approaches to development and social progress in the underdeveloped third world/Global South. Early on, radical critiques and “antidevelopment” arguments were leveled at Westernization, imperialism, modernization, and neocolonialism by third world spokesmen (Brookfield 1975). Following on from such critiques, the remaining sections detail a representative set of critical, alternative models that their advocates felt could, or would, bring about “development from below.” These alternative approaches culminate in people-centered development models, which combine development ethics, sustainability, capacity building, and community empowerment as progressive means for the poor majority to achieve development (see Figure 2 for an overview).

Marxist critiques and models of (under)development

Anti-imperialist, radical approaches which were offered as alternatives to the Western, Eurocentric, modernity-based models were idealistic geopolitical works, often written by Soviet scholars. Often, translations of Lenin’s writings, as well as those of Karl Marx and Frederick Engels, were used as theoretical and ideological guidance by local revolutionaries and radicals. Some
academically trained political leaders of newly independent countries, such as Tanzania’s Julius Nyerere and Ghanaian Kwame Nkrumah, wrote their own nationalistic manifestos advocating socialist, revolutionary models.

Other radical authors penned socialist and Marxist critiques that became well cited authoritative texts for decades. For example, Paulo Freire’s Pedagogy of the Oppressed (1970) became an extremely influential text for educators interested in improvements in literacy in third world territories. Another seminal work was Frantz Fanon’s The Wretched of the Earth (1967), a devastating critique of colonialism’s damaging impacts on the psyche of colonized societies’ new political elites. Particularly perceptive was Fanon’s portrayal of the imitative attitudes of the new generation(s) of third world leaders, which would lead them to continue on similar paths of servitude and deference to “outsiders” in their postcolonial relations with their ex-imperial masters.

Perhaps most influential of the radical critiques came from a group of scholars based or working in Latin America in the 1950s and 1960s. In his influential 1957 anticapitalist treatise, Marxist economist Paul Baran “explained” Latin America’s underdevelopment as a consequence of Western nations forming special partnerships with powerful elite classes in the continent’s underdeveloped countries. These special partnerships perpetuated the ability of advanced capitalist, core countries to maintain traditional systems of surplus extraction and ensured that the periphery’s domestic resources would be continuously available. By such means, they made economic development unlikely, since any surplus generated was appropriated by these client-elites. This would then enable core countries to keep their own monopoly of power over cheap primary resources (Baran and Sweezy 1966).

Notably, these alternative anticapitalist ideas on uneven exchange, surplus redistribution, and expropriation would be further developed by a chorus of voices from below: Latin American political-economy scholars, collectively known as Economic Commission for Latin America (ECLA) structuralists, and later as dependistas, or the “Dependency School” (Brookfield 1975). Although their neo-Marxist critiques differed in theoretical detail, all argued that Latin America’s historical marginalization and resultant underdevelopment were perpetuated by unequal commercial arrangements with core countries (particularly, the neighboring United States) who benefited at the expense of Latin America.

André Gunder Frank would soon become the main advocate for the dependistas. Focusing on the dependent character of peripheral Latin American economies, Frank coined the term “development of underdevelopment” to characterize the capitalist dynamics that developed core countries while at the same time ensuring underdevelopment and dependency within Latin American countries. Frank’s conceptual framework therefore explained the dualistic capitalist relations which had occurred, and which he felt would continue to occur between Latin American and core counties. In line with other anticapitalist critiques, Frank (1966) argued that any change for the better would be possible only through revolutionary action, which would install socialist ideals within the political systems of Latin America’s dependent countries and usurp the client-elites’ special relationships.

Another highly influential and passionate third world voice was that of Samir Amin, who expanded on the notion of dependencia in his critical examinations of the dependent economic relations of Africa since European colonization, and found them to be the destructive and disastrous consequences of that continent’s peripheral
development of capitalist condition. His writings, spanning the 1970s to the present (and dying in 2005), are voluminous and provide a highly persuasive case for the failure of the “development project” all over Africa (Adil 2007). Amin goes so far as to propose extreme measures to overcome the decades of failed Eurocentrism and its modernization approaches, such as “delinking” from Europe and starting afresh (Amin 1990).

Broadening the global scope of this Latin American-based construct, Wallerstein (1974; 1980) adapted dependency notions to comment on the commercial relations between the core countries and Latin America, and examined world historiography in terms of the dominant and subordinate relations that successive emerging cores, their peripheries, and semiperipheries experienced. Wallerstein’s account starts with the long sixteenth century and passes through successive eras of capitalism to arrive at the present neoliberal era of globalization (the post-1980s). His world-systems theory complements and expands on Frank’s ideas, providing a more comprehensive global stage appropriate for understanding the wider reach and more diverse spatial realignments of commercial capitalist relations in contemporary times. More recent world-systems explanations of geopolitical eras detail the transformations of the world’s hegemonic relationships of core–periphery relationships to the present global era in the second decade of the twenty-first century (e.g., see Sparke 2013).

Alternative approaches: development from below, empowerment, and capacity building

Western modernity’s externally driven strategies to bring about economic, social, political, and cultural changes in the underdeveloped client states of the third world came to be characterized as development from above by its radical critics. They criticized its import-substitution form of industrialization and its monolithic and uniform value systems, which lauded capitalist entrepreneurialism, urban living and intellectual “modernity.” In contrast, development from below sought to replace such externally driven dependent and subordinate relations and to increase equity rather than inequity across regions. It sought the creation of dynamic development impulses within regions that were less favorable to urban industrial growth and “growth pole” development, such as rural peripheries and remoter hinterlands.

From its inception in the 1970s, development from below was an activist agenda with a progressive message on the need for radical change in policy formulation and implementation. Succinctly, it argued that economic growth from above needed to be supplanted by development from below so that the basic human needs of all people could be met, rather than just those of elite minorities and the powerful classes. Development policy needed to be reoriented toward basic needs service provision at the territorial scale. Rural and village development should be as much a focus as urban-located development. Labor-intensive activities and micro-economic small businesses and projects should be favored over high-technology entrepreneurialism. Such territorially integrated development should aim to provide full employment, and regional human, natural, and institutional resources should be mobilized. Following Schumacher’s vision that Small Is Beautiful (1973), it was argued that intermediate technologies, small- and medium-sized projects, and participatory and locally designed and executed projects should be integral parts of the mix of development strategies (Chambers 1997). For the next three decades, from the 1970s to the 1990s, a mix of alternative approaches,
most seeking similar objectives for their humanistic development paths, would be launched in academia and in policy circles.

A key contemporary successor to development from below alternatives is the notion of people-centered development (Eade 1997). People-centered approaches explicitly blend concerns about societal and environmental sustainability and promote capacity building and community empowerment as progressive means through which the poor majority may achieve development. Saliently, civil society, NGOs, and other grassroots philanthropic organizations are embracing this notion. Proponents of people-centered approaches maintain an ethical position that is aligned with Goulet’s (1971) development ethics, and which seeks to make the real-life economy of people more human and humane, as well as ecologically sustainable (Korten and Klauss 1984). Decrying the roots of the current “global problematique” – Western modernity, scientism, and developmentalism, the nation-state’s political limits – Ekins (1992) sought a new world order of “democratic popular mobilisation” in which “real-life economics” was to be a radical alternative to positivist development economics. Given the structural rigidity and persistence of the stark global problems presented by war, insecurity, and militarism; the persistence of poverty; the denial of human rights; and the occurrence of environmental degradation, Ekins believed this radical option would be better able to bring about peace, human dignity, and ecological sustainability for all.

At the same time, “real-life economics” was never offered as a complete alternative to conventional political-economic orthodoxy. The structural geoeconomic forces at work globally pose so many diverse problems that such a local-level perspective can scarcely be viewed as an all-encompassing equivalent. Rather, the approach was an intellectual, theoretical alternative, and the basis for a new interpretation of people’s day-to-day economic strategies. In support, Ekins and Max-Neef (1992) marshaled a community of like-minded scholars to offer their alternative views on “living economics” in which real-life economics, ethics, environmentalism, and sustainable development should be considered inseparable dimensions in tomorrow’s thinking about human life and livelihoods. They warned that the market, the state, and civil society, singly or in combination, are all potential creators or destroyers of wealth. They cautioned that monetary concerns are not the only motivations for economic choice, and that livelihood behaviors do not have to be rational. Rather, habits, intuition, and bounded “satisfying” rationality guide human decision-making. Love, altruism, duty, respect, dignity, and obligations are all meaningful influences equal to economic rationality and wealth accumulation.

Gender and development debates also drew on this discourse, but with a focus on the issue of women’s empowerment, which subsequently became important for development institutions such as the UNDP and Oxfam, and for Global South government agencies and NGOs (Momsen 2004). Driving this initiative was a growing concern that gender equality would not be achieved unless women could challenge patriarchy and global inequality. Most importantly, scholar-activists from the Global South raised crucial questions as to how this objective might be attained, with a rising tide of female advocates supporting its goals. While women’s greater representation in government circles was seen as an achievable objective (and successfully accomplished in time), the plight of poor women and their empowerment and representation in decision-making processes were viewed as much more difficult to achieve. It was also clear that the reality of many poor women’s
struggles for empowerment and local communal authority required the analysis of local complexities within their regional, national, and global contexts. It was eventually realized that empowerment projects and their local advocates needed to pay attention to the ways institutional structures and discursive frameworks shaped the possibilities and limits of individual and group agency and choices. Furthermore, the inclusion of men and masculinity is a welcome expansion of the empowerment project and a key to more comprehensively addressing gender empowerment in our contemporary globalizing world (Parpart, Rai, and Staut 2002).

Bringing NGOs and the wider plural, civil society into the picture, Friedmann (1992) viewed “alternative development” as an expression of militant and activist civil society, also known as NGOs. Upholding universal human rights had to be the prime political objective, especially the rights of the disempowered poor majority of the Global South. For Friedmann, the empowerment approach is fundamental to such an alternative approach because it places emphasis on autonomy in the decision-making of territorially organized communities. It thereby brings about local self-reliance (but not unrestricted power), direct (participatory) democracy in communal affairs, and experiential social learning. With people empowered by participatory democratic involvement at the local level, they will become engaged in the larger processes of representative governance, and experience progressive “alternative development.”

The central idea behind local institutional capacity building is about how outsiders such as NGOs and private–public partners in development projects can help foster community authority and managerial expertise in the provision of social capital, social welfare and services, agricultural extension services, urban neighborhood community projects, and the like (Eade 1997). With governmental involvement as an enabler, rather than an active participant, or as leader and manager, this people–centered strategy seeks to strengthen institutional and social capacity to support greater local control, accountability, transparency, initiative, and self-reliance for urban and rural communities of the Global South (Korten and Klauss 1984).

A reflective assessment by Oxfam GB summarized the potential of this current development mantra: “The concept and practice of capacity building has to be tested against whether it can contribute to creating the synergy between different actors which can confront and challenge existing imbalances of power” (Eade 1997, v)

All told, people–centered approaches seek to keep hope alive in the face of the seeming impossibility of achieving human development for all, utilizing current paradigms. They seek progressive change, and a path forward that builds self-reliance, empowers men and women, and puts the first last in terms of which people are the most deserving: principally, the poor (mostly rural) majority who have been left out of the development project to date (Chambers 1997). They challenge conventional economic thinking and seek humanistic solutions, in which development is conceived in terms of people’s sovereignty. Succinctly put, peoples’ participatory involvement in controlling resources, thereby, gives rise to individual, familial, and communal empowerment. Furthermore, and with reference to his conceptualization of development ethics, Goulet (1971) also recognized three social freedoms as crucial quality-of-life components of a people–centered development that follows ethical tenets: (i) life sustenance and the provision of basic needs; (ii) self-esteem acquired through self–respect and independence, or autonomy; and (iii) freedom for people to determine their own future.
Transnationalism and the migration–development nexus

At the global scale, both temporary and relatively permanent international cross-border movements have taken on more diverse and complex characters, with migrants undertaking what has come to be labeled “transnational migration” (Vertovec 2009). Conceived as a transformative spatial and mobile option for an increasing number of international migrants, transnationalism constitutes the social, economic, and cultural experiences and practices of living “in and between two worlds, or sometimes three or more” (Conway and Potter 2009, 227). Over time and through their life course, transnational migrants undertake repetitive mobilities, circulate regularly, maintain close global social contacts with family and friends, and exchange information, knowledge, monetary remittances, goods in kind, technologies, and innovations via information technology and other communication media, or personally through visits. Such mobilities, exchanges, and circulations can contribute in significant ways to livelihoods and other development outcomes at the household and community scales.

Serving to greatly alleviate poverty in the Global South, monetary remittances are overseas migrant donors’ transfers that meet the immediate needs of family recipients, but also contribute to savings and investment when sustained. With remittances being spent locally to buy goods and services, this has multiplier effects in the recipients’ communities. Remittances represent a direct investment in human capital as well as needed consumption, and, with further migration being financed as well, the reproduction of flows of remittances can be assured. Remittances can also be altruistic transfers to community and family to fulfill obligations, without repayment. On the negative side, however, remittances may create migration dependency, encourage the conspicuous consumption of luxury imports, and contribute locally to wealth inequalities, where among recipient communities the haves and have-nots are distinguished by their overseas connections and their family members’ propensities, not their own industry and work efforts while staying home (Newland 2007).

Diasporas, however, do contribute far more to development than just the monetary support they send home to families and communities. Diaspora donors’ collective flows are a major source of foreign direct investment (FDI), market development (both for exports and for the outsourcing of production), technology transfer, philanthropy, tourism, political development, and more intangible flows of knowledge, new attitudes, innovations, and cultural influences. Consequently, recognition of return migrants’ considerable stocks of social and human capital has developmental implications, as so-called social remittances can serve as vehicles for people-centered organizational management and capacity building in the Global South. Migration–development relationships such as these that occur at microscales, therefore, appear to show promise for migrant families and local communities, although regions and more geographically marginal areas may not be so fortunate (Newland 2007).

Sustainable development

The notion of sustainably balancing economic growth with environmental protection and resource conservation emerged out of growing concerns about the pace and scale of the destructiveness of modernization in the latter half of the twentieth century (see Adams 2009 for a detailed review). With their early roots in more radical perspectives on limits to growth and deep ecology, such concerns manifested themselves in mainstream development objectives with the
DEVELOPMENT

now famous World Commission on Environment and Development’s (WCED) Brundtland Report, *Our Common Future* (WCED 1987). At the core of the mainstream sustainability agenda was the notion that economic growth should be managed in a manner that facilitates development while ensuring that there will be sufficient resources available for future generations to achieve their own development objectives. Achieving sustainable development thus entails balancing growth with environmental and social considerations through technological, market, and institutional changes that improve the efficiency and distribution of resource use and access, incentivize environmentally friendly behavior, and protect vital ecosystem services (e.g., hydrological cycles) such that these are able to sustain and enhance socioeconomic development. At its core, however, such a view maintains a *weak* perspective on sustainability in that it does not fundamentally question the paradigm that growth equals development, but calls instead for incremental, largely technocratic, adjustments to the capitalist system.

In contrast, *strong* perspectives on sustainability challenge directly the often used conflation of development and growth as if they were the same process. Ecological economists and limits-to-growth proponents argue that it is untenable to pursue a capitalist hard-growth strategy, but that a soft-growth path should be followed instead, in which environmental and quality-of-life concerns are fully integrated and prioritized. *Hard* (quantitative, physical) *growth* involves an increase in the size and scale of consumption and production, intensifying the exploitation of natural resources and pressure on the natural resource base. Conversely, *soft* (qualitative, nonphysical *growth*) is based on improvement, social efficiency, and the qualitative “unfolding of potentialities” (Daly 1996). While the first type of growth is severely constrained by natural limits; the second would be potentially sustainable (Daly 1990).

Given the North/South divide, strong sustainability proponents argue for limits to growth and a global “steady-state economy,” but acknowledge that poverty need to be addressed for sustainability at the global scale to be possible. To balance the global account, therefore, Goodland, Daly and El Serafy (1992) propose that an increase in hard economic growth for the poor be offset by a corresponding decrease for the rich, who predominantly reside in the Global North. Once the offset is rectified, a steady-state sustainable system can then be maintained through economic and ecological tradeoffs that carry the world “beyond growth” (Daly 1996).

**Beyond development? Postcolonial, postdevelopment, and antidevelopment thinking**

Since the 1980s, the shortcomings of modernization theories, as abstract and ahistorical depictions of developed and underdeveloped societies, have been contested through postmodernist and poststructural deconstructions and sociocultural critiques. Generally speaking, these approaches offer alternative humanistic interpretations of self, identity, and plurality of meaning, and reject the objectivity and scientific rationality of modernism and structuralism. In doing so, it critically examines the systems of knowledge that produce human and societal objects that need to be “developed” or that represent or lack “development.” Such perspectives privilege agency and subjectivity, and often ignore or decenter structural factors such as those derived from political-economic relations, hierarchies and other power relationships, markets, and class systems. This postmodern or poststructuralist turn in the social sciences inspired three
reconceptualizations of development theory and thinking – postcolonial, postdevelopment, and antidevelopment theory (see Power 2003; McEwan 2009).

Postcolonial theory is more than “after colonialism”; it seeks to critique the complexity and ambiguities of the imperial encounter in crucial ways. It shares with the postmodern critique a focus on relationships between the colonizers and the colonized, and challenges the legacies of Western Enlightenment discourses that promoted the prevailing modernist (and Western) histories of progress and advancement. Moreover, a postcolonial discursive lens contributes to the development discourse by enabling us to understand the material and cultural geographies of colonialism (and neo-colonialism) and the politics of identity and belonging as it varies across spaces (Power 2003). Postdevelopment and antidevelopment thinking share similar concerns and approaches, albeit with a more conscious intent to promote alternative means to achieve distribution, progress, justice, and equality through grassroots social movements based in the Global South (Escobar 1992). The wider point of these approaches is to overcome the constraints imposed by the mainstream development project and to develop alternative, egalitarian, and perhaps radical means by which communities might prosper and progress. As Escobar argues (1992, 48–49):

> In the long run, it is a matter of generating new ways of seeing, of renewing social and cultural self-descriptions by displacing the categories with which Third World groups have been constructed by dominant forces, and by producing views of reality which make visible the numerous loci of power of those forces; a matter of “regenerating people’s spaces” or creating new ones, with those who have actually survived the age of modernity and development by resisting it or by insinuating themselves creatively in the circuits of capital and modernization.

While many welcome such perspectives, others have criticized postcolonial, postdevelopment, and antidevelopment theories as being overly abstract, academic theorizations and as radical reactions to the hegemonic notion behind the modernization construct that a Western middle-class lifestyle is not at all a desirable or realistic goal for the majority of the world’s population (Pieterse 2000). Moreover, some have argued that any concern for the relationships between such theories and the political economy of development under contemporary globalization omits considerations of the historiography of European colonialism and US neocolonialism, and instead serves as umbrella terms for the experiences of decolonialization (Power 2003).

The contemporary impasse: neoliberal capitalism as antidevelopment

Just as Keynesianism prevailed as the dominant political, economic, and industrial mantra during the post-1945 era, the ensuing era of globalization that followed the deep recession of the early 1980s continues to be dominated by neoliberal capitalism and its free-market ideological faiths. The elimination (in 1999) of the venerable Glass–Steagall Act, devised in the Great Depression era to separate American commercial and investment banking, allowed for the centralization of financial institutions and for corporate mergers to flourish. Further deregulation and rollbacks allowed finance capital to tighten its control over global financial markets and to usher in an era of so-called casino capitalism (Stiglitz 2013; Strange 1986). There was extensive refashioning of the United States and global regulatory regimes, driven by US geopolitical expediency and leadership as an “American project” to foster globalization (Agnew 2005), such that the
international mobility of finance capital became free and unfettered. This enabled finance capital to play a dominant role in the global economy, and raised its rates of profit, while Global South and North capital stocks, sovereign funds, and financial profits were channeled into sustaining US and European fiscal and balance-of-payment deficits. The exponential expansion of this global financial credit economy created debt peonage within advanced capitalist countries and mounting debt burdens internationally. International debt combined with the absence of capital controls accentuated boom and bust cycles of debtor countries, leading to recurrent international debt crises which are still with us today (Verick and Islam 2010).

Despite its crisis tendencies, and the precariousness of an increasingly financialized global economy, neoliberalism persists, extending its global reach and scope, and remaining extremely powerful as an organizing logic for contemporary capitalism. Viewing the 2008–2012 global crisis and its aftermath from a progressive perspective, Peck, Theodore, and Brenner (2012) find that neoliberalism not only survived but that it subordinates collective, communal rights to the dominant power of market exchange and individualistic accumulation of wealth, even underwriting the justification for excessive militarism and warmongering (Amin 2007). For his part, Harvey critically condemns neoliberalism’s structural power as:

creative destruction, not only of prior institutional frameworks and powers (even challenging traditional forms of state sovereignty), but also of divisions of labor, social relations, welfare provisions, technological mixes, ways of life and thought, reproductive activities, attachments to the land and habits of the heart … It holds that the social good will be maximized by maximizing the reach and frequency of market transactions, and it seeks to bring all human action into the domain of the market. (2005, 3)

While the neoliberal era arguably represents a particularly problematic period in world history, it also highlights the continued failure of capitalism as a system to deliver development widely, justly, equitably, and in the broadest sense of the term. Reminding us of the repetitive flaws of capitalism’s economic outcomes, John Maynard Keynes observed that “The outstanding faults of the economic system in which we live are its failure to provide for full employment and its arbitrary and inequitable distribution of wealth and incomes” (1935, 372). In other words, capitalism still divides and conquers, extracts surplus from labor, and accumulates at the expense of those still too powerless to democratically negotiate for social justice and basic human rights. A radical overhaul of the system is past due (see Harvey 2010, 2014), one that can redirect national and global economies such that they are fairer, more just, and sustainable.

SEE ALSO: Dependency theory; Gender and development; Marxist geography; Migration: international; Modernity; Modernization theory; Participatory development; Postcolonial geographies; Postdevelopment; World-systems theory

Dedication

While writing this entry, my dear friend and co-author, Rob Potter died at the age of 64. Rob was an eminent geographer and the founding editor of Progress in Development Studies. As his ideas remain central to this assessment of the field of development studies, I feel it is only proper that this final work of ours be dedicated to his memory (Dennis Conway).
References


http://www.migrationpolicy.org/research/
circular-migration-and-development-trends-
policy-routes-and-ways-forward (accessed April
7, 2016).
Parpart, J., S. Rai, and K. Staut. 2002. Rethink-
ing Empowerment: Gender and Development in a
“Neoliberalism Resurgent? Market Rule After the
Great Recession.” South Atlantic Quarterly, 111(2):
265–288.
Press.
Potter, R.B., T. Binns, J. Elliott, and D. Smith.
2008. Geographies of Development: An Introduction
to Development Studies, 3rd edn. Harlow, UK:
Pearson-Prentice Hall.
In The SAGE Handbook of Geographical Knowl-
dge, edited by J.A. Agnew and D.N. Livingston,
595–610. London: SAGE.
Potter, R.B., D. Conway, R. Evans, and S.
London: SAGE.
London: Routledge.
Goals.” In The Companion to Development Studies,
Abingdon, UK: Routledge.
Sen, A. 1999. Development as Freedom: Human Capa-
Serra, N., and J.E. Stiglitz, 2008. The Washington Con-
sensus Reconsidered: Towards a New Global Gover-
Sparke, M. 2013. Introducing Globalization: Ties,
Tensions, and Uneven Integration. Oxford: Wiley-
Blackwell.
Stiglitz, J.E. 2013. The Price of Inequality: How Today's
Divided Society Endangers Our Future. New York:
Norton.
Blackwell.
Thirlwall, A. 2015. Essays in Keynesian and Kaldorian
UNDP (United Nations Development Programme).
New York: Oxford University Press.
of 2008–2009: Causes, Consequences and Policy
Responses.” Discussion Paper No. 4934. Bonn:
IZA.
Vertovec, S. 2009. Transnationalism. Abingdon, UK:
Routledge.
Capitalist Agriculture and the Origins of the European
World-Economy in the Sixteenth Century. New York:
Academic Press.
Wallerstein, I. 1980. The Modern World System,
vol. 2, Mercantilism and Consolidation of the European
World-Economy, 1600–1750. New York: Academic
Press.
WCED (World Commission on Environment and
Oxford University Press.

Further reading

World Bank. 2016. Migration and Remittances
Fact Book 2016, 3rd edn. Washington, DC:
Migration and Remittances Unit, World
migrationremittancesdiasporaiissues (accessed
March 17, 2016).
Developmentalism is a theory of economic development targeting developing nations that proposes the intervention of the state in economic and social institutions in order to achieve “development.” Developmentalism is an approach to understanding economic differences between countries based on the idea of a ladder of growth and progress (Taylor and Flint 2011). While developmentalism generally refers to a set of practices that involves the cultivation of a strong domestic market and local industry, and the development of social institutions amenable to capitalism, there is considerable divergence on the genealogy and application of the term.

Watts (2009) identifies four philosophical roots of developmentalism: Keynesian-style state planning, free markets and property rights, Marxist critique of capitalist accumulation, and development institutions’ agendas. It is generally agreed that contemporary forms of state developmentalism emerged after World War II, when aid to newly decolonizing nations was consolidated around the objective of addressing the “underdevelopment” of these states. New institutions, such as the International Bank for Reconstruction and Development, were directed to helping impoverished former colonies emulate the economic success of Western countries. Countries that have adopted varying strategies of developmentalism since the 1950s include India, Singapore, South Korea, Taiwan, Botswana, and more recently Thailand and Vietnam. Also known as developmental states, these states are characterized by their ability to regulate capital and class power to meet long-term social, economic, and strategic goals (Olds and Yeung 2004).

A model that best exemplified postwar developmentalist thinking was Walt Rostow’s Stages of Economic Growth. Countries move through a linear sequence of stages from “traditional society” through “take-off” phases before reaching the stage of “high mass consumption.” Advanced economies such as that of the United Kingdom are presumed to occupy the top of the ladder, while developing countries attempt to catch up from the bottom rungs by transforming subsistence agrarian societies into export economies. Modernization theories have been seriously challenged by Immanuel Wallerstein and other world-systems theorists. World-systems approaches view developing countries as locked in dependent relationships with industrialized countries, resulting in cycles of poverty and economic stagnation rather than economic advancement. Critiques have also been lobbied from postdevelopment and postcolonial scholars such as Escobar (1995), who argues that development aid perpetuates linear views of progress while maintaining unequal power relations between giver and receiver. Postcolonial geographers argue that developmentalist thinking is rooted in colonial projects and in the very language we use for describing countries and cultures (McEwan 2003; Robinson 2003). More recently, the spread of neoliberalism as economic doctrine through the mechanisms of the Washington Consensus of the International Monetary Fund and the World Bank has led some to argue that neoliberalism...
DEVELOPMENTALISM

has become the latest strand of developmentalism (Harvey 2005).

SEE ALSO: Aid; Development; Economic geography; Neoliberalism; Postcolonial geographies; Postdevelopment; States and development; World-systems theory

References


Diaspora

Elaine Lynn-Ee Ho
National University of Singapore

Diasporas are formed when people move from one part of the world to settle elsewhere but retain a deeply felt sense of belonging to their place of origin, often referred to as the homeland. Derived from Greek etymology, the label “diaspora” in its earlier usage referred mainly to populations that were exiled and dispersed in other lands, such as the Greek, Jewish, or Palestinian diasporas. But, over time, “diaspora” has been increasingly used to describe other types of emigration. Opinions differ on the criteria that should be used to delimit whether an emigrant population can be described as a diaspora. This is partly because of different disciplinary approaches to the study of diasporas but diasporas as an object of study are also protean-like and evolve dynamically over time. The journal *Diaspora: A Journal of Transnational Studies* was established in 1991 as a publication dedicated to this topic, attesting to not only its wider scholarly significance but also the proliferation of views concerning the study of diasporas. Several other journals exploring diaspora as a subject of inquiry have also emerged since then.

The political scientist William Safran (1991) suggested that a diaspora community would have experienced dispersion from or loss of a homeland, but its members retain a collective identity and desire to return to the homeland because they believe that they are not or cannot become fully integrated into their host society. For Safran, members of a diaspora feel a sense of ethno-solidarity with their homeland and are committed to its welfare. This rather narrow definition of diaspora, however, belies the variety of ways in which the label is applied loosely to different types of emigrant populations now. Robin Cohen, a prominent scholar in diaspora studies, proposed a broader typology of diasporas in his book *Global Diasporas*, which was first published in 1997. This influential book has been reprinted multiple times and translated into several languages. For Cohen, diasporas are characterized by the shared link their members have with a certain migration history and by their solidarity with those in the homeland. In the latest edition of the book, Cohen (2008) classifies diasporas in the following ways: (i) victim diaspora; (ii) labor diaspora; (iii) imperial diaspora; (iv) trade diaspora; and (v) deterritorialized diaspora.

Interlinking analyses of nation, diasporic identities, and homeland

Notions of belonging and collective identity situated in a national homeland are commonly invoked whenever the label “diaspora” is deployed. Diaspora revolves around imaginative and material links to a home and homeland, framed in terms of belonging to and identifying with a nation. These tenets of diaspora are capable of mobilizing emotive feelings of patriotism among diaspora populations. Critical race scholars argue that diasporas should not only be treated as a descriptive typology but also examined conceptually. Among them is Brah (1998) who suggests that the “homing desire” underlying diaspora demands critical interrogation from a historically informed perspective rather than
be accepted as a primordial condition. This is because the genealogies and social relationships underpinning diaspora contribute to subjectification and identity-making. These in turn create categories of inclusion or exclusion in conceptions of the nation and diaspora.

Another group of scholars in diaspora studies, influenced by the writing of cultural theorists such as Paul Gilroy (1993), emphasize how diasporas function as sites of hybridity, creativity, and even resistance. Gilroy’s signature book, *The Black Atlantic*, argues that transnational bonds in the case of the African diaspora foment hybrid cultures that destabilize the hegemony of indigene claims. In contrast, other scholars dispute the hype of hybridity or resistance associated with diasporas. Historian Arif Dirlik (2004) argues that, inasmuch as diasporas can be celebrated for countering parochial framings of “race,” ethnicity, and nation, diasporic imaginaries are also vulnerable to reification and appropriation by chauvinist agendas. Scholars like Dirlik caution against essentializing the ethnic markers contained within diasporas.

For example, portrayals of Chinese capitalists as “flexible citizens” strategizing their immigration pathways in exchange for citizenship rights (Ong 1999) sideline the heterogeneous class backgrounds of members in the Chinese diaspora. Also contributing to distinct homeland identities within the Chinese diaspora are the geopolitical relationships of countries with majority (e.g., China or Taiwan) and minority Chinese populations (e.g., Southeast Asia). Likewise, what is commonly referred to as the South Asian diaspora is in fact made up of distinct homeland identities associated with the geopolitical entities of Bangladesh, India, Pakistan, or Sri Lanka. Within any of these South Asian diasporas, there are further nationality, religious, ethnicity, class, and gender divisions. Intersecting axes of identity such as those related to gender, class or nationality divisions thus differentiate between subgroups contained in diasporas. The intersectionality of these identities can reinforce domination or subordination, affecting resource allocation and life chances.

There is growing recognition of the internal divisions and contradictions within diasporas. Diasporas do not exist a priori but are constituted by the interplay of power relations internally and externally. The external influences that contribute toward how diaspora claims are mobilized by different groups at the community, national, and international levels are significant for understanding diaspora formation in space and over time. Brah (1998) proposes the term “diaspora space” to draw attention to not only border crossings by migrants but also their copresence with those that stay put. This underlines the multiple power relations between and within social groups in spaces of diaspora.

**Spaces of diaspora, territory, and extraterritorial reach**

Diasporas identify with a spatial center considered the homeland. The metaphor of “homeland” situates diasporic identities and belonging in actual or sought-after territorial space. Narratives associated with particular landscapes situated in the homeland can become intertwined with personal memories or family stories of migration. Migrants recreate home and homeland identities during their diasporic journeys through the use of oral histories, visual cultures, and material artifacts. The intersection of geographies of home and spaces of diaspora reflects contested and multiply refracted notions of identity, belonging, and diaspora. As such, nationally framed portrayals of diasporas can be complicated by not only the dual affiliations that diaspora populations have with their countries of origin and arrival but, in the
context of postcolonial migrations, wider colonial and postcolonial histories that also inflect on diasporic identities and belonging.

While the anchoring of diaspora in homeland space and the recreation of homeland in spaces of diaspora are already prominent themes in geographical research, the concept of diaspora also intrigues geographers because of the way it destabilizes the fixity of territory that has become ensconced with nation and nationhood. The dispersal of populations identified or identifying with a national territory evokes spatial imaginaries of deterritorialization, referring to how territory is decoupled from its presumably stable links with nation, identity, and community. Classical diasporas are often portrayed as enclosed social formations characterized by homogeneous culture and identity. But contemporary conceptualizations of diasporas are more likely to emphasize their fluid nature, internal tensions, and hybrid identities that in turn trouble homogeneous portrayals of diasporic space or territorial delineations. Diasporas can be studied both as a social condition and as a social process that plays out in space and across geographical scales and networks.

Inasmuch as the territorial imaginary associated with nation can be destabilized through diaspora formation and deterritorialization, reterritorialization takes place alongside deterritorialization processes. Examples are of political mobilization and lobbying by diaspora actors aimed at influencing developments in the homeland, or remittances sent to the homeland for family support or development purposes. These practices by diaspora populations reinforce the interlocking of nation and territory. While some of these activities are initiated by the diaspora, migrant-sending states are increasingly asserting extraterritorial claims on their emigrant populations. Their goal is to reinforce national belonging and to solicit contributions from the diaspora to the homeland. These state-led policies extending the reach of the nation-state over diaspora populations are collectively referred to as diaspora strategies.

**Diaspora strategies and geographical inquiry**

The “intrinsic and contradictory” (Dirlik 2004, 491) relationship between diaspora and nation is widely acknowledged today, but entangled with it are the impacts of globalization and late capitalism. Whereas the mobilization of diaspora populations for development purposes have long been associated with poorer countries that tap their diasporas for remittances or small-scale investments, today such state-led policies are just as likely to be implemented by economically advanced countries. Nation-states looking to gain a competitive advantage in the global economy are tapping on their diaspora populations for highly skilled human capital, international networks, and global capital.

While diaspora studies emphasize the homeland orientation of emigrants, transnationalism approaches recognize that migrants are simultaneously anchored in their countries of immigration and emigration. Transnationalism and diaspora can thus be read as an analytically distinct phenomenon. But migrant-sending states are enacting diaspora strategies to capitalize on the transnational attachments and mobilities of their emigrant populations. International institutions such as the World Bank and the Asian Development Bank, as well as migrant-receiving countries like the United States or European Union countries, also promote diaspora strategies as a means to further development in poorer migrant-sending countries. Their interest in using diaspora strategies to effect development elsewhere is prompted
partly by a global leadership role but also out of national interest. Promoting development in countries where their immigrant populations originate is a means of deterring further economic migration that could strain the social stability of the migrant-receiving country. The joint international and national interests in diaspora strategies have triggered a global dialogue for the knowledge sharing of best practices through international nongovernmental organizations like the International Organization of Migration and the Migration Policy Institute.

This policy trend has influenced the direction of geographical research on diaspora strategies in several ways. One approach critically interrogates the neoliberal ideology underpinning the “development” agenda of diaspora strategies (e.g., Larner 2007). Such analyses posit that new forms of global governance and subjectivities are culturally constructed through diaspora strategies that constitute certain ways of participating in the global economy. Another approach considers how political institutions and governance mechanisms evolve as migrant-sending countries deploy diaspora strategies to connect with their emigrant populations (e.g., Gamlen 2008). Citizenship is one such political institution that has become redefined spatially when membership, rights, and responsibilities are selectively extended beyond the national territory to diaspora populations. This reconstitution of citizenship creates sites where some emigrants are privileged over others, leading to the marginalization of certain diaspora groups and even contestations of the state’s portrayal of identity and belonging (e.g., Ho 2011). Diaspora strategies can trigger spaces of uneven development when diaspora strategies entrench capital and other accumulation in particular sites, resulting in greater dispossession elsewhere.

When migrant-sending states portray emigrant populations as diasporas that belong to them, the claim exerted by the nation of departure is privileged over the nation of immigration. In this way, diaspora strategies can have politically debilitating effects if they cast doubt over the attachment and commitment of immigrant groups to their countries of reception. Tensions between ethnic, national, and diasporic identities can be heightened by the demands of diaspora strategies when they are juxtaposed against the parallel onus of assimilation or integration borne by immigrants in their countries of reception. The following section expands this discussion by highlighting the geographical implications of overlapping national claims experienced by diasporas when migrants undertake transnational sojourning between countries of origin and countries of reception.

Multiple diasporic identifications and proliferating emigrant mobilities

Increasing migration flows and more complex geographical trajectories are likely to engender multiple diasporic identities among emigrants or their descendants as they journey between the different countries they simultaneously consider home. Geographers have studied transnational sojourning trends that indicate that, even though migrants embark on return migration, they intend to subsequently remigrate back to the immigration country they have left (or to another country). Diasporas are thought to long for return to their places of origin but during transnational sojourning there can be several forms of “return,” thus challenging the premises of diaspora identity and belonging. Migrants may have naturalized in the immigration country and developed legal and emotional attachments to it. As immigrants they were considered the diaspora
of the country of origin, but during return migration they adopt an additional diasporic identity and form the secondary diaspora of the immigration country they have left. They identify as the diaspora of more than one country, thus complicating nationally framed definitions of diaspora membership.

Such diaspora populations are able to capitalize on their transnational affiliations if the countries in question view their national interests as complementary, which may even trigger the coproduction of diaspora as both countries leverage their diaspora populations to mutual advantage. But, should the countries of origin and immigration have conflicting interests, then either or both of those countries can view their diaspora populations suspiciously and question where their loyalty lies. Proliferating emigrant mobilities present fruitful avenues for geographers to explore the tensions wrought when different migrant-sending states, whether the country of origin or country of secondary migration, assert overlapping claims in relation to what they both consider are their diaspora populations.

Also of interest to geographers is the reverse migration of diasporic descendants to their ancestral homelands. These studies illuminate the structural pressures placed on the children of immigrants to assimilate or integrate into immigration countries. Yet their ethnic identities, family migration histories, and transnational ties set them apart from mainstream society. In some cases, experiences of discrimination or nonbelonging in immigration countries precipitate nostalgia for the ancestral homeland and a desire to be among coethnics whom they perceive to be more similar to them. In other cases, diasporic descendants have been compelled to leave their settlement countries because of conflict situations or ethnic persecution. Ancestral homelands that prioritize shared ethnicity and kinship obligations could implement ethnically privileged migration policies to give coethnics in distress preferential rights to entry and resettlement. Ethnically privileged migrations may even intersect with development agendas, such as manifested by diaspora strategies that target coethnics and diasporic descendants abroad.

When diasporic descendants relocate to the ancestral homeland they are presumed to cohere more easily with the resident population because of assumed ethnic similarities, but this is usually not the case. Diasporic descendants are likely to find that the expectations they had of life in the homeland are different in reality. This is because the ancestral homeland has changed compared to the memories and family stories shared with them by an earlier generation of emigrants. Moreover, when diasporic descendants live in proximity with coethnics in the ancestral homeland, they become conscious of differences in habits, mindset, language, or accent. Likewise their coethnics may treat them as insiders because of their shared ethnicity but simultaneously view them as outsiders because of cultural differences. This could bring about another layer of social segregation that multiplies the spaces of exclusion experienced by diasporic descendants even if they are in their purported ancestral homelands.

Diaspora is, as Clifford (1994, 302) describes it, “a traveling term.” Not only does it denote human mobility; it is also a practice and concept that takes on certain emphases in different spaces. Intrinsic to the social processes and relationships constituting diaspora is its relationship to the nation and state. Diasporas can be protective of their perceived homeland identity, and the homeland, equally, can mobilize the diaspora for the benefit of the nation. Geographical approaches to the study of diasporas thus contribute to identifying how diasporas change the meanings associated with places and territoriality, and further the impact of such spatial concepts on diaspora formations.
DIASPORA

SEE ALSO: Citizenship; Identity; Intersectionality; Migration: international; Nation-state; Postcolonial geographies; Race and racism; Territory and territoriality; Transnationalism

References

**Difference**

Deborah P. Dixon  
*University of Glasgow, UK*

As with all manner of key concepts, difference has been used to underpin a wealth of geographic thought, analysis, and writing while eluding precision, in terms of a circumscribed definition, and specificity, in terms of what it appropriately illuminates. What is more, this geographic engagement itself rubs up against explorations and deployments of the term within other disciplines in the natural and social sciences, as well as the arts and humanities. Difference is itself differently understood as an object of analysis, but also (and often implicitly) as a means of sorting and ordering other objects such that they lend themselves to analysis. As such, these differences are themselves indicative of much broader conceptual and methodological shifts in geography and beyond.

Prior to the emergence of spatial science in the mid-twentieth century, difference as a concept had largely been wrapped up in regional geography. Geographers sought to lay out the singular collection of spatially overlapping physical and social phenomena that, taken together, both differentiated regions from each other, and allowed an understanding of their holism. With the arrival of spatial science – an approach that very much took the world as there for all to see – an objectivist gaze was coupled with an abstract, isotropic, and rectilinear spatiality, one that relied on both subject/object separation and the unambiguous location of objects of analysis. While there were certainly diverse understandings of space and spatiality at work in spatial science (e.g., topographic versus topological), this approach became firmly associated with difference as a matter of both variation across a gridded, Euclidean space, and the positing of objects and events as discrete entities. For critically minded geographers, both of these assumptions have been problematic. Numerous lines of inquiry have sought not only to bring to light an underlying essentialism and transcendentalism within spatial science, but also to proffer alternative understandings of what difference is and how it matters.

Though difference has certainly been implicated in geographic inquiry for centuries, its importance as an analytic concept was subordinate to human–environment relations. Spatial science was the first geographic approach to place difference squarely at the forefront of analysis as a means of sorting and ordering phenomena in the form of objects and events. Here, objects/events are distinguishable from one another in ways that permit scientists to measure variation between them, and to construct elaborate systems for analyzing and explaining such variation. This distinguishability is itself made possible through a particular understanding of difference as discreteness. That is, objects/events are presumed to have an essential core of attributes (size, weight, color, duration, intensity, etc.) that afford them stability and fixity, and on which a host of powerful scientific analyses can be constructed and carried out. These attributes are crucial constitutive components of the object/event under analysis, but, importantly, their presence and intensity can be compared and contrasted between phenomena because they lend themselves to being observed by an objective observer and
measured according to universally applicable criteria. Within spatial science, this approach to difference is most explicit in the construction of data matrices, wherein various measured attributes – systematic, temporal, and spatial variables – are taken to be representative aspects of the real-world object/event under investigation. It is also apparent in topographic accounts of objects/events that seek to map the putative influence of discrete objects on each other as relations that unfold across space. The methodological rationale for this approach to difference, wherein everything has its place, is rigor, which guards against any ambiguity that might undermine scientific analysis of cause and effect.

The most significant implications of this ontology revolve around the presumed discreteness of objects/events. At one level, such sharp-edged distinctions permit the separation of academic disciplines into categories charged with understanding the systematic (e.g., biology or economics), temporal (history), and spatial (geography) characteristics of objects/events. At another level, these distinctions mean that scientists do not have to approach the world as an undifferentiated, chaotic, and interconnected jumble that requires that each new study reconceptualize that reality. By assuming that objects/events are distinguishable, and that their characteristics are divisible into the three dimensions, scientists can direct most of their conceptual efforts into devising the most appropriate concepts and measurements for these dimensions.

Although spatial science still underpins a great deal of geographic inquiry, particularly with regard to the earth sciences, human geography has been radically transformed by a more critical accounting for scientific knowledge and associated practices, combined with a philosophical interest in the crisis of representation. That is, and in contrast to the above, Marxism, feminist theory, postcolonial theory, and poststructuralism have largely worked to a pluralist epistemology, wherein the distinctions drawn between objects/events are not taken for granted as the starting point for analyses, but are positioned as socially constructed. None of the distinctions – or differences – between individuals and groups, made and natural objects, types of experience, and aspects of meanings, are understood to be naively given as unmediated parts of reality. These are positioned, rather, as compartmentalizations that allow the world, as well as the place of human beings within it, to be made sense of. Compartmentalization relies on the use of rigidly fixed and policed boundaries around the meaning of objects/events, such as natural and cultural, masculine and feminine, human and animal, and the putative relations between them.

Geographers working within several conceptual approaches have developed numerous lines of inquiry that address the larger discourses within which such distinctions are posited, brought to bear, and challenged. In the process they have drawn inspiration from a range of social theorists from Roland Barthes and Michel Foucault to Jean Baudrillard and Jean-François Lyotard, each concerned, although with varying emphases and intent, with the production of knowledge. Although discourses are enabled and reproduced through language, to geographers interested in the making of distinctions “discourse” is a term more complicated than its everyday use as “mere words,” for it refers not only to the processes of categorization but also to everyday social practices – from raising children to dancing – that, like language, are imbued with meaning and hence also signify something about the world. Through discourse human subjects come to understand where things fit in the world, literally and figuratively. They also come to comprehend the relationships between
cultural motivations behind these, have been taken up by geographers working across the discipline’s subfields. And, in the process, these concerns have led geographers to consider the utility of literatures on psychoanalytic issues, emotion, and affect, in conjunction with methodological treatments of reflexivity and situatedness.

This move from difference as a given ordering of the world to difference as socially constructed has had three broad implications for geography. First, and at the level of the discipline’s objects of analysis, constructivism suggests that geographers scrutinize the ways in which discourses, and in particular binary logic frames, predetermine their conceptualizations of objects of analysis. Indeed, it has been pointed out that binary thought seeps through epistemology and into the discipline’s ontological categories, as can be seen in the distinctions between social and physical phenomena, or between individual agency and sociospatial structures, just two of the more prevalent. Second, and at the level of geographic thought more generally, these two approaches provide contrasting views of the history and development of the discipline. On the one hand there exists a “paradigmatic” view that posits distinct bodies of inquiry shaped by different philosophical frameworks, methodologies, substantive domains, research questions, and methods. Paradigms are predicated on the assumption that they can be compared and contrasted – for example, environmental determinism versus regional geography – as if they were discrete bodies of thought. And, by implication, paradigmatic thinking also assumes that there exists a body of thought external to others, from which the historian of geographic thought might evaluate the rise and fall of paradigms. On the other hand, a constructivist approach sees these as embedded in a relational field of methodological constructs that have been sorted and afforded significance, such as pattern versus
DIFFERENCE

process, science versus art, truth versus fiction, nomothetic versus idiographic, objectivity versus subjectivity, quantitative versus qualitative, explanation versus understanding, and mapping versus writing. The borders between these concepts are not given, but are maintained by various discourses that center paradigms and their competitors. In poststructural accounts, for example, modern-day geography emerges as a metaphysical enterprise whereby seemingly secure ontological categories, such as presence, essence, existence, cause, origin, substance, subject, truth, God, and man, are pivotal in the formulation and development of geography’s paradigms (Doel 1999). Third and finally, since compartmentalization relies on excluding those objects/events that do not fit easily into place, the very construction of such an edifice is made possible by what lies outside of the category. Thus, as many geographers working with post-colonial theories argue, the outside, or other, is always already constitutive of the inside, leaving its trace within the boundary of the category (Nash 2002). Certainly, in many human geographic syllabi, a key pedagogic component is now to make students aware of the need to acknowledge otherness in the formation of often taken-for-granted social categories, such as “European” or “white,” the everyday operation of inclusion/exclusion, and the ethical appeal of diversity (McLeay and Chalmers 2011).

Taken as a whole, the drive to query difference as socially constructed has led geographers to engage in all manner of epistemological debate, and to practice a linguistically driven, power-sensitive deconstruction of meaning that has often been referred to as the “cultural turn.” And the dangers of assuming difference to be a matter of the essential (discrete, static, and fixed) attributes of phenomena are well signposted. Over the past two decades, however, the discipline has also been witness to a variety of efforts to ontologize difference in a manner that does not fall back into such an essentialism. That is, efforts have been made to pay attention to difference not as a social construct, but as a fundamental part of the world, much of which lies outside of human purview. What is at stake here is the charge of transcendentalism, whereby the criticism is made that geographers, including those working within the critical traditions noted earlier, have tended to prioritize analyses that simply ignore (as inaccessible or as unimportant) the workings of a world outside of human understandings of it.

Confronting this ontological challenge, geographers inspired by the work of Derrida, Gilles Deleuze and Felix Guattari, and Bruno Latour especially have mapped out an immanent ontology bereft of socially constructed, organizing concepts such as scale. Immanent here refers to a self-organization, as opposed to the work of some form of external, organizing agency. This ontology is characterized by what has been termed a “radical alterity,” wherein phenomena are understood not as discrete, static, and fixed, or as the end products of a process, but as existing, rather, in a continual process of becoming (Dixon, Woodward, and Jones 2009). Becoming is made possible by the entanglement of phenomena with a host of others, each affecting and being affected by the rest. What is of interest in such an approach is not the current material configuration of an object/event, which might be summed up as a series of attributes, but the relations that bind and cleave phenomena. These relations are in turn understood as a reference for various forces that, with varying speeds and intensities, constitute (as opposed to direct or emanate from) the material composition of phenomena. Difference thus becomes a matter of differentiation, wherein phenomena are refashioned time and again through their connections with other phenomena.
Geographers interested in difference as an ontological condition of the world have explored a variety of concepts, such as network, assemblage, folds, and strata, that more usefully capture this sense of connectivity between phenomena and their ongoing co-constitution. Each of these takes space not as a dimension across which variables can be measured and events unfold, but as a fundamental component of the objects/events in question. Often referred to as a “counter-topographic geography,” spatiality becomes here a capacity, alongside other capacities, that helps to constitute phenomena and which, as those phenomena are refashioned, is also caught up in a continuing process of differentiation. Geographers have also turned to concepts such as queering and hybridizing to emphasize something of the momentum that takes place as differentiation occurs. Nevertheless, although there is a shared concern to understand how phenomena are refashioned, politically and ethically, there are debates over how refashioning is to be accomplished, insofar as the theories brought to bear do not add up to provide a correct or universally relevant accounting for an immanent ontology. Such conceptual approaches must be considered, rather, as in a productive dialogue that highlights some continuing problematics for geographers, such as the manner in which a human/nonhuman boundary is composed and blurred, and the nature of relations that compose as well as undercut such boundaries, but that also offers up new lines of geographic inquiry. Currently, there are three broad areas of debate.

First and foremost, while there is a concern in the immanent approaches noted above to emphasize the self-organizing of phenomena, and to map out how self-organization is materially differentiated, there is also a concern to recognize the various limits of this movement. For some geographers, the danger inherent in such approaches is that they lend themselves to an understanding of phenomena as always connected, and always fluid and malleable, such that they are constantly in the process of transforming into the new. Taking a cue from Deleuze’s (1994) notion of “difference with repetition,” these geographers have sought to emphasize in turn the blockages, fractures, displacements, and antitheses that characterize ontology time and again. To date this effort has been particularly manifest in work that situates geography as the analysis of sites (Woodward, Jones, and Marston 2010) that are considered dynamically composed aggregates constituted from emergent and self-organizing processes, and that can be “mapped” according to their own internal logics of composition. In disciplinary terms, spatiality is one such logic. We should not be surprised, therefore, to find that the spatialities of sites rub up against each other; indeed, this polycentrism may well be a defining characteristic in understanding how a site works.

A second and related debate concerns how the “becomingness” of phenomena is to be understood. Some geographers have alleged a tendency, despite the emphasis on phenomena as self-organizing, to read objects/events as the end products of a combination of processes. The worry here is that in attending (often implicitly) to the connections between phenomena as processes there is an accompanying assumption that their sum operation, which must extend beyond the materiality of the phenomena in question, are generative of a material recomposition. For critics, the holism underpinning such an approach is an unwarranted distribution of causality. In part, this is because, it is argued, this conflation of process with causality glosses over the “thingness” of the very phenomena under scrutiny. Often resorting to “object-oriented” philosophies to work through what is taken to be a backdoor transcendentalism, geographers...
have emphasized differentiation as the work of “force-full” phenomena (Shaw and Meehan 2013). While scholars may speculate over the capacities of such phenomena, there is simply no final accounting for such capacities; a lingering reticence must be accorded the objects/events that draw so much attention.

Third, and taking the emphasis on the materiality of phenomena, and their connections (and nonconnections) noted above, there is considerable debate as to how the human subject is to be constituted in such analyses. Geographers working under the rubric of a posthumanism have looked to illustrate the problematic character of differences that promised to render the human subject unique – such as reason, sense-making, and even artistry – and have emphasized the transfers of energy and matter between and among bodies and environments in the form of, for example, affect. The question remains, however, as to whether certain capacities are to be accorded the status of a fundamental human/nonhuman distinction (Dixon, Hawkins, and Straughan 2012).

Regardless of such conceptual debate, future work on difference is likely to track back and forth across the iterative construction of difference, and the ontologizing of difference, in the pursuit of new lines of inquiry regarding how particular phenomena are composed and afforded significance, and how they work both materially and discursively within a particular context. In the process, one concern that has so far remained submerged and will no doubt come increasingly to the fore is the tendency for geographers working with and on difference to assume that doing so is an inherently productive endeavor. Long skeptical of the political inequalities forged and maintained by homogenization, stasis, and sameness, geographers have turned time and again to difference as a means of querying and even disrupting such dynamics.

In large part, this normative impetus explains why difference has remained a key concept in geography. And, yet, there remains much more reflective work to be undertaken on the question of why critical analyses should help to proliferate difference as an output of the research process; and, moreover, why the difference posited by the inaccessible is the most appealing of all.

SEE ALSO: Critical geography; Cultural turn; Feminist geography; Poststructuralism/poststructural geographies; Social constructionism

References


Further reading


Digital divide

Dorothea Kleine  
*University of London, UK*

Sammia Poveda  
*United Nations University Institute on Computing and Society, China*

**Digital divide/digital divides**

The term “digital divide” has existed for some decades and was perhaps most famously defined by Pippa Norris in 2001 as “a multidimensional phenomenon encompassing three distinct aspects. The global divide refers to the divergence of Internet access between industrialized and developing societies. The social divide concerns the gap between information rich and poor in each nation. And finally within the online community, the democratic divide signifies the difference between those who do, and do not, use the panoply of digital resources to engage, mobilize, and participate in public life” (Norris 2001, 4).

Journalistic shorthand sometimes refers to the so-called haves and have-nots of the digital world. This binary is, however, empirically and conceptually flawed, first, because access and use opportunities differ for different digital media and 2.9 billion people now have some form of access to the Internet (ITU 2014a). Thus many people’s reality situates them somewhere on a digital technologies continuum. Second, some “not-having” is based on a deliberate choice by users and can thus be better described as “not-choosing-to.” Therefore, instead of “digital divide” (singular), “digital divides” (plural) has been proposed (Warschauer 2004) and this is conceptually a more accurate term.

**Identifying the gap**

At the 2003 UN World Summit on the Information Society (WSIS) in Geneva, the member countries endorsed a set of 67 principles, indicating their commitment to “turning this digital divide into a digital opportunity for all, particularly for those who risk being left behind and being further marginalized” (Principle 10, WSIS 2003). Namely, the WSIS declaration identified those at risk as those living in rural areas or marginalized urban areas, poor, elderly, or indigenous people, women, people with disabilities, “marginalized and vulnerable groups of society, including migrants, internally displaced persons and refugees, unemployed and underprivileged people, minorities and nomadic people” (Principles 11, 12, 13, 14, 15, and 16, WSIS 2003).

Institutions such as WSIS view “universal, equitable and affordable access to ICT infrastructure and services” as a central development objective for all stakeholders building the information society (WSIS 2003). As a consequence, universal access has become a key policy objective in most countries’ national ICT strategies, and in some countries, such as Finland, France, and Estonia, legislation has been passed which makes access to the Internet a basic right.

However, social scientists have frequently pointed out that the digital divide per se is nothing new but often another manifestation
of existing social inequalities along education, gender, income, class, or ethnicity lines. The relationship with age is far more complex, as many studies have shown that young people are frequently skilful adopters of new digital technologies if they have access to them.

The digital divide is measured, at a relatively basic level, as numbers of people with access to the Internet or a mobile phone, households with Internet access, or subscribers to mobile phone providers, captured by counting subscriber identity module (SIM) cards or devices. Yet, this is frequently criticized for not taking into account that many people with low incomes hold multiple SIM cards and that increasing numbers of digital devices in households will soon make the counting of devices problematic as an indicator. At international level, there have been longstanding efforts to measure ICT and country development and trace the digital divide (see Figure 1). The UN’s International Telecommunication Union (ITU) has developed the ICT development index (IDI) which has evolved from previous indices developed by ITU, Orbicom–UNESCO (United Nations Educational, Scientific and Cultural Organization), and the United Nations Conference on Trade and Development (UNCTAD) and has been published annually since 2009. The IDI relies on 11 indicators across three components: access (lines and subscriptions), use, and skills (literacy, secondary and tertiary education enrolment). Other agencies collect data on specific themes; for example, UNCTAD collects data on “information economy” indicators such as the relative size of a country’s ICT sector. The data have significant limitations, and yet it is through these statistics that the “digital divide” is most commonly measured.

Beyond the indicators and indices, however, there are theoretical debates about the nature of access and which dimensions should be taken into account. Some of these dimensions are not directly visible in the indicators.

Dimensions of access

Access can be conceptualized in three dimensions: (i) availability of Internet connectivity, (ii) affordability of access, and (iii) skills needed for use (Gerster and Zimmermann 2003). Other aspects include having content that is relevant, meaningful, and in a language that the relevant user group can read (Warschauer 2004); being
able to create content (Van Dijk and Hacker 2003); and having access to content without censorship or paywalls causing a “data divide” (Gurstein 2011).

Other axes of exclusion affect other dimensions of access. Regarding gender, how different members of the family “ought to use their time” and social norms on the use of space may impact on girls’ and women’s ability to use the Internet (Kleine 2013). Studies have explored key cultural factors that may be sustaining the gender gap in ICT or mobile phone access, ownership, and usage. For instance, Buskens and Webb (2009) indicate that male relatives may be surveilling/prohibiting the ICT/mobile use of their wives and daughters. According to the 2012 GSMA mWomen report, “82% of married BoP [base of the pyramid] women who own mobile phones [agreed that] ‘it makes my husband suspicious’” (3).

Initiatives to combat the digital divide

Initiatives have evolved along with the pace of technological change and the understanding of the divide. A new field of research and practice, ICT4D (Information and Communication Technologies for Development), emerged to explore the use of ICTs in development work (Unwin 2009). Typical initiatives in the 1990s and 2000s were telecenters, or communal Internet access spaces, for example in schools or libraries, which provided access at subsidized rates or free of charge. Efforts here are ongoing, however, because as home and mobile Internet access increases (see Figure 2), telecenters have had to reinvent themselves as multipurpose community hubs which combine the access to information and communication with services and advice. This is provided by intermediaries and can in some cases create

Figure 2  ICT-related indicators globally over the last 15 years. ITU 2015, 2. Reproduced by permission of ITU.
DIGITAL DIVIDE

opportunities for collective empowerment, which is an aim stressed by the Community Informatics movement.

Another front for digital inclusion has been IT training, or more in-depth, critical digital literacy training. Although IT skills have instrumental value, researchers have been emphasizing the need for critical thinking to promote meaningful uses that support people’s own empowerment. On content, there are organizations such as the Wikimedia Foundation or the World Wide Web Foundation, who campaign to keep the Internet publicly accessible, while human right groups, such as FreedomInfo and the Open Rights Group, campaign against government censorship.

Diverging visions for the future of the Internet

Basic connectivity is improving. Technological innovations such as the battery-based Internet router BRCK by Ushahidi offer new possibilities, especially for rural communities that still lack basic electricity. Also, in 2009, underwater fiber-optic cables skirting the African continent docked at the Kenyan coast. According to Graham and Mann (2013), this process is producing significant effects in “cost, speed, and the ability to access information and communicate” in East Africa, due to a combination of technological improvements and people’s powerful optimism which impacted policies, practices, and expectations. However, structural factors governing national economies and the global trade system continue to limit the transformative power of such access.

While Internet backbone lines will still be cable-based, Internet use is increasingly taking place on mobile devices such as smartphones and tablet computers. Mobile applications, such as Facebook Zero, provide optimized speed and zero data charges as a way to attract more users in low-income countries. These initiatives, and a reduction in mobile data costs in low-income countries, would make it likely that first-time Internet users in these countries encounter the Internet on mobile devices rather than via a stationary computer.

Globally, the current prevailing business model allows free access to content in exchange for collecting and exploiting user data for commercial purposes. Commercial search engines (such as Google or Bing), email or social media platforms (such as Facebook and Twitter), and others allow easy and simple ways in which ordinary users can generate and share content, which then is held and regulated by the platform. The risk is that most users lack critical understanding of, or feel helpless to resist, the imposed data gathering, as well as surveillance by firms or governments.

The digital divide debate, which started as the question of how the poor might gain access to the Internet, is increasingly overshadowed by the global struggle between different visions of the future Internet. The discussions around open access to data and content online are set to continue, and have received a significant impetus by some governments committing to open data policies for their own work and government research councils, for example in the United Kingdom, making open online access to research outputs compulsory from 2016. Open access and digital rights activists fear that, if a vision of the internet as a largely privatized and commercialized space, rather than as a global public commons, prevails, then, by the time some of the last currently digitally excluded people gain access, the Internet itself will be much impoverished.

SEE ALSO: Development; Digital Earth; Information and communications technology; Technology and development
References


Gurstein, M. 2011. “Open Data: Empowering the Empowered or Effective Data Use for Everyone?” First Monday, 16(2).


Further reading


Digital Earth

Changlin Wang
Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences

Digital Earth is an evolving global initiative in which spatially referenced information – historical, present, and modeled – is integrated to create a three-dimensional representation of the Earth’s physical and social environments for use by researchers, decision-makers, and citizens (Craglia et al. 2012). Digital Earth extends beyond the visualization-centric virtual globes of the early twenty-first century, for example, Google Earth, to include connected infrastructures and participative information resources needed for people to compute, manage, and communicate big Earth data. The Digital Earth concept is promoted and developed globally by public, private, and third-sector organizations, usually within the contexts of sustainability and earth system science.

The concept of a computer-driven, virtual globe is not a recent development. Buckminster Fuller’s Geoscope is an early example. In his 1962 book Education Automation, Fuller described “a large two hundred-foot diameter (or more) lightweight geodesic sphere ... a miniature Earth. Its entire exterior and interior surfaces will be covered with closely packed electric bulbs, each with variable intensity controls” (Fuller 1962). The lights on the globe, controlled by computers, could display a variety of data, including simulated “consequences of various world plans.” The concept has also been described in science fiction. In the novel Snow Crash, Neal Stephenson imagines a digital globe “about the size of a grapefruit, a perfectly detailed rendition of Planet Earth, hanging in space at arm’s length in front of his eyes ... It is a piece of CIC software called, simply, Earth. It is the user interface that CIC uses to keep track of every bit of spatial information that it owns – all the maps, weather data, architectural plans, and satellite surveillance stuff” (Stephenson 1992). He describes it as having very high resolution with an ability to zoom in on specific areas.

Former vice president of the United States, Al Gore (1992), first proposed the idea of setting up a Digital Earth program in his book Earth in the Balance: Ecology and the Human Spirit. It would help organize and interpret the enormous volumes of Earth observation data, turn raw data into understandable information, build a new global climate model capable of receiving data from different sources, and improve the quality and usefulness of the models by comparing predicted results based on information from known climate records and what has actually happened. The concept of Digital Earth was then popularized by Gore’s speech, “The Digital Earth: Understanding Our Planet in the Twenty-First Century” (Gore 1999). He defined Digital Earth as a multiresolution, three-dimensional representation of the planet into which vast quantities of georeferenced data could be embedded. Gore’s vision predicted that with technological innovations and advances in information acquisition and processing, unprecedented amounts of information about our planet and a wide variety of environmental and cultural phenomena could be captured, stored, processed, and displayed digitally. He sketched a scenario where a young girl in an immersive environment would be able to
don a head-mounted device, and enter a virtual environment that would offer a “magic carpet ride” over the Earth’s surface, zooming to a sufficient resolution to explore and learn about the Earth and its human and physical environments from global to local scales over the Internet.

Although there is no singular definition of Digital Earth as it is understood today, Digital Earth in a very broad sense is an extensive integration of other advanced technologies including Earth observation, geographic information systems, global positioning systems, communication networks, sensor webs, electromagnetic sensors, virtual reality, and computational grids. Digital Earth can be generally understood as putting a representation of the Earth, and activities on the Earth, including temporal and spatial changes in the environment into a networked computer system, thus enabling information to be stored and transmitted as a service. This kind of digital replica of the planet can potentially reproduce all that is known about the real Earth. It could be presented as one of a class of services termed Digital X, a digital mirror of some complex system X, embedded in space and time, including a detailed representation of X, and software for acquiring, storing, manipulating, visualizing, and archiving that representation (Goodchild 2012). This notion of a digital world could be used to describe all digital types and forms – digital countries, cities, land, oceans, heritage, or events – showing a variety of applications of digital technologies for modeling, simulating, predicting, and representing the Earth globally and locally. The implementation of Digital Earth is highly associated with global and national information infrastructure development. The global, regional, and national development of Digital Earth may vary in different forms, such as improvements in spatial data infrastructure (SDI), and e-government initiatives. Furthermore, Digital Earth is seen as a potential global strategic contributor to scientific and technological developments and is promoted as a catalyst in finding solutions to international scientific and societal issues.

At the end of the last century, the vision of Digital Earth seemed almost impossible to achieve due to the limitations of the required high-resolution imagery, high-speed computer processing, broadband Internet, interoperability of systems, and – above all – data organization, storage, and retrieval. Today, however, the Digital Earth concept has come closer to reality. Vast, unprecedented quantities of spatiotemporal data are increasingly available at ever-accelerating rates due to the rapid development of Earth observation technologies, the increasing sophistication of new airborne and spaceborne sensors, new ground-based sensor networks, and crowd-sourced information. Achievements in computer technology, data processing, multimedia, and data storage and retrieval capabilities make it possible to realize 3-D presentations and descriptions of the Earth at multiple resolutions, scales, times, and categories.

Digital Earth initiatives

In 1993, the Clinton administration of the US government started construction of the “National Information Infrastructure” (NII), or as it was called the “Information Super Highway” for building an “Information Society.” In April 1994, President Clinton signed Executive Order 12906 for coordinating geographic data acquisition and access, establishing the National Spatial Data Infrastructure (NSDI) and paving the way for nascent Digital Earth initiatives.

In response to Al Gore’s 1998 speech, NASA and its federal partners initiated the Interagency Digital Earth Working Group (IDEWG) in 1999, in which 17 government agencies were
coordinated, including the National Oceanic and Atmospheric Administration (NOAA), Federal Geographic Data Committee (FGDC), US Geological Survey (USGS), Environmental Protection Agency (EPA), and National Science Foundation (NSF), to sustain development of applications for the Digital Earth enterprise. The initiative attempted to accelerate technological developments in standards, protocols, and tools toward the full realization of the Digital Earth vision.

NASA’s leadership for Digital Earth had waned by 2001, however, owing to a change in the US administration. The coordination of related activities was taken up by the Geospatial Applications and Interoperability (GAI) working group, a part of the US FGDC. Several outcomes were achieved in this three-year effort, including the development of interoperability standards for geospatial data sharing, the Web Map Service (WMS), the Digital Earth Reference Model (DERM), the NASA Digital Earth Testbed, the SRI International Digital Earth project, and a series of “Digital Earth prototypes,” such as Digital Earth alpha version applications for climate and weather tailored to four audiences: museums, education, governance, and journalism. An early academic example of Digital Earth research was the Alexandria Digital Earth Prototype Project (ADEPT), a “virtual learning system” developed at the University of California, Santa Barbara (UCSB), between 1999 and 2004. Further studies of Digital Earth at UCSB have since focused on analyzing Google Earth’s functionalities in comparison to the Digital Earth vision, and defining a Digital Earth system (Grossner, Clarke, and Goodchild 2008). Under the Obama administration, Digital Earth–related issues have been represented by the e-government initiative and US Group on Earth Observation with the vision of enabling a healthy public, economy, and planet through an integrated, comprehensive, and sustained Earth observation system.

Outside of the United States, Digital Earth concepts have attracted the attention of decision-makers, scientists, and entrepreneurs, and have been followed by worldwide discussions on this important topic. An INSPIRE (Infrastructure for Spatial Information in Europe) initiative was commenced and operated by 28 countries of the European Union to support environmental policy and overcome barriers affecting the availability and accessibility of relevant data (Annoni et al. 2011). The Digital Asia project was launched by Keio University in Japan to develop geospatial infrastructure for data sharing via the Internet and to provide people and communities in the region with easy access through the collaboration of potential agencies who participate on a voluntary or best-effort basis. Virtual Australia and New Zealand (VANZ) is an ambitious long-term vision to gradually create a post–Google Earth “system of systems” to monitor and simulate both countries’ environmental and social conditions. Scientists in China have paid much attention to building “Digital China” and have designed its framework.

China has been particularly active in its pursuit of Digital Earth, including the development of a digital country, digital cities, digital oceans, digital heritage, and other digital representations. It has proposed to build China’s digital resource platform by integrating and harnessing a vast variety of data, information, and knowledge by taking the whole of China as an object with geospatial coordinates as its base. “Digital China” would have features such as georeferenced layers of socioeconomic and environmental data; remote sensing imagery base maps; interface features such as virtual reality, seamless contiguity, smooth zooming, accurate querying, and measurability; and mechanisms for rapid or real-time updates and sharing of information. Digital provinces and cities have already been developed in China. “Digital Beijing” and “Digital Fujian”
DIGITAL EARTH

are examples of provincial Digital Earth projects initiated by the Beijing municipal government and Fujian provincial government, respectively.

In November 1998, the Geoscience Division of the Chinese Academy of Sciences (CAS) organized a high-level scientific forum in Beijing on “Environmental Resource Information and Digital Earth,” and then submitted “A Proposal for Developing China’s Digital Earth” to the state council in January 1999. The First International Symposium on Digital Earth (ISDE), themed “Towards Digital Earth,” was initiated by CAS and held from November 29 to December 2, 1999, in Beijing, thereby providing a venue for extensive international cooperation in implementing the Digital Earth vision. Approved by some 500 participants including scientists, engineers, educators, managers, and industrial entrepreneurs from 20 countries and regions, the 1999 Beijing Declaration on Digital Earth called for close cooperation and collaboration between governments, public and private sectors, non-governmental organizations, and international organizations and institutions to ensure equity in the distribution of benefits derived from the use of Digital Earth in developed and developing countries. The declaration also proposed that a symposium should be organized and held biennially in rotating locations by interested countries or organizations.

Guided by this declaration, a nonpolitical, nongovernmental, and not-for-profit international organization called the International Society for Digital Earth (also abbreviated as ISDE) was established in May 2006 in China. ISDE principally promotes academic exchange, scientific and technological innovation, education, and international collaboration on Digital Earth in order to further enhance the understanding of Digital Earth’s evolution, to keep pace with the progress of science and technology, and to advance Digital Earth worldwide.

In March 2008, ISDE in cooperation with Taylor & Francis based in the UK, launched an official journal named the International Journal of Digital Earth. The journal was accepted by Thomson Reuters for the Science Citation Index Expanded in August 2009, and reached an impact factor of 3.291 for 2014, ranking it fourth among the 28 journals in remote sensing. It was the first journal dedicated to the field of Digital Earth, and has become one of the world’s leading journals in geospatial information.

At the ten-year anniversary of Digital Earth, a group of scientists gathered together under the Vespucci Initiative in 2008 to re-evaluate Digital Earth in light of the many developments in the fields of information technology, data infrastructures, and Earth observation that had taken place. The group published “Next-Generation Digital Earth” as a position paper inviting comments and feedback.

The 2009 Beijing Declaration on Digital Earth was another milestone for advancing Digital Earth, evaluating the progress of Digital Earth’s development and recommending the use of integrated technologies to construct Digital Earth applications in the fields of global climate change, disaster prevention and response, alternative energy source development, agriculture and food security, and urban planning and management. In November 2009, ISDE was accepted by the Group on Earth Observations (GEO) as a participating organization, so as to jointly work together on implementing a ten-year plan to build a Global Earth Observation System of Systems (GEOSS).

In March 2011, ISDE organized the “Digital Earth Vision 2020 Workshop,” the outputs of which were “Next-Generation Digital Earth” and “Digital Earth 2020: Towards the vision for the next decade,” respectively published in the Proceedings of the National Academy of Sciences (Goodchild et al. 2012) and in the International
Journal of Digital Earth (Craglia et al. 2012), providing a big-picture outlook for Digital Earth’s development.

By the end of 2013, ISDE had organized eight international symposia and four summits on Digital Earth worldwide; additional details about these events are available on the ISDE website (www.digitalearth-isde.org). In addition, nine Taipei International Digital Earth Symposia have been held since 2003 by the Digital Earth Research Center, Chinese Culture University.

Components of Digital Earth

Digital Earth is a framework comprising conceptual and technological components that together form a system of geographically linked research and applications in the physical and social domains of the Earth, a digital modeling platform to monitor, measure, and forecast natural and human activity on the planet, and a visualization of the world. It is at least three-dimensional, four-dimensional if a temporal monitoring component is added, and five-dimensional if scale is treated as a variable instead of a set of discrete steps.

The vast availability of digital resources around the planet is the most basic component of the Digital Earth paradigm. Specific key technologies needed to build a Digital Earth are: computational capacity, mass storage, satellite imagery, broadband networks, interoperability, and metadata. Numerous techniques such as image fusion and automatic image interpretation are also used to facilitate the realization of Digital Earth. The NASA Digital Earth Office has identified four components of the Digital Earth system, including digital resources, interoperability, tools and technologies, and applications. National spatial data and information infrastructures, and Earth observations are also fundamentals for building Digital Earth.

The interoperability of the Digital Earth system defines the standards, terminologies, and protocols for communication between heterogeneous data sets. People from public and private sectors using desktop geobrowsers are also a component of Digital Earth, communicating with an infrastructure of high-performance servers over a high-speed communication network. They collect data and utilize the other components to enable communities to better understand and model the Earth.

Digital Earth platforms

A Digital Earth platform is a system that embodies a vision of Digital Earth. Taking spatial information as a theoretical basis, Digital Earth platforms aim to build a manageable system for visualization, quantitative analysis, and simulation of the Earth. Virtual globes and their portals (geobrowsers) are specific kinds of Digital Earth platforms. Virtual globes allow interaction with a 3-D representation of Earth (or another world), and provide the user with the ability to freely move around in the virtual environment by changing viewing angle and position. Compared to a conventional globe, virtual globes have the additional capability of representing many different views of the surface of the Earth. These views may be of geographic features, man-made features or abstract representations of, for example, demographic quantities.

A geobrowser, first developed in 1993, is a client in a web browser or a stand-alone application that is capable of accessing georeferenced data and providing multidimensional visualizations to end users. The advent of the Digital Earth vision has advanced web mapping technology and has propelled geobrowser research to a completely new phase characterized by applications that strive for greater realism through the
use of 3-D globes and immersive environments. The Digital Earth conceptual framework model (Foresman 2008) consists of a foundation of standards, disciplinary content and applications, and the aggregation of the three pillars for sustainable development (economic, social, and environmental) through 3-D geobrowsers. In December 2002, the Chinese Academy of Sciences hosted the first International Workshop on Digital Earth 3-D Geobrowsers at the Institute of Remote Sensing Applications, and the second 3-D geobrowser workshop was convened at the University of California, Santa Barbara, in March 2003.

The emergence of Web 2.0 (including volunteered geographic information) feeds virtual globes with rich, user-controlled information, including images, videos, sounds, and written content. Virtual globes use a variety of hierarchical tiling structures known as discrete global grids (DGG) to enable rapid zooming, and tiles pre-computed on servers to avoid extensive local computation. Sophisticated level-of-detail management allows the field of view to be refreshed at video rates (Goodchild et al. 2012). DGG systems usually contain increasingly finer resolution grids that represent the Earth with a tessellation of nested cells designed to ensure a repeatable representation of measurements that are better suited to today’s requirements and technologies than our legacy coordinate systems that were designed for repeatable navigation and manual charting. For example, the PYXIS Digital Earth Reference Model is such a system for geometric coordination and digital indexing on an optimized discrete global grid. Another example is from GeoFusion, Inc., whose technology includes features such as a fast and versatile rendering engine, dynamic terrain tessellation on unique global grids, high-quality image and terrain processing, and an application programming interface (API).

On March 12, 2014, the Open Geospatial Consortium announced the formation of a new Technical Committee Standards Working Group to explore and propose terms for a standard to enable interoperability through the use of Discrete Global Grid Systems (DGGS), which would better serve the design of virtual globes used to interact with Digital Earth.

Generally speaking, Digital Earth systems could be classified into two types: scientific and commercial. Scientific systems mainly contain information about research from different fields and present analytical results to the public. NASA World Wind, released in 2004, is both a geobrowser for public-domain scientific data, as well as a platform for land, sea, air, and space mission operations. It provides graphic access to terabytes of imagery and elevation models for planets and other celestial objects including satellite and other data of the Earth, moon, Mars, Venus, and Jupiter, as well as astronomical data made available through the Sloan Digital Sky Survey. The Eyes on the Earth software of the NASA Jet Propulsion Laboratory allows users to travel in time and explore NASA satellite visualizations in 3-D. It presents recent data on air temperature, carbon dioxide, carbon monoxide, sea level, ozone, ice, and water. It follows NASA satellites and provides information about how they collect critical data about the Earth’s atmosphere, land, and oceans, and also provides views of Earth from space, with daily updates. NASA Hyperwall, first developed in 2002 by NASA’s Advanced Supercomputing Division, is now capable of rendering one quarter billion pixels. The system helps researchers display, analyze, and study high-dimensional datasets in meaningful ways, allowing the use of different tools, viewpoints, and parameters to display the same data or datasets. The CAS Digital Earth prototype system released in 2006 is a platform for conducting applied research in Digital Earth.
Digital Earth

Sciences and demonstrating results in a virtual reality environment (Guo, Fan, and Wang 2009). The Earth Simulator developed in Japan consists of a 640-node super-computer that is used in various fields of research including climate change projections and modeling solid Earth interior dynamics. The Earth Simulator has enabled scientists to run weather forecasts and climate projections at more than 1000 times the resolution of conventional simulations. In addition to the above, Australia initiated the BLUELink program for creating a global ocean forecasting system that provides information on ocean conditions to help manage Australia’s diverse area of maritime operations, and a Glass Earth project for using 3-D visualization and geological modeling technologies to unveil 1-km sections beneath the Earth’s surface. The new “UNCOVER” initiative is underway to help Australia’s geosciences community understand the Australian crust and its mineral systems beneath the regolith – the under-explored layer of unconsolidated rocks and soil covering about 80% of that continent.

Many commercial virtual globes are available today. Most of these globes share similar functionalities, although they differ in some features. Digital Earth systems with general applications are often free to the public; however, enhanced versions with more powerful functionalities must be purchased. Popular software has included Google Earth, Skyline Globe, Bing Maps (formerly Microsoft Virtual Earth 3D), ArcGlobe, Leica Virtual Explorer, Yahoo Maps, Marble, Unidata’s Integrated Data Viewer, Diginext’s VirtualGEO, EV-Globe, GeoGlobe, GeoBeans, and SuperMap GIS. Among these systems, the most well-known is probably Google Earth. Released in 2005, Google Earth’s innovative techniques for organizing and presenting data have enabled hundreds of millions of people of all ages worldwide to view and share scientific information about the planet Earth, without the need for prior GIS training or software purchase. It allows access to geospatial information, with the ability to show fine-scale features in high-resolution images. Its functionality for zooming from space right down to the street level, flying over terrain and structures, and showing a high level of detail – even individual people – is very reminiscent of the “magic carpet ride” found in the 1998 Gore speech. With Google Earth, people are able to explore the Earth, moon, Mars, sky, ocean, and 3-D buildings, as well as landscapes of the past with historical images. Google’s other widely used geobrowser product is Google Maps. The publication of the Google Maps API has enabled users to become producers of geographic information and contributors in the development of Digital Earth initiatives.

Many countries use virtual globes as their portal for managing national spatial data. In April 2007, Geoportail was developed on the basis of Skyline Globe and launched by the French government as its national geospatial information network and distribution system through a 3-D virtual globe platform. Although similar to Google Earth, it provides much more detailed information for France with higher resolution images than those provided by Google Earth. Such national geoportals can be found for Finland, Ireland, China, India, Luxembourg, Italy, Philippines, and several other countries. Each of these geoportals contributes to the development of their national spatial data infrastructure.

At a global level, the Group on Earth Observations developed a portal (GEOPortal) for GEOSS. The GEOPortal, an operable Digital Earth, offers a single web-based access point for users seeking data, imagery, and analytical software packages relevant to all parts of the globe. It connects users to existing databases.
and portals and provides reliable, up-to-date, user-friendly information that is vital for the work of decision-makers, planners, and emergency managers. It also allows users to search datasets or contribute to the portal’s resources and data discovery tools, and link worldwide communities practicing in the nine societal benefit areas defined by GEO: disaster, health, energy, climate, water, weather, ecosystems, agriculture, and biology.

These virtual globes have contributed enormously to the practical implementation of many features of the Gore vision, but whether Digital Earth has been realized through the achievements of various virtual globes has been questioned in the literature. Goodchild (2008) provides an analysis of the uses of Digital Earth with comparison of the content of Gore’s speech and the features of Google Earth. Google Earth has a passing resemblance to the vision of Digital Earth and has realized some functions of Digital Earth (Grossner 2007), however, it is only a simplification and approximation of Digital Earth. Virtual globes can “embed” information and media for visualizing data from various sources, but the mechanisms for collective discovery are largely absent. In geobrowsers there is a lack of ability to analyze and process integrated data, and there has been a failure to harness grass-roots efforts toward generating the “vast quantities of georeferenced data” that should be available in a Digital Earth. Furthermore, the functions of reconstructing past scenarios and predicting the future in Gore’s vision have not been realized.

Digital Earth prospects

Gore’s Digital Earth aimed to develop a single geoportal that would present an integrated view of everything that is known about the planet’s surface and near-surface, along with views of the past and predictions about the future (Goodchild 2012). The next generation of Digital Earth would be distributed, with multiple connected infrastructures based on open access and participation across multiple technological platforms to address the needs of different users (Goodchild et al. 2012). In addition, Digital Earth in five to ten years would have the following features: dynamic and interactive exploitation of the full range of information flows from sensors and people; synthesis of heterogeneous information; metrics of quality and trust of both data inputs and outputs; mechanisms for being more participative as people become geoinformation providers and analysts; ease of use with different levels of functionality available to different audiences; and a ubiquitous frame of reference with information to be continuously updated (Craglia et al. 2012). Newly emerging citizen science, such as crowd-sourcing and volunteered geographic information, encourages people to become actively involved in the acquisition of data on many aspects of Earth-related phenomena, such as disasters or disease outbreaks. The use of citizens’ data to complement official data on forest fires in Europe, for example, has already been undertaken (Craglia, Ostermann, and Spinsanti 2012). In the future, citizens will more likely improve their understanding of the forces that shape society, such as measures that ensure global environmental sustainability.

The literature on the future of Digital Earth also addresses ethics and education. Ehlers et al. (2014) developed a set of guiding principles for realizing Digital Earth: (i) All humankind must have unrestricted access to Digital Earth; (ii) Digital Earth must be developed to improve the welfare of humankind as a whole; and (iii) Digital Earth must always be “on” – immediately accessible, with known precision and accuracy, comprehensive, fully interactive, and available anywhere at any time. In addition, it is foreseen that a set of underlying principles, that
is, Digital Earth ethics, to guide the development, implementation, and application of Digital Earth should be proposed so that Digital Earth can deliver its anticipated benefits for all humankind. Digital Earth education has been put on the agenda in Europe (www.digital-earth.eu). A Digital Earth curriculum has been proposed to promote the Digital Earth vision to the general public, and the idea of Digital Earth as a “knowledge engine” makes it particularly relevant to education.

SEE ALSO: Discrete global grid systems; Geoportals; Interoperability of representations; Spatial data infrastructures; Virtual reality; Volunteered geographic information; Web-mapping services

References


A digital elevation model (DEM) is one of the most important spatial datasets in many geographical information systems (GIS). It is defined as an ordered or unordered digital set of ground elevation (spot height) for terrain representation. In literature, there are three commonly used terms related to this, namely, digital elevation model (DEM), digital terrain model (DTM), and digital surface model (DSM). The distinction among the three terms is not clear and universally agreed, but some common tenets may apply, as follows.

- A DEM is a “bare” land surface model, which is supposedly free of trees, buildings or other “nonground” objects.
- A DSM is an elevation model that includes the tops of everything, including buildings, treetops, and ground where there is nothing else on top of it.
- A DTM is a more generic term referring to a DEM with one or more types of terrain information, such as terrain morphological features, drainage patterns, and soil properties. When dealing with only one terrain information type (i.e., height), this is a DEM. Obviously, DEMs is a subset of DTMs (Li, Zhu, and Gold 2005).

Tasks of digital terrain modeling

A few major tasks can be identified in the study field of DTM, namely, terrain data capture, DEM generation, DEM manipulation, DEM interpretation, DEM visualization, and DEM application (Hutchinson and Gallant 1999). The objective of data capture, DEM generation and manipulation is to achieve digital representation of terrain. The ultimate outcome from these tasks is an error-free and “hydrologically sound” DEM at the right scale. The rest of the tasks, DEM interpretation, visualization, and applications, can be categorized in a general field of terrain analysis, which aims to derive terrain parameters, morphological features, and other terrain-controlled environmental factors from DEM. The overall relationship between the DTM tasks is illustrated in Figure 1.

Topographic data and topographic map

The foundation of a DTM is the topographic map. A topographic map is a tool for the transformation and dissemination of geographical information of a region, and has a long-standing and ascendant place in human history. One reason for this is that a topographical map is an image of the real world. Without leaving home, a person can read the map of a region of interest and understand its natural environment, society, economic conditions, and terrain. Another reason is that the natural and social conditions reflected in the topographical map can, through
the interpretation and discovery of numerical and graphic data, form a basis for inferring further information not directly displayed on the map.

Traditionally topographical data are acquired through field survey, by measuring horizontal and vertical positions. The outcomes of such field survey are typically represented as contour maps. In the digital era, the large collections of contour-based topographical maps naturally become the major source for generating a DEM, typically through time-consuming high-cost digitization processes. More advanced technologies have been developed to acquire digital elevation data, including digital photogrammetry, satellite remote sensing, LiDAR (a portmanteau of “light” and “radar”) and InSAR (interferometric synthetic aperture radar), and so on.

Digital elevation data acquisition techniques vary in performance, cost, time, and accuracy, with advantages and disadvantages in different aspects. A simple comparison between the techniques is shown in Figure 2.

**Figure 2** Comparison between different DEM data acquisition techniques (the color shows the speed of data acquisition, the darker the slower) (adapted from Zhou and Liu 2006).

**Digital terrain representation**

Whatever the data acquisition method employed, the resulting data are merely a set of discrete sample points (spot height). They are independent of each other without any inherent relationships among them. In order to fully represent the terrain surface, a model has to be established to define the relationships between these points so that a continuous surface can be formed in a 3-D space.

Three data models are commonly used to construct a DEM:

1. regular grid structures (grid, lattice, raster);
2. triangulated irregular networks (TINs);
3. contour structures.
Regular grid

With the regular grid data structure, a terrain surface is represented by a matrix of spot heights, assuming there are slopes between these spot heights, thus forming a continuous surface (Figure 3, Figure 4, Figure 5). The position of any location is specified by its relative position to the origin in terms of column and row numbers, thus the positional accuracy of a given DEM is directly linked to the grid resolution (Figure 3).

Regular grid data structure is currently widely used for DEMs. It is simple to structure and compute, facilitating the automatic generation of photogrammetric and remote sensing imagery. Its structural characteristics, such as its fixed resolution and regular arrangement of data points, however, limit its flexibility and accuracy in representing topography. At the same time, producing the grid with randomly distributed sample data (e.g., samples acquired by ground survey, analytical aerial mapping, and map digitization) requires interpolation technologies. This not only causes the loss of terrain features by the smoothing effect but also results in non-natural phenomena, such as pseudo-depressions, increasing the complexity of later applications.

Triangulated irregular network

Triangulated irregular networks (TINs) use non-crossing, nonoverlapping triangular networks to simulate topographical surfaces. The size and shape of the triangles depend on the position...
Figure 5  A continuous surface is formed assuming there are slopes among spot heights.

<table>
<thead>
<tr>
<th>Triangle table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id#</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>H</td>
</tr>
<tr>
<td>I</td>
</tr>
<tr>
<td>J</td>
</tr>
<tr>
<td>K</td>
</tr>
<tr>
<td>L</td>
</tr>
<tr>
<td>M</td>
</tr>
<tr>
<td>N</td>
</tr>
</tbody>
</table>

Figure 6  The TIN data structure.

<table>
<thead>
<tr>
<th>X-Y coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node#</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Z coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node#</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>11</td>
</tr>
</tbody>
</table>
and density of randomly distributed samples. The horizontal and vertical positions of the three vertices of a triangle determine the extent and orientation of a triangular facet in the 3-D space (Figure 6) and the interconnected facets then forms a representation of the terrain surface (Figure 7).

The advantage of a TIN model is that it allows stratified sampling of ground elevation. Where the terrain surface is flat, less sample points can adequately represent the surface. In rugged terrain, on the other hand, more sample points may be used to preserve details of the terrain features. Additionally, a TIN can take into account various topographical feature points (e.g., peaks, pits, and saddle points) and lines (e.g., ridges, valleys, and cliffs), allowing it to accurately approximate complex topographical surfaces from relatively few sampling points. A TIN, however, has its drawbacks. Its storage and management, and data manipulation process are relatively complex, costing significantly higher in computation compared with grids, particularly when a large dataset is used.

**Contour**

A contour-based DEM is composed of a set of contour lines that are “tagged” with their corresponding elevation values. The contour lines are stored by ordered coordinates and can be thought of as polygons or polygon-arcs containing elevation values (Burrough and McDonnell 1998). The contour and flow path (i.e., the normal of the contour line) can be used to divide the topographical surface into irregular polygons (Figure 8), which can simplify the analysis and computation of hydrological models. This type of structure is, therefore, relatively common in hydrological modeling and related analytical geosciences.

The various topographical factors and features in DEM applications can be produced from each of the three DEM structures, although the grid DEM is the simplest, and relatively more efficient. Because of this, at present, DEMs have been commonly defined as grid structures (Theobald 1989) and many countries provide DEM data only in the regular grid matrix format. Advantages such as the high degree of
automation of data collection, the simplicity of computation and database management, and the ease of integration with remote sensing and raster GIS have given grid DEMs a dominant position in applications (Tang 2000). Nevertheless, digital terrain modeling data structures are not static and a large number of algorithms allow for conversion between different structures of DEM. Therefore, in selecting data structures for digital terrain modeling, the data source, technical requirements, and the application aims must all be considered.

Figure 8  A contour-flow path network.
Digital terrain analysis

Digital terrain analysis (DTA) is defined as the information extraction method and technique that derives terrain parameter computation and feature extraction from DEMs. The DTA is also termed “geomorphometry” in literature (Hengl and Reuter 2009). The core task of DTA is to find a balance between the influencing factors in the course of complex real-world geography and simple, clear, efficient, and easy-to-understand abstractions through computer realization. There are many parameters that can be used to describe terrain features and spatial distribution, and the classifications of terrain attributes vary across different disciplines and fields. Within the scope of the earth sciences, terrain attributes can be separated into ordinary terrain attributes (e.g., elevation, slope and aspect) and hydrological attributes (e.g., terrain structure recognition, watershed segmentation). In terms of the relationship between topographical elements, there are single attributes and compound attributes. The single attributes are derived directly from DEM, whereas compound attributes are functions of several single attributes, often used to describe spatial variation, and usually constructed by modeling empirical relationships or simplified natural mechanisms. In terms of computation, terrain attributes can be divided into local terrain attributes and nonlocal attributes.

Based on the relationship and computation characteristics of terrain elements, the terrain attributes can be grouped into four types: surface parameters (parameters), morphological features (features), statistical attributes (statistics), and compound attributes, as summarized in Tables 1–4 (Wilson and Gallant 2000; Zhou and Liu, 2006).

Terrain surface parameters (Table 1) have clear mathematical expressions and physical definitions, such as slope, aspect, and curvature, and can be directly calculated on a DEM. The results of parameter calculations have actual physical meaning and are scalar quantities, and can be directly assessed by the measurements in the field or on a topographical map. The calculation of parameters is usually realized through differential or surface fitting techniques in a local window of a grid DEM.

Terrain morphological features (Table 2) are qualitative expressions of surface morphology and characteristics, and can be directly extracted from a DEM. The distinguishing quality of features is that while their definition is clear, their boundary conditions have ambiguous qualities, making them difficult to represent using mathematical expressions. An example is the division of watershed units, as it is often difficult to define the watershed boundaries. In accordance with the spatial features and relationships in surface morphology, the surface’s morphological features can be extracted and used to determine terrain elements, such as topographical feature points, geomorphological feature units, hydrological elements, topological structure lines, and visibility. Terrain morphology features are commonly extracted in accordance with simplified models of the analysis of spatial distribution relationships between elevation points or the motion mechanisms of surface material, through simulation algorithms and realizations, such as the water-path algorithm used to determine hydrological elements. Results obtained through terrain morphological features are commonly expressed in the form of categorizations; statistical methods can be used to undertake testing of these categories.

Terrain statistical attributes (Table 3) refer to the statistical properties of the terrain surface in a region. Statistical methods are applied to the DEM to derive measurements of terrain nature, such as correlation, regression, trend surfaces, and spatial clustering. The statistical attributes are
<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Symbol</th>
<th>Unit</th>
<th>Description and expression</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$H$</td>
<td>m</td>
<td>The vertical distance from a surface point to the Geoid</td>
<td>Relief, potential, climate, vegetation types</td>
</tr>
<tr>
<td>Relief</td>
<td>Absolute relief</td>
<td>$\Delta H$</td>
<td>m</td>
<td>The difference between the highest and lowest points in a region</td>
<td>Relief, soil erosion, terrain complexity</td>
</tr>
<tr>
<td></td>
<td>Relative relief</td>
<td>$\delta H$</td>
<td>m</td>
<td>The difference between the local ridge line and valley line</td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>Projected distance</td>
<td>$L$</td>
<td>m</td>
<td>The projected distance between two points</td>
<td>Terrain complexity, civil engineering</td>
</tr>
<tr>
<td></td>
<td>Slope distance</td>
<td>$L_p$</td>
<td>m</td>
<td>The distance along the slope between two points</td>
<td></td>
</tr>
<tr>
<td>Terrain surface</td>
<td>Area</td>
<td>$S$</td>
<td>m²</td>
<td>Projected area of a terrain unit</td>
<td>Civil engineering</td>
</tr>
<tr>
<td>statistics parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Surface area</td>
<td>$S_S$</td>
<td>m²</td>
<td>The actual surface area of a terrain unit</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Volume</td>
<td>$V$</td>
<td>m³</td>
<td>The volume between the terrain surface and a given plane</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Surface roughness</td>
<td>$K_R$</td>
<td>—</td>
<td>The complexity of surface in a terrain unit, $K_R = \frac{S_S}{S} = \frac{\sum S_i \sec \beta_i}{\sum S_i}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slope parameters</td>
<td>Slope</td>
<td>$\beta$</td>
<td>°</td>
<td>The angle between the normal vector and zenith at a point, $\beta = \arctan \sqrt{p^2 + q^2}$</td>
<td>Flow velocity, vegetation, rainfall, geomorphology, soil moisture, land suitability</td>
</tr>
<tr>
<td></td>
<td>Aspect</td>
<td>$\alpha$</td>
<td>°</td>
<td>The angle between the projected normal vector on the horizon and the true North, $\alpha = \arctan \left( \frac{2}{p} \right)$</td>
<td>Flow direction, solar radiation, soil moisture, evaporation, vegetation distribution</td>
</tr>
<tr>
<td>Curvature</td>
<td>Mean curvature</td>
<td>$C_M$</td>
<td>m(^{-1})</td>
<td>The arithmetic mean of the maximum and minimum curvatures, $C_M = \frac{r^2 + 2pq + \frac{q^2}{r}}{2(1 + r^2 + q^2)^{3/2}}$</td>
<td>Divergent/convergent area, soil moisture, soil property, landslide distribution, flow acceleration erosion/deposition rate, geomorphological feature, vegetation distribution, mass movement</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------</td>
<td>-------</td>
<td>------------</td>
<td>----------------------------------------------------------------</td>
<td>-----------------------------------------------------------------</td>
</tr>
<tr>
<td>Profile curvature</td>
<td>$C_p$</td>
<td>m(^{-1})</td>
<td>The curvature along the slope, $C_p = \frac{r^2 + 2pq + \frac{q^2}{r}}{(r^2 + q^2)(1 + r^2 + q^2)^{3/2}}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contour curvature</td>
<td>$C_C$</td>
<td>m(^{-1})</td>
<td>The curvature along the contour line, $C_C = \frac{q^2 - 2pq + \frac{p^2}{q}}{(r^2 + q^2)^{3/2}}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tangential curvature</td>
<td>$C_T$</td>
<td>m(^{-1})</td>
<td>The curvature along the direction perpendicular to the slope, $C_T = \frac{q^2 - 2pq + \frac{p^2}{q}}{(r^2 + q^2)^{3/2}}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hydrologic parameters</th>
<th>Upslope slope $\beta_U$</th>
<th>°</th>
<th>The mean slope of the upslope area</th>
<th>Flow velocity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upslope length $L_U$</td>
<td>m</td>
<td>The mean flow length of the upslope area to a given point</td>
<td>Flow acceleration, erosion rate</td>
<td></td>
</tr>
<tr>
<td>Upslope area $S_U$</td>
<td>m(^2)</td>
<td>The area of the upslope region</td>
<td>Flow volume, stable flow velocity</td>
<td></td>
</tr>
<tr>
<td>Dispersal slope $\beta_D$</td>
<td>°</td>
<td>The mean slope of the dispersal area</td>
<td>Soil infiltration rate</td>
<td></td>
</tr>
<tr>
<td>Dispersal length $L_D$</td>
<td>m</td>
<td>The length between a given point to the boundary of the dispersal area</td>
<td>Soil drainage resistance</td>
<td></td>
</tr>
<tr>
<td>Dispersal area $S_D$</td>
<td>m(^2)</td>
<td>The area of the dispersal region</td>
<td>Soil drainage velocity</td>
<td></td>
</tr>
<tr>
<td>Flow path length $L_F$</td>
<td>m</td>
<td>The maximum flow path to a given point in a catchment</td>
<td>Soil erosion, sediment and mass movement</td>
<td></td>
</tr>
<tr>
<td>Catchment slope $\beta_C$</td>
<td>°</td>
<td>The mean slope of the catchment</td>
<td>Time for mass movement and deposition</td>
<td></td>
</tr>
</tbody>
</table>

(continued overleaf)
### Table 1 (Continued)

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Symbol</th>
<th>Unit</th>
<th>Description and expression</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrologic parameters (continued)</td>
<td>Catchment length</td>
<td>$L_C$</td>
<td>m</td>
<td>The distance between the highest point to the catchment outlet</td>
<td>Surface flow damping</td>
</tr>
<tr>
<td></td>
<td>Catchment area</td>
<td>$A$</td>
<td>m$^2$</td>
<td>The projected area of the upslope catchment at a given point or segment of contour line</td>
<td>Runoff volume, runoff depth</td>
</tr>
<tr>
<td></td>
<td>Specific catchment area</td>
<td>$A_s$</td>
<td>m$^2$</td>
<td>The ratio of the catchment area and contour length, $A_s = \lim_{L \to 0} \frac{A}{L}$</td>
<td>Runoff volume, runoff depth, stable flow velocity, soil moisture, geomorphological features</td>
</tr>
<tr>
<td></td>
<td>Flow accumulation</td>
<td>$V_A$</td>
<td>m$^3$</td>
<td>The sum of upslope flow at a given point, $V_A = A \cdot R$ (R = flow depth)</td>
<td>Runoff volume, drainage network, geomorphological feature</td>
</tr>
</tbody>
</table>

Terrain surface polynomial functions: $z = f(x,y)$, $p = \frac{\partial z}{\partial x}$, $q = \frac{\partial z}{\partial y}$, $r = \frac{\partial^2 z}{\partial x^2}$, $t = \frac{\partial^2 z}{\partial y^2}$, $s = \frac{\partial^2 z}{\partial x \partial y}$. 

---

Hydrologic parameters (continued): 

- **Catchment length** ($L_C$): The distance between the highest point to the catchment outlet.
- **Catchment area** ($A$): The projected area of the upslope catchment at a given point or segment of contour line.
- **Specific catchment area** ($A_s$): The ratio of the catchment area and contour length, $A_s = \lim_{L \to 0} \frac{A}{L}$.
- **Flow accumulation** ($V_A$): The sum of upslope flow at a given point, $V_A = A \cdot R$ (R = flow depth).
Table 2  Terrain morphological features.

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Description and expression</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terrain feature points</td>
<td>Peak</td>
<td>A point on a convex surface (higher than all surroundings), ( r &gt; 0, t &gt; 0 )</td>
<td>Terrain structure and features, geomorphological unit classification, soil moisture, land use</td>
</tr>
<tr>
<td></td>
<td>Ridge</td>
<td>A point on a convex line, ( r &gt; 0, t = 0 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pass</td>
<td>The intersection between the perpendicular convex and concave lines, ( r &gt; 0, t &lt; 0 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Plane</td>
<td>Points on a plane, ( r = 0, t = 0 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Value</td>
<td>Points on a concave line, ( r &lt; 0, t = 0 )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pit</td>
<td>A point on a concave surface (lower than all surroundings), ( r &lt; 0, t &lt; 0 )</td>
<td></td>
</tr>
<tr>
<td>Geomorphological feature units</td>
<td>Divergent shoulder</td>
<td>Upper part of divergent slope</td>
<td>Soil landscape classification, soil moisture, geomorphological unit mapping, geomorphological unit mapping, delineation of drainage network, hydrological modeling</td>
</tr>
<tr>
<td></td>
<td>Convergent shoulder</td>
<td>Upper part of convergent slope</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Divergent backslope</td>
<td>Middle part of divergent slope</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Convergent backslope</td>
<td>Middle part of convergent slope</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Divergent footslope</td>
<td>Lower part of divergent slope</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Convergent footslope</td>
<td>Lower part of convergent slope</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sink</td>
<td>The area without outlet</td>
<td></td>
</tr>
<tr>
<td>Hydrological features</td>
<td>Flow direction</td>
<td>Surface flow direction at a given point</td>
<td>Hydrological modeling, soil, geographical unit, catchment geomorphology</td>
</tr>
<tr>
<td></td>
<td>Catchment</td>
<td>Catchment unit</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Drainage network</td>
<td>Surface drainage network</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Drainage node</td>
<td>Intersection of streams</td>
<td></td>
</tr>
</tbody>
</table>

(continued overleaf)
<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Description and expression</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terrain skeleton lines</td>
<td>Ridge line</td>
<td>Water divide, $V_A = 0$</td>
<td>Surveying, hydrology, soil, geomorphology, geography</td>
</tr>
<tr>
<td></td>
<td>Valley line</td>
<td>Water accumulation line, $V_A &gt;$ threshold</td>
<td></td>
</tr>
<tr>
<td>Terrain profile</td>
<td>Longitudinal section</td>
<td>Relief along a given direction</td>
<td>Regional relief, geological and hydrological features, soil, vegetation, civil engineering, land use</td>
</tr>
<tr>
<td></td>
<td>Cross-section</td>
<td>Relief along the direction perpendicular to the longitudinal section</td>
<td></td>
</tr>
<tr>
<td>Visibility</td>
<td>Intervisibility</td>
<td>Visibility between two given points on the surface</td>
<td>Site selection, telecommunication, navigation, surveying, forestry</td>
</tr>
<tr>
<td></td>
<td>Viewshed</td>
<td>Visible/invisible area from a given surface point</td>
<td></td>
</tr>
</tbody>
</table>
Table 3  Terrain statistical attributes.

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Symbol</th>
<th>Unit</th>
<th>Description and expression</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation and relief</td>
<td>Mean elevation</td>
<td>$\overline{H}$</td>
<td>m</td>
<td>$\overline{H} = \frac{1}{N} \sum_{i=1}^{N} H_i$</td>
<td>Average elevation of a region</td>
</tr>
<tr>
<td></td>
<td>Difference from mean elevation</td>
<td>$\Delta H$</td>
<td>m</td>
<td>$\Delta H = H_0 - \overline{H}$</td>
<td>Local relief, groundwater, forestry</td>
</tr>
<tr>
<td></td>
<td>Standard deviation of elevation</td>
<td>$\sigma$</td>
<td>m</td>
<td>$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (H_i - \overline{H})^2}$</td>
<td>Local relief diversity, roughness of geomorphological unit</td>
</tr>
<tr>
<td></td>
<td>Elevation range</td>
<td></td>
<td>m</td>
<td>The difference between the maximum and the minimum elevation in a local window</td>
<td>A general description on relief</td>
</tr>
<tr>
<td></td>
<td>Deviation from mean elevation</td>
<td>$D\overline{H}$</td>
<td>m</td>
<td>$D\overline{H} = \frac{\Delta H}{\sigma} = \frac{H_0 - \overline{H}}{\sigma}$</td>
<td>Identifying abnormal elevation points</td>
</tr>
<tr>
<td></td>
<td>Percentile</td>
<td>pctl</td>
<td>%</td>
<td>$pctl = \frac{100}{N} \sum_{i=1}^{N} \mathbb{1}(H_i &lt; H_0)$</td>
<td>Relative range in the regional relief</td>
</tr>
<tr>
<td></td>
<td>Percentage of elevation range</td>
<td>pctg</td>
<td>%</td>
<td>$pctg = \frac{100}{H_{\text{max}} - H_{\text{min}}} \sum_{i=1}^{N} \mathbb{1}(H_i &lt; H_0)$</td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>Mean aspect</td>
<td>$\overline{\alpha}$</td>
<td>°</td>
<td>$\overline{\alpha} = \tan^{-1}\left(\sum_{i=1}^{N} \sin \alpha_i \sum_{i=1}^{N} \cos \alpha_i\right)$</td>
<td>Average aspect of a region, reflecting the general geomorphological feature</td>
</tr>
<tr>
<td></td>
<td>Composed length</td>
<td>$\overline{R}$</td>
<td>—</td>
<td>$\overline{R} = \frac{1}{n} \sqrt{\left(\sum_{i=1}^{N} \cos \alpha_i\right)^2 + \left(\sum_{i=1}^{N} \sin \alpha_i\right)^2}$</td>
<td>Distribution of aspect, geomorphology</td>
</tr>
<tr>
<td>Elevation distribution</td>
<td>Normal distribution</td>
<td></td>
<td>—</td>
<td>$p = \frac{\Delta H \rho^\alpha}{\sqrt{2\pi} \sigma^\beta} e^{-\frac{(H_i - \overline{H})^2}{2\sigma^2}}$</td>
<td>Distribution of elevation, geomorphological modeling, cartography, engineering design</td>
</tr>
<tr>
<td></td>
<td>Pearson III distribution</td>
<td></td>
<td>—</td>
<td>$p = \frac{\Delta H \rho^\alpha}{\Gamma(a) \delta^\beta (H_i - \overline{H})^{\alpha-1} \beta(H_i - \delta)^{\alpha-1} e^{-\beta(H_i - \delta)}}$</td>
<td></td>
</tr>
</tbody>
</table>

(continued overleaf)
<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Symbol</th>
<th>Unit</th>
<th>Description and expression</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gully density</td>
<td></td>
<td>$D$</td>
<td>m$^{-1}$</td>
<td>$D = \sum \frac{L}{S}$</td>
<td>The sum of gully length, reflecting erosion intensity, soil erosion, civil engineering</td>
</tr>
<tr>
<td>Trend surface</td>
<td></td>
<td>$\hat{Z}$</td>
<td></td>
<td>$\hat{Z} = \sum_{i,j=0}^{N} a_{ij} x^i y^j$</td>
<td>The general trend of terrain</td>
</tr>
<tr>
<td>Spatial autocorrelation indices</td>
<td>Morán’s I</td>
<td>$I$</td>
<td></td>
<td>$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} (x_j - x_m) (x_j - x_m)}{\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} (x_j - x_m)^2}$</td>
<td>Spatial autocorrelation of elevation data</td>
</tr>
<tr>
<td></td>
<td>Geary’s $\gamma$</td>
<td>$\gamma$</td>
<td></td>
<td>$\gamma = \frac{(n-1) \sum_{i=1}^{n} w_{ij} (x_j - x_m)^2}{\sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} (x_j - x_m)^2}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Semivariance</td>
<td>$\gamma(h)$</td>
<td></td>
<td>$\gamma(h) = \frac{1}{2n} \sum_{i=1}^{n} [z(x_i) - z(x_i + h)]^2$</td>
<td></td>
</tr>
</tbody>
</table>
also used to find the variation patterns or implied links between factors, and to select appropriate factors or parameters to construct geographical models, so that the spatial patterns governing the evolution of terrain can be explored at a deeper level.

Compound attributes (Table 4) are typically computed based on surface parameters and terrain morphological attributes, using models of relevant applied science (such as hydrology, geomorphology, or soil sciences). Unlike the surface parameters and terrain morphology characteristics previously mentioned, compound terrain attributes are not directly computed or extracted from a DEM, hence they are named “secondary topographic attributes” (Wilson and Gallant 2000). Compound terrain attributes have practical significance to a number of fields of application, but it is difficult to test the results of their practical applications. Their physical implications also require professional measurements and interpretation within the field of application.

Tables 1–4 show the direct and indirect links between compound attributes and surface parameters such as slope, aspect, catchment area, and specific catchment area (SCA). These parameters have broad applications in such analytical geosciences as hydrology, geomorphology, or soil sciences (Zhou, Lees, and Tang 2008). For example, slope can be used to calculate flow direction, which can be used to calculate SCA. The SCA can further be used to discern geomorphological structure lines (ridges and valleys), and used with slope as the input parameters of topographical wetness indexes, flow strength, and so on.

Besides numerical analysis, another important area of DTA is the DEM visualization. DEM visual representation is perhaps the most impressive task of DTM, and it is clearly a focus of development for commercial GIS software. In addition to rapid, realistic, accurate, and clear visual presentation of original DEM or derived data, the visual analysis in DTA is also focused on enhancing the visual impact of terrain features of interest, and conveying information on surface parameters, morphology, and compound attributes.

**Summary**

Digital terrain modeling is an important component of GIS. Almost all current GIS software has some kind of DTM function in its functional capability specifications. In recent years, the rapid development of elevation data acquisition techniques has been witnessed. Numerous public-accessible global and national DEM databases have become available at various spatial scales. These advancements will certainly encourage more advanced applications in a broader application domain.

DTM, in a broader sense, is not only the data representing terrain surface but also the provider of geographical analytical functions that derive and extract geographical/geomorphological parameters and features from digital elevation data. In many GIS application projects, particularly those dealing with environment, natural resources, and urban lives, use of DTM data and analytical functions has been a common practice.

In the foreseeable future, it is expected that DEMs will progress in the following aspects:

- more rapid and cheaper acquisition of high-resolution digital elevation data;
- more advanced generalization/extraction methods and algorithms in constructing multiscale DEMs from very large elevation datasets;
- broader adoption and integration of DTM data and techniques in applications of specific application fields;
### Table 4  Compound attributes.

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Description and expression</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topographic wetness index</td>
<td>$\omega$</td>
<td>$\omega = \ln \left( \frac{A_s}{\tan \beta} \right)$</td>
<td>Flow accumulation, soil moisture, groundwater level, soil evaporation, soil organic matter, pH, sediment, vegetation ecological distribution</td>
</tr>
<tr>
<td>Stream power index</td>
<td>$\Omega$</td>
<td>$\Omega = A_s \tan \beta$</td>
<td>Potential erosivity, soil organic matter, pH, sediment, vegetation ecological distribution</td>
</tr>
<tr>
<td>Sediment transport capacity index</td>
<td>$T_C$</td>
<td>$T_C = \left( \frac{A_s}{22.13} \right) \left( \frac{\sin \beta}{0.0896} \right)$</td>
<td>Potential erosivity, soil organic matter, pH, sediment, vegetation ecological distribution</td>
</tr>
<tr>
<td>Solar radiation models</td>
<td>$R_{\text{dir}}$</td>
<td>$R_{\text{dir}} = \frac{R_{\text{dirh}}(1-a)}{\sqrt{1+p^2+q^2}} (\sin Z - p \sin \alpha \cos Z - q \cos \alpha \cos Z)$</td>
<td>Surface solar radiation intensity ($p$ and $q$ are defined in Table 1)</td>
</tr>
<tr>
<td></td>
<td>$R_{\text{diff}}$</td>
<td>$R_{\text{diff}} = \nu R_{\text{dirh}}$</td>
<td>Diffused radiation on nonflat land surface</td>
</tr>
<tr>
<td></td>
<td>$R_{\text{ref}}$</td>
<td>$R_{\text{ref}} = (R_{\text{dirh}} + R_{\text{diff}})(1-v)a$</td>
<td>Reflected radiation on nonflat land surface</td>
</tr>
<tr>
<td></td>
<td>$R_{\text{ts}}$</td>
<td>$R_{\text{ts}} = (R_{\text{dir}} + R_{\text{diff}}) \left[ \frac{n}{N} + \left( 1 - \frac{n}{N} \right) \xi \right] + R_{\text{ref}}$</td>
<td>The total received shortwave solar radiation in a period of time (including topographic and cloud corrections)</td>
</tr>
<tr>
<td></td>
<td>$L_{\text{in}}$</td>
<td>$L_{\text{in}} = \varepsilon_s \sigma T_s^4 + (1-v)L_{\text{out}}$</td>
<td>Incident and atmospheric longwave radiation</td>
</tr>
<tr>
<td></td>
<td>$L_{\text{out}}$</td>
<td>$L_{\text{out}} = \varepsilon_s \sigma T_s^4$</td>
<td>Outgoing longwave radiation</td>
</tr>
<tr>
<td></td>
<td>$R_n$</td>
<td>$R_n = (1-a)R_{\text{ts}} + \varepsilon_s L_{\text{in}} - L_{\text{out}}$</td>
<td>Net radiation and surface energy balance in a period of time</td>
</tr>
<tr>
<td>Temperature model</td>
<td>$T$</td>
<td>$T = T_0 - \frac{T_{\text{topc}}(H-H_0)}{1000} + G \left( \xi_R - \frac{1}{\xi_R} \right) \left( 1 - \frac{LAI}{LAI_{\text{max}}} \right)$</td>
<td>Estimating landscape and surface temperature using nearby station data</td>
</tr>
</tbody>
</table>
• faster and interactive DEM visualization hardware and software to support real-time decision-making for a large region.

SEE ALSO: Data structure, raster; Data structure, vector; Geocomputation; Geodesign; Glacier modeling; Landscape; LiDAR; Spatial modeling; Voronoi diagrams

References


Digital soil mapping and pedometrics

Bradley A. Miller
Iowa State University, USA
Leibniz Centre for Agricultural Landscape Research (ZALF), Germany

Digital soil mapping (DSM) is the computer-assisted production of maps that present either a single soil property or specified combinations of soil properties, based on a defined classification system. Digital soil maps differ from soil maps traditionally produced on paper by being a visualization of a geographically referenced database. Although a digital soil map can be directly generated from a paper map, digital soil maps are increasingly being created from digital data by processes that are explicitly defined. The format of the spatial data created by DSM is particularly useful as model parameters for precision agriculture, land management, and other environmental issues.

Due to the extent and complexity of the soil landscape, creating soil maps requires some form of spatial prediction and generalization. Methods of prediction commonly utilize the concepts of spatial association and spatial autocorrelation. DSM applies the data and tools of geographic information systems (GIS) to improve the accuracy and efficiency of soil landscape models. GIS can quantitatively utilize data from digitized legacy maps, field samples georeferenced by global positioning systems (GPS), and data products from remote and proximal sensing to model the soil landscape.

Predictor variables (base maps)

DSM makes use of parameters from a large variety of sources, each of which has strengths and weaknesses. With the current state of technology, however, they all have roles in the calibration and validation of digital soil landscape models.

GIS can utilize legacy maps through digitization, but spatial accuracy is a major concern when utilizing maps that were not initially created for the GIS environment. One of the reasons for this is that the meaning of cartographic scale has changed between maps produced on paper and those produced digitally. With paper maps, the display size is fixed, meaning that the map audience is most likely viewing the map at the same scale that the mapper did. This limitation creates a natural linkage between map detail, accuracy, and cartographic scale, which is no longer the case in GIS environments. In a GIS environment, the user is able to freely change the cartographic scale by zooming in or out, allowing the map to be viewed at a greater (or lesser) level of detail than the map maker intended or than is supported by the original data.

Another source of information for DSM is direct points of observation. Soil samples from the field can have their location recorded accurately by GPS. The attributes of that location can be based on field observations or lab analysis. These point observations are commonly coupled with the strategies of spatial association or spatial autocorrelation to fill in the areas between observations for a more complete coverage of the map area.

Proximally and remotely sensed datasets are collected by noninvasive methods that provide
DIGITAL SOIL MAPPING AND PDEOMETRICS

greater areal coverage than data collected by point sampling. Proximal sensing commonly uses geophysical techniques such as electromagnetic induction. Remote sensing uses aerial sensors to detect the spectral distribution of electromagnetic radiation being emitted or reflected by the soil surface. Both of these technologies detect characteristics that can be directly correlated to soil properties, and thus, enable DSM. Because proximal sensing tends to detect belowground features and remote sensing tends to detect surface features, they sometimes offer different perspectives on the soil environment.

In addition to model variables that are directly measured, useful information can be generated from the analysis of variables to create new parameters. Land-surface derivatives, for example, are parameters that are beneficial to DSM. Elevation alone can be used as a predictor of soil properties. However, calculated derivatives of elevation, such as slope gradient, slope curvature, and location on the slope (Ruhe and Walker 1968; Walker and Ruhe 1968), can all provide even more and potentially better information about contextual characteristics of the soil environment.

Apparent soil patterns differ due to analysis scale (i.e., modifiable area unit problem), unique combinations of processes, and the soil attributes of interest. For example, a continental scale map of soil organic carbon shows that carbon relates broadly to climate and the associated ecological communities (Figure 1a). At more local scales, climate is modified by topography, which produces patterns of soil organic carbon that are better distinguished by terrain parameters (Figure 1b). However, the influence of topography can be diminished in some landscapes where parent material or climate characteristics negate topographic processes, for example when the erosion of soil material is suppressed due to high infiltration rates, or in arid climates where all soils are low in organic matter. Additionally, contextual parameters can be calculated at virtually any scale. Depending on the phenomenon being modeled, certain analysis scales will be better suited for the spatial prediction of soil properties. The vast quantity of parameter options and unique conditions at different sites make DSM a challenging, stimulating, and potentially enlightening endeavor.

DSM models

Traditional soil maps are made by synthesizing information gathered in the field with a variety of pre-existing maps, such as topographic maps, geology maps, and aerial photographs. These pre-existing maps are called base maps because they provide spatial reference for the mapper to transfer their mental model of the soil landscape into a graphical representation of the soils they see. However, in addition to spatial reference, the base maps can provide information about the environmental context or other attributes that are spatially associated with soil properties. Due to human limitations, only a few base maps can be layered at a time. Also, when the maps are overlain, there is a potential for misalignment. DSM improves upon this process by enhancing spatial accuracy, quantifying the synthesis of base maps, and increasing the number of base maps that can be synthesized at once, all in a more time-efficient manner. The quantitative synthesis of a series of base maps is particularly important as it allows for a fully documented inclusion of many data layers. The analysis of different spatial datasets (aka maps) together is accomplished digitally within a geographic information system. The challenge is to know how and which layers to synthesize, in order to make the best map for the desired purpose.
Figure 1  (a) Global map of soil organic carbon generated from the Harmonized World Soil Database version 1.2 (FAO/IIASA/ISRIC/ISSCAS/JRC 2012). (b) Local scale map (approx. 40 km²) of soil organic carbon generated by digital analysis of remote sensing data. The local scale map is shown in a three-dimensional perspective to illustrate the relationship with terrain. However, the modifying effect of land use is also affecting the spatial pattern at this resolution. Source: (a) FAO/IIASA/ISRIC/ISSCAS/JRC. 2012. Harmonized World Soil Database (version 1.2).

Recognizing the value of the vast body of knowledge that has been accumulated by field soil scientists using traditional soil mapping methods, expert knowledge models attempt to capture and convert the logic in the expert's mental model into digital techniques. Expert knowledge capture includes the selection of model parameters and analysis structure.
DIGITAL SOIL MAPPING AND PEDOMETRICS

Strategies for capturing expert knowledge range from directly questioning the expert, to analyzing the map produced by the expert. Asking the expert to explain their mental model is problematic, however, because tacit knowledge is not always completely quantitative or explicit. Conversely, analysis of the expert’s map is complicated by the multitude of considerations that were used to construct it. For example, different map units can be based on different soil attributes as determined by perceived value, or unique soil areas could be omitted due to issues of scale.

Whether based on tacit knowledge, driven by sample data, or some combination of the two, some type of synthesis of the available data is always needed to accurately model the soil landscape. Popular modeling approaches include regression, kriging, cluster analysis, decision trees, and machine learning (Scull et al. 2003). Because DSM is a rapidly evolving field, new versions of these approaches are continually being developed. Although there is constant innovation in DSM, modeling approaches generally apply the concepts of spatial association, spatial autocorrelation, or some combination of the two.

Fundamental strategies of DSM

Spatial association is the similar distribution of attributes within geographic space. In cases where there are known spatial associations, the spatial distribution of a measured attribute can be used to predict the distribution of an unmeasured attribute. This is particularly useful when certain attributes are more easily measured than the desired attribute. For example, soils in prairie or steppe ecosystems tend to have a thicker horizon of topsoil enriched with organic carbon. Also, the soil particle size distribution can be used to predict soil hydraulic properties. In soil science these relationships are commonly referred to as pedotransfer functions (Bouma 1989). Computer models that utilize this concept are generally referred to as soil inference systems. A challenge for DSM strategies utilizing spatial association is the selection of the optimal parameters and models for the soil properties of interest. Not only is there an ever-increasing variety of measurements available, many of those measurements are scale dependent, which multiplies the number of parameters to choose from.

A classic example of spatial association in soil mapping is Dokuchaev’s (1967/1883) factors of soil formation: climate, organisms, relief, parent material, and time. Because each of these parameters describes the environment affecting soil development, this approach to soil mapping can be described as environmental correlation. The majority of parameters used in DSM models can be categorized under these factors. However, related soil properties that are more easily measured than the property of interest are also commonly used. Some examples of these kinds of pedotransfer functions are the relationship between clay content and electrical conductivity, or predicting soil erodibility from the combination of particle size and organic matter content data.

Spatial autocorrelation is the covariation of properties within geographic space. For DSM, spatial autocorrelation relies on the increasing similarity of soil properties with increasing proximity. Most spatial interpolation methods utilize this concept to predict attributes between measured observations. Various versions of kriging, a type of spatial interpolation based on a Gaussian process, have become popular in DSM because of their internal quantification of spatial variability. Because many soil properties and conditions are difficult to observe at all locations, especially those below the surface, some form of interpolation is often needed to estimate attributes between points of observation. The challenge to this problem lies in the requirement of adequate sample density or support for interpolating
across the desired area. Different soil properties have different ranges of spatial autocorrelation, requiring different sampling densities for properly constructing an interpolation model.

**Representation**

A digital soil map needs to be stored and visualized in one of the data models available in GIS. The most common of these are the vector and raster data models. Vector data models are the most similar to traditional soil mapping and are thus most often used for digitizing legacy (paper) maps. Raster data models can store data more efficiently, but are restricted to a grid of rectangular delineations, or cells (Figure 2). Raster data models are becoming increasingly popular in DSM for two reasons. First, better data storage efficiency allows for faster data processing and soil maps of larger extents, which are particularly needed for environmental models. Second, increasing the resolution of the raster (decreasing the cell size) brings the map representation closer to the theoretical concept of soil as a continuous surface.

An additional strategy to represent the soil landscape as a more continuous surface is fuzzy logic (Burrough 1989). With fuzzy logic, a location can have partial memberships in multiple attribute classes, that is, a map may contain delineations that have partial membership in soil A and B, whereas on a traditional map the delineation would have been forced to be either soil A or B. Typically the degree of similarity of a location to a class is expressed on a numeric scale from 0 (nonmembership) to 1 (complete membership). Classes can then be created from the data (e.g., fuzzy k-means) or based on predefined classification systems. Fuzzy memberships can be used throughout the DSM process and can be simplified at any time. The process of converting fuzzy memberships into a Boolean classification is referred to as hardening.

**Uncertainty (error)**

Uncertainty is an important aspect of digital mapping and modeling that will need to be added to future soil maps. In this context, uncertainty is a quantified expression about the limit in precision for the map data. Information about uncertainty must be sufficient for the user to determine the fitness of the soil map for the intended purpose. Uncertainty in a soil map can propagate from measurement inaccuracies as well as from spatial and temporal variation (Heuvelink 1999). Errors in spatial position and soil property analysis both contribute to measurement error. To simplify the problem, spatial position is commonly assumed to be correct, allowing the error to be more concisely expressed as the difference between the modeled value and the measured value at particular locations. For soil attributes, this error can be communicated stochastically or summarized for the entire map area with statistics like the coefficient of determination ($R^2$) or root mean square error (RMSE). Because soil classification is a categorical attribute, tests of digital mapping accuracy in this case tend to express error as percent agreement with validation points, a confusion matrix, or a kappa statistic. Geostatistics (e.g., kriging) has recently become popular because of the need to quantify uncertainty and its ability to model error across the map area.

Spatial heterogeneity is also an important consideration for determining the fitness of a spatial dataset for a given use. Generally, interpolation methods estimate an attribute value for a central point within a map delineation and assign that value to the entire delineation area. For this reason, spatial interpolation approaches
Figure 2  Example of (a) a traditional soil map (Dideriksen 1986) side by side with (b) a digitized version of the same map in a polygon (vector) data model (Soil Survey Staff 2013). In (c), these data have been converted to a raster data model. In the digital format, the spatial soil data can be easily incorporated into models for a variety of purposes. The goal of digital soil mapping is to leverage new digital data and techniques to improve the soil map. Source: (a) Reproduced from Dideriksen 1986. © US Government; (b) and (c) Produced from data provided by the Soil Survey Staff 2013. © US Government.
seek to increase the resolution to as high as can be supported by the data and computational constraints. However, it is difficult to have a sufficiently high density of data points covering large spatial extents, and there are few technical restraints for interpolating a resolution finer than that supported by the data. In contrast, traditional soil mapping uses spatial association to divide the landscape into as many pieces as possible. The size of map delineations is limited by the resolution of the base maps, thus the size of the delineations reflects spatial uncertainty. Attribute uncertainty is then expressed by a range of values associated with the unique combination of predictors observed within the delineation area.

Pedometrics

Pedometrics is the application of mathematical and statistical methods for the study of the distribution and genesis of soils, often with a focus on quantifying uncertainty (De Gruijter, Webster, and Myers 1994). The establishment and growth of pedometrics in the early 1990s followed the post-quantitative revolution study of pedology. As digital techniques of soil mapping are by definition based on quantified datasets and procedures, DSM is nearly synonymous with pedometrics. However, pedometric analyses tend to focus on the use of geostatistics. Those who study pedometrics advocate for the use of geostatistics because it is based on random function theory, which enables the modeling of the uncertainty associated with spatial estimation. Like pedology, in addition to soil mapping, pedometrics includes the study of soil formation.

Although the spatial distribution of soils and their formation factors are intimately linked, they do not necessarily have to be studied together. For example, geostatistics can interpolate between sample points without regard to pedogenesis. Conversely, soil research can focus on vertical processes in the soil profile without regard to the horizontal dimension. However, each research concentration benefits from the consideration of the other. This is illustrated by the use of Dokuchaev’s soil formation factors, which began as a tool for spatial prediction, later applied as a method to reduce independent variables in pedologic research (e.g., Jenny 1941).

Information about both individual pedons and their place in the soil landscape is essential for understanding the soil system. The innovation of geographic information systems and other geospatial technologies has reinvigorated interest in soil geography, leading to a recent explosion in ideas and research for improving soil maps.

SEE ALSO: Data structure, raster; Data structure, vector; Scale; Soil mapping and maps; Spatial analysis; Uncertainty

References


FAO/IIASA/ISRIC/ISSCAS/JRC (Food and Agriculture Organization of the United Nations/ Institute for Applied Systems Analysis/Inter-


Further reading


Disability

Edward Hall
University of Dundee, UK

Disability is generally understood as a physical or mental impairment that constrains an individual’s participation in the activities and sites of “normal” society; this biomedical definition is now challenged by a social constructionist conceptualization that locates the “problem” of disability in the exclusionary landscapes and attitudes of mainstream society. Human geography’s long-standing engagement with disability reflects this shifting debate: from studies of the spatial distribution of medically defined conditions, such as multiple sclerosis, to identifying the physical and attitudinal barriers to the accessing of public space, and from mapping the everyday social geographies of people with intellectual disabilities to, more recently, examining the role of online environments for those on the autistic spectrum. In short, disability geography studies the interactions between people with impaired bodies and minds and social and spatial contexts; a spatialized perspective has made a unique and incisive contribution to the understanding of disability and the lives of disabled people.

Disabled people have experienced much positive change since the mid-1990s, with an increasing presence in public spaces, workplaces, education, sport, and the media. Antidiscrimination legislation in many countries and United Nations support for the human rights of disabled people, along with a growing disability political movement, have contributed to this gradual transformation. The emergence of an aging and increasingly chronically ill population in the Global North has emphasized the growing commonality of the experience of disability; however, many disabled people continue to experience discriminatory attitudes and inaccessible environments, and hence limited opportunities, in many spheres of life; there is also a growing awareness that in the Global South, where the majority of disabled people are located, everyday lives remain extremely challenging.

The study of disability in the discipline has its roots in medical geography; indeed, geographical analyses of disability have, perhaps unsurprisingly, reflected changes in social science and popular conceptualizations at large. Medical geography understood disability to be within its remit as a set of clinically defined conditions, and applied the techniques of epidemiology to map and study its incidence and to examine causation; examples are studies of multiple sclerosis and schizophrenia and their possible environmental causes. While such studies are now questioned for their uncritical adoption of a medical model of disability, they did succeed in identifying the unequal spatial distribution of disabled people, and emphasized the importance of location for the production of disability. Another quite different research strand that emerged from a broader social geographical concern with welfare and the “crisis” of urban deprivation was the deinstitutionalization of disabled people (including people with mental health conditions and intellectual disabilities) from asylums and other large care sites in the 1970s and 1980s. This research focused attention on the role of spatial context in shaping the lives of disabled people, including barriers to access in urban
environments and transport systems for people with mobility problems (Hahn 1986); and the discrimination targeted at people with mental health conditions, together with the absence of care and support services, in local communities (Dear and Wolch 1987).

The phrase “asylum without walls” was coined to reflect the experiences of many disabled people, now spatially included yet remaining socially excluded, in community settings. Important historical geographical work was also undertaken in this period on “lunatic” asylums, which described how their location and internal spatial organization reflected both a medical demand for classification and treatment and a social fear of those with impairments. The interaction between disabled people and urban spaces was also studied, most notably through descriptions of inaccessible streetscapes and the possibility of using tactile maps (and, later, GPS systems) for people with visual impairments in negotiating city environments. Humanistic research, prominent in geography in the 1970s and 1980s, also touched on disability, with studies of the sensory “life worlds” of people with impairments.

The geographical study of disability has reflected broader academic and sociopolitical developments. The discrimination, exclusion, and lack of services experienced by the many disabled people now living in mainstream communities inspired political activism in the 1970s, drawing on other rights-based movements. In the United Kingdom, for instance, the Union of the Physically Impaired Against Segregation (UPIAS), one of the first disability rights organizations, defined the principles that would later inform the social model of disability: “What we are interested in are ways of changing our conditions of life, and thus overcoming the disabilities which are imposed on top of our physical impairments by the way this society is organised to exclude us” (1976). UPIAS and other emerging groups, allied with a small but influential group of disabled academics in the social sciences, including Mike Oliver, Vic Finkelstein, and Jenny Morris, identified and sought to dismantle the dominant impairment-determined biomedical model of disability and to establish in its place a social constructionist model, disentangling the experience of impairment from the “disabling” nature of “normal” social and spatial environments. The broad acceptance of the social model in policy and practice is testament to the tenacity and political empowerment of some disabled people.

The emergence of the social model encouraged social (and later cultural) geographers to consider disability as a topic of study. An exchange in Transactions of the Institute of British Geographers (1993–1996) marked the emergence of the “geography of disability” as a subdiscipline. Ruth Butler (1994), Rob Imrie (1996), and Brendan Gleeson (1996) critiqued the impairment-focused work of Reginald Golledge (1993), advocating instead a concern with the exclusionary nature of everyday environments and the role disabled people should play in the design and practice of academic research.

Three themes emerged. First, building on earlier work on barriers to access in urban spaces, critical studies identified the discriminatory attitudes and structures in public spaces, workplaces, institutions, and other sites, and further how such environments were inherently “disablist” in their design and operation. Second, a series of studies considered the experience of bodily impairments and the everyday social geographies of disabled people. Significantly, this work saw a critique of the social model of disability, arguing that its focus on social structures and environments had distanced it from the realities of pain and discomfort, and
the challenges of visual impairment and mental illness, experienced by many disabled people (Hall 2000). Through these studies, geography made a significant contribution to the development of an embodied model of disability. Third, there was the development of methodological approaches to include the voices and interests of disabled people to ensure that research is both ethical and empowering. Significant challenges remain, however, including developing appropriate means of gaining permission from those unable to give conventionally understood informed consent.

While the foregoing research strands continue, the study of disability in geography has taken a series of new directions, both broadening its scope and adding further complexity to its conceptualization. An edited collection *Towards Enabling Geographies* (Chouinard, Hall, and Wilton 2010) argued for the extension of the concept of disability – and the insights of disability theory – to include other “non-normal” bodies “disabled” by dominant sociospatial practices, such as older people, “fat” people, and young people with emotional and behavioral problems; in addition, the inclusion of research on people with chronic illness hinted at a fruitful reconnection between disability geography and medical geography. However, to some activists and scholars, the inclusion of other bodies of difference in disability geography is a potentially controversial move because of the risk of undermining disability politics and the social model. A counterargument is that a truly emancipatory politics of disability will be inclusive of all forms and experiences of disablism, whatever the nature of impairment. There is also much to be gained analytically: studies of the everyday geographies of people with intellectual disabilities (a group long excluded from the social model of disability and geographies of disability), for example, reveal the intensely relational nature of the construction, maintenance, and contestation of disabling spaces and attitudes, in particular the central role of friends (with and without a disability), family and supporters, and particular sites. Recent studies of the use of online spaces by deaf people and people on the autistic spectrum (many of whom choose not to adopt a disabled identity) have shown the empowering potential of relationships with others with the same impairment, and the enabling value of being able to interact in spaces separate from the mainstream (online and otherwise). In addition, in studies of participation in the creative arts, cultural geographers have revealed the importance of disabled people’s embodiment and emotions in the ongoing relational production of “disability.”

Together, these studies constitute the emergence of a contextual model of disability, placing material embodied experiences of impaired bodies of all forms within the varied and dynamic landscape of social, cultural, and political constructions of disability. This sharpens the critique and insight offered by disability theory into the experiences of disabled people (broadly defined), the disabling (and enabling) nature of local spaces, the discrimination of institutional and state policies, and the meaning of “disability” itself.

The emotional lives of disabled people, in particular people with mental health problems (broadened to include those with phobias and anxiety), have been the subject of recent studies. Building on earlier geographical work on mental health, these studies argue that becoming attuned to emotions permits a fuller understanding of how people with mental ill-health seek to locate themselves within often challenging everyday environments and, further, how identities characterized by difference can be developed. This work has made three key
contributions: first, it has challenged the binary construct of mental health/mental illness, revealing the widespread experience of mental health problems in contemporary society; second, it has shown how mental wellbeing is, at least in part, a product of an individual’s emplacement within local contexts and sets of relations (and, in turn, can shape these contexts); and, third, it has revealed the limitations of the dominant notion of social inclusion for those who cannot fulfill the ideals of active citizenship and paid employment.

The geography of disability has also contributed to the recent expansion of the debate around care and caring. As with the focus on impairment, this debate has generated some negative reaction from disability campaigners who view “care” as part of the disabling biomedical institutional structures that have taken so long to (partially) dismantle. However, it can be strongly argued that care and support are integral to many disabled people’s lives, and, as we include more people with impairments into the disability realm, including older people, we can see that the spaces and relations of care must be central to geographical study (Power 2008). For many disabled people, in particular people with intellectual disabilities and older people, the support of family members, friends, voluntary helpers, and professional carers is often crucial to their wellbeing and their ability to live an independent life. In addition, the continuing deinstitutionalization and personalization of care in many countries in the Global North, for example the closure of adult daycare centers in the United Kingdom, has raised questions of where caring should take place (in an institution, in a community facility, or in an individual’s home) and who should be responsible (family and friends, voluntary organizations, or the government). For many disabled people this period of transformation has caused anxiety; for others, new possibilities for more informal, empowering, and ethical caring have opened up in local community contexts. There is the opportunity here for dominant discourses of social inclusion to be contested by embodied and emotional senses of belonging founded on connections with others in local places, shared activities (e.g., creative arts) and friendships. Further, the assumption that inclusion in the mainstream spaces of education and employment is best for the wellbeing of disabled people has been challenged by geographers, who reveal the benefits for some of separation, at least for part of a person’s everyday life.

There are a number of directions for future research in the geography of disability. First, spatially informed scrutiny of disabling barriers must continue; despite legislative protection and marked cultural change in attitudes in many countries, disabled people still endure significant discrimination in many realms of society. This work can be strengthened by the recognition of the intersecting characteristics of disability identities (i.e., ethnicity, gender, sexuality, and age). Second, a broadened concept of disability to include older people and people with chronic illness can encourage a productive (re-)engagement between the geography of disability and medical and healthcare geography, kept apart for many years by the constructionist underpinning of the social model. Critiquing the social care of older people from a geography of disability perspective, and a medical geography-informed analysis of the wellbeing of people with intellectual disabilities, could prove equally enlightening. An engagement with medical geography would also provide an opportunity to critique the ongoing expansion of genetic medicine that, through improved testing, offers the possibility for the elimination of impairment, a development seemingly widely welcomed by many (yet contested by some
disabled people). Third, as many governments in the Global North make radical changes to welfare provision, the category of “disability” and the status of disabled people have come under intense scrutiny; at the same time, the disabled population continues to grow, with increasing numbers of older people and people with chronic illness. Geographers can play a central role in critiquing the consequent shifts in sites and practices of care and support, and the place of disabled people in the spaces of everyday life. Fourth, there remains a significant absence of work on disability in Global South contexts, despite the majority of disabled people residing in developing countries. The conceptual debate in the Global North, in particular the widespread adoption of the social model, has arguably marginalized the experiences of many disabled people in developing countries, where medically framed attitudes and policies remain dominant. Western geographers should establish enduring connections with academics and organizations in countries in the Global South to ensure that the experiences of many more disabled people in a wider range of contexts are represented. Fifth, the geography of disability has, arguably, been an early adopter of empowering and participatory research methods; indeed, the ethics of research were central to the social model of the 1990s. Innovative techniques, such as “photovoice,” “go along” interviews, and participatory mapping, have emerged in an attempt to better convey people’s experiences. There remains much to be done, however, to rework the relations of research into that of true coproduction, and further to ensure that the outcomes of research are effectively communicated and acted on, in collaboration with people with disabilities to ensure positive outcomes for people’s lives.

**SEE ALSO:** Aging; Caregiving; Health and wellbeing; Mental health geographies

**References**


Further reading


The term “discourse” is very much dependent upon “the disciplinary context in which the term occurs” (Mills 1997, 3). Mills explains how the “term ‘discourse’ has become common currency in a variety of disciplines [including]: critical theory, sociology, linguistics, philosophy, social psychology and many other fields, so much so that it is frequently left undefined, as if its usage were simply common knowledge” (Mills 1997, 1). While the term “discourse” is ubiquitous, it can be argued that the academic signifier and signified have a specific history that has developed through a large number of (sometimes interlocking) theoretical disciplines. Four schools are salient to the discipline of human geography: (i) a critical theory school; (ii) a neo-Foucauldian school (or what we might call a practice and performativity school); (iii) a critical linguistic school; and (iv) a discursive psychology (or reification) school.

The critical theory school

In the field of critical theory, studies of discourse are often related to what Jørgensen and Phillips have described as “critical research … [which] investigate[s] and analyse[s] power relations in society” and which “formulate[s] normative perspectives from which a critique of such relations can be made with an eye on the possibilities for social change” (Jørgensen and Phillips 2002, 2).

As a starting place it can be argued that theories of discourse in this tradition actually owe their beginnings to neo-Marxist reinterpretations of the concept of ideology. The work of the French neo-Marxist Louis Althusser (1918–1990) and his discussion of ideology has had an impact on a range of critical theories of discourse. Althusser suggests that ideology in a capitalist society is produced through the state and its institutions: what Althusser called ideological state apparatus (ISAs). Althusser also suggests that “ideology works through constituting (‘interpellating’) persons as social subjects, fixing them in subject positions, while at the same time giving them the illusion of being free agents” (Fairclough 1994/1992, 30).

Without a doubt, these early ideas of ideology had a major impact on a range of poststructuralist thinkers who developed new theories of power, discourse, and identity. One major commentator influenced by Althusser’s ideas was the French poststructuralist philosopher Michel Foucault (1926–1984). As opposed to viewing power as emerging from a singular source, Foucault argued that power is dispersed and is tied up with institutions. For Foucault, institutions refer to historical blocks of knowledge or epistemes. Foucault contends that an episteme is a system of knowledge or more appropriately a site (or framework) of discursive production. In short, epistemes, or discursive frameworks, define what can be understood as knowledge in a particular time period. Moreover, epistemes produce multiple discourses (as opposed to ideologies) that cannot be read as sites or conduits of bourgeois hegemony, but are sites of production and reduction. Thus, Foucault contends that power is not only a weighty force – it is not only a force
that prohibits – it is also a force that produces, such as knowledge and pleasure.

Foucault outlines a theory of power as dispersed and heterogeneous. In the first volume of *The History of Sexuality* he contends that “power is everywhere; not because it embraces everything, but because it comes from everywhere” (Foucault 1987/1976, 93). Thus, and moving away from foundationalist ontologies that locate power in one or a specific number of sites, power emerges from “innumerable points, in the interplay of non-egalitarian and mobile relations” (Foucault 1987/1976, 94). At each of these points (or discursive formations), then, discourses locate, position, and produce subjects. In developing this theory, Foucault is interested in the way subjects are located in historical frames or discourses. Like Althusser’s concept of ideology, discourses are not simply linked to the production of ideas; rather, they are also linked to the process of producing subjects, or subject positions.

**Foucault and geography**

Foucault’s ideas of discourse have had a very strong influence on geography, and geography’s poststructuralist turn in particular. During the late 1980s and early 1990s, Foucauldian ideas of discourse, and particularly Foucault’s genealogical work, had a reflexive influence on the way geographers began to look back at the history of the discipline. During this period scholars began to examine the role of imperialist, sexist, racist, and competitive discourses in the construction of the human geographical field. Commentators such as Jonathan Crush also explored discourses of progress in human geography and began to unpack the writing formulas that geographers operationalized in their texts.

By the early 1990s, more and more geographers began to utilize Foucauldian notions of discourse to reread ideas of space, place, and landscape. In the seminal *Writing Worlds* (1992), Barnes and Duncan began reframing ideas of landscape through Ricoeurian notions of text and Foucauldian notions of discourse. Importantly, the editors note that “Texts, in the broadest sense that we construe them, are constitutive of larger, even more open-ended structures termed discourses” (Barnes and Duncan 1992, 8). Arguably Barnes and Duncan’s textual-discursive readings of landscape set a precedent for a new era of human geographical work. Alongside these textual concerns, Foucauldian theories of discourse also had a large impact on development studies. Notably, the Colombian-American anthropologist Arturo Escobar views the concept of development itself as a discourse, as a “system that allows the systematic creation of objects, concepts and strategies; it determines what can be thought and said” (Escobar 1995, 40). In his text, *Encountering Development*, Escobar views development as a socially constructed knowledge which has maintained the hegemonic power of Western nations.

Foucauldian approaches to discourse (and particularly Foucauldian genealogical epistemologies) have also had a broad impact on the terrain of historical geography. Implicitly Foucauldian discourse analyses of modern English landscapes were a strong feature of research by historical geographers in the late 1990s, including the writings of David Matless and Catherine Brace. In the field of geopolitics a series of writers have sought to map out Foucauldian definitions of discourse to analyze foreign policy and/or political disputes. In the field of economic geography, Foucauldian methods of genealogical discourse analysis have developed in relation to the work of Gibson-Graham and the seminal text *The End of Capitalism (As We Knew It)*: A Feminist Critique.
of Political Economy (1996). Gibson-Graham traces the historical formation of discourses that enforce and maintain the capitalist system. Moreover, in the arena of spatial and environmental policy, Foucauldian discourse analysis (and particularly the idea of governmentality) has played a major role in understanding the way in which spatial and environmental texts are produced and constructed.

Foucauldian discourse analysis and geographical methods

Foucauldian definitions of discourse analysis have also played a part in the development of methodological discussions in contemporary geography. Notions of discourse analysis have featured as a discrete method in a number of geographical methods books. One innovative use of Foucauldian discourse analysis can be found in the work of Gillian Rose and her book Visual Methodologies, where Rose explicitly unpacks two forms of Foucauldian discourse analysis that she believes are pertinent to the practice of geography. The first of these, discourse analysis 1, she describes in terms of the “notion of discourse articulated through various kinds of visual images and verbal texts” (Rose 2016, 192). The second form of analysis, discourse analysis 2, Rose describes as analysis that pays “more attention to the material practices of institutions than it does to visual images and verbal texts. Its methodology is usually left implicit. It tends to be more explicitly concerned with issues of power, regimes of truth, institutions and technologies” (Rose 2016, 192). While many geographical methods textbooks gloss over Foucauldian discourse analysis in a vague way, Rose can be credited for sketching out the nuances of the analyses and the ways in which these theoretical tools can be applied to geographical and visual contexts.

Neo-Foucauldian schools: practice and performativity

Foucault’s ideas on discourse continue to be taken up by scholars in philosophy and the wider social sciences today. But of all the theorists who have utilized Foucault, Ernesto Laclau and Chantal Mouffe are regarded as key interpreters of Foucault’s original project. First, like Foucault, Laclau and Mouffe challenge traditional Marxist notions of ideology and reject the idea that society can be divided into “a surface or appearance of society [ideology] and an underlying reality [the truth of the capitalist system]” (Laclau and Mouffe 1985, 22); second, Laclau and Mouffe also reject the traditional Marxist contention that economic relations “determine political and discursive processes” (Howarth 2000, 100). Drawing upon Foucault (and Derrida and Lacan), Laclau and Mouffe define a discourse as “the fixation of meaning within a particular domain” (that is, the fixation of meaning in a field of meanings or truths) (Jørgensen and Phillips 2002, 26).

In this regard, “discourse is a reduction of possibilities. It is an attempt to stop the sliding of … signs in relation to one another and hence to create a unified system of meaning. All the possibilities that the discourse excludes Laclau and Mouffe call the field of discursivity” (Jørgensen and Phillips 2002, 27; emphasis in original). But, as Laclau and Mouffe contend, it is exactly “because a discourse is always constituted in relation to an outside, [that] it is always in danger of being undermined by it, that is, its unity of meaning is in danger of being disrupted by other ways of fixing the meaning of the signs” (Jørgensen and Phillips 2002, 27). Laclau and Mouffe suggest that discourse has a relational side: dominant discourses are always dependent upon other discourses (binary others), which unwittingly allows for openness, multi-meaning,
and instabilities in social and practical situations. But, discourse always works continually to try and close elements: these “are the signs whose meanings have not yet been fixed; signs that have multiple, potential meanings (i.e., they are polysemic)” (Jørgensen and Phillips 2002, 27).

Second, if the idea of a field of discursivity is important in Laclau and Mouffe’s theory, so is the notion of practices; indeed, in explaining Laclau and Mouffe’s ideas, Howarth has referred to the example of Thatcherism. Thatcherism represents a discursive set of ideas but, importantly, Thatcherism also refers to “a set of practices (‘strong leadership’ and ‘entrepreneurship’) and involved attempts to transform institutions and organizations, such as the British Conservative Party and the British state” (Howarth 2000, 103). Finally, and drawing upon Lacanian theory, Laclau and Mouffe also view subjects as split. That is: “The subject is fragmented or decentred; it has different identities according to those discourses of which it forms part” (Jørgensen and Phillips 2002, 43). As Jørgensen and Phillips contend, in Laclau and Mouffe’s ontology subjectivity is contingent: that is, subjects always have “the possibility to identify differently in specific situations” (Jørgensen and Phillips 2002, 43).

### A theory of discourse and performativity

If Laclau and Mouffe develop Foucauldian analysis through the idea of practices, the work of Judith Butler (2007/1990) has also been important to another very popular reading of Foucault’s work, and the idea of discourse in particular. Like Laclau and Mouffe, Butler also draws upon wider thinkers including Freud, Lacan, Kristeva, and Austin. In her classic work *Gender Trouble* (2007/1990), Butler critiques the idea of heteronormativity, a hegemonic discourse which prohibits, denies, or renders nonheterosexual sexualities as “illegitimate” or even polluting. In her study of the reproduction of heteronormativity (or compulsory heterosexuality), Butler suggests that discourses of gender intertwine with compulsory heterosexuality to reinforce the concept of heterosexuality as the only “natural” or “normal” form of sex.

Importantly, however, Butler argues that while heteronormative ideas of gender and sexuality are very powerful, they are haunted by sites of “disorganization and disaggregation … [which] disrupt the regulatory fiction of heterosexual coherence” (Butler 2007/1990, 185). Because of the continual instability of the fiction of gender coherence, Butler claims that subjects are always already involved in a never-ending drama of repetition and re-inscription whereby gendered identities must be continually performed. Rather than the hailing of a subject into a position, the subject in Butler’s ontology is always trying to maintain the idea of an “internal” (or essential) “core” or substance by continual acts of performativity, repetition, and “ritualized” legitimation (Butler 2007/1990, 191). Like Laclau and Mouffe, Butler also transforms Foucault’s original concept of discourse through the idea of practices, more specifically performativity.

### Practices, performativity, and geography

In recent years more and more geographers have moved beyond standard Foucauldian discourse analysis. Theories of discourse proposed by Laclau and Mouffe have featured in writings on qualitative geographical methods and geopolitics. In his article on “The disciplinary effects of communicative planning in Soufriere, St Lucia,” Pugh utilizes Laclau and Mouffe’s discursive approach to unpack “the importance of the Contingent articulation of a discourse in specific places at specific times” (Pugh 2005, 310). Here, Laclau and Mouffe’s utilization of the concept
of contingency allows Pugh to pay attention to the way “in which language brings into being signed differential subject positions, partially, but never completely, fixing identity around nodal points” (Pugh 2005, 311).

Butler’s notions of performativity have also had a huge impact on cultural geographers as well as geographers in the fields of gender and sexuality studies. Articles in the journals Gender, Place and Culture, Environment and Planning D, and Progress in Human Geography have often taken up discussions of performative theory in relation to issues of gender and sexuality; however, more recently discussions of performativity have also entered into debates concerning the nonrepresentational (and more than representational) in cultural geography; thus, writers such as Catherine Nash and others have pointed to the relevance of performative theories in reorienting “cultural geography towards practices rather than representations” (Nash 2000, 660).

The critical linguistics school

So far we have examined the field of critical theory and have drawn attention to a series of Marxist, poststructuralist, and performative theories. A third school of discourse theory and analysis belongs to those writers who have drawn upon the writings of scholars within traditional linguistics and the critical social theoretical schools discussed above (schools 1 and 2). Loosely speaking, in traditional linguistics the idea of discourse refers to the way subjects utilize the knowledge they already possess about language. Critical linguistics (CL), or the critical discourse analysis (CDA) school, has close links to these traditional linguistic concerns, but often charts its critical social theoretical influences back to the writings of Russian theorists such as Valentin Voloshinov (1895–1936) and Mikhail Bakhtin (1895–1975) who sought to conceptualize language in terms of wider social processes. Moreover, commentators in the CDA school also refer to the Frankfurt school and the work of Antonio Gramsci as influences on their work.

The CDA school began to take shape and develop in the late 1960s and 1970s, and can be associated with the work and writings of Michel Pêcheux, Roger Fowler, Bob Hedge, Gunther Kress, Toney Trew, Norman Fairclough, Teun A. van Dijk, and Ruth Wodak, among others. Like the critical theorists of the Frankfurt school, the CL or CDA school aims to unpack the relations between language and power and to understand the way hegemonic groups normalize their own interests within language. Also, like the Frankfurt theorists draw inspiration from, the CDA school advocates a transformative approach to contemporary society and the hegemonic institutions within society. For writers in the CDA school, the critical discourse analyst must provide alternative solutions to dominant discourses and hegemonic society. One influential theorist in this school is Norman Fairclough, whose version of CDA often focuses on a technologization of discourse: that is, in a very Foucauldian way, a reading of discourse (and, importantly, of language and words) as a mechanism of social control in the “engineering of social and cultural change” (Fairclough 1994/1992, 8). Fairclough is often concerned with the naturalization of discourse and its “widespread acceptance” in the social world (10). In this regard, Fairclough’s research has often investigated the commonsense nature of language and the way institutions such as the media construct their texts as neutral (or objective) (see Fairclough 1994/1992, 108).

Critical linguistics and geography

Geographers have also started to explore the work of writers and thinkers in the CDA school.
Commentators pointed to a “linguistic” turn in urban studies inspired by the work of Fairclough, Wodak, and Van Dijk. In her article, “Urban Geography: Discourse Analysis and Urban Research,” Lees talks about a distinct “discursive turn” based upon poststructuralist Foucauldian theory and Gramscian notions of discourse. In relation to the latter, Lees maintains that the Gramscian school (which we have called the CDA school) constructs the idea of discourse as being “almost synonymous with ideology itself in so far as it functions to conceal the power of vested interests and to induce the consent of the dominated to their own domination” (Lees 2004, 102). What she suggests here is that often this research has “emphasized the role of ‘discourse coalitions’ in urban politics and policy” which draws upon a CDA style methodology; that is, research that “involves the close semantic scrutiny of rhetoric and turns of phrase to discover particular narrative structures, issue framings and how storylines close off certain lines of thought and action at the expense of others” (Lees 2004, 102).

The discursive psychology (or reification) school

A final large school in the field of discourse theory emanates from the discipline of discursive psychology. This school can be associated with the work of Michael Billig, Jonathan Potter, Margaret Wetherell, and Derek Edwards, among others. While a number of key writings have been associated with this approach, Jonathan Potter and Margaret Wetherell’s Discourse and Social Psychology: Beyond Attitudes and Behavior (1987), has been regarded as a seminal text. Like traditional ethnomethodologists, Potter and Wetherell draw upon the idea of language as a form of action or commitment; however, where ethnomethodologists and conversational analysts see language as a tool to refer to entities in the “real world,” Potter and Wetherell start from the Foucauldian notion that language is discursively generated. In a form of ethnomethodological logic, Potter and Wetherell are interested in processes of reification: that is, the reality-generating aspects of discourse. Rather than simply analyzing the psychology of language, Potter and Wetherell explore the way in which reality becomes solidified and reified by subjects. As they state at the end of Discourse and Social Psychology, discourse analysts ask questions such as: “What procedures are used to authorize factual accounts? How are factual accounts used to perform a specific act? How is the effect of ‘mere description’ generated in discourse?” (Potter and Wetherell 1987, 182). Furthermore, as commentators have suggested, their work is not only concerned with the construction of reality and the authorization of factual accounts, but it is also concerned with the undermining and unsettling of reifying discourses.

The discursive psychology “reification” school and geography

It might seem that the discursive psychological school has made little impact on human geography, where figures like Foucault, Laclau and Mouffe, and schools like the CDA school have made significant inroads; however, a closer examination of a range of papers within the field, over the last 20 years, demonstrates that the discursive psychology school has indeed appeared in a number of geographical papers and writings. References to the work of Potter and Wetherell have been taken up by numerous authors in various subdisciplines of the subject. In their article, “Naming as Norming,” Berg and Kearns utilize Potter and Wetherell’s approach to
discourse because it allows for “a double movement between styles of reading that emphasize the constitution of subjects and objects, and those that emphasize the ideological work of discourse” (Berg and Kearns 1996, 103). Berg and Kearns have shown that in their ideological and reality-producing guise, Potter and Wetherell’s insights are really valuable to human geographers who have sought to look at commonsense notions around race, gender, and sexuality.

SEE ALSO: Bodies and embodiment; Critical geography; Environmental discourse; Governmentality; Identity; Orientalism/Occidentalism; Poststructuralism/poststructural geographies; Social constructionism; Subjectivity

References


Discrete global grid systems

Perry R. Peterson
The PYXIS Innovation, Canada

Discrete global grid systems (DGGS) are spatial references that use a hierarchical tessellation of cells to partition and address the entire globe. Figure 1 shows an example, using the ISEA3H (icosahedral Snyder equal area aperture 3 hexagonal) grid of nested hexagons. DGGS are designed to portray real-world phenomena by synthesizing digital values on a common discrete geospatial data structure. They are commonly used to create virtual globes. The adoption of an optimized DGGS is considered a favorable choice for the establishment of distributed Digital Earth information systems (Goodchild 2000).

DGGS differ from conventional geographic coordinate reference systems. A DGGS is designed to be an information grid, not a navigation grid. Conventional coordinate reference systems address the globe with a continuous lattice of points suitable for repeatable point-to-point navigation and analytical geometry. DGGS, however, address the entire planet by partitioning it into a hierarchical tessellation of nested cells, thereby providing a reference frame for repeating the location of measured Earth observations, feature interpretations, and extrapolated predictions.

A DGGS provides a digital framework for geospatial information. Conventionally referenced geospatial information is quantized or sampled into a discrete, aligned, discontinuous representation of the source information. This analog to digital conversion provides benefits exemplified by other digital signal technologies such as digital imagery, digital audio, and digital video. Data values are independent of the original data source – spatial reference system, spatial scale, or data type – vector or raster imagery. Information integration, decomposition, and aggregation can be optimized in the hierarchical structure, which can be exploited to support data processing, storage, discovery, transmission, visualization, computation, analysis, and modeling.

Formal development of DGGS began in the 1980s with the promising value of global analysis coinciding with the increased use of geographic information systems and availability of global mapping data and positioning systems. DGGS are characterized by their tiling designs, cell address, quantization strategy, and associated mathematical functions. The tessellation of cells provides discrete, nearly uniform fixed areas to assign values that describe the Earth – from backyard birdbaths to entire continents.

Spatial resolution is explicit as every item of information in a DGGS is associated with an area. The grid resolution can be designed to represent arbitrarily sized locations. Properly indexed and processed, data stored as a multilevel global grid can integrate geospatial data layers captured at different scales and resolutions at appropriate degrees of accuracy. Meaningful use of data stored at different resolutions can be accomplished without the need to conflate between scales. For the end user, combining or integrating layers becomes a free operation because items of information are automatically aligned. Topology is explicitly defined within the relationship of cells and their values within a DGGS.
There are early records of using grid patterns to organize measurements of land parcels, astronomical observations, and cartographic information. Most charting and map projections rely on transforming a grid on a globe to a plane. Artillery surveys provide accurate targeting on grid quadrants. Requirements for digitally organizing national maps produced innovation in map tiling and spatial data indexing systems that now drive spatial databases and web maps. At a global level, computational grids developed for fluid flow models supported climate, weather, and ocean current predictions. However, none of these approaches to Earth gridding evolved directly into the information-centric Digital Earth reference of a DGGS.

In 1986, Waldo Tobler and Zi-tan Chen identified an opportunity to create a global grid, the primary purpose of which was information storage. Tobler – of Tobler’s first law of geography: “Everything is related to everything else, but near things are more related than distant things” – argued that as geographic information systems are primarily for planning, analysis, and inventorying of geographic phenomena, it follows that “coverage must be uniform and that every element of area must have an equal probability of entering the system. This suggests that the world should be partitioned into chunks of equal size” (Tobler and Chen 1986). Tobler suggested an approach with a cube as a base polyhedron and dividing it into rectilinear quadtrees to create successive subdivisions with unlimited resolution. Tobler’s paper referenced Geoffrey Dutton, a software researcher working at the Laboratory for Computer Graphics and Spatial Analysis (LCGSA) at Harvard Graduate School of Design, who had done similar work, first using squares, then triangles.

Harvard’s LCGSA was a fertile birthplace of ideas related to digital thematic mapping, spatial analysis, and what is now called geographic information systems (GIS). Early work at LCGSA set some groundwork for using discrete models to characterize geospatial information. From 1967 to 1972, under the direction of William Warntz, LCGSA produced a series of papers on geography and the properties of surfaces that would lay a foundation for digital mapping, including several notable papers examining the properties of mixed hexagon tessellations to represent central places and networks. In 1967, Carl Steinitz, then an associate professor at Harvard’s Graduate School of Design, applied LCGSA-developed computer mapping software – SYMAP (Synteny Mapping and Analysis Program) – to an aggregation of various vector maps to facilitate an urban planning study of the Delmarva Peninsula. Steinitz found while integrating the various vector-based data sources...
that gridding the data was more efficient. A new raster-based module to SYMAP, coined GRID, resulted. GRID went on to become IMGRID – the basis of work by another Harvard student, Dana Tomlin, the originator of Map Algebra, whose raster-based analysis software would be integrated into ERDAS (Earth Resources Data Analysis System) Imagine and the US Army Corps of Engineers’ Map Analysis Package (MAP) in GRASS (Geographic Resources Analysis Support System).

Geoffrey Dutton came to Harvard for a master’s degree in urban planning in 1966 and worked as a research assistant at LCGSA until 1989. Dutton focused much of his attention on raster-based data structures and three-dimensional (3-D) terrain visualization. Using a time series of 2.5-D gridded surfaces he created the first animated holographic map. In 1984, Dutton published his work on assembling and managing global terrain data on a triangular global grid. His global geodesic elevation model (GEM) started with a cuboctahedron connected into a rhombic dodecahedron and recursively divided the initial 12 triangular faces into refinements of nine partially nested equilateral triangles. Elevations were assigned to the faces of each successive triangle.

Dutton’s paper did more than lay out the mathematics for a global grid; it emphasized why he felt such an effort was necessary: “we live on one planet, a world which may be regarded as an organic entity. The wholeness of the Earth – self-evident to many non-western cultures – largely escapes the western culture of science, even the so-called earth sciences. This peculiar myopia is reflected both by our disciplinary specialization and by territorial concerns.” Dutton believed assembling relevant data in a holistic, globally useful framework could act as a unifying model: “[it is] only through using technology that we have any hope of comprehending the nature of the problems we have wrought for the planet” (Dutton 1984).

A few years later, Dutton modified the GEM model to a simpler structure called quaternary triangular mesh (QTM). QTM was based on an octahedron embedded in the Earth with a vertex at each pole and four vertices at the equator.

In 1975, Denis White, who had been a computer scientist at the Massachusetts Institute of Technology, joined LCGSA at Harvard to co-design and develop the vector-based GIS, Odyssey. During that time, White learned of Michael Woldenberg’s work on mixed hexagonal hierarchies and collaborated with Geoffrey Dutton. White was familiar with Dutton’s vision for a global data structure when he left Harvard to work at the US Environmental Protection Agency’s (EPA) ecological research lab in Corvallis, Oregon.

A cluster of research offices in Corvallis, Oregon, which included the University of Oregon, US-EPA, and US Department of Agriculture-Forestry Service, fermented and fostered the next concerted development to formalize global grids into a system. The basic idea was driven by a need to develop an integrated approach to monitoring aquatic and terrestrial resources in the United States, the objective of the EPA’s new Environmental Monitoring and Assessment Program (EMAP). The concept of using an area frame as the basis for a survey design was proposed by Scott Overton, a statistical ecologist at the University of Oregon. Statistically valid sampling was a very important criterion for EPA. White, working with Overton, proposed that EMAP use a global system of hexagonal cells. They shared their plans with Jon Kimerling, a professor of geography at Oregon State University who had extensive experience with map projections. Kimerling suggested a global grid that started with a truncated icosahedron (White, Kimerling, and Overton 1992).
By June 1989 the three convened a meeting in Corvallis to discuss the geometry of sampling designs for EMAP and included Tobler, Dutton, Hrvoje Lukatela, a Canadian scientist who had developed a global model using Voronoi areas, and Tony Olsen, a statistician at the Pacific Northwest Laboratory. Design criteria for a DGGS began its formalization. Michael Goodchild, of the National Center for Geographic Information and Analysis (NCGIA) and University of California, Santa Barbara, had met Dutton when he presented QTM at an NCGIA specialist meeting on spatial accuracy at Montecito in 1988. Goodchild could not attend the Corvallis meeting, sending instead a paper authored with his student Yang Shiren suggesting modification to Dutton’s QTM labeling. Dutton would later collaborate with Shiren to show, as is often the case with DGGS indexing methods, that their respective models were isomorphic, permitting mappings between the two numbering systems.

The choices of three regular tilings for a global grid – quadrilateral, triangular, and hexagonal cells – were represented at the 1989 Corvallis meeting with the challenges for a further decade of global grid research established.

Discrete global grid system design

The Open Geospatial Consortium (OGC) proposes a formal definition for a DGGS “as a spatial reference system that uses a hierarchical tessellation of cells to partition and address the globe. Discrete Global Grid Systems are characterized by the properties of their cell structure, geo-encoding, quantization strategy and associated mathematical functions.” For an Earth gridding system to be compliant with the OGC’s standard definition of a DGGS it must be a fixed tessellation with its cells covering the entire Earth at multiple levels of granularity. The system must include a method of addressing each cell, a mechanism for sampling data values into the cells, and known algebraic operators for expressing and transforming these values.

There are many possible DGGS, each with their own advantages and disadvantages. Criteria for choosing an appropriate tessellation include properties of shape, adjacency, connectivity, orientation, self-similarity, and decomposability, as well as packing properties. There are only three shapes that provide regular tiling of the plane: quadrilateral, triangle, and hexagon. All methods must deal with the fact that a perfect regular partition for the surface of a sphere does not exist, except in the form of the five Platonic solids, each with a single level of granularity: the tetrahedron, cube, octahedron, dodecahedron, and icosahedron.

Quadrilateral cells are an attractive choice as they are in common use within raster imagery and textures, hardware devices, and image-processing algorithms. They are widely adopted in raster data structures. Triangular cells can be rendered very efficiently and are supported by many built-in graphic mesh functions. As a lattice they provide an efficient network connectivity for interpolation of values such as elevation models. Hexagonal cells are favored for their statistically optimal sampling, close packing, and uniform adjacency. They are a preferred shape for modeling dynamic environmental systems. Quadrilaterals and triangles can be decomposed or refined into self-similar congruent shapes to form hierarchical structures. Triangular lattices and hexagonal tessellations form Delaunay/Voronoi duals and so transformations between them provide dual utility. Essentially, squares are familiar, triangles are fast, and hexagons are the finest fit. Each has its own
advantages within the context of DGGS (Sahr, White, and Kimerling 2003).

The generation of a tessellation on a spheroid is mathematically intensive. Most methods start with a regular polyhedron and then project the cells to the spheroid. Preserving cell size or shape is an inevitable compromise in spherical geometry because projected shapes cannot be both of equal area and conformal. Preservation of cell area is a preferred trait when a DGGS is intended to represent information consistently across the entire globe at the same resolution. There has been some work developing voxel-based 3-D DGGS; others rely on the attribution of data values to extend an elevation or time dimension.

Each cell must have a unique identifier. Hierarchy-based, space-filling curves, and axes methods of indexing have been used to uniquely address cells. Hierarchical data structures that provide the basis for nearest neighbor, fast linear ordering, and parent–child relationships are the most common. Though choices can be limited by the type of tessellation, geocoding types can generally be transformed from one to the other (Mahdavi-Amiri, Samavati, and Peterson 2015).

Strategies for sampling content into cells are an important aspect of encoding information to a DGGS. There is normal loss when quantizing a continuous signal to a discrete digital signal. This principle also applies to DGGS. Some DGGS choices are therefore optimized for high fidelity and statistically valid sampling. For example, one cannot assume that quadtree indices will be more accurate with each additional refinement; centroidal points used to define vector geometries (lines and polygons) can move further away from a real value with a refinement. However, when centroids of child cells align with a parent’s vertices, as they do on some hexagonal tessellations, the resulting hierarchical indices will converge monotonically as it refines. This gives those tessellations the characteristics of a Cartesian coordinate lattice where each decimal place can indicate a level of accuracy. Similarly, the finer the change in cell area between resolutions the generally more efficient the sampling rate can be to meet the Shannon–Nyquist theorem on digital representation of an analog signal.

DGGS provide a uniform environment to integrate and visualize both vector geometry and raster-based geospatial data sources in much the same way that information within a computer graphics pipeline becomes the pixels on a computer screen. Efficiencies are gained as operations and algorithms can be created on the data structure itself, independent of the data sources. Typically, a DGGS will include Boolean, relational, and image algebra, mathematical morphology, frequency transforms, and other image processing operations. The structure of the DGGS provides opportunities to simplify most spatial relationships into set theory operators. Topological relationships between two regions can be described within an efficient binary matrix such as the Dimensionally Extended 9 Intersection Model. Predictive modeling techniques such as finite element, agent-based, or cellular automaton can rely on the cell structure, operations, and data integration characteristics of the DGGS.

The common conceptions of a grid can lead to misidentifying any general gridding schemes as a DGGS. As an example, single resolution computational grids cannot be considered to qualify as DGGS, although a DGGS could serve the dual purpose of a computational grid. Similarly, spatial data indexing used to organize map tiles or optimize rapid spatial analysis cannot be considered to qualify as a DGGS in and of itself; although DGGS often utilize hierarchical indices to identify a cell, the primary feature of the DGGS is the cell geometry, not the optimization of a spatial query.
DISCRETE GLOBAL GRID SYSTEMS

A decade of development

In March 2000, the pioneers of DGGS convened at the 1st International Conference on DGGS in Santa Barbara. Notable additions to the main group included Tony Olsen, Kevin Sahr, and Dan Carr.

Tony Olsen came to the EPA in 1990 to lead the statistical design of EMAP. Dan Carr, a statistician from George Mason University, had developed innovative data binning techniques using hexagons and contributed to the cartography of hexagonally partitioned data. White and Olsen worked closely on implementations of the original EMAP grid system, but after several years the disadvantages of the truncated icosahedron hexagonal grid encouraged further research and development. The importance of an equal area cell had become apparent, so Kimerling turned to John P. Snyder, the clever self-taught geodesist and author of the definitive text on map projections, for a method of projecting equal area cells from the face of a polyhedron to a sphere (Snyder 1992). Subsequent work at the EPA and Oregon State then focused on the icosahedron as the preferred solid model upon which to develop global grids.

Funding was made available to add a computer scientist to compile and test the ideas that were being generated for a truly global equally valid sampling grid. Kevin Sahr had completed his master’s degree in computer science from the University of Colorado in 1995. Sahr was an admirer of Buckminster Fuller, his Dymaxion technology, and his global vision of “Spaceship Earth.” Kevin had picked up on the military use of hexagonal tessellations while serving in the Reserve Officers’ Training Corps in Army intelligence and was a keen champion of a hexagon Earth. Sahr coined the term “discrete global grid” and would become a passionate advocate and technical expert on DGGS. “The term ‘discrete global grid’ was proposed by Kevin Sahr to identify the types of global partitioning systems based on continuous coverages of polygons. It was first used in print in a draft technical report from the OSU Geosciences Department Terra Cognita Laboratory, where the EPA funded projects on biodiversity and global grids were located. Kevin wrote the report, dated 30 May 1996, entitled ‘Terra Cognita Technical Report 96001: Discrete Global Grid.’ In September 1996, Kevin and others wrote a description of 14 desirable properties of discrete global grid that were proposed by Michael Goodchild in the 1994 meeting in Santa Barbara” (White 2014).

In 1988, the US National Science Foundation established the National Center for Geographic Information and Analysis to conduct basic research to advance geographical information systems. Many technical impediments existed. The geographer Michael Goodchild, serving as one of the NCGIA’s founding co-directors, produced a prolific series of papers, reports, and vision documents that earned him the reputation as the pre-eminent thought leader within geographical information science. In 1994, Goodchild had constructed a list of criteria describing a global grid.

The EPA research team at Corvallis had advanced significantly by the 1st International Conference on DGGS in 2000. At Oregon State University, Kimerling had received the Milton Harris award in 1996 for his exceptional achievement in basic research on developing global grids. A number of research areas had been explored and the preferred solution was identified as the ISEA3H grid (Sahr and White 1998). Dutton had advanced QTM and completed a doctoral degree at the University of Zürich with a dissertation thesis on the subject (Dutton 1999).

Goodchild’s research into DGGS systems followed Dutton’s QTM approach; however, his interest aimed at a more generalized and
lofty utility than Dutton’s original objective of global terrain modeling. Goodchild could envision an Internet subsystem for organizing and aligning all geospatial information which was sufficiently simple that a child could use it to explore the Earth. His vision was articulated in a 1998 speech, *The Digital Earth*, written with substantial input from Goodchild, that was to be delivered by Vice-President Al Gore (Gore 1998). For Goodchild, *The Digital Earth* provided the actionable requirements needed to respond to Al Gore’s vision of a digital replica of Earth proposed in his 1992 book *Earth in the Balance* (Gore 1992). Goodchild articulated the connection between DGGS and Digital Earth at the 1st International Conference on DGGS: “[Digital Earth] can clearly benefit from developments in discrete global grid, which can provide the georeferencing, the indexing, and the discretization needed for geospatial data sets. They have properties, in particular hierarchical structure, uniqueness, explicit representation of spatial resolution, and consistency, that make them superior to any single alternative” (Goodchild 2000).

Keith Clarke, a geographer at the University of California, Santa Barbara, prepared and presented at the conference a comparison of criteria for assessing a DGGS (see Table 1).

### Applications for discrete global grid systems

The growth of the web as a primary source of information dissemination has naturally led to the evaluation of different approaches to efficient online cartography and spatial analysis. Microsoft, as an example, used Dutton’s QTM as a base structure for their early digital mapping program Encarta and by 2003 they had investigated other DGGS as central data structures for a web-based media geocoding standard they called the Worldwide Media Exchange (WWMX). However, the opportunity for fast display of global maps on the web led Google to adopt a much more simplified approach to digital mapping. In 2005, Google Maps was released using Web Mercator — a spherical variant of the Mercator projection suitable for partitioning and indexing map tiles for rapid transmission and display. Despite its controversial nonconformance, it has become the de facto standard for web maps. Similarly, speed and visualization trumped analytical capacity when Keyhole (now Google Earth) introduced a virtual globe that used mip-map to create mapping tiles for fast rendering of a multiresolution 3-D globe.

Trends indicate emerging requirements that cannot handle these solutions. Geoscientific data and mass-market location data of petabyte scale are rapidly heading toward the exabyte-scale range. Decision-makers expect fast access to big, distributed, and dynamic geospatial data archives in a form that answers their geospatial questions without the assistance from a middle data integrator. Combining multiple sources of geospatial information — a necessary and expensive step in the geospatial intelligence cycle — is not solved by these advances. Traditional approaches to the acquisition, management, distribution, and application of mapping data are not able to meet the Digital Earth vision that presupposes a young child searching and exploring a virtual globe for meaningful insights on demand.

The deployment of DGGS is providing a more comprehensive and scalable solution for distributed Digital Earth information systems. By conversion of traditional Earth observation data archives into standardized DGGS, massive amounts of data are made available to scientists as timely decision-support products in a transparent and repeatable fashion. Very large multiresolution and multidomain datasets are aligned and ready for distributed and/or high-performance
### Criteria for DGGS characteristics proposed by both Goodchild and Kimerling, à la Keith Clarke 2000.

<table>
<thead>
<tr>
<th><strong>DGGS criteria in Goodchild (Goodchild 1994)</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G1. Each area contains one point.</td>
<td></td>
</tr>
<tr>
<td>G2. Areas are equal in size.</td>
<td></td>
</tr>
<tr>
<td>G3. Areas exhaustively cover the domain.</td>
<td></td>
</tr>
<tr>
<td>G4. Areas are equal in shape.</td>
<td></td>
</tr>
<tr>
<td>G5. Points form a hierarchy preserving some (undefined) property for ( m &lt; n ) points.</td>
<td></td>
</tr>
<tr>
<td>G6. Areas form a hierarchy preserving some (undefined) property for ( m &lt; n ) areas.</td>
<td></td>
</tr>
<tr>
<td>G7. The domain is the globe (sphere, spheroid).</td>
<td></td>
</tr>
<tr>
<td>G8. Edges of areas are straight on some projections.</td>
<td></td>
</tr>
<tr>
<td>G9. Areas have the same number of edges.</td>
<td></td>
</tr>
<tr>
<td>G10. Areas are compact.</td>
<td></td>
</tr>
<tr>
<td>G11. Points are maximally central within areas.</td>
<td></td>
</tr>
<tr>
<td>G12. Points are equidistant.</td>
<td></td>
</tr>
<tr>
<td>G13. Edges are areas of equal length.</td>
<td></td>
</tr>
<tr>
<td>G14. Addresses of points and areas are regular and reflect other properties.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>DGGS criteria in Kimerling (Kimerling, Sahr, White, and Song 1999). Goodchild’s numbers given in parentheses</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>K1. Areal cells constitute a complete tiling of the globe, exhaustively covering the globe without overlapping. (G3,G7)</td>
<td></td>
</tr>
<tr>
<td>K2. Areal cells have equal areas. This minimizes the confounding effects of area variation in analysis, and provides equal probabilities for sampling designs. (G2)</td>
<td></td>
</tr>
<tr>
<td>K3. Areal cells have the same topology (same number of edges and vertices). (G9, G14)</td>
<td></td>
</tr>
<tr>
<td>K4. Areal cells have the same shape. Ideally a regular spherical polygon with edges that are great circles. (G4)</td>
<td></td>
</tr>
<tr>
<td>K5. Areal cells are compact. (G10)</td>
<td></td>
</tr>
<tr>
<td>K6. Edges of cells are straight in a projection. (G8)</td>
<td></td>
</tr>
<tr>
<td>K7. The midpoint of an arc connecting two adjacent cells coincides with the midpoint of the edge between the two cells.</td>
<td></td>
</tr>
<tr>
<td>K8. The points and areal cells of the various resolution grids which constitute the grid system form a hierarchy which displays a high degree of regularity. (G5,G6)</td>
<td></td>
</tr>
<tr>
<td>K9. A single areal cell contains only one grid reference point. (G1)</td>
<td></td>
</tr>
<tr>
<td>K10. Grid reference points are maximally central within areal cells. (G11)</td>
<td></td>
</tr>
<tr>
<td>K11. Grid reference points are equidistant from their neighbors. (G12)</td>
<td></td>
</tr>
<tr>
<td>K12. Grid reference points and areal cells display regularities and other properties which allow them to be addressed in an efficient manner.</td>
<td></td>
</tr>
<tr>
<td>K13. The grid system has a simple relationship to latitude and longitude.</td>
<td></td>
</tr>
<tr>
<td>K14. The grid system contains grids of any arbitrary defined spatial resolution. (G5,G6)</td>
<td></td>
</tr>
</tbody>
</table>
parallel-processing computer environments. Geospatial data integration on demand is a grand challenge that DGGS have solved.

A recent survey of DGGS indicated that there are at least 10 DGGS in active development and use. PYXIS WorldView, developed in collaboration with Natural Resources Canada and Defence Research and Development Canada, is one example. The web-based client application uses the ISEA3H DGGS to enable complex spatial queries over multiple geospatial data streams on demand. SCENZ-Grid DGGS, developed by Landcare Research New Zealand, is a rectangular DGGS based on 3×3 tessellations of the six faces of a hierarchical equal area iso-latitude pixelated ellipsoidal cube (HEALPix) designed for use in high-performance computing and cloud architectures for interdisciplinary environmental modeling.

**SEE ALSO:** Data structure, raster; Digital Earth; Geocoding; Geodesy; Geographic information science; GIS: history; Indexing; Map projections and coordinate systems; Open Geospatial Consortium standards; Representation: 3-D; Spatial analysis; Spatial data infrastructures; Spatial organization and structure; Spatial tessellations

**References**


**Further reading**

DISCRETE GLOBAL GRID SYSTEMS

Disease and illness, political ecology of

Ezekiel Kalipeni
*University of Illinois at Urbana-Champaign, USA*

Jayati Ghosh
*Dominican University of California, USA*

Joseph Oppong
*University of North Texas, USA*

Political ecology is an analytical social theory framework that focuses on the relationships between political, economic, and social factors and environmental and health issues (Peet and Watts 1996, 6). It was developed to examine a wide ranging set of phenomena spanning the human–environment nexus and interactions and has led to better systemic understanding of health and disease. This framework has become an increasingly attractive approach because of its capacity to handle specific questions that deal with complex historical and contemporary issues at varying scales of analysis. Its power lies in its ability to combine the concerns of ecology on the one hand and political economy on the other. As such, political ecology can be defined as cultural ecology plus political economy. Steward (1972) defined cultural ecology as “...the study of the processes by which a society adapts to its environment.” Indeed, it is the study of the ways in which different human groups interact with, and relate to, their environment in order to survive and/or change. Culture is viewed as the primary mechanism by which human societies adapt to the environment. With reference to disease, different cultures develop different adaptations to their environment that produce or prevent the spread of disease.

Examples of such cultural adaptations might include “eating bush meat” and the onset of the HIV and/or Ebola epidemics, the drinking of hot tea in India to prevent outbreaks of cholera, or houses in Indonesia built on stilts to keep malaria-causing mosquitoes at bay. In Africa, bush meat is meat from wild animals hunted in the forests and consumed for sustenance. The wild animals are from a range of species, which include great apes, bats, porcupine, African squirrels, and so on. Many of these animals are known to harbor various pathogens, including the simian virus and Ebola. Research has shown that human immunodeficiency virus (HIV), the virus which causes AIDS, originated from a similar virus in primates called simian immunodeficiency virus (SIV); researchers believe that HIV probably initially jumped into humans after people in Africa came into contact with infected bush meat. Similarly, Ebola is believed to have spread from bats to humans through the consumption of Ebola infected bats. This is an example of a case of disease exposure due to human consumption of wild meat. The second example of cultural adaptation is the customary drinking of hot tea in China and India. Because water is often contaminated with cholera-causing bacteria, drinking hot tea is safer. The health benefits of drinking hot tea have made it popular in India today and definitely increased longevity by keeping cholera, dysentery, and typhoid at bay. In Southeast Asia, particularly Indonesia, stilt houses built primarily as a protection against flooding also serve to keep out vermin and insects such as mosquitoes. Living in a raised house is...
DISEASE AND ILLNESS, POLITICAL ECOLOGY OF

a cultural adaptation that lowers the risk of mosquito bites and the associated risk of malaria.

These are examples of cultural adaptations that may result in the spread or control of the diseases in question. On the other hand, political economy is the study of the interaction of politics and economics to explain how political institutions, the social and political environment, and economic systems influence each other. The central argument of political economy is that resources should be distributed based on social, economic, or environmental justice. Thus, people deserve a share of resources because of long-standing property claims/customary use, the prior investment of labor, and/or citizenship status. Those few members of society that control the means of production do well while most of the poor suffer from the effects of poverty, including ill or poor health. With reference to political economy and how it relates to disease, the behavior and dynamics of visible and invisible actors within the health-care sector can only be understood in terms of their power and class position in the larger social system (Meade and Emch 2010). Those who apply the political economy approach to health-care problems often look to the political, economic, structural constraints, and historical contexts to explain changes to health and social issues. Indeed, for the less fortunate such as poor farmers, health problems and poverty intersect and create complex causal chains with synergistic effects. For example, diseases such as schistosomiasis, tuberculosis and Ebola may result in debilitation or even the death of the main provider in the household, resulting in poor human nutrition. On the other hand, poor human nutrition and poor soils will undermine productivity, which may undermine the household economy and the ability to afford medical treatment, with spiraling deleterious consequences on the household and its well-being. It is in this regard that political ecology of disease is defined as political economy plus cultural ecology. In a nutshell, political ecology of disease is the study of how power relations and political and economic factors (political economy) impact environmental and health-care issues (ecology) and the ways people interact with the environment (culture).

Terminology of public health, epidemiology, and medical geography

Researchers that employ the political ecology approach to explain both distal and proximate causes of epidemics and diseases usually place their research within the disciplines of medical geography, public health, and epidemiology. Meade and Emch (2010) define medical geography as a discipline that uses concepts and techniques of geography, an integrative, multidisciplinary, and holistic discipline that deals with human–environment interactions. There are many facets of medical geography that examine the geographical patterns of health and disease from geographic and population points of view, including health geography, disease ecology, epidemiological geography, public health geography, and so on. As a matter of fact, medical geography can be considered the spatial side of public health but with emphasis on population, place, and time. Nevertheless, medical geography distinguishes itself from public health and epidemiology by acknowledging the distinctiveness, uniqueness, and importance of “place,” integrating its research at multiple scales and addressing the core questions of geography, that is, “where” and “why there.”

The World Health Organization (2014) defines public health as all organized measures (whether public or private) to prevent disease, promote health, and prolong life among the population as a whole. Its activities aim to provide conditions
in which people can be healthy and focus on entire populations, not on individual patients or diseases. As such, public health is concerned with the total system and not only the eradication of a particular disease. Further, the World Health Organization defines epidemiology as the study of the distribution and determinants of health-related states or events (including disease), and the application of this study to the control of diseases and other health problems. In essence, epidemiology is the cornerstone of public health; it informs policy decisions and evidence-based practice by identifying risk factors for disease and targets for preventive health care. While medical geography is intricately related to public health and epidemiology, its emphasis is on spatial or geographic aspects of disease. It is, therefore, important to define briefly several public health, epidemiology, and medical geography terms that are used to illustrate how the political ecology framework is applied to diseases.

The three disciplines, public health, epidemiology, and medical geography, use several medical and environmental terms to illustrate the different ways that diseases and interrelated health conditions, such as malnourishment, are transmitted or acquired. The case studies and ensuing discussion in this entry make reference to some of these medical terms, which need to be defined in advance, for example, the terms communicable or infectious disease, infectious agent, host, vector, reservoir, zoonosis, waterborne and airborne, toxic exposure, endemic, epidemic, and pandemic. A vector is a living agent that transmits pathogens, for example, arthropods such as ticks, mites, fleas, and lice and insects such as mosquitoes. On the other hand a fomite is an inanimate object that can transfer pathogens to a host when contaminated (e.g., surgical instruments). A reservoir of pathogen or disease agent can be a geographic site (natural nidus) or an animal host that carries the pathogen but does not get the disease. Once identified, natural reservoirs provide the complete life cycle of infectious diseases, providing effective prevention and control. For example, raccoons, foxes, and bats are natural reservoirs for rabies while rats and squirrels are natural reservoirs for bubonic plague. Another related term is zoonosis, which is a disease that occurs primarily in animals but can be transmitted to humans, for example, swine influenza virus (swine flu) transmitted from pigs to humans or Ebola transmitted from bats and other wild animals.

The transmission of several other diseases does not depend on vectors; rather they are waterborne and airborne or generated through toxic exposure. Waterborne diseases are caused by pathogenic microorganisms and most commonly transmitted in contaminated water, with infection occurring through bathing, washing, and drinking. A good example of a waterborne diarrheal disease is cholera. Airborne diseases are caused by pathogenic microbes that can be discharged from an infected person via coughing, sneezing, and close personal contact. Examples of such diseases include severe acute respiratory syndrome (SARS) and tuberculosis. Environmental pollution due to toxic chemicals and respiratory irritants can lead to poor health
through inhalation of the irritants. Toxic chemical compounds and aerosols can cause harm by a variety of different mechanisms, and the extent of injury can vary widely, depending on the degree of exposure and on the biochemical properties of the inhalant. People exposed to toxic chemicals are likely to develop varied ailments, including cancer, bronchitis, pulmonary hemorrhage, nasal inflammation, and pneumonitis.

Finally, endemic implies a disease is constantly present in a given locality, usually in low numbers, for example, chicken pox. But it can also be in sizable numbers, such as malaria in sub-Saharan Africa. An epidemic is when the occurrence of the disease is unusually high in a localized population. A pandemic is when a disease becomes a worldwide epidemic or epidemic that has spread to new geographic areas such as HIV/AIDS and the 1918 influenza outbreak. Emerging and re-emerging diseases, in contrast to endemic diseases, are those diseases that have recently appeared in a population or have existed but are rapidly increasing in incidence or geographic range (emerging). Re-emerging diseases are those diseases that once were health problems, and then declined, but are again health problems for a large proportion of a population, for example, tuberculosis. As Meade and Emch (2010) note, emergence is about changing and expanding spatiotemporal patterns of diseases.

Examples of the Aral Sea in Asia Minor and schistosomiasis in Senegal

Demonstrated in the remainder of this entry are how social, economic, and political influences have shaped land-use patterns, human–environment interactions, and disease emergence and spread in specific locations using the political ecology approach. We weave together a story of examples from different regions of the world that demonstrate clearly the workings of political ecology of disease. Two notable examples include the impacts on disease of dam building in Senegal and the disappearance of the Aral Sea in Asia Minor. During the mid-1980s, the Senegalese government decided to increase the production of rice by building the Diama dam at the mouth of the Senegal River. It had tragic ecological consequences, culminating in the outbreak of schistosomiasis (Bilharzia), a water-based disease. In Asia Minor, the Aral Sea Basin has seen its waters dwindle by 90% because of irrigation for cotton, a situation exacerbated by the five countries that find themselves in this basin. These countries have different sizes, political orders, conflicting political and economic interests, particularly in natural resource use.

These two examples offer two different case studies, one straightforward in its application of political ecology and the other more complex. The case of schistosomiasis in Senegal is a seemingly more straightforward example, with a single disease that has a well-defined mechanism for transmission, which is enhanced by a single development intervention (the dam). It illustrates clearly both proximate and distal causes of the disease, and the related potential solutions to the problem. By contrast, the second example of the Aral Sea is more complex in the sense that it does not highlight a single health problem at all; nor are the mechanisms of transmission all so straightforward. Instead, a macroscale environmental intervention produces a suite of complex often interrelated health problems due to environmental exposure. The proximate and distal cause distinction holds, but the epidemiological story is more complex. It is for this reason that we begin with the simpler story and deliberately move to the more complicated one. These two examples are discussed briefly before
moving to a general discussion and application of this framework.

**Schistosomiasis and the Diama Dam in Senegal**

Schistosomiasis, also known as Bilharzia, is a disease caused by a helminth (parasitic worm) of the *Schistosoma* type. Three types of these helminths are responsible for this disease, namely, *S. mansoni*, *S. haematobium*, and *S. japonicum*. Schistosomes are waterborne parasites that require aquatic snails as important intermediate hosts. The snails release free-swimming schistosomes into water; these are then attracted to the human skin. They often infect the urinary tract or intestines and symptoms include abdominal pains, diarrhea, bloody stool, or blood in the urine. In those who have been infected a long time, liver damage, kidney failure, infertility, or bladder cancer may occur. More than 700 million people are at risk of schistosomiasis worldwide, mostly children in poor, rural villages; Africa contains 97% of the world’s infected population (Hunter 2003). An estimated 200 million people worldwide are infected with the disease and over 20,000 deaths annually are associated with the severe consequences of infection.

There is convincing evidence that dam construction increased the habitat for, and decreased native predators of, the intermediate host snails in Senegal, contributing to the epidemic emergence of human schistosomiasis (Kuris 2012). In order to increase rice production, the Government of Senegal embarked upon the building of a dam at the mouth of the Senegal River in the early 1980s. The Diama Dam drastically changed the ecology of its water upstream, which resulted in the reduction of salinity and change from acidic to alkaline environment, promoting the growth of freshwater snails and affecting their natural predator, the prawns which all but disappeared. Creation of new irrigation canals and expansion of the rice fields provided new habitats for the snails, intermediate hosts of the schistosomes, to colonize; hence, a schistosomiasis epidemic in the affected areas.

The ecological damage caused by human activity in Senegal has unquestionably influenced the rise of schistosomiasis and adversely impacted human health. With reference to scale, decisions taken at the national level have impacted the rise of disease at the local scale. In spite of the knowledge that the disease outbreak was caused by ecological damage of the environment, public health campaigns rarely turn to ecological solutions to mitigate infectious disease risk, instead favoring traditional approaches such as vaccine development or drug treatment. The case of Senegal is an excellent example of how the political ecology approach can explain the rapid proliferation of disease and inform policy with reference to novel approaches to disease control strategies for schistosomiasis, one of the most significant parasitic diseases in Senegal today. It is possible to devise a strategy to mitigate the ecological damage caused by dam building in Senegal, thereby reducing the transmission of this disease that emerged and spread to epidemic levels following dam completion. Indeed, experiments are ongoing in Senegal to develop an ecologically sound intervention strategy that combines bio-economics, aquaculture, social science, ecology, and epidemiology to reverse the spread of schistosomiasis. As Jullien (2013) indicates, a harvestable native crustacean (prawns) suffered dramatic declines in the Senegal River ecosystem due to dam construction. These very same prawns also happen to be the natural predators of snails. Scientists are in the process of environmental re-introduction of the native prawn, using village-based aquaculture, as
a socially and economically sustainable control strategy for schistosomiasis (Jullien 2013).

Aral Sea

These two examples illustrate clearly the issue of scale and history in political ecology and set health issues within a broader social and economic context. Health problems at a local scale need to be connected to national and global decisions about health policies, funding allocation, land use and management. In other words, there is a need to examine carefully the effects of state policies on land-use change/management by starting at the “micro” or local scale and examining the historical context and local impacts of global and national level policies. Let us examine the case of the Aral Sea using the political ecology lens. The often asked question is what happened to the Aral Sea? If we go back to 1960, the Aral Sea was the world’s fourth largest inland water body. Although from the start the water environment was brackish and inhabited by mostly freshwater species, the lake supported major fishery operations and offered a key regional transportation route for the five countries that surround it, that is, Afghanistan, Kazakhstan, Tajikistan, Turkmenistan, and Uzbekistan.

By 2003, the Aral Sea had lost approximately 90% of its volume. Salinization caused massive loss of biodiversity and species. The Aral Sea Basin countries were part of the former Soviet Union before it broke up and during Soviet rule the allocation of water was based on the development master plan instituted in 1987 by the former Soviet Union’s Water Management Ministry. The intention of the former Soviet Union was to use the waters of the two major rivers (Syr Darya and the Amu Darya) that drain into the Aral Sea for irrigation purposes to increase cotton and food production (Kasperson, Kasperson, and Turner 1995). While in the Soviet era peace and calm was maintained from Moscow, the break up has resulted in rising tension regarding water allocation, water delivery, and sustainable development. There is no interstate agreement between the countries with regard to the Aral Sea Basin to address water use and regulation.

The countries are ravaged by drastic economic problems and the primary interest is to increase agricultural production, which it is hoped will have a major impact on the economy. As such, each one of the former republics is striving to expand irrigation, mainly to meet food needs for a growing population. This has necessitated heavy application of insecticides, pesticides, herbicides, and defoliants on irrigated farmland. It is estimated that about 150 000 tons of toxic chemicals are used annually, making it hard to find clean drinking water because of the contamination by agricultural chemicals (Ataniyazova 2003). Ecological consequences include the collapse and disappearance of the fishing industry as a source of food and income. In addition, the soil is so salty that crops cannot easily be grown, and agricultural output has declined by 30–50%. An estimated 40 000 km² of the heavily saline sea bed is now exposed and is blown out by frequent winds, resulting in the air becoming salty and dusty, leading to respiratory illnesses (Micklin and Aladin 2008). Also, the local climate is becoming hotter and drier as the effect of the Aral Sea is reduced. Health has also been affected in numerous ways with diseases increasing, particularly rates of anemia, tuberculosis, kidney and liver diseases, respiratory infections, allergies, and cancer (Ataniyazova 2003). Rates for these diseases in this region far exceed those of the former USSR and present-day Russia. Average life expectancy in the Kzyl-Orda region of Kazakhstan has declined from 64 to 51 years and one in every 20 babies is born with a birth defect, a statistic five times higher than European countries.
DISEASE AND ILLNESS, POLITICAL ECOLOGY OF

In short, the Aral Sea perfectly illustrates the use of political ecology to explain ecological damage and disease proliferation. Historically, the former Soviet Union put in motion a strategy of agricultural development that relied on the waters of the two major rivers to increase cotton production. Once the former Soviet Union broke up, the new independent states began to pursue their own resource use interests, which often were in conflict with those of the other states in the region. This has resulted in ecological consequences that have had a major impact on health issues in the Aral Sea Basin. Thus, the political ecology framework is useful for explaining the fate of the Aral Sea with reference to issues of scale (decisions taken by the former Soviet Union at higher levels with disastrous effects at the local scale) and the historical and economic contexts.

General discussion of the application and usefulness of the political ecology framework

As these two examples illustrate, understanding newly emerging and re-emerging diseases demands a political ecology approach. The traditional human ecology model of disease can tell us what conditions a disease requires to develop, but it cannot explain why conditions have changed. To draw out causative factors the political, historical, and economic story must be told. We can add little to the vast literature on the human ecology of tuberculosis; however, the reappearance of the disease in the Ukraine after 50 years requires researchers to address the effects of Soviet rule, man-made disasters, structural adjustment policy, cultural tension between east and west, inadequate private health care, and rising rates of HIV/AIDS due to neglected family planning programs and sex trafficking. Thus, from a political ecology perspective one of the oldest known human diseases still has a rich and exciting story to tell.

Emerging diseases are diseases that were previously unknown, either globally or locally (Wang et al. 2013). These may be diseases that appear in a new geographic area, that have adapted to new hosts, or which have never been seen before. The most obvious example is the human immunodeficiency virus (HIV). Virtually unknown before the late 1970s, HIV is now the sixth leading cause of death worldwide (World Health Organization 2013). On the other hand, re-emerging diseases are those diseases that were once eradicated or controlled but have since reappeared (Wang et al. 2013). Re-emerging diseases usually come about when the environment conducive to disease spread is recreated, either biologically, physically, socially, or behaviorally. Many of these realignments come from economic and political forces beyond the individual’s control, but some develop at the local level. Wherever a re-emerging disease is present a political ecologist is sure to find something fascinating to explore. In the modern world most of our changing disease patterns are anthropogenic; humans learn to control diseases but then the diseases bite back, exploiting many of the projects intended to make human life better as the case of dam building in Senegal testifies.

One way humans create and enable disease spread is through physical changes to the environment, such as dam construction, deforestation, agriculture, and wetland drainage. These environmental changes are usually a direct result of ongoing government policy or neglect. Furthermore, economic challenges resulting from government economic policy and global and regional economic fortunes fuel the environmental changes, which ultimately result in the proliferation of certain diseases. Global climate change is sure to contribute to disease spread by making previously uninhabitable areas attractive
to disease vectors such as mosquitoes. For example, Kyasanur forest disease is a tick-borne hemorrhagic fever that was endemic only to 300 square miles (777 km²) of forest in South India in 1957. However, in the 1980s a major epidemic occurred in villages that had recently been cleared within previously undisturbed forests. Currently about 100 people are infected a year and the number continues to increase. Similar origin stories exist for Leishmaniasis in Latin American rainforests, and perhaps even Ebola outbreaks in Uganda and the Democratic Republic of the Congo (Yale, Bhanurekha, and Ganesan 2013). Of course, people are moving deeper into the forest because large-scale agriculture and limited land require people on the fringes of society to venture farther into the forests to eke out a living. In the Indian case most of the land had been cleared to produce cashew, a cash crop. Cash crop farming has been particularly detrimental in the developing world because it breeds reliance on outside goods, services, and funds. In most cases, it is the poorest people who have the fewest resources who endure the first waves of emerging and re-emerging diseases caused by the changing physical environment.

One more example of how political ecology contributes to our understanding of emerging diseases is through the increasing mobility and interconnectedness of modern life. Some of these changes are purely recreational. When an American citizen with multidrug resistant tuberculosis wished to attend a wedding in Europe he evaded the CDC quarantine directives, possibly spreading an uncontrollable and highly infectious disease to people in the airports, planes, and venues he attended. Many more population migrations are economically and politically forced, however. Refugee camps have notoriously poor health outcomes and can lead to re-emerging diseases like cholera from poor infrastructure, but also to emerging psychological illness such as shared night terrors caused by fear of ethnic cleansing (Kalipeni and Oppong 1998). Migration into large cities creates slums, where every disease caused by unsafe water, poor ventilation, malnutrition, and overcrowding will flourish if given the chance. These diseases do not stay in the slums and can create epidemics throughout the region. This migration is due to the same forces that push poor people deeper into the forest primarily due to economic desperation.

Likewise, the spread of HIV in Africa is largely driven by the massive numbers of migrant laborers who must trek to cities and mines to make a living, leaving their families at home for many months at a time. Once infected, many do not have access to anti-retroviral treatment and the pandemic continues to spread unabated. Indeed, the historical and geographical context of Africa is critical to the understanding of the economic conditions, migration, and spread of the HIV/AIDS epidemic. Colonial powers were attracted to southern Africa because of the presence of rich soils and minerals, which included gold and diamond. As the European powers carved out colonies and began to settle they also created pockets of economic development that, in turn, triggered migration of labor from economically depressed areas such as rural parts as well as land-locked countries such as Malawi to resource-rich areas such as South Africa. This, in turn, contributed towards the spread of diseases, including sexually transmitted diseases (Sawers and Stillwaggon 2010). It is pertinent here to reiterate the distal and proximate causes of epidemics and diseases. Globalization has so shrunk the world that a disease can be transported from a distant part of the world to some far-away place in a matter of days, for example, from Africa to the United States. Proximate causes of the disease like Ebola might include people eating bush meat such as infected bats and contracting Ebola by doing so. A distal cause might be an
infected person in Liberia traveling by air to Nigeria bringing with them the infection and spreading it among the populace in Nigeria.

Take the case of Malawi, for example, a country that is no exception to the severity in HIV prevalence although recent estimates indicate a declining trend. The first case of AIDS was diagnosed in 1985 and in 1998 the country reported a prevalence rate of 26%. USAID estimates that at the end of 2007 there were 930,000 adults and children in Malawi living with the HIV. It is currently estimated that about 10% of the adult population aged 15–49 in Malawi is infected with the virus (World Health Organization 2013). Among the population group affected by the AIDS crisis, young women seem to be disproportionately affected. Nearly 50% of all new infections occur among women between the ages of 15 and 24 years. This age group also reports a prevalence rate of 9%, which is substantially higher than the 2% reported for males within the same age group.

In Malawi, migration from rural areas to urban areas and international migration to other surrounding countries, such as South Africa, Zimbabwe, and Zambia, facilitates HIV transmission. Migration is in response to lack of sustainable economic conditions in rural areas. The migrant population includes truck drivers, female sex workers, traders, fishermen, mine workers, military personnel on peacekeeping missions, refugees, and seasonal workers. These population groups tend to engage in high-risk behaviors, which increase their vulnerability to infection. In general, many of these migrants have limited education and training, and thus have few marketable skills; as a result, they represent the country’s marginalized population who are subject to discrimination.

It is also important to note that an individual’s vulnerability not only depends on the economic conditions but also social position. Women in Malawi tend to have limited entitlements, generally earn low wages, and do not benefit from international aid programs. Furthermore, women in general have less access to education beyond the primary level, which restricts them in their choices of occupation, formal sector employment, and sustainable wages. According to the political ecology approach, the history of labor migration and the inequality among men and women influence the power structure and vulnerability to HIV and other equally important diseases in Africa and elsewhere.

SEE ALSO: Biosecurity; Climate change and health; Environment and disease; Health geography; Health and wellbeing; Political ecology

References


Disease diffusion

Daniel J. Exeter  
The University of Auckland, New Zealand

Clive E. Sabel  
University of Bristol, UK

Spatial diffusion theory seeks to explain the movements through time and across space of goods, people, innovations, or ideas. A diffusion process involves numerous events on all spatial levels from intercontinental transport and climatic variations to the micro level of homes, workplaces, schools, and hospitals. In geographical terms, a diffusion process involves a place of origin (a source) and one or more receiving locations, pathways, and modes of diffusion, as well as a force driving the diffusion.

Applied to disease, diffusion theory can help describe how, where, and when the spread of infectious disease occurs. Studies of the diffusion of infectious disease draw heavily on the work of health geographers and spatial epidemiologists who are interested in how a disease spreads and where cases are likely to occur. The potential for spatial diffusion depends on the character and ecology of the causative agents, such as their transmission modes, mobility, reach, and viability, as well as various determinants. Exposure (i.e., the contact between causes and susceptible individuals) is crucial for transmission and the start of a pathological change in a human or other living being.

Disease diffusion depends on exposure – the innumerable microscale events that bring individuals in contact with health hazards. Occupational conditions, daily personal routines, and attitudes may be decisive for exposure. Social norms and cultural practices play an important role for either spreading or barring infections. For example, promiscuity or fidelity, polygamy, circumcision, and other practices can influence the spread of sexually transmitted diseases (STDs).

Like other geographical phenomena, sources of disease diffusion may assume point, line, or area shapes. Infected wells or minor ponds, perhaps food stores, or even kitchens could be considered point sources. Line-shaped potential sources could be rivers or riverbanks where mosquitoes breed. Area sources of disease diffusion may be endemic regions and natural reservoirs in lakes, swamps, forests, or paddy fields.

Early models of disease related to the impacts such ailments had on the demographic structure of society, or dealt with the rates of disease transmission. These models included the classic three-compartment Susceptibles-Infectives-Recovereds (SIR) model and the use of neural network models that “learn” the underlying processes of the phenomena. These temporal models typically resulted in logistic growth curves, showing the accumulation of cases (events) over time, or normal bell curves for frequency of disease over time.

The examination of phenomena moving through space over a period of time was conceptualized by Hägerstrand (1968), whose work focused on the acceptance of new innovations...
DISEASE DIFFUSION

within a community. He contended that different spatial patterns arise due to the different methods of transportation/communication that carry information regarding the innovation and the barriers that restrict the flow of such information. When a new innovation is publicized, early adopters will begin using the technology. They share their experience with their friends and colleagues, and with continued publicity the number of people adopting the new innovation increases. As this process continues over time, information about the innovation travels to new communities – through different sectors of society, between cities and countries – until the innovation becomes “mainstream.” The speed at which the innovation spreads is dependent on the way information travels from person to person (and indeed place to place). Additionally, other factors such as the cost of purchasing the innovation, which is also governed by supply and demand, are also influential. Consider the spatial diffusion of condom usage and compare its evolution to the uptake of tobacco cigarette smoking. They each experienced spatial diffusion, although the communication networks, barriers, and associated costs were very different!

To health geographers a disease or illness (e.g., measles) is typically defined as the innovation and the time an individual is exposed to the illness represents the acceptance component of Hägerstrand’s model. Contagious diffusion processes are driven by proximity, and typically spread out from the source of infection to adjacent locations. As the epidemic progresses over time, contagious diffusion could be thought of as a band of concentric circles spreading out from the core. By contrast, hierarchical diffusion epidemics often “bounce” from the origin to (often) nonadjacent locations based on spatial interactions of importance. For example, a disease may spread from one city to a very distant city as a person with measles travels from London to New York. Although the distinction between contagious and hierarchical diffusion is made for modeling purposes, in reality, diseases are spread through a combination of these processes, especially as populations become increasingly mobile. In addition, epidemics of infectious diseases often occur in cycles in which the incidence of disease increases and wanes over a period of time. Although these “waves” tend to occur with relative frequency, different waves may behave differently or display both “contagious” and hierarchical phases simultaneously or sequentially as parts of a larger diffusion process.

In order to investigate the diffusion of measles and influenza, Iceland has been considered as a natural laboratory in the sense of being a large, effectively “closed” island population (Cliff et al. 1981). Epidemic models have been linked with spatial theory (notably spatial autocorrelation) to help better describe the flow or movement of contagious disease through time and space. Patterns of both contagious and hierarchical diffusion are evident in the modeling of Iceland’s measles epidemic of 1950–1952 and movement of disease from the capital, Reykjavik, to outlying districts around the island along transportation networks.

Understanding the geographical diffusion of diseases assists long-term prediction and containment of diseases. Adequate prediction and surveillance is necessary for public health intervention in emergencies. Vigilance and early detection of disease epidemics are important for national and global security in cases of serious infectious epidemics, regardless of whether they follow natural “waves” (e.g., measles) or hypothetically result from deliberate terrorist-induced attacks. In the past 20 years, there has been an alarming re-emergence of infectious and vector-borne diseases, such as tuberculosis (TB) and malaria, which are returning to parts of the world in which these diseases have been absent
or confined for decades. Furthermore, there is concern that some diseases such as TB have mutated and become resistant to existing medical treatments. In addition, emerging diseases such as severe acute respiratory syndrome (SARS) and avian influenza have prompted many national and international initiatives to improve surveillance and control of communicable diseases.

The influenza virus regularly appears in new strains, preventing lifelong immunity in the population, while the measles virus remains stable. Current health threats are mutations giving rise to new varieties and drug-resistant forms of causal organisms, for example, methicillin-resistant *Staphylococcus aureus* (MRSA) and TB.

Mapping disease epidemics provides a visual summary of their progression, but often these maps are not prepared in a timely manner, or estimates of the disease distribution are crude and inaccurate; however, tools have been developed for monitoring, surveillance, simulation, and prediction, based on better understanding of the spatial and temporal behavior of diseases. Geographical information systems (GIS) can be usefully applied to aid disease surveillance, simulation, and analysis, provided that cases and risk factors can be mapped with high spatial and temporal accuracy. Remote sensing can be used for surveillance of ecological conditions and habitats for disease agents and vectors. GIS-supported simulation has been used to forecast the diffusion of HIV/AIDS (human immunodeficiency virus/acquired immune deficiency syndrome) and legionnaire’s disease. With sufficiently accurate data, simulation can be applied on any level down to blocks and workplaces. Cluster detection software such as SaTScan™ has been extensively developed in order to be integrated with spatially referenced infectious disease notification registers and other health databases to either prospectively or retrospectively identify clusters of disease efficiently and rapidly.

There will no doubt be increased use and sophistication of spatial technologies such as GIS, remote sensing, and Global Positioning Systems (GPS), and the integration of such technologies with disease infection models in the future. Obtaining the detailed information regarding where individuals move, who they interact with, and for how long, on a daily basis required to better understand the spatial diffusion of disease is extremely complex and difficult. However, it is now possible to both model and use GPS sensor technology to understand social processes and diffusion of disease progression.

Place effects on health emerge from complex interdependent processes in which individuals interact with each other and their environment, and in which both individuals and environments adapt and change over time. Traditional epidemiologic study designs and statistical regression approaches are unable to examine these dynamic and interactive processes. Agent-based models (ABMs) and other system-dynamics models may help to address some of these challenges, including within the infectious disease epidemiology field (Yang and Atkinson 2008). ABMs are microsimulation models where individuals in a population are called “agents.” ABMs allow researchers to change the characteristics of agents, such as their age, gender, ethnicity, and socioeconomic position, or whether they are carriers of a disease. By running the microsimulation and allowing the agents to interact with other agents in the population, models of diffusion or indeed the impact of immunization programs may be developed to inform policy.

In summary, this entry has described the processes used by health geographers and spatial epidemiologists to model the spread of disease across space and through time, with a particular focus on infectious diseases. Disease diffusion modeling has been influenced by Hägerstrand’s (1968) time-geographic work, with most
DISEASE DIFFUSION

models representing hierarchical and/or contagious diffusion processes. Understanding these processes is invaluable in disease surveillance and in the detection of infectious disease clusters. While attention has historically focused on the eradication of infectious diseases associated with the poverty and squalor of the nineteenth century such as cholera and plague, there has been a recent emergence of new diseases and significant developments in the way people travel across large distances, such as by high-speed train and air-travel, increasing population mobility. Together, such developments will challenge our ability to model diffusion processes and to respond in a timely and appropriate manner.

We have focused on biomedical approaches to modeling diffusion in this entry; however, it is important to stress that applying our knowledge of spatial processes such as human interaction and mobility is vital to increasing our understanding of the diffusion process. Consider, for example, the rise in measles cases, and deaths, in the United Kingdom in 2008 due to insufficient population-wide vaccinations rates, to understand this. While measles is well understood medically, clearly there are social factors contributing to its diffusion which can result in diseases remaining endemic or even re-emerging in the future as epidemics or global pandemics.

SEE ALSO: Exploratory spatial data analysis; Geographic information science; Microsimulation; Spatial analysis

References


Further reading


Dispersal, diffusion, and migration

D.F. Greene
Concordia University, Canada

M.J. McCavour
University of Quebec at Montreal, Canada
Concordia University, Canada

Until a few decades ago, the movement of animals or plant propagules, especially long distance dispersal (LDD), remained a sideshow in ecology, of little interest except to a few specialists. Although the process of distant dispersal was clearly important in, for example, the new biogeography pioneered by MacArthur and Wilson in 1969, as well as in our understanding of the history of taxa at large temporal and spatial scales in a world of moving plates, it nonetheless garnered little attention from geographers and ecologists. Recently, however, LDD has earned considerable interest because of two applied issues (Nathan 2006). The first is related to fast climate warming and poses the question of whether species can migrate sufficiently rapidly to keep pace with the projected velocity of climate change induced by greenhouse gases. The second issue concerns invasive species: are there traits that allow us to distinguish potentially noxious invasives from those that would either be very unlikely to establish in a new area or else will, if established, have small populations with little effect on the native biota? If so, we could make our border control protocols far more efficient.

Migration has been defined in two ways. First, many animal species migrate diurnally (e.g., most plankton species have predictable vertical movement within the water column) or annually (famously, many bird species) in response to changes in environmental stimuli. This meaning will not be dealt with here. Instead we focus on migration as a dramatic alteration of range limits at a time scale of thousands of years, a spatial shift imposed by sustained, directional climate change. The primary driver typically is change in temperature rather than in precipitation. In particular, the migrations that have been studied most intensively are the poleward range shifts of the Holocene at higher latitudes and elevations, with most of the research in mountains also occurring well away from the equator. The primary reason for this focus on the higher latitudes is merely that the main empirical tool available for studying population dynamics in the distant past – ancient pollen grains in lake sediments or deep mosses – is useful only for wind-pollinated plant species, and aside from grasses and sedges these are quite rare near the equator.

Climate-induced migration has been studied for a century (Faegri and Iversen 1989). After early geomorphologists such as Agassiz in the
mid-nineteenth century made clear that much of the landscape of Europe and North America made sense only with an assumption of recent, widespread glaciation. Scandinavian investigators argued that the pollen grains in bogs showed predictable changes vertically, with the deeper layers of course indicating the species composition at earlier times, and further, that the plant species presently well to the north of a site had shown up lower within the core. For example, willow (**Salix**; an important constituent of tundra vegetation) pollen being replaced locally by aspen (**Populus**) and pine (**Pinus**) pollen was clearly an indicator of sustained warming. At first, there was little interest in the species themselves; they were proxies for measuring the kind of radical climate change that had caused the waning of Agassiz’ glaciers. Until radiocarbon dating became common in palynological work in the late twentieth century, no dates and therefore no velocities could be assigned to temporal changes at a site or to spatial changes among a network of sites.

By the 1990s there was a sufficiently dense constellation of palynological sites with dated cores in the higher latitudes that one could chart the migrational velocity of the northern or southern range limit for wind-pollinated species. Generally, beginning about 18 kya, species in the northern hemisphere moved rapidly in a northward direction until approximately 3–9 kya at which point the great majority of plant species had approximately reached their present northern range limits (Delcourt and Delcourt 1991).

Within the last two decades four generalizations have emerged from this dense literature. The first was that when a species arrived at a site, the pollen abundance increased in a logistic manner, with doubling times on the order of 50–200 years, values broadly similar to those seen in studies of contemporary invasive populations undergoing fast increases in abundance. Presumably, this represents the competitive advantage enjoyed by the invader as it keeps pace with its realized climate niche while the other species who have been there longer are no longer well-adapted to the recently emerging combination of precipitation and temperature. Subsequently, at this site the same species declines in abundance as its climatic niche moves further poleward and thus it becomes less competitive.

The second conclusion, and far more surprising, was that species moved independently of one another; species that presently have more or less coincident ranges typically began the migration 18 kya with far less overlap in ranges, and moved at different rates and along different axes. In short, each assemblage of plant species at a place was ephemeral, and seldom corresponded to our expectations of the floristic composition of “tundra” or “boreal forest” or “deciduous hardwood forest.” This finding effectively ended the still-popular assertion of F.E. Clements from a century previously that organisms within a climax association were so finely co-evolved that they must have remained together as a discernable unit for millions of years.

The third finding – quite unexpected – was that migrational velocities seemed to be far too fast – on the order of 50–200 m per year. Given that for example trees with ample light do not begin to produce large numbers of seeds until they are several decades old, these seeming rates mean that we are actually imagining that each generation of trees has long distance dispersal (LDD) events on the order of a few kilometers. Given that most contemporary studies of seed dispersal have been limited to a few hundred meters from a focal plant, and found vanishingly small numbers of seeds at even this modest distance, it seemed extraordinary that tree species could routinely manage this magnitude of LDD over the course of a few decades. This issue
remains unresolved – because of the logistical problem of finding needles in distant haystacks – and ultimately has been dealt with only via models. Encouragingly, it is becoming clearer from these models that dispersal of small seeds on the exterior of animals and of larger seeds within the intestinal tract of animals could easily permit a dispersal distance of several kilometers for a non-negligible fraction of a seed crop. We note that some authors, mainly in Europe, have argued that refugia at hypothetical ice-free alpine sites would have greatly aided poleward migration, and thus the required LDD events need not be as high as a few kilometers per generation to explain observed migrational velocities; this contention, however, has not led to any consensus on the existence or importance of the putative pockets of survivors among the glaciated landscape.

The final result was also a surprise: all species, regardless of whether they were dispersed by animals or wind, seemed to migrate at roughly similar rates. How could this be? The obvious explanation was that the limiting factor was not LDD but rather the rate of climate change. That is, the recent Holocene example was as if a set of runners of varying abilities were all simultaneously pushing up against the same slowly yielding barrier. Testing this hypothesis has proved difficult given that invariably the plant data are used to reconstruct the rate of change of climate. In any case, there is now a burgeoning literature based on the speculation that species may not be able to track CO₂-induced warming because (i) the required migrational velocity will be far more rapid than what was experienced in the early Holocene, and (ii) species will move more slowly than previously because they must migrate through intensely fragmented terrain. In turn, this notion of a potential migrational deficit has led to arguments for “assisted migration” – that is, moving propagules from likely genotypes well-north of their present range limits because their dispersal capacity is presumed to be incommensurate with the task.

The capacity of the more mobile animals (large mammals; birds) to track migrating plants is not doubted. But for smaller animals such as herbivorous insects it is not clear if they invariably lag near or very far behind the plants. A crucial question is the ability of pollinators and frugivores to keep pace with the plant species upon which they are dependent. There is a potentially important link: the migrational velocity of the dependent plant may be limited by the velocity of the fastest symbiont for the guild of pollinators or fruit dispersers upon which it is dependent. Unfortunately, there is as yet no dense spatial network of insect macrofossils using, for example, moss cores, and thus the question must remain unanswered at present. Nonetheless, there are studies at single locations using insect head capsules and in some cases distinctive feces to examine temporal changes in abundance. An eventual network of similar sites will, of course, as with the analogous data for pollen, be limited to those species that are most commonly represented in the moss cores in a region, and unfortunately these species may seldom be important mutualists but rather high-abundance herbivores such as pine beetles or spruce budworm.

We limit the meaning of diffusion in this entry to the spread of a species across new areas, with the range change induced by a driver other than climate. Thus, here the focus is on invasions promoted by the paucity of parasites and other enemies in the new area; the rate of climate change imposes no limit on the rate of spread. Borrowing from chemistry the argument for a random walk of particles, J.G. Skellam (1951) and others had by the mid-point of the twentieth century developed simple models of species’ spread across landscapes. In particular, such models tended to focus on the rate of spread at the range margin, and, while not explicit, it
was understood that the spread rate was greatly controlled by the capacity for LDD. While the very simple diffusion model of Skellam, for example, featured local exponential population growth and predicted a constant rate of range expansion, more recent models have included selection for enhanced dispersal capacity among range edge populations. There is no shortage of good empirical data sets of species spread rates from the last 100 years available for model testing, especially for animals (e.g., muskrats in Europe or starlings in North America).

The Convention on Biological Diversity defines invasive species as those spreading outside of their normal range and causing substantial harm. Globally, the damage caused by invasive species has been estimated as £1 trillion per year – close to 5% of global Gross Domestic Product (GDP). In developing countries, where agriculture accounts for a higher proportion of GDP than elsewhere, the negative impact of invasive species on food security, as well as on economic performance, can be even greater. Indeed, invasive species are typically viewed as primarily an economic issue rather than a biological one (despite the frequently dire effects of invasives on native species, especially on small islands).

Much work has been invested in enumerating the traits that predict invasion probability. For both plants and animals an essential characteristic of successful invaders is, relative to other members of the taxon, rapid individual growth rate. An equal role for dispersal capacity is less clear. Where dispersal traits can be at least modestly quantified with commonly available metrics (e.g., for wind or water dispersal by plant seeds, the ratio of propagule mass to propagule area is key), better dispersed species are indeed more likely to be successful invaders (e.g., Moravcova et al. 2010). But for many other kinds of plants, for example, those dispersed inside the intestinal tracts of animals, there is no such simple measure. Thus, while the more complex diffusion models make clear that LDD is a crucial determinant of invasion success and subsequent spread rate, nonetheless the empirical evidence remains largely unhelpful.

Uninformed by any but the simplest kinds of biological reasoning about dispersal, governments have developed policies for reducing invasions, and here they focus upon humans as the dispersal vector. For example, one cannot bring firewood into the United States as it might contain saprophagous insects such as pine beetles. Additionally, voluntary certification protocols have begun to play a role in reducing invasions. For example, there is now a certification program in North America governing the sale of “clean” nursery stock for gardens. A certification approach is of course highly useful as it capitalizes on the interest and education of the potential dispersal agent.

**SEE ALSO:** Biodiversity; Island biogeography; Paleoeconomy; Zoogeography

**References**


Further reading


The term distance decay has been used to describe the effect of distance on interactions between two separate locations. Although the term might also be used in ecology or habitat of plant and animal species, the description here is confined to human activities in the disciplines of urban, economic, and cultural geography, that is, a concept linking geography and the occurrence or frequency of a pattern of activity. Distance decay is an important precept of spatial analysis, especially for spatial interaction models and notions of cultural diffusion; it is inherent in various applications of gravity models (Hanks 2011).

As a term that describes the decline of influence as distance increases, the term is related to Waldo R. Tobler’s first law of geography that “All things are related but near things are more related than far things.” Two terms need to be qualified: distance and decay. Traditionally, distance means the physical map and, more realistically, topographic surface distance. Assuming a flat surface where accessibility from a center point to all directions is equal, lines of equidistance will display a circular pattern around that point. This was the key assumption in the von Thünen rural land-use model and in Christaller’s central place theory at the time when the road network was not yet well developed and animal transportation was still dominant. The term decay refers to the diminishing influence of a phenomenon, attribute, or activity when two locales are far away from each other. For example, there is less interaction between distant places because accessibility and land value tend to decrease from a center.

Distance decay is commonly represented graphically by a curving line that swoops concavely downward as distance along the x-axis increases. It can be mathematically represented by the expression $I = 1/d^2$, where $I$ is a measure of spatial interaction and $d$ is distance (The Stands4 Network 2013). Borrowing the idea of Isaac Newton’s equation defining gravitational force, distance decay bears a strong relationship to the gravity model in that a precise mathematical relationship exists between the strength of a spatially expressed phenomenon and the distance over which it is distributed (Hanks 2011). For instance, a superstore tends to have a lower distance-decay function than a neighborhood convenience store (Figure 1). Geographers using gravity theory in their work attempt to construct a similar model analyzing labor or trade flows; transportation; shifts in population; and diffusion of languages, religions, and technologies.

The bid–rent theory is an important element in the gravity model, in which the concept of accessibility plays a decisive role. The rationale is that due to land-use competition in a free market, the land will be allocated to the highest bidder that is able to derive the greatest utility from the most accessible land. The saving in transportation cost can be traded off for extra rent payment. As the distance from the city center increases, the accessibility decreases, resulting in the decrease in derivable utility and rent. Hence, distance and land rent bear a negative
relationship termed as distance decay. Different functions have different bid–rent curves due to differences in height (a measure of the utility) and gradient (an indicator of sensitivity to accessibility). Commercial land use, especially for high-order retail business and service, earns the highest bid, as it is supported by a large market threshold with longer range. Industrial land use has smaller utility, as the return per unit area is low due to its large space requirement. Residential land use has the least utility, so its bid–rent is the lowest. The distance-decay curve, therefore, turns out to be steepest for commercial use and the gentlest for residential use. In urban geography, a city or town center is usually explicitly defined, from which all attributes, such as population density, land value, and quality and quantity of services, decline as the distance from it increases.

Although the distance between two places need not necessarily include such a center, it bears the same underlying assumption that declining accessibility will lead to reducing interactions for places farther apart. In many cities, the highest land values occur in the city center, but prices drop off rapidly with increasing distance. Local increases to the prevailing trend occur at outlying subcenters, forming secondary peaks, which are probably also major freeway and arterial highway locations, forming ridges of higher land value (Figure 2). Values at suburban downtown locations again reach the peaks found in the downtown areas, but drop off rapidly at short distances.

It has always been controversial whether the concept is universally applicable and the relationship is quantified correctly. With the advent of transportation and communication technologies, the frictional effect of distance has largely been reduced. Personal or social group interactions need not be actually meeting one another at the same place and at the same time. Besides, speculation for high-class residential blocks in space-limited cities has drastically raised its bid–rent ability. It may be more accurate to conclude that physical or topographic distance is no longer a prime determinant of the degree of interactivity between places and people. Yet, they still dictate a spatial variation of interactivity between places, whether at the local, regional, national, or global level.
A new perspective of distance decay

Interactions between places and people are still very much dependent on available connectivity and the efficiency in transporting people and goods or transmitting information. It can be a passenger-traveling network for land, sea, and air transport, a walking path network for short-range travelers, or even a wireless network for mobile communication. As such, the term “distance” needs to be defined with a new perspective. Empirical studies have shown that the following have become key variables instead of mere physical distance.

1. **Time distance** – with a road network for driving by car or walking as a pedestrian, the time taken to travel between two places normally increases in proportion to distance; but additional factors, such as waiting at traffic lights, the number of stops, number of lanes, density of cars or commuters, and speed restrictions, will all affect the travel time. This also varies temporally for day and night, normal and rush hours, weekday and weekend, low season and peak season. Time is an important determinant for everyday life activities, public administration, and commercial decisions. A 5–10 minute longer travel may stop people from buying daily necessities from shops that are further away. It is uncommon to spend over three hours’ journeying to enjoy just an hour’s hiking in the countryside in a day; postal delivery is less frequent the greater the distance from the post office; shopping malls and business activities will orientate themselves toward the biggest potential markets.

2. **Cost distance** – this includes fuel cost, toll fee, or even administration cost at custom borders. Despite the general truth that it is more costly to travel a greater distance, in most cases cost in monetary terms is not proportional to distance. In particular, for public transport, fare is nonadditive. Zonal fares, section fares, concessionary fares at selected stops can apply in many different ways for different modes of transport in different cities. These may alter the pattern, in that people may be willing to travel a little further distance for greater benefits or leisure. In Hong Kong, ever since the government introduced a subsidy for senior citizens to travel anywhere for just HK$2 (equivalent to one US quarter), mobility for this age group has increased significantly. Such spatial expansion is especially obvious for people who rate fare more importantly than time.

3. **Convenience in terms of alternatives** – this refers to reaching a place directly or having to make one or more transfers. The proximity to car parks, terminals, stops, or transport nodes can also affect the desire to travel.

4. **Behavioral preference** – despite shorter time and lower cost, some prefer to travel on certain modes of transport or to certain destinations for leisure or business activities due to other nonquantifiable reasons, such as security, fame, or personal preference.

**Rate of decay**

Traditionally, the rate of decay – or distance lapse rate in some literature – was plotted inversely with one variable distance, with the gradient dependent on the phenomena and studied area being considered. From a strictly normative path set by the original retail gravity models in the 1930s, consumer behavior was considered predominantly uniform and deterministic until the 1960s. For instance, land value has a steeper downward sloping curve for cities with less developed transportation or more rugged
DISTANCE DECAY

terrain. Today, distance decay may still hold for long distances. With the advent of transportation and communication technologies, the increasing diversity of means of transportation, the introduction of various fare structures, and the increasing demand for better living conditions, the effect of distance on human interactions has become less and, thus, the curve deviates significantly from the traditional trend. Models explaining human activities between places should take into account not only typical normative criteria but also other environmental and people-based factors. One of the major problems associated with traditional gravity models is lack of parameters on intuitive understanding of interaction behavior.

Many studies have, therefore, worked on computing mobility patterns to see the extent of distance or other factors in governing human behavior. In particular, with the advent of geoinformation technology, mass data can be collected and analyzed. To compute or predict the number of interactions between two spatially separated places, one way is to evaluate their accessibility with various types of activity, facility, or service. The impact of distance on a particular activity may be linear or nonlinear and can be modeled with a distance-decay function. Computation accuracy is enabled with network analytical functions in a number of commercial geographical information systems (GIS). The accessibility of each origin or center is computed with regard to a number of destinations or service points, for example, school from residential blocks, supermarkets from housewives’ residences, resorts from potential sources of visitors. On the side of origins, both distance (including also traveling parameters such as time and cost) and attractiveness can be taken into consideration. Together with the demographic data of the service points, potential trips or interactions are predicted.

When evaluated by empirical data, the effect of distance decay can be assessed. However, to accurately model the degree of interactions between places, it is necessary to formulate very clear objectives. The type of interaction, for example, working, schooling, or shopping, and whether it is a single or multipurpose trip has to be identified first. Much essential data have to be collected including:

1. places of origins and destinations, which can be scaled from a building to a city level or even national level;
2. demographic composition and characteristics (e.g., certain age groups, income level or socioeconomic status) at origins that are essential to the type of interactions;
3. services at various destinations, which can be a long and complex list, such as the type and order of services, cost, popularity, and so on, from which an index of attractiveness is computed;
4. individual trip data obtainable from the more traditional but reliable questionnaire survey and activity log sheet to the more advanced and automatic mobile tracking data and credit/debit card records; traffic flow data in between places might also reflect the degree of interactions made at a generalized level.

On the other hand, to find out if these activity patterns are related to distance or other variables, data on the journeys made are important. These include not only the more quantifiable and objective parameters of the above-mentioned physical distance, time distance, and cost distance, but also the more subjective behavioral issues like safety and familiarity, road quality, and scenic value. The latter are perhaps overriding but not accounted for very well in traditional path-finding models. The main
reasons are that these spatiotemporal variants are not easy to quantify and precompute and are perceived differently from individual to individual.

The collection, manipulation, and analysis of voluminous data have become feasible with advances in positioning technology, CCTV image capturing, and geoinformation technology. The current way forward is to develop appropriate techniques for big data mining and to generalize patterns for developing reliable models, especially in the area of human behavior, of the activity patterns.

SEE ALSO: Accessibility, in transportation planning; Central business district; Central place theory; Distance; Geographic information science; Geographic information system; Network analysis; Spatial interaction; Tobler’s first law of geography; Transport and development; Transport networks; Urban geography

Further reading


References


Distance

Anthony C. Gatrell
Lancaster University, UK

Distance is surely one of the defining variables in geographic inquiry. Indeed, 60 years ago the Scottish geographer Wreford Watson referred to geography as a “discipline in distance.” As location theory (whether concerned with agriculture, settlement, industry, or retail) took hold on the discipline in the 1960s (ushering in a preoccupation with the relative location of agricultural land use, central places, firms, or shops) distance as a variable assumed a crucial role. The relative location of consumers and producers mattered. And despite writers such as Cairncross (2001) asserting the “death of distance,” it still structures much behavior and interaction. People do still want to meet in geographic as opposed to virtual space; people do still want to shop in stores on the high street or in the shopping mall, despite the massive growth of online retailing; and many consumers care that their food is grown locally rather than shipped across the globe and thereby eating up many “food miles.”

We can conceive of distance, or spatial separation, in a number of different ways. For example: as a straight line or perhaps “great-circle” distance that we can measure once we know the locations of two places, given by their respective latitude and longitude; the route length that is measured along a set of roads, streets, or paths; the time it takes to get from one place to another; the monetary cost incurred in making a trip; and, perhaps, the sense we have of how far one place is from another. Such “distances” can be both precise and absolute (given identifiable and fixed or stable locations), but also sometimes vaguely specified and perhaps having a frequency distribution. A train journey from one station to another might be scheduled to take 81 min. if there are no stops, but if the train is delayed or requires a connecting journey it will take longer. The flight will cost more at particular times of the year, and depending on the carrier. If engrossed in a good book or movie the flight may seem a lot shorter than on other occasions, while if we ask a set of people to estimate the distance between, for example, Paris and Berlin we will get a variety of estimates. And while we may measure distance on an interval scale (kilometers, minutes, dollars) we often use nominal or ordinal scales (“it’s close by,” “it’ll take you quite a while to get there,” “it’s closer than you think,” and so on).

Distances are structured by economic and social factors, some of which are under individual control, some of which are not. Such factors change over time. While the great-circle distance between Paris and London remains fixed, journey times and costs do not. While the physical distance between Chicago and Los Angeles is the same for the wealthy and the disadvantaged traveler, the cost in monetary terms may be very different if one travels business class and the other relies on a cheap, last-minute deal. Conversely, the real cost to the traveler on a low budget might seem considerably more than that for the business traveler who can either easily afford the trip or has someone else covering that cost. Social factors matter too; if accompanied on a flight by an intimate partner the journey may appear much shorter than if traveling alone.
or with small children who require frequent attention and amusement.

Distances are invariably thought of as measures of the spatial separation of two discrete locations on the Earth’s surface. To measure, for instance, the distance from Los Angeles to Miami requires us to specify where each city is located; but since cities have no fixed “center” the distance is somewhat arbitrary. Further, while it is not unreasonable to ask “how far is it from California to Florida?” the answer will depend on what we take to be their respective “locations”: the geographic centers of each state, the state capitals, the largest cities, the east-most point on the boundary of California and west-most equivalent in Florida, and so on.

It should be clear from the above that distance can get quite complicated, conceptually and empirically. Distance is more than a simple variable that matters only to those pursuing a spatial analytic approach to geographic inquiry. Some of these conceptual and measurement issues are explored below before considering how distance has been used in some branches of geography.

**Formalizing distance**

In a very general sense, distance can be defined as a measure, often (though not necessarily) quantitative, of the space that separates two points in that space. It is a relation between those two points. This is a sufficiently general definition that it takes into account different types of spatial separation and also recognizes that while we may make precise measurements, we can also think of distance in qualitative terms.

More formally, consider two urban locations (i and j) that have known coordinates; denote these locations as \((x_{i1}, x_{i2})\) and \((x_{j1}, x_{j2})\) respectively. Simple geometry gives the distance between these points as:

\[
d_E(i,j) = \left[ (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 \right]^\frac{1}{2}
\]

This is the straight-line or Euclidean distance (hence the subscript “\(E\)” in the equation) between the two points, determined by Pythagoras’s familiar formula. Euclidean distance is useful in many contexts but it is of little use if measuring distances that have to take into account the curvature of the Earth’s surface. In this case, assuming the latitude and longitude of each point are given it is quite straightforward to calculate the length of the shorter arc on the great circle that joins these locations.

Euclidean distance forms what is known in mathematics as a metric. Formally, this means that \(d_E(i,j)\) has the following properties: it cannot be negative; it can only be zero if the locations are identical; it is symmetric (the distance from \(i\) to \(j\) is the same as that from \(j\) to \(i\)); and it obeys the triangle inequality (this means that the distance from \(i\) to \(j\) must be less than or equal to the sum of the distances from \(i\) and \(j\) to another location, \(k\)). Note further that the locations are assumed to be in a two-dimensional space; a third dimension will require another coordinate.

Euclidean distance is, however, not the only metric. Imagine a city whose streets are set out in a grid. Measuring the straight-line distance (distance “as the crow files”) from one location to another might help the crow but it will not help someone walk or drive from one location to another. But the equation above is readily modified:

\[
D_M(i,j) = [(x_{i1} - x_{j1}) + (x_{i2} - x_{j2})]
\]

and the subscript \(M\) is used to denote this as the “Manhattan” metric. This replaces the square of the coordinate differences, and the square root of their sum, by unity. More generally, an entire family of metrics (so-called Minkowski metrics) arises if these differences are raised to the power \(“p”\) and, after summation, raised to the power...
1/p. Some of these ideas were explored by Gatrell (1983), though Miller and Wentz (2003) provide a more recent perspective.

Proximity is the converse of distance. Locations are proximal if they are near to, or close to, each other. But proximity can also refer to areas or regions; for example, California is closer to Arizona than it is to Oklahoma. This proximity relation can be represented as a topological graph in which two states are connected directly if they are adjacent (contiguous). Such a simple representation of spatial separation (as adjacency) has often been exploited in spatial analysis, particularly in creating spatial weights matrices as used to construct measures of spatial autocorrelation.

While distance is considered above in terms of spatial separation it can also be conceived in terms of the difference or dissimilarity in the characteristics or attributes of particular locations or places. For example, if two countries share identical or similar languages, religious and ethnic composition, and so on, they can be thought of as “close” in terms of cultural “distance.” In this case we could construct a matrix of countries and attributes and assess the extent to which pairs of countries were similar or dissimilar across a range of attributes.

Relative distance

Distance, therefore, can be conceptualized in ways other than as physical distance. These include measures of cost, time, cognition, and social separation. None of these is fixed, so instead of speaking of the absolute distance between a pair of fixed locations we refer to relative distance instead (Abler, Adams, and Gould 1971). More recently, a relational, as opposed to a relative view of space, has emerged, asserting that objects can be understood only in relation to other objects. Yet while the relations matter, it is material things (people, stores, parks, hospitals, and so on) occupying specific locations that usually give rise to such relations.

Cost distance

Since distance is a constraint on some forms of activity (the so-called friction of distance) there is a cost incurred in overcoming such distance. Classical location theory (such as that of Von Thünen) posited an isotropic or uniform plane in which cost distance and physical distance were interchangeable. But since geographic space is not homogeneous or uniform in character the straight-line distance across such space ignores any underlying landscape heterogeneity or even – in flight – the variable pressures that may be encountered. As a result, geographers have sought to identify “least-cost” paths that seek out regions of low cost and avoid zones that “cost” more to traverse (Gatrell 1983, 52–53; Miller and Wentz 2003, 582). A simple example would be route construction, where constructing a road or rail track across marshland or steep terrain would be prohibitively expensive compared with drier, flatter land. To take a very different example, the perceived “cost,” to personal safety, of taking a direct route through a park at night might be very high and it may well be a safer option to take a longer, better-lit route home. Cost does not always have to be monetized.

Economists interested in international trade have shown that physical proximity matters: about one quarter of trade flows are between countries sharing a common border, and half is between countries separated by less than 3000 km. Giuliano, Spilimbergo, and Tonon (2013) look at the costs of freight transport among European countries using data from the International Monetary Fund’s Direction of Trade Statistics. They correlate these cost distances with a measure of genetic distance – the
DISTANCE

extent to which the populations of pairs of countries are genetically similar (Figure 1). More generally one can think of the economic distance – the disparity in wealth or economic size between countries or regions.

Time distance

Time distance can be considered as the average journey time from \( i \) to \( j \), or perhaps the fastest travel time between two places. As a measure, it does not meet the definition of a distance metric (see above) since it is unlikely to be a symmetric relation (the travel time from \( i \) to \( j \) may well be different from that of \( j \) to \( i \)). Travel time is shaped by material factors, such as mode of transport and level of technology. Further, any trip involves time spent in situ. For example, a short-haul flight takes perhaps an hour to get from one landing strip to another. But this bears little relation to the time taken to get from the origin (perhaps home) to the destination (a downtown hotel, perhaps), when the relatively unproductive waiting times (for passport control, security, luggage, taxis, and so on) add greatly to the real time-distance (and, of course, its perceived cost).

Journey times from one place to a set of other locations can be displayed graphically as a set of isochrones, lines joining points of equal travel time from a chosen location. In an urban context the spatial configuration of the street network generates isochrones that are quite convoluted, perhaps with inversions which suggest that locations that are geographically further apart are, in terms of time distance, somewhat closer. At a regional, or even international, scale there may be a broadly linear relation between travel time and geographic (or network) distance but there will be inversions here too, and no single time distance from one location to a set of others. A compelling historical example, due to the German geographer Eckert in the early twentieth century, showed time distance from Berlin to parts of Africa and the Middle East in 1900; while the travel time to the south coast of Africa might have taken 25 days by boat, it would have taken over 40 days to reach inland parts of North Africa (Figure 2).

Geographic space is therefore warped by travel time. But it is a one–many relation, showing time distance from one location to several. In order to represent travel time between many pairs of locations geographers have used multidimensional scaling (MDS) to create maps depicting the relative location of places in time space. More recently “spatialization” is a set of techniques where cartographers have played a prominent role. The underpinnings are essentially the same as MDS, although they seem to favor self-organizing maps (SOM) over MDS methods. MDS is often referred to as “ordination” in other disciplines (such as ecology).

Of considerable interest is that time (and cost) distances themselves change over time, as transport technologies and routes improve or are upgraded. Locations therefore “move” relative
Figure 2 Travel time from Berlin to Africa in 1900. Source: Gatrell (1983). Reproduced by permission of Oxford University Press.
to others. This process is called time–space convergence. Such convergence is not a simple linear, or even continuous, process since it is shaped by the introduction of new technologies that may result in a sudden and dramatic reduction in travel time. For example, the introduction of diesel trains, and then high-speed trains, dramatically reduced journey times in the United Kingdom during the latter half of the twentieth century, while transatlantic journey times by air have reduced sharply over the last 60 years, reflecting new engine technologies. But such convergence is itself geographically differentiated; while some places get brought closer together, others are bypassed, with potentially serious effects on their economic prospects.

Cognitive distance

Cognitive distance is the implicit or explicit judgment made about the spatial separation of locations that are not directly observed. For example, respondents might be asked to estimate distances between a set of landmarks in a city or the spatial separation of world cities. Judgments of distances between multiple pairs of places may, via MDS, be used to construct cognitive maps. As “behavioral geography” developed as a subdiscipline in the 1970s a substantial literature emerged – both in geography and psychology – that examined the relationship between cognitive and physical distance (invariably shown to be nonlinear) and what the correlates were of cognitive distance. Such correlates depend on the spatial scale of analysis but in a within-city context might include whether places are linked directly by major routes, or if there are spatial barriers (such as rivers). Other variables affecting distance estimation relate to the properties of the places themselves and how these are assessed. If one pair of places is “valued” more than another then distances are underestimated, and vice versa.

There is an inverse relation between “emotional involvement” and cognitive distance, neatly captured in German research which showed that distances between cities in the former East and West Germany were over-estimated by those opposed to German reunification – an effect poignantly labeled a “mental wall” (Carbon 2010).

Sociocultural distance

Intuitively we can think of the cultural and social distance between countries or social groups and individuals. Two countries made up of predominantly different religious or language groups are culturally and perhaps politically distant even though they might share a common border (India and Pakistan, or Israel and Jordan). Such differences, whether in religious beliefs, language, or ethnicity, can create distance between two countries. Trade among countries with common language is, other things being equal, three times the volume of those without, and countries linked as colonizer–colony (and therefore politically “proximate”) see trade increased nine-fold. And, of course, trade and tariff barriers create distance between potential partners.

The distance between social groups or individuals can be thought of in terms either of volumes of interaction or their social difference. At an individual level there are sets of interpersonal distances according to whether the interaction is intimate, personal (e.g., conversation between two colleagues), social (perhaps within a small group), or public (e.g., an address to a small audience). These sorts of interactions are mediated by culture, gender, physical setting, and other factors. More generally, an entire subdiscipline of social network analysis exists (and is expanding rapidly) that looks at the structure of interpersonal contact networks and the relations that define such networks (whether simple knowledge of each other, acquaintance, friendship,
and so on). Clearly, it is also of interest to know whether such networks are also structured by geographical proximity.

Despite the huge growth in social media, where it matters not a jot whether interaction is with a near neighbor or with someone the other side of the globe, distance does still structure human interaction. People often want to meet in real space, not only cyberspace. An important family/social occasion such as a wedding or a civil partnership, or even a funeral, usually expects or requires copresence. In both cases, distance still matters but is socially structured; an invitation to attend will depend in part on the social relationship to the couple or the deceased.

Despite the huge growth in online shopping (retailing) many consumers want to visit stores in person, while charges for local delivery to the front door will, in some cases, depend on distance from the store. In addition, social media themselves may call for spontaneous interaction in real space (flash mobs) while there are many examples of meetings that are assemblies of concerned citizens protesting about political issues. Think, for example, of the “Arab Spring” uprisings in countries such as Egypt, where crowds have gathered on many occasions (in Tahrir Square, Cairo, for example) to call for changes in government.

Examples

As noted in the introduction, distance was the crucial variable in classical location theory. For example, Von Thünen’s agricultural location theory expressed rent as a function of distance (cost of transport) to the market; Alonso’s urban land-use theory took the same stance; Weber’s industrial location theory relied on measuring distances between sites of extraction and of consumption, while central place theory (due to Christaller and Lösch) was concerned with the spacing of settlements at different levels of the urban hierarchy.

Distance also became the defining feature in classical studies of movement – whether of human beings, freight, or information. In the late nineteenth century, studies of migration by Ravenstein suggested that most migration (at least, in England and Wales, his area of study) comprised short distance moves, often between neighboring counties, and that longer-distance migration was to larger towns and cities. Such an acknowledgment of the role played by population and distance prefigures work 40-50 years later by researchers (such as William J. Reilly, John Q. Stewart, and William Warntz) into “social physics.” Here, the fundamental principle (Abler, Adams, and Gould (1971) was that spatial interaction is shaped by factors operating at the origin and destination but, crucially, by the distance separating them. This became enshrined in the notion of a “gravity model” that borrowed directly from Newtonian physics, where attraction is directly proportional to the product of masses of “bodies” and inversely proportional to the square of the distance separating them. Later reformulations and generalizations of such modeling (leading to classes of “spatial interaction” models rather than simple gravity formulations) offered much greater sophistication, though the importance of distance as a variable that constrained human interaction was still crucial. It remains an important component of any attempt to measure spatial accessibility (for example, to schools or health centers) though more sophisticated analyses use likely travel time along street networks as a measure of spatial separation, while other studies use route distances rather than travel times and some question the added value of using travel times rather than simple Euclidean distance. Distance also remains an attractive variable to economists attempting to model international trade flows as a class of
Distance

Spatial interaction (with national income used as a measure of “mass” and intercountry distance in the denominator). Income and distance explain a good deal of the variation in international trade.

At much the same time as social physics was developing, the Swedish geographer Torsten Hägerstrand was embarking on another classic study of spatial interaction, namely, the spread or diffusion of agricultural innovations. The key feature of his model was the “mean information field,” a construct which proposed that the ability of one farmer in rural Sweden to influence others was constrained by physical distance. Farmers who lived close to one another were more likely to meet and exchange ideas and information than those who lived further away. To the modern reader this may seem quaint, but the historical context (1940s rural Sweden) and Hägerstrand’s own empirical research in the 1950s bore witness to the fact that distance did indeed impede social interaction. The mobile phone, the tweet, and Facebook may well have eroded the importance of such interaction, but such devices and media were undreamt of 60 (even 25) years ago.

Distance in contemporary spatial analysis and GISci

The development of geographic information systems (GIS) and its accompanying scientific underpinnings (GISci) over the past 25 years has reasserted the importance of distance and location. The basic building blocks here are the accurate recording of locational data, and these data, coupled with the measurement of (usually physical) distances between objects, permit a considerable array of spatial analyses to be performed. A basic GIS procedure is the definition of “buffer zones” around point, line, or area objects and these require the setting of a distance limit; for example, a buffer zone of 50 m either side of streets in order to assess air quality, or circular regions of, say, 250 m around a set of schools in order to count the number of fast-food outlets within such catchment areas.

Geographers in the early 1960s drew on work by ecologists on the spatial distribution of plant species. These ecologists had devised a method called nearest neighbor analysis which sought to assess whether plants were distributed randomly in geographic space, or instead showed a tendency to cluster or alternatively display a more regular spatial pattern. Since empirical testing of central place theory was required, an obvious use of such methods was to detect departures from nonrandomness in settlement distributions. This required the measurement of distances from one settlement to its neighbors. More recently, much more sophisticated techniques for exploring spatial point patterns have become available, including the estimation of K-functions (to detect clustering) and kernel smoothing (to create an estimate of the density or intensity of the point pattern) – in effect, a continuous surface (O’Sullivan and Unwin 2010). These still require distance calculations. Such methods are now widely adopted in spatial epidemiology and spatial analysis of crime data. A critical issue is that these methods assume that the point locations (in human geographic applications, often residential locations) are fixed and meaningful. But people occupy activity spaces rather than being rooted to fixed locations, the distances between which may have little or no intrinsic significance. And other concepts of distance may mean more.

Physical (Euclidean) distance also features strongly in the problem of spatial interpolation. Given attribute data (perhaps precipitation, mineral ore content, or particulate air pollution, for example) sampled at a set of discrete locations it is often of interest to construct a best estimate of the continuous field from which these samples are taken, and in some cases to predict attribute...
values at locations which were not sampled. There are a variety of methods or techniques for doing this. The simplest case is to lay a regular grid over the study region and estimate values at grid points based on a weighted average of known values, where the weights are a function of distance; intuitively, it makes more sense for locations nearby to “contribute” more to the estimates than samples taken further away. More sophisticated methods of interpolation, such as kriging, make use of spatial autocorrelation in the given data, and also permit estimates of the uncertainty that attaches to each of these (O’Sullivan and Unwin 2010).

The classic work by Hägerstrand on the spatial diffusion of innovations was highlighted above. Other geographers have used some of these ideas in order to understand the spread of disease. Imaginative research on the spread of cholera in nineteenth-century America (Pyle 1969) showed how in 1832 disease spread was constrained by distance from the point of entry (New York City) but that a later epidemic in 1866 spread in a way that was structured more by distance in a space shaped by the recent rapid expansion of the rail network. In epidemiology distance, either in absolute or relative terms, remains relevant today, as seen by the spread of diseases such as SARS (severe acute respiratory syndrome) and new strains of influenza. SARS emerged in Guangdong, China in 2002, almost certainly having crossed the species barrier from civets to human because of the close proximity of animals to human populations and the keeping of animals in cramped conditions (Gatrell 2011, 142–144). The disease spread at a microscale within a Hong Kong hotel, but given Hong Kong’s location as a hub in the international airline network it was almost inevitable that the disease would diffuse quickly to other world cities (such as Toronto) that are either directly or indirectly connected to the source. Physical proximity still matters, certainly at a very localized scale, as on an airplane (Baker et al. 2010; see Figure 3). Yet distance in spaces that are structured by social and economic relations, and particularly by the nature of transport networks, matters more.

Distance matters in other public health-related contexts as, for example, in studies of the “walkability” of urban neighborhoods. Neighborhoods are “walkable” if they have good levels of street connectivity, a mix of land uses and residential densities, and facilities such as urban parks that make walking more attractive. A key element of walkability is the distance between residences, shops, and workplaces and geographers have made use of GIS to analyze variations in walkability and consequent variations in walking behavior. Given the rising levels of obesity in many countries this research is important.

Conclusion

Seen simply in terms of physical separation on the Earth’s surface, distance, both as a key geographic concept and as a factor in human life, has weakened over the years. Forty or fifty years ago, when the dominant paradigm in human geography was positivist spatial analysis, distance was perhaps the key variable that exercised the geographer’s interest. In general, and excepting the study of geographic information science, this is no longer the case. Relatively few contemporary students of geography engage with location theory, spatial interaction, spatial diffusion, and other subjects that were once part of a core geographic curriculum.

In part, this declining role in academic inquiry reflects its lessening importance as a factor influencing human behavior and socioeconomic life. Fifty years ago authors spoke of the “tyranny of distance” (Blainey 1966); now, others speak of the “death of distance,” though Cairncross
DISTANCE

(2001, 5) is quick to assert that while “the death of distance loosens the grip of geography ... it does not destroy it.” Blainey’s book referred to how the relative isolation of Australia – its distance relative to other parts of the world – had shaped its history and development. Now, cities such as Sydney and Melbourne are truly world cities that are connected into a global network. But they are still far away from Europe and North America and journeys to those continents are not a trivial undertaking. Despite the rapid, recent growth in communications technologies, if face to face meetings are required, or people wish to visit real places, distance still matters.

Over 40 years ago Waldo Tobler invoked geography’s first “law”: that everything is related to everything else, but near things are more related than distant things. It still holds. But we must acknowledge that there are alternative concepts of distance, and these matter a great deal – to geographers and many others working in different disciplines. The physical separation of places does not imply separation in other kinds of space. Given the complexity of modern communications and travel networks, simple conceptions of distance are unlikely to tell the whole story; physical separation is not always a good proxy for social, cultural, economic, or other kinds of separation.

SEE ALSO: Cartesian coordinate systems; Distance decay; Measuring spatial dependence; Spatial analysis; Spatial concepts; Spatial weights; Spatialization; Time geography and space–time prism; Tobler’s first law of geography

Figure 3  Plan of section of aircraft showing passengers according to their H1N1 (“swine flu”) infection category and seating position. Source: Baker et al. (2010).
References


Website

[www.csiss.org/classics](http://www.csiss.org/classics)
Earth system science

Noel Castree
University of Manchester, UK

Earth system science (ESS) goes back 30 years and today has considerable momentum. As one commentator puts it, ESS “embraces chemistry, physics, biology, mathematics, and applied sciences in transcending disciplinary boundaries to treat the Earth as an integrated system and seeks a deeper understanding of the physical, chemical, biological, and human interactions that determine the past, current and future states of the Earth” (Ruzek 2013). ESS thus has grand intellectual ambitions. As a science it uses evidence, logic, and physical laws to identify patterns, trends, and explanations that aspire to accurately represent the interactions between the components of the Earth surface (namely, the atmosphere, hydrosphere, lithosphere, cryosphere, and biosphere). However, unlike the specialist sciences (such as marine biology), ESS aims to “join the dots” by identifying relations and feedbacks between Earth surface phenomena often studied in relative isolation. Because the Earth surface comprises a set of interlinked “open systems” existing at a range of spatiotemporal scales, ESS’s quest to be interdisciplinary has faced some formidable intellectual challenges, as this entry will explain (see Environmental uncertainty). Geologist Richard Alley (2000, 7) likened it to creating the world’s most detailed ever “operator's manual” for a “machine” of unequalled complexity.

In its early years, ESS was global physical geography by another name – although relatively few geographers were actively involved in promoting it, for reasons explained below. Before and immediately after the millennium, ESS tended to bracket the human causes of global environmental change and focus largely on biophysical effects. More recently, however, its ambitions have broadened (see Global environmental change: human dimensions). Today, ESS looks more like the sort of environmental geography that, for some of academic geography’s founders (such as Halford Mackinder), defined the discipline’s mission as a “bridge” between the physical sciences (on the one side) and the social sciences and humanities (on the other) (see Geography and the study of human–environment relations). ESS increasingly considers two-way human–environment interactions rather than only humans’ biophysical impacts on Earth. However, again few geographers have championed ESS in its current form, although several have been involved in parts of its grand research agenda.

The why and how of this involvement is detailed in the final part of this entry. Before this, the history, institutional configuration, and evolution of ESS is described in three sections. Along the way some evaluative comments will be offered on successes and challenges. Thereafter, ESS’s relations and nonrelations with research by physical and environmental geographers is explored. Because this entry includes a great many acronyms, Box 1 itemizes them alphabetically to aid the reading process.
The origins of ESS

The precursor to Earth system science was the idea of an “operator’s manual for spaceship earth” formulated in 1968 by the remarkable American architect-inventor Richard Buckminster Fuller. Fuller criticized academic specialization and challenged his readers to aspire to “total knowledge” of a world they were fast-changing. He wrote at a time when humans, courtesy of the Apollo space missions, had seen their planet from space for the first time. The term Earth system science originated seventeen years later with a British mathematician and modeler of ocean and atmosphere dynamics, Francis Bretherton (1985). A University of Wisconsin professor, in 1983 he was appointed by the United States’ National Aeronautics and Space Administration (NASA) to chair a committee convened to consider how the Administration could most effectively fulfill its mission to observe the Earth surface. The committee comprised 16 members covering meteorology, atmospheric chemistry, marine biology, plant ecology, soils and vegetation interactions, agronomy, and geophysics (among other fields).

After a series of meetings in 1984 and 1985, the committee produced a report entitled *Earth System Science Overview* (NASA 1986). Page one described ESS’s goal as “understanding the entire

**Box 1** A list of acronyms used in this entry.

AIMES, Analysis, Integration and Modeling of the Earth System  
ESS, Earth System Science  
ESSP, Earth System Science Partnership  
FE, Future Earth  
GCRP, Global Change Research Program (USA)  
GECAFS, Global Environmental Change and Food Systems  
ICSU, International Council for Science  
ISSC, International Social Science Council  
IGBP, International Geosphere-Biosphere Program  
IHDP, International Human Dimensions Program  
NASA, National Aeronautics and Space Administration (USA)  
NERC, Natural Environment Research Council (UK)  
NOAA, National Oceanic and Atmospheric Administration (USA)  
NSF, National Science Foundation (USA)  
PAGES, Past Global Changes research project  
SDSN, Sustainable Development Solutions Network  
START, Global Change SysTem for Analysis, Research and Training  
UNEP, United Nations Environment Program  
UNESCO, United Nations Educational, Scientific, and Cultural Organization  
UNU, United Nations University  
WCRP, World Climate Research Program  
WMO, World Meteorological Organization  
WSSR, World Social Science Report
Earth System on a global scale by describing how its component parts and their interactions have evolved, how they function and how they might be expected to continue to evolve on all time scales.” On the same page, ESS’s “challenge” was to “predict those changes that will occur in the next decade to a century, both naturally and in response to human activity” (NASA 1986). The report included a diagrammatic presentation of the Earth system that would subsequently circulate widely in the research communities devoted to understanding global environmental change (Figure 1). The diagram expressed the holistic view recommended twelve years earlier in a (now quite famous) scientific paper. That paper proposed that Earth is seen as a self-regulating system and was authored by James Lovelock and Lynn Margulis (1974). Lovelock, a former NASA employee, went on to promote the so-called “Gaia hypothesis.” One aspect of this, which few took seriously in the early 1970s, was that humans’ combined activities might be sufficiently powerful to set in train alterations to the Earth system as a whole, not just one or other component of it.

As this diagram made clear 30 years ago, a number of important ingredients would be required if ESS was to become more than a grand, possibly idealistic, research idea. First, mechanisms for fostering cross-disciplinary exchange among the various geosciences would be essential. Second, high-resolution data about Earth surface patterns and processes at a range of spatiotemporal scales would be required. Third, there would be a need to ensure international academic cooperation among researchers in all parts of the world. Finally, there would be a need for computational and analytical techniques able to portray (and predict) complex Earth system dynamics across geographical space and through time. Recognizing this, NASA quickly used the Bretherton committee report to involve the USA’s National Science Foundation (NSF) and the National Oceanic and Atmospheric Administration (NOAA) in its aspiration to make ESS a reality.

One result of this outreach was the creation of the federal Global Change Research Program (GCRP) in 1989, underpinned by the 1990 Global Change Research Act. This program, involving multiple federal government agencies, committed considerable financial resources to promoting integrated environmental research. Its original strategic vision was threefold: “1. Establish an integrated, comprehensive monitoring program for Earth system measurements on a global scale, 2. Conduct a program of focused studies to improve our understanding of the physical, chemical, and biological processes that influence Earth system changes and trends on global and regional scales; and 3. Develop integrated conceptual and predictive Earth system models” (Committee on Earth Sciences 1989, 11–12). This trio of aims was framed by the overarching desire to “establish the scientific basis for national and international policymaking related to natural and human-induced changes in the Earth System” (Committee on Earth Sciences 1989, 9). Because of the United States’ leading role in academic and applied research globally, the creation of the GCRP had an almost immediate impact on research programs into anthropogenic environmental change elsewhere. For instance, the UK’s Natural Environment Research Council (NERC) announced an ESS initiative of its own from 2001. Around the same time, leading German physicist Hans Joachim Schellnhuber (1999) made Earth system analysis central to the Potsdam Institute for Climate Impact Research (founded 1992 and today a “hotspot” of ESS). Meanwhile, the term “Earth system” is routinely found in the latest strategy documents and research funding streams associated with the GCRP (National Science and Technology Council 2012).
Figure 1  Conceptual model of Earth system processes and interrelations. Source NASA 1986, 24–25.
ESS’s international evolution

The Bretherton committee report was more-or-less coincident with the creation of four global research programs focused on anthropogenic environmental change. These were the International Geosphere-Biosphere Program (IGBP, launched in 1987), which followed the World Climate Research Program (WCRP, created 1980), and which, in turn, was followed by the International Human Dimensions Program (IHDP, created in 1990, re-launched in 1996), and DIVERSITAS (launched in 1991, focusing on global biodiversity and biogeography). The four programs enjoyed support by numerous governments and became a vehicle for collaborative research among geoscientists across international borders. They emerged because of a widespread concern, articulated in the 1987 Bruntland Report on Sustainable Development, that humans were instigating planetary-scale environmental changes. That concern was expressed politically in 1992 at the United Nations “Earth Summit” held in Rio de Janeiro. Of all the programs, the IGBP took the Bretherton report’s aspirations to heart the most. From its inception, it tried to develop models – based on physical laws and mathematical logic – that could represent accurately the intricacies of the Earth system.

Despite their professed commitment to “integrated environmental analysis,” members of all four programs soon realized that some key issues were falling between the proverbial cracks. Well aware of the ESS goals articulated by Bretherton, senior members agreed to form a partnership in 2001 – the Earth System Science Partnership (ESSP). They also enrolled a capacity building and networking initiative designed to reach researchers in the Global South (called the Global Change SysTem for Analysis, Research and Training: START). Program resources were then used to co-design and implement a set of joint research projects, such as Global Environmental Change and Food Systems (GECAFS) (Table 1). Running for a decade (2001–2011), GECAFS’ aim was to “to determine strategies to cope with the impacts of global environmental change on food systems and to assess the environmental and socio-economic consequences of adaptive responses aimed at improving food security” (www.gecafs.org/). As this objective indicates, it sought to link research into climate change, soils, and vegetation to social scientific inquiry into farming, food storage and distribution systems, and much else besides. It was, in short, “intersectional” in focus. The system concept proved useful for enabling interchange between researchers from a diversity of disciplinary backgrounds. At the same time, researchers benefited from more numerous and higher resolution Earth observation technologies. This and other joint projects were eventually steered by an ESSP science committee (from 2007). The partnership also created a peer review journal, Current Opinion in Environmental Sustainability. Under its auspices, two manifestos-cum-bibles of ESS were published in 2004: one was the jointly authored Global Change and the Earth System: A Planet Under Pressure (Steffen et al. 2004); the other was an edited book Earth Systems Analysis for Sustainability (Schellnhuber et al. 2004).

Despite progress being made in GECAFS, specific separate projects within the IGBP, WCRP, IHDP, and DIVERSITAS have also contributed to the ESS aspiration for holistic analysis of global environmental change and its regional components. For example, the IGBP’s AIMES (Analysis, Integration and Modeling of the Earth System) project made great strides in representing computationally the dynamic interrelations between atmosphere, water, and land. Unlike the traditional image of a system as closed and self-regulating (unless externally forced beyond...
Table 1  A lexicon of global environmental change research organisations and initiatives. Adapted from Ignacuiik et al. 2012, 149, with permission.

<table>
<thead>
<tr>
<th>Definitions</th>
<th>Organizations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global environmental change programs</strong></td>
<td>DIVERsITAS, IHDP, IGBP, WCRP, FE (from 2014 onwards).</td>
</tr>
<tr>
<td>Programs are legally recognized scientific organizations that coordinate GEC research. They are co-sponsored by major agencies, such as the International Council for Science, the United Nations Educational, Scientific and Cultural Organization and the World Meteorological Organization.</td>
<td></td>
</tr>
<tr>
<td><strong>Partnerships</strong></td>
<td>ESSP</td>
</tr>
<tr>
<td>Partnerships are in-formal arrangements established by the GEC research programs to exchange ideas, synthesize and communicate integrative GEC research findings and conduct interdisciplinary research.</td>
<td></td>
</tr>
<tr>
<td><strong>ESSP joint projects</strong></td>
<td>GCP, GECAFS, GWSP &amp; GECHH</td>
</tr>
<tr>
<td>Joint Projects are sponsored by at least three GEC research programs, promoting interdisciplinary research across disciplinary boundaries (natural and social science). The ESSP Joint Projects are designed to directly address the two-way interaction between GEC and global and regional sustainability issues. The Joint Projects also benefit from the expertise and synthesized knowledge of the Core Projects and the GEC research community.</td>
<td></td>
</tr>
<tr>
<td><strong>Core projects</strong></td>
<td>For example, bioGENESIS (DIVERsITAS); Integrated Land Ecosystem-Atmosphere Processes Study (IGBP); Urbanization and Global Environmental Change (IHDP); Stratospheric Processes and their role in Climate (WCRP).</td>
</tr>
<tr>
<td>Core projects are disciplinary enterprises sponsored by one GEC research program, designed to research one specific field/scientific challenge.</td>
<td></td>
</tr>
<tr>
<td><strong>Regional networks</strong></td>
<td>Asia-Pacific Network for Global Change Research (APN), Inter-American institute for Global Change Research (IAI), and global change SysTem for Analysis, Research and Training (START).</td>
</tr>
<tr>
<td>Regional networks provide opportunities to enhance GEC research and networking capacity, particularly in developing countries.</td>
<td></td>
</tr>
</tbody>
</table>
operational boundaries), AIMES has made use of wider scientific thinking about complexity, nonlinear behavior and irreversible tipping points. The new incarnation of AIMES aspires to incorporate human behavior and responses, and to help decision-makers anticipate serious socioenvironmental problems before they appear (van der Leeuw 2013).

This aspiration leads back to the original inspiration for the creation of the ESSP. Many projects within the four global change research programs have arguably advanced “narrow” interdisciplinarity rather than “wide” interdisciplinarity. For instance, computer models of the sort used in many WCRP and IBGP projects have thus far left out many social scientists whose research into “human dimensions” of environmental change focuses on politics, social power, societal resilience, and community capacity building as they relate to avoiding, or adapting to, biophysical hazards. This is, perhaps, consistent with the externalization of those dimensions evident in the 1985 Bretherton diagram (to the right in Figure 1). Even the joint projects conducted under ESSP auspices have had their limits. Consider the Global Carbon Project (established in 2001. Its professed aim is to offer a “complete picture of the global carbon cycle, including both its biophysical and human dimensions together with the interactions and feedbacks between them” (www.globalcarbonproject.org/). However, it has been dominated by geoscientists and, among the social scientists involved, economists have loomed large. This means that other disciplines focused on other human dimensions of carbon acquisition and use have been rather side-lined. Meanwhile, some nationally funded ESS projects have remained unapologetically focused on physical dimensions – such as the QUEST (Quantifying and Understanding the Earth System) project funded by the UK’s NERC. What is more, some have wanted to take ESS in an applied direction by treating coupled human–environment systems as objects for a new generation of engineering and management approaches. For instance, in his manifesto for ESEM (Earth Systems Engineering and Management), Brad Allenby argued that “the anthropogenic Earth is a difficult, highly complex, tightly integrated system that challenges society to rapidly develop tools, methods, and understandings that enable reasoned responses. Engineers in general, and civil and environmental engineers in particular, must be a critical part of any such response” (Allenby 2007, 7961). This chimes with Richard Alley’s earlier mentioned notion of ESS as providing an “operator’s manual” if it achieves sufficient analytical precision and predictive accuracy.

Why the limited focus on human dimensions, going back to Bretherton – notwithstanding the creation of the IHDP? Arguably, it reflected the wider, long standing “divide” between the natural sciences and the disciplines studying people – a divide British novelist and scientist CP Snow (1959) famously described as academia’s “two cultures problem.” In essence, the divide reflected the ontological differences between the biophysical world and the world of people. Unlike rocks and rivers, people are creatures of both reason and emotion, able to be proactive and reactive in relation to nature and other humans. Consequently, “scientific” approaches are not wholly appropriate to understanding human thought and behavior. As far back as the 1960s, the German social theorist Jurgen Habermas (1968; English translation 1972) argued that “hermeneutic” and “critical-emancipatory” approaches were equally important. The former elucidates various human values, beliefs, feelings, and desires using methods like ethnography, interviews, and focus groups. These methods do not aim for objectivity and precision in the same sense as most
natural science research techniques. The latter highlights the inequalities rife in various societies and reveals the values and goals associated with marginalized or oppressed social groups (with a view to achieving them in practice). Again, where natural science aims for objectivity, critical-emancipatory research is overtly political and researchers may align their own activities with those of the disadvantaged people they are researching (see Participatory action research).

The limitations in applying a scientific approach to understand human dimensions help us understand why ESS, on the international stage, is now morphing into something less dominated than heretofore by climatologists, oceanographers, and other geoscientists. This mirrors changes in the way research into climate-society interactions has changed since the Intergovernmental Panel on Climate Change published its fourth global assessment in 2007. The most recent report (published in 2014) makes a considerable effort to focus on human responses to climate change and policy options for better mitigation and adaptation measures.

However, the intellectual and practical barriers to success in broadening ESS’s intellectual reach remain extremely high. Despite (i) improvements in the quantity and quality of available Earth surface data, (ii) increased computational power and sophistication, and (iii) success in international cross-disciplinary team working, problems remain. First, most geoscientists are interested in identifying what has happened and will happen to the Earth’s environment. However, many social scientists (and humanists) are concerned with normative issues – that is, with debates over what could and should happen, depending on the values and institutions that might influence human decision making looking ahead. Normative questions can be logically debated; but they are not amenable to measurement and modelling. They speak to humans’ capacity to change the way they think and act, perhaps dramatically so in the face of perceived existential threats. Large – perhaps unprecedented – societal changes would radically alter the future course of the Earth system.

Secondly, practical issues arise in trying to synthesize knowledge among researchers who have different “worldviews.” For instance, where environmental economists might be keen to put a price on carbon dioxide emissions, an anthropologist might want to focus on the ensemble of institutions, social relations, and cultural norms that dispose whole societies to burning fossil fuels on a mass scale. Finally, practical issues also arise in assembling data on current patterns of human behavior. Such behavior cannot be monitored by satellites in way Amazonian forest cover can. Instead, different sorts of data from a myriad of sources (governmental and academic) need somehow to be synthesized to better understand patterns, variations, and transitions in how people influence and are influenced by global environmental change. One complication here is that past and present human behavior is not governed by a universal rationality; instead, it is profoundly conditioned by the norms, relationships and power dynamics of different societies. Added to this, predicting such behavior is very difficult when compared to the challenging, but more tractable, problem of predicting future environmental change in light of current biophysical knowledge.

ESS today: the Future Earth framework for global change research

As noted already, ESS is now transitioning into a more transdisciplinary endeavor. A recent definition by some of those involved captures an intellectual broadening that is central to this. According to Leemans et al. (2009, 5), ESS today
EARTH SYSTEM SCIENCE

aims to “observe, understand and predict global environmental changes involving interactions between land, atmosphere, water, ice, biosphere, societies, technologies and economies” (emphasis added). The ESSP has come to an end, along with the IHDP. With the three other global change research programs and START, the ESSP has fed into a new initiative called Future Earth (www.futureearth.org/). Future Earth (FE) was launched at the 2012 Earth Summit (otherwise known as Rio+20). It has high level support globally. It is sponsored by the Science and Technology Alliance for Global Sustainability comprising the International Council for Science (ICSU), the International Social Science Council (ISSC), the Belmont Forum of national funding agencies, the new Sustainable Development Solutions Network (SDSN), the United Nations Educational, Scientific, and Cultural Organization (UNESCO), the United Nations Environment Program (UNEP), the United Nations University (UNU), and the World Meteorological Organization (WMO).

Several joint projects and separate projects from the ESSP and the four global change research programs are rolling forward under FE auspices. However, more than in previous research, FE aims to foster “wide-angle and relevant” inquiry that speaks more to (i) human dimensions (including development in the global South) and (ii) the needs of decision makers (in politics, commerce and civil society) and stakeholders. To quote from its strategy document, FE places “a strong emphasis on full integration among scientific disciplines, on engagement with societal partners in co-designing and co-producing knowledge, on international collaboration, on producing knowledge that is valuable to decision-makers, and on generating the solutions that society needs” (Future Earth 2014). Indicative of this, only the first of its three overarching themes speaks to the “pure research” interests of many geoscientists (the themes are, respectively, “dynamic planet,” “global sustainable development,” and “transformations towards sustainability”). This said, it will be several years before projects more focused on the human side of the human–environment relationship come to fruition. In turn, such fruition will only be possible if sufficient funds are made available to pay for the time and travel needs of myriad researchers from all parts of the globe. Additionally, the intellectual and practical challenges of achieving “wide interdisciplinarity” – noted previously – will remain. In the meantime, various environmental scientists aspire to deepen the complexity and predictive power of their Earth system concepts and models (Liu et al. 2015).

To summarize, ESS is not so much a single coherent research approach as a set of approaches to research, supported by various funding streams and institutions. These approaches, sometimes overlapping, sometimes sharing a family resemblance, together aim to provide a holistic analysis of Earth surface change (including prediction of future states in some cases). For many years dominated by geoscientists favoring large-scale Earth observation technologies and computational models, it can be seen that ESS is now broadening out. In the process, the notion of an Earth system is likely to serve as little more than a metaphor for many social scientists (and humanists) who are less disposed than some geoscientists to see humans as one (albeit important) “component” within a wider set of parts and connections.

This is evidenced by the recent world social science report (WSSR) sponsored by the International Social Science Council (2013). Entitled Changing Global Environments, it does not foreground the Earth system concept, even as it accepts the need for “joined up” analysis across the natural sciences and the “people disciplines.” Instead, it accents the diversity of human values and ambitions, the need for radical change to
our fossil-fuel driven way of life, and the role of social science in such change through fostering democratic debate about alternative future pathways for humanity. (A similar emphasis attaches to recent social science discussion of “socio-environmental resilience,” thereby complicating a concept – resilience – that began its life in ecology: see Social resilience and environmental hazards). Where the “Earth system” idea invites the image of an omniscient engineer “fixing” the machine, the WSSR promotes the idea that any “fix” will need to be socially negotiated between diverse cultures, governments, NGOs, researchers, and so on. In this context, geoscientists will be servants of diverse political agendas not “value free” experts presenting “objectively” relevant facts and “neutral” technological solutions. The seeds of this alternative approach were already sown in some projects sponsored by the IHDP, notably the Global Environmental Change & Human Security Project (1999–2010) (O’Brien and Barnett 2013).

ESS and geography

How have geographers responded to ESS over the 30 years or so since NASA’s Earth System Science Overview was published? It might be assumed that geographers were, if not ESS pioneers, then at least enthusiastic followers and contributors in the years after the Bretherton report. After all, as geographer John Wainwright noted, “ESS sits in the nexus of the physical and human worlds, which was the traditional definition of the discipline” (Wainwright 2009, 162). Indeed, one non-geographer said of ESS some years ago that it “… is actually twenty-first century geography: it encompasses the study of the environmental physical and life sciences and engineering, coupled with an analysis of human constructs and political and economic policies. It employs space-age technologies to identify, measure, and manage diverse global databases that serve as a framework and foundation for a coherent discipline.” (Ernst 2000, 520). However, as will be seen below, only a relatively small number of geographers have been actively involved in the various ESS developments detailed above. The question then arises: What explains the “missing geographers”? The answer lies with the particular evolution of Anglophone academic geography after 1945.

Briefly, once the World War II finished, Anglophone geography began to become a more internally fragmented discipline, albeit with aspirations to specialize topically using a shared “scientific” approach across the discipline as a whole (see the entry Nature). One result of this was that physical geography moved away from large-scale landscape studies and focused more on in-depth local scale research into Earth surface processes and forms. Another result was that a great deal of human geography began to bracket environmental issues, focusing on the locational aspects of economy, politics, society, and culture with little reference to the things physical geographers investigated. Consequently, despite landmark volumes such as Man’s Role in Changing the Face of the Earth (Thomas 1956), relatively few practitioners occupied the “middle ground” between human and physical geography by the mid-1970s. One underlying reason for this was the embarrassment of pre-1939 “environmental determinism” in geography. For several decades, leading geographers had tried to explain the relations between diverse human societies and their environments. For instance, in the United States, practitioners like Ellen Semple and Ellsworth Huntington were apt to argue that certain physical environments produced human “races” less intellectually or physically capable than Europeans. These arguments were
presented as descriptions of “historical accidents” that had blessed some races but not others. Indeed, both Semple and Huntington on occasion claimed that even “Europeans” would “decline” if they had transplanted a region like West Africa. After 1945, many English-speaking geographers were understandably determined to fashion a more precise and less contentious geography that had no truck with unverifiable generalizations about environmental causation (see Environmental determinism).

As a result of all this, by the mid-1980s academic geography had many capable specialists in the various branches of human and physical geography. However, relatively few “environmental geographers” existed, let alone ones working at a global scale on complex society–nature dynamics across space and through time. Interestingly, some leading geographers had – in the late 1960s and early 1970s – used systems thinking to provide a common analytical approach for human and physical geography alike (Chorley and Kennedy’s 1971). However, this approach gained little traction outside parts of physical geography. One reason is because “humanism” became influential in parts of human geography, and focused on the feelings, norms and values of diverse human agents – thereby challenging the idea of a universal “rationality” (see Humanistic geography); another reason is that the rise of Radical geography produced criticisms of system thinking as a means of controlling not only environmental phenomena but human behavior (Olson 1972). As a result, the idea that there are objectively knowable socioenvironmental systems had been called into question by the time ESS was “launched” in 1986.

Even so, several geographers have played a role in ESS since its inception, working alongside researchers in various other disciplines. Some are physical geographers involved in the IGBP, such as Quaternary scientists Frank Oldfield and John Dearing. Oldfield was co-author of among the first educational fruits on an ESS approach, the earlier mentioned book Global Change and the Earth System: A Planet Under Pressure (Steffen et al. 2004). Meanwhile, Dearing was an active researcher in the long-standing past global changes (PAGES) project of the IGBP, which continued under Future Earth auspices. Its stated aim was to “support research aimed at understanding the Earth’s past environment in order to make predictions for the future” (the IGBP has now been terminated). Meanwhile, others are human–environment geographers focused on contemporary environmental changes. Examples are Eric Lambin, Billie Lee Turner II, and Diana Liverman. All three have been key figures in developing Land systems science, an attempt to systematically understand the drivers and effects (on nature and humans) of alterations to terrestrial vegetation, soil, water, and so on. Turner was (with Oldfield) co-author of the above mentioned book by Steffen et al. (2004), while Lambin coedited a major synthesis book about land use and land cover change (Lambin and Geist 2006). During the same period, Liverman was chair of the scientific steering committee of the earlier mention ESSP GECAFS project. She coedited a major edited collection entitled Food Security and Global Environmental Change (Ingram, Ericksen, and Liverman 2010).

These are just a few examples of those geographers directly involved in the joint and single projects organized by the ESSP and the various global change research programs. Others have undertaken theoretical and empirical research that resonates with the ESS agenda without much formal involvement. Sometimes their research has been funded by the national research councils mentioned earlier in this entry. A good example is Tim Lenton, a natural scientist who models ocean–atmosphere interactions and is now based
in geography at the University of Exeter. Lenton has, among other things, advanced the idea of “planetary boundaries” (see Anthropocene and planetary boundaries).

Through these sorts of involvements, a set of physical and human–environment geographers have made honorable contributions to the evolution of ESS. Others, however, have either pursued similar projects by others means or else expressed misgivings about the aims of ESS. Perhaps the best example of the former is the Sustainability science advocated since the late 1990s by American human–environment geographer Robert Kates. It emerged formally in 2001 out of the same academic conference (held in Amsterdam) that inspired the creation of the ESSP. Kates was frustrated that more progress had not been made in integrating knowledge across the “human–physical” divide in research among the environmental disciplines. In the original manifesto for sustainability science, it was described as seeking to “encompass the interaction of global processes with the ecological and social characteristics of particular places and sectors. The regional character of much of what sustainability science is trying to explain means that relevant research will have to integrate the effects of key processes across the full range of scales from local to global” (Kates et al. 2001, 641). The overt reference to sustainability gave it an applied, normative feel that was rather muted in the early statement about the aims of ESS (see Sustainable development). Today sustainability science attracts a number of devotees from across the disciplines, although is rather dominated by science and technology-minded researchers.

In contrast to Kates and those directly involved in ESS, some geographers have questioned it keenly. This may seem odd given that ESS, as noted, shares the integrative, holistic ambitions many attribute to the discipline of geography. However, not everyone in geography approves of the particular vision of integration found in the first 20 years of ESS. For instance, the physical geographers Nick Clifford and Keith Richards (2005, 379) argued in that “In its appeal to a combination of abstract generalization and complex modelling, and to a technocentric fusion of the physical and social worlds, ESS may yet prove to be a re-invention of scientific privilege and practice, with both a language and mode of expression which are restricted rather than general, and where the claims to apprehend the real and complex are, once more, by the few on behalf of the many.” In their view, geographers should resist the siren-song of ESS and aim for more plural forms of “joined-up” analysis alive to geographical variability and the role of things not exclusively amenable to “scientific analysis”. They make a particular pitch for local scale knowledge integration on the basis that anthropogenic environmental change unfolds in geographically varied and particular ways that are felt differently by various stakeholders (Richards and Clifford 2008). They called this LESS (local environmental systems science) as opposed to ESS.

In the years since they published this critique, a number of geographers have contributed to the broadening of ESS described earlier in this entry. Among the most prominent is Karen O’Brien, a human–environment researcher at the University of Oslo, Norway. She now sits on the science committee of Future Earth. In a series of powerful articles (e.g., O’Brien 2011), she has enjoined geographers to show how predicting the future of the “Earth system” opens crucial questions about the sort of world we want our descendants to inherit. For her, there is not one “correct” way to manage a planet under pressure from unprecedented human impacts. Instead, she argues, geographers should illuminate the diverse ways that different communities might want to fashion a new Earth less burdened by endless commodity.
production, consumption, and disposal. For her, human and physical geographers can combine their knowledge of biophysical phenomena and society in ways that resist the notion of single “fixes” at any one spatiotemporal scale. This speaks to a conception of the “unity” of geography as a discipline rather different to those fed to geographers in print by senior figures down the years (on which see Castree 2015). It calls for plural forms of “intradisciplinarity” that see human and physical geographers collaborate to analyze a range of socioenvironmental challenges mindful of the various ways those challenges can be interpreted culturally and politically. However, there are risks. For example, some might worry that research findings will be seen as less credible if value questions and normative concerns are brought into the heart of the exercise.

**Summary**

ESS is today a term that circulates far and wide among the various geosciences, reflecting its origins in the Bretherton committee “wiring diagram” of 1986. There are nowadays various programs, professorships, and even departments that have ESS in their titles, usually in faculties of earth and environmental science. There are also several introductory level science texts with “Earth system” or ESS in their main or subtitles. The language of the “Earth system” now peppers publications by, and discussions between, researchers involved in the global change programs that have fed into Future Earth. However, ESS is now expanding beyond the geosciences and trying to better connect with the various “people disciplines.” This not only reflects its holistic intellectual aspirations but also the important fact that people are arguably the most important “component” at stake if the Earth system is to be managed more sustainably. As ESS broadens, questions arise about whether social scientists participating in it, so too humanists, will be comfortable retaining the ESS label to describe their collaborations with geoscientists.

Several geographers, while not always lead players, have played an important role in evolving ESS. That more practitioners have not joined them reflects the particular condition on academic geography when the NASA report was released in the mid-1980s. Even so, geographers continue to enter the fray, most recently those outside physical geography concerned that a richer conception of “human dimensions” informs initiatives like Future Earth. Geography is well placed to make more substantial contributions that challenge the assumption that there is one way to describe, explain, and predict the behavior of the Earth system. As such it can help democratize the essential question of what sort of Earth we should bequeath our successors, including nonhumans.

**SEE ALSO:** Anthropocene and planetary boundaries; Environmental determinism; Environmental uncertainty; Geography and the study of human–environment relations; Global environmental change: human dimensions; Humanistic geography; Land systems science; Nature; Participatory action research; Radical geography; Social resilience and environmental hazards; Sustainability science; Sustainable development

**References**


Further reading


Earth’s energy balance

Tristan S. L’Ecuyer
University of Wisconsin–Madison, USA

The Earth’s weather and climate are strongly influenced by the exchanges of energy (fluxes) between the sun, the atmosphere, and the surface. Collectively referred to as Earth’s energy budget (EEB), these energy flows and their geographic variations establish the long-term average temperature structures of the atmosphere and oceans, drive global wind patterns and ocean currents, and define where clouds and precipitation are most likely to occur. Most other aspects of our climate follow from these basic quantities. Quantifying the factors that influence Earth’s energy balance is, therefore, of paramount importance to understanding the climate system and predicting and mitigating the effects of global climate change.

In the broadest sense, “Earth’s energy balance” refers to the balance between the incoming energy from the sun and the thermal energy emitted by the planet. The sun closely resembles a blackbody at a temperature of approximately 5778 Kelvins (K). At an average distance of 149.598 million kilometers from the sun, this amounts to an energy density of 1360 Watts per square meter, a value now known with high precision thanks to measurements from the Total Irradiance Monitor (TIM) on the Solar Radiation and Climate Experiment (SORCE) (note that fluxes of energy in the Earth’s climate system are generally reported in units of Watts per square meter, or Wm$^{-2}$, that represent the amount of energy that flows through a 1 m$^2$ surface in 1 second). Because the Earth is approximately spherical, this energy illuminates a surface area four times larger than the area intercepted by a flat disk of the same radius. The resulting energy density at the top of the Earth’s atmosphere is, therefore, 340 Wm$^{-2}$. About 29.4% ($\pm 0.3\%$) of this incoming energy is reflected, primarily by clouds and bright surfaces such as deserts and the polar ice caps, while the remainder is absorbed to heat the planet.

Thus the Earth receives a total flux of 123.5 PW (1 PW = 1 petaWatt = $10^{15}$ Watts) from the sun. For perspective, this flux produces more energy in three days than has been consumed by human activity since the dawn of civilization based on International Energy Agency (IEA) estimates. This incoming solar energy is largely balanced by 123.2 PW of thermal emission such that the Earth appears to an outside observer as a blackbody with a temperature of $-18^{\circ}$C. The 0.3 PW residual represents the excess energy trapped by the atmosphere as it adjusts to higher concentrations of greenhouse gases from anthropogenic emissions and associated climate feedbacks. Tracking the fate of this excess energy is a more challenging problem. Measurements of changes in ocean heat content from a large network of buoys known as the Argo Array suggest that approximately 90% of this excess energy is absorbed in the upper layers of the world’s oceans while the remainder contributes to melting ice sheets, warming and moistening the atmosphere, and heating the deep ocean.

The global balance between incident solar radiation and Earth’s emission is only a small part of the story. Local imbalances owing to variations in the amount of solar radiation received, thermal radiation emitted, and the partitioning
of energy between the atmosphere and the surface in different regions are responsible for the different climates encountered on Earth. Over the course of a year, the tropics receive about six times more solar radiation than the poles but only emit about twice as much thermal radiation. The stronger latitude-dependence of incoming sunlight leads to a surplus of radiation near the equator and a deficit of radiation at the poles. These imbalances induce large circulations in the atmosphere and oceans that redistribute heat and restore local equilibrium on multiyear timescales. Peak poleward heat transport occurs at the latitude where the annual mean incoming solar radiation exactly balances the thermal emission. This occurs at 36.9°S in the Southern Hemisphere and 35.0°N in the Northern Hemisphere (Stephens and L’Ecuyer 2015). Between these latitudes, the atmosphere and oceans transport approximately equal amounts of heat out of the tropics while large storms, called extra-tropical cyclones, transport heat poleward at higher latitudes. The resulting zones of convergence and divergence in the tropics and storminess at mid-latitudes define the typical weather conditions experienced around the planet. Beyond sea level rise and melting ice sheets, altered global weather patterns that could result from the changing regional energy gradients that accompany climate change (whether natural or human-induced) have profound implications for life on Earth.

**Historical perspective**

Characterizing energy exchanges between the surface, atmosphere, and space from observations has been the subject of vigorous research for more than a century. Hunt, Kandel, and Mecherikunnel (1986) summarize many early efforts to reconstruct Earth’s energy budget from ground-based observations at a number of locations scattered around the Northern Hemisphere and carefully thought-out balance arguments. It wasn’t until the development of weather satellites as part of the space race in the latter half of the twentieth century, however, that we obtained a true global picture of the EEB. New space-based observations acquired in the 1960s revolutionized our understanding of Earth’s radiative balance by providing a unique global perspective on the spatial distribution of incoming and outgoing radiation at the top of the atmosphere. In just a few years following the flight of the first radiation budget instrument aboard the Explorer VII satellite we learned that the Earth was darker and warmer than previously believed and that the gradient of absorbed solar energy between the tropics and the mid-latitudes was much larger than earlier estimates (Vonder Haar 1994). In the decades that followed, satellite observations with improved calibration, increased spatial and temporal resolution, and increased sensitivity to surface and atmospheric properties led to regular refinement of Earth’s energy budget.

The launch of three satellites that collectively comprised the Earth Radiation Budget Experiment (ERBE) from 1984 to 1986 marked another important milestone in our understanding of the Earth’s energy balance. Each ERBE instrument collected observations of solar, thermal, and total fluxes radiating from the top of the Earth’s atmosphere with sufficient accuracy to establish (with the aid of ancillary cloud information) the partitioning of fluxes between the surface, troposphere, and stratosphere. At the same time, rapid increases in computing power enabled the development of complex numerical models capable of representing the dynamic and thermodynamic processes that govern the three-dimensional flows of energy around the globe. These early climate models and the observation-integrating models (termed
“reanalyses”) that evolved from them in the 1990s simulated all components of Earth’s energy balance, including surface turbulent heat fluxes, enabling complete reconstructions of Earth’s energy budget and explicit calculations of the amount of heat transported by atmospheric circulations and ocean currents.

The influence of clouds on Earth’s energy balance

New observational and modeling tools developed near the end of the twentieth century also spawned several new areas of investigation devoted to quantifying the factors that influence Earth’s energy balance. These include aerosol particles, ocean surface temperature, snowpack and sea ice, vegetation, surface winds, and variable trace gases such as water vapor. Of particular significance is a large branch of research that centers on quantifying the role of clouds in modulating regional energy imbalances and, in turn, atmospheric circulations and ocean currents. Clouds are especially important players in the EEB since they influence both the amount of solar radiation the Earth absorbs and the amount of thermal radiation it emits. Since clouds are generally brighter than the Earth’s surface, they increase the amount of sunlight reflected back to space, cooling the planet in the same way they provide us with relief on a hot sunny day. High ice clouds, on the other hand, can be several tens of degrees colder than the average emitting temperature of the Earth, thus reducing the amount of thermal energy emitted to space. In this way clouds can also be thought of as blankets that trap heat and keep the planet warmer. This is the reason cloudy nights tend to be much warmer than clear nights in winter.

The net effects of any particular cloud on the Earth’s energy budget depend on numerous factors including its height, the amount of condensed water it contains, whether it is composed of liquid or ice, the position of the sun in the sky, and the brightness of the underlying surface. Since no two clouds are exactly alike, quantifying the effects of clouds on Earth’s energy budget is a challenging problem that only became possible in the satellite era. Harrison *et al.* (1990) demonstrated that when the effects of all clouds around the globe are summed over the course of the full year, clouds approximately double the amount of sunlight the Earth reflects from about 15% in clear skies to the observed 29.4%. This represents a reduction of incoming solar energy of 49 Wm$^{-2}$. Similarly, clouds reduce the amount of energy emitted by the Earth to space by about 28 Wm$^{-2}$ (equivalent to an 8°C change in the Earth’s emitting temperature relative to a cloudless planet). Thus, on aggregate, clouds cool the Earth by approximately $21 \pm 4$ Wm$^{-2}$ relative to the cloud-free atmosphere.

It should not come as a surprise that the influence of clouds varies significantly with location and atmospheric conditions. Figure 1 shows a map of the annually averaged cloud influence on radiation at the top of the atmosphere from the successor to ERBE, the Clouds and the Earth Radiant Energy System (CERES) instrument that flies aboard multiple Earth observing satellites (Wielicki *et al.* 1996). This figure highlights the vastly different effects clouds have on the environment in different regions, including their much stronger cooling effect over subtropical oceans and their warming effects over bright surfaces such as deserts and ice sheets. It also highlights a somewhat counterintuitive fact about the role of clouds in Earth’s energy budget: although the deep cumulonimbus clouds (thunderstorms) that occur in the tropics are the most impressive and imposing cloud systems on
the planet, these clouds do not play as important a role in defining the climate as their much more benign, shallow counterparts that reside in the stratuscumulus decks off the west coasts of most continents. This puzzling result can be understood when one considers that while cumulonimbus clouds are bright and reflect a lot of sunlight back to space, they also grow very deep and trap a considerable amount of thermal emission owing to their high, cold cloud tops. From a top-of-atmosphere energy balance perspective, these effects approximately cancel one another, limiting the influence of these intense cloud systems on the net radiation balance. Shallow clouds, on the other hand, tend to be much warmer and are often composed of liquid droplets that reflect solar radiation very effectively but emit almost as much thermal radiation as the surface. These clouds are, therefore, primarily responsible for cooling the Earth relative to cloud-free conditions.

Earth’s energy budget at the start of the twenty-first century

With the successful launches of numerous Earth observing satellites in recent years, it can be argued that the early twenty-first century was a golden era of Earth’s energy budget measurements. The combination of radiative fluxes from CERES and cloud information from three new cloud sensors, the Moderate Resolution Imaging Spectroradiometer (MODIS) and, more recently, CloudSat and the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO), for example, provides unprecedented data for determining the role of clouds in radiative exchanges between the Earth and space. Observations in the microwave region of the electromagnetic spectrum have provided new insights into the global distribution of latent heat release in precipitation, while algorithms designed to infer near-surface temperature and humidity have facilitated the inference of evaporation and sensible heat fluxes from the surface through bulk formulas. Growing networks of surface-based radiation measurements and recent advances in computing power have also provided substantially better constraints on surface radiative and turbulent heat fluxes and atmospheric heat and moisture transports.

Together, this combination of satellite observations, in situ measurements, and models are providing the most comprehensive description of Earth’s energy budget to date. A depiction compiled by L’Ecuyer et al. (2015) from state-of-the-art National Aeronautics and Space Administration (NASA) Earth Observing Satellite (EOS) data sets is presented in Figure 2. These estimates invoke new techniques to impose known closure relationships and derive associated error bars that represent our current level of understanding of each component of Earth’s energy budget. Figure 2 shows that every square meter of the Earth’s surface is warmed by $164 \pm 5 \text{ W}$ of solar energy and an additional $341 \pm 5 \text{ W}$ of thermal energy emitted by the atmosphere to the surface. These sources of surface heating are offset by $81 \pm 4 \text{ W}$ of latent heat
Figure 2 A global energy budget reconstruction derived from satellite observations over the first decade of the twenty-first century (adapted, with permission, from L’Ecuyer et al. 2015). All fluxes are expressed in Watts per square meter (Wm$^{-2}$). The value of each flux represents the aggregate from all surfaces around the globe, but longwave and shortwave fluxes are plotted over land and ocean regions, respectively, for convenience. Alternative reconstructions based on satellite observations, ground-based measurements, and global model reanalyses can be found in Trenberth, Fasullo, and Kiehl (2009), Stephens et al. (2012), and Wild et al. (2013).

Transfer from the evaporation of water (primarily from the oceans), 25 ± 4 W of sensible heat transfer (heat mixed into the atmosphere by the wind blowing over the surface), and 399 ± 5 W of cooling by thermal emission from the surface. Thus the energy absorbed at the surface is very nearly balanced by heat transfer to the atmosphere. Independent estimates of changes in ocean heat content from buoy measurements suggest that the Earth’s surface is currently absorbing about 0.5 Wm$^{-2}$ of energy due to the trapping of thermal energy owing to the increases in carbon dioxide and water vapor concentrations in the atmosphere that have occurred since preindustrial times. While this may not seem like a lot of energy, when multiplied over the entire area of the global oceans it amounts to about 300 trillion Watts, or almost 17 times the world’s annual energy consumption estimated by the IEA. Careful efforts to track this energy suggest that 90% is absorbed in the oceans while the remainder is taken up by melting of the large Greenland and Antarctic ice sheets, melting of sea ice, and warming of land and the atmosphere.

Future challenges

While our accounting of the energy flows that make up Earth’s energy balance has improved considerably in the last century, there remains a lack of consensus among the recent energy balance reconstructions, indicating there is room for improvement (e.g., Trenberth, Fasullo, and Kiehl 2009; Stephens et al. 2012; Wild et al. 2013; L’Ecuyer et al. 2015). Furthermore, uncertainties remain in satellite observations of reflected solar and emitted thermal radiation at the top of the atmosphere as well as in buoy measurements of ocean heat content. As a result, the magnitude of present-day energy imbalances owing to increased greenhouse gas concentrations is only known to an accuracy of about 50% (Trenberth, Fasullo, and Balmaseda 2014). Better absolute calibration of satellite-based measurements of thermal and reflected solar fluxes and longer, more accurate estimates of ocean temperature structure are required to more precisely establish the magnitude of energy imbalances owing to human-induced climate change.

The primary influence of energy flows on the climate is through the regional gradients that drive the large-scale circulations that transport heat from regions of surpluses to regions of deficits. Assessing the influence of changing energy imbalances on the climate therefore requires a detailed accounting of local thermodynamic processes and feedbacks including evaporation, precipitation, and cloud cover. These processes are difficult to observe globally yet they represent the aspects of global climate
models that are most important to improve. As a result, the spread of predicted regional climate sensitivities to human activities in today’s climate models remains unacceptably large and continues to be linked to differences in the way models simulate cloud responses to temperature, humidity, and aerosol concentrations. Accurately representing these so-called cloud feedbacks in climate models is a long-standing concern of the climate community, and documenting model biases continues to be the focus of a large body of research. In fact, the Intergovernmental Panel on Climate Change has stated that uncertainty in simulating cloud feedbacks in response to increasing greenhouse gas concentrations likely remains the largest obstacle for accurately projecting future climate with today’s climate models (IPCC 2013).

Based on these considerations, the World Climate Research Program (WCRP) identifies clouds and their role in the Earth’s energy balance as one of the “Grand Challenges” or major areas of scientific research required to advance climate science in the coming decade. To address this challenge and improve climate model predictions, the WCRP further identifies three questions that must be addressed to gain a better understanding of the central role clouds play in defining Earth’s energy balance:

1. How do clouds couple to circulations in the present climate?
2. How will clouds and circulation patterns respond to global warming or other forcings?
3. How will clouds feed back on climate through their influence on Earth’s radiation budget?

Meeting these challenges will require a combination of improved modeling, observations, and analysis techniques. Given the important role satellite observations have played thus far, it is likely that continued refinement of space-based observing systems will play an instrumental role in advancing our understanding of Earth’s energy balance. Continued support for future missions with new technologies for improving the precision to which individual component energy fluxes are measured and quantifying the factors that modify them is critical. Particular attention should be given to quantifying biases in component fluxes on regional scales, and it is anticipated that analysis of data from the Global Precipitation Measurement (GPM) mission, Soil Moisture Active Passive (SMAP), and Suomi NPP (National Polar-orbiting Partnership) satellites will facilitate progress toward this goal. Proposed future satellite missions dedicated to observing cloud and precipitation processes on global scales with sophisticated, multifrequency, Doppler radars offer perhaps the greatest potential for revolutionizing our understanding of the factors that define Earth’s energy balance in the future.

SEE ALSO: Atmospheric/general circulation; Climate change, concept of; Clouds; Earth system science; Water budget

References


Ecogeography/macroecology (range and body size)

Marta A. Jarzyna
Brian A. Maurer
Michigan State University, USA

Patterns of biodiversity have intrigued scientists for centuries and, since the 1980s, there have been significant advances in describing and explaining species distributional patterns, their geographic ranges, and their statistical distributions. Historically, these issues have been approached by many scholars, resulting in the creation of several different subfields such as geographical ecology, areography, dynamic biogeography, ecography, and ecogeography, to name a few. While somewhat disparate in their focus, these different domains of interest largely center on the issues of what today is called “macroecology,” and frequently overlap. Macroecology is a nonexperimental, statistical investigation of patterns of abundance, distribution, and diversity of species in order to understand the processes and mechanisms controlling the structure and dynamics of regional and continental biotas (Brown and Maurer 1989).

The foundation for macroecology was provided by some genuinely transformative insights from the fields of ecology, evolutionary biology, and biogeography. Near the end of his career, R.H. MacArthur laid the groundwork for macroecology by discussing the mechanisms underlying species distributional patterns across geographic space. The equilibrium theory of island biogeography, developed independently, first by Munroe (1948) and then by MacArthur and Wilson (1976), provided the starting point for the theory of the assembly of discrete biotic communities. Later, Rapoport (1982) developed a comparative approach to geographic ranges, while Price (2003), among others, examined patterns of the distribution and abundance of organisms in light of evolutionary processes.

The intent of this entry is to examine prominent macroecological patterns found in nature and to discuss the processes and mechanisms that act as primary drivers of these patterns. Species–area relationships, latitudinal diversity gradients, range size and abundance relationships, and patterns in the distribution of body size among species are some of the most general and widely studied patterns of biological diversity.

Biodiversity patterns

Species–area relationship (SAR) is one of the best-known generalizations in ecology. In its simplest form, SAR states that larger regions tend to contain more species than smaller regions (Figure 1). The study of SAR dates back to at least the mid-nineteenth century and the work of H.C. Watson, whose species–area curve of vascular plants in Great Britain was the first known illustration to be published of the species–area relationship. Currently, SAR is probably the most widely researched pattern in macroecology, encompassing thousands of studies of a wide variety of taxa and geographic extents. Despite the fact that many mathematical functions can estimate SAR, the power law (and its log-linear transformation) remains the most commonly used model both for fitting species–area curves.
ECOGEOGRAPHY/MACROECOLOGY (RANGE AND BODY SIZE)

and for the development of mechanistic theories of species diversity. It has been observed, however, that the log-linear SAR is not maintained across all spatial scales. For instance, the log-linear relationship can often be obtained only for the intermediate (i.e., regional) scales, but it does not hold for small or very large (i.e., continental or intercontinental) scales, where the increase in the number of species with area is much faster. Such triphasic SAR implies that potentially different mechanisms underlie the species–area relationship at different spatial scales. For very large areas, the steep SAR is likely a result of evolutionary dynamics, while smaller regions will lack species that are ecologically similar and evolutionarily distant, thus resulting in a flatter SAR. At very small scales, the increase in species richness is faster again because of the higher likelihood of encountering new habitats, and thus new species, as the area increases.

Latitudinal diversity gradient (LDG), another strikingly general pattern in macroecology, refers to the increase in species richness that occurs along the latitudinal gradient from the poles toward the equator. In other words, regions at lower latitudes generally contain more species than locations at higher latitudes. The LDG seems to hold for different spatial extents, habitat types, and various taxonomic groups (e.g., marine and terrestrial groups, ectothermic and endothermic animals). Yet, despite the ubiquity of the pattern, a universal explanation for the latitudinal diversity gradient has yet to be found.

The statistical distribution of abundance among species within a taxon has also received substantial attention in the macroecological literature. Comprehensive datasets on the relative abundances of species indicate that most species abundance distributions have long and positively skewed tails, indicating that communities usually consist of many rare species and only a few common species. Mathematical functions proposed to describe this species abundance pattern include Fisher’s log series and the most frequently used Preston’s log-normal, though recent theoretical work suggests that the log-normal distribution fails to properly account for the asymmetric shape of many species abundance distributions (Hubbell 2001).

Within a species, abundance tends to be the highest in the center of the species’ geographic ranges and declines steadily toward the margins of its range boundaries (e.g., Maurer 1999). It has been suggested that such pattern reflects the response of populations to local environmental conditions. Environmental conditions tend to be spatially autocorrelated (i.e., locations in close proximity tend to have more similar abiotic and
Thus, the abundance will also vary across space from high values, where the conditions are the most appropriate, to zero, where conditions preclude survival and reproduction. From a multispecies perspective, the abundant species tend to have broader geographic ranges, whereas the rare ones tend to be narrowly distributed. This pattern is the most apparent for closely related or ecologically similar species, pointing to the importance of both evolutionary processes and ecological constraints as mechanisms shaping macroecological patterns.

Another common pattern in macroecology is the tendency of species’ geographic ranges to increase with increasing latitude, formally known as the Rapoport’s rule (Rapoport 1982). Unlike the previously discussed patterns, the generality of the Rapoport’s rule has been much debated, with some studies finding convincing empirical evidence for the pattern, while others argue that it is a local rather than a global phenomenon.

In addition to patterns of species distributions, abundance, and range, macroecology also concerns itself with distributions of body size. Quantifying patterns of body size is of utmost importance because practically every aspect of the ecology of individual organisms (e.g., metabolic rates, life expectancy, range size) is related to body mass (e.g., Maurer 1999). The frequency distributions of body masses of North American terrestrial mammals vary with the spatial scale. At the largest spatial extents, body mass distributions tend to be positively skewed (i.e., medium-sized species are the most common, while there is a long right tail containing the largest species), whereas this distribution flattens with decreasing spatial scale. At local extents, the species–body size distributions are on average statistically indistinguishable from log-uniform distributions. Similar patterns have been found for other groups of organisms (e.g., Maurer 1999).

### Processes and mechanisms

Patterns of species diversity arise from mechanisms that tie evolutionary adaptations of species with the variation in environmental conditions. Because of the complex interaction of these different processes, explanations of specific biodiversity patterns continue to be conceptually challenging. However, several mechanisms have been proposed to explain the patterns discussed, and these include: (i) ecological niche theory, (ii) ecological neutral theories, (iii) metabolic theory of ecology, and (iv) phylogenetic constraints.

Ecological explanations of species distributional patterns are based on the assumption that species are limited by a unique set of environmental and biotic factors, which taken together constitute a species’ ecological niche. This ecological niche theory captures the uniqueness of each species. In contrast, ecological neutral theories consider communities to be open, continuously changing, nonequilibrium assemblages of species, in which species presence, absence, and abundance are governed by random speciation, dispersal, ecological drift, and extinction (Hubbell 2001). Neutrality means that all species are ecologically symmetrical, that is, they have identical ecological properties. Several neutral theories have been proposed to date, including MacArthur and Wilson’s theory of island biogeography and Hubbell’s unified neutral theory of biodiversity and biogeography (MacArthur and Wilson 1976; Hubbell 2001). While controversial, neutral theories have been successful in predicting some of the most fundamental patterns of biodiversity. For example, Hubbell’s neutral theory is able to successfully generate species abundance distributions as well as explain the triphasic species–area relationship as consequences of the speciation–dispersal–extinction
process rather than a manifestation of ecological niche (Hubbell 2001).

Distribution of body size among species (i.e., intermediate-sized species are the most common) has in the past been explained by invoking physical constraints and energy requirements, both of which increase as species depart from the medium body size (Maurer 2003). This notion has served as a precursor to the metabolic theory of ecology (Brown et al. 2004), which links the variation in the rates of metabolism between different kinds of organisms to the structure and dynamics of populations, communities, and ecosystems. In addition to distributions of body size, the metabolic theory of ecology provides an explanation for other common biodiversity patterns. For example, the theory attributes the latitudinal diversity gradient to the kinetic effect of higher temperatures (e.g., Brown et al. 2004). The implications of this theory have yet to be fully realized, but may lead to new insights into the complex set of processes that underlie patterns of species diversity.

Recent work has cast the study of macroecological patterns in a macroevolutionary framework. The macroevolutionary perspective argues that evolved traits of different organisms (i.e., morphology, behavior, life history) influence their ecological relationships, and thus patterns of species distribution and abundance. For example, a phylogenetic constraints hypothesis (Price 2003) argues that certain sets of characters, common to closely related species, limit the major adaptive developments in a taxon, and thus its ecological options. Minor adaptations within the phylogenetic constraint are allowed, and those have ecological consequences that are reflected in patterns of distribution, abundance, or population dynamics (Price 2003). Incorporating phylogenetic information into macroecological explanations synthesizes ecological and evolutionary concepts into a framework for understanding how the ecological properties of species evolve within higher taxonomic levels (Price 2003).

What is on the horizon for macroecology?

Macroecology has provided a synthesis of perspectives from ecology, evolutionary biology, and biogeography that has helped generate new explanations for widespread patterns that large continental biotas exhibit. A great deal of effort has gone into generating data on a variety of groups of species, and there are opportunities to develop more rigorous general explanations for these patterns. The recent activity to develop explicit theories for macroecological phenomena using models that assume ecological symmetry between species is a promising start. However, these theories don’t go far enough, because it is unlikely that species can be assumed to be ecologically equivalent. The evolutionary adaptations that drive population dynamics of different species across geographic space are ultimately responsible for ecological asymmetries between species. For example, closely related species often have different geographic range sizes that are directly tied to how their different ecological adaptations interface with available resources across geographic space. Theories based on ecological symmetry may explain how the “average” species behaves dynamically, but are ultimately unlikely to explain the range of macroecological attributes found within real groups of species.

The incorporation of evolutionary information via well-supported phylogenies can help with the development of more general theories for macroecological variation within species assemblages, but this is only part of what is needed. One of the strengths of the metabolic
A key approach to macroecology is that it has the potential to link fundamental design constraints imposed on the evolution of attributes of species with the ecological consequences that those design constraints impose on species assemblages across space. These insights may yet provide the basis for generating more general theories that can be used to understand the growing body of data accumulating in studies of assemblages of species at continental scales.

**SEE ALSO:** Biodiversity; Biogeography; Geography of evolution; Island biogeography

**References**


Further reading


Ecological footprint

Andrea Collins
Cardiff University, UK

Ecological footprint analysis was initiated in North America in the early 1990s by William Rees and Mathis Wackernagel (see Wackernagel and Rees 1996). The ecological footprint is an aggregated indicator which provides a measure of the global ecological impact of human resource consumption, and is analogous to gross domestic product (GDP) as a representation of the dimensions of the financial economy (see Wackernagel and Rees 1996). The starting point for the concept is that there is a limited amount of bioproductive land on the Earth to provide for all human resource demands. Sustainable development requires living within the carrying capacity of the Earth, allowing our economies to develop, but also ensuring that human needs are being met.

The ecological footprint adds together the uses of six types of bioproductive land (carbon, grazing, forest, fishing, cropland, and built-up land) required to provide the resources consumed by a reference population and assimilate their waste. The ecological footprint’s unit of analysis is the global hectare (gha) and is usually expressed in global hectares per capita for a given population (gha/capita). The measure is based on a strong sustainability position, as it considers the ecological limits to resource consumption and economic growth. By comparing the area of land required to support a certain level of consumption with what is currently available, the ecological footprint is considered a valuable approach for assessing whether or not consumption is ecologically sustainable.

The ecological footprint has been advocated by the World Wildlife Fund for Nature (WWF), and since 2000 has published through its biennial Living Planet Report the “state of the environment.” The 2012 edition reported that, in 2008, the Earth’s available biocapacity was 1.8 gha/capita. In comparison, the ecological footprint was 2.7 gha/capita and so exceeded the Earth’s available biocapacity by 50 percent, thereby indicating a situation of ecological “overshoot” as consumption patterns were not sustainable and resources were being used at a rate faster than they could be generated (WWF 2012). Over- shoot may reduce the Earth’s ecological capacity permanently and this is a key concern for sustainability.

Traditionally, national ecological footprint accounts have been calculated based on a nation’s domestic production and imports and exports of primary and secondary products. Recently, however, Wiedmann et al. (2006) developed a method whereby national ecological footprints can be disaggregated to provide information by industry sector, final consumption categories, subnational area and socioeconomic groups. This approach uses input–output tables (supply and use tables) to allocate footprints to consumption categories. The strengths of this methodological development is that it permits greater consistency across footprint estimates; the footprint results can be used to provide better-quality information for scenario development and policy formation on sustainable development patterns; footprint results can be allocated to different final consumption categories and so can highlight consumer responsibilities more clearly;

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0018
and footprints can be more easily estimated at subnational level.

The first academic article on the ecological footprint concept was published in 1992 (see Rees 1992). Since then the concept has spread rapidly around the world. At the end of 2013, Google delivered almost 3 million hits and Google Scholar 73,500 hits for the term “ecological footprint.” As a source of information on environmental impacts the footprint has traveled relatively quickly, and, within the space of 20 years, studies have related to five world continents. The rapid transfer of environmental ideas on resource use and its impact, and growth in the application of the ecological footprint, have followed a similar patterns to how ideas on environmental impact assessment spread from the United States in the late 1960s to other developed countries such as Canada, the United Kingdom, and Europe in the 1970s and 1980s, and then to developing countries such as Africa and South America in the 1990s. There are three main reasons for the level of interest generated by the ecological footprint. The first is its contribution as a method for measuring biocapacity and assessing the environmental impacts of natural resource use. Second, it is often considered as providing an innovative approach to communicating messages about the ecological impact of current resource consumption. Finally, it has potential value as a policy tool and link to practice.

Despite the ecological footprint’s popularity as a measure of ecological impact, it has received a number of criticisms: it does not reflect the full impacts of consumption activity; it provides inadequate material for use in policy formulation, although improvements led by Wiedmann et al. (2006) to link footprints with industrial sectors and final consumption categories have gone some way to address this problem; the algorithms used to link production and consumption with notional global land areas may not be appropriate or sufficiently transparent; the method does not allocate the responsibilities for negative impacts correctly; in terms of deriving policy foresight, there is a strong underlying assumption that an equal distribution of global resources is a desirable outcome; and, finally, few studies have attempted to deal with the underlying uncertainty surrounding footprint estimates.

Although the ecological footprint has received criticism, it has continued to be popular and has received a wide range of interesting and novel applications including individuals, households, products, services, organizations, educational institutions, energy and fuel, food and agriculture, tourism, travel, and major events. These academic applications have helped to enrich debates on the environmental impact of resource use and also to provide alternative perspectives and a more comprehensive assessment of the impacts of resource use. However, in more recent years, with increased policy attention being paid to resource limit issues such as peak oil and peak food, and debates around climate change, the ecological footprint has been increasingly sidelined by the carbon footprint owing to the latter’s direct relevance to policy debates on climate change.

**SEE ALSO:** Consumption; Environment and consumption; Environment and development; Environmental impact assessment; Environmental issues and public understanding; Environmental policy; Resources and development; Sustainable development

**References**


The term “ecological imperialism” comes from Alfred Crosby’s book of that title. The book was hailed by some as a landmark text in environmental history when first published in 1986. It followed Crosby’s earlier, rather more discursive exploration of an equally big idea, The Columbian Exchange (1972). In that volume, he had discussed the movement of biota between Europe and the Americas in the centuries after Columbus. Ecological Imperialism represents a conceptual shift from exchange to empire. It sought to address the question of why Europeans and their descendants are found “all over the place” and to ask, in characteristically vivid prose, what “is the reason that the sun never sets on the empire of the dandelion” (Crosby 1986, 7)?

Many years later, Ecological Imperialism reads as a fairly conventional Eurocentric history, and one that did not actively engage with work in other disciplines. Nonetheless it contained a seductively constructed argument, and was innovative in its broad sweep, global focus, and introduction of the role of biological agents in the imperial incorporation of what Crosby called the “neo-Europes,” that is those lands thousands of kilometers from Europe but with majority European populations. As a result, the term ecological imperialism has entered the academic lexicon, but more often than not without critical assessment. It deserves scrutiny both in its own terms and from the perspective of later work by a range of scholars, including geographers, whose insights question important aspects of Crosby’s argument.

Precursors

People have long carried biota around the world for sustenance, trade, and settlement. Captain Cook’s expeditions landed pigs, goats, wheat, potatoes, and many other European vegetable and root crops on islands in the Pacific as well as accidentally allowing ship rats to escape. The consequences of lack of intentionality were often noticed. A famous example was Charles Darwin’s observation of “the invasion of the cardoon,” or artichoke thistle (Cynara cardunculus), on both sides of the Andean cordillera, in Chile, as well as the pampas of Uruguay and Argentina (Darwin 1839, 151). Brought to the New World as part of the Columbian exchange in the late eighteenth century, cardoon was a fairly popular vegetable in European gardens. In their new habitat, the plants quickly escaped cultivation and reverted back to their wilder and spinier forms.

The language of “invasion” permeates the various interpretative literatures that have developed since Darwin’s time to explain the sort of phenomenon that he observed in South America. To ecologists and biogeographers, invasive species are taxa that have or are spreading rapidly into habitats and places in which they were previously absent. Normally these are alien species introduced from another place or part of the world, although many introduced taxa never become invasive, just as there are native species that invade, that is, expand their distributions as happens in response to environmental change (Blumler 2011). The first full-length treatment
of these processes was Charles Elton’s book *The Ecology of Invasions by Animals and Plants*. He opened by proclaiming that “Nowadays we live in a very explosive world,” giving then topical examples such as the influenza virus, a fungus like potato disease, a plant such as prickly pear, and an animal like the gray squirrel (Elton 1958, 15).

Elton’s argument was that in the mid-twentieth century “the mingling of thousands of kinds of organisms from different parts of the world [was] setting up terrific dislocations in nature” (1958, 18). He used a series of well-mapped case studies, such as the spread of chestnut blight, *Endothia parasitica*, brought from Asia on nursery plants that progressively infected American chestnut trees in the first half of the twentieth century, and of the European starling which in about sixty years established itself across much of the United States from an initial stock introduced into New York’s Central Park in 1891. The book illustrated a range of “successful species” in different geographical contexts, such as continents, islands, and the sea. But he argued that “there are enormously more invasions that never happen,” that is introductions that never become well established or spread (Elton 1958, 109). He addressed the question of why this might be, notably with the hypothesis, never since clearly verified, that species-rich communities will be more resistant to invasion than less diverse ones.

Elton’s ecology was strongly humanized, both in his concern for conservation of richly varied landscapes for economic and social reasons, but also in its portrayal of the human context of species explosions due to trade and travel. Even so, he was criticized for assuming that humans will eventually take all species everywhere, while there has since been much less focus on “the actual historical geography of invasion” (Blumler 2011, 515). There was, however, an American geographical tradition of research into the intentional spread of agricultural species, in the New World and Old, led by Carl Sauer at Berkeley, whose best known work was *Agricultural Origins and Dispersals* (1952). One of Sauer’s doctoral students, Andrew Hill Clark, undertook his field research in New Zealand, publishing the results as *The Invasion of New Zealand by People, Plants and Animals* in 1949. New Zealand was also to be used as a case study by both Elton and Crosby.

Clark’s historical geography was concerned very much with intentionality. He described it as focusing “on a revolutionary change in the character of a region,” with one of the key factors in this being “the invasion of the area by armies of plants and animals, which, with the help of man, mingled with or displaced the native flora and fauna” (1949, v). Unlike Elton, he was mainly interested in the settlement of people and their agricultural plants and animals, such as potatoes, wheat, grasses, clovers, sheep, and cattle. But he did not ignore animals that had become destructive of farmed as well as natural landscapes, such as rabbits and deer, nor invasive plants such as gorse and blackberry, the seeds of which were originally sold for hedging and fruit respectively. *Invasion* was intended to be the first of a series of studies “dealing with similar problems of the development of patterns and practices of land use in mid-latitude areas overseas which were settled by folk from the shores of the North Sea” (Clark 1949, vi). Although Clark subsequently published two volumes on Atlantic Canada, others projected for Australia and South Africa were never completed. He did not use the term neo-Europes but in his systematic interest in those parts of the world he was an intellectual forerunner of Crosby.

**Crosby’s ecological imperialism**

Alfred Crosby retired from the University of Texas at Austin in 1999 as professor emeritus
of geography, history, and American studies. By his own account, he was a radical in his early career, being involved in the civil rights movement, teaching black studies and leading anti-Vietnam War demonstrations. It was this background that shaped his interest in the histories of peoples who were invaded and exploited by European imperialism. It did not, however, lead to a clear political economy framework for the argument developed in *Ecological Imperialism*, even though the book begins with a quote, among others, from Marx and Engels’s *Communist Manifesto*. The book therefore does not connect with work in human geography at the time, and it also neither discusses nor references the work of Elton or Clark. It is best portrayed as standing to one side of other literatures, although in spirit it displays rather more of an interdisciplinary inclination than is usually the case in historical writing. It was republished in an unchanged second edition in 2004, its appeal underlined by translations over the years into nine European languages, including Turkish, as well as into Chinese, Japanese, and Korean.

This appeal reflects the quality of Crosby’s writing and use of metaphor in a field that, as the work of his predecessors shows, is prone to this. But the key was the way in which he pulled together a range of evidence about the purposive spread of people, plants, animals, and diseases in the service of the development of the neo-Europes. His was a grand narrative that convincingly inserted the role of biological factors into an explanation of the effects of imperialism at the same time as it effectively sidestepped any encounter with the economic. And while colorfully portraying “invasions” as ecological, it did not attempt to engage with the complexities of ecological processes. Hence it embodies the characteristics of all grand narratives, being simultaneously plausible and partial.

Central to Crosby’s ecological imperialism was what he called the “portmanteau biota” with which the neo-Europes were colonized. This is a more analytical category than Clark’s invading armies, but serves a similar purpose. He traced it to the Old World Neolithic Revolution and the assemblages of people, plants, and animals — including those that people did not want, the weeds and vermin — that developed in close proximity, initially in Mesopotamia but then spreading through the Middle East and into Europe. Crucially, the synergistic effect of these different species existing side-by-side produced new diseases and variants of old ones. Then, continuing the military analogy, he analyzed the fate of “the full platoon of the Old World Neolithic” (Crosby 1986, 37) taken by the Norse to Iceland, Greenland and beyond, as well as by the Crusaders into the Levant. These early geographical lunges were unsuccessful: supply lines were tenuous, migrant numbers small, and environments unsuited for crop plants. Isolation rendered small Norse settlements prone to diseases evolving in Europe, and there were never enough Crusaders to acquire immunity to those such as malaria.

The core of Crosby’s analysis, however, concerned colonies at a greater distance, once Europeans began to discover the sea: with this the “seams of Pangaea were closing, drawn together by the sailmaker’s needle” (Crosby 1986, 131). He employed this analogy to describe how the continents and their biota, isolated in deep geological time by plate tectonic boundaries, were knitted together with the new technologies of shipping and navigation. In a detailed analysis of the impacts of European settlement, cropping and diseases in “the Fortunate Isles,” the Azores, Madeira, and Canary Islands, he argued that these “were the laboratories, the pilot programs, for the new European imperialism” (Crosby 1986, 100). The lessons learned were, first, that
simplified versions of the portmanteau biota could flourish—wheat, pigs, goats, honeybees, sugarcane too—and, second, that indigenous peoples could be subjugated, as in the Canaries by epidemics of European disease and the use of the horse in contests for territory.

Crosby concluded that the essential conditions for successful European colonies of settlement were that, first, land and climatic conditions had to be similar to some part of Europe; second, remoteness from Europe would mean few predators or diseases adapted to prey on Europeans; and, third, such remoteness would also ensure that indigenous inhabitants would be without defenses to introduced diseases and lack “servant species” such as horses or cattle. The tropics did not meet the second condition, and the lands of East Asia met neither the second nor third and were too well populated and organized. The neo-Europes on the other hand are all in similar climatic zones to Europe, but with biota that had evolved separately on the other side of the “seams of Pangaea.” Drawing on work about waves of extinctions of large land animals and birds that coincided with the first movements of people into the Americas and Australasia in pre-history, Crosby suggested that when Europeans followed millennia later, they found vacant ecological niches ready to occupy with their tame and feral animals. Often their livestock and plant pathogens were slower to arrive, if at all, but epidemics of introduced human diseases had dramatic effects on indigenous peoples.

This led to one of the key elements of Crosby’s argument: that the members of the portmanteau biota “did not function alone, but as a team” (Crosby 1986, 287). He used as an example the co-evolution of pasture grasses and livestock in Europe, such that when native grasses in the neo-Europes, often unaccustomed to heavy grazing, were slow to recover, Old World weeds and forage crops spread quickly. Newly “Europeanized” landscapes, being constantly disturbed and remade, provided environments in which newcomers could thrive. Hence the profusion of Old World cardoon in the heavily stocked grasslands of South America that Darwin saw.

Critique

Crosby’s thesis was a stimulating addition to understanding the nature of European expansionism but said little about other forms of imperialism, or how the ecological might intersect with cultural, political, or, indeed, economic imperialisms. His discussion of the dynamics of imperialism was limited and questions of motivation were glossed over. Expansionism to him was the product of five factors, the first of which, left undeveloped, “was simply the emergence of a strong desire to undertake imperialistic adventures overseas” (Crosby 1986, 105). The other four were technical: suitable vessels, way-finding methods, weaponry, and energy. Analytically this is close to technological and ecological determinism, and does not explain a lot, as Crosby himself implied when he wrote of “the puzzle of the rise of the Neo-Europes” (1986, 270).

The book did, however, conclude with the observation that the Americas and Australasia “have provided windfall advantages to humanity twice, once in the Paleolithic and again in the last half millennium” (Crosby 1986, 307). Here there was a hint of a political economy of imperialism, the second set of “windfall advantages” representing the expropriation of surplus from the colonial periphery as recognized by Marx. This was the product of “accumulation by dispossession,” driven by enclosure of lands removed from their indigenous owners. A system of individualized land titles and tenure provided the spatial and social building blocks of new colonial
ECONOMIC IMPERIALISM

ecologies in the neo-Europes. Enclosure was enforced using technologies of surveying and fencing, allowing the effective management and containment of flows of people and animals alike. These technologies delineated private property rights in territory at the same time as excluding those dispossessed (Pawson and Christensen 2014). Fences also enabled the parceling of landscape for removal of indigenous land cover and its replacement by familiar crops to generate export of food and raw materials for metropolitan centers of capital. Rather than “ecological imperialism,” it would be more convincing to assess Crosby’s contribution as highlighting some ecological components of imperialism, which in turn might situate environmental history more strongly in the mainstream of critical scholarship.

If explanation of European imperialism was a weakness in Crosby’s argument, so too was the portrayal of the other. With qualification, this applies to both his representation of people and biota. Crosby mused that maybe “European humans have triumphed because of their superiority in arms, organization, and fanaticism” (1986, 7). While the discussion of the effects of Old World diseases on the indigenous peoples is persuasive, although hardly novel, he came close to portraying those peoples as passive victims, rather than active contestants of processes of colonization. His use of the term “Stone Age” marginalized the cultural attributes of Native Americans, Australian Aboriginals, and Māori as shapers of landscape as hunters, fishers, and gardeners. Nor were their histories of resistance, now clear from several decades of subaltern scholarship, recognized. This was perhaps inevitable given the manner in which the narrative was framed through the metaphor of “invasion.”

In some places, Darwin’s theories were read as supporting the apparent success of introduced plants and animals. The naturalist W.T.L. Travers wrote from Wellington to J.D. Hooker at Kew about how New Zealand’s indigenous plants “appear to shrink from competition with these more vigorous intruders” (in Crosby 1986, 255). Crosby’s position effectively endorsed this view, and it lay at the heart of his argument. Despite stating that there “is little or nothing intrinsically superior about Old World organisms compared to those of the neo-Europes,” he then wrote that “Old World organisms are almost always “superior” when the competition takes place in their home environment,” as well as “wherever colonial environments have been Europeanized.” “Europeanized” in this context meant in “a condition of continual disruption: of plowed fields, razed forests, overgrazed pastures, and burned prairies” (Crosby 1986, 291–292).

In this one passage, Crosby repeated two venerable but frequently disputed ideas in biogeography: first, the inherent competitive superiority of organisms sourced from Europe over those of other continents, and, second, that these “invaders” are more effectively disturbance-adapted than those from elsewhere. It is this position that underlay the asymmetry at the root of his understanding of European ecological imperialism. He asserted that only a very small number of weeds, animals, and pathogens from the neo-Europes have naturalized in Europe, quoting Hooker about “this total want of reciprocity in migration” (in Crosby 1986, 167). But this is a product of how the argument is framed and the evidence assembled: reference to Elton (1958), for example, would have produced a different picture. As an ecologist, Elton made no such spatial assumptions and produced many examples that counter any “want of reciprocity.” These include the devastating effects of the American vine aphid, *Phylloxera vitifolii*, on European vines, the impact of the Colorado beetle on the cultivated potato in Europe, and *Elminius modestus*, a barnacle from...
Australasia that by the 1940s had established itself as an intensive fouler of ships’ hulls along the coasts of southern England, northern France, Belgium, and the Netherlands.

Crosby’s argument was therefore ideologically freighted, in ways that might have been tempered had he taken account of pre-existing work in ecology. Instead he could claim that in a “Europeanized” Australia, “dandelions have a more secure future than kangaroos” (Crosby 1986, 292). This was consistent with his position: it is also not correct, given that the larger species of kangaroo have considerably increased in numbers within and beyond their range since European colonization (Flannery 2004). Behind that simple oppositional binary, of dandelion and kangaroo, lay another assumption. This is that it is possible to denote “a portmanteau biota” that is identifiably “European,” rather than a hybrid assemblage sourced from all the places with which Europeans and their networks were in contact. Yet Clark (1949), for example, had illustrated the colonial reach of Californian and Australian species, including the Monterey pine and Monterey cypress (Pinus radiata and Cupressus macrocarpa), as well as many eucalypts. It is therefore necessary to reconsider Crosby’s implicit geographies a little more deeply.

Networks and agents

The sort of spatial imagination that Crosby exhibited, of a dichotomy between metropolitan “core” and colonial “periphery,” has come under increasing challenge from postcolonial scholars, actor-network theorists, geographers, and historians (Pawson 2008). In their formulation, space (including imperial space) can be understood as a sphere of multiple trajectories, some stronger and more formal than others, and described using metaphors of webs or networks (Lester 2004). It is through these webs that projects such as the attempted “Europeanization” of overseas territories were mobilized, and the biotic materials upon which this depended were assembled and dispersed. The webs themselves are seen as connecting metropolitan, colonial, and trading sites in complex meshes, fragile yet dynamic, and influencing the ways in which those sites, or places, were continually reconstituted. From this perspective, flows of material, things, ideas, and practices go in multiple directions. This, although he did not use an explicit spatial imaginary, is consistent with Elton’s (1958) account of species movements.

In a world of webs and networks, the predominantly asymmetrical dynamic of a metropole that shapes “other” places is replaced by the potential for transformation generated by sites throughout the network. In other words – depending on the relation being examined – any number of places can function as metropoles, or act as sources of generative change. They can be simultaneously nodal in some of their relations and peripheral in others. Mid-nineteenth century Melbourne was a colonial city yet at the same time played metropolitan roles within Australasia. More broadly still, its botanic gardens acted as a site of dispersal of Australian plant species, notably many eucalypts and acacias, around the Pacific and Indian Oceans. With plants being shared across networks between far-flung places, this again questions the concept of a European “portmanteau biota.” For example, hundreds of plants were introduced into England from the seventeenth century, sourced from progressively more distant places such as the Americas, China, and Japan. Many of these had significant effects on its landscape, such as species of conifers.

A network perspective on biotic change echoes calls from biogeographers and ecologists to pay more attention to historical and contemporary geographies of dispersal through an emphasis
on trade routes and travel patterns (Blumler 2011; Olden, Lockwood, and Parr 2011). It also requires an understanding of the role of a range of agents, human and nonhuman, in activating and mediating movements along such routes. In this respect the nodal points that controlled flows of information and exchange, or centers of calculation, exercised the power to assemble, order, and disperse knowledge and material. The functions of botanic gardens and zoos as collectors, sharers, and acclimatizers of species have been widely discussed. Kew Gardens in London was “the pivot of a colonial complex of gardens in Calcutta, Jamaica, Singapore, St Vincent, and Mauritius that it orchestrated from its metropolitan vantage point” (Livingstone 2003, 172). But so too for their own distinct array of exchanges were gardens like those in Cape Town. The activities of commercial agents have been less well recognized, for example, nurseries sending plant collectors on long distance gathering trips, such as into the Himalayas from Europe, and seed companies, like Suttons, which claimed to “seed the world” with its tins of grass seed dispatched from Reading, outside London (Pawson 2008).

The landscape impacts of such extensive exchanges of biotic material across the world have been extremely complex. Even if the agents of change since Europe began to discover the sea have often been European, to describe the effects as “Europeanized” – as did Crosby – risks removing both historical and ecological complexity from the understanding of agency. The impacts of “competitive” alien pathogens and species in Europe have already been observed. Two of the most invasive plants in parts of Europe (and elsewhere too) are *Rhododendron ponticum*, native to Spain and parts of Asia, and Japanese knotweed, *Fallopia japonica*. Both were introduced as ornamental species. Many plants recognized at the time as weeds were, however, carried to the neo-Europes unintentionally, as impurities within cheap consignments of seed (although companies like Suttons prided themselves on minimizing such risks). The New Zealand landowner and naturalist, Herbert Guthrie-Smith, observed the consequences of using poor quality grass seed on his own property, and surmised that a lot of other weeds arrived adventitiously in the hooves and hides of stock and horses (Wynn 1997).

Then there is the problem of the kangaroo, which Crosby thought faced an insecure future. Since the European colonization of Australia many smaller species of kangaroo have been in steady decline, and some have become extinct. But the larger species (notably Eastern Greys and also Reds) have flourished in many places, sometimes excessively. The reasons are not straightforward. The provision of introduced pasture and waterholes for stock have long been given as reasons, but others include the shooting of dingoes by shepherds. For millennia Aboriginal peoples also regulated kangaroo populations by hunting, a form of check that diminished as Aboriginal numbers declined from disease and displacement (Flannery 2004).

**Reflections**

There is much to prompt reflection in Crosby’s grand narrative, even if his analysis of “ecological imperialism” is rather too straightforward ecologically, spatially, and historically. It retains some purchase with respect to the impact of pathogens transferred by Europeans. But even here the portrayal of asymmetry was overplayed with networks of flows being far more complex than his model allowed. This has been even more so for insects, plants, and animals. One consequence is that hybridized landscapes are now the norm almost everywhere, the product
of both human intentionality and ecological process with increasing homogenization of biotas, although non-natives may boost overall species richness in some regions (Olden, Lockwood, and Parr 2011). Questions arise as to what is “native” and what is “alien” and the extent to which it is productive to attempt to retain or recover some sort of imagined edenic state of landscape that pre-dates what Elton described as the worldwide “bombardment” of every place by introduced biota (1958, 29–30).

What are left are paradoxes, not certainties. Species identifiably associated with particular places often turn out to be recent arrivals from quite different sources. The emblematic eucalypts of coastal California are from Australia, the Torbay palm of southwest England is New Zealand’s *Cordyline australis*, and the tumbleweed that blows across the sets of American western movies came into South Dakota as a contaminant of flax introduced by Mennonite farmers from Russia. And the confidence that accompanied the introduction of so many species not only in the neo-Europes but also in Europe and other parts of the world to gain “windfall advantages” from the appropriation of nature has been replaced today by biosecurity measures designed to manage the risk of further invasions. Blacklists enforced at state borders attempt to exclude known pests; whitelists allow introduction of species considered unlikely to become invasive (Blumler 2011). But wherever species are introduced, the potential for dislocation as well as for evolutionary adaptation, as the history of biotic exchanges shows, renders such attempts at control problematic.

**SEE ALSO:** Actor-network theory; Alien and native species; Biosecurity; Environmental history; Historical geography; Imperialism; Landscape

---

**References**


Pawson, Eric, and Andreas Aagaard Christensen. 2014. “Landscapes of the Anthropocene: From dominion to dependence?” In Invasion Ecologies from the Environmental Humanities, edited by Jodi...
Frawley and Iain McCalman, 64–82. London: Routledge.

Ecological modernization

David Gibbs
University of Hull, UK

A growing recognition that the result of human impacts on the environment is a key challenge for global society has led to a number of responses from academics and policymakers alike. However, although there is broad agreement that the world is experiencing major environmental changes as a consequence, notably enhanced global warming and climatic change, there is much less agreement on how these problems should be tackled. There is a spectrum of opinions here, from that of deep green ecologists, who advocate wholesale restructuring of the economy and society, through to some economists who see the market as being capable of restoring environmental equilibrium. An approach to addressing environmental problems that takes a position between these two is ecological modernization. Unlike the more radical green approaches, ecological modernization suggests that the ecological crisis can be reconciled with current forms of economic development and institutions to form a new mode of development for capitalist economies. However, in some versions it also sees a greater role for purposive intervention than just relying on market forces.

Ecological modernization privileges the role of technology, especially the role of innovation, and market dynamics as drivers for change. While they may have been the source of past environmental degradation, it is argued that, when directed toward solving environmental problems, innovation and markets can be combined with the modernization of political institutions and processes to steer a change toward environmental (and economic) improvement. Competitiveness and economic growth are seen as compatible with environmental protection, with solutions to the major global problems such as climate change found in more modernization rather than in reduced industrial production or economic growth. Ecological modernization frames environmental problems in such a way that they can be solved politically, economically, and technologically in the context of existing institutions and power structures and of continued economic growth. Ecological modernization, at least in some forms, appears to allow us to have it all—not only does it deliver environmental improvement, but it also does so without seriously challenging existing economic practices (see Figure 1). Indeed, ecological modernization suggests that addressing environmental problems can produce win–win outcomes that increase firms’ profits as well as improving the environment. It is this seeming compatibility with mainstream economic activity, if only we could be forward thinking and modern enough to make this ecological shift, that helps to explain the appearance of ecological modernization in policy.

In policy terms, ecological modernization aims to move beyond “environment versus economy” debates whereby environmental regulation is seen as a hindrance to business. Rather, environmental regulation can be seen as a necessary catalyst for innovation and the development of new markets. The tradeoff thus turns into a mutually beneficial outcome where no one
ECOLOGICAL MODERNIZATION

Figure 1  Ecological modernization in action: solar roof panels on conventional housing.

loses, or in economists’ terms a “positive sum game.” In this form ecological modernization appears in the guise of a pragmatic program for business and government. In the absence of regulation and incentives, it is argued, companies will have little incentive to develop ecologically desirable (and profitable) innovations. By introducing environmental regulation, governments can encourage innovations and products that lead to high environmental standards and ensure the future competitiveness of their industries in global markets. As an approach, therefore, ecological modernization represents not just an academic and theoretical analysis, but also a practical guide to appropriate responses. These two approaches have also been characterized as the “analytical and descriptive” versus the “normative and prescriptive.” While ecological modernization may appear to offer hope for those committed to stronger environmental protection measures, critics have argued that ecological modernization can equally serve as a cover for business-as-usual with a slight green tinge, or “greenwash.” In this sense, the term could serve to legitimize the continued destruction of the environment and to foreground the industrial and technocratic discourses of modernity over more critical ecological ones.

Ecological modernization as theory

Ecological modernization has a relatively optimistic and constructive view of the potential for society to find solutions to environmental problems. The approach gives a central role to science, technology, and the state and proposes not only that economic development, social welfare, and environmental protection can be reconciled, but also that positive synergies for future development can be generated in the process. It focuses on technological solutions and eco-efficiency gains that can be introduced, and managed, with a minimal amount of change in existing institutional arrangements. One of the central arguments of ecological modernization is that our current institutions can be progressively transformed and modernized in order to avoid an ecological crisis.

The concept was first developed in the 1980s by Joseph Huber (1982) and Martin Jänicke
(1985). In Huber’s (1982, 1985) view, industrial society needed to experience a transition away from the current basis of industrial society toward a new relationship between economy and ecology to create a more ecologically rational organization of production. Huber called this transition an “ecological switchover” and, using a biological metaphor, suggested that in the process “the dirty and ugly industrial caterpillar will transform into an ecological butterfly” (Huber 1985, 20). An ecological modernization approach would involve structural change at both the macroeconomic level, through broad sectoral shifts in the economy, and the microeconomic level, for example, through the use of new and clean technologies by individual firms. As these shifts develop over time there will be a growth in eco-efficiency as the amount of raw material incorporated into finished products diminishes and waste streams are reduced, in terms of both their quantity and their toxicity. This view has a strong faith in the ability of the capacity of markets to deliver environmental technologies to solve major environmental problems.

In theoretical terms, the concept has been developed to analyze the necessary changes to the central institutions in modern society in order to solve the growing ecological crisis. Ecological modernization involves the transformation of institutions and social practices in industrialized societies, where the motivation to make this transformation comes from the growing recognition of environmental problems. Ecological modernization indicates the possibility of overcoming environmental crises without leaving the path of modernization. This use of the concept exemplifies Huber's ecological switchover whereby ecological modernization represents a major transformation of current industrial and economic systems into a form whereby its environmental impact is reduced and the resource base for future development maintained. However, ecological modernization does not assume that this will involve radical changes to our ways of living. The assumption is that processes of production and consumption can be restructured on ecological terms through the institutionalization of ecological aims. However, it is of key importance that environmental factors are not just taken into account in this restructuring process, but are considered as central to the future development of both production and consumption. The rationale for this is not to assert that ecological issues have primacy over economic rationales, but to give both equal weight in order to achieve the desired outcomes. At the heart of the ecological switchover concept are three central points:

- Production and consumption can be restructured toward ecological goals. This involves the development and diffusion of clean production technologies and decoupling economic development from its resource inputs, resource use, and subsequent emissions – that is, expanding economic growth is not accompanied by increases in either inputs or emissions.
- The “economization of nature,” through placing an economic value on nature, including the introduction of structural tax reform and a shift away from taxing environmental “goods” such as employment toward environmental “bads” such as pollution and carbon emissions.
- The integration of environmental policy goals into all other policy areas.

The development of ecological modernization theory

Ecological modernization theory emerged in a European context from the 1980s onward, in part as a practical application of the discourse
ECOLOGICAL MODERNIZATION

of sustainable development and in part as a counterweight to calls from radical ecologists for wholesale social and institutional change. From these initial formulations in the 1980s, a number of approaches have developed in ecological modernization theory (for overviews, see Mol and Spaargaren 2000; Spaargaren, Mol, and Buttel 2000; Mol and Sonnenfeld 2000). Three broad phases can be identified in the development of the theory. In the first phase, developed by Huber (1982, 1985) and Jänicke (1985), there was a heavy emphasis on the role of technological innovation, a critical attitude toward the state, and a belief in the power of market forces and actors to deliver change. Huber’s perspective was that ecological modernization offers a way out of ecological crisis through more industrialization, albeit with changed production and consumption. However, this view has been criticized for overemphasizing the industrial and technological aspects and neglecting the social context within which they occur. In Huber’s initial work, the ecological switchover is a logical, necessary, and inevitable stage in the development of the industrial system. From this perspective, technological developments are largely autonomously determined and lead to change in industrial systems and their relations with the social and natural environment. The key role of technology as a propulsive force in Huber’s work means that the state has a limited role in redirecting the processes of production and consumption. More generally, while a central argument of ecological modernization at this stage was that modern institutions can be restructured on ecological lines and away from a purely economic rationale, as a theory it had little to say on the extent to which institutions can be reformed – this was open to empirical investigation.

The second phase of work, from the late 1980s onward, placed less emphasis on technological determinism as a driving force, had a more balanced perspective on the role of state and market forces in the process of ecological modernization, and emphasized institutional and cultural dynamics. Hajer (1993) proposed two interpretations of ecological modernization. The first of these is Huber’s “techno-corporatist” interpretation, which emphasizes the “economization of nature” and elitist decision-making structures. A second interpretation shifts the focus from being simply on the necessary production and consumption changes, and extends it to encompass the need to incorporate greater democratization, a redistribution agenda, and ideas of social justice. Christoff (1996) characterized these two interpretations as “weak” and “strong” versions of ecological modernization (see Table 1). Drawing on Ulrike Beck’s work on risk society, Hajer further developed the second interpretation of strong ecological modernization as reflexive ecological modernization. In this formulation, political and economic development proceed on the basis of critical self-awareness involving greater public scrutiny and democratic control. Conversely, weak ecological modernization is seen as a lifeline for capitalist economies threatened by ecological crisis and unlikely to lead to the kinds of substantive change needed to avoid this.

A third phase of work on ecological modernization theory devoted more attention to consumption processes and attempted to deal with criticism that ecological modernization is a Eurocentric approach, given that the theory was developed in Germany and subsequent work focused largely on Germany, the Netherlands, and Scandinavian countries. Thus research has examined developments in a diverse range of nation-states including Australia, Brazil, China, South Africa, Russia, New Zealand, and the United States in attempts to see how far the theory can be used to explain a range of environmental policy developments in these countries.
Table 1 Characteristics of “weak” and “strong” ecological modernization.

<table>
<thead>
<tr>
<th>“Weak” ecological modernization</th>
<th>“Strong” ecological modernization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technological solutions to environmental problems</td>
<td>Broad changes to institutional and economic structure of society incorporating ecological concerns</td>
</tr>
<tr>
<td>Technocratic/corporatist styles of policymaking by scientific, economic, and political elites</td>
<td>Open, democratic decision-making with participation and involvement</td>
</tr>
<tr>
<td>Restricted to developed nations which use ecological modernization to consolidate their global economic advantages</td>
<td>Concerned with the international dimensions of the environment and development</td>
</tr>
<tr>
<td>A single, closed-ended framework on political and economic development</td>
<td>A more open-ended approach with no single view but multiple possibilities, with ecological modernization providing orientation</td>
</tr>
</tbody>
</table>

Source: Derived from Christoff (1996).

(see Scheinberg 2003; Jepson, Brannstrom, and de Souza 2005; Kotilainen et al. 2008; Milanez and Bührs 2008; Schlosberg and Rinfret 2008; Curran 2009; Howes et al. 2010; Lai et al. 2012; Long and Patel 2011; Memon, Kirk, and Selsky 2011). While this third phase still conceptualizes environmental problems as the key motivation for the introduction of reforms in social, technical, and economic spheres, there are some additional changes. In particular, nation-states shift toward more decentralized and consensual styles of governance and there is a greater incorporation of social movements which become more engaged with the state’s reform agenda and less involved in confrontation with the state. Such reforms should increase the involvement of nonstate actors via negotiations, market mechanisms, and dynamics. Thus recent approaches to ecological modernization reforms of environmental policy involve a shift away from a model of environmental change that is hierarchical and dictated by the state. In the same vein, the technological aspects of reforms to environmental policy are not limited to addressing individual products, emissions, or production process steps, but should also include higher levels of aggregation, production–consumption chains, and networks.

Ecological modernization and the role of the state

Most work on ecological modernization has focused on the potential for environmental reform at the scale of national governments, environmental movements, and business enterprises. While other approaches to environmental policy, such as sustainable development, envisage the “hollowing out” of environmental policy, with competencies moving up to the international scale and down to the local scale, ecological modernization does not necessarily require de-emphasizing the nation-state. In the initial formulation of ecological modernization by Huber, the state plays little or no role in the ecological switchover. Indeed state involvement was seen as hindering the development and diffusion of clean technologies. However, subsequent research has argued that this is an overly technical perspective and that it is difficult to imagine any moves toward ecological
ECOLOGICAL MODERNIZATION

modernization taking place without some state intervention at various levels.

This may not necessarily involve the need for a strong, bureaucratic state – the suggestion is that the role of the state should change from being reactive toward acting in a preventive manner, to move toward participatory policymaking methods, to become decentralized, and to steer developments contextually rather than direct them in a top-down fashion. Thus ecological modernization, it is argued, requires political commitment to a longer-term and more holistic approach to the integration of the environment with economic development rather than a dirigistic approach, where government has a strong directive role. Thus it is argued that those nation-states where we can see ecological modernization in operation (at least in its weak form) are those which have consensual forms of government and decision-making arrangements that encourage more collaborative relationships, such as the Netherlands, Germany, Sweden, and Japan. Despite this, Huber in his later work has argued that, despite significant progress in these countries, they remain in the early stages of ecological modernization and that further progress has been repeatedly delayed. There is thus a need to explain why progress has been delayed and to develop further conceptualizations of the state in ecological modernization theories.

A key area of work which has sought to explain the potential for nation-states to undergo ecological modernization has been undertaken by Martin Jänicke. Overall, Jänicke argues that Huber’s emphasis on corporate ecological modernization is necessary but not sufficient to lead to environmental improvement. He suggests that the ability of nation-states to undertake holistic and longer-term environmental policy approaches is important and that this results from the interplay of their economic performance with their capacities for innovation, strategic thinking, and consensual government. From this he develops the concept of “environmental capacity” in order to understand both the conditions that encourage nation-states to address environmental problems and those that can subsequently result in the implementation of successful policy forms. He then introduces a second concept of “strategic capacity” which refers to the ways environmental protection can become a function that goes across the whole administrative apparatus of government. This means that different government departments should work together, rather than in separate “silos,” so that potential disagreements and contradictions can be defused and dismantled.

Jänicke then builds on this theory of strategic capacity to produce a model of policy explanation. He argues that the capacity for action on environmental issues sets the structural conditions for successful environmental policy, as well as its limitations. Capacity in this case therefore refers to both the opportunities and the barriers that may arise. In this model, solutions to environmental problems are developed within structural framework conditions and within what Jänicke calls “situative contexts,” involving actors and their strategies, together with institutional, economic, and informational factors. Strategies are defined as the general approach taken toward a problem – for example, using environmental policies to address problems and achieve goals. Actors, who include individuals, pressure groups, and third parties as well as government, are the opponents and proponents of special issues. The latter have relatively stable general interests and core beliefs, and their capacity for action depends largely on their strengths and competencies. An important dimension to this model is that the ability of actors to develop strategies can be significantly influenced by the situative context, such as an economic recession or a
major environmental disaster or pollution event, which could hinder or encourage policy action. The other components of the model comprise the “structural framework conditions,” which provide the backdrop to the situative context, actors, and strategies. Together these form the broad conditions for environmental action and give rise to an opportunity structure for actors. This consists of:

- the cognitive-informational framework, that is, the conditions under which environmental knowledge is produced, distributed, interpreted, and applied;
- the political-institutional framework comprising the institutional and legal structures and institutionalized rule and norms in society;
- the economic-technological framework which includes economic performance, technology levels, and sectoral composition.

The resulting capacity of actors to act (and the success or otherwise of these actions) will be influenced by the interplay of these three frameworks in any given situation. Jänicke suggests that utilizing this model to understand the implementation of environmental policy helps us gain an understanding of how the potential for ecological modernization may vary across different nation-states and at other scales. The ability to successfully implement policies will depend on specific situative contexts and structural frameworks, as well as on the composition of local actors and forms of strategies. However, while in theory ecological modernization is thus useful in helping us to think through some of the changes that need to be made to economies through political programs, in practice most subsequent research has rarely addressed the social processes involved through notions of situative contexts and the constellation of actors at different spatial scales.

EcoLOGICAL MODERNIZATION

Ecological modernization as political program

Ecological modernization is also used to describe a more pragmatic political program to redirect environmental policymaking. As a political program, ecological modernization has three linked components which are related to the theoretical arguments:

- compensation for environmental damage and the use of environmental technologies to minimize the effects of expanding production and consumption on the environment;
- the introduction of policies to encourage a shift in production and consumption processes toward the use of clean technologies and greater use of economic valuation in environmental policies;
- the deindustrialization of economies and a shift toward small-scale units of production, as well as a closer link between production and consumption.

While perhaps few policymakers would recognize the term, approaches based on ecological modernization are increasingly being utilized within policy which seeks to integrate economic development and the environment. Proponents of ecological modernization have argued that the process of shifting from the first to the second of the points listed is already underway in countries like Japan, the Netherlands, Germany, Sweden, and Denmark as material flows have become delinked from economic flows, with a simultaneous decline in the use of natural resources and emissions. This has involved moves to close substance cycles and the chain from raw materials through to the production process and waste and recycling, conserving energy and improving the efficiency and utilization of renewable energy.
sources, and improving the quality of both production processes and products.

As a pragmatic political program, ecological modernization approaches suggest that these changes will engender support from private sector businesses, given that they can have beneficial financial outcomes—ecological modernization is a response to environmental issues that stresses win–win outcomes. It is claimed that business can gain advantages in a number of ways: first, through greater business efficiency as a consequence of reduced pollution and waste production; second, through the avoidance of future financial liabilities, such as the potential future cost of the cleaning up of contaminated land; third, the creation of a better work environment has positive impacts on recruitment and retention of the workforce; fourth, there is greater potential for business through increasing sales of “environmentally friendly” products and services; and, finally, there are opportunities for the sale of pollution prevention and pollution abatement technologies to other companies and utilities. In total, then, it could be argued that ecological modernization has a well-developed perspective on how to conceptualize economy–environment relationships and a set of policy prescriptions that, taken as a whole, can engender the development of a new economic trajectory. Concepts derived from work on ecological modernization have certainly become incorporated into the policy agenda, even if their actual implementation has been limited and they are drawn from the “weak” conceptions in Table 1.

**Ecological modernization: a critique**

In recent years environmentalism has become institutionalized—environmental agendas and objectives are part and parcel of most nation-states’ policies and major corporations devote space in their reports and on their websites to trumpet their contributions to sustainable development and corporate social responsibility. Although it is clear that institutional changes and environmental improvements have occurred and that the changes advocated by the proponents of ecological modernization are desirable, there are a number of problems with ecological modernization as a theory and as a policy program.

First, one criticism of ecological modernization as both theory and discourse is that, in its weak forms, it can help to legitimize an environmental policymaking culture that absolves private sector businesses and major corporations of their broader environmental responsibilities. This is perhaps not surprising given its emphasis on technological innovation, on creating markets for new goods and services, and on the role of large producers and retailers in developed economies as a force for change. Indeed, this may be one reason for its widespread popularity, and some sectors of business have seen substantial profits from improved environmental technologies and stricter global environmental regulation, for example in the emerging clean tech sectors. For the advanced capitalist nations, which are struggling to remain competitive against the rise of emerging economies in Asia, imposing strong environmental regulations which require high-tech solutions has provided not only a competitive advantage for their own domestic industries but also a strong export market for these same environmental technologies. As a counter to this, developing economies have been critical of such initiatives, seeing them as a new form of protectionism and as “environmental imperialism.”

Thus while ecological modernization may be predicated on the potential transformation of capitalist economies, it is also liable, as a discourse, to be co-opted by powerful economic
actors such that there is greater dominance of global resources by transnational industry, national governments, and “big science.” Certainly the ecological modernization discourse has been incorporated into both corporate reporting and the agendas of global economic institutions without much apparent change to everyday practices. The underlying conceptualizations of ecological modernization can therefore act to perpetuate the dominant forms of economic power. At the same time there is little recognition in ecological modernization approaches that some industries may find it impossible to innovate to reduce their impacts – it may be too expensive or technologically impossible to do so. In these cases, for example in the fossil fuel industries or the automotive industry, an alternative strategy may be to lobby for exemptions or to attempt to dilute environmental policies. There are thus limits to the extent to which changes can be made within the modernization framework – not all sectors may be able or willing to make the shift. Moreover, other developments may represent a shift away from ecological modernization processes. For example, the increasing use of fracking for gas, especially in the United States, appears to be creating a new model for energy production, albeit aligned to the existing carbon-based system. The environmental impacts of this, through both increased carbon emissions and more localized concerns over water table pollution, are considerable and controversial (see Figure 2).

Second, despite a focus on the institutional reorganization of society, ecological modernization approaches often contain little detailed analysis of the forms of institutional adaptation or change that are required. It can be argued that this is a product of the undertheorization of the state in ecological modernization and oversimplifying assumptions about the role of states in ecological transformations. As was pointed out earlier, in Huber’s work the role of the (central, regional, or local) state was seen as minimal. In later accounts, there has been a more sophisticated understanding that the state performs an enabling and contextually steering role. Often, though, the appearance of institutional changes is presented as evidence that “ecological rationality” has started to challenge “economic rationality.” Although Jänicke’s concept of the “environmental capacity” of states focused on their ability to make the shift to ecological modernization, and others have argued that the integration of social and environmental movements into the policymaking process is driving the process forward, there is still little detail of the form of institutional adaptation or change required at the nation-state level, let alone at subnational scales.

One argument is that ecological modernization involves the development of an “enabling state” as the institutional response that will secure the efficient functioning of the market economy, but within a clear framework of state regulation. In this view, the enabling state will deliver ecological modernization through corporatist relationships between government and industry, although co-opting environmental movements when necessary. Certainly, many accounts have recognized the continued importance of the nation-state, albeit one that is no longer solely responsible (if it ever was) for environmental governance. This failure to adequately conceptualize the state and the social processes at work means that the type of embedded cultural transformations that will sustain factors, such as environmental improvements, reduced consumption, and greater equity, are unlikely to be realized.

Third, ecological modernization lacks a theory of power relations. In some accounts, the assumption often appears to be that the logic of ecological modernization is so obvious (and
profitable for business) that its widespread adoption is simply a matter of time so long as a mix of international civil society and environmental nongovernmental organizations keep up the pressure. In a situation where nobody seems to oppose it and where the theory lacks any scope to encompass power relations, there is little perceived room for opposition, and ecological modernization is assumed to occur almost automatically. In other accounts, ecological modernization will occur only if there is sufficient societal, political, administrative, and organizational capacity. Even then, this needs to be accompanied by the presence of a range of variables, such as strong corporatist institutions, innovative legal and informational systems, and a proficient regulatory system.

However, the reasons why such changes should occur, let alone when and where, are rarely addressed. The implementation of ecological modernization, as with any policy, however, is about the exercise of political and economic power, and, although this is rarely made explicit, ecological modernization is a fundamentally political concept. Whether it can be successfully introduced depends on questions such as who is in control, who sets agendas, who allocates resources, who mediates disputes, and ultimately who sets the rules of the game. Ecological modernization is thus as much an ideological and political issue as it is an ecological and economic one.

While the proponents of ecological modernization are not necessarily unaware of these issues, the driving force behind the development of “environmental capacity” and institutional reorganization is rarely made explicit, although there is a strong emphasis on market-based measures and pressures from civil society. Similarly, while some commentators have argued that ecological modernization is not a seamless progression toward a better future and that “ecological subversion” may reverse the semipermanent character of the ecological

Figure 2  Anti-fracking protest in France.
Some critics have identified ecological modernization as being part of the wider move toward the neoliberalization of society and economic policy, such that real transformative change is impossible. Indeed, ecological modernization has been criticized by both neo-Marxist and deep green critics who argue that the current capitalist system is inherently unable to reform itself along ecological lines beyond a weak “light green” reform program of “greenwashing” and that more fundamental structural changes are needed in the economy and society. In the process, the logics of neoliberalism and ecological modernization are said to have depoliticized environmental policy through scientific and “econocentric” discourses to exclude the consideration of alternative strategies. Certainly advocates of ecological modernization have had little to say about the impact of neoliberal economic policies on the environment or about the ability of markets to resolve environmental crisis.

Such a critique would suggest that ecological modernization needs a more radical edge if it is to lead to any substantive changes. This would require addressing the key underlying processes at work, including the associated power structures, social relations, institutional configurations, discourses, and belief systems, that currently lead to environmental and social injustice. At the stronger end of ecological modernization, the incorporation of environmental justice would require considering alternative modes of production, consumption, and distribution to current forms if ecological modernization is to be a more radical program of political action. To some extent, this point is recognized in the work of Hajer (1995) and his “argumentative approach.” His approach includes politics in the analysis in the form of a struggle for discursive hegemony whereby the actors involved attempt to secure support for and to impose their particular definition of environmental policy. In this case, the discourse of ecological modernization, at least in its weak version, appears to be politically attractive as it provides a means to both accommodate radical environmental critics and to revive developed nation economies through a new round of capital accumulation based on environmental technologies and products. However, ecological modernization is not (yet, at least) a distinct social theory; to become so, it needs to adequately incorporate theories of politics and the state.

Ecological modernization in geographical uses

Some geographers have drawn implicitly on the concepts of ecological modernization. Soyez (2002), for example, used the term “eco-modernization” as a synonym for ecological modernization. More recently, Bailey and Wilson (2009) and Bailey, Gouldson, and Newell (2011) use ecological modernization as a way to frame debates around climate change. However, despite the growing importance of ecological modernization as both a theory and a pragmatic program, relatively few geographers have explicitly engaged in the ecological modernization debate to advance its theoretical development compared to those working in the field of environmental policy and politics (for an exception, see Murphy (2000)). However, strong parallels with ecological modernization can be seen in the work of Hayter and Le Heron (2002) and Hudson (2000, 2001). Hayter and Le Heron drew on research by Freeman (1992) on techno-economic paradigms (TEPs) to develop his suggestion that a “green paradigm” will form the basis of future economic development. This paradigm, they suggested, will involve both technological and institutional changes. It will revolve around the
ECOLOGICAL MODERNIZATION

dematerialization of the economy, the internalization of environmental values by industry, and the prioritization of the environment within research and development (R&D), as well as take-back strategies and a shift toward selling services rather than products. In Freeman’s original theorization, new TEPs arise when the economy is confronted by crises that cannot be solved by the existing TEPs. However, far from being a technological determinist argument, in a similar manner to ecological modernization, emphasis is placed on the development of a set of matching institutional forms, including business organization, labor relations, R&D structures, and international regulatory forms.

While Hayter and Le Heron (2002) did not use the term, there are strong parallels between their view of a green TEP and ecological modernization. They argue that in a green TEP environmental imperatives form the motivation for more widespread systemic change, and priorities for innovation are geared toward reducing energy and materials use of energy and materials in transportation, construction, and manufacturing systems. Similarly, Hudson (2000, 2001) developed the idea of “eco-Keynesianism” as a way of resolving the competing pressures of generating profits, providing work, and protecting the natural environment. He argued that this would be an attempt at radical reformism which combined environmental and social sustainability while respecting the profitability imperatives of a capitalist economy (Hudson 2001). Although his “sustainable eco-capitalism” has a much stronger emphasis on social justice, as with ecological modernization, it similarly involves the widespread use of clean technologies and environmentally friendly production.

SEE ALSO: Environment and the state; Environmental governance; Environmental management; Environmental policy; Green capitalism; Modernity; Neoliberalism and the environment

References


Economic development zones

Jinn-Yuh Hsu
Ling-I Chu
National Taiwan University

An economic development zone (EDZ) is, by definition, a special zone designed and established for the purpose of economic development. The establishment of EDZs is the process whereby the state assigns to specific pieces of land regimes of governing and regulation that are distinct from the normalized rules prevailing in the national territory (Ong 2006). Within EDZs, exceptional packages of incentives and preferential privileges are granted to induce investments. Compared with more traditional sector-oriented policies, these spatially specific strategies represent different state forms and powers to promote economic development. Normally, special zones are designed with the aim of achieving specific policy goals, including attracting foreign direct investment, alleviating unemployment, supporting economic reform strategies and industrial upgrading, and serving as experimental laboratories for the application of new policies and approaches (Farole and Akinci 2011). But, in practice, the performance of the zones is ambiguous, and needs to be examined case by case.

The history of EDZs

As shown by Easterling (2014), the historical roots of EDZs can be traced back to ancient free ports, medieval free cities, or maritime colonies. The Roman port of Delos, which is frequently cited as the primordial free port, as well as other Mediterranean free ports for trade, flourished in antagonistic Italian, Phoenician, and Muslim civilizations in the first millennium. In the Middle Ages, a network of “free cities” such as Hamburg, Bremen, Lübeck, and Cologne collaborated as the Hanseatic League to evade the jurisdictional power of monarchies, and later on to confront the sovereign authorities of rising nation-states. The European powers, including Portugal, Spain, Holland, France, and Britain on the one hand, extended their trading networks by establishing a series of colonies from the Middle East, Africa, and America to Far East Asia; the Chinese and Japanese emperors, for their part, developed sanctioned trading systems in Canton and Nagasaki respectively to cope with their encounter with the West. Some of these legacies, like Hamburg and Hong Kong, endured and were transformed into global hubs in the twentieth century.

It has been only in recent decades that the EPZ (export processing zone), a major kind of EDZ, has emerged as a powerful global form, even though its roots are ancient. In 1934 the implementation of the Foreign-Trade Zones Act in the United States marked the start of the new age of EDZs in their modern forms: freedom of trade was represented as a freedom from taxes. In the authorized zone areas, custom-free goods were sorted, refined, and processed, and then re-exported. From Ireland (Shannon Customs-Free Airport) to Mexico (the maquiladoras on the United States–Mexican borderland), this spatial technique was widely adapted in the postwar era. At the same time,
ECONOMIC DEVELOPMENT ZONES

More tax-exempt measures were adopted to attract manufacturing and service industries. To fire up development, latecomer states in particular elbowed their way into the global economy with EPZs, in which foreign capital was welcomed to take advantage of cheap labor and land.

At present, EDZs can be found all over the world, and the number has grown rapidly. For example, in 1986 the International Labor Organization’s (ILO) database of SEZs (special economic zones) reported 176 zones in 47 countries; by 2006, this number had risen to 3500 zones in 130 countries (Boyenge 2007). Among these newly established zones, most are located in developmental states. Despite the historical and current link between zones of exception and the ideology of free trade, they can also serve the interests of the developmental state, particularly in East Asia and Latin America.

The first East Asian EDZ was established in Kaohsiung, Taiwan, in 1965. Following it, Masan EDZ was set up in South Korea in 1970. As an agent of economic territorial nationalism, the East Asian developmental state was, on the one hand, inclined to make a distinction between domestic industries and foreign capital and, on the other, could not but take measures to connect the former with the latter in order to promote the national economy. Instead of being a passive administration that views its responsibility as simply providing a sound regulatory environment, the developmental state is interventionist and plays an active role in targeting strategic industrial development. Given these circumstances, EDZs became a crucial part of the developmental strategies that utilized their cheap labor and land and their loose regulation to attract foreign investment and transfer technology and international expertise, while maintaining the integrity of the national economy to a considerable degree.

Among the zoning programs in East Asia, China’s Special Economic Zones (SEZs) Policy is spectacular in comparison to its forerunners. Since the early 1980s, when post-Mao China decided to adopt an open policy, the state has kept on creating new capitalist spaces on the socialist mainland by establishing waves of SEZs. Deng Xiaoping considered the establishment of special zones as replicating new Hong Kongs along the Chinese border, acting as both an economic bridge and a political window to the outside world. Therefore, all the first four SEZs were located in border regions: Shenzhen adjacent to Hong Kong; Zhuhai across from Macao; Xiamen across the strait from Taiwan; and Shantou, which had strong traditional connections with Chinese communities in Southeast Asia. Once market development in the coastal SEZs gained momentum, different kinds of free trade zones, custom-free warehouses, export-processing and high-tech industrial parks were established in coastal and interior cities to attract foreign investment and accelerate the development of hinterlands.

EDZs in China have evolved into a unique system not only for export-oriented industrialization but also for transformation of the state itself. Ong (2006) argues that China’s opening and market reform policies relied not on unbundling or denationalizing sovereignty but on the production of new spaces of exception and border-crossing powers. According to Ong, the Chinese state’s flexible zoning practices create not only political spaces and conditions of variegated sovereignty aligned with an axis of trade, industrialization, and knowledge exchange, but also the regional space informally called Greater China in order to integrate disarticulated political entities economically as part of a path to eventual political integration.

Another key player in “zone fever” is India. The government of India has made various
efforts in order to promote waves of EDZs since
the 1960s. The Kandla Export Processing Zone
in Gujarat was established in 1965. This was fol-
lowed in the 1970s by the creation of the Santa
Cruz Electronics Export Zone (SEEZ), which
was expanded in the 1980s to include gems and
jewelry. The 1980s also saw the creation of EPZs
in Noida, Chennai, Cochin, and Falta (West
Bengal), followed by the Visakhapatnam EPZ in
1994.

But India’s push toward a more comprehen-
sive SEZ policy began in earnest following a
visit by the then commerce minister Murasoli
Maran to China in 2000. Hugely impressed by
what he had seen in China’s SEZs – and by his
discussions with Chinese officials – Maran acted
quickly to initiate a change in India’s policy
regime (Das 2009). This change took the form
of new SEZ rules notified in the Commerce
The new rules were a precursor of what would
later become the 2005 SEZ Act, which allowed
the private sector to assemble land and develop
the zones, unlike China. The main difference
between EPZs and SEZs is in their compre-
hensiveness. Whereas EPZs can be thought of
as industrial estates, SEZs typically contain a
full array of social facilities – housing, hospitals,
schools, and retail developments – that make up
a small city.

A spectacular characteristic of the new gen-
eration of EDZs, like the SEZ, is the rise of
zone-cities, or urbanization through special
zoning, particularly in China and India. One of
the most famous cases is Shenzhen, an “instant
city” which became a boomtown because of
its designation as an SEZ and saw sustained
rapid growth due to the extended implementa-
tion of favorable policies and the rapid and
continued build-up of state-financed infra-
structure (Chen and de’Medici 2010). These
special zones mostly turn a practical space for
the production of exports into connected places
of service provision and, most importantly,
urban infrastructures (Bach 2011; Easterling
2014). Free zones such as the EPZ, prior to the
last few decades, were usually fenced enclaves
for warehousing and manufacturing which
offered exemptions from customs or taxes. Yet
the zoning form that was largely relegated to
the backstage has recently taken a center-stage
position to become a primary feature of global
urbanism and a world city paradigm. Usually
located around harbors or airports, an SEZ
involves massive land development, and even
claims to be a city itself. The initiation and
operation of many SEZs are no longer solely
funded by government, as with EPZs, but are
jointly developed and managed with the private
sector in public–private partnerships. In practice,
the government provides the necessary planning
and construction, such as the infrastructure in
the harbor or airport areas, leaving the facilities
and particularly the urban infrastructure in the
SEZs and neighboring areas to the private sector.
Zone-cities are a key location for understanding
the social and cultural impact of globalization
on overlapping territories and urban space (Bach
2011), assuming the form of a contemporary
capitalist utopia and heterotopia, and of an
interface between geographies of management
and of imagination.

The main reasons for the wide adoption of
EDZ as a developmental strategy are that it pro-
vides a gateway to the international community,
lower operational costs, demonstration effects,
and, more importantly, a compromise between
liberal and protective regimes (Amirahmadi and
Wu 1995). EDZ creation may be driven by
political reasons. In countries with privileged
corporate groups which monopolize industry,
a separate economic zone rather than complete
liberalization may be more politically acceptable
to the elite, while also promising economic
mobility to the local masses (Amirahmadi and Wu 1995). Even though foreign investments and aid are welcome on condition that they create employment opportunities and stabilize the trade balance, the host states also have to find ways of ameliorating the negative impacts of liberalizing capital flows. Hence EDZ can be seen as a political-economic compromise involving the implementation of partial liberalization policies to protect the domestic economy so that it can avoid facing direct foreign competition in the early industrialization era. In order to find the right balance, the state experiments with zoning technologies, on the one hand, to exploit its cheap labor and land in attracting capital investment and, on the other hand, to avoid fully exposing the national economy to a relatively unstable internationalization process. As a result, the EDZ typically provides cocktails of exemptions from broader national rules that may include tax breaks, foreign ownership of property, streamlined customs arrangements, and deregulation of labor or environmental laws.

EDZs and liberalization

Heated debate about the role of the EDZs as free-market islands or bridges to structural transformation driven by economic liberalization always haunts developing countries. Globalizing processes intensify the spreading of EDZs. While a number of analysts believe that EDZs will provide their impoverished host country governments with a gradual, two-track alternative to neoliberal “shock therapy,” and therefore welcome their growing popularity (Rodrik 1999), others believe that they are second-best options to promote exports, in particular regions of developing countries, without having to undertake the challenge of broader structural adjustment. The real liberalization impacts of EDZs will depend on the particular national contexts, so there is no consensus between advocates of these perspectives (Schrank 2001).

The rapid growth of EDZ programs around the world since the 1990s, and their success in contributing to export-led growth in regions like East Asia, are due in part to the unprecedented globalization of trade and investment that has taken place since the 1970s and that accelerated during the 1990s and 2000s. Moreover, establishing linkages between local clusters of firms and global value chains (GVCs) has become a key development strategy for countries embarking on latecomer industrialization and seeking to engage in industrial upgrading (Humphrey and Schmitz 2002). In most cases, EDZs, and particularly EPZs, serve as the designated local industrial districts to host segments of global value chains. One of the major governance mechanisms is the statecraft of the developing countries. Meeting the challenge of global connection, the developmental states bring forth the modes of interscalar connection as a key dimension in exploring how economic and extra-economic regularities are materially, spatially, and temporally related. The state spatiality of EPZs can therefore be seen as special regulation spaces that cushion the tensions associated with global–local connections.

But participating in global value chains both enables and restricts the capability of the state to manage external connections, and produces different state spatialities that reshape the spatial forms and functions of EDZs. At first, the developmental states of developing countries strive to focuses on manufacturing roles, while product design and final markets are located in advanced countries. Since demand is often out of their control, it is even more imperative for latecomers to be “competitive” in factor supply. In this stage, the EPZs are the one of typical
spatial forms adopted by the state. However, in engaging more deeply in the global integration of economic activities, the state can no longer focus on the reduction of the cost of production factors, but has to provide a liberal and friendly business climate to attract high value-added investments. To fulfill the demands of liberalization, a special zone of a liberalized economy, such as South Korea’s Incheon Free Trade Zone, Taiwan’s Free Economic Pilot Zone, and China’s Shanghai Free Trade Zone, assumes the role of a bridgehead, connecting with and redirecting international economic flows. Since there are different types of global–local connections, the forms of zoning vary. EDZs, as a mobile technology, spread and evolve internationally in ways that correspond with distinctive geographical features of the industrialization process. The state can, albeit reluctantly, liberalize capital flows and legalize risky industries, such as those associated with financial derivatives, via EDZs, to create a friendly business climate and enhance competitiveness. With the establishment of EDZs, states are better positioned to act as agents of liberalization with a view to enhancing global competitiveness.

Given the state’s intention to mobilize zoning technology to cope with the pressures of globalization, the performance of the East Asian EDZs is ambivalent. It is widely recognized that the EDZs in South Korea and Taiwan, as well as in Malaysia to some extent, have performed relatively well in meeting the needs of industrialization, creating employment opportunities, and engendering technology transfer. Some EDZs have evolved into high value-added industrial estates. It is argued that domestic linkages between firms operating in EDZs and the local economy are crucial. These positive externalities could, but do not necessarily, spread locally. If the EDZs’ level of integration into the local economy is low, the spillover effects and externalities are likely to be limited, and the EDZ would be directly exposed to the potential volatility of global capital flows. For “footloose” transnational corporations (TNCs), substantial linkages may not develop because TNCs may prefer, for strategic reasons, not to become deeply embedded in the host country’s economy. Particularly for assembly-type work using inputs sent directly from the parent company, assembling in the zone and then shipping out, backward linkages with the domestic economy may not be generated. In some countries such as Madagascar, spillover effects and externalities are probably limited by the EDZs’ low level of integration into the local economy (duty-free imports act as import subsidies and prompt minimal local integration), the low skills of the labor employed, and the potential volatility of foreign investment (Cling, Razafindrakoto, and Roubaud 2005). Successful cases of national integration with the EDZs, such as South Korea, Taiwan, and now China, are cases where EDZ policy is accompanied by strategic state policy designed to encourage domestic firms to articulate with the GVCs, generating spinoffs from the TNCs in the EDZs. For instance, in the case of China, the state has a strong card in the shape of access to the huge domestic market to influence the TNCs in the zones to engage in technology transfer to local firms. Thus EDZs may be regarded as growth centers which are expected to generate economic multiplier effects, with employment and subcontracting work trickling down to surrounding regions.

Governing territory, borders, and sovereignty in EDZs

In contrast to the idea that EDZs are part of liberalizing processes and escape the control of the nation-state, regulation theorists such as Bob
ECONOMIC DEVELOPMENT ZONES

Jessop, Martin Jones, and Neil Brenner argue that EDZs actually reflect state accumulation strategies designed to strengthen the national economy. This line of research sees EDZs as a component of state strategies which represent initiatives to mobilize state institutions toward particular forms of socioeconomic intervention with a tendency to privilege particular social forces, interests, and actors over others. Jones (1997) showed that state institutions are endowed with distinctive spatial selectivities, which refer to processes of “spatial privileging and articulation” in which state policies are differentiated across territorial space in order to target particular geographical zones and scales. In this vein, they regard the partial insertion of the national economy into truncated GVCs as the main driver behind the formation of EDZs.

From a subtly different angle, Cameron and Palan (2004) argued that a bounded political economy is in fact a prerequisite for the regulation of all forms of international trade, and that international trade is a concept that would have no meaning if the world economy were not divided by national borders. In spite of being assembled to allow the diffusion of liberalization, the EDZ contrarily creates the “domestic” and the “international” economies as discrete spaces by visualizing the state border as if it were an economic boundary. The liberalization measures incorporated in EDZs reflect the state’s use of deregulation as a new means of reregulating economic flows, particularly finance and services. It is noteworthy that deregulation and reregulation processes are embedded in a series of social-political relationships of bordering, networking, and interscalar connections. Therefore the EDZ, one of the particular state forms, can also be seen as an arena where actors, including the state and firms, struggle over the construction of political-economic relationships, governance structures, and institutional rules and norms.

Another line of research, through a Foucauldian lens, considers the EDZ as a governmental technology in which market-driven strategies of spatial fragmentation respond to the demands of global capital for diverse categories of human capital, thus engendering a pattern of noncontiguous, differently administered spaces of “variegated sovereignty” or “graduated citizenship” (Ong 2006, 7). These scholars think of the differently administered spaces as the praxes of liberalization or neoliberalism that are put into effect as exceptions to the normal governmental rule that is supposed to prevail in the national territory. It should be noted that EDZs are used to reinforce rather than weaken national sovereignty, because the centralized state can define the general national interest by liberalizing part of its territory and people to articulate with international flows.

Since EDZs are further conceived as gateways to international economic flows, the discourse of inside/outside can be clearly defined. A wide-open door symbolizes a friendly business environment welcoming transnational corporations to establish a foreign–domestic nexus. The metaphor that the state can decide how wide the door should be opened also works well in symbolizing the capacity of filtering and selecting what is desirable and nondesirable. In manipulating zoning technologies, geo-economic hope arises in the discourse of (imagined) economic separation or interconnection. EDZs can be seen as a compromise between national security and economic development.

Zoning strategies always involve a balancing of geopolitical fears and geo-economic hopes (Sparke 2006). On the one hand, the zones can attract and retain liberal flows of capital, technology, and people to develop the national economy; on the other hand, these kinds of internally established borders bring about new security issues. Consequently, border control becomes
one of the key concerns with the setup of EDZs. An ideal pattern of border control will expedite the global economic flows to realize time economies, and at the same time filter out the “dirt” associated with unexpected or unwanted flows of goods and people and pre-emptively keep them outside the national territory (Sparke 2006). Advanced biometric technologies are utilized to monitor and govern the mobilities of people and goods; in so doing, states patrol and police the dividing lines between inside and outside to tackle the threats posed by external forces. Given that EDZs are regarded as the interface along the inside/outside line, they involve spatial practices of debordering and rebordering.

The contradictions of EDZ policies

Because zoning policy implies differentiating the treatment of land and people, it always generates social and spatial tensions. Some EDZs restructure the local economy in a way that is adverse for the poor, especially those in the agriculture sector. There is social outcry over land expropriation for the construction of EDZs and the resultant displacement of existing settlers in developing countries such as India and China, among others, as the displaced people consider the land that is essential for their livelihood or the compensation for land appropriation to be inadequate. In investigating India’s controversies over “land grabs,” Levien (2013) argued that, while the earlier models of company towns and industrial estates reflected a regime of land for production with pretensions of inclusive social transformation, the EDZ model nowadays represents a neoliberal regime of land for the market in which “land broker states” have emerged who indiscriminately transfer land from peasants to capitalist firms for real estate projects. The special zone policies in countries such as India often become the basis of a regime of “accumulation by dispossession” closely associated with the multiple forces seeking to turn land and other resources into capital. Levien (2013) argues that this type of real estate-driven agrarian transformation which the EDZ policy generates in the surrounding countryside will lead to land speculation that amplifies social inequalities in novel ways, marginalizes women, and creates an involuntary dynamic of agrarian change that is ultimately impoverishing for the rural poor.

As for the labor condition in EDZs, low wages are critical in attracting foreign capital, and employment is usually precarious or informal as a result of the EDZs’ setting aside of national labor regulations. People working inside EDZs are governed solely in relation to their potential for growth and productivity. Ong (2006) argues that these governance arrangements effectively suspend workers’ political rights, that individual workers themselves have an unclear political status, and that workers can be exposed to the worst abuses, including murder, with virtual impunity. However, Cross (2010) disagrees with the language of exceptionality by arguing that the EDZs should not be regarded as closed or bounded systems but should be structurally embedded in the wider economy. In other words, the labor condition in the zones is nothing more than the formalization of the informality that has existed outside the zones for a long time in industrializing countries, particularly in the Global South. Accordingly, the presumption of a Western system of political rights is not appropriate in explaining EDZs in the context of emerging economies where the political status of working subjects is never that of inalienable rights and entitlements. In this sense, the most significant achievement of the zones is that the new employment opportunities they provide render visible the insecure working condition in the wider economy.
The controversies about land appropriation, labor abuse, and environmental degradation never fade away in the construction of EDZs. In most cases, that the zones discriminate against certain groups of citizens, exempt capital from taxes and regulations, and force farmers to move out of their inherited land may in the end see the rise of various social movements. Consequently, states that are eager to carry out a liberalizing policy by spreading EDZs are very likely to be authoritarian regimes that are able to control social unrest. Certainly, in recent years, SEZ policy has been hotly debated. Those who stand to lose their land are organizing huge protests, and there have been scathing critiques of SEZs by politicians, scholars, media, and civil society (Aggarwal 2006).

### Conclusions

By and large, the EDZ started out as a state’s special and experimental strategy to explore the emerging opportunities of internationalization, and was gradually transformed into a major channel of global connections for the state. Situated in changing political and economic climates, two major trends in the institutionalization of EDZs can be noted. One is the “universalization” of EDZ exceptionalism across developing countries; the other is the utilization of special zones as a liberalization toolkit by developmental states. Rather than integrating EDZs into the general economy within national boundaries, states use them to offer special treatment as a way of exploring new windows of business and technology opportunities. Zoning technologies proliferate and extend to create a network of subcontracted parts covering much of the country. Ultimately, zoning became a part of the assemblage of globalization and has spread across the world (Easterling 2014; Ong 2006).

### References


### SEE ALSO:

Cities and development; Export processing zones; Free trade zone; Space of exception; States and development

---

**ECONOMIC DEVELOPMENT ZONES**

---

SEE ALSO: Cities and development; Export processing zones; Free trade zone; Space of exception; States and development

References


Economic geography: spatial interaction

Jiaoe Wang
Chinese Academy of Sciences

Spatial interactions consider the dynamics of flows of people, freight, services, energy, or information between locations generated by economic activities. They are demand–supply relationships expressed over a geographical space, and usually refer to a variety of movements such as tourism, commuting, migration, international trade, and the transmission of information or capital.

The consideration of spatial interactions has played a significant role in understanding regional development, particularly as the global mobility of people and goods has increased substantially since the 1990s. For instance, the total number of international tourist arrivals worldwide reached 940 million people in 2010, 2.2 times the 1990 level. Meanwhile, the volume of world merchandise exports reached 15,274 billion dollars in 2010, 4.4 times the 1990 level. The case of China is particularly illustrative as the mobility of people and freight sharply increased between 1990 and 2010. Passenger traffic volumes in 2010 (32.695 million persons) were 4.2 times the 1990 level, and freight traffic volumes (32.418 million tons) were 3.2 times the 1990 level.

The theoretical origins of spatial interactions are attributed to the American geographer E.L. Ullman (1957, 1980), who first formally defined the concept and proposed complementarity, transferability, and the lack of intervening opportunities as three interdependent conditions for a spatial interaction to occur.

Complementarity

Complementarity refers to a demand for or a deficit in a product in a place and a supply or surplus of the same product in another place. For instance, a city has a demand for vegetables and milk while the surrounding rural areas may provide them. Spatial differentiation does not necessarily correspond to complementarity as the former is only a necessary, but not a sufficient, condition for a spatial interaction to occur. For example, some European Union (EU) countries may have a shortage of coal, while China has a large surplus, but coal trade between the EU and China is limited. This is because the energy consumption of EU countries mainly relies on other sources, namely oil, natural gas, and alternative energy. Therefore, the surplus–deficit complementarity is commodity-specific and related to economic development.

A supply–demand pair is necessary for a spatial interaction to occur, which implies that a location generating a supply provides surplus products while a location generating a demand has shortage of them. The direction, distance, and route of an interaction depend on the respective locations of supply and demand. Due to complementary relationships, spatial interactions may occur over long distances, such as the flow of petroleum between the Middle East and the United States, or over much shorter distances, such as the flow of commuters from their places of residence to their workplaces.
Complementarity is a function of physical, cultural, and technical differences between regions, which are also closely related to the level of economic development. For example, China, Australia, Brazil, and India are the top four iron ore producers. China, however, is an iron ore importer while the three other countries are exporters. In 2010, China imported 619 million tons of iron ore from the other three countries due to high demand from the steel industry, an outcome of the growing demand from its construction, infrastructure, and manufacturing industries. Therefore, the top four iron ore producers are well endowed in iron ore reserves, but they remain complementary because of differences in their respective demand driven by economic development.

Cultural differences can also be a factor for complementarity. The United States imports soy sauce from China and curry powder from India, two products consumed in large part by Chinese and Indian Americans. Meanwhile, some international trades are generated because of technical differences. For instance, the United States, Japan, and EU countries export core computer components, mobiles phones, and cars to the whole world because of their technical advantages in related fields. Finally, David Ricardo’s classical economic concept of comparative advantage provides a relative measure of the degree of economic complementarity between two countries based on their opportunity costs. More recently, global garment manufacturing companies have been bypassing China and electing for lower-cost locations in Southeast Asia, such as Vietnam and Thailand, because of rising input costs in China, such as the cost of labor.

Intervening opportunity

The second interaction factor explains the absence or insufficiency of intervening opportunities between two complementary locations (Stouffer 1940). Complementarity will only generate a flow if there is no intervening or closer location (Rodrigue, Comtois, and Slack 2013). The flows that would otherwise occur between two complementary locations may be diverted to a third location if it represents an intervening opportunity, such as a closer complementary alternative with a cheaper overall transport cost. For instance, in order to have an interaction between a customer and a store, there should not be a closer store that offers a similar array of goods at the same price.

Figure 1 depicts how intervening opportunities change the spatial interactions between two locations. In Figure 1a, locations X and Y are complementary in terms of supply and demand of a good and there is no intervening opportunity between them. Therefore, an interaction of 3 units takes place between them. In Figure 1b, location Z, which acts as an intervening opportunity, joins as a demand market located closer to location X. As a result, the demand–supply relationship between X and Y is replaced by X–Y and X–Z.

![Figure 1](https://via.placeholder.com/150)

**Figure 1** How intervening opportunity influences spatial interactions.
Some interactions are also diverted from X–Y to X–Z. Location Z captures 2 units from location Y due to its shorter distance from X. Figure 1c shows another scenario where location Z has a similar supply to location X. In this extreme case, location Z substitutes for X and becomes the only supplier for location Y.

Under some circumstances, intervening opportunity may help to create interactions between distant supply and demand pairs. As Ullman (1980) noted, the trade-diverting effect of an intervening opportunity could eventually facilitate interaction between more distant complementary locations. He used the example of a mill (point a in Figure 2) where the nearest source of wood would justify the construction of a railway segment from the mill to the forest.

Interactions between point a and the more distant complementary locations such as point c and point d might never have been established had the transportation infrastructure not been constructed in a series of incremental extensions servicing intervening opportunities.

Transferability

Transferability refers to the possibility of interactions occurring between locations by overcoming distance, time, and cost. Even though a complementary supply–demand relationship exists between locations, no interaction will take place if the transfer cost is higher than the benefits derived (Rodrigue, Comtois, and Slack 2013). The cost of overcoming distance is known as the friction of distance, which is subject to factors such as existing transportation technology and the cost of energy. The shorter the distance between supply and demand, the higher the interaction.

The rule of distance decay describes the decrease of interaction as distance increases, often as a nonlinear function. For instance, although the Middle East ranks first in the world for oil production and exports, the United States imports more crude oil from Canada due to the shorter transport distance. Also, different goods have their own suitable transport distance. For example, high-value low-weight goods such as high-tech products are transferable at a global scale while heavy, low-value goods such as construction materials are usually consumed very close to where they were produced. In other words, high-technology products have a flatter distance-decay function than that of construction materials.

Transferability can be accomplished by different transport modes depending on the weight and value of goods as well as the distance involved. Steel factories and petroleum refineries are more inclined to locate in port cities where iron ore and petroleum are commonly imported. Undoubtedly, maritime transportation is the cheapest mode for carrying heavy cargoes in large quantities over long distances. Moreover,
transferable distance could be extended by improving the capacity and efficiency of the transport system.

Distance, expressed as time or cost, has played an important role in shaping spatial interactions. However, regulations and policy also have an impact. For instance, Canada and the United States are similar countries in terms of culture, language, and institutions. The national border between them is relatively innocuous but still exerts a decisive impact on continental trade patterns. Distance matters, but political barriers do too.

SEE ALSO: Spatial organization and structure; Tobler’s first law of geography; Transport geography; Transport networks

References


Further reading


The subfield of economic geography refers to research that examines how economic processes intersect with the geographical organization of society. It is now a large international field of study that can be traced back to the early years of the discipline in Europe, and is found in most countries. Over the last 60 years, at least in Anglo-American geography, some of the most important disciplinary debates about theory and method were played out within economic geography. The first attempts to construct a logical empiricist, quantitative human geography were undertaken by economic geographers beginning in the mid-1950s. Political economy as an approach, along with the emergence of radical/critical geography, was principally worked out in economic geography from the late 1960s. More recently, economic geographers have played key roles in debates about realism, feminism, post-prefixed approaches, and the relative merits of qualitative and quantitative methodologies. Given the authors’ limited expertise, this entry is confined to anglophone economic geographical research. Future entries are anticipated that move beyond the anglophone discipline alone, allowing a global understanding and appreciation of the field. For example, the 60-year-old Japanese Association of Economic Geographers has several hundred (mostly male) members and a rich tradition of indigenous theoretical concepts. Or again, a large cluster of economic geographers in the Chinese Academy of Sciences has played a key role in shaping the geographical development of post-Mao China.

In recent years, the term “economic geography” has become widely used also outside geography as a discipline (e.g., World Bank 2009). In particular, stimulated by the work of the Nobel prize-winning economist Paul Krugman, economists have developed a rich body of research dubbed the “new economic geography.” This work has close affinities with the location theory tradition within economic geography, particularly influential in the 1960s. The economists’ “new” economic geography (new for them), or more accurately labeled geographical economics (Overman 2004), is closely aligned with contemporary mainstream economic theory; that is, it is grounded in the rational choice of individual self-interested economic agents, whose actions are coordinated by the market, and produce equilibrium outcomes.

While engagement between economic geography and geographical economics is important, in this entry the former is described as it has evolved within geography (cf. Barnes, Peck, and Sheppard 2012). Economic geography’s considerable vitality is matched only by its variegation. Nevertheless, there are four features that hold it together as a distinctive subdiscipline. First, economic geographers conceptualize the spatialities of economic activities as coevolving with those activities: economic processes produce emergent spatialities that themselves shape the evolution of those processes. Economic geographers study how space, place, scale, connectivities, and networks are produced, but also how these
affect the evolving geographical organization of the economy. Second, economic geographers examine how economic processes coevolve with political, cultural, and biophysical processes. Commodity production and market exchange do not operate in a vacuum, nor are they foundational to other processes; they are shaped by (even as they shape) political processes, discourses, and identities, as well as by the material world (which economic activities seek to transform into forms that humans can use). Third, economic geographers argue that the seeming contemporary predominance of capitalist economic activities should not be mistaken for their ubiquity, inevitability, or desirability. More-than-capitalist economic processes coexist with capitalism, and may be necessary, as well as preferable. Questions of morals and ethics are therefore placed front and center within economic geography. Finally, economic geography is a pluralist field, with space for all manner of philosophical assumptions, scales of inquiry (from the globe to the body), theoretical languages, and methods of investigation (from qualitative to quantitative and everything in between).

In the words of Sheppard, Barnes, and Peck, “Economic geography has become a peculiarly open-ended subdiscipline, one that has tended to privilege the analysis of rapidly changing phenomena, studied in real time. It is an anti-canonical project; it is open-ended and will remain so, repeatedly breaking out of the boundaries created for itself” (2012, 18). The remainder of this entry seeks to summarize how such a discipline has come into being by providing, first, a genealogy of its anglophone intellectual development (see the section “A genealogy of anglophone economic geography”), and second, selected examples of research themes to illustrate the breadth of the field and its current predilections (see the section “Current research foci”).

A genealogy of anglophone economic geography

In tracing the shifting mix of approaches that have been gathered into the big tent of anglophone economic geography over more than a century, two points need to be borne in mind. First, changes over time are not simply a succession of approaches, with new replacing old along a trajectory of progress toward the best possible approach. Instead, the disciplinary landscape is like a palimpsest, with past versions of the discipline still partially visible, not completely erased, continuing to contribute to the subject’s vitality and present form. It’s messy, with different approaches jostling and rubbing against one another, with no external criteria to determine which one is “best practice,” but it makes for a lively subdiscipline (economic geography is always living in “interesting times”). Second, as geographers it is important to attend to the geography of this knowledge production: within our self-imposed geographical restriction to the anglophone world, different approaches have gained traction in different spatiotemporal contexts. An approach that makes sense in one context may not translate well to others, further undermining the possibility of any singular best practice.

From commercial to regional geography

Like the broader discipline, economic geography was historically tethered to nineteenth-century European imperialism. Initially called commercial geography, the subdiscipline’s charge was to provide practical geographical knowledge to the agents of empire: the military, the colonial bureaucracy, and the business classes. Focused on the commodity, economic geography delivered meticulously detailed information about where different goods were produced, and could
be produced, the necessary conditions of their production, types of regional specialization, trade patterns, and circuits of available transportation. While early economic geography might not have been intellectually exciting — one of its early exponents, George Chisholm, said, “if … there is some drudgery [to it], I see no harm” (quoted in MacLean 1988, 25) — it provided the commercial scaffolding of empire. The first-ever English language textbook in economic geography was Chisholm’s (1889) *A Handbook of Commercial Geography* (earlier ones had been published in German). Jammed with maps, tables, and economic geographical facts and figures, the *Handbook* was designed both to give Britain’s business classes an applied education and a competitive edge, and to provide its imperial civil servants with knowledge of the globe that they administered.

Chisholm’s book was also a volume of scientific knowledge. In 1882 the German geographer Wilhelm Götz distinguished between commercial geography, which “chiefly served practical ends,” and economic geography defined by “the scientific task of dealing with the nature of world areas in their direct influence upon the production of commodities and the movement of goods” (Götz quoted in Sapper 1931, 627). Chisholm took that scientific charge to mean positing “nature” as a central causative agent, arguing that the environment made each place in the world uniquely fitted to undertaking a particular type of economic geographical activity. This view became extreme, however, colored by racism, through the work of a number of environmental determinists working in economic geography at the beginning of the twentieth century. The Yale geographer Ellsworth Huntington (1915) was the most notorious, arguing that labor’s “mental” and “physical efficiency” peaked in temperate climatic regimes found in western Europe, parts of North America, and white settler colonies. In contrast, mental and physical efficiency fell to disastrously low levels in tropical regions, condemning them to “backwardness,” making them “The White Man’s Burden.”

By the end of World War I, economic geography had become well entrenched in both UK and American universities. It was human geography. During the interwar period, with a less rampant form of imperialism, and growing criticism of environmental determinism for its blatant racism and shoddy scholarship, economic geographers moved away from emphasizing global commodity production to stressing local economic interconnections, taking the form of unique regions.

The turn to the region was readily seen in economic geographical textbooks published in North America from the mid-1920s onwards. It involved characterizing regions using a common typological scheme made up of such categories as “leading industries,” “natural resources,” “modes of transportation,” and so on. Once all regions were classified, their differences, and thus their individual uniqueness, were immediately evident by reading along any given row of the typology. For example, Vernor Finch and Ray Whitbeck’s (1924) *Economic Geography* used a fourfold classification scheme for each of the regions the book covered: agriculture; minerals; manufacture; and commercial trade, transportation, and communications. Another example was Clarence Jones’s (1935) textbook *Economic Geography* that deployed an eightfold typology. In both cases, the typology functioned as a grid into which regional facts, variously mapped, drawn, photographed, and exhaustively described, were slotted. By typologically comparing the facts of different regions, economic geographical difference became transparent, and regional uniqueness shone by its own light.

The intellectual justification for this regional approach wasn’t articulated until just before World War II. The American geographer
Richard Hartshorne argued that by its very nature each region was a collection of exceptional features, describable only in their own terms. Consequently, no generalizing theories, no scientific laws, were possible. As Hartshorne (1939, 44) wrote, “Regional geography, we conclude, is literally what its title expresses. ... [I]t is essentially a descriptive science concerned with the description and interpretation of unique cases.”

Spatial science and the turn to theory

Hartshorne’s book was published at exactly the wrong time. The regional descriptivism it proposed for geography, and economic geography in particular, was quickly out of step with a set of larger changes pointing in exactly the opposite direction. From the beginning of World War II, a number of social sciences, and even some humanities, were transformed from the type of descriptive approach championed by Hartshorne to one that instead emphasized scientific generalization and explanation, and was designed to accomplish practical ends. In part because of Hartshorne’s influence, economic geography initially resisted that impulse, but by the mid-1950s it began to join in. The resulting shift to spatial science, represented by geography’s “quantitative revolution,” profoundly altered economic geography (Burton 1963). It swept away talk of ideographic regional uniqueness, replacing it with scientific forms of general theorizing and rigorous statistical techniques of description and analysis.

Spatial science was defined by five main features. First, and foremost, was its use of formal theories and explanatory models. Necessarily, these were imported from elsewhere: from economics (rational choice theory, general and partial equilibrium, and German location theory), and perhaps less likely, from physics (gravity and potential models, and later the entropy model). Second, was utilization of an increasingly sophisticated arsenal of quantitative methods. At first they were off-the-peg inferential statistical techniques. Later, specialized statistical measures and methods were designed in-house to meet the peculiar features of geographical data (e.g., techniques of spatial autocorrelation). Third was the deployment of computers. At first they were very crude and limited, but within a decade they performed hitherto unimaginable calculations, for example, the inversion of large urban and regional economic input-output matrices that would have taken a lifetime to calculate by hand. Fourth, a philosophical justification was made for spatial science based on positivism, the idea that only scientific knowledge is authentic knowledge. Fred K. Schaefer (1953) provided an early influential justification that was later broadened and deepened by David Harvey (1969). Finally, there was a focus on abstract spatialities. Regions remained part of the economic geographical lexicon, but conceived utterly differently. Regions were now explanatory, theoretical, and instrumental, spatial units to achieve policy and planning objectives (brilliantly realized in the parallel movement of regional science). Finch, Whitbeck, and Fielden Jones (and their respective textbooks) were simply no longer recognizable as part of economic geography.

Like most intellectual revolutions, spatial science began at only a few sites, from which it diffused. In Europe it was associated with Cambridge University, Bristol University, and Lund University; within North America, it started at the Universities of Iowa, Washington, Chicago, and Toronto, as well as Ohio State and Northwestern. In each of these places, groups of young, bright, ambitious, competitive, and almost exclusively white male students gathered to participate.

More than anything else the revolutionary act was the introduction of theory, making
economic geography for the first time a social science. It was, as Kevin Cox (2014, 145) calls it, “the great hinge of 20th-century geography.” While economic geography subsequently radically changed, the tie binding all subsequent versions was the generation of theory. While contemporary economic geographers might now reject the substance of spatial science, through their continuing theoretical sensibility they are heirs to that earlier tradition.

Letting many flowers bloom – post-spatial science

The first signs of trouble for spatial science came in the early 1970s when David Harvey (1973) disavowed his earlier support, declaring that the quantitative revolution had run its course, telling us less and less of any import. Harvey was the first of a series of high-profile defections. In retrospect, the problem was the mathematically abstract and narrow conception of economic geography that spatial science proffered. It was not true to economic geography’s own variegated disciplinary history; not true to the historical moment of the early 1970s that was increasingly politicized and drawn to issues of social relevance; and not true even to its own scientific logic, as assorted logical contradictions, inconsistencies, and gaps revealed.

The subsequent 40 years has seen the ceaseless proliferation of theoretical alternatives to spatial science. For the most part these alternative theories are not couched in the scientific hypothetico-deductive form. Nonetheless, they are used to explain, and in some cases to intervene within, a robustly drawn and variegated empirical world. They are not new Kuhnian paradigms that replace the old, constructed to cope with emerging anomalies. Indeed, there is no agreement in economic geography about what constitutes an anomaly.

In the early 1970s Harvey introduced classical Marxism as his theoretical alternative to spatial science, from which he continues to draw and find creative inspiration. Yet Marxist theory is only one component of a larger political economy tradition that has more or less dominated economic geography since that time. By political economy we mean the tradition that harks back to classical economics (Adam Smith, David Ricardo, and Karl Marx), emphasizing that economic processes cannot be understood without attention to politics and the state, that capitalist commodity production and exchange are not simply the result of self-regulating efficient markets. To reference economic geography’s common ground of political economy is not saying a lot, though. Political economy’s strands are multifarious, and heterodox. In rough chronological order, after Harvey’s classical Marxism, political economy has been taken up in economic geography as Doreen Massey’s Althusserian-inspired spatial divisions of labor thesis; as the locality approach; as regulation theory, taken primarily from French economists and influential in understanding both the decline of Fordist industry in the United Kingdom and the northeastern United States, and the rise of post-Fordist growth dynamics, especially in California; and an institutional approach concerned with identifying formative mechanisms of industrial districts, applied particularly to high-tech agglomerations.

Yet this list takes us up only to the mid-1990s. Since then it has become only more complicated and variegated. Uneasily bolted on to political economy was a cultural approach (“the cultural turn”), influenced by poststructuralism, and focused on how both discourse and gendered, raced, and sexualized identities shape economic processes. Examining gender became a disciplinary theme from the 1970s even within spatial science. Massey brought it to the fore in her
work on spatial divisions of labor, but it became even more central in a series of works published in the second half of the 1990s emphasizing culture, and drawing on poststructural feminist theory (McDowell 1999). Gibson-Graham’s (1996) *The End of Capitalism (As We Knew It)* was perhaps the most influential of those works, eclectically piecing together post-Marxism, feminism, and poststructuralism into a critique of political economy that proposed a new substantive disciplinary agenda of research on alternative or “diverse” economies.

Even more recently, Asian and German geographers have advocated for a “relational turn,” stressing the increasingly networked nature of economic activities. German geographers originating in Frankfurt, drawing on science studies, have explored the performativity of markets. Dutch and other European geographers based at Utrecht have catalyzed evolutionary economic geography, drawing on evolutionary economics to model how diverse populations of firms change, and the implications for the regions in which they are located. European and American geographers interested in questions of development and the Global South have advocated for a postcolonial approach to economic geography – one that seeks to understand distinctive economic practices emerging from beyond the North Atlantic realm, such as Islamic finance (Yeung 2005; Boschma and Martin 2010; Pollard, McEwan, and Hughes 2011; Berndt and Boeckler 2012).

Last but not least, European and North American scholars based (albeit sometimes intermittently) at the London School of Economics (LSE) have reconstructed a version of spatial science that is formal, resting on neoclassical micro-economic principles, deploying rigorous econometric testing, and aspiring to influence government policy. Its intellectual basis was the 1990s work of Paul Krugman that turned on his innovative theorization of the relation between spatial growth, trade, and agglomeration economies. That work catalyzed a “new economic geography” or geographical economics (Brakman, Garretsen, and Von Marrewijk 2009) leading to attempts to strike up a disciplinary conversation between geographers and economists, including a new journal, *the Journal of Economic Geography* (launched in 2001). The reconfigured LSE Department of Geography and Environment was another such outcome. In terms of impact and unity of purpose, it is now likely the most successful grouping of economic geographers in the United Kingdom.

In short, in less than 150 years, anglophone economic geography has moved from a handmaiden to empire to “physics envy” to an intellectually open, eclectic, pluralist, and possibly chaotic discipline straddling the humanities and social sciences. Anything goes, and sometimes does. Its seeming lack of a theoretical core can be disconcerting. Critics complain that the result is anarchy, eclecticism in which nothing fits, producing flighty, sloppy, and sometimes incomprehensible works. One result perhaps is Balkanization, a discipline of solipsism and solitudes. The counter response is that fragmented theorizing is necessary to understand the increasingly fragmented geographical economy in which we live and study. It is exciting, however, with no chance of contemporary economic geography “boring its audience to death” (Nigel Thrift’s fear at the end of the 1990s; Thrift 2000, 692). The unusual challenge for this subdiscipline will have less to do with breaking new ground, which it seems to do on a daily basis without breaking sweat, and more with holding its existing ground. This diversity and flexibility can be a weakness as well as a strength. Economic geographers must be willing to make space for diverse perspectives, promoting a culture of constructive and rigorous engagement and mutual learning (Barnes and Sheppard 2010).
From theoretical to methodological diversity

Underlying this efflorescence of variegated theoretical traditions are questions of method – of how to apply these perspectives to, and assess their validity against, “really existing” economic geographies. The term “underlying” is used here because questions of method have not received the attention they deserve (Tickell et al. 2007). Issues of methodology have been subject to a “don’t ask, don’t tell” policy. Economic geographers of all stripes insist on the importance of empirical methods, concerned with representing an empirical object. But what counts as empirical information, the methods used to collect, assemble, and interpret data, and rules for determining the validity and power of explanatory frameworks all have significantly shifted and multiplied.

Under commercial geography, the concern was the assiduous collection of information that was as precise as possible: ideally, numbers and statistics. Under spatial science the bar was raised to aspire to logical empiricism: the rigorous statistical testing of theory-driven hypotheses against the “facts.” The turn away from spatial science ushered in a “qualitative revolution,” however: a move to methods concerned with gathering and analyzing non-numerical data. This was motivated by trenchant criticisms of the so-called positivist philosophy underlying spatial science – criticisms that stressed the theory-laden nature of observation. It is not simply “just the facts.” Data are not out there, like apples, waiting to be picked; what is available depends on the theories and agendas of those (including economic geographers) collecting, ordering, and distributing data, and what economic geographers see depends on the interpretive and theoretical framework they bring to empirical work.

The “qualitative revolution” began with “intensive case study research,” pioneered by proponents of critical realism in the early 1980s (Sayer 1984). Qualitative methods have multiplied since, now including in-depth interviews, focus groups, oral histories, ethnographies, participant observation, discourse and textual analysis, following networks, action research, and more besides. In this new methodological environment, nothing is proscribed; everything is permitted, if not equally valued. Critics complain that the turn to qualitative methods in economic geography has induced slapdash and superficial research. In response, qualitative researchers stress the difficulty of, and care necessary to undertake, rigorous case study and qualitative research – that rigor means different things in different contexts. Methods must be appropriate to the necessary data, and if that is qualitative, reflecting what is important for understanding the object of inquiry.

The upside is diversity and rapid change. Most of the qualitative techniques listed above would have been viewed as beyond the pale, or at best, suspiciously avant garde, when first introduced. The downside has been the derogation of quantitative analysis and loss of associated skills. Lacking the training, economic geographers are increasingly unable to undertake, or knowledgeably critique, statistical and numerical analysis. There is emergent interest in mixed methods approaches, combining different (qualitative but also quantitative) methodologies in the spirit of engaged pluralism. Doing this effectively is even more challenging than rigorous quantitative or qualitative methods, however. First, research teams are increasingly necessary because researchers too often specialize in qualitative or quantitative methods, however. Second, developing mixed methods research designs that rigorously tackle the questions at hand, rather than mixing approaches together haphazardly, is still very much in the experimental stage in the social sciences.
ECONOMIC GEOGRAPHY

Current research foci

To illustrate the kinds of research currently receiving significant attention from economic geographers, and how this reflects the subdiscipline as it is described here, summaries of eight areas of focus are provided, organized around aspects of the spatial economy. It is not a definitive list, however (no definitive list is possible). Rather, these are simply current major areas of research in contemporary anglophone economic geography.

Worlds of commodity production: industry and services

From its early days, economic geographers have been interested in what is produced where, and how. While industries and services have different kinds of locational requirements (e.g., access to resources versus access to consumers), contemporary research focuses on theories that apply, in different ways, to both metasectors (some question whether the distinction remains meaningful in today’s globalized world). Within its current heterodox political economy phase, three major areas of work stand out, at the scales of the firm, the region, and the globe.

At the scale of the firm, the emphasis has moved away from explaining the rationality of location decisions to understanding how firms make decisions about what to produce and how, and the role that space plays in these processes. A particularly rich area of contemporary research is known as evolutionary economic geography. Drawing analogies between competition and the survival of firms and Charles Darwin’s theory of evolution, the paradigm of generalized Darwinism developed in evolutionary economics has been applied to explain why firms are founded, closed, and grow or decline in success relative to their competitors. Unlike economics’ focus on an idealized representative agent making rational decisions, and equilibrium outcomes, it is recognized that any regional population of firms, also those producing the same commodity, will vary in terms of size, technology, strategy, and labor relations. Hypothesizing that each firm has a production routine, analogous to genes in organisms, it is argued that the dynamics of a population of firms will depend on three processes: variation (between firms), selection (of firms, through competition), and retention (the propagation of routines within and between firms). A key question has been how firm dynamics relate to regional economic dynamics: How do some regions experience lock-in, whereby their economic structures stagnate and decline, whereas others remain dynamic, or attract emergent growth sectors? This is argued to be related to a region’s related variety: the more similar firms are, the greater are localization economies and innovation, within a particular sector. When they are less similar (unrelated variety) there are fewer such advantages, but the region is more open to qualitative change rather than path dependence (Boschma and Iammarino 2009; Boschma and Martin 2010).

At the regional scale of metropolitan areas, catalyzed by research on industrial districts, scholars have identified what they believe are keys to the success or failure of territorial economies, particularly in a world where globalization is always a threat to regional economic cohesion. The key is agglomeration economies: the relational assets in a region that facilitate both economic growth and resilience to change. One potential asset is related variety; others include reduced transactions costs, tacit knowledge or local “buzz,” the sociocultural milieu, and the presence of a “creative class.” Each of these has been identified as enhancing local economic dynamism and flexibility under the right conditions, and leading to regional decline under the wrong ones, fostering...
place-based explanations of regional economic growth. Key kinds of territorial economies identified include industrial districts, learning regions, and city-regions (Scott 2006; Storper 2013).

At the global scale, it has been recognized that many economic activities are increasingly spatially decentralized, with different locations around the world playing different roles in the production or assembly of a commodity. Multinational corporations, surrounded by an ecosystem of subcontractors and licensees, thus spread production around the world in what Richard Baldwin (2006) has dubbed “the great unbundling” of production. If the regional-scale research stresses the importance of local networking, geographers studying this unbundling seek to understand it in terms of global production networks (GPNs). They not only examine the economics and power relations constituting these networks, with some firms in a position to shape outcomes and others just bit players, but also trace how the territorial strategies of states, labor unions, and multilateral governance organizations each affect the structure and evolution of GPNs (Coe et al. 2004).

**Worlds of consumption**

Consumer markets traditionally were studied in terms of the location of retailers, consumers’ behavior, and spatial price gradients. Beginning with central place theory, this led to a more general theorization of the monopolistic nature of spatial markets. It was determined that geography challenges conventional economic wisdom. For example, spatial competition need not result in spatial price equilibrium and may not be socially beneficial (Sheppard and Curry 1982). In the past two decades, geographers studying retailing have turned their attention to the spatial strategies of corporate retailers, including their roles within global production networks. This recent research aligns with how economic geographers characteristically begin with geographies of production, unlike mainstream economics’ focus on markets and consumption, extending this to the production of services. Retail geographers now study: the impact of corporations on retailing, and their capacity to drive small retailers out of business and dramatically reshape geographies of consumption (Coe and Wrigley 2009); the role of geodemographic marketing in shaping consumption patterns and norms, segmenting markets, and customizing marketing to wherever consumers find themselves (location-based services); and the emergence of e-commerce and its material as well as virtual geographies.

With the poststructural and feminist turns to issues of identity, however, considerable research has examined the relationship between retailers, consumption, and identity, with particular attention to issues of social difference: the intersectionality of gender, class, race, sexuality, and geographical location (Jackson and Holbrook 1995; Cook and Woodyear 2012; Mansvelt 2012). This examines how consumers’ desires emerge from and intersect with their sociospatial positionality, and how these issues relate to marketing strategies (branding, store location, and spatial organization). The role of citizens as consumers also has received attention, examining how households are expected and induced to consume to foster economic growth, particularly in postindustrial territorial economies.

Most recently, geographers have taken up the question of marketization: how spatial markets emerge and the implications thereof. Geographers have found approaches in economic sociology more compatible with their sense of how capitalism works than economists’ categorizations of ideal typical market structures. They examine how information networks unequally shape participants’ ability to take advantage
of markets (“the strength of weak ties”), how trading technologies shape market structure and performance, and how markets themselves are produced. Market formations are regarded as emergent features, shaped by theoretical predispositions, ideology (e.g., neoliberalism), interests, technologies, and geographies. If perfect markets emerge, this is because the participants believe in their desirability and/or actively create them. Turning to the uneven geographies of marketization, geographers have examined the construction and elimination of boundaries, both those separating markets and those separating what is sold as a commodity and what is not – such as markets in waste, genes, and human organs (Grabher 2006; Mackenzie 2009; Berndt and Boeckler 2012; Gidwani 2012; Parry 2012).

Labor, work, and bodies

From the beginnings of the discipline, labor was considered an important economic geographical factor influencing the location of firms and industry. But there was no analysis of labor markets as such, and nothing on either the conditions defining work, or especially the kinds of bodies doing the work. That changed in the 1970s when radical economic geographers began emphasizing the centrality of labor within production and the role of a distinct social and political class, the working class (Lier 2007). The 1970s, though, was exactly the decade when old verities about labor, work, bodies, and the working class changed. Hitherto labor in the Global North was conceived as expending brawn working in a factory (engaged in “execution” rather than “conception,” to use Frederick Taylor’s famous pair of terms). The bodies were conceived predominantly as male bodies, and the social class to which they belonged, the working class, was taken by the Left as the vanguard of progressive politics. This became less and less true as the larger processes of deindustrialization, neoliberalism, and globalization gained purchase in the 1980s. Consequently, it became necessary to retheorize labor geography. While elements of radical geography’s approach were retained in that retheorization, many new elements were added, drawn from poststructuralism, especially feminist and postcolonial theory, and emphasized identity, embodiment, and transnationalism.

From the 1970s, deindustrialization in the Global North decimated traditional male Fordist industrial jobs along with their privileges (e.g., the “family wage”). It was the end of “Old Father Ford,” provoking a crisis in masculine identity, especially for young males (McDowell 1991). Furthermore, with the rollout of neoliberalism from 1980 onwards, traditional unions were also undermined, along with their associated left-wing politics. Traditional manufacturing jobs in the Global North were moved offshore, often into the Global South, to sites now linked by unfolding GPNs. Operating from free trade zones, large multinational corporations drew on abundant, very cheap labor, including pools of young women, who labored excessive hours and were subjected both to work injuries and to extreme disciplinary measures both inside the factory and outside. They became the world’s disposable workforce (Wright 2006).

Some in economic geography argued that labor should now be conceived as possessing agency, as having an ability to determine their geographical fate (a “labor geography” rather than a “geography of labor,” Herod 2001). But much recent empirical work examining low-paid service workers in the Global North, especially women, international migrants, and workers of color, shows the reverse. Employment is increasingly precarious, creating a “precariat,” with almost no opportunity to assert agency. This form of employment is defined by part-timeism, short-term contracts, and the absence of benefits...
(Peck and Theodore 2001). That’s less true at the opposite end of the labor market, where highly paid knowledge workers make up Richard Florida’s (2002) creative class. While they also may be on contract, carrying out project work, they are often well paid, enabling them to assert some agency by choosing the country, city, or neighborhood in which they live. In both cases, though, the larger point is that labor and work are not neutral. Workers always are more than a technical input into production. The social, cultural, and geographical contexts in which work is practiced leaves its mark (sometimes literally), shaping social identity, bodily comportment, and the life led, including where it is led geographically.

**Governance**

Within the political economy tradition dominating anglophone economic geography, there is consensus that markets cannot function effectively in isolation: issues of state regulation and governance also must be addressed (bringing formal politics into economic geography). Regulation refers to actions taken by state and quasi-state agencies, at various scales, constraining the actions of both producers and consumers. Governance embraces the broader sense of how commodity production and consumption are (increasingly) governed also by and through all kinds of nonstate institutions.

Much attention has been paid to how different forms of capitalism emerge in different territorial contexts, and how these change. Economic geographers have drawn extensively on French regulation theory, which argues that a functioning territorial capitalist economy must combine a regime of accumulation (how commodity production is undertaken) with a mode of regulation (how the state regulates the market to manage the national balance between supply and demand) (Dunford 1990).

Originally developed to explain the emergence and problems of Fordism as a national territorial economic system in the Atlantic realm after 1945, geographers have extended it to consider geographical scale and spatial variegation. They examine how regulatory regimes vary across space at the national scale, reflecting the persistence of national cultural and political traditions – variegated capitalisms (Peck and Theodore 2007). They also argue that local governments within a national territory may develop regulatory norms that differ from those at the national level, depending also on the central–local state relations in a given national context (Tickell and Peck 1992).

Particularly at the subnational scale, much attention has been paid to how local states act to foster investment in their territory. Urban entrepreneurial strategies include both inward investment strategies – subsidies and improvements to the local “business climate,” designed to persuade firms (and knowledge workers) to move in (or stay) – and incubator strategies – seeking to create conditions that foster local innovation, dynamism, and agglomeration economies (Hall and Hubbard 1998).

The rise and geographical differentiation of neoliberalism has received particularly close attention. Neoliberalism has become something of a leitmotif of anglophone economic (indeed, human) geography since 2000. This research includes analyses of the nature and diffusion of neoliberal governance, debates about how neoliberalism came to replace seemingly impregnable state-led modes of regulation, and explanations of why neoliberalism “in the wild” is persistently variegated and never converges on a pure, ideal–typical neoliberal model, notwithstanding the relatively ubiquitous presence of processes of neoliberalization (Larner 2000; Harvey 2006; England and Ward 2007; Peck 2010).
Turning to the question of how neoliberalization diffuses across space and time, economic geographers also study policy mobilities: how neoliberal principles become “best practice” policies that take flight, mutating as they do so, rapidly moving between localities (Peck and Theodore 2010). Important debates remain, particularly about whether neoliberalism has become ubiquitous, whether political economic accounts suffice, and whether and how it can be contested (Leitner, Peck, and Sheppard 2007; Barnett et al. 2008; Wray et al. 2013), but the uneven geographies of this shift from state regulation (e.g., Fordism) to market-centric governance are now quite well understood. There also have been many case studies documenting its impact – enhancing inequality within cities, regions, and nation-states, and between world regions.

In terms of governance more broadly, economic geographers have examined the role, influence, and evolution of institutions on territorial economies. They have unraveled the conditions under which institutions may facilitate or block regional and urban economic dynamism, examining how they can channel regional economic development along certain paths (path dependence), some of which may be highly beneficial (with institutions helping put in place “relational assets” that foster innovation and economic growth), and others of which may result in stagnation, when institutional lock-in prevents a region from rebounding from a negative shock. These are important issues to understanding the resilience of territorial economies (Martin 2000).

Finance

Economic geographers have paid some attention to money and credit over the years, but recently particularly to financialization. Research into forms of money and credit explored how the geographical extension of trade and capitalist economic relations required the development of forms of money and credit that enable exchange to occur across space as well as time. When non-geographers claimed that the globalization of finance markets, facilitated by cyberspace, is creating a world where geography no longer matters, geographers responded by studying why financial activities increasingly concentrate in a handful of centers of global finance (New York, London, Frankfurt, and Tokyo, now supplemented by Hong Kong, Singapore, and Shanghai). Face-to-face communication remains essential to take advantage of the tacit knowledge (e.g., insider information) essential to gaming the market; digital networks have resulted in an extreme spatial concentration of high-frequency trading servers, to gain a millisecond advantage over competitors. They also examined the different spatial scope of different kinds of lending, uneven geographies of how finance is governed, and the role of such institutional actors as pension funds and real estate investment trusts (Leyshon and Thrift 1997; Clark and Wojcik 2007).

Turning to the relationship between financialization and globalization, geographers have examined the role of uneven national regulatory and governance structures, such as Asian and offshore banking systems; connections across scales, as supranational finance institutions shape lending possibilities, and conditionalities attached to lending, in nation-states and localities; the emergence of novel financial practices in certain regions (e.g., Islamic finance, sovereign wealth funds) and at particular scales (e.g., micro-finance, Ithaca dollars); and uneven geographies of credit, lending, and indebtedness at every scale (from financial exclusion in neighborhoods to the lack of lending in sub-Saharan Africa). They study the financialization of nature, and the geography of finance crises – notably the
patterns of uneven geographical development underlying the 2007 global financial crisis, and its spatially differentiated consequences. Echoing related scholarship in the humanities and anthropology, arguing that financialization is as much a cultural as an economic phenomenon, economic geographers also examine how financialization and its emergent pro-market norms have become increasingly central to everyday cultural practices. They examine the increasingly pervasive nature of discourses and daily practices that reflect as they also disperse these norms, and the multiple ways in which investment and credit underwrite all kinds of activities (Pike and Pollard 2010; French, Leyshon and Wainwright 2011; Christophers 2013; Clark, Dixon, and Monk 2013).

Material worlds

The material worlds of nature and resources have been the focus of animated disciplinary discussions from the beginning, and for good reason. Currently around 50 trillion tonnes of the Earth’s material resources are annually appropriated for one human use or another, transformed into primary commodities by the “metabolic engine” of the economy (Bridge 2009, 1222).

When economic geography was first institutionalized in the late nineteenth century the discipline was primarily about material worlds, codifying meticulous lists of natural resources, their geographical distribution, and corollary environmental conditions. Interest waned in the 1920s and 1930s, and under postwar spatial science the topic of “natural resources” became an intellectual backwater. Drawing upon Marx, David Harvey (1974) reinvigorated the subject. His was not nature “red-in-claw” but social-nature, “the production of nature.” One of capitalism’s imperatives, Harvey argued, is to transform original “first” nature to produce “second nature” (Smith 1984). In the process, nature is commodified, subject to the market, becoming hybrid: nature and society join, forming “socio-nature” (Castree and Braun 2001). But to become a commodity, conditions have to be just right. As Erich Zimmerman (1951, 814–815) famously put it: “resources are not: they become.” Coal is merely a rock that makes your hands dirty until capitalism invents the steam engine. But once capitalism emerges, pre-existing nature is transformed. Nature still exists, but how it is exploited, used, thought about, and represented is irredeemably and sometimes cataclysmically altered.

With Harvey linking nature to political economy, economic geography overlapped with political ecology, which similarly brought a political economic understanding to resource use and production within, initially, the Global South. The exemplar is Michael Watts’s work in Nigeria, showing how social relationships, first under imperialism then under independence, made the country’s natural resources neither natural nor even a resource, creating instead maldevelopment, inequality, and social mayhem (Watts 2004).

Another political economic approach, Immanuel Wallerstein’s world systems theory, provided a different, macro-framework connecting economic geography and material worlds: the global commodity chain. By following the “thing” – a papaya, a cut flower, a barrel of oil – as it moves and is transformed from first producer in one place to final consumer in another reveals: (i) the differential forms and effects of materiality; (ii) the diverse geographies inherent in all commodities; (iii) the asymmetrical sociogeographical relationships within the production process; and (iv) the uneven effects of state regulatory regimes (Cook 2004).
ECONOMIC GEOGRAPHY

The most recent body of work about resources elaborates precisely on this last point, focusing on how one regulatory regime in particular, the regime of the last 30 years, neoliberalism, has stamped nature in its likeness. For examples as different as water, fish, trees, ores, and CO₂, under neoliberalism nature becomes privatized, commodified, priced, and marketed. Geography, though, remains vital. Each place does neoliberalism differently, creating not one seamless landscape of indistinguishable process but a patchwork quilt of variegated resource regimes (Heynen et al. 2007). Neoliberalism has also been the context for understanding the development of new markets for either goods or services ostensibly meant to improve environmental conditions. There are markets for carbon reduction, body parts, biodiversity enhancement, and more generally for the provision of environmental services.

Globalization and development

For many years, self-styled economic geographers in the anglophone realm concentrated their effort on the parts of the world they inhabited, those of the “first world” – Europe, Japan, and the white settler colonies of North America, Australia, South Africa, and New Zealand. With the implosion of the Soviet sphere, attention turned to questions of “transition” toward capitalism in the post-Soviet world. Yet the “third world” – those countries whose inhabitants suffered from colonialism and still struggle to overcome the economic disadvantages and global peripheralization they inherited from colonialism – was largely left to what was styled “development geography.” During this past decade, recognizing that problems of underdevelopment in the postcolonial world differ in scale rather than kind from those of impoverished regions in the first world, economic geographers increasingly are turning to study the post-colony, integrating development with economic geography (Slater 2004; Lawson 2007).

In so doing, economic geographers have focused on development at scales ranging from the globe to the village, tracing out how these are interconnected. They also are taking up the question of what is meant by development and what is at stake in the development discourses that circulate alongside global policymaking, in such forms as sustainable development and Millennial Development Goals.

At the global scale, geographers examine uneven geographies of economic globalization, considering such questions as: Why is it that economic inequality remains so persistent notwithstanding the increasing interconnectedness of the global economy? Why has China prospered, India struggled, and sub-Saharan Africa stumbled? Does free trade, and the unrestricted movement of direct and portfolio investment, level the global playing field, or does it reinforce pre-existing inequalities? Challenging the tendency in development economics to attribute success or failure to local, place-based characteristics (culture, physical geography, governance institutions), economic geographers argue that the uneven relations connecting places – the flows and mobilities of commodities, finance, ideas, and people, not to mention water and climate – may be equally important in determining the wealth of territories (Sheppard 2011). They examine not only how local possibilities are shaped by broader-scale processes, processes that localities seemingly have limited influence over, but also how local events can trigger significant broader-scale consequences. At regional and neighborhood scales, they examine why some localities prosper whereas others do not, and how these dynamics are shaped by both the colonial past and contemporary processes.
At the local scale, in fields, villages, factories, and urban neighborhoods, economic and development geographers study the embodied practices of individuals working to improve their livelihood prospects, but also how these connect to the households and other collectivities to which they belong. They show that what others dismiss as “traditional” practices, represented as being in need of development to modernize, are often better adjusted to the circumstances in which they emerged than are the practices of development that were introduced to replace them. Such analysis centers on the diversity of economic practices, their relation to the biophysical environment, and the role of power and constructed identities (race, age, class, caste, gender, etc.) in shaping these practices and livelihood chances. Substantive topics range from peasant farming and subsistence, to informal housing and economic practices, to formal employment in export processing zones, and connections between these places and practices (Rankin 2004; Porter 2006).

In terms of what is meant by development, research by economic geographers, documenting ways in which policy interventions informed by European notions of development undermine livelihood possibilities in postcolonial societies, raises questions about the ubiquity and universality of such notions of development. If the development trajectory that has brought prosperity to the Global North also brought impoverishment and environmental destruction to the Global South, then the problem may lie in this very notion of development, rather than in the failings of southern places and populations. Thus economic and development geographers have turned to interrogating northern definitions of development – as the product of globalizing capitalism – distinguishing between “development,” as the set of practices developed in a particular geographical context that have enabled its residents to live well in their own terms, and “Development” as a universal definition, to be foisted on all people and places, and against which their progress should be judged (Hart 2002). Geographers raising these questions draw on postcolonial theory, questioning whether northern theory (also in economic geography) and development principles are applicable across the Global South (defined as those, everywhere, who find themselves living precariously).

More-than-capitalist economies

Dating back to commercial geography, economic geographers have long taken an interest in the different economic practices to be found in different parts of the world. As ideas of development spread from the North Atlantic realm to be applied everywhere, noncapitalist practices came to be looked on as traditional and backward, as in need of replacement by integration into globalizing capitalism. Even economic geographers critical of capitalism and its consequences, documenting how it creates accumulation by dispossession or increased sociospatial inequality, have tended to accept capitalist practices (in some form or other), by now globally hegemonic, as the only kinds of economic activity worthy of their attention.

This default position has been influentially questioned over the past two decades, with economic geographers turning their attention to community and diverse economies. In their book, The End of Capitalism (As We Knew It), Julie Graham and Katherine Gibson, writing as J.K. Gibson-Graham (Gibson-Graham 1996), forcefully argued that many noncapitalist economic practices continue to coexist with capitalism, and in its very geographical heart (e.g., New England). Connecting Marxism with feminist and poststructural theory, they argued that such “community economies” practices,
going on under our noses, deserve far more attention from economic geographers than they receive, as a locus for potential alternatives to globalizing capitalism.

While such practices may have seemed obvious to development geographers, working in societies where the informal economy and subsistence practices predominate, their presence in the Global North proved something of a revelation. The result is a branch of anglophone economic geography now called the “diverse economies” school, focusing on variegated noncapitalist practices, and ways that they are tangential to or simply separated from capitalism. These include: local economic and trading systems, barter systems, labor sharing, workers’ democracies, and so on. Beyond the study of such practices, economic geographers have worked with communities. They have both brought an awareness about such practices, and fostered and expanded them using novel forms of academic research. This work has been criticized for focusing on seemingly insignificant and marginal practices, and for presuming that they can exist outside capitalism (they are more-than-capitalist rather than noncapitalist). Nevertheless, these ideas have brought attention to the different forms both that the economy can take, and through which neoliberalism, and capitalism more generally, can be contested (Gibson-Graham, Cameron, and Healy 2013).

Conclusion

The ongoing vitality and diversity of anglophone economic geography that circulates around geographical political economy promises a bright future. But its practitioners must be willing to engage constructively with one another and with scholars in cognate disciplines examining related questions. Geographers examine how economic processes shape, and are shaped by, geographical space (place, scale, connectivity, mobility); they examine how economic processes coevolve with political, cultural, and biophysical processes; they study how subject formation and identity at one end, and global political economic structures at the other, are interrelated; they take an interest in noncapitalist economic processes, critical of the proposition that regulated capitalist markets optimize societal wellbeing and ecological sustainability. Real-world-oriented, economic geographers study some of the major challenges facing contemporary human society and its relation to political economic processes: global production networks, the politics and culture of consumption, work and labor, the role of state regulation and governance institutions, nature–economy relations, financialization, globalization and development, and diverse economic systems.

Nevertheless, looking forward, it is possible to identify priority areas for new research. Economic geographers have paid too little attention to the geography of logistics – the transportation and communications networks – which has coevolved with globalizing capitalism, shaping its emergent geography at all scales. Notwithstanding the scholarship on material worlds, economic geographers have paid far too little attention to arguably the most urgent emergent aspect of the geographical economy: global environmental change, carbon economies, and ecosystem services. The wealth of economic geographical research about firms, consumers, and territories has yet to be complemented by systematic empirical analysis at the global scale (taking advantage of large historical databases).

It is hoped that these and other issues will be taken up in the spirit of engaged pluralism across different approaches within economic geography, and across different national communities of economic geographers. It is also our hope
that, through such engagement, economic geographers will continue to interrogate the moral foundations and implications of their analysis for the more-than-human world that we all inhabit.

SEE ALSO: Development; Industrial geography; Labor geography; Producer services: definition and classification; Regional development models; Resources and development; Transport geography

References


Ecoregions

Tyra A. Olstad
State University of New York at Oneonta, USA

Ecoregions are geographic units distinguished by their place-specific ecological characteristics, including local climate regimes, physiographic features, soil types, ecological history, and, especially, communities of life. They differ from biomes in size as well as specificity: unlike large, general vegetation categories, ecoregions are classified into progressively smaller yet relatively homogenous units in which ecosystems may function similarly to those in neighboring ecoregions or ecosystem types in other biotic provinces, but are unique to their location. For example, the “Adirondack High Peaks” feature a unique mix of mountainous terrain, anorthosite bedrock, foggy, windy, and icy weather patterns and distinct assemblage of boreal and alpine species found only in one particular area in upstate New York; nearby “Highlands” and “Mixed Forests” enjoy milder conditions, while other alpine zones in New England have different geologic histories and compositions.

Delineation and purpose

Government officials – Robert Bailey of the United States Department of Agriculture’s Forest Service (USFS) and James Omernik of the US Environmental Protection Agency (EPA) – began delineating nationwide ecoregions in the late 1970s and 1980s, with the intent to “assist managers of aquatic and terrestrial resources in understanding the regional patterns of the realistically attainable quality of these resources” (Omernik 1987, 118) and provide a new spatial framework for managing the country’s natural resources and protecting environmental quality. Rather than being forced to rely on political borders, agency districts, or single factors such as a watershed or a species distribution, which do not follow or fully reflect the extent and complexity of ecosystems and thus provide poor ecological and spatial context for management decisions, Bailey and Omernik each sought a more holistic, integrative, scalable perspective that would help officials to make decisions based on knowledge of key localized characteristics, driving local forces, and connections between single places and much larger areas. In mapping “ecoregions,” they aimed to identify and account for relationships between biological, geological, hydrological, and climatological phenomena and processes in individual places, in adjacent and/or similar places, and at several different scales. Resulting maps help land managers more efficiently collect samples, estimate ecosystem service productivity, and make decisions based on informed predictions of how key species, physical features, and/or ecological processes will respond to local and/or landscape-scale natural and/or anthropogenic changes.

In the late 1990s and 2000s, nongovernmental organizations such as the World Wildlife Fund (WWF) and The Nature Conservancy (TNC) also took an “ecoregional” approach to mapping global biogeographic units, with the intent to inform conservation efforts and prioritize terrestrial and aquatic “hotspots” (areas that boast high species richness and endemism and face severe threats). Although the WWF and TNC’s maps of
“Terrestrial Ecoregions of the World” (TEOW), Freshwater Ecoregions of the World (FEOW), and, most recently, Marine Ecoregions of the World (MEOW) are, by virtue of objectives and expertise, more focused on issues of biogeography than with general ecosystem services or environmental conditions, they do account for place-by-place variation in and connections between abiotic and biotic (including human) forces – key ecoregional characteristics.

Government officials, conservation organizations, and independent scientists worldwide now use ecoregion maps and GIS shapefiles developed by the USFS, EPA, WWF, TNC, and other agencies and organizations to situate, integrate, ecologically synthesize, and geographically bound their work. Recent work based on ecoregions has included studies of area-source pollution, general anthropogenic impact on specific places, regional comparisons of species assemblages and health, evaluations of the usefulness of ecoregion classification in Europe, Oceania, South America, and southern Africa, and attempts to integrate geographic information systems and remote sensing technology into the processes of ecoregion delineation and monitoring.

Variations

The inherent complexity and dynamism of ecosystems make them difficult to classify and regionalize. In particular, ecoregion mappers struggle to account for microhabitats, ecotones, and dominant processes or factors. Depending on the scale, microhabitats, or localized variations in abiotic conditions and vegetation types, may be acknowledged or separated out into more internally homogenous units, but are generally treated as part of the “patch dynamics” or “mosaic” of larger, more heterogeneous ecoregions. Similarly, ecotones – areas of transition between ecoregions – are occasionally treated as separate entities but are, more often, split between neighboring units. Depending on how and where individuals chose to draw lines, this has resulted in slightly different borders on different ecoregion maps. Most significantly, from both a philosophical and practical perspective, there is no single accepted method for eco-regionalization; methods and methodology vary according to the interests and needs of those who plan to use the resulting maps (Olson et al. 2001; Omernik 2004).

Both the USFS and the EPA use nested hierarchies to inform management decisions on several scales, but they use different techniques for eco-regionalization to fulfill different agency needs. Robert Bailey (with the USFS) took an objective, universally replicable factor-based approach, beginning by identifying areal differences between climate types, then through “evaluation and integration of physical and biological components including climate, physiography, lithology, soils, and potential natural communities” (McNab et al. 2005). The USFS now recognizes four large domains (polar, humid temperate, dry, and humid tropical) in the United States, which are classified into divisions (with more fine-tuned climatic patterns and separate mountain ranges), then provinces, and finally sections (and subsections). Descriptions of the 190 separate ecological subregions in the conterminous United State, which range in size from thousands down to hundreds of acres, also include reference to unique hydrological features, disturbance regimes, land-use patterns, and cultural ecology. Overall, this hierarchical, nested approach allows land managers the flexibility to analyze, manage, and monitor natural resources and ecosystem services within forest landscapes, forest-wide, regionally, and/or nationally.

James Omernik (with the EPA) also took a hierarchical, nested approach to delineating
ecoregions, but he developed and used a more subjective weight-of-evidence methodology that considers how features and processes such as geomorphology, geology, soils, hydrology, climate, dominant vegetation, wildlife, and/or land-use history may play different dominant roles in different places. Using this approach, the EPA, and subsequently the international Commission for Environmental Cooperation (CEC), a partnership between Canadian, American, and Mexican parties, now recognizes 15 Level I ecoregions in North America (based on a mix of broad major vegetation types, landforms, and climatic factors), 50 Level II regions, and 182 Level III regions. Officials with the EPA, other US federal agencies, and state land management agencies have collaborated to further distinguish 968 Level IV relatively small ecoregions in the conterminous United States according to fine-tuned knowledge of localized ecological characteristics and driving forces. Although this place-by-place variation reflects a more nuanced and detailed understanding of each ecosystem and enables environmental protection officials to formulate more place-specific management strategies, it required laborious, less-replicable, expert collection and interpretation of field data (Omernik 2004, S34).

Most recently, researchers with the World Wildlife Fund, The Nature Conservancy, and partnering organizations have undertaken the task of delineating ecoregions worldwide in order to inform the conservation and/or sustainable use of natural resources. They have identified 867 TEOWs within 14 different biomes and 7 biogeographic realms (Afrotropical, Australasia, Indo-Malayan, Nearctic, Neotropical, Oceania, and Palearctic), 426 FEOWs within 12 major habitat types, and, most recently, 232 MEOWs within a hierarchy of 12 realms and 62 provinces. Their approaches were intended to provide “enough detail to be useful in … regional conservation … planning efforts” (Olson et al. 2001) and, especially, to help researchers and conservationists analyze biogeographic patterns and prioritize biodiversity conservation efforts on a comprehensive, global scale. Since completing the foundational work of mapping global ecoregions, the WWF has focused attention on a “Global 200” (now 238, to include aquatic ecoregions) and 36 WWF “Priority Places,” selected according to overall species richness, number of endemic species, taxonomic uniqueness or “extraordinary” adaptations, and global representativeness and rarity, as well as ecological history, ownership, and current level of threat. As Olson et al. (2001, 936) insist, “[c]onservation strategies that consider biogeographic units at the scale of ecoregions are ideal for protecting a full range of representative areas, conserving special elements, and ensuring the persistence of populations and ecological processes.”

Now that the USFS, EPA, WWF, and other organizations have proven the usefulness of defining, mapping, and using ecoregions for management and conservation efforts, many scientists and governmental officials in different countries are beginning to delineate ecoregions in order to address domestic research and management needs on a more place-specific level (on the scale of Bailey’s subsections or even Omernik’s Level IV). In addition to choosing between regionalization methods (nested hierarchy or not, factor-based or weight-of-evidence, what degree of heterogeneity to accept), many are also integrating remotely sensed imagery and geographic information systems into the classification process. While use of these tools and technologies may prove to be more cost-effective and efficient, there are questions as to whether simple overlay procedures are nuanced enough to precisely identify and delineate ecoregion boundaries, especially when operators still have to decide what degree
of internal heterogeneity (“patches”) to accept and how to deal with ecotones.

Concerns and future research

As more scientists, environmental protection and management officials, and conservation organizations use an ecoregion-based perspective to help inform and prioritize decisions, there are lingering concerns to consider. First, the multiplicity of available ecoregion maps (and corresponding shapefiles) provides a great deal of information and affords users flexibility and choices in matters of ecoregion sizes, borders, and distinguishing characteristics, but these options are only beneficial if users are aware of the existence of, and trade-offs between, different versions. While ecoregion maps can help provide a spatial and ecological framework for planning or decision-making, drawing borders and defining characteristics can also give the impression that ecoregions are formal and fixed when they are, in fact, complex, fluid, and dynamic percepts.

Second, ecoregions, by and large, reflect potential natural conditions. Places that have been heavily impacted or modified by humans may no longer feature the original soil type, hydrology, or species assemblages. Even when ecoregion descriptions include references to ecological history and current land use, ownership, and/or threats (especially, agricultural practices and large-scale resource harvesting), classification schemes may not always identify human activities as a dominant or driving factor. “Major Land Resource Areas” designated by the USDA Natural Resources Conservation Service bear similarities to USDA and USFS ecoregions, but are, as yet, separate spatial entities.

Third, as climate patterns change, species invade or go extinct, humans modify and/or attempt to restore ecosystems on local and global scales, and people develop new understandings of ecology and new environmental management practices, ecoregion maps will need to be continually redrawn and their uses reconsidered. In particular, scientists and environmental managers will continue to ask and try to answer questions such as: what tools and techniques for eco-regionalization are most accurate and effective? Can geographic information systems and remotely sensed imagery be used to precisely identify and delineate ecoregion boundaries, and/or how can they help scientists and land managers monitor changes within and between ecoregions? Is it necessary to consolidate current maps into a single, universal classification scheme or is it more useful to have multiple versions with different emphases? Are ecoregion-based management plans proving practical and effective? And, most importantly, how can eco-regionalization help people better understand relationships between multiple complex and changing geographical and ecological processes and phenomena?

SEE ALSO: Areal differentiation (or chorology); Biogeography; Borders, boundaries, and borderlands; Complexity in biogeography; Ecosystem; Environmental management; Land systems science; Regional definition and classification; Spatial concepts

References


**Further reading**


Ecosystem services

Jessica Dempsey  
University of British Columbia, Canada

Ecosystem services, commonly defined as the benefits provided by ecosystems to humans, are an increasingly dominant way of conceptualizing relationships between humans and ecosystems. Many advocating this approach argue that focusing on the benefits that ecosystems provide to people will enlist more individuals and institutions in the conservation of biodiversity and in the project of environmentalism. By focusing on changes in ecosystem services and, ideally, economically valuing them, the argument is that decision-makers (i.e., governments, consumers, businesses) will make more sustainable and economically efficient decisions. For example, by calculating ecosystem services provided by a forest – including water purification services and carbon sequestration – those responsible for decisions about that particular forest will be able to better weigh the tradeoffs between different courses of action, between, for example, approving or rejecting a new soya plantation. Such calculations also prepare the ground for policy options that can facilitate payments between humans that “provide” or “secure” ecosystem service delivery to those who “consume” them.

The ecosystem services revolution, then, aims to make visible and ideally quantifiable the often underappreciated and undervalued ecological structures and processes necessary for human life on Earth. While sometimes oriented toward the production of new markets and commodities, ecosystem services are always matters of calculation, of creating qualitative distinctions in ecosystems (i.e., categories and classifications), differences that can be evaluated quantitatively. The point of this calculability is to organize ecosystems in ways that can render what they do for humans (i.e., the services they produce) equivalent and comparable, with the hope that these quantitative abstractions will incite sustainable and efficient decisions. This is by no means easy, demonstrated by three decades of conceptual and calculative debate on economic valuation of ecosystems.

Ecosystem services are not simply a new name for what ecologists have always studied, but rather are an increasingly dominant approach to ordering and organizing human and nonhuman (or ecosystemic) relations, setting norms for the management and governance of socioecologies. Ecosystem services must be seen as political, an approach that increasingly defines what is of value, contributing to the production of present and future natures. As with many contemporary political, economic, and cultural formations that define how we ought to manage and live, ecosystem services – as a metaphor, scientific practice, model, and policy approach – are largely formulated by experts located in scientific and governmental institutions of the Global North.

Ecosystem services seem to be poised to become a key universal guiding how we ought to live with, manage, and produce nature in the twenty-first century, in the so-called Anthropocene. But as with other “universals,” ecosystem services are highly variegated in definition and in practice, and marked by debates and contradictions.
The rise of ecosystem services

Ecosystem services, as a concept, emerged out of the implosion of two disciplines: economics and ecology. Erik Gomez-Baggethun and colleagues (2010) traced the history of the general notion of ecosystem services to the era of classical economists in the nineteenth century, including Ricardo, Mill, and Malthus, all of whom were explicitly concerned with the services humans receive by “natural agents” (p. 1211). The marginalist revolution of the twentieth century re-oriented focus to the sphere of exchange values, turning away from the non-priced or nonmarketed contributions of ecosystems. This ushered in the heyday of economic arrogance regarding the substitution of goods in face of increasing scarcity (and correlated high prices). The rise of environmental and resource economics in the 1960s provided some correctives to these blind spots and created methods to value environmental costs and benefits, including those without a market price. Departing from this significantly in the 1970s, heterodox economists with systems ecology backgrounds began to actively challenge the substitutability of ecosystems, arguing that there are “limits to growth” (Georgescu-Roegen 1971; Daly 1977). Ecological economics, as the discipline came to be known, advanced the notion that nature is a stock of “natural capital” that provides ecosystem services – a stock that is limited and being depleted.

For ecologists, a focus on ecosystem services began in the early 1980s along with deepening concern about ecological change and biodiversity loss (although thinking in terms of ecosystem functioning has a longer history in ecology). Ecologists Paul and Anne Ehrlich (1981) use the concept of ecosystem services to explain why all of humanity ought to be concerned with biodiversity loss. Their 1981 book *Extinction* put forward the argument that species diversity is central to the Earth’s life support systems, as it provides a basis for healthy ecosystem functioning, which in turn delivers the ecosystem goods and services humanity needs. Their book approvingly cites ecological economist Herman Daly’s focus on the “limits” to growth and the need for a steady state economy and pits this against traditional economic thinking and a global political economy that puts economic growth above all else.

In the 1990s, the concept of ecosystem services became central to conversations between leading ecologists and economists as they attempted to bridge the divide between the concerns of ecologists and the approach of mainstream economics. Many conversations take place at the Beijer Institute on Ecological Economics in Sweden, which undertook a project on the economics of biodiversity in the early 1990s. Within this project, ecosystem services are a key concept facilitating dialogue between ecologists and economists about biological diversity, especially for translating the problem of biodiversity into mainstream economics. The consensus emerging from the Beijer Institute is less focused on limits to growth and more focused on how to optimize and economically manage life on Earth.

This concept reached its high point in 1997. Ecological economist Robert Costanza and his co-authors published an article valuing the world’s ecosystem services at US$33 trillion (Costanza et al. 1997), a paper now cited over 10,000 times. The article led to enormous debate, largely focused on the methodological approach. The same year, ecologist Gretchen Daily (1997), also a participant in the Beijer meetings, published the edited collection titled *Nature’s Services: Societal Dependence on Natural Ecosystems*, a book cited over 4000 times.

Despite this growing elaboration of the concept in the academy, ecosystem services remained largely on the outside of much environmental
ECOSYSTEM SERVICES

law, policy, and advocacy, especially conservation policies and politics. Decisions made under the Convention on Biological Diversity scarcely mention the term until 2006 (see Table 1). Large environmental nongovernmental organizations (NGOs) such as Conservation International and The Nature Conservancy maintained their focus on biodiversity conservation through hotspots and other biodiversity related criteria (although this may be changing). The practice of biodiversity conservation by NGOs, biologists, and ecologists in the 1990s focused on the establishment of parks and protected areas and, in the so-called developing world, on what are often termed “integrated conservation and development projects” – projects that emphasize joint objectives of environmental protection and economic development. While focused on conservation and poverty alleviation, these approaches were not guided by quantitative assessments of ecosystem services.

The broader context of environmental law and policy, however, does take an economistic, market-oriented – often termed neoliberal – turn at this time, when governments began to place increasing faith in markets for achieving wealth and environmental governance. This included the creation of the 1990 Clean Air Act in the United States, establishing a cap and trade system to deal with emissions leading to acid rain. One early, possibly the first, ecosystem service market emerged to enact the broad principles of the Clean Water Act. Wetland banking – a scheme that allows for those who impact wetlands to purchase credits from firms who produce functioning wetlands – emerged to meet the broad principles of this act, while not impeding development (see Robertson 2004). Perhaps most famously, the 1997 Kyoto Protocol, which established binding reductions for developed countries, followed in the footsteps of the Clean Air Act by allowing for “developed countries” to exceed their reductions by buying offsets produced in the Global South through the Clean Development Mechanism. The CDM, while mostly producing credits related to increased efficiencies and fuel switching, does include credits produced via the carbon sequestration services of living systems, defined technically as “land use, land-use change, and forestry” offsets.

In the late 1980s and early 1990s, the elaboration of tort law in the United States to include damages to natural resources is another institutional driver of economic valuation of ecosystem services, as polluters like BP are required to pay for the cost of environmental

<table>
<thead>
<tr>
<th>Conference of the parties</th>
<th>Number of decisions mentioning “ecosystem services”</th>
<th>Number of times words “ecosystem services” found in decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>COP 5 (2000)</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>COP 6 (2002)</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>COP 7 (2004)</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>COP 8 (2006)</td>
<td>7</td>
<td>79</td>
</tr>
<tr>
<td>COP 9 (2008)</td>
<td>10</td>
<td>52</td>
</tr>
<tr>
<td>COP 10 (2010)</td>
<td>22</td>
<td>118</td>
</tr>
<tr>
<td>COP 11 (2012)</td>
<td>16</td>
<td>83</td>
</tr>
<tr>
<td>COP 12 (2014)</td>
<td>13</td>
<td>57</td>
</tr>
</tbody>
</table>
ECOSYSTEM SERVICES

damages like the Deepwater Horizon accident in the Gulf of Mexico, or the earlier Exxon Valdez spill. For example, in order to capture the “passive use values” – meaning the existence value, or sometimes called “nonuse” value of ecosystems – damaged by the Exxon Valdez oil spill, the state of Alaska commissioned a large contingent valuation study that led to a national commission on the methodology a few years later, headed by Nobel prize winning economists Kenneth Arrow and Robert Solow (see Carson et al. 2003 on the Exxon, and Kornfeld 2011 on the Deepwater Horizon). The notion of fungability in ecosystems was also a part of the European Union habitats directive (92/43/EEC) in 1992, which requires protected areas that are degraded to be replaced elsewhere with their equivalent (for a recent review of this policy and its elaboration see McGillivray 2012).

Millennium ecosystem assessment

The Millennium Ecosystem Assessment (MA), released in 2005, is a key turning point in the life of ecosystem services. A monumental work involving more than 1300 scientists over five years, the MA shifted the approach to international environmental assessment by focusing on auditing changes in ecosystems and biodiversity in relation to human wellbeing. The summary findings from the MA are grim. The MA found that 60% (15 out of 24) of the ecosystem services examined are being degraded or used unsustainably, including fresh water, capture fisheries, air and water purification, and the regulation of regional and local climate, natural hazards, and pests. Many of these ecosystem services were found to be degraded as a result of increased supply of other services, like food (MA 2005, 1). Importantly, the MA found that the effects of ecosystem service degradation are felt most by those with the lowest incomes, results that would not surprise geographers studying processes of uneven development.

Besides this assessment, the MA concretized a framework for understanding ecosystems as a stock of natural capital that provide ecosystem services, producing the now ubiquitous categorizations of services: supporting, regulating, provisioning, and cultural (see Figure 1). These categorizations are now commonly used to delineate what nature does for humanity. However, these definitions of ecosystem services are not without debate even among proponents of the approach, as some economists argue that the definition is too broad, and not appropriate for facilitating appropriate economic accounting (see review in Dempsey and Robertson 2012).

For the scientists and policymakers involved in the MA, proper management of ecosystem services is not about tracking human encroachment into some “pristine nature.” More than anything, the central governance problem emerging from the MA centers on measuring and managing tradeoffs between different ecosystem services, for example, impounding streams for hydroelectric power may have negative consequences on fish populations and thus on provisioning services. Or outputs of provisioning services such as timber (e.g., as a result of increased rates of logging) can have a negative impact on the regulation of another service, like the flood prevention service or carbon sequestration service that is provided by forest cover. These tradeoffs are complex and must consider temporal and spatial scales as well as degrees of reversibility (see Figure 2). For example, a tradeoff between maximizing a service like food production now which will impact the service provision later, or maximizing a service in one particular place (upstream) such as hydroelectricity is a tradeoff with the provision of fish in a different place (downstream).
The project of ecosystem services as it is articulated through the MA is most centrally about producing the conditions – in this case, the knowledge conditions or framework – where the value of different courses of action can be quantitatively assessed. The hope is, of course, that such knowledge will lead to optimal decisions and policy options, ones that consider not just traditional commodities, but also previously excluded ecological services. In focusing on tradeoffs among many ecosystem services, together, the approach appears to undermine a focus on achieving “maximum sustained yield” for a single commodity (i.e., maximizing fisheries, agricultural output, or timber production), and rather to maximize yield for a broader array of services that ecosystems provide, beyond what is traditionally priced. Ideally such modeling would also consider the way that consumption of ecosystem services at one time has effects for future generations and to consider the way that the consumption of ecosystem services in one place has effects for people and communities in other places (see Figure 2).

While many governmental, NGO, and private sector institutions welcomed the MA, the reports also generated debates (i.e., McCauley 2006; Reid et al. 2006) and pointed to challenges (i.e., Carpenter et al. 2006). Environmentalists, indigenous communities, and organizations...
as well as governments raised concerns about the utilitarian and economistic framing. For example, writing during a negotiation of the Convention on Biological Diversity just after the release of the MA, a collective of NGOs and activists noted this shift in tone of the negotiations toward the language of the World Bank and World Trade Organization, arguing that the focus should be on the public benefits that ecosystems provide rather than services that might fall under the WTO General Agreement on Trade in Services (ECO 2005).

Critiques come from inside the academy, too, as ecologists and economists alike express concerns. Reflecting on the outcomes of the Millennium Ecosystem Assessment, ecologists, including Steve Carpenter (one of the lead scientists in the MA), note that a barrier to the ecosystem service approach sits with a paucity of ecological knowledge, especially related to how changing quantities of biodiversity, or changing ecosystem structure, relates to changes in ecosystem functioning and services (Carpenter *et al.* 2006). Richard Norgaard (2010) provides a more fundamental critique of the ecological-economic model of the MA, taking issue with the way the MA conceptualizes nature as a “stock” that “flows” ecosystem services to humans. Norgaard asserts that the social and ecological problems of our time require a diversity of ecological approaches, further arguing that the stock-flow model fits too well with status quo institutional and deeply uneven power relations, and fails to question the broader political-economic obsession with economic growth. Many others, including ecologists (i.e., Adams and Redford 2010) and political ecologists (i.e., Sullivan 2009), focus on the perils of economic metaphors and practices that they say will lead to further inequities.

**Economic valuation: toward “the economics of ecosystems and biodiversity”**

Dogging natural resource, environmental and ecological economists for three decades is the question of how to value nonmarketed aspects of ecosystem services. The debate following the release of Costanza *et al.*’s (1997) paper that valued all of the Earth’s ecosystems demonstrates this. Costanza and his colleagues’ impulse was aggregative and summative, which led to loud criticisms from marginalist economists who saw his 33 trillion as “breaking the rule”: who could buy the whole Earth? Rather, these critics argued that valuation must focus on establishing credible market prices based on supply and demand (see Dempsey and Robertson 2012). There remains debate within the ecosystem services literature about valuation, including, for example, debates over the ethical implications of discount rates and the role for nonmonetary valuations (i.e., Heal 2000; Ludwig, Brock, and Carpenter 2005; TEEB 2010; Luck, Chan, and Fay 2009).
In response to these debates and gaps in knowledge, and also following on the heels of the Stern report and the Millennium Ecosystem Assessment, in 2007 the G8 environmental ministers created The Economics of Ecosystems and Biodiversity (TEEB) project to study the economic impact of the global loss of biodiversity. The head of the initiative was Pavan Sukhdev, a former international banker, whose writing and talks sometimes sound revolutionary. For example, he states that “the root causes of biodiversity loss lie in the nature of the human relationship with nature, and in our dominant economic model,” going on to say that our current economy “promotes and rewards more versus better consumption, private versus public wealth creation, human-made capital versus natural capital” (TEEB 2010, xviii, emphasis in original). Yet, despite this widespread criticism of contemporary Western culture and economies, in the next breath Sukhdev states that the main problem is that of externalities and market failure: “there are no “markets” for the largely public goods and services that flow from ecosystems and biodiversity” (TEEB 2010, xxi).

The first report emerging from TEEB in 2008 valued the cost of inaction on biodiversity loss and ecosystem change – in terms of lost ecosystem services – at about 50 billion euros per year. If current rates of ecosystem change and biodiversity loss continue, the report estimates that by 2050 (from a base of the year 2000) the opportunity cost in terms of lost ecosystem services would weigh in at 14 trillion US dollars. TEEB ended in 2010 after producing a number of reports that drew attention to the global benefits of biodiversity and the costs of biodiversity loss. Reports targeted specific end users: international and national policymakers, subnational policymakers, business and enterprise, and individuals.

These TEEB reports recognize the challenges, uncertainties, and politics of economic valuation. Yet they unequivocally state that economic valuation “of Nature’s public goods and service flows is both necessary and ethical” (p. xxi) in order to communicate with decision-makers “using the language of the world’s dominant economic and political paradigm” (p. xix). In pursuit of better valuations, TEEB produced a book (TEEB 2010) that synthesized the foundations of ecological-economic valuation. While this text recognized the wide variety of approaches to valuation, overall it largely consolidated a neoclassical (preference-based) approach to valuation, which is amenable to assessing tradeoffs between different uses of land and ecosystems (recall the role of tradeoffs in the Millennium Ecosystem Assessment).

A crucial debate in valuation is how marginalist approaches can be used to value the resiliency of ecosystems in conditions that are increasingly nonlinear with changing climates (i.e., that have thresholds where the system changes dramatically) (see TEEB 2010, ch. 5). Gretchen Daily (1997) and her co-authors note the limitations of measuring the additional value (or the reduction in value) in ecological systems that are “highly interdependent” and how “seemingly small changes in one place cause large impact on the overall system” (p. 396). When ecological thresholds or boundaries are approached, then changes in ecosystem services cannot be measured using traditional marginalist approaches: how does one value even a small change that leads to the collapse of an ecological system? Given this, TEEB (2010) warns that traditional valuation is not appropriate when approaching “ecological thresholds” (p. 223), and Ring et al. (2010) argue that there are cases where the “the basic theorems of welfare economics are not valid” (p. 17). These contradictions are recognized by scholars and experts, but often...
ECOSYSTEM SERVICES

not discussed when economicist approaches are embraced in the mainstream.

And how do new environmental markets appear in the work of TEEB? The written books are careful to note that the project is not about market-making, for example, stating that “[p]lacing blind faith in in the ability of markets to optimize social welfare, by privatizing the ecological commons and letting markets discover prices for them, is not at all what TEEB is about” (TEEB 2010, xxiv). Yet geographers Ken MacDonald and Catherine Corson (2010) argue TEEB is indeed part of a project of environmental market-making. Based upon participant observation at international meetings, they find that TEEB is central to “legitimating and circulating the narratives, images and ideas of nature essential to these new speculative nature markets” (p. 181). With the current decline in international carbon markets, it is unclear what the next decade will bring in terms of market-making and growth, and it seems crucial to recognize that ecosystem services does not equal market-making.

Emerging practices of ecosystem service assessment and policy

While there are ongoing debates and challenges to the concept, ecosystem services are now firmly found within the discourse and increasingly in the practice of the world’s largest conservation organizations, known as BINGOs (big international non-governmental organizations): the World Wildlife Fund, The Nature Conservancy, and Conservation International (CI). But taking ecosystem services from a model and conceptual framework into practice, policy, and politics is an ongoing project and one with enormous variation.

Calculating ecosystem services

Even if it is becoming common to speak in the language of ecosystem services, calculating those services with any ecological credibility is a whole other matter. To convince decision-makers to invest in ecosystems (i.e., protect or restore them), ecosystem services need to be more than a heuristic; they need to be incorporated into the cost–benefit or due diligence models of decision-makers (governmental, corporate). Ecosystem services must be turned into quantified flows that can be represented in numbers, ideally dollars and cents. This problem is widely recognized in the ecosystem services literature by proponents of the concept and approach. In response, academics (including not least ecologists and economists), conservation organizations, and international institutions are creating spatially explicit models that assess the flows of services from particular land and marine assemblages, biophysically and economically. This includes the integrated valuation of environmental services and tradeoffs (InVEST) tool produced by the Natural Capital project (a collaboration between Stanford University, the University of Minnesota, and two conservation organizations: The Nature Conservancy and WWF-US). InVEST is an open source computer modeling software tool that quantifies, maps, and values ecosystem services. To refer back to the conceptualization of the Millennium Ecosystem Assessment (above), InVEST aims to make quantitative the tradeoffs associated with different courses of action; it seeks to organize and structure encounters between humans and the “calculative goods” of ecosystem services – in the present and future – to compare and contrast different courses of action in land and marine use.

With the (idealized) knowledge of how increasing one particular ecosystem service (e.g., soya production) impacts on other ecosystem services (e.g., carbon sequestration, water quality,
pollination services) in the short and long term and over various spatial scales, this integrative framework is meant to help decision-makers optimize allocations of resources. In rendering the costs and benefits of different land uses visible and quantifiable, InVEST aims to give the state, or decision-makers, the tools to govern environmentally and economically at the same time, to design policies that could maximize value. InVEST is now being used throughout the world, for example, for water security in Latin America, to design coastal protection in the Gulf of Mexico, and aiding China’s new systems of protected areas.

Given the difficulties of understanding ecosystem functioning and services, one significant question is how such models can provide accurate quantitative assessments (see Johnson et al. 2012) especially with the uncertainties of climatic changes. This is a tension in the turn to ecosystem services. While ecologists increasingly adhere to nonlinear and surprise-ridden models of ecosystems (especially with changing climates), the turn to ecosystem services requires them to nail down interactions and outcomes – ecological and economic – that can be rendered quantitative now and into the future.

Ecosystem service policies and markets

Given these internal debates on how to define, calculate, and value ecosystem services (economically and ecologically), it is not surprising that the practice of ecosystem services policy is difficult to characterize. Implementing the ecosystem service framework now includes national accounting strategies all the way to full markets in carbon sequestration ecosystem services. The most developed policy framework is known as PES – payments for ecosystem services – a term that refers to monetary incentives given to landowners (or tenure holder) to manage their land for a particular ecosystem service (e.g., carbon sequestration or water filtration). PES programs are sponsored by a wide range of institutions from international development agencies to ENGOs to governments and for-profit firms. Although often characterized as market-based, not all PES are market-oriented, and many operate more like state subsidies. For example, these exist in the form of agri-environmental subsidies schemes, where farmers get paid in Europe producing landscape “services” rather than only crops. Such schemes are quite different than markets in wetlands such as those in the United States (discussed above) or markets in forest or ecosystem carbon sequestration where there are buyers and sellers and stronger assumptions about ecological substitutability and fungibility. This has led some to categorize policies in terms of “payments” and “markets” for ecosystem services – PES and MES, respectively. For further discussion on the distinctions between incentives and markets see Pirard (2012), Muradian et al. (2013), Vatn (2014), all of whom argue that markets are a poor way to describe much of what is going on in ecosystem service policymaking.

Indeed, many policies seem to be a mix of state-funded incentives and market-like governance. For example, McAfee and Shapiro (2010) describe Mexico’s PES program as hybrid neoliberalism in that it combines market norms with state-sponsored rule making and institution building (p. 586). Rather than straightforward commodification or the operationalization of market efficiency, McAfee and Shapiro find that Mexican PES policies are complex and contradictory and morphed into a hybrid program due in large part to the efforts of activists and campesinos who wanted to see the program move in a pro-poor direction. In a detailed examination of Costa Rica’s PES program, Fletcher and Breitling (2012) argue that it
operates more like a subsidy, as a way to support forest conservation through government funds distributed to forest owners (see also Vatn 2014).

Ecosystem service policy design, particularly in the Global South, depends on the goals of the program: are practitioners trying to achieve “efficient allocations” of services, or trying to alleviate poverty? Or both? Are they a part of a global resource transfer, or a way to deal with national or regional issues? Within critical geography and other social sciences, much work focuses on empirically examining and theorizing the terrain of payments and markets for ecosystem services. There is enormous variation in approaches and sites of empirical studies, but many researchers are now focusing on examining new relations of property emerging within payments and markets for ecosystem services, drawing out new (and old) dispossessions and hierarchies in the wake of such policies (see review in Dempsey and Robertson 2012).

Conceptually, much critical social science examination of ecosystem services draws from Marxist analysis and on the production of new commodities. For example, Kosoy and Corbera (2010) argue that payments for ecosystem services should be understood through the lens of Marx’s commodity fetishism. They draw out how payments/markets for ecosystem service projects disregard ecosystem complexity in pursuit of market transactions, obliterating other social and ecological qualities, values, and relationships, producing (not ameliorating) new socioeconomic hierarchies. Robertson (2012) draws from Marxist conceptions of value, arguing that the turn to ecosystem services must be understood as a transformation akin to that of labor under the advent of capitalist relations.

Proponents of ecosystem services are also expressing doubts about the feasibility of market policies. The Millennium Ecosystem Assessment led Reid et al. (2006) to note the limits of environmental markets. With the small scale of many ecosystem service markets, and high transaction costs, some wonder if the market approach achieves any efficiency gains (i.e., Kroeger and Casey 2007; Muradian et al. 2010).

National accounting standards

For years environmentalists have bemoaned the narrowness of gross domestic product (GDP) for measuring the health of nations and citizens. Drawing on the work of economic valuation, there are moves afoot within the World Bank and many governments to account for a broader range of ecosystem service values in national balance sheets. In 2012, governments adopted an internationally agreed upon method to account for natural resources like minerals, timber, and fish (known as the UN Statistical Commission of the System for Environmental and Economic Accounts (SEEA)). Going beyond natural resources, the World Bank is now leading the “Wealth Accounting and the Valuation of Ecosystem Services” (WAVES) partnership that aims to bring a broader suite of ecosystem services into national economic accounts and to promote and support this within developing nations.

At Rio +20 in 2012, 68 countries signed the Natural Capital Accounting declaration along with many multinational corporations (such as Walmart and Nestlé). Over 44 financial institutions and firms have now signed the Natural Capital Declaration, a finance sector initiative to integrate natural capital considerations into loans, equity, fixed income, and insurance products as well as in accounting, disclosure, and reporting frameworks. However, incorporating ecosystem changes into the quantitative models of the financial and corporate worlds is incredibly challenging (Dempsey 2013).
Ecosystem services in the big picture

A new universal

For many, ecosystem services represent a kind of new universal, a unifying concept that can finally bring environmentalism and conservation into the financial calculations of the private sector and the deliberations of governments, operationalizing the ever-elusive “green economy.” Ecosystem services, as Tallis and Kareiva (2006) argue, are “one of the few ideas that resonate in corporate and governmental board rooms, on stock exchanges and in farmhouses, mud huts, eco-tourist lodges and palm palapas” (p. 748). Proponents of ecosystem services often promote the approach in relation to the failures of biodiversity conservation: an approach that they say does not resonate with anyone but the converted or that separates humans from ecosystems.

Yet there are crucial commonalities between ecosystem services and other material-semiotic orderings, like biodiversity and even wilderness. All are concepts still largely produced in the same institutions: academics based (or trained) in the Global North, conservation organizations whose modus operandi is still oriented in the Global North, and institutions like the World Bank. In other words, the production of a world of ecosystem services is, in many ways, a reiteration of the (colonizing) concept of Western science and experts to discern from a distance the laws of, in this case, socioecological natures. As Sullivan (2009) asks, “what knowledge and experiences are being othered and displaced through the parlance and practice of ecosystem services?” (p. 23). In other words, what new and old orientations of power and knowledge does the material-semiotic object of ecosystem services facilitate? The turn from biodiversity toward ecosystem services ushers in new hierarchies and rankings with implications for what kinds of socioecologies will be invested in, or not. Will these be less violent? And, if so, for whom?

In addition to the work of critical scholars, there continues to be social movement and activist opposition to the ecosystem services turn. At World Trade Organization negotiations in late 2012, a statement released by the Indigenous Peoples Movement for Self-Determination and Liberation called for “the halt of all policies controlling the reproductive capacity of Mother Earth through market-based mechanisms that allow for the quantification and commodification of the natural processes of Mother Earth being branded as ecosystem services” (IPMSDL 2013). And there are signs that proponents of ecosystem services recognize this conceit. Indeed, a recent set of articles in ecology draw out some of these questions, focusing on the ethics of ecosystem services, noting explicitly that ecosystem services are only one metaphor or approach, and it may not be “culturally appropriate” in all cases (Raymond et al. 2013). Yet, despite these challenges, ecosystem services are being embraced in international policy, national governments, multinational firms, and financial institutions. This suggests that the approach is increasingly in the hands of institutions that have benefited most from ecosystem service decline, raising serious questions about how this particular metaphor/model can avoid the colonizing, violent and reductionist tendencies that hampered other environmental metaphors and models.

Will it work?

Linked to the above, there is the question of whether the approach can contribute to lessening ecological impoverishment; in other words, if this new environmental formation of ecosystem services will change decisions toward sustainability and if it can make the green economy a
realism. Bringing the changes in ecosystem service into conversation with political-economic trends tells us that it is certainly no panacea. Indeed, this is what a broadly based geographical political economy brings to the question of environmental change. When one examines power relations globally, one has to wonder if the ecosystem services turn results in little more than “tinkering” as Norgaard suggests (2010). Decisions – for example, to build new energy infrastructure to extract fossil fuels (pipelines, refineries) or the continued destruction of mangroves for shrimp farms are not likely to be reversed due to new calculative figures about the costs (for example, the reduction of ecosystem services resulting from ocean acidification and mangrove destruction). While economically “stupid” or “irrational” decisions may haunt us in the future, those stupid economic decisions pay in the present. A recent study commissioned by The Economics of Ecosystems and Biodiversity (TEEB) is revealing: the study tallied up the globe’s total unpriced natural capital (ecological materials and services for which businesses currently do not pay), like clean water and a stable atmosphere. It found that none of the world’s biggest businesses would be profitable if they had to pay for those services (Trucost 2013). “Irrational” economic decisions continue not only because we lack prices and values of natural capital, but also because many large, powerful (and currently profitable) industries, institutions, and governments depend deeply on these externalizations. Changing these externalizations means coming up against these deeply embedded structures.

Put another way, a major problem facing ecosystem service tradeoffs remains the same as it was for biodiversity conservation: who will pay the price of shifting to better, more “rational” decisions? How can we convince the state, for example, to stop approving business-as-usual energy infrastructure? A big question for ecosystem service proponents – scientists, experts, conservationists – is whether, and if so, how, this particular concept and model can be mobilized to break these deep political-economic entrenchments, and/or help us deal with “humanity’s shriveled ecological options and gross social injustices” (Norgaard 2010, 1224).

SEE ALSO: Biodiversity; Conservation and capitalism; Environment and resources, political economy of; Environment and the state; Environmental governance; Environmental knowledges and expertise; Environmental policy; Environmental science and society; Environmental valuation; Environmentalism; Environmentality and green governmentality; Green capitalism; Nature conservation; Neoliberalism and the environment; Political ecology

References


ECOSYSTEM SERVICES


“Ecosystem” is used in three ways. Following Pickett and Cadenasso (2002) these are the dimensions of meaning, model, and metaphor (these follow the three identified by Golley 1993, 72, as the early, undifferentiated uses: a particular object, an emphasis on energy and material transfer, and another word for ecology). First, it is a term that can mean a type of environment with characteristic plants, animals, soil and climate, or a particular instance of this type in a particular area. In the general part of that usage it is similar to the definition of biome and is used interchangeably, as the tropical savanna biome/ecosystem. In a particular instance “biome” would not be used, but such individual usage is common for areas such as the ecosystem of a particular lake, valley, or other geographic area. This usage is part of a hierarchy of terms related to environmental biology. Second, the “system” is emphasized and the word would apply to the structure of interactions and flows of energy and matter among the biotic and abiotic components of a particular area. Models are necessary to apply ecosystem to a particular instance. As metaphor, ecosystem is a rhetorical device to emphasize connectivity of humans with the environment. All usages derive from the original coinage (as a historical footnote, Willis (1997) reported that a young colleague of Tansley, A.R. Clapham, claimed decades later to have coined “ecosystem” some time before 1935) of the term by Tansley (1935; italics in original).

I have already given my reasons for rejecting the terms “complex organism” and “biotic community.” Clements’s earlier term “biome” for the whole complex of organisms inhabiting a given region is unobjectionable and for some purposes convenient. But the more fundamental conception is, at it seems to me, the whole system (in the sense of physics), including not only the organism–complex, but also the whole complex of physical factors forming what we call the environment of the biome – the habitat factors in the widest sense. Though the organisms may claim our primary interest, when we are trying to think fundamentally we cannot separate them from their special environment, with which they form one physical system.

It is the systems so formed which, from the point of view of the ecologist, are the basic units of nature on the face of the earth. Our natural human prejudices force us to consider the organisms (in the sense of the biologist) as the most important parts of these systems, but certainly the inorganic “factors” are also parts – there could be no systems without them, and there is constant interchange of the most various kinds within each system, not only between the organisms but between the organic and inorganic. These ecosystems, as we may call them, are of the most various kinds and sizes. They form one category of the multitudinous physical systems of the universe, which range from the universe as a whole down to the atom.

The hierarchical use, although Tansley might have objected, is still close to “biome” but with the climate and soils of the biome included (perhaps as a modern usage of “biome”); “biome” and the major biomes of Earth have separate entries in this encyclopedia). The narrow systems use has a history of development and a determination of scientific research all its own. This usage is conceptually important because it is a constant reminder of the connections of the
ECOSYSTEM

biotic and abiotic components. With this connection ecosystem is convenient as a metaphor when the connections may not be obvious but need to be reiterated as in some instances of environmental impact communication.

A biological hierarchy

Tansley did not specify all elements of a hierarchy, but he placed “ecosystem” somewhere between atom and universe. Odum (1959, 6) place ecosystem more specifically in a ten step hierarchy from protoplasm to biosphere, near the top: … – Community – Ecosystem – Biosphere.

Here ecosystem fits where others might place “biome,” but unlike biome it is not restricted to a large spatial extent or similar spatial scale. Ecosystem becomes multiscale because it is at the transition point between a biological hierarchy and a geographical one, and as a result is used in either sense and scale is dropped. Populations and communities can be aspatial, but in the hierarchy an ecosystem cannot (but in systems terms, below, it can be) – but the space can vary. In this sense the ecosystem of a nearby lake has a clearly defined location and extent and, while the location of the tropical savanna ecosystem may not be clearly defined, it is conceived as occupying some location of great extent on the planet. The ecosystem of a droplet of water may have a specific location and clearly has a specific extent. Ecosystem becomes a useful summary of the organisms and their environment and whatever scale.

The number of such ecosystems (or biomes) is undecided. Biogeographers will count roughly 7 to 15 biomes depending on predilection for fine division. Whittaker (1975, 135–161), blurring a distinction, identified 36 “biome types” as “major kinds of communities,” but many of these seem more like ecosystems, and include:

27 Fresh-water lentic communities (lakes and ponds) …
28 Fresh-water lotic communities (streams) …

Because ecosystem can be used at many scales, perhaps counting is irrelevant (if any water droplet has its own ecosystem). Given current interest in microbes, and microbiomes, the smallest extents of ecosystems may be shrinking. Whether we will examine the ecosystems of larger planets can be left to science fiction, for now.

Systems ecology

The narrower usage developed through a focus on the “constant interchange” in a system in question. The ecosystem of a particular lake may be in question, or a generalized ecosystem concept for lakes, but in either case a model system is defined. The interchange became focused on the flows of energy and matter among components in which the biological populations and communities were made abstract – much more so than in the population ecology based view. The types of flows can be seen in abstractions of the models, as in Figure 1, historically among the most influential.

The systems being flows of energy and matter may in part have been shaped by the interests in the potential applications of ecosystem concepts to the development of nuclear weapons and energy: much of the early funding (and relatively high funding within ecological research at the time) was from the US Atomic Energy Commission and was related to bomb testing and the development of experimental nuclear facilities. Tracing radioactivity in the environment was a concern and a systems approach in which biotic populations and communities were abstracted (and thus to some degree became abiotic, although Golley pointed to the biological
orientation as reason for why other disciplines, most notably geography, did not adopt the approach; see Malanson 2011) addressed this concern with models. This direction in research was enabled by the development and accessibility of computers in the 1950s. The models could be quantified and used to project outcomes for different scenarios. By the early 1960s, systems ecology could be touted as the “new ecology” with notable disdain for the more biological population-based ecology and especially for natural history.

This early period of growth was sustained by the International Biological Program, which fostered a systems approach in ecosystems (as defined above) for explanation and comparison. Under this program new computer simulations developed that are the foundations for current systems ecology. Today, systems ecology is used to address more complicated ecosystems (e.g., urban ecosystems; see Pataki et al. 2013), but the focus is still on the flux of energy and nutrients. Ecosystems scientists keep track of energy, often using carbon (e.g., gC m$^{-2}$ year$^{-2}$) as the currency of measurement because it is easier than explicit measures of joules, and nutrients (e.g., nitrogen) in increasingly complicated forms and processes.

Core ideas in systems ecology include (i) energy flow – the flow of energy (flow, given the first and second laws of thermodynamics)}
leads to structure as organisms evolve to use it to reproduce. The ecosystem structure that develops is called a food web. Food webs, not so identified, were studied before the term ecosystem appeared (Summerhayes and Elton 1923; see Figure 1), and the complexity led to a simplification in which the multitude of species were grouped into a hierarchy or trophic structure with a simpler food chain. The trophic structure is based on net primary productivity, primarily through photosynthesis, by the group of autotrophs. Higher levels, of heterotrophs, in the structure are occupied by primary consumers (herbivores) and secondary, tertiary, and so on, consumers (carnivores and omnivores), but with a limit to the number of levels because the amount of energy that is available to a higher level drops by 90% (a rough estimate that will vary widely among ecosystems) due to entropy, with loss of energy as heat through the processes of living.

(ii) Material cycles – the flow of energy drives cycles of matter. Ecosystem scientists concentrated on the cycle carbon because it was a way to track energy (entering the biota through photosynthesis), but the carbon cycle has become important in its own right because of its connection to climate change. Nutrient cycles have been a core study of ecosystems with macronutrients such as nitrogen phosphorous receiving early attention, but micronutrients now being examined. The detailed transformation of the molecular characteristics of the elements in question and how these are produced by or affect organisms is quantified. The biotic part of the hydrological cycle has also been a major component of ecosystem science. This systems focus tied closely to developments in other environmental sciences, notably hydrology.

Systems ecology is closely linked to modeling, as the transfers of energy and material are tracked and then projected in computer simulations. Intensive modeling efforts were the core of the International Biological Program, but were taken up by other projects with different conceptual structures and core epistemologies (Golley 1993).

Metaphor

The popular use of ecosystems is important because it is popular in the sense of reaching out to all (a Google search returns >50 million hits, although many are used as analogy as with business ecosystems or the Android ecosystem). With an increasing awareness of environmental consequences of human activities (perhaps part spurred by the atomic research) that developed in the 1960s, inspired by Rachel Carson’s Silent Spring and leading in the United States to Earth Day in 1970 and several related pieces of legislation, ecosystem became a shorthand for the connectedness in the environment at many scales. Although a kind of imprecise hand waving, this usage has rhetorical weight because it reiterates Tansley’s view of connectedness. In this sense, even short of what is seen for analogies, ecosystem is metaphorical.

This use is important for nonscientists to understand the place of people, every individual’s place, in the environment. At one end of the scale used, people are important components of the Earth’s ecosystem. While people are tied to smaller scale ecosystems of their locales these can be difficult to see in an urbanizing world, but the connection to a global ecosystem, and the power and importance of the metaphor, are strengthened in a globalizing world.

SEE ALSO: Boreal forest ecosystems; Mediterranean-type ecosystems; Polar region ecosystems; Temperate forest ecosystems;
Tropical rainforest ecosystems; Tropical savanna ecosystems

References


Further reading

**Edge city**

**Richard P. Greene**  
*Northern Illinois University, USA*

An *edge city* is an employment center on the edge of an urban area. The term was first coined by Joel Garreau, a *Washington Post* journalist who ran a series of articles focused on edge cities including Tysons Corner Virginia, which was the model for Washington DC. He then published the book *Edge City* (1991) in which he says that an edge city:

- has 460,000 m² (5 million ft²) or more of leasable office space;
- has 56,000 m² (600,000 ft²) or more of leasable retail space;
- has more jobs than bedrooms;
- is perceived by the population as one place; and
- is not considered part of a “city” as recently as 30 years ago.

To Garreau, who identified edge cities for many urban areas in the United States, an *edge city* looks and feels like a new city. It is often found in the suburbs near a highway interchange, having office and retail space, and perceived as one place. Edge cities take many shapes and sizes, sometimes forming linear series or corridors — for example, one outside Chicago, which encompasses O'Hare International Airport, Woodfield Mall, and many office parks (see Figure 1). Its status was elevated in 1991 when Sears moved its corporate headquarters and 5000 jobs to the Hoffman Estates community from the company’s former headquarters in the 110-story Sears Tower (Willis Tower today) in downtown Chicago. People are now accustomed to edge cities because besides being places where people work, they provide diverse shopping and entertainment opportunities. In an *edge city*, regional malls might be intermixed with commercial corridors and upscale hotels, often present for out-of-town visitors who are on business in the *edge city* or even for those who live in the region who are looking for a weekend escape. Many of these characteristics sound much like the ones ascribed to big downtowns because, in fact, edge cities are providing something of a downtown atmosphere.

Although Garreau coined the term *edge city*, the concept itself had been studied earlier. For instance, the multiple nuclei model of urban development by Harris and Ullman (1945) recognized the phenomenon. Their model depicted growth and development occurring around multiple centers of economic activity in a metropolitan area, rather than around a single central business district (CBD). Urban geographers in the 1970s, 1980s, and 1990s were also active in monitoring both the evolving forms and the impacts of these outlying employment centers. Baerwald (1978) provided precise delimitation of an evolving suburban employment center outside Minneapolis through the use of aerial photographs and other data. Muller (1981) conducted a broad and comprehensive analysis of the trends of suburban nucleation with a case study on Philadelphia’s King of Prussia. Erickson (1986) modeled the process of multinucleation and found that the spatial distribution of employment was becoming more random. Cervero (1989) examined a number of traffic congestion-related issues for the 50 largest suburban employment centers and thus situated the concern in the urban planning literature.
Figure 1  Top 12 employment centers in Chicago area ranked by jobs in 2010.
Contemporary studies have attempted to develop better empirical methods for defining and comparing edge city growth through time (Giuliano and Small 1991; Fujii and Hartshorn 1995; Greene 2008; Antipova and Ozdenerol 2013). Giuliano and Small (1991) delimited suburban employment centers in Los Angeles using a job-density criterion – their cutoff to identify an employment center was a zone with at least 10 jobs per acre, or 6400 jobs per km². Another useful measure to delimit edge cities is the ratio of employment to resident workers (E/R ratio), reflecting the balance of workers and jobs. The ratio is defined as follows:

\[
\frac{E}{R} = \frac{\text{Number of workers working in area}}{\text{Number of workers residing in area}}
\]

Edge cities have been identified in terms of places with high E/R ratios, a criterion that corresponds to the third and fourth parts of Garreau’s (1991) definition.

Edge cities play an important role in helping direct the future of sprawl as they represent independent and competing employment hubs to the long-established downtowns of metropolitan areas (Shearmur and Motte 2009). As edge cities continue to grow and expand, the urban–rural fringe has the potential for expanding much farther outward than it did in previous cycles of growth, when the downtown dominated. This expansion is occurring as an individual figures his or her acceptable commuting range not to the long-established CBD, but to an edge city where he or she holds a job.

SEE ALSO: Cities and development

References


Egypt: Al-Jam‘iyya Al-Jughrafiyya A-Massriyya (Egyptian Geographic Society)

Founded: 1875
Location of headquarters: Cairo
The first work of electoral geography is generally considered to be André Siegfried’s 1912 Tableau politique de la France de l’ouest sous la troisième république (A Portrait of Politics in Western France under the Third Republic), although the Anglo-American tradition often identifies Edward Krebheil’s 1916 analysis of British parliamentary elections as the earliest work. Both Siegfried and Krebheil ask similar questions and employ similar methods: What is the spatial pattern of a vote, and what explains such patterns? Both ultimately sought to make a causal argument about the source of political preference and voting behavior. In essence, they compared maps of election returns with physical or social factors with similar spatial variations, and interpreted similarities in spatial patterns as evidence of causality. Siegfried focused on social variation caused by different physical environments (geology), while Krebheil identified variations in occupation as the key factor.

Subsequently, in the United States during the 1930s, John K. Wright introduced the mapping of legislative (congressional) votes, in addition to the mapping of election returns. This was an important conceptual innovation because it focused concern on political relationships per se, rather than using election results as a proxy for public opinion.

Early electoral analyses were undertaken before the advent of modern survey methods and before the invention of computers and information technology powerful enough for systematic spatial analysis. This meant that election returns were the best – if not the only – source of information about public opinion and its variation. Without systematic data on individual responses, however, causality could only be inferred by spatial coincidence. Hence Krebheil interpreted his findings (that industrial and less wealthy agricultural regions tended to support the Liberal Party while wealthier agricultural areas were Conservative bastions) to mean that there was a causal relationship between occupation and political preference. This relationship could be affected, however, by party positions on specific political issues; for example, support for Home Rule in Ireland explained the higher than expected support for the Liberal Party in southern Ireland. Such analyses were not statistical, however, and relied primarily on the visual inspection of maps. They also suffer from the ecological fallacy, inferring individual characteristics (and causal mechanisms) from aggregate data.

While techniques of electoral analysis have grown more sophisticated, the basic approaches, questions, and limitations in contemporary electoral geography remain similar to these early works.

The geography of voting

Until the advent of sufficiently powerful and inexpensive computer technology (roughly until the early 1990s), the task of mapping election returns remained relatively difficult. This gave academic geographers something of a monopoly on electoral cartography. Work in the 1930s by Wright showed maps of US presidential
ELECTORAL GEOGRAPHIES

elections, with some inferences about the explanation of persistent patterns. The main thrust of such work, however, was the visualization of election results and the identification of spatial patterns. This approach enjoyed a revival in the early 1980s with the work of geographers such as J. Clark Archer and Peter Taylor which emphasized the persistence of regional cleavages in American presidential elections. Indeed, Archer and Fred Shelley coauthored a “Resource Publication” of the Association of American Geographers (Archer and Shelley 1986, 87) that declared: “the task of the political geographer is the identification and interpretation of the sectional manifestations of political issues.”

Even at that time, however, academic geographers were losing their monopoly on electoral mapping, and high-quality electoral maps became common in the popular print media. Although the scholarly works use sophisticated statistical analyses to identify spatial patterns, they remain primarily concerned with where votes are cast. As information and GIS technology improved through the 1990s, it became even easier for media organizations to create maps showing election returns, often in real time. The ubiquitous red–blue map of the 2000 US presidential election cemented this practice.

Moreover, gender – a major determinant of political preference – tends to be spatially uniform and thus invisible to this kind of electoral analysis. With polling and survey data, the influence of these various factors can be assessed directly, so drawing causal inference from spatial patterns becomes a poor alternative. Consequently, scholarly attention to individual causality has shifted largely to the analysis of survey data in political science. Although Ron Johnston and his collaborators use polling data extensively for electoral analyses, the practice is limited among other electoral geographers.

Moving beyond the compositional effects shown by the geography of voting has been one of the major challenges for electoral geography. As Agnew wrote, “‘Mapping politics’ can offer more than cartographic illustrations that decorate more compelling aspatial accounts of electoral geography, but only if we work harder at understanding the roles of context and showing in what ways it counts” (1996, 144).

Geographic influences on voting

A second major theme in electoral geography is geographic influences on voting. Interest in this issue appears in the earliest works in electoral geography: What causes people to vote differently in different areas? As geography embraced statistical and quantitative techniques in the 1960s, it became possible to answer that question in a more systematic fashion. Johnston (2005, 581) has repeatedly articulated a contemporary version of this question: Do “similar people vote in the same way wherever they live?” Indeed, the question of whether place matters is one of the major points of dispute between political science and political geography (Johnston and Pattie 2006, 40–43). Political scientists such as Gary King argue that political behavior and
preference can be explained by purely individual factors, and that unexplained spatial variation is merely the consequence of imperfect knowledge about individual voters, that is, compositional effects. John Agnew (1996), Charles Pattie and Ron Johnston (2000) – the latter two typically working together – have carried on a spirited debate with this position, arguing for the contextual approach and for the existence of the neighborhood effect.

The contextual perspective asserts that an individual’s political preferences and behavior cannot be understood or explained by looking simply at her or his individual characteristics (income, occupation, age, etc.) but depend to some degree on place and location. For example, early work – ironically by political scientist V.O. Key – found that support for candidates was relatively greater in areas close to the candidate’s homes, an approach and finding that is still replicated in contemporary geography.

Contextual effects go beyond mere distance, however. Johnston and Pattie (2006, 44) identify at least eight scales (from the household to neighborhood to country) that may constitute the place of an individual and that can influence his or her vote. One can imagine, for example, that a booming economy may encourage a favorable view of the incumbent party among voters in general, but may have little effect if one’s spouse is recently unemployed. Most commonly, however, contextual effects are investigated at the neighborhood scale, hence the so-called neighborhood effect. Such an effect can result in greater homogeneity through person-to-person interaction, the adoption of common role models, and the influence of institutions (common schools, churches, etc.) Work by other scholars shows, for example, that the turnout of minority groups increases as the proportion and absolute number of minorities in the neighborhood increases. Alternatively, residents of a neighborhood may compete with each other for resources, or may judge their interests and status relative to their immediate neighbors. These latter two may produce greater heterogeneity of preferences than otherwise expected. Pattie and Johnston (2000) use data from the 1992 British Election Study (BES) to show that voters who spoke together tended to change their party preferences to match each other. Such face-to-face conversations are a consequence of living in the same area, and cannot be reduced to individual-level characteristics. At the same time, the effect of such information sharing has not always been shown to be significant, so the relative importance of the neighborhood effect remains an active area of research.

Beyond face-to-face interactions, political geographers have generally focused on political party activities as a contextual force. In his study of Italian elections, for example, Agnew (1996, 139) identifies two contextual effects that are closely tied to party activity and presence. First, the slate of parties may not be the same in every constituency, so that voters in different areas “consider the political options and what they mean in different ways.” Second, parties may have a more or less effective presence in different areas, with some having “familiar faces to present, congenial stories to tell, and the means to communicate them both” where others do not.

In a slightly different fashion, spatial analyses of campaign spending also reflect the position that location makes a difference in political preference. The mechanism in this case is slightly different. Political parties may invest resources in particular areas, building the kind of long-term presence that may produce local exceptions to national patterns. In electoral systems that use territorial constituencies, such as Canada, the United Kingdom, and the United States, parties typically vary campaign spending between constituencies. As Pattie and Johnston have
ELECTORAL GEOGRAPHIES

shown in the United Kingdom, parties typically maximize expenditures in their most competitive constituencies, where their candidate has a relatively even chance of winning or losing. Parties can thus create spatial variation through both long- and short-term activities.

With only a few exceptions, electoral geography has focused on election and referenda analysis, and have not pursued Wright’s investigation of legislative (roll-call) votes. This is unfortunate because such analyses can reveal the dynamics of the relationship between representatives and their constituents in a geographically sensitive manner.

The geography of representation

The third major area of electoral geography concerns the analysis of electoral systems. The preponderance of this research concerns so-called first past the post (FPTP) or plurality systems using territorial constituencies. Typically in such systems, two or more candidates compete for a single seat, and the voting constituency is a contiguous geographic area (“districts” in the United States, “ridings” in Canada, and simply “constituencies” in the United Kingdom). The candidate who obtains a plurality or majority of votes wins the seat. Outside the Anglo-American core, however, proportional representation (PR) is far more common. In PR systems, an elector casts a vote for a party rather than a candidate, and parties obtain seats (more or less) in proportion to the number of votes they receive.

The reasons for the focus on plurality systems are, arguably, twofold. First, the constituency system provides data that easily demonstrate spatial variation. The more than 300 ridings of the Canadian Commons, the 435 Congressional districts of the US House, or the 650 seats of the British House of Commons, for example, all provide scholars with multiple simultaneous elections with thousands of candidates, multiple parties, and a wide variety of socioeconomic and geographic contexts. Modern computer and GIS (geographic information system) technology has increased the availability and detail of such data to an extraordinary degree; in Quebec, for example, election results are available for over 20,000 geographically defined precincts. These data not only lend themselves easily to spatial analysis, but election outcomes in FPTP systems are highly dependent on the geographic distribution of the vote.

Second, Anglo-American geographers tend to live in states that use plurality systems. To the extent that scholarly interest and research funding stay close to home, the focus on such systems is hardly surprising. Indeed, Ronald Johnston and his collaborators are so dominant that their analyses of the British system constitute a plurality of publications in the subfield. Nonetheless, the collection edited by Warf and Leib (2011) includes case studies from Europe and Asia in addition to those of the United Kingdom and the United States. There is also periodic attention to historical cases, and to elections in former communist states.

Unlike PR systems, FPTP elections rarely result in proportionality, and the relationship between the proportion of seats won and votes received is called the seat–vote ratio. Table 1, showing the results of the 2011 Canadian federal election and the distribution of the 308 Common seats between five parties, illustrates this phenomenon. In the 2011 election, for each 1% of the popular vote, the Conservative Party obtained 1.4% of seats in the Commons, whereas the Liberal Party obtained only 0.6% of the seats for each 1% of the popular vote. While such results may be counterintuitive, they are easily explained by the nature of plurality systems. Candidates compete in separate constituencies,
Table 1  Canadian federal election (House of Commons), 2011.

<table>
<thead>
<tr>
<th>Party</th>
<th>No. of seats</th>
<th>% of seats</th>
<th>% of vote</th>
<th>Seat–vote ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservative</td>
<td>166</td>
<td>53.9</td>
<td>39.6</td>
<td>1.4</td>
</tr>
<tr>
<td>New Democratic Party</td>
<td>103</td>
<td>33.4</td>
<td>30.6</td>
<td>1.1</td>
</tr>
<tr>
<td>Liberal</td>
<td>34</td>
<td>11.0</td>
<td>18.9</td>
<td>0.6</td>
</tr>
<tr>
<td>Bloc Québécois</td>
<td>4</td>
<td>1.3</td>
<td>6.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Green</td>
<td>1</td>
<td>0.3</td>
<td>3.9</td>
<td>0.1</td>
</tr>
<tr>
<td>Others</td>
<td>0</td>
<td>0.0</td>
<td>0.4</td>
<td>0</td>
</tr>
</tbody>
</table>

and in principle only need to win by a single vote. If a party’s candidates, for example, come in second place in all districts, the party will not win a single seat even if they obtain a significant fraction of the vote. More generally, FPTP systems can generate a large number of “wasted votes,” votes cast for losing candidates.

Seat–vote ratios are dependent on both the boundaries of constituencies and the geographic pattern of the vote in each election. For example, in the 2008 Canadian federal election, run with the same district boundaries, the Bloc Québécois enjoyed a seat–vote ratio of 1 : 6. In general terms, such plurality systems penalize parties whose support is spread relatively evenly, unless it is high enough to constitute a plurality in many districts (as was the case with the Conservatives in 2011). These systems can provide opportunities, however, for parties or minority groups that have significant geographic concentrations and that form local pluralities in one or more districts (as was the case with the Bloc Québécois in 2008).

From a party’s perspective, the ideal set of constituencies is one that produces the most efficient distribution of their supporters, one that results in the fewest wasted votes for their candidates. In theory, the most efficient distribution of votes is one where a party wins each victory by one vote, and thereby maximizes its seat–vote ratio. In practice, of course, parties do not attempt to cut their margins of victory so thinly.

Redistricting and gerrymandering

Jurisdictions using plurality systems typically redraw their constituencies periodically in the process of redistricting (Morrill 1981). Such exercises are often highly controversial because they provide opportunities for political parties to create gerrymanders. While there is no single measure of what constitutes gerrymandering, it is generally understood to mean the configuration of election districts that confers an unfair advantage (or disadvantage) to a political party or minority group. Although proportionality is an attractive baseline for measuring gerrymandering, the seat–vote ratios inherent in plurality systems mean that even districting plans created through a nonpartisan process will typically produce nonproportional election results. The determination of what constitutes an “unfair” advantage (or disadvantage) is ultimately a political decision rather than a scientific one.

There are three principal techniques of gerrymandering. The first is malapportionment, or the unequal distribution of electors between constituencies. While the “rotten borough” (a highly underpopulated constituency)
was a long-standing British tradition, most states now limit the degree of population deviation they permit between constituencies. The United States has among the strictest standards, which in practice require the equal distribution of population between congressional districts within a particular state. Canada, in contrast, permits up to ±25% population deviation between ridings within a province as a matter of routine. One consequence is heavily populated urban ridings in Canada that dilute the voting strength of visible minorities as well as of city dwellers more generally. Like most federal states, however, the distribution of seats between provinces and states means that some degree of malapportionment is unavoidable.

Second, if a group is relatively small compared to the size of constituencies, it can be split between two or more districts. Such “cracking” works by lowering the number of a political or ethnic minority in a district below plurality, ensuring that they can never elect a representative of their choice.

Finally, if a minority group is too large to be split effectively between districts, it may be “packed” into one district to create a supermajority and thereby waste as many votes as possible. Methodologically and legally, the evaluation of packing can be difficult. The difference between a packed district and one with a “safe” majority is blurred, and depends on many factors, including the volatility of the electorate, turnout, citizenship rates, and population movement. Ethnicity and party may also intersect. In the United States, a district with a bare majority of African Americans may have a large Democratic majority; it would be competitive in racial terms but packed from a partisan perspective.

Jurisdictions have grappled with the partisan nature of redistricting in various ways. Canada and the United Kingdom, for example, use arm’s length boundary commissions in redistricting. Even such commissions, however, may be susceptible to partisan influence. In contrast, the redistricting process in the United States is typically under the jurisdiction of legislatures and thus tends to be a highly partisan affair. The actions of legislatures are limited through the legal system, and litigation is a normal step in the US redistricting process. For example, 31 of the 50 states had their redistricting plans challenged in court by the end of 2011 during that round of redistricting. Consequently, the electoral geography of redistricting in the United States tends to engage closely with constitutional law. In such studies, the primary questions concern not whether a districting plan is biased, but how the court system judges the plan in constitutional terms.

The centrality of racial conflict to the politics of representation also makes the study of redistricting in the United States distinct from other jurisdictions. Elsewhere gerrymandering and bias are typically discussed only in terms of political parties, but in the United States the representation of racial minorities, particularly African Americans and Latinos, is of at least equal concern.

SEE ALSO: Democracy; Neighborhood, conceptual; Political geography; Power; Representation; Spatial context

References


Further reading


Elite labor

Sarah Hall
University of Nottingham, UK

Elite labor emerged as a central component of the global economy following the deindustrialization and resulting tertiarization of the economies of Western Europe and North America from the 1970s onwards. Following the decline in the relative contribution of the manufacturing sector to the gross domestic product of these economies, a range of business services became increasingly important including both knowledge-intensive advanced producer services (such as finance, legal services, accountancy, and management consultancy) and creative industries (such as design, architecture, and advertising). For all of these services the knowledge and expertise of their employees represent the central “raw material” of such activities; these individuals collectively make up elite labor. As such, elite labor is typified by a high skills base that combines holding highly valued academic credentials (often at graduate level including masters and PhD degrees) with an appreciation of the taken-for-granted norms and cultures associated with each sector.

Given the close relationship between elite labor and the rise of advanced producer service firms and activities within the creative industries, early work within economic geography and cognate social sciences focused on examining it as a factor of production within these firms. Scholars emphasized how the nature of the knowledge embodied within elite labor shaped the organizational and geographical operations of these firms, particularly as they expanded their operations internationally. More recent work has developed this approach by focusing more explicitly on elite labor itself. This work has examined the operation of elite labor markets including their gendered nature, the role of corporeal mobility within them, and their geographies in terms of the dynamic relationship between different global cities in which these labor markets are concentrated.

Within this increasingly broad literature, it is important to note that the definition of elite labor itself is not widely agreed upon. These definitional issues stem, at least in part, from the genealogy of studies of elites and elite labor themselves. Most notably, the development of work on elites is characterized by its episodic advancement. For example, central social theorists of the nineteenth century such as Karl Marx positioned the ruling classes at the center of their analyses of the development of economic life. However, this interest in the power of elites diminished in the twentieth century. Moreover, those theorists who continued to work on powerful elites, notably Pierre Bourdieu, moved away from a focus on economic elites and elite labor to foster an interest in cultural elites, including academics and civil servants. More recent work, such as that by the sociologists Manuel Castells and Leslie Sklair, has understood elites through the lens of globalization, but has said comparatively little about the microsociological practices of elite labor and its wider socioeconomic implications. The result is a disparate literature on elites and elite labor with some work emphasizing their control over (economic) resources at the expense of other groups from a predominately Marxist perspective, while
other work adopts a more cultural and socially inflected account to examine how elites seek to reproduce their powerful position in society. This diversity in understandings of what elites are is reflected in the development of geographical literature on elite labor. However, while the theoretical standpoints of work on elites remain varied, the field can be characterized by a concern with understanding the ways in which elite labor emerges as powerful through its control of resources (be that economic, cultural, or financial), based significantly on their knowledge and expertise in the contemporary global economy.

Elite labor and service firms

In contrast to the manufacturing sectors in which the raw materials of production include not only the skills of workers employed within them but also material elements involved in activities such as car assembly, advanced service sector firms and the creative industries are characterized by the fact that the knowledge and expertise of individuals working within them are their central raw materials. For example, in corporate legal services, solicitors in law firms are expected to tailor their knowledge of the law of the jurisdiction in which they are operating when offering advice, often drawing on their previous working experience and their knowledge of the corporate culture of the firm in which they work. Reflecting this understanding of the centrality of the knowledge of individuals to the operations of their employing firms, early work on elite labor conceptualized it as a central factor of production for these firms that had important implications for understanding the geography of these sectors.

This literature specifies the different types of knowledge provided by elite labor to these firms. A key distinction is made between tacit and explicit knowledge. Drawing on the work of Polanyi (1944), tacit knowledge refers to embodied knowledge acquired through experience, while explicit knowledge is the formal knowledge that is widely agreed upon within any given sector (such as particular approaches to corporate valuation within financial services that are learnt through standard university courses and widely described in key textbooks). While both forms of knowledge are central to elite labor within highly skilled service sectors, tacit knowledge is seen as being particularly important. The cultural sensitivities and expertise derived from previous working experience that are central to the formation of tacit knowledge are seen as being particularly valuable in differentiating between different firms that are competing within any given sector.

The significance of varying types of knowledge within highly skilled service sectors has given rise to considerable debate concerning the different geographies associated with these different types of knowledge and the implications of this for the locational decision-making of firms. Building on the seminal contribution of Gertler (2003), scholars initially emphasized the need for material geographical proximity in order to share the tacit knowledge that was vital for corporate success through face-to-face interaction between individuals within elite labor markets. In this work, it is argued that shared cultural norms are needed to develop the trust-based interpersonal networks through which tacit knowledge can circulate. The geography of tacit knowledge embodied within elite labor thus gives rise to the geographical colocation of similar knowledge in highly skilled services. Examples include the cluster of high-tech firms in Silicon Valley in California, financial services complexes in financial districts such as Wall Street in New York and the City of London, and the media cluster in Leipzig.
However, following work on communities of practice (Amin and Cohendet 2004), scholars have subsequently examined the ways in which different forms of relational proximity can exist in which individuals may be geographically distant but relationally proximate through shared experience of, for example, working on the same projects for short but intense periods of time or through being embedded in the same corporate culture while working in different geographical locations (Amin and Cohendet 2004). This work has suggested that, using these relational proximities, tacit knowledge may be shared between individuals who are geographically disparate. This approach has led to a greater emphasis on conceptualizing elite labor and the firms in which it operates as forms of relational networks that cross-cut geographical space but are grounded in particular places at particular times (see for example Faulconbridge 2006). It has also facilitated a greater emphasis on the nature of elite labor in and of itself and elite labor markets more generally rather than placing elite labor as a factor of production within wider work on the geographies of advanced producer service firms and creative economy.

Elite labor and elite labor markets

The growing interest in elite labor and elite labor markets reflects the ways in which elites have become the source of considerable political, popular, and media attention following their central role in choreographing the financialized boom of the 2000s and the ensuing global financial crisis and recession. The centrality of elites to these significant economic changes has been reflected in a resurgence of work on elites within the social sciences (Savage and Williams 2008) that has sought to move beyond an emphasis on the formation of elites based on their educational and social background to include analysis of their practices and the ways in which these have been used to shape the global economy in particular ways. In so doing, this work includes analyses of the operation of elite labor markets and the implications of these for individuals within them.

The most notable feature of elite labor markets has been their dramatic expansion in recent years in light of the expansion of advanced producer services and the creative economy. As a result, traditional recruitment methods based around educational background at a limited number of exclusive schools and universities (see Bourdieu 1996) are no longer able to meet the demand for elite work. While recruitment has expanded, research demonstrates that entry into these labor markets remains very competitive. In response, individuals seeking to work in elite labor markets adopt a number of strategies in order to secure themselves positional advantage in relation to other potential new recruits in an attempt to mark themselves out as distinctive, and hence potentially valuable to employing firms. Strategies adopted in this respect include not relying solely on a first or undergraduate degree but also developing an extensive portfolio of work experience in their chosen field or undertaking an international degree program that they feel will have a scarcity value in the labor market, as well as carefully selecting the university and degree program they undertake. Research into these activities demonstrates how individuals seek to strategically accumulate credentials and work experience in order to secure entry into competitive elite labor markets.

These positional competitions within elite labor markets continue once an individual has secured their first position. At this stage, elite labor markets are often, implicitly at least, framed as global labor markets in which there is a global “war for talent” within so-called knowledge-based economies. At one level, these
ELITE LABOR

discourses suggest that global labor markets offer considerable opportunities for upward mobility of elite labor. For example, work on education and learning within elite labor markets has demonstrated how discourses of knowledge-based economies are used to encourage elite labor to invest in their economic potential and human capital through the continuous acquisition of new skills and credentials. However, Thrift (2005) highlights that this form of what he terms “soft capitalism” has “hard edges” such that not all elites are equally able to benefit from the opportunities within elite labor markets.

In this respect, work in economic geography on segmented labor markets is instructive. Through research largely conducted into the labor market experiences of minority groups, this literature has demonstrated how an individual’s experience of labor markets is shaped by the local contexts in which they operate and an individual’s experiences of them. This work has been usefully developed in order to consider one of the “hard edges” associated with elite labor that focuses on its geography. In much of the popular and academic literature that surrounds elite labor markets, elite labor is positioned as being global in geographical scope, with individuals within this group having the potential to move relatively freely between leading centers of economic activity, particularly global cities. However, by using work on segmented labor markets, it has been shown that elite labor is more geographically nuanced than has often been assumed to be the case. In this work, elite labor is characterized by its trans-local qualities such that while individual elites have the potential to render themselves mobile through migration and expatriation, their working practices are often tailored to the cultural norms and regulatory practices of the specific geographical location in which they work. This can limit their ability to move geographically since such mobility requires an understanding of the different working cultures associated with different geographical locations.

Building on this appreciation of the geographically nuanced nature of elite labor markets, a considerable body of research has emerged that looks at the mobility of elite labor and the implications of this for the nature of elite economic practice in different economic clusters, particularly global cities. Much of this work has documented how migration and expatriation of elite labor from established global cities such as London and New York to emerging centers, particularly in Southeast Asia, are used by firms in knowledge-intensive business services such as finance and law to aid in their internationalization activities by circulating taken-for-granted ways of working as they seek to tailor these on entry to new markets. However, these mobility patterns are becoming more complicated. For example, migration from (rather than to) emerging global cities such as Beijing and Shanghai is becoming increasingly common. These types of mobility will be potentially very significant in the future since they are being used by domestic firms in rapidly growing economies such as China to develop expertise within their own domestic elite labor markets, with implications for elite labor in both advanced and emerging economies. Moreover, the variegated impacts of the 2007–2008 global financial crisis and ensuing recession have created several new important migratory flows of elite labor within Europe and North America, such as the large elite French population living and working in advanced business services in London, partly in response to political changes in France targeting elite labor.

Developing largely in parallel to this work on elite labor mobility, geographers have been at the forefront of work that has revealed the gendered nature of elite labor, in terms of both the composition of elite labor markets and the
embodied performances and workplace practices of elites themselves. This can be viewed as another “hard edge,” in Thrift’s terminology, of elite labor markets. Much of this work builds on McDowell’s (1997) seminal analysis of the nature of gendered elite financial labor in London’s financial district in the 1990s. Here, she revealed that despite claims from employers within elite labor markets that the Americanization of the City of London following the deregulatory changes of the 1980s and the associated increase in the number of overseas firms within London had led to a more meritocratic City, including the greater representation of women within highly skilled financial labor markets, women continued to be underrepresented within financial services work in London. Moreover, elite workers employed within finance undertook a number of gendered performances pointing to the existence of multiple masculinities as well as femininities within these labor markets, ranging from the aggressive, ambitious trader to the “gentlemanly capitalist” of the boardroom. The significance of this work has been underscored by the financial crisis of 2007–2008 about which policymakers and the media argued that the underrepresentation of women with elite labor had contributed to the crisis because of the inherently greater level of risk taking associated with masculine performances in the workplace, claims that have been strongly critiqued for their essentialized understandings of gender by feminist scholars.

Taken together, work on elite labor has advanced considerably in recent years, moving beyond an initial focus on elite labor as a factor of production within knowledge-intensive business services and the creative industries to develop understandings of elite labor markets that have distinctive entry requirements and operate in geographically nuanced and gendered ways. In so doing, this work has demonstrated that understanding elite labor is important in terms of wider discussions about the reproduction of a variegated global economy as the nature of elite labor market practice is shaped by the cultural norms and regulations associated with the local contexts in which it takes place. This work has also drawn on and developed important methodological debates within economic geography and the social sciences because one of the explanations for the comparative neglect of elites in the latter part of the twentieth century has been the argument that securing research access to these groups is difficult, particularly for early career researchers. Moreover, even if access is granted, the power relations involved in elite interviewing are asymmetric, thereby shaping the kinds of knowledge and understanding that are produced. Economic geographers have developed a number of strategies in attempts to overcome this, such as fostering close research collaborations with the economic elites that they are researching. Meanwhile, in the wider social sciences, research into the geographies of money and finance has increasingly relied upon former financiers writing up their own personal accounts of working in the international financial system as a form of research method. While both of these approaches have resulted in rich insights being gained into contemporary economic elites, they also raise important questions concerning the importance of the positionality of the researcher in accounts of elite labor.

The development of work on elite labor in geography and the wider social sciences also has important implications beyond theoretical debates and methodological concerns. In particular, in policy circles, much has been made of the emergence of more meritocratic labor markets that embody the theory of human capital in which individuals progress following investment in their own skills and knowledges, with the onus on individuals increasingly to do this. However,
geographical work has revealed the limitations of this position, with career progression within elite labor markets continuing to be shaped by the geography in which an individual operates and other socioeconomic characteristics, notably gender. In the future, these debates will need to be extended to provide greater understanding of the emergence and development of elite labor in rapidly growing economies, particularly in the Asia-Pacific region and the BRIC economies (Brazil, Russia, India, and China). Indeed, an important part of this work will be not only to understand the emergence of domestic elite labor in these regions but also to analyze the implications of their growth for the more established location of elite work in North America and Western Europe that have been the main geographical focus of attention thus far.

**SEE ALSO:** Corporate spatial organization and producer services; Economic geography; Financial geography; Global cities; Human capital; Interviews; Labor geography; Labor market segmentation; Migration: international

**References**


**Further reading**

Emerging market: southern corporation

Pádraig Carmody
Louis Brennan
Trinity College Dublin, Ireland

Transnational corporations (TNCs) or multinational enterprises (MNEs) are perhaps the defining institutions of globalized capitalism and have been characterized as “the most significant actors in today’s globalised world” (Rugman and Verbeke 2004, 3). While their emergence is often thought to be a relatively recent phenomenon they have been in existence for centuries. Early examples of these would include the British and Dutch East India companies, which received their charters in 1600 and 1602, respectively.

While Stephen Hymer’s (1976) theory of foreign direct investment (FDI) examined the ways in which these companies were implicated in northern imperialism, more recently there has been a substantial growth of corporations originating in and from the Global South, beginning in the 1970s with those from the newly industrializing countries of Asia, such as Hong Kong (Yeung 1998). Traditionally FDI flows were dominated by north-north and north-south directionality with considerably lower levels of south-south flows, generally involving investment in proximate markets. However, the new millennium, in particular, has been marked by a dramatic increase in both south-south and south-north flows of FDI. Flows from the developing world have been increasing faster than those from the developed world, accounting for an ever-greater proportion of total FDI flows and stocks. For example, in 2012, developing economies generated almost one-third of global FDI outflows, and recent data reveals China to be the third-largest source of outbound foreign direct investment (OFDI) in the world; moving from sixth to third in 2012, behind the United States and Japan. By way of further example, Tata, originally an Indian conglomerate, is now Britain’s largest manufacturing company. Recent decades have also seen the emergence of sovereign wealth funds from the Global South, some of which now also hold substantial investments in northern government bonds and other investment classes. The emergence and expansion of these companies has implications for global development and consequently theories of FDI. For the purposes of this entry, “emerging market” and “southern” are treated as synonymous, although countries in the former category are often held to be more developed than the least developed economies and countries.

The rise of the South?

The emergence of southern corporations is both a driver and reflective of “the Rise of the South” (United Nations Development Programme 2013). In the Forbes Global Fortune 500 index of the biggest companies in the world, 89, mostly state-owned, are now from China. In 2012, of the eight Indian companies that made the list, five were state-owned. The growth of these companies is reflective of both the spatial dialectics of globalization and the “success” of particular state-directed strategies of late-development and the interaction between
EMERGING MARKET: SOUTHERN CORPORATION

them. These dynamics have given rise to some of the largest companies in the world. For example, according to Forbes, the Industrial and Commercial Bank of China is now the world’s largest company.

The growth of emerging market corporations is also both reflective and constitutive of the ongoing rebalancing of the global economy and in particular the shift in its center of gravity from the Global North to the “Global East,” as China is now the world’s second-largest national economy. The Chinese government has a declared “go out” policy for its corporations to encourage them to invest overseas to tap new markets, gain experience of overseas operations, develop scale economies, and access new technologies and other forms of knowledge. As stated by then Premier Wen Jiabao in March 2012: “This year, China will further its strategy of going abroad. The nation is at an important stage of accelerating the steps it has taken to make outbound investments. China will guide all sorts of companies to make orderly investments in the energy, raw materials, agriculture, service and infrastructure industries, using mergers and acquisitions” (quoted in China Daily, March 6, 2012). The economic reform program of the new leadership, announced in 2013, further promoted Chinese outward FDI by again relaxing some of the controls on outbound investment. The “global” financial crisis (GFC) was also seen by some companies as a buying opportunity and, partly as a result, Europe is now the largest regional recipient of Chinese FDI.

The fastest-growing economies in the world have recently been in the developing world. For example, Africa hosted up to seven of the world’s fastest-growing economies in the early years of the new millennium, although given rapid population growth and growing inequality this has not resulted in commensurate increases in per capita income or poverty reduction. Nonetheless, certain companies and conglomerates have been able to take advantage of relatively rapid economic growth to expand their operations. The best known example is perhaps the Dangote Group, headquartered in Nigeria. This is a diversified conglomerate engaging in the production of a range of commodities from cement to food. Whereas Dangote’s operations are regional rather than “global” in scope, some (South) African corporations have established a substantial extra-regional presence, such as Nando’s (fast food) having operations in 30 countries on five continents.

The pattern of regional investment is mirrored in Latin America by the multilatinas, which are corporations which have invested in multiple countries in the region, with the more successful of these expanding globally over the past decade. For example, Vale from Brazil is now one of the world’s largest natural resource corporations, with substantial, and at times controversial, operations in Africa and Canada, where it has been fined for fatal safety breaches.

Manufacturing firms

There is a variety of major transnational manufacturing corporations headquartered in the Global South. Sometimes these are diversified conglomerates, involved in seemingly disparate economic activities. ZTE, for example, a Chinese company best known globally for its operations in mobile telephony, is now a foreign investor in African land.

Examples of southern multinationals acquiring established brands abound, including that of the well-known British breakfast cereal brand Weetabix, now owned by the Chinese company Bright Foods. The bias toward asset exploration (acquiring strategic assets available in a host country) in the developed world on the part
EMERGING MARKET: SOUTHERN CORPORATION

of these emerging MNEs is reflected in the distribution of their investments between acquisitions and greenfield-type investments. During 2000–2013, the value of acquisitions by Chinese firms in the United States to greenfields was over 7:1 (Rhodium Group 2013). Chinese investment into Europe is similarly skewed toward acquisitions, particularly in recent years with distressed European firms representing good buying opportunities. According to Rhodium Group estimates, annual Chinese OFDI flows to Europe grew from less than €1 billion per year from 2004 to 2008 to €1.9 billion in 2009 and €2.7 billion in 2010, almost tripling to €7.6 billion in 2011. In 2012, Chinese firms invested a further €7.8 billion in the European Union, including large-scale acquisitions in utilities (Energias de Portugal), consumer products (Weetabix), industrial machinery (Putzmeister), and infrastructure (Heathrow Airport) (Rhodium Group 2013).

Agro-conglomerates

Agro-conglomerates from developing countries have now also become major investors in parts of the Global South. For example Karuturi Global, an Indian company, is now the largest producer of cut roses in the world, exporting 1.5 million stems a day to Europe from Kenya alone. With the global food price spikes of 2007–2008 and 2011, Indian corporations and the government have also been active in sourcing land in Africa to grow food crops to export back to India and to sell on international markets.

Service providers

While many southern corporations are now transnational in scope, home country advantages may remain important for their operations. For example, Infosys in India is one of the world’s largest providers of “back office” functions to transnational corporations. This company is able to leverage its international connections and the low labor costs and relevant English speaking skills of the Indian workforce to grow domestically, while also expanding internationally.

Generally, growing domestic markets, fueled by global resource demand, particularly from China, have also enabled the development and expansion of other service providers in parts of the Global South. For example, Ecobank is a pan-African banking conglomerate with operations in 33 countries on the continent and in a further four countries overseas. This bank is of particular interest not only for the scale of its operations but because it was founded on the basis of a joint initiative between the Federation of West African Chambers of Commerce and Industry and the Economic Community of West African States.

Sovereign wealth funds

Kuwait was the first country to establish a sovereign wealth fund, in 1953. Sovereign wealth funds take assets, typically from natural resources, and then invest them in other asset classes in order to secure “permanent income” from other sources once natural resource reserves run out (Collier 2010). A number of oil-rich states in Africa, such as Ghana and Chad, and some oil-rich Gulf states have followed the Kuwaiti example and established sovereign wealth funds in recent years. China also has over 3 trillion dollars in foreign exchange reserves, some of which are invested overseas and domestically through the China Investment Corporation. In some instances sovereign wealth funds serve as
EMERGING MARKET: SOUTHERN CORPORATION

holding companies. For example, Emirates, the largest airline in the Middle East, is owned by the government of Dubai through its sovereign wealth fund.

State-owned enterprises and natural resource companies

As noted earlier, while it is often thought that globalization is a private sector-led phenomenon, some of the largest southern corporations, from China in particular, are wholly or partially state-owned. Despite a general tendency of neoliberalization in the Global South over the last number of decades, southern states often maintain a direct presence, interest, or ownership in and of their energy sectors given their strategic importance. Strategies to ensure energy security have ranged from investment in overseas resources to the development of new domestic resources. For example, as a result of production by companies such as Copersucar, Brazil is now the second-largest producer of ethanol globally, behind the United States.

In the third world, emerging market corporations are increasingly being used to source critical or strategic natural resources to fuel national economic growth and meet security concerns in the context of growing ecological scarcity. Thus, depending on the sector in which they are involved, and whether this is deemed important to national and economic security, and whether or not they are state-owned, emerging market corporations may receive different levels of encouragement and support from home country governments. The three biggest oil companies in the world in terms of reserves are not well known “supermajors,” such as Exxon Mobil, but state-owned companies from Saudi Arabia, Russia, and Iran.

Are southern multinationals different?

Whereas extant theories of FDI examined the reasons why developed country TNCs might establish operations in the Global South, such as labor cost differentials, flows of foreign investment from developing to developed countries follow a different logic in a number of respects. Traditional theories of the internationalization of the firm suggest a gradual, incremental process starting with geographically (and perhaps culturally) close host countries before advancing to more distant locations (Buckley and Ghauri 1999). Yet emerging market or southern multinationals have tended to expand geographically at a much faster rate than extant northern corporations, often entering into distant markets in the early stages of their internationalization. One of the most prominent motivations of southern corporations to invest in northern markets is often brand acquisition, although in some cases they also succeed in establishing their own brand identity in those countries, as in the case of Nando’s discussed above. Also, while northern investing firms tended to engage more in asset exploitation (the transfer of a firm’s proprietary assets to overseas operations), the newcomers have been observed to be engaging more in asset exploration. This helps to explain why these firms move to, or establish operations in, high (labor)-cost locations in the developed world. By seeking to take advantage of knowledge in the new recipient country locations, acquire advanced technologies, and buy established brands, they aim to leapfrog ahead in their development, thereby avoiding the normally lengthy process of generating knowledge, developing new technologies, and establishing brand names. The Chinese computer company Lenovo’s acquisition of IBM PC is one early notable example of such an acquisition.
The case of Chinese MNEs and their investment strategy deserves special attention for a number of reasons (Brennan 2011). First, there is the primacy attached to the goal of facilitating China’s next stage of domestic economic development in relation to Chinese OFDI. Second, there is the degree to which China’s MNEs remain deeply embedded in China’s cadre capitalism. The Chinese model of OFDI is different from that of other countries, with 93% of Chinese investment in Europe now coming from state-owned corporations, for example. This represents a new model of globalization in contradistinction to the private sector-led one of western corporations. As China liberalizes its economy domestically, the power of the Chinese state is growing overseas through this and a number of other channels, throwing into question the ontological distinction between “states” and “corporations” as separate social forces in this instance. Thus, rather than globalization reducing the power of the Chinese state, it is respatializing it (Carmody, Hampwaye, and Sakala 2012). The growing significance of this trend has been underlined: “In no dimension of China’s economic engagement internationally is the interaction between the economic and political systems more prominent and important than in respect of the surge in China’s overseas direct investment (ODI)” (East Asia Forum Weekly Digest, April 2, 2012).

Third, there is a contrast between OFDI on the part of Chinese MNEs and some of the established theories. State-owned corporations are run differently, respond to strategic state imperatives, and have longer-term time horizons than the quarterly stock market returns that drive publicly listed corporations. Particularly as they may be able to access substantial foreign exchange reserves, this means that they can sometimes deploy countercyclical strategies. For example, when the price of copper collapsed in the wake of the GFC, and many western and other companies consequently were divesting from copper production in Zambia, Africa’s second largest producer, some Chinese companies took advantage of falling asset prices to buy out previous incumbents.

Given the sensitivity of some sectors in host economies, these corporations sometimes become embroiled in geopolitics, as when the US Congress voted 398–15 against the Chinese National Overseas Oil Corporation’s acquisition of a US oil company in 2005. This resistance is sometimes cited as one of the reasons for the rapid expansion of Chinese oil companies in Africa. While Europe’s receptivity to nontraditional sources of investment has been varied and fragmented, the Eurozone crisis resulted in a greater receptivity to such investment on the part of those actors who were previously apprehensive about southern MNEs (Brennan and Kim 2013).

Sovereign wealth funds from the South have sometimes raised concerns among policymakers and market players in the developed world. These stem from the dominant role played by national governments in the management of these large funds and from a lack of transparency and accountability arising from this. These concerns have led to the promulgation of the Santiago Principles, voluntary “best practices” for sovereign wealth funds, from an International Working Group established by the International Monetary Fund, representing 26 countries.

The future of southern multinationals

It can be anticipated that southern multinationals will be an increasing presence on the global landscape. Their early and rapid expansion into developed markets poses challenges for both the new entrants themselves and for incumbents. Perhaps the greatest challenge for the new
EMERGING MARKET: SOUTHERN CORPORATION

entrants is learning to operate in environments that are culturally and institutionally distant from their home environments. This is critical for such firms if they are to overcome the “liability of origin” and gain greater legitimacy in the developed world (Brennan and Kim, 2013). For host countries the challenge will be to ensure that investment by southern multinationals brings benefits by contributing to growth and employment.

SEE ALSO: Brands and branding; Corporations and the nation-state; Economic geography; Firms; Globalization; Industrial geography; Internationalization; Mining and mineral resources; State-owned enterprise

References


Further reading


Emotional geographies

Deborah Thien
California State University, Long Beach, USA

The field of emotional geographies comprises interdisciplinary scholarship focused on theoretical and substantive considerations of emotion, space, and society. Major influences on this work include feminist scholarship, humanistic and phenomenological approaches, health and wellbeing literatures, and the spatial turn in cultural theory. The resulting scholarship on the spatialities and socialities of emotion, affect, and feeling has emerged from diverse disciplinary perspectives both within geography (from feminist, health, social, cultural, critical race, and other geographies) and without, including long-established work in sociology, psychology, philosophy, and psychotherapy, as well as examinations from economics, neuroscience, literature, and political ecology, among others. These collective and diverse considerations have contributed to an “emotional turn” in geography.

Thinking about feeling

Emotion encompasses mental and corporeal, social and cultural, and spatial and historical elements. Scholars have sought to define, delineate, understand, and contest the specifics of each of these elements as well as their intersections and interactions. Socially and culturally inclined definitions of emotion regard it as a dynamic process, subject to shifting relations of power and to dominant views of emotion in any particular time or space. Emotion, then, can be understood as the continual taking place of dynamic, felt relations between bodies, happenings, spaces, and things. Bondi, Davidson, and Smith (2005, 1) emphasize the importance of “a non-objectifying view of emotions as relational flows, fluxes or currents, in-between people and places rather than ‘things’ or ‘objects’ to be studied or measured.” Such understandings of how the world is mediated by feeling attempt to make sense of emotion as a way of being, knowing, and doing in the world. Relational models of emotion incorporate both embodied and psychological dimensions. Such definitions, which build on the relations of body and mind and their attendant, shifting, and power-laden spatialities, are indicative of contemporary thinking about emotion and stand in contrast to those shaped by Descartes’s seventeenth–century theorizing of a mind–body separation.

Within the Cartesian framework, emotion is defined predominantly by its otherness from reason. Examples abound which suggest the disordered nature of emotion: “boundless grief,” “wild with anger,” “uncontrollable fury,” and “overjoyed.” Scientific inquiry has distinguished itself epistemologically in the practices of rationality, intellectual rigor, and (self-)discipline; in this tradition, science as a set of laws dispenses with that which oversteps the bounds. This binary is further cemented through emotion’s association with the body, in opposition to the reasoning mind. To be emotional in these terms is to be thought-less. Discursively and in practical terms, the power of reason has been a wide-ranging masculine domain, while women have represented, and have been represented by, the intimate messiness of the body and its feeling states (Harding and Pribram 2002).
EMOTIONAL GEOGRAPHIES

A consequence is that knowledge itself has been positioned as a desired/desiring masculine property, while emotion has been categorically restricted to the spaces of the personal. Another consequence of these authoritative discourses has been to make the study of emotion oxymoronic: it is seen as unreasonable and at odds with scientific objectivity. Scholarship around emotion can be read, therefore, for reiterations of power inequalities in the production of knowledge. Feminist scholarship has issued an important challenge to these Enlightenment binaries and has asserted new perspectives on emotion.

Feminist perspectives and influences on studies of emotion

The increasing legitimacy of emotion as an area for scholarship is often attributed to feminist critiques of masculinist scientific paradigms; in particular, feminists have challenged the legitimacy of binary oppositions, claims of objectivity, and the disembodiment of knowledge. The feminist deconstruction of binary oppositions has allowed for understanding the reason–emotion opposition as both an influential construct and a false divide. Emotion yoked to reason is no longer simply a feminized negative to masculinist rationality; rationality itself is reread as infused with emotionality. For example, the amplified dialogue about emotion within academia serves to counter the cool spaces of the neoliberal university in which students are positioned as rational consumers and instructional staff as dispassionate suppliers of goods and services. Emotions have been both feminized and privatized within contemporary academies, serving to contain emotions (Harding and Pribram 2002). Safely contained, emotions can be set aside while the “real” academic labor goes on. Paradoxically, emotion is figured as out of control and yet is tightly controlled.

Feminist scholars have contested masculinist claims of objectivity by arguing for the tentative, partial, and situated nature of knowledge. In contrast to the expert account that claims value-free results, many feminist social scientists have championed the use of reflexivity, an explicit practice of situating knowledge by openly acknowledging the place of the researcher in a given research context. Many feminist and geographic scholars have been inspired by Donna Haraway, who argues for “the view from a body, always a complex, contradictory, structuring, and structured body, versus the view from above, from nowhere, from simplicity” (1988, 589). Many scholars have read this as an invitation to write in the first person, which serves to highlight the ways in which knowledge is always embodied, always situated, and, indeed, always emotional. This practice has both personalized and politicized scholarship in complicated ways and has engendered its own feminist critique (Rose 1993). While the term “emotional geographies” is not synonymous with feminist or geographic research, there is a distinct vein of emotional geographies work that is influenced by feminist theory and practice and that prioritizes the embodied, relational, and placed experiences of emotion.

Feminist versions of emotional geographies scholarship acknowledge and address the uneven terrain of embodied spatial experience, that is, they understand emotion as embodied, not in universal or even readily recognized ways, but rather as always intermeshed with bodies and places and not in neutral ways. Matters of difference are thus vital to consider. Looking closely at embodied, emplaced emotional experience encourages thinking through not just what binds or sticks people together in and through place, but also how emotional disarticulations and disconnections can tell us something about one’s place (or lack of place) in wider networks and
how dominant emotional geographies shape social, cultural, and economic practices.

These relational and politically charged understandings of emotional geographies have implications for further study. For example, nonhegemonic emotional geographies can be explored, including indigenous geographies, which incorporate the effects and affects of colonial contexts and the associated power dynamics, as well as making space for non-dominant emotional expression. In this sense, acknowledging the place of emotion can be a progressive and political act.

**Emotion and affect**

A key debate that characterized early discussions of emotional geographies revolved around the terms “emotion,” “affect,” and their respective positionings with the field (Anderson and Harrison 2006). In broad strokes, those claiming emotional geographies tended to focus on issues such as relationality, subjectivity, and embodiment from a distinctly feminist stance. This typically feminist scholarship has sought to consider how feeling mediates experiences, to challenge the perception that emotion is “out of place” in scholarship, and to argue for an understanding of emotion and place as mutually constitutive and differentially felt. Those exploring affect have drawn inspiration from nonrepresentational theory (Thrift 2004), a post-structuralist approach that encourages scholars to focus on the practice and embodiment of knowledge. Feminist critiques of this work have noted a masculinist foregrounding of the body, one that is apolitical and universal. In her critique of geography’s humanistic history, Gillian Rose (1993, 11) described claims such as “aesthetic masculinity,” which “establishes its power through claiming a heightened sensitivity to human experience.”

This debate continues to be referenced, but typically now in the service of advancing conversations that move beyond the past divide, whether staking space for “more-than-representational” (Lorimer 2008) geographies, or actively reflecting on the differences and the similarities of these approaches and exploring the political implications for the study of emotion. Pile (2010), for example, has drawn on psychoanalytic geography as a way to combine the examination of emotions in emotional geographies and the consideration of affect in affectual geographies.

**Empirical dimensions: the health and wellbeing example**

Studies of emotion, space, and society have produced not only theorizations of affect, emotion, and feeling but also numerous empirical (typically qualitative) investigations that focus on the embodied, relational, and placed experiences of emotion. One substantial area of this work investigates issues of health and wellbeing, from an ethics of care to therapeutic landscapes. For example, scholarship on homelessness as a felt experience has sought to render the emotional valences and “woundedness” of chronic homelessness. Geographers working at the intersections of gender, health, and emotion have offered empirical explorations of agoraphobia, delusion, post-traumatic stress disorder, and HIV-AIDS.

Emotional geographies scholarship offers major contributions toward research at the intersections of gendered realities, the specifics of place, and the diversities of health experiences. First, the emphasis on the placed and embodied nature of emotion works to interrupt a neoliberal representation of a health consumer who is in command of not only (his) health and wellbeing but also (his) wallet. Instead of depicting health
EMOTIONAL GEOGRAPHIES

and wellbeing as a series of self-evident decisions made by a rational, autonomous, and economically empowered actor, particular and diverse health circumstances can be analyzed as subjected to and by social, cultural, political, and economic conditions, such as occupational segregation by gender, made manifest in localized bodies, such as the lone woman in a boardroom. Second, emotional geographies research offers a critique of emotion as the simple and feminized opposite to masculinized reason, and this critique has led to a further challenging of the emotional and gendered character of the spaces and subjects associated with health and wellbeing. For example, a nurse is readily identified as a female because of women’s assumed nurturing and empathic qualities; however, the loving care expected of nurses is complicated by the biomedical training they receive, by the bodies nurses inhabit, and by the dominantly masculinist spaces of medical facilities where cure is prioritized over care. An emotional geographies perspective encourages attention to these complexities and leads to valuable insights into the geographies of gender and health (Thien 2014).

Current and future emphases: emotion matters

In examining the academic landscape over the past decade, a clear progression of interest in and legitimacy of all things emotional can be readily identified. From geography, two edited collections have been produced (Davidson, Smith, and Bondi 2005; Smith et al. 2009). Emotional geographies conferences have been staged in the United Kingdom, Canada, Australia, and the Netherlands, with international attendance. The journal Emotion, Space and Society has shaped a dedicated and interdisciplinary space for this scholarship. Emotion, Space and Society went to press in 2008, and as of early 2015 has published close to 400 research articles. Special issues in the journal have ranged widely and include some of the following topics and themes: transnational migration; affective ontologies; activism and emotional sustainability; intimacy and embodiment; researcher trauma; ecology and emotion; geographies of trauma; gendered spaces; researching intimate spaces; emotions and motivations; urbanity and fear; children; movement; sound; feeling differently; practicing emotional geographies; scales of belonging; emotional and affective methodologies; touch; and emotional geographies of education. As these themes and topics suggest, the journal has been distinctly interdisciplinary, reflecting the breadth of the field.

Evolving and ongoing work in the areas of racialization, migration studies, sexualities, participatory geographies, geographic information systems (GIS), and sensory dimensions suggests rich future studies, both empirically and theoretically. Considerations of the emotional health of the academy (professors, students, and so on) are also underway. Methodologically, there is increasing attention to innovative ways of answering questions of emotion and affect.

SEE ALSO: Affect; Difference; Feminist geography; Feminist methodologies; Gender; Health and wellbeing; Imaginative geographies; Nonrepresentational theory; Participatory geographies; Phenomenology; Poststructuralism/poststructural geographies; Psychoanalysis/psychoanalytic geography; Reflexivity; Therapeutic landscapes

References


Smith, Mick, Joyce Davidson, Laura Cameron, and Liz Bondi, eds. 2009. Emotion, Place and Culture. Farnham, UK: Ashgate.


Emotional labor

Joanne Entwistle
King’s College London, UK

Emotional labor is a concept used to describe the requirement that workers manage their emotions in face-to-face interactions at work. It is a form of labor that is critical to many occupations and professions in the health and service sectors and has been a critical concept within geography. Indeed, feminist geographers have been key in developing the concept and extending its application. The significance of emotional labor was first detected in developed economies in the North and West with restructuring away from manufacturing toward service sector jobs. This service economy is now global, and thus emotional labor is no longer seen as the preserve of Western economies. Emotional labor is different from emotion work; it refers to the overt management of emotions under conditions of marketization. It is labor required of many workers in the labor market that employers seek to harness and control as part of the overall “impression management” of their service or brand, and may be scripted (as in a call center, for example) or highly regulated and monitored (as in many hospitality environments). Feminist geographers have for many years addressed the changing nature of paid employment, and noted the increasing forms of new service work provided in private homes (McDowell 2009). Much of this work in the private spaces of the home is performed for “love” by relatives, as McDowell (2009) notes, though it may also be commodified as low-paid “caring” work. Further, some of this less formally regulated work falls into the category of sex work performed by strippers, prostitutes, and masseurs, which similarly demands emotional labor. What all this work shares is a concern to manage the emotions, though it may entail forms of emotional labor that are less scripted than in the corporate office or hotel (Brewis and Linstead 2000; Wolkowitz et al. 2013).

Arlie Hochschild’s (1983) study of air stewards in *The Managed Heart* has been particularly influential in proposing the importance of emotional labor and defining its characteristics. The bodily countenance of workers – their emotional labor – is often an essential aspect of that service; it is utilized by companies to gain competitive advantage and frequently features as part of their marketing and advertising (“service with a smile”). Emotional labor can take two forms, with different microgeographies: it can be performed as a form of “surface acting,” when emotions are displayed on the surface of the body (e.g., through a smile, kindly expression, soft voice) without being felt by the worker. Alternatively, laborers may have to work on their inner emotions to produce the desired surface and make it look authentic, through the suppression of unwanted emotion or the manufacture of the desired emotion. There is a gendered dimension to this emotional labor, with the particular requirements of the performance dependent on whether the emotional labor is performed by a man or a woman. Since much service work (though not all) is female-dominated, the emotional labor required is based on normative gendered expectations as to appropriate feminine appearance (Mears and Finlay 2005; Tyler 2009). Complex geographies of racial identity come into play as well.
EMOTIONAL LABOR

While this emotional labor enables companies to do their business more successfully and to gain competitive advantage, it may take its toll on the individual. The labor of emotional labor comes from the effort required by workers to work on, and contain, their emotions. It colonizes the inner life of the worker, who may experience a dissonance between outward expression and inner emotional life that can cause them to feel alienated from their own labor. This alienation is not unlike that which Marx describes in relation to nineteenth-century laborers, except that it takes the form of alienation from emotions generated at work.

Relations of emotional labor are ever more spatially complex under conditions of contemporary capitalism, in which increasing globalization has led many service sector employers to move their services to developing countries. This introduces different demands on workers who are located far from their clients/customers, as in the case of call center workers in India, who have to be English or American at night and Indian by day to accommodate a North–South geo-timescale (Bryson 2007). Moreover, complex patterns of migration and displacement have occurred in the domestic labor markets which mean that many domestic workers have to do emotional labor while simultaneously distanced from loved ones at home (Yeoh and Huang 1998). These complex geographies of employment extend the reach of emotional labor beyond Western economies and require an understanding of emotional labor from within a geographic perspective.

For some, the term “emotional labor,” while useful, places too much emphasis on the cognitive aspects of labor – the thinking or thoughtful processes – to the detriment of a more embodied understanding of the analysis (Witz, Warhurst, and Nickson 2003). Recent attempts to reclaim the embodied aspects of work have come through discussions of “aesthetic labor,” which refers to particular “embodied capacities and attributes” that enable employees to “look good and sound right” for the job (Warhurst and Nickson 2001, 13, 2; see also Warhurst et al. 2000). There is a more expansive geography associated with aesthetic labor: while emotional labor refers to specific workplace temporally and spatially bounded interactions, aesthetic labor goes further to examine all the ways in which workers have to maintain themselves to perform effectively at work. Thus, to be an effective worker who has to look right might involve keeping slim through dieting and exercise that take place outside workplace interactions. Emotional and aesthetic labor are not separate moments for the worker but conjoined; all emotional labor involves the creation of a particular aesthetic appearance, yet the effort to maintain the body is, at times, emotional (Entwistle and Wissinger 2006).

Conclusion

Emotional labor remains an important concept within organizational studies and geographies of the workplace and home, as it attends to the particular ways in which the capacities of individuals are harnessed and marketized by companies and other employers. Many economies now rely on “soft” qualities, like emotions – as opposed to brute physical strength or dexterity required in manual work. Today emotional labor remains a salient feature of the modern workplace (in its various locations) and one that continues to have academic viability within human geography.

SEE ALSO: Domestic workers; Emotional geographies; Feminist geography; Gender, work, and employment; Labor migration; Race, work, and employment
References


Further reading


Empire

Daniel Clayton
University of St Andrews, UK

Empire denotes an extensive, hierarchically structured, and usually militaristic polity encompassing diverse peoples and territories that is ruled more or less directly and authoritatively by a supreme power or ruler (emperor, monarch, state, or centralized elite), and that revolves around a spatialized logic of “inside” and “outside.” It is a transhistorical concept, and has long framed debates in the social sciences and humanities about the nature of cultural change, development, society, sovereignty, and territoriality. The term is derived from the Latin imperium (sovereign authority, and extensive command and influence), and is entwined with other ancient concepts (democracy, freedom, power, the state). Since the nineteenth century it has been closely associated with the terms “imperialism” and “colonialism.” Empires are sprawling political entities (such as the Roman empire, British empire, and Soviet empire), which subsume and subordinate the fortunes of alien and distant peoples and domains to the drives and dictates of a dominant center, culture, or figurehead. But the term is also used more colloquially, to evoke transcendent power or oppressive influence (as in corporate empire, empires of vision, and mosquito empire).

Geography and empire

Empires have been greatly shaped by geographical understandings, traditions, practices, concepts, and institutions – of frontiers, the exotic, and the unknown; expeditionary traditions of discovery, observation, mapping, description, and classification; colonizing practices of appropriation and dispossession; cosmographic and secular systems of geographical knowledge, and doctrines of environmental determinism and geopolitics; and learned societies such as the Royal Geographical Society (founded in London, 1830) and the National Geographic Society (founded in Washington, DC, 1888) and its popular magazine National Geographic. Felix Driver (2001) coined the expression “Geography’s empire” to capture the recognition that from the late eighteenth century imperial expansion went hand in hand with the promulgation of geography as a professional discipline. From ancient times through to the present – from geographical luminaries such as Strabo and Herodotus, to Dee and Mercator, Cassini and Humboldt, and Mackinder and Bowman – projects of “earth-writing” (geography’s Latin root) have more often than not fired and been fired by imperial imaginations and ventures. What is more, imperial imagery of intrepid (usually cast as “manly”) explorers, and the allure of the exotic, still permeate public perception of the subject. Cartography has been a mainstay of “Geography’s empire” (and remains so today in the guise of new
geographical information systems), inviting and choreographing imperial expansion, fueling myths of empty space, charting fecund lands ripe for European penetration, targeting alien territories for military intervention, and framing the racialization and sexualization of empire.

Questions might be raised about why an avowedly critical literature on geography’s complicity in empire only arose within the discipline in the 1990s, years after decolonization. We might also ask about why this complicity has been prosecuted (by and large) through a selective focus either on the world of European empires (and the diverse colonial domains within them) from the mid-eighteenth to the mid-twentieth century, or on the political economy and geopolitics of imperialism today (Butlin 2009). With some notable exceptions, geographers have shown less interest in Eastern and ancient empires, or, until very recently, in post–World War II geography’s connections with the Cold War and decolonization. Did Western geographers, many of whom had served country and empire in World War II, lack the desire to launch what would be a soul-searching critique of their discipline’s makeup because the imperial past was still too (uncomfortably) close to the present? Or was it the case that the theoretical language with which to advance such a critique was not in the air until later? The stubborn presence of empires in world history undermines the fantasy that academic knowledge is impartial and operates in a political and historical vacuum, and attests instead to its partial, power-laden, and often compulsive qualities. This recognition impinged on the fields of anthropology, history, sociology, and literature from the mid-1970s (and, since the 1980s, on political philosophy), before it did geography and was instrumental in fashioning new multidisciplinary critical projects of cultural and postcolonial studies.

Awareness of geography’s past and ongoing complicity in empire has played a major role in promulgating a critical human geography which, among other things, embraces questions of cultural difference, otherness, and exclusion; probes connections between power, knowledge, and geography; is antiracist and antisexist; and alights on the Western (Eurocentric) blinkers that have crimped the discipline’s outlook and the way its history has been told. Given the general contextual significance of empire to geography, however, it is surprising how few geographers have engaged in general terms with how empires are spatially organized. Historians and sociologists have taken the lead in this regard, and have shown that geography’s imbroglio with empire is not just a disciplinary affair but needs to be projected onto a much larger screen.

History, meaning, and character of empire

Empire can involve diverse forms of power, subordination, and dependence, and important distinctions have been drawn between its economic, political, ideological, and technological dimensions (how empires work as markets, projects of rule, cultures of dominance, and modes of territorial organization and transformation). There have been over 70 empires in history, and they have been typologized in terms of their declared aims, expansionist ethos, sources of social power, modes of command and control (formal and informal), systems of rule (direct and indirect), geographical character and reach (maritime and land-based), economic and political organization (into colonies, protectorates, mandates, trusteeships, trade monopolies, and domains of foreign direct investment), socio-economic configuration (mercantile and settler, tributary and territorial), and benefits and costs
EMPIRE

(in perceived and demonstrable terms). From the time of the Roman empire (Western, ca. 27 BCE–476 CE; Eastern, 330–1453 CE) and the early Chinese empire (Qin 221–206 BCE; Han 202 BCE–220 CE), through the modern era of European colonial empires (ca. 1450–1945 CE), to current debate about nature or the existence of an American empire, empires have been regarded as forces of both good and evil. Some see them as fostering collaboration, toleration, order, belonging, and modernization. Others have derided them as the incarnation of economic exploitation and xenophobia, and as the harbingers of a world divided by civilization, class, nation, race, and religion.

The modern (post-1400) era of empire building, starting with the Portuguese and Spanish (in Central and South America and the Indian Ocean), and followed by the French, British, and Dutch (in North America, the Caribbean, and Asia), has been closely linked with the advent of modernity and a “world-system” of capitalist production and exchange. During the second half of the nineteenth century empire became virtually synonymous with the terms “imperialism” and “colonialism,” which were coined to denote a European ethos of territorial expansion, exemplified by the imperial partition of Africa between 1885 and 1914, and which involved both intense national rivalry and the flowering of a more broadly European sense of superiority and justification (or smoke screen) for empire as a “civilizing mission.” By 1914 over three-quarters of the world’s land surface was under the imperial aegis of a small number of European powers, plus Japan and the United States. Marxist revolutionaries such as Lenin and Luxemburg argued that this imperialism was precipitated by the contradictions of monopoly capitalism and was a necessary stage leading to capitalism’s eventual demise. Liberals such as Hobson and Schumpeter were less sanguine about the latter, arguing that imperialism was bound up with a more convoluted set of relations between capital, class, nationalism, and political factionalism, and was a cyclical element of “great power” politics.

During the mid-twentieth century, and particularly during the postwar era of anticolonial struggle and decolonization, “empire” became a term of abuse. In the 1980s the *bête noire* of a new critical project, dubbed postcolonialism, returned to the scene of Europe’s garbled and bloody retreat from empire and probed how the hopes and promises of freedom and autonomy accruing from the transfer of power faded and became corrupted, and, in some parts of the world, were overshadowed by territorial partitions and violence. Independence and liberation became what Fanon (1967, 183), in the early 1960s, identified as a “zone of occult instability” rather than a clean break with the past. In 1965 Nkrumah deployed the expression “neocolonialism” to describe lingering imperial and emerging postcolonial structures of dependency on the West. Historians have since shown that, after World War II, Western (and especially American) intellectuals and politicians were convinced that they could steer the development of the decolonizing world, avert revolution and instability, and fashion a new global order of liberal-democratic-capitalist states that would serve Western interests. Key to this project was a paradigm of modernization geared to developing agriculture, promoting trade and its liberalization, and harboring pro-Western (if in some places authoritarian) political regimes through the transfer of Western capital and technology, and the use of diplomacy. In short, the fortunes of postcolonial nations were still fettered by Western values and norms. Western interests were bolstered by the coterie of international institutions created at Bretton Woods in 1944 (and centering on the United Nations), which
were initially geared to defending empire but soon had to adapt (some argue quickly, others say uneasily) to the travails of national sovereignty and impress of Cold War antagonisms on the “Third World,” which was a term originally used (in 1952) to characterize those nations (including many in the decolonizing world) that were nominally nonaligned with the capitalist or communist blocs.

Since the mid-1990s, academic and political consensus about the deleterious character of empire has become fractured. Following the late 1980s collapse of East European and Soviet communism, and latterly since the attacks on New York and Washington of September 11, 2001, and the financial crash of 2008, empire has returned as a way of talking about and dealing with global crisis and conflict (as has modernization theory). Empire is back in vogue as a live analytical and political concern. The postcolonial desire to consign empire to the past now seems misplaced. And the urge in other – “neoliberal” – quarters to champion the global ascent of an international liberal–democratic–free market system as the final and mature stage in the evolution of economic and political community, seems deeply divisive. The number of state-based civil wars (both with and without foreign intervention) has increased sharply since the early 1970s, and there is now a much greater range of opinion about empire than there was 50 years ago. Burly reaffirmations of empire’s modernizing and stabilizing influences jostle with singeing critiques of the death, destruction, and double standards wrought by “international” (military and humanitarian) interventions in different parts of the world. Within and beyond geography, the latter critical concern hones in on the appearance of new spaces of violation, vulnerability, and “exception” to national sovereignty and international law (e.g., the American detention camp at Guantanamo Bay).

As this discussion suggests, and as Colás (2007, 3) and other theorists stress, “the notion of empire is mixed up with all sorts of conflicting political sentiments and social processes,” making the search for “a normative or analytical equivalence between arguments for and against empire” more complex and perhaps more ambivalent than ever. However, it is still possible to generalize about empire. Colás and a longer line of theorists have viewed expansion, hierarchy, and order – or, more precisely, the ways in which they have been sought and the extent to which they have been realized – as the core attributes of empire, and as explanatory keys with which to compare different imperial formations, and to gauge how the meaning of empire has varied over time and from place to place.

While other types of polities share one or two of these core attributes with empire, it is their combination in empire that distinguishes empire from other forms of political community and imagined community (such as nation-states, city-states, fiefdoms, tribal federations, and ethnic homelands). Yet, while empires and nation-states are different political entities, their fortunes have long converged and clashed. As Burbank and Cooper (2010, 8–9) observe, when “nation-states … based on the idea [however illusory and divisive in practice] of a single people in a single territory constituting itself as a unique political community … did become meaningful political units, they had to share space with empires” whose boundaries were porous, whose resources were more diverse, and whose extensive command of space was intimidating. Today, as in the past, the specter of empire raises important questions about whether nations can be self-sufficient, and whether they are any better than empires at respecting ethnic pluralism or democracy or human
rights. “As long as diversity of political ambition exists,” Burbank and Cooper (2010, 9) continue, “empire-building is always a temptation,” yet because empires, by definition, impose, subsume, and dispossess, “there is always the possibility of [their] coming apart” as a result of opposition and overexpansion.

There is currently considerable debate – including within geography – about the recent emergence of a “liberal empire” spearheaded by the United States, buttressed by the United Nations, and cloaked as a new collective and consensual internationalism that intervenes in the world’s trouble spots in the name of global order and universal human rights. This empire is geared to policing and enforcing legitimate state authority, civil liberties, democratic rights, and humanitarian standards, and to rooting out terror and threats to “homeland security.” It is putatively consensual rather than coercive – supported by an array of public and private actors and institutions that have a collective interest (as well as their own vested interests) in making the world a secure and peaceful place. Some see in this liberal empire an older imperialism of free trade, and regard the United States as a hegemonic (consensus-oriented) power, coordinating an international community that wields a “higher” moral authority. American hegemony has been dubbed “soft power” (an ability to attract and co-opt rather than coerce) and “empire lite” (bending world affairs to the will of the US and international community with diplomacy and quick-fix military interventions).

Hardt and Negri (2000, xii) take this problematic of internationalism to its conceptual conclusion, arguing that,

In contrast to imperialism, Empire [today] establishes no territorial center of power and does not rely on fixed boundaries and barriers. It is a decentered and deterritorializing apparatus of rule that progressively incorporates the entire global realm within its open, expanding frontiers.

They are as concerned with resistance to the territorial elusiveness of empire as they are with the discourses of might and right that course through it. In destabilizing modern imperialism’s territorial fixations, their synopsis leaves room for new forms of identity and resistance to Western hegemony. Yet, at the same time, their empire denotes a supranational expansionism that is geared to “enlarge the realm of the consensus that support[s] its power” (Hardt and Negri 2000, 15, 190–194). Critics suggest that Hardt and Negri underestimate both the continuing significance of nation-state territorial sovereignty in a globalized world, and the need for global markets and multilateral flows of people, information, business and finance, and cultural influences to be politically and militarily regulated from recognizable power bases. Indeed, it is the attempt by those on the receiving end of this purportedly decentered empire to name and locate (reterritorialize) it – as Anglo-American, revanchist, and militaristic – and to bear witness to the dagger behind the cloak of internationalism that is one of most salient aims of resistance in many parts of the world today.

Another criticism of Hardt and Negri’s thesis must surely be that they underestimate the ways in which this decentered empire is wrapped around an American military empire – one of bases rather than colonies. The US military is currently deployed in more locations around the world than at any time in its history, and since the mid-2000s its military spending has accounted, on average, for almost half of global military expenditure. The global ubiquity of the United States’ strike force is part and parcel of what Gregory (2011) terms an “everywhere war” – a war that the United States deems a “virtuous” war because of its use of smart bomb and drone technology, ostensibly to make enemy targeting more precise (minimizing civilian casualties and
wanton destruction) and to forestall the need for sizable and costly troop deployments.

Marxist critics see US military expansionism as a reaction, in part, to what Harvey (2004, 63, 68) describes as “a chronic and enduring problem of overaccumulation since the 1970s,” and as having resulted in capitalist devaluation (especially in Latin America, East Asia, and Russia). Recent US military escapades overseas have been bound up with geoeconomic concern over control of the world’s energy resources and have been read as attempts to stop recession and devaluation coming its way. Harvey has dubbed this articulation of economics and militarism a “new” imperialism that “attempts to accumulate by dispossession,” although he frames the phrase with scare quotes because this imperialism looks a lot like that analyzed by Luxemburg and Marx’s “primitive accumulation” (the modes of imperial plunder and appropriation needed, Marx argued, to kick-start capital accumulation). The 2008 financial crash, however, exposed the effectiveness of this imperialism by exposing how colossal military expenditure was tied up with the precarious securitization of US government debt. In other words, the United States has sought to combat one type of risk by ramping up another.

The spatiality and tensions of empire

Colás (2007, 31) notes that, from ancient Greece to Guantanamo Bay, “processes of imagining and managing space are crucial in identifying the distinctiveness of empires, both among themselves and vis-à-vis other forms of rule.” Empires revolve around a hierarchical (inside–outside, center–periphery) spatiality, and a series of tensions and oppositions that this geography brings into play: between imperial expansionism and the need to delimit empire for the purpose of jurisdiction; between ruler and subject, and colonist and native; and between more intricate and diffuse gradations of rank and status across the metropolitan–colonial reach of empire. All empires have faced the problem of how orders and influences are transmitted, and how material and worldly goods are circulated. They have sought means of speeding up imperial flows and making them more secure, and technological change has been pivotal in this regard. In the modern era, and with growing intensity since the end of the eighteenth century and the coming of the Industrial Revolution, empire building has been aided and abetted by a string of inventions and innovations in shipping, navigation, banking, accounting, engineering, weaponry (including, by World War I, air power), communications (print, steamship, railway, telegraph, and radio technologies), visualization (photography and film, and museums and world exhibitions), medical science, and urban design. Technological change was integral to expansionism, furnishing the means of stretching imperial systems over more far-flung areas without losing control. But it was equally important to an increasingly ambivalent imperial existentialism, with rapid growth in metropolitan knowledge about the world, perceived speedup in pace of life wrought chiefly by communications technologies, and the growing material and cultural proximity of people and things that were once deemed alien and faraway, provoking both awe and disorientation.

All empires have also been concerned with how their imperial agents (settlers, traders, administrators, military personnel) are transplanted overseas from the mother country (and remain loyal to it), and how to accommodate alien and diverse populations and territories within imperial structures of command and control. Concomitantly, all empires have faced the questions of where and when it is best to centralize or devolve power, how to adapt imperial ambitions and ideals to colonial contingencies
and pressures, and how to avert and manage resistance and colonial malfunctions. For example, the ancient Chinese empire, Roman empire, and early modern Ottoman empire had centralized fiscal systems yet decentralized administrative structures that incorporated elements of local custom and tradition. From early days, then, the creation of imperial hierarchies involved two-way (albeit still unequal) and diverse mediations between center and margin rather than a one-way imposition. The sixteenth-century Spanish developed a system of trusteeship (encomienda), whereby Spanish settlers in the New World were granted rights to the labor of indigenous people and to keep the fruits of some of that labor, in return for loyalty to the crown and a duty to convert their native charges to Christianity. By contrast, the late nineteenth-century British espoused an empire of free trade and developed a system of “indirect rule” in Africa (governing through local intermediaries and alliances), although neither protectionist inclinations nor military might were faraway, whereas the more assimilation-minded French in West Africa and Southeast Asia failed to implement the metropolitan republican principles (of free expression, mass education, social justice) that were enshrined in official imperial rhetoric. Finally, while military aggression and intervention have long been seen and used as key power brokers of empire in the last instance, the postwar promulgation of guerilla warfare as a viable strategy for small numbers of insurgents to resist and disperse considerably larger and better equipped conventional military forces. The Vietnam War was the apogee of such “asymmetric” warfare and raised a gamut of questions – persisting to this day – as to whether militarism is a self-defeating imperial strategy, and whether different kinds of militarisms are suited to different types of terrains and imperial situations.

These and other imperial questions and processes comprise a large, complex, and contradictory spatial repertoire of imperial dominance and colonial governance from which present-day analysts, policymakers, and opponents of empire (and capitalism and globalization) alike draw “lessons” and identify general patterns and trends. Two trends in the history of modern empire building have been elucidated. First, a triumphalist “follow my leader” ideology, with the West arrogating to itself the power to decide on what counts as right, normal, and true (and what does not), and to define itself as the most advanced and civilized part of the world. Modern empires marshaled a Eurocentric (diffusionist) imperial worldview – overriding national animosities – in which all “good things” (reason, civilization, progress, etc.) were deemed to flow from “the West” outward, to be imbibed and emulated by “the rest.” This ideology made empire building a project of epistemological violence (working at the level of language, knowledge, and culture) as well as a palpably material project of geographical violence. To reiterate, geographical practices of observation, mapping, classification, territorial appropriation, and settlement are central to empire on both counts. Sometimes brashly and sometimes surreptitiously, geographers and those working within what became characterized as its traditions (explorers, cartographers, administrators, and so forth) extolled the “superiority” of Western knowledge (as objective, universal, and comprehensive) and denigrated “other” (native, distant, local) ways of knowing.

Second, the seventeenth-century advent of modes of sociospatial reproduction based on the coalescence of territorial sovereignty and political authority (often dated to the 1648 Peace of Westphalia) have been seen as heralding a general shift in the spatiality of empire away from a focus in ancient empires on people (on
the extraction of tax and tribute, the control of labor, and a proselytizing concern with conversion) to a modern concern with the use and administration of territory, resources, markets, and trade (in short, a focus on land and the creation of wealth). This modern territorial ideology became enshrined in British and French Enlightenment thought and a political philosophy of “improvement” (from whence came theories of development and a Western fixation with progress), and the territoriality of nation and colony both facilitated the accumulation of capital (not least, through the export of capital to colonies rather than wealth simply being extracted and drawn back to the imperial center) and created what Marx regarded as constitutive “limits” and “barriers” to it. This imperial ideology also coded imperial and colonial subjects in territorial terms, with figures of class, nation, gender, sexuality, and race assigned to fixed domains in a distinct imperial pecking order. Empire both traverses these domains and constitutes them by so moving, creating a politics and aesthetics of fixity and motion that parallel the workings of capital.

Yet there are significant exceptions to, and irregularities within, these two trends. The early modern Spanish and Ottoman empires are often invoked as both ancient and modern. And the British used the very ancient institution of slavery to build their characteristically modern empire; and then, during the early nineteenth century, built a new scientific racism (“hierarchy of races”) with one hand, as it was abolishing slavery in its empire with the other. Postcolonial thought probes such inconsistencies and paradoxes in imperial formations. To gloss just one important critical position, Bhabha sees a fundamental ambivalence at the heart of both colonial power and postcolonial identity, which is marked by hybridity and belatedness. Decolonization, he argues, was unable to proceed by returning “the eye of power to some prior archaic [precolonial] image or identity,” because the cultural identity of the colonized was mixed up with that of the colonizer. Such a quest for postcolonial authenticity was also belated, because the imperial and colonial forms from which the colonized sought to extricate themselves only became authoritative after the “traumatic scenario of colonial difference, cultural or racial.” Such a quest – for a clear break with the past, and to find some precolonial cultural or national essence and read it into the origins and history of anticolonial struggle – could “neither be ‘original’ – by virtue of the act of repetition that constructs it – nor identical – by virtue of the difference that defines it” (Bhabha 1994, 153).

Empire also has an ambivalent relationship with capitalism, and the history of that relationship reverberates through contemporary debate about globalization and the current economic crisis. Are the interests of nation and capital symbiotic? The US government’s recent unwillingness to fully bail out its “toxic” financial sector has brought this question to a head. And it can be found in Harvey’s analysis of the “new” imperialism, a much older question about whether capitalist accumulation relies on state protection and regulation, and military violence. Both questions lay at the heart of the activities of the world’s two utmost commercial agents of empire, the English and Dutch East India companies (founded in 1600 and 1602 respectively). Both of these trading monopolies had complex and at times stormy relationships with their respective states, and posed the enduring question of whether, for capitalism to thrive and for private interest to serve some national public good, violence needs to be subordinated to the rational and orderly pursuit of profit.

Another enduring tension, this time bridging the spatial and the cultural, is between the ideas of empire as expansive and as delimited – with
an “inside” and an “outside,” and with the edges of the known world manifested in fuzzy imperial frontiers peopled with monstrous and threatening figures of otherness. The imperial contours of this tension can be traced back to ancient Greece and Herodotus's idea of oikoumene (the humanly made, known, and interconnected world), and since ancient times cartography has been a prime means both of substituting fact for uncertainty and of underscoring the elusiveness of empire’s edges. Imperial and colonial sites (such as Madrid and Mexico City, London and Lagos, Paris and Phnom Penh), and imperial frontiers (such as Central Asia, the American West, and Equatorial Africa) have been key sites at which this tension between the known and unknown, and its cultural cipher as an opposition between civilization and barbarism, have been exhibited, negotiated, and contested. Furthermore, the origins of modern diplomacy can be located in state concern with how to manipulate not just geographical knowledge about distant places and obscure events, but also geographical ignorance emanating from the frontiers of empire, and how to translate it into sovereignty.

A considerable critical literature on the “culture of empire” examines how fantasy, speculation, and egregious assumptions about what was “out there” sparked and shaped myriad imperial conflicts, and also how imperial desire is bent on putting hierarchies “in place” (i.e., overcoming uncertainty and fluidity). Said’s 1978 book Orientalism was seminal in showing how these two elements of the spatiality of European empires combined (and often clashed) to generate potent “imaginative geographies” of how the world was presumed to be. Said averred that images and the geographical imagination of cultural difference and European superiority were as important to empire as guns and armies. Empires have been built and sustained with diverse texts, representations, and spectacles of power – maps, travel narratives, world exhibitions, museums, and imperial parades. Meanings and images coming from scholarship, literature, religion, and the arts and sciences were directly implicated in the exercise of power, and Said’s concern with how empires thus operate as texts (in the broadest sense) made a deep and lasting impression on geographers (Said 1993/1978).

Said demonstrated that imaginative geographies have a predilection to harden distinctions between “us” and “them” into binary oppositions. Yet he also recognized that such binaries often struggled to work their Manichean magic, and that imperial and colonial categories could be fluid and were in constant need of reiteration (citation) if representations of otherness were to become accepted as reality. On the one hand, binaries of “us” and “them” (Bhabha’s trauma of colonial difference) were enunciated and performed in and through abject (and acutely raced and sexualized) material spaces and spaces of representation – plantations, missionary stations, native reservations, and the racially segregated districts of colonial cities; colonial schools, courtrooms, and censuses – and in artistic and scientific representations of the natures and cultures of the colonized and the edges of empire as seductive and supine and in need of domestication and control. Reworking ancient celestial conceptions of imperial order and harmony between heaven and Earth, a modern imperialist ideology, and practice of sequestration commanded that everything “had a place,” and as Fanon (1967, 35–38) proclaimed, the place the colonized occupied was set apart from and beneath that of the colonizers. But, on the other hand, by the nineteenth century, spatial boundaries between center and margin were being lowered by the commercial traffic of empire and rapid technological change. Colonial divides were being crossed too (and
in the most intimate, libidinal ways), yielding fear and apprehension about the physical and moral degeneration (especially in hot climates) of the imperial self and colonial populations. Expansionism and speed made for anxiety as well as indeterminacy.

Said (1993/1978), and now a much wider literature on “tensions of empire,” argue that the increasing accessibility and experience of empire to a growing number of Europeans through travel and work, combined with the growing (if limited) accessibility of European ideas and Europe itself to the colonized and mounting metropolitan awareness of colonial conditions and (importantly) of resistance to empire, conspired at once to “produce” European culture (spur its arts and sciences, and infuse its political and popular culture), and to foster an existential crisis within it. Postcolonial thought alights on this motility and breakdown to advocate what Said (1993/1978, 97) describes as a “contrapuntal” reading (a term borrowed from the Cuban thinker Ortiz) of empire that is “inherently spatial”: to ignore or otherwise discount the overlapping experience of Westerners and Orientals, the interdependence of cultural terrains in which colonizer and colonized co-existed and battled each other through projections as well as rival geographies, narratives, and histories, is to miss what is essential about the world in the past century [sic]. (1993/1978, xxiii)

A colonial world “out there” wound up “in here,” and made a vital and lasting contribution to the metropolitan world, albeit, Said points out with reference to the novel as a nineteenth-century literary form, in diminished, distorted, and sublimated ways.

A related, and final, point addresses the spatiality and tensions of empire. Critical emphasis on the reach and swagger of imperial power runs the risk of lengthening the shadow that empire casts over the present – empire can appear ineluctable; and there is thus another, equally compelling, side to postcolonial critique, which is about showing – pace Fanon, Said, and Bhabha, among many others – how imperialism and colonialism have been challenged and compromised, and have morphed into novel and unexpected (“creole”) forms. Since the beginnings of modern European imperial expansion, the “placing of empire” has worked through a twin psychospatial tension: first, between imperial might and right (the impulse to expand and the need to justify that expansion); and second, and as Bhabha stresses, between mastery/fantasy and anxiety/dislocation (fear over a loss of control and identity). Furthermore, these two tensions revolve around two (vexed) geographies of identity formation: first, between metropole and colony (how they related to and addressed one another); and, second, between colonizers and the alien lands and peoples they encountered (how colonizers adapted to new physical and cultural landscapes from European backgrounds that had already been shaped by imperial desire and ideas about wilderness and savagery). Such tensions and geographies animate Said’s “contrapuntal” envisioning of empire, and make the task of postcolonial criticism not just one of exposing and challenging imperial hierarchies and modes of othering, but also about recovering and extolling countergeographies, counterhistories, and spaces of cultural invention and mixing, and thus about building a world with fewer barriers and borders in it.

With reference to the first tension and geography, European nations went to great lengths to justify imperial expansion, both to one another and to the peoples they were colonizing. Moreover, their fraught recourse to natural and divine law remains a vital historical reference point in legal and political battles over indigenous rights and sovereignty, not least in former British
EMPIRE

settler colonies such as Australia, Canada, and New Zealand. A fast-growing literature now demonstrates that the boundaries and differences between European and indigenous laws and practices were not as tightly defined as colonists and colonial administrations presumed, and that to continue to insist otherwise is an imperialist act and fiction.

If, as this literature suggests, the political burden of sovereignty finds its main ramifications in settler-colonial domains, then a second tension and geography, and its legacy, is more keenly felt in India (which has been an important fount of postcolonial theorizing) and other colonial situations where colonists were vastly outnumbered by their colonial subjects. By the mid-nineteenth century, the West was justifying colonization on the grounds that the beneficent/enlightened colonizer was bringing “gifts” of civilization (reason, science, and progress) to “the native” (in some discourses it is the native who beseeches them). For colonialism to endure and remain legitimate, however, it could never admit that it had completed its task – admit that “the native” was now “civilized enough” to run her or his own affairs. There was a constant need for the colonized to be placed in a subordinate position, to be “worked on” by European culture and reason. A “rule of colonial difference” thus cemented itself and came to haunt the nationalist imaginaries of postcolonial nations such as India. Independence did not bring full decolonization, this postcolonial story goes, because formerly colonized peoples and nations were still enjoined to pursue and accept Western standards and expectations – to sign up to a constitutively colonial modernity. While it was desirable to accept the “good” bits of modernity (its veneration of liberty, justice, and equality) and to jettison the “bad” bits (naked individualism, for instance), the “decolonized” could only ever become modern in a secondary way – after Europe and on its terms.

Conclusion

Admired and vilified, empire is a key comparative framework for geographical, historical, social, and political analysis, and is now central to what some see as a contextual and historical turn in economic and political theory. The word conjures up a long-lived arsenal of ideas and practices that do not belong to a bygone era but are thoroughly mixed up with the predicaments of the present. The spatiality of empire is integral both to the hope and the despair of the postcolonial condition, and to contemporary conceptions of resistance and the political. Lastly, critical exposure of the complicity of geography in empire is very much a matter of the present and not just something that concerns the past and the discipline’s heritage. The whole idea of empire remains key to how geography operates as a situated knowledge and discipline that deals with the world, and geographers are as vigilant today as they have ever been about the need to combat actually existing imperialisms and colonialisms.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Geopolitics; Imperialism; Postcolonial geographies; Race and racism

References


Further reading

Endemism

Kenneth R. Young
University of Texas at Austin, USA

Endemism is a measure of the uniqueness of the biota of a particular place. It is one of the basic ways used by biogeographers to characterize biological diversity across the surface of the Earth. Early explorers documented species that were new to science and evaluated their ranges, ecological tolerances, and potential usefulness. As they did so, the proportion of species unique to an area became one way to assess how different that place was from others previously explored. These place-to-place differences in species composition became included in the recognition of biogeographic provinces and regions. In addition, high endemism was also understood to indicate the presence of biogeographical barriers, which isolate species and restrict the establishment of new populations. Endemism can be used to refer to other taxonomic levels, for example, in referring to families or genera found restricted to a particular part of the world. The degree of endemism nowadays is used for biodiversity conservation planning, and places with high values have special significance given their distinctive taxa. The kind and degree of endemism gives insights into the historical biogeography of particular places, and helps in understanding the origin and evolution of the biota. It may also help in carrying out nature conservation under future conditions.

All life on Earth is endemic, as it is found nowhere else; as a result, the endemism of the planet would be 100%. Smaller extents of the Earth’s surface would have less percentage endemism, with many continents having around 50% to 80% endemic species, with the remainder being widespread, for example cosmopolitan species found on multiple continents. Places within large uniform parts of the world, including the oceans, share many species, and so would have a small percentage of endemism, and relatively low beta diversity (a measure of place-to-place change). Oceanic islands, on the contrary, have many unique species, and so typically have high endemism. In an ecological sense, isolation serves to limit or even prevent dispersal. As can be seen in these examples, endemism is both scale-dependent and place-specific.

Charles Darwin found unique finch species on the Galapagos Islands (Grant and Grant 2008). From the distribution patterns revealed by his collections and by looking at morphological details of the birds’ beaks, he developed an important line of evidence for showing how natural selection acts over time to result in speciation and adaptation. Islands such as the Galapagos have high endemism for two reasons: they are isolated by distance from the mainland and their biota has evolved further under those isolated conditions. Species colonizing an island or island chain are subjected to new evolutionary pressures. As a result, they have sometimes speciated in an adaptive radiation manner wherein they diversify to use available resources, especially if competitors or predators found on the mainland are not present.

Besides islands, high endemism would also be expected in places with significant dispersal barriers and/or numerous different microenvironments. Highly fragmented environments often have restricted range species. For example,
mountains have many limitations to dispersal given the rough topography and abrupt slopes that act to create different biophysical conditions; collectively these work to restrict species to relatively small areas, especially for habitat or elevation specialists, and for species with low vagility (capacity to disperse) that are not able to fly, swim, walk, or otherwise disperse across those barriers. There is a set of plant species that are restricted to growing on certain soil types; the distribution of the soil types would then determine where those specialized species can survive, creating habitat islands. A consequence of these kinds of controls on species distributions is high endemism for mountain regions or for other places with much natural variation in soils or topography, in contrast, for example, to lowlands with relatively uniform biophysical conditions.

Countries with mountains would likely have more endemic species than similar-sized countries with no mountains. Thus, it can be said that endemism also has a political dimension, as unique species are a particular conservation responsibility for the country in which they are found. Island nations, for example, need to devote considerable resources to protecting their unique species, which are often also prone to displacement by aggressive species introduced by humans. As another kind of example, South Africa has such unique plant lineages associated with a Mediterranean climate that the Cape flora is considered globally distinct. There is a trend of global biotic homogenization in the last several centuries, whereby humans reduce place-to-place differences in species composition through land change and the introduction of exotic species.

Systematic planning for biodiversity conservation uses endemism, along with other variables such as species rarity and overall diversity, to choose the best portfolio of lands to protect, given costs and land tenure. This helps with the decision-making process because it weights the degree to which endemism, rarity, charisma, usefulness, or other species traits are used in determining nature conservation strategies as networks of protected areas are assembled.

Relatively new techniques from molecular biology can now reveal the timing of speciation events, allowing for reconstruction of past evolutionary processes. For example, differences between closely related species in their DNA can reveal the order in which the different species evolved from a common ancestor. Further analyses can suggest the timing of those speciation events, which can then be tied to independent sources of information on Earth history from sedimentary records and fossils. Additional information on species distributions can be used to elucidate the influences of past plate tectonics, climate change, and species dispersal events on the history of the biota. Places with high endemism, for example Australia, have a history of isolation from other land masses, and this can be assessed using historical biogeography. The kangaroos and echidnas found there today live nowhere else naturally, although their fossils occur in other places. Australia was once connected to other land masses in the supercontinent Gondwana. In other words, endemism is a crucial data source for reconstructing the history of life on Earth.

Endemics are among the most threatened species, as they are rare by definition in having a small range. Endemics with small populations would be of the highest concern for conservation interventions. An endemic species with an ancient evolutionary lineage is defined as a paleoendemic; this kind of species would be of great interest for conservation because its loss would represent loss of information of an entire lineage. Neo-endemic species are relatively recently evolved, and their presence points to places having high rates of speciation
in the Quaternary period. Documentation of range-restricted species in the past may reveal previously unknown dispersal barriers, while widespread species in the past point to the presence of dispersal corridors or shifts in the locations and connectivity of the land masses.

Biophysical parameters in the next 100 years will be quite different from today. In fact, Williams, Jackson, and Kutzbach (2007) have shown that much of the world in the future will have “no analog” conditions, unlike any found today. Specialized species, especially if they are found in small areas today and if they have limited dispersal and colonization abilities, will be the most threatened and least likely to prosper. Hence, the study of endemism is also important for carrying out research and management of the biogeography of the future. Biogeographic shifts can be predicted using ecological and species distribution modeling, while management will require planning for conservation corridors and other interventions that permit the movement of populations to new sites.

**SEE ALSO:** Biodiversity; Biogeography; Biogeography: history; Climate change and biogeography; Dispersal, diffusion, and migration; Ecogeography/macroecology (range and body size); Geography of evolution; Habitat destruction and fragmentation; Island biogeography; Land change science; Mountain biogeography; Nature conservation; Plate tectonics in biogeography; Zoogeography

**References**


**Further reading**


Millington, A., M. Blumler, and U. Schickhoff, eds. 2011. *The SAGE Handbook of Biogeography*. Lon-
don: SAGE.
Historically, countries across the globe have exploited and consumed ever-increasing quantities of energy resources to satisfy the growing demand for goods and services and to promote jobs and economic development. The historic pattern of progress has involved an ever-increasing and diversifying portfolio of energy resources. Since the prosperity and security of countries has been intimately tied to their access to energy, this entry begins with an overview of the geography of energy resource endowments. With the growing reliance on electricity to power economies and the need to meet increasing levels of reliability, it next turns to a discussion of how electricity is generated globally, and in particular in a subset of regions—the European Union, the United States, China, and India. As concerns over the negative externalities of fossil energy consumption have grown, analysts are increasingly focused on the demand side of the equation, aiming to reduce the amount of energy consumed per capita and per dollar of GDP (National Research Council et al. 2010). Overlaid on these transitions is the need for governments to develop energy policies to effectively direct investments toward sustainable, affordable, and reliable energy choices. These are the four topics covered in this overview of the geography of energy resources and use.

Energy resources

Understanding how the actions of humans affect the Earth’s landscape has long driven the discipline of geography. The demand for ever-expanding sources of energy has been one of the most studied causes of landscape transformation. The ensuing economic, political, and societal consequences of this transformation have led to the sobering realization that humans have grown dependent on energy in general, and fossil fuels in particular. This dependence has created a growing concern about the security of energy supplies.

The International Energy Agency (IEA) defines energy security as the “adequate, affordable, and reliable supplies of energy.” Today, countries from Japan to the European Union and South Africa are grappling with the fact that they are highly dependent on foreign sources of petroleum and natural gas. The economic burden that this reliance imposes on these countries hinders their efforts to become players in the global economy.

Today, the 15 nations with the largest energy reserves are geographically dispersed, located on six of the world’s seven continents (Figure 1). The United States, Russia, China, and Australia have the largest coal reserves, spanning four continents. The greatest oil reserves are found in the Americas and the Middle East, with the largest reserves in Venezuela and Saudi Arabia. These cover politically volatile regions where nationalized oil companies operate: the 12 largest oil companies are all state-owned and they control roughly 80% of petroleum reserves. At the top of
**ENERGY RESOURCES AND USE**

![Diagram showing global distribution of energy resources](image)

**Figure 1** The global distribution of energy resources: the 15 nations with the largest fossil energy reserves (Reproduced from US Energy Information Administration (EIA)).

This list are the Saudi Arabian Oil Company and the National Iranian Oil Company. Russia, Iran, Qatar, and Saudi Arabia host the largest natural gas reserves.

Interestingly, countries that are well endowed with one or two energy resources still rely on imports of other energy sources. For example, Saudi Arabia is the largest exporter of crude oil but has to import refined gasoline. Russia exports natural gas but has to import uranium. The United States is a net exporter of coal but imports oil. This interdependence explains why the challenge of energy security must consider the interactions between countries as much as it considers the resources of individual nations.

At the same time, the global energy map is changing, with potentially far-reaching consequences for energy markets and trade. It is being redrawn by the global spread of unconventional gas production, and by the continued rapid growth of renewables – especially wind and solar technologies.

In the United States, significant energy reserves can be found from Alaska to Pennsylvania (Figure 2). Coal is the largest American energy asset and is widely available in multiple states, but the largest coal reserves reside in Montana. In fact, if Montana were a country, it would rank in the top 10 based on coal reserves. Texas is home to the largest reserves of both oil and natural gas.

While the geography of shale gas outside of the United States is not well reported by the energy statistical agencies, shale gas resources in the United States are increasingly well documented. Shale gas production is progressing at a dizzying rate across North America, expanding almost twentyfold from 2005 to 2013, making the United States a net exporter of natural gas in 2009 (Logan et al. 2012). The US supply of shale gas is anticipated to double again by 2015 and to treble by 2030, as new investment flows into the sector (Deutch 2011; Zou 2012). Globally, the International Energy Agency (IEA) expects production from unconventional gas resources to double from 2010 to 2035 (IEA 2012). British Petroleum expects global shale gas production to grow sixfold from 2011 to 2030 (Hughes 2013). The result may be “the greatest shift in energy-reserve estimates in the last half century” (Deutch 2011).

In the United States, Texas is endowed with significant unconventional gas reserves, as is Pennsylvania. Low-cost natural gas is reducing coal use in the United States, freeing up coal for export to Europe (where, in turn, it has displaced higher-priced gas). The development of unconventional gas resources may also transform energy markets worldwide, especially for the developing world. Trillions of dollars in resources lie buried in the backyards of many of the world’s poorest citizens; if these were to be effectively and accountably managed, the economic development of these countries could surge.

Still, the extraction of energy resources such as oil, gas, coal, and uranium will inevitably decline despite improvements in efficiency and
advancements in technology. As a result, resource depletion is a fundamental issue that societies must consider. Are there meaningful peaks? Will depletion occur soon enough to be of concern? And does it in fact matter?

Affordable energy is another energy security concern. Many countries deal with this issue by instating energy consumption subsidies. The IEA estimates that fossil fuel consumption subsidies worldwide totaled $548 billion in 2013, with Iran, Saudi Arabia, India, Russia, and Venezuela topping the list (IEA 2014). The International Monetary Fund (IMF) estimated that in 2013 energy subsidies across 176 countries amounted to $1.9 trillion. On a post-tax basis, this is equivalent to 8% of all global government revenue for that year (IMF 2013). In most non-OECD (Organisation for Economic Co-operation and Development) countries, at least one fuel or form of energy continues to be subsidized, most often through price controls that hold the retail or wholesale price below the true market level. According to the Economist (2015), such handouts have caused extra consumption, which has been responsible for 36% of global carbon emissions between 1980 and 2010. Widespread fuel poverty has been a key driver of energy cost subsidies. Most of these countries have announced their intention to phase out consumption subsidies eventually, though at varying rates (IEA 2012). In the meantime, energy consumption subsidies are strong deterrents to energy efficiency by making investments in energy efficiency unattractive.

Electricity generation

Coal, natural gas, nuclear energy, and renewables are the principal energy sources used to generate electricity worldwide. The relative reliance of each energy source varies geographically, as shown in Figure 3. Coal is overwhelmingly the main source of electricity in China, and also the largest source in the European Union and the United States. While natural gas is the second largest provider of electricity generation in the United States, it ranks only third in the European Union, and provides just a fraction of China’s energy budget. Today, nuclear energy is the second largest source of energy for electricity in the European Union and the third largest in the United States, but accounts for just a fraction of electricity generation in China. The nuclear accidents at Three Mile Island, Chernobyl, and Fukushima have drawn public attention to the human costs associated with nuclear power. Geographic discussions have centered on the siting of nuclear plants, the dispersal of downwind radioactivity, and the proliferation of weapons-grade nuclear fuels on a global scale (Blowers, Lowry, and Solomon 1991).

The use of renewable energy – such as wind, solar, geothermal, hydroelectric dams, and
bioenergy – is growing rapidly in the electricity sector. Still, as can be seen in Figure 3, it remains a small player in the electricity markets of many of the world’s largest economies. Renewable energy is the third, fourth, and fifth largest source of electricity generation in China, the European Union, and the United States, respectively. Across all three regions, wind is the largest nonhydro renewable source of electricity generation, and the fastest-growing renewable energy resource on the planet. Wind turbines are being deployed across many countries, and are multiplying particularly in China, Denmark, and the United States. Unlike most energy resources that are hidden from the eyes of the average person, wind turbines are visible and serve as a reminder that all energy originates from a source. Wind turbines emit no greenhouse gases, are easily installed and decommissioned, and have prices comparable with conventional energy sources. The visibility of wind power is also a double-edged sword since the presence of a wind turbine alters the appearance of the landscape.

The national electricity demand of the United States could be fulfilled through about 3% of the land in the contiguous 48 states being used for wind energy. Unlike a coal plant, the land can also have other uses since only about 5% of land is used for access roads, electrical equipment, and wind turbines with wind energy (Pasqualetti 2000; 2011; Warren et al. 2005).

One of the most prominent hurdles to the diffusion of renewable electricity that exists today is the required backup. In addition, renewables are impeded by social and environmental obstacles, as well as costs.

How energy is used

With the rapid growth of energy consumption over the past half century, it has become increasingly clear that both the problems and solutions of energy demand growth are intertwined with geography. In addition, consideration must be given to the transportation of fuel, the generation of electricity, and the resulting impacts on the environment. In China and India, people are suffering from the pollution created by their coal plants and increasing numbers of motor vehicles. Air quality regulations in these countries are less stringent than in the United States and the European Union. Further, greenhouse gas emissions are being capped in the European Union and the United States, but China and India have not yet signed up to defined CO₂ emission reduction targets. Access to electricity and commercial energy fuels is also an issue in many countries, particularly those with significant energy poverty: at least 1 billion people worldwide are off the grid.

China and the United States are the world’s two largest energy consumers, together accounting for 38% of the world’s energy consumption in 2010. Germany and the United Kingdom are
two of the largest energy consumers within the European Union and are responsible for 31% of the its total energy consumption.

The focus will be on these four countries as well as the European Union in discussing energy consumption patterns and energy policies. These countries have been selected because they are important players in the energy arena: the United States, United Kingdom, Germany, and China account for nearly 40% of the energy used worldwide in residential and commercial buildings. Additionally, there are significant distinctions between these four countries and the European Union in terms of their economic composition, energy consumption patterns, policies, and political systems.

Table 1 summarizes the per capita GDP and energy intensity of commercial buildings in these four countries and the European Union. Of these countries, the United States has the highest per capita GDP and the second-highest energy intensity in commercial buildings. Germany and the United Kingdom are also affluent, but Germany is more efficient in using energy for commercial buildings: they are both wealthier than the EU average. Although China is another leading world economic power, the average Chinese citizen is significantly less wealthy than the average American, reflecting China’s developing country status. The energy intensity in Chinese commercial buildings is the lowest among this selected group by a large margin. Other research suggests that China’s intensity is driven by a low penetration of space cooling technologies and, to a lesser extent, by unmetered district heat. Incidentally, commercial buildings in China provide less thermal comfort than comparable buildings in the West. A close examination of the building energy codes in these areas illustrates the path dependency of regulations and policy innovation in different economic contexts.

In all three regions, industry is a large end-use energy consuming sector. In China, the majority of energy is used in the industry sector. Energy use is more uniformly distributed across end-use sectors in the European Union, with the four sectors each receiving at least 20% (services), and no sector receiving more than 29% (industry) (see Figure 4). The United States experiences a similar distribution, with the sectors ranging from 19% (services) to 32% (industry).

The value of energy products tends to be a large share of domestic economies across the world. In the United States, for example, energy expenditures were about 9% of GDP in 2012, or about $4000 per capita in 2013 (EIA 2015f). Almost half of these expenditures are on transportation fuels, which are about 95% petroleum-based (Brown and Kim 2015).

Energy demand has increased across the planet since the late 1970s. Geography was the key to coping with the skyrocketing demand for energy. The location of energy sources had implications not only for the surrounding environment but also for transporting the energy around the world (Pasqualetti and Brown 2014). Looking to the future, the IEA forecasts a 40% increase in worldwide energy demand, driven largely by China, India, and countries in the Middle East. According to the International Energy Outlook 2014 (EIA 2014), Brazil, China, India, and Russia will be responsible for nearly 70% of the growth in non-OECD energy usage over the next several decades. These shifts in energy consumption will also result in an evolving trading pattern that will financially benefit energy exporters (Bradshaw 2010; Pasqualetti and Brown 2014).

As a social science, geography is instrumental in understanding interactions and their effects on economic and political stability. It is no coincidence that throughout history the most bustling centers for economic trade have been
ENERGY RESOURCES AND USE

Table 1  Energy consumption and intensity in buildings, 2013.

<table>
<thead>
<tr>
<th>Country</th>
<th>2012 GDP per capita* ($)</th>
<th>Energy intensity in commercial buildings (MJ/m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>51734</td>
<td>1203</td>
</tr>
<tr>
<td>European Union</td>
<td>32787</td>
<td>–</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>41886</td>
<td>1342</td>
</tr>
<tr>
<td>Germany</td>
<td>39065</td>
<td>896</td>
</tr>
<tr>
<td>China</td>
<td>6093</td>
<td>400</td>
</tr>
</tbody>
</table>

*Adjusted for purchasing power parity (PPP).

Source: Sun et al. 2015. Reproduced by permission of John Wiley & Sons, Ltd.

...the cities with the greatest energy resources. This relationship can be seen from the Industrial Revolution in London, with its access to coal, to today’s modern cities such as Abu Dhabi and Baku, which are surrounded by oil fields.

The geography of energy policy over the years illustrates the race-to-the-bottom principle, whereby localities, states, and nations strive to avoid increasing environmental regulation because of the anticipated loss of potential economic gains. Locations with less stringent environmental regulations and policies become pollution havens, encouraging businesses to provide economic benefits to their region at the cost of the health of residents. Today, a race to the top is emerging in some affluent countries where a premium is being placed on environmental protection and the value of maintaining ecological services (Pasqualetti and Brown 2014).

A review of the building energy codes has identified many common features. For example, they all include an efficiency requirement for building envelopes, space heating, and ventilation. However, depending on the different climate situations, space cooling is regulated in countries like the United States and China, which have large areas with warm climates, while the United Kingdom and Germany do not. The political system also affects building energy regulations. Because the US Constitution prohibits the federal government from passing...
laws to regulate building energy efficiency, those responsibilities are shifted to state and local officials. The relationship between building energy codes and climate change also varies from country to country. In the United States and China, the adoption and implementation of building codes has been spurred by the concern for rising energy consumption. Although building codes offer CO2 reduction as a co-benefit, it has not been the primary driver. In contrast, European countries have used building codes to tackle both energy consumption and climate change mitigation issues, in part because of their obligations to reduce their greenhouse gas emissions under the Kyoto Protocol. As a result, the goals of building codes are more aggressive in Germany and the United Kingdom: the aim is to make all new buildings built by 2020 consume nearly zero energy. In order to achieve that goal, European building codes not only have efficiency requirements, but also explicitly encourage the use of renewable energy. The United Kingdom goes even further by moving from regulating energy use to regulating building CO2 emissions. Table 2 compares the key features of building energy codes across countries.

The most frequent and successful policies to promote renewable energy employed by governments across the globe are feed-in tariffs (FITs). By the end of 2013, 98 countries, states, or provinces had FITs, increased from 34 in 2004. The strengths of FITs include guaranteed grid access over a prolonged period (typically 15–20 years), with prices based on the cost of generation plus a reasonable rate of return. FIT payment is typically administered by the utility and derives from an additional charge for electricity imposed on national or regional customers. The German Electricity Feed-In Law, for instance, adopted in 1991, was instrumental to the advancement of renewable energy in Germany during the 1990s. It required public utilities to purchase renewably generated power from biomass, hydro, solar, wind, and landfill gas sources. Nine years later, the Renewable Energy Act was adopted, promoting renewable energy mainly through stipulating feed-in tariffs under which companies producing renewable energy are guaranteed that they can sell their “product” at fixed prices for 20 years. A surge in the production of clean energy has followed. This act also helped cover over 10% of Germany’s total energy consumption from sources like wind, solar, water, or biomass, and made it the world leader in photovoltaics. As a result, since 1990, the installed power capacity from renewable electricity in Germany has increased many times over (Frank 2014), reaching about 27% in 2014, nearly three times the share in the United States. Fixed prices are set for purchases of renewable

Table 2  International comparison of building code policies.

<table>
<thead>
<tr>
<th>United States</th>
<th>Germany</th>
<th>United Kingdom</th>
<th>China</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad technology coverage</td>
<td>Excludes space cooling</td>
<td>Excludes space cooling</td>
<td>Excludes lighting and water heating</td>
</tr>
<tr>
<td>Regulate energy efficiency</td>
<td>Regulate energy efficiency</td>
<td>Regulate CO2 emissions</td>
<td>Regulate energy efficiency</td>
</tr>
<tr>
<td>Implemented by states</td>
<td>Includes renewable energy and certification program</td>
<td>Includes renewable energy and certification program</td>
<td>Does not include renewable energy or certification program</td>
</tr>
</tbody>
</table>

Source: Sun et al. 2015. Reproduced by permission of John Wiley & Sons, Ltd.
electricity at a rate above the retail market price for each unit of electricity fed into the grid in Germany’s FIT. Additionally, German FIT requires that power companies purchase all electricity from eligible producers in their service area at this premium rate over an extended time period. These tariffs are guaranteed for 15–20 years, and differentiated by type of renewable electricity, location, and size. Germany’s FIT has created jobs and built the country’s solar and wind industries.

Figure 5 compares the electricity consumption, prices, and bills of households in nine industrialized nations. The North American nations of the United States and Canada have the greatest household electricity consumption and the lowest electricity prices. However, household electricity bills in the United States are the second highest of the nine nations. Further inspection reveals the highest prices and smallest consumption rates are found in the European nations of Denmark, Germany,
and Italy. Despite these differences, the resulting electricity bills of the North American and European nations are comparable, since bills are a function of consumption and prices. The high levels of energy efficiency in Italy, Germany, and Denmark enable bills to remain affordable even with relatively high prices. Thus, while the policy-driven surge in renewables has escalated electricity prices in many European countries, their strong building codes and other energy-efficiency policies have curbed consumption, resulting in affordable electricity bills.

Energy transition at the national scale intertwines consumers, policymakers, and energy providers; to be successful, it requires a persistent and predictable policy commitment. The term “energy transition,” or *Energiewende*, originated in 1980 in Germany alongside calls to end the use of nuclear and petroleum energy. Germany’s energy transition toward cleaner renewable energy benefited from a decentralized energy market and focused primarily on transforming the supply side of the energy equation. In Germany the increasing focus is now turning to energy efficiency and the demand side. While energy transition has been a success in Germany, neighboring Netherlands has not experienced the same level of success. The Dutch transition to alternative energy sources has failed to attain widespread use in the Netherlands in large part because of frequent shifts in energy policy and delays in implementation (Geels 2014; Verbong and Geels 2007).

Understanding energy transitions through the lens of polycentrism has been gaining traction; it is an approach to energy and climate policy that taps into the strengths of geographic analysis. The existence of multiple authorities with overlapping jurisdictions operating at multiple scales of action – engaging multiple stakeholders – can provide a powerful policy environment for enabling clean energy transitions (Brown and Sovacool 2011). This multilevel perspective may explain why, despite evidence indicating the need to increase the use of clean renewable energy, coal’s contribution to electricity generation grew from 2000 to 2010. Following the 2015 Paris Climate Conference, countries across the world are beginning to decarbonize their energy systems by implementing an array of new policies. Explaining the diffusion of such policy systems across time and space (Matisoff and Edwards 2014) is a critical need which can capitalize on the geography profession’s strong analytic and conceptual tools.

References


EIA (US Energy Information Administration). 2015a. “Crude Oil Proved Reserves, Reserves Changes,


Access to modern energy services is essential to all aspects of human wellbeing; therefore, guaranteeing secure and affordable supplies of energy is a key concern of the state. Energy security is best conceived as a property of an energy system. Thus, a national energy system can be more or less secure and there are a multitude of factors that determine how resilient a particular energy system is to shocks that might threaten its integrity and thus the security of energy services to industries, households, and individuals. This “whole systems approach” highlights the complexity of energy security as a system property in an increasingly globalized and interconnected global energy system (Mitchell, Watson, and Whiting 2013).

The contemporary concern about energy security is not new. Daniel Yergin (2006, 69) identified Winston Churchill’s decision of more than a century ago to shift the British Navy from coal (which Britain had in abundance) to oil (which made Britain dependent on oil imports from what was then Persia) as marking the origins of energy security as a strategic concern for the state. For Churchill, “Safety and certainty in oil,” he said, “lie in variety and variety alone.” Today we still see physical security of supply as the core challenge to energy security and – in the absence of self-sufficiency – diversity or variety is still seen as the solution to import dependence. However, in the contemporary context energy security is a much more complex notion than just physical security of supply.

Since Churchill’s decision, access to oil has become a strategic and economic necessity and a major cause of conflict. This empowerment of petroleum has also bestowed geopolitical influence on the major oil-exporting states (LeBillion and Bridge 2013). The geopolitical use of the “oil weapon” in the mid-1970s by the Arab members of the Organization of Petroleum Exporting Countries (OPEC) resulted in the creation of the International Energy Agency (IEA) to coordinate the interests of the Western industrially developed economies (members of the Organization for Economic Development and Cooperation, or OECD). The initial purpose of the IEA was to coordinate Western energy policy and manage the Strategic Petroleum Reserve, with each member required to hold oil stocks equivalent to at least 90 days of net imports and to maintain emergency measures for responding collectively to disruptions in oil supply of a magnitude likely to cause economic harm to its members. The actions of OPEC proved counterproductive as high oil prices triggered economic recession and promoted energy conservation measures that reduced demand for oil. It also encouraged the OECD states to develop their own oil and gas reserves in Alaska, the Gulf of Mexico and the North Sea. After a second oil shock in 1979–1980 caused by a revolution in Iran, the 1980s and 1990s were a period of relative calm with more than enough oil to satisfy demand. In early 1999, the front cover of The Economist Magazine carried the headline that the world was “Drowning in Oil” and pointed out that in real terms oil was cheaper than in 1973. They speculated that the oil price
ENERGY SECURITY

might fall as low as $5 a barrel. They proved to be wrong, and less than a decade later the oil price hit a record level of $147 a barrel.

In the first decade of the new century, the geographical locus of growth in energy demand shifted eastward with the rise of China. The decade saw China become the world’s top oil-importing country. China’s rise was part of a wider wave of emerging economies that fueled sustained growth in demand for oil and gas. The consequence of the previous decades of plenty and the associated low price was a lack of investment in new production, and the global oil industry struggled to keep pace with demand. The situation was compounded by continuing instability in the Middle East: the loss of production in Iraq due to conflict, in Iran due to sanctions, and in Libya due to the Arab spring and civil war. The resultant period of high oil prices stimulated the development of unconventional oil and gas in North America, as well as substantial new investment in more expensive reserves, often described by the industry as “the end of easy oil.” The advocates of “peak oil” have argued that the global oil industry is close to its maximum rate of production and that there will soon be a significant reduction in production. However, the optimism surrounding unconventional oil and gas and the prospects of increased production from Iraq and Iran have quieted such concerns, for the moment at least. At the same time, the 2008 global financial crisis dampened energy demand and it has yet to recover. Thus, a combination of confidence about future production and the slowing rate of demand growth have relaxed concerns about peak oil, yet the oil price remains historically high, with analysts talking about a price of around $100 a barrel being the “new normal.” History shows that oil prices are volatile and in the summer of 2014 oil prices fell dramatically as supply outstripped demand and Saudi Arabia decided not to cut back on production, seeking to test the resilience of US tight oil production. For oil-exporting states, a prolonged period of low prices presents a huge political challenge as many of the world’s more autocratic regimes are supported by the rent derived from oil exports.

Discussion so far has focused on the traditional concerns related to the geopolitics of oil, but there is an additional challenge that complicates the contemporary definition of energy security, namely climate change. Benjamin Sovacool (2011, 3) has reviewed the policy and scholarly literature and identified 45 different definitions of energy security. He distilled those definitions into four dimensions: availability, reliability, affordability, and sustainability. Thus, energy security is about physical availability, reliability or continuity of supply, at prices that are affordable, and in forms that are environmentally sustainable. The latter links energy security to the need to reduce the carbon emissions associated with the global energy system, which is the single largest source of greenhouse gas emissions. The tensions between the different dimensions are also clear to see. Energy security comes at a cost and a tradeoff must be made between the level of security desired and the ability of the consumer to pay the price. Equally, low-carbon energy sources, at present at least, are more expensive than the incumbent high-carbon systems. Thus many in the energy industry talk of a “resource trilemma,” the tension between energy security, affordability, and decarbonization. But the “low-carbon” transition brings with it new energy security concerns; for example, in relation to the security of supply of the materials, rare Earth minerals and elements such as lithium that are essential to these new technologies. The intermittent nature of most renewables is also a new source of insecurity: backup is required when the sun is not shining, when the wind does not blow, and
when there is insufficient water to power the turbines or cool the reactors. Thus, the current “energy transition” retains all of the traditional energy security concerns associated with the hydrocarbon economy that will continue to account for the vast majority of energy production for the foreseeable future, and introduces a new set of energy security concerns that are related to the introduction of low-carbon energy sources. This demands a more integrated conception of energy security that recognizes the complex interrelationship between energy security and climate change. Some have called this the “new energy paradigm.”

The energy security literature has a clear supply-side bias reflecting the needs of the industrialized energy importing states, but the importance of “security of demand” for energy-exporting states is also an increasingly important consideration. If energy-exporting states are to invest in developing new fields and building new infrastructure they need to be confident that there will be demand for their production at a price that they consider to be fair. “Transit security,” or security of transportation, raises a third area of concern, with international trade in oil and gas being vulnerable to disruption at key “choke points,” such as the Straits of Hormuz and the Straits of Malacca, or to political disruption, as in the case of the Russia-Ukraine gas disputes in 2006 and 2009.

A final dimension that is now center stage on the policy and research agenda is that of energy access. The United Nations has declared 2014–2024 the decade of Sustainable Energy for All in recognition of the fact that there are still 1.4 billion people on the planet that do not have access to modern energy services. Although energy access was not a Millennium Development Goal, it is now included as Sustainable Development Goal 7 to: “Ensure access to affordable, reliable, sustainable energy for all.” This places the energy security concerns of low-energy societies in global context and poses a global energy dilemma: “Can we have secure, affordable and equitable supplies of energy that are also environmentally benign?” (Bradshaw 2014, 21). Addressing this challenge requires a much broader conception of energy security based on an understanding of the socioeconomic processes that drive demand for energy services.

SEE ALSO: Climate change policy; Globalization

References

Enlightenment geographies

Charles W.J. Withers
University of Edinburgh, UK

The Enlightenment, usually taken to be an intellectual phenomenon and a historical period, broadly equivalent to the “long” eighteenth century in Europe, should also be understood as a matter of geography with global reach and significance. The term “Enlightenment geographies” embraces three related concerns: the importance of where questions, and matters of scale, space, and place, in explaining what the Enlightenment was; the explanation of philosophical inquiries concerning human difference, cosmopolitanism, sociability, and comparative cultural development with reference to geographical factors, such as climate; and the content and purpose of geography as a subject and as a set of practices in the Enlightenment. Thinking geographically about the Enlightenment in these ways has challenged essentialist conceptions of the Enlightenment, and nation-based interpretations of the Enlightenment as only a European enterprise. Explanation of the Enlightenment’s spatial dimensions extends from the broader “spatial turn” in the humanities and social sciences since the 1990s, including evident recognition of the importance of geography in the history of science in particular, and in studies of colonialism and of imperial history in general.

For most contemporaries then and for modern scholars now, the Enlightenment was both a historical moment – by convention, the “long” eighteenth century in Europe (ca. 1680 to ca. 1815) – and an intellectual movement characterized by challenges to established authority, advances in natural philosophical (“scientific”) method, and new ways of thinking about the world that were distinguished by belief in the power of critical reason harnessed for human moral and economic improvement (Livingstone and Withers 1999). Considerable attention has been paid to the Enlightenment’s biographical, historical, and intellectual dimensions: to its who, its when, its what, and to its lasting legacy, the Enlightenment’s so what? Since the mid-1980s or so, a key distinguishing feature of Enlightenment studies has been recognition of the fact that the Enlightenment and its ideas, motivating personnel, and material artifacts did not “float free,” somehow “above and apart from” their spatial setting and social context. Locatable as well as datable, the Enlightenment importantly had a where (Withers 2007). This view has its parallels in work in Enlightenment studies that has sought to de-essentialize the Enlightenment as a single unitary intellectual phenomenon (Mayhew 2010; O’Brien 2010). It is also a reflection and consequence of that broader “spatial turn” in the humanities and the social sciences since the 1990s: as Brewer put it in 2004, “A glance out across fields of knowledge quickly reveals that ‘space’ became the master metaphor of late twentieth-century epistemology” (Brewer 2004, 171). Consideration of the Enlightenment’s spatial dimensions and of the content and purpose of geography in the Enlightenment has challenged established orthodoxies and proposed new ways of understanding the period and its intellectual concerns. Three main and connected themes may be distinguished.
Enlightenment geographies as matters of scale

The Enlightenment in national context

The nation has been the commonly used frame of reference in examining the Enlightenment’s geographies. The idea of the Enlightenment understood in national terms has two main variants. The first is signaled by “the Enlightenment in France,” “the Enlightenment in America,” and so on. The second is evident in “the Italian Enlightenment,” “the Scottish Enlightenment,” and the like. In the first, the Enlightenment is held to have both defining characteristics and variant national expression by virtue of the content and material form of philosophical or scientific ideas and their presumed unproblematic movement from one nation to another. In this view, the Enlightenment travels over time and across and between national space and, often, does so from an originating “core,” such as France or the Low Countries, to a “periphery” elsewhere. The second variant is suggestive of more distinctive Enlightenment characteristics resulting from given national circumstances. Yet, in this perspective, differences within the nation are obscured, and shared experiences above the nation may be neglected. For these and other reasons, these historiographical perspectives have been criticized and revised in recent years (Kors et al. 2004; Withers and Mayhew 2011).

What matters is not that there could have been a shared and single Enlightenment apparent at the national level: it is just that scholars should not take that interpretation, and the question of national scale, for granted. Persistent and uncritical use of the definite article has encouraged the presumption that there was such a thing as the Enlightenment, and that how we understand its geographical dimensions and implications is, simply, to trace its origins, expression, and dissemination in a national context. Omitting the definite article — and thinking of diverse enlightenments rather than any single essential Enlightenment — is to caution about the importance of differences within and above the nation. The Enlightenment understood as a single intellectual movement is undermined by differences between nations and the diversity within nations. The idea of the nation was anyway relatively new in the eighteenth century. In France from the 1680s, for example, map-makers and military surveyors were centrally involved in knowing the extent of their nation: in the Enlightenment, the nation of France was constantly being put to shape geographically. Further, to understand the Enlightenment’s geographies as only a national matter neglects vital social distinctions within the nation. In Portugal, for example, Enlightenment was characterized more by foreign intellectuals working within that nation than it was by nationals’ expertise.

Because Enlightenment ideas and ideals did not easily respect national boundaries, and rather than think of the Enlightenment as having only national expression, we should also approach its geographies in terms of transnational connections.

Cosmopolitan networks: the Enlightenment as a Republic of Letters

In eighteenth-century Europe and beyond, what sustained critical commentators in their inquiries over the power of reason and the improvement of the human condition was the spoken and the written word. Writing, reading, recording, and reporting bound people together in networks of communication and exchange. Ideas about natural diversity and human difference were shaped by the receipt of new data – plant species,
Cultural artifacts, and, on occasion, living human specimens.

Considered above and beyond the nation, the Enlightenment was a Republic of Letters (and of artifacts), defined by correspondence, a cosmopolitan intellectual outlook, practices of material and intellectual exchange, and networks of institutional communication. Such transnational geographies are apparent, for instance, in the makeup of the personnel involved in the expeditions that undertook to measure the transits of Venus in the 1760s. They are apparent in the difference between the publication of influential works – Adam Smith’s *Theory of Moral Sentiments* (1759), for example – and their reception and, often, later translation, and the mediated engagement with their cognitive content in different intellectual and institutional settings (Butterwick, Davies, and Sánchez Espinosa 2008). Thinking about the Enlightenment’s geographies in terms of cosmopolitan networks is well illustrated by the case of the Enlightenment in Latin America, or, more properly, in Portuguese South America and in Spanish South America. Enlightenment in Spanish South America centered not on European theories of reason or in the mobility of European texts but, rather, on the interpretations that could be placed on Aztec monumental epigraphy, on the dynamics of local Mexican scholarship, and on the efforts of the local intelligentsia in resisting attempts to transport the monuments and the scholarship back to Europe. This work has been supplemented by studies in the history and geography of Latin American Enlightenment science that point to the making of science as a form of local public good and to the importance of natural history expeditions and indigenous medical knowledge. The result is that the Latin American Enlightenment has been both “denationalized” and “creolized,” and Enlightenment in Spain and in Iberia has been rethought (Canizares-Esguerra 2001).

Yet, the Republic of Letters was not, as it were, a free country with open borders. Correspondence and material exchange took place principally between people deemed credible authorities by virtue of their social status. Those accumulative networks of correspondence and exchange that sustained the Republic of Letters centered on the intellectual interests of metropolitan Western Europe more often than they did on the peoples of the Pacific, or of America, or even of northern Europe. The facts of social and geographical variation do not weaken the idea of the Enlightenment as a dynamic and cosmopolitan phenomenon. They remind us of the need to pay attention to the global but unequal and asymmetric reach of Enlightenment ideas, and to the strength, direction, and cognitive content of the networks sustaining the Enlightenment at scales above and beyond the nation. This being so, the ways in which the Enlightenment’s cosmopolitan and national geographies were grounded in specific local settings can also be examined.

Local sites and active settings

The Enlightenment was always an actively local affair. Consideration of the Enlightenment’s geographies at a local scale is wholly consistent with its study at national and transnational levels and is in keeping with that more general de-essentialism that is a characteristic of recent Enlightenment historiography (Mayhew 2010; Withers and Mayhew 2011). Working with the Enlightenment at scales above and below the national reveals a series of situated local settings and practices – polite conversation in salons, instructive rhetoric in lecture halls, instrumental practice on board ships and in chemical laboratories, contemplative private
reading in domestic drawing rooms, taxonomic ordering in herbaria, and so on. Looking at the Enlightenment’s geographies in terms of one setting or as a series of local sites, most commonly in an urban context (see Elliott 2010 on Enlightenment England), discloses articulations of practice and performance, and often the presence of competing ideas that never attained widespread currency for one reason or another. Certain venues, such as learned academies and societies, and to an extent the universities, were key sites for the presentation and criticism of new knowledge. An overemphasis on local settings and the dynamics of individual places alone runs the risk of diminishing the shared dimensions of intellectual culture and popular practice and of failing to examine how Enlightenment ideas moved over space.

In sum, the Enlightenment operated, and may be studied, at different geographical and social scales. Understanding these scales as important, distinct, and yet related is not to accord a primacy of one over another. What matters is the relationships between scales: national pictures always had variant local expressions, cosmopolitan networks were always “earthed” in specific social settings and epistemological concerns; the practices of (say) Enlightenment botany in one place may not have been the same in another. Recognition of these different scales neither privileges one over another, nor sees any one topic as intrinsically more important than another. Rather, what is emphasized is that the Enlightenment should be understood as about processes of intellectual and material production and exchange in places and over space, not just as a period in time or as a philosophical movement. The result overall of such geographical thinking is to displace the idea of the Enlightenment as an essential thing by views of it as multiple and variegated, with, always, meaningful geographical qualifiers.

Enlightenment geographies as matters of practice

Exploring, mapping, and enlarging the globe

Geographical exploration in the Enlightenment aimed to bring the globe under the sovereignty of science, accumulating new knowledge through empirical inquiry and speculative theorization. In the Pacific, exploration revealed what was talked of as a new “Fifth division” to the world, Europe, Africa, and Asia having provided the first three and the Americas the fourth (Withers 2007, 89). Oceanic navigation gave the world a new and more certain shape, at least in terms of its continental margins. But Enlightenment voyages of discovery did not reveal or represent this geographically more diverse world easily, equally, or at once. Several explorers died undertaking such work. Others were imprisoned, yet others being either underfunded or overwhelmed by data. Continental interiors remained largely unknown. China was described in travelers’ accounts and depicted on the maps of Jesuit missionaries, but much was based more on hearsay rather than hard fact. The Arctic and Antarctic regions were not explored until much later.

Geographical discovery revealed a world of different peoples and diverse cultures, and an unimagined natural plenitude. The result was not always precision and certainty – although these were intentions commonly promised in the wording of explorers’ narratives – so much as a sense of wonder, even of astonishment. Contemporaries had to choose their words and actions with care: cultural contact commonly began with moments of linguistic and gestural incommensurability. Map-makers had to rethink their works. Guides to travel and to scientific procedure looked to establish methodological
rigor and descriptive standards in order to effect meaningful comparative analysis. The Enlighten-
ment even brought new types of map-makers into being. Most European nations had a mix of navigator-map-makers, geographers, commercial publishers, and land surveyors. The Enlightenment was distinguished by the rise to prominence of the professional military or state surveyor. France took the lead among the nations of Europe in being systematically surveyed in the Enlightenment under the direction of the Cassinis. In Bengal, military surveyors such as James Rennell produced atlases and maps that promoted and reflected the commercial expansion of Britain’s empire. Enlightenment map-making emphasized empirical observation and quantifiable measurement in the service of the state.

For many of the peoples encountered through geography and represented in mapping, such cartographic delineation often silenced or marginalized them. Symbols of geographical engagement, Enlightenment maps are often also works of cultural effacement. So, too, many Enlightenment painted depictions of “foreign” landscapes are framed by European conventions – of sublimity, of the primitivism of native otherness, or of the boundless luxuriance of a nature later subject to the remorseless logic of resource extraction. In the Enlightenment, moves toward realism in representation were associated with the development of standards in measurement, observational training, systematic collection, and comparative explanation in science. Yet, geographical encounter with the faraway and with the exotic nearer at home – of the Lapps, for example, or the Scottish Highlanders – was most often undertaken on Europe’s terms. What commonly followed in the wake of explorers’ ships were unequal commercial and cultural relationships (Carey and Festa 2009).

Knowing the world: the rise of authoritative science

The new facts obtained by Enlightenment exploration challenged established views of the age and origins of the Earth and the authority of religious interpretations, as well as providing the foundation for modern scientific disciplines extending from that more general pursuit, “natural philosophy.” Two examples, geodesy and botanical classification, illustrate these claims.

In the early Enlightenment, the shape of the Earth was hotly contested. To adherents of Newton’s gravitational theory, the Earth was elliptical in shape, flattened at the poles. To proponents of Descartes’s ideas about vortices, the Earth was a different shape entirely – extended at the poles, flattened at the equator. The solution to these contrasting views over the geography of the whole Earth demanded empirical testing in those parts of the world that would permit proof, namely on the equator, and if not at the North Pole itself, then as far north as possible. Two expeditions were undertaken to this end in the 1730s. The first, in 1733, under the guidance of the Newtonian and French mathematician Pierre-Louis Moreau de Maupertuis, went to Lapland. The other, in 1737, under the direction of Charles-Marie de la Condamine, was based in the Viceroyalty of Peru. The results of the expeditions proved Newton and Maupertuis right. In such ways, geographical fieldwork, often from local sites but with global consequences, could test, corroborate, and refute book-based reasoning. Mathematics, and not the scriptures, became the language through which nature’s secrets were revealed and explained.

From the mid-1730s, botanists increasingly put the world of plants to taxonomic order. In the schema proposed by the Swedish naturalist Carl Linnaeus, the botanical world was classified on the basis of plants’ sexual characteristics, and involved a standard nomenclature of generic and
of specific names. Linnaeus also identified five global climate zones by which to explain the distribution of plant types, and he experimented with the acclimatization of species from one zone to another as part of his interests in economic and colonial development. For Linnaeus and many others, botanical inquiry with a utilitarian end in view was central to enlightenment as a process of human improvement. It is important to note, however, that his taxonomic ideas, formulated in Uppsala and based on fieldwork in Dalecarlia, were not accepted everywhere. Even as the empirical natural sciences became established, the map of Enlightenment scientific activity looked different in different places: the Enlightenment natural sciences were formulated within and reached across particular geographical circumstances (Clark, Golinski, and Schaffer 1999).

Encountering human difference

Enlightenment geographical discovery was greatly concerned with the “shock of the human new”: reports about human cultural variety flooded in – of pygmies, “cannibals,” Pacific natives, North American tribes, Australian aborigines – and challenged long-standing explanations on the basis of first-hand observation. New evidence about human variety helped establish what Enlightenment contemporaries termed the “science of man,” that is, they brought questions of science as observational and deductive method and of utility as outcome to bear on the moral and the human world, not just on the physical and the natural world.

Conceptual models of human development commonly placed “European man” at the head of sequential and progressive stages of social development and positioned other peoples and cultures beneath them, at different developmental stages. The Enlightenment was thus characterized by geographies of human difference in time, wherein ideas about different social systems were used to advance theories about relative social development, and by geographies of human difference over space, in that these different social systems and stages of human development were present simultaneously across the world.

Climate was often invoked in explanation of these different human geographies. An important figure in this regard was Montesquieu, whose 1748 *L’esprit des lois* (*The Spirit of the Laws*) is a key work on the effects of climate as a physical determinant of the moral bases to human behavior. What geographers and naturalists such as Linnaeus took to be the “temperate zone” Montesquieu saw as a space of moral and cultural equability. But to the north and, especially, in the warmer south, climate determined different degrees of cultural capacity vis-à-vis the “norm” of temperate Europe. Montesquieu was in part advancing an environmental explanation based on others’ theories. But he was also providing, broadly, an attempt to explain the nature of the human condition and its geographical variability. Others thought his emphasis on physical causes misplaced: for Voltaire and Hume, for example, social agency and not temperature or latitude was the basis to human difference. In such debates, Enlightenment commentators were developing and employing new analytical and rhetorical instruments by which to explain human difference. Even the terms they employed – “race,” “nation,” “variety,” and “species” – changed over time and were differently used in different Enlightenment settings.

In his initial classification in 1735 of the genus *Homo* and its single species, *sapiens*, Linnaeus incorporated four *varietate* (varieties): “Europaeus albus” (white European man), “Americanus rubens” (American red man), “Asiaticus fus-cus” (Asian yellow man), and “Africanus niger” (African black man). His classifications changed as new data were acquired. Linnaeus took the
world to be made up of different fixed varieties. Others saw more dynamic relationships within la variété humaine. At the same time, ideas of the progressive, stage-by-stage, development of human society from hunting, pastoralism, and agrarianism to commercialism, helped position the “West” in relation to the “rest.”

What people then understood as the “science of man” was in several senses a “geography of man.” Human geography was a matter of location and distribution (who was where?). Inquiries into human distribution as a matter of geography (why were those peoples there?) sought explanation in climate and latitude. For others, geography was not a fixed natural base against which human cultures were determined, but rather something that could be molded by human agency. For that reason numerous works of economic commentary and of philosophical speculation in the Enlightenment were concerned with improvement as the geography of the future—not with what was and has been, but with what could be.

Enlightenment geographies and the subject of geography

Geography’s print cultures

As Enlightenment geographies revealed the world to be bigger and more diverse, geography’s print cultures incorporated and disseminated that new information. Alongside mapping, these print cultures took, broadly, two forms: books of geography and geographical periodicals, and texts in which geographical reasoning featured as part of the production, mobility, and reception of the Enlightenment’s ideas.

English-language books of geography produced in Britain tended to position Britain as the leading nation in advancing geographical understanding, the French doing likewise for their geography texts. Geographical authors were seldom geographers by background or training, being usually compilers of others’ facts. Some authors were abreast of Enlightenment theorizations about human difference. William Guthrie’s A New Geographical, Historical and Commercial Grammar (1770) explicitly linked geography to theories on the stage-by-stage development of human societies (Mayhew 2000). In North America, Jedidiah Morse’s Geography Made Easy (1784) and American Geography, or A Present Situation of the United States of America (1789) were the first books in Enlightenment America to shape its national consciousness through geography. Before and after independence from Britain, the new geography of the North American continent represented challenges to established knowledge and to political administrators alike. Mapping made the new republic visible as a geopolitical space, albeit slowly. For geographical authors, new information was needed if the emergent nation was to be understood and governed properly. Morse’s works are part of a distinctive geographical print culture in early America. Morse reversed established publishing practice, putting America before—and, by implication, above—Britain and Europe. The fact that he regarded Connecticut (his home state) as the most virtuous in the new nation, and saw the Southern states as being prone to indolence through drink and the effects of climate, illustrates geography’s moral dimensions as a basis to national identity (Mayhew 2000; Withers 2007).

The Encyclopédie, ou Dictionnaire Raisonné, edited by Denis Diderot and Jean d’Alembert and published in Paris between 1751 and 1765, has been seen by later commentators as the Enlightenment’s manifesto. Its editors considered it to provide a new map of the intellectual world in which the connections between categories of knowledge were presented. Geography had
close links with mathematics and with history and was divided into several parts, natural geography and political geography being the main divisions. If the Encyclopédie’s production and reception are considered, rather than its content, the Enlightenment may be seen to be a geographical phenomenon embodied in print. The Encyclopédie was published in Geneva, Lucca, Leghorn, Neuchâtel, Berne, and Lausanne, in addition to Paris. Many of the authors of the Paris edition were members of local academic institutions. Whereas its authorship was largely Parisian, the Encyclopédie’s sales reveal a different geography. Sales were generally higher in provincial France than in the nation’s capital, and in towns with an administrative and cultural focus. Patterns of readership and of authorship disclose different enlightenments.

Medical space

In France in the Enlightenment, significant geographical work was undertaken by the ingénieurs-géographes, military engineers trained in trigonometry and in survey techniques in order to realize the shape and content of their nation. In France and elsewhere in Europe, topographic mapping was paralleled by work designed to fix the position of the prime meridian, either at Paris or at Greenwich. In several countries, medical map work aimed to link topography and human wellbeing, often by association with local variations in climate and soil type. The Bolognese geographer and naturalist Luigi Marsigli prepared a plague map of the eastern Adriatic coast in 1700 in which, among other features, he calculated and represented the time geographies associated with distance from the center of the outbreak. Large numbers of map-based medical topographies were produced in France in the 1770s and 1780s, works that were, in effect, syntheses of meteorological, climatic, and demographic data geared toward more effective political administration. In Germany, the physician Leonhard Ludwig Finke outlined in 1772 what might be thought of as a “proto medical geography” by drawing on his experiences as a regional medical officer. In New York in the 1790s, physicians used dot maps to aid their treatment of the disease and speculate on its causes. For these and other philosophers and for officials at work on behalf of the state, their reasoning was geographical, the end in view political. Geographical thinking, and geographical techniques such as mapping, were tools for the practical improvement of the human condition.

Sociability and sites of geographical education

The Enlightenment was characterized by considerable civic engagement with geography in schools, learned academies, and universities, and in public lectures. There were seven géographes du Roi listed as members of the Parisian Académie des Sciences between 1699 and 1793, and a further 17 persons there recorded “geography” as their principal disciplinary interest. But, as noted earlier, many botanists, mathematicians, and physicians either employed geographical methods in their work or saw improvements in a nation’s geography as a worthwhile end in view. Geography was widely taught in Europe’s universities in the Enlightenment, in association with civil history, mathematics, and moral and natural philosophy. Because instructors in different institutions placed different emphases on different parts of geography, local geographies of geographical teaching can be sited. In Königsberg, Immanuel Kant concentrated on physical geography. In Göttingen, under Tobias Mayer and others, mathematical geography was to the fore.
There were over 20 different public geography classes in Edinburgh between 1708 and 1800, a story echoed in London and overseas in Europe and North America. In school, in public classes, and in the home, geography was spoken about as well as read and drawn or committed to memory and was learned through play involving geographical games and “dissected maps” (jigsaws in modern terms), as well as via formal instruction. For women and children especially, the home was commonly a site of geographical learning and of sociability.

The continued importance of the Enlightenment’s geographies

The Enlightenment’s where is as important as its when, what, who, and why. In the Enlightenment, the Earth was revealed as an object of geography through the geographical practices of exploring, mapping, and reporting. Geographical knowledge was a key component of Enlightenment theories about human culture. Geography the subject was a species of Enlightenment knowledge, evident in texts, instruments, polite conversation, and cultures of formal instruction (Livingstone and Withers 1999). The too often privileged geographies of metropolitan Europe were connected with and dependent on other parts of the world. Particular places – the Americas, Oceanic islands – were key sites for the making of world knowledge, not margins to an Enlightenment “core.” These facts of geographical difference require us now to dismiss the idea of the Enlightenment as an essential feature, explicable only at the scale of the nation. Uncritical (and too often repeated) notions such as “the Enlightenment project” should be dispensed with.

The Enlightenment – and its geographical differences – is with us still. That is, the Enlightenment continues as an object of modern study because of its political implications, its ideals, and its languages, and not least because it is unfinished (Schmidt 1996; Carey and Festa 2009). The Enlightenment continues as a scholarly object because its unequal legacy continues still. Progress is always shadowed by its “other,” exclusion and persecution: Enlightenment always had a dark side. Scholars continue to ask what’s left of the Enlightenment, and to inquire about its longer-reaching consequences, the Enlightenment’s so what? Because we live with the Enlightenment’s consequences, it is important that we continue to address the Enlightenment’s geographies.

SEE ALSO: Cores and peripheries; Cosmopolitanism; Empire; Environment, nation, and “race”; Historical geography

References


ENLIGHTENMENT GEOGRAPHIES


Environment and celebrity

Dan Brockington
University of Sheffield, UK

There are many different varieties of environmentalism. Different environmental movements can be fundamentally opposed to each other and pit different social groups against each other. Perhaps the starkest examples occur when environments championed as wilderness, and suited best for luxury safaris for the international rich, are at the same time claimed as the dwelling places of much poorer rural residents of poor countries. Appreciating these varieties of environmentalism is essential to understanding the work of celebrity in environmental affairs, because we have to see not just how the support for environmental issues comes to be “celebrified,” but what sorts of environmental causes are espoused by celebrities, and what sorts are not.

As argued below, there are different kinds of celebrity at work in environmental causes, and that, for all of them, the nature of their involvement has changed recently as a more general professionalization of celebrity advocacy has taken place. The changes, for environmental causes, are bound up in the close relationship that mainstream environmentalism enjoys with contemporary capitalism. This also results in new forms of exclusion for other environmental causes. The entry ends with a conclusion on thoughts as to what further study of the work of celebrities and the environment can entail.

As described elsewhere (Brockington 2009), there are three types of celebrity at work in environmental affairs. First, there is the category of people who are already famous, and who lend their fame to conservation causes, with varying degrees of commitment. Some of the most prominent to do so might include Harrison Ford (an actor) who supports the work of Conservation International or Cate Blanchet (an actress) who supports a variety of environmental causes. In this category royalty have a particularly long tradition of supporting particular environmental and conservation causes. The Duke of Edinburgh (husband of the British queen) for example was involved in the establishment of the Worldwide Fund for Nature (WWF). His son, Prince Charles, and grandson, Prince William, have recently helped to set up United for Wildlife to campaign against poaching. The late Prince Bernard (of the Netherlands) was similarly active in his support of the WWF and conservation organizations in South Africa.

Second, there are the wildlife filmmakers and documentary makers who come into the public view because of the way that they speak for and represent nature. The most famous are Sir David Attenborough and the late Jacques Cousteau, others include the late Steve Irwin, Saba Douglas-Hamilton, or Charlotte Uhlenbroek. The films this group made were not necessarily focused on promoting conservation messages per se, at least not at the outset. Instead, this genre was used simply to capture the awe-inspiring beauty of nature and provided a source of entertainment in its own right. Conservation films, by contrast, were typically poorer in quality and had the unfortunate tendency to harangue tired viewers. More recently, the goals of nature filmmaking and conservation have drawn closer, however, as the filmmakers began to recognize the need to conserve scenic
ENVIRONMENT AND CELEBRITY

habitats and the charismatic species that formed the subjects of their documentaries. There is now more support for conservation within the sector as a whole.

The third group of celebrities working for environmental causes are those who have become famous because of their environmentalism. Some may only be famous within particular environmental movements, but others are now more widely known public figures. They include Henry Thoreau (a posthumously famous US environmentalist), John Muir (a champion of protected areas in the US), Gray Owl (a Briton who successfully masqueraded as an Ojibwa elder in Canada for decades), Sir Peter Scott (who co-founded the WWF), Dianne Fossey and Jane Goodall (both for their work with primates), and Jonathon Porritt and George Monbiot (both British environmentalists).

Note that fame is working in slightly different ways across these categories. This has been effectively brought out in the work of Olivier Driessens (2013). Drawing on Bourdieu’s work on different forms of capital and field theory, Driessens distinguishes between symbolic capital and celebrity capital. The former is recognized authority that is bound to specific social fields. The latter is recognizability and is not restricted to any particular field. This, Driessens observes, allows celebrities to migrate from one social field to the other, for example, from musical entertainment to rainforest protection (as has the musician Sting).

There are different sorts of such migration visible in the three categories of environmental celebrity offered above. The first group are using their celebrity capital in different domains – in environmental fields – from those where they first earned it. The second probably involve the least migration. From the public point of view these people make ideal spokesmen and women for nature. After all, the role of championing environments is precisely what their day job consists of; the authenticity appears to be natural. The third group have to try to allow their visibility within the environmental field to cross over into other domains. Particular individuals (such as Jane Goodall) have been especially successful in adopting this public intellectual role, particularly after taking on the role as a UN Ambassador for Peace in the 1990s.

It might be possible to approach this categorization of celebrity interest in the environment from the other way around and categorize not the celebrities involved, but the sorts of environmentalism for which they speak. There is a tendency for the rich and famous to support the landscapes and wildlife the rich and famous get to enjoy. One of the oldest conservation organizations, established in 1903 as the Society for the Preservation of the Wild Fauna of the Empire (now Flora and Fauna International), was popularly known as “the penitent butchers” because of their strong representation of the interests of wealthy big game hunters. Prince Bernhard’s support for, and secret funding of, violent antipoaching activity in southern Africa is very much in that tradition of aristocratic support for landscapes the wealthy love. Another tendency is that animal rights and welfare causes often tend to be popular among the glitterati; we could count Virginia McKenna’s Born Free Foundation as a prominent example. This introduces interesting tensions within the conservation movement as the animal rights movement is not necessarily the same thing as environmental conservation, which is more concerned with landscapes, species, and ecological systems than with individual animals. Thus, for example, Born Free activists have named and otherwise identified individual lions and will go to great lengths to protect each one, even in circumstances where the funds expended might arguably generate a
greater conservation benefit if the focus was the protection of whole ecosystems.

What is missing, perhaps unsurprisingly, from this list are the “environmentalisms of the poor.” These are the concerns for resources – firewood, thatching grass, grazing, and access to the same – which tend to characterize the lives of rural people in poorer parts of the world. There are relatively few nongovernmental organizations (NGOs) who champion these causes, at least few of the sort which work with celebrity, and fewer celebrities who can lend their name to this sort of environmentalism with integrity and authenticity. Vandana Shiva (an Indian environmentalist) or the late Wangari Mathai (a Kenyan Nobel peace prize winner) might count among that number, but they are a notable minority.

The history of fame and celebrity is a little unusual. As the list of famous environmentalists above demonstrates it has been a more or less continual feature of public life since the late 1890s. There has been a steady small stream of public spokespeople for natural causes for a long time. This is the not the case in other causes. International development, for example, was one of the main causes of fame in the Victorian era, but produced no famous figures in the first decades after World War II (Brockington 2014).

Recently, however, that has changed, and this is to do with the general rise of celebrity advocacy across all causes. It is undergirded by the rise of NGOs more generally from the late 1980s onwards. As neoliberalism took hold and faith and investment in states declined so NGOs stepped up, and forward, to become the vehicles of conservation, development, and environmental projects. They have mushroomed in number and size, and with it has their need for good publicity and celebrities to support their causes.

Mirroring that increase has been the rise of celebrity itself. This is not a recent phenomenon – historians of celebrity date its significance as a social phenomenon from at least early Victorian times (Boorstin 1992/1961). Contemporary forms emerged with the rise of cinema in the early years of the twentieth century. Celebrity flourishes in part where different media provide new ways of being seen and new potentially lucrative sources of income. Accordingly when the forms of media multiply – with cable and satellite TV, new glossy magazines, and the Internet – then the number of celebrities also multiplies. This is what happened in the 1990s (Marshall 1997). Thus the rise of celebrity advocacy is partly due to the rise of celebrity itself.

But we should not think that technology determines celebrity. Celebrity is also a means of organizing fame and appearance (Gamson 1994). And this is particularly visible in the last 15 years when there has been a reorganization of the relationship between the celebrity industries and the NGO sector. This is now coming to drive the relationship between environmental causes and celebrity.

In its simplest terms, the work of the famous for good causes, including environmental issues, is now much more systematic and organized than before (Brockington 2014). Celebrity liaison officers have full time posts in most of the major charities in the United Kingdom; these have generally been established since 2000. Three of the largest talent management companies in Hollywood have established foundations to manage the charitable interests, and appearances of their clients, also since 2000. There are now websites (Look to the Stars) which document the charitable activities of the famous and newspaper columns doing the same. Celebrity advocacy has become a specialist niche of the celebrity economy more generally. Companies (such as the Global Philanthropy Group) exist to advise celebrities on their philanthropy and help them to give wisely and build their brand while doing so.
ENVIRONMENT AND CELEBRITY

Environmental causes feature prominently in these developments. Ecorazzi was one of the first websites documenting the work of the famous for environmental causes and, as its name suggests, specializing in environmental causes. There are also specialist nonprofits which work to link environmental causes for the famous. Consider, for example, the Environmental Media Association (EMA) whose “generation E” campaign involves partnering with environmental groups with EMA’s role being to secure “talent,” that is, celebrities, as well as all of the other paraphernalia of successful media events and networking to support media rich campaigns.

Undergirding this professionalization of celebrity advocacy in environmental causes is an important fact: the work of celebrity for environmental causes is part of a more general trend – the integration of capitalism with environmentalism (see also Conservation and capitalism). There is a tradition of seeing environmentalism as opposed to commerce and business, of resisting the advances of development, use, and modification of the environment. But, as we have just seen, there are many varieties of environmentalism. These days the dominant form of environmentalism, mainstream environmentalism, enjoys a close working relationship with corporate capitalism.

Celebrity environmentalism is part of that trend in several ways. In the first instance it is a business in itself. Celebrity environmentalism makes commercially valuable films, books, and images. This has been true for a while. Gray Owl rose to prominence as a great publishing success and made wildlife films. The Adamsons’ work with lions was also a publishing success story (although that was not the reason driving their work). But it is central to the business model of some wildlife and nature film presenters that they use their fame and public prominence to promote other businesses and products. The Irwins, an Australian family, are probably the most successful at this – even after the death of the main presenter Steve Irwin (famous for his Crocodile Hunter series). Others will sell access to their expertise and knowledge of nature in the form of safaris and holidays for the privileged.

But, more generally, the rise of celebrity advocacy within the NGO movement as a whole has been fueled by corporate hunger for access to celebrity. Celebrity sponsorship for commercial purposes is expensive. Associating with celebrities via NGO intermediaries is free. Therefore many companies are seeking links to NGOs, as part of their corporate social responsibility agenda, but are doing so partly on the grounds of the celebrity support to which they could then get access. It is routine these days for NGOs to advertise the possibility of access to celebrity patrons on their corporate sponsorship web pages. Indeed, NGOs have to be careful that their celebrity patrons do not inadvertently endorse companies’ products. Free association is allowed, but free endorsement would threaten the celebrity business model.

Celebrity environmentalism is therefore both manifestation and lubrication of the close relationship that many environmental causes enjoy with capitalism. At the more powerful end of the spectrum, where the more influential groups are gathered, the displays resulting are, literally and deliberately, spectacular. Prince Charles can muster the world’s leading financiers in the heart of the City of London in support of his campaigns to make the carbon in rainforests valuable. The Conservation Caucus Foundation brings together the world’s largest NGOs, companies like ExxonMobil, International Paper, BP, DuPont, and Walmart, over one hundred of the top politicians in the United States, and leading celebrities like Harrison Ford in prominent displays of concern and networking opportunities for environmental causes.
“Spectacular” here does not just mean an extraordinary event that draws the eye. It has a more formal meaning which further deepens the intertwining of capitalism and environmental causes, and is best explained in Jim Igoe’s work, drawing on the writings of Guy Debord (1995/1967; Igoe 2010). Debord wrote critically of how commodified images – what he called “spectacle” – shape social relationships in late capitalism. He argued that such spectacular images, and the processes that bring them to our attention through popular media in the first place, effectively obscure as much as they reveal. Igoe (2010) argues that this plays a vital role in the promotion of conservation and environmental causes. Using the example of the African Wildlife Foundation (AWF), he shows that images of happy and successful conservation initiatives, and the wielding of powerful brands by conservation NGOs, conceal the patterns of inequality and local impoverishment that often beset human populations living in or near conservation areas as well as profound and unexpected transformations to ecosystems. As Igoe puts it: “The fetishization of connections and relationships through spectacle thus shields Western consumers from the more complex and problematic web of connections and relationships in which they are actually enmeshed” (Igoe 2010, 389).

How can we better understand the role of celebrity in environmental affairs? There are three possibilities. In the first instance a constant gap in celebrity studies is the relative lack of audience studies. Outside of business schools’ studies of marketing and advertising (which produces samples largely composed of US undergraduates) audiences are relatively neglected. It has been suggested that celebrity advocacy is marked by the lack of purchase from which it suffers in many audiences (Brockington 2014). Many people are just not that interested. But this study was of British audiences and celebrity cultures and responses to celebrity vary enormously from country to country. Exploring the contours of this and the ways in which celebrity can be important to identity work would be important contributions.

Second, as Igoe’s work shows, it is important to follow particular consumer goods through society and explore the social relations they embody. Ethnographic approaches can be used to examine the creation, production, and consumption of different celebrity-environment spectacles and show what is elided, omitted, or concealed by them as much as what is supported.

Finally, Holmes has suggested, following Sklair, that we can speak of a transnational conservationist class (Holmes 2010). He has also shown that we cannot assume that this is a united class, but that there will be important divisions and differences within it based on, for example, national identity. Celebrity support for environmental causes is part of this elite formation. Accounts of how these elite networks function and what power they wield would be a valuable addition to this growing field.

SEE ALSO: Environment and the media; Environmentalism; Nature conservation; Neoliberalism and the environment

References

ENVIRONMENT AND CELEBRITY


Environment and consumption

Laura Pottinger

University of Manchester, UK

Awareness of the environmental consequences stemming from processes of production and consumption has grown in recent decades, as consumers become increasingly anxious about the negative effects of contemporary lifestyles. Persuading individuals to shift their consumption practices is a core strand of contemporary environmentalism, but the increasing popularity of politicized consumption is not without critique. Theorists in geography and beyond have paid close attention to the impacts of key spheres of consumption such as food, energy, transport, fashion, and tourism, and have sought to understand how environmentally damaging processes can be mitigated or reversed. Changing consumption behavior may be one solution that enables environmental resources to be better managed or protected.

The subset of consumption practices aiming (or claiming) to be more environmentally or socially responsible are variously referred to as “ethical,” “green,” “sustainable,” or “political” consumption. The diverse set of approaches encompassed by these labels are also sometimes referred to as “alternative” because of their oppositional or separate relationship to mainstream markets, though the degree to which individual approaches reject market relations varies. Fairtrade, for example, utilizes capitalist commodity markets and global supermarket chains to sell products in order to provide producers with a social benefit alongside profits. Other approaches entail a more fundamental rejection of consumerism, the attachment to materialistic values and possessions, and a growth-oriented economic model based on consuming in progressively greater amounts. “Voluntary simplicity” approaches, for example, instead advocate a reduction in overall consumption and simpler lifestyles. Broadly, green consumption as an environmental solution implies that disparate, individual purchases and everyday actions are significant when considered together, and can be connected to form a powerful political voice capable of reforming global actors such as corporations and governments. Critics, however, note its limitations for environmental protection, suggesting that consumer capitalism and environmentalism are fundamentally opposed.

Environmental implications of consumption

Historically, environmental resources were assumed to be inexhaustible inputs into capitalist systems of production and consumption. The ways these systems of commodity production functioned and evolved depended on treating the environment as an externality ripe for human exploitation. In the latter half of the twentieth century, this instrumentalist view shifted toward seeing the environment as fragile and in need of protection, and recognizing that finite natural resources could not be exploited indefinitely. Yet accelerated consumption patterns continue, particularly in Western economies and newly industrializing nations with burgeoning middle classes. This is having dramatic consequences...
ENVIRONMENT AND CONSUMPTION

for the climate, ecosystems, and biodiversity, as a result of raw material and energy input into production, distribution, use, and disposal of commodities.

Modern agriculture, industrial production, and personal transport, for example, are still largely dependent on oil and other finite, nonrenewable energy sources. This creates carbon emissions that contribute to anthropogenic climate change, and there are often controversial consequences of extraction, such as oil spills or violent disputes. Pollution of the biosphere, deforestation, and overfishing of marine resources can all be seen as stemming from the ways our food, clothing, houses, holidays, cars, computers, and so on are produced and consumed. These damaging impacts are exacerbated by how we consume and what we expect from our lifestyles, or are persuaded we deserve. Affluent consumers see buying fresh food and flowers out of season, disposable “fast” fashion, foreign holidays, and individual car travel as the norm, and global brands hotly compete to satisfy these desires. The rate and quantity at which consumers buy and throw away commodities has negative environmental consequences, particularly within the industrialized economies of the Global North. In the United Kingdom, for example, household purchase and disposal of uneaten produce accounts for millions of tons of wasted food each year, which has significant impacts not only for landfill (Figure 1) but also in terms of the energy and water used up in production.

History of consumption research

Consumption is defined as the using up of a good or service, and incorporates selection, purchase, use, upkeep, repair, and disposal as waste or recyclables. Processes of consumption closely overlap with those of production, and together these are fundamental elements of social life. Commodities, which are things produced for the purpose of exchange, can hold important social and symbolic value, and are utilized to display class, taste, gender, or group membership. In the consumptive act of eating food, the human body is sustained and biologically (re)produced, and consumption is also implicated in the production of meaning and identities.

However, the spheres of consumption and production were traditionally separated in academic research, with spaces of production historically receiving more attention than those of consumption, particularly in Marxist and economic theory. Early social theorists such as Theodor Adorno and Max Horkheimer in the 1940s took a moralistic stance, focusing on the negative sides of consumption, and conceived consumers as “passive dupes,” controlled and manipulated by consumer capitalism and advertising. In the second half of the twentieth century, Pierre Bourdieu drew attention to the role of gift-giving and “conspicuous consumption” of luxury goods in defining identities and maintaining class distinction. From a more optimistic perspective, later cultural theorists, like Stuart Hall in the 1970s, emphasized the creativity and resistance enabled by consumption practices, suggesting that consumers possessed the capacity to subvert and appropriate meanings associated with commodities.

The latter half of the twentieth century has been characterized by accelerated levels of human consumption and production accompanied by growing interest in the dynamics of consumption and its environmental impacts in popular writing and media coverage. Environmental concerns have moved up the agenda since the 1970s, with companies like the Body Shop among the first to recognize the potentials of green capitalism by aligning business with an environmental/social justice ethic. In one way,
the current fascination with green consumption marks the success of environmental campaigns and green agendas of the late twentieth century, which questioned the economic centricity of development paradigms. The Brundtland Report, published in 1987, redefined sustainable development as necessitating a balance between the economic, social, and environmental, implying that, rather than acting in opposition to environmentalism, economic growth can be compatible with preserving diverse ecosystems and protecting the planet (WCED 1987).

**Perspectives on green consumption in geography**

Until recently, consumption has been relatively ignored within geography. However, interest in the subject has grown in the last few decades, with research into consumption drawing on two major approaches. First, some work is driven by economic geography perspectives, including that on global value chains (GVC) and global production networks (GPNs). This research has furthered understandings of the
spatialized relations in networks of production and consumption and how they touch down and are embedded in communities, environments, and localities. Research on GVCs and GPNs has both informed and interrogated an array of social and environmental codes of conduct that seek to modify and adapt unsustainable production processes, such as the Ethical Trading Initiative, which is a multistakeholder alliance of companies, trade unions, and nongovernmental organizations (NGOs).

Second, other work reflects the “cultural turn” more strongly, and considers how consumption practices are made meaningful by consumers. This research draws economic relationships between production and consumption together with critical analyses of the cultural and material dimensions of consumer conduct. It addresses consumption as an emotional and affective practice, which can be motivated by desires for novel experiences or authenticity or by concern for others, for example.

A prevailing view across approaches within geography is that environmentally damaging processes persist because of a spatial and cognitive distance between spaces of consumption and production (Scales 2014). Powerful global firms like multinational supermarkets extend across the globe through vast networks of national suppliers, producers, and subcontractors. The effect of the increasing globalization of production practices, supply chains, and consumer brands is that consumers are dislocated from the impacts of their consumption on the environment. This means that they are less able to see the environmental effects of the products they buy, either because production is hidden behind the factory door or because waste products are exported to another country for disposal.

Some geographers have drawn on the theory of “commodity fetishism” proposed by Karl Marx, which suggests that a veil is drawn between commodities and the conditions of their production, concealing the social and environmental source of the value of commodities from the consumer. From this perspective, the disconnection between spheres of consumption and production allows “enchantment” with commodities to be created; their magical, alluring qualities are achieved by masking the true nature of their origins. For example, the glossy, multimillion dollar campaigns advertising the latest smartphone do not portray the long working hours factory employees endure or the conflict minerals used in their fabrication. The removal of consumers, both geographically and ecologically, from spaces of production makes them unaware of the environmental impacts of waste, resource depletion, pollution, and so on. This lack of understanding means that they may be less likely to change damaging environmental practices.

Informing consumers is a popular solution aimed at overcoming the dislocating effect of globalization and, in turn, combatting some of the negative environmental impacts of consumption. Underlying many campaigns that promote green consumption is the notion that consumers would act more ethically if only they were made aware of how the commodities they buy are produced. The thinking here is that consumers who lack knowledge can be educated, which will cause them to change their behavior, and that this in turn will persuade companies or governments to reform. Consequently, the way green consumption is mobilized often depends on increasing consumers’ understanding of environmental issues and the impact of their individual actions.

NGOs are among those trying to achieve this, including ethical consumption campaigning organizations and conservation charities, for example World Wide Fund for Nature (WWF), Greenpeace, and the Fairtrade Foundation. They
facilitate green consumption by providing consumers with information and advice, developing labeling and certification schemes, lobbying governments, and engaging with companies through multistakeholder groups. As a result, many corporations have also begun to recognize the value in “greening” their image, not only to mollify those customers dissatisfied with a poor environmental track record, but also to actively attract eco-friendly consumers and generate new markets. The rationale behind green consumption is that consumers wield the power to challenge global brands by voting with their cash, thereby persuading environmentally destructive corporations to change. Supermarket customers may be able to buy air-freighted asparagus all year round, but consumers can also choose to purchase less environmentally damaging products as a way of persuading companies to behave more ethically. However, while some argue that these increasingly complex and spatially diffuse supply chains are driven by consumer choice, others reject what they see as an oversimplified relationship between supply and demand, and highlight the ways that demand and desire are created and manipulated through advertising and branding.

There are two main ways in which NGOs and other campaigning organizations aim to reeducate dislocated consumers and to promote green consumption: first, by providing information; and, second, by reconnecting consumers with practices and spaces of production.

Providing information

Providing consumers with accurate, easy-to-interpret information is often seen as a panacea for encouraging behavior change and “better” consumption practices: buying eco-friendly products, boycotting environmentally damaging brands, or making lifestyle changes like recycling or using public transport. This may take the form of state- or NGO-led campaigns, or sensationalist exposés of unsustainable or undesirable practices. For example, the popular documentary film *Food Inc.* (2008) scrutinizes the corporate food industry, exposing the harmful outcomes for animal welfare, human health, and the environment resulting from the practices of major companies producing cheap food in the United States. Perhaps the most prominent method of information provision is via ethical product labeling (Figure 2), such the Soil Association’s Organic food certification and Forestry Stewardship Council (FSC) labeled timber.

While the assumption that educating discerning consumers will generate more environmentally friendly consumption behaviors has
permeated campaigns, media coverage, and academic literatures, it has also been criticized. The idea of the consumer who makes self-interested “rational choices” about the environmental impact of their purchases based on available information comes from classical economics. However, it has several flaws. First, consumers are not always rational in their purchasing decisions. Second, it assumes a straightforward transfer of information from the producer to the supposedly ignorant consumer, which overlooks the ways in which information is received, reworked, or even ignored by individuals. Third, it also fails to account for constraints such as time, money, or access that might prevent consumers acting on information provided.

As a consequence, people’s actions when consuming often do not match their stated pro-environmental intentions, producing what has been termed the “attitude–behavior gap.” The issues confronting “ethical” consumers are complex, often contradictory, and rarely are they easily encapsulated in product labels. For example, consumers may find it difficult to weigh up the negative carbon emission impacts from air-freighted fresh produce against the social justice imperatives of fairtrade that stimulate economies in the Global South. More recently, the idea of the “rational” consumer has been criticized for assuming that consumption practices take place in a vacuum rather than being embedded in social networks and contingent on complex, wide-ranging factors such as economic capacity and commitments to friends or family members. Research into ethical consumption has shown that much of what we may consider to be ordinary consumption actually has very little to do with choice. The justifications people give for not consuming ethically should perhaps be understood as expressions of genuine frustration at being burdened with too much responsibility (Barnett et al. 2011).

Reconnecting consumers with practices and spaces of production

A second approach that uses knowledge to promote green consumption is that of “reconnecting” consumers through localized or alternative consumption practices. This can be challenging, particularly for global issues such as climate change, where individuals are unlikely to directly register the impacts of their consumption on carbon emissions. Attention to moral geographies of consumption has raised the question of whether ethical actions are only possible in local places that we know or feel some connection toward. Since space or distance is often understood as a barrier to responsibility, one solution to the problems associated with consumers’ increasing separation from production has been the idea that we should “think globally, act locally.” This assumes that people are more likely to support their own communities, worry about the plight of local farmers, or defend nearby green spaces under threat of development, and so on, than to care for distant others.

Buying commodities produced locally or sold by small independent businesses is presented as one way consumers can overcome the alienating effects of globalized production. Many alternative food networks (AFNs), for example, are premised on reconnecting consumers with local agricultural producers. AFNs may include local vegetable box schemes (Figure 3), community gardens (Figure 4), or farmers’ markets, which provide food differently from mainstream market relations, often with a significant emphasis on community, trust, and place-based production. Sourcing locally grown food is appealing because, in addition to reducing the environmental impact of the transportation of food by cutting down “food miles,” it gives the impression that the network spanning production to consumption can be more easily known.
While buying local makes sense for consumers on many levels, uncritical championing of localization is problematic, especially where it works by identifying local and rural environments as bounded places under threat from outside influences such as global corporations. This presents the global/local as a binary opposition in which the local is the privileged location for trust and responsibility, while the global is seen as threatening. This can lead to defensiveness and intolerance of outsiders or new ideas. It also limits the scope of moral duty and action to immediate locations, suggesting that we can or should only care about those who are in close proximity. Furthermore, the ability of such reconnected consumption practices to expose the true nature of fetishized commodities is questionable. Eco-labeled commodities and ethical food are still marketed and made enchanting, just in a different way. The example of how food is offered within farmers’ markets (Figure 5) illustrates how ethical consumption itself can fetishize commodities. Physical proximity, interaction between producer and consumer, the aesthetics of the market, and promotional material that draws on geographical imaginaries of local places all help to repackage ethical food as “connected,” adding mystique and allure through ideas about authenticity.

**Politics and governmentality in green consumption**

Though we may have become more aware of the impacts of our consumption practices since the late twentieth century, researchers in geography are divided over the effectiveness of green consumption for protecting the environment. There are two main perspectives on political agency in green consumption. The more optimistic perspective highlights how consumption has become an important mode of political action that is democratic and accessible to a wider demographic than traditional party membership or political lobbying, and symbolizes a flourishing engagement in political and environmental issues by a wide range of people in their everyday lives. Theorists writing from this perspective (e.g., Barnett et al. 2011) stress the participative and collective dimensions of what they call “political consumption” (rather than “ethical consumption”). The implication is that naming consumption activity as “political” redefines it as action that makes collective claims.
on states or institutions, forcing them to act on environmental issues. Consumption is frequently understood in terms of voting, and this analogy seems to stitch together the seemingly individualistic nature of consumerism with the collective ideals of citizenship. Deirdre Shaw’s (2007) interviews with ethical consumers, for example, found that participants routinely expressed their purchases in terms of “votes,” and felt a reduction in their overall consumption would limit their power to effect change. The idea that we can register a vote for organic agriculture or rainforest protection with each purchase we make is compelling. Although some strands of ethical consumption seek to reduce overall levels of consumption, ethical consumption is attractive to individuals because on the whole it is not based on asceticism, constraint, and consuming less; rather, it is about consuming better.

This perspective argues that, as consumers, we want to feel empowered by this sense of agency rather than seeing ourselves as passive dupes, constrained by hegemonic marketing forces. By celebrating the citizenship dimensions and political dynamics of consumption, “political consumption” to a certain extent overcomes the moralizing tendencies displayed in previous academic and societal treatments of consumption. Importantly, individual acts of shopping
or purchasing can be joined together through ethical consumption to create a powerful collective voice. This perspective argues that the policies and marketing strategies reconnecting consumers spatially and cognitively, such as labeling, do mean that consumers are more aware, and in some cases environmentally irresponsible corporations have been persuaded to clean up. The logic that companies will respond if their action equates to more sales makes green consumption seem like a win–win solution.

The second perspective sees green consumption less positively as reformist, neoliberal, and symptomatic of citizens’ disengagement with traditional modes of politics. Critics suggest that the increasing turn to lifestyle politics demonstrates disillusionment and apathy with traditional forms of political lobbying and campaigning. Some have noted that this heralds the increasing commodification of nature and the social realm, as entities that were previously viewed as unsaleable are transformed into goods that can be traded. An example is emissions trading, an approach adopted by various countries and aimed at controlling pollution, which uses a market-based system of providing economic incentives for reducing pollutant emissions. Carbon credits can be bought and sold by governmental bodies, effectively enabling the purchaser to buy the right to pollute.

In contrast to prevailing paradigms of sustainable development, this more skeptical perspective argues that green consumption can only reinforce rather than challenge the logics of neoliberalization and relentless economic growth, which are fundamentally incompatible with environmentalism, limiting political activism to questions of what to buy (or not buy). This perspective sees the politicized representation of consumption—as voting as problematic in a number of ways. First, by aligning consumption with political clout, it seems to encourage increased levels of consumption, which both fails to address the corporate interests this serves and the material environmental impacts of overconsumption, such as resource depletion, landfilled waste, and energy usage. It also sidesteps the fact that, while democratic principles allocate each citizen with the same number of votes, not everyone can participate equally in a “market democracy,” because it privileges those with more money, resources, or human capital.

Second, the rhetoric of consumer choice is used to stimulate awareness and evidence support for green consumption, but emphasizing choice displaces responsibility into the hands of individual consumers. This prevents claims being made on the real culprits: more powerful actors, such as states, institutions, or corporations, who are better equipped to mitigate environmental ills. Third, focusing on consumer choice and “ethical” consumption invokes the “unethical,” “irresponsible” consumer who must be persuaded to act differently and make better decisions. This could be seen as a renewed moralism and judgmental stance that is both elitist and exclusionary. From this perspective, ethical consumption is too tied up with questions of taste, class, and status displays of eco-awareness, which perhaps reflect the expendable wealth of the consumer more than their environmental friendliness.

One way to analyze green consumption from this more pessimistic perspective is using theories of governmentality to address the role of identity and subjectivity in environmentally conscious consumption. Drawing on Michel Foucault’s work, this theorizes the intersection of the “consumer-citizen” and neoliberal forms of governance, in which responsibility is seen to be devolved to individuals whose preferences are expressed through the marketplace rather than by making claims against a social state. These approaches aim to understand how
consumers are governed and controlled, and have aided political economic theorizations of the connection between macro-level attempts to shift from state to market and the everyday processes at work at the individual level. These approaches often argue that “ethical” consumers may inadvertently be complicit in maintaining systems of inequality or environmentally damaging practices.

Drawing on governmentality theory with reference to alternative food networks and organic agriculture, Julie Guthman (2008) laments the way that what counts as political action today has been reduced to the individual purchases made by consumers. For Guthman, considerations of what to eat have taken precedence over collective boycotting. She argues this has been damaging to environmental movements seeking change in agricultural practices, and has “made agro-food politics so anemic” (Guthman 2008, 1181). Efforts to build what she deems marginal alternatives not only distract environmental activists from the big issues, but they foreclose collective options and are in fact very closely aligned with the neoliberal discourses of consumer choice, localism, entrepreneurialism, and self-improvement.

Under this perspective, even radical alternatives like organic food can be seen to change in character as they are mainstreamed into conventional consumption channels such as major supermarket chains. Both the scale and methods employed in organic agricultural production have been dramatically transformed to enable consumers to buy fresh organic produce year-round, bringing what began as a resolutely alternative food institution in line with conventional industrial agriculture and diluting the progressive environmental intentions of organic producers. In other sectors, radical alternatives have been co-opted superficially to “greenwash” the image of less sustainable companies, through corporate social responsibility initiatives or the creation of new “green” products. Paradoxically, promoting green consumption can be seen as creating new markets and industries rather than reducing consumption (Luke 1997).

New directions in theorizing environment and consumption

These critiques may imply that consumption-based approaches rarely accomplish more than assuaging an individual sense of inertia, making us feel better about our impact on the planet while substantively achieving relatively little. To overcome this, some geographers suggest that we move away from thinking about individual consumers and, instead, pay attention to the intermediary networks of campaigning groups and NGOs that mobilize ethical consumption (Barnett et al. 2011). What these organizations are very good at is connecting purchasing practices to wider themes of environmental stewardship or solidarity by providing people with resources to facilitate their engagement. This interpretation sees consumption not as an individualizing and depoliticizing practice, but as one of many activities linked together in a broad, collective effort that forms part of contemporary environmentalism.

What this more recent work shows is that individuals rarely identify themselves as “consumers,” but instead use purchasing practices to further their commitment to existing environmental concerns. Understanding individuals’ narratives about their prevailing commitments and routines, therefore, has become an important object of research. Contemporary studies of consumption have moved beyond conspicuous green identities and choices to address the everyday ethics at play in more mundane practices of “ordinary” consumption such as showering or
using electricity (e.g., Hall 2011). This atten-
tion to practices of consumption marks a move
from understanding agency in the purchases of
rational decision-makers, to recognizing how
consumers’ desires and behaviors emanate from
everyday practices and ordinary routines. This
area of research pays attention to the material
and social settings in which consumption prac-
tices are embedded and shows how attempts at
changing behavior are framed by infrastructural,
institutional, and cultural contexts.

In contrast to the moralizing tendencies
sometimes associated with ethical consump-
tion, researchers have found fertile ground in
exploring the embodied dimensions of green
consumption, looking at the role of bodily
sensations, feelings, and emotions in fostering
environmental commitments. Geographers and
anthropologists have addressed the pleasurable
and aesthetic dimensions of sustainable
consumption, particularly around alternative
models of food distribution and consumption
such as organic vegetable boxes, Slow Food,
and Community Supported Agriculture. These
studies have elucidated the tangible, tactile mate-
rialities of alternative food, and have revealed
how consumers experience these types of green
consumption. They highlight the importance of
looking at the roles played by moods, aesthetic
sensibilities, and pleasure in nurturing ethical
and political behavior, alongside individuals’
verbally articulated rationales for involvement.

While mobilizing the pleasurable dimensions
of alternative green consumption may be a wor-
thy pursuit for ethical consumption campaigner-
s and academics alike, critics note that this
emphasis on pleasure compromises the efficacy
of consumption as activism, as these practices
are inseparable from questions of taste, quality,
and distinction. Alternative food networks are
perhaps too entangled with making choices
about what is “good” to eat, and alternative
food itself is often valued by consumers as much
for its flavor, health benefits, and freshness as for
its contribution to challenging environmentally
damaging practices.

Alongside renewed interest in consumers’
routines and experiences, promising avenues
have opened for exploring moral geographies
of consumption and how ethical consump-
tion practices are distributed across space.
Studies of alternative consumption such as eco-
tourism and farmers’ markets, for example, have
demonstrated how geographical imaginaries of
destination/producer countries are commodi-
fied and marketed as environmentally conscious
(Coles and Crang 2011). These studies draw
attention to the way space is consumed and
culturally encoded, and to how the practices of
consumers both in space and of space itself are
tied up with values of quality, taste, and distinc-
tion. Importantly, researchers have highlighted
concerns that spaces of alternative consumption
such as farmers markets can be elitist enclaves
that dissuade nonwhite and lower-income
groups from participation, and that they may
be implicated in the potentially exclusionary
processes of urban gentrification.

Geographers are also moving beyond notions
of the environment as a valuable external
resource and essential input into processes of
production, to consider how consumption takes
place in space and how we consume space
and place. Emerging research into the tactile,
emplaced, and experiential facets of consump-
tion, such as studies exploring urban food grow-
ing, community gardening, or foraging where
the distinction between producer/consumer
becomes blurred, for example, is beginning
to overcome problems associated with view-
ing the environment as an externality that
individuals must be persuaded to care for at a
distance. With an increased sensitivity to the
interconnected nature of environment–society
relations, researchers are gaining a clearer picture
of how environmentally conscious behavior
ENVIRONMENT AND CONSUMPTION

is experienced and may be motivated in the everyday, ordinary consumption practices of individuals.

SEE ALSO: Alternative food movement; Consumption; Corporate environmental responsibility; Environment and waste; Environmental certification and eco-labeling; Environmentalism; Environmentality and green governmentality; Green capitalism

References


Further reading

Recent decades have witnessed a surge of interest among scholars, policymakers, and activists in the linkages between environment and development. Growing recognition of the multiple environmental crises facing humanity as well as increased awareness of the connections between human welfare and environmental integrity have made clear the need to study issues of environment and development from integrated, holistic perspectives. At the same time, environmental concerns have been progressively incorporated into development policy and planning at local, regional, national, and international scales.

The “greening” of development thought and practice is most clearly evidenced in the rise of the sustainable development paradigm, which development agencies, lending institutions, non-governmental organizations, and governments around the world have embraced as an approach for uniting social, economic, and environmental aims in development agendas.

Human–environment geography has been uniquely positioned to probe the environment–development nexus given the field’s long tradition of research on the dynamic and recursive relationship between the nonhuman, biophysical world and processes of social and economic change. Nonetheless, delimiting an “environment and development” field of research in geography is a challenging assignment. The difficulty derives, in part, from the complex and shifting meanings of the two terms. As Cowen and Shenton (1996) explain, “development” tends to be conceived of in two broad ways. On the one hand, it is understood as an immanent process, one that unfolds historically without explicit direction (as in “the development of capitalism”). On the other hand, development is thought of as a positive transformation – typically economic or social in character – that is deliberately fostered. This latter notion of development as an intentionally pursued improvement has become increasingly predominant in the post–World War II “age of development.” Nevertheless, disagreement continues over how exactly development should be defined and measured. Sachs (1999, 7) has gone so far as to claim that as development agencies have diversified their emphases and strategies – moving away from a traditional focus on economic growth and incorporating themes such as injustice, gender, and the environment – the result has been “semantic chaos,” with development becoming “a shapeless amoeba-like word.”

Probing the term “environment” also reveals ambiguous and competing meanings. Eden (2007, 62) notes that while the environment once signified, in a relatively unproblematic manner, “the surroundings that we seek to manage or that influence us,” the term no longer enjoys such definitional modesty. In the academic literature, ambiguity regarding the meaning of environment is related to the fact that it tends to be used interchangeably with “nature” – a term laden with its own history of complex and disputed meanings. This word-switching occurs despite some generally, if implicitly, agreed-upon distinctions between the two terms: nature, for instance, is typically thought of as organic and
environment as inorganic (Eden 2007). Beyond these academic ambiguities, social constructivist research shows that understandings, valuations, and representations of the environment shift over time and space, and that the environment’s multiple constructions – for example, as a natural resource repository, as an assemblage of “goods” and “services,” as a web of interconnected living and nonliving things, and as a threatened landscape – are the products of contextual factors including media portrayals, economic imperatives, scientific understandings, and prevailing value systems.

How “environment and development” is bounded as a field of study, then, is a function of how the two terms are understood. As the content of this entry demonstrates, most research in geography that explicitly adopts an “environment and development” framing is concerned with the theories, policies, and practices of intentional development in its assorted forms. The entry also, however, engages relevant geographical scholarship on the relationship of environment (broadly conceived) to more general processes of capitalist development. The picture that emerges is of a robust, multifaceted area of study, one notable for maintaining a critical awareness of development’s complex and uneven socionatural implications.

**Development through resource mobilization**

One of the principal ways that the nonhuman, biophysical world intersects with – and enters into – development processes is in the form of natural resource inputs. Indeed, strategies to “mobilize” resource endowments have figured prominently in a wide variety of development policy frameworks. In the post-World War II years, as the emerging field of development economics came to focus on promoting economic growth and modernization in the “underdeveloped” regions of Asia, Africa, and Latin America, intensified resource exploitation was often seen as a critical means for achieving development ends. Rostow’s influential linear stages of growth model is illustrative (Rostow 1960). Contending that “it takes more than industry to industrialize,” Rostow posited that for a traditional society to develop economically it was necessary to boost productivity in agriculture and the extractive sectors via the application of modern science and technology, the rationale being that increased output from these resource industries was needed to provide the food, raw materials, and capital for industrial expansion (Rostow 1960, 22). Ideas like these animated the work of the World Bank, International Monetary Fund, and other international development organizations in the post-World War II decades, resulting in a stream of development funding into resource-intensive megaprojects. At the same time, intensified resource exploitation also played a central role in the state-led development and modernization programs of twentieth-century socialist governments (Khrushchev’s Virgin Lands Campaign being a notable example), as well as in the autarkic development schemes of postcolonial governments in the Global South. Meanwhile, the broad transition toward a neoliberal development paradigm in the last decades of the twentieth century frequently resulted in a “doubling down” on natural resource extraction, as free-market policies opened countries’ resource sectors to foreign investment, and international financial institutions encouraged the expansion of resource industries, including as a means of debt repayment.

Thematically, work in geography on the role of resource mobilization in development and
modernization programs has proceeded along various fronts. One line of investigation has focused on the large-scale water-infrastructure re-engineering schemes that have been planned and carried out by governments across the Global North and South. Such projects have often involved the construction of large dams, which have been extolled by proponents not only for their practical benefits (e.g., electricity generation, urban water provisioning, flood control, agricultural expansion), but also as symbols of modernity. Another body of research has centered on programs of agricultural development and modernization, including those associated with the “green revolution.” Coined in the 1960s, this term refers to an assemblage of agricultural technologies, inputs, and practices (including hybrid, high-yielding seeds; synthetic pesticides and fertilizers; and intensive irrigation) that came to be strongly promoted by international development agencies as a means of boosting agricultural production – and averting food crises – in “Third World” countries. A third line of research has focused on the efforts of political leaders to promote development and modernization by exploiting subsoil resources like oil and minerals. Programs designed to capitalize on these subterranean “riches” through extractive-industry expansion have commonly held the twin objectives of sparking regional and national economic growth, while also providing revenue for states to pursue other development initiatives – an approach on display in Venezuela’s long-standing attempts to “sow the oil” and in Bolivia’s recent efforts to use natural-gas revenues for social programs.

At a conceptual level, work on resource-based development strategies has benefited from engagement with Marx’s historical materialism, which has provided geographers a theoretical framework for understanding capitalist economic development as a continuing process of resource mobilization. Marx considered the rise of capitalism to have been predicated on a process he referred to as “primitive accumulation.” This involved the “freeing” of land and other natural resources from the fetters of noncapitalist social formations, thereby allowing for their incorporation into capitalist circuits of accumulation – through, for instance, enclosure and privatization. Marx’s treatment of primitive accumulation focused on its role in England’s historical transition from feudalism to capitalism. Recent scholarship in geography and cognate fields, however, has stressed that key aspects of primitive accumulation remain central to present-day, “advanced” capitalism, as witnessed in the recurrent efforts of capitalist firms to appropriate publically or communally held natural resource stocks.

Work on these ongoing efforts to make land and resources available for capitalist development dovetails with scholarship – in both Marxist geography and the field of industrial ecology – that engages the economy as a “metabolic engine” through which labor is continually applied to elements of nature in the production of commodities. Such conceptualizations of the economy as a process of material transformation are, in turn, helpful for appreciating why such a wide variety of development models, from across the political and ideological spectrum, have included strategies of intensified resource exploitation. These understandings also serve to underscore the basic fact that development is inherently an environmental – and, thus, geographical – process, one that involves the reshaping of landscapes, the reorganization of human–environment interactions, and the generation of new socionatures. Research on development’s material-geographical transformations further highlights the extent to which development is a highly contradictory process,
one that entails social and environmental costs and benefits, which tend to be distributed unevenly in both social and spatial terms.

Development’s contradictions

The idea that development gains inevitably flow from expanded natural resource production—an idea underpinning much post-World War II development thinking—has been increasingly called into question by scholars from a variety of fields. A multidisciplinary body of work on the “resource curse” has been especially notable for casting doubt on conventional assumptions regarding the role of resource mobilization in development. In particular, studies in this area have found that natural resource abundance is often, in fact, associated with poor economic performance, along with various other social and political maladies. While the resource curse literature has assorted strands, research tends to center on the consequences stemming from resource-rich countries’ overreliance on export-centered primary commodity industries, such as fishing, forestry, hydrocarbon extraction, and mining. This overreliance, resource curse theorists aver, leads to political and economic “distortions” that ultimately undermine the contribution of natural resource exploitation to development.

While the resource curse thesis has gained traction in development policy debates, human–environment geographers have pointed to several limitations of this body of literature (Bridge 2008). These include a bias toward national-scale analyses, which leaves research ill-equipped to account for the multiscalar and spatially discontinuous socionatural impacts of resource exploitation, as well as a tendency to privilege physical resource availability as the primary determinant of development outcomes—a tendency that may obscure important social, political, and economic factors. Thus, while research in geography has also called attention to many of the social and environmental costs associated with development through resource mobilization, geographical scholarship, by and large, has favored in-depth studies that capture the socially and spatially complex ramifications of these activities. This work has underscored that while some people and places may reap the rewards of new patterns and trajectories of natural resource exploitation, others bear the negative social and environmental implications of these activities—in the form of, for instance, displacement, environmental degradation, marginalization, and pollution.

A leading example of scholarship in this vein is Hecht and Cockburn (2010). This influential work within the field of political ecology offers a detailed, critical analysis of the efforts of Brazil’s political leaders and economic elites to develop the country’s Amazônia region. Training their sights on the large-scale development and modernization schemes of Brazil’s military dictators of the 1960s and 1970s, Hecht and Cockburn document the environmental destruction and social dislocation that these programs unleashed, while also showing how the negative socionatural implications of development policies have been experienced especially acutely by the poor and indigenous residents of the Amazônia.

Geographical research has also brought to light the sociospatially uneven impacts of large-scale water development. This research shows that while grand projects of waterscape re-engineering may benefit some actors by fostering certain activities (e.g., industrial expansion or urban growth), other groups typically bear the costs of these projects (e.g., displacement, livelihood destabilization)—a dynamic that has been on display in cases like the Sardar Sarovar Dam (Narmada River, India) and the Three Gorges Dam (Yangtze River, China). In
addition, scholarship on the political ecologies of agricultural development – including green revolution initiatives – has drawn attention to the environmental impacts of the transition to “modern,” resource-intensive agriculture (e.g., reduced genetic diversity, unsustainable water usage, agrochemical contamination), as well as its social ramifications, such as the marginalization and disempowerment of small-scale farmers. Meanwhile, work on oil-based development strategies has emphasized that such efforts are frequently predicated on the creation of “sacrifice zones” – areas such as the northern Ecuadorian Amazon, the Niger Delta, and northern Veracruz (Mexico), where the negative socionatural effects of intensive oil extraction are concentrated.

Through scholarship like this, geographers have contributed to a more general critique of the post-World War II development project that has been mounted in recent decades by scholars and activists, including those associated with the “postdevelopment” school of thought and political action (e.g., Sachs 1999). By recording and explaining the ground-level effects of resource-intensive development activities, geographical scholarship has called into question the extent to which the forms of development promoted by the postwar development apparatus actually constitute “improvement” and “progress,” thus contributing to a destabilization of the very meaning of development. Moreover, the uneven, contradictory, and often counter-productive character of resource-intensive development suggests, for many, a fundamental tension between environment and development. For Marxist scholars, the idea of such a tension is captured in the concept of the “metabolic rift,” which refers to the ecological disruptions and crises that result from capitalist expansion.

Work on development’s social and ecological contradictions also raises questions about how such programs have been – and continue to be – promoted across various social, political, and geographical contexts. In this respect, geographical scholarship on the discourses and ideologies of development is insightful. Influenced by post-structuralist approaches, especially Foucauldian discourse theory, geographical research has shown how development actors mobilize particular historical narratives, spatial imaginaries, and scientific “truths” in their efforts to advance their development agendas, in the process making some development trajectories appear preferred or predetermined, and others impossible. In their introduction to the influential edited volume Liberation Ecologies: Environment, Development, Social Movements, Peet and Watts (1996) deploy the concept “regional discursive formation” to refer to the constellation of ideas and imaginaries that run through a region’s discursive history and create the conditions for particular development interventions to be “thinkable” – often at the expense of alternative development visions. Such scholarship helps to elucidate the stakes involved in development-related sociopolitical conflicts. As well as being material struggles over productive resources (and the distribution of costs and benefits generated by their exploitation), these conflicts often have strong symbolic dimensions, with the actors involved contending for “control over the human imagination” (Peet and Watts 1996, 37).

Development environmentalized?

Within the broader arena of world politics, serious problems with orthodox, resource-intensive development had become progressively more apparent in the last decades of the twentieth century. One set of concerns centered on issues of natural resource availability. The notion that the Earth imposes “natural limits” on human
societies has deep roots in Western thought, with one especially influential origin being Malthus’s writings from the late-eighteenth and early nineteenth centuries on the relation between population growth and food supply. A number of studies in the 1960s and 1970s, meanwhile, rekindled alarm over the Earth’s capacity to supply resources for continued demographic growth and economic expansion. These included The Limits to Growth, a study that was commissioned in the early 1970s by the Club of Rome, a nonprofit research foundation. Based on the results of a computer simulation that modeled the interaction of natural and human systems, the study’s authors suggested that population growth and economic expansion would, in the not-too-distant future, lead human societies to “overshoot” the world’s carrying capacity, resulting in the collapse of the global system.

At the same time, concern also revolved around the impacts of modern industrial development on environmental conditions and human health. Rachel Carson’s Silent Spring, which dealt with the detrimental effects of pesticide use on the environment, raised awareness after its publication in 1962 regarding the ecological implications of industrial production, especially in the United States. Public awareness regarding the negative side effects of modern industry was further heightened by a number of high-profile cases of industrial pollution in countries of the Global North and South, including the Love Canal disaster (United States), the Bhopal gas tragedy (India), and the nuclear catastrophe at Chernobyl (Soviet Ukraine).

In the context of these mounting concerns, a new paradigm, sustainable development, took hold in global policy circles. As Adams (2009) details, the concept of sustainable development has a complex provenance, with roots in various scientific, intellectual, and political-institutional currents – colonial nature conservation and scientific resource management among them. Meanwhile, the core ideas of sustainable development – including the notion that environment and development are reconcilable – emerged from the 1972 United Nations (UN) Conference on the Human Environment, held in Stockholm, and the term itself was codified in 1980 by the International Union for the Conservation of Nature in its The World Conservation Strategy. Then, the World Commission on Environment and Development’s (WCED) 1987 report titled Our Common Future – also known as The Brundtland Report, after the WCED’s chair, Gro Harlem Brundtland of Norway – propelled the concept into prominence on the global stage. The UN had established the WCED in 1983 and charged it with assessing and proposing potential solutions to the environment and development problems facing world society. WCED findings were published in Our Common Future, which underscored the need to move toward sustainable development, for which it gave the following, now well-recognized definition: “development that meets the needs of the present without compromising the ability of future generations to meet their own needs.” The term swiftly migrated into a wide variety of policy circles, and the UN has subsequently organized several major conferences aimed at implementing sustainable development, including the 1992 Conference on Environment and Development in Rio de Janeiro, which produced Agenda 21, an action agenda for sustainable development in the twenty-first century; the 2002 World Summit on Sustainable Development in Johannesburg; and the 2012 UN Conference on Sustainable Development in Rio de Janeiro.

Geographers have engaged the discourse and practice of sustainable development in a number of ways. Like numerous other observers, geographers have drawn attention to the amorphousness and malleability of the concept,
which helps to account for the extraordinarily broad range of political and economic actors that have “gotten on board” with sustainable development. Critics contend that while the idea of sustainable development holds radical potential as an alternative to traditional development models, this potential has been diluted as powerful actors have laid hold to the concept and defined it in ways that buttress current power relations and political-economic trajectories. This tendency was on display at the 2002 World Summit on Sustainable Development, which produced agreements that entrenched the idea that neoliberal, market-based strategies should be prioritized in efforts to achieve environmental protection and social development (Mansfield 2009). It is also seen in the efforts of private firms – including mining, oil, and gas companies – to co-opt and deploy discourses of sustainability as a means of justifying the continuation of their activities.

Given the manifold uses to which ideas of sustainable development have been put, understanding the implications of the “sustainability turn” in development thought and policy requires that researchers investigate: what exactly is being sustained; why it is being sustained; to whose benefit; and to whose disadvantage. Geographical research that has addressed these issues in the context of concrete sustainability initiatives has revealed sustainable development to be a highly political – and often strongly contested – endeavor (for valuable examples of such research, see chapters in Krueger and Gibbs 2007). Such accounts of the politicized nature of sustainability run counter to the overwhelmingly technocratic and managerial framings that characterize mainstream discussions of sustainable development (Mansfield 2009). In this sense, work by human–environment geographers on the late-twentieth-century “greening” of development intersects with the field’s now significant body of literature on the histories, practices, and discourses of nature conservation, including work on protected-area formation in various historical-geographical contexts. While nature conservation initiatives, like sustainable development programs, are frequently framed in apolitical terms (e.g., as noble efforts to preserve “at risk” landscapes), work by political ecologists and others has exposed the uneven power relations and complex political-economic agendas shaping the rollout of these initiatives – laying bare, for instance, the ways in which the creation of protected areas in diverse world regions has been entangled with capitalist expansion, the consolidation of political authority (including that of colonial powers over colonized lands and peoples), and state-based projects to order and rationalize rural landscapes and their inhabitants. Further, research has revealed the often-harmful implications of conservation-area expansion for local populations, who have commonly found their resource and territorial rights restricted by interventions aimed at “saving” the nature on which they depend for their livelihoods.

Significantly, the late twentieth-century rise of the sustainable development paradigm took place in conjunction with a reorientation of conservation strategy within global environmental policy circles. With the negative socionatural consequences of conventional “fortress conservation” models increasingly apparent, international conservation organizations came to champion “participatory” approaches that could meet the social and economic development needs of people living in areas targeted for conservation – the result being a bevy of what have often been referred to as “conservation-with-development” initiatives (Perreault 2009). Thus, the “greening” of development has been associated with a concomitant “socialization” of nature conservation. While these trends have intersected in several
important ways, a shared emphasis on livelihoods has constituted a key point of convergence.

The livelihoods approach and the question of scale

Increasing understanding within the development community of the limitations of traditional, large-scale development programs for improving the welfare of the poor (who are often, in fact, further marginalized by these programs) has contributed in recent decades to a growth of interest in “bottom-up,” or “grassroots,” development approaches. In general, these approaches are strongly actor-oriented, often placing the household or the community at the center of development strategy; seek to foster the participation and empowerment of the poor, while also raising their standard of living; and place emphasis on incorporating local or indigenous knowledge into development policies. These “neopopulist” approaches, as De Hann (2000) explains, place faith in people’s adaptive capacities and, in particular, their capacity to create enhanced and sustainable livelihood systems even in the face of changing environmental, social, political, and economic circumstances. The result is that such approaches embody a relatively optimistic view of the environment–development nexus. By emphasizing human agency, neopopulist developmentalism has countered the determinism of orthodox Marxist development theory (e.g., dependency theory). It has also challenged the tendency of mainstream development to blame the rural poor for environmental degradation.

This shift toward “people-first” development policies has coincided in the academic literature with the rise of the “livelihoods perspective.” As Perreault (2009) discusses, though the concept of livelihood has become a mainstay of environment and development geography, there exists little agreement regarding the exact meaning of the term. Broadly, however, “livelihood” is used to refer to the ways that people make their lives – in social, economic, and cultural terms – as well as the capabilities, relationships, and assets (including natural resource endowments) that allow them to do so. The appeal of the livelihood perspective, according to Scoones (2009, 172), lies in its rather straightforward call to “look at the real world, and try to understand things from local perspectives.” Much of the work on livelihoods that has been undertaken in human–environment geography has sought to understand the complex and diverse ways in which individuals, households, and communities are able to adapt to and make the most out of their situations in the context of structural constraints and different forms of political-economic and environmental change.

One of the great benefits of the livelihoods perspective is that it counters a bias toward national-scale analysis that is present in much environment and development research. In addition to the resource curse literature cited above, many sustainability metrics take the nation-state to be the preferred scale of analysis, as do several other theories and models that have been developed to explain the relationship between socioeconomic development and environmental change, such as the environmental Kuznets curve and the forest transition theory. Studies adopting livelihoods approaches, in contrast, provide significantly more fine-grained accounts of environment and development dynamics given both the explicit focus of these studies on local-scale processes as well as a tendency for researchers to adopt qualitative methodologies. The result is that such studies typically allow for greater attention to social and spatial variation in environment and development dynamics at subnational scales.
There is a risk, however, that research adopting a livelihoods approach becomes overly localized and fails to account for the diverse ways in which people’s practices of production and social reproduction are shaped by – and depend upon – extra-local processes and systems. The rural development literature is indeed now replete with evidence of the various ways in which livelihoods, even in the more remote parts of the world, have become diversified in ways that make the inhabitants of rural spaces increasingly reliant on nonfarm, nonlocal forms of income, such as wage labor and migration remittances. In this context, it is important to resist the assumption that livelihood systems are spatially bound – that is, simply “local.”

Human–environment geographers have been, on the whole, cognizant of this risk and have guarded against it. This is especially the case among political ecologists, who have long insisted on the need to undertake multiscalar analyses that contextualize the decision-making of resource users and environmental managers within wider-level processes and power relations, from the household-scale to the global-scale. McSweeney (2004), through a case study of the dugout canoe trade in Central America’s Mosquitia, provides a valuable analysis of the complex and multiscalar spatialities of rural livelihoods. With the objective of understanding the lived experience of the Mosquitia’s rural poor, McSweeney’s analysis of the historical influence of extra-local forces, namely international capital, on the dugout canoe trade draws attention to the ways that supposedly isolated livelihoods in the region have long been shaped by interactions with broader-scale networks and economies.

Another potential peril associated with the “localization” of environment and development policy and research relates to a tendency to conflate “local” with just, democratic, and sustainable. In an important intervention, Purcell and Brown (2005) refer to this assumption as the “local trap” and contend that it implicitly or explicitly runs through academic and policy discussions. As noted above, the turn toward “bottom-up” development strategies can be understood as a response to the “top-down” character of traditional development. Purcell and Brown recognize that many calls for more localized decision-making in development stem from the real problems associated with neoliberal globalization and mega-development projects (which tend to be managed by national- and international-scale actors). They argue, however, that there is nothing inherently more socially just or environmentally sustainable about the local. Moreover, the evidence suggests that, in practice, efforts to localize environment and development decision-making have had mixed results. Drawing on work from critical geography on the social construction of scale, the authors note that scales of decision-making should not be understood as having any inherent qualities in themselves, but rather “as strategies that are pursued by and benefit social groups with particular social and environmental agendas” (Purcell and Brown 2005, 280). The task for researchers, then, is to pay close attention to the specific agendas of those actors who are empowered by efforts to “re-scale” environment and development decision-making toward the local level – for it is these agendas that will determine the socionatural outcomes of localization strategies.

Differentiated communities, heterogeneous environments

In-depth, place-based studies by human–environment geographers have also helped
to make clear the importance of taking into account both social and environmental difference in analyses of the environment–development nexus. Regarding social difference, research has “opened up” the household and the community in order to better understand the role of gender, class, ethnicity, race, and other forms of social differentiation in shaping systems of resource access and environmental decision-making, including in the context of development interventions. A valuable example of scholarship in this vein is Carney (2004) on the gendered impacts and dimensions of wetland development – namely, irrigated rice schemes and horticulture projects – in Gambia. Carney documents how these development programs have led to repeated gender conflicts due to the fact that they have increased women’s labor burdens, while at the same time weakening women’s customary rights to rural lands and resources. Meanwhile, scholarship has also broadened the analytical focus to examine not only how social difference shapes environment and development trajectories, but also how sociocultural identities are themselves constructed through initiatives of resource development and environmental management. This work has provided valuable and nuanced insights into the lively interrelations between environment, development, and identity.

Just as the category of “society” has been unpacked within the environment and development literature, so too has that of “environment.” In particular, geographical scholarship has worked to elucidate the complex ways in which the biophysical qualities of particular resources or environments shape development processes. Efforts to grasp how nature’s heterogeneity influences socioeconomic change have a long history. The Marxist theoretician Kautsky, for instance, in his turn-of-the-twentieth-century writings on agrarian change in Germany, identified a number of ways in which the variability of the natural environment shaped the form and trajectory of capitalist development in agriculture. In human–environment geography, the long shadow cast by early twentieth-century environmental determinism led many researchers to abstain from affording nature an active role in development, instead focusing on the often-dramatic ways in which human action – including via development interventions – transforms the environment, in the process “producing” new natures. Meanwhile, theories giving analytical priority to physical geography features in shaping development outcomes have often been roundly critiqued. Recent years, however, have seen an expansion of efforts to understand and theorize the recursive relationship between environmental conditions (understood as heterogeneous and dynamic) and socioeconomic change. The result has been a greater appreciation of both nature and the technologies through which nature enters into social life as actors – or “actants” in the language of actor-network theory – in development.

Useful examples of research in this vein come from recent scholarship on “development waterscapes.” Sultana (2013), for instance, explores the water-technology-development nexus in her examination of efforts to develop safe drinking water in the Bengal Delta. Sultana’s analysis highlights the active role of specific forms of nature and technology in development processes. In particular, she highlights the ways in which the qualities of the region’s groundwater resources (including the widespread presence of arsenic in these waters), as well as the characteristics of the technologies that have been promoted to extract groundwater for human consumption (namely, tubewells) have actively contributed to the creation of a new set of “technonatural assemblages” that both enable and challenge development.
Conclusion

Development has long been a central concern for human–environment geographers. This is especially the case for political ecologists for whom issues of resource use and environmental management in rural areas of the Global South have been a traditional focus. In recent years, however, the field of political ecology has undergone a process of change, as researchers deploy novel theoretical frameworks, expand into previously unexplored thematic areas, and broaden the geographical scope of their investigations, for instance, into cities of the Global North. Nevertheless, despite this process of scholarly variegation, a strong interest in the environment–development nexus remains. Looking forward, it is possible to identify topics that are likely to garner increasing attention from geographers interested in the dynamic relationship between socioeconomic change and the nonhuman, biophysical world.

One is the expanding role of private-sector actors in the development enterprise. As the social and environmental practices of transnational firms, particularly those operating in poorer areas of the Global South, have come under increasing scrutiny and criticism, one response has been for these corporations to become more directly involved in social and economic development efforts, typically as part of their so-called corporate social responsibility agendas. This is especially the case for resource companies, such as mining and hydrocarbon firms, many of which now carry out development programs in their areas of operation, at times in partnership with state agencies and nongovernmental organizations. While this trend toward corporate-led development has been discussed in the academic and policy literatures, its ramifications for environment and development trajectories remain underexplored.

Another topic likely to attract increasing attention in upcoming years is “degrowth.” As economic growth is increasingly recognized to be environmentally unsustainable (given the close and difficult-to-sunder linkages between economic production and energy and raw materials usage), there has been a burgeoning of interest in sustainable degrowth, a concept that typically is used to refer to the intentional scaling-down of production and consumption in ways that enhance human wellbeing and environmental conditions. Meanwhile, numerous areas of the world have experienced imposed degrowth in the context of urban and regional economic decline— the city of Detroit being a leading example. While a variety of nongrowth-oriented economic paradigms have been developed in recent decades (e.g., the steady-state economy model of Herman Daly and others), the crisis of climate change, coupled with the ongoing economic crisis, has, for many, made the need to transition to such a paradigm all the more urgent. Human–environment geographers, with their extensive tradition of research on the environment–development nexus, are well placed to contribute to studies of degrowth— immanent and intentional— and its, no doubt, complex socionatural ramifications.

SEE ALSO: Development; Livelihoods; Natural resources; Natural resources development; Nature conservation; Resources and development; Sustainable development

References

Bridge, Gavin. 2008. “Global Production Networks and the Extractive Sector: Governing


Environment and disease

James D. Tamerius
University of Iowa, USA

Throughout human history, humans have attempted to mitigate the detrimental effects of environment on health. We have constructed shelters, developed farming practices, domesticated animals, and relocated to more amenable sites. These developments have – in part – led to decreasing mortality rates and have enabled global populations to surpass 7 billion. However, increases in population size, animal husbandry and farming practices, land-use change, and industrialization have generated new environmental sources of disease. Indeed, environment remains a large factor in human health, particularly among low-income populations in developing countries.

“Environment” is a term used to describe the objects, soil, air, and water that surround humans and their physical domain. Environmental factors can cause disease either directly or indirectly. An example of a direct effect occurs when a human is exposed to extreme temperatures and cannot maintain normal body temperature (hyperthermia or hypothermia). The environment can also modulate disease indirectly through interactions with a transmissible disease system. For instance, surface water and suitable climatological factors can enable the maintenance of biological vectors, such as mosquitoes, that transmit pathogens to humans.

Disease risk varies strongly across space and time, partly due to the variability of natural environmental factors. Local hydrological, biological, climatological, and geological characteristics modulate exposures to various physical, chemical, and biological elements that cause disease. The vulnerability of a population to disease is also a critical factor determining the presence of a disease in a population and its burden on a population. The vulnerability of a population is closely tied to social, economic, and political power. Indeed, the social movement and the field of study called “environmental justice” emerged as research demonstrated that poor and minority communities in the United States were disproportionately affected by environmental disease. This trend is observed worldwide where low-income communities in developing countries are particularly vulnerable to diseases mediated by the environment as a result of inadequate infrastructure, increased susceptibility to disease, and limited access to proper health care.

Here we discuss the environmentally mediated systems and processes that disproportionately contribute to the global burden of disease: air pollution, diarrheal disease, and vector-borne disease. We also discuss other topics such as heat waves and emerging diseases because of their strong associations with environment and environmental change. Finally, we describe the expected effects of global climate change on disease.

Air pollution

Air pollution refers to the presence of any physical, biological, or chemical agent that contaminates either the outdoor or the indoor atmosphere. Air pollution can be generated from anthropogenic or natural processes. Major air
pollutants include microscopic particulate matter (PM), ozone ($O_3$), nitrogen dioxide ($NO_2$), sulfur dioxide ($SO_2$), and volatile organic compounds (VOCs). Currently, tropospheric $O_3$ and PM are considered to be the most important pollutants in terms of health impacts worldwide (Haerens 2011). Exposure to air pollution has been tied to the development of chronic obstructive pulmonary disorder (COPD), asthma, lung cancer, cardiovascular disease, and severe lower respiratory infections. Indeed, air pollution is related to more than 3 million deaths worldwide every year and accounts for 3.2% of the total global burden of disease (WHO 2013).

Historically, the primary anthropogenic source of air pollution has been smoke from fires used for cooking and heating. This trend has continued in modern times where approximately 3 billion people use solid biomass fuels (wood, animal dung, and crop waste) for heating and cooking. The concentration of PM$_{10}$ (PM up to 10 $\mu m$ in size) in homes using biomass fuels can be 10–50 times greater than concentrations considered safe. The inhalation of these soot particles makes individuals, particularly children under the age of 5 years, more susceptible to severe lower respiratory infections caused by bacteria and viruses such as influenza, respiratory syncytial virus, pertussis, mycoplasma pneumoniae and human metapneumovirus. Exposure to indoor PM from stoves can also increase the odds of developing COPD. The health burden of indoor PM is greatest among women and children due to the large amount of time they spend inside near the stoves. Indoor air pollution is not as problematic in developed countries where cleaner fuels and improved ventilation limit exposure.

Outdoor pollution accounts for about one-third of the disease burden of air pollution. Industrialization and the burning of fossil fuels dramatically increased the scale and geographic scope of air pollution beginning in the eighteenth century. As the negative effects of air pollution have become more apparent and new technologies have emerged, the most prosperous societies have developed and implemented air pollution mitigation strategies. As a result, emissions of major air pollutants have been declining in most countries in North America and western Europe since the 1970s via controls on emission sources, the use of cleaner fuels, and closures of outdated factories. However, outdoor air pollution remains particularly problematic in rapidly developing countries where energy consumption and the number of motor vehicles (especially unregulated two-wheelers) are increasing. Indeed, outdoor PM concentration in many urban areas of the developing world often exceeds three times the concentration considered safe. In many of these countries, measures aimed at reducing air pollution are not instituted because of their costs.

Local environmental factors can exacerbate the effects of air pollution. For instance, photochemical smog can become especially severe in areas that are prone to temperature inversions (atmospheric phenomenon where cool air is trapped at the surface below warmer air), which limit vertical mixing in the atmosphere and concentrate pollutions near the surface. Valley regions are particularly susceptible to this phenomenon because surrounding mountains can also limit horizontal movement of polluted air. Indeed, photochemical smog is particularly problematic in Los Angeles, Beijing, and Mexico City due to high emissions and favorable atmospheric and terrain factors. Further, SO$_2$ can interact with water droplets in humid environments and generate “classical” smog which caused the “Great Smog” event during the 1950s in London. This event was particularly severe because of a temperature inversion that trapped pollutants near the surface, and because it was
associated with cold weather that prompted people to heat their homes (wood fires), which intensified emissions (Fenger 2009).

**Water pollution**

Water pollution is the contamination of natural water systems by chemical, physical, or biological agents. Water quality is affected by both natural and anthropogenic sources. The most important sources of water pollutants include agricultural and mining activities, landfills, and industrial and urban wastewater, as well as natural geologic emissions. In addition to pathogenic microorganisms that cause disease (see diarrheal diseases, below), contaminants such as heavy metals (e.g., lead, mercury, arsenic) can cause cancer and brain damage and affect child development.

In developed countries, water supplied to the public is treated by means of filtration and disinfection processes. These treatment processes eliminate a vast majority of contaminants, but some pollutants, such as arsenic, lead, and microorganisms, may remain in the water. In developing countries, untreated waste water is often discharged directly into surface waters. Furthermore, natural sources of arsenic in some regions, such as Bangladesh and West Bengal, have contaminated drinking wells and caused arsenic poisoning in a significant proportion of these populations. Ironically, the drinking wells in these regions were created to mitigate the effects of contaminated surface water.

Heavy metals from industrial processes can also accumulate in surface waters and become concentrated in fish and shellfish. Humans can become exposed to the metals if they rely on these fish and shellfish for their diet, and be afflicted by cancer and birth defects. Further, fertilizers can contaminate freshwater and marine ecosystems, and lead to algal blooms. Algal blooms can produce toxins that are harmful to humans, thereby making water unsuitable for drinking. Algal blooms can also harm fresh and saltwater fisheries and result in malnutrition for people that depend on local fish for their diet.

**Diarrheal diseases**

Diarrheal diseases are primarily caused by bacteria, virus, protozoa, and parasitic worms. The most common agents of diarrheal disease are rotavirus and *Escherichia coli*. Diarrheal diseases are the second leading cause of death in children under 5 worldwide, and account for 1.5 million deaths annually (WHO 2012). Like most environmental diseases, the burden of diarrheal diseases exists almost entirely in developing countries due to poor infrastructure and increased susceptibility resulting from malnutrition, and existing and prior infections. Although diarrheal disease is common in developed countries, deaths from such diseases are rare.

People typically become exposed to diarrheal pathogens from contaminated water, although transmission pathways are multifaceted and complex. Following the excretion of human and animal waste, diarrheal pathogens can be transmitted to a new host through several environmentally mediated pathways. Diarrheal pathogens frequently contaminate surface and groundwater via open defecation and faulty or inadequate sewage systems. People in low-income areas that do not have access to treated water can become exposed to the pathogens by drinking contaminated water from rivers, streams, lakes, and wells. Diarrheal pathogens can also be transmitted when food becomes contaminated while it is being grown (contaminated irrigation water), shipped, handled, or prepared. Although boiling water is an effective measure for sterilizing water, the costs for fuel can be prohibitive.
in low-income populations. Washing hands with soap is also an effective measure to prevent the spread of diarrheal diseases, but access to handwashing facilities can be limited. Bathing in contaminated waters such as rivers and lakes is common practice in many developing countries and contributes to pathogen exposure.

Environmental factors are involved in the transportation of diarrheal pathogens, their survival in the environment, and human behaviors. For example, incidence of cholera in coastal areas is strongly associated with sea surface temperature, sea surface height, water nutrient levels, and solar radiation, which act to amplify the concentration of the reservoir (e.g., zooplankton) and the pathogen in coastal waters, and transport them into human environments. People may also respond to their environment by modifying their sources of drinking water, which can cause incidence of disease to vary by time. For example, fresh rainwater may be used for drinking water during the rainy season, whereas people may resort to drinking stagnant and contaminated surface waters during dry seasons. Temperature and humidity are also known to modulate the survival of diarrheal pathogens on surfaces and in water, which potentially modulate their transmission between hosts.

Vector-borne diseases

Vector-borne diseases are infections that are transmitted to humans by infected arthropod species. The most common vectors include mosquitoes, flies, ticks, mites, fleas, and lice. Major vector-borne diseases worldwide include malaria, dengue, yellow fever, and Japanese encephalitis. Malaria is by far the most significant cause of disease and mortality in humans worldwide and contributes to 1 million deaths per year (Murray et al. 2012). Vector-borne diseases primarily burden developing countries in tropical environments, although some vector-borne diseases such as West Nile virus and Lyme disease are prevalent in developed temperate countries.

Environmental factors can modify vector-borne disease systems by affecting the vector, the host (e.g., humans), or the pathogen. For example, reproduction in arthropods is temperature dependent and, despite favorable summertime conditions in temperate regions of the world, cold winter temperatures can interrupt disease cycles by terminating the reproductive cycle of vectors, such as mosquitoes. For this reason, most vector-borne diseases spread by mosquitoes are most prevalent in tropical regions that remain warm and humid year-round. On the other hand, diseases transmitted by ticks can remain prevalent in temperate regions despite harsh winter conditions because they can transmit pathogens, such as viruses, into their eggs. The eggs can remain viable over winter and then hatch when temperatures increase, thereby continuing the disease cycle. Further, some pathogens have disease reservoirs such as birds that can winter in warm locations where the disease cycle can continue uninterrupted. During the spring, the birds harboring the pathogen (e.g., West Nile virus) can return and initiate the disease cycle in a temperate region.

Environmental control is one of the primary methods to control and eradicate vector-borne diseases, in particular, permanently modifying environmental characteristics that favor the propagation of vector populations. Vector control strategies will vary depending on characteristics of the vector and the local environment, and include improved drainage, leveling of land, filling depressions, lining canals, adding gutters and roof drains onto structures, implementing wet-dry irrigation strategies, and use of pesticides.
Emerging diseases

An emerging disease is a disease that has recently appeared in a human population, or a disease that is rapidly increasing in incidence or geographic range. Recent emerging diseases include HIV (human immunodeficiency virus), Ebola, the Nipah virus, West Nile virus, SARS (severe acute respiratory syndrome), and influenza A/H1N1pdm. Most emerging diseases are zoonotic agents that are transmitted to humans through close contact with another species.

Increased disturbance of natural environments by humans has been accelerating the rate of emergence of new infectious diseases (McMichael et al. 2004). There are several human-induced environmental changes that facilitate the emergence of novel infectious agents, such as tropical deforestation, building of roads, irrigation, dam building, crop and animal production, urban sprawl, poor sanitation, and polluted coastal areas. For example, deforestation can result in habitat fragmentation and increase “edge effects,” which promote greater interaction between hosts, reservoirs of novel pathogens, and disease vectors. Natural variability of the environment can also facilitate the emergence of novel agents. For instance, in the early 1990s increased winter precipitation in the south-western United States associated with an El Niño event generated an abundance of food and led to a 10–15-fold increase in deer mouse populations, which carried a unusually virulent strain of hantavirus. As a result of the deer mouse population increase, contact between humans and deer mice intensified, leading to an outbreak of hantavirus in humans.

The emergence of novel pathogens is often complex, with both natural and anthropogenic environmental changes interacting to generate disease. For instance, in the late 1990s in Malaysia, slash-and-burn deforestation caused a thick haze to form over the region. The fires and the haze were intensified by drought conditions related to El Niño. The haze limited photosynthesis in fruiting and flowering trees (food sources) and caused fruit bats to seek out sources of food in human environments. Ultimately, this sequence of events enabled the transmission of the Nipah virus into domestic pigs and, subsequently, into the human populations (Bing et al. 2002).

Heat waves

Although many environmental disease systems are highly complex and include interactions between numerous biological and environmental agents, the environment can also act directly on humans and generate disease. Heat waves are characterized by prolonged periods of elevated temperature and humidity and are often associated with increases in morbidity and mortality. Exposure to high levels of heat and humidity limit the body’s ability to dissipate heat and can result in elevated body temperature and heat stroke. In addition, heat waves are associated with increases in respiratory and cardiovascular events. Indeed, in the summer of 2003 a heat wave in western Europe caused as many as 70,000 deaths (Robine et al. 2008).

The severity of a heat wave is dependent on many environmental and social factors. The increasing proportion of elderly populations in developed countries has increased the vulnerability of these countries to heat waves given the susceptibility of the elderly to heat (Sheridan et al. 2012). Increased temperatures due to the “urban heat island” effect place urban communities at greater risk for the effects of extreme heat. The rapid urbanization that is occurring in developing countries will likely increase vulnerability to heat waves in the coming decades.
Further, global climate change is expected to increase the severity and frequency of heat waves and overall temperatures. However, there are several behavioral, technological, and societal strategies to mitigate the effects of extreme heat. For example, regions characterized by hotter and more humid environments typically have higher tolerances for extreme heat conditions than do cooler areas. This suggests that humans can adapt to increasing temperatures. However, it is unclear whether human adaptation can keep pace with rising temperatures associated with global climate change.

Climate change

Increases in the concentration of greenhouse gases caused by anthropogenic emission has altered the radiative balance of the Earth. In response, global temperatures have been increasing since the early twentieth century and are expected to continue to increase. In addition to increasing temperatures, climate change is associated with changes in atmospheric moisture content, precipitation, and the frequency and intensity of storms, droughts, floods, and heat waves. Accordingly, climate change will have significant impacts on environmental factors and processes that mediate disease.

Direct effects of climate change will include an increase in the frequency and intensity of heat waves, which can cause a significant number of deaths and create health emergencies in both developing and developed countries (see heat waves, above). Increasing temperatures and humidity levels will encourage a geographic shift in the distribution of vector species, such as ticks and mosquitoes, to higher latitudes and elevations. Further, existing drainage and storm water systems in some regions may not be sufficient to manage the influx of water from extreme rainfall events. These surface waters may promote the growth of vector species and lead to increased incidence of vector-borne disease. Likewise, these events could increase exposure to diarrheal pathogens if wastewater treatment plants become overrun and increase opportunities for exposure in undrained environments. On the other hand, some areas will become more prone to droughts, which will increase the scarcity of fresh water and force people to utilize contaminated water sources. Agricultural production will also be negatively affected in some regions and could cause malnutrition and increase susceptibility to disease.

There is significant confidence that climate change will modify the incidence and geographic distribution of diseases. Our ability to predict the effects of climate change on disease is limited by our inadequate understanding of disease systems, and the variability of social, economic, and political factors which modulate population vulnerability to disease. Indeed, the effects of climate change on disease will not be uniformly distributed across space, and will likely disproportionately impact low-income communities in developing countries.

Conclusions

Environmental factors and processes significantly contribute to human disease. These diseases are typically generated through a combination of local environmental factors and social vulnerability. Thus, we observe significant variability over space and time with regards to the type and intensity of diseases mediated by the environment. As human populations have increased in size and density, and new technologies and lifestyles have emerged, we have changed the way we interact with the environment. Most of these changes have been beneficial to health, although others have created new sources of disease.
The disparity in social, political, and economic development contributes significantly to the variability of disease burden that is observed worldwide. Environmental change and new demographic profiles will continue to generate new sources of disease.

SEE ALSO: Climate change and health; Environment and health; Environmental (in)justice; Health geography

References


Further reading


Environment and everyday life

Alex Loftus
King's College London, UK

Intellectual and social context

The relationship between the environment and everyday life has not received the attention it deserves within either environmental scholarship or activism. In recent years, nevertheless, a critique of everyday life, a theorization of its conditions of possibility, and a growing recognition of the empirical complexity to researching everyday life has begun to emerge. In part, these three concerns reflect the influence of the environmental justice movement on dominant environmentalisms as well on geographical perspectives on the environment. In part, they also encompass a view that mainstream environmental movements have made catastrophic mistakes: in contrast rich possibilities might be found within a critique of everyday life.

Major dimensions

Environmental justice

As indicated above, the recognition that everyday life is crucial to the practice of an environmental politics comes through most clearly within the environmental justice (EJ) movement. For Whitehead (2009, 664) the EJ movement “reflects the apotheosis of second-wave environmentalism, combining as it does a concern with questions of nonhuman environmental justice alongside a strict focus on the everyday spaces of city life.” Here the environment has been understood as that which is lived and practiced, that which is breathed, cycled over, and played upon. Rather than focusing their energies on the preservation of green spaces for the benefit of a narrow elite, activists have demonstrated the ways in which environmental injustices are produced and reproduced at an intimate, everyday, and bodily level through the perpetuation of structural inequalities. Thus, for Pulido (2000) the ongoing reproduction of hegemonic white privilege within the United States ensures the continued exposure of raced communities to toxic waste. Pulido explores the everyday aspects of such environmental racism through a focus on the geographies of Los Angeles.

The roots to the environmental justice movement are often traced back to antiracist and working class struggles within North America. Perhaps the two sets of events most readily cited in relation to the genesis of the movement are the struggle against toxic waste in Warren County (Bullard 1990) and the struggle for relocation among homeowners sited on the toxic dump known as Love Canal in New York state. When told solely from these two North American narratives, accounts tend to overlook the importance of environmental justice as a force within locations across the Global South. In post-apartheid South Africa, however, the environmental justice movement has been one of the most important sources for alternative political imaginaries over the last twenty or so years. For McDonald (1998, 2004) EJ activists...
were able to effectively position the delivery of basic services as the central post-apartheid environmental question. In one of the most profound reinterpretations of environmentalism, many South Africans began to view the environment less as the National Parks from which so many were excluded or displaced under apartheid but rather as the sewage being handled by the municipal worker, the garbage being reclaimed by men, women, and children from the municipal dump, and the contaminated air being breathed by the asthma sufferer in a landscape liberated from, but now reinscribed by, the injustices of apartheid. Myers has noted that an overemphasis on the successes of the South African EJ movement can lead to a neglect of the unevenness of a movement which is still notable for its absence in many sub-Saharan African countries, nevertheless EJ has had an active presence within many Latin American and Asian struggles over the last three decades.

**Against the abstractions of the global**

Environmental justice activists are not the only ones to challenge the more abstract, global concerns of mainstream environmentalism. Illich (1973), Schumacher (1973), and Mander and Goldsmith (1996) have all in very different ways made a case for the importance of the local, very often positioning this against a universalizing and abstract environmental politics. Small is Beautiful, Schumacher’s paean to “enoughness” in the face of claims that “bigger is better” has been seen as the basis for an understanding of “intermediate” or appropriate technologies. The development interventions of NGOs such as Practical Action have sought to build on such a philosophy and construct a model of sustainable development informed by existing practices of everyday life in historically and geographically specific contexts. Local Agenda 21, one of the key policy innovations to emerge from Our Common Future and the subsequent Rio Earth Summit provided a more formal recognition that achieving global environmental sustainability would require shifts at a local scale, and that many of the key actors in environmental sustainability had not been adequately recognized. Through Local Agenda 21 the abstractions of the global became something experienced at a far more humble level. Environmental interventions needed to ensure improvements in the spaces that people cycled or walked. The resurrection of a canal towpath thus came to represent the movement for sustainable development at a far more mundane, everyday level. Nevertheless, even if such interventions are experienced most directly at the level of the everyday and are informed by a more nuanced scalar politics than prior manifestations of environmentalism, the everyday is rarely discussed explicitly within such work. Instead, local institutions, civil society, communities, and small-scale entrepreneurs become the key actors within what can sometimes become an atavistic and romanticized counter to the modern abstractions of the global.

**Between the marvelous and the mundane**

Elsewhere, geographers have borrowed from both actor network theory and science studies to better understand the more mundane – but no less important – politics of urban wildlife. Seeking to attend more closely to the intermingling of human and nonhuman within urban environments, Hinchliffe et al. (2005) have emphasized what Stengers refers to as cosmopolitical experiments. Practices that are “unnoticed by urban politics and disregarded by science” become central to an analysis that
is more sensitive to the myriad ways in which urban ecological assemblages come to be produced within everyday encounters. Others, such as Michael (2006) have stressed the relationship between technoscience and everyday life and sought to think through an environmental politics that is more explicitly concerned with such interrelationships.

Less theorized accounts of environmentalism and the everyday are abundant – one needs to make only the briefest of searches on YouTube – but these are often premised on a simplistic reading of the idea that change needs to come from the individual, an unproblematic reading of the subject and a complete divorce from the broader contextual factors that provide opportunities and obstacles to an environmental politics in differing historical and geographical contexts. In the effort to shift attention from spectacular manifestations of environmentalism, the result is often a bland appeal to the micro and an individualistic claim that the acts of many single actors will add up to the change necessary for an environmental revolution.

These more naive claims to the everyday, or to a politics of the grassroots, have, not surprisingly, attracted criticism from a range of different quarters. While deeply sympathetic to many aspects of environmental justice activism, and seeking to learn from the organizational success of movements to organize across raced, classed, and gendered constructions, David Harvey (1996) has questioned the ability of the environmental justice movement to move beyond a militant particularism in order to achieve the “global ambitions” that it often seems to set itself. Elsewhere, Harvey (2008, 29) writes that “Paul Hawken … in ‘Blessed Unrest’ makes it seem as if social change in our times can only emanate from the practical engagements of millions of people seeking to transform their daily lives in creative ways (a position that is sometimes taken in the works of Gibson-Graham and Escobar).” Instead, Harvey calls for a dialectical approach that is attendant to how social change works across and through several different moments simultaneously: everyday life is one of these moments alongside relations to technology, relations to nature, social relations, and relations of production and ideas.

Whitehead (2009) considers these questions in more explicit reference to the manifestations of environmentalism in the contemporary moment. Although heavily influenced by the environmental justice movement, Whitehead notes his disappointment at the inability of the environmental justice movement to live up to its own claims to be shifting the practice of environmentalism to a more intimate sphere. Instead, as he notes, the focus has often been on the spectacular manifestations of environmental injustice, the most extreme cases of poisoning from toxic waste and the most devastating health consequences of environmental racism. In becoming increasingly preoccupied with such catastrophic events EJ becomes increasingly divorced from the very intimacies from which it claims to emerge. Paradoxically, it speaks less to those affected on a day-to-day basis by ongoing environmental injustice and environmental racism. Whitehead’s response is to develop a far more rigorous “critique of everyday life.” In an early take on such everyday environmentalisms, he explores the meaning of such a politics in the post-communist city of Katowice (Whitehead 2005). Claiming that environmentalism often finds itself caught “between the marvelous and the mundane” or rather between spectacular claims around a looming apocalypse – often used as the basis for large scale environmental summits (Death 2008) – and a more banal claim that everyday acts, such as turning off the light switch will make all the difference, Whitehead seeks to develop an alternative basis for an everyday environmentalism. Such a critique requires a
thoroughgoing analysis of the conditions of everyday life and a critique that emerges immanently from the processes, relationships, and paradoxes through which that everyday life is practiced. Thus, Whitehead (2005, 2009) turns to a more thorough reading of Henri Lefebvre, whose three-volume *Critique of Everyday Life* provides one of the most far-reaching theorizations of the conditions of possibility immanent to a practice of everyday life. In a more recent paper, he sees such an immanent critique of everyday life emerging in disadvantaged groups’ engagements with the Black Country Urban Forest in the West Midlands of the United Kingdom.

The resources to be put to work from Lefebvre’s three-volume *Critique of Everyday Life* are numerous. Outside of these three volumes, Lefebvre made an effort to distinguish between some of the elements of his critique: “daily life” (*la vie quotidienne*), everydayness (*la quotidienneté*), and the everyday (*le quotidien*). Whereas daily life has always existed as the rhythm that structures humans’ existence (permeated with historically and geographically specific cultures, myths, and values and deeply rooted in connections with nature), the everyday marks the movement of daily life into the era of modernity (Lefebvre 1988, 89). Everydayness becomes the bland homogenizing effect of this entry and the disciplining of a variety of patterns into a more mundane repetition. The everyday, however, remains open as a space from which critique might be possible. Thus, for John Roberts (2006, 67) “if everydayness designates the homogeneity and repetitiveness of daily life, the ‘everyday’ represents the space and agency of its transformation and critique.” One might think of the disciplining effects of the alarm clock or the daily commute and yet one might also think of the situated knowledges emerging as an emancipatory critique within a worker’s understanding of her labor process. This crucial category of mediation permits a way out of the trap that Whitehead identifies in which environmentalism is trapped between the marvelous and the mundane. A radical critique becomes possible within the homogenizing of everyday practices.

**Everyday environmentalism**

Lefebvre is one of several theorists who figure prominently in Loftus’s (2012) *Everyday Environmentalism: Creating an Urban Political Ecology*. Loftus begins with a recent perspective on the nature of the environment. Eschewing the kind of dualistic understandings that have been propagated within mainstream environmentalism, Loftus seeks to develop an environmental politics as it emerges from the practices of everyday life. Criticizing mainstream environmentalism for the disempowering way in which apocalyptic visions are so frequently precursors to the ceding of power to technocratic elites, Loftus seeks to grasp the fragmented and incoherent environmental knowledges that emerge from historically and geographically specific practices. Although it is now commonplace within academic writings to reject dualisms, the implications for a critique of everyday life have not been explored. However, if the practical activity of everyday life is seen to be woven out of and weaving together particular socio-natural constellations, conditions of possibility begin to emerge not only for a fundamentally different environmentalism, but also for a broader transformative project.

Loftus’s empirical examples are drawn from very different contexts. First, he draws on the everyday practices of provisioning a household with water in post-apartheid informal settlements and formal townships. These acts make and remake what many have referred to as the “waterscape” and within them the conditions of possibility are forged for a critique of both the
practices and the produced environment – or waterscape – itself. Thus, a rapprochement becomes possible between the kinds of scholarly critique developed within urban political ecology and the wellsprings of anger that have been tapped when waters run dry, when environmental injustices are reproduced in the post-apartheid period and when everyday access to the means of existence are frustrated by the mediation of the exchange abstraction. Second, Loftus draws from the practices of artists in London as the socio-natural fabric of the city becomes the material out of which alternative urban futures come to be forged. In both London and South Africa, everyday environmental praxis comes together to provide conditions in which new worldviews are incubated and also the materials and practices out of which new environments might be forged.

The theoretical resources that might be drawn upon within such a praxis-based understanding are many, ranging from geographical thinkers such as Harvey and Smith to feminist writers on situated knowledges. What unites these thinkers is a recognition that the quotidian acts of producing and reproducing the world contain with them conditions of possibility for thinking about that world differently. Ontology and epistemology are thus seen to be related, just as theory and practice come to be identified in a mutually symbiotic manner.

Future directions

If there has been an increasingly explicit discussion of the ways in which everyday life matters for an environmental politics – building on what had been an implicit recognition previously – there is still much work to be done on how this might be put to work in generating new understandings of the possibility for more just and democratic socio-ecological relationships. First, more empirical work is needed to analyze the relationship between historically and geographically specific acts of making distinctive environments and the forms of environmental knowledge implicit within these acts. Second, the relationship of such research with actually existing environmental movements needs to be considered in far more depth. Distinguishing everyday environmentalisms from more mainstream environmentalisms, Cooper (forthcoming) suggests further research on nonenvironmental environmentalism. Others have pointed to the need for a more explicit recognition of the identities that emerge within such everyday environmentalisms. For those seeking a more explicitly normative agenda, there is the possibility of considering how such environmentalisms might be shared among different communities and the ways in which they might provide a basis for future possibilities. Overall, a new field is being defined and the potentials for a new environmental politics to emerge on the basis of such discussions are great.

SEE ALSO: Environmental (in)justice; Production of nature

References


Cooper, T. forthcoming. “‘Everyday Environmentalism’ in Historical Context: The Case of Waste in Twentieth-Century Britain.” *ACME*.


ENVIRONMENT AND EVERYDAY LIFE


Environment and gender

Andrea J. Nightingale  
*Swedish University for Agricultural Sciences, Sweden*  
*Norwegian University of Life Sciences, Norway*

For some readers the topic “environment and gender” invokes women’s roles in and knowledge of environmental issues. Yet scholars have long argued that the importance of gender in relation to environment extends far beyond women or even gender relations. Like the social sciences more widely, over the past four decades there have been significant shifts in how environment and gender are conceptualized, both as separate analytical constructs and in relation to each other. Gender is only one dimension of power, but has served as a critical entry point for rethinking environment–society dynamics, or what is perhaps best thought of as “socio-natures.” Therefore, “gender” requires one to not only analyze relations based on biological sex within society, but also to analyze how such relations are constructed, contested, and internalized, and how these processes of power are co-emergent with ecologies and environments. Similarly, “environment” is not a taken-for-granted entity; rather, like gender, it is a socially constructed domain that is dynamic biophysically and also in terms of what we imagine it to be. When these insights are taken seriously, they transform how all scholars approach environment–society questions and provide potent analytical tools for understanding the operation of power in relation to environment.

In this entry, the evolution of thought on environment and gender is traced and this history is placed within debates about gender and environment as separate intellectual agendas in the social sciences. It is shown on the one hand that theoretical advances concerning environment and gender in geography have served to transform the wider debate, and on the other hand, at times, that geographers have been slow to incorporate analytical advances from other debates. Many of these conversations have taken place within feminist geography and feminist theory, but by no means are all contributions to the environment and gender debate self-styled as “feminist.” The entry tries to distinguish between feminist thinking and other work that attends to gender, acknowledging how important all of them have been to the overall understanding of environment and gender. The following key concepts are brought forward as being particularly important contributions to social science understandings of environment–society dynamics more generally: politics, intersectionality, and co-emergence.

Politics refers to contestations and collaboration between people and between institutions that serve to order and govern everyday affairs. In this sense, politics encompasses government, but also signals relations beyond governments that are instrumental in framing society–environment dynamics. Intersectionality is a term used to capture the operation of power and how it frames social difference (gender, race, ethnicity, caste, class, and disability, among other markers of difference). Intersectionality highlights the ways that different dimensions of identity come together to shape how social difference operates and forms everyday experience for both
individuals and groups. Co-emergence captures the way that societies and environments are intertwined and internally related. The boundaries between them are a product of how we think rather than the way the world is. Separating what is society and what is environment becomes nearly impossible once one focuses on how they are interrelated. The idea of relationality is core here. When phenomena or objects are relationally produced, it means that one does not exist without the other. Any change in one will result in a simultaneous change in the other. Further elaboration on all of these concepts follows below.

The entry begins with the history of environment and gender in the 1970s, with feminist critiques of Enlightenment thinking. In this work, feminists argue that the oppression of women is linked to the overexploitation of environment and as such is inherently a political question. The marginal position of women within patriarchal societies must be addressed in order to shift environmentally damaging practices. This analysis swiftly diverged into two related, but somewhat separate, strands of thought, both of which are alive and well today. First, ecofeminists argued that women have innate, privileged knowledge of environments, which inspires in them a desire to protect nature. Second, other scholars explored politics more explicitly and probed how the social construction of women and nature facilitates certain kinds of oppression of women and overexploitation of environments. In other words, the link between environment and gender is not fixed, but rather a constructed and contested relation. They criticized ecofeminists for assuming an essential link between women and nature, and instead focused on how, and with what consequences, women become associated with nature. The idea that links these two debates, however, is the notion that gender politics and environmental issues are not separate questions, but rather are inextricably linked. And, as the entry explores more fully below, taking this critique of the Enlightenment seriously requires a reformulation of the environment–society nexus itself.

In keeping with the focus on politics, and closely related to the critique of ecofeminism, a number of scholars in the 1980s and 1990s drew from historical materialist thinking to argue that women’s productive relations bring them into culturally, historically, and geographically specific relations with environments. These relations can give women important knowledge of environmental questions, but that knowledge arises from experience rather than innate inclinations. This work insists on a more nuanced understanding of social relations and how they link societies to environments through productive (and extractive) relations. It is closely related to debates in political economy that seek to show how class relations play out within environmental issues, a field now known as political ecology. Therefore, rather than overexploitation of environment being a necessary outcome of patriarchy, the critique here shows how the organization of society is linked (dialectically related in historical materialist terms) to the political economies of environmental use. Historical materialist work on gender was thus crucial in pointing to the multiscalar social relations that mediate the society–environment nexus and cautioning against assuming that class alone was adequate to explain how social inequalities arise in environmental management. The idea that gender, class, race, and ethnicity combine to shape productive relations was an important precursor to later work on intersectionality in relation to environment.

The debate was far from finished, however, and in the 1990s and 2000s a number of scholars drew from insights in feminist theory and the emerging intellectual agenda in geography
around social construction of environment to explore how gender and other social relations (intersectionality) are constituted in relation to environments. There are two contributions from this debate. First, gender needs to be understood as an outcome of the operation of power and thus a dynamic concept that serves to mark differences between men and women. Taking account of gender therefore requires scholars to look within and beyond household- and community-level relations, as well as probing how forms of social difference intersect to shape the ideas and practices that generate inequality. And perhaps most importantly in terms of a political ecology perspective, there are political economic consequences of gender in terms of accumulation, property, dispossession, and control over surplus labor, all of which are driven by, as well as being manifest upon, environmental conditions. Intersectional social differences, in other words, are not merely unfortunate by-products that need to be managed by special attention to disadvantaged people. Rather, power operates to create inequalities based upon gender and other social differences. These in turn result in unequal access to land and property, differences in wealth, and a greater ability by some people to “get ahead” (control their surplus labor). These kinds of inequality are often manifest on the landscape as degraded land (marginalized people forced to overexploit their resources), fertile land (wealthy people being able to invest in land improvements), protected areas (local people kept off the land to preserve it), and so forth. Below these ideas are taken up more fully.

Second, environment is similarly a dynamic and contested analytical construct. Within most of physical geography and some parts of human geography, “environment” refers to the natural or material world, assumed to be unproblematically distinguishable from “society,” and, perhaps most importantly, to be a real entity that pre-dates its social construction. Many human geographers, however, have argued that such a taken-for-granted understanding of “environment” masks two important processes. One, what counts as environment and how it is distinguished from society are political and, as such, historically and culturally contingent. Second, by assuming environment is a nonhuman domain, it closes down an analysis of processes and objects that are not clearly social or natural. Rather, this work argues that environments and gender (society) are relationally produced, or co-emergent.

Thus the environment and gender debate has come a long way from insisting on women’s role in environmental management. Now it captures a range of scholarship that seeks to hold both gender and environment as contested, political and unstable analytical constructs that are perhaps only partially useful as starting points for a socio-natural analysis of power, social inequality, and environmental change. The remainder of the entry elaborates on this history of “environment and gender,” and shows the key contributions to wider understandings of environment–society (socio–nature) that have emerged from a concern with gender, namely politics, intersectionality, and co-emergence.

Ecofeminism: linking social systems with human use of environments

Ecofeminists are in some sense the originators of the environment and gender debate, and while parts of their analysis have been largely discredited, they made important contributions that continue to be influential today. First, they placed a concern with how social relations extend to environmental domains firmly on the intellectual agenda. They argued that the logic which serves to oppress women is the same logic that leads to overexploitation of the environment, and this
logic is closely connected to science and capitalism. Second, they argued that women have different knowledge of environment from men, and we need to harness such knowledge to find less damaging society–environment interactions.

The basis for these two contributions is grounded in what has become an important feminist critique of Enlightenment and Cartesian thinking. The Enlightenment is associated with the rise of science and the analytical separation of environment from society. Cartesian thinking refers to work inspired by Descartes who reframed the environment as a machine, departing from his contemporary metaphysical counterparts who conceptualized the world as a living being. This mechanistic model paved the way for the scientific method, which prioritizes examining components of natural systems in isolation from other components in order to understand them. Feminists in the 1970s critiqued this dominant “scientific” paradigm on two grounds. First, Enlightenment thinking analytically separates environment from society, producing a host of other binary associations that seem to logically follow: rational–emotional, mind–body, men–women, heterosexual–homosexual, white–people of color, and culture–nature, to name some of the more persistent formulations. Second, feminists argued that Enlightenment thinking associated women with one half of these binaries: emotion, body, and nature. This second argument in many respects led to a core split within feminist thinking on environment: those who embraced such associations and sought to politically mobilize them and those who argued that binary thought needs to be rejected entirely. Ecofeminists, who embraced the idea that women are closer to nature, explored how women are more likely to protect the environment than men due to their role as mothers – a position popularized by Vandana Shiva (1988). They also advocated for a more connected or spiritual approach to environment–society dynamics in order to overcome the mechanistic model perpetuated through science (Diamond and Orenstein 1990).

Before moving on to critiques of ecofeminism, it is important to note the crucial contributions it has made, not only to environment and gender debates, but also to social science and geographical scholarship more widely. First, feminist thought continues to be strongly informed by a critique of Enlightenment and Cartesian thought. Many feminist geographers begin their analyses across a range of topics with the need to unpack binary distinctions, the most relevant here being the opposition of environment with society. A more elaborate discussion on the implications of this move follows below. Furthermore, feminists have been at the forefront in critiques of reductionism and the scientific method, highlighting the limitations of these approaches for understanding some of the most pressing environmental problems of our time. Following Merchant (1982), they have argued that “nature” cannot be reduced to a set of interlocking parts, but rather needs to be analyzed as a relationally constituted whole. This has fed into recent debates in science studies and nature–society geographies on how to theorize “environment–society” and how to develop new methodologies to capture their nexus. In this sense, contributions from environment and gender scholarship have been foundational in helping to steer the current direction of nature–society geographies.

Second, while most feminists are wary of essentialism, there continues to be a tradition of ecofeminist thinking that begins with an understanding that the oppression of humans and overexploitation of nonhumans are linked. Most recently, queer theorists have entered the debate and insisted on nature as more than a stage for the playing out of social relations. This
body of scholarship argues for a fundamental transformation in how we understand our relationship to the rest of the world socially and environmentally. Some in this tradition (re)turn to spirituality to find other, nonreductive and nonmechanistic paradigms for framing the place of humans in nature (Anderson et al. 2012). This strand of scholarship has perhaps fewer adherents within geography, but is closely related to deep ecology and also to the growing field of ecocriticism within the humanities.

**Historical materialism and the critique of essentialism**

Despite these important contributions from ecofeminism, the idea that women are closer to nature was criticized almost as soon as it was put forward. In particular, the reduction of all women to one inherent nature was deemed to be problematically essentialist. Essentialism refers to the idea that particular groups of people possess fundamental or essential characteristics. Attributing essential characteristics to a group or category of people ignores important differences between members of the same category, and as such can mask some of the dynamics that serve to keep some people “closer to nature” than others.

Feminists of color were at the forefront here, criticizing the white, middle-class bias in much of feminist thinking at the time, and demonstrating the importance of foregrounding differences between women (Mohanty, Russo, and Torres 1991). In addition to race and class, other forms of difference also serve to shape who has what experiences and what knowledges of environment, including sexuality, disability, and geographical location (i.e., north–south). Therefore the question is not whether women are closer to nature than men, but rather how some women (or men) come to be materially and symbolically associated with nature, and the implications of those associations socially, politically, economically, historically, and materially.

Speaking explicitly to the environment and gender debate, the Indian feminist Bina Agarwal (1992) drew from her background in political economy to argue that the focus needs to be on relations of production and reproduction that serve to bring (some) women closer to nature than others. Women who engage in agricultural labor, or are responsible for gathering essential natural resources (wood, water, etc.) to support household reproduction, do indeed gain crucial knowledge of environmental processes. They are more likely to be invested in long-term ecosystem health, but the crucial insight here is that a propensity to protect the environment arises from the material conditions of everyday life, rather than from innate tendencies. These insights have been taken up within development practice, which now targets women for various conservation schemes. So while, on the one hand, feminists celebrate how development practice has recognized the need to understand intra-household dynamics when designing development interventions, on the other hand, they are wary of how development targets women as a source of competent and willing labor for environmental management schemes.

Within academia, Agarwal’s work fed into what is now known as feminist political ecology, although at the time Agarwal named it “environmental feminism.” Feminist political ecology has perhaps been most productively developed within geography, as elaborated upon below. The historical materialist lens, Agarwal suggested, provided an important set of tools for unpacking the category “women” and recognizing how people with different class and associated productive relations necessarily come into different relations with environment.
ENVIRONMENT AND GENDER

This laid a foundation for thinking about intersectional social relations in feminist political ecology: although now, the structuralist framing of gender, class, and race in environmental feminisms has been replaced by a performative understanding of gender and social difference within “new” feminist political ecologies. And while Agarwal herself never took the analysis this far, her work begins to show how social difference is in part generated from production and thus is co-emergent with environment. This last point is elaborated on below in “the production of socio-natures” section.

Cyborgs and socio-nature: retheorizing society–nature

While the debates between ecofeminists and their critics were in full flow, Donna Haraway (1991) argued for a reconceptualization of both environment and gender. She drew from early feminist critiques of science to imagine a post-Enlightenment understanding of environment–gender. Her account tackled many dimensions of science and our conceptualizations of environment–gender at once. She has been profoundly important in changing the entire character of environment–society debates, most recently with her work on human–animal relations. The entry will attempt to summarize and make sense of her rich argument by framing it around two core contributions: (i) reimagining the entry points for probing environment–society dynamics, and (ii) reimagining how we can know and make claims from our scholarship.

Enlightenment thinking opened up methodological space for research techniques wherein detached observation and attempts to isolate parts of the whole were valued. This is recognizable to most readers as the “scientific method” wherein variables are isolated and tested one by one against hypotheses. Haraway rejected this kind of detached scholarship and reconceptualized environment–society by insisting on its social construction and relational emergence. She turned to metaphor and figuration to reimagine, in an “ironic and blasphemous manner,” the interconnections between conceptualizations, discursive constructions, political economies, and the materialities of different bodies and species. She used the figuration of the “cyborg” to review the boundaries of human/nonhuman and to playfully show the (re)production of those boundaries and what it means to be human. She argued that humans are relationally produced from the encounters we have on an everyday basis, and thus the very meaning of human (or the social) emerges from how we define ourselves against the nonhuman (or nature).

In subsequent work, Haraway developed this critique further to probe the boundaries of kinship, bringing back and also going beyond questions of gender. She explored how difference is constructed, not only between men and women (although this was a crucial starting point for her), but also between humans and other species. She argued that our scholarship – and the political economies of science and technologies that result – goes astray when we forget that our analytical constructs are not “life itself.” She used the example of genetics to show how “genes” are an abstraction of very complex biochemical reactions that are not fixed in time and space. They are “real” in the sense that we can create them through genetic technologies, we can even move them around across species, but “genes” are not life itself. They are one lens through which we try to understand life, and a lens that has very specific biological, political, economic, and societal consequences. Thus one of the core contributions to come from the environment and gender debate has been to show the interconnection
between how we understand “life,” and the lives that become possible as a result. Perhaps most importantly, these lives are necessarily already bound up in social-political as well as ecological dynamics that cannot be separated from our analytical constructs of “life itself.”

Furthermore, if we embrace such a relational emergence of environment and society, then we have to develop new methodologies to study these dynamics. Haraway made a profound methodological intervention by adding her voice to other feminist critiques of neutral observation. She demonstrated through the example of primate behavior research that the background of the researchers – particularly their gender – was instrumental in shaping how they interpreted the behavior they observed (Haraway 1991). Therefore, she argued that detached observation was not possible. The expectations, assumptions, and background, or “positionality,” of researchers shape how they design, analyze, and interpret their research. She rejected “the god trick,” her name for detached observation, as being impossible to achieve, and rather argued for a commitment to “situated knowledges.” Situated knowledges is an epistemological and political stance that demands one be accountable for “how one learns how to see,” arguing that the way one sees is at least in part derived from one’s positionality. She joined others in the social sciences at the time, including Foucault, in arguing that knowledge is contingent on the people and the conditions under which it is generated. Therefore, science needed to shift from attempting to reduce observer “bias” to creating objectivity by acknowledging the position from which we are able to see.

Haraway’s work, which began in the environment and gender debate (in her words, a socialist, materialist feminism), has provided a foundation for an entire generation of scholars to research the implications of environment and society conceptualizations. These contributions have emerged in conversation with parallel work in science and technology studies and geography to encompass a wide variety of perspectives trying to make sense of “socio-natures.” An elaboration on how this debate has been taken up by “new feminist political ecologists” follows.

**Feminist political ecology**

While Haraway’s ideas were almost immediately influential within feminist theory, some of the core dimensions of her argument were slow to be taken up by geographers working on environment and gender. Rather, most scholarship on environment and gender continued to (ontologically) treat the two as separate, interacting phenomena. The goal was to refine an understanding of how gender relations shape environmental issues. Within this work there were two parallel strands: the first was a merging of the ecofeminist debate with historical materialism to show how environmental questions are “gendered.” The second was to take the historical materialist debate further and think more carefully about the way that social relations have political economic effects which are both driven by, and manifest within, environmental governance contexts. There are not clear lines between these two strands, but the first has a more explicitly feminist agenda than the second. Linking them is a concern with how usufructuary rights and forms of property are mapped onto access to and control over resources.

Diane Rocheleau, Barbara Thomas-Sluyter, and Esther Wangari’s 1996 book, *Feminist Political Ecology* (1996), best exemplifies the first strand. They sought to push forward the environment and gender debate by arguing that environment is “gendered” in three key ways:
ENVIRONMENT AND GENDER

1 gendered knowledge, or how access to scientific and ecological knowledge is structured by gender;
2 gendered environmental rights and responsibilities, including differential access by men and women to various legal and de facto claims to land and resources; and
3 gendered politics and grassroots activism, including an examination of women’s roles within environmental movements.

The concern with gendered knowledge builds directly from historical materialist accounts to probe how men and women have differential knowledge of natural resources. Examples from development practice abound here. In development, men generally have access to agro-forestry extension workers, new training opportunities, and other knowledge associated with “science,” whereas women have knowledge gained from their activities as subsistence providers in households. By focusing on how people gain knowledge, this framing helps to avoid lumping all women together as having innate understandings of environment. It also brings into view the processes and contexts—here development interventions and training opportunities—through which such differentiation is perpetuated and reinforced.

The second theme, gendered rights and responsibilities, points to unequal access to land and resources by men versus women. While Rocheleau, Thomas-Slayter, and Wangari’s (1996) book emphasizes examples from developing countries, other feminist political ecologists reviewed how similar processes are at work in so-called developed countries as well (Seager 1993). The regularization of land tenure through the distribution of land titles most often assumes that male heads of households are the “rightful” owners of land, undermining women’s de facto claims to use-rights and control over land. In some contexts, development projects seeking to change land use intersect with these conflicts over land and resources, to the detriment of women. Researchers therefore need to probe how land use and rights intersect with gender relations to disadvantage some and privilege others.

The third theme, gendered politics and grassroots movements, examines social movements, highlighting how women have taken leadership roles and gained self-confidence within community struggles for control over natural resources. This interest in social movements resonates with the hope in the wider political ecology debate that environment could become a rallying point for transformational politics. Although certainly not without their contradictions, environmental social movements seem to have potential for the emancipation of women and impoverished communities, in addition to environmental protection (Peet and Watts 1996; for an alternative argument, see Reed 2000). On a conceptual level, environmental movements help to show how social difference is related to environmental knowledge and desires to protect nature. Social movements can help to provide people with new knowledge of nature, in addition to environmental issues acting as a focal point for collective action across forms of social difference including gender, race, and class.

The themes in Rocheleau, Thomas-Slayter, and Wangari’s (1996) book developed in conversation with a growing political ecology literature in geography that does not necessarily have an explicitly feminist agenda (Peet and Watts 1996). Here, concern with gender relations is an empirical question: rather than beginning with gender as a focus, this work looks into which social relations are most important in any given context. Many environmental conflicts that most intrigue political ecologists are strongly shaped by gender relations as well as class. Several studies explore the way that gender, as historically, culturally, and
geographically specific male–female sex roles, combines with class and race to structure access to particular types of knowledge, space, resources and social–political processes. Conceptually, this work contributes to debates on property rights and their relationships to social inequality and environmental outcomes. In particular, it probes how gender and class shape property relations and usufructuary rights to land and resources. Well-intentioned development projects that seek to mobilize women’s labor, or to make particular lands more productive through new agricultural technologies, often lead to dispossession of resources for women and marginalized community members. In other words, social relations of power cannot be undone by development interventions, but rather social relations powerfully influence control over, access to, and distribution of resources and the environmental consequences of these relations and development interventions.

Political ecology has been particularly important in showing that attention to gender should not be confined to the household and community level only, but rather also needs to be researched at the level of policy. For example, Schroeder’s (1999) work in the Gambia traces shifts in how development projects imagine the roles women should play and the ways these assumptions intersect with gender relations within households and communities. An inadequate understanding of local gender relations and how they structure control over land and resources and shape production often results in reversing gains made in earlier development interventions. An analysis of environment and gender therefore needs to account for not only how men and women are differentially engaged in environmental conflicts and development programs, but also how conceptions of gender frame the design of interventions. In other words, social relations are not separate from environmental governance, but are foundational to which schemes are put forward as important, how successful they may be, and how they may fuel existing conflicts.

It is important to note here that, within both feminist and wider political ecology, an analytical separation between environment and gender (or society) is retained, and this separation helps demonstrate how environment is a potent terrain within which struggles over access to, control over, distribution of, and knowledge of resources are played out. Today, many scholars continue to investigate these four core aspects of environmental conflict: access, control, distribution, and knowledge. A crucial contribution from the environment and gender debate has been to insist that each of these processes is structured by social relations within households, communities, and across scales to the global.

Intersectionality and new feminist political ecologies

The environment and gender debate was far from finished, however. A number of scholars were uncomfortable with the emphasis on women in feminist political ecology. This conflation of women and gender was not consistent with other developments in feminist theory. Rocheleau, Thomas-Slayter, and Wangari (1996) insisted that their interest was in gender as socially defined roles for men and women, but in many places the book slips back into an uncritical focus on women. In part this occurs because of their desire to spotlight the contributions of women who are often invisible within development projects and the global spheres wherein a book of that type circulates. They therefore attend to the knowledge, struggles, and activities of locally based, often illiterate, women. The danger with such slippage, however, is in reverting to an essentialized understanding of women as having
a privileged understanding of nature simply by virtue of their biological sex. In addition, the slippage closes down a careful analysis of how political, economic, cultural, and symbolic processes serve to *define* gender. But perhaps most significantly, conflating gender with women masks how gender is implicated in the construction of “environment” itself. In short, what is not sufficiently probed is how gender comes to be relevant in the first place.

Such an uncritical focus on women as a unified category was not consistent with concurrent advances in feminist theory. Chandra Mohanty, Judith Butler, and others argued that biological sex was an inadequate basis for understanding social difference (Butler 1990; Mohanty, Russo, and Torres 1991). Rather, social relations, first, “add up” to structure how individuals (in this case) have access to, control over, and knowledge of environments. Second, the production of social difference is performed in the everyday and as such is constantly up for negotiation and transformation. This formulation is often referred to as “intersectionality” to capture how class, race, ethnicity, gender, and other forms of social difference are not separable, but rather emerge together from the operation of power to shape which social relations are relevant and contested. Feminist geographers apply this understanding to space to show how space and social relations are mutually constituted within everyday activities (performances) (Massey 2005).

A new generation of feminist political ecologists took up these ideas to retheorize gender within the environment and gender debate. These “new feminist political ecologists” place more emphasis on intersectionality and processes of subjectivity in relation to environment than they do on women per se. Subjectivity comes from Foucault’s (1995) and Butler’s (1990) conceptualization of the operation of power. Subjectivity emerges from the exercise of power and the internalization of power by the subject, and as such is dynamic and continually (re)constituted. For example, people who are biologically female are often subjected as “women,” but what being a “woman” means and the extent to which it reflects different dimensions of biological sex is contextually contingent. At the same time, those subjected as women internalize as well as resist these notions of “women,” serving to transform (or cement) what it means to be “women” (Butler 1990). While these ideas are highly abstract, they have also been very important in moving beyond a focus on women within environment and gender debates. Attention to subjectivity demands an exploration of the processes, contexts, and moments wherein social difference such as gender, race, and class are invoked and (re)produced, and the consequences they have for both social relations and environments.

New feminist political ecologists thus argue that research needs to pay more attention to how the performance of intersectional social relations is crucial in defining “men,” “women,” “villagers,” and so on. For example, Harris (2009) explores how gender is crucial in the design of large-scale dam projects in Turkey. Project designers worked from particular assumptions about how gender structures household labor, water needs, and the ability to retain new “knowledge” of irrigation and other technologies. These assumptions shape how the dams function and for whose benefit. In addition, masculinities and femininities are transformed by shifting relations to water as a result of the dams. Environmental issues, therefore, are not simply about environment, with perhaps some unfortunate social justice implications. They are foundational to how inequality is conceptualized and maintained within societies and across scales. What makes this kind of analysis “new” is the attention to how gender itself is transformed through environmental issues, as well as gender
relations shaping the struggles that emerge around environment.

Conceptually, this opens up space to understand how changing environments and changing social relations are co-emergent. Any attempts at environmental transformation will lead to transformations in how social difference is defined and performed. This last point distinguishes the new scholarship from earlier historical materialist accounts. Historical materialists are similarly concerned with how environmental projects serve to reshape social relations around those environments. New feminist political ecologists attend to these dynamics, but do so through a more fundamentally relational perspective that seeks to illuminate how the very meanings of “women,” “men,” “race,” or “ethnicity” are transformed by environmental projects (Sundberg 2004).

It is in this sense that this entry argues that gender and intersectional subjectivities produce environments and ecologies (Nightingale 2006). Social relations are not simply played out in environmental projects; they are mapped onto the landscape in ways that have tangible nonhuman effects, and which in turn serve to reshape or cement social relations. For example, land use is framed by intersectional social relations and land use itself serves to transform which species are present, how soils develop, and other dimensions of ecologies, all of which then have implications for what kind of work is performed by whom as ecologies change or remain the same. This kind of argument needs to be distinguished from environmental determinism that suggests environments form societies. The difference lies in how both “environment” and “gender” are conceptualized. A determinist perspective assumes environments have the same influence on all people, with the same outcomes. A new feminist political ecology perspective recognizes both as dynamic assemblages of humans and nonhumans that relationally emerge. So it is not a case of environment shaping society, or society impacting upon environments: rather, the “new” environment and gender debate takes both as contingent, dynamic, and linked material-symbolically: in short, co-emergent.

The production of socio-natures

Through the idea of co-emergence, new feminist political ecologists have thus taken up the challenge originally posed by Haraway to retheorize environment itself. Inspiration for this move was also fueled by debates within geography on social constructions of nature, materiality, actor networks, and posthumanism (Castree and Braun 2001). These debates share a common concern with how society and environments are relationally produced, and a desire to find new theorizations and vocabularies to avoid making analytical separations between them.

When environment and society are kept as separate analytical categories, this leads to a Cartesian separation of humans from their environment, as if somehow we could ever exist outside of it or have no influence within it. This kind of Cartesian thinking has spawned innumerable studies of “human impacts” on environments that fail to account for how what is considered important and worthy as “environment” is relationally constituted with society. In other words, what we value and do not value in nonhumans reflects what we value in society. Again, these debates can seem very abstract, but they have important empirical implications. For example, environmental conservation continues to be justified in the name of protecting a variety of charismatic species, such as tigers, rhinoceroses, or pandas. By focusing on these species, people do not value, and in fact most often do not see, other species and habitats that are equally
important to overall ecosystem integrity, such as micro-organisms, noncharismatic species such as rats, or even humans themselves. This lack of a wider focus sometimes has disastrous consequences for conservation goals, but, most importantly, it also justifies an uneven valuing of different people and their different levels of engagement with environments. The idea of co-emergence draws attention to places, spaces, and species that often fall outside of our consideration as “environment.” Environmental justice activists – many of whom are women – argue precisely that. When “nature” is assumed to exist outside of places where most humans live, it is difficult to bring urban pollution concerns forward as, for example, “environmental issues.”

New feminist political ecologists seek to capture these relational dynamics by keeping in view the networks, relations, and assemblages of humans and nonhumans that serve to constitute “environment” as well as the social-political consequences of how socio-natural difference is conceptualized and becomes important. As such, the environment and gender debate continues to add cutting-edge conceptual contributions to wider debates around environment and society. Taking account of women remains important, but by no means the only, or even necessarily core, agenda of the environment and gender debate. Rather, the debate now retains an emphasis on the politics of environment, adding new tools and new insights from an intersectional and co-emergent conceptualization of “environment and gender.”

SEE ALSO: Environmental determinism; Feminist geography; Feminist political ecology; Gender; Indigenous knowledge; Intersectionality; Marxist geography; Nature; Political ecology; Political economy and regional development; Power; Social constructionism

References


Environment and health

Susan J. Elliott
University of Waterloo, Canada

How is health shaped by the environment? How does the environment shape health? Contrary to what is portrayed by media reports, these are not contemporary concerns. Today’s media are full of references to climate change, chemicals in the water we drink, pollution in the air we breathe, contaminants in the food we eat, inexplicable increases in (re-)emerging environmentally related illnesses (e.g., asthma, food allergies, and infectious diseases such as the West Nile virus, dengue, and chikungunya), but documented interest in the study of how the environment shapes health – and vice versa – began as early as 400 BC with Hippocrates’s famous treatise On Airs, Waters, Places, in which he wrote:

Whoever wishes to investigate medicine properly, should … in the first place consider the seasons of the year … Then the winds, the hot and the cold … the qualities of the waters … and the ground, whether it be naked and deficient in water, or wooded and well watered, and whether it lies in a hollow, confined situation, or is elevated and cold; and the mode in which the inhabitants live, and what are their pursuits, whether they are fond of drinking and eating to excess, and given to indolence, or are fond of exercise and labor, and not given to excess in eating and drinking … For if one knows all these things well, he [sic] cannot miss knowing … the diseases peculiar to [a] place. (Hippocrates 1923, 71)

Essentially, Hippocrates was saying that where we live, and how we live in those places, shape our health. Historical evidence has borne out this assertion. Thomas McKeown (1975) clearly illustrates through historical data that major advances in health in the nineteenth and twentieth centuries had much more to do with lived environments than advances in medicine and/or clinical care, as is often thought. A classic example is tuberculosis, a well-recognized disease of poverty. McKeown shows that the decline in mortality from this infectious disease began long before the disease mechanism was identified (1882), the development of an antibiotic to use in treatment (1947), and the introduction of a vaccine (1953). So what was the cause of the decline? McKeown (1988) explains that the introduction of basic population-level public health measures (e.g., better nutrition, access to safe water and sanitation) led to better health conditions in general, particularly for the urban poor. Indeed, in 2007 over 11,000 readers of the esteemed British Medical Journal voted for advances in water and sanitation as the greatest medical advance of the previous 150 years.

Clear articulation of what is meant by the complex concept of health is a useful point of departure for understanding how it is shaped by the environment; defining environment is just as complex. The definition of health has undergone several iterations in the postwar era. The evolution of health from more than simply the absence of disease to complete physical, social, and emotional wellbeing set the stage for a revisiting of the biomedical paradigm and for the development of models of health and
wellbeing that emphasize not only multiple determinants of health but also the role of the social, cultural, political, physical, and economic environments within which health and well-being are shaped. Alternative health frameworks began to emphasize the role of socioenvironmental factors, in addition to lifestyle and biology, in the production of health and illness, while at the same time diminishing the role of the formal health-care system. Health is now considered a resource for everyday living that allows individuals to manage, cope with, and even change their environments. What is environment? One plausible response is Einstein’s view: the environment is everything that is not me.

In the context of environment and human health, we often think of the physical environment (e.g., air quality) and its impacts on human health (e.g., asthma). But what of the social environment (e.g., social capital), the economic environment (e.g., unemployment rates), the cultural environment (e.g., the role of women), or the political environment (e.g., access to opportunities and institutions)? These environments, too, have been shown to influence human health, which is why the World Health Organization, in a meeting in Sofia, Bulgaria, in 1993, crafted a very broad definition of environmental health that includes quality of life in the concept of health and refers to the influences of the physical, biological, chemical, social, and psychosocial environments. This definition has a further, prescriptive component addressing the correction, control, and prevention of those factors in the environment that can (potentially) adversely affect the health of present and future generations. Regardless of how long these issues have been under consideration or how they are defined, they remain at the forefront of concern for researchers, policymakers and the general public alike, despite the fact that skeptics still exist (see, e.g., Harrison 2000).

Measuring environment and health

The classic example of an environment and health relationship found in virtually every health geography textbook involved the ingenuity of Dr John Snow of London, England. Snow solved the mystery of the cholera outbreak of 1858 by mapping the location of deaths due to this infectious, water-borne disease. At the time, of course, most households did not have piped running water. Rather, standing water pumps, from which local residents drew their daily requirements, were scattered throughout the neighborhoods of inner-city London. When cholera started claiming many lives in the mid-1850s, there was no apparent explanation. At this time, there was little knowledge of the links between cholera and contaminated water. There were, rather, ill-defined notions that cholera – and other infectious diseases – were spread by “miasma,” or bad airs. Snow was suspicious of these notions and decided to map the deaths from cholera occurring across the city. What he found was that there had been a cluster of deaths in the vicinity of one particular pump located on Broad Street. The local water supply was drawn from the River Thames, which was essentially an open sewer for both industrial and human waste. Several water companies drew their water from the Thames, and “treated” the water before distribution to local households; clearly, some companies did a better job than others. Snow ceremoniously cut the handle off the Broad Street pump in order to seal his discovery.

Snow’s action was, however, based on pure speculation: the data that existed were descriptive (a mapping of the deaths), and there was no science (i.e., an understanding of the biological mechanism causing the disease) to substantiate his claim. Some were skeptical (Johnson 2006), and this illustrates the pervasive “rub” in environment
and health: we suspect these relationships, but measurement – of exposures and of outcomes – is elusive. Concomitantly, however, in the absence of strong science, decisions must still be made by politicians and policymakers. Should Bisphenol A (BPA), used in the modern day production of plastics, continue to be used in baby bottles? Or the millions of water bottles that adults use? What type of mass-produced energy source is really “safest” – coal, nuclear, or gas? What about fracking? (Jackson et al. 2014). We also have to make these decisions at the individual and/or family levels. Should I use pesticide on plants in the back garden? Should I or we eat organically grown food even though it’s much more expensive? Where is the evidence and how can we go about getting it?

Design

There are several different study designs to choose from in any health research endeavor. The strongest of these is the randomized controlled trial (RCT). In such a study, the researcher has control over both the exposure of interest as well as the sample exposed. Typically, RCTs are undertaken to test a new treatment for a known illness, often a newly developed pharmaceutical. The researcher would select a control group (which did not get the newly developed pharmaceutical or would receive a placebo) and an intervention group (which would receive the newly developed pharmaceutical), and assess the impacts (ideally in a double blind manner; that is, neither the researcher nor the subject would know if they were receiving the newly developed pharmaceutical or the placebo). Widely heralded as the strongest study design and yielding the most credible results, RCT is impossible to implement in an environment and health study. For many reasons, including ethical ones, it would be impossible to expose, say, a group of children to lead for a period of time and then assess the impacts compared to a control group not similarly exposed. As an aside, we often have access to “natural laboratories” in environment and health research. For example, the 1976 explosion at a chemical factory in Seveso, Italy, allowed researchers to follow the workers and residents of the surrounding community for the next 30 years to assess the long-term health impacts; however, because this was not an RCT, the researchers could not control for all of the potential confounding factors that may also have influenced the health outcomes of interest (e.g., whether residents smoked, and whether they were exposed to other sorts of toxins in the residential environment).

Since we cannot run RCTs in environment and health research, we are limited to observational studies – essentially, observing what has already happened. One form of observational study is the case control design. In this type of study, a number of cases are identified – for example, a cluster of leukemia cases – and these individuals are studied for their exposure histories. This was the study design used in the famous case of Woburn, Massachusetts. In this case, a seemingly inexplicable cluster of childhood leukemia cases occurred in a residential neighborhood. Area residents suspected the local water supply to be the culprit. For some time, the water had smelled of chemicals and led to rashes and burning on human skin. Residents worked to convince local authorities and physicians that there was indeed a link between the leukemia cases and the water supply. A local industry was subsequently convicted of contaminating a city well. (The story was later made into a Hollywood movie entitled A Civil Action, starring John Travolta.)
A second type of observational study is the cohort study. In this approach, sampling is on the basis of exposure, as opposed to outcome (as in the case control study). Cohort studies can be one of two types: retrospective or prospective. In the former, a cohort of individuals is identified who are thought to have experienced an exposure. The cohort is then studied to determine what, if any, adverse health effects they might have experienced in the past. This is the typical type of study done in environment and health research, often because a community realizes it may have been exposed to something (e.g., an air pollutant, a leaking landfill, or a contaminated water supply) as a result of some ill-defined symptomatology or an inexplicably large number of cancer cases in a neighborhood. This self-identified cohort may then be studied to assess potential ill-health effects experienced in the past during the exposure. An example here is the Sydney Tar Ponds in Nova Scotia, Canada. This site, known as the most heavily contaminated industrial site in Canada, saw approximately 100 years of coal smelting and steel production, characterized by inappropriate waste disposal practices. The site is now contaminated with 700,000 tons of known carcinogens. Residents living within two miles of the site have been very concerned about the impacts of the exposure on their daily lives as well as anxieties around reproductive health and decision-making. In most instances, a mixed methods approach will provide the most comprehensive picture (Gatrell and Elliott, 2014).

Measurement

Complicating matters further are issues of measurement: it is no easy task to measure exposures and outcomes, and to ascertain the pathways through which individuals are exposed. Each of these aspects of measurement is elaborated on in what follows, drawing on the environment and health literature for illustrative examples.

Exposure

What are we exposed to and how do we measure it? With respect to the former question, toxins in our environment that have long-term, irreversible effects are of most concern. For example, neurotoxins (e.g., pesticides and heavy
metals like lead, mercury, and chromium) attack the nerve cells. Radiation can cause genetic damage to the self and to one’s offspring. Teratogens (some solvents and heavy metals) can cause birth defects and carcinogens (e.g., PCBs) cause cancer. Another environmental health concern plaguing us now is endocrine disrupters. The first compelling evidence for this concern comes from animal populations that have shown changes in fertility and other reproductive problems, including the feminization of the male of the species (Bergman et al. 2013); however, when individuals report to their friends, family, physician, or local public health official that they feel they have been exposed to some environmental hazard, it is very difficult to measure that exposure. First, the exposure happened in the past, and recall bias will impact on how accurately we are able to measure the exposure (do you remember, for instance, how often you spent time in your basement in the last five years? How much time you spent in the shower or bath? How much tap water you consumed? How much time you spent in the garden or yard of your house?)

Second, often the exposure comes from an illicit, undocumented source, for example, chemical contamination by an industry to the public water supply in Woburn, Massachusetts. Third, we rarely know what the exposure contains. For example, when Lois Gibbs first discovered yellow ooze seeping into her basement in her Love Canal neighborhood home, all she knew was that it was yellow ooze. Toxicologists then had to discern the composition of the ooze, which is very difficult to do. When toxicologists investigate a toxin, they need to know what they are looking for so that it can be identified. Are we looking for polyaromatic hydrocarbons? Dioxin? Dieldrin? Trichloroethylene? Each of these toxins has a unique identifying character. We cannot simply run yellow ooze through a chemical scanner and identify its contents.

Fourth, we know very little about the synergies between chemical contaminants. That is, we may know with some certainty the relationship between trichloroethylene (a chemical primarily used in the process of dry cleaning) and bladder cancer, but what if the trichloroethylene is mixed with polyaromatic hydrocarbons? Is there a synergistic effect that can actually cause more harm to human health than if one were exposed to these chemicals in isolation? Finally, all of this science is very new. Until as recently as the 1960s, chemical companies invested in television advertisements promoting the wonders of chemicals and how they were going to change our lives. A current challenge is that technology can outpace knowledge. For instance, we can now measure contaminants in exceedingly small amounts, yet we have very little if any knowledge of the impact of exposures at that level on human health.

Outcome

If we can ascertain what an individual or a community has been exposed to, we must then determine what health outcomes we should be documenting. This is a very difficult task, for a number of reasons. First, we simply do not know how community-based (as opposed to occupational) exposures to these chemicals will affect human health. Therefore, we do not know what to look for. In the highly regarded Upper Ottawa Street Landfill study undertaken in Hamilton, Ontario, Canada, researchers documented the entire health histories of residents who had unknowingly been exposed to (illegally dumped) liquid industrial waste for over five years. Every organ system was examined. This process was extremely time-consuming and resource-intensive. But, not knowing what to look for, these researchers looked for everything. Ultimately they discovered very little: some...
red, itchy eyes; and some psychosocial stress. Why might this have been the case? It could be that the dose that residents were exposed to was insufficient to have an impact on their health (as the sixteenth-century Swiss German physician Paracelsus remarked, it is the dose that makes the poison). It could also be that, as in many cases, the health impacts experienced are ill-defined symptoms that are difficult to trace to any specific exposure. For example, we often hear reports of headaches, rashes, stomach problems, sleeplessness, fatigue, and psychosocial distress. These symptoms could be due to any number of factors; again, how do we control for all the confounding factors that exist in a community “laboratory”? The third reason why these researchers, despite their very thorough examination of the population, may have found very little in the way of health effects, relates to the issue of latency. That is, if we are concerned about cancers or some other type of chronic illness, it takes time (perhaps 15–20 years) for such illnesses to develop. Following an exposed cohort for 20 years is very expensive research and is fraught with difficulties (e.g., control for confounders, sample attrition).

The abovementioned concerns are not to belittle feelings of anxiety and concern around exposure — that is, the psychosocial impacts of exposure. Psychosocial impacts can be defined as the combination of distress, dysfunction, and disability resulting in social, behavioral, and psychological impacts on individuals, groups, and communities as a consequence of actual or perceived environmental contamination. These are real impacts, worth studying primarily for two reasons. First, they are health impacts in and of themselves (recall the earlier definition of health). It is difficult to have full health when one is anxious, worried, or distressed. Indeed, despite whether the exposure is actual or perceived, these are real effects that have real consequences. As the sociologist W.I. Thomas expounded in the 1960s, if a situation is defined as real, it will be real in its consequences. For example, the study of the Sydney Tar Ponds found no excess levels of adverse reproductive outcomes (i.e., stillbirths, miscarriages, low birth weight); however, the perceptions held by local residents of the impacts of the tar ponds on reproductive health was so strong that they were making decisions about fertility based on this perception. For example, some families were putting off having children until the site was cleaned up. Others were deciding not to have children at all. Still others were choosing to adopt from outside the community.

The second reason why it is important to study the psychosocial impacts of exposure is their relationship with physiological health. For example, when the chemical factory in Seveso, Italy, blew up in 1976, and workers and local residents received high doses of dioxin exposure, researchers who followed this cohort for over 30 years expected to find elevated rates of cancers, particularly liver cancer. They did not, but they found a fivefold increase in cardiovascular disease. The researchers speculated that the exposed cohort worried so much about developing cancer that they developed cardiovascular disease as a consequence.

Pathways

There have been many situations in which an actual documented environmental hazard exists, and individuals and communities are worried about their health. But before we can claim a relationship, there has to be a clear pathway for the exposure. For example, the contamination from the Sydney Tar Ponds site was to a tidal estuary. A moratorium on fishing was established, and area residents did not obtain their water from this source, so, if there had been an
exposure, what was the pathway? Such pathways are clearer in the case of Bhopal, where a train derailment led to the release of 40 tons of methyl isocyante into the air. Individuals could then have breathed it in. This could have been a pathway for exposure. Another clear example is the radiation exposure at Chernobyl. It was relatively straightforward to assess immediate exposure through documentation of radiation sickness, and later through the consumption of contaminated crops and livestock.

Evaluation

Once exposure and outcome have been measured, the existence and strength of the relationship between them need to be evaluated. This evaluation can be undertaken by calculating odds ratios, that is, the likelihood of an individual experiencing the health outcome of interest given the exposure of interest. For example, at Seveso, the researchers documented an odds ratio of 5 for cardiovascular disease, indicating that those who had been exposed were 5 times more likely to develop cardiovascular disease than those who had not been exposed. An odds ratio of 5 in environmental health research is quite high; typically, we see odds ratios between 1.5 and 2. While these odds ratios may be statistically significant, they may or may not be clinically significant. For example, studies of the relationship between air pollution and premature mortality in Montreal, Canada, indicate an odds ratio of about 1.03 of dying prematurely on a “bad air” day. While this relationship is only marginally significant in statistical terms, it might be regarded as clinically significant because the odds ratio is being applied to a large urban population (i.e., approximately 2 million people). This environmental exposure may not be a high risk but affects a large number of people, potentially legitimating public health response.

Even if a relatively robust odds ratio is apparent, there is still a need to ascertain whether there is a causative relationship between variables. In the 1970s Sir Austin Bradford Hill gave an after dinner speech wherein he outlined what are now euphemistically known as the Bradford Hill criteria or the “9 diagnostic tests of causation.” Briefly, Bradford Hill asserted that one could not claim causation unless most, if not all, of these criteria were met. They include strength, consistency, and specificity of association; temporality (i.e., that the cause comes before the effect in time); existence of a biological gradient or dose–response curve; biological plausibility of the association; coherence of the relationship with what we already know; experimental data; and, analogous evidence in the literature (for further information, see Bradford Hill 1965).

What does the future hold?

Two fundamental aspects of environment and health will guide geographical and related public health research for at least the next couple of decades. The first is how we view issues of human health and the environment. In this context there is movement toward the “One world, one health” movement, which brings all actors in the ecosystem (animals, humans, natural environment) together into an integrated whole. This development is implied in John Last’s definition of sustainable health:

If human affairs are conducted in such a way as to sustain life-supporting ecosystems in a stable state of equilibrium, then humans will survive and flourish. If, however, a local, regional or global ecosystem is degraded beyond the point where it is sustainable [think Aral Sea], then the health and survival of humans in that ecosystem cannot be sustained. (Last 2014, 1175)
ENVIRONMENT AND HEALTH

There is no question that the greatest environment and health question currently facing human kind is global environmental change, including climate change. Again, the health effects of global change are often indirect and difficult to assess, and the availability and quality of evidence vary widely. Malaria is increasing – is that due to climate change? The West Nile virus is found in places where it has never been found before (North America) – is that due to climate change? And will the climate warm by 2°C or 4°C and what havoc will the latter wreak? (New et al. 2011). Faced with this prediction, most scientists have been literally stunned into silence and have been paralyzed by their incapacity to envision this climatically changed world. In the meantime, others continue to ponder the capacity for adaptation and resilience. In this regard, Brown (2014) and O’Brien (2012a; 2012b) lament the lack of focus on true transformation that would benefit those who are most vulnerable. With respect to this last point, as is the case with all health outcomes, it is invariably the most vulnerable (the poor, the marginalized) who continue to be most impacted by the environmental determinants of health. Much work therefore remains to be done to link the domains of environmental health and social justice (Walker 2012).

SEE ALSO: Climate change and health; Environmental health; Environmental issues and public understanding; Geography and the study of human–environment relations; Health geography; Health inequalities; Vulnerability

References


Websites


Environment and law

David Correia
University of New Mexico, USA

Environment and law have multiple meanings. Environment, in all uses, refers to the surrounding arrangements, relations, and factors required for a given system, organism, or object. Its common usage refers to the biophysical foundation for human and nonhuman life, but its use extends to nonbiophysical domains such as the built environment, the social environment, the digital environment, and so on. Environment as the biophysical serves as a synonym for one definition of the word “nature.” For Raymond Williams (1985) “nature” is among the most difficult words in the English language. Often it describes an essential or inherent quality of something, as in the phrase “human nature,” which is used as a way to describe the condition of being human. Law too has no singular meaning. Legal scholars generally understand law as a way of ordering and making sense of the world. Law is a collection of negotiated rules that govern behavior and are enforced by a state or through social institutions. Law is spatial. It draws literal and figurative boundaries in the world. It is a set of historically produced, unevenly mediated, social and spatial practices driven by logics of boundary-making and categorization that produce the effect of law as a form of social domination claimed as legitimate.

According to legal geographers like Nicholas Blomley and David Delaney, the law reflects and produces sociospatial relations (see Blomley 2008; Delaney 2003). As a mode of inquiry, legal geography maps the social relations and spatiality of law.

For legal scholars influenced by liberalism, the domination that law requires serves as a means to a just end. Law defines and enforces a particular and just social order (Dworkin 1986). For critical legal scholars, however, the force found at the heart of legal authority can also serve as “an apparatus of violence that disorders, disrupts and repositions preexisting relations and practices all in the name of an allegedly superior order” (Sarat 2001, 3). To law, nature is the material, nonhuman, primordial stuff of the world; and law makes sense of this world, translating the nonhuman into legible forms and stable categories. So it is through the categorical logics of law that the irreducible complexity of “nature” becomes the legible and orderly category of “environment” placed in managed relation to the human. Making nature legible as “environment” reflects a central epistemological effect of law, a shift from the radical otherness of nature to the ordered legibility of environment is perhaps law’s most important accomplishment. Law works to reduce the irreducible complexity of nature into discrete juridical categories or objects of interest. Environment in this sense is made legible to and by the state, courts, and legal authorities.

While this diversity in use and meaning complicates any straightforward definition, two broad domains of environment and law emerge. The phrase “environment and law” can be examined as, first, a reference to natural law or the perceived intrinsic or innate laws of systems understood or depicted as natural (whether biophysical or social) and which serve as one origin of sociolegal relations; and, second, as the manner in
which legal reasoning influences environmental management and nature–society relations more broadly.

Natural law

As a reference to fundamental properties, natural law refers to universal claims to legitimate law whose legitimacy exists as a function of its basis in nature or in the divine. This is the idea, in other words, that law emerges from nature and, as such, exists as an objective, universal, and neutral source of legitimate authority. Various forms of natural law can be found throughout history, including in Greek and Stoic philosophy among others, and in all cases it serves as a means to understand the complex, shifting, and porous boundary between the natural and the social. Carolyn Merchant (1980) in her influential book The Death of Nature examined the relationship between Renaissance views of nature and social organization. For example, one common view of nature, both in the Renaissance and in contemporary environmental thought, is that nature is best understood as a designed order arranged in hierarchical relationships. This view, according to Merchant, serves as a powerful and conservative rationale for resisting change to a particular social order. In other words, if nature is orderly and hierarchical, society should be too, and thus society must not depart from a natural order. This depoliticizing logic is recast when nature is depicted as an active unity of opposites in dialectical tension. Where such a view is influential, so too is the idea that change serves a progressive function in society. In other words, political change is natural and thus imperative. This has made its way into modern legal and political theory though a number of avenues. Thomas Hobbes (1994) elaborated laws of nature that described the rights and obligations of humans that emerge through submission to a sovereign authority and serve as the means to overcome a state of nature he depicted as “nasty, brutish and short.” Hobbes’s “state of nature” is a condition of constant and universal jeopardy that exists without the possibility of human cooperation and, therefore, security. We can escape the anarchy of nature by submitting to a voluntary social contract based on law and order that is by definition not natural. Hobbes offers a paradox in which nature’s laws cannot operate in a state of nature and thus require law and order as much as we do.

According to Jean-Jacques Rousseau, the governed social world of legal standards and codes does not save us from nature but rather serves as the origin of social inequality. In contrast to Hobbes, Rousseau found in nature not a world of violence and misery, but a state of human equality corrupted only with the emergence of a social contract that imposes social inequality. He famously wrote in Discourse on Inequality:

The first man who, having fenced in a piece of land, said “This is mine,” and found people naïve enough to believe him, that man was the true founder of civil society. From how many crimes, wars, and murders, from how many horrors and misfortunes might not any one have saved mankind, by pulling up the stakes, or filling up the ditch, and crying to his fellows: Beware of listening to this impostor; you are undone if you once forget that the fruits of the earth belong to us all, and the earth itself to nobody. (Rousseau 1984/1754, 109)

In both cases, and despite their differences, Hobbes and Rousseau turn to natural law as both the measure and the means of legitimate sociolegal relations.

Both understand the origins of legitimate social liberty and political freedom as expressions of natural law – nature and its intrinsic properties as society’s wellspring. This is particularly explicit
in the writing of John Locke, who located in nature the very laws from which he described the terms of an emerging, and unequal, social order. According to Locke (1976/1690), by natural right, nature, once transformed through human labor, becomes the private property of whoever transforms it. Thus natural law emerges in order to impose a social order that legitimizes the very social inequality Rousseau, for example, condemned. It may be unequal, in other words, but it is found in nature and thus is natural and therefore legitimate. For Locke nature’s law serves as the means through which common rights become private rights. This right to the private accumulation of wealth becomes the basis for social inequality and serves as the key element in liberal formulations of law depicted by Locke as “life, liberty and property.”

Where natural law theorists posit law as emerging from nature or the divine, positive law theorists argue that law must be imposed by a sovereign politico-legal authority. Positive law makes normative claims regarding law’s legitimacy without recourse to morality. Likewise common law is understood as imposed, although, unlike positive law, common law refers to a legal system organized by principles established by the precedent of previous decisions by judges.

Where environment and law collide

Law is a key means through which human uses of nature are understood and regulated. This takes a number of forms. Environmental legislation and statutes such as the Clean Air Act (1970) and the Clean Water Act (1972) in the United States regulate how natural resources are managed for commercial and industrial uses. A series of industrial accidents and influential books on environmental issues shifted public perception of nature in the 1970s, gave momentum to a series of laws that marked a shift in how natural resources were managed for commercial and industrial uses, and paved the way for such legislation. The publication of Silent Spring by Rachel Carson in 1962 drew in sharp relief the human and nonhuman cost of the widespread release of chemical pesticides into the environment. These concerns were heightened in the late 1960s as a series of industrial accidents gave further momentum to environmental concerns over resource management. In January 1969 a Union Oil platform off the California coast exploded, sending an estimated 100,000 barrels of crude oil into the Santa Barbara channel. Months later, the heavily industrialized Cuyahoga River near Cleveland caught fire (and not for the first time). In 1971 Barry Commoner, a professor of biology at Washington University in St Louis, Missouri, published The Closing Circle. In it he argued that US society should be restructured to conform to what he called the four laws of ecology: (i) everything is connected to everything else; (ii) everything must go somewhere; (iii) nature knows best; and (iv) there is no such thing as a free lunch.

The US Environmental Protection Agency (EPA), created in 1970, enforces laws such as the Clean Water Act through the promulgation of administrative law, which governs the ways in which the EPA enforces environmental regulation. The Endangered Species Act (1973), for example, established a legal framework controlling economic activity in habitats deemed critical to the survival of threatened species. The Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) of 1980, commonly known as Superfund, created a legal mechanism and fund to clean up major contaminated sites. These environmental laws defined the purpose of the EPA and transformed government agencies; for example, the United States Forest Service, which during the first half of the twentieth century managed forest resources
entirely in service to an industrial timber industry, was turned into an agency that was required to mitigate the negative effects of those uses.

Following the rise of environmental regulation by federal and state agencies via administrative law, *civil law* tort actions by environmental organizations became a frequent tactic used by a variety of environmentally minded plaintiffs to force state agencies to limit industrial uses of nature.

The arena of *international law* has also become a key site of struggles over the environment. Signatories to the 1989 Montreal Protocol agreed to prohibit the use of ozone-depleting substances. A multilateral fund was created to finance conversion efforts. To date it is the only treaty universally ratified by member countries of the United Nations. The 1992 Framework Convention on Climate Change, or the Kyoto Protocol, sought to resolve the problem of anthropogenic climate change by stabilizing and eventually reducing global greenhouse gas emissions. In contrast to the Montreal Protocol, which relies on mandatory reductions to accomplish its goal, Kyoto is based on a cap-and-trade scheme. Limits on greenhouse gas emissions were established for each signatory and these countries then establish limits for different industries. Individual firms are given a legal right to emit an amount of greenhouse gas no greater than an industry average, thus providing the firm with a legal claim to a portion of the atmospheric commons. If a firm is able to lower its emissions, often through the application of clean technologies, it can sell its unused claim to the atmosphere on a carbon market or to other firms. To date, 192 countries have ratified the agreement; the United States and Canada are not among them.

International law has also been a key site for struggles around the rights of nonhuman nature in legal questions. In US case law, nonhuman nature has no standing and therefore cannot claim legal rights. Acknowledgment by US courts of the possible standing of nonhuman nature has been rare. In the landmark case *Sierra Club v. Morton* 405 U.S. 727 (1972), the Sierra Club, a membership-based US environmental organization, sued the US Forest Service to stop the development of land near Sequoia National Park. The US Supreme Court determined that the Sierra Club lacked standing (the ability to demonstrate a connection to harm as a result of the activity in question), but the ruling included a dissenting opinion by Justice William O. Douglas regarding the question of legal claims to nature, including standing for nature itself. In an expansive opinion, he wrote:

> The river as plaintiff speaks for the ecological unit of life that is part of it … The voice of the inanimate object, therefore, should not be stilled … before these priceless bits of Americana (such as a valley, an alpine meadow, a river, or a lake) are forever lost or are so transformed as to be reduced to the eventual rubble of our urban environment, the voice of the existing beneficiaries of these environmental wonders should be heard.

While nonhuman nature has no legal standing in US common law, the formal extension of legal rights to nature are found in legal documents worldwide. The Swiss Constitution requires that consideration be given to the animals, plants, and other organisms. Chapter 7 of Ecuador's 2008 Constitution is titled “Rights of Nature” and begins with Article 71 “Nature, or Pacha Mama, where life is reproduced and occurs, has the right to integral respect for its existence and for the maintenance and regeneration of its life cycles, structure, functions and evolutionary processes.” In 2012 the Waitangi Tribunal, a special jurisdiction unit of New Zealand’s Ministry of Justice charged with investigating and making recommendations on claims brought by Māori, heard a claim by the indigenous Whanganui River Iwi,
which asked the tribunal to grant legal personhood to the Whanganui River. The subsequent agreement established a legal framework that extended legal standing to the river.

**SEE ALSO:** Environmental regulation; Environmentalism; Legal geography; Nature; Nature conservation; Property and environment; Socio-nature

**References**


Since the mid-1980s the relationship between human migration and the environment has become a widely debated topic among policy experts, political actors, and activists, and in the media. This attention has fueled a renewed interest in the issue by geographers, as well as by other social and natural scientists. It is in the context of global environmental change, but essentially in the context of anthropogenic climate change, that the relationship between the environment and human migration has been politicized, debated, and examined. As a result of these debates, environment and migration has become as much a topic of scholarly interest as a discourse used by a wide variety of actors which sometimes bears little relation to actual natural and social processes.

Earlier interventions have offered forecasts of hundreds of millions of “environmental refugees” or “climate refugees” (people forced to relocate because of environmental and climate change) as a way to draw attention to the consequences of environmental change such as sea level rise, droughts and desertification, or cyclones, heavy rains, and floods. The claims that environmental change would induce such a massive displacement of people were criticized from a variety of disciplinary and political positions which underlined the impossibility of singling out the environment as a sole driver of migration, and castigated the crude environmental determinism informing the environmental refugees thesis. Subsequently, in the late 2000s, in response to a knowledge imperative, to funding opportunities, and to policy demands from international organizations (such as the International Organization for Migration and the World Bank), Northern governments, and nongovernmental organizations, major research projects have been undertaken in order to offer a more refined picture of the relations between human migration and the environment. These have led to a much more nuanced, dynamic, and complex understanding of these processes, and have emphasized methodological issues. Although most participants in the debate share the assumption that environmental change can somehow cause human migration, they reject the existence of a linear causal relation between these two phenomena. The projects have also offered new normative visions and policy proposals.

The definition of what constitutes “environmental migration,” “environmentally induced migration,” “environmentally induced population movements,” “climate migration,” “climigration,” and so on is strongly contested, and disagreements over definitions are constitutive of the history of the field (Gemenne 2012; Piguet 2013; Morrissey 2012). Generally speaking, the “environment” has been understood in a rather narrow and naturalistic sense in these debates. It relates to the nonhuman biophysical world and includes some elements of socially modified nature, most notably climate change. The whole debate is still largely informed by neo-Malthusian conceptions of the Global South (Morrissey 2012). Its focus has been essentially on Southern countries, rural areas, agricultural livelihoods, and poor vulnerable
ENVIRONMENT AND MIGRATION

people. It has also often been characterized by an understanding of environmental migration (especially that caused by climate change) as something happening in the future. At the same time, population movements caused by industrial hazards, urbanization, or deindustrialization in Northern countries have generally not been considered as falling within the realm of environment and migration. Similarly, research on migration and the environment has tended to discuss temporary or permanent displacement from degraded environments, but has rarely considered the environmental amenities that may motivate people to settle into new environments. For instance, suburbanization is generally not considered as an issue of environment and migration, although, arguably, many people choose to relocate in suburbs away from a noisy and polluted inner-city environment in order to enjoy environmental benefits such as a garden.

Nor are European colonial settlements understood as earlier examples of environment and migration in spite of the fact that these processes had environmental causes and (often deleterious) environmental consequences. Wealthy North American citizens who move to Florida after retirement are generally not considered to be environmental migrants.

Indeed, environment and migration is still essentially considered in policy and scientific debates as an issue, a problem, or even a looming crisis. In these assumptions, it is a problem that happens in poor Southern countries and that needs to be solved, or at least managed, if possible, at an international level. Thus, so far, managerialist and policy-driven perspectives have dominated this scientific field, with many contributors to the academic debate also working as experts and consultants for a variety of governmental or international institutions. Indeed, it has been argued that scientists working in this field should be seen as “policy entrepreneurs” rather than as neutral observers (Gemenne 2012). Most contributors to these debates tend to agree that the environmental changes are real enough and that, although they may not directly produce migration, they tend to reinforce factors affecting human vulnerabilities.

Some radical geographers and political ecologists have questioned the framing of the whole environment and migration debate, unearthing some implications of the environment–migration nexus, especially in terms of racialized, gendered, and political-economic processes. These have notably argued that putting too much emphasis on environmental change obscures the real social processes leading to situations of poverty, exploitation, and dispossession. Rather than being the effect of unintended environmental change, human displacement may actually reinforce existing structures of power such as capital accumulation, patriarchy, and state domination.

Earlier conceptualizations of environment and migration relations

The institutionalization of the social sciences, including geography, as academic disciplines at the end of the nineteenth century in Europe and North America was closely linked to the development of nation-states as territorially bounded and competing entities. Knowledge of population and the control of population movements are important features in the constitution of a territorialized sovereignty. Early social science contributions to the study of migration shared some of the territorial assumptions of this emerging paradigm of sovereignty: mobility and migrations, rather than fixity, were in need of explaining.

In an insightful article Etienne Piguet (2013) has shown that the link between migration and environment was an important concern for the
founders of migration studies at the turn of the nineteenth and twentieth century. The influence of environmental factors in explaining patterns of migration was quite widespread then. It was used, for instance, to explain the great “barbarian” invasions and the fall of the Roman empire. Policy-oriented knowledge was also important, for instance, in order to select places appropriate for European colonists to settle in. This is hardly surprising considering the origins of academic geography as a naturalist and determinist form of knowledge and its institutional contribution to the consolidation of the power of emerging national states.

However, studies on the influence of the environment on migration almost totally disappeared from the literature between the beginning of the twentieth century and the mid-1980s. According to Piguet (2013), there were different reasons for this. The growing domination of natural landscapes by industrialized societies led to the idea that natural factors had less of an influence on human actions. Crude physical or environmental determinism was increasingly rejected in the social sciences. Human and physical geography became dichotomous disciplines with different sets of explanations, and human migration was regarded as a social phenomenon to be explained by social, rather than by environmental, factors. The growth of an economic paradigm in migration studies further entrenched the separation between the natural environment and migration. Finally, so-called refugee studies or forced migration studies developed from the 1980s, predicated on the political definition of refugees in international law, which revolves around the notion of persecution and is thus hardly compatible with environmental factors.

Accordingly, the reappearance of the relation between human migration and the environment was not triggered by migration scholars or geographers. Rather, it was the work of environmentalists and policymakers which sought to politicize the consequences of global environmental change. It was thus strongly impregnated with determinist assumptions about the causal influence of the environment on human societies: in spite of strong criticisms, these assumptions have proved quite resilient in subsequent debates.

The environmental refugees thesis

The renewal of studies on the environment–migration nexus was predicated on what has come to be known as the environmental refugees controversy. Although this controversy was always more complex and fine-grained than later accounts of it, it is generally considered to have set up two opposing theses, variously described as the maximalists versus the minimalists, the alarmists versus the skeptics, or the proponents of the environmental refugee thesis versus its critics (for detailed accounts see Gemenne 2012; Morrissey 2012).

The notion that environmental degradation could lead to population displacement was entertained from the 1970s by prominent environmentalists such as Lester Brown, founder of the Worldwatch Institute. Yet, it is commonly agreed that the foundational document for the revival of the environment–migration linkage is a policy document entitled Environmental Refugees, written by Essam El-Hinnawi and published under the auspices of the United Nations Environment Programme (UNEP) in 1985. It offers the first definition of environmental refugees and a rudimentary typology of the environmental causes of displacement: it distinguishes between those “temporarily displaced because of an environmental stress,” such as earthquake, drought, or flood, those “permanently displaced
and re-settled in a new area” because of development projects, and those who are victims of slow-onset environmental degradation such as deforestation and desertification. As the choice of the word “refugee” suggests, El-Hinnawi insists that those whom he thus categorizes are victims and are forced to move. But he also hints that those refugees from the third category are partly to blame because of their supposed “mis-use” of the environment through, for instance, slash-and-burn agricultural practices. This first intervention was thus instrumental in framing the debate and construing environmental refugees as both victims and perpetrators of environmental degradation.

The work of biologist and environmentalist Norman Myers has proved to have the greatest influence on the revival of the migration and environment debate. From the early 1990s and into the early years of the twenty-first century, both alone and in collaboration, Myers has repeatedly offered the most gripping account of the potential migratory consequences of environmental change (e.g., Myers and Kent 1995). He is also responsible for offering quantified projections of future environmental refugees. His prediction that there would be 200 million environmental refugees by the year 2050 has been very widely repeated and even amplified, generally quite uncritically. It was, for instance, cited in the authoritative Stern Review on the Economics of Climate Change, published by the UK government in 2007.

This strand of studies was infused with a strong positivist perspective and a globalist bias. It framed climate or environmental change as a “global” problem that would therefore have global migratory consequences, which could be quantified. Often carried out by environmentalists and natural scientists with little training in migration studies, such research assumed that migration would be the quasi-automatic response to a degraded environment. It forecast future migration patterns by combining maps of populations and estimates of environmental impacts, especially the impacts of climate change. For instance, densely populated coastal areas (notably in Asia) or low-lying island states (such as the Maldives and Tuvalu) were flagged as potential victims of climate change–induced sea level rise. These projections were underpinned by a form of neo-Malthusianism, whereby environmental degradation and poverty were explained with reference to population rather than to social, political, and economic relations. In this sense, environmental refugees appeared as the logical result of global processes of environmental degradation.

The limited literature in geography, migration, and development studies which dealt with migration as a result of environmental variations, especially in the context of drought and desertification, offered little support for these estimates. Rather, local and regional empirical studies tended to show that migration was only one possible coping strategy in the face of environmental change, that such migrations tend to be short-range rather than international, and that it is extremely difficult to disentangle environmental causes from other drivers of migration. Furthermore, empirical studies highlighted the importance of the differential vulnerabilities created by social relations along gender, class, ethnic, and age lines. Whereas the impact perspective projects environmental hotspots onto whole territories which are thus classified as at risk, more concrete analysis avoids such sweeping generalizations and displays a much more nuanced context- and place-specific analysis.

The environmental refugees thesis thus came under sustained attack from the late 1990s and early 2000s, especially from geographers and migration scholars (for an influential example, see Black 2001). These authors, although
they shared the concern of the maximalist/catastrophist school about the linkages between environmental change and migration, nevertheless expounded the idea that this relation was direct, monocausal, and linear. Rather, they insisted that migration decisions are essentially multicausal and that the environment cannot be abstracted as their primary cause; that migration (especially temporary and circular migration) has long been a response to environmental uncertainties (a form of livelihoods diversification) and thus should not be seen as something radically new and catastrophic; and that most vulnerabilities which may lead to forced migration are the result of sociopolitical (mal)development projects rather than of an apolitical or ahistorical environmental change. Some of these “skeptics” thus reproached the “proponents” for their purported depoliticizing agenda, for ignoring the political economic contexts of migration, and also for describing as “migrants” people who should simply be regarded as “refugees” and in need of protection.

Because of the widely shared dissatisfaction with the environmental refugees literature, and as a result of the constant calls for more empirical research and greater theoretical clarification, a number of important works have recently been published on the topic. They have also benefited from public funding made available by national governments and international organizations in order to produce policy-oriented research. This has been the case, for instance, with the research program EACH-FOR funded by the Commission of the European Union from 2007 to 2009, and with a major report commissioned by the United Kingdom’s Government Office for Science (Foresight 2011). Over 70 peer-reviewed studies have been commissioned and some 350 experts and “stakeholders” have been involved in creating Migration and Global Environmental Change. The report is targeted at a wide readership in academia and in policymaking circles at both national and international levels. The theoretical underpinnings of the report have been published in influential scientific journals and may be regarded as the new orthodoxy on the topic (Black et al. 2011).

Instead of focusing on a typology of environmental migrants or on finding a simple correlation between environmental stress and migration, Black and his colleagues (2011) insist on the complexity of this relation and that migrations are always multicausal. Migrant should be conceptualized as having agency and not simply as passive victims of structural forces. Thus, rather than identifying types of migrant (whether or not environmental), these authors define five drivers of migration (economic, political, demographic, social, and environmental) which may act as barriers or as facilitators of movement. They then argue that it is the interaction between these drivers which should be the focus of researchers’ attention. Environmental change does not produce migration as such, but it may reinforce or inhibit existing migration patterns. For instance, climate change may reduce crop productivity in some places and thus affect the income of rural households. This in turn may affect (as an economic driver) the choice of these households to migrate, but it may also reduce their ability to migrate (by reducing their assets), should they choose to do so. Some critiques of this approach have taken issue with the apparently arbitrary division between these five drivers, and hence the circularity of the argument (why, for instance, should land productivity be considered as an environmental driver rather than as an economic one?). Nevertheless, this new approach has proved to be quite authoritative.

Whereas previous interventions, especially by environmentalists, tended to present migration as a failure to mitigate environmental change,
the new orthodoxy insists on the fact that migration should actually be understood as a strategy of adaptation to climate and environmental change. Migration is thus understood as a risk management strategy by rural households, albeit one which is not universally accessible. Indeed, mobile households show their resilience to adverse conditions such as climate change, and may even increase their resilience by moving, as their migration allows them to send remittances, act as knowledge transfer agents, and so on. In this perspective, the problem actually becomes the fate of those unable to move because of their lack of resources or because of barriers to migration. “Trapped populations,” rather than environmental refugees, are increasingly becoming the new topic of concern, which warrants increasing migration management intervention. Indeed, the new orthodoxy in environment and migration research broadly fits the migration management agenda of international organizations, whereby these institutions seek to organize human mobility in order to make it productive for economic accumulation (Felli 2013).

Normative and legal responses to the environmental refugee thesis

The plight of environmental refugees has also attracted the attention of political scientists, scholars of international relations and international law, and normative political theorists and philosophers from the early 2000s onward. Most normative discussions dwelled on the rights of those (potential) migrants, who are portrayed as victims of environmental change, and on the responsibility of Northern industrialized states in this situation. In keeping with the tone of the debate, most of these interventions took for granted some of the catastrophist predictions of an impending refugee crisis, or at least accepted the framing of the issue as an international problem in search of normative and legal solutions.

A vexed question from the beginning was the use of the word “refugee.” Although it was purportedly used in order to suggest a process of forced displacement, some scholars opposed it on the ground that it created an unfortunate association with the juridical language of refugee protection (Black 2001; see Gemenne 2012; Morrissey 2012). At a time when migrants and refugees increasingly came under attack in Northern countries, the forecast of millions of environmental refugees was thought to fuel the anti-immigrant backlash. Notwithstanding this critique, scholars of international law (often of international environmental law rather than of refugee or migration law) advocated various legal schemes in order to offer environmental migrants or refugees protection under international law, through the extension of existing refugee conventions or the creation of new international legal instruments. Although these proposals have never gained traction in international law, more deormalized norms have emerged, related to the protection of people “displaced by natural hazards,” such as the Nansen Initiative, jointly launched in 2012 by the Norwegian and Swiss governments. This muddled legal situation is compounded by jurisdictional conflicts over environment and migration between various international organizations such as the United Nations High Commissioner for Refugees (UNHCR) and the International Organization for Migration (IOM), which compete with each other for funding, expertise and legitimacy in this area.

Political scientists have sought to overcome the so-called fragmentation of international environmental governance, and accordingly have devised various institutional forms to govern environmental migration at an international
level. Sometimes these legal and normative endeavors were justified with reference to Northern countries’ responsibility for causing dangerous anthropogenic climate change; taking in refugees was presented as a form of compensation for the environmental harm thus caused. Other scholars eschewed the language of responsibility altogether and merely emphasized the humanitarian needs of environmental or climate refugees. Some scholars have reflected on the legal and normative implications of the physical disappearance of a sovereign state due to climate change (the notion of a sinking island) and the fate of its deterritorialized citizens.

The increasing characterization of environmental refugees in terms of a security issue, or even a threat, is a recent concern of some political scientists and scholars of international relations (White 2011; Bettini 2013). The catastrophist discourse of an impending environmental refugee crisis has been seized by various security actors in the NATO (North Atlantic Treaty Organization) states and in international organizations. In this narrative, environmental degradation, especially climate change, will trigger mass displacements of impoverished populations from Southern to Northern countries. Such visions build on and elaborate some powerful narratives which gained momentum in the aftermath of the fall of the Berlin Wall, when environmental change, rather than Soviet-style communism, became the common enemy of capitalist societies. In the early 1990s the idea that environmental conflicts would be triggered by increasing resource scarcity and environmental degradation started to be heavily discussed in policymaking circles as well as in academic publications. Some scholars explored the possibility that environmental change could lead to violent conflicts by creating human displacement. Such neo-Malthusian perspectives also circulated in more popular outlets and in the mass media, and they remain part of the political “common sense” about climate change. For instance, in 2007 the Secretary-General of the United Nations, Ban Ki-moon, suggested that the violence in Darfur was attributable to resource scarcities caused by climate change.

By framing climate change and environmental degradation as a threat to national (state) security, these actors have sought to “securitize” environmental migration; that is, turn it into a security issue. Traditionally, security matters are considered as more important than “softer” issues such as the environment. By securitizing the issue of environment and migration, these actors have sought to bring it into the realm of high politics and to the attention of policymakers. But, precisely because it is framed as a national security issue, this move implies tighter control of international migration, the reinforcement of borders, more authoritarian governance, and the reassertion of an exclusivist understanding of sovereignty (that is, a neat separation between insiders and outsiders). For instance, some semiperipheral states strategically positioned in the transit of migrants from Sub-Saharan Africa to Europe seem keen on reinventing themselves as migration managers vis-à-vis the European Union in a context of growing environmentally induced migrations. Thus the securitization of the environment and migration has certainly given these issues a higher profile in international and national politics, but this may be detrimental to the situation of the migrants themselves, who are cast as a threat to state security (White 2011).

**Current research in environment and migration**

The literature discussing linkages between the environment and human migration has grown exponentially since the mid-2000s and includes
different focuses and methods. It is possible to identify different types of study within the field.

Quantified estimates of existing or future climate or environmental refugees

As mentioned earlier, these studies are often done by projecting the forecasted impacts of climate change onto population growth (e.g., Myers and Kent 1995). Although these studies are now widely discredited in the scientific literature for their reductionist understanding of human migration, they have proved instrumental (even if problematic) in drawing attention to this issue, and often remain the source of much public debate on this topic, as their estimates of tens of millions of climate refugees regularly resurface in the media, or in reports by international organizations or NGOs.

Conceptual clarifications, typologies, and literature reviews

These are quite prominent in the field. Important distinctions have been offered on the causes of migration and the forms of migration, and on the analytical and political consequences thereof. Classical distinctions were made between catastrophic events (e.g., earthquake) versus slow-onset degradation (e.g., desertification), between long-term and short-term migrations, between forced and voluntary migrations, and between local, internal, and international migrations, and so on. Yet, these typologies were often not backed up by dedicated empirical research and have been criticized for not conceptualizing the mechanisms through which environmental change actually leads to decisions to migrate. More recent theoretical interventions have, instead, insisted on the multicausal nature of migration and on the complex linkages between the environment and other drivers of migration.

Empirical quantitative studies

These studies try to assess the relationship between environmental variables and human migration based on data aggregated at a national or regional level. Various methods and sets of data are used, and generally other variables (economic, demographic, social, etc.) are controlled for. The usual debates on the choice of methods, variables, and indicators apply. Some researchers use existing datasets while others also produce their data, for instance through household surveys, interviews, focus groups, and mixed methods.

Some studies are interested in the interaction between the exposure of specific places to environmental stress or hazards and the adaptive capacity of the population; both exposure and adaptive capacities are unevenly distributed and migration is only one possible outcome of a specific set of socioenvironmental relations. Environmental stress tends to reinforce dominant social relations and to shape migration decisions through complex political-economic processes.

Other authors posit migration as a risk management strategy when households are faced with uncertainties stemming from environmental changes. From this perspective, temporary migration increasingly appears as a coping or an income diversification strategy in rural areas faced with environmental uncertainties such as rainfall variability, but it is restricted to households with sufficient resources. Those moving may increase their adaptive capacity, thus becoming more resilient. On the other hand, poorer (or more vulnerable) populations may simply lack the resources for such a strategy and thus become trapped in their territory. The normative challenge from this perspective is to lower the barriers to migration for vulnerable households.

Some researchers are turning to agent-based migration models, which simulate autonomous
actors’ decisions and interactions in order to understand the aggregate effect of such interactions. Such models are held to be better at predicting future migration patterns than simple projections, yet they do not explain migrations. Similarly, past migration patterns are being re-examined (e.g., by means of event history analysis) in order to take into account possible environmental drivers of (non)migration. These studies tend to confirm that households respond quite differently to environmental change and that migration is by no means the favored option.

For instance, in their quantitative study of out-migration in a valley in Nepal, Bohra-Mishra and Massey (2012) use event history analysis in order to account for the impact of environmental degradation (mainly deforestation) on human mobility while controlling for economic, social, and demographic variables. Their essential conclusion is that, while environmental degradation is indeed associated with greater mobility, such mobility remains strongly local within the valley investigated, rather than internal or international. While the authors are careful to restrict their conclusions to their specific case (one involving slow degradation of resources), nevertheless they point out that the quasi-absence of international migration as a result of environmental degradation may simply be because those affected lack the resources to undertake such migrations.

Empirical case studies of specific places/regions

These studies tend to be of a more qualitative sort and are generally influenced by a political ecology perspective. Depending on the particular theoretical inclinations of the researchers and on the subject at hand, social, political, and historical factors such as unequal statuses, social positions and power, access to resources, social relations of production, land tenure, sexual division of labor, local institutions, livelihoods, and so on are given explanatory priority over environmental determinism. Within specific regions, not all inhabitants have the same access to resources or are equally able to shield themselves from natural hazards. A closer study of the local structure of power is warranted in order to understand these differentiated vulnerabilities and their relation to whether or not it is possible for some groups to move. Indeed, power inequalities (resulting, for instance, in denial of land tenure), rather than environmental stress per se, impact the propensity to migrate. Environmental transformations are merely one of several constraints and opportunities which these studies take into account in order to explain human (im)mobility, but also the identity, self-understanding, and representations of the subjects of their research.

Emphasis on the structure of power leads researchers to point to the ability of dominant groups to create security for themselves being predicated on their creating insecurities for other groups. Migratory practices are highly dependent on such socially, economically, and politically created (in)securities. Indeed, environmental stress may lead to forms of migration that in turn reinforce the power structures that were responsible for the uneven vulnerabilities of inhabitants in the first place (e.g., by further concentrating land property into the hands of a minority).

Because these studies insist on the local particularities of their case studies and their specific complexes of social, political, and economic relations, they tend to avoid drawing general conclusions from their empirical findings. Such studies often avoid quantifying in- and out-migration, and are more interested in underlining the socioecological relations at work in specific settings. For instance, Carr (2005) discusses how environmental and economic changes in three villages in rural Ghana (notably
changes in off-farm employment opportunities because of environmental degradation) have led to changes in gender relations within the household, and how, in turn, these changes have influenced a pattern of out-migration over nearly four decades.

Some authors argue that the dominant apolitical framing of the environment–mobility nexus is actually blind to the issue of power, and may thus advocate solutions that could reinforce existing inequalities and oppressions. These sociopolitical features are overlooked when environmental drivers are accorded explanatory primacy, and managerialist solutions are offered. In such cases only structural changes can resolve the crisis. Furthermore, climate change or environmental degradation may be instrumentalized by powerful actors, such as state actors or large landowners, in order to obscure the real (economic and political) reasons for displacing people or resettling them onto less productive land.

Normative, legal, and institutional questions

Contributions from political scientists and international legal scholars remain important in this field. Yet, whereas earlier contributions took for granted the catastrophist account of millions of environmental refugees and sought to offer them legal protection in international law (as discussed earlier), more recent research focuses on improving existing legal standards, “soft” laws, and the governance processes that deal with displaced people. Grand schemes for governing environmental migration at an international level are becoming less prevalent, and more piecemeal approaches are on the rise.

Discursive analyses

An important feature of recent publications on migration and the environment is the growth of discourse analysis, which seeks to reveal the hidden assumptions of much of the literature on this topic. This work by critical geographers, often underpinned by a radical or poststructuralist perspective, has helped to redefine and problematize the whole issue. The catastrophist tone of the early debate had led to both national security and humanitarian perspectives, both of which actually work to deny migrants’ agency and to obscure alternative futures. They therefore act as depoliticizing devices which impede transformative practices. Climate or environmental refugees have sometimes been pictured as passive victims, but sometimes also as a threat to Western modes of life and consumption, thereby reinforcing racial and gender- and class-based stereotypes. Furthermore, framing migration as a strategy of adaptation to climate change reinforces the dominant neoliberal perspective in contemporary environmental governance (Bettini 2013; Felli 2013).

Phenomenological analyses

Finally, some studies are starting to explore the lived experiences of those compelled to move because of environmental transformations. Geographers have shown that the proverbial inhabitants of “sinking islands” in Tuvalu or Kiribati did not essentially picture themselves as potential climate refugees but rather, when confronted with the effects of climate change on their environment, sought to find ways to mitigate climate change and to stay on their land (McNamara and Gibson 2009). Migration is not the primary choice for those faced with environmental degradation. Instead, concerns about “sinking islands” express Western anxieties about climate change rather than the actual experience of their inhabitants, and thus betray a neocolonial gaze.
A potentially fruitful avenue for research may have been opened by Rob Nixon, a scholar of environmental humanities, and his concept of “displacement without moving,” which describes the estrangement caused by environmental degradation, including in the absence of migration. Nixon (2011, 19) discusses a more radical notion of displacement, one that, instead of referring solely to the movement of people from their places of belonging, refers rather to the loss of the land and resources beneath them, a loss that leaves communities stranded in a place stripped of the very characteristics that made it inhabitable.

SEE ALSO: Climate change adaptation and social transformation; Climate and societal impacts; Environment and development; Environmental determinism; Labor migration; Migration: forced; Migration: international; Refugees; Relevance debates; Social vulnerability and environmental hazards

References

Further reading


Environment and regional development

Clint Ballinger
Independent scholar

From the earliest geographical writings through to the beginning of the twentieth century, it was common to attribute differences between social regions to variations in environmental and geographic factors such as climate, soil types, or terrain differences, for example, from Strabo, al-Jahiz, and Ibn Khaldun, through to Friedrich Ratzel, Ellsworth Huntington and Ellen Churchill Semple (note that, in economics, examples of Ricardian comparative advantage also often hinge on environmental differences). However, for a variety of reasons, including associations of this type of geography with colonialism and imperialism, determinism, and relying on overly simplistic correlations, it became controversial by the early to mid-twentieth century and fell from favor in human geography.

After this period human geography was dominated by a succession of intellectual approaches: descriptive regional geography from the 1920s, abstract spatial science from the 1950s, and the various social “turns” from the 1970s to the present: radical, cultural, postmodern, poststructural, cultural economy, and related approaches. Crucially, however, although these approaches varied substantially on many issues, they nevertheless shared the view that environmental factors could not or should not be used to explain variations in regional development. They rejected environmental factors in explanation of social outcomes as either overly simplistic or deterministic or both.

However, in the 1970s and 1980s, in part building on the tradition of Fernand Braudel and similar historical work, a number of influential historians such Edward Whiting Fox, Alfred Crosby, Philip Curtin, and William McNeil integrated environmental factors into their explanations of uneven development. These works were built on in the 1990s by other nongeographers, perhaps most notably Jared Diamond (1997) and economist Jeffrey Sachs, who, along with colleagues associated with the Center for International Development at Harvard University, produced a series of influential working papers that in turn generated significant additional research into the impact of geographic factors on regional development differences. These works by historians and economists were significant for specifically integrating, as causal variables, environmental factors such as spatial patterns of endemic disease, soil quality, climate, terrain, landforms (from continental to local), and other environmental factors into their explanations of variations in regional development.

These modern theories differed in important ways from the earlier use of environmental factors. Critics of earlier uses of environmental factors argued that they were overly simplistic because (i) they often attributed differences in economic and political development to environmental factors alone and (ii) they often made extremely direct associations of climate with human behavior and social outcomes (e.g., the
tropics made people lazy and therefore their societies were undeveloped).

However, modern research seeks to integrate environmental factors into the complex social science understanding of regional differences. The effects argued for are not simple and direct, but are due to the cumulative effect over long time periods of relatively small constraints and incentives on individuals that may have a cumulatively large spatial influence on development outcomes. Some of these explanations involve relatively short-term, direct influences, such as the direct social costs of poor soils or tropical diseases, and possibly the amplification of these through the effect of the resulting poverty and mortality rates on schooling decisions, population levels, and ability to invest in further economic growth. Others are longer-term and/or more socially mediated. For example, rich agricultural areas generally experienced substantially higher degrees of urbanization for longer periods, which in turn fostered the development of institutional structures that in the long run may have been conducive to modern economic development, and that are lacking in areas or cultures that did not experience such long urban histories. Similarly, climate influences the spatial patterns of endemic diseases, which in turn influenced the spatial patterns of migration and colonialism, and thus the spatial distribution of institutions and regional development.

Environment and explanation

The role of the environment in explaining regional development differences is a specific example of the more general philosophy of science question on the role of context in explanation. In physics, there are laws and constants that are regular, such that \(x\) always causes or relates to \(y\) with no exceptions. Discovering these was a crucial step in the development of the sciences, and became a cornerstone in the concept of explanation. Another fundamental concept of explanation is that of description and categorization. The natural sciences, for example, made great advancements not through the discovery of laws of nature but through the acquisition of new data and the systematic classification of these data. These two extremes—nomothetic (“lawgiving” approach) and idiographic (“writing about the unique”) formed much of the foundation for concepts of explanation from the early development of the sciences and social sciences through the twentieth century. This was reflected within human geography by recurring debates on the proper role of nomothetic versus idiographic approaches to geography.

However, among practitioners of the historical sciences such as astronomy, cosmolology, geology, biology, paleontology, and history, the concept of explanation increasingly transcends the nomothetic–idiographic debate. In these fields events do not of course break the laws of physics, but neither can they be explained by them. Explanations in the historical sciences rely on antecedent spatiotemporal facts. For example, in biology, a major question concerns the reasons for the K–T event, the extinction of the dinosaurs. The answer seems to be a meteorite impact and/or the global effects of the Deccan lava traps of India. These are unique spatiotemporal events that altered the context of biological processes, changing their subsequent development trajectory and modern biogeography.

Variations in the spatiotemporal contexts of social processes may, likewise, have been the underlying cause for modern spatial variations in development. This may be especially true given the path-dependent nature of development,
whereby relatively small changes in context can result in widely divergent outcomes. For example, variations in flora and fauna may have affected spatial patterns of urbanization; climate-related endemic disease patterns and poor soils may have affected agriculture and, in turn, the related institutions of what was for many centuries the primary means of production of many societies; disease patterns are known to have deeply affected migration patterns and, thus, the modern distribution of political and economic institutions. The importance of ongoing knock-on effects from the role of environmental factors as “spatial anchors” in spatial explanation is discussed in greater detail in Ballinger (2011).

Criticisms of neo-environmental determinism

The renewed integration, often by nongeographers, of environmental factors into explanations of social processes and regional development outcomes has largely been received negatively by academic geographers. This is evident in the significant number of articles by geographers specifically and strongly criticizing the use of environmental factors in explanations of regional development, what they frequently term “neo-environmental determinism” (e.g., Merrett 2003; Sluyter 2003; Coombes and Barber 2005; Judkins, Smith, and Keys 2008; O’Keefe et al. 2010; Radcliffe et al. 2010). Overall, these criticisms by geographers have several common themes, the most important of those relating to environmental explanations being (i) that they are simplistic, that is, there are not simple causes between some factor such as a warm climate or poor soil and some social outcome such as poverty; and (ii) that they are deterministic, when it is argued that the world, or at least the social realm, is indeterministic and that determinist views are immoral or unethical because they preclude or downplay human agency.

Replies to criticisms of environmental factors

There are a number of replies to these criticisms. In relation to the charge of being overly simplistic, modern studies do not argue that the environment alone is somehow all-determining, nor do they usually posit direct effects of climate on individual behavior. Modern research often focuses on indirect effects and long-term, cumulative, and path-dependent outcomes, and integrates these into the modern complex understanding of society and social change.

Overall, modern studies that integrate environmental factors into their explanations of regional development differences deal with a greater number of factors—environmental as well as sociocultural, political, and economic—than those that exclude environmental factors on a priori grounds. Hence they can be argued to be more, not less, complex than approaches that a priori exclude the large set of environmental factors. It might also be argued that these types of studies are more complex because they deal to a greater degree with the complexity arising from the interactions of social processes that operate on very different scales, the very large-scale and long-term with the local and short-term.

There are a number of possible objections to criticism based on determinism, including the following.

1 It is not clear why the use of environmental factors in geographical explanation is “deterministic” (however defined) when
other factors are not. In the nature versus nurture debate in psychology, for example, a perceived over-reliance on innate biological or genetic factors in the explanation of an individual’s behavior is often characterized negatively as “deterministic” and as needing to be counterbalanced by a greater emphasis on nurture (environment). It is not clear why the nature/nurture argument should reverse on a larger-scale view of explanation, that is, why environmental factors do not have larger-scale cumulative effects on society if they also have strong influences on individuals.

2 If environmental factors are indeed used in a deterministic way, this is not necessarily an error in explanation. A common criticism of neo-environmental determinism is that the world, or at least the social realm, has been proven to be indeterministic, so deterministic arguments must be flawed. However, this rests on the assumption that determinism has somehow been disproven. Yet, in the fields most closely related to the study of determinism – physics, philosophy, and biology/neurology in the life sciences – there is no consensus on the ontological status of determinism/indeterminism (Ballinger 2008a and citations therein).

3 In criticisms of the use of environmental factors as “environmental determinism,” it is frequently implied that this somehow makes these explanations immoral or unethical. However, among philosophers, there is no consensus on the moral implications of determinism (Ballinger 2008b).

Crucially, in points 2 and 3, critics are making claims to knowledge that they cannot possess unless they claim special knowledge concerning determinism, and/or the ethical implications of determinism, beyond that of the physicists, biologists, and philosophers who specialize in these questions. Given the current state of knowledge, either indeterministic or deterministic arguments could conceivably be correct, and neither can be dismissed solely for being one or the other.

SEE ALSO: Anthropogeography; Biogeography; Cities and development; Economic geography; Environmental determinism; Historical geography; Path dependence; Political economy and regional development; Regional geography

References


DOI:10.1046/j.1467-8330.2003.00352.x.


Environment and resources, political economy of

Matthew T. Huber
Syracuse University, USA

A curious thing happened over the course of the twentieth century. Political economy, or what is increasingly known as “economics,” stopped paying attention to the relations between capitalism and nature. As Timothy Mitchell (2011, 234) argues, “the availability of low-cost energy allowed economists to abandon earlier concerns with the exhaustion of natural resources and represent material life as a system of monetary circulation – a circulation that could expand indefinitely without any problem of physical limits.” In hindsight, the oversight was relatively brief in historical terms. The age of “abundance” in the postwar mid-twentieth century was “shocked” into confronting the deep, and precarious, entanglements of capitalist economies with flows of energy, materials, and wastes during the 1970s. Since the 1980s, the discipline of economics, and several other social sciences, most notably geography, have attempted to provide theoretical clarity and empirical detail on the precise relations between capitalism and nature. This entry will attempt to review the varied ways in which this relationship has been conceptualized over the past three decades or so.

The entry will proceed in four parts. First, in a theoretical introduction, it will argue that the major perspective on nature and capitalism in geography can be traced back to Neil Smith’s (2009/1984) landmark “production of nature” thesis. Second, it will review the historical context of neoliberalism and the rise of free market environmentalism since the 1970s – and geography’s critical engagement with such policy. Third, it will argue that theories of economy–environment relations in geography often were contingent upon which part of nature was being examined – ecosystems, resources, energy, and biological life itself. Those focused on the ecological “conditions” of production tended to employ theories based on Karl Polanyi’s (1944) concept of “fictitious commodities” and James O’Connor’s (1998) “second contradiction of capitalism.” Critical resource geography built upon Erich Zimmerman’s (1951, 15) famous pronouncement, “resources are not; they become,” to examine the material and sociocultural politics of naming certain parts of nature as resources. Increasingly, those who study energy have focused on the centrality of energy to the social reproduction of capitalism as a whole. This approach sees “nature” as internal, rather than external, “conditions” or resource “inputs.” Finally, this entry will cover recent debates over the relationship between capitalism and “life itself” – that is, new regimes of accumulation based upon the sociotechnical production of biological life forms.

Theory: nature as socially produced

“What jars us so much about the idea of the production of nature is that it defies the conventional, sacrosanct separation of nature and society, and it does so with such abandon and without shame” (Smith 2009/1984, 7). It could
be argued that over the last three decades critical environmental studies have been wholly focused on breaking down this “sacrosanct separation of nature and society.” Certainly Neil Smith was not the only (or first) one to offer this critique – feminist (Merchant 1980; Haraway 1991), actor-network (Latour 1993), and critical race theory (Moore, Pandian, and Kosek 2003) have also done this work – but what Smith’s “production of nature” approach offered was a decidedly political economic approach to understanding the relations between nature and capitalism as a historically specific mode of production. As Smith (2009/1984, 7) pointed out, “It is capitalism which ardently defies the inherited separation of nature and society, and with pride rather than shame.”

Smith’s contribution centers upon the fundamental tension in capitalist societies between use-value (e.g., direct use of nature for subsistence) and exchange value (e.g., the extraction of natural resource commodities for profit). For Smith, only in capitalism, where wage-labor becomes generalized and accumulation becomes the overarching goal of production, does exchange value begin to ultimately dominate the social relation with nature. In other words, if capitalist production is bent toward growth and accumulation (and not simply exchange of commodities), this drive necessarily expands the drive of capital to access nature as critical inputs of raw materials and, perhaps more importantly, seek natural sinks for the voluminous (and often inevitable) wastes/pollution from production. There is a geography to this expansion. As value relations tie global commodity markets together, “capital stalks the earth in search of material resources … [T]he appropriation of nature and its transformation into means of production occur for the first time on a world scale” (Smith 2009/1984, 71). For Smith, most controversially, this global scale channeling of nature into accumulation ensures that most parts of the planet we may classify as “nature” have been “altered through human agency” (2009/1984, 81). This might appear self-evident in our current age of “the Anthropocene” and global climate change, but it certainly was not in 1984.

This thesis could be applied to countless empirical “case studies.” A monoculture of cash crops suddenly was seen as a produced landscape – not only through the direct labor of cultivation, but also through scientific knowledge systems that design hybrid seeds, chemical fertilizers, and pesticides. What we imagine as “pristine” nature – say a national park – is itself a political project to create quarantined spaces of “nature protection” as a hedge against urban industrial capitalism’s expanding encroachment – a political project that almost always involved massive violence and displacement of actually existing communities in the name of creating an unpeopled natural space. Ultimately, Smith used the concept of “production” for specifically political aims. If nature is produced, it also means we can change our social relation with nature: “Truly human, social control over the production of nature, however, is the realizable dream of socialism” (2009/1984, 91).

This perspective on the “production of nature” was surely revolutionary in geography and beyond in understanding capital’s relation to nature. Yet, it was also very broad and lacked specifics. Meanwhile, despite its critical insight and efforts to explode the nature–society dualism, its publication in 1984 was concomitant with a more mainstream way of thinking about the economy and the environment as separate. For an insurgent camp of economists there was the discovery of a whole host of environmental processes and functions that were conceptualized as “outside” the market system.
Historical context: neoliberal natures

Since the 1970s, environmental social movements and forms of knowledge in ecological science have increasingly recognized the planet itself as an interconnected set of biogeochemical processes. On the one hand, what constitutes environmental problems has shifted from predominantly local concerns of landscape destruction or air/water pollution, to include “global” problems of ozone depletion, climate change, and biodiversity. On the other hand, this global ecological outlook has led to increased recognition of a broad set of “services” provided by interconnected ecological systems – from biospheric processes of atmospheric circulation to the hydrological functions of wetlands; from nutrient cycling in soils to the metabolisms of human and nonhuman organisms. Of course, of central importance was the general recognition of the importance of these “services” to the maintenance of life on Earth. Yet, even more important was the idea that these “services” were valuable even if their value was not captured by the market system.

These ideas coalesced around the idea of “ecosystem services” in the subdiscipline of ecological economics, with some audacious scholars even attempting to estimate the global monetary value of such services (e.g., Costanza et al. 1997). While the concept of “externalities” – the idea that the market fails to internalize ecological costs into the price mechanism – has long roots in economic thinking (see Coase 1960), this particular form of critique recognized not only the “costs” of pollution, but also the values provided by ecological systems. Whether one was focused on internalizing the “ecological” costs of various economic activities, or integrating ecological services into market-based logics, the common thread was that market systems could be deployed to solve environmental problems. This ideology created the conditions for a multiplicity of market-based environmental policies: emissions trading (“cap and trade”), pollution taxes, wetland banking, biodiversity offsets, and so on.

For critical environmental scholars in geography it became increasingly obvious that, first, these market-based environmental policies were reflective of the larger political shift toward neoliberalism (Heynen et al. 2007). If neoliberalism as a whole is disdainful toward government regulation, neoliberal environmentalism situated what they called “command and control” environmental rules and sanctions as inflexible, centralized mandates from above. Market schemes were seen as more flexible, allowing the private sector to choose their own pollution abatement schemes, or purchase “offsets” on the market when they engaged in (inevitable) environmental destruction. Second, and more provocatively, McCarthy and Prudham (2004, 276) suggest we need to conceptualize neoliberalism itself as a “distinctively environmental project.” This argument suggests that some concepts at the heart of neoliberalism – such as property rights – found their most logical application in specifically environmental projects such as privatizing water supplies (Bakker 2003). Moreover, just as neoliberalism has been defined through its “roll back” of Keynesian-era regulations and social welfare systems, “assaults on Keynesian-era environmental regulation have been [just as] central…” (McCarthy and Prudham 2004, 278).

There has been no shortage of “case studies” examining the implementation of neoliberal policies in the environmental fields. The definitive collection is Heynen et al. (2007) (see Castree (2008a, 2008b) and Bakker (2010) for discussion of the analytical limits of the empirical “case study” approach). Robertson (2012) examines the contradictions of “wetland banking” as a
project to create homogeneous abstract values for the complex, heterogeneous ecological relationships that wetlands represent. Mansfield (2004) shows how privatizing fisheries (through individual transferable quotas on catches) stands in contradiction to the materiality and mobility of fish themselves. As Castree (2008a, 2008b) has warned, such a focus on “case studies” of particular “natures” subject to neoliberal reforms has led to lack of clarity on the larger explanatory and theoretical conclusions of this literature. A common theme in this literature is not the retreat of the state, but rather the reformation of state power toward creating the conditions for markets to function – via licenses, emission credits, privatization programs, and so on. For the purposes of this entry, one possible lacuna is perhaps linking the historical specificity of neoliberal natures with the broad general “production of nature thesis” reviewed in the previous section. It is clear that the neoliberal natures literature is informed by theories of the production of nature, but exactly how neoliberal capitalism produces nature is left unclear.

One general insight from work on neoliberal natures is that the materiality of nature matters in social efforts to “produce” it in particular ways (Bakker and Bridge 2006). For example, Bridge (2003) shows how the invisible and intangible nature of natural gas sets particular constraints on transport networks. Although Smith (2009/1984) clearly argues “production” does not mean control or domination of nature, he did not pay particular attention to how the materiality of nature together with humans coproduces socio-natures. Moreover, the materiality of nature matters in the kinds of theoretical approaches to understanding nature–capital relations. The rest of this entry will review four different material forms of nature – ecosystems, resources, energy, and biological life – to show how each material form requires different kinds of theories to explain the social production of nature under capitalism.

The nature of capitalism

Ecological conditions of production and the necessary role of the state

As mentioned above, the last four decades have seen science and policy fixate on the concept of “ecosystem services” as a bundled set of “values” that somehow escape the price or market system (see Dempsey and Robertson 2012). While neoliberal economists were somehow sanguine about the market system’s ability to “internalize” such values, critical environmental scholars (including many geographers) tried to think more deeply about how capitalism’s inability to “see” these values represents a larger contradiction. James O’Connor’s (1998) theorization of the “second contradiction” of capital argues that capital systematically destroys its own ecological conditions (in addition to human/labor and communal/infrastructural conditions) precisely because there are no market values attached to things such as wetlands, carbon cycles, and soil fertility. At the heart of O’Connor’s theory is Karl Polanyi’s (1944) notion of nature as a “fictitious commodity” – in other words, many of the resources and life forms capital relies on are not produced like commodities for sale, but rather are products of these larger (and often long-term) ecological and biogeochemical cycles that escape the value form.

The main way in which capital systematically destroys its own ecological conditions is, of course, through pollution of air and water and the exhaustion of resource stocks (whether forests, mines, or soil fertility). For O’Connor, this systematic destruction will ultimately yield an “underproduction” crisis that undermines
the conditions of accumulation for capital as a whole. Another key aspect of the theory is the role of the state as a mediator between capital and the conditions of production (vis-à-vis environmental and labor regulations that limit capital’s self-destructive tendencies). Through environmental regulations, setting aside lands for preservation, and restrictions on extraction, the state attempts to save capital from its own environmentally destructive self. O’Connor’s theory is not only about environmental conditions; following Marx, it is also about personal conditions (social reproduction of labor) and communal conditions (infrastructure and other public resources). Yet, O’Connor’s theory of the second contradiction ultimately argues that the state can only “mediate” so much and that it is an inadequate basis for solving the crisis tendencies inherent in capital’s destruction of its own conditions.

O’Connor (and Polanyi’s) theory is useful to an extent. For example, McCarthy (2004) has used it to understand how the conditions of production become privatized through trade agreements which protect capital’s right to pollute and avoid environmental regulation. Essentially, the plethora of work on various efforts to “marketize” ecological services, such as the carbon cycle through REDD (Reducing Emissions from Deforestation and Forest Degradation) and other forestry projects, wetland banking, and biodiversity offsets, is all examining the conditions of production in O’Connor’s terms. Yet, this theory has its limits for understanding the full spectrum of nature–capital relations. First, it systematically reproduces the nature–society dualism by theorizing “nature” as external to the economy and capital (O’Connor literally uses the words “external conditions of production”). (To be fair, O’Connor might argue it is capital that views nature externally, not him.) Second, this “external” framing ignores the parts of nature that are internal, either as resource commodities or as internal “natural” processes central to the production process itself. These two will be discussed over the next two sections.

Resources: becoming a commodity

Of course, some parts of nature do become commodities. They are called “natural resources.” Indeed, neoliberal environmentalism is itself a project to turn the entire wealth of the environment into “resources” called “ecosystem services.” But, what parts of nature become resources? A strict Marxian approach might flatly say those parts of nature that require expenditures of human labor to extract become “values” (i.e., commodities). While the factory may harness “natural” falling water for free, it must pay by the ton for coal mined from the Earth by human labor. It also must pay for the waterwheel (crafted through timber, iron, and manufacturing sectors) that harnesses this “free” falling water.

A more nuanced view of “natural resources” acknowledges that it is not only strict labor expenditures that determine the commodification of nature, but also culture, historical circumstances, and systems of knowledge. (Indeed, the nuanced view would see all these factors as inherently part of “labor.”) This is summed up by resource geographer Erich Zimmermann’s (1951, 15) famous line, “resources are not; they become.” Take the example of crude oil. It is now seen as the most important commodity fueling the entire global economy. Its central refined products are gasoline and other transportation fuels. Its “use value” – its status as a natural resource – is a product of culture in the sense that systems of automobility are themselves wrapped up in ideas of freedom, national identity, gender, and race (Huber 2013). History matters immensely as well. Historians
of oil always point out that ancient cultures used crude oil for everything from waterproofing and medicine to embalming and flammable weapons. Even in our current history of the Industrial Revolution, crude oil has become useful historically in two different contexts. First, crude oil was a source of illumination (kerosene) before the advent of electric light (and just after the near exhaustion of sperm whale stocks). As a matter of interest, another refined product, “gasoline,” was seen as a dangerous, volatile, and flammable substance that was best deposited into local waterways as a “waste” product. Second, it was during the twentieth century, of course, that the age of automobility transformed oil into a source of transportation fuel and, eventually, a hydrocarbon stew of molecules that could be manipulated into petrochemicals.

Of course, the “becoming” of oil as a resource is also directly a product of systems of extraction and knowledge systems. In recent years, most commentators agree that cheaper-to-extract oil (easy oil) has become harder to find. Thus, oil prices boomed from 2005 to 2014 before collapsing under the weight of mass production in the United States, among other places. The “boom” of high oil prices made difficult-to-access oil (tough oil) economical to extract – that is, new reserves “became” resources (e.g., the tar sands of Alberta reportedly need at least $40/barrel oil to allow profitable extraction). Moreover, as global oil capital has expanded into new frontiers of extraction (e.g., deepwater offshore, tight oil accessed through hydraulic fracturing), so have knowledge systems of petroleum engineering and computer-based exploration techniques.

Several resource geographers have pointed to the historical, social, and ultimately political relationships which shape the extraction and use of resources. A recurrent theme in the literature is that no social form of resource appraisal is purely social – resources are always defined through material forms and processes (Bakker and Bridge 2006). Despite this turn toward the material, and similar theories around ecological conditions, there persists a curious dualism between social (cultural, economic) systems that “define” resources and the material, biophysical aspects of those systems themselves. Perhaps this is because much of this work in resource geography is centered upon the moment of extraction – that is, the moment where “raw nature” and human labor appear to be mostly distinct. If we start to examine the ways in which resources flow throughout society, it becomes much more difficult to maintain these clear nature–society divisions. The recent work on energy in geography has developed a more “internal” theory of economy–environment relations.

Energy and the social reproduction of capitalism

As reviewed above, most approaches to the political economy of the environment approach nature as either a set of processes external to market valuation (ecological conditions) or a discrete set of commodifiable “things” that attain the status of “resources.” The problem with such approaches is they imagine a wholly social world of “capitalism” confronting some external realm of ecosystems and resources. But, as Moore (2011, 34) has recently forcefully argued, we need to see capitalism itself as already fused with ecological relations – “Capitalism doesn’t have an ecological regime; it is an ecological regime.” In other words, too often we imagine a “social system” of factories or mining enclaves extracting from or degrading a “natural” system.

Scholars at the intersection of feminist and political economy have recently focused on the concept of “social reproduction” to explain
the political struggles involved on the terrain of everyday life – simply “reproducing” lived existence entails struggles over food, transportation, care work, and what is increasingly called the “work–life” balance (a capitalist social construction if there ever was one) (Mitchell, Marston, and Katz 2004; Huber 2013). It is this terrain of everyday life that Lefebvre (1973, 89) referred to as, “the very soil on which the great architectures of politics and society rise up.” Therefore it is an important concept to understand the “internal” ecology of capitalist social relations.

When we approach energy it is obvious why social reproduction is a useful concept. Energy systems are ingrained in the basics of lived existence. Whether the focus is on peasant struggles for wood (fuel) resources, or industrial consumers calling for public ownership of electric utilities, energy systems are also usually highly politicized. Of course, energy is not the only topic one can approach from this perspective. Timber products, minerals, and food (also a form of energy) are all deeply embedded in the reproduction of everyday life and capitalist social relations. The point is, this kind of “ecological” analysis is different from that which might look at the conditions of extraction or environmental degradation. Moreover, particular conditions of extraction and degradation are relationally linked to these same geographies of social reproduction. Sometimes political struggles over/against forms of extraction neglect these connections. This was David Harvey’s (2003) critique of politics centered on “accumulation by dispossession” (which often are environmental struggles against dispossession of land and resources). He argued that such a politics must link with struggles on the “inside” (for lack of a better term) of the “expanded reproduction of capital.”

A focus on the ecologies of social reproduction can also emphasize the cultural and political aspects of resource or energy consumption. As several authors have recently pointed out (e.g., Boyer 2011; Mitchell 2011; Huber 2013), the very basis of ideas of modernity, democracy, and freedom emerge out of socioecological regimes of massive fossil fuel consumption. Chakrabarty (2008, 208) puts it most succinctly: “The mansion of modern freedoms stands on an ever-expanding base of fossil-fuel use.” This approach, what Boyer (2011) terms “energopolitics,” sees energy as central to wider forms of politics. Yet, even with these “internal” social reproduction approaches there still persists a kind of nature–society dualism where the world of “things” (energy resources, food, materials) is seen as formative in creating “social” worlds of urban, industrial life. Part of the problem is that these natural “things” are too often imagined as “inanimate” objects such as iron ore and coal, but are also often living biological systems (such as wheat fields, forests, fisheries). Examining the intersection between capitalism and biological life raises different questions.

### Biological capital

Even the “production of nature” thesis does sometimes revert to dualistic thinking: capitalism does the producing; nature becomes the product. In reflecting on the thesis over 20 years later, Neil Smith (2006) suggested the production of nature has intensified even more. Through the manipulation of “life itself” via biotechnology, DNA coding, and other genetic tools, Smith suggests the production of nature has proceeded “all the way down” to the very basics of biological existence. Boyd, Prudham, and Schurman (2001) developed an interesting theoretical approach to this shift. Following Marx’s distinction between the formal and real subsumption of labor under capital, they suggest two forms through which nature is subsumed under capital.
The formal subsumption of nature is a basically “extractivist” model wherein capital extracts “things” from nature such as trees, minerals, crops, and so on. On the other hand, the “real” subsumption of nature occurs when capital actively produces the biological content of certain resources themselves. Thus, over the last decade, some major theoretical work has grappled with the relations between capital and the broad field of biotechnology and efforts to produce “life itself” (see Rajan 2006; Cooper 2008).

Many have observed the capitalist production of biological life. Kloppenburg’s (1988) classic work examines not only the historical geography of the genetic manipulation of seed technology, but also how control over seeds allowed for the concentration of wealth and power for capitalist agribusiness. Prudham (2005) looks at how the process of commodification in forests also creates particular strategies for producing plantation-style forests. Kosek (2011) examines the honeybee as a critical “technology” in the production of other kinds of life in agriculture. Life is also examined via struggles over viruses, bacteria, and what Braun (2011) calls “the molecularization of life.” Of course, the ultimate way to break down the nature–society dualism is to examine the human body itself as a particular kind of produced nature via foods, chemicals, genes, and sexual reproduction (Guthman and Mansfield 2013). The politics of the body often reproduces larger power relations and geographies of inequality and difference based on race, gender, and class.

**Conclusion**

In the end, it should be re-emphasized that the problem of segmenting nature from society is not even primarily a problem for academic theorists or intellectuals. It is a problem that is itself produced through capitalist social relations. It is capitalist value systems that do not take into account larger “ecological” processes—and appropriate those systems for “free.” It is capitalists who view resources as a stock of discrete “things” that can be transformed into commodities and, ultimately, money. It is capitalism that requires the majority of us (who are wage workers) to rely on a bundle of commodities torn from nature simply to reproduce our lives. It is capitalism that is now looking toward genes, nanotechnologies, and biological life itself as a new frontier of accumulation. Moving forward, geographers will continue to do research to uncover the illusory nature of these nature–society divides that capitalism insists upon.

**SEE ALSO:** Commodification of nature; Energy resources and use; Marxist geography; Neoliberalism; Political ecology; Socio-nature

**References**


ENVIRONMENT AND RESOURCES, POLITICAL ECONOMY OF


State environments and environmental states

The Somerset Levels provide a valuable geographical setting in which to begin our discussion of environment and the state. The Somerset Levels is a coastal plain and wetland area that is located in the southwest of England. As with many low-lying coastal areas, the Somerset Levels are prone to flooding. Since medieval times the Somerset Levels have been synonymous with attempts to settle and cultivate the area alongside regular inundation events. The heavy storms that hit the UK during late 2013 and early 2014 served to bring local flooding in the Somerset Levels to national attention. Significantly, in the context of this entry, these floods exposed some of the complex ways in which states and the environment are connected.

In total, the flooding of the Somerset Levels affected 35 homes, 17,000 acres of land, and led to the closure of a busy section of the national rail network for two months. The severity of flooding resulted in many villages being cut off and having to be abandoned for days on end. While the British Prime Minister was quick to establish a connection between the storms and floods that affected the Somerset Levels and climate change, public and media attention quickly focused on the role and responsibility of the state in managing flood events. At a local level, the UK government’s Environment Agency was criticized for failing to adequately dredge the rivers that run through the Somerset Levels. At a national level, the flooding of the Levels (and the wider storms of which they were a part) led many to question the 25% cut in flood defense spending the Coalition Government had imposed in 2010 (Carrington 2014).

The case of the Somerset Levels exposes the complex ways in which states and the environment come together. States are seen to have a responsibility for protecting national environmental assets – such as the farmland of the Somerset Levels – as part of its broader remit to support economic development. The UK state has also been instrumental in establishing the 32 sites of special scientific interest that are scattered throughout the Levels. These sites have been formed to protect the flora and fauna of this wetland environment. While protecting the flora and fauna of the Levels, the local and national state has also been instrumental in helping to drain large sections of the Levels so that they could be put to productive use. The recent flooding of the Somerset Levels also reveals the state’s assumed responsibility for protecting people from environmental risk and perturbations, such as storms, floods, and inundations. In these contexts it becomes clear that the state occupies a difficult position between the priorities of environmental conservation and securing the resource needs of its population, and of protecting the environment from people and people from the environment.

This entry seeks to explore in greater detail the nature of the relationship between states and the environment. In doing so it will reflect
ENVIRONMENT AND THE STATE

upon the history of state–environmental relations in different geographical contexts and introduce key theoretical frameworks that can help us understand these relations in new ways. Because there are many different ways in which both the state and environment can be defined, this entry explores the various ways in which it is possible to understand states and environments and how these different perspectives can enable us to see the various ways in which environments and states are connected.

While this entry deliberately avoids firm definitions, it is helpful to propose some loose frames of reference for understanding what is meant by state and environment. This entry understands states as collections of institutions (such as parliaments, departments, and courts of law) which tend toward forms of central coordination, have a responsibility for a territorially demarcated area, and maintain monopolies in relation to rule making and the legitimate use of physical force (see Mann 1993, 185). While this may be a useful working framework within which to comprehend states, it is important to note that many argue that the operation of states is not limited to the work of formal state institutions (in the case of the Somerset Levels, for example, the area was initially drained by private interests who were working under license from the king); that many forms of state power do not radiate out from a center of power; that state influence is neither absolute within a territory nor limited to by territorial borders; and that legitimate rule making and violence are no longer the exclusive preserve of states.

The environment can be understood in two broad ways. First, environments can often simply refer to contexts and can take the form of built/manufactured structures (cities, houses, offices, roads) or more natural settings (forests, wetlands, savannah). More usual, when people refer to environmental issues they have in mind specific ecological understands of the environment. Second, notions of ecology (or the economy of nature) tend to see environments less as contexts and more as processes. Ecological processes could, for example, refer to the greenhouse effect, nitrogen fixing, or the water cycle. It is important to note that when people refer to the environment in these more ecological terms there can be a tendency to understand related processes in isolation from human activity (a so-called external views of nature, see Smith 1984). What is now clear, however, is that in the context of growing human influence on ecological systems (from climate change to the acceleration of the nitrogen cycle that is product of the global use of inorganic fertilizers), ecological systems cannot be understood as separate from the human world. In these contexts, this entry understands the environment as both a series of contexts and ecological processes that are the product of both the work of nature and humanity.

This entry explores state–environment relations through three interconnected perspectives. The first section of the entry considers the historical relations between states and environments. In this section particular attention is given to the role of states in shaping human knowledge of environmental systems, and the importance of environmental resources and issues in justify expanded forms of state power and influence. The second section of this entry considers the emerging role of the state in managing environmental externalities. Environmental externalities refer to situations were environments and the people who depend on them are subject to forms of pollution and exploitation that have been caused by a third party who does not bear the costs of such activities. This entry will show that environmental externalities have been central to the expanded powers and responsibilities of the modern state. The final section
of this entry considers the constitutional implications of emerging forms of state–environment relations. Particular attention is given in this section to the ecological responsibilities of states and emerging understandings of environmental rights. Throughout these sections, different ways of understanding both states and environments are introduced, but consistent emphasis is brought to the contradictory nature of the historical and contemporary interactions of states and environments.

Environmental history of the state and governmental history of the environment

One way in which it is possible to understand the connections between states and environment is through the lens of history. In many ways, modern understandings of the state and environment share common origin stories. These stories help us understand that centrally orchestrated systems of territorial government were developed in order to exploit environmental systems and resources (such as the Somerset Levels). At the same time, these stories reveal how modern understandings of environments (particularly as scientifically quantifiable objects) are products of emerging state systems.

Prehistories of states and environments

In order to understand the coevolutionary history of states and environments it is important to consider the prehistory of states and environments. The term prehistory referred to here means the periods that predate the emergence of modern understandings of states and environments. The term modern is used here both to denote our contemporary “of the moment” understandings and to refer to something more historically specific: modernity. Modernity refers to a scientifically oriented shift in socioeconomic and political life that began in Europe in the sixteenth and seventeenth centuries. This societal shift was broad ranging and affected moral codes, industrial practices, politics, economic thinking, and technological development. At its center, modernity involved the (partial) rejection of religiously oriented modes of understanding of the world and associated theological ways of organizing society.

Before the advent of modernity political society was organized in ways that are very different from contemporary state systems. While power still had a central focus, it was not organized around a bureaucracy, but instead focused on the figure of the monarch or emperor. The prince or Caesar who sat at the center of premodern systems of government justified and secured their power on the basis of divine purpose and their position in the cosmic order of things. The purpose of premodern systems of government was also very different from modern states. While modern states have a governmental responsibility of care toward their people and territory, premodern monarchs were primarily concerned with securing their leadership through the use of military power and influence (Foucault 2007). In addition to its underlying purpose, the geographical form of premodern states was also very different from its modern counterparts. While modern states are based upon clearly demarcated forms of territorial power, premodern states were much more geographically nebulous. Not only were the boundaries of premodern states in a state of regular flux (as military struggles saw borders shift), but there was much more limited understanding of what actually existed within the boundaries of a fiefdom or empire.

Premodern understandings of the environment (or more accurately creation) were also at odds with contemporary sensibilities. Indeed
ENVIRONMENT AND THE STATE

in the premodern world it appears that the category of environment (at least as something that is separate from humans) may not even have existed. Certain premodern worldviews saw nature as a holistic realm of existence through which all things were connected together within a broader religious order of things. On these terms it was believed that the environment could be read much like a book in order to reveal God’s purpose. As a realm of divine creation, premodern environments were places of mystery and discernment, which demanded the respect of humans who were themselves part of the broader web of life of which the environment was a central component.

The rise of modernity: timber, food, and water

The rise of a modern, more scientifically oriented society in the sixteenth and seventeenth centuries had a significant impact on state-environmental relations. The rise of modernity represented something of an existential threat to states. As science increasingly exposed the laws of nature and challenged religious orthodoxies it became increasingly difficult for sovereigns to justify their existence on the basis of divine authority or the eternal order of nature. In this context, states became increasingly governmental in nature and sought a more practical basis for their existence. This practical basis was to be found in providing coordinated care, protection, and sustenance to national populations. This newfound purpose within states had significant implications for the relationship between states and the environment.

In his classic analysis of the emergence of the modern state, Seeing Like a State, James C. Scott exposes how the emergence of modern manifestations of governmental power was intimately tied to the control and management of the environment (1998). In order to consolidate their territorial power and provide care and sustenance for their population, Scott describes how early modern states developed new bureaucratic systems and sciences that enabled them to understand and manage their environmental resources better. These new bureaucratic sciences of environmental management are most clearly expressed in the emergence of modern state forestry. Timber resources were vital to early modern states. Timber was a key source of fuel, was widely used in building, and, through its use in shipbuilding, was a vital part of a nation’s commercial and military security. But premodern forests presented thickets of confusion for state administrators. The organic nature of woodland made it very difficult for states to effectively estimate the amount of timber resources they had in reserve. It was in this context that governments supported the emergence of more scientific systems of forest management (Scott 1998). Scientific forestry involved the gradual replacement of organic woodland with the regimented spaces of tree plantations. With their geometrically ordered form and straight lines of carefully planted trees, scientific forests could be administered more effectively. At the heart of this administration process was the goal of legibility (Scott 1998). Legibility, or the ability to see things clearly, lies at the very heart of all forms of governmental power (from secret service surveillance to tax returns). Scientific forests meant that woodlands could be measured more accurately and became legible entities within the economic planning of the state (see Demeritt 2001; Whitehead, Jones, and Jones 2007).

Ultimately, making the environment more legible was connected to the procedures through which states made the environment quantifiable. In the context of woodlands, the increased legibility afforded by scientific woodland enabled nature to be more easily quantified. Instead
of vague estimates of available woodlands, state-sponsored scientific forestry was central to the emergence of more universal measures of forestry that indicated the acreage and tonnage of available timber. The scientific quantification of nature in this way was not limited to woodland. Early modern states were also keen to measure agricultural lands more accurately. During the early modern period many states conducted extensive land-use surveys that sought to accurately map the territorial scope of available environmental assets (Whitehead, Jones, and Jones 2007). The mapping of land resources not only facilitated the spatial quantification of the environment, it was also pivotal to the consolidation of states’ control over the territories of which they often only had an opaque awareness. Ultimately, accurate land-use surveys helped states to estimate available agricultural resources and estimate likely harvest yields. They were, however, also vital instruments for the administration of taxation throughout different territories.

What the emergence of scientific forest and land-use surveys have in common is that both changed the ways in which states understand and relate to the environment. Ultimately, they resulted in a narrowing of vision in relation to the natural world as the environment was simplified and quantified. This simplification process is seen nowhere more clearly than in the notion of environmental resources. The accurate quantification of environmental resources such as timber and harvests are sine qua non to the establishment of states as large-scale territorial systems for organizing society. But the very notion of an environmental resource tends to axiomatically define the environment on the basis of its socioeconomic utility. On these terms it is clear that the history of the state is a history that has seen human understandings of, and relationships with, the things of nature become increasingly divorced from the broader web of ecological processes of which they are a part. The history of state–environment relations is essentially a history that connects the emergence of scientific forestry and land-use surveys with the declining awareness shown by many children of what crops produce certain food types.

The modern history of state–environment relations is, however, not restricted to territorial phenomena such as trees and crops. The historian Karl Wittfogel, for example, has drawn attention to the connections between early state systems in Asia and the establishment of complex water management and distribution systems. In his much-discussed book Oriental Despotism: A Comparative Study, Wittfogel argues that the construction of the extensive irrigation systems upon which many Asian societies depend required large-scale systems of bureaucratic coordination (Wittfogel 1957). Such bureaucratic coordination, which involved connecting water supply networks over large-scale areas, required the organization (and manipulation) of cheap labor and depended on large-scale systems of government that were often despotic in nature. In this context, the systems of government outlined by Wittfogel appear to embody a mix of premodern sovereignty and control alongside an evolving bureaucracy that is more typical of modern states. While the despotic states of Wittfogel are very different to modern systems of government, they do reflect precursors to modern states. What is clear, however, is that throughout history states have continued to play a pivotal role in the consolidated supply of water to national populations. In particular, states have been instrumental in the production of supply networks that could carry water from where it was most abundant to where it was most scarce (Swyngedouw 2007). It is interesting in this context to note that during World War II the British government sought to combine efforts to drain the Somerset Levels in order to secure
valuable agricultural land with a plan to facilitate enhanced forms of water supply to a local munitions factory. The building of the artificial Huntspill River in the Levels thus sought to meet draining and water supply needs and was seen as a war priority by the state (Williams 1970).

What these examples illustrate is that the history of the modern state is intimately tied to the environment and modern understandings of the environment are products of their coevolution within territorial governments. To put things another way, histories of human–environmental relations that ignore states, or accounts of the history of states that neglect the environment, are likely to be very partial accounts of the respective pasts of these complex entities.

Externalities: the state and the rise of environmental regulation

Introducing environmental externalities

If the longer-term history of state–environment relations is one that centers on the construction, production, and management of environmental resources (such as timber, food, and water), recent history has focused on another environmental phenomena: externalities. Externalities are actually economic phenomena that pertain to situations where a particular process or transactions causes harm to an innocent third party who receive no form of compensation. Environmental externalities can take two basic forms. First of all they refer to situations when members of a society are subjected to forms of environmental pollution they had no direct role in producing. These forms of environmental externality can be witnessed by those who live along busy transport routes, have made their homes near to waste treatment plants, or who happen to inhabit areas that are downstream or downwind of major industrial facilities. The second type of environmental externality relates to the uncosted exploitation of environmental systems and resources. These forms of externality are produced when chemical plants release their untreated waste products into rivers and oceans or timber companies raze woodlands and destroy the habitats that they constitute at no additional costs. The case of the Somerset Levels provides us with an interesting example of an environmental externality. As with many modern agricultural areas, the Somerset Levels developed on the basis of the application of inorganic fertilizers. These inorganic fertilizers, which contain nitrates and phosphates, have supported the expansion of agricultural yields in the Levels, but they have had environmental consequences. The inorganic fertilizers that run off the agricultural lands of the Levels enter local river systems and eventually make their way to the ocean. Elevated levels of inorganic fertilizers in river and ocean waters lead to the production of blooms of algae (whose growth is accelerated by the fertilizers). These blooms can reduce the amount of sunlight penetrating the water’s surface and have detrimental impacts on the ecologies of the water systems where they are found. As with many forms of environmental externality, the impacts of the use of inorganic fertilizers are often experienced at great distances from the original point of pollution, and those who have benefited financially from the pollution event do not meet the ecological costs.

There are two things that it is important to note about externalities. The first is that they are clearly unjust, with those suffering the effects of externalities not being able to reap any of the benefits associated with their production. The second is that they are an inevitable outcome of unregulated market systems that promote economic freedom and the pursuit of profits. As an institution that sits partially outside of the market
system (for a more detailed discussion of this issue see below), the state has played an important role in regulating a series of externalities.

The state as environmental arbiter

If externalities are a product of market failure (or success, depending on which side of the externality relation you are on), and if the state has a role in regulating externalities, it is important to consider the relationship between the state and the market. At one end of the spectrum are those who advocate a strict Marxist interpretation of the state–environmental externality relation. These perspectives on the state claim that states are largely in the service of private sector interests, and will, where and whenever possible, limit the amount of environmental regulating it completes in order to preserve market efficiency and profit. There are others, however, who argue that states are able to achieve a degree of partial independence from the market system. In this context it is argued that states are able to position themselves within the relationship between markets and the environment in order to pursue the just treatment of the environment and a more equal distribution of the benefits that emerge from its utilization (Johnson 1996).

The modern state is able to exercise some form of regulatory power over the social uses of the environment for three broad reasons. First, many states have been founded on a constitutional commitment to equity and justice among its citizens. As the narrow pursuit of environmental gain, which mostly only benefits the few, can often lead to the unequal distribution of ecological harm, the state has a moral imperative to act. Second, as one of the key roles of the modern state is to think about the longer-term needs and welfare of its people, it has an inherent commitment to ensure that the narrowly focused individual pursuit of profit does not compromise the long-term sustainability of a nation’s resource base. Third, the state is arguably the only organization that has the administrative reach and expertise to be able to regulate complex environmental systems in a fair and sustainable way. In relation to these three contexts, Rudolf Bahro has suggested that states have the capacity to act as “elite stewards” of the environment, providing expert care for environmental resource to ensure long-term social and ecological justice (Bahro 1984).

The history of the regulatory relationship between the state and environment suggests that there has been an almost constant ebb and flow between the state’s prioritization of the interest of the economy and the environment. While states have constitutional commitments and bureaucratic capacities that impel them to control the use and pollution of the environment, they must also support the economic vitality of the nation over which they govern. It is often assumed that the state’s role as an arbiter of environmental externalities is a fairly recent phenomena. In reality, however, the state has been intervening in environmental externalities for some time. It was not, however, until the nineteenth century that it is possible to observe the more strategic, national level regulation of environmental externalities. It was during the industrial revolution of the nineteenth century that many industrial cities started to see the debilitating spread of environmental pollution. From power plants, foundries, and domestic chimneys releasing smoke and sulfur dioxide, to chemical plants releasing harmful alkalis into the atmosphere, the industrial city quickly become an unhealthy and unsafe place to live (Whitehead 2009). Early attempts to regulate the pollution of industrial cities was pursued through the enacting of public health legislation, which attempted to set standards of air and water quality and provision more comprehensive water treatment and sanitation systems. In the twentieth century public health
legislation was supplemented by clean air and water acts that not only sought to improve the health of the population, but also to protect terrestrial, riparian, and marine ecosystems which act as sinks for unwanted pollution.

When it comes to more economically developed countries (MEDCs) it would appear that governments have been successful in addressing the production and spread of environmental externalities within their territories. On many measures people living in MEDCs are now able to breathe much cleaner air and drink much purer water than they have ever been able to do since the emergence of the industrial revolution. But this apparent regulatory success has only been possible because many of the activities that produce environmental activities have moved from states with strict environmental regulations to less economically developed countries (LEDCs), which do not subject companies to such stringent regulations and can furnish them with much cheaper labor.

In the context of environmental externalities, it is important to distinguish between two different forms of state activity. In certain instances states have adopted a role as neutral arbitrator between competing private interests. In this context, the state is most obviously present in the development and ratification of legal frameworks relating to environmental nuisance, negligence, and trespass, which can help to arbitrate between the competing demands that are being placed on the environment. In other contexts, states have developed more overtly regulatory capacities in relation to environmental management. The regulatory state apparatus can be seen in relation to statutory prohibitions against certain environmental activities such as air and water pollution, which are designed to maintain certain pre-established thresholds of environmental quality. The first major acts of statutory prohibition against environmental pollution emerged in the decades following the end of World War II. In order to support environmental prohibitions many states formed environmental agencies and ministries. These state organizations were given responsibility for monitoring compliance with environmental law and ensuring that established levels of environmental quality were maintained. In more recent years, the command-and-control strategies of regulatory states have given way to more market and incentive-oriented state policies. In relationship to climate change, for example, certain states (and collections of state) have sought to create markets within which carbon credits can be traded and addressing climate change is financially incentivized. In other instances, states provide direct financial incentives to agriculturalists that farm in ways that do not undermine the broader ecosystem services that their land provides. These forms of post-regulatory policy reflect the fact that generating incentives for people to protect the environment are often more effective, and more easy to deliver, than heavy-handed strategies of regulation.

The limits of the liberal environmental state

The changing nature of environmental externalities has raised questions about the ability of states to effectively manage social relations with the environment. Some have argued that the planet has now entered a new geological era in which the power of humans to transform the global environment rivals the power of the natural world. This new area has been termed the Anthropocene. The Anthropocene is characterized by global-scale changes in the ecological balance of the planet, which include climate change, transformations in the nitrogen cycles (associated with the use of artificial fertilizers), acid rain, and the pollution of marine ecosystems. What unites these varied processes is that
they tend to transcend the jurisdictional boundaries of states. These are forms of environmental externalities that often involve situations where the perpetrator of environmental pollution is located in a different state to those who are directly affected by pollution. In the context of such transboundary issues, the environmental movement and national governments have often had to turn to international level organizations such as the United Nations to establish agreements on the fair and just treatment of ecosystems. Establishing effective environmental agreements at this level has proved difficult with many states being unwilling to sacrifice their long-held sovereign rights to use and exploit their national resources.

William Ophuls has suggested that the extensive use and exploitation of environmental resources (including agricultural land, water, and minerals) may ultimately result in new forms of environmental state (2011). According to Ophuls (2011), liberal systems of government prioritize the relatively free use of environmental resources, except when the use of those resources significantly impinges on other people’s well-being or freedom. Ophuls argues that liberal systems of government have only been able to be sustained because of the relative abundance of environmental resources and the ongoing ability of natural systems to absorb the environmental externalities that humans produce. In the context of relative environmental abundance, the use of a resource by one person is not seen to have a particularly detrimental effect on the ability of other people to use that resource. Ophuls claims that as resources become scarce and environmental systems become less able to absorb environmental pollution (as appears to be the case with observed climate change), governments will inevitably have to become more authoritarian.

The state regulation of environmental externalities clearly reflects a changing environmental ethos in the government. Over the last 150 years, states have not only facilitated the human use and exploitation of the environment, they have also sought to ensure a fairer and more sustainable relationship between society and environments. Despite this changing emphasis within environmental government, many claim that current environmental problems (from desertification to climate change) indicate the failure of states to effectively control the unsustainable use of the planet’s environmental resources. Some have claimed that the failure of states to adequately manage human–environmental relations stems from the fact that states are ultimately anthropocentric. That is to say that they tend to prioritize the rights and needs of people over those of nature (Bullen and Whitehead 2005; Eckersley 2004).

In her groundbreaking work on the relationship between states and the environment, Robyn Eckersley argues that in order to effectively address the environmental challenges that are associated with the Anthropocene more ecocentric forms of government (2004) are required. According to Eckersley, the modern state is constitutionally ill equipped to deal with the complex, environmental challenges that the planet now faces. The state is constitutionally limited by the fact that it only defines and protects the rights of people and not those of nature. This bias will always ultimately lead to the prioritization of the rights of humans over those of the natural world, even when the long-term wellbeing of humanity is indisputably connected to ecological sustainability. Furthermore, the territorial nature of the state means that governments tend to be locked into
the narrow defense of national environmental interests even when the environmental systems upon which those interests depend constantly transcend sovereign boundaries.

In her book, The Green State: Rethinking Democracy and Sovereignty, Eckersley outlines a new vision of the state, which she terms the green democratic state. This green state is different from the modern state to the extent that it is interested not only in the rights of people, but also of nature. It is also defined by a concern not only with the rights of those living within a given territory now, but also of the needs and rights of those living in other territories and who will live there in the future. Eckersley’s vision of the green state is deliberately idealistic as she seeks to explore what it may take to develop systems of government that are fit for purpose within the Anthropocene. Notwithstanding the deliberately idealistic nature of Eckersley’s work, it is important to note than some aspects of the green state are starting to come into existence. For example, in 2010 the Bolivian government passed a new law entitled the Ley de Derechos de la Madre Tierra: Law of Rights of Mother Earth. The law declares that “life systems” and “Mother Earth” (which are inclusive of human and natural systems) are titleholders of rights specified in national laws. This law is important because it is considered the first to give official rights to nature and, in so doing, make it possible for people to legally sue others who perpetrate actions that affect the sustainability and integrity of life systems. At present it is difficult to assess whether Bolivia’s Law of Rights of Mother Earth reflects the first step in the much broader evolution of green states throughout the world. What is clear is that Bolivia’s attempt to develop a greener form of state is more ethically palatable than Ophuls vision of an eco-authoritarian future.

Conclusions

This entry has charted the different ways in which it is possible to think about and observe the relationship between states and the environment. An historical perspective reveals how states and the environment have been in a constant state of coevolution, with the use and exploitation of the environment being central to emerging power of states and the practices of the state being crucial to the ways in people collectively understand and relate to environmental resources and systems. Considering state–environment relations through the lens of externalities reveals the ways in which states have been positioned at the intersection of the relationship between people, economies, and environments. Certain theories suggest that when it comes to managing environmental externalities that states tend to favor the interests of economic growth over those of the natural world. Others suggest that the state is able to act as an independent referee or arbiter of socioenvironmental relations. A constitutional perspective on the relationship between states and environments reveals how states have been historically constituted in ways that tend to favor the rights and needs of people over those of the environment. Constitutional perspectives raise the interesting question as to whether the formation of a green state, which recognizes and protects the rights of nature, could provide a long-term context within which to sustainably manage socioenvironmental relations. What this entry has illustrated is that the relationship between states and environments has been central to broader social relations with the natural world. Whether it is in the context of a new green state or eco-authoritarian form of government, it appears that state–environment relations are only likely to grow in significance as the twenty-first century progresses.
SEE ALSO: Environmental citizenship; Environmental governance; Environmental policy; Environmentality and green governmentality

References


Media communication on environmental issues

Media range from entertainment to news media, spanning traditional or mass media such as television, films, books, flyers, newspapers, magazines, and radio, as well as new media such as the Internet in general, Web 2.0, and social media. Traditional media rely on one-to-many (often monodirectional) communications and are sometimes referred to as “mass media,” whereas new or social media involve many-to-many, more interactive, webs of communications. Since the 1990s, the shift from traditional to new media has signaled substantive changes in how people access and interact with information, who has access to it, and who are considered “authorized” definers (e.g., actors with more power and influence than others) of the various dimensions of environmental issues. It is argued that new and social media have democratizing influences, as these channels of communication often offer a platform for more people to become content producers, and therefore have the potential to more readily shape the public agenda.

In all media, actors such as publishers, editors, journalists, and other content producers such as online bloggers generate, interpret, and communicate images, information, and imaginaries for varied forms of consumption. These “media representations” are therefore critical inputs to what becomes public discourse on today’s environmental issues.

As an example, climate change as a highly politicized media topic, especially in the United States, illustrates how (powerful) groups with diverging political ideologies, worldviews, or economic interests heavily influence the public debate on climate change. Recent studies on worldwide media coverage of climate change (Boykoff et al. 2015; see Figure 1), as well as on climate discourse and the interconnection of media, politics, and public opinion, suggest that media agendas match public agendas on the perception of climate change and policy implications (Hmielowski et al. 2014; Brulle, Carmichael, and Jenkins 2010; McCright and Dunlap 2011; Boykoff and Roberts 2007; Boykoff and Boykoff 2004; Weingart and Engels 2000). Through a web of interactions, the media have thereby influenced a range of processes from formal environmental policy to informal notions of public understanding about the environment.

Illustrating how this influence has changed over time, Figure 1 shows media attention on the terms “climate change” and “global warming” in English- and Spanish-language newspapers around the globe. The attention spikes can be attributed to certain events, for example the publication of the Stern Review and Al Gore’s film An Inconvenient Truth in 2006, the Fourth Intergovernmental Panel on Climate Change (IPCC) Assessment Report and the Bali Summit on Climate Change in 2007, President Obama’s inauguration and making climate change a
ENVIRONMENT AND THE MEDIA

Figure 1  World newspaper coverage of climate change or global warming in 50 newspapers across 25 countries and six continents, 2004–2015 (Boykoff et al. 2015; reproduced from Center for Science and Technology Policy Research International Collective on Environment, Culture & Politics).

political issue, the Climate Change Conference in Copenhagen in December 2009, and the United States–China joint announcement on climate change and green energy cooperation by the end of 2014. However, the figure also shows that the Fifth IPCC Assessment Report, published in 2014, had lower media visibility and public attention compared to the Fourth IPCC Report, although it provides better knowledge on the causes and effects and on the predictions of climate change in the future (Fernández-Reyes, Piñuel-Raigada, and Vicente-Mariño 2015).

Media coverage and journalistic norms

Mass media follow ethical codes of pursuing fair, accurate and objective journalistic work. These codes, mainly referred to as professional norms, consider journalism as a platform for an open and transparent discourse between different sectors of society. In this context, media strive for independence, truth, and accuracy in coverage (ASNE 2002), which are reflected in similar values and attitudes toward their professional work (Bennett 1996). However, these ethical codes or norms can also be considered as styles of storytelling that focus on how these rules critically shape media content and become an inherent part of communication (Boykoff and Boykoff 2007), also known as “news values” (Galtung and Ruge 1965). In this context, journalists follow the newsworthiness of a message according to a set of criteria such as familiarity, negativity, meaningfulness, unexpectedness, personalization, conflict, and others. Galtung and Ruge call these criteria the “conditions for news,” which turn facts or events into media messages. These norms also intersect with the journalistic norm of balance, that is, the common practice of
providing both sides of any dispute with roughly equal attention (Boykoff and Boykoff 2004). This is an activity that often appears to fulfill pursuits of objectivity (mainly prevalent among US media). In coverage of complex issues such as stem cell research, nuclear power, or genetic engineering, balance can provide a validity check for reporters who are on deadline and do not have time nor scientific understanding to verify the legitimacy of various truth claims about the issue (Dunwoody and Peters 1992).

Media coverage is also informed by authority-order bias, where journalists tend to rely more heavily on authoritative, legitimate, and official sources. While in some cases these authorities step in to restore order, at other times they serve to increase political concern.

Media interventions seek to enhance understanding of complex and dynamic human–environment interactions such as climate change. However, the characteristics of such interactions often run contrary to journalistic norms and values like personalization or novelty. As a result, vague and decontextualized reporting confuses rather than clarifies understanding and engagement on environmental issues. The *New York Times* journalist Andrew Revkin (2007) has referred to reporting without context as “whiplash journalism.” Context helps sort out marginalized views from counterclaims worthy of consideration on various aspects of environmental issues.

Values and ideological influences on media messages

As scientific understanding improves, it often unearthed new and more questions to be answered (Sarewitz 2004). What seems like a simple process to define what constitute “environmental problems” is actually influenced by priorities, ideologies, experiences, and perspectives. In other words, anytime the biophysical is captured and categorized, it undergoes varying degrees of interpretation, as influenced by power and scale via temporal and spatial contexts. The media play an inherent role in representing certain interpretations of the biosphere.

For example, in the case of climate change, media coverage of environmental issues is thus not a simple collection of news articles and clips produced by journalists; rather, media coverage signifies key frames derived through complex and nonlinear relationships between scientists, policy actors, and the public that is often mediated by journalists’ news stories (Trumbo 1996). These frames emerge in media representations regarding a certain issue to make it “more salient in a communicating text, in such a way as to promote a particular problem definition” (Entman 1993, 52). Asymmetrical influences also feed back into these social relationships and further shape emergent frames of “news,” knowledge, and discourse. For example, frames regarding climate change coverage in the media can be associated with different actors by emphasizing problems and causes (scientists) or judgments and remedies (politicians). The number of scientists in the media as news sources may decline “as the issue becomes increasingly politicized” (Trumbo 1996, 269). As another example, positivist approaches work to understand and interpret already existing social reality; meanwhile, constructivist positions emphasize interrogations regarding how power and scale construct, reflect, and reveal varied and complex phenomena such as language, knowledge, and discourse (Forsyth 2003). With these varied approaches into the complex and nonlinear interactions shaping public perceptions of environmental issues, geographical research has converged on the notion that media representations and its framings are not simply translations.
Humans contribute to climate change

US federal cap-and-trade legislation will have a discernible influence on the wider economy

Anthropogenic climate change has contributed to increases in hurricane intensity

UNFCCC architectures optimally address Global North/Global South issues

Figure 2  This schematic shows the distribution of relevant expert agreements and disagreements on climate science and governance issues (over time from left to right). The curve illustrates the relative strength/weakness of agreement or disagreement. The figure is adapted from Boykoff (2011) and comments by New York Times journalist Andrew Revkin at an annual Society of Environmental Journalists meeting.

of the “truth.” This becomes even more evident in the context of climate change and its polarized perception between different societal groups. Taking climate change as an example of a key environmental issue, Figure 2 shows facets of science and environment, where agreement is strong and others where there is disagreement.

Consider panel (a): “Humans contribute to climate change.” Over the past decades, reports and findings have increasingly signaled a broad scientific consensus – despite lingering uncertainties regarding the extent of attribution – that human activity has significantly driven climate changes in the past two centuries, and that climate change since the Industrial Revolution has not been merely the result of natural fluctuations. In other words, detection (of climate change) and attribution (to human activities) research has improved significantly. Noting this improved understanding, the United Nations’ IPCC has articulated this evidence-based view through multiple assessments of emergent
peer-reviewed climate research and many stages of consensus-driven processes. The steady flow of IPCC reports since the 1990s has represented “critical discourse moments” (Carvalho 2005) that describe happenings within an established discourse (e.g., on climate change) that may challenge the dominant and established positions on the topic. Those critical discourse moments have solidified a narrative of consensus, supported too by similar declarations from national science academies and other scientific groups over time. Despite this convergence, when mass media report on this issue, excessive attention can be paid to the tails in this schematic: outlier viewpoints at the ends of the distribution, rather than those under the bell curve that converge on agreement, have actually been found to have received amplified attention in media representations in particular country contexts such as the United States and United Kingdom (Boykoff 2011).

Panel (b) considers relevant expert-based views on the statement that “US federal cap-and-trade legislation will have a discernible influence on the wider economy.” As shown, a more flat and wavy line most accurately depicts the relative strength of agreement from “positive effect” through “no effect” to “devastating effect” (in schematic from left to right). In other words, panel (b) shows that there are a variety of legitimately divergent views on the potential effects that the implementation of cap-and-trade legislation may have on the wider US economy. Panels (c) and (d) illustrate further climate science and governance questions that have a range of perspectives, views, and opinions. The bimodal distribution of panel (c) captures that the relevant expert community working on questions of links between hurricane intensity and anthropogenic climate change may cluster around two peaks of consensus rather than one; that is, this represents what can at times be an issue where there are two convergent and rival explanations for an issue, within legitimate expert communities of researchers.

Overall, broad-brush treatment by mass media can then both privilege marginal views as legitimate, by giving them media coverage although they lie far from the main consensus (see panel (a), and unduly dismiss legitimate claims where consensus is less strong (exemplified by panels (b) and (d)). Numerous factors – within the issues themselves, as well as external contextual factors – contribute to the changing shape of these distributions over time.

This shows that fair, accurate, and precise media portrayals of environmental issues become even more perennial, central, and fundamental challenges. By more accurately, precisely, and fairly portraying the contours of the varied aspects of environmental change, understanding, meaning, and potential public engagement have greater opportunities to succeed.

Media and cultural politics of the environment

Media and environment interactions – from processes to effects – are usefully situated in a wider cultural politics of the environment. Cultural politics refer to processes involving how meaning is constructed and negotiated across space and place. This involves not only the representations and messages that are present in media discussions but also those that are absent. These discussions then shape how members of the public perceive possible actions and social practices at the human–environment interface. In other words, media frames influence the ways that the environment is perceived and discussed and how the public then view environmental issues, from formulations of what are “problems” to considerations of potential, feasible, or desirable ways to alleviate problems (sometimes referred to as “solutions”). These
ENVIRONMENT AND THE MEDIA

elements are also inextricably shaped by ongoing environmental processes themselves.

Mass media representations arise through large-scale (or macro) relations, such as decision-making in a capitalist or state-controlled political economy, and individual-level (or micro) processes such as everyday journalistic practices. Whether media are state-run or corporate-run shapes media coverage differently in countries and contexts around the world. While the main principle of democratic news production has been that media organizations then serve as a check on the state, in practice, corporate-controlled media have been argued to have acted systematically in the service of state power (Curran 2002).

For example, in the United States, Fox News (owned by Rupert Murdoch’s News Corporation) was seen to have inordinate power over the Republican primary elections leading up to the 2016 presidential election to succeed Barack Obama. While over 17 candidates declared that they were running for the Republican nomination at the time of the first televised debate, Fox News declared that only 10 candidates were going to be able to participate in the first Fox-televised debate. Consequently, candidates were held hostage to Fox’s power to determine the rules of selection of candidates to participate in this high-profile event. Meanwhile, billionaire brothers Charles and David Koch – who own Koch Industries, a conglomerate of oil and gas interests – budgeted to contribute nearly $900 million to candidate coffers and effectively influence the overall 2016 election, in which this episode took place. In this context, the Koch brothers have been prominent climate contrarians, calling into question the wisdom of regulatory interventions to address twenty-first-century climate changes as well as whether humans contribute to climate change at all, and their moneyed influence has particularly tilted the US Republican Party away from broad scientific evidence as well as public opinion on the subject of climate change. Together, Fox News’ and the Koch brothers’ influence has meant that candidates vying for the Republican nomination and eventually the US presidency may act in the service of corporate and corporate-controlled media power in order to meet their objectives. As such, productive public discourse on climate change – via media and elsewhere – suffers.

Over time, numerous researchers have explored how economic pressures and ownership structures have impacted news production (Carvalho 2005). Environmental journalism around the world is fraught with capacity challenges (in terms of time, personnel, and financial resources) to collectively cover complex and dynamic stories at the human–environment interface. Journalists, producers, and editors striving for fair and accurate reporting get swamped by these large-scale (macro) political economic pressures. Decreased mass media budgets for investigative journalism have adversely affected the communication of scientific information in that complex scientific material has often been oversimplified in media reports. Moreover, critical environmental issues have failed to garner coverage at all. In the name of efficiency, many reporters have increasingly covered a vast range of topics (called “beats”) under tighter deadlines, making it as difficult as ever to satisfactorily portray the complexities of environmental issues amid numerous demands. Moreover, content producers in publishing organizations that have withstood newsroom cuts and shortfalls have faced increased competition from other information platforms, especially from social media (video, audio, and text, along with blogs, Twitter, Facebook, Instagram, YouTube postings, etc.).

These numerous political economic challenges have damaged communication of environmental issues. For example, in many places in the Global
South, journalists often lack the capacity and training to cover the intricacies of environmental science, politics, and governance, as well as access to clear, timely, and understandable environment-related resources.

Other individual-scale (micro) factors include the mobilization and deployment of journalistic norms. The tendency to personalize stories means that coverage focuses on individual claim-makers and sensationalized stories, often subsuming deeper structural or institutional analyses. This connects to dramatization, where coverage of dramatic events tends to downplay more comprehensive analysis of the enduring problems, in favor of covering the surface-level movements. Novelty is important. Commonly, journalists mention the need for a novel “news hook” in order to translate an event into a story. These “new” things are actually novel ways of portraying or depicting already existing things, in the context of ongoing storylines and historicized or pre-existing norms and pressures. In tandem, journalistic valuations of drama, personalities, and novelty can serve to trivialize news content, as it can also lead to the blocking out of news that does not hold an immediate sense of excitement or controversy. However, pursuing these norms is not necessarily linked to reduced coverage.

An example of a dramatic, personalized, and novel event that generated tremendous news coverage is Hurricane Sandy, which struck the East Coast of the United States in late October 2012. Despite scientific uncertainty regarding links between hurricane intensity and anthropogenic climate change (see Figure 2), the event nonetheless spurred coverage focused on conflict and debate, and political actors as well as journalists pointed out that more has to be done in terms of disaster risk reduction, climate mitigation, and adaptation (Eilperin 2005).

In summary, media practices powerfully shape and negotiate meaning, influencing how citizens make sense of and value the world. Media representations thereby bridge different ways of knowing about the environment, and often mediate public perceptions, attitudes, perspectives, and behaviors related to environmental issues. This can have far-reaching consequences in terms of ongoing environmental scientific inquiry as well as policymaker perceptions, understanding, and potential decision-making. Media representations are at the same time shaped by framings, journalistic norms, and cultural politics of media economy that are inextricably linked with each other.

Going forward, the stronger convergence of traditional and new media means that we should rethink how or if media can still be considered as the democracy watchdog in times of digital turmoil. Climate change as a media topic has so far vividly illustrated how polarizing opinions from traditional media, and increasingly from new media beyond professional journalism, influence public perception and the agenda on the issue. It will therefore become more crucial for geographical research to consider and analyze future roles that various claim-makers have in the creation, maintenance, or silencing of discussions of environmental issues.

SEE ALSO: Environmental governance; Environmental issues and public understanding; Environmental management; Environmental policy; Environmental science; Political ecology

References


A priori, it is not clear why international trade should have a major environmental impact. The vast majority of pollution is generated at the point of producing and consuming goods and services. The transport of goods and services across country borders via trucks, trains, ships, or aircraft does generate air, water, noise, and other pollution, but it is relatively minor compared to total pollution levels. So why has there been so much activist, media, and academic attention on the environment and trade relationship? Key to understanding this is that, while the direct environmental impact of trade is small, there are myriad indirect effects which can be more significant. Analytically, it is useful to distinguish six effects of trade liberalization on the environment: scale, compositional, transport, static efficiency, dynamic efficiency, and governance effects.

Trade liberalization is supposed to boost the size of the economies participating in the liberalization, making people richer on average. In fact, this seems to be the main reason why policymakers sign up to relevant treaties. The scale effect could be static, that is, a one-off increase in economic size, or it could be dynamic, that is, increasing the growth rates of participating economies. Theoretically, the case for a sizable scale effect is well established since David Ricardo (1817) proposed the theory of comparative advantage, though some economists point out that, under special circumstances, trade liberalization may not only benefit some countries more than others but also positively harm some countries (Samuelson 2004). There is no consensus in the empirical literature on the size of the scale effect, however. In part this is because it is extremely difficult to establish the counterfactual, that is, what would have happened in the absence of trade liberalization. Inasmuch as there is a sizable scale effect, the impact on the environment is complex since the relationship between economic growth and the environment is complex. Some aspects of the environment seem to become better with higher per capita income (e.g., access to clean water and safe sanitation), while others become worse (e.g., carbon emissions). Some pollutants seem to follow an inverted U-shape with emissions first increasing with higher income up to a maximum point after which they tend to decrease again, a phenomenon known as the environmental Kuznets curve (see Neumayer (2013, 81–94) for a review of the theoretical and empirical literature on economic growth and the environment).

Trade liberalization impacts the compositional structure of participating economies. It typically spurs compositional changes already underway; for example, it allows postindustrial countries like Western developed countries to specialize even faster in the services sector since they can import manufactured goods from elsewhere. Mining and agriculturally oriented economies can specialize in this primary sector since liberalized trade allows them to sell their products in a bigger market. The environmental consequences are very much country-specific and a function of where in the process of compositional change a country is. Put simply, liberalized trade benefits the environment in countries that specialize
in the production of services since services are cleaner in production and manufactured goods, which are dirtier in production, are imported from elsewhere. Conversely, the massive increase in pollution in places like China and others is in part the consequence of a compositional change from agriculture to heavy industry and manufacturing, which in turn is partly spurred by trade liberalization allowing Chinese producers to sell their products globally rather than nationally. The environmental effects of countries specializing in agriculture will depend on the type of agriculture and whether it results in deforestation (e.g., palm oil plantations in Indonesia or soybean plantations and cattle farming in Brazil). The effects of countries specializing in mining are likely to be predominantly negative.

Understanding the transport effect is straightforward: more international trade requires more transport, which creates more emissions. Shipping is less environmentally intensive than moving goods via freight trains, which in turn is less intensive than trucking or flying cargo, but this direct effect of trade on the environment is unambiguously negative. The problem is exacerbated by the fact that often transport is either subsidized or any tax imposed on it is typically insufficient to cover the full environmental external cost.

All other things remaining equal, trade allows countries to specialize in the production of goods and services in which they have a comparative advantage, that is, that use factors of production intensively that are abundant in a country. This static efficiency effect should, on balance, have positive, if small, environmental consequences. For example, fewer chemical and energy inputs are needed to grow oranges in sunny climates than in gloomy and cold northern Europe. More generally, if goods and services are produced according to comparative advantage then no resources are wasted and no environmental pollution is created unnecessarily.

More important than static are dynamic efficiency effects of international trade. Producers in economies that are open toward trade can continue to produce profitably only if they operate at what economists call the production frontier, which implies that they need to keep up with the latest technological advances and to update their capital stock continually. Globally integrated markets also facilitate the more rapid diffusion of modern technologies to latecomers (Perkins and Neumayer 2005). Conversely, producers that are shielded by protectionist barriers can get away with hugely outdated capital stocks and technologies, as the experience of South America and India before the gradual opening of markets in the 1990s testifies. More modern capital stocks and the latest technologies tend to be less resource- and environmentally intensive. All other things being equal, this effect should therefore benefit the environment.

The effect of trade and trade liberalization on governance is perhaps the most complex of all the effects. First, environmentalists fear that trade liberalization takes away power from sovereign nation-states to enact strong environmental policies as these might clash with the obligation of countries to abide by “free trade” rules (Neumayer 2001, ch. 8). Much of this suspicion and criticism has been raised at the World Trade Organization (WTO) and its predecessor, the General Agreement on Tariffs and Trade (GATT), but similar fears have also been voiced with regards to the North Atlantic Free Trade Agreement (NAFTA) and other regional and bilateral free trade agreements. A related second fear is that free trade rules will inhibit countries from inserting trade measures into multilateral environmental agreements (Neumayer 2001, ch. 9). Third, environmentalists are concerned that trade liberalization will deter countries from
further raising their environmental standards out of fear that doing so will induce footloose capital to invest in other places with lower environmental standards—a phenomenon known as “regulatory chill” (Neumayer 2001, ch. 4). In the extreme, mobile capital would be drawn to countries with particularly lax environmental standards (so-called pollution havens), which may in turn force countries with higher standards to lower theirs in order to stay competitive—a phenomenon known as the “race to the bottom” (Daly 1993; Daly and Goodland 1994; Neumayer 2001, ch. 3; Eskeland and Harrison 2003; Clapp and Dauvergne 2011). Trade represents an important causal mechanism in this concern because it is liberalized trade that allows producers to take advantage of lower environmental standards in foreign places and yet export their products to places of higher environmental standards where consumers tend to be richer. It should be noted, however, that it is not legitimate to attribute lower environmental standards (or less enforced standards) to a deliberate intention to provide a pollution haven. Differences in existing amounts of environmental pollution, differences in the pollution absorptive capacity of environments, and differences in the intensity of environmental preferences can all cause differences in environmental standards between countries. Only where environmental standards are set inefficiently or suboptimally—where efficiency or the social optimum is defined as the level of pollution where the marginal social cost of pollution is equal to the marginal social benefit—or are deliberately not enforced in order to attract foreign capital is the charge justified. This is more likely to be the case in autocracies where free media and civil society groups are repressed, and in countries where polluting business groups can buy significant influence over the political process, such as in economies heavily dependent on fossil fuel extraction.

Of all the effects of trade and trade liberalization on the environment, the governance effect has probably attracted the most attention, particularly outside academia. As concerns the first fear, there is very little evidence that the WTO or other “free trade” institutions have actively restricted countries’ sovereign right to enact strong environmental regulations, as long as these are not aimed at process and production methods (PPMs) abroad, that is, as long as these do not aim at imposing trade sanctions on products imported from countries that produce these products in a way that is polluting or harmful to animals such as dolphins and sea turtles (Neumayer 2004; Charnovitz 2008). Where the adjudication panels of these organizations have ruled against a trade restrictive measure, it has typically been because the policy was not fairly and evenhandedly designed or applied, discriminating against foreign producers rather than imposing the same requirements on both domestic and foreign producers, or because the scientific basis for a health protection measure is thin (e.g., the European Union ban on beef imports from cows raised with growth hormones). That countries cannot, in general, impose their own environmental standards on the process and production methods outside their own jurisdiction may disappoint environmentalists, but this is not surprising: sovereign nation-states do not like to be told by other countries what to do within their own borders. Even here, however, WTO panels have allowed restrictive trade measures under stringent specified conditions, as the long-lasting WTO dispute between Southeast Asian countries and the United States, which imposed import bans on shrimp and shrimp products from Southeast Asian countries whose harvesting methods incidentally killed endangered sea turtles, testifies. Essentially, if the species meant to be protected by the import restriction is
endangered, the import-restricting country has shown good faith through efforts to negotiate a multilateral environmental agreement and has offered technical assistance and aid to exporters in a fair and nondiscriminatory manner, then trade restrictive measures aimed at PPMs abroad can be WTO-compliant. Moreover, countries are allowed to use labeling and similar measures to make consumers aware of different PPMs.

As concerns the second fear, anecdotal evidence seems to suggest that negotiation parties do worry about potential clashes of any considered trade measures in multilateral environmental agreements (MEAs). Those opposed to a MEA find it convenient to cite potential clashes with rules in trade treaties to buttress their opposition. However, there is no evidence that such worries are anything but marginal reasons why more MEAs do not exist and why any existing ones are not more ambitious. Rather, the lack of political will on the part of participating countries is to blame for that. One also has to keep in mind that several existing MEAs contain fairly strong trade-restrictive measures – for example, the Montreal Protocol and its amendments phasing out ozone-depleting substances, the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), the first amendment to the Basel Convention that essentially bans trade in hazardous waste between developed and developing countries, the Rotterdam Convention which regulates industrial chemicals and pesticides, the Agreement on Persistent Organic Pollutants, and the Cartagena Protocol on Biosafety.

The fear that countries would engage in a race to the bottom in environmental regulation are certainly unfounded, as are similar concerns of a race to the bottom in the different area of labor standards. There is not much evidence that any high-standard countries have actually lowered their environmental standards as a response to trade liberalization. That foreign direct investment increasingly goes to low-standard developing countries and that the exports of these countries become increasingly pollution-intensive need not be a result of the pollution haven phenomenon but may be the consequence of shifting comparative advantage, compositional changes in the economies of countries, and the fact that economies of the developing world tend to grow faster than economies of the developed world, a phenomenon known as economic convergence. One also has to keep in mind that the share of costs spent on complying with environmental regulations are typically dwarfed by costs spent on labor and capital, such that the incentive to channel investment to a particular place is typically not dominated by environmental cost-saving considerations though these can additionally play a role.

The regulatory chill concern is not easily dismissed, however, and may well be a real and very significant effect. Corroborating empirical evidence is hard to come by since the concern makes a counterfactual argument, namely how countries would have raised their environmental standards further were it not for the fear of capital flight. Arguably, European Union countries may well have enacted stricter carbon reduction policies were it not for the concern that such stricter policies would drive both existing and, more importantly, new direct investments abroad into unregulated countries (Neumayer 2001, 69–71). Naturally, it is easier to channel new direct investment abroad than to try and relocate existing investments (direct investments consist of cross-border mergers and acquisitions or the setting up of new plants as opposed to the much more short-term and speculative portfolio equity investments). As the example of greenhouse gas emissions shows, unregulated countries posing a regulatory chill threat on regulating countries like the European Union countries can include
even developed countries like the United States, which in many ways provides an ideal substitute for foreign investors because of its similar political stability and economic conditions.

It would be wrong, however, to assume that trade and trade liberalization can only have negative effects on environmental governance. Liberalization in the trade of environmental pollution control technologies and environmental clean-up services will benefit the environment as will the reduction of environmentally damaging production subsidies that sometimes forms part of trade liberalization. The same might be true for the increased flow of information that typically accompanies trade liberalization: countries open toward foreign trade are typically open toward many things. Finally, countries exporting to high-standard countries may over time adopt the higher product standards of the foreign countries they export to; a phenomenon known as the trading-up effect (Vogel 1997). This is because exporters may lobby the domestic government to ratchet up domestic environmental standards since producing a single product for both home and export markets allows firms to benefit from greater economies of scale, and tightening domestic environmental product standards may grant exporters a commercial advantage over their home market competitors lacking the requisite compliance technologies by raising the latter’s relative costs. Additionally, governments may embrace regulatory tightening as part of a strategic industrial policy, raising domestic emission standards in order to increase the competitiveness of domestic firms in higher-regulating export markets. For economies of scale to kick in, the higher-regulating foreign country market should be significantly larger than the domestic market; for example, German car manufacturers’ exports to the higher-regulating United States market induced them to lobby the German government to raise domestic emission standards in the 1980s.

The trading-up effect points in the opposite direction of the pollution haven argument. Note, however, that any trading-up effect is likely to be restricted to product standards (e.g., the exhaust pipe emissions of cars) rather than process and production method standards (e.g., the environmental standards faced by a factory producing cars). For process and production method standards there could potentially be an investing-up effect as transnational corporations (TNCs) bring more environmentally advanced technologies with them and may lobby governments in countries in which they invest to ratchet up domestic environmental standards. TNCs often have access to environmentally superior technologies, which they may deploy to their foreign affiliates and subsidiaries in lower-regulating countries. This is because it may be costly to substantially re-engineer technologies to suit different environmental regulatory requirements and higher-profile TNCs can face criticism from civil society if they are seen to deploy environmentally inferior technologies in host countries (Zarsky 2008). In turn, the transfer of standardized, beyond-compliance technologies creates incentives for TNCs to lobby for the upward harmonization of environmental standards in order to ensure regulatory consistency across the different markets in which they operate. Higher emission standards are likely to be especially advantageous for TNCs because they help to remove one of the key competitive advantages enjoyed by indigenous manufacturers, that is, their ability to produce low-cost, albeit higher-polluting, products using vintage technology.

In sum, with multiple effects pointing in different directions and having different strengths for different pollutants and in different countries, no verdict is possible as to whether trade liberalization is unambiguously positive or negative for the environment. However, on balance, one can
be fairly confident that the overall effect is overwhelmingly negative. This is based on the plausible presumption that, in most countries and for most aspects of environmental degradation, environmental regulations are insufficiently stringent (they fail to fully internalize the environmental externalities) because of the larger lobbying and other political power of polluters. Environmental regulations are bound to be suboptimal, in particular, as relates to transboundary and global environmental pollutants, since the benefits of regulation have to be shared with others, whereas the costs are borne domestically. Given that environmental policies are suboptimal, trade liberalization will act like the equivalent of a fresh breeze of wind on a house that is already set on fire – it exacerbates the negative environmental consequences of suboptimal policies, predominantly because, with policies being suboptimal, the negative effect of increased transportation and scale as well as, depending on the country, compositional change will dominate any positive effect from static or dynamic efficiency. Liberalized trade in turn may render it more difficult to achieve more stringent environmental policies as policymakers shy away from enacting policies that move closer to the efficient level of environmental pollution, where environmental externalities are optimally internalized, for fear of their economies losing competitiveness and footloose capital moving elsewhere.

Even if trade and trade liberalization can be expected to often result in more environmental degradation due to suboptimal environmental policies in trading partners, it does not follow that trade restrictions are the best answer. In fact, the optimal solution comes from more efficient environmental policies and trade liberalization. Also, there are still many win–win situations, where measures could be taken that are good for both trade and the environment. For example, more progress could be made toward removing trade barriers that are detrimental to the environment as well as fishery, agricultural, energy, and transport subsidies and toward removing restrictions on trade in environmental goods and services such as air, water, and noise pollution abatement technologies, and waste management techniques.

More progress could also be made by “greening” the WTO as well as regional and bilateral trade agreements. However, particularly at the WTO level, the prospects for this are slim. While practically all developed countries are in favor of some greening – partly by conviction, partly as a result of pressure from civil society – developing countries do not trust the alleged idealistic intentions and suspect that the greening of WTO rules is old protectionism in a new environmental guise. Regional and bilateral agreements fare a bit better, but not much (Neumayer 2001, 190–193).

This impasse tempts some to call for unilateral action by powerful countries with high environmental standards, imposing trade restrictions on lower-standard countries (Daly 1993). Yet, such unilateral action is not only in violation of WTO and other free trade agreement rules. It also poses the danger of protectionism in green disguise, opening the floodgates for trade protectionists to capture the environmental agenda and further their own selfish material interest. Another problem is that, once the floodgates are opened for unilateral action for some aspect of animal welfare or environmental protection, there will no longer be any holding on the slippery slope toward widespread use of unilateral action, not least because of the support by protectionist groups. There are no limits to the well-meaning idealist in search of targets on which to impose unilateral sanctions. Unilateral action also poisons the international diplomatic climate and runs counter to the spirit
of cooperation, which is needed for multilateral environmental agreements. This leaves the environment between a rock and a hard place. Unilateral action may be more effective in the short run but is undesirable, whereas multilateral action is desirable but hard to achieve, particularly in the short run.

SEE ALSO: Corporations and global trade; Environment and resources, political economy of; Environmental policy; Environmental regulation; Trade, FDI, and industrial development; Transnational environmental governance; Transport geography

References


Environment and waste

Stewart Barr
University of Exeter, UK

Histories of “waste”

For many readers of this encyclopedia the notion of waste is a familiar one. Our lives are seemingly dominated by the accumulation of products and packaging, much of which contains the necessities of everyday life: the food we eat, the liquid we drink, the cosmetics we used to cleanse ourselves, and so the list goes on. Nearly all of these products and their packaging are based on the abundance of cheap and easily available raw materials that can be manufactured into packaging and, given their abundance, can be discarded without the apparent need for reuse or recycling. So it is that we live in what some popular commentators have referred to as the “throwaway society.” This society, or system, is defined by a linear process of material accumulation and discarding, rather than a circular one, in which materials are extracted, manufactured, used, reused, recycled, and used once again. It is the hallmark of contemporary living in most developed nations and is illustrative of a radical shift in our relationship with materials and the environment that provides such resources.

Identifying the point at which this system emerged is not easy, but we can note several key moments in history that have progressively nudged us toward our present position. Not surprisingly, the industrialization of society is an important pretext for the era of mass consumption that we experience now. The discovery of abundant resources and their effective exploitation and manufacture during much of the nineteenth and early twentieth centuries has created a sense that raw materials are unlikely to run out any time soon. Oil, for example, the staple which we use in the production of plastic goods, has become the “black gold” of consumer societies and has facilitated many of the major technological developments of our age.

Built upon this context of abundance, a number of economic and social trends can be identified that have formulated our current relationship to resources and our propensity to discard them after use. The first of these is the economic growth model that has been adopted in many developed nations during the twentieth century. This is founded on the principle of a consumption-based society in which much of the economic wealth is generated by the demand for consumer goods (anything from clothing and footwear to kitchen white goods to mobile phones and computers). A slump in consumer demand is one of the main precursors of economic recession and so nation states have for a long time now set in train policies that attempt to maintain economic growth by encouraging consumers to spend and to create employment that will provide them with the necessary disposable income to do so.

In line with this economic model, the growth of an enriched middle class in many developed (and now developing) nations has been closely associated with the important signs and symbols of what it means to be successful. There has been the rapid development of social norms and expectations surrounding the material culture of being middle class, affluent, and successful, which often requires conforming...
to social practices of consumption that are implicitly wasteful. Such practices include the regular replacement of consumer goods (such as clothing, motor cars, and kitchen products) for reasons of trend and fashion, the consumption of exotic food products (often flown in from abroad) and the over-provision of goods to meet social expectations.

Indeed, the development of wasteful practices in the form of over-consumption is something which contemporary living and working arrangements have afforded many of us to engage with all too readily. As societies have become progressively sub-urbanized and individualized, and as long as commutes and shifting family relationships have put pressure on personal time budgets, manufacturers and supermarkets have created opportunities for “easier” shopping for goods that can be kept longer and are more convenient. The days of shopping in specific stores for particular goods has now largely been replaced by visiting one supermarket and purchasing goods packaged in set (and often inappropriate) sizes. Such shifts in practices have resulted in rising volumes of packaging waste and, through the promotion of bulk buying, increases in food waste.

Finally, increasing health and safety regulations surrounding the preparation and storage of food have increased the volume and diversity of materials that are discarded. Outside of the domestic context, tighter regulations about food preparation and hygiene have often resulted in manufacturers using greater amount of materials to protect food, in particular when longer life is required from perishable products. Within the home, the progressive improvement in cleanliness standards afforded by household cleaning products and a general shift toward reducing the risks associated with harmful bacteria has led to an unwillingness to store compostable materials in spaces like kitchens. Indeed, much of the traditional infrastructure supporting waste management for households has now disappeared. For example, “rag and bone” collectors, farmers collecting food scraps for pig feed, and scrap merchants are now banished from neighborhoods, where once they provided a way of householders providing material for reuse or recycling.

In these varied but interlinked ways, our relationship with materials and their destiny has changed over the last century or so, resulting in a situation where many of the goods we purchase are not immediately reusable and where there is every incentive not to reuse or recycle, but to buy new products, with the associated social and cultural benefits associated with being a good consumer. Within this context, this entry now explores three different lenses through which to view waste and its environmental implications.

The political ecology of environment and waste

Understanding how we have arrived at the situation where we are regarded as such a wasteful society – and why this is so problematic – can only be partly appreciated by a historical narrative. There is an underlying question that we need to address, which is to examine why waste has become such a potent political problem and why, within an environmental context, it has become such an icon of environmental degradation in contemporary society.

To understand the role of waste within the environmental protection movement, we need to engage with the initial stirrings of environmental protestors and protagonists in the 1960s, responding to the work of authors like Rachel Carson (1962) and the scientific arguments of writers like Garrett Hardin (1968), both of whom argued that rates of materials exploitation and the impacts of industrialization on the environment...
were likely to have severe consequences. These popular scientific arguments were institutionalized in the Club of Rome’s (Meadows and Club of Rome 1972) report, *The Limits to Growth*, in which the combined factors of resource exploitation and resultant pollution would have the overall impact of population collapse, a thesis constructed on neo-Malthusian principles.

What such arguments did was to capture the attention of governments and in particular the United Nations, which, during the 1970s initiated a range of programs (including the World Environment Programme and the Man [sic] and the Biosphere Programme) (McCormick 1989). Within these initiatives, the United Nations highlighted the importance of conserving renewable resources, reducing the exploitation of and reliance on non-renewables and controlling waste. Accordingly, during the 1970s, the economic and social practices underlying the so-called throwaway society were explicitly recognized and responses began to be institutionalized in both developments in international law and national regulations.

Yet such developments at a global and national scale were founded on a political framework that was focused largely on the relationships between government and industry, which, in the 1960s and 1970s, were often closely intertwined through formal ownership or political interests. Indeed, the political philosophy of the time was one dominated by a top-down, regulatory regime, in which governments relied on the “hard” policy measures of legislation and public ownership to deliver political goals.

The consequence of this governmental regime was that although significant improvements were made in the production of materials and associated reductions in pollution emitted from industrial processes, the basic economic model on which the consumer society was founded did not change. Indeed, through political developments in the 1980s and 1990s, the role of consumption became even more influential. This process is often referred to as the neoliberalization of society and has placed a much greater emphasis on “soft” policy approaches for achieving governmental goals, such as the use of financial disincentives, incentives and, most notably when it comes to waste management, the application of exhortation as a way of managing the environment. Such political approaches place the individual consumer, rather than the state or industry, center stage in the process of achieving desired outcomes.

It is the political philosophy of neoliberalism that has enabled waste management to come to the fore in the way it has since the 1990s. As Clarke et al. (2007) have noted, the move to a system where individuals are simultaneously regarded as consumers (with the expectations associated with maintain economic growth through increased consumption) and citizens (responsible for exercising sound environmental stewardship) has created a hybridized “citizen-consumer” in which the goals of a neoliberal society can be achieved through a form of ecological modernization. In this way, consumption patterns can be maintained through the use of more efficient and intelligent technologies and materials can be recycled using highly mechanized and efficient processes.

Waste management through recycling has, therefore, become an icon of the ecologically modern society in which economic growth can be maintained and which relies on the goodwill of individuals to sort and specify materials for recycling. Indeed, it is notable that the priority for much of the recent decades has not been on reducing the need to waste nor indeed direct reuse (both of which run counter to a neoliberal economic framing), but rather the reliance on mechanized and industrialized processes of
remaking products through recycling. Such approaches exemplify the broader approach to environmental management in advanced Western societies, which is to find “smarter” ways of managing resources and maintaining consumption levels, with the expectation that individuals can make relatively minor adjustments in behaviors to achieve specific policy outcomes. Within this context, this entry will now move on to explore the academic perspectives on waste as an environmental issue, focusing on the increasingly important role of behavioral change.

Academic perspectives on waste

Although research on waste is implicitly interdisciplinary, there are two broad approaches that can be identified across the spectrum of academic subjects. In the first instance, waste has for a long time been regarded as a technical issue, related to the chemical and physical processes associated with managing materials and their production. In this way, materials engineering has been concerned with exploring the potential for creating products that can be efficiently recycled and remanufactured. Such approaches focus on product life cycles and reusability, as well as the application and development of universal processes for treating discarded materials, usually in mechanized materials reclamation facilities (MRFs).

However, from a geographical perspective, a second body of research has become focused on what we can term the social dimensions of recycling. Set across a range of disciplines, this research tradition focuses on the ways in which waste is defined, the waste management practices that lead to high levels of discarding and the factors that influence the development of such practices. Broadly termed “environmental social science,” there are a number of important and often sharp divisions in the approaches adopted by social scientists for understanding how individuals, communities, and businesses deal with waste.

The first perspective we can identify within environmental social science is dominated by an approach that has a tradition of working on waste and recycling from the 1970s and is characterized by studies of psychological factors influencing consumer choices. Research within social psychology has defined the waste “problem” as one firmly rooted in individual attitudes and resultant behaviors. In this way, reducing waste and promoting reuse and recycling is a function of individual decisions that need to be understood as a precursor to promoting shifts on behavior. Such research adopts a logical positivist perspective, in which the broad epistemological assumptions are that: existing models of behavior can be utilized and adopted to understand attitudes toward waste management; it is possible and desirable to measure attitudes and behaviors and to quantify these measurements through extensive research approaches; the key unit of measurement is the individual on whose psychology the outcome is based.

This epistemological tradition has therefore focused on the application of specific psychological models and frameworks to the waste problem. Although there are many frameworks that could be mentioned here, three are worthy of note. First, Fishbein and Ajzen’s (1975) theory of reasoned action has been adopted in numerous cases to examine the relationships between stated attitudes, social norms, behavioral intentions and reported behaviors. A second, and related, model is the theory of planned behavior (Ajzen 1991), which promotes the importance of “perceived behavioral control” as a predictor of behavioral intention. Finally, Schwartz’s (1977) norm activation theory has been used extensively to examine the ways in which social
norms become mobilized to promote or prevent specific behaviors.

Much of the research that utilizes a social-psychological perspective is concerned with identifying key factors that determine behavioral commitment and an example of this type of approach, although somewhat dated now, is Schultz, Oskamp, and Mainieri’s (1995) comprehensive meta-analysis of factors influencing recycling behavior. In this paper and in research published subsequently, scholars have highlighted key factors relating to situational or structural characteristics (such as demographic characteristics, access to services and facilities, and explicit knowledge and behavioral experience), psychological factors (such as social norms, perceptions of convenience, response efficacy, self-efficacy, and attitudes), and underlying social and environmental values (such as the propensity of individuals to express environmental concerns). Barr (2002) subsequently developed a framework for understanding the relationships between these key factors in a framework of pro-environmental behavior, which illustrates some of the complexities of the psychological approaches for identifying and characterizing waste behaviors.

It is without doubt that the majority of social science research on waste has been from this social-psychological perspective. However, since the late 1990s, there has been growing unease within other parts of social science at the apparently deterministic and objectivist approaches that characterize this kind of approach. Most notably, Owens’s (2000) critique of rationalistic perspectives on issues like waste management has spurred social scientists from disciplines like human geography and sociology to explore waste and other environmental practices from the perspective of interpretivist and intensive social science perspectives.

In a ground breaking study, Gregson, Metcalfe, and Crewe (2007) highlighted one of the key critiques of such deterministic approaches by highlighting the very understanding that many social scientists have of the term “waste.” Building on research from the Waste of the World project, they argue that problematizing “waste” is the wrong approach, because it simplifies how human societies relate to material possessions and the ways that they have done so over time. In particular, they argue that linear conceptions of materials entering the home context and leaving as worthless discards hides the complexity of households’ interaction with “things” – possessions which are differently valued and often reused, reclaimed, reformed and “ridded” within the home. At the heart of their argument is for researchers to recognize the relationship between what appears to be individual behavior and broader social practices, which are routine-driven and historically shaped acts shared by groups (households, communities, classes). This, they argue, enables a richer understanding of how particular ways of living and being become shaped across society and which are re-enforcing, making the process of behavioral change much harder.

The intellectual tradition underlying this second approach is epitomized by Shove’s (2010) call for a deeper understanding of social practices that are relevant to the environment. In this and her earlier work, she calls for a focus on the broader underpinnings of contemporary practices, which are intricately related to the economic system, driven by a consumer-based culture. In this way, the focus needs to be less on the environment and more on the factors that shape everyday practices. So, for example, one could use this approach to argue that our focus should be less on the factors that shape negative attitudes toward recycling as a specific act, but more on the influences that have led to “wasting” in contemporary society and the systems that promote increased consumption and......
the establishment of profligate resource use as a social norm.

Epistemologically, this approach utilizes intensive and qualitative research as a means to explore broad social phenomenon in particular contexts, resulting in deep and rich understandings of practices at particular times. As a result, such research is often based on methods that draw on ethnographic research traditions and which utilize grounded theory approaches, which focus on the depth and richness of research data as a pathway for building a theory from bottom up, rather than the imposition of an objectivist and theory-led perspective.

Whichever of these two broad traditions are followed in the social sciences, the emphasis has very much been on an appreciation of waste in social context. Yet there is a third strand of research within and beyond the academic context, which has sought to explore how change might be effected. It is to this policy context that this entry now turns.

Policy responses to the “waste problem”

The political interest in waste as an environmental problem has emerged in recent years as a potent force in local government decision-making and has, particularly in countries like the United Kingdom, been illustrative of many of the conflicts associated with the promotion of sustainable development. Indeed, studies from North America have illustrated the ways in which waste has become implicated in struggles for environmental justice in marginalized communities.

Many of the political drivers for focusing on waste as an environmental problem have been framed by supranational aspirations to attain certain environmental targets and achieve reductions in harmful carbon emissions, which contribute to anthropogenic climate change. For example, the UK’s recent moves to reduce waste going to landfill sites has largely been the result of European Union directives, which have set out ambitious climate change emissions reductions targets and, more specifically, reductions in certain wastes being sent to landfill sites. The impact of these directives has been to spur national and local governments into developing specific targets for waste diversion and recycling, and in the UK this has resulted in a number of strategies since the 1990s to promote change.

Waste policy has conventionally been formulated in a technocratic environment, where focus has been placed on resource efficiency and mechanized recycling. However, since the 1990s, there has been recognition that greater reliance needs to be placed on reducing the potential for waste arisings in the first place. Accordingly, many policy makers now work with what is generally termed the “waste hierarchy,” in which waste reduction and minimization are placed as the top priority. For those materials that are produced, the second preference is for reuse without the need for major energy input (for example, the resale of clothing or washable nappies). In cases where materials cannot be directly reused, mechanical recycling is the next preferred option, highlighting the role for household collections of recyclables. Finally, if certain products cannot be recycled, the use of waste to produce energy, or “energy from waste,” is the preferred option. It is only when these options are exhausted that waste should be directed to landfill.

Within this framework, policy options have largely followed the political framework that was outlined earlier in this entry. In this way, softer policy options have included encouragement for both producers and consumers to reduce their reliance on packaging and to promote both reuse and recycling. In reality, this has resulted in
a series of voluntary codes of conduct for manufacturers to ensure that materials can, where possible, be recycled and that labeling reflects this. Nonetheless, most policy focus has been on the role of consumers in promoting recycling. Indeed, the way in which this has been achieved is through a very particular framing of consumer engagement that has focused on the need for behavioral change.

Within nations like the United States, Canada, and the UK, promoting behavioral change has been rolled out through programs that have drawn largely on the social-psychological tradition of focusing on the ways in which individuals can be persuaded to change their behaviors. Although policies have traditionally focused on approaches that try to fill the “deficit” in perceived consumer knowledge about recycling, there is broad recognition that there is often a discrepancy between stated attitudes toward recycling and behavioral commitments. This so-called value-action gap has been identified among many policy makers as a key priority and one of the ways in which practitioners are seeking to address this problem is to use approaches from market research to explore how consumers can be engaged to recycle more of their waste. Referred to as “social marketing” (Peattie and Peattie 2009), this approach utilizes the commercial marketing principles of audience segmentation, marketing mixes and setting specific behavioral goals to promote a social (or environmental) good and has been widely applied in health promotion campaigns. Essentially, the principle underlying social marketing is that messages and interventions to change behaviors need to be positive and easy to adopt.

Critically, social marketing is concerned with advocating change within existing consumption frameworks and as such needs to adopt the logic of choice (rather than compulsion) to engage consumers. In this way, it is closely related to a recent theoretical and pragmatic development from the United States, known as Nudge (Thaler and Sunstein 2008). Nudge theory asserts that changing behavior requires alteration of the “choice architecture” surrounding people as they make decisions and in the context of waste management, this would refer to the structures in which recycling behaviors operate, such as waste collection systems and infrastructure and the architectures within domestic settings that promote or prevent change.

The use of social marketing and nudge theory is not without controversy and for many social scientists and critics of instrumentalist policy making, these kinds of intervention are designed to be overly prescriptive and to crowd out alternative and more creative ways of reducing waste. In this way, critics argue that such approaches perpetuate the consumer-based society that has created the waste problem that we are now seeking to deal with.

This concern, which is driven by a different social and environmental ethic, has also been reflected in wider commentaries of the political ecology of waste systems within specific policy contexts. For example, in North America, much of the literature centered on environmental justice has emphasized the ways in which underlying social and economic inequalities can be reflected in and perpetuated by the politics of waste. Pellow’s (2002) compelling account of what his term “garbage wars” in Chicago illustrates are the ways in which established political systems can reinforce inequalities by continually marginalizing communities through the siting and out-working of waste infrastructures, such as MRFs. Put simply, Pellow explores why it is the case that black and minority ethnic communities are more likely to have an MRF situated in their neighborhood and why this often leads to exploitation and environmental injustice.
Waste in emerging economies

Although this entry has largely focused on the role of waste management and recycling within developed world contexts, there have been significant changes in the ways that developing economies have tackled the issue of waste arisings, especially in emerging economies where a growing middle class has greater access to consumer goods. Castillo Berthier (1983) examined the relationship between rapid growth in a transitioning economy and the increase in municipal waste. As the author noted 20 years on, when reflecting on the impact of his work (Castillo Berthier 2003), two key trends have emerged in emerging economies as they seek to deal with the environmental consequences of waste. First, increasing levels of formal household waste collection gradually replace informal ways of dealing with materials at the local level. In so doing, local authorities introduce mechanized and progressively more technical ways of removing and treating wastes, thus replacing previously vibrant and often essential informal economies of exchange for materials within neighborhoods. The enhanced role of the local state in managing waste is rationalized in part by the need to provide greater levels of environmental health and normally results in waste being removed to large municipal dumps and landfill sites on the edges of cities. However, as Castillo Berthier (2003) and Wilson, Velis, and Cheeseman (2006) highlight, the creation of major waste dumps has created new informal economies of scavenging and exchange that have shifted such economic practices away from neighborhoods toward landfill sites. In this way, an informal sector has emerged for creating value out of the growing number of consumer items discarded by growing numbers of middle class households.

Although the emergence of new and informal forms of economic exchange have followed state-led attempts to effectively manage waste, there are concerns in some developing economies that such practices can have negative impacts on the environment and human health. Indeed, several notable and high profile cases have emerged in recent years in China, which have highlighted the health impacts of unregulated recycling of products, many of which have been sent for recycling by developed nations and which have been processed in dangerous and often toxic conditions. Such examples serve to highlight that although waste is often dealt with in the localities in which it is produced, there is a growing globalization of waste and recycling that often means the “problem” of unwanted materials becomes one that is “out of sight and out of mind.”

Conclusion

Waste is both a contested and problematic term in social science and geographers have contributed a great deal to the understanding of how human societies have come to relate to and deal with materials that transform from having value to becoming value-less. Much of the focus of recent research has been on how to reach specific environmental targets associated with the diversion of waste from landfill and to increase recycling rates. Yet there are disagreements within social science about the most effective way to understand how practice of waste management develop and the ways in which researchers can use such understandings to inform and influence policy. Indeed, the ways in which waste is defined and managed are politically implicated, with a strong focus on soft policy approaches for attaining environmental goals. In conclusion, waste is a topic that is both academically and politically
complex and reflects the often stark differences in opinion about how we should manage the environment for future generations.

SEE ALSO: Behavioral geography; Climate change policy; Consumption; Ecological modernization; Environmental (in)justice; Neoliberalism

References


Further reading


To the casual observer, towns and cities around the world often appear as heavily degraded environments. Open rubbish dumps, palls of smog, long lines of traffic, and dirty bodies of water all contribute to the common perception that urban areas are causing environmental problems. But the relationship between urbanization and the environment is far more complex than this. Many of the most significant environmental consequences of urbanization are felt far outside city boundaries, whether in river basins that have been affected by water extraction to meet the needs of towns and cities or at the scale of global climate change driven by emissions from (and associated with) city activities. At the same time, while the concentration of settlement, industry, and infrastructure in urban areas may result in localized pressures on the environment, towns and cities have the potential to generate significant environmental achievements through economies of scale and agglomeration, through being sites of innovation of new responses to environmental challenges, and through acting as leaders to other cities and countries.

This entry reviews the evolving ways in which the relationship between the environment and urbanization has been understood and addressed in empirical and policy-oriented circles, while making some reference to more theoretical approaches. It does this through describing the characteristic environmental challenges facing cities, assessing recent policy efforts to address these, and describing some of the trends in academic thought that have engaged with these processes. It closes by identifying some of the key current and emerging issues that seem likely to influence the relationship between the environment and urbanization in the coming years and decades.

The extent and nature of urban environmental challenges

One key starting point for understanding the relationship between the environment and urbanization is to look at global patterns of urbanization. In 1950 only 29.4% of the world’s population (or 745.5 million people) lived in urban areas – by 2010 that had increased to 51.6% (3.56 billion people), and by 2050 it is expected to rise to 67.2% (6.25 billion). It is not only the global proportion of urban residents that has changed significantly, but also the distribution of these across continents. In 1950 14.4% of Africans (or 33.0 million people) lived in towns and cities – this had grown to 39.2% (400.7 million people) in 2010; for Asia the 1950 figures are 17.5% (245.1 million people) and the 2010 figures 44.4% (1.85 billion people). Any conclusions about the effects of urbanization on the environment should therefore deal increasingly with how this is played out in the towns and cities of the Global South (all figures UN DESA 2012).

A second starting point is to frame the significant environmental issues related to urbanization at different scales. Cities occupy only a small
amount of the world’s land area – approximately 2.8%, according to the Millennium Ecosystems Assessment (McGranahan and Marcotullio 2005) – and many of the environmental impacts of urbanization are concentrated within city boundaries. This is particularly true of the direct activities that take place in urban centers: the use of water, the production of sewage and solid waste, and the generation of air pollution. But, as cities become more successful and wealthier, many of these immediate impacts on the environment are reduced – through improved sanitation and solid waste management, through regulations that improve local water and air quality, and through the movement of “dirty” industry to locations with cheaper land and labor and less stringent regulations. But cities that appear “clean” on the surface may exert even more considerable environmental impacts at a global scale: while increased affluence can help to remove many local problems, it is associated with higher consumption of energy, goods, and services which has the potential for greater negative impacts on global environments and ecosystems.

Taken together, the increasing scale of urbanization and the increasing reach of environmental problems generated by towns and cities force a re-examination of the relationship between the environment and urbanization. This can no longer be seen simply as the observed environmental changes within municipal boundaries, but rather needs to be considered as the way in which activities in urban centers shape natural systems both in and far beyond their immediate surroundings. These impacts are frequently negative but, as cities increasingly demonstrate leadership in policies and practices relating to the environment, there is also considerable opportunity for positive outcomes.

Urban programs to address environmental concerns

Many of the key issues and debates about the relationship between urbanization and the environment focus on how different levels of wealth and poverty shape this. Several sets of ideas developed during the 1980s and 1990s explain this in different ways – sometimes at the level of the individual or community (the local environmental consequences of urbanization), sometimes at the level of the city (and how particular cities function), and sometimes at a global level (the way in which cities address one set of environmental challenges while making contributions to another).

Perhaps the most significant linkages between the environment and development in recent decades emerged in the wake of the United Nations Conference on Environment and Development (the 1992 Rio Earth Summit), much of which was reflected for the urban context through Local Agenda 21. Various planning guides, including a highly influential publication from the International Council for Local Environmental Initiatives (now ICLEI – Local Governments for Sustainability) (ICLEI 1996), were produced and implemented to various extents in hundreds of cities around the world. Other international programs initiated in this period included the Sustainable Cities Programme (UNEP and UN–Habitat), Localizing Agenda 21 (UN–Habitat) as well as several bilaterally funded programs (e.g., Danida’s Green Cities program), while many countries also developed their own versions of Local Agenda 21 to support their local authorities in addressing the new sustainable development agenda. Each of these had their particular features, but all attempted to include a concern for the environment and sustainability into urban planning, and to include civil society and the private sector.
in the planning process. Many of them also fell under the broader banner of “green urbanism” or “sustainable cities.”

These urban concerns for global environmental issues were sometimes criticized for prioritizing newly emerging ecologically oriented concerns over the long-standing environmental health and pollution problems that were still prevalent in many cities in low- and middle-income countries, and for lacking sufficient relevance to low-income groups. An alternative approach that also achieved prominence in the 1990s was to categorize environmental concerns in urban areas as being part of either a “green” or a “brown” agenda. The green agenda represents a more traditionally environmentalist approach to assessing the environmental consequences of urbanization – highlighting the impacts of urban areas on ecosystem health on a regional and global scale. In contrast, the brown agenda emphasizes not only local air and water pollution, but also the effects of environmental problems on low-income groups within cities (Table 1).

However, the “green” and “brown” agenda approach unnecessarily polarizes a set of environmental priorities that can be dealt with more effectively in tandem. This can be seen particularly in the emerging issue of climate change, in which the more “global” concern of mitigation (reducing the atmospheric concentration of greenhouse gases) directly shapes the extent to which adaptation (adjusting to moderate the harm of actual or expected changes in the climate) will be necessary at the “local” level. At the same time, however, by highlighting the environmental burdens of poverty, engagement with the brown agenda contributed significantly to keeping these issues in the realm of research and urban policy. A more sophisticated approach to these different dimensions of urban environmental challenges is the model of urban environmental transitions, which posits that, as cities become wealthier, the environmental benefits that they generate are gradually displaced (McGranahan and Marcotullio 2005) – from being primarily experienced at the local level (particularly inadequate access to water and sanitation, associated with informal settlements) to the city-regional scale (particularly air and water pollution associated with rapid industrialization) to the global scale (greenhouse gas emissions and high levels of resource consumption).

Table 1  Emphases of the brown and the green agendas in relation to key urban environmental concerns.

<table>
<thead>
<tr>
<th></th>
<th>Brown agenda</th>
<th>Green agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>Inadequate access and poor quality</td>
<td>Overuse; need to protect water resources</td>
</tr>
<tr>
<td>Air</td>
<td>High human exposure to pollutants</td>
<td>Acid precipitation and greenhouse gas emissions</td>
</tr>
<tr>
<td>Solid waste</td>
<td>Inadequate provision for collection and removal</td>
<td>Excessive generation</td>
</tr>
<tr>
<td>Land</td>
<td>Inadequate access for low-income groups for housing</td>
<td>Loss of natural habitats and agricultural land to urban development</td>
</tr>
<tr>
<td>Human wastes</td>
<td>Inadequate provision for safely removing fecal material (and waste water) from living environment</td>
<td>Loss of nutrients in sewage and damage to water bodies from its release into waterways</td>
</tr>
</tbody>
</table>

Integrated approaches to environmental management in urban areas lost momentum in the 2000s to more issue-specific approaches (such as water and sanitation, or climate change). While many of them had been successful at generating dialogues about how environmental issues should be addressed in urban centers, they had made fewer inroads in shaping the flows of investment that significantly influence the environmental futures of cities. In addition, they were often perceived of as being externally driven by “environmental” actors – while city officials were more concerned with more developmental social agendas or with attracting external investment. More recently, City Development Strategies (widely supported by the World Bank and Cities Alliance) have encouraged integrated, participatory, and sustainable approaches to planning that recognize the pressures that economic growth places on the environment while simultaneously offering opportunities for resource efficiency. However, neither Local Agenda 21 and its associated programs nor the City Development Strategies has explicitly engaged with the underlying issues of the urban political economy (Dodman, McGranahan, and Dalal-Clayton 2013).

Research trends on urbanization and the environment

In contrast to this, much research and academic analysis of the relationship between urbanization and the environment has taken a more strikingly radical approach. Political economists began to consider environmental concerns seriously, and to break the nature/society dichotomy – starting with Neil Smith’s analysis of *Uneven Development* (2008/1984), and reaching its apogee with the two influential volumes on society and nature by Noel Castree and Bruce Braun (Braun and Castree 1998; Castree and Braun 2001). These ideas were consciously absorbed and developed by urban scholars including David Harvey, who recognized that “the distinction between environment as commonly understood and the built, social and political-economic environment is artificial and that the urban and everything that goes into it is as much a part of the solution as it is a contributing factor to ecological difficulties” (Harvey 1996, 435; emphasis original). By the early 2000s, urban political ecology had become a recognizable field of academic endeavor, bringing together analysis of “the power-laden socioecological relations that shape the formation of urban environments” with a political program to achieve “a more equitable distribution of social power and a more inclusive mode of environmental production” (Swyngedouw and Heynen 2003, 898).

Some of the first research that explicitly adopted urban political ecology approaches and terminology involved detailed examination of the politics of water supply and distribution in Latin America (Swyngedouw, Kaika, and Castro 2002; Swyngedouw 2004) and southern Europe (Kaika 2005), as well as broader examinations of urban environmental processes and politics in Toronto, Los Angeles (Keil and Desfor 2004) and New York City (Gandy 2002). By 2003 there had been sufficient work under this general banner to warrant a series of “progress reports” (Keil 2003; 2005) – an effort that has recently been revisited (Heynen 2013). In this report, Heynen (2013, 2–3) lists a range of “vectors” that have been used to interrogate urban political ecology during the intervening years, ranging from air pollution, urban forests, gentrification, gardens, and waterfront transformations. He further suggests that more recent research aligned with the broad project of urban political ecology has moved from “resource domains” into “more traditional urban geographic themes” – although
much of this does retain an explicitly Marxist orientation (Loftus 2012).

A more policy-engaged and practically oriented stream of work on this topic is also represented in the journal *Environment and Urbanization*, first published in 1989. The publication has an explicit intent to fill a perceived gap in the coverage of urban issues in Africa, Asia, and Latin America, and has two sister journals (*Medio Ambiente y Urbanizacion*, with a focus on Latin America, and *Environment and Urbanization Asia*) which engage more directly with particular geographical regions. In recent years, a growing number of titles have sought to cover this area, including (but by no means limited to) *Local Environment* and the *International Journal of Urban Sustainable Development*.

The more general description of relationships between urbanization and the environment has become less of a research focus in recent years, with work tending to focus either on specific issues (such as climate change) or on some of the emerging themes discussed later in the entry (such as the urban green economy). The second half of this entry attempts to identify current issues and trends in environment and urbanization. Some of these (particularly around unmet basic needs) are residual agendas which, despite decades of attention, are yet to be resolved, while others (particularly the urban green economy) anticipate what may become key concerns in coming years.

**Current and emerging issues**

**Unmet basic environmental needs**

Despite the wide range of policies and programs mentioned earlier, many of the most basic local environmental needs remain unmet for residents of urban areas in the Global South. The United Nations Water Conference in 1977 declared 1981–1990 the International Drinking Water Decade, with the aim of ensuring universal access to clean drinking water; yet, by the time the Millennium Development Goals (MDGs) were declared, this commitment had been diluted to only halving the proportion of people without access to improved sources of water. Despite some progress in this area, 2.5 billion people in developing countries still lack access to improved sanitation facilities, and 768 million people remain without access to an improved source of drinking water. Similarly, many urban residents in low- and middle-income countries remain in a state of food insecurity. In urban areas this is primarily a challenge of access to and affordability of food rather than an issue of agricultural productivity. Whether this is seen as an environmental issue in cities is very much a matter of definitions, but it is a critical component of wellbeing for hundreds of millions of city dwellers. These examples of water, sanitation, and food illustrate that many of the issues that have long been seen as central to local environmental improvement in urban areas remain critical – indeed, they are frequently reframed as “rights” issues – and require ongoing research and policy attention even as new agendas are developed.

**Urban green economy**

One of the most significant emerging policy discourses linking environment and development concerns is that of the green economy. This is increasingly applied to urban contexts – indeed, a chapter of UNEP’s significant report *Towards a Green Economy* was dedicated to this (Rode and Burnett 2011). While there is not a single coherent definition for an urban green economy, most of the definitions build on the local environmental concerns identified earlier, but with an explicit effort to explain how these can be achieved through particular forms of economic activity. An urban green economy will
therefore contribute to cities achieving lower levels of pollution and carbon emissions, lower energy and water consumption, improved water quality, lower waste volumes, higher recycling rates, greater areas of green space, and reduced amounts, of agricultural land loss.

Advocates for urban green economies see the potential for these to simultaneously address economic development, poverty, and environmental improvement – including in low-income cities. For example, Simon (2012) argues that this can be achieved in African urban areas through the creation of “green jobs,” environmental rehabilitation that can enhance ecosystem services, expanding access to energy through greening energy supplies, recycling initiatives that reduce the amount of solid waste while providing employment, and greening urban areas.

Cities are also seen as being important sites for driving the green economy more generally. Rode and Burnett (2011) suggest that this is the case for three key reasons:

1. the proximity, density, and variety intrinsic to cities deliver productivity benefits for companies and help stimulate innovation;
2. green industries are dominated by service activity – such as public transport and energy provision, installation, and repair – which tends to be concentrated in urban areas where consumer markets are largest;
3. some cities will also develop high-tech green manufacturing clusters in or close to urban cores, drawing on knowledge and skill spillovers from universities and research labs.

Many of these principles are also evident in the discourses of transition towns and resource efficiency, which link some of the ideas of the green economy with responses to climate change. Transition towns originated in the United Kingdom, but have now spread to over 300 communities in the United States, Australia, Japan, and Chile; yet the diverse range of communities and interests found in the urban arena means that the underlying principles of resilience and self-sufficiency have been challenging to achieve (A. Smith 2010). Resource Efficient Cities is an approach championed by the United Nations Environment Programme (UNEP) which aims to reduce the total environmental impact of the production and consumption of goods and services in order to achieve financial savings and reduce pollution levels – while simultaneously opening up opportunities for the expansion of services to underprovided groups within the city. In contrasting ways, these two approaches address the concerns of the overuse of resources in urban areas, and similarly link this to broader economic concerns around affordability and competitiveness.

Cities and climate change

The release of the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) in 2013 and 2014 has been paralleled by a growing focus on the role of cities in contributing to and responding to climate change. One strand of research and practice focuses on the role of urban areas in generating greenhouse gas emissions and the potential of cities to reduce these. Some of this analysis has attempted to quantify the relative “responsibility” of cities for greenhouse gas emissions (e.g., Hoornweg, Sugar, and Trejos Gomez 2011), while another focus has been on the governance relationships that can contribute to their reduction (e.g., Bulkeley and Castán Broto 2013). In the arena of policy and practice, there have been considerable efforts to develop city-based greenhouse gas inventories that mirror the “measurable, reportable and verifiable” characteristics
required for nation-states under the United Nations Framework Convention on Climate Change, such as the carbon climate change registry (http://carbonn.org/) under which (as of April 2016) more than 600 reporting entities in more than 60 document and report their emissions.

The IPCC Fifth Assessment was the first version of this report with a dedicated chapter assessing the impacts of, vulnerability, and adaptation to climate change in urban areas (Revi et al. 2014) (the Fourth Assessment had addressed this in the broader sense of “Industry, Settlement and Society”). In contrast to the previous assessment which highlighted the vulnerability of cities, there is growing research and reflection on specific measures that are being taken by cities to reduce risk. Many of these have been carried out under the framework of “resilience” (Tyler and Moench 2012) or “transformation” (Pelling 2010), recognizing that, for the rapidly growing urban centers in the Global South, maintaining existing functions in the face of an uncertain climate is an insufficiently ambitious goal.

Key frameworks for understanding mitigation and adaptation in cities include focusing on the role of different stakeholders and governance arrangements. However – in common with the analyses of other environmental problems in cities – much of this has not engaged explicitly with power relations and political dynamics, and how they shape the vulnerability of cities or their ability to contribute to solving climate change.

Urban governance, global leadership, and the environment

The ways in which city governments function have also evolved substantially over recent decades, at the same time as the urban environmental agenda has become increasingly prominent. This is encapsulated in the shift from “government” to “governance,” which has been reflected by city governments becoming increasingly purchasers rather than providers of services. To an extent, this has removed some of the potential for direct action on the environment by municipal authorities by reducing the amount of control that they may have over waste management or other environmental services. At the same time, however, this has encouraged a new wave of entrepreneurial action, with cities actively competing to attract external investment – and environmental improvements or leadership can be one component of this.

Cities and city leaders have become increasingly vocal about the role of urban areas in engaging with global concerns, including those related to the environment. While environmental treaties remain the domain of the nation-state, groups of mayors and cities have coordinated their voices and actions to demonstrate leadership in this area. Significant efforts include those coordinated by ICLEI – Local Governments for Sustainability and C40 Cities Climate Leadership Group. In contrast to the international environment and urbanization agendas of the 1990s, these are driven in a more endogenous manner, and seek to use the mobilizing power of cities to encourage other cities to engage with environmental issues, and to demonstrate potential areas for action to national governments.

Conclusion

Scholarship and practice in recent years have significantly shifted perceptions on two key issues relating to the environment and urbanization. First, the environment and the urban environment are no longer seen as being distinct but as being closely related to each other. Second, towns and cities are recognized as not only
harming the environment but as representing significant opportunities for engaging with some of the most substantial global environmental challenges. As the world’s urban population continues to increase – indeed, almost all net population growth in the coming decades will be in towns and cities in the Global South – the environmental agenda will of necessity have to continue and deepen engagement with urban concerns. Equally, the decisions that are made and the actions that take place in urban areas will be the main features shaping the future of the global environment.

The continued growth of urban areas, and the ongoing (and in many cases worsening) environmental problems they face, mean that linking research to policy and practice in this area is increasingly important. At the same time, urban residents in many places have become increasingly involved in planning, managing budgets, and setting priorities for their cities through the opening up of urban governance systems and efforts to foster broader citizen participation. The potential – perhaps even need – therefore exists for greater coproduction of strategies and responses to address environmental challenges in towns and cities, particularly in the Global South. As urban political ecology and sustained political engagement have both shown, this can be achieved only through simultaneously engaging with the social relations and inequality that cause different groups to gain and lose from contestations over the urban environment.

Acknowledgments

This entry has benefited substantially from discussions with colleagues in the Human Settlements Group at the International Institute for Environment and Development (IIED). Any omissions and errors remain, of course, my own.

SEE ALSO: Sustainable cities; Urban political ecology

References


Environment, democracy, and public participation

Patrick Devine-Wright
University of Exeter, UK

Introduction

Public participation can be broadly defined as the ways that the public are enrolled into procedures of policy formulation and decision-making on environmental issues. Over the past 20 years, public participation has become a cornerstone of many different sectors of policy and decision-making, encompassing the environment and sustainable development, science and technology, spatial planning, and more recently climate change. The United Nations-led Aarhus convention introduced a legal commitment upon individual states to enable public access to environmental information, participation in environmental decision-making, and access to justice on environmental matters. To date, it has been ratified by 46 countries, chiefly in Central Asia and Europe (United Nations 2014).

Public participation is related to a broader term – public engagement – that encompasses diverse methods of educating the public (e.g., awareness raising campaigns), consulting and involving the public (e.g., through surveys or deliberative discussions) as well as engagement by the public (e.g., petition signing or protest activities) (Walker et al. 2010). Advocacy for greater public participation stems from several opinions. First, it is based upon positively valuing lay knowledge, seeking to bring this into formal procedures of decision-making or policy formulation. This positive valuation also critiques public engagement that is undertaken to remedy a presumed “information deficit” or lack of knowledge held by the public about environmental problems (Owens 2000). Second, advocacy for public participation can be based upon a lack of public trust in the role of experts or concerns about the legitimacy of expert contributions to environmental decision-making (Collins and Evans 2003).

Types of public participation and engagement

Rowe and Frewer (2000) distinguish three types of engagement activities: communication, consultation, and participation. These differ by the flow of information occurring between the parties and by significance in the policy or decision-making process. Communication is the most basic type, involving one-way information flow from the “sponsor” (typically a government organization or private company) to the public. In this context, the public are simply informed of policymaking or planning decisions with no power to influence these. Consultation involves two-way information flow between “sponsor” and the public, but the information flows back without any dialogue. Participation involves a two-way exchange of information between sponsor and public with the possibility for transformed opinions in both parties. This typology builds upon the classic framework proposed by Arnstein (1969) involving a ladder of public participation from “manipulation” at the bottom to “citizen control” at the top. Participation is
often regarded as the optimal form of public engagement since it provides participants with a greater degree of influence or control by comparison with communication or consultation that can be viewed as tokenistic or manipulative.

Multiple rationales for public engagement

Public participation can occur for many reasons. The literature identifies three predominant rationales or motivations: substantive, normative, and instrumental (Stirling 2005). The substantive rationale means that public participation is undertaken because it is believed to improve the quality of policy formulation or decision-making. This links to the valuing of lay knowledge mentioned above. The normative rationale for public participation is the view that involving members of the public should be undertaken because it is the right thing to do, for example, in a democracy it can be thought right and proper to seek citizens’ views before making decisions on their behalf. Finally, an instrumental rationale means that public participation is undertaken because it is believed to increase the likelihood of achieving a predetermined goal, for example, to promote greater social acceptance of a specific policy. This is based upon the presumption that the public will be more likely to accept certain policies or planning decisions if they are invited to participate in formulating those. Attempts to engage with the public for the purpose of dispelling ignorance and misunderstanding – most common under the instrumental rationale – may therefore presume that this will then lead to greater social acceptance (or at least less opposition) to policies or decisions. By contrast, the substantive rationale presumes that lay knowledge is a valuable resource that can be usefully incorporated into decision-making procedures, particularly those involving scientific and technological innovation associated with uncertain or risky outcomes.

Normative and substantive rationales for public participation are central to the concepts of sustainable development and sustainable communities. These rationales claim that citizens in a democratic political system should be empowered to participate in decision-making processes, that resulting decisions are better quality, that citizens’ consent and trust can be gained in addition to fostering change to attitudes and behaviors. Policymakers have emphasized the value of inclusive engagement strategies that aim to secure input from marginalized or vulnerable social groups on the grounds of equity, fairness, and justice, and promoting community involvement in local planning decisions. Thus a sustainability perspective would generally opt to “open-up” discussions on policy or planning (Stirling 2005) by means of participatory, two-way mechanisms of public engagement (see the Rowe and Frewer typology above), rather than “close-down” debate through the use of one-way mechanisms of communication.

Recognizing these diverse rationales for public participation in environmental policy and decision-making is important, since a single participatory initiative may be delivered or supported by organizations holding quite different rationales which, left implicit, can create tensions and difficulties. At minimum, it suggests the value of a critical approach to rhetorics of democracy or public participation in environmental decision-making propagated by powerful actors.

Methods and timing of participation and engagement

Methods of public participation and engagement are multiple and varied, including:
one-way information provision (e.g., leaflets, posters, newspaper adverts, and web-based materials);

two-way consultation (e.g., public exhibitions where the public can learn about proposals and register their opinions about it, phone lines, and questionnaires);

deliberation (e.g., citizens panels or juries where individuals have more time to learn about the issues at stake, to question experts, and to discuss potential outcomes).

In the United Kingdom, the National Consumer Council (NCC 2008) considered deliberative engagement valuable for improving public services, social cohesion, and democracy. An engagement process was defined as deliberation if it (a) involves discussion between participants; (b) involves a range (diversity) of people and information sources; and (c) has a clear task or purpose relating to, for example, a policy or project question or topic.

Three main types of deliberative public engagement are currently used in the United Kingdom:

- deliberative research, which builds on market research techniques (e.g., national citizens’ summits and policy consultations);
- deliberative dialogue, which builds on dialogue and consensus-building techniques, to develop an agreed view or set of recommendations (e.g., national dialogues on science and technology); and
- deliberative decision-making, which builds on partnership methodologies to enable participants and decision-makers to decide jointly on priorities and programs (e.g., participatory budgeting exercises) (NCC 2008).

Each of these forms of deliberative public engagement has been practiced in relation to environmental policies and decision-making. Both deliberative research and dialogues have been used to inform policymaking (e.g., a citizens’ summit to inform development of the Climate Change Bill in the UK).

The principles that should be applied when practicing deliberative public engagement are that the process (i) makes a difference; (ii) is transparent; (iii) has integrity; (iv) is tailored to circumstances; (v) involves the right number and right types of people; (vi) treats participants with respect; (vii) gives priority to participants’ discussions; (viii) is reviewed and evaluated to improve practice; and (ix) participants are kept informed (NCC 2008, 6).

The timing of engagement with the public is important in determining how it is perceived and responded to, and what opportunities are open to affected groups (e.g., communities in a locality) to materially influence the issues at stake. “Upstream” engagement implies actions taken to engage with communities at an early stage of project development; “downstream” engagement implies engagement taken at a later stage where little opportunity to substantially revise proposals may exist. There is considerable advocacy for “upstream” public engagement – for example, of communities affected by planning decisions – yet there is also some concern among academics that this will continue to be problematic unless differences in power and status between developers and communities is addressed. Upstream engagement is also challenging in contexts of emergent technologies where there may be very low levels of public awareness and knowledge. In such contexts, what works when engaging the public will have to be tailored to the specific technology, and more deliberative methods of engagement are especially useful to solicit informed opinions and judgments from public participants.
Imaginary publics, rationales, and methods of engagement

The siting of low carbon energy technologies is one area of environmental policy where public participation has been strongly encouraged, often from an instrumental perspective of lessening “NIMBY” (Not In My Back Yard) conflicts with local communities. Research has shown that energy companies typically conceive the public as emotional individuals that are deficient in technical knowledge (Cotton and Devine-Wright 2012), leading to engagement by “communication” (Rowe and Frewer 2000) that aims to undermine “myths” about certain technologies and address public concerns. Such ways of thinking lead to the use of certain engagement methods (e.g., exhibitions, internet communication, and focus groups) and the avoidance of others (e.g., public meetings), in order to diffuse and render invisible conflicts between companies and residents (Barnett et al. 2012). As a result, opportunities for two-way dialogue and the empowerment of communities are overlooked.

Evaluating public participation in environmental decision-making

Although judging the quality of public participation is far from an easy task, in principle such judgments refer to the quality of the inputs (e.g., information provided), processes (e.g., how participatory procedures are organized), and the outcomes or outputs produced. Rowe and Frewer (2000) identified a set of generic criteria for judging the quality of participatory processes, distinguishing criteria for public acceptance and good process. Public acceptance criteria included:

- the representativeness of the participants and their views;
- independence – participation should be practiced in an independent, unbiased manner;
- early public involvement – as soon as value judgments became important with regard to the issue under question;
- influence – where the output of the procedure genuinely impacted on policy; and
- transparency – for both involved and interested parties.

Process criteria comprise resource accessibility, including information, and time to enable the public to achieve objectives; and task definition, which refers to how well specified or defined the tasks for participation are with regard to the scope of the exercise, its expected outcome/s, and mechanisms of procedure. In addition, participation in decision-making requires mechanisms for structuring and displaying the decision-making process (for example, flow-charts or pictures), efficiency and transparency of process and documentation of process.

On the value of public participation in the environmental domain

One of the most comprehensive reviews conducted on the role and value of public participation in the environmental domain was conducted by a panel convened by the US National Research Council (Dietz and Stern 2008). The review was based upon a series of workshops, meetings, and working papers that sought to integrate both academic and practitioner expertise, and to analyze empirical evidence of processes of public participation and relevant social science theories and concepts including governance, conflict resolution, risk
perception, group processes, and public understanding of science. The principles advocated by the authors were drawn from a broad range of environmental domains including pollution, forestry, watershed management, and climate change.

The review argued that public participation should be judged on three criteria: improving the quality of assessments and decisions, enhancing their legitimacy among those involved and potentially affected, and increasing understanding and decision-making capacity among all those involved. The three criteria (quality, legitimacy, and capacity) may also be used to evaluate public participation in environmental assessment and decision-making; in particular, they suggest the following indicators of success.

- **Quality**: public concerns are addressed, more information is considered and innovative ideas generated, outputs reflect a broad view of the situation, conclusions reflect best available evidence;
- **Legitimacy**: mistrust among participants is reduced, participants accept the decision process as fair even if they do not agree with the recommendations for action;
- **Capacity**: participants (including scientists) are better informed about environmental (or related) issues and gain a better understanding of each other, organizers and participants develop skills in participatory decision-making.

The evidence for the effects of public participation was examined, drawing upon practitioners’ experiences with dispute resolution, planning and environmental assessment, and upon empirical studies (single and multicase studies, experimental and quasi-experimental research) of environmental public participation processes. They concluded that public participation does improve decision-making along the three dimensions of quality, legitimacy, and capacity. Not only this, the evidence shows that these criteria are positively correlated: processes that enhance decision quality simultaneously tend to improve legitimacy and enhance capacity. Thus, there need not be a trade-off between making “better” and “more acceptable” decisions.

The review also presents key principles for effectively managing and organizing public participation in the environmental domain and for integrating science within this process. A well-managed participatory process is defined as having a clear purpose, adequate resources, commitment from those involved to support the process and take the results seriously, and a focus on learning and implementation. Five key attributes were identified: breadth, intensity, openness of design, influence, and integration of science. In terms of breadth, a more successful process is based upon the involvement of a full spectrum of both interested and affected parties. Intensity refers to the length of time participants spend in the process and particularly the amount of face-to-face contact: in general, there is a positive association between intensity and desired results, but this can depend on the particular context or aims of the process. Influence and openness of design are closely related concepts, and imply collaborative problem formulation, transparency of process, and good faith communication, all of which improve the success of decision-making processes.

In terms of the optimal timing of public participation, involving participants early in the process is beneficial for a number of reasons. It allows for collaboration in problem formulation and option generation and the “frames” or lenses through which new information is interpreted and are co-constructed to reflect the broad values of all participants (not only the agency or sponsor organizing the process). Integrating public views
from the outset improves the substantive quality of decision-making (by illuminating the problem from a range of perspectives) but also enhances its legitimacy by representing a range of interests, which in turn means the public is more likely to see the process as fair and accept its outcomes. From an instrumental perspective, early involvement tends to avoid trouble later on whereby participants seek to express their concerns outside the participation process (for example, through litigation or protest). Finally, there should be iteration between analysis and deliberation, so that participants shape the focus of scientific analyses and in turn able to evaluate (as in an extended peer review) and use that technical information in their decision-making. From this perspective, public engagement is not merely an add-on to scientific investigation; rather, the authors argue that integration of science and public deliberation is mutually beneficial.

The review addresses why public participation sometimes fails to produce expected instrumental, substantive, or normative outcomes, and identified several challenges that must be overcome to deliver a successful participatory process. The review does not conclude that there are any environmental issues or problems, in principle, that are not amenable to participatory engagement. However, it does recognize that the more complex, uncertain, invisible, or remote an issue, the more difficult it will be to implement participatory processes effectively and successfully, and climate change was identified as an example of an issue that posed challenges of this kind.

To this end, two categories of challenges are identified: issue-related and people-related. Inclusive participation is harder to achieve with global scale issues, where everyone is potentially affected, or in cases where there are disparities in power or resources (e.g., finances, time, education). Similarly, motivation to participate is likely to be lower where a problem is intangible or invisible. Scientific uncertainty and complexity may also make decisions more difficult to agree upon. However, the authors concluded that these challenges are not insurmountable. For example, the participation of geographically dispersed communities may be facilitated by convening online participation or workshops in multiple locations. Incentives to participate in the form of assistance grants and the use of collaborative design requiring participants to work together help to address inequalities in power or status. Formally characterizing uncertainty and seeking to communicate in plain language may help when the scientific evidence is uncertain or complex.

The review concludes by noting the deficiency of the evidence based on public participation. In particular, the authors advocate the development of a “science” of public participation that seeks to purposefully experiment with different aspects and methods of participation to better identify how best they should be designed and evaluated. They are critical of the current research because of a preponderance of single case study analyses, and a dearth of experimental designs.

Problematising public participation

Skeptics of the increasing use of public participation have claimed that it rarely leads to the empowerment of participants as is often presumed. Cooke and Kothari (2001) denigrated a “tyranny of participation” in which the involvement of the public has become little more than a box-ticking exercise, formulaically conceived and delivered, with minimal chance to achieve lasting change. Hindmarsh and Matthews (2009) decried the use of “deliberative speak” in Australia, where public participation was paid lip service in contexts of renewable energy planning
and decision-making. Without a genuine attempt to involve communities in order to build consensus about where wind farms should be sited, a situation of “placation” was diagnosed, echoing Arnstein’s judgment of “tokenistic” participation (Hindmarsh and Matthews 2009). Moreover, use of the very concept of “the public” (as well as related terms such as “individuals,” “households,” or “communities”) can easily lead to overtly homogeneous conceptions of the people targeted by specific engagement initiatives, underplaying their diversity and overlooking marginalized social groups.

These tensions between seeking to enhance and deepen public participation on the one hand and, on the other, to reduce it to a shallow, formulaic approach that leaves power relations or environmental competencies undisturbed are increasingly prevalent and discussed in the literature. Recent UK planning law has been said to simultaneously centralize power and to localize participation. The publication by government of national policy statements has provided a backdrop that strongly supports low carbon energy generation, making it difficult in principle to oppose certain forms of low carbon energy. Local authorities no longer make decisions about major energy projects (e.g., wind farms over 50 MW); instead decisions are made by the Secretary of State, without the need for locally based planning tribunals. Concurrently, the Localism Bill (2011) proposed the devolution of powers to local communities, including the retention of a proportion of business rates at the local level, amidst rhetoric of a “Big Society” that extols the virtues of volunteerism, communitarianism, and local empowerment. Yet such rhetoric has been criticized for being undermined by other policies to reduce local funding or to centralize decision-making.

Increasing efforts are being made to “marketize” aspects of environmental planning and local impacts of development projects in particular through the provision of “community benefits packages” that are distributed by project developers to locally affected residents in an effort to redistribute the costs and benefits of development. Whether local impacts (e.g., landscape change) can or should be equated with a monetary value has been strongly queried and may inflame rather than reduce public objections. For example, it has been shown that offering a financial incentive can reduce the moral motivation to support specific goals (Walter 2012). Such packages are just one example of attempts to supplement traditional hierarchies of environmental decision-making with numerous additional “instruments” that involve information provision, communication, and incentives in order to better ensure the delivery of national policies at the local level.

Environment, democracy, and public participation beyond national boundaries: the case of climate engineering

While much of this discussion has been oriented toward national policies and local development projects (e.g., wind farms) as important environmental domains for public participation and deliberation, these present only a partial picture of the contexts of environmental decision-making unfolding in the twenty-first century. Given the globalization of economic, cultural, social, and political processes over the past few decades, it is pertinent to “upscale” this discussion with reference to a challenge that goes beyond any particular local or national interests – climate change. In particular, the emerging possibility of large-scale, intentional modification of the Earth’s climate in order to prevent harmful impacts of increasing greenhouse gas
emissions raises profound and complex social, moral, legal, and ethical uncertainties that are beyond issues of technical feasibility. Of relevance is the question of how such modification should be governed and how democratic and participatory principles can inform both the research and potential implementation of climate engineering in a just and legitimate manner.

Aside from whether suitable institutional structures exist at the global level to govern climate engineering, efforts to encourage public participation on this subject will face significant challenges, including the fact that this is an emerging set of technologies with low levels of public awareness. Given this, methods of engagement such as public dialogue are more suitable than questionnaire surveys, for example, since they allow the possibility to both raise awareness and solicit informed responses from participants. Two studies applying deliberative methods have been carried out involving the structured presentation of factual information on climate engineering to small samples of the public (Pidgeon et al. 2013; Macnaghten and Szerszynski 2013). These studies suggest a position of “qualified” or conditional acceptance in which support for small-scale research does not correlate with support for large-scale deployment, as well as strong concerns about the virtue of “interference with nature,” issues of reversibility, and control. Furthermore, informing the public regarding more critical discourses, including past attempts to shape weather and climate and allowing deliberative discussions of the implications, led to a lessening of support for research (Macnaghten and Szerszynski, 2013). This demonstrates the significance of “framing effects” in public deliberations: public acceptance of climate engineering will be influenced by how it is presented, what issues or technologies it is associated with, and so forth.

While these studies are a useful start, the fact that both were conducted in the United Kingdom shows a clear limitation in terms of geographical scope – there is a dearth of social science research on climate engineering conducted outside of developed country contexts. But specific forms of climate engineering (e.g., ocean iron acidification) raise questions about who should participate and how they might do so. Who are the affected public that should be engaged with regarding either research or actual deployment of such techniques in a context such as the Pacific Ocean? And what mechanisms could be employed to enable their engagement? This is an important issue since physical science modeling of the potential impacts of climate engineering shows the possibility of negative climate impacts on other global regions; for example, research has found that releasing aerosols into the atmosphere in the Northern Hemisphere (a technique known as solar radiation management) could lead to drought in Sahelian regions of Africa. It is unsurprising, then, that research on climate engineering has revealed public concerns about the potential for international conflict following unilateral actions by nation states, together with skepticism concerning the ability of national governments and international institutions to effectively govern interventions such as solar radiation management in light of stuttering progress on climate mitigation (Macnaghten and Szerszynski, 2013).

It is also evident that new mechanisms – likely involving the use of Internet and digital media that can enable the participation of both larger and more diverse samples of the public – will be required to feed into and inform future efforts to govern climate engineering at the supranational scale. One exemplar that
could be followed is “World Wide Views,” an initiative that emerged from methods of technology assessment developed in Denmark and which has been implemented on environmental issues such as global warming (see http://www.wwviews.org/). The deliberation on global warming involved 4400 participants from 38 countries. Although it was successful in involving and collating the opinions of citizens at an unprecedented scale, it raises many challenges, such as the method and scope of evaluation, sample representativeness (e.g., both China, with a population of 1.3 billion and St Lucia, with a population 161 000, each had roughly 100 participants), standardization across diverse economic, political, and cultural contexts, and a lack of country-level information. It is also costly – the tentative budget for an envisaged 2015 global deliberation on energy and climate is 7.2 million euros, with the aim of enabling participation of a total of 10 000 individuals in 100 countries. Nevertheless, as an experiment, it does point to ways that future public deliberations on global scale environmental problems and potential solutions to these problems might be implemented.

SEE ALSO: Environment and the state; Environmental citizenship; Environmental governance; Participatory geographies

References


Environment, nation, and “race”

Lisa Palmer
University of Melbourne, Australia

The intertwined concepts of environment, nation, and “race” have long fascinated and preoccupied geographers. As a trio of concepts, their historical contingencies and interrelationships provide critical insights into geography’s core disciplinary concerns of place, social and cultural transformation, and human–nature relations. Always being negotiated as a result of social and political change and power, the inherent instability of ideas of environment, nation, and “race” remain central to twenty-first-century processes of environmental governance.

Writing nature, controlling environments

The first geographers, loosely defined as those who sought to write about the world (geo = earth, graphy = description), were travelers and explorers like Herodotus, Marco Polo, Columbus, and perhaps most famously the lettered entourage accompanying the eighteenth-century “scientific expeditions” of James Cook. Handmaidens of empire, the men who followed set out to describe, catalog, and order new nature and new peoples for an attentive and knowledge-hungry European public, fashioning the world less through imperial plunder and more through what Pratt (1992) refers to as a scientific “anti-conquest,” or the hegemonic possession of other people and places through the ideas and gaze of objective science. This period between the mid-eighteenth and the late-nineteenth centuries was the triumph of an empirical naturalism which produced systematic accounts of nature and humans (Pratt 1992). Meanwhile, the early nineteenth-century accounts of the German geographer Alexander von Humboldt, and his enormously influential South American travel accounts of “wild and gigantic nature,” introduced the European public to an entirely new style of nature discourse to narrate the science of the natural world. Humboldt’s writings, which fused science with the Romantic literary tradition, conjured up “a dramatic, extraordinary nature, a spectacle capable of overwhelming human knowledge and understanding” (Pratt 1992, 120). His accounts stimulated an interest in travel among many other artists and naturalists, most notably Charles Darwin, and legions of other men who set out to document and describe the environments and peoples they encountered in these new worlds. By their contributions to “knowledge,” they also enabled the possession of immense territories, fertile abodes awaiting the imprint of “civilized” man.

This modern scientific inquiry and method expounded by Enlightenment thinkers influenced the responses of explorers to the non-European peoples they encountered (Thomas 1994). Whereas the period up to Columbus had identified the peoples of other lands as “barbarians,” this was not a difference certified on racial or national grounds. Rather, the distinction was drawn on religious grounds, in comparison with Christianity, where there was always the possibility that the “barbarians” would be converted. However, modern ideas “entailed
new models for constructing otherness; these were couched in a narrative of natural history rather than salvation and privileged distinctiveness in character and physique rather than faith” (Thomas 1994, 77). Gradually, through the investigation of scientists and others, Europeans became aware of factors of physical and social differentiation between people, and this awareness corresponded with the emergence of the organizing idea of race.

In the late nineteenth century social Darwinists decreed that Darwin's theory of natural selection also applied to the social and political realms. This led to the emergence over time of another key geographical concept, that of environmental determinism or the idea that the “environment” has a dominant or controlling effect over human characteristics and society. This idea justified the dominance of imperial powers over colonized regions as a result of their own superior environmentally conditioned racial characteristics. Colonial underdevelopment and the moral backwardness of local populations could, it was hypothesized, be overcome via colonial improvement to these physical and social landscapes.

Meanwhile in Europe, the ascendancy of the concept of the nation-state in the sixteenth century saw the meaning of landscape (Landschaft), understood until then as the commonplaces of “a polity’s area of activity,” gradually change to mean scenic spaces unified under the single body politic of the nation-state (Olwig 2002). By the nineteenth century the doctrine of the nation-state excluded recognition of alternative traditions of governance and sovereignty, including the customary polities of those encountered in the new world who had their own relationships with and configurations of landscape. This positivist legal doctrine of civilized nations became the primary intellectual instrument of nineteenth-century colonialism, and up to today the nation-state remains the normative model for the recognition of sovereignty in global politics.

In the 1970s, reflecting on geography’s complicity in imperial discourse and in the colonial disciplinary machinery, human geographers began to critically address the historically contingent imaginings, impact, and continuing import of the intertwined concepts of environment, nation, and “race.” Out of the postcolonial concern to theorize ongoing processes of neocolonialism emerged powerful geographical critiques of empire, nature, and cultural difference (see Postcolonial geographies). In the vast “emptiness” of the settler nations of Australia, the Americas, and Africa these geographers and other cultural theorists argued that landscapes remained colonized as much by settler national imaginaries of “nature” and “culture” as by legal and political institutions. Material possession was shown to be conjoined with powerfully wondrous, subliminal, and Romantic colonial discourses and images. These tropes of Romantic redemptive nature and noble savages were, moreover, shown to still permeate the imaginaries of a new breed of late twentieth-century nature-loving activists and “explorers.”

“Race,” nation, and primordial nature

As a part of the cultural turn in socio–nature studies, geographers increasingly paid critical attention to the lingering effects of the Enlightenment and the associated Romantic othering of nature and primitives, to the production of the idea of natural landscapes and environments, and eventually to a critique of settler tropes such as wilderness. They interrogated the ways in which the changing value of land and human labor under industrial capitalism in the eighteenth-century European “home” environment saw an emergent
ideological appreciation of wild lands. As the moral order of society, configured as the domain of social relations, was displaced by an alienated and “unnatural” economic order, the idea emerged that if morality was to be discovered it must be in “pure” nature, where human society had not intervened. A central part of the English celebration of national pride was shown to have developed in the late eighteenth century through the English Romantic poets, for example Wordsworth, who venerated the landscape of the Lake District, and the picturesque rural idyll combined with a mountainous wild nature. In the United States, geographers developed the thesis that a Romantic appreciation of nature was an integral part of the forging of national identity, and natural vast landscapes known as “wilderness” became a dominant imprint of (white) national identity on the American mind. These Romantically venerated environments in Europe, the Americas, Australia, and elsewhere became something to be seen and appreciated by vista-seeking tourists, a process Matless terms “a pilgrimage of scenery” (1995, 107).

Writing from the edge of empire, Jacobs argued that in the twentieth century, particularly in the West, nature became a sought-after commodity: “Part of the legacy of the cocktail of Enlightenment thinking and the transition to capitalism was the invention of ‘external,’ ‘primordial’ Nature” (1996, 135). By the end of the century this fantasy of the primordial was played out through capitalism’s appropriation and mediation of nature in the discourses of conservation, sustainability, tourism, and environmentalism (see Environmental discourse). In this popular reality, nature and associated images of wilderness landscapes were treated by capitalist cultures as a commodity that was both good for people and good for business. Capitalism created objectivist categories, such as primordial nature, and encouraged the culturally naive acceptance of an objective or reified world existing outside social relationships.

Within the global capitalist imaginary, images of nature and of primitive peoples continue to be subjected to a process of ideological mystification that Marx called “the fetishism of commodities.” Postcolonial geographers and others have shown how settler nation-states have drawn on such reified “natural” assets as unique flora, fauna, and landscapes, as well as emblematic images of “first peoples,” to provide their nation with a sense of national antiquity, purity, and distinctiveness. These now highly valued “indigenous” attributes have become central to new environmental moralities which view the presence of introduced species (and nontraditional practices) in wilderness as mongrelizing the environment. In contrast, many writers have documented the ways in which indigenous peoples embrace this transformed ecology in their “wilderness” homelands, understanding introduced species as coming to “belong” through social relationships and historical contingency. At the same time, settler societies in their new environmentalist guise have been shown to place such ideological value on pristine environments that they are “reluctant to grapple with the conceptual demands and responsibilities of new ecosystems, couching many environmental debates in terms of returning things to a previous state” (Head 2000, 231). As nature is valued in its “primordial” state, so, too are indigenous peoples valued in their primordial state as “timeless” hunters and gatherers who speak “authentically” for the unique landscape of the nation. These voices are, however, authorized to speak (politically and culturally) only from outside of history and preferably in translation. In this process of translation, others have shown how “indigeneity” is made and remade over and over in a process of intertextual dialogue, imagination, representation, and interpretation (Langton 1993). These dialogues,
drawing on the colonizing acts of explorers and pioneers, reconstitute the fetish of indigeneity in the contemporary preservationist ethos of the environment movement. As Jacobs argued, this discourse of primordial nature cemented its “othering” by placing indigenous people back in First Nature as the First People of the world who were “then reabsorbed into a global chronology of planetary survival, which begins with ‘them’ but ends with an environmentally sound ‘us’” (Jacobs 1996, 137).

“Race,” nation, and environmental governance

By the late twentieth century, environmental governance in the (neo)liberal public sphere of nation-states was critiqued as underpinned by the powerful effects of a racialized and naturalized concept of nature, what Braun (1997) termed a “culture of nature.” In such Western scientific commonsense categories of “nature” and “resources,” natural landscapes were shown to be an empty space to be accounted for, and the authority of Western science became the basis for moral and political ideas and practices of resource use and conservation. In this schema indigenous and local people’s concepts of nature (replete with integrated social systems) linked with local cultural and political contexts were (and are) displaced in favor of the authority of the “nation” or the “public” to speak for and manage an unmediated nature (Braun 1997). This managerial Western culture of nature underpinned by science and the “anti-conquest” came to constitute the “commonsense” view of environmental governance. Recognition of “culture” in such “natural environments” was ceded only as an exotic overlay on a universal and pre-existing nature.

In settler societies such as Canada and Australia, despite legal regimes which seek to recognize indigenous property rights, the pervasiveness of this “culture of nature” continues to impede the recognition of indigenous regimes of social and environmental governance and indigenous peoples’ full participation in local, regional, and national natural resource management regimes. Simultaneously, as many geographers have noted, the ongoing reification and racialization of indigenous “subsistence” communities in such contexts creates them as naturalized prepolitical and pre-economic spaces excluded from the politics of governance and the full recognition of their economic rights.

Others have argued that since the late twentieth century an emerging multiscalar values-based politics has resulted in the rise of civil society and nonstate actors who are actively challenging the nation-state’s dominance on issues of governance and sovereignty, particularly in relation to environmental issues. Governance, they argue, is now less linked to territorial control. Yet, as other scholars have noted, many indigenous peoples continue to construct their worlds around vibrant socio-natures and deep relationships with particular places. Such relationships to place are configured through a politics of unique eco-social rights and responsibilities enacted in a multitude of dynamic, practical, and creative ways. The continued observance of many indigenous peoples to socioterritorial forms of political formation challenges both the ideology of the nation-state and the democratic ideals of an emerging multiscalar (nonplace-centered) civil society.

In this sense, a range of theoretical contributions have elucidated the political inadequacy of democratic projects that seek to merely create spaces of recognition and inclusion for indigenous peoples within existing nation-state and local-level environmental governance arrangements. In contrast to such inclusionary social justice models, these theorists argue that more
thoroughgoing political engagements are needed to refashion the norms and principles of existing environmental governance arrangements. Especially in the context of postcolonial settler societies, geographers have highlighted the importance of understanding so-called natural environments as political, as embedded in social relations. This body of research shows that indigenous peoples’ relations with place and nonhuman nature are not only expressions of deeply felt attachment to particular environments, nor are they relics of a conservative, closed, and parochial past; rather, these relations are shown to be central to the constitution of dynamic and engaged indigenous lifeworlds. It is argued that it is only because of postcolonial contexts of unequal symbolic and economic power that the formation of distinct fixed identity boundaries such as indigenous/nonindigenous has emerged as a necessary political strategy for cultural and economic survival (see Ethnicity).

Others have argued that, viewed from the perspective of diverse societies, liberal, nationalist and communitarian theories of modern constitutionalism serve to reinforce the hegemony of the European nation-state tradition, giving only the appearance of constituting diversity. As a result of these failures, Borrows (2002), in his book on the resurgence of Indigenous law in Canada, considered the need for refashioning the underlying principles and normative features of the Canadian Federation. He argues that a limited settler conception of the jurisdictions of indigenous polities has contributed to the failure of the settler state to fully incorporate indigenous governments as active and meaningful participants in the Canadian Federation, and suggests that a useful starting point in the process of creating a shared and multidimensional citizenship would be for non-Aboriginal Canadians to recognize and incorporate the indigenous notion of “landed citizenship” wherein “the land is animate and perceived as having rights and obligations in its relations with humankind” (Borrows 2002, 146). Arguing that for indigenous peoples the politics of place-making is based on the need to recognize and negotiate relationships with significant others including nonhuman nature, Borrows writes that Aboriginal people’s values where the land is concerned “could to be entrenched in Canada’s governing ideas and institutions” (2002, 146).

Indeed, as many geographers have shown, it is often through environmental governance issues that indigenous peoples are starting to refashion their stake in the governing ideas and institutions of regional, provincial, and national polities. As these institutions of the nation-state engage in a new multiscalar and values-based cultural politics, risk and opportunity abound for indigenous peoples. These new spaces of environmental governance can enable productive encounters between diverse political actors, creating places where indigenous peoples can begin to engage in conversations about their rights and ways in which these rights can reconfigure the whole. In contrast to the suppression of indigenous peoples’ moral and political subjectivities in state-centric environmental governance arrangements, these more open environmental governance approaches signify a political moment for new arrangements of power to be created.

Critical ecologies and socio-natural difference

With the ascendancy of new environmental governance processes, the concepts of environment, nation, and “race” are also perpetually in motion, responding to and reasserting themselves in the light of historical contingency. Meanwhile, a new critical ecologies literature works to denaturalize both the environmental
ENVIRONMENT, NATION, AND “RACE”

governance project and nature itself. Given the increasing recognition that to naturalize “nature” is to paralyze politics, in these new multinatural configurations of “nature,” geographers have begun revisiting the cultural turn in socio-nature studies (Lorimer 2012). They seek to weave its key ideas into a plurality of socioecological worlds and to explore the diverse ways of being (ontology) and knowing (epistemology) which frame environmental governance possibilities. Going beyond conventional Marxist notions of political ecology (underpinned by a concern to interrogate matters of political economy as they relate to natural resource use), these critical ecologists are interested in “the multiple trajectories along which any ecology might evolve and the various ways in which they can be sensed, valued and contested” (Lorimer 2012, 594). Many of these writers are concerned with the question of “what it means to be and become human today, in dynamic relationship with non-human worlds” (Sullivan 2009, 24). Such a question is profoundly connected, as Kay Anderson (2008) has shown, to the history of the idea of race. As well as moments for asserting racial rule and Enlightenment othering, colonial encounters with nature and the racialization of aboriginal peoples in places like Australia were critical to the ways in which European classificatory schemas grappled uneasily with ideas and assertions of humans as nature-transcending and as distinct from the nonhuman world (Anderson 2008).

Wishing to bring the new posthumanist critical ecologies literature into conversation with practices of environmental science and governance in the Anthropocene, Lorimer (2012) proposes the necessity of developing new biogeographies through an engagement with critiques of nature which are sensitized to nonhuman difference and to ontological politics and frictions. The aim, he argues, must be to think beyond the preservationist species presentism, or steady statism of biodiversity conservation and the social and economic distributive justice concerns of political ecology to include accounts of the agency of nonhuman nature. At the same time, in order to sustain critical attention on the politics of human difference, Lorimer acknowledges that there is a need to retain the “sophisticated understandings … of political and regulatory contexts” developed in the work of political ecologists and others (2012, 606).

Taking up the challenge of tracing an ontological politics, Lorimer writes that emerging multinatural geographies are concerned with the posthumanist ways in which context-specific forms of ethical and political action and responsibility produce “respect for the radical alterity and unpredictability of organisms, their ecologies and the multiple constituencies who have a stake in their conduct” (2012, 604). His discussion explores the way that an emerging posthumanist ethics strives to recognize a concern for “flourishing,” even seeking in the tradition of ecocentricism to hold firm to a distinction between the human and nonhuman as an ethical space of action. Yet, in this new critical ecological sensitivity to diverse ontologies, we are also cautioned to remember that specific ontologies are not theories of an imagined world but specific ways of enacting and bringing a world into being. As such, in their search for new social and political imaginaries, social studies of science theorists and others from geography and anthropology have argued that theories of ontological pluralism are the praxis for radically decolonizing thought.

While the new critical ecologies literature clearly breaks with past deployments of the intertwined concepts of environment, nation, and “race,” posthumanist thought often has little to say about existing forms of human and nonhuman sociality within alternative and long established forms of property relations. While
foregrounding the importance of nonhuman agency reconceptualizes both nature and the environment as a part of an ethical constituency, it also risks recreating a homely universal nature that has little to say about socio-natural configurations wherein nonhuman agency and property already link together humans and nonhumans in specific rights-based configurations. For example, in their ethics and politics of recognizing nonhuman presence and agency, many indigenous groups already have clearly defined rights-based pathways for communicating with and relating to nonhuman nature. As such, newer understandings of posthumanism propounded by the critical ecologies literature need to be integrated with careful ethnographic work which draws out in a place-specific sense what forms and trajectories of difference matter, who decides, on what grounds and through what processes (Lorimer 2012). As we look to the future, it is also clear that we need to look to the past and carefully understand the present to ask questions about the impact and beneficiaries of different multinatural futures.

New environmental governance possibilities

The approaches of the critical ecologies literature raise important questions as well for how nation-states and transnational conservation organizations recognize rights to “speak for” nature. Here, more than ever, others have argued that it is how we engage in relations that is critical to the reconfiguration of historical concepts of environment, nation, and “race.” In spite of the formal reserve focus of global conservation and environmental governance, most “protected areas” in the world are not a formal part of the global protected areas network; rather, they are areas which have long been managed and conserved by indigenous and local peoples. Tallied up, these community conserved areas would exceed the size of the total area of formal protected area reserve globally (Stevens 2014). While community-based natural resource management has now long been trialed by an array of international conservation and environmental governance initiatives, there have been very mixed results. According to Stevens (2014), the failure of these processes to generate success in the area of conservation (and their linked agenda of development) is because these programs are mostly concerned with incentivizing people to follow the conservation agenda of others and lack a lexicon of keywords to transform lingering colonial relations, including “justice,” “rights,” “equity,” “collective tenure,” “restitution,” “self-governance,” “cultural integrity,” and “indigenous knowledge” (Stevens 2014). Escaping from the legacy of its racial origins, twenty-first-century global conservation policy has at least now shifted toward a recognition of community-controlled nature conservation which has at its core a rights and social justice agenda. While this change was brought about by the skillful political intervention of indigenous peoples themselves, there is, according to Stevens, still much “political work ahead” (2014, 80).

As others show, the insidious history of colonial relations is such that emerging environmental governance regimes are often (unthinkingly) reinstating past practices of domination and control over racialized others. Baldwin (2009) has coined the term “carbon nullius” in his consideration of the implications of apolitical and ahistorical approaches to climate mitigation in Canada’s boreal forest. He argues that emerging carbon sequestration discourses and practices have manipulated concepts of “race,” nation, and environment, enacting “a political geography of racial difference, one that seeks to
accommodate an imagined mode of traditional aboriginal life to the exigencies of global climate change mitigation and, importantly, to a logic of global capital now well into its ecological phase” (Baldwin 2009, 231). Examining this shift to market-based environmentalism and the continued centrality of forests to the global environmental imaginary, Baldwin’s work draws out the ways in which racialized understandings of nature and its others inform national responses to climate change mitigation through an engagement with new capitalist logics such as carbon markets. He shows how, as Canada’s boreal forests enter global carbon markets as carbon reservoirs and wilderness zones, Indigenous peoples and their practices of environmental governance are alternatively elided as absent or contorted through a veil of managerial environmental governance which cedes them a history but not a politics or an economic stake. He writes that the “imperial racism of terra nullius [the empty land] is re-enacted at the very moment the space in question is reordered through a discourse of climate change, carbon and the corresponding taxonomy of exchange value” (Baldwin 2009, 241). This suggests that “carbon valorization, racialization and dispossession are here not just coincident processes but constitutive features of the same process” (Baldwin 2009, 242). Such instances of racial violence in emerging modes of market-based environmentalism (or ecological modernity) have given rise to a far-ranging geographical critique, much of which addresses the dangers of the marketization of nature and its failings, particularly for local and indigenous peoples. At the same time, other writers and activists, conscious of heterogeneity and the ambivalence of these trends and their applications in diverse environments, have also been keen to interrogate such processes for their social, economic, and environmental possibilities and for their potential capacity to circumvent dominant concepts of environment, nation, and “race.”

What is clear is that, in the most pressing realms of global environmental governance, such as climate change and associated policy debates, immutable understandings of “nature,” while useful for some agendas some of the time, continue to reify the world, excluding and dominating its others. As the world embraces a late capitalist ecological phase of market-based environmentalism, Neil Smith (2007, 38) has powerfully argued that “Capital is no longer content simply to plunder an available nature but rather increasingly moves to produce an inherently social nature as the basis of new sectors for production and accumulation.” In these ongoing processes, for better or worse, the entwined concepts of environment, nation, and “race” will continue to be materially and culturally reconfigured in old, new, and creative ways.

SEE ALSO: Environmental discourse; Environmental governance; Ethnicity; Nature; Nature conservation; Postcolonial geographies; Posthumanism; Socio-nature

References


Environmental (in)justice

Noriko Ishiyama
Meiji University, Japan

Environmental (in)justice is a philosophical and political notion created and espoused by activists, policymakers, and academics to articulate important connections between environmentalism and place-based struggles for social justice. The sites of environmental (in)justice occur at different geographical scales ranging from the individual bodies of humans and nonhumans to global ecological arenas past, present, and future. The concept of environmental (in)justice challenges the narrowly defined notion of environment focused solely on wilderness in a US cultural context, and is also embedded in geographical and historical processes, social relations, and political-economic power structures, which are themselves intertwined with issues of class, race, gender, sexual orientation, and colonization. Scholars and activists involved in the field of environmental (in)justice analyze the ecological dilemmas experienced and reproduced at the sites of daily living—in other words, the spaces of home in which we actually live. Their aim is to enable peoples and groups with different cultural backgrounds to live their lives without worrying about environmental threats and risks relating to their socioeconomic status, as well as to participate in environmental decision-making forums for addressing and resolving issues of environmental (in)justice. The following introductory story illustrates the complexity, difficulties, and possibilities inherent in environmental (in)justice struggles.

In 1975 Dr Masazumi Harada (1934–2012) visited Grassy Narrows and White Dog, First Nation reserves in Ontario, Canada, for the first time. Harada was a medical doctor and scholar from Japan who specialized in Minamata disease, a neurological condition caused by methylmercury that had been dumped into the sea as industrial wastewater from the 1940s to 1968 by the Chisso Corporation, located in Kumamoto Prefecture. The discharged mercury accumulated in fish and shellfish that were then consumed by coastal communities around Minamata Bay, in the southern part of Japan. Far away from the political-economic center of the nation, Minamata was a place inhabited predominantly by low-income fishermen and their families, who fell ill as a result of their exposure to mercury and yet remained invisible in the political processes concerning environmental decision-making at the local, regional, and national levels.

Across the globe, in small indigenous communities in Ontario, Harada found a number of patients suffering from neurological, physical, and mental disorders also caused by mercury that had been emitted from the caustic soda plants operated by the Dryden Chemical Company, part of the local paper mill located upstream along the Wabigoon River, between 1962 and 1970. At the request of the First Nation communities, Harada and his team from Kumamoto Gakuen University conducted follow-up clinical surveys in 2002, 2004, and 2010. Their studies observed the prevalence of symptoms similar to those that had been observed over several generations in the Minamata region (Harada et al. 2011).

In both cases, the livelihoods of the surrounding communities depended heavily on
ENVIRONMENTAL (IN)JUSTICE

fishing. Mercury poisoning via water devastated public health, local economies, and cultures in these already socially marginalized populations. The sick individuals remained unrecognized, neglected, and almost abandoned by the municipal and national governments, which for many years prioritized corporate interests and industrial growth over environmental protection. Recalling his transnational experiences as a medical professional, Harada later told a newspaper reporter based in Tokyo: “I used to think that Minamata patients were discriminated against because they were sick. However, after visiting a number of contaminated sites all over the world, I now realize that environmental pollution haunts places where discrimination already exists” (Harada 2011). With his clinical expertise, as well as first-hand experience of diagnosing and offering medical treatment to thousands of patients, which included testifying for the victims in numerous lawsuits filed against Chisso Corp. and the prefectural and national Japanese governments, Harada knew exactly what was going on in Ontario – namely, that it was another case of environmental injustice.

Even though Harada did not use this particular vocabulary to articulate his struggles as a doctor and an advocate for grassroots movements, what he had witnessed and experienced on various geographic scales brought to the forefront issues of environmental injustice across time and space. The stories arising in Minamata and the Canadian First Nations highlighted the intertwining dynamics involved, such as ecological devastation in relation to human as well as nonhuman bodies; anxiety over reproduction, local food and cultures, livelihoods and resources, and the political economy on a larger scale; state support for the accumulation of corporate capital; lack of societal recognition of issues and peoples at the social periphery, as well as the structural limitations on their participation and capabilities in terms of environmental decision-making; classism, racism, sexism, heterosexism, and colonialism underlying the unequal power structure that developed spatially through the course of regional and national history; international and intergenerational ecological problems that threatened sustainable futures, as well as the growing networks among activists and scholars. All of these factors contain elements that inform the notion of environmental (in)justice, an issue which is profoundly geographical.

Since the concept of environmental (in)justice was first recognized and gained legitimacy in the United States through activists, public policy, and scholarly circles, as well as through media coverage, this entry primarily addresses US cases. It will review the emergence of the environmental justice movement, which aims to resolve dilemmas concerning environmental (in)justice, and the historical roots of the movement; the environmental justice movement as an antithesis against traditional environmentalism; policy development; and scholarly exploration of environmental (in)justice as a theoretical framework. The final section on academic theorization clarifies the pitfalls of the distributive paradigm combined with the overt racism prevalent in early studies, and then introduces the growing field of critical studies on environmental (in)justice.

Emergence of the environmental justice movement in the United States

As exemplified in the story of Minamata disease, communities with diverse histories and cultures all over the world have been fighting for centuries against environmental injustices that threaten their welfare, even though they have not referred to it by this particular term. Nonetheless, the development of environmental (in)justice as an activist agenda and a theoretical
framework is relatively new. The origins of the movement, along with the creation of this specific term, are found in the United States. Many activists and scholars point to the non-violent civil disobedience against a major case of environmental injustice – the construction of a hazardous waste landfill in Warren County, North Carolina, in 1982 – as the official birth of the environmental justice movement. The story dates back to 1978, when truckloads of liquid contaminated with polychlorinated biphenyl (PCB) were illegally dumped along a rural state highway. Consequently, 400,000 cubic yards (305,822 m$^3$) of soil along 240 miles (386 km) of road in 14 counties were contaminated with PCBs, and the state government was forced to take the lead in cleaning up the mess. Eventually, the government designated Afton, in Warren County, as a future dumping site, but made this decision without local consent. One of among 90 proposed sites, Afton was an economically depressed town in a county with the highest percentage of African Americans in the state of North Carolina. In the following year, angry residents and their supporters filed lawsuits to block the disposal plan. Their litigations failed and, as a result, hundreds of trucks loaded with hazardous waste started coming into Afton in the fall of 1982.

Lying on the ground in front of the trucks, residents and their allies tried to stop the disposal plan, which constituted a major health threat to the surrounding communities. The civil disobedience continued for six weeks, but eventually the police were called in and more than 500 demonstrators were arrested. While the construction project proceeded and the landfill was eventually built, the nonviolent protests, initiated by local blacks and supported by a number of civil rights and antitoxics activists, were reported on a large scale. Moreover, the Warren County event rallied a number of people who considered environmental destruction as a violation of basic human rights to live, work, and play in safety (Bullard 1990; Taylor 2011).

At the request of the leaders of the Warren County protests, in 1983 the United States General Accounting Office (known as the US General Accountability Office (GAO) after 2004) conducted a sociodemographic study in eight southeastern states, including North Carolina, of the relationship between the location of hazardous waste facilities and the racial and economic characteristics of the neighboring communities. The research concluded that hazardous waste landfills were located disproportionately in low-income black communities (US General Accounting Office 1983). Four years later, in 1987, United Church of Christ published another now classic study of environmental (in)justice, which revealed a national pattern in which race was determined to be the major factor in the siting of commercial hazardous waste facilities (UCC 1987). The Warren County protests and the subsequent GAO and UCC studies gave credibility to the grassroots activists’ claims. These developments marked the beginning of the environmental justice movement throughout the United States.

One highlight of the early environmental justice movement was the First National People of Color Environmental Leadership Summit, held in October 1991. More than 1000 environmental justice activists and academics gathered in Washington, DC, to share their experiences of living with environmental injustice and to pool creative ideas to examine environmental issues in a social justice framework. Rejecting top-down and elitist approaches, the summit’s participants emphasized the importance of grassroots activism and personal accounts. Simply put, they believed that those who had been exposed to risks at pollution sites should also be present at the table of decision-making. Because these bottom-up
tactics have been prioritized and respected, the environmental justice movement has never been exclusively represented by particular individuals or institutions in authority at national or international levels, although a number of umbrella organizations do exist.

At the end of the 1991 summit, the participants, who had cultivated deep connections with local geographies and were beginning to form wide-ranging networks, drafted a list of 17 principles of environmental justice, which consisted of the specific goals they were committed to. The 17 principles embrace such diverse issues as ecological unity and the interconnectedness of all species; rights to self-determination, safe work environments, and health care; and the enforcement of informed consent in medical procedures. They object to the environmental damage and social injustices generated by multinational corporations and military occupation, and by the subjugation and abuse of local peoples and other life forms. The principles support education focused on social and environmental issues, with an emphasis on cultural diversity, while encouraging people to consume the least amount of natural resources and to produce as little waste as possible (Delegates to the First National People of Color Environmental Leadership Summit 1991). As clarified by Pezzulo and Sandler (2007, 5), the principles thus “called for a robust activist agenda and a wide range of spiritual, ecological, sustainable, educational, and social justice commitments. They articulate a desire for universal protection and self-determination domestically and internationally.”

This remarkable event garnered much attention, and encouraged policymakers at local, state, national, and even international levels to recognize the necessity of engaging with this issue. It is now seen as a symbolic turning point in the history of US environmentalism. In particular, the 17 principles became a landmark of a new course of direction in the environmental justice movement, inspiring a variety of groups and peoples at risk.

The summit and the principles presented an inclusive master framework that incorporated, inspired, and resonated with a number of social movements, thus going beyond the boundaries of particular regions or nations. Nevertheless, some scholars pointed out that they remained reformist in terms of setting specific agendas, and argued that the reformist, as opposed to the revolutionary, approach would not challenge the globalized capitalist political economy that perpetuated environmental injustice. For instance, environmental sociologist Robert Benford (2005, 50–51) presented a critical analysis of the movement’s initial reluctance to emphasize the need for radical social changes: “Rather than following the logic of its diagnostic framing by advocating sweeping systemic changes, the movement’s prognoses remain focused on tinkering with a system constructed from slave labor and predicated on the exploitation of disadvantaged people.” While these critiques raise important issues for further consideration, they lack historical perspective. In fact, environmental (in)justice struggles have a long history of challenging existing social systems and power structures.

Environmental (in)justice as a concept for societal change: historical perspective

The concept of environmental (in)justice was directly born from antitoxics and civil rights struggles in the United States in the late 1970s and early 1980s. Among those who participated in the Warren County demonstrations, for instance, were members of the United Church of Christ (UCC), which had been instrumental in the Civil Rights Movement in the 1960s, as well as Lois Gibbs, a mother from a
working-class white neighborhood who stood up and became a leader in the widely publicized Love Canal struggles in the mid-1970s. In order to understand the depth and complexity of environmental (in)justice as a concept, however, its history should be traced further back.

Sociologist Dorceta Taylor (2011) explores the movement’s origins, such as public health efforts in early American cities, antislavery activism, resistance against the appropriation of tribal land in the establishment of national parks, and efforts to develop antizoning and residential segregation in the nineteenth century. She astutely points out that environmental activism among people of color in the United States has often been overlooked “because their discourses and actions are not always couched in environmental terms” (2011, 280). Those who participate in social activism as activists, policymakers, or scholars too often do not realize that what they think is quite innovative and new has actually been explored by others years before in different temporal and spatial settings. It is, therefore, important to acknowledge the historical context of recent developments in the environmental justice movement and to clarify the sort of productive strategies that should be adopted for articulating solutions and accomplishments.

Both now and in the past, environmental justice activism initiated by various groups has consistently challenged unequal power structures as well as legal systems dominated by racist, classist, sexist, heterosexist, and colonial ideologies. Moreover, with their powerful and diverse stories and experiences, their activism has generated multiple meanings in different historical and geographical contexts. In short, at their core, concepts of environmental (in)justice embody an idea of political dissent against oppressions and exploitations in our daily lives as well as an idealistic commitment to achieve social change, with a special focus on those who are vulnerable and at risk of being overlooked or ignored by those participating in environmental decision-making processes.

Indeed, the achievement of environmental justice means that everyone, regardless of socioeconomic and cultural background, has access to safe and happy lives, as well as the hope of an even better future for the generations to come. Geographer Gordon Walker argues: “Looking across academic, activist and policy literatures, environmental justice is most often defined in terms of an objective, something that is sought after and for which certain conditions are specified” (2012, 8). Activists and scholars engaged with the environmental justice movement have been exploring a variety of ways to reconstruct bodies, homes, neighborhoods, and communities on a daily basis, by looking both back and into the future.

One pioneering scholar, sociologist Bunyan Bryant (1995, 6), defined this multifaceted term as follows:

[Environmental justice] refers to those cultural norms and values, rules, regulations, behaviors, policies, and decisions to support sustainable communities, where people can interact with confidence that their environment is safe, nurturing, and productive. Environmental justice is saved when people can realize their highest potential, without experiencing the “isms.” Environmental justice is supported by decent paying and safe jobs; quality schools and recreation; decent housing and adequate health care; democratic decision-making and personal empowerment; and communities free of violence, drugs, and poverty. These are communities where both cultural and biological diversity are respected and highly revered and where distributed justice prevails.

This notion of environmental (in)justice has thus developed in tandem with progressive politics aimed at empowering the socially vulnerable in order to safeguard and improve their everyday lives and spaces. It is clear that the geographical
spheres elucidated in Bryant’s definition of the term are different from those established by more conventional environmental movements.

The uneasy relationship with conventional environmentalism

Environmental justice activists have generally set distinctly different goals from those involved in traditional environmental activism aimed at the preservation of wildlife and the conservation of natural resources. Rather than looking at human experience in dialogue with what surrounds us, preservationists and conservationists have been seen as tending to objectify wilderness and/or natural resources as something to be worshipped, protected, or managed with professional knowledge and technologies. This result was often considered as a detachment of environmentalism from the social and political-economic processes that determine human and nonhuman lives, relations, and cultures.

An innately complicit relationship between environmentalism and eugenics has been an important part of the national identity of the United States. As noted by historian Jonathan Spiro (2009), many of the pioneering conservationists in the late nineteenth century were eugenicists who played instrumental roles in contributing to ideologies of scientific racism. While arguing that it was the duty of the superior white race to maintain the purity of the wilderness, they impeached people of color, new immigrants, and gays and lesbians for polluting pristine and safe environments, and also viewed such populations as a threat to US national security.

In consequence, although both conventional environmentalism and the environmental justice movement shared the core objective of preventing the destruction of environment, their agendas and participants did not necessarily overlap, nor they did always agree on interpretations of the concept of the environment itself. In his provocative and often cited essay, William Cronon (1995, 85), one of the most prominent environmental historians, reflects on this blind spot in mainstream environmentalism, and deconstructs the notion of wilderness:

But the most troubling cultural baggage that accompanies the celebration of wilderness has less to do with remote rain forests and peoples than with the ways we think about ourselves – we American environmentalists who quite rightly worry about the future of the earth and the threats we pose to the natural world. Idealizing a distant wilderness too often means not idealizing the environment in which we actually live, the landscape that for better or worse we call home.”

In order to fill in this gap, environmental justice advocates seek to change the social dynamic itself, by reconstructing spaces of home, rather than limiting their focus exclusively to physically and metaphorically distant mountains, lakes, or rivers that are visited on vacation or that are simply natural resources to wisely manage and use. Needless to say, this dynamic has generated dialectical relationships with political-economic power structures, as well as with prevalent ideologies nested at different geographical scales. For instance, “home” has therefore meant the lead-poisoned public housing in inner cities, pesticide-contaminated farmland worked by undocumented immigrants, deserted buildings and streets in post-Katrina New Orleans, urban parks from which LGBT people have been excluded and/or put under surveillance, and radioactive well water adjacent to abandoned uranium mines on and near southwestern Indian reservations. “Home” also encompasses radioactive land in Fukushima whose residents were forced to evacuate after the nuclear power plant disaster, and the atoll islands that have disappeared as a result of drastic climate change. At
the same time, in many parts of the world, the wilderness preserved in the national parks visited by thousands of nature-loving tourists has always been home to indigenous peoples. These are places where peoples with diverse ethnic, racial, and national backgrounds have lived or have wished to live their everyday lives.

Conceptualizing the environment as “home” enables us to see the reciprocal relationships that humans have developed with their surroundings which offer opportunities and capabilities. The economist and philosopher Amartya Sen has an illuminating interpretation of the values of the environment. He emphasizes the value of the environment as “the opportunities it offers to people” rather than as a matter of what exists (Sen 2009, 248). Another key issue Sen observes is that “the environment is not only a matter of passive preservation, but also one of active pursuit.” He stresses the possibilities of human power to “enhance and improve the environment in which we live” (249). Indeed, these are the values that have been appreciated and sought after by environmental justice advocates through the years.

Through a wide range of place-based experiences, environmental justice activists have observed and apprehended the interconnectedness of their homes with ecological risks, as well as with issues such as social invisibility, the globalized capitalist political economy, classism, racism, sexism, heterosexism, and colonialism, issues which are simultaneously material, institutional, and ideological. These are the environmental issues they confront, whose solutions require fundamental changes in existing social norms and practices. Indeed, in the words of David Harvey (1996, 372), environmental conditions represent “beliefs, institutions, social material practices, and relations, [and] forms of political-economic power.” As a consequence, critical perspectives on environmental (in)justice have redefined the conceptual framework of the environment as something that is constantly reproducing and restructuring social processes and relations.

Environmental justice advocates have criticized the conventional environmental movement for being dominated exclusively by middle- and upper middle-class white males and for having overlooked ecological issues threatening socially marginalized populations. One year prior to the First National People of Color Environmental Leadership Summit in 1990, supporters of environmental justice such as artists, writers, academics, students, activists, and representatives of churches, unions, and community organizations signed a public letter addressed to the “Group of Ten” national environmental organizations, including the Sierra Club, Wilderness Society, and National Audubon Society. They pointed out that these mainstream organizations lacked accountability with regard to communities of color, and that they received financial support from such polluters as ARCO, General Electric, Exxon, and Coca-Cola, multinational corporations notorious for disregarding the wellbeing of their workers and neighboring communities, as well as for destroying ecological health worldwide.

Sociologist Robert Bullard, often called the father of environmental justice, opened his first book, Dumping in Dixie: Race, Class, and Environmental Quality, with the following lines:

The environmental movement in the United States emerged with agendas that focused on such areas as wilderness and wildlife preservation, resource conservation, pollution abatement, and population control. It was supported primarily by middle- and upper-middle-class whites. Although concern about the environment cut across racial and class lines, environmental activism has been most pronounced among individuals who have above-average education, greater access to economic resources, and a greater sense of personal efficacy. Mainstream environmental organizations were late in broadening their base of support to include blacks and
other minorities, the poor, and working-class persons. (1990, 1)

It is noteworthy that the introduction of Bullard’s monumental work on environmental (in)justice, which has been quoted countless times in academic, activist, and policymaking circles for more than two decades, describes conventional environmentalism as a movement of the privileged.

The majority of the environmentalists under scrutiny, however, have espoused liberal politics alongside a critical view of the massive accumulation of capital resulting from rapid industrialization that comes at the expense of ecological health, particularly since the 1970s. Indeed, the organizers of the 1970 Earth Day, an event which helped to establish the Environmental Protection Agency (EPA) and led to the amendments of the Clean Air Act (1970) and the establishment of the Clean Water Act (1972), rejected donations from corporations. The 20 million participants in this extraordinary Earth Day event were not necessarily hostile toward low-income people of color; most of them did in fact support liberal political ideologies. However, from an environmental justice standpoint, those who had been endowed with upper- and middle-class white privilege, and had not shared the geographies of everyday livelihoods, struggles, and memories of the other, were seen as simply indifferent to or ignorant of the ecological problems that existed in social/locational peripheries.

In response, the environmental justice movement has challenged the existing environmental movement for its limited definition of environment and for its exclusion of the underprivileged from its leadership, agenda setting, and practices. Prompted by such harsh criticism, the Natural Resources Defense Council, Greenpeace, and the Sierra Club, along with other Big Ten organizations have been trying to improve communication with the environmental justice movement, to diversify their staffs and members, and to broaden their missions.

Some decry this trend to broaden focus, arguing that conventional environmentalism should not be conflated with the anthropocentric environmental justice movement but should instead remain focused on the protection of wilderness. After all, in their view, local and international efforts to protect wilderness areas are in crisis because of globalizing capitalism and, therefore, environmental movements cannot afford to be divided or distracted from their original objective, which is to speak for nonhuman nature and wildlife. While recognizing the accomplishments of the environmental justice movement, Kevin Deluca (2007 argues that this new activism does not support environmental efforts to protect nonhumans that may have negative outcomes for human rights or interests. What is even more problematic for Deluca is that “environmental justice groups, not content to have their own ‘environmental movement’ focused on people, directly challenge and berate the environmental movement for not focusing on people and their problems, in other words, for being environmental groups” (2007, 30). Hence, Deluca considers the recent policies initiated by large environmental organizations such as the Sierra Club and Greenpeace to integrate environmental justice issues with their objectives to be premature (27–28). Referring to the environmental philosophies elaborated by Henry David Thoreau and Aldo Leopold, Deluca emphasizes the importance of “recognizing that wilderness/nature grounds and circumscribes the human” (41). This leads to a claim that respecting wilderness/nature, which supports the very existence of humanity, should inform the central tenets of environmentalism.

Deluca raises important questions that could contribute to productive dialogue between environmentalists who maintain different and
sometimes contradicting agendas. His criticism of the anthropocentricity of the environmental justice movement remains unanswered. Yet Deluca’s anxiety about mainstream environmental organizations losing their focus on wilderness if they incorporate environmental justice activism reflects that dominant mindset that emphasizes a dichotomous relationship between humans and nonhumans.

Environmental justice advocates would say that it is precisely this artificial dichotomy that should be challenged. Moreover, they stress that environmental organizations that feel the need to speak only on behalf of wilderness underestimate their own vital agency. Those who are committed to environmentalism and/or environmental justice, they argue, should consider ways to deconstruct the simplistic dichotomy between wilderness and society, as well as that between human and nonhumans.

Environmental justice policy in the United States

The vocal activism of the environmental justice movement prompted US governmental agencies to reconstruct their policies, although critics complain that progress has been painfully slow and that much still has to be done to accomplish the goals set forth in the 17 principles ratified over two decades ago. The history of the US government’s efforts to engage with issues of environmental (in)justice reveals how difficult it is for state entities to (i) define this complex term beyond the distributive paradigm, and (ii) support the processes that would bring about structural and institutional change within a society dominated by a neoliberal political economy. Focusing mainly on the path taken by the EPA, whose role is to actively initiate and advocate for environmental justice, this section on the development of US environmental justice policy outlines both the progress and unresolved dilemmas.

Calls for environmental justice first gained the attention of policymakers at various governmental levels in the 1980s and 1990s, which was also when the concept of multiculturalism was spreading across the United States. As mentioned earlier, the GAO (US General Accounting Office 1983) and UCC (1987) studies revealed the connection between hazardous waste facility locations and host communities’ racial and economic backgrounds. As more stories of local dissent against environmental injustice made headlines, governmental officials attended conferences and meetings organized by activists and academics. Finally, in 1990 the EPA created an Environmental Equity Workgroup, which was followed by the establishment of the Office of Environmental Equity in 1992. The symbolic use of the term “environmental equity” has both laid the foundation for, and stymied, the US government’s efforts to correct the unequal allocation of environmental goods and bads. The EPA was forced to function within a strict and limiting framework of distributive justice, even though it had been trying to broaden its missions and scopes.

As it employed the term “environmental justice” more frequently, the EPA created an independent advisory committee in 1993 called the National Environmental Justice Advisory Council (NEJAC), which consisted of a wide range of stakeholders who provided the agency with advice and recommendations. Members of the NEJAC included, for example, community-based activists, officials from the state, local, and tribal governments, scholars, and individuals from businesses, as well as nongovernmental organizations engaged with indigenous issues, civil rights, and environmentalism.
Many celebrated the establishment of the NEJAC for offering a space where various stakeholders could make their voices heard. Environmental justice advocates would argue that it did not challenge the fundamental dilemmas embedded in the neoliberal political economy. Some scholars realized the limited usefulness of what they regarded as simply cosmetic changes in the policymaking arena. Pellow and Brulle (2005, 10), for instance, question the “participatory schemes that a neoliberal US EPA hatched during the 1990s to address EJ demands,” pointing out that the NEJAC and other advisory committees and task forces encourage a “lack of political power, inequalities among participants, [and] a drain on energy away from grassroots issues.”

While the EPA developed the organizational task force, President Bill Clinton issued Executive Order 12898 on February 11, 1994, which marked a major victory for environmental justice advocates. The executive order encouraged each federal agency to “make achieving environmental justice part of its mission by identifying and addressing, as appropriate, disproportionately high and adverse human health or environmental effects of its programs, policies, and activities on minority populations and low-income populations.” With the organizational support of the Clinton administration, the Office of Environmental Equity eventually grew into the Office of Environmental Justice (OEJ) in 1995. The executive order and the OEJ, however, do not have the power to compel federal institutions to actually implement environmental justice policies. Institutions are encouraged to incorporate environmental justice ideals into their projects, but they are not necessarily penalized for failing to do so or even for leaving the issue entirely unaddressed.

As noted by a scholar of environmental public policy, Edwardo Lao Rhodes (2003, 86), who conducted a personnel survey and examined an employment census at the EPA in the early 1990s, the majority of employees initially gave “little more than lip service to the concept of environmental justice.” The passiveness of the agency may partly be explained by the social and technical backgrounds of its staff. According to Rhodes’s study, whites made up almost 74% of the total EPA workforce and accounted for 95% of the executive ranks and 84% of the more senior ranks. Most of these were white men with college degrees in science, engineering, or law (Rhodes 2003, 94). Rhodes concluded that the poor record of the EPA in implementing environmental justice “was not so much the result of deliberate obstruction of justice or planned discrimination” as a “simple result of culture,” one that did not emphasize the social implications of their missions and policies (2003, 92). In short, the majority of the EPA workers with privileged backgrounds were not equipped with the firsthand experience to understand environmentalism as a social struggle to protect the spaces of home. In other words, they were not the ones whose bodies, livelihoods, and homes were at risk on a daily basis.

As a result, since the early 1990s, the agency has made little progress in the arena of implementing its environmental justice policy. In addition, eight years of the Bush administration’s disregard for environmental issues in general undermined the EPA’s capability in terms of human resources and budget. Under the Obama administration, however, the EPA has been trying to renew its mission to further engage with issues of environmental justice. In January 2010 EPA Administrator Lisa P. Jackson, the first African American woman to serve in this position, declared that expanding the conversation on environmentalism and working for environmental justice should be one of the EPA’s priorities.
In a proposed plan entitled EJ 2014, which also recognizes the twentieth anniversary of the signing of Executive Order 12898 on environmental justice, the EPA declared that it would attempt to: “protect the environment and health in overburdened communities; empower communities to take action to improve their health and environment; and establish partnerships with local, state, tribal and federal governments and organizations to achieve healthy and sustainable communities” (EPA 2011, 1). However, although the EPA emphasizes the growing need for strategic implementation, these newly proposed goals contain nothing new, either in terms of policy or philosophy, compared to those specified in the 17 principles compiled over two decades ago.

The administrative history of newly created offices and task forces, the 1994 executive order, and the recent EJ 2014 project all show that the federal government has officially recognized the issue of environmental (in)justice, as well as its historical negligence in addressing it. Dominated by a neoliberal political economy, however, policies have been restricted within a distributive justice paradigm to help overburdened communities to reconstruct their surroundings to have better and safer lives.

The administrative history is at odds with the potential inherent in the more progressive ventures to dissolve unequal power structures, oppressive social relations, and various ideological isms, all of which are fundamental to and inherent in the geographies of environmental (in)justice. Keeping in mind the political dilemmas clarified earlier, the next and final sections of this entry present the development of environmental (in)justice scholarship. Academic endeavors to articulate and pursue environmental justice have intertwined with the history of the movement and policymaking in interesting and instructive ways.

Emergence of environmental (in)justice scholarship

Since the dramatic birth of the US environmental justice movement, a number of important scholarly works have espoused the progressive visions championed by activists at the sites of resistance and struggle. As exemplified by Robert Bullard, a sociologist and former NEJAC member who is both a prolific writer and a passionate advocate of outreach activities, many academics working within a variety of disciplines and methodologies have played crucial roles in making environmental (in)justice issues visible to the general public and encouraging local and federal officials to engage with the prevailing problems.

At the same time, however, the majority of the early literature remained either uncritically celebratory of the new social movement, or superficial, addressing the problems in terms of the unequal distribution of risks combined with intentional racism. Some studies accused corporations and political institutions of being overtly racist and imposing unwanted facilities on impoverished and powerless communities of color, while others were preoccupied with deciding whether it was race or poverty that played a bigger role in the disproportionate mapping of risks. While these studies succeeded in raising provocative questions and conveying powerful statements of protest, they nevertheless remained theoretically underdeveloped.

As noted by a number of geographers, philosophers and others (Lake 1996; Ishiyama 2003; Schroeder et al. 2008; Schlosberg 2009; Holifield, Porter, and Walker 2010; Walker 2012), the distributive justice paradigm initially dominated the discourse on environmental (in)justice, in both academic and political arenas, in a powerful and problematic way. Needless to say, unequal distribution is a critical issue to be


**ENVIRONMENTAL (IN)JUSTICE**

explored and analyzed but it is only one issue among many.

Discussions about distributive justice tended to theoretically reduce environments into entities that could be divided between members of communities, societies, and nations. This reduction opened up the growing field of environmental (in)justice to easy criticism. For example, some academics questioned the legitimacy of the environmental justice movement in terms of a classic chicken-and-egg dispute: which came first, poor communities of color or hazardous facilities? Others tried to quantitatively prove that the aforementioned GAO (1983) and UCC (1987) studies, for instance, were invalid, and that hazardous facilities were not disproportionately sited in low-income and/or minority neighborhoods.

The mechanical quantification and mapping of the final outcomes of distribution neglected the underlying stories that could not and should not be sealed in a fixed and monolithic container. Accordingly, these endeavors ended up presenting rather sterile information and empty discussions based solely on demographic patterns, which were disconnected from the actual struggles of people at various sites. Schlosberg (2009, 45) points out that

> the literature and demands of environmental justice movements, in both the United States and globally, reveal that these movements are less enthralled with defining justice as solely distributional than are many theorists. A critique of the distribution of environmental goods and bads is certainly central to environmental justice movements, but unlike liberal theorists, movements tend to offer a more expansive, plural, and pragmatic notion of justice.

Activists realize that the simple redistribution of risks does not resolve the institutional and structural problems inherent in the processes that create environmental injustice. As Schlosberg points out, there is not one reference to an equal and fair distribution of environmental hazards anywhere in the 17 principles of environmental justice written in 1990 (2009, 49). Even with their liberal politics, early studies obsessed with the conceptual framework of distributive justice overlooked the movement’s core philosophies, which had, historically and geographically, arisen out of marginalized people’s everyday experiences.

Another persistent problem in the first wave of environmental (in)justice studies was a limited understanding of racism. As noted by Benford (2005, 40), the analytical framework of environmental racism was “evocative, provocative, and innovative,” as well as catchy, in terms of commanding the world’s attention. Nonetheless, the emphasis on the dichotomy between a society dominated by whites on the one hand, and victimized communities of color and the poor on the other, neglected the complex dilemmas relating to power structures and social relations among the people concerned. In contrast, people’s struggles to pursue environmental justice for the sake of social change generated a greater number of diverse and complex stories, which were grounded in multiple and anomalous identities.

For example, geographer Laura Pulido’s (1996) case studies on pesticide campaigns and grazing battles against mainstream environmentalists, both of which were initiated by subaltern Chicano communities, and Ishiyama’s (2003) research, focused on the desperate efforts of an American Indian tribe to accept a high-level radioactive waste storage facility for the sake of political-economic survival, reveal the marked and at times contradictory politics of identities, which play important roles in interpreting as well as reconstructing environments. The stories can be even messier, in fact, because justice for one group quite often means injustice for another, even within certain communities categorized
by race, class, gender, and sexuality. Oversimplified good-and-bad stories do not explain the impact of transformative identity politics on the multilayered pursuit for environmental justice.

Moreover, as some scholars have pointed out, the majority of the initial studies portrayed racism as a set of explicit actions that were fixed and without anomaly, but this portrayal neglected how racism works as an ideology, as a sociohistorical process, and as a form of cultural domination (Pulido 1996; Ishiyama 2003; Benford 2005). With the significant exception of Bullard (1990), who from the outset emphasized the institutional aspects of racism, many scholars were initially preoccupied with elucidating the unequal pattern of environmental hazards that victimized poor communities of color, and which were the result of intentional racism instigated by corporations and governments.

This perspective concluded that fundamental causes were to be found in the racist actions (and/or nonactions) of corporate and governmental decision-makers, who naturally denied such accusations. In other words, no environmentally destructive corporation would willingly admit that they were deliberately targeting certain communities to suffer for the benefit of others. As a result, the rather ridiculous question could be raised: If racist bigots were removed from decision-making processes, or if educated and privileged elites with political-economic power became more considerate toward the poor minorities, could environmental racism be eradicated and environmental justice eventually be achieved? The answer is obviously no. In addition to being characterized by overt gestures, racism also assumes various systematic forms rooted in ideas and structural forces. In the geographical and historical processes of environmental (in)justice, racism functions and is reproduced in the same way.

Critical environmental (in)justice scholarship

Moving beyond the scheme of distributive justice in tandem with overt racism, recent scholarship on environmental (in)justice has engendered more critical perspectives. As geographers Ryan Holifield, Michael Porter, and Gordon Walker (2010, 2) observe that scholars of environmental justice today “are situating their work with respect to far broader cross-disciplinary debates about knowledge, representation and meaning, engaging more substantially with explanatory social theory and utilizing a wider diversity of methodologies in investigating the material and political content of socio-environmental concerns.” This section introduces the expansion of critical environmental (in)justice studies arising from its intersections with theoretical considerations of procedural justice, along with recognition and capability, critical race theory, gender and sexuality, just sustainabilities, international/globalized environmental issues, and ecological justice.

Geographer Robert W. Lake was one of the first scholars to articulate the structural issues neglected in the distributive justice paradigm and to propose an alternative perspective with an emphasis on procedural justice. As he writes (1996, 196), “redistributing outcomes will not achieve environmental justice unless it is accompanied and, indeed, preceded by a procedural redistribution of power in decision-making.” Analyzing environmental (in)justice regarding the siting of hazardous waste facilities, Lake stresses the importance of investigating how communities were capable of making their own choices if they were given control over the production of the waste in the first place. In other words, clarifying the social processes, political-economic structures, cultural identities,
and ideologies through which communities participate in or are excluded from decision-making articulates the determinants of environmental (in)justice (Ishiyama 2003, 122). Mathematical readings of the final distribution of hazards mean little when we compare them with careful examination of the spatial dynamics that have created the oppressive conditions and hegemonic ideologies underlying certain forms of inequity.

In his groundbreaking theoretical analysis of the multifaceted concept of environmental justice, David Schlosberg (2009) highlights the procedural interplay of equity, recognition, participation, and capabilities, at both individual and group levels. As an environmental philosopher, Schlosberg emphasizes the importance of considering the complex characteristics of the environmental justice movement, since “one simply cannot talk of one aspect of justice without its leading to another” (2009, 73).

In addition to equitable distribution, conceptual discussions about environmental (in)justice should incorporate the notion of recognition. According to Schlosberg (2009, 16),

- the key to understanding recognitional injustice lies in understanding the social norms, language, and mores that mediate our relation between those who are denigrated and so less well-off in the scheme of justice. The argument is that mis-or malrecognition is a cultural and institutional form of injustice.

Various social movements have thus demanded the recognition of and respect for socially marginalized communities, whose cultural identities have been dominated, undervalued, and/or neglected. The environmental justice movement in particular confronts the connection between a lack of social recognition and environmental risks threatening certain cultural and political-economic groups. In other words, environmental justice is ultimately aimed at enhancing the respect for the place-based identity and survival of communities through institutional changes.

In reference to the important work of Amartya Sen, among others, Schlosberg introduces the capability approach as another useful tool for understanding environmental (in)justice. An analysis of the notion of justice focused on an individual’s and/or community’s capacity to function and to do the things they choose to value is specified here. In *The Idea of Justice*, published in 2009, Sen provides the following explanation:

> The capability approach focuses on human life, and not just on some detached objects of convenience, such as incomes or commodities that a person may possess, which are often taken, especially in economic analysis, to be the main criteria of human success. Indeed, it proposes a serious departure from concentrating on the *means* of living to the actual *opportunities* of living. (2009, 233)

Besides distribution and recognition, this approach examines the process by which goods are transformed into flourishing and well-functioning individuals and communities (Schlosberg 2009, 4).

The examination of what people or communities are capable of doing, whether or not they decide to make use of the opportunities, echoes Bunyan Bryant’s (1995, 6) definition of environmental justice cited earlier: “Environmental justice is saved when people can realize their highest potential, without experiencing the ‘isms.’” Indeed, the ideologies and spatial processes of racism, classism, colonialism, sexism, and heterosexism have historically inhibited the ability of socially vulnerable peoples and communities to function and survive. Because environmental injustice arises when the ability to live safely at home is impeded, the processes and practices that create the underlying ideologies of institutional and structural oppression must be clarified.
This discussion leads to yet another essential perspective for the development of critical environmental (in)justice research, namely its engagement with critical race theory. As Pulido has argued, race is not isolated in a monolithic category; rather, it takes multiple forms in constructing unjust landscapes. Explicating the intersection of capitalism and racism as sociospatial relations and processes is vital for the development of critical environmental (in)justice studies. Moreover, a number of neoliberal states are inherently racial, and therefore play prominent roles in constructing as well as controlling the racial categories in the context of capitalist political economies. In her comprehensive review of the existing literature, Hilda Kurtz (2010, 95) asserts: “Without a more sophisticated consideration of the racialized nature of the state, our understandings of racially oriented EJ activism – its successes, failures and remaining possibilities with regard to the state – will remain woefully incomplete.” Critical race theory will thus “deepen our understanding of the interplay between state apparatuses and institutions and social processes and social movements that delimits the possibilities for achieving EJ” (Kurtz 2010, 112).

Gender and sexuality present another set of vital perspectives that meaningfully contribute to the scholarly and activist conversations on environmental (in)justice. Since environmental (in)justice addresses a diverse range of health and safety issues at home, and since women have been tasked with maintaining this domain in many parts of the world, they have taken crucial roles in activism. Women of color and working-class women, in particular, have taken action to protect their families and communities (Stein 2004, 2). Having firsthand experience of patriarchal, colonial, and racial domination over their bodies, minds, and territorialities, women of color easily recognize the intersection of environmental destruction and issues of family health and community survival.

In addition, Stein (2004, 7) argues that reframing sexuality issues as environmental justice issues enables us to see that “people of differing sexualities have the human right to bodily sovereignty and the right to live safely as sexual bodies within our social and physical environments” and, even more importantly, to advocate broader alliance with related movements led by men and women with various sexual identities to initiate resistance against the oppression over their bodies, peoples, and lands. Building on the theoretical frameworks elucidated by Stein’s edited volume, Montimer-Sandilands and Erickson (2010, 39) elaborate on an innovative term, “queer ecology,” which “embraces deviation and strangeness as a necessary part of biophilia, sexual pleasure and transgression as foundational to environmental ethics and politics, and resistance to heteronormativity as part and parcel of ecological science and green strategy alike.” The conceptualization of nature from a viewpoint of queer sexualities challenges and liberates heteronormative bodies and spaces, radically reconstructing the ideas and politics of environmental (in)justice.

Looking at bodies as procedural and transforming sites of domination and dissent from a feminist perspective, geographer Giovanna Di Chiuro (2008) explores the intersections of environmental and reproductive justice. Without their informed consent, the bodies of impoverished women of color have too often accumulated toxins and become agents of reproductive risk for future generations. By interpreting the notion of social reproduction as “the intersecting complex of political-economic, socio-cultural, and material-environmental processes required to maintain everyday life and to sustain human cultures and communities on a daily basis and intergenerationally,” Di Chiuro...
ENVIRONMENTAL (IN)JUSTICE

(2008, 281) presents a compelling argument that all environmental issues, both local and global, are after all reproductive issues, and, moreover, that environmental justice struggles ultimately seek to achieve healthy social reproduction at and across different geographical scales.

The pursuit of reproductive justice is directly related to what Julian Agyman and his colleagues have clarified as “just sustainabilities” in their attempt to make a connection between the discourses and movements of sustainable development and those of environmental justice. Defined by Agyman, Bullard, and Evans (2003, 5) as “the need to ensure a better quality of life for all, now, and into the future, in a just and equitable manner, whilst living within the limits of supporting ecosystems,” the concept of just sustainabilities looks broadly at our responsibilities in intergenerational and ecological spaces. A framework of just sustainabilities has thus added another important perspective to broaden the sphere of environmental (in)justice.

Furthermore, the growing field of critical environmental (in)justice studies has also considered international and transnational perspectives. Geographies of environmental (in)justice cover a wide range of spatial scales, from individual bodies, which could be those belonging to humans, fish, or even grains of rice, to transnational and global ecological systems. Indeed, environmental (in)justice issues, activism, and scholarship have not been limited to those existing within the national boundaries of the United States. Many cases of environmental disaster and resistance occur in and across different locations, and possess a variety of historical and cultural backgrounds, which are embedded in the capitalist political economy and intertwined with unequal and oftentimes colonial power structures and relations.

For example, multinational corporations have initiated and managed labor-exploitative and ecologically devastating natural resource development projects in cooperation with international organizations and local governments that could be labeled as oppressive. A number of toxic industries originating in North America, Europe, and Japan have relocated their production sites to other locations where they can secure cheap labor and where environmental regulations tend to be less stringent. The crisis of climate change is most apparent in the Arctic marine ecosystem, and currently endangers the livelihoods of the ecosystem’s residents, both human and nonhuman. In reality, however, the entire planet is affected, and the effects of climate change will only grow more severe in the future. The geographical inequity of who generates the causes for climate change and who is or will be affected has historically been related to the global political economy. In consequence, environmental (in)justice issues are broadening and growing more intricate in terms of social and geographical processes and international relations. The complexities of spatial politics pose difficult but important questions to be addressed by critical environmental (in)justice scholars. As Schroeder et al. (2008, 548) write:

Ever-increasing capital mobility, adjustments in trade governance, agrarian restructuring, neocolonial conservation projects, expanding suburban frontiers, and displacements induced by war and violence have also contributed to a sense that environmental justice concerns are increasing in scope. In effect, environmental justice scholars have been confronted with circumstances that require them to reconsider their core analytical constructs – across scales, within spaces defined by new concepts of territoriability, and in the context of complex and shifting network of global interconnections.

As these geographers specializing in political ecology assert, “[T]he core issues at the heart of environmental justice struggles are universal”
(Schroeder et al. 2008, 554). Needless to say, they are also transnational, crossing national borders. The Minamata disease story described earlier, which impacted both fishing villages in Japan and the First Nations in Canada, attests to these universal and transnational dynamics quite eloquently.

Finally, but most importantly, environmental justice movements and studies have searched for explicit and integral connections to ecological justice. Humans and their everyday lives are part of broader ecological systems at both the local and the global levels. Despite the fact that, with its emphasis on the security of daily livelihoods, the concept of environmental (in)justice possesses an anthropocentric implication, human lives are inherently ecological and are intertwined in reciprocal relationships with their surrounding landscapes and various other forms of life. In reference to indigenous environmental activism as well as food and climate justice movements, Schlosberg (2009, 6) argues that “we can draw parallels between the application of notions of justice as distribution, recognition, capability, and participation in both the human and nonhuman realms.” There is no clearly definitive boundary between human and nonhuman worlds, and, moreover, every living thing, be it a plant or a dog, has its own spirit and agency. Acknowledging the integral link between ecological and environmental (in)justice at the sites of reproducing risks seems to be important for us.

Future prospects for critical environmental (in)justice scholarship

Concepts of environmental (in)justice have covered a remarkably diverse range of social and ecological issues relating to struggles for place-based self-determination and livelihoods, all of which are inherently geographical. There are disputes over the extent of the movement itself, as some are concerned that proponents of environmental (in)justice have broadened their scope too far, and therefore fail to embrace its plurality. Regarding the possible diffusion of the movement, Pellow and Brulle (2005, 16) remark that “the difficulty in drawing boundaries for the EJ movement may stem from its multi-issue focus, its multi-ethnic and multi-racial composition, its multi-national scope, and its origins in multiple related movements. EJ activists must bound and limit the purview of their concerns.” If they do not set some limits on their agendas, these authors warn that the movement “will lose its explanatory (and mobilizing) power.” They propose a move toward critical environmental (in)justice studies that emphasizes evaluating the effectiveness and achievement of the movement in terms of its original goals.

According to Pellow and Brulle, critical environmental (in)justice studies should be focused on linking theory to practice in order to accomplish a more effective movement to create just and sustainable societies (2005, 17–18). Their objection to the further expansion of the environmental (in)justice framework is based on their frustration with the painfully slow implementation of environmental justice policy and the persistent spatial injustice that lingers in impoverished communities of color across the globe. And indeed, their claims make sense when we try to prioritize and set practical and realistic agendas.

Nonetheless, this approach may encumber the fundamental core of environmental justice ideas, whose authors, cultivated in multiple and complex geographies, explore societal changes with great creativity. Refuting Pellow and Brulle’s approach, which was derived from pragmatic reasoning, Schlosberg (2009, 8) presents the following compelling argument:
The proposition here is that a more thorough definition of justice – one that encompasses the expressed concerns of environmental justice groups, the conception of justice to the non-human world, and the recent contributions of justice theory – can offer a broadly accessible, plural and workable frame. I am not arguing for a single, all-inclusive, holistic theory of environmental and ecological justice; rather, the point is to expand the discourse of justice, and legitimize the use of a variety of tools and notions as they apply to various cases.

Understanding the multiplicity and complexity of environmental (in)justice studies and activism would further help in developing our ability to reconstruct just and healthy spaces in which all the living things, including humans with various political-economic, racial, national, gender, and sexual identities, as well as animals, plants, and other forms of life, are capable of living happily, and with the hope of a better future.

SEE ALSO: Environment, democracy, and public participation; Environment and everyday life; Environment and gender; Environment, nation, and “race”; Environment and waste; Environmental racism; Political ecology

References


Environmental (in)security

Simon Dalby
Wilfrid Laurier University, Canada

Security and insecurity

Humans face many environmental threats to their wellbeing and, in some cases, directly to their survival. These come both from such things as droughts which disrupt the regular agricultural activities of societies and impact food supplies, as well as more immediate hazards from storms or earthquakes which directly imperil lives. Insecurity is a matter both of environmental context and of the social organizations that facilitate certain modes of human life in those particular circumstances. As humanity’s agricultural and subsequently industrial capabilities have expanded, vulnerabilities to natural environments have become increasingly mediated by economic and political factors. Food supplies remain a major concern, as they have been through most of the modern period, but now there are repeated discussions of shortages of many other resources such as recent scares about petroleum and rare Earth metals. While huge inequalities in the human condition persist in what has become a global economy, universal norms that encourage aid to the most vulnerable in conditions of natural hazard or famine have emerged too.

To speak of such things specifically in the language of security, as is frequently done in contemporary policy debates, complicates matters because of the multiple meanings of the term “security” and its interconnection with the provision of particular forms of modern social order. Security is a key term in the contemporary political lexicon. It is politically effective because it simultaneously invokes protection from danger and the perpetuation of existing social arrangements that can supposedly provide safety. In terms of national security, it refers to the protection of state institutions, traditionally from external military threats and internal political subversion. Global security loosely means preventing major wars, with their potential, now, for large-scale destruction caused by nuclear weapons, as well as maintaining the conditions for the perpetuation of the international state system. This is now dependent on a functioning global economy and, as at least some policymakers are slowly realizing, a relatively stable climate system, a functional stratospheric ozone layer, and oceans that are not too acidic for complex marine ecosystems.

There have been many prior discussions of the human predicament, and concerns about both shortages of various resources as threats to social order and the dangers of rapid changes to natural systems caused by humans. Concerns about scarcities of one form or another as a source of social conflict go back a long way in Western thought, although Thomas Malthus’s (1970/1798) famous “Essay on the Principle of Population,” first published at the end of the eighteenth century, is usually cited as the key work that clarified the theme. European imperial policymakers frequently worried about resource shortages, and geopolitical conflict has long been tied to matters of access to key supplies for military and economic activities (Le Billon 2012). In the 1960s environmentalists worried about overcrowding and food
supplies as a source of conflict, and frequently (if vaguely) suggested that scarcities were a cause of twentieth-century wars (Robertson 2012). Richard Falk’s (1971) overview of these matters, and his explicit attempts to link environmental change to geopolitical problems of world order, which Jon Barnett (2001) usefully suggests is a key precursor to the contemporary discussion of environmental security, didn’t actually make an explicit connection to formulations of security.

Environmental security

Discussing these things specifically in terms of security is a relatively recent development, one that has generated a huge policy literature that is only sometimes directly connected to academic research. As the Cold War ended in the late 1980s, growing alarm about environmental change and potential conflicts that might arise over resource shortages generated a complicated policy and academic discussion about what became known as environmental security (Renner 1989). While such themes have been part of discussions of economics, geography, and other disciplines for much longer, the heightened attention to environmental concerns in the aftermath of the Bhopal chemical disaster in India in 1984 and the Chernobyl reactor meltdown in 1986, attempts to deal with stratospheric ozone depletion as well as tropical deforestation, in particular in Brazil in the late 1980s, when linked to matters of security, meant that they became high profile. The World Commission on Environment and Development’s report in 1987 on *Our Common Future* warned of conflict over scarce environmental resources, which briefly related these themes to security, led the way to the major Earth Summit meetings at the United Nations Conference on Environment and Development in 1992.

While alarms over potential conflict were good at getting attention, there was little evidence of causal links between the environment and warfare or large-scale violence. That people were insecure in many ways was clear, but quite how this was to be understood as a matter of either policy or academic research was anything but simple. The discussion about these relationships continues apace with policy and academic studies intertwined in ways that are frequently confusing (Floyd and Matthew 2013), not least because of the contradictory claims made by academic studies and the misleading headlines that appear when studies become a matter of media attention. Most recently all this has become a high-profile discussion once again as fears of climate change have reinvigorated policy discussion, academic investigation, and media coverage.

The early discussion

Arguments about resources shortages and the potential for conflicts over water suggested that, as Cold War confrontation was reduced, new sources of conflict would arise. Hence, as *Our Common Future* had suggested, sustainable development was needed, not just because of the obligation to help poorer societies become more affluent states, but because the failure to anticipate and deal with deleterious environmental changes might have more serious consequences. Implicit in the discussion was the assumption that scarcity might lead to violence, although quite how was frequently not specified. As with earlier environmental advocacy in the 1960s, when population concerns were rather incoherently linked to crowding as a source of social friction, the relationships between resource shortages and conflict was assumed to be obvious. If this were so, the logic frequently went, then it must be
a matter of national security which is at the heart of modern state power (Mathews 1989). But quite how scarcities would lead to conflict, or what the standard institutions of national security, the intelligence agencies, police, and military might do about environmental issues frequently wasn’t clear.

Critics of the argument about the utility of national security means as a way of dealing with environmental difficulties made a number of key points. First, Daniel Deudney (1990), in a widely cited critique, made the simple but important point that military institutions are often serious polluters and, even in training, are very destructive of the environment. Explosives and heavy equipment are directly damaging to environments, while the chemicals used in military operations are frequently toxic. Huge clean-up operations were necessary in the 1990s to deal with the toxic residues from many Cold War vintage bases. Matthias Finger (1991) bluntly argued that the industrial infrastructure that generated the war machines in the first place was a key part of what caused environmental damage.

Nor was it clear in the early discussion exactly how resource scarcities would turn into military conflicts, especially between states. While wars in which oil was a major factor were obviously possible, as the 1990–1991 Gulf War apparently showed, they were not about spreading deserts, disappearing forests, or water supplies. How societies suffering acute shortages might mount military campaigns to rectify the situation wasn’t clear even if small-scale violence over access to such things as grazing land was plausible. Prins and Stamp (1991) posed the problem of pollution as an issue in contrast to high-technology warfare in the pithy title of a television documentary, *Top Guns and Toxic Whales*. The top gun fighter pilots were of little use in either preventing whales from absorbing toxic effluents from sea water or disposing of the hazardous corpses of dead ones that washed up on beaches.

**Academic skepticism**

While the policy discussion continued, academics posed questions about the causal links between environmental changes and conflict. Given the all-encompassing and hard to define nature of the term, Thomas Homer-Dixon (1999), in particular, focused on acute conflict rather than security. Looking at cases in various parts of the world, a series of research initiatives under his direction in the 1990s tried to track the causal pathways from environmental change to direct violence. These studies suggested that, while such connections could be found, the social context and actions (or failure to act) of political elites were an important part of the picture, not least because sometimes the actions of elites in a crisis made access to resources more difficult. Resource capture by elites and the resultant ecological marginalization of vulnerable peoples was often part of the problem. Assumptions that simple relations between scarcity and violence were commonplace were not supported by his empirical investigations. Guenter Baechler (1998) led a broadly similar set of investigations in Europe in the 1990s which concluded that environmental issues were related to violence in circumstances where they were superimposed on discrimination and restrictions of access to resources that followed other social cleavages. Baechler’s research emphasized that environmental matters were sometimes related to violence in situations of what he called mal-development, and that this was likely to take place in marginal lands including mountainous regions where development was causing disruptions to traditional ways of life.
ENVIRONMENTAL (IN)SECURITY

Despite the caveats and cautions in the scholarly work, Robert Kaplan used Homer-Dixon’s early papers to write an essay for the *Atlantic Monthly* in February 1994 that ominously announced “The Coming Anarchy.” In this dystopic account of current conflicts, Kaplan melded discussions of the clash of civilizations with ideas about militias, civil wars, and the changing nature of warfare and concerns about resource shortages, environmental change, and violence into an alarming vision of the near future. It was widely circulated and discussed at the time, and many thought it provided an explanation of the genocide in Rwanda that took place a few months after its publication. This single article galvanized attention to the discussion of environmental security in Washington in particular and influenced policy debates in the Clinton administration. NATO undertook a series of studies into various relationships between the environment and security problems and developed a series of analyses that suggested that environmental matters were related to politics in numerous different patterns, only a few of which syndromes fitted the classic discussions of environmental change causing violence fairly directly (Lietzmann and Vest 1999). Many of the others involved disputes over development projects, pollution, and campaigns to reduce the damage caused by industrial expansion.

Methodological debates

While the discussion of environmental security and its policy ramifications was underway, a number of methodological debates took place among scholars trying to tease out how these relationships might be understood and researched. Three are noteworthy, and all have a bearing on how the subsequent more recent discussion of climate change is playing out.

First is the question of what counts as evidence in the discussion of causal relations between environmental change and violence. While both Baechler and Homer-Dixon looked to field investigations to find cases where these links could be demonstrated, others suggested that large-scale statistical investigations were more appropriate methods for investigating the relationships. If there is a correlation between episodes of violence and documented environmental change, then, so the argument goes, a scientific case can be made for the relationship. Individual field investigations, focusing on particular situations rather than general patterns, cannot effectively demonstrate general causal patterns. Second, without detailed fieldwork, so the counterargument goes, correlations between the two might be entirely spurious because of other factors that are not clear from statistics; only practical investigations on the ground can show what causal mechanisms are in play.

Third, this controversy was aggravated by some geographers’ and anthropologists’ concern that the case study research ignored previous careful ethnographic research into the changing relationships of people to environments in particular places, research that suggested that the relationship of political violence related to environmental change was much more complicated than simple measures of resource scarcity (Peluso and Watts 2001). These critics pointed out that this political ecology literature, and what it had to say about the disruptions caused by processes of development in particular, had been given short shrift by political scientists. Globalization and the rapid expansion of resource extraction, the spread of corporate and commercial agriculture, and related development projects such as dam and road building are, in this scholarship, much more important than the scarcity variables considered in the political science literature.
An additional important argument emerged in the late 1990s from development studies suggesting that violence was more obviously related to resource abundance than to scarcities and shortages (Bannon and Collier 2003). The discussion of the resource curse and related phenomena suggested that, in places where poverty was widespread and few economic development options were available, it might be more sensible on the part of those looking for wealth to use violence to seize control of mines and forestry resources and control the rent streams instead of following the long slow road to economic development. Related to this, the resource curse is also connected to the phenomenon of inflated currency prices due to the effect of commodity exports, the so-called Dutch disease, which also complicates economic development strategies. Conflict resulting from attempts to gain instant wealth from selling resources was, so the argument went, an alternative interpretation to those suggesting that rural resource shortages lead to violence (Le Billon 2012). As with the political ecology studies, the resource abundance argument suggested that interconnections between the global economy and particular places were an important part of the puzzle of how resources were linked to violence.

**Human security**

In addition to the policy debates about environmental scarcities and the academic arguments about the appropriate mode of investigation of these issues, the United Nations’ agenda in the 1990s included a much broader consideration of social matters and human welfare than the Cold War period had permitted. Matters of environment, gender, development, and related matters were the focus of a series of large conferences. This led to various attempts to develop new policy agendas to deal with societal matters and extend human rights and development not just to states and their elites but to peoples, and eventually to a much larger academic discussion of how to rethink these issues in light of environmental change and globalization (Brauch et al. 2008).

Among these initiatives was the high-profile formulation of the idea of human security in the 1994 United Nations Development Report. Instead of focusing on security in the narrow sense of military danger and direct threats to states and social order, the authors of the report suggested a much broader understanding of security related to the wellbeing of people. Focusing on the conditions necessary for development to take place, the report also suggested that insecurity for many people was not the result of direct political threats from other states but of unanticipated and unintended consequences of the huge transformations underway globally as economic growth, urbanization, and globalization caused rapid change across boundaries. Human security was formulated in anticipation of such consequences, with measures being taken to protect people from hurtful disruptions; unlike traditional military security, human security was about preparation rather than violent response after the fact.

While the human security formulation dealt with political, economic, and other changes, it also included environment as a key factor that makes people feel insecure. What it did not offer was an explanation of how this came to be, beyond some obvious concerns about ozone depletion and pollution in particular places. Nonetheless, as a matter that concerned people’s safety, the environment is clearly a factor to be considered in discussion of the necessary prerequisites for development and the expansion of freedom for people, especially in the poorer parts of the planet (O’Brien, Wolf, and Sygna...
Subsequently, some of these discussions have been extended in the debate about the now widely adopted principles of responsibility to protect as governing state legitimacy. While environmental degradation is not listed as one of the key factors that trigger international action to protect people where their state has manifestly failed to do so, there have been discussions as to how these principles may be extended to help in emergency situations caused by environmental factors such as floods, storms, or droughts.

**Climate change**

While matters of environmental security were a low priority in many parts of the world in the early stages of the American “war on terror” in response to the attacks of 9/11, they have since found their way back into discussions of security. While there was great reluctance on the part of the Bush administration to deal with matters of climate change, by 2007 a number of high-profile reports were appearing, which raised the alarm about traditional national security implications of a rapidly changing climate (CNA Corporation 2007; Campbell et al. 2007). This has a number of aspects, not least the very direct one that the American military is concerned that its facilities, and its ability to operate them, might be vulnerable to climate change. Low-lying runways and port facilities are obviously in danger of rising sea levels and hurricanes might directly impact bases in many ways. The military is also usually the first agency called upon to aid in disasters, not least because of their ability to move emergency supplies and operate outside normal civilian supply systems. For all these reasons, many militaries are starting to take climate change seriously as a factor that may affect their operational capabilities and as a threat that they cannot directly control.

One formulation that has appeared repeatedly portrays climate change as a threat or conflict multiplier. The suggestion is that either climate change in general, or immediate disruptions caused by storms, can aggravate social tensions and in turn provide a pretext for insurgency, civil unrest, or terrorist action. If states are unable to cope with climate change, then, according to this argument, political disruptions and violence are likely to be more frequent, some of which may involve military interventions, and all of which will make global security more tenuous. Hence a sensible security policy requires dealing with these situations before they can cause destabilization. The worst case scenario of state failure may require military action to contain terrorist, criminal, or piracy threats that might result.

This discussion has revived the earlier scholarly debate about the relationships between environmental change and insecurity, albeit with a more direct and narrower focus on climate and its specific consequences. But, insofar as climate change affects basic environmental factors such as rainfall, water availability, droughts, and storm damage, it involves a variety of matters similar to those in the earlier discussion. Once again scholars have looked to the empirical record in terms of climate change and social consequences; a formal German advisory council itemized many of the studies in terms of climate change as a security risk (German Advisory Council on Global Change 2008).

At the macroscale the discussion about climate engages with themes of historical crisis and the collapse of civilizations that were unable to cope with resource shortages or changing weather patterns for agriculture (Diamond 2005). The archaeological record is not clear in many cases and, even if climate changes have been exacerbating factors, in the absence of clear historical documentation, how social changes happened,
and where and how these were unsuccessful, simply are not known in most cases.

Historical research into more recent events is beginning to trace connections between many episodes in recorded history and their climatological circumstances. The seventeenth century, in a period sometimes known as the Little Ice Age, is particularly pertinent because there are numerous clearly documented situations from across the globe that allow connections between disruptions in food supplies, in particular, and subsequent political events, including widespread warfare, to be traced. While climate wasn’t the only factor in play in this tumultuous century, political elites made decisions in the context of climate variability which drastically upset traditional agricultural calculations. What is clear, as Geoffrey Parker’s (2013) synthetic study suggests, is that attempts to deal with such problems by going to war were usually counterproductive.

The direct relevance of historical studies to present circumstances, with its very different global economy, a gradually emerging global civil society, and an international food system, is questionable but highly suggestive nonetheless.

Recent social science work on climate change and potential political difficulties has in some ways replicated the discussions of the 1990s. Some researchers have once again looked at large-scale databases on conflict and tried to link these with weather conditions to draw statistical connections between the two. While these are contentious, and some media headlines have drawn alarmist conclusions from the published studies, data from the last few decades has provided few clear indications that climate change has caused conflict. While some statistical correlations may link meteorological factors with civil unrest, many of these studies use national statistics, which are at a level of generalization that doesn’t effectively explain specific political developments in particular places. Attempts to disaggregate data to focus on smaller areas have not definitively proved causal connections between climate and conflict.

Other studies which have investigated particular cases where environmental change might be related to conflict or violence have concentrated on development and, in work loosely consistent with Gunther Baechler’s 1990s studies, suggested that economic structures are more important than environmental matters in explaining how environmental circumstances might be related to conflict. But these too have made it clear that the relationship is complicated and that contextual factors have to be considered in any attempt to connect the two; conflicts over water happen in the context of complicated institutional arrangements and in conditions of structural violence which have to be understood for the dynamics of conflict to be grasped and policies put in place to deal with these difficulties. Clearly, media preoccupations with “water wars” notwithstanding, interstate wars caused directly by water issues simply haven’t happened. Structural inequalities make many people insecure, and access to water is a very important matter for human security, but it does not lead to security problems in terms of overt warfare (Zografos, Goulden, and Kallis 2014).

In the case of the Arab Spring food, prices were in some senses a precipitating factor in the political upheaval, and the rise in food prices was related to climate factors in the global agricultural system. In the civil war in Syria journalists have repeatedly suggested that the drought in the years prior to the outbreak of violence was a significant factor in setting the stage for the subsequent war. While climate change might have been a contributing factor to the drought, which caused people to migrate to the cities, it is clearly not a direct cause of the civil war. Political institutions and their ability or inability to handle rapid environmental and social change are the key
to whether conflict results. Larger geopolitical calculations and support for factions in struggles are also a key part of most conflicts; environmental scarcities are neither the sole cause of conflict nor the cause of other forms of insecurity.

**Earth systems science**

As global change accelerates, the questions facing human security increasingly relate to the artificial circumstances that humans find themselves confronting in what is now frequently called the Anthropocene. The majority of people now live in urban areas in dependence on the surrounding rural areas, and sometimes very distant locations, for supplies of food as well as other necessities of life. Globalization means not only the interconnection of economies and production systems, but also the transformation of environments directly to supply materials and absorb wastes. It also works indirectly through climate change which is raising sea levels and upsetting assumptions about food production in numerous places. Insecurities are, hence, increasingly artificial and, as such, encompass the social provision of the necessities of life (Dalby 2009). While great care has to be taken over extrapolation from the history of elites who failed to take seriously their responsibility to people in their charge and ended up with collapsed societies, this theme needs serious research attention in our interconnected world marked by great inequalities.

This is especially important now as Earth system scientists trace the scale of the transformations that humans have set into motion and emphasize that current changes to the earth system are faster than anything in the geological record. Insecurity is now of our collective making. Human insecurity is now connected to key decisions about the future of energy systems that power the global economy. The environment is not an independent variable that causes human insecurity, but is now partly a human artifact, and as such a matter for political deliberation and decision (O’Brien, Wolf, and Sygna 2013). So far political attention has focused on matters other than the future configuration of the climate system, but this may change in the coming decades as changes multiply and climate adaptation requires more attention. Given that climate change is being caused by human actions, it is no longer appropriate to think of the environment as a residual category that is separate from humanity. Climate change now makes it clear at the macroscale, as it has long been evident in political ecology research at the microscale, that humans are transforming environments and in the process changing who is insecure where.

What research has so far shown is that the failure to think carefully about our ecological context may lead to unwise or counterproductive attempts to mitigate the effects of or to adapt to climate change. The purchase of offsets for the use of carbon fuels have sometimes aggravated land-use conflicts in the Global South, whereby forestry plantations certified as carbon sinks require numerous ecological changes that undercut their ostensible purpose. Land grabs are part of national and corporate strategies to secure future supplies of food, but if they displace local inhabitants or cause disruptions in land markets they may cause conflict in the process. Rather than environmental change causing conflict, it may be that attempts to cope with that change will stimulate upheavals and violence (Dalby 2013). The failure to understand the complex interconnected natural and artificial systems that are now humanity’s context will result in simplistic, counterproductive security policies. If the perils of such misguided policies are to be avoided, considering how economies are remaking environments should form a key
part of current research agendas (Brauch et al. 2008). All of this requires careful analysis of how particular places are connected to the rapidly changing larger ecological contexts of our times.

SEE ALSO: Anthropocene and planetary boundaries; Climate change adaptation and social transformation; Development; Earth system science; Environmental futures; Geopolitics of the environment; Globalization; Natural resources and human conflict; Political ecology; Security; State, the; War; Water conflicts; Water security

References

ENVIRONMENTAL (IN)SECURITY


Environmental assessment techniques

Luuk Fleskens
University of Leeds, UK
Wageningen University, Netherlands

Environmental assessment techniques comprise a range of methods and tools, either applied in isolation or in combination, to qualify or quantify environmental changes and/or the environmental impacts of human interference. Environmental assessment is a systematic process for evaluating and documenting information about natural systems and resources that may be used in planning for sustainable development and environmental decision-making by different actors, for example, governments, companies, and so on (Sadler 1996). Whereas environmental assessment techniques can be deployed autonomously, that is, as a stand-alone tool, they are often used institutionalized within legal frameworks for environmental impact assessment or environmental accounting procedures. When institutionalized, certain standards need to be adopted, a combination of assessment techniques may be required, certain procedures may need to be complied with, and a certain sequence may be prescribed.

There is a distinction between “environmental assessment,” which focuses on the effects and consequences of specific human undertakings, and “environmental monitoring,” which takes the wellbeing of the whole system (ecosystem, land-use system, etc.) as its starting point (Kessler et al. 2002). The wider system provides the context within which decision-making processes operate and is normally more difficult to assess as sources of information are outside the direct sphere of influence of the decision-maker. For example, monitoring the quality of a watershed for water resources development requires information on a range of parameters, many of which (e.g., meteorology, land use, soil types) are not directly collected by a water resources authority. Notwithstanding, monitoring this context is important as it is here where impacts from the decision manifest, and also because changes in the context may interfere with the possibilities to reach objectives.

Reconciling performance monitoring and measuring of environmental change requires selecting appropriate environmental indicators (Kessler et al. 2002) and environmental assessment techniques can essentially start this reconciliation process from either end. This means that environmental assessment can include environmental monitoring although this may not always be feasible or appropriate. Conversely, environmental monitoring could point to effects of management interventions although it is often difficult to attribute change to the effectiveness of measures.

The deployment of environmental assessment techniques should be carefully organized within analytical frameworks. The ecosystem functions and services framework (De Groot et al. 2010) offers such a comprehensive framework that explains how changes in environmental systems affect human wellbeing. This framework has been widely adopted for environmental assessment, including the Millennium Ecosystem Assessment (MA 2005). The focus in the
ENVIRONMENTAL ASSESSMENT TECHNIQUES

ecosystem services assessment framework is on increasing understanding of the linkages between nature and society and raising awareness of the value of the natural environment for society.

The DPSIR (drivers–pressures–states–impacts–responses) framework (European Environmental Agency (EEA) 1999) provides an alternative structure to assess how human undertakings affect the environment and is also frequently used to structure environmental assessments, for example, in the EEA’s State of the Environment Reports. The DPSIR framework allows indicators to be presented in a way to enable feedback to policymakers on environmental quality and the resulting impact of the political choices made or to be made in the future.

Historical evolution

Environmental assessment techniques may go back a long way, for example, meteorological records are available since the 1730s, but their integrated use for environmental assessment is more recent. One of the first methods for environmental assessment was environmental impact assessment (EIA) which was developed in the 1960s (e.g., Leopold et al. 1971). EIA at that time was carried out using a checklist matrix to identify the importance and magnitude of impacts of future projects and plans. Such matrices replaced earlier ad hoc tools that were criticized for being subjective and descriptive and not accounting for spatial and cumulative effects of identified impacts (Munier 2004). Checklist matrices, however, lacked the necessary guidelines for estimating the degree of impacts, which affected the transparency of the EIA process. In response, more elaborate techniques such as the semiquantitative rapid impact assessment matrix technique (Pastakia and Jensen 1998) and environmental risk assessment were developed. Such techniques deepened and broadened the assessment in order to get more comprehensive data for project planning while at the same time being able to demonstrate better the evidence on which the assessment was based to multiple stakeholders.

With increasing attention to long-term trends of global change effects and evolution of computers, model-based scenario analysis was added to the toolbox. New methodologies such as strategic environmental assessment (SEA) emerged that complemented EIA by focusing on the environmental impacts of policies and strategies for which EIA was ill-equipped (Therivel 2004). It was also recognized that the portfolio of assessment techniques required to plan for environmental change needed to be augmented to respond to different information needs at different levels of decision-making and involve the potentially affected stakeholders. This led to the introduction of multistakeholder multiscale assessment techniques (e.g., Reed et al. 2011; see also Figure 1). More and more, use is being made of remote sensing and (web-based) participatory monitoring techniques. In several fields, the integration of several types of information gathered at multiple levels by different actors is currently being advocated to advance environmental assessment. For example, policymakers at national level require information on the future likelihood of extreme events affecting the country in order to develop appropriate policies. District-level authorities require assessments of which adaptation options will be required where. Farmers need seasonal forecasts in order to anticipate adaptation choices. Citizens need early warning signals to respond to threats to their safety not attenuated by the adaptation measures implemented.

For measuring the progress toward long-term environmental goals, systematic accounting techniques have also developed over time. Natural resource accounting has gained prominence
since the early 1980s and evolved into the first system of environmental-economic accounting (SEEA) in 1993. SEEA allows integration of the environment into national accounts, and enables environmentally adjusting macroeconomic indicators such as GDP. Mainstreaming of the ecosystem services concept (MA 2005) gave a further impetus to development of sophisticated environmental accounting techniques for tracking the contribution of the environment to economic activities. Roughly at the same time, Rockström et al.’s (2009) concept of planetary boundaries reiterated the need to closely monitor environmental change to avoid crossing resilience thresholds with potentially severe implications for humanity.

**Techniques**

Environmental assessment techniques constitute a wide range of tools to assess or monitor a variety of environmental conditions and indicators. Among the most prominent of these techniques are expert judgment, checklist matrices, environmental risk assessment, mapping and spatial analysis, impact prediction, impact evaluation, remote sensing, natural resource accounting, participatory techniques, and crowdsourcing.
ENVIRONMENTAL ASSESSMENT TECHNIQUES

Expert judgment

Expert judgment is one of the most frequently used techniques, especially in the context of EIA/SEA applications. This recruits a team of experts covering multiple specialisms and asks them to rank alternative projects or plans, and/or suggest mitigation measures. Procedures include (group) discussions, sometimes in multiple rounds, to generate data, judgments, or suggestions. The main advantages of using expert judgment is that it is quick and cheap, able to consider multiple types of information, can lead to innovative solutions, and contributes to information sharing and education. Its drawbacks are that it is sensitive to biased opinions based on selected experts and is a nonreplicable and unscientific process (Therivel 2004).

Checklist matrices

Checklist matrices are used to qualify the importance and magnitude of environmental impacts expected from future projects and plans and constitute the most commonly used technique applied in EIA. A number of variants have been developed. In the simplest form, checklists matrices present a list of environmental qualities against which to score the importance and magnitude of impacts of future projects and plans. A more sophisticated matrix is the rapid impact assessment matrix (RIAM) proposed by Pastakia and Jensen (1998). RIAM uses a structured matrix to allow for semiquantitative impact judgments, specifying importance of the environmental condition (ranging from no importance to importance to national or international interests) and magnitude of effect (from major disbenefit to major benefit), together with judgments on the permanence, reversibility, and cumulative nature of changes.

Environmental risk assessment

Risk assessment refers to a formal process of evaluating the consequence(s) of a hazard and their likelihoods/probabilities. The process followed resembles an EIA, with components of formulating the problem, carrying out an assessment of the risk, identifying and appraising the management options available, and addressing the risk with the chosen risk management strategy. In the risk assessment phase, however, frequently attempts are made to quantify the consequences and probabilities of risks as qualitative and semiquantitative assessments would remain subjective and inconclusive. Quantitative assessments can, for example, use quantitative exposure assessments, simple deterministic risk estimation, or Monte Carlo simulation techniques.

Mapping and spatial analysis

In environmental assessment, there is often a need to understand spatial patterns of threats, impacts, applicability of mitigation options, and so on. A wide range of practical tools fall into this technique. Overlay maps can be produced for each specific factor of interest and combined (e.g., flood risk and asset value maps) to quickly identify hotspot areas. Geographic information systems (GIS) have generated rapid advancements to mapping and spatial analysis, and are now the tool of choice for analyzing spatial aspects of environmental assessments. Commonly, field data and other assessment data collected outside of the GIS environment are geo-referenced and linked in spatial geodatabases to accommodate mapping and spatial analysis. An example of a spatial analysis in environmental assessment is land-use partitioning analysis which is used to assess the effect of habitat fragmentation resulting from infrastructure projects. It involves the identification, assessment, and recording of landscape fragmentation effects.
Impact prediction

For informed decision-making, environmental impacts frequently need to be anticipated. This can be achieved by impact prediction techniques which compare environmental changes with and without implementing a strategic action. Of interest may be whether a strategic action’s impact is expected to be positive or negative, large or small, short term or long term, reversible or irreversible, and likely or unlikely. While such assessments can be based on expert judgments or checklist matrices, there is often a need for more comprehensive quantitative assessment based on modeling. A huge variety of environmental model types exists, for example, through Bayesian belief networks (BBNs), a probabilistic type of network analysis, it is possible to capture the probabilities about the performance of an outcome indicator (e.g., farm income) based on different state variables (e.g., rainfall and soil quality). BBNs are versatile in that the conditional probability tables used to populate the model can be filled with different types of information (qualitative judgment, measured or modeled data). Life cycle analysis (LCA) is another modeling tool in which environmental impacts of a product are quantified by considering all resources consumed and emissions to the environment during all stages of its life cycle. In the framework of impact prediction, models are generally used for scenario analysis of different strategic options. For example, Basset-Mens and van der Werf (2005) report on a scenario-based assessment of different pig production farming systems in France using LCA.

Impact evaluation

Taking decisions on whether or not to implement strategic actions requires impact evaluation, that is, interpreting the information from impact predictions to know which option is the most preferred. Impact evaluation techniques consider environmental impacts, often along with other (economic, social) criteria, but the way in which this is done varies. In cost–benefit analysis, an attempt is made to value all costs and benefits of implementing a strategic action in monetary terms relative to not implementing the action. An alternative tool is multicriteria analysis, in which strategic options can be compared based on multiple criteria that can remain in different units, for example, considering environmental impact, economic return, employment generation, and health benefits. Impact can also be evaluated based on solely environmental grounds, for example, using tools such as carrying capacity or ecological footprints. Impact evaluations can also be made spatially explicit, for example, Fleskens, Nainggolan, and Stringer (2014) provide a framework for spatially explicit financial cost–benefit analysis of sustainable land management options.

Remote sensing

Given considerable resources required for environmental assessment when based on field data collection, and in response to a need to extend assessments over larger areas and longer time spans, a variety of remote sensing techniques has gained prominence in environmental assessment. Remote sensing applications include satellite-based monitoring of terrestrial and environmental changes (e.g., land cover, vegetation greenness) as well as observation of different characteristics using ground-based methods (e.g., laser to assess micro-topographical changes due to tillage erosion) or, increasingly, drone-based methods (e.g., sensors to detect pollution, video cameras to detect forest fire spread). Satellite-based products such as Landsat and MODIS have already proven invaluable for environmental assessment, whereas development
ENVIRONMENTAL ASSESSMENT TECHNIQUES

of assessment techniques with drones is a hot topic in many environmental science disciplines.

Natural resource accounting

Natural resource accounting is a technique aimed at monitoring stocks of environmental resources and understanding how dependent economic growth is on depletion of such resources. It was introduced in response to demands from policymakers to have systems in place to fulfilling these purposes since the manifestation of environmental crises and in particular the realization of limits to growth in the 1970s. The island state of Nauru is a case in point where natural resource accounting demonstrated high dependency of its GDP on a dwindling natural capital (phosphate). Statistical standards for the assessment of natural resource stocks became common ground with the SEEA environmental accounting system. The existing standards are continuously being improved and extended; for instance, experimental statistics about the flows of ecosystem services are currently being developed and piloted in several countries.

Participatory techniques

Public participation has been advocated since the inception of the EIA process, but is in practice often better qualified as consultation. Arguments in favor of involvement of non-expert participants revolve around normative, substantive, and instrumental rationales (Glucker et al. 2013). Participatory techniques range from focus groups and citizen’s juries to more specific techniques with the purpose of eliciting specific knowledge from stakeholders, including participatory indicator development, participatory scenario development, and participatory technology appraisal (for examples in the context of dryland management, see Reed et al. 2011). One technique which has been used to make the voice of marginalized groups heard (stressing the normative rationale) is participatory mapping, for example, forest resource mapping by indigenous people to contest logging concessions.

Crowdsourcing

In crowdsourcing, sometimes also referred to as citizen science, public participation becomes the driving force of environmental assessment. It refers to (online) collaborative efforts either initiated by scientists or by the public themselves. The aims for crowdsourcing can differ, but it calls upon specific expertise or local knowledge as the only or most cost-effective way to generate or validate large datasets. Increasingly mobile apps are used for this purpose, such as in the EU Horizon project Interactive Soil Quality Assessment for Agricultural Productivity and Environmental Resilience (iSQAPER, www.isqaper-project.eu).

Applications

Environmental assessment techniques are being applied in multiple contexts, such as in EIA/SEA, environmental governance, disaster risk management, adaptation planning, environmental markets and compensation schemes, environmental monitoring, environmental accounting, scientific research, and knowledge management.

EIA/SEA

In the screening and scoping phases of EIA/SEA and other strategic planning decision support tools to assess the environmental impacts of activities, projects, plans, programs, and policies assessments seek to identify the most critical potential environmental effects and to predict
and evaluate which alternatives or options may mitigate such impacts. Environmental assessment in EIA and SEA is characterized by a streamlined application of techniques to fit windows of opportunity for influencing decisions. For example, EIA of large infrastructural development projects, such as dams, are in many countries required by law. The purpose of environmental assessment is to understand how construction of a reservoir might affect water provisioning and regulation downstream in the river basin, under both dry and wet conditions. The environmental assessment should take a broad view and, for example, offset hydropower generation against alternative sources of energy. The assessment could lead to alterations in the design, for example, by identifying a need to prevent sedimentation of the reservoir by afforestation of its catchment area. Ultimately, a profound and well-carried-out assessment could lead to project plans being shelved. Complex large-scale projects may not fail directly on environmental criteria, but can be particularly controversial when it comes to social criteria such as displacement, prospects for development of different social groups, or environmental impacts disproportionately borne by a specific group.

Environmental governance

Environmental assessment techniques directly inform decision-makers about the need to adjust policies or regulations, or enforcement officers about infringements. An example of the first is assessment of water availability for abstraction in catchments in the United Kingdom. Based on climate and hydrological monitoring information, so-called environmental flow indicators are calculated to inform a color-coded system of implications for water licensing. An example of the latter is the Brazilian satellite-based real-time system for detection of deforestation (DETER) which provides fortnightly high-resolution updates of deforestation hotspots, which has greatly facilitated law enforcement officers to act upon illegal logging and has received credit for reducing deforestation rates in the Amazon since the early 2000s.

Disaster risk management

Environmental assessment techniques are increasingly deployed in a range of early warning systems, varying from linked observational–prediction networks for natural hazards (e.g., the Japanese earthquake early warning system which issues warnings along with guidance on how to react in case of observed seismic activity, or more commonly, weather forecasts including severe weather alerts) to multifaceted systems that predict how dynamics in environmental and social systems may culminate in disasters. A much-lauded example of environmental assessment information in predicting disasters is the famine early warning system network (FEWSNET) which combines remotely sensed information on climate (e.g., droughts, vegetation greenness), agriculture production, prices, trade, nutrition, and other factors, together with an understanding of local livelihoods in order to predict areas at risk of food insecurity and famine. Starting in sub-Saharan Africa in the 1980s, the program has recently expanded its monitoring activities to nontraditional regions at risk of food insecurity such as Central America and Central Asia (see Figure 2).

Adaptation planning

Planning for uncertain future conditions such as climate change requires multiple environmental assessment data and a combination of risk assessment, scenario modeling, and mapping techniques to understand patterns of exposure.
and sensitivity in order to increase adaptive capacity. Assessment of exposure requires understanding the severity of threats such as sea level rise, heatwaves, droughts, and floods. At national level, countries are developing national adaptation plans with a view to increase preparedness for future climate change.

**Environmental markets and compensation schemes**

Environmental assessment techniques are used to devise and manage environmental markets where providers and buyers of environmental services need information on the effects of their interactions. Monitoring systems that allow for credible measurement, reporting, and verification (MRV) of activities are essential in underpinning such schemes. For example, new assessment techniques in carbon markets are being developed and need to be approved before they are allowed to enter the marketspace of schemes such as the Clean Development Mechanism or Verified Carbon Standard. Payments for other environmental services, such as watershed services (e.g., Green Water Credits) or enhanced ecological functioning (agri-environmental measures in the EU Common Agricultural Policy) have a similar need for compliance and effectiveness monitoring to prove that they are doing what they claim.

**Environmental monitoring**

Governments have agreed to monitor and report several indicators in the framework of United Nations conventions, either directly linked to commitments or for general monitoring of the state of the environment. Similarly, companies and organizations engage in sustainability
reporting in which the environmental impacts of operations are presented.

Environmental accounting

An arena of application which is rapidly gaining attention is ecosystem accounting, which deals with integrating complex biophysical data, tracking changes in ecosystems, and linking those changes to economic and other human activity (UNSTATS 2014). Starting with techniques such as natural resource accounting and environmentally extended input–output matrices, accounts of environmental flows and changes are being targeted.

Scientific research

For monitoring, environmental change data are often required over large spatial areas and long periods of time. Hence there is a need for long-term multisite data collection using the same or similar environmental assessment techniques. The datasets that are generated by such initiatives are important for scientific research. An example of this is the Long Term Ecological Research Network (LTER, www.lternet.edu) created by the National Science Foundation in 1980. Recognizing that long-term data across sites allows for continental-scale questions to be addressed, the LTER Network makes data collected by all LTER sites broadly accessible to other investigators.

Knowledge management

Increasingly, attention is also paid to the fact that environmental data needs of different stakeholders vary and that knowledge is fragmented across these actors. Knowledge networks need to be created to unlock this knowledge about the environment. Environmental data portals and hubs where users can collaboratively construct, combine, and discuss about data obtained from different assessment techniques can form a useful starting point for this purpose. Specific portals of this kind for environmental assessment include the International Institute of Applied System Analysis-initiated Geo-Wiki site (www.geo/wiki.org) where citizens are invited to contribute to address global land-cover issues, and the World Overview of Conservation Approaches and Technologies (WOCAT, www.wocat.net) dedicated to uniting efforts in knowledge management and decision support for upscaling sustainable land management among all stakeholders.

Challenges and outlook

Environmental assessment is hampered by a number of challenges including lack of data, tempo-spatial variations in environmental qualities, scale effects, and normative aspects. EIA judgments are often claimed to be subjective, based on a limited evidence base, or because they are done too superficially. This could in part be due to a dominance of expert judgment-based or other qualitative methods. However, even quantitative methods require judgments which may be subjective due to spatial variations and uncertainties in risk assessments.

Inherent uncertainties would imply that ideally, environmental assessment should be periodically repeated, at least for complex projects. When new evidence becomes available, this should then feed into the decision-making process to consider readjustments. In fact, SEA offers a framework for such more-embedded assessment procedures. However, this does raise the bar for demands on the systematic recording and reporting of considerations, and also of subjective
judgments, as the people involved in conducting the assessment may change over time.

The role of participation in different assessment procedures and techniques is also a challenging issue that continues to provoke debate. Engaging people in assessments is often easier said than done and the outcomes are not necessarily leading to more sustainable decisions in the long term and may give rise to dominant “not in my backyard” notions. In case of conflicting stakeholder interests, finding middle ground using transparent procedures may be illusive.

Nexus thinking gives rise to increasingly complex issues for which an increasing amount of data is required and, accordingly, an appropriate set of assessment techniques needs to be selected. Good systems knowledge is required to understand the linkages between different environmental processes and welfare outcomes. Roles of different stakeholders may need to be rethought as, for example, nonexperts will probably have an increasing role to play not only in the decision-making, but also in environmental monitoring. Operationalizing indicators for assessing planetary boundaries, and transforming these into national targets, may be one task for environmental assessment techniques and a simultaneous bottom-up process for building and managing knowledge of what approaches and strategies work where and why will be crucial to meet human needs in a changing environment.

SEE ALSO: Ecosystem services; Environmental change and social learning; Environmental impact assessment; Geography and the study of human–environment relations; Global environmental change: human dimensions; Monitoring and evaluation; Qualitative data, acquisition; Spatiotemporal analysis

References


Environmental certification and eco-labeling

Tad Mutersbaugh  
University of Kentucky, USA

Certification networks are fundamentally concerned with the management of extrinsic qualities, that is to say, qualities which are not detectable in the final product but whose existence is important to the buyer. To give an example, organic products provide an exemplar of a certified, extrinsic product: although it may be possible for a laboratory to detect trace quantities of synthetic chemicals banned in organic food production, this cannot tell the consumer whether the organic product was produced in fields where the integrity of soil and water were protected and biodiversity nurtured, whether organic seed or rootstock was utilized, or indeed whether any banned chemicals were used since trace chemicals found during testing might be the result of contamination by overspray from a neighboring “conventional” field. Under these conditions of uncertainty, certification offers a package of field, packing-shed, transport, and storage inspections that trace an organic product from point of origin to consumption, providing an “audit trail” of the “chain of custody” that serves to justify the issuance of an organic label that provides a guarantee of product integrity.

Even a brief inspection of the number and types of “green” environmental and ethical certifications will show that this constitutes a rapidly expanding product category. For instance, the eco-label index (www.ecolabelindex.com) provides a listing of 458 certified products that include consumer-oriented consumption items such as bird-friendly® coffee, wildcrafted and organic products from shrimp and timber, carbon (sequestered), and industry-oriented services such as waste management, building codes, mining protocols, product lifecycle assessment, water management, and ecotourism. However, even this extensive “eco-label” product list excludes environmental products such as GIs (geographical indicators, e.g., wine, cheese, and other terroir products), ISO 14000 environmental management processes, and sanitary (e.g., phytosanitary, bioterror) certifications.

Origins and rationales of certification

Although certification has been important for commodities such as agricultural seed since the early twentieth century, the modern certification framework came into being with the establishment of ISO (the International Organization for Standardization) after World War II. This 1940s institutionalization took on initial tasks of governing standards such as the length of a meter and introduced the contemporary relationship between standards, certifications and accreditation. In these early years standards adoption was for the most negotiated and democratic, and featured limited state involvement, voluntary agreements, and few third-party inspections. Since the early 1990s, however, ISO has grown to include 165 member bodies such as US ANSI (US American National Standards Institute), and now takes a rather more activist approach, enforcing standards compliance via the TBT (technical barrier to trade) agreement, which requires...
World Trade Organization (WTO) signatory nations to comply with global standards.

Within this ISO context, one of the most important standards is the 17065 certification standard “Conformity assessment – Requirements for bodies certifying products, processes and services.” By establishing this standard, ISO has become the de facto governing body for global certification practices, since any organization that sells a globally traded, certified product (such as certified organic coffee) is required to undergo a lengthy (and costly) audit to demonstrate that their inspections and accreditation practices comply with ISO norms. Without this ISO audit, an organization is unable to process certified goods for international markets due to the TBT restrictions noted above. These ISO certification practices are rigorous and require the establishment of stable links between certifiers, standards makers, and accreditors as noted in Figure 2.

This rapid expansion, wide purview, and emergent complexity of ISO-inspired certification have prompted a wide range of research focusing on certification’s positive and negative consequences. These practices make the certification process much more transparent and the application of standards more robust from the standpoint of auditors (e.g., the US Department of Agriculture, which audits US-destined organic certifications). However, ISO-based practices also increase certification costs and make it difficult to implement and opaque from a producer standpoint, which results in the exclusion of many poorer producers such as small farmers (Mutersbaugh 2004). Below the entry assesses three aspects of certification research as noted in Figure 1: (1) the global certification economy, (2) the impact of certification on environments and social movements, and (3) the effect of certification on the everyday experiences of certified product producers and inspectors.

Certification forms a global industry

Certification’s global industry is represented in the diagram by the ISO emblem and the combined images – found on ISO World Standard’s Day (WSD) posters – that depict the effects of standards on labor and the environment. These images point toward an understanding of certification as a global network comprised of transnational institutions, certifiers, certified producers, consumers of certified goods, in essence a global industry of certifications, standards-making, standards compliance, and audits that involves – for our particular focus on environmental certifications – hundreds of thousands of “nature workers” that include farmers, foresters, fishers, and craftspeople who labor to comply with “environmental” standards and the tens of thousands of personnel working in a burgeoning global certification service sector involving certification agencies, nongovernmental organizations (NGOs), national regulatory boards, and ISO working groups.

Certification networks are, then, rapidly expanding to form a global network that comprises a set of practices that shape an everyday, lived “certification” experience, which in turn affects people, environments, and economies. Nonetheless, place still matters within this global setting: the processes of the “harmonization,” through which all global standards are rewritten to conform to a single standard, and “routinization,” through which all certification practices must also conform to a single standard, both accompany network expansion. However, as examples in this entry will demonstrate, place matters because local social movements, unions, and governments have played a role in how these norms and standards are enacted and linked to other institutions. In some cases concerned parties have challenged, modified, and resisted...
the application of ISO-style certification standards and practices. In others, ISO certification mechanisms have been replaced by alternative, locally controlled forms of certification (Schewe 2011; Kimura 2012; Gonzalez and Nigh 2005).

As a result of these limits and challenges, the global certification network is uneven, with some places and practices strongly affected and others less so. Current research on global certification networks considers three questions. The first question asks, “what are the parameters of this global certification framework, and what should be excluded and included from analysis?” An article on the “NGO-industrial complex” noted a rapid expansion in those forms of “private governance” in which products were increasingly subject to varying forms of certification (Gereffi 2001).

- First party intra-firm assessments, such as factory quality control or academic teaching evaluations.
- Second party verification arrangements (e.g., Starbucks practices) in which an outside auditor verifies a product’s compliance with firm-specified standards (that is, not set by an independent standards board).

Figure 1 Three research perspectives on certification geographies.
Third party or “Type 1” certification, which is defined as a “tripartite” relationship consisting of three components: an independent standards board, a certifying agency that is legally separate from both producers and the standards board, and an accreditation process to assure that certifiers are qualified to assess quality.

Fourth party governmental regulation of firm practices (e.g., meat-packer audits by the US Department of Agriculture).

Here we are principally concerned with 3rd party certification, which has become the “gold standard” or preferred, most rigorous form of certification (i.e., short of 4th party governmental regulation). Recent research on 3rd party certification has both examined the tripartite relations between aspect of 3rd party certification, such as audits and inspections, and also expanded the focus to examine the wider “metrology” and “audit culture” or set of sociotechnical practices and discursive understandings within which certification is based. These studies have focused also on the certification governance frameworks, for example, multistakeholder initiatives, sustainability networks, or global production networks (Freidberg 2010; Ponte and Cheyns 2013; Loconto and Busch 2010; Baird and Quastel 2011).

The shift to view certification within a governance framework has led scholars to consider an expanded number of “sites” that support certification practices. Figure 2 provides an overview of some of sites that might nominally
be included in an expanded network for certifying a “simple” cup of organic coffee. In this diagram, the tripartite relation is depicted in the center, but the diagram also includes a very large number of financial and labor subsidies from other organizations upon which certification is economically dependent. For example, certification is supported by Mexican government agencies, such as the Conservation and Forestry and the Agricultural Ministries that provide direct governmental subsidies for forest management and export promotion and also manage foreign subsidies such as those linked to REDD+ (reducing emission from deforestation and forest degradation) reforestation, and by national and transnational banks that support a coffee producer credit union, micro-lending, and electronic funds transfers.

Several recent studies have assessed the role of governmental agencies in the “public promotion of private [neoliberal] governance.” In many cases, such as the Mexican case depicted in Figure 2, governmental support has subsidized and supported certified markets; in a few instances, governments have become directly involved in fostering public–private certification initiatives – and the Chinese state has gone so far as to initiate “state-led” NGOs, “capturing” certification institutions to better promote exports (Buckingham and Jepson 2013; Foley and Hebert 2013; Jaffee 2012).

A second question on global certification networks inquires into the origins of globalized certification, asking “does the rapid expansion in certification mechanisms reflect a specific, contemporary conjuncture of global capitalism?” Geographers in particular, have explored the idea that certification reflects a reconfiguration in global production networks, commodity chains, or rent relations within the context of existing neocolonial relations (Baird and Quastel 2011; Loconto and Busch 2010).

A third and related question seeks to ascertain which theoretical perspectives will best explain the social and economic dynamics within globalized certification. Certification scholars have applied a number of theoretical perspectives concerned to understand the “certification and auditing” relations depicted on the lower left of Figure 2, asking how private and public forms of governance operate to shape legitimacy, autonomy, and democracy, how the social and economic relations depicted in Figure 2 are best assessed and understood, how embodied practices such as inspections are experienced, and how engagement rules and audit cultures (Ponte and Cheyns 2013; Buckingham and Jepson 2013; Konefal and Hatanaka 2011).

Certification shapes environments and social movements

The lower right compartment of Figure 1 highlights another arena of certification studies, namely, the role of certification practices in shaping environments and social movements. Recent research into the environmental effects of certification has sought to determine: (i) the specifics of existing environmental standards and certification frameworks, for instance, what agricultural inputs are permitted by organic farming standards, and what practices are required by organic farming inspections; (ii) the geographic distribution of environmental standards – that is to say, the location of the fields, forests, and aquatic environments that are subject to certification; and (iii) whether standards and certification practices have the desired impact on the environmental and production activities that they are designed to regulate.

For example, the Marine Stewardship Council (MSC) was analyzed using a political economy perspective, which found limited evidence that MSC certification brings environmental
benefits, and rather more evidence of geographical unevenness with few “sustainable fisheries” located in the Global South, suggesting that MSC certification produces a “sustainable fish market,” which is to say an opportunity to market fish that are certified as “sustainable” rather than improve the ecological sustainability of fish stocks in the sea (Ponte and Cheyns 2013; Eden and Bear 2010).

In many cases the environmental practices associated with certified environmental products provide uncertain benefits, and in some cases may damage the environment. Reasons for the mismatch between standards and ecological realities may include inadequate or politicized standards, and a centralization of standards-making by boards located in sites – such as the National Organic Program in Washington, DC – that introduces a social distance and complicates communication between standards boards and producers such as farmers or fishers located in other countries. In either case this “environmental distance” may result in detrimental effects: for instance, environmental inspectors may suggest practices, such as the planting of exotic erosion control species, that address the environmental concerns of soil erosion and yet introduce another ecological problem by introducing nonnative species; or farmers themselves may prefer ecologically sound practices that allow for greater productivity yet contradict accepted environmental standards, such as reducing the number of shade trees or applying night soil (Elder, Zerriffi, and Le Billon 2013; Gonzalez and Nigh 2005).

In addition to shaping environments, certification also shapes the character of environmental and social justice movements. As social movement NGOs such as the organic agriculture movement or antisweatshop campaigns have created certified standards in a bid to make these practices legible and consistent, certification practices have, during the same period, become subject to new pressures. Perhaps the most important overall has been changes to ISO certification norms, which has in recent years declared the practice of writing standards to be in conflict with standards certification. As a result, NGOs have had to eliminate standards-making activities or stop certification activities. A second source of pressure has come from states, which have in some cases, such as organic agriculture, carbon sequestration, and sustainable forestry, exerted at least partial state control over NGOs involved in setting standards and certifying these activities. A third source of pressure has come from industries that produce certified products. In some cases industry-sponsored certifications have been created to compete with NGO-sponsored standards, whereas in other cases industry actors have sought to alter certification standards to weaken standards to make them more amenable to industry practices. Taken together, these pressures have had the effect of reducing the ability of social movements to shape standards (Mutersbaugh 2004; Gereffi 2001; Jaffee 2012; Buckingham and Jepson 2013).

Within certified organizations, certification requires changes to the organizational configuration of participating unions, families, and communities. For NGOs involved in making ethical standards, ISO specifications require a separation between standards-making and certification activities, for instance, Fairtrade International was forced to create a new “FLO-CERT” organization to handle certification. For producer organizations such as farmers unions, ISO norms require the hiring of technical staff to supply certification services such as pre-inspections, extension officer training, and farmer training and accreditation exams (each organic farmer must be accredited according to ISO 17061 guidelines).
Within communities, the complexities of certified production (see Figure 2) shape social relations. Negative effects include increased economic and organizational costs that cause poorer certified organic farmers to quit certifying (Méndez et al. 2010). Positive effects include new skilled labor opportunities such as community inspection jobs, stronger unions due to increased income from premiums and higher quality coffee, and procedures that allow women farmers to receive their coffee and subsidies payments via direct deposit to personal bank accounts, thereby bypassing male-dominated unions (Elder, Zerriffi, and Le Billon 2013; Lyon, Bezaury, and Mutersbaugh 2010; Mutersbaugh 2004).

Certification as an everyday experience

The image of a field inspection (Figure 1, panel 3) emphasizes that certification alters the social relations of production in the field by reworking the relations of inspectability and surveillance. This has been described by coffee producers as “ecological neocolonialism” because under certified international standards, the decisions about what standards such as organic, fairtrade, or bird-friendly (to pick a few of many possible examples) mean, what they entail in terms of production norms, are set by multiple tiers of Global Northern institutions including ISO – which sets certification and accreditation norms – and standards bodies such as, respectively for our examples, the USDA National Organic Program, Fairtrade International, or the Smithsonian Migratory Bird Council. Producers in the Global South are required to comply with these standards without having much say in crucial aspects of standards and certification, including, for instance, how standards are formulated, what is required to meet the standards, how certifications are performed and compliance determined, how much certifications cost, how producers and producer organizations are audited, and how accreditation is performed among other aspects. In this sense, then, it is neocolonial in the sense that certification reproduces the pattern whereby Global Southern producers are required to conform to the dictates of Global Northern institutions, and ecological in that these standards shape human practices that in turn affect ecological relations in fields, forests, and fisheries.

Despite strong producer sentiments regarding ecological neocolonialism and the attendant dissatisfaction with the top-down and opaque relations between standards makers, certifying agencies, and producers, the social relationships and experiences of producers in fields, fisheries, and forests are an understudied aspect of global certification. Research that has been undertaken from a producer standpoint paints a picture of adversarial inspections processes in which trained, accredited inspectors in the field, survey in hand, require producers to answer questions regarding their farming practices, even when the inspector already knows the answer. In addition, under ISO certification norms, inspectors are not supposed to inform the inspected parties regarding how to correct errors that they uncover, thereby creating an additional source of frustration that contributes to the sense of frustration and disenfranchisement felt by inspected parties (the producers). However, it is important to note that not all certifications are alike: to take the case of organic agriculture, although the recent history of certification has led to a loss of the peer and participatory character of earlier forms of organic certification, such as those undertaken by the Organic Crop Improvement Association (which was displaced by the USDA National Organic Program), some forms of agricultural certification such as the Demeter
ENVIRONMENTAL CERTIFICATION AND ECO-LABELING

biodynamic approach still offer greater participatory engagement and higher levels of producer satisfaction (Schewe 2011; Mutersbaugh 2004; Gonzalez and Nigh 2005).

This everyday experience of certification as adversarial and imposed is also in part a consequence of the style of environmental certifications in the Global South. In order to defray high certification costs that arise from the expense of visiting very small and remote plots, producers generally undertake “group” certifications rather than certify each farmer individually. The groups that are formed typically undertake “internal” inspections in which persons within the producer union inspect each farm in preparation for “external” inspections, performed by an inspector appointed by the certifying agency, that inspect a randomly selected sample of 10–20% of plots. This leads to significant cost savings, but exposes producers to decertification since the failure of a few group members to comply with certification standards can, in some instances, result in a loss of certification for members who do comply. It also creates a situation in which some group members are subject to peer inspections by “farmer inspectors.” This can introduce significant dissent if a group member is “failed” by the group’s farmer inspector, and can also introduce new forms of identity and social status into groups. In this fashion, the specifics of certification practice, constrained both by ISO dictates and economic realities, can introduce significant tensions that call into question the notion that certification will in each case necessarily introduce greater transparency and accountability (Klooster 2010; Mutersbaugh 2004).

As Figure 2 demonstrates, as certified production networks have grown, they have become increasingly linked to supporting relationships such as financialization, state subsidies, and commercial product differentiation efforts. This linkage of course shapes the experience of certification for producers who find it difficult to determine which requirements and demands come from certifiers and which from other sources including NGOs, inspectors, financiers, state agencies, or companies seeking to commercialize the product in niche markets (Dolan 2010; Wilson and Curnow 2013; Marston 2013).

Conclusion

From the standpoint of those who are likely to read this, the question “what is certification?” might prompt them to hold up a cup of certified organic coffee or bar of fairtrade chocolate. The farmer, however, might just as well point out that she is the consumer of certification and accreditation services, for which she must pay each and every year, a certified product inspector might point to the accreditations for which they must endure training each year, and the coffee roaster or chocolatier might signal the product label (e.g., certified organic, bird-friendly) embossed on the bag or bar. Each, as we have seen, is correct. Although these multiple meanings of consumption and production may seem to confuse any tidy schema, they also signal a key point regarding the broader certification network. Insofar as “final” consumers in cafés and shops, NGOs or governmental agencies seek to use certification as a tool to promote social and environmental goals, each must also be aware of the complex geographies of global certification that contribute to uncertain and often contradictory outcomes. As the global certification network expands to include new products, people, places, and, of course, subsidies and financing, it will also become increasingly integral to state agencies, NGOs, unions, and confederations. Although we do not yet know how these expanded networks will shape our
future understandings of environmental conservation and social action, current research shows notable effects upon environments and society, and the importance of continued research (Guthman 2007; Goodman 2010; Jaffee 2012; Wilson and Curnow 2013; Naylor 2014; Freidberg 2010; Konefal and Hatanaka 2011).

SEE ALSO: Commodification of nature; Ecological imperialism; Environment and consumption; Environment and the state; Environmental governance; Environmental issues in rural areas; Environmental knowledges and expertise; Environmental performance, practice, and affect; Environmental regulation; Environmentality and green governmentality; Governmentality; Monitoring and evaluation; Nature conservation; Neoliberalism and the environment; Political ecology; Production of nature; Rural policy and politics; Sustainable development

References


ENVIRONMENTAL CERTIFICATION AND ECO-LABELING


Environmental change results from natural processes such as droughts, floods, and wildfires, and from human activities. Land transformation is the most common form of human-influenced environmental change. Humans have been transforming the Earth’s land surface since ancient times. This started with the use of fire to open up natural savannas and grasslands for hunting and early forms of agriculture. There is evidence that, as early as 8000 BP, human activities began to transform the landscapes of Europe and Asia with settlements and agriculture, albeit on a local scale. From 1500 onward, the European colonization of the New World, Africa, and Australia resulted in the clearing of vast areas of forest and woodlands for timber, cultivation, and livestock pastures. This process accelerated after the Industrial Revolution with the global area of cropland increasing fivefold and the land area used for grazing livestock increasing tenfold over the last three centuries (Klein Goldewijk et al. 2011). Urbanization is a more recent driver of land transformation. While urbanization currently occupies less than 5% of the global land area, the demand of urban areas for food, building material, products, and water has severely impacted on environments elsewhere.

Environmental change also involves the physical and/or chemical change of soils, water, and the atmosphere resulting from human activities. Air and water pollution came to prominence in the Industrial Revolution, with the unregulated burning of coal and the discharge of waste polluting the atmosphere and waterways of Britain and Western Europe. More recently, the pioneering work of Rachel Carson’s Silent Spring (1962) exposed the environmental contamination caused by the pesticide DDT, raising public awareness of the environmental cost of technological progress and the need for regulation, and setting the stage for the modern environmental movement. The rapid rise in anthropogenic greenhouse gases such as carbon dioxide (CO₂) and methane (CH₄) since the Industrial Revolution, and especially since World War II, has resulted in the warming of the atmosphere and oceans, the melting of snow and ice, and a rise in sea levels. Each of the decades since the 1980s has been successively warmer at the Earth’s surface than any preceding decades since 1850. This acceleration in global warming is a pervasive form of environmental change, and represents a major challenge for humanity and our ability to survive on the planet in the long term.

Since the 1960s, scientists and policymakers have struggled to respond to environmental change, relying mostly on solutions designed by technical experts and implemented using top-down control. While these approaches proved appropriate to simple environmental problems (e.g., point-source pollution), environmental change over time has become more complex (e.g., diffuse pollution, climate change),
triggering a paradigm shift in the way environmental management is approached (Holling and Gunderson 2002). It is now recognized that top-down technical solutions are often inadequate and that a social-ecological systems approach, which recognizes the social and institutional dimensions shaping ecosystem processes and dynamics, is required (Folke et al. 2005).

Within this approach, the concept of social learning has gained prominence. Learning can be defined as a change in behavior or attitude due to critical reflection. Social learning is therefore understood as a change in the understanding of individuals that becomes accepted as a wider community understanding through social interactions (Reed et al. 2010). Social learning is a concept that initially emerged from psychology and organizational theory, which emphasized the context-dependent nature of knowledge (Bandura 1977; Argyris and Schön 1978). The concept has gained traction in environmental management since the mid-2000s, with a substantial effort devoted to understanding the way in which social learning builds social-ecological resilience and adaptive capacity.

Social learning (in the environmental management literature) can be defined as the process of learning from and of responding to environmental experience. It is not the domain of a particular group of experts or stakeholders, but of all stakeholders and all institutions at all scales. It involves learning not only about the environmental or natural resource problem at hand, but also about one another in a social or relational sense, based on the premise that knowledge is heavily context-dependent. The involvement of all stakeholders and all types of knowledge thus increases both the efficiency and legitimacy of the solution. Stakeholder collaboration, and the recognition of different types of knowledge (e.g., expert, local, indigenous), are thus the critical prerequisites for social learning to occur. Institutions (legal, organizational, cultural, socioeconomic) that support social learning are also crucial.

Social learning is understood to encompass learning at multiple levels (individual, management, social network, organizational) (Berkhout, Hertin, and Gann 2006). It is an ongoing dynamic process, which may emerge in a locally generated or self-organized form triggered by crisis, or from part of an active adaptive comanagement strategy. It can occur across time scales, whereby instrumental single-loop learning occurs within short to medium time frames, and more radical double-loop learning occurs over longer time scales. Instrumental single-loop learning only informs and changes the most immediate technical operations, while deeper learning may change governance procedures at the organizational level, and even overarching values and norms at the policy level (Argyris 1976).

The conceptual framework of multilevel social learning for natural resource and environmental management is captured in Figure 1 (Pahl-Wostl et al. 2007). Formal and informal multiparty collaboration leads to social learning which is influenced by the socioecological context, and produces outcomes that may lead to changes in that context and thus to a cyclic and iterative process of change. These arenas of collaboration are referred to as “communities of practice.”

Early proponents claimed that facilitating social learning in environmental management would undoubtedly have positive effects on the outcomes. However, experimentation has revealed that social learning also comes with costs, risks, and pitfalls. For example, the high levels of sustained stakeholder participation required for social learning can result in stakeholder fatigue or burnout. In cases where social learning processes have failed to reconcile highly contested
environmental values and worldviews, participants have been left feeling alienated. Social learning will therefore not happen automatically. A number of preconditions need to be met, including an active civil society, receptive and willing participants with a capacity to learn, and available platforms for interaction. It is also important to note that social learning cannot be guaranteed even if these preconditions are met. This is because the factors affecting the transferal of lessons upward in organizational and policy hierarchies are often far beyond the control of those who facilitate social learning. Factors such as broader constitutional rules, governance settings and political ideologies may inhibit the effective upward transfer of important lessons (Adger et al. 2009). Finally, while social learning has been proven to assist in the solution of clear-cut environmental problems, it has yet to solve more intractable problems such as climate adaptation, where there is little agreement on both the causes and the solution.

As a concept, social learning also presents challenges. It has been used vaguely to refer to both individual and group learning, or interchangeably with stakeholder participation. While progress has been made in refining the concept and separating the social learning concept from the stakeholder participation concept, there also remains a lack of parameters to measure social learning. Furthermore, while there has been some important theoretical and single case study work in this domain, there has been very little systematic comparative work on how social learning can effectively respond to increasing inequity, complexity, and change.

Social learning is also slow, whereas environmental change can be both slow and rapid, requiring rapid assessment and implementation of the lessons of new insights. As a consequence, research and practice have scaled up the focus of analysis from the design of learning processes for individuals and groups to multiscale governance regimes that can facilitate social learning, with new insights from the fields of political geography, sociology, public policy, institutional economics, and public administration (Morrison 2014). Further development of systematic work in this domain is focusing on the role and tools of the state, and other “metagovernors,” in organizing both the ecosystem-level conditions for social learning and the higher-level enabling fiscal, administrative, and democratic preconditions.

SEE ALSO: Climate change, concept of; Community-based natural resource management; Environmental governance; Governance and development; Multilevel governance; Sustainability science

References


Further reading


Environmental citizenship has gained increasing prominence in policy-making processes and academic work over the last two decades, as a means of promoting the goals of sustainability and environmental protection, and integrating environmental concerns into political theory and modes of political engagement. However, there is no clear definition of the term; rather it aggregates the two concepts of “environment” and “citizenship,” both highly contested in their own right. A key element of citizenship is the sense of belonging to a particular community or geographically defined area (such as a city or nation-state), with implications for how citizens understand and perform their own identities. The notion of environmental citizenship in particular raises questions about where the bounds of this community or environment of belonging begin and end, as definitions of the environment multiply, are contested, and challenge assumptions about the appropriate scale of analysis and action.

Citizenship entails a conditional set of rights (freedom of speech, for example) and obligations (to keep abreast of political developments or to protect the freedoms of others, for instance) as part of this community. In the case of environmental citizenship, the primary focus is on regulating and reducing the environmental impacts of individual actions, however defined. Consequently, definitions of environmental citizenship tend to emphasize obligations over rights, challenge the geographical definition of the community to which a citizen belongs, and commonly require the cultivation of particular skills or virtues, such as scientific literacy. Environmental citizenship is an idea associated with both conservative and radical political projects, from behavior change programs which aim to tweak current consumption habits, to experiments in making alternative communities and ways of living. Debates over the proper definition of environmental citizenship and how to inculcate its associated virtues within particular populations have become a significant element in environmental politics and policy, particularly in Western democracies. Furthermore, environmental citizenship has been a key site of struggle and contestation through which broader issues of political engagement and process have played out, and where visions of the future have been co-constructed.
ENVIRONMENTAL CITIZENSHIP

(Mitchell 2009). In the twenty-first century, early developments include: the inception of environmental virtues and responsibilities as a part of citizenship; the emergence of transnational or global forms of citizenship; increasing recognition of the plurality and diversity of citizens; and changes in the categories of who or what might be considered citizens.

Many of the concepts and commitments associated with environmental citizenship have a long history, stretching back to the conservation movements of the late nineteenth century – from George Perkins Marsh to John Muir and the Sierra Club – and finding clear political expression in the environmentalism of the 1960s onwards. Pro-environmental virtues and responsibilities were articulated first by these popular movements and advocates, through treatise writing, through the creation of voluntary associations and charities, through social movements, and through protests and opposition to environmental harms. Such virtues and responsibilities included a general desire to live in harmony with the natural environment, the conservation of “natural” landscapes or endangered species, and concerns about the depletion of natural resources.

The emergence of environmental citizenship as an important concept in environmental policy, politics, and scholarship is closely linked to the discourse of sustainable development. From the early 1980s both concepts were debated at an international level, gradually filtering into national policies as well as the discourse of nongovernmental organizations (NGOs), civil society groups, and businesses. As a consequence, environmental citizenship has become closely allied with the goal of sustainable development to marry the apparently conflicting aims of human development and environmental protection, and has also been central to much academic analysis.

Changing definitions and understandings of environmental citizenship are also linked to changing expectations about citizen participation and engagement, which have been particularly marked in the context of environmental politics and decision-making, in the Global South as well as in the North. Initially governments saw environmental policy-making as a primarily technical business and thus envisaged little role for citizens to play, beyond upholding the principles expected in a representative democracy. However, repeated local planning and land-use conflicts, protests about government inaction on environmental issues like climate change, and knowledge controversies around environmental policies have all been interpreted by governance actors and social scientists as indicators of the need for greater involvement of lay citizens in environmental decision-making.

Methods for achieving this encompass opinion polling, citizen science – typically scientific projects involving lay people in data collection or analysis, for example species counts – and deliberative processes – decision-making processes based on giving participants time to learn about the issues under discussion and the points of view of others in order to reach an informed decision – amongst others. Such methods were typically justified as leading to better decisions through the inclusion of different (local) knowledges and values, helping to improve the legitimacy of policy decisions and processes, and making environmental policy more democratic in an era when science and environment-based decisions regularly have important consequences for citizens’ everyday lives (Fiorino 1990). These methods of citizen engagement have gradually become institutionalized, especially in the West, allowing for their broad use in different policy- and decision-making contexts.
Creating environmental citizens and consumers

Geographers and others have been interested in investigating how environmental citizens might be created. This was an important principle in Tim O’Riordan’s seminal book *Environmentalism* (1976), where he draws his narrative to a close with a hopeful passage about the potential for a better educated and more environmentally aware citizenry, which may put into practice many of the values he describes. More recently such sentiments have been strongly echoed in the science communication and education literatures, which emphasize the need to inform citizens and have increasingly adopted the language of behavior change. With this forming a significant part of government environmental policies, the techniques developed to promote behavior and attitude change have become increasingly sophisticated. Conventional social marketing campaigns have been replaced, for example, by the promotion of community projects, from Transition Town groups to local energy initiatives, as a behavior change and environmental citizenship tool. Likewise the concept of “nudge,” which aims to subtly influence citizens’ decision-making contexts and thus promote certain forms of behavior over others, has gained popularity in policy-making in the United Kingdom, the United States, and beyond.

In debates about the responsibilities of environmental citizens, tensions between so-called deep green and more consumerist or market-oriented models of environmental citizenship are played out, in politics and in academic work. It has been argued by many that mainstream practices of environmental citizenship, as promoted by many governments, NGOs, and businesses, tend to cast citizens in the role of consumers. Such campaigns, policies, and initiatives tend to focus on information giving – to make sure that consumers make informed choices – and price changes to alter consumer behavior in order to reduce environmental harms.

However, it has been strongly argued in many areas of the social sciences that this “rational actor” model of human behavior is flawed, and that there are many other social, emotional, and contextual influences on the consumption choices people make. Furthermore, imagining environmental citizens as consumers is seen as potentially limiting, as it constrains the responsible actions of environmental citizens to activities around their consumption, excluding activities which challenge or operate outside of conventional modes of consumption, such as protests or attempts to build low-consumption communities around housing, energy, or food cooperatives. It has also been suggested that this citizen-consumer model takes for granted the market-based logic of much policy-making and business activity, despite the environmental harms this mentality potentially creates – from negative externalities such as air pollution to the idea that societies may become locked into an economic model dependent on the continual growth of consumption.

The political scientist Andrew Dobson (2003) argues that a model of ecological citizenship is potentially more productive than one of environmental citizenship, which he feels has become too closely associated with market-based mechanisms for pro-environmental behavior change. Dobson sees ecological citizenship as representing a greater ideological challenge to the liberal state, by emphasizing responsibilities and virtues, sometimes in opposition to market logics. The political scientist John Barry (2006) has suggested the notion of sustainability citizenship as a useful alternative to environmental citizenship. For Barry, sustainability citizenship challenges consumption-based definitions of environmental citizenship by requiring an engagement with...
the underlying structures driving environmental degradation and injustice, rather than inculcating pro-environmental behaviors. Like many other academics in this field, both are keen to point out the political economies and infrastructures underlying apparently unsustainable or un-environmentally friendly behaviors. In doing this, they suggest that initiatives which fail to challenge these underlying structures, such as those focused on consumption or market-based mechanisms, will be unsuccessful or even potentially damaging.

Though many civil society groups and NGOs have to an extent bought into market or consumer-oriented models of environmental citizenship, some have also sought to challenge this model. Behavior change policies in particular seek to control citizen behaviors sometimes at the micro-level, for example through surveillance by smart technologies or attempting to more subtly influence attitudes and values. These policies are potentially constraining for the more radically minded civil society actors, such as those opting to live in eco-homes or even alternative communities or those who are part of activist and social movements, for whom marketing campaigns to get them to turn the heating down a few degrees or to take fewer car journeys rather miss the point. Furthermore, the behaviors of these more alternative environmental citizens may actively contradict some of the aims of mainstream behavior change policies, such as getting people to upgrade to more efficient and therefore newer appliances. Or more alternative environmental citizens may be held back by existing infrastructures, for example for public transport provision or the capacity for electricity grids and energy technologies to support personal electricity generation, in their attempts to build more sustainable lifestyles.

In contrast to the rational-actor model of environmental citizenship assumed by most governments, environmental citizenship could also be thought of as an ethic of care, evoking an emotional as well as rationally informed concern for and connectedness to distant others. These attitudes again challenge some of the assumptions underlying mainstream government policies in this area, yet they can potentially encompass a larger spectrum of pro-environmental actions, from mundane everyday practices such as taking shorter showers to more radical experiments in alternative living. Academics too have become increasingly interested in looking at how everyday actions and practices are linked to societal and individual attitudes to environmental protection, often concluding that household technologies and the way people learn to use them are equally as important as pro-environmental beliefs and commitments in the development of more sustainable ways of living (e.g., Shove 2010).

Scales of environmental citizenship

Geographers have long sought to challenge the assumption that the nation-state is the only or primary site relevant to questions of citizenship. As processes of globalization became increasingly apparent through the 1980s and 1990s, the idea that citizens’ rights and obligations need only apply at a country level became increasingly untenable. Individual actions, from our consumption choices to use of resources in the home, and political allegiances unavoidably impact on distant citizens of other countries and continents, and increasingly on environments and environmental processes apparently unconnected to the initial actions. It has been recognized that environmental issues do not respect national boundaries, and are often bound up in broader transnational processes associated with globalization, such as international travel or lengthy supply chains for consumer goods (e.g., Dobson 2003).
This more transnational or global view of citizenship has been characterized as *cosmopolitan citizenship*, because of the implied sense of connectedness to and responsibility for geographically distant others. Furthermore, cosmopolitan citizenship entails the conferral of political authority and administrative capacity away from the nation-state to other scales of governance and organization. On the one hand, environmental policy-making has appeared to increasingly play out on a global or transnational level – with a predefined constituency of global or transnational citizens – with faith placed in international negotiations and standard-setting by international bodies. On the other hand, in response to the apparent failures of international environmental governance, there have also been moves to promote and study what are seen as more pragmatic and manageable forms of environmental governance at continental, regional, city, and community levels, engaging with different communities of citizens. Some scholars have also noted the increasing importance of virtual networks, particularly for environmental activism and protest, which tend to be fluid and cannot be confined to conventional understandings of territory and scale.

Some have suggested that the multiscaled nature of environmental citizenship means that it can no longer be based on the exclusive membership of a particular territorial community but must be based rather on general rules and principles which are applicable to diverse settings (Held 2010). This leaves us in uncharted political and ideological territory, and attempts to formulate a political and academic project of cosmopolitanism have faced robust criticism. Despite occasional successes in international environmental negotiations and in the creation of international institutions such as the IPCC (the Intergovernmental Panel on Climate Change) and more recently the IPBES (the Intergovernmental Panel on Biodiversity and Ecosystem Services), attempts to create a global environmental constitution or bill of rights and responsibilities have thus far been unsuccessful. The challenge is to create a meaningful sense of solidarity and responsibility at multiple scales, which engages with citizens’ actual experiences of their environments and of environmental problems. Whilst the idea of cosmopolitanism brings attention to the multiple scales and networks beyond the nation-state which are important to understanding environmental politics and citizenship, its challenge to the sovereignty of the nation-state might also have the effect of appearing to support so-called *neoliberal* policies and initiatives. Cosmopolitan expressions of citizenship rights and responsibilities have thus been criticized for appearing to tacitly support the deliberate attempt by many governments to shift the burden of responsibility for environmental ills onto individuals, through the behavior change agenda.

It is not only the spatial scales of environmental citizenship that are being challenged through academic work, environmental politics, and emerging environmental problems, but also its temporal scales. The discourse of the Anthropocene, which claims that human impact on the environment has become so marked that it represents a new geologic epoch, also has ramifications for how we understand environmental citizenship. First, the Anthropocene represents a blurring between the categories of the human and the natural, the environmental and the political, which are conventionally assumed to be separate domains in definitions of environmental citizenship. This has the effect of potentially bringing the environment into the political sphere in a more radical way than simply promoting pro-environmental behaviors. By casting human life in the context of geological time, the concept of the Anthropocene also challenges the temporalities usually associated
with environmental citizenship, concerning everyday decisions and practices, electoral time spans, the event horizons of particular protests and movements, or the span of the individual life. The geographer Kathryn Yusoff asserts the need, in the wake of the Anthropocene, to engage not only with the biopolitics or biology of human life but also with the geopolitics or geology of life (Yusoff 2013), from renewed social scientific interest in the formation and effects of fossil fuels and other fossilized minerals, to a consideration of the fossil residues contemporary human societies will leave behind. Following this argument, perhaps it is more appropriate to talk of geologic citizenship rather than environmental citizenship, evoking responsibilities to and the rights of temporally distant others.

Engaging with environmental citizens

Modes of engaging with and including environmental citizens in decision-making are also constantly evolving. Following the institutionalization of deliberative modes of public participation in environmental policy, they have increasingly become the subject of academic critique (Chilvers 2009). Some scholars have been concerned with the increasingly powerful role played by the mediators or facilitators of deliberative processes, who are able to design and control modes of citizen engagement and speak on behalf of citizens in policy-making contexts. Furthermore, it has been argued that the way environmental problems are framed strongly influences the kinds of discussions that are possible during deliberative processes, and even which citizens are permitted to be involved in the first place. For example, if climate change is framed as a purely technical problem of regulating the level of greenhouse gases in the atmosphere to avoid more than 2°C warming at a global level, then only those with technical expertise related to climate and to reducing greenhouse gas emissions can be involved in deliberating over policy solutions. If it is framed as an issue largely requiring dramatic behavior change to reduce heavy emitting behaviors then lay citizens must be involved in deliberations over solutions, but they are only permitted to contribute insights on the drivers of their consumptive behaviors rather than challenging the infrastructures or political systems which may also underlie this behavior. If climate change is framed as primarily an issue of equity which requires societies to work together to ensure that the burdens of mitigating and adapting to climate change are distributed fairly and not only felt by the poorest in society, then all citizens deserve to have a say in decision-making and should be permitted to make normative judgements and draw on all aspects of their own knowledges and experiences of climate change.

Usually only a small number of methods for citizen engagement become institutionalized in governance mechanisms, and these methods then become highly standardized and are promoted across the board regardless of the specific issue under discussion. This can be seen with the popularity of a particular kind of public opinion polling in government where attitudes are elicited through survey methods from a statistically representative sample of the population. Similarly, certain highly standardized forms of deliberative methods of public engagement have been used in policy-making, planning, conflict resolution, and development projects, with subtle variations across different domains and national contexts. In more recent years there has been an increasing turn to internet-based modes of engagement, in formal policy-making as well as in activism. These new practices, including the opening up or crowd-sourcing of large data sets to be analyzed and interpreted by citizen
data scientists or the use of algorithms to analyze citizen sentiments expressed through social media, have the potential both to empower and to exclude and require new skills and understandings of citizens, just as we have seen with other forms of democratic engagement such as deliberation or opinion polling.

Geographers and science and technology studies (STS) scholars have taken a close empirical interest in how ideas about environmental citizenship and environmental citizens themselves are constructed in particular contexts. This strand of scholarship significantly departs from political science approaches to citizenship which aim to identify and define, instead shifting the emphasis to the practices, processes, and performance of (environmental) citizenship. For example, the conventional citizen imagined and therefore summoned by powerful actors around public engagement and participation related to biotechnology has been characterized as the “innocent citizen” (Irwin 2001), an individual assumed to have little subject-specific knowledge or political preferences related to the issue under discussion – and therefore in need of proper information and education – in contrast to the prominent role played by certain vocal activist groups around these issues. Such work highlights how certain kinds of citizens can be excluded from or cast as invisible to environmental policies, but also how certain visions of environmental citizens can be brought to life and reinforced through the performance of deliberative processes around environmental issues. Such accounts have also observed how debates around environmental issues, from nuclear power to climate change or the introduction of genetically modified foods, have played out very differently in different national contexts due to different national modes of engaging with citizens, institutional contexts, and histories of dealing with environmental problems and controversies.

Scholars have also been interested in studying attempts to create environmental citizens through top–down, usually policy-led approaches, from the rolling out of smart home technologies, to practices of public deliberation, behavior change policies, or housing regulations (e.g. Hobson 2013). It is argued that the embedded assumptions about environmental citizenship, and views about what environmental citizens should be and do, can become internalized by citizens themselves when they are the subjects of these top–down policies, such as behavior change. So citizens will use their homes differently when they feel they are under surveillance by the smart meters and other smart technologies which have been installed, or consumer-focused behavior change initiatives may cause citizens to internalize the role and message of sustainable consumption, neglecting the other possible roles they could play. However, it has also been suggested that these top-down policies are not absolute determinants of the performance and practice of environmental citizenship. Rather these top–down impositions can be contested and altered from the bottom up through the practices of environmental citizens, who use their smart home technologies in ways other than those expected by the designers, who subtly undermine or challenge discussions in the context of deliberative public engagement processes, or who find alternative modes of consumption away from the narrow sustainable consumption measures promoted in government behavior change initiatives.

An ongoing challenge for democratic theorists and those interested in public engagement has been how to engage environmental citizens in decision-making justly and in a sustained manner, without citizens’ voices being co-opted or people simply becoming tired of being constantly consulted. Recent scholarship has attempted to take a more systematic or organization-wide view of how participation processes play out and
how they interrelate with and feed into other political processes, for example highlighting the role played by social protest, such as climate camp or protests against the third runway at Heathrow in the United Kingdom, as well as formal deliberative processes in influencing government policy. Another tactic has been to point out how citizens are always already engaged with aspects of environmental politics through their everyday lives and practices, from the consumption decisions they make about buying fairtrade, local, or organic products or through resource use in the home. Citizens also engage heavily with environmental issues through the media, and through diverse forms of protest and dissent, from demonstrations to signing online petitions. This moves the focus away from the sometimes narrow view of how environmental citizens can be engaged with environmental policy and politics, towards an interest in how institutions and decision-makers themselves listen and respond to citizen voices. A further strand to this debate has been the promotion of the view that environmental decision-making is always partial and incomplete, requiring constant debate, development, and monitoring amidst moments of settlement and decisiveness.

Redefining environmental citizens

With the advent of universal suffrage, at least in Western liberal democracies, it would initially appear that the question of who has the right to be considered a citizen has been definitively settled. One exception to this rule is the case of the rights of marginalized indigenous peoples to be considered environmental citizens. Whilst indigenous communities are thought to broadly live in a more harmonious and interdependent relationship with their surrounding environments, official environmental policies can sometimes work to exclude and harm such peoples, for example by preventing access to important food-gathering or hunting areas despite the minimal environmental impacts of such activities.

Political developments have enshrined the rights of nonhuman nature in law in some countries, for example Bolivia which passed the Ley de Derechos de la Madre Tierra (Law of the Rights of Mother Earth) in 2010, which gives Mother Earth certain inherent rights as a subject of public interest. Bolivia has also been instrumental in inserting such perspectives on the rights of Mother Nature into the conceptual framework of the newly created IPBES. In academic work, developments in political philosophy and in animal or more-than-human geographies have further complicated this dimension of environmental citizenship. Such approaches emphasize the capacities of the nonhuman, from individual animals to plants or even bacteria and viruses, to sense and to have political significance and effects, and call for the redistribution of political agency, with implications for the practice of democratic politics and science (e.g., Whatmore 2006). Furthermore, it is not just the citizen-like qualities of individual entities that have been emphasized; there has also been a move to consider the rights and responses of systems or assemblages, for example in the growing importance of the ecosystem (including the concept of ecosystem services) as an object for policy-making and critical inquiry. It has also been claimed that the difficult moral dimensions opened up by environmental policy-making require a reconsideration of the citizenship rights of as yet unborn humans and nonhumans, opening up similar challenges for political engagement and the definition of citizenship rights.

These developments raise serious questions for those concerned with environmental citizenship
ENVIRONMENTAL CITIZENSHIP

around how the eligibility of nonhumans for citizen rights might be defined, and how they can be engaged in political processes. New dimensions for defining the right to citizenship have been suggested, including the ability to sense, the ability to suffer, or the ability to have significant effects on human communities or networks. Several geographers have adopted experimental approaches in attempting to engage with and take seriously these potential citizens. A current UK project funded by the Arts and Humanities Research Council (AHRC) and hosted at the University of Edinburgh (More-than-Human Participatory Research 2014) has experimented with adapting co-design and other participatory procedures in order to have conversations with nonhumans, including trees, dogs, bees, and water. The point of the project is to engage with and reflect on both the practicalities and the politics of attempting to adopt a broader definition of citizenship and democratic rights.

SEE ALSO: Animal geographies; Anthropocene and planetary boundaries; Behavioral geography; Citizenship; Cosmopolitanism; Participatory geographies; Sustainable development

References


Further reading


ENVIRONMENTAL CITIZENSHIP


Environmental degradation

Salvatore Engel-Di Mauro
State University of New York at New Paltz, USA

Definition and brief historical overview of the concept

A term frequently used but rarely defined, environmental degradation, in the current scientific and broader institutional mainstream is considered a specifically human-caused negative change in the biophysical (i.e., nonhuman) environment (e.g., UNEP 1997) or a human-induced decline of environmental health. The same concept may also be expressed by different labels, as ecological degradation or as ecological or environmental decline, deterioration, destruction, or disruption. Irrespective of semantic manifestation, the matter is posited as a reduction in the quality of the environment and is imputed to direct or indirect human impact, also referred to as human disturbance. Excluded from definitions of environmental degradation are nonhuman sources of environmental change that is deleterious to people (e.g., natural hazards). This does not mean that those who study environmental degradation bracket or dismiss nonhuman sources (drivers) of environmental change. Such instances are regarded as processes at most impinging on environmental degradation as either part of its effects or exacerbation of existing nonhuman drivers. For instance, accentuation of landslide frequency and magnitude in areas susceptible to such hazards may be traceable to or increased by anthropogenic slope deforestation.

Concerns over the quality of the physical environment and the state of other organisms have been expressed in diverse ways in different places over millennia, whether orally, textually, or pictorially (see e.g., Marks 2012). Such preoccupation has directly involved shifting ideas about nature and attendant diversity in worldviews more broadly, including, among others, questions of value, ethics, and ways of knowing. Describing undesired changes in the nonhuman natural world as environmental degradation appears traceable to the late 1960s, during a time of popular uprisings and of the concurrent rise of mass environmental movements in many parts of the world. As in earlier periods, matters of meaning, normativity, and other social questions have pervaded the use of the concept even while public and private institutions representing a variety of interests have tended to treat the large and expanding number of problems under the banner of environmental degradation in technical terms, requiring the application of expert knowledge certified by governing national and international organs. This latter technocratic approach has emerged some centuries ago and very unevenly with the development and especially expansion of certain national states, largely in Western Europe.

It is still an often Eurocentric technocratic approach that reigns in the framing of research questions, techniques and data gathering and analysis used, and discussions about environmental degradation. For instance, soil erosion is still considered solely in negative terms and a priority in the fight for soil conservation even though it is not a main concern for many farmers and in spite of instances where erosion can be beneficial to downslope ecosystems (Brookfield 2001).
ENVIRONMENTAL DEGRADATION

This does not imply that other perspectives are completely shut out or have been ignored within predominant institutions (such as the UNEP, World Bank, European Union, or environmental or resource-conservation branches of the world’s most powerful governments). Over the past few decades, for example, the importance of “local knowledge” to environmental monitoring and reconstructing past environments has gained considerable ground, thanks in part to scientists working within the same technocratic framework or institutions. To a moderate degree more attentiveness has been given to noninstitutional understandings of environments as well, so that there has been rising documentation and discussion of various resources as they are defined and categorized by nonexperts or, for example, indigenous peoples. This is reflected in the various subfields established with an “ethno-” prefix (e.g., ethnopedology). There are also overlaps among types of knowledge, either, among other factors, because of cross-cutting species-specific needs that to some extent transcend social differences (e.g., food and water needs) or because of increasing sensitivity by institutional actors to more grassroots concerns or because of perspectival convergence resulting from socially wide-reaching education systems, or combinations of all three.

On the main, within current mainstream institutional frameworks, environmental degradation is understood in terms of generic effects on human health and/or economic processes, but it is increasingly extended to other forms of valuation, such as aesthetic appreciation and, potentially, relative cultural framework (perhaps, eventually, even according to different social position, given the greater sensitivity to matters of environmental justice). The concept is thereby also defined by such notions as natural resource availability, productivity, and development, as well as cultural amenities. It is not uncommon to find the depletion of resources, the impairment of biomass or crop production, and implications for various indicators of development among the topics embedded in discussions of environmental degradation. Still, the concept is treated as if it were possible to free it from social specificity.

Measurement and interpretation of environmental phenomena

Environmental degradation encompasses a wide range of processes, including depletion (e.g., deforestation, soil erosion, water withdrawals faster than replenishment rates), disruption of biogeochemical cycles (e.g., alteration of atmospheric and oceanic chemical composition, soil acidification), species extinctions, and pollution (e.g., overloading of environment with trace elements or synthetic compounds). Observations and measurements of environmental degradation involve many different kinds of methods and data, in line with the great diversity of processes studied. There are direct and proxy sources of data that can be assessed qualitatively and/or quantitatively and gathered through a variety of techniques. Direct sources are those entailing measurements (e.g., rainfall gauges, field and lab soil pH analyses) or documented observations (e.g., interviews, archives, diaries, landscape paintings). Proxy or indirect sources require inference about past environments from nonhuman material remains (e.g., fossils and subfossils, buried soils, ice cores) or artifacts whose meanings are unknown to analysts (e.g., cave pictographs, sculptures, reliefs) or whose likely uses, composition, or attached residues aid paleoenvironmental interpretation (e.g., sherds, stone tools, pieces of worked metal, construction ruins). The enormous diversity in the types and sources of environmental data is reflected in the variety of quantitative and
qualitative techniques that can be employed, ranging from, for example, remotely sensed data through satellite imagery and magnetometry to more direct data acquisition through interviews and textual analyses.

Once acquired, data can be compiled and analyzed to glean changes over time and space so as to detect trends and thereby ascertain whether environmental degradation has occurred or is occurring. For instance, remote sensing instruments (e.g., aerial photography, satellite imagery) are often used to estimate the extent and type of forests over time in different places. Such data can be tested against the results of fieldwork-based inventories. For longer-term considerations, pollen extraction from nearby sediments (often in lake beds) and tree rings of existing trees can be examined to supplement current observations or carry the analysis further back in time. These research activities have been important to identifying deforestation problems and measuring rates of deforestation in different parts of the world.

Indices can and have also been devised and applied to facilitate interpretation of environmental change. The normalized difference vegetation index (NDVI), based on red and infra-red reflectance by vegetation, gives an indication of the amount of photosynthesis in an area and thereby enables interpretations about the state of plant life for a given area. Other indices have been developed that attempt to estimate human impact per se, on the basis of biophysical observations, inferred data, and indices. One recently popularized index is the Ecological Footprint, the total biologically productive land and marine area implied by resource consumption and waste generation by a given human population. Another is the human impact index (I=PAT), measured by multiplying figures for human population, affluence, and consumption rates relative to technologies used.

Involving the study of many different types of data and resulting from multiple inter-related processes, environmental degradation presents a major challenge in merely describing the phenomenon, regardless of time period considered. A common approach to classifying forms of environmental degradation is on the basis of geospheres or sets of interactions, a method that roughly corresponds to the existing division of labor among the biophysical sciences. Thus, one may describe a problem as pertaining to air (atmosphere), water (hydrosphere), land (lithosphere), and/or organisms (biosphere). Sometimes additional areas of research are specified and separated from these four geospheres, such as soils (pedosphere) and ice-covered regions (cryosphere).

There may be advantages to splitting descriptions of degradation relative to geosphere, such as avoiding the complete redrawing of disciplinary boundaries and the overhaul of education systems that would imply, but this conventional splitting can be at the expense of integrating data and analyses to detect mutually altering effects between geospheres. For example, acid precipitation can alter both soil and groundwater quality and thereby affect biodiversity in and geochemical composition of surface waters. Such environmental degradation is clearly described more effectively by combining analyses of atmospheric, soil, hydrological, and biological processes. A multidisciplinary approach is therefore necessary just to reach a basic comprehension of the problem. The Millennium Ecosystem Assessment (Hassan, Scholes, and Ash 2005) provides an example of how multiple sources of environmental monitoring and studies can be integrated to produce a broad understanding of environmental degradation. The exercise involved hundreds of scientists from disparate fields convened to discuss and carry out analyses that could put together findings in

ENVIRONMENTAL DEGRADATION
such a way as to render the results intelligible for the purpose of detecting major planetary trends as well as reaching a wider public.

Grasping the full extent of the consequences of environmental degradation is even more involving than the already challenging process of gathering analytical results from disparate fields (using diverse approaches and methodologies) into a coherent estimation of the status of biophysical environments. There are issues of complexity, cumulative repercussions, thresholds, and distance-related time lags, to cite some of the major aspects. Relationships among different biophysical processes are complex in that a change in one can have multiple and cascading effects upon others. For instance, in regions characterized by acid sulfate soils, the building of dams, and the installation and operation of deep-pumping wells, for instance, can lead to the lowering of water tables and the oxidation of underlying parent material, triggering a rapid acidification process affecting surface water quality and thereby also aquatic life forms. Heavy metals embedded in soils may eventually be released into nearby streams and bioaccumulate to the point of threatening the health of people consuming affected aquatic species. Consequences can also be cumulative and involve time lags and thresholds, such that an environmental degradation process may not be discernible until an irreversible change has occurred, as in the activation of acid sulfate soils. There are, in other words, threshold conditions that may be lurking and may not be clear until after the problem emerges. This is further complicated by the reach of some human impacts at great distances from a source area, as in the case of acid precipitation. Emissions of sulfur and nitrous oxides from coal-burning power plants can lead to acid precipitation hundreds to thousands of kilometers from emission sites. In this case, there is a time lag in that, for example, it takes time for emitted compounds to work their way through atmospheric processes and then precipitate in acid form. There is also a potential cumulative impact as, over time, acid precipitation can change the water chemistry of receiving lakes and more sensitive soils (e.g., those with little acid-buffering capacity).

It can also be no straightforward matter distinguishing anthropogenic from nonhuman sources or drivers of environmental degradation. The fact that multiple concatenations of biophysical processes are typically involved, as in the acid sulfate soils example, already makes it evident that environmental degradation entails more-than-human causation, even if it is evident that such problems are human-induced. This multiplicity of interlinked processes also reduces the possibility of gauging environmental degradation comprehensively, as one impact can have repercussions in many different components of ecosystems, not all of them necessarily negative. Much more difficult to interpret relative to drivers of environmental degradation are instances in the remote past, such as Late Pleistocene megafaunal extinctions, and in cases where changes are initiated unknowingly, such as the multiple introductions of new species that become invasive and, in turn, supplant one another in predominance. Sometimes, human-induced environmental change (degradation) bears results virtually identical in frequency and/or magnitude to that induced by nonhuman forces, as in the case of distinguishing subsidence due to karstic processes from those traceable to over-pumping of groundwater in largely limestone bedrock environments. Biophysical environments involve interdependent processes of which human agency constitutes but one (or possibly several, if one wishes to distinguish among types of societies), so there is an added complexity to pinpointing anthropogenic impact in that one must gather many different kinds of data (not always feasible) to enable the
discernment of human from nonhuman impact in general. Finally, but not exhaustively, there can be lags in time and space between putatively destructive human impact (e.g., reduction of soil cover) and its effects (e.g., sedimentation in streams and lakes) in different ecosystem components.

Data and indices can be plotted and analyzed statistically against values ascribed to a societal variable (often economic figures such as gross domestic product (GDP)) to begin formulating hypotheses regarding the causes of environmental degradation. To some extent, this is the basis for establishing causal relationships, but since there are always multiple human and nonhuman factors involved, it is seldom the case that straightforward cause–effect relationships can be established. Nevertheless, broad explanatory frameworks have been developed that guide much policymaking and enliven academic debate. Major governmental and international institutions and often large environmental organizations subscribe to multifactorial models that posit different social dimensions as influential in shaping human impact, such as demography, economy or wealth (measured according to specifically capitalist frameworks), formal politics, religious beliefs or cultural values, and technological characteristics. Some academics, activists, and policymakers emphasize one or another of these dimensions and it is often the case that population growth or poverty looms large in such arguments as a primary driver of environmental degradation. Problems also tend to be framed as a matter of maintaining resources (presumed the same for all societies) or “natural capital” within the strictures of some predetermined “stocks and flows” or “carrying capacity” (e.g., Ecological Footprint analyses) or as a set of “ecosystem services” where policy choices amount to tradeoffs among such services to a generic humanity.

Critical appraisals of environmental degradation

Yet even within such a narrow scope of institutional viewpoints, the concept remains vague and its definition, as mentioned at the outset, rarely formulated. Moreover, there has even been institutional inconsistency over the meaning of environment. Early on, by the early 1970s, the term was beginning to encompass issues such as housing and sanitation within an overall attempt at reconciling development with humans’ interdependence with ecosystems (Ambio 1979). By the late 1980s, with the popularization of the Brundtland Report, it appears that environment took on (again) the meaning of nonhuman nature in contrast to what is human-made, as in more widely accepted notions. The predominant view continues to be of environment as external nature and therefore of environmental degradation as the ruination of biophysical environments as if separable from human-constructed ones, like cities.

To complicate matters further, the term degradation is even more intrinsically malleable because the referent is an unspecified set of criteria that supposedly constitute a coherent basis to rank environments as improving or declining in quality. This is made evident in the application of the concept in more specific instances of environmental degradation, namely land and soil degradation, which have received much greater attentiveness regarding their definition. Attempts to define such processes have evinced both overlaps and substantive divergence in the content and objectives of such formulations. This is due to the inherent contingency of any such definition on the perspective of those establishing it. For instance, soil salinization is disastrous to most people involved in food procurement, but could be a boon to salt miners if salt accumulation is sufficiently extensive. The introduction of
an invasive grass species can be problematic for one community where efforts are dedicated to preserving a native grassland ecosystem, but the change may be instead welcomed in another community where there is reliance on herding and the invasive species improves on existing pastures. Or, to apply the same principle to a process more commonly regarded as worrisome, deforestation is a form of degradation to those concerned with, for instance, global warming and habitat preservation, or to those relying primarily on forests for, say, their livelihood or political prestige, but it is arguably an environmental improvement to those whose livelihood is largely dependent on growing shade-intolerant crops. Different criteria may apply relative to which purpose or need one deems a priority. The issue of social context-dependence or specific human objectives or needs becomes more knotted if one considers different species and their widely divergent survival requirements (e.g., between aerobic and anaerobic organisms), as what is degradation for one species may be beneficial for another. In any case, it seems that by and large it is a single species (some might say a single type of society) that serves as the main if not the only (often tacit) referent in defining degradation.

This is not to suggest that there are no generalizable cases of degradation whose impacts cross species and/or social needs or purposes. There are overlaps in what is covered by the term, regardless of criteria, and in how environmental degradation is more broadly understood across and beyond institutions. After all, we are all organisms enmeshed in ecosystems and there are certain basic physical processes that serve to enable all life forms, even if organisms are differently positioned within and differentially affected by ecosystem dynamics. Still, degradation implies a ranking system that points to valuation criteria. Since people have differing perspectives (and members of other species or physical forces are not involved in any discussions on the topic in any direct manner), it cannot be presupposed that ranking criteria will be the same for everyone or in every instance. The concept of degradation implies the application of a value system involving people’s varied understandings and, to go further analytically, their social position. The general use of the term may presently refer exclusively to destructive human impact (disturbance), often regarded as “artificial” or “nonnatural,” yet what qualifies as destructive remains contingent on one’s point of view and on the social relations that favor the emergence of such a perspective. Hence, environmental degradation should be understood as change that is perceived as negative or undesirable. This has been a conclusion reached, at least by some, even within the environmental sciences (Johnson et al. 1997), but without an ostensible awareness of the social processes and contextualization involved and analytically necessary to consider.

This more socially contextualizing understanding has yet to reach most academic and wider institutional circles. Environmental degradation continues to be discussed as if relative perspective has no impact on policy formation and application or on the development of research questions and the analysis and interpretation of environmental data. It is assumed that, for example, human-induced deforestation is necessarily a form of degradation even if there are cases of biodiverse forms of farming replaced by tree monocultures. Likewise, soil erosion is considered as harmful regardless of context, as if there were no instances where human-induced erosion contributes to raising soil nutrient status downslope (e.g., Brookfield 2001, 165–166). Arguably, the majority of instances of deforestation are problematic relative to loss of species, regional climate, and other deleterious effects (the case of soil erosion is rather murkier; see,
for example, Stocking 2006). However, lack of contextualization may lead to overestimations of a problem and the misidentification of areas requiring policy attention.

Geographers, especially environmental and human geographers, have been particularly attentive to this problem and have developed ways of including a critical appraisal of the variable meaning of environmental degradation, examining the social dynamics involved as well as perceptual aspects, without denying its occurrence. To some extent, concern over environmental degradation has deep roots in geography, traceable to the field’s very beginnings in the 1860s as an institutionally recognized field of knowledge (Goudie 2006, 3–11). Geographical work on the topic has also shaped the study of human impacts, as in the influential 1955 meeting and derivative volume titled “Man’s Role in Changing the Face of the Earth.” The project involved human geographers as main organizers and it was inspired by the writings of geographer George Perkins Marsh (1801–1882). It was among the first international and interdisciplinary meetings aimed at a comprehensive assessment of human impact and environmental degradation.

Nevertheless, studies or entire research agendas focused on expressly social explanations of environmental degradation are of more recent vintage. One can find precedents in approaches such as cultural landscapes and cultural ecology a century ago, which largely addressed human shaping, not degrading of environments. There are even earlier, prescient contributions, mostly of a speculative and prescriptive character, in late nineteenth-century anarchist geographers’ writings, as well as a few other precursors, but these were largely ignored (Robbins 2012). Sustained, critical analyses and explanations linking social relations of power to environmental degradation and to its socially uneven effects did not come about until the 1970s, when earlier marginalized anarchist geographical approaches were recovered and newer Marxist and feminist, and eventually antiracist and various poststructuralist perspectives emerged, influenced by intellectual currents and social movements outside geography and the academic setting more generally (e.g., works on environmental racism). Some of these reside at the margins of academia and exert greater influence in wider social movements, such as in the case of indigenous environmentalisms, environmental justice/racism, ecofeminism, and eco-Marxism/socialism. The ferment of the latter half of the twentieth century emerged as several strands of thought, from within as well as beyond geography, coalescing about a paradigm that came to be known as political ecology by the late 1980s. This paradigm, especially influenced by Marxist theories (including world-systems perspectives) in its early phase, was infused with and enriched by the aforementioned perspectives (not so much anarchism, though). There have also been separate, parallel Marxist dialectical currents in human geography that intertwine with parts of political ecology, aside from similar and copious social science scholarship emerging especially in the fields of anthropology and sociology. There have also been critical empirically focused interventions from physical geography, though more recently and rarely (see, e.g., Kiage 2013).

These varied but mostly complementary approaches have contributed fundamental correctives to the explanatory frameworks briefly described above and that dominate policymaking and mainstream environmental activism. Featured among the major contributions are demonstrations of the importance of multiscale analysis, the corollary consideration of interlinkages among places, a focus on social differentiation, and attentiveness to the dangers of common-sense ideological constructs (e.g.,
environmental narratives, received wisdom). Multiple-scaled and interlinkage analyses, for example, show how environmental degradation in one place cannot be explained and hence cannot be resolved by pretending that it is due solely or even largely to local human impacts. This understanding can be translated even to studies solely concerned with biophysical phenomena. Perspectives and conclusions about the occurrence and severity of environmental degradation may vary according to an analyst’s scale of analysis and the manner in which data are spatially bounded (e.g., how data are aggregated, pixel resolution). Studies centered around social differentiation underline the fact that social processes cannot be treated as independent entities in explaining what motivates human actions leading to environmental degradation. They also uncover major explanatory weaknesses in conventional explanations that treat humans as homogeneous (as if we all have the same political leverage, for example) and that fail to grasp the inherently inter-related character of social processes (e.g., economic growth is never divorced from political decisions over what counts as economic growth and how it is distributed socially). Therefore, such conventional explanations cannot account for the large social unevenness in both the causes and effects of environmental degradation and for the multiple social processes that combine to promote different kinds and magnitudes of human impact. Finally, taking apart common-sense views can be crucial in devising constructive policies. For instance, political ecologists have demonstrated how often received notions about past environments have been constructed as part of legitimizing the displacement of communities or in justifying colonial or other sorts of authoritarian interventions. Others have shown how technical terms are infused with ideological commitments to wresting resources from some communities and handing them over to private interests, as in the conflation of commons with open-access resources or the untenable use of “carrying capacity” to people-environment relations (since, for example, long-distance trade obviates its applicability). Still others have pointed to constructs, like “ecosystem services” or “wetland mitigation banking” that buttress attempts to commodify and thereby render alienable resources necessary to human survival.

The outpouring of these studies over the past few decades has yet to inform conventional policies in fundamental ways, but they point to ways forward in making of environmental degradation a context-sensitized and hence more viable concept as well as a means to gauge environmental degradation more accurately. Some of the results of such critical scholarship have found their way to conventional projects and have been shown to provide innovative ways of producing knowledge about environmental degradation that overcome technocratic or decontextualized understandings of environments. An example is the UN-sponsored project titled “People, Land Management, and Environmental Change,” a farmer-centered participatory approach that led to identifying knowledge and technologies that sustain agrobiodiversity and reduce destructive impact. This is one way not only of raising the relevance of environmental degradation to communities otherwise excluded from policy development and implementation, but also developing appropriate interventions addressing instances where they are actually useful and constructive both socially and environmentally.

**SEE ALSO:** Ecological footprint; Ecosystem services; Environmental health; Environmental knowledges and expertise; Environmental valuation; Environmentalism; Indigenous technical knowledge; Land degradation; Political ecology
References


Further reading

Environmental discourse

Tom Mels
Uppsala University – Campus Gotland, Sweden

Environmental discourses are frameworks of meanings about the biogeophysical world and its attendant natural and social qualities. Each environmental discourse is part of an exchange of knowledges, imaginations, assumptions, and judgments. In this exchange, the individual and combined notions of environment and discourse themselves remain highly contentious and appear in various political, cultural, and economic contexts, and webs of historical and geographical association. For example, the environment understood as nature or aesthetically pleasing landscapes differs from the environment understood as a scientifically demarcated ecosystem or toxic waste site. And discourse understood in narrow terms of conversation, writing, and language use differs from discourse understood in a wider sense as particular ways in which distinctions between truth and untruth or right and wrong are made, such as through scientific research. These wider, more political and critical, notions of discourse are of central importance to social scientists. They draw attention to the variable degrees to which different environmental discourses are expressions of power and authority, and come to pervade both social ways of seeing and material practices in the landscape, such as choosing which crops to grow where. Control over how and why particular configurations of representations and understandings are constructed and given authority, while others are downplayed, rejected, or excluded, is therefore often a vital feature of environmental discourse analysis within geography and the social sciences.

Work on environmental discourse by geographers was prompted by a wide-ranging surge of critical and notably poststructural theory within the discipline, and of an outpouring of social science literature on environmental issues since the mid-1980s. As in other fields of research, the term “discourse” originally appeared as a critical analytical tool, designed to reveal connections between, on the one hand, bodies of knowledge (rhetorical strategies, vocabulary, background assumptions) and, on the other hand, practices of social power, projects of domination, and systems of control. A cultural and linguistic turn in research insisted on the textuality of objects of geographical knowledge, including images, maps, descriptions, and whole landscapes and environments. Instead of accentuating the material world, this textual approach treated language as a conduit for the expression of social meaning about spaces and environments.

The accepted notion of textuality of the environment therefore contrasts with the intuitive notion of environment as natural and taken for granted. Drawing on Michel Foucault’s highly influential analyses of modernity, discursive practices are often theorized as naturalizing the socially made, normative expression of particular interests. The creation of a modern wilderness park, for example, may be promoted through very partial understandings of the park space as uninhabited and strictly natural, and, while this may be in the interest of the nature tourism business whose customers take this naturalness for granted, it may simultaneously entail a ban against established human practices in the new park, such as husbandry or hunting.
ENVIRONMENTAL DISCOURSE

Such discursive formations encourage particular social patterns of behavior and institutional practices, and establish particular orders of truth in the form of powerful, taken-for-granted ways of understanding. To talk of these orders in Foucauldian terms of discourse does not aim to reveal their falsity or to expose a deliberately designed “dominant ideology,” or hegemonic forms of consent, as in Marxist versions of ideology critique. It aims, rather, to display discourses as human-made constructs, emanating as they do from various practices and sources of sociopolitical life rather than from nature. Congruently, this approach encourages us to understand environmental discourses not as ideological processes expressing the struggle over political economic power, social justice, and class-based ontologies prevailing in capitalist society, but instead as the markers of shifting categories, classifications, and subject positions of regimes of truth about the environment. In this vein, several thinkers of more forthright postmodern inclination emphasize the multivalent and unstable qualities of discourse and the shifting situatedness of individuals in relation to different discourses. For yet others – including Gayatri Chakravorty Spivak, Ernesto Laclau, Chantal Mouffe, and Edward Said – the borders between discourse and ideology critique are more permeable. They emphasize the normative qualities of power and (opposing) discourses, in tandem with developing a keen interest in social justice. There is, for example, a world of difference between categorizing large-scale corporate control over Ethiopian fields as an unjust process of “land grabbing,” resulting in the “dispossession” of local farmers, and categorizing it through a discourse of legal “land acquisition” based on “modern agriculture” and farm labor.

Although Marxist ideology critique and discourse analysis differ in theory and intent, both are fundamentally driven by an impulse to denaturalize, resocialize, and repoliticize what was traditionally, or at least according to conceived stereotype, considered to be products of nature. Both would indeed question the naturalness of the aforementioned “wilderness” park. The introduction of discourse theory into environmental studies thus brought a profoundly social agenda to the critical scrutiny of the production of knowledge. The natural science pursuit of gaining objective knowledge about “the environment” was instead seen as affected by the shifting theories, terms, concepts, conventions, and accepted methods (the structured codes, or epistemes) that organized the scientists’ work and directed their possibilities of gaining knowledge. While environmental science may carry an aura of scientific precision and neutral inquiry, it was increasingly understood as a political rather than an objective field of research. To speak of environmental discourse expressed the recognition of this awareness.

Environmental discourse has subsequently been widely adopted to highlight the social construction of (knowledge about) nature, and to disclose how dominant discourses about the environment are reinforced by existing systems of education, law enforcement, media coverage, scientific research, and other practices. The importance of the concept of discourse was partly influenced by worldly developments such as the realization that uneven geographies of ethnicity, race, gender, and class across the world mapped closely onto environmental health problems. Also, a multiplicity of often conflicting interpretations about the societal causes and consequences of environmental change on all geographical scales have erupted since the 1980s. Global warming, genetically modified foods, threats to biodiversity, toxic waste, air pollution, deforestation, the nuclear disaster in Chernobyl,
Bhopal’s gas tragedy – an endless list of emblematic places and issues of environmental ruin have become vitally important to the politics of sustainable development and a globalizing environmental justice movement. For example, Hurricane Katrina was not just a “natural” cyclone that heavily damaged and polluted particular parts of New Orleans in 2005, but a disaster exacerbated by and deepening already existing social inequalities of race, class, and gender that have been structurally present in the built environment, and that continue to shape the cityscape and its protection. Contested discourses of race, class, and gender were thus of vital importance to understanding the environmental effects of Katrina’s impact and aftermath, during which the environment was profoundly politicized in terms of social inequality, demands of corporate responsibility, the causes and consequences of uneven systems of power, and ever-present struggles over interests.

Environmental discourse in academic research

There are several different (sub)disciplines that have taken up the concept of environmental discourse in the social sciences and humanities. First, within environmental sociology, interventions based on, inter alia, critical realism, social constructivist theory, and visions of reflexive modernization have gained prominence through the work of scholars such as Ulrich Beck, Peter Dickens, Anthony Giddens, John Hannigan, Michael Redclift, John Urry, and Graham Woodgate. These unsettled and subverted the traditionally dominant objectivist or what was seen as naive realist standpoints within sociology, with their reliance on the apparent naturalness of the environment, and emphasized instead sociality and contested claims made about the world.

Second, within environmental ethics, controversies arose about the intrinsic or inherent value of nature and social science notions of the social construction of nature in environmental discourses. Seeing the environment as a more or less natural entity, external to society, was challenged by more discursive approaches to wilderness and the nonhuman world. A much cited case in point was environmental historian William Cronon’s dissection of wilderness as a category of knowledge, resting on sharp but ultimately untenable distinctions between nature and culture. Cronon not only saw the idea of wilderness as a particular human construct, but also questioned the strong orientation toward uninhabited landscapes within environmental discourse, which often failed to recognize nature in inhabited landscapes such as cities. What Cronon and his fellow contributors in the seminal collection *Uncommon Ground* (Cronon 1996) gestured at was a concept of nature as fully implicated in social life.

Third, environmental historians such as Richard H. Grove, Ramachandra Guha, Donald Worster, Carolyn Merchant, and Richard White also examined ecological discourse and environmental change. Among other things, this work contributed much to deeper understandings of the character and profound consequences of colonial and gender relations embedded in past environmental discourses.

Fourth, green varieties of historical materialism also accentuated a normative edge to nature and the environment. Ecological Marxism (developed by Elmar Altvater, Ted Benton, John Bellamy Foster, and James O’Connor, among others) emphasized the social and environmental consequences of capitalism and ontological realism rather than representational practices and histories of ecological thought. Such work did develop an interest in an environmental discourse of sorts, by examining the politico-economic
ENVIRONMENTAL DISCOURSE

assumptions, ideological tactics, and various valorizations within capitalism of nature as resources.

Finally, environmental discourse also emerged as a major concern within political science where, for instance, Maarten Hajer’s work on ecological modernization and discourses of acid rain, and Karen Litfin’s exposure of global cooperation around ozone depletion, made renowned contributions to the politics of environmental discourse during the 1990s. On a meta-level, the political theorist Richard Dryzek (1997) developed a taxonomy of competing interpretive repertoires of environmental concern, identifying several environmental discourses that challenged the no-limits, Promethean discourse of industrialism that dominates mainstream society. All versions of environmental problem-solving, he argued, reject radical change and instead recognize experts, markets, or democratic processes as agents of change. What he calls “discourses of sustainability” (sustainable development and ecological modernization) aim instead at dissolving conflicts between economic development and environment. Finally, versions of green radicalism propose a complete transformation of modern society.

Environmental discourse and geography

Geographers also contributed in distinctive ways to this extensive intellectual and practical politicization of environmental discourse. With an understanding of all sorts of knowledge in terms of discourses, geographers were drawn to highly specific forms of human reasoning, learned ways of seeing, arbitrary political rationalities, and cultural preferences through which the world is made accessible and comprehensible. Mapping how and to what end particular discourses came to possess power in society was an interest shared with other scholars. Yet geographers devoted particular attention to the spatialities of environmental discourse, or the way certain more or less strategic forms of representation on a variety of scales emerged, vanished, and were implicated in the making of specific landscapes and places. For example, the European discourse of wilderness was applied in colonial Africa to make spaces available for the creation of national parks, by supplying a rationale for the exclusion, repression, and eviction of indigenous people from such spaces.

Geography’s engagement with environmental discourse appeared more or less simultaneously in two multidisciplinary fields: environmental justice and political ecology. Environmental justice studies, as developed in the work of scholars like Julian Agyeman, Brendan Gleeson, Nicholas Low, Laura Pulido, Petra Tschantz, and Gordon Walker, have, since the mid-1990s, gone through a fundamental shift from explaining environmental inequalities simply in terms of the distribution toward explaining them in terms of how difference is produced and normalized. This emphasizes both the various material-discursive processes underlying disparities and encourages emancipatory calls for environmental justice as recognition and participation in public life, environmental planning, and policymaking (Walker 2012).

In political ecology since the 1990s, discourse analyses were frequently combined with other qualitative and quantitative methods of inquiry to grasp changes in biogeophysical environments (Peet, Robbins and Watts 2011). Part of this work is discontented with any geographical analysis that tends to separate natural processes from social processes (culture, politics, economy) in environmental change.

Any discourse research, therefore, encouraged a rethinking of the material efficacy and entanglement of discourse: that is, how knowledges
ENVIRONMENTAL DISCOURSE

and modes of representation can transform and reproduce spaces and the biogeophysical world. The problem of discourse is thus at the heart of the matter of nature and related mind/body, self/world, nature/culture, material/mental dualisms. In this context discourse is used to explore a wide variety of research questions, including those concerning the material properties of nature, the cultural predispositions that accompany our concepts of environment, and the role of an array of nonhuman things and processes.

An ongoing conversation about the discursive construction and materiality of nature was inaugurated during the 1990s, involving poststructuralist deconstruction and ecological Marxism (Braun and Castree 1998). This claimed that the environment always appears in socially interpreted guises and shapes, and that, therefore, it is crucial to scrutinize discursive conventions, silences, and affordances (the possibilities of action made available through discourse), and also to trace how those guises influence ongoing environmental change on the ground. However, it has been argued that what is at stake is not just a question of discursive mediation, but of coproduction through practices and performances of nonhuman life. From a relational vantage point, environmental discourses at all times actively constitute embodied and enacted, socially specific modes of existence.

Under the combined influence of Michel Callon, Bruno Latour, and other actor-network theorists, and so-called new materialist perspectives developed by Jane Bennett, William E. Connolly, Manuel DeLanda, Gilles Deleuze, Félix Guattari, Isabelle Stengers, and others, the constructivist focus on thinking and representing human subjects is largely rejected. Geographical interpreters such as Ben Anderson, Nick Bingham, Bruce Braun, Steve Hinchliffe, Hayden Lorimer, Nigel Thrift, and Sarah Whatmore lament what is seen as the privileged position given to humans, their agency, practices of representation, and discourse. Instead, we need to go beyond discourses to look at nondiscursive, tangible, and material things too. A common commitment in this work is thus to shift attention to the constantly changing exchange of things, technologies, and forms of life (Whatmore 2002). Within these combined human–nonhuman processes, new assemblages and unstable hybrids are both steadily vanishing and being made. No entity, concept, or action can be privileged over others, and therefore what are perceived as the increasing abstractions or, in some cases, self-referential tendencies of discourse analysis remain unable to give matter and becoming its due. In that manner, the new materialism calls not so much for the abandonment of environmental discourse and entanglements of power in environmental research, but for a recontextualization of discourse in the ongoing processes of environmental change.

Although many researchers have exposed the limitations of discourse analysis, they remain important to geographical work on the environment. Against what is taken to be the idealist tendencies of discursive construction and the monopoly of agency by humans, Marxian scholars like Karen Bakker, Gavin Bridge, Noel Castree, George Henderson, Becky Mansfield, Thomas Perreault, Scott Prudham, and Erik Swyngedouw have also positioned discourses as one way in which the biogeophysical elements of social and economic life are constituted. Unlike the new materialists who build on a flat ontology of relations, Marxian scholars emphasize the importance of capital in shaping society–nature relations, although across very diverse discourses and contexts (Harvey 1996). In recent years, this take on environmental discourse has been pursued in critical geographies that analyze neoliberalism as an environmental project. The
construction and mobilization of discourses—such as about the compatibility of economic growth, capitalist competition, and environmental care—are seen as feeding neoliberalism’s unhealthy and unjust appetite for new modes of appropriating natural resources (Heynen et al. 2007). The significance and agency of environmental discourses are particularly well recognized in literature on the neoliberalization of nature that focuses on appropriating biophysical phenomena and resources for environmental ends. According to these scholars, discourses about ecosystem services, ecotourism, biofuel production, biodiversity conservation, biocarbon sequestration, and the like alter material access to, and use of, resources, and also invent new forms of value, meaning, and intent. In particular, green discourses of commodification (i.e., discourses that are instrumental to bringing ecologies, aesthetic experience, or CO2 to the market as commodities) remain entangled within the network of capitalist economies and their commercial rationales and discursive values. For example, the producers of biofuel for cars use discourses about the nature-friendliness of biofuel to justify their use of rural spaces and resources that might have been used for food crops instead. Such discursive maneuvers contribute to the commodification of nature and are indispensable to convince customers and to secure market shares in a capitalist economy.

SEE ALSO: Commodification of nature; Construction of nature; Discourse; Environmental ethics; Environmental history; Environmental (in)justice; Environmental performance, practice, and affect; Environmental science and society; Neoliberalism and the environment; Poststructuralism/poststructural geographies; Power; Reflexivity; Representation

References


Further reading

Environmental education

Lindsey Atkinson
University of Hull, UK

Environmental education (EE) can be defined quite broadly as education about, from, and for the environment. Education about the environment seeks to develop knowledge and understanding about the environment and the skills needed to benefit the environment. The environment here includes the living (biotic) and physical (abiotic) environment, both natural and built. Environmental education can also include education from the environment, using the outdoors as a learning resource to enhance learning (Palmer and Neal 1994) (Figure 1). Education for the environment aims to develop a sense of awareness of and responsibility for the environment and thus aims to influence attitudes and behavior.

While environmental science is largely associated with the disciplines of geography and the sciences, environmental education is viewed as interdisciplinary (UNESCO 1978), incorporating social, economic, and political perspectives. Environmental education takes place in both formal and nonformal education settings, through governmental and nongovernmental organizations. Teaching methods place an emphasis on fieldwork to develop awareness and skills, and also on participatory learning to engage learners with the subject matter. Environmental education is carried out at different ages and levels, and is also very much considered as part of lifelong learning.

Topic areas that may typically be covered include materials and resources, in particular energy, climate, water, plants and animals, buildings and industrialization, and people and communities. In early childhood, environmental education is likely to be topic-based and to focus on the development of basic skills such as numeracy, problem-solving, and communication. At this level, outdoor education is often used to enhance all areas of the curriculum. In secondary education most environmental education takes place in geography and the sciences, and some aspects may be covered in history. Cross-curricular opportunities may be more limited at secondary level as a result of timetabling and curriculum constraints, although there is still scope for individual schools or teachers to develop learning opportunities outside of the formal curriculum. In higher education, although most environmental education will still occur in geography and the sciences, some aspects may be covered in many other disciplines, such as law, politics, literature, business, and health studies. Schools, colleges, and universities may also act as environmental role models, demonstrating, for example, recycling, energy conservation, and the promotion of biodiversity. Outside of the education system, informal environmental education may occur through visits to museums, wildlife reserves, and so on and through organizations such as the Field Studies Council in the United Kingdom.

Environmental education has its origins in the study of natural history, followed by the development of the discipline of ecology from the late nineteenth century. Around the same time, Patrick Geddes (1854–1932), a Scottish town planner whose interests spanned biology,
Prior to the 1960s and 1970s environmental education was an activity carried out in a number of different disciplines, including conservation, environmental and urban studies, and outdoor education, but the term itself was not widely used. For instance, at the first meeting of the IUCN at the Conference of Fontainebleau in 1948, the importance of the role of education in conservation was noted as either “conservation education” or education for the “protection of nature.” In the 1960s environmental education began to establish an identity of its own in the literature with the setting up of the Journal of Environmental Education (formerly Environmental Education) in 1969. In the first issue William Stapp and others put forward an early definition of environmental education as “aimed at producing a citizenry that is knowledgeable concerning the biophysical environment and its associated problems, aware of how to help solve these problems and motivated to work toward their solution” (Stapp et al. 1969).

A fuller definition was agreed in 1970 at the IUCN’s International Working Meeting on Environmental Education in the School Curriculum: this definition acknowledged the interrelatedness between people and the environment to a greater extent. Following Recommendation 96 from the United Nations Conference on Human Environment (the Stockholm Conference) held in 1972, the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the United Nations Environment Programme (UNEP) held an International Environmental Education Workshop in Belgrade in 1975, at which the International Environmental Education Programme (IEEP) was launched, with Stapp as its first director. Resulting from this meeting, the Belgrade Charter, “a global framework for environmental education,” put forward a goal and objectives for environmental education.
which noted the need not only for awareness and knowledge but also to develop attitudes and skills.

The first Intergovernmental Conference on Environmental Education was held in October 1977 at Tbilisi (UNESCO 1978). This conference referred back to the sense of urgency about environmental problems expressed at the Stockholm Conference and emphasized the important role that education has to play in addressing these problems. The meeting set out the purposes of environmental education, expanding on the Belgrade Charter with three goals relating to awareness, skills, and attitudes. The declaration from the Tbilisi Conference also built on the guiding principles for environmental education put forward by the Belgrade Charter. According to these principles, environmental education should be for all ages, lifelong, and interdisciplinary, and should cover all scales from local to international. In addition, environmental education should cover a breadth of teaching and learning approaches, focusing on practical problems and with learners participating in planning their own learning. As environmental problems are often complex, there is a need to develop critical thinking and problem-solving skills. The meeting also considered that environmental education should be integrated into the curriculum, rather than being taught as an additional subject. Since Tbilisi the momentum has been maintained with subsequent international conferences (e.g., International Conference on Environmental Education (ICCE); World Environmental Education Conference (WECC)) and the establishment of international journals on the subject: the Australian Journal of Environmental Education was established in 1985, followed by International Research in Geographical and Environmental Education, Environmental Education Research, and the Canadian Journal of Environmental Education in the 1990s.

There has been increasing emphasis on sustainability and sustainable development since the 1980s. Consequently, in the 1990s there was a shift toward environmental education for sustainable development (EEFSD) or environmental education for sustainability (EEFS), and later to education for sustainable development (ESD) or education for sustainability (EfS). ESD incorporates ideas from both environmental education and development education (DE). While there is overlap between environmental education and ESD, the two have different roles: environmental education tends to focus more on the knowledge and values of the individual, whereas ESD looks more at a move to a sustainable society (Tilbury and Cooke 2005).

The importance of education was again emphasized in the Rio Declaration, which resulted from the UN Conference on Environment and Development (UNCED, also known as the Earth Summit, held in Rio de Janeiro, 1992) and the subsequent action plan for advancing sustainability, Agenda 21. However, the extent of progress since Tbilisi continued to be questioned. Recommendations from the UN World Summit on Sustainable Development in 2002 included a call for a Decade of Education for Sustainable Development, which ran from 2005 to 2014 with the aim of promoting the role of education and training in working for a sustainable future.

There are two main areas of criticism of environmental education. First, there may be issues about how it is presented, such that it can be seen as advocacy rather than unbiased presentation of factual information. The use of the preposition “for” in “education for the environment” may in itself be problematic and may appear to be value-laden (Scott and Oulton 1999). Jickling and Spork (1998) view “education for the environment” or “education for sustainable
development” as a slogan and argue that, while slogans may be useful for providing a focus, they can also develop into doctrines. Complex arguments about environmental issues may be simplified, potentially leading to misconceptions, and in some cases the presentation may be alarmist (e.g., on overpopulation, depletion of oil reserves). These alarmist approaches may be counterproductive when the predicted events do not occur, resulting in uncertainty and lack of trust in the sources and deliverers of the information. To counter this, students need to be made aware of the complexity of environmental issues and to develop skills in critical thinking and problem-solving (Palmer and Neal 1994). They will need to consider plural understandings rather than privileging certain viewpoints (Scott and Oulton 1999) and to apply systemic thinking (Sterling 2001).

The second area of criticism, according to some views, is that environmental education has been relatively ineffective and its outcomes uncertain (Scott and Oulton 1999; Sterling 2001). This may have resulted from a top-down approach based largely on knowledge deficit, along with a transmissive model of education, that is, one focused on the one-way delivery of information by experts. Sterling (2001) argues instead for the need of a transformative model of education to effect change. This would take a democratic, bottom-up approach, with a focus on participatory learning. Education may also be seen to play a role in developing environmental citizenship as a route to sustainability.

Despite these criticisms, environmental education has become more established and has broadened in scope since the 1960s. As ideas around sustainability have continued to develop, the concept of resilience has been explored in the context of the relationship between learning, society, and sustainability (Krasny, Lundholm, and Plummer 2010). Thus environmental education may contribute to preparing for and managing complex change, through building resilience from the level of the individual to that of social-ecological systems.

SEE ALSO: Environmental citizenship; Environmental knowledges and expertise; Environmental science; Environmentalism; Social resilience and environmental hazards; Sustainable development

References


Environmental ethics

Emma Roe
University of Southampton, UK

Environmental ethics

Why study environmental ethics?

Toward the end of the twentieth century, environmental ethics emerged to address the mounting environmental problems. The hope was and still is that, through addressing humanity’s responsibilities to care for, protect, and offer adequate stewardship of the nonhuman “natural” world, humanity might learn to change its ideas and practices and thereby tackle environmental problems. Environmental problems are those that arise from human dealings with the natural world and its physical processes. The natural world supports human life, through the production of food and clothing and through providing shelter, water, the air we breathe, and the fuel we consume. This treatment of the nonhuman world for human exploitation to meet humanity’s needs is called an instrumentalist view of nature. The instrumentalist viewpoint treats nature as existing exclusively to meet human ends, and is an attitude that has dominated for centuries. Thus instrumentalism is also an anthropocentric view stemming from the belief that the whole of creation exists for the sake of humanity and that humanity was given dominance, “rightful mastery” over all fellow creatures. However, in the last few centuries there has been a change in the extent and rate at which the natural world is being used as a resource for humanity. Not only are we using nature, but we are also producing “nature” (Smith 1984) in novel forms. As we enter the twenty-first century, the scale, penetration, and domination of humanity’s activities on the planet are radically altering human relations with the environment as a sustainable resource that supports life. Despite the principles of an instrumental relation with “nature” having been established millennia ago, from which the practice of treating the natural world as a resource derives, it is the extent to which “nature” has become a resource under industrial capitalism in the twentieth century that is of urgent concern. Smith (1984) argues that capital produces “nature” itself in material form. Thus, it is widely accepted that not only is human activity affecting the nonhuman natural world, contributing to climate change, the extinction of species, the pollution of water supplies, and the destabilizing of fragile ecosystems, but it is also manipulating life at the genomic level, giving life to un-“natural” beings in animal and plant form. Equally, concerns are raised about how we look after nonhuman life that we control and manage, for example animals produced for human use on farms, in laboratories, and as domestic pets. This range of human activities raises serious questions about the inclusion of the nonhuman “environment” in its diverse forms within politics and how to adequately address humans’ ethical relationship to it.

What is ethics? Ethics is the branch of philosophy concerned with how we should decide what is morally wrong and what is morally right. As humans we are faced with moral decisions about the right thing to do. These may be matters of great concern or, more often, small matters:
ENVIRONMENTAL ETHICS

Should we buy eggs from hens kept in battery cages? How should we travel to reduce the negative impacts to the environment – by bicycle, car, or plane? We may give a lot of thought to some questions and very little or no thought at all to others. Ethics is a specific discipline that probes the reasoning and the practices that engage us in a moral life. It does this by critically examining and analyzing the thinking that is or could be used to justify our moral choices and actions in particular situations. Geography, as the discipline that specializes in studying the relationship between humans and the environment, is particularly engaged with addressing environmental ethics through empirical research and theoretical and methodological innovation.

How should we foster practices that are more in sympathy with sustaining a quality of life for both humans and the rich diversity of nonhuman animals and plants that inhabit this planet? Can these ethical considerations be extended to inanimate natural processes and geologic life, with which humans, animals, and plants have codependent life-supporting relations?

Anthropocentrism

During the Enlightenment period from the mid-seventeenth to the eighteenth century, when the authority of God and monarch was decentered and human individuals were positioned at the center of the social world, the grounds for establishing a moral ethical framework based on the human capacity for reason and agency took hold. The consequence of this shift was to emphasize humans’ exceptionalism as beings capable of reason and agency, and to dismiss the nonhuman world as lacking reason and cast (wrongly) as insensible and inert. The great variety of beings and entities that fell into this category of the nonhuman natural world was not differentiated. Plants, animals, rocks, and rivers were marveled at for being the antithesis to humanity and for the qualities of naturalness that were “untouched” by humans. Human civilization in Europe at the time witnessed increasing numbers of people living in industrial, polluted towns and cities through the eighteenth, nineteenth, and twentieth centuries. Notably, this dualistic representation of the nonhuman world of “nature” on the one hand and the representation of humanity on the other was a further extension of the anthropocentric position. The modern thinking that produced an ontological divide between the world of nature and the world of society continues to have a tenacious grip on the cultural meanings of nature and, therefore, approaches to environmental ethics. Accompanying this thinking was the belief that it was morally appropriate to consider only the interests of humans, thus excluding the nonhuman world from moral judgments. Consequently, the nonhuman world of animals and plants was considered outside the remit of ethical concerns and the scope of human morality in Western society. In summary, the division between society and nature characterized much post-Enlightenment thought. The ecofeminist Val Plumwood (1993) has argued vigorously against this mode of thinking, in which the essence of human nature is located in a masculinist, rational human mind, in opposition to the physical, embodied, and emotional nature of the feminized human.

Despite post-Enlightenment thinking being still very much alive today – for example in how the term “natural” is used to describe an aesthetic that implies the absence of human intervention (e.g., natural birth, natural food, natural landscapes) despite this being far from the case. Take, for example, organically certified milk: despite controls over the use of antibiotics on organic cattle, unnatural practices are still at large in the
industry, where artificial insemination is used to breed the dairy cow. Various events through the last 200 years have heralded changes to the level of concern human society expresses about its relationship to “nature” and the environment. To illustrate, the following are a selection of the historical events that have paved the way for animals to become included within Western society’s circle of ethical reasoning.

Close relations have formed between humans and animals over millennia, and yet it is only in recent centuries, as human society has moved away from an existence based on subsistence farming to be able to focus on more than food production, that concerns about the plight of animals has come to the fore. Jeremy Bentham wrote about animals in his *Introduction to the Principles of Morals and Legislation* in 1789: “But a full-grown horse or dog is incomparably more rational and conversable than an infant of a day, or a week, or even a month old. Even if that were not so, what difference would that make? The question is not, Can they reason? Nor, Can they talk? But, Can they suffer?” (1907/1789, 144).

In the United Kingdom the Royal Society for the Prevention of Cruelty to Animals (RSPCA) was established in 1840. In 1954 Ruth Harrison published *Animal Machines*, which critically examined and exposed the conditions of animals raised in the newly established industrialized large-scale farming systems in post-World War II Britain. Swiftly following Harrison’s pioneering book was the first UK government inquiry into the wellbeing of farmed animals, the Brambell Report, published in 1956. The report outlined recommendations for the living conditions of farmed species. It included the Brambell version of the five freedoms: “freedom of movement to be able without difficulty, to turn around, groom itself, get up, lie down and stretch its limbs” (Brambell 1965, 13). These have since gone on to form a key pillar in applied animal welfare ethics globally, and continue to inform production standards for farmed animals. In the second half of the twentieth century, an ethical literature dedicated to animal rights became more widespread and affected public consciousness. Most notably, Peter Singer in his pivotal 1975 book *Animal Liberation* argued for animals to have rights, claiming that the boundary between human and animal is merely arbitrary. Animals can be intelligent and have reasoning and/or linguistic capacities; like humans, they also have the capacity to feel, most notably to suffer and experience pain (Singer 1975). Since the mid-1980s animal welfare science, a discipline that provides scientific knowledge specifically on the capacity of animals to suffer, to experience pain and positive emotions, has become established. This science has supported moves to change the legal status of animals. In 1997 the European Amsterdam Treaty legally recognized animals as sentient beings, and by 2009 the United Kingdom Farm Animal Welfare Council (FAWC) was looking to a future beyond the five freedoms, where all farm animals would have a life worth living, from the animal’s point of view, and where an increasing proportion of farm animals would have a good life.

These various historic milestones both chart the development of normative challenges to the abstract moral philosophies that had placed human interests central in all ethical concerns, and mark shifts in society’s cultural thinking about animals. Changes to human cultural thought and practice continue to alter the meanings given to the nonhuman entities that make up “the environment.” These changes are guiding the direction of scientific funding to ask different questions about nonhuman entities and ecosystems. Together new knowledges, scientific or otherwise, and the impact of particular events inform the development of environ-
Environmental ethics

mental ethical frameworks that guide environmental policy in its widest sense from the care for farmed animals to the management of ecosystems in coastal waters. Some readers may think the illustration of the development in animal welfare to be something quite different to the ethical work related to thinking of the environment. It is notable that animal-welfarists tend to prioritize the value of the wellbeing of individual animals, whereas environmentalists think holistically about the whole ecosystem – plants, animals, and biophysical environment – and how its collective health can be sustained. This leads to value attributed to continuing the existence of species, and maintaining biodiversity whereby the individual is valued in relation to its contribution to maintaining the whole. Indeed, while both fall under the broad category of environmental ethics, animal welfare and ecological environmental approaches often have quite different orientations and, at times (though this may not always be the case), this has led to different types of ethical reasoning. For example, deontological ethics, which is concerned with rights as advocated by Tom Regan and Peter Singer, is easier to apply to animals than to a river. An animal can be given rights to feed and water, and someone can have a duty to provide that feed and water. But what rights does a river have? Does it have a right to flood annually and to change its course slowly over time without interference from humans? Are those what a river needs to have a life worth living? Instead, it may perhaps be more beneficial to place the river within an ecosystem that is a habitat that supports nonhuman animal and plant life. The “happy” state of the river is then judged by the flourishing of animal and plant life in its waters and on its banks and floodplain. It is these types of questions that are actively discussed.

An alternative approach to environmental ethics is one that puts it in broad alignment with deep ecological perspectives and attempts to address the complexity of ecosystems. These approaches have sometimes been informed by Lovelock’s Gaia hypothesis which dates from 1979. This theory conceives of the Earth as a self-regulating system, maintaining the conditions that support life, and as such it is a “superorganism.” These ecological perspectives share an impulse to extend human consciousness to include recognition of our embeddedness in constitutive relations with the nonhuman world, in the commitment that humans should not be the sole measure of ethical considerations. However Plumwood identifies a humanism at the center of these nonanthropocentric enterprises, which links ecological imperialism with the “master model of self” (1993, 145), because they privilege the self–world relationship and deny difference in terms of gender and race in their conceptualizations of self–world relations. Despite attempts at an intersubjective conception of ethical agency in these deep ecological perspectives, the nonhuman world is still conceived as an extension of the human being rather than as having value that is independent of human life.

Environmental ethics in the Anthropocene

Bill McKibben argued in 1990 that we have reached the end of “nature” as a separate entity untouched by human life (McKibben 1990). More recently, Crutzen and Stoermer (2000) have proposed that we are now living in the Anthropocene – an era in which the biogeochemistry of the planet itself is being altered and the climate is being destabilized, because ultimately humans are influencing coevolution at a planetary level. As a consequence, it no longer seems feasible to continue to place humans at the center of an ethical framework as this
immediately places limits on the capacity for ethicists to properly consider the interests of nonhumans and to recognize their unique and highly differentiated capacities and affects in the anthropogenic environment. It is for these reasons that environmental ethicists have sought to move beyond anthropocentric perspectives on the nonhuman natural world, and rather to meet the demand to engage with anthropocentrism as a material fact in the Anthropocene. This point is made to stress that it is not only about how humans represent and give meaning or values to the nonhuman world that needs to be written into our ethical frameworks but how the environment is materially produced and internally transformed by the activities of humanity.

The challenge to include how humanity has materially transformed the environment has led some environmental ethicists to move away from the historic ontological framework of division between society and nature, human and animal, to adopt frameworks of ontological relations. Debates about ontology are interested in exploring what the nature of being is, and refer loosely to what constitutes the physical stuff in the world. Geographers have made important contributions to the field of a material, relational environmental ethics through developing a number of strands of thinking that work to develop environmental ethical frameworks around an ontology of relations between the heterogeneous entities, beings, and persons in the world, rather than ontologies that divide the world up into categories of different entitlements for ethical consideration. The implications are that ethical frameworks revolve around what ontological choices one makes and the consequences of who or what qualifies for ethical consideration. Does the environment — as a place separate from humans — still exist? And, if it doesn’t, this demands a radical shake-up of what environmental ethics can be. Indeed, the debates within geography and other disciplines about ontology have heralded the advent of nonanthropocentric ethical frameworks for studying environmental problems. But, as various commentators have made clear, the ethical framework needs to be supple, sensitive to contingency and inclusive.

What has influenced this turn toward a relational ontology? Perhaps most influential is feminist environmentalist ethical thought that built on the thinking in Plumwood (1993). Plumwood posed the challenge to contend with a humanism that went beyond engagement with the human characteristics dominated by the white rational male, to include human characteristics of emotional vulnerability and gender and racial differentials. To address this challenge, there has been an important shift toward studying humans as embodied beings that are practically engaged with the world around them to the extent that it is less clear where the human ends and the nonhuman world begins. The term “more than human” has become popular to gesture to this sense of walls coming down between self and other, the other being nonhuman. Sarah Whatmore (2002), the clearest advocate in geography for this form of thinking, argues that we should focus our attention on how materiality connects humans and nonhumans together and do this via mapping out imaginative cartographies of these connections to articulate a material ethics.

Relational material ethics

Whatmore’s influential book *Hybrid Geographies* (2002) puts into practice an ontological politics based on the inseparability of nonhumans from what should constitute the social, and by doing so opens up a different way to explore environmental ethics. Hybrid geographies are mapped to disrupt categories of humans and nonhumans and hence the “purification” of the
world of nature and the world of society. There is a strong sense in this work of the world not staying still. This is a world that is always being caught up in dynamic processes and practices that undermine the stasis of pregiven entities, and therefore looks toward the immanent potential of something very different possibly coming into being. This move is significant for how it reorders our ethical community beyond the universal human subject as through the cartography of networks (or rhizomes), a vast range of “affective actants-in-relation” (Whatmore 2012, 251) perform to make and hold their shape. The shapes they make and hold are precarious achievements whose durability and reach are either threatened or supported by humans. For example, the market for free-range eggs is a sociotechnical achievement made up of hens, charities, supermarkets, science, and compassion, whose precarity is underlined by the fact that its market share has only really become established in UK supermarkets in the first decade of the twenty-first century. There was a time when the free-range egg shelf in a supermarket did not exist and had not been imagined. It has become possible to sell free-range eggs at a higher cost than barn or caged eggs because shoppers feel compassion toward the plight of the battery hen through the campaigns of animal welfare charities raising the profile of hens in the media; the effective use of findings from hen welfare science to give credibility to the claims made, and ultimately the growth of support for criticism of caged hen production systems; and the use of EU legislation, specifically the 2012 caged system ban, which encouraged the egg industry to invest in changes to their production systems.

Crucial to Whatmore’s approach is the attention given to what Michel Serres names as “extra-linguistic forces, bonds and interactions,” acknowledging the impossibility of language being the mode of engagement between humans and nonhumans. It is the ways that humans and nonhumans interact, and affectively bond with or force their presence on each other, that direct her cartographic methodology. Returning to the free-range egg market example, there are emotions and affects that shape its precarious achievement: humans responding with concern about the experiences of hens living indoors or outdoors, in a cage or in a barn; the use of pictures and drawings on egg boxes of imagined rural landscapes with free-range hens happily clucking around. Drawing on Whatmore’s thinking, it could be argued that the ethical connectivity between hens and humans is oversimplified, whether this market were to be understood only in utilitarian ethical terms, or merely as an example of hens being given more rights. Whatmore’s approach enables spatial imaginaries and spatial practices to be witnessed that challenge “myopic parameters of ethical connectivity” (2002, 158) to complicate geographies of intimacy and affect that have, to date, configured conventional understandings of what the worldly responsibilities of humans are and should be. For, to give a different example, when we eat meat in the form of a burger, it can be easy to forget about the animal from whose body the meat came from. Why should we care about the life of the animal that has become meat? Or why care what food the animal from which our meat came from ate? Bovine somatic encephalopathy (BSE), or “mad cow disease,” and the pathological connection with Creutzfeld–Jakob Disease (CJD) – a brain disease in humans – demonstrates the significance of mapping the practices, processes, and events in the life of a nonhuman actant. In this case the prion remains intact through the metabolic pathways between beef fed to cows, which are then fed to humans. Eating prion-infected meat can make people susceptible to the brain disease. Thus cartographic practices of this kind can
emphasize the relevance of exploring new spatial imaginaries and spatial practices as they can make distant, remote practices relevant and ultimately engage ethical responsibilities in new directions.

Relevant to Whatmore’s theorization of hybrid geographies is the need to begin from an embodied and practically engaged human self. Feminist phenomenological thinkers have developed a theory of the flesh and a theory of intercorporeality to indicate more carefully how a material ethic can evolve from this material, bodily engagement with the self. For example, the embodied practice of eating directly places the bodily flesh of human in intimate relation with the bodily flesh of another body, whether human or animal, when eating food. Thus, when we eat foodstuffs with different material provenances, how we engage with the edibility of an egg from different farm production systems – caged, barn, free-range, or organic – and specifically how the material aesthetic plays a part in the practice of eating, are of ethical relevance. The affective response that leads to the quality of an egg from a battery cage system as being not as “tasty” as one from an organic free-range system, illustrates an intercorporeal ethic in operation. Whether or not there is a discernible difference purely in the aesthetic taste or appearance between one egg and another is a different matter. The aesthetic appreciation of the eggs as a result of the ethical connections between egg, hen, and human are evident and are bundled into the eating experience for some consumers.

A second influential theme for Whatmore, which emerges again from feminist ethics, is the praxis of care that is elaborated through the concept of an intersubjective situated self. Recent advances in this element of thought can be found in Haraway’s (2008) concept of “sharing suffering” as an embodied intersubjective ethical position outlined in her book *When Species Meet*. Here she explores possible alternatives to interspecies relations as they are currently known, arguing for a new style of ethical consideration that, instead of drawing on ethical principles to explain and guide decision-making, seeks to understand ethics as practiced in the everyday work of caring. In terms of addressing how we share suffering, Haraway argues that it is through a performative approach – in other words what we do practically, as opposed to abstract reflection – to ethics that we can *respond* to another’s suffering rather than *attend* to it. In this way Haraway’s concept of “response-abilities” is a way to understand “ethics in practice” as more than representational. This approach focuses less on making decisions, showing good judgment, and suppressing gut feelings, and more on the ability to feel and respond, which modern moral frameworks have often sought to drill out of subjects. The hope is that we can better enact the principle of minimizing harm and suffering through learning how to become more responsive to the nonhuman material forms with which we live.

The object of ethical concern therefore affects how we perform ethically, how we establish more care-full and more response-able relations. For example, paying attention to how somatic sensibilities can offer insights into how to respond to the suffering or needs of a mammalian nonhuman animal may face fewer immediate challenges than to the needs of fish, with whom humans share fewer corporeal experiences. And, yet, when it comes to the constituent beings of the environment, it is of course very important not only to be attracted to caring for those most like us, or most charismatic, but to be aware of the need to be open to respond with care in diverse and contradictory circumstances. Thus, the attention to the praxis of care focuses attention on how ethical responses can be aligned with learning how to be affected and how to do the right thing. It is not a straightforward route
to prioritize attending to opportunities to learn how to respond differently, or to make the world different from how we have come to know it. Karen Barad (2007) argues that, depending on the “cuts” we necessarily make into the world to observe any phenomena, the world opens up to us differently or, in other words, what we learn about the world will change, depending on how we make the cut between the subject and the object. Thus the ethical emphasis here, as in Haraway, is on what we do rather than on our capacity to be self-reflective about ethical principles. For example, new understanding about animal lives has led to a legal recognition of animal sentience – that is, that animals have emotional experiences in response to the environment they are in. Interestingly, the experiences of sentience of farm animals does not register only in their living bodies but also affects how their bodies can be used as meat. Meat science now recognizes that the treatment of a farmed animal prior to slaughter directly affects the material quality of the animal’s flesh, now meat, postslaughter. Following Barad’s thinking, what is happening here is an interaction between how the body registers stress responses in pigs, and how the physiology of the pig’s body responds differently if it is stressed at the point when oxygen leaves the living breathing body through the process of dying and becoming dead. However, the effect of this process on the actual flesh of the pig has arguably always occurred, but, in the contemporary age of concern for animal welfare, knowledge and awareness of animal welfare supports the interpretation of the material change in the pig’s flesh in a new way, and in the process articulates another way to make an ethical relation between human and nonhuman.

Looking toward the future and continuing the work to develop nonanthropocentric approaches, there are some notable fields of work that use participatory methodological approaches for working with nonhumans of diverse kinds. Participatory approaches have developed in geography in relation to working with human individuals and communities to achieve social justice. By including nonhumans as constitutive members of the social, following Bruno Latour (1993), how can participatory approaches support social justice for the nonhuman world? In many respects this work is still in its early stages of innovating methodologies and theories appropriate for nonhuman participatory research practices. One participatory approach is to use the human imagination to explore what it is like to be a plant, an animal, or a river through the use of thick descriptions derived from scientific knowledges and lay knowledges and creative thought, to encourage humans to explore becoming cow-ness, or becoming river-ness as an embodied experiment of becoming. An advocate of this approach is Joana Formosinho (2014), who is exploring methodologies for encountering otherness and acquiring knowledge of things in themselves as a route to learning how to see and respond differently. A second approach is to try to engage in “conversation” with nonhumans. Michelle Bastian and colleagues have been pioneering this experimental approach More-than-Human Participatory Research. For example, this might include asking how the nonhuman world can be included within our ethical protocols, thus questions we might ask include: How do you ask or seek consent from a river to change their management plan? How do we invent ethical protocols that encourage asking questions differently to take into account the interests, capacities, and potential disruptive elements of the nonhuman world in environmental governance practices?

SEE ALSO: Animal geographies; Commodification of nature; Construction of
nature; Cultures of nature; Environment and consumption; Environmental governance; Nonrepresentational theory; Poststructuralism/poststructural geographies; Production of nature

References


Further reading


Environmental futures

Geographers from Alexander von Humboldt onwards have played leading roles in developing scientific methods of observation, data collection, and deriving and testing hypotheses that have proved effective in increasing understanding of environments past and present. However, key questions preoccupying contemporary society also relate to what our past and current actions might mean for future environments and generations. Explicit attention to environmental futures is therefore becoming an increasingly familiar practice in research and policymaking. Thinking critically about environmental futures is not, however, a recent phenomenon. There is a long and diverse history of literary, artistic, and scientific work that provides the foundation for current thinking about environmental futures. While based on different drivers and motivations, adopting divergent methods and approaches, and culminating in distinct and not necessarily comparable outputs, the broad field of environmental futures research nonetheless encompasses core geographical concerns including matters of scale and place and most fundamentally of nature–society interactions.

Much of the historic literary work in the field of environmental futures is focused on all-encompassing visions or blueprints of sustainable societies, often gathered together under a banner of the “utopian tradition.” The utopian tradition can be traced back to Plato, through the political philosophies of Thomas More (Utopia, 1516), to William Morris (News from Nowhere, 1891) and on to the work of anarchist geographer Peter Kropotkin (The Conquest of Bread, 1892, and Fields, Factories and Workshops, 1904) and planning visionary Ebenezer Howard (Garden Cities of Tomorrow, 1902). At the same time, dystopian visions of future environments have also been prominent features of politico-literary texts, as seen in the work of Aldous Huxley (Brave New World, 1932) and George Orwell (Nineteen Eighty-Four, 1949) and increasingly through contemporary science fiction novels, films, and computer gaming. These visions, both utopian and dystopian, have been widely critiqued and categorized according to the proposed solutions for society’s ills, be that extended state control or anarchy, private or public ownership, and the degree of democracy and equality. Visions explicitly concentrating on environmental futures emphasize matters of resource use and consumption, typically focusing on problems of scarcity and the search for some attainment of “the good life.” Yet solutions remain diverse, some focusing on sufficiency and limiting consumption based on basic needs, others placing great faith in technological progress to enable continued production and consumption of goods and services.

Envisioning environmental futures is not just the preserve of utopian/dystopian literary authors. Quotidian processes of land-use planning, for example, frequently embody the creation of plans for how neighborhoods, cities, regions, and nations might evolve in the future. Yet these plans often depict near-term futures
ENVIRONMENTAL FUTURES

(around ten years in most cases), and are constrained by pragmatic resource limitations and the realpolitik of policymaking. Yet, as Marius de Geus (1999) argues, imaginative visions for future societies as produced by the utopian authors may work usefully together with planning – at local, national, regional, and global scales – as a means to provoke discussion about longer-term developments, providing alternative ideas, perspectives, and questions about decisions to be made and resources to be allocated.

Modeling global environmental futures:
trends, forecasting, and scientization

In April 1968 the Club of Rome, an international think tank, commissioned a team of system scientists from the Massachusetts Institute of Technology (MIT) to model possible future environmental scenarios based on five factors seen as limiting growth: population, agricultural production, natural resources, industrial production, and pollution. The model predicted global economic collapse and ultimately population decline by 2030 if trends in production and consumption were to continue. Other scenarios identified possibilities for economic growth, but only if systems of governance were developed that both regulated resource consumption and invested in technologies to control humanity’s ecological footprint. The research was published in 1972 as “Limits to Growth,” and while heavily critiqued for its assumptions and limited data, it had international impact selling more than 12 million copies in more than 30 languages. It was followed in the same year by “A Blueprint for Survival” by Edward Goldsmith and colleagues (1972). This document set out a manifesto for addressing the challenges identified in “Limits to Growth” and included supporting signatures from 36 experts from biology, chemistry, genetics, natural history, and medicine as well as two British geographers, S.R. (Bob) Eyre from Leeds University and George Melvyn Howe from Strathclyde University. Central to the authors’ vision for a survivable future were radical shifts in the ways society, as experienced within Western industrial countries at the time, would function. The vision was to create stability through decentralized, self-governing, and self-sufficient communities where diversity rather than uniformity could be fostered. While technology was perceived to play a role in this stabilized future it would only do so when the full cost of its development and use was proven to be less than the benefits it provided. Likewise, the Blueprint focused on the recognition of what was termed “real value” rather than economic value, with an emphasis on quality rather than quantity in design, and a reclamation of value for skilled craftsmanship and the arts. However, how such transitions to the stable, survivable society envisaged by Goldsmith and colleagues were to be enacted were not outlined.

Many of the concerns of the Club of Rome regarding an impoverished environmental future, biodiversity loss, environmental pollution, and unsustainable consumption, remain concerns for global environmental governance institutions in the twenty-first century. Indeed, little progress has been made in redirecting the trajectory of trends identified by the MIT scientists back in the late 1960s. There has been significant progress, however, in terms of the computing capability of the program used to model possible environmental futures, the quality and quantity of data to underpin those models, and the transnational collaboration between experts to refine both data and modeling attributes. A frontier activity in this regard is following the global emissions scenarios of the Intergovernmental Panel on Climate Change (IPCC). The IPCC was established to assess scientific, technical,
and socioeconomic information relevant to the understanding of humanly induced climate change, potential impacts of climate change, and options for mitigation and adaptation. In 1992, the IPCC released the first global scenarios to provide estimates for the full suite of greenhouse gases (GHG), but by 1996 an even broader set of emission scenarios were developed. The new scenarios included improved emission baselines and updated information on economic restructuring globally, incorporating rates and trends in technological change and expanding the diversity of the economic-development pathways considered. Modeling these broader parameters indicated that different social, economic, and technological developments strongly influence emissions trends, thereby providing significant insights into the interlinkages between environmental quality and development. Following the rules and procedures of the IPCC, multidisciplinary core writing teams, including geographers, draft reports which are then sent to international experts and member governments of the IPCC. For example, more than 2400 experts were formally invited to review the 2007 IPCC Synthesis Report and reviewers included geographers Thomas Spencer from Cambridge University, UK, and Georg Kaser, Institut für Geographie at the University of Innsbruck, Austria.

During the following decade, other international organizations followed the IPCC in its environmental futures activities. The World Water Commission produced their “World Water Vision,” while agricultural scenarios were developed by the International Assessment of Agricultural Science and Technology for Development, and scenarios of global ecosystems services resulted from the work of the Millennium Ecosystem Assessment. Meanwhile, transectoral environmental scenarios have also been produced by both the Organisation for Economic Co-operation and Development (OECD) and the United Nations Environment Programme (UNEP). The analysis for the OECD Environmental Outlook combines economic and environmental modeling frameworks developed at the OECD and the Netherlands Environmental Assessment Agency (PBL). Specifically, a global dynamic computable general equilibrium model “ENV-Linkages” is used to describe how economic activities are linked between sectors and across regions. It also links economic activity to environmental pressures, specifically to emissions of GHG. These links are then projected several decades into the future, in an attempt to illuminate potential impacts of environmental policies. Outputs from ENV-Linkages are combined with the IMAGE (Integrated Model to Assess the Global Environment) models run by PBL. A dynamic integrated assessment framework to model global change, IMAGE is underpinned by modeling of global land allocation, subject to production of food, feed, timber and water, and projects emissions, mapped onto a grid of the world. IMAGE has been used for other key global environmental assessments such as the Global Environmental Outlook (GEO) by UNEP. In contrast to the OECD approach, GEO aims to be a more consultative, participatory process focused on building capacity for conducting integrated assessments reporting on the state, trends, and outlooks of the environment. Based on integrated assessment and reporting approaches developed by a team of authors, including geographer Jill Jäger, the UNEP GEO approach also aims to make its reports scientifically credible and policy relevant, supporting multi-stakeholder networking and intra- and interregional cooperation to identify and assess key priority environmental issues at the regional levels. To this end a worldwide network of partners has been developed which
allows governments and other stakeholders to nominate experts to the process, advisory groups to provide guidance on scientific and policy issues, and a comprehensive peer review process.

However, while there have been increased efforts to ensure greater interdisciplinary input into models of environmental futures, academics including geographers David Demeritt and Michael Hulme, have demonstrated that they remain predominantly informed by natural rather than the social science research (Hulme and Mahony 2010). As well as marginalizing attention to important social components of global environmental change, this has also meant less critical attention to the philosophical and methodological aspects of prediction and forecasting. That there remains gender and spatial bias in terms of the inputs to many of these models also suggests that the visions of environmental futures portrayed are both narrowly conceived and highly bounded; there are what Richard Powell (2007) calls geographies of science even in the modeling of big data with respect to global environmental change. Equally, as with the Club of Rome, such modeling provides little assistance in terms of imagining how society might develop governing pathways to manage and construct less destructive environmental futures.

**From forecasting to backcasting:**

**the participatory turn in environmental futures research**

Within much standard environmental modeling work, natural phenomena are assumed to behave in the same way, whenever and wherever they occur. This assumption extends also to natural processes that represent a departure from long-term average conditions, and to nonlinear processes, such as a collapse of the thermohaline circulation (Rahmstorf 2000). However, it is argued that this assumption of continuity cannot extend to the social sciences where innovation and discontinuity have been identified as key features in the development of social systems. While many significant social structures, processes, and norms may remain relatively stable for periods of time, it is also the case that they can and do evolve, either suddenly or gradually, as a familiar outcome of economic and social activity. So, in contrast to the forecasting approaches adopted by the IPCC, OECD, and UNEP reports where past trends and current “rules” are iterated into the future, a body of work collectively called backcasting has evolved that explicitly adopts a normative scenario building approach (Dreborg 1996). The process is shaped around the question, “where do we want to go and what actions do we need to take in order to get there?” Essentially, backcasting assumes that futures can be created. This is its relative strength when compared with forecasting approaches although there are also dangers of overstating the capacity of actors to influence the future and clearly not all actors have equal influence in terms of shaping change. In response to these concerns, a third way between forecasting and backcasting has been framed as exploratory scenario building. In this context a range of underlying socioeconomic conditions are used to generate alternative futures as a means of mapping “possibility spaces.” Within this exploratory framing, the future is imagined through the development of multiple alternative states in which social agents can respond according to their own interests albeit with limited control. It is the emphasis on the importance of adaptation that is a defining feature within this arena of environmental futures work.

Of particular importance in many of the backcasting activities are the interlinked features of expanded participation, innovation, and creativity. Including diverse interests and
expertise and explicitly focusing on a future that is far enough away to free participants from the constraints of everyday living, but close enough to imagine themselves or their children inhabiting, is thought to offer more opportunities for combining knowledge, stimulating creative thinking, and developing innovative visions. Such transdisciplinarity in these normative environmental futures exercises is seen as essential to address the complexity of environmental challenges now and into the future. Backcasting scenarios are not theoretically based or empirically validated models, but heuristic tools to enable experts and others to engage in a collective process of deliberation. For example, the generated scenarios are frequently accompanied by a narrative, and sometimes even visualized, storylines elucidating the imagined futures. As such their value is seen to lie predominantly in their capacity to bring together disparate actors and agencies in a common process of envisaging desirable as well as possible futures; creating spaces for sustainability learning (Davies, Doyle, and Pape 2012). The development of scenarios is thus a process through which imagined futures can be revealed, ordered, and analyzed although uneven patterns of power and participation in those visions remain.

Collective visioning experiments have been undertaken by different environmental sectors from energy and city planning, to water resource management and sustainable consumption and at different scales, from the household scale to national planning fora. The Netherlands in particular has adopted backcasting as a feature of its strategic environmental decision making with visioning applied within the main Dutch research organization for agriculture and rural areas. Elsewhere experiments with participatory backcasting for supporting sustainable environmental futures have been undertaken in Sweden, Germany, New Zealand, and Canada.

Environmental futures: themes and roles

It is clear that thinking about environmental futures has a diverse lineage and continues to be practiced in many ways. Environmental futures activities can be constructed by individual visionaries or transdisciplinary collectives and can address different scales, from the modeling of global futures produced by the IPCC to the small-scale utopian ecological communities crafted by Peter Kropotkin. Such imagining occurs both in terms of near environmental futures, the ten year development plan within a local government district for example, as well as in terms of the far futures of science fiction, way beyond the lifespan of current populations. While literary visions of environmental futures have tended to be characterized as utopian or dystopian, the scientization of thinking through environmental futures sought to create positive rather than normative fields of possibility for environmental futures. Yet, even these increasingly sophisticated scenarios based on complex models map only certain dimensions of those futures, for example, predicting weather patterns or rainfall frequency and intensity based on lessons from the past and the present to estimate future patterns. How those altered environmental futures might affect the ways in which people live across the globe and under what governing arrangements is not addressed. Yet, environmental change scenarios involve much more than the scientifically based models that underpin them. Scenarios, whether they are developed by the IPCC or any other organization involve processes of selection and negotiation between different interests and stakeholders. For as Michael Hulme and Suraje Dessai (2008) outline, fundamental choices are made about the types of models used, the ways in which uncertainty is assessed and communicated, and the development pathways and
emissions scenarios on which to focus. So, while there is widespread skepticism of the utopian grand plans of individual visionaries or the singular blueprint of the future proposed by elite groups such as the Club of Rome, even the most scientifically robust vision of environmental futures are constructed following debate and contestation. Recognizing such characteristics does not diminish the usefulness of scenarios, as heuristic, motivational, and informative tools for design, planning, and policy development. Environmental futures experiments of all kinds provide spaces to debate important questions about the kinds of environments we might wish to inhabit in the future. While clearly multidisciplinary endeavors, the fundamental attention to nature–society interactions at the very roots of imagining environmental futures means that geographers (both human and physical) have a key role to play within them.

SEE ALSO: Big data; Environmental change and social learning; Environmentalism; Global climate change; Global climate models; Global environmental change: human dimensions; Imaginative geographies; Participatory modeling; Sustainability science

References


Further reading

Scholars have noted an incremental shift over the past few decades in the way that the environment is governed. While originally the preserve of local and then national governments, environmental governance (as distinct from “government”) now encompasses the activities of multiple actors, including states as well as also industry, nongovernmental organizations (NGOs), and the public. Indeed, the term environmental governance has been coined by scholars to denote such patterns of interaction that derive from the strategies of these various actors to govern or “steer” society towards achieving environmentally-related objectives (Adger and Jordan 2009). States and/or governments have traditionally used regulation and the courts to govern the environment in a relatively top-down fashion, but in the last two decades non-state actors, including businesses and communities, have become much more directly involved in governing, necessitating the use of other approaches that rely on market forces and/or networks of local actors.

At first, the literature on governance was rather disjointed, but by the 2000s it had consolidated around three core perspectives: empirical (i.e., describing how society is governed and the changing form and role of the state in contemporary industrialized societies); theoretical (i.e., seeking to explain how it is governed and predict future changes); and normative (i.e., indicating how society should be governed and the different forms of environmental governance required to solve complex, cross-scale and cross-societal issues such as climate change and sustainable development). Normative in this sense refers to how governance should be organized in practice. One of the most well-known formulations is provided by the World Bank, which publishes good governance principles for implementation by developing countries and also a wide array of other public and private bodies. In this context, “good” refers to a set of normative standards against which governance can be benchmarked.

A short history of environmental governance

Environmental problems are as old as humanity, along with societal attempts to address them. One of the earliest recorded environmental laws was a proclamation in 1272 by Edward I of England banning the burning of sea coal in London to prevent smoke pollution. Air quality issues then became progressively more important during the Industrial Revolution. Noxious gases from the production of chemicals in Victorian Britain led to the introduction of the national Alkali Act 1863, along with a government inspectorate to reduce emissions. However, government intervention did nothing to combat chronic air pollution in Britain’s rapidly growing cities, with
abatement measures, mostly regulatory in nature, primarily left to local authorities. This situation was replicated in other countries such as the United States, where cities such as Chicago and New York introduced their own clean air controls in the 1800s. Concerns over polluted drinking water, sanitation, and waste management then became salient public health issues in many other countries during the late nineteenth century. Again, governance was invariably predicated on action being taken at local scales. For example, in the United Kingdom, municipal authorities assumed most responsibilities for the supply of drinking water and sanitation, while catchment drainage and fisheries boards performed related management functions such as land drainage.

Thereafter, a wave of limited centralization spread across the western world, with central governments becoming gradually involved in environmental management in the early twentieth century. Six national parks had already been created by the US Congress before 1900, and a federally funded Soil Conservation Service was then established to reduce agricultural erosion. The creation of the Tennessee Valley Authority by Congress in 1933 became the prototype for integrated water management systems worldwide (Benson, Jordan, and Smith 2013). In the United Kingdom, the centralization of water management started in the 1940s, via the national Rivers Board Act 1948. After a succession of severe smog incidents in London, Parliament introduced the Clean Air Act (1956) that significantly reduced the burning of coal in the city. But most environmental tasks remained the preserve of subnational governments in many countries well into the 1960s. And, crucially for those seeking to understand the changing governance of the environment, legislation and other forms of regulation remained the chief instruments of environmental policy, and very much the responsibility of national governments.

Environmental decision-making in the late 1960s and 1970s became increasingly dominated by central governments. Across the developed world, chronic air and water pollution problems became coupled with mounting threats from toxic waste, species decline, habitat loss, and damage to marine environments. The transboundary nature of many of these problems meant that the local level actors responsible for them found it difficult to govern effectively, thus providing a strong rationale for the centralization of many environmental governance tasks. This process was first visible in the United States, where the inability or unwillingness of state governments to intervene in solving pollution issues, despite some federal clean air and water legislation, led to a shift in the national public consciousness, which then precipitated demands for government intervention (Andrews 2006). Adoption of the landmark National Environmental Policy Act (NEPA) by the US Congress in 1969 for the first time anywhere established environmental protection as an objective of a central government. Other major federal legislation on air pollution, water quality, species protection, and toxic waste clean-up followed quickly afterwards. The federal Environmental Protection Agency (EPA) was also established to implement these measures. Taking a lead from the United States, many national governments worldwide then acted to introduce environmental legislation and implementing institutions. During this era of expanding internationalization, national governments also cooperated multilaterally to adopt several major environmental regimes or international agreements, such as the 1973 Convention for the Prevention of Pollution from Ships (MARPOL) and the 1979 UN Convention on Long-range Transboundary Air Pollution.

The shift towards state-led environmental government was then followed by moves towards forms of governing that were less reliant on the
introduction of new legislation. During the 1980s, a regulatory backlash occurred in the United States, United Kingdom and other countries, reflecting wider ideological and discursive shifts globally towards a more neoliberal political agenda. More decisions were taken after comparing economic costs and benefits, rather than following political demands for legal intervention (Jordan and Turnpenny 2015). This discursive shift promoted government deregulation and a greater role for markets in governing environmental problems. In the United States, for example, the Reagan presidential administration sought to undermine the federal environmental protection structure, primarily through defunding the EPA and attempting to decentralize its powers back to the state level (Andrews 2006).

Governments began to share even more powers with nonstate actors after the emergence of sustainable development as a political discourse, following the United Nations’ Brundtland Report (WCED 1987). Sustainable development recognizes that environmental, social, and economic issues need to be addressed holistically. This notion underpinned the UN 1992 Rio Earth Summit, which implored national governments to introduce wider ranging and nonlegal strategies for environmental governance. Agenda 21, the main implementing agreement of Rio, also envisaged new roles for nonstate actors, such as businesses, local governments, and communities, in governing for sustainability. These principles, to varying degrees, became integrated into environmental decision-making by different governments. In particular, they found a receptive audience in “greener” European countries, such as Germany, Denmark, and the Netherlands, via the discourse of ecological modernization. Premised on the notion that economies can be reformed along more ecologically-friendly lines, this approach involves government intervention (i.e., “steering”) to empower societal actors such as consumers and businesses in promoting environmental protection, often through the creation of new markets and technological innovation. Less regulatory or even nonregulatory ways of environmental protection (such as environmental taxation and emissions trading) were actively promoted by economists. Their views encouraged many states to adopt what came to be known as “new” environmental policy instruments (Wurzel, Zito, and Jordan 2013).

At the same time, many countries, such as the United States, Canada, and Australia, witnessed moves towards more collaborative and/or participatory forms of environmental governance based at the local or ecosystem level, involving state and nonstate actors cooperatively managing resources such as watersheds, forests, and oceans (Benson, Jordan, and Smith 2013). “Co-management,” or community-based environmental management, also found favor in developing countries, through involving localized stakeholders in government-led efforts to protect environmental resources. New policy formulation tools were developed to provide more anticipatory and participatory responses (Jordan and Turnpenny 2015). During the 2000s, the transition from national or central government-led environmental protection towards more governance-based forms continued. Intensifying globalization of socioeconomic processes and ongoing trade liberalization meant that governments were less and less able to deliver sustainable development on their own. Moreover, the increasing complexity and cross-scale characteristics of environmental problems, such as climate change or biodiversity loss, necessitated the involvement of multiple actors at different institutional scales. The follow-up conference to Rio, in Johannesburg in 2002 (the UN Rio +10), was, therefore, dominated by commercial and NGO interests, rather than by states. Public–private
partnerships, the main policy outcome of the conference, reflected the relative disengagement of governments in the sustainability agenda and also the greater emphasis being placed by this point in time on alternative governance forms at the global level.

As a result of these developments, the current situation is perhaps best encapsulated by the concept of multilevel environmental governance, in which governing occurs through many actors. The locus of environmental decision-making has expanded, vertically and horizontally, from local or subnational levels prior to the late 1960s to national and international scales in the early twenty-first century. At the global level, environmental governance now takes the form of multifarious regimes agreed between states – many brokered by the United Nations – that seek to address transnational issues as diverse as marine pollution, ozone depletion, deforestation, biodiversity loss, nuclear threats, and climate change. Regional scale initiatives are also prevalent, most notably in the European Union (EU), which, over several decades, has developed a broad-ranging body of environmental policy that now determines action in some 28 Member States (Jordan and Adelle 2012). Nonetheless, national governments arguably remain the most significant actors in managing the environment globally, although they share responsibilities with supranational bodies, subnational and local governments, businesses, NGOs, civil society, and communities in the pursuit of environmental objectives. One defining feature of this multilevel era is also the expansion of bottom-up, community-based initiatives in many states, often introduced reactively in response to the failure of more top-down and centralized approaches of environmental government of the past. Another characteristic, in the post-2008 “age of austerity,” has been the relative downgrading of the environment as a policy issue in national government policy as economic priorities have taken precedent (Russel and Benson 2014).

The main features of environmental governance

Research into environmental governance reflects this evolution through time, with geographers amongst others detecting changes in the focus, modes, and means of governing. First, scholars employ the term governmentality to denote the way in which power is exercised by the state (Evans 2012). Here, what became widely perceived as localized problems in the Victorian era, soliciting subnational institutional responses, gradually changed over the twentieth century as knowledge about the transboundary and socioeconomic effects of environmental problems increased. Concerted government intervention in the late 1960s and 1970s constituted a new knowledge system, which prioritized greater centralized state powers and regulatory solutions. In the 1980s, the adoption of more neoliberal approaches meant governments relied much less on regulation. Although sustainable development became an important feature of government approaches in the 1990s, by the late 2000s the governmentality of problems altered again, as national economic priorities became more pressing than environmental or developmental ones in the wake of the global recession.

Second, another feature of environmental governance is the increasing interconnection across hierarchies, markets, actor networks, and communities in developing and implementing policy (Evans 2012). As with multilevel governance, this involves the dispersal of powers formerly the preserve of the state in favor of government interaction within a nested hierarchy of multiple institutional levels or scales. For example, the governance of climate change now
ENVIRONMENTAL GOVERNANCE

Involves actors at the international level (the United Nations), regional supranational entities (the EU), national governments, subnational and local governments (Connelly et al. 2012).

Governance also now features many economic actors, with markets emerging as a significant mechanism for achieving environmental objectives, despite economic pressures often being the root cause of many problems. According to this logic, if markets can be harnessed properly they can provide powerful incentives to producers and consumers to behave in environmentally friendly ways. Here, markets can help “internalize” the negative externalities or impacts of economic activities, such as pollution, by placing a value on them. A well-known example is the establishment of emissions trading systems in the European Union and elsewhere that force industrial polluters to purchase permits to emit carbon dioxide, thus encouraging them to reduce emissions in order to save money. One of the features of environmental governance is, therefore, the prevalence of market-based instruments (or MBIs for short), as discussed below.

Third, environmental governance is closely associated with a growth in networks that span horizontal and vertical scales of interaction, and are sometimes composed entirely of nonstate actors. Such networks can operate within nation-state contexts or be transnational in nature, such as an “epistemic community” (Haas 1990), comprised of scientists and technical experts, bound by common belief systems concerning specific issues and often formed for advocacy and initiating policy change. Within environmental governance, examples could include the Union of Concerned Scientists, the Intergovernmental Panel on Climate Change (IPCC), and the Stockholm Environment Institute.

Another type of network operates transnationally (Bulkeley et al. 2012), in a more horizontal form that connects local actors across an international space in order to exchange knowledge and engage in learning about solving environmental problems. Such networks may be “self-organizing” to the extent that they are not necessarily reliant on central government steering or resources for their operation. One widely-known example is ICLEI, a global network of cities, regions, and towns devoted to transferring knowledge on local sustainability best practice. Other networks operate across more vertical scales, linking levels such as the international, national, and local.

As identified already, environmental governance has also involved greater roles for the public in decision-making. Agenda 21 strongly promoted citizen participation in sustainability, as did the UN Aarhus Convention in 1998. Public engagement by governments is thus encouraged via processes such as environmental impact assessment (EIA) in land use planning and strategic environmental assessment (SEA) of policies. Collaborative environmental governance is also a growing feature worldwide, particularly within integrated water resources and coastal zone management.

Finally (and as noted previously), environmental governance is itself associated with and constituted by the emergence of “new modes of governance,” such as taxes, voluntary arrangements, and information provision (Wurzel, Zito, and Jordan 2013). Early government intervention in managing environmental problems during the 1970s and 1980s was characterized by so-called “command and control” regulations, typified by US legislative measures such as the federal Clean Air and Clean Water Acts. Such mechanisms impose specific targets or objectives that societal actors must meet, with implementation often achieved through punishing noncompliance. The European Union has also introduced many regulatory measures as part of its wider environmental policy. These measures can be contrasted with market-based instruments that impose costs on environmentally damaging actions,
but which also establish economic incentives for actors to modify their behavior. Examples include environmental taxation, which levies charges on actors for the right to pollute and also subsidies for encouraging environmentally friendly approaches. In the United Kingdom, each ton of waste sent for disposal in landfill is subject to a tax, with some of the revenues dedicated to supporting local environmental projects. Meanwhile, the United Nations has introduced mechanisms such as the reducing emissions from deforestation and degradation (REDD) program that employs market-based instruments to support forest protection in developing countries.

As an alternative to the market, informational instruments provide information to influence environmentally friendly behavior through extending consumer choice. One well-known example comes from Germany and Austria, where governments have introduced eco-labels for some consumer products that provide details on their environmental performance. Another is the Forest Stewardship Council self-certification approach for labeling sustainably produced timber, in what has been called “nonstate market-driven” governance (Cashore et al. 2007). Environmental management and audit systems also allow businesses to assess the impacts of production cycles and hence reduce them.

Another “new mode” of environmental governance is the increasing use of voluntary agreements between governments and “target” groups such as industry. Agreements can either be self-imposed via declarations of intent by social actors to act in environmentally responsible ways, or they can be negotiated with governments. An example of the latter could include covenants, or formal legal contracts, struck between the Dutch government and business groups in the 1990s. Such “new modes” could additionally include the creation of individual institutions to achieve specific environmental objectives. For example, the global diffusion of integrated water resources management (IWRM) norms has led to the establishment of task-specific governance institutions at the river basin scale, such as the river basin districts introduced to implement the EU Water Framework Directive (Benson, Jordan, and Smith 2013).

**Theorizing environmental governance**

As well as describing the changes outlined so far, scholars have sought to use the term governance in a more theoretical manner, in attempting to explain some of these empirical patterns. For example, scholars have employed theoretical arguments to understand governance by focusing on different networks of actors, both in the European Union and in national contexts (Schout and Jordan 2005). Another example comes from the work of the OECD, which employs a particular network-based mode of “good governance” known as the “OECD technique” (Lehtonen 2007). This technique involves states benchmarking or reporting their implementation of environmental policy against OECD performance reviews and checklists. They then share best practice with other states in order to improve implementation. Other theories draw upon the literature on public participation to explain the collaboration of different actors in the management of resources such as river basins, forests, and coastal zones (Benson, Jordan, and Smith 2013). One argument made for greater collaboration by theorists is that it supports information sharing, trust, and learning amongst actors, which helps environmental governance become more adaptive and resilient in response to rapidly changing threats such as climate change. Finally, some political geographers refer to the ongoing neoliberalization of environmental governance.
to explain the growth in market-based mechanisms and deregulatory processes (Benson, Jordan, and Smith 2013).

Despite many attempts to test different types of theory, there is still no single “theory of environmental governance.” The vast majority of accounts of how the environment and sustainable development has been governed make little or no reference to what currently passes for “governance theory.” Mostly these studies import what theory or theories they need from the broader field of social science research, with a pronounced emphasis on those that feature networks, learning, collaboration, and voluntarism.

Environmental governance as a normative strategy

Environmental governance has undeniably also become somewhat of a normative strategy in itself. Although there are many arguments about how governance should be organized, there are five key areas of debate around: (i) global governance; (ii) market liberalism; (iii) the green economy; (iv) the green state; and (v) localized or decentralized models.

1 **Global governance.** Some authors see global governance as an antidote to the failings of state-led environmental government (Biermann 2006). Since the early 1970s, states have cooperated to form environmental regimes – sets of norms and institutions – aimed at resolving different problems, with varying degrees of success. Evidence from successful environmental regimes, such as the 1987 Montreal Protocol on ozone depletion, would suggest that internationalization can provide solutions to collective-action environmental problems by increasing transparency and trust amongst state actors. Yet this success is rare; more often than not, state self-interest dominates, as demonstrated by the past impasse in the UN Kyoto Protocol process or the inability of the UN Forum on Forests to counter chronic global deforestation. Environmentalists, meanwhile, remain skeptical over whether nonstate actors such as multinational corporations (MNCs) should be included in global environmental governance, seeing them as part of the problem rather than the solution to greater sustainability (Connelly et al. 2012).

2 **Market liberalism.** From this perspective, sustainable development is best achieved through limited government, less regulation, and greater engagement with economic actors. This view places great faith in markets and economic actors to deliver environmental objectives through their ability to value environmental resources. But the combined record of market liberalism and international capitalism in promoting more sustainable forms of global economic development is hardly encouraging, primarily because issues such as environmental justice, equity, and resource limits are often sidelined in market-based approaches. Some environmental resources, such as the Amazon rainforest or the global climate, also remain difficult to value in pure monetary terms, meaning that they can be overexploited. In response, some authors, such as Newell and Paterson (2010), call for reformed market-based approaches to addressing global warming, that is, “climate capitalism,” by suggesting that rather than eschewing business, environmental governance should accommodate it better through new economic growth models.
ENVIRONMENTAL GOVERNANCE

3 **Green economy.** The notion of the green economy (Barbier and Markandya 2013) builds on earlier work that sought a “greening” of economies to make them more sustainable and promotes the use of economic instruments, valuing natural capital and providing more socially equitable solutions via green employment and poverty reduction. Here, governments act as central facilitators in setting economic policy working in conjunction with nonstate actors such as business and communities to achieve collective objectives. While still very much a normative perspective, green economy, green growth, or green capitalism ideas have nonetheless permeated institutional responses at national and international levels, although progress has to date been globally variable (Russel and Benson 2014).

4 **The Green State.** Another normative argument is that the state should itself be reformed to promote more ecologically friendly forms of development that involve different actors. This chimes with the idea (Connelly et al. 2012) that state-centered governments are paradoxical in the way that they simultaneously generate environmental problems (through support for economic growth) and seek to solve them (through environmental policies). Governments are important in this respect, as they alone retain sovereignty to act on environmental issues within their territories, despite also allocating some powers to nonstate actors. Hence, Robyn Eckersley (2004) has argued for a greener state, through more inclusive “ecological democracy” based on greater public participation and integration of environmental and social norms into decision-making.

5 **Localized governance.** Finally, some scholars have gone further still in outlining more localized models of governance, which seek to greatly decentralize existing state powers. Collaborative environmental governance could help reframe government–society interactions through creating new responsibilities for actors, alternative forms of democratic engagement, and innovative mechanisms for managing resources. Green commentators have also called for more consideration of justice in environmental governance by shunning globalization, demanding controls on corporations, and promoting localization (Connelly et al. 2012). This localization of governance, with its rejection of centralized state-led steering, has in places become a partial empirical reality, with community-based economic activity emerging, such as the Transition Towns movement in different countries. More radical approaches still demand the rejection of environmental government entirely, favoring forms of eco-anarchy based on bioregions, rather than geopolitical borders, and exclusive community control.

“Governance” is a vital aspect of all these debates, suggesting that, after 40 years of governing the environment, the eagerness amongst geographers to understand and discuss its empirical, theoretical, and normative dimensions is as great as ever.

SEE ALSO: Climate change, concept of; Community-based natural resource management; Ecological modernization; Environment, democracy, and public participation; Environment and the state; Environmental certification and eco-labeling; Environmental impact assessment; Environmental management; Environmental policy; Environmental regulation; Governmentality; Green capitalism;
Intergovernmental Panel on Climate Change (IPCC); Localization/delocalization; Multilevel governance; Neoliberalism; Power; Public policy; River basin management and development; Scale; Sustainable development

References

Environmental hazards

Ilan Kelman
University College London, UK
Norwegian Institute of International Affairs (NUPI), Norway

Environmental processes and phenomena are everywhere around planet Earth, affecting life and livelihoods, and sometimes leading to death and damage so they are seen as environmental hazards. That does not mean avoiding or preventing the hazards. Instead, society can use knowledge of environmental processes and phenomena to gain advantages from them while avoiding many hazardous aspects.

Definitions

Hazard

The word “hazard” tends to mean some form of potential danger, difficulty, or peril – actual or metaphorical. The word “risk” is often used in English to be synonymous with “hazard,” but in disaster studies the words are subtly different. “Risk” combines hazard and vulnerability into an expectation or probability of harm occurring. “Hazard” is a phenomenon or process with the potential to interact with vulnerability, thereby exposing the vulnerability. It is the vulnerability which causes harm in a disaster. Examples of environmental processes and phenomena with the potential to become hazards are earthquakes, lightning, comets, icebergs, floods, bacteria, and parasites.

Hazards are sometimes described as being foreseeable or unavoidable. Neither descriptor is necessarily correct in the context of disasters. Hazards can be unforeseeable if, within the bounds of knowledge, no attempt was made to examine them or the knowledge was not fully understood. For example, prior to the start of the 1991 cataclysmic eruption, it is unclear whether or not Mount Pinatubo in the Philippines had erupted before settlement of the area, so foreseeability of the 1991 eruption is disputable. In contrast, Soufriere Hills volcano on Montserrat appears not to have erupted since Europeans arrived in the Caribbean, but the volcano was studied with the hazards documented in scientific papers and reports to government. The eruption sequence starting in 1995 and its consequences were largely foreseeable. Regarding unavoidability, theories and realities of influencing environmental hazards are described later in this entry.

Characterizing environmental hazards uses parameters and descriptors including the following.

- A physical process explanation such as through motion, energy, or chemistry. Examples are acceleration (change in velocity, encompassing magnitude and direction) during seismic shaking, velocity (speed and direction) of a landslide followed by forces imparted by the landslide’s material, heat from lava, volumetric flow rate of floodwater, and mass of a meteorite or of locusts in a swarm.
- Magnitude, intensity, and scale values of specific parameters.
ENVIRONMENTAL HAZARDS

- Time characteristics, which could cover the time required for a hazard to emerge and decay, the duration during which the hazard threatens, the frequency with which a hazard manifests, and the time periods during which a hazard might be expected (for instance, specific months of the year for tropical cyclones for different locations).
- Space characteristics, describing the linear extents, areas, and volumes which the hazard affects along with distribution patterns over space.

All the above characteristics can be described in terms of predictability, quality of predictions, and what is known and unknown – all of which are essential parts of trying to understand and respond to environmental hazards.

Environmental

The environment tends to refer to the four “spheres” of nature, none of which is created by humans: the atmosphere (air), the hydrosphere (meaning water – salt, fresh, and frozen – such as oceans, rivers, lakes, glaciers, and ice sheets), the lithosphere (the Earth’s surface, such as soil and rock and everything below the surface, down to the core at the planet’s center), and the biosphere (plants, animals, and ecosystems). The environment outside planet Earth could also be considered to be a sphere of nature. The spheres are rarely distinct, instead having blurry boundaries and extensive overlaps such as tidal zones. Although humanity and society are excluded from this strict definition, it is difficult to delineate a definitive boundary between the environment and society, with environment–society enmeshment being a starting axiom for many scientific disciplines.

As an example, human-induced alteration of the atmosphere’s composition is leading to rapid climate change affecting the planet’s surface, especially manifesting through changing weather and a changing biosphere. Similarly, many human-made chemicals persist in the environment for long periods, with chlorofluorocarbons staying in the stratosphere for decades and affecting natural ozone there. The loss of stratospheric ozone leads to increased natural radiation intensity from the sun reaching the Earth’s surface, potentially harming the biosphere. As such, the environment and society cannot be separated; they are heavily interconnected and continually influence each other.

Nevertheless, many hazards tend to originate in the spheres of nature, such as volcanic eruptions, precipitation, bacteria, and solar flares – even if sometimes affected by society such as climate change altering precipitation or genetic engineering producing bacteria. Other hazards clearly originate from society only, such as chemical releases, the stock market, vehicles, and vehicle drivers. These societal hazards are not considered in this piece.

Some argue that the environment never produces hazards, only processes and phenomena which can become hazardous due to human choices, actions, or inactions. When a ship strikes an iceberg, when a plane crashes after being hit by lightning, or when a petrochemical tank explodes after being struck by a landslide, it is possible to argue that the hazard originated in the environment, in society, or both.

If rain comes through a window ruining the carpet inside because the homeowner did not close the window, then the rain in itself is not necessarily a hazard although it has potential for hazardousness. The human decision made the rain a hazard because the rain damaged a human creation due only to a human decision. If a housing development is approved in a known earthquake zone without earthquake-resistance measures being factored in, then the earthquake itself is not necessarily the hazard although it has
potential for hazardousness. Instead, the human decision made the earthquake a hazard because the earthquake damaged a human creation owing only to a human decision. An environmental phenomenon or process can become a hazard to society because human decisions led to damage of a human creation.

The environmental processes/phenomena and the circumstances leading to manifestation of hazardousness are often distinct. A “chain of hazard emergence” could be postulated in which “environmental” refers to only the phenomenon or process from nature which might be modified by society. Then, it encounters a societal creation – such as a process, structure, person, or network – so that damage to society could result, at which point it becomes an “environmental hazard.”

The importance of the hazard emerging only upon interaction with society is that it is rare that environmental processes or phenomena are deemed to be hazardous to the environment itself. In ecology, a “disturbance” is a phenomenon or process which kills organisms or forces them out of their habitat. The disturbance opens up opportunities for other organisms to colonize the location. Windstorms, ice storms, and wildfires kill trees and knock off branches, sometimes with animal mortality as one consequence. But they are essential ecological processes, being part of the ecosystem and contributing to biospheric life cycles. Many tree species have evolved under the assumption that fires occur periodically, so those disturbances are part of the species’ reproductive cycle (Reice 2001). How would an environmentalist perspective view disaster impacts – and disaster risk reduction efforts to reduce those impacts – when the environmental phenomenon or process becoming a hazard for society is also essential for a healthy ecology?

The “chain of hazard emergence” helps to answer that question by explaining that societal interaction with an environmental process or phenomenon sometimes creates the environmental hazard. That societal interaction might be deliberate modification of the environmental process or phenomenon. Wildfires can be started by a cigarette butt or arson. For several decades in the twentieth century, the United States’ government managed “natural” forests to avoid fires. The policy changed following the realization that natural fires were inevitable, but suppressing fires in the short-term leads to larger and more intense fires in the future which could be outside the parameters within which species evolved. The fire is a “natural” environmental process, but human management transformed the specific parameters into non–natural fires which were an environmental hazard to both society and possibly the ecosystem.

Environmental hazard

An environmental hazard is a phenomenon or process with root origins in the environment which is interpreted by or interacts with society to pose a threat or danger to society. That is, it has the potential to damage or cause harm to society due to society’s actions. This definition gives environmental hazards a negative connotation, but threats and dangers to society can be opportunities, trivialities, or major concerns. The influences depend on the exact environment–society contexts and perspectives in which the environmental phenomenon or process occurs and is viewed.

Russia has been saved at least three times from invading armies by the winter weather contributing to the invaders’ defeat, although that does not diminish the military mistakes made by the invaders and the military strategies of the defenders. Examples of invading
leaders are Charles XII of Sweden from 1708 to 1709, Napoleon Bonaparte at the end of 1812, and Adolf Hitler from 1941 to 1943. The Russian defenders used the winter weather to their advantage by implementing a scorched Earth policy as they retreated. Meanwhile, the victory of the English navy in the Camperdown Campaign in 1797 has been attributed as much to weather as to military tactics (Wheeler 1991) which is the same for the Spanish Armada’s attempted invasion of England in 1588. In all these cases, one side of the battle saw the weather as being an environmental hazard while others saw it as being an environmental boon and opportunity.

On April 14–15, 1912, in the northwest Atlantic Ocean, the ship Titanic struck an iceberg and sank with over 1500 passengers and crew dying. It is suggested that the captain knew that icebergs were in the area but chose to travel at a fast speed, in addition to the ship being constructed of low-quality steel and not having enough lifeboats for everyone on board (Smith 1994). The iceberg’s hazardousness emerged due to only the long “chain of hazard emergence” involving the environmental phenomenon of the iceberg coinciding with human decision-making from design through to operation of the ship, producing iceberg forces on the ship which the ship could not withstand.

Consequently, the labeling of an environmental hazard can be subjective, because (i) the hazardousness can depend on the viewpoint and (ii) the hazardousness might emerge due to human action. While these points do not apply to every case, when analyzing or labeling an environmental hazard, it is important to consider how much these points do apply and what humanity might be able to do in order to reduce the hazardousness to society of an environmental phenomenon or process.

Categories

Environmental hazards have numerous modes and levels of classification (IRDR 2014). Here, four groups are selected: astronomical, biological/ecological, geological, and hydro-meteorological.

Astronomical

Astronomical or extraterrestrial hazards are defined as hazards with origins in space, that is, hazards with extraplanetary or non-Earth origins.

A common form of astronomical hazard is geomagnetic storms, which are high-energy particles, such as flares emanating from the sun, that meet the Earth’s magnetic field. These are most often seen as beauty rather than hazard in the form of the aurora, most visible at high latitudes. They can also be highly disruptive to life on Earth, with geomagnetic storms damaging satellites, interfering with communications networks affecting everything from mobile phones to bank machines, inhibiting power networks leading to blackouts, and forcing airplanes to re-route in order to avoid high radiation levels.

Near-Earth objects are another prominent example of astronomical hazards. They are objects in space such as asteroids and comets which could collide with the Earth. Such collisions occur continually, but most objects burn up in the atmosphere, almost unnoticed. Frequently, we see larger objects burn up as meteors, poetically termed shooting stars or falling stars. When the object is large enough to make it through the atmosphere and hit the ground, wide-scale destruction can occur. Humanity has been fortunate thus far that human injuries and deaths from near-Earth objects are rare. There is speculation but no proof either way that many people were incinerated by the June 30, 1908,
Tunguska explosion in Siberia and by the Kaali meteorite strike on Saaremaa, Estonia, several thousand years ago. In addition, isolated claims appeared through the centuries of people being struck by celestial bodies.

No injuries or deaths from a near-Earth object were confirmed until February 15, 2013, when a meteor burned through the atmosphere over the Chelyabinsk region in Russia. The shockwave shattered glass in thousands of buildings injuring over 1200 people. This event might be behind concerted efforts afterwards to establish a global approach to near-Earth object monitoring, warning, and response, but many scientists had been trying for decades beforehand.

A large near-Earth object strike has the potential to be cataclysmic for the planet, with such strikes identified as likely causal factors in planet-wide mass extinction events including the death of the dinosaurs approximately 65 million years ago. Other potentially cataclysmic astronomical hazards for the planet which could threaten humanity if they occur within several light-years’ vicinity include a supernova – an exploding star – or a gamma ray burst which would release huge amounts of lethal radiation.

**Biological/ecological**

Biological or biospheric hazards are hazards with origins in living organisms, ecosystems, or other levels of the ecological hierarchy. These hazards tend to be listed as only biotic (living) without fully accepting abiotic (nonliving) biospheric hazards, an example of which would be wildfire (also called a forest fire and a bushfire) as an ecosystem process. Here, only biotic biological/ecological hazards are considered. Wildfire is classified as hydrometeorological, since this environmental process usually starts from lightning with dry weather conditions abetting it.

Microbial pathogens and parasites represent the main biological/ecological hazard, encompassing, for example, archaea, bacteria, prions, protists (including protozoa and some algae), rickettsia, and viruses. They provide numerous services to ecosystems, such as breaking down dead plants, and to individuals, such as bacteria in our gut aiding digestion. They also affect people detrimentally through diseases, some examples of which are the prion causing Creutzfeldt-Jakob disease, the bacterium Esherichia coli O157:H7 causing food poisoning, and HIV (Human Immunodeficiency Virus) causing AIDS (Acquired Immunodeficiency Syndrome).

Microbial pathogens and parasites are perhaps the most lethal environmental hazard, not just through epidemics and pandemics but also through regular death rates. More than 600 000 people a year die from malaria (WHO 2013) and up to one-third of that toll die annually from cholera. The black death, being different forms of plague caused by a bacterium carried by fleas which live on rats, killed perhaps 200 million people in the fourteenth century including one- to two-thirds of Europe’s population. Infectious diseases – including cholera, colds, measles, plague, smallpox, tuberculosis, and typhoid – brought by Europeans to the Americas played a major role in destroying indigenous populations and permitting Europeans to settle there, with some indigenous groups experiencing 100% mortality due to microbial pathogens (Alvarez 2014). The most widespread modern epidemic, HIV/AIDS, kills approximately two million people a year with infection rates of 15% in some sub-Saharan countries (Bongaarts, Pelletier, and Gerland 2011).

In addition to microbiological hazards, macrobiological hazards – namely plants and animals – can pose hazards to people and society through mechanisms including poison, crushing when falling on people or buildings, trampling,
ENVIRONMENTAL HAZARDS

constricting, and biting. Swarms of insects, such as locusts, can devastate crops. Some animals which commonly kill many people each year are alligators, ants, bees, dogs, elephants, scorpions, sharks, and snakes. Many more people die when vehicles crash into elk, kangaroos, and moose. Vegetation most commonly causes fatalities when people ingest toxic varieties or when a large plant, such as a tree, falls on people as often occurs during storms.

Geological

Geological or lithospheric hazards are hazards with origins in the Earth or on the Earth’s land surface. Earthquakes and volcanoes tend to be seen as representing geological hazards, but they have a wide scope beyond the stereotypes of the Earth’s surface rupturing in an earthquake and a mountain exploding for a volcano.

Many tremors occur hundreds of kilometers below the surface or along the ocean floor. Shaking can occur in any direction, so infrastructure must withstand combinations of horizontal and vertical shaking. Two distinct sets of earthquake waves strike, with the faster waves generally being weaker than the slower waves. Earthquakes have numerous associated hazards, most prominently tsunamis, which are a series of waves caused by sudden displacement of large amounts of water. An earthquake-generated tsunami can be more lethal than the original shaking. Tsunamis can also be generated by a near-Earth object landing in water, linking tsunamis with astronomical hazards, as well as by mass movements such as landslides. Mass movements have multiple origins (see later in this section) including earthquakes. When El Salvador was struck by two earthquakes on January 13 and February 13, 2001, approximately half the fatalities were attributed to landslides.

Volcanoes present a wide range of hazards across all states of matter and their combinations as well as massless hazards (Figure 1). Here is some terminology is explained.

Fumaroles: openings emitting gases.
Jökulhlaups: in a volcanic context, glacial melt floods when a volcano erupts underneath.
Lahars: mudflows, often from ash deposits.
Pyroclastic density currents: hot, fast flows of gas and ash.
Tephra: solid material ejected into the atmosphere, including rocks and ash.
Vog: fog and smog resulting from volcanic emissions of gases and aerosols.

Large volcanic eruptions sending ash high into the atmosphere can alter weather for several years afterwards. When Tambora, Indonesia, erupted in 1815 it affected weather worldwide, especially leading to a damp and cold summer in the northern hemisphere which allegedly led to Mary Shelley writing *Frankenstein*. The impact of volcanoes on climate is disputed apart from in extreme cases. An example of the latter is basaltic flood eruptions in which one or several fissures release lava and gases over vast areas – with the eruptions sometimes lasting millennia. They have been implicated as potential causal factors in mass extinctions.

Other geological hazards include erosion, ground subsidence, quicksand, sandstorms and dust storms, sinkholes, and slides and mass movements (such as landslides, rockslides, mudslides, and debris flows). Mass movements have different origins, including earthquakes, rainfall (linking them with hydrometeorological hazards), accumulation of too much material, and destabilization by human activity such as cutting a road through a slope.

Ice ages are a global, catastrophic geological hazard which has occurred many times during
ENVIRONMENTAL HAZARDS

Gas hazards:
- fumaroles and other emissions
- high air temperatures

Liquid hazards:
- jökulhlaups
- tsunamis

Solid hazards:
- debris and ejecta, including tephra
- slides

Gas/liquid hazards:
- aerosols
- vog

Liquid/solid hazards:
- lahars
- lava
- wet tephra

Gas/solid hazards:
- aerosols
- pyroclastic density currents

Massless hazards:
- air shocks
- fire
- ground deformation
- lightning

Figure 1 Volcanic hazards.

Massless hazards:
- ground deformation
- tremors

The Earth’s past but not during humanity’s reign. Ice ages can bring glaciers several hundred meters high into the locations of current-day cities such as Toronto and Moscow. The glaciers weigh down the land, pushing countries lower and reshaping topography, yet because so much water is frozen in the ice, ocean levels dip by dozens or hundreds of meters.

Hydrometeorological hazards, including atmospheric and hydrological/hydrospheric hazards, are hazards with origins in the water (hydrological) or air (meteorological), joined into a single category because they are frequently linked. Examples of hydrometeorological hazards are avalanches, droughts, floods, fog, glacial surges, icebergs, lightning, precipitation (with different forms including freezing rain, hail, ice, rain, sleet, and snow), storm surges, temperature extremes or fluctuations (cold and hot), tornadoes (which come in many varieties including waterspouts), tropical cyclones (referring to cyclones, hurricanes, and typhoons but not forgetting cyclonic storms outside of the tropics), waves, and wind.

Weather-related and climate-related hazards are often conflated and many languages do not differentiate between weather and climate. Nonetheless, English does make the distinction, with IPCC (2014, 5) stating “Climate in a narrow sense is usually defined as the average weather, or more rigorously, as the statistical description in terms of the mean and variability of relevant quantities over a period of time ranging from months to thousands or millions of years … Climate in a wider sense is the state, including a statistical description, of the climate system.” IPCC (2014) does not define “weather.” IPCC (2012, 114–115) explicitly treats weather and climate events as being the same, again without defining “weather,” while also stating that time scale differentiates them in that (somewhat circularly) “An extreme weather event is typically associated with changing weather patterns, that is, within time frames of less than a day to a few weeks [and an] extreme climate event happens on longer time scales.” (p. 117).

Converging dictionary definitions assist: weather describes or parameterizes the atmosphere’s state while climate is the usual weather or a statistical description of weather. Since the weather–climate differentiation is a linguistic-cultural construction which does not always make an operational difference in understanding and dealing with hydrometeorological hazards, that latter phrase is used here rather than categorizing the phenomena and processes as either climatological or meteorological. The lack of distinction supports inclusion of environmental processes and phenomena which can fall between the weather time scale of
Environmentally, hazards can last from minutes to weeks and the climate time scale of decades and more—such as droughts.

Droughts are effectively the absence of water, due to lack of precipitation or increased evapotranspiration, but most droughts result more from human decisions than from precipitation variation (Glantz 1976). Examples are extraction of too much water for drinking, hygiene including flush toilets, irrigation, manicured lawns, or watering non-native trees. When lack of precipitation or excessive evapotranspiration does produce a meteorological drought—so it is a true environmental phenomenon with the potential to become an environmental hazard—it can be described by absolute or relative metrics. An absolute drought metric is a specific parameter, such as rainfall over a time period. A relative drought metric is an index indicating the difference or deviation from a predefined “normal,” such as soil moisture availability.

With regards to environmental processes and phenomena involving too much water, storms can be brief and highly localized or, as with tropical and extra-tropical cyclones, can last days and affect thousands of square kilometers. Floods appear in numerous varieties, from rivers slowly rising over days to flash floods through ravines and canyons which can be devastating within minutes of an upstream cloudburst. Large-scale freshwater floods include Pakistan in 2010 during which approximately 20% of the land area flooded with some areas still underwater several months after the initial inundation and China in 1931 with death toll estimates exceeding one million people. Storm surges are low air pressure and winds across a stretch of water raising the sea level and reaching the coastline as a wall of water, the height of which can be exacerbated by coinciding with high astronomical tide. Hundreds of kilometers of coastline can be inundated over several hours.

Hot and cold temperatures can lead to significant numbers of fatalities which are difficult to tally. In Europe, during the summer of 2003, tens of thousands of people died during a heat wave, but no exact compilation of numbers could be completed because it is not always feasible to attribute an individual’s death directly to heat. Calculations are usually conducted by considering excess mortality, examining death rates above the standard rates and correlating that with heat-related measures, involving combinations of air temperature and humidity or associated metrics such as dew point (e.g., humidex). Aside from heat waves, outdoor laborers in the tropics experience heat-related illnesses, leading to high numbers of lost work hours as well as some deaths.

Similar issues emerge with respect to cold as an environmental hazard. Compiling the impacts is fraught with difficulty due to attribution problems, but studying mortality patterns reveals the high rate of excess mortality during cold spells. As with heat, it is not just the cold temperature which contributes to the hazard, but also relative humidity and wind (e.g., the wind chill factor) combining with cold temperatures.

The environmental hazard of temperature has two noteworthy aspects. First, the view of temperature effects as an environmental hazard has a cultural component. People who have grown up in tropical climates feel extremely uncomfortable in temperatures such as 10–15°C which are considered balmy in mid-latitude winters. Meanwhile, 30°C with high relative humidity is standard fare for the tropics, but can be dangerous for people not used to it who overexert themselves.

Second, thresholds in temperature in addition to extremes are important as environmental hazards. One significant threshold is water’s freezing point. Precipitation falling through air around water’s freezing point is far more
Table 1  Examples of overlaps and connections among environmental hazards.

<table>
<thead>
<tr>
<th>Environmental process or phenomenon</th>
<th>Mainly classified as</th>
<th>Could also be classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avalanches, icebergs, and glacial surges</td>
<td>Geological</td>
<td>Hydrometeorological because they are mainly water and are linked to the weather. Small avalanches called sluffs can be triggered by rising air temperature</td>
</tr>
<tr>
<td>Jökulhlaups (glacial floods)</td>
<td>Geological when triggered by a volcano erupting underneath</td>
<td>Hydrometeorological when glacial water becomes a flood</td>
</tr>
<tr>
<td>Ice ages</td>
<td>Geological even though they involve water</td>
<td>Astronomical because they have occurred due to the Earth's position in space</td>
</tr>
<tr>
<td>Dangers from extraterrestrials</td>
<td>Biological</td>
<td>Astronomical</td>
</tr>
</tbody>
</table>

hazardous for the unprepared than rain (above freezing) or snow (below freezing) because, around water’s freezing point, it can come down as ice, leading people and cars to slip while accumulating on and breaking trees and power lines. Additionally, water infiltrates into masonry and, if that freezes, it expands causing cracking and damage. The number of times that masonry experiences a freeze/thaw cycle dictates its lifetime more than specific air temperatures.

Cascading and coincident environmental hazards

The abovementioned categories have overlaps and connections (see Table 1) as seen through cascading environmental hazards and coincidental environmental hazards. Cascading environmental hazards occur in sequence, with one triggering the next. Tsunamis are representative because they never occur in isolation. There must be a triggering phenomenon or process such as an underwater volcanic eruption or earthquake, a landslide underwater or into the water, or a meteorite striking water. As an example of a series of cascading hazards, for a volcano near a coast, an explosive eruption would be accompanied by earthquakes which could trigger a landslide into the sea generating a tsunami. In 1980, Mt St Helens in Washington, USA, was ready to erupt with most volcanologists expecting a near-vertical explosion. Magma movement inside the volcano generated an earthquake that caused a large slope failure on one of the volcano’s flanks, relieving enough pressure for the magma to burst out in a lateral (sideways) eruption. A hypothetical sequence of cascading hazards would be a meteorite striking land leading to seismicity and landslides as well as sending tremendous amounts of dust into the stratosphere which influences weather and, potentially, climate.

Biological/ecological hazards can be influenced by the consequences of geological and hydrometeorological hazards through microbiological hazards emerging during or after a disaster. The direct link from disaster to epidemic is rare (Floret et al. 2006) but diseases which are already endemic can manifest further
ENVIRONMENTAL HAZARDS

if disaster response measures are inadequate. Increases in rates of dengue, hepatitis A, leptospirosis, malaria, typhoid, West Nile Fever, and yellow fever occur after poorly managed floods where those diseases were already present. One example where a long-absent disease re-emerged postdisaster was cholera appearing in Haiti for the first time in decades following the January 12, 2010, earthquake in the Port-au-Prince area. The disease was brought in by United Nations troops who were assisting with the recovery, but who came from a country where cholera is endemic and they did not enact proper sanitation and hygiene methods in Haiti.

Coincidental hazards happen at the same time and might or might not be related. In June 1991, Mount Pinatubo in the Philippines exploded in one of the largest volcanic eruptions of the twentieth century. The cataclysmic eruption coincided with the passage through the area of Category 3 Typhoon Yunya. Volcanic ash accumulated on roofs and readily absorbed water from the typhoon’s rain. The weight of wet ash collapsed the roofs and killed dozens of people, including places where people had sheltered because the load calculations had assumed dry ash. Pandemics and droughts can last for months, leading to the possibility of other hazards occurring during that period.

In terms of coincidental hazards which are related, many volcanic eruptions are accompanied by tremors while landslides frequently occur during earthquake-related shaking as well as during rainfall. Poor response to a drought might reduce a population’s nutrition levels making them more susceptible to diarrheal outbreaks which could then become an epidemic coinciding with the drought. Around Taiwan, typhoons have been shown to trigger “slow earthquakes” in which seismic energy is released over hours rather than seconds (Liu, Linde, and Sacks 2009).

Climate change and environmental hazards

Today, one notable influencer of many environmental hazards is the climate changing globally. The climate has always changed, throughout humanity’s and the planet’s history, including long-term trends, shifts in the state and baseline, variabilities, and cycles. Contemporary global climate change is marked by human influence through two main mechanisms. First, the release of greenhouse gases such as methane and carbon dioxide which trap heat and warm the planet’s average temperature. Second, anthropogenic changes to the Earth’s surface which reduce the planet’s ability to absorb the greenhouse gases emitted by human activities. The most prominent of these land-use changes is deforestation, since trees are an excellent source of uptake and storage of carbon dioxide.

Contemporary climate change influenced by human beings has two principal definitions. The main scientific body responsible for assessing and synthesizing climate change science for governmental approval is the Intergovernmental Panel on Climate Change (IPCC). The first IPCC assessment was published in 1990 with the latest one, the fifth assessment, published in 2013–2014. The IPCC’s (2014, 5) definition of climate change is “Climate change refers to a change in the state of the climate that can be identified (e.g., by using statistical tests) by changes in the mean and/or the variability of its properties, and that persists for an extended period, typically decades or longer. Climate change may be due to natural internal processes such as modulations of the solar cycles, volcanic eruptions, and persistent anthropogenic changes in the composition of the atmosphere or in land use.” Meanwhile, the main international treaty for addressing climate change is the United Nations Framework Convention on Climate Change (UNFCCC) which defines climate change to be “a change of
climate which is attributed directly or indirectly to human activity that alters the composition of the global atmosphere and which is in addition to natural climate variability observed over comparable time periods” (UNFCCC 1992, art. 1, para. 2). The difference in definitions is that the science examines all climate changes irrespective of the cause of the change while the international policy processes and measures consider climate change from human origins only. Both the IPCC and the UNFCCC agree that the human influence on the climate seems likely to push the planet into a climate regime that humanity has never before experienced.

Contemporary climate change influences many hazards, exacerbating some and diminishing others. Climate change is, in effect, a potential hazard driver or potential hazard diminisher, rather than being a hazard itself. The complexities of the interactions between climate change and specific hazards in specific locations sometimes make attribution and projections challenging. The IPCC (2012) attempted to systematically analyze links between climate change and hydrometeorological hazard trends and extremes. Few strong associations were found, with daily high temperature being the most prominent. It is conceivable that IPCC’s (2012) method might not be robust enough to discover correlations or lack thereof by (i) highlighting climate change perspectives and theories without equal balance from wider disaster and development topics and (ii) aiming to analyze connections that would require centuries of consistent and comparable hydrometeorological data from around the world – data which do not exist. Nonetheless, individual studies provide insight into possible influences of climate change on various environmental processes and phenomena.

For tropical cyclones, it appears as if – particularly for Atlantic hurricanes – climate change could lead to reduced frequency of storm formation but increased intensity once a storm forms (Knutson et al. 2010). There are feedbacks among sea surface temperatures, stratospheric winds, and the lessening temperature differential between the poles and the tropics (because the poles are warming faster than the tropics), in addition to other factors, which make analysis challenging. It is much clearer that storms in the Norwegian Sea termed “polar lows” are expected to decrease in frequency due to climate change because the North Atlantic’s ocean water is projected to warm more slowly than the air above it, leading to more atmospheric stability which discourages polar low formation (Zahn and Von Storch 2010).

Nonetheless, precipitation is expected to become much more intense in many locations around the world due to climate change. Warmer air can hold more moisture meaning that it can rain much more in terms of quantity and intensity. That seems likely to lead to more rainfall-induced flooding overall. Meanwhile, as sea ice around the Arctic diminishes due to climate change, storms can produce more wave energy which exacerbates coastal erosion. Communities such as Newtok, Alaska, are already experiencing this problem and are preparing to resettle inland due to it. Meanwhile, sea levels rising due to climate change are impacting low-lying islands, especially in the tropics, through worsening flood, erosion, and water salinization leading communities such as Takuu, Papua New Guinea to plan resettlement. Again, interactions are complex and climate change is not certain to increase floods in all locations. As a specific example, climate change is expected to decrease winter flood frequency in central Europe’s Elbe and Oder rivers due to fewer ice jams (Mudelsee et al. 2003).

Such complexities impact other hazards. Increased precipitation (and even increased earthquakes) under climate change could
be expected to increase the frequency and magnitude of landslides, but it is not straightforward. For a mass movement to occur there must be material to slide. An initial increase in frequency of mass movements due to climate change has the possibility of using up much of the slidable material, leading to smaller movements in the future because less material can build up before it moves. With avalanches, increased precipitation could lead to more snowfall accumulation and hence more avalanches while air temperature rise could reduce avalanche magnitude if the precipitation falls as rain rather than snow. The effects of climate change on avalanches are expected to be highly localized.

Costello et al. (2009) summarize the impacts of climate change on microbiological hazards by focusing on rodent-borne and vector-borne diseases. As temperatures increase, vectors and parasites tend to breed and mature more quickly. More life cycles are permitted within a given time frame along with an increased rate of biting, each of which supports the spread of vector-borne diseases. In addition to these time factors, vector density over a given area tends to increase with temperature increasing. Meanwhile, vectors are able to survive at higher altitudes and latitudes than before because the climate in the new locations matches what the vectors are used to in the original locations. Populations living at higher altitudes and latitudes often have never before had to deal with these vectors or pathogens, so their immunity and their knowledge of countermeasures is limited.

Costello et al. (2009) particularly highlight mosquito-borne and tick-borne diseases such as malaria, encephalitis, and dengue fever, but their arguments likely apply to many other diseases such as Lyme disease, leptospirosis, and West Nile Fever. The authors, though, indicate some factors inhibiting vector-borne and rodent-borne diseases due to climate change. For instance, where hydrometeorological hazard frequency or intensity increases, or where the environmental hazards change in nature, vector eggs and larvae could be harmed. Heavy rains can wash away vectors at many life cycle stages. Salinization of water due to sea level rise and coastal inundation could inhibit vectors which need fresh water or could force them to move inland or to higher elevations.

Prospects for geological hazards being influenced by climate change are discussed by McGuire (2013) who provides geophysical explanations for how volcanic eruptions, tsunamis, and earthquakes have the potential to be augmented. Simplistically, as sea levels rise and as glaciers melt, a shifting of the weight on the Earth’s crust has the possibility for affecting seismicity and underwater landslides, with subsequent effects on tsunamis. Similar impacts could affect volcanic eruptions. There is much speculation and many uncertainties regarding the interaction between climate change and geological hazards. In contrast, since climate change by definition affects only planet Earth, there seems little scope for impacts on astronomical hazards.

Climate change is an example of a major environmental hazard driver and diminisher, indicating the intricacies and complexities involved in trying to understand the overlaps, connections, and interactions among environmental hazards. Other examples of environmental hazard influencers are cycles such as the El Niño Southern Oscillation and the North Atlantic Oscillation.

**Influencing environmental hazards**

A major portion of human–environment interaction involves attempts to prevent or control environmental hazards. Aside from appealing to deities or nature to actively avoid environmental phenomena or processes which can be hazardous
to humanity, human efforts tend to involve technology which actively changes the environment. Such actions frequently create, augment, or exacerbate an environmental hazard—sometimes even making an environmental process or phenomenon become hazardous, as discussed above.

**Astronomical**

Astronomical processes and phenomena have perhaps been the least affected by human efforts to influence environmental hazards because of the logistical and financial difficulties of space-related operations and the scales involved. Preventing geomagnetic storms or solar flares would involve manipulating the sun, an act which is barely conceivable with current technology. Shielding the Earth from these phenomena is more realistic, but would still be an exceptional undertaking. Instead, the successful efforts shield individual pieces of infrastructure such as satellites or computers, preventing the harmful radiation from reaching damageable components.

Many realistic and achievable plans have been proposed to monitor space for near-Earth objects on a collision course with the planet and to prepare a response plan in case of a threat (Bobrowsky and Rickman 2007). Few have been implemented. The suggested responses are not as dramatic as depicted in Hollywood films portraying the use of nuclear weapons or lasers to destroy the offending object. Instead, small-scale interventions with enough lead-time could deflect an object's trajectory by a tiny fraction of a degree leading to it bypassing the Earth rather than striking it.

**Biological/ecological**

Biological/ecological hazards are influenced inadvertently and deliberately through habitat destruction and species extinction. Sometimes specific ones are targeted. Malaria and similar diseases were eliminated from the southern United States and from the United Kingdom (where it was called ague) by draining swamps and killing the mosquito vector. One microbial pathogen has been deliberately eliminated by humanity: the virus causing smallpox, although a few laboratories retain stockpiles of the virus. Many other microbial pathogens and parasites are currently the target of elimination campaigns, including dracunculiasis (Guinea worm disease) caused by a parasitic worm, measles caused by a virus, and polio also caused by a virus. More common, everyday approaches to influencing biological/ecological hazards are seen in the form of mouse traps, pesticides and insecticides used in buildings or for crops, and vaccines and antibiotics.

Influencing sentient biological hazards, such as keeping polar bears out of Arctic communities and elephants out of communities in India, is attempted, usually by discouraging the animals from entering the area rather than by killing them. That leads to questions about why preventing microbiological hazards by eliminating them is acceptable. Intense moral debates ensue on whether or not laboratory stockpiles of the virus causing smallpox should be destroyed. Vaccines and antibiotics have undoubtedly saved many lives, but their overuse and misuse is problematic, leading to stronger microbial agents. The same occurs for drug-resistant malaria parasites and insecticide-resistant mosquitoes carrying the malaria parasite. In many cases, short-term preventive measures have the potential for making longer-term prevention even harder.

Instead, interventions tackling the root source of biological hazards are often suggested. Separate water and wastewater systems alongside proper treatment and management of drinking water and wastewater provide a cleaner, healthier
ENVIRONMENTAL HAZARDS

environment and help to prevent cholera. Cleaning up the environment by avoiding stagnant water collected by discarded tires and coconut husks can reduce the places where eggs are laid by the mosquito vector of the dengue fever virus. Proper waste management, especially of food, can discourage mice and rats, reducing the chance of flea-borne diseases and of leptospirosis which is a bacterium carried in the urine of many wild animals. Constructing properties with good ventilation reduces dampness and condensation, inhibiting mold formation. Overall, preventing many biological hazards follows creating and maintaining a healthy and clean environment.

Geological

Geological hazards are difficult to influence. Certain earthquakes might be controllable, with case studies of reservoirs, dams, nuclear explosions, groundwater removal, oil extraction, or fracking (using high pressure fluids to break underground rock for releasing fossil fuels) apparently increasing small-scale seismicity (see the review by Ellsworth 2013). So far, no large earthquake has been definitively shown to have been caused or triggered by human actions. That has not stopped conjectures about injecting water or other fluids into fault lines, to induce numerous low-magnitude earthquakes for avoiding a large shock (or even perhaps to create a large, damaging shock). A danger exists of having appeared to have caused a large earthquake whether or not that was indeed the case.

Volcanically induced seismicity could conceivably be controlled by providing an artificial outlet for the volcanic gas or magma which causes earthquakes. Excavating surface material or drilling holes are possible techniques for creating the outlet. The amount of material needed to be moved and the depth of the holes required would often be far beyond that which could be achieved quickly, cheaply, and safely, so success might be limited. There is also the danger of inducing an unwanted volcanic explosion since complete control of the volcanic processes could never be attained.

Lahars (mudflows) and water from jökulhlaups can be directed by hydrological engineering techniques, including dams, levees, and floodways. Lava flows can be diverted with analogous barriers and diversions. Additionally, an advancing lava flow can be doused with water, cooling it and forcing the lava to spread sideways. Even if such techniques do not divert the lava flow, they usually slow it down, providing more time for evacuation and for natural cooling of the lava. All these techniques are controversial with regards to analyzing their success or failure.

Barriers and landscape engineering could be used to affect avalanches, landslides, rockslides, mudslides, and other such mass movements as well as diverting poison gas emissions from volcanic lakes away from settled areas. Some experiments have been completed on degassing volcanic lakes to prevent a sudden, catastrophic gas release. An example of such a gas release killed approximately 1700 people in Cameroon in 1986. The effectiveness of these approaches cannot be entirely assured. Barriers cannot always block hazards completely, since gas pressure can build up behind a barrier while snow, rocks, Earth, mud, and water can pile up behind and then overtop a barrier. Jökulhlaups and other flash floods sometimes occur because water built up behind a barrier which then fails suddenly.

Hydrometeorological

Attempts at weather modification have a long history, from efforts to seed clouds to yield rain to efforts to change the strengths and tracks of storms including hurricanes. The effectiveness of
these measures has always been disputed, especially because it is difficult to prove what might have happened in the absence of cloud seeding. In 1947, the American Army and General Electric Corporation seeded a hurricane which suddenly altered its direction and struck land in Georgia and South Carolina. Damage resulted, so people affected blamed the seeding. Overall, controlling rain, hail, and storms is fraught with difficulties and it is rarely straightforward to attribute cause and effect, although cloud seeding is still commonly carried out around the world from the Philippines to the United Arab Emirates to the United States.

Floods and droughts are easier to influence. Structural measures such as dams, levees, floodways, dredging, and reconstructing the courses of waterways are used for influencing these hazards. Dams such as the Aswan Dam along the Nile River and China’s Three Gorges Dam were built for several reasons, some of which were political, but incorporating statements that they would help agriculture by controlling water flow meaning flood and drought prevention. A floodway was built in 1968 to send overflow waters from the Red River around Winnipeg, Manitoba, rather than having the city center flood regularly. London’s Thames Barrier and Singapore’s Marina Barrage assist in keeping periodic floodwaters out of those cities’ streets.

But every structure has design limits and those limits can be exceeded, leading to a catastrophic flood. The flooding of New Orleans, Louisiana, in 2005 from Hurricane Katrina; along the Saguenay River, Québec, in 1996; and in the town of Morbi, India, in 1979 resulted from failures of the structural measures designed to prevent flood hazards.

In many cases, the efforts to use structures to prevent flood hazards end up creating or exacerbating flood hazards (Etkin 1999). Engineering waterways can cause waters to flow faster. More insidiously, people tend to settle in floodplains which have been dried out due to walls, levees, and barrages, as with New Orleans and London, meaning that extensive infrastructure is built because the area is deemed safe. In addition to more assets being damaged when the design limits of the structural defense are exceeded by a flood, the urban buildup augments run-off, raising the water level and increasing the flood hazard. Consequently, efforts to prevent the flood hazard actually exacerbate the flood hazard and the disaster is much worse than it could have been in the absence of the structural defense.

Living with environmental hazards

So far, environmental hazards have been defined and described along with their influence on, and how they are influenced by, society. To a large degree, the hazardousness is socially constructed, owing to either the viewpoint taken or to human actions which make an environmental phenomenon or process hazardous. An implicit baseline has been present, from the beginning, where the environmental phenomenon or process is neutral, but society can make it negative, that is, a hazard. Society also reaps rewards from environmental phenomena and processes, even when they are hazards.

Advantages of environmental processes and phenomena

The previous section on the influencing of environmental hazards identified significant problems when humanity tries to control environmental hazards. Influencing environmental hazards predictably is not straightforward, with the potential for dangers to emerge from those efforts. That is especially the case when environmental phenomena and processes are important to the
ENVIRONMENTAL HAZARDS

environment, to society, or to both, being part of what nature needs while providing society with opportunities for life and livelihood. This does not support zero interventions, but instead suggests a balance of actions which complement each other, including but not limited to environmental hazard prevention.

For example, orogeny – the process of mountain building – is an environmental process emerging from tectonic forces, illustrating the continual balance of creation and destruction achieved by nature, a balance which cannot and should not be prevented. The island of Surtsey, Iceland was created by a series of volcanic eruptions in 1963 – although this creation process “destroyed” a section of the ocean. The March 27, 1964, earthquake in Alaska changed the elevation of more than 250,000 km² of land, with large portions of land dropping into or rising from the ocean. Seismic zones frequently make groundwater accessible so that urban centers, such as in Iran, grow across or near earthquake faults because of (not despite) the earthquake hazard (Jackson 2001).

Volcanic ash blankets land resulting in a barren landscape – until the nutrients enter the ground and create fertile soil, amply demonstrated in many parts of Indonesia. Volcanoes contribute significantly to biogeochemical cycles of metals, including aluminum, arsenic, cadmium, copper, iron, lead, mercury, and zinc; assist the sulfur, chlorine, and nitrogen cycles through gas emissions; and contribute to the atmosphere’s evolution into its breathable-for-humans state today. Tectonic subduction and spreading processes cycle minerals and elements by eliminating old rock and creating new rock. All these environmental phenomena and processes make the world livable for humanity, so despite the hazards which might result, they are essential and are to be welcomed.

Hydrometeorological processes and phenomena involve the transfer of mass, heat, and linear momentum through the atmosphere and hydrosphere, and so are responsible for air flows, ocean currents, weather, and climate. Winds knock down dead trees permitting the wood to decompose, adding nutrients to the soil, while leaving space for new growth. When humans or their infrastructure sit below that tree, the wind appears hazardous, but it is actually a needed environmental phenomenon which contributes to a healthy ecosystem. Glacial movement has been responsible for the good farming till, along with much of the topography, in places such as southern Ontario and Orkney, Scotland. The water cycle, incorporating precipitation, affects most biogeochemical cycles and makes available fresh water needed for human survival. Some Caribbean islands receive a major proportion of their annual fresh water from storms.

Hydrometeorological processes and phenomena also contribute to ecological and geological processes. Floods scour out stagnant waterways and inundate tracts of land, often causing damage to the land and vegetation, but depositing nutrients so that people settle in floodplains to farm the land – a process which makes life both feasible and hazardous on the island chars of Bangladesh. In the African Sahel, many farmers traditionally use “flood retreat farming.” When the usual annual floodwaters retreat from the floodplain, the farmers plant seeds in the floodplain so that the moisture left behind can be used by the seeds before the moisture evaporates or drains (Matlock and Cockrum 1976).

Biological/ecological processes and phenomena have many advantages, leading to a major concern about the speed and extent of human-induced ecosystem destruction and species extinction. Food production, medicine, recreation, and materials development continually
experience advances from new species discoveries along with new ways of using a species without devastating it or its habitats. In fact, “bioprospecting” – the search for new species products which can be exploited commercially – tends to find its greatest rewards in biodiversity conservation hotspots. Many animals and plants have direct advantages for society, irrespective of their potential hazardousness. Venom from a saw-scaled viper kills people and is used to produce drugs combating heart attacks. The poisonous nightshade is applied to detoxify PCB (polychlorinated biphenyl) chemicals. Plenty of meat and hide can be obtained from large and often dangerous game animals, which could be sustained as long as they are bred and hunted in such a way that does not harm the species or their ecosystems.

Astronomical processes and phenomena have brought major advantages to humanity. Most accessible precious metals are theorized to have arrived due to meteorite strikes (Willbold, Elliott, and Moorbath 2011). The bombardment occurred long before humanity existed, meaning that society can exploit the advantages of the astronomical processes and phenomena without worrying about the associated hazard.

Managing hazardousness

Given these advantages across environmental processes and phenomena – even those which are made to become hazardous – the view should never be about humanity versus nature, with humanity needing protection against terrible environmental hazards and therefore building up protection or defenses. This is not a game during which “nature bats last” or during which each side gets a turn to try to harm the other. Dealing with environmental processes and phenomena and reducing hazardousness is not about surviving despite adversity or countering abnormal, unpredictable, unwelcome extremes (see also Hewitt 1983).

Instead, humanity and society could decide to learn to live with the environment in all its forms. This approach does not mean harmonious living where everyone adores nature and nature returns that love. No one expects a monthly ritual of tree worship or dismantling our homes and settlements to live exposed to the elements, clothed in only leaves. “Living with environmental hazards” means recognizing the intertwined resource and hazard aspects of environmental phenomena and processes and using those as much as feasible to create a society in which damage from the environment is minimized and gains are maximized. Dangers can never be eliminated, nor should they be. Hazardous aspects will and should always remain, in the same way that gravity as an environmental phenomenon is essential for living but is also frequently hazardous, through falling off a cliff or dropping a glass. Nevertheless, it is possible to live in such a way that environmental processes and phenomena, such as gravity and precipitation, are integrated into daily life and livelihoods in such a way that people and society can deal with them without being harmed or causing harm.

That can be difficult when an environmental process or phenomenon has not manifested in living memory or even since humans settled the area. Conversely, we know enough to be able to assess any location and to act on those assessments in order to reduce the potential hazards while exploring and applying the benefits. Rather than “living with environmental hazards” being harm avoidance and nothing else, gaining knowledge and using the opportunities is necessary too, through managing the hazardous aspects. That also means wider reduction of risk and vulnerability in terms of addressing human behavior such as population numbers,
consumption rates, inequality, injustice, and wealth/asset distribution.

Irrespective, living with environmental phenomena and processes by just ignoring the hazard potential, and without any form of separation between humanity and nature, is not possible. Humans need daily shelter from, for example, precipitation and extreme temperatures. Where a volcano continually erupts, such as in Montserrat, or where degassing a volcanic lake cannot be completed with full confidence, which might be the case in Cameroon, then the main option for ensuring that people live with – rather than die from – environmental hazards might be protection. In Montserrat, people and communities have been evacuated out of the main danger zones. In Cameroon, resettlement in consultation with the communities might be the best option as well – or it might be feasible to design a warning system giving enough time to escape when a gas release occurs. All suggestions need to be considered carefully to avoid any potential snags or shortcomings. Often, the key is not to rely on a single approach, but instead to combine elements of avoidance, warning, protection, designing infrastructure, and other suggestions, overall achieving a balance of multiple approaches.

Environmental phenomena and processes – and their potential hazardousness – are based on balances among, interaction between, and cycles throughout the atmosphere (air), the hydrosphere (water), the lithosphere (Earth’s surface and interior), and the biosphere (life) – while not neglecting the environment outside the Earth, namely space and the rest of the universe. Humans must live with and within that environment. Environmental processes and phenomena are unavoidable and are needed components of the Earth, making the planet habitable. Even if it were possible, preventing all potential hazards from the environment would severely impact the resources available to society, particularly over the long term. Environmental processes and phenomena are frequently damaging, but it is the damage to be avoided rather than the environmental processes and phenomena themselves.

**SEE ALSO:** Climate change, concept of; Climate and societal impacts; Disease diffusion; Disease and illness, political ecology of; Environment and disease; Geomorphic hazards; Glacier lake outburst floods; Glaciers; Global climate change; Mass movement processes and landforms; Paleofloods; Social resilience and environmental hazards; Social vulnerability and environmental hazards; Tropical cyclones and tropical climate; Volcanic processes and landforms; Vulnerability; Weather, extreme

**References**


**Further reading**


Environmental health

Brian King
Pennsylvania State University, USA

Environmental health addresses the quality and sustainability of the natural environment in order to understand the impacts for human health and wellbeing. This requires a historical and spatial approach that is sensitive to the complex interrelationships between human and nonhuman species, and within natural and physical settings. Environmental health is a field within public health that examines the ways in which the natural and physical, or human-constructed, environment shapes human health and wellbeing. These interrelationships can be complex and historically mediated as well as varying across different contexts.

Environmental health studies how different population groups are unequally exposed to pollutants and other conditions that shape wellbeing. Environmental negatives, such as proximity to landfills and polluting facilities, are differentially experienced based on race, socioeconomic class, gender, generational differences, and ethnicity. Similarly, environmental positives are not universally experienced, with some individuals and communities more likely to be located in close proximity to green space for recreation, bike lanes for alternative transportation, and full service grocery stores that offer healthy food and produce. Much of the awareness of these environmental inequities has come from activist and policy communities that use the concept of environmental justice to show how some groups are more vulnerable to health hazards than others.

A standard definition of environmental health is provided by the World Health Organization (n.d.), which emphasizes all the physical, chemical, and biological factors external to a person, and all the related factors impacting behaviours. It encompasses the assessment and control of those environmental factors that can potentially affect health. It is targeted towards preventing disease and creating health-supportive environments. This definition excludes behaviour not related to environment, as well as behaviour related to the social and cultural environment, and genetics.

The World Health Organization notes that millions of deaths can be prevented each year by addressing environmental hazards, such as unsafe drinking water or exposure to air pollution. Within developing countries, the main environmentally caused diseases are diarrheal disease, lower respiratory infections, unintentional injuries and accidents, and malaria. Within wealthier and industrialized countries, cancer, cardiovascular disease, asthma, lower respiratory infections, and traffic injuries are among the primary environmental health hazards (see Figure 1). According to the World Health Organization, 85 out of the 102 categories of diseases and injuries listed in the World Health Report are influenced by environmental factors (Prüss-Üstün and Corvalán 2006). This contributes to the assertion from multiple agencies and policymakers that healthy and sustainable environments are needed to reduce mortality for human populations. These environments also influence human morbidity, which is the incidence or prevalence of disease within a population.

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0754
In considering environmental health, it is helpful to distinguish between infectious and noninfectious, or noncommunicable, diseases. Infectious diseases are often the product of social and ecological conditions that facilitate their spread. As one example, *Anopheles* mosquitoes require specific ecological conditions in order to survive and to transmit the malaria parasite (see Figure 2). Recent scientific studies have shown these mosquitoes to be highly dependent on the natural environment, specifically in terms of precipitation and temperature which can alter the environmental conditions that allow for their survival. Recent work has shown that daily fluctuations in temperature affect parasite infection, the rate of parasite development,
and the components of mosquito biology that combine to determine the intensity of malaria transmission.

In addition to the natural environment, the spread of malaria is also facilitated through host organisms, also known as disease vectors, which help transmit the virus. Human populations that reside within malarial zones are placed in direct contact with mosquitoes, but cultural patterns and behavioral practices, such as the adoption of a bed net, can reduce susceptibility to infection. As a result, the spread of an infectious disease, like malaria, should be understood through a diverse mix of social and ecological conditions that combine to produce the necessary conditions for exposure.

**History of environmental health research**

Research that addresses the relationships between human health and the environment has been advanced by a number of disciplines. One influential example was the work of Dr John Snow, a British physician and early leader in the emerging field of epidemiology. Snow was a skeptic of the dominant miasma theory which held that diseases such as cholera or the Black Death were caused by pollution or a noxious form of “bad air.” An outbreak of cholera in the Soho district of London in 1854 allowed Snow to test his belief in germ theory, an alternative that was not widely accepted at the time. Cholera is an infectious disease that causes severe diarrhea,
ENVIRONMENTAL HEALTH

which can lead to dehydration and, in extreme cases, death. It is caused by the consumption of food or drinking water contaminated with a bacterium called *Vibrio cholerae*. Through careful research Snow was able to demonstrate how cholera was spread through the environment. In speaking with those affected by the outbreak, he was able to locate the source of cholera infection as a public water pump on Broad Street. Snow's study helped to establish the field of epidemiology, which addresses how diseases are spread, and to demonstrate how the built environment has a direct bearing on the health and well-being of human populations as well as giving greater credence to germ theory within public health.

Related work on environmental health has come from the field of environmental history. Environmental historians have shown how the management of environmental landscapes to reduce the spread of infectious disease has resulted in distinct infrastructure and development projects. The eradication of malaria in the continental United States was the result of aggressive efforts to remove the environmental conditions that enabled the spread of the *Anopheles* mosquito. In *Inescapable Ecologies*, Linda Nash (2006) provides a compelling examination of changing understandings of environmental health within the United States. At the time of European expansion into California, the dominant understanding of environmental health was based on humoral theory, which asserted that the blood and other bodily fluids, regarded as “humors,” had to remain in proper balance for health to be maintained. Humoral theory depended on a “constant exchange between inside and outside, by fluxes and flows, and by its close dependence on the surrounding environment” (Nash 2006, 12). This led to European settlers being concerned about remaining in balance with the natural environment through the human skin, which was understood as a porous boundary that enabled interactions between the individual and the environment. Nash suggests that the dominance of germ theory in the late nineteenth and early twentieth centuries shifted attitudes so that the skin was seen as a layer of protection from outside threats. This would influence new ways of viewing nature and its impact on human health.

Understanding of the changing relations between humans and the natural world shifted in the twentieth century. A clear example of this was the work of Rachel Carson, who used scientific methods to demonstrate the relationships between human health and the natural environment. Carson was a leading scholar and environmentalist who helped raise public awareness about the hazards of pesticides. The publication of *Silent Spring* in 1962 made the use of pesticides and other chemicals a major concern. Carson argued that chemicals, such as dichlorodiphenyltrichloroethane (DDT) were becoming widely used for controlling disease and pests. Modern agriculture relied on pesticides, as well as on single-crop farming and imported species. The combination of careful scientific analysis and Carson’s compelling writing led to a ban on the use of DDT in the United States, the formation of the Environmental Protection Agency, and the implementation of other environmental acts such as the Federal Insecticide, Fungicide, and Rodenticide Act, which regulates the use of these chemicals.

In the early twenty-first century, increasing industrialization and urbanization within the United States has contributed to a perception that humans have become more disconnected from the natural environment. In *Last Child in the Woods*, Richard Louv (2008) argues that the current generation is experiencing a new relationship with the natural world that
is characterized by five elements: (i) severance of the public and private mind from our food origins; (ii) a disappearing dividing line between machines, humans, and other animals; (iii) an increasingly intellectual understanding of our relationship with other animals; (iv) the invasion of our cities by wild animals (even as urban/suburban designers replace wildness with synthetic nature); and (v) the rise of the suburb as a new form of spatial organization. Louv asserts that these dynamics reduce the direct contact between humans and nature in ways that have profound effects for how people understand their position within the system.

These contributions help advance environmental health by arguing that health and wellbeing can be produced by direct contact with the natural environment; in other words, environmental health is generated through positive encounters with nature that provide therapeutic benefits. This idea has been promoted by landscape architects and health geographers for some time. Landscape architects and urban planners work to create landscapes that have beneficial elements for human health, such as green space or bike lanes to encourage exercise and alternative commuting options. Health geography has worked with the concept of a therapeutic landscape to show that people’s connection with certain places can be beneficial to their health and to their relationship with the external environment.

It is recognized that the disciplines of environmental epidemiology, toxicology, and exposure science contribute significantly to the field of environmental health. Environmental epidemiology studies the relationship between environmental exposures, such as exposure to chemicals, radiation, microbiological agents, and human health. Toxicology studies how these exposures lead to specific health outcomes. Finally, exposure science examines human exposure to environmental contaminants by both identifying and quantifying exposures. In addition to these research fields, scholars and practitioners within social epidemiology, medical anthropology, and public health work to demonstrate the interconnections between human populations and the natural and built environment.

**Causes of environmental health patterns**

The incidence of infectious disease is not equally experienced around the world and different factors are important in different places. Developing countries, such as those in Africa, face distinct challenges in addressing the underlying environmental health conditions that shape the spread of infectious disease or management of noncommunicable disease. More than 90% of the estimated 300 million to 500 million malaria cases worldwide every year occur in Africa, mainly in children under five years of age, but many governments are moving toward better treatment policies (WHO 2006).

The United Nations has emphasized the need to concentrate on noncommunicable human diseases, which were identified as the primary cause of death and disability in the world. Roughly two-thirds of annual deaths are caused by cardiovascular diseases, cancers, chronic respiratory diseases, and diabetes. Because many of these can be prevented, public health experts emphasize the importance of reducing tobacco and alcohol use, improving unhealthy diets, and addressing a lack of exercise. There are also a variety of behavioral and physical environmental factors that make such noncommunicable diseases more common. Obesity and type 2 diabetes are often identified as being caused by the relationships people have with their surrounding environment. Built infrastructure, which encourages
driving rather than walking or cycling, is often blamed for giving rise to poor health outcomes. Employment occupations that require more sedentary lifestyles can contribute to decreased physical exercise. The role of diet and nutrition are also central to current understandings of noncommunicable diseases.

Another factor affecting environmental health is large-scale, industrial agricultural production. In the United States this affects the production and consumption of food, which have consequences for human health. Popular authors such as Michael Pollan (2006) and Eric Schlosser (2001) point to shifts in the diet of Americans toward heavily processed foods, meat products, and fast foods, arguing that these have transformed the social environment and led to dramatic consequences for economic systems, cultural patterns, and nutrition. Pollan and Schlosser assert that the expansion of industrialized food production has led to an overreliance on pesticides and reduced crop diversity which make ecosystems vulnerable. Other authors warn that increasing exposure to pesticides through food production and consumption patterns could damage human and environmental health. These shifts in agricultural production, while notable in the increased production of food per unit area, are also credited with disrupting human and environmental interactions in ways that can reduce human health and environmental quality.

Human health is the product not just of the natural environment. Environmental health is the consequence of policies, planning mechanisms, and political and economic processes that give rise to places and landscapes in which the conditions that contribute to noncommunicable diseases are produced. These same conditions also result in differential vulnerability to infectious diseases within places and populations.

Environmental health futures

Environmental health represents one of the central challenges of the twenty-first century. Anthropogenic climate change is expected to rework and disrupt human and environmental systems, and therefore to affect the spread of infectious disease. Shifts in precipitation and temperature dynamics may facilitate new health challenges for nonhuman species. The Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (Smith et al. 2014) includes a chapter that concentrates on the links between global climate change and human health. The IPCC identifies changes in temperature and precipitation, and the occurrence of heat waves, floods, droughts, and fires, as having direct effects for human populations. It is also anticipated that indirect effects to human health may result from ecological disruptions caused by climate change, such as crop failures or shifts in disease vector patterns. Scientists point to the spread of malaria in parts of Africa and the West Nile virus in the United States as examples of how infectious disease can accompany changing weather patterns, demonstrating that changes to the underlying environment can have repercussions for human wellbeing.

The IPCC report (Smith et al. 2014) also notes that there can be social effects that influence health, such as the displacement of populations due to sustained drought or declines in agricultural productivity. It is believed that, until the middle of the twenty-first century, climate change will largely intensify already existing health problems, with the largest risks to populations that are most directly affected by climate-related diseases. This could occur with shifting disease vectors or the negative health effects from increased food insecurity. The IPCC emphasizes the importance of reducing vulnerability to these changes by ensuring sustainable
public health measures that increase access to clean water and sanitation, health-care services including vaccination and child health services, the reduction of socioeconomic poverty, and improvement in the ability of populations to respond to disasters.

Another factor shaping the possibilities for healthy people and environments is economic development. As evidence of this, the United Nations Millennium Development Goals emphasize health and environment as key challenges for the global population. Goal 6, designed to combat HIV/AIDS, malaria, and other diseases, identifies the need to expand access to treatment for HIV/AIDS while also altering the conditions that facilitate the spread of malaria, which causes the deaths of hundreds of thousands of people each year, largely in the developing world. Goal 7 emphasizes the importance of environmental sustainability. In addition to addressing biodiversity loss and sustainable policymaking, this goal seeks to reduce the proportion of the population living without sustainable access to safe drinking water and sanitation. The urban population residing within informal and slum conditions was also identified as a key environmental health factor. According to the United Nations, 863 million people are estimated to be living in slums in 2012 compared to 650 million in 1990 and 760 million in 2000. However, the proportion of urban slum residents in the developing world declined from 39% in 2000 to 33% in 2012, and more than 200 million of this population gained access to improved water sources, improved sanitation facilities, or durable or less crowded housing, thereby exceeding Millennium Development Goal 7 (United Nations 2013). This demonstrates an earlier point that investments in natural and physical environments can have positive impacts for human health and wellbeing.

The concept of sustainable development is widely acknowledged as being integral to environmental health. This is particularly the case given that mainstream economic development is generally credited with improving environmental health; however, there can be unintended consequences. The construction of hydroelectric dams in parts of the developing world disrupts biophysical conditions, and has made individuals and communities vulnerable to schistosomiasis and malaria. The increasing use of chemical fertilizers and pesticides in global agriculture has increased the risks to agricultural laborers and ecosystems. Sustainable development is defined in multiple ways, but is generally understood as addressing the social and economic needs of the present generation without compromising the ability of future generations to meet their own needs. Global conferences, such as the 2002 World Summit on Sustainable Development held in South Africa, were intended to make concrete links between economic and environmental wellbeing and to encourage policies that ensure sustainable practices across multiple settings and scales. Declarations from the 2002 World Summit emphasized that sustainable development cannot be achieved as long as there is a high prevalence of illness and poverty. Additionally, it is understood that environmental degradation, the unsustainable use of natural resources, and unhealthy consumption patterns and lifestyles impact health. These impacts can contribute to a downward spiral whereby poor health reduces the capacity of individuals and populations to move out of socioeconomic poverty.

In addition to existing factors, a public health bulletin from the World Health Organization (2011) emphasizes that new health hazards could develop in an increasingly urbanized world. Road traffic injuries are the ninth leading cause of death globally, with most occurring in low-
ENVIRONMENTAL HEALTH

and middle-income countries. It is estimated that urban air pollution kills around 1.2 million people globally each year, primarily due to cardiovascular and respiratory diseases. Infectious diseases, such as tuberculosis, can concentrate in large urban centers. In New York City, the tuberculosis incidence is four times higher than the national average. Lastly, the World Health Organization indicates that urban environments are not always designed to support physical activity or promote healthy food consumption. Participation in physical activity is constrained by a variety of urban factors including overcrowding, high-volume traffic, heavy use of motorized transportation, poor air quality, and lack of safe public spaces and recreation facilities.

Geographic approaches to environmental health

The discipline of geography has a well-established interest in human–environment interactions and the ways in which they shape the lived experiences of people in various settings. Geographic research on human disease and health has been centered largely in the subfields of medical geography and health geography, although emerging research within development studies, political ecology, and feminist geography has advanced the frontiers of environmental health. Much of this scholarship contributes to environmental health by showing the social and ecological determinants that shape the spread of human disease and differential vulnerabilities across time and space.

Geographic scholarship has considered how environments are produced with either direct or indirect consequences for human health. As one example, medical geography focuses on disease ecology, mapping of disease, spatial analysis, and disease diffusion, in addition to locational analysis and the regionalization of health services. Medical geographers often use quantitative data that can be analyzed statistically. Geographic information systems (GIS) are also used to evaluate spatial patterns. Disease ecology is one of the earliest academic fields focusing on the interconnections between social and environmental systems in contributing to disease transmission. As Mayer (1996, 441) explains, disease ecology examines how “humanity, including culture, society and behavior; the physical world, including topography, vegetation and climate; and biology, including vector and pathogen ecology, interact together in an evolving and interactive system, to produce foci of disease.” Additionally, the underlying biophysical conditions that contribute to making individuals and communities vulnerable to disease are addressed to identify the spread of disease and vulnerabilities of human populations. Disease ecology has been foundational in showing how the relationships between humans and the natural environment can determine the spread of disease.

Other studies have shown how individual decision-making is simultaneously shaped by external political and economic systems. One geographic subfield that has addressed environmental health in this way is political ecology. From its beginnings, political ecology has provided a rigorous analysis of how political economy structures the decision-making opportunities for local actors and contributed to environmental change in various settings. As one example, deforestation in the Brazilian Amazon is presented not as a discrete act in one location but as the product of complex and interlinked systems that intersect across spatial and temporal scales. In the now classic Land Degradation and Society, Piers Blaikie and Harold
Brookfield (1987) demonstrate the relationships between social and biophysical processes, linking local management practices to broader processes that showed how choices to transform the landscape were inherently rational and the product of political and economic systems. While not directly focused on human health, these studies explained how individual actors and decisions are linked to broader systems that spread across multiple spatial and temporal scales. As such, to better understand the drivers of environmental change, greater detail on the determinants of human health and wellbeing is needed.

Other studies at this time demonstrated how decision-making was tied to external structures and capitalist relations of production that constrained the available opportunities to local populations. In *Silent Violence*, Michael Watts (1983) examines how famines serve as an interconnected process between society, political economy, and the environment to shape the periodic collapse of rural production and distribution systems in northern Nigeria. Watts considers various forms of surplus extraction and state intervention that increased local vulnerability to drought within rural households, and thereby provided a significant challenge to the view that famines result from an environmental collapse and subsequent decline in agricultural productivity. More recently, work within political ecology has shown how economic neoliberalism has produced new pressures for local populations and environments, either by decentralizing responsibility for management or by commodifying natural resources to justify particular management strategies that benefit external actors. All of this research, while different in their object of study, shows how environmental health is produced through multiscalar dynamics that shape the possibilities for human health and wellbeing.

**Environmental justice**

The concept of environmental justice has been utilized to emphasize how different groups are made vulnerable to health hazards emanating from toxic facilities, landfills, incinerators, or chemical plants. Environmental justice activists have therefore argued that environmental health is differentially experienced on the basis of socioeconomic class, race, gender, and age. An early example was the case of Love Canal, which showed how exposure to environmental toxins could have significant health implications. Love Canal is a community in upstate New York that, at the time of the exposure, was largely middle-class and white. The community was built on top of 21,000 tons of toxic waste that had been dumped by Occidental Petroleum Corporation, previously Hooker Chemical Company. In 1978 rainfall well above the historical average caused the improperly stored containers to leak their toxic contents into the basements of 100 homes and a public school. This generated widespread concern because the landfill was believed to contain 82 different compounds, of which 11 were suspected carcinogens at the time. Love Canal attracted national attention that crystallized through the formation of the Love Canal Homeowners Association, which pressured the state to relocate residents (see Figure 3). It represented a landmark event for environmental health by demonstrating how communities could unwittingly be exposed to toxic pollutants. It also helped establish the United States Environmental Protection Agency’s Superfund legislation, which identifies toxic waste sites for remediation.

The Love Canal case showed the American public how environmental health could be dramatically disrupted by unsustainable environmental practices. It also alerted the general population to the fact that they could be
exposed to toxic pollutants. Other cases showed the inequities in exposure to pollutants within the United States. Specifically, a series of cases involving the siting of landfills helped generate interest in the concept of environmental justice. In 1978 it was announced that a landfill to store polychlorinated biphenyls (PCBs) was to be placed in Warren County, North Carolina. The community directly impacted was Afton, an area of high socioeconomic poverty with a largely African American population. Community residents were strongly opposed to the siting of the landfill and employed litigation and nonviolent protests to try to prevent its opening in 1982. While they were unsuccessful in stopping the landfill, the Warren County case had lasting effects for the environmental justice movement in establishing links between exposure to pollution and social and spatial patterns.

Similarly, a proposal in the early 1980s to create the Whispering Pines sanitary landfill in a predominantly African American neighborhood in Houston showed how race, class, and environmental health are closely integrated. While African Americans made up 25% of the city’s population at that time, the five city-owned landfills (100%) and six of the eight city-owned incinerators (75%) were in African American neighborhoods. From 1920 to 1978, 11 of the 13 city-owned landfills and incinerators (84.6%) were built in these same neighborhoods (Bullard 2005). A lawsuit was initiated to prevent the Whispering Pines sanitary landfill from opening. Bean v. Southwestern Waste Management Corp. was the first lawsuit in the United States to argue environmental discrimination in waste-facility siting under the Civil Rights Act. As with the Warren County landfill, the lawsuit was unsuccessful in preventing the opening of Whispering Pines; however, it broke new ground in advancing claims of environmental justice through the application of civil rights legislation.

In addition to inequitable exposure to toxins and pollutants, environmental justice theory and practice have also been concerned with differential proximity to environmental conditions that can contribute positively to human health and wellbeing. Within urban settings in the United States, for example, there can be differences in terms of where greenspace, such as parks and bikeways, are located. This can be the product of long-standing urban planning as well as being driven by socioeconomic dynamics that benefit some urban residents and not others. Another example is variations in access to certain types of grocery stores that offer fresh produce and
healthy foods. Scholars and practitioners have used the concept of a food desert to refer to rural and urban locations that lack access to full-service grocery stores. Both of these dynamics show the relationships between environmental health and environmental justice in highlighting inequities between population groups. They also attest to the fact that environmental health is produced through a number of factors, whether it be capital and investment or deliberate planning processes. Human health, therefore, is not just a product of the natural environment; rather, it is produced through planning and political and economic processes that generate places and landscapes that produce inequities in infectious disease exposure or the conditions that contribute to noncommunicable diseases.

At the same time, healthy environments can be directly impacted by human disease. Recent reviews on the ecological effects of HIV/AIDS suggest that affected populations utilize natural resources in the short term through unsustainable practices (Talman, Bolton, and Walson 2013). Land-use systems can be modified or transformed such that agricultural land remains uncultivated, thereby disrupting food security. Additionally, research is showing that disease can pressure currently protected natural resource areas as a result of the increased need to mitigate the costly impacts. For example, some studies suggest that there is an increase in the hunting of wild animals or the collection of medicinal plants or wood as a strategy to offset the costs of managing health. These patterns demonstrate that healthy environments do not just produce health outcomes for human populations, but that disease and poor health can similarly produce unhealthy environments. Therefore environmental health must be seen as a series of interconnected and reciprocal relationships between humans and their natural environment. Environmental health necessitates attention not only to social processes but also to the ecological dynamics that shape the possibilities for human wellbeing and environmental sustainability.

SEE ALSO: Environment and health; Environmental (in)justice; Environmental movements and protest; Health inequalities; Neighborhoods and health; Political ecology; Therapeutic landscapes; Urban political ecology

References


Further reading


Environmental history

Eric Pawson
University of Canterbury, New Zealand

Andreas Aagaard Christensen
University of Copenhagen, Denmark

The term “environmental history” has come into widespread use since the 1970s, initially among historians in the United States, and more recently in continental Europe and elsewhere. The first World Congress of Environmental History was held at Copenhagen in Denmark in 2009 to explore the historic relationship between people and the environment over time. The breadth of this ambition indicates that the practice of environmental history cannot be, and never has been, the purview of one discipline, and arguably history itself came late to the scene. Rather, environmental history is an interdisciplinary pursuit, with roots in geographical writing and contributions from a wide range of other subjects, such as anthropology, ecology, ecological economics, and environmental philosophy. It has developed as a form of environmental conscience that seeks to counter an increasingly powerful, forward-looking liberal theory of the environment.

The first section of this entry explores some of the early contributions to the field from geographers. It shows how these introduced the now recognized themes of environmental history: the human impact on the Earth and the ways in which this has been shaped by human understandings of nature and land – in other words, the relations between environmental ideas and materialities. The prophetic overtones of environmental history have, however, rarely been assembled into a coherent global-scale assessment of the accelerating environmental impact of people since the mid-twentieth century. The second section sketches how this might be achieved, while the third discusses the types of manufactured environmental risks that have become widespread as a result. The sort of narratives employed to explain changing people–environment relations are then discussed, before considering what environmental history can contribute in the future.

Geographical foundations

In 1955, as the pace and extent of environmental impacts in the postwar world were becoming clear, a now famous symposium was held at Princeton, New Jersey, in the United States. It brought together 70 scholars from the social, earth, and natural sciences to discuss the theme of “Man’s Role in Changing the Face of the Earth.” One of the symposium chairs was the eminent cultural geographer Carl O. Sauer; also participating was Clarence J. Glacken, a colleague of Sauer’s from the Berkeley campus of the University of California. Glacken later came to prominence as the author of one of geography’s seminal books, about nature and culture in 2000 years of Western thought. Towering above the Princeton meeting, however, was the figure of the nineteenth-century American geographer George Perkins Marsh: indeed, Sauer’s hope was for a “Marsh Festival” (Thomas 1955, 49).
ENVIRONMENTAL HISTORY

Marsh, who had been born in Vermont in 1801, was a nineteenth-century polymath and legislator, who also served the American government overseas, in Turkey and in Italy. In 1864 he published *Man and Nature; or, Physical Geography as Modified by Human Action*. Another edition, ten years later, was subtitled *The Earth as Modified by Human Action*. The book remained in print for prolonged periods. It first appeared at the height of colonial resource optimism, and through force of evidence and example argued against the prevailing belief that nature existed for people to use at will. Marsh’s experience of the degradation of Mediterranean environments made him anxious for America’s future. “Man has too long forgotten,” he wrote, “that the earth was given to him for usufruct alone, not for consumption, still less for profligate waste” (Marsh 1965/1864, 36). His practical impact at the time was on forest policy and watershed protection in his homeland and in British India; his broader intellectual contribution was to shift the perspective from humans endowed with a limitless nature to a longer-term understanding of their use and abuse of, and the need to care for, the Earth.

Carl Sauer’s admiration of Marsh owed much to this latter point. Sauer’s early experience with the Michigan Economic Land Survey revealed that land use was not always wise and that people could damage as well as improve an area. His subsequent fieldwork in Mexico was driven by his curiosity to discover the ways in which, at different times and in different places, culturally differentiated landscapes were produced. He worked closely with anthropologists at Berkeley to achieve a greater understanding of place, process, and period. His perspective was shaped by a strong anti-utilitarian sensibility; he believed that the variety and wealth of human inventiveness manifest in the landscape was in danger of being sacrificed to the short-term growth agenda of a modern liberal economy. His involvement in the Princeton meeting gave him the opportunity to frame the debate not in terms of technical solutions for the future but by asking “How did we get to where we are?” and “How can we construct an intelligible description of where we are?” (Williams 2014, 155).

Clarence Glacken’s great work, *Traces on the Rhodian Shore* (1967), also took the long view into the past. It is the most complete survey we have of three persistent ideas that have characterized human thinking about nature since the time of classical Greece. The first is the one that Marsh challenged, that of a “designed Earth” purposively made for human use. The second is the influence of the environment in molding the nature of individuals and the character of culture. This was a perspective once popular in American geography as “environmental determinism,” and roundly rejected by the young Sauer. The third, which Glacken concluded was not as well formulated in antiquity as the other two, was that which motivated both Marsh and Sauer: charting the human transformation of the world.

These three ideas were for a long time not seen as contradictory, which is why each persisted with varying emphases through time. A designed Earth could also be one in which all life was understood to shape itself to purposefully created harmonious conditions, just as human beings also fulfill their role of finishing creation, bringing order to nature as God’s stewards. Marsh saw that human actions need do no such thing, with his *cri de coeur* that the “earth is fast becoming an unfit home for its noblest inhabitant” (1965/1864, 43). But his choice of words is consistent with Glacken’s observation that, as early as the Hellenistic period, “the fundamental cleavage of human from other forms of life” (1967, 708) was recognized. Over time this cleavage came to legitimate the role
of people as supreme in creation. It fashioned seventeenth- and eighteenth-century Enlightenment thinking that to understand the world is the way to reshape it for our own ends. The anthropocentrism that underlies the ambition of human activity and its designs on nature thus has deep cultural roots in the West, even if its destabilizing potential has been realized much more recently.

The great acceleration

The sense of crisis that pervades writing in environmental history reflects concerns in a growing number of disciplines that there have been both qualitative and quantitative shifts in human impacts on the Earth during the Anthropocene period. This term was coined around the turn of the present century to describe the recent era of environmental change in which people have become global change agents in their own right. There is some debate about an appropriate starting date for the Anthropocene. But there is broad agreement that it was during the late eighteenth century that fundamental alterations in the relationship between people and the environment were set in place, as the Enlightenment encouraged the development of scientific and technical knowledge, and industrial culture. It is in the second half of the twentieth century, however, that levels of resource use took off, with every indicator of human activity undergoing a sharp rise from about 1950. This increase is known as the “great acceleration” (Steffen et al. 2011).

Human population reached 1 billion around 1820, but over the next two centuries it climbed precipitously to more than 7 billion, with energy use growing about 40 times and economic output about 50 times. The upturn began in northern Europe in the eighteenth century with the development of capitalist forms of production and the spread of private means of landownership. Agricultural experimentation allowed more people to be fed in rapidly growing towns, and gradually improving nutrition enabled them to survive and to live longer. Simultaneously, the energy bottlenecks of the preindustrial world were broken with the increasing use of coal. Preindustrial enterprises had relied on the power of human and animal bodies and inefficiently harnessed wind and water; new capitalist firms were powered by machines capable of converting into mechanical energy the biomass stocks that had accumulated over millions of years of geological time (McNeill 2000).

These changes depended on a new conception of land and of the ownership of resources. As formulated by English liberal writers, but usually ascribed to the late-seventeenth-century author John Locke, the right to property depended not on custom but on human toil. Resources held in common, or for which no ownership regime was deemed to apply, such as indigenous territory, were considered by the application of personal labor to be “improved,” or removed from a state of nature. This was the essential condition for private ownership, which in turn was marked on the ground by the process of enclosure, and protected in the legal domain by right of personal security. English law, in contrast to Roman law or Chinese and Islamic custom, did not subject these rights of ownership to countervailing social rules; instead it gave liberty to use the land’s resources and legal protection to the private owner (Linklater 2014). In these ways, the holders of enclosed lands were guaranteed the yield of their agricultural improvements as well as the right to exploit the energy stocks that lay beneath them. The resulting surplus enabled the spatial extension of capital beyond Europe into the neo-Europes of the Americas and Australasia. New fields of resources from which to profit were opened up to produce huge quantities
of food and oil. The private property calculus thereby underwrote an immense expansion of enclosure, “improvement,” energy exploitation, and population growth.

This dynamic explains how the drivers of the Anthropocene initially occurred in very uneven ways across the globe. But if the surge in human numbers began with European populations, by the 1950s these had passed through a demographic transition from high to low fertility and mortality rates. Paradoxically, it was at this point that the great acceleration began. Consumption levels soared, and with this demands for natural resources and levels of waste generation. The Western model of development was exported to other parts of the world as surplus capital sought wider arenas for investment, along with modern health-care systems and a “green revolution” in food production in Asia, Latin America, and Africa. With populations beginning to age in Europe and North America, the number of human beings in the world boomed as developing nations moved into the expansionary stages of the demographic transition. In 1950 the population of Africa was about half that of Europe, whereas today they are about equal, and by 2050 Africa’s population is projected to be three times that of Europe.

If it was the Industrial Revolution and the international expansion of capital that led to the dramatic shift in human relations with the natural world, then it was the great acceleration after 1950 that normalized it. An ever-expanding global population expressed growing wants as well as needs, consuming resources and producing waste in ways never before seen. The rise in the use of inanimate energy, notably fossil fuels, illustrates this. It has tracked a more than fivefold increase in gross world product since 1960, and is directly reflected in the rise of carbon dioxide in the atmosphere. Emissions of methane, an even more potent greenhouse gas, have risen faster still. It is no coincidence that the term “Anthropocene” was popularized by Paul Crutzen, an atmospheric chemist (Steffen et al. 2011). In parallel, carbon sinks have been undermined as tropical rainforests are converted to farm and plantation land to supply products such as palm oil that are ubiquitous components of food manufactured on industrial scales. Today, therefore, we “live in a global system in which our most critical problems go well beyond regional and national borders” (Costanza et al. 2007, 522).

It is climate change that has brought into focus the ability of humans to reshape their environments not just in immediate and discernible ways but, rather as George Perkins Marsh feared, at the scale of the Earth as a whole. Humans have significantly altered other biogeochemical cycles, such as those for nitrogen and phosphorous; strongly modified the terrestrial water cycle through intercepting river flows to the sea; have caused widespread land cover change; and are “likely driving the sixth major extinction event in Earth history” (Steffen et al. 2011, 843). These changes are a product of the need for food and resources for growing populations that are simultaneously increasingly urban. In 2007 the number of people living in towns and cities worldwide moved into a majority for the first time. The logistical and infrastructure networks of city systems have been built to supply the growing appetite for material consumption that accompanies urbanization.

The imprint of human enterprise on the Earth–atmosphere system is therefore now unmistakable, and shows no signs of moderating as the locus of growth in economic activity shifts from Europe and what were its overseas settlements to the countries of East, Southeast, and South Asia in particular. China’s carbon emissions exceeded those of America in about 2007, and by most measures its economy will
be the larger of the two within a few years. But such observations are still usually made outside of an environmental frame of reference, and “much discussion of human–environmental interactions continues to lack a long-term, temporal dimension” (Costanza et al. 2007, 522).

One environmental historian summed up the history of the twentieth century in saying that the “human race, without intending anything of the sort, has undertaken a gigantic uncontrolled experiment on the earth” (McNeill 2000, 4).

**Environmental risk and hazard**

It is a paradox that, as people have seemed to move further from nature with the making of what is sometimes called “second nature,” or nature reshaped by human action, we have created what Anthony Giddens (1999) has termed a “runaway world.” The economic, cultural, and technological processes of globalization simultaneously generate highly interconnected problems. This is an essential difference between social experience today and that before the Anthropocene. In the past, when civilizations were threatened or collapsed, they did so with few, if any, implications for people living elsewhere on the planet. That is no longer the case. The distinguishing feature of industrial culture is that it has been framed by risk and opportunity. The embracing of risk has been the driving force for the creation of wealth in modern economies. Risk management has been the means by which industrial societies have endeavored to determine their own futures, rather than leaving things to the forces of tradition or of nature. They have invested in future opportunity, through calculations of profit and loss, at the same time as seeking trade, resources, and capital accumulation in distant places. With this has come a new set of regional, and even global, environmental problems.

This forward-looking, interdependent world has therefore not eliminated risk. Rather, the nature of risk has changed, from what Giddens describes as “the predominance of external risk to that of manufactured risk” (1999, 26). If external risks are those coming from nature, then manufactured risks are those generated by the increasingly assertive and widespread patterns of capitalist activity that so concerned Marsh. A good example comes from studies of nineteenth-century Western agricultural expansion into the lands of the European new world. Often such areas, in the western United States, Canada, and Australia, were resettled by people driven from Europe by external risks such as famine, hunger, and disease. In these new environments, they faced new risks. The story of the American Dust Bowl in the 1930s, for instance, “is less about the failures of nature than about the failures of human beings to accommodate themselves to nature” (Cronon 1992, 1348). The progressive assumptions of a forward-looking people, that the “wilderness” would inevitably yield to linear “improvement,” came into conflict with the cyclical environmental rhythms of these dry grassland areas.

A classic of geographical writing in this vein is Donald Meinig’s (1962) study of the wheat frontier in South Australia, *On the Margins of the Good Earth*. Much of the area inland from Adelaide was “mallee country,” a scrub forest with thick shallow roots and a low– to medium–height canopy. Mallee clearance required a big investment of time and work, taking the chance that the effort would pay off in the future. The environment, however, was uncertain and unpredictable, and in 1865 the surveyor–general G.M. Goyder was sent through the outlying districts to identify and demarcate those pastoral areas then in drought. What became known as Goyder’s Line was
taken subsequently to represent and formalize the outer limit of agriculture, and it came under severe pressure from settlers wanting to move beyond it when wetter years returned in the 1870s. But they were confounded by the recurrence of severe drought in the 1880s.

Attempts to transform climatically volatile regions into cropping areas were driven by the opportunities of feeding new industrial cities. A prominent theme in environmental history has been the problems and politics of urban pollution from industrial and infrastructure development. This is a clear case of manufactured risk. There have been increasing attempts to manage pollution in Western, and now Asian, cities during the great acceleration. These have been characterized as the product of a change in environmental values as higher standards of living have prompted a search for improved environmental quality and higher levels of amenity (Hays 1987). The search has had varied outcomes, not least a growing focus on air and water management, and the paradox of highly consumptive but outwardly clean suburbanization.

Urban environmental issues of this sort, although experienced by cities throughout the world, are essentially the product of local-scale activities and are amenable to local solutions. The more intractable sources of manufactured risk are those expressed at much bigger spatial scales. A good example is the threat to coastal cities from tidal flooding and extreme weather events as climate change becomes more pronounced, at the same time as more people seek to live or develop trading facilities by the sea. The cities most in jeopardy are those in Asia, such as Shanghai and Mumbai, and those on the Gulf and East Coasts of the United States. Hurricane Katrina, which devastated New Orleans in 2005, and Hurricane Sandy, which badly affected New York City in 2012, are widely seen as warnings. Often this is imagined to be a case of external risk rather than one of manufactured risk on coastlines made vulnerable through human action.

In an analysis of the New Orleans situation, Freudenburg et al. (2009) take a particular view of this issue. The title of their book describes Hurricane Katrina as a “catastrophe in the making,” in which the “growth machine” of industrial capitalism set in train destructive processes that led to an inevitable outcome. These included the extension of the city into low-lying, flood-vulnerable districts; the driving of the main export canal through the Mississippi Delta into the teeth of Gulf hurricanes; and, critically, the damaging intrusion of oil industry infrastructure and wastes across the deltaic wetlands, undermining their buffering functions to storms. For them, “New Orleans is by no means unique in the ways in which local leaders have increased their communities’ vulnerability to ‘natural’ disasters” (Freudenburg et al. 2009, 11). They see the story of Katrina as a parable of the consequences of manufactured risk. It also highlights the key difference between the modern preoccupation with short-term decision-making and environmental history’s focus on long-term analysis.

Stories in environmental history

The most common form of environmental history is narrative, or the organization of understandings about people and environment into story form. Storytellers select certain events, processes, and places to emphasize, and particular relations to highlight. At the same time, other players and connections are downplayed, and “the discontinuities, ellipses, and contradictory experiences that would undermine the intended meaning” of the story are hidden (Cronon 1992, 1349–1350). The most effective environmental history works by telling stories about stories about nature.
The analysis of the impact of Hurricane Katrina on New Orleans illustrates this point. Freudenburg et al. (2009) argue that engineering evaluations of the causes of that extreme event, including that of the US Corps of Engineers which had built the walls and levees that failed, avoided any attribution of blame to the construction of the main canal through the delta. The consequences for the wetlands were ignored, including the destruction of large areas of swamp cypress that had earlier mediated the impact of floods, as was the “levee effect,” or the acceleration of urban construction behind flood defenses that were assumed to give greater security than they could ever have done. Instead, for the engineers, the disaster was the result of structures that performed inadequately. The larger point is that the story that frames this engineering perspective is one that, like the “growth machine” that supports it, embodies Enlightenment notions of progress and “improvement.”

William Cronon (1992) identifies two main “plot lines” for stories in environmental history, illustrating these with a range of narrative accounts of settler encounters on the Great Plains. The most famous of the progressive narratives in this context is the frontier thesis of Frederick Jackson Turner, which described a rising arc of landscape transformation from wilderness to trading post, farm, and boomtown. Turner’s account, unproblematic in that it did not highlight struggles and setbacks, was followed by many other progressive story lines that did, in which the narrative role of such problems was “to play foil to the heroes who overcame them” (Cronon 1992, 1353). At the same time, there have been competing stories that trace more negative outcomes, stories that are tragic or declensionist in form. Meinig’s (1962) interpretation of the South Australian frontier is a story about both types of story – about eager settlers in times of rain, and anxious administrators aware of cyclical drought.

Such narratives often silence other human or nonhuman actors. Since the mid-1980s there has been some correction of the omission of indigenous voices from these accounts; experiences from the other side of the frontier are now more conspicuous than before, as are environmental histories of non-European parts of the world. This has been matched by the use of different forms of narrative, including critical engagements with cartography and photography. There have been experiments with re-photography to illustrate the nature and extent of change over time, and also with environmental art. The Strata project, about Puritjarra in western central Australia, engaged Aboriginal artists with an artist of Australian settler meanings, alongside an archaeologist and an environmental historian. Their purpose was to uncover different kinds of knowledge and ways of knowing place and, through co-understanding, to access something of the deep time that is lost in Eurocentric stories of colonizer experience (Martin, Robin, and Smith 2005).

The future

In his reflection on the value of stories, William Cronon suggests that “the task of environmental history is to assert that stories about the past are better … if they increase our attention to nature and to the place of people within it” and encourage us to look at these in new ways (1992, 1375). Such stories are told at a range of scales, from the local to the global. The canvas is both ideas and materialities, or how people know nature as well as the material outcomes of their interactions with it. The aim is to understand environmental changes in the long run, such as the great acceleration, and to analyze how
past variability in natural systems exposes the assumption that has held sway for much of the Anthropocene: that conditions favorable to human activity will always prevail.

The best environmental history therefore reveals the consequences of the spread of liberal capitalism and the unexpected opposition that its short-term futurism has prompted from the planet itself. At this juncture, rather than “grit our teeth and hope for a miracle,” as Bruno Latour (2013) puts it, “we could inquire into what this modern project has meant so as to find out how it can be begun again on a new footing.” This raises two points that only a few environmental historians (e.g., Dovers 2000) have attempted to explore. First, what does an understanding of past environmental transformations mean for the analysis of society today, its priorities, structure, and agency? And, second, how can such knowledge be translated into solutions and policies? These questions pose extended interdisciplinary challenges for environmental history. But doing the work toward answering them may encourage a wider and more reflective social engagement with nature, in keeping with Marsh’s view of the human being as earth’s “noblest inhabitant.”

SEE ALSO: Anthropocene and planetary boundaries; Berkeley School; Globalization; Historical geography; Imperialism; Landscape; Natural hazards and disasters; Population and natural resources

References


Environmental impact assessment

Richard K. Morgan
University of Otago, New Zealand

Environmental impact assessment (EIA) is a structured process for considering the implications of proposed actions for people and their environment, while there is still an opportunity to modify (or even, if appropriate, abandon) the proposals. In principle, it can be used at all levels of decision-making, from policies and plans through to specific projects; in practice, project-level application has dominated its use around the world.

Brief history of EIA

The institutionalized forms of EIA now so evident around the world had their origins in the late 1960s with the enactment of the National Environmental Policy Act (NEPA) in the United States. The requirement to produce an environmental impact statement (EIS) was a late addition to the act, to provide an enforcing mechanism that ensured federal agencies showed explicitly how they were implementing the environmental policy in their own major policy and project initiatives. The very process of producing an EIS achieved its own identity and was adopted, within a few years, as an environmental management tool by a number of other countries, including Australia, Canada, and New Zealand. The following decades saw the process, by then more generally known as environmental impact assessment, spread to more and more countries. Following the 1992 United Nations Conference on Environment and Development (UNCED), all UN agencies developed internal impact assessment procedures; this was a significant but perhaps relatively unheralded expansion of the reach of the process. In addition, impact assessment has gained greater explicit recognition in international law, the more notable being the Convention on Transboundary Environmental Impact Assessment; the Convention on Wetlands of International Importance; the Convention on Access to Information, Public Participation in Decision-Making and Access to Justice in Environmental Matters; the United Nations Framework Convention on Climate Change; the United Nations Convention on the Law of the Sea; and the Protocol on Environmental Protection to the Antarctic Treaty. Whether through national statutes or international legal instruments, impact assessment procedures are now recognized by virtually every member country of the United Nations (Morgan 2012).

A significant milestone in recent years has been the development of the Equator Principles (EP) to establish guidelines on the use of EIA by private-sector financial institutions. As funding of major development projects by the members of the World Bank group declined significantly as a proportion of total global funding, it became obvious that the bank’s Safeguard Policies were in danger of being marginalized. Therefore, in 2002 the International Finance Corporation (IFC) met with several major banks and initiated discussion that led to the launch of the Equator Principles in June 2003; by 2014, 81
private-sector financial institutions had signed up to the principles. The EP are based on the IFC’s social and environmental performance standards, and the World Bank’s environmental and health and safety guidelines, all of which give EIA a prominent role in the decision-making process.

Nature and purpose of EIA

The purposes of IA are:

- to provide information for decision-making about the biophysical, social, economic, and institutional consequences of proposed actions;
- to promote transparency, and the participation of the public, in decision-making;
- to identify procedures and methods for the follow-up phase (e.g., monitoring and mitigation of adverse consequences) in policy, planning, and project cycles; and
- to contribute to environmentally sound and sustainable development.

The process of IA benefits proponents, stakeholders and local communities, and decision-makers. Although it has been criticized since its introduction for being a technocratic tool, based on rational decision-making models (Richardson 2005; Weston 2010), much of the contemporary literature on IA principles tends to emphasize a participatory, inclusive approach, which recognizes different types of knowledge and the importance of representing the views of different groups in society, regardless of their economic and political status (e.g., Spaling, Montes, and Sinclair 2011). As such, environmental and social justice considerations, including the consideration of human rights, are important guiding principles of impact assessment (Kemp and Vanclay 2013).

In recent years, IA has become strongly linked to sustainability concepts, and some would argue that sustainability is an integral part of all IA activities (Gibson 2013). On the other hand, it is possible to conduct impact assessments within
other policy contexts, without direct or indirect reference to sustainability thinking.

**Practice of EIA**

EIA involves the identification and characterization of the most likely impacts of proposed actions (impact prediction/forecasting) and an assessment of the social significance of those impacts (impact evaluation). Most methodologies break these two basic components into a series of steps comprising some or all of the following:

- **Screening**: Should an impact assessment be carried out? Many countries use lists of activities that require an EIA, perhaps supplemented by lists of those activities that may require an EIA if they meet certain size/capacity characteristics.
- **Scoping**: The core of the impact assessment process, scoping involves the identification of significant potential impacts and establishing a work program to investigate these further.
- **Impact prediction**: The phase during which potential impacts are investigated by appropriate methods, to assess their likelihood and the consequences of their occurrence. Methods will vary according to the impact type (e.g., social versus ecological), and also according to cultural context; indigenous environmental knowledge can often provide valuable insights based on a long history of resource usage that inform predictions, which also embed social and cultural values.
- **Significance evaluation**: The impacts that are likely to occur need to be evaluated for their social significance. Many systems leave this to the public submission process during formal decision-making, but both monetary and nonmonetary evaluation methods can be used within the EIA process to capture significance.
- **Impact management provisions**: Significant impacts will require some form of response to avoid or mitigate the impacts. This may be through design changes to a proposal, by instituting measures to protect people and/or the environment, or by compensating affected parties. Managing impacts through the life of the proposal is an important part of EIA, so the development of impact management plans, together with monitoring provisions to ensure compliance and the effectiveness of those plans, is critical to the whole process.
- **Reporting/communication**: Effective communication of the information generated through the EIA to the people who need to use the information is vital. Potentially affected communities and other stakeholders need the information in a form that enables and empowers them to participate in decision-making processes. And decision-makers need the information in a form that allows their decisions to be fully informed.

Some jurisdictions use a narrow definition of “environment” to mean just the natural environment; others consider environment to also include people and their activities and structures. The trend in the international research and practitioner community, as exemplified by publications from the International Association for Impact Assessment (IAIA), is to follow the second of these interpretations.

An important stimulus for the development of EIA was the failure of existing procedures (typically dominated by cost–benefit analysis) to anticipate the less obvious environmental consequences of developments. Prominent examples from the 1960s include the effect of
the agricultural pesticide DDT on bird-of-prey populations (biomagnification of DDT residues resulting in fragile eggshells and hence decreased reproduction rates), publicized by Rachel Carson in *Silent Spring* (1965); and the increased incidence of schistosomiasis among local people in tropical environments due to changes in water snail habitat following the construction of hydroelectric dams (Edington and Edington 1977). In both examples, dramatic impacts resulted from changes that had not been anticipated and that involved a series of intermediate stages before the significant impact was manifest.

Significant direct impacts of major activities are usually well known and, in most cases, environmental engineering solutions have been developed. For example, modern dams usually have structures (e.g., fish ladders) to allow fish migration upstream to continue once the dam is completed. The more problematic issues are the indirect impacts, which typically result from more complex cause–effect pathways, and may be separated in space and/or time from the original action. An important purpose of EIA, then, is to recognize the possibility of indirect impacts and to attempt to predict what they might be, and the likelihood and implications of their occurrence.

Another important consideration is cumulative impact. This concept recognizes that, in many situations, other previous activities (typically projects) have already created a legacy of effects on the local environment. If a further activity is allowed to proceed, the analysis of cumulative impacts considers how the impact of the proposed activity will add to and perhaps interact with existing conditions. Given the complexity of such questions, it is unsurprising that cumulative effects assessment is generally seen as one of the more problematic and least developed of the impact assessment methods. There is also a view that such considerations are best managed not on a project-by-project basis but at a regional or national level by setting limits to development or establishing some form of carrying-capacity thresholds on resource use.

Public participation is now given prominence in EIA writing. Many evaluations of public participation in EIA have identified similar, largely mechanistic, procedural problems, including proponent reluctance to engage with local communities, poor public understanding of statutory processes, barriers that limit public access to information, and lack of decision-maker sympathy for public involvement (Hartley and Wood 2005). In contrast, more recent studies have tended to place greater emphasis on evaluating practices against the expectations of deliberative democracy or collaborative planning (Petts 2003). The postrationalist thinking of recent years emphasizes the development of constructive relationships between public, proponents, and decision-makers, and changes in the structure and power relationships of decision-making (Diduck *et al.* 2007). However, social, political, and cultural settings and traditions are important determinants of the development and practice of these processes, especially in those countries where such practices are still new. Therefore, concepts of social and organizational learning are also important considerations in evaluating the progress with, and nature of, participatory processes in impact assessment (Sinclair, Sims, and Spaling 2009).

**Forms of EIA**

Ideally, impact assessment will be an integrated treatment of the likely impacts on any of the facets of the natural and human-modified environment, including people, communities, and their wellbeing. However, different sectors of the environment can pose rather different methodological and technical challenges for
impact assessment, resulting in communities of practitioners who focus on the use of specific methods for the assessment of particular impact phenomena. Any given proposal may require a different array of impact assessment methods, reflecting the range of potential impacts associated with the proposal. The challenge for the impact assessor is to bring all that information into an integrated assessment.

Under the umbrella of EIA a number of specific forms have become firmly established since the 1970s, including social impact assessment (SIA), ecological impact assessment (EcIA), health impact assessment (HIA), and strategic environmental assessment (SEA). To some extent, each tends to have arisen through some level of dissatisfaction with EIA as it has been practiced – SIA, for example, developed strongly in the late 1970s and 1980s as EIA, especially in the United States, was considered to have a strongly biophysical emphasis and to neglect social impacts. SEA has been vigorously promoted as a way to extend impact assessment to higher-level decision-making at policy, program, and plan levels, a reaction to the project orientation of most EIA applications.

Theory and EIA

Dissatisfaction with the lack of serious critiques of EIA as a process gained momentum in the late 1990s, as the influence of debates in related disciplines finally began to reach the impact assessment community. The resulting literature has explored the historic roots of EIA in the rationalist decision-making model, and examined the way critiques of that model in related fields such as planning, environmental management, decision theory, political science, sociology, and economics have provided alternative ways to think about EIA.

One important source of thinking represented in subsequent literature has been the link to models and theories of planning and decision-making. Bartlett and Kurian (1999) identify six models they consider to have been implicit in discussions of EIA in the literature:

- the information-processing model: essentially the rationalist, decision-support model;
- the symbolic politics model: EIA is used to suggest accordance with certain values but not necessarily for holding to those values;
- the political economy model: EIA is used by the private sector to reduce financial risk, and if possible increase financial opportunities by internalizing environmental externalities;
- the organizational politics model: changes occur in the internal politics of those organizations required to use EIA;
- the pluralist politics model: EIA process is used to open opportunities for negotiation and compromise between different interest groups; and
- the institutionalist politics model: political institutions are changed significantly by the effect of EIA on values, actions, and perspectives in their policy-making processes.

Leknes (2001) uses a simpler threefold categorization of decision-making approaches: the rational, the new institutionalist, and the negotiation perspectives. The common theme in these and similar contributions is the critique of the rationalist model of planning/decision-making, and by implication of EIA, and the exploration of alternative models that embrace new thinking about planning and decision-making processes in their wider social, cultural, political, and economic contexts.

The rationalist model rose to dominance after 1945 in organizational and decision-making thinking, and is characterized as having a strong
ENVIRONMENTAL IMPACT ASSESSMENT

technical emphasis, with planners and other professionals acting as neutral processors of information, producing independent evaluations of alternatives to be provided to the decision-makers. As the basis of the rational comprehensive planning theory which dominated strategic and development planning in many Western countries in the 1960s and 1970s, this model has been the subject of significant criticism. One theme has been the impossibility of recognizing all possible alternative solutions, from which to select the “best” solution; hence, more constrained and practice-informed models of the rationalist approach have emerged (such as the bounded rationality model, the incrementalist model, and various other pragmatist models). However, they still carry the rationalist imprint, and they have continued to attract criticism for their failure to recognize the political and value-based nature of decision-making, and especially issues related to the use of power (Richardson 2005). This has encouraged the promotion of deliberative and collaborative approaches to planning and decision-making processes, drawing on the work of Habermas: bringing stakeholders and communities into the processes, emphasizing the importance of communication as a means of negotiating a consensus solution that captures the values of those participants, and moving the professional technocrats from a controlling to a facilitating role in the decision-making process.

These collaborative and consensus-oriented models have themselves been criticized by those influenced by Foucault for not recognizing sufficiently the issue of power relations between participants, which will inevitably affect the ability of different groups or individuals to enter social negotiations in an equitable way (Richardson 2005).

Evidence suggests that the center of gravity of EIA thinking, and especially practice, is still firmly rooted at the rationalist end of that spectrum (Morrison-Saunders and Sadler 2010). But it is important to recognize that EIA thinking since the beginning of the twenty-first century has been affected by wider debates more clearly seen in the major disciplines, and that these ideas are being explored and reinterpreted in ways that will encourage EIA practitioners to reflect on their practices.

The prevailing view in the theoretical literature would seem to be that EIA is in essence a rationalist, information-based process that is used to support decision-making, but that the value dimension of the process has to influence the way impact assessment is designed and carried out, and that the context within which EIA is used must be recognized as influencing its effectiveness. In particular, wherever significance judgments are to be made, the process has to accommodate the values of those potentially affected by the proposed activity, and that must include as a minimum, for example, the scoping phase as well as the impact evaluation phase. How this involvement should be managed to promote inclusion and ensure equitable participation in a socially and culturally appropriate way has been the subject of a burgeoning public involvement literature in recent years. It reflects an increasing awareness of and sensitivity to, among other things, the inherent power relations found in rationalist decision-making processes, so one challenge has been to develop new approaches to impact assessment that address these problems (Morgan 2012).

Effectiveness of EIA

The final report of the International Study on the Effectiveness of Environmental Assessment in 1996 concluded that, while EA had made its mark since it was introduced 25 years earlier, it would be necessary to maintain the
efforts to improve its performance if it was to make a substantive contribution to the goal of sustainable development (Sadler 1996). The theme of effectiveness of EIA has been ever present in the literature since then, but as Cashmore et al. (2004) observe, the bulk of that literature addresses procedural issues, with a much smaller proportion concerned with substantive issues. Both are important parts of the overall assessment of effectiveness but, while the procedural aspects are more amenable to study and analysis, substantive considerations raise more difficult questions.

Where EIA has been incorporated into environmental or resource management, or planning legislation, it supports the wider policy aims of those statutes. Effectiveness then has to be judged in large part in those terms, and will vary from country to country. Yet there are still criticisms of EIA for falling short of other aims, such as sustainability goals. The wider policy process within which resource exploitation takes place must be the focus of scrutiny and criticism, not the enforcing mechanism per se. Using the political models of EIA suggested by Bartlett and Kurian (1999) – or indeed the various planning or decision-making models – effectiveness can be seen from a number of different, politically oriented perspectives. Has the process opened opportunities for local people to be more involved in decision-making? Have companies become more aware of environmental issues through EIA and modified their practices accordingly to gain competitive advantage? Has change been brought about in government bodies dealing with, say, natural resources, to internalize EIA thinking? Do decision-makers and other stakeholders understand and use the EIA information provided to them?

SEE ALSO: Community-based natural resource management; Decision analysis; Environment, democracy, and public participation; Environment and development; Environmental assessment techniques; Environmental (in)justice; Environmental management; Environmental planning; Environmental regulation; Indigenous knowledge; Multicriteria decision-making; Participatory development; Social justice; Sustainable development

References


ENVIRONMENTAL IMPACT ASSESSMENT


Further reading


IAIA (International Association for Impact Assessment) wiki: http://www.iaia.org/iaiawiki/.


Public understanding of environmental issues does not draw simply on scientific facts and expert information, but on a range of familiar knowledges, personal values, commitments, and responsibilities. People learn about environmental issues through everyday social practices at work, at home, and at school as well as through media reporting on local and global problems. And “the public” are not a homogenous, undifferentiated lump, but differ greatly in terms of age, gender, class, race, outdoor activity levels, educational attainment, and geographical location. This means that public understanding of environmental issues is complex as people can hold multiple beliefs at any one time; beliefs that can be contradictory, in opposition, or scientifically impossible.

Learning about environmental issues can happen in situ and ex situ. In situ learning takes place in the environment that is being studied, perhaps as part of a fieldwork program or during a holiday or hobby activity. For example, Figure 1 shows a park ranger in Yellowstone National Park, USA, explaining to visitors how geysers work and how they are damaged by people throwing in litter. Public understanding may thus develop through embodied practice and sensory experience of unfamiliar environments while a tourist, or of familiar environments while walking to work or school as part of everyday life. Ex situ learning takes place outside the environment being studied, for example, in classrooms through formal education or more informally in homes when people read about environmental issues, watch television documentaries about wildlife, or surf the Internet for fun.

One challenge for public understanding is that the global scale of many environmental issues, such as climate change, make them more difficult for members of the public to grasp and engage with than more local, more visible, and everyday issues, such as taking more environmentally friendly forms of transport to work.

Another challenge is that discussion of public understanding remains largely in the Global North, with little attention paid to this important question in the Global South. Postcolonial geographies of public understanding are required to avoid an Anglo-American focus on global issues. The public understanding of environmental issues is thus inherently geographical in both the ways that it is influenced by context – environmental, social, cultural, and personal influences on the public – as well as the ways in which those environmental issues will influence people’s lives into the future.

Science, the media, and pressure groups as influences on public understanding of environmental issues

Public understanding of environmental issues also depends upon claims made by powerful actors
ENVIRONMENTAL ISSUES AND PUBLIC UNDERSTANDING

Figure 1  Park ranger and visitor near Old Faithful geyser, 2002. Photo by Sally Eden.

about what issues matter and why. Specialists or elite groups such as scientists, businesses, governments, nongovernmental organizations (NGOs), and pressure groups are powerful actors in identifying and articulating environmental issues through making claims about their science, severity, and possible solutions.

Many environmental issues originate in scientific work, which often identifies and measures the issue, for example, stratospheric ozone pollution or climate change. This is one reason why opinion polls often report that the public trust scientists more than any other profession in society, including business and government, to tell the truth about environmental issues. Because of their specialized training and expertise, scientists are often regarded as being knowledgeable about environmental issues and as having solutions or answers to environmental problems as well as pursuing facts, data, and truth for its own sake without being affected by the need to make a profit as a business might do.

Scientists and scientific writings, such as journal articles or reports, can be particularly important in influencing public understanding of environmental issues because many environmental problems cannot be perceived by ordinary people, maybe because invisible pollutants are involved that can only be detected with specialist equipment and/or their effects occur beyond normal human perception. For example, in the 1980s, stratospheric ozone depletion many kilometers up in the skies above the Antarctic was detected using specialist monitoring and computing equipment and climate change today may be invisible to the average person because the effects are often felt far away from home and the timescales of change can be longer than a human lifetime.

But the success of scientists in influencing public understanding of environmental issues depends also upon media coverage of their work because very few ordinary people encounter science directly. Many environmental issues are therefore communicated to the public through media coverage of science; coverage that often tags them with memorable names as well, such as “the ozone hole” or “global warming.” In the case of stratospheric ozone depletion, there was rapid and unequivocal coverage in the mass media of the thinning of the ozone layer over the Antarctic using NASA data and visually striking maps. This led to a rapid rise in public awareness and, subsequently, international policy to reverse the damage. In contrast, in the case of climate change, public understanding of climate change has been more difficult to build with different scientific claims made for and against the evidence for human-induced climate change at a global scale.

In addition, media coverage of environmental issues is diverse. News reporting, representations of nature in the form of television documentaries about wildlife made by Disney or David Attenborough, movies about environmental catastrophes from *The Day After Tomorrow* to *Wall-E*, and even adverts that use images of lush green landscapes to promote corporate reputations can all
contribute to public understanding and appreciation of environmental issues (e.g., Burgess 1990).

Pressure groups and nongovernmental organizations (NGOs) are also important in raising public understanding of environmental issues. In the past, these groups used newsletters, rallies, petitions, and, especially, symbolic embodied protests, such as Greenpeace sailing dinghies near whaling ships, which were then communicated to a mass readership through media coverage in newspapers and television news. Today, social media and self-publishing on the Internet mean that NGOs and campaigning organizations of all kinds can publish their awareness-raising materials themselves in order to communicate more directly with the public. As with scientists, opinion polls often report high levels of public trust in what NGOs say about environmental issues, usually much higher than the trust given to business and government as providers of information about environmental issues.

These processes of making and judging claims to knowledge can have a big effect in terms of public understanding and behavior. For example, research studying two communities in the United Kingdom and Netherlands revealed that an inability to establish the legitimacy of expert claims, confusion around environmental facts, and a distrust of government led to a reduction in people taking personal responsibility for environmental issues (Burgess, Harrison, and Filius 1998).

Public understanding of science (PUS) and environmental issues

Research into the public understanding of environmental issues has been informed by work on the public understanding of science (PUS) more generally in the social sciences. This work began in the 1980s when the public understanding of science was seen as limited because scientists were regarded as knowledge-holders or “experts,” whereas ordinary or “lay” people were regarded as holding little or no scientific knowledge. Scientists, policymakers, and governments commonly thought of ordinary people as having insufficient knowledge about science to be able to understand the debates on how to govern it, with public perceptions of new technology or scientific claims often being dismissed as uninformed, misunderstood, or even nonsensical.

The assumption was often made that the public should be given more information so that they could not only understand new science and its risks for human health, for example, from vaccination, genetic modification, or stem cell research, but also learn to support decisions about how new science and its risks were governed. This argument also emphasized “scientific literacy,” that is, the level of knowledge that people have about basic scientific principles, such as laws of motion, energy, and genetics. Again, the assumption was that, without this basic level of scientific literacy or knowledge, people would be unable to debate the issues that arose around new scientific findings or technological applications and unable to make informed decisions.

Social scientists working in PUS were critical of these assumptions, referring to the whole approach as the “deficit model,” because it assumed a knowledge deficit in public understanding of science and because it kept (supposed) experts and nonexperts apart. They argued that this model failed to comprehend fully how ordinary people understand science and the consequences of that assumption on their behavior and feelings about scientific applications and governance. These arguments were applied particularly to environmental issues such as climate change and genetically modified foods in the 1980s and 1990s.
The deficit model was challenged by alternative models, such as that of “post-normal science” (Funtowicz and Ravetz 1993), which argued that environmental issues such as climate change or genetic modification were too high risk and too uncertain to be dealt with through traditional, “normal” scientific methods. Instead, such issues required science not only to change itself by moving into a “post-normal,” reflexive mode of thinking about its own assumptions, but also to change its relationship with others to a more participatory, dialogue-based model, which would include “lay” public in debates. This “extended peer community” would include a wide range of participants beyond the traditional scientific “experts.”

Such alternative theorizing can be seen as part of a democratic turn toward promoting the participation of citizens in science and environmental issues, as well as part of critiquing (and thus attempting to replace) the deficit model. Scientists and other elites also criticized such proposals in turn, fearing the dilution of science’s traditional authority by forcing scientists to share it with what they saw as less (or un-)qualified publics, and also as politicizing issues that should remain outside the realm of politics.

Lay knowledge

Academics and others who rejected the deficit model have been arguing for some time that public understanding of environmental issues can be as useful and valid as scientific or expert knowledge. Ordinary people are argued to have important “lay knowledge” that can feed into understanding of environmental issues, knowledge that arises not from formal education or training, but from everyday life and experiences.

For example, people are often very knowledgeable about their local environments, but because such knowledge is not usually framed in the objective and rational way expected by science, it has often been discounted as worthless or at least unreliable. Recent social science work has, however, begun to re-value such knowledge, from using a person’s weather diaries to reconstruct climate change in the nineteenth century to incorporating local people’s folk memories of flooding into scientific models of hydrological management (e.g., Lane et al. 2011). Such local or “lay” knowledge will also take into account the emotional and subjective encounters, responses, and debates surrounding environmental issues as part of the context of learning and evaluating environmental change.

People may also have specialized knowledge about everyday practices because of their work or hobbies. A famous example analyzed by Wynne (1996) is that of sheep farmers in Cumbria, northern England. Following the Chernobyl nuclear accident in 1986 in Ukraine, the UK government banned the movement and sales of sheep from key areas, thus threatening farmers’ incomes, already fragile in a precarious industry. Government decisions were largely based on scientific modeling of the movement of radio-caesium and its movement through vegetation, soil, and the bodies of sheep grazing on the hills, modeling which initially predicted that contamination of sheep would only last for three weeks. However, when measurements of the sheep showed that contamination lasted far longer than predicted, the ban was extended from three weeks to become indefinite and there was also talk of a widespread cull. The modeling had assumed that clay soils were prevalent in the grazing areas, whereas peaty soils also existed and these mobilized caesium back from the soil into the grass upon which the sheep continued to graze. In contrast, the farmers argued that the contamination was not solely due to the fallout from Chernobyl in Ukraine, but from normal...
operations of the nearby nuclear power plant at Sellafield in England.

Wynne argued that the problem was not only that the scientists claimed complete certainty over their environmental modeling, a certainty that was clearly unrealistic, but also that they refused to listen to the farmers’ arguments, arguments based on local knowledge about where sheep grazed in comparison to where the highest levels of contamination had been measured, or about how the metabolism of sheep would respond to being penned up during scientific experiments. He concluded that the two “knowledge cultures” – lay farmers and professional scientists – made very different assumptions about agency, control, and predictability and also valued each other’s forms of knowledge very differently.

As a consequence of this kind of analysis, researchers have since argued that lay knowledge, often built up tacitly, without professional qualifications, without being written down or peer-reviewed, but instead embodied in everyday lived practices and experience of people’s local environments, as well as through folk memories, nevertheless has value in understanding environmental issues.

However, a tension between scientific knowledge and lay knowledge about the environment remains. Scientific knowledge, even where it is not strictly “universal,” is still assumed to be generalizable and big-pattern thinking, whereas laypeople’s knowledge is assumed to be much more narrowly scaled or piecemeal in its spatial coverage, focusing on the local rather than the global, the visible rather than the remote. Hence, science is often associated with higher scales of environmental knowledge and public views with much lower scales which are not seen as transferable to other places either. The hierarchy of knowing is therefore based not on what is known but how it is known – for example, formal qualifications, professional positions, and other credentials more frequently legitimate environmental knowledgeability than do recreational experiences.

Moreover, the validity of “lay knowledge” is also criticized when it is invoked in controversies over environmental issues. For example, local protests about the damage that new developments, such as an onshore wind farm, might wreak in a landscape are often denigrated as merely self-interested NIMBY (not in my back yard) concerns, undermining the legitimacy of local people’s connections to and understanding of their local environment. A more critical geographical perspective would be to understand such protests as reflecting lay knowledge arising from and supporting attachment to local places (e.g., Devine-Wright 2013).

### Indigenous knowledge

One particular form of lay knowledge is “indigenous knowledge,” which refers to knowledge held by aboriginal (and usually marginalized) groups in the Global South about their local environments and how best to use environmental resources for agriculture, medicines, education, and other benefits. Indigenous knowledge has been gained through living in those environments for generations, not through traditional scientific research, and is passed down the generations through oral traditions and rituals, rather than through written histories or formal education.

For example, indigenous people living in tropical countries may know about and make use of the medicinal benefits of plants unknown to modern science as part of what is called “ethnobotany.” Pharmaceutical companies try to tap into this indigenous knowledge in order to develop new medicines, making it potentially
very valuable and unfortunately therefore ripe for exploitation, particularly by companies from the Global North (e.g., Agrawal 2002).

In environmental issues, indigenous knowledge is often argued to be intrinsically in tune with natural processes, using fewer resources less intensively in order to support a particular way of living or a cultural group, and thus more sustainable than more “modern” methods of growing crops or managing the environment. Again, this emphasizes that environmental knowledge does not have to come out of the laboratory to be valid or useful, although critics have argued that such claims can romanticize some forms of environmental management which were necessarily exploitative (e.g., fire being used in hunting) rather than conservationist.

Citizen science

As well as lay and indigenous knowledge, both of which are argued to develop over time and be handed down through the generations, attempts have recently been made to increase public understanding of environmental issues through directly involving lay people in producing environmental knowledge in collaboration with professional scientists. Bridging recent work by science and technology studies (STS) and environmental geography, this coproduction of knowledge is another more recent approach to public understanding of science and the environment that rejects the deficit model. This approach argues that knowledge, understanding, and decision-making are no longer the sole domain of so-called experts, but produced in tandem with a wide range of stakeholders. As a result knowledge-making is a collaborative and joint activity between partners. An example is the role of coproduction of knowledge in Canada’s Arctic, bringing together knowledges from science and the indigenous First Nations to enable adaptation to environmental change.

Sometimes called “citizen science,” this takes many forms, including making science more available to lay people through experimental forms of science–society dialogue and interaction, such as “science shops” (Irwin 1995). Recently, the term has become more specific in referring to processes which involve lay people in the production of scientific knowledge. In this guise, “citizen science” has been heralded by academics and policymakers as a good way to engage citizens in environmental debates, as well as improving scientific literacy through education, training, and outreach, although so far empirical evidence for this is limited.

Involving ordinary citizens as “amateur” scientists is increasingly useful for scientists as a cheap way of generating a lot of data. Enrolling enthusiastic amateurs as (paid or unpaid) volunteers can extend the amount and scope of environmental fieldwork that conservation and environmental science can do, such as amateur naturalists collecting biodiversity data for the state (Ellis and Waterton 2004, 2005).

There are numerous recent examples. The 2013 OPAL (Open Air Laboratories) Tree Health Survey involved two government science agencies recruiting and training a cadre of citizen experts to notify the authorities of early sightings of tree pests and diseases that threatened both landscape and economy in the United Kingdom (Figure 2). Members of the Climatological Observers Link (COL) record weather and climate in great detail and for many years as a hobby, but have been approached by scientists to contribute their data to wider research programs, often for free.

Citizen science has also been adopted by NGOs with large memberships but small budgets. The United Kingdom’s Royal Society for the Protection of Birds runs Garden Birdwatch, with up to 280,000 people participating by reporting bird
counts and, in the USA, the Cornell Lab for Ornithology and the National Audubon Society run the Great Backyard Bird Count, with over 80,000 counts submitted, as part of their “citizen science” program (Bonney et al. 2009).

As well as engaging in the scientific process, it is argued that citizen scientists can gain scientific knowledge, explore the environment, reflect on science, and develop positive attitudes toward science. At best, citizen science can be led by citizens from the bottom up, prompted by community concerns such as monitoring water pollution in local rivers, rather than by more esoteric academic questions. This can challenge what constitutes expertise and also open up research and policy processes to on-the-ground challenges faced by citizens. In this sense, citizen science offers hope for expanding public understanding of environmental issues and reflexively shaping the agenda of environmental science itself in future, not through reducing scientific authority but through increasing and redistributing it (Lave 2015) among more diverse knowledge producers.

But the idea of bringing science out of the laboratory into the public sphere has also been criticized. First, citizen scientists may be well aware of how the data they submit is used to make environmental management decisions and may intentionally modify it accordingly. For example, anglers and hunters have been shown to adjust the catch numbers they report to state agencies, numbers which often form the basis of national wildlife statistics, in order to strategically influence subsequent management policies.

Second, the citizens involved may be exploited, not only being unpaid but receiving little in return for their data, not even information or support. Third, citizen science initiatives may simply reinforce the traditional hierarchy of professional scientists telling lay people what to do, with little reward for the latter (Ellis and Waterton 2004, 2005; Lave 2015).

Methods for analyzing public understanding of environmental issues

Another area of critique surrounds the methods used to measure the public understanding of environmental issues. Surveys are often used to identify what the public know about scientific or environmental issues and to compare levels of knowledge across time and space, especially between countries. Surveys and opinion polls are often conducted by government bodies, providing a snapshot of public attitudes at one point in time, a snapshot that can prove powerful when it is used to shape future policy. For example, the United Kingdom’s Department of Energy and Climate Change surveys the general public regularly about whether they think natural processes or human activity are more important in causing climate change.

Academic researchers have also employed quantitative surveys to measure public understanding of environmental issues. For example, Whitmarsh, Seyfang, and O’Neill (2011) surveyed over 500 people in England by post about...
ENVIROMENTAL ISSUES AND PUBLIC UNDERSTANDING

their understanding of the main causes of climate change and what “carbon” means, analyzing the results in terms of what are scientifically “correct” or “incorrect” answers. They interpreted linking ozone depletion with carbon dioxide emissions as a “misperception” of the scientific consensus, rather than an attempt by the public to make sense of highly technical terminology.

However, surveys are not the only or even the most meaningful way of evaluating public understanding of environmental issues. Polling methods do not account for the multiple and complex ways in which people engage with environmental issues because surveys represent a one-off, one-time capture of data with little opportunity for a more nuanced understanding of the answer and little understanding of the attachment, emotion, and feeling experienced as a result of participation, nor of how environmental issues are embedded within wider cultures of value.

Other ways of researching public understanding of environmental issues have therefore been developed. The work done by geographer Jacquie Burgess and colleagues has particularly promoted qualitative methods in this context since the late 1980s. In a time when qualitative methods were largely distrusted in comparison with quantitative ones, methods such as focus groups and workshops (Burgess et al. 1998; Burgess, Limb, and Harrison 1988; Burgess 1996) and deliberative mapping (Burgess et al. 2007) produced a richer, more culturally sensitive analysis of how people understood and responded to environmental issues. Such methods have now become more widely used in human geography to analyze the complexities in public understanding of environmental issues (e.g., Macnaghten and Urry 1998).

SEE ALSO: Environment and the media; Environmental knowledges and expertise; Environmental science and society; Ethnobotany; Indigenous knowledge

Dedication

In memory of Sally Eden. Sally was a generous and supportive colleague. Without her kindness this entry would not have been possible. Sally offered “another way round it” and finished the entry with me. She will be sorely missed (Hilary Geoghegan).

References


Further reading


Environmental issues in rural areas

Geoff A. Wilson
Plymouth University, UK

Human–environment interactions in rural areas over time

While urban areas have received much academic and media attention with regard to environmental issues and threats, it is often thought that rural areas have been less affected by recent environmental challenges (Woods 2005). Yet, most rural areas in a globalizing world have faced unprecedented environmental risks in recent years that are threatening the survival and resilience of many rural systems (Wilson 2007; 2012).

In the past, environmental issues in rural areas were closely associated with environmental impacts of agriculture — a human activity predicated on transformation of the environment through the replacement of one ecosystem with another, with a focus on maximum production of food for human consumption (Pretty 1995). In the preindustrial era, therefore, environmental pressures in rural areas were almost entirely associated with transformation of the environment from relatively undisturbed ecosystems to agricultural land through the burning or clearing of vegetation or the draining of wetlands (Simmons 1996). In the preindustrial era pollution from agriculture was almost nonexistent as all agriculture was organic and based entirely on human/animal labor and organic fertilizers such as manure or plant biomass.

With the onset of industrialization from the nineteenth century onward, and the associated emergence of large agribusinesses and “industrial farming” characterized by intensification and specialization, the environmental impacts of agriculture began to shift toward issues of pollution (pesticides, herbicides), the eutrophication of watercourses (overuse of nitrate-based fertilizers), health issues (e.g., antibiotics and hormones in meat; impact of pesticides on human health: Carson 1962), and the reduction in biodiversity (especially in large-scale monocultures) (Wilson 2007). In the twentieth century, this led to a realization that environmental problems were not only an urban issue but that they also severely affected rural areas. Especially during the 1970s and 1980s, critical human geographers began to highlight the major environmental challenges in rural areas caused by what has been termed the “productivist” era of mass food and fiber production based on large multinational agribusinesses that operate in even the remotest corners of the globe (Wilson 2001; 2007). The productivist era has been closely associated with environmentally harmful practices, especially increased soil and water pollution, further pressure to remove “unprofitable” and “unproductive” vegetation remnants (e.g., woodlands, hedges, ponds), and the substitution of farming focused on local and regional markets with globalized agrocommodity chains characterized by complex and multilayered stakeholder networks and capitalist interests (Robinson 2004). A particularly lively debate has emerged around the replacement of traditional crops with high-yielding and genetically modified (GM) varieties in productivist farming.
ENVIRONMENTAL ISSUES IN RURAL AREAS

systems and its environmental repercussions. While some commentators have argued that GM crops have resulted in the reduced need for fertilizer, pesticide, and herbicide applications (as plants can be tailored more specifically to local climatic and edaphic conditions), others have warned about “genetic pollution” with hitherto largely unknown environmental consequences and, in particular, the path dependency induced by farmers adopting GM crops who are increasingly locked into multinational agribusiness and knowledge trajectories over which landholders have less and less direct control (Wilson 2007).

Environmental implications of productivist, nonproductivist, and multifunctional pathways

Since the 1970s there have been substantial shifts in the role of agriculture as an economic sector, with new environmental challenges emerging for rural areas. Most of the developed world has witnessed two distinctive pathways: on the one hand, new technologies and increasing mechanization have led to an intensification of agricultural production in the most productive and globalized agricultural areas (e.g., the American Midwest, the Paris Basin, most of the Netherlands, Emilia-Romagna in Italy, the Australian wheat belt), while, on the other hand, geographically marginal and upland areas (especially in the developed world) have often seen a relative withdrawal of agriculture and extensification of production with associated “deagrarianization” (Robinson 2004). Critical commentators have referred to this process as multifunctional territoriality (Wilson 2001), where some agricultural territories have become more specialized with regard to intensive production (productivist globalized areas), while others have embarked on nonproductivist pathways characterized by environmental conservation, extensification, and production of specialist and (usually) environmentally friendlier high-quality foods (Marsden 2003; Wilson 2007).

However, these processes have not been spatially uniform and human geographers from Australia and New Zealand, in particular, have highlighted that multifunctional agricultural and rural development pathways have taken different form in the Antipodes characterized by neoliberal ideologies and ever more globalized agrocommodity chains (Dibden and Cocklin 2009).

Nonproductivist pathways characterized by the extensification of agriculture have often been the result of external pressures in marginal rural areas rather than endogenous forces. This has been particularly the case in areas where outmigration and a lack of willingness of young people to take over their parents’ farms (e.g., in most of Europe), or where mass rural–urban migration (e.g., China where 300–400 million rural people have moved to urban areas for better job opportunities) have eroded social and economic capital in rural communities, leading to “enforced” nonproductivist pathways based on pathways of deagrarianization (Rigg 2006; Wilson 2012). As a result, in many nonproductivist rural areas the most severe environmental threats no longer emanate from a dwindling agriculture but from other pressures associated with rapidly globalizing societies. Human geographers have, therefore, increasingly focused on processes such as counterurbanization and urban sprawl as new environmental threats in many rural areas, in particular through the increasing loss of fertile agricultural land near rapidly growing megacities (e.g., thousands of square kilometers of fertile agricultural land lost in China every year to rapidly expanding urban conurbations), but also with regard to the loss of productive agricultural land to “counterurbanites” who purchase agricultural land for...
recreational purposes rather than for food production (Robinson 2004). Similarly, increasing attention is paid to the impacts of climate change in rural areas, especially in arid and semiarid rural areas which rely on constant water supplies for the continued production of food and fiber (e.g., Australia, parts of the American West Coast, the intensively farmed areas of the Ukrainian and Russian steppe). Research is increasingly suggesting that climate change, in particular, is often exacerbating existing socioeconomic problems in rural areas (e.g., compounding the loss of land through urbanization/counterurbanization, increased tensions over the demand for water from both agricultural and counterurbanization processes, outmigration of young people), often leading to a loss of resilience in already vulnerable rural communities (Wilson 2012).

In an academic research context, the changing role of agriculture as an increasingly less important component of rural society has led to a change in focus away from “pure” agricultural research that predominated until the 1930s (e.g., the now largely defunct subfield of agricultural geography) to an increasing focus on rural issues (rural geography), where agriculture is conceptualized as only one of many processes influencing rural development and environmental pathways – possibly best illustrated by the emergence of specialist academic journals such as *Rural Sociology* (established 1936), *Sociologia Ruralis* (1960), and the *Journal of Rural Studies* (1984), where issues of environmental threats to rural areas feature more and more prominently (Woods 2005; Cloke, Marsden, and Mooney 2006). Rural geography research is, therefore, also increasingly acknowledging political issues in rural areas linked to tensions between multiple and multilayered stakeholder groups with vested interests in rural areas (e.g., farmers, developers, planners, multinational corporations) and associated environmental repercussions (Winter 2004).

Particular attention has been given to increasingly complex tensions between environmental conservation groups, landholders, and developers in the context of protected areas management (e.g., national parks, biodiversity reserves) and environmental management issues (e.g., ground and surface water protection, soil management, rural ecosystem services).

**The changing rural environmental policy context: internal and external drivers**

These environmental issues in rural areas are influenced by both internal and external processes. Much research has been conducted on internal drivers of change linked to understanding the environmental attitudes and perceptions of landholders. There is general consensus that, in general, landholders are increasingly appreciating the importance of environmental conservation on their farms (Wilson 2007). This is highlighted in the continuing importance of organic farming and what has been termed “alternative agri-food networks.” Although driven partly by market forces (e.g., higher prices for organic products) and a relative leveling off of organically farmed land since 2000 at about 10% of globally farmed land, organic farming and alternative agri-food networks suggest an increasing appreciation by both farmers and consumers of the negative environmental and health-related side effects of modern productivist agriculture (Goodman 2004). Yet, as Burton and Wilson (2006) have highlighted, farmer environmental attitudes remain complex and multilayered and can change rapidly from environmental conservation to capitalist pragmatism based on market conditions and personal circumstances. There is no doubt, however, that the professionalization of farming in the developed world (and increasingly...
in the developing world), associated with better access to environmental education, has led to a recognition by landholders of the threats linked to climate change, productivist agricultural pathways, and resultant land degradation.

*External* processes influencing environmental management in rural areas are closely associated with state policy and planning. To be sure, regulation has been part and parcel of human attempts to regulate human–environment interactions in rural areas since the dawn of agriculture, epitomized by, for example, strict regulation of water use in early “hydraulic” societies to safeguard precious water resources and provide equitable distribution of water to affected stakeholders, or, more recently, the establishment of national parks and protected areas to protect vulnerable fauna and flora threatened by agricultural and urban expansion or the implementation of green belt policies in rural areas near large conurbations to prevent urban sprawl and impacts on remnant biodiversity (Simmons 1996; Wilson 2007).

Possibly the most important suite of external regulations attempting to protect the rural environment has been related to *agri-environmental policies*. Today, such policies are devised nationally by most nation-states, but the most comprehensive agri-environmental policy package comes from the European Union’s CAP (the so-called Pillar Two policies) in the wake of substantial CAP reform since the mid-1980s (Buller, Wilson, and Höll 2000). Agri-environmental schemes (e.g., Environmentally Sensitive Areas Scheme in the United Kingdom, the MEKA Programme in Germany, the ÖPUL Programme in Austria) reward farmers financially for protection of the environment. While they take many different guises based on national environmental priorities, they usually encourage farmers to protect remnant wildlife habitats on farms, to extensify production (often with a payment per hectare to farmers for income foregone), to reduce livestock densities, to plant or maintain private forests, or to reduce or abandon the use of environmentally harmful fertilizers, pesticides, and herbicides. However, academic research (with the involvement of many human geographers) has shown that the success of agri-environmental policy has been questionable in environmental terms, especially as many agri-environmental schemes are rewarding farmers for carrying on with business as usual—in other words, while many studies have shown that agri-environmental schemes have benefited the environment by preventing further productivist intensification and possibly helping change farmers’ attitudes toward conservation, the schemes have rarely led to a substantial improvement of the environmental condition of rural areas (Buller, Wilson, and Höll 2000).

Since the early 1990s, a wide-ranging ideological debate has emerged between different rural environmental policy priorities in Europe and other developed countries. Critical commentators have argued that environmental trajectories in rural areas may be different in Europe compared to New World countries because of the different historical pathways of human landscape evolution, thereby generating different policy needs. In Europe, for example, the focus has been more on the preservation of cultural landscapes with a complex mixture of productivist and nonproductivist (or, often, less productivist) territories through policies emphasizing the multifunctionality of the rural landscape, suggesting that productivist production and environmental preservation are not incompatible (Wilson 2007). In the New World, meanwhile, policy has tended to focus more on the setting aside of areas in which agriculture and human settlement is largely prohibited (e.g., national parks in the United States and Australia), while the environmental regulation of agriculture in nonprotected areas has been
less stringent. Inevitably, these processes closely intertwine with philosophical-political debates related to dominant ideologies, with many arguing that neoliberal agendas in countries with large agricultural export markets such as Australia, New Zealand, and the United States are the key explanation for less stringent rural environmental policies compared with less neoliberalized (and less agricultural export-oriented) European countries (Dibden and Cocklin 2009). These often incompatible policy ideologies have come to a head in attempts to harmonize trade in food and fiber through recent rounds of the World Trade Organization, and increasing demands for developing countries to provide more environmental protection for their rural areas. The outcome has been, for the near future at least, a relative stalemate between advocates of more protectionism (EU countries in particular) and those advocating free-market ideologies (e.g., United States, Australia, New Zealand) (Potter and Tilzey 2005; Rigg 2006).

Future challenges for managing environmental issues in rural areas

The discussion has highlighted the complexity of environmental issues affecting rural areas around the globe. These challenges are only likely to increase in the future because of several interlinked factors. First, global human population growth is forecast to peak at approximately 9.5 billion around 2050, with the addition of over 70 million new mouths to feed per year for the foreseeable future. While Malthusian and neo-Malthusian arguments about maximum global human carrying capacity in relation to food and “people overpopulation” have been largely discredited on the back of continuing rising agricultural productivity linked to technological innovations (Green Revolution, genetically modified crops, high-yielding crops), changes in diet toward increased meat and dairy product consumption in high population countries such as China and India (“consumption overpopulation”) are putting further pressures on global food systems (Wilson 2007). Second, these processes will be exacerbated by further urbanization and increasing urban sprawl, with over 50% of the global population now officially living in urban areas, leading to the further loss of productive land, more pressure for rural–urban migration with associated deagrarianization, and the possible loss of agricultural productivity. Third, and possibly most importantly, IPCC predictions increasingly highlight further severe impacts of climate change on agricultural productivity in most rural areas of the globe (2–4°C temperature rise by 2100 now most likely). Although these impacts will be geographically highly diverse, with northern latitudes likely to benefit from a milder climate, climate change will particularly threaten the livelihoods of people living in arid and semiarid rural areas that are already facing severe water shortages (Mestre-Sanchis and Feijoo-Bello 2009).

These processes combine to put additional pressures on often already strained agricultural systems. Likely scenarios involve pressures to further intensify food production on ever-shrinking agricultural land (squeezed by urban sprawl, climate change, and, in places, irreversible land degradation), with an emphasis on intensifying productivist pathways that will leave less room for environmental conservation. These processes have been termed either “neoproductivist” in relation to rural areas that have recently witnessed extensification and deagrarianization but are being recruited back into the productivist fold, or as “super-productivist” in the case of already highly productivist areas that will be forced to further intensify food and
fiber production to meet rising global demand (Wilson 2007; Wilson and Burton 2015). The key question over the next few decades will therefore be whether this still leaves room for multifunctional or non-productivist pathways, with an emphasis on environmental protection of rural ecosystems, or whether environmentally unsustainable productivist pathways will prevail.

SEE ALSO: Environmental degradation; Environmental governance; Environmental management; Environmental policy; Environmental regulation; Rural citizenship; Rural geography; Rural policy and politics; Rural society in the Global North

References


Environmental knowledges and expertise

Carol Morris

University of Nottingham, UK

The production of knowledge about natural environments is one of the major concerns of the discipline of modern geography. This disciplinary focus recognizes that it is necessary to know in order to make assessments about the kinds of environmental problems that confront society and how these might be resolved. However, since the 1990s environmental knowledge has itself become the object of geographical research interest for a range of geographers identifying mostly, but not exclusively, with one or more of the human geographical sub-disciplines. For these geographers, questions of environmental knowledge have come to be seen as central to examining the governance and management of environmental issues. However, the drivers of what can be called this “cognitive turn” in the Global South are distinct from those in the Global North. In the former, the study of environmental knowledge has been stimulated by the emergence of participatory approaches to development in which “indigenous environmental knowledge” is taken more seriously than in traditional, top-down models of development. This way of thinking has also influenced research into environmental knowledge in the Global North which is the focus of this entry. However, more significant in the Global North have been other interrelated societal debates leading to the notions of citizen science and of public involvement in environmental decision-making as characterizing the approach to the study of environmental knowledge.

The first of these debates concerns the relationship between science and society, in which declining public trust in expert institutions and the knowledge they produce about environmental and other issues has prompted interrogation of scientific knowledge claims, how these circulate and are contested by “lay publics” that claim to know environments differently. A widely cited example of this is the clash between sheep farmers and scientists working for the United Kingdom’s Ministry of Agriculture, Fisheries and Food in the management of the Cumbrian hills following the radioactive fallout from the Chernobyl nuclear plant accident (Wynne 1996). The farmers drew on their unaccredited, experience-based expertise in mountain ecology and sheep behavior to challenge the inappropriateness of the management prescriptions developed by the scientists. A second debate is about broader shifts in the political landscape toward more participatory forms of democracy which have opened up opportunities for a range of noncertified knowledges in environmental decision-making. For example, the knowledge about fish stocks that anglers develop through their fishing practices on particular rivers is utilized by the public organizations that are responsible for managing rivers in the United Kingdom. Third, and increasingly pronounced since the financial crisis of 2008, is the reduction in public sector funding for science which has undermined the institutional capacity of professional environmental knowledge production. In these conditions scientists have begun to recognize the potential of volunteers and
amateurs/enthusiasts as legitimate providers of environmental knowledge within both the context of crowdsourcing initiatives and more formalized participatory monitoring networks. Geographical exploration of these developments has been conducted alongside and in dialogue with a range of cognate social sciences, among which science and technology studies (STS) have been particularly influential.

Defining, conceptualizing, and categorizing environmental knowledges

Knowledge is a spatially and temporally dynamic means of representing the world that is developed through observing and interacting with people and the nonhuman world (Castree 2005). Important within recent conceptualizations of knowledge is the idea that knowledge is a relational phenomenon, produced in association with others, both human and nonhuman. It is also suggested that knowledge exists as more or less established bodies of knowledge that are shared by distinct groups of people, leading to the use of the term “knowledges” to capture the notion that there are multiple forms of knowledge. Beyond this broad definition, three distinctions are often made in debates about environmental knowledge that are used to help understand the character of that knowledge.

The first distinction is that all knowledge has a point or points of origin (i.e., institutions, groups, or individuals who promulgate a particular body of knowledge or specific knowledge claims); a referent or referents (i.e., those particular things referred to in any knowledge claim or body of knowledge and can either be material things or other bodies of knowledge); an addressee or addressees (i.e., the intended audiences of a body of knowledge or the consumers of knowledge). This trio of origin(s), referent(s), and addressee(s) enables one body of knowledge to be distinguished from another (Castree 2005). For example, a university-based marine biologist who undertakes research into the management of fish populations will know about fish in ways that are likely to be distinct from individuals and communities who rely on fishing for their livelihoods but who may not have received any formal training or education.

The notions of tacit and codified knowledge bring us to a second distinction that is often drawn in debates around environmental knowledge. Tacit knowledge is informal and “ordinary” knowledge that is made use of in everyday life and so deeply internalized that it is taken for granted or regarded as common sense. It is a form of knowledge that emphasizes knowledge as practice as well as cognition, and explains the use of the term “knowledge practice” and an emphasis on knowing rather than knowledge. This concept has similarities with Ingold’s (2000) notion of “environmental perception” in which the environment becomes meaningful only within the context of everyday acts and direct experience of other people and the nonhuman world. Understanding knowledge in this way suggests that knowledge of the environment is in large part likely to be situated and context-specific – hence “local” and “indigenous” knowledge – emerging through routinized, site-specific practices and produced in close association with the wider objectives and identity of a particular social group. Meanwhile, codified knowledge is formal, explicitly articulated knowledge which can be standardized and systematized, written, stored, and easily transferred from one location to another. Codified knowledge is typically technical and specialized in nature such as the environmental knowledge
produced and promulgated by science and formal systems of education. It is recognized that all forms of knowledge can have a tacit dimension, including those that are codified.

The third distinction recognizes that a body of environmental knowledge comprises four different elements. First, know what refers to knowledge about facts and is an element of knowledge that is largely codified, involving the systematic and formalized collection of information. Second, know why is the knowledge of principles and rules, and is a significant feature of science and technology. Third, know how refers to the practical ability to enact something and implies a set of skills, while, fourth, know who is the knowledge of people who can provide assistance in undertaking tasks. There is some debate around whether there are distinct differences between these components of knowledge or whether they are complementary and relational (Ingram 2008). For example, know what may provide the basis for know why, and know why, although mostly codified, can be seen to rely considerably on tacit knowledge for interpretation particularly at the level of individual understanding. It has been suggested that know how is particularly significant within tacit knowledge, while know what and know why are key elements of codified knowledge (Ingram 2008).

Research on forms of knowledge continues to be important because of the ongoing problem of nonexpert knowledge being ignored and devalued in processes of development. However, some geographers have questioned the essentialism (i.e., fixed or given) that is implied in the prefixes “expert,” “lay,” and “local.” This has been prompted in part by the efforts of scholars whose geographical focus of interest is outside the Global North and who have undertaken considerable conceptual work in problematizing the divide between “scientific” and “indigenous” knowledges in natural resource management (e.g., Agrawal 2002). Focusing on types or forms of knowledge runs the risk of reifying and fixing that knowledge. Instead, there is a need to conceptualize knowledge as dynamic and fluid, being continuously reproduced through ongoing social interactions. Moreover, the content and epistemology of knowledge is believed to be less significant than the links between power and knowledge, or the politics of knowledge. As such, what is required is examination of the social, historical, and institutional relations in which knowledge develops and is recognized as valid or legitimate (e.g., the knowledge of government scientists) while other claims to knowledge (e.g., those made by sheep farmers) are delegitimized.

The concept of a knowledge culture has been developed as one response to these concerns and applied in the context of the environmental management of agriculture in the United Kingdom. Central to the notion of knowledge culture is that an understanding of an environment or environmental process is collectively achieved or produced through routinely executed knowledge practices and discourse, for example within the context of farming, veterinary science, or policymaking. This shared understanding – a knowing from within as it is characterized by Tsouvalis, Seymour, and Watkins (2000) – enables people within that knowledge culture to relate and respond to one another. In order for knowledge cultures to become established and be sustained, they develop sets of rules about what counts as legitimate knowledge practices, for example, within policymaking this is likely to be scientific knowledge of one form or another, while within farming experiential and practical knowledge are likely to be emphasized. These rules can be challenged both from within and without the knowledge culture (e.g., through contact with other knowledge cultures associated with
new technologies or policies, or through wider societal demands), and the ensuing knowledge struggle can lead to the evolution or dissolution of a knowledge culture. The concept of a knowledge culture does not mean that it is no longer possible to talk about different knowledges; rather, the emphasis is on exploring how these knowledges come into being, come to be recognized as particular forms of knowledge, and how they are reproduced through complex social processes and power relations within the context of knowledge cultures.

Although knowledge cultures are understood as dynamic, based on ongoing social interaction and exchange and, as already implied, subject to contestation, they can become very durable, with their boundaries becoming relatively fixed. This stabilization takes place either through a process of institutionalization, that is, where the “rules” of knowing become formalized and codified and/or when a knowledge culture starts to constitute individual or group identity, that is, I/we know – in particular ways – therefore I am/we are. The identity of a particular community, and the individuals that make up this community, are part and parcel of its knowledge culture. For example, within the context of a farming community what it means to be a “good” farmer has come to be associated with knowledge of intensive agricultural techniques and being seen to farm in a “tidy” fashion. This identity can be brought into relief at times when a knowledge culture is challenged. For example, EU policies since the mid-1980s have incentivized farmers to use more environmentally friendly farming practices but these can challenge the pre-existing “agrarian” knowledge culture because they demand engagement with new knowledge practices that appear antithetical including encouraging farmers to manage their land in ways that may appear to be “untidy” but are beneficial for biodiversity. One advantage of a knowledge culture approach is that it does not assume at the outset the particular character of knowledge(s) that distinguish the knowledge culture. However, this perspective has been less sensitive than some others to the role of nonhumans, both living and artificial, in the relational production of environmental knowledge.

Noncertified environmental knowledges

In spite of some of these concerns, the most prominent theme within geographical research on environmental knowledges has been noncertified knowledges, that is, those produced outside formal, official, or professional processes and institutions of knowledge production. These are variously labeled lay ecologies, local, indigenous, traditional, practical, and experiential environmental knowledges. As some of these monikers suggest, research has examined the environmental knowledges produced by indigenous communities, for example First Nations peoples in Canada or Aboriginal groups in Australia, as well as those produced by the other residents of particular places and by specific occupational and recreational communities such as farmers, anglers, and amateur naturalists and meteorologists (e.g., Endfield and Morris 2012). It is typically assumed that livelihood dependencies, recreational activities embedded in specific localities, or long-term residence result in a very intimate relation between humans and (aspects of) the nonhuman world. Such is the closeness and sustained nature of this relationship that people develop a very detailed knowledge of local environmental conditions. Such knowledge accumulates and can be handed down through generations by cultural transmission but is also understood to be dynamic, as evolving over time through adaptive processes. Understanding how local and indigenous environmental knowledges
are produced and evolve has been one concern of this research.

A number of other key features of this scholarship can be identified. First, researchers have been at pains to draw attention to the ways in which local environmental knowledges are often ignored or marginalized in modernization and development processes, both in the Global North and Global South, but also within the development and implementation of environmental policy. Second, the differences between noncertified and expert knowledges have been explored, and how these can lead to tension and struggles over both the identification and definition of environmental problems and, by extension, the formulation of solutions to those problems. This concern for the politics of environmental knowledge takes a slightly different form in work that examines the treatment by professionals of amateur knowledge producers whose environmental data collection has become even more vital in an era of significant constraints on public finances (Lawrence and Turnhout 2010). Here, exploiting volunteered environmental data and the volunteers who provide them risks discouraging these actors from continuing to contribute to environmental knowledge production. Third, in spite of an identified “epistemological anxiety” concerning noncertified environmental knowledge among both experts and nonexperts, and the persistence of an expert deficit model of lay knowledge that is based on the idea that the public misunderstand environmental issues, research continues to assert that this way of knowing can and should make a valuable contribution to the governance and management of natural environments, including the resolving of specific environmental problems. Noncertified environmental knowledge, it is argued, needs to be taken more seriously by environmental “professionals” including policymakers and scientists.

Another theme within environmental knowledges research pertains to the mechanisms by which the integration of noncertified and expert environmental knowledges could take place, or, in the language of recent studies, the means of their “coproduction.” Specifically, attention has begun to be given to a range of participatory or collaborative approaches to environmental governance, research, and management which, in theory, provide more of an opportunity for local/noncertified environmental knowledges to be integrated into decision-making. For example, Landstrom et al. (2011) explore the potential of “competency groups” in which members of the public worked with scientists to coproduce new environmental knowledge about flood risk and its management. Competency groups are distinguished from some other models of participation in environmental decision-making as they are not constituted by representatives of institutional stakeholders or groups of local residents who may have particular, established agendas or political aims. Instead, competency groups require “participants to be willing to engage as individuals, with their own particular knowledge, skills and expertise” (Landstrom et al. 2011, 1619). As a result, this approach diversifies the range of publics with whom scientists have previously been prepared to collaborate. Expertise, so this study argues, can be effectively redistributed between science and affected publics through this new form of “research collective.” However, it involves a repositioning of scientists with respect to their flood modeling practices which may be threatening to their professional identities.

One of the challenges identified in the coproduction of environmental knowledge is the importance of an ongoing relationship between lay publics and experts to ensure that the tacit knowledges of the former are given the attention they deserve. It also concurs with research that
calls for a focus on “learning” among environmental actors as a means of addressing the challenges of integrating local environmental knowledge into environmental governance. Framing the task in this way aims to avoid the problem of differentiating between forms of knowledge and demarcating which forms are perceived as valid and which are not in environmental decision-making.

**Expert environmental knowledges**

A distinct form of research on environmental knowledge to that discussed in the previous section focuses on formally accredited environmental expertise and the activities of associated experts, such as environmental advisers, identified as expert through their scientific training and/or occupation. Some of this work involves geographical historiographies of certified expert environmental knowledge domains such as global environmental change (e.g., Hulme 2010). Other research relates to that discussed in the previous section when it examines how, where, and why certified expert knowledge wins out over noncertified environmental knowledges in the context of resource management disputes. One specific process of interest here is “boundary work,” a concept coined by Thomas Gieryn (1983) to make sense of the efforts to distinguish one domain of knowledge or expertise from another, notably scientific knowledge from non-science. For Gieryn, rather than trying to understand the characteristics of science that demarcate it from other intellectual activities, the focus should instead be on the ideological efforts by scientists to distinguish their work from non-scientific intellectual activities, that is, how scientists attribute selected characteristics to science that enable them to draw a social boundary around their work that sets it apart from non-science to make it legitimate and authoritative, particularly in the public domain. Gieryn suggests that it is helpful to explore and conceptualize boundary work through a focus on the rhetorical style of scientists’ ideological statements. Gieryn argues that boundary work is seen as a particularly useful stylistic resource for any profession or occupation when the aim is expansion of authority/expertise into domains claimed by other professions or occupations; monopolization of professional resources and authority; and/or protection of autonomy over professional activities.

While Gieryn’s concept was developed within the sociology of scientific knowledge, it has been deployed by geographers interested in environmental knowledge. For example, Eden, Donaldson, and Walker (2006) explore processes of boundary-making by environmental nongovernmental organizations (NGOs), conceptualized as both scientific and political actors and involved, therefore, in both the consumption and the production of scientific knowledge within the context of wider processes of democratizing expertise. The winning of boundary disputes by NGOs enables them to gain influence and access to resources. However, sometimes the boundaries around scientific environmental expertise are drawn very tightly and at other times much more loosely, depending on the particular agenda that is being pursued by an NGO. This leads Eden, Donaldson, and Walker (2006) to conclude that boundary work is highly contextual, tactical, and adaptable.

A further dimension of the utility of boundary work is the demarcation of disciplines or theoretical orientations within science. This is illustrated in Bell and Sheail’s (2005) historical geographical study of ecology as a discipline which explores when, where, and how an ecological expertise came to be defined in the early twentieth century. They identify that it was both accredited and nonaccredited groups, in specific geographical locations, that were important in
this process. The study reveals the fluidity and imprecision of the boundary between specialist and lay knowledge of ecology and in so doing reinforces the conclusions of Eden, Donaldson, and Walker (2006).

Another theme is the doing or practicing of formal environmental expertise in the field (as opposed to other spaces associated with this expertise such as the laboratory, zoological/botanical collections, or scientific societies) by particular types of knowledge worker. Here, the focus of concern shifts away from controversy and dispute in environmental governance and management to examine the “how” of environmental knowledge coproduction. The particular fields of interest have included forests being managed under certification schemes that promote sustainability (e.g., Eden 2008) and farms participating in agri-environmental schemes (e.g., Proctor et al. 2012). The knowledge actors investigated are field-level advisers who mediate between scientists and actors such as farmers and foresters, and who work with specialized environmental knowledge to facilitate the implementation of certification standards and policy prescriptions. Concepts from STS are here mobilized to reveal the heterogeneity of field-based environmental expertise, that is, its production in association with many other human actors, including, in particular, other environmental experts but also the diversity of plants and animals comprising farms. Also revealed are the uncertainties of environmental science but these are rendered beneficial within the field context because they provide knowledge workers with an important degree of flexibility as they negotiate and enable environmental management tasks to be realized in practice. Another important finding of this work, and one which deserves further investigation, is that intermediary knowledge workers are vital actors in environmental governance and management, and yet their role can be significantly constrained. For example, in the public policy context of agri-environmental schemes it has been shown that field advisers are limited in the extent to which they can encourage a wider environmental learning among farmers as opposed to simply ensuring that scheme objectives are met.

The ongoing commitment to participatory forms of decision-making among policymakers suggests that the challenges involved in coproducing environmental knowledge will continue to exercise geographers. Future debates about environmental knowledge will also need to be extended to provide greater understanding of particular categories of environmental knowers who have been relatively neglected in research to date including, for example, children, young people, and women; and of the politics of environmental knowledges associated with institutional contexts beyond farming and forestry that constitute major uses of land and present significant environmental challenges, for example resource extraction industries and sport.

SEE ALSO: Environment, democracy, and public participation; Environmental change and social learning; Environmental citizenship; Environmental governance; Environmental issues and public understanding; Environmental policy; Environmental science and society

References

ENVIRONMENTAL KNOWLEDGES AND EXPERTISE


Environmental management

Robert A. Francis
King’s College London, UK

Definition and overview

Environmental management essentially encompasses the different types of human intervention put in place to ensure that environmental forms, processes and functions (i) are not degraded to an unacceptable level by human action and/or (ii) do not pose an unacceptable hazard or risk to human societies. It is therefore the management of our impact on the environment, and the environment’s impact on us, rather than direct management of the environment itself. Most specifically, management refers to the methods and processes used to predict, measure, and mitigate environmental impacts. Management is usually performed to address environmental legislation or policies, which are laws, regulations, or guidelines put in place to ensure that impacts are limited to an acceptable level. Improvement of degraded conditions may also be considered a form of management, though usually the terms restoration, rehabilitation, or improvement are used in this context.

The concept and practice of “management” not only implies but necessitates control of something; in this case, the environmental system, subsystem, component, or process under consideration. Management of the environment therefore presents a major challenge, particularly given the complexity of environmental systems, our limited understanding of both the systems themselves and the impacts we may cause, the many different demands that we require the environment to meet, and our limited ability to guide environmental processes in the ways we desire. It includes the application of a wide body of knowledge housed within a range of disciplines, including environmental science but also political ecology, environmental law, information theory, environmental engineering, and environmental economics.

Many activities could be considered environmental management at some level and the principles, methods, and processes of management are therefore broad. For example, the term may be used to refer to “sustainable” or environmentally conscious business practices such as low-impact resource utilization and production (reducing water use or carbon emissions in production of industrial materials) as well as to attempts to maximize environmental resource production (such as the use of pest control techniques to improve crop yields). Over recent decades, environmental management has moved from a focus on individual environmental components or processes to broader environmental systems.

A brief history of environmental management

Although some management of environmental impacts (both to and from humans) has been performed for millennia, much of our understanding and application of management techniques has emerged relatively recently. This
has followed increased environmental awareness, originating in the nineteenth century in response to increasing industrialization and the impacts it generated, particularly in terms of air and water pollution and the degradation of traditional landscape aesthetics. General public health and aesthetic concerns led to early forms of environmental regulation and legislation, such as the various clean air and water acts passed in the nineteenth and twentieth centuries (e.g., the Public Health Act 1875, River Pollution Control Act 1876, and Clean Air Act 1956 in the UK).

A shift in focus from human impacts resulting from environmental degradation to the environment itself as an entity deserving of protection came about following recognition of extensive land-use change and declines in wildlife populations in the late nineteenth and early twentieth centuries, and is closely linked to the development of the conservation movement (Adams 2003). The rapid increase in land-use change and the development and use of chemical fertilizers, pesticides, and herbicides in the mid-twentieth century, highlighted by landmark publications such as Rachel Carson’s *Silent Spring* (Carson 1962) led to an increase in environmental concerns and the widespread development and formalization of “environmentalism” as a social movement. These concerns led to the development of key regulating agencies, organizations, or branches of government, such as the US Environmental Protection Agency (1970), or the UK Department for the Environment (1970). Over the last five decades, the number and extent of environmental organizations and agencies set up to help determine policy and guide environmental management efforts has increased dramatically, to the extent that the vast majority of countries now have environmental organizations that guide management practices and enforce regulations, though the authority and efficacy of such organizations naturally varies.

In recent decades, further changes have increased the need for effective environmental management. The overall quality of the global environment has continued to decline despite some local improvements, while at the same time both human populations and our reliance on environmental resources and ecosystem services have increased and will likely continue to do so (Wackernagel et al. 1999). As a result both the potential for and consequences of environmental mismanagement have been amplified. This may be seen in global increases in both frequency and impacts of natural disasters such as flooding and ongoing failures to curb global biodiversity loss (Butchart et al. 2010; Jongman, Ward, and Aerts 2012).

Alongside this, novel forms of environmental hazard have emerged that have required new approaches to management. Increasingly complex or unknown forms of pollution, such as (to draw on two simple water-based examples) endocrine-disrupting hormones or plastic debris in aquatic ecosystems as well as further environmental changes (for example, anthropogenic climate change, greater socioeconomic disparities between the poor and wealthy), have magnified the risks of both environmental degradation and human impacts.

This increasing need for management has led to the further development of national and international environmental legislation and policy to set targets and guide management efforts. Globally relevant examples of such instruments include the Convention on Biological Diversity (CBD), which requires all member states to sustainably manage their biological resources; the UN Framework Convention on Climate Change, which requires member states to limit atmospheric greenhouse gas concentrations and therefore prevent anthropogenic climate change; and the Convention on Long-Range
Environmental management is based around the principle that while some environmental impacts are inevitable as we use environmental resources, these should be limited as far as possible so as to ensure that such resources are not permanently depleted and so that wider effects that may be detrimental to both the environment and people do not occur. It is therefore a weak form of the precautionary principle, which holds that as impacts are so complex and potentially irreversible the safest thing is to prevent impacts at all. In reality, complete prevention of impacts is not possible and so the aims of management efforts must be practical and achievable. Where management is performed to minimize risk to humans from environmental hazards, again there is recognition that some form of exposure to risk is unavoidable, but that the environmental processes can be managed to keep these at an acceptable level.

All forms of environmental management essentially proceed through potentially cyclical stages of assessment, planning, action and monitoring, though the emphasis will vary from case to case (Figure 1). The broad stages are as follows.

1. **Assessment of current environmental condition and potential impacts.** Any attempt at management must be based on an understanding of the environmental system and impact(s) that is to be managed. This is a fundamental aspect of management planning. Where impacts are yet to occur (for example, when building development is being planned), pre-impact assessments are used to determine the value of the environment and the potential risks from the development. These are usually termed environmental impact assessments (EIA). These have emerged in many countries in the past 40 years, and are essentially procedures wherein proposed actions (e.g., building construction) are assessed for their environmental impacts in order to inform decisions prior to commitments being made, and to ensure potential impacts are addressed either pre- or postaction, where possible (Morgan 2012).

EIA utilize a substantial skills base and require those involved to be well trained in the environmental sciences, including experts in ecological and environmental surveying, monitoring, data analysis, and sometimes environmental modeling. Knowledge of national legislation and policy is also important to ensure that regulations are followed. Individual countries have their own specific operational processes and legislation.
related to EIA, but the broad principles of EIA are relatively universal. Nearly all of the 193 member nations of the United Nations have either domestic EIA legislation or have signed international agreements that incorporate EIA in some form (Morgan 2012). Many international funding bodies, such as the World Bank, require EIA to be implemented for major projects that they support financially, regardless of location.

But EIA can have limited efficacy, usually stemming from limited practitioner availability and training or a lack of resources required to fully establish the potential extent of impacts; but also sometimes because of a lack of a wider understanding of environmental systems and an inability to accurately quantify indirect, cumulative, and multiscale impacts. There can also be significant pressure for regulators to prioritize development before environment and make EIA cursory rather than comprehensive, further undermining the process. In situations where impacts have already occurred, the assessment process may determine the extent and severity of these as well as current environmental conditions to inform subsequent management activities. Where the assessment stage involves planning for prevention or mitigation of environmental hazards (e.g., flooding), the assessment may involve the prediction of probability of an event of a given size happening over a given time frame (e.g., a one-in-hundred flood event, which has a 1% probability of occurring in any given year) as well as potential environmental and societal damage. In such cases, expertise in, for example, economics, engineering, and disaster planning can become particularly important to the assessment process.

Overall, the assessment stage requires a substantial skills base and utilization of resources to determine potential impact, along with clear guidance relevant to the particular type of assessment being conducted (e.g., construction planning, limiting flood risk) and national or regional legal requirements. It is the foundation upon which all further efforts are built.

2 Planning of management activities. Based on the findings of the assessment stage, a plan is formulated to limit impacts. The plan will be guided by relevant legislation and policy and should have achievable and quantifiable objectives. A relatively universal example is biodiversity action plans (BAPs). The Convention on Biological Diversity (1993), which 191 nations have ratified, requires member nations to sustainably manage their biological and ecological resources at a range of scales, from genetic variability within species up to entire ecosystems. Biodiversity action plans, or their equivalents, are the primary mechanism that some (though by no means all) countries have chosen to achieve this. The plans are informed by an assessment stage that involves species inventories and measures of geographical distribution and abundance, and review of the conservation status of species. They are essentially plans for sustainable management of species (including conservation and restoration). These are then used to highlight potential impacts as well as to guide management activities that are intended to benefit species considered to be of conservation concern.

As with the assessment stage, the planning stage also involves a range of expertise, often involving modeling and prediction of the likely impacts of different management actions, again supported by knowledge of relevant legislation. Crucially, the planning
stage is where aims, objectives, and relevant timelines for project completion are set; these are the criteria through which the success (or otherwise) of management activities will subsequently be measured.

3 Management actions. The plan is put into action in the next stage. This may be construction of flood defenses, changes to the organization or behaviors of a company with regards to waste disposal, construction of buildings with mitigation strategies such as biodiversity offsetting, and so on. This may occur over several years and involve distinct phases of activity, including some ad hoc revision of the management plans if it rapidly becomes apparent that they are not achievable. As an example, management efforts to achieve “good ecological status” or “good ecological potential” in rivers throughout Europe in response to the Water Framework Directive (2000/60/EC) have often been revised or adjusted due to complications in identifying appropriate reference systems, improvement in methods and techniques, quantifying uncertainty and appropriate environmental indicators, and overly ambitious timescales (Hering et al. 2010). Usually specialists are responsible for enacting the management activities, guided by the prepared plans and other relevant experts.

4 Monitoring the effects of management actions. It is important that the management actions that have occurred are in some way monitored to ensure that their success (or lack of it) is based on evidence. Monitoring essentially involves reassessment of the environment and impact to determine if the outcomes are what was expected. If outcomes are not as expected, the cycle continues through further planning and intervention if resources permit. Monitoring requires expertise on the environmental system being evaluated as well as the indictors used (for example, changes in biodiversity of a river if water quality is improved or the ability for a flood defense wall to remain functional over time following flood events). The expertise and long timescales involved mean that this is the stage of management most often under-resourced and therefore compromised. Ideally, monitoring efforts should be supported by publication of findings so that they may inform the planning stage of other management efforts.

These four stages are demonstrated in environmental management systems (EMS), which are most specifically aimed at ensuring that (usually manufacturing or servicing) organizations are compliant with regulations that seek to limit environmental impact, and that both resource use and production of waste are minimized. With environmental management systems, the organization’s systems of operation are usually assessed to determine where production processes can be made less environmentally harmful (e.g., by using chemicals that are less likely to impact the environment, use of biodegradable materials, reducing resource use, or recycling materials). Plans are made based on this initial assessment, taking into account cost–benefit analyses and environmental regulations. Actions are usually linked to changing the behavior of the organization and may range from relatively simple changes, such as where manufacturing materials are sourced from, to large-scale changes that radically alter, for example, a manufacturing process or the way a service is delivered. Finally, monitoring of the changes against set objectives determines the effectiveness of the actions to inform decisions on whether further changes should be made – for example,
whether changes in manufacturing materials resulted in a reduction in carbon emissions during the manufacturing process, and whether this was at the level expected.

Environmental management systems are often informed by the International Organization for Standardization (ISO) 41000 standards. These are a group of voluntary standards that are aimed at helping organizations to set up effective environmental management systems, and (because they are universal) allow comparisons of best practice between organizations. The standards also include a commitment to continual improvement, so that the processes of environmental management are revisited to allow further progress as new challenges and opportunities (e.g., technological developments) arise. Conformity to ISO 41000 standards is usually validated by appropriate external parties (not the ISO), and helps to show environmental awareness and responsibility on the part of the organization implementing the environmental management system. The ISO 41000 standards have been an important part of European Union’s Eco-Management and Audit Scheme (EMAS), the latest version of which (EMAS III, which again is universal and not limited to the EU) was initiated in 2010. This represents the most recent phase in the developments of universal environmental standards for organizations.

All of these management activities, while guided by appropriate legislation and policies, can sometimes have limited effectiveness and may face many challenges in achieving their objectives. In many cases this is simply because limited knowledge exists on the environmental systems and processes that we are attempting to manage, and they are more dynamic and complex than is often assumed.

Environmental management has been limited by a perception of the environment being formed of relatively simple and predictable systems composed of components that may interact, but can be understood in relative isolation. This has stemmed from a reductionist approach to understanding natural systems, which proceeds from the view that the behavior of a system can be understood by investigating individual components or subsystems in isolation and then fitting them together, like a machine. Some scientific disciplines (e.g., biology, chemistry) have historically been firmly grounded in reductionist approaches and have restricted fields of acceptable enquiry that limit more holistic understandings of systems at higher levels of organization. Though ecology and environmental science have never had fiercely reductionist approaches to scientific investigation, early studies did emphasize particular ecosystem components or processes and their capacity for prediction. An example can be found in early ecosystem concepts such as succession, which originally emphasized predictability of ecological community composition and dynamics, with a climax state that would remain stable unless disturbed. Presently, the many subdisciplines of ecology that focus on specific ecosystem components (mammal ecology, landscape ecology), biogeographical contexts (tropical ecology, desert ecology) or applications (restoration ecology, ecological engineering), and the many branches and scales of investigation found within environmental science, to some extent reflect this tendency for reduction.

This influenced many early forms of environmental management, which tended to focus on only a limited number of system components or processes as with the following two examples.
1 Management of a species population that is harvested on a regular basis, such as marine fisheries (Gaichas 2008). A reductionist approach using relatively simple single-species population models would often be employed, sometimes supported by population monitoring, to provide a measure of the maximum sustainable yield (MSY) that could be removed from an area within a given timeframe (a year, for example) without causing decline or collapse of the target population. However, such approaches failed to account for variability in population dynamics and wider (nonharvest) impacts on fish populations, with the result that target populations in some fisheries have crashed (Gaichas 2008).

2 The setting of critical loads for contamination by environmental pollutants. Critical loads for pollutants are those below which there is considered to be no significantly harmful environmental effect (Payne et al. 2013). Determining the critical load of a pollutant within a given environmental system, and then setting industry standards for pollution emissions, is difficult. Critical loads that were set based on a limited set of indicators and scales of investigation (such as individual species or environmental systems) have repeatedly been shown to miss chronic impacts that may result from accumulation over time (e.g., lead accumulation and adverse health effects in humans due to environmental exposure) and the breaching of thresholds for significant change (e.g., eutrophication in lakes due to nitrogen and phosphorus inputs). This approach also failed to recognize the differential sensitivity of ecosystems to pollutants, such as the varied responses of particular ecosystems or plant communities to nitrogen deposition (Payne et al. 2013). Consequently, critical loads are frequently shown to be limited in their capacity to prevent or mitigate environmental impacts from pollutants.

The “traditional” forms of management exemplified above generally assumed that (i) the component, process or subsystem under consideration was understandable and manageable in isolation from the wider system, and (ii) that an environmental system that was disturbed in some way would return to an equilibrium state and therefore ecological components and processes (which we now recognize as forming the basis for ecosystem services) were infinitely recoverable or replaceable when impacts ceased. But initial faith in these approaches was lost partly because they relied on modeling and all models are simplifications of complex systems with limitations that, when not fully appreciated, may create false confidence in predictive ability. This is true at a range of spatial and temporal scales; for example, early attempts to model ecosystems more holistically and enhance predictability met with limited success due to over-simplification, though of course wider understanding of ecosystem process and function was gained. As an example, management of biological resources often requires an understanding of food webs and the interdependencies of species. Despite ongoing research spanning the best part of a century and the establishment of general laws relating to food web structure, processes, and characteristics, food web models still retain substantial uncertainty and limited predictive capacity in relation to environmental change, somewhat restricting their potential as a management tool (Chiu 2013).

The reductionist approach to environmental phenomena gradually began to be replaced by systems thinking, which emphasizes patterns, relationships, and interactions between components, rather than the isolated study of particular
components or processes. With the advent of complexity theory, chaos theory, nonequilibrium thermodynamics, and self-organization from the 1970s, environmental systems increasingly began to be seen as complex adaptive systems, wherein the broader system emerges from multiple and complex interactions between components and subsystems occurring at more local scales, and is therefore dependent on the outcomes of those interactions (see Levin 1998).

Environmental management consequently became much more complex because a narrow focus left substantial uncertainty that can limit effectiveness. It is increasingly recognized that systems thinking that considers complexity and dynamics is important for guiding management efforts and this has been something of a shift in management principles.

Adaptive cycles and resilience in environmental systems

It is now well established that environmental systems demonstrate notable dynamism and nonlinearity in their functioning. For example, woody species exhibit increased growth as atmospheric CO$_2$ levels increase, but at some point a threshold is reached whereupon further additions have no effect. Likewise, a temperature threshold exists at which permafrost in the soil thaws, releasing stored carbon and leading to a positive feedback in atmospheric warming. Increasingly, coupled human–environmental systems across a range of scales have been found to display nonlinear processes and threshold effects. All of the major global environmental issues that humanity is faced with in the Anthropocene (including climate change, biodiversity loss, freshwater use and land-use change, as examples), display such characteristics, which creates substantial uncertainty in their understanding and management.

These characteristics are also illustrated in the concept of the *adaptive cycle* that environmental systems experience (in particular socioecological systems, which are composed of the biophysical environment and the human activities and societies that are linked to, rely upon, and fundamentally influence, that environment) (Holling 2001; Figure 2). The adaptive cycle represents several phases through which a system progresses. Some phases exhibit the relatively steady accumulation of “resources” (e.g., biophysical aspects such as species, nutrients, soil quality alongside social aspects such as knowledge, productivity, quality of life), while others display a rapid release of resources and system reorganization. The four phases are the following.

1. Exploitation ($r$), wherein a system is in an emergent state and available resources (e.g., nutrients, seed banks, financial capital, labor) are rapidly being utilized.
Conservation ($K$), wherein the system experiences some dynamic stability, and resources that have been utilized in the exploitation phase are again accumulated (nutrients are returned to the soil, crops are produced, money is made, quality of life increases).

Collapse or release ($\Omega$), wherein the system experiences some form of disturbance (whether externally or internally induced) and is pushed over a threshold into a new phase or “state” (see below), and accumulated resources are released.

Reorganization ($\alpha$), wherein a new system or state begins to form and potentially new components, processes and configurations may emerge, prior to the beginning of the exploitation phase.

The adaptive cycle recognizes that environmental systems can go through significant changes over time and may therefore exist for long periods of time in relatively stable “states” that differ substantially from each other. As environmental stresses or disturbances occur, the current system configuration is pushed toward a threshold and the system becomes less stable and more “precarious,” until finally the threshold is exceeded and the system changes state into a configuration no longer representative of the key characteristics, functions, or processes that originally defined the system. It is important to recognize, however, that the adaptive cycle is not “fixed” so that the same states will cycle indefinitely, but rather that random or unpredictable processes operating throughout the cycle will influence the states that may emerge. As a result, there are multiple possible trajectories of change within the cycle and there is a dependency of a given state on what has happened in the past (sometimes considered a form of “path dependency”). Consequently, two conservation ($K$) phases may be quite different, representing different stable states based on, for example, past changes to the components present within the system.

The capacity of a system to remain within a given phase or state despite disturbance is termed “resilience” and this has recently become an important concept in environmental management (Groffman et al. 2006). Human impacts on the environment tend to reduce resilience and accelerate change, pushing the system toward thresholds where it may move to another phase and/or another stable state. So much environmental management is fundamentally aimed at ensuring that a particular environment or socioecological system remains within the phase/state that is most appropriate for human use. This is usually the conservation phase ($K$) wherein resources are accumulated. Though management is rarely expressed using resilience terminology, recognition of this is useful for forming more comprehensive approaches to management. Historically, the focus of management has been on maintaining environmental stability and preventing impact where possible, but there is now greater emphasis on allowing flexibility in management approaches, recognizing that systems are complex and that responses may not be as anticipated. There is increasing focus on not just avoiding “critical loads” but also improving adaptive capacity, or flexibility of response, within environmental systems as a form of management. An example is response to the potential impacts of climate change in cities, which has moved beyond simple local or regional mitigation of greenhouse gas emissions to urban planning efforts that incorporate green building design and use of green space with an anticipated level of climate change to ensure that cities are “future-proofed” against, for example, increasing temperatures and reduced water availability (e.g., Hamin and Gurran 2009). Ultimately, systems theory has led to a paradigm...
shift in environmental management principles, though it is taking some time to work through to practice.

New approaches to environmental management

Environmental management is increasingly moving away from simple regulatory exercises and the imperfect practice of the precautionary principle. New ways of approaching and understanding environmental management from a systems perspective have emerged, such as incorporation of a wider range of interdisciplinary expertise in the management process through the application of postnormal scientific approaches, the practice of adaptive management, and incorporation of a sustainable development framework.

Postnormal science is a modern approach to addressing environmental issues that require action, but which involve significant complexity and uncertainty so that management outcomes are not always obvious or easy to predict. It encourages the contributions of a wide range of stakeholders who may hold forms of knowledge that are relevant to the environment system to be managed (e.g., local environmental knowledge or the desirability of management outcomes for the local community), and therefore adds to any scientific expertise that is brought to bear on the issue. This wider group of stakeholders is termed the “extended peer community.” These different levels of expertise fit into the management stages described above and may help to ensure success by utilizing local knowledge, harnessing societal resources (such as the use of citizen science for assessment or monitoring), and ensuring that management aims and objectives are realistic and appropriate.

Postnormal thinking has been applied to environmental management issues such as biological conservation, climate change, agroecosystem health, and sustainable development (e.g., Turnpenny, Jones, and Lorenzoni 2011). Importantly, it involves an acknowledgment that the complexity of many environmental issues means that multiple management responses and outcomes are possible, and that these may not remain static over time. As a variety of management outcomes are possible, these will be valued in different ways by stakeholders, and it may be difficult to find solutions that satisfy everybody. Because of this, it is important that the “extended peer community” includes stakeholders who can both contribute useful knowledge and work with others toward satisfactory management outcomes for the community as a whole.

Within this context, interdisciplinarity is increasingly defining modern approaches to environmental management. Mismatches between disciplinary principles and practices remain challenges that need to be overcome to maintain progress, but this remains an area where lots of expertise is being combined, facilitated by the extended peer community. For example, wildlife management efforts may be more effective when modern scientific approaches (e.g., population monitoring and modeling, systems analysis) are combined with sociocultural understanding, such as local ecological knowledge that informs on how and why the species under consideration may be valued by communities, and what localized actions are most likely to be successful (e.g., Pretty 2011).

These elements are brought together to some extent in the concept of adaptive management, which holds that management is a process of continued learning and response, without necessarily a neat solution or stable endpoint and which relinquishes some ideologies of control over the system (McLain and Lee 1996). It also embraces
different types of knowledge and recognizes that different stakeholders may hold varying types of expertise, opinion, and objectives, which may need to be reconciled in management. Adaptive management approaches are increasingly being applied to environmental problems and ultimately accept the dynamism and resilience inherent to environmental systems and thus that management practices must become less rigid and more accommodating of uncertainty, including of outcomes of interventions. The shifting template for urban river improvement in response to the Water Framework Directive mentioned above is an example of adaptive management. Likewise, attempts at control of sewerage discharge into urban rivers during high rainfall events, for example, within an adaptive management framework, would incorporate the recognition that changes in factors such as urban weather patterns, population and building density, and river hydrology may mean that some discharge is unavoidable on certain occasions and adapt management objectives and actions with these limitations in mind.

Linked to adaptive management is adaptive co-management. In this case, the continued learning and response of adaptive management is tied to the principles of co-management, wherein government agencies or other authorities share the rights and responsibilities of management with the stakeholders who use the environmental system or its resources. This may be exemplified in the Biodiversity Action Plan framework in the United Kingdom, in which several governmental and nongovernmental partners are responsible for managing particular species or habitats, with further assistance from local stakeholders such as conservation volunteers. Adaptive co-management allows for management plans and actions to be tailored to local requirements and for rapid exchange of information between stakeholders as well as a sharing of power and responsibility that ideally encourages more effective collaboration.

Environmental management is also central to the concepts and applications of sustainable development, which represents an increasingly ubiquitous and normative approach to the environment – that our interactions with the environment should not compromise its ability to function to an appropriate level and therefore it may support future generations. Within a socioecological system, sustainability may most accurately be considered the creation, validation, and maintenance of adaptive capacity over time. It therefore means the sustainability of the system’s capacity to cope with change, including change that may be unpredictable or unknown. Environmental management for sustainability should operate within this framework and not simply be interpreted as ensuring that environmental resource use can continue indefinitely. It requires an understanding of how pressures on interacting components may have wider impacts on the behavior and functioning of the system and is increasingly becoming the focus of management efforts.

Increasingly, a geographical perspective is recognized as being important for environmental management, for two main reasons. First, any environment must be understood within the particular context of its spatial extent and boundaries, its situation within hierarchies of scale, and the place that it represents to the humans that inhabit it – all of which are major geographical concerns. Second, geographers are primarily concerned with human–environment interactions, appreciation of which is essential for understanding how any form of management will influence the environment, both positively and negatively.

SEE ALSO: Environmental assessment techniques; Environmental hazards;
ENVIRONMENTAL MANAGEMENT

Environmental history; Environmental impact assessment; Environmental regulation; Environmental restoration; Environmental risk analysis; Environmental uncertainty

References


Environmental movements and protest

Anthony Bebbington
Clark University, USA

While “social movement” can be a slippery term, it is generally taken to refer to forms of collective action that are sustained over time in pursuit of goals that have some sort of political and normative resonance and are linked to a notion of rights. Thus, a self-help housing process might have very practical and material goals (e.g., to provide housing and basic sanitation), but when the process is also bound up with claims about rights to shelter, to the land on which to build that shelter, and to basic environmental services, or about the failure of government to meet citizens’ basic needs, then most observers would have no qualms at all about calling the process a movement (as, for instance, in the case of the Shack/Slum Dwellers International movement, or SDI: www.sdinet.org).

On the basis of this broad conception of social movement, environmental movements would be those whose agendas have some relationship to a politicized conception of nature (Peet and Watts 2004). That still leaves much latitude of course, reflecting the diversity of politicized “environmentalisms” that circulate. This entry first addresses the range of environmentalisms that might be mobilized through such sustained forms of collective action. It then discusses examples of phenomena that might qualify as environmental movements, drawing attention to the fact that while each is politicized, their political orientation is not necessarily the same. The examples also serve to reflect on movement structures and strategies – which can range from overt and violent protest through to the most muted forms of critique that can all but disappear behind an emphasis on negotiation and dialogue. These diverse strategies constitute part of the broader context of environmental governance in the areas and domains in which they unfold, and the final section comments briefly on the nature of protest and its possible interactions with other governance arrangements.

Environmentalisms

The environment becomes politicized in many ways and these are reflected in different forms of environmentalism. One possible mode of classifying such environmentalisms might be to distinguish among those that emphasize conservation, deep ecology, justice, livelihood, and nationalisms. Of these, conservationist environmentalisms are perhaps the most clearly present in the public sphere and are visible in the form of advertising, media campaigns, and iconic features in the landscape such as protected areas, nature reserves, or private conservation areas. While often cast as conservative (perhaps unsurprisingly given that the concern is to conserve), these are nonetheless sets of ideas that politicize species, biodiversity, and notions of access to the environment (not least because in many instances they strive to impede human access in order to conserve species). Deeper ecologies are similarly conservationist in intent, but for differing reasons. Deep ecological environmentalisms draw on much more profound commitments to
ENVIRONMENTAL MOVEMENTS AND PROTEST

the rights of species and the absence of hierarchy between the rights and needs of nonhuman parts of nature and those of humans.

Other environmentalisms are much more explicitly social in how they politicize the environment. Again at a popular level, perhaps the most widely recognized variant of such environmentalisms is that grounded in ideas of environmental injustice (and its embodiment in environmental justice movements). With particular roots in the United States, environmental justice agendas draw attention to the structurally unequal distribution of environmental costs and benefits, frequently drawing attention to the concentration of environmental damage caused by noxious facilities, waste, highways, and the like in areas occupied by populations that are poorer and frequently not white (Pellow 2007). Environment is politicized here through focusing on the ways in which it is structured and distributed. The “environmentalism of the poor” (Martínez Alier 2002) is also a form of environmental justice, but places much more emphasis on questions of livelihood and access to environmental resources that can in some way enhance the quality, security, and sustainability of poor people’s lives. Finally, there are those bodies of thinking that emphasize the relationship between “nature and nation” (Perreault 2013) and the idea that the environment is central to nationalist projects. There are many forms that this coupling of nature and nation can take ranging from those underlying disastrous forms of politics (as, for instance, in the place of nature within Nazi-ism) through to resource nationalist arguments seeking greater sovereign and social control over environmental “assets” that have been previously controlled by private and generally foreign interests.

These “environmentalisms” cross a very broad spectrum of politicizing discourses on nature and clearly pull in different directions ideologically. The centrality of the “natural” environment is also clearly not the same in each of them. Indeed, in some cases, critics who would seek to delegitimize those using the more socially explicit discourses on nature will argue that the concerns at stake are not environmental at all, and that the term “environment” is being used as no more than a cover for other agendas. However, these are all environmentalisms in that they each politicize the environment and constitute an axis around which sociopolitical movements emerge and organize.

Environmental movements

Environmental movements have emerged for reasons that range from resistance to the destruction or enclosure of natural resources, to struggles to secure access to resources, or to overturn the practices, policies, and structures that drive climate change, through to the desire to build alternative environments and relationships between environment and society. These imagined alternatives can themselves be diverse, including anything from visions for the expansion of green space in cities through to efforts to shift the trajectory of whole political economies toward postextractivist relationships with nature.

Some environmental movements are quite explicitly grounded in conservationist concerns. One example would be the sustained and far-reaching global mobilization surrounding biodiversity conservation and protected area management. At a transnational level, this movement is arguably anchored in large non-governmental organizations (NGOs) such as The Nature Conservancy, Conservation International, Worldwide Fund for Nature, Wildlife Conservation Society, various zoological and botanical societies, and the sorts of foundations that support them (the Gordon and Betty Moore
ENVIRONMENTAL MOVEMENTS AND PROTEST

Foundation, Macarthur Foundation, etc.). But the network is also made up of all the NGOs and groups outside the OECD who work with these organizations, the myriad professionals and scientists that work in or alongside them, networks of givers and volunteers, and so on. While this sort of composition does not look much like the sorts of social movement invoked by the contentious politics literature, it is hard to suggest that something of this size and scope, that has grown over time and secured such presence in public life does not constitute a broad-based movement that has sought to make biodiversity and species protection a global political issue.

That said, there are other similarly global and national apparatuses that also pay attention to conservation but which do look more similar to the types of political actor invoked by literatures on contentious politics. This would be the case, for instance, of the movement constituted by the loose, sometimes organic, sometimes tense, relationships among organizations like Friends of the Earth International, Greenpeace, International Rivers Network, Rainforest Action Network, Amazon Alliance, the Goldman Environmental Prize, the many national and local NGOs, social movements, public interest lawyers and activists that work with them, and the foundations and individuals that mobilize resources to support the activities of these organizations and activists. In some sense, this network is also a movement of movements.

The increasingly visible and important movement against climate change shares similar characteristics: indeed it involves many of the actors noted above while having iconic organizations of its own (such as 350.org). In the final instance this is a movement about conservation – of both human and species’ occupation of a habitable Earth – but with notions of global conservation that are more and less assertively politicized. Degrees of contentiousness vary across the movement, but as a whole the actors involved in this struggle seek to change the beliefs and ideologies that legitimate, deny, or ignore climate change as well as transform the material structures and practices that drive the production of greenhouse gases. Some understand these structures as problems of capitalism, other as problems of technology, and others as problems of culture, and as a result the tone and targets of their interventions vary – from stopping coal production worldwide through to promoting broad-based “green” technological change and much else.

While these examples suggest a transnational dimension, most movements are, of course, national and subnational in scope. Such movements also manifest a similar type of networked structure embracing local groups, NGOs, activists, scientists, lawyers, and citizens. In their analysis of how Brazil slowly but surely saw a greening of its government agencies, laws, and public debate, for instance, Kathy Hochstetler and Mimi Keck (2007) draw attention to the cumulative effects of just such a loose network of activists and organizations running the range from the local to the national scale. Their study also draws attention to the deep historical roots of such processes.

Just as transnational movements are constituted by actors operating at different scales, so too national environmental movements tend to be made up of a mix of national organizations and subnational movements. While such subnational movements may well have narratives on national and global environmental politics, their primary concerns tend to be with locally and regionally specific environmental issues. The literature in political ecology, especially within geography and anthropology, explores many examples of such movements. Once again, the nature of such movements and the issues that they politicize are diverse. Some of these movements address highly localized issues such as specific dams, mining
projects, oil contamination, or urban waste management. Others, however, may address issues that while still “subnational” reach over wide areas. The US “Wise Use” movement, for instance, extended across a large part of the American West (McCarthy 2002).

As noted earlier, a characteristic of many such movements is that, reflecting the underlying environmentalisms at play, they are often as much about race, ethnicity, and class (less often gender) as they are about the environment. That is, deep down the issues of contention that move them are ones of social injustice as much as of environmental quality. In part for this reason some observers prefer to speak of socioenvironmental movements rather than purely environmental movements – both to be more descriptively accurate and to draw attention to the fact that the issues at stake in environmental disputes are very often ones of rights, justice, and distribution.

In a similar sense, “environment” can often become a part of the narrative of movements that have emerged because of other agendas and then begin to take on environmental issues. Indigenous and aboriginal movements, for instance, typically emerged to address the systematic infringement or complete disregard for their rights to land, territory, language, or existence. With time, however, many such movements have also taken up environmental issues because their territories and livelihoods hinge around forests, rivers, or rangelands that have become subject to enclosure and encroachment by other interests. In New Caledonia, for instance, Kanak people have mobilized around conservation but primarily “as a means of reinforcing their cultural identity through preservation of their cultural heritage, grounded in the landscape” (Horowitz 2008, 258). Indeed, with time some indigenous movements have seen the advantage of assuming this environmental mantle as a way of relating their specific concerns to broader concerns for environmental quality, biodiversity, or climate change and in that way widening the purchase of their narratives and the alliances that they are able to build. While such movements are not environmental movements, environment has become central to their agendas.

In other instances, an apparently environmental mobilization in contexts where there is no prior indigenous movement might still not be best understood as “environmental.” Thus, for example, when the Ok Tedi mine in Papua New Guinea dumped tailings into riverine systems and so induced the mobilization and protests of communities downstream in a process that ultimately built a wide network of transnational alliances and led to lengthy legal disputes in Australia, it is an open question as to how far the initial mobilization was because of the environmental impact qua environmental impact, or because this caused massive disruption to modes of livelihood and living that were central to the identities of groups living downstream (Kirsch 2014). While at one level this distinction may seem like splitting hairs, analytically it is important because it has implications for how these movements are spoken of and how the larger issues at stake in such protests are understood.

In contrast to indigenous movements, human rights movements have been more ambivalent to incorporating environmental concerns into their agenda. In Peru, for instance, the national human rights coordinator (an umbrella group within the human rights movement) was relatively slow to accept the membership of environmental defense organizations that were offering legal support to communities affected by extractive industries. The rationale appeared to be that traditional human rights concerns (incarceration, post-internal conflict reparations, political liberties, etc.) were still too far from being secured to justify a broadening of the coordinator’s agenda. In a somewhat similar
way, environmental rights are not especially visible in the campaigns of those transnational organizations such as Amnesty International that have traditionally anchored the global human rights movement. While this is changing slowly, reflecting the increasing coupling of natural resource enclosures and severe human rights abuses, what is especially interesting is that some human rights organizations based outside the OECD but with ambitions of greater global projection do pay more attention to environmental rights (see, for instance, the Colombia-based organization, Dejusticia: www.Dejusticia.org).

Finally (for the purposes of this section, certainly not for the breadth of the issue) are those movements organized around principles of deep ecology, nature’s rights, and animal rights. Here the focus is much more clearly on environment qua environment, with movements that are both national and local in scope. Examples of the sorts of organization that anchor such movements might include Earth First! or PETA (People for the Ethical Treatment of Animals), though the ideological orientation and political motivations of people involved in such movements can again run a very wide range from those who are radically committed to direct action to people who are more likely to vote conservatively in national elections but who are deeply committed to the wellbeing of nonhuman species.

“Environmental movement” thus covers a very wide canvas. This reflects, perhaps, the many ways in which “environment” is present in agendas that are not self-consciously environmental, the fact that an environmental ethic can range from the deeply conservative to the equally deeply militant, and also the depth of human transformation of nature meaning that environmental problems, crises, and injustices present themselves in so very many domains of social life. This diversity makes simple classification of movements risky and suggests the importance of understanding member motivations in detail and of recognizing that multiple motivations and identities are often at play in any one given movement. Another implication of such diversity is that approaches to “protest” also differ wildly among these different movements. First, though, this entry presents a commentary on the ways in which such movements are structured.

Environmental movement structures

The organizational and social structures that can sustain “forms of collective action that are sustained over time in pursuit of goals that have … some relationship to a politicized conception of the environment” are diverse. These structures may, or may not, reach across different scales, may or may not involve formal nongovernmental organizations, may or may not incorporate people with presence in government, and may be more, and less, formalized. That said, much literature would suggest that most movements – environmental or not – have to involve “social movement organizations” (McCarthy and Zald 1977) that are able to serve as vehicles that coordinate actions, keep movements moving during periods of relative quiescence, or steward the physical and financial resources that are so essential for movement survival and growth. These more formal organizational roles have been played by NGOs, religious organizations, student associations, political parties, or environmental clubs, among others.

The role of such organizations within environmental movements is critical but awkward. Because they tend to centralize and anchor they can easily become dominant within movement processes, leading easily to tensions with other parts of movements especially when the movement organizations use this dominance to pursue their own institutional agendas and views on
movement strategy. This has been a recurring problem with large international environmental NGOs who have been the object of much criticism in that they have imposed their agendas and strategies within global and national conservation movements and in the process have excluded the views of local communities, indigenous organizations and other environmental NGOs (Chapin 2004). This is a particularly acute problem, but the issue at stake is a more general challenge for environmental movements: to hold in check the centripetal forces surrounding the movement organizations whose capacities are so vital to movement strength but also so potentially disruptive to movement cohesion.

A second generic structural issue for environmental movements derives from themes raised in the preceding section and relates to the question of boundaries: where does one movement end and another movement start, and when is a particular organization part of a movement or just an ally of the movement around specific issues? Is an indigenous federation that is concerned with forest protection part of a broader environmental movement, or merely an ally on specific issues related to forest governance? Can an indigenous peoples’ federation be part of both an environmental movement and an indigenous movement at the same time? At one level these questions can again seem like splitting hairs, but they also matter for how these movements are talked about, how strong one takes the “environmental” movement to be, and the sorts of environmentalism more likely to dominate the ways in which movements frame their protests and the alternatives that they pursue.

Protest, strategy, and environmental governance

The title of this entry combines two phenomena that are not necessarily associated with each other: not all environmental movements are involved in the business of overt protest and not all protests necessarily involve movements. Some protests can reflect relatively spontaneous mobilizations in the face of some sort of environmental concern, mobilizations which can then unravel almost as quickly as they form; and some movements can emphasize negotiation and persuasion rather than confrontation and protest as strategies to pursue their ends. This then makes the conjunction of environmental movement and overt protest especially interesting, raising questions as to what constitutes “protest,” what leads environmental movements to engage in more explicit protest, what the effects of protest might be and the factors that might lead protest to have effects on environmental valuation and governance that are in line with the objectives of those protesting.

Protest can range from everyday foot-dragging (Scott 1985) through to the coordination of collective, publicly visible, and violent direct action. Across this spectrum of protest options, movements evidently fall under the collective and coordinated. Some movements, and especially some social movement organizations, prioritize direct action: blocking access to laboratories doing animal testing, camping out in trees slated for clear-cutting, closing access roads to mine sites, organizing mass climate protests in downtown New York, and more generally confronting the objects of their (socio-)environmental grievances head on. At the other extreme, some movements – and again, in particular some movement organizations – prioritize avoiding confrontation and violence at all costs. Here the emphasis is instead on finding ways to negotiate with, coax, and cajole other parties to change their environmental behavior. This is a strategy that favors dialogue, behind the scenes discussions, and joint implementation of activities oriented toward changing behavior,
among other options. Such strategies might still be thought of as protest, insofar as they reflect some level of complaint and dissidence at the current state of affairs. But these are actions of protest in spirit more than in action.

Protest can also be pursued through the crafting of alternative narratives as much as through overt contention. Indeed, there is a strong strand in social movement thinking that would say that the primary contribution of movements, indeed the feature that defines a phenomenon as a movement, is the crafting of discourses that contest taken for granted ideas and build different ways of viewing the world. In the domain of environmental movements, an example of such crafting of narratives would be the very act of building and communicating ways of understanding nature, and the relationships between society and environment that imply significant rethinking of the organization of the economy, for instance. Another example would be the movements that have steadily fostered a reframing of climate change debates in the public sphere.

Where movements fall across this range of strategies (dialogue, contention, narrative) depends very much on their origins, their relationships to private finance (movement organizations receiving corporate contributions are much more likely to focus on negotiation than confrontation), their internal culture, and their assessment of the political environment in which they operate (what some refer to as the “political opportunity structure”). In some instances, even for militant organizations, protest may seem simply too dangerous and likely to elicit repression. It may also be the case that some forms of direct action (such as road blocking) have been criminalized, leading some movements to step back from them. Other movement organizations may come to the view that it makes more sense to work with change agents within the companies or government agencies that they seek to affect. Whether they come to this view or not depends also on the existence or not of such change agents, and indeed other organizations may come to opposite conclusions and decide that there are so few opportunities for change within existing structures of private and public governance that they are left with few options but to protest openly.

Another calculation facing movements regards whether to use violence or not. While it may well be that movement leaders and organizations are unable to control the mood of mobilizations or the provocations of other parties, with the effect that violence occurs anyway, calculations over the role of violence can be very real discussions within socioenvironmental movements. One line of reasoning draws on the historical record of contentious politics to argue that at certain times certain forms of violence have been critical in catalyzing progressive institutional change. Other lines of argument reason that acts of violence serve to undermine the legitimacy of the movement’s agenda (which is often one of its most important resources) and almost always induce a hardening of state authority and restrictions on freedoms of expression. There can clearly be no simple response to such debates, and at times both arguments can seem simultaneously correct.

As one example, a large scale socioenvironmental protest in lowland Peru in 2009, led by indigenous movement organizations but involving many more than just indigenous people, culminated in the death of 33 people, 23 of whom were policemen. Many have argued that the Peruvian government’s passage of legislation to enforce prior consultation with indigenous peoples before large scale investments can proceed in their territories was a direct result of this conflict. But it is also true that the Peruvian state has used the violence as a pretext to weaken movements,
with 53 indigenous leaders now being tried for crimes that the public prosecutor claims were committed during the protest. Some of these leaders face the possibility of life sentences.

A further strategic question facing environmental movements is how far to pursue their objectives domestically or internationally. This has much to do with movement structure on the one hand and political opportunity structures on the other. Movements which have only weak, or no, international members or allies almost by definition focus their attention on national and subnational levels, seeking to pursue their goals through changing national legislation, or the actions of corporations, governments, and others through national campaigns. Where national political opportunities limit what can be achieved domestically, movements have sought to exercise pressure outside their countries in the hope that this would ultimately lead to subsequent pressure on their own governments. Keck and Sikkink (1997) referred to this as the “boomerang” strategy, illustrating it with cases from environmental, women’s, and human rights movements. In this strategy, movement organizations that are unable to influence their own state directly, seek through their own action or that of allies to exercise influence indirectly via influencing other states, intergovernmental, or international bodies that have influence over their home state. For instance, Latin American environmental movement organizations have tried to influence their own states’ environmental policies by mobilizing the Inter-American Commission on Human Rights, the international Roman Catholic Church, International Financial Institutions, the UN, front page stories in The New York Times, or other similar international vehicles deemed to have leverage. In the private sector, a structurally similar strategy involves trying to influence activist shareholders or the corporate headquarters of international companies operating within their country. Other strategies, frequently led by international NGOs and legal support bodies, include efforts to influence home state legislation or regulation on environmental standards.

The drawbacks of such international strategies are many. National movements can easily lose control of them, as organizations based in Washington DC, Brussels, New York, London, and elsewhere take over. Perhaps more importantly, the decision to bypass national judicial, legal, and political institutions may reduce the scope for organizational learning and the extent to which these institutions ultimately change. The jury is very much out (and likely always will be) as to which of these options is “better,” not least because the answer to that question clearly varies depending on the case, the national context and the moment in time, and many would argue that the best strategies are those that address environmental problems by intervening across the whole reach of the global production network or global policy network at stake. By the same token, the choice between direct action and negotiation, though a real one for individual movement organizations, may be a moot one at an aggregate level where it is often argued that the best movement strategies are the “good cop/bad cop” ones which combine reform specific support and cajoling from good cops promoting changes within those private and public organizations that feel obligated to respond to the head-on and generalized pressure from the bad cop.

SEE ALSO: Environment, democracy, and public participation; Environment and development; Environmental (in)justice; Environmentalism, grassroots; Participatory geographies; Regional political movements; Social justice; Social movements
References


Environmental performance, practice, and affect

Gordon Waitt
University of Wollongong, Australia

Environmental performance, practice, and affect is a rich and variegated field of research that critiques dominant ways of thinking found within environmental science that conceives the environment as objective, external to people and inanimate. For most environmental scientists, the environment is “out there” waiting to be discovered, measured, and observed in a variety of ways. This notion of the environment as separate from and subordinate to people as well as having machine-like qualities is associated with the scientific method advanced in seventeenth-century Europe by Rene Descartes. This recourse to split the environment from humans is an example of the dualisms of much enlightenment thinking that privileges the mind over the body, solids over fluids, humans over nonhumans, and men over women.

Much work conducted by environmental scientists and environmental managers continues to draw on dualist thinking. Instead, critical geographers engaging with the environment seek to provide accounts that consider the different agencies, materialities, and dynamics. Environmental performance, practice, and affect is therefore a field of research that emerged within the context of poststructuralist approaches that advocate for new approaches that trouble understanding of the environment as pure, stable, and singular form. Instead, this field of research embraces the notion of the environment composed of multiple events, trajectories, and rhythms both human and nonhuman, over different scales, and comprised of a multiplicity of forces (social, cultural, and affective). Hence, the environment cannot be known in advance. Instead, the coming into being of the environment is conceived as an immanent process that involves human and nonhumans. The next section outlines how this field has opened up new conversations around environmental management, politics, and ethics. To conclude, similarities and difference are noted between two strands within this field (i) poststructuralist feminism and (ii) more-than-representational geographies.

Embodied practices and performed environments

What are the implications arising from a starting point of enquiry for environmental geography that does not begin with an already preconfigured environment, “out there,” waiting to be discovered? What if this account of the environment was just one of many social constructions of the nonhuman world? Nigel Thrift, Donna Harraway, and Bruno Latour have propagated lines of research challenging the purification of the world into two distinct categories, culture and nature, and advocate for an approach that rethinks the environment, subjectivities, and society, as always a relational achievement between human and nonhuman components. Environmental geographers now discuss “more-than-human” and “multinatural” worlds, epitomized by how places for nature and subjectivities

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0366
are co-constituted, multiple, immanent, and enlivened through practice, performance, and affect. For example, recent explorations of the practice of walking by many geographers have sought to avoid erasing the agency of nonhuman life and instead explored the more-than-human spatialities performed by people, plants, and animals. These “more-than-human” accounts of the practice of walking moves geography toward understandings of the environment that is as impure, multiple, immanent, embodied, situated, and performative. For environmental geographers, an approach that shifts agency beyond human-centered accounts has important conceptual implications.

The first point is that there is no preconfigured environment, or indeed walker. Instead, there are embodied practices and emergent configurations of self and the environment. People actively become walkers as they make, and remake, places understood as appropriate for particular styles of walking. Such configurations in part rely on learning particular social norms; characterized by styles of dressing bodies, particular walking techniques and technologies, and following selective routes. In part, learning to walk illustrates the importance of knowledge-in-practice. Only through regular practice do bodies become attuned or trained to particular styles of walking. For geographers, particular styles of walking, like rambling in England, bushwalking in Australia, and tramping in New Zealand, may be understood as a place-and boundary-making process by helping to zone off particular locations as places for nature that may be spoken about as “natural environments.” At the same time, places designated as suitable for particular walking practices, such as national parks, nature/urban reserves, and beaches are often entangled within particular classed, gendered, and racialized geographies of belonging. For example, Waitt, Gill, and Head (2008) discuss how, in Australia, it is impossible to overlook the historical weight of colonial ideas that embed places designated as urban nature reserves, National Parks, or the “outback” in British and European settler ideas of wilderness and pristine nature. Yet, these are places Indigenous Australians continue to call home over thousands of years.

The second point is that the particular understanding of the environment configured by walking is always more than a human activity. As Wylie (2005) demonstrates there is the gradient, weather, and temperature to consider; as well as the sight, touch, and smell of animals, plants, rocks, and footpath. A performed and practiced environment must equally embrace both human and nonhuman life in a process of creativity, improvisation, and emergence. Here, attention turns to the affective energies of walking; characterized as the change in intensity of a non-conscious force or energy that flows within, between, or through bodies that guides human/nonhuman connections or disconnections. Fundamental to what motivates, or discourages, a particular practice of walking are non-conscious affective intensities that may be conveyed through a wide array of emotions including joy, love, wonder, awe, and frustration. Affectivity intensity is characterized as the ebb and flow of a force that helps inform how we become differently present by aligning us to our embodied histories, and telling us about our connection or disconnection with the present. There is, therefore, the possibility of an affective politics at work in the everyday choices people make that then play out in terms of, say, carbon emissions through modes of transport, heating, cooling, and eating.

The third and related point is that, while different practices of walking are embedded in particular essentialized ideas of the environment such as patriotism, Romanticism, the wilderness ideal, the frontier or the rural idyll that
“pure” spaces for nature, there is always the possibility of creativity, surprises, inventiveness, happenstance as nonhumans enroll the attention of walkers, not to mention the pain of the footsore. The sum effect is to sharpen us toward thinking about the environment not as a static backdrop of flora and fauna “out there” separate from humans. Instead, the environment is conceived as a relational achievement, a process of becoming – among a set of located, reciprocal, dynamic, and embodied connections and disconnections between humans and nonhuman entities. As such, the dynamics of the process of (dis)connections between human and nonhuman entities is conceived to make, remake, and undo a spatial order in the world. Hence, while people may draw on ideas that configure environments as seemingly pristine – such as a nature reserve, national park, or urban reserve – at the same time through the embodied skilled practice of walking, people may incorporate a whole set of practices and objects as enhancing or detracting from their experience – including mountain bikes, dams, bridges, music, footpaths, litter, rubbish-bins, signposts, benches, and restaurants. The taken-for-granted understanding of what, or who belongs, or is excluded from a particular spatial context may be made resilient or stretched.

A fourth point is that embodied environmental knowledge derived from daily walking through places designated as urban parks and reserves has important practical implications for environmental managers. Detailed ethnographic investigations reveal the multinatural modes of making sense of, and relating to the environment. In contrast to the conventional conversation story of human history that positions people as an environmental threat, a more-than-human approach has enabled possibilities to discuss the different ways in which people care for, and are attached to what they may make sense of as the environment. Such a conversation is possible by how this field of work abandons notions of a pure environment. Consequently, geographers are now thinking through the implications of environmental policies informed by the notion of multinatural approach, opening possibilities to acknowledge and appreciate the important role of people in the protection of plants, animals, rivers, and places. For example, environmental managers may better motivate people into conservation programs by tapping into the affective and emotional relations of awe, care, enchantment, loss, mourning, joy, and wonder; rather than programs characterized by the dispassionate technical ways of talking about the environment such as “buffer zones,” “green belts,” and “endangered species habitat.” At the same time, important questions are being raised in thinking through the practical application of a multinatural approach. Through what processes are people recruited? For what purpose? Who makes the decision? What ideas matter from the past?

Environmental politics as everyday practices

Environmental politics as everyday practice is emphasized in recent work that pays attention to affect, and investigates how bodily judgments inform everyday choices like laundering clothes, showering, waste disposal, refrigeration, cooking, eating, and shopping. Here, environmental performance, practice, and affect is located within a politics of change that centers around bodily judgments, “gut reactions,” and the visceral, rather than disciplining the mind. For example, many authors illustrate how the environmental politics, at various parts of the food chain, plays out around the bodily capacity to be affected and to affect. In grappling with what foods become accepted and or rejected from our diets, and thus shapes the environment
through agricultural practices, many geographers have chosen to emphasize the materiality of food (including texture, taste, smell, and color) to emphasize what food “does” through practices of production, marketing, shopping, preparation, plating-up, and ingestion, rather than focusing solely on what food “means.”

For example, Hayes-Conroy and Hayes-Conroy (2010) deployed a visceral approach to explore the politics of the “Slow Food” movement. Slow Food is often promoted as a rational environmentally sustainable choice. The pleasures of eating are aligned with being sensitive to the planet while supporting “local” business. The Slow Food movement exemplifies one way in which ideals of healthy bodies and sustainability plays out as a badge of social distinction for “good” consumer citizens to seek out what to eat. Hayes-Conroy and Hayes-Conroy drew attention to what they termed the “visceral politics” of food, characterized by bodily judgments commonly referred to as moods, sensations, or “gut reactions.” Ingesting food is shown to be how representations of food exist materially in the body. They reconceptualized taste within the visceral realm, recognizing that a taste for something cannot be viewed apart from the political locations we inhabit; for example, how people live their lives according to social constructions of class, gender, ethnicity, and age. The point is that visceral experiences of the textures, tastes, smells, and colors of foods (such as disgust or delight), provide important bodily clues to how everyday food choices are not simply an intellectual project, but also a corporeal one, and part of how a person inhabits the world. So, visceral responses to eating are a dynamic and differentiating force between bodies and they thus shape environmental politics through our everyday food choices.

Another example is kangaroo meat, which is rarely cooked in Australian homes despite environmental and climate change scientists supporting it because the digestive tracts of kangaroos produce less methane than either cows or sheep. Paying attention to the affective and emotional relations of shopping, cooking, and eating kangaroo shows how neoliberal climate change policies fail to connect with how food is bought, prepared, and eaten in everyday life. Embodied geographies provide vital clues to how the sensual body, affect, and emotions work alongside sets of ideas about food, parenting, home, and travel, to better understand the choices that compel participants to eat or not eat kangaroo in different places. Visceral disgust at the sight, smell, texture, and taste of kangaroo allowed gendered, classed, and racialized subjectivities to be felt, expressed, and questioned. Remaining alert to “gut reactions” provides valuable insights to the challenges of a changing climate by moving beyond highly simplified assumptions in neoliberal policies of people as rational “consumers.”

Rethinking environmentalism

Rethinking environmentalism is a central theme of the field of environmental practice, performance, and affect. Focusing on the embodied knowledge of practices that involve human–nonhuman animal relationships, this field questions human-centered ethical and political traditions of environmentalism, namely ecocentrism (ideas that emphasize the intrinsic importance of nature for humanity, such as wilderness or the sublime) and technocentrism (ideas that emphasize the instrumental importance of the environment and suggest human political, scientific, or technological ingenuity can resolve environmental challenges). Inspired by the work of Donna Haraway on interspecies relations, this works calls for rethinking environmentalism as an ethics of shared suffering and response abilities with nonhuman others, such
as cats, dogs, sheep, elephants, trees, and seeds. Finding ethical ways of living with plants and animals is conceived as an everyday practice. Animal/plant encounters are understood to comprise skilled practices that orient, dispose, make, and unmake people in affective ways. Through practice and physical engagement with nonhuman others, not only do we learn about difference, but also learn something about their world and find better ways of living with nonhuman others.

For example, Jamie Lorimer (2012) researched human–elephant relationships and their implications for rethinking environmentalism. Lorimer makes use of poststructuralist ideas that do not neatly align touch with texture, but rather conceives of touch (and indeed smell, sound, taste, and sight) as always multisensory, part of the process of making, remaking, and unmaking relationships across difference into a sociospatial formation that makes sense as a coherent whole. First, Lorimer demonstrates how the affective capacities of touching elephants, and the reciprocity of being touched by elephants, are an integral part of an experiential ecotourism economy; alongside touching other animals/birds framed as charismatic including koalas, kangaroos, whales, penguins, and dolphins. Lorimer discusses the tensions between Western volunteers and scientists who travel to Sri Lanka as “scientific ecotourists” in national parks. These are people who have paid for organized corporeal encounters; and often expect their bodies to be affectively charged through encountering an elephant in terms of risk and adventure. Lorimer argues how paying for encountering elephants operates as a form of entitlement. Through the workings of the experiential economy, private desires and access that help in part fashion becoming a scientific ecotourist may trump the ethics of park management schemes of environmental science that advocates against deliberate proximate encounters between humans and elephants. For these reasons, the concept of affect provides insights to the ethics of human–elephant relationships in the context of volunteer nature tourism where people pay for organized touching encounters.

Second, Lorimer illustrates another modality of the reciprocal touch of elephants, that of the embodied knowledge of an experienced mahout, a person who has normally worked with elephants in the forest since childhood. The mahout’s embodied knowledge of individual elephant’s sounds, smells, and touch illustrates how attunement, or training, is crucial to the logic of affect and interspecies communication. Over a lifetime, mahouts learn to attune to an elephant’s oral, haptic, and olfactory clues about its mood, thermal comfort, or appetite, while at the same time learning to be affected by their behavior. Lorimer illustrates how the interests of captive elephants are not well served by either intrinsic environmentalism that neatly categorizes elephants as wild or instrumental environmentalism that categorizes them as working animals. Instead he advocates for a model of companionship and a more-than-human account of environmentalism, arguing that “human-elephant flourishing must recognize how elephants are touched by humans and how, in turn, elephants touch us” (Lorimer 2012, 182). Paying attention to embodied knowledge, the circulation of human and nonhuman affects and knowledge, and the practices through which humans and nonhuman affect and learn to be affected in moments of encounter, are crucial to thinking anew about environmentalism of which geographers write.

**Multinatural ontologies**

There are a diverse array of research strands that bypass enlightenment thinking of the environment by drawing on different philosophers. One
ENVIRONMENTAL PERFORMANCE, PRACTICE, AND AFFECT

strand draws on feminist scholarship including Donna Haraway, Val Plumwood, Elspeth Probyn, Elizabeth Grosz, and Rossi Braidotti, while another resonates with Nigel Thrift’s (2007) non-representational theory (NRT) or affective geographies.

There are similarities but also important differences between these conceptual positions of rethinking the environment in more lively ways. Both approaches share an interest in: (i) relational over oppositional modes of power, which enables a future that is imagined as open to possibilities rather than predicted by uneven power relations; (ii) the significance of everyday practices; (iii) the located body as active in measuring, feeling, shaping, and reshaping our sense of self and place; and (iv) the capacity of affect to shape and transform our self and the world we live in. Equally, both approaches share conceptualization of the body not as a bounded solid whole, but rather as porous, permeable, and active. Strongly influenced by Deleuzian concepts, each emphasizes thinking through the ongoing, provisional process of bringing together diverse elements (bodies, ideas, things, and intensities of affect) into a working order that makes coherent sense (assemblages), where agency does not lie solely within the body-subject. Assemblages hold together, or fall apart, depending on how their diverse elements relate. Finally, both approaches call for methods that trace the textures of everyday life; injecting living, breathing, sensual, emotional, and affective bodies into geographical enquiry.

Nonetheless, there are important differences that can be identified within and between research strands into environmental performance, practice, and affect; particularly around mode of relating to human agency, the body, affect, and emotion (see Lorimer 2008). First, drawing on the work of Massumi and Thrift, non-representational accounts try to avoid thinking around bounded interacting bodies. In contrast, following Haraway one strand of ecofeminist geography emphasizes the organism as the entry point for analysis. Second, non-representational accounts distance themselves from research which privilege subjectivity and nameable emotions and focus on the capacity of bodies to affect and to be affected that precedes, and exceeds any stratified formations of power knowledge. Whereas, following Elspeth Probyn (2000), alongside the capacities of bodies to affect, and be affected, bodies are always located and entwined within discursively produced relations of power, including the most intense tracing of affect in and through the body communicable as emotions. Third, feminist geographers argue that affective geographies often ignore gendered bodily differences, and de-emphasizes the humanity of embodied experiences articulated as socially and spatially mediated emotions. This debate centers on how affective geographies reinstate dualist thinking of mind and body (see Thien 2005). In summary, Bondi (2005, 438) argued that “feminist geographers find research informed by non-representational theory too abstract, too little touched by how people make sense of their lives, and therefore too ‘inhuman,’ ungrounded, distancing, detached and, ironically, disembodied.”

SEE ALSO: Affect; Animal geographies; Bodies and embodiment; Emotional geographies; Feminist geography; Geography and the study of human–environment relations; Hybridity; Identity; Nonrepresentational theory; Phenomenology; Poststructuralism/poststructural geographies; Zoogeography

References


Environmental planning

Graham Haughton
University of Manchester, UK

Environmental planning involves working within both statutory planning systems and wider systems of environmental management, such as water quality and flood risk management. It operates at and across a wide variety of scales, from the site through the local to the regional and beyond, with different types of approach relevant at different scales.

Reflecting its role at the interface of statutory and nonstatutory ways of planning and managing the environment, a central concern of environmental planning is the need to acknowledge and deal with multiple environmental and administrative geographies. Environmental geographies might include floodplains, watersheds, marine and terrestrial ecosystems, landscape character designations, and recreational parks. The geographies adopted by environmental organizations sometimes deliberately involve fuzzy and overlapping boundaries – ecosystems and river basin catchments, for instance, do not lend themselves to clear lines on maps. As a result we find that from forestry authorities to those concerned with species protection or soil preservation, many environmental organizations will create their own geographies that quite avowedly do not coincide with the political-administrative boundaries of local and metropolitan government. The importance of this is that it is the territorial units of government that have provided the mainstay of most statutory planning systems, where the legal and regulatory powers of planning most clearly come to bear.

We can begin to see from this how environmental planning exists as part of the statutory processes of land-use planning, and also in relation to a wide variety of other sectors and organizations – such as water and energy utility companies, state forestry commissions, and catchment partnerships. Each organization will have its own priorities, strategies, sets of stakeholders, funders, and preferred geographies. The challenge for environmental planners then becomes how to reconcile all these different possibilities, from the statutory to the nonstatutory, from the formal to the informal, mediating between often conflicting objectives, all the while seeking to find a balance between protecting private interests and promoting the public good.

Because an essential part of planning involves allocating or denying development rights of different kinds, it is also necessarily political since it involves redistributing power and resources. This has always been true, but it has been brought to the fore by recent work on urban political ecology, which strenuously opposes attempts to think in ways that present humans and nature as somehow separate, arguing instead that they are symbiotically interrelated (Swyngedouw and Heynen 2003). This framework leads to the inevitable conclusion that far from being a neutral science or arbiter of neutral expertise, environmental planning is always a political act, involving a redistribution of winners and losers. As recent scholarship on river basins has demonstrated, whilst those supporting work using river basin geographies can draw on the appeal of their seeming apolitical “naturalness,” in reality...
the decisions about empowering certain scalar fixes over others always involves deeply political choices, for instance privileging certain resource uses or ecological features over others (Molle 2009; Cohen and Bakker 2014).

In related vein, environmental planning requires a continuous negotiation between different understandings of the environment. As such it must find ways of mediating between the many actors who lay claim to privileged knowledge and understandings of how environmental systems work and their impacts on humans, from scientists of various kinds, to lay experts and local communities, each of which can make different types of knowledge claims, often in emotionally charged ways (Castree 2014). For instance, lying behind debates such as how to preserve “pristine nature” in national parks are difficult issues such as whether pristine nature is any longer a meaningful concept and how preserving one particular moment in a landscape’s evolution as somehow “best” privileges certain interests over others.

The need to understand debates around climate change provides an excellent example of how environmental planners must also work with multiple time horizons, dealing with the past, present, and near and distant future. Despite an inevitable obsession with maps and boundaries, environmental planning also requires thinking beyond horizontal planes and surface processes, looking deep into the Earth and oceans and up into the atmosphere – it requires three-dimensional thinking as well as thinking about the intersection of environmental, social, cultural, political, and economic processes.

Environmental planning then is always an exercise in complexity management, dealing with multiple spatial scales, complex interactions between highly variable relational and territorial geographies, different professional and scientific traditions, and disputed knowledge claims, working across a wide range of policy sectors and multiple time frames.

As a discipline, environmental planning has continued to evolve, reflecting wider societal debates and improving environmental knowledge and understanding of nature–society relationships. The evolution of environmental planning ideas is important precisely because they have been hugely influential in shaping our urban and rural landscapes for well over a century. To give a broad sense of how these ideas have developed over time, the remainder of this contribution focuses on two attempts to look at how dominant themes have emerged in different contexts. The first of these is based on a periodization provided by Thomas Daniels (2009) to explain the evolution of environmental planning in the United States – this sets out five broad eras of thinking, which have been refined a little here by reflecting on the emergence of environmental planning in the United Kingdom. The second approach, urban environmental transitions (McGranahan et al. 2001), looks more critically at the evolution
of environmental policy in relation to economic development, providing a framework more oriented to the problems faced in poorer nations and cities.

Daniels calls the first era in his periodization “Getting on The Green Path,” stretching from the nineteenth to the early twentieth century. This was a hugely important period in terms of generating ideas that would continue to influence environmental planning over the next century. In the United States key unifying themes that emerged at this time included the city beautiful movement, and later garden cities and suburbs, the urban parks and playground movement, wilderness protection and conservation of natural resources. Starting with Yellowstone in 1872, the designation of national parks was inspired by a desire to protect areas deemed to be of particular natural or cultural value, with the emphasis initially at least on what were seen to be wilderness areas relatively unblemished by industrial progress. In both the United States and Britain new conservationist movements sprang up in this period pushing for the conservation of land, such as the Sierra Club, founded in 1892 in the United States, and the National Trust, founded in 1895 in Britain.

From the British perspective, mid-nineteenth century legislation to improve the condition of slums in the industrial cities was particularly important. National legislation allowed local authorities to intervene in the market, enabling them to require landlords to make improvements and where necessary to demolish unfit housing. The importance of this legislative intervention was that it demonstrated a key role for central government in regulating to improve the urban environment, acknowledging the limits of unregulated markets, and in creating a new profession, environmental health. During this period too municipal authorities became increasingly active in investing in clean water supplies and in urban drainage and sewerage. Despite limited scientific underpinnings, much of the thinking that emerged during this era still resonates with some of the key themes in environmental planning today – clean water supplies, sanitation, protecting nature, providing places for recreation, and better quality residential areas in cities.

A second era, running from the 1920s to the 1960s, is broadly defined by Daniels as “Regional ecological planning,” involving an increasingly scientific approach to environmental management and the rise of techniques such as environmental impact assessment. In the United Kingdom, the key environmental themes of this period would include the emergence of green belt as a tool that has been ascribed various roles, preventing urban sprawl and the coalescence of neighboring towns, protecting the identity of historic towns, ensuring access to the countryside for urban residents, and encouraging urban regeneration. Following major disputes over allowing public access to the countryside during the 1930s and the changed national mood after World War II, new legislation was introduced in 1949 to allow the designation of National Parks and Areas of Outstanding Natural Beauty (AONBs).

Daniels refers to the third era in US environmental planning (1960s–1970s) as “The Birth of Modern Environmental Planning,” driven particularly by a desire to regulate better against industrial and agricultural pollution in the wake of Rachel Carson’s (1962) influential book *Silent Spring*. This led to the development of federal level legislation, including a Clean Air Act (1970), Water Pollution Control Act (1972), and Safe Drinking Water Act (1974), plus new initiatives for clearing up contaminated sites and addressing the disposal of solid wastes. In the United Kingdom, the great smog of December 1952, which resulted in around 12 000 deaths in
London, kick-started a growing environmental awareness and the emergence of new legislation, including the Clean Air Acts of 1956 and 1968.

By the early 1980s a backlash began to emerge against the earlier regime, which was deemed by some to be an overly centralized “command and control” approach, leading to inflexible and costly uniform standards that did little to incentivize the search for market solutions. This led to the fourth era in Daniels’s schema, “Backlash or Bridge to Sustainability” (1982–2008), during which new market-based approaches emerged, such as the “cap and trade” approach introduced to encourage innovation in how firms sought to reduce greenhouse gas emissions. In this period, economic growth was given greater primacy in policy, with environmental policies expected to demonstrate that they would not prove a drag on the drive for growth. This said, the 1980s and early 1990s also gave birth to the environmental justice movement, with its central concern about the socially and geographically inequitable burdens of polluting activities (Bullard 1994), providing at least some counterbalance to the growing “economization” of environmental planning.

Moving to the present, the dominant themes in environmental planning in the United States are characterized as “Planning for Sustainability and the Global Environment.” In Daniels’s schema, this era (1992–present) overlaps with the previous one, a device that might usefully be adopted for all the periods as there are inevitable overlaps and interplays as new dominant approaches emerge. The year 1992 is significant because of the Rio Earth Summit, which set out a global agenda for addressing sustainable development into the next
century. An equally significant date, however, was 1987, which saw the publication of *Our Common Future*, widely known as the Brundtland Report (World Commission on Environment and Development). This provided a definition of sustainable development that is still in common use today, namely that sustainable development must meet the needs of the present without compromising the ability of future generations to meet their own needs.

There have been many other attempts to define the concept, typically emphasizing the need to give equal consideration to environmental, social, and economic issues, or identifying core principles, such as intergenerational equity, intragenerational equity, and geographical equity (Haughton 1999). In 2004 the UK planning system was for the first time given a clear statutory purpose, to support the pursuit of sustainable development. This gives a sense of the importance of the sustainability debate for planning at this time, with attempts to devise win–win–win solutions (development decisions that provided environmental, social, and economic benefits, rather than trading one against another) gaining considerable professional attention. More recently there has been some disillusionment with the way in which sustainable development has been incorporated into the formal decision-making processes of planning; the fuzziness of the concept has allowed powerful lobby groups and politicians to argue for the primacy of economic growth over the environmental and social dimensions of sustainability in planning decision-making. For some commentators, the fuzzy nature of sustainable development and its associated technocratic apparatus of sustainability appraisals and so forth turned the decision-making process into an apolitical bureaucratic negotiation dominated by the views of those with powerful commercial interests, rather than a process in which open and meaningful debate is fostered about how best to improve environmental and social conditions (Swyngedouw 2007; Allmendinger and Haughton 2012).

This brief chronological account has focused on the US and UK experiences, so it is inevitably much less helpful in understanding events in other countries. Of particular interest here is the growing critique of postcolonial planning, referring to the ways in which Western ideas have continued to hold sway in poorer nations, sometimes being transferred in inappropriate ways, such as recent attempts to masterplan eco-neighborhoods for African cities (Watson 2014). Environmental discourses in such models are used to justify a form of planning which favors those pursuing a high-growth economic model that accepts or even accentuates growing social inequalities, where environmental benefits fall mainly to the rich.

It is in this context that it is helpful to examine the Urban Environmental Transition model (McGranahan et al. 2001; Figure 2). Broadly speaking this model shows how particular environmental problems tend to dominate policy concern at particular moments in specific local contexts, involving a broad distinction between “brown” and “green” agendas in environmental policy. For poorer neighborhoods in poorer cities, the main environmental problems faced tend to be immediate, highly localized in scale, and related to environmental health – poor access to clean water and adequate sanitation for instance leading to a range of debilitating illnesses that in turn impair ability to work, providing a mutually reinforcing downward spiral of poverty and disease. Here the need is for improved collective infrastructure, in particular, clean water, drainage, and sanitation. For cities moving through an industrial phase of economic growth, environmental problems are more typically manifest at city-wide and regional level,
with rising air pollution a particular problem. For wealthier cities, where polluting industries have tended to move out or have been subject to stringently observed regulation, the problems of consumption begin to dominate political discourse, including contributions to global resource depletion and global environmental change. Local problems will not have disappeared, they will, however, be less dominant in policy framing.

The value of this model is its sensitivity to spatial and temporal scale – from the local and immediate to the global and long-term. Particularly valuable is the differentiation between the so-called brown agenda, bugs and diseases associated with poverty, malnutrition, and poor infrastructure, and the green agenda, with its concerns about the long-term sustainability of society. Like all transition models there is a danger of overstating movements and of reading off trajectories as somehow universal, unavoidable, and reflecting some desirable end point. The authors of this model are very clear that their work is not intended to be interpreted in such ways. Rather it is a broad heuristic, which recognizes that even the wealthiest cities will have areas of concentrated poverty and very immediate issues of health and wellbeing, just as poorer cities will have wealthy areas which are relatively well-provisioned and very different from the conditions in nearby poor areas.

To conclude, environmental planning has long attracted the interest of both physical and human geographers and there is a considerable degree of critical interchange of ideas between geography and planning in terms of both academic debate and engagement with practice. In recent years the growing political interest in “sustainable development” has at one level helped encourage a growing interest in environmental planning. At another level, the fuzzy concept of sustainable development could be said to have diluted the focus of planning on environmental issues, as the term sustainable development has been appropriated in part by powerful interests pursuing a “business as usual” high-growth economic model. There is no doubt that environmental planning merits and indeed needs continuing critical scrutiny from geographers (Gleeson 2012) and others to ensure that policy in this important arena is continuously and rigorously critically examined.

SEE ALSO: Environmental (in)justice; Nature; Sustainable development; Urban political ecology

References


Environmental policy

James R. Palmer
University of Oxford, UK

Environmental policy may refer to any consciously planned human activity that is designed to alter the conditions under which ostensibly natural (or, increasingly, nonhuman) landscapes, systems, and processes exist and operate. It is often but not always intended to reduce the impacts of deleterious human activities upon nonhuman forms of existence, ranging from the protection of sentient organisms from extinction through to the stewardship of nonsentient resources such as clean, safe drinking water, or a stable global climate system. Environmental policies can be conceptualized, implemented, and evaluated at a range of scales, from the local to the global. In formal terms at least, explicitly environmental policymaking can be traced, in both the United States and the United Kingdom, to 1970. Earlier policies in both countries, however, had borne characteristics that would today be recognized immediately as environmental (such as the Clean Air Act of 1956 in the United Kingdom and the Air Pollution Control Act of 1955 in the United States). In the twenty-first century, they assume a wide diversity of forms, and are undertaken not only by state governments, but also increasingly by a range of nonstate actors including local communities, civil society organizations, and supranational bodies of various kinds.

Human and environmental geographers, largely of the Anglo-American variety, have taken a diversity of approaches to engaging with environmental policy, particularly since the emergence of political economic analyses of capitalist wealth accumulation in the 1970s, and of more poststructuralist forms of analysis in the 1980s. Geographical work on environmental policy, whilst sometimes “applied,” has more often taken a critical approach to the complex processes by which knowledge is produced, policy objectives are formulated, and desirable environmental futures are identified in this field. This entry has two main sections. The first provides an historical overview of environmental policy, drawing principally from the post–World War II experience of Anglo-American polities. The second, meanwhile, offers an overview of the diverse geographical approaches to analyzing and engaging with environmental policy that have gained influence and credibility within the discipline since the 1970s. A brief final section then speculates on the future directions that geographical work on environmental policy might take, with a particular emphasis placed on the development of more indigenous geographical theory, as well as the attainment of greater policy impact.

The history of environmental policy

Environmentalism and the establishment of a “pollution control” agenda

Though today it is a taken-for-granted concept and domain of human concern, “the environment” arguably came to represent a distinct focus for political activity in Western countries only from the 1960s onwards. In the case of certain problems, such as air pollution, whilst public and private concerns were historically
well established, key “focusing” events in the postwar era were ultimately still needed to help bring remedial legislation into being. In the United Kingdom, for instance, London’s so-called Great Smog of December 1952, caused by a combination of unusual weather conditions and the widespread burning of coal for heating in homes, led to thousands of deaths, and served as the catalyst for policy designed to prevent a repeat of the situation. The resulting Clean Air Act of 1956 effectively enabled authorities to create “smoke-free zones” in the city, outlawing all emissions of black smoke and obliging local residents and factory operators to convert to smokeless fuels.

Beyond such “focusing” events, a far broader driver of environmental policy in this period was the emergence of an environmental movement, epitomized perhaps most famously in Rachel Carson’s 1962 publication of *Silent Spring*. This seminal text drew on many years of Carson’s own research to outline the negative impacts of indiscriminate pesticide spraying on the environment, and particularly on bird species. It is widely credited not just with introducing environmental concerns to much of the American population in the early 1960s, but also with bringing about a shift in American federal pesticide legislation in 1972, forcing a phaseout of the use of DDTs (dichloro-diphenyl-trichloroethanes) in agriculture. In the years immediately following *Silent Spring’s* publication, a highly combative and well-organized environmental lobby emerged on both sides of the Atlantic, embodied in the establishment of the Environmental Defense Fund in 1967 in the United States, and Friends of the Earth in 1971 in the United Kingdom.

By the end of the 1960s Western governments were coming under a great deal of pressure to respond to increasingly widespread environmental concerns. This pressure ultimately told in 1970, a year that witnessed both the bringing into effect of a new National Environmental Policy Act in the United States, and the formation of the first Department for the Environment in the United Kingdom. These new policy institutions and initiatives identified the environment as a significant additional area of responsibility for Anglo-American governments, beyond conventional state obligations. Initially however, their remit was confined almost exclusively to the reduction of gross pollution – whether of air, water, or land – by industrial contaminants of various kinds. In broad terms, a “pollution control” paradigm therefore prevailed over environmental policy in this period, with practitioners adopting a so-called end-of-pipe approach to reducing the impacts of industrial development. “Sound science” was here assumed capable of identifying the assimilative capacities of various sinks in the environment (e.g., oceans, lakes, rivers, the atmosphere), and so long as these capacities were not exceeded, industrial processes and products could be left fundamentally to continue as before. The definition and implementation of pollution standards was also characterized by close relationships between regulatory bodies and the industrial groups subject to regulation, with few apertures for environmental groups to become involved. Environmental policy institutions and initiatives of the 1970s, in short, were not wont to view environmental degradation as a structural component of wider socioeconomic practices and lifestyles.

**Ecological modernization and the rise of “integrated,” sustainable environmental policy**

Though Anglo-American environmental policies of the 1970s scarcely considered issues beyond the control of gross pollution, there was – in wider circles – much consternation...
over the more systematic problems of global population growth and natural resource use. The publication of *The Limits to Growth* by the Club of Rome in 1972, for instance, typified this overtly Malthusian trend by raising the question of whether continuous economic growth could be sustained within the confines of a single planet. In the same year, the first United Nations Conference on the Human Environment, held in Stockholm, had also generated 26 principles for governments dealing with environmental issues, many going far beyond questions of pollution control to embrace a more comprehensive view of the potential remit of environmental policy.

Though these new ideas were important and influential, the OPEC (Organization of the Petroleum Exporting Countries) oil crisis of 1973, and subsequent global economic downturn, ultimately conspired to suppress the political profile of environmental issues in many Western countries for the rest of the decade. Not until the mid-1980s can any real shift away from the paradigm of end-of-pipe pollution control be legitimately discerned. A succession of controversial environmental issues – some “transboundary” (e.g., acid rain and stratospheric ozone pollution), others rife with uncertainties (e.g., concerns over the possible impacts of nuclear power and lead in petrol) – contributed to a change in the lens through which environmental problems as a whole were interpreted. From this new perspective, the causal mechanisms leading to environmental degradation were widely recognized as complexly intertwined with existing modes of economic development (and their associated social practices). The end-of-pipe mentality of the 1960s and 1970s slowly began to give way to one of ecological modernization, under whose terms economic development and environmental protection would have to be fundamentally reconciled, rather than simply traded off against one another.

In contrast to previous, institutionally incoherent and fragmented approaches to environmental issues, ecological modernization served to promote the integration of environmental policy concerns across diverse areas of governmental activity. At the UK level, this was initially evidenced by the formation of Her Majesty’s Inspectorate of Pollution in 1987, a new body responsible for overseeing pollution control across multiple media. At the European level meanwhile, the passing of the Single European Act in 1986 saw the integration of environmental protection measures as a statutory requirement for all new policy proposals. Beyond this integrative trend moreover, ecological modernization was also promoting more anticipatory and precautionary approaches to pollution control. The idea that “pollution prevention pays” thus began to gain considerable credibility in this period, as did a sense that policymakers should consider the rights of both present and future generations (both human and nonhuman) to an unspoiled environment. Overall, ecological modernization encouraged a far more systematic approach to environmental policy, a stance reinforced in the United Kingdom with the establishment of the Environment Agency in 1995, and at the European Union level in the shift towards more preemptive modes of ex ante environmental assessment and policy appraisal from the mid-1990s onwards. Despite its apparent promise, however, ecological modernization has more recently come to be associated with the dilution and mollification of radical, environmentalist rationalities under more conservative, economic modes of thinking. Whilst its influence over environmental policy today therefore remains strong (see Market-based measures and the rise of global environmental issues, below),
the extent of its success in actually ameliorating environment policy outcomes remains contested.

Beyond its tangible impacts upon environmental policy from the 1980s onwards, one can also discern within ecological modernization the seeds of an even more powerful concept – one that would arguably become the most influential of the 1990s – *sustainable development*. With the publication in 1987 of the UN World Commission on Environment and Development’s now famous Brundtland Report, approaches to environmental policy came increasingly to be influenced by this concept, defined as “development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (WCED 1987). The subsequent formalization and promotion of sustainable development at the Earth Summit – a UN Conference on Environment and Development held in Rio de Janeiro in 1992 – marked a watershed moment for Western environmental policy. The promotion of Agenda 21, for instance, served to institute sustainability as a routine concern for policymakers operating across diverse spheres, at the local, national, and global levels. Beyond this, the Earth Summit also opened three legally binding agreements on major environmental issues for signature by nation states: the Convention on Biological Diversity, the United Nations Framework Convention on Climate Change, and the United Nations Convention to Combat Desertification. These latter initiatives epitomize two broad shifts in environmental policy that have been discernable since the 1990s. The first, typified by climate change in particular, involves the move towards assessing and addressing major environmental problems at the “global” level, often through market-based measures. The second, meanwhile, has involved the establishment of environmental protection as the responsibility of both governmental *and* nongovernmental actors of various kinds (a change that has come to be labeled widely as a shift from government to governance). The following sections now deal with each of these transformations in turn.

**Market-based measures and the rise of global environmental issues**

For the past two decades, it is undoubtedly the issue of climate change that has come to function as an “emblem” of the “environmental problematique at large” (Hajer 1995, 5). Climate change’s profile owes much to the work of the Intergovernmental Panel on Climate Change (IPCC), a body established jointly by the World Meteorological Organization and the United Nations Environment Programme in 1988. Since the publication of its first report in 1990, the IPCC’s status as the foremost source of expert scientific knowledge and evidence pertaining to the magnitude, causes, and implications of climate change has grown substantially. Its most recent Fifth Assessment Report (5AR) of 2013–2014 delivers the most unequivocal verdict yet on the extent of man’s imbrication in the processes that lead to climate change. Despite this consensus, however, policy interventions designed to mitigate climate change have enjoyed only mixed success. Whilst the 1997 Kyoto Protocol saw many advanced industrialized countries (though notably not the USA or Australia) adopt legally binding targets for the reduction of greenhouse gas (GHG) emissions, many countries have failed to meet these in practice. Moreover, whilst 195 countries signed the Paris Agreement in December 2015, agreeing on paper to reduce their GHG emissions “as soon as possible,” questions remain about the extent to which individual states’ commitments in this regard will prove legally binding. Even in the United Kingdom, where targets subscribed to under the Kyoto Protocol were...
met, this achievement can be attributed almost in its entirety to the so-called dash for gas of the 1990s, whereby a newly privatized electricity industry replaced large numbers of coal-fired power stations with cleaner gas-turbine equivalents. The passing of a UK Climate Change Act in 2008, which stipulated targets for significantly reducing carbon dioxide emissions compared to 1990 levels before 2050, is likely to pose a much tougher long-term challenge. Whether implemented nationally or negotiated internationally, GHG emissions targets however represent only one of many potential tools available for addressing climate change. Indeed, and as theorists of ecological modernization would expect, more obviously economic (or market-based) approaches to tackling climate change have also become prominent in many developed countries since the mid-2000s. So-called cap and trade carbon markets, for instance, aim to mitigate climate change by forcing companies to buy GHG emissions allowances that they can then trade with one another as desired. The most significant such market, formed in 2005, is the European Union’s Emissions Trading System (EU ETS), which covers around 45% of the European Union’s emissions today. By putting a price on carbon, these schemes effectively seek to correct a perceived “market failure” in which the costs of environmental degradation are externalized and borne by wider society, rather than retained internally, to be borne by those who instigate the degradation in the first place. The distinctly economic take on climate change that is offered by carbon markets has been reinforced elsewhere too, most notably in the United Kingdom by the Stern Review of 2006. Commissioned by the UK government’s Treasury, this influential report argued that to deal with climate change sooner would cost developed economies far less in the long run than delaying action into the future – in effect echoing the mantra that “pollution prevention pays,” another longstanding principle of ecological modernization. The pervasive infiltration of economic rationality into attempts to tackle climate change is instructive for those studying other spheres of contemporary environmental policy too. In urban environments, for instance, attempts to enhance air quality have sometimes been predicated on congestion charging, which effectively prices large numbers of motorists out of the market for road space in major cities like London, Stockholm, Singapore, and Milan. In the realm of conservation policy, meanwhile, many contemporary initiatives aim to meet targets and objectives through schemes that provide conditional “payments for ecosystem services” to the owners of environmental goods. Ecosystem services are understood to include a diverse portfolio of benefits that accrue to human populations from the functioning of natural (or nonhuman) systems, whether regulatory (e.g., maintaining soil quality and nutrient cycling), provisioning (e.g., supplying clean water or agricultural crops), or cultural in nature (e.g., sustaining unique habitats or species for future generations to experience). In 2010, the publication of a major report entitled The Economics of Ecosystems and Biodiversity (TEEB) served to reinforce this economic approach to conservation, in much the same way as the Stern Review had done for climate change four years earlier. More recently, the foundation of a new Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES) in 2012 stands as a testament both to the intuitive appeal of market-based approaches for conservation, and to the influence and credibility that has been gained by cognate, global-level organizations in other fields, most notably the IPCC.
ENVIRONMENTAL POLICY

From government to governance: public engagement and the reconfiguration of expertise

A further broad set of developments in environmental policy that can be discerned, particularly since 2000, is found in the increasingly routine engagement of wider stakeholders in decision-making processes. Whilst mechanisms designed to permit stakeholder involvement in environmental policy have a long history, historically such practices were highly constrained, limiting input to formal consultations whose remit, duration, and relationship to actual decision-making were all determined exclusively by governmental actors. Today, considerable moves have been made towards more inclusive and deliberative modes of governance that engage wider actors throughout the policy process, from initial problem definition and policy design through to implementation, monitoring, and evaluation. Such a trend arguably stems from two broad sources. First, the cumulative effects of neoliberalism – including deregulation, privatization, and budget cuts – have in many Western countries opened the way for corporations and consumers themselves to wield power over environmental policy. Second, the application of continued pressure from the environmental movement, facilitated both through traditional campaigning and through newer forms of protest and activism using social media and the Internet, has itself served to secure greater influence for nongovernmental organizations of various kinds.

One area where these broad shifts can be discerned particularly clearly in the United Kingdom is in the practice of environmental risk assessment, particularly relating to the introduction of new technologies such as genetically modified crops, or to food-related scares such as the “foot-and-mouth” and BSE (bovine spongiform encephalopathy) crises in the United Kingdom. Initial approaches to engaging wider stakeholders in risk assessment in these fields were almost always informed by an “information-deficit” model, in which public resistance to, and mistrust of, policy decisions was deemed to result exclusively from a lack of “rational” scientific knowledge and understanding. More recent initiatives have, however, moved away from this stance, focusing instead on facilitating what is frequently termed “upstream” public engagement, where the definition of who is to count as an environmental expert in the first place is kept open. Here the implicit assumption is that wider stakeholders might offer productive new perspectives and rationalities with which to approach the management of complex environmental problems, beginning from the definition of what constitutes an unacceptable risk in the first place. Consensus conferences, citizens’ juries, and other forms of deliberative experiments have therefore been deployed throughout many European countries and in the United States, in areas relating to genetically modified crops, nanotechnology, and nuclear power, amongst a host of other areas. Despite the shift in philosophy associated with these schemes, however, questions remain about the extent to which inclusive deliberation genuinely impacts, positively or otherwise, on eventual environmental policy decisions.

In the case of more systemic environmental issues such as climate change and biodiversity conservation, wider stakeholder and public engagement has arguably been manifest in even deeper ways, with responsibility for policy implementation itself increasingly being shared amongst governmental and nongovernmental actors. The UN’s Reducing Emissions from Deforestation and Forest Degradation (REDD) program, for instance, which encompasses a large number of policy measures ranging from carbon payments through to the clarification of land
tenure arrangements, is premised on the substantive involvement of a diverse set of stakeholders in implementation efforts at both the national and the global level, including forest-dependent and indigenous groups. Other examples, meanwhile, point to even more fundamental shifts in the way that the basic building blocks of environmental knowledge are constructed, with an increasingly significant role played by non-scientific actors of various kinds. The establishment of The People’s Agreement at the 2010 World People’s Conference on Climate Change and the Rights of Mother Earth, held in Bolivia, for instance, has served to counteract the globalizing thrust of the IPCC’s Assessment Reports on climate change, prioritizing instead modes of governance that are “in harmony with Mother Earth and appropriate to local cultural contexts.” The emphasis on place-specific knowledge and cultural sensitivity is not insignificant here. Indeed, as the twenty-first century unfolds, producing genuinely meaningful knowledge about many of the ostensibly global issues that have dominated the environmental agenda since the 1990s seems to be destined to require ever more place-specific, bespoke modes of inclusive, deliberative governance (Hulme 2010).

Geographical approaches to environmental policy

Geographical work on environmental issues today bears a heterogeneous array of theoretical and empirical characteristics. Indeed, the question of what makes for a quintessentially geographical interpretation of environmental policy today is remarkably difficult to answer (Castree 2004). Arguably, this is so not only because the very concept of the environment itself permits multiple epistemological (and indeed ontological) interpretations, but also because much work undertaken by geographers addressing the subject has inherited ideas, concepts, and theoretical approaches from other disciplines (including amongst others environmental sociology, science and technology studies, philosophy, and anthropology). For these reasons, any attempt to categorize major areas of geographical work addressing environmental policy will always, in some senses at least, be flawed or incomplete. The approach taken here represents just one of many possible heuristics for grasping diversity in the field today.

“Applied” environmental policy research

Applied investigations into the impacts and consequences of environmental policies in particular geographical contexts have formed a key research interest now for many years. The focus for these inquiries is diverse, ranging from local initiatives (such as those affecting waste management or household energy use) through to international-level environmental negotiations over complex, “transboundary” issues. Given its prominence as an emblematic environmental issue, another major focus is found in climate change adaptation policies, with many choosing to explore the prospects for more sustainable modes of consumption and ecological behavior patterns in the future (Pelling 2010). In line with the emergence of complex governance networks, particularly since the 1990s, the focus of this research encompasses not just state governmental actors, but also businesses, consumers, and other civil society groups.

Research in this strand does not explicitly critique the interpretive frameworks through which environmental problems are identified and addressed in wider society. Rather, it can be viewed as contributing greater levels of sophistication to conventional environmental impact assessment processes, in the sense that it supplies detailed and nuanced evidence about the
consequences of particular measures, as well as explanations of how and why these consequences vary over space. An implicit aim, therefore, is to identify more efficient, fair, or otherwise desirable means of implementing existing environmental policies in the future. In this respect the potential audience for such work goes beyond academic communities to include the very groups and actors whose attempts to intervene in society–environment relationships are being studied. Future research in this area is likely to continue to address a diverse range of environmental practices, including the “greening” of the private sector, the inducement of more environmentally friendly behavior amongst individuals and communities, and the impacts of various state interventions designed to help promote more sustainable forms of socioeconomic activity.

**Political economic approaches to environmental policy**

A more critical take on environmental policy is offered by research that adopts the tenets of political economic analysis, drawing directly from earlier Marxist and neo-Marxist work in economic and political geography. Here, environmental transformations of various kinds are seen as inextricably intertwined with power relations, social inequalities, and livelihoods. Research focuses on environmental policy and its effects in both the developed and the developing world, but also goes beyond the mere assessment of those effects to question the deeper worldviews and value judgments that underpin dominant problem definitions and preferred policy solutions in particular social contexts.

Perhaps the most significant recent strand of research in this guise explores the ways in which neoliberal market principles have been applied to contemporary environmental governance (Castree 2008). Scholars here have sought, for instance, to explore how, and with what consequences, natural resources are “produced” (or socially constructed) by capitalist societies. The transformation of previous public goods such as drinking water into private, ostensibly scarce commodities, a process labeled by Harvey (2003) as “accumulation by dispossession,” is particularly well studied in this regard. Additional research examining extractive industries (whether relating to fossil fuels or other biophysical resources) has, similarly, illuminated the inherently unequal social and economic effects of capitalist resource governance regimes. In many cases such research has also sought to document how the irrevocably material character of biophysical resources is itself consequential for the ways in which they are then appropriated, commoditized, and marketed. Since the very status of natural resources as *nonhuman* elements of the production system often precludes their full integration into capitalist production and distribution systems, researchers have been prompted to explore in depth the complex politics of resource access, use, and environmental impact in different settings.

A further and related area of work in this strand, though arguably with a longer lineage and a strong bias towards American contexts, is found in environmental justice research. This overtly political subfield explores the social processes, political interests, and institutional structures that collectively determine the distribution of negative environmental externalities associated with industrial capitalism. These externalities include toxic pollution of rivers, groundwater, or air, or else relate to the siting of potentially dangerous industrial facilities (such as nuclear power stations), usually in locations inhabited disproportionately by groups with low levels of resilience and insufficient resources to move elsewhere. Having long been associated with the environmental justice movement in
the United States, such research aims not just to document the unequal distribution of these environmental impacts, but also to advocate for measures that would serve to redress them, whether through holding polluters more firmly to account for reducing their waste outputs, or through better representing the interests and rights of vulnerable groups in society in the environmental planning process.

In addition to these streams of activity, critical research of this kind has also been undertaken in developing world contexts. Sometimes referred to as “third world political ecology,” this work seeks, at the broadest of levels, to document the interrelations of environmental policy, development, and livelihoods in non-Western societies, particularly in the rural context. A particularly fecund strand of work, dating back to the 1980s at least, has for instance explored the politics of conservation initiatives in the developing world, and particularly their consequences for local communities and land users (Adams 2009). More recent research, meanwhile, has centered on the governance of food and bioenergy production systems. Policy measures ranging from biofuel targets in the Global North, through to renegotiated land tenure arrangements and even intellectual property laws, are frequently interpreted here as “neoliberalizing” tools designed to concentrate the ownership of land, water, seeds, and other agricultural inputs into the hands of a few powerful actors. Particularly in the context of anthropogenic climate change, the consequences of these arrangements for local and regional food security in the Global South, as well as for land rights, are argued to be potentially profound.

Environmental policy as the coproduction of knowledge and social order

In a manner quite distinct from Marxist political economy, a further set of critical approaches to environmental policy serves to examine the construction, and political effects, of knowledge about the environment itself. From this perspective, knowledge is viewed not as an epiphenomenal variable in environmental decision-making, but as an independent variable of considerable consequence, particularly where the initial meaning of an issue is socially contested, or otherwise obscured through epistemic uncertainty. The fundamental starting point for all of these strands of work is thus the outright rejection of a view in which science (knowledge) and politics (power) can ever be kept separate from one another. Researchers instead regard environmental knowledge and wider policy objectives as fundamentally coproduced. Whilst the theoretical terminology used in this field is often overtly poststructuralist – key works have explored the influence of “discourse” and “framing” over environmental decision-making, for instance – it stops short of endorsing outright relativism. Facts and values may well be inseparable, but this does not mean that all environmental knowledge is fictitious. The claim made is simply that the definition and meaning of all environmental problems necessarily emerges from the complex interaction of knowledge and evidence with other more material and institutional factors, and that the terms of this interaction are influenced by wider social, cultural, and political contexts.

Much geographical work in this subfield has drawn on theoretical ideas developed in the policy sciences, science studies, and science and technology studies to explore the role of knowledge in environmental policymaking processes in the developed world. One particularly well-explored question concerns how social considerations and political interests shape the working assumptions, and thus the outputs, of environmental science, often in pervasive ways. Here work addressing the production of
knowledge about new or “risky” technologies such as genetically modified crops and nuclear power has proven influential, as has work examining the treatment of uncertainty and complexity in climate change science and global climate modeling (Demeritt 2001). Closely related work has also explored the social and political processes through which environmental expertise, as a supposedly superior category of environmental knowledge, is established in relation to particular problems. The rise of market-based measures in the environmental sphere, for instance, has witnessed a concomitant increase in the credibility of knowledge produced by environmental economists and ecological economists, particularly within policymaking circles. Expert or “policy-relevant” knowledge, for this perspective, cannot be defined in a vacuum, but gains its status in relation to wider power relations that serve to perpetuate particular sets of normative beliefs (Owens 2005).

The potential ramifications of uncertainties for environmental decision-making, both epistemic and ontological, have also served as a clear focus for research in this field. Work on risk management practices in the context of natural hazards, for instance, has documented both the intensely subjective ways in which uncertainties are interpreted, and the sociocultural factors that influence what counts as a tolerable risk for different groups in different contexts. More recent work has subsequently extended these insights to the governance of wider environmental risks, highlighting “lay expertise” (or citizen science) as a source of knowledge about complex problems that is no less valid or rational than science itself. A key contention implicit in much of this work is therefore that environmental policymaking should be institutionally recast to facilitate more participatory and inclusive deliberation, in which multiple rationalities and perspectives on problems are valued equally. Instead of privileging only scientific knowledge and evidence, diverse groups of actors should be allowed collectively to define both the initial objectives of policy, and the best means of achieving those objectives. Recent transdisciplinary and experimental research work undertaken in the context of flood risk in the United Kingdom, for instance, has outlined the potential value of precisely these types of institutional reconfigurations (Lane et al. 2011). Despite such advances, however, significant difficulties remain in determining what counts as a successful or effective mode of participatory governance, just as they do in bringing about genuinely inclusive deliberation on environmental issues at scales beyond the local.

More-than-human approaches to environmental policy

Despite providing a wealth of insights into the nature and consequences of environmental policy processes, geographical work in the previous two strands has been criticized by some for focusing too exclusively on human accounts of the environment, and for privileging human agency over that of other, nonhuman actors. An increasingly prominent strand of research that counteracts this trend has therefore emerged, particularly since 2000, with its roots to be found initially in actor-network theory (Latour 2005). At the most fundamental level, research of this kind seeks to account for the agency of the “more-than-human” world in determining the outcomes of environmental policies, with the modernist dualism of society and nature being rejected outright in favor of a new, “hybrid” ontology.

This hybrid geographical approach has significant implications for our understanding of the objects, and objectives, of environmental policy. Categories labeled by humans as “society” or
“nature,” “human” or “animal,” far from being ontological givens, become from this perspective the contingent outcome of a network of interrelations amongst diverse actors (or actants). Work exploring the governance of genetically modified crop trials, for instance, might pinpoint not just scientific research and human intentionality, but also bees, butterflies, and bacteria as independent, autonomous agents whose interrelations with one another, and with humans, have the potential to generate entirely unexpected or undesirable outcomes (Bingham 2006). By exploring specific, situated human–animal relationships, a burgeoning new literature on “animal geographies” has also sought to demonstrate the diverse ways in which nonhumans might impact profoundly on the outcomes of conservation and agroecological policies, in a range of settings. Far from imposing a fixed, human-centric view of what counts as pristine wilderness (or simply what counts as natural) upon particular animals, such work contends that conservationists would gain much from embracing the openness and contingency of nonhuman species’ “lively biogeographies” (Lorimer and Srinivasan 2013). In the urban context, for instance, an appreciation of the more-than-human agencies that enable some species to adapt to ostensibly “unnatural” city environments forces us to re-evaluate dominant conservation agendas, from both an ethical and a political perspective.

Progressing environmental geography: the need for indigenous theory and policy impact

Over a decade ago, Castree (2004) remarked of his surprise that geographers had not interrogated more deeply the implications of the term “environment,” choosing instead to direct critical scrutiny towards its close correlate – nature. Though it has often been taken for granted as a category of academic inquiry that is distinct from the economy, culture, politics, or society, in reality – and as the brief account of the geographical research agendas outlined above has made clear – environmental issues are always inextricably economic, cultural, political, and social issues as well. Today, it is arguable that only more-than-human approaches to understanding society–environment relations have made real headway in exploring the full implications of this complexity. Even here, moreover, and despite geography’s claims to be the first truly environmental discipline, the most progressive and influential contributions frequently draw on theoretical approaches derived from other parts of the social sciences and humanities.

Future geographical work addressing environmental policy certainly has a rich array of contemporary problems and issues to address, particularly in an era that is almost certain to remain dominated by global climate change. The emergence in the past 10 years of the concept of the Anthropocene – a supposedly new geological era in the Earth’s history characterized by un fettered human influences on the environment with potentially irreversible consequences – has in some senses begun to initiate a paradigm shift in the way that policymakers, scientists, and other academics conceptualize society–environment relations. This is an era ripe with opportunity for geography to establish itself as a discipline with something unique and crucial to say about changing environments, how they might best be governed, and to what sorts of equitable and sustainable ends. Contributions of uniquely geographical provenance to these debates, however, remain rare. One clear exception is found in Mike Hulme’s (2009) groundbreaking book, Why We Disagree About Climate Change, which calls for a “re-culturation” of climate policy to
ENVIRONMENTAL POLICY

take better account of the diverse, place-specific ways of knowing and living with changing climatic conditions. Whilst these ideas may take some time to impinge on dominant, broadly global approaches to understanding climate change, they should nonetheless be applauded for demonstrating the special contribution that geography as a discipline can offer in the context of this complex environmental issue.

Ultimately, whether the issues addressed by future work in geography are familiar (involving questions of environmental justice, of pollution control, or of engendering more sustainable lifestyles) or novel (centering on geoengineering, on the prevention of new biohazards resulting from climate change, or on the political implications of scientific efforts to map so-called planetary boundaries), for researchers to be content with the mere expansion of pre-existing, insular literatures on environmental policy would represent a major dereliction of their wider social responsibilities. Pushing for greater policy impact on environmental issues must become a central priority for geographical researchers in the coming years and decades. Though their work on environmental policy to date has seen much progress forged in academic circles, for geographers to make the leap towards real, tangible progress and practical influence of this kind will arguably require them to “fly the nest” offered by wider social scientific theory and develop their own, indigenous approaches to understanding and resolving the most pressing issues of the twenty-first century.

SEE ALSO: Climate policy; Ecological modernization; Environment, democracy, and public participation; Environmental discourse; Environmental governance; Environmental impact assessment; Environmental knowledges and expertise; Environmental planning; Hybirdity; Political ecology; Sustainable development

References


Further reading


Environmental racism

Laura Pulido
University of Southern California, USA

The concept of environmental racism emerged in the United States in the 1980s to refer to the socially uneven distribution of pollution and resources along racial lines. The term “environmental justice” initially referred to the movement that arose to confront environmental racism, but has since expanded to encompass multiple forms of environmental inequities and problems. The term “environmental (in)equality,” which is also used in relation to environmental racism, refers to the goal of a more equitable distribution of pollution and environmental amenities, and is frequently used by policymakers and some researchers. Many activists see the term “environmental equity” as a depoliticized concept devoid of a larger critique of structural racism, economic inequality, and environmental degradation. Accordingly, both “environmental racism” and “environmental justice” are more oppositional terms than “environmental equity.” Some have aptly described environmental justice as a marriage of environmentalism and social justice.

Origins of the term

Most scholars trace the origins of the term “environmental racism” to Warren County, North Carolina, in 1982. Trucks had been dumping soil tainted with polychlorinated biphenyl (PCB) along the side of the road in the area. Subsequently, officials decided to site a dump in Warren County. The county, which contained both African American and white residents, mobilized in opposition; however, a number of the African American leaders involved in the struggle had long histories in the Civil Rights Movement and consequently drew on its tactics, such as civil disobedience, and applied a racial frame to this particular problem, which became known as “environmental racism.” The Reverend Benjamin Chavis is often credited with having coined the term. Scholar and activist Robert Bullard (2000/1990, 98) defines environmental racism as “any policy, practice, or directive that differentially affects or disadvantages (whether intended or unintended) individuals, groups, or communities based on race or color.”

Although the struggle of Warren County was extremely important, it must be placed within its larger historical context. In fact, Warren County was just one of many communities experiencing extreme pollution problems. Other communities that attracted widespread attention at the time include Love Canal, New York, and Times Beach, Missouri. Activists from these communities began connecting and developing their own identity and discourse as the “anti-toxics movement.” Indeed, as Luke Cole and Sheila Foster argue in their book From the Ground Up: Environmental Racism and the Rise of the Environmental Justice Movement (2001), the movement arose from multiple strands, including the Black Civil Rights Movement, the labor movement, Native American struggles, the mainstream environmental movement, and the antitoxics movement. Of these various strands, environmental racism arguably became

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0453
the most prominent strand within the larger environmental justice movement.

The environmental justice movement

While many activists identified with and laid claim to the term “environmental justice,” the section of the movement focused on racism was able to gain the most traction. This was reflected, for example, in the National People of Color Environmental Leadership Summit in 1991, a research agenda focused heavily on race, Executive Order 12898, and the fact that the Environmental Protection Agency established an Office of Civil Rights. Although these initiatives did not serve only communities of color, there was no doubt that they had been inspired by charges of racism and inequality. During the 1980s and early 1990s the last embers of the Civil Rights Movement were still burning, and evidence of racial inequality was sufficient to trigger a reaction. Moreover, activists emanating from the Civil Rights Movement and other antiracist struggles built on earlier strategies that they felt had been successful, such as the Civil Rights Act and the Voting Rights Act, or at the very least were familiar. In many ways, environmental justice can be seen as one of the last struggles when the state gave at least lip service to racial equality – before neoliberalism became fully entrenched. In contrast, discriminating against communities because they are poor has never been illegal or even recognized as a problem in US society and culture.

Regardless of this tension, all environmental justice activists shared an outsider’s position, as they were not part of the mainstream environmental movement. Focusing on other priorities, the mainstream movement was slow to get involved in environmental racism. Environmental justice activists did not hesitate to critique the elitist and exclusionary nature of the mainstream movement as well as state regulators. They demanded transparency and accountability on the part of the state and polluters; challenged the privileged place of science in the policy process, especially when it overlooked or dismissed local knowledge; insisted on the need to prioritize local, everyday environments; and challenged the perceived elitism of mainstream environmentalism. This culminated in a letter sent to 10 leading environmental organizations in 1990 not only charging them with being racist, but also making clear that they did not speak for those communities in which they were not embedded, namely working class and communities of color. The idea “We speak for ourselves” emerged as a powerful theme of the environmental justice movement.

As increasing numbers of communities were impacted by pollution, they began to form networks to share knowledge and support each other. These networks were both geographically based, such as the Southern Organizing Committee for Economic and Social Justice and the Southwest Network for Environmental and Economic Justice, as well as racially based, as can be seen in the Indigenous Environmental Network, the Asian/Pacific Environmental Network, and the National Black Environmental Justice Network. In addition, several resource centers emerged around the country, such as the Environmental Justice Resource Center at Clark Atlanta University, which serves local communities battling environmental hazards. A second People of Color Environmental Summit was held in 2002, and numerous states have implemented environmental justice initiatives. Thus, there are a variety of structures and positions that characterize the movement.

Throughout the 1980s and into the 1990s the movement grew rapidly, and activists developed a philosophy and culture that distinguished it
from mainstream environmentalism. As previously mentioned, activists forcefully challenged conventional scientific practices and the right of mainstream environmentalists to speak for aggrieved communities. Another distinguishing characteristic of the movement was its rearticulation of the environment as “where we live, work, and play.” In so doing, activists crafted an environmental ethic centered on urban and poor places, in contrast to the environmental concerns of the mainstream movement, which were characterized as centered on wilderness and wildlife. While this was not necessarily an accurate characterization of the mainstream movement of the time, there is no denying that it was overwhelmingly white and paid less attention to the urban, public health, and land-use issues which have become associated with the environmental justice movement.

A key point in the development of the movement came in 1991 when the First National People of Color Summit was organized in Washington, DC. This event united hundreds of activists from across the United States to strategize and pressure the federal government to act. Subsequently, in 1994 President Clinton signed Executive Order 12898, which mandated federal programs to design features to help them achieve environmental justice. While it did signal progress, the executive order did not offer a new remedy to fight environmental injustice. Accordingly, some activists have been pressuring for Executive Order 12898 to be transformed into an actual law. The other important achievement of the summit was the establishment of the Principles of Environmental Justice, which embody the politics of the movement in the words of the activists themselves (Box 1).

As can be seen, the principles articulate a fairly radical critique of industrial capitalism. The concerns reflect the history and experiences of people of color, poor people, and native communities. However, it has been criticized as being overly anthropocentric. Indeed, within the last five years or so there have been growing criticisms of the environmental justice movement in terms of both its efficacy and its politics. While some maintain that the movement needs to be more inclusive of the middle class, others assert that it has become far too accommodating and has lost its political edge. For example, Benford (2005) argued that the environmental justice movement has exchanged its emphasis on racism for a more generalized focus on justice. This shift made the movement more inclusive, both in terms of the antitoxics movement in the United States but also globally, since a racial analysis does not apply to all vulnerable communities experiencing environmental injustice. However, the drawback, as Benford points out, is that the discourse of “justice” is much less confrontational than mobilizing against “racism,” and such a shift has contributed to the waning political vigor of the movement. Others have suggested that the movement has become far too cozy with the state, and thus increasingly conservative. Indeed, in some assessments of the environmental justice movement it has been suggested that the civil rights framework has not led to significant improvements in the environmental quality of vulnerable communities and that a new strategy and framework are necessary.

Whatever the organizing problems of the movement, however, one cannot overlook the profound shift in US politics over the past several decades. Not only have antiracist initiatives come under attack, but so too has environmental regulation. There has been a concerted effort by corporations and right-wing activists to regain control of both the state and the larger political culture in order to create an environment more favorable for capital accumulation and white supremacy.
Box 1 Principles of environmental justice.

1. Environmental justice affirms the sacredness of Mother Earth, ecological unity and the interdependence of all species, and the right to be free from ecological destruction.

2. Environmental justice demands that public policy be based on mutual respect and justice for all peoples, free from any form of discrimination or bias.

3. Environmental justice mandates the right to ethical, balanced, and responsible uses of land and renewable resources in the interest of a sustainable planet for humans and other living things.

4. Environmental justice calls for universal protection from nuclear testing, extraction, production, and disposal of toxic/hazardous wastes and poisons and nuclear testing that threaten the fundamental right to clean air, land, water, and food.

5. Environmental justice affirms the fundamental right to political, economic, cultural, and environmental self-determination of all peoples.

6. Environmental justice demands the cessation of the production of all toxins, hazardous wastes, and radioactive materials, and that all past and current producers be held strictly accountable to the people for detoxification and the containment at the point of production.

7. Environmental justice demands the right to participate as equal partners at every level of decision-making, including needs assessment, planning, implementation, enforcement, and evaluation.

8. Environmental justice affirms the right of all workers to a safe and healthy work environment without being forced to choose between an unsafe livelihood and unemployment. It also affirms the right of those who work at home to be free from environmental hazards.

9. Environmental justice protects the right of victims of environmental injustice to receive full compensation and reparations for damages as well as quality health care.


11. Environmental justice must recognize a special legal and natural relationship of Native Peoples to the US government through treaties, agreements, compacts, and covenants affirming sovereignty and self-determination.

12. Environmental justice affirms the need for urban and rural ecological policies to clean up and rebuild our cities and rural areas in balance with nature, honoring the cultural integrity of all our communities,
Environmental justice calls for the strict enforcement of principles of informed consent, and a halt to the testing of experimental reproductive and medical procedures and vaccinations on people of color.

Environmental justice opposes the destructive operations of multinational corporations.

Environmental justice opposes military occupation, repression and exploitation of lands, peoples and cultures, and other life forms.

Environmental justice calls for the education of present and future generations which emphasizes social and environmental issues, based on our experience and an appreciation of our diverse cultural perspectives.

Environmental justice requires that we, as individuals, make personal and consumer choices to consume as little of Mother Earth’s resources and to produce as little waste as possible; and make the conscious decision to challenge and reprioritize our lifestyles to ensure the health of the natural world for present and future generations.


Scholarship on environmental racism

Scholarship on environmental racism can be divided into three categories. First, there is an established body of research that examines the relationship between race and proximity to environmental quality. This is the largest literature and what most people think of when they hear the term “environmental racism.” Second, there is a slightly smaller body of work on the efforts of activists to fight environmental racism. This literature, which includes contributions by activists, scholars, and journalists, offers many case studies and analyzes the politics, values, strategies, and cultural practices of the environmental justice movement. And, last, there is a still smaller body of work that explores the relationships between race and the environment more generally. One of the distinctive features of the first two sets of scholarship is the degree to which activism and scholarship are connected.

As part of the early effort to frame the distribution of pollution as racially discriminatory, studies were commissioned that analyzed the distribution of pollution and key variables, namely, race and income. The first two of these studies was the US General Accounting Office’s (GAO) Siting of Hazardous Waste Landfills and their Correlation with Racial and Economic Status of Surrounding Communities in 1983 and the United Church of Christ’s (UCC) Toxic Wastes and Race in the United States in 1987. These were the first systematic attempts to map known pollution sources with demographic data. The UCC study examined multiple cities across the United States, whereas the GAO study focused on the southeastern United States. These studies identified uncontrolled hazardous waste sites as well as hazardous waste facilities and landfills in selected cities and states across the country, and
compared them to the distribution of minority populations and income groups. *Toxic Wastes and Race*, in particular, was extremely significant not only in bolstering the claims of the environmental justice movement, but also in giving rise to an explosion of studies on environmental racism and inequity, more generally. While there had been previous research hinting at this relationship, such as work by Brian Berry and Susan Cutter in the 1970s, it did not invoke such a framework. The UCC study essentially announced a new field of study.

Geography contributed in numerous ways to this burgeoning research field. For example, geographers working within the tradition of hazards and risks sought to understand pollution patterns within the framework of hazards, while historical geographers provided insights into how such patterns had evolved. Political geographers raised essential questions regarding the state, rights, justice, and what constitutes fairness. Urban geographers sought to understand both the distribution of pollution and environmental justice activism in terms of larger urban processes, including gentrification, segregation, and neoliberalism. Economic geographers situated the problem of environmental injustice within a political economic framework. One subfield that was slow to embrace environmental racism and justice, but that had much in common with it, was political ecology. Perhaps because the roots of political ecology were based in the Global South and tended to be rural, whereas environmental racism had emerged in the Global North and was heavily urban, the two streams were slow to find each other. But they did and have since forged a vibrant relationship. Another closely related area of inquiry is the whole field of sustainability studies. Early on, scholars such as Berman-Santana saw the link between environmental racism and efforts to build more socially and ecologically sustainable communities.

One of the richest connections to emerge between geography and environmental racism was through geographic information sciences (GIScience). As it happened, research in the burgeoning field of environmental justice, specifically environmental racism, coincided with major advances in GIScience and geographers, among others, contributed new methods and techniques to help identify and analyze the spatiality of pollution and hazards. This includes research on such things as the best unit of analysis; clustering; and reconciling changes in census tracts over time. However, it is also true that this literature is characterized by a diversity of methodological approaches, which often makes drawing any definitive conclusions difficult.

While much of this research was concerned with testing hypotheses and new methods, there is also no denying that there were political motivations. Just as some scholars hoped to challenge the claims of environmental justice activists and show that racism and/or classism was not a factor, others were equally committed to showing that they were significant variables (recall Bullard’s definition of environmental racism). This contestation was the most palpable around racism. As previously mentioned, targeting poor communities does not seem to be politically problematic in the United States, but targeting communities based on race is recognized as unacceptable. The work of sociologist Robert Bullard deserves special mention in this context. Bullard’s book, *Dumping in Dixie: Race, Class, and Environmental Quality* (2000/1990), was the first scholarly text examining the question of environmental racism in the United States. Focused in the southeastern United States, Bullard’s work analyzed various forms of pollution and employed multiple methods to explore both the demographics and politics of environmental racism. His work was generative in that, along with *Toxic Wastes and Race in the United States*,
it not only inspired many other studies but also identified a series of issues which scholars would subsequently address, including: What is the best spatial unit of analysis? What constitutes a risk? What is the process by which people of color and working-class people become disproportionately exposed? What is justice? What does an environmental justice victory look like?

As noted, one of the key themes of this early literature was an intense focus on the relationship between race and class. One reason for this debate was the way conceptions of racism and economic processes were articulated. On the one hand, neoclassical concepts dominated the literature, in which race and class could be understood as distinct and separate variables, whereas studies rooted in critical human geography were more apt to see them as social processes. Another key theme was the question “Which came first, the hazard or the people?” This question reflected the position of some scholars, who maintained that environmental racism could exist only if hazards were placed near already existing communities, in effect, targeting them. In contrast, others maintained that such intraurban migrations, in which people move toward a hazard, could be understood only in light of larger racialized processes. While all of this research enhanced our understanding of the landscape of environmental racism, scholars were also talking past each other to a certain extent. As a result, while most studies of environmental racism are empirical in nature, there are also a number of conceptual and theoretical studies that grapple with these issues. Some of the more recent theoretical innovations include applying color-blind racial ideology to environmental justice struggles, exploring the racial state, and expanding the definition of environmental racism to include acts of racial violence occurring on the land.

Significant methodological progress has been made in empirical studies since the mid-2000s. Most of the early studies relied solely on data collected by the government, despite the fact that researchers have documented weaknesses in such data. In response to these problems, research teams, such as the Los Angeles Collaborative for Environmental Health and Justice (2010), began engaging in “ground truthing” with local community organizations to get a more accurate and robust understanding of local pollution sources and hazards. In this case, they trained local residents to identify potentially hazardous and vulnerable land uses in their communities that would otherwise be overlooked. Another related weakness in the data (and regulation) was an almost nonexistent ability to account for the cumulative effects of local pollutants and toxins, as most data are simply collected and counted by individual facilities. Once again, researchers have taken the first steps toward developing models that allow us to appreciate the cumulative impacts of pollution, in some cases creating “hotspots.” Hotspots refer to concentrations of pollution due to multiple sources and kinds of pollution. Researchers have long acknowledged that hotspots exist, but have not been able to map them rigorously. Building on this trajectory, and through continued innovations in GISciences, researchers have also begun to include vulnerability in such mappings, thereby portraying a more complete portrait of the hazardous landscape. Although we are only in the first generation of such work, this is one of the frontiers of environmental racism research (Pastor, Morello-Frosch, and Sadd 2013). Developing such a screening device can enable policymakers to better protect the most vulnerable communities before they are burdened by additional environmental problems and have to wage lengthy and costly battles against them.

The second pillar of environmental racism research centers on the environmental justice
movement itself. Scholars across the social sciences have studied the origins, motivations, identities, tactics, and challenges of the environmental justice movement. Early research was initially concerned with simply documenting its emergence. Its very existence was considered noteworthy because in the 1980s environmentalism was assumed to be a strictly middle-class, quality-of-life affair and not something in which the poor and people of color participated. Although most of the scholarship on the movement is rigorous, there is some work that tends toward the celebratory. This reflects the excitement of seeing historically marginalized people challenging the state and polluters, as well as the fact that researchers are often writing for dual audiences: scholars and activists. During the 1990s identity became an important topic of inquiry within social movement scholarship, and this is reflected in the literature on the environmental justice movement. Numerous scholars explored the kinds of organizations people of color created, how they understood and articulated environmental concerns, how they organized themselves, and how they interacted with the local state and polluters. While the vast majority of environmental racism research prioritizes race and class, there has been growing attention to the role of gender, in terms of both the gendered dynamics of the movement and vulnerability. Drawing on feminist and critical race theory, some scholars have begun employing intersectional analyses.

Research on the political demands, strategies, tactics, and philosophies underlying the environmental justice movement also falls into this category. One topic that both scholars and activists have discussed at length is the breadth and fluidity of the concept of environmental justice itself. On the one hand, activists deliberately challenged and broadened the conception of the environment as articulated by the mainstream movement. This expanded notion is embodied in the idea that the environment is “where we live, work, and play.” One of the strengths of this formulation is that it challenges us to never separate the environmental sphere from the social realm, so that we are always cognizant of the multiple forces constituting a subject or place. On the other hand, however, sympathetic critics have pointed out that the term is so expansive that any kind of injustice can theoretically be termed “environmental racism.” The question then becomes: Is such a broad definition helpful? Does an effective movement require a tighter point of intersection, or is it enhanced by such breadth?

Another important topic that both activists and scholars have addressed centers on the interrogation of “justice” itself. In his book *Defining Environmental Justice: Theories, Movements, and Nature* (2007), David Schlosberg identifies three forms of justice that run through the environmental justice movement: distribution, recognition, and participation. Distribution refers to the spatiality and proximity of pollution; recognition refers to the need for acknowledgment and respect; and participation is the degree to which affected communities can contribute meaningfully to the regulatory process. His work, along with that of scholars such as Ryan Holifield, has underscored that environmental justice is about much more than just the location of pollution. Equally important are the processes leading up to such patterns and the path(s) to resolution. This focus on ethics has greatly enriched the academic literature, the policy arena, and activists’ praxis.

Yet another strand of scholarship that has developed recently is cultural studies of environmental justice. This scholarship moves beyond typical social science analysis and is more rooted in the humanities. Examples of such research include analyses of toxic tours, the role of memory in environmental struggles, and the way in
which environmental problems are represented by writers and artists. This branch of work is linked to the larger world of the environmental humanities, which is currently undergoing rapid development.

As the movement has evolved, however, there have been a growing number of critiques of both the movement and its associated scholarship. As previously mentioned, some have chided the movement for becoming less oppositional. One arena where this can be seen is in terms of both capitalism and the state. One of the initial critiques of the environmental justice movement levied by corporations and others was that activists simply did not want locally undesirable land uses (LULUs) in their backyards. Activists countered that they sought to ensure that LULUs were not put in anyone's backyard. There have been, and continue to be, impressive examples of solidarity between communities to ensure that an objectionable project was not simply relocated to another vulnerable community. This was fairly radical, as activists were essentially challenging the private decision-making nature of capitalism. Accordingly, early writers saw within environmental justice the roots of a powerful critique. Some of this work has continued with a greater emphasis on the globalized nature of the economy, by highlighting economic and material flows, the uneven integration of places across the planet, and the dominance of corporate power (Faber 2008; Pellow 2007), although this research strand is not dominant. In fact, some have argued that the movement itself has lost touch with its more radical roots and is operating primarily within the space defined by the regulatory environment. There is growing evidence that, while activists spend enormous resources seeking relief from the state, and while the state has certainly acknowledged environmental racism, it has not assisted activists in a meaningful and significant way. A recent edited volume by David Konisky (2015) indicates that the federal government has failed to address environmental racism in a meaningful way. Indeed, researchers have even found racial disparities in how the Environmental Protection Agency (EPA) protects local communities. A pivotal ruling occurred in 2001 in Alexander v. Sandoval, where the court ruled that plaintiffs could no longer bring suits against entities using federal funds based on discriminatory outcomes; plaintiffs must now prove discriminatory intent, which is extremely difficult. One alternative idea that seeks to move beyond traditional remedies is the idea of environmental reparations districts. In this scenario, local communities would have powers comparable to those of historic preservation districts to ensure that vulnerable communities are protected.

Another critique stems from the fact that the environmental justice movement has not done an adequate job of connecting with labor. Despite the fact that movement activists and workers may come from similar communities and social positions, there have been relatively few times when community activists and workers were able to unite in demanding that a polluter clean up its act, while protecting workers' jobs. In many ways the classic “economic blackmail” dynamic still operates across the environmental justice landscape. This leads to yet another criticism that is actually related to, and pertains to, class. Observers have noted the movement's hesitance to take on class issues, particularly within communities of color themselves. Such tensions may stem from conflicts between workers and residents, the class position of organizers versus the larger community, and/or economic diversity within communities. The specific intersection of racial and class dynamics presents a range of challenges that require careful negotiation.

Partly as a response to these shortcomings, as well as other challenges, such as globalization,
Pellow and Brulle (2005) have called for the development of “critical environmental justice” studies, which tackle these issues directly. Such a call suggests the maturing nature of the field. Signs that the call is being heeded are evident in some of the most recent work, which not only draws on sophisticated theoretical insights but moves beyond hagiography and does not hesitate to critique the movement itself.

The third and last area of scholarship centers on the question of race and the environment more generally. Essentially, the environmental justice movement, by focusing on race and inequality, challenged dominant narratives of human–environment relations and essentially called for a rewriting of environmental history. This is a vast project that entails rethinking people’s historical and contemporary experiences with the environment in light of inequality and power. Although the term “environmental racism” only came into being in the 1980s, it does not mean that societal inequalities were not manifest in the environmental arena earlier. Indeed, one could argue that all indigenous history, as well as the history of slavery, for example, can be seen as environmental history (see, e.g., Carney 2002). While social inequality never characterized environmental studies in general, some scholars have acknowledged such social relations – but they lacked a systematic framework and scholarly context for discussing them. Increasingly, scholars are going back and reinterpretting such fields as environmental and urban history (Taylor 2009), in light of new understandings of the connection between race, inequality, and the environment. This project is still in its infancy.

One example of this kind of work is by sociologist Devon Peña, who has studied Chicana/o agricultural communities in southern Colorado. His focus on land encroachment, the commons, acequias (communally maintained water systems), and heirloom seeds has comprised an important stream of research that has not only challenged traditional assumptions about the nature of family farms, but also insisted that the environment be taken seriously by Chicana/o studies (Peña 1998). While Peña does not hesitate to show the structural challenges these farmers encounter, his research is extremely important in that not only does he discuss a community of color far beyond the realm of urban toxics, but he indicates the breadth and depth of environmental issues for all communities. Carolyn Finney (2014) has also pursued a distinctive path with her work on African Americans and the National Park Service. She has explored not only how African Americans have historically experienced parks, but also how they are represented within the larger context of wilderness. Like Peña, she extends her analysis back hundreds of years and compels us to reconsider what we think we know about social forms of domination and the environment.

One of the findings to emerge from this scholarship is that important differences exist between the relationship of various communities of color with the environment. While much of the environmental justice literature refers simply to “people of color,” this term overlooks important distinctions. For instance, while indigenous people are always included as a group facing environmental racism and have been extremely active within the environmental justice movement, they have not always embraced the term “environmental racism.” Some argue that native peoples are distinct from racial minority groups in that they are nations, and the concept of racism is not adequate to the task of describing or analyzing their situation. Accordingly, some have applied the concept of settler colonialism to better understand the environmental hardships and struggles that native peoples have experienced. Settler colonialism
not only elucidates the multiple injustices and power relations embedded, for example, in the proposed Tar Sands pipeline (Preston 2013), but also draws attention to the fact that the land on which most US national parks and wilderness sit once belonged to native peoples who were forcibly removed in order to make room for new owners and uses.

Asian/Pacific Islanders also have a very distinct relationship to the environmental justice movement. Almost all studies of environmental racism, that is, disparities, center on African Americans, Latinas/os, and indigenous peoples. Only rarely are Asian/Pacific Islanders included, and, consequently, there has been little evidence that they are impacted in a comparable way, primarily because of spatial and economic patterns; however, Julie Sze (2011) has argued that even though Asian/Pacific Islanders don’t register significantly in terms of distributitional justice, the story changes if we switch the lens and consider procedural justice, in which Asian and Pacific Islanders, especially immigrants, are routinely overlooked and excluded. Sze also reminds us that Asian/Pacific Islanders have been an essential part of the environmental justice movement, as individual activists, scholars, and policymakers, through the Asian Pacific Environmental Network (APEN), and in multiracial formations.

In contrast, the environmental disparities and challenges that both African Americans and Latinas/os face have been more thoroughly documented. As large, primarily urban groups, both populations have been studied in terms of their geographic proximity to pollution. Because these two groups are highly urbanized, disproportionately low-income, and residentially segregated, they register far more frequently as impacted communities. In addition, given their concentration as low-wage workers performing some of the most dangerous work, there has also been some attention given to the workplace hazards they encounter, especially Latinas/os.

A key event that enabled many people to see the deep and pervasive nature of environmental racism was Hurricane Katrina. When Katrina struck New Orleans in 2005, it became apparent to millions of people exactly how poverty, residential segregation, and racism operated to create a disaster resulting in the disproportionate deaths of African American residents, and also how racism has shaped the rebuilding of the city. Whereas previously many were willing to see racism as a hostile, discriminatory act, Katrina showed how systematic neglect and disinvestment could produce landscapes of environmental racism. Many scholars and activists took the opportunity to study Katrina and to use it as a symbol of environmental racism, with a particular focus on the structural disadvantage that characterizes much of African American life.

While there is no denying the tremendous impact that the environmental justice movement has had in producing a new paradigm centered on inequality and justice that has resonated far beyond narrow conceptions of the environment (e.g., climate justice, food justice, transit justice, energy justice, etc.), the future of environmental racism as a concept is uncertain. As previously discussed, given the political shift toward the right in the United States, it has become increasingly difficult to make claims based on race in the courts, and large swaths of environmental law and regulation have been seriously weakened. In response to this situation, some have suggested that a human rights framework may be more effective and that activists need to operate more in the international arena. Indeed, globalization presents both opportunities and challenges to environmental justice activists. On the one hand, the concept of environmental justice, broadly defined, has found a global audience and the term is frequently used to denote instances when
poor, vulnerable, and nonwhite communities are being exploited, killed, displaced, and polluted. For instance, the struggle of the Ogoni people of Nigeria against oil producers is generally recognized as an environmental justice struggle, as is the Union Carbide disaster in Bhopal, India, and its aftermath. While the injustice is obvious, these examples are not typically conceptualized as racial. Thus, there is a clear need for activists and scholars to work on reconceptualizing the key terms and structures of the movement. What is the relationship between environmental racism and environmental justice, given that the latter has expanded significantly? How do these various pieces function together? While US activists generated the concepts of environmental racism and environmental justice and offered them to the world, they have not been as open to learning and drawing from global experiences, despite the fact that environmental struggles on the part of the poor and dispossessed are rampant today. In many ways, the concept of environmental racism remains very much rooted in the US experience. Although there are many opportunities to open up a dialogue and to move toward a more global movement, it remains to be seen where the environment justice movement will go.

SEE ALSO: Environmental hazards; Environmental (in)justice; Environmentalism; Environmentalism, grassroots; Race and racism; Social movements

References


Environmental regulation

Richard Perkins
London School of Economics and Political Science, UK

Defining environmental regulation

Environmental regulation refers to the means, mechanisms, and actions used to achieve behavioral control and influence over target groups in the field of environmental protection. Traditionally, environmental regulation has been interpreted to mean forms of regulatory authority promulgated by state actors through detailed, prescriptive, and legally binding rules. Yet this narrow perspective has increasingly given way to a much broader conception of environmental regulation which encompasses a wider range of regulatory actors, approaches, and forms of control (Gunningham 2009; Fiorino 2006). This expanded interpretation of environmental regulation reflects a more sophisticated scholarly understanding regarding the “who” and “how” of environmental regulatory influence – that is, the actors who promulgate regulatory control and the instruments through which they do so. Moreover, it reflects ongoing shifts in an underlying reality, most commonly associated with a growing role for nonstate actors in environmental regulation, and the adoption of so-called smart regulatory strategies by state ones.

Typologies of environmental regulation

Environmental regulation has been categorized in a number of ways. One common way is according to the actors and sources of authority. A distinction is therefore frequently drawn between state and nonstate forms of regulation – also framed as between public and private regulation. As its name suggests, public environmental regulation involves the promulgation of regulatory control, influence, and actions by governmental actors. Along similar lines, private regulation involves nonstate authority, wherein environmental standards, expectations, and activities are developed, implemented, and enforced by private actors. Private/nonstate actors, in turn, can be further subdivided into civil and market actors. Civil actors comprise individuals, community groups, and nongovernmental organizations (NGOs), while market actors include firms and business associations. Regulatory influence and control exercised by these three groups are referred to as government (state), civil, and market (self-) regulation, respectively.

More recently, attention has focused on so-called hybrid and co-regulatory forms of regulation. Underpinning this interest is the observation that environmental regulation may also be achieved by combinations (or “constellations”) of different actors who interact, cooperate and collaborate to realize collective goals (Steurer 2013). Three main co-regulatory arrangements exist: (i) public co-regulation between state and market actors, also variously
labeled government–business and public–private partnerships; (ii) public co-management involving state and civil actors; and (iii) private co-regulation, also known as private–social and business–NGO partnerships, involving joint action by market and civil actors.

A second way in which environmental regulation has been classified is in terms of regulatory instruments (Wurzel, Zito, and Jordan 2013). Instruments can be understood as the tools or techniques through which regulatory control is exercised and policy objectives are achieved. Although a number of competing typologies exist, four broad categories of instruments are commonly identified: (i) directive-based standards; (ii) market-based instruments; (iii) voluntary approaches; and (iv) information-based approaches. The last three categories are described as new environmental policy instruments (NEPIs).

Directive-based standards attempt to achieve regulatory goals through the promulgation of technical rules (or standards) by the state, backed up by various administrative or legal penalties to incentivize compliance. Also known as command and control (C&C) instruments, they comprise a command (i.e., what a particular regulatory target must achieve) and a control (i.e., a sanction for noncompliance). The actual command can take a wide range of forms from ambient and emission standards, technology-based standards, through to product- and process-based standards. In practice, C&C can be anticipatory in nature, requiring regulatory targets to achieve certain standards prior to the commencement of an activity (e.g., through permitting processes), as well as continuing to control activities once they are operational. To take one example, in order to receive clearance for sale in domestic markets, many countries require new motor vehicle types to meet prescribed tailpipe emission standards, measured in a test facility over a specific driving cycle (i.e., the anticipatory control). Once on the road, individual vehicles are invariably subject to further rules, obliging owners to obtain a test certificate on a periodic basis to demonstrate conformity with in-use emission standards (i.e., the continuing control).

Directive-based standards provide a transparent, familiar framework of environmental regulation which, assuming compliance by regulated parties, offers timeliness and certainty of outcomes. This makes C&C particularly well suited to regulating substances where an immediate, outright ban is required, such as in the case of highly toxic chemicals with potentially serious environmental or human health effects. However, a frequent criticism of standards is that they are overly detailed, inflexible, and ultimately expensive. All regulated parties are required to comply with the same standard, irrespective of their abatement costs or availability of alternative abatement options, meaning that C&C may not achieve environmental goals at the least overall cost for society. Directive standards have also been criticized for failing to offer incentives to abate beyond prescribed requirements and for discouraging ongoing innovation. Additionally, critics have pointed to the high monitoring and enforcement requirements of C&C standards (particularly in the case of multiple, distributed sources of degradation), and their potential to create adversarial relationships between government regulators and regulated parties.

Market-based instruments (MBIs) are a second set of regulatory instruments. MBIs use financial incentives (and disincentives) – as opposed to prescriptive standards – to influence environmentally relevant behavior. There are two main types. The first are so-called price-based instruments which can be further broken down into eco-taxes and subsidies. Eco-taxes seek to reduce environmental “bads” by internalizing the costs of environmental damage into market
users are required to pay a tax for the use of the environment (e.g., as a resource in production or as a sink for pollution) – thereby making environmentally damaging activities more expensive. Categories of eco-taxes include user charges (e.g., taxes on waste sent to landfill), emission charges (e.g., on NOx emissions), sales taxes (e.g., on pesticides used in agriculture), and deposit refund schemes (e.g., for drinks containers). The aim of subsidies is to increase the supply of environmental “goods” by paying actors for environmentally beneficial choices, activities, and investments. Residential feed-in tariffs, whereby households are paid a preferential rate for supplying electricity from small-scale renewables to the grid, are one contemporary example of an environmental subsidy.

Quantity-based instruments are a second type of MBI which, in the realm of environmental regulation, mainly comprise tradable permit (TP) schemes. These schemes involve the creation of a right – either to consume a particular resource (e.g., water) or to emit a pollutant (e.g., SO2) into the environment – which is subsequently traded within an administered market. Two subcategories of TPs exist: cap and trade and baseline and credit. Under cap and trade, a specific quantity of rights (also known as permits) equal to the overall cap are allocated to regulated parties through one of a number of possible mechanisms, for example given away for free based on historic usage/emissions or auctioned off. Individual parties are prohibited from using more resources/emitting more pollutants than the quantity of rights they currently hold, with penalties for noncompliance. However, rights can be bought and sold by parties, constituting the trade. Regulators can increase the stringency of environmental requirements by lowering the overall cap (e.g., on the total amount of pollutant releases permitted in a particular year) over time. Established cap and trade schemes include Australia’s Hunter River Salinity Trading Scheme (regulating saline water discharges), the European Union’s (EU) Emissions Trading System (regulating greenhouse gas emissions), and the United States’ Acid Rain Program (regulating SO2 emissions). In baseline and credit schemes, regulated parties are assigned a baseline, defining the maximum amount of emissions/resource usage permitted during a particular period. Actors who emit/consume less than this baseline can generate reduction credits which can be sold to actors who exceed their baseline.

According to their proponents, an important advantage of MBIs lies in their flexibility, in that they allow regulated parties to decide the actions they wish to take in order to contribute towards environmental goals. For example, a polluter faced with a carbon emissions tax could either pay the tax or chose to reduce emissions through one of a number of abatement channels (e.g., switching to a less carbon-intensive fuel or investing in more carbon-efficient process equipment), selecting whichever option is cheapest. Precisely because of this flexibility, and the ability of the “invisible hand” of the market to distribute environmental effort towards actors who can undertake environmentally beneficial actions most efficiently, it is suggested that MBIs are cost-minimizing for society as a whole. MBIs also create ongoing incentives for low-cost abatement and for cost- and environment-efficiency-enhancing technological innovation. Critics have countered this positive assessment. Amongst others, they have pointed to the difficulties of effectively administering, monitoring and enforcing MBIs; the possibility of market manipulation by more economically powerful parties; and problems of environmental “hot spots” inadvertently emerging where polluters with high abatement costs are concentrated in the same area.
ENVIRONMENTAL REGULATION

A third category of instrument, voluntary approaches (VAs), brackets a diversity of different regulatory techniques. What unites VAs, however, is the fact that they are not mandated by law, with participants voluntarily engaging in measures aimed at protecting the environment. Several different types of VA are identified in the literature: unilateral initiatives, codes of conduct, public voluntary challenges, and negotiated agreements. Unilateral initiatives are voluntary environmental actions taken by individual firms. A case in point is 3M’s Pollution Prevention Pays (3P) program comprising an “in-house” corporate initiative aimed at supporting projects which save money through reduced pollution at source and improved energy/resource efficiency. Codes of conduct, on the other hand, are institutionalized forms of collective action orchestrated by industry associations, NGOs, and standard-setting organizations (Eden and Bear 2010). Adopters commit themselves to a set of principles, policies, and practices governing their environmentally relevant behavior. Certification schemes are a particular type of code, where participants meeting predefined process and/or product-based requirements gain recognition in terms of an eco-label or publicly recognized environmental standard. Cashore (2002) has famously invoked the concept of nonstate market-driven (NSMD) governance systems to describe such schemes in which the institution of the market provides the necessary incentives for participation and compliance. The Forest Stewardship Council (which certifies wood sourced from sustainably managed forests) and ISO14001 (a standard for certifiable environmental management systems) are well-documented examples of NSMD schemes.

Two other types of VAs involve a more explicit role for the state. Under public voluntary schemes (PVSs), the government defines a set of voluntary standards or requirements, which participating firms pledge to follow. In return for compliance, private parties may receive public recognition in the form of a state-sponsored eco-label, or else be rewarded with other inducements (e.g., regulatory “relief” from reporting requirements or public subsidies). An example of a PVS is the Costa Rican Tourism Board’s Certification for Sustainable Tourism, which provides participating tourism businesses with a certified evaluation of their beyond-compliance sustainability-related practices and performance. A fourth type of voluntary approach is that of negotiated agreements (NAs), which are bilateral agreements struck between governments and private actors. The precise nature of particular agreements and their participants vary. However, they typically involve some form of environmental target (e.g., for pollution reduction), together with a timescale over which either individual firms or entire industry sectors agree to meet the target. The Netherlands has made extensive use of NAs, with the Dutch environmental ministry agreeing a series of covenants with different industry sectors and their organizational representatives as a key vehicle to realize public environmental goals.

Much like MBIs, the strengths of VAs are assumed to lie in their flexibility and creativity. VAs harness the self-regulatory and innovative capacities of private actors to foster cost-effective regulatory actions which go beyond the “one size fits all” of directive standards. Moreover, they appeal to the logic of self-interest, providing participants with positive market incentives (e.g., such as differentiation advantage for their products), and reducing the threat of potentially costly, state-imposed regulation. By relying on voluntarism, rather than on the legal sanction of government regulators, proponents have also suggested that VAs engender a sense of positive (shared) environmental responsibility which sustains ongoing commitment. Voluntary
ENVIRONMENTAL REGULATION

approaches have nevertheless been heavily criticized, first and foremost on the grounds of their limited substantive effectiveness. A common assertion is that many VAs are vehicles for corporate “greenwash.” That is, voluntary regulations allow firms to create the impression that they are contributing to environmental goals, without necessarily bringing about significant and corresponding improvements in underlying environmental performance. The empirical literature provides some support for this claim – at least for weaker codes or PVSs with limited substantive performance requirements and/or lacking monitoring and sanctioning mechanisms (Borck and Coglianese 2009).

A final set of instruments are premised on influencing behavior through the provision of information to various external parties (Mol 2008). The underlying logic is that actors lack sufficient information – about environmental quality, the environmental performance of firms and their products, or even their own environmental impacts. Information-based regulation addresses informational failures and, in doing so, empowers various stakeholders to take regulatory actions. As citizens, for example, communities can use information to put pressure on polluters, politicians, and/or environmental agencies. As consumers, individuals can use information to guide their purchasing decisions, providing firms in relevant product markets with an incentive to improve their performance. As investors, actors can penalize poorly performing firms (e.g., by selling their stock), and reward those with superior environmental performance (e.g., by lowering the cost of capital).

One of the most widely documented forms of informational regulation is public disclosure, of which there are two main types. Pollutant Release and Transfer Registers (PRTRs) make data available to the public on the volume of specific pollutants emitted and/or transferred off-site by particular firms. Two prominent cases are the United States’ Toxic Release Inventory and the EU’s European Pollutant Release and Transfer Register (formerly European Pollutant Emission Register). Performance evaluation and ratings programmes (PERPs) go one step further, in that regulators use data to produce publicly available environmental ratings of individual firms. Indonesia’s Program for Pollution Control Evaluation and Rating (PROPER) falls into this category, with public regulators assigning a graded color code to firms corresponding to their level of environmental performance. Disclosure-based initiatives have also increasingly been enacted by private actors, including business groups and NGOs, in areas such as climate change. A high-profile example of the latter is the CDP (formerly the Carbon Disclosure Project), a nongovernmental organization sponsored by major institutional investors, which collects and reports information on the greenhouse gas performances, policies, and practices of thousands of private companies across the globe.

Eco-labels constitute another type of information-based regulation. Eco-labels can be voluntary, yet labeling can be mandated by governments through regulations requiring producers to provide environmental information on their products to relevant stakeholders. A growing number of countries therefore require manufacturers of various electrical goods to provide standardized energy labels for their products. Governments can also empower citizens through freedom-of-information legislation, requiring public bodies to disclose environmental information that they hold in response to citizen requests.

Themes in environmental regulation

Underpinning much of the recent work on environmental regulation is the idea that it is undergoing significant change, characterized
ENIRONMENTAL REGULATION

by three ongoing dynamics: (i) the reform of public environmental regulation; (ii) a growing role for nonstate actors; and (iii) the rescaling of environmental regulation. The rest of this section outlines these dynamics.

Reform of public environmental regulation

A first key theme in the literature is the reform of public environmental regulation and the shift from a first to a second generation of regulation (Fiorino 2006). First-generation regulation, which came to the fore in countries such as Germany, the United States, and Japan during the 1960s and 1970s, is commonly described according to several distinguishing characteristics. One is that it was predominantly concerned with controlling local, visible, and point sources of environmental degradation (e.g., air pollution from power plants). Second, it was predicated on centralized, hierarchical regulatory power, exercised by state institutions. Hence first-generation regulation was largely achieved through media-specific C&C instruments, implemented and enforced by newly created environmental regulatory agencies. Third, it was technocratic, underpinned by a belief in expert science, and the idea that environmental problems could be solved through modern technological solutions. Many of these solutions involved end-of-pipe technologies (e.g., catalytic converters fitted to vehicle exhausts), which capture pollutants after they have been produced, rather than prevent their generation in the first place.

The reality of first-generation environmental regulation was – perhaps unsurprisingly – more complex and geographically variable than these stylized descriptions suggest. To take one example, important differences have been documented in the way that C&C was implemented by regulatory agencies across countries. An early contrast was made between the US approach, described as rigid, sanction-based, and adversarial, and the UK one, characterized as flexible, compliance-based, and cooperative (Gouldson 2004). Typically, such variations have been explained in terms of so-called policy or regulatory styles, indicative of underlying institutional traditions. Significant differences have also been noted in the degree to which public regulators made use of instruments other than C&C.

Notwithstanding these differences, observers have pointed to a shift towards a second generation of environmental regulation in many developed countries. This second generation seeks to address a new set of environmental challenges. Amongst others, these include forms of environmental degradation which are more diffuse in origin (e.g., agricultural runoff); lack the visibility of many first-generation pollutants (e.g., volatile organic compounds); are characterized by scientific uncertainty, ambiguity, and contestation (e.g., genetically modified organisms); and are closely bound up with the daily lifestyle choices of citizens (e.g., CO₂ emissions).

Another distinguishing characteristic of second-generation regulation is that it has attempted to move away from a predominant reliance on technical rules, legal coercion, and state monitoring systems. Instead, the emphasis has been on the development of more cooperative, less prescriptive forms of regulation which leverage the self- (and co-) regulatory capacities of nonstate actors. To this end, regulatory agencies have innovated, experimented, and applied a range of more flexible, collaborative forms of NEPI. The result is a policy instrument mix that includes not only C&C, but also a variety of MBIs, VAs, and informational approaches. The strategic use and orchestration (by government regulators) of different combinations of regulatory instruments which additionally
harness the regulatory capacities of nonstate actors is often described as smart regulation (Gunningham 2009).

A further characteristic of second-generation regulation is its emphasis on prevention. Inspiration for this preventive logic has come, in part, from the precautionary principle which has emerged as a key principle guiding environmental regulatory decision-making over recent decades. However, it also reflects growing recognition that first-generation C&C approaches did not always solve problems, but often simply transferred them across media. One practical example of this preventive thinking is the growth of integrated permitting whereby regulators issue a single permit for an entire plant, considering together all air, water, and waste issues so as to minimize overall environmental harm.

The shift from first- to second-generation regulation has formed part of a wider effort to reform – or literally reinvent – the way in which governments attempt to regulate environmental issues. Pressure to reform has been fueled by growing resentment towards C&C, both by regulated parties faced with escalating compliance costs, and by regulators themselves struggling to enforce rules. Yet, beyond these practical considerations, regulatory reform has been profoundly shaped by ideological considerations associated with the rise of neoliberalism in the 1980s and 1990s. The influence of neoliberal principles in environmental regulation – often framed as market environmentalism or neoliberal environmental governance – has been widely discussed in geography (Heynen et al. 2007). Crucial to this work is the notion that markets, market principles, and market actors are increasingly being deployed in natural resource and environmental governance. To take one example, TP schemes are predicated on processes of commodification whereby an environmental good or service (say, CO₂ or H₂O) is converted into a tradable commodity which can be exchanged for profit in a market setting.

Whilst public regulation has evolved, substantial variations have been observed in the way that second-generation environmental regulation has unfolded across space. Much of the comparative work in this area has focused on the EU, with studies revealing significant differences in the adoption of NEPIs across member states over the past two decades (Wurzel, Zito, and Jordan 2013). National contextual factors, which influence the costs, benefits, and feasibility of different choices, have figured significantly in explanations of these cross-national differences. Another related concern voiced by scholars has been the degree to which NEPIs have supplanted more conventional, directive-based approaches. A recurrent conclusion is that C&C remains a central component of the instrument mix in the majority of developed countries, although the way it is configured and implemented has become more flexible and responsive to regulatory realities (Fiorino 2006; Gouldson 2004).

Approaches by governments to environmental regulation have also been evolving in countries outside of the developed world. The literature in this regard largely examines the experience of emerging economies which, following in the footsteps of developed countries, have experimented with a range of NEPIs (Caffera 2011; Blackman 2008). A number of factors have been implicated in this trend, including: external support for new regulatory approaches from international organizations such as the World Bank; lesson-drawing and inspiration from, and emulation of, past adoptions in other countries; and the search for regulatory alternatives against a backdrop where the effectiveness of C&C has sometimes been undermined by low levels of enforcement and compliance.

A common observation – across developed, emerging, and developing countries – is that
new regulatory instruments are far from a panacea to the shortcomings of more traditional, directive-based approaches. In fact, despite their appeal (in terms of flexibility, cost-efficiency, etc.), the actual performance of NEPIs is heavily influenced by two fundamentals. One is the design of individual NEPIs in practice. There are multiple ways in which instruments such as TP schemes, codes of conduct, and disclosure schemes can be set up, configured, and put into operation. Evidence suggests that the actual effectiveness of NEPIs crucially depends on features such as the stringency, specificity, and bindingness of performance obligations together with the quality of monitoring and enforcement mechanisms (Borck and Coglianese 2009; Fiorino 2006).

Another fundamental which is known to significantly impact the performance of NEPIs is the capacity of actors to perform various regulatory functions. This includes administrative capacity, which is necessary in order for public regulators or private actors to devise appropriate regulatory schemes, effectively oversee their implementation, resolve any problems, and ensure compliance. Yet it also includes the capacity of actors to meaningfully and effectively participate in environmental regulatory schemes, systems, and programs. To take one example, the ability of civil stakeholders to act as environmentally effective surrogate regulators in the case of public disclosure schemes is known to depend on the existence of civil capacities, such as the technical knowledge and skills needed to effectively process disclosed information, as well as the resources, freedoms, and channels available to civil actors to exert regulatory pressure on firms (Mol 2008). The importance of capacity goes some way to explain why the performance of NEPIs has often fallen short of expectations in various emerging economies where they have been implemented (Caffera 2011).

The growing role of nonstate actors

A second major theme has been the growing regulatory role of nonstate actors. Within this context, an especially provocative thesis to emerge is that a transition is underway from an era of “environmental governance by government” towards one of “governance without government,” as private actors increasingly engage in the setting and implementation of regulation. The regulatory function of nonstate actors has been widely discussed in relation to market actors. Their enhanced role reflects the fact that firms are increasingly being enrolled by government regulators as part of ongoing regulatory reforms. However, market actors are also actively becoming regulators in their own right, orchestrating and participating in a wide variety of private voluntary initiatives.

A significant body of work, much of it couched within a framework of corporate environmentalism and corporate social responsibility (CSR), has documented the drivers underlying firms’ growing self-regulatory activities (Blowfield and Murray 2014). One is heightened stakeholder expectations regarding firms’ environmental performance and the fact that they can no longer simply rely on compliance with government regulation for social legitimacy. Another strand of literature emphasizes market benefits, focusing on how beyond-compliance strategies and investments can help firms to reduce their costs, increase their revenues, and mitigate business risks (Meckling 2011). Parallel to these developments, the growth of market self-regulation has frequently been associated with globalization, with voluntary schemes pursued by firms as a means to manage their cross-border environmental impacts and safeguard their reputations.

Civil actors and NGOs, in particular, have additionally played an increasingly important regulatory role over recent decades. They have been instrumental in defining and articulating
informal standards and norms of environmental responsibility. NGOs have performed a monitoring function, collecting information on governments’ and firms’ compliance with treaty obligations and public regulations, as well as the performance of firms in relation to their own commitments and societal norms of corporate responsibility. Allied to this role, they have acted as enforcement agents, using a range of tactics to bring regulatory targets into compliance with formal or informal standards. NGOs have assumed the role of private policymakers, participating in efforts to formulate, promulgate, and enforce voluntary initiatives, both by themselves and in conjunction with market actors (Eden and Bear 2010).

Recent work has also explored how individuals are being engaged in self-regulatory activities. One stream of literature has documented how governments have encouraged individual citizens to contribute towards public environmental goals through behavioral change. Amongst the domains of “pro-environmental behavior” (PEB) that have been discussed as objects of public policies are recycling, modes of travel, household energy use, and the purchase of environmentally friendly goods and services. Work has also examined the rise of ethical consumption and how individual citizens are demonstrating their normative commitment through consumption practices.

The growing involvement of private actors has been met with a mixed response. On the one hand, some have welcomed their expanded role as regulatory actors, particularly in areas where traditional forms of state-centric regulation have proved ineffective. Nonstate actors bring new resources, incentives, and regulatory architectures to the pursuit of collective environmental goals. Yet many others have been more critical. A recurrent focus of this critique has been on civil regulation and its representativeness, legitimacy, and accountability. Certain environmental NGOs have been accused of lacking a democratic mandate, articulating the interests of narrow publics through their regulatory actions, whilst proving inattentive to others. Questions have also been raised about the transparency and inclusivity of the (“closed door”) processes through which private codes of conduct have been made by NGOs and market actors.

Another oft-made charge is that, rather than a progressive response to evolving environmental demands, the growing involvement of market actors is a covert effort to manipulate those demands. The literature has therefore documented how certain firms (sometimes in collaboration with NGOs, as part of so-called bootlegger-and-Baptist coalitions) have lobbied for public environmental regulations as a means to disadvantage their competitors who face higher compliance costs. Critics have further suggested that market actors have purposely sought to narrowly enframe and shape environmental regulatory solutions in ways which do not threaten their interests and which prevent the emergence of more binding, stringent forms of regulation. As an example, advocacy by market actors has been implicated in the emergence of TP schemes in carbon. Carbon trading has allowed major polluters to accommodate growing demands for climate action, but to do so in a way which permits ongoing capital accumulation, and avoids costly carbon taxes (Meckling 2011).

Rescaling environmental regulation

A third important theme in environmental regulation has been rescaling. A core idea of this literature is that the traditional model of territorial environmental regulation is being reworked and reconfigured in multiple ways (Reed and Bruyneel 2010). More specifically, it is suggested
that regulatory authority no longer exclusively resides in a centralized state, which exercises hierarchical power vertically. Instead, authority in environmental regulation is increasingly being dispersed across a wider range of spatial scales, as well as to different actors within and between these scales. Rescaling is a consequence of a purposeful strategy of national governments to cede sovereignty – upwards to supranational bodies and downwards to governmental actors at the local level – in order to more effectively achieve their policy goals. Yet it also reflects attempts by actors to challenge the monopoly of the nation-state, to expand their regulatory competence, and to address real or assumed governance gaps in the area of the environment.

Earlier work tended to focus predominantly on the vertical aspects of rescaling. Scholars of global environmental politics have therefore examined the rescaling of regulatory authority in the intergovernmental realm. Particular attention has been paid to interstate cooperation, and the creation of international environmental agreements (IEAs), regulating issues ranging from whaling and hazardous waste through to climate change. A considerable body of work also exists on the upward rescaling of authority to supranational governmental bodies. By far the most significant of such bodies is the EU, which has emerged as an important regulatory actor over recent decades. The literature on vertical rescaling has additionally focused on the decentralization of regulatory authority to subnational governmental and non-governmental actors. Community-based natural resource management initiatives represent one example of vertical rescaling, realized through a partial transfer of regulatory authority from the state to local-level civil actors.

The past decade has further witnessed a flourishing of work on the horizontal rescaling of regulatory authority. Much of the focus has been on transnational (governance) networks which straddle national borders and, directly or indirectly, perform regulatory functions. By definition, these networks go beyond those exclusively constituted by national governments in the intergovernmental realm, in that they involve a prominent role for nonstate and/or subnational governmental actors. Within the private sphere, geographers have shown particular interest in transnational certification networks, through which market actors and/or NGOs seek to regulate environmental and social aspects of production processes and/or products across national borders (Eden and Bear 2010). Work on horizontal rescaling involving governmental actors has, amongst others, focused on transnational municipal networks comprising local authorities and municipalities (Bulkeley 2005).

**Continuity amidst change**

Although much of the recent literature on environmental regulation emphasizes change, caution should nevertheless be exercised not to overstate these dynamics. Certainly, new regulatory instruments are being deployed more widely, nonstate actors are assuming an expanded role, and regulatory authority is being spatially diffused. Yet the overall story is just as much about continuity as it is about change. Hence directive-based standards continue to play a prominent role within public environmental regulatory systems and, in the majority of countries, the state remains the single most important environmental regulatory actor. In fact, despite claims of environmental governance without government, a great many NEPIs are reliant on governments defining and enforcing rules, and even largely private forms of regulation can often be seen as operating “in the shadow of hierarchy” (Gunningham 2009). Recent dynamics are therefore best understood as a reconfiguration of state regulatory authority, with governments...
Performing regulatory functions in conjunction with a growing range of actors, and furthermore at and between a range of scales.

See Also: Environment and the state; Environmental certification and eco-labeling; Environmental governance; Environmental policy; Neoliberalism and the environment

References


**Environmental restoration**

**Sally Eden**  
*University of Hull, UK*

“Environmental restoration” and “ecological restoration” refer to a range of methods that aim to repair human damage to ecosystems and landscapes. Like restoring a damaged painting or piece of furniture, the main idea is to remove traces of damage and put an area back to how it looked or functioned before humans changed it – what is often referred to as the “pre-disturbance” state. Methods used to do this include removing species, reintroducing species, recontouring and redirecting river channels, re-establishing flooding regimes on wetlands, replanting forests, and decontaminating land.

Sometimes restoration projects aim for a complete redesign of an ecosystem, including its ecological and hydrological functioning. More commonly, though, restoration projects aim only for a limited or partial improvement in a landscape or they focus on particular functions, for example, flooding, prey–predator relationships; in such cases, restoration is sometimes referred to as “environmental rehabilitation” instead. Where environmental restoration is implemented or studied by professional ecological scientists, it is sometimes referred to as “restoration ecology” or “ecological restoration,” because it is being used not only to try to repair environmental damage, but also to test scientific theories of environmental change and ecological development (e.g., Palmer *et al.* 2005).

So the term “restoration” may refer to any one of various levels of intended intervention, such as:

- creating new ecological functions or whole landscapes;
- restoring or (re)creating (some) pre-disturbance or historically defined functions or landscapes;
- reassembling (some) pre-disturbance functions;
- rehabilitating (some) functions, by repairing or reclaiming (parts of) landscapes;
- enhancing or improving functions or aspects or appearance, that is, very partial modification or minimal repair.

One example of restoration is at Wicken Fen, a national nature reserve in Cambridgeshire, England, that has been owned and run by the National Trust, a charity, since 1899. In recent years, the Trust has expanded the wetland area by digging out surrounding farmland (see Figure 1), aiming to restore it to its traditional state, that is, the type of reed wetland existing before the region was extensively drained in the 1940s. This example shows that restoration can sometimes involve *changing* one landscape (e.g., farmland) into another (e.g., wetland), using bulldozers and other earthmoving equipment or fire. This differs greatly from other forms of conservation which usually focus on maintaining a landscape in its current state and thus preventing change.

**Origins**

Restoration is the latest “wave” of environmental protection methods, following on from nineteenth century efforts at preservation...
(e.g., through designating national parks and nature reserves) and twentieth century efforts at conservation and sustainable development (see Sustainable development). Sometimes referred to as “creative conservation” (e.g., Adams 2003), restoration seeks not only to protect valued environments or features, as preservation or conservation methods also try to do, but also to repair damaged environments and, in the more extreme cases, to create new ecosystems and landscapes following human use and abandonment, such as restoring an old quarry pit into a lake for waterfowl.

One of the earliest and most famous examples of restoration began (and continues) at the University of Wisconsin-Madison’s Arboretum in the United States. Aldo Leopold, a major figure in environmental conservation and professor of game management at the University, began to restore prairie vegetation in the 1930s and 1940s by cutting, seeding, and setting fires to cultivated land. His aim was to restore the prairies to their ecological state in the 1840s, before Wisconsin was settled, cultivated, and therefore environmentally transformed by incoming Europeans.

But this pioneering effort was isolated and not promoted further after Leopold’s untimely death in 1948. It was not until the 1980s that restoration was more widely advocated and popularized as a solution for human-related environmental damage and loss of valued landscapes and ecosystems. A key figure was a successor to Leopold: William R. Jordan III arrived at the University of Wisconsin-Madison Arboretum in the late 1970s and began to practice and promote ecological restoration through his scientific and conservation activities. He helped to set up the journal Restoration and Management Notes (now Ecological Restoration), to promote the ideas and practices of restoration internationally. Through his writings and advocacy of what he has referred
to as “the restoration movement” (e.g., Jordan 2003), he encouraged environmental restoration not only as a way to improve environments and their ecological functioning, but also as a way to bring humans closer to nature through rehabilitating their local landscapes in a more “creative” form of conservation (Adams 2003) and one which Jordan (2003, 3) referred to as “a kind of gardening.” Jordan also explicitly linked “nature” and “culture” through the meanings and practices of ecological restoration and argued that this was a more positive, hopeful approach to conserving all manner of landscapes and ecologies. He also used this approach to criticize the emphasis on technical aspects of restoration in the literature and therefore to emphasize also the sociocultural performance and ritual involved in doing restoration work.

Reflecting growing interest in the 1980s, the Society for Ecological Restoration (SER) was set up in 1987 to bring together people and organizations interested in and actively practicing ecological restoration across the world. It now publishes two journals: Restoration Ecology – published by Wiley and including peer-reviewed papers on the scientific and technical aspects of restoration – and Ecological Restoration – published by the University of Wisconsin Press and including articles on the practical and professional issues facing restoration practitioners. SER also publishes an International Primer and Guidelines to encourage the normalization of restoration practice across the world. The Primer states that: “Ecological restoration is the process of assisting the recovery of an ecosystem that has been degraded, damaged, or destroyed.” The Guidelines stated that:

Ecological restoration is the process of assisting the recovery of an ecosystem that has been degraded, damaged, or destroyed. It is an intentional activity that initiates or accelerates ecosystem recovery with respect to its health (functional processes), integrity (species composition and community structure), and sustainability (resistance to disturbance and resilience).

SER considers that the acid test of a successful restoration is therefore that the restored ecosystem should continue to function ecologically, even after restoration projects have stopped and deliberate management has ceased.

There are many other organizations that now promote various methods for restoration. For example, in the UK, the River Restoration Network was set up to promote river restoration by acting as an information exchange for case studies and developing best practice and linking practitioners to share their experiences and expertise. Now called the River Restoration Centre, this brings together river restoration experiences across Europe and runs conferences for interdisciplinary learning between scientists and practitioners of river restoration specifically.

Ideas of restoration often challenge existing and institutionalized methods of environmental management and conservation. For example, river management by the state in the late twentieth century tended to focus on engineering-led approaches to securing water supply and preventing flooding through damming, deepening, straightening, and dredging rivers and removing wetlands. Rising interest in river restoration among practitioners in the 1990s directly challenged such traditions by arguing for more sinuous river channels (see Figure 2) and, most controversially, more flooding of river floodplains. This new paradigm did not completely replace the established ways of managing environments, as it continues to be contested, but it did expand the range of methods employed by state agencies and shifted thinking more toward “working with nature” in the interest of sustainability (e.g., Adams, Perrow, and Carpenter 2004).
Recently, restoration has increasingly been referred to as part of the notion of “rewilding” landscapes. Introduced in the 1990s, the idea of rewilding focuses specifically on restoring species and areal interrelationships, such as reintroducing predators to a landscape and connecting habitats through wildlife corridors. In the Netherlands, the organization Rewilding Europe was initially set up by four conservation organizations (WWF-Netherlands, ARK Nature, Wild Wonders of Europe, and Conservation Capital) to build a network of rewilding projects and

Figure 2  Meanders recreated on the River Skerne, northern England. With permission from AirFotos Ltd, Northumberland Water, and the River Restoration Centre.
promote the ideas and practices more widely. However, the idea of “rewilding” relies upon the contested concept of “wildness” and this makes the term both more immediately graspable in the popular imagination and also more controversial than the seemingly less provocative concept of “restoration” (see below).

So, the initial idea of restoration has taken various forms and different names in different times and circumstances, emphasizing that restoration is often based on an ideal, rather than a reality, and also that it is prey to the vagaries of both the politics of and the changing trends in environmental management.

Controversies

Since the 1980s, environmental restoration has become controversial for both practical and philosophical reasons. As noted above, proponents of environmental restoration have argued that it can have positive benefits through testing and increasing scientific knowledge and also through connecting people with nature. However, critics have argued that not only is environmental restoration impossible in practice, but also that it threatens more legitimate forms of conservation and even may be used to justify causing more environmental damage.

Is restoration possible?

There are several reasons why restoration of environments and landscapes is often regarded as impossible in practice. First, it is argued that no one can truly recreate lost ecosystems, because our knowledge of the environment and how it functions is simply not good enough to correctly design restoration and ensure that it will work. Consequently, mistakes will be made and the resulting landscape will not be what was intended; instead, greater problems and damage may occur, especially where species are being introduced or removed. Evidence of this comes from many ecological interventions in the past, for example, introducing predator species to solve an existing pest problem, only to find that the predator species itself grows to become a problem in turn. And although many examples of restoration have been implemented and studied by scientists, often this is short-term and few are monitored long-term to produce convincing evidence of the success or otherwise of restoration work on ecological functioning (Palmer et al. 2005).

Second, it is difficult to select appropriate and clearly defined historical benchmarks or reference ecosystems that restoration is trying to reproduce, especially when landscapes have changed in different ways across the centuries. In other words, defining the “pre-disturbance” state as the goal for restoration is very difficult and contentious in practice. In the United States, pre-disturbance is often taken to mean “pre-European settlement,” which immediately and problematically “naturalizes” the activities of people living in North America before the Europeans arrived, thus devaluing both their impact (e.g., through deliberate use of fire) and their status as humans. In Europe, such a historical benchmark, however controversial, is in any case unavailable, so restoration projects there tend to hark back to various historical periods of ecological functioning, including preindustrial (i.e., pre-nineteenth century), preclearance or pre-agricultural intensification (i.e., pre-nineteenth century), or even postglacial (i.e., 10,000 years ago).

Third, some valued landscapes have actually been caused or created by human “disturbance” historically, for example, through grazing, tree clearance, excavation, and so on, especially in Europe. In England, some landscapes partly created by human activity are not only distinctive
and valued, but also designated to be protected and preserved, for example, the North York Moors and the Norfolk Broads. Attempting to “repair” such landscapes under the rubric of restoration would thus change and possibly threaten environments that are currently valued and protected.

Is restoration authentic?

A fourth and very different reason why restoration is argued to be impossible in practice is philosophical: even if we do have sufficient knowledge to recreate an ecosystem in all its complex functioning, even after centuries of human damage, and even if the resulting landscape looks and works exactly the same as another such landscape that has never been damaged, the “restored” area can never be an authentic environment because it has been created by humans, not by nature. Critics argue that the value of a restored environment will always be less than the value of the naturally created (but now lost) environment, even if the two environments cannot be distinguished from each other in terms of how they look or how they function ecologically. Elliot (1997) referred to the idea of restoring natural value as the “replacement thesis,” to argue against it on exactly this point. This philosophical argument about how far restoration can truly restore “nature” is really about whether we understand “nature” to be essentially defined as “without human interference,” and this assumption itself is very questionable (see below).

Does restoration threaten conservation?

A second line of opposition to restoration is that it may threaten or undermine other forms of conservation. If we accept that landscapes and ecosystems can be restored after damage, the implication is that damage is less of a problem and this will undermine efforts to prevent damage in the first place. And if we accept that restored environments can replace value that has been destroyed elsewhere, then this could be used to justify further damage.

Restoration has been used controversially in planning cases in precisely this way. At Twyford Down, England, a new motorway bypass round Winchester was partly justified by proposing to relocate valued areas of grassland and its non-human inhabitants such as butterflies to adjacent sites before blasting open the underlying rock to create a new cutting for the road and to restore the old road (now redundant) to grassland. This was an example of what Elliot called the “replacement thesis” being used by the state to support development over environmental protection. A huge and well-publicized campaign against the new road objected both to the destruction of the grassland to dig the bypass in the first place, but also to the claim that the recreated habitats would be equivalent to those lost.

Under Section 106 of the UK’s Town and Country Planning Act 1990, “s106 agreements” can be used to make a development proposal acceptable to planners by mitigating (that is, preventing or softening) the most damaging aspects of a proposal. Referred to as “developer contributions,” these may include restoration payments or works, for example, to restore a development site to its original functioning and design after use, such as when quarrying has ended, or to create comparable landscapes or ecosystems elsewhere to compensate for the loss at the development site. Referred to also as “planning gain,” such habitat “creation” has been used to enable developers to justify new developments by arguing that restoration elsewhere will compensate for damage in situ (e.g., Cowell 2003).
Using restoration as “compensation” or “mitigation” is thus a highly political move. Another example of restoration being used to undermine other forms of environmental protection is where the UK regulatory body, the Environment Agency for England and Wales, used restoration ideas to justify its changing approach to sea defense. The rising cost of defending the coastline against encroachment and the threat of climate change was used by the agency to argue that instead of rebuilding and maintaining sea walls against rising water levels, tidal flooding should be allowed and intertidal wetlands (re)created. At Alkborough Flats in Lincolnshire, England, agricultural land was bought and sea walls deliberately breached by the agency to allow the sea to flood a low lying area delimited by higher ground. Referred to originally as “restoration,” the agency later referred to such works as coastal “retreat,” “managed realignment,” and “sustainable flood management,” reflecting how ideas of restoration are not neutral, but become politicized and are used to justify contentious decisions about environmental management and to represent potentially negative change or damage in a more positive light.

As well as being part of the state’s planning processes, restoration has also been used in the interests of private capital. In the United States, restoration ideas have been developed into a commercial industry of “mitigation banking” (Robertson 2006). In the 1990s, proposals to destroy or damage wetlands were required to include provision to create the equivalent amount of wetland functions (an “ecosystem service,” see Ecosystem services) elsewhere. Private sector interests developed a new industry by making wetlands, having them approved by regulatory agencies as providing those functions, and then selling them as “credits” to developers; rather than creating their own wetlands, developers could therefore simply buy such “credits” and refer to them in their proposals to meet the planning requirements.

Such transfers of equivalent value (monetary and ecological) are highly contentious and have been criticized by geographers and others for attempting to create habitat in the same way that commodities (such as food or cars) are created, by exposing how parcels of land, presence of species, or ecological functions are commodified and used to justify further development in the interests of capital. Drawing on political economy, such approaches focus upon the role played by restoration in creating and substituting forms of “natural capital” and critique the notion that “nature” can be detached, moved, and replaced elsewhere.

As well as these practical and philosophical questions, in many cases restoration is controversial because it threatens other interests, for example, agriculture and hunting. A good example is the reintroduction of wolves to Yellowstone National Park, USA, in the 1990s. Wolves had been almost entirely eradicated from the United States by the mid-twentieth century because they were considered to be a predatory threat to livestock and government provided incentives in the form of bounty payments per head. But changing ideas of ecology and conservation raised interest in using predators to control expanding populations of prey and led to calls to reintroduce wolves to Yellowstone to control the rising elk population by predation, rather than by human-led culling. Since wolves were released in Yellowstone in the late 1990s, a stable population has been established, despite some having been shot, probably by ranchers concerned that they were preying on livestock animals, rather than elk.

In Scotland, a plan to reintroduce beavers to the rivers and streams in Argyll was held up for years by arguments over the likely effects,
leading to some beavers which had been captured for later release then dying in their pens. After a five-year trial was agreed in 2009, a small number of beavers were released and have since begun breeding. Beavers are particularly good examples of restoration by nonhuman agency because not only do they interact with other species, but they literally rebuild environments through damming waterways.

As well as reintroducing species, some restoration projects also aim to remove species from a landscape. For example, gray squirrels were introduced to the UK in the late nineteenth century and have since outcompeted red squirrels, which have been in the UK since the end of the last ice age, ten thousand years ago. This has prompted campaigns to “restore” red squirrels by culling grays. There are also many examples of plant species that are targeted for removal in order to restore environments, often because they were deliberately introduced by humans for aesthetic or commercial reasons, but later expanded and outcompeted other species, for example, Himalayan Balsam, introduced to the UK in the mid-nineteenth century.

Both reintroducing and removing species are controversial because decisions about which species should be present in a particular area often fall back on problematic definitions of “native” species that need to be retained versus “alien,” “invading,” or “exotic” species that need to be removed. For example, red squirrels are classed as “native” to the UK, but grays are classed as “invasive non-native.”

From a geographical perspective, such arguments about “native” and “alien” are problematic. They would no longer be tolerated in a discussion about human migration across national borders and they are becoming increasingly untenable in discussion about nonhumans, as we learn more about how plant and animal species have evolved, moved across the globe and interacted with each other over time. Definitions are also highly time-dependent: for example, what is “native” in a landscape during an ice age may disappear from the same landscape during a warmer interglacial period, so defining “nativeness” as an essential quality of any species may be a rather pointless exercise. And the very idea of being able to clearly distinguish the “native” from the “alien” illustrates a problem with environmental knowledge underlying restoration: nature is not easily compartmentalized and often our taken-for-granted ways of thinking about the environment in such categories fall apart under academic scrutiny.

Should we restore “nature”? Finally, even if all these problems about knowledge, definitions, and competing interests are solved, an important ethical argument remains: do we have the right to experiment with landscapes in this way? Some have argued that any attempt to restore an environment is yet another example of human arrogance in assuming we can somehow “fix” the natural world, an arrogance that is even more pointed if we remember that the damage that restoration tries to repair was induced by human activities in the first place. The ethics of restoration and especially what this means for our understanding of nature (see Nature) is thus an important topic for research.

Environmental philosophers (e.g., Elliot 1997; Katz 1996; Light 2000) have debated the meanings of “restored” environments in terms of authenticity and fraud. If someone sees a “restored” environment but does not know that it has been deliberately restored by environmental managers following human-induced damage, does that mean that the person is deceived into thinking that it was produced solely by “nature”? And if they are deceived, does that matter? Some have argued that any restored environment is
fraudulent because attempting to erase human activity through human intervention renders that environment inauthentic, because it has been shaped by human agency.

Others have argued that this debate over authenticity is misguided because it falsely and unhelpfully separates humans out from “nature” and assumes a dualism between nature and (human) culture which separates the two and values nature over culture as a source of environmental transformation. Instead, it has been argued that we should appreciate that humans are also environmental agents, just as other species are; after all, beavers change the environment by building dams, herbivores change grasslands by grazing, and predators change the ecological balance of populations by catching prey.

In geography, the rise of research approaches variously called “cultures of nature,” “hybridity,” and “actor-network theory” have increasingly challenged the idea that nature can be defined or understood as nonhuman. For example, the recent idea that we now live in a geological era that can be called “the Anthropocene” emphasizes that all Earth’s systems have now been altered (even if only slightly) by human activity and that no “pure” nature, no “wilderness” strictly remains. Such discussions both emphasize how tightly humans, nonhumans, and inorganic environments are entwined but also reject the idea of trying to repurify this entanglement in order to “restore” a pure nature that has supposedly been lost. Here, restoration is in any case a coproduction of humans and nonhumans and thus no less or more “authentic” than any other such coproduction.

The other side of this argument is that geographers have also challenged the focus of most restoration projects on the nonhuman aspects of redesigning landscapes, to the neglect of human aspects. This is because the literature continues to be dominated by natural science perspectives and technical information, which argue that restoration needs to be strongly science-led to be successful and to counter opposition from (nonscientifically trained) critics, such as residents in areas surrounding restoration (e.g., Whalen et al. 2002). As a consequence of such arguments, there remains little social science research into how ideas of environmental and ecological restoration are understood and how far restoration projects are meeting the needs of local human stakeholders in practice. Human geographers (e.g., Adams, Perrow, and Carpenter 2005; Eden and Tunstall 2006; Petts 2007) have drawn attention to this problem and emphasized the importance of local community involvement in restoration projects.

Also, the literature on restoration and examples of restoration projects tend to be confined to areas in the Global North, giving a strong geographical bias to debates in restoration research and its practices. This is perhaps partly because of the longer history of dramatic environmental transformation in many northern countries but also because of the increasing budgets and attention given to conservation there, ironically only often after considerable species and habitat loss.

SEE ALSO: Actor-network theory; Alien and native species; Commodification of nature; Cultures of nature; Ecosystem services; Hybridity; Nature conservation

References


This entry charts the evolution of environmental risk analysis and management. Particular attention is given to the evolution of thinking and analysis, especially in the United States and Europe, and particularly to the role of geography in this evaluation. The entry begins with a discussion of early thinking in hazard analysis, and particularly the role of natural hazards. The discussion then goes on to the analysis of environmental risk, and particularly the structure of such risks and how they relate to the options in risk management. Particular attention is given to uncertainty, the social amplification of risks, the importance of place, and risk communication.

Natural hazards and disasters

The study of natural hazards internationally dates back to 1945 when Gilbert F. White (1945) in the United States published his PhD dissertation entitled Human Adjustment to Floods. This was followed several decades later by work in which White continued to explore how hazard management could be improved by the adoption of an enlarged range of options seeking to reduce the toll of natural hazards. But the first overall assessment of the incidence and effects of natural hazards occurred in 1975 with White and Haas’s (1975) Assessment of Research on Natural Hazards. In this work, they made a general argument that hazard management could employ better planning, land-use controls, and other mitigation measures (Mileti 1999). The work reflected the efforts of over 100 experts (mostly from the United States) to take stock of past, present, and future natural hazards and associated disasters. A decision model, linked with a human “adjustment” model, generated a five-step strategy for coping with natural hazards in the United States and abroad:

1. assess hazard vulnerability,
2. examine possible adjustments,
3. assess human perceptions and estimates,
4. analyze the decision process,
5. identify the best decisions and adjustments.

In particular, a broad approach was mandated, including a prominent role for the social sciences in North America. In short, there was not a new hazards paradigm, but a broader and more integrative and interdisciplinary approach emerged.

Gilbert White continued to spearhead many of these changes in the United States. During the 1960s and 1970s at the University of Chicago he trained a number of geographers who became leaders in hazards research, notably Robert Kates and Ian Burton. Subsequently, he moved to the University of Colorado where he founded the Center for Natural Hazards Research and Applications Information, whose annual meetings became a meeting ground for scholars and practitioners. Meanwhile, the US National Science Foundation established a natural hazards research program which provided stable funding for the evolution of natural hazard and disaster research during the 1970s, 1980s, and 1990s, and continues to this day.
ENVIRONMENTAL RISK ANALYSIS

In 1992, at a Boulder workshop, leaders of the field supported a second assessment which would review the burgeoning natural hazards research that had emerged since the first assessment was conducted. The result was *Disasters by Design: A Reassessment of Natural Hazards in the United States* (Mileti 1999). Not only did this assessment mark the notable developments in the field during this period, but it set forth six major principles by which natural hazards management could move from an adjustments approach to the realization of a sustainable hazards mitigation orientation.

1. Maintain and, if possible, enhance environmental quality.
2. Maintain and, if possible, enhance people’s quality of life.
3. Foster local resiliency to and responsibility for disasters.
4. Recognize that sustainable, vital local economics are essential.
5. Identify and ensure inter- and intragenerational equity.
6. Adopt a consensus-building approach, starting at the local level.

A somewhat different path for the development of natural hazards and natural disasters research was proposed by sociologists at the Disaster Research Center at the University of Delaware. (For an anthropological review of the emergence of this perspective and impact on social policy, see Quarantelli 1998.)

Risk analysis

Early risk analysis approaches

The early history of risk analysis focused on the probability of events and the magnitude of consequences and grew out of the expansion of nuclear power in the 1950s and 1960s – largely centered on reactor accidents. Critical to this was the so-called Rasmussen Report, the *Reactor Safety Study*, WASH-740, published by the US Nuclear Regulatory Commission in 1975, which examined the numerical estimates of the probabilities for different accident scenarios. In the years that followed, a deluge of plant-specific probabilistic analyses occurred. By early 1987, some 15–20 large-scale assessments had been conducted (Kasperson and Kasperson 1987). Similar studies were also conducted in Germany and Sweden. Over time, the function of risk assessment came to be understood as achieving sound risk characterization, understood as the final step in a process that involved hazard identification, exposure assessment, and dose-response assessment, culminating in risk characterization which provided the science needed for decision-making. Consideration of benefits was also recognized as important. The objective of risk assessment research over the early decades, then, was no less than to depict a particularly hazardous situation in as accurate, thorough, and scientifically objective a manner as possible. This process was enshrined in the United States in a National Research Council (1983) report which came to be known as the Red Book. Since that time, and despite its narrow view, it has become the bible of federal agencies and other risk managers in the United States and abroad.

World Commission on Environment and Development

An early departure from the world of the Red Book occurred when the World Commission on Environment and Development (1987) report appeared. The Commission set the goal as realizing sustainability rather than limiting itself to risk and disaster management. It defined
sustainability as “development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (188). This definition has become widely used though sustainability has remained largely a philosophical principle that has not been comprehensively adopted in the environmental risk community.

Critique of the mainstream

Even in the latter part of the last century, not all environmental risk analysts espoused mainstream thinking and models emanating from the early work on natural hazards and nuclear technology. Geographers were prominent in this critique, including Hewitt (1983), who called attention to the vulnerability associated with poverty and social class. In At Risk: Natural Hazards, People’s Vulnerability, and Disasters (Blaikie et al. 1994; see also the second edition, Wisner et al. 2004) and earlier work on social erosion geographers who worked out of a neo-Marxist and development studies approach presented a different picture of the ongoing toll of hazards in the so-called third world.

Enlarging the environmental risk paradigm

Although it has received much less publicity, the National Research Council published a 1996 report entitled Understanding Risk: Informing Decisions in a Democratic Society (Stern and Fineberg 1996). Going beyond the Red Book, this report argues that the process used for characterizing risk must come from a combination of analysis and deliberation. In this context, deliberation involves such processes as discussion, reflection, and persuasion to communicate and consider issues, while increasing understanding and informing decisions. As such, the expanding framework for risk analysis and decisions involves such criteria as:

- getting the science right,
- getting the right science,
- getting the right participation,
- getting the participation right, and
- developing an accurate, balanced, and informative synthesis (Stern and Fineberg 1996, 6–7).

Ironically, although the report mostly enlarges the framework used in the Red Book, it has not been widely adopted by risk practitioners, principally, perhaps, because of the enlarged role given to deliberation as opposed to expert technical knowledge.

Uncertainty

Few issues are more perplexing for environmental risk analysts than uncertainty (Jager et al. 2001). Examples of uncertainty problems abound. If we take a familiar problem, such as the continuing toll of automobile accidents, the causality of collisions is generally well understood, even including how human behavior interacts with the physical system so that events (collisions) can be minimized through traffic regulations and highway and vehicle designs. Compare this with climate change where an array of changes in temperature, precipitation, sea level rise, and severe storms may extensively alter the nature of ecosystems, disrupt coastal systems, and threaten human health. As yet, the timing and spatial distribution of such changes remain beyond our assessment capabilities.

At the outset, it should be clarified what is meant by uncertainty and how it enters into environmental risk analysis and decisions. In cases of uncertainty, we usually mean that the direction of change is relatively well known but the probability of events and the magnitude of consequences, and perhaps even the receptors at risk, cannot be estimated with any precision.
ENVIRONMENTAL RISK ANALYSIS

Three major sources of uncertainty exist: where insufficient data exist to estimate probabilities and magnitude; where insufficient models exist to characterize risk genesis and consequences; and where basic processes in physical and human phenomena are not understood and even the direction and causes of change are not known to science, and thresholds and nonlinear relationships are present but not understood.

The most difficult source of uncertainty is what is typically termed “deep uncertainty,” or alternatively “ignorance” or “epistemic uncertainty.” These are uncertainty situations in which, typically, the phenomena posing potential threats to human societies are characterized by high levels of ignorance and are still only poorly understood scientifically, and modeling and subjective judgments must substitute extensively for estimates based upon experience with actual events and outcomes, or ethical rules must be formulated to substitute for risk-based judgments.

For those who must make decisions, deep uncertainty can be a major swamp – and, paradoxically, an opportunity as well. On the one hand, deep uncertainty is a field for creativity and experimentation. Well-established approaches may or may not be appropriate, and so arguments may well be made for initiatives that strike out in new directions. Timing is also uncertain, so one can wait for political “windows” of opportunity. On the other hand, escape from the imperative to come to decision is also abundantly there. In the political context, delay is a major resource and deep uncertainty provides wonderful cover. For those who seek to escape the exigencies of a decision whose outcome and political fallout remain highly uncertain, or to postpone to the next lucky incumbent, abundant rationale can be made to delay for further research and assessment with the admonition that “we will know better.”

It is important to note that whatever the political objective, delay itself further adds to uncertainty just as the science to narrow uncertainties proceeds. Exporting decisions to unknown future political environments inevitably adds new and often unforeseeable elements to the decision context.

While there is little question about how difficult the determination of risk uncertainty is for the scientific community, it is not an issue for the scientists alone. As Box 1 shows, principles have been identical to guide analysis. Uncertainty also has much to do with the differential pattern of people’s vulnerability, to nature and to human communities, and to the complexities involved in determining optimal management approaches and interventions. To take a prominent example, risk assessment assumes that sufficient knowledge and quantification can be achieved so that command-and-control strategies and regulation can be employed. Much depends on how large the residual uncertainties are and how they affect the acceptability of the risk. Where deep uncertainties exist, other management approaches may be called for, such as adaptive management.

A causal model of risk

Risk assessment is at least a four-step process involving risk identification, assignment of priorities, risk evaluation, and risk management.

Risk identification

Hazard managers do not like surprises. What is unacceptable, indeed downright dangerous, to the continued wellbeing of the manager is to miss the existence of a risk completely. Risk managers have available to them a variety of methods for identifying hazards, including research, engineering analysis, screening, monitoring, and diagnosis. Some of these sources of information are incidental to risk identification: thus, new carcinogens may be found as byproducts of
cancer research, product failure mechanisms may be recognized in engineering analysis performed for other reasons, and surprise hazards emerge as outbreaks or clusters. Other sources of information are the product of purposeful efforts: thus, screening of chemicals provides early warning of toxicity, and environmental monitoring of pollutants affords estimates of potential health effects.

Despite an enlarged capacity, some technological risks escape timely identification, of course, and become known through outbreaks or experienced consequences. Thus, the threat of buried chemical wastes at Love Canal in the United States was unrecognized until severe winter storms raised the water table and injected noxious chemicals into the basements of residences (Ember 1982).

Assigning priorities

Risk managers cannot, of course, deal simultaneously with all identified risks within their domains of responsibility. Somehow the risk domain must be ordered and priorities attached to the many candidates competing for managerial attention. There are choices to be made—choices between risks with better-known acute consequences and those with poorly understood chronic consequences, or choices between attending to serious risks with few available sources of control and attending to lesser risks with effective available means of control. The criteria for establishing priorities are laden with value considerations: Is it the aggregate risk or the distribution of risk that is more important? Should ecological risk receive lower priority than health risks? Should children enjoy a higher priority for protection than adults? Should present generations be valued higher than future generations? Inevitably, establishing priorities requires trading off some values to achieve others. Perhaps one of the most value-laden decisions a risk manager, or monitor for that matter, makes is the initial one of what to work on.

The message is clear. Effective risk management requires a well-ordered risk domain. Since creating such a structure is intrinsically normative as well as scientific, it should be rationalized openly and in close consultation with the various interested stakeholders (Renn, Webler, and Weidemann 1995).

Risk evaluation

Once a risk has been identified, the next steps in assessment are: (i) to estimate and characterize scientifically the probabilities of specific events and related consequences and (ii) to evaluate this characterization in social terms.

Unfortunately, there is no simple relation between scientifically estimated risks of death and injury and the social valuation of a given hazard. In fact, much to the chagrin of many scientific risk analysts, the public apparently does not respond equivalently to equal threats of mortality. A high-mortality hazard such as an auto accident sometimes provokes relatively little fear, whereas some low-mortality hazards (e.g., botulism or nuclear power) evoke great anxiety. A great deal of confusion and conflict in risk management arises from this conundrum.

Beyond the question of how best to characterize risks, risk evaluation presents other serious problems for the management process. There is the recurrent need to attend to secondary and tertiary effects, or, equivalently, to define the full range of possible consequences. A nuclear accident, for example, may produce fatalities, injuries, property damage, and a high level of anxiety. But it can also lead to regulatory change that result in subsequent shutdown of all similar plants, producing disruption, further anxiety, and possible power outages. Such secondary consequences are rarely predicted and infrequently
Box 1 Recommended principles for uncertainty and variability analysis.

1 Risk assessments should provide a quantitative, or at least qualitative, description of uncertainty and variability consistent with available data. The information required to conduct detailed uncertainty analyses may not be available in many situations.

2 In addition to characterizing the full population at risk, attention should be directed to vulnerable individuals and subpopulations that may be particularly susceptible or more highly exposed.

3 The depth, extent, and detail of the uncertainty and variability analyses should be commensurate with the importance and nature of the decision to be informed by the risk assessment and with what is valued in a decision. This may best be achieved by early engagement of assessors, managers, and stakeholders in the nature and objectives of the risk assessment and terms of reference (which must be clearly defined).

4 The risk assessment should compile or otherwise characterize the types, sources, extent, and magnitude of variability and substantial uncertainties associated with the assessment. To the extent feasible, there should be homologous treatment of uncertainties among the different components of a risk assessment and among different policy options being compared.

5 To maximize public understanding of and participation in risk-related decision-making, a risk assessment should explain the basis and results of the uncertainty analysis with sufficient clarity to be understood by the public and decision-makers. The uncertainty assessment should not be a significant source of delay in the release of an assessment.

6 Uncertainty and variability should be kept conceptually separate in the risk characterization.

Source: NRC 2009, 120

analyzed, even though in many cases they are the most important consequences of a particular event.

Following the risk assessment, control analysis judges the tolerability of the risk and rationalizes the effort that is made in preventing, reducing, and mitigating a risk. The key link between risk assessment and subsequent initiation of control actions is the judgment of whether a hazard is tolerable or not. One of the most perplexing issues facing hazard managers, it has been mislabeled the acceptable risk issue (Fischhoff et al. 1981). It is unlikely that any risk is “acceptable” if it is unaccompanied by benefits or is susceptible to easy reduction. Many risks are imposed upon individuals, often without warning or information. Such risks are better thought of as “tolerated”; they are suffered in practice, not accepted.

For identifying means of control, the causal structure of the hazard becomes the central concern. Complex cases require a full fault- or event-tree analysis. Whether simple or complex, an analysis of causal structure must be broadly based and include potential control actions that span the range from human needs and wants to exposure, consequences, and mitigation.
of consequences. To this end, it is useful to contemplate the control structure for the traditional technological risk, within the seven-stage model of causal structure. The potential control actions include: (i) modify wants; (ii) choose alternative technology; (iii) prevent initiating events; (iv) prevent releases; (v) restrict exposure; (vi) block consequences; and (vii) mitigate consequences.

In addition to identifying the technical means of control, each control action can be implemented in a number of different ways. There are three major modes of inducing society to undertake the control action. Society can: (i) mandate the action by law, administrative regulation, or court order and thereby ban or regulate the product or its use or distribution; (ii) encourage the action through persuasion or by providing incentives, penalties, or insurance; or (iii) inform those creating or suffering risk, allowing them voluntarily to reduce or tolerate the hazard. At any moment all of these modes may be utilized in connection with a specific hazard. Thus, the hazard of driving is controllable by 37 different “highway safety countermeasures” and a comparable number of vehicle safety standards, each involving one or more stages of the causal structure of the hazard. For any given case, a lively ideological debate may erupt over which implementation modes are the most desirable.

An important aspect of control analysis is to inquire into the relative cost of control interventions (be they technical, behavioral, or informational) and modes of implementation. Known as cost-effectiveness analysis, this approach permits the risk manager to select the most efficient actions available.

Risk management

Equipped with a risk assessment and a control analysis, and assuming the risk has been judged intolerable and thus requiring action, the manager next must design a risk by society management strategy, consisting of an overall management goal. Such a strategy needs to take into account the need for “thick analysis” of longer-term issues in sustainability (Adger et al. 2003). The control package will specifically include both control interventions (oriented toward the causal structure of hazards) and modes of implementation (oriented toward alternative institutional means for control).

Risk acceptance may be achieved by providing compensation, as through higher wages for riskier work, or by seeking informed consent, as in informational or warning labels on hazardous products. The purpose of the latter is to make the risk more voluntary by increasing information related to technology choice. It applies to a broad range of risks.

Risk spreading seeks to transform a maldistribution of risk into a more equitable one, through redistribution of the risk over social groups, regions, or generations. The new distribution may also seek to equalize experienced risk, to make risk concordant with benefits or with the ability to bear risk. An interesting example of risk spreading is the introduction of tall stacks to transform a local pollution problem into a regional one. Here, by all appearance, the principal regional hazard, acid rain, was initially unsuspected, and only at a later time became recognized as a serious ecological and health threat.

Risk reduction, in contrast to spreading and acceptance of risk, involves decisive intervention in the causal sequence of hazards. It is therefore a step that may in some circumstances curtail the benefits of technology. An extreme case of reduction is aversion, as exemplified by the total ban on a technology.

Risk mitigation includes a variety of ways of modifying hazard consequences once they
have occurred. Typical actions include disaster relief, medical intervention, family assistance, and compensation through insurance or other means. Risk mitigation is often an initial societal response when risks have not been anticipated or when the causal chain of the hazard is poorly understood, as exemplified by the thousands of court cases now pending against asbestos manufacturers.

Although presented here as an integrated managerial approach, strategies usually develop piecemeal, frequently lack internal logic, and may appear only through trial and error over time (Perrow 1984). However they develop, they must eventually strike a balance between reaping technological benefits and reducing unwanted risks. In addition, governance needs to be adaptive and to seek longer-term resilience (Folke et al. 2005). Control actions will range along the causal chains of hazards, reflecting optimal points of intervention. In contrast, for preventing the disruption of the ozone layer, a feared and intolerable consequence caused in part by a minor technology (aerosol cans), upstream intervention based on banning is both reasonable and acceptable. For most hazards, midstream and downstream strategies of intervention are appropriate. They interfere less with benefits and can be directed at specific targets. Typical of such interventions are the use of filters on cigarettes and the wearing of seatbelts in cars. Neither strongly affects the benefits, and each reduces hazard consequences. When the causal structure is poorly understood and unpredictable, society by necessity concentrates on mitigating consequences. An example of this is the case of environmentally caused cancer, where both agents and mechanisms are to a large extent unknown.

A mature hazard management strategy is one which over time seeks a preferred path between realizing technological benefits and reducing unwanted risks. It will normally employ a complex set of interventions along the hazard chain and utilize a variety of managerial modes of implementation, and includes both “normal” and “post-normal” risk assessment (Funtowicz and Ravetz 1992). Such a system evolves partly through improved knowledge and partly through trial and error.

Vulnerability

The concept of vulnerability, defined here as “the differential susceptibility to loss from a given event,” is a central but still incompletely developed concept (see Adger 2006). It is abundantly clear that the impacts of environmental change or stress lie partially with the character of the extent and exposure – and partially with human and social vulnerability systems, people, and social organizations affected. Sen (1982) has shown powerfully that food crises can occur even with food available, finding their roots instead in changing social and economic entitlements. Environmental risk, in short, is an interactive phenomenon involving both nature and society, particularly inequality and a lack of buffering against environmental threats.

Although the specific terms, concepts, and criteria used to address vulnerability differ, basic concepts are consistent: the ability to continue function within a normal range despite perturbation and the ability to overcome perturbations that substantially disrupt the normal functioning of the system. These two concepts address internal characteristics of the affected system. Economic, political, and biophysical changes may interact to alter the vulnerability or resilience of the system. Differing efforts at defining vulnerability reflect the character of the concept. Vulnerability is a useful general
concept that depends on the definition of event and social context to provide insight into specific key situations. Accordingly, the factors influencing vulnerability vary widely among scales of analysis, so that scale-sensitive analyses of environments remain a pressing need. Essential steps in assessment can nevertheless be defined (Schröter, Polsky, and Patt 2005).

The structure of environmental risk, then, reflects human driving forces, patterns of environmental change, the fragility of ecosystems, the vulnerability of social groups, and the adverse effects that result (Pelling 2011). The stages in the evolution of risk consequences are highly related and interactive. The magnitude of the impact will depend not only on the extent of the environmental change but also on the character of social institutions and relations contributing to the resilience or capacity of a group in the region to be harmed (i.e., its vulnerability).

Obviously, different attributes of an environment are highly relevant to analyses of human-induced environmental stresses. Fragile environments degrade more readily under mismanagement and exact higher societal costs for risk management or for substitution in production. More robust environments do not degrade so rapidly and respond to substitutes and mitigation interventions more economically. The projected rapid rate of change relative to the rate at which species reestablish themselves, the isolation and fragmentation of many ecosystems, the existence of multiple stresses, and limited adaptation options suggest that some ecosystems (e.g., forests, mountain systems, and coral reefs) are particularly vulnerable to climate change.

Economic marginality describes a situation in which the returns to an economic activity barely exceed the cost. Again, it may be that small changes in physical conditions affecting economically marginal practices will result in great changes. Uncertain changes in environmental patterns may remove some of the buffering offered by established coping strategies and institutions based upon “peasant science” and the local tradition of indigenous knowledge and experimentation (Watts 1987, 179). The concept of entitlements – access to resources and assistance – suggests a wide range of circumstances surrounding everyday life that may influence vulnerability (Sen 1982). Decisions about risk management also depend on the social context, such as information channels, types of entitlement available, psychological factors of risk heuristics and biases, and how decisions and risk problems are framed (Kahneman, Slovic, and Tversky 1982).

The nature of social and economic institutions is yet another factor that affects vulnerability. The infrastructure of social institutions pervasively shapes the breadth and depth of response capability. For institutions, the ability to benefit (or to avoid harm) rests largely with the ability to react rapidly to opportunities. The ability of social institutions to provide warnings, diagnosis, insurance, and planning support is a key element in shaping successful societal coping.

Demographic and health considerations among human populations represent an area of specific concern and investigative depth in relation to a number of hazards. Population growth in marginal areas – such as coasts, drylands, and degraded areas – may expose more people to additional stresses. Other demographic and social characteristics (such as age and occupation) of those populations will also be important in evaluating overall levels of vulnerability. Special needs populations (the elderly, those with disabilities, lactating women, children under five) are among the most vulnerable persons in natural disasters. Vulnerability, in short, leaves a fingerprint that becomes increasingly complex as more social relations and scale disaggregation are
integrated into the analysis. This is even more the case when ecological effects are included.

One of the most perplexing problems in risk analysis is why some relatively minor risks or risk events (as assessed by technical experts) often elicit strong public concerns and result in substantial impacts on society and the economy. Such concerns and impacts are typically the result of social amplification – changes in risk perception and response based on psychological, social, institutional, and cultural processes. Social amplification is most likely to flourish when the risks are serious and the situation is fraught with uncertainties.

Social amplification of risk

The social amplification of risk framework seeks to understand and account for public attitudes toward risk. The framework links the technical assessment of risk with psychological, sociological, and cultural perspectives of risk and risk-related behavior (Kasperson et al. 1988; Pidgeon, Kasperson, and Slovic 2003). The main thesis of the framework is that risks interact with these perspectives in ways that may amplify or attenuate public responses.

Risk amplification typically occurs at two stages in a risk scenario: in the transfer of information about risk and in social response mechanisms. Signals about risk are both transmitted and processed by individuals and social entities, called amplification stations. The individual might be a scientist, for example, who communicates the risk assessment; a social entity might be the news media, a cultural group, or an interpersonal network. The perceived amplified risk leads to behavioral responses that result in secondary impacts or ripples.

Social amplification may qualitatively and quantitatively increase not only the perception of risk but also the risk itself and its consequences. For this reason, social amplification of risk must be included in analyses of public and regulatory reactions to risk events.

The key amplification stages are listed below:

- filtering signals (only a fraction of all incoming information is actually processed),
- decoding and reframing signals,
- processing risk information (e.g., drawing inferences),
- attaching social values to information as a basis for drawing implications for management and policy,
- interacting with cultural and peer groups to interpret and assess the validity of signals,
- formulating behavioral intentions about whether to tolerate a risk or take action against the risk or risk manager,
- engaging in group or individual actions to accept, ignore, tolerate, or change the risk.

The starting point in the social amplification framework is a risk event, which might be an actual or hypothesized incident (or even a new report about a known risk) and which may be minimal, largely irrelevant, or localized in its impact, unless it is observed and communicated to others and thus amplified. The characteristics of the risk are then portrayed through communication signals that interact with psychological, social, institutional, and/or cultural processes in ways that intensify perceptions of the risk and its manageability. The experience of risk thus involves not only concern about potential physical harm but also interpretations of risk by groups and individuals. The social amplification of risk framework enables effective assessment of this multidimensional risk experience, secondary and tertiary consequences, and the actions of risk managers, stakeholders, and the public.

The term “amplification” refers to the process of various social agents generating, receiving,
ENVIRONMENTAL RISK ANALYSIS

interpreting, and passing along risk signals, which are always changed in the process. In fact, risk signals are subject to predictable transformations as they filter through social and individual amplification stations. The social amplification stations generate and transmit information via communication channels such as mass media, social media, letters, telephones, and face-to-face conversations. The transformations may increase or decrease the volume of information about an event, heighten the salience of certain aspects of a message, or reinterpret and elaborate available symbols and images, leading to particular interpretations and responses by those who next receive the information.

Individual amplification stations are affected by risk heuristics (i.e., qualitative aspects of the risk and context such as attitudes, blame, or trust). Individuals are also members of cultural groups and other social units (social stations of amplification) that codetermine their risk perception. Individuals in groups and institutions do not simply pursue their personal values and social interpretations; they perceive risks, those who manage them, and the risk problem through the lens of values of the organization or group and, perhaps, its cultural biases. Thus, there is often an “organizational amplification” of risk that must be considered.

Social amplification also accounts for the secondary and tertiary consequences, or ripples, of some events. Like ripples in a pond, they may spread far beyond the initial point of impact and may even affect previously unrelated groups or institutions. Imagining ripples in a pond is a good way to think about how impacts associated with the social amplification of risk spread outward from those directly affected (or first notified) to the next, institutional level (a company or an agency), and, in some cases, to other parts of an industry. For example, in the wake of the Deepwater Horizon explosion in April 2010, the effects spread from the drilling rig to the rest of the Gulf of Mexico, including the wetlands and beaches in all of the adjacent states, and then to politicians and petroleum industry representatives who were compelled to reconsider their plans to expand offshore drilling.

The concept of rippling impacts suggests processes that can extend (in risk amplification) or constrain (in risk attenuation) temporal, sectoral, and geographical impacts. It also illustrates that each order of impact, or ripple, may not only have social and political effects but also trigger (in risk amplification) or hinder (in risk attenuation) managerial interventions to reduce risk. Such secondary effects include market impacts (e.g., consumer avoidance of a product or related products), demands for regulatory constraints, litigation, community opposition, loss of credibility and trust, stigmatization of a facility or community, and investor flight. They may also include some or all of the following effects:

- enduring changes in perceptions, images, and attitudes (e.g., anti-technology attitudes, alienation from the physical environment, social apathy, stigmatization of an environment or risk manager),
- losses in local business sales, lower residential property values, and lower levels of economic activity,
- political and social pressure (e.g., political demands, changes in political climate and culture),
- changes in the nature of the risk (e.g., feedback mechanisms that heighten or lower the risk),
- changes in training, education, or required qualifications for operations and emergency response personnel,
- social disorder (e.g., protests, riots, sabotage, terrorism),
ENVIRONMENTAL RISK ANALYSIS

- changes in risk monitoring and regulation,
- higher liability and insurance costs,
- repercussions on other technologies (e.g., lower levels of public acceptance) and on social institutions (e.g., erosion of public trust), as when the 1984 explosion at the chemical plant in Bhopal, India, raised concerns about the possible failure of “fail-safe” systems at nuclear power plants.

Once secondary impacts are perceived by social groups and individuals, they may lead another stage of amplification and tertiary effects that may affect other parties, more distant locations, or future generations. Each order of impact may also trigger (in risk amplification) or hinder (in risk attenuation) positive changes for risk reduction. Examples of positive changes were apparent in the wake of the Fukushima nuclear accident of 2011 in Japan when Germany restructured its energy system and the United States (among other countries) launched a major review of all its nuclear plants.

Risk and place

Geographers have been particularly interested in how the nature of a particular place affects the risk experience. The hazardous facility siting literature illustrates how new facilities and technologies can change how people feel about a place because of worries about environmental contamination and health risks – people may feel the place is now unsafe (Slovic, Flynn, and Gregory 1994). Studies can show that places can be associated with fears about risks to health and valued amenities and the place can become stigmatized, affecting retirement plans and tourism. Places can be compared by their risk portfolios and by public concerns about risks (Boholm and Löfstedt 2004).

Facilities, as well as associated infrastructure such as electricity transmission lines, may also be perceived as changing the character of an area or of a community. For example, siting of new energy sources can be seen as a taking of public space or the industrialization of the seascape.

There are some, however, who caution against emphasizing visual (and noise) concerns too much, even when they are clearly salient, as there may be other reasons for the strength and direction of opinions (Firestone and Kempton 2007). They suggest that a broader set of reasons are often important to negative opinions, including potential impacts on tourism, recreation, and the environment, as well as factors such as perceived inadequacy of the decision process and prior experience with facilities and developers.

Social trust and risk

Trust is related to expectations about others and orientations toward an uncertain future. The “others” of whom individuals and groups have expectations include actors at various scales, such as individuals and institutions. In the context of risk communications and management, trust has been the subject of extensive research by geographers and others (Siegrist 2010). Discussions have focused variously on general trust, interpersonal trust, social trust, and institutional trust. Interpersonal trust refers to the expectations that individuals have about others, and it emerges primarily through direct interaction. On the other hand, social trust is a feature of more distant or indirect relations between people and others. In complex sociotechnical systems the focus of managing social trust (or distrust) often refers to the expectations that people have about institutions and the agents acting on their behalf. Kasperson, Kasperson, and Golding (1999, 37) argue that social trust is “onion-like”:
the deepest level, the core of the onion, involves trust in the political community that underlies the constitutional structure of politics and democratic institutions; next is a layer of trust in the political regime; then comes trust in governmental and other political institutions; and finally, there is the level of trust that relates to particular representatives of the institutions (Kasperson, Golding, and Tuler 1992). The importance of social trust for risk communication arises for a variety of reasons. First, social trust helps to create the conditions by which people with conflicting preferences or values can collaborate to manage risks – through dialogue, problem-solving, and negotiation.

While social trust can facilitate agreement, acceptance, and support, it cannot guarantee these outcomes, although this is often an implicit hope of those who seek it. Trust is particularly important for stable arrangements in contexts of unequal power. Risk-related controversies can be a function of ways that people think about the competence or fairness of risk managers (Kasperson, Golding, and Tuler 1992). Opposition to siting of hazardous facilities often grows from negative judgments about the competence or fairness of those seeking to place risks on a host community.

While there is considerable discussion of the nature of trust in the context of risk management and communication, the general conclusion is often that social trust is lacking. Also, it is clear that there are many tools for risk communication, but they carry a trust ingredient. Moreover, a long-term shortfall in public confidence in social institutions has been documented in numerous polls and surveys.

There is scant evidence that the rebuilding of trust can be accomplished in the near term, despite the view of some risk managers that they will quickly build trust where it is lacking. Actions that might increase social trust among some groups erode social trust in others. To the extent that social distrust is systemic in nature, system change is required for changes in the level and objects of social distrust. Thus, there is no simple answer as to how to rebuild social trust in many pluralistic (or even fragmented) societies.

A second reason is that changing negative views can be very difficult. The “affect heuristic” suggests that when people like an activity or technology, they tend to view it as having high benefit and low risk (Slovic 2012). On the other hand, if they dislike it, they see benefits as low and risk as high. Thus, information intended to educate or persuade is all too often impotent. In contexts of information complexity and uncertainty, the affect heuristic can be a powerful mechanism for making judgments about risks and those that managed them, in much the same way as trust is used to make judgments about those who manage risks in contexts of information complexity and uncertainty.

A third reason is that our understanding of which events and activities build social trust and which erode it is still limited. Slovic (1993) found that events and activities presumed to erode trust can have a stronger impact on overall levels of trust than do those thought to strengthen social trust. This is often referred to as the “asymmetry of trust,” and is the source of the aphorism that “trust is easy to lose but hard to gain.”

Although there are many challenges to rebuilding and maintaining social trust, there is evidence that institutions responsible for planning, operations, and risk communication can be designed to function in the presence of social distrust, rather than seeking trust before moving forward. Social distrust can serve as an important social and political factor in risk communication (Pidgeon, Kasperson, and Slovic 2003). Functional distrust, or “critical trust,” can ensure oversight and vigilance. Critical distrust can
ENVIRONMENTAL RISK ANALYSIS

hold in check the growing power of elites and technical expertise, generate alternative control mechanisms in democratic institutions to maintain social order, and encourage realistic appraisal of the operation of elites, and the actions of democratic institutions, without fostering the withdrawal of mandates. So, the pressing need is not to maximize trust, but to concoct the mixtures of trust and distrust that should prevail within a political risk management institution.

Risk communication

The appearance of Effective Risk Communication (Árvai and Rivers) in 2014 indicates how important risk communication is in environmental risk assessment and management. In 1986, the first National Conference on Risk Communication, attended by 500 experts, was held in Washington, DC. This review was intended to take stock of the burgeoning field of risk communication and the explosion in the early 1980s of books, scholarly articles, and handbooks purportedly providing guidance on “how to do it.” The conference dissected the level of thinking and practice.

A first major finding in risk communication is the importance of persistence. It is not unusual for either the public sector or the private sector to provide a modicum or token of support for risk communication at the end of a lengthy risk assessment process. The success of the anti-smoking campaign in the United States was due in no small part to its decade-long effort to convey a wide array of information and warnings of the risks of smoking. Effective risk communication requires sustained effort throughout the risk deliberation process, with learning along the way. This means that the resources that are needed are often far larger and the time required much longer than many risk managers anticipate or undertake.

Also, as Fischhoff, Bostrom, and Quadrel (1993) pointed out 20 years ago, the scope of the communication must be broad and comprehensive enough to embrace the issues of concern in the policy or decision arena. Part of the success in the anti-smoking effort was due in no small part to the fact that the communicators took head-on the notion that attractive women look so sophisticated with a cigarette dangling from their graceful fingers. The notion that smoking was a dirty habit may have been as important as pointing out the added cancer risk. Similarly, the failure in radioactive waste disposal to broaden the debate to include civil liberty and equity issues (such as the distributions of risks and benefits across generations and the location of waste generation versus the location of risk-bearers) has excluded important issues of public concern.

An indispensable element of risk assessment and management is characterizing and communicating uncertainty. Uncertainty comes in a wide variety of types and flavors. And neither decision-makers nor publics like it very much. Despite decades of work on communicating uncertainty, the track record is not encouraging. Variability, involving different risk exposures and sensitivities, is still often confused with uncertainty and the unknowns in the structure and experience of risk. And even risk experts often confound different sources of uncertainty which arise alternatively from data inadequacy, limitations in model parameters, and imperfect understanding of the scientific phenomenon in question. As a result, risk managers often see more data or “working harder” as the best answer to reducing uncertainty despite accumulated experience that points out that more scientific work often uncovers new uncertainties while reducing existing uncertainties, making management no easier. Then again, value issues often creep into the identification and assessment of uncertainty, making decisions no less
problematic. Accordingly, the issue of how to communicate uncertainties to decision-makers and publics remains problematic and risk experts often content themselves to say that society and decision-makers must learn to live with uncertainty, a stance that remains no less problematic than it was at the first national risk communication conference nearly 30 years ago.

So the question remains as to what can be learned from this record of grappling with uncertainty. First, it is clear that decision-makers and publics do not require the full catalogue of uncertainty that science can create. Indeed, sharing everything that experts know about risk uncertainty may be highly counterproductive for decision-makers and publics. Rather, they do need to know those uncertainties that really matter to the magnitude of the risk and its management. Equally important, uncertainties lodged in the value systems of those who will bear the risks need to be brought to the fore and integrated into the assessment and management processes. It is also clear from “mental models” research (Morgan et al. 2002) that understanding how perceptions are lodged in cognitive structures can be very helpful in risk communication. And there is a need to move to resilience and even beyond (Pelling 2011). Otherwise, important uncertainties will remain unknown but quietly sabotage decision-making.

**Conclusion**

While environmental risk analysis has attracted much attention from both scholars and decision-makers, the field is poised for yet new topics and directions. The pre- eminent role of the United States and the place of natural hazards are undergoing major transformations. Climate change is revealing the global scale of the issues that confront the next generation of geographers and environmental risk analysis. The speed of economic development and technological change poses new challenges for analysis and management, and above all for international governance and cooperation. Other sciences will be increasingly important and new theoretical and management integration will be required. The new emphasis upon building resilience will draw geographers and others into the theoretical developments and planning initiatives. Meanwhile, the scope of the analysis and those bearing future risks will inevitably shift to Asia, Africa, and South America. New thinking and analysis will be needed as the scale and importance of these changes become clearer and more salient.

**SEE ALSO:** Climate change adaptation and social transformation; Environmental change and social learning; Environmental hazards; Social resilience and environmental hazards

**References**


Environmental risk analysis


The debate about environmental science and society refers to the social influences on knowledge about environmental changes and risks. Many orthodox analysts claim that scientific inquiry is a representative and trusted means of understanding environmental risks in a way that is free from social influence. Increasingly, however, various social scientists have argued that scientific knowledge can never be free of social influence, and that, consequently, it is difficult to identify natural processes and objects that are not in some way framed by social thinking. The differences between these perspectives have important implications for understanding the nature of environmental risk, and for how disagreements about risk should be settled. The debate has also influenced parallel discussions about the nature/society divide, and the politics and authority of expertise.

Environmental science and knowledge

Science is usually understood to be trusted and accurate knowledge that has been created using established methods by trained analysts. The term “scientific” is often used to denote knowledge that is different from individual opinion or political influence. The debate about environmental science and knowledge partly involves a criticism of scientific methods for complex social phenomena, as well as how scientific knowledge is a site of political contestation and power itself.

First, “science” itself is a varied and changing practice. Most scientific research employs the term “positivism” to describe a scientific method that seeks to make generalizations about trends or objects in rigorous and trusted ways, rather than on the basis of individual cases or personal opinions. Positivist science has been used with great success for various outcomes in human history, but social scientists have also questioned its appropriateness for explaining social phenomena. Positivism has changed over time. Early positivism, such as by the physicist Ernst Mach (1838–1916), looked for patterns in datasets. The Vienna Circle of logical positivists in the 1920s advanced this approach by seeking to verify observed patterns by looking at other datasets. Karl Popper (1902–1994) went further by arguing that theories should be considered true until empirical tests were able to falsify them. At each stage, the predictive potential of positivism has increased. But social scientists have argued that it simplifies how far generalizations can be made, or how far datasets can actually test assertions. Most analysts also believe that scientific progress does not always occur by empirical research, but by social factors driving new “paradigms.”

Second, social scientists also point out that much discussion of “science” in public and policy debates does not always reflect recent scientific research, but instead is better described as knowledge that maintains authority despite the existence of counter-evidence. This type of knowledge owes its authority to cultural,
political, and social factors that lie outside of scientific methods. A key concern here is the long-standing debate about the relationship between facts and norms. Commonly, many policymakers or scientific advisers seek to justify normative decisions by asking “What are the facts?” Social theorists, on the other hand, have often argued that social norms drive the generation of facts — and that existing facts, therefore, reflect historic norms, which in turn reflect unequal power relations in society.

Debates about environmental science have also shown these divisions. Many researchers adopt a so-called knowledge deficit model that argues there is a need to resolve uncertainty about environmental risk by seeking more facts and information. Sociologists of scientific knowledge, however, have argued that it is also important to note that certainty also implies the absence of contestation, and hence analysts should ask what social factors lead toward some questions being identified as certain and others as uncertain. Often, certainty also implies the creation of networks of expertise or arenas of debate that are not open to public participation (Funtowicz and Ravetz 1993).

A related debate also concerns the ability to label objects or scientific causality as either “natural” or “social.” Many discussions about environmental risks engage with biophysical hazards or environmental changes that are commonly labeled as natural, or as uncontrolled by human agency. Research about so-called natural hazards has indicated that many risks posed by events such as landslides or flooding can affect vulnerable social groups more than others, and hence society can influence risks. More generally, sociologists of science have also questioned the possibility of calling anything “natural” because all knowledge is generated through some kind of social gaze (Latour 1993).

Some physical scientists have expressed concern that these comments might reduce attention to natural concerns beyond human interests (Soulé and Lease 1995), or that “social constructivist” approaches to knowledge are culturally relativist because they reduce the ability to make truth claims to the predominance of different social positions. But, against this, social scientists have pointed out that the intention in discussing the social influences on environmental knowledge is to reach a more useful and inclusive understanding of environmental risks rather than to deny that they exist. Moreover, they argue that there is a need to develop new approaches to understanding environmental science that acknowledges the historically contextual nature of knowledge about the environment, and the political arenas in which knowledge is generated and used.

Indeed, social scientists have expressed concern that much currently accepted science might provide simplistic explanations or facts by which to measure or manage environmental problems. For example, analysts have argued that current debates about anthropocentric climate change tend to focus too exclusively on atmospheric greenhouse gas concentrations as the means by which to drive climate change policies. Global climate models focus on atmospheric changes such as in greenhouse gas concentrations or radiative forcing, and climate policies emphasize relative rates of fossil fuel use. Fossil fuels, or carbon, are an important part of the climate change problem, but focusing on these aspects alone might lead to the avoidance of questions about how and why climate change is considered a problem in different contexts or the means by which people can reduce social vulnerability to it (Demeritt 2006). Indeed, in many developing countries, climate change is a problem because of a lack of social resilience, which might be achieved through the appropriate use of fossil
fuels. Consequently, social scientists have argued that measurements of environmental changes per se, such as increasing greenhouse gas concentrations, should not be conflated with the risks and vulnerabilities of people who experience these changes, and that, hence, environmental science needs to seek more holistic and inclusive ways of measuring risk.

There is therefore a need to understand the social contexts within which both environmental changes are considered meaningful and some environmental facts are considered representative, and to find ways to generate new, and more representative, environmental facts. Some analysts use the term “coproduction” to indicate the mutual generation of knowledge and social order (Jasanoff 2005). Social science analyses of environmental science might therefore seek to understand which social outcomes are supported (cognitively or tacitly) by existing environmental knowledge, and what alternative outcomes might arise from new knowledge, and vice versa.

**Theorizing science and society**

Various approaches attempt to explain how social forces shape environmental science and facts. These approaches have developed over time. One early approach, known as cultural theory, held that the production of authoritative knowledge lay in the essential differences between different individuals in all societies (Thompson, Ellis, and Wildavsky 1990). They identified four key worldviews – individualism, egalitarianism, hierarchism, and fatalism – which could be associated respectively with the social positions usually adopted by businesses, environmental activists, governments, and powerless citizens and workers. Cultural theory argued that the common disagreements about environmental problems usually came down to the generation of knowledge and the respect given to knowledge by different worldviews, whose approaches to problems were very different from each other. This approach has been used to explain the differences between environmental skeptics and activists, as well as the widely diverging statistics generated on environmental problems such as deforestation.

Hence, cultural theorists argued that we should not ask “What are the facts?” but, instead, “What would you like the facts to be?” Successful environmental governance does not lie in asking which perspective is right but in acknowledging that they all contain elements of truth.

During the 1990s, analysts began to adopt the insights of more poststructuralist theorists such as Foucault and Latour. This kind of research analyzed the specific contexts where social forces have acted cumulatively to make science socially selective and unchallenged. For example, analysts have used the term “environmental narratives” to describe convenient explanations of environmental problems that are adopted in common discourse or by specific organizations, and which organize complex biophysical changes simultaneously with the creation of social roles of blame and responsibility (Hajer 1995; Forsyth 2003). For example, the problem of desertification in Africa has been defined, and redefined, over a period of years, often in ways that allow states to legitimate environmental policies that would otherwise be challenged, such as controlling pastoralists or smallholder agriculturalists. Narratives might be “stabilized,” or held in place, by a combination of alliances between different actors such as governments and environmental activists who might agree on certain environmental facts or framings for policy discussion. Narratives might be destabilized by demonstrating that smallholders might not threaten environments, or by proposing new framings for problems. Some development organizations have argued...
that the word “desertification” itself gives the wrong impression of land being permanently damaged by human action. Instead, a more useful and less discriminating form of policy for vulnerable people might result if it were called “drought-proofing.”

Reframing environmental science might also occur through active consultation of, and participation by, diverse social groups in the formation of environmental facts and norms. Feminist social movements, for example, have been partly responsible for diversifying the understanding and treatment of breast cancer during the 1970s and 1980s. Health activists in the United States have also persuaded scientists conducting research on HIV/AIDS (human immunodeficiency virus/acquired immune deficiency syndrome) to look at ways of enhancing the lives of people living with HIV as well as seeking to prevent its transmission. The generation of knowledge on environmental or health risks, therefore, reflects the social solidarities that perceive it to be important, or that are socially empowered to influence data collection or the objectives of research.

The growing field of science and technology studies (STS) is increasingly influencing the analysis of environmental science and society. STS scholars often study the “boundary work,” or means by which various actors might separate science (or knowledge that is publicly unquestioned) from nonscience (or knowledge that lacks social legitimacy). Boundary work is an important concept that explains how facts or narratives become stabilized, or how experts gain legitimacy. Actor-network theory (ANT) has also been used to analyze how the objectives and assumptions of scientific inquiry attribute agency to nonhuman items (Braun and Whatmore 2010). ANT reveals how scientific explanations can be upheld only if underlying framings of research are maintained, and hence how scientific explanations and models can be transported between different contexts if those same framings are maintained in each location.

**Governing expertise**

The debate about environmental science and society also implies that science gains authority according to the organization or individual that presents it. Much discussion about scientific authority in environmental policy adopts the concept of epistemic communities, which was originally defined to mean networks of scientists who had established authority, and who were also willing to communicate important scientific concerns to policymakers (Haas 1992). Perhaps the best-known example of an epistemic community is the Intergovernmental Panel on Climate Change (IPCC), which has published regular synthesis reports of climate science in order to explain priorities to policymakers since 1990.

STS scholars, however, have made various criticisms of this model of expertise. First, expert knowledge is rarely communicated without the influence of local experiences and cultural norms. Knowledge therefore does not travel cleanly to new locations but is consumed and respected according to pre-existing norms and concerns. Second, risk assessments and official communications are sites that build or destroy trust between citizens and officials. For example, research on government advice following the radioactive fallout from Chernobyl during the 1980s made simplistic assumptions about which people and activities were at risk, and resulted in local people rejecting formal advice. Other examples include the attempts of governments to reassure or warn citizens about mad-cow disease, HIV, or wearing car seat belts (Wynne 1996).

Third, epistemic communities themselves undertake various acts of boundary work to
define which problems are to be assessed, and by which kinds of expert. The IPCC, for example, generally defines risks arising from anthropocentric climate change as resulting from each unit of additional atmospheric greenhouse gas concentrations. It also seeks to define these risks by speaking with one voice through official summaries partly negotiated by national political representatives. Critics have suggested that this form of expertise tends to set itself up for rejection by reducing the range of risks experienced under climate change, and by its manner of communicating to diverse social actors (Hulme 2009).

And, fourth, the criteria for authoritative epistemic communities vary between different contexts. It is sometimes difficult to distinguish the epistemic authority of different organizations. In the case of genetically modified organisms, for example, there is no equivalent of the IPCC, and much public debate has been influenced—rightly or wrongly—by newspaper reports or non-governmental organizations. Some analysts have proposed the concept of *civic epistemologies* to indicate the specific national contexts by which knowledge or expertise is considered authoritative. In the United Kingdom, for example, debates about genetically modified organisms have been stabilized more through the influence of nominated experts and judiciary bodies than in the United States, where policy science and litigation have more influence (Jasanoff 2005).

Questioning the epistemic authority of expert organizations does not necessarily mean denying the existence of an environmental problem. Rather, it seeks to draw attention to which questions are being asked, or not asked, and which social outcomes are being achieved, or not. Pointing out how the IPCC, for example, defines climate risk, or seeks to define certainty, does not deny climate change but instead aims to improve how complex phenomena such as risks arising from climate change can be understood and to make scientific knowledge more relevant to a wider set of people. For example, the Research Program on Climate Change, Agriculture and Food Security (CCAFS) of the Consultative Group on International Agricultural Research (CGIAR) is devising ways of implementing climate change policy among low-income small-holders by integrating climate change mitigation and adaptation. Instead of framing climate change policy only in terms of reducing the risks posed by additional greenhouse gases, this network is trying to reduce emissions simultaneously with increasing food supply and social resilience.

A more decentralized form of expertise is *citizen science*, which seeks to harness the knowledge and concerns of local citizens. Citizen science also offers the opportunity to build social trust in risk assessments, as well as flexibility in identifying problems. In developing societies, this form of analysis has been used under frameworks such as *community-based adaptation to climate change*, which seeks to develop participatory responses to new and unexpected environmental risks in ways that address local concerns and knowledge. In other locations, forms of citizen science allow local people to feel included in the framing and implementation of research about environmental problems such as changes in wetlands, bird populations, or pollution. By being involved, citizens also are more likely to trust scientific expertise and to participate in resulting policies.

**SEE ALSO:** Actor-network theory; Climate and societal impacts; Desertification; Discourse; Environmental change and social learning; Environmental governance; Environmental issues and public understanding; Environmental knowledges and expertise; Environmental risk analysis; Environmental uncertainty; Environmentality and green governmentality; Feminist political ecology; Global
environmental change: human dimensions; Indigenous knowledge; Indigenous technical knowledge; Intergovernmental Panel on Climate Change (IPCC); Land systems science; Participatory development; Participatory geographies; Political ecology

References


Environmental science

Sally Eden
University of Hull, UK

Environmental science is a wide field of research that examines all aspects of how environments develop, change, and are impacted by human activity. This includes work on the basic principles of environmental processes and change, considering both organic and inorganic elements of the natural environment. It also includes more applied work and research into policy-relevant environmental issues, such as climate change and pollution, as well as human–environment research into how human activity impacts on the environment and how this impact might be ameliorated.

Origins of environmental science

The origins of environmental science lie in the development of ecology, geology, and other sciences in the eighteenth and nineteenth centuries. At first they were driven by amateur naturalism in the “Arcadian tradition” (Worster 1994), where catching and classifying beetles or gathering fossils was enjoyed as a hobby by wealthy gentlemen (and some ladies) of leisure.

But scientists increasingly sought universal laws to explain how the environment worked, not merely to understand and enjoy nature, but also in order to use, manage, and exploit natural resources, to dominate and control environmental processes. Rather than looking in wonder, this approach often regarded environments pragmatically in terms of machine-like processes that, given sufficient knowledge, could be more effectively managed and exploited in the interests of industrialization and European imperialism in the nineteenth century (Bowler 1993; Worster 1994).

This mechanistic view of the environment supported the separation of processes for analysis, so that different aspects of Earth surface functions and ecosystems became the focus of different scientific disciplines: geology, ecology, geomorphology, atmospheric meteorology, and physical geography all developed their own disciplinary perspectives on aspects of environmental processes and change through the nineteenth century and into the twentieth century. These early efforts at understanding environments often focused on how they functioned in terms of natural cycles of carbon, water, and other nutrients, food chains/webs, weather patterns, stream evolution, and positioning of the continents. But the disciplinary diversification that facilitated such specialization later proved problematic when the search for more integrated and holistic understanding to tackle environmental problems in the twentieth century was hampered by the disciplinary boundaries that were by then entrenched.

Against the rising power of human technology and the dominant Enlightenment ideal of progress, more and more commentators began to argue that environments were being damaged by human exploitation. A key geographical work, Man and Nature by George Perkins Marsh, was published in 1864 in the United States and is sometimes cited as the first “modern” expression of such environmental concern. Marsh’s book
was pioneering in challenging the common assumption that environmental resources were inexhaustible and the environment was resilient to human impact, but the agenda it set was not pursued until the twentieth century.

Environmental science and environmental concerns

It was only in the mid-twentieth century that environmental science really began to address what we now think of as “environmental issues,” prompted especially by Rachel Carson’s (1962) *Silent Spring*, which used scientific evidence to argue that pesticide use was taking a huge toll on bird populations. The rise of ecology as a scientific discipline in the twentieth century, as well as computer modeling of environmental futures, also contributed to the rise in environmental concern and attempts to protect environments from further harm. Both *Silent Spring* and *The Population Bomb* (Ehrlich 1968) were politically influential books written by ecologists that prompted both high-profile debate over and more research into human–environmental futures. The environmental movement thus both contributed to and benefited from the expansion of environmental science since the 1960s, with environmentalism serving as “a midwife” in legitimating new scientific agendas and providing an audience eager to hear their results (Hart and Victor 1993).

So, by the 1960s, the label “environmental science” had come to mean scientific research into basic environmental processes, but also more applied research that sought to inform environmental protection by improving our understanding of how environments work and how they are changing. Geographers and historians of science have thus shown that environmental science is not neutral, but was shaped by the rise of the environmental movement, as well as by policy demands for environmental knowledge for decision-making. For example, scientific work on the carbon cycle in the 1960s was not connected with climate change until the 1970s, when this had become an issue (Hart and Victor 1993). In turn, emerging scientific results have shaped public awareness of, and policy agendas to tackle, environmental problems such as global climate change (Taylor and Buttel 1992).

Some environmental scientists embraced this policy relevance and usefulness; for example, British ecologists in the 1940s attempted to define nature “conservation” as their own scientific and professional remit, and thus to exclude amateur naturalists who were working instead on nature “preservation” (Bocking 1993). But others sought to maintain their scientific credentials as separate from policy demands; for example, the Ecological Society of America sought to present ecology “as an objective, value-free science,” rather than a political movement (Kinchy and Kleinman 2003, 891).

Geography as a discipline missed many of the benefits from the “rising tide” of environmental concern in the 1960s that underpinned the expansion of environmental science elsewhere, especially in ecology. This is remarkable given the potential of geography to combine insights from physical and human geography to advance research into the human–environment interactions. At the time, the discipline was undergoing divisions between physical geography and human geography, with the latter increasingly influenced by humanism and Marxism into the 1970s, with little attention paid to environments, and experiencing the rise of spatial science in physical geography in particular. By the 1990s, however, geography had begun to develop this area, so that “environmental geography” was recognized and defined as “the interface” between physical and human geography by the
United Kingdom’s Subject Benchmark Statement in 2000, which defined the discipline in higher education nationally.

The range of environmental science today

Today, “environmental science” is a widely recognized label for both basic and applied research into environmental processes and problems and encompasses a vast field of topics. For example, the subject category of “Environmental Sciences” in Thompson Reuters ISI Citation Reports included 223 journal titles (as of October 2015) including  

- Aerosol and Air Quality Research
- Arctic, Carbon Management, Conservation Biology, Ecotoxicology, Energy Policy

Currently topping the ranking in this category (as measured by Impact Factor) is Energy & Environmental Science, launched by the Royal Society of Chemistry in 2008, which is emblematic of the range of environmental science today, describing itself as a “forum for researchers investigating a more energy-sustainable planet.” The journal editors emphasize collaboration and cross-fertilization between disciplines such as chemistry, engineering, and global atmospheric science (Energy & Environmental Science 2015, 14). This reflects both the breadth of environmental science today and also its emphasis on addressing environmental problems through issue-oriented research and innovation.

Environmental science thus contributes to practical measures for redesigning environments through restoration, remediation, rehabilitation, reclamation, and species reintroductions. Bodies such as the Intergovernmental Panel on Climate Change have been established precisely in order to identify the best scientific knowledge (as defined through peer review) on the processes in, impacts of, and options for mitigating climate change, and to communicate this knowledge to global environmental policymakers under the auspices of the United Nations.

Is environmental science multidisciplinary, interdisciplinary, or transdisciplinary?

This breadth of scope means that, to address environmental issues fully, environmental science needs to draw on knowledge traditionally claimed by different disciplines and integrate diverse expertise. For example, to research ways to reduce waste pollution on land, a research team may need expertise in microbiology, to analyze how mixed waste decomposes after dumping, in hydrology, to understand how polluted waters from dumps move into and through groundwater systems, in engineering, to explore ways to tap gas released from the dump for energy recovery, and in economics, to analyze the costs and benefits of different technologies to address this.

The need for such integrated expertise has caused environmental science to be labeled multidisciplinary, interdisciplinary, or transdisciplinary by different commentators.

“Multidisciplinary” means that researchers work collaborate by pulling together knowledge and methods from different disciplines, and sharing expertise to address problems that
ENVIRONMENTAL SCIENCE

involve more than one discipline. For example, the journal *Science of the Total Environment* defines the total environment as the atmosphere, hydrosphere, biosphere, lithosphere, and anthroposphere, and only publishes “multidisciplinary” work which focuses on more than one of these spheres. The journal lists 19 relevant topic areas for articles, including agriculture, air pollution, ecosystem services, ecotoxicology, environmental management, groundwater hydrogeochemistry, human health, nanomaterials, noise, remote sensing, marine ecology, and waste and water treatment (*Science of the Total Environment* 2016). This again suggests the great breadth of environmental science today and that multidisciplinary collaboration can be as varied as engineering working with hydrology to address water management, or atmospheric science working with medicine to monitor the spread of air pollution and its effects on human health.

“Interdisciplinary” goes further and means that researchers from different disciplines work together by developing new knowledge and methods; that is, they do not simply combine expertise from multiple disciplines (as in multidisciplinary research), but collaborate more creatively to innovate and develop new ways of addressing problems that exist between and across disciplines. “Transdisciplinary” means a yet further move to a holistic approach to researching environmental problems that borrows freely from other disciplines, but works across disciplinary boundaries without being confined by disciplinary agendas.

These three terms are often used interchangeably and it is difficult to pin down the differences precisely, because they often function mainly as normative ideals of how research such as environmental science should progress, rather than as mutually exclusive descriptions. Moreover, achieving interdisciplinary and transdisciplinary research in practice remains difficult, because traditional disciplinary boundaries are still used to organize academic departments, training and recruitment of researchers into academia, and research funding and publishing regimes, for example, through the remits of funding bodies and peer-reviewed journals. This means that work that aims to go beyond disciplinary agendas is often pulled back into them when it comes to academic publishing or teaching, although it may have more success influencing practical work outside academia, which is less bound by such disciplinary “silos.”

And there is a further challenge: many environmental issues require analysis not only of myriad organic and inorganic environmental processes, but also of complex environment–society interactions, for example, for climate adaptation modeling. This requires the environmental natural/physical sciences to integrate with the social sciences addressing those same problems from very different perspectives, but is hampered especially because of deep divides in methodological expectations. Environmental social science has come to value qualitative methods recently, because they are seen as more able than quantitative methods to get beyond simple, post hoc rationalizations of environmental actions to reach more contextual, socially embedded reasons, especially to expose social norms and routinized practices. However, qualitative methods are often seen by environmental natural/physical scientists as woolly, anecdotal, and lacking in scientific rigor, thus hampering interdisciplinary working.

Some have argued that because of these problems of integration, what is needed is not merely environmental science but “sustainability science” (see Sustainability science), taking a transdisciplinary approach to fully integrating research from the natural, biological, social, and engineering sciences in order to use findings directly to shape policies and actions for sustainability.
Criticisms of environmental science

Although often presented as neutral, objective, and value-free, environmental science has also been shown to be irreducibly social and political (e.g., Taylor and Buttel 1992; Demeritt 2001). For example, analyses by social scientists of global warming science and climate change modeling suggest that scientific research is defined not solely by scientists but also through their interactions with other interests, especially policymaking (e.g., Shackley and Wynne 1995). But acknowledging this risks providing ammunition to those opposed to mitigating climate change, who then claim that scientific results are distorted and corrupted by political pressure or personal interests.

Environmental science has also been criticized for not being able to address the complex environmental problems of today, such as biodiversity loss and climate change, problems which have high uncertainty in terms of what we know, but also high risks in terms of their potential impact on human activities. Because of this uncertain but risky character, some argue that these problems require a new kind of science, a science that goes beyond the traditional approach of providing objective, certain knowledge to policymakers, a “post-normal science” (Funtowicz and Ravetz 1993; Ravetz 1999) that draws on scientific knowledge but also incorporates other ways of knowing, thus democratizing participation in decision-making through involving the “extended peer community” outside professionally accredited scientists.

This offers a normative analysis of what science should become, in order to tackle complex environmental problems where “facts are uncertain, values in dispute, stakes high, and decisions urgent” (Ravetz 1999, 649). This exhortation argues that, where there are few “hard,” objective scientific facts available to make decisions, “soft” subjective value-judgements become more important. But some scientists see this as threatening their scientific authority and professionalism by allowing nonscientists to be accepted as valid contributors of knowledge on the basis of experience or their position as a stakeholder affected by environmental change, rather than on the basis of research or training.

SEE ALSO: Environmental knowledges and expertise; Environmental policy; Environmental risk analysis; Environmental science and society; Environmental uncertainty; Environmentalism; Global climate models; Global environmental change: human dimensions; Intergovernmental Panel on Climate Change (IPCC); Sustainability science

References


Environmental uncertainty

Simon J. Shackley
University of Edinburgh, UK

Environmental systems are typically complex making them difficult to understand and subject to high levels of variability and uncertainty. Predicting how such systems might change in the future is a precarious task, a classic example being the difficulty of weather forecasting beyond a week ahead due to the chaotic properties of weather (requiring great precision in establishing initial conditions, which in practice is currently not possible).

Types of environmental uncertainty

Understanding of scientific uncertainty has been revolutionized in the past half century through mathematical advances in understandings of nonlinear dynamic processes and systems, variously known as chaos, complexity, and indeterminacy (Smith 2007). But there remains vigorous debate about how best to analyze environmental uncertainty. Some scientists maintain belief in the capacity of classical scientific experiments, whereby underpinning mechanisms and processes are neatly parcelled up and investigated one by one through establishing dependent and independent variables. Those working in the complexity paradigm, meanwhile, point to the prevalence of interactions between multiple, simultaneous mechanisms/processes, requiring extensive use of surrogate and composite variables, parameterizations, analogs, and more holistic, system-wide investigation, involving inter alia statistics and simulation modeling tools but also recognition of the value of qualitative methods.

Uncertainty is often described in terms of “risk” but environmental uncertainty differs from risk in that risk is quantifiable in a probabilistic sense but uncertainty is not, because the probability of an event happening and the magnitude of its effects are unknown and therefore not quantifiable. An example of risk is a 1:10 chance of a flood happening. But in many real-world settings, such a flood risk is based on historical records, which are usually insufficiently long or accurate to generate quantitative predictions; furthermore, it assumes that other parameters influencing flooding are either not changing or are changing in known ways. Predicting flooding is therefore more about uncertainty than about risk.

As well as uncertainty, there is also indeterminacy (Wynne 1992), whereby further uncertainties are introduced when people and institutions attempt to apply knowledge of (and in) complex environmental systems. Indeterminacy usually arises where contingent sociotechnical practices and behavioral and collective responses brush up against the presumptions of scientific and engineering analysis. It is assumed, for instance, that agricultural workers who apply pesticides carefully read and understand the health, safety, and environmental instructions for appropriate application; hence the manufacturer and regulator can feel confident that due diligence is undertaken. Yet, the realities of real-world application of pesticides soon reveal a different story, one that requires regulators and
Environmental uncertainty

scientists to appreciate the sensitivity of their risk assessments to questionable assumptions and to perhaps ask a different kind of question – not so much “how safe is chemical X?” but rather “under realistic use conditions, what is the risk of overexposure by the operator, how certain is our understanding of the harmful impacts of chemical X, and is that risk acceptable to users and operators?”

As well as uncertainty, risk, and indeterminacy, understanding and predicting environmental change also involves ignorance, where processes, mechanisms, interactions, and feedbacks that are relevant to a given system are simply not known about at a given time (they are “unknown unknowns”). An example in climate modeling is the realization over the past 20 years that human-induced biomass burning (associated with deforestation and agriculture but also cooking) has impacts upon global climate change through the production of black and organic carbon particles. An example in the social sciences is the unexpectedly rapid pace of economic change in China over the past two decades. Even in 2000, the experts on greenhouse gas emission scenarios for the Intergovernmental Panel on Climate Change (IPCC) so underpredicted China’s growth that the IPCC’s highest emissions scenario was outpaced by real events in less than a decade.

Environmental uncertainty as a problem, resource, or construction?

Environmental uncertainty is still widely regarded as a “problem” by policymakers and by many scientists alike, the prevailing assumption being that reducing uncertainty will illuminate the correct policy pathway. This so-called linear model ties in closely with modernist assumptions of progress which have proven so controversial amongst social scientists. Yet, scientists also know only too well that science-in-practice generates new uncertainties at the same time as answering existing questions and, occasionally, paradigm shifts change the set of questions entirely. Environmental uncertainty is also used by researchers and other stakeholders to argue for more expenditure on research and development, and in policy debates to argue against implementing a policy or technology (e.g., via the Precautionary Principle, where the onus is on avoiding environmental harm, but also by climate change sceptics who continue a long tradition of proposing “more research” before major policy interventions are undertaken). Environmental uncertainty is rarely “neutral” but becomes embroiled in – and fuels – political and policy debates and controversies.

Douglas and Wildavsky (1982) reframed the entire domain of environmental risk and uncertainty as being far more about cultural and social constructions of what is regarded as an acceptable risk than about merely quantifying risk; the role of technical knowledge in that social process varies but is frequently minimal while the role of values is given greater prominence. Because of technical uncertainty, the same corpus of knowledge can readily support a number of distinctly different positions on complex environmental issues. Cultural theorists have therefore argued that values and associated commitments and preferences come to shape the way that uncertain knowledge is constructed and its consequences articulated. Some advocate a specific grid-group theory to account for the mutual reinforcement of patterns of social interaction and social construction of risk and uncertainty (e.g., Verweij et al. 2006). But this has proven controversial, in part because scholars have been unable to agree on a single theory of social construction, resulting in widely diverging
and incommensurable understandings of how environmental uncertainties are constructed.

**Coproduction of environmental knowledge and policy**

Underpinning notions of uncertainty is the *conditionality* of scientific knowledge. While scientific knowledge claims can be validated to some extent though use of controlled (usually laboratory) conditions, beyond such confines claims to knowledge are harder to validate. A classic example is the long-term fate of radioactive fallout from the Chernobyl nuclear power plant accident of 1986 within upland peaty soils in northern England (Wynne 1998). The soil scientists used models of soil dynamics which failed to take account of the particularities of peaty soils and of sheep grazing practices. This inappropriately extended the scope of their existing numerical tools into untested domains, including socioecological practices that were not incorporated into their soil models. Similar examples can be provided for many other types of models used in geography, for example in studying climate change, hydrology, geology, and infectious diseases.

In understanding why such extension work happens, scholars have developed the concept of the coproduction of science for policy. “Science” provides answers to technical questions faced by policymakers, while policy decisions help scientists by articulating technical questions, helping to bring closure on what issues they should address and how. Uncertainty for both science and policy is reduced by this coproduction – a self-reinforcing cycle which only comes unstuck when it bumps up against some disruptive event. A good example is the risk assessment surrounding “mad cow disease” (bovine spongiform encephalitis, or BSE) in the United Kingdom, whereby key scientific questions faced by the scientific advisory committee established to advise the government on the risks of BSE were answered in part on the basis of the implicit preferences of senior policymakers. In the pre-1996 period, this advice was that the risk of the infective agent responsible for BSE moving from cows to humans (hence the risk of eating BSE-infected meat) was negligible; that advice changed in March 1996 when the UK government stated that there was firm evidence that the infective agent causing BSE had moved from cows to humans. Politicians subsequently explained to the public that their pre-March 1996 decisions were taken on the advice of the scientific experts, ignoring or obscuring the political decisions in choosing which questions were (and were not) addressed (and how) in the first place (Van Zwanenberg and Millstone 2005).

**Responses to environmental uncertainty**

It has been proposed that policymakers shift focus away from “predicting” or “anticipating” the future because of endemic uncertainty (Collingridge 1980). Even if it is dressed up as “analysis of future scenarios,” the temptation to select more or less “likely” scenarios tends to prove irresistible to scientists and policymakers alike. An alternative is a “fallibilistic” approach (Collingridge 1980), based upon prudent “trial and error,” minimizing the error costs (that is the cost of an incorrect policy or business decision) and maximizing the rapidity and quality of feedback on the impacts of an intervention. The fallibilistic approach tends to encourage more flexibility and changeable options and responses, frequently implying smaller-scale interventions, at least initially until the complexity of the system and its sensitivities are better understood. This is because scale effects, nonlinearities, and
indeterminacies always limit the ability to predict the impact of an intervention going forwards.

Views based on the construction or the coproduction of environmental uncertainty have prompted arguments for greater transparency and openness in how values and environmental facts become intertwined. New institutions and processes for bringing together different perspectives to create policies which appeal across distinct value clusters have also been proposed. For example, participatory, dialogue, or deliberative democracy approaches are now widely researched and practiced in areas of environmental uncertainty. In development studies, uncertainty in the socioeconomic and cultural context of rapidly changing economic, environmental, and socioecological conditions has been better understood by using more qualitative, dialogue-focused methods such as interviews, focus groups, and ethnography – as opposed to traditional surveying. Still, the relative undercapacity of research institutions in many developing countries has hindered attempts to understand and interpret rapidly changing societies and their relationship with the environment, encouraging an opportunistic or “scattergun” approach by governments, nongovernmental organizations (NGOs), and donor organizations aiming to encourage more sustainable development pathways. All too frequently, interventions are poorly designed, are not integrated with other ongoing programs and developments (including commercially focused projects), and the outcomes are not properly or independently recorded. Nonetheless, new technologies for collecting and sharing information may assist the further development of progressive mutual learning participatory methods, in turn providing a better understanding of the uncertainty space within which interventions are proposed.

In industrialized countries, experience of participatory approaches with publics has revealed a greater awareness of the conditionality of scientific knowledge amongst the public than was perhaps initially presumed by the scientific community. Some skepticism of the motivation of scientists and engineers, of their funding sources and suspected “hidden agendas” has emerged during focus and in-depth discussion groups, while many participants have also found face-to-face interactions with practicing scientists to be highly rewarding due to the openness and transparency with which scientists tend to discuss uncertainty in such an intimate context. The problem is in scaling up such experiences beyond the very small samples in a focus group and also in embedding the outcomes of such processes in real-world policymaking, thinking, and practice. In reality, many senior policymakers are resistant to the premise behind deliberative democracy, limiting the potential opportunity for uncertainty to be viewed as a constructive source of dialogue and negotiation rather than as a “problem” that needs to be “fixed.”

SEE ALSO: Environmental risk analysis; Environmental science; Environmental science and society; Participatory modeling; Visualizing uncertainty

References


Environmental valuation includes different methods for estimating the economic values associated with ecosystems. Monetary values for ecosystem “goods” (such as clean air, wildlife, biodiversity, and scenic views) and “services” (such as the carbon sequestered by forests or the habitat provided by wetlands) are, however, difficult to calculate because such goods and services are not normally traded on the market. There is, thus, much contention over the best method for deriving monetary values and debate about whether ecosystems can be responsibly represented by prices.

Origins of environmental valuation

Environmental valuation originates in the field of economics where there have been wide-ranging attempts to assign monetary values to ecosystems since the late 1940s. However, it was not until the 1990s that systematic valuation frameworks were devised in the discipline. Such valuation frameworks allowed the practice to be taken up outside of the academy by policymakers and public agencies. In the absence of actual markets for environmental goods and services that would determine prices based on demand and supply, economists now use two main approaches in order to arrive at monetary values. The first approach encompasses stated preference methods, including the most commonly used contingent valuation method in which respondents are surveyed about their willingness to pay for environmental improvements or to avoid environmental loss. For example, respondents might be surveyed about how much they are willing to pay for a specified reduction in air pollution or to conserve a particular species. The second approach encompasses revealed preference methods, whereby environmental prices are calculated using market information for similar market-traded goods and observed human behavior. In hedonic pricing, for example, holding constant other possible variables, the difference between the price of a house with a mountain view and one without is approximated as the environmental value of the view itself.

While the term “environmental valuation” most commonly refers to the practice of assigning monetary values to nature, the question of how nature should be understood, treated, and valued (i.e., what significance and importance it should be accorded) has been discussed by multiple disciplines through varied approaches. From Marxist perspectives, authors expose the poverty of capitalist economies that can only value nature in terms of its ability to circulate on markets, thus negating the cultural and moral values associated with all economic activity (see Eaton 2011). Other authors suggest that nature can and should be valued through nonmonetary approaches that focus on more pluralistic conceptions that account for cultural, moral, and ethical understandings of nature. For example, O’Neill, Holland, and Light (2008, 200) propose
that environments should not be understood as “mere bundles of resources” but, rather, as “places to which humans have a lived relation of struggle, wonder and dwelling.” For these authors there is no single value (such as dollars or pleasure) to which all others are reducible. There is thus no credibility to the idea of trading off environmental values, because value is specific to time, place, and positionality. Instead, these authors call for pluralistic conceptions of environmental values that are rooted in everyday encounters between humans and nonhumans, and privilege history and narrative in the valuation of the environment. Such an approach requires policymaking practices that see decisions not as discrete events but as historical patterns of choice that express the character of the institutions that make them. Certainly conceptions of value in relation to the environment are wide and diverse in geography and many have arisen precisely as criticisms of the process that has come to be known as environmental valuation.

Once environmental goods and services have been assigned monetary values, they can be incorporated into resource management and policy decisions, and traded on markets as commensurable quantitative amounts. For example, lost habitat, increased air pollution, and aesthetic changes are rendered as quantitative sums that allow them to be compared to the projected economic benefits of a proposed industrial project in a cost–benefit analysis. In the case of externalized costs associated with environmental degradation, policymakers can transfer the costs associated with a particular activity to the perpetrator through legislation, taxation, or levies. For example, the environmental costs associated with absorbing the waste products of a particular commodity can be internalized into the price of the good to be paid by consumers, or into the price of the inputs to be paid by producers. Payments for maintaining ecosystem services such as habitat can also be made by states wishing to promote environmental stewardship to private citizens and other organizations that manage private resources. Farmers, for example, have benefited from such payments to conserve habitat on their lands. Finally, environmental valuation enables practices of compensation for environmental losses and the offsetting of environmental impact in one location through investment in environmental goods and services in another. Once a price is determined for a wild lake bearing fish, for example, an industrial developer can be made to provide compensation to a local community for the impact of its development on the lake, or it can be asked to offset the environmental damage it will cause by investing in “equivalent” environmental goods and services in another location.

While academics have been developing methodologies of environmental valuation since the 1940s, its application in public policy and private markets has been more recent and limited. Environmental valuation techniques have been used in damage assessment and resource compensation in the United States, and in valuing the benefits to human health from changes in environmental quality in the United States, Canada, and Europe. Environmental valuation has also been used in cost–benefit analyses of water and forest resource planning and for state-run payments for ecosystem services (PES) programs, whereby farmers or other stewards are paid to preserve an ecological function like species habitat.

More recently, environmental valuation has been increasingly used to value ecosystem services with the aim of incorporating them into various programs of trading and offsetting. At the international level, since the 2000s the United Nations Environment Programme, the G7 environmental ministers, the Global Environmental
Facility, and the European Union have been involved in several initiatives to assess and document the costs and benefits of biodiversity. This international assessment of environmental benefits and costs is informed by multidisciplinary debates that surrounded Costanza et al.’s (1997) valuation of the entire Earth (its “natural capital” and environmental services) at $33 trillion (Figure 1). Especially contentious is the accuracy of calculating a total value of the Earth rather than a marginal value of a unit of nature. Further debate surrounds the best practices for the valuation, including the standardized classification of environments.

At national scales, trading and offsetting have been explored by Hillman and Instone (2010) in their analysis of Australia’s NSW Threatened Species Conservation Amendment (Biodiversity Banking) Act 2006, which allows landholders to establish values for the biodiversity of a site and offer the values as credits that developers can purchase in order to offset biodiversity losses caused by their actions elsewhere. Robertson (2007) has examined the US Environmental Protection Agency and the US Department of Agriculture’s creation of water quality and wetland credits. In these banking and offsetting schemes, environmental goods and services are brought into markets; their prices are, thus, forged through the practice of market trading rather than assigned on the basis of stated or revealed preferences or intrinsic biophysical worth.
Engagement in human geography

Human geographers have engaged with the practice of environmental valuation through both application and criticism. Some authors have adopted it as a mechanism through which to incorporate a concern for the environment into policymaking. For example, looking at China’s housing market, Jim and Chen (2007) compared stated and revealed valuation techniques in a study of environmental externalities, such as access to green space and views of water and parks in Guangzhou. They argue that municipalities could mobilize the positive economic values accorded to environmental attributes in order to encourage urban green spaces and nature conservation in the development of housing. Heidkamp (2008) argues that geographic factors such as scale and proximity should be incorporated into environmental valuation techniques, and thus into cost–benefit analyses. According to Heidkamp, spatial proximity should be factored into the relative scarcity of environmental goods and services. For example, trees may be plentiful globally, but their being scarce locally might make for a higher price in a particular market. Moreover, transaction and interaction costs increase over distance, which means that, in addition to temporal discount rates, spatial discount rates ought to be applied in valuation methods. Heidkamp suggests that spatial variables such as scale and proximity be integrated into cost–benefit analysis through the use of geographic information systems (GIS).

Geographers have also been investigating environmental valuation through the use of qualitative methods. In their study of a contingent valuation survey used by the Pevensey Levels Wildlife Enhancement Scheme in the United Kingdom, Burgess, Clark, and Harrison (2000) found that respondents’ answers to the survey questions were very context–dependent because respondents came to the surveys with different knowledge and experience of the environment they were asked to value. They also found that the presentation of information about the good that was to be valued significantly affected, indeed framed, the respondents’ willingness to pay. Finally, Burgess, Clark, and Harrison reported that once respondents had engaged in their in–depth discussion groups and learned more about contingent valuation, they felt deceived about the purpose of the survey and were resistant to the idea that a monetary value could even be associated with nature. In a second study Clark, Burgess, and Harrison (2000) used qualitative methods to assess whether the willingness–to–pay figures that survey respondents provided were valid. They found that respondents called into question both the bases and the relevance of their own willingness–to–pay figures when they were engaged through in–depth discussion after having taken the survey. These studies suggest that contingent valuation methods don’t simply reveal underlying and unchanging values, but rather that environmental values are constructed through interactions, that they are changing, and that they are specific to contexts like place and experience.

Criticisms of the practice of valuation can be grouped in two main categories. First, geographers have challenged the methods of valuation and questioned the commensurability of different bits of nature across space. Much of this criticism is focused on revealed and stated preference models for deriving prices for nature. As critics point out, such methods are focused on individuals’ willingness to pay, which says little about the value of the piece of nature in terms of its ecological, cultural, social, or even economic functions. In other words, just because individual consumers are willing to pay only a small amount to conserve water quality does not mean that water quality is of little value to a local economy.
When preference-based approaches are used for deriving value, only those bits of nature for which individuals are willing to pay receive investment for conservation. However, other methods for valuation exist. For example, Costanza et al.’s (1997) valuation of the Earth’s ecosystem services used a benefits transfer approach that priced the benefits of natural ecosystems to society by transferring calculations based on willingness to pay from one particular study region to various other regions and contexts. This allows the calculation of value regardless of whether there is a willing buyer of such services, and recognizes that ecosystems do not stop transferring valuable benefits to society if there are no willing buyers. However, at root, this approach is still based on calculations of willingness to pay.

Techniques rooted in biophysical approaches provide a quite different method for valuation. According to the Economics of Ecosystems and Biodiversity (TEEB) (Kumar 2010), a prominent global initiative focused on the economic benefits of biodiversity, biophysical approaches can be used to measure the value that flows from underlying physical parameters. For the TEEB, biophysical indicators such as diversity, quantity, condition, and pressures facing ecosystems should be the main factors in valuation. In the case of timber production, for example, dry matter productivity can be modeled and mapped using remote sensing imagery and meteorological data to measure the extent and quantity of dry matter. Here monetary values are attached to biophysical characteristics rather than to human preference.

Geographers have also questioned whether it is even possible in principle, and therefore desirable, to reduce complex ecosystems or bits of nature to quantitative prices and commensurable units. According to Robertson (2011), in order for ecosystem services to be incorporated into markets, and thus to circulate as value, a variety of “socially-necessary” abstractions must occur. For example, wetlands have to be classified into types that are described by relatively static and governable qualities, but these types can’t capture the kind of dynamism of natural landscapes whose qualities are constantly changing over time and space. Moreover categories of value related to ecosystem functions are created that can’t fully express the complexity of wetland functions. In other words, the removal of wetlands in one location may not be ecologically equal to the preservation of the same quantity of wetland in another location, even if they are classified and categorized as equals, and circulate on markets as the same quantitative sum of money.

Another example of the reduction and simplification processes at the heart of environmental valuation is the ways in which the removal of different parts of a particular habitat may have differential effects that valuation techniques will miss. For instance, if one hectare of wetland habitat is removed from a larger mass, the impacts will be less than the removal of one hectare of the same habitat that serves as a small connecting patch to a wider network. The removal of the latter endangers species capacity for migration and movement. Valuation techniques could be modified to explicitly value the connectivity function of habitat, but it is unlikely that valuation can ever account for all the various different functions that nature and ecosystems provide. At the very least, such functions have to be known, understood, and capable of fitting into categories constructed by the humans doing the valuation. Not only will some environmental functions be missed, but others may not fit into valuation frameworks at all. Noninstrumental values, that is, values that are not useful for other (often human) purposes, such as filtering water or pest control, are the most difficult subjects of valuation. Such values are often called intrinsic values because they are ends in themselves rather
ENVIRONMENTAL VALUATION

than useful functions. For example, it is difficult to figure out how a cultural attachment to place can be captured by valuation. Indeed, the conceptualization of ecosystems as bundles of discrete functions itself is a socially constructed and instrumentalist imposition that belies the interconnected and complex ways in which ecosystems work.

Typically, geographers have understood the process of valuation as a practice of social construction rather than as the quantification of objective biophysical processes and traits. Managers, scientists, and economists must first name, describe, and explain the functions and worth of bits of nature before they can be classified and categorized discretely and assigned monetary values. In this line of critique, it is not just that humans might miss the full complexity of ecosystems in their classification schemes and categorization projects, but that nature can only be understood through limited human conceptions that are constituted by prevailing social, economic, and political discourses. Such a social constructionist view of valuation recognizes that values are not given or objectively measured; they are, thus, open to contestation and prone to becoming sites of resistance over political–ecological relations.

The social practice of environmental valuation has also been criticized as Eurocentric because the simplifying practices outlined earlier can be understood as privileging certain knowledges, peoples, values, and landscapes over marginalized others. Gibbs (2010) examines values associated with water in central Australia and finds that resource managers’ reliance on environmental valuation practices marginalize Aboriginal values like the variability of water regimes and the interconnections between water and the rest of the world. According to Gibbs, concepts of diversity, change, complexity, and interconnection rather than separation, compartmentalization, and domination should be at the center of resource managers’ valuation practices. Both the spatial and the temporal variability of water regimes contribute to the diversity of values attached to these variable conditions. For example, pastoralists are dependent on the growth that is stimulated by the episodic flooding of floodplains. Similarly, patterns of rainfall and river flow affect how water resources are used; waterholes may be present for only three months of the year or they may be permanent features of the landscape. Even though economic valuation practices have attempted to price a diversity of values, such as recreation value, aesthetic value, and Aboriginal heritage value, among others, according to Gibbs, these interpretations are often oversimplified and rooted in Eurocentric assumptions. After all, whose practices of recreation, aesthetic, and Aboriginal heritage are valued, and what does this do to the diversity of understandings and uses of each?

Finally, geographers have been wary about the consequences of representing nature in monetary units; many authors see it as the first step in the commodification of nature and the incorporation of nature into neoliberal and capitalist logics. However, analytical distinction can be maintained between the framing of the environment in economic terms (for example, as costs and benefits, as providing services, or as natural capital), the monetization or pricing of nature, and the commodification of nature. According to advocates of environmental valuation, the assigning of monetary values to nature and its functions can be achieved without the circulation of that nature as a commodity on markets. For example, environmental valuation for the purposes of cost–benefit analyses does not necessitate commodification. In fact, environmental valuation as a practice was developed in order to assign values precisely to those bits of nature that did not have market prices because
they were not exchanged on markets. On the other hand, the economic framing of the environment and the monetization of nature and its functions are, indeed, necessary prerequisites for commodification. As Gomez-Baggethun and Ruiz-Perez (2011) argue, conceiving of human–nature relations as relations of utility and exchange invites the logics of capital and profit-making. Given an economic framing of the environment, decision-makers increasingly make choices within a narrative of scarcity, efficiency, and profit.

Although the economic framing, pricing, and commodification of nature can be held apart analytically, in practice monetary valuation has often been the first step in a process that ends with the incorporation of nature into capitalist processes. Since the rise of ecosystem service valuation in the 1990s, conservation has been practiced more and more through market instruments with the goal of realizing the monetary values calculated by practitioners of environmental valuation. As Robertson and Wainwright (2013) show, the state plays a prominent role in pricing nature and ensuring the expansion of capitalist value. According to these authors, it would be a mistake to prize apart the two processes of pricing nature and profiting off nature because the practice of environmental valuation only makes sense within a system that confronts nature in socially abstract value forms.

SEE ALSO: Biodiversity; Commodification of nature; Conservation and capitalism; Ecosystem services; Green capitalism

References


Environmental determinism

Tom Johnston
University of Lethbridge, Canada

Environmental determinism is an idea that occupied a central position in geography during the early part of the twentieth century. It holds that the physical environment shapes the course of human economic, cultural, and social development, and that this effect is independent of space and time. Many of the core ideas held by environmental determinists can therefore be linked to the theory of uniformitarianism and to Darwin’s conceptualization of evolutionary processes. Clearly, environmental determinism drew inspiration from outside geography and must be situated within several broader intellectual and scientific traditions.

It was an idea that was very much a product of its time. According to at least one prominent chronicler of the history of geographic thought, it represents the first attempt by geography to develop an overarching, unifying philosophical framework and fashion the subject as a science in search of theories and laws. Geography was a relatively new discipline at the time; it had only been formalized institutionally in the late 1880s and was thirsty for a focus that would provide an identity for its practitioners and would garner respect from other scholars. Further, through an explicit characterization of Europeans and North Americans as superior, owing in part to a more bracing climate, it provided a theoretical and moral justification for colonial expansion and exploitation. In addition environmental determinism was used in support of false claims about the existence of “different races” but more insidiously of the presumed moral superiority of some so-called races and of the potential for moral decay faced by members of the “white race” who spend extended periods of time between the tropics of Capricorn and Cancer.

Within geography, environmental determinism is closely associated with the German geographer Friedrich Ratzel (1844–1904), who summarized his ideas about environmental controls in a two-volume offering entitled Antropogeoghie published in 1882 (vol. I) and 1891 (vol. II). Ratzel, who is also credited with codifying the concept of “Lebensraum” (living space), an idea that formed the theoretical foundation for Nazi territorial expansionism, influenced a number of other scholars associated with this school of thought. Most notable among Ratzel’s followers were Americans Ellen Churchill Semple (1863–1932) and Ellsworth Huntington (1876–1947). Thomas Griffith Taylor (1880–1963), whose appointment to the University of Toronto in 1935 marks the formal establishment of the geography department there, was also a prominent voice for environmental determinism.

During an unquestionably important period in the historical development of geographic thought, the advocates of environmental determinism failed to cultivate universal acceptance for their ideas. In his Presidential Address to the 1922 Annual Meeting of the Association of American Geographers, for instance, published a year later in the association’s journal (Barrows 1923), Harlon Barrows advocated a vision of geography in which the aim was
to “to make clear the relationships between natural environments and the distribution and activities of man” and “to view this problem in general from the standpoint of man’s adjustment to environment, rather than from that of environmental influence” (p. 3). Barrows went on to urge geographers “to minimize the danger of assigning to the environmental factors a deterministic influence which they do not exert” (p. 3). Two years after Barrow’s address, Carl Sauer launched a less diplomatic attack on environmental determinism, referring to the search for geographic controls as “facile” and stating that “the geographer has taken the affirmative side in a debate on environment and has therefore failed to maintain the objective quality of the scientist” (Sauer 1924, 18).

In response to mounting criticism, geography witnessed the emergence of “possibilism,” which was really a form of modified determinism. Under possibilism the environment is viewed as offering a range of options from which humans select a preferred one based on cultural differences, available technology, and so forth. As Jones (1956, 369) notes, the options available to humans were “circumscribed” by the environment but not “fattily determined.” Possibilism failed to address adequately the central concerns surrounding environmental determinism and was eventually dismissed for many of the same reasons that environmental determinism was; it too was regarded as fundamentally deterministic, it ignored the role of what later came to be called “human agency” and the importance of locally constructed realities, and with the regional perspective gaining prominence, a non-nomothetic approach, many geographers of that time questioned the legitimacy of human geography searching for general theories and laws in the first place.

In the telling and retelling of the story of environmental determinism, the views of its advocates might have been characterized somewhat more rigidly than may have been the case. For instance, Spate (1952, 407) observed that while Ellsworth Huntington’s determinism was “avowed, and clearly based too much upon a part only of the immense complex of factors that which have gone into the making of history.” He added a qualification: “though it should be added that he is aware of this, and it is his disciples rather than himself who threw discretion to the winds.” And in a review of Semple’s (1911) *Influences of Geographic Environment: On the Basis of Ratzel’s System of Anthropo-Geography* Maret (1913, 282) noted that “Miss Semple possesses in a high degree the saving grace of moderation” and is “fully aware as she is of the fact that anthropogeographers in the past have occasionally brought more zeal than discretion to their task, she does not put her strengths into ventilating the larger and vaguer claims of her subject.”

The acceptance of environmental controls in the context of human development was not limited to geography nor was it limited to the period during which the idea occupied a position of prominence within the discipline. As noted by many commentators, discussion of the link between environment and humans can be traced to Greek scholars and to such works in particular as Aristotle’s *Politics* and *On Air, Water and Places* by Hippocrates. Environmental controls were also advocated by the Arabic scholar Ibn Khaldun (1332–1406), and according to Sommer (2007) a range of post-Medieval scholars, including Machiavelli (1469–1527), Montesquieu (1689–1755), Jean Bodin (1530–1596), and the German philosopher Johann Gottfried Herder (1809–1882) whose ideas about human adaptation to the environment are echoed today in the concept of cultural adaptation. During the modern era, environmental determinism has also enjoyed general acceptance within other disciplines, most
notably in anthropology, where the enthusiasm for environmental determinism may have been more deeply entrenched in theory and certainly lasted longer, but also within history, economics, and political science.

In recent decades, support for environmental circumstance as a nontrivial influence in the course of human activities and development has enjoyed in resurgence of sorts, especially in studies of regional economic development. This thinking can be seen, for instance, in Diamond’s Pulitzer Prize winning *Guns, Germs and Steel*. Such views have not been received with much enthusiasm in some quarters of contemporary geography, especially although not exclusively among critical theorists. And so environmental determinism has not enjoyed any kind of a renaissance within geography.

Once a core theory in geography, environmental determinism is today, much like William Morris Davis’s “Cycle of Erosion,” a subject covered in courses on the history of geographic thought rather than a functioning analytic framework. But geography has a long history of reinventing itself and the study of human–environment relationships remain a central part of the discipline’s focus. We see this interest manifested in many research initiatives, most notably in the contributions geographers continue to make to the study of natural hazards and global environmental change.

**SEE ALSO:** Anthropogeography; Geography and the study of human–environment relations; Human ecology; Human geography

**References**


Grassroots environmentalism is that element of the wider environmental movement that engages in forms of contentious politics to develop knowledges about environmental issues, challenge cases of environmental destruction, and construct alternative visions and examples of ways of living that grassroots environmentalists argue are superior to contemporary unsustainable, unequal or oppressive ways of organizing society. Grassroots environmentalism involves ordinary people and activists who develop forms of contentious politics that they put into action themselves, using the resources and skills they have to-hand and guided by strategies they develop themselves. Grassroots environmentalism involves “contentious politics,” a feeling that something is not right, and that something must be done about it. This may be direct action, creating alternative visions and knowledge, or creating alternative institutions. Grassroots environmentalism does not limit itself to lobbying, to making demands or requests that others – businesses or the state – should take action to change policy. If demands are made, they are backed-up by action undertaken by activists, either to raise issues or to develop alternative models of how society should be organized. It is this focus on activists that makes grassroots environmentalism distinct.

Thus, grassroots environmentalism can be contrasted with other forms of environmentalism, such as conservation and community-based resource management, which do not have that element of contentious politics or the generation of alternatives. A focus on the grassroots excludes “elite activist” environmentalisms, such as that of Greenpeace, where the role of the grassroots is to provide money and/or resources to fund activities aimed at raising media awareness carried out by a small, select group of activists, perhaps in clandestine ways, or to fund lobbying. It can be contrasted with electoral politics through green parties. It can be contrasted with policy activism, aimed at changing government policy through lobbying alone. It must include some element of contentious politics.

Values

Grassroots environmentalism shares a number of value systems with the wider environmentalism, but tends to align itself with more fundamentalist or deep green, or ecocentric values (which see humans as just another species of equal value with other species) than with more pragmatic, cornucopian or human-centered values (which stress human primacy). Grassroots environmentalists share environmentalisms’ concerns about the questionable value of industrialization, centralism, and the big technological projects associated with contemporary capitalism but are also likely to be critical of the large scale, technological or market-based fixes to environmental problems, such as genetic modification, geoengineering or nuclear power, associated with cornucopian or promethean conceptions of...
ENVIRONMENTALISM, GRASSROOTS

environmentalism. Grassroots environmentalists are likely to be skeptical about the power of these “technological fixes” to solve pressing environmental problems, to question whether the resources needed to manufacture them on a large scale are are likely to continue to be available, and are concerned that these innovations are an attempt to maintain an unsustainable economy for one species – humanity – at the expense of others. Grassroots environmentalists are more likely to identify with the Luddites, those early opponents of capitalism who broke machinery that was putting them out of work, than with those who see technology as part of the solution. Some grassroots environmentalists are attracted to the primitivist preindustrial politics of green anarchism, arguing that given that humanity began to degrade its ecosystem when it first embraced settled agriculture, fossil fuel industrialization is unsustainability encapsulated.

Thus, grassroots environmentalists are more likely to take an ecocentric or deep green philosophical position, seeing humans as but one species amongst many others of equal value. Often attracted to anarchist thinking, they generally favor small scale and appropriate technological solutions because they argue that these enable humans to live within the wider ecosystems that nurture them, not attempt to dominate them. They are more likely to want to value local and indigenous knowledges, perhaps based on a deep understanding of the ecosystems in which indigenous people have lived for generations, and the knowledges that have grown up over time that have been constructed by a society adapted to that ecosystem. They look to value what Martinez Alier (2002) has called the environmentalisms of the poor and support an ethos of environmental justice, opposed to the destruction of environments inhabited by poor people and people of color. In cities, grassroots environmental struggles may take the form of urban social movements or community activism against, for example, a decision to site an incinerator next to a poor community.

Grassroots environmentalism may take the form of a political ecological struggle over the use of water resources or about land use. It may be a struggle over the industrialization of the countryside, over plans for, for example, a wind farm or fracking shale gas. In the latter case, grassroots environmentalisms can be dismissed by opponents as “nimby,” arguing that development should be somewhere else, “not in my back yard.” Grassroots environmental actors can, indeed, be relatively globally privileged actors defending their property values in an area of natural beauty valued for its recreational value. Or they can take a more ecocentric position, defending the rights of other species for their own sakes, opposed to development anywhere: not on Planet Earth (“nope,” not “nimby”). Here, disputes about the value of methodologies for valuing the services provided to humans from the natural world, so-called payments for ecosystem services, come into play. Some grassroots environmentalisms would welcome the chance to draw the value of the natural world into land-use decision-making. Others would oppose what they would see as the “commodification of nature.”

Grassroots activists may active in a number of social movement organizations at the same time, as well as in noncontentious environmentalisms, such as conservation societies, and vote for green parties. They are part of the wider environmental movement, not separate from it. Movement between the different elements of the wider movement is free and easy. They may be active in animal liberation politics, wishing to value other species in their own right. They are also likely to be engaged in women’s liberation, the peace movement, LGBT politics, and/or race-based movements, groups between which
again they would not wish to draw a wall. They see themselves as part of the wider movement for global social justice, with the proviso that theirs is a politics from the bottom up, from the grassroots, periodically coalescing into convergence spaces (Routledge 2003) in which they come together to act in concert, before dispersing.

**History**

Some would claim that grassroots environmentalism can be discerned in political actions going back to the Diggers of the English revolution tilling the land in common on St Georges Hill in Surrey, through direct action against industrialization from the Luddites to attempts to establish alternative or intentional communities in the United States and elsewhere in the nineteenth and twentieth centuries. Contemporary grassroots activism emerged in the 1960s as activists in the wider antiwar, countercultural, and student movements in the United States began to engage with the effects of environmental issues on people’s daily lives, and with wider debates about the environmental dangers of nuclear power and nuclear weapons in an era where nuclear weapons were still tested in the atmosphere in the South Pacific.

The issue of hazardous waste in poor communities was encapsulated in the struggles over Love Canal in Niagara Falls, New York, where hazardous chemicals had been dumped for decades in an abandoned canal that was later sold for a dollar for the construction of a school. The resulting health problems amongst local residents sparked grassroots mobilization to acknowledge the problem, close down the school, and relocate the community. The aptly named Love Canal became the first environmental cause celebre, inspiring many community protests against toxic chemicals that would have been tolerated before.

The environmental justice movement was born. Activists also pointed to the issue of environmental racism: the strong correlation between race and the location of toxic waste dumps.

Other grassroots environmentalists in the United States were inspired by Ed Abbey’s (1973) novel, *The Monkey Wrench Gang*, which popularized the tactic of “monkey wrenching,” the clandestine sabotage (ecotage) of industrial machinery that was engaged in ecological destruction of wilderness. The book inspired Earth First!, a loose network of activists who, adopting the slogan “do or die,” engaged in semilegal or illegal monkey wrenching, at times at great cost to their members (EF! member Judi Bari survived a car bomb). Others, inspired more by Gandhi than by Abbey, reacted against what they perceived as red neck misanthropy, rejecting a politics that seemed antihumanist. Some Earth Firsters, especially in the United States, argued strongly that humanity had become unsustainably populous and needed to die off: might AIDS be nature’s way of fighting back? Others identified a racist trope – were there not too many people but too many brown skinned people? Many were repelled by such antihuman views (Scarce 1990).

European and British grassroots environmentalism owes more to the peace and anti-nuclear movement and to feminism than to the preservation of wilderness given that it is almost impossible to find areas of Europe that have not in some way been altered by humanity. The Campaign for Nuclear Disarmament emerged in the late 1950s, and antinuclear activists opposed both civil and military applications of nuclear technologies. Large scale protests against nuclear power, especially in Germany, when allied to the threat to German forests from acid rain led to the formation of the German Green Party, which specifically saw itself as an antiparty party linked to grassroots environmental and peace activism. Friends of the Earth formed specifically
as a result of the failure of conservation groups to oppose nuclear technologies. Nuclear accidents at Sellafield (UK), Three Mile Island (USA), and, of course, later the catastrophic accident at Chernobyl in the USSR were huge mobilizers of protest against nuclear power. The decision by NATO to base ground-launched Cruise missiles in western Europe in the early 1980s was a similar mobilizing event, sparking mass marches and the long lasting peace camps at US bases in the United Kingdom.

The ideas and tactics of Earth First! crossed the Atlantic in the early 1990s as a result of the decision of the then UK government to engage in a major expansion of road building (Wall 1999). Protest camps were established to oppose a number of road building projects that activists claimed destroyed ecosystems and communities to promote unsustainable and antihuman “car culture.” “Reclaim the Streets” and “Critical Mass” organized occupations of road space for street parties and cycling festivals: large numbers of cyclists would congregate at rush hour and proceed through the city at a leisurely pace, arguing that they were not blocking traffic, they were traffic. This movement won a limited, if short term, victory in that a number of planned road schemes in the United Kingdom were cancelled. A new raft of innovative grassroots protest techniques such as “locking on” (activists using bicycle locks to attach themselves to machinery, property or trees scheduled for demolition) and constructing tree houses or sleeping platforms on stilts became a regular feature of grassroots environmental activism.

Other grassroots environmental activists experimented with the construction of localized economies. They established a number of local exchange trading schemes (LETS), networks sharing skills through the use of a community currency that members created and agreed to share between themselves. They established housing cooperatives and other forms of cohousing, and independent media. They set up bike repair workshops and social centers where activists could meet and exchange ideas, often in squatted buildings. These grassroots networks formed some of the currents from which emerged the anti-globalization and anti-capitalist movement that emerged at the turn of the last century until the events of September 11, 2001 and the resulting war on terror gave birth to a significant antiwar movement worldwide, but diverted attention away from environmental issues.

While long-term anthropogenic global warming had been identified in scholarly papers for many years, Al Gore’s (2006) film “An Inconvenient Truth” and a series of high temperature events and disasters like Hurricane Katrina contributed to a new form of grassroots environmental activism in the early years of the twenty-first century, climate activism (North 2011). Concerns about the implications of dangerous levels of climate change, “peak oil” (a time when easily extractable and cheap fossil fuels become more scarce), and other forms of resource constraint, and, after 2008, the global financial crisis intertwined to give a spur to a new wave of alternative institution building in the form of transition towns and low carbon communities. Activists met to educate themselves and their fellow community members about the implications of these three threats to continued global prosperity, and to take steps to create what they argued was the alternative: a community-run, convivial, localized grassroots economy. They established local currencies, community-owned ways of generating renewable energy, local food, local forms of production, and cooperative housing in order to bring these localized economies into being (Hopkins 2008).

Other grassroots climate activists, concerned that the transition movement was apolitical,
argued that more robust direct action was necessary to confront those entrenched powerful groups that, they argued, were responsible for global warming. Short duration protest camps were established outside power stations at which activists experimented with alternative ways of living in low carbon ways, and carried out direct action against polluters often in highly imaginative ways. For example, some climate activists formed the rebel clown army, using clowning as a way to communicate the politics of the movement and to disarm the states’ repressive apparatus: who would hit a clown? (Mason 2012). In the United States, high profile activism opposed the building of pipelines to carry oil from Canadian tar sands, and against coal-fired power stations. In the United Kingdom, communities opposed “fracking,” injecting fluids under high pressure to release natural gas from shale.

Drawing what they argue are the lessons of the failure of global elites to respond to the multimillion strong global protests against the forthcoming attack on Iraq in 2003, grassroots climate activists can be skeptical about the power of demonstrations to change policy, preferring direct action. They point to the success of the struggles of the suffragettes in early twentieth century Britain, which, they argue, forced the issue into the news and alerted an apathetic population about the nature of the problem. And, as was the case with Earth First! in the United States, climate direct action has drawn the attention of the security forces to the movement, with the high profile unmasking of police informers who had been embedded in the movement for many years, even fathering children with activists.

The response of the environmental movement to plans for a major expansion of nuclear power in the United Kingdom is unknown at the time of writing, although some high profile environmental activists have argued that given the danger of climate change and the need to keep the lights on, nuclear power should be embraced. Others, less convinced, continue to argue for grassroots, convivial, community-based local economies powered by renewable energies. They continue to be skeptical about high tech solutions to the climate crisis.

Current dilemmas

A number of questions face grassroots environmentalism in the early twenty-first century. Should the response to environmental and, more recently, climate crises work with or oppose the grain of currently hegemonic neoliberal conceptions of how to create prosperous economies, have faith in technological solutions (and in the human ingenuity capitalism unleashes), and see mass consumption lifestyles as inevitable? Who, some asked, ever rioted for austerity? Should the movement aim to convince large numbers of people to change their unsustainable practices, or concentrate on communicating the problem to the media and developing sustainable ways of living through what Seyfang and Smith (2007) call grassroots innovation niches? Transition towns activists are particularly critical of the likelihood of large numbers of people embracing the deeply ecological lifestyles pioneered in the protest camps.

Is grassroots environmentalism anti-capitalist or not, or, in the case of accusations of nimbyism, even progressive? Are radical environmentalists, as some neoliberal commentators and climate skeptics argue, “watermelons” (green on the outside, red on the inside), or wealthy people protecting their way of life in the face of what needs to be done to keep the lights on? For grassroots environmental activists associated with the peak oil movement, there is no point in actively confronting a system that is doomed to
collapse anyway as a result of resource constraint. Mass high consumption lifestyles, they argue, are unsustainable, have no future, and so the task for twenty-first century grassroots environmentalism is to use the same ingenuity that was used to build fossil fuel capitalism to build sustainable alternatives. While they would argue that, of course, it is necessary to oppose egregiously unsustainable practices like the construction of new fossil fuel power stations, they argue that too often direct action convinces few to change their livelihoods, and will be easily contained by the state – as the levels of police infiltration into the movement demonstrate. What is necessary is the patient work of construction of alternatives to unsustainable capitalism, which in many cases means recovering local knowledge about what food can be grown locally, how to reuse and repair what we have, and develop renewable energy projects, cooperatives, and local currency schemes. Anti-capitalist advocates of direct action counter that transition towns activists are naïve in expecting powerful elites to enable convivial and egalitarian grassroots localized economies to grow up as a viable alternative to the system from which they benefit.

Another question concerns the continuing nature of the grassroots environmental movement. It continues to be dogged by accusations that is white, globally privileged, and in danger of being seen as defending itself from developments that it still does not want in its back yard – an issue from fracking and opposition to wind farms especially. On the other hand, others argue that globally privileged actors engaging with the geographies of responsibility for their engagement with 200-year-old unsustainable fossil fuel capitalism is to be applauded. Those who accuse the western environmental movement as globally privileged ignore ways that networks of activists in the Global South and North can collaborate, or the rise of environmentalisms in the global south, especially in China (Watts 2011). Grassroots environmental activists continue to oppose the dumping of hazards on poor communities. But what has yet to be constructed is a global grassroots movement capable of opposing unsustainable practices, and bringing into being a sustainable economy that balances the right to development of all the world’s people in a climate constrained world where humanity has changed the planet to such an extent that real wilderness, untrammeled by humanity, cannot really be seen to exist. We now live in a period where humanity has changed the planet to such an extent that we are now living in a new geological age, the anthropocene. Grassroots environmental activists now need to develop an ecological and economic ethics for this new era (Gibson-Graham and Roelvink 2010).

SEE ALSO: Anarchist geography; Anthropocene and planetary boundaries; Climate and societal impacts; Energy resources and use; Environment and resources, political economy of; Environmental discourse; Environmental (in)justice; Indigenous knowledge; Political ecology; Social justice

References

Mason, K. 2012. “No Names, No faces, No Leaders: The Risible Rise and Rise of CIRCA, an


In simple terms, environmentalism can be understood as the political and personal actions that are taken to justify and enable the protection and transformation of the environment in different ways. There is, however, a challenge associated with trying to develop an effective understanding of environmentalism. Environmentalism is associated with such a bewildering array of processes, ideas, and practices that its nature is difficult to grasp.

There are several different ways in which environmentalism can be defined. It can be understood in relation to that with which it is associated, namely, the environment. Environmentalism can thus be seen in the action of ecologically oriented protest groups, the electoral politics of green parties, the planning processes that seek to protect parts of the natural world, and the laws that seek to reduce the human impact on the global environment. Environmentalism is also evident in the energy company that fits pollution filters to its chimneys, the airlines that offset the impacts of flying by planting new forests, and the supermarkets that support organic food production. One can also define environmentalism by what it seeks to do. This might be protecting a piece of apparently unspoiled nature, reducing the amount of greenhouse gases that are entering the atmosphere, raising awareness of environmental change, opposing the damaging environmental activities of governments and large corporations, carefully monitoring the processes of ecological change, or providing education about how people can reduce their own impact on the environment.

But there is more to environmentalism than its goals. First, environmentalism often offers a form of metaphysical perspective. Metaphysics is concerned with the fundamental nature of being and seeks to highlight processes that may not be immediately obvious in any given moment. In this context, environmentalisms of different forms theorize the underlying processes that inform human–environmental relations in different ways. For eco-Marxists, environmental exploitation is connected to the broader processes of capitalism and its tendencies to externalize environmental costs (Burkett 2014). For deep green philosophers, human–environmental interactions are defined by an emerging disconnect between human consciousness and the natural wisdom (Luke 2002). Other branches of environmentalism emphasize the role of religion, technology, urbanization, gender relations, and race in shaping emerging human–environmental relations (Pepper 1996).

Second, environmentalism can be understood as an ethical standpoint that attempts to define what is just, right, and proper in environmental affairs (Fox 2006). This is not to suggest that there is necessarily any agreement on what is ethically good and bad within environmentalism. While certain thinkers and activists will argue...
that environmental justice is defined by limiting human interference in the biosphere and allowing ecological systems to return to a state of nature, others claim that such idealistic thinking is impossible and that humans have a collective duty to manage the global environment sustainably in order to balance the needs of economic growth, social welfare, and ecological protection.

Third, environmentalism marks out a form of political position and mode of representation. In the nineteenth century, this involved initial systematic attempts to protect nature from the harmful impacts of human development. The second key period in the development of environmental politics was the 1960s and 1970s, when environmentalism emerged as one of a series of radical movements (alongside feminism and antiwar campaigns) that sought to question the values, logics, and consequences of the Western military-industrial complex. As a political position, environmentalism is expressed in the treehouses of road campaigners and the meetings of the Intergovernmental Panel on Climate Change; antifracking camps and the electoral politics of green parties; animal rights campaigners and antinuclear power groups. What unites these diverse actors and interests is a desire to bring nature and the environment into political debates and consideration.

This entry explores environmentalism in three different forms and draws particular attention to the connections between environmentalism and geography. The Prussian geographer Alexander von Humboldt famously defined geography as the study of the Earth as the home of the human race. This deceptively simple definition reveals that geography finds its origins in the ways in which environmental and human systems interact. As a discipline that has brought together those keen to document changes in the planet’s environment (including hydrologists, glaciologists, pedologists, geomorphologists, meteorologists inter alia) with those who seek to understand the particular nature of the human condition (including anthropologists, urbanologists, behavioral scientists, and historians), geography has occupied very similar practical and conceptual spaces as environmentalism.

The connections between geography and environmentalism can be understood in three main ways. First, the research of environmental geographers into the processes of environmental change has actively contributed to scientific, public, and political awareness of environmental transformations that have inspired different forms of environmentalism to emerge. Second, the work of certain forms of critical geography (particularly in the context of Marxist, feminist, and various strands of poststructuralist geography) have helped to expose the broader processes that generate environmental change and prevent effective action being taken to prevent ecological harm. Third, geographers have been important chroniclers of environmentalism, and have explored the ways it has emerged in particular places, the specific spatial tactics it has deployed to pursue its political and ethical goals, and the impacts of different policy regimes and modes of socioecological practice.

Environmentalism in history

Environmentalism is now a mainstream phenomenon. This can be attested in governments, where environment ministries exist throughout the world; in the media, where broadsheet newspapers have dedicated environmental sections and TV channels employ environmental correspondents; and in politics, where green parties vie for our votes on a regular basis. As with many contemporary phenomena, however, there is a tendency to ascribe much too short a history to the notion of environmentalism. Given the
rapid rise of environmentalism over the last half century, there is a collective propensity to think of it as an invention of our time. In reality, environmentalism is much older than that. Although environmentalism as a formal political movement may have relatively recent points of origins (dating back to the late 1960s and 1970s) as both a metaphysical perspective on socioenvironmental affairs and an ethical standpoint with regard to human–ecological relations, environmentalism is tied into a much longer human past. While they may not have voted for a green party, or campaigned against the impacts of climate change, early human societies developed complex understandings of the relationships between people and their surrounding environment, and orchestrated sophisticated systems and rituals for controlling these relations. By briefly exploring the complex historical evolution of environmentalisms of different kinds, the next section reveals some of the historical ideas that continue to inform contemporary manifestations of environmental thought and practice.

Premodern environmentalisms

One of the most valuable insights into the history of environmentalism is provided by the geographer Clarence Glacken. A professor of geography at the University of California, Glacken is today most widely remembered for his book *Traces on the Rhodian Shore* (1967), which explored the history of Western thought concerning culture and nature. The notions of “designed Earth,” “environmental influence,” and humans as “geographic agents” have, according to Glacken, been characteristics of thought since at least the time of the ancient Greek world. These systems of thought echo the ethical dimensions of modern environmentalism to the extent that they raise the question of how humans impact on environmental systems (in relation to the notion of humans as agents of environmental change), and imply the existence of a natural environmental order in the world (environmental influence). They also reflect the metaphysical dimensions of later environmentalism in the ways they encourage a search for deeper, hidden meanings in the observation of environmental systems (in relation to the notion of a divinely designed Earth).

Recognizing the ancient history of environmentalism is helpful on at least two fronts. First, it enables us to see more clearly the evolutionary roots of our environmental concerns and to understand their place within the broader human condition. Second, being cognizant of the ancient history of environmentalism illustrates how environmental thought has changed over time, and that our own environmental assumptions may not be as solid as we believe. A closer look at the prehistory of environmentalism reveals the selective translations of certain ideas into different strands of contemporary environmentalism. In his celebrated discussion of the medieval roots of modern environmentalism, for example, David Pepper (1996) provides a captivating account of premodern visions of nature. According to Pepper, the medieval period exhibited forms of environmentalism that resonate strongly with the two dominant strands of modern environmental thinking: ecocentrism and technocentrism (1996, 37). For Pepper, ecocentric perspectives emphasize the rights of the environment and the importance of observing and learning from the natural world. Conversely, technocentrism suggests that human science and technology have the capacity to deal with environmental problems without the need for society to significantly limit its socioeconomic ambitions.

The medieval world was characterized by two ways of thinking about the environment:
anthropocentrism and organicism (Pepper 1996, 124–165). Anthropocentrism was a medieval worldview based on the assumption that the Earth was at the center of the universe and constituted a divinely created terrestrial dwelling for humans. Anthropocentrism led not only to the common belief that the environment’s purpose was primarily one of human use, but also to the common practice of explaining natural phenomena through human-centered modes of understanding. Consequently, in the medieval period it was common to hear accounts of the Earth and environmental systems that likened it to a human body (Pepper 1996, 130). The idea of anthropocentrism continues to inform technocentric versions of modern environmentalism. Technocentrism thus typically places human needs over those of the intrinsic rights of nature, and tends to project technological solutions onto environmental systems that fail to operate as social systems require.

While medieval cosmologies tended to place humans at the center of creation, they were also based on a firm belief in the organic unity of the total environment. In the medieval world recognition was constantly given to the mutual interdependencies of all living and nonliving matter (including humans). Organicism was associated with popular and scientific beliefs in a Great Chain of Being, within which all things, from the lowly truffle to the mighty whale, formed part of an integrated organic whole (Pepper 1996, 132). In certain versions of the Great Chain of Being, humans were located between monkeys and angels! Organic visions of the Earth continue to inform ecocentric environmentalisms today. In the defense of environmental rights and consistent attempts to recognize humanity as part of, and not separate from the environments on which they depend, ecocentrism supports an essentially organicist worldview.

Modern environmentalism

Although environmentalism has ancient points of origin, the environmentalism we are most familiar with today has a more recent point of origin. The immediate post–World War II period is associated with an unprecedented period of economic growth and prosperity for many living in more economically developed countries. This period witnessed the rise of mass consumption, the popularization of the motor car, the emergence of new forms of comfort in the home (particularly in relation to central heating, labor-saving devices, and refrigerators), the rising availability of fast and cheap food, and the broader rise of the suburbs. This period of economic growth has been described by some as the “Great Acceleration” (Whitehead 2014) and was associated with new forms of often unnoticed pollution that were products of the industry on which economic growth was based. It is in this context that a new generation of ecological scientists, operating both within and beyond the discipline of geography, played a critical role in alerting the world to the environmental impacts of the mass consumption society. The most celebrated of this new generation of scientists was undoubtedly Rachel Carson.

Carson was an American biologist, conservationist, and nature writer. She is now best known for her 1962 book *Silent Spring*, which explored the impacts of the expanded postwar use of artificial pesticides (such as DDT) on wildlife and broader environmental systems (Carson 2002/1962). While artificial pesticides had enabled harvests to increase greatly and had helped to keep the costs of food mercifully low, Carson revealed that many of the chemicals in circulation were in fact carcinogenic and were having a devastating impact on wildlife populations. *Silent Spring* led to outright bans on certain pesticides in many countries, and
laid the foundations of the modern environmental movement which arose in the 1960s. In many ways, Carson’s work was emblematic of a broader field of ecological science that had been emerging since the nineteenth century. This was a form of science that lent itself less to technocentric solutions to environmental problems and more to illustrating the complexity and fragility of environmental systems. Ecological science would ultimately offer a scientific basis to the forms of ecocentric environmentalism that had historically been based on myth and conjecture.

During the 1960s scientifically grounded, ecocentric forms of environmentalism flourished and grew. What is now described as modern environmentalism grew out of the connections that were forged in the 1960s between concerns over environmental destruction and broader currents of countercultural protest. Countercultural movements in the 1960s questioned the orthodoxies of mass production and consumption and persistent forms of conflicts (focusing specifically on the American war in Vietnam and the proliferation of nuclear weapons). This was a form of politics that resonated strongly with environmentalists’ concern over the forms of social values that were generating environmental degradation. By the late 1960s and early 1970s modern environmentalism was evident in the presence of the first green political parties, the rise of various governmental departments that were devoted to environmental protection, and various green nongovernmental organizations.

Nongovernmental organizations (NGOs) have been central to the emergence of modern forms of environmentalism. The first environmental NGOs can be traced back to the nineteenth century and the establishment of prominent nature protection organizations such as the Sierra Club in the United States and the Royal Society for the Protection of the Birds (RSPB) in the United Kingdom (which was established in 1889 as the Society for the Protection of Birds). The RSPB was formed in order to oppose the use of bird feathers as plumes on hats. As a society with a royal charter, many of whose members were from the upper and middle classes, the RSPB represents what is often termed a reformist environmental group. Reformist environmental groups emphasize gradual change within existing social and governmental structures as an effective route to addressing social problems. Since the 1960s and 1970s a new breed of environmental NGOs has emerged which have called for more fundamental changes in social relations with the natural world. So-called radical environmental groups such as Greenpeace and Earth First! argue that only significant changes in the nature of the capitalist industrial order will lead to the protection of the biosphere over the long term.

A defining moment in the history of modern environmentalism came in 1972 with the first international conference being convened to discuss emerging environmental problems: the United Nations Conference on the Human Environment, which was convened in Stockholm (Adams 2002). While in many ways it reflected the culmination of centuries of struggle and intellectual evolution within the field of environmentalism, it was also a key turning point in environmental debates. In Stockholm it quickly became apparent that addressing many of the most pressing environmental issues facing humanity could not be achieved in isolation from the pursuit of economic development in many of the poorest parts of the world (Adams 2002). During the 1970s and 1980s the tensions between the demands of environmentalists and the need for economic growth were tentatively reconciled in the now popular (if highly contested) concept of sustainable development (Whitehead 2007).

Drawing on the insights of sustainable yield science, sustainable development suggested
that it is possible to combine environmental protection with continued economic growth, so long as certain ecological systems on which humanity depends were managed in a sustainable way. In many ways, it was the idea of sustainable development that enabled the broader goals of environmentalism to enter the political mainstream, as national, regional, and local governments and corporations signed up to its core principles. But, rather than reflecting a new era for environmentalism, sustainable development appears to have provided new contexts within which the competing visions associated with environmentalism could be expressed and contested. Writers have, for example, observed that there are two broad forms of sustainable development: weak and strong (Neumayer 1999). While strong forms of sustainability suggest that there are natural limits within environmental systems that human development should observe, weaker versions of sustainability suggest that long-term sustainability can be achieved through the substitution of human capital (in the forms of new technological developments) for natural capital. It is clear that strong interpretations of sustainability are grounded in ecocentric thought, while weaker interpretations of sustainability are indebted to technocentric ideas.

In his celebrated analysis of the history of environmentalism, O’Riordan (1981, 5) claims that there are actually important subdivisions within ecocentric and technocentric forms of environmental politics. According to him, ecocentrism is composed of deep ecology and self-reliance movements. The deep ecology movement emphasizes the complex web of interconnections and mutual dependencies that characterize life on Earth, and advocates that nonhuman life should be given rights and protection in a similar way to human life. Self-reliance initiatives are less ecologically oriented but emphasize the importance of living at small scales in close proximity to the environmental resources on which a community depends. What deep ecology and self-reliance movements share is a suspicion of large-scale modern technologies and cultures of mass consumption (O’Riordan 1981).

O’Riordan argues that technocentric thinking can be divided into two main groups: environmental managers and cornucopians. What unites both of these perspectives is a belief that economic growth and environmental resource exploitation can continue without posing a significant long-term threat to the viability of the biosphere. The environmental managerial perspective recognizes that, in order to protect the environment, it is important that new laws, taxes, and restrictions are put in place and that there is careful scientific monitoring of environmental pollution. Cornucopians, however believe that humans will always find a way out of the difficulties that environmental change may precipitate and that the best way to release creative energies is to promote deregulated market systems (O’Riordan).

It is not difficult to find examples of the different forms of ecocentric and technocentric environmentalism identified by O’Riordan in the world around us. Geoengineering, for example, appears to take environmental management to new levels. Geoengineering involves the large-scale manipulation of environmental systems by humans in order to tackle problems such as climate change or nitrogen pollution in marine systems. In relation to tackling climate change, climate engineering could involve spraying the upper reaches of the Earth’s atmosphere with sulphur particles to produce an artificial cooling effect. Many have been critical of the idea of geoengineering, partly because it is claimed that it does not address the root causes of contemporary environmental problems, but also because of the often unknown nature of the environmental consequences of the geoengineering interventions themselves.
In opposition to geoengineering, groups have advocated more radical transformations in the nature of the relationship between humans and the environment. Prominent among these initiatives has been the degrowth movement, which exhibits many of the characteristics of O’Riordan’s self-reliant branch of environmentalism. The movement calls for a change in social attitudes toward economic growth. While the necessity of growth has been an axiomatic assumption within capitalism since its inception, advocates of degrowth claim that, in future, society must find a way to develop more stable and sustainable levels of economic development. The degrowth movement is based on the fact that there are biophysical limits to the total growth of the global economy and that economic growth does not necessarily lead to enhancements in human wellbeing. In this context, contemporary manifestations of ecocentric environmentalism are increasingly exploring the ways in which it may be possible to develop socioecological prosperity without having unregulated economic growth (Jackson 2011).

**Environmentalism across space**

As well as historical perspectives on environmentalism, spatial analyses of environmentalism are also important for two broad reasons. First, they draw attention to the in situ practices that characterized environmentalisms of different kinds. By focusing on what could be termed “actually existing environmentalisms” (the forms of environmentalism that are actually delivered on the ground rather than expressed in ideas and discourse), geographers have exposed the inevitable compromises that have to be sought between ecocentric and technocentric brands of environmentalism.

Second, a spatial perspective reveals the important role that space plays in the constitution of environmentalisms of different kinds. If space is taken to refer to various aspects of proximity, location, scale, and boundary, it can help to expose the ways in which environmental thought gets repeatedly expressed, realized, and contested in the world around us. As the rest of this section will show, these forms of spatial perspective on environmentalism can shed valuable light on its metaphysical, ethical, and political dimensions.

**Proximity, scale, and boundary**

There is something of a geographical paradox at the heart of ecocentric versions of environmentalism (see Sandbrook 2015). On the one hand, these environmentalisms emphasize the importance of being in close proximity to the natural world. Through proximity to nature, many in the environmental movement argue, people can both enhance their own wellbeing and develop a better understanding of, and respect for, the workings of environmental systems. The idea of the value of environmental proximity is often traced back to the nineteenth-century writings of the American author Henry David Thoreau, who famously spent two years conducting an experiment in simple environmental living on Walden Pond in Concord, Massachusetts (Thoreau 1996). The environmental proximity principle is also part of a much broader attempt to redefine ethics in relation to the land and landscape. In his celebrated book *A Sand Country Almanac* (1968), the American ecologist Aldo Leopold suggested that an environmental ethic should be based on the extension of the ethical community from people to the ecological systems they inhabit. The land ethic suggests that moral judgment about right or wrong actions can only be made with reference to understandings of the actual consequences of those actions for the social and biotic communities in which they are made.
ENVIRONMENTALISM

At the same time, many environmentalists believe that the key to protecting the environment is to ensure that it is kept separate from human activity and intervention (Sandbrook 2015). The separation of environmental spaces, landscapes, and ecosystems from human activity can be seen in a range of spatial designations including national parks, wildlife reserves, and the greenbelts that encircle many cities. Most recently it has been observed in attempts to “rewild” certain environmental areas. These spatial strategies put physical distance and functional separation between human and environment systems, and while they may help to protect nature they can serve to work against the development of forms of environmental understanding and appreciation that rely on human proximity to nature.

The practices of human environmental propinquity and separation find expression in two of the most common strategies in environmentalism: conservation and preservation. The conservationist movement has consistently emphasized the ways in which careful human management of the environment can enable natural and human systems (including agriculture, tourism, forestry, and fisheries) to coexist without compromising ecological sustainability. This type of thinking lay at the heart of the conservation ethic promoted by the American forester Gifford Pinchot. Pinchot suggested that it would be possible, through the careful accumulation of scientific knowledge of the forests (and broader environmental systems), to use timber resources without compromising ecological integrity of the woodland systems from which they were taken. Preservationists, however, argue that any form of human intervention in environmental systems represents a disruption to that system and that the best way to protect the ecological integrity of different environments is, where possible, to prohibit human activity within them.

Interestingly, questions of proximity and separation are important features within geographical and environmental debates around the issue of scale. Although scale has various meanings within contemporary geographical scholarship, the term is used here to refer to the spatial extent of social and environmental organization. Many in the environmental movement have been critical of the ways in which industrial capitalism has resulted in the globalization of human life. Globalization is seen as problematic in environmental (and more specifically ecocentric) terms for three reasons: (i) because it is associated with increasing demands for finite energy resources to facilitate the global transfers of goods and services; (ii) because the global transfer of goods and services contributes to the production of rising levels of pollutants, in particular greenhouse gases; and (iii) because it results in the increasing separation of people from the environmental systems on which their day-to-day life depends. Significantly, those who adopt a more technocentric approach to environmental matters tend to be less worried about globalization. To them the free movement of trade across international boundaries provides new incentives and opportunities to develop the economic efficiencies and technological advances that will enable more sustainable forms of environmental relations to emerge (see Kahn 2010).

In response to the concerns that ecocentric thinkers have with the environmental impacts of globalization, certain branches of the environmental movement have called for the rescaling of socioecological life. Two prominent initiatives within this field have been the relocalization and bioregionalist movements. The relocalization movement, such as transition culture initiatives (Mason and Whitehead 2012), work toward the principle of subsidiarity: the idea that social and economic processes should be carried out at as local a level as possible. Relocalization
movements thus claim that if energy, food, and water production and supply can be carried out at community levels, the impacts of these processes on the environment will be greatly diminished. **Bioregional** movements assert the same ethos of localization, but tend to focus less on the community and more on larger regional scales somewhere in between the local and the national. What is significant about bioregionalism is that its advocates suggest that it is nature and not humans that should set the boundaries within which bioregional communities work. So, while bioregionalists claim that regions should be self-sustaining in their use and reuse of environmental resources, they also hold that regional boundaries should make ecological sense. Bioregions thus tend to be defined by river catchment areas, mountain ranges, or other ecological and geomorphological markers. Echoing the principles of ecocentric environmental thought, bioregionalists claim that if society is organized in environmental spaces it is much more likely that humans will understand and respect the ecological systems on which they depend.

Technocentric critics of relocalization and bioregional movements claim that these geographical expressions of environmentalism tend to generate arbitrary spatial boundaries in social and economic life. It is further argued that these boundaries not only inhibit personal freedom, but also serve to stifle the creative energies and exchange of ideas that are vital in the pursuit of solutions to some of our most pressing environmental problems.

**Environmental sites and locations**

The previous section illustrated that environmentalism has implications for how we understand the spatial organization of human societies, and that the decisions we make about the scales at which society operates have important implications for our relations with the environment. But there are other spatial factors that are important in the constitution and practice of environmentalism. The changing sites and locations within which different forms of environmentalism are practiced can tell us a great deal about their ethical and political dimensions.

Historically, many ecocentric forms of environmentalism have been located on the margins of mainstream society. The establishment of environmental communes, eco-villages, and other forms of green utopianism has embodied an attempt to provide examples of how humans can forge different types of living arrangements with each other and with the environment. While often criticized for constituting easily ignored forms of environmentalism, eco-utopianism has used its metaphorical and literal position on the margins of industrial society to demonstrate the forms of community that can emerge when environmental values are aligned in more ecocentric directions.

While environmental protests take very different forms, they often involve the deliberate occupation and reuse of prominent spaces. For example, the direct action environmental movement, which became popular in the 1990s, strategically occupied spaces in order to delay or prevent the building of new roads and other environmentally damaging infrastructures. More recently the rise of climate camps has seen the establishment of large-scale but temporary communities, in protest against global inaction in the face of climate change.

Perhaps the greatest locational challenge faced by modern environmentalism has come in the form of the complex processes associated with urbanization. Historically, many branches of environmentalism have either ignored or actively opposed the processes of urbanization. In these contexts, urbanization was seen as either the geographical antithesis of where
environmentalism’s concerns were located (namely, in the exurban spaces of nature), or as a spatial threat (particularly in the forms of transboundary pollution and suburbanization) to the ecologies that environmentalism was dedicated to protecting (Gottlieb 1993). Since the early 1990s there have been significant challenges to the urban blind spot associated with environmentalism (Harvey 1996). Groups such as the Environmental Justice Movement and the Cittaslow initiative have sought to actively change the geographical focus of environmentalism from nature to the everyday places where people work, rest, and play (Gottlieb 1993). The Environmental Justice Movement emerged from concerns over the consistent location of communities of color and low income near to sites of environmental pollution production and toxic waste disposal. The movement has sought to shift the locational politics of environmentalism from exurban nature and toward urban living. The movement has also sought to fuse concerns over environmental degradation with those of social injustice, and to forge a common politics of radical ecological resistance.

Interestingly, the Environmental Justice Movement, and the varied forms of radical political ecology with which it is associated, has also challenged the metaphysical, ethical, and political dimensions of environmentalism. At a metaphysical level the Environmental Justice Movement asks new questions about the ways in which the often obscured spatial relations of racism, sexism, and class exploitation generate social and ecological injustice at a range of scales. In relation to ethics, the practices of environmental justice seek to extend the moral domains associated with environmentalism. While environmentalism, historically, sought to extend ethical concerns from human to environmental realms, the Environmental Justice Movement, and associated forms of radical ecology, seek to extend the political concerns of environmentalism to marginalized socioeconomic groups. What was once a movement devoted primarily to the protection of nature has become one that recognizes that the protection of the environment cannot be achieved without addressing the spatial relations of global economics, racial exploitation, and neocolonialism.

**SEE ALSO:** Anthropocene and planetary boundaries; Environment and urbanization; Environmental ethics; Environmental (in)justice

**References**


Environmentality and green governmentality

Stephanie Rutherford
Trent University, Canada

Environmentality, or green governmentality, is a neologism that offers an extension of Michel Foucault's concept of governmentality (Foucault 1991, 1997, 2003, 2007, 2010). Defined very broadly as “the art of government” (Foucault 1991) or “the conduct of conduct” (Gordon 1991), governmentality takes aim at the modalities of modern rule. More specifically, studies in governmentality examine how power is not only repressive but also productive, eliciting and promoting particular knowledge systems, techniques for regulation, and subject positions that work to the best end of a governing authority (for further explanation, see the entry on governmentality in this volume). As an elaboration on governmentality, environmentality seeks to understand the ways in which this analytic of power can be applied to how we understand and act upon questions of environmental management, from global climate change negotiations (Bäckstrand and Lövbrand 2006) to individual habits of recycling (Darier 1999). It has become a useful concept to help think through the ways that the environment is not only a biophysical reality, but also a site of power, where truths are made, circulated, and remade.

As originally articulated by Foucault, governmentality did not give any weight to the environment as a domain of rule, even though he wrote at a time of eco-cultural foment. Indeed, Foucault’s first articulations of governmentality were largely preoccupied with questions of the modern state and the birth of liberalism, considering how various political authorities worked to govern their human populations to achieve the best end of the state. This focus on the practices of the state has largely remained as Foucault’s interlocutors have applied his insights to a range of cases (Rutherford 2007). However, as one can see from the two broad definitions of governmentality given above, there is room to extend the terrain that governmentality considers. Recognizing this possibility, scholars began to think about how the kinds of questions that governmentality asks about the world might successfully be applied to the environment.

For example, in the late 1990s, American political scientist Timothy W. Luke (1995, 1999) wrote a series of articles and book chapters that used the terms environmentality and green governmentality to explore the ways in which the environment (and the study of it) can be a site of political calculation. He is the scholar most closely associated with the coining of the term. For example in 1995, Luke examined the Worldwatch Institute (www.worldwatch.org/), a nongovernmental research organization concerned with global environmental issues, from biodiversity loss to climate change. Through the organization’s efforts to collect the latest data about the state of the planet’s environment, it measures and forecasts environmental harm and provides solutions to avert environmental crises. In so doing, Luke argued that the Worldwatch Institute is at the forefront in generating “the eco-knowledge of modern governmentality” (1995, 70). In later writings he went beyond the Worldwatch Institute to assert that “most
environmentalist movements now operate as a basic manifestation of governmentality” (Luke 1999, 121). Around the same time, Eric Darier’s edited volume *Discourses of the Environment* (1999) was published, containing a series of chapters that expanded the relationship between governmentality and nature even more fully. This book offered an early (and still salient) assertion that we are witness to the biopolitical management of all life, where nature is rendered into populations of resources to be mapped, measured, and managed. Humans are also arranged in various populations through environmentality, and as such, their interactions with the environment are framed as in need of management.

From these initial articulations, environmentality, or green governmentality, has been tested in a wide variety of cases – from biodiversity protection (Youatt 2008) to the American Museum of Natural History (Rutherford 2011) – and places – from the rural Philippines (Dressler 2014) to Amazonia (Cepek 2011). Although there are myriad ways this analytic of power might be applied, there are three interrelated areas that are included in any effort to understand the workings of green governmentality.

First is the production of rationalities of rule. Foucault suggests, “… power produces; it produces reality; it produces domains of objects and rituals of truth” (Foucault 1995, 194). Hence, the rationalities that underpin green governmentality suggest the forms of knowledge/power that open up particular grids of intelligibility while foreclosing others. Dean (1999, 23) has talked about governmental rationality as operating through “characteristic forms of visibility, ways of seeing and perceiving” that coalesce into regimes of truth. Perhaps the most salient form of rationality or way of seeing and perceiving the environment is through the lens of science (Rutherford 2007). The supposed impartiality, objectivity, and dispassion of science lend it an authoritative voice to speak for the wellbeing of nature on a range of scales from the local to the planetary. Statistical measurement, graphic representation, modeling, and forecasts all work to generate a broad picture of nature to be managed. Demeritt (2001) offers a potent example of how this happened in the Progressive-era United States by sketching the emergence of the “normal forest” through scientific forestry. Demeritt shows how the visual representation of the nation’s forests as both under threat and in need of management worked to enframe nature, making what was once heterogeneous appear to be coherent and unified. In so doing, the forests of the United States were abstracted from their specificity and made into a single national and natural resource, one whose depletion marked both the progress and the cost of civilization. By representing nature in this way, new categories and means of assessment became possible – such as annual allowable cut and maximum sustainable yield – and new authorities became empowered to manage these categories. And so, in the case of the “normal forest,” power was productive: it generated new ways of seeing, invented categories of analysis, normalized particular strategies for environmental management, and authorized experts (foresters and conservationists alike) to act on that management.

Second is the question of strategies of intervention. As Demeritt’s example demonstrates, the rationalities at play in representing environmental problems generate specific ways of seeing and hence offer a particular range of solutions for their amelioration. In a broad sense, the calculations offered through power/knowledge regimes produce a domain of objects – the environment – that is then taken up by various authorities – nongovernmental organizations, scientists, corporations, institutions like the World Bank, all levels of government, and even celebrities – to propose solutions to environmental problems.
Particular strategies of intervention are elicited, from establishing biodiversity preserves to geoengineering to composting, as mechanisms to deal with the environmental problem in question. One approach to assessing potential environmental harm that has gained broad acceptance in Western liberal democracies is the environmental impact assessment (EIA) (see Doyle and McEachern 2008). As these authors demonstrate, the application of a cost–benefit analysis through government-mandated EIAs generates forms of rationality read through expertise that seeks to measure and weigh the environmental impacts of economic activity. EIAs also invite public participation in the decision-making process, bringing both proponents and critics under its normative banner, where the proceedings produce new kinds of knowledge, but within a delineated (and often neoliberal) framework. In the end, these environmentality strategies follow the logic of the rationalities that suggest them; put differently, the solution necessarily follows the discourse used to frame the problem.

The third dimension central to environmentality is generation of specific kinds of self-governing subjects. Technologies of the self, or those tactics and practices that allow subjects to constitute themselves as particular kinds of people in and through power, are important to the practice of rule. Various discourses have produced the conditions of possibility for us tell to the truth about ourselves in a range of circumstances: for example, voter, activist, parent, child, student, or teacher. The multiple identifications with and performances of these subject positions, in turn, work to regulate the ways in which people present themselves to the world. In the case of environmentality, the subject position most often taken up is that of green citizen. This can mean different things at different times, but at its core, the green citizen manages her or his behavior, and the behavior of others, to promote the health of the environment. For example, Agrawal explores the idea of “intimate government” (2005, 193), where villagers in Kumaon, Northern India, became environmental citizens through a process of devolution of power, becoming experts in the management of their own forests through conservation initiatives targeted at their own practices as well as the practices of others. In a different vein, Rutherford (2011) had examined how a documentary like Al Gore’s An Inconvenient Truth can work to produce a particular kind of environmental citizen: a subjectivity formed through the rites of green consumption. In any incarnation, the green citizen is an excellent example of governing at a distance, whereby the state (or any authority) can rely on subjects to regulate themselves.

The three dimensions of environmentality elaborated above might be made clearer with the help of a brief example, one that also points to an emerging field in environmentality: the issue of climate change or climate governmentality (Stripple and Buckley 2013). Increasingly, both the political context and discursive framing around climate change have moved towards emissions trading as the mechanism through which global climate change can be managed. For example, a global cap-and-trade system through the now defunct Kyoto Protocol was normalized as the appropriate means to ameliorate climate change, rather than a carbon tax, or perhaps more radically, the payment of an ecological debt. However, this approach (or any approach) requires experts and specific technologies to measure either failure or success and determine the optimal results of a strategy of intervention. In climate governmentality, an important knowledge/power nexus that shapes the terrain of debate is found within the profession of accountancy. One might wonder what accountants have to do with what many would consider primarily to be a problem of the
ENVIRONMENTALITY AND GREEN GOVERNMENTALITY

planetary ecosphere. But the framing of climate change has emerged in distinctly financial terms. As Lovell and MacKenzie (2011) point out, accountancy has become increasingly central to this normalized way of framing both the problems and the solutions to the climate crisis. As such, the example of climate accounting provides an important and specific articulation of the key aspects of green governmentality explored above. Climate accounting produces a particular mentality or rationale for intervening in climate politics. Lovell and MacKenzie show that the accounting profession offers a reading of the climate problem that both narrows the terrain of intervention while inserting their expertise as central to its solvability; for accountants, “climate change is seen as a corporate problem, which is solvable with careful application of existing accounting approaches and techniques” (2011, 725). Lovell and MacKenzie further contend that accounting, accelerating with the rise of the European Union Emissions Trading Scheme and other mechanisms to trade carbon, has intervened with very specific policy solutions in a political terrain that asserts that climate change is a problem of calculation, measurement, and trading of carbon rather than its production. These authors suggest that the epistemic community of accounting, which emphasizes their specific professional expertise and technical strategies of calculation, offers itself up as the logical body of knowledge to deal with climate mitigation. Calculation here is key as it defines the manner in which problems can be assessed and acted upon by, following Murray Li (2007), rendering the problem in strictly technical terms, a hallmark of governmentality. And, although Lovell and MacKenzie do not make this assertion, a subject position could be put forward – the climate accountant – whose expertise and practical skills allow them to engage in policy construction and implementation on this question. As such, accountants not only respond to the question of climate change within their professional frame, but actively produce the broader discourse and policy action on this environmental problem, through the application of accounting rationales, the practice of particular technologies, and via their epistemic community. In this way, the seemingly apolitical discipline of accountancy constitutes and shapes the responses that can be taken to climate change; it is a productive force for environmental management. Put another way, the accounting profession is a key actor in the production of green governmentality on the question of climate change.

SEE ALSO: Biopolitics; Biopower; Cultures of nature; Environmental governance; Governmentality; Poststructuralism/poststructural geographies; Power

References


Eolian erosional processes and landforms

Julie E. Laity
California State University, Northridge, USA

Wind erosion is an important geomorphic process in locations with abrasive sediment, strong winds, and sparse vegetation. The sediment, largely sand, is supplied from rivers, beaches, and alluvial fans. As a result, eolian erosion is principally confined to desert, coastal, and cold (periglacial and paraglacial) environments. In deserts, owing to the pervasive aridity, lack of vegetation, free sweep of the wind over large areas, and abundant sand, a complete suite of eolian landforms develops, including ventifacts (wind-eroded rocks), yardangs, inverted relief, and deflated depressions. These landforms are best expressed in the very arid deserts of South America, Asia, and Africa, with the smaller and more humid deserts of North America having fewer features (ventifacts, small yardangs, and pans), and Australia appearing largely free of abraded landforms, although pans are present. Similar landforms have been identified on Mars from both satellite and rover imagery. In terrestrial coastal and cold environments, ventifacts are recorded, but yardangs are generally absent. During the Pleistocene, widespread abrasion occurred at the fringes of ice sheets, with many reported localities in Europe and North America.

The landforms of wind erosion vary greatly in scale, from small pans to enormous depressions, and from individual yardangs to immense meso-scale ridge systems. Erosion of ventifacts and yardangs occurs as sand is transported from source areas towards sinks. These landforms thus mark the path of moving sand and can be used to reconstruct pathways and, by inference, contemporary or paleo-circulation patterns. Landforms of eolian erosion are largely stationary, unlike many depositional landforms, and thus represent a time-integrated record of wind direction and intensity. Larger landforms, such as yardangs, range in age from thousands to millions of years, and thereby experience periods of climatic change, during which time they are modified by water as well as wind. Indeed, fluvial processes are often important precursors to wind erosion, supplying sediment, affecting surface erosivity in pan and depression formation, and providing conduits for wind erosion in yardang fields. Furthermore, water modifies yardang surfaces through solution, rilling, gullying, and mass movement. Climate change also affects the biomass, the height of the water table, and the circulation of the atmosphere. Wind erosion is thus strongly interconnected with many other elements of the Earth system environment.

Relative to landforms of deposition, such as dunes, those of erosion have received less attention. In part, this stems from the relative remoteness of many of these landforms from major research institutions and the difficulty in traveling to such regions. Many questions remain to be answered concerning the mechanisms of micro-feature formation on ventifacts and yardangs, the physical interaction between landforms and the wind, and the evolutionary history and rates of formation of landforms of wind erosion.
Processes of eolian erosion

Two principle processes operate to erode a landscape: abrasion and deflation. **Deflation** involves the removal by the wind of noncoherent, fine material from the surface. The material has often been pre-weathered by salt weathering or other processes, facilitating erosion. This process is most significant in areas with poorly lithified rocks or unconsolidated sediments. **Abrasion** results from the mechanical wear of coherent sediments or rocks by the impact of particles in saltation, chipping away the underlying surface. The relative significance of these processes of erosion is not well understood. Abrasion appears to be the key player in the development of several important landforms, including ventifacts and yardangs, whereas deflation contributes to the development of surface depressions, both large and small.

The process of abrasion is often likened to sandblasting. It is affected by the properties of the particles and the rock, the nature of the environment, and the timescale. The effectiveness of abrasion is related to sand grain composition, size, and shape. In terms of composition, the density and hardness of the particle are significant. The density determines the mass contribution of kinetic energy. Hardness influences whether kinetic energy is transferred to the ventifact or the grains become deformed themselves. Quartz grains dominate on Earth, with basalt sand present in some areas. On Mars, most sand is basaltic. As particles become larger in diameter, their kinetic energy for a given velocity increases greatly. However, it becomes increasingly difficult for the wind to entrain large grains. Therefore, while large grains are more abrasive, smaller sand grains may do more of the work, because they are more easily entrained (Bridges and Laity 2013). Grain shape is generally less of a factor in erosion because most sand grains are well rounded. Nonetheless, recently formed angular grains have the potential to cut and gouge rocks more effectively, enhancing abrasion.

Environmental factors also influence wind erosion. These include wind speed and shear stress; wind direction; particle supply, wind frequency, and integrated flux; local topography; and local rock distribution. As noted earlier, abrasion requires that winds be strong and frequent and that there be an adequate supply of sand. Thus, a region that is windy, but lacks loose sand, will not show signs of abrasion. Furthermore, abrasion may lessen over time as sand supply becomes depleted (commonly the case in glacial environments) or as the winnowing of fine sand leaves behind a coarse, protective lag (common in desert river systems). A continuous supply of abrasive sand may require surface disturbance, such as flooding, to destroy surface lags and expose underlying sand. These disturbance events promote and sustain the abrasion process.

On level ground, the height of sandblasting is typically limited to 1–2 m above the surface. This value is sometimes exceeded on hillslopes or in locations where sand ramps or beveled ventifacts propel sand higher into the airstream, impacting downwind rocks at a higher level.

Fossil ventifacts and yardangs mark regions where the conditions for abrasion have ceased. Relict landforms may appear weathered, stained, varnish–covered, or veneered in lichens. The lithology of the landform affects weathering rates. Yardangs formed of weak sedimentary rocks can show signs of mass wasting and gullying, and solution modifies limestone yardangs or those with a significant salt component. Salt weathering of ventifacts adjacent to playas results in rapid physical changes that cause many rocks to disintegrate. In deserts, some rocks, such as granites, weather relatively rapidly and do not provide a long record of abrasion on ventifacts; others, notably basalts, preserve features in fresh
condition over thousands of years. The partial abrasion of varnish, as observed on some ventifacts, indicates that morphologic development may occur over multiple cycles.

The processes associated with deflation have received less attention than those of abrasion. Deflation involves the action of the wind alone. In order for fine particles to be picked up and removed, the rocks or sediments must be pre-weathered. This can occur through salt weathering or the action of grazing animals, which disturb the surface by trampling, rendering it more susceptible to wind erosion. Deflated material is transported as fine grains in atmospheric suspension.

Landforms of eolian erosion

Desert depressions and pans

The wind plays a role in excavating large-scale desert depressions and forming pans, small-scale features that are widely distributed in arid lands. Desert depressions frequently have a polygenetic origin, involving wind erosion, block faulting, broad shallow warping, stream erosion, karst activity, mass wasting, salt weathering, and zoogenic processes (Aref, El-Khoriby, and Hamdan 2002). They include the large enclosed basins of Africa: Siwa, Qattara, Baharia, Farafra, Dakhla, and Kargha.

Pans are widespread in southern Africa, on the High Plains of the United States, and in western and southern Australia (Goudie and Wells 1995). Several factors predispose a region to pans, including a dry climate, a vegetation-free surface, a low water table, and materials susceptible to deflation, such as poorly consolidated sediments, shales, and fine-grained sandstones. Feedback mechanisms play an important role in pan growth and enlargement. As water accumulates in depressions, salts are left behind, which retard future vegetation growth, weather sediments, and attract large animals, rendering the surface more susceptible to wind erosion. The position of the water table influences both depositional and erosional processes in pans through complex interactions with vegetation and sediments.

Yardangs

Yardangs are the most impressive landform of eolian erosion, being immediately visible on aerial or orbital images of Earth and Mars. They form vast arrays of parallel forms, as observed, for example, in the Dasht-e Lut Desert of Iran. Around the Tibesti Mountains of Africa, wind-eroded ridge systems are of such extent that they follow the deflection of the trade winds and cross ancient impact craters. In large systems, barchan dunes may occupy the swales between yardangs, abrading the yardang flanks and the inter-yardang corridors.

In their most classic form, yardangs appear as elongate, streamlined ridges, resembling an inverted ship’s hull. The windward face is typically blunt-ended as a result of direct abrasion, whereas the lee end may taper to a point. However, the yardang assumes many forms and may also appear flat topped or short and stubby in appearance. The length-to-width ratios of yardangs average 3:1 or greater, a form that minimizes wind resistance. However, this value is influenced by rock material, time of exposure to the wind, and wind direction. In Peru, streamlined yardangs range in length-to-width ratio from 3:1 to 10:1. The length of yardangs ranges from meters to kilometers; those developed in basaltic lava flows of Holocene age in Argentina are up to 10 km in length (Inbar and Risso 2001).

Yardangs occur in the most arid regions of Earth, where wind processes dominate. Their formation is assisted by a favorable material, weathering, and channels formed by rock fractures or fluvial processes that are exploited.
by unidirectional or seasonally reversing winds. Prominent fields in Africa and Asia include those in Egypt, Libya, Chad, Niger, the Arabian Peninsula, the Namib Desert, the Taklimakan Desert, the Qaidam Basin, and the Lut Desert. In South America, yardangs occur in Argentina and along the coastal desert of Peru. There are minor groups in North America. Paleo-yardangs, relict conditions that were possibly drier and colder, are recorded in Europe, both in Spain and in Hungary. Owing to the remote nature of most yardang fields, there has been little comprehensive study of the processes of yardang formation, and meteorological information is often lacking. The relative role of abrasion and deflation remains unstudied and probably varies according to the material and the weathering environment.

Yardangs form in a broad range of geologic materials. These include claystones, sandstones, limestones, dolomites, granites, schists, gneisses, and volcanic ignimbrites and basalts. Yardangs in the Qaidam Basin of China are covered in thick crusts of salt (Rohrmann et al. 2013). In layered rocks, the form of the yardang is influenced by differential erosion of materials that vary in their resistance. Steps, fins, and shelves may result.

Between the yardangs are spaces referred to as troughs, couloirs, corridors, swales, or boulevards. When the yardangs are close together, a U-shaped corridor is common; as the distance widens, the couloirs become flattened. Most of the geomorphic work creating yardangs appears to be concentrated within these corridors, where ripple trains or barchans migrate downwind, abrading longitudinal striations and creating shallow erosional basins.

In addition to abrasion and deflation, other geomorphic processes modify the windform. Fluvial erosion helps to initiate yardang fields by creating stream courses that are refashioned and enlarged by the wind. In weak materials, yardang slopes are often gullied by rainfall and the corridors eroded by fluvial action. Mass movement modifies yardangs and slump blocks are observed, particularly near the undercut noses of the form.

The rate of yardang formation is strongly affected by the relative hardness of the substrate. On playas, yardangs can develop within a few thousand years, sometimes facilitated by a drop in the water table, which forms a base level to wind erosion. In the Bodélé Depression, Chad, 4 m-high yardangs were estimated to form in playa sediments within 1200–2400 years. In Argentina, yardangs 2–3 m in height formed in basalt in less than 10000 years. However, large bedrock yardangs along the coast of Namibia may have taken millions of years to develop (Goudie 2007).

**Inverted topography**

Inverted topography develops where areas of the landscape that were previously low in elevation, such as river systems and deltaic distributary systems, are left standing in high relief owing to their relative resistance to wind erosion (Bristow, Drake, and Armitage 2009). The resistance of the channels may result from the induration of sediments by chemical precipitates. Inverted topography is most often evident in yardang fields, with the former point bars and meanders of channel systems now marked by remnant hills and ridges. Examples of inverted topography include raised channels in Oman, Egypt, and China, as well as on Mars.

**Ventifacts**

Whereas yardangs are predominantly found in highly arid climatic conditions, ventifacts occur in desert, coastal, and cold (periglacial and paraglacial) environments. Ventifacts are also widespread on Mars, where they were identified, first tentatively from Viking images, and
then unequivocally from the Pathfinder and Mars Exploration Rover (MER) images. The requisite environmental conditions for ventifact formation are strong winds, a supply of abrasive sediment (dry sand), vegetation cover that is absent or scant, and a relatively stable surface covered by boulders or bedrock. Ventifacts occur both as active and as fossil forms. The geological record indicates that ventifacts are extensive in Quaternary deposits and may preserve evidence of intense wind activity in even older rocks, including deposits of Permian to Triassic age where the facets, grooves, and polish that exemplify ventifacts remain remarkably well preserved.

Wind erosion is pervasive in cold environments and ventifacts are found from sea level to altitudes as high as the Qinghai–Tibet Plateau (5200–5600 m a.s.l.) and the Argentine Puna Plateau (4500–5000 m a.s.l.). Modern ventifacts are reported from Antarctica and ventifact sites relict of Pleistocene ice age conditions are found widely across Europe and North America. Ventifacts form near glaciers because of the high sediment yields associated with retreating ice and the strength of katabatic or other downslope winds. Development occurs close to the glacial margin and ceases within a few decades of deglaciation, reflecting the relative speed with which vegetation recolonizes surfaces and abrasive sediments are winnowed away by the wind.

Coastal ventifacts have been described from many localities, including New Zealand, the United States, and the United Kingdom. However, it is difficult to determine from the literature how widespread they are and whether they are more common in temperate climatic zones than elsewhere. Abrasion may be enhanced during periods of lower sea level, when sand is exposed on coastal plains.

Ventifacts appear to be relatively common in arid and semiarid environments, but there has been little comprehensive or basin-wide mapping of these features, and thus their distribution here is also poorly known. They have been recorded in most of the world’s deserts, with the exception of Australia. It is likely that ventifacts are more areally extensive in deserts than elsewhere owing to the greater supply of dry, transportable sand. In coastal and paraglacial environments, by contrast, ventifacts are often limited to narrow fringes. Periglacial ventifacts may be more widespread, but their distribution has not been effectively mapped.

As Mars is a dry and windy planet with abundant sand and widespread rocky surfaces, it is not surprising that ventifacts are commonly observed in surface images. The environmental conditions differ substantially from those on Earth, however. The atmosphere is thinner (≈1/100), the gravity is lower (about 1/3), and the abrasive particles are largely basalt, rather than the dominantly quartz–feldspathic terrestrial sands. Nonetheless, the morphology and surface textures of ventifacts on Mars are broadly similar to those on Earth (Bridges and Laity 2013) and these features have been used to infer paleo-wind directions.

Field evidence and theoretical considerations indicate that sand is the most effective agent of abrasion in all environments. The efficacy of dust, ice, and snow particles has not yet been demonstrated. The process of abrasion usually occurs within the lower 1–2 m of the atmosphere. Kinetic energy (KE) increases with distance from the ground, usually peaking at approximately 10–40 cm, and then declining. As a result, abraded rocks evolve with a ramped facet up to the height of maximum KE. In very tall rocks, a notch can develop where abrasion lessens with height above the maximum KE point. In the field, conditions are often complex, with several conditions impelling saltating grains to elevations >2 m. These include: (i) hard surfaces, which impart greater grain bounce, (ii) moat and ramp structures in front of ventifacts – sand
travels up the ramp to abrade the upper surfaces of rocks, and (iii) in distributed rock fields, sand which flows up beveled ventifact faces and is propelled into the airstream, hitting rocks down flow on their upper surfaces. The rate of abrasion is strongly dependent on wind speed and frequency, in addition to characteristics of the particle and the rock.

Ventifacts are recognized by their unique morphology, which is often characterized by (i) facets, (ii) polish, (iii) textural features (pits, grooves, flutes, helical scores), and (iv) textural changes (etching, fretting, and knobby texture). Each ventifact has a unique shape, a function of the lithology and size of the rock, the size and nature of abrasive material, the duration of abrasion, wind strength and frequency, wind direction, and rock exposure (influenced by burial by sand, sheltering effects of other rocks, and flow interactions within the boulder field) (Laity and Bridges 2009). Textural features are universal in nature, such that a flute may form on volcanic tuff, basalt, granite, or marble. Textural changes are more dependent on the initial lithology of the rock, with etching, for example, resulting from layered rocks of different hardness. The orientation of linear features such as flutes and grooves parallels the direction of the highest velocity winds in a region.

Many ventifacts show facets – planar surfaces that are largely perpendicular to the wind, formed on the windward face of a rock. More than one facet may form if there are bidirectional or multidirectional winds. For small rocks, new facets may form when the rock is subject to shifting or overturning associated with earthquakes, frost heave, undermining by wind or water action, and disturbance by animal movement or humans.

Local topography influences wind speed and direction, and hence the location of ventifacts, the intensity of their development, and the orientation of features such as flutes, grooves, and facets. Wind is commonly accelerated through topographic constrictions and up the windward flanks of hills. As a consequence of this increase in wind speed, sand transport is increased. As a result, ventifacts are commonly found near ridgecrests and on the upper slopes of hills, and in notches and structurally controlled valleys. The intensity of abrasion, as measured by factors such as groove width and length and pit diameter, often increases with elevation up a slope.

The relationship between the form and texture of a ventifact and the wind direction may be used to infer atmospheric circulation patterns at the surface. If the ventifacts are relict, they serve as a proxy for paleocirculation reconstruction. Efforts to map at a regional scale require many observations because, in small areas, the orientation of grooves and other features may be strongly influenced by local topography. The results of mapping over wide-ranging areas can be related to models of atmospheric circulation to infer changes in wind pattern or intensity over time.

SEE ALSO: Geomorphic systems; Landforms of other planets; Soil erosion and conservation; Weathering processes and landforms

References


Further reading


Estonia: Eesti Geograafia Selts (EGS) (Estonian Geographical Society)

Founded: 1955
Location of headquarters: Tallinn
Website: www.egs.ee
Membership: 213
President: Mihkel Kangur
Contact: mihkel.kangur@tlu.ee

Description and purpose

The EGS is a nonprofit organization for geographers and enthusiasts of geography. Its principal goals are to develop the science of geography, to support education of geography, to popularize geography and concepts of sustainability, and to collaborate with scientific societies and institutions in Estonia and abroad.

The EGS provides a place where representatives of different divisions and groups of the geographical sciences can meet, providing favorable conditions for discussing wide-ranging geographical problems. To this end, an important role is played by conferences, seminars, and publications that report the results of scientific research conducted by its members.

Journals or major publication series

Eesti Geograafia Seltsi aastaraamat (Yearbook of the Estonian Geographical Society)
Estonia. Geographical Studies

Current activities or projects

The main field of activities by the EGS are: scientific research by members; publication of scientific and popular scientific works; organization of conferences, symposia, seminars and other scientific undertakings (expeditions and tours); and dissemination of geographical knowledge.

Brief history

Important in the founding of the Russian Geographical Society in 1845 were geographers from the Baltic province with close relations to Estonia, including academics K.E. von Baer, F.G.W. Struve, W. Dahl, and G. Helmersen; the ocean explorers A.J. von Krusenstern, F. von Wrangell, and F. Lütke; army leader F.R. Berg; and others. The idea of founding a geographical society in Estonia as a branch of the Geographical Society of the Soviet Union originated in the mid-1950s. By that time a considerable number of specialists had graduated from the Geographical Department of Tartu State University, a necessary requisite for establishing the organization. The resolution for the founding of the EGS was passed by the Scientific Council of the Geographical Society of the Soviet Union on November 15, 1955. The assembly elected Vello Tarmisto (Cand. Geogr., the originator of the idea for the EGS) as president.

Since 1992 and the 27th International Geographical Congress in Washington, the EGS has represented the Estonian National Committee in the International Geographical Union.

Submitted by Tiit Vaasma
Ethics in geography fieldwork

Jenny Lunn
American Association of Geographers, USA

An examination of the ethical principles that apply to geography fieldwork, as well as their application in the past, present, and future, must be preceded by a brief discussion of what is meant by “the field” and “fieldwork,” and what is particular to “geography fieldwork.”

“The field” is both a physical and conceptual space. While the term itself is contested, a simple interpretation situates the field as a counterpoint to the classroom, the laboratory, and the library as a place where information exists that can be collected and studied. While the field is outside of these other places of learning, it is not necessarily a distant or unknown place.

In turn, “fieldwork” is an act of entering the field to gather information which then can be processed, mapped, cataloged and analyzed in order to understand places, people, and phenomena. In the context of education, fieldwork is also a pedagogic technique for offering a hands-on experience of the subject matter, as well as teaching and practicing different methods and skills. Thus fieldwork is an activity that is carried out by both individuals and groups, as well as by experienced professionals and school and university students.

The sheer breadth of geography as the study of Earth and everything in it makes fieldwork very varied, from collecting ice cores on a glacier to interviewing children in an African village. Fieldwork involves dealing with landscapes, landforms, and environments; people, communities, and organizations; and statistics, images, and documents.

Although fieldwork is a defining feature of geography and an important mode of learning for students, it is not unique to the discipline. Fieldwork is practiced in many closely allied disciplines including anthropology, archaeology, biology, ecology, geology, and sociology, all of which go outside the classroom, library, or laboratory to collect data about things in the world. The distinctive nature of geography fieldwork lies in the core concepts of geography – space, place, scale – which frame investigations.

Although the field can be defined as a place where a geographer goes to gather information about people, places, or phenomena, raw data does not actually exist per se, it is created by the intervention of the researcher putting value on something as being of investigative worth. For example, water will flow along a river channel at a particular speed whether or not it is being measured by a researcher’s instrument, and forest villagers will collect firewood for their needs whether or not a researcher asks questions about their community resource management strategies. Thus data is shaped by the interaction between the researcher and the research subject, including how this relationship is established, conducted, and sustained.

This leads us to the very heart of the ethical dimensions of geography fieldwork: it is in the way that the researcher seeks, collects, analyzes, presents, and disseminates data that creates ethical issues to be negotiated and resolved. Furthermore, ethics not only relates to data on a specific research topic, but also to the wider moral and
ETHICS IN GEOGRAPHY FIELDWORK

social responsibilities of being a researcher in the field. This makes fieldwork not just a pedagogical or professional undertaking, but also something personal and political.

Key principles for ethical research

The key principles of ethical research practice are respect, beneficence, and justice (after the 1979 Belmont Report), and there are a number of practical ways in which these can be applied. It is important to note that these principles emerged in relation to research on human subjects but are also applicable to research on nonhuman subjects. The focus in the literature and in training tends to be how ethical principles relate to the relationship between the researcher and the research participants but there is also an ethical dimension to the relationship between the researcher and co-researchers or other professionals, the institutions that support and fund the research, the government of the home nation of the researcher, and the government of the nation where the research is carried out (if different).

“Respect” refers to the general attitude that the researcher should have toward the research subject and its wider context, thus it encompasses respect for people (both individuals and groups) and the organizations and structures set up by people, as well as respect for landforms, ecosystems, biodiversity, and natural resources.

A key aspect of practicing respect with human subjects is “informed consent,” which involves giving adequate information about the research aims, methods, and outcomes to potential participants, ensuring that they understand the implications of their involvement and gaining their voluntary consent to take part. Additional attention needs to be given to people from vulnerable groups or those with diminished autonomy (such as children or those with disabilities) to ensure that they or their proxy fully understand. Informed consent is also ongoing: participants have the right to withdraw at any time. Informed consent is a fairly uniform requirement with the exception of some covert research. In the case of nonhuman subject research in realms such as ecology, environmental science, and geomorphology, informed consent may involve dealing with the people who inhabit, own, or manage the habitats, resources, or landscapes being studied.

Another aspect of respect is the protection of privacy, which includes maintaining confidentiality (i.e., not sharing information about the research subjects), anonymization (i.e., removing names and other identifiers), and safe management and storage of data. Again, there are some exceptions; for example, some participants may not want to be anonymous and can be offered the choice if they understand the implications of being identified, and confidentiality can be breached if someone (especially a child) is at risk of physical harm. Privacy is just as relevant to the physical environment; for example, a study to map the location of a rare species should protect its location from public knowledge.

A further aspect of respect is cultural sensitivity, which means being aware of cultural codes, religious norms, and local beliefs so that the researcher is less likely to cause offense or distress by their behavior or attitude. While particularly pertinent for research carried out in a foreign country, this can be equally significant much closer to home.

The concept of “beneficence” refers to the welfare of the research subject being paramount. This means protecting them from any harm or distress that may be caused by participating in the research. Beneficence is usually considered in reference to the research subject, but protecting the welfare of the researcher themselves can be
just as important, particularly when the research involves dangerous locations, vulnerable people, or emotive matters. Beneficence is also relevant to protecting landscapes, environments, and nonhuman organisms.

Researchers seek to ensure beneficence through systematically assessing the nature and scope of both risks and benefits, and implementing measures to maximize possible benefits and minimize possible harms. Harm encompasses both the physical (e.g., physical injury, threat of violence) and psychological (e.g., stress, risk to reputation, loss of dignity, loss of self-esteem), and it is vitally important to consider the potential for harm both during the research and afterwards when the researcher may no longer be present.

The principle of “justice” means an equitable distribution of both the benefits and risks of research. This applies to all stages of the fieldwork process from fair procedures in the selection of research subjects, to ensuring that voices and perspectives are equally heard and represented, and the sharing of research results with individuals and communities affected by the research in a way that is appropriate and accessible.

Justice also involves reciprocity. What may finally result in a single-authored scholarly work may have been impossible without support and participation from funders, supervisors, gatekeepers, informants, interpreters, and many others. All such contributions to the research should be duly acknowledged and they should receive compensation of an appropriate kind.

In addition to these three core principles, ethical research practice also means complying with appropriate laws and policies (which vary by country and context) on issues such as data protection, environmental protection, land rights, criminal disclosure, and conflicts of interest.

Conducting research in another country can add further ethical considerations and requirements. For example, researchers should show respect for the host nation by complying with any legislation and procedures for outsiders conducting research (e.g., research permits) and seek to ensure that the research benefits the host country as much as themselves once they return home.

The key principles of ethical research practice—respect for people and places, maximization of benefits and minimization of possible harms, equity in the benefits and risks, and compliance with laws and policies—are pertinent to all geography fieldwork. However, there is little to suggest that ethics affect geographers any differently to anthropologists, ecologists, or geologists conducting fieldwork. Perhaps what makes geography distinctive is the sheer diversity of contexts in which geographers carry out fieldwork and the complex webs of social and physical interactions to which ethical principles should be applied. In addition, many geographers are engaged with contemporary issues—such as social justice to environmental protection—and may feel a moral conscience to use their work to make the world a better place; this blurring between professional and personal interests creates further ethical considerations.

Historical overview of ethics in geography fieldwork

During the eighteenth and nineteenth centuries, a diverse range of travelers, explorers, soldiers, scholars, missionaries, surveyors, social reformers, imperial diplomats, writers, traders, naturalists, journalists, and others observed, mapped, and collected information about different places, people, and phenomena. This can be loosely categorized as early geography fieldwork.

The dominant paradigm was a detachment between the researcher and the field. The field was “out there,” somewhere else. It was
often overseas and the public “at home” reve-
eled in reports of the distant and the exotic. When the field was domestic it was also pre-
sented as an “other place,” as exemplified by nineteenth-century social reformers who cat-
aloged and mapped conditions of poverty in industrial cities such as London.

Instruction manuals for travelers had been in circulation since the sixteenth century but proliferated during the nineteenth century. The Royal Geographical Society added to the assem-
blage with “Hints to Travellers,” first published in 1854 and with many subsequent revised and expanded editions. Contents included suggested equipment that a traveler should take and how it should be packaged; instructions for taking accurate measurements and drawing maps; and advice on appropriate clothing and medical care in the field. A more populist handbook called The Art of Travel was published by RGS Fellow, Francis Galton, in 1855. It contained practical advice on expeditions (such as making camp and purifying water) and suggested what equipment to take. The content of both publications focused on practical and methodological issues with simply no reference to suggested behavior in the field in relation to the local environment or people.

While motivations for fieldwork varied – from adventurous exploration to scientific investigation and from military surveying to religious crusade – the common thread in each was a utilitarian ethic, collecting whatever was useful for the particular purpose. Public interest in, and respect for, these endeavors lent legitimacy to further expeditions and field trips. The unquestioned right to do whatever they wanted and wherever they wanted was based on the assumed supremacy of Western civilization and the unshakable belief in science and progress. However, built into the relationship between fieldworkers and the field were power and hierarchy, prejudice and racism, that is clear in the language used to describe “other” people and places (e.g., uncivilized, savages, cannibals).

Analysis and reflection of (un)ethical practice in the field during this period has only been car-
rried out more recently, particularly in the context of postcolonial critiques. Today the behavior of these “early geographers” would be classified as gross malpractice. There are plenty of examples of taking local people as slaves, plundering vil-
lages, sexual exploitation of native women, and excessive violence. For example, on one of Stanley’s expeditions in Africa, the theft of a few items from his field camp was countered by shooting dozens of locals dead. With widespread demand from museums and anthropological institutes for human specimens and cultural artifacts, naturalists and ethnologists dug up skeletons from graves and took material objects without permission. It was also popular practice to take one or two “natives” back home to be publically displayed at exhibitions and circuses as live specimens of exotic and savage peoples.

While some showed utter contempt for the subjects of their investigation, others displayed some cultural sensitivity and respect in the field. Many were driven by a moral, sometimes religious, imperative to bring “civilization” and “progress.” The goal of some fieldwork and the publication of findings was to improve condi-
tions (e.g., urban poverty) or campaign against mistreatment (e.g., slavery). However, there sometimes seems to be a contradiction between the humanitarian mission and the imposition of power and authority.

Overall, it is difficult to make generalizations due to the sheer diversity of people conduct-
ing fieldwork, and their different motivations and attitudes. In the field there is a lack of evidence of ethical principles such as gaining the consent of participants, protecting them from harm, and fair representation of different views. Likewise, the majority of post-fieldwork
publications – whether formal reports, scholarly papers, or populist articles – describe methodology and findings but do not demonstrate ethical considerations or engage in critical self-reflection. However, all of this must be seen within the context of the times when there was a complete separation of science and ethics.

From the establishment of the academic discipline of geography in the nineteenth century through to the first half of the twentieth century, there were a succession of major trends – environmental determinism, regional geography, and the quantitative revolution. In each case the methodology of fieldwork and data collection was influenced by the prevailing epistemology, but a lack of reference to ethical principles continued indicating that it was just not part of research practice. For example, early guides for practitioners such as the “Outline for Field Work in Geography” (Jones and Sauer 1915) focused on practical issues and data collection methods with no reference at all to ethical principles or behavior. This does not necessarily mean that all fieldworkers were unethical in their practices but neither can it be claimed that they were ethical.

From the 1970s onwards new trends and topics in geography created a context for ethical issues to be considered. The emergence of critical geography and the humanistic turn, increased awareness of environmental problems and global issues, feminist critiques and postcolonial discourses, postmodern ideas, and participatory methodologies have all in some way challenged assumptions about Western hegemony in science and the production of knowledge, particularly contesting the ability of geographical investigations to be objective. In this context a considerable amount of reflexive writing has emerged by geographers questioning issues such as identity and positionality, politics and power-relations, voice and representation. These socially and politically situated understandings of research and knowledge production provided the context for a more ethical approach to geography fieldwork.

Geography was slow compared to other disciplines to develop ethical guidelines for research practice. It was seen as the responsibility of the individual researcher or teacher to ensure that ethical principles were upheld in fieldwork. Some learned societies and professional bodies produced codes of good practice but these were nonbinding. Among the ethical guidelines available to geographers are the Association of American Geographers’ “Statement on Professional Ethics” (2009), the American Geographical Society’s “Guidelines for Ethical Conduct of Foreign Field Research” (2009), and the Royal Geographical Society’s “Code of Practice for the Grant Programmes” (2006). Depending on their subdiscipline, geographers can also refer to guidance from the bodies of allied disciplines, such as the Geologists’ Association’s “Code for Geological Field Work” (1975), the British Sociological Association’s “Statement of Ethical Practice” (2002), or the British Ecological Society and Young Explorers’ Trust’s “Environmental Responsibility for Expeditions” (2002).

In addition to voluntary ethics codes, mechanisms such as research training, student supervision, and peer review provided a framework for research ethics to be taught and monitored, but, by and large, ethics in geography fieldwork was still self-defined and self-regulated until its more recent formalization in the ethics review process.

**Formalization of ethics in geography fieldwork**

The recent introduction of a system of external accountability for ethics in geography fieldwork needs to be contextualized in the emergence of research ethics in wider academic
ETHICS IN GEOGRAPHY FIELDWORK

and professional practice since World War II. While the ancient civilizations of Egypt, Greece, and Rome set down some ethical principles in relation to medical science, there were some appalling examples across the centuries of some very unethical medical practice.

The history of medical science is littered with examples of testing on people, often marginalized or vulnerable groups whose lives were perceived as having less value. For example, physicians developing vaccines in the late seventeenth and eighteenth centuries used slaves as their test subjects. During World War II Nazi scientists conducted medical experiments on thousands of concentration camp prisoners, inflicting wounds, transmitting infections, and administering experimental drugs so that the effects could be observed. Likewise, the Japanese used captured prisoners of war for testing biological and chemical weapons. In the United States, the twentieth century saw government-sponsored research into the effects of untreated syphilis in African American men and the effects of radiation on human beings using cancer patients, pregnant women, and military personnel as test subjects. In each of these cases the research was carried out either without the knowledge or without the consent of the participants.

The war crimes tribunal that charged 23 of the abovementioned Nazi physicians led to the establishment of the Nuremberg Code in 1948 which set out guidelines on the ethical conduct of research, emphasizing that human participants should give voluntary consent and that the benefits of research must outweigh the risks. This was followed by the 1964 Helsinki Declaration on Ethical Principles for Medical Research Involving Human Subjects. Under these codes, the responsibility for carrying out research in an ethical manner lay with the individual researcher and it was not obligatory until governments stepped in to establish regulations.

The United States was the first, in response to a series of scandals including the Thalidomide tragedy and the Tuskegee experiments. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research was established and published a set of ethical principles and guidelines in 1979 known as the Belmont Report, which became the standard for ethical review in the United States of America and beyond. The 1970s and 1980s also saw Canada, Australia, the United Kingdom, and other countries adopt the practice of ethical review of medical research by independent committee. This served to standardize various voluntary codes and provide a mechanism for peer and public accountability.

By the turn of the twenty-first century, the process of ethics review had expanded from the medical and behavioral sciences to human subject research across the social sciences including human geography. Many geographers are now required to undertake an ethics review prior to fieldwork. Formats and requirements vary by institution and country, but usually an Institutional Review Board or Ethics Committee will consider a researcher’s proposal and the measures put in place to ensure ethical practice. In some cases they may request a revision of methodology before giving permission for the fieldwork to go ahead. Some institutions also require an end of project ethics report which details whether ethical considerations changed during the research process, how they were addressed, and how this affected the research.

Some critiques of ethical codes and the ethical review process for geography fieldwork

The improvement of ethical standards in fieldwork should be lauded. Voluntary ethics codes
set by disciplinary bodies and funding organizations, as well as compulsory ethics reviews imposed by academic departments and educational institutions, have been implemented largely with good intentions. The professional practice of today’s geographers in the field is far removed from the invasive, exploitative, insensitive, patronizing, racist, and selfish behavior seen in previous centuries. Hopefully it has also improved the quality of scholarship and the reputation of the discipline. However, ethics codes, the teaching of ethics, and the ethics review process have not been without criticism for a number of reasons.

First, voluntary codes are seen as too vague while the ethics review process is too restrictive. Given the nature of the discipline, geography fieldwork is tremendously diverse in terms of subject matter, location, and methodology thus any ethics code applying to all members of the discipline will have to be broad in scope. However, reducing core principles to the lowest common denominator can have the effect of distilling their value and leaving them so open to interpretation that almost anything goes. Furthermore, concepts such as “integrity,” “justice,” and “transparency” are abstract ideals and not readily translated into practice in the field. On the contrary, the ethics review process can be very prescriptive but the principles and practices developed to govern medical research are not necessarily suitable for geographical research whose aims, methodologies, relationships, and risks are quite different. For example, informed consent is not always morally appropriate or practically possible for geographers conducting fieldwork in some settings, such as ethnographic studies, yet ethics review committees may insist upon it. This may be partly caused by the cross-disciplinary composition of ethics review committees where most of the members are not geographers and do not necessarily understand epistemologies or methodologies specific to geography.

Second, ethics is compartmentalized and presented as a discrete aspect of fieldwork. The process of ethics review is one aspect of preparing for fieldwork, a task to be ticked off the to-do list along with a risk assessment, testing equipment, booking travel, and so on. Similarly, in most fieldwork guides ethics is a single chapter among other issues to be considered such as research design and data collection methods. Likewise in student training, ethics tends to be one self-standing lecture or workshop in a series on research methodology. However, the reality is that ethics are integrated into every aspect of the fieldwork, before, during, and after. The literature, training process, and formal procedure for ethics approval should better reflect the integrated, cross-cutting, and enduring nature of ethics in practice.

Third, codes of good practice and the ethics review process tend to present ethical principles as a series of issues each with an attendant list of solutions, but the reality is not so clear-cut. The field is a constantly shifting environment where the researcher is interacting with the research subject and location. Ethics are messy, contested, and negotiable, requiring ongoing reflection and responsiveness. Even the most carefully prepared researcher can face ethical issues in the field that were not anticipated in the ethics review processes and which may require decisions not covered by ethical codes. To resolve an ethical dilemma often involves violating one ethical principle in order to preserve another. Thus the fieldwork process is a minefield of ethical “gray areas,” some of which may never be fully resolved. As every fieldwork situation is unique, ethics codes should be used as tools rather than rules.

Finally, while ethics codes and ethics reviews are now widely adopted, they are no guarantee
of ethical research. It is possible for researchers to meet all the standards on paper but to flout them in practice, and there is very little real punishment if this is discovered. Despite much talk about “accountability,” there is a lack of clarity in terms of accountability to whom. This stands in contrast to health and safety aspects of fieldwork which are now taken very seriously, with maleficence holding real consequences. Critics may say that the ethics review process provides a buffer for educational institutions and funding bodies in the case of misconduct, but while medical science, business, law, and other fields have faced cases of misconduct and malpractice, geography has been largely free from scandal and enjoys the public trust (perhaps apart from the recent criticisms of the American Geographical Society’s Bowman Expeditions). At the end of the day, without a system of guaranteed checks and punishments, the responsibility for ensuring ethical practice in fieldwork lies with the researcher’s own integrity and conscience.

Despite great advances in improving ethics in geography fieldwork, an endpoint has not been reached. As geography and geography fieldwork continues to evolve, new ethical issues will emerge. For example, new technologies such as remote sensing, GIS, and three-dimensional geo-visualization have opened up different ways of looking at, measuring, monitoring, and understanding the Earth and geographical processes; the insertion of a computer screen between the researcher and the field does not remove the ethical dimension but rather changes it. Likewise, the emergence of Web 2.0, social media, and mobile technology has enabled the use of citizen science to collect spatial data; but opening up of the category of “researcher” to anyone with access to the Internet poses difficulties of enforcing ethical practice. Meanwhile, virtual field trips now available to geography students via digital technologies also change traditional methods of real-life interaction between the researcher and the field, thus altering ethical dynamics.

SEE ALSO: Ethics in GIScience; Fieldwork in human geography; Geography education: fieldwork and contemporary pedagogy

References


Further reading


Considering that the possibilities of geographic information systems (GIS)/GIScience as an enabling science and technology of human endeavor are limitless, ethical issues do and will often arise in their development and use. Actual developments and uses, and potential developments and uses, lead to significant responsibilities and choices that involve considerations of values. Ethical issues connected to GIS/GIScience are inseparable from science and technology and involve considerations of scientific and technological as well as philosophical and practical aspects. Reflecting this range of issues, a high degree of intellectual and cultural diversity characterizes both academically published and unpublished work aimed at professionals and specific audiences on GIS/GIScience ethical issues. Throughout all engagements with ethical issues three questions repeatedly come to the fore:

1. What can I/we do?
2. What could I/we do?
3. What should I/we do?

By way of introduction, this entry turns to an underlying question in GIS/GIScience ethical discussions and interventions: the question about the scope of ethics and determinations of the philosophical degree deemed necessary for meaningful ethical engagement. Taken in reverse order, since the understood meaning of ethics defines its scope, the observation that philosophy is too important to be left to the philosophers offers a relevant starting point. Paralleling the range of philosophical enquiry, the issues that ethics encompasses can touch all imaginable pursuits of humankind. The question arises as to where one can start to consider these issues. This vast number of theoretical developments and uses, as well as applications, suggest that starting from a field of science and engineering can delineate a scientific and professional scope and result in a useful framework for organizing ethical engagements. While “meaning of life,” “moral rigor,” and even many “right and wrong” ethical and moral questions may lie beyond the resulting scope, the development of disciplinary ethical enquiry to study ethical issues develops helpful guidance for thinking about scientific and professional issues involving conduct, value-connected choices and responsibilities in one’s work, and responsibilities arising in scientific and professional contributions.

We can start from a primary question, and perhaps the most essential ethical question for scientists and professionals working with GIS/GIScience, of what I/we should do with my/our skills, knowledge, and technology. Our lives and the world are increasingly shaped by computer-based information technologies. But these technologies do not come from another planet or mysterious network of laboratories,
but are infused with human values. Still it often seems that information technologies lead to situations in which common moral rules and principles don’t seem to apply or are unhelpful in assessing what should be done. Spatial IT scenarios have rich ethical dimensions, as illustrated, for example, in such questions as whether retailers should be able to track store visitors using the unique code associated with their mobile phones; whether data collected by laypeople for a scientific project should be used for research and distributed without assessment of its positional and attribute accuracy; or whether GIS professionals should redistribute data from an analysis that relies on copyrighted data. The ethical aspects of imaginary and real GIS/GIScience scenarios provide practical references for what otherwise can become very abstract questions about what to do with skills, knowledge, and technology.

In other words, GIS/GIScience ethics emphasizes engagements with practices and draws only in limited degrees on theoretical ethics. This reflects the central importance in GIS/GIScience of assessing and analyzing the values that impact research and professional activities (Calkins and Obermeyer 1991). Going beyond the “rights versus wrongs” canard, applied approaches to ethics emphasize practical thinking about the consequences of what GIS professionals and/or GIScientists are doing, could do, or should do. These questions have an ethical component related to the value-based choices individuals and groups make when developing, using, and applying GIS/GIScience. The outcome and orientation of the questions and discussion, and later their actions, are often connected to an activity they individually face or could face. The degree to which people take up philosophical theories varies – approaches of practical ethics most often proceed from an implicit theoretical engagement and take up the practical challenges without theorization but can be very significant in improving scientific and/or professional activities.

The remainder of this entry on GIS/GIScience ethics proceeds by reviewing important academic and professional work on ethics associated with GIS, GIScience, geography, and computers. This is the basis for a summary of current issues in the following section and some concluding observations in the last section of the entry about developments of GIS/GIScience ethics. It aims to complement entries in the Encyclopedia on privacy, surveillance, and other issues with ethical dimensions.

Previous academic and professional work

Ethics offers powerful means to approach the roles of values in the developments and uses of GIS/GIScience. Previous academic and professional work has thoughtfully engaged the limitless breadth of ethical approaches and provides important touchstones for considering the ethical dimensions of GIS/GIScience.

A very good entry point in contemporary scholarship to consider GIS/GIScience ethics is the essay by James Proctor, “Ethics in Geography: Giving Moral Form to the Geographical Imagination” (1998). Proctor situates ethics in science and philosophy, suggesting that geography’s contributions to the study of ethics lie in the twofold consideration of geography’s ontology and epistemology. Distinguishing practical scientific questions arising in the conduct of science from moral questions about the justification behind different emphases in scientific fields and their role in the development of Western societies, Proctor accepted widely followed distinctions and separates theoretical ethics, systemic individual reflection on normative concerns, from applied ethics which focuses on specific moral
concerns. Since a major concern of ethics over the millennia has been the alignment of intellectual endeavors with the goals and ideologies of the regime, much work on moral philosophy deals with abstract means and ends questions and formulation of normative positions. Since the enlightenment and the rise of the modern state, the theoretical frameworks for philosophical ethics moved from teleological (ultimate purpose) to deontological (duty and obligation) approaches. With socioeconomic developments of the nineteenth century, work on ethical issues burgeoned and continued to do so through the twentieth century into the contemporary era. Proctor summarizes the ecumenical shift in ethical philosophy which includes the establishment of intellectual queries into gender, modernity, development, and community that connect to work in geography. Ethical enquiry in geography itself has largely taken the form of “thick descriptive ethics” (Proctor 1998, 10), which can be the foundation for much ethical enquiry in geography and bridge professional and substantive concerns. Proctor points to the importance of querying the ontological and epistemological dimensions and making sense of the central space, place, and nature metaphors in geographical knowledge-building. These queries become more focused in studying the technical implementation from the geographical concept. Because ethical issues pervade geography, Proctor concludes that considerations of geographical knowledge-building can offer valuable benefits for the discipline (Proctor 1998, 15). The questions, “What is the place of ethics in geography?” and “What is the place of geography in ethics?” point to important ways in which such endeavors can enrich the fields. Michael Curry’s contributions offer significant insights along these lines into the limits of GIS, questions for GIScience, and the reflections on the developments that come with the widespread use of spatial technologies and development of digital places and individuals (Curry 1998).

Jeremy Crampton provides helpful engagements with these ethical concepts and brings them to bear on important practical aspects. In “The Ethics of GIS” (1995) he points to a lack of substantial discussion of ethical issues in geography, which has led to the dominance of conduct standards and simplified normative concepts of good behavior. In response, Crampton develops a contextualized approach to GIS ethics that relates internal practices to external contexts. While resting on a normative concept of ethics, his dialectic approach to considering local relations and wider forces interactions helps to move past the local/global duality. Drawing on Brian Harley’s work, he refers to the different perspectives as internal and external, following Harley’s perception that cartographers were unwilling to accept that maps can follow an external and ideological agenda (Crampton 1995). Crampton in later work suggests turning to the question, “Under what circumstances is a map authored?” (Crampton 2001, 243). While more hermeneutic in emphasis than Blomley’s invocation of critical geography’s Marxist mission (Blomley 2006), Crampton’s later work develops and applies a Foucauldian concept of power to assess how maps align heterogeneous perspectives and thus offer the state enhanced means of control over its now calculable territory (Crampton 2010).

During this period, GIS was a rapidly growing field and going through a process of professionalization (Obermeyer 1994) which included addressing ethical issues (Huxhold and Craig 2003). Professional work on ethics had of course started earlier, but in parallel to academic discussions and with only limited exchange and interests in crossing between the two fields during this period. The creation of a Code of Ethics and Rules of Conduct (URISA 2003).
culminated in some senses the professional discussions of that time with a normative formulation of conduct standards (Craig 2003; Huxhold and Craig 2003). Neither code nor rules of conduct meant the end of professional discussion and refinements. For instance, a 2004 workshop convened by the Association for Geographic Information (AGI) in the United Kingdom sought ways to go beyond an implicit emphasis in codes and rules on sanctions and focus on guidance for professionals (Blakemore and Longhorn 2004).

Such events point to the ongoing discussion of ethical issues in formal professional and scientific associations and at professional and scientific meetings. As GIS has become an important information technology in other fields, the discussion of ethical issues has expanded into specific domains, yet in the context of changing societal values. All academic disciplines reflect larger societal values. Ethical issues in GIS/GIScience necessarily consider and refer to values created outside of GIS/GIScience.

One significant uptake of ethical issues for GIS and GIScience led to the creation of case study materials and guidance for using those materials in both education and professional development (DiBiase et al. 2012). Case studies offer scenarios for teaching that challenge students or participants to analyze ethical problems and to develop solutions. Case studies make it possible for professionals to connect didactic materials to practical experiences. The well-known “case method” is a common pedagogical technique for strengthening the moral reasoning skills of students in business, medicine, law, engineering, and computer and information science (Davis 1999; Keefer and Ashley 2001; Quinn 2006). Discovery and deliberation of case studies draws heavily on Davis’s (1999) “seven-step guide to ethical decision making” (outlined in Box 1). Similar ethical analysis models have been suggested by Keefer and Ashley (2001) and others.

While case studies provide an important didactic instrument for engaging GIS and GIScience ethical issues at a variety of levels, work from computing ethics offers numerous other pedagogical concepts and discussions of ethical issues that have relevance for GIS/GIScience engagements with ethical issues. The virtue approach, which involves considering concepts and testimonials from professionals who exemplify virtuous values in their work, offers students material with which to engage with the complexity of professional and scientific activities in their field and professionals ways to reflect on values and consider theoretical situations (Huff and Martin 1995; Huff, Barnard, and Frey 2008). It also offers practitioners ways to connect ethical concepts to their experiences and hypothetical scenarios (Huff 2008) and move beyond the relativism and sublimation that often stymie people engaging with theoretical ethics (Critchley 2007). The rich discussion of computing ethics offered by Deborah Johnson in Computer Ethics (2009) covers a gamut of issues arising in the development of new technologies in the vacuum of policies and amid the conceptual muddles that so often dominates our knowledge and understanding of GIS and GIScience.

In particular, the distinction she makes between the “standard account” and “socio-technical systems” approaches to engagements with computing ethics offers a number of highly relevant starting points to consider the development of engagements with ethical issues in GIS and GIScience. Perhaps the most important starting point is recognizing that technology and society co-shape each other. That is, ethical engagements with information technologies have to consider the complex and multidimensional relationships between technology and society. Indeed, while the use of computing technologies raises many ethical issues, including a good number of unique issues, many of these
Box 1 Seven-steps approach.

Step 1 State problem. For example, “There’s something about this decision that makes me uncomfortable” or “Do I have a conflict of interest?”

Step 2 Check facts. Many problems disappear upon closer examination of the situation, while others change radically.

Step 3 Identify relevant factors. For example, persons involved, laws, professional code, other practical constraints.

Step 4 Develop list of options. Be imaginative, try to avoid “dilemma”; not “yes” or “no” but whom to go to, what to say.

Step 5 Test options. Use such tests as the following. Harm test: does this option do less harm than alternatives? Publicity test: would I want my choice of this option published in the newspaper? Defensibility test: could I defend choice of option before Congres- sional committee or committee of peers? Reversibility test: would I still think choice of this option good if I were adversely affected by it? Colleague test: what do my colleagues say when I describe my problem and suggest this option as my solution? Professional test: what might my profession’s governing body or ethics committee say about this option? Organization test: what does the company’s ethics officer or legal counsel say about this?

Step 6 Make a choice based on steps 1–5.

Step 7 Review steps 1–6. What could you do to make it less likely that you would have to make such a decision again? Are there any precautions you can take as an individual (announce your personal attitude on such questions, change job, etc.)? Is there any way to have more support next time? Is there any way to change the organization (for example, suggest policy change at next departmental meeting)?

Source: Davis 1999.

issues cannot be separated from complex ethical issues arising in the use of all technologies that have ever “stirred fear and apprehension as well as fascination and hope” (Johnson 2009, 6).

In what she refers to as the “standard account,” IT creates new possibilities for individual action and new forms of interactions. The focus of work on ethical issues is on these new possibilities through questions about who, what, how, and when. Ethical analysis can consider questions of what should be done and help evaluate how the possibility fits practices, values, cultural concepts, and so on. Computer ethics thus focuses on filling policy vacuums, for example in copyright laws that proceed from an assumption that all copyright material is provided on physical media, thus creating a vacuum for streamed movies and downloaded spatial data.
ETHICS IN GISCIENCE

In contrast, the “socio-technical systems” approach focuses on the co-shaping of technology, people, and society and rejects technological determinism. Strongly influenced by science and technology studies (STS), this approach’s rejection of technological determinism means connecting technological developments to societies and ideologies and looking at how technology, people, and society co-constitute each other. Thus, instead of a “standard account” analysis that may look for parallels arising in the adoption of Internet-based communications in each country that implements Internet technology, the “socio-technical systems” approach considers how societal influences alter the design, implementation, and use of Internet-based communications. For instance, the concerns about location privacies expressed by many people in Germany reflect a cultural history of state surveillance. In Brazil, in contrast, tracking is felt by many to offer a new level of security for individuals and families. Perhaps the biggest difference between the two approaches is that “socio-technical systems” avoid focusing on material technological objects and resulting impacts and instead consider how technological artifacts are embedded and created by intentional social activity as a social product that is also part of a social system. For example, a cell phone is not just the phone, but a combination of artifact and a network of people and other technologies.

This aspect of the “socio-technical systems” approach leads to significant consequences for engagements with ethical issues. Traditionally, ethicists have focused on human behavior and action and left out technology. Technology was largely understood to be neutral and its use depended solely on decisions of a person, group, or institution. The resulting emphasis on the material creates obstacles to understanding how artifacts and people interact, how people influence artifacts, and how artifacts influence people. Values are infused in the technology. The social practices that Crampton points to (2001) are integral to ethical engagements of GIS and GIScience.

An example of the “socio-technical systems” approach to this can be found in applied ethics (Harvey 2012). Applied ethics is an area of philosophy that approaches ethics with the idea that philosophy has to be relevant and engage actual issues for an audience that includes the public. Among philosophers, the public, scientists, and professionals, applied ethics has received considerable attention in recent years. Applied ethics echoes the principles of the “socio-technical systems” approach in its emphasis on engaging actual issues and being relevant to a larger public. Reflecting a scientific emphasis on moving past intellectual reflection, the emphasis of applied ethics is on developing an empirical foundation for ethics. Instead of looking for ethical theories and principles to explain choices, applied ethics involves considering both ethical concepts and the actual situations with actual consequences. In other words, ethical engagement involves clear connection to facts. For example, a GIS analyst working on a mapping project using census data needs to consider requirements of the project but also the larger situation.

Applied ethics involves considering the situation and internal and external relations and facts. Beyond these considerations applied ethical deliberation also reflects on the perspective and assumptions of different actors and ourselves in analyzing a situation – what many feminists refer to as standpoint theory (Walker 1998).

Persistent issues for GIS/GIScience ethics

GIS and GIScience have introduced, developed, and refined powerful technologies, skills, and
knowledge for the representation of geographic processes and patterns observed or measured in the world. Both professionals and scientists draw on ethics to help make decisions arising in the plethora of issues complicated by the vacuum of policies and the conceptual muddles that accompany the increasing ubiquity of location technologies (Sui 2004; Wilson 2012). These involve new applied ethical challenges.

While a particular application involves the specific concerns raised in other entries in this encyclopedia, several ethical issues of GIS and GIScience are of broad significance. Four such issues that deserve consideration here are:

1. the question of place and the geographical dimensions of ethical issues
2. ethical concerns with representation
3. spatial analysis and ethical responsibility
4. consideration of values in professional and scientific activities.

Place has been a central metaphor for geographic knowledge production. Informing the ontological scope of geography, it is used to organize geographical knowledge and make connections between materialist concepts of space and the places people live in and experience. Ethical issues arise in living and conceptualizing places. In geographic interventions of Bunge, Olsson, Wood, and others, whether ethics are place-bound or fundamentally universalist is questioned. The ethical issues arising in GIS and GIScience involve a great diversity of places and leave much latitude in interpreting this question.

Likewise, ethical concerns with representation arise frequently in making decisions in the design, implementation, and use of GIS and in advancing GIScience. The spread of GIS and its growing accessibility mean more people are thinking about representations of territory and questioning the canonical representations produced by state agencies and other actors. Issues of accuracy also reflect ethical concerns with representation and the need to assess whether spatial data is sufficient – the fitness for use criteria (Chrisman 1984; Comber, Fisher, and Wadsworth 2004; Goodchild 1998).

Similarly, the broadening use of spatial analysis leads to concerns with the ethical responsibility arising in the use of GIS and GIScience developments. It increasingly behooves GIS professionals and GIScientists to understand how data are collected, avoid the pitfalls of objectivism, and account for related ethical risks. Marc Armstrong draws on deontological ethics and the legal concept of malpractice which arises in the failure to follow accepted professional standards through negligence when the professional is required to act (Armstrong and Ruggles 2005). While GIS misuse rarely results in grievous consequences that directly endanger people, the consequences of slapdash work or unquestioning use of data can result in analysis results and information that lead to negative consequences and risks.

Accordingly, Armstrong and others argue for training and continuing education focus on consequences that arise from choices that can change results and thus sway conclusions. These issues are of course diverse yet also specific, making them a key focus of more recent attempts to advance ethics training (DiBiase et al. 2012). Case studies of various lengths and the seven-step method provide ways to learn how to incorporate values into GIS/GIScience activities and also provide a method for addressing ethical dimensions.

Conclusions

As the enabling science and technology in geography and many other fields, a primary, and perhaps the most essential ethical question about GIS/GIScience for professionals and scientists working with it is that of what
ETHICS IN GISCIENCE

I/we should do with our skills, knowledge, and technology. Given the limitless potential of GIS/GIScience’s support for human endeavors, opening the underlying issues, assessing and developing different pedagogical approaches for students and continuing professional development, and providing counsel have been the significant contributions of GIS/GIScience ethical engagements.

This entry provides an overview of these issues and approaches to engage ethical issues that GIS professionals and GIScientists have developed. Given actual and potential uses of these skills, knowledge, and technology, the more widespread distinction of theoretical ethics that deals with systemic and abstract reflections and applied ethical engagements with specific moral concerns degenerates into conduct standards and simplified normative concepts associated with maxims that become challenges to realize in the real-world complexity of science and technology. Instead, considering the geography of knowledge-building (Proctor) facilitates the engagement with actual situations and the co-shaping of GIS/GIScience socio-technical systems.

The ethical issues connected to GIS/GIScience accompany the developments in these domains. And they may even, from time to time, become the center of interest. The scientific, professional, technological, philosophical, and practical aspects of these ethical issues are ever present in a complexity matching the potential uses of GIS/GIScience. Considering the ethical dimensions of scientific and professional conduct, values in work-related choices, and responsibilities for the use of outcomes and products moves potentially abstract discussions into three questions:

1. What can I/we do?
2. What could I/we do?
3. What should I/we do?

Engagements with these questions and the situations, internal relations, external relations, and facts make up a common thread in GIS/GIScience ethics discussions and will continue to be a part of these fields.

SEE ALSO: Critical GIScience; Environmental ethics

References


Ethnicity

Emily Skop
University of Colorado Colorado Springs, USA
Wei Li
Arizona State University, USA

Ethnicity is a social construct that divides people into varying groups based on characteristics such as ancestral geographic and religious links, a shared sense of belonging, cultural traditions, values, political and economic interests, and historical ties. At the center of ethnic identities are the building blocks of language, dialect, faith, literature, folklore, music, food preferences, and clothing, along with kinship, neighborhood, community, and territorial links, and/or migratory status. These identifiers can serve as important referents to self–other identification, and ethnicity may influence both individual and group behavior in various areas of social life. This is especially true in multiethnic societies where the ethnic tie can be significant in shaping both primary social relationships, including one’s choice of close friends, marital partner, and residence, as well as broader collective political relationships, whereby a strong group consciousness can result in benefits and privileges within prevailing social systems.

While ethnicity can be a key source of social–psychological attachment, ethnicity is not the only way in which individuals see themselves and how others view them. Ethnicity typically interacts and intersects with other social constructs such as race, gender, class, social status, sexual orientation, age, and disability to form mutually constituted identities that cannot be examined in isolation from one another. At any time, an individual has multiple identifiers; for example, a Somali immigrant, mother-in-law, friend, and Christian may also be an office manager. Each one of these identifiers may be more personally relevant depending on the situation, but when they intersect they can profoundly affect individual lives.

Most scholars argue that certain identifiers are more dominant and powerful in society, and that these identities need to be studied together to fully understand how intensely they impact everyday life. Studying race and ethnicity in concert, for instance, is a key way forward to understanding how whole systems of privilege are granted to particular sectors of the population at the expense of other sectors. Since racial identification obscures ethnic group identification and vice versa, individuals and groups are placed in multiple social positions, which in turn become institutionalized in systems of inequality. In the United States, for instance, focusing only on race would hide the important differences between and within Chinese, Asian Indians, Filipinos, Koreans, or Japanese – all of which are Asian American ethnic groups. Then again, identifying individuals or groups solely by their ethnic identities would also obscure the relevancy of race in society. Thus, the mutual construction and intersection of these identities demonstrates how different people relate to one another based on varying social interactions.

Indeed, when varying social identities intersect, they form matrices of privilege and oppression whereby different social locations, based on ethnicity, race, gender, class, sexual orientation, age, disability, and other identifiers, combine
to shape individual lives and create systems of inequality. Experiences of oppression and privilege create social differentiation, whereby people come to be identified and ranked against one another based on physical and social characteristics. The combined aspects of one’s identity can either intensify (through multiple oppression) or buffer (via privilege) the full force of oppression that individuals and groups of people experience.

Because of the costs and benefits that can result, ethnic identities will be displayed to a varying degree by individuals and groups based on both in-group and out-group responses, and reflected at both individual and group levels. Ethnic identities can be self-claimed or externally imposed, depending on circumstances. Individuals typically self-identify as a member of a particular ethnic group to assert their sense of belonging, to maximize the benefits associated with this identity, to retreat from a group to minimize the disadvantages, or to disavow an ethnic identity for fear of persecution. However, it is the externally imposed ethnic identity that reflects broader societal impacts. For instance, using “Asian American” as the basis for categorization may result in “racial lumping” by the dominant society. In turn, population groups caught in this system of ascriptive racial categorization may react by rallying together to (re)create emotional links and form alliances based on their ethnic ties. In other words, ethnic identity formation may be the result of a group’s attempts to withstand structural disparities or to gain privileges that might otherwise be denied to them based on other social identifiers.

The end result of social differentiation is social stratification, which creates unequal access to valued resources, services, and positions in society. Discrimination, residential and occupational segregation, and economic isolation, as well as educational and health disparities, are the result of social differentiation. Indeed, matrices of privilege and oppression based on particular identifiers dictate the allocation of justice, the distribution of societal benefits, indeed the entire spectrum of life chances. In extreme cases, differential ethnicity is used to justify ethnic cleansing or wars against humanity. Ethnicity is one of the key elements used in social stratification, and thus it has very real consequences in people’s everyday lives.

While in the past essentialism was the theoretical model that dominated discourse surrounding ethnicity, the social constructionist model is currently the chief theoretical framework that structures the idea of “ethnicity.” Essentialism boiled ethnicity down to hereditary, biological, and primordial ties but social constructivism illustrates how individual and broader societal action and inaction institutionalize ethnic identities in particular places and situations. In the social constructionist model, individuals typically have multiple social identities which they use to define themselves, at the same time as dominant societal values play a critical role in the construction and classification of social groups. In other words, societal notions of what is considered normative, normal, or accepted frame personal identities at the same time as personal identities can reshape dominant ideologies through reframing stereotypes.

Indeed, social constructivism argues that differentiation and stratification vary across cultures and through time, so that the treatment of individuals and groups at both the individual and the institutional level changes, depending on context. In other words, it is possible to tell if some identifier like ethnicity is socially constructed if that same identifier is framed differently in another time frame or culture. In the United States, for instance, many groups that are now defined as ethnic groups, including most European immigrant groups, were in previous historical periods defined as races. The Irish,
Jews, and many other European groups came to be known as “ethnics” only in the early twentieth century and were thus identified by those of the dominant Anglo-Saxon ancestry groups, who often used the term pejoratively to distinguish and separate those groups from them. In fact the term “ethnicity,” or the condition of belonging to an ethnic group, did not even appear in scholarship until the early twentieth century, thus illustrating the socially constructed nature of this particular identifier.

Within geography, studies of ethnicity have reflected scholarly paradigm shifts over time. Earlier geographic research on ethnicity was strongly influenced by the Chicago School of sociology, with a focus on (i) the “race relations” cycle whereby all groups eventually assimilate to Anglo-Saxon Protestant American norms; and (ii) the “invasion–succession” model whereby a newly arrived ethnic group invades an inner city neighborhood and ultimately succeeds the group living there as that group moves outward toward the suburbs.

In recent decades, geographers have increasingly analyzed ethnicity from the perspective of dynamic structural opportunities and restraints, with a focus on antiracist and racialization theory, instead of the previous essentialist angles of the ascribed linguistic and cultural traditions alone. The retheorizing of ethnicity, in addition to the multiscalar analysis of individuals, families, groups, neighborhoods, city-regions, nations, and supranations, have reinvigorated the field of ethnic geography and made such research increasingly influential.

The early twenty-first century brings a new complexity to the scholarship surrounding ethnicity. As the cross-national movement and connections of people, goods, information, and financial capital becomes more prevalent, the idea of transnationalism has come to the fore. Transnationalism, the process by which migrants forge and sustain multistranded social relations that link together societies of origin and settlement, provides an alternative mechanism through which migrants forge their identities. If in fact individuals have multiple matrices of privilege and oppression to navigate, what will be the result for ethnic identities? One might argue that, in a world of globalization and diminishing spatial barriers to communication, movement, exchange, originality, exoticism, and authenticity will become prized possessions, and thus ethnic identities will become even more important and salient in the coming years. Yet others might contend that the entire process of social differentiation will become unhinged and that current mappings of ethnicity will be destabilized.

Whatever the outcome, the consequences of future ethnic geographies are likely to be remarkable. Indeed, the study of ethnicity may advance the goal of building toward more just and peaceful societies in the world. Therefore, future research on ethnicity should be multifaceted and multiscalar, including but not limited to change over time and across space, considering variations as a result of class, gender, age, generation, or migration status, and analyzing different subnational, national, supranational, and transnational contexts.

SEE ALSO: Antiracist geography; Citizenship; Diaspora; Difference; Identity; Inequality; Intersectionality; Migrant settlement; Multiculturalism; Race and racism; Social constructionism; Transnationalism; Whiteness

Further reading

**ETHNICITY**


Ethnobotany

Robert Voeks
California State University, Fullerton, USA

Ethnobotany is the study of the dynamic relationship between plants and people. It traces its early history to the colonial quest for precious spices – cinnamon, clove, nutmeg, and other tropical treasures. As colonial merchants and settlers unwittingly spread tropical microbes to the far corners of the Earth, the search expanded to botanical remedies for the flotilla of new and geographically expanding diseases. Entrepreneurs and men of science sought fame and profit as they scoured these newfound lands for “green gold,” and so was born the ethnobotanical enterprise.

Ethnobotanists continue to search for novel botanical products, especially drug plants. These bioprospecting efforts have yielded considerable pharmaceutical success, but often at the expense of indigenous intellectual property. Beginning in the 1960s, the primary objectives of ethnobotanical research became more conceptual and problem oriented, as researchers shifted away from useful plant inventories to explore the complex relationship between people and plants. Ethnobotanical inquiry likewise became more interdisciplinary, involving especially anthropologists, archaeologists, botanists, chemists, and geographers. Laboratory and field methods were adopted from cognate disciplines, while data analysis became increasingly quantitative. The time frame of inquiry ranges from the late Pleistocene to the present, while the scale of investigation ranges from genes to landscapes.

Ethnobotanical research involves any dimension of people–plant relationships, but most current research focuses on several “projects.” Ethnobotanists seek to understand the process of domestication and diffusion of crop plants. They study the strategies employed by traditional societies to manage plant species, and the degree to which these actions are environmentally sustainable. They investigate how species are cognitively categorized, ranked, named, and assigned meaning. And they explore the impact of globalization on knowledge and use of plant resources.

The origin and spread of agricultural crops had long been a focus of interest, but in recent years there has been an expansion of inquiry beyond the big commercial staples – rice, maize, wheat, and potatoes – and toward understanding the patterns and processes of lesser known “wild” and semidomesticated species. Management involves selective wild harvest, sparing of perennials, pruning, weeding and fertilizing wild populations, burning competing vegetation, distributing seeds, and protecting important taxa through community regulation. Species managed by these strategies often exhibit significant differences in seed size, seed number, and other morphological features compared with unmanaged populations. In Mesoamerica, for example, 5000–7000 plant species are exploited by people, most are in some way managed, and roughly 200 are in some stage of incipient domestication (Casas et al. 2007). When wild species are introduced to home gardens, even more intense management occurs. Home
gardens are crucial resources for rural people, especially women, who use them for food, medicine, fuelwood, fiber, and shade, as well as for ornamental and ceremonial purposes. They are important repositories of rare wild species, particularly as their natural habitats disappear, and they are hotbeds of incipient domestication. These studies suggest increasingly that plant domestication should be seen not as a binary beginning and endpoint, but rather as a continuum, from totally wild species to highly modified modern cultivars.

Species management is often associated with intense manipulation of nature, and so investigation of species domestication often segues into the question of the “pristine myth.” Due to long-term encouragement or sparing of useful tree species, many so-called natural landscapes are better termed “cultural landscapes” or “domesticated landscapes.” The ancient Mesoamerican Maya, for example, planted old-growth fruit trees in fallows and created orchard gardens (*pet kotoob*) in the forest. Consequently, many of the otherwise pristine forests of Guatemala, Belize, and Mexico’s Yucatan today are dominated by useful wild trees that were planted and tended by people over a thousand years ago (Ross 2011). Amazonian forests likewise exhibit significant species distribution and composition features that are consistent with human management. To this day, indigenous people clearly distinguish between the varying utility of forest stands that are more or less anthropogenic. Even the distribution of the endemic Brazil nut tree (*Bertholletia excelsa*), which for many is iconic for the Amazonian forest primeval, is the result of centuries of movement and likely planting by forest folk.

As pointed out early by Carl Sauer, the use of fire to manage species and vegetation was widespread in temperate latitudes. Recent ethnobotanical research has confirmed many of these observations. In California, large expanses of oak woodlands were burned by native peoples to reduce insect predation on acorns and to reduce competition from invading conifers. In the southern Andes, monkey puzzle (*Araucaria araucaria*) was managed by fire to enhance its growth and productivity. Aboriginal peoples of southeastern Australia employed burning to enhance the growth of edible herbaceous plants at the expense of trees. And Canada’s First Peoples used fire liberally to enhance the quality and quantity of edible berry and tuberous species.

Rural subsistence farmers often maintain huge lexicons of plant names. But while they universally name and classify familiar species, all taxa are not treated equally. Staple crops, such as corn (*Zea mays*) and potatoes (*Solanum tuberosum*), are specified by dozens of subspecific epithets – the blue type of corn, the red type of corn, and so on. But less useful species often receive simply a nondescript gloss – it’s a weed, or it’s a liana. Species salience appears to be the primary naming feature; manioc (*Manihot esculenta*), for instance, is more salient to Amazonian subsistence horticulturalists than a rare vine that invades the swidden. Traditional societies organize their ethnofloras into ranks, often including kingdom, life form, genus, species, and variety; ranks tend to be similar across cultures, regardless of cultural significance (Berlin 1992). The genus in particular represents the most recognizable rank of plant perception among humans, and in most classification schemes finds its parallel in the Linnaean system.

The field of ethnobotany developed around understanding the relationship between indigenous people and native floras in rural, often isolated settings; the more sedentary and the more natural, the better. Yet in spite of preconceptions that people lose their ethnobotanical skills in the process of migration, it is clear that immigrants go to extreme lengths to continue using their traditional foods and healing
floras. This is particularly true when these plants represent distinctive cultural markers for oppressed diaspora communities. In the case of enslaved Africans, for example, there was a powerful impulse to recreate a semblance of their ancestral flora in the Americas. And in spite of the enormous barriers, significant elements of their useful flora arrived in the New World as part of the Columbian Exchange. This botanical homogenization facilitated continuity of African plant-based culinary and healing traditions among their New World descendents, thus providing avenues of cultural resistance to Euro-American hegemony (Voeks and Rashford 2013). Migrant people today likewise encounter considerable obstacles to international transfer of useful plants. Many avoid customs restrictions by surreptitiously bringing plants or seeds with them. In some instances, particularly with medicinal plants, the homeland species are weedy and cosmopolitan, and so often precede the arrival of immigrants. When other strategies fail, immigrants either learn the identities of replacement species in their new home, or simply abandon the use of useful plants in favor of supermarkets and pharmacies.

Subsistence harvest of traditional plant products in many cases has shifted to commercial extraction. Exotic foods, fiber, medicines, latex, and sundry other plant and fungal products – termed non-timber forest products (NTFPs) – are finding ready markets in the more developed world. By the 1990s, many nongovernmental organizations (NGOs) and government agencies began to view NTFPs as a potential rural development-resource conservation win-win. NTFPs can open routes to livelihood improvement among marginalized, rural communities, and, unlike alternative destructive forms of land use, the effects of petty extraction on the structure and function of forests is much less than other uses. Research on this subject is ongoing, and the results are mixed. In Latin America and Africa, earnings represent an economically justifiable use of gatherer time, but this is not the case in the poorest countries of Southeast Asia. Regarding ecological sustainability, most studies report that current levels and intensities of harvest are not threatening the biological replacement of individuals and populations, nor is the ecological integrity of the relevant ecosystems threatened. Regardless, commodification of previously subsistence plant products is dramatically changing many aspects of people–plant relations. Young people abandon all but commercially valuable species, gendered relations with nature change as women and men redefine their roles in the community, and the cultural significance of plants shift as botanical products are increasingly defined by their monetary value (Shackleton and Pandey 2014).

SEE ALSO: Agricultural environments; Agricultural geography; Biodiversity; Globalization; Nature conservation; Sustainable development

References


Shackleton, Charlie, and Ashok Pandey. 2014. “Positioning Non-Timber Forest Products on the
ETHNOBOTANY


Further reading

If you were to snap a picture of the stereotype of the ethnographer, it would be a landscape of his or her “field site”: perhaps a jungle clearing with thatched huts, and the lone field worker, writing inside a journal in the tent. Specifically relevant to the geographer-ethnographer, I think also about the image of a boat reaching ashore, the explorer discovering a “new” land. If not a jungle, then the boat is arriving on windswept rocky shores – or slides through an icy landscape of the polar regions. Sometimes those scenes have other, “local” people in them – but more often those landscapes are empty of everything but non-human biota. Ethnographers write the world in text and visual imagery, not without extensive field work – while also referencing the geographical imagination of the culture who will consume the ethnography. All knowledge is partial. By the stereotype’s definition, this means an ethnographer studies places and phenomena that are not within the researcher’s own home. There is a Far Side cartoon in which the viewer is privy to the living room conversation of a tribal family – who, upon seeing a canoe with people disembarking, cry “Anthropologists! Anthropologists!” And rush about the room, hiding their television and high-tech video equipment. The joke is of course playing on these stereotypes, and implies that the researcher is hoodwinked by their own fetishization of the “primitive.” I think of these stereotypes in my current research situation, in a small village of Athabascans in Nulato, Alaska. I hoped that my tent, situated for the summer outside the Chief’s front door, was merely repeating the placement and not the position and practice of the ethnographer in this community. In the first place, I told the Chief that what I hoped was to collaborate on a project as equals, that he as tribal leader could direct the purpose of the ethnographic research. Now I have left to wonder what sort of cartoon could be crafted about me. (Memo 3 on positionality, July 2, 2011)

Ethnography is both a method of data gathering and a representational practice of the qualitative scholar. Within geography it is often used as an inductive approach to empirical and emotional data, especially fieldwork data. Its attention to holism is perhaps a reason why geograph-ers, who literally “write the world,” are attracted to using ethnographic methods. Described by anthropologists as a practice of “thick description,” ethnography has been used to elucidate phenomena and causal relationships at scales such as a “culture” or “community,” particular places, or a particular event. Within geography, ethnography has been used to understand how humans create and relate to their material environments, as well as how power relations create and sustain social spaces, networks, and the literal movement of material bodies. Current trends in geographic theory, including critical geography and the “postmodern turn” across the social sciences, have influenced the topics, scope, and practice of ethnographic work. This entry outlines both the continuities and the changes in how ethnographic methods are employed by geographers.

The data used in ethnography comes from primary data gathering, often taking the form of field notes or visual and/or audio records of their fieldwork and/or participant observations. Ethnographers also generate data via interviews, surveys, photographs and other visual data, archival research, mapping, or other qualitative...
methods. An ethnographer’s primary sources are often complemented by secondary sources, such as census or geographic information systems (GIS) data. Nonspatial data is often analyzed discursively or via a coding scheme, with this textual analysis shaping the final ethnographic narrative – the “results” of the inductive study. There are a multitude of sources and analytical approaches available to ethnographers; many of these sources and approaches are helpfully explained through texts such as Crang and Cook (2007), Watson and Till (2010), and Herbert, Gallagher, and Myers (2005). As shown through their and other examples, this means that ethnographies can take a variety of forms, from a variety of research practices.

Via fieldwork, Herbert (2000) writes that ethnographers are able to understand what people do, because people cannot always consciously describe their actions and beliefs. Ethnographic research can also access the lived realities of populations and places that cannot be well understood simply by looking at aggregate quantitative social data. Aggregate survey data measuring the efficacy of public transportation, for example, may erase the problems experienced by minority groups, such as the specific transportation difficulties faced by poor women in urban areas. There are many kinds of experiences that are better accessed via ethnography than through other (quantitative or qualitative) methods; through the use of ethnography geographers can make sense of their literal “field” of inquiry, and generate new research questions that can be pursued through further qualitative or quantitative research.

Importantly, while there is no single way to conduct an ethnography, it is nevertheless possible to ascertain a good- from a poor-quality ethnography. For researchers do not merely “collect” data; as with every research methodology, ethnographers have developed rigorous approaches to analyzing their data and representing results. In the first place, ethnographers engage theory in their fieldwork differently from physical geographers; an instructive comparison is found in the chapter by Herbert, Gallagher, and Myers (2005), where geographers of different subfields describe their use of theory in the conduct of fieldwork. The physical geographer continually measures the “field” (via formal or informal hypothesis testing) against a larger body of theory of physical systems – often called an “objective” approach to the empirical world. In contrast, the ethnographers in Herbert, Gallagher, and Myers (2005) and elsewhere describe a process of using theory to inform their field practices as well as to make sense of their data, and they also describe their iterative approaches to analyze field experience while it was happening.

In this way, ethnography can be said to be both iterative and “intersubjective” – and thus the process of memo-writing and field note taking at large is a key way ethnographers maintain rigor in their work. As shown in the example memo at the beginning of this entry, and explained elsewhere (Crang and Cook 2007; Watson and Till 2010), a memo does not have to be a direct observation with the five senses, but more importantly it is a reflection on theory, method, and/or the researcher’s own participation in gathering data. A memo may record some ideas or observations quickly, or discuss at length the issue under focus and develop preliminary analyses. Memos, like all field notes, may be coded for patterns of thought and/or developed into a discourse analysis. They may be quoted in a finished piece of writing, like the aforementioned memo, as a sample of ethnographic data. Quoting memos and field notes is also a way to communicate or explain an issue in ways that are more accessible to multiple readerships, given that well-written (and first-person) field notes can at times read like a novel.
Rigor in field note practices includes constantly examining the researcher’s role in the collection of empirical data – called an exercise in “reflexivity.” One example of ethnographic reflexivity in geography is the collection of feminist research edited by Nast (1994) in the *Professional Geographer*. As demonstrated in that collection, to be reflexive a researcher must examine how they are sometimes “outsiders” or “insiders” in relation to a particular phenomenon under study – often their positionalities remain *in between* the “field” and “home.” The assumption is that researchers have an effect on the data regardless of the method, whether qualitative or quantitative, inductive or deductive. Thus writing about positionality provides rigor to an ethnographer’s empirical representation of the world; both the researcher and the reader would be made aware of the perspective through which the researcher has recorded and analyzed the empirical data. But, to be clear, these reflections are not mere personal opinions or journalistic observations. Rather, to be reflexive involves not only assessing the researcher’s emotions and personal experiences in relation to the issue/object being observed, but also engaging with theories of research ethics and positionality. Therefore, the actual act of being reflexive in the act of taking field notes allows a researcher to perhaps change their practices to more successfully collect data throughout the remaining time on a given project. It is a method for the ethnographer to check their own assumptions and to forge ways to build trust with the communities within which they work. The modern ethnographer does not want to be duped by his/her own desires, as the researchers in the cartoon were.

Reflections on positionality evolved to respond to what is called the “crisis of representation,” which greatly affected work across social sciences. As noted in Nast (1994), researchers began to recognize the “politics of representation”: where the *etic*, or “objective,” perspective was revered as the ideal of all Enlightenment-based sciences, ethnographers especially had to confront how such a practice can erase local expertise, called the *emic* perspective. The Western scholar, in accruing expertise on a community as an object of knowledge, thereby also disenfranchised that community from speaking for themselves in political life.

Watson and Till (2010) describe how geographers thereby became concerned with “decolonizing” their methods. To be seen as legitimate experts, ethnographers have traditionally emphasized the long-term nature of their projects – long periods of time to gather and interpret data, to immerse themselves in the lives of those they work with, and to reinterpret data. Important to this practice is an assumption that the ethnographer was studying an *other* society – via a more *etic* perspective – and thus required the time to immerse themselves in their new milieu. While the time of immersion is still of key import to the rigor of ethnographic work, today’s theory and practice of ethnography does not always require extensive temporal commitment in the field. In fact, a rigorous ethnographic practice today may emphasize less time spent in the field observing, and more time instead *collaborating* with those who hold an *emic* perspective – regardless of whether the ethnographer is “from” that context.

Because of the “crisis of representation,” geographers began to use traditional ethnographic methods instead to “study up” in elite Western societies. For example, some ethnographers studied practices of the state in its administration, in its policing, or in creating museums. Rather than focusing exclusively on other or minority populations who had less ability to represent themselves, many ethnographers studied up to access the lived experience of the “dominant” culture of which these researchers were often a part.
Meanwhile, geographers have also contributed to changing ethnographic practices of representation to continue to work with those people once constituted as “subjects” of research. Based on the assumption of the partiality of all knowledge, critical researchers from the fields of feminist and indigenous geography began to develop more deeply collaborative methods (Benson and Nagar 2006), and publish ethnographic accounts co-authored with those heretofore subjects (Watson and Huntington 2008). In this and other ways, the practice of ethnography has evolved in relation to the spread of postmodern and critical ideology and the evolution of participatory methods of research.

The practice of collaborative ethnography emerged also in response to traditional limits of participant observation. Instead of being limited from studying large-scale events (like a global environmental conference), ethnographers can create a model of team-based research similar to that of their physical geography counterparts. For a team of ethnographers who carefully coordinate their observations not only provides more data for a project, but the team model also provides structured checks on the data.

The ethnographers’ objects of knowledge have thus changed in tandem with changes in geographers’ objects of knowledge. Anderson (2012) describes how recent ethnographic works by geographers do not have to treat categories such as culture or community as static. Both informed by and informing geographic theories of globalization, Anderson (2012) writes that ethnographers understand communities as made or remade through uneven “spatial-social processes,” and that process-centered approaches can better understand and represent the lived experiences of globalization. Thus, not just cultures, places, and small events, but processes have also become the objects of knowledge of the geographer using ethnography. For example, multisited ethnographic studies (studying events located in different places) allow researchers to study the processes of globalization and the literal movement of material life.

Like the study of globalization, the new technology of the Internet and emergent digital communities have similarly created a challenge to traditional ethnographies of place, with carefully circumscribed boundaries. But, as noted in recent nonrepresentational theory, it is not just a place, but how we are in a place that constitutes that place. Thus an ethnography of cycling up a mountain, for example, demonstrates how movements in a place create meanings of that place.

Ethnography has been an important methodology for geographers whose objects of knowledge have included emotional experience. Via the exercise of reflexivity, some of this emotional data can contribute to understanding the role of subjectivity in the creation of knowledge. But, in other cases, access to emotional worlds is relevant to policy purposes – such as an ethnographic study of the experience of transnational migrants, or other groups whose life experience is not directly informing policy, but whose lives are nevertheless the topic of policymaking.

Geographers have also explored the connections between ethnography and the arts to engage the emotional world, many projects being visual in nature but some textual – such as those marrying ethnography with creative fiction. Unlike realist ethnographies, ethnographic fiction engages humanistic and nonrepresentational geography to create narratives whose emotional content is communicable to nonacademic audiences. What is to be gained ultimately by accessing emotional worlds is an understanding of the human values that motivate human action – data that are often not obtainable via other methods. Ethnography remains an important method to access information for which people are less immediately conscious.
As another visual method of representing the world, GIS and other spatial technologies and analytical methods are increasingly a part of the ethnographer’s methodological toolbox (e.g., Matthews, Detwiler, and Burton 2005). A growing number of realist ethnographers are conducting participatory mapping projects or cataloguing spatial data from field notes to process in a GIS. These projects are happening both with indigenous peoples, over large areas of territory, and in urban areas at small spatial scales. Such projects set apart the contributions of geographer-ethnographers from ethnographers in other disciplines.

It has been four summers since I first set up my tent in the Chief’s front yard. We developed this work iteratively … One day I was shocked to hear the Chief call me a “modern Christopher Columbus.” Shocked, and at the same time pleased – because as a critical geographer, I never wanted to deny my roots as an explorer … So this inductive, exploratory approach allowed me to include the Chief steering this ethnographic research: not describing “Nulato” or “Koyukon” as objects of knowledge … To do critical ethnography means to sacrifice one’s ego – to be humble – and to make room for voices besides my own in the ethnographic narrative. The effort has variable “success” – whatever that means. Four years into this current project, I am feeling success without even having yet published one word on this work. (Memo 1 on the use of critical ethnography, November 11, 2014)

Both the continuities and changes – and the challenges – in the practices of ethnography are referenced in this memo. Ethnography is most utilized within geography in fields such as cultural ecology, development studies, feminist studies, and social, political, cultural, and nature–society geography. Yet ethnographic methods need to be better understood across the discipline of geography; fieldwork, being a large component of ethnographic data, is also constitutive of geography as a discipline. In this way the inductive practices of ethnography may be usefully studied by any geographer wanting to explore a terrain that is unfamiliar or to better inform the research question. While ethnography is not applicable to all research questions, perhaps in future ethnography in geography will be more formally recognized as used by geographers everyday.

SEE ALSO: Cross-cultural research; Cultural turn; Emotional geographies; Feminist methodologies; Fieldwork in human geography; Qualitative data; Reflexivity; Subjectivity

References


ETHNOGRAPHY


Further reading


Ethnophysiography

David M. Mark  
*University at Buffalo, USA*

Andrew G. Turk  
*Murdoch University, Australia*

**Ethnophysiography: definition and scope**

**Definition**

Ethnophysiography studies how people conceptualize the natural landscape, especially landforms and water bodies. David Mark and Andrew Turk coined the term in 2003 to refer to an ethnoscience of natural landscape features. Ethnophysiography relies heavily on ethnography as a method for obtaining information through language community participation in interviews and discussion. Ethnophysiography focuses on kinds of things in the landscape, such as hills, rivers, and vegetation assemblages and aims to document in detail the terms in a language that refer to the landscape and its parts. It also examines how landscape plays a role in oral history and culture and how people are cognitively and spiritually attached to landscape. By documenting landscape concepts, ethnophysiography also may provide foundations for culturally appropriate indigenous GISs.

The Ethnophysiography Hypothesis proposes that people from different places and cultures use different conceptual categories for features of the landscape, as reflected in the generic terms in their languages for features and in their toponyms.

The meaning of the term *landscape* is contested. Sometimes the term “geographic” is used for the domain of ethnophysiography. The term “ethnophysiography” was chosen to reflect a focus on the world of physical geography, and so far, ethnophysiography excludes the built environment and administrative units at landscape scale. However, it does include toponyms and cultural and spiritual associations with landscape (Turk, Mark, and Stea 2011).

**Motivations**

Is GIS culturally biased? In the 1960s and 1970s most GISs were automated versions of analog practices in cartography, transportation, urban planning, and related fields. By the late 1980s, however, the GIS research community began to seek more comprehensive theories of spatial relations and geographic data models. This new research drew on formal mathematics and on artificial intelligence models of spatial phenomena. Several publications in the late 1980s and early 1990s raised questions of universality and possible linguistic or cultural biases in GIS software (Frank and Mark 1991; Turk and Mackaness 1995). If people in different cultures or languages conceptualize geographic or landscape phenomena differently, then GIS software is likely to be biased toward the languages and cultures that produce it, namely an Anglo-Germanic European worldview. An ontological turn in geographic information science (GIScience) directed further attention toward these issues.

Is this a language issue? Do cross-cultural differences in landscape terms and their meanings indicate differences in *thinking* about landscape,
ETHNOPHYSIOGRAPHY

or do they only reflect language use at a surface level? The well-known yet controversial Sapir-Whorf hypothesis, or Sapir-Whorf conjecture, claims that people think and reason using language, and thus predicts that if terms divide up a domain differently, this may lead to different reasoning and decision-making. An extreme version of this hypothesis implies that it may be impossible to fully transmit meaning from speakers of one language to speakers of another. A more probabilistic version of Sapir-Whorf would predict that terms will not “line up” across languages in a one-to-one fashion, but their meaning can still be expressed in other languages by using longer compound expressions.

Language, culture, cognition, environment, and history all play a role in the terms and categories that are used to communicate the conceptualization of a domain. The deeper structure of a language may influence the nature of categories and their labels and may constrain the nature of terms for landscape entities.

Interestingly, it was a geographic information (GI) and society research thrust, rather than cognitive and linguistic research themes, that brought issues of language-specific concepts to the forefront. A workshop on GI and society held in Friday Harbor, Washington, in 1993 led to a special issue of a journal that included the seminal paper by Robert Rundstrom (1995), entitled “GIS, indigenous peoples, and epistemological diversity.” Land rights and land management are particularly important to indigenous peoples, so if GIS is used in land disputes or land management, it is crucial to assess whether cross-cultural differences in conceptualization of landscape lead to conceptual biases in GIS. So far, the authors know of no serious efforts to redesign GIS software to accommodate such differences, although some practical GIS have attempted to include cultural information through their ontology and/or via multimedia elements, for example, the work of Wellen and Sieber (2013) with the Wemindji Cree in Northern Quebec. Research results from ethnophysiography should enable further advances in culturally appropriate indigenous GIS.

Components of ethnophysiography

Ethnophysiography differs from fields such as ethnobiology in more than just subject matter. In ethnobiology, scientific categories often are used to form a baseline or etic grid for the study of emic (within culture) categories across languages. A critical issue for ethnophysiography is that natural inorganic domains are not organized by nature into kinds to the same degree as for biological entities. Thus scientific geomorphology and hydrology do not provide ethnophysiographic researchers with an adequate etic grid of categories in the same way that biology often does for ethnobiology. People may cut the continuous landscape into different spatial parts, and then classify the parts differently as well. This provides a methodological challenge for ethnophysiography due to the lack of an independently defined baseline of categories. It also explains why languages, cultures, and individuals appear to differ much more in their categorization of very similar landscape elements than they do for plant and animal categories. Thus ethnophysiography may extend ethnoscience by more than just adding another domain.

Landscape terms and their associated concepts form a key component of the ethnophysiography of a language or culture. Most ethnophysiography or landscape-in-language projects have restricted the domain of study to terms for “natural” landscape features, including landforms, water features, and vegetation assemblages.

One issue is whether the “primary” denotation of a particular term is a feature at landscape scale. For example, the Yindjibarndi people
of northwestern Australia have a term “yirra,” which is used at landscape scale to denote the sharp edge above a cliff, where there is a particular risk of falling. However, the same word is also used to refer to other sharp edges, including the edge of a tooth or of a knife blade. A similar pattern is observed in Yindjibarndi for several other terms whose primary meaning seems to designate shapes, but which also denote objects or features having that shape. Research could determine whether landscape is the primary denotation of such terms, but to understand how language subdivides landscape phenomena into types or kinds, it makes sense to include terms that are frequently used for landscape even if their primary denotation lies elsewhere.

Toponyms (place names) are linguistic or symbolic labels for particular places or landscape features (including linear features, such as trails, and groups of features). Toponyms are one of the important subclasses of proper nouns and form an important area of ethnophysiography research. As symbolic references for particular places, toponyms may be a human universal. However, linguistic systems for toponyms vary widely across languages and cultures. To some extent, toponyms are “encapsulated,” by which it is meant that the terms (vocabulary) and constructions used within individual toponyms may differ from how the same concepts are expressed in that language more generally. In many, but not all, languages toponyms for some classes of features contain a generic component that indicates the kind of feature being named (e.g., Richmond Hill). In English and many other Indo-European languages, names of populated places often lack such a generic term while most other feature types do include generics.

Traditional Environmental Knowledge (TEK) is a term that is often used to focus on the value of traditional knowledge that is not captured by “scientific” or first-world knowledge. For example, TEK applies to types of plants and their uses, as food or medicine. Traditional geographic knowledge, on the other hand, seems mostly to be knowledge about places, which are instances or particulars, such as knowledge of what is where, and of how to travel from one place to another. Because geographic knowledge emphasizes instances, applying the TEK approach to geographic phenomena may be difficult, however, it is critical to include such investigation in ethnophysiography. People’s dwelling relationships with landscape are triggered predominantly by their temporal patterns of regular activities concerning utilitarian and social needs within what social anthropologist Tim Ingold (1993) calls a “taskscape.”

Landscape ethnoecology provides an approach to generalized places as containers for useful types of plants or animals (Johnson and Hunn 2010). Landscape ethnoecology focuses on relatively homogenous patches of vegetation. The approach is similar to land use and land-cover mapping via remote sensing within conventional science, but is based on traditional (“ethno-”) concepts of habitat types or kinds of places. The patch approach allows knowledge of plants and animals to be “scaled up” to the landscape scale.

Ethnophysiography recognizes that the landscape conceptualizations of a speech community extend beyond physical aspects of land and encompass their cultural and spiritual dimensions: memories, beliefs, fears, and so forth. Such considerations may be most evident for indigenous peoples but apply everywhere. A landscape term may incorporate spiritual aspects; for example, a permanent pool (yinda) in Yindjibarndi always includes a spirit (warlu). Basso’s (1996) book is a seminal publication on this topic and work by philosophers of place, such as Jeff Malpas (2006) and Ed Casey (1996), are reviewed in Turk (2011).
ETHNOPHYSIOGRAPHY

Tuan (1974, 4) discussed environmental perception, attitudes, and values about place and coined the term “Topophilia” to refer to “the affective bond between people and place or setting.” Within ethnophysiology, people’s emotional relationships with landscape can be considered via “intentionality” (phenomenology) or through the recent turn to “affect” in geography, literature, social sciences, and cultural studies.

What is excluded from ethnophysiology?

The current notion of the domain of ethnophysiology concentrates on “natural” environments and on terms for entity types, toponyms, and beliefs about landscape in general. Thus it does not include terms or linguistic constructs for spatial relations or for directional terms, nor choices and priorities for spatial reference frames, which are scale-dependent in some cultures. Also excluded are terms for artifacts (constructed entities) at landscape scale, such as buildings, settlements, roads, dams, canals, bridges, and geopolitical regions and administrative units. Separation of “constructions” and “nature” is itself somewhat artificial and the methods of ethnophysiology could usefully be brought to bear on geographic-scale artifacts. Perhaps this will demand a new term for the field or a broadened definition of ethnophysiology.

History of ethnophysiology

Research topics and collaborations

David Mark and Andrew Turk initiated the field named “ethnophysiology” in 2002; David Stea joined the project the following year. Seeds for the ethnophysiology topic were sown in their earlier work (summarized in Turk and Stea 2014) and germinated when Turk and Mark participated in a workshop on geographical ontology held in France in 2000. In 2002 Mark began a sabbatical with a month at the Max Planck Institute for Psycholinguistics (MPI) in Nijmegen, where he interacted with researchers examining space in cross-cultural context. Then Mark joined Turk in Western Australia where fieldwork with Yindjibarndi speakers launched the project; a description of that first case study was published (Mark and Turk 2003) and that paper introduced the term “ethnophysiology.”

Collaboration by ethnophysiology researchers with MPI linguists in their preparation of a special issue of Language Sciences (Burenhult 2008a; Burenhult and Levinson 2008) led to an international transdisciplinary workshop (on the Navajo Reservation) on “Landscape in Language” in 2008 (Mark et al. 2011) and a European Science Foundation Exploratory Workshop in Spain in 2012. Burenhult established the Language, Cognition and Landscape: Understanding Cross-Cultural and Individual Variation in Geographical Ontology (LACOLA) research group in Lund (Sweden). Turk is assisting linguist Clair Hill with one (regarding the Manyjilyjarra language) of about eight LACOLA case studies of landscape in language in different parts of the world.

Highlights of some case studies in ethnophysiology

By now, at least thirty detailed ethnophysiology-related studies of landscape in language have been published. These include the nine articles in Burenhult (2008a), several chapters in Johnson and Hunn (2010), and 14 case studies reported in Mark et al. (2011). These studies provide very strong support for the ethnophysiology hypothesis stated above and indicate that factors motivating landscape
ETHNOGRAPHY

categorization are perceptual salience, human affordance and use, and culture and social organization. Brief summaries of several of the case studies follow.

The first ethnophysiography case study was of Yindjibarndi, a Pama–Nyungan language (Turk, Mark, and Stea 2011, 26–31). Yindjibarndi country in northwestern Australia includes tablelands, broad valleys, ranges, rolling hills, extensive flats, and some cliffs and rock outcrops. The climate is hot and dry. There are no permanent rivers or creeks, with water only running after major rainfall events (cyclones), however, there are some permanent pools, springs, and soaks. The vegetation is grass, spinifex, low scrub, and scattered trees. Yindjibarndi people traditionally lived a seminomadic hunter-gatherer lifestyle, but have been progressively displaced by European colonization. Most Yindjibarndi people now live outside their traditional territory but retain strong cultural links to their land, visiting for hunting, gathering, and ceremonies. Many Yindjibarndi terms subdivide and classify landscape features differently than English. For example, the term marnda covers a range of features that in English would be called mountain, hill, ridge, rock, and so forth; related terms (bargu, burbaa) do not simply equate to hill or rise. Some Yindjibarndi terms effectively refer to shape rather than topographic objects, for example, burbaa, and may also refer to non-geographic entities of that shape. Compound phrases are used to denote kinds of marnda, some of which would be denoted by single words in English, for example, marnda marlirri for mesa. Yindjibarndi hydrologic terms refer separately to water, lying in pools, flowing, or flooding (bawa, yijirdi, mankurdu) and longitudinal depressions, small and large (gaga, wundu), along which water sometimes flows, and thus contrast with English and other European languages where terms such as “river” and “creek” incorporate both water and channel. Similarly, thardarr is a place where water sometimes falls down a cliff; if water is present it is referred to by the flow magnitude words above, so there is no Yindjibarndi term for a waterfall per se.

Turk, Mark, and Stea (2011, 31–36) conducted a study with Navajo, a Native American language in Arizona, New Mexico, and Utah in the United States. Navajo (or Diné Bizaad) is an Athabaskan language, having many verbs, few nouns, and no word-class corresponding to adjectives in English. Until recently, the Navajo people led a hunter-gatherer lifestyle, but then adopted corn growing and sheep rearing. Their land is semi-arid with a few permanent rivers and streams crossing or bounding Navajo country, but most watercourses are usually dry. Physiography is basin-and-range with flat-lying rocks and volcanic necks forming buttes, mesas, and plateaus. Some Navajo landscape terms encode the material of which the feature is composed, for example, bikooh is the term for a canyon or gully through soil and tsékooh for similar features cut into rock. Navajo toponyms often are descriptive, sometimes making it difficult to distinguish toponyms from mere landscape descriptions.

O’Meara and Bohnemeyer (2008) reported on landscape terms in Seri, a language isolate spoken in the Sonora region on the Pacific Coast of Mexico. Seri territory includes an arid coastal plain, some large islands, and the sea. The Seri were traditionally seminomadic hunter-gatherers and their primary food resource was the green sea turtle, but some terrestrial animals, such as deer, and also wild plants were eaten. Seri classify landscape in terms of material consistency and spatial properties such as shape, orientation, and mereological relations. Complex landscape terms are the rule rather than the exception and are based on one of four Earth substance terms:
ETHNOGRAPHY

hast (stone), hax (fresh water), xepe (seawater), and hant (ground, land), in combination with a definite article with posture semantics, a nominalized form of an intransitive verb, an adjective, or a relational noun.

Levinson (2008) discussed landscape terms in Yéli Dnye, from Rossel Island in Papua New Guinea. The island is mountainous with tropical rainforest, beaches, barrier reef, lagoon, rocks, and so forth. Speakers’ lifestyle includes gathering nuts and game, growing taro and coconuts (via swidden cultivation), and fishing. Terms are for seascape as well as landscape. To the Yéli Dnye, a watercourse consists of three segments: an upstream part in the mountains, a middle segment on lower land, and a third part being the stream of river water across the sea lagoon to a reef opening. The two upstream portions, taken together, would be termed a “river” in English. Parts of the lagoon have terms but not the whole. Toponomy seems to follow terms for landscape features (rather than being a complementary system) and toponyms form a syntactically distinct word (or phrase) class.

Burenhult (2008b) reported on language and landscape for the Jahai people from Peninsula Malaysia and adjacent parts of Thailand. Jahai belongs to the northern Aslian branch of the Mon-Khmer language family. Speakers are hunter-gatherers (and occasional swidden cultivators). The landscape is dominated by mountains drained by swift-flowing streams in steep-sided valleys. Rainforest covers foothills and lower mountains with occasional bare rock outcrops, some of mythological significance. Metaphor from animate domains, mainly human, is used in landscape terms; body parts (creating partonomy) and kinship terms (creating size taxonomy) are used to refer to streams and valleys. Toponyms seem exclusively linked to notions of drainage with permanence of water being crucial.

Enfield (2008) summarized research on landscape terms in Lao, a Southwestern Tai language. The landscape is flat, well-watered land with mountains (adjacent and distant). Vegetation is thick and tropical, and low land is cultivated. There is a general term for water – and rivers can be just referred to by that term – or if large by another term meaning “water mother.” Lao has a basic word phuu2 often translated as “mountain,” but is more accurately glossed as “mountain terrain.” Thus, phuu2 muaj1 nii4 (mountainous_terrain clf_unit this) means “this unit of mountainous terrain.” (The numbers denote tones, and “clf” indicates a “classifier.”) There is no Lao word for valley.

Holton (2011) compared landscape terms from two unrelated North American indigenous language groups, Athabaskan and Eskimo, living in adjacent areas of western Alaska, a landscape of large rivers, taiga woodlands, and broad valleys. This study demonstrates clearly that Tanacross Athabaskan and Central Yup’ik Eskimo speakers conceptualize landscape in different ways, such as different systems for categorizing elevated features. Living near the coast, Eskimo marine and coastal terminology is naturally much more elaborate than that of the inland Athabaskan. Conversely, hydrological terminology is more developed in Athabaskan with differences reflected in toponymy.

These case studies, and more than twenty others, clearly demonstrate the variability between language/cultural groups in the way that continuous landscape is broken into features and the sorts of cultural and spiritual associations with place. Concepts of landscape features may be expressed in generic nonlandscape terms, such as those for shapes. Also, nonspatial metaphors may be utilized to organize and represent aspects of landscape. There are also considerable differences in ways of forming and using place names.

Based on the published landscape-in-language case studies, Turk defined the ethnophysiography
ETHNOPHYSIOGRAPHY
descriptive model (Turk, Mark, and Stea 2011).
It lists key differences in the way landscape is
treated in different languages and cultures and
possible factors that might lead to these differ-
ences, although not suggesting that deterministic
causation is necessarily involved.

Methods

Ethnographic methods

Ethnophysiography case-study methods (Turk,
Mark, O’Meara and Stea 2012) vary according
to the physical and social environment, for
example, remoteness, vehicle access, vegeta-
tion, weather, availability of consultants, and
organizations to assist (especially representing
indigenous consultants). However, case-study
research in ethnophysiography usually includes
the following stages:

1 dictionary work and representative photo
collection;
2 field interviews with consultants;
3 photo interpretation sessions;
4 semi-structured follow-up to clarify confu-
sions, and so forth; and
5 reporting initial results back to community
members and getting feedback.

Case-study circumstances and pragmatic con-
siderations may mean particular techniques may
be used in a different order and/or iteratively.
All ethnographic procedures should be used
in a “reflective” manner with the fieldworkers
continually questioning the validity of their
results and ensuring ethical practice.

Other specific techniques that can be used
include:

• questionnaire for eliciting landscape terms
  and toponyms;

• consultants narrating a video (or animation)
  presentation featuring a person (or animal)
  moving through the landscape; and

• “director/matcher” technique – two speak-
ers are separated by a visual barrier, with one
describing a landscape image well enough for
the other to choose the correct image from a
duplicate set.

Some of the most valuable results have been
obtained by showing landscape photographs to
speakers. One can ask the participants simply to
report what they first notice in the photographs.
Of course, the result will depend on the partic-
ular photographs shown. Alternatively, partici-
pants can be asked to arrange photographs into
groups and perhaps describe each group verbally.

Surveys, questionnaires, and experiments

For some languages and cultures, information
about landscape categories and other concep-
tualizations can be obtained from surveys and
experiments. Surveys can provide valuable infor-
mation for larger groups especially in literate
majority cultures. Of course, for indigenous
populations, elders, and other marginalized
groups such experiments or surveys might not
be appropriate and ethnographic methods may
be preferred.

A valuable early survey approach can be to ask
participants to list terms of types of landscape
features. The exact wording of the question may
have significant influence on the results. The
superordinate category might be “landscape” or
“geographic” or perhaps “environment.” The
noun might be “object,” “feature,” or similar
words. Questionnaires and surveys can be used
to ask for typical terms of various superordinate
categories or to elicit definitions or distinctions
between neighboring categories. These different
types of data might be collected in writing or by
ETHNOPHYSIOGRAPHY

audio recording. If a large number of speakers can be surveyed, statistical methods of analysis can be used to determine, for instance, the frequency of subjects listing each term.

Ethical issues

When working with indigenous groups, language preservation often is a key consideration, however, there is also some element of “appropriation,” of “taking away” from the culture, so asking permission and involving communities in ethnophysioigraphy studies is required. It is also essential to understand the worldview of a specific indigenous group before attempting to interpret linguistic data.

The rights of collaborators and language communities must be respected by:

• ensuring permission is obtained for all fieldwork;
• making appropriate payments to collaborators;
• using both male and female researchers to handle gender-sensitive issues;
• not eliciting, recording, storing, or publishing secret/sacred information; and
• acknowledging collaborator contributions in publications.

Threats to validity

When designing or evaluating ethnophysioigraphy case studies one must consider threats to validity arising from methods of data collection and/or interpretation (Turk et al. 2012), which include but are not limited to:

• terms may relate to fairly equivalent concepts between languages, but landscape feature examples fitting the concepts may differ;
• the physical extent of a feature covered by a term may vary;
• different whole-part concepts may apply, for example, river water/channel;
• terms may refer primarily to shape rather than landscape features as entities;
• the physical point of view may influence feature conceptualization;
• terms may refer to spatial areas (places) rather than features (objects);
• participants responding to landscape photographs may misinterpret the scale, focus of attention, or content of the photographs;
• cultural and spiritual aspects of place may be misinterpreted, or not discussed;
• rare or exceptional features may be referred to only by proper names and not by generic terms; and
• generic parts of toponyms and landscape terms may have different meanings in more general contexts.

Such threats to validity of interpretation of landscape terms can be reduced through: use of multiple methods (“triangulation”) and iterative procedures; using a diverse range of collaborators; not “leading the witness”; being open to alternative interpretations; and becoming as intimate as possible with lifestyle and social structures of the language community.

Implications for geographic information systems and science

GIS software

As noted above, GIS may be biased toward the culture and society that produced the software systems. Languages of spatial relations show differences in the verbal expression of relative locations and spatial relations, but the relations as expressed in language seem largely to be based on common underlying topological and geometric principles. Ethnophysioigraphy research
has assisted in understanding this problem and providing solutions, potentially via contingent feature classification, leading to culturally specific GIS.

Bias in spatial data infrastructures?

Results from ethnophysiography clearly demonstrate that terms for categories of geographic entities do not “line up.” Spatial data infrastructures (SDIs) are usually produced by national mapping agencies or large corporations and are largely based on standard sets of entity types (feature codes). It is not surprising that the categories recognized by SDI developers are usually based on the primary languages of the countries or companies producing the databases. Mark and Turk (2003) showed that Australian government landscape categories for hydrographic features (from AUSLIG) had limitations for GIS for indigenous land management in Western Australia due to the mismatch of categories. Ethnophysiography results from other languages suggest that similar limitations will be the rule rather than the exception.

Information retrieval and natural language

Much of information retrieval in current search engines such as Google is based on text, words, and languages. If the meanings of terms from language to language do not “line up” or have a word-for-word translation then there will be impediments to users from one natural language trying to retrieve information on the web that is described in another natural language. This issue is broader than ethnophysiography, but there are probably unique issues of category correspondence for the landscape domain. The Semantic Web is intended to overcome such interoperability issues but may have just the same difficulties.

Implications for geography

Geography as a discipline has struggled for centuries with the relative importance of the “geography of the particular” (instances) and the “geography of the general.” In 1650, through *Geographia generalis*, Bernhardus Varenius launched a scientific geography. Findings from ethnophysiography remind geographers and those from related disciplines that attention must be paid to the particular and that cultural differences and some aspects of “general geography” are not universal. Ethnophysiography exposes a more subtle point – what is general within one culture may not be general within another. This recognition reminds the discipline of the importance of a cross-cultural and cross-linguistic perspective and raises some skepticism about truly general findings and human universals for the landscape domain.

Summary and conclusions

Ethnophysiography is an emerging field that can make important contributions to geography and to geographic information science as well as to linguistics, anthropology, cognitive science, philosophy, and other fields. The topic occupies a promising overlap between three main areas of contemporary geography: human geography, Earth systems science, and geographic information science. Ethnophysiography is positioned to provide answers to some of the questions of bias raised by critical social theorists. Results can form the basis of culturally specific information systems that would allow indigenous groups and others to address geographic problems from within the conceptualizations of their own cultures.

Given all this, it is somewhat surprising that cultural differences in geographic concepts have
ETHNOPHYSIOGRAPHY

received little attention until recently. Cultural studies were an important part of geography before the so-called Quantitative Revolution. Critical scholars reacted against the quantitative approach but did not restore cultural studies to their former prominence within the discipline. Ethnophysiography is a step in that direction.

SEE ALSO: Critical GIScience; Cross-cultural research; Ethnography; Geographic information science; Indigenous knowledge; Landscape; Ontology: theoretical perspectives; Spatial feature classes; Toponymy

References


Everyday life is all around us. We engage with it and construct it every day of our lives. We inhabit everyday spaces and do everyday things—repetitively, routinely, continually, rhythmically. Everyday life has been defined and described as banal, the taken-for-granted stuff of life, the “leftover” or remaineder, the unnoticed, the unremarkable spatial-temporal practices of ordinary people taking place in everyday spaces. At the same time, everyday life is fascinatingly rich and varied; it can be punctuated by the extraordinary. Aspects of everyday life can be rendered remarkable and exceptional by some kind of change in our social environments: a freaky day of weather; a dramatic frightening event; a chance meeting with a good friend; an accident; news of a baby’s birth; being fired; a first kiss. These events or instances make the day, time, and place memorable for either negative or positive reasons. They can be interruptions that are triggered by global processes (climate change, economic collapse, war) or intimate, localized processes (a car accident in the neighborhood, a stranger’s kindness). Hence everyday life is a conundrum, at once humdrum but also with the potential for drama. As Ben Highmore states, “the everyday offers itself up as a problem, a contradiction, a paradox: both ordinary or extraordinary, self-evident and opaque, known and unknown, obvious and enigmatic” (2002, 16).

Everyday life, described as the “landscape closest to us, the world most immediately met” (Highmore 2002, 1), is made up of repeated actions, recurrent journeys, and inhabited spaces that constitute the day-to-day stuff of living and being. The notion of “everydayness” is interpreted often as ordinariness, boredom, monotony, tediousness, involving dull routines, and lacking excitement. Such social ennui around the everyday is frequently exploited for commercial ends through the advertising of products that offer excitement, newness, and difference. However, even without the advertiser’s hype, a key feature of everyday life that philosophers and geographers alike emphasize is that the exceptional can be found within the everyday and that everydayness has the potential for the spectacular. We even articulate this in the uttered expression when something unusual, bizarre, or shocking is witnessed: “Now there’s something you don’t see every day.” This surprise element is considered to have a political and critical potential for social change. Feminists, in particular, have focused on the everyday (particularly the domestic everyday) and captured this in the pithy slogan “The personal is political” to demonstrate the potential of, and possibility for, the everyday to generate new political forms.

Everyday life as an entity has been with us as long as there has been human life; yet, as a concept, it did not appear as an analytical construct until the 1920s but it became increasingly significant post-World War II. Key Western philosophers who have developed analytical approaches to the everyday include Georg Lukács and Georg Simmel, both of whom have been particularly important in sociology and cultural studies, although less so in geography. Henri Lefebvre’s and Michel de Certeau’s theorizations have been more significant in geographical work.
EVERYDAY GEOGRAPHIES

on the everyday because they developed more critical insights of the everyday as problematic, bringing “the everyday into awkward focus” (Highmore 2002, viii), and for their analyses of the spaces and places of everyday life.

Lefebvre made the critique of the everyday a lifelong project, publishing three volumes on Critique of Everyday Life in 1947, 1962, and 1981, only the first of which has been translated into English (1991/1947). He was influenced by Marxism and surrealism, and driven by a desire to challenge the modernization processes at play in France and its colonies. Through observations of social formation, Lefebvre developed an analysis of the everyday that contained a critique of the US-driven processes of modernity and the continued colonial projects of the French state which were rapidly changing postwar life and creating states of alienation for ordinary people. For him everyday life was being produced according to the rhythm of capital, which penetrated every detail of everyday life. He identified urban space as a site of the intensification of everyday life but believed that everyday life held the possibility of its own transformation through processes of dealienation. For Lefebvre such processes would push toward the achievement of a politics rooted in the modification of the everyday. Lefebvre thus identified the everyday as a site of resistance, revolution, and transformation. As he worked with the concept of everyday life, Lefebvre also developed a critique of modernity and its negative impact on social differences alongside an analysis of the spatial forms that modernity creates. Despite his valuable insights, he openly acknowledged that his critique constantly lagged behind the persistent dynamism of the reality of everyday life.

Michel de Certeau’s key work, The Practice of Everyday Life (1984/1980), is located in a different intellectual and social climate. De Certeau was interested in the imaginative and inventive practices of everyday life while recognizing that these cultures were submerged below a level of textual and social authority. He argued that, while everyday practices and cultures remained invisible and unrepresentable, they performed resistive practices against these authorities. De Certeau was particularly interested in the metaphor and materiality of journeying and its notion of becoming rather than being, one of the ways in which he understood and analyzed everyday life. He wanted to capture the ways in which people operated and practiced everyday life, and recognized that, while people make do with ready-made culture (akin to Lefebvre’s culture of modernity) they also made with such a culture and demonstrated inventiveness and ingenuity. De Certeau wrote of this as a creative arrangement and rearrangement and so focused on the “creativity of appropriation, a poetics of experiences and a style of everyday life” (Highmore 2002, 148). It has been recognized that de Certeau’s is the most wholehearted attempt to develop an approach to the everyday from the material provided by the everyday itself. Critics have argued that his arguments roam and are unsystematic; however, that is the patterning of everyday life he worked to capture and present as a kind of poetics. For him it was essential to make the everyday visible, to be seen as more than remaineddered, because only after the everyday is allowed to surface can a politics of the everyday become possible.

Despite the framing insights and conceptual toolkits provided by these social theorists, contemporary scholars in the social sciences and humanities continue to grapple with definitions of everyday life. Geographers have found it a tricky concept to define – slippery and dynamic at the same time as it is routinized and predictable – but have nevertheless worked with the everyday for decades (particularly within a spatial-temporal context) with different
EVERYDAY GEOGRAPHIES

philosophical framings and methodological techniques that have proven valuable in the study of “geographies of everyday life” (Jarvis 2010). In an essay on the geography of everyday life, John Eyles (1989) provided an early argument as to why geographers should examine the everyday. He proposed that understanding and being aware of the commonplace were important for the development of a geography that was human-centered. It was seen as an essential way to explore the relationships between people and their environments at a time of rapid change (in his case time–space compression, but we might now name globalization, neoliberalism, terrorism and conflict, or climate change). Feminist geographers have made important contributions to furthering and stretching analysis of the everyday through researching women’s lives and articulating place- and space-based knowledge production. A key political aspect of feminist geographical work about the everyday is the goal of keeping women visible despite the processes of rapid global change, which tend to hide their social, economic, and political contributions. Holloway and Hubbard (2001) have drawn together a comprehensive analysis of what they call “the extraordinary geographies of everyday life” in order to examine the fundamental elements of human geography, people, and place. Most recently, Horton and Krafṭl (2014) have dedicated a whole chapter to “Everyday Geographies” in their Cultural Geographies: An Introduction. In a section working through the complexities of how to define the everyday, they conclude: “so whereas ‘everyday life’ refers to life as it is lived and experienced, ‘the everyday’ might be understood as the form that life takes after it is captured and transformed by a system of representation” (Horton and Krafṭl 2014, 183).

What is certain about the complexities and unpredictability of everyday life is that they will always be with us and, as human geographers, we will continue to try and understand and analyze it.

SEE ALSO: Consumption; Cultural geography; Feminist geography; Humanistic geography; Phenomenology; Social geography; Urban geography

References


Evolutionary economic geography

Eike W. Schamp
Goethe University Frankfurt am Main, Germany

Evolutionary economic geography has emerged fairly recently as an alternative approach to explaining change in spatial and regional structures through endogenous technological innovation. Thinking about economic dynamics “from within” gave rise to a renewed debate in heterodox economics concerned with understanding the emergence of novelties that compete with each other and with incumbent structures as the main drivers of change. Charles Darwin’s theory of the evolution of species has been acknowledged as a generic concept here. There is considerable disagreement, however, on how far the analogy should go for socio-economic processes, particularly as Darwin’s “broad” concept has stimulated further refinements both in biology and in the social sciences. Much imprecision still prevails in economics about what “evolution” means. It is fair to say that the same applies to the current debate in evolutionary economic geography.

Evolutionary thinking has been discussed by economists since the end of the nineteenth century, for instance by Thorstein Veblen and Joseph Schumpeter. The debate gained momentum with Richard R. Nelson and Sidney G. Winter’s seminal book *An Evolutionary Theory of Economic Change* (1982). Evolutionary thinking in economic geography started in the late 1990s and achieved much attention when Ron Boschma and Koen Frenken (2006) picked up a question similar to the one asked by Thorstein Veblen 100 years earlier in economics: “Why is economic geography not an evolutionary science?” Currently, the perspective is predominantly oriented to evolution in economic geography as a competing or complementary conceptual approach to the study of regional change. A considerable proportion of scholarly debate in economic geography still draws its basic ideas from Nelson and Winter’s book. However, Richard Nelson considered evolutionary economics to be at an embryonic stage of development in the mid-1990s. Although very lively in recent years, evolutionary economic geography has not yet passed beyond that initial stage.

Some basic dimensions of evolutionary theory

Two questions arise when transferring a theory from the natural to the social sciences. First, how are the focal concepts translated and interpreted in accordance with the understanding of deliberate human agency in social sciences? Second, in what sense is the theory of evolution different from “competing” approaches relating to similar empirical problems in social sciences in general and in economic geography in particular? These questions will be answered briefly before turning to specific aspects of evolutionary economic geography.

The principal analogy draws on the question of how a new population of species emerges from the variation and replication of its “genes,” how a given population copes with environmental change by adaptation, and how it survives under
EVOLUTIONARY ECONOMIC GEOGRAPHY

selection pressures from environmental change. While variation and retention refer to genes, their carriers are the species of a given population, so-called phenotypes, interacting with and being selected by the environment. More recently, phenotypes have been termed “interactors,” whose capacity to adapt and survive in changing environments depends on the variation and replication of genes, hence the term “replicators.” This basic reasoning offers fascinating perspectives on the explanation of change and restructuring in economies if one translates the interactor into “organization or firm,” the replicator into “routines and technologies,” and the population into “sectors or industries.” This was the seminal step taken by Nelson and Winter (1982) in their book. Since then, however, a continuing debate has run as to whether the use of Darwin’s theory and terminology in economic and economic geography thinking involves analogies, metaphors, or mere “lip service.” Thus there is no clear-cut theoretical base for evolutionary economic geography but much diversity in the approach.

According to Witt (2008), at least three different interpretations can be identified in evolutionary economics, two of them relevant to economic geography. First, the most popular seems to be to use Darwin’s theory in a metaphorical way as a heuristic device in empirical research. This applies to the neo-Schumpeterian approach to evolution in economics, with Nelson and Winter’s (1982) book as its main proponent. Herein, entrepreneurial innovative activities at the firm level create variety in new technologies which, if successful (surviving), drive economic change. Many geographers of what could be called the Utrecht school of evolutionary economic geography around Ron Boschma and Koen Frenken followed this interpretation. Second, the naturalistic Darwinian view has been more or less rejected in economics. Third, universal or generalized Darwinism has been suggested as an abstract generic base for any empirical evolutionary research in the social sciences. In economic geography, Jürgen Essletzbichler and David Rigby (2007) have repeatedly discussed generalized Darwinism as a principal approach to evolution.

There is, nevertheless, some common understanding of basic dimensions in evolutionary thinking in economic geography. First, the trinity of variation, selection, and retention of technologies and organizations forms the starting point of dynamics in economic evolution. Second, the basic mechanisms of these dynamics are creativity and competition in particular environments. In the Schumpeterian sense of creative destruction, competition on innovation is the main driver of the emergence and creation of new “routines” within and among firms, where new routines, including the use of new technologies, enfold new organizational forms of both firms and sectors that have to persist on markets. Evolution then is the outcome of selection processes where the capability to persist depends on the diversity of innovations generated by the purposive agency of firms and the capability of actors to alter the rules that decide what is selected. This interpretation can obviously and easily be related to other theoretical strands in heterodox economics, in particular to neo-Schumpeterian innovation theory, Penrosian theory of the resource-based firm, and Simon’s concept of bounded and procedural rationality. In a wider sense, each consideration of knowledge creation and learning as purposeful action in trial and error that starts from heterogeneity of resources and finalizes in heterogeneity of results can be incorporated into evolutionary economic geography thinking. Altogether, the continuing mutation of routines and technologies through creation (and finally innovation) and imitation explains the extant
heterogeneity and variety of products, processes, and their firms.

Third, changes in a firm’s environment may stimulate novelties and contribute to their selection, that is, co-determine the survival chance of some competing routines and technologies and firms. While markets fundamentally form the firm’s selection environment in economic thinking, both upstream and downstream, geographers highlight place and region as important parts of the selecting environment. However, this brings several problems to the fore because markets as well as place and region are constructed (sometimes not deliberately) through the purposeful actions of competing firms, state regulators, and civil society and often do not exist in advance but may coevolve simultaneously with new technologies and sectors. As will be discussed later, this opens the debate to the concept of coevolution and to the need for incorporating institutional and political economic approaches into evolutionary economic geography.

Fourth, the emergence of new technologies, new firms, and new sectors is considered to take “historical” time. Current processes are irreversible, therefore, and embedded in or dependent on past processes. The perspective is of years if not decades in the practice of empirical research in evolutionary economic geography. Although historical events and contexts are unique, from this perspective, evolutionary thinking in economic geography looks for generic mechanisms that steer spatial and regional evolution.

Fifth, the evolutionary economic approach suggests building a bridge between the micro-foundations of variety creation, selection, and retention at the level of the individual or organization and the macro-perspective on (regional and national) economic change. While it is acknowledged that agency at the micro level (individuals, firms) is the main driver, evolutionary thinking wishes to use a meso analytic lens when explaining macro processes of economic change. This meso-level perspective refers to the emergence and survival of “populations” in their (changing) environment, be it organizational or technological routines, firms and sectors, or complex organizational forms such as clusters or regions.

Taking all these dimensions together, the second question about competing approaches can be answered quite quickly. Evolutionary thinking refuses the fundamental assumptions of neoclassical economic theory, such as the presumed rationality of human agents, the timeless and abstract market model as selection environment, and the search for a state of economic equilibrium, which for a long time formed the basis for spatial analysis and many regional growth theories in economic geography. Nor is evolutionary economic thinking similar to idiosyncratic thinking in economic history or older landscape approaches in geography. Evolutionary economic geography focuses on exploring generic spatial principles in the emergence of novelties and the dynamics of economic life.

Evolution in regions, an exercise in population thinking

Broadly speaking, evolutionary economic geography suggests different interpretations of Darwin’s theory and tackles a variety of problems with a variety of methods in practice. On various occasions, Jürgen Essletzbichler has convincingly differentiated between evolution (of technologies, firms, and sectors) in regions and evolution of regions. The first relates to evolving populations of technologies and firms in their respective local or regional environment, a perspective that is predominantly chosen in
the neo-Schumpeterian approach to evolution and realized in the majority of empirical studies in evolutionary economic geography (Boschma and Frenken 2011). These studies focused almost exclusively on the evolution of firms and sectors in a spatial and regional context. The evolution of regions has been much less approached empirically, and is considered a more complex problem that may call for the systemic perspective of generalized Darwinism (Essletzbichler and Rigby 2007).

Often implicitly, both the neo-Schumpeterian and the generalized Darwinist approach take up earlier concepts for analysis of the dynamics of economic evolution, in particular Gunnar Myrdal’s cumulative causation and Raymond Vernon’s product life cycle model. The development of novelties, firms, and sectors is said to pass through several life cycle stages in a seemingly well-defined order, that is, from birth or creation to growth, then maturation and, if there are no opportunities for life extension, decline. The same applies to higher-order arrangements of industries, basically to clusters which generally emerge around a focal industry, and networks (Glückler 2007). With reference to the basic geographical perspective on spatial unevenness and its reproduction in the world, the crucial questions concern, first, in what sense location has a diversifying impact on the emergence of novelties and, second, how far the evolution of novelties results in spatially uneven geographies. Hence the focus is on the interplay between the evolving population and its geographically defined environment.

Being less interested in the diffusion of novelties to extant firms by imitation and adaptation, the emergence of firm populations in an uneven spatial environment and the subsequent uneven spatial development of the population are at the heart of empirical research in evolutionary economic geography. Interestingly, empirical research has mainly focused on the spatial evolution of manufacturing industries either in “old” sectors such as the automobile and dairy industries taking a long-term perspective or in the recent formation of “high-tech” sectors such as laser, optoelectronics, video game, and software industries. Early studies of the evolution of industrial sectors such as the automobile industry have clearly revealed spatial unevenness in the formation stage of the sector, in the likelihood of firm entry and survival, and changing geographies of the sector during its life cycle. As the emergent spatial pattern in these industries is mostly in clusters, economic geographers have also turned to understand clustering as an evolutionary process (Boschma and Frenken 2011). The transfer mechanisms of new routines from an incubator firm to others mainly highlight the role of local spin-offs and labor mobility in evolutionary cluster models.

Over time, more detailed concepts have been introduced in order to explain the uneven geography of the evolution of firms and sectors and the role that local externalities play therein. Michael Storper and Richard Walker (1989) introduced the concept of a window of locational opportunity that was favorable to the emergence of novelties. Favorable environments seemed to be characterized by the lack of obstructive regional institutions – often caused by incumbent sectors in old industrialized regions – and the diversity of urban agglomeration externalities of the so-called Jacobs type. While this may be the case for disruptive technologies and sectors, in the Schumpeterian sense, recent research in economic geography has turned to more incremental technical change in looking for the roots of new technologies, with the concept of technological relatedness among firms and sectors (Boschma and Frenken 2009). In line with the literature on proximities in learning and knowledge development, it is claimed that the
evolution of new technologies and firms at a specific location is often related to previously accumulated knowledge there. As a result, the likelihood of the creation and survival of firms in a new sector may depend on local relatedness. It is a matter of empirical research, however, to depict that part of existing knowledge in incumbent sectors which generates opportunities for further knowledge creation and the formation of new firms and sectors. There certainly is a research frontier for the moment. Geographers have recently termed the local formation process of new industries that are technologically related to older ones “regional branching.” Branching may occur as a process of outsourcing new technologies from an old industrial sector or as a process of recombining competences from different sectors at a location in new firms. Current empirical research in evolutionary economic geography uses indirect measures of knowledge flows between industries that are assumed to be technologically related, such as patent use, proximity in industrial classification, and co-location. It seems evident that branching at certain locations may happen while it is at the same time truncated at other locations. Less is known in detail about knowledge communication resulting in the emergence of a new sector in situations in which the knowledge-base of different firms is related but nevertheless different. The crucial question then is when and how does heterogeneity among firms in variety creation lead to a new sector, hence heterogeneity between sectors. Another issue concerns the context in which bifurcation of sectors (or the emergence of a new variety of sectors) occurs. As the formation of new sectors is seen as both place dependent and place forming, further research is needed to understand the mechanisms of “lock-in” of the emergent firms and sectors and their local conditions.

EDITIONARY ECONOMIC GEOGRAPHY

The role of institutions in evolutionary economic geography

Since the beginnings of evolutionary thinking in economics and economic geography, there has been a controversy about the meaning of institutions for economic evolution. In the neo-Schumpeterian perspective, institutional change can be seen as another explanatory approach to economic change, separate from and supplementary to evolution. This is how Ron Boschma and Koen Frenken (2006) pointed to ontological differences and looked for “interfaces” between the institutional and evolutionary perspectives. These authors understand the (spatial) evolution of routines and firms as an outcome of industrial dynamics, and claim that territorial institutions do not sufficiently explain the evolution of both. However, they acknowledge a higher significance of sectoral institutions as part of sectoral and technological systems of innovation approaches. In contrast, critics claim that evolution of technologies, firms, and sectors cannot be fully explained by the economic assumptions underlying industrial dynamics alone but that agency embedded in institutions and local institutional settings as a selecting environment must also be considered. There is, however, also much controversy about the precise meaning of the term “institution.” Discussants generally apply an “old” understanding of institutions, as promulgated by Douglass C. North as the “rules of the game,” although concepts of new institutional economics such as transaction costs can also be used to explain evolutionary processes such as branching.

In fact, many concepts dealing with the evolution of firms and sectors in the neo-Schumpeterian interpretation of evolutionary thinking at least implicitly refer to the presence or absence of institutions related to technology. This is the case with places where windows of
locational opportunity open up when hampering institutions of old industries are absent or weak, or places favorable to the emergence of new firms such as urban agglomerations and certain clusters where institutions are present in knowledge communication between diverse (Jacobs externalities) and related technologies and sectors, enabling local relatedness of industries and branching.

The emergence of or changes in institutions can also be seen as an indispensable part in the explanation of evolution when adopting perspectives on the coevolution of two populations in mutual symbiosis, that is, an industry and related institutions. Here, a bidirectional causality through various self-reinforcing feedbacks brings about simultaneous growth in the evolution of two (or more) populations, for instance the co-emergence of an industrial sector and its related institutions of knowledge creation (as in the case of the rise of technical universities and interpersonal relationships between firms and universities in nineteenth-century Germany). Coevolution has become an important notion and issue in evolutionary research. But again there is often fuzzy use of the concept, despite the quest for a clear definition of bidirectional causality in empirical research.

The development of evolutionary paths

Mutations of routines, technologies, firms, and sectors – and regions – are assumed to happen continuously in the capitalist world of competition through innovation, steadily producing heterogeneity. For a long time, social scientists focused on trajectories or development paths of these units in order to explain the dynamics of uneven spatial development, widely adopting the concept of path dependence and lock-in in a life cycle model. From the evolutionary perspective, selection and retention of the selected variety are necessary for an evolutionary path to emerge, not least because this limits the kind and number of competing varieties that emerge. A path can be considered dependent when past occurrences and given specific social mechanisms condition future choices and the direction of future change, a process called “positive” lock-in of a specific technological, sectoral, and regional trajectory. Interestingly, economic geographers have based the concept of lock-in on arguments both from heterodox economics, such as increasing returns to scale, sunk costs, agglomeration, and network externalities, and from social sciences, such as institutions or other untraded interdependencies (Michael Storper). Path creation in technologies, firms, and sectors – in competition with incumbent structures and nearby alternative novelties – has largely been considered a local process, partly due to an enhancement by powerful actors such as policy agents, risk-taking entrepreneurs, customers, and institutions or certain regulations. Evolutionary economic geography claims that proximities and institutions in some places work in favor of the creation of specific sectors. As mentioned before, some evolutionary geographers have introduced concepts such as technological relatedness and regional branching based thereon.

Once a path has been locked in, questions arise as to how and how far retention of this path really is fixed. Time and again caveats have been expressed related to path dependence being overdetermined, and to possible degrees of freedom in path creation, path bifurcation, and path contingencies. More open concepts of development such as path diversity and path plasticity have been repeatedly suggested. The notion of lock-in has been more fundamentally blamed for referring to mechanisms of continuity rather than change, for failing to differentiate analytically between exogenous and endogenous pressures, and for favoring an understanding of strong
inertia and unexpected disruption/destruction instead of incremental change (Martin 2010). Furthermore, queries arose about the idea of inevitability in the life cycle concept as paths may be subject to rejuvenation or ongoing change. As a result, Ron Martin and others called for a rethinking of the concept and a closer theoretical social scientific underpinning that includes institutions and political economy.

While regional path creation in technologies was a major subject in early evolutionary economic geography with the example of emerging high-tech regions, the decline of technological, sectoral, and regional paths has been less analyzed. At the end of a life cycle, lock-in mechanisms may become a negative factor that protects the old and hampers the emergence of something new. For example, Gernot Grabher (1993) pointed to cognitive, functional, and political lock-ins through strong ties quite early on when explaining resistance to structural change in the Ruhr area. It is, however, far from clear what a life cycle means for a complex system such as a region. From an evolutionary perspective, the mechanisms of path decline or path rejuvenation (Martin 2010) and resilience still remain underexplored in empirical research.

Evolution of regions in a generalized Darwinist approach

Much recent criticism from an institutional or political economy perspective has focused on regional development as an evolutionary process. Generalized Darwinism has been suggested as another evolutionary approach for explaining the evolution of complex systems in the social sciences. The idea seems obvious for economic geography as well, as regions can be seen as complex adaptive systems of multiple populations in variegated environments.

Generalized Darwinism considers variety, selection, and retention as generic principles applicable to any evolutionary analysis of any (social) scientific issue. In this very abstract sense, complex systems may evolve similarly to populations of species through endogenous mechanisms (such as self-organization). However, approaching a specific issue with generalized Darwinism requires a careful translation of abstract principles into operational dimensions such as corresponding replicators, and of the interactors involved. “Higher-level groups” such as regions and cities or other groups such as class or gender may be considered as interacting complex populations in social sciences (Essletzbichler 2012). For these complex systems, individual agency is no longer relevant; instead cooperative and collective agency based on shared visions matters, and therefore the mechanisms of group selection are different. Jürgen Essletzbichler (2012), for example, suggests customs and conventions, routines and institutions, and regional policy as “component replicators” for the analysis of regions as social interactors.

Understanding regions as complex systems touches upon an important query in evolutionary economic thinking, namely the relationship between the micro, meso, and macro levels of analysis in the explanation of economic and regional change. While the neo-Schumpeterian perspective on evolution, as exemplified by Nelson and Winter (1982), has sometimes been mistakenly blamed for its focus on the micro level of the firm only – which nevertheless was intended to explain meso level change in sectors – the generalized Darwinism approach promises to emphasize the meso level, simultaneously respecting its interlinkages with the micro and macro levels. Naturally, these levels can be conceptualized as different geographical scales in economic geography. Hence the argument is for a multiscalar geographical perspective that
ECONOMIC GEOGRAPHY

also integrates institutional change at various levels into evolutionary analysis. This focus calls for careful reflection on social mechanisms that connect the different scales to each other.

The generalized Darwinist approach in evolutionary economic geography is a challenging endeavor that so far is mainly only a desideratum. Empirical evidence is urgently needed. In terms of methodology, generalized Darwinism demands a quite sophisticated systemic approach to evolution. However, generalized Darwinism is also a somewhat contested concept, both in the social sciences in general (Levit, Hossfeld, and Witt 2011) and in economic geography in particular. On the one hand, other partly competing partly overlapping concepts of the evolution of complex systems have been recently suggested, such as path dependency, complexity theory, and even panarchy. On the other hand, in the current debate on the dynamics of regions, “mid-range” issues such as regional adaptation and adaptability, regional resilience, and the development paths of regions do not seem to require generalized Darwinism as a basic concept (Simmie and Martin 2010).

Evolutionary economic geography: an ongoing construction site

Evolutionary economic geographers have presented substantial new insights into the dynamics of economic change and uneven spatial development within quite a short period of time. However, evolutionary economic geography is still a rather heterogeneous approach that calls for much further refinement. As a result, some evolutionary economic geographers conceive evolutionary economic geography as a metatheoretical framework for empirical analyses. Specifications are required for various dimensions. First, there is much overlap with other heterodox approaches such as industrial dynamics, firm demography analysis, innovation systems theory, or general systems theory. Subjects such as knowledge creation, innovation, learning, absorptive and adaptive capacity, life cycle, path creation, path dependence, path decline, lock-in, regional innovation system, learning cluster, and, more recently, resilience, can all be seen in very different perspectives, so that it seems easy to simply use the term “evolution” in a metaphorical sense. Suffice it to repeat, however, that the basics of evolutionary theory are emerging varieties, selection between competing varieties, and retention of the selected varieties, terms that lead to considering and analyzing related social mechanisms and processes of change.

Second, there is much scope for further refinement of arguments, both in the neo-Schumpeterian and the generalized Darwinian perspective on evolution. In the neo-Schumpeterian interpretation of the evolution of firms in a region, queries arose about the appropriate unit of analysis. While some researchers seem to take the firm as interactor theoretically for granted others just see this as a deliberate choice for empirical research. However, the firm is an ambiguous unit, in particular in the form of the “one plant one product” company often implicitly considered in evolutionary economic geography. Taking the multilocational and multiproduct firm into account would require a multiscalar perspective on firm agency that cannot be linked easily to economic evolution in and of regions. Furthermore, coexisting and co-emerging technological paths and their interrelatedness within the firm would require another perspective on time, more on simultaneous than on sequential processes. In analyzing the evolution of technologies at the plant level in regions, some authors avoided these pitfalls (Rigby and Essletzbichler 1997), but this choice
may often be limited by data availability. Moreover, most evolutionary microanalyses focus on the interactors (firms) but neglect to analyze the emergence of replicators (routines and technologies) per se. Technical novelty is generally taken as occurring between firms, not within firms. Consequently, mechanisms of replication and diffusion of routines and technologies take place between firms, in the general understanding, through labor mobility, spin-offs, and – less analyzed – supplier–customer relationships in value chains or mediated by intermediaries such as consultancy and training services or the media. In fact, there is ample evidence in the recent global value chains literature on local economic change driven by technological upgrading through top-down diffusion as opposed to bottom-up creative processes. Having said this, the specific definition of interactors and indicators of replicated variety (as, for example, the incidence of spin-offs, patent use, etc.) in the neo-Schumpeterian view of evolution in economic geography has paved the way to quantitative analysis and model building in empirical research (Boschma and Frenken 2011).

Generalized Darwinism requires higher complexity in empirical research and even stimulates further queries. Obviously, the definition of higher-level groups as the target of empirical analysis calls for an appropriate definition of the micro, meso, and macro scales, which in turn influences the kind of multiscalarity and interscalar interactions involved. For instance, firms simultaneously form part of regions, clusters, sectors, and networks, that is, groups at a meso scale which are not independent of each other. These groups may emerge differently, may be subject to different institutional settings, and may develop routines differently. As a result, mechanisms of variety creation, selection, and retention must be considered to be much more complex. This poses a problem for the methodology of empirical research, in favor of a qualitative approach to nested systems. Much criticism has evolved of generalized Darwinism as a workable approach in economics, blaming it for its imprecision in defining the core terms and explaining economic mechanisms in detail, and for being generally too abstract and imposing its concepts in a top-down manner onto reality in social sciences (Levit, Hossfeld, and Witt 2011). Recent discussions in biological sciences have also put a question mark over the current understanding of the fundamental principles of variety, selection, and retention and the use of generalized Darwinism, making the practice of a generalized Darwinist approach in evolutionary economic geography even more challenging.

Such claims for further development of the evolutionary approaches in economic geography are not exhaustive. Conceptual debate on evolutionary economic geography (or geographies, as some authors prefer to say) is lively. One of the most recent suggestions attempts to avoid fundamental debate about neo-Schumpeterianism “against” generalized Darwinism and prefers a practically oriented approach to subjects of evolutionary dynamism. Referring to Trevor Barnes’s plea for engaged pluralism in economic geography, Robert Hassink (Hassink, Klaerding, and Marques 2014) proposes the incorporation of institutional as well as relational economic geography concepts into evolutionary approaches to strengthen the role of agency and institutions in evolutionary explanation. Some weaknesses in the current empirics of the neo-Schumpeterian approach could be avoided, such as understanding the firm as an isolated actor and the coevolution of institutions as merely occurring intraregionally. While analysis of micro processes would benefit from a relational perspective, the institutional perspective would support the meso level in evolutionary analysis and help to clarify its interlinkages to the macro level of the economy.
References


Further reading


Existentialism

Tom Mels

Uppsala University – Campus Gotland, Sweden

Existentialism is a diverse, anti-essentialist tradition of thought and engagement, incessantly seeking full immersion in everyday life. It claims that any effort to grasp existence entails a concern with the human being in its multifaceted bodily, emotional, cultural, and societal situation. Existentialism therefore insists that human experience, practical life, personal awareness, and the study of life from the inside need to be a fundamental part of scholarly engagement. A wide-ranging, yet distinct conceptual and categorical repertoire is commonly associated with existentialism’s approach to human existence. Notable among these are the intimately related notions of alienation, anguished responsibility, authenticity, commitment, irrationality, boredom, and situatedness. These motifs are recognized as descriptive of the human condition and the complexities of freedom.

Existentialist concerns were developed historically as a critical response to various kinds of deterministic thought. It thus provided an oppositional vocabulary challenging varieties of cultural, economic, religious, and racist essentialisms prevailing in European society and political life of the nineteenth and twentieth centuries. Within geography, the chief interpretation of existentialism has occurred in conversation with versions of phenomenology and various postpositivist currents. Although a heterodox existentialist allure can arguably be felt throughout the discipline’s research on place as a repository of meaning and experience, it remains most frequently associated with the advent of humanistic scholarship in the 1970s. By that time, existentialism and its multiple intellectual and cultural ramifications in visual art, literature, and cinema, had already enjoyed widespread popular acclaim in Europe for several decades.

The emergence of an extended associative community and intense scholarly conversation with Marxism and phenomenology help explain the reputation of existentialism as a loosely defined intellectual movement, presenting nevertheless a passionate sensibility of revolt against social and scientific categorizations and structures. While nineteenth-century philosophers Søren Kierkegaard and Friedrich Nietzsche have been identified as precursors of the movement, and Martin Heidegger and Maurice Merleau-Ponty (both of whom incidentally disavowed the term) pioneered some of its core themes, it was during the 1940s and 1950s that classic contributions to existentialism were made in the philosophical and literary work of Jean-Paul Sartre and his coeals Simone de Beauvoir and Albert Camus. Martin Buber, Karl Jaspers, Gabriel Marcel, José Ortega y Gasset, and Lev Shestov are some of the major figures staking out different varieties of twentieth-century existentialism internationally.

Existentialists typically insist on the deeply social and conventional ground of values, keeping at arm’s length mainstream philosophical positions in which disinterested science and moral claims to what is good and right take center stage. This may best be clarified by taking a brief look at the distinctive ways in which the aforementioned conceptual repertoire of existentialism has been construed.
EXISTENTIALISM

For existentialists, conventional appeals to pre-existing essences (divine will or abstract reason) deny that human existence is always in the making, a constant embodied being-in-the-world, rather than the result of a deterministically fixed framework. Meaning and value rather emanate from life itself, from personal engagement in the world. This prioritization of existence over any pregiven essence confronts us with a kind of historically, socially, and bodily situated freedom. Existentialists think of the committed personal affirmation of these conditions of existence as central to the notion of authenticity.

Although authenticity and freedom remain contentious terms in existential deliberation, they are usually not seen as human faculties or powers, but as processes or achievements. In various ways, they describe pathways by which one appropriates the possibilities of one’s social, historical, and spatial situatedness, achieving authentic, owned existence and freedom or fails to do so. For that reason the terms certainly do not hold any triumphant promise of liberty, let alone any easy justification of existing structures of oppression and domination. Rather to the contrary, for they mark out the burden, anguish, and anxiety of having to be free, of taking responsibility for our self-formation (Sartre 1943). Through engagement and identification with the projects conducted in our lives – rather than the “bad faith” of acting as interchangeable figures in roles dictated by the world – we have a means of starting to recover ourselves from alienation. Since this is always a recovery and self-making in situation, it is necessary to acknowledge how freedom, authenticity, and commitment are at all times entangled in various kinds of circumstances, ranging from the space of the body to geographical and historical context, the conditions of human sociality and forms of collectivity. Heidegger’s famous term for Being, Dasein (to-be-there), not only presupposes a self-interpretation, but also connotes an understanding of the contextual spatial and social situatedness of Being-in-the-world (Heidegger 1927). The possibility of freedom and authenticity, bound to such situations, includes a tight relation to the formation of others. Existence as a form of seeing beyond oneself achieved in actual human experience thus opens the prospect of authentic existential commitment in the world and of bringing into being a human community.

Authentic existence and recovery from alienation also raises the question of history (the temporalities of past, present, and future), and an empirical diagnosis of everyday life in the modern world. By extension, this question more immediately immerses existentialism in the practical matter of politics. Existentialist writing tends to diagnose existence in the face of the modern world as steeped in apprehension, confronting freedom and authenticity with instrumental reason, faceless technocratic rule, religious dogma, capitalist ideology, specialization and fragmentation, and the division of labor. Under such circumstances, genuine individuality will be substituted with what in various ways is described as the inauthentic herd mentality of humanity. Not unlike the Frankfurt School’s critical theory, Sartre was drawn to Marx’s critical analysis of alienation and capitalism as offering a singularly compelling understanding of the lived experience of unfreedom, oppression, and conflict under capitalism. Heidegger, meanwhile, for some time supported the Nazi movement that he thought of as announcing the advent of authentic selfhood, a genuine manifestation of an inner essence of Being in actual history, to later abandon politics for a more spiritual and poetic mood.

Despite the demise of the existentialist cultural movement, its core motifs continue to have prominence in contemporary philosophy and social science, ranging from poststructuralism to
humanist Marxism. Simultaneously, existentialism has been caught with an unclear position in crossfire between structuralist, phenomenologic, and positivist thought. Critics have taken existentialism to task for what is seen as its bleak image of alienated despair and moral defeat. They argue that as long as these conditions of existence are seen as universal rather than as the historically specific condition of capitalist modernity, existentialism naturalizes the very existence it sets out to condemn. Any vision of solidarity with other human beings then gets necessarily thwarted. Another strain of discontent concerns the adoption of phenomenological methods. This resulted in a narrowing of the existential to a notion of nontransferable, unique selfhood, and human consciousness. From a Marxist point of view, existentialism has been found to fetishize freedom and to mystify the construction of reality behind a jargon of authenticity that insufficiently addresses the class-based structural oppressions of capitalism (Adorno 1964).

Notwithstanding such concerns, clear traces of existential philosophy were concurrently bequeathed by existentialism to Marxian scholarship. Moreover, the bulk of this critique can be traced more particularly to the development of existential phenomenology. This is also crucial for what remains a partial reception of existential philosophy within geography.

Although the origins of its philosophical venture emanated from the European contexts and languages of French and German intellectual debate, the first interpretations within geography were made in the Anglo-American context. In the early 1970s, existentialism was partially absorbed in the undogmatic eclecticism of the humanistic turn within geography, joining an increasingly forceful critique by Marxists and other radicals against the prevailing positivism within the discipline. Not unlike the early and mid-century existentialist movement, humanist geography was an existential response born of discontent with the various determinisms it identified in modern geography. With the positivism and abstractions of spatial analysis, the essentialist approaches within cultural geography and the determinism of structural Marxism were seen as drifting further and further away from the complexities of human agency.

According to the proponents of humanism in geography, these explanatory methods, models, and law-finding ideals were a far cry from the complex lived experiences and relationships between people and their environment. They claimed that positivism had unduly limited geographical research to quantitatively measurable, material spaces. It had reduced the wealth of place to the diagrammatic simplicities of abstract space and reduced human complexity to the robotic schematics of economic man. As a corollary, many humanistic geographers also rejected the avowed neutrality of scientists with their ideals of distant objectivity and elitist views of scientific knowledge. Instead, combining elements of existentialism with a philosophical and methodological menagerie of hermeneutics, idealism, and phenomenology, they brought out a fuller spectrum of subjectivity in knowledge production. Meaning, values, experiences, and the imagination became keywords in their vocabulary and tools for the qualitative exploration of the sense of place, everyday life, the lifeworld, home, rhythms, and existential insideness and outsideness.

Existential phenomenology was most commonly used to explore the experiential depth of practically living in places: the sense of place shaped through our immersion in the everyday geographies of houses, offices, schools, and parks. For instance, Yi-Fu Tuan (1974) explored an array of human awareness of, affective and bodily ties with, and responses to the environment. Merleau-Ponty’s approach to embodied presence
Existentialism

and Heidegger’s example informed his method to uncover meaning and values associated with place. Embodied experience and the pre-reflective body-subject were treated with considerable detail in David Seamon’s work (e.g., 1979). Anne Buttimer’s existentially aware geography drew inspiration from architectural theorist Christian Norberg-Schultz and Alfred Schütz’s phenomenology of the social world, but also from the French geographical tradition on social space. Her work addressed, among other things, questions of place, rhythms (temporal and spatial zones of reach), and ideas of dealing with the everyday lifeworld and embodied knowledge. Buttimer’s work (e.g., 1993) pioneered bringing out the humanity of the researching geographer, with meditations on values, dialog and responsibility, meaning construction, and other epistemic and ethical questions. A strong sense of sociality was also negotiated by David Ley (1974), whose urban social geography opened windows on the interpersonal and intersubjective values and meanings were instilled into and distilled from existential space. In the late 1970s, Marwyn Samuels developed an explicitly existential geography influenced by Martin Buber’s spatial thinking about existential nearness and detachment, measured not in actual physical distance, but in terms of the perceived distance of human relationships. Existential geography would be about the interpenetration of people’s subjective attachments and historical relationships with the landscape.

Meanwhile, such empirical and theoretical exertions were not infrequently coupled to the ambition of developing a sustained critique of technocratic societies and their urban planning ideals and alienating practices. Edward Relph (1976) developed the argument that in advanced industrial society, authentic senses of place have by and large been replaced by inauthentic modes of existence. The result is an alienating placelessness in which a burgeoning existential outsideness – espoused by the dominant values of modern technologies, the mass media, commerce, state authority, and a space of uniformity – takes precedence over erstwhile conditions of existential insideness, which was felt through rich symbolic meaning, identity, and belonging to place. Existential critique is also prominent in Gunnar Olsson’s long-standing critique of taken-for-granted meanings and power imbuing positivist knowledge, modern planning, and the bureaucratic state. His enduring engagement with art and literature provided sources for critical inquiry into the interstices between individual and collectivity, mind and materiality, and notions of marginality, the body, alienation, and the absurd (Olsson 2007).

These existential forays into the geographical embeddedness of human subjects were derived from an array of philosophical sources – ranging from structuralism to phenomenology – and occasionally also from the intellectual heartland of existentialism. Although it hardly came to represent an agreed-upon approach, Edmund Husserl’s phenomenology may be singled out because it attracted considerable approval from existentialist quarters. Within the social sciences and humanistic geography the idea of consciousness as an immediate, meaningful openness to the world proved attractive. Conversely, Husserl’s bifurcated image of cognitive consciousness (the transcendental ego) standing outside the active, lived world of embodied, human existence remained unacceptable for many. Roughly along the lines of such a digestion of Husserl, existential phenomenology within geography redirected attention to the meaning and concrete reality of the human lifeworld. It prioritized intuitive methods and phenomenological description (devoid of the preconceptions of scientific understanding or common sense) over observable sense data and scientific explanation.
Unsurprisingly, in its often contrasting searches for the rethinking of key geographical methods and concepts such as place, space, and landscape, the existential response within human geography has been subjected to critique. The first line of critique is leveled on existential phenomenology’s methods. The bifurcated staging of, on the one hand, phenomenological description and intuition, and, on the other hand, scientific observation and explanation has been questioned. This critique often blends with the perennial philosophical controversy around the viability of a return to original experience and direct description of the world, devoid of preconceptions. By extension, the intuitive aspects of existential phenomenology’s method have been rejected as excessively subjective and relativist.

A second line of critique is related to these questions of method, but also reaches considerably beyond them. In particular from Marxian and feminist vantage points, the privilege given to everyday experiences, intentionality, and subjective interpretations has been questioned. Such privileging tends to underestimate the degree to which unequal and unjust social, political, and economic circumstances and structures – the capitalist mode of production, ingrained gendered inequalities – guide everyday life, self-understanding, and the like. According to this appraisal, the focus on place experience in existential phenomenology rested on an essentialist view of the human subject that in turn diverted attention away from structuring forces of class, gender, ethnicity, and other unevenly reproduced power relations. What transpires is not a critical geography of place and the exposure of various kinds of marginalization and oppression, but a romanticized, idealistic, and nostalgic view of place (often in the form of “dwelling”) contrasted with an alienating present. In this context, the contrasting pairings of place – space, insider – outsider, authentic – inauthentic in some existentialist writings have been perceived as expressions of a dualistic worldview. Equally, scholars identifying their work with existential phenomenology have rebuffed many of these issues as misconceptions of the existentialist intent.

Both genres of critique have implications for the awareness of values in geography. On the one hand, the humanist preoccupation with values in research processes has been accused of ignoring how societal pressures shape research agendas. But on the other hand, these disputes over method, values, and epistemology were of seminal importance to subsequent attention to the partiality and situatedness of scientific knowledge production and representation within geography. With its interest in the complexities of emotional response, and how feelings and values imbue experience and human agency, existential geography shared important traits with early feminist geographies.

Today, existentialism within geography remains primarily associated with the existential phenomenology of the 1970s humanistic turn. It is important, however, to recognize that the selectivity of this particular embrace easily underestimates the alternative routes through which the existentialist movement has entered geography. Although shared themes need not entail shared diagnosis, fascination with modernity, alienation, freedom, and situatedness all bear testimony to the circuitous influence of some of its core insights and analyses into the plethora of positions within postpositivist geography and during the 1980s and beyond. In recent years, the development of qualitative methods for a wide variety of human experiences has attracted renewed attention. Existentialist heritage can arguably be identified in the increasing amount of scholarship that is now being devoted to Heidegger’s work and Merleau-Ponty’s thorough reinterpretation of embodied experience, under
EXISTENTIALISM

the heading of emotional geographies, nonrepresentational theory, and studies of perception.

SEE ALSO: Bodies and embodiment; Emotional geographies; Humanistic geography; Identity; Phenomenology; Place; Radical geography

References


Further reading

Exploitation

Jim Glassman  
*University of British Columbia, Canada*

Exploitation is a term that owes much of its importance to the varied and sometimes incompatible ways in which it is used, ranging from highly charged moral claims about human conduct to technical claims about the labor process. The range of these claims ensures that exploitation is a performatively significant term within political debates.

In its broadest and most morally charged uses, exploitation implies abuse of another, taking more from another than is normally considered socially acceptable. Frequently, this notion is applied to the use of another person’s labor without adequate compensation being given. Thus, wages that are extremely low compared to what many would find adequate are referred to as exploitative, as sometimes is the disproportionate performance by women of unpaid household labor. In this broad moral sense, the abuse of nonhuman nature by humans can also be seen as exploitative.

Within Marxist theories, the term “exploitation” takes on a more technical meaning. Marx did not regard the extraction of surplus value from wage labor to be abnormal under capitalist conditions of production, and in this sense exploitation was for Marx an analytical category, not a moral one (Heinrichs 2004; Harvey 2010). His argument was based on the notion that labor is the sole producer of value in capitalist economies, and that the surplus controlled by capitalists in the form of profit must be derived, ultimately, from wage labor. Exploitation thus refers to the difference between the total amount of time workers labor and the socially necessary labor time required for workers to reproduce themselves at a historically determined subsistence wage, which Marx assumed to be what workers normally received (Marx 1977). That difference is the surplus labor time appropriated by capitalists, or surplus value, and any increase in the amount of surplus labor time relative to the labor time necessary for reproduction is an increase in exploitation, independently of whether or not the historical subsistence wages paid workers are considered by some observers to be morally acceptable or unacceptable.

While this notion of exploitation within the labor process has been central to most Marxist analyses, the geographies of labor and surplus-value production associated with globalizing capitalism complicate estimates of who is being exploited and to what degree. If all labor took place within one capitalist firm and one location, it would be relatively straightforward to estimate a rate of exploitation. But if capitalists produce in different places around the world, if they pay workers differentially for the same amount of labor and similar levels of productivity, and if they can shift the surplus from one location to another while potentially paying more to workers in certain locations, the possibility arises that geographic transfer of surplus will allow workers in one location to reduce their levels of exploitation at the expense of other workers (Emmanuel 1972; Sheppard and Barnes 1990). This possibility of some workers benefiting from exploitation of other workers underpins notions of imperialism through unequal exchange, championed by certain neo-Marxists, such as world systems theorists (Taylor and Flint 2000).
EXPLOITATION

SEE ALSO: Imperialism; Marxist geography; Surplus labor; World-systems theory

References


Further reading

Exploration can be broadly conceived as a process of building new knowledges through searching for and encountering new environments, people, and places. Although processes of exploration are integral to other disciplines as well, exploration has been most closely connected to geography through time; however, exploration does not begin with the discipline of geography, nor does it even necessarily begin with modern humans. The earliest forms of hominid exploration central to the migration and settlement of the human species took place at least 90,000 years ago when early hominids (*Homo erectus*) left East Africa. Taking into consideration the fact that every continent except Antarctica was settled over a 75,000-year span from those initial migrations, our ancestors were undoubtedly ambitious explorers. The methods of prehistoric migration, whether by land or sea, are still a subject of some dispute—indeed, some have argued that Australia was populated by ocean voyagers as early as 45,000 years ago, though early archaeological remains that could bolster this claim are unlikely extant. What is clear is that many of the earliest recorded historic voyages were undertaken at sea. As early as 670 BCE, the Egyptian Pharaoh Necho II dispatched a Phoenician expedition to circumnavigate Africa over three years.

The enduring relationship between geography, exploration, and science began forming early in the “Age of Discovery” from the fifteenth to the seventeenth centuries. This period of exploration may not have been any more ambitious than earlier eras, but manifested in large-scale public displays of power through financial investment, deployment of new technology, and, perhaps most importantly, a machine of publicity to document it all. The age of exploration and empire, perhaps most closely associated with the scientific voyages of Captain James Cook, relied not just on embodied exploration and encounter but on systems of documentation such as writing, painting, and mapmaking to establish colonial territories, filling in knowledges on what Europeans perceived as *terra incognita*. Geographers were complicit in contributing to European military expansion as part of the larger Enlightenment cultural project. Over this time, military and religious conquests installed European nation-states over much of the world, including the Americas, the Pacific, and Asia.

In the late nineteenth century, increasing pressure through competition to colonize more of the globe led to what has been termed “the Scramble for Africa”: the invasion, colonization, and annexation of the African continent from 1881 to 1914 (Driver 2001). Many of these African expeditions were sponsored by the Royal Geographical Society (RGS) in the United Kingdom (founded in 1830). Similar expeditions have been sponsored by the National Geographic Society in the United States of America (founded in 1888). Both institutions, as a result, have been embroiled in debates over the ethics of their involvement in imperial projects. It became clear to many in the
EXPLORATION
twentieth century that the tools of geographic exploration were the tools of empire. Because of these associations with governments, military, and spectacular cultural tourism, exploration as a concept became increasingly stripped from academic geography, seen as anti-intellectual endeavor. This view was reinforced by explorers themselves, who accused geographers attempting to theorize practice of being “armchair” academics, lacking the experiential credibility to guide narratives of exploration. Whatever the politics behind such expeditions, they were certainly successful from the point of view of European powers, leading to suggestions by the early twentieth century that there was nothing left to explore in the world; that humankind had entered the age of Geography Triumphant (Driver 2001, discussing a 1924 article in National Geographic magazine by Joseph Conrad).

Inevitably, however, a second land scramble of a slightly different nature began in the twentieth century and geographers were once again at the forefront of exploration under the guise of objective science. James Cook’s second expedition from 1772 to 1775 had fueled curiosity about Terra Australis, a new continent in the Southern Ocean. Between 1898 and 1910, seven major expeditions by European powers were launched onto the Antarctic landmass, the only continent on Earth without an indigenous population (Dodds 2012); however, this did not mean these explorations were without conflict. As with explorations in earlier ages, these expeditions were heavily financed by governments with political motivations leading to serious disputes over territorial claims. As with the scramble for Africa, in this period, the claims of nations were often pinned on “heroic” explorers. Norwegian Roald Amundsen was the first to reach the South Pole in 1911, to the dismay of British explorers who arrived just a few weeks later. Amundsen may have also been the first person to reach the North Pole in 1926 via airship. With both poles then mapped, the “final” pole sought was the highest. Mount Everest (the highest mountain from sea level but not the tallest from the sea floor) was summited by Tenzing Norgay and Edmund Hillary in 1953, who captured photos as evidence of their accomplishment. As with the mapping of Africa, many once again declared the “end” of exploration by the 1950s.

Debates and issues around exploration

As on Everest, visual representations have always been important to exploration. Many prehistoric cave paintings may have been depictions, or visual maps, of explorations into new territory. In the age of discovery, representations took the form of resident artists (painters) hired to depict journeys. With the invention of photography in 1839, the medium was recognized almost immediately, in the context of the expansion of European overseas empires, as an asset to explorers seeking to document their journeys and achievements (Ryan 1998). From the mid-nineteenth century to today, exploration photography has become ubiquitous and expected. Photographs were thought to capture something of an objective “truth” of an expedition, event, or encounter, perhaps more accurately than painting where the painter’s subjective interpretation of events comes through more clearly in the framing, textures, rendering of expression and gesture, and choice of color. Indeed, from the mid-nineteenth century, it was difficult to conceive of an exploration expedition without a photographer, or at least a camera (still or later video) as part of the field kit, a relationship that remains to this day. And although many still consider images “proof of achievement,” more contemporary geographic work has considered overlooked accounts of
images, the neglected home archives, and potential for additional mining of known archives to tell stories of individuals not foregrounded or those just out of frame.

These stories of exploration, while inspiring, are also stories of exclusion. The history of exploration until the twentieth century is a story largely told from a Eurocentric, androcentric, and science-centric point of view. Sustained critique followed from within geography to celebrate indigenous accounts of exploration and accounts of European exploration from indigenous perspectives. These narratives, many scholars argued, had been written from perspectives that failed to articulate the meanings and logics of non-Western peoples. Demands were also made to recognize the active exclusion of women from these expeditions and to rewrite history where women were involved and subsequently written out of history. In the history of Victorian exploration, for instance, as Mona Domosh (1991) convincingly argues, there were a number of women explorers written out of history because their explorations were not “goal oriented” or because their views and activities did not accord with geographic and scientific standards outlined by institutions such as the RGS. The RGS actively excluded women from membership and decision-making; women like Mary Kingsley, Mary Gaunt, and Isabella Bird were accused of being “travelers” rather than “explorers,” a gendered language divide still evident today. Interestingly, in the act of privileging process over goals, as many women who explored were accused of doing, these explorers often contributed more deeply to ethnographic and cultural understanding through their “subjective goals” than their male counterparts. The traveler/explorer divide highlights the always-present truth that exploration is as much internal as external, that exploration of the world is always an exploration of the self.

New themes in exploration

As a result of these sustained and important critiques, exploration fell even more deeply out of favor within geography. Few geographers embedded themselves on expeditions and even fewer labeled themselves “explorers” for fear of being perceived as anti-intellectual or, even worse, racist, sexist, or neocolonial. However, an American geographer called William Bunge disrupted this view in the 1950s; in short, Bunge argued that exploration could once again be a positive force, turned on ourselves. In the midst of stark racial and spatial division in 1969 Detroit, Bunge began undertaking what he called “expeditions” into low-income parts of the city under the auspices of the Detroit Geographical Expedition and Institute (DGEI). He argued that complete and full exploration of the urban environment would lead to democratization of space and a cultivation of deeper senses of community. Bunge offered tuition-free urban geography classes to low-income African-Americans and bused middle class white students into poorer areas of Detroit so that they could see how urban segregation was shaping their worldview and reinforcing inequality. A radical explorer for his time, Bunge was dismissed from multiple academic posts for his explorations. However, his message, that geographers should actively participate in the communities they are studying and strive to enact positive social change, has had lasting effect on the discipline and our considerations of the role exploration can once again play in geography.

The related philosophies of the Situationist International (SI), active in the 1950s and 1960s in Paris, were founded on the notion that due to the relentlessness and pervasiveness of commodity and administration under late capitalism, urban citizens had become passive spectators in a world presented to them as a
EXPLORATION

spectacle (Pinder 2005). Their solution to this problem was to reinstate active participation through explorations of the city in a process that might be considered a “counter-mapping” of place. Explorers were encouraged to undertake dérives or drifts through the city in the spirit of nineteenth-century flâneurs who refused to adhere to notions of “proper” places. The dérive opens out, in the spirit of exploration, chance for happenstance, which may lead to unexpected spatial discoveries and social encounters. Importantly, these explorations seek to redress earlier problems in exploration stemming from perceived overly objective and scientific goals by embracing the subjective process of exploration. And importantly, like Bunge’s expeditions, these were explorations not of distant lands or “foreign” people, but of the “home” and the “everyday.” The maps of the SI, drawn up from exploration of the exceptional in the everyday, were meant not to document and rationalize space but to disrupt social ordering, to draw up hypotheses for new constitutions of place that prefigured the social before the economic (see Bunge 1969).

The SI praxis was further drawn out through urban exploration, a globally recognized and socially significant practice from about the year 2000 to today (Garrett 2013). Urban explorers appropriated the language and imagery of colonial exploration and applied it to contemporary metropoli, which they claimed had become colonized by forces of capitalism, surveillance, and health and safety to the point of existential threat. Urban explorers called for subversion of dominant narratives of spatial regulation by undertaking explorations in cities without permission, sneaking into abandoned buildings, under construction skyscrapers, and subterranean tunnel systems, photographing the “discoveries” and distributing those photos publically online. The emulation of the colonial “explorer” ethos in the culture extended into the declaration of “Golden Ages of Exploration” (some lasting only months), the naming of locations by explorers of “virgin territory,” and the aesthetic coding of the photography used to “document” their explorations. The appropriation of these languages and aesthetics is both a serious political statement about access to and control over space and a flippant sapping of colonial practices. It is in this collapse of binaries that urban exploration finds its form, being both an objective technique of observation that reveals and comments on the “hidden” in the tradition of exploration and an inclusive, internal subjective search for meaning. Urban exploration is closely allied with practices such as skateboarding, graffiti and parkour (free running), which also constitute modern forms of exploration of cities. These practices similarly colonize and occupy spaces in the context of popular practice rather than colonial exploitation, being undertaken by local groups and individuals without state or institutional support. Ironically, the state and connected institutions, who clearly sponsored (and continue to sponsor) far more militant and violent forms of exploration, often condemn these exploratory practices “at home,” as being inappropriate, illicit, and even illegal. It would appear that not much has changed since Bunge’s time – exploration is still meant to remain “Out There,” among “The Other.”

Within geography there has been a preoccupation with mapping, Cartesian coordinate systems and “accurate” photographic depictions of places and landscapes that amounts to, in essence, a two-dimensional imagination, which is being challenged as geographers explore the vertical world through practices such as caving, scuba diving, tunneling, mining, air travel, ballooning, and vertical architecture. Ambitious as these studies may be, attachment to the “geo” has meant geographers often still remain remiss in
their attention to exploration of the deep sea and outer space, indications that perhaps vertical imaginaries have not stretched far enough. With the boundaries of deep-sea exploration continuously advancing through the use of robotics, there remains vast work to be undertaken on oceanic and maritime exploration. And beyond soil and sea, decades of concerted space exploration have also opened out new possibilities for discovery beyond Earth’s exosphere. Overwhelming public interest in the photographs taken by the Mars Rovers, life aboard the International Space Station, and the possible colonization of Mars in the next decade should have geographers working overtime to consider the geopolitics of space exploration and the importance of the current political moment in the galaxy, where control over and exploration of outer space is being handed over from governments to private companies. Given the discipline’s relationship with historic precedent, and the particular sensitivities that has engendered, geographers are well poised to critically comment on such activities. Additionally, attention to new spaces of exploration such as these, in the context of work by people like Bunge and the SI, should be seen as an opportunity for researchers to embed themselves on such missions and tell new ethnographic tales of exploration, critically filtered through our fraught past with exploratory practices (Naylor and Ryan 2010).

We might also consider the limits of exploration in a more (or less) phenomenological sense. For the endless accumulation and storage of information, again another proclivity of living under scientific frameworks emphasizing collection and retention, have also created new realms for exploration in the archive. It may be easy to imagine the physical archive or library as a place of exploration, a “great indoors,” digging through old manuscripts, maps, and photographs; indeed, many have written evocatively about the process of “delving” for material and having unexpected encounters with fragments of the past (Lorimer 2009). Further interrogation of the relationship between fieldwork and archives or what Felix Driver termed the field and “the cabinet” (Driver 2001) leads to a suggestion of a particular hybridity at work. Exploration of the archive rehabilitates archival fragments and lends itself to performance of landscape, place, and memory that might also be considered a form of activism. Archival exploration can answer the call of Domosh (1991) to recall the hidden histories of those who have been inadvertently or purposefully written out.

Archives can be “delved” beyond the obvious in a number of ways. First, one might consider, as Lorimer (2009) suggests, how Internet-based shopping on sites like eBay allows for collection of archives we may have never had access to previously and may encourage people to explore in an effort to find saleable cultural remains such as family photograph collections (another area of heated ethical debate). Second, the Internet most of us know and use, which is essentially an infinite archive, is only 0.03% of what is on the web. Behind, under, and entangled within this archive is a “deep web” where information accumulates without being categorized, tagged, organized, listed, and publically distributed. The daring virtual explorer or hacker may make forays into this zone, sometimes at great personal risk, to recover social and cultural data. Digging even deeper, one may find oneself in the “dark web,” an ethically and legally murky landscape where important social research can and should be undertaken. Consideration is also due to increasingly multisensory inhabitations of virtual avatars and virtual environments. Virtual “sandbox” environments in games, for instance, many of which take months or years to explore, may soon become boundless. Within the next few decades, it may become increasingly difficult
to distinguish between “real” and “virtual” experiences as our “second life” online becomes indistinguishable from the first, changing our notion of exploration indelibly.

Finally, we might consider new research into processes of nonhuman exploration. Now that machine capacity is outstripping human capacity to explore, and humans increasingly embody machines to explore inhospitable environments, the boundaries of exploration will inevitably change. But this is nothing new, for as we have seen, if anything defines exploration, it is its constantly changing nature. If we knew what was going to happen, it wouldn’t be exploration.

SEE ALSO: Archival and document research; Borders, boundaries, and borderlands; Geopolitics; Historical geography; Imperialism

References


Exploratory spatial data analysis

Diansheng Guo
*University of South Carolina, USA*

Geographers acquire new knowledge by analyzing observation data, searching for patterns, testing hypotheses and formulating theories with insights from data. Spatial data analysis methods can be divided into two general groups: confirmatory analysis methods and exploratory analysis methods. Confirmatory methods normally require an a priori model, a statistical measure or an hypothesis, and use observational data to configure the parameters of the chosen model or statistically test the hypothesis. For example, it is possible to configure the coefficients of a linear regression model or statistically test a specific measure, for example, Moran’s I statistics, to detect the presence of a certain type of patterns in the input data. One shortcoming of confirmatory analyses is that they depend on the user to choose a model based on domain knowledge and the analysis cannot suggest alternative models if the chosen model does not correctly represent the phenomenon being analyzed. In practice, however, it is often not known what the true model is or which hypothesis should be tested. Therefore, it is more challenging and important to find the appropriate hypothesis or model (Tukey 1977, 1980).

Exploratory spatial data analysis (ESDA) uses visual graphics (including maps), exploratory statistics, and computational algorithms to enable the analyst to interact with the data, identify patterns, search for hypotheses and discover unknown or unexpected information. Exploratory analysis is often an iterative process, as preliminary findings may prompt the analyst to change data input, perform new analyses, and revise hypotheses or models. The importance of exploratory spatial data analysis is rooted in the complexity of geographic problems. Geography concerns the understanding of the dynamics and interactions of the Earth’s landscapes, peoples, and environments, which form a highly complex system. Therefore, geography heavily relies on the analysis of observational data to acquire new knowledge and to understand geographic processes, which inevitably involve many known and unknown factors and may also change over space and time. Spatial analysis by nature is an exploratory process that incorporates domain knowledge and analysis methods to comprehend the complexity of geographic phenomena with limited (although big) observational data.

In statistics, exploratory data analysis (EDA) was promoted by John Tukey to explore data and formulate hypotheses that could lead to new findings and experiments (Tukey 1977). Commonly used statistical graphics and methods include, for example, box plot, histogram, scatter plot, stem-and-leaf plot, and parallel coordinate plot, and multidimensional scaling. ESDA is an extension of EDA for the detection of spatial patterns and formulation of hypotheses based on the geographic distribution and relationship of the data. Traditional exploratory spatial data analyses are primarily visual and centered on a map. One classic example is John Snow’s study on cholera, in which he mapped the locations of deaths related to cholera during an epidemic in London in 1854 and found a high concentration...
of deaths near the water pump at the intersection of Cambridge and Broad Streets. Following Snow’s suggestion, the officials removed the handle of the Broad Street pump that supplied water to this neighborhood and the epidemic was contained. Today, the literature of exploratory spatial data analysis has become very broad and diverse, with methods developed in a variety of domains including geography, statistics, computer science, data mining, information visualization, human–computer interaction, and a wide range of related spatial sciences such as biology and geology.

Spatial data types

Before introducing various exploratory spatial analysis methods, it is important to understand spatial data, which can be classified into four major types from the analysis perspective: lattice data, geostatistical data, point event data, and spatial interaction data. Different data types require different methods for analysis.

Spatial lattice data are counts or averages of a measurement for a finite set of spatial units within a geographic area. Examples of spatial lattice data include remote sensing images (i.e., a set of pixels with observation values), cancer mortality rates for states in the United States, the presence or absence of a plant species in square blocks laid out in a forest, or the population of each block in a city. The spatial units in lattice data are fixed and represent a non-overlapping partition of the study area. The analysis of lattice data focuses on the attribute values, multivariate relationships, and their spatial distribution patterns.

Geostatistical data in theory can be observed at any location (thus form a surface), although the data are usually available for a set of selected locations, such as weather stations. The analysis of geostatistical data include, for example, the interpolation of the continuous surface and the prediction of values at unknown locations based on the limited observations.

Point event data are a collection of events that occurred at a discrete location and time, such as crimes, trees of certain species, and disease cases. The spatial locations of point events are often the most important piece of information and the focus of analysis, such as point pattern analysis or spatial clustering. Point event data may be aggregated to lattice data given a set of spatial units (such as the number of disease cases for each county) or be transformed to a surface (e.g., crime density surface).

Spatial interaction and mobility data represent the movements and flows from among a set of locations, such as migration, commodity flows, animal movements, vehicle trajectories, and human daily activities. With the ubiquitous location-aware technologies, such as smart phone and GPS, big data on geographic mobility become increasingly available, which present both unprecedented opportunities and challenges for spatial data analysis.

Challenges for spatial data analysis

Spatial data have a number of unique characteristics that pose problems and challenges for analysis, which nonspatial analysis methods often fail to address. A brief introduction is provided here to several of the main problems and challenges that spatial data analysis often attempts to address.

The spatial dependency or autocorrelation problem refers to the fact that spatial data often
EXPLORATORY SPATIAL DATA ANALYSIS

exhibit a certain degree of spatial autocorrelation, where the measurement at one location is similar to that of nearby locations. On one hand, the existence of spatial autocorrelation violates the data independence assumption (i.e., observation data at different locations are independent of each other) that many nonspatial analysis methods assume. For example, regression analyses that do not take into account spatial dependency can produce unstable parameter estimates and biased models. On the other hand, the detection of spatial autocorrelation by itself can be an important finding for applications that look for spatial distribution patterns such as disease clusters or crime hotspots. A number of spatial statistics and analysis methods have been developed to detect the existence of spatial autocorrelation or to carry out analysis that exploits spatial autocorrelation as a source of information, such as spatial regression methods and spatial interpolation methods.

The second challenge is the spatial heterogeneity or nonstationarity problem, which means that a relationship, model, or pattern may change over the geographic space. For example, the parameters for a regression model may change from place to place, or the values of a statistical measure (e.g., Moran’s I) may be different in different areas. To address the spatial nonstationarity problem, various local spatial analysis methods have been developed, such as local point pattern analysis, local spatial statistics, and local regression analysis, for example the geographically weighted regression (GWR).

The small-area problem in spatial data analysis is also well known. The term “small area” refers to a geographical area of either small size (e.g., area) or a small demographic count (e.g., population). For example, disease rates (i.e., the number of disease cases divided by the population at risk) estimated for units with a small population are usually unstable and exhibit much higher variance than those of larger units, which consequently cause difficulties for discovering true spatial patterns and relationships. The fundamental question is how to produce reliable estimate measures for smaller areas, such as means, counts, and rates, based on very small samples taken from these areas. The general solution to the problem is to borrow information of related data from spatial neighbors. Confirmatory approaches to small area estimation include various model-based methods such as multilevel regression models and Bayesian methods (Pfeffermann 2002). Exploratory spatial methods that address the small area problem include, for example, spatial smoothing, spatial interpolation, and kernel density estimation.

The modifiable areal unit problem (MAUP) has long been acknowledged in the literature of spatial analysis (Openshaw 1984). For example, for confidentiality reasons, spatial data are often aggregated to high levels based on administrative boundaries (e.g., counties) or other spatial tessellations. MAUP refers to the fact that spatially aggregated measurements (e.g., median income or racial percentage) and their spatial distribution patterns may change dramatically if different aggregation boundaries are used, which is caused by both scale and aggregation effects. The scale effect relates to the size of the spatial units being used (e.g., county vs state) and the aggregation effect arises from how the boundaries are drawn. Changes in either scale or unit boundaries will lead to changes in the data distribution and, thus, may lead to different analysis results. MAUP can be considered as a special case of the ecological fallacy, which concerns the potential problem for making inference about individuals based on aggregated data, for example, by ethnic groups, socio-economic classes, or geographic regions (Robinson 1950; Greenland and Morgenstern 1989).
In addition to the above well-known problems, there are also a number of major challenges that exploratory spatial data analysis can help address in analyzing the fast-growing complexity and volume of spatial data. First, spatial datasets are often of high dimensionality (i.e., having a large number of attributes or variables), which can cause serious problems for most analysis methods. For example, there are dozens or hundreds of variables in the census data and/or other data sources, which, for example, can be analyzed to examine how socioeconomic factors affect health outcomes. However, including too many irrelevant variables in a model will lead to unsatisfactory results and miss the true patterns. Most confirmatory analysis methods depend on the user to select factors for model construction or hypothesis testing. Since it is not often known beforehand which factors are important, exploratory spatial data analysis can be used to help identify important factors through: (i) visual approaches, such as univariate, mapping, multivariate mapping, and interactive feature selection assisted by statistical measures; and (ii) computational analysis, such as automatic feature selection methods, to search and evaluate subsets of variables. Feature selection methods are traditionally used to select a subset of variables for supervised classification problems (Liu and Motoda 1998).

The second general challenge in spatial data analysis concerns the complexity of potential patterns and various forms that they may take. For example, lung cancer incidence rates may be related to the interaction of behavior choices (e.g., smoking), local climate (e.g., temperature and humidity), socioeconomic status, access to health care, and other factors. Moreover, the relationship among these factors may be linear or nonlinear and may also change over the geographic space. All possible relationships or patterns form a hypothesis space. Ideally, it would be possible to identify the best model in the hypothesis space and then configure the model with appropriate observational data. Of course, this is practically infeasible and most (if not all) analysis methods limit or compress the potential hypothesis space by assuming a simple form of model, which can be configured with several parameters. For example, a linear regression analysis assumes that the relationship is of a linear form, that the observations are independent from each other, and that the relationship does not change over the geographic space. Exploratory spatial data analyses attempt to search in the hypothesis space without imposing an a priori hypothesis and assess a large number of hypotheses or patterns simultaneously.

The third challenge for spatial data analyses is the massive volume of today’s data with (i) details at a fine spatiotemporal scale, (ii) a large number of observation variables, and (iii) dynamic connections among locations and multiple attributes. Moreover, many research problems and tasks demand near real-time insights, such as for emergency response and decision-making. To achieve the capability of processing large data volumes efficiently and detecting complex patterns accurately, exploratory spatial analysis methods often integrate computational algorithms with visualization approaches. Such integration leverages the computing power and accuracy of algorithms with the inference and problem-solving capability of human expertise to addressing complex problems.

**Smoothing, density estimation, and interpolation**

**Spatial smoothing**

Spatial smoothing can be applied to spatial lattice data, such as area-based attribute data and
pixel-based images, to re-estimate observation values of small areas by considering the values in their neighborhood. For example, smoothing can be used to obtain a more reliable disease rate for each area by recalculating the rate using the total number of disease cases and the total population at risk in its neighborhood, which should be sufficiently large in terms of population or the number of cases. Broadly defined, spatial smoothing also covers various image filters, such as low-pass and high-pass filters, which are commonly used to extract different features from images. In general, a smoothing method involves three components: a neighborhood (e.g., defined by a bandwidth, first-order neighbors, or k-nearest neighbors), a weighting kernel (e.g., uniform, triangular, or Gaussian kernel) (Figure 1), and a filter operation (e.g., weighted average, mode, or other transformations).

Spatial smoothing can help remove spurious data variation, improve the signal-to-noise ratio, and thus help discover true patterns or important features in data. This benefit is maximized when the neighborhood size and shape is the same as the scale and shape of the true patterns. The bandwidth of the kernel can exhibit a strong influence on the smoothing result and, thus, should be configured carefully. Since the true scale of patterns is usually unknown, a sequence of different bandwidths may be tried to examine how patterns change across a scale. A larger bandwidth can help discover generalized high-level patterns at the sacrifice of spatial resolution and the risk of missing local patterns.

There are two main types of bandwidth: fixed and adaptive. With a fixed bandwidth, the radius is normally a geographic distance that is assumed to be the same throughout the dataset. An adaptive bandwidth allows the radius to vary from one data point to another. It is widely acknowledged that a fixed bandwidth causes biased estimations for most spatial datasets, as the underlying data density often exhibits significant spatial heterogeneity. Various adaptive bandwidth selection approaches have been developed to help configure bandwidth, which can be categorized into model-based and domain-based approaches. With the model-based bandwidth selection, a statistical model or criterion is often used to provide guidance on selecting an appropriate bandwidth value. Cross-validation (CV), Akaike information criterion (AICc) and Bayesian information criterion (BIC) are among the commonly used criteria. Domain-based bandwidth selection approaches usually rely on domain knowledge to choose a relevant attribute (e.g., population) to determine the bandwidth. For example, to adapt with the underlying population distribution, a population threshold can be used to determine the adaptive bandwidth, which stops expanding when the threshold is reached.

**Figure 1** Three kernel functions. (a) Uniform: \( W_{si} = 1 \text{ if } d_{si} \leq B_s \), else 0, where \( B_s \) is the bandwidth at location \( s \) and \( d_{si} \) is the distance between \( s \) and its neighbor \( i \). (b) Triangular: \( W_{si} = 1 - | d_{si} / B_s | \), if \( d_{si} \leq B_s \), else 0. (c) Gaussian: \( W_{si} = \exp \left( -(d_{si} / B_s)^2 \right) \) if \( d_{si} \leq B_s \), else 0.
Although in most smoothing applications the neighborhood is defined as a circular shape (where the bandwidth is the radius), it can be extended to accommodate different contexts. For example, the neighborhood can be defined as the area within a travel distance on a network or the first-order (or higher-order) neighbors based on contiguity. Spatial smoothing mostly uses the weighted average of neighboring values (including the observation value to be smoothed), where the weights are determined according to the kernel and the bandwidth. This can be customized for different applications to extract a surface of patterns and features, such as the low-pass and high-pass transformations in image processing.

**Spatial kernel density estimation**

Spatial kernel density estimation (SKDE) is often applied to point event data, such as geocoded crimes and disease incidences, to estimate a density surface of events and help identify the spatial structure of event distribution. Originated in statistics, kernel density estimation (KDE) is a nonparametric way to estimate the probability density function of a random variable based on a finite set of data samples. The simplest approach to density estimation is through a histogram (for 1-D data) or quadrat counts (for 2-D data). To obtain a more robust and smoothed density surface, kernel-based approaches are preferred. Let \( S = \{s_1, \ldots, s_n\} \) be a set of point events, where \( s_i = (x_i, y_i) \) is the location of an event. The kernel density surface \( f(s) \) over the geographic space can be estimated with equation 1:

\[
   f_h(s) = \frac{1}{n} \sum_{i=1}^{n} K_h(|s - s_i|) \quad (1)
\]

where \( f(s) \) is the probability of event occurring at location \( s \), \( h \) is the kernel bandwidth, \( |s - s_i| \) is the distance between the event location and \( s \), and \( K \) is a kernel function that is symmetric, non-negative, and integrates to one. \( K_h \) is called the scaled kernel and defined as \( K_h(|s - s_i|) = 1/h \cdot K(|s - s_i|/h) \). The kernel density estimation constructs a kernel centered at each event and the probability of observing an event at a location \( s \) is the average probability of all events in the data at that location.

Similar to spatial smoothing, a range of kernel functions are commonly used in kernel density estimation, including uniform, triangular, Gaussian, Epanechnikov, and others. While the choice of the kernel function \( K \) is not crucial, the choice of the bandwidth can have a strong influence on the estimated surface since it controls the amount of smoothing effect. A variety of automatic and data-based methods have been developed for selecting the bandwidth (Park and Marron 1990). The bandwidth may also be chosen with domain knowledge and empirical considerations.

**Spatial interpolation**

Spatial interpolation predicts the attribute value at a location based on a limited number of observed data points at other locations, such as estimating the temperature at one place (where there is no temperature measurement) using observed temperature values at nearby weather stations. Spatial interpolation is suitable for exploring the trend surface in geostatistical data where the observations are limited samples of an unknown continuous surface over the geographic space, such as rainfall, chemical concentration, noise level, and so on. Essentially, spatial interpolation methods assume that nearby locations are more likely to have similar observation values than distant locations. One of the common methods for spatial interpolation is inverse distance weighting (IDW), which calculates a weighted average of nearby values as
the estimation for a given point. To estimate the value at location $S_0$, the weight (or contribution) for data at $S_i$ is:

$$w_{i0} = \left(\frac{1}{d_{i0}^k}\right) / \sum_{i=1}^{n} \left(\frac{1}{d_{i0}^k}\right)$$

where $d_{i0}$ is the distance between $S_i$ and $S_0$, $n$ is the number of observed data points, $k$ is power parameter, and

$$\sum_{i=1}^{n} w_{i0} = 1$$

The interpolated value at $S_0$ is

$$x_0 = \sum_{i=1}^{n} w_{i0} x_i$$

A larger value for the power $k$ will introduce a stronger distance-decay effect, meaning that distant locations have much less contribution (weight) for the interpolated value.

While it is theoretically possible to interpolate values at any location, the estimation is more reliable for locations that have observed data points nearby and in different directions. If all observation points are all far away or on one side of the interpolation location, the result will be less accurate. Therefore, the interpolation is usually constrained within the bounding box of the input data. One common method to evaluate interpolation results is cross-validation, which withholds a sampled observation one at a time, performs the interpolation to obtain its estimation, and compares it to the true value. This approach may also help configure the power parameter $k$.

One of the major limitations of IDW is that it does not consider how sample data points are spatially distributed and correlated. Alternatively, geostatistical approaches (known as kriging methods) use variograms to evaluate the structure of spatial autocorrelation in the sample data, construct a formal function to describe the structure, and calculate the weights of observation points for interpolation. Since kriging methods take into account the spatial distribution and autocorrelation structure of observation data, they usually produce better interpolation results than the IDW method. For more information on spatial interpolation, see related entries Interpolation: areal, Interpolation: kriging, and Interpolation: inverse-distance weighting.

### Local spatial statistics

As introduced earlier, the existence (or nonexistence) of spatial autocorrelation by itself can be an important finding when the goal is to identify spatial distribution patterns. A number of spatial statistics have been developed to quantify spatial autocorrelation, that is, the degree to which a random variable at a specific location is dependent on the values at its neighboring locations. Such spatial statistics can be classified into two types: global statistics and local statistics. Global statistics calculate only one statistical value, the whole dataset, which can be tested to indicate whether the data have significant spatial autocorrelation or not. Local statistics focus more on the local variation of spatial autocorrelation and, therefore, calculate a local statistical measure for each location, which can be used to identify pockets of spatial association, such as unusually high concentration of disease cases. Therefore, global spatial statistics are considered confirmatory approaches that only test a single hypothesis while local statistics are more exploratory and search for significant local patterns within a hypothesis space.

### Global statistics

Moran’s I and Geary’s C are the most commonly used statistical indices for global autocorrelation.
EXPLORATORY SPATIAL DATA ANALYSIS

analysis, which calculate one statistical value for the entire dataset to measure how nearby locations are similar or dissimilar for a given attribute. The statistics value is then tested to determine whether the data exhibit significant spatial autocorrelation (or spatial clustering). The Moran’s I statistic is calculated using equation 2:

$$I = \frac{n \sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\left( \sum_i \sum_j w_{ij} \right) \sum_i (x_i - \bar{x})^2}$$

(2)

where $n$ is the total number of locations, $x_i$ and $x_j$ are the attribute values for area $i$ and $j$, $\bar{x}$ is the mean of the attribute values, and $w_{ij}$ is the weight for the pair of areas. The weights $w_{ij}$ are normally set as 1 if $i$ and $j$ are spatial neighbors and 0 otherwise. The neighborhood can be defined as data points within a fixed distance, first-order neighbors, or according to other definitions. Moran’s I values range between $-1$ and 1, where 1 indicates the strongest positive spatial autocorrelation (with high values near high values and low values near low values) and $-1$ represents the strongest negative spatial autocorrelation (with high values near low values). If the observed Moran’s I is close to 0, it indicates spatial randomness, that is, absence of spatial autocorrelation.

Similar to Moran’s I, Geary’s C measures spatial autocorrelation by directly examining the differences in attribute values among observation points:

$$C = \frac{(n - 1) \sum_i \sum_j w_{ij}(x_i - x_j)^2}{2 \left( \sum_i \sum_j w_{ij} \right) \sum_i (x_i - \bar{x})^2}$$

(3)

Geary’s C values range between 0 and 2 with $C = 1$ indicating spatial randomness (i.e., no spatial autocorrelation), $C = 0$ represent the strongest positive spatial autocorrelation, and values above 1 shows increasingly negative spatial autocorrelation.

Local statistics

Local spatial statistics are exploratory approaches that focus on the identification of local patterns of spatial association.

The local $G$ statistics (Getis and Ord 1992; Ord and Getis 1995) measures the degree of association that results from the concentration of high or low values of a given variable. Given an area with $n$ spatial units, where each unit $i$ has a value $x_i$ for a variable, the $G_i(d)$ statistic tests hypotheses about the spatial concentration of the sum of $x_j$ values within the neighborhood of location $i$. The neighborhood is defined as the points within a distance $d$ to the unit $i$. The $G_i$ statistic is calculated as:

$$G_i(d) = \frac{\sum_{j \neq i} w_{ij}(d)x_j}{\sum_{j \neq i} x_j}$$

(4)

The summations over $j$ may or may not include $i$; $w_{ij}$ is one if unit $j$ is within distance $d$ to unit $i$. The $G_i(d)$ statistics detects whether high values or low values cluster in the neighborhood surrounding unit $i$. A high $G_i$ value indicates that high values tend to be near each other and a low $G_i$ value indicates that low values tend to be near each other. Note that the variable must have a natural origin and is positive, and that transformations of the variable values will result in different values for the test statistic. For example, it is inappropriate to use $G_i(d)$ to study residuals from a regression analysis.

Local versions of Moran’s I and Geary’s C, and a more generalized class of local indicators of spatial association (LISA), are developed to measure
the association between the value at \( i \) and values of its nearby areas (Anselin 1995). With the same notation for Moran’s I (equation 5), a local Moran’s I statistic for spatial unit \( i \) is defined as

\[
I_i = (x_i - \bar{x}) \sum_j w_{ij} (x_j - \bar{x})
\]

where the summation over \( j \) only includes values in the neighborhood through the configuration \( w_{ij}, w_{ii} = 0 \). A positive \( I_i \) means positive spatial autocorrelation at location \( i \), while a negative \( I_i \) means either a low value surrounded by high values or a high value surrounded by low values. Local Moran (and local Geary C) cannot distinguish high–high clusters from low–low clusters, unless values across clusters are compared after calculating the indices.

Local spatial clusters, sometimes referred to as hotspots, are those locations that have significant local statistics values. To test the significance of local statistics values, a Monte Carlo permutation approach is commonly used, especially when the theoretical distribution curve of the statistics is hard to obtain. The permutation process holds the value at a location \( i \) fixed and randomly permutes the remaining data values over the locations in the dataset. Each permutation creates a resampled dataset for location \( i \) as a realization of the null hypothesis (i.e., no spatial association at \( i \)). Repeating this process many (e.g., 999) times and calculating the statistics for each repetition creates an empirical distribution curve for the statistics, which provides the basis for significance testing of the actual local statistics value at \( i \). This will be repeated for every location in the dataset to obtain a map of local statistics values and their associated \( p \)-values.

However, when a large number of significance tests is carried out, there is a multiple-testing problem. It refers to the fact that the probability of incorrectly rejecting a null hypothesis is much higher than the confidence interval \( \alpha \) indicates when many tests are performed at the same time. For example, if a test rejects a null hypothesis at a confidence level of 0.05, it means that there is a 5% possibility that the null hypothesis is incorrectly rejected. If 100 such tests (which, for simplicity, are assumed to be independent from each other) are performed at the same time, then it is almost certain that one or more (actually the expected number is 5) null hypotheses will be incorrectly rejected even if there is no significant pattern in the data. In this case, it is difficult to tell whether the rejected hypotheses are merely out of chance or indeed significant. Various approaches for addressing the multiple-testing problem are introduced in the next section.

**Multiple-testing problem and correction**

In the literature of statistics, a number of methods have been proposed to control the multiple-testing problem. Two groups of methods are commonly used: (i) controlling the family-wise error rate (FWER) or (ii) controlling the false-discovery rate (FDR).

The family-wise error rate (FWER) is the probability of erroneously rejecting even one of the true null hypotheses among \( m \) independent tests. If the FWER is to be controlled at some level \( \alpha \), then each of the \( m \) tests has to be carried out at much lower levels. The simplest FWER approach is the Bonferroni method, which simply sets the level for each test at \( \alpha / m \) and is the most conservative correction approach for the multiple testing problem. Holm (1979) proposed a step-down procedure to control FWER. Consider \( m \) hypotheses \( H_1, H_2, \ldots, H_m \), with corresponding \( p \)-values \( P_1, P_2, \ldots, P_m \). Let \( P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)} \) be the ordered \( p \)-values and denote their corresponding hypotheses by \( H_{(1)}, H_{(2)}, \ldots, H_{(m)} \). Holm’s method starts testing from the smallest \( p \)-value \( P_{(1)} \). Let \( k \) be the smallest \( i \) for which \( P_{(i)} > \alpha \/(m - I + 1) \), then reject all \( H_{(i)}, i < k \).
Benjamini and Hochberg (1995) proposed an alternative approach that controls the false discovery rate (FDR), which is the proportion of rejected hypotheses that are incorrectly rejected. Consider testing $H_1, H_2, \ldots, H_m$, based on the corresponding $p$-values $P_1, P_2, \ldots, P_m$. Let $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ be the ordered $p$-values and their corresponding hypotheses be $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$. To control the FDR at a level $\alpha$, and let $k$ be the largest $i$ for which $P_{(i)} \leq \alpha(i/m)$, then reject all $H_{(i)}, i < k$. FDR-based correction is less conservative and may reject more null hypotheses than an FWER-based method does as long as the proportion of incorrect rejections is controlled.

Both of the above correction procedures (FWER or FDR) assume that the hypotheses are independent from each other. When dependence exist among the hypotheses to be tested, which is the case for most local statistics as their neighborhoods overlap and spatial autocorrelation may exist, the above procedures may need further adjustments (Getis and Ord 2000; Benjamini and Yekutieli 2001). To correct the Bonferroni method for the overlapping problem, Getis and Ord (2000) use $\alpha/v$ (instead of $\alpha/m$) as the control level, where $v$ is the estimated number of nonoverlapping local regions.

**Spatial and space–time scan statistics**

Spatial scan statistics is a special type of local spatial statistics that is applicable for point event analysis to detect significant spatial clusters of events, such as hotspots of crimes or unusual concentration of disease cases. Spatial scan statistics uses both computational algorithms and statistical assessments to find areas with a significant excess of observed event points, for example, disease incidents.

There are three major differences between a spatial scan statistics and the local spatial statistics (e.g., Moran’s I and Geary’s C) introduced earlier. First, spatial scan statistics is mainly applied to point event data while local spatial statistics is mostly for spatial lattice data. Point event data can be aggregated to lattice data, for which spatial scan statistics is still applicable. Second, spatial scan statistics compares the total number of events in a neighborhood with its expected count, while local statistics examines the similarity of attribute values within the neighborhood. In other words, spatial scan statistics treats a neighborhood as a new and single unit. Third, spatial scan statistics attempts to enumerate all possible neighborhoods at different scales (i.e., with different bandwidth values), while local spatial statistics primarily focus on one scale (which is determined by the given units and the bandwidth selection). Theoretically, spatial scan statistics avoids the MAUP problem to a large extent and is able to find the appropriate scale for the strongest pattern.

A spatial scan statistic usually consists of the following steps:

1. Define a computational strategy to enumerate a large set of local neighborhoods (areas) that are candidates for spatial clusters. For example, a neighborhood can be defined as the area within a distance to a location. By varying the distance threshold, neighborhoods of different sizes can be created. Theoretically, all possible neighborhoods can be created and tested. However, to achieve computational tractability and statistical test power, circular or eclipse-shaped neighborhoods are most commonly used.

2. Calculate a test statistic for each local area based on the number of events and (optionally) the background population (such as population at risk for a disease). As such, the test statistics takes into account the
neighborhood size so that statistics values of different neighborhoods can be compared.

3 Test the significance level of the test statistic values, one for each neighborhood. For most applications, the theoretical distribution of the test statistics under the null hypothesis is not known and, therefore, a Monte Carlo permutation approach is often used. The multiple-testing problem should be addressed in the testing process.

An earlier example of such spatial scan statistics is the geographic analysis machine (GAM) by Openshaw (Openshaw et al. 1987; Openshaw, Cross, and Charlton 1990), which follows the above general steps and uses a Monte Carlo approach to perform significance test for each local statistics. The GAM defines 2-D grids covering the study area, sets a minimum and a maximum circle radius, and a size increment. For each grid point and a specific radius, compute the test statistic value (which is the count of events in the circle). To obtain a significant level, the test statistic value is then compared with the statistic values from a large number (e.g., 500) of permutation datasets, each of which is generated separately under the null hypothesis. To generate a permutation dataset, the events (e.g., disease cases) in the data are randomly re-assigned to the population at risk. A potential problem with the GAM is that it is difficult to adjust for the multiple-testing problem (Rogerson and Yamada 2009).

The spatial scan statistics developed by (Kulldorff 1997) calculates a likelihood ratio for each neighborhood. Let $p$ be the risk within a neighborhood $Z$ and $q$ the risk outside the neighborhood $Z$ in the study area. The null hypothesis is that $H_0$: $p = q$, and the alternative hypothesis is $H_A$: $p > q$ (or $p < q$). With the Poisson model, the test statistic, likelihood ratio ($\lambda$), for neighborhood $Z$ is defined as:

$$\lambda = \frac{\binom{O_Z}{P_Z} \binom{O_W - O_Z}{P_W - P_Z}^{O_W - O_Z}}{\binom{O_W}{P_W}^{O_W}} I \left( \frac{O_Z}{P_Z} > \frac{O_W - O_Z}{P_W - P_Z} \right)$$

(5)

where $O_Z$ and $P_Z$ denote the count of observations (e.g., disease cases) and the population within $Z$, respectively; $O_W$ and $P_W$ are the counts of observations and population for the whole study area. The likelihood ratio calculation based on the Bernoulli model is also provided in Kulldorff (1997). Different from the GAM, the spatial scan statistic by Kulldorff uses the maximum likelihood ratio of all enumerated neighborhoods as the test statistic. Therefore, the method only reports the most likely cluster, although a set of secondary clusters is also provided (but not declared significant). This can effectively avoid the multiple-testing problem, at the sacrifice of missing additional clusters. To derive the significance level, a Monte Carlo procedure is used to generate a large number (e.g., 999) of permutation datasets under the null hypothesis. For each permutation dataset, the test statistic (i.e., the maximum likelihood ratio) is found over all enumerated neighborhoods. Then the actual test statistic value is compared to the test values of all permutation datasets to derive the significance level for the most likely cluster. A significance level can also be attached to secondary clusters but they will not be declared significant even if their $p$-values are less than the critical level (e.g., 0.05). Later Kulldorff et al. (2006) extended the circular scan statistic by using ellipses as scanning windows; this has more varying parameters, including major axis, minor axis, and angle. The new likelihood ratio also considers a noncompactness penalty. Readers are referred Kulldorff (1997) and Kulldorff et al. (2006) for a detailed methodology explanation.
The spatial scan statistics can be extended to space–time event data with an added temporal dimension. Space–time scan statistics analyzes the incidents using both spatial and temporal information, in which the scan “window” (or neighborhood) is a cylinder with three dimensions (Figure 2). The base of a cylinder represents a 2-D circular window in the geographic space, while the height of the cylinder represents a time range. Similar to space scan statistics, the radius of the base and the height of the cylinder vary systematically, so that the cylindrical windows cover each possible combination of location and time. There are two types of space–time scan statistics: prospective space–time scan statistics and space–time permutation scan statistics. The prospective analyses only evaluate the cylinders that the ending time is “current,” that is, each cylinder (potential cluster) is still active (Kulldorff 2001). Prospective space–time statistics can be used to monitor and detect disease outbreaks early by comparing current pattern with historical trends. Alternatively, space–time permutation statistics primarily examines the space–time interaction without using population at risk, which is often unavailable, such as the population at risk for patients visiting a hospital with flu symptom (Kulldorff et al. 2005).

### Geographically weighted regression

Different from the above methods that detect univariate or point event clusters in space, regression analysis is commonly used to detect multivariate relationships among spatially referenced variables. However, there are a number of assumptions underlying the basic regression model that spatial data may not satisfy. One is that the observations should be independent of one another, which is often not true with spatial data due to the existence of spatial autocorrelation. This means that the regression model’s residuals might exhibit spatial dependence and lead to biased parameter estimates, which are either too high or too low as an estimate of the unknown true value. To address this issue, a number of spatial regression models have been developed, such as the spatial error model and the spatial lag model, which incorporate dependent variable values or residual values in each location’s neighborhood to explicitly model and use spatial autocorrelation for prediction. These models are considered model-based and confirmatory methods, and thus are not covered here. Details on spatial regression models are given in Cressie (1993) and other related papers.

Spatial heterogeneity is another challenge for regression analysis of spatial data. Generally, a single regression model is configured with all input data, which thus assumes that the relationship being modeled is the same everywhere within the study area. However, this is often not the case with spatial data, as the relationship might vary across space, which is referred to as spatial heterogeneity or spatial nonstationarity, as explained earlier. An early example of a regression model that attempts to deal with spatial
heterogeneity is the spatial expansion method (Casetti 1972), in which the model parameters are functions of location. Geographically weighted regression (GWR) is a more recent method for addressing spatial nonstationarity in regression analysis. GWR extends traditional regression analysis by allowing local rather than global parameters to be estimated (Brunsdon, Fotheringham, and Charlton 1996; Fotheringham, Brunsdon, and Charlton 2002). For each point in the geographic space, a regression analysis is performed on the data in the neighborhood around that point, where the neighborhood (or bandwidth) can be defined in several different ways, similar to the neighborhood model used previously in spatial smoothing and kernel density estimation methods. In estimating the regression model at a location \( u \), the observations that are closer to \( u \) will have a greater weight in the estimation than observations further away from \( u \). The equation for GWR is:

\[
y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} + \varepsilon_i
\]

where \( y_i \) is the prediction for the dependent variable at location \( (u_i, v_i) \) and \( \beta_k(u_i, v_i) \) is the parameter for the \( k \)th independent variable at the location. The bandwidth in the kernel determines the number of observations and their weights being used in estimating the regression parameters.

As an exploratory spatial analysis method, GWR outputs a surface of estimates for each parameter and diagnostic (e.g., standard error and residual), which can help understand how the parameters change over space. With the kernel-based weighting and smoothing, nearby locations will have similar parameter estimates. As with other kernel-based approaches, the choice of kernel does not substantially affect the result while the selection of bandwidth can affect the outcome significantly. Therefore, it is important to choose an “optimal” bandwidth. When the spatial units are not evenly distributed in space, an adaptive bandwidth should be used, where each neighborhood will have the same number of spatial units and, thus, the same number of sample points. To find an optimal value for the bandwidth (i.e., number of sample points), GWR uses the corrected akaike information criterion (AICc), which compares different models considering the estimate of the standard deviation of the residuals and a penalty for the model complexity. The bandwidth with the lowest AICc is used in the estimation of the model parameters. The user may also choose a different bandwidth based on practical reasons or prior knowledge.

GWR provides an effective approach for examining spatially varying regression models and explores how independent variables affect the dependent variable in different ways from place to place. In the application of GWR, there are a number of aspects that need careful attention. First, to test the significance of parameter estimates in each local regression model, the multiple-testing problem should be taken care of according to the discussion in the section on Local statistics. Second, while examining the surface of parameter estimation is useful, it is also important to distinguish whether such variation in space is significant or not, that is, does GWR make a significant improvement in prediction over the global regression model for the specific analysis task? The AICc measure can be used for this purpose (Fotheringham, Brunsdon, and Charlton 2002). Third, GWR addresses the spatial nonstationarity problem by constructing local regression models. However, each local model may still suffer the spatial autocorrelation problem, that is, observation data are not independent from each other. Fourth, GWR allows parameter estimates of a regression model change over space while assuming that the model form does not change, that is, a linear
relationship for every neighborhood. Practically this is reasonable as nonlinear relationships are more difficult to model within a local neighborhood. Variable transformations, such as a logarithmic transformation, may be applied before the application of GWR to better model nonlinear relationship if there is prior knowledge of the relationship form.

Clustering, spatial clustering, and regionalization

General-purpose clustering

The previously introduced methods of local spatial statistics and scan statistics are to detect spatial areas (clusters) with excessive concentrations of point events or high (or low) attribute values, where the clusters only represent a portion of the input data. This section introduces another type of clustering analysis, which is to organize or partition the input data into a number of clusters based on the similarity among data items. Data items in the same group (cluster) are more similar to each other than those belong to different groups. The clusters together cover all data items, that is, every item in the data belongs to a cluster. Clustering analysis is also called unsupervised learning or numerical taxonomy, which is to discover the inherent structure (grouping) of observed items and group them accordingly to extract high-level labels and descriptions. Clustering has been studied extensively in many disciplines, such as statistics, computer science, biology, psychiatry, psychology, archaeology, geology, geography, and marketing. General clustering methods can be broadly classified into two groups: hierarchical clustering and partitioning clustering. Figure 3 shows an overview of different types of clustering methods, some of which are briefly introduced below.

Hierarchical clustering organizes observations into a hierarchy with a sequence of nested partitions or groupings. Commonly used hierarchical clustering methods include the Ward’s method, single-linkage clustering, average-linkage clustering, complete-linkage clustering, and model-based clustering (Fraley 1998; Jain, Murty, and Flynn 1999). Hierarchical clustering results are normally visualized with dendrograms, which are useful to help users examine patterns (clusters) at different scales and determine how many clusters should be obtained. To build the cluster hierarchy, a method often follows one of two general strategies: (i) a bottom-up agglomerative process – each observation starts in its own cluster and the most similar pair of clusters is merged as the hierarchy moves up; or (ii) a top-down divisive approach – all observations start in one cluster and splits are performed recursively when moving down the hierarchy. During the merge (or division) process, an objective function or a heuristic can be used to determine the best merge (or split). Figure 4 shows the clustering results of two hierarchical clustering methods: single linkage and complete linkage. The two methods are both agglomerative methods and only differ in their definition of “distance” between two clusters. The single-linkage method uses the distance between the nearest pair of points that connect two clusters as the distance between the two clusters, while the complete linkage method uses the distance between the farthest pair of points. This difference leads to the different results shown in Figure 4.

Partitioning clustering methods divide the input set of observations into a number of nonoverlapping clusters. A data item is assigned to the “closest” cluster based on a proximity or other dissimilarity measure. Commonly used partitioning clustering methods include the k-means and the expectation maximization (EM) methods. For example, k-means clustering
Partitioning clustering
- Distance-based methods (e.g., k-means)
- Model-based methods (e.g., expectation maximization)
- Density-based methods (e.g., DBSCAN)
- Neighborhood-based methods (e.g., OPTICS)
- Grid-based methods (e.g., CLIQUE)

Hierarchical clustering
- Distance-based methods (e.g., single linkage, average linkage, and complete linkage)
- Model-based methods (e.g., Fraley 1998)
- Density-based methods (e.g., expectation maximization)
- Neighborhood-based methods (e.g., OPTICS)
- Grid-based methods (e.g., grid clustering)

**Figure 3** A classification of clustering methods. Adapted from GeoInformatica, ICEAGE: Interactive Clustering and Exploration of Large and High-dimensional Geodata, 7(3), 2003, 229–253, Guo, D., D. Peuquet and M. Gahegan © With permission of Springer.

**Figure 4** Hierarchical clustering results of the single-linkage method (left) and the complete-linkage method (right), using the same 1-D dataset of 10 data points.

is an optimization problem that tries to find \( k \) cluster centers in the data and assign each data item to the nearest cluster center such that the total squared distances from data points to their corresponding cluster center are minimized. However, to be able to process large data in practice, the k-means algorithm uses a heuristic-based and iterative procedure to find a local optimum. Most partitioning clustering methods require that the number of clusters \( k \) be given, which is one of the disadvantage of these methods. Self-organizing map is a special partitioning clustering method, which not only segments data into clusters but also orders the clusters in a 2-D layout so that nearby clusters are similar to each other, which partially avoids the need for specifying an exact number of clusters beforehand. A comprehensive review of different clustering methods can be found in Jain, Murty, and Flynn (1999).

**Spatial clustering and regionalization**

There are three main strategies to incorporate spatial information in a spatial clustering process.
The first strategy derives clusters based on multivariate data with nonspatial clustering methods and then maps the clusters in the geographic space to understand the spatial meaning and distribution of clusters. This strategy is straightforward and requires minimum modification to the clustering method. However, it cannot capture spatial patterns in the clustering process and can only rely on user interpretation to link clusters to their spatial context.

The second strategy is to directly use spatial coordinates or spatial distance definition in the clustering process, which may solely be based on geographic coordinates (e.g., clustering of spatial point events) or a combination of spatial and nonspatial information (e.g., clustering spatial objects with multivariate information and spatial locations). This type of approach represents the majority of spatial clustering applications, as it can incorporate various context-dependent geographic metrics, such as network distance, travel time, barrier and impedance factors, into the clustering process. The challenge with these approaches, particularly when spatial and nonspatial information are both considered, is how to determine an appropriate way to weigh and combine spatial similarity with nonspatial (attribute) similarity, which subsequently may cause difficulties in interpreting the clusters. Similar to general-purpose clustering methods summarized in Figure 3, spatial clustering methods can also be partitioning, hierarchical, density-based, or grid-based. Readers are referred to Han, Kamber, and Tung (2001) for a comprehensive review of various spatial clustering methods.

The third type of approach imposes geographic constraints (such as geographic contiguity) in the clustering process. Regionalization is a special form of spatial clustering that seeks to group spatial objects into spatially contiguous clusters (based on a contiguity definition) while optimizing an objective function. Regionalization has been widely used in many application problems, such as the delineation of climatic regions, defining eco-regions, and map generalization. Existing regionalization methods that are based on a clustering concept can be classified into three groups: (i) multivariate (nonspatial) clustering followed by spatial processing to rearrange clusters into regions; (ii) clustering with a spatially-weighted dissimilarity measure, which can consider spatial distance as a factor in forming clusters but cannot guarantee that each cluster is contiguous in space; and (iii) contiguity-constrained clustering that enforces spatial contiguity during the clustering process. Recent reviews of different regionalization can be found in Duque, Ramos, and Surinach (2007) and Guo (2008). Regionalization methods are often extensions of a general-purpose clustering method by explicitly enforcing a spatial contiguity constraint in the process. Therefore, regionalization methods can be grouped into partitioning and hierarchical methods as well, which also allows the use of multiple variables to define cluster similarities and the incorporation of different contiguity definitions (such as contiguity on a road network).

Challenges in clustering analysis

As evident in the introduction above, there are a large number of clustering algorithms in the literature. Actually, an existing clustering method can easily be customized or a new clustering method designed by modifying one or more components in the clustering process, such as data representation, distance measure, and clustering procedure. As such, it becomes a major challenge for the user to choose (or design) an appropriate clustering method suitable for the problem at hand. Each method can be unique in a number of aspects:
the input data type, which may be numerical, categorical, network, spatial, image, or a mix of different types;
• the definition of “cluster,” which can be hierarchical, density-based, connection-based, distance-based, distribution-based, or a combination of them;
• the definition of distance (i.e., dissimilarity), which should effectively enable the construction of clusters;
• the strategy to group or divide data into clusters based on the above definitions;
• the number of parameters in controlling the above definitions and clustering process.

Different methods may produce dramatically different results. Even for the same method, different data inputs, data transformation or parameter setting can lead to quite different clusters. Therefore, it is critical in clustering analysis to carefully address the following questions:

• What data attributes should be included and how should the data be normalized and weighted?
• What is considered a cluster and what similarity (distance) measure is appropriate for the given situation?
• How to construct clusters given the data and definitions? How to balance efficiency in processing the data and the accuracy in finding clusters?
• How should domain knowledge be integrated in a particular clustering problem?
• How many clusters should be obtained and how to validate the clusters?
• Application-specific requirements, constraints, interpretation, and validation.

Given a data input, all clustering algorithms will produce clusters regardless of whether the data contain inherent clusters or not. Therefore, it is important to assess the output clusters, which may take one of two types of validation approaches: internal evaluation and external evaluation. Internal evaluation examines the clustering result based on the input data, which assign the best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters. Commonly used measures or indices for internal evaluation include the Davies–Bouldin index, Dunn index, and the silhouette coefficient. On the other hand, external evaluation assesses the clustering result based on external benchmark datasets with known cluster labels, which measure how close the clustering outcome is to the predetermined benchmark classes. This type of evaluation, however, should consider the fact that the performance not only depends on the clustering algorithm but also on the input data. If the input data do not contain the needed attributes to distinguish the clusters, then, no matter which method is used, the clustering result will not be very useful. Commonly used external evaluation measures or indices include the Rand measure, F-measure, Jaccard index, and the Fowlkes–Mallows index.

Geovisualization and visual analytics

Geovisualization concerns the development of theories and methods to facilitate knowledge construction through visual exploration and analysis of geospatial data, and the implementation of visual tools for subsequent knowledge retrieval, synthesis, communication, and use (MacEachren and Kraak 2001). Mapping is essential in geographic data visualization, which is often coupled with other information graphics, such as tables, histograms, scatter plots, charts, and more sophisticated multidimensional visualization techniques. The main difference between
traditional cartography and geovisualization is that the former focuses more on the design and use of maps for the communication of known information while geovisualization emphasizes the discovery of unknown information with highly interactive maps and computational algorithms for data exploration, hypothesis generation, and knowledge construction. For example, there are many geovisualization methods for multidimensional spatial data analysis, for example, scatterplot matrices, pixel-oriented approaches, parallel coordinates plots (PCP), and the dynamic linking between one or more of these nonspatial multivariate representations and geographic mapping. With linked representations (i.e., multiple views), users can construct a complex and comprehensive understanding of multivariate, spatial, and spatiotemporal patterns.

To cope with today’s big and complex spatial datasets, geovisualization involves active research to address several major challenges: (i) processing big datasets efficiently; (ii) handling multiple perspectives and many variables simultaneously to discover complex patterns; and (iii) designing effective user interface and interactive strategies to facilitate a user-centered discovery process. Due to limited visual space (e.g., monitor screens), large data volume can cause cluttered visual displays (e.g., points overlapping in a scatter plot) and make it difficult (if possible at all) for the analyst to visually perceive patterns. Large amounts of data also require a considerable amount of time to process, which is a huge challenge for interactive analysis. One solution is to rely on the user to dynamically filter, select, zoom, and adjust detail levels in the visualization, meaning that only a small subset of data is visualized at one time. Another type of solution is to combine efficient computational methods, such as clustering, classification, and association rule mining, with geovisualization, with the former finding patterns quickly and the latter helping users explore and understand the patterns.

To visualize multiple perspectives, such as space, time, and multivariate dimensions, it is often necessary to couple multiple visualization components (one for each perspective), integrate with dimension reduction techniques (to help compress high-dimensional data to a low-dimensional representation), and allow user interactions to seek patterns across perspectives. Different visual components are chosen according to the specific type of data or perspective to be visualized. For example, a map is needed for presenting spatial distribution and location patterns; a trend plot is needed for temporal patterns; and multidimensional visualization techniques are needed for understanding multivariate patterns. These components have to be linked to allow a holistic understanding of patterns across different perspectives; this is often referred to as multiple linked views. The “linking” among different views can be achieved through two main strategies: linking through visual identifiers such as color (e.g., the same object is assigned the same color in different views) and linking through user interaction (e.g., when a user selects a data object in one view, the same object will be highlighted in all other views). Each strategy has its advantages and limitations. Visual linking allows a quick overview of holistic patterns without user interaction but it relies on a small set of a clearly defined identifier (such as different colors or labels), thus is unable to simultaneously support linking too many patterns. Therefore, dimension reduction and computational grouping are often needed to extract major patterns and skip details, using methods such as multidimensional scaling, principal components analysis (PCA), self-organizing maps, or clustering methods. Interactive linking, on the other hand, can allow more details to be presented upon request but
cannot support a quick overview. In practice, these two strategies are often used together, following the well-known visual information seeking mantra: “overview first, zoom and filter, then details on demand” (Shneiderman 1996).

Research efforts for the third challenge have emerged as an active field called visual analytics, which is “the science of analytical reasoning facilitated by interactive visual interfaces” (Thomas and Cook 2005). Visual analytics is an interdisciplinary field that involves computer science, information visualization, cognitive and perceptual sciences, interactive design, graphic design, and different application domains such as social sciences. The design of the visual analytics tools and techniques is based on cognitive and perceptual principles, with a goal of supporting analytical reasoning and engaging human judgments in analyzing data and addressing complex tasks. Related to, but different from, scientific visualization and information visualization, visual analytics is especially concerned with coupling interactive visual representations with underlying analytical processes to address high-level and complex tasks, such as sense making, reasoning, and decision-making. Geo-visual analytics specifically focus on geographical data and problems, with strong involvement of methodologies and research advancement in the geographic information and spatial sciences, such as geovisualization, geospatial semantics and knowledge management, geocomputation, spatial data mining, and various spatial domains such as emergency management, public health, and citizen participatory decision-making.

Geovisualization methodologies may also be highly specialized for different data types, such as multivariate spatial lattice data (such as climate data), space–time event data (such as crimes, traffic accidents and jams), spatial interaction data (i.e., origin–destination data such as migration), trajectory data (such as human daily activities, vehicle trajectories), and georeferenced documents and social media data (such as web documents and tweets). The methodologies may also differ for different analysis tasks, such as real-time monitoring, spatiotemporal pattern discovery, participatory decision-making, or emergency response support. Two examples of geovisualization methodology are briefly introduced here; one is multivariate mapping for climate data analysis, the other is flow mapping for migration data analysis.

### Multivariate mapping

Methods for multivariate mapping can be classified into three main types. The first type, such as composite glyphs, depicts each variable with a certain visual attribute (such as color, shape, size, or orientation) and then integrates all variable depictions into one symbol to plot on a map. For example, the Chernoff face approach maps multivariate data by relating variables to different facial features (such as mouth, eye, and eyebrow) to create a face icon for each data object and then draws each face icon on a map at the location of the data object. However, this type of approach can only work with small datasets, as too many icons on a map will make the map cluttered and ineffective to allow pattern discovery. The second type uses multiple linked views (including maps) that show one (or more) variables per view and different views. The third type summarizes the multivariate data to a lower-dimensional space through clustering analysis and then visualizes the clusters with maps, plots, and other visual forms.

Figure 5 shows an example of the SOMVIS multivariate mapping (Guo et al. 2005, 2006), which couples a self-organizing map, multidimensional visualization (a parallel coordinate plot), and a map component. The method is a combination of the second and third types introduced above. Using the the monthly mean
surface air temperature for 60 years (January 1948–December 2007), an anomaly value, that is, the difference between the 40-year average (1948–1987) and the latest 10-year average (1998–2007), is calculated for each month and each grid cell in order to analyze temperature change patterns globally. The self-organizing map groups all grid cells into 49 clusters based on their twelve anomaly values (one for each month). The clusters are then colored with a systematic coloring scheme to support visual linking among multiple views. The parallel coordinate plot and the map visualize the multivariate meaning and spatial distribution of these clusters, respectively. For example, the red/reddish clusters, primarily located in the Arctic area, have high positive values for winter months (November, December, January, and February) and around zero (i.e., no change) for summer months (June, July, and August). This means that during the past decade, the Arctic area was much warmer than before in winter.
but relatively stable during summer. In contrast, the green/light-green clusters, mainly located in the Antarctic area, have high positive values for its winter months (April–September) and negative values for its summer months (January, February, November, and December). This means that during the past decade, the Antarctic area was much warmer in winter but cooler during summer. These patterns involve space (i.e., place-to-place differences in climate change patterns) and time (which is treated as multivariate information, i.e., 12 months as 12 variables), which cannot be easily discovered without the combination of clustering, multidimensional visualization, and mapping. The three views also support interactive linking; a grid cell or a cluster can be selected in any view and the selection will be highlighted in other views.

### Flow mapping

Flow mapping has long been used in a wide range of applications, such as human migration, transportation, commodity flow, and commuting (Tobler 1981, 1987). Two of the major challenges for flow mapping are: the visual cluttering problem – maps become illegible with too many flows plotted on top of each other; and the modifiable area unit (MAUP) and unit size problem – flow volumes are highly correlated with unit sizes, for example, population.

To address the visual cluttering problem, a number of approaches have been proposed based on: location aggregation, surface generation, or edge rerouting. A review of aggregation methods for movement data can be found in Andrienko and Andrienko (2002). However, aggregation will inevitably cause a significant loss of information, skip flow patterns at local scales, and suffer from the modifiable areal unit problem. Surface generation approaches produce a vector flow surface that only maps flows between geographically adjacent places (Tobler 1987), where a long-range flow is decomposed to a sequence of short flows. The limitation is that the origin and destination information of each particular flow is lost. The third group of approaches focuses on minimizing edge crossing (and thus reduce cluttering) in flow maps through edge rerouting (Phan et al. 2005) or edge bundling (Cui et al. 2008; Holten and van Wijk 2009; Verbeek, Buchin, and Speckmann 2011), which reroute or bundle edges to improve the visual clarity of flow maps. These methods are effective in producing aesthetic representation of flow data, especially for small datasets. On the other hand, the main limitation of this type of approach is that bundled edges make it difficult to perceive the actual connection between specific pairs of origin and destination.

There is also a variety of methods for flow visualization based on nonspatial views, such as ordered matrices, combinations of maps and matrices, interactive OD maps, and exploratory visualization. Normally, interactive visualization systems do not intend to summarize the entire dataset in a single flow map. Instead, they provide a nonspatial view (such as a matrix) and rely on user interactions to select data to map and explore flow patterns through an iterative process. These nonspatial approaches avoid, to a certain degree, the visual cluttering problem but they cannot provide a clear overview of spatial flow patterns. There are also methods for summarizing flow properties for each location using graph measures such as net migration ratio, centrality, and flow density (the number of flows passing a pixel) (Rae 2009). The kernel-based smoothing and density estimation methods introduced earlier can be extended or customized for mapping locational characteristics of flows (but not the actual flows) (Scheepens et al. 2012).
Different from the above location-based summary measures, a flow density estimation and flow mapping method has recently been developed to map flow patterns; it can address both the cluttering problem and the modifiable unit problem (Guo and Zhu 2014). This approach first estimates a normalized flow volume (or density) between each pair of origin and destination; then selects major representative flows to filter out less-important details and produce a flow map of overall flow patterns. For example, the United States county-to-county migration data collected by the Census Bureau contains migration flows among over 3000 counties, with more than a million flow paths. The flow volumes are not comparable to each other since county populations vary dramatically. Moreover, when focusing a specific age group (e.g., senior migrants of age 65–69), the flow from one county to another county can be very low or even none. Therefore, it is necessary to go beyond the default geographic units and examine flows between meaningful and comparable areas. Figure 6 shows the smoothed migration map for senior migrants of age 65–69; it shows that senior migrants in the east had a strong preference for Florida (particularly West and South Florida) while those in the west were more likely to move to Arizona or its surrounding area. For the metropolitan areas in the north (such as Minneapolis, Detroit, and Boston) there is also a senior migration trend towards the areas further

Figure 6 Flow map of US county-to-county migration for migrants of age 65–69. Guo and Zhu (2014). Adapted by permission of IEEE.
Spatial data mining and geographic knowledge discovery

Nowadays the availability of big and high-resolution spatial and spatiotemporal data provides opportunities for gaining new knowledge and better understanding of complex geographic phenomena and addressing urgent real-world problems, such as human–environment interaction, geosocial dynamics, global climate change, and pandemic flu spread. However, most traditional spatial analysis methods were developed in an era when data were relatively scarce and computational power was not as powerful as it is today (Miller and Han 2009). Existing methods often focus on a limited perspective or a specific type of model. With newly emerged data types, such as mobility, trajectories, social media, and volunteered geographic information on the web, there is a strong need for new approaches to analyze big and heterogeneous data sources and discover complex structures and new types of patterns (Guo and Mennis 2009).

To address these challenges, spatial data mining and geographic knowledge discovery has emerged as an active research field, focusing on the development of theory, methodology, and practice for the extraction of useful information and knowledge from big and complex spatial databases. Spatial data mining has deep roots in traditional spatial analysis fields (such as spatial statistics, analytical cartography, and exploratory spatial data analysis) and various data mining areas in statistics and computer science, such as clustering, classification, association rule mining, information visualization, and visual analytics. As an expansion of traditional exploratory spatial data analysis, spatial data mining and geographic knowledge discovery has several new perspectives. It tends to focus on much larger and more complex datasets and research problems. Data mining and knowledge discovery emphasize an iterative process that involves multiple steps, including data preprocessing, incorporation of prior knowledge, computational analysis, visual exploration, formulation of hypotheses, evaluation of results, and feedback on next iteration (Fayyad, Piatetsky-Shapiro, and Smyth 1996). It is exploratory in nature but more inductive and data-driven than traditional methods. For more information, readers are referred to the entry on Geographic data mining.

SEE ALSO: Artificial neural networks in geospatial analysis; Density estimation; Geographic data mining; Geostatistics; Geovisualization of social media; Interpolation: areal; Interpolation: inverse-distance weighting; Local statistics and place-based analysis; Point pattern analysis; Spatial analysis; Spatiotemporal analysis; Trajectories: analysis; Visualization; Visualizing uncertainty

References


EXPLORATORY SPATIAL DATA ANALYSIS


Export processing zones

Dennis Arnold
University of Amsterdam, Netherlands

In the past half-century major transformations have occurred in the global economy as production networks originating in “core” capitalist economies have expanded their sourcing networks into new frontier regions of production across Asia, Latin America, Eastern Europe, and Africa. Foreign direct investment (FDI) has been the most important vehicle for bringing goods and services to foreign markets and for integrating national and subregional production systems. Assembly-based export processing zones (EPZ) and more recently special economic zones (SEZ) have remained a central means to attract and contain FDI in developing and emerging market economies. While the global trade and investment context from which assembly-based EPZs emerged has changed significantly in recent years, so too has zone development and administration.

SEZs are defined as geographically delimited areas, frequently physically secured, that are usually, but not always, outside the customs territory of the host country. SEZs are generally implemented to meet fiscal, social, and infrastructure policy rationales. The most important fiscal goal of an SEZ is to facilitate economic growth through the use of reduced tariffs and more efficient customs controls. They are also tools for companies seeking to cut costs and improve inventory efficiency, and they often address developing nations’ inefficient trade policies and physical infrastructure. They range in size from single factories to large cities. SEZs are under single management, either government or private sector. Businesses located within SEZs are normally eligible for benefits such as duty and tax exemptions on goods based on the fact that they are physically located within the zone. SEZ has become the generic term encompassing a broad range of economic zones, including free ports, free zones, foreign trade zones, industrial processing zones, science and industrial parks, logistics hubs, and EPZs, the focus of this entry. Many countries employ their own variations of these special enclaves, and use their own terminology. For example, Mexico refers to its SEZs as *maquiladoras*, Ghana, Cameroon, and Jordan have industrial free zones, and the Philippines calls its economic zones special export processing zones.

EPZ and SEZ definitions are, for the most part, synonymous. EPZs specifically focus on manufactured goods that are largely aimed for export markets. EPZ and assembly-based SEZ objectives are to create new jobs, boost growth in trade, exports, and foreign exchange earnings, facilitate economic diversification and industrialization, and provide access to foreign technology and management expertise. They provide a functional advantage to investors seeking to capitalize on the economies of scale that a geographic concentration of production and manufacturing may bring to a trade region. They compete primarily on the basis of fiscal and tax incentives, differential labor regulation mechanisms such as reducing or altogether banning trade union activity, relaxed environmental regulations, and/or proximity to major markets. For the latest generation of zones, EPZ is increasingly an anachronism, as most countries
So label manufacturing and assembly zones as SEZs, an obvious association with China’s zones.

The experience of SEZs and the benefits that they bring is a subject of much debate. While these zones can create jobs, promote investment, and catalyze regional development, they can also be associated with labor abuse, limited social and economic upgrading, low levels of investment, and excessive establishment costs (Nel and Rogerson 2013). Labor activists campaigned against restrictions on freedom of association, collective bargaining, and other basic rights in SEZs. Dispossessed farmers and peasants have mounted large-scale protests over state practices in acquiring land for the zones. The world’s most successful zones have been established in East Asia’s newly industrializing countries, and their achievements have not been uniformly replicated in the developing world’s less prosperous environs, particularly countries with small markets. SEZs tend to benefit FDI and larger domestic investors most in the short term; they are not a direct solution for the development of local small and medium enterprise development. While certain zones have demonstrated more positive long-run economic impacts, many neoclassical and neoliberal thinkers and institutions criticized zones for hindering economy-wide liberalization (Farole 2011; FIAS 2008). By the late 1990s it appeared that SEZs would be relegated to a footnote of history.

Despite criticisms, the number of zones has continued to expand over the past 40 years. From 1975 to 2006 the number of SEZs increased from 79 in 25 countries, to approximately 3500 zones in 135 countries, clustered mainly in Asia and the Pacific and the Americas (FIAS 2008). Globally, zones account for approximately $200 billion in gross exports per annum and directly employ some 66 million workers (FIAS 2008). EPZs and assembly-oriented SEZs remain central to many late industrializing countries’ development models (Figure 1).

**Historical evolution of zones**

Free zones have existed for centuries. They were originally established to encourage entrepôt trade, and mostly took the form of citywide zones located on international trade routes. Examples include Gibraltar (1704), Singapore (1819), Hong Kong (1848), Hamburg (1888), and Copenhagen (1891). By 1900, 11 free trade zones (FTZs) existed globally; of these, seven were in Europe and four in Asia (Farole 2011). Manufacturing entered the realm of free trade zone activities in the twentieth century. The Spanish FTZs established in the 1920s were among the first to accommodate industrial production, hosting one of the first Ford Motors plants in Europe. This and other examples of early manufacturing programs in FTZs were of limited scope and the focus remained very much on trade. The next major “innovation” in zone administration and policy was the introduction of export-orientation in developing countries. The 1948 Operation Bootstrap in the US Commonwealth of Puerto Rico was a key moment in changing SEZ policy and orientation. The program offered low cost labor to attract US firms to set up manufacturing operations to serve the mainland US market.

From the mid-1960s on, economic development strategies in industrializing countries took one of two forms (with combinations beyond the scope of this entry): import substitution and export-oriented industrialization. EPZs acted as globalization catalysts. The modern variation of the manufacturing-oriented zone is traced back to Shannon, Ireland, where an EPZ was set up in 1958. The Shannon approach was original because it combined the attributes of
the free trade zone with those of the industrial park into a single, integrated investment, industry, and trade development instrument (Farole 2011). It was followed by EPZs in Kaohsiung, Taiwan, and Kandla, India, established in 1965. These early EPZ policies were replicated in numerous developing countries, perhaps not coincidentally typically in authoritarian states, in the subsequent decades. South Korea opened its first zone, Masan, in 1971. Indonesia, Malaysia, the Philippines, Thailand, Singapore, and Sri Lanka all developed zones in this period as they shifted from import substitution to export-led development.

Mexico’s *maquiladora* program was another early and prominent assembly zone. Its overarching goal was to compensate for the termination of the US Bracero Program in 1964, a program initiated in 1942 which governed the flow of contract laborers from Mexico to the United States. Its termination left Mexico with a severe shortage of employment opportunities. The new program made surplus labor available to American-owned companies. The *maquiladora* scheme entailed legislation to allow and encourage American manufacturing firms to invest and operate along the border, marking a key moment in the liberalization of Mexico’s economy.

China’s remarkable economic growth over the last 30 years has been partially attributed to its SEZ model – in fact China is the first country to use the term. China’s approach to SEZs has been the most influential application of the policy, and it is not surprising that many developing countries are anxious to emulate at least part of China’s economic success, by adapting SEZs to their countries (Nel and Rogerson 2013). China’s initial SEZs were political experiments in response to stagnant economic growth; the first four zones were initiated in 1980, all near large urban areas pursuing similar objectives, namely to facilitate broad-based comprehensive

---

**Figure 1** EPZs and similar economic zones. Source: Boyenge (2007).
development (Yeung, Lee, and Kee 2009). This evolved into the “open door" policy leading to the establishment of SEZs in 14 coastal cities. Today, China has more than 200 zones of various types, sizes, and sectoral concentrations.

China’s SEZs afforded entry into the world markets, initially in labor intensive light manufacturing industries, and more recently in higher value-added manufacturing and services. For foreign investors, the size, location, flexible labor laws, and stable politics have been among the key factors driving investments. During the 1980s, SEZs drew significant investment from neighboring Hong Kong and Taiwan, but now that investment opportunities are made available throughout China, the current value of SEZs is questionable. Admission to the World Trade Organization (WTO) has further reduced the special attributes of many of the zone policies, as the WTO required that similar policies must apply to the whole country in terms of trade liberalization and foreign investment (Nel and Rogerson 2013). As a result, China has developed a range of new policies to stimulate economic growth. Thus, SEZs were not a “stand-alone” policy option fixed in time and space, rather in China SEZ policy was an element of a much broader policy package of economic reform, economic development, and global engagement relative to the needs of the Chinese economy at that time (Nel and Rogerson 2013). In addition, there were clear efforts to spread the economic benefits of SEZs outside of the privileged zones such that their operational evolution reflected global economic and trade policy shifts (Yeung et al. 2009).

In sum, the global economic context from which assembly-based EPZs emerged in Ireland and transitioned to China’s massive-scale SEZs has changed significantly. From the mid-1960s through the 1990s zones were central components of export-led development and were key to the impressive growth in East and Southeast Asia, and to a lesser extent Mexico. Many have argued that SEZs were losing relevance in the neoliberal era, as trade and investment barriers were disassembled through the WTO and other mechanisms. Meanwhile, global production networks have expanded and export-led development has become global orthodoxy. Global replication of zones has diminished their “special” characteristics. However, the number of SEZs has continued to increase dramatically. Many countries have implemented and replicated assembly-oriented SEZs in the hope of emulating China’s success, catalyzing a global “zone fever.” The following section reviews contemporary zones – their rationale and the challenges they present for labor rights, women, the environment, and land use.

Recent trends and debates

Late industrializing countries, those most likely to initiate the EPZ or assembly-oriented SEZ over the past decade, face an altogether different competitive environment from that faced by the earlier zones. The sheer number of states competing for investment is increasing, while almost all are offering similar incentives, limiting the capacity of all zones to attract FDI. For example, the average number of firms per zone is 35 in Africa compared to 300 in Asia and Latin America (Nel and Rogerson 2013). More specifically, the entrenchment of “factory Asia” leads to higher entry barriers to global production networks (GPN) in textiles and garments and electronics for other world regions; the expiry of the Multi Fibre Arrangement in 2005 has led to greater concentration of textile and garment sourcing, heavily favoring Asian-based producers; and recently slowing demand in traditional export markets has only heightened
global export competition – in which speed to market and production quality play fundamental roles in sourcing decisions, along with labor and other factor costs.

In certain countries and regions the speed with which SEZs have been approved is staggering. For example, the Indian government formally approved 462 SEZs between the enactment of the SEZ Act in 2005 and May 2008 (Banerjee-Guha 2009). The number of SEZs has mushroomed across Africa and zones now exist in over 20 countries (Nel and Rogerson 2013). In Continental Southeast Asia 49 SEZs were approved between 2005 and 2010, with 73 expected by 2015 (Arnold 2012). Recent World Bank research argues that contemporary SEZs have the potential to act as catalytic exclaves that both announce and prepare for liberalization under certain circumstances and provided certain prerequisites have been addressed (Farole 2011). If not, SEZs may, at best, provide limited economic benefits for a limited period and, at worst, may turn into welfare-reducing enclaves that restrict countrywide liberalization. In short, most zone programs continue to be designed as instruments for trade and investment; they continue to be built around low labor costs, trade preferences, and fiscal incentives.

Numerous SEZs have cost the host country more to build than they bring in trade revenues. This has contributed to the rise of privately developed zones and zone development as a form of real estate speculation. In the 1980s, the first privately developed SEZs were created in the Caribbean and Central America to compete with government trade zones that were deemed inefficient with cumbersome regulations. SEZ ownership has gone from 100% government owned and operated in 1975, to roughly 25% privately owned and operated zones in the 1980s, to 62% privately owned and operated zones in the mid-2000s (FIAS 2008).

The increasing privatization of zones does not necessarily entail a withdrawal of the state from zone promotion and regulation. In practice, the most “successful” zones are those intertwined with the national economy, the national investment environment, and the capacity of the government. This gives rise to government offering “one-stop service centers” on site, including customs and import-export officials and labor ministry representatives, for example. At the same time, many programs allow private zone developers to supply electricity and other utilities to the zone, and SEZ programs include provisions for commercial and professional activities. In sum, the increasingly complex intertwining of public and private concerns in zone development and administration may come at the expense of government control and oversight over economic development. Yet it offers a wide range of new opportunities for state involvement, including the finance and real estate speculation of land procured for zones, new forms of territorial control, and the increasing prevalence of politically connected domestic elites turned zone developers, thus combining neoliberal rationality with authoritarian political styles (Arnold 2012). In the past decade zones have been utilized as geopolitical tools, with private and public investors from China expanding its SEZ model globally by establishing “economic cooperation zones” in Africa, and SEZs in Southeast Asia. Zones continue to play a clear role in new forms and experiments with liberalization, land and labor regulation, and urban and regional planning, while social and environmental concerns persist.

Relaxed environmental standards are an incentive to increase profitability, often with severe environmental damage occurring as a consequence. This provides an incentive for relocating production to places where environmental regulation is less stringent. This response to
environmental regulation by enterprises may weaken the state’s determination to enforce environmental regulations. In Tijuana, Mexico, the Rio Grande is so polluted from maquiladora waste that it has caused an increased risk of Hepatitis A. In China, it is estimated that more than 80% of coastal water and about 70% of its rivers are polluted with industrial waste, raw sewage, and agricultural runoff.

Numerous EPZs have thrived as a space for globalized production – as a capital-led strategy for low-cost exports and as a state strategy to absorb surplus labor and attract FDI – yet they have typically failed to create decent jobs (McCallum 2011). Work in EPZs and assembly-based SEZs is synonymous with low pay, excessive overtime, and restrictions on trade union activity. Employment is gendered in the labor intensive sectors such as toy, textile and garment, and electronics manufacturing, the mainstay of EPZs over the past 50 years. The typical zone worker is a young female, rural-urban or international migrant, employed in factories for a relatively short period of time. Once these female workers have children, and/or lose the mental and physical faculties for which they were initially employed, they are dismissed and younger substitutes replace them. In other words, over time workers become a form of industrial waste, at which point they are discarded and replaced (Wright 2006). Yet the contradiction is these workers produce much prosperity through their own industrial destruction (Wright 2006). Despite these and other hardships, EPZ workers are not passive victims of structural forces. China’s light manufacturing workers at the heart of the “world’s factory” have led numerous and large-scale wildcat labor protest for better social and economic entitlements. This has contributed to substantial wage increases over the past 10 years.

Displaced farmers and peasants have also mounted numerous protests in different parts of the world, often in response to insufficient compensation for their land appropriated for an SEZ. This reflects development processes that rest on displacement of the poor to make way for SEZs, and the interrelationship that exists between spatiality of capital and the opening up of new economic spaces (Banerjee-Guha 2009). Displaced farmers not only lose their economic means of survival, but generally lack the skills required to be employed in construction of the zones and supporting infrastructure, and in eventual manufacturing, service, or other industries located in the zones. Yet dispossessed peasants and exploited workers employed in the zones are not mutually exclusive – the relation to one other, as well as state practices and global supply chain dynamics, are crystalized in zone development (Mezzadra and Neilson 2013).

Conclusion

EPZs have changed over time, from manufacturing enclaves effectively shuttered off from a country’s wider political economy, to incubators of economy-wide liberalization. The early EPZs were key components of East and Southeast Asia’s rapid industrialization and economic growth, and later SEZs in China proved to be an experiment with liberalization on the leading edge of an unprecedented social and political economic transformation. In recent years late industrializing countries are keen to replicate China’s SEZ “success,” initiating their own assembly-oriented SEZ programs. Yet the social and environmental costs often outweigh the economic gains. Rather than become a footnote to history, the number of zones has expanded, and along with it questions on the sustainability of dominant economic growth models that give
rise to economic spaces characterized by labor rights violations, environmental degradation, land confiscation, and gendered violence.

**SEE ALSO:** Global commodity/value chains; Rights, labor; States and development; Trade, FDI, and industrial development

### References


**Externalization**

W. Richard Goe  
*Kansas State University, USA*

Externalization refers to the use of market contracting by a firm to procure a good and/or a service (i.e., a product) from another independent firm as opposed to producing that product internally within the former’s organizational hierarchy. This phenomenon is also referenced by other concepts used in academic discourse in economic geography, regional science, economics, and economic sociology, including outsourcing, vertical disintegration, and flexible organizational forms. Research on externalization within these fields has predominantly focused on market contracting between: (i) manufacturing firms and their suppliers, and (ii) the headquarters of corporations and firms in the producer services sector.

**Externalization in manufacturing**

The decade of the 1980s witnessed the onset of a period of rapid and extensive “restructuring” in the economies of the developed nations that coincided with the rise of neoliberal, conservative political movements. One dimension of this process was the development and implementation of new models of business organization and strategy by multinational corporations (MNCs). An important prong in these strategies was the growing use of externalization to procure needed goods and services. Within the United States, for example, manufacturing firms increased their use of externalization while “downsizing” by closing facilities and/or eliminating jobs (also called vertical disintegration) in response to competitive pressures and/or the opportunity to elevate profits and stock valuations by reducing costs. This strategy had the additional benefit of transferring the investment risk associated with the production of needed goods and services to supplier firms. It was argued that the large, vertically integrated firm described by Alfred Chandler (1977) as the hallmark organizational form of twentieth-century capitalism was being superseded by new forms of organizing production that involved externalization and the formation of subcontracting networks between manufacturing firms and their suppliers. A second important dimension of change was the continuing rise in the competitiveness of Japanese and European firms across an expanding range of manufacturing industries, who utilized business strategies and modes of organizing production that were different from those used by manufacturing firms in the United States. Finally, the decade of the 1980s was not only a period of heightened corporate consolidation via mergers and acquisitions, but was also marked by the development and increasing use of important innovations in computers, robotics, and other forms of information technology, corporate finance (e.g., junk bonds), and employment practices (e.g., contracting for temporary workers), among others.

Economic geographers, regional scientists, economists, and economic sociologists developed stylized theoretical models that described the key features of the form of economic organization resulting from these changes. Building on concepts developed by the French school of regulation theory, it was posited that these new
forms of production organization were providing a key dimension to a new “mode of regulation” that would allow capitalist market economies to resume a long-term, stable period of accumulation. One such model that was subject to extensive theoretical debate and empirical research was flexible specialization, brought to the attention of the English-speaking world by Piore and Sabel (1984). A key feature of this model is the spatial clustering of small and medium-size firms to form a flexibly specialized, industrial district where they engage in stable, long-term subcontracting relationships. Firms within these industrial districts cooperate by sharing information and technical expertise, engaging in joint planning and contract bidding, and sharing technical, financial, and educational services from common providers. These subcontracting arrangements allow production costs to be spread across firms in the industrial network. Moreover, the sharing of information and technical expertise promotes interfirm learning and the more rapid development and introduction of new products. It was contended that the fragmentation of mass markets and shorter product cycles requires firms to respond with greater flexibility by more rapidly developing new products and producing a greater diversity of products with the help of computerized industrial machinery and robotics, in addition to other technologies. Empirical evidence to support the flexible specialization model was drawn primarily from central and northeast Italy (knitted goods, special machinery, ceramic tiles, textiles, agricultural implements, hydraulic devices, shoes, and electronic musical instruments), the Baden-Württemberg region of West Germany (textiles, apparel, textile machinery, machine tools, and automobile components), the “Second Denmark” region in Jutland (textiles, apparel, furniture, machine tools, and shipbuilding), and the motion picture industry in the Los Angeles region.

One point in the critique of the flexible specialization model was that the concept encompasses a diverse range of business practices, some of which may actually serve to reduce flexibility. For example, Sayer (1989) distinguished between functional flexibility (a firm changing its product configuration) and numerical flexibility (in output and employment). In turn, these are facilitated by flexible labor markets (internal and external), flexible working practices (e.g., job rotation), flexible machinery (e.g., robotics), flexibility in restructuring (e.g., corporations paring off and selling nonprofitable divisions), and flexible organizational forms (e.g., networks of specialist producers). Flexibility in restructuring, for example, could limit investment in flexible machinery due to increasing debt, which could in turn reduce flexibility. A second important criticism was that unless they grew into large firms (thereby presumably decreasing their flexibility), the firms that made up flexibly specialized industrial districts were not likely to become dominant competitors that exerted a strong influence over the field of practice within their respective industry(ies) in the global economy. Nonetheless, in the two decades that have followed this research program, the use of subcontracting networks in manufacturing, the ability to reconfigure production lines to produce a greater diversity of products, and the ability to more rapidly develop and introduce new products have become important features of manufacturing in the global economy.

A second model that was subject to extensive theoretical debate and empirical research was the Japanese system of manufacturing, which was described through such metaphors as flexible rigidities and innovation-mediated production (Dore 1986; Kenney and Florida 1993). The Japanese system was also characterized by a spatial clustering of firms that were engaged in subcontracting relationships. A key difference from
flexible specialization was that the subcontracting network (e.g., *keiretsu*) was anchored by a large manufacturing firm that produced the final assembled product. Similar to those in the flexible specialization model, however, the relationships between firms in the network of subcontractors were long term, relatively stable, and cooperative in nature. For example, Dore (1986) describes the relational contracting arrangements between manufacturers and suppliers in the Japanese textile industry as being “moralized trading relationships based upon mutual goodwill.” Trading partners forego the use of any advantages in bargaining superiority, fairly share the gains and losses during periods of growth and decline, and forego any short-term price advantages offered by alternative suppliers. In effect, while anchoring firms have greater economic power than supplier firms, opportunism is avoided in favor of the longer-term benefits that may potentially be realized from a trading partner in the future. Supplier firms deliver parts to the factory(ies) of the anchoring firm on a “just-in-time” (JIT) basis as they are needed on the production line. The long-term benefits of the subcontracting networks in the Japanese JIT system include decreased inventory, heightened utilization of equipment, minimal scrappage or rework of parts and supplies, higher quality, and increased technological efficiency. Increased technological efficiency is achieved through rapid, systematic transfer of information between firms in the JIT network which facilitates joint learning by doing, coordinated investment, joint product planning and design, and the diffusion of new techniques across firms in the network (Dore 1986; Kenney and Florida 1993; Sayer 1989).

As a result of the competitive advantages of the Japanese system, the dominance of the Japanese economy appeared imminent to some analysts in the 1980s. Changes in economic conditions in the 1990s, however, undermined these advantages as the Japanese economy entered a long-term period of economic stagnation and malaise. This ostensibly allowed surviving competitor firms in the United States and other developed nations to catch up and regain their level of competitiveness. While few would now contend that Japan will become the next hegemonic power in the global economy, elements of the Japanese system have been widely adopted as technologies in manufacturing, including the use of subcontracting networks for JIT delivery of parts and supplies, work teams, job rotation, and total quality management to continually improve product quality.

Externalization and producer services growth

Downsizing by firms in the manufacturing sector (i.e., deindustrialization) in the 1980s, combined with rapid employment growth in other economic sectors, produced an ongoing decline in the share of total employment accounted for by manufacturing in the economies of most developed nations. As the economic expansions of the 1980s and 1990s unfolded, the vast majority of new jobs created were in service sector industries, furthering the perception that the developed nations had undergone a shift to service-based economies. Closer examination of this growth revealed that the largest and most important source of job growth was in the “producer” services. In theory, producer services are services that are ultimately used by businesses and other productive enterprises. In practice, the producer services sector has been operationally defined as including all or a subset of: banking and financial services, insurance, real estate, business services, legal services, engineering services, accounting, and architecture. The business services, in particular, subsume a diverse range of
work activities including advertising and marketing, management consulting, computer and IT services, private security, and employment agencies, among others.

The growth of the producer services sector stimulated an extensive research program by economic geographers, regional scientists, economists, and economic sociologists to better understand the causes and implications of this growth. A key issue concerned the fact that from the demand side, producer services growth was being driven by the externalization of service functions by business firms, government, and other nongovernmental agencies. The growth of producer services employment was found to be concentrated in the metropolitan areas of the developed nations, which also served as the locations for the headquarters of MNCs and other large corporations. Decisions to externalize producer services are primarily made by managers at the headquarters of the outsourcing firm. Externalization therefore leads to the formation and growth in density of corporate complexes, which are interfirm networks, anchored by the headquarters of a firm, which engages in market contracting with one or more vendors of different types of producer services.

The potential of realizing agglomeration economies would suggest that the headquarters of a firm would choose to contract with nearby producer services firms that are located within the same metropolitan area. However, empirical studies of producer services firms indicated they extensively engage in interregional and international trade outside of the metropolitan area in which they are located (for example, see Beyers, Alvine, and Johnson 1985). This indicated that corporate complex networks frequently extend beyond the boundaries of the metropolitan area in which the headquarters of the contracting firm is located. Empirical research on the geography of network linkages within corporate complexes indicated that corporations tend to contract with producer services firms located in the same metropolitan area when the producer services being purchased require frequent contact between the contracting firm and vendor. The probability of contracting with a distant vendor increases when contact requirements are low (for example, see Goe et al. 2000).

A point of debate concerned the extent to which the growth of producer services was tied to the changes in the manufacturing sector described above. The industrial paradigm of economic development emphasized the role of manufacturing as the central lynchpin in economic growth. Service industries were viewed as parasitic to manufacturing and dependent upon the export of goods by manufacturing industries for the revenue needed to sustain their growth. One perspective of producer services growth argued that continued deindustrialization in the developed nations would result in the eventual decline of their producer services sector as well because of its strong direct dependency on the manufacturing sector. In contradiction to this perspective, a substantial number of empirical studies of producer services firms, conducted in a variety of metropolitan regions and nations, found that firms in other service sector industries were providing the primary source of demand for the producer services firms that were studied (Goe 1990; Illeris 1989; Ley and Hutton 1987). These findings suggested that producer services growth could be stimulated within a metropolitan region without a direct, primary dependency on manufacturing firms for revenue.

The externalization of producer services was also situated within the flexible specialization model. Coffey and Bailly (1991) contended that externalization by producer services firms to procure inputs provides them with flexibility in controlling their costs and externalizing the risks associated with the production of the inputs they
require. Similarly, the externalization of producer services by the headquarters of contracting firms allows them to realize these forms of flexibility in control over their production and administrative tasks. A related point of debate concerned whether the growth of producer services was being primarily driven by vertical disintegration (also termed unbundling) as firms externalized producer services functions and eliminated their internal capacity to provide these services. Empirical studies of firms who externalized producer services suggested this was not the case. Rather, employment in producer services occupations was found to be growing internally within these firms in conjunction with their use of externalization (for example, see O’Farrell, Moffett, and Hitchens 1993). These findings suggested there was likely to be some form of functional complementarity between the internal producer service activities and those that were externalized and/or contracting firms which used externalization to handle temporary work overloads in regard to specific internal activities.

Empirical research conducted on firms externalizing producer services has also found that externalization decisions are not based on pure cost considerations. Goe et al. (2000) found that contracting firms were willing to forego higher costs if the service being externalized was deemed to be of sufficient strategic importance to the firm. Coffey and Drolet (1996) found that the lack of sufficient in-house expertise to provide a high quality service and infrequent and/or irregular demand for a service were much more important in motivating externalization than cost considerations. These findings highlight the role of many producer services (e.g., management consulting, advertising, and marketing) in transferring strategic knowledge and expertise to contracting firms (see Harrington and Daniels 2006). Lindahl and Beyers (1999) found that the competitive advantage of producer service firms lies predominantly in their ability to differentiate themselves through their creativity, their ability to conduct research and development, and their ability to respond quickly and provide personal attention to their clients. Taken together, these studies underscore the role of knowledge creation and innovation as factors underlying the externalization of producer services and the growth of the producer services sector.

Conclusions

In sum, over the past three decades, the research programs on flexible specialization, the Japanese system of manufacturing, and producer services drew attention to the phenomenon of externalization. Each of these programs presented models of economic organization where firms favor the use of market contracting to externally procure parts and supplies, jointly manufacture a product, or purchase services used as business inputs. This body of research situated externalization as an element in the broader organizational strategies being employed by firms in response to the exigencies of international economic competition. During this time frame, manufacturing systems have become increasingly international in scale and scope as part of the so-called process of globalization. There has been a growing use of subcontracting for production work as China and other developing nations (e.g., Mexico) have become offshore manufacturing platforms for firms located in developed nations. A strong case could be made that an important characteristic of globalization is that production work has become a low value-added labor activity that is now externalized by MNCs and predominantly carried out in the low-wage environments of developing nations. Labor activities required for the management of MNCs as well as the invention, development, design, and marketing of
products represent higher value-added activities that are predominantly located in the developed nations. The acquisition of services to facilitate the management and operation of these complex organizations thus accounts for a component of producer services growth in the developed nations. The networks that bind corporate complexes have also become increasingly international in scale and scope. India has developed as an important offshore platform for firms in the developed nations to purchase software and other types of business services (e.g., call centers), thereby devaluing certain types of work in the producer services. An increasing number of firms in producer services such as management consulting, software development, and legal services, for example, have evolved into MNCs as the markets for their service products have expanded. At this juncture in time, externalization continues to increase as an element in global economic competition.

SEE ALSO: Corporate spatial organization and producer services; Economic geography; Flexible specialization; Industrial agglomeration; Industrial districts; Industrial restructuring; New economic geography

References


Factors of production

Dean M. Hanink
University of Connecticut, USA

At a basic level, any good or service is the result of the purposeful combination of its factors of production. The two elemental factors of production are capital and labor. At times land is also considered an elemental factor of production, but unless agriculture is being considered it is usually treated as a component of capital. By broad definition, capital is also taken to include natural resources such as mineral ores or water. Capital is typically considered in fixed form, such as the buildings and machinery used in production. It is rarely considered in financial form. Capital is often more finely defined with respect to its private or public ownership; publicly owned capital is often referred to as infrastructure. Labor is typically counted as number of workers or hours worked in production. It may be differentiated by skill level. Broadened definitions of capital and labor sometimes overlap in the factors of human capital and social capital. Human capital takes the form of education and health characteristics in the labor force that contribute to production. Social capital contributes to production through cooperation in social networks of producers. Technology is sometimes considered an independent factor of production and sometimes it is viewed as being embodied in capital or in labor or in both factors.

Production requires that both capital and labor be combined in their application as factors. Their combination and its effect are often described in production functions. The Cobb–Douglas production function (Douglas 1976) takes the form

\[ Q = (TL^\alpha K^\beta) \]

where \( Q \) is the maximum level of output, \( L \) is labor, \( K \) is capital, and \( T \) is total factor productivity, which accounts for the output not directly attributable to capital or labor singly. Elasticities of output to capital and labor are defined by the values of \( \alpha \) and \( \beta \), respectively. If their sum is less than one, then there are diminishing returns to scale, meaning that output is increasing on the margin at a slower rate than the combined factors are being used. If their sum is equal to 1, then there are constant returns to scale, and if their sum is greater than 1, then there are increasing returns to scale.

Goods and services can be classified by their factor intensities. Labor-intensive goods and services use relatively more labor inputs than capital in their production. Capital-intensive goods and services use relatively more capital than labor in their production. Variations in relative factor intensity among products and geographical variations in the distribution of the factors, or factor endowments, underlie a theory of location and trade developed by the economist Bertil Ohlin (1933). That theory suggests that trade that occurs among places is based on their relative production efficiencies, so labor-intensive products produced in places with rich labor endowments are traded for capital-intensive products produced in places with relatively large capital endowments. The factor content of a unit of production can be calculated by a method that uses an input–output table, as described by the economist Wassily Leontief (1986). That method is often used to analyze trade flows in the context of Ohlin’s factor endowment theory.
FACTORs OF PRODUCTION

SEE ALSO: Human capital; Regional and interregional trade in producer services; Social capital

References

Famine

Stephen Devereux
University of Sussex, UK

Famines used to be understood as natural disasters but are increasingly recognized as products of human action or inaction. Despite being seemingly straightforward events, only recently has an operational definition of famine that triggers a humanitarian response been internationally agreed. One reason is that famines are actually extremely complex phenomena, and have been analyzed from various disciplinary perspectives, including climatology, demography, economics, geography, and political science.

After reviewing alternative approaches to defining famine, this entry provides a brief historical overview, showing how famines have been geographically concentrated in a few regions – Europe, South and East Asia, and sub-Saharan Africa. The causes of twentieth-century famines are analyzed under four “c’s” – climate, conflict, colonialism, and communism – highlighting the role of politics in creating or exacerbating weather-related triggers. The recognition that politics is central to understanding recent famines is also seen in a review of the evolution of famine theories – from a focus on demographic and climatic factors towards the economics of famine, and “new famine” thinking about the failure of responsible actors to deliver humanitarian relief.

The four mass mortality twenty-first-century famines to date – in Ethiopia in 2000, Malawi in 2001/2, Niger in 2005, and Somalia in 2011 – are examined individually and patterns are identified that suggest why similar famines are likely to recur in the future. Four common factors are weather shocks, conflict or civil insecurity, weak or absent democratic institutions, and unaccountability of national and international actors with responsibility for famine prevention.

Defining famine

Famines are easy to recognize but challenging to define. Commonly understood as the death by starvation of large numbers of people, academic definitions have struggled with issues of scale (how many hunger-related deaths constitute a famine), duration (when a famine begins and ends), even causality (most famine victims die of diseases, not starvation) and outcomes (whether or not a famine process that stops short of mass mortality nonetheless constitutes a famine).

The Somalia famine of 2011 was unique in that it was the first to be officially declared a famine by the United Nations (UN), using an operational definition called the Integrated Phase Classification scale (IPC 2012), which identifies five escalating levels of acute food insecurity, where level 5 is a “Humanitarian Catastrophe/Famine.” The key indicators defining a level 5 situation for a defined population include a global acute malnutrition (GAM) rate above 30% and a crude death rate (CDR) above 2/10000/day. In July 2011, nutrition and mortality surveys conducted in 16 locations in southern Somalia found that the GAM and CDR rates exceeded the IPC level 5 thresholds in 11 and five locations, respectively. These findings provided the basis for the UN declaration of famine, which triggered a massive international humanitarian response, though too late to prevent a quarter of a million deaths (see the discussion of this famine below).
This episode highlights both the value and the limitations of a technocratic approach to defining famine. There is no doubt that having international consensus on an objective, measurement-based definition of famine is a major advance on the past, when disagreements between governments and aid agencies about whether famine conditions had yet been reached often led to fatal delays in relief efforts. On the other hand, two risks follow from defining famine in this narrowly technical way. The first is that no action will be taken until the thresholds for declaring a famine have been exceeded, meaning that large numbers of people are already starving, destitute, or dead before an emergency food aid appeal is launched. The second is that the focus on preventing mass deaths by starvation could divert policy attention away from less visible manifestations of hunger and food crisis, including the millions of malnutrition-related child deaths that occur every year in South Asia and sub-Saharan Africa.

People who have actually experienced famine identify different levels: minor famines cause hunger, severe famines cause destitution, and only catastrophic famines result in deaths. These insider definitions see famine as a continuum rather than a discrete event, but this subtlety is overlooked by binary definitions based on thresholds for food security indicators which trigger humanitarian responses that are invariably late. If the objective measurement approach was combined with the subjective perceptions of local people themselves – which could be validated by monitoring their coping strategies as food security deteriorates – and the IPC scale was relabeled with varying degrees of famine (instead of “stressed,” “crisis,” “emergency,” “famine”), this could ensure that responses are mobilized early enough to prevent not only mass deaths but also widespread destitution and unacceptable levels of hunger.

Geographical trajectories

Famine has a history as long as that of humankind itself. Famines were recorded in Egypt more than 6000 years ago, mostly triggered by erratic weather. The Bible reports conflict-related famines such as the siege of Jerusalem. Western Europe was famine prone at least from the time of the Roman Empire. China, India, and Russia – known historically as lands of famine – have each suffered famines for thousands of years. More recently, over 70 million people died in famines during the twentieth century (Devereux 2000). Although famine has been perceived as an exclusively African phenomenon since the Great Ethiopian Famine of 1984, the overwhelming majority of famine deaths since 1900 occurred not in Africa, but in Asia and Europe. China and the Soviet Union alone accounted for over 80% of twentieth-century famine mortality (Devereux 2000, 9).

In the past, when agrarian communities were isolated and where rural families depended on rain-fed agriculture as their principal source of food, a single harvest failure was often enough to trigger a food crisis. Geographers identified two climate-related “famine belts.” Northern Europe, the Soviet Union, and China were vulnerable to crop failures because of cold, excessive damp, and short growing seasons. Sahelian Africa, the Horn of Africa, and South Asia were vulnerable to crop failures because of drought or flood, often triggered by El Niño events (Cox 1981).

In the late nineteenth century a sequence of massive famines struck India, China, Ethiopia, and other countries in Asia and Africa. Although the Americas have been virtually famine-free, even Brazil was affected. These “Late Victorian Holocaus ts” killed between 30 and 60 million people (Davis 2001). They were the product of
a lethal combination of climate shocks, colonialism, and the introduction of capitalism and market-driven economic policies that included exporting food during famines.

In the early twentieth century, a series of small-scale famines struck countries across sub-Saharan Africa, mainly triggered by drought and colonial policies. In the 1920s and 1930s a sequence of famines struck China and the Soviet Union. In the 1940s, disrupted global food systems and the use of siege tactics as a weapon during World War II triggered numerous famines across Europe, Asia, Australasia, and Africa. The 1950s and 1960s were relatively quiet decades for famines, with the significant exception of the Great Leap Forward famine in China, the worst famine in history in mortality terms.

The late 1960s also saw a decisive shift in the prevalence of famine, away from South and East Asia and towards sub-Saharan Africa. Between 1968 and 1999, 13 of 17 major famines occurred in Africa, including three each in Ethiopia and Sudan and two in Somalia. Most of these were triggered by conflict and/or drought. The four remaining famines were all in Asia (Bangladesh, Cambodia, Iraq, and North Korea). There have been four twenty-first-century famines to date, all in Africa (Ethiopia, Malawi, Niger, and Somalia) – these are discussed individually below.

Causes of famine

Famines occur at the intersection of bad weather and bad politics. Most famines since the late nineteenth century can be attributed to four “c’s” – climate, conflict, colonialism, and communism – often operating in lethal combinations.

It is unfashionable to attribute famines to climate shocks (drought, flood) or environmental stress (Malthusian population pressure). However, climate shocks continue to be trigger factors, including for all four twenty-first-century famines to date. Climate shocks have the effect of undermining the subsistence basis of agrarian economies that depend on agriculture for their food supply. A common feature of famines that are triggered by natural disasters or weather shocks is that they affect mainly rural communities. It is a paradox that people who grow food – farmers, agricultural laborers, and livestock herders – are most susceptible to food crises. Droughts, floods, and cold spells – most prevalent in sub-Saharan Africa, South Asia, and Northern Europe, respectively – trigger harvest failures that can leave entire communities with no food crops to eat and no cash crops to sell for income to buy food.

Secondary effects include loss of livestock (or depleted value of livestock if they lose weight due to lack of grazing and water) and loss of local employment opportunities (because farmers have no need to hire labor for weeding or harvesting, while demand for services collapses because of the income shock). In isolated rural areas with fragmented markets, food prices typically rise as market supplies are depleted and demand for food surges, when normally self-provisioning producers unexpectedly become market-dependent consumers. The collapsing terms of trade as food prices rise and asset (e.g., livestock) prices fall leave the poorest unable to buy the food they need. Unless they find alternative sources of food and income (including by migrating out of the famine zone) or humanitarian relief arrives, mass starvation will follow.

Key characteristics of famine-prone communities that are vulnerable to climate shocks include: local livelihoods are undiversified, so a harvest failure cannot be compensated with income earned from other sources; local economies are weak, so traders do not react promptly to move
food from surplus to deficit areas; insurance markets are effectively nonexistent, so farmers have no possibility of insuring against crop losses; and local communities are geographically isolated and politically marginalized, so food aid is either not mobilized or arrives too late. The point is that climate shocks do not cause famines, but vulnerability to climate shocks does. In countries like Ethiopia, Niger, and Sudan, economic and political vulnerabilities persist into the twenty-first century, which is why the deep rural areas of these countries remain vulnerable to famine.

In this sense, the eradication of famine from large parts of the world, especially in Europe and most parts of Asia, can be seen as a by-product of normal economic development processes. In countries like China, India, and Russia in the late nineteenth and early twentieth centuries, rapidly improving transport and communications infrastructure had the effect of integrating previously isolated communities into larger – national and even global – economies. Livelihood options expanded, trade flows into and out of rural communities increased, and governments and traders were aware earlier and had more capacity to respond promptly to signals of distress from districts where slow-onset food crises were developing. In northern China, for instance, a protracted drought in the 1870s resulted in an estimated 9–13 million famine deaths, but in the next 40 years 6000 miles of railway were constructed and, when a comparable drought struck the same region in the 1920s, relief supplies were delivered by rail from other parts of China and mortality was contained to approximately half a million.

A second major trigger factor for famine is conflict, including the use of food as a weapon of war, which has been recorded since Biblical times, when cities such as Jerusalem were besieged and their inhabitants were deliberately starved. During World War II this tactic was responsible for mass death by starvation in several European cities – in Greece (Athens), Poland (Warsaw), the Soviet Union (Leningrad, Kharkov, Kiev), and the Netherlands (Amsterdam, The Hague, and Leiden). These famines were atypical in that they affected urban centers more than rural areas, and wealthy people were as vulnerable as the poor. Vulnerability in blockade famines is related to the separation of urban populations from food supplies. Mass starvation is caused by blocking flows of food into the city and preventing movements of people out of the city.

The scale of famine-related mortality during World War II has only recently become evident. Collingham (2011) claims that at least as many people died during World War II from starvation, malnutrition, and associated diseases (20 million) as from military action (19.5 million). Apart from deliberate blockades, the causes of famine deaths in World War II ranged from disruptions of food production and supplies, to systematic starvation of prisoners of war, to failure to respond to natural disasters because all efforts were focused on the war.

Conflicts have the potential to disrupt all components of the food system: production, trade, and relief. In rural areas, conflict causes farmers to be displaced from their land, conscripted into armies, disabled or mutilated by landmines and militia, or killed. Often standing crops are burnt, granaries are raided or destroyed, water sources are poisoned, and livestock are stolen or slaughtered. All of these side effects of conflict undermine agricultural production. Conflict also disrupts trade in food and other commodities. Roads become unsafe or blocked by military checkpoints, so traders either cannot purchase food or cannot transport food to conflict zones. Markets are targeted by militia or military aircraft: during the Ethiopian famine of 1984, many rural markets operated only at night, to avoid being
bombed. Humanitarian relief operations are also compromised. Governments or rebel forces sometimes ban humanitarian agencies from operating in conflict zones, or agencies withdraw their staff, either fearing for their safety or to prevent relief supplies and vehicles being conscripted by the warring parties. In Somalia in 2011, for instance, famine-affected communities were “no go” areas for most humanitarian agencies. As a consequence, people living in conflict zones become cut off from external food supplies, and when their own stocks are depleted, raided, or destroyed, they face the threat of famine.

Many of the worst twentieth-century famines occurred in totalitarian regimes: the Soviet Union, China, Cambodia, North Korea, and Ethiopia. Policies such as collectivization of agriculture reduced farmers’ incentives to produce food, excessive extraction of food through quotas reduced farmers’ access to food, pressure on local officials to demonstrate success reduced their incentive to report food shortages to the central level, and an authoritarian regime that was unaccountable to its citizens had little incentive to admit to policy errors or to respond to prevent famines.

Conversely, a powerful antidote to famine is democracy. Amartya Sen (1999) argued that the defining features of democratic systems, notably free and fair elections and a free press, help to prevent famines. Regular elections give citizens an opportunity to throw out a government that fails in its duty to prevent famine. A campaigning press exposes problems such as evolving food crises to public scrutiny and allows the electorate to hold their leaders to account, unlike in totalitarian states where the media are strictly controlled, civil society is suppressed, and leaders are unaccountable. More recently, political scientists have nuanced this argument, pointing to the limitations of democracy as a famine prevention mechanism (Rubin 2011), especially where it is weakly embedded institutionally, or if demographic structures generate ethnic majorities that allow elected political leaders to neglect marginalized minority groups with impunity. This is a feature of recent African famines, which have often occurred in countries like Ethiopia and Malawi that are nascent democracies.

Famine theories

Crudely speaking, the evolution of famine theorizing has reflected shifts in disciplinary emphasis, starting from geography (climate and environment), demography (Malthusianism), and then economics (Sen’s entitlement approach), and most recently political science (social contracts). However, a comprehensive but cogent political theory of famine has yet to be elaborated.

The simplistic view that famines are natural disasters, caused by environmental catastrophes or population pressure on overstressed ecosystems, dominated famine thinking and analysis until the 1980s, when Sen’s “entitlement approach” shifted the focus of analysis from geography and demography to economics, and the 1990s, when political scientists drew attention to the politics of famine processes.

Famines were initially conceptualized as failures of food supply or availability: a harvest failure that leaves too little food (climate theories), or population growth in a resource-constrained environment until there are too many people to feed. Malthusian demographics even characterized famine as a stabilizing mechanism, balancing the carrying capacity of the natural resource base with the consumption needs of human and animal populations (Ghatak and Ingersent 1984). More recently, political scientists and geographers increasingly saw this environmental determinism as inadequate.
FAMINE

After Amartya Sen demonstrated that there was adequate food at national level in at least three of four twentieth-century famines that he reanalyzed, the focus shifted from “food availability decline” to failures of effective demand or access to food. Sen’s entitlement approach refers to people’s ability to acquire food based on their resources (own production, labor, purchase, or transfers), rather than to moral or legal claims on the state to secure their right to food. Sen (1981) pointed out that famines never affect entire countries, and that rich people rarely die during famines, except in wartime. It is poor, vulnerable, and marginalized population groups who are most likely to lose access to available food and to face starvation.

Building on this, Watts and Bohle (1993) defined vulnerability to hunger and famine in terms of social spaces, not just physical spaces. They argued that analysis of hunger and famine should examine not only “the particular distribution of entitlements and how they are reproduced in specific circumstances,” but also “the larger canvas of rights by which entitlements are defined, fought over, contested, and won and lost (i.e., empowerment or enfranchisement); and the structural properties (what we shall call crisis proneness) of the political economy which precipitates entitlement crises” (Watts and Bohle 1993, 44).

These insights shifted the focus of famine analysis from aggregate supply failure to group-specific demand failure, and initiated a paradigm shift in the famine discourse. Before Sen, most proposed solutions to famine focused on increasing foodcrop yields to increase national food availability, whereas the implications of the entitlement approach are that access to food must be protected for specific groups of people who are vulnerable to a collapse in their livelihoods following a shock to the (local, national, regional, or global) food system.

“New famine” thinking (Devereux 2007) argues that a second paradigm shift is needed, to reflect the reality that most contemporary famines could and should have been prevented. If local food supplies are inadequate, and if impoverished people lack resources to buy the food they need for subsistence, national and international duty-bearers have a moral responsibility to protect the basic human right to adequate food for all. Alex de Waal (1997) has argued that the most effective famine prevention system is an “anti-famine political contract” between states and citizens, which makes famine politically unacceptable. De Waal points out that there has been no major famine in India since the Bengal famine of 1943. Successive post-independence governments have implemented interventions such as the Scarcity Manuals, the Employment Guarantee Scheme, and the National Food Security (“Right to Food”) Act, knowing that allowing another famine to occur will not be tolerated by India’s robust media and activist civil society.

The capacity to intervene has increased exponentially since World War II, with rapid advances in crop yields, global transport and trade networks, famine early warning systems, and the international humanitarian system. Famines happen because they were not prevented. This is a political failure (or political intent) for which national governments, rebel militia, or humanitarian relief agencies should be held accountable.

Twenty-first-century famines

The four famines that have occurred since the year 2000 give insights into why famine has not yet been eradicated, despite unprecedented improvements in global technical and institutional capacity to do so in recent decades. These famines also give insights into the likely risk factors for famine recurring in the future, and
therefore into how they might be prevented. It is no coincidence that the four African countries that experienced famine between 2000 and 2012 are among the poorest in the world and have weak political institutions, with citizens whose livelihoods are dependent on rain-fed agriculture which is highly susceptible to weather shocks. However, contemporary famines are analyzed as failures not only of food production (the focus of earlier natural disaster and (neo-) Malthusian theories), but also of access to food (due to market failure, entitlement failure, or poverty) and of humanitarian response (food aid).

The famine that afflicted Somali region, Ethiopia, in 2000 caused 70,000–122,000 deaths. This was a slow-onset crisis, triggered by a protracted drought that started in 1997 which decimated livestock herds, the main source of both food and income for pastoralists. The drought was compounded by a ban on livestock exports to Saudi Arabia due to a Rift Valley fever outbreak in East Africa in 1998, which halved exports to this lucrative market from 3 million to 1.5 million head and cost pastoralists and traders an estimated US$100 million in lost income. So this was simultaneously a production and an income failure. Unable to sell their animals, many pastoralists watched them die and were left with no income or assets to exchange for food.

Ethiopia has had a famine early warning system since the 1970s, but it monitors the crop farming highlands rather than the pastoral lowlands, which were the epicenter of the famine in 2000. Nonetheless, anecdotal evidence about deteriorating food security conditions in Somali region was ignored by the federal government, which was preoccupied with a border war with Eritrea at the time, and by the donor community, which was concerned that food aid intended for the famine-affected south might be diverted to feed soldiers fighting in the north. By the time the international media drew attention to the famine and food aid was delivered, three-quarters of famine deaths had already occurred. The food crisis in Somali region might have been triggered by drought and livestock disease, but the ultimate cause of the famine was the politicization of the humanitarian response.

An agricultural drought in Malawi in 2001—a break in the rains during the growing season—reduced the national maize harvest by 32% and was blamed for the estimated 47,000–85,000 excess deaths that followed. However, this harvest followed two bumper years and was 6% higher than the 10-year average, so food availability decline is an inadequate explanation for this famine. The famine was concentrated in the densely populated southern and central districts, where landholdings are too small to produce a full year’s food supply so smallholder families are market-dependent for several months. The reduction in maize supplies in local markets triggered an increase in retail prices of up to 500%, making the acquisition of adequate food during the hungry season impossible for poor smallholder families, even as they sold off their livestock and other assets at distress prices to finance food purchases. There was too little work and the oversupply of labor caused agricultural wages to collapse. Some analysts argued that the food price spike rather than the harvest failure was the main factor explaining how a moderate production shock became a major food crisis.

Malawi’s famine early warning system failed to predict the food shortage in 2001, instead forecasting that any maize deficit would be more than covered by drought-tolerant cassava. These inaccurate projections prompted the World Bank and International Monetary Fund (IMF) to advise the government to sell off its strategic grain reserve (SGR) and replenish it after the harvest. This left Malawi with no buffer stocks to distribute, and when suspicions arose that government officials had profiteered corruptly.
Famine

from sales of SGR maize, the donors demanded an explanation, which fatally delayed their humanitarian response.

The food crisis in Niger in 2005 seems like a classic food production failure. Drought and locusts destroyed food production across the West African Sahel. Niger’s grain harvest was only 11% below the five-year average, but this decline was not evenly distributed across the population, and 13,000–47,000 people died in the worst affected rural communities. On the other hand, similar crop failures were recorded in neighboring countries, but only Niger experienced a famine, so once again food availability decline explains only the trigger, not the entire famine process. Rapidly rising food prices – staple cereals increased by 200–300% between the harvest and the peak famine months – were mirrored by rising malnutrition rates. Smallholder farmers, pastoralists, and farm workers all adopted coping strategies which included selling their assets for food. The livestock/grain terms of trade collapsed to one-quarter of pre-drought levels.

The Niger government was aware of the evolving food crisis, but its appeals for aid were initially ignored. Unlike in neighboring Mali, where a famine was averted through distributions of free food to drought-affected families, the government of Niger chose a market-led response, by selling grain at subsidized prices. The intention was to minimize “dependency syndrome” and market destabilization, but prices were too high for people who were already destitute, and the subsidy was largely captured by nontarget population groups. Niger’s development partners endorsed this market-led approach. The International Monetary Fund enforced tax hikes on food staples as a condition for budget support, in the famine year.

The famine in Somalia in 2011 was also triggered by a severe drought – the lowest rainfall in 50 years in the 2010 deyr season – that caused harvest failures and livestock deaths and reduced employment opportunities in the crop farming and agropastoral southern regions. The cereal harvest in 2011 was just one-third of the 2006–2010 average. The drought was accompanied by a dramatic rise in food prices, partly because of reduced food supplies in local markets and partly because of global food price rises at the time. These factors reduced both availability of food and access to food for farmers, agropastoralists, agricultural laborers, and internally displaced persons (IDPs).

The main reason why the food crisis in southern Somalia became a famine was protracted conflict and civil insecurity, which had displaced more than a million people by 2009. The area was controlled by the militant group Al-Shabaab, which was declared a terrorist organization by the United States government in 2008, prompting USAID to cut off food aid and the World Food Programme to withdraw from southern Somalia. Al-Shabaab also forced CARE to pull out in 2008, leaving the local population with no access to humanitarian relief when harvests failed in 2010. The politicization of aid by both the donor community and Al-Shabaab was responsible for a response failure that persisted until the famine was officially declared in July 2011. Even though donor pledges escalated immediately, the relief operation was too late to prevent an estimated 258,000 deaths, and was further compromised when aid workers were killed and Al-Shabaab closed down several programs run by nongovernmental organizations.

Future famines

Conflict is likely to continue being a driver of most famines that occur in the future. The 2011 famine in Somalia is the most recent example of a “complex emergency” famine, where food production was first disrupted by drought, then
trade and aid responses were undermined by violence and insecurity. The combination of a weak national state, the obstructive practices of Al-Shabaab, and international donors that withdrew under Al-Shabaab pressure and United States counterterrorism laws that effectively criminalized aid flows into southern Somalia, all contributed to a failure of humanitarian response, despite adequate early warning information. Future famines that are triggered or compounded by political instability and civil insecurity will be exacerbated by similar “no go” geographical spaces that prevent food aid from reaching starving people. In such situations, airborne food drops might by necessity become a more common way of delivering humanitarian relief than truck convoys.

Weather shocks will continue to trigger harvest failures for as long as farmers practice rain-fed agriculture, and this trend could be exacerbated by climate change in fragile agroecologies. In poor agrarian communities where markets are weak, food prices will spike and asset prices and wage rates will collapse as families sell off their possessions to buy food. So food availability will fall for vulnerable groups, and their access to food will be compromised. However, whether these episodes progress towards full-blown famine will depend crucially on the response of governments and the international community – whether aid is mobilized and delivered promptly – which in turn will depend on the nature of relations between governments and citizens and between national governments and their development partners.

Contemporary famines are preventable, and most recent famines could have been prevented. In all four twenty-first-century famines to date, it is the failure of the humanitarian response to production and market failures, by governments and agencies, that has tipped affected communities over the edge from food crisis to famine. Unless and until accountability for allowing famines to occur is enforced, it seems likely that these avoidable tragedies will recur in future, not because of natural disasters but because of politics.

SEE ALSO: Environment and development; Food security; Livelihoods; Natural hazards and disasters; Scarcity; Social vulnerability and environmental hazards

References


Further reading


**Feminist geography**

Robina Mohammad  
*National University of Singapore*

Feminist geography is plural, heterogeneous, and fragmented, intersecting with many fields within and across geography, the social sciences, and humanities. Feminist geographers are thus located within multiple intersecting webs of knowledges that connect geographers not only to feminist scholarship within the academy but also beyond it, such as to those active within the new social movements and nongovernmental agencies. There are, therefore, many feminist geographies and many ways to document such a historiography.

Feminist geographies bring feminist concerns together with a geographical focus. They can be defined as those “Perspectives that draw on feminist politics and theories to explore how gender relations and geographies are mutually structured and transformed” (Pratt 2000, 259). Feminist geography only started to emerge as a field of geographical inquiry from the early 1980s. In the decades since then, it has grown to become much more than a subfield of the discipline, but rather permeates every aspect of human geography and, significantly, mounts a critique of the geographical tradition. Feminism’s impact on the discipline of geography, as with other academic disciplines, was supported by post-World War II social movements. The women’s movement and the Civil Rights Movement, in particular, brought a focus to different structures of inequality and how they intersect to shape experiences of power and social exclusions, setting the sociopolitical context in which feminism was put back on the political agenda in North America and Europe, informing activism on the ground, as well as extending its influence to the academy, particularly to sociology and anthropology and more recently to geographical scholarship. Dialogues between feminism and geography, and feminist geographies and post-modernism/poststructuralism, have undergone three major shifts: first, from geographies of women to feminist geographies (it may also be viewed as a move from description to grand theory); second, from a focus on equality (between men and women) to a concern with differences and diversity between women, examining the ways that gender intersects with other forms of social oppressions; and, third, an epistemological shift that is sensitive to plays of power.

**Conditions of possibility**

Feminism is a set of philosophical perspectives as well as a social movement that gathered pace in different places and at different times, shaped by national, socioeconomic, and political contexts and serving particular political agendas. The early twentieth century had witnessed women’s militancy against political disenfranchisement, leaving the issue of wider gender inequalities unchallenged.

In some places, feminism was a top-down phenomenon, connecting to struggles for independence (colonial India, colonial Egypt, and post-1923 Turkey) and/or struggles for democracy (Franquista Spain, 1939–1975). The “woman question,” the role of women, and their rights and obligations were linked to issues
of governance and development. The feminist movement that resurgence in the Anglo-American world in the post–World War II period, referred to somewhat problematically at times as the “second wave” of feminism (Nicholson 2013), was a grassroots, albeit middle-class, phenomenon that has challenged the core basis of gender differences that translate into inequalities.

In postwar North America liberal feminists (of whom one of the most visible was Betty Friedan) pointed to the home and the private sphere as the basis of women’s social marginality. Women’s wartime, large-scale, cross-class entry into the public sphere, as part of their role as a reserve army of labor, and the opportunities it brought to enter occupations that were historically viewed as men’s work, served to some extent to destabilize the gender division of labor and the ideology of separate spheres. It was not until after World War II that feminists began to formulate critiques of the naturalization of the spatialized gender divisions of labor. Part of the feminist critique was the insistence that the “personal is the political,” which challenged the idea of politics as formal governance in the public sphere and, by the same corollary, of the private sphere as a domain outside of politics. Feminists asserted that intimate relations in the private sphere connecting with fertility, reproductive rights, and heterosexual coupling were also underpinned by power, defined as the “arrangements whereby one group of persons is controlled by another” (Millet 1970, 23).

The sex/gender distinction is key to contesting the ideology of biological determinism, which holds that biology determines women’s place in the home. This binary distinguishes natural and given attributes of sex, that is, biological attributes and the social meanings ascribed to them (Rubin 1975). As French feminist Simone de Beauvoir (1989/1949, 267) pointed out, “One is not born but rather becomes a woman.” For de Beauvoir, the meaning of “woman” is not innate. It does not have an inherent nature; rather, it is socially constituted in relation to the meaning of man. By separating the natural and the social, feminists could challenge the givenness of the meaning of the feminine. If women’s position in society is not natural but rather socially produced, then it can be changed. Yet as Donna Haraway (1991) points out, while feminists mounted fervent critiques of binary oppositions such as culture/nature they were reluctant to undo the sex/gender binary because of its centrality in confronting the pervasive biological determinism that formed the basis of women’s inequality and disadvantage. Thus feminist scholarship of the 1960s and 1970s made no attempts to bring the category of nature, whether external or in terms of biology/sex, into culture by historicizing it, which meant that the categories of woman and man continued to be treated as essential. In this way the category of nature continued to threaten the notion of gender through slippage as sexual difference. As Haraway (1991, 28) explains:

In the political and epistemological effort to remove women from the category of nature and to place them in culture as constructed and self-constructing social subjects in history, the concept of gender has tended to be quarantined from the infections of biological sex.

Biology continued to be treated by feminist scholars as a study of the natural world rather than a social discourse through which the category of nature becomes constituted. It was not until the 1990s that deconstruction was applied to the sex/gender dualism by the post-structuralist critiques of scholars in the field of queer and transgender studies. This was a necessary starting point for developing theories of heteronormativity, but it served to trouble the category of woman in the process (Butler 1999).
Butler (1999, xv) and others argued that the heterosexual matrix is a prerequisite for gender. In this view gender becomes a fiction that is produced through performativity, so

that what we take to be an internal essence of gender is manufactured through a sustained set of acts, posited through gendered stylization of the body … [so that] what we take to be an “internal” feature of ourselves is one that we anticipate and produce through certain bodily acts.

Challenges to the category of “woman” were a source of concern for feminists whose project depended on her emancipation. The impact of postmodernism and poststructuralism on theories of subjectivity, identity, conceptions of power, and, as part of this, the deconstruction of dualisms would bring about epistemological shifts in feminist scholarship but not without great anxiety.

Spatiality was implicit in Anglo-American feminist critiques and countercritiques of the gender order and hierarchy, but it was not subject to explicit analysis. It would be almost a decade before geographers began to turn their attention systematically to the study of gender. The “radical,” or perhaps more appropriately “critical,” turn in the discipline saw geographers such as David Harvey turn to Marx for the analytical tools to examine the role of power in sociospatial relations and in structural inequalities. By the late 1970s the Palestinian American scholar Edward Said’s influential work on “Orientalism,” combined with the burgeoning field of queer studies, opened new directions for geographical inquiry to focus on forms of marginality that were not directly related to class, setting new agendas for feminist geography.

In the historiography of geography, feminist geography as a distinct perspective is narrated as emerging much later than in the other key social science disciplines, including sociology. In the United States during the late 1960s, the gains made by social movements began to affect academic disciplines with the implementation of antidiscrimination legislation targeting gender as well as racial discrimination (Rose 1993). Policies of affirmative action saw the establishment of quotas for hiring underrepresented, minority populations underlining their lack of presence and invisibility within the academy. Their absence was particularly notable in geography, historically a masculinist discipline linked to imperial practices (Rose 1993). The androcentrism of geography had been left unchallenged by the critical turn in geography, in which the gender blindness of spatial science, which privileged a disembodied, rational, economic man, was emulated by gender-blind Marxist and humanistic geographies. An obvious example was the taken-for-grantedness of the use of “man” to refer to all of humanity. Courses with titles such as “man and the environment,” “man and nature,” or even “man and culture” could still be found on the geography curricula (WGSG 1984). Rose (1993, 2) points out that masculinism in geography is not restricted to teaching and research but can be found in the everyday forums of disciplinary spaces, and that it delineates its borders:

[It] can be seen at work not only in the choice of topic made by geographers, not only in their conceptual apparatus, not only in their epistemological claim to exhaustive knowledge but also in seminars, in conferences, in common rooms, in job interviews.

Masculinism was accentuated, and some would argue supported, by the paucity of women faculty and graduate students. Some departments, such as the one at Clark University, accepted higher numbers of women for postgraduate studies but those graduating struggled to find jobs in the discipline. The pioneering feminist geographer and former president of the
American Association of Geographers (AAG) Susan Hanson has narrated in a session at the AAG annual meeting in Boston (2010), and documented in a special issue of Gender Place & Culture, as well as in other disciplinary sites, how both she and her husband, Perry, graduated from Northwestern at the same time (1973), yet, while he gained employment immediately, Susan was not even able to get an interview despite sending out many applications for positions in geography. That the geographer was normatively male and heterosexual was underlined to her again when attending the AAG’s annual meeting in the early 1970s where women geographers would routinely get mistaken for “wives” of the geography delegates and were expected to attend activities laid on especially for ladies. Women’s lack of disciplinary presence is confirmed in quantitative terms by examining the gender of the authors of geographical scholarship, the membership of scholarly societies (some of whom had, historically, formally excluded women), and faculty of institutions of higher education. Gillian Rose (1993) shows how, until the 1970s, the number of papers published by women scholars in the American flagship disciplinary journal Annals of the Association of American Geographers was just over 2%; similarly, women made up only a little above 10% of the membership of the AAG, something that was also reflected in their relative absence from American higher education institutions as teachers of geography.

Yet a focus on the numbers of women within the discipline neglects those women located outside of it by particular constructions of the geographical tradition, thereby legitimating certain kinds of knowledge and knowing subjects who were not feminine. At the same time Rose (1993, 4) suggests that the subject position of those women working within the discipline, such as Monica Cole and Alice Coleman, was inherently masculine, for to do geography, to think within the parameters of the discipline in order to create knowledges acceptable to the discipline, was to occupy a masculine subject position. Women’s presence within the discipline certainly says little about the presence or absence of a feminist perspective since being a woman geographer does not necessarily equate with being a feminist. Early women geographers mostly worked within the androcentric parameters of the discipline and made few attempts to contest, resist, or rework these as feminist geographers aimed to do.

Feminist voices began to emerge during the 1970s and would gather ferment by the end of the decade. Prominent were those of Susan Hanson and Sophie Bowlby, both of whom graduated from Northwestern University and gained employment in geography departments during the early 1970s. They were practitioners of a feminist geography that maintained a dual focus: to attend to the geographies of gender and to the gender of geography. The aim was to rectify the neglect of women as subjects capable of knowing and producing geographical knowledge by promoting equal opportunities for women as practitioners. The absence of women was regarded as one of the main reasons why feminism was so late in having an impact on geography. The presence of women was viewed as key to bringing feminism and geography into a productive dialogue for the transformation of geographical scholarship and teaching. Another aim was to rectify the neglect of women as the objects of geographical research by advocating a woman-centric focus, documenting and explaining women’s experiences of inequality from their own perspectives.

Feminist research remained situated within the disciplinary paradigmatic context as well as influenced by the wider social and political milieu within and beyond the academy. In keeping with this, feminist geography at the time was largely descriptive and positivist in
Feminist geography. The implicit assumption that feminist geographies were geographies of women seemed to be validated by the work that was being done. “Geographies of women” drew on the prevailing approaches of humanist, Marxist, or socialist and welfare geographies, extending and developing these toward a focus on gender. In this way they brought attention to areas that had hitherto not been considered worthy of attention by mainstream or “malestream” geography to animate the silences around women’s life experiences, social marginality, and disadvantage vis-à-vis men. Marxist approaches to feminist geographies were also informed by the theories of socialist feminists who interrogated the gender blindness of Marxism to examine the ways in which patriarchy intersected with capitalism to impact on women’s social position. Feminist geographies examined women’s socioeconomic position within capitalist societies through a focus on the spaces of women’s paid and unpaid labor and its social and economic contribution. Other work examined the historical separation of production and reproduction, and the emergence of a spatialized gender division of labor and its inscription into the urban form (Bowlby et al. 1989).

Welfare and neo-Weberian approaches, by contrast, focused on the spatial, material, and ideological constraints experienced by women in their gender roles as mothers and wives. Jackie Tivers drew on Hägerstrand’s time–space model to examine the impact of women’s mothering and domestic roles on their spatiotemporal mobility and how their activities might affect their access to paid work (Bowlby et al. 1989). A humanist approach explored women’s perceptions of space, place, and environment through a focus on their experiences of social and physical landscapes, to demonstrate how women’s experiences differed from those of men. This work tended to valorize women’s difference, constructed in terms of their sensitivity and perceptiveness (Bowlby et al. 1989). This risked not only essentializing women’s difference but also reaffirming the male norm, from which women were seen to deviate. Similarly, by documenting and describing the lives of women as they were being lived – as they performed activities such as domestic labor, provision of care, and nurturing children in spaces culturally associated with femininities – feminist geographers risked reinforcing the naturalness of women’s place and role. In addition, there was a notable lack of attention paid by feminist geographers to the variations of women’s lives across places, either intra- or internationally, which is surprising given that areal differentiation was a significant paradigm within the discipline (Bowlby et al. 1989).

The move from documentation and description toward analysis and explanation led to a growing recognition that the notion of gender roles was a conceptual cul-de-sac. Gender roles could not offer an account of how change might occur; how, for example, they might be negotiated or resisted; and how their meaning might change through such practices. In this way they seemed once again to lock women into particular positions, just as the focus on sexual difference had done earlier. By contrast, the notion of gender relations contained explanatory potential, drawing attention to the social structures through which gendered patterns of behavior, inequality, and disadvantage were constituted.

While feminist geographers were grappling with questions of how to approach the geographies of women/gender and to address the masculinism within the discipline, the importance of feminist approaches to geography was affirmed with the establishment in 1980 of the Women and Geography Study Group (WGSG) within the Institute of British Geographers (IBG). The IBG was a professional body for critical geographers maintaining, at the time, a
Feminist Geography

distinction from the Royal Geographical Society, which was associated by many with geography’s complicity with imperialism. The WGSG’s objectives were:

- to encourage the study of the geographical implications of gender differentiation in society and geographical research from a feminist perspective;
- to encourage and facilitate the exchange of information and ideas with reference to research and teaching in these areas, and to disseminate information through its publications.

These aims reflect a shift away from the “geographies of women” approach toward feminist geographies. As the WGSG (1984, 20) explains, a feminist perspective requires more than just to “increase … the number of studies of women per se in geography but an entirely different approach to geography as a whole.” Hanson (1992, 570) elaborates this point in her 1992 presidential address to the AAG: “feminism looks at the world through the lens of gender (note: not women or femaleness) while seeking to build a world in which gender is no longer a key dimension along which life’s possibilities are defined and resources are allocated.” This shift in focus and feminist geography’s stronger institutional presence were confirmed in 1994 with the establishment of a new interdisciplinary feminist journal *Gender, Place & Culture* (which recently celebrated 20 years), creating a dedicated (inter)disciplinary space for feminist scholarship.

Spatiality and gender relations

This shift did not mean that “geographies of women” disappeared entirely, but rather it began to mutate into a feminist geography that could take forward the achievements of the earlier focus on women to bring about a broader focus on gender. By the late 1980s, as the United Kingdom and other Western economies experienced huge economic restructuring, the issue of gender could not be ignored by Marxist and neo-Marxist economic geographers studying industrial change and uneven spatial development that normatively centered analyses on class relations. Their attention to gender was forced by the changing gender composition of the labor market in regional, local, and urban economies, with the simultaneous growth of women’s presence in paid work and men’s unemployment. Yet it was feminist geographers who examined more fully the relationship between the economy and patterns of gender relations to demonstrate how different regional economies produced distinct gender relations and divisions of labor. Moreover they suggested that historical patterns of gender relations and divisions of labor shaped the changing composition of regional economies. These kinds of studies foregrounding the scale of the locality allowed for a broader focus on gender relations than a narrower labor market focus.

Much of the feminist scholarship of the 1970s sought to explain women’s oppression by theorizing sex/gender systems through reference to Marxism in combination with psychoanalytic theories. The gender division of labor and the psychology of desire were posited as translating into a compulsory (Rich 1980) or obligatory heterosexuality (Rubin 1980; Wittig 2013) that held the key to women’s subordination by men. Yet Bowlby et al. (1989) note how feminist geographers were not examining the source of women’s subordination, which was implicitly or explicitly understood as emerging from a capitalist social system that was inherently patriarchal such that it could be referred to as a capitalist patriarchy. In an early exception to this assumption, Foord and Gregson drew on a realist approach to
Feminist geography examine the conditions for the development of patriarchy. They supported the idea of a dual system of capitalism and patriarchy each with its own distinction. This position was rejected by McDowell, who argued that patriarchy is inherent to capitalism and serves to resolve its internal contradictions (Bowlby et al. 1989).

By the early 1990s the critiques of white, Western feminism taking place in other disciplines and outside the academy began to filter into geography, articulating with postcolonial critiques from literary theorists, such as Edward Said, to draw attention to an issue that hitherto had been conspicuous as an absent presence: “race” and difference. Registering difference and diversity in turn opened up epistemological questions as well as ethical concerns, bringing under critical scrutiny the power relations and exclusions marking research and representations.

From the 1970s onward, the women’s movement had become increasingly divided by feminisms of different persuasions, with multiracial feminism also beginning to gather force alongside its radical, socialist, difference, liberal strands (McCann and Kim 2013). Despite the diversity of these ideas, primacy was still given to the division between men and women. Feminist geographers also neglected attention to other dimensions of social exclusion such as race/ethnicity, postcoloniality, and nonheteronormativity. Lesbian feminists’ 1970s critiques of feminist privileging of heteronormativity were followed in the 1980s by African American and black British feminists’ voices in powerful critiques of white, Anglo-American feminism with charges of a totalizing universalism, the same charge that white feminists had earlier made to androcentric knowledges. Women of color from both sides of the Atlantic joined voices to call into question universalizing concepts such as patriarchy and global sisterhood, pointing out that a feminist theory that does not take account of race cannot fully explain their experiences of oppression. To substantiate their point, they drew attention to the idea that it is not only black women who are oppressed but, in a world of white, capitalist supremacy, black men are also subordinated. In this way they problematized feminism’s goal of equality with men. Hooks (2000, 19) posits the question, since all men are not equal, “which men do women want to be equal to?” She uses the term “white supremacist capitalist patriarchy” to define the experience of multiple interlocking forms of oppression. The experiences of oppression of women of color highlighted the ethnocentrism of the sex/gender division which constructed the body as located in nature rather than culture, a body that is not only gendered but also raced, classed, and sexed. Hooks (2000) notes how a text that is considered as a founding text of the women’s movement, Betty Friedan’s Feminist Mystique, remains silent on the experiences of women who are the most oppressed. These are the women who have toiled in low-paid jobs outside of the home and are not fortunate enough to have the luxury of a gender division of labor that locates them in the home: the sphere of social reproduction. Women of color argued that the home has different meanings for them: in alliances with men of color, the home space can be a site of resistance and solidarity as well as a refuge from racial violence in the world outside. These voices both supported and were supported by the epistemological shift from modernism to postmodernism and its emphasis on difference, plurality, and multiplicity. As Linda McDowell (1993, 307) notes in her reflections on the trajectory of Anglo-American feminist geography, there has been a shift in focus from “equality towards that of difference.”

The 1990 meeting of the WGSG at University College London is insightful about feminist geographers’ concerns about, and responses to, postmodernism and difference. One of the three
issues for discussion at this meeting was postmodernism’s relationship to feminisms and the implications of difference for feminist politics. Informal presentations by Linda McDowell, Eleanore Kofman, and Gillian Rose explored whether postmodernism posed a challenge to feminist geographies or if it was complementary, particularly around the issue of difference and diversity. These presentations were subsequently summarized in *Antipode* (1992). Jan Penrose (*Antipode* 1992, 220) reflects on how “support for postmodernism – the theoretical framework that purports to celebrate difference … has become one of the most contentious and divisive issues within the WGSG.” Yet Monk, McDowell, and Kofman insist that there is nothing new for feminists about the concepts identified with postmodernism. While feminists were slow to recognize diversity among women as a group, they had been at the forefront in highlighting the processes of subjectification, the recognition of difference, polyvocality, plurality, and deconstruction of dualisms. Moreover, they were also the first to critique “classic modernist notions of rationality, equity, fairness, citizenship” (*Antipode* 1992, 221). Thus Penrose (*Antipode* 1992, 220) suggests that feminists’ contentious attitude toward postmodernism arises from the “new paradigm’s usurpation of long standing feminist tenets without any acknowledgement of its debt to feminist thought and practice.”

McDowell echoes wider concerns of feminists when she points to a number of problems with the postmodernist paradigm. First, although postmodernism recognizes and even celebrates the plurality of difference, such differences continue to be located in a hierarchy, constructed in relation to a norm that is reaffirmed by the presence of difference. Second, the emphasis on diversity serves to relativize gender, making it just one among many significant differences. This claim is refuted by feminist assertions that gender “is not one relevant strand but our uniting perspective” (McDowell in *Antipode* 1992, 226). Third, in addition to the category of “woman” being fractured, it is under threat of annihilation by the dissolution of the subject (for a detailed discussion see Butler 1999) – a source of great feminist anxiety because, as Pratt (1993, 56) in a paper published in the same journal two years later pointed out, “the dangers to feminism are clear enough: feminism threatens to self-destruct as feminists deconstruct its central analytical category.” Finally, postmodernism’s abandonment of the notion of a singular truth for multiple truths was said to risk the relativization of truth. Kofman (*Antipode* 1992, 229), who refers to postmodernism as a “male-created theorization,” reiterates wider feminist critiques when she suggests that, just when the truth installed by white Western men was being called into question by women and other “others,” postmodernism asserts that there is no truth to be discovered.

**Doing feminist geography: difference, polyvocality, and situating knowledges**

Assertions of the importance of difference and diversity among women, from feminists across a range of academic fields, gained further ground through articulation with the postmodern/cultural turn’s refusal of grand narratives and universality that also foregrounded difference, plurality, and multiplicity. Moreover, poststructuralism’s relativization of truth worked together with the decentering and dissolution of the subject, the destabilization of social categories and meaning, and modernist certainties to create epistemological concerns, anxieties, and ambivalence for feminist geographers. These concerns and anxieties prompted a rich debate on the nature of feminist epistemologies and
ontologies and the politics of fieldwork praxis, representation, and authorial claims (Nast 1994).

It is worth reminding ourselves that the focus on difference, diversity, and polyvocality pertains in large part to the demands of previously excluded groups to be heard and to have their claims accepted as authentic and legitimate. Feminists’ voices, alongside the voices of nonwhite, non-Western, postcolonial subjects, pointed to the nexus of power–knowledge in their critiques of androcentric, ethnocentric, rational, Western knowledges and their claims to universality.

Neopositivism, a system that served as a guarantor of the validity of hegemonic knowledges, was vociferously critiqued. Neopositivist ontologies expressed in the aims for discovering universal laws pertaining to the social world were abandoned and epistemological claims of objectivity, value freedom, neutrality, and replicability were posited as myths that merely served to cloak the biased, partial, and contingent nature of all knowledges. These critiques were instrumental in shaping the contours, content, and goals of feminist methodologies with the aim of producing knowledges supporting the emancipation of women and the researched through an emphasis “on the mutuality of the research process; intersubjectivity not objectivity and dialogues in place of monologues” (Gilbert 1994, 90).

A central plank of feminist epistemologies was an acknowledgment of the limits of seeing as the basis of the capacity to know. Feminist geographers asserted that – by contrast to the false universalities propped up by neopositivism’s attempts at a god trick, a view from nowhere, a disembodied, omnipresent vision – true objectivity lies in the recognition that all knowledges are embodied, situated, contextual, and therefore partial. In more recent years critical feminist modes of reading the visual, as well as feminist visualizations as method, have been advocated.

This has included an exploration of the possibility of recuperating geographic information systems (a military technology) for visualizing qualitative data.

Situating knowledges as a practice translated into a self–reflexive approach that would make transparent researcher biases and the processes by which knowledges become enframed and staged, making visible the relationship between the researcher and researched (Gilbert 1994) as part of making the researcher accountable by acknowledging power relations involved and the exclusions that their account rested on. In keeping with wider debates among feminists in other disciplines, feminist geographers also pointed to the distinctiveness of women’s sociospatial experiences from those of men and their capacity to know. The focus on difference and the politics of representation opened up two key questions: Who can speak for whom? With what authority? For example, on the question of speaking for others, England (1994, 81) asks whether, “if in a rush to be more inclusive and conceptualize difference and diversity might we be guilty of appropriating the voice of others.”

In resolving these questions, greater emphasis was placed on the researcher’s social location, positionality, and personal biography (see Mohammad 2001). These became regarded as key, in the academic community as well as in the field, to establishing the researcher’s capacity to know, the validity of his/her accounts, as well as his/her right to study particular groups. Identity on the basis of structurally similar social positioning or sameness of color/culture/gender and/or sexuality was seen as conferring an “insider” status, translating into a greater capacity to know and to understand the researched than that of those who attempt to study a group from the position of an “outsider” (England 1994). There was also an implicit acceptance that the position of an “insider” ensured greater equality between
Feminist geography

researcher and those researched (see England 1994).

In contrast to the privilege enjoyed by white, middle-class, European masculinities as knowing subjects, marginality was seen as conferring a moral authority and a privileged way of knowing. This recognition led to the emergence of a form of identity politics in which differences were placed in a hierarchy of oppressions ... [and] multiple oppressions came to be regarded not in terms of their patterns of articulations/interconnections— but rather as separate elements that could be added in a linear fashion, so that the more oppressions a woman could list the greater her claims to occupy a higher moral ground. (Brah 1992, 135–136)

The additive approach to identity critiqued by Brah was also problematized for suggesting that those who were marginally located along two or three axes of identity suffered double or triple oppression, neglecting the ways in which multiple marginality qualitatively changes the experience of oppression.

Feminist geographers have drawn on the concept of betweenness to deconstruct the insider/outsider dualism. In practical terms it offers a means of negotiating difference in the field. Nast (1994, 57) elaborates on how betweenness highlights the fact that we can never not work with “others” who are separate and different from ourselves; difference is an essential aspect of all social interactions that requires that we are always everywhere in between or negotiating the worlds of me and not-me.

This point also underlines the recognition that identity is fluid, relational, and constituted through a range of social narratives so that one can also be simultaneously an insider and outsider (Nast 1994; Miraftab 2004).

It is notable that, in contrast to neopositivist knowledges, many feminist geographers’ situated and reflexive accounts of the politics of fieldwork point to myriad complexities entailed in their embodied presence, bringing them into confrontation with different kinds of differences. These accounts are frequently saturated with emotions, concern, fear, guilt, and anxiety at possible or potential unequal power relations and exploitation of the researched. These feelings risk reading the relationship with the researched in binary terms: the all-powerful, active researcher presiding over an absolutely powerless and passive researched, negating the agency of the latter. Yet Miraftab (2004) draws on her experiences in the field to foreground the dialogic nature of interactions through which both the researcher and the researched encounter, read, and position each other. She points out how, despite the researcher’s often heightened sense of his/her own privilege, he/she is open to being positioned in very different ways from that imagined by the researcher, including being positioned as impoverished and even oppressed on different axes of identity.

Despite the best intentions of feminists, critiques of feminist methodologies point to the ways in which the emphasis on closeness, trust, and sharing in feminist methods can often carry greater potential for exploitation and even betrayal of the researched, who may come to see the researcher as a friend.

Some ways in which such feminist agendas of the politics of research and representation have been taken forward since the 1990s include participatory action research, in which the research is driven by those who have a stake in the project rather than by an outside researcher. It is democratic, collaborative, and designed to achieve some form of action that benefits the participants. The ethical turn during the 2000s has also configured the politics of research through the institutionalization of ethical frameworks within...
universities and funding bodies across Europe, North America, and increasingly beyond. All postgraduate, and in some institutions even undergraduate, research must have ethical approval prior to commencement; thus ethics committees have become gatekeepers, enabling, blocking, and modifying research projects. Yet feminist scholars continue to reaffirm the importance of self-reflexivity, positionality, and critical attention to the circulation of power in the field in achieving ethics objectives. At the same time, postcolonial feminist geographers have argued that sensitivity to the politics of representation has encouraged an “impasse” in feminist geography now, where fears of (mis)representation and (in)authenticity have led to a general withdrawal from fieldwork in the Global South, which means that fewer scholars are engaged in research that can be politically and materially useful for the poor in the Global South. (Sultana 2007, 375)

They also suggest that ethical guidelines of universities in the Global North may not be adequate for research settings in the Global South. There is a need to pay attention to the greater complexities involved in the research process introduced by “The complex movements of footloose researchers’ in this era of globalization” (Miraftab 2004, 597) and the impact on fieldwork dynamics.

Feminist geographies of the new millennium reflect the theoretical and substantive impacts of the cultural turn, articulating with the wider social context of deindustrialization, intensification of globalization and mobility, and important geopolitical shifts. Three emergent areas of study that have become established during this time are the focus on the body as part of the corporeal turn; geographies of masculinities; and Muslim geographies of gender as part of the new subfield of geographies of religion. This broader focus of feminist geography is reflected in the changing of the name of the Women and Geography Study Group in 2012 to Gender and Feminist Geographies Research Group.

In its 40 years of existence feminist geography has negotiated a range of institutional as well as epistemological challenges to become a significant institutional presence, transforming the discipline from within in terms of ontology, epistemology, and methodologies, constantly broadening and extending its focus in new directions.

**SEE ALSO:** Bodies and embodiment; Citizenship; Cultural politics; Cultural turn; Difference; Discourse; Domestic spaces; Domestic workers; Feminist methodologies; Gender; Gender and development; Gender, work, and employment; Habitus; Identity; Life course; Modernity; Patriarchy; Postcolonial geographies; Postmodernity; Representation; Sexualities; Social capital; Social reproduction; Subaltern; Visuality; Work–life balance

**References**


**FEMINIST GEOGRAPHY**


Feminist methodologies

Linda J. Peake
York University, Canada

There is genuine confusion and misunderstanding about the term “methodologies” among and beyond feminists. A search of the term yields definitions that range through vague (“general research strategy,” “the design process for carrying out research”) to inadequate (“the study of methods”), the latter equating methodology with techniques for data collection. While the term is mercurial, it still requires definition and is taken here to refer to the theoretical underpinning of the practices, procedures, and rules used by feminists to conduct research. As such it accounts for many aspects of how research is conducted, but it also explains “the literature the researcher is using, the language and terminology, the other theories and explanations being used, the methods and the type of analysis that will be used to interpret the data and information collected” (Rangahau n.d.), as well as the formulation of research questions. And yet, despite its central importance to conducting research, to producing new practices of knowledge production, and to linking the intellectual and political, feminist geographers, surprisingly, have had much less to say on the geographical tradition of fieldwork than feminist anthropologists (see Wolf 1996), the last major intervention being in *The Professional Geographer* in 1994 and the most recent a 2016 theme issue published by *ACME*. Feminist geographers, however, have made significant interventions into methodological debates in other parts of geography, in particular, the development of visual methodologies (Rose 2016), critical geographic information science (GIScience), and indigenous methodologies.

Understandings of what is feminist about methodology are contested and have changed over time, paralleling the struggles of second and third wave feminisms. In the early 1970s, the political goals of the women’s movement in countries of the Global North were central to feminist research and this was when feminist scholars first started to challenge processes of knowledge production. In particular, the dominant and androcentric positivist scientific method was critiqued for its inability to address the experiences of women (Harding 1987). Work by feminists of color (Baca Zinn 1979; Hill Collins 1990) served as a corrective to the ethnocentricity of these critiques and brought the experiences of women of color into knowledge production. Poststructural feminists’ questioning of the unified subject of women 2002), whereas in sociology, for example, feminists have been publishing extensively on issues of methodologies and pioneering interventions into feminist methods since the early 1980s (see, e.g., Annie Oakley’s (1981) seminal piece on feminist interviewing; Dorothy Smith’s (1987) work on institutional ethnographies).
and postcolonial feminists’ broadening of the remit of feminist knowledge production, to encompass not only women but all marginalized groups, further expanded the grounds of feminist methodologies. By the 1990s, an analytical approach of Intersectionality (Crenshaw 1993), which highlights the relational connectedness of race, class, gender, and sexuality in examining women’s lives, was widely adopted, but in becoming increasingly divorced from its roots in explaining the oppression of African American women, intersectionality is coming under criticism. And the important ethical and political questions raised by antiracist and indigenous perspectives on decolonizing methodologies (Smith 2012) have served to highlight questions around the fixity of power and the limits of feminist critiques of knowledge production from the predominantly white Western academy. Interrogating the feminist grounds of knowledge production thus depends on the analytical and political positioning of feminists, which currently range through liberal groups, mostly in the Global North, with interests in women’s positioning in relation to men, to radical groups, both in the Global South and North, whose interests in women form part of a broader concern with community and communities, to those whose focus on gender cannot be separated from interests in colonialism, race, and/or sexuality.

The particular contributions made by feminist geographers to debates on feminist methodologies relate to the intersections of space, power, and knowledge. They can be seen as fourfold: spatializing existing feminist approaches to knowledge production; new spatialized contributions to feminist methodologies; engagements with methodological practices that speak to the valorizing of epistemology over praxis; and debates on methods, particularly in relation to quantitative versus qualitative approaches.

Spatializing existing feminist approaches to knowledge production

Inserting notions of spatialities into work on the production of knowledge and of subjectivities, feminist geographers have drawn especially on the work of two feminist scholars, Sandra Harding and Donna Haraway. A feminist philosopher, Sandra Harding’s interest in how relations of power shape the production of knowledge in different contexts led her to argue for a self-reflexive and collective approach to theorizing in which the viewpoints of all those involved in knowledge production processes must be acknowledged (Harding 1987). In this call, she went to the heart of a critique of hegemonic Western Enlightenment knowledge, built on the supposed bedrock of complete “objectivity,” that is, that the application of objective rules of science leads to the “truth.” She came to call her approach “strong objectivity,” in contrast to the unreflective positivist approach based on the distanced and top-down viewpoints of only a few so-called experts. Moreover, Harding emphasizes a relationship that proponents of the scientific method either see as irrelevant or seek to neutralize, that is, the centrality to strong objectivity of the relationship between the subject and object of inquiry, which highlights the importance of subjectivity. Harding (2004) also supported the epistemological stance of standpoint theory, in which Marxist feminist standpoint theorists, such as Nancy Hartsock, argue that women can produce more complete and less distorted knowledge than men because of their sex–class position. The universalism presumed by feminist standpoint theory was rejected by the poststructural feminist scholar Donna Haraway, however, who argues that neither women nor men can ever have total knowledge. She claims that the understanding of complete objectivity should be redefined and replaced by
that of “situated knowledge,” which acknowledges not only the partial and situated nature of all knowledge production but also its embodied nature, grounded in real bodies (Haraway 1988).

Understandings of all knowledge as situated, embodied, and geographically placed have come to characterize feminist geographic knowledge production. Gillian Rose in her landmark book *Feminism and Geography: The Limits of Geographical Knowledge* (1993) was arguably the first to develop a specific feminist critique of the “transparent space” of masculinist geographic knowledge production in which claims to know, divorced from experiential and emotional knowledges, transcend the specificities of the body and are equated with (violent) claims to space and territory. In contesting both what counts as legitimate geographic knowledge production and the masculinism of geographic discourses, Rose argues that “various forms of white, bourgeois, heterosexual masculinity have structured the way in which geography as a discipline claims to know space, place and landscape” (Rose 1993, 137). Drawing on the work of other feminists, her aim is to challenge the epistemic violence of the masculinist geographical desire to know by exploring feminist constructed spatialities that acknowledge and allow for difference. Utilizing the work of Teresa de Lauretis, who draws on the spatial imagery of “elsewhere” to describe “resistance to hegemonic identities of the subject of feminism” (Rose 1993, 139), and Adrienne Rich’s spatialized notion of a “politics of location” in which Rich explores her body as female, white, Jewish, and as a body within a nation, allows Rose to connect thinking through the body to scales both bodily and geopolitical, and to refuse the divide between metaphorical and real spaces.

These spatialities form the basis of Rose’s notion of the subject of feminism occupying a space that is “multidimensional, contingent and shifting” and that is also “paradoxical,” in the sense that “spaces that would be mutually exclusive if charted on a two-dimensional map – center and margin, inside and outside – are occupied simultaneously” (1993, 140). Constituted through, but also working to destabilize, the same–other masculinist discourses of the master subject that limit what women are and can be, the subject of feminism is simultaneously in two places. Rose suggests a different territorial logic to the transparent space of dominant subjectivities, one in which geographic imaginaries are built on a refusal of territoriality (namely, the claiming of territory), thus allowing for difference. Indeed, feminist geographic imaginaries are infused with multiplicities of situated differences that speak to class, race, sexuality, ablebodiedness, religion, and multiple other relations of power and relational understandings of subject formation.

Within the discipline of geography it is situated racialized differences that have most exercised issues of knowledge production within feminist geography (see Women in geography). The hegemonic discourse of whiteness speaks to epistemic exclusions that limit feminist geography’s ability both to speak across differences and to occupy an emancipatory paradoxical space. Drawing on the work of the African American feminist scholar Patricia Hill Collins (1990), women of color in Anglo-American geography have taken their simultaneous occupation of center and margin as an “outsider within” position (Mullings 1999) to describe their paradoxical position within feminist geography, of being inside the same and outside as other same/other. Such paradoxical positions are characterized by their troubled relation to the hegemony of whiteness and the hard-won political spaces from which challenges to it are launched.
New contributions to feminist methodologies

Attempts have been made not only to apply but also to add original contributions to feminist knowledge production, perhaps most clearly in Cindy Katz’s theoretical construct of “counter topography” (2001a; 2001b). Katz argues that Haraway’s concept of situated knowledge is often taken to refer to knowledge from a single site and a knowing subject, and, while situatedness may imply locale, it is most commonly constituted as a subject position – as Katz (2001a, 1230) puts it, “a space of zero dimensions,” located nowhere specifically. Situatedness, Katz suggests, implies location in abstract location to others, but not any specific geography leading to a “politics of ‘sites’ and ‘spaces’ from which materiality is largely vacuated” (2001a, 1230), erasing the effect of specific historical geographies and the difference that space makes. She also argues that situated knowledges are “simultaneously universal … and specific. … The politics of extension and translation, from the site (point) to the global is too easily assumed in the insistence on situatedness, when, of course, that is what has to be explained” (Katz 2001a, 1231).

Katz’s response to the inherent universality of situated knowledge has been to redefine it as the “local particularities of the relations of production and social reproduction” (2001a, 1230) and to employ the method of topography to show how social relations are far from abstract, but rather are “sedimented into space” (1229). Turning from the abstract topological representations of space inherent to Haraway’s formulations she gives ontological priority to topographical (territorial) space. Katz’s intention is to reinsert materiality into feminist theorizing through an insistence on studying the processes of global capitalism in particular places, each with its own specificity but also with its connections to other places. Drawing on the metaphorical idea of contour lines, which are lines of the same elevation that connect places at the same altitude, to describe how places are connected to each other, she imagines each contour line as representing a process, such as deskilling. Tracing these contour lines across places shows how they are materially connected to each other by the same processes. Countertopography, then, is a means of recognizing the historical and geographical specificities of particular places while also inferring their analytic connections in relation to specific material social practices. It is these material connections based on common interests, as opposed to shared identities, that serve as the basis for a feminist politics of connection. The intent underlying Katz’s notion of countertopographies is to suggest the importance of situated knowledge for a gendered oppositional politics that “works the ground” of and between multiple social actors in a range of geographical locations – crossing space, boundaries, and scales – in order to disrupt hegemonic understandings of capitalist processes. Katz eschews any suggestion of being able to predetermine how such politics may play out. She argues that, for politics to be effective in combatting globalizing capitalism, it is not simply a matter of building coalitions between places (vital as they are), but of building a politics sensitive to global processes while being grounded in the local.

The work of imagining knowledge production, of bringing together the intellectual and the political through the mobilization of situated solidarities, has been practiced by a number of feminist geographers and is most comprehensively documented in the work in Uttar Pradesh, India, of the Sangtin Writers Collective and Richa Nagar (2006). They argue that praxis – the simultaneous and continual interplay between thought and action – through the interweaving
of theories and self-reflexive practices of knowledge production via collaborative dialogues, provides a way to radically rethink existing approaches to knowledge production and issues of subalternity, voice, authorship, authority, and representation. Nagar and Lock-Swarr’s (2010, 2) work on critical transnational feminist praxis also speaks to the importance of recognizing the collaborative nature of knowledge production in developing feminist alliances across difference. Collaboration can be thought of then as a methodological tool that bridges the gap between (academic) feminists engaged in theorizing the complexities of knowledge production across borders and those (activists) concerned with imagining concrete ways to enact solidarities across nations, institutions, sociopolitical identifications, and economic categories and materialities.

**Valorizing epistemology over praxis**

Although praxis involving alliances across difference lies at the heart of feminist geographic knowledge production for many feminist geographers, such alliances have been marginalized, some would argue increasingly so, in the neoliberal academy, which values the fast creation of outputs best secured from short-term research engagements that limit the longer-term engagements that praxis usually requires (Peake and de Souza 2010). Throughout the 1990s and 2000s, it could be argued that feminist geographers’ interests in how the practices of research influence knowledge production were dominated by analyses of how the researchers’ identity and experience shape the research process. Praxis, for example, is a term that rarely appears on the pages of the feminist geography journal *Gender, Place & Culture (GPC)*. Instead, it has been a much narrower range of epistemological issues, namely, questions of Positionality, self-reflexivity, representation, and accountability, that have come to occupy feminist geographers, threatening to obscure feminist knowledge production as a political act and divorcing it from questions of praxis.

Positionality highlights how people, including researchers, see the world from different social, cultural, and geographical locations, with those dimensions of a researcher’s lived experience most commonly assumed to influence the research process being her gender, race, class, and sexuality. Building on the work of Gillian Rose, feminist geographers have highlighted how the positionality of the researcher is itself both contextual and mercurial. Awareness of one’s positionality has been most commonly explored via the methodological tool of self-reflexivity that Kim England (1994, 82) defines as “self-critical sympathetic introspection and the self-conscious analytical scrutiny of the self as researcher” (emphasis original) or being aware of the partiality of one’s own social location and knowledge. Indeed, reflexivity is critical to the conduct of research; it induces self-discovery and can lead to insights and new hypotheses about research questions. Self-reflexivity has also been critiqued, however, for its uncritical adoption, for being seen as an end in itself, as a reflexive accounting of individually held beliefs that all too often turns to introspection divorced from the situated power relations within which the researcher operates, with the effect of recentering the concerns of the researcher. Moreover, as Nagar and Lock-Swarr (2010, 7) have pointed out, understandings of positionality “often assume transparent reflexivity in ways that the very desire to ‘reveal’ the multiple, complex, and shifting positionality of the researcher freezes identities and social positions in space and time,” preventing analysis of the ways in which the embodied subjectivities of the researcher and the
researched are mutually constituted, interactive texts through which research-based knowledge production is produced and negotiated.

Feminist researchers’ engagement with what can count as evidence of explanation and how we can know if it is “correct,” or the extent to which it holds in the social world, has focused on challenging the authority of the researcher’s voice, which positivist epistemology, in its dichotomous representation of object and subject as a prerequisite for objectivity, has presented as that of an omnipotent expert. Feminist geographers have sought to dismantle their authoritative positions, giving voice to participants by respectfully engaging them in the decision-making processes of the research, conducting research with as opposed to on participants, experimenting with writing styles, diversifying practices of dissemination, producing materials that speak to nonacademic as well as academic and policy audiences, and sharing research products, such as interview transcripts and publications, with participants. In doing so, feminist geographers are also engaging in the practice of accountability, in pushing boundaries of knowledge production, yet it remains unclear just how systematic or influential such practices are; given the constraints of time and resources, they may not often be rigorously employed. And as Nagar and Lock-Swarr (2010) state, the limitation of these epistemological concerns is that, despite their progressive intentions, they do little, if anything, to dislodge the hierarchy of knowledge producers or to trouble the notion of the capacity of a researcher to “empower” research participants or the orientation of processes of praxis toward progressive humanization. Similarly, Smith (2012) urges researchers to take deliberate care to examine the effects of knowledge as a legitimating process, and to consider whose agendas and beliefs are legitimized in any research project.

Debates on methods

Initially, methods in feminist geography were influenced by the second wave women’s movement and quantitative research (most of its practitioners in the 1970s had come up through geography, at least at the tail end of the quantitative revolution). Feminist geographers documented the low numbers of women in the discipline, mapping their locations and documenting spatialized dimensions of their experiences. It was this early work that not only showed women as subjects of geographic research but also served to make women geographers into legitimate producers of knowledge (Thien 2009). By the mid-1980s, feminist geographers had moved on from such empirically oriented work to engaging further with theoretical perspectives and the masculinist underpinnings of the subfields of human geography. In moving away from quantitative methods, feminist geographers were part of the general shift in research practices in human geography to qualitative analyses. The arguments feminist geographers developed against the use of quantitative techniques are well rehearsed. Implicitly masculinized, quantitative techniques were seen as incapable of reflecting the complexity and richness of women’s lives, with many aspects of experience being difficult if not impossible to quantify, and as producing disembodied and abstracted knowledge that can only speak to predetermined categories that may neither “fit” women’s experiences nor capture interrelationships. The most common objection has been cast at an epistemological level, namely the association of quantitative research with positivism. Positivist knowledge production is critiqued for its assumption of universality, what Donna Haraway (1988) refers to as the “god-trick,” the “view from nowhere.” It is another sleight of hand altogether, though,
that has relegated quantitative techniques to an association *solely* with positivism.

Tainted by their association with positivism, quantitative methods today, such as statistical analyses of survey or secondary data and geocomputation and geovisualization based on geographical information systems (GIS), are not absent (Mattingly and Falconer Al–Hindi 1995) but they are much less likely to be employed by feminist geographers. The epistemological shifts in understanding gender – from its being a role to a power relation, to its identification with identities with different meanings and spatialities, to being something processual, something that subjects do that is considered as embodied, performative, and spatialized – have methodological implications and have served to promote the use of qualitative methods. Such methods are considered more open to relationships, to encouraging engaged dialogue between researchers and communities, to what can count as data – perceptions, experiences, representations – and to what counts as ethically sound research practice.

Indeed, a rich variety of qualitative methods are employed by feminist geographers, including interviews, focus groups, participatory action research (not to be confused with praxis), ethnographies, discourse analysis, archival analysis, visual techniques such as photography and video, plays, oral and life histories, and personal narratives. Notwithstanding the use of multiple qualitative methods, and triangulation between them, however, the overwhelming method of choice is the interview. A review of all the articles based on primary research (from 1993 onward) in *GPC* shows that by far the most frequent, and often only, method employed is that of the “in–depth” interview (a term that has become virtually meaningless as the meanings attached to it are rarely defined, and they may vary tremendously in length from 30 minutes to two– to three–day events) usually employed with small samples of under 50 participants. Within the pages of *GPC*, moreover, the overwhelming focus on qualitative methods and the virtual absence of quantitative research (less than 5% of all research articles) have allowed a vacuum to occur, legitimizing the absence of any reference to broader questions of the politics of knowledge production. Thus the extent to which debates by feminists in other disciplines about quantitative versus qualitative methods have been recognized as a red herring is still unclear, with many of the critiques of quantitative methods no longer being applicable when they can be deployed within a feminist research design. It is as if the recognition that the association of quantitative techniques with positivist science has been socially and historically constituted, and is moreover an ideological position that is neither necessary nor inevitable (Lawson 1995), has bypassed submissions to the journal.

Since the early 2000s, a small but increasing number of contemporary feminist geographers have been using postpositivist geospatial technologies (GIS, Global Positioning System (GPS), remote sensing), and it is in these feminist applications that some of the most innovative possibilities of feminist methodologies are being explored. The work of the feminist geographer Mei-Po Kwan serves as an example of hybrid geographies combining methods and perspectives commonly thought to be incompatible, integrating spatial analytical methods with feminist theory. For example, in her study of the impact of anti-Muslim hate violence in the United States after 9/11, Kwan attempts to produce a counternarrative to recover the lived affective experiences of American Muslim women in Columbus, Ohio, in order to understand the short–term and long–term impact of their fear of being attacked while going about their daily activities and using public space, and
the strategies they have adopted to cope with it. In her mixed-methods study she utilized interviews, mental mappings, digitally recorded oral histories, and GIS-based visual narratives using ArcScene (the 3-D geovisualization environment of ArcGIS). The video she produced is based on the oral history of a Muslim woman in Columbus, Ohio, which

not only shows the routes and the spaces her body moved through, but also tells her story through the images and her oral narrative as she recalls what happened to her life and how she negotiated the hostile urban spaces after 9/11. The video shows what she saw and experienced from her personal point of view (i.e., from the position of a driver who was traveling along various roads in the study area). It is a powerful form of individualized storytelling based on her personal movements, memories, feelings, and emotions. (Kwan 2007, 27)

The woman herself is absent from the video, unable to be watched by spectators; taking the viewpoint instead of the protagonist, the videographer serves to contest the objectifying gaze of dominant practices of geovisualization. Kwan argues that such alternative uses of geospatial technology not only have relevance for a world characterized by natural disasters, global-level conflicts, and wars through, for example, community-based mapping and progressive uses of Big data, but also to contest the dominant uses of geospatial technologies to combat their usage for violent and oppressive ends. She argues that a moral use of these methods requires that geospatial practices are embodied and attentive to emotions, claiming that feminist geospatial technologies (GT) practitioners “can appropriate the power of GT, contest the dominant uses of these technologies, and reconfigure the dominant visual practices to counter their objectifying vision … and explore the possibilities of performing (practicing) GT as resistance” (Kwan 2007, 23–24). Kwan’s work shows how quantitative data and methods can be powerful ways of initiating progressive social and political change, revealing not only broad but also micro-contours of difference and similarity and speaking to policy. The reality, however, is that many feminist geographers cannot take up these challenges because they lack training in quantitative analysis, which reduces the scope of the research questions that many feminist geographers can ask.

**Whither feminist methodology?**

While feminist methodology served initially to critique the epistemological claims of masculinist knowledge production and bring into view the female subject and embodied knowledges, currently it is the substitution of the use of qualitative methods for discussions about feminist methodologies, a restricted engagement with praxis and its replacement by questions of epistemology, and the reduction of these questions predominantly to those of reflexivity and positionality that have come to define the parameters of debates about the current production of feminist knowledge in geography. These are narrow parameters and we need to be wary of the limitations they place on feminist geographical knowledge production and on geographical imaginaries. For feminist geographers concerned with the intellectual and political practices of knowledge production – what constitutes knowledge, what knowledge is considered legitimate, and how it is produced – there is a need to move on from the dead-end debate about quantitative versus qualitative methods, which serves only to replicate dualist thinking; to ask broader methodological questions that include a place for praxis and a recognition of the need for much greater flexibility over suitable methods, including tolerance
and respect for a variety of methods; and to understand that no method is inherently feminist and that all data are representations. Furthermore, with big data rapidly gaining traction and with some of the most exciting methodological innovations in geography coming from geospatial techniques, a monistic take on qualitative methods by far too many feminist geographers runs the risk of reducing the epistemic authority of feminist work and the dynamism of the field. Clearly, it remains critical for feminist researchers to acknowledge and continually to problematize power relations in research.

SEE ALSO: Affect; Bodies and embodiment; Critical spatial thinking; Emotional geographies; Feminist geography; Fieldwork in human geography; Focus groups; Gender; Geographic information science; Intersectionality; Interviews; Participatory action research; Positionality; Power; Qualitative data; Quantitative methodologies; Reflexivity; Representation; Subjectivity

References


Feminist political ecology

Juanita Sundberg
University of British Columbia, Canada

Feminist political ecology (FPE) is a subfield that brings feminist theory and objectives to political ecology, which is an analytical framework built on the argument that ecological issues must be understood and analyzed in relation to political economy (and vice versa). Feminist political ecologists hold that gender—in relation to class, race, and other relevant axes of power—shapes access to and control over natural resources. FPE also demonstrates how social identities are constituted in and through relations with nature and everyday material practices. FPE builds bridges between sectors that are conventionally kept apart—academia, policymaking institutions, activist organizations—thereby connecting theory with praxis. In addition, FPE weaves threads between sites and scales to produce nuanced understandings of the socioecological dimensions of political economic processes. Rooted in feminist critiques of epistemology (the study of how knowledge is produced and legitimized), FPE asks compelling questions about who counts as an environmental actor in political ecologies and how ecological knowledges are constituted. As such, FPE has made substantive, epistemological, and methodological interventions in political ecology, environmental studies, and gender studies.

Sites of inspiration and formation

Feminist political ecology was forged out of feminist and women-centered scholarship and activism in environmental and livelihood/quality of life issues. Inspired by feminist movements of the 1970s, many scholars and activists began to approach nature–society issues with a feminist sensibility, characterized by a persistent linking of the personal and the political. Such feminist environmental engagements brought the feminist movement’s diverse political objectives to bear on the most intimate sites of daily life including relations between humans and nonhumans, food consumption, and corporeal wellbeing. Feminist scholarship in this vein both elaborated critiques of research that excludes women, and advanced alternative theoretical framings to account for women (Haraway 1991; Seager 1993). This now extensive and theoretically varied body of work asks fundamental questions about the relationship between forms of oppression and the domination of nature as manifest in environmental degradation, species extinction, industrial slaughter, toxic contamination, and so on. Feminists also advanced alternative ethical framings built on concepts such as relationality, care, responsibility, and friendship (Cuomo 1998).

Feminist political ecology emerged from this arena of lively debate and theorizing. Three bodies of work are particularly relevant to the consolidation of FPE as a subdiscipline: ecofeminism, feminist science studies, and feminist...
FEMINIST POLITICAL ECOSYSTEM

critiques of development. Ecofeminists point to links between the oppression of women and the exploitation of nature, although how such links should be analyzed and acted on is highly debated (for an overview, see Diamond and Orenstein 1990). Although some suggest that women are closer to nature because of their biologically constituted corporeal experiences, the majority of ecofeminist scholars turn to historical shifts in Europe, including the scientific revolution, capitalism, and colonialism, to demonstrate how and why women in Western societies (and their colonies) are so frequently associated with nature, as well as how nature is feminized (see Merchant 1980). For example, environmental philosopher Val Plumwood (1993) traces associations between women and nature in Western societies to oppressive material relations – for example, sexism, colonialism, anthropocentrism (a belief that humans are the most important entities) – that have left their mark on epistemology (or ways of knowing) in the form of a network of dualisms. Accordingly, the human has been framed in opposition to nature in Western thought, with the human capacity for reason and abstract thought as the grounds for transcendence and domination of nature. In turn, reason is framed as masculine through its opposition to and domination of all that is associated with nature, the body, reproduction, emotion, and ultimately the feminine. Plumwood’s work demonstrates how such dualisms underpin oppression.

Postcolonial feminist scholars have criticized Western ecofeminism for its narrow focus on the philosophical or conceptual dimensions of oppressive relations as well as its neglect of the political economic arrangements – at multiple and intersecting scales – that constitute actual ecological relations in particular places (see Shiva and Mies 1993). Debates in ecofeminism continue to inform feminist political ecologists’ interest in how women and men’s relations with the natural world, in particular, played an important part in defining gender norms, such as notions of appropriate femininity and masculinity.

An equally important arena of inspiration for the emergence of FPE is feminist critiques of science and epistemology. Sandra Harding (1986), Donna Haraway (1991), and others argue that patriarchal gender norms inform basic conceptions of who counts as a knowledge producer, what counts as knowledge, and how knowledge is produced. Scholars in this vein demonstrate how women and other marginalized groups are systematically disadvantaged by conventional scientific practices that exclude them as knowers, while producing knowledge that renders their experiences invisible or represents them as inferior. As such, feminist studies of science problematize the concept of objectivity. Conventionally framed as a value-free view from nowhere, objectivity is predicated on the assumption that the researcher’s mind is separate from his or her body, social position, and geopolitical location. Feminists argue that, historically, claims to objectivity masked and protected what were actually the partial perspectives of dominant social groups, specifically European or white, heterosexist, bourgeois men. Hence, the aura of objectivity is an achievement, derived from denying or concealing the researcher’s embodied subject position. In addition to these critiques, feminists introduced various alternatives to masculinist forms of objectivity. For instance, Haraway’s (1991) concept of situated knowledge suggests that knowledges emerge in relation to embodied social locations. Harding’s (1986) proposal for partial objectivities takes subjective or local knowledges seriously by developing methods to verify and validate them within specific contexts of shared experience. Theirs are not calls for relativism but for responsibility and accountability in practices of knowledge production.
production. Feminist political ecologists build on these conversations to address how research practices are implicated in (re)producing and contesting power relations.

A third body of scholarship important to FPE is feminist critiques of development, which demonstrate how women have been excluded from or exploited by (sustainable) development and conservation projects (Shiva and Mies 1993). Feminist postcolonial scholars such as Chandra Mohanty (1991) complement this work, exposing how Western feminists leading development projects tend to depict what she calls “Third World women” as victims in need of Western help; such homogenizing portrayals deny the diversity of women’s locations, experiences, and knowledges. Scholars working in this field address the ways poverty is deepened and feminized when women are neglected as agents of environmental transformation (e.g., as managers of natural resources) and environmental knowledge bearers/producers. For instance, Judith Carney (1992) revealed how gender differences in land use, labor obligations, and crop rights articulate with development in The Gambia, Africa. International donor projects that introduced irrigation systems and improved rice production packages to male household heads resulted in women’s loss of access to land and, in some cases, income. Richard Schroeder’s (1999) research, also in The Gambia, centers on conflicts between men and women sparked by international donor projects in the 1970s, which were designed to include women in development by supporting women’s expansion of market gardening. When donor interests shifted to environmental concerns in the 1980s, however, men were encouraged to engage in agroforestry on the same plots of land as the gardens. Consequently, men’s and women’s crop production systems came into conflict. Ultimately, as scholars have documented, if development agency personnel and researchers consult only men, then the relevance of particular resources, women’s specific knowledge of them, and women’s livelihood strategies are made invisible. This, in turn, generates resistance among women toward development and conservation interventions. The importance of this body of work is evident in FPE’s ongoing emphasis on the potentially devastating consequences for women and their dependents when gender differences in resource management and land-use practices are neglected.

Building on these three bodies of work and debates, Dianne Rocheleau, Barbara Thomas-Slayter, and Esther Wangari (1996) put forward feminist political ecology as an integrative conceptual framework in the edited volume Feminist Political Ecology. The book situates gender as a crucial variable – in relation to class, race, and other relevant dimensions of political life – in shaping environmental relations. Rocheleau, Thomas-Slayter, and Wangari (1996) suggest that gender norms result from social interpretations of biology and socially constructed gender roles, which are geographically varied and may change over time at individual and collective scales. As such, the editors shift away from essentialist (i.e., one-dimensional and universalizing) constructions of women found in some ecofeminist work to treat gender differences and gender relations as constituted in and through material political ecological relations. The book’s conceptual agenda advances three primary areas of research: (i) gendered environmental knowledge and practices; (ii) gendered rights to natural resources and unequal vulnerability to environmental change; and (iii) gendered environmental activism and organizations. And, the editors outline an exciting call for research that connects the local and global, urban and rural, North, South, East, and West, through close analysis of everyday experiences and practices of gendered
environmental risks, rights, and responsibilities. Chapters feature case studies from across the globe rooted in collaborative and activist methodologies. Authors address issues such as the struggles of women on the front lines of the rubber tappers’ movement in Brazil, women in environmental justice organizing in West Harlem, New York, as well as women dealing with industrial waste in Spain. Feminist Political Ecology marks a noteworthy moment in environmental studies by demonstrating the analytical purchase of feminist political ecology to identify how inequality is (re)produced when women’s environmental engagements, knowledge, and activism are neglected. Recent work in feminist political ecology continues to engage with the agenda and debates outlined in the book.

Sites of intervention and contribution

Feminist political ecologists have produced a vibrant body of work that significantly enriches understandings of the political–ecological nexus. Moreover, researchers’ substantive contributions have prompted epistemological shifts and methodological innovation. Simply by engaging women as political actors, agents of environmental change, and bearers/producers of environmental knowledge, feminist political ecology revolutionized research in political ecology. While seemingly straightforward, considering women has far reaching consequences, for it is not possible to simply add women to existing frameworks and proceed as before. Indeed, to disrupt conventional assumptions about men as the primary environmental actors is to ask fundamental epistemological questions about how knowledge is produced and legitimized. For instance, feminist political ecology challenges claims to objectivity by pointing out that if researchers only engage men in any given site (as if they represent the only or primary actor), then their results are partial rather than neutral or unbiased.

Dianne Rocheleau and David Edmunds (1997) make this point by showing that women in many rural areas manage spaces – along with specific natural resources – that are nested in or between spaces controlled by men. Their analysis of the gendered dimensions of tree tenure around the world sheds light on the complexity of customary laws that grant men and women differing rights and responsibilities to multidimensional fields with distinct and overlapping species. For example, women often have customary rights to species above, below, or between men’s crops or trees; as such, they are subject to men’s decisions about changing the species they plant or tend (see Schroeder 1999). As this research highlights, attending to women as resource managers reveals the limitations of existing two-dimensional concepts of land and landownership, which are based on fieldwork with men only. These concepts do not account for the multidimensionality of species management by men and women.

Likewise, the personal experiences of white middle-class Western feminists/scholars may restrict their interest in or attention to particular spaces or activities, which, in turn, has the effect of shaping knowledge production. Maria Elisa Christie (2008) makes this point in relation to the kitchen, which is often framed as a principal site of women’s oppression in Western feminism. Christie’s close engagement with women’s “kitchenspaces” in central Mexico demonstrates the importance of food preparation in the enactment of rituals and fiestas that sustain extended family and kinship networks as well as unique skills and knowledge. In short, FPE demonstrates that political ecological stories are implicated in power relations, and researchers risk reproducing gender inequalities if and when women are left out as agents of environmental change.
Accounting for women as actors brings about additional epistemological shifts. Since many women around the world labor in social spheres that, historically, have been excluded from analysis, addressing the particularities of their knowledges and practices requires asking questions about what scales of political ecological life count as relevant. Building on feminist economics and feminist geography, research in FPE draws attention to everyday intimate and embodied practices along with household micropolitics. The scale of the everyday is where social reproduction takes place, where subject identities and social orders are brought into being and contested. Attending to daily life allows FPE to shed light on otherwise neglected dimensions of environmental engagements. For instance, Shubhra Gururani’s (2002) research with women collectors of fuel and fodder in the Kumaon Himalayas suggests that forests are sites of emotion, memory, and meaning. Women engage the forest as much more than simply a backdrop or site of resources where they meet livelihood needs, Gururani argues; indeed, women’s everyday material engagements constitute but also challenge culturally specific gender norms. As such, Gururani’s findings contest predominant utilitarian and mechanistic assumptions of human–nature relations in political ecology.

Likewise, Farhana Sultana (2011) examines how natural resource access is mediated through emotions, which are defined as intersubjective (e.g., produced in relationships between people or people and nature) rather than as individual mental states. In rural Bangladesh, where drinking-water wells are contaminated by naturally occurring arsenic, women’s relations with water are saturated and constituted by emotions, particularly suffering. Thus, Sultana suggests, women’s daily lives are configured not solely by struggles to obtain safe drinking water for their families but also by emotional distress; these emotions, in turn, shape women’s decisions about how to negotiate the power relations that constitute water access and control. As Leila Harris (2015) notes, attention to emotions allows feminist political ecologists to demonstrate not only that resource access is important for livelihood and health but also people’s sense of dignity and belonging.

Even as FPE legitimizes the everyday as a significant scale of analysis, researchers also excel at demonstrating how the intimate connects with other scales such as the nation or global political economy. For instance, Yaffa Truelove’s (2011) research on women’s water-collecting practices in Delhi, India links the body to city and state. While city planners look to market mechanisms to fulfill their vision of a modern city with efficient services, Truelove shows how the establishment of metered water sources creates a whole range of “illegal” water practices. Such legal mechanisms particularly affect women in slums without legal water connections, as they must engage in time-consuming, dangerous, and illegalized activities just to procure water for daily needs. As a consequence, young girls in marginalized communities are often kept out of school because of the amount of time required to meet family water needs; this, in turn, limits their life opportunities but also their sense of belonging in a city with global aspirations. For Harris (2015), the importance of research such as Truelove’s is to challenge existing claims made by state and nonstate actors (such as the World Bank) that the commodification of water leads to increased efficiency. As Harris contends, addressing embodiment and the scale of the everyday serves to demonstrate how capitalist logics privileging efficiency ignore nonproductive needs and uses associated with health, poverty reduction, or cultural and spiritual values (e.g., preservation of heritage seeds/crops).
Another important epistemological intervention stemming from the seemingly straightforward act of accounting for women relates to how the subject or person is conceptualized in political ecology. Historically, political ecologists have tended to assume that subject identities are narrowly defined based on taken-for-granted or congruent notions of class position, sex, or race. Juanita Sundberg (2004), Leila Harris (2006), and Andrea Nightingale (2006) draw from feminist poststructural theory to outline anti-essentialist framings of the political ecological subject. Judith Butler’s (1999) work is particularly significant here. Butler argues that gendering practices are not simply built on sex difference; instead, bodies are gendered in and through the regulatory practices of disciplining institutions such as the family, along with medical, educational, and religious institutions. In other words, gendered bodies have no natural foundation (in sex) but are constituted in and through gendering practices that are reiterated or performed in daily life. For Butler, everyday performances produce gendered subject positions rather than simply reflect them.

Sundberg, Harris, and Nightingale build on Butler’s work to insist there is no necessary or pregiven relation between men or women and the environment; rather, such relations are forged through geographically contingent, power-laden practices. Sundberg (2004) analyzes how conservation discourses, practices, and performances in Guatemala are instrumental in mapping gendered and racialized ways of life. In the process, Sundberg also reflects on her research collaboration with an indigenous women’s group to highlight how research practices are constitutive of gendered and racialized performances that (re)produce asymmetrical geopolitical relations. Likewise, Nightingale (2006) treats gender as a process to show how performances of masculinity, femininity, and caste are constituted in community forest management in Nepalese villages. Harris (2006) demonstrates how differences between men and women are (re)cited and naturalized in relation to new irrigation economies and ecologies in Turkey; gender comes to matter to irrigation practices, she argues, through the regulatory insistence on difference.

Feminist critiques of knowledge production also prompt methodological innovations in FPE so as to include previously excluded actors and to account for their knowledges as well as how they come to know their environments. Women and other marginalized groups may consider themselves or their work to be unimportant and their life experiences may lie beyond those of researchers. Moreover, as noted, women’s spaces of work are often nested in those controlled by men. Examining what was made invisible or neglected requires methodological creativity. Many feminist political ecologists work with feminist participatory or collaborative methodologies to enable research that supports feminist political objectives. In this context, feminist scholars tend to conduct qualitative research from the bottom up by privileging the experiences, spaces, and categories of marginalized people. Along these lines, Louise Fortmann (1996) specifically addresses strategies for ensuring that women’s distinct experiences with trees, plants, and animals are included in natural resource mapping. For example, forming separate groups of men and women while undertaking natural resource mapping helps to ensure that women have the space to express themselves freely (see also Sundberg 2004).

Some feminist political ecologists suggest that qualitative methodologies need not be the only ones appropriate to feminist research. Rocheleau (1995) pioneered the development of methodologies to triangulate data derived from
In her discussion of research evaluating the results of a forestry and agricultural initiative in the Dominican Republic, Rocheleau notes that gender-informed quantitative analysis contradicted predominant assumptions of women as auxiliaries to men; in addition, counter-mapping – map-making that starts with rural people and their homes – produced images that resulted from the mixing of local people and researchers’ specific skills and knowledge. Relatedly, Nightingale’s (2003) study of a community forestry program in Nepal combined aerial photo interpretation with ecological oral histories to analyze the effectiveness and sustainability of community forest management. Each of these two methods is rooted in a distinct epistemological tradition and, therefore, produces distinct kinds of knowledge. Working with Haraway’s concept of situated knowledge, Nightingale (2003) treated both aerial photo interpretation and oral history collection as partial yet internally valid methods of generating distinct stories about forest change. Rather than triangulating data, Nightingale attended to the inconsistencies between the data, thereby producing new insights about the pace and location of forest regeneration as well as how and why local people claimed the community forestry program as a success. In so doing, she also framed local people as legitimate producers of environmental knowledge.

In short, research that accounts for women necessitated epistemological innovation, and feminist political ecologists have been at the forefront of developing new theoretical and methodological tools. Nonetheless, the contributions of FPE tend to be assimilated into mainstream political ecology with little explicit acknowledgment. Indeed, in the recent trend to canonize political ecology through the publication of textbooks and edited collections, feminist political ecology is only marginally addressed. And yet, Rebecca Elmhirst (2011a) suggests, political ecology owes an epistemological debt to feminist theory for the range of fresh perspectives it offers. Nonetheless, many scholars whose work articulates with the political and theoretical objectives of FPE do not identify as such. Thus, a review of recently published research demonstrates that the field of gender and environment is flourishing although few identify as feminist political ecologists, leading Elmhirst (2011a) to ask if FPE is a disappearing subject. The response to her question is evident in renewed attention to FPE along with debates about its analytical purchase.

Sites of challenge and debate

In part, the apparent disappearance of FPE is due to the emergence of anti-essentialist framings of gender, which have destabilized assumptions about who counts as the (natural) subject of feminist-oriented research. In addition to Butler’s argument, noted earlier, postcolonial scholars have challenged homogenizing views of women as a pregiven, coherent category that is studied using similar theoretical frameworks the world over (see Mohanty 1991). Such critiques lead to a crucial question: if women are no longer the organizing purpose of feminism and gender is no longer its central analytical category, then what is the point of FPE?

A new generation of feminist political ecologists responds to the destabilization of gender by emphasizing intersectionality as the primary method of addressing how social subjects are constituted in and through diverse and interlocking processes of differentiation such as gender, sexuality, race, ethnicity, class, and livelihood. In other words, new FPE seeks to account more fully for the ways systems of power articulate
in time and place. Farhana Sultana and Andrea Nightingale advance the concept of intersectionality by explicitly considering how subject identity is constituted in and through material ecological relations. While FPE has long treated the natural environment as a constitutive element of political subjectivity, this dimension is often neglected in feminist theory more generally and is in need of further theorization. Sultana’s (2011) analysis of gender–water relations in Bangladesh highlights how the geologic distribution of arsenic in the local aquifer plays a crucial role in configuring gendered subjects. By and large, the contamination of water sources and the resulting need to travel longer distances to fetch safe water has worked to entrench the notion that masculinity is not compatible with water collection. Nightingale (2011) examines how imaginaries of gender and caste boundaries are materially enacted in postconflict Nepal. Normative femininity, she notes, requires Hindu women of a particular caste to be spatially segregated during menstruation because their bodies are considered polluting and therefore damaging to the environment. As such, appropriate performances of femininity are enacted in and through such spatial moves. Nightingale found that the Maoist insurgency disrupted gender and caste performances by enacting shifts in embodied spatial practices like sitting and eating in mixed caste and gender groups.

Sharlene Mollett and Caroline Faria (2013) present a strident critique of new FPE, suggesting that researchers too often continue privileging gender without also giving full consideration to the ways it intersects with race. Race is a crucial variable in subject formation, the authors suggest, while racial thinking constitutes the very categories used to name and order the modern world (e.g., racial labels such as “European” or “African” along with binaries like civilized/primitive, modern/traditional, formal/customary). In other words, the environment and environmental politics are not raceless. Mollett and Faria (2013) point to whiteness as an institutional factor that shapes the production of knowledge in FPE; the predominance of whiteness in the Western academy works to normalize the absence of critical race perspectives. Mollett and Faria (2013) call for a postcolonial intersectional approach that situates patriarchy and racialization as entangled in postcolonial genealogies of nation building and development.

Even with these critiques and reflections, some feminist political ecologists stress the continuing relevance of gender as a key variable due to the persistence of masculinist forms of objectivity and ongoing neglect of women as environmental agents. For instance, Aya Hirata Kimura and Yohei Katano (2014) suggest that performances of gender are at stake in times of crisis or disaster, such as Japan’s Fukushima nuclear reactor accident. Their study highlights how gender norms informed perceptions of risk in the aftermath of the nuclear disaster. Political elites called on binary constructs of appropriate masculinity and femininity to manage the disaster; in emphasizing the need for patriotism, normalcy, and safety, citizens concerned about radiation were feminized as irrational or hysterical. For her part, Elmhirst’s (2011b) study of forests in Indonesia introduces queer theory, which examines how normative gender categories are produced and contested. Elmhirst demonstrates how the Indonesian state manages and controls access to natural resources by privileging heterosexual conjugal couples. In other words, heteronormative marriage becomes an important conduit for resource access and therefore affects women as well as men. Elmhirst calls on political ecologists to question the naturalness of categories such as conjugal relationships and heterosexuality as they are deployed in the practices of knowledge production.
Future directions in feminist political ecology

Even as feminist political ecologists clearly demonstrate the ongoing importance of gender relations in natural resource struggles, feminists work on a range of topics wherein gender is not the primary analytical variable. In other words, feminist scholarship is not restricted to analyses of gender. This is evident in recent FPE scholarship centering on the body as the primary analytical category and site of analysis (Sultana 2011; Truelove 2011). In this vein, Jessica Hayes-Conroy and Allison Hayes-Conroy (2011) elaborate a political ecology of the body framework to account for the intersection of material and affective/emotive practices. Intended to facilitate analysis of food–body relations, especially how schools seek to promote healthy eating habits, the framework insists on considering the articulation of variables at multiple scales: structural factors that (re)produce inequality and therefore access to particular foods; discursive practices that constitute imaginaries of health and good food; and the material interactions that shape the emotive and bodily experience of eating. Hayes-Conroy and Hayes-Conroy’s framework is attuned to the unpredictability of bodily dispositions and potentialities and, as such, makes space for explanations that are complex, partial, and unfinished (as called for by feminist theories of knowledge production). In many ways, their approach is in line with Harris’s (2015) appeal for an FPE centered on the everyday, embodied, and emotional aspects of society–nature engagements.

Another exciting new direction in FPE is evident in recent efforts to more actively consider relations between humans and other-than-human beings such as animals. Here, two concerns found in ecofeminism are given new life: the connections between different forms of oppression; and, proposals for a feminist ethics of care. Building on work that registers the active presence of other-than-humans in coproducing our world, as well as ongoing feminist concerns about who counts as a political actor and what counts as politics, Kirsty Hobson (2007) argues for the inclusion of animals as political actors in political ecology. As Hobson notes, political ecologists risk reproducing oppressive relations between humans and nature by treating animals as mere objects over which people struggle rather than as living beings whose ecology, behavior, and wellbeing are caught up in (shaping) political ecological outcomes. These concerns are taken up in Sundberg’s (2011) elaboration of a more-than-human methodology to consider other-than-human beings as actors in geopolitical processes. As Sundberg demonstrates, desert soils, thornscrub landscapes, and ocelots (a small feline) constitute, inflict, and disrupt the United States’ enforcement of its southern boundary, forcing state actors to call for more funding, infrastructure, and boots on the ground. Sundberg tells alternative stories about the escalation of US boundary enforcement strategies, stories that refuse the US government’s narratives of mastery over borderland environments. With its unique focus on oppressive formations, corporeality, and the politics of knowledge production, FPE is ideally positioned to make innovative contributions to the shift away from treating nature as backdrop and toward an understanding of agency on the part of other-than-human actors.

Finally, recent work suggests that FPE is moving in the direction suggested by Rocheleau (1995) over two decades ago: to undertake research touching on gender, class, and other systems of difference from a position of affinity as opposed to identity. If identity politics implies assuming that women share concerns as women, affinity politics entails situating ourselves and research participants in webs of power and identifying research questions on the basis of issues
FEMINIST POLITICAL ECOLOGY

of shared concern, such as neoliberalization, environmental degradation, and imaginative geographies of distance and difference. A useful template for the establishment of research collaborations across sites and scales is Cindi Katz’s (2001) concept of counter-topographies, which entails tracing lines between places to show how they are constituted in and through the same processes of development or environmental change. In this vein, Roberta Hawkins (2012) forges new ground in her critique of ethical consumption campaigns that position Northern (female) consumers as saviors of (feminized) people and environments in the Global South. Approaching consumption as a gendered and environmental act that connects the intimate and global across geopolitical space allows Hawkins to chisel away at entrenched binaries such as North/South and researcher/researched that continue to structure political ecology.

Likewise, Harris (2014) considers the implications of Western models of environmentalism in Turkey through a framework she terms imaginative geographies of green, which builds on postcolonial and intersectional analytics. Harris examines how everyday narratives of environmental politics in Turkey articulate differences between East and West and, in so doing, evoke painful legacies of colonialism. Harris calls on scholars to problematize what counts as appropriate environmental politics so as to refuse the West as the primary or only legitimate point of reference and thereby initiate the process of decolonizing conceptualizations of green politics, citizenship, and subjectivities. Research that begins from a position of affinity rather than identity promises to shift political ecology away from studies that examine the concerns of distant and different others and toward research that is accountable to the many ways in which scholars are entangled in and complicit with the very webs of power, privilege, and oppression they seek to analyze.

As a style of research, FPE works with feminist concerns about how oppressive relations are (re)produced at various scales of everyday life and makes significant epistemological and methodological interventions in feminism and political ecology alike. Working at the nexus of nature, power, and knowledge production, FPE promises to continue supporting broader feminist political objectives for more equitable and ecologically viable futures.

SEE ALSO: Bodies and embodiment; Environment and gender; Feminist geography; Feminist methodologies; Gender; Gender and development; Identity; Intersectionality; Natural resources; Political ecology; Race and racism; Scale

References

Elmhirst, Rebecca. 2011b. “Migrant Pathways to Resource Access in Lampung’s Political Forest:
FEMINIST POLITICAL ECOLOGY


Fertility

Yvonne Underhill-Sem
University of Auckland, New Zealand

Fertility is understood most widely in geography as one of the three determinants of population change alongside migration and mortality. Everyone is born, everyone moves geographically in some manner, and everyone eventually dies: however, such self-evident understandings belie the importance of the conceptual framing of these processes, and none more so than fertility. A fertile concept, idea, or process is one that promotes further inquiry, knowledge, and understanding. Yet, within geography, connecting these understandings of fertility through research and analysis is neglected. Instead fertility is incorporated in a routine, uncritical way, with the result that geographic understandings of fertility remain largely taken-for-granted and under-researched. A partial reason is that within geography there has been a preoccupation with migration as the main demographic process that distributes people spatially. Yet this preoccupation overlooks the vital fact that one needs to be born to be able to move, and birth is something that occurs in diverse ways and with potentially problematic consequences for all involved. A more complete understanding of fertility involves inquiry into the spatiality of vital population acts (such as giving birth), performances (such as becoming pregnant), and institutions (such as bride wealth exchanges). Such inquiry would provide insight into areas such as reproductive practices, sexualities and embodiment, and notions of family and motherhood.

Developing critical analysis from the foundation of classic fertility narratives, statistical examination of survey and census data, and multilevel interdisciplinary methodologies can provide more nuanced insight into the dynamics of population change. Such insight promises to contribute to debates both within and beyond geography. For instance, the global mobility of domestic workers into Singapore can be understood as groups of women moving to care and nurture the aged and young in distant places. This practice can also be understood as sexualized subjects carrying with them particular fertility potentials that change their childbearing possibilities. Such a view allows a recognition of new fertility patterns and their implications. For instance, in what ways is their fertility controlled in new workplaces? In what ways might it appeal to have children born in places where they might have new possibilities for citizenship? And in what ways might new, informal sexual relations in different places affect formal sexual relations in home countries? Similarly, with new reproductive technologies available in low-fertility countries such as those in Europe, the social relations between the different biological contributors to a particular pregnancy are increasingly complex and variable. In what ways are new configurations of parents affecting understanding of households and families? In what ways does the particular scale of new reproductive technologies affect analysis of small-scale population dynamics, and in what ways does this development produce new options for citizenship and inheritance? Furthermore, how have historical patterns of
FERTILITY

reproduction, such as China’s one-child policy, or localized environmental events like nuclear contamination in the Ukraine, or large-scale political catastrophes like genocide in Cambodia, affected the fertility of people and their procreative practices?

Mainstream understandings of fertility and reproduction

Fertility routinely refers to the ways in which groups of people produce offspring. Some demographers argue that the term “fertility” refers to a more refined analysis of “natality,” which refers more broadly to the role of births in population change and human reproduction, and invites a more comprehensive analysis. This framing has not, however, flowed into other disciplines, such as geography or anthropology, which also appropriate demographic analysis. Instead universal assumptions about biology mean that similar measures of fertility are used in nonhuman populations such as plants and animals. Population geography, however, offers distinctive possibilities of shifting analysis across scales between bodies and populations and back again. In human populations there are many measures of fertility and each one attempts, at increasing levels of “accuracy,” to represent the rate at which a population is growing by relating the number of live births to a specific population and time period. Unfortunately, even identifying a live birth, let alone ascertaining how many there are, is difficult.

The simplest measurement is the crude birth rate, which represents the number of live births in a given year by the number of people in that population in the middle of the year. Slightly more refined and commonly used measures include general fertility rates (the number of births by the number of women of childbearing age), age-specific birth rates (the number of births by the number of women in a given age group), and total fertility rates (the sum of age-specific birth rates over all ages of childbearing women). More sophisticated measures include completed fertility rates, which show the fertility progress of a cohort through the childbearing years, and age-specific birth probabilities. The former provides some insights into historical shifts in the timing of child-bearing and the latter provides for the calculation of population projections.

The variability of fertility among subgroups of a population means that it is important to refine measures of fertility specific to known age, nuptial status, and other social and temporal features. Further, because of the total absence or unreliability of records about birth dates or total populations in many parts of the world – especially in the demographically largest and fastest-growing parts of the economic south such as sub-Saharan Africa and China – continual efforts are made to develop the best techniques to estimate fertility rates. Despite ongoing and often justified skepticism as to the accuracy of fertility rates, they are still ubiquitously used and are relevant for comparative accounts of progress toward social equity. All fertility rates have temporal and spatial dimensions. Used carefully, with due regard to the assumptions and methods of data manipulation, the comparative analysis of fertility rates provides a convenient representation of change. For this reason it is useful to calculate fertility rates, refining them as best one can, but keeping in mind that they are often a crude tool for analyzing more complex processes.

Closely related to the notion of fertility is the term “reproduction,” or the extent to which a group of people are replacing themselves by natural processes. The term “reproduction” has
FERTILITY

A wider scope than the term “fertility,” and is defined demographically as the act or process of producing offspring, giving birth, bringing forth offspring, producing, yielding, making, manufacturing, working, and/or creating unity. Used in demography and population studies, including population geography, this term is often limited to the biological event of producing babies. It clearly has a wider meaning linked to the production of human beings and is the focus of a range of social sciences, especially those interested in class differentiation. The main problem with this concept is that, like “fertility,” biology tends to be treated as an uncontested universal concept that takes agency away from people, and especially women.

Central to the shortcomings of these mainstream conceptualizations of fertility and reproduction is the unproblematic treatment of the body as exclusively a biological given. Bodies are now accepted as being both biologically configured but also socially, culturally, and politically constituted by places and people. This recognition provides for nuanced analysis of the complex layering of meaning regarding how groups and individuals produce offspring.

Reconceptualizing fertility:
of epistemologies, bodies, and desires

Interest in fertility within demography is matched by a similar longstanding interest within anthropology in documenting ethnographic accounts of childbirth in different places. In the process of documenting the many different cultural interpretations of the “universal” process of birth, however, it becomes clear that childbirths are as different as the overlapping and frequently contradictory discourses that are used to describe them. For example, in the English-speaking world, it is not unusual to hear fathers-to-be saying “we are pregnant.” Clearly, there are solid grounds for reconceptualizing another epistemological position from which to interrogate notions of fertility.

Population geographers increasingly examine the implications of an epistemology in which demographic events and process are socially constructed and therefore partial. Following the work of feminist anthropologists and demographers, population geographers also consider how “places” construct bodies so that they are not simply understood as a biological given. To do so requires a research process that combines critical ethnographies and geographies with statistical analysis.

This process begins with recognizing that all categories are “marked” in one way or another. Feminist scholars have long drawn attention to such marking by showing how the default body in social theory is often a white male body. More robust scholarship would specify more precisely what bodies are being discussed. Thus even fertility analysis would require consideration of what precise body is being represented in mainstream analysis. An embodied analysis would give corporeal form to otherwise abstract definitions, and doing so would subvert taken-for-granted understandings whose privileging has suppressed other ways of knowing. Such analysis is likely to include other discursively constructed realities such as the existence of “specialists,” “sorcerers,” “deceased ancestors,” and even God. Placing the analysis of fertility within specific historical contexts and places gives explanatory preference to the complex interplay of structural features like the dynamics of social, cultural, and political issues. Thus a study of fertility would not simply focus on the event of the birth of a child but would also examine how the mother and the others involved talk about the child, the pregnancy, and the birth. In this way fertility can be understood as relating to bodies as biological
FERTILITY

constants, but also bodies situated in historical processes in particular places.

Such understanding requires looking beyond mainstream measures such as total fertility rates to also examining the discursive processes that constitute the body and particular subjects. This approach opens up the possibilities of considering subjects such as, for instance, the bodies of trafficked women, of sexually active people who live with physical impairments, disabilities, or chronic illness, or women living in isolated villages in the economic south.

The intersection and interrelationships of material and discursive fields in relation to bodies and places is a critical place from which the analysis of fertility can expand. When one begins to capture the workings of communities and family groups, enmeshed in the overall agreements as to the aim of particular fertility outcomes, different gendered versions of these agreements will emerge. For instance, the reasons for women wanting children, and especially sons in some places, may not be only to maintain lines of male descent but also to provide comfort, loyalty, and security in one's daily life. These fertility desires and sentiments last as long as mothers and caregivers do, but they play an important part in understanding otherwise contradictory fertility practices.

Depoliticizing fertility: abortion, the International Conference on Population and Development, and HIV/AIDS

Fertility studies in geography are an important route into working with contemporary social justice issues around women’s sexual and reproductive health and rights. Abortion remains a controversial and complex issue that incorporates moral positioning and political responsibility in ways that can threaten the overall development of a country as well as inflame debates around when a fetus acquires the right to life. The situations leading to pregnancy, the availability of safe abortion services, and the legal definitions of the fetus are entangled in ongoing political debates globally. In some places, such as the United States and the Caribbean, this entanglement is combined with the increasing use of the morning after pill, “virgin clubs” (i.e., young people committed to saving sex for marriage), and the return of traditional marriages to recast societal norms around families and households. In other places, such as the Sudan and Central African Republic, death and chronic illness as a result of repeated unwanted pregnancies – often in places of civil war and disasters, with the subsequent displacement of domestic life – also have the effect of shifting social norms around families and households. So too has the high incidence of HIV/AIDS, with a growing number of deaths of young adults, especially women, threatening to radically alter notions of family and household because different members of the household become the ones whose fertility is most sought after – especially young uninfected girls. Combined with practices such as female genital mutilation, sex work, and the cross-country adoption of the girl child, the examination of fertility from a geographic perspective is sorely wanting.

In the global policy arena, the twentieth anniversary of the International Conference on Population and Development and its outcome, the Cairo Program for Action, is poised to promote a comprehensive sexual and reproductive health and rights agenda. This has happened even with strong arguments to return to the demographic target-setting of population processes pre-1994. The decennial international population conferences before 1994 (Bucharest in 1974, Mexico in 1984) reflected many struggles
because the fertility of women—especially from the economic south—was the primary concern. Over time the debates have moved from Malthusian-inspired theories that targeted family planning to neo-Marxist influenced theories of economic development. In 1994, however, the relationship between demographic theory and population policy, especially as it concerned women’s fertility, witnessed a remarkable ideological and political shift. The more explicit human rights focus of the Cairo Program of Action emphasized that empowerment of women is the way to reduce gender and other social inequalities. This has been a well-recognized milestone moment in both the history of population and development and in the history of women’s rights.

Few fertility studies within geography work explicitly with this framework. Instead there is an apparent comfort in adhering to conservative gender agendas and continuing to work on descriptions of population dynamics that are immune to the politics of fertility research. For instance, there are defendable empirical and political reasons for such research in South Asia where specific countries make key contributions both to global population dynamics and to geopolitical dynamics. However, when the analysis of marital unions in Europe is undertaken with little consideration given to sexuality, it is less clear why conservative advice prevails. This disinclination is constraining the potential for geographic thinking on the fertility process to contribute to critical research-based policy debates at national and international levels.

While few geographers have actively engaged with the national, regional, and/or global policy processes around the International Conferences on Population and Development, many more have taken on issues concerning HIV/AIDS and new reproductive technologies. Closely linked to patterns of mortality, the analysis of HIV/AIDS provides fresh insights into the ways in which fertility intersects with the mortality of sexually active populations. For instance, reduced fertility can be due to shifts in traditional partnering patterns, biochemical reactions to medication, or changing sociocultural values over sexual reproduction such as polygamy, slavery, bride-stealing, or sexual trafficking. These are issues that are experienced and embodied in very diverse places and provide huge possibilities for new approaches to fertility analysis.

**Conclusion**

The traditional assumption that groups of people are best understood when they are studied as collective outcomes of individual population actions is challenged by contemporary fertility research in geography. New approaches to analyzing fertility might begin by mapping or recording patterns of fertility using standard demographic techniques; however, other texts, such as maternity histories, life course analysis, analysis of religious records, and analysis of naming practices or burial practices and sites also need to be incorporated. This is not just because it is highly likely that standard techniques cannot capture complex human processes, but also because there is greater access to such texts. Drawing on interdisciplinary insights from, for instance, anthropology, history, and literature will ensure geographers are more informed of the limitations of particular texts or measurements.

Patterns of fertility are not only the result of biological processes. They also represent complex and interesting social, cultural, and political processes that create new subjects such as adopted children, orphans of war and HIV/AIDS, rural bachelors, “tiger parents” (a phrase associated with parenting within Asian–American families), and children conceived by in vitro fertilization.
or surrogacy. These new subjects are constituted differently across time and place, and many more will emerge on the global population landscape. Critical geographies of fertility have much to consider and, as with population geography in general, the analysis of fertility needs to move beyond method and concern over technique and toward asking critical questions about contemporary issues: for instance, what is the relationship between pro-natal policies in Iran and antenatal tendencies in the Ukraine? In what ways do impoverished families negotiate fertility futures for themselves and their children? What are the multiple connections between environmental degradation, menstruation, and young women’s reproductive health in boarding schools? The opportunities are wide-ranging for critical population geographers informed by new social theory, to interrogate fertility comprehensively and justly.

**SEE ALSO:** Gender; Health and development; Health and wellbeing

**Further reading**


Fiat and de facto objects

Thomas Bittner
The State University of New York at Buffalo, USA

Spatial entities are located at three-dimensional regions of space (Casati and Varzi 1999). Ontologically this category of entities is characterized by mereological (Simons 1987; Varzi 1996), mereotopological (Casati and Varzi 1994), and mereogeometrical (Borgo and Masolo 2010) properties and relations. Here the focus is on ontological makeup of fiat and bona fide spatial entities. As illustrations, three bona fide and fiat boundaries and objects in three kinds of geographic environments are discussed: built environments, large-scale geographic environments with individual objects such as mountains and valleys, and systems of regional partitions that structure the surface of the Earth. Ontologically relevant aspects of each kind of environment are discussed in the context of a prototypical running example: (i) a parking lot as an example of a human-made built environment of intermediate scale (Figure 1a); (ii) Europe as a large-scale geographic entity with natural phenomena, cultural and human resources, and a political subdivision into nation-states (Figure 1b) (Burrough and Masser 1998); and (iii) categorical classification and delineation systems (Figure 2).

Mereology, mereotopology, and mereogeometry

Mereological properties and relations characterize the ways in which wholes are made up of parts (Simons 1987). For example, parking spot 6 (P6) as well as the service booth (SB) are parts of the parking lot in Figure 1a; P6 and SB are disjoint, that is, they do not have a part in common. Part-of and disjoint are examples of mereological relations (see Simons 1987 for more examples). At the topological level there are two classes of fundamental distinctions (Casati and Varzi 1994). First, there is the distinction between entities that are singly connected (e.g., parking spots) and entities that have disconnected parts. For example, (i) there is the union of all parking spots in Figure 1a that are labeled by an odd single-digit number; or (ii) there is the United Kingdom with its multiple disconnected parts (Figure 1b). Second, there is the mereotopological distinction between parts of spatial entities that form their interior (P6 is an interior part of the parking lot) and parts that form their boundary. Boundaries demarcate spatial entities from their surrounding environments (Casati and Varzi 1994; Smith 1995); for example, a boundary separates P6 from P7, P14, P15, SR, and MR. Singly connected entities have a continuous singly connected boundary. By contrast, entities with disconnected parts are such that every mereologically maximal disconnected part has its own continuous singly connected boundary. Boundaries have parts; for example, the boundary separating P6 and P15 is a part of the boundary of P6 as well as a part of the boundary of P15. Boundaries are dependent entities (Smith 1995): there cannot be a boundary without some spatial entity to bound.

Mereogeometric properties establish the distinction of entities and their boundaries according to their shape and size (Borgo and Masolo 2010): boundaries can be straight versus curved,
and entities can be convex versus concave, spherical versus nonspherical, and so on. Scale is an inherently mereogeometric phenomenon.

Bona fide and fiat boundaries

Boundaries fall into the major categories of bona fide boundaries and fiat boundaries (Smith 1995). Bona fide boundaries are boundaries in the things themselves and exist independently of all human cognitive acts. They are a matter of qualitative differentiations or discontinuities of the underlying material reality. Examples are surfaces of extended objects such as cars, walls, and the floor of the parking lot in Figure 1a. Projections of bona fide boundaries onto the floor are marked by bold solid lines in the figure. Examples of bona fide boundaries of entities of larger scale include the boundaries that separate the mainland of the British Isles from the surrounding North Sea (Figure 1b).

There are also boundaries that do not correspond to qualitative differentiations or discontinuities of the underlying material reality and that may lie skew to boundaries of the bona fide sort. Examples include the boundaries of parking spot 6, P6, in Figure 1a, large parts of the boundary between Germany and The Netherlands, and large parts of the boundary of the Alps along the foothills (Figure 1b). Such immaterial boundaries either (i) exist by virtue of certain geometric features of their environments or (ii) are brought into existence by human fiat.

As an example of an immaterial boundary of type (i), consider the parking lot in Figure 1a. By virtue of its rectangular shape it has parts that are separated by geometric boundaries. Due to the orientation of its representation in the figure, those purely geometric parts of the parking lot can be referred to as left/right half, top/bottom.
half, lower-left quarter, top-right quarter, and so on. Similarly, the equator is a purely geometric boundary that divides the surface of the Earth into two parts: the Northern Hemisphere and the Southern Hemisphere.

Fiat boundaries, on the other hand, exist by virtue of demarcation and are affected by human cognitive acts (Smith 1995). Prototypical examples are the boundaries of the parking spots in Figure 1a. They were brought into existence by the cognitive act of a planner when dividing up the space of the parking lot into parking spots.

The classification of spatial boundaries generalizes to a classification of spatial entities (Smith 1995): bona fide objects are spatial entities that have a single topologically closed bona fide boundary. Examples include the service booth, SB, in Figure 1a, or the mainland of the United Kingdom. By contrast, fiat objects have fiat boundary parts. Examples include the parking spots in Figure 1a and the political units in Europe in Figure 1b.

Properties of fiat boundaries

Bona fide and fiat boundaries have very different properties. These include properties such as observability, co-locatability, being a barrier, and many more. While bona fide boundaries have their properties by virtue of the material structure of physical reality, the properties of fiat boundaries are often established by conventions.
FIAT AND DE FACTO OBJECTS

Observability

Fiat boundaries are not directly observable because they do not correspond to discontinuities in the underlying material reality. Therefore the location of fiat boundaries must be either inferred from the location of other observable features or determined by measurement (e.g., surveying).

Consider parking spot 15, P15. It is located in the middle of the parking lot where there are no bona fide boundaries nearby (except the floor). The parking spot, in conjunction with its boundaries, was brought into existence by a cognitive act of a planner when dividing up the space of the parking lot into parking spots. In order to make the demarcated parking spots perceivable, the back, left, and right boundaries had to be first identified by surveying procedures and were then marked in paint so that they would be directly observable. Examples of explicitly marked boundaries are the thin solid lines in Figure 1a.

In certain situations cognitive agents are able to infer the location of immaterial boundaries from the structure of the material features of a given environment. For example, the front boundaries of the parking spots in Figure 1a are not marked, but nevertheless human agents are able to infer their location as coinciding with the shortest line connecting the ends of the left and right boundaries. Similarly, the location of the upper and lower boundaries of the side road, SR, toward the main road, MR, are to be inferred. Nonmarked, and hence invisible, fiat boundaries for which location is to be inferred are marked by dashed lines in Figure 1a.

Inferring the location of the boundaries of large-scale objects is usually more difficult than for small-scale objects because (i) there are fewer cues in the environment (ignoring cases such as the Berlin Wall or parts of the border between the United States and Mexico, etc.); and (ii) boundaries of fiat objects such as “The Ruhr,” “The Paris–Brussels Axis,” “The Sunshine Coast,” and “The Alps” in Figure 1b are subject to vagueness (see “Vagueness of boundary location”).

Co-location and overlap

Two spatial entities are co-located if and only if they are located at exactly the same three- or two-dimensional region of three- or two-dimensional space at the same moment in time (Casati and Varzi 1999). It is a defining property of bona fide objects that they exclusively occupy a given location. This means that no distinct bona fide spatial objects can be co-located or overlap.

Moreover, there seem to be good reasons to assume that distinct fiat objects of the same ontological kind cannot overlap and therefore cannot share their location (not even partially) (Smith 1995). Examples that illustrate this point include my land property and your land property, your parking spot and mine, the objects forming a political subdivision such as countries, and so on.

The situation is different for fiat objects of a different kind and for fiat boundaries in general. Fiat objects of an ontologically different kind can be co-located (Casati and Varzi 1999, 1994). For example, the “City of Vienna” is co-located with the “Federal State of Vienna” of the Republic of Austria. Similarly, fiat boundaries of neighboring fiat objects of the same and of a different kind can be co-located (Casati and Varzi 1999, 1994); for example, P6 and P15 share parts of their boundary and so do P6 and MR.

Moreover, some boundary parts of bona fide and fiat objects can be co-located; for example, the back-boundary of P1 is co-located with a part of the bona fide boundary of the outer wall of
the parking lot, which faces the interior of the parking lot.

According to what was stated above, fiat objects can have bona fide boundary parts. It may be more precise to say that fiat objects have fiat boundaries, parts of which may be co-located with parts of boundaries of bona fide objects.

Temporal properties

Although physical reality changes over time, such changes happen relatively slowly. By contrast, the location and properties of fiat boundaries can change instantaneously. As an example consider the intersection of 5th Avenue and 110th Street, which are located on the Manhattan island of New York City. They are fiat objects because they have fiat boundary parts, for example, at their intersections. Moreover, both streets are spatial entities of the same kind (streets).

Fiat objects of the same kind cannot overlap. For this reason there are traffic lights at the intersection of 5th Avenue and 110th Street which determine when the intersection of the two streets is a part of 5th Avenue and when it is a part of 110th Street. Therefore, the streets do not overlap because there are no parts of them located at the same region of space at the same moment in time (see Simons (1987) for an extended discussion of the temporal properties of the parthood relation).

The temporal changes that occur when the traffic lights change are instantaneous: the mereotopological makeup of both streets changes instantaneously when one street gains a part while the other loses that part; the topological structure of one street changes instantaneously from a singly connected whole to a whole with multiple disconnected parts, while the other street undergoes a corresponding change in the opposite “direction.”

Boundaries, environments, and conventions

Fiat boundaries that are co-located with bona fide boundaries “inherit” some of the properties of the bona fide boundary they are co-located with (Smith 1995). In an obvious way, this includes the property of being a barrier to the movement of other material entities.

Immaterial boundaries by themselves cannot affect the physical degrees of freedom (such as the movement) of material entities. Nevertheless, some fiat boundaries – in conjunction with systems of conventions – seem to create affordances in a Gibsonian sense that do affect the degrees of freedom of cognitive agents. Consider your favorite parking lot. Marked fiat boundaries seem to afford agents (in cars) not to cross despite the fact that there is no physical barrier. By contrast, nonmarked boundaries seem to afford crossing in the sense that the nonmarked boundary of an empty parking spot “invites” you to cross this boundary and park your car at this spot. Presently there are no fully developed theories of how the structure of an environment affects the properties of fiat boundaries.

In many cases there are no obvious affordances in the environment that enable human agents to perceive or to infer specific properties of fiat boundaries. For this reason, humans have designed complex systems of signs and corresponding conventions that regulate the properties of fiat boundaries and thereby affect the degrees of freedom of the movement and other activities of cognitive agents. Consider the boundary of the side road (SR) in Figure 1a. Traffic rules in conjunction with stop signs require one to give way to cars on the main road, MR, when leaving the side road, that is, when crossing the boundary between side and main road. Other examples include conventions for one-way
FIAT AND DE FACTO OBJECTS

streets, posts or signs demarcating property lines, nonsmoking zones, and so on.

Vagueness of boundary location

Large-scale natural geographic objects such as mountains, hills, valleys, ridges, and capes are fiat objects because they have fiat boundary parts that demarcate them from their surroundings (Smith and Mark 2003): many mountains are separated from their surroundings by mostly fiat boundaries along their foothills, many valleys are separated by mostly fiat boundaries from adjacent hills, and many capes are separated by mostly fiat boundaries from the mainland. Such large-scale fiat objects are well defined in the sense that there are determinate facts of where the top of a mountain or the end of a cape are located. That is, there are determinate facts that identify a given mountain or a given cape. By contrast, there do not seem to be determinate facts about the exact location of the boundaries that demarcate those geographic objects from their surrounding environments. That is, the location of the boundaries of certain kinds of spatial objects of geographic scale is subject to vagueness.

The notion of vagueness is complex and there are a number of different views of what vagueness actually is (Keefe and Smith 1996). On the semantic view of spatial vagueness, certain concepts and associated language expressions provide determinate means to identify the objects that fall under the given concept, but they do not provide means to deterministically pick out the location of the boundaries of those objects. For example, it is determinate that Mount Everest (ME) falls under the concept of a mountain. Moreover there is a determinate fact about the location of the highest mountain peak on the surface of the Earth which identifies ME. By contrast, there is no determinate fact about the location of the boundaries of ME along its foothills. Similarly, there are determinate facts that identify the objects marked by the bold ellipses in Figure 1b but there are no determinate facts about the location of their boundaries.

Regional partitions

Regional partitions are collections of pairwise disjoint regions that jointly cover a given part of the surface of the Earth. Such regional partitions are created, for example, by categorical classification and delineation systems, administrative subdivisions, or purely geometric subdivisions of the surface of the Earth into raster cells.

Categorical classification and delineation systems

In categorical classification and delineation systems, regional partitions of the surface of (parts of) the Earth are created by delineating spatially (i.e., mereologically) maximal regions that are homogeneous with respect to the categories of some classification system. The boundaries that delineate homogeneous regions are located where there is inhomogeneity in the spatial distribution of the qualities categorized in the underlying classification system. Such boundaries may or may not correspond to discontinuities in the underlying material reality. This is because the underlying classification that identifies the relevant qualities may be of the fiat sort and the delineation itself may be subject to vagueness.

Consider a classification system according to which the quality of temperature is subdivided into three (bona fide) intervals: the interval below the freezing temperature of water, the interval above the boiling temperature of water,
and the interval between the freezing and boiling temperatures. Suppose further that, to provide a more detailed classification system, the last interval is further subdivided by fiat into 10 equally sized intervals. The boundaries of the corresponding partition of the surface of the Earth according to some temperature distribution during the summer months will be of the fiat sort (Figure 2b).

By contrast, the boundaries of regional partitions based on the subdivision of the surface of the Earth according to kinds of land uses (urban or built-up land, agricultural land, forest land, etc.) will be often of the bona fide sort. This is because land use affects the material makeup of the surface of the Earth and thus boundaries delineating regions of different land uses are likely to correspond to discontinuities in the underlying material reality.

The nature of the boundaries of the regional partitions created by ecosystem classification and delineation systems (e.g., Figure 2a) is to a certain degree scale-dependent. The boundaries of systems that focus on the delineation according to locally observable land-surface forms, vegetation, and so on are (ideally) of the bona fide sort (e.g., Omernik 2004). By contrast, ecosystem classification and delineation systems, which are based on subdivisions with respect to highly nonlocal climate qualities, are to a larger degree of the fiat sort (e.g., Bailey 1996).

**Administrative subdivisions**

Administrative subdivisions such as subdivisions of the surface of the Earth into postal districts, counties, federal states, states, and so on are prototypical examples of subdivisions for which boundaries have been created by fiat for political and administrative purposes. As such they are not the result of classificatory activities that result in categorical classification and delineation systems. Historically, older administrative and political boundaries tend to be aligned with discontinuities of the underlying material reality such as rivers, mountain ranges, and so on. Examples include political subdivisions into national states in Europe (Figure 1b). By contrast, more recent administrative and political boundaries tend to be purely geometric in nature and lie skew to the discontinuities of the underlying material reality (e.g., Smith 1995). Examples include many of the political boundaries in the United States (Figure 2b) and in North Africa.

**Geometric partitions**

The boundaries of many regional partitions are purely geometric in nature and are often created by measurement processes. For example, clocks partition the timeline into intervals separated by clock ticks. The spatial equivalent of the partitioning of the timeline into intervals is the subdivision of space into raster cells (Figure 2c). More specifically, the georeferencing of remotely sensed images creates (roughly raster-shaped) partitions of the surface of Earth.

**Conclusions**

The aim of this entry was to give examples that illustrate the interesting and complex relationships between the fiat boundaries that structure geographic reality and the acts of cognitive agents that bring such boundaries into existence, discover and measure their location, and possibly change their properties over time. Bona fide boundaries are directly observable and restrict the degrees of freedom of other material entities by virtue of their physical properties. By contrast, fiat boundaries are not directly observable but, nevertheless, are capable of restricting
FIAT AND DE FACTO OBJECTS

the degrees of freedom of cognitive agents in conjunction with complex systems of signs and conventions. The signs and conventions presuppose that cognitive agents are able to perceive or infer the location of fiat boundaries despite the fact that they are not directly observable.

Presently there does not exist a comprehensive theoretical framework of the complex relationships between the material components of a geographic environment, systems of signs and conventions of cognitive agents, and the location and properties of fiat boundaries that structure geographic environments. To illustrate the challenges that are faced by those who aim to develop such a framework, three classes of geographic environments of different nature and scale were discussed: (i) built environments, (ii) large-scale geographic environments with individual objects such as mountains and valleys, and (iii) systems of regional partitions that structure the surface of the Earth. The cognitive acts that bring the fiat boundaries in those environments into existence range from conceptualization to design to classification and demarcation. Vagueness is a defining and not yet well understood phenomenon that illustrates the complexity of the interrelations between cognitive activities in the form of conceptualization and classification and the corresponding fiat boundaries in the respective geographic environments.

SEE ALSO: Built environments; Cognition and spatial behavior; Ecoregions; Ontology: theoretical perspectives; Qualitative information: representation; Spatial concepts

References


Further reading


Fieldwork in human geography

Cynthia Pope
William Price
Central Connecticut State University, USA

History of fieldwork and methods through the 1960s

Until the early twentieth century, human geography fieldwork was based on exploration and describing in detail a certain area or region. It must be noted that geographic knowledge and geographic philosophies in Western geography have been heavily influenced by European individuals and thinking. Geography as a discipline may be traced back to Greek philosophers and mapmakers, such as Plato and Miletus. Learning about different global regions occurred concurrently with the Age of Exploration and the expansion of European colonial empires throughout the world. Ideas emerging from the Enlightenment and Scientific Revolution became the foundation of fieldwork during the nineteenth and early twentieth centuries and are highlighted by the influential books written by the Prussians Alexander von Humboldt and Karl Ritter in the 1800s that described international explorations.

Humboldt, in particular, is often identified as the quintessential example of an individual who personally explored the world. He collected data points on physical geography and cultures he encountered during his numerous expeditions. In doing so, he produced a voluminous body of knowledge about places that were heretofore unknown or unexplored by Europeans. Humboldt’s five-volume work Kosmos was encyclopedic for its time and sought to establish universal laws for nature, particularly in Latin America and Asia.

Observation, description, and categorization were considered objective ways of viewing the world during this period, and were integral in the development of geographic field methods, including Darwin’s work documenting the role of evolution and nature in the Galapagos Islands. This work, in the latter half of the nineteenth century influenced Carl Sauer, founder of the Berkeley School. For Sauer, observation was critical to understanding the landscape and the human–environment linkages in it. He looked specifically into the ideas of Earth and its life forms, the study of life forms in the environment, and the study of differing habitats of the Earth. Sauer argued that each place has physical qualities, known as “landscapes,” which are connected by certain types of phenomena, such as population and land use. From his observations came the term “cultural landscape,” which has been a backdrop to human geography fieldwork ever since. How did the physical environment shape cultural practices? How did a population create practices to fit in, or alter, their environment? In essence, what is the morphology of these landscapes? This idea of human–environment relations led to work in environmental determinism and possibilism, some of which, while international in nature, often veered toward racial and ethnic stereotypes – such as the work of Ellen Churchill Semple. One of Sauer's most influential works, Morphology of Landscape (1925),
shaped the international fieldwork approach utilized by many subsequent human geographers, particularly those conducting fieldwork in Latin America.

Applied geography took center stage in the discipline during the two world wars as geographers’ skills with mapmaking and their knowledge about other parts of the world were used in war efforts by different countries. With the dawn of the quantitative revolution in geography in the 1950s and 1960s, geographers integrated statistical analysis into their fieldwork to gauge international trends, such as migration. This work was generally conducted with large data sets and involved little, if any, interaction between researchers and participants. During this time period, fieldwork in human geography was at its nadir.

The cultural turn in geography starting in the 1970s

The “field” in human geography can be the international arena where one emerges oneself for months (and perhaps years) at a time, living among those being studied. It is not unusual for the researcher to become directly involved in participants’ lives and/or communities. In the last few decades, however, there has been a theoretical shift concerning how and why a place is defined as worthy of being studied. Throughout the evolution of human geography as a discipline it is important to note how the idea of “the field” has changed from a “taken-for-granted” (Gregory et al. 2009, 253) place to a subject that is complex and an object of study in its own right. Traditionally, as in all social sciences, the idea of objectivity was the goal of geographic fieldwork, stemming from a physical sciences tradition of observation and repetition of studies. In the last several decades, however, a critical geography has emerged that integrates ideas from other social sciences, humanism, and the general zeitgeist in the humanities and social sciences of the post–1960s era of community experiences, multicultural perspectives, and multiple vantage points in both shaping research topics and carrying out the projects.

This critical turn in geography introduced newer methodologies and questioned assumptions of the past about objectivity and the belief in an absolute truth. Ideas concerning fieldwork have changed from the notion of a universal truth and order to the world that could be understood through a researcher’s observation and categorization to an understanding that all knowledge is situated. Though, given the breadth of geography as a discipline, often these two ideas (one based on Enlightenment notions of rationality, reproducibility, and a natural order or truth and the competing philosophy based on situated knowledges, process, places, and the partial nature of information) are often at epistemological odds with each other.

The guiding questions of the critical turn in geography, and the accompanying field methods and internationalization of the discipline, are: Whose Science? Whose traditions? Whose knowledge? (Harding 1998; Haraway 1988). As such, a cornerstone of the post-quantitative period of research is the problematizing of the idea that an unbiased perspective is even possible, given unlimited material interactions and information. Power dynamics and subjectivity became the foundation of much of human geography inquiry and subsequent methodologies during this period.

Advances in social theories – mapping how social forces work as a whole – have led to a critique of what has come to be called the modernist view of the world. This view, rooted in the European Enlightenment and Scientific Revolution, values human rationality over emotion
and the individual over the collective. This view has also separated human society from nature in an explicit division, putting humans “outside” of nature. While this view did, and has, helped advance humanity and our understanding of the physical world in innumerable ways, it has also come under heavy scrutiny for: (i) homogenizing human experience when vast differences occur among peoples’ experiences, shaped by gender roles, racial histories, colonial histories, sexual identities, among other important factors; and (ii) denying the interconnectedness of humans and the planet (Urbanik 2012).

The integration of open-ended interviews, oral histories, documentary work, participatory action research have all taken positions of prominence in human geography fieldwork with the advent of the paradigm shift from quantitative geography in the 1950s and 1960s to more critical geographic inquiry in the 1960s and 1970s to present day. Del Casino (2009, cited in Gomez and Jones 2010) outlines the trends in human geography fieldwork in the past decades.

Humanism updates the cultural landscape tradition by critically examining “places of meaning” or “sense of place.” Instead of viewing space as a two-dimensional entity that can easily be bounded, this viewpoint expands space to reveal that particular objects are placed within a landscape in order to represent feelings and emotions, thus contributing to a “sense of place.” This “sense of place” is highly individualistic and differs across cultures and time periods. This approach leads to field researchers developing empathy with their study participants and contemplating what the research means to the individual conducting the research. Instead of being an objective, nonpartisan observer, the researcher is enmeshed in their methodology. In the humanistic tradition, oral histories of individual life stories, archival research, in-depth interviews, and cultural immersion all have a role in conducting fieldwork. As an example of the increased influence of humanistic interpretations in fieldwork, tourism studies have grown within human geography. In this subdiscipline, there is an increasing recognition of the power of the “tourist gaze,” a term popularized by John Urry (1990). Tourism geographers have utilized participatory methodologies and applied semiotics to understand the difference between how tourists and residents perceive and encounter places and the role of mediums such as the photograph and promotional material.

Structuralism, or critical realism, is another contemporary tradition that has influenced human geography fieldwork. The structuralist assumption is that structures (such as education, health insurance, political systems, and economic systems) vary across societies and locales to the extent that research is necessarily context-specific, and therefore case studies are necessary. Structuralist methodologies are varied, but there is an emphasis on ethnographic data, structured interviews, and participant observation. This work reveals the ways individuals interact with social structures and the case studies allow for concrete insight into different cultures. Structuralism is based on the assumption that individuals, both participants and researchers, have situated knowledges that shape the way they understand the world. The researchers’ “positionality,” how their own worldviews influence interpretation of data and outcomes, becomes an important consideration.

Poststructuralism is “interrogating the production and import of social discourses.” While researchers in this trajectory also examine their role in creating knowledge, they arguably go one step further in self-understanding than structuralism. The researcher her/himself becomes an object of analysis, a process termed “reflexivity.” In this perspective, the research cannot be viewed as detached from the researcher. In
FIELDWORK IN HUMAN GEOGRAPHY

this method, textual analysis, participant observation, ethnographies, and cultural immersion are essential. It is the way in which people speak and write about objects that give those objects meaning. Discourses are the socially constructed ways we use to define the world, and this way of thinking has opened the possibilities for understanding how the researcher and researched are intertwined in the research process when creating a set of meanings about the world. It also opens the research process to understanding how human and nonhuman are embedded in wider worlds of meanings and relations.

In contemporary human geography, there is an increasing recognition that working with people is an opportunity for empowering the subject. Researchers recognize the increasing interconnectedness of the world and the analyzed topics reflect this. International work and global trends have always been important for geographers. Immigration, for example, has been studied in its multiple dimensions across various cultures and working in the “in-betweenness” of sending and receiving areas, an idea that considers how migrants may not feel at home anywhere, whether that be Africans migrating to Europe, Latin Americans to North America, or Pacific Islanders to Australia and New Zealand. Trends in global health, particularly with HIV/AIDS, expanded geographers’ fields to areas of high prevalence of HIV, such as sub-Saharan Africa, the Caribbean, and Brazil, a country with a breadth of knowledge and research on the virus that has informed many North American geographers. HIV research in particular has shown how the discipline moved from mapping disease (an outcrop of the quantitative revolution) to theoretical and international investigations into the gendered and national power relations that lead to transmission.

Human geographers investigate how different cultures define and interpret such “objective” things as disease, natural resources using a variety of methods and vantage points. However, human geographers have shed light on how Eurocentric definitions of traditional research topics have been biased and need to be re-examined in a more contemporary and multicultural light. Increasingly vital in recent fieldwork is calling attention to whose geographies are traditionally not told. Thus, we see more fieldwork on the level of the body and corporeality, queer studies, and geographies of emotion. Indeed, much of the foundation of current human geography fieldwork asks the perennial question about whether the subaltern can speak, put forward by Spivak (1988). By analyzing power relationships (whether they develop through gender, age, economic, ethnicity, physical ability, health, or other factors) human geographers are ensconced in revealing the thoughts, motives, and spaces that influence those individuals or groups who were traditionally excluded from research and political and economic power. In this way, human geographers have often become involved in social justice movements whereremedying these power inequities are important.

Participatory action research has been important for researchers across the globe, nationally and locally when geographers have taken up notions of spatial injustices. This may be manifest in the unequal distribution of wealth across city neighborhoods or in the unequal access to resources at the global scale. It is this social justice motif and the quest to expose and remedy these situations that intersects many subdisciplines of human geography – such as feminist geography, queer geography, medical/health geography, sustainability studies, urban geography, and environmental geography. David Harvey’s social activist work in Baltimore and human geographers founding and participating in nongovernmental organizations (NGOs) around the world highlight the intersection and
the blurring of boundaries between researchers and participants. It is often an impulse to improve the living conditions of the researched community that leads geographers to take an active role in the participants’ lives and communities.

More and more, human geography fieldwork methodologies and research topics are cutting across subdisciplines and scales, such as urban geography, development studies, geopolitics, health geography, economic geography, GIS, and historical geography. Human geographers address global issues, such as gendered mobility, women’s travel writing, racism in the academy, tourism in the Global South, responses to environmental change, and refugees and asylum-seekers, among many others – constraining and enabling spaces, queering inquiry, and coloring research methods (Moss and Falconer Al-Hindi 2014).

Feminist perspectives question the assumption about the researcher’s ability to see a particular truth that participants cannot see, to judge the authenticity of participants, and to ascribe a kind of subjectivity to research participants. The role of gender became increasingly visible during the critical turn and continues to be one of the axes on which geographers understand power relations throughout the world. In this approach, importance lies not only in the topics and places that are studied, but also with who is doing the research. In Anglo and North American geography, editors are increasingly seeking contributions from scholars and activists with different life experiences and worldviews. A growing female student body at all levels of higher education and from diverse backgrounds has resulted in a changing dynamic in human geography fieldwork that can be seen around the world. As more women are entering the field, more research has taken on the roles of gender and sexuality and revealing situated knowledges. Increasingly, topics concerning women’s lives around the world are more common in geographic publications including, for example, research into reproductive health, gendered migration patterns, political violence, and the role of the body in society.

Human–environment relations has also been one of the main themes in fieldwork, but the perspective has changed quite a bit from the Carl Sauer and Ellen Churchill Semple period of the early twentieth century. Much contemporary fieldwork is interested in exposing struggles based on power and powerlessness at global, national, and local scales and understanding interactions between individuals and their spaces, whether that is at the geopolitical or personal level. Within the human—environmental theme, political ecology has been a framework that has cut across human geographers’ research in many international contexts.

We can see this political ecology framework at work where traditional foods are still eaten, despite the fact that their numbers have dwindled, such as turtle soup in Central America and the Amazon. Political ecology is an approach that has been particularly useful when looking at issues of nature—society relationships. From bear bile in traditional Chinese medicine to Amazonian dark earths to whale killing in the Northern Atlantic to whale hunting as a mode of cultural survival in Canadian First Nations’ lands, the link between nature and culture has long been a focus of human geography fieldwork. Geographers have used ethnography in the US–Canadian border, as well as in Scandinavia, to weave narratives about traditional whaling whereby a community wanted to revive the whale hunt as part of their cultural survival, but with contemporary weapons such as guns. This contrasted with the environmental groups who were more interested in the harm of an animal than in the cultural survival of a nation. This shows the divide, or the intricacies, that exist in fieldwork when we see...
the tensions between sustainable resource use, cultural survival, technology, and even animal geographies.

Participatory action research allows participants to record their own experiences of the world and their everyday lives through uses of mental maps, diaries, paintings, and more recently photography with disposable cameras, which has proven especially useful when dealing with the growing field of child and youth geographies. Using cameras literally changes the viewpoint of the subject into fieldwork. This has been particularly useful on work with children on the US–Mexico border. Autoethnographies are another way to understand the researchers’ stance and inherent biases that come into a project, but also allow participants themselves to describe and reflect on their lives. Indeed, the way individuals make their way through places has been important for geographers who examine disabilities and who focus on the everyday life paths of individuals. In this way, human geographers expose here a critical eye toward how humans define and use their space, particularly those who are in wheelchairs and others who navigate spaces with physical limitations in mind.

Recent and future trends in fieldwork places and methods

Fieldwork takes place in different types of environments and the definition of both words are being expanded and re-examined on a continual basis by geographers. In this section, the entry reflects on several of the current research trends in human geography fieldwork. Being physically in the field, whether that means the university campus, the local neighborhood, national parks, or in international arenas, has always been a cornerstone of human geography fieldwork and exploration. In the last couple of decades we have also seen a growing confidence in using secondary data sets in creative ways to describe how humans interact with their environment. Although these data sets traditionally have been used in quantitative analysis and associated more with perhaps physical geography or even with demography, by using these data sets in a visual sense (vis-à-vis geographic information systems) geographers have the capacity to be in the space of the home or office and use the data in a creative way to enhance the themes of human geography fieldwork. Indeed, “while often associated with standard analytical techniques, secondary databases might useful be thought of as vast territories to be explored, visualized, and understood using new critical and creative approaches” (St Martin and Pavlovskaya 2010, 191).

Human geography research has taken a qualitative turn, but there is now a movement to integrate newer technologies, such as GIS, to combine with traditional qualitative methods to create a more nuanced interpretation of various situations. GIS has evolved from a set of technical tools into an integral part of mixed methods in fieldwork and is utilized to examine topics that are much more cultural and qualitative than ever before. We are at a point in human geography fieldwork where varying topics at various scales can be examined by integrating the cultural trends that qualitative methods can reveal through the heavily data-driven methods of GIS and statistical analysis. Geographers, including Kwan (2002), are examining how traditional quantitative methods can be used to highlight trends that can actually reveal power relations, which are typically the domain of quantitative field methods. These traditional quantitative methods include examining secondary databases (such as national census data) that can be used to create data layers in
GIS to highlight spatial correlations between phenomena. The participatory GIS method underscores the increasingly blurred boundaries between applied research and participatory action research. Examples include helping tribes to map traditional boundaries and settle longstanding disputes (Ghana, for example) to understanding the role of racial profiling in police stops in Canada.

For example, feminist geographers have been at the forefront of examining women’s everyday lives as an aspect of the field. The field can be the individual’s experience with space and place and the emotions that one develops in reaction to the world around him or her. Thus, this inner space of emotions becomes the field, in essence, and the fieldwork techniques are centered around trying to expose those private emotions in a more public platform. New spaces, such as those exposed in medical ultrasounds, reveal the body and society interaction at the most personal scale. Geographies of emotion highlight an openness in the field to expanding boundaries of what are considered “acceptable” topics for field research. In particular, the spaces of the body, whether how the individual body moves through or reacts to space or how large-scale political processes affect individual bodies (through migration policies, ethnic segregation, war) is apparent. However, the individual scale has been taken to an extreme and looks at the interior of the human body – at one point, looking at the emotions of a person and in another direction even looking at the space of the body that viruses inhabit.

Defining the field, and the reasons for the parameters, is an important element for geographers to consider. The field can be a piece of land or a specific place with defined boundaries, but it can also be something as amorphous as cyberspace. This new space has been used as both a focus of study for geographers and also as a means by which to elicit responses from participants to conduct research from a wide array of international sources and to show disparity in resources through a disparity in access to the Internet or to cyberspace. It is important to see how the definition of the field is evolving. It does not need to be a place with defined borders and in places that are outside the lived experiences of the researchers.

The emerging fieldwork “space” of cyberspace is increasingly important because of the notion of positionality. The researcher and the participants, when using online methods, become placeless. This can have both advantages and disadvantages. While online questionnaires can reach many people, the disembodied nature of them provides an anonymity that many research projects cannot (or do not even desire). Another way of thinking about this, however, is that this space is both an opportunity to create bonds (however superficial) with others in the same situations. The Internet is growing in popularity as a research site and is often framed as the next frontier in human subjects research. Online research presents a way to cross international boundaries and perhaps may allow access to gaining global information even for those researchers with not enough resources to travel. It also allows participants in these projects to create global linkages while being in their home communities. The Internet provides the opportunity for researchers to engage in participants’ lives from a literal distance.

This new research environment raises questions for researchers concerning: (i) interpreting politics and visibility in online spaces; (ii) researcher positionality across virtual and material study sites; and (iii) subjectivity and power in online research ethics (Oona, Hawkins, and Kern 2014). Online research opens up new ways of conceptualizing central ideas within human geography, including politicization and power.
FIELDWORK IN HUMAN GEOGRAPHY

Cyberspace is providing the opportunity for geographers to reflect on and challenge their traditional place-based stance that tends to assume face-to-face contact and “field” settings.

Another example of contemporary fieldwork issues includes climate change and sustainability. Both of these very broad topics are being examined at the local scale by human geographers in diverse places such as Micronesia, the Sahel region in Western Africa, Greenland, and even university campuses. Around the world, geographers are increasingly working with Indigenous communities to integrate traditional ecological knowledge into climate change and resource use models. Utilizing participatory methodologies and emphasizing capacity building approaches, geographers are building mutually beneficial relationships with communities. In this way, geographers seek to not only further their understanding of the mechanisms and extent of environmental change, but also to gain insight into potential human responses and adaptations, ranging from the application of sustainability principles to migration.

Human geography is growing to encompass the world beyond humans. Indeed, the field of animal geographies presents newer frontiers for defining the field, the ethics of research, and what issues merit study. This geographic perspective forces the researcher to dig even more deeply into understanding the interconnectivity of perspectives in fieldwork and question power dynamics. Many of the power relations (patriarchal, heterosexist, racist, neoliberal) that structure our daily lives in spaces have not improved dramatically for the better since the mid-1990s. As social theories have moved into a “postmodern” period that differentiates humans’ subjective experiences, so too are they moving into a “posthuman” period with challenges to notions of what has historically been seen to separate humans from animals coming almost daily from academic disciplines like biology. From the social sciences, posthumanism has emerged as a challenge to what constitutes a subject and how we might live in a world that includes not only humans and the appropriate methodologies to investigate it. Animals are now being no longer only objects to be studied and categorized by the academy, but their experiential lives and humans’ experiences with them are becoming more popular subjects of research and fieldwork (Urbanik 2012).

SEE ALSO: Cross-cultural research; Intersectionality; Mixed-method approaches; Subjectivity

References


Moss, Pamela, and Karen Falconer Al-Hindi. 2014. “Feminisms in Action: Who We are, What We Do and How We Do It. An Introduction to a Gender, Place and Culture Reader.” Gender, Place, and Culture, 1–4. http://www.tandf.co.uk/journals/


**Further Reading**


In the last three decades there have been five major edited collections giving form to the breadth and trajectory of film geography research. Uniting each of these works is the lurking reified construct of the real/reel, a binary understanding of film and media as a re-presentation of reality. With the release of Burgess and Gold’s Geography, the Media and Popular Culture in 1985 a sustained inquiry into film by geographers had begun. This was a landmark book in geography because it broke from the normative belief that media and popular culture are mere entertainment. Although this and Zonn’s book, Place Images in Media (1990), had some chapters on film, it was not until Aitken and Zonn’s Place, Power, Situation and Spectacle (1994) that a full volume was dedicated to film geography. Whereas Burgess and Gold (1985) and Zonn (1990) had unwittingly operated through a real/reel perspective, it was not until Aitken and Zonn (1994) that the binary was identified and challenged. In Engaging Film (2002), Cresswell and Dixon argued that the textual metaphor, or a hermeneutic approach, had become hegemonic in film geography. This assertion was significant because it pointed to the way that the text/context couplet that the textual metaphor relies upon reinforces the binary logic of re-presentation. In The Geography Cinema: A Cinematic World (2008), Lukinbeal and Zimmermann argue that while film geography includes research on text, it also includes a focus on film form and the affects generated by film, which provide new avenues of research beyond representation. The history of the real/reel binary sketched here provides an entry point to key debates within this subfield including hermeneutics and other supporting binaries such as material/nonmaterial, direct/indirect experience, and subject/object.

Beginning in the late 1980s and early 1990s geographers began discussing issues associated with the cultural turn and crisis of representation. Of importance to film geography is the questioning of the reification of a series of binaries ingrained in geographic thinking that control and delimit the terrain of representational studies and how film and media are positioned within them. First were the debates over the cultural turn in geography, best illustrated by the Price and Lewis dialog in 1993 in the Annals of the Association of American Geographers (hereafter Annals). Whereas those cultural geographers following the Berkeley School tradition privileged material artifacts as indicators of the transformation of natural landscape to cultural landscape, the new cultural geographers questioned the discursive construction of categories such as “natural” or “cultural,” thus shifting the conversation to landscape (and other primitives) as nonmaterial representations. It is difficult to imagine the subfield of film geography appearing if not for the passionate offense of the new cultural geographers in their then-unorthodox assertion of the legitimacy of the nonmaterial as an object of geographical inquiry. This debate between traditional and new cultural geographers was not felt only in cultural geography, however, but paralleled the normative belief throughout
geography during the 1980s that film and media
do not have the power to effect true change in
the material conditions of everyday life and are
mere entertainment, not a part of high culture.

Despite these important dialogs, the legacy
of the turn was the production of a politi-
cized binary empowered to police and regulate
knowledge production. The result of the cul-
tural turn for film geography was the reification
of the material/nonmaterial binary. Taken in
conjunction with the crisis of representation, the
cultural turn helped generate a second binary,
that between direct and indirect experiences.

The crisis of representation blurred any clear
distinction between a direct and indirect expe-
rience of the world by challenging our ability
to access a reality not mediated through cultural
ways of knowing. The roots of this binary can be
traced even further, however, to environmental
perception and humanist geography’s rebellion
against the mimetic beliefs of the quantitative
revolution. The solidification of this binary at
that time was significantly aided by Lowenthal’s
essay in the Annals in 1961, which emphasized
sensory perception as the dominant mode of
understanding. Both the material/nonmaterial
and direct/indirect binaries position film as
holding less value in terms of knowledge pro-
duction because research on film is perceived as
not dealing with things that are real or directly
experienced. Equally problematic is that they
fuel film geography’s recurring ontological
distinction between the real and the reel.

The real/reel binary is the idea that film is a
representation of reality, a mode of theorizing
that has two important outcomes. First, there
is a continual deferral of meaning produc-
tion away from film and onto seemingly more
important thematic areas like gender, sexuality,
race, colonialism, or class. Second, the emphasis
placed on epistemology by poststructuralism
reinforces hermeneutics as the dominant mode
of film analysis, thus keeping the focus of film
geography on reading narratives and exposing
underlying or hidden meaning. Hermeneutics,
or textual analysis, can be understood as the
author-text-reader (ATR) model. Research
using this approach frequently focuses on under-
standing the films’ constructed meanings and the
power relations that produce them by focusing
on one of the three given modalities.

An author-centered approach focuses on how
the text is influenced by the conditions of
production, including the filmmakers, cast and
crew, location of production, and sociohistorical
milieu. Dixon and Grimes (2004), for example,
use good and bad dialectics as a trope to explore
the production site of Tarzan’s Secret Treasure.
They argue that while a bad dialectic actively
deploys the real/reel binary, leading to essen-
tialized identities, a good dialectic sees the real
and reel as relational, leaving these categories
open. Using transactionalism and psychoanalysis,
Aitken and Zonn (1993) interrogate gender
and landscape in the films of Peter Weir. Through an
auteurist reading they find that these films res-
sonate with themes of ecofeminism and how class
and gender construct cultural identity. Lukinbeal
(2012) explores the production of the cinematic
landscape of San Diego using spatial analysis and
performance theory. Rather than an examination
of a single text, Lukinbeal shows how a land-
scape is a palimpsest, continually produced by
individuals working in the regional film industry.

A text-centered approach critiques the produc-
tion of reality and seeks to uncover naturalized
ideological claims. Natter and Jones (1993),
for instance, investigate how class and power
lead to the imaging of public/private spaces in
Michael Moore’s Roger and Me. Using a dialec-
tic understanding of the real/reel binary they
contest the notion of objectivity and narrative
neutrality in documentary films. Lukinbeal
and Aitken (1998) explore the power relations
embedded in masculinity through films by Gus van Sant. Through a framework of psycho-analysis and scale they show how masculinity is a battleground where hysteria is ensconced. Dando (2005) examines the film *Boys Don’t Cry*, looking at how gendered landscapes are constructed, policed, and transgressed. She shows that the plains are more than a mere frontier zone dividing civilization from the wilderness. The plains are also a place where meaning and identity oscillate and are constantly reworked.

The reader-centered approach focuses on the diversity of reception. In their book, *The Place of the Audience*, Jancovich, Faire, and Stubbings point out four areas of film reception studies: one, the audience as market; two, the situatedness of reception; three, ethnographies of reception; and four, exhibition. The audience as market examines preference, demographics, and how the film industry conceives the audience. The situatedness of reception is archival in nature, as it explores movie reviews, marketing material, and other media to contextualize the reception of a film to a social group within a given era. Ethnographies of reception engage the everyday practices and motivations of viewing and the social activity of cinema going. Finally, exhibition studies examine the historically and geographically contingent sites of reception, emphasizing the architecture within which films are viewed. Bypassing these four approaches, Kennedy (2008) takes an autobiographical look at how film has impacted her own life, suggesting that our lives are so mediated it is difficult to distinguish between real and reel.

Inherent to reader-centered analyses are the issues of relativism and essentialism. Where essentialism appears when found meaning is granted the authority to speak for everyone’s viewpoint, relativism exposes that essentialism has been naturalized. Essentialism further underlies the assumption that there is an archetypical (male, white, young, middle class) text-centered subject that consumes a film’s meaning, when in fact there is an infinite array of possible outcomes. While the poststructuralist belief in the “death of the author” has been much discussed in theory, there is a lack of research by geographers on understanding the relativism of meaning reception.

The power of the ATR model lies in its ability to help methodologically organize an investigation of a film; however, when the ATR model is constructed through linguistics and optical theories, an epistemological trap is created that prevents one from getting outside of the text, thus reinforcing the real/reel binary. The problem with this conceptualization is that images are not text and film is more than an image. In the book *Picture Theory* (1995) W.J.T. Mitchell points out that images and texts are two different types of representations; we have very few theories to explain images. Furthermore, we are undergoing a pictorial turn, where the era of language is being swept aside by images. At issue here is how we conceive of film and cinema: when we conceptualize film as representation there is an assumed sensory understanding that privileges vision and immobility, a heterology.

A heterological viewer, or voyeur, separates the subject (the viewer) from the object (the film). Rather than conceiving the spectatorial event as a segregation between subject and object, in *Atlas of Emotion* (2002) Bruno argues for a haptic mobilization that turns the voyeur into a voyageur; the spectator is no longer chained in Plato’s cave, but free to wander. Haptics focuses on corporeality and the porous boundary between inside and outside, the skin, thereby repositioning our attention to the sense of touching and being touched. This redirection occurs not merely as an add-on to vision, but as an emotional resonance affecting the body. Etymologically, emotion can be traced to both feelings
and movement. A focus on optics positions the viewing subject as occupying a reel or real space, which produces an indexical relationship between image and reality. The move from optics to haptics mobilizes the spectator and shifts the attention from sightseeing to site-seeing. The focus on site is not about a real signified grounding the reel, but rather is about a reconfiguration of place through embodied movement.

SEE ALSO: Cultural turn; Poststructuralism/poststructural geographies

References


Finance and the financial system are key components in the workings of a capitalist market economy and of capitalist firms. The absence of finance or an inefficient financial system may be detrimental to economic growth and development (Levine 1997). This is relevant to economies across the income spectrum from poor to rich. Over the last few decades, developing countries have undergone considerable financial development, which includes establishing securities markets and growing the banking system, often at the behest of multilateral agencies like the World Bank and the International Monetary Fund. While such development has had positive effects, it has also brought problems due to inadequate institutional supports (e.g., supervisory capabilities). Keeping in line with the removal of borders to international trade in goods and services, developing countries have become more integrated in global financial markets, opening themselves to greater flows of foreign capital and allowing greater provision of financial services by foreign financial institutions. Again, while some of this integration has brought positive benefits to developing countries, often to the urban middle classes and economic elites, it has also brought growth-reducing crises, which affect the poor the most. Microfinance has grown exponentially in the last decade, but many of the world’s poorest still experience financial exclusion, and whether the access that is provided produces better developmental outcomes is debatable.

This entry first considers debates over the direction of causality of finance in driving economic growth and development, or rather, if finance leads or follows entrepreneurial activity. The second section discusses global financial integration and convergence of financial systems and the implications for economic growth and development, while covering the rise of Islamic finance and sovereign wealth funds. The entry concludes with a discussion of finance and the poor, and issues of financial exclusion.

Does finance lead or follow development?

That finance and the financial system is an important component of any developed market economy is not a question of debate. Yet, at what point in the chain of development finance and the financial system enter is debatable. For some, finance leads development. For others, finance follows development. The former point of view, often attributed to Joseph Schumpeter, contends that financial intermediaries, such as investment banks, identify and finance entrepreneurs and technological changes that lead to economic growth and development. In that respect, the quality and sophistication of the financial services industry as a real sector is important. Financial intermediaries emerge in advance of the demand for their services, assets, and liabilities. The industry is hence an active participant in economic growth and change, as it researches, identifies, and finances the most promising sectors, firms, and managers, while providing
the critical functions of savings mobilization, risk management (e.g., portfolio diversification), and monitoring and control. In effect, this view assumes that finance is supply-leading.

The latter point of view, often attributed to the post-Keynesian economist Joan Robinson, contends that the financial system simply follows in the wake of enterprising firms, providing a conduit through which savings can be transferred from slow-growth sectors to high-growth sectors. Put simply, the financial system is responsive to the demand for its services. As entrepreneurs and firms identify new opportunities, which cannot be financed with their own internal resources, the financial system responds by providing services to see these opportunities to fruition. In comparison to the supply-leading view, this demand-following view sees a much less important role for the financial services industry as a real sector. It is much less critical to the growth and development process. The industry is a passive actor. This does not mean the financial system and financial intermediaries are not important, however. Absent a functioning financial system and financial intermediaries, growth and development could be constrained, as entrepreneurs and firms are unable to realize promising opportunities.

In reality, it is unlikely that financial systems, particularly in more developed economies, are either characteristically supply-leading or characteristically demand-following (Patrick 1966). More developed financial systems are likely to demonstrate characteristics of both, where their interaction drives and sustains development. For example, consider that supply-leading finance may identify promising but untried opportunities that once they take hold are sustained by demand-following finance. The latter is crucial in channeling resources from slow-growth sectors to the new higher-growth sector, or from economically mature regions and countries to high-growth emerging regions and countries. Even if demand-following finance continues as a passive vector sustaining economic growth and development, this does not mean that supply-leading finance is not looking for new opportunities and growth industries it can support.

In no case on the spectrum of supply-leading to demand-following finance is growth and development guaranteed, however. On the supply-leading side of finance, intermediaries may not be very good at identifying the most promising sectors, firms, and entrepreneurs. Just like any other real sector in an economy, they may underperform. This may be due to any number of reasons, from the expertise and experience of individuals working in the sector, to excessive risk aversion or excessive risk-taking. On the demand-following side of finance, an underdeveloped financial system may have high transaction costs, and asymmetric information problems may not be sufficiently mitigated and managed such that savers may be reluctant to provide resources to borrowers. In other words, savers may not trust the system to operate effectively and efficiently enough to part with their savings.

Even if this dichotomy is not so strict in practice, especially in developed regions of the world, the causal significance of finance as an initiator of growth and development remains uncertain. As the above argument suggests, supply-leading and demand-following finance work in parallel in cumulative causation reinforcing the economy’s growth and development. But for developing countries and regions with undeveloped financial systems and capital scarcity (or unaccounted for economic resources), it is not clear if finance should be a leading vector of development, or if it is, in what form, and with what supports. In other words, financial development may in theory boost growth and development, but such development may require progress elsewhere in
a country or region’s institutional development (McKinnon 1991).

For example, a stock market may affect economic development through liquidity creation; investors can buy and sell shares of companies, as and when they require their savings, yet the companies benefit from the permanent access to the capital raised through the issuance of those shares. This may be the case in developed countries, where markets are underwritten by strong regulations and the rule of law. Indeed, the largest and most liquid capital markets, such as in the United States, have very strong legal protections for minority shareholders. Firms and other entities, such as governments, that issue debt or equity securities are required to disclose regular information on their activities, and management and the directors of companies are subject to particular legally defined fiduciary duties or codes of conduct to ensure they act in the best interests of the company. Although malfeasance and misfeasance may still occur and do occur, the existence of strong investor protections underwrites market efficiency, building trust among participants and therefore liquidity. Consequently for firms, the cost of capital is likely to be lower, making investment opportunities and projects more feasible.

In a developing country or region, however, this beneficial dynamic may not hold, as regulations and enforcement mechanisms that underwrite an efficient stock market may not be well developed. The stock market may therefore be more of a casino rather than a mechanism for transferring economic resources across space and time. Because of weak institutions, firms may face a higher cost of capital and investors may shun only but the most seasoned and well-known firms. Hence, stock market development does not produce widespread development, but rather supports existing concentrations of capital and power. Moreover, weak institutions may lead to greater market volatility and asset price bubbles that inflict macroeconomic harm on the real economy once they burst. In effect, financial development, particularly when not accompanied by wider institutional development, may be detrimental to development, causing more harm than good. While not applicable to all, this was the experience of many developing countries in the 1980s and 1990s, as governments heeded the advice at the core of the so-called Washington Consensus that market-based financial development and liberalization would boost growth. Yet, insufficient development of lending expertise, mechanisms for monitoring, and regulatory and supervisory capabilities eroded the potential benefits of financial development. If premature financial development at home has clear domestic consequences, the liberalization of cross-border capital movements during the period brought further problems for the unprepared.

The globalization of finance

Since the early 1990s cross-border capital flows have grown increasingly large, particularly among advanced economies but also including emerging market economies (Lane and Milesi-Ferretti 2008). For the former, much of this growth is attributable to European integration. For the latter, this is a reflection of the removal of barriers to cross-border capital, as emerging markets become increasingly integrated into the global economy and multilateral agreements on trade and capital movements. In effect, the growth of global financial relations reflects the expansion of global trade and economic integration. As such, capital markets have become increasingly global in scope, offering opportunities, in particular, to large multinational firms and other large capital users, but also increasing opportunities for diversification for savers. But even if the benefits are
clear for some, particularly those in more developed economies, most of the world’s poor remain excluded and in some cases adversely affected.

The removal of barriers to the flow of international capital has also coincided with the removal of operational barriers for the financial services industry. Large investment and commercial banks can more freely operate in other jurisdictions, establishing subsidiaries that bring greater competition to previously protected financial services providers. In some countries, such as many of the postsocialist countries of Eastern Europe, foreign banks dominate the local retail and wholesale provision of financial services. For these countries, the entrance of foreign financial services providers and foreign capital beginning in the 1990s helped spur local economic development, as accessing capital became easier. But problems have arisen, particularly at the retail level, where consumers took out mortgages denominated in foreign currencies (e.g., Swiss franc, euro). A collapse of local currencies has constrained the ability to repay foreign-denominated mortgages, undermining the developmental outcomes from increased foreign capital and access to foreign financial services providers.

With the globalization of the financial services industry, common financial practices have become normalized across countries, thus lessening previous national distinctiveness in the provision of financial services. National distinctiveness has further eroded as banking and accounting standards have become increasingly harmonized at the international level. The most prominent being the Basel Accords on banking regulation and the International Financial Reporting Standards, which are required or permitted in over 100 countries worldwide. The latter facilitate international investment, as investors are better able to compare firms from different jurisdictions.

Despite such harmonization, there are important pockets of dissenting activity. Although representing a very small portion of global financial assets (less than 1%), Islamic banking and finance (IBF) has grown in recent years. IBF involves various restrictions on usury, speculation, excessive uncertainty, and activities inconsistent with Shari’a law (e.g., alcohol consumption). Moreover, there is an emphasis on sharing risks and benefits, and on procedural and contractual transparency (Pollard and Samers 2013). If small, IBF, like other nonconventional yet comparably small financial activity (e.g., socially responsible investment, ethical investment), signals alternative possibilities for how and through which means economic resources are managed and transferred across time and space. Put slightly differently, such nonconventional modes of finance emphasize that finance should be socially useful – a point that was laid bare in the recent global financial crisis.

In theory, the removal of barriers to cross-border capital flows and the globalization of the financial services industry, all underwritten by multilateral agreements and the normalization of financial and corporate accounting practices, are supposed to facilitate the transfer of economic resources across space and time, where capital flows to those regions that are growing and need it most, while increasing opportunities for diversification and thus risk management globally. As such, the integration and convergence of financial systems is supposed to facilitate growth and development, igniting opportunities through supply-leading finance and providing needed capital and demand-following financial services to rapidly growing emerging market economies. It would be inaccurate to contend that this has not occurred, but the increasing liberalization of global capital flows has not been without negative consequences. Capital has not necessarily flowed to areas where it is most
needed; many low-income developing countries have not derived many benefits, and crises have resulted. It has also facilitated the growth of offshore tax havens, which deprive governments, rich and poor, of tax revenues that could go to supporting economic development (Sarre 2007).

What has been learned through increasing liberalization at the global level and the liberalization of financial systems locally, particularly in developing countries, is that growth and development is not necessarily guaranteed. The deepening of credit provision and access to financial intermediation in developing countries does spur development, but it must be underwritten by adequate regulatory oversight, effective private sector credit evaluation capabilities, and a robust legal system and accounting infrastructure. Rapid or excessive financial deepening weakened banking systems and brought inflationary pressures, which contributed to growth-inhibiting financial crises (Rousseau and Wachtel 2011). In effect, without well-developed institutions the finance-growth nexus is weakened significantly. The global financial crisis of 2008, which was a developed world financial crisis, demonstrated that this principle applies to developed economies as well. Although the causes are numerous, the crisis occurred arguably because of institutional weakening and a passive attitude to financial regulation in the largest capital markets of the world.

For developing economies, the liberalization of global capital flows has posed problems because of so-called hot money. Hot money refers to short-term speculative capital flows from low interest rate yielding countries to high interest rate yielding countries. As this money can move quickly in or out of a country, it can cause instability and ultimately harm to a country’s economy. For example, if hot money leads to a rapid increase in a country’s currency in foreign exchange markets, this can have a negative impact on a country’s ability to export. Hot money can also contribute to an asset boom, by supplying excess liquidity. Once the hot money begins to flow outward, the asset boom collapses, which can lead to a widespread financial crisis. This is what happened, for example, in the 1997 East Asian Financial Crisis.

In response to that crisis, many countries in the region have accumulated vast foreign exchange reserves, partly reflecting active management of their currencies to support export-led growth, and an effort to insure against future balance of payments crises. While much of these reserves is held at central banks, governments have increasingly funneled capital into separate government investment management agencies, which now are referred to as sovereign wealth funds (SWF). One partial reason many countries have established stabilization funds, a form of SWF, is to limit the need to access external help, in the event of a macroeconomic shock or crisis, from multilateral institutions such as the International Monetary Fund, as such external intervention comes with conditionalities requiring macroeconomic and fiscal reforms that may be overly harsh, potentially unsuitable for local conditions, or politically undesirable. Importantly, these government-owned or parastatal investment funds are increasingly taking a seat at the table, making acquisitions in major Western firms, buying resource assets in other countries, or financing important infrastructure projects. They are doing this in part to earn much higher investment returns than are possible by holding cash or government bonds, usually US dollars and US Treasuries. Many of these SWFs have set up offices in the world’s major financial centers and they are employing experienced asset managers. Some SWFs – particularly the more sophisticated of them – cannot be thought of as yet another government agency or just a department in a country’s central bank. There are now more
than 60 SWFs in operation around the world, with more than 6 trillion dollars in assets under management (Clark, Dixon, and Monk 2013).

Although some SWFs are created from the surpluses of export-led growth, most are based on commodity revenues. The largest of these are associated with oil and gas producers in the Middle East and Norway. As commodity prices have boomed over the last decade, more and more resource-rich developing countries are establishing SWFs. For instance, in the coming decade, Africa may become the largest sponsor of SWFs on the planet. Estimates vary, but the number of funds in Africa alone range from the low to mid-teens, with sponsors from Algeria and Botswana to São Tomé & Príncipe and Sudan. Add to that the number of countries considering or constructing new SWFs, such as Angola, Ghana, Mauritius, Mozambique, Nigeria, Rwanda, South Africa, and Zimbabwe, and Africa could soon be home to upwards of 20 or more SWFs. Given the financial resources that SWFs have, and that many come from emerging and developing economies, the pressure to focus on development (broadly speaking) is undoubtedly high. On the one hand, an SWF can facilitate development within its home country by investing in infrastructure projects, investing in industrial development to help diversify the economy, and smoothing the effects of economic crises by acting as a lender of last resort. For resource-rich countries, in particular, the SWF can help provide stability to the often volatile revenue streams that the government receives from natural resource rents.

Nonetheless, SWFs are not a panacea for the myriad challenges facing resource-rich developing countries. While these special-purpose vehicles may provide some optimism to the resource wealthy, as a means of mitigating the so-called resource curse (see Ross 2012) by managing the volatility of government spending or even directing dividends to individual citizens, SWFs are not a replacement for broader institutional development. Such optimism for the potential of SWFs in resource management should be tempered, moreover, because implementing and sustaining the SWF solution itself is inherently complex. On the one hand, governing, managing, and operating an SWF as an institutional investor or as an agency charged with delegating asset management to external private sector fund managers requires expertise and organizational capabilities not necessarily readily available in many developing countries. Even if such challenges were overcome, an SWF, reflecting existing structures of political authority in a country, may serve the interests of the economic and political elite, providing little by way of social and economic progress to the wider citizenry. Instead of overcoming the resource curse, the SWF simply reinforces it.

Financial exclusion and microfinance

Even if finance appears to have a ubiquitous presence in developed economies and urban life in particular, there are still serious issues of financial exclusion (Leyshon, French, and Signoretta 2008). For many of the world’s poor, basic financial services such as savings accounts or access to credit do not exist. As a result, the poor are forced to rely on informal means of financing, which can often be exploitative or have excessively high costs. According to the World Bank Global Financial Inclusion Database, roughly 2.5 billion adults are without access to a formal bank account. Even though most of these are in developing countries, financial exclusion does exist in high-income countries as well. Financial exclusion often occurs in rural areas, where provision of financial services may be sparse. But many urban poor can face financial exclusion, even though there is likely to be greater provision
of banking and other financial services. Likewise, there are disparities with age and gender, where older individuals make use of formal finance, partly as a function of having greater income and wealth, and men have greater access than women by 6–9% (Demirguc-Kunt and Klapper 2012).

The causes of financial exclusion are largely structural. Banking for the poor runs up against issues of scale. For banks, servicing the poor may not be cost effective, as the savings and loans involved are small and the number of individuals large. In the case of the rural poor, there is the added cost of branch banking, which explains an urban bias in the availability of financial services to the poor. As formal finance has been unavailable to the poor, particularly in developing countries, the poor typically have relied on informal financial arrangements, which can be classed as either autonomous or reactive. The autonomous sector includes pawnbrokers, indigenous bankers, and mutual, or club-based, arrangements like ROSCAS (rotating savings and credit associations), where a group of individuals comes together for a fixed period of time to save or borrow together. The reactive sector arises in reaction to deficiencies in and control over formal finance. This sector includes private finance companies, urban curb markets, and other fringe entities. Despite the apparent dichotomy between formal and informal finance, the line between the two is actually blurred, as informal arrangements have become institutionalized in developing countries, not least because informal finance often accounts for a large share of outstanding household debt.

The revolution in microfinance, which is reaching more and more of the world’s poor, is in part attributable to the work of Muhammad Yunus, an economics professor in Bangladesh who began making small loans to villagers in the 1970s. From this experience Yunus founded Grameen Bank, which works on the principle that making loans to the poor is preferable to charity as a means of disrupting poverty. The Grameen model has been replicated around the world, and even in developed economies (Armendariz and Morduch 2010). Different and comparable models have also developed, such as Kiva, a nonprofit organization that leverages the Internet and a worldwide network of microfinance institutions. Mobile phone banking, such as M-Pesa in Kenya, has helped open up access for the financially excluded in some countries where mobile phone ownership is high. The latter example is evidence of the formal sector making greater efforts to reach the poor. Although microfinance has grown in popularity around the world, it is not without its critics. Some see microfinance as another neoliberal fantasy. Typical arguments are that microfinance institutions are not self-financing; free of any charity as claimed; that loans are taken to cover consumption rather than to make investments in productive activities; and, that the small enterprises financed through microlending are not successful over the long term (Bateman and Chang 2012).

Despite the structural causes of financial exclusion relating to the cost of provision, the political and institutional causes are significant. Financial exclusion reflects wider social and economic exclusion. Irregular work and below-subsistence income, for example, reduce demand for financial services. Likewise, illiteracy and inadequate formal education are further barriers to accessing and understanding financial products and services. This does not mean that financial inclusion is a worthless effort. Rather, financial inclusion is likely only to be successful in the context of wider social and economic inclusion.

SEE ALSO: Corporate financialization; Development; Financial geography;
Globalization; Institutions and development; Newly industrializing economies (NIEs); Regional finance

References


Today, the existence and management of the natural environment are increasingly integrated into the systems of financial capital. Finance refers both to money (lent and/or invested, i.e., “capital”) and to the allocation of assets and liabilities over time. Finance plays a significant role in sustaining the global economy by mobilizing capital for energy, infrastructure, manufacturing, and technological innovation. As a consequence, patterns of capital investment amongst governments, companies, and consumers influence the environmental consequences of economic life.

Christopher Wright (2013) provides an excellent summary of the conventional ways in which the natural environment and financial capital are intertwined. The globalization of finance coupled with innovations in securitization (the financial practice of pooling various types of contractual debt, such as mortgages or credit card debt obligations, and selling their related cash flows to third-party investors as securities) and technology has given rise to new forms of environmental investment as well as increased financial instability and short-term investing. Derivatives (financial contracts which derive value from the performance of an asset or commodity) have the potential to undermine sustainable land and resource use by distancing shareholders of capital from environmental impacts. While corporate environmental reporting and dedicated environmental investment practices help mitigate the negative impacts of financial transactions on the environment, these developments have created an uncertain and unstable condition in which governments, companies, and investors struggle to make long-term decisions and plan for the future (Wright 2013).

A relatively recent trend in the relationship between finance and the environment is the practice of using the environment as a means or medium of financialization. For the sake of clarity, financialization is defined here as the process of translating value that is exchanged (whether tangible or intangible) into financial instruments or derivatives of financial instruments. As such, financialization can be considered a process of wealth accumulation that derives profit from financial channels (the circulation and exchange of capital) in addition to commerce or commodity production. The financialization of the natural environment is commonly referred to as environmental finance, wherein the existence and quality of natural resources are priced and exchanged as instruments in financial systems.

Economic productivity and value in relation to environmental finance

The contemporary role of finance in marshaling capital for the material transformation of the natural environment – and thus the use of the environment to generate economic growth and productivity – raises the question of the ultimate source of value in economic productivity. This is a widely debated topic between several streams of thought from neo-Marxist political ecology
to environmental history (studies of human and environment history coming predominantly from environmental sociology and social studies of science). An overview of some of these debates can be found elsewhere (Mol and Spaargaren 2000). Because value plays a central role in Marxist economic thinking, it serves as a useful starting point to consider the issue of the relationship between the environment and value. Scholars following from Marx assert that the source of value in economic productivity is labor (i.e., the fruits of workers’ activities for their employers). From this simple premise three important points of departure regarding the relationship between Marxist thinking on value and the environment can be derived: historical materialism, the general laws of production, and the issue of metabolic rift.

Historical materialism is the idea that human history is the story of the evolution of the “means of production” (the physical, nonhuman inputs used in production, such as machinery, tools, factories, infrastructure, and natural capital) and the “relations of production” (relations between the owners of the means of production and the laborers). Specifically, with the concept “historical materialism” Marx argued that the means of production (the development of new technologies and greater productive capacity) eventually advance to the point where they revolutionize the relations of production, transforming societies from slavery to serfdom, to capitalism (Marx 1993; Marx and Engels 1965). The materiality of changing infrastructure, technologies, and even natural capital such as source of energy, transforms the relationship between owners of capital and laborers and the system of production, all the while advancing the privatization of the means of production (land, labor, and capital). The history of the evolution of production from serfdom to capitalism is not linear but, rather, results from the interaction and evolution of the means and relations of production. From Marx’s perspective, a historical understanding of the evolution of political economy therefore requires an understanding of the transformation of the social relations of production, particularly the power relationships between the owners of the means of production and the suppliers of labor (Marx 1993). Surplus accumulation accrues from the transformation of commodities according to the formula M–C–M′: money (M) transformed into commodities (C) transformed (through the labor of the proletariat) into surplus value (M′) (Marx 1867). The wealth of capitalism, and for that matter every other system of production, is, therefore, built through social labor. For this reason, a primary focus in Marx’s later works (Capital) is the social relations of production.

Political ecologists and environmental scholars argue against this overemphasis on social relations, pointing out there is an underlying metabolism of natural resources in production that is also essential to the generation of profit – an issue Marx himself acknowledged. For one thing, commodities are themselves derived from natural resources. Although Marx rejected natural determinism as an explanation for the function of capitalism (Harvey 2011), he did directly address the decline of the natural environment with concepts like “the metabolic rift” – the idea that increasing the productivity of large-scale agriculture and industry ultimately leads to a decline of the long-term fertility and productivity of the soil (Foster 1999). However, the ways in which ecological conditions impact the evolution of the forces and relations of production, as well as the implications of production on environmental decline, were secondary to the concerns and issues of the social relations of production in Marx’s scholarship.
Environmental historians such as William Cronon (1990, 1993) take issue with the overemphasis on labor in neo-Marxist conceptions of the production of value and seek instead to integrate both the natural environment and cultures of ecological interaction into the understanding of production and consumption decisions at various points in time. Cronon (1993), in particular, highlights the ways societies make production decisions based on both the ecology of the environment in which they live and culturally driven perceptions of society–environment relations. He argues that the evolution of social history can be understood from the standpoint of “modes of consumption” (the varied ways things are consumed) rather than modes of production (the varied ways things are produced). The value of economic processes is a product of both socioeconomic relations of production and socio–natural relations of consumption (Cronon 1990).

While neo-Marxists tend to focus on the social nature of value and ecologists on natural and physical manifestations of value, both describe elements of a common material phenomenon: the cycling of value through various economic spheres and processes. Additionally, both are concerned with value creation and the ways in which capitalism and financialization transform social and natural materiality. The two different views of materialism can be perhaps reconciled with the introduction of a typology that considers the space and time of production and consumption processes as well as their material constitution (Figure 1).

The matrix in Figure 1 represents the spatial and temporal dynamics of four distinct types of value (use, exchange, derived, and external). On the horizontal access time is divided between the present and the future. The vertical axis divides space into socioeconomic – symbolizing social interaction and economic structuring or sociomateriality – and socioenvironmental – symbolizing the interaction of human and environmental systems or physical materiality (a detailed overview of the two concepts is given by Bansal and Knox–Hayes (2013)). The horizontal axis divides time into present (symbolic of realized economy) and future (symbolic of potential economy). The lower left quadrant, socioeconomic transaction in the present time, is where use value is created and consumption takes place. The upper left quadrant, present socioeconomic transaction, is where exchange value is created and commerce takes place. The upper right quadrant, future or potential socioeconomic transaction, is where derived value (value derived from goods and services, e.g., derivatives such as wheat options or mortgage-backed securities) is created and where finance exists. Finance derives its value from underlying resources and services (or commodities and social relations). Finally, the lower right quadrant, future or potential socioenvironmental transactions, is where external value (the value of externalities, or value “external” to present use and exchange) is created and where environmental finance is being created.

In the center of the diagram the transition of value according to the cycle of financial production is represented with black arrows (M–D–C–M′) as money (M) leveraged through derivative finance (D) to generate commodities (C) and sell these for surplus capital (M′). A new inverse cycle of the “internalization” of externalities (M–E–C–M′) leverages money (M) through the priced externality (E) to invest in commodities (C) and generate profit. Internalization is generated through environmental finance and is considered here as the process of pricing externalities to change production decisions. For example, accounting for the priced externality of carbon dioxide released from the burning of fossil fuels might lead to the production of solar
panels due to new relative price competitiveness of solar energy over conventional thermal energy combined with a carbon price. The externality (E) is introduced to represent certificates and other commodities, such as carbon credits, that are used to price externalities.

The model is useful for situating the debates between neo-Marxist scholars and environmental historians over the source of value. Although the neo-Marxian scholarship first conceptualizes the cycle of production, its concern with respect to the generation of value is fundamentally with labor relations within commerce and finance, and the attendant creation of exchange value from social labor. Put in the context of the framework outlined above, the theories from neo-Marxian scholarship fundamentally engage with the creation and cycling of exchange value. In contrast, although environmental historians are also concerned with the broader cycles of production, by emphasizing the human–environment interaction they describe the use of natural resources and the impact of consumption on the natural environment. In so doing, environmental historians primarily engage with the generation and cycling of use value, and particularly with concepts such as “modes of consumption.” Therefore, rather than being a dispute about the ultimate source of value, the debate between these scholars is about the role of different but interconnected sources of value related to different but parallel economies.

The notion of parallel economies indicates several issues at play. First, parallel economies refer to the creation of systems of externality pricing intended to balance the environmental impact of economic production. Production either in the form of \( M-C-M' \) or with the use of financial derivatives \( M-D-C-M' \) can arguably be short-circuited (as with some forms of derivatives trading) and money can be used to directly generate surplus value without ever mobilizing physical commodities \( M-D-M' \). Nevertheless, the ultimate goal is to accumulate value in finance. This surplus value is extracted from both natural resources and social labor. As Marx acknowledged with concepts like “metabolic rift,” the production cycle by its very nature depletes the natural resources and social systems upon which it relies.

The pricing of externalities is designed to balance this circuit by forcing value from finance through externalities into the creation of environmentally beneficial technologies like renewable energy. In theory the circuit of internalization should close the loop and return value from finance or surplus accumulation to improve the quality of the natural environment or at the least to improve socioenvironmental relations. In this framework, environmental finance generates a parallel economy (parallel to the economy of production) that internalizes externalities and transforms production decisions to return value to the natural environment.

The notion of parallel economies also represents the ability to extend the value of physical
resources from their present objective spatial and temporal connotation across space and time. Figure 1 illustrates both present and future or potential value with individual columns. While present value is singular (commodities and services exist in particular places and have value either in use or exchange at any given moment), future or potential value is unlimited. Figure 1 could, therefore, illustrate future value with innumerable columns moving to the right across time. Because future or potential value is subjective, within finance (and indeed even environmental finance) there exists the potential to create numerous layers of value from singular physical resources through the structuring of potential value (Knox-Hayes 2013). For example, consider a commodity such as a bushel of wheat. It can be used through consumption to generate use value (lower left quadrant). It can also be first exchanged (upper left quadrant) and then consumed to generate use value. However, as long as the value of the bushel of wheat is located with its physical existence it generates single sources of value (it can be either used or exchanged), but its value is very much in its present place and time. In contrast, the subjective value of the bushel of wheat can exist in numerous places and times. For example, finance can derive value from the wheat through the creation of a variety of “options” to buy or sell the wheat at a specified price at various points in time. Multiple options (derivatives) for the single bushel of wheat can exist simultaneously because they represent future or potential sources of value spread across time. But the options instruments themselves have financial value that can be traded in the present.

The subjective nature of future and potential value creates concerns for the accounting of financial value and, particularly, for the ways in which financial value becomes represented in present time. Modern economic systems recognize all types of value as commensurate through pricing, regardless of qualifications regarding how, when, and where value actually exists. A wheat option might generate a quantitatively similar price to a bushel of wheat, but the quality of value contained within the two as well as the consequences of that valuation is considerably different. The bushel of wheat can feed hungry people. The wheat option merely provides the opportunity for a potentially advantageous economic transaction in the future. This is not to say that a financial option is without value, but rather that its value is fundamentally different from the value of a commodity itself. Therefore, in addition to offering an understanding of the relations of various economic spheres and processes, the topology in Figure 1 also suggests that the spatial and temporal nature of value is significant. In practice, there are several challenges with the ability of externality pricing to balance the social and environmental detriments of production.

To operate effectively, environmental finance must create a truly parallel economy, of similar scale and scope to the productive economy, that internalizes the externalities of production and operates according to the specified pathway of finance with an aim toward generating environmentally beneficial commodities. The problem is that just as finance can short-circuit the cycle of production by creating paths for money to generate surplus value irrespective of underlying physical commodities, so too can environmental finance shortcut the cycle of externality pricing that should generate new technologies. Environmental finance can, instead, derive surplus value solely from the circulation of environmental externalities through financial transactions. In the subsequent sections, the structure of environmental finance is further detailed and some of the concerns for accounting the value of environmental assets are outlined.
Development of environmental finance

While natural resources have always been incorporated as productive inputs in economic systems, environmental finance has the potential to transform the way in which natural resources are valued in economic systems. Environmental finance manages natural resources through the pricing and trading of positive and negative externalities – environmental benefits or costs accrued from economic activities that are not accounted in economic transactions. For example, the burning of fossil fuels results in the emission of carbon dioxide and other greenhouse gases; these are not accounted for in the price of energy unless a carbon-pricing scheme is put in place. To price externalities, the value of environmental resources is shifted from use value – the value of using a good or service grounded in objective space and time – to external value derived from the good or service, which is value liberated in subjective space and time. The conversion of the use value of the natural environment (which is deeply embedded in physical materiality) to financial or external value (value external to the consumption or exchange of the resource) through the pricing of the externalities presents some concerns that environmental finance will ultimately serve to devalue and destabilize – rather than preserve – natural environmental systems (Knox-Hayes 2013). These concerns primarily stem from the mismatch between the scale of financial productivity that markets achieve and the material impacts they generate for natural resources. Environmental finance will undoubtedly have profound implications for energy and environment use in coming decades, particularly as financial markets are adapted to manage new sets of environmental asset classes.

Negative externalities

The idea that market mechanisms utilizing well-defined property rights could be used to price and overcome the problem of negative externalities is a theory derived from Ronald Coase (1960). Early applications of Coase’s theory of externalities, particularly the EPA’s acid rain trading program in the United States, demonstrated the ability of markets to successfully control levels of pollution from sulfur (SO$_x$) and nitrogen (NO$_x$) oxides. Under the EPA program, yearly caps on SO$_x$ and NO$_x$ were established and heavy emitters were allocated a limited set of allowances to emit of SO$_x$ and NO$_x$ (also called permits). The emitters could meet the cap either by reducing their emissions of SO$_x$ and NO$_x$ in line with the permits they had been allocated or by buying additional permits from more efficient emitters who had reduced their levels of pollution below their established targets. The acid rain trading program was seen by many to be a success in the United States because SO$_x$ levels fell. The perhaps unintended consequence of the program, however, was the rise in coal use (since scrubbers were introduced to remove SO$_x$ and NO$_x$) as electricity generation demands increased.

The general success of the acid rain trading program established the viability of using market-based approaches for addressing negative externalities. The precedent the program set established the foundations for treaties addressing climate change under the United Nations Framework Convention on Climate Change (UNFCCC). Climate change is framed as first and foremost an economic problem rooted in the failure to correctly price environmental externalities. The 1998 Kyoto Protocol introduced market mechanisms through which countries could cooperate to meet their national emissions reduction targets. Emissions trading in the Kyoto Protocol functions was along lines similar to
the trading systems set up under the acid rain trading program, wherein economic actors can buy and sell emissions permits under an overall emissions cap.

Both within and without the framework of the Kyoto Protocol, the capping and trading of emissions has become a common policy response around the globe to mitigating climate change. This policy response is designed to create parallel economic activity that internalizes the cost of externalities through the creation and transaction of priced externalities such as a carbon emissions credits (Knox-Hayes 2010; Bumpus 2011). Emissions trading systems have been set up in the European Union, in California and the northeastern United States, in Australia (although the system was repealed a few years after initiation), China, Japan, and South Korea. Additional systems are planned in the Ukraine, Turkey, Brazil, and Thailand. A comprehensive overview in the development of emissions trading schemes is given in the World Bank’s annual State and Trends of Carbon Pricing Report (Kossoy et al. 2014).

Positive externalities

In addition to mitigating greenhouse gases, carbon emissions markets also serve as demonstration markets: if greenhouse gases can be successfully regulated through market mechanisms, then so too can virtually every positive and negative externality. Pilot programs have been trialed to manage other negative externalities, including water pollution. Additionally, a series of markets to manage positive externalities such as forestry and biodiversity are being created. For example, the reducing emissions from deforestation and degradation (REDD) program is being negotiated under the UNFCCC. This program would finance the positive externalities (including carbon syncing and biodiversity preservation) of the world’s forests (particularly rainforests) with funding from emissions trading programs and private and public environmental funds. The establishment of markets to trade carbon emissions, forestry, and biodiversity demonstrate the widespread belief in the potential of environmental finance to manage positive and negative externalities.

Gretchen Daily and her colleagues (1997) coined the term ecosystem services to describe the ways in which the positive externalities generated by the world’s ecosystems are capital assets. The logic of pricing environmental externalities can be extended to suggest that the function of the biosphere itself could be priced. Ecosystems can be priced to manage the flow of vital services, including the production of goods (such as seafood and timber), and life support processes (such as water filtration and pollination) (Daily et al. 2000). Moreover, ecosystems have value in terms of the conservation of future use options (such as preserving genetic diversity for future medicinal use). Despite these potential benefits, it is challenging to directly price ecosystems. Relative to other forms of capital, ecosystems are poorly understood, scarcely monitored, and (in many cases) undergoing rapid degradation and depletion. Market prices often do not reflect the full social costs of production and most ecosystem services are presently not traded on markets. Additionally, methods of indirect revealed preference (for example, valuing clean air by comparing land rents in clean versus polluted areas) are not relevant to setting a value on the existence of certain assets (such as the satisfaction derived from contemplating the existence of a tropical rainforest or essential functions like pollination). Finally, measuring the value of ecosystem services based on avoidance of costs provides only partial indications of value, especially for services without adequate substitute.
Undeterred by these challenges, Robert Costanza and his colleagues (1997) have estimated the financial worth of the biosphere to be in the range of US$16–54 trillion per year. To put this in perspective, Costanza and his colleague highlight the fact that the global gross national product in 1997 was US$18 trillion. In other words, the unaccounted economic functions of the biosphere (the sum of positive externalities) are on a yearly average twice the size of the global economy. The pricing of ecosystem services suggests that there is tremendous untapped value to be gained from the pricing of the functions of the ecosphere. In combination with programs such as sulfur trading and carbon trading that value the avoidance of negative externalities, these efforts to price the value of ecosystem services demonstrate the scale and potential at which environmental finance could operate.

Nevertheless, the productivity arising from integrating ecosystem services is not really based on the creation of new resources but rather is based on the valuation of existing environmental resources in parallel ways (Figure 1). First, natural resources are valued in the conventional sense of economic inputs (lower left quadrant in Figure 1); second, the generation of positive and negative externalities from the environmental resources are valued (lower right quadrant in Figure 1). In other words, the valuation and integration of ecosystems services into financial systems has the potential to create parallel economies (economies of use and economies of quality of natural resources), increasing not only the scope but also the scale of financial incursion into socioenvironmental relations. For environmental finance to achieve its full potential, however, it must transform the natural environment into a new source of assets. Ecosystem service pricing demonstrates the value of environmental positive externalities and the potential for the marketization of externalities to create parallel economies that magnify financial value.

Creation of environmental assets and liabilities

The creation of environmental finance requires that the interactions of human industry with the natural environment (as either positive or negative externalities) must be accounted for as assets and liabilities. The principle of environmental finance is based on the creation of parallel economies, which generate economic activity not from the pricing of resources but rather from the pricing of positive and negative impacts on/from the existence of environmental resources. Take, for example, forestry. The first layer of financial value generated from the forest is the collection and consumption of environmental resources directly from the forest: fruit, sap, and timber (lower left quadrant in Figure 1). The second layer of financial value is derived not directly from the use of forest resources, but rather from the preservation of the future quality of the natural resources and the services they provide through the pricing of externalities (lower right quadrant Figure 1). This second layer is primarily where environmental finance operates, through the extraction of the financial value of positive and negative externalities from the forest.

Continuing with the example, leaving the trees standing rather than harvesting them for timber avoids negative externalities, such as carbon dioxide emissions from the burning of the trees, as well as soil degradation from the denuded forest floor. The standing forest also creates positive externalities, or provides ecosystem services, by sinking excess greenhouse gases, protecting the biodiversity, and maintaining the integrity of the hydrological balance within the ecosystem. In order to financially account for these positive
and negative environmental services, a complex infrastructure of governance and economic accountability must be created. The first step is to privatize – define property rights or titles for the ecosystem services and environmental liabilities – in order to commodify the existence of the natural environment (Castree 2003). The next step is to make environmental assets and liabilities internally and externally commensurate with other financial instruments. Finally, the exchange of environmental goods and services must be integrated into the operation of financial services through intermediaries and exchanges.

Privatization

Privatization is the assignment of the legal title of a specific thing to a named individual, group of individuals, or an institution. The property right or title grants the owner exclusive rights to utilize the resources of title however they choose. Privatization is essential for financialization: without it, resources cannot be freely exchanged in a market system. The establishment of titles and environmental property rights begins with the privatization of land, water, and air, and progresses to the privatization of finer aspects of land, water, air, and organic materials. Privatization of land, for example, has progressed from title to the use of the surface of the land, to the minerals contained in the land, to the species, and perhaps even genes of species, living on the land.

Privatization proceeds through several steps. First, the resource itself must be legally defined by establishing spatial and temporal boundaries. For example, the parameters of ownership of a forest must be defined spatially on the basis of the landmass on which a forest exists. Ownership of the forest may be distinguished from ownership of the land and other associated resources (minerals, water, etc.). Defining the time frame over which the forest is owned is also essential, particularly if the forest is disassociated from the land upon which it exists. The property right itself relies on social convention: the established means by which the property comes to be understood to exist and the legal infrastructure through which the right can be represented in the form of a title (MacKenzie 2009).

Once the parameters of ownership are set around an asset, the value can be defined. However, the financial value of an asset exists only in exchange or through the potential for exchange (Knox-Hayes 2013). For these reasons, Castree (2003) suggests that alienability is a critical aspect of the creation of assets and liabilities. The resources, or the facets of resources that are privatized, must in some way be alienable – isolatable – so they can be exchanged. The title that grants ownership of a certain number of hectares of forest in a specific place assumes that the hectares of forest are alienable from the broader ecosystem (land, species, hydrology) within which they exist. Specifically, the owner of the forest has the right to cut it down for timber or to extract the minerals in the soil beneath the timber because – in line with resource disaggregation – doing so is understood to have little or no impact on neighboring resources.

The process of defining environmental property rights with respect to positive and negative externalities becomes more problematic because the rights must often define something intangible beyond the environmental resource itself. For example, leaving a forest standing avoids the release of greenhouse gas emissions into the atmosphere. Emissions reduction credits can then be generated from the existence of a forest and defined as property rights. The ownership of the carbon credits is derived from the existence of a market that values the absence of the negative externality (emissions) and from the event of avoiding emissions by leaving the forest standing.
for a certain period of time. The ownership is tied into a certificate, which defines the avoided externality in terms of tons of carbon not emitted into the atmosphere (carbon emissions credits).

The credits are anchored in the spatial and temporal context of an extent forest but their value can easily be moved to another spatial and temporal frame, where they can be retired or cancelled against carbon productive activities. For example, the owner of one ton of CO₂ equivalent produced by a forest standing in Peru can exchange the credit for one ton of carbon dioxide produced by a thermal power plant in Germany (Bansal and Knox-Hayes 2013). Avoiding the negative externality (cutting down the forest) is represented by the credit, which is issued to an original owner, who then holds the right to a defined unit of emission. The carbon credit allows the owner to move the generation of a negative externality to any time or place because the carbon credit represents the capture of the reduction of an emission.

Critically, the value generated from the exchange of the emissions credit does not guarantee the preservation of the forest. Its financial value is generated from a parallel economic activity distinct from the use of the forest as a resource in and of itself. As with the forest, the value of the credit must likewise be defined and is established through a contract of exchange. The contract allows the credits to be valued as a source of exchange in advance of their production. The exchange value of the credits is derived from the establishment of a title over current and future potential activities. For example, carbon credits allow for the exchange of the positive externalities embodied in standing forests (carbon sinks) for the negative externalities embodied in productive activities (emissions from a power plant).

The privatization of externalities of environmental resources has tremendous repercussions for the use and exchange of those resources. Scholars have raised concerns that the value derived from the use of a resource (harvesting of the forest) cannot be reconciled with value derived from positive externalities of the forest left standing (Altvater 1993; Miles and Kapos 2008). There are inevitably sources of conflict from the two types of property rights. The standing forest requires that the land not be used for purposes other than to hold the forest. As such, financial value tied to the use of the land will conflict with the financial value derived from the property rights tied to positive externalities of a standing forest.

Commensuration

Commensuration is the process through which resources are made accountable by defining the resource according to a common metric (Callon 2009). Financialization requires the creation of a standard against which the resources can be compared. Castree (2003) identifies two processes of commensuration: individuation and abstraction. Individuation refers to the representation and physical act of separating a specific thing or entity from its supporting context: establishing the forest as a legally owned entity. This involves putting legal and material boundaries around the forest so that it can be bought, sold, and used. Abstraction is a process whereby the specificity of any individualized thing (an oak forest or an oak tree) is assimilated to the homogeneity of a broader type of things (a “forest” or a “tree”). Abstraction is both functional – a pine forest becomes abstracted as a “forest” – and spatial – a forest in Peru is abstracted as equivalent to a forest in any other location (Robertson 2000). The goal of commensuration is to make the commodities uniquely identifiable (individuation) but completely fungible (abstracted).
Commensuration enhances the scale at which assets and liabilities can be marketed as well as the rate at which they can be exchanged. Commensuration is achieved through the standardization of resource metrics or of the information contained in representation of the resource (such as a carbon credit or other financial certificate) (Lohmann 2009). Standardization requires agreements on the format and content of the commodities, as well as for the contracts of trade. Here the critical objective is to determine which pieces of information are included in the representation of an asset, and which pieces are not. For a pine forest to be made commensurate with an oak forest the common metrics of evaluation might be the number of hectares of land covered by the forest, the weight of the biomass, or even a metric that evaluates the capacity to sink carbon dioxide emissions. Note that equally important to what is included in the metric of evaluation is what is left out. From the standpoint of standardization in the context of climate change policy, the role of the oak forest as a habitat for particular bird species is immaterial, and thus that characteristic is stripped from the metric. Assets are made homogenous through this process of standardizing and reducing information content.

For positive and negative externalities, commensuration similarly requires the definition of metrics of evaluation and variables of equivalency. With respect to carbon credits, all of the information of the process of reducing emissions (the maintenance of a forest or the installation of a scrubber on a thermal power plant) is reduced to the metric of one ton of carbon dioxide equivalent. In this respect an oak forest can be made commensurate with a pine forest by considering the relative amount of CO2 equivalent it absorbs every year. The credits are indistinguishable from their counterparts and erase the specificity of other qualities of the forests from which they were derived.

Finally, the creation of financial instruments enhances the spatial and temporal liquidity of environmental assets (both direct and derived from externalities) by enabling them to be treated as commensurate with any other financial instrument. Environmental assets acquire a financial value that makes them exchangeable with any other financial asset. The metric is their financial value. The final stage of creating environmental assets is integrating them into financial infrastructure through intermediaries and exchanges.

Intermediation and the infrastructure of environmental finance

Intermediation is the process of transferring the value and liability associated with a good or service between two places and two individuals, a service which is usually provided by banks, brokers, and other financial service providers. Intermediation is a crucial element of the integration of environmental assets into financial services. This is accomplished through the construction of new standards and conventions, the establishment of infrastructure and systems of trade, and the creation of new intermediary firms (Knox-Hayes 2013; Labatt and White 2003). Intermediaries provide custodial services, transferring environmental assets from their origination source to the end buyer.

The valuation and trade of assets and liabilities constructed from environmental externalities require considerable social standardization and connectivity. Several types of firms and organization are essential to building services and infrastructure of environmental finance, including legal and accounting firms, aggregators, development funds, registries, banks, brokerages, and exchanges (Knox-Hayes 2009). Legal firms define environmental assets as legal commodities and tradable instruments, and are often
consulted in writing new regulatory frameworks to provide legal scope and a market for the exchange of environmental assets and liabilities. Accounting firms develop the standards and metrics through which environmental assets and liabilities are measured and accounted for. Environmental asset aggregators originate the environmental projects in partnership with leading legal firms, and aggregate the credits from these projects for use in developed countries. Development funds facilitate the implementation of environmental projects.

Additionally, a range of private intermediaries have become involved in environmental finance, particularly investment banks, which provide most of the finance to develop environmental projects. Banks hold portfolios of carbon credits and other environmental assets, which they sell to their clients and bring to market through carbon brokerages and exchanges. Environmental brokerages link buyers and sellers of the environmental assets. Environmental exchanges provide a forum for buyers and sellers to meet, but also hedge the delivery risk associated with the creation of financial assets from externalities. Registries enhance the mobility of financial assets and liabilities created from externalities, such as carbon credits, by digitally connecting producers and consumers of the credits.

As initiators, custodians and mobilizers of environmental assets, banks and other intermediaries occupy a critical junction point in the process of environmental financialization and are able to shape it in profound ways. In particular, intermediaries enable environmental assets to be structured in the same format as other financial assets, regardless of the unique characteristics of the material environment from which the environmental assets were drawn. This creates the ability to securitize environmental assets like carbon, forestry or biodiversity credits and to exchange them for other financial assets or environmental liabilities as bankers seek to accelerate the rate at which the credits produce financial value.

Avenues of investigating environmental finance

Sitting at the interface of human and environment interaction, and socio-natural and socio-economic productivity across time, scholarship on the structure and function of environmental finance is of increasing importance to the discipline of geography. Scholars approach the subject through a range of lenses, critiquing issues of valuation and accounting, the integrity of financialized environmental resources, and the hybrid nature of socioenvironmental and socio-economic systems. In addition, analysis of the spatial and temporal dynamics and interactions of finance and the environment offers the potential for new insights.

Economic geographers like Neil Smith (2008) and David Harvey (1996) criticize the production of nature, arguing that environmental commodities become a means to the end of accumulation for accumulation’s sake irrespective of the qualitative particularities of environmental commodities. Similarly, John O’Neill suggests that “the neoclassical project of attempting to cost all environmental goods in monetary terms becomes an instance of a larger expansion of market boundaries” (O’Neill 2007, 21). These studies are complemented by work from scholars that challenges the ways in which environmental resources become transformed through the process of commodification. Scholars such as Bakker (2004) have studied the demise of water as a “public good” as it is privatized and transformed into a salable commodity. Similarly, Ralph Kloppenburg (2005) demonstrates how the privatization of crop varieties limits diversity and
vitality of crops as companies gained proprietary access to hybrid crop varieties.

In general, these studies illustrate the ways in which the natural environment has been transformed from open access public goods to a set of private resources. William Boyd (2001) demonstrates this through his investigation of the poultry industry: chickens as natural entities become physically altered to suit the profitability requirements of the agri-food companies that manufacture them. Gavin Bridge (2008) highlights the ways in which economic and political institutions of commodity production and consumption shape the ecology and society of resource production, particularly the processes and cultural practices through which natural environments become enacted as resources. The work of these and other scholars suggests a need to understand the transformative capacity of environmental finance.

Additionally, there is a growing interest and need within geography to evaluate the spatial and temporal dynamics of environmental finance. One of the strengths of geography as a discipline is its ability to bring scale into analysis. There is an organic connection between space and time within capitalism, which in turn means that space–time plays an important role in the dynamics of financialization (Castree 2009). Finance is predominately motivated to accelerate the production of financial value from physical or intangible assets. Environmental finance is about the integration of natural systems into economic systems, and the acceleration of the production of value from environmental assets. It does not guarantee, or even necessarily strive towards, the protection or preservation of the natural environment. The inability of financial actors and institutions to make long-term decisions and plan for the future is critical to the challenges of environmental finance. Space and time are, therefore, central to understanding the intersection between environmental finance and other spheres and processes of economic productivity (as in Figure 1) but also critical lenses for understanding the broader function of environmental finance.

The application of financial markets to the management of environmental systems suggests dangers of distortion. Because markets account for neither information complexity nor spatial scale and temporal duration under which environmental systems operate, they produce, at best, a very partial picture of the real world value of environmental assets and can introduce temporal mismatches as natural and economic processes move at very different speeds. For example, the financial value of forests under REDD is assessed by measuring the amount of CO₂ reduced per area of forest standing (Miles and Kapos 2008). Yet, the age and diversity of the forest area do not factor into the credit. Monocultures, new growth, and old growth forests can be considered comparable in their ability to reduce carbon if they are of similar density and size. Furthermore, in reducing the forests to this commensurable metric, their history, vitality, and diversity become irrelevant to calculations of their value.

The creation of parallel economies removes the material context of the forest resource, including its use value, from the calculation of the exchange value of the financial instrument. Consideration of the spatial and temporal nature of value highlights the danger that financial products create the illusion that socially meaningful value exists without context. Financialization transforms the intrinsic and even instrumental value of the forest to financial or external value. There is a real potential for the financial product to leave the forest and its spatial and temporal complexity behind. Environmental finance creates the sense that a forest that took 10 000 years to evolve in a particular geography can be
exchanged for a forest of similar size and density that was grown in 30 years in a location of choice. Distortions in spatial and temporal scale heighten the demand for accelerated rates of resource production. While socioeconomic systems can to some extent adjust to the demands of accelerated production, the function of socioenvironmental systems is deeply embedded in space and time, and cannot easily accommodate temporal acceleration and spatial compression (Bansal and Knox-Hayes 2013). Spatial and temporal dynamics are essential to environmental finance and offer promising avenues of investigation.

SEE ALSO: Climate policy; Ecosystem services; Environment and resources, political economy of; Environmental governance; Environmental management; Environmental valuation; Financial geography; Green capitalism; Intermediaries; Time–space convergence

References


Financial geography

Karen P.Y. Lai
National University of Singapore

Financial geography is concerned with the roles of finance, money, and markets in the restructuring of contemporary capitalism. Research focuses on how such restructuring has uneven impacts across space, and also on the ways in which space and place are mobilized in the production of financial markets and processes. Some examples of the spatialities of finance include financial exclusion, regional inequalities due to financial activities, differential connectivity of global and local financial flows and networks, and the rise of new onshore and offshore financial centers. More than acting as a neutral lubricant in the economic system, financial markets, actors, and instruments shape the development and life chances of people and communities across regions and localities. While some have argued that globalization and new information and communication technologies (ICT) have rendered geography irrelevant, financial geographers contest that space and place remain crucial to the financial system in shaping financial decisions, allocation of capital, regulatory frameworks of global financial markets, and their unequal outcomes.

Economic geographers and regional scientists have long been preoccupied with industrial landscapes, especially in terms of industrial development and agglomeration. Little consideration has been given to the impact of money and finance on spatial processes and impacts. The late 1980s, however, saw a wave of studies on the spatial dimensions of money, establishing the theoretical and empirical importance of finance in shaping the economy, regional development, and place-making at different spatial scales. This partly reflected a broader research interest in service sectors in the 1980s and 1990s, following deindustrialization in many developed economies and increased attention to the role of finance and other advanced producer services in driving economic growth in cities and regions. The impetus for financial geography also arose from a critical inquiry into how financial flows and structures shape and contribute to uneven development and instability, following the excesses of the high-rolling 1980s in the financial centers of London and New York.

Initial attention was given to the regulation and governance of the international financial system and debates around the dynamic relationships between finance capital and nation-states. This political economy approach is complemented by more social and cultural explanations for analyzing the persistent and increasing importance of international financial centers (IFCs) as strategic basing points for global capital. The territorial embeddedness of specific localities in the production and reproduction of global finance is explained by particular social and cultural constructions of people and places. While initial studies analyzed individual cities, relational economic approaches became more important in explaining how and why such financial centers develop or maintain their roles and prominence as IFC networks. More in-depth studies examined the workings of specific financial markets, the influence of global financial markets on corporate governance and national policies, and broader implications of financial logics for
governing economic and social lives. More recently, financial geographers are paying more attention to emerging economies and alternative financial networks as a critical response to the domination of Anglo-American economies in financial geography research.

**Political economy approach**

Early research in financial geography explores the changing spatial structures of the international financial system and the financial industry. The role of money and finance is given center stage in examining the circulation, accumulation, and regulation of capital and their spatial implications. While advancements in transportation and ICT technologies have reduced the relative distance between places, this “shrinking” of the world is also part of the general process of capitalist commodity production and capital accumulation, as capitalist producers are locked into a perpetual race to exploit new markets and to reduce the turnover time of capital (Harvey 1982). This time–space compression reshapes relationships between space and capital, and presents new regulatory challenges with the international monetary and financial system constituting a kind of regulatory space holding together broader processes of accumulation. This perspective draws attention to geographical and structural shifts in power between states and markets in the regulation and operation of markets as the globalization of capital and ideas disrupted traditional forms of sovereignty and international regulation. Studies examine how regulatory changes, technological innovations, and the globalization of markets and monetary flows are remapping the financial landscape towards increasing deterritorialization, but states remain important arbiters in determining the shape and speed of such processes, both within their national space economy and on the international arena.

An important theme in this area of research is the notion of hegemonic stability – the idea that one powerful state will impose order on the international financial system such that it provides a platform upon which more general processes of capital accumulation follow. Studies on earlier systems of international regulation in the nineteenth and twentieth centuries largely focus on British financial capitalism and the pivotal role of the City of London in the construction of the gold standard in the nineteenth century, and the rise and fall of US financial hegemony during the postwar period. The construction and subsequent collapse of the US-led Bretton Woods system demonstrate the limits of national markets for accumulation strategies as economic activity becomes increasingly internationalized. The development of Eurodollar markets, for example, threatens the control that national governments exercised over national space–economies and symbolizes the emerging power and international mobility of private capital. This gives rise to the concept of posthegemony, whereby different state actors linked into different policy communities increasingly operate within transnational networks that comprise counterparts abroad as well as global financial institutions. In this sense, hegemonic power resides not with one territorial state but in a number of nodal points or flows. While this decentered hegemonic core is more spatially diffused than before, modes of regulation are still anchored by various actors, including supranational organizations such as the World Bank and the International Monetary Fund (IMF), large private capital, and particular communities (be they policy, finance, or other interest groups) located primarily in the United States, Europe, and Japan.

Another systemic concern is that of regulatory arbitrage. The spatially bounded nature of
regulation (often organized according to political-territorial boundaries) makes it possible, particularly with the aid of advances in ICT, for economic agents to move between different “regulatory spaces” in order to circumvent constraints associated with a particular regulatory regime. Regulatory arbitrage and reregulation are key themes as scholars analyze changing power relations between states and markets and their implications for the global financial system. With this practice of spatial-regulatory arbitrage, national economic policies become increasingly sensitive to the power of international money and orientated to the pursuit of low inflation and relative price stability in the constant bid to retain such “hot money.” State regulatory functions are further challenged by the inherent tendency of financial institutions for financial innovation, as new financial instruments open up new areas of financial activity and profits as yet unconstrained by regulatory control. The power of finance capital thus pressures national and local governments to accommodate global capital in ever more earnest ways through deregulation of financial markets and institutions. The winners of this process have been particular global cities that have become attuned to the rhythms of the global economy and adept at pinning down hypermobile capital through deregulation and specific forms of urban and economic development policies. A related strand of work explores the transformation of financial space through the growth of “offshore” financial centers (OFCs), where more permissive regulatory environments serve as havens from the restrictions that national governments place on financial activities. The linkages between onshore and offshore regulatory spaces are important in understanding the broader financial system, such as the development of international banking facilities in the United States and its relationship to offshore financial activities in the Bahamas and the Cayman Islands.

Even as the shift in power from states to markets is a key research theme, others point to the continued relevance of the regulatory role of the state in the international financial system. The state is clearly vital for creating and maintaining the physical infrastructure that is necessary for a city’s high standing within the network of global cities, but national and local state agents are also highly implicated through their contributions to the institutional infrastructure of these cities. State regulation, in particular, is an important part of “institutional thickness” – a product of the interaction and collective strategies of institutions of all kinds, including firms, financial institutions, trade associations, non- or intergovernmental organizations, and economic development agencies, as well as state and local governments – that appears to differentiate successful places and regions in the global economy. The growing importance of supranational organizations and international forums of monetary and economic coordination, such as the G7 and G20 country groups or the IMF, is also noted. Conventional understandings of state power and economic boundaries are blurred and works on geographies of money and finance become more concerned with flux, fluidity, and the geography of flows.

International financial centers and networks

The persistence, and increasing importance, of particular localities (in the form of IFCs and OFCs) as strategic basing points for global capital features prominently in research on the geographical relationships, processes, and impacts of finance. The role of economic centrality in the development of IFCs is prominent in global cities research. The transformation of the world economy since the 1980s is connected to an economic
shift towards services and finance, which has consolidated the importance of major cities as sites for certain types of activities and functions. These global cities are centers of command in the organization of the world economy, as key locations, marketplaces, and major sites of production and innovation for finance and specialized services. Localities such as New York, London, and Tokyo are leading examples of cities that have evolved into strategic sites for the development, production, and supply of innovative financial and business services. Some of the work illustrating this could be categorized as social constructivist in approach, inspired by the work of sociologists and anthropologists who treat money as a contingent and socialized phenomenon. This rejects the idea of a single, centered model of money in favor of viewing money as constituted through different networks of human practices. In order to explain why IFCs have not simply melted into a generalized and electronic “space of flows,” research highlights the influence of a wide range of social and cultural factors that contribute to their success and survival. This departs from a more abstract political economic perspective by conceptualizing financial centers as active and dynamic ensembles of networks – of institutions, individuals, machines, concepts, interpretations, and information.

Information is particularly important as the role of a financial center is to generate, capture, and interpret the vast amounts of monetary information that flows through every day. Rather than merely being physical locations of financial markets and institutions for the global finance industry, IFCs are centers of representation and chief points of surveillance, centers of expertise where product innovation and marketing are conducted, and centers of social interaction amongst finance workers. Stepping outside such networks and localities could be extremely costly in a financial world that has become increasingly complex and hypermobile. Studies of trust show how the business of finance still thrives on close interfirm and interpersonal relationships, such that physical proximity within an IFC is integral to the operation and reproduction of the financial system itself. Although formal institutions and mechanisms, such as credit rating agencies, exist to underwrite trust in finance, informal and personal relationships continue to be vital in lubricating the wheels of international finance. As the financial system becomes increasingly complex, trust relationships have also become central strategies for reducing risk through the development and maintenance of interpersonal relationships, which are more easily performed in relatively circumscribed spaces such as IFCs.

Attention to such social and cultural constructions of IFCs has important implications for understanding urban and regional development. Active branding exercises conducted by state actors and institutions have become critical to the imaging of particular global cities as having the right credentials and expertise to claim the status of IFC. The City of London, for example, consciously plays to the role of cultural authority and expertise in global financial services through a rediscovery of tradition. Such branding exercises are more than symbolic reconstruction; they have material consequences for the reconfiguration of city centers to meet the demands of capital. The developments of the London Docklands, Battery Park City in New York, and Shanghai Pudong New Area are the material expressions of a shift from organizing the city as an assortment of concrete production spaces towards visualizing it as a coherent space of financial processes, consumption, and political ideology. Studies in this vein range from the iconography of the City of London’s built environment, to analyses of media representations of financial crises, and issues of gender and
embodiment in understanding the practices and spatialities of finance.

Even as particular cities (such as London and New York) are seen as top-ranking global cities and preeminent IFCs far ahead of other competing centers, global cities must inevitably engage one another in fulfilling their functions of economic coordination and control, as their very success is sustained by their roles within a network of cities. Earlier studies divide cities into different tiers of importance according to their degree of integration into the global economy, and their roles as basing points in the spatial articulation of global production, finance, and consumption. IFCs are ranked according to variables such as number of financial headquarters and branches or volumes of financial transactions. The roles of producer services in supporting IFC growth are also noted, whereby particular cities function as sites of production for services and financial products that coordinate global economic activities. Since the early 2000s, geographers have highlighted the value of a network approach. Studies of IFCs and financial services explore more complementary forms of relations and alliances alongside the notion of hierarchy and competition between global cities. The theorization of a network of global cities inspired geographers to analyze intercity networks based upon flows and connections with the establishment of the Globalization and World Cities (GaWC) research network based at Loughborough University in the United Kingdom. These relational approaches highlight the roles of “lesser” IFCs in sustaining the overall global financial architecture, such as Frankfurt’s growth as a complementary financial center to London, the IFC network of Hong Kong, Shanghai and Beijing in Greater China, and the consistent flows of capital, labor, and knowledge that have sustained both New York and London as pre- eminent global centers (often referred to as the NY-LON connection).

An IFC network approach has further developed in recent years through connecting developments between IFCs and OFCs. This problematizes the divide between what is seen as legitimate onshore financial activities in major financial centers (located within global cities) and less regulated and slightly shady transactions handled by offshore jurisdictions (often depicted as small island economies). Such offshore jurisdictions are more important to global economic transformations and regional development than previously recognized; they have become a vital component of business and financial practice for firms, individuals, and other institutions (Coe, Lai, and Wójcik 2014). Concomitantly, offshore functions in the interest of regulatory arbitrage are also widely practiced in many IFCs, with the rise of so-called midshore financial centers that have both onshore traits (sophisticated financial markets, strong legal and regulatory systems) and offshore features (low taxes and secrecy), such as Singapore and Hong Kong.

Financial markets and financialization

The workings of specific financial markets, the influence of global financial markets on corporate governance and national policies, and the broader implications of financial logics for governing economic and social lives have become more important in the past decade. With the scaling back of state welfare functions in Anglo-American economies, the privatization of pension funds, growth of unit trusts and other investment vehicles, and their management and investments have become significant topics of research into how these institutional investors impact financial markets development, national development, and firm strategies. This is related
FINANCIAL GEOGRAPHY

to the spatialities of stock markets and how different national market structures and regulatory systems (such as in the United States, the United Kingdom, France, and Germany) have evolved in response to global financial imperatives. These strands of research highlight the increased importance of global financial markets for corporate governance, as accounting strategies have changed in response to the interests of global portfolio managers, leading, in turn, to the transformation of European corporate governance. The rise of stock markets and the changing relationship between bank financing and capital markets financing also have an important influence on corporate governance, regional development, and global production networks. As long as financial decision-making and the allocation of capital remain concentrated in major IFCs, regions and locations that lack such centers are disadvantaged in terms of access to finance, particularly in areas such as venture capital. The risky nature of the sector, and the trust-based and relational nature of venture capital investment, could lead to clustering of funding to firms and individuals located near to IFCs. This could account for spatial variations in the levels of success shown by innovative enterprises.

This leads to a more recent topic termed “financialization.” This refers to the increasing importance of financial logics in different aspects of economic, political, and social life. This diverse and rapidly expanding literature ranges from regulation theory and critical social accountancy approaches to more sociocultural perspectives. Coverage ranges from the growing and systemic nature of finance and financial engineering and their impacts on global governance, through national and firm level analyses that examine patterns of accumulation through financial channels rather than through trade or commodity production, to more individual and household level analyses of the transformation of everyday habits of savings and borrowing. Financial geographers have focused mainly on analyses of the processes and impacts of financialization on firms and regions, and have been particularly influential with respect to culturally inflected sociological research on how finance shapes everyday life within contemporary capitalist societies. Individuals are being conditioned to take on greater financial responsibilities and risks as Keynesian-style state social support is replaced by private sector variants such as personal pensions, private insurance, and investments. The transformation of everyday practices of borrowing and saving is also seen in the rise of credit card and other debts, and savings being channeled into various forms of investment products rather than conventional bank deposits. While some view this as a democratization of finance and investment to a broader public, others see it as the creation and extension of new risks with spatially uneven impacts. Instead of an abstract and monolithic entity, the financial system is reframed as a coalition of smaller constitutive ecologies, such that distinctive groupings of financial knowledge, practices, and subjectivities emerge in different places with uneven connectivity and material outcomes. This financial ecologies approach shows how households and individuals are changing their savings/investment practices and are being drawn into different financial relationships delineated by distinctive sociocultural demographics, and with unequal outcomes. The overall impact of financialization is a decisive shift in the way that financial futures and risks are conceptualized and managed by individuals and households. There is an engagement with, and acceptance of, investment in capital markets in order to secure individual financial futures. As global finance becomes increasingly influential in driving state policies, firm practices, and individual actions,
money is being made not just from efficient allocation of funds to profitable ventures in the “real economy” but by buying and selling of financial instruments. Such trends increase the vulnerability of the real economy to financial crises, and also raise questions about conceptualization of “real” and virtual/fictitious/financial economies.

**Alternative financial networks**

While research approaches and topics in financial geography have diversified over the past few decades, the geographical focus has been largely dominated by Anglo-American economies, with some attempts at incorporating certain financial markets in Europe and Japan. The scrutiny of financial markets and activities in US and UK financial centers is perhaps unsurprising given the importance of service sectors (and advanced producer services in particular) in those economies, following deindustrialization since the 1970s. However, as relational approaches to urbanization and capital accumulation become popular in studies of global city networks and IFC development, the geographical scope of financial geography begins to incorporate new nodes and networks in the global economy beyond the premier centers of London, New York, and Tokyo. With cycles of economic crises and recovery, financial markets and regions that used to be deemed unsuitable for investments are reframed as investment alternatives, offering opportunities for higher returns or as strategic assets for risk management. Such studies of “emerging markets” critically examine their discursive construction and mobilization by fund managers, analysts, rating agencies, and even nation-states as they seek ever higher returns on capital, especially when faced with recessionary pressures or slow growth environments in developed markets. The growing significance of IFCs and financial markets in East and Southeast Asia, the Middle East, and Eastern Europe has important implications for understanding the geographical scope of global finance, the entanglement of different types of financial markets and financial centers in strategies of capital accumulation, and how dominant discourses about the world economy have real impact on capital flows and their uneven distribution across countries and regions.

The worldwide rise of sovereign wealth funds (SWFs) also has important implications for understanding the complex entanglement of nation-states, geopolitics, development, and global finance. While their motives, intentions, and operations may vary, wealthy economies in East Asia and the Gulf states using SWFs share common goals of managing rates of return on accumulated financial assets through investment in a broad range of asset classes and regions around the world. Research explores how these SWFs act as strategic instruments of states in maintaining legitimacy and furthering geopolitical interests, which combine both domestic concerns and international relations. As a result of these global investment footprints, even small countries with large and globally oriented SWFs can wield significant influence in the global political economy.

Analyses of the constructions and practices of financial logics based on Islamic religious principles and ethical guidelines have also been incorporated in order to better understand so-called mainstream finance. They reveal the use of different types of economic theories beyond the orthodox financial economics that dominate “Western” economic systems and a suite of different financial centers and networks previously unaccounted for. The growth of Islamic banking and finance in Muslim-dominated countries in the Middle East, Southeast Asia, and parts of Africa, as well as in non-Muslim countries (with London being an important global center for
Islamic financial products), also points to the entangled nature of “alternative” and “Western” financial systems. This is particularly significant as even conventional financial institutions are now looking to “alternative” financial products and markets in the incessant search for returns on capital, be they emerging markets, Islamic finance, or carbon trading markets. This underscores the importance of engaging in more intensive research on financial geographies beyond established financial centers, in ways that are sensitive to geographically variegated systems and practices, as well as on the ways in which different circuits of financial knowledge and practices are entangled and mutually constituted along uneven power geometries.

SEE ALSO: Corporate financialization; Corporate spatial organization and producer services; Economic geography; Finance and development; Global cities; Global production networks; Mainstream and shadow banking

References


Further reading


Geographers often need to make a systematic model that describes some phenomenon of interest. These models frequently involve numerical values, and give relationships between those values as a set of equations to solve. In many cases these solutions are not available in analytic form, and instead require a computational approximation to find an answer. Offered here is a brief overview of how such answers may be obtained; more detailed specialist texts are indicated in the section “Further reading.” After a short introduction, the basic ideas of using finite differences to find approximate solutions to differential equations are described. There follows an illustrative example of using finite elements, which can deal with more complex geometries and capture a range of geographic phenomena that can be represented as spatially continuous. A final section illustrates how granular systems (or those better approximated using discrete entities) can be approached using the discrete element method.

The spatial distributions or temporal evolution of some geographical systems often have some aspects that are, to a good approximation, continuous, in time or space, or both. Physical examples are fluid flows in the atmosphere or ocean, surface flows in lakes and rivers, coastal phenomena including tides, storm surges or tsunami, and subsurface water fluxes including both near surface saturated and unsaturated layers, or deeper groundwater flows. Volcanic plumes or lava flows, or in colder areas snow-avalanches and, at longer time-scales, flow of glaciers may also be included. Surface deformation such as soil creep or, over geological timescales, the movement and deformation of the crust of the Earth may also fall into this category. Although, ultimately, the flowing materials are atomic in nature, at scales of geographical interest the material may be treatable as essentially smooth and continuous. However, even in systems of a noticeably discrete nature, if numbers are sufficiently large and space scales sufficiently great, it may be appropriate to treat ecological and human systems as having some nearly continuous properties. Examples include population growth, predator–prey interactions (Gurney and Nisbet 1998), spread of disease (Keeling and Rohani 2008), physiology of trees and animals, demographic and economic “Limits to Growth” (Meadows, Randers, and Meadows 2005), or diffusion processes in cities, such as suburbanization or intraurban migration (Batty 2007). Conversely, in granular systems comprising many discrete individual particles, interactions may be treated as continuous, even when contacts between particles are discontinuous both temporally and spatially (Munjiza 2004). In all these cases the most natural way to formulate a dynamic mathematical model of the continuous system is to use differential equations. Unfortunately, while it might be possible to write these equations down, they are often sufficiently difficult that they cannot be solved analytically (i.e., using exact mathematical techniques) but instead must be solved by other means, including numerical and computational approximations. In Box 1 a few
Finite-element and finite-difference modeling

Box 1 Some simple illustrative equations

**Exponential growth/decay**

Population $N$ with birth rate $b$ and death rate $d$:

$$\frac{dN}{dt} = (b - d)N = \lambda N$$

**The logistic equation – bounded growth**

Population $N$ with birth rate $r$ and “carrying capacity” $K$:

$$\frac{dN}{dt} = rN \left(1 - \frac{N}{K}\right)$$

**Nonlinear oscillations – the Lotka–Volterra predator–prey equations**

Prey species $p$ born at rate $\alpha$, is eaten by predators, $e$, at rate per capita $\beta$, leading to predator births at rate per capita $\gamma$. Otherwise predators die at rate $\delta$:

$$\frac{\partial p}{\partial t} = \alpha p - \beta pe$$

$$\frac{\partial e}{\partial t} = \gamma pe - \delta e$$

**Compartmental model of disease spread – the “SIR” equations**

Susceptible population fraction $S$, Infected $I$, Recovered (or dead) $R$, with infection rate $\beta$ and recovery rate $\gamma$:

$$\frac{\partial S}{\partial t} = -\beta SI$$

$$\frac{\partial I}{\partial t} = \beta SI - \gamma I$$

$$\frac{\partial R}{\partial t} = \gamma I$$

**Diffusion in one space dimension**

Quantity $S$ diffuses over space $x$ with time $t$ with diffusion constant $k$:

$$\frac{\partial S}{\partial t} = k \frac{\partial^2 S}{\partial x^2}$$

**Shallow water equations in one space dimension – conservative form, no friction or viscosity**

$H$ is the water depth and $h$ the height of the water surface level (above the horizontal), velocity $u$, time $t$, space $x$, acceleration due to gravity $g$:

$$\frac{\partial h}{\partial t} + \frac{\partial (Hu)}{\partial x} = 0$$

$$\frac{\partial (Hu)}{\partial t} + \frac{\partial (Hu^2)}{\partial x} = -gH \frac{\partial h}{\partial x}$$

Illustrative examples are shown, from the simplest one-dimensional forms where we have access to the closed-form solution, to those cases where a small degree of nonlinearity leads to complex dynamics that must instead be approximated. These examples will prove useful in illustrating computational ways forward.

Finite differences

Solving sets of equations has a long history. The volume “Institutionum Calculi Integralis” (Foundations of Integral Calculus), published in 1768 by Leonhard Euler, offers methods for obtaining approximate integrals of differential
equations. Fortunately these techniques have been elaborated sufficiently often since then that we do not have to resort to the primary text. The basic idea is that starting from a known point, and assuming that rates of change are constant over a small interval, neighboring values of variables can be approximated by moving forward a finite step. In doing so it is hoped that an accurate estimate is made by choosing a step that is sufficiently small. Essentially this reverts to the original way the derivative is defined using a straight line to approach a tangent to a curve (Figure 1).

Given some function $S$ for which we know the rate of change, $F$, as a function of $S$ itself and a coordinate $q$, and a starting value $S_0$ at $q_0$, that is:

$$\frac{dS}{dq} = F(S, q), \ S = S_0 \text{ at } q = q_0 \quad (1)$$

a new value of $S$ at $(q_0 + dq)$ can be estimated by using the rate of change $F(S_0, q_0)$ thus:

$$S(q_0 + dq) \approx S(q_0) + F(S_0, q_0) dq \quad (2)$$

which is generally known as the forward Euler method. So our differential equation is replaced with an approximate algebraic expression that is easier to evaluate. Subsequent values of $S$ at \{q_0 + dq, q_0 + 2dq, q_0 + 3dq \ldots \} can be found sequentially, stepping forward to find the new values of $S(q)$ and thus $F(S,q)$, so that we can find a complete estimate of the behavior of $S$ over some range (Figure 2).

Even in a simple case this procedure is tedious and error prone to carry out by hand, although as pointed out by Richardson (quoted in Hunt 1998) such methods are often simpler, more systematic and easier to check than many analytic methods. Richardson’s own pioneering attempt at a numerical weather forecast in the early twentieth century (Richardson 1922), based on finite differences, laid the groundwork for later development of weather forecast models in the 1940s and 1950s. His own first attempt lacked accuracy because the initial conditions for the forecast needed to be filtered to match the timescales present in the numerical scheme – otherwise errors in the initial data can trigger large-amplitude high-frequency gravity waves that are usually suppressed in the real atmosphere (Lynch 2006). However, the practicality of this approach lay in doubt because of the effort required – Richardson’s own vision of a room filled with people each tasked with the calculation of part of the forecast needed 64,000 people to keep pace with the global weather (presaging the modern method of using many parallel computational cores). An automatic method is needed to make finite-difference calculations realistically possible.

As early as the 1800s the need for mechanization to improve reliability of finite-difference methods was recognized, leading to Babbage’s design for the difference engine, essentially a mechanical computer. The target was, in that

Figure 1 The rate of change of $S$ at $q_0$, $F(S_0, q_0)$, is given by the red tangent line. For a finite change $dq$, the actual change in $S$ is shown by the green dotted line. Extrapolating $S(q_0)$ to an estimate of $S(q_0 + dq)$ using the tangent at $q_0$ we make the error relative to the exact value shown by the blue dashed line, which, for a sufficiently smooth function, we can make as small as we please by reducing the size of $dq$. 


Figure 2  Successive approximations to $S$ starting from $S(q_0)$ at each of the points $\{q_0+dq, q_0+2dq, q_0+3dq \ldots \}$ using the forward Euler method. Notice that the error in each step accumulates to give a global error after many steps that is larger than each individual error, although in the case illustrated the error per step is decreasing as $q$ increases.

case, the calculation of series approximations, dependent on the idea (due to Newton) that successive differences of a polynomial sequence become constant once the number of differenc- ings matches the polynomial order. Although such automation was not realized until the advent of the digital computer in the mid-twentieth century, the mechanical nature of the difference engine makes it clear that the calculation has finite precision. Digital computers, being by their very nature finite, lend themselves to finite-difference approaches almost by default. While it is possible, with modern machinery, to evaluate arithmetic expressions to arbitrary precision, the time required increases with the number of significant digits, and at some point a truncation in the representation of numbers is needed. So every numerical quantity in a computation usually has a fixed fractional precision: a typical value would be about 15 places of decimals. Multiplication of two such numbers cannot have better precision, and extra digits that might be present in an exact calculation get discarded. This rounding error is independent of the above errors made using a finite-difference scheme, but may be important, as these small imperfections can be amplified by unstable numerical methods.

Desirable properties for numerical solution schemes

The three main desiderata of finite-difference methods are consistency, stability, and convergence. For a method to be consistent, the error in making a single step from any starting point should reduce to zero as the step size reduces. The rate at which this local error reduces, expressed as a power of the step size, determines the order of the method – the forward Euler method is linear (i.e., first order) and will give exact answers only for linear functions. For numerical efficiency it is advantageous to have a faster reduction of local error with step size, either quadratic (second order) or better, although the great effort needed to work with order higher than five usually means that very high order methods are not often used. However, this by itself is not sufficient to ensure that the global error be controlled – for this the method has to also be stable. In general, if the method is both consistent and stable, then it will also be convergent, in the sense that the global error will reduce to zero as the step size is reduced, and the numerical solution converges to the solution of the original differential equation.

To look at this more closely, consider some of the example equations from Box 1. The conditions of consistency and stability demand that the global error decrease with step size $dq$, so that in the limit the exact answer is recovered.
For exponential decay this can be shown to be the case by solving the finite-difference equation from the forward Euler method directly. The original equation is shown in equation (A):

\[
\frac{dN}{dt} = (b - d)N = \lambda N \Rightarrow N = N(0) \exp(\lambda t)
\]  

(3)

where \(N\) is a population number, \(b\) the birth rate, \(d\) the death rate, and \(\lambda = b - d\), so that for deaths exceeding births we get exponential decay. Using the forward Euler method with timestep \(h\) gives equation (4):

\[
\frac{N(t + h) - N(t)}{h} \approx \lambda N(t)
\]

\[
\Rightarrow N(t + h) = N(t)(1 + \lambda h)
\]

\[
\Rightarrow N(nh) = (1 + \lambda h)^n N(0) = \left(1 + \frac{\lambda t}{n}\right)^n N(0)
\]

(4)

where \(t = nh\) denotes the total time represented by \(n\) steps of length \(h\). If \(t\) is kept constant by letting \(h \to 0\) and \(n \to \infty\), the final discrete expression of equation (4) reduces to the exponential function. Hence, the solution converges to the solution of the original equation in the limit of small \(h\). However, something different happens when \(h\) becomes large. Whereas for \(-1 < \lambda h < 0\) a monotonic decreasing behavior is obtained, in \(-2 < \lambda h < -1\) the solution still decreases, but oscillates, and once \(\lambda h < -2\) the oscillations grow geometrically and there is no possibility of a solution. This kind of rapid unwanted growth is typical of unstable numerical methods – usually there is a threshold step value, beyond which instability sets in and the algorithm “blows up,” often crashing the program that tries to implement it. Some methods, however, can be unstable for any timestep value. For example, we can try to increase the order of the method by using the so called “leap-frog” scheme. Suppose a Taylor expansion of \(N(t + h)\) and \(N(t - h)\) is written (equation (5)):

\[
N(t + h) = N(t) + \frac{dN}{dt} \bigg|_t \bigg(h + \frac{1}{2} \frac{d^2 N}{dt^2} \bigg) h^2 + O(h^3) \ldots
\]

\[
N(t - h) = N(t) - \frac{dN}{dt} \bigg|_t \bigg(h + \frac{1}{2} \frac{d^2 N}{dt^2} \bigg) h^2 - O(h^3) \ldots
\]

(5)

Subtracting the two expressions in equation (5), the terms in \(N(t)\) and those proportional to \(h^2\) cancel to give:

\[
\frac{dN}{dt} \bigg|_t = \frac{N(t + h) - N(t - h)}{2h} + O(h^2) \ldots
\]

(6)

This result is exact at point \(t\) for both linear and quadratic curves, and, therefore, the single-step error should reduce more quickly as \(h\) is reduced. Using this approximation in the exponential equation gives:

\[
N(t + h) = N(t - h) + 2\lambda h N(t)
\]

(7)

Note that two starting values are now needed: if \(N\) is known at \(t = 0\) and \(t = h\), then \(N\) can be found at \(t = 2h\). Usually a first-order step (e.g., forward Euler) is used to acquire \(N(t + h)\) as part of the process of solution. Initially the result is more accurate than the Euler method (Figure 3), but after a while initial small oscillations (kicked off by the first step!) begin to grow and shortly are well out of control. Reducing the timestep just postpones the inevitable failure to find a solution.

For this reason, many alternative schemes have been developed for higher order accuracy. Runge-Kutta schemes are often employed for their relative simplicity and robustness, although multistep (or predictor-corrector) schemes may give high order with good stability (Press et al. 2007). Often, though, in more complex
Figure 3  Trying to solve the exponential decay equation using a leap-frog scheme. The exact solution is the black curve, and the numerical scheme is in red. After a while exponentially growing oscillations set in: the solution is unconditionally unstable.

cases, finding the right timestep can require some experimentation. Careful analysis of the expected errors and their rate of growth may give a condition on the timestep controlling the solution. Using a Taylor series expansion to analyze growth of errors can help, for example, although a Fourier analysis is often more informative (Morton and Mayers 2005). In nonlinear cases this typically means first making a linear approximation to the equations to make the analysis tractable – even so, for nonlinear problems it can be necessary to introduce some artificial damping terms that control the error growth even though they are not strictly implied by the original equations.

As noted above, the attempted solution with the leap-frog method is also oscillatory, a behavior that is not evident at all in the original solution – such spurious “numerical modes” can be present even when the method is otherwise stable, and can be quite difficult to suppress. For example, the SIR equations (per Box 1, a model of disease spread) can have such oscillatory errors for a range of different numerical schemes (Keeling and Rohani 2008). Here the logistic equation is looked at as a slightly different example. This equation has often been used to describe a population initially growing exponentially, but limited as to resource so that it has maximum carrying capacity $K$ – the solution can be found analytically with a little algebra (Gurney and Nisbet 1998), as shown in equation (8):

$$\frac{dN}{dt} = rN \left(1 - \frac{N}{K}\right) \Rightarrow N = \frac{KN(0)}{N(0) + (K - N(0)) \exp(-rt)}$$

(8)

which has the rather smooth behavior shown in Figure 4. Note that for an ordinary differential equation of this kind it can be shown that the solution is unique for given $N(0)$, $r$, and $K$.

In finite-difference form, using the forward Euler method gives:

$$N(t + h) = N(t)(1 + rh) - N(t)^2 rh/K$$

(9)

For small timesteps this reproduces the analytic form reasonably well. However, as the timestep increases, bounded oscillations appear: the solution alternates about the final analytic value without growing. Increase the timestep still more and the oscillation begins to vary in size and become unpredictable, in the sense that a small change in the initial condition leads to a large difference in the solution. The difference equation shows deterministic chaos (Ott 2002). Here the point is to illustrate that the differential equation and the finite-difference version can have completely different behavior, and for nonlinear cases it may be difficult to anticipate the parameter ranges over which such phenomena might appear. The issue of chaotic nonlinear behavior, which renders significant forward prediction in time impossible, is a frequent problem.
Figure 4 Solution of the logistic equation with $N(0) = 1$, $K = 100$, $r = 3$ (black) and the discrete version from equation (9) with $h = 0.975$. For $h > 1$ the solution is unstable and plunges off toward $-\infty$.

Typically such a form leads to more stable expressions for the difference equation, but now it is necessary to solve for $S(q)$ at each timestep, since on the right-hand side there are terms that are evaluated at the forward point $q + dq$. In cases where $F$ is nonlinear this may be nontrivial: typically it at least requires solution of a set of linear equations to find the value of $S(q + dq)$. Hence, these methods are usually called implicit schemes, as opposed to the more straightforward explicit schemes that have been looked at above. For these reasons it may not be worth the expense of using an implicit method in order to gain stability – a shorter timestep may be easier and may possibly suffice. However, for certain types of equations, particularly those that are known as stiff, where the existence of a wide range of timescales in the solution makes a stable scheme difficult to achieve (Acton 1970), implicit schemes may be a necessity. In simple cases, it may be possible to find a solution directly – for exponential growth the solution is shown in equation (11):

$$\frac{N(t + h) - N(t)}{h} = \lambda N(t + h)$$

$$\Rightarrow N(t + h) = N(t) \frac{1}{1 - \lambda h}$$

yielding the backward Euler method, which is stable for any step size $h$ when $\lambda < 0$, although large values of $h$ will give large errors. Further elaborations are possible using semi-implicit schemes that use a linear combination of forward and backward differences, as shown in equation (12):

$$\frac{S(q + dq) - S(q)}{dq} = \theta F(S, q + dq) + (1 - \theta) F(S, q)$$

where $0 < \theta < 1$. These do not imply less work than a strictly implicit scheme, but it may be possible to obtain a higher order method than either a strictly explicit or implicit method alone, whilst also giving improved stability.

Including space

In order to solve initial value problems that involve both time and space it is necessary to difference the spatial part of the system. As an example consider the one-dimensional diffusion equation:

$$\frac{\partial S}{\partial t} = k \frac{\partial^2 S}{\partial x^2}$$

Notice that an approximation is now needed for the second derivative, which can be obtained from the Taylor expansions of equation (5),
written for space coordinate $x$ rather than time $t$. Adding rather than differencing the two, the first order derivatives cancel to give:

$$\frac{\partial^2 S}{\partial x^2} \approx \frac{S(t, x + \delta x) - 2S(t, x) + S(t, x - \delta x)}{\delta x^2}$$  
(14)

for small spatial difference $\delta x$. For time differencing the Euler method can be used (the leap-frog method would, in this case, be unstable for all $\delta t$ again), as shown in equation (15):

$$S(t + \delta t, x) = S(t, x) + \frac{2k\delta t}{\delta x^2}(0.5(S(t, x + \delta x) + S(t, x - \delta x)) - S(t, x))$$  
(15)

On the right-hand side of equation (15), the second term shows that some fraction of the value at point $x$ is replaced with the average of the values to either side: a smoothing of the spatial values is obtained as time advances, as expected from a diffusive process. Note that this term is now weighted by a ratio of the timestep to the square of the spatial step – the stability of the scheme depends on this ratio being sufficiently small: a tiny difference in $\delta t$ can again lead from a stable solution to exponentially growing oscillations. Implicit or semi-implicit schemes are typically therefore preferred in this case. Notice also that information is needed across the entire boundary in $x$ in order to start the solution off. This includes the edge points where the above approximation for the second derivative cannot be used, and the behavior must be prescribed at all times. The dependence of the solution on the entire initial boundary is characteristic of this type of parabolic equation.

By contrast hyperbolic equations, such as the shallow water equations, used to model fluid systems where the horizontal scale is much larger than the vertical, have associated with them a finite propagation speed. This means that the domain of dependence for any spatial point at a given time includes only those other spatial points that are close enough for a signal to have traversed the intervening distance since $t = 0$. Here, for stability, the ratio $\delta t/\delta x$ needs to be small enough that the numerical domain of dependence includes that of the original equation – this corresponds roughly to the notion that no propagating disturbance should be able to travel more than $\delta x$ in under a single timestep.

To extend the method to two space dimensions, a regular rectangular grid of cells extending in the $x$ and $y$ directions can be used. Since the axes are orthogonal we can introduce independent differences in each space dimension. For example, for the diffusion operator in 2-D taking differences separately in $x$ and $y$ is shown in equation (16):

$$\nabla^2 S = \frac{\partial^2 S}{\partial x^2} + \frac{\partial^2 S}{\partial y^2} \approx \frac{S(t, x + \delta x, y) - 2S(t, x, y) + S(t, x - \delta x, y)}{\delta x^2} + \frac{S(t, x, y + \delta y) - 2S(t, x, y) + S(t, x, y - \delta y)}{\delta y^2}$$  
(16)

Numerical stability now depends on which of the two spatial dimensions has the more restrictive condition on the timestep. Spatial considerations can lead to new kinds of instability, however – for example, centered spatial schemes (similar to leap-frog, but now for spatial differences) can lead to a decoupling of solutions between alternating spatial cells and lead to a characteristic “checkerboard” pattern that may, again, grow exponentially with time.

These spatial finite-difference approximations are closely related to cellular automata (CA), the main difference being that here there is the specific goal of approximating a continuous equation. In the abstract, cellular automata do
not normally need to impose stability or convergence criteria on changes between steps, unless the intent is to model a real physical process such as fluid flow. On the other hand, the set of allowable states in a cellular model is typically smaller, but the range of phenomena covered draws on a broader class than that which is usually thought to be covered by differential equations. However, it is useful to bear the connection in mind, as the large body of literature that has grown up around finite-difference methods may be informative for CA models.

**Finite volumes and finite elements**

One difficulty with the use of orthogonal grids for spatial differences is that the grid may not fit complex geometries. In such cases it may be appropriate to allow for other cell shapes. While it is common to write cellular automata models on a hexagonal rather than a rectangular grid, solution of differential equations typically needs more flexibility. Geographers are familiar with the idea that triangulated networks are useful for tiling domains with irregularly spaced points, and this can be exploited in creating a mesh that can efficiently cover complex boundaries. The problem is to find schemes that can deal with arbitrary geometry in a well-controlled way. A whole area of research has developed around generation of meshes that have sufficient smoothness and regularity to be stable for a given solution scheme.

For flow problems, where it is possible to write the governing equations in conservative form, that is, where time derivatives depend on spatial fluxes (e.g., the shallow water equations, Box 1), finite volume methods can be used. The discretized equations are written in terms of cell-average values, with fluxes between cells specified on cell boundaries. This has the advantage that the cells can have any shape, and that the conservation properties of the system are good, since fluxes subtracted from one cell are guaranteed to be added to neighboring cells and vice versa (subject, of course, to some rounding error). This method is frequently used for shock-capturing schemes in fluids, since the conservative flux-form of the equations gives better accuracy in following moving discontinuities.

*Finite-element* techniques are an alternative set of methods arising originally from the need to solve boundary value problems in engineering: in these cases, rather than the initial value problems that have been dealt with above, the values of a function are specified along an entire boundary, and the solution has to be found everywhere in the interior of the bounded domain simultaneously. Equations of *elliptic* type fall into this category, an example being the steady-state version of the diffusion equation, with zero time derivative, leading to the Laplace equation, \( \nabla^2 S = 0 \). However, finite-element schemes can also cope with time dependency. To deal with irregular cell shapes, the fact that any function can be expanded in a series of linearly independent basis functions is used. By choosing these basis functions to be nonzero only outside some small spatial range around each cell, an integral form of the differential equation can be converted to a set of algebraic equations (Figure 5). For simple cases this often reduces to an equivalent finite-difference method, but the ability to handle arbitrary cell shapes, and to use relatively high order polynomials as basis functions, makes the extra effort required to set up the method worthwhile when boundaries have a lot of structure that needs to be captured. For example, where the boundary has sharp edges or rapidly changing properties, mesh refinement can be used, covering areas that have more rapidly changing properties with a larger number of smaller cells.
Finite-element and finite-difference modeling

Figure 5 Linear finite elements \(e_1, e_2\) and \(e_3\) defined on range \(x_0\) to \(x_4\), \(e_2\), shown in red, has the form:

\[
\begin{align*}
  e_1(x) &= \frac{x-x_1}{x_2-x_1} \quad x_1 < x < x_2 \\
  e_2(x) &= \frac{x-x_3}{x_2-x_3} \quad x_2 < x < x_3 \\
  e_3(x) &= 0 \quad \text{otherwise}
\end{align*}
\]

And similarly for \(e_1\) and \(e_3\), so products of \(e_1\) and \(e_2\) will only be nonzero when \(x_1 < x < x_2\).

An example of one version of the finite-element method is shown in Box 2. This simple case can be generalized to more space dimensions, and adjusted to accommodate more complex equations and more general boundary conditions, while maintaining the sparse matrix form, so that solutions are computationally efficient.

Discrete elements

The methods described so far are helpful when the behavior of a system is well approximated by aggregate or averaged values as continuously varying functions over space and time. For some cases this may be an inadequate description, or it may not be clear how to construct such aggregate equations in a way that is known to be consistent with the underlying dynamics.

As an example, rock fall or granular flow, such as sand grains moving against each other, can be considered. Here there are a large number of discrete particles distributed over a range of sizes, and possibly made from different materials, that may only interact fleetingly. Although it may be possible to write linear equations to describe the elastic interactions between the particles when they are in contact, the nonlinear nature of contact–changes over time can give rise to complex emergent behavior for the whole system that might not be anticipated from the individual interactions. One approach to this problem is to try to model every individual particle, allowing those not in contact with each other to move freely (possibly accelerating under body forces such as gravity), but solving the elastic equations when particles are in contact. Thus, a set of equations is obtained as shown in equation (17):

\[
\begin{align*}
  m_i \frac{d^2 r_i}{dt^2} &= \sum_j F_{ij} + B_i \\
  I_i \frac{d\omega_i}{dt} &= \sum_j \Gamma_{ij} + \tau_i
\end{align*}
\]  

(17)

where \(m_i\) is the mass of particle \(i\), \(r_i\) is a three dimensional vector giving the position of the particle center of mass, \(F_{ij}\) are the forces acting between particle \(i\) and \(j\), and \(B_i\) are the other forces acting on the particle, including gravity. \(I_i\) is the particle moment of inertia, \(\omega_i\) the angular velocity, \(\Gamma_{ij}\) the torque between particle \(i\) and \(j\), and \(\tau_i\) any torque acting other than those due to particle interactions, and this equation is written rotated into the principle axes of body \(i\). The difficulties now lie in efficiently tracking particle contacts, and in finding and specifying the force laws that govern the particle interactions. The force equations can then be solved at each contact point using finite differences. To find contacts the simplest method is again to use a regular grid.
Box 2 Finite-element example

Assume we wish to find $S$ as a sum over a set of functions $e$ so that:

$$S = \sum \alpha_j e_j$$

where the values $\alpha_j$ are values to be determined. The $e_j$ need to form a complete set over our domain. Usually it is sufficient to use polynomials in the coordinates for this purpose, but we have to truncate the set in order to make a usable approximation – so, for example, linear, quadratic or cubic polynomials might be used, depending on the degree of accuracy required.

Suppose we want to solve:

$$\frac{\partial^2 S}{\partial x^2} = F(x)$$

with some boundary conditions, say $S = 0$ over the whole of the boundary of the domain.

First we write our equation in the weak form, using some set of functions $\nu(x)$:

$$\int \frac{\partial^2 S}{\partial x^2} \nu(x) dx = \int F\nu(x) dx$$

which will generally be satisfied, given the original equation.

Assuming that we can choose $e_j$ to satisfy the same boundary conditions as $S$, and that we can also use the $e_j$ as our functions $\nu(x)$ (this sometimes being known as the Galerkin method):

$$\int \frac{\partial^2 \sum \alpha_j e_j}{\partial x^2} \nu_i dx = \int \frac{\partial \sum \alpha_j e_j}{\partial x} \frac{\partial \nu_i}{\partial x} dx = \int F\nu_i dx$$

where we have integrated by parts and used the fact that the $e_j$ are zero at the boundary. Interchanging the sum and integral we can now write a set of linear equations for $\alpha_j$:

$$\sum_j K_{ij} \alpha_j = f_i$$

where the stiffness matrix

$$K_{ij} = \int \frac{\partial e_j}{\partial x} \frac{\partial e_i}{\partial x} dx$$

and

$$f_i = \int F e_i dx$$

By choosing $e_j$ carefully so that they are only nonzero in a small range near to a given grid-point $j$, the matrix $K_{ij}$ can be made sparse, which is to say only a few of the elements are nonzero, typically mostly those close to the leading diagonal.

Particles can then be searched for contacts with those that lie in neighboring cells, rather than having to search the entire domain. The simplest model represents all particles as rigid spheres, but approximates contact forces using linear springs, where the particles are allowed to overlap by a small amount to represent spring compression at the surface. Forces producing rotational torques reduce to tangential friction in this case, and can be represented by a spring that stretches over time up to some limiting length beyond which slip occurs. Thus:
with normal force $F_{i,j}$, where $k_{i,j}$ is the stiffness of the spring acting to repel the two particles, and $\Delta r_{i,j}$ the particle overlap. $u_{i,j}$ is the normal component of the relative velocity, leading to a damping term $C_u_{i,j}$. Similarly, $k^T_{i,j}$ is the tangential spring stiffness and $Dv_{i,j}$ damps the tangential motion given tangential relative velocity $v_{i,j}$. $F^T_{i,j}$ is the tangential force, $\mu$ is a friction coefficient and $\hat{t}$ is a unit vector along the direction of the tangential spring (Cleary and Prakash 2004). The timestep for stability now depends on the factor

$$\sqrt{\frac{k}{m}}$$

as the spring constant for real rocks is extremely large (and as we would also like to limit the particle overlap to a small fraction of the radius) this necessitates a very small timestep – typically microseconds. Because any real system comprises many billions of particles, the method is very computationally expensive: usually only a small fraction of the material can be dealt with, either by using a very small domain, or neglecting most of the smaller particles. However, spheres typically flow more easily than irregular shapes, and using nonround particles may be required. This can be done by bonding together groups of spheres and treating these groups as a single rigid body, but this may leave artificially lumpy particle boundaries. At the other extreme, each particle can be modeled as an elastic solid, using finite elements, and the internal propagation of elastic waves and shear deformation can then be added to the solution. This combined finite-discrete element should then give a more accurate representation of the bulk properties that can emerge from individual particle-scale interactions (Munjiza 2004). Either way, the solution of the contact problem for nonspherical particles is even more expensive, especially if particles are nonconvex and can have multiple contact points with each other.

Challenges

This brief examination of discrete methods has touched upon a range of problem domains, each with their own challenges for future development. Even where a good solution method is known for a well-established set of equations, we often have poor knowledge both of the appropriate parameters and of the relevant boundary conditions for an unambiguous solution. In hydraulic models, for example, the subsurface conditions are typically not known, and models will only give reasonable results once calibrated to flow data – this can mean several different set of parameters might fit the surface flow data equally well (the “equifinality problem”) (Beven 2006). In the atmosphere nonlinear processes that occur at microscopic scales cannot be resolved by a grid that can cover the global domain, yet can have a significant effect on the solution (cloud processes and surface hydrology being examples). These subgrid processes have to be modeled independently (usually referred to as a “parameterization”) to try to approximate effects of the unresolved dynamics. Even so, fluid advection tends to produce narrow structures that move downscale and cross the boundaries of fixed grids, leading to their being artificially diffused away by the numerical scheme. This is particularly a problem for traces gases or pollutants that are being carried along by the fluid and can add to the difficulties of simulating atmospheric chemistry (which often has “stiff” numerical
behavior, in that solutions require implicit methods). One possible way of dealing with these issues is using adaptive mesh refinement, adding extra resolution dynamically to the model where needed. As a further difficulty, the initial conditions are often poorly known, and the equations chaotic, so that current practice is to produce not just one model output from given initial data, but an ensemble that covers plausible variations in the initial state that are consistent with the discrete model setup, and therefore produces a range of possible future atmospheric conditions.

At smaller scales, free surfaces in complex domains with moving boundaries are a challenge for fixed-grid methods. As of the time of writing, for example, in flood models, stage-dependent flow of water past arbitrarily shaped flexible vegetation is just beginning to be addressed. The case where the fluid is filled with floating debris, as is typical for any real flood, has yet to be approached effectively. When suspended and moving sediments are included then modeling becomes still more difficult, especially where a discrete element model of sediment is to be combined with the fluid flow. The large size range that typically occurs in such situations, from fine sand and silt up to cobbles and boulders, represents a further challenge, although recent advances in the use of methods based on parallel graphics-processing-units (GPU) make the treatment of large numbers of particles and the solution of the contact problem over large size ranges more accessible.

Moving to the ecological and human realm, the use of aggregate population–based equations has its own problems. Unrealistic results can come from the assumption that population densities are continuous down to the accuracy of the computational representation – for example, a continuous model of the spread of rabies in foxes allowed for the resurgence of the disease in areas where the population density suggested only $10^{-18}$ of a fox could be present (hence referred to as the atto-fox problem) (Mollison 1991). It might be questioned whether the differential equation approach is really appropriate for this kind of case. As the logistic equation example shows, discrete and continuous approximations can have quite distinct dynamical results – it may be that in these kinds of cases an explicitly discrete approach that takes into account the “graininess” of the real system is more appropriate and realistic. For granular flows the discrete element method is one step in this direction, and the equivalent for ecosystems (individual-based models) and social systems (agent-based models) may be better suited to addressing the full complexity that these systems can display.

SEE ALSO: Agent-based modeling; Cellular automata; Geocomputation; Glacier mass balance; Global climate models; GPU computing; Hydrologic flow models; Models in geomorphology; Models and simulation in biogeography; System dynamics

References


Euler, Leonhard. 1768. *Institutionum Calculi Integralis* [Foundations of Integral Calculus]. Petropoli Impensis Academiae Imperialis Scientiarum.


Description and purpose

Suomen Maantieteellinen Seura is a group of geographers and those interested in geography. The official languages are Finnish and Swedish. The purpose of the society is to promote research in geography and related sciences at home and abroad. It also promotes the use and visibility of geographical knowledge and understanding in tackling societal and environmental challenges. The society also makes statements on topical issues and processes related on, for example, research and education politics and regional planning. Through its activities the society promotes interaction between geographers nationally and internationally.

Current activities or projects

Geography Days are twice-yearly meetings of geographers and geography enthusiasts organized in collaboration with university geography and regional studies units. They are open to all society members; five to six board meetings are held in a year.

The society organizes international collaboration and meetings with other geographical societies especially from neighboring countries (for example, the biannual Nordic Geographers Meeting (NGM)). The society prepares statements on topical processes in the society (e.g., related to science politics and geography education in national school curriculums).

Brief history

The Geographical Society of Finland (founded 1888) and Finnish Geographical Association (founded 1887) were united in 1921 into a single organization, the Geographical Society of Finland. One of the cornerstones of the society’s publishing has been the National Atlas of Finland. The first edition of the Atlas was the first of its kind in the world and raised international interest (e.g., achieved a gold medal in the Exposition Universelle of 1900 held in Paris, France). Since then, cartographic communication and map-making methods have changed drastically. The maps of the previous editions are currently available in the geography.fi web pages for everyone to use and enjoy.

Submitted by Sanna Mäki
Firm foundation and growth

Godfrey Yeung
National University of Singapore

The firm acts as an intermediate agent to reduce the transaction costs incurred in dealing with individual economic actors. It also acts as a nexus of contracts to allow transactions to be internalized and coordinated efficiently in the circular flow of production and consumption, where trust between economic agents could become institutionalized within the economic system.

Firm foundation is about how to govern the operation of a firm, ranging from official rules and regulations to unofficial norms, customs, and practices. Firm growth is about the development of a firm, including the expansion of its spatial reach through mergers and acquisitions and/or greenfield investment in local and overseas markets. An understanding of the firm, its foundation, and its growth is crucial to comprehend how economic agents and actors are interrelated and contribute to economic development across space.

Economic geographers highlight the interlinked relationships between institutions (the formal rules and regulations and informal norms) and the best practices of industrial sectors and how these shape firm behavior. Institutions not only regulate industrial best practices and conventions regarding the organization of work and labor-management relations in the workplace, but also shape the values, expectations, and even the decisions of economic actors through the functioning of labor markets, regimes of industrial relations, and corporate governance. Economic geographers therefore contribute to the debate about globalization by providing solid empirical evidence that it is generating heterogeneous rather than homogeneous processes (or outcomes) of firm foundation and growth. As clearly illustrated by the two case examples of law firms and staffing agencies in this entry, the world is not “flat” and there is no convergence of best practices across space as the transnational corporations (TNCs) cannot just adopt their operational templates from home countries in their subsidiaries without adjusting their corporate cultures to the national laws and regulations and the specific local culture and practices in host countries. Geography does matter for producer service sectors, notably consultancy, accountancy, banking, and finance, as partly illustrated by the high-profile global advertising campaign slogan of the banking giant HSBC: “The world’s local bank.” In other words, diverse national institutions create heterogeneous industrial practices across space, and their reproduction and modification results in distinctive national/regional modes of governance and trajectories of growth in firms rather than their global convergence.

An overview of the theory of the firm proposed by Ronald H. Coase will be followed by a brief outline of various perspectives on firms, including the crucial contribution of Oliver E. Williamson on the importance of interfirm (market) and intrafirm (hierarchical) transactions in the functioning of the market economy. A detailed explanation of Meric Gertler’s concept of the formal rules and regulations and informal value and norms which contribute to the formu-
FIRM FOUNDATION AND GROWTH

Firm

Ronald H. Coase, a Nobel laureate in economics, initially developed the theory of the firm. In his classic article, Coase (1937) introduced the concept of transaction costs to explain the nature and limits of firms. He asked why, instead of economic actors trading with each other directly, economic transactions have to be channeled through firms in a presumably efficient market economy. The crux of the issue is that the functioning of the price mechanisms is not cost-free, as almost every transaction involves some information costs (for example, finding out the prices of raw materials to make a commodity) and coordination and agency costs, including the negotiation and conclusion of a separate contract for each exchange transaction, for example, from the purchasing of raw materials to measuring the separate contribution of individual workers in the production line, and so on. All the information and enforcement costs involved in the execution of economic transactions are called transaction costs.

Instead of dealing with every individual economic actor in each transaction and incurring the corresponding high transaction costs, Coase (1937) argued that the firm acts as an intermediate agent – between employer and employees on the production side, and wholesale/retail on the consumption side – to internalize the production of goods and services within a circular flow of production and consumption in an economy. The employment contract regulates the relationship between employers and employees: for a specified remuneration (which may be fixed in the case of monthly salary or vary according to certain predetermined parameters, such as piece rate wages), one agrees to obey the instructions of an entrepreneur within certain limits. In other words, firms exist to reduce the transaction costs of economic transactions in a market economy.

Various views of the firm

Coase’s pioneering work laid the foundations for our understanding of firms. A number of scholars have further elaborated the concept of the firm and the five most important interpretations are outlined below.

Resource-based: The firm is an administrative organization that aims to make profitable use of its resources. The differences in how resources and services are utilized explain the heterogeneity of firms with a similar level of endowments.

Knowledge-based: The firm is an efficient mechanism for the creation and transformation of knowledge into economically viable products and services. Temporary geographical proximity is vital for economic actors to exchange business information. Smaller firms in the same sector tend to agglomerate in industrial clusters, given their limited financial or human resources do not allow frequent commuting. Larger firms, in contrast, tend to locate themselves near subcontractors or research laboratories operating at different stages of the production processes. They are more “footloose” in their location and could find it easier to relocate in areas with better endowments and preferential policies offered by local governments.

Entrepreneurial: This approach focuses on the role of entrepreneurship where judgmental decisions concerning the mobilization of resources by opportunistic risk-takers are crucial for
the establishment and development of firms. Schumpeter argued that the “creative destructive” nature of entrepreneurship is illustrated through the introduction of (i) a new good or the new quality of a good, (ii) a new method of production, (iii) a new market, (iv) a new source of supply of raw materials or components, or (v) the new organization of any industry.

Evolutionary: Firms are seen as systems of distributed knowledge whereby established ones have developed a set of routine procedures and processes from their past successes. These routines could be modified or even replaced as a response to competition from new market opportunities and/or product solutions. Evolutionary economic geographers, notably Ron Boschma and his associates, have adopted such an analytical approach by emphasizing the importance of path dependency and technological lock-in to explain spatial variations in the formation and development of industrial clusters.

Governance-based: Coase’s pioneering work laid the foundation for the later important contribution of Oliver E. Williamson, another Nobel laureate in economics, on transaction costs and the firm. In his seminal work, Markets and Hierarchies (1975), Williamson further elaborated Coase’s ideas on transaction costs by showing how intrafirm (hierarchical) transactions could mitigate imperfections in the market economy. He argued that transaction costs accompany both interfirm (market) and intrafirm (hierarchical) transactions, and both markets and hierarchies are important coordinating mechanisms for economic transactions.

Due to the specific nature of human beings as economic agents, Williamson (1975) emphasized that in reality the market economy may not function effectively. He pointed out that economic actors would behave opportunistically and in their own self-interest (called opportunism, such as shirking their responsibilities and free-riding on the back of others’ work) or even by making irrational decisions due to lack of access to information (bounded rationality due to informational asymmetry). Instead of relying on market mechanisms to minimize the costs of interfirm transactions, a firm as a nexus of contracts can allow entrepreneurs to organize business transactions between employers and employees in a vertical hierarchical organizational structure to reduce transaction costs. Only by so doing can individual firms survive market competition from various other firms conducting their transactions through the market. Moreover, firms can allow transactions to be internalized and coordinated efficiently where trust between economic agents can become institutionalized within the economic system: that is, the economy can function efficiently when employment contracts are the norm and where both employers and employees know their obligations and rights. Therefore, the creation of hierarchical organizations and institutions is a rational response to the transaction costs problems caused partly by human nature (bounded rationality and opportunist) in a pure market economy.

The institutional turn in new economic geography

Two important perspectives in the development of new economic geography arose in the 1990s. In addition to the organizational dimension, economic geographers examined the impact of globalization on regional economic development, arguing that relational or organizational proximity is at least as important as geographical proximity. Other economic geographers
have focused on the supporting regional social and institutional contexts when examining the organization of production. One of the most notable contributors to the latter is an institutional approach as part of the new economic geography initiated by Meric S. Gertler.

Adding to the earlier work of Douglass C. North on new institutional economics, Gertler (2004) proposed a new conceptualization of industrial practices and the behavior of firms. He argued that the institution is the main driving force behind the organization of production through examining the influence of cultures on the practices of firms and the trajectories of regional and national economies.

Like North (1990, 3), who argued that institutions embody “the rules of the game in a society” and shape and constrain the behavior of individual economic agents, Gertler (2004) defined institutions as the formal rules and regulations and the informal norms that regulate the behavior and shape the values and expectations of individual economic actors (firms, managers, investors, and workers) through the labor and capital markets, corporate governance, and education and training systems, among others.

For the interlinked relationships between institutions and industrial practices, Gertler emphasizes the importance of maintaining a clear distinction between institutions at the societal level, attitudes and values at the individual level, and economic behavior as expressed in the industrial practices of firms (Figure 1).

Institutions have a twofold impact on industrial practices. On the one hand, they regulate industrial best practices and conventions regarding the organization of work and labor: that is, management relations in the workplace. On the other hand, institutions also shape the values, expectations, and even the decisions of individual economic actors (unconsciously) through the functioning of the labor market, industrial relations regimes, and corporate governance. These interactive relationships between institutions and industrial practices could have an impact on their own reproduction and modification over time. Thus, all firms are embedded within the matrix of their institutional environment, which incorporates formal rules and regulations, and informal customs and social norms in the context of a specific region or country.

In common with North (1990), Gertler (2004, 2010) also argues that market mechanisms play an important role in the continuation of industrial best practices and their further refinement. Economic actors in a competitive market could phase out inefficient industrial practices. Nonetheless, the scope and pace of any change in industrial practices is constrained by the existing institutional framework in the specific national context and the locally embedded formal and informal rules that have been reproduced over time. This is especially the case if changes in industrial practices have involved changes in the institutional environment that have had an impact on local, regional, or global politics. It is one of the reasons why institutional change is a path-dependent process and involves considerable transaction costs.

Using a comparison of the practices, norms, and expectations of German and North
American advanced machinery users, Gertler (2004) demonstrated that the industry-specific shared routines and industrial best practices – including the internal and interfirm organization of production, the adoption of technologies, and their affiliated knowledge and skills – in each region are strongly influenced by its corresponding institutional matrix (regional-specific social and industrial contexts). The subsequent development of industrial best practices and economic activities can, in turn, influence institutional development. Through these case studies, he demonstrated that it is essential to understand the way that place-specific institutions in the form of regulations and norms can define the corresponding values and attitudes. A detailed understanding of the relationship between local institutions and work cultures or industrial practices can demonstrate how workers should behave and how firms could be organized and operate. This in turn informs best practices for people in the workplace: that is, the industrial practices of workers in specific industries and in specific regions.

Instead of a convergence of best practices across space, as suggested by the leading proponents of globalization such as Francis Fukuyama and Kenichi Ohmae, Gertler (2010) and other economic geographers’ work has demonstrated that the distinct local institutional environment at various levels interacts with individual actors, organizations, and institutions, and regional institutions impact on the structure and operation of firms. This creates heterogeneous industrial practices across space. In other words, differences in local institutional settings affect industrial best practices, including differences in knowing how to work in specific professions in different regions. Due to the resistance to change from vested and dominant interests, and the existence of dominant social mores and traditions, the specific setting of each institutional environment in a region is a combination of elements that could both promote and hinder industrial and regional development. Consequently, the existence of institutional inertia (path dependency) supports the existence of distinctive national “models” rather than the global convergence of mode of governance and trajectory of growth in firms (as argued by the conventional thesis of globalization, in which geography no longer matters).

Two service sectors as illustrative examples

The implications for research into new economic geography are that the diversity of institutional regimes across regions accounts for significant variations in work practices. Instead of implementing the same business model and adopting an identical corporate culture all over the world, TNCs may therefore have to allow their local subsidiaries to adopt their own sets of best practices and corporate cultures in order to function effectively. In other words, TNCs are effectively not only dealing with differences in best practices, but also differences in corporate identity over space, as corporate culture is a set of social conventions, norms, and standards that underlie the social interactions within a firm.

Using a data reduction analysis (principal components analysis) of office networks, Taylor et al. (2013) documented the overlapping and permeable regional geographies of 175 leading advanced producer service firms in 138 cities in 2010. These major producer service firms extended their market reach via global cities in order to serve major clients in manufacturing sectors. The internationalization of law firms and staffing agencies are just two examples that are often used to illustrate how diverse national institutions can shape the development strategies of transnational service firms: that is, they show
FIRM FOUNDATION AND GROWTH

the importance of regional differences in the institutional environment and how transnational firms have to work with local management in overseas subsidiaries.

Law firms

Law firms provide interesting case studies of how regional institutions can have an impact on local work practices, partly because of the restrictive nature of a profession where professional associations play a key role in the establishment and enforcement of barriers to entry, and partly because of governance whereby legal firms are effectively co-owned and managed by (senior and managing) partners. Such de-equitization is an increasingly widespread trend in transnational law firms, especially after the bankruptcy of Dewey & LeBoeuf (New York) in 2012. Some partners in these law firms are called “service” partners, who are de facto employees as they are nonequity holders and have no clients they can claim as their own. It has been suggested that Americanization of legal services would be a feature of the globalizing era, including the spread of adversarial legalism. However, instead of the automatic transplantation of Anglo-American mega-law firms’ culture through American and British law firms, some studies show how their arrival in Germany, for example, in fact led to the development of German-type corporate mega-law firms which transformed the legal service sectors that had been dominated by small firms and individual lawyers.

As with other professions where educational institutions and professional associations have played a significant role in training and regulation, both the legal profession (lawyers and solicitors) and law firms are highly regulated at national and even regional levels. National (and regional) educational institutions, especially professional associations, thus have a profound effect on the establishment and development of specific cultures of professional practice through various entrance qualifications, examinations and licensing, in-house and/or often mandatory training (as interns or trainees) after graduation, and so on.

The independence of lawyers and the remuneration structures are two major areas of difference in the legal profession across space. The autonomy of lawyers is partly determined by local legal traditions: common law versus civil law. In the traditional Anglo-American common law jurisdictions, lawyers work in teams to deal with the complexities of regulation. American lawyers, however, have long been independent from the state, promoting their clients’ interests through the judicial process, and large corporate law firms were established in the early 1900s. British lawyers have been following this trend since the 1960s, and large law firms with more than 25 partners have been allowed since 1967. In continental European civil law jurisdictions, lawyers work as autonomous technical experts advising their clients of the relevant law and its specific requirements rather than using the law (and its potential loopholes) to defend or promote their client’s interests through the judicial process. In Germany, large corporate law firms (with more than 10 lawyers) have only been allowed since 1989. Law schools also instill specific local cultures in their graduates. For instance, individualism and competition are integral to American law schools, where their year group ranks graduates, while their British counterparts are not ranked. The Law Society in England stipulates that every law school graduate must complete a two-year training program with a law firm under the mentorship of a partner before she or he can be registered as a licensed lawyer, while their counterparts in the United States only have to pass the bar examination in their local state, and there is relatively little
interaction between partners and new recruits. In Germany, potential lawyers have to undergo a five-year degree program, pass the first state examination, undergo two years’ training in a law firm, and then pass the second state examination before finally being allowed to register as a lawyer (see Faulconbridge 2008).

In terms of the remuneration structure, it is largely about profit-sharing (or “lockstep”: all the partners share the firm’s revenue, with the specific percentage of their share determined by their seniority within the firm) versus the merit-based model (an individual lawyer’s income is based on the revenues she or he generates for the firm). In England and Australia, lawyers in the same firm normally work as a team to defend or advise a client, and the profit-sharing model of remuneration is adopted. Merit-based remuneration is normally used in the United States as an individual lawyer defends or promotes the client’s case to a judge through the judicial process. For instance, Clifford Chance and Linklaters, two of the largest British-based law firms, have adopted the profit-sharing model of remuneration, while American counterparts such as Skadden, Arps, Slate, Meagher & Flom or Jones Day use the merit-based model of remuneration (Faulconbridge 2008, 505). To maintain an integrated partnership across different offices, certain transnational law firms have adopted a hybrid model of remuneration that incorporates elements of both models, but uses the merit-based model for exceptional performance by individual lawyers. Latham & Watkins and White & Case, both US-based leading law firms, have adopted this hybrid mode of remuneration.

There can be diverse practices among the legal profession within the same country, partly to address the needs of their clients and partly as a reflection of their specific corporate cultures. In the interests of their brand and reputation, transnational law firms have an incentive to standardize their practices across space. Spatial heterogeneity in the culture of work between different regional offices could impact on the financial performance and even reputation of transnational law firms; their TNC clients expect the same standards as they enjoy in their home countries. This is especially the case for mergers and acquisitions deals across several jurisdictions, say from the United States or Europe to either Asia or Latin America. Senior partners at the headquarters normally develop and adopt certain “gold standards” in their internal training manuals so that new recruits (and junior partners) know what is expected as best practice. Moreover, transnational law firms will assemble a legal team across different jurisdictions to address the expectations of their clients. Practice groups, where all lawyers with the same legal specialization in different jurisdictions gather together, are the main strategy deployed by transnational law firms to address the need for teamwork and to form a common corporate culture. As a partial reflection of their legal work, each practice group within the same law firm will have its own subculture. For example, corporate lawyers tend to be involved in negotiation and debate, while litigators tend to have more ruthless work styles (and sometimes, attitudes) (see Beaverstock 2004).

Institutionalization of work cultures across different offices requires senior partners in global law firms to act as cultural entrepreneurs who impose their standards upon new recruits in regional offices. The common strategies range from mentoring trainees, forming communities of practices and being role models, enrolling in in-house and external professional development programs (such as enrolling in Master of Business Administration (MBA) programs), and specifying the criteria for promotion (especially for becoming a partner), to selective recruitment.
FIRM FOUNDATION AND GROWTH

(by recruiting lawyers with similar educational and/or work experience). A typical example of recruiting lawyers with similar educational background is the Tokyo office of Clifford Chance, where all but one partner has qualifications gained in countries other than Japan, with English and American law or MBAs being the most common (Faulconbridge 2008, 508). The international bent of the lawyer profile is partly a reflection of the profile of their major clients, many being TNCs based in Europe or North America. Lawyers trained in Anglo-American educational institutions have a closer cultural distance and are likely to be more effective in communicating and understanding the expectations of clients with similar backgrounds.

As the commercial success of transnational law firms is closely linked to their long-term business relationships with major clients, the expectations and demands of TNCs have an impact on shaping the cultures of practice in transnational legal firms and even the legal profession itself: that is, in defining the norms of the profession. There is a counterargument that the best practices of transnational law firms are increasingly influenced by the rise of emerging economies and their specific demands. Non-Anglo-American legal systems may place importance on proficiency in foreign (local) languages, and the specific relationships between the legal profession and local states. For example, lawyers normally have a more passive role when advising their clients on the interpretation of local commercial law in China than their counterparts in the United States.

Tension between the headquarters and regional offices of transnational law firms can occur if they have to recruit local lawyers for their local offices. These local lawyers have already been molded or socialized into certain work practices and thus have their own forms of territorial and societal embeddedness, either through formal training at (local) educational institutions or on-the-job training (see Faulconbridge 2008). The resultant best work practices are therefore a reflection of the symbiotic relationships between the senior partners and the managing partners in the home and host countries, who have to negotiate and resolve potential conflicts and tensions between the different cultures of work. This is especially the case when the remuneration systems adopted in different offices (e.g., the merit-based model in the New York office but the profit-sharing model in the London office) and the recruitment of star lawyers from rival legal firms cause tension.

Staffing agencies

The staffing industry is another case illustrating the importance of institutions for cultures of work. The staffing industry is inherently geographical, as physical proximity, local knowledge of the labor market, and an institutional environment are crucial to the delivery of its core services. Staffing agencies are (largely) privately owned companies acting as economic intermediaries that recruit permanent and/or contract workers for their clients for a fee. The involvement of professional staffing agencies in the recruitment of (senior) permanent executives can defuse potential accusations of nepotism or the influence of “old-boy networks,” especially at listed companies where the governance of senior management is under public scrutiny (see Beaverstock, Faulconbridge, and Hall 2010). In the case of contracted staff, the employment relationship exists between the staffing agencies and the workers, but the clients determine the work relationship— the staffing agencies pay workers. The highly flexible arrangements for contracted staff allow firms to more easily cover seasonal fluctuations in market demand. This contributes to the popularity of staffing
agencies; for example, the US-based Manpower, one of the major players in the staffing industry, increased the number of countries in which it has offices from 52 to 82 countries between 2000 and 2009 (Coe, Johns, and Ward 2011).

Staffing agencies can be classified into two main categories: general and specialist. The former are largely traditional staffing agencies providing temporary contract staff to their clients, while the latter specialize in recruiting temporary and permanent staff in specific sectors, especially information and communication technology and the financial service sectors. Examples of major general staffing agencies are Adecco (Switzerland), Manpower (United States), and Vedior (Netherlands), while Hays (United Kingdom), Michael Page (United States), and Robert Half International (United States) are the leading international specialist staffing agencies. As a partial reflection of their businesses, general staffing agencies normally have a greater level of geographical presence than specialist staffing agencies. For instance, Adecco has 6600 offices worldwide while Robert Walters, a specialist staffing agency, has only 38 offices, mostly in the global cities where its potential clients are located.

Due to the specific nature of the staffing industry, a decentralized mode of operation is commonly adopted. Marketing, information and communication technology, and human resources (providing the “matching” service to clients) are the key departments in local offices. Some of these offices can be larger than the headquarters of transnational staffing agencies: for example, one of the top five transnational staffing agencies has only 80 staff in its global headquarters (Coe, Johns, and Ward 2011, 64). There are, however, variations among the major players. Vedior is one of the most decentralized staffing agencies; its managers have a high level of autonomy to deal with changing market demands (some of them even adopt a multibrand approach to cater for different market segments). Manpower, meanwhile, has a greater degree of centralization of its management structure, branding, and marketing strategies.

As with the law firms, practices in the staffing industry are shaped by market demand and the institutional environment in the host countries. This includes their regulatory authorities and corresponding welfare regimes (including medical and work insurance), the role of labor unions and industry-wide codes of conduct, as well as the lobbying efforts by relevant actors concerning the regulation of work hours and the work environment, special regulations for short-term employment, and so on.

Variations in national welfare and labor regimes shape the development of the staffing industry in different countries. An understanding of the local job market and institutional environment, including the work culture, is therefore vital. The local regulatory and welfare regimes range from the liberal labor markets of the United States and the United Kingdom, the liberal but still regulated market of Australia, and the highly regulated markets in Sweden and Japan, to the liberalizing former socialist markets in the Czech Republic and Poland (see Coe, Johns, and Ward 2011). For instance, the labor market in Australia is normally regarded as neoliberal. However, it cannot be assumed that the labor market, and thus the staffing industry, is the same as in the United States or the United Kingdom, partly due to its legacy of a relatively egalitarian wage arbitration system for male employees (breadwinners) and the crucial role of the labor unions in collective bargaining and subsequent wage resolutions. Despite the liberalization of the labor market since the 1980s, provisions for certain welfare conditions in employment contracts, such as tea breaks, sick leave, and holiday leave, are still common in Australia.
In Sweden, the development of the staffing industry is heavily influenced by the social democratic welfare state regime and the roles played by different actors. Although the regulation of the industry is liberal, the Swedish labor market is still highly regulated, with its legacy of full employment, the active role of labor unions in collective bargaining, and the resources committed by the state to achieve such goals, from the provision and promotion of full employment to staff training and retaining. These labor market regulations actually facilitate the development of the staffing industry whereby the demand for temporary staff provides a higher level of flexibility for employers in their management of human resources (because of the rigidity of the redundancy processes) and yet fulfills the labor regulations that require, for example, cover for the maternity leave of permanent staff. Instead of lifetime employment (for males) in Japan, the deregulation of employment relations by the developmental state regime and the decades-long economic recession, plus partisan lobbying from labor unions and local agencies, have played important roles in the development of staffing agencies.

In the Czech Republic and Poland, wide-ranging deregulation prior to accession to the European Union (EU) has had a profound impact on the development of various industries and subsequent economic restructuring. Development of the staffing industry has been shaped by the liberalized labor market and the norms, expectations, and economic structure of their respective economies. While both countries have allowed the establishment of transnational staffing agencies – either through greenfield investment or mergers and acquisitions – there are strategic differences in their deregulation. The Czech Republic has adopted the Dutch model for the regulation of the staffing industry, but Poland has liberalized the industry without reference to the experience of any one country.

The territorial embeddedness of the staffing industry is demonstrated by its resistance to the standardization of practices across space and the centralization of control. As staffing agencies need to respond to sector- and region-specific market demands, those operating in transnational markets cannot always use the same operational template everywhere in the world (Coe, Johns, and Ward 2011). This explains the entry of transnational staffing agencies into the Czech Republic and Poland via mergers and acquisitions. While local managers are allowed a high level of autonomy to reflect the importance of local expertise, there is a tension between this and the importance of the global strategic presence of a brand in newly established markets, in that certain aspects of the headquarters’ corporate cultures and practices central to the brand name may not always be adequately represented.

Conclusion

The understanding of the importance of firms has improved significantly since the seminal work of Ronald H. Coase on the theory of the firm. He characterized the firm as an intermediate agent for reducing the transaction costs incurred in dealing with individual economic actors. Oliver E. Williamson emphasized the importance of interfirm (market) and intrafirm (hierarchical) transactions for the functioning of the market economy, and Douglass C. North introduced the new institutional economics.

Based on these significant theoretical foundations, economic geographers such as Meric Gertler have highlighted the interlinked relationships between institutions (formal rules and regulations and informal norms) and industrial
practices. The former not only regulate industrial best practices and conventions regarding the organization of work and labor-management relations in the workplace, but also shape the values, expectations, and even the decisions of economic actors through the functioning of labor markets, regimes of industrial relations, and corporate governance. These diverse institutions create heterogeneous industrial practices across national space, and their reproduction and modification illustrates the existence of distinctive national modes of governance and trajectories of growth in firms rather than any globally convergent world.

The cases of law firms and staff agencies clearly illustrate the central argument made by economic geographers, namely that diverse national formal and informal institutions create a heterogeneous culture of work across space. In so doing, economic geographers have contributed to the ongoing debates about the impact of globalized forces and varieties of capitalism in general, and corporate cultures and regional development in particular, on the issue of whether globalization does or does not lead to uniform patterns of firm foundation and growth.

**SEE ALSO:** Evolutionary economic geography; Institutions and development; New economic geography

**References**


**Further reading**


Firm migration

Jouke van Dijk
Piet H. Pellenbarg
University of Groningen, Netherlands

The location decisions of firms are crucial for the economic development of countries and regions. Therefore, insights into the determinants of firm migration are very important for policy purposes. However, although firm migration occurs frequently worldwide, it is only a minor field of study in the research practice of economic geography and spatial economics. This is surprising because determining the optimum firm location is a key issue for these disciplines, and a firm migration can be understood as a relocation, following a new decision about the best place to establish the firm’s activities when location conditions change over time (the terms “firm relocation” and “firm migration” are used synonymously in this entry). Relocation decisions might even be considered as the most interesting type of firm location decision. In the initial stages of new firm formation, many location choices are hardly felt as real choices. New entrepreneurs often take their hometown as a natural start-up location. In a relocation, the situation is usually completely different. For whatever reason (firm growth, lease expiration, new infrastructure conditions, shifting spatial patterns of demand and supply, labor costs, or labor shortages), the existing location is now the main problem, and the thinking about where to find new and better location conditions is of prime concern. An interesting question is why in the literature other firm demographic events such as firm start-ups and closures, firm growth and decline, and firm spin-offs and spin-outs receive much more attention. Part of the answer is that, in the systematic exploration of the effects of firm demography factors, new firms are assumed to have a larger impact on local and regional economies than migrations.

Yet another weighty reason is the complexity of the subject. The complexity is in the multitude of possible interpretations of locational changes as proper relocations. A lock, stock, and barrel move of the whole firm to another location is easily understood as a proper firm migration; and, when only a part of the original operations is moved to a new location while other parts stay at the existing site, this can be understood as a partial migration. But what about situations where the transfer of parts of the production to another place is framed as contracting out to another independent firm? Or where two firms from two locations merge and restart in a third place? Or where a firm closes down in one location and opens up a new plant making the same product in a different location under a different firm name? In all such cases, part or all of the initial firm’s activities move to another location, but it is more difficult to label them as real firm migrations. The next problem is the relative scarcity of sufficient and reliable data. Firm registration systems tend to be different in type and quality between countries, and sometimes even between regions and cities, which is a problem for comparative research in firm demography. But the more serious problem is that many firm registration systems aim at
Firm migration, spatial scale, and spatial theories

Spatial moves by firms can be classified into three broad categories: intraregional, interregional, and international (Pellenbarg, Van Wissen, and Van Dijk 2002). Intraregional moves mainly concern the industrial suburbanization around the larger urban agglomerations, while interregional migration involves industrial decentralization from the economic core areas to peripheral areas. Intraregional moves can be related to the incubator hypothesis developed by Leone and Struyck, which describes the process of firm birth in central urban areas, with a “natural” follow-up of firm migrations from these central areas to urban peripheries, where the surviving and then growing start-ups seek space for their expansion at a location that is more accessible for their clients and suppliers. Firms moving over smaller distances within urban areas tend to be young and small, and the migrations are not partial but integral. Interregional moves may also be integral if firms move to areas with lower land and/or labor costs, but often take the form of partial moves when subsidiaries are opened in other parts of the country. International migration can be integral or partial for small and medium-sized firms that are undergoing an internationalization process similar to that of multinational (MNL) enterprises. The spatial behavior of large companies is more complicated. They try to realize ownership, location, and/or internationalization advantages according to the OLI framework developed by Dunning by means of export, foreign direct investment, or relocation, or a combination of them. Their relocation behavior is driven either by their search for new international markets and the acquisition of new technological, financial, and marketing competences, or by their continual search for low operating cost locations, especially low labor cost locations. Low labor cost definitely links the international long-distance firm migration story to Gunnar Myrdal’s classic theory of cumulative causation.

On reflection, the juxtaposition between long (international) and short (local and regional) distance moves is not an absolute one. From Table 1 it is clear that there is a gradual transition between the extreme spatial types of global and local firm migration, which can be correlated to the various organizational types of migration: integral move, partial move, setting up of branch plant and joint ventures, and contracting out.

The range of spatial theories that may be linked to the phenomenon of firm migration is much broader than the theories of Myrdal, Dunning, and Leone and Struyck. Investigators seeking to describe, explain, and predict firm migration processes at the different spatial levels have been using neoclassical, behavioral, institutional, evolutionary, and organizational network approaches, more or less in that order. A more detailed description of these five theoretical approaches to firm migration can be found in the elaborate surveys of firm relocation literature by Pellenbarg, Van Wissen, and Van Dijk (2002) and Pellenbarg and Knoben (2012). A relative new approach is followed by Baldwin and Okubo (2014) who introduce firm migration into a combination of the heterogeneous firm trade (HFT) models and the new economic geography (NEG) models introduced by Krugman, where increasing returns to scale create forces...
Table 1 Firm migration and spatial scale.

<table>
<thead>
<tr>
<th>Type of move</th>
<th>Spatial scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Local/regional</td>
</tr>
<tr>
<td>Integral move</td>
<td>SME moves within countries</td>
</tr>
<tr>
<td>Partial move</td>
<td>SME subsidiaries within countries</td>
</tr>
<tr>
<td></td>
<td>SME cross-border moves</td>
</tr>
<tr>
<td></td>
<td>TNC and SME moves from west to east, or from north to south, in Europe, America, and Asia</td>
</tr>
<tr>
<td>Branch plants and joint ventures</td>
<td>SME and TNC cross-border moves</td>
</tr>
<tr>
<td></td>
<td>TNC and SME moves from west to east, or from north to south, in Europe, America, and Asia</td>
</tr>
<tr>
<td>Contracting out</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FDI by TNCs from Europe or USA to Asia</td>
</tr>
<tr>
<td></td>
<td>International purchase of goods and services</td>
</tr>
</tbody>
</table>

Note: FDI = foreign direct investment. SME = small and medium-sized enterprises. TNC = transnational corporation.

that foster spatial agglomeration. In a recent survey paper Arauzo-Card, Liviano-Solis, and Manjón-Antolin (2010) argue that, although most studies relate the choice of the explanatory variables to neoclassical, behavioral, and/or institutional factors, the link with the associated theories is usually weak. They conclude that the analytical framework has remained very much the same since the early 1980s, but that the studies of relocations lack specific theoretical foundations and are essentially empirically driven.

Origin and development of firm migration research

An extensive survey of progress in firm migration research since the 1950s can be found in Pellenbarg and Knoben (2012). This section presents an adapted and abridged version of this survey.
FIRM MIGRATION

Research projects of the 1950s and 1960s in the United States and the United Kingdom

McLaughlin and Robock’s (1949) study Why Industry Moves South was the earliest published study on firm relocation. It describes the mid-century migration flow of manufacturing industry in the United States from the northeastern to the southeastern states, where labor costs were lower and trade unions less active. Several classical firm migration studies concerning the United Kingdom were published in the 1950s and 1960s as well (e.g., Keeble 1968). In these studies, external relocation factors dominate the explanation of the phenomenon, although some attention is given to firm-internal relocation forces, especially the push impulse resulting from output growth. A core interest of the researchers in this period was the impact of firm relocation on local and regional economic growth and decline, and the possibility of influencing firm moves with local and regional policy instruments. Firm relocation was regarded as a means to transfer work and prosperity to lagging regions while at the same time easing the congestion, labor market, and space capacity problems in core regions. Instead of reducing regional disparities by means of a “workers to the work” policy, firm relocation was welcomed to facilitate a “work-to-the-workers” policy. Summing up, the early studies shared a prime interest in the external causes of relocation (push and pull) and the impact of relocation on local and regional economies, including the policy options to stimulate these economies. Several of the early studies also tried to evaluate such policy measures, with mixed results (e.g., Cameron and Clark 1966).

British research dominance in the 1970s

In the 1970s most international publications on firm relocation were British-based. The interest of British researchers in firm relocation coincides with the heydays of British regional policy, which relied heavily on guiding manufacturing firms to assisted areas and using location controls, capital subsidies, and labor subsidies. These firm relocation studies are part of a bigger research wave in this period which sought to measure the effect of regional policy instruments on the economy of the assisted areas. Naturally, the focus of the analyses in this period remained on external relocation factors. The availability of reliable (national) data on firm relocation in the United Kingdom was a great stimulus for firm migration research. Using these national data, Keeble (1976) argued that earlier research had underestimated the importance of firm relocation. In terms of employment, one-third of all industrial movement in the late 1960s in the United Kingdom as a whole is categorized by Keeble as short-distance “spillover” from the large conurbations and 50% as long-distance “between regions” moves to the periphery (Keeble 1976). Keeble also concluded that “regional policy was the single most important variable influencing the sub-regional pattern of migrant industries between 1966 and 1971,” a conclusion that contradicts Cameron and Clark’s (1966) more cautious conclusion about the same issue for the previous decade. Later in the 1970s, Townroe (1979) added to the abundant information about industrial firm migration in the United Kingdom new information about the United States. Interestingly, Townroe was the first to analyze firm relocation from a behavioral theoretical perspective. His study was a first attempt to present a general descriptive model of the location decision-making process. Furthermore, he was the first to suggest that firm relocation research should be integrated with studies of the birth, growth, decline, and death of firms to form a so-called components-of-change
approach, later to be labeled the “demography of firms” approach.

Firm migration research in other European countries

The second half of the 1970s witnessed firm migration studies from many more European countries. Both policymakers and regional scientists were reflecting the same curiosity as their colleagues in the United States and the United Kingdom about the nature of the process of spatial–economic change and the possible results of regional policies, which in the post-war period were revived all over Europe. The late 1970s may be considered the florescence of classical firm relocation research. In the Netherlands an early industrial relocation report was published by SISWO (1967), followed by a national survey study by Pellenbarg (1985). Information regarding France, Italy, Ireland, Belgium, Denmark, and Greece became available as well (for references see Pellenbarg and Knoben 2012). An interesting result of the European survey by Klaassen and Molle (1983) is that in all countries there is a short-distance “industrial sub-urbanization” around the larger urban agglomerations on the one hand, and a longer-distance “industrial decentralization” from economic core areas to peripheral and/or development areas on the other. Most of the European research projects of the late 1970s were still concentrated on the manufacturing sector, which was apparently the most mobile sector of the 1960s and 1970s and, at the same time, the main target sector for regional development policies in most countries. All European studies of this period agree about the nature of the driving forces behind the firm relocation processes. These are lack of space, transport issues, and labor market problems. The first two are the main causes for short-distance moves in and around the urban agglomerations, while the third is the dominant motive for long-distance movements to development areas in national peripheries. So, in line with UK and US studies from the 1970s, external factors were still found to be the main drivers of spatial firm mobility. However, in the course of the 1970s the importance of the labor market push motive gradually decreased. The economic recession of the mid-1970s caused higher unemployment levels, and labor shortages disappeared in many core regions. After this, the magnitude of long-distance relocation decreased.

The 1980s: from the regional to the urban level

In the 1980s international publications on firm migration decreased and the interest of migration research switched to the urban level. Policies such as environmental policies, urban renewal policies, and local development policies were expected to result in firm relocations as well (Pen 2002). For most of these new policies, the urban agglomeration was the most relevant playing field. Despite the change in the spatial level of analysis, the focus of the analyses from this period remained on external relocation factors. The new interest in the urban (i.e., local) level of analysis in this period coincided with a strong increase in short-distance firm relocation. This can be related to the process of the suburbanization of firms, propelled by space shortage and increasing land prices in large cities, parking problems, and growing congestion on city roads and beltways. The urban firm exodus consequently grew, and city governments worried about the erosion of urban employment. Some of the government’s own policies, especially environmental policy and urban renewal policy, reinforced the pressure on firms to leave urban locations. In this context, it was easily understood that firm relocation
FIRM MIGRATION

studies from the 1980s onward concentrated more heavily on urban agglomerations. In the meantime, the majority of firms which left the city during this time were no longer industrial or wholesale firms, but business services.

The 1990s to the present: new approaches and perspectives

In the 1990s and 2000s a few more regional and national studies appeared that repeat the approach of the classics in firm migration studies, that is, monitoring regional or national (manufacturing) firm migration patterns and explaining these on the basis of external factor-oriented interview programs or regression analyses. After the 1980s the interest in firm relocation as a direct remedy for regional development and/or labor market problems disappeared in most countries. Modern regional policy views now start from the premise that regions should independently create the conditions for economic growth and innovation based on their own strengths; regional subsidies should be given to develop these strengths. Key concepts underlying the new policy agenda are regional agglomeration effects, endogenous growth, regional knowledge networks, learning regions, sustained economic growth, and place-based smart specialization.

Firm demography

According to current views, firm relocation is still a relevant issue, as the interesting consequence of the coming of age of generations of new firms developing various locational strategies for coping with sustained growth. Consequently firm relocation is no longer treated as a separate event, but incorporated in what Townroe labeled the “components-of-change approach.” This is in fact a demographic approach to firm dynamics, which gradually gained popularity among regional scientists during the 1990s (Van Dijk and Pellenbarg 2000a). The approach, which is followed by geographers, sociologists, and economists, is variously termed “industrial demography,” “demography of the firm or enterprise,” or “economic demography” (Van Wissen 2002).

Firm-internal factors

Another new research theme that emerged in the 2000s is the focus on the role of firm-internal characteristics as a cause of firm relocation. In this approach, micro-level models are estimated that try to explain firm relocation by incorporating various firm-internal, locational, and firm-external variables (Van Dijk and Pellenbarg 2000b; Brouwer, Mariotti, and Van Ommeren 2004; Knoeben and Oerlemans 2008). Van Dijk and Pellenbarg (2000b) were the first to econometrically show, using firm-level data from the Netherlands, that firm-internal variables, such as firm size, economic sector and migration history/age, locations stress at the present location, and ownership of the building are better predictors of firm relocation than locational and other external factors. Brouwer, Mariotti, and Van Ommeren (2004) performed a similar analysis but included more firm-internal variables and utilized data from more countries. Their results also indicate that firm-internal variables are the most important cause of firm relocation. Besides size and age, they find in particular that changes in the size of the firm (both positive and negative), as well as mergers and takeovers, are strongly associated with firm relocation. De Bok and Van Oort (2011) confirm from an analysis of the Dutch province of South-Holland that firm relocation behavior is affected much more by firm-level attributes (size, age, and growth
rate) than by agglomeration and accessibility attributes.

Impact of firm migration

In recent years studies of the possible impact of firm migration as one of the factors (in addition to, for example, new firm formation) influencing regional economic performance have become more important. The focus is no longer on the migration of manufacturing firms to lagging regions, but on a broader policy perspective of promoting regions as attractive locations, including environmental policies that aim to influence the location decisions of a polluting firm. Cabus, Horemans, and Van Haverbeke (2008) conduct a comprehensive study of the location behavior of firms in Belgium. They start with a traditional investigation of the location determinants, taking into account the characteristics of both the firm and the region. They then investigate the implications of this behavior for the firm and for the region, and end with a series of policy recommendations. Van Oort et al. (2007) analyze how spatial dynamics in firm behavior affect local and regional employment in the Netherlands. They show that the contributions to employment growth from new firm formation and from firm migration are of a comparable magnitude. Furthermore, migrating firms show employment growth that is almost twice as high as nonmigrating firms. In contrast to this, Neumark, Zhang, and Kolko (2006) find that, in the United States, new firm formation is a much more important generator of new employment on a regional scale than the interstate relocation of firms. The difference in spatial scale may cause this difference, because interstate moves in the United States may be comparable to migration between countries in Europe. Daunfeldt, Elert, and Rudholm (2013) analyze the differences between start-ups and firm in-migration using Swedish data on wholesale industries in the period 2000–2004, and conclude that the determining factors are largely the same but that the magnitude differs. Manjón-Antolín and Arauzo-Carod (2011), using data for Catalonia, find that most factors affecting start-up rates are different from those affecting relocation rates, while those that affect both rates do so with different intensities. In addition, start-ups and relocations are positive but are asymmetrically related.

International firm migration

In a study about international relocations of Belgian firms Sleuwaegen and Pennings (2006) find that wages and the market potential of host regions are important determinants for the location choice. With regard to firm characteristics, they show that large firms more often relocate to remote countries. Public aid only plays a decisive role for relocations to adjacent countries with similar factor conditions. This suggests that this type of relocation aid is only distorting competition instead of investing in facilities that improve a region’s attractiveness. Baldwin and Okubo (2014) argue that, until now, international firm migration has been ignored in trade models. In a theoretical model that combines the HFT model and the NEG model, they show that, by allowing firm migration whereby the most efficient firms move to the larger countries, both the large and the small countries are better off. International firm migration also becomes relevant in a world of outsourcing and offshoring, and more recently also reshoring, where firms are looking for optimal locations all over the globe in order to maximize the contribution to the global value chain (Contractor et al. 2010). In addition, climate change and more frequent extreme weather conditions may create the need for firms to relocate away from highly affected regions (Linnenluecke, Stathakis, and Griffiths 2011).
The relocation decision process

As a consequence of the new conception of firm relocation as a specific locational strategy, firm migration research shifted from its long-term one-sided focus on the external causes of firm relocation to the study of the internal causes of firm relocation, and also developed an interest in the structure of the relocation decision process. The new interest in the process of decision-making in firm relocation situations was an important renewal. The identification of external push, pull, and keep factors, which was customary in pre-1990 research, provided only a superficial explanation of firm relocations. In the 1990s there was a desire for a deeper explanation. Earlier interest in decision processes was already observable in the behavioral approach to firm migration. As early as the 1970s Townroe, representing this approach, developed a model with five successive decision stages: (i) stimulus; (ii) problem definition; (iii) search; (iv) formulation and comparison of alternatives; and (v) choice and action. There was little application of his scheme in empirical research, but the 1990s saw a revival of interest in the process of firm relocation decisions. Louw (1996), for example, empirically studied the location choice behavior of (relocating) large offices in the Netherlands. He conceptualized the decision-making process as a sequence of three phases: an orientation phase, a selection phase, and a negotiation phase. His findings indicate that spatial factors (i.e., geographical position, accessibility, parking possibilities, proximity of facilities, public transport, and quality of the spatial surroundings) play an important part in the first two phases, whereas financial and contractual factors are more important in the third phase, when it comes to negotiating a result. Louw’s process-oriented approach of firm relocation decisions has been followed up by Pen (2002). In a study based on an extensive inquiry among over 1000 individual firms, Pen considers relocation as one of six types of location strategies. The average length of the decision-making process proves to be much longer than is often assumed (on average two years) and shows that the number of stages in the decision process varies from a minimum of three to a maximum of seven. Greenhalgh (2008) analyses the firm relocation decision process based on interviews in Tyne and Wear, in the northeast of England, from the perspective of the property market using a three-stage decision model: (i) trigger (stimuli or catalyst provoking initiation of process); (ii) analysis (pursuit; influence of people and factors on process); (iii) outcome (decision made and process concluded). They find that organizations adopt varying degrees of sophistication when making relocation decisions; small firms are more inclined to make decisions based on constrained information whereas larger organizations adopt a more complex approach. Regardless of firm size, key individuals exert considerable influence over the decision-making process and its outcome. In a recent study based on various data sources for the Tokyo Metropolitan Area, Nguyen et al. (2013) study the firm relocation process using a model that also consists of three levels: (i) moving probability; (ii) region choice probability; and (iii) zone choice probability. The results indicate that large firms and older firms have a lower probability of relocation and that migrating firms are more attracted to regions with high accessibility. Finally, spatial factors and land prices in a given zone strongly affect the zone choice decision-making process of all firms. Using individual business surveys for 2006, Hu et al. (2008) examine US firms’ decisions to relocate or expand in the past as well as their intention to relocate or expand in the future. The results indicate that factors related to firms’ internal characteristics, features of location sites, and
the general economic environment may affect firms’ past and future decisions. These factors are found to be generally consistent in their impact on past and future decisions with several noticeable differences. The hypothesis that firms that have moved once will move again in the future (footloose firms) is supported by this study.

Recent new developments

A more specific new strand of research deals with firm relocation at the neighborhood level because the out-migration rates of firms show large differences between regions. However, based on Dutch data for 2010, Sleutjes and Völker (2012) find that neighborhood characteristics only play a modest role, while growth ambitions and the size and quality of the business premises are most important. In line also with Greenhalgh (2008), this indicates that the real estate market is likely to be an important factor for firm relocation. Risselada, Schutjens, and Van Oort (2013) focus especially on the role of real estate and conclude that real estate factors are important determinants of the likelihood of firms relocating. Like Van Dijk and Pellenbarg (2000b), they find that property ownership has a negative effect on relocation. Residing in high age and large property sizes also limits firm relocation.

With regard to the international migration of MNL enterprises, Lampón, González-Benito, and García-Vázquez (2015) analyze relocation processes in the Spanish automobile components industry for the period 2001–2008. While acknowledging that location advantages play a relevant role, they show that international relocation changes in production geography can, to a great extent, be explained by corporate strategies, other decision-making mechanisms, and firms’ characteristics. Knoben and Oerlemans (2008) add to other studies, like Van Dijk and Pellenbarg (2000b), by explicitly taking into account the relations between a firm and other organizations in its environment (i.e., relational variables). They show that, compared to a baseline model including internal, external, and locational variables, adding relational variables significantly enhances the explanatory power of the models, and that this varies according to the strength and geographical distance of their relationships. Bodenmann and Axhausen (2012) show that local taxes have a very positive pull effect on firm relocation for the St Gallen region in Switzerland over the period 1996–2006. Their findings also confirm the role of distance, and they find significant differences by sector. Weterings and Knoben (2013) analyze Dutch data for 1996–2006 and confirm the classic distinction between short-distance moves caused by lack of space and that long-distance relocations are mainly influenced by regional characteristics. They also find that the spatial concentration of similar or related firms, a higher level of urbanization, and research and development (R&D) intensity keep firms in a region, but that firms leave regions with a higher share of innovative firms. A recent study by the City of London Corporation (2014) on the effects of the current economic crisis shows that firm migration is acting as the key component of change, compensating to a large extent for the negative effects of job losses due to layoffs and the bankruptcy of existing firms in the City.

Theory and methods

The location decisions of firms are crucial for the economic development of countries and regions and, therefore for policy purposes, insights into the determinants of firm migration are very important. In a recent survey paper Arauzo-Cardó, Liviano-Solis, and Manjón-Antolín (2010) try to shed light on this in a critical assessment of the methods and empirical
Firm migration results of about 50 studies on location behavior, including relocations. They conclude that for policymakers there is room for action in areas such as education, infrastructure, and taxation. Future research is needed, especially in the analysis of behavioral factors as underpinning for the theoretical foundations for firm relocations and of the impact of public policies. They also argue that more sophisticated econometric specifications, using different levels of geographically aggregated (panel) data and/or spatial econometric techniques, are needed to provide further insights.

Conclusion and research frontier

Firm migration or relocation is a very important determinant of regional economic dynamics. This entry provides an overview of the present theories and empirical findings on this topic. The spatial scale is important because the causes and effects differ substantially for firm relocations within urban areas and between regions within countries, and for international relocations between countries or even continents. Over time the research focus shifted between the spatial scales. Firm migration research started in the 1950s and 1960s by studying the relocation of manufacturing firms in the context of the regional development policies. Early on, firm-external factors were the main explanatory factors, but later on the role of firm-internal factors and the phases in the decision-making process became more important as better data and methods were available. In addition, firms in other industries were analyzed and the differences in the determinants and effects of new firms compared to relocating firms were the subject of study. Most recently, external factors like neighborhoods, real estate, and the variety in the regional economy have been linked to firm migration. With regard to the methodology, there has been a shift from purely descriptive studies to the use of sophisticated econometric methods.

The location decisions of firms are crucial for the economic development of countries and regions and, therefore for policy purposes, insights in the determinants of firm migration are very important for developing effective policy measures for regional economic development, but also for decision-makers in firms who have to survive in an increasing competitive world market. Although substantial progress has been made since the mid-1950s, there is still room for further development. Firm migration is now studied in different fields and by researchers from different disciplines. The theoretical foundations of firm relocation could be enhanced by greater integration of traditional location theory and regional development theory, and by insights from behavioral and evolutionary theories the organizational network approach, trade theories, and management science. Empirical questions like the effects of firm migration on firm growth, its impact on regional economic development (such as the employment), a firm’s territorial embeddedness in a location, further firm growth after migrating, corporate strategies of MNEs and so on, deserve further study. More attention is also needed on the interaction with other firms and the real estate market, and the impact on regional development in the context of new policy concepts like place-based smart specialization. This can be facilitated by the increasing availability of employer–employee linked micro-data and big data to apply more sophisticated methods like spatial and panel econometric methods, GIS, and methods for network analysis.

SEE ALSO: Economic geography: spatial interaction; Firms; Geography and the study of human–environment relations; Industrial geography; Industrial location theory;
Location-allocation models; Location and multiplant firms

References


**Firm Migration**

*The Recent Location of Manufacturing Plants in the South.* Washington, DC: Committee of the South, National Planning Association.


SISWO. 1967. “Verplaatsing industriële bedrijven” [Relocation of Industrial Firms]. Deelrapport 1, 2. SISWO, Amsterdam.


Business firms in their infinite variety are situated at the heart of the modern economy. They are the creators and destroyers of products, services, jobs, and markets. As the pivotal economic actors driving the rise and fall of cities, regions, and nations, firms are a core interest of economic geographers. A general understanding of the firm includes that of its existence, boundaries, organization, and competence. Geographers are interested in the spatial dimensions of these issues. The focus on place, space, and scale is a key difference between geography and the other social sciences in the way the firm is theorized (see Taylor and Oinas 2006). Economic geography has emphasized location decision-making, the value added of co-location, localized collective learning and knowledge spillover, interfirm networks and local–global dynamics, and the impacts of formal and informal institutions at multiple geographic scales as they shape business transactions, organizations, and performance. However, a comprehensive, consistent, geographic conceptualization of the firm that integrates key insights from the various theories of the firm has yet to be established.

Every firm has its unique geography of creation, expansion, integration, and differentiation. Geographers see firms as intrinsically spatial and territorial, and are deeply concerned by the firm–territory nexus. As human–devised mechanisms for coordinating and motivating the activities of individuals and groups, firms are internally diverse, complex, multidimensional, and spatially variegated. A business firm is generally characterized by: (i) an ownership structure, an employer–employee relationship, and a management hierarchy; (ii) market-oriented, for-profit production; (iii) product or market specialization and internal division of labor and tasks; (iv) durable association and interaction of a group of people; (v) embeddedness in the network of suppliers and customers; and (vi) a life cycle from birth to death. Relatedly, indispensable functions of the firm include acquiring and processing information, discovering knowledge, allocating resources, eliciting efforts, coordinating tasks, evaluating performances, settling disputes, and governing conflicts (Dosi and Marengo 2007). Broadly conceived, a business firm’s core functions can be categorized as either incentive alignment (motivational governance) or problem-solving (cognitive competence), both of which are geographically mediated.

Surprisingly, interest in the firm in economics and the social sciences in general has been relatively new, partly because the mainstream has, for a long time, adopted a rational choice perspective based on the *Homo economicus* assumption: that humans are hyperrational and narrowly self-interested actors who have complete knowledge and the ability to make instant and perfect judgments. Accordingly, the firm has been viewed from a production function perspective, as a “black box” that transforms inputs into outputs. Similarly, a “complete contracts” approach has dominated, which assumes that all economically significant contingencies can be foreseen and incorporated into enforceable contracts.
The emphasis on rationality and de-emphasis on uncertainty are two sides of the same coin. The *Homo economicus* assumption rules out the relationship of the limited human mind to the complex environment. When there is ubiquitous uncertainty, imperfect information, and cognitive inadequacy, the heterogeneity of firm–environment interplays comes to the fore, and the firm-specific centralization of decision-making – a process termed “cephalization” by Knight (1921) – becomes imperative and inevitable. Knight drew attention to the role of the entrepreneur in attempting to reduce uncertainty. The entrepreneurship-based view of the firm focuses on the firm’s entrepreneurial, judgmental decision-making under conditions of uncertainty for opportunity recognition and resource mobilization. Because markets for entrepreneurial judgment are closed, one must first start a firm before one can exercise judgment; moreover, judgment implies asset ownership and organizational control. Economic geography still underplays the significance of entrepreneurship processes, especially how firms are first created, thus missing the first crucial point in the development of a firm and cluster.

The transaction cost approach pioneered by Ron Coase (1937) and elaborated by Oliver Williamson (1975) also attempts to leaven the rationalist, maximizing assumptions of the complete contracts model by introducing the concept of bounded rationality, wherein firms do not have all the necessary information and calculative ability with which to make decisions, especially given behavioral opportunism. In this view, the use of the price mechanism can be costly, and transactions that are recursive and associated with substantial uncertainty and transaction-specific investments are more likely to take place within hierarchically organized firms rather than spontaneously coordinated markets. The transaction costs model implies that contracts can never be complete, because it is either impossible or too costly to specify all the unforeseen contingencies in a transaction. Geographers have generally emphasized the role of spatial proximity and social networks in reducing transaction costs, but the precise mechanism and causality have yet to be further specified. The transaction cost approach places exclusive emphasis on incentive alignment; thus, it has also been termed the governance-based view of the firm. Even with a relaxed assumption of rationality, the focus is on the moral rather than cognitive aspect of uncertainty, and the firm is still assumed to be striving solely for profit maximization and productivity optimization.

In contrast, the behavioral theory of the firm, popularized by Cyert and March (1963), drops the assumptions of profit maximization and perfect knowledge, and focuses on the internal decision-making structure in firms. The firm is seen as a coalition sitting within a network of relationships and a site of decision-making in the face of uncertainty, conflict, problem-driven search, learning, and adaptation. It is argued that the decision-makers of firms are satisfiers rather than maximizers. The firm’s decision-makers often have diverse and conflicting goals, expectations, and opinions under uncertainty. When goals are not met, search behavior is instigated, initiating learning and adaptation. The behavior approach was introduced into economic geography in the 1960s by Alan Pred, who regarded firms as locational decision-makers constrained by their access to, and their capacity for processing, information. But there is as yet no systematic framework integrating behavioral factors into economic geography.

The evolutionary view of the firm is inspired by the behavioral perspective, but it has the dynamics of knowledge generation at its core, along with Darwinian notions of variety, inheritance, retention, and selection. The emphasis
is on firm-specific knowledge that is complex, tacit, distributed, and context-dependent. The analytical entry point is the variety arising from the microbehaviors of firms, directed by durable, inheritable, selectable, quasi-genetic routines. Drawing on the behavioral approach, the evolutionary view conceptualizes firms as satisficers rather than maximizers: a “search” for new routines and technologies is triggered when performance falls below the satisfactory level. It holds that superior routines are selected through market competition. Evolutionary economic geography, developed on the basis of an evolutionary view, argues that largely localized processes of search, learning, spinoffs, and labor mobility are prime vehicles for routine replication and the formation and persistence of clusters (Boschma and Frenken 2006).

From a different angle, the resource-based view (RBV) of the firm seeks to account for the creation, maintenance, and renewal of firm-specific competencies by emphasizing the resources of firms (Penrose 1959). In contrast to the knowledge-centric view, access to and adaptive accumulation of key pecuniary assets is seen as being just as important as the knowledge used to deploy them in production. Although RBV is criticized for its tautology, it does remind geographers of the importance of control over and access to nonorganizational, nonknowledge, and location-specific resources.

The dynamic capabilities framework synthesizes RBV and the evolutionary view of the firm. It defines the firm as an ecosystem for sensing, seizing, and managing market and technological opportunities. In this view, market competition is centered on control over firm-specific resources in a dynamic setting, emphasizing dynamic competencies based in “layer upon layer” of path-dependent development involving complex interactions between resources, strategies, and routines (Teece 2007, 1327). In addition to efficient routines, privileged access to and control over strategic, complementary, and pecuniary resources that have been inherited from the past and/or are aided by corporate and state power can be at least as important. Geographically, this means that, to a certain extent, the facilities, assets, and human capital created by previous rounds of politically mediated investment and development in a region constitute a suitable basis for subsequent production, innovation, and profit-making. How conflicts are triggered and mediated, and mitigated by territory-specific institutions, is an important subject for future research.

Regardless of the focus on problem-solving or incentive alignment, geographers increasingly take a relational embeddedness perspective, seeing the firm as a constellation of socially constructed networks controlled by social actors with activities spread across space but embedded in particular places. The network ties are theorized, vaguely, as conducive to the building of mutual trust and understanding, thus helping to reduce both communication and transaction costs, while making the firms’ boundaries open and porous. The emerging literature on global production networks takes these networks further as the foundational unit of analysis for understanding the global economy.

However, when the boundaries of the firm are seen as fuzzy, permeable, and multiple, it risks making the firm an awkward if not irrelevant analytical category, notwithstanding the legal/contractual structures that embrace and surround it. As an economic institution grounded in the laws and regulations of its particular institutional environment, the firm has boundaries that are delineated and precise. Firms are not only the creation of social relations and networks, but also the creation of government and politics. It is up to the state to create stable and reliable conditions under
which firms organize, compete, cooperate, and exchange. How property rights as social relations are defined, enforced, and contested is a geographic variable and needs to be explored. The size of the risk premium associated with each transaction varies in accordance with the territorial-specific characteristics of the contract enforcement mechanism. A particular institutional context offers a unique payoff structure and gives rise to a particular kind of demand structure for different kinds of knowledge and skills. The kinds of information and knowledge required by the entrepreneurs and managers are in good part a consequence of a particular institutional context. Given the distributed nature of knowledge, firms need not only to develop a better absorptive capacity internally but also to build and maintain an increasing number of external “knowledge nodes” with user communities, public and private universities, technical service institutes, and partnering firms, whose generation and interaction are shaped by state-engineered institutions. The question of how the microbehavior and evolution of firms are shaped by and coevolve with macro, multisclar societal institutions, structures, and relations has yet to be answered by geographers.

SEE ALSO: Economic geography; Evolutionary economic geography; Global commodity/value chains; Global production networks; Industrial agglomeration; Industrial districts; Industrial geography; Industrial location theory; Industrial upgrading; Innovation and regional development; Institutions and development; Labor geographies and the corporation; Local embeddedness; Technology and development; Uneven regional development

Further reading


Flexible labor markets

Tara Vinodrai
University of Waterloo, Canada

The rise in flexible labor markets is an outcome of globalization, deregulation, technological change, and the transition toward a knowledge-based economy which encompasses new forms of work and organizational practices. Characteristic of the rise of flexible labor markets are higher turnover or churn rates, decreased union densities, eroded worker protections, and the increased use of subcontracting and temporary, part-time, and casual work arrangements. Workers who are exposed to these practices represent a growing pool of contingent labor that endures high levels of precariousness and risk. This contingent labor pool is composed primarily of people in marginal groups, including women, visible minorities, immigrants, and migrant workers. However, the precise nature of flexibility in local, national, and global labor markets is a product of the actions of, and the negotiations and struggles between, firms, workers, and other actors, which in turn are also shaped and constrained by institutions organized at a variety of geographic scales.

Theoretical perspectives on the study of flexible labor markets have evolved over time. Initial work on flexible labor markets drew on early theories of labor market segmentation, highlighting the dual nature of the labor market, which included primary and secondary labor market segments. In the dual labor market model, the primary labor market segment was generally characterized by good employment and working conditions, high wages, and stability achieved through well-defined rules and structures for advancement. The secondary labor market included lower-wage workers, with fewer prospects for advancement, less job security, and poor working conditions. Under the dual labor market model, the secondary labor market was characterized by flexibility, in contrast to the more rigid and stable structures of the primary labor market.

However, as scholars began to study and document the transition from a Fordist economy toward an economy labeled as “flexible specialization,” “post-Fordism,” or other monikers, it became clear that flexibility was increasingly pervasive throughout the labor market. In other words, both high- and low-skilled workers faced flexible labor market conditions. Moreover, scholars also observed that the degree of flexibility and the nature of labor markets varied geographically and by industrial sector. Thus, throughout the late 1980s and 1990s, economic geographers and other scholars contributed to our understanding of how restructuring in various industries affected labor markets at the national, regional, and community levels. Most of this work grounded the shift in the desire of firms to achieve flexibility in their production processes, as well as the concurrent responses of governments to enable and support business competitiveness under globalization. This work highlighted the array of organizational practices used by firms, including subcontracting arrangements, flattened organizational structures, teamwork, as well as the use of automation and other technologies to become flexible in their responses to changing market conditions. Moreover, earlier theoretical work tended to
be informed by the experiences of advanced, Western economies. Subsequent work noted the important shifts in the social regulation of labor markets and, therefore, the markedly differential effects felt by different groups of workers. Indeed, some observers have suggested that flexibility has always been a characteristic of labor markets, but its reach has extended beyond marginal groups (e.g., women, visible minorities, etc.) to affect those most often privileged in the labor market. Thus, while there continues to be an emphasis on the structural elements of labor market flexibility, more recent work also highlights the experiences of workers in the labor market and their role as active geographical agents that shape and rework labor market structures and employment relations in an effort to improve wages and working conditions. Recent work also has been more sensitive to the commonalities, differences, and relationships between the Global North and the Global South, as well as reflecting a greater awareness of social differences.

As a result of this evolution in how flexible labor markets are conceptualized, the contemporary geographic literature on employment and flexible labor markets emphasizes several important dimensions from both a theoretical and a research standpoint: (i) the role of deregulation and the pervasive influence of institutions; (ii) the organizational practices of the firm; (iii) the emergence of new forms of work; and (iv) the growing importance of intermediaries as agents for labor market change. Each of these is discussed in detail.

First, scholars emphasize that labor market flexibility is shaped by institutions and government policies organized at a variety of geographic scales, as well as by employment regulations and standards that govern how labor markets operate. In response to competitive pressures under globalization, governments and businesses have attempted to remove rigidities in the labor markets. Underlying the motivation toward deregulation is a belief that this will improve competitiveness and reduce costs for business by allowing firms to easily respond to changing marketing conditions and reduce their overhead costs. This is often achieved through weakening the power of labor organizations such as unions, reducing broader welfare provisions, and changing (un)employment policies, often involving workfare schemes. However, the nature and degree of flexibility in labor markets is both historically and geographically contingent. For example, the erosion of social welfare provisions and protections afforded to workers in earlier periods toward workfare arrangements is particularly prominent in the United States. Moreover, even within the United States, state-level legislation, such as right-to-work laws, results in unevenness in terms of how flexibility can be achieved. In this latter case, right-to-work legislation effectively limits the formation of unions, leaving power primarily in the hands of the employer. At the other end of the spectrum, in Scandinavian countries where there have been long-standing social welfare traditions, “flexicurity” arrangements enable the state to provide security to individual workers if they become unemployed while simultaneously allowing firms to achieve flexibility. In many countries in the Global South, where historically workers’ rights have been weaker, the ability of unions and other labor organizations to improve conditions has been greatly influenced by regional and national policies and institutional arrangements. Overall, these observations emphasize that institutions organized at a variety of geographic scales shape the nature of flexible labor markets.

Accompanying these shifts in governance and (de)regulation has been the decline of permanent, stable full-time jobs with benefits, as firms engage in a variety of organizational practices,
including outsourcing and subcontracting, as well as adopt and implement technologies intended to reduce labor costs and achieve production efficiencies. For this reason, scholars have often examined labor market flexibility through the lens of the firm. Indeed, the main advantages of a flexible labor market are realized by firms that are able to achieve both functional and numerical flexibility as a means to reduce costs and improve efficiency. Functional flexibility refers to the ability of firms to deploy workers to meet particular demands. It is often associated with teamwork, multiskilling, and other practices that allow firms to adjust their workforce internally based on their current needs. At the same time, numerical flexibility allows firms to adjust the size of their workforce through the use of a number of employment practices including the use of short-term contracts; subcontracting arrangements; casual, temporary, or part-time employees; and other arrangements. While the use of these nonstandard forms of employment can confer benefits on firms, there are clear disadvantages for workers themselves, who are exposed to high levels of personal risk. In aggregate this has led to the rise of a contingent workforce that collectively has little job security and access to few benefits. Overall, the prognosis is that the contemporary flexible labor market is highly unequal and segmented, leaving both high- and low-skilled workers in a precarious position.

While an assumption in earlier work was that it was primarily low-skilled workers who were exposed to the risks of the flexible labor market, the contemporary reality that flexibility is a common characteristic throughout the economy brings with it other observations. The transition to a knowledge-based economy has meant that new forms of work are emerging, including occupations where rules, standards, and norms are not well established. For instance, work in relatively new creative industries such as digital media is often precarious and exposes individuals to high levels of economic risk, involving nonstandard forms of work including self-employment, freelancing, and short-term contracts. Work in digital media and other creative industries is often project-based. However, it should be noted that project-based work has become widespread across other sectors of the economy. As an organizational form, it is inherently flexible as it allows the assembly of teams with different skill sets on an as-needed or temporary basis, and often involves the regular movement of workers between projects and employers. This is highly beneficial from the point of view of innovation and learning as it allows for experimentation and the ongoing creation of novelty. However, it also confers considerable risk on individual workers and their long-term career development, as this type of work is not stable and comes with neither long-term security nor collectivized health and pension benefits. Moreover, project-based work often relies on participation in social and professional networks that allow individual workers to gain access to important knowledge about upcoming projects and opportunities, as well as to learn various trade secrets and industry-specific practices. As such, project-based work often requires that workers trade on their personal identity characteristics, in addition to skill. As scholars have noted, the social orientation of these emerging flexible labor markets means that they are deeply imbued with power relations, have a higher chance of being exclusionary, and lead to unequal access to opportunities, depending on who you know rather than what you know. The emerging evidence suggests that these power relations reinforce exclusions based on gender, race, class, and ethnicity.

Finally, the shift toward flexibility in both high- and low-skill employment has led to
the growth of labor market intermediaries, organizations that help to mitigate labor market risk by matching employers and employees. While this means new and expanding roles for traditional intermediaries such as unions, guilds, and professional associations, it has also meant the growth and emergence of new forms of intermediaries ranging from temporary staffing agencies to online job clearinghouses to various community-driven worker collectives. In addition to intermediaries that directly broker the employer–employee or job seeker–employer relation, emerging intermediaries may also provide services, assist in skills and career development, facilitate networking activities, monitor changes in practice, and engage in advocacy. However, these intermediaries are not always benign. Certainly, the immense growth of the global temporary staffing industry is a symptom of an increasingly flexible labor market; but, as profit-seeking organizations, they often actively reinforce labor market conditions that leave workers facing precarity and risk in their everyday lives. Moreover, these organizations mediate supply and demand in local, national, and global markets, thereby shaping global migration flows and networked relations between places in the Global North and Global South.

Overall, current research accepts flexibility as a core feature of the contemporary labor market. However, there remain open questions about how these processes differentially affect particular groups with shared identities based on race, gender, ethnicity, class, and other markers of difference; how different social groups organize at and across different geographic scales to change their circumstances; what new or different intermediaries, networks, and governance arrangements are emerging that either mitigate or reinforce the risks and issues associated with flexible labor markets and produce new geographies and networks spanning the Global North and Global South; and how these are embedded within and vary between different geographic and institutional contexts.

SEE ALSO: Deskilling; Flexible specialization; Intermediaries; Labor market; Labor market segmentation; Precarious work

Further reading


Flexible specialization

Godfrey Yeung
National University of Singapore

Flexible specialization refers to the phenomenon of a closely linked network of firms, from suppliers to distributors, that engage in the production and distribution of nonstandard, specialized products to cater for ever-changing consumer demands. It is a partial reemergence of the craft model of production, which consists of flexible, specialist firms, based on flexible technologies, skilled workers, and a supporting industrial community. The transformation from the mass production of standardized products under Fordism to the volume production of nonstandardized products under flexible specialization illustrates how different forms of production organization can shape the characteristics of regional economic activities over time.

Fordism and post-Fordism

Fordism is the notion of a modern economic and social system based on the mass production of standardized products in large volume using purpose-built machinery and unskilled/semiskilled labor, a form of production initially developed by Henry Ford for the production of the Model T automobile in the Ford Motor Company in 1908. By utilizing purpose-built machinery to break down complex manufacturing processes into simpler, repetitive, and routine tasks that could be performed efficiently by unskilled/semiskilled workers along assembly lines driven by conveyor belts, the Ford Motor Company enjoyed economy of scale in the production of standardized products in large volumes—all Model Ts were black—and marketed Model Ts as the first affordable passenger vehicle by undercutting other automobile manufactures through a (much) lower retail price.

When parts and components are delivered in large and infrequent batches from suppliers, some of which could be located far away from the assembly plant, the assembly plant has to build large warehouses nearby to hold a large inventory of parts and components under the “just-in-case” supply distribution system to support large-scale production and avoid potential downtime in assembly lines. The firm normally keeps a distant working relationship with its large number of suppliers, and procurement is largely based on pricing: that is, there may be no long-term working relationship with suppliers (Dicken 2011, 147).

The dominance of the Fordist mode of production came to an end during the mid-1970s, when the developed economies ground to a halt due to rising inflation and growing unemployment after the 1973 oil crisis. During the recession, producers were left with plenty of unsold and out-of-fashion products, especially products with short product life cycles, such as consumer electronics. Neo-Marxists (including regulation theorists) argue that Fordism is based on the unsustainable regime of the accumulation of domestic mass production and mass consumption that drove post-World War II economic growth in capitalist countries (Jessop 1992).

Distinct from the neo-Marxist emphasis on the intrinsic crisis of the regime of accumulation, Michael J. Piore and Charles F. Sabel (1984),
among others, argued that the forms of the organization of production have shaped the characteristics of regional economic activities and their performance. The move away from the mass production of Fordism to a flexible form of organization of production was a direct response to the change in consumer demand for differentiated and “trendy” products in an increasingly globalized world (Scott 1988; Storper and Scott 1989). Moreover, manufacturers in developed countries were no longer able to compete on cost alone, as generic products were readily available from emerging economies at even lower costs.

Flexible specialization

The debate on regional modes of production was largely triggered by the publication of Piore and Sabel’s (1984) modern classic, *The Second Industrial Divide: Possibilities for Prosperity*. They described the second industrial divide as characterized by a transition from the technological paradigm of Fordism to flexible specialization. The economic slowdown was associated with the crisis of the Fordist production system, as the mass production of standardized products with dedicated machines and unskilled workers was unable to meet consumer demand for specialized and distinctive goods in the globalized economy.

The reemergence of some forms of the craft model of production occurred as a result of a demand and supply-side response in the changing global economy, whereby producers employed skilled workers and installed highly adaptable machinery to cater for consumer preference for customized products in various forms of product mix. Piore and Sabel (1984) termed this new mode of production *flexible specialization*: a network of flexible, specialist firms, with flexible technologies, skilled workers, and a supporting local industrial community. In the twentieth century, the Fordist mass production era signaled the end of small-scale craft production. Piore and Sabel (1984) argued that abandoning the mass production of standardized products and the return to craft methods of specialized production in the 1980s represented a second industrial divide (which is also the title of their book) in industrial societies.

The main distinctive features of flexible specialization in industrial districts compared to the Fordist mode of production are (Dicken 2011, 101) as follows.

- While the Fordist mode of production used complex purpose-built machinery to produce standardized products in large volumes, flexible specialization is characterized by highly flexible production methods using modular component systems deployed by a number of independent small-scale family-owned and craft-based firms, each of which tends to cover an individual phase of production and is normally coordinated by a larger (sometimes trading) firm to produce economy of scope, rather than of scale.
- Instead of using largely unskilled or semiskilled workers to perform specific, repetitive, routine tasks as in Fordism, workers (or workers-cum-owners in the case of small-scale family businesses) in flexible specialization have multiple production and practical skills, including the skills to maintain their machines. The highly flexible modular production system allows the local network of firms to adjust its product specifications, or even shift to new products without the costly retooling of machinery, to produce a wide range of different products of the volume required to meet effectively the ever-changing market demand.
- Rather than keeping a large inventory at warehouses near the assembly plant under
FLEXIBLE SPECIALIZATION

the “just-in-case” supply system of Fordism, firms in industrial districts are connected by specialized transaction networks of trusted suppliers and distributors. Local suppliers are crucial in the “just-in-time” delivery system, whereby specialized and quality parts and components have to be delivered in frequent, small-volume batches to the assembly plant to minimize the inventory and its associated costs (hence, the costs of switching to new products or specifications is lowered). The firm also has a close working relationship with its suppliers. The significant advancements in information communication technologies, especially the adoption of ERP (Enterprise Resource Planning), have facilitated logistical management by the provision of real-time demand information for supply and distribution chains.

Critics of flexible specialization have pointed out that the empirical foundation of research into this new organization of production is largely based on West European industrial districts, typically textiles in Prato and the engineering and ceramics sectors in Emilia Romagna of the Third Italy, according, for example, to Piore and Sabel (1984), and Russo (1985).

Indeed, the practicality of the flexible specialization mode of production is now in question due to the recent economic crisis in the Third Italy. In addition to the wider economic crisis in the European Union and rising production and other miscellaneous costs in Italy, the decline of industrial districts in Italy is partly due to the apparent revival of mass production (hence, lower costs) in emerging economies. This is especially the case in the “fast fashion” industry, as Chinese manufacturers can produce (very) similar products as quickly as Italian artisans/designers and at much lower costs (fast and large-scale production), partly through imitation or even counterfeiting originals.

SEE ALSO: Fordism/post-Fordism; Industrial agglomeration; Industrial districts; Just-in-time production system

References


Further reading


Fluvial depositional processes and landforms

Paul F. Hudson
LUC The Hague, Leiden University, Netherlands

Fluvial deposits and landforms constitute a broad and diverse range of Earth surface processes, and are intricately related to human interests and influences. Fluvial sediments are deposited by overbank or lateral accretion processes which result in a complex assemblage of deposits organized within a floodplain mosaic. The dominance of specific types of depositional processes depends greatly upon the fluvial sediment and discharge regime, which should be considered from a drainage basin perspective. Knowledge of fluvial sedimentary processes and their depositional landforms is important to a variety of disciplines, especially sedimentology and geomorphology, engineering, hydrology, ecology, and archaeology.

Earth scientists study fluvial deposits to gain insights into patterns of flood sedimentation and relations with climatic or anthropogenic forcings. Recognition of distinct types of fluvial deposits is essential to understanding the evolution of river valleys and their adjustment to sea level fluctuation. Geotechnical properties of fluvial deposits are essential for engineers to consider, such as in the management of storm runoff and the design of flood control systems, as well as in transportation infrastructure. Fluvial depositional processes also create significant floodplain topography, which is important from the standpoint of flooding. Fluvial landforms, such as natural levees, have been utilized for millennia by ancient and prehistoric civilizations for settlement and agriculture, and archaeologists are interested in active and relict fluvial deposits in view of their geoarchaeological potential. Because overbank fluvial landforms are created from numerous individual flood events, they represent a climate- and landscape-sensitive archive with great promise for furthering our understanding of watershed-scale hydrologic variation in response to climate and anthropogenic change.

There are a variety of ways to classify fluvial depositional landforms, including mechanisms of sediment transport, morphology, hydraulics, and whether sedimentary deposits are cohesive or noncohesive. But fluvial deposits should always be viewed from the perspective of the drainage basin in which they have been constructed. Here the important watershed characteristics that influence fluvial deposits include scale (basin area), discharge regime (magnitude, frequency, seasonality), sediment load (suspended, bedload dominated), lithology (crystalline, calcareous), land use and land cover, and drainage basin morphometry (slope, shape, relief). An important drainage basin control is stream power, and specific (or unit) stream power has been shown to be an effective means to discriminate general styles of fluvial deposits (Nanson and Croke 1992). Similar to other components of fluvial systems, specific fluvial deposits are scale dependent and increase in size with drainage area due to mutual increases in sediment load and streamflow (energy). The scale dependence of fluvial deposits suggests that the concept of effective discharge applies to different types of fluvial landforms, especially point bars, crevasses,
and natural levees because of their relation to channel morphology.

From the perspective of a methodological framework, fluvial sedimentology and geomorphology provide useful approaches for examining the processes controlling the development of floodplain deposits and landforms, and in relating them to the broader topic of river metamorphosis and evolution. The science of fluvial sedimentology is mainly conducted by geomorphologists and sedimentologists within university physical geography and geology programs, with government agencies having an important role for much primary data collection (e.g., deep cores, soil mapping, geospatial data, hydrologic data). Understanding active sedimentary processes requires fieldwork and instrumentation, such as the measurement and description of bedforms, topographic surveying, coring, and the mapping and collection of flood deposits. Laboratory work includes chemical and physical characterization of field samples, with the latter utilizing the modified Wentworth scale to formally classify the particle size of sedimentary deposits. The use of color-infrared imagery, combined with the increasing availability of high resolution LiDAR digital elevation models (DEM), greatly facilitates the identification and mapping of fluvial depositional features such as crevasses, old channel belts, and terraces. Establishing the chronology of fluvial deposition and landform evolution requires a combination of relative and absolute dating techniques, with the latter most commonly including radiometric approaches or field-based instrumentation. Finally, indices of the above noted data types are commonly analyzed by mathematical and statistical approaches using computer-based hydraulic models (e.g., Hydrologic Engineering Center – River Analysis System (HEC-RAS)) within the framework of a geographic information system (GIS).

**Form and process**

Conventionally, fluvial landforms are created by two major sets of depositional processes: lateral accretion and overbank sedimentation. Within these broad categories, however, are a range of specific processes and depositional features. Sediment sorting during overbank conditions results in a lateral fining of flood deposits whereas sediment sorting associated with channel migration and bar formation results in a vertical fining-up sedimentary sequence.

**Lateral accretion processes**

In laterally migrating channels, coarse-grained (noncohesive) bedload is deposited along channel margins, and bedforms are the key sedimentary signature where median particle size ranges from coarse silt to gravel. Deposition occurs when the critical shear stress required to entrain particles of a specific size exceeds boundary shear stress. As such, the pattern of channel deposits approximates that of the distribution of channel boundary shear stress (Gomez 2006). Thus, the coarsest channel sediments are deposited within the channel thalweg, with sequentially finer sediments deposited outwards towards the channel margins as channel depth and velocity decrease. A typical sequence along an asymmetric channel cross-section (alternate bar or point bar) is that coarse gravel or cobbles are deposited within the channel thalweg at the cutbank, with sand to coarse silts deposited within shallow reaches towards the channel margins (inside of bend). Lower flow regime (Froude number <1) bedforms include ripples and dunes, which are commonly well preserved in the sedimentary record. The spacing and height of dune crests are proportional to flow depth, and therefore of keen interest to sedimentologists and geomorphologists to analyze the flow conditions.
during the moment of deposition. The transition to upper flow regime (Froude number >1) includes upstream migrating antidunes, which are rarely preserved in the sedimentary record (Bridge 2003). Between low and high flow stage, ripples and dunes provide sedimentologic evidence of recent events and the hydraulics of sedimentation. Phases of point bar deposition can often be distinguished by thin (≈1–3 mm) “clay drapes,” consisting of fine-grained cohesive wash load sediments deposited during slackwater conditions during the waning stage of the discharge hydrograph.

Channel bars

Channel bars develop as ripples and dunes are successively deposited along the flanks of shallow channel margins, resulting in lateral accretion. The process of lateral channel bar aggradation results in a distinctive fining-up trend in particle size, and is consistent with the concept of helical flow and a laterally migrating channel, which is a key sedimentary signature to identify past fluvial processes. The development of point bars within meandering rivers is among the most studied sedimentary features. Under an assumption of equilibrium, alternating channel bars and point bars adjust in association with lateral migration and maintain a consistent size and morphology. Bar deposits represent the largest percentage of overall accommodation space within a river valley. The recognition of floodplains infilled with channel bar deposits is an important indication of former channel activity, which may provide evidence for prior hydroclimatic conditions. Coarse-grained floodplain deposits are important natural resources which are heavily exploited by humans. Within dry regions, floodplain agriculture is often dependent upon alluvial aquifers for irrigation, particularly during extended low-flow conditions. Channel bar deposits are heavily exploited for aggregate, especially for construction of urban areas and associated transportation infrastructure.

Overbank sedimentary processes

Overbank sedimentation is an important mechanism for floodplain creation. Overbank deposits bury coarse-grained channel bar deposits, and in especially wide valleys cohesive fine-grained backswamp deposits constitute a considerable proportion of overall accommodation space. In addition to backswamps, other fluvial sedimentary features formed by overbank processes include natural levees, crevasses, and infilled channels (Bridge 2003).

Overbank sedimentation includes coarse sediment (sand/coarse silt) transported as bedload and fine-grained sediment transported in suspension, especially wash load (clay/fine silt). Although large floods receive a great deal of attention from the standpoint of hydrology, sedimentologically there is no relationship between flood magnitude and the rate of floodplain sedimentation. Individual flood deposits vary in thickness from millimeters to as much as half a meter, exhibiting pronounced lateral fining in thickness and particle size with distance from the active channel. Overbank flow velocity rapidly decreases upon exiting the channel. This causes sand and coarse silt to fall out of suspension and to be deposited along the channel margins, or transported across the floodplain as bedload, which may then be preserved as dunes and ripples as flow conditions transition from upper to lower flow regime (Bridge 2003). The distance that coarser sediments are transported across the floodplain surface as bedload is controlled by local floodplain hydraulics, which are influenced by the floodplain topography, the growth of vegetation, and floodplain hydrology. The overbank sedimentation process can be further considered...
FLUVIAL DEPOSITIONAL PROCESSES AND LANDFORMS

Based on whether advective or turbulent diffusive processes are dominant (Pizzuto 1987). Turbulent diffusive processes occur when floodplain basins (e.g., backswamps) are inundated prior to the river being overbank. This produces a steep suspended sediment concentration gradient across the floodplain, and induces abrupt sedimentation and high amounts of coarse sediment to be deposited along the channel margins. Advective sedimentation occurs when the river floods an otherwise dry floodplain, and enables coarser suspended sediment to be transported further from the channel margins. Along the distant margins of natural levees and in backswamps, slackwater deposition of clay is the dominant mode of sedimentation and results in thin clay laminations perhaps a millimeter in thickness (Aalto et al. 2003).

**Natural levees**

Natural levees represent the aggregate construction of numerous individual sedimentation events and require hundreds of years to form. Natural levees form along channel-floodplain margins and slope towards floodplain bottoms. Natural levees are located along meandering, braided, anastomosing, estuarine, crevasse, and distributary channels of alluvial fans and deltas. They are especially associated with flood-prone rivers dominated by suspended sediment transport, and coastal plain meandering rivers in particular. The size (height and width) of natural levees is related to flood stage variability and the available sediment for construction (Cazanacli and Smith 1996). The stratigraphy of natural levees at cutbanks includes individual strata commonly ranging from 1 cm to 10 cm in thickness, decreasing in thickness with distance from the active channel bank.

The slopes of natural levees between the channel and flood basins are related to sediment size, with coarse sediments associated with steep natural levees while fine-grained sediments are more associated with broad low sloping natural levees (Cazanacli and Smith 1996). Along a given river reach, the largest natural levees are located at the outside of meander bends, suggesting the reworking of floodplain deposits during large events, but also that crevasse mechanisms are important to natural levee development. In the case of laterally stable channels, minor natural levees may form on the inside of meander bends (burying point bar deposits). Along a river, the size of natural levees increases downstream of tributaries, especially with a large increase in discharge and sediment load. Natural levees are largest towards the lower reaches of large coastal plain alluvial valleys, but then decrease in size as they flow within deltaic settings and sediment load is exhausted (Hudson and Heitmuller 2003). From the standpoint of humans, natural levees are among the most important landforms within an active floodplain, as their higher surface and permeable soils permit settlement and, especially, agricultural activities with minimal flood risk.

**Crevasse splays**

A crevasse is formed during overbank conditions because of flood waters scouring a trough into the channel bank. A crevasse splay is a fan-shaped deposit with an axial scour channel extending at a tangent from the main river channel and natural levee crest, along the outside of a meander bend (Bridge 2003). Crevasses have a dendritic pattern because of smaller bifurcating channels, which are primarily depositional. Sediment is deposited mainly by sheet wash, and includes minor natural levees formed along the banks of the larger secondary channels (Cazanacli and Smith 1996). Crevasses are important mechanisms for distributing coarse sediments from high-energy settings to low-energy settings, which are
FLUVIAL DEPOSITIONAL PROCESSES AND LANDFORMS

otherwise dominated by fine-grained cohesive deposits. The formation of crevasses is especially associated with flood duration, although flood frequency is also important.

Crevase splays provide an indication of the flood activity of a river, and in remote locations should be seen as a good indicator of flood hazards. In regions with considerable economic and human welfare interests, river management agencies are tasked with managing crevasses. Restoration of floodplain ecosystem services along some lowland rivers includes the explicit breaching of river dikes during flood events to simulate crevasse processes. The ensuing hydrologic and topographic variability restores important elements of a river’s geodiversity, stimulating biogeographic complexity (Florsheim and Mount 2003). The recognition of distal floodplain sedimentation by crevasse-like processes is especially important in rivers with a high flood duration, such as humid tropical rivers which undergo an extended season of floodplain inundation. In such cases, crevasses continue to deliver large amounts of coarse silts to flood basins within a network of interconnected floodplain channels (e.g., Aalto et al. 2003).

Infilled channels

Many floodplains contain evidence of past river channel dynamics in the form of old channels, including abandoned river courses, meander cutoffs, and sloughs. Old channels become infilled with overbank deposits, primarily fine-grained cohesive deposits. Sedimentologically, these are often termed “clay plugs,” especially infilled oxbow lakes. This is somewhat of an oversimplification, however, as most infilled channels also contain lenses of coarser deposits (fine sand and coarse silt) which provide an indication of the frequency of prior flood events, especially large floods. Radiocarbon dating organic material obtained from field coring is a common approach to identifying the timing of past river channel changes, as well as paleoflood history.

Infilled channels exist within active meander belts, but are also preserved in abandoned meander belts. The timescales for channel infilling depend greatly on the main-stem sediment load and flood regime, as well as on the floodplain topography and distance of the old channel from the active channel. Over long periods (10^−3 years) infilled deposits become compacted and dense, and form clay plugs. Old clay plugs, however, have contemporary relevance because of being resistant to bank erosion, which thereby influences subsequent patterns of river meandering. The presence of hard clay plugs is especially common where an active channel is reworking an older channel belt, and should be considered in the development of channel engineering and bank protection management plans.

Backswamps

Beyond the limits of coarse sediment delivery, overbank flooding results in the transport and accumulation of thick fine-grained cohesive deposits. Backswamps are usually associated with larger river valleys which have sufficient accommodation space to store old deposits, and which are distal enough from the active channel that they are dominated by low-energy conditions and fine-grained sedimentation for extended periods (10^−3 years). Sedimentation is dominated by the deposition of wash load sediments (fine silt/clay) within a slackwater environment. The presence of “stringers” of coarser sediments (silt/fine sand) transported by prior large flood events is occasionally noted in deep floodplain corings. The thickness of backswamp deposits varies mainly according to the size (depth and width) of the river valley, but commonly exceeds 10 m (Bridge 2003).
Alluvial fans

Alluvial fans are formed where fluvial sediments are deposited as a fan-shaped landform. The essence of the feature is that multiple distributary channels extend outwards in a dendritic pattern from the fan apex, which is located where the longitudinal profile exhibits an abrupt reduction in slope. The channels deposit sediment at widely different rates because of being in different stages of formation and function. Two important controls on alluvial fan formation include an abrupt reduction in stream gradient and an increase in the zone of deposition. These boundary conditions occur where a narrow river valley enters a larger valley along bluff lines or fault-controlled mountain flanks. Several additional drainage basin controls on the formation and morphology of alluvial fans include (i) sediment load, (ii) hydrologic regime, (iii) gradient, and (iv) drainage area. Alluvial fans are located across all climatic regions, although a common misconception is that they are mainly associated with dry climates. This is likely because the morphology and vegetation pattern of alluvial fans is more discernible and distinctive within dry environments. The groundwater within coarse alluvial fan deposits often represents an important aquifer to small communities, and especially for irrigation of associated agriculture.

Sedimentary deposits associated with alluvial fan formation include a combination of channel and overbank deposits, with the latter being dominant under the influence of extensive sheet flow conditions during flood events, as well as debris flow deposits. As such, the coarse particle size of alluvial fans exhibits a distinctive fining trend with distance away from the fan apex, and sediment sorting also increases. The timing of alluvial fan sedimentation should be sensitive to exogenic controls, although linking their activity to climate change can be problematic in tectonically active regions because of base-level changes (Harvey 2002). Where alluvial fans are closely spaced they may coalesce, resulting in a larger depositional landform termed a “bajada.”

Fluvial terraces

Fluvial terraces are relict fluvial landforms situated above the active floodplain surface. Their morphology consists of distinctive elongated flat surfaces (treads) defined by erosional bluffs (risers). Terraces are formed when channel incision hydrologically disconnects the floodplain surface from active fluvial processes. Relict fluvial deposits then undergo reworking and modification by other types of surficial processes, including eolian, runoff, weathering, and pedogenic processes. Many river valleys are defined by specific terrace sequences, with progressively higher (and older) terraces displaying evidence of a greater degree of reworking and pedogenesis. In relation to humans, fluvial terraces are among Earth’s most productive lands. River terraces have a variety of qualities that lend themselves to human exploitation, including (i) freedom from flood risk, (ii) deep fertile soils, (iii) level surfaces, and (iv) proximity to water resources, making them invaluable for occupation and agricultural activities.

The study of river terraces has a long tradition within physical geography and geomorphology, and the mapping of river valleys into active Holocene surface bounded by fleets of relict Pleistocene terraces (i.e., $t_0$, $t_1$, $t_2$, $t_3$, etc.) is among the earliest traditions in landform mapping. Unfortunately, this rather simplistic perspective does not recognize geomorphic variability during the Holocene and the formation of low Holocene terraces, or effective terraces created during the Anthropocene. Earlier – and
persistent – ideas concerning the origin of terraces attempted to “fit” each specific fluvial terrace to a glacial cycle, an idea that originated in central Europe with the study of Rhine River terraces in relation to glaciation in the Alps. This conceptual framework was also applied in North America, especially in relation to continental ice sheets, sea level fluctuation, and Mississippi River terraces. Modern ideas concerning the origin of river terraces recognize a greater range of climatic variability across the oxygen isotope record, changes in sediment load, and especially tectonic controls.

**Fluvial deltas**

Deltas form at the terminus of a watershed where a river discharges into a standing body of water, and result in coastal environments with considerable geomorphic and biological diversity. The classification of deltas is based upon their morphology, which primarily relates to their planform geometry. The shape of a delta is determined by the style of deltaic sedimentation, with three controls on sedimentation being tides, waves, and fluvial regime, with the latter control being defined by river discharge and sediment load. Fluvial deltas are located within lakes and marine settings having low wave energy or tidal range relative to fluvial inputs. Fluvial deltas are characterized by an elongated channel, which is formed as a prograding river mouth advances into a receiving basin. Fluvial delta evolution occurs in association with river mouth progradation, and especially includes river mouth bifurcation and the formation of distributary branches and subdeltas created by crevasse events during flooding, which collectively form a delta lobe.

The process of delta lobe formation is initiated upon the debouching of sediment at a river mouth, which abruptly results in sediment sorting and deposition. Sediment sorting at the river mouth produces deltaic sedimentary facies comprising three distinctive depositional units: (i) bottom-set – horizontal bedding consisting of soft fine-grained cohesive deposits derived from the suspended load; (ii) fore-set – inclined bedding consisting of coarse-grained deposits derived from bedload; and (iii) top-set – horizontal bedding consisting of medium- to fine-grained deposits derived from overbank processes with organic material representative of wetland and marsh environments (Bridge 2003). Continued river mouth progradation results in dense coarse-grained fore-set bedding burying and compacting soft bottom-set bedding, producing a distinctive coarsening-up vertical sedimentary sequence which defines deltaic facies.

**Floodplain deposits and pedogenesis**

Understanding the development of floodplain soils in relation to sedimentary parent material requires consideration of sediment structure, floodplain hydrology, weathering, and bioturbation. The length of time for which sedimentary structure is preserved depends not only on primary sedimentology, particle size, and bedding, but also on pedogenic regime. Floodplains are often considered a single depositional environment, but pedogenic characteristics vary spatially, away from the channel. A soil catena, for example, may occur along the natural levee profile, which represents a systemic topographic transition from higher well-drained channel bank margin to lower and poorly drained backswamps. Sandy sediment structure is usually preserved longer than fine-grained muddy bedding, with individual sedimentary laminae possibly
destroyed after a year or so because of bio-
turbation by plants and burrowing organisms, 
especially in the humid tropics. Soils developing 
on thicker, sandier, and well-drained surfaces 
of channel bars and natural levee crests fre-
quently have oxidized and leached horizons with 
abundant manganese nodules and iron oxide 
stains. These soils may have medium to stiff, 
mottled gray, tan, and brown silty clay, sandy 
clay, or silty sand, and overlay darker fine-grained 
soils with considerable organic matter, devel-
oped in floodplain bottoms. The most common 
soil profile along a channel cutbank is an A–C 
horizon, which occurs because of insufficient 
time for eluvial (E) or illuvial (B) horizons to 
develop. High rates of overbank sedimentation 
result in floodplain topsoil typically lacking dark 
organic matter. With increasing distance from 
the active channel, however, the maturity of 
soils increases, as does bioturbation. Soils within 
poorly drained flood basins typically lack strat-
ification and have better developed illuvial (Bt) 
horizons. Backswamp soils are frequently gleyed 
(iron reduction) because of a persistent high 
water table. In the case of a high water table, 
the plant root zone is near the surface, whereas 
in floodplains with high seasonal fluctuation 
in the water table the root zone extends much 
deeper, which reduces the preservation of thinly 
laminated sedimentary structure.

Conclusions

Fluvial deposits are associated with a diverse 
range of surficial environments and have been 
important to humans for millennia for settle-
ment, agriculture, and natural resources. Because 
of their sensitivity to climatic and anthropogenic 
influences, in an era of global environmental 
change it remains essential to understand the 
fundamental controls influencing Earth’s fluvial 
sedimentary processes.

SEE ALSO: Fluvial erosional processes and 
landforms; Geomorphic hazards; 
Geomorphological mapping and geospatial 
technology; Rivers and streams; Water resources 
and hydrological management

References

Aalto, R., L. Maurice-Bourgoin, T. Dunne, 
et al. 2003. “Episodic Sediment Accumu-
lation on Amazonian Flood Plains Influenced by 
El Niño/Southern Oscillation.” Nature 425: 
493–497.

Bridge, J.S. 2003. Rivers and Floodplains: Forms, Pro-

Cazanacli, D., and N.D. Smith. 1996. “A Study of 
Morphology and Texture of Natural Levees – 
Cumberland Marshes, Saskatchewan, Canada.” 

in Lowland Floodplain Sedimentation Processes: 
Pre-disturbance to Post-rehabilitation, Cosumnes 

Transport.” Proceedings of the National Academy of 
Sciences USA, 103(46): 17170–17173.

in the Dissection of Alluvial Fans: Case Studies 
from Southeast Spain and Nevada.” Geomorphology, 

Watershed-Scale Controls on the Spatial Variability 
of Natural Levee Deposits in a Large Fine-Grained 
Floodplain: Lower Pánuco Basin, Mexico.” Geom-

Classification of Floodplains.” Geomorphology, 4(6): 
459–486.

Pizzuto, J.E. 1987. “Sediment Diffusion during Over-
FLUVIAL DEPOSITIONAL PROCESSES AND LANDFORMS

Further reading


Fluvial erosional processes and landforms

Ross H. Martin  
Texas State University, USA

Fluvial erosion can be found in varying forms and rates throughout a watershed. From raindrop-impact erosion and upland denudation to undercutting and mass wasting, fluvial processes are responsible for moving large amounts of sediment across landscapes. Fluxes of sediment create various physical signatures on the landscape. River planform and morphology, and other landforms associated with fluvial processes are determined by the characteristics of fluid flow and the properties of the erodible material (Knighton 1998). The interaction of fluid flow with the Earth’s surface leads to an understanding of how sediments are initially entrained. Erodible material varies greatly, from fine-grained cohesive sediments to bedrock to loosely consolidated alluvial material.

The scale of assessment of fluvial erosion spans from understanding the physics of how a single particle is entrained to mass wasting events along river banks to landscape evolution. Fluvial processes occur from the top of the watershed to the bottom, beginning as rain falls on a landscape. Rain-splash erosion occurs as rain falls on bare ground. In upland areas, overland flow can denude hillslopes. As surface water flows down a watershed it accumulates into streams, then rivers, creating potential for the erosion and transport of sediment.

For the purpose of general understanding, the watershed can be divided into three areas based on the dominant process – erosion, transport, and deposition (Schumm 1977). Erosion is dominant at the upstream most portion of the watershed, where steep slopes provide greater potential energy for sediment mobility. The middle zone is dominated by transport. In the transport zone, sediment deposition from upstream acts in quasi-equilibrium with erosion. The most downstream portion of a watershed can be characterized by deposition; here, the slope decreases and the stream’s ability to move sediment is reduced. In reality, geomorphic response is a system of complex responses depending on many variables at many scales.

Flow dynamics

The flow of water in open channels is primarily dependent on the interaction of two factors, the force of gravity acting in the downslope direction and friction (Knighton 1998). This relationship can be approximated by the Reynolds number (Re):

\[ Re = \frac{VR}{v} \]

where \( V \) is equal to velocity (m s\(^{-1}\)), \( R \) is the hydraulic radius of the flow (m), and \( v \) is the kinematic viscosity (m\(^2\) s\(^{-1}\)).

The velocity profile of flow in a stream shows that the velocity decreases as it approaches the boundary layer. In systems with a low Reynolds number the flow is laminar, which means that the velocity of each “layer” of water is constant and there is little mixing between the layers. In these cases, the amount of shear stress available to transport sediments is proportional to the velocity difference between the “layers” of flow.
(Petts and Foster 1985). On the other side of the spectrum, systems with a high Reynolds number are characterized by turbulent flow. Turbulent flow results in eddies and diffusion of water particles, which provides an additional viscosity factor to the system, termed eddy viscosity (Petts and Foster 1985). This results in greater shear stress available to transport and erode sediment. In reality, Reynolds numbers of typical stream channels are somewhere in the middle of the two extremes (laminar and turbulent flow).

Large rocks, large woody debris, weirs, dams or other features can create momentary or singular disruption to the flow and, thereby, can create irregular erosive patterns. A hydraulic drop in the stream’s path can create a transition from subcritical to supercritical flow. Most streams with gradually varied flow exhibit subcritical flow, where the gravitational force dominates. Supercritical flows occur where inertial forces dominate, move rapidly, can overshoot bends in the channel, and, therefore, have the potential to be highly erosive (Charlton 2007). The Froude number \( Fr \) is the ratio of the gravitational and inertial forces. Where the Froude number is \( Fr > 1 \) the flow is supercritical, where it is \( Fr < 1 \) it is subcritical, and if it is equal to one it is considered transitional:

\[
Fr = \frac{v}{\sqrt{gd}}
\]

where \( Fr \) is the Froude number, \( v \) is velocity \( (m s^{-1}) \), \( g \) is the gravitation constant \( (9.8 \text{ m s}^{-2}) \), and \( d \) is depth \( (m) \).

The surface of the channel provides resistance to flow, in varying degrees, depending on the roughness of the channel bed and walls. The flow exerts a shearing action on the channel boundary (bed and walls). Therefore, the flow in the channel is slower near the boundary and faster in the middle of the channel, where there is little resistance. It is typical to measure average water depth at 0.6 of the water depth in the center of the channel (Petts and Foster 1985). Several equations have been formulated to approximate the velocity of a flow. The velocity is dependent on channel roughness or resistance to flow, slope, and cross-sectional form. The Chezy equation, Manning’s equation, and the Darcy–Weisbach equation relate flow velocity to bed roughness. In general, roughness and flow resistance are greater where the ratio of sediment size to flow depth is greater.

### Thresholds for initial motion of sediments

Erosion of material by fluvial processes begins with the movement of sediment. Sediment resting on a stream bed or hill slope requires force to be coaxed to move. Work is done when a force moves an object over a given distance. Power is defined as the rate at which work is done. Bagnold (1966) suggested that stream power be used to evaluate the capacity of a flow to mobilize sediments:

\[
\Omega = \rho g Q S
\]

where \( \Omega \) = stream power \( (\text{wattsm}^{-1}) \), \( \rho \) = density of water \( (\text{kgm}^{-3}) \), \( g \) = gravity constant, \( Q \) = discharge \( (\text{m}^3 \text{s}^{-1}) \) and \( S \) = slope \( (\text{m/m}) \).

Stream power can also be expressed as specific stream power \( (\omega) \):

\[
\omega = \frac{\Omega}{w}
\]

where \( w \) = width. Specific stream power \( (\omega) \) can also be related to bed shear stress \( (t) \):

\[
\omega = \frac{t}{v}
\]

where \( t \) = bed shear stress and \( v \) = average flow velocity. Bed shear stress \( (t) \) is the amount of shearing force placed on the bed surface by the flow of the water. It can also be written as:
\[ t = \rho ghS \]

where \( \rho \) = density of water (kg m\(^{-3}\)), \( g \) = gravity constant, \( h \) = height or depth of water (m) and \( S \) = slope (m m\(^{-1}\)).

Sediments are mobilized when the shear stress induced by the flow of water overcomes the forces that are keeping the particles in place. Larger particles require more energy and, therefore, flows of a greater discharge to be mobilized. The amount of surface area that is exposed to the flow is proportional to the drag force (or shear force) the grain endures. The degree to which a particle protrudes into the flow plays a major role in how much force it takes to dislodge and mobilize that sediment. On a stream bed with similar sized sediments individual grains will be equally sheltered from flow; whereas larger grains lying on a fine-grained bed will protrude into the flow and shelter smaller grain sizes downstream.

Wiberg and Smith (1987) showed that a major component of initial motion is the pivot angle, which is how much a grain must rotate to roll over the next downstream grain. The degree of bed grain size homogeneity plays an important role. It takes more force to pivot a large grain from between other large grains, whereas a larger grain resting on a fine-grained bed will have a very small pivot angle.

Critical shear stresses for mobility are related to the force of gravity, the size and weight of the grain in relation to the shear stress on the bed, and the kinematic viscosity. Shields (1936) plotted critical shear stresses against Reynolds number. This relationship is known as the Shields function or parameter. Once rearranged, this function becomes a dimensionless critical shear stress that is equated to the ratio of shear stress exerted on the bed to the submerged weight of bed surface particles:

\[ \theta_{cr} = \frac{t_{cr}}{g(\rho_s - \rho)D} \]

where \( \theta_{cr} \) is the Shields parameter (dimensionless shear stress), \( t_{cr} \) is the critical bed shear stress, \( g \) is the gravitational constant, \( \rho_s \) and \( \rho \) are the densities of sediment and water, and \( D \) is grain size.

The Shields parameter is used in many sediment transport equations. A commonly cited threshold value of the Shields stress is 0.03 (Buffington and Montgomery 1997) and Bagnold (1980) used a critical Shields value of 0.04. Many studies have related critical shear stress with relative grain size. In these power relationships, the range of the coefficients and exponents is determined by bed sorting, methodology used, and other factors inherent in natural streams or flumes.

**Sediment sources and erosion**

Sediment sources for alluvial streams include predominantly bed incision and bank erosion. The makeup of the substrate that underlies the stream bed will, to a large extent, determine the sediment size distribution of grains on a stream bed. Low to moderate flows may only entrain some of the sediment that is already resting on the stream bed; however, a larger flow, which can entrain most or all of the sediment on the stream bed, may erode or incise the underlying substrate. Bank erosion is complicated by root systems that hold sediment together. Thus, riparian vegetation is an important factor in fluvial erosion. Bank failures induced by direct action of the water include mass wasting and weakening of bank material. Mass wasting on banks includes slumping, rotational failures, and undercutting failures, which are the result of particle entrainment from the bank. Processes such as frost action and ice wedging can weaken the bank, making it more susceptible to failure and erosion. Slaking occurs when dry sediments
are rapidly inundated and air gets trapped in the sediment, which results in buoyancy forces dislodging sediments.

Bedrock channels are eroded by quarrying, abrasion and corrosion. Quarrying includes the breaking apart and movement of blocks or aggregates of bedrock. Blocks can be broken off by wedging, where fine grain particles enter a crevasse during a high flow and over time the crevasse enlarges, similar to ice wedging. Cavitation occurs only during high flows, when shock waves are generated by high flows with large pressure difference (Knighton 1998). Abrasion occurs when fine-grained sediments suspended in the flow scour, scratch, and abrade the bedrock surface. Abrasion becomes concentrated in weak areas, and acts in a positive feedback loop. Once abrasion scours a weak area a pothole or pit can form, which results in an eddy within the pothole. Sediments can become trapped, which then increases the rate of abrasive scouring (Charlton 2007). Corrosion of bedrock material is the result of dissolution of soluble bedrock, commonly along fracture planes.

Transport

The transport of sediment can be divided into three categories: bedload, suspended load, and dissolved load. Bedload is defined as the sediment that moves by rolling, sliding or saltation (bouncing) along the bed during transport. It generally contains the coarsest size fraction of mobile particles. This portion of sediment transport is primarily dependent on flow competency, and the largest particles are likely just beyond their thresholds for mobility. The bedload material can also protect the underlying substrate from erosion. Flows can winnow smaller grain size fractions, leaving only coarser gravels behind.

This condition, known as armoring, results in a protective layer that can withstand moderately high flows. If the armor layer is entrained, the river will experience a sudden and dramatic flux of sediment, as all of the grain sizes, including the smaller once protected grains, mobilize at once. Estimates of bedload transport rates are varied and complicated. Armoring is one of many factors that make empirically estimating bedload difficult. Log jams and large woody debris can also interfere with flow regimes. Cohesive sediments make estimating bed load difficult, too, as they tend to erode as aggregates instead of as individual particles. Bedload transport is primarily dependent on flow competency, so there are many transport equations that relate volume of bedload transport to discharge.

Suspended load comprises the sand and silt-sized sediments that are suspended in the flow, which will settle as soon as the discharge decreases. The size, shape, and density of a particle and the fluid density of a flow will determine the fall velocity for that particle (Charlton 2007). Slow fall velocities (smaller grains) and a turbulent flow result in the suspension of sediment. Since this sediment remains in suspension, it is able to be transported over long distances. The amount of sediment transported as suspended load is often determined by the availability of sediment in the watershed and is sourced from hillslope erosion, bank erosion, and, to a lesser extent, the stream bed. Since suspended load discharge varies with sediment availability, it often exhibits hysteresis, in which there are different quantities of sediment transported for a similar discharge. Because of the variability of suspended load it is difficult to estimate. Sediment rating curves that relate suspended load to discharge can be created but, as suspended load is dependent on vegetation cover, soil properties, and topography, among other factors, those rating curves are typically calibrated for a watershed or region. Often,
the suspended load comprises the majority of the total sediment yield from a basin. Very fine grained clays and silts, and biogenic sediments are included in suspended load. Turbidity, a common water quality parameter, is a measure of water cloudiness, and suspended sediment by proxy.

Dissolved load includes solutes from chemical weathering, atmospheric inputs, and anthropogenic sources (Knighton 1998). Dissolved load transport is not confined to surface water flow and is not a contributor to the mechanics of river adjustment and migration. However, chemical denudation of rock by surface water is a source of dissolved load and an appreciable factor of geomorphic change in some environments, such as karst. Chemical denudation is complex and dependent on water chemistry, among other factors (Petts and Foster 1985).

**Bedforms**

The flow dynamics and availability of certain sizes of sediment will dictate bed forms. In homogenous sand channels, as the Froude number and velocity vary the channel bed will exhibit different formations on the bed surface. At lower velocities and Froude number, ripple patterns will be dominant on the bed surface. As both the Froude number and velocity increase, dunes will begin to form and boils may appear in the water surface. Both dunes and ripples form when the upstream face is eroded and sediment is deposited on the downstream slope face. As the Froude number increases to one, the bed surface will flatten, with no ripples or dunes, as the flow velocity is too high to allow sediments to be deposited on the downstream slope face. Further increasing the Froude number and supercritical flow will result in standing waves, and the formation of anti-dunes on the bed channel.

Channel beds with mixed sediment sizes will exhibit sediment sorting. Pebble clusters may form when a large particle extends into the flow, and because it locally interrupts the flow, it initiates deposition on the upstream side of the large clast. Often pebble clusters can exhibit imbricated structures (Charlton 2007). Smaller sediments are deposited in the sheltered area downstream of the large clast. Shear stress variations on the bed surface sort the available grains. In pool-riffle sequences, coarser grains are found in the riffles and finer grains are found in the pools, where they are allowed to settle. In meandering channels, particles are sorted through the meander bend (Clayton 2010). Gravel bed channels tend to form armor layers where finer sediments are winnowed away, leaving a layer of coarse-grained sediments on the surface of the bed.

In some channels, especially after high flow events, erosion creates zones of high bed slope. These local increases in gradient are known as nickpoints. Nickpoints can occur in both bedrock and alluvial single thread channels, and migrate upstream because of shear stress increases associated with localized increase in stream gradient. Layers of more and less resistant rock or soils can also create nickpoints as the underlying less resistant rock erodes.

**Channel planform**

There is a continuum between a handful of dominant variables that will dictate channel planform. Those include sediment supply and grain size, slope, discharge, flood frequency, riparian vegetation, and geologic setting. The most obvious distinction is that channels are either single thread or braided. Schumm (1977) showed that slope and sediment load correspond to channel planform. A simplistic model is
that low sediment loads and low slopes result in straight channels, high sediment loads and high slopes result in braided channels, and meandering channels exist with intermediate slopes and sediment loads. A stream’s energy can be linked to channel planform and plays a major role in determining the size of sediment that can be mobilized. Potential and kinetic energy perform work via fluid flow, which results in turbulence and associated viscosity, a flow’s ability to overcome friction at the boundary layer, erosional shear, and capacity to transport sediment (Knighton 1998).

Higher-energy systems tend to be bedload dominant and can be characterized by braided planforms and large mid-channel bars. Braided planforms are associated with abundant bedload sediment, erodible banks, and highly variable discharge, conditions which are often present in proglacial environments. The abundance of bed material can be the result of aggradation, in which the river cannot transport the material. Braiding is an adjustment so that the river can transport the bed material. According to the stream power theory of sediment transport suggested by Bagnold (1977), bedload transport is more efficient in channels with higher width to depth ratios (Knighton 1998), such as braided channels. Single-thread streams in high slope environments may form cascades or steps and pools as a mechanism to dissipate energy. The energy is dissipated because of the increase in boundary roughness as the water flows over the large rocks or logs that make the steps and falls to the pool below.

With only moderate energy and mixed loads, the stream may begin to meander. Single thread channels can be described by their degree of sinuosity, which is equal to channel length divided by valley length. Flume studies have shown any degree of sinuosity can be accomplished by varying slope, discharge, and sediment supply (Knighton 1998). Meanders, functionally, provide a mechanism for the dissipation of a stream’s energy by reducing the stream’s slope. Meandering is the result of the feedback processes that cause cut banks to erode and point bars to aggrade. Cut banks are found on the outside of the meander bends. In these areas, as the name implies, lateral erosion of the bank material is occurring. This results in the meander bend migrating. Point bars are locations of deposition that act in quasi-equilibrium with the cut banks. High sinuosity and undeterred meanders result in meander bends being cut off as two meanders migrate together. This results in oxbow lakes and scarred landscapes. Streams with coarse grain availability and with varying degrees of sinuosity will form pool and riffle sequences. Riffles are typically coarser grained with shallower, faster moving water and pools are deeper with slower flow over finer grains.

Lower energy streams dominated by suspended load will vary from meandering channels with tortuous sinuosity to anabranching channels. Anabranching channels are associated with very low slopes and energy. The anabranching nature of these channels is caused by avulsions during peak flows. Avulsions are erosional episodes where high flows jump established banks and incise new channels. Wandering gravel channels and anastomosing channels are the primary types of anabranching channels documented. Anastomosing channels are characterized by very low gradients and very low rates of bedload transport. They are similar in appearance to braided channels, except that they are usually more stable and have clayey or silty banks. Wandering gravel channels are characterized by coarse sediment input but lower transport rates than braided channels (Charlton 2007). Wandering gravel channels tend to be more stable than braided channels because often the mid-channel bars are vegetated.
Hillslopes and interfluvies

Fluvial processes are not confined to channels. As rain falls, the impact of the rain drop can mobilize sediments, known as rain-splash erosion. Overland flow of water is also responsible for eroding and transporting sediments on slopes. Both of these processes act as sediment sources to streams.

Raindrops fall and accelerate to terminal velocity, which is dependent on drop size. The sediment particle size that can be mobilized by a falling rain drop is dependent on momentum. On a flat surface, raindrop impact results in a random distribution of sediment. As the slope increases, more of the mobilized sediment is displaced down the slope. Rain-splash erosion is complicated by vegetation canopy, which can intercept rain, reducing total volume, and can prevent rain drops from reaching terminal velocity (Petts and Foster 1985). Root systems and ground cover also play an important role in determining the amount of sediment available for mobilization.

Once the rain has fallen, it collects and flows over the surface, a process termed Hortonian overland flow. Hortonian flow can entrain sediments after critical shear stresses are overcome. Sediment is mobilized, which results in changes in the microtopography of the hillslope. Rill formation is dependent on the vegetation cover and soil or sediment type. Water flow concentrates in the depression, where erosion is accentuated, resulting in rills. Gullies are similar to rills but are larger and tend to be permanent channels for ephemeral flow. However, gullies are distinctly different from stream channels, with lower width:depth ratios. Gullies, like rills, are highly dependent on the lack of vegetation cover and tend to form in arid environments. Gullies enlarge and deepen as headcuts or nickpoints concentrate the erosion and migrate upslope.

Arroyos are gully-like features that form in arid environments with abundant sediment.

SEE ALSO: Fluvial depositional processes and landforms; Geomorphic thresholds; Hydrologic flow models; Rivers and streams; Soil erosion and conservation; Watersheds

References


FLUVIAL EROSIONAL PROCESSES AND LANDFORMS


Further reading


Focus groups

Peter Hopkins
Newcastle University, UK

Focus groups are now an established qualitative research method in the discipline of geography and are regularly used within many subfields of human geography including environmental geography, social and cultural geographies, population studies, and children’s geographies (Burgess 1996; Hyams 2004; Skop 2006). As a method, focus groups are about producing data based on a one-off group discussion on a specific topic that is provided by the researcher, who often takes on the role of the moderator (Hennink 2014; Morgan 1998). The main idea behind focus groups is, therefore, to generate group interaction and contestation over ideas and opinions that cannot readily be found in individual interviews (Krueger 1994; 1998). Key factors to consider about focus groups include group interaction, the topic, the role of the moderator, and the size of the group (Hopkins 2007).

Focus groups can be used in a self-contained manner as the main source of data. They can also supplement another method such as interviews, or be used in a multimethod study alongside a host of other research methods. Focus groups can work usefully alongside interviews, ethnography, questionnaire surveys, as well as participatory action research approaches. It can also be useful to consider the potential benefits of using focus groups at different stages of a research project. For example, focus groups may be used at the start of a research project to orient the researcher to a new field, to generate ideas and themes for further exploration later in the study, or to generate hypotheses that can be systematically tested through further research. Focus groups may also be used later on in a study to discuss research outcomes in groups or to verify the findings of the researcher.

The increasing use of focus groups is largely due to the many benefits they can bring to research (Barbour and Kitzinger 1999; Goss 1996). As well as being helpful for those who want to orient themselves to a new field, focus groups can be particularly useful for providing insights into how groups develop, challenge issues, and focus their arguments around particular points. This can often lead to the generation of new ideas, the evolution of a creative process, and a dynamic research agenda that may encourage researchers to be reflexive about their perspectives and approach. Indeed, the group dynamic may be pertinent to the research content. Focus groups can also be more inclusive, offering participants safety in numbers and perhaps even recruiting those who are “hard to reach”; some would therefore see focus groups as being more democratic and empowering to the participants and less threatening and power-laden compared to individual interviews. Focus groups can arguably also be cathartic experiences for participants, allowing and enabling them to raise and debate issues that they rarely have the opportunity to discuss (Johnson 1996), thus contributing toward social change and transformation.

That being said, focus groups are not without their weaknesses and potential pitfalls. Those who participate in focus groups may give “public” accounts as they present themselves to other group members. Certain types of participants
FOCUS GROUPS

may also dominate the discussion and silence those who are less verbal or confident, or, worse still, there may be saboteurs who are determined to refocus the discussion around their ideas and agenda, thereby overlooking the wishes of the moderator. Furthermore, focus groups can be hard to recruit beyond the usual suspects. One of the most significant criticisms levied at focus groups is that they are difficult to run and, therefore, rely very heavily on skilled moderators (Krueger 1998).

The development of focus groups as an established qualitative research method is somewhat contested, however. Focus groups are often linked with interviews (and, in particular, group interviews) and most strongly associated with market research where a group of people are consulted about a product, event, or issue. It is important to differentiate between focus groups – which are one-off group meetings – with in-depth groups that meet on more than one occasion (and often on many occasions over a period of time) and have their origins in training, educational, and therapeutic contexts. Generally, the intentions of both methods are very different. Focus groups aim to gain one-off insights into a particular issue in a group; although this may also be one of the aims of an in-depth group, it is likely that in-depth groups will have additional primary aims such as training people in a particular skill or technique, or acting as a therapeutic context in which people can improve their wellbeing over a period of time. As such, comparing focus groups to university tutorial or seminar groups is unhelpful, as often these are not one-off and the overall aim is to improve students’ knowledge and understanding over a period of time rather than having only a one-off group discussion.

Given the significance of the role of the person who moderates a focus group, much of the literature often focuses on the skills required to be a good moderator (e.g., Krueger 1998). Issues here include how to gain the consent of and other information from each person, as well as how to facilitate the introduction of group members to each other and how to generate the initial discussion. With some focus groups, it may not be clear how many people will actually turn up, so the moderator may need to be prepared for different sizes of groups and the different skills required. During the focus group discussion, the moderator also needs to balance how often they intervene to steer the discussion and how they manage group members who are either very quiet or who dominate the group. Furthermore, unless the moderator is working with another researcher who observes the discussion and takes notes, they will need to try to keep track of who says what to facilitate the process of transcribing the focus group and correctly allocating statements to participants.

There are a number of methodological considerations to take into account when employing focus groups as a method of data collection. These can include, but are by no means limited to, the number of participants in the group; the composition of group members; location, timing, and context; as well as other factors including the sensitivity of the topic, the age of the participants, and the positionalities of the researcher and the participants. The main factors that often determine the success of a focus group are the focus of the researcher/moderator and the interaction of the group. These additional considerations are important for those aiming to conduct rigorous and ethical research.

Most definitions of focus groups include some reference to the number of people needed for the group and these tend to range from as low as 4 to as high as 12. However, the most important aspect of the focus group method is group interaction in discussing a particular topic rather than the number of participants.
Depending on the particular topic and the social group participating, the optimum number of focus group members will vary. For some topics, it will be very challenging to moderate a focus group discussion with 12 people, whereas for others it may be challenging to generate a discussion with only four people. A useful example that challenges the focus on numbers is research about pregnant women’s experiences of public spaces in New Zealand, in which Longhurst (1996) conducted three focus groups, to which only two women turned up. These ended up being a useful form of data collection and offered the researcher insights into the research topic despite their having been initially categorized as “failed” focus groups. In a second example, Hopkins (2007) found that, in working with Muslim young men, larger groups were more difficult to moderate compared to those with only three or four participants, because the smaller groups provided the opportunity for each participant to voice their opinion, which was not always offered by the larger groups.

In terms of group composition, some of the focus group literature claims that focus group participants should have shared characteristics in terms of the criteria for selection (e.g., gender, age, ethnicity, location) but should not know each other (Tonkiss 2004). This is problematic as it is based on the principles of survey logic and makes assumptions about who does and does not know each other. Holbrook and Jackson (1996) observed that focus groups with people who already know each other and group discussions with people who have never met before have different strengths and weaknesses. Focus groups where participants are unknown to each other may require more work from the moderator to facilitate introductions and generate initial discussion compared to focus groups with people who already know each other. On the other hand, focus groups with those unknown to each other may result in people being more considered about their opinions, whereas focus groups with people who know each other well may be informed by subtle power struggles that may shape the discussion in particular ways and that remain undisclosed to the moderator. Researchers using focus groups as a method of data collection should therefore consider these important methodological issues when recruiting and moderating focus groups.

Location, timing, and context are also important considerations when planning and conducting focus groups and when analyzing the data that emerge from them (Hennink 2014). The physical location of the focus group discussion often requires careful consideration to ensure that participants feel comfortable and able to discuss the topic. Related to this, the timing of the focus group needs to be thought through. For example, office workers who regularly work from 9 to 5 are unlikely to turn up at an afternoon focus group, and some older people may not want to attend a focus group that takes place late in the evening when it is dark. Finally, there may be significant contextual matters to consider such as political events or local activities that might shape the focus group discussion. This may include issues such as local or national elections, protests, strikes, or community events that may limit focus group participation or result in the focus of the discussion being shaped by specific events that are happening at the time (Hopkins 2007).

There are a range of additional issues that human geographers should consider when conducting focus groups, many of which are related to the points already raised. One factor to consider is the sensitivity of the topic. While some topics may be so sensitive that it would be ethically inappropriate to discuss them in a group setting, others may usefully be explored in a focus group; sensitivity should be demonstrated
FOCUS GROUPS

in the choice of the size and composition of the group, as well as the location and context of the group discussion. For example, the small size of the group may be one of the reasons why the focus groups that Longhurst (1996) conducted with pregnant women were so successful. This is a very personal topic and some of the participants may not have been as comfortable sharing their experiences in a large group. Related to this, the age of the participants and the positionalities of the researcher in relation to the research participants may also be important considerations (Hopkins 2007).

SEE ALSO: Interviews; Qualitative data

References


Food security

Colin Sage
University College Cork, Ireland

The origins of the term “food security” are often traced to the 1974 World Food Conference. This was called in the wake of a series of humanitarian disasters affecting the Horn of Africa, South Asia, and the Sahel (1972–1973); the Soviet Union’s secret emergency purchase of US grain following its own harvest failures; and the escalating global increase in cereal prices between 1971 and 1974. The conference gave rise to the first widely adopted definition of food security as the “availability at all times of adequate world food supplies of basic food stuffs … to sustain a steady expansion of food consumption … and to offset fluctuations in production and prices.” Yet, despite an apparent consensus by scholars as to this origin, it has been argued that food security can be traced back to Europe and World War I. Edelman (2014) asserts that the term was strongly linked to concern for national security and to the capacity of each country to produce its own food such that it would reduce its political and military vulnerability. This framing resonates strongly with contemporary discussions of the related concept of food sovereignty, to which this entry will return. It is clear that food security does have a longer history than many acknowledge and, indeed, featured in statements made during the 1943 Hot Springs Conference of allied governments which was regarded as the first step in founding the Food and Agriculture Organization of the United Nations (FAO).

The 1974 definition was naturally a consequence of its time and is clearly framed by the imperative of production and supply that strongly resonated with the prevailing agricultural modernization discourses of the era, best exemplified by the Green Revolution model of technology transfer. The development and dissemination of high-yielding varieties of wheat and rice as part of a “package” of technologies that sought to increase scale, levels of mechanization, and market orientation while successfully increasing “output” were to have enormous unanticipated consequences for rural societies. While aggregate output rose in those regions endowed with the necessary resources (soils, water, and infrastructure) required for the technology (principally Southeast Asia, the Punjab, parts of Latin America), other regions (especially in sub-Saharan Africa) were marked by deepening hunger, malnutrition, and disease. Understanding of the complexities of food insecurity and vulnerability was aided by local and regional studies conducted by geographers and other rural development specialists.

However, probably the single most important contribution to shifting the prevailing view of food insecurity was the publication of Amartya Sen’s Poverty and Famines (Sen 1981). Here Sen forcefully demonstrated that hunger and starvation are not an inevitable consequence of a decline in the availability of food, but rather reflect the circumstances of people not being able to secure access to food. For Sen, this can be explained by understanding people’s entitlement relations, which might derive from endowments in land, other assets, or their labor power and which provide the basis for survival under “normal” conditions.
Changing circumstances, such as the occurrence of drought, may reconfigure entitlements as the abandonment of agricultural production in the absence of rainfall, for example, renders rural wage laborers vulnerable to unemployment and hunger. Livestock owners face a similar predicament as, in the absence of adequate grazing, animals weaken and their value drops, while grain prices invariably soar. Entitlement theory consequently contributed to a shift in food security thinking away from the macroscale and toward the individual and household level.

Food security was, moreover, once synonymous with the supply of high-calorie staples such as cereals and tubers to resolve problems of protein-energy malnutrition. The logic of the Green Revolution was to deliver increased calories above all. But calorie intake says little about nutritional status, and by the late 1980s health and nutrition research better understood that minerals, vitamins, and trace elements – including iron, iodine, folate, and vitamins A, B complex, and C – play a critical role in human wellbeing. Globally, today, “over 165 million children are stunted and two billion people lack vitamins and minerals for good health” (Khoury et al. 2014). Indeed it has been estimated that 2 billion people are anemic – over 30% of the world’s population – making it the most widespread nutritional disorder in the world.

By the time of the World Food Summit in 1996 the definition of food security had evolved to reflect issues of access and the variety of analytical scales from the household to the global level, as well as social and cultural influences over food preferences. Thus food security is said to be achieved when “all people, at all times, have physical and economic access to sufficient, safe, and nutritious food to meet their dietary needs and food preferences for an active and healthy life.”

Clearly, then, the notion of food security is a storied term for which it has been estimated there are more than 200 definitions and 450 indicators. In part, the multiplicity of meanings attached to food security can be attributed to the boundary nature of the concept: it has crossed many academic disciplines besides geography, and been a focus for applied work in the agricultural, environmental, and health sciences. Yet, if there is one pillar that has anchored food security discourse over the past half-century or more, it is the central role of technology-led production as the key to food availability.

**Food security and productivism**

From the second half of the twentieth century, astonishing increases in agricultural productivity were achieved as a consequence of a series of interconnected technological developments, which Mazoyer and Roudart (2006) conceptualize as:

- motorization: the development of increasingly powerful tractors and engines fueled by oil, substituting capital for labor;
- mechanization: increasingly complex and effective machinery streamlining production throughout the food chain; and
- chemicalization: the deployment of growing volumes of synthetic fertilizers and pesticides, as well as associated developments in plant and animal genetics.

This model, which has come to be labeled *productivism* for its singular focus on output, has enabled the world today to produce sufficient food by volume of crops to meet the needs of the world’s growing population. Of course, there are an estimated 870 million malnourished people whose present state of food insecurity is a consequence not so much of output as of a failure
of access, compounded by locally contingent variables such as affordability or conflict. Yet productivism, together with the development and intensification of global trade, has proven the axiomatic principle promoted by international organizations and national governments by which food security would be achieved.

The development of the global food system since the 1980s has witnessed the emergence of new trading patterns that have had huge repercussions for some of the poorest and least food-secure countries. For example, the rise of the new agricultural exporting countries such as Kenya has led to their engagement with nontraditional, high-value fruit, flower, and vegetable contract production for major supermarket chains in the North. Meanwhile other countries have witnessed extensive cereal and oil seed cultivation in response to the rising global demand for livestock feed and biofuels. Such large-scale commercial developments have often served to marginalize millions of small farmers through dispossession of land, depletion of water resources, or loss of local market share. At the same time, closer economic integration leads to increased exposure to international market forces for hundreds of millions more.

During the first decade of the twenty-first century, cereal prices rose to their highest levels in real terms since the early 1970s, reaching a peak in 2008, which was followed by significant price volatility. Such dramatic increases triggered violent protests and widespread civil unrest across the world as household budgets could no longer meet basic dietary needs and the number of food-insecure people reached 1 billion. Unsurprisingly, attention turned to identifying the drivers of these price rises, and their continuing volatility. Among the drivers for which analytical models demonstrated a link with food prices were: financial speculation on global commodity markets; extreme weather events and potential climate perturbations; declining grain stocks, possibly linked to falling public investment in agriculture over two decades; and rising consumer demand, as a consequence of demographic increase and also of changing dietary composition. However, the clearest correlation with the behavior of food commodity prices was that of hydrocarbon energy prices which rose steadily through the decade, reaching a peak in July 2008 before falling back. For many analysts it is higher oil prices and the associated expansion of the biofuels sector that has been most responsible for undermining food security worldwide.

A consequence of tightening oil markets as demand outstrips supply capacity has resulted in a huge boom for biofuel crops capable of being converted to either ethanol or biodiesel. By 2010, the combined amount of maize and sugarcane from the United States and Brazil respectively used as feedstock for ethanol amounted to 460 million tonnes, or 6% of global crop production by mass (Cassidy et al. 2013). It has been estimated that the volume of maize currently diverted to ethanol distillation in the United States would be sufficient to feed 400 million people for one year. The setting of mandatory targets for utilization of renewable energy sources in the transport sector by the European Union, the United States, and other countries has driven the expansion of maize and sugarcane, as well as palm and other vegetable oils, for conversion to biodiesel, across arable land. The essential fungibility of agri-commodities like these suggests that the food needs of the poor are subordinate to the energy requirements of the more powerful. This is best illustrated by the appropriation of land currently underway throughout the poorest countries of the Global South as powerful states and corporations engage in deals to acquire land and water resources. It has been estimated that up to 227 million ha – an area equivalent to the
total arable land of Western Europe – has been identified as suitable for occupation by foreign investors. Two-thirds of this land grabbing is in Africa, and 60% is destined for biofuel crops.

The logic of productivism, then, is to raise output and trust in the power of markets to enable people to get access to food. Yet the multiplicity of challenges – such as those regarded as drivers of the food price spikes of 2008 and beyond – demonstrate the increasing vulnerability of the food system. Indeed, there are particular anxieties around the impacts of climate change, freshwater depletion, impairment of ecological services, and the consequences of tightening energy markets that are leading many to believe that we need to move away from the productivist model. While policy analysts may argue over the rank ordering of drivers, it is clear to many that the global food system has become increasingly sensitive to a variety of short-term episodic shocks and lacks the resilience to cope effectively with such events. The task now facing us, therefore, is to optimize production across a far more complex landscape involving environmental sustainability, social justice, and nutritional security.

**Sustainability, sovereignty, security**

The first decade of the twenty-first century has consequently witnessed a reinvigorated global debate about food security, such that it has been said to constitute one of the key master frames of public policy. Rising and volatile food prices and associated political instability, together with projections over future needs, have raised the stakes in formulating legitimate solutions to food security. In particular, framing the problem as one of pressing future demand emerged from the 2008 Rome Summit on World Food Security. Here, two key projections were presented that have since become central pillars of global food security discourse: (i) that food production needs to increase by 50% by 2030 to meet rising demand; and (ii) that it needs to double by 2050 to feed a world population of 9 billion (Maye and Kirwan 2013).

Focusing on these projections has been a particular feature of those who favor treating food security as a scientific problem requiring technical solutions and where agricultural biotechnology and genetically modified (GM) seeds are regarded as the most effective way to feed the poor in developing countries. Proponents of this approach most often focus on the low yields of indigenous farming systems and argue that modern technologies through a new Green/Gene Revolution could do much to achieve food security across the poorest countries. Critics, on the other hand, have argued that not only does this emphasis on untried, imported technologies smack of a neocolonial attitude that undervalues the knowledge and adaptability of small farmers, but it runs the risk of deepening vulnerability to distant input markets as well as conspicuously failing to appreciate changing local environmental circumstances. Perhaps the most comprehensive case for a change of direction from high-tech productivism was set out by the International Assessment of Agricultural Knowledge, Science and Technology for Development (IAASTD 2009), which argued that industrial agriculture alone will not resolve the problems of hunger and poverty, and that it is incapable of adapting to the environmental challenges that lie immediately ahead.

A more pragmatic approach emerged in the wake of the 2008 crisis from a UK Government Foresight exercise that served to advance the concept of sustainable intensification (SI). Essentially, SI means producing more food from the same amount of land but with less environmental impact, a notion that has led its critics...
to suggest that SI means we can have our cake and eat it too. However, SI is concerned with halting and even reversing the continued expansion of agricultural land, recognizing the “new fundamentals” associated with climate change, freshwater depletion, and other resource limits. Hence its emphasis rests on the intensification of existing agricultural land, although this should not be pursued without due regard for other multifunctional benefits such as carbon sequestration and biodiversity conservation. Nevertheless, it is a concept that remains largely focused on the two Rome Summit key projections and that promotes a largely technology-centered approach where bioscience, nanotechnology, and information and communication technology (ICT) are regarded as keys to future food security.

If the development of a reinvigorated global debate about food security has revealed a sharply fractured consensus, it has also enabled a strong counterhegemonic discourse to be heard. Emerging from a highly heterogeneous constituency – from peasant movements in the South to food activists in the North, together with some very eminent global policy experts and leaders – this alternative discourse encompasses many divergent aspirations but shares a rejection of the technology-centered, business-as-usual approach in favor of an overarching goal of making agri-food systems more resilient. Questions of social justice, the cultural value of food, public health, and sustainability also crowd in to this alternative conceptualization of food security in an environmentally uncertain and changing world.

Among the most important claims proposed by this broad movement is that of food sovereignty. This is a term that also has a more obscure origin than is generally acknowledged, but it is most commonly associated with the international peasant movement Via Campesina, and is rooted in a rights-based framework that insists on food being treated as a basic human right. Food sovereignty has been widely proclaimed and reaffirmed, and in 2012 the FAO’s Committee on World Food Security agreed to its adoption. Indeed, the ascendency of food sovereignty as a potential policy proposition has helped it to move from being considered an entirely oppositional term associated with an antiglobalization agenda to one that is now entangled with food security. By insisting on the relational and rights dimensions of food, the notion of food sovereignty has begun to establish not only the human right to eat but also the right to produce. In this regard it challenges the existing food system built on the disembedded production of tradable commodities and transformed by large-scale manufacturing into cheap, processed, global brands. For, as diets worldwide become more homogeneous, they also appear to be less healthful, with a higher consumption of processed and convenience foods rich in saturated fats, sodium, and sugar. It is these growing dietary practices (malconsumption) that appear to account for an estimated 2 billion people worldwide being considered overweight and obese and susceptible to a host of medical problems that are thought to cost the global economy up to $2 trillion.

Food security, then, has become an increasingly contested term and one that is challenged by more intractable and interconnected threats than “simply” global hunger. The interlocking of food and energy markets; climate change and freshwater depletion; and diet-related ill-health all serve to make the solution a great deal more complex than one of “doubling output” (Sage 2013). Reframing food security in relation to food sovereignty may serve to demonstrate the vital importance of policy measures at all levels – from local to global – that will be required if we are to find ways to enable everyone to eat sustainably, healthily, equitably, and securely into the future.
FOOD SECURITY

SEE ALSO: Agriculture; Environmental (in)security; Famine; Food security; Global environmental change: human dimensions; Security; Sovereignty; Water and climate change

References


Further reading


Fordism/post-Fordism

Tu Lan
University of New Hampshire, USA

Fordism was a model of economic development characterized by mass production and mass consumption in the United States and Western Europe between the 1930s and 1970s. Post-Fordism includes a number of alternative development models flourishing after the crisis of Fordism in the 1970s. Both concepts have been widely used by geographers, particularly of the Marxist tradition, to study changing labor relations and associated social formations in major capitalist economies.

The term “Fordism” was originally used by Antonio Gramsci (1971, 277–316) to describe the forms of production pioneered by Henry Ford’s automobile industry. Gramsci attributed the higher productivity and relatively more stable labor relations in the United States to Frederick Taylor’s reinvention of assembly lines, a rational demographic composition, a higher wage, and “a specific mode of living and of thinking and feeling life” (1971, 302).

On the basis of Gramsci’s work, the French regulation school developed a conceptual framework to analyze Fordism. It argues that capitalism has no permanent solution for its inherent contradictions of overproduction and falling rate of profit, and that any stable period of capitalist accumulation is therefore historically contingent on a regime of accumulation and a mode of regulation.

According to Alain Lipietz (1987) and Bob Jessop (1991, 136–137), Fordism has four major dimensions. First, Fordism was made possible by the labor process of Taylorism, in which unskilled labor specializing in simple, repetitive work on the assembly lines was separated from skilled and engineering labor at the managerial level. Taylorism, together with rapid mechanization, significantly raised the productivity of standardized commodities, such as Ford’s automobiles. It was the precondition for mass production.

Second, Fordism needs a regime of accumulation which concerns the allocation of social resources between different departments of production and between production and consumption. In order to maintain the rate of profit, increasing mechanization needs to be balanced out by the rising productivity of equal rate, while the rising productivity has to be absorbed by the rising purchasing power. Therefore in Fordism, mass production necessarily leads to mass consumption.

Third, the Fordist regime of accumulation requires a specific mode of regulation of wage relations. In particular, it must ensure a rising wage level for the working class to buy the commodities that are put out by mass production. Regulationists believe that Fordist wage relations are based on an institutional arrangement which is a result of a consensus within the ruling bloc of the individual states. Consequences of the consensus include the welfare state, nationalized trade unions, and a centralized banking system based on the gold standard of the US dollar.

Finally, Fordism also had widespread consequences in the broader society, such as a lifestyle of nuclear families who consume standardized goods, which Jessop called “a general pattern of social organization” (1991, 137).
Fordism went into a crisis signaled by the end of Bretton Woods system in 1971. The crisis was due to three main reasons. First, productivity failed to catch up with the rate of mechanization, leading to the rising organic composition of capital since the 1960s. Therefore, the overall rates of profit in major capitalist economies continually fell, making mass production untenable. Second, multinationals started to outsource manufacturing jobs to newly industrialized countries (NICs), leading to the stagnation of the wages of the working class in the United States, which made mass consumption unsustainable. Finally, short-term incidents such as the oil embargo by Arab countries in the 1960s and 1970s also severely halted the capitalist accumulation of major capitalist economies.

There have been multiple post-Fordisms, that is, alternative development models in response to the Fordist crisis (Amin 1994). In geography, by far the most influential has been flexible production or flexible specialization. This work is closely related to the literature on Italian industrial districts. In contrast to the mass production of vertically integrated corporations in the United States and Western Europe, industrial districts in Italy were renowned for their just-in-time production of small and medium-sized enterprises (SMEs). It is argued that changes on the demand side from standardized to diversified goods in Western consumer markets necessitated the transition from Fordism to flexible production (Piore and Sabel 1986).

For geographers, the transition to flexible production in each of the regions has always been specific. For example, Allen Scott (1988) and Michael Storper (1997) showed the importance of territorial embeddedness in just-in-time production. Focusing on the cultural dimension of Fordism, David Harvey (1991) argued that the flexibilization of labor led to the fragmentation of postmodern cultures in the West. Lipietz (1997) observed that, in certain capitalist economies such as Germany and Japan, the main response was a “negotiated involvement of the workers,” whereas in the United States the primary response was neo-Taylorism, in which flexible production was achieved by labor outsourcing and a rollback of social welfare. More recently, Jessop (2013) summarized two post-Fordist regimes of accumulation as the knowledge-based model in Europe and the finance-dominated model in the United States and the United Kingdom, both of which have been discredited after the global financial crisis in 2008.

Many features of post-Fordism, such as flexible specialization of labor and the cutting of the social welfare, have been articulated into the broader project of neoliberalism. However, for Marxist geographers, none of the alternative models is likely to permanently solve the inherent contradictions of capitalism.

**SEE ALSO:** Economic geography; Industrial districts; Labor geography; Labor market; Marxist geography; Neoliberalism; Regulation/deregulation

**References**


In a seemingly unassuming article published in *Science* in 1967, Mandelbrot (1967) posted a question that geographers and cartographers had been puzzling about: How long is the coast of Britain? The answer to this coastline measurement problem, as most geographers had known, would depend on the length of the divider used to traverse the zigzagging coastline on the map, and the total coastline length would be the product of the number of divider steps needed to traverse the line multiplied by the divider length. The intriguing part of this measurement problem is that, given a contorted coastline, the smaller the divider length used, the longer the coastline is. Further, the more contorted the coastline, the more drastic the difference in the resultant length measurements from using varying divider lengths. This phenomenon, also called the Steinhaus paradox, has long been observed by geographers and researchers alike. But Mandelbrot was able to capture and quantify the relationship between coastline length, its measurement scale (divider length), and the complexity of the line by using a new concept called fractals.

The coastline measurement problem was just one of the phenomena that prompted the development of fractals. Through observation of real-world objects, Mandelbrot found that most spatial patterns of nature, such as clouds, terrains, forests, and mountains, look so irregular and fragmented that classical geometry is not efficient in describing and quantifying their complexity. In classical (Euclidean) geometry, the objects are defined by an integer dimension, which is also called the topological dimension. A point has a topological dimension of zero, a line is one-dimensional, a plane has two dimensions, and a cube three dimensions. In Mandelbrot’s fractal geometry, objects have fractional dimensions called the fractal dimension \( D \). A point can have a fractal dimension ranging from zero to one, a line from one to two, and an area from two to three. The higher the fractal dimension of the object, the more complex is the geometry of the object. Real-world coastlines have fractal dimension values typically around 1.2–1.3, and topographic landscapes around 2.2–2.3. Dimension values for satellite image surfaces have been reported to be much higher; depending on the type of landscape examined, they can be as high as 2.7–2.9 (Lam 1990). This major paradigm shift in geometry thinking was further elaborated and crystallized in Mandelbrot’s subsequent publications, which led to many followers and applications.

Since Mandelbrot’s first book *Fractals: Form, Chance, and Dimension* published in French in 1975 and translated in English in 1977, fractal analysis has blossomed and dominated the scientific literature for over two decades (Mandelbrot 1977, 1983). During the 1980s and 1990s, studies on fractals can be found literally in every major discipline, ranging from physics, chemistry, biology, engineering, and geography to music and arts. Scientists found fractals and fractal geometry an attractive concept that could be used to explain the complexity of the world. The movie and video game industry (such as “Star Wars”) utilized the fractal technique to
enhance the simulation and visualization, and this helped popularize fractal analysis to the non-science communities and the public during those years. The significance of the fractal concept has been said to be comparable to the concepts of entropy, chaos, or Gaussian distribution. It was once proclaimed that no one would be considered scientifically literate if the meaning and scope of fractals was not understood (Batty 1985; Lam 2009).

Fractal analysis has a long tradition in geography. Geographers started examining the fractal concept and its potential applications as early as the 1980s (Goodchild 1980; Lam 1980; Arlinghaus 1985; Batty 1985; Clarke 1986; Goodchild and Mark 1987). The fractal concept addresses fundamental issues that have long plagued geographers and spatial scientists, such as the issues of scale, map generalization, and landscape measurement and characterization. While fractal analysis has generated many successful applications and valuable findings, there are also major misunderstandings, problems, and misuses of fractals. Literature on fractals and fractal analysis is abundant across disciplines. This entry focuses on fractals as applied in geography. The goal is to provide a better understanding and appreciation of fractal analysis in geography. The basic concept of fractals is first explained and some key methods for computing fractal dimension and simulation are introduced. Then, a number of fractal applications are highlighted, with issues related to fractal analysis summarized. The entry concludes with remarks on issues and future development of fractal analysis.

Fractal measurement

In defining the fractal dimension, Mandelbrot employed a concept called self-similarity. Many spatial forms such as curves and surfaces are self-similar, meaning that each portion can be considered as a reduced-scale of the whole. The degree of self-similarity, expressed as a similarity ratio, is used to define the theoretical fractal dimension $D$ (equation 1):

$$D = \log N/ \log (1/r) \quad (1)$$

where $r$ is a similarity ratio (i.e., a scale reduction factor) and $N$ is the number of steps needed to traverse a curve. To illustrate how to determine the fractal dimension of a curve, consider a straight line and a divider step size of 1/4 of the line length (i.e., $r = 1/4$), it will take four steps ($N$) to traverse the straight line. Applying equation 1, $D = \log 4/ \log (1/(1/4)) = 1.0$. If the complexity of the line increases such that it takes six steps to traverse the line, then $D = \log 6/ \log (1/(1/4)) = 1.29$. As the complexity of the line increases, the fractal dimension increases, until it reaches a theoretical maximum fractal dimension of 2.0 for a line.

Since most natural forms, curves, and surfaces are not strictly self-similar but rather statistically self-similar, the fractal dimension will need to be estimated statistically through the regression technique. Regardless of the difference among different fractal measurement algorithms, the regression technique often involves three steps: (i) measure an attribute of the object (e.g., length, area) using various step sizes; (ii) plot the logarithm of the attribute of the object against the logarithm of the step size, also called the fractal plot, and fit a linear least-square regression line through the data points; (iii) estimate the slope of the regression line, which is used to determine the fractal dimension. For a curve, the fractal dimension is determined by equations 2 and 3:

$$\log L = K + B \log S \quad (2)$$

$$D = 1 - B \quad (3)$$

where $L$ is the length of the curve, $S$ is the step size, $B$ is the slope of the regression, and $K$ is a
constant. These equations show that the more irregular the curve, the greater the increase in length as step size decreases, hence the steeper the negative slope, leading to a higher fractal dimension. The $D$ value of a surface can be estimated in a similar fashion, but using equation 4 instead of equation 3:

$$D = 2 - B$$  \hspace{1cm} (4)$$

For a reliable regression estimate, at least five step sizes should be used. This means that in a logarithmic scale, if the initial step size is 2, then the subsequent step sizes would be 4, 8, 16, 32, and so on, which can quickly reach the limit of the line length. Therefore, in reality, the fractal plot seldom has a lot of data points for regression. The goodness-of-fit of the regression, the $r^2$ value, is used to indicate if the estimated $D$ value is reliable. Table 1 compares the estimated and theoretical $D$ values for three curves using the walking-divider method described above. The circle and the Koch curve were generated by computer routines, whereas the Louisiana coastline was digitized from a 1:1,350,000-scale map that resulted in 485 digitized $x,y$ coordinate pairs.

As can be seen from Table 1, the measured $D$ values are very similar to the theoretical $D$ values. However, it is also observed that the $r^2$ value of the circle measurement is very low ($r^2 = 0.489$), implying that the fractal plot is not linear. In fact, a number of studies have already found that fractal plots of empirical curves are seldom linear for all step sizes, indicating that true fractals with self-similarity at all scales are uncommon. This observation shows that fractals and “scale,” another central concept in geography, are intimately related. Instead of including all the data points to create a single regression line when they do not behave linearly, breakpoints can be used to delineate a specific range of scales at which the fractal dimension is measured. In other words, the fact that self-similarity exists only over a limited range of scales could be utilized to summarize the scale changes of an object or a pattern (Lam and Quattrochi 1992). This scale issue subsequently leads to an extension of basic fractal analysis into multifractal analysis to examine multiscale objects and patterns and local fractals to capture local variations.

Measuring the fractal dimension of curves is relatively straightforward, even though a number of arbitrary decisions have to be made, such as the choice of the initial step size, the number of step sizes used, and the range of scales (i.e., breakpoints) used to derive the slope of the regression. Extending the fractal dimension measurement to surfaces, however, is not straightforward, and many more arbitrary decisions are needed to be made. Numerous methods have been developed to calculate the fractal dimension of a surface, which unfortunately often lead to very different results. Fractal dimension is defined differently in different algorithms by different

<table>
<thead>
<tr>
<th>Object</th>
<th>Theoretical $D$</th>
<th>Measured $D$</th>
<th># Data points</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circle</td>
<td>1</td>
<td>1.0084</td>
<td>6</td>
<td>0.489</td>
</tr>
<tr>
<td>Koch snowflake</td>
<td>1.2619</td>
<td>1.2033</td>
<td>6</td>
<td>0.982</td>
</tr>
<tr>
<td>Louisiana coastline</td>
<td>$1 \leq D \leq 2$</td>
<td>1.2007</td>
<td>8</td>
<td>0.976</td>
</tr>
</tbody>
</table>

Source: Adapted from Lam and De Cola 1993, 25. Reproduced by permission of Blackburn Press.
FRACTAL ANALYSIS

researchers. Some algorithms use only a single measurement instead of multiple step sizes and regression analysis to derive the dimension. For example, FRAGSTATS, a free software package to compute landscape indices, defines the fractal dimension as the ratio between the perimeter and its area of a patch, which is different from many other fractal surface measurement algorithms. Previous research compared several algorithms, such as box-counting, variogram, isarithm, and triangular prism, and found that the latter two methods are more accurate. These two algorithms are briefly described here to illustrate their different approaches, so that a better understanding of the fractal surface measurement can be gained (Lam 1990; Lam et al. 2002).

The isarithm method for measuring the surface dimension is intuitively simple; it can be considered as a direct extension of the walking-divider method, except that the curves measured are now isarithms derived from the surface (e.g., contours in relief surface). In extending the method from curve to surface estimation, two additional considerations are needed – how many isarithms should be used to represent the surface and how should these isarithms be averaged to derive a composite $D$ value? Because of these ambiguities, another method, the triangular prism method, is often preferred.

The triangular prism method follows the same logic but uses a different strategy. It defines the surface by a mesh of grids and measures the area of the surface at different grid step sizes. The surface area at each step is the sum of the surface area of the triangular prisms constructed by connecting the height (i.e., $z$ values) of the four corners of each grid step and its center. The logarithm of total surface area is then regressed against the logarithm of step size (not step size squared as documented in Clarke 1986), and the fractal dimension is calculated as two minus the slope of the regression line as in equation 4.

The triangular prism method has been improved over the years and previous research found the method robust and accurate. A version of the improved triangular algorithm, plus several other fractal measurement algorithms, can be found in a software package called ICAMS (image characterization and modeling systems), which was developed by the authors and collaborators and is available for the public to download (www.rsgis.envs.lsu.edu/icams.asp) (Lam et al. 2002).

Fractal simulation

Another hallmark contribution of fractal analysis is its ability to simulate real-world objects and patterns. Enhanced with computer graphics color rendering, images generated from the fractal technique are often so striking and compelling that filmmakers, artists, and musicians use the method to produce extraterrestrial landscapes, fractal music, and design patterns (Peitgen and Richter 1986; Peitgen and Saupe 1988). The Mandelbrot set was reportedly the most complex mathematical set in the literature and has since become an icon of fractals. In geographic analysis, measurement using the fractal dimension is the first step towards scientific understanding of complex forms and processes, whereas simulation that mimics reality often leads to further understanding of spatial processes.

In searching for a stochastic model that can simulate fractal patterns closely, Mandelbrot employed the Brownian motion model in physics. Brownian motion refers to the irregular movements of small particles, the time series of which often shows a highly autocorrelated sequence, indicating a long-time dependence on previous movements. Brownian trails are fractal sets, with a fractal dimension of 1.5 for a Brownian curve and a fractal dimension of
2.5 for a Brownian surface. Mandelbrot combined Brownian functions with the concept of self-similarity and developed a new model called the fractional Brownian motion (fBm) functions. From these functions, curves and surfaces of other dimensions can be generated, hence they are also called the fractional Brownian motion (fBm) curves and surfaces. Several algorithms can be used to generate fractal curves and surfaces, including the shear displacement method, the modified Markov method, the inverse Fourier transform method, and the recursive subdivision method. Figure 1 displays three simulated surfaces in 3-D and the corresponding 2-D images generated by the shear-displacement method, which shows clearly that as \( D \) increases, the complexity of the surface increases (Goodchild 1980; Lam et al. 2002). The algorithm is also available in the ICAMS software previously mentioned.

Fractal applications

Fractal applications in geography generally fall into two groups. The first group uses fractals to generate real-world objects and patterns for both analytical and visualization purposes. The second group uses the fractal dimension or the power function as an aid in describing the complexity of spatial forms and patterns. Such applications can be further classified into three categories: applications in geospatial analysis, analysis of physical processes, and analysis of human phenomena. Each type of application is briefly described here.

In simulation applications, the fractal model provides a theoretically sound approach to simulate curves and surfaces for scientific analysis. The range of surfaces and curves simulated with a prescribed fractal dimension value is an ideal test-bed for evaluating geographical models and algorithms. For example, simulated fractal surfaces have been used as test datasets to examine the performance of various spatial interpolation methods and the efficiency of different spatial data structures. Another major type of applications is to use the fractal simulation technique as an interpolation method to generate coastlines and curves based on a few sample points. This curve generation process is, in fact, an inverse of curve generalization, and such algorithms could be implemented in a geographic information system for automated map displays at varying levels of cartographic generalization. In physical and environmental applications, researchers developed elaborated algorithms for deriving river networks from simulated fractal surfaces to help understand the various properties of geomorphic processes. In human geography, a number of studies applied the fractal model to simulate central place hierarchy, urban morphology, and urban growth. Last but not the least, researchers have used simulated fractal surfaces as a null-hypothesis terrain or norm, whereby further simulation of various geomorphic processes can be made to investigate the effects of different landform processes.

There are numerous applications of the fractal model in measuring and evaluating spatial phenomena. In the subfield of geographic information science and remote sensing, Goodchild (1980) in a pioneer article demonstrated that the fractal dimension can be used to predict the effects of cartographic generalization and spatial sampling on some common measures such as length, area, and point estimation. The dimensionality of the object or pattern can be input into an equation to estimate the relative benefit of using finer grid cells or denser sampling. This finding helps in determining the optimum resolution of pixels and polygons used in remote sensing and GIS studies.
Figure 1  Three simulated surfaces and corresponding images; from top to bottom, $D = 2.1$, 2.5, and 2.9. Adapted from Lam et al. (2002). Reprinted by permission of Taylor & Francis (http://www.tandfonline.com).
Fractal analysis of physical processes takes many forms. Researchers have used the fractal dimensions computed for different parts of a coastline to infer the underlying coastal processes that shape the coastline. A smooth coastline suggests the dominance of large-scale marine process such as ocean currents, whereas a contorted coastline is likely a result of local deltaic deposition and land subsidence. Other studies use the fractal plots to describe the power function relationships between physical variables, such as the power function between stream length and basin area and the time series of river bedload transport. In the latter case, the fractal structure of the time-series data of river bedload transport offers insight into the underlying mechanics of bedload movement and provides information that can be used to establish guidelines for bedload sampling (Lam and De Cola 1993). In ecological applications, the fractal dimension has been used together with other spatial indices to characterize the different tree types as manifested in remote sensing images. The fractal dimension has also been found useful to indicate the degree of human disturbance in natural landscapes; the higher the fractal dimension yielded, the least disturbed the landscape by humans.

Many human-dominated processes exhibit fractal patterns. The city’s rank–size rule is a well-known example of the fractal power function. Also, researchers have linked the classical Christaller’s central-place hierarchy in urban geography with fractals and found that a $K = 3$ central place hierarchy has $D = 1.26$, whereas $D = 1.13$ for a $K = 7$ hierarchy. For a $K = 4$ hierarchy, which is the most complicated boundary configuration, $D = 1.58$. It has been suggested that if two urban areas have the same fractal dimensions, it may mean that similar planning actions could be applied (Arlinghaus 1985; Lam and De Cola 1993). Batty and his collaborators have conducted extensive research on fractal cities (Batty and Longley 1994). Their research suggests that urban growth follows a form of fractal process, called the diffusion-limited aggregation (DLA) process, and that it is possible to classify a city’s development history based on the fractal dimension.

Issues and development

After the first wave of fractal applications in the 1980s–1990s, a number of problems associated with fractal analysis arose; further development of the fractal model and technique has since been made. First, the fractal model has been criticized by some researchers because self-similarity across all scales rarely exists in natural phenomena and reducing the complexity of spatial patterns into a single dimension runs the danger of oversimplification. The subsequent development of multifractals and local fractals has made the fractal analysis more flexible and applicable. Furthermore, instead of regarding the scale–dependency of fractals as a shortcoming, researchers have utilized the information on the changes of the fractal dimension with scale to help interpret the underlying spatial processes at specific scale ranges.

The second problem of fractal analysis refers to the inconsistency of measurement. Different researchers employ different algorithms and their measurement results often differ, sometimes substantially. How to determine which computed value is the most accurate one? And how do we know when a computed fractal dimension value is statistically significant? Consider two surfaces, one with a $D$ value of 2.2 and the other with a value of 2.3, are they statistically the same or different? Associated with the measurement problem is that most fractal curve and surface measurement algorithms have not been tested systematically. Also, these algorithms were seldom designed for general use or made available
FRACTAL ANALYSIS

to researchers. The difficulties and time involved in programming their own algorithms have prevented more geographical studies from using fractals.

Subsequent advances in the fractal studies have led to more algorithms and software being made available for use by the public. The development of a software package called the image characterization and modeling system (ICAMS) is one example (Lam et al. 2002). Funded by US federal agencies such as the National Aeronautics and Space Administration (NASA), ICAMS was developed to provide algorithms for computing fractal dimensions for surfaces and images, generating fractal surfaces, as well as computing a number of related spatial/textural statistics such as wavelets, lacunarity, spatial autocorrelation, and variograms. Benchmark studies comparing the fractal measurement algorithms implemented in ICAMS have also been conducted, which led to further improvement of the algorithms. The modified triangular prism method previously described is an improved algorithm in ICAMS. It was a direct outcome of making the software available for many researchers to use so that they can focus on better analysis rather than spending time to program the algorithms from scratch.

Significant advances have been made recently in applying the fractal model in image processing and remote sensing. The medical community increasingly uses fractal and multifractal analyses to characterize a wide range of images and signals, including electrocardiogram (ECG) and electroencephalogram (EEG) signals, brain imaging, mammography, and bone imaging. In geographical image processing, the fractal dimensions computed for local moving windows have been used as a texture layer for better image classification and change detection. This new local fractal approach is designed to overcome the problems of using a single fractal dimension to represent an entire study area. By applying the calculation in a small window area, the variations and complexity at fine scales can be better captured. Including the local fractal dimensions in image classification has resulted in a dramatic increase in classification accuracy. Moreover, unlike many spatial indices used in landscape ecology, fractal dimension can be computed directly from unclassified images. This property makes fractal analysis a very useful tool for change detection, which then paves the way for future automation in image classification and change detection.

Finally, recent studies have advocated using the fractal dimension as part of the metadata for spatial data such as remote sensing images. The argument is that if means and standard deviations are basic nonspatial descriptors for spatial data, then why not add the fractal dimension as an index of spatial complexity to fill in the missing basic spatial information that conventional nonspatial statistics do not provide? Such information should help speed up the mining of large image databases and the selection of appropriate imagery for many applications.

Conclusions

Fractals were literally everywhere in the 1980s and 1990s. Studies utilizing the fractal model and techniques have made tremendous impacts on many aspects of science, art, and technology, including advancing geographical sciences. However, the explosive number of studies in the first two decades is in marked contrast with the relatively few studies in recent years. This significant decline in the number of studies involving fractals can be seen as a natural life cycle of any academic research topic and may indicate that the study topic has now reached a certain level of maturity. Given the tremendous potential of the fractal model, it is useful here
to identify the research gaps and suggest future research using fractals that will help further our understanding of spatial phenomena.

At least four future research directions can be identified. First, more research is needed to develop the confidence limits and statistical significance of the fractal dimension measured. Second, the fractal dimension is a useful and competent index, but the fractal dimension alone is not sufficient to represent all aspects of spatial complexity; additional spatial indices such as wavelets, lacunarity, and spatial autocorrelation are needed to help in describing complex systems. Thus, the complementarities among these various indices should further be explored. Third, future studies should focus more on the validity of the research design and the interpretation of the fractal analysis results. Many previous fractal studies used complicated algorithms and data to generate results but stopped short in making a full and useful interpretation of how the patterns could be related to the underlying processes and how fractals have helped in explaining such patterns. Fourth, existing geographical applications of the fractal model have been mostly confined to explaining the present forms and patterns; its use in predictive process modeling has seldom been explored. The fractal model would seem to be a useful analytical addition to current research frontiers, including research on modeling the complex interactions between natural and human processes, agent-based modeling, big data mining, and system resilience and sustainability. The fractal model is a fantastic model of complexity; its use in explaining and modeling complexities is very promising and should be further explored.

SEE ALSO: Geographic information science; Measuring spatial dependence; Scale; Spatial analysis

References

FRACTAL ANALYSIS


The term “fragmentation of production” describes the distribution of the production process across different countries. It is an important feature of globalization. Products become internationalized; national markets are more closely linked to markets abroad; economy-wide behavior patterns converge; and the domain of overlapping policy interests expands (Arndt 2010).

The nature of fragmentation

The term “fragmentation” was first used in this context by Jones and Kierzkowski (1990) to explain the growth of service sectors related to the increase of intraindustrial international trade. They introduced two key concepts in their explanation: “production blocks” and “service links.”

The production blocks refer to the technical possibility of a production line being “fragmented” into small parts, so as to be carried out in different locations. The development of transportation and communication technologies significantly decreases the cost of coordination between production blocks that are dispersed in terms of distance. It becomes not only feasible but also rational to involve producers operating in different countries or even different continents in supply chains for certain products because the firm can take advantage of location-specific advantages in production such as proximity to resource supply, easy access to market, and convenient transportation. For example, an airplane may have its wings manufactured in Japan, its doors in France, and its seats in the United Kingdom. All the components are shipped to the United States, where they are assembled and sold as the final product.

In order to coordinate all the activities scattered in different locations, service links are crucial to facilitate the exchange of information between different blocks. Jones and Kierzkowski (1990) provided a framework to analyze the process of fragmentation (see Figure 1). In a traditional production process, inputs are organized and combined to generate final outputs. All aspects of production are conducted in one place, the proximity helping to reduce the cost of communication. As production becomes increasingly fragmented, the coordination is more involved and complex, and calls for more service links between the production blocks. A service link consists of activities such as transportation, insurance, telecommunication, quality control, and management coordination to ensure that the production blocks interact in the proper manner. The higher the speed and efficiency of the service link, the better the capability for production to take advantage of location-specific resources through fragmentation.

Fragmentation of production enables manufacturers to exploit component-specific economies of scale through vertical specialization. If the producer of a component can serve competing producers of final products, they may expand production for greater economy of scale, which in turn contributes to the growth of international
production sharing. With the establishment of service links, the specialized component producer will invest in technological innovation for fragmented production; this is not likely to take place in an integrated production system (Arndt and Kierzkowski 2001).

Beside the technical feasibility of international fragmentation, the separability of ownership, or controllability, can also affect the organizational structure. The firm has to make a choice of fragmenting production across both geographic and organizational boundaries, and may trade off fragmentation along one dimension for integration on the other (Kimura and Ando 2003). In order to organize the production blocks dispersed in different locations smoothly, firms provide service links for communication and coordination to facilitate the efficient and on-time transportation of parts and components within the network. Therefore, fragmentation will take place only when savings in production cost exceed the costs associated with service links. Besides economic concerns, the choice of ownership arrangement also depends on legal and institutional considerations especially when dealing with cross-border business; it involves trust, enforceability of contracts, and protection of intellectual property (Navaretti and Venables 2004).

The fragmentation of production used to be limited to within a country because domestic service links were easier to establish and

---

**Figure 1** Fragmentation of production and service links. Source: Jones and Kierzkowski (1990). © Reproduced by permission of Oxford University Press.
cheaper to operate. However, in recent years, technological innovations, the liberalization of international trade in services, convergence in legal and regulatory systems, and increased freedom of establishment have significantly reduced international coordination costs. As cross-border service links have become cheaper, more reliable, and more accessible, the locational options for production blocks have become global in scope (Arndt and Kierzkowski 2001). As mentioned earlier, multinational companies (MNCs) trade off between integration and fragmentation along both geographical and organization dimensions. Their decisions have impacts on the global pattern of international trade and foreign direct investment (FDI).

The analytical issues of fragmentation

Fragmentation developed well between advanced and emerging economies. It has grown fast in industries such as electronics, machinery, and textiles and apparel. In developing countries, the domestic value added to a country’s exports is often significantly less than the value of those exports. What factors determine the role of different countries in the division of labor in international production sharing? The phenomenon of cross-border fragmentation may be explained from a variety of theoretical perspectives related to international trade and FDI (Deardorff 2001).

First, the Ricardian classical theory of international trade provides a simple model based on absolute difference in technological capability between countries to explain cross-border fragmentation. According to this model, countries have different capabilities in technology, which determine intraproduct specialization among countries. Countries with limited technological development will engage in low-tech production, while more advanced components and assembly will be produced in countries with more sophisticated technologies. With advantages in technologies, FDI will move from technologically advanced to technologically emerging economies. But this theory cannot explain how the differences in location-specific technologies among different countries came to be.

Second, the Heckscher–Ohlin (H–O) framework attributes the basis for specialization to the differences in factor endowments of each country. Countries will specialize in activities that make intensive use of locally abundant factors of production. That is, other things being equal, labor-intensive component production and assembly will be carried out in labor-abundant countries, while skill- and capital-intensive activities will be performed in more advanced countries. This factor-proportions view provides a good explanation for vertical intraindustrial specialization, and predicts that FDI will tend to flow from capital- and skill-abundant, labor-scarce countries to labor-abundant economies. However, the H–O model cannot explain the dynamics of country-specific factor endowment. The shifting of FDI from developed to developing countries can help provide up-front capital for an industrial facility and technological capability in some developing countries, which could increase the skill and capital intensity of the local industrial base through the multiplier effects of industrial localization and human capital investment.

Neither the Ricardo nor the H–O framework is capable of handling scale economies and externalities related to the localization of industrial investment, which is important in the development of fragmentation. The imperfect competition literature provides important insights: that consumers value variety, and that monopolistically competitive firms specialize in producing variety. There is increasing return
to scale: economies of scale come into play on both the production side and in the provision of service links between production blocks. The concentration of production of each part or component in one location permits firms to exploit internal economies of scale at each location. Meanwhile, the agglomeration and clustering of producers generate economies that are external to the plant and firm.

The imperfect competition model provides a micro-economic mechanism for the fragmentation of the global production network. However, it simplifies most of the location-specific factors that can shape the production network in various ways. The factor endowments are only one type of location advantage, and therefore cannot fully explain the pattern of international trade and movements of FDI. The gravity model offers a way of accounting for additional location-specific considerations – including distance, country size, border effects, and various institutional and policy-related factors. It has been employed in the context of production networks to test the effects of distance on the trade and FDI in terms of both horizontal and vertical specialization. It suggests that FDI responds positively to distance. In horizontal specialization FDI is the substitute for exports. As distance and cost increase, exports are replaced by on-site production for the local market and hence FDI flow rises. At the same time, vertical FDI moves to support production sharing. When distance raises the cost of service links, the production sharing becomes unprofitable and FDI declines (Kimura, Takahashi, and Hayakawa 2007).

The effects of fragmentation

MNCs have played a key role in the evolution of cross-production networks. They reshape the local industrial structure through direct investment and purchasing activities. The initial organizational structure follows a hub-and-spoke pattern, with affiliates in several countries trading directly with the parent company and less with one another. As interactions become more sophisticated and infrastructure is developed, networks evolve in which trade takes place between affiliates located in various countries, as well as between affiliates and local firms. This pattern is exemplified by Japanese multinationals operating in Southeast Asia (Kimura and Ando 2003; Cheng and Kierzkowski 2001) and MNCs based in Singapore (Yeung 2001).

The geographical agglomeration of certain fragmented stages of production can generate positive externalities. The expanding market helps to develop specialized local industrial clusters, which attract and train workers, improve access to finance and other services, and generate knowledge spillovers that raise productivity and reduce costs. The local government can help the development of such agglomerations with supportive public policies and investment in infrastructure. Local producers can enter the industry as suppliers of goods and services to the multinationals, which opens the path to industrialization for developing countries.

The provision of service links can significantly benefit from economies of scale, too. Larger numbers of units and firms can share the fixed costs associated with establishing a communications network as more and more local participating entities join the network. Cooperative policies between countries to reduce the barriers in service sectors will benefit the development of agglomeration. Some empirical studies also show that production sharing tends to reduce the sensitivity of trade flows to financial fluctuation (Okubo, Kimura, and Teshima 2014) or exchange rate change (Arndt and Huemer 2007).
Anxieties about the spread of fragmentation relate to its impacts on jobs and wages in advanced countries, as a result of the offshoring of manufacturing, as well as services like call centers, help desks, programming, and ticketing operations. Production sharing with low-wage countries may increase competition for lower wages in a “race to the bottom.” It affects low-skilled workers more than the skilled workers. However, the real effects could be complex. Although jobs directly attached to shift industries have been lost, the cross-border sourcing of components could raise competitiveness and lower the prices of final products or services, thereby increasing sales and thus output. Thus, more jobs could be created elsewhere in the industry or more generally in the economy. And even unskilled workers in developed countries can enjoy price decreases as consumers. Empirical studies show mixed results in different sectors in relation to the skill ratios in affected industries, and the relationship between labor and capital. There is a trade adjustment problem related to fragmentation, for jobs lost in one sector must be balanced against jobs gained in others. Some policy interventions are needed to match unemployed workers with emerging jobs (Feenstra 1998).

Regional trade agreements can affect the development of fragmentation of production. Free trade agreements within preferential trade areas (PTAs) can help the development of production sharing in member countries, while rules of origin restricting the use of components from outside the PTA can significantly divert trade flows. At the same time, fragmentation of production is changing the nature of regional integration. It requires deeper integration beyond the removal of tariffs and other restrictions on the flow of goods and services. It calls for liberalization of investment and for the cross-border movement of persons, as well as the harmonization of technical standards, regulatory policies, and dispute settlement procedures. Cross-border coordination of production networks promotes convergence of business cycles and requires greater regional cooperation, including on monetary policy and market governance.

SEE ALSO: Corporate spatial organization and producer services; Global production networks; Producer services: definition and classification; Trade, FDI, and industrial development

References


France: Association de Géographes Français (AGF) (Association of French Geographers)

Founded: 1920
Location of headquarters: Paris
Membership: 220 (as of March 2015)
President: Roland Pourtier
Contact: nrpourtier@wanadoo.fr

Description and purpose

The Association of French Geographers (AGF) organizes four full-day public research seminars per year. Most meetings take place in Paris but sometimes they take place in other cities as well. Peer-reviewed proceedings are published in the bulletin, which also welcomes articles on any number of themes or topics. All PhD students are strongly encouraged to participate in AGF activities.

Paper sessions may focus on any aspect of our discipline, from geomorphology to cultural geography, regional to epistemology. Recent themes include global warming, heritage and territory, peri-urban agriculture, geography of the high seas, the United States, hazards in tropical environments, city dynamics in Southeast Asia, geography’s turning points, and re-industrialization in France.

Journals or major publication series

Bulletin de l’Association de Géographes Français

Current activities or projects

The association is a partner with the CNFG (French National Committee for Geography–IGU) and has organized joint seminars with specialized commissions of the CNFG. It is also a partner of the APHG (Association of History and Geography Teachers).

The board of the association is made up of 15 elected members (elected for three years, with an annual turnover of five members).

Brief history

The AGF was created in 1920 by Emmanuel de Martonne, who later became president of the International Geographical Union. Its headquarters are at the Institute of Geography, located in Paris’ Quartier Latin.

Submitted by Yves Boquet
France: Comité National Français de Géographie (CNFG) (French National Committee of Geography)

Founded: 1920
Location of headquarters: Paris
Website: www.cnfg.fr
Membership: 350 (as of 2014)
President: Richard Laganier
Contact: president@cnfg.fr

Description and purpose

The main aim of the CNFG is the coordination of French geographical activities (i.e., it launches or directs research which it deems needed or useful). It also serves as the French representative organization to the International Geographical Union (IGU) and mediates the communication between the IGU and French geographers. The CNFG also maintains communications with other international geographical organizations. Its 24 subcommittees pursue research and collaborations in various subdisciplines of geography. The CNFG also publishes a newsletter, issues awards and prizes on an annual basis, and organizes meetings and conferences throughout the year.

Journals or major publication series


Current activities or projects

The committee organizes and participates in a range of meetings, seminars, and conferences. It publishes the newsletter and represents French geography at the IGU meetings.

Brief history

The CNFG was created in 1920 under an initiative of the French Academy of Science. The association is now a member of the French Committee of International Scientific Unions (COFUSI) and represents French geography at the International Geographical Union, where it has a vote in the general assembly.

The CNFG has been a major actor in geography in France and abroad, notably for the French-speaking geography outside of France. It has organized the 1984 IGU Congress in Paris, and is planning to organize the IGU’s Centennial Congress in 2022, also in Paris.

Recently, the association has been trying to open up to new generations of geographers and to new methods and developing ideas in geography by creating new subcommissions and relaxing its membership conditions.

The CNFG maintains strong links with other French geographical associations and international events.

Submitted by Antoine le Blanc
France: Société de Géographie (Society of Geography)

Founded: 1821
Location of headquarters: Paris
Website: www.socgeo.org
Membership: 914 (as of December 31, 2013)
President: Jean-Robert Pitte
Contact: jean-robert.pitte@wanadoo.fr

Description and purpose

The Société de Géographie, established by royal decree on December 14, 1827, is a scientific and cultural nonprofit organization. The society's mission is to promote the knowledge and popularity of geography. The organization also aims to appeal to specialists as well as the general public by promoting the understanding and use of geography in all its diversity.

Journals or major publication series

La Géographie
Bulletin de liaison des members de la Société de Géographie

Current activities or projects

The society pursues multiple activities: it organizes monthly conferences, alternating them with those of the Explorers’ Society, which it hosts and with which it actively collaborates. The society also organizes one or two round tables each year and publishes edited papers from these events. It also serves its members through two publications: the journal La Géographie, intended to more widely disseminate the work of university academics or writers concerned with the larger contemporary questions in geography; and the Bulletin for society members, which publish information on the life of the society and its members’ activities. The two publications share a review of other works or publications that contribute to the advancement of geographic knowledge. The society's valuable photography collection is well-illustrated in a collection of albums, co-published with éditions Glénat. The society also organizes expeditions and voyages. The most recent such trips include a crossing of the Morvan, a cruise on the Saône, a long journey through Uzbekistan, a cruise in the Venetian Lagoon, and an educational field trip to northern Cotentin, following an address by Barneville-Carteret. Lastly, once annually, at the end of November, a Day of Geography allows the society to gather a large number of members who come to listen to presentations, join in debates, see films, and more. The society's awards are distributed at the close of this special day.

Brief history

Founded on December 15, 1821, the Société de Géographie is the oldest in the world. It was founded 7 days before its counterpart in Berlin and 9 years before the Royal Geographical Society in London.

Its first bulletin was published in 1822 and it was declared a public service by Charles X in 1827. Geographers and cartographers of this period were mostly dedicated to the creation of maps, atlases, and geographical dictionaries. They did not possess a monopoly on the description of the world regions that were published by navigators, the military, religious orders,
and natural scientists. The first president of the society, the Marquis Pierre-Simon de Laplace was a mathematician, physicist, and astronomer, a member of the Académie des Sciences since 1773 and of the Académie française since 1816. He never authored documents related to geography, but he was a respected scientist who legitimized the creation of an organization dedicated to this branch of human knowledge.

The society encouraged numerous explorers, contributed financially to their travels, and also remunerated them upon their return home. The society has one of the most important specialized libraries in the world, collecting manuscripts, archives, and photographs. Since World War II the collection is hosted at the Department of Maps in the Bibliothèque nationale de France, which ensures its conservation and access for the public.

Today, the society has approximately 1000 members and about 100 honorary members residing on all the world's continents. Among these are Prince Albert II of Monaco and Prince Charles of England, both passionate promoters of environmental and natural resources management.

Submitted by Jean-Robert Pitte
Free trade zones (FTZs), also known as commercial free zones and free commercial zones, are geographically delimited, usually physically secured (fenced-in), duty-free areas, offering warehousing, storage, separate customs, streamlined procedures and distribution facilities for trade, trans-shipment, and re-export operations. They are defined spaces where laws and regulations, especially those related to trade, are more liberal than in the rest of a country. Typically, the territory set aside is considered outside a country’s customs regime. Normal trade barriers such as tariffs and quotas are eliminated and bureaucratic requirements are lowered to attract new business and foreign investments. FTZs are often located at or near a port of entry to facilitate import and export or an underdeveloped part of the host country to boost employment and economic growth.

In FTZs, trade-related activities have become increasingly integrated with manufacturing activities. FTZs of the United States, for example, are import- and trade-oriented, but they also permit manufacturing. Most export processing zones of the world have also incorporated the functions of FTZs but are primarily export- and manufacturing-oriented. Here a broad conceptualization is adopted and FTZ treated as a generic term that may (partially) encompass many practical variants in the name of free export zones, duty-free zones, industrial free zones, export processing zones, special economic zones, industrial export zones, free ports, enterprise zones, bonded warehouses, and so on. The different terms over time and space often reflect the specific activities carried out within a particular zone.

While nation-states have developed “exceptional” spaces of economic activity for over a century, the recent phenomenon and proliferation of FTZs refer to a period beginning in the late 1960s, when developing countries sought to attract investment through territory-based concessionary incentives. The typical FTZ policy package includes: (i) partial or complete exemption from export taxes, duties on imports of raw materials or intermediate goods; (ii) exemption from direct taxes such as profits taxes, municipal and property taxes; (iii) exemption or deferral from payment of indirect taxes such as value-added tax and goods and services tax on domestic purchases; (iv) exemption from national foreign exchange controls; (v) free profit repatriation for foreign companies; (vi) provision of streamlined administrative services with less bureaucracy and red tape; (vii) free or subsidized provision of enhanced physical infrastructure for office space, utilities, production, transport, logistics, and business services; and (viii) other incentives (often offered by regional/local governments) to attract jobs and investment.

Free trade zones in a broad sense have existed for centuries. They were originally established to encourage entrepôt trade, and mostly took the form of citywide zones located on international trade routes, for example, Gibraltar (1704), Singapore (1819), Hong Kong (1848), and Hamburg (1888). An entrepôt was a central...
point where goods were assembled and physically traded before being exported to final destinations. Closely related to the entrepôt function is re-export and trans-shipment trade.

The development of the US foreign trade zone program in 1934 was the next major milestone in the history of free zones. At the time tariffs were as high as 60% for goods entering the United States, so the foreign trade zone program was perceived as a counterweight to offset the harmful effects of trade protectionism. The program was also expected to create and retain jobs and investments, and to help major US trade centers to compete against their international rivals such as London.

The first modern industrial free zone was the Shannon Free Zone (SFZ), established in Ireland in 1959. A further extension of the free zone concept, known as the export processing zone (EPZ), was developed in Kaohsiung, Taiwan, in 1965, and subsequently became the role model in many developing countries. The EPZ is a combination of old customs areas and export-oriented manufacturing. EPZs combine tax-free holidays with other incentives for foreign investors to set up factories that were expected to attract foreign direct investment, produce export goods, train low-skilled workers, and facilitate technology transfer. Before the 1970s, most zones were clustered in industrialized countries, primarily in Western Europe. Inspired by the role model of Shannon and Kaohsiung, EPZs subsequently proliferated throughout Asia, as well as in Latin America and in the Middle East. In the 1980s the pace of zone development increased and expanded to new regions, including South Asia, South America, and sub-Saharan Africa. The number of zones – especially EPZs – has grown dramatically, particularly since the mid-1990s.

Overall, enhancing trade efficiency and manufacturing competitiveness is the principal rationale behind free zone programs in most industrialized countries. Revitalization of economically distressed urban and rural areas is the motivation behind the many free zone programs in developed countries. The main rationale for the Shannon Free Zone in Ireland was to establish a “growth pole” in the economically distressed southern part of the country. For developing countries, free zones have traditionally had both a policy and an infrastructure rationale. The policy rationale is associated with the expected economic benefits from FTZs, which commonly include: attracting foreign direct investment, creating employment and generating income, earning foreign exchange, developing and diversifying exports while maintaining protective barriers, piloting new policies and approaches, transferring knowledge to local firms through backward linkages, and catalyzing entrepreneurship and innovation in the local community. The infrastructure rationale is aimed at enhancing the competitiveness of manufacturers and service providers through fully serviced sites with purpose-built facilities for sale or lease. It is also intended to realize agglomeration economies and more efficient environmental controls from concentrating industries in one locality.

Today, a relatively small number of countries account for the majority of worldwide zone activity. Zones are concentrated in Asia and the Pacific (mainly China), Latin America, and Central and Eastern Europe and Central Asia. The Asia–Pacific region has been at the forefront of zone development over the last three decades, led by the “Asian Tigers” and China. The development of free zones has been the epitome of China’s market-oriented reforms and opening up since 1979, as well as a major catalyst of this process. The opening of China and its market reform policies have relied heavily on all kinds of zoning technologies and the production of new spaces of exception (Ong 2006). This strategy allows the authoritarian Chinese party-state
to create and accommodate islands of distinct governing regimes, or laboratories of capitalism, within the broader landscape of at least nominal, enduring socialist rule.

Traditional free zones have played a key role in the process of industrialization, especially the development of export sectors. Newly industrialized Asian economies are often pointed to as models in utilizing their zones to both promote and diversify their export bases. These economies have to a certain extent succeeded in moving up the value chains and encouraging exports in a wide range of industries. Nevertheless, most zone enterprises worldwide are still engaged in labor-intensive, assembly-oriented activities in low value-added sectors such as apparel, textiles, toys, footwear, electronics, and auto parts. African FTZs are generally marked by low levels of investment and exports, marginal employment impact, limited attraction of FDI, and an absence of linkages with the domestic economy. Widespread failure has raised serious doubts about the desirability of the free zone programs.

FTZs have frequently been criticized for negative socioeconomic impacts, for example, exploitation of women, suppression of labor standards and core labor rights including trade unionization, low wage levels and poor employment conditions (work hours, health, and safety), and lax environmental standards. The attraction of the FTZs to foreign investors is frequently the ready availability of cheap but disciplined labor. FTZs have often been characterized by a lack of social dialogue between employers and the representatives of workers, many of whom are migrants or working in industry for the first time. Foreign enterprises often take advantage of the weak enforcement of safety, health, and environmental regulations in the zones.

Many have also cautioned against a race to the bottom in relation to social and environmental standards, and of a race to the top in terms of incentives to investors of the FTZs. The package of fiscal incentives offered by FTZs is increasingly similar and more generous around the world. These incentives have imposed significant costs on government budgets but are increasingly finding it difficult to live up to public expectations. The long-term success of EPZ-led development rests on the ability of nation-states to seize foreign demand and continuously direct resources to higher-productivity areas. To a great extent, the likelihood that FTZs will contribute to wider economic and social development hinges on their embeddedness in the host country’s economy, through forward and backward linkages between zone enterprises and local suppliers of goods and services. Linkages are usually dependent on the host economy having an existing industrial base to produce the necessary absorptive capacity. Some countries lack the fundamental investments in human capital necessary to enable the FTZ to be a positive experience. Poor education and health problems are certainly fetters on the likelihood of expanding the career potential and job satisfaction of workers that lead to improved productivity and stability. Others face strong political barriers to building free, strong, and representative workers’ organizations that can play a major role in improving working conditions and increasing productivity and competitiveness. These are usually larger societal challenges that are difficult to rectify in the short run.

SEE ALSO: Corporations and global trade; Economic development zones; Export processing zones; Global commodity/value chains; Global production networks; Human capital; Industrial districts; Industrial linkage; Industrial upgrading; Industrialization; Infrastructure and regional development; Institutions and development; Labor geographies and the corporation; Migrant labor;
FREE TRADE ZONE

Newly industrializing economies (NIEs); Science and technology parks; Skill; Space of exception; Trade, FDI, and industrial development; Trade and regional development

Reference


Further reading


Freshwater resources: past, present, future

Lene Petersen
Institute of Environmental Engineering, ETH Zurich, Switzerland
World Wide Fund for Nature (WWF), Switzerland

Martin Heynen
Institute of Environmental Engineering, ETH Zurich, Switzerland

Francesca Pellicciotti
Institute of Environmental Engineering, ETH Zurich, Switzerland
Northumbria University, UK

Definition, facts, and figures

Freshwater, also called sweet water in contrast to saltwater, is defined as water having a concentration of less than 500 ppm (parts per million) dissolved salts. Of all the water on Earth, 97.5% (approximately 1.41 billion km³) is saltwater and only 2.5% is freshwater – but this is of great importance and plays a crucial role in human activities and natural ecosystems (see the next section). Freshwater comprises all natural water on the Earth’s surface in the form of lakes, rivers, ponds, glaciers, ice and ice sheets, and other wetlands and beneath the surface in the form of groundwater and soil moisture as well as the water contained in the atmosphere. Almost 70% of the freshwater is stored in the icecaps and as soil moisture or is present as groundwater in deep aquifers not accessible for human use. Less than 1% of the global freshwater (which corresponds to roughly 0.007% of all water on Earth) is accessible for direct human use. This is the water of lakes, rivers, reservoirs and aquifers shallow enough to be tapped. This water is renewed by precipitation on a regular basis and therefore available for sustainable use. Groundwater is by far the most abundant and readily available source of freshwater, followed by lakes, reservoirs, rivers, and wetlands (UNEP 2008). The sources of freshwater and the corresponding amounts are summarized in Figure 1 (WWAP 2006, based on data from Shiklomanov and Rodda 2003). As with most estimates at this scale, there is some uncertainty associated with them, especially for the estimate of fresh groundwater resources.

Freshwater amounts vary considerably in space and time. In space, its availability is unevenly distributed. The global freshwater resources as volume per continent are shown in Figure 2. Most freshwater is stored in the ice caps, with Antarctica containing by far the largest amount. For all continents, the major amount of freshwater is groundwater and nearly all of it is fossil, thus not suitable for renewable water consumption. Only very little freshwater is readily available and can be used in a sustainable way. This is significantly less in the Southern Hemisphere (Africa, South America, and Australia) than in the Northern Hemisphere.

The hydrosphere is all of the Earth’s water, including surface water (water in oceans, lakes, and rivers), groundwater (water in soil and beneath the Earth’s surface), snowcover, ice, and water in the atmosphere, including water vapor. In time, the amount of water in the hydrosphere varies due to the exchange between ocean,
land, and atmosphere, the so-called hydrological cycle. The water cycle plays manifold roles in the climate, chemistry, and biology of the Earth and freshwater ecosystems, being of great importance to maintain freshwater. Figure 3 shows the scheme of this cycle. The water moves from one reservoir to another, for example, from river to ocean or from the ocean to the atmosphere, by the physical processes of evaporation, transpiration, sublimation, condensation, precipitation, infiltration, runoff, and subsurface flow, thus going through the different phases of liquid (water), solid (ice), and gas (vapor). The main driver of these processes is the radiative energy from the sun, which heats water in oceans, lakes, and wet soils. Water evaporates as water vapor into the air. Ice and snow can sublimate directly into vapor. Evapotranspiration is water transpired from plants and evaporated from the soil. Rising air currents take the vapor up into the atmosphere where cooler temperatures cause it to condense into clouds. Air currents move water vapor around the globe, cloud particles collide, grow, and fall out of the upper atmospheric layers as precipitation. Some precipitation falls as snow, hail, or sleet and can accumulate as ice caps and glaciers, which can store frozen water for thousands of years. Thus mountains, glaciers, and ice caps play an important role in many areas of the globe as they store water and release it through melting during the warmest seasons. Most water falls back into the oceans or onto land as rain where the water flows over the ground as surface runoff. A portion of runoff enters rivers with stream flow moving water toward the oceans whereas some water infiltrates into the ground and into aquifers. Runoff and water emerging from the ground (groundwater) may be stored as fresh water in lakes. In river valleys and floodplains there is often continuous
water exchange between surface and groundwater beneath and alongside a stream bed. Over time, the water returns to the ocean continuing the water cycle.

Vörösmarty and Sahagian (2000) estimate the total renewable water supply to 39,600 km³ with about 75% accessible to humans, the majority coming from forests and mountains. The global fraction of total runoff accessible to humans shrinks to 53% if uncaptured flood flow is not considered. About 80% of the global population live downstream of renewable freshwater services while 20% remain without sustainable supply (UNEP 2005) and rely on unsustainable water resources. Forest and mountain ecosystems serve as source area for the largest amount of (renewable) freshwater supply. With 57% of the total, and 28% of the renewable runoff, they supply at least 4 billion people with freshwater (UNEP 2005). More than 90% of a river’s flow may be derived from catchment headwaters and the hydrology of downstream areas depends directly on the water coming from upstream. The connection of upstream and downstream areas is very important and threatened by many factors. The transmittance of upstream hydrology to downstream is crucial for the functioning of the ecosystem and thus the maintenance of ecosystem services. Global surveys like the Millennium Ecosystem Assessment (UNEP 2005) or the Global Biodiversity Outlook identify freshwater ecosystems as having suffered from greater degradation than any other global ecosystem (see below).
Freshwater ecosystem services – why do we need freshwater

Water is essential to all forms of life and is needed in almost all human activities, from basic drinking water to food production and health, from energy production to industrial development, from sustainable management of natural resources to conservation of the environment. Water also has a cultural and religious value. Together with energy and nutrients, water is a core piece for the delivery of all ecosystem services to humankind. All species on Earth, including humans, depend on clean freshwater from freshwater ecosystems, which in turn depend on healthy rivers and basins. Maintaining healthy ecosystems is thus crucial for the maintenance of the services they provide. Scientists have discussed ecosystem services implicitly for decades but the ecosystem services concept was popularized by the Millennium Ecosystem Assessment (UNEP 2005) and is since then widely used, even though it is still subject to criticism, especially for the concept of cultural ecosystem services. The ecosystem services literature focuses on the benefits for humans but discussions about the intrinsic value of nature are persistent, arguing along the lines that nature has a value beyond money and beyond direct human services. Also the value as a resource base beyond attaching ecosystem services to it should be considered. However, estimates have been made in attaching a monetary value to ecosystem services,
Table 1  Freshwater ecosystem services, adapted from UNEP (2005) and Le Quesne et al. (2010).

<table>
<thead>
<tr>
<th>Provisioning services</th>
<th>Regulating services</th>
<th>Cultural services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Products obtained from ecosystems</td>
<td>Benefits obtained from regulation of ecosystem processes</td>
<td>Nonmaterial benefits obtained from ecosystems</td>
</tr>
<tr>
<td>Water (quantity and quality) for consumptive use</td>
<td>Maintenance of water quality</td>
<td>Recreation</td>
</tr>
<tr>
<td>– drinking water</td>
<td>– natural filtration</td>
<td>Tourism</td>
</tr>
<tr>
<td>– domestic use</td>
<td>– water treatment</td>
<td>Existence value, aesthetic value</td>
</tr>
<tr>
<td>– agricultural use</td>
<td>Flood and erosion control</td>
<td>Cultural, spiritual and religious value</td>
</tr>
<tr>
<td>– industrial use</td>
<td>– flow regulation</td>
<td>value</td>
</tr>
<tr>
<td>Water for nonconsumptive use</td>
<td>– sediment transport</td>
<td></td>
</tr>
<tr>
<td>– power generation</td>
<td>Flow to marine systems (e.g., mangroves) and prevention from saline intrusion</td>
<td></td>
</tr>
<tr>
<td>– transport/navigation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fiber and fuel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biochemicals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetic material</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biodiversity (aquatic organisms for food and medicine)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supporting services</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Services necessary for the production of all other ecosystem services</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Role in nutrient cycle (primary production)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biodiversity maintenance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predatory/prey relationships and ecosystem resilience</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

reporting an amount of US$6.5 trillion per year as the value attached to freshwater ecosystems (Costanza et al. 1997). The concept of attaching a monetary value to ecosystem services is also subject to ongoing discussions and criticism. Table 1 summarizes the ecosystem services of freshwater divided into provisioning services (products obtained from ecosystems), regulating services (benefits obtained from regulation of ecosystem processes), and cultural services (nonmaterial benefits obtained from ecosystems). In the Millennium Ecosystem Assessment (UNEP 2005) freshwater is mainly classified as provisioning service referring to the domestic, agricultural, and industrial use as well as for power generation and transportation.

A main service considered in many studies is the use of freshwater by humans for domestic, agricultural, and industrial purposes and geographic distribution of human settlement can be directly linked to the availability of freshwater. The annual global freshwater use is reported to be approximately 3600 km$^3$ corresponding to 25% of the annual total runoff of the total accessible runoff of 53% (Shiklomanov and Rodda 2003). Nearly half of the global total is used in Asia and a third in OECD (UNEP 2005). As freshwater availability varies in space and time, overuse,
resources depletion, and water scarcity are important subjects related to freshwater use. The related problems are mainly aquifer depletion, soil salinization, and competition for water. The nonrenewable use is affected by large uncertainties and is reported to lie between 5% and 25% (UNEP 2005). All continents record overuse, predominantly the arid regions of North Africa and the Middle East and 20% of the global population remains without enough water, thus relying completely on unsustainable water resources. Impacts of unsustainable water use can directly be linked to food security. Globally the major part (70%) is used in agriculture, followed by industrial use (22%) and domestic use (8%). These figures differ significantly for high-income countries, which use the greatest amount for industrial purposes, and low/middle-income countries, using an even higher percentage for agriculture. The use in the agricultural sector is mainly defined by irrigation and leads to an irretrievable loss of a third of the global water use (UNEP 2005). The efficiency of agricultural irrigation varies greatly and losses through evapotranspiration over irrigated cropland are reported to be on average 50%. Water stress or water scarcity is the lack of sufficient accessible water to meet the demands within a region. Two common indicators are the amount of water available per person and the ratio of volume of water withdrawn to volume of renewable water available. According to the Falkenmark Water Stress Indicator (Falkenmark and Lindh 1976) values between 1000 and 1700 m³/annum per capita indicate water stress and values less than 1000 m³/annum per capita indicate extreme water stress. Figure 4 shows water stress per country calculated as the ratio of water withdrawal and potential renewable supply. However, these indicators give only partial indications of water resources pressures in a country or region because the consequences of water stress also depend on how water is managed. Another approach therefore divides the water scarcity into physical and economic

Figure 4 Water stress (ratio of water withdrawal and renewable supply) per country. Graphic: Martin Heynen.
water scarcity. Physical water scarcity indicates that the water resources are approaching or have exceeded sustainable limits whereas economic water scarcity indicates that malnutrition exists because human, institutional, and financial capital limit access to water even though water in nature is available locally to meet human demands (International Water Management Institute 2007). The Human Development Report (Watkins 2006) states that water scarcity already affects every continent. Around 1.2 billion people, or almost one-fifth of the world’s population, live in areas of physical scarcity, and 500 million people are approaching this situation. Another 1.6 billion people, or almost one-quarter of the world’s population, face economic water shortage.

Worldwide 1–2 billion people are reported to live under water scarcity conditions, and the problems associated with water stress and scarcity are expansive in the areas of human health, food production, and economic development. Also, the water scarcity coping capacity varies significantly with the level of economic development, education, and governance (also see UNEP 2005).

Past, present, future – how is freshwater changing?

Past

Since Antiquity, drainage, irrigation, and impoundment have been the major human activities impacting water. Since at least 5000 years ago, most water has been used for irrigated agriculture. L’vovich and White (1990) report a fifteenfold increase in global water withdrawal between 1800 and 1980 and the overall increase since 1900 is reported to have been sixfold (WMO 1997). In the late nineteenth century and early twentieth century dams became a popular water management tool and to date around 45 000 bigger dams were built, most between 1950 and the early 1980s associated with an increase in global water withdrawal since the 1950s due to economic development and industrial water consumption as well as an expansion of irrigated areas. Freshwater withdrawals have almost doubled since 1960 and more than half of all accessible freshwater runoff is currently used by humans. As a result, the dynamics in time and the spatial distribution of water are no longer determined only by natural climate variations (Shiklomanov in Gleick et al. 1993). Together with anthropogenic changes of the hydrological cycles alteration of the water quality also occurs, which is severely affected by industrial and agricultural chemicals. It was not before the Clean Water Act (first federal law governing water pollution in the US) in 1972 that humans officially and globally recognized the need to protect water. The index of health of freshwater ecosystems shows a decline of 50% between 1970 and 1995 and extinction rates of freshwater species are reported to be much higher than those of other species (terrestrial, coastal marine). Rivers and their floodplains, lakes, and wetlands have undergone more dramatic changes than any other type of ecosystem in the recent decades due to multiple human activities.

Present

The greatest present and future challenge to freshwater systems is the management of human and ecosystem needs. The biggest threats to freshwater ecosystems are land-use disturbances, hydrological alterations, and the introduction of non-native species. The first two can be closely linked to population growth and economic development. Land-use practices and changes (e.g., deforestation, expansion of agricultural
and irrigated land) affect freshwaters by modification of nutrient and sediment loads, changes in water temperature, and increased pollution due to fertilizers, pesticides, and waste. Land gets degraded leading to water stress and increased water scarcity. The main drivers of hydrological alterations are dam constructions for hydropower use, diversion of water for multipurpose use, the draining of wetlands, and constructions for flood control and agriculture. Potential changes in the hydrological regime due to impoundment, channelization, water withdrawal, or catchment conversion have an influence on key environmental parameters: water quality (water temperature, oxygen level, nutrients), energy sources, and also the physical habitat structure. They can also lead to fragmentation of aquatic habitat and groundwater depletion. A river’s natural flow dynamics maintain and support key ecological processes and vital life-cycle stages, thus alterations can significantly change an ecosystem or its functioning (Poff et al. 1997). As the flow regime influences all other key environmental factors such as water chemistry, physical habitat, and biological composition and interaction, “flow regime” has been described as the “master variable” by scientists (e.g., Poff et al. 1997) and the primary determinant of freshwater ecosystem functioning. Excessive changes to the natural flow regime can cause river ecosystems to degrade with high costs for both biodiversity and society (Richter et al. 2003). The five components of the natural flow regime defined by Poff et al. (1997) are magnitude, timing, frequency, duration, and rate of change. Maintaining the full range of variation in natural hydrological regimes is vital to sustain the integrity of aquatic ecosystems. As the ecosystems’ need for water competes with the human needs, they require protection in order to maintain their capability to provide goods and services to people, now and for future generations. An understanding of freshwater ecosystem integrity is therefore essential to maintain the range of goods and services that underpin livelihoods and communities, especially in developing countries (Le Quesne et al. 2010).

Humans now capture more than 50% of freshwater runoff, river systems have been fragmented by approximately 1 million dams globally and reservoirs trap 25% of the global sediment load before it reaches the oceans (Vörösmarty and Sahagian 2000). The Flowing Forward report by Le Quesne et al. (2010) concludes that by 2020 there will be few, if any, untouched ecosystems and that there are many water bodies likely to be profoundly transformed in their key ecological characteristics. These will inevitably have crucial impact on livelihoods and human development.

Future

Climate and freshwater are interconnected in complex ways and climate change should be seen as part of a broad set of pressures on freshwater systems. Observational records and climate change projections provide evidence of freshwater resources being vulnerable and having the potential to be severely impacted by climate change with wide-ranging consequences on human societies and ecosystems. Most impacts on freshwater are expected to occur as a result of climate change with other anthropogenic pressures, such as changes in population, lifestyle, economy, and technology and climate change is not necessarily the predominant driver of freshwater ecosystem degradation. However, it will endanger human use of freshwater and freshwater species, driving engineering responses such as new dams, dredging, levees, and water diversion to improve water security that will further threaten freshwater ecosystems and the human responses to climate change could cause effects as large or larger than the effects of climate change itself. As climate change is not a
smooth and gradual process, but very complex and affected by large uncertainties, so will be the impacts of it on freshwater ecosystems. But they will lead to changes in quantity, quality, and timing of water and thus to the entire freshwater flow regime. Not all freshwater ecosystems will be affected in the same way due to the variability of climate change in space and time leading to diverse degrees of vulnerability and resilience. Particular elements of the ecological (freshwater) system will be at risk at particular points in time and space, and to particular kinds of changes and stressors (Le Quesne et al. 2010). Significant changes in water availability and use or the hydrological cycle (affecting water supply and floods) require adaptation in the management of water resources. Most impacts are expected in the most vulnerable regions of the world and already sensitive ecosystems. The key variables for change (following Le Quesne et al. 2010) are as follows.

- **Temperature**: increased temperatures, especially over land, leading to changes in volume and timing of water availability and changes in water temperature.
- **Precipitation**: increased global precipitation highly variable in space and time leading to changes in timing and amount of precipitation, increase in number of extreme events and thus an increase in droughts and floods.
- **Evapotranspiration and sublimation**: increased evapotranspiration as a function of humidity, radiation, wind and temperature as well as increased sublimation from ice and snow can be seen as direct loss of water in the basin.
- **Runoff**: as a combination of precipitation and evapotranspiration also changes in runoff such as increased low-flow episodes and associated water stress or shifts in amount and timing of floods will occur, which are geographically highly variable. Also changes in seasonality occur, especially in mountainous regions where less solid precipitation is falling due to higher temperatures. Sediment and pollution load will change following the runoff.

Gradual changes (changes around mean values, e.g., increase of extreme events) are significant for ecosystems (species and people) and hardest to predict, but there are also modal changes (e.g., periods of rapid shifts like glacial melting). IPCC has concluded that the atmosphere holds more water vapor and that global precipitation seems to be increasing. However, this is not uniform over the globe, but unevenly distributed in space and time. The main impacts of climate change can be summarized as follows (compare to Le Quesne et al. 2010).

- **Altering timing/availability of freshwater**: fast melting of glaciers, changing precipitation patterns, changes in groundwater recharge pattern, changes in amount of solid precipitation and timing of melt, increased evaporative losses.
- **Increase/shift in timing of extreme events**: not enough water (droughts, food crisis) versus too much water (floods, storms, water-borne diseases), changes of habitat structures.
- **Loss of biodiversity due to decline in habitats**: increasing temperatures, increasing pollution, changes in sediment load, water temperature, and so forth.
- **Increased water stress**: longer low-flow periods (reduced habitats, increased temperature, and pollution).

Water requirements for agriculture could double by 2050 without even considering climate change (International Water Management Institute 2007). Traditional pressures with
uncertain future trajectories are expected to be superimposed by climate change pressures, which are hard to predict. Attempts of human societies and economies to adapt to climate change may decrease the ability of ecosystems to adapt (Le Quesne et al. 2010).

With anticipated population growth and economic development and associated phenomena, a further increase of water stress and scarcity is very likely. UN reports that projections of water use are to increase by 50% for developing and 18% for developed countries and that by 2025, 1800 million people will be living in countries with absolute water scarcity and two-thirds of the world’s population will be affected by water stress conditions.

The future rate of freshwater species’ extinction is predicted to be almost five times greater than for terrestrial animals and three times greater than for coastal marine species. Conservation strategies need to focus on hydrologic and water chemistry regimes linked to species diversity, population dynamics, and habitat quality.

Outlook

Freshwater ecosystems are currently under severe threat. In order to maintain healthy freshwater systems that are able to serve as resource base and provide the necessary services to livelihoods, the reduction of the existent pressures and the protection of freshwater ecosystems are crucial. The reduction of existing pressures and thus the maintenance of ecosystem functions, while still allowing development, is a key challenge. A focus point of sustainable resources management remains the tradeoff between water used to satisfy human demand and the water needed for sustaining aquatic ecosystems. A good definition of the goal of sustainable water management is given by Richter et al. (2003): “Ecologically sustainable water management protects the ecological integrity of affected ecosystems while meeting intergenerational human needs for water and sustaining the full array of other products and services provided by natural freshwater ecosystems. Ecological integrity is protected when the compositional and structural diversity and natural functioning of affected ecosystems is maintained.” Sustainable management of freshwater resources has gained importance at the regional and global scale and “integrated water resources management” (IWRM) has become the corresponding scientific paradigm. The integrated basin-wide planning considers the whole catchment as management unit and respects the interdependences of the freshwater system (upstream–downstream, groundwater–river, etc.). It tries to involve all stakeholders in the management process and decisions, thus being an iterative approach to water management. The key to translate the latest knowledge into action is institutional capacity and good stewardship. Frameworks and institutions need to be enabled and policies and laws at the international and national level in the water sector as well as general environmental programs need to be strongly formulated.

SEE ALSO: Ecosystem services; Environmental change and social learning; Global climate change; Global environmental change: human dimensions; Hydrologic cycle; Natural resources; Rivers and river basin management; Sustainable development; Water and climate change; Water resources and hydrological management

References

FRESHWATER RESOURCES: PAST, PRESENT, FUTURE


Further reading


Friendship, geographies of

Tim Bunnell
National University of Singapore

Friendship refers to a broad range of interpersonal relations that are entered into and may be left voluntarily. The voluntary character of friendship distinguishes it conceptually from kinship ties (that exist and persist regardless of personal choice) and marital bonds (that require legal and/or religious procedures to be broken), although in practice the distinction between friends and kin is often blurred and people may also regard their spouses as friends. Geography is implicated in different ways in friendship formation, maintenance, and dissolution. Advances in technologies of communication are extending possibilities for people to maintain and even to forge friendships at a distance, yet sites of schooling, work, and other physically copresent activities continue to be important spaces of friendship. As is also the case when marital relations end, the dissolution of friendships can have profound implications for personal geographies, reconfiguring the kinds of social spaces and networks that a person is able to or wishes to access.

The sheer diversity of relations to which the term “friend” can refer – ranging from a recent acquaintance to a lifelong confidante – poses a significant challenge to generalization of its geographies, as does contextual variability in the meaning and scope of “friendship.”

Although Western thinking and writing on friendship can be traced back to the classical scholarship of Aristotle and Plato, until recently the term had attracted only sporadic interest in geography and most other social science disciplines. In part, this lack reflects a tendency to trivialize the interpersonal and informal character of friendships. In sociology, for example, friendships have often been perceived as being of merely (inter)personal significance and, correspondingly, as having little in the way of wider social or structural consequence. Anthropology, meanwhile, has tended to focus on consanguinity (relation through kinship) and formalized affinity (e.g., relationship through marriage) to the neglect of non- or more-than-kin relationships, particularly in studies of traditional societies. The most established and sustained interest in work on friendship is to be found in social psychology. Interpersonal attraction, including but not limited to issues of friendship formation, has been analyzed by social psychologists since the 1960s; however, while geographical dimensions such as the significance of residential propinquity have often been incorporated into such studies, they have had little resonance in academic geography.

Growing interest in friendship among geographers has been bound up with a shifting evaluation of its significance in disciplines such as sociology and anthropology, where the concept has not traditionally been a focus of attention, rather than a reflection of any strengthening of geography’s engagement with social psychology. In particular, it has increasingly been recognized that friendship is shaped by and, in turn, (re)shapes wider conditions of social, economic, and political life. The rise to prominence of work on social capital, for example, not only highlights the importance of social networks for individual wellbeing, but has also meant that friendship and other informal social relations evoke alternative imaginings of social structure.
and organization. Highly influential sociological work has characterized social networks in terms of “status homophily,” meaning that friendships may actively reinforce and reproduce wider structural distinctions rather than merely reflect them. Conversely, it has been recognized that friendships can generate or sustain alternative subjectivities and hidden solidarities that challenge existing social and/or political formations. In her geographical ethnography of leaf-collecting work by girls in the Indian Himalayas, Jane Dyson (2010) showed how friendship was a medium for the reproduction of dominant gender norms and ideologies, but also allowed experimentation with identities that did not necessarily conform to adult expectations. In geographical terms, it is significant that Dyson emphasized the physical location of girls’ work and play – the forest as “a space figuratively and physically remote from the adult domain of the village” (2010, 493) – in her description of their irreverent and mischievous behavior. While sociologists have written about the “context” or “backcloth” of friendship formation, geographic scholarship is distinguished by its emphasis on “places and spaces of friendship” (Bowlby 2011, 612) and associated spatial practices.

While speaking to sociological debates about the social, political, and even structural significance of what was once considered to be “merely” personal, geographers’ interest in friendship also reflects disciplinary recognition of the intimate as important in its own right. Friendship is a means by which people across the world maintain intimate relations both proximate and, increasingly, at a distance. Examination of such intimate human geographies positions work on friendship in humanistic, feminist, and poststructuralist strands of the discipline, in contrast to the more positivist or behaviorist approaches conventionally favored by social psychologists. There is considerable potential for research connecting geographies of friendship to the wider growth of disciplinary interest in issues of human emotion and affect (Bunnell et al. 2012). More so than forms of intimacy associated with kinship or civic unions, friendship requires ongoing and reciprocal emotional work by virtue of not being based on blood ties or formal contracts. The production and reproduction of, for example, mutual trust, fondness, and emotional solidarities through friendship implies both geography and affect: the social geographies of embodied encounters and the performative aspects of friendships that have often been overlooked in categorical accounts of human identity (as being classed, racialized, sexed, gendered, etc.); and the precognitive and often inexpressible affective “substratum” or “bedrock” that circulates between and connects bodies. While existing geographical work has referred to “affective structures” (Valentine 2008) that underpin intimate ties, the term “structure” arguably sits rather uneasily in studies of the relational and fluid “doing” of friendships. The affective dimensions of friendship and its geographies thus form part of a wider questioning of the ontology of the social in human geography.

Although the inter- or transpersonal may be conceptualized in a variety of ways, voluntary relations such as friendships are widely understood to be assuming increased importance in contemporary life. The concept of friendship may be said to reconcile imaginings of late modernity in which individuals have historically unprecedented freedom to enter and leave relationships, on the one hand, with recognition of the continued significance of intimacy, on the other. Friendship is a form of intimacy that appears increasingly important in our urbanizing, mobile, and interconnected world. For purposes of review, it is possible to disentangle three main strands of geographical work that attend to friendship as part of historically shifting social
and spatial relations. The first has to do with geographies of intimacy in contemporary urban life. Geographers such as Nigel Thrift have highlighted the role of friendship and conviviality in “keeping cities resilient and caring” in the twenty-first century (Thrift 2005, 146). In the face of the supposed demise of sociality in alienated Western cities, Thrift considers the institution of friendship to be widespread and even “thriving.” Furthermore, in contrast to loving or romantic couple relationships, friendship, according to Thrift, offers a “light-touch” model of intimacy and compassion that is realistic and achievable today.

Second, friendship is arguably assuming greater social significance in the context of increasing human mobility and long-distance connections that characterize lives in the twenty-first century. Existing practices of connection, community, and intimacy that have been sustained primarily through geographies of stasis are thrown into flux under contemporary conditions of globalization and transnationalism. While this flux has implications for a range of intimate social relations, friendship has gained particular prominence in recent geographical research on migration and mobility. Katie Walsh’s work on British expatriates in Dubai, for example, begins by examining transient heterosexual intimacy but concludes by noting the prevalence of platonic (friendship) ties (Walsh 2007). The young migrant women in Walsh’s study often found (mainly same-sex) friendships to be more meaningful and dependable forms of intimate attachment than those with their sexual partners. Friendship may assume even greater importance for migrant populations at the other end of the socioeconomic spectrum, in the lives of men and women whose social geographies tend to be much more restricted, whose sexual relations are often policed and regulated, and who are unlikely to be able to afford to pay for professional help in dealing with personal or psychological problems. Friendship networks can serve as resources for such migrant men and women in terms ranging from day-to-day practical support (e.g., wayfinding in new environments) to emotional support and care (e.g., for those suffering from depression or struggling with alcohol addiction).

Third, and relatedly, friendship is bound up with shifting geographies of care. In addition to the ways in which migration and transnationalism have reconfigured geographies of caregiving among kin relations, sociological literature has made much of a wider shift from “families of fate” to more voluntarily formed and individualized “personal communities.” Geographer Sophie Bowlby (2011) has highlighted the need for greater attention to be given to the role of friendship in research on care, while resisting notions of a wholesale shift away from “traditional” family-based relations and associated caring practices. Empirical evidence for such a shift is inconclusive at best, with family and other kin members continuing to be important caregivers in at least some life stages, including in those Western contexts on which sociological generalizations have mostly been based. In addition, however, Bowlby (2011, 610–611) points to a blurring of the distinction between family and friends, including situations in which “some family relationships take on many of the attributes of friendships and a few friends may become more like ‘family members.’” The different combinations of friends and/or family implied in “caring for” and “caring about” people, and the ways in which these vary across life stages, make for a rich vein of potential future scholarship in human geography.

Running across all three of these intertwined strands of research are wider geographic issues concerning the significance of bodily copresence for friendships, the extent to which friendship practices and spatialities are being reconfigured
with advances in information and communications technologies, and whether extended possibilities for real-time interaction with distant others are making friendship more or less important compared to other forms of social relations. Do the “light-touch” forms of urban intimacy in which Nigel Thrift has expressed hope require face-to-face interaction, or can they be mediated through technologies of communication? Does messaging with mobile phones and Internet technologies negate the need for convivial spaces of association among migrant friends, or is the coordination of such association among their most important functions? What variation is there in the relative importance of friendship in forms of care that are provided in situ versus those provided at a distance through technologies of communication? Much more empirical work is needed to answer such questions in different social and spatial contexts. The prevailing consensus is that the maintenance of most friendships requires at least occasional copresence (Bowlby 2011), and highly influential research on the Internet-based social media network Facebook has shown that online “friends” mostly emerge from prior offline relations (Ellison, Steinfield, and Lampe 2007). The pace of technological change renders any such conclusions tentative at best, however, and emerging technologies of friendship may form the basis for a transdisciplinary field of research in itself. At the same time, it is important not to presume that technologies of friendship at a distance are new to a twenty-first-century world of high-speed Internet connectivity. Work by historical geographers has highlighted letter-writing and correspondence as technologies of friendship at the turn of the twentieth century (Thomas 2004). As that century progressed, letter-writing became a technology of long-distance social relations that extended beyond political or aristocratic elites. Through the emergence of phenomena such as “pen pals,” it became possible for “friendships” to develop in the space of correspondence between people who had never met face to face.

One established area of geographic research on friendship that is being impacted profoundly by technologies such as Facebook is geographies of children and young people. Not only is friendship conventionally assumed to be of particular importance to children and young people, but they are also often the most enthusiastic adopters of new social networking technologies. To what extent does access to specific technologies exacerbate the impact of socioeconomic status on children’s identities and the friendship groups in which they can participate? Does the use of such technologies mean that moving house or moving school no longer spells the end of friendship networks and associated senses of belonging? Existing research points to the importance of human mobility as a motivation for the formation and maintenance of online networks, with social network sites (SNS) often serving to offset the “friendsickness” associated with moves away from old friends (Ellison, Steinfield, and Lampe 2007). It is important to note, however, that work addressing friendship in geographies of children and young people has certainly not been entirely subsumed by issues of information and communication technologies. Helen Wilson’s (2013) recent ethnographic work on playground encounters, for example, extends established geographic themes concerning how, where, and with whom children forge friendships.

Work on children and young people is also a key field of geographic research in which the positive valence of friendship has been subjected to critical scrutiny. While friendship is generally considered a positive form of sociality, it also has a dark side. Existing work has noted how friendships among children and young people can be important sources of self-esteem and collective identity, but also that bullying and
the production of spaces of fear can be carried out through gangs of friends. Similarly, although friendships can provide opportunities for mutual learning and the broadening of children’s lived spatialities, expectations of conformity among friends can result in children being “led astray.” Such issues arguably apply to the lives of adults as well as to those of children and young people, although they are particularly noteworthy among the latter group precisely because of the way in which childhood friendships can shape opportunities into adulthood. Research that is more specifically concerned with friendship among adults, meanwhile, has included consideration of how the institution reinforces social and spatial inequalities. Although it is possible for friendships to form and be maintained across sociospatial difference, the tendency toward status homophily means that friendship is bound up with ways in which geographies of social inequality are reproduced. In addition, the power associated with trust between friends is clearly open to abuse and may be cast in terms of vulnerability. The informal labor recruitment industry that forms part of a wider migration industry from the Philippines, for example, rests in large part on appropriations – and misuse – of trust in friendships. Sallie Yea’s interviewees who migrated to Singapore for work opportunities as waitresses or hostesses had often been deceived by trusted friends about the nature and conditions of their employment, and this placed them in situations of debt bondage and lack of choice in performing sexual labor (see Bunnell et al. 2012).

Issues of interpersonal trust and power compel many geographers and other social scientists to reflect critically on friendships in their research practice. Clearly, this reflection applies to research that is not explicitly concerned with friendship in either empirical or theoretical terms (and it is worth noting that the terms “friend” and “friendship” appear more commonly in the preface or acknowledgments sections of books and journal articles than in the main body of social science publications); however, even limiting consideration here to scholarship in which friendship does feature as a central concept, the role of friends in carrying out one’s research varies according to methodological choices and preferences. In work that aims to make statistical generalizations about friendship patterns and preferences, or that seeks to map friendship patterns rather than examining the meanings of sociospatial linkages for the individuals concerned, friendships between researcher and researched are unlikely to assume methodological significance (indeed, they are perhaps unlikely to form at all). In other methodological approaches to geographic research, such as in forms of ethnography, not only is it likely that friendships will form during fieldwork, but friends and friendship can become central to the research process. In feminist and postcolonial geographic scholarship, friendships can be vital to research solidarities and collaborative praxis (Sangtin Writers and Nagar 2006).

Finally, there is a pressing need for forms of postcolonially inflected research that takes seriously contextual variation in the meanings and practices of friendship. Such variation expresses perhaps the most straightforward geography of friendship. To date, however, geographers and other scholars working on friendship have largely focused their attention on experiences in the West and the wider anglophone world. If sociological generalizations about shifting social relations, in particular, are notoriously Western- or even US-centric, then an important contribution that geographers can make is to provincialize such generalizations through comparative consideration of friendship in other regions of the world. Commonly cited attributes of friendlike relations in the West, such as the absence of instrumentality in friendship making
and maintenance, do not necessarily apply in other sociocultural contexts. Importantly, such contexts should not be viewed merely in terms of “traditional” cultures or as societies that are undergoing transformations (such as industrialization and urbanization) that have already occurred in the West. One example of how such teleological thinking can be avoided or even inverted concerns the role of friendship in informal networks of care. If these are expanding in many Western societies to compensate for diminishing levels of care provided by the state, perhaps emergent geographies of friendship and care in such contexts are becoming more like societies that have never had extensive state welfare provision. Thus, in addition to being important in their own right, studies of “friendship” in non-Western regions of the world – quotation marks here denote the fact that most people in most regions of the world would not use the English-language term – compel definitional rethinking of the concept and its wider social or structural dimensions. Similarly, critical engagement with terms in languages other than English could help to identify specific forms of friendship. While adjectival differentiations such as “best friend” or “true friend” have been noted in English, for example, other languages (such as Korean) have specific terms for forms of human relations that are conventionally given the generic label of “friendship” in English.

SEE ALSO: Affect; Community; Ecological footprint; Human geography; Modernity; Networks, social capital, and development; Place; Social capital

References


The frontier is one of the most anarchic, unstable, liberating, fertile, and, because of this bewildering ambiguity, also one of the most contested concepts in geography. Within classical border typology, a frontier is a particular kind of border and thus, in its most basic aspect, a criterion for spatial differentiation, both material and ideological. A primary and crucial distinction is the one made by the Polish geographer Ladis Kristof (1959) in his seminal work on the nature of boundaries and frontiers. He scrutinized frontiers through the revealing method of contrasting them to their most antagonizing opposite: unlike the boundary, which summons images of well-defined lines and limits aimed at fixing and disciplining bodies and minds, the frontier is a blurred area of “variable width” (see also Semple 1911, 208) whose moving contours remain beyond anyone’s control, open to all kinds of either friendly or violent encounters. While the boundary is a manifestation of the state’s “centripetal forces,” the frontier is not the imprint of a political project’s claim over space but a “phenomenon of the ‘facts of life,’” “a manifestation of centrifugal forces,” a decentralizing circumstance and thus a challenge to the nation-state (Kristof 1959, 270, 273).

The frontier is what lies “in front” and thus a world yet unknown (Kristof 1959, 269–270). As the vision of what lies ahead, the frontier is precisely the antithesis of a limit: a permanently open horizon; a world more firmly grounded in our imagination and aspirations than in reality — the world but not yet. The frontier is promise, hope, and the opportunity of forward-moving space, which are alluring enticements for the pioneer and the adventurer. A frontier hence suggests a zone of transition whereas a limit marks clear delimitations decreeing the end of something and the beginning of something else. Frontiers’ outer orientation “toward the outlying areas that are both a source of danger and a coveted prize” (Kristof 1959, 271) may be a remedy against the claustrophobia of the boundary: the “desire to distance oneself from the other in order to uphold the (fantasy of the) self during feelings of fear and anxiety” (van Houtum 2005, 677). The frontier stirs vibrating feelings that pulsate to the beat of foreign influences reverberating across space in waves of cultural shudders (Kristof 1959, 271; see also Ferguson and Raffestin 1986). Frontiers give rise to raptures of improvisation, spontaneity, and resourcefulness. In opposition to the promised immobility of the boundary, the frontier threatens with imminent change. While the boundary’s aim is to erect a self-fulfilled prophecy of clearly marked difference and incompatibility, the frontier’s confusion prevents the formation of the very categories that make such division clear-cut or even thinkable. Seen in this way, the frontier may be a conceptual antidote against pernicious expressions of ossified politico-territorial affiliations that find their epitome in the nation-state, as well...
as in the clearly defined identities it promotes and the neatly severed allegiances these identities inspire.

Although it is as geopolitically malleable as the boundary, the frontier is arguably a better representation of incessant flows and thus a more accurate conceptualization of the continuous transformations that are the concern of human geography. Despite the exact nation-state boundaries that make up the basic grid of our modern maps, by zooming from the dishonestly large scale of maps into the smaller scales of spatial interaction, it is hard not to feed the suspicion that cartographical borders would better illustrate the morphology of lived borders with the rough brushstrokes of a frontier than with the terribly executed portraits of linear boundaries (Febvre 1947, 204). Thus, although “For convenience’ sake, we adopt the abstraction of a boundary line … the reality behind this abstraction is the important thing in anthropo-geography” (Semple 1911, 205).

Their partnership with openness is thus a liberating aspect of frontiers: they endow eyes and feet with the feeling that they can wander aimlessly for days or months without finding a drastic change in the spatial-cultural landscape. Arguably, its imprecision and relentless space of possibility make the frontier’s morphology more exciting than the disciplining customs’ booths, passport controls, and waiting lines associated with the highways and airports of boundaries.

Frontiers are also present in the mesmerizing space-altering consequences of the body’s political implications. Markers of socioeconomic status such as skin color and phenotype; accents and language; religiosity and its behavioral manifestations; fashion, diet, manners, material possessions are sociological borders with spatial contexts and consequences. Their political interpretation may turn them into expected, manageable, or even desirable elements of socialization – for example, multiculturalism – or into obstacles for people’s spatial and social movement – for example, xenophobia or racism – and so transform a frontier’s embracement of inevitable variation into the boundary’s objection of difference (van Houtum 2010a). The political implications of thinking more in terms of boundaries than of frontiers are thus far-reaching, for frontiers animate the sort of imagination that boundaries may smother.

Yet, the free-spirited and bohemian interpretation of the frontier should not seduce us into ignoring the concept’s pitfalls. As with the terms “borders” and “boundaries,” which feature a Janus face (van Houtum 2010a), showing excluding closure on one side and resourceful opening on the other, the term “frontier” clearly has another side too. The frontier has been used to justify brutal expansionism, predatory colonization, and exploitative imperialism. The frontier’s inherent vagueness has dark drawbacks. The unsettledness and lack of rigor to establish the limits of space, culture, identity, and ultimately political space and emotional allegiances may give free rein to the worst tendencies, namely the colonial explorations and oppression brought by the experiences of colonization, African enslavement, and the world wars of the twentieth century.

Perhaps the theory that best illustrates the emancipatory and oppressive tension contained in the frontier is Jackson Turner’s highly influential work. In his collection of essays, The Frontier in American history (Turner 1921/1893), he repeatedly referred to the western American frontier as “an area of free land.” Turner’s could be either an oppressive or an emancipatory idea depending on whether you were an American pioneer in search of property, fortune, and religious freedom or the unlucky Native American, the demise of whose lands, community, and culture was necessary for the constitution of the
pioneer’s apparently “free” land. The tension between liberation and oppression contained in the concept of the frontier is related to another tension about its population density. Unlike Turner’s empty frontier, authors like Kristof considered liveliness a *sine qua non* characteristic of the frontier, which he thought of as a place swarming with people and which would be “inconceivable without frontiersmen” for “an ‘empty frontier’ would be merely a desert” (Kristof 1959, 272). The debates about whether the frontier is liberating or oppressive, empty or populous, show that, like any other concept in geography, the historical interpretations, morality, and political implications of the frontier are contested by power.

**Imperial frontiers**

As already hinted at, the term “frontier” has been often associated with empires seeking to legitimize their expansionism or universal ambitions (Kristof 1959, 271). In the frontier thesis of Jackson Turner, for instance, the western frontier of North America provided the space for a rugged individualism to prosper, and thus the basis of democracy and the American political ethos. The frontier was a metaphor for a steady movement away from the influence of Europe and thus the detachment necessary to justify the American exceptionalism Turner saw in America and its development.

A more critical gaze at the frontier thesis would, however, render a different analysis. The frontier in Turner’s sense provided the distinction between civilized and not civilized, inhabited or “empty,” wild land. The geographical imagination evoked by the idea of “free land” behind Turner’s “frontier thesis” glorified the creation of the American West by providing a morally digestible geopolitical narrative for what more impartially could be told as the forceful colonization of Native American territory.

Because of its neglect of the frontier’s meaning for Native Americans, Turner’s understanding of the moving and pioneering frontier across the North American West as the ethos of American democracy has acquired a disreputable rust. This, however, has not prevented the imperial logic underlying Turner’s frontier thesis from becoming widely adopted by political projects with expansionistic ambitions even up to this day. The geopolitical rationales that the United States, Israel, Russia, and even the European Union, for example, present as vindication for the seizure and control of their respective spheres of influence typically characterize their ambitioned territories as lands that lie “in front” and whose emptiness in physical or ideological terms requires – or at least justifies – the kind of management that would be unacceptable in populated lands – for example, drone strikes, armed invasions, fences. Awareness about the frontier’s colonizing pedigree is crucial for its historical understanding and political analysis. Frontiers have been associated with freedom as much as they have been associated with colonizing notions of people without history, without culture, and consequently with no right to space. Holding one or another passport can either increase or decrease the number of frontiers in people’s minds. The carrier of a privileged passport whose country is seen as part of a familiar world can travel unhindered and easily experience the sensation of a seamless world of frontiers as open spaces of possibility, whereas someone who, as a result of a misfortune dealt by the lottery of birth and cannot travel unhindered, may more likely experience a world of prejudice and obstacles in whose boundaries he or she is trapped (van Houtum 2010b). This dark side of imperial frontiers divides the world into geopolitical allies, enemies, and
buffer zones and has helped justify unjust wars and war crimes, ethnic cleansing, genocides, and the systematic oppression of peoples and cultures whose second-class humanity empires have deemed unworthy of owning the very lands they inhabit and even the very lives they embody.

Frontiers of imagination

Apart from its imperialistic notion, in a geopolitical sense the frontier is also a geographical imagination. Frontiers are as material as they are imaginary. Frontiers can be recognized and studied not only on the basis of their location but, perhaps more importantly, on the basis of their perception. Space is more than a physical location; it is also a political and an ideological position. There is a whole travel, touristic, and migratory industry for which these imaginary worlds matter. Also geopolitically, frontiers of imagination have far-reaching, life-altering consequences for the self and outer perception of entire nations. The Americas and Russia, for example, have been European frontiers as much as Europe has been theirs in each other’s cultural imaginary. Ukraine has been a frontier for both Europe and Russia, a vague transition zone between their mutual frontiers. The United States has been a frontier of economic opportunity for poor Mexicans searching for better conditions that their own country fails to offer them. Africa was the frontier for the expansion of the European mission civilisatrice during the nineteenth and twentieth centuries as much as Europe was a frontier of conquest for the medieval Arab caliphates that colonized its south. Scandinavia was a frontier of the unknown for the Mediterranean civilizations of antiquity as much as the Mediterranean was a frontier of plunder and trade for Nordic raiders during the Middle Ages – when Scandinavia had not even been detected or charted on maps.

The frontier is related to geographical exploration and a sense, an acknowledgment, of ignorance and wonder about what lies beyond the ecumene, that is, the “known world.” Since frontiers are characterized by a “certain width,” the question then is how much width these zones of transition can bear before the places they separate are imagined to be so far apart as to be noncontiguous. This idea of frontiers as moving zones of influence can and has been most dominantly applied in physical-geographical terms to denote spatial contiguity, in the sense of natural frontiers, but even more so in political geography to depict ideological, geopolitical, and culturally moving zones of power.

Yet, frontiers can sometimes be as wide as the Atlantic. The American continent was the first European periphery – in the sense that it was the first significant other from which the idea of Europe as a cultural and political unity distinguished itself when it first emerged in the sixteenth century (Wintle 1999) – and was considered as “difference within sameness” (Mignolo 2000, 58). The anthropocentric focus of geography and cartography has played a crucial role in the definition of the frontier. The frontier, being what lies beyond, only makes sense once a center has been defined. The frontier is thus culturally contextual and as such can be subordinated to particular kinds of knowledge, dependent on civilization and a manifestation of ethnocentrism.

Perhaps some of the most fascinating frontiers are those of a dislocated nature whose lack of geographical contiguity makes them among the most complex. They are best understood as what Michel Foucault called heterotopias, that is, a sort of “effectively enacted” utopia (Foucault 1986, 24): places within places whose atmospheres carry our senses into faraway locations.
They are factories of disconnected experiences whose atmosphere bears little or no relation at all to the contexts in which they are inserted. They snatch our emotions the moment we go past their gates into an ambience different from that we left behind.

Colonies, schools, hospitals, cafes, and brothels are examples of such frontiers. Certain means of transportation, such as trains and particularly boats, as can be seen in the refugee crisis of the 2010s, are among the most epitomical heterotopic frontiers:

the boat is a floating piece of space, a place without a place, that exists by itself, that is closed in on itself and at the same time is given over to the infinity of the sea and that, from port to port, from tack to tack, from brothel to brothel, it goes as far as the colonies in search of the most precious treasures they conceal in their gardens ... the boat has not only been for our civilization, from the sixteenth century until the present, the great instrument of economic development ... but has been simultaneously the greatest reserve of the imagination. The ship is the heterotopia par excellence. In civilizations without boats, dreams dry up, espionage takes the place of adventure, and the police takes the place of pirates. (Foucault 1986, 27)

The boat as a heterotopia is a frontier that, even if rarely conceptualized as such, has persistently awakened the fascination of geographers: “After the discovery of fire the next most important step in the progress of the human race was the invention of the boat. No other has had so far-reaching results” (Semple 1911, 332–333).

Since the boundary resulted from advancements in the techniques for territorial demarcation, the frontier is intimately related to the history of technology. The frontier is a historical concept whose disappearance from the map follows cartographical development, which is a testament of frontiers that have been and are no more – at least on maps and in geopolitical discourse. Map-making is the technique that has tried to put the taming leash of boundaries around frontiers. Medieval Muslim cartography is proof that frontiers, not boundaries, were once the borders represented on maps. Muslim cartography from the ninth to the fourteenth centuries shows boundaries between distinct political projects as “zones of transition” and of “uncertain sovereignty” rather than “sharply defined boundary lines” (Brauer 1995, 5–6).

The frontier belongs to a time whose technology did not allow for precise territorial demarcation, or perhaps to cultures whose environments, such as mighty indistinctive deserts, made the very notion of clearly defined boundaries pointless. Beginning in the fourteenth century, improvements in navigation and world travel by the Muslims, the Chinese, and, most influentially, the Europeans expanded the physical frontiers of each of these civilization’s known worlds and shook the very foundations of mental frontiers around the world. Perceptions of identity and
FRONTIERS

politics, race and religion, and even of humanity itself were altered by the revelation of uncharted lands and peoples.

It would be a mistake, however, to consider the frontier as a thing of the past and a relic of politico-territorial arrangements that did not have the technology to clearly demarcate the legal-territorial boundaries of their power. Not only is the European Union a political project that has found a way to maintain the national boundaries of its member states while opening frontiers between them, but its most powerful foreign policy tool has been the frontier-based geopolitical promise of enlargement. This propensity to expand has been the core of the European Union’s diplomatic and geopolitical strength; and, in itself as well as because of its destabilizing consequences in preventing the coagulation of an EU-wide identity, it remains a robust dam against the recurrent waves of nationalism.

In spite of geographical exploration and the amazing sophistication of instruments of navigation, frontiers have not disappeared. Many of the features and perceptions associated with frontiers remain. For tourists, the frontiers of either new (e.g., a road trip to an unknown destination) or mass-produced happenings (e.g., having one’s photo taken by the Eiffel Tower) are the coveted experiences of their travels; however, the ubiquitous and highly detailed worldwide cartography provided by global positioning systems (GPS), mobile phones, and the Internet may have anesthetized the sense of wonder and appeased the hunger for untrodden paths by providing the misleading impression that maps can capture the world and that there is nothing left to discover.

Since antiquity, concepts like terra nullius, terra incognita, and hic sunt dracones have denoted emptiness, a zone of transition between settled and unsettled, cultivated and uncultivated, civilized and uncivilized. Historically, the frontier has conjured the unknown. When the world was limited to the realm of the ecumene (the Greco-Roman known world), the military imaginary perceived the front as the unexplored danger, excitement, and possibility. The frontier is perhaps the most exciting of borders because it is always expectant. Frontiers still evoke a sense of both danger and adventure, glory and perdition.

Fortunately, the uncertainty of the frontier is still attainable for those who crave it. It is hard not to imagine the sense of wonder, spaciousness, controlled uneasiness, and exciting desperation when traveling to less familiar or unfamiliar places. It would be a mistake to believe that the full mapping of the world has brought about the end of geographical exploration (as Mackinder assumed in 1904). Frontiers as large, wide, empty spaces provoking fear, anxiety, enthusiasm, and whatnot have not disappeared. The possibility to imagine frontiers has been stifled with the advent of maps that create the misperception of a fully settled, fully bordered, fully inhabited, and fully controlled world. Yet, while their morphology has multiplied, frontiers still exist.

Frontiers of integration

Precisely because the term “frontier” is reserved for a not yet, a fearful yet tempting there, some authors see frontiers in terms of spatial integration. Much emphasis has been placed on borders of separation, and less (but growing) on borders of integration. Although the recognition of underlying domination structures within discourses has an emancipatory power, the perception of affinity has an indisputable emancipatory power. Much weight has been laid on coercion while bonds have been overseen. It is important to remember that frontiers are
integrating elements (Kristof 1959, 273). We recognize borders not only when we are confronted with difference and exclusion, but also when we come across unforeseen compatibility and inclusion. Whenever we unexpectedly find people with compatible or complementary beliefs, mores, phenotypes, architecture, food, and relationships to ours, we find frontiers of familiarity. When we find compatibility hidden behind an apparently different facade, when we recognize our endophenotype in someone in whom we did not expect to find ourselves, we feel suddenly rebordered into an abrupt familiarity. These are frontiers of inclusion. They are frontiers hidden in plain sight.

Seas are perhaps some of the most interesting among such eclipsed frontiers. In ancient cartography, the massive number of interactions that occurred between continents and civilizations was artistically depicted by traces of routes (such as in portolan charts), as well as by galleons and other kinds of ships, visually suggesting the migration, connection, and trade routes between merchant cities and between metropolises and their colonies. Today's satellite maps present seas as a mere background, thus dispossessing them of their historical and actual importance.

In the same vein, there are frontiers of transmission and translation, that is, contextually divergent practices and perceptions that nonetheless follow a similar spatial logic or entail an analogous sociological meaning. The illegal border crossing of migrants across both the Mediterranean and the Mexican–American border are a good example. The economic and living-standard disparities, as well as the imaginations of empowerment and emancipation that their destinations offer the migrants, speak of a shared geographical imaginary despite the migrants’ disparate heritages and locations. The treatment African and Mexican migrants receive, the dramas and suffering they experience, the prejudices of which they are the target, and the policy debates their movements arouse may be similar and subject to translation. Socioeconomic disadvantages among the black communities of America, as well as their plights as a minority and ways of coping with them, could be translated to the Turkish German communities of Berlin – a certain kind of protest music already does it. A translation implies an abstraction to identify commonalities. Yet, the uniqueness and diversity of border dramas makes any frontier translation and its validity inevitably controversial.

**Conclusion**

Precisely because of its conceptual openness as well as its conceptual relation to exploration, the frontier has been used dominantly as an oppressive term. Increasingly, however, the term’s potential for constructing a more just interaction among a humanity that has become inexorably and irreversibly globalized has exposed the other side of the frontier’s Janus-faced nature (van Houtum 2010a). As naturally curious explorers of the world and its relentlessly changing spatial and cultural sceneries, geographers have a hard time resisting this double-sided, discovery-arousing charm of the frontier. It is hard not to be conceptually allured by the range of possibilities evoked by the untamable attractiveness of the frontier’s drama-laden history and its constantly mutating meanings. This attraction should be a motive for rejoicing among geographers, for a world of enthralling fields of joyful academic discovery lie in front of the further conceptualization, discussion, and exploration of the frontier.

**SEE ALSO:** Borders, boundaries, and borderlands; Cartographic design; Colonialism,
decolonization, and neocolonialism; Critical geography; Empire; Geopolitics; Governmentality; Imaginative geographies; Imperialism; Political geography; Postcolonial geographies; Territory and territoriality

References


Further reading

Fuzzy classification and reasoning

Suzana Dragicevic
Simon Fraser University, Canada

Geographic entities from the real world are commonly represented as vector or raster geospatial data models in computer databases. In the vector data model, reality is represented by the geometric primitives of points, lines, and polygons. In this case, feature boundary and identification are well defined. In the raster data model, reality is represented by geometric area primitives such as squares or hexagons. In this case, feature boundary and identification are not always well defined. Moreover, while it is possible to transform vector data into raster data and vice versa, there is a penalty incurred in precision and accuracy of locations and attributes.

This representation of geographic entities on maps was traditionally based on Boolean logic and crisp sets that use binary terms to classify an element to either one or other set or class (Hailperin 1986). This logic has been transposed to geographic information systems (GIS) and geospatial databases that are suited to use Boolean logic and bivalent/crisp representation of object boundaries in digital GIS data layers. However, many geographic phenomena are difficult to represent in a crisp manner. For example, the boundary between forested areas and grassland, or between water bodies such as coastline, lakes, rivers, and shores which change gradually (Burrough and Frank 1996); in addition, these boundaries change over time. When classifying soils or forest types, the sharp boundary representation does not capture the real-world situation (McBratney and de Gruijter 1992; Brown 1998). Moreover, when remote sensing images are used to extract objects, several features on the Earth’s surface are often captured within a single pixel or raster cell, making it difficult to assign a single class corresponding to only one object (Wang 1990). The problem of bivalent representation has been addressed with the theory of fuzzy sets and reasoning developed by Zadeh (1965). Consequently, fuzzy reasoning has been introduced into geography, GIS theory, and spatial databases (Robinson 1988) to replace the traditional Boolean logic and area-class maps representations (Mark and Csillag 1989), and to represent temporal geographic events that change their physical characteristics or attribute descriptions.

Fuzzy set theory is a general specification of classical set theory. It is based on fuzzy membership functions and the gradual values of membership of an element to one or several sets or classes. The mathematical description of a fuzzy set \( A \) is given by:

\[
A = \{ (x, \mu_A(x)) \mid x \in A, \mu_A(x) \in [0, 1] \}
\]

The quantity \( \mu_A(x) \) is called the membership function and it specifies the degree to which an element \( x \) belongs to the fuzzy set \( A \). The elements have continuous membership values between 0 and 1 such that the value 0 means not belonging to a particular class or a set and 1 represents completely belonging to a class or set. The way an element is classified is based on the characteristic of the fuzzy membership function. Fuzzy membership functions can have various forms, such as linear, triangular,
FUZZY CLASSIFICATION AND REASONING

Figure 1 Example of a triangular fuzzy membership function.

Fuzzy set theory later led to possibility theory, which extends the theory of probability (Zadeh 1978), giving the capability to address the uncertainties raised by the problems of crisp classifications and Boolean reasoning. Fuzzy reasoning is based on fuzzy sets theory and enhanced by artificial intelligence and various information processing areas, such as logic, graph theory, topology, and optimization (Pappis and Siettos 2005). Fuzzy sets and their use in GIS have been addressed in detail by Robinson (2003).

Fuzzy set theory and reasoning are used for a variety of spatial applications that include the definition of a class or a cell state for positional or attribute uncertainty. Fuzzy representation of geographical boundaries and geographic objects are discussed in Burrough and Frank 1996; Wang and Hall 1996; Cheng, Molenaar, and Lin 2001. Examples of the spatial application of fuzzy logic have been widespread and include areas such as: forestry (Lowell 1994), soil science (McBratney and Odeh 1997; Burrough 1989), climate classification (McBratney and Moore 1985), land evaluation and suitability (Sui 1992; Jiang and Eastman 2000; Prato 2009), remote sensing (Wang 1990; Fisher and Pathirana 1994; Foody 1996), landscape ecology (Salski and Kandzia 1993; Marsili-Libelli 1989; Fisher et al. 2006; Foody 1986), landslides (Muthu et al. 2008), geology (An, Moon, and Renz 1991; Luo and Dimitrakopoulos 2003; MacMillian et al. 2000), and spatial interpolation (Bogardi and Kelly 1990; Dragicevic 2005). In addition, handling the temporal component in GIS databases using fuzzy sets has been reported by Dragicevic and Marceau (2000a). These resulted in the use of fuzzy reasoning for modeling dynamic spatiotemporal processes (Dragicevic and Marceau 2000b; Dragicevic 2004a) and the development of fuzzy cellular automata approaches for representing land-use change and urban growth (Wu 1998; Liu and Phinn 2003; Dragicevic 2004b; Al-Ahmadi et al. 2009), propagation of forest fires (Mraz, Zimic, and Virant 1999), forest insect infestation (Bone, Dragicevic, and Roberts 2006) or invasive species (Dragicevic 2010). Fuzzy reasoning was applied to agent-based modeling to mimic the decision strategies that agents, in this case squirrels, implement in order to move in the landscape as part of their food searching process (Graniero and Robinson 2006). Fuzzy kappa statistics and measures have been developed to explain and compare map patterns arising from complex spatiotemporal processes (Power, Simms, and White 2001; Hagen 2003; White 2006; Hagen-Zanker 2009).

Fuzzy classification methods

The goal of classification is to decrease the dimensional complexity of a dataset by grouping statistically similar values into a smaller number of classes or categories. In the commonly used crisp classification method there is a one-to-one correspondence between the measured value and the class into which it is assigned. Class
membership values can be either “1” for complete belonging to a class or “0” for not belonging.

In the case of fuzzy classification, a measured value can have membership in various classes and with different degrees. Fuzzy classification groups values into a fuzzy set where the membership function is defined by the truth value of a fuzzy membership function. In crisp classification, a class is a set of values with specific properties and every measured value that satisfies those properties is a member of the class. The classification algorithm or rule then evaluates the measured values and determines which class they belong to. However, for fuzzy classification there are some modifications to this generality. A detailed explanation of fuzzy classification methods has been presented by Robinson (2008). Ultimately, the fuzzy membership function defines how an element belongs to one or several sets and thus gets classified to a particular class. The key aspect of fuzzy classification is the way in which one or several membership functions are specified to describe the problem, and so can be subdivided into specification approaches that use prior knowledge or those that are data driven.

The fuzzy classification with prior knowledge relies on the knowledge of domain experts to construct fuzzy membership functions that determine the membership values. Aggregation (Burrough and McDonnell 1998) or rule-based methods (Cannon et al. 1986) allow fuzzy membership functions to be combined in specific ways so that an output map can be created with the fuzzy membership values that describe a particular problem. Figure 2 presents an example depicting the stages of generating an urban land suitability map for potential new urban residential development. The analysis is based on three fuzzy membership functions representing the criteria of elevation, proximity to schools, and proximity to roads with their corresponding fuzzy maps. The first fuzzy function and derived map describe the suitability of the land based on the elevation factor. From expert opinion (such as urban planners, urban developers, urban policymakers), the elevations from 0 to 50 m are considered the most suitable for urban development, and thus have the membership value of 1 (purple on the map). Then the elevation value decreases linearly until it reaches 1000 m (red to dark blue on the map). From 1000 m, the land is considered to be unsuitable for any further urban development thus the membership function is 0 (black on the map). Similarly, the second and third functions describe the best distances to schools (500–1500 m) and to roads (100–500 m) where the fuzzy membership function will indicate the value of 1 as the most suitable. The overall fuzzy suitability map (the top right map on Figure 2) for urban land use is generated by combining the individual map layers using the GIS overlay operator. This overall map provides the combined grades of membership function indicating the level of suitability for new urban development at any locations in the Metro Vancouver area based on the three suitability criteria (i.e., fuzzy functions of elevation, proximity to schools, and proximity to roads). The most suitable and least suitable locations for new urban development are presented as purple and black, respectively, on the map.

When expert opinion and knowledge are not available, then data-driven methods can be used (Luo and Dimitrakopoulos 2003). Various techniques can be used, such as fuzzy k-mean clustering methods (McBratney and de Gruijter 1992; Wang 1990; Cannon et al. 1986) or fuzzy neural networks inference systems to guide fuzzy classification (Foody and Boyd 1999). These methods depend on available datasets to derive the rules prior to the classification procedure.
Conclusion

There are many advantages of using fuzzy sets and reasoning in spatial applications and their widespread use in the GIScience community is strong confirmation of their relevance and usefulness. Exact classification methods lose information and increase errors while fuzzy approaches can address the complexity of the typical geographical and environmental data that change gradually over space and time. However, the main disadvantages of fuzzy approaches are the reliance on the choices of the fuzzy membership functions, mainly by expert opinion, as well as the choice of the parameters for the clustering or other data-driven approaches used for fuzzy classification. In addition, the choice of the de-fuzzification technique to classify the output fuzzy function into meaningful categories can be somewhat subjective. Nevertheless, fuzzy sets and reasoning are valuable analytical tools for GIS and geographical analysis.

SEE ALSO: Cellular automata; Cluster detection; Data structure, raster; Data structure, vector; Geocomputation; Geographic information science; Geographic information
system; Interpolation: kriging; Land change science; Landscape; Representation; Space; Spatial analysis; Spatial database; Spatiality; Uncertainty

References


FUZZY CLASSIFICATION AND REASONING


Gated communities

David Wilson
University of Illinois at Urbana–Champaign, USA

Current gated communities across America, Europe, and beyond are isolated residential enclaves designed to be exclusive, exclusionary, and physically detached from the immediate city or community. To achieve this status, gated communities typically contain controlled entrances, central open courts or streets to enable tenants to perform a maximum amount of traffic surveillance, and class-based meanings embedded in architecture to accentuate who should and who should not be at these microsites. They are also often marked by closed perimeters of walls, fences, and barricades. These communities can vary in size from one or two buildings to coordinated blocks of similar housing and internal retail and recreational facilities. For smaller communities, carefully monitored parks and common areas are common. For larger communities, there are often ample facilities and social activities for residents to stay within the community for most daily activities. Such common interest developments are proliferating across the globe today, fueled by the increased desire of the affluent to separate themselves from other segments of the population.

Economic segregation, of course, is not new. A multiplicity of measures, from zoning to housing covenants to racial steering, has existed for decades across the world to ensure class exclusivity. But gated communities go further, in particular privatizing public space and deepening patterns of social surveillance, in taking segregation to a greater level.

This balkanizing of people into compartmentalized spaces has enormous policy consequences. With people isolating themselves from active social contact with others, understanding of others and recognition of city and community diversity break down. The traditional notions of social contact and class intermixing that have been at least a professed ideal in many cities and communities disappear. As Blakely and Snyder (1997) ask, what is the measure of nationhood when the divisions between neighborhoods require armed patrols and electric fencing to keep out fellow citizens? In their view, when social space is dramatically splintered and the community of responsibility stops at gated community entrances, society gives up too much, and we should not casually disregard this new development. In short, can any community, city, region, or nation fulfill its social contract in the absence of social contact?

SEE ALSO: Residential segregation

References

Further reading


Gender and development

Ann M. Oberhauser
Iowa State University, USA

Gender identities and women’s positions in society represent critical dimensions of development theory and practice. In recent decades, international organizations, policymakers, scholars, and activists have increasingly recognized how inequality in the development process contributes to issues such as low literacy rates among women, gender-based violence, the exploitation of women in low-wage jobs, and underrepresentation of women in leadership positions. These conditions affect the progress not only of individuals and communities but of entire regions of the world. Critical perspectives in the gender and development arena, however, often question whether or not the resources and energy that development institutions and policy have invested to advance women and achieve gender equality have made significant impacts on the overall development of countries, especially those from the Global South.

This discussion outlines six themes that highlight major theories and approaches to gender and development, as well as practical and applied aspects of this area of study. The first theme examines how women’s and gender issues in the development framework gained global attention through efforts such as the establishment of the United Nations Commission on Women in 1946 and the UN Decade for Women. The field shifted from an approach that essentialized the category of “woman” to frameworks that examine gender and its intersections with other social categories such as race and ethnicity, class, caste, age, and sexuality (Visvanathan et al. 2011). This shift laid the foundation for a movement that is grounded in voices and perspectives from developing countries and that challenges patriarchal and Western biases in development theory and practice. Furthermore, recent work in postcolonial and transnational feminism critiques the hegemonic discourses and inequities that have been historically constructed in both colonial and capitalist systems.

Economic globalization is the focus of the second theme on women and gender in the development process. Women and men are impacted differently by the global economy, and thus globalization has gendered implications for livelihoods and economic strategies. The third theme outlines local organizing and empowerment as the root of grassroots and community-based initiatives in development work. The discussion includes analyses of the institutions, alliances, and livelihood strategies that constitute women’s efforts and often embody their gendered roles. The fourth theme highlighted later in the entry examines gendered aspects of national and global health issues that impede development and impact families and communities. Gender plays an important role in the overall health of a society, partly due to women’s responsibilities in the areas of caretaking, family nutrition, and reproduction. Of note in this discussion is growing attention to sanitation, and especially access to toilets and clean water in developing regions.

The dynamic and gendered relations between nature, the environment, and development are the focus of the fifth theme. Nature and the environment are often situated at the core of the development process, and therefore gendered access to and control over natural resources and
the environment contribute to inequality in development. The discussion addresses sustainability, feminist political ecology, and climate change in the context of gender and development. The conclusion addresses future directions in the field of gender and development that include information technology, global security, and sexuality.

Conceptual approaches to gender and development

Changing perspectives on women, gender, and development have created a dynamic and somewhat contested field in geography and related disciplines. The shift from a focus on women in the development process to gender and other social dynamics is evident in the conceptual paradigms as well as the institutions within which the field of development is based. Initial efforts to address women and inequality in the 1970s adopted a welfare approach with an explicit focus on women in their reproductive roles as wives and mothers. Specifically, women in development (WID) draws from the modernization framework which was modeled after Global North capitalist theories of development. Ester Boserup’s seminal work revealed that not all people benefited equally from development efforts which often had negative impacts on women. This period also saw global efforts to mobilize and draw attention to the status of women in developing regions through initiatives such as the UN Decades for Women, which began in 1975 with the Nairobi Conference.

Subsequent approaches to development critiqued both the overarching focus on women and the Western bias in theorizing women and gendered aspects of this field. The women and development (WAD) framework was one such approach and drew from Marxist critiques of capitalism to examine multiple aspects of women’s productive and reproductive labor in the context of international capitalism and unequal class relations (Kabeer 1994). This expanded view of women’s roles in society coincided with an emphasis on gender relations at multiple scales of the household, community, and workplace. Furthermore, gendered approaches to development allowed a much greater awareness of how different women experience gender. These approaches highlight women’s active engagement in change that is meant to disrupt power relations between men and women and to enhance women’s legal rights in areas such as inheritance and land reform. Moreover, Moser’s (1993) work emphasizes gender planning and policy by articulating two aspects of empowerment in the development process. First, strategic needs such as legal rights, equal wages, and domestic violence relating to women’s subordinate position in society are a means of addressing gender inequality. Second, practical gender needs relate to immediate necessities such as water, health care, and employment. Further efforts to situate development in regions most affected by the development process arose with initiatives such as the Development Alternatives with Women for a New Era (DAWN) in the mid-1980s. This organization is a feminist research network focused on widespread education and consciousness-raising to empower women in the Global South (Sen and Grown 1987).

The focus on women’s experiences in developing regions is also illustrated by the women, environment, and development (WED) approach, which addresses unequal access to and use of the environment. In particular, ecofeminism draws parallels between men’s control over nature and over women (Mies and Shiva 1994). These views of environmental and patriarchal exploitation are grounded in critiques of Western science and masculinist models of industrialization. In contrast to the ideologically based ecofeminist
approach, Agarwal (1992) developed feminist environmentalism as a way to emphasize more material approaches to the link between gender and the environment. This perspective provided the basis for critical activism and community efforts aimed at sustainable development in fields such as forestry and agriculture.

More recent theoretical approaches in this field include ongoing challenges to the Western bias in development discourse by women and feminists from developing regions. For example, postcolonial feminists critique Western models of development, patriarchy, and feminism that ignore the voices of Global South women or essentialize women as a homogeneous category. In *Third World Women and the Politics of Feminism*, Mohanty, Russo, and Torres (1991) emphasize the role of colonialism and hegemonic discourses that discriminate against women and marginalized voices from developing countries. They cite Western feminist representations of Global South women as universalizing and largely derogatory. Overall, theories and approaches in the field of gender and development have dramatically shifted since the mid-twentieth century from a focus on the status of women in traditional reproductive roles to a broader analysis of gender and other social identities, inequality, and power in diverse geographical contexts.

**Gender and economic globalization**

The approaches to gender and development examined so far are grounded in several themes that have shaped the advancement of this field. The first theme, economic globalization, underlies many aspects of gender and development theory and practice. The erosion of social services under neoliberal capitalism, the concentration of women in low-wage manufacturing, and gender bias in agricultural reform are some of the processes that illustrate the connection between globalization and gender in contemporary society. According to Beneria (2003, 77), gender and globalization are “shaped by the interaction between these forces and the different ways through which gender constructions have been reconstituted during the past three decades.” In particular, neoliberal globalization translates to the institution of laissez-faire economic strategies that promote privatization, free market forces, and reduced government roles in many areas of the world. In numerous cases, these strategies are the results of structural adjustment programs implemented by international financial and development organizations such as the World Bank and the International Monetary Fund (Beneria 2003). The co-constitution of structural adjustment and gender is evident as women experience higher rates of poverty than men, unequal access to well-paid and secure employment, and increased vulnerability in times of economic cutback. Given the gendered divisions of labor in the reproductive arena, women are often at a disadvantage as governments decrease support for social services such as education, health care, and child care (Momsen 2010).

Economic globalization also expands employment opportunities in many developing countries, as capital invests in areas where costs of production are considerably lower than in the Global North. Women form a disproportionate share of the workforce in low-paid and insecure jobs because of gendered practices that undervalue and exploit women’s labor (Runyan and Peterson 2014). While there are negative aspects of women’s employment in what Elson and Pearson (1981) refer to as the “global assembly line,” many believe that these jobs benefit women by expanding their income generation and independence. Kabeer (1994), for example, argues that women have increased autonomy and bargaining power as a result of
their engagement in waged labor. Work in the informal sector is also impacted by the rise of globalization in both industry and the service sector, which employ women in highly marginal and exploitative conditions (Beneria 2003).

In addition, the intersection of globalization, gender, and development encompasses many aspects of women’s labor and rights within the agricultural sector. Boserup’s work on women in agriculture, Agarwal’s (1994) research on gender and land rights in South Asia, and studies about the impacts of neoliberal policies on women and rice production in the Gambia (Carney 1992) demonstrate the gendered nature of agriculture and agrarian transformation in developing regions. These studies highlight the social aspects of what constitutes women’s and men’s labor in agriculture, as well as gender differences in the rights to land and other rural assets. Moreover, women’s involvement in subsistence production is essential to rural households in times of fluctuating income, climate change, and political conflict.

Finally, globalization is the focus of critical discourse surrounding the body and gendered violence in areas where foreign capital has invested in the manufacturing and service sectors. Wright (2006) develops a compelling story of women in the border industries of northern Mexico who are subject to violence as a consequence of gendered, racialized, and ethnicized discourses of female workers. This research demonstrates that the Mexican and US political economies construct the notion of disposable bodies among poor women in this area. Similarly, Gibson-Graham’s (1996) work on feminist political economy shows that gendered constructions of women in the hegemonic discourse of globalization are related to cultural and economic identities. They argue for an anticapitalist imaginary that represents a politics of economic transformation with an array of diverse economies at multiple scales.

**Community-based development**

Another theme in the field of gender and development focuses on local and community-based models of development designed to empower marginalized women. These models draw from gendered dimensions of economic and social empowerment that rely heavily on women’s active engagement with, and input to, the development process. Two areas of community-based economic organizing and empowerment are microfinance and microenterprise initiatives. Microfinance models such as the Grameen Bank in Bangladesh exemplify initiatives to raise people, and especially women, out of poverty in developing countries. The small-scale savings and investment programs are often successful, partly because women have higher rates of repayment than men, and have been replicated in many developing countries.

Additionally, microenterprise and community-based livelihood strategies are the focus of considerable research in many parts of the developing world. Oberhauser and Yeboah (2011), for example, argue that women’s roles, and especially their access to assets and other resources, shape their livelihood strategies in West Africa. Many women have developed microenterprises and small-scale businesses in both urban and rural areas as a means of supporting households (Figure 1).

These livelihoods in turn inform policymakers and development workers about effective and appropriate support for women’s empowerment strategies. Critics of microfinance efforts by global and national institutions, however, cite the potentially exploitative nature of this model, where women are forced into repaying high-interest loans in order to invest in these
small-scale activities. According to some critics, women are also subject to strict timetables for repayment of these loans and are thus at risk of increasing their debt load.

Other areas of community-based development that incorporate gender identities are the cross-border practices and collaboration situated within transnational feminism. Alexander and Mohanty (2010, 40) describe this approach as solidarity work (albeit tense and uneven) that is “brought into ideological and geographic proximity with one another in ways that produce connectivity and inter-subjectivity.” Research in this area involves praxis or a combination of practice and theory that destabilizes the binaries often established by the academy on one hand and community-based organizations on the other. For example, Pratt’s (2012) work with Filipina migrant laborers in Canada, and Peake and de Souza’s (2010) research with Red Thread on violence against women and community organizing in Guyana, are illustrations of transnational praxis. Additionally, India’s Self Employed Women’s Association (SEWA) is a group that has organized women in informal and marginalized work of all kinds in a feminist trade union model. Thus, local organizing and collective action across national and transnational boundaries serve a critical role in advancing gender equality in a variety of contexts.

Gender and health

The health and wellbeing of people in developing countries is a growing area of emphasis in the gender and development field. International and intergovernmental organizations such as the United Nations and the World Bank, along with nongovernmental organizations such as Médecins Sans Frontières (Doctors without Borders), have implemented extensive initiatives and invested considerable resources in order to help eliminate diseases that continue to plague developing countries. Global indicators and comparisons of fertility rates, life expectancy, and infant mortality are commonly used to identify gender inequality and often relate to other indices of wellbeing. According to the Earth Policy Institute, for example, high fertility rates are linked to low levels of education among girls (Fitzgerald Reading 2011). Figure 2 illustrates that girls and women with higher levels of education tend to have fewer children, which makes it easier for them to engage in paid work and to make choices that will improve their overall wellbeing (UNDP 2013). In developing countries, however, girls are less likely than boys to be in school, partly because of the cost of enrollment and expectations that girls contribute more to families by working at home. These gendered aspects of health-related issues have direct outcomes for the social and economic wellbeing of these societies.

Gender also plays a role in the provision of health care in families and communities. Women are often responsible for prevention
Figure 2  Female secondary education and total fertility rates (Fitzgerald Reading 2011). Reproduced from Earth Policy Institute.

and for taking care of family members, and therefore shoulder a disproportionate amount of the emotional and physical burden. The rise of HIV/AIDS in developing countries highlights this gendered aspect of health care. Women are often primarily responsible for care of HIV-infected family members, which impedes their economic activities and increases their risk of contracting the virus (Schoepf 1997). The historical legacy of male migrant labor systems in regions such as sub-Saharan Africa has affected the transmission of this disease, especially from urban areas to rural villages and households. The high incidence of HIV/AIDS in many developing regions also contributes to labor shortages due to poor health and high morbidity among the working-age population (UNAIDS 2013). Finally, females tend to have a lower but more rapid increase in their rate of HIV/AIDS infection than men in many regions, partly because of risky sexual behavior between men and women, such as that associated with the sex industry in these regions (Momsen 2010).

Another topic that has brought attention to health issues in developing countries is research into sanitation and toilets. The spread of disease and the overall health of a population are closely linked to sanitary facilities such as toilets; therefore strategies to improve health must involve the provision of these services. Many researchers examine the rise of neoliberal policies and specifically cutbacks in public services as negatively impacting the health of impoverished people. Furthermore, Gershenson and Penner (2009) show that access to and use of public toilets is gendered, as women tend to value the availability and quality of these services more than men.

The final health-related topic in this discussion is gender-based violence and its impact on development. Interpersonal and self-inflicted violence
tend to reflect power dynamics and coercion that contribute to injury, death, or psychological harm. Interpersonal violence, which includes family and partner harm that is often embedded in gendered relations, is the leading cause of death in persons aged 15 to 44 years (Krug et al. 2002). Gender-based violence is considered a human rights violation and is gaining global attention, with recent incidents related to what some refer to as the “rape culture” in India and “honor killings” of women in some South Asian countries. Cultural patterns that tolerate domestic violence often make such violence difficult to report and consequently lead to the widespread neglect and physical and sexual abuse of not only women but also children. The World Health Organization has extended the negative impact of violence to national economies that suffer from losses to health care, law enforcement, and worker productivity (Krug et al. 2002).

**Women, gender, and the environment**

Access to and control over natural resources and the environment have a long history in the field of gender and development. In many cases, women don’t have equal access to or use of energy resources, food, or raw materials, and have therefore organized around these issues at both the local and international scales to expand awareness of and attention to gender and the environment. The Earth Summit in Rio de Janeiro in 1992 was one global forum that launched efforts to mobilize groups around topics that included gendered aspects of biodiversity, sustainability, and climate change (Momsen 2010). It led, for example, to the creation of the Women’s Environment & Development Organization (WEDO) as an international network of women’s environmental groups.

Subsequent global forums such as the Rio + 20 United Nations Conference on Sustainable Development in 2012 reflect growing concern about effective ways to reduce the impact of environmental degradation on impoverished people in developing countries, and to produce a more sustainable and equitable use of resources. In addition to these efforts, women are active at the local scale organizing against the destruction of forests: for example, the Chipko Movement in India (Shiva 1988), the Green Belt Movement in Kenya (Momsen 2010), and action against water contamination in Bangladesh (Sultana 2009). Critical feminist analyses of gender and the environment examine how gendered subjectivities “influence the responsibilities, roles, rights and norms that constitute the relations that men and women have to natural resources” (Sultana 2009, 427). This increased focus on the gendered aspects of nature–society relations is partly due to the contributions of feminist political ecology (Rocheleau, Thomas-Slayter, and Wangari 1996). According to this approach, gender, as well as other social and economic power relations in society, influences access to and control over resources. Both activists and scholars in this field make a point of grounding their research in a local context.

The field of environment and development also includes growing attention to climate change, and specifically increased variability in rainfall and temperatures which leads to flooding and drought, rising sea levels, and erratic weather events (Momsen 2010). These climate fluctuations and their unpredictability have consequences for livelihoods, food security, the spread of disease, and infrastructure. Some scholars and activists argue that gender is a central concern in this field because women are more vulnerable to the impacts of climate change, especially given the gender divisions of labor in agricultural production and gathering resources such as water and fuel wood (Dankelman 2010).
Additionally, a gender-sensitive approach to mitigating the negative impacts of climate change applies to the implementation of agricultural practices that include planting drought- and flood-resistant crops above rising flood waters in coastal and low-lying areas. Thus numerous aspects of nature–society relations, environment, and development are closely intertwined with gender roles and access to resources.

Future directions of gender and development

Looking forward, the field of gender and development will build on research, policies, and activism that have increased awareness of and improved conditions for many women and girls in the developing world. Critical issues in this field, however, highlight persistent inequalities in social, economic, and political arenas that reflect institutional biases concerning gender and other social axes of power. This discussion concludes with analyses of three issues that will most likely shape the future of gender and development – information and communication technology, global security, and sexuality. These topics are by no means exhaustive, but reflect some of the ways that feminist approaches to development will advance gender-sensitive and other forms of equality and empowerment.

The growing use of information and communication technology (ICT) has the potential to impact economic activities for women at all scales and in a variety of sectors. For example, increased access to ICT, especially through the use of mobile phones in many developing regions, reflects and changes gender dynamics in the household and the workplace. This expanded access entails not only job creation but also efforts to organize women across social and spatial boundaries through networks and support groups. Furthermore, widespread access to social media and other forms of information is a means to diffuse consumerism and cultural practices. However, the rapid growth of ICT also has the potential to maintain social and economic inequalities in terms of access to resources and skills needed to further women’s advancement in society (Asiedu 2012). Negative aspects of what some refer to as digital inequalities expand beyond spatial divides to include diverse socio-economic divisions such as gender, class, caste, and race/ethnicity. These inequalities are evident in the disparities between the ability of countries to invest in infrastructure and communications. Growing attention to the role of ICT in the development fields focuses on digital inequalities as a more holistic way of examining the complexities of the use of digital technologies.

Second, militarization and security issues are gaining attention in the global political and economic arena, with widespread implications for women and gender roles in society (Enloe 2007). Terrorism, military intervention, and despotic governments construct discourses of resistance to power and domination that are embedded in racial, ethnic, and gendered forms of repression and marginalization. For example, certain masculinist identities of Middle Eastern men are depicted by Western media in ways that reinforce stereotypes of extremism and violence. Likewise, the post-9/11 world has impacted transnational migration with mechanisms of exclusion and gender inequality (Marchand 2009). Feminist postcolonial analyses of the “war on terror” effectively situate these gendered representations within sociospatial inequalities and uneven development in the Global North and South.

Third, axes of difference based on sexuality draw attention to a form of discrimination and
harassment that is constructed through histories of colonial and postcolonial relations. Homophobic laws and attitudes are evident in certain areas of the Global South, with sometimes brutal forms of marginalization. Growing mobilization against these attitudes and practices often situates the lesbian, gay, bisexual, transgender, and queer (LGBTQ) community in alliance with postcolonial feminism in order to advocate for equality and social justice. For example, queer theorists argue that many homophobic practices stem from colonial attitudes such as British laws in colonial India that made nonheterosexual practices illegal. These histories of sexual norms often counter claims that homosexuality is a practice imported by Western ideals and norms. Thus, feminism and sexuality provide a framework for rethinking power relations and identities in gender and development.

In conclusion, pressing issues in gender and development require feminist perspectives and gender-based analyses to further understand inequality and uneven spatial patterns in the developing world. Public policy, scholarship, and activism must draw from rigorous research and grounded practice in this field to establish a socially just and sustainable world.

SEE ALSO: Feminist geography; Feminist political ecology; Postcolonial geographies

References


Further reading

Gender, work, and employment

Helen Jarvis
Newcastle University, UK

While the definition of employment is well grounded in labor process theory and law, what counts as “work” is socially and culturally constructed. How we talk about women and men and the work they perform, and how this discourse appears to judge and prescribe “proper” behavior in a given context, have to be understood through social norms and conventions of language as much as through laws and state regulation. The market for labor is also rooted in socially constructed ideas, such as those of money, capital, and trade: an individual sells his or her time, sweat-effort, and skills in exchange for payment. Yet, inasmuch as these earnings are taxed and recorded as a barometer of national competitiveness, in the form of gross domestic product (GDP), paid employment tends to be recognized in a way that informal and unpaid activities of work are not. Efforts to critique and transform systematic gender bias in the international standard of measuring economic growth, and the ways in which women’s unpaid work have been excluded from what counts as “productive” in the economy, have contributed to a significant and growing body of feminist scholarship on gender, work, and employment (Waring 1990; Campbell 2014).

It is important to differentiate and qualify the terms and concepts used in academic debate and policy because both work and employment are segmented, structurally and culturally, in diverse ways to reflect persistent inequalities between women and men. This is to say that the prevalent notion of work as employment overlooks a wide range of tasks performed as a function of love, duty, subsistence, and voluntary association, or as reciprocal exchanges between friends or neighbors. Nor is this distinction limited to money; more crucially, it involves identity, status, and “worth.” In societies where a measure of gender equality is perceived to exist, financial reward continues to signal to society that work as employment is “real” and “useful,” in contrast to the concept of the “unproductive” and “dependent” housewife. This is readily apparent in the way that people introduce and define themselves by a paid occupation, by what they are qualified and paid to do in contractual terms. Similarly, to be unemployed is to be branded “workless” even though lack of income obliges the unemployed single mother, for example, to work very hard to get by and to make do on limited means. In societies where gender roles and divisions are entwined with religion and tradition, where daughters are valued less than sons, these distinctions between “productive” and “reproductive” roles heavily circumscribe women’s life chances and status. Viewed this way, activities and sites of work in the broadest sense span multiple and diverse economies that variously trade through currencies of money, sex, bartered goods, reciprocal assistance, gifts, extortion, and exploitation.

This introduction demonstrates that “work” and “worth” are circumscribed in dominant public discourse by a gender bias in the way economically productive activity is defined and measured. Consequently, it is not possible to make sense of what constitutes work and
employment without first drawing attention to unequal power relations between women and men. For this purpose, gender is widely used as a concept that accounts for the varied and complex arrangements that encompass the organization of reproduction, the sexual divisions of labor, and cultural definitions of femininity and masculinity (Bradley 1996).

Gender is distinguished from sex because the latter confines social difference to biological characteristics. Explaining the differences between male and female lived experience and behavior by recourse to sexual (genital) characteristics is widely dismissed as a form of essentialism whereby male and female characteristics (dressing up in blue or pink, fixing things or talking about them, exhibiting strength or nurturing) are read off from sex as “in essence” “already there” from birth. By contrast, feminist scholarship has developed a more culturally nuanced frame of analysis that draws attention to the multiple and fluid performance, language, meaning, display, and popular representation (on TV and in magazines, for example) of different forms of femininity and masculinity. Feminist scholars challenge the policy discourse that treats labor as a generic, disembodied commodity.

While gender is widely recognized in feminist scholarship as a major axis of social difference, one that especially affects the segmentation of work and employment, it is important to recognize the significance of other intersecting social axes, such as class, race, ethnicity, and sexuality. It has not always been the case, for instance, that labor market research has taken adequate account of gender inequalities or of difference and diversity. A crude distinction is usefully made between the “structural” accounts of sex/gender systems of discrimination and constraint which dominated Euro-American scholarship in the 1970s and 1980s, and poststructural and postcolonial accounts that emerged from a splintering of radical, moderate, and “new” feminisms in the 1990s.

Classical studies from the 1970s, especially those in economic geography and sociology, tended to focus narrowly on those social changes associated with male-dominated major industrial sectors. This emphasis on working-class paid employment overlooked the significance of gendered work considered outside the relations of production, including the messy, fleshy lived experience of diverse service occupations and domestic work. Labor market segmentation was explained by grand narratives of patriarchy (literally the rule of the father) and the dual systems of capitalism and patriarchy. Subsequent decades witnessed growing confidence among feminist scholars to challenge androcentric and ethnocentric bias and to “recover” excluded and marginalized subjects (women, ethnic minorities, plural and fluid femininities and masculinities, “bodywork” and emotional labor) through poststructural theory and methods. Women of color were especially critical of structural theories that privileged certain types of (white, middle-class) knowledge and experience by employing gender as an imperializing category.

Poststructural critique has led to greater awareness of intersecting axes of social difference. For example, Geraldine Pratt (2004) introduces the figure of the racialized Filipina maid as a way to challenge white middle-class feminists to face their privilege, and draws attention to the gains in gender equality in affluent societies that often rest on the availability of poorly paid domestic labor as migrants from lower-income regions. Nevertheless, “gender” continues to be central to debates on work and employment as a means to challenge the dominant discourse on what counts as economically productive activity. This deployment of gender effectively opens up the black box of socially constructed ideologies
of work and worth; and it does not preclude recognition of multiple axes of exploitation.

The global economy has been fundamentally transformed in recent years by the increased international mobility of people, products, and money around the world. The term “globalization” is widely used to describe the growing interconnectedness of people and places on a global scale, while the term “economic restructuring” describes the transition to a different phase in the what, how, and where of GDP. Economic restructuring is widely used as shorthand for a shift from making things to selling services: while this top-down view of the world has a tendency to desex the social and spatial profile of inequality, these trends are intimately gendered in the cultural norms and microsociological practices enacted within and between households, firms, local labor markets, industrial sectors, and regional economic specialization.

Around the world, but especially in the Global North, there has been a dramatic reduction in the proportion of well-paid, skilled jobs in male-dominated manufacturing and primary industry (from 1 in 4 in the 1970s to fewer than 1 in 10 by 2010) (BLS 2012). Parallel growth in poorly paid and elite jobs in the service sector has been reinforced by hugely significant advancements in technologies associated with cheaper and more efficient transport alongside new and extended capacity for real-time, virtual, and time-shifted interactions via digital media and mobile telecommunications. It was in 1992, for instance, that the World Wide Web (WWW) was made available to home users for the first time, ushering in the rapid expansion of Internet use in all aspects of work including not only paid work at a distance but also the remote coordination of coworkers and people and activities in public and domestic spheres of daily life.

In affluent economies, the restructuring of work and employment tends to be understood in terms of laissez-faire market forces. By contrast, the restructuring of economies in the Global South has been, to a considerable extent, imposed by the policy decisions of supranational bodies such as the World Bank. Offers of financial assistance in the form of structural adjustment programs (SAPs) have come with harsh conditions attached. These conditions reflect systematic gender bias in that they do not recognize or protect the productive aspects of women’s unpaid work. It is assumed that the unpaid work done by women to care for the family, including raising children, gathering fuel, nursing the sick, preparing and processing food, will continue regardless of deep cuts in public expenditure and the pressure for women to participate in greater numbers in low-paid employment in the export sectors. There is evidence that male bias results in disproportionately high rates of poverty among women in the Global South, especially women-headed households. This is because cuts in public expenditure (state support for health, education, transport) intensify the workload for women, who are likely to sacrifice their own health and wellbeing to feed and educate their children (Perrons 2004, 57).

The number and quality of jobs created in affluent economies have been permanently altered by economic restructuring, whereby the ideal of a “job for life” has largely disappeared. This new precariousness can be understood in terms of four distinct employment trends, all of which reflect persistent gender inequalities in the social construction of work and associated identity and reward: (i) a more competitive international division of labor (multinational employers; jobs and production facilities outsourced to low-cost countries); (ii) an emphasis on flexibility (deregulation that enables firms to hire and fire employees based on their willingness to learn new skills and adapt to changes in working conditions); (iii) “feminization” of
job descriptions that emphasize “high-touch” provision of care services (interactive bodywork, washing, dressing, caring); and a casualization of the terms of employment (e.g., freelance portfolio workers replacing full-time employees, and “zero hours” contracts replacing fixed and stable working hours).

While the feminization of new service sector jobs appears to favor women over men, more women than men are concentrated in poorly paid, temporary, and insecure jobs. This concentration of women into particular sectors (e.g., hospitality, health care, education, factory production) and occupations (waiting on tables, nursing, primary education, component assembly) describes a horizontal segregation of jobs constructed as “women’s work” or as “mother-friendly” (e.g., based on part-time hours or shifts that can be reconciled with a “second shift” of child care and housework). Women who participate in paid work outside the home frequently work alongside other women in jobs that look remarkably like those of caring, domestic, and emotion work long performed inside the home as labors of love and duty. Vertically, too, women are concentrated in jobs or cultures of work that frequently offer limited access to promotion. This apparent barrier to women achieving equality of seniority and decision-making power is popularly described as a “glass ceiling” and it is readily observed in the absence of women from annual report photographs of the boards of directors of major firms. The glass ceiling derives from multiple causal factors, including the impact of maternity leave on opportunities to demonstrate commitment to career, as well as tacit sexism in the cultures and practice of recruitment and promotion.

A persistent gender pay gap is evident, which reflects the impact of both horizontal and vertical structures of gender inequality and segregation. Current wage data indicate that women in full-time management, professional, and related occupations in the United States are earning 72 cents for every dollar earned by men in equivalent employment. Across Europe the annual gender pay gap between men and women in full-time employment is stuck at a similar point and the gender gap in pension earnings is even greater at about 50% (Campbell 2014). This contributes to a pronounced feminization of poverty and vulnerability in later life. Further structural inequalities in pay are attributed to segregations by race and ethnicity as well as gender. This intersection adds to the complexity of the picture; for example, in 2012 Asian women who were full-time wage and salary workers had higher median weekly earnings than women of all other races/ethnicities, as well as African American and Latino men (Catalyst 2012). While feminist scholars acknowledge that gender functions alongside race and class in mutually reinforcing ways as part of a “wicked triangle” of structural inequality, these intersections reflect complex geographies of inequality (Grünell and Saharso 1999).

The jobs that have been created since the 1990s typically represent two different extremes of financial reward, and similarly contrasting experience of fulfillment, autonomy, and status. On the one hand, new jobs have been created in areas of high-technology, new digital media, and the knowledge economy. For the (predominantly male) workers in the high-tech sector, these new jobs are skilled and generally fulfilling, offering high wages to compensate for expectations of long and typically unsocial hours, as well as time spent away from home working on projects overseas. Research and policy in this area draw attention to issues of masculinized institutions, self-exploitation, and the negative impact on health and family life of intensely work-centered, goal-oriented lifestyle practices and identifications. This serves to reinforce divisions between
working mothers’ time and men’s time whereby it is rarely possible for a dual career working family household to practice a dual carer arrangement, except perhaps in Scandinavia where shared care is more likely to be supported by progressive child-care and egalitarian parental leave policies (Gornick and Meyers 2003).

On the other hand, a substantial and rising proportion of jobs have been created in service sectors that are notorious for low pay, temporary and seasonal working, variable hours, and poor working conditions. Examples include the hospitality, retail, health care, and personal service sectors. While these poorly paid jobs are classed as “unskilled,” the (mostly female) workers recruited to them are increasingly required to perform intimate bodywork (cleaning, caring, serving) and emotion work (dealing with other people’s feelings), which represent undervalued (underpaid) “soft skills.” Feminized interactive service employment tends to offer not only low wages but also poor job satisfaction and little influence over the conditions of speed, stress, and safety under which tasks are to be performed. Research and policy action in this area draw attention to the human costs of very low wages and the erosion of autonomy and dignity. Some of the emerging issues have been brought to light by ethnographic methods of documentary journalism involving firsthand insights of the social indignities and economic realities of women living and working on minimum or “poverty” wages. Two examples are Barbara Ehrenreich’s (2001) searing account of working as a waitress, hotel maid, house cleaner, nursing home aide, and Wal-Mart salesperson in the United States and Polly Toynbee’s (2003) equally vivid portrait of life in low-pay Britain.

Economic competitiveness and growth have become paramount in public discourse on the demand for, and supply of, labor. The imperative that states and firms are globally competitive has repeatedly been used as justification for dismantling the terms of employment protection that were previously negotiated in struggles between organized (mainly male) labor unions and the interests of capital (firms and the state). Arguments about the impact of economic restructuring on new articulations of patriarchy through socially stratified gender divisions of labor tend to distinguish between optimistic and pessimistic narratives.

The optimistic narrative is of “detraditionalized,” mobile, educated portfolio workers who use social networks and technologies to escape from conventional gender and class divisions. Thus Beatrix Campbell (2014) observes that neoliberal competitiveness tolerates girls being astronauts or bankers but resists dismantling deeper structures of sexism and misogyny. The pessimistic vision points to a disillusioned and exploited “reserve” labor force (women working part-time, women of color, migrant men and women who are poorly represented by trade unions) and, beyond that, a permanently excluded underclass of NEETS (Not in Employment, Education, or Training). For example, there are competing viewpoints concerning the extent to which “flexible” labor market practices allow workers to choose when and how much to participate in employment whereby so-called zero hours contracts allow employers to hire staff with no guarantee of work: employees are only paid for the hours when their services are required (often at short notice), without incurring overheads for the workers on standby. Critics point out that promoting these contractual arrangements as mother-friendly attempts to disguise new articulations of modern slavery.

Economic restructuring has been accompanied by a parallel trend of diversity and change in the social characteristics of the labor force and the structure and geography of household employment. The quantity and quality of
spouse employment tend to be multiplied within working family households whereby work-rich (time-poor) dual career, dual income, or multitask households are significantly advantaged compared to households where none of the working-age adults is earning a living from paid employment. A combination of increased job insecurity and escalating housing costs motivates the majority of couples to believe that both partners have to work. In households with young children it is increasingly common for both parents to participate in paid employment but the cost of commercial child care imposes significant constraints on the working hours and times that mothers, in particular, can balance or reconcile with family care. It is not uncommon for parents of young children to work opposite shifts (such as night and day) or for mothers to fit part-time paid employment around the demands of child care, whether by working from home in the hours when children are sleeping or by working when older children are at school. Moreover, the hours that people work vary far more than in the past: nonstandard schedules are increasingly associated with a 24/7 economy whereby large grocery stores remain open 24 hours a day to service a customer base made up of workers who are also working shifts that would otherwise prevent them from shopping for their families during traditional 9 to 5 retail opening hours.

Time is a significant resource with respect to the coordination and reconciliation of the activities and identities of work with the rest of life – which is not gender-neutral. For many years the overwhelming preoccupation has been with time allocation as a function of the organization of employment, where time is conceived according to a featureless plane – of working hours (length of the working day) and working times (extended office hours and nonstandard shift arrangements), using categories such as full-time and part-time to differentiate levels of career salience and labor market participation. This emphasis on time as the essential currency of production coincides with a “workification” or work-centeredness of contemporary life. Attention has turned more recently to the gendered experience of time, such as feeling harried. Inequalities are variously explained by the pressure, intensity, and disruption of time that women feel more than men. Time pressures are imposed by scheduling constraints for specific tasks (such as collecting a child on time from daycare) and in relation to managing child-care providers and children’s daily schedules. While men are observed to help out and fill in with child care and to organize time in chunks to undertake a discrete domestic project, women typically experience time with greater intensity: activities are typically being done in combination or all at once. Women are also more likely than men to have their leisure time interrupted by domestic tasks. At the same time, social changes that have developed in affluent economies alongside popular struggles for work–life balance (e.g., “wraparound” child care and online shopping) further contribute to the socially unequal gendered experience of time.

A significant trend in recent years is the prevalence of in-work poverty, which is associated with problems of underemployment (part-time or variable hours) combined with low pay, especially where wages do not cover the cost of child care and the rising costs of living, notably housing and food. This is in a context in which the assumption in neoliberal states is of a “universal worker” model of citizenship. In order to maintain national productivity and international competitiveness, governments assert that all prime-age women and migrants should participate in paid employment on whatever terms the market will bear: this pressure is reinforced by the dismantling of protective legislation which in the
past served to reward one member of a house-
hold (typically male) with a family wage that
was sufficient to compensate another household
member for unpaid work at home. Economic
restructuring has exerted a downward pressure
on wages in nonprofessional service sector jobs,
especially where previously secure public sector
jobs (catering, cleaning, and security) have been
transferred out to private subcontractors who
typically rehire staff at a reduced rate of pay. In
some sectors and places this new paradigm of
insecure employment has triggered the forma-
tion of new grassroots community unionism and
campaigns and coalitions (scholars, activists, and
nonprofit organizations) to raise awareness of
issues concerning corporate social responsibility,
for example campaigning for a living wage can
be an efficient and effective tool for reducing
gender pay inequality (Wills 2009).

Evidence that women tend to be confined
to the most vulnerable categories of employ-
ment both in the formal interactive services of
affluent economies and within poorly paid and
informal sectors in developing countries has
motivated action for change within women’s
groups and transnational feminist movements
around the world. The Global Women’s Strike,
for instance, is a transnational network of direct
action intended to seek recognition for women’s
unwaged caring work in different settings around
the world, building on the seminal radical cam-
paign Wages for Housework, orchestrated by
Selma James in the 1970s (Jarvis, Kantor, and
Cloke 2009). Promoting decent and productive
employment and income opportunities equally
for women and men is one of the key priorities
of the International Labour Organization’s (ILO)
Decent Work Agenda. Another example is the
Clean Clothes Campaign, which raises awareness
of women’s working conditions in the global
garment industry, notably in Bangladesh. The
vast majority of workers in garment factories
worldwide are women, and the campaign is an
international alliance of trade unions and labor
rights activists working together to demand that
garment workers in Asia are paid a living wage
(Clean Clothes Campaign 2014).

Despite a huge expansion of academic analyses
in the field of waged work, a range of questions
arguably remain unresolved, most notably con-
cerning the intersection and co-constitution of
multiple dimensions of paid and unpaid work.
Gender is a highly politically charged concept
around which the notion of patriarchy has been
used to describe the systematic privilege and
domination of men over women. The language
and theory of patriarchy are less widely used in
feminist scholarship today because gender and
poststructural theory have a greater capacity to
reflect fluid and multidimensional relations that
are constructed, performed, and inhabited in
complex and changing ways. This is not to deny
the enduring structures of constraint in women’s
access to power and representation. Women
remain underrepresented in parliaments, judicial
systems, and the senior civil service.

Indeed, since the development of detailed
who-does-what studies of housework, child
care and money management in the 1970s,
time budget surveys have consistently revealed a
gendered sociospatial distinction between indoor
“wifework” and outdoor, “macho” home con-
struction and maintenance, typically involving
machinery (Nelson and Smith 1999, 125).
Enduring norms of behavior are reproduced
through the gender stereotypes represented in
magazines on cookery, cars, home decoration,
needlecrafts, home improvement projects, and
power tools. According to 2012 data from the
American Time Use Survey, on an average
day, women in the United States spent more
than twice as much time preparing food and
drink, almost three times as much time doing
interior cleaning, and four times as much time
GENDER, WORK, AND EMPLOYMENT

doing laundry, as men did. Men spent more than twice as much time doing activities related to lawn, garden, and houseplants, and doing interior and exterior repairs and decoration as women did (BLS 2015). While survey data continue to be reported in the news media, the focus of academic research has shifted away from time budget surveys toward more culturally nuanced ethnographies of how bodies, places, and biographies are gendered today.

Conceptual developments in gender performativity have been especially fruitful in recent studies of the workplace. Focusing on how gender identities are constructed through daily interactions at work allow questions to be asked about workplace cultures and the policy changes that are needed in order to challenge tacit as well as overt discrimination and disadvantage. From the mid-1980s a growing number of studies on gender and work (gender at work as a process) began to explore the ways in which gender-specific traits and characteristics are attributed to men and women through the work they do in particular settings. Increasingly it was understood that jobs are not gender-neutral but instead constructed as appropriate for men or for women. Likewise, gendered identities are created and recreated at work, rather than being held as fixed and unchanging attributes that individuals bring with them to the labor market. This gives rise to the potential in the workplace context for workers’ bodies to be controlled and regulated to create a particular brand or commodity that reinforces expectations of hypermasculinity, for instance, and the sexually available female body. An example of the latter would be the US sports bar Hooters, which requires female waiting staff to wear provocative cropped tops and short shorts. Further illustrating this point, a recent recruitment drive for an international clothing retailer specified that the new store they were opening would be employing “store models” rather than shop assistants. From the words and images used in the advertisement, the underlying message was clear: only young, slender, bright, beautiful people need apply (Jarvis, Kantor, and Cloke 2009, 18). In the new 24/7 consumer-led service economy the physical attributes of the body providing a service are part of the exchange that occurs at the point of sale (McDowell 2009).

Another significant body of recent research explores the intricate chains of exploitation binding some individuals in servitude while at the same time linking them, indirectly, to more affluent households (especially working parents in global cities), who themselves seek solutions to time pressures and the squeeze on social reproduction work. Motivations and driving forces behind the growth of private personal services (help in the home) can be attributed to the rise in women’s paid employment, such as with the normalization of a dual income household, as well as the growing complexity of home-work–family life (including long-distance commuting) balance. In her book The Time Bind, Arlie Russell Hochschild (1997) explored the repercussions of a work-driven lifestyle, highlighting the marital tensions in dual income couples where parents worked longer hours and husbands rarely contributed their share of the housework, all in a climate of rising cultural expectation of intensive parenting, labor-intensive home decoration, and high standards of cleaning and garden maintenance. This is the demand-side explanation for increasingly complex chains of exploitation. The supply-side explanation has also attracted considerable attention through issues of migration, low pay, human trafficking, and bonded labor.

One of the largest domestic maid services, Molly Maid, operates through thousands of franchised businesses in towns and cities across Europe and North America. Every aspect of
the brand (pastel pink with a frilly maid’s cap) is gendered in a stylized way. A number of handyman agencies have also emerged (notably in Australia) which similarly adopt stereotypes that reinforce the idea of a new division of labor. The question of who does what in the home is no longer limited to a man or a woman but is instead extended to “a cleaner.” The cleaner is positioned in this new division of labor as of such low status that he or she is desexed, less than man or woman – barely a human machine (Ehrenreich and Hochschild 2003). Marriage counselors popularly recommend that dual career couples hire a cleaner (a cleaning lady) to shore up the myth of gender democracy in their relationship and as an alternative to squabbling over whose turn it is to empty the dishwasher. These are profoundly gendered systems of occupational segregation at work – but they are global, intersectional, and increasingly also dehumanizing (Jarvis, Kantor, and Cloke 2009).

The concept of emotional labor is often used to help explain gender divisions of labor in the service sector that emphasize the significance of social characteristics including ethnicity, skin color, body shape, and other markers of difference. Examples of emotion work may include a parent’s efforts to keep children happy, a nurse comforting a bereaved relative, or a trade union leader rallying members to strike for fair wages. Historically the display of emotions has been denigrated as a feminine attribute. With economic restructuring and the proliferation of service sector employment, workers are increasingly expected to display (or convey) tightly scripted emotions as part of their job. Examples of workers whose job it is to follow a fixed script and embodied performance include call center employees “smiling down the phone” and shop staff wearing the shoes and clothing that are for sale in the store – masquerading as living mannequins.

Work and employment can be viewed in a number of ways, as structured events in people’s lives and as powerful sites of identity formation. Terms like “transition,” “reconciliation,” and “coordination” have been used as a way of thinking about multidimensional and dynamic identities and events. Viewed this way, the movement by an individual in and out of employment may correspond with movements in and out of education or to accommodate life events (such as taking time out to care for young children or older relatives) or because of injury or long-term limiting illness. The picture of individual employment is made more complex by home and market lives that are interlinked through family and household relations: the nature and extent of work and employment that any one individual undertakes have a significant bearing on the quality of life not only of that person but also of family members and other people in the same household. Indeed the changing landscape of work and employment is often seen as a challenge for family life, but largely for women as mothers because of the complicated and difficult acts of juggling competing demands of making a living and caring for a home and family. It is widely acknowledged that, even as more women have entered the paid labor force alongside men and are working increasingly similar hours over a lifetime, men have not shifted to an equivalent extent to pick up the tasks of unpaid work required at home (Presser 2003).

There are a number of specialist academic journals devoted to the study of gender, work, and employment in various combinations of emphasis: Work, Employment & Society (a British Sociological Association publication); Gender, Work & Organization; Gender, Place & Culture; and the Labor Studies Journal. These journals and the continuing production of knowledge and debate in this field attest to continuing speculation about the future of work (Green 2005).
GENDER, WORK, AND EMPLOYMENT

As a general trend, academic interest in work has become more integrated with ethnographic studies of social difference, social justice, and social movements. This has led Halford and Strangleman (2009) to conclude a broad trend of marginalization with respect to the academic (especially theoretical) interest in the way work is rooted in classical notions of social change. They argue that the study of work has become dis-embedded from wider social theory. A number of new research directions appear to contradict any suggestion of declining scholarship. These include questions associated with the social and economic impact of the 2008 global financial crisis and the decade of recession and state austerity measures that followed. The scarring effect of unemployment is one area of growing policy relevance, alongside new concerns around issues of intergenerational equity and the inter-connection of income and employment, debt (including student debt), and the uneven quality of individual life chances (health, notably mental health) where these are co-constitutive with diverse household employment structures and unequal access to livelihood amenities including housing, wealth, social capital, and welfare.

SEE ALSO: Care work; Domestic spaces; Domestic workers; Emotional labor; Flexible labor markets; Patriarchy; Restructuring; Social justice; Social reproduction; Unemployment and “underclass”; Uneven regional development; Work–life balance

References

Gender, Work, and Employment


Further reading


Gender

Robyn Longhurst
University of Waikato, New Zealand

Geographers have employed the concept of gender differently over time and space. It is a concept that is fluid and overlapping, therefore, presenting a chronological narrative that represents categorically different and progressive understandings of gender throughout the history of feminist geography is not without difficulty. Having said this there are some general trends worthy of note.

In the 1970s the concept of gender was introduced into the study of human geography. While sex was understood to describe fixed biological differences between men and women, gender was used to refer to socially constructed differences that are amenable to change. A small but growing number of geographers were becoming increasingly interested in how gender differences impacted on people’s, especially women’s, lives. Gender, it was argued, could be used as a concept not just to comprehend differences in men’s and women’s identities but also to recognize that societies and spaces are structured in a manner that is patriarchal. Patriarchy, like the term gender, has a variety of meanings but feminist geographers in the 1970s and 1980s were using it mainly to refer to men’s domination of women. The stereotypical roles played by women (e.g., homemaker) tended to be devalued compared to the stereotypical roles played by men (e.g., breadwinner). Topics such as gendered inequalities in relation to access to facilities, women and development, and the interdependences of men’s paid and women’s unpaid labor began to be addressed. Also, with a growing awareness of gender issues, geographers began to critique sexism and anthropocentrism in the discipline. Women were largely absent as the subjects of research. They were also absent as geography teachers, lecturers, and researchers.

Toward the end of the 1970s and early in the 1980s gender geography groups started to form. For example, in 1981 the International Geographical Union (IGU) held a session at a meeting of the Commission on Rural Development at California State University, Fresno, in which seven papers on women were offered. It was not until 1988, however, that the IGU Executive Committee finally approved a Study Group (later to become a Commission) on Gender and Geography. While the IGU represents an international group, national groups were also forming. For example, in the United Kingdom in the 1970s an informal group, and then “working party” within the Institute of British Geographers (IBG) was formed called the Women and Geography Study Group (WGSG), a group which became formally recognized in 1980 (Morin 2009). These groups were important in helping to constitute a new “field” known as gender or feminist geography. Gender was, and still is, at the heart of feminist politics and scholarship, including feminist geography.

Over the decades, however, understandings of the concept have changed. In the 1970s and 1980s when gender tended to be used in contrast to the concept sex, the emphasis fell heavily on the way men and women are “made” or socially constructed providing a route for challenging what seemed like essential and immutable differences in the stereotypical roles...
played by men and women. Women, it was argued, occupy widely differing positions in the roles they perform, both within and outside the home. Women are often the caretakers of children, elderly, and other family members, responsible for housework, active workers in various community roles and engage in paid work. Regardless of the varied roles women take on, however, they tend to be seen as less important than those played by men. It became apparent that in many areas such as geographies of employment, urban geography, and political geography there was a tendency to adopt perspectives that subsumed women’s activities into those of men. This recognition of women’s invisibility gave rise to a number of studies that used the concept of gender to identify the nature of women’s contribution to the home, workplace, and community.

Documenting the different and unequal gender roles played by men and women challenged the assumption that men’s work, lives, and the public spaces they occupy are the norm against which women’s work and lives are measured. This research paved the way for further scholarship in the 1980s and 1990s that focused less on gender roles and more on gender relations. Rather than understanding gender in terms of socially ascribed roles, the concept of gender relations enabled a deeper examination of the reasons why roles take the form they do, and how roles are contested and change over time and space. Gender came to be seen as a relational concept. It was no longer just about women per se but about power relations between men and women.

This shift from gender roles to gender relations in the mid-1980s led some to draw not just on feminist but also Marxist concepts to tease out intersections between gender and class politics. Feminist geographers began to question Marxist and political economic understandings of the labor process arguing that the position of women needed to be conceptualized not just in terms of labor relations but also gender relations. Arguments about the merits of this approach were rehearsed in 1986 and 1987 when geographers debated the concept of patriarchy in the journal Antipode (McDowell 1986). It was important to understand not just how capitalism alone creates and perpetuates unequal power relationships but also how it intersects with patriarchy to oppress women.

From the mid-1980s, then, geographers became increasingly interested not just in understanding the roles played by men and women but in the systematic gendered and capitalist power relations produced through, and producing, particular constructions of femininity and masculinity. Research addressing contexts such as the home, community, workplace, city, and rural spaces was carried out in an attempt to understand more about the gender relations and identities constituted within these spaces. It became apparent that gender- and class-based power relations result in women having for the most part access to fewer resources and opportunities than men. Their disadvantage in the labor market combines with responsibilities for childcare limiting women’s opportunities, especially in full-time and professional employment. Focusing on gender relations enabled geographers to understand how patriarchy and capitalism within different contexts leads to women’s inequality. Feminist geographers aimed to draw attention to the unvalorized role of the reproductive sphere within the largely hidden worlds of home, childcare, elderly care, and other community spaces. Gender relations began to be theorized as spatially variable and at a range of scales, not just regions, cities, and nations, but also the home and community.

In the late 1980s and early 1990s feminist geographers started to conceptualize gender not just as a role or a relation but as a complex and shifting
subject position. Feminist poststructuralist work of theorists such as Judith Butler (1990) gained in popularity prompting debate about issues such as performativity, embodiment, essentialism, representation, discourse, and heteronormativity. It was becoming increasingly evident that the way in which some women’s identities differed from the assumed dominant form (heterosexual, white, Western, middle class, able-bodied) mattered. Butler’s argument that gender emerges out of repeated gender performances, brought into effect through discourse, captured the interest of geographers who were keen to show how these repeated and unstable practices are enacted in different ways, at different times, and in different places.

In early feminist geographical scholarship there tended to be an assumption that whiteness and heterosexuality were the norm; however, in the 1990s there was growing recognition that not only do gender differences between men and women matter but so too do differences among women. Feminist geographers moved from the using “grand theories” to explain women’s “universal” oppression to instead examining smaller-scale, more localized expressions of gendered subjectivities. It was becoming increasingly apparent that gender is a fluid concept marked by some similarity but also by difference. Geographers started to question what were seen as the priorities of white, middle-class, Western feminists in the academy and instead encouraged recognition of the multiple and varied lives of women marked as “Other” (Kobayashi and Peake 1994).

The 1990s saw increasing diversity of representation as women of color and researchers from outside the Anglo-American center began to add their voices to the body of work on feminist geography. That is not to say that all parts of the world are now represented, far from it (Garcia Ramon and Monk 2007); however, as feminists from different countries, notably India and more recently Japan, and regions, particularly continental Europe, Southeast Asia, and Latin America, began to discuss gender issues the extent to which place and positionality shape knowledge became increasingly apparent (Monk 1994). Questions began to emerge about the Anglo-American hegemony of feminist geography including use of English as the *lingua franca* in publication outlets. Questions were also raised about what constitutes useful feminist theory or challenges inequalities in different places and spaces. The universal and essentialist category of “Woman” had by the 1990s been subject to critique and could no longer be used unproblematically as the foundation for feminist politics. Criticisms of feminism as a white, Western, and middle-class phenomenon made it clear that differences among women were at times more important than similarities.

Also in the 1990s and 2000s it became increasingly obvious that the concept of gender is not just about women and femininity but about men and masculinity. There was a growing body of work in feminist geography on masculinities, men, and male subjectivity. Examples of topics addressed include hegemonic masculinity, alternative masculinities, masculinities at work, men and empire, soldiers and masculinities, and masculinities and disability. There was also a great deal of research focused on sex and sexuality, especially gay male identity.

As the concept of gender was re-theorized by feminist poststructuralist scholars in the 2000s it became increasingly problematic to prize apart gender and sex. Butler (1990, 33) argues gender is “the repeated stylization of the body, a set of repeated acts within a highly rigid frame that congeal over time to produce the appearance of substance, of a natural sort of being.” When gendered and sexed performances in different times and places do not align in ways that are
GENDER

considered normative, when people “do” their gender differently, the result is “gender trouble” (Butler 1990). Many feminist geographers began to embrace this definition of gender as something that people “do” recurrently, embodied acts and gestures when repeated come to take on the appearance of the “real” or an “essence.” Not only gender but also sex and sexuality became increasingly evident in human geography in the later 1990s and 2000s (Domosh 1999). It was argued that if gender and sex are understood to have ontological status but are produced by repeated performances (Butler 1990) heterosex is clearly constructed rather than “natural.”

Another linked strand of scholarship that began in the 1990s is that of embodied geographies (Longhurst 1995). The body’s association with femininity has provided a route for feminist geographers to deconstruct gendered binaries such as sex/gender, mind/body conceptual/corporeal, and essentialism/constructionism thereby disrupting hegemonic masculinist structures of knowledge production (Rose 1993).

Over the past decade as feminist geographers have increasingly paid attention not just to gender per se but to the body and how gender intersects with class, race, ethnicity, masculinity, sex, sexuality, and more recently health, (dis)ability, body size, and age, questions have been raised about according primacy to gender as the central analytical category and whether there are instances in which gender ought to be decentered. Critical geography, which often addresses the intersections between different subjectivities, began to gain ground in the late 1990s. The Inaugural International Conference of Critical Geographers (IICCG), held in August 1997 in Vancouver, British Columbia, Canada helped propel interest. In 1999 Routledge began publishing a new series entitled “Critical Geographies.” In 2001 a new online international journal ACME An International E-Journal for Critical Geographies began publishing. While these initiatives did not signal the end of gender as an important analytical category they did prompt discussion about an exclusive focus on gender.

Gender, for more than four decades has been a central concept in shaping how feminist geographers think about place and space. Given, however, that feminist geography is a heterogeneous subject area that encompasses multiple understandings of gender these understandings do not necessarily map neatly onto different decades and spaces although there are some trends which are worth reflecting upon. To recap, in the 1970s gender tended to be equated with women and was understood in terms of roles. In the 1980s the focus shifted to gender relations in the hope it would offer more explanatory potential for understanding patriarchal relations and the oppression experienced by many women. In the 1990s geographers, informed by feminist poststructuralism, began to conceptualize gender as an embodied performance that could not be separated from the sexed body. Geographers increasingly questioned the priorities of white, middle-class, Western feminists instead recognizing the multiple and varied lives of women. The 1990s and 2000s saw increasing diversity of representation beyond the Anglo-American center. Masculinity, sexuality, and the body were also put on the agenda. Today, gender is still an important concept but questions have been raised about when it ought to be accorded primacy as the central analytical category and when it ought to be decentered or seen as intersecting with other subjectivities such as class, “race,” ethnicity, sexuality, age, and ability.

SEE ALSO: Antiracist geography; Critical geography; Difference; Domestic spaces; Feminist geography; Gender, work, and employment; Intersectionality; Patriarchy;
Poststructuralism/poststructural geographies; Queer geographies; Race and racism

References


Further reading

Gentrification

Brenda Kayzar
Kate Derickson
University of Minnesota, USA

Gentrification is the transformative process whereby lower-income and working-class households are subjected to higher rents and eventual displacement as higher-income households purchase and rehabilitate inner-city residences. In similar fashion, lower order businesses are pushed out as the character of the neighborhood changes and commercial space is renovated. From its introduction, however, gentrification has been a politically charged and less than simple term.

Concern over the displacement of lower-income households from London’s older inner-city neighborhoods prompted Ruth Glass (1964) to coin the term gentrification five decades ago. Her intent was to describe succinctly not only the outcome of the acquisition of inner-city residences by middle- and higher-income homebuyers and subsequent demographic shifts, but the impacts upon the former occupants of these properties as well. As Slater (2011) suggests, because the term “very powerfully captures the class inequalities and injustices created by capitalist urban land markets and policies” (2011, 571), this transformative process has been the subject of ongoing inquiry and debate.

Advocates for greater equity in property markets have sought to maintain Glass’s focus on the class shift and loss of affordable housing and commercial space inherent in gentrifying processes. They argue gentrification should be a planning and social concern. The loss of affordable housing in gentrifying communities is not offset by the production of affordable housing in proximate communities, making the advent of displacement even more arduous for the lower-income households and small businesses that are forced to relocate. Advocates for revitalization of the inner-city have sought to lessen this focus by highlighting the benefits of reinvestment such as aesthetic improvement to the neighborhood and increased property values and tax revenues. Gentrification is reconfigured and offered as a solution to the degradation of disinvestment in older parts of the city.

A myriad of arguments populate the continuum between these two perspectives and include disagreements pertaining to displacement and neighborhood improvement measures, the character of the actors involved, and the proper methodological approaches to understanding processes of neighborhood change. The contentious and often intense discourse around gentrification has drawn together scholars from political science, sociology, geography, anthropology, housing studies, planning, and economics. In the main, their collective efforts have sought to define gentrification, explain why it occurs, describe who is impacted and in what way, and to suggest what market, policy, or social justice effort presents the best mitigating outcome.

Production and consumption

There are two primary points of entry to the study of why gentrification occurs. The first focuses on the political economic processes
that produce the market conditions conducive to gentrification ("production"); the second focuses on the cultural and social relations that render the inner-city a newly desirable place for the middle-classes to consume ("consumption"). Work that focuses on production is typically rooted in Marxist theories of land markets. Defined by Smith (1996) as a rent gap, this perspective suggests gentrification is a result of uneven development, in which capital is withheld from inner-city properties as the appropriateness of their use diminishes. This capital is invested elsewhere as demand and profit opportunities dictate (i.e., suburbs, commercial property). Disinvestment, which reduces property values over time, enables reinvestment. Devalued properties can be purchased with less capital outlay for redevelopment to the benefit of inner-city property owners, investors, and developers. Work that focuses on the consumption of the city seeks to understand the process whereby a new middle class eschews the suburbs and seeks an alternative aesthetic and lifestyle in the city (Ley 1997). A succession of gentrifying actors have been introduced as a result of studies employing this demand-side approach from the early pioneers looking for more affordable housing with historic character, shorter work commutes, and racial and class diversity, to artists in search of affordability and the ability to practice their craft without complaint from neighbors, to the less expected middle-class minority and family households. While some scholars have resisted integrating these competing explanations, Lees, Slater, and Wyly (2008) note that by the fourth decade of the gentrification debate most scholars were accepting of a multiplicity of causal factors. Competing understandings of the causes of gentrification have generated a wide range of methodologies for studying it, leading to different and often contradictory outcomes.

Methodologies and outcomes

Bondi (1999) notes the consumption and production perspectives of gentrification processes represent “cultural and economic imperatives” that remain important frameworks of scholarly inquiry (1999, 261). Within these frameworks, scholars have proffered numerous theoretical and methodological approaches in an effort to enrich the discourse regarding the sustained transformation of older neighborhoods.

Various data resources such as census and government surveys have been used to discern the existence of gentrification and its spatial extent, as well as the scope of demographic change. Pull and push factors ranging from proximity of amenities to the tenure status and incomes of existing residents have been examined to discern predictors of gentrification. Interview and ethnographic methods have provided further insight into the motivations of new, existing, and departing residents. Collective resistance has also been studied to comprehend the success or failure of these mitigation efforts.

While economic theories such as the rent gap continue to be revisited, there has also been an effort to broaden the theoretical discussion with regard to race and gender. Recognition of the imprint of historical exclusionary practices on policy and the material landscape offers an important contextualization to the current gentrification debate as it demonstrates the impact of power relations favoring white middle and higher-income households. Past practices fostered inequality in housing opportunity and segregated populations. Disinvested neighborhoods, depicted as frontiers of financial opportunity for wealthier majority households, cannot be viewed without contemplation about who already exists in – or was relegated to – these spaces (Smith 1996). This figurative “emptying” of the urban landscape intimates...
revanchist desires or a recolonization of the city for the privileged classes, at the expense of lower-income and often, minority households.

Scholars have discussed the oppression of suburban development and the emancipatory space of the city for many women. Research on gentrification also addresses the role of women, noting that single-woman and female-headed households with children have been evident in gentrifying neighborhoods and that the rent gap presented an opportunity for career women to build sweat equity in real estate while residing near their employment. Bondi (1999) provides a comprehensive review of this research through the end of the millennium and offers her own empirical analysis of three communities in varying stages of change; interweaving life course into the equation of gender and class as an additional structural factor in women’s participation.

In general, the differing measures and deeper contextualization of the past five decades of research further established the complexity of gentrification. While there is wide-ranging agreement around the consideration of a multiplicity of structures and contingencies, there remains disagreement around the positives and negatives related to who is impacted and in what way.

For example, the results of a study and subsequent book by Freeman (2006) proposed that the high rate of mobility of low-income households was due to their already tenuous economic status and was not increased by gentrification. Freeman’s argument against displacement, along with studies signifying deconcentration as a positive outcome for poor households are antithetical to Glass’s aim, but they also reflect a shift in the way gentrification is perceived by policymakers as well as some scholars. While early studies and policy actions resonated with Glass’s concern about the potential for displacement, and thereby continued segregation and social disconnection for lower-income groups, Lees, Slater, and Wyly (2008) note that gentrification is now embedded as a goal in urban policy in North America and Europe under the assumption that it will foster less segregated and more sustainable communities. Cautionary narratives “have been overtaken by a global neoliberal discourse of regeneration and renaissance” (2008, xvii).

USA Today’s interpretation of Freeman’s study resulted in the pro-gentrification headline “Studies: Gentrification a Boost for Everyone” in 2005. Terms such as revitalization, renaissance, and redevelopment have replaced gentrification in policy and scholarly lexicons, and they imply progress while obfuscating the reality of displacement and social disruption as patronage to a pro-growth agenda. The pro-growth argument draws on deindustrialization as the impetus for cities to re-envision economic futures as centers of consumption and entertainment. Historic preservation, smart growth, and sustainability efforts further bolster the discourse in favor of reinvestment in older housing stock and the design and aesthetic appeal of older neighborhoods. Further, cosmopolitan and creative workers in the new knowledge and creative economies are expected to demand revitalized urban centers in which to live, work, and play, prompting “how to” guides for civic leaders and planners. Wilson (2007) argues that concern about gentrification has been recomposed as a civic necessity, conflated in globalization and competitive city rhetoric while the spaces of low-income minority households are further compromised.

**New spaces of gentrification**

Recently, research has sought to broaden our understanding of the actors involved and the places impacted by housing mobility and displacement, rethinking the process of
GENTRIFICATION

gentrification as displacement by the white middle class, and from the urban core. Singles, dual-income households, empty nesters, artists, the creative, and gays have all been suggested as consumers of gentrifying neighborhoods, and scholars are noting a further evolution in the gentrifying population. More recent studies explore the return of black middle-class households to urban communities in decline, challenging the commonly held belief that gentrification is driven by white middle- and higher-income households.

The remaking of inner-city neighborhoods as historic and cultural centers has been expanded to include studies of the postindustrial rural west and older industrial suburbs and urban districts. Lees, Slater, and Wyly (2008) explore the way in which the notion of gentrification has been expanded, “mutated,” and applied to spaces beyond the urban core, fostering different types of displacement. For example, they discuss the liability of students on a place, the role of new construction, as well as forms of mobility and displacement generated by hyper-financialization. They argue, however, that all of these “mutated” forms have in common “a socio-economic and indeed cultural transformation due to middle-class colonization or recolonization” (Lees, Slater, and Wyly 2008, 133).

Finally, scholars of gentrification have turned their attention to the increasing pace of gentrification beyond world cities, to the more quotidian spaces of the Global South, as a generalized strategy of urban policy and development. According to Smith (2002), gentrification is “no longer isolated or restricted to Europe, North America, or Oceania, the impulse behind gentrification is now generalized; its incidence is global, and it is densely concentrated into the circuits of global capital and cultural circulation” (Smith 2002, 202, 427). Lees (2012) proposes a postcolonial approach to studying “mega-gentrification” in the Global South and suggests productive intersections between new work on assemblages and policy mobility and the rise of gentrification in the Global South.

A return to the social justice agenda

While ways of knowing about gentrification and associated scholarly inquiry differ, much work (though not all) in this tradition in geography is united by a critique of the enduring vulnerability of the working classes and minority households to displacement, whether due to circuits of capital running their course or changing preferences of the middle- and upper-classes. This critique is borne out in the activism of the “Right to the City” movement, which has strong ties to academic geography. David Harvey, Peter Marcuse, Nik Theodore, Don Mitchell, Ed Soja, Mark Purcell, and Susan Fainstein have all been actively involved in collaborating with activists and citizens’ groups to produce the political language and scholarly studies that critique class-based urban precarity. A phrase attributed to Henri Lefebvre regarding the “right to the city” asserts the rights of everyone, including the poor, marginalized, and disenfranchised, not only to live in the city free from the threat of displacement, but to participate in shaping or “making” the city.

SEE ALSO: Class; Housing; Neighborhood; Social justice; Urban uneven development

References


Geoarchaeology

Lisa Maher
University of California, Berkeley, USA

Definition

The subdiscipline of geoarchaeology has come a long way in the last few decades in terms of research scope and focus and, today, it frequently plays a central role in archaeological research design. Much ahead of his time, British prehistorian Colin Renfrew noted that “… since archaeology … recovers almost all its basic data by excavation, every archaeological problem starts as a problem in geoarchaeology” (Renfrew 1976, 2). Now, archaeology as a discipline widely recognizes this statement. As such, geoarchaeology today is defined as the study, at all scales, of the direct interrelationships between past humans and the landscape/environment, where the landscape can be viewed as the intersection of physical processes and cultural activities. Notably, this definition includes the social, economic, technological, and ideological (including symbolic) aspects of human interactions with the environment. It covers a wide range of topics and deals directly with: (i) material culture (how physical objects were made, obtained and/or used); (ii) cultural attitudes and perceptions of landscape (such as the symbolic status of artifacts or landscape features); (iii) landscape modification and other human impacts; and (iv) issues of subsistence (i.e., how resources were culturally manipulated; how economy affects and is affected by the patterning of resources and their human and nonhuman exploiters), settlement (i.e., how relationships between patterning in landscape and patterning in distributions of settlements vary over time), daily activities (i.e., identifying single and repeated activities or routines), and sustainability (i.e., examining the temporality of specific human activities). The focus is placed on human–environment interactions and cultural aspects of the environment such that research is driven by questions about human behavior, activities and practices, and related social phenomena. A key aspect of geoarchaeological research today is that consideration of these issues is highlighted in research design from the outset.

Geoarchaeology has become a subdiscipline recognized as necessary to any study of the human past (Goldberg and Macphail 2006, ix). Archaeologists realize the importance of understanding the context within which the material culture record is found and understood. Environmental scientists, ecologists, geologists, and geographers likewise realize the role humans have played in landscape creation, modification, and management over millennia and the importance of reconstructing human–environment interactions today and in the past. With the realization of the long-term impact that people had on the landscape and issues of climate change, environmental sustainability, and resilience being on the forefront of scientific research today, the contributions of geoarchaeology as a science based on human–environment interactions over time cannot be understated. Recognition of the importance of the subdiscipline within anthropological research comes in many forms. These include the creation of several interest groups within larger archaeological or other
science societies, such as the geoarchaeology interest group of the Society for American Archaeology, the archaeological geology division of the Geological Society of America, and the developing international geoarchaeology (DIG) group. Geoarchaeology-related research features prominently on the websites of the National Geographic Society, Royal Geographic Society and the US Geological Society. The establishment of geoarchaeology laboratories within university departments is becoming commonplace, such as Boston University’s microstratigraphy laboratory, the Charles McBurney laboratory for geoarchaeology (University of Cambridge), the geoarchaeology laboratory at the University College London, the scientific archaeology research group at the University of Reading, the laboratory of geoarchaeology and soils at the University of Haifa, and the geoarchaeology laboratories of Washington State University, University of Washington, Washington University at St Louis, University of Arizona, and the University of California, Berkeley. Annual workshops, such as the international soil micromorphology working group, University College London’s archaeological soil micromorphology training course, and the annual International Symposium on Archaeometry, provide training and conference venues for geoarchaeological research. Finally, there are now several peer-reviewed publication venues for geoarchaeology-focused research, including *Journal of Archaeological Science*, *Environmental Archaeology*, *Archaeometry*, *Journal of Human Evolution*, and *Geoarchaeology*.

**Historical background and modern perspectives**

As a subdiscipline of environmental archaeology, geoarchaeology historically sat at the crossroads between the earth sciences and archaeology. However, early geoarchaeological research followed a somewhat different trajectory than other types of environmental archaeology, such as archaeobotany, which always kept human activity at the forefront of research design. Borrowed from geology, early focus was placed on stratigraphy and aimed at clarifying relationships between and among strata at an archaeological site to establish the contemporaneity of deposits and material culture associations, as well as identify various site-formation processes in order to provide an environmental backdrop to culture change. Early practitioners of geoarchaeology were typically earth scientists or archaeologists with some earth sciences training, brought into archaeological projects already underway to answer specific questions about sediment matrix, stratigraphy, and natural processes operating on the archaeological record. Therefore, early conceptions of geoarchaeology focused on the application of earth science techniques to the solution of discrete archaeological problems. Essentially, it was archaeology done with the use of geological methods.

Early practitioners of geoarchaeology came to the field from archaeology, geography, geology, environmental science, and pedology and their research reflected these divergent backgrounds and covered a range of topics. In North America, early archaeology was heavily influenced by a number of scholars who bridged both geology and archaeology. Ernst V. Antevs was an American glacial geologist largely responsible for compiling United States and Canadian varve sequences that comprise the varve chronology used to reconstruct Pleistocene glaciations (Antevs 1954). Beyond this, he was also acutely interested in the relationships between glacial episodes and the peopling of North America. Along with K. Bryan and E.H. Sellards they instigated a long tradition of research on early
human–environment relationships and the study of early Palaeoindian sites in America (Haynes 1990).

Much of this early work was focused on chronology and stratigraphy. In the 1950s and 1960s, earth science approaches were being better integrated into archaeological lines of inquiry beyond dating and identifying deposits. While developments in dating techniques were proving revolutionary, archaeologists were also focused on provenance and petrographic studies for examining lithic and ceramic raw materials to address production and trade and exchange. Theoretical stances were also being taken on the necessity of integrative collaboration between geologists and archaeologists (Rapp and Hill 2006). Researchers began to explore the role of earth science approaches to understanding the different contexts of archaeological material and sites and for reconstructing human adaptations to changing environments. This work went under the guise of many names, including archaeological sedimentology, archaeological geology, geological archaeology, Pleistocene ecology, Quaternary geography. However, the goals were always the same – relating archaeology to the environment over time (Rapp and Hill 2006).

Beginning in the 1960s, C. Vance Haynes Jr brought geoarchaeology to prominence in the American Southwest and has remained active in research throughout North America ever since. Early geoarchaeological work in Europe included the study of Pleistocene geochronology by F.E. Zeuner (1959) at University College London. Ian W. Cornwall’s work on the interpretative contributions of sediment and soil analyses from archaeological sites was one of the earliest systematic attempts at geoarchaeology (Rapp and Hill 2006). In Soils for the Archaeologist (1958) Cornwall argued for the importance of context within the framework of understanding the sedimentological and soil processes responsible for the creation, modification and preservation of the archaeological record. Taking this even further, his work is considered by many today as revolutionary for the introduction of micromorphology, then a strictly pedological method developed by W. Kubiena (1938), into archaeology (Cornwall and Hodges 1964). Francois Bordes brought the study of geological context to Palaeolithic archaeology in France (Bordes 1954).

From the 1960s to the 1980s, archaeologists were concerned with understanding the depositional contexts of sites. However, it was the work of Karl Butzer, C. Vance Haynes Jr, G.W. Dimbleby, and a few others, who worked towards developing programs of research focused on geoarchaeological questions, as well as developing the appropriate methods and techniques to undertake this research (Goldberg and Macphail 2006). Haynes, appointed jointly in the departments of geology and archaeology at Arizona State University, continued to do much to improve our understanding of American Southwest Palaeoindian geochronology. G.W. Dimbleby (1962) reconstructed past vegetation and soils to understand Bronze Age monument construction. Working on valley fill sequences in the Mediterranean Basin and palaeosols at prehistoric sites in Italy, respectively, C. Vita-Finzi (1969) and M. Cremaschi (1987) explored the interrelationships between settlement and land use and landscape change. The work of Davidson and Shackley (1976) is particularly relevant as they are often credited as the first to pull together various aspects of geoarchaeological research under one name and published their approach in the book Geoarchaeology: Earth Science and the Past. Since the 1960s, Karl Butzer has focused on human–environment interactions in

Geoarchaeology today is a multidisciplinary approach to understanding the archaeological record, its environmental context, and site formation processes. Taking a broad approach to human landscapes and archaeological sites, geoarchaeology has come into its own with microscale approaches to the archaeological record. Explorations of microstratigraphy (Matthews et al. 1997; Matthews 2012) and microarchaeology (Weiner 2010) now highlight the invaluable contributions of geoarchaeology to reconstructing daily activities, individual events, and cycles and episodes of use and re-use of features at levels of detail unknown in the discipline previously. As a result of the new, integrative questions geoarchaeologists now address, rather than being a rare specimen who managed to navigate the blurred intersections between the hard and social sciences, geoarchaeologists are growing in number and many academic programs offer training in geoarchaeology from the undergraduate level. Goldberg and Macphail (2006, 2) situate geoarchaeology as providing “... the ultimate context for all aspects of archaeology from understanding the position of a site in a landscape setting to a comprehension of the context of individual finds and features. Without such knowledge, even the most sophisticated isotope study has limited meaning and interpretability.”

Scales of analysis

Perhaps one of the most significant points to be taken from the definition of geoarchaeology provided above is that geoarchaeologists operate at a variety of analytical scales, often sliding between macro-level regional and landscape studies and micro-level scales of investigation that involve examination of features, deposits, particles, and even individual molecules and elements in high resolution (Figure 1). The techniques employed for each scale are somewhat different, but the research questions often remain the same. Notably, whether we are discussing macroscale or microscale issues, we are really interested in understanding dynamic human–environment interactions in both spatial and temporal dimensions.

Macroscale studies usually involve regional landscape studies, palaeolandscape reconstruction, identifying landforms, mapping site locations, reconstructing climate change, and so on. These studies often incorporate a range of techniques designed to detect and map archaeological sites and landscape features, such as remote sensing, geophysical prospection, and ground surveys designed to identify and sample landforms and depositional environments, map and define soils and soil catenas, and locate and characterize archaeological sites. With these techniques the geoarchaeologist can establish the relationships between deposits and human activity to track landscape transformations and land use and settlement change over time. This often involves creating local and regional stratigraphic relationships between locales within a survey frame and correlating these deposits across space. With a theoretical shift towards human impact on the landscape, geoarchaeology intersects with historical ecology (Crumley 1994; Balée 2002) and human eco-dynamics (Kirch 2007).
Micromorphological analysis of archaeological deposits requires analyzing deposits at a variety of scales, from documentation and sampling of sections in the field (a) to the description, analysis and interpretation of thin-section slides (b) at a variety of magnifications (c) (Tell Seker al-Aheimar, Syria; images courtesy of Y. Nishiaki, University of Tokyo Museum).

**Figure 1**

Intermediate and microscale analyses involve investigation at the site level and below, such as individual archaeological features (i.e., hearths), particles contained within features (i.e., ash), and even molecular and elemental levels of analyses (i.e., elemental compositions of obsidian). Here there is overlap in techniques used at the macro and microscale, including the use of geophysical prospection to locate features within a site (e.g., houses, burials), sedimentological and pedological analyses of on-site deposits to establish stratigraphy and to assess preservation potential of deposits. The recent book *Microarchaeology* by S. Weiner (2010) is an excellent example of the sliding microscales of analyses performed by the geoarchaeologist, amongst other specialists. Similarly, W. Matthews’ (2005, 2012) and Twiss et al.’s (2008) approach to building construction, maintenance, and abandonment at Çatalhöyük provide a benchmark for the kinds of high-resolution interpretative power of microstratigraphic analyses, as does the examination of Paleolithic bedding material at Sibudu Cave, South Africa (Goldberg et al. 2009).

**Methods and techniques**

Geoarchaeology is both a field and laboratory-based endeavor. Geoarchaeologists employ methodologies and techniques from a wide range of fields, including archaeology, geology, geography, geomorphology, ecology, pedology, climatology, geochronology, geochemistry, and geophysics. Geoarchaeological methods can also include the reconstruction of past human activity through the analysis of biological datasets, such as archaeobotanical remains or marine and freshwater shells. Integrated organic and inorganic datasets allow palaeocological reconstructions that shed light on past human land use and practices, and landscape transformation at the
GEOARCHAEOLOGY

hands of humans. Many geoarchaeological tech-
niques focus on examining the composition and
origin of material culture, including ceramics,
metals, lithics, and building materials (e.g.,
plaster, mudbrick, adobe, stone, and cement).
Those who employ geoarchaeological methods
and techniques include researchers in university
settings, museums, and governmental agencies,
cultural resources management companies, and
private sector geoarchaeology consultants.

Perhaps one of the most long-standing and
well-known applications of geoarchaeology is
identifying, documenting, and analyzing strati-
graphy, sediments, and soils. Geoarchaeologists
make a clear distinction between sediments and
soils and, thus, require a firm understanding of
the different processes responsible for each and
how each interacts with and impacts the archae-
ological record. For sediments this includes
knowledge of the processes of sediment produc-
tion, erosion, transport, and deposition in order
to observe and record their characteristics and
as a means to understanding landscape history
and depositional environment (Goldberg and
Macphail 2006). As clastic, chemical, and bio-
logical sediments are all relevant to reconstructing
the contexts of human activity, geoarchaeologists
study colluvial deposits, riverine, estuary, and
alluvial deposits, aeolian environments, coasts,
and caves and rock shelters. Documentation
and interpreting sediments involves assessments
of their composition, color, grain size, shape,
roundness, sphericity, sorting, density, surface
texture, fabric/packing, and structure (bedding,
bedforms). Field techniques include trench and
column sampling, augering, coring, and archae-
ological excavation. Laboratory techniques
include particle size analysis (granulometry),
bulk density and mineralogy, organic matter (loss
on ignition), calcium carbonate content, iron
or magnesium content, magnetic susceptibility,
phosphate content, trace element geochemistry,
clay mineral analysis, and analyzing biological
and bioclastic inputs (e.g., ostracods).

Soils are usually formed over long periods
and conditions that entail some degree of land-
scape or landform stability and, thus, can prove
extremely important for understanding human
settlement and interpreting human–environment
interactions. Geoarchaeology thus often includes
a basic understanding of the processes of soil
formation and how humans can complicate
these processes, including anthropogenic inputs
and impacts. Activities such as terracing slopes
to control erosion can deepen soil profiles and
increase moisture content, and stock manage-
ment in highland soils can alter soil catena
development. In many past societies, especially
agricultural ones, soils were essential resources
that were mined, managed, and modified in a
variety of different ways. They can be seen as a
key factor affecting human occupation and land
use and, in some areas, particular soils were likely
foci for early farming activities. Hunter-gatherers
were also impacted by variations in soil, which
affected settlement location and instigated land
management practices. Palaeosols have some-
times proven useful indicators of past land
surface stability and can serve as likely indicators
of human occupation potential (Maher 2011).
Beyond the exploration of human–environment
interactions, soils, like sediments, can be involved
in the preservation of archaeological sites. Field
and laboratory techniques for soil analyses often
overlap with those conducted for sediments,
but also include the identification and docu-
mentation of soil profiles, soil (ped) structure,
boundaries between soil horizons, as well as the
analysis of pH, fertility (cation-exchange
capacity, N, K, and P content), sesquioxide
accumulation, and magnetic susceptibility.
Methods include the chemical, elemental and
mineralogical analysis of samples through X-ray
diffraction (XRD), electron microprobe analysis,
X-ray fluorescence (XRF), instrumental neutron activation analysis (INAA), atomic absorption (AA), Fourier transform infrared spectrometry (FTIR), Raman spectrometry, and inductively coupled plasma atomic emission spectrometry (ICP-AES). Micromorphology has proven to be indispensable to geoarchaeology for elucidating the complex relationships between sediments, soils, and anthropogenic activities.

Palaeosols can overprint archaeological stratigraphy and can preserve archaeological sites and, conversely, their formation can be sped up or slowed down by anthropogenic activities. An excellent example of the contributions of a geoarchaeological approach to soil analysis and the impact of anthropogenic input into soil forming processes is the study of dark earth (Macphail 1994; Macphail, Galinié, and Verhaeghe 2003), an organic-rich soil found at many Medieval archaeological sites in Europe. Despite being found in archaeological-rich urban contexts, these soil deposits exhibited a complete lack of stratigraphy and the processes responsible for their formation were largely unknown. Detailed microscale analyses of these deposits through micromorphology have shown them to be anthropogenic soils formed over very short periods of time as the addition of high levels of organic matter compounded by microfauna activity (namely, earthworms) accelerated the processes of soil formation and obliterated all pre-existing archaeological stratigraphy.

Characterizing on-site and off-site deposits allows the geoarchaeologist to identify and disentangle natural and cultural processes operating on the archaeological record and establish stratigraphic relationships between deposits, including correlating them across a site or several sites in a landscape. Together, the analysis of soils and sediments from on-site and off-site contexts provides a means for reconstructing palaeoenvironments and palaeolandscapes, identifying the possible locations of archaeological sites (by identifying landforms likely to contain archaeological material of particular ages), identifying human impacts on a landscape scale, and characterizing deposits and features within sites at high resolution. These direct and proxy data are powerful tools for contextualizing human activity across space and time.

Remote sensing and mapping

As technology becomes more sophisticated, portable, accessible, and affordable, field techniques for detecting and mapping landforms and sites are greatly improved and now almost standard preliminary survey methods. Remote sensing and satellite imagery, including open source platforms such as Google Earth, have aided immensely in landscape characterization and targeting specific landscape features (Comer and Harrower 2013; Ur 2006). Geophysical prospection, namely ground penetrating radar (Conyers 2013), electrical resistivity, and magnetometry (Aspinall, Gaffney, and Schmidt 2009), to name but a few, have revolutionized the way landscape potential is assessed and features located within sites to help target and refine excavations. Geographical information systems (GIS) are proving to be an invaluable tool for building models of land use and landscape change and developing and testing specific hypotheses regarding human settlement patterns and landscape modification (Conolly and Lake 2006).

Dating techniques

Chronological assessment and control, both in terms of relative (e.g., stratigraphic correlations) and absolute (e.g., radiometric methods) dating
have long been a pursuit of geoarchaeology (Aitken 1990). Radiocarbon dating, uranium series dating, K–Ar dating, optically stimulated luminescence (OSL), thermoluminescence (TL), electron spin resonance (ESR), and palaeomagnetism, in particular, are useful for archaeological deposits over a wide range of ages, from the very recent (Kirch and Sharp 2005) back to fossil-bearing deposits over one million years in age (Liritzis, Singhvi, and James 2013). Commercial and university-based radiocarbon laboratories, such as the Oxford Radiocarbon Unit, Weizman Institute, University of Arizona, and Beta Analytics, permit high-quality dates to be obtained at affordable prices with quick turn-around times. With interest in the age of landscapes, the age of sites contained within them, and the age of components within sites, geoarchaeologists employ dating techniques that date a wide variety of different types of materials at different resolutions (with different errors). All of these techniques require an in-depth understanding of archaeological context, laboratory preparation techniques and analysis, and how to interpret the results. With this knowledge, the quality of the sample collected and its appropriateness for a particular research question can be determined. In particular, it must be ensured that the sample material and technique used will not result in a date with uncertainties (errors) that are larger than the chronological resolution needed to solve the research question.

**Microscale analyses and daily practices**

Sometimes traditional geoarchaeological methods are not appropriate for archaeological deposits on a small scale. With developments in analytical techniques that permit examination of increasingly small samples, it is now possible to take microscale approaches to the archaeological record that can detect and identify human activity at the level of individual events. Rather than only assessing broad patterns in human activity and behavior within a site or feature from sedimentary or soil units, specific events, repeated activities, and daily practices within these individual units can now be discerned (Boivin 2000; Matthews et al. 1997). These microscale approaches fall under many names, including microstratigraphy, micro-artifact or microdebris analysis, micromorphology, and molecular analyses (such as isotopic analysis or residue analysis). These microstratigraphic approaches are unique in their ability to reveal specific human activities, such as in situ occupation, construction, and maintenance of floors, hearth cleaning, and dumping.

In particular, the introduction of soil micromorphology to archaeology has revolutionized the way archaeologists approach microscale analyses. Micromorphology is the examination of the composition, texture, structure, and arrangements of sediments, soil, and archaeological deposits at a variety of scales, mostly of high resolution. It involves the collection of undisturbed blocks, approximately 10×10 cm in size, of a deposit of interest. These blocks are then impregnated with a hardening agent (often a clear epoxy resin) in order to allow the blocks to be prepared into thin sections. Thin section slides are usually 5×7 cm in size (although this varies) and 30 μm thick, and are then examined and described at a variety of magnifications (Figure 1), often with binocular and petrographic microscopes. Like with artifacts, the context, distribution and composition of archaeological sediments is informative. With this technique, the geoarchaeologist considers sediment as material culture for the information it can provide on human activities, behaviors, and practices. A micromorphologist examines the composition...
and character of the coarse and fine fractions of a deposit, voids, organic matter, pedogenic activity, and anthropogenic input or traces. The advantage of micromorphological analyses over field observation of anthropogenic deposits is that human activity can be identified and examined at a completely different scale, often invisible in the field. Focus is placed not just on the identification of individual features or components within a slide, but on understanding the arrangement of these components with respect to each other for the information they can elucidate on the nature and relative sequence of processes of deposition, erosion, post-depositional alterations, and human activities (i.e., individual episodes of floor re-plastering).

Of particular relevance to archaeology, micromorphology allows one to look at occupational deposits, material culture, patterns in the use of space and object discard, and site-formation processes at a very high resolution usually invisible in the field. Micromorphology allows a high resolution view of time and space for understanding the activities of people within houses and settlements, including daily, seasonal, annual, and life cycle timescales. Building phases, levels, and repeated daily activities can be examined in more detail. Beyond the site, micromorphology contributes to studies of landscape and landform change due to human impact (e.g., farming practices like irrigation, manuring of fields, intentional burning, deforestation, overgrazing). In essence, it can provide a variety of types of information on human activity at a variety of scales. It has proven particularly useful for identifying and interpreting anthropogenic deposits, such as the construction, maintenance, and abandonment of buildings (resurfacing) (Matthews 2012; Rosen 1986), activity areas (food processing, flint knapping, sleeping, storage, middens), and changing use of areas/features over time (clean versus dirty spaces; sweeping) (Matthews et al. 1997), combustion features (Albert, Berna and Goldberg, 2012; Berna et al. 2012; Goldberg et al. 2012), bedding and matting (Wadley et al. 2011), human-induced modifications to soil regimes (agricultural practices) (Beach et al. 2006; Fisher et al. 2003; Macphail 1990), and site-formation processes (Rosen 1989; Tsatskin and Nadel 2003).

Key issues/current debates

Perhaps one of the most visible and common issues in geoarchaeology, as with any archaeological research design, is the issue of sampling strategy – deciding where, how, and how much to sample. Deciding on an appropriate sampling strategy depends on the research questions and is an issue at all scales of analysis, from regional landscape studies to reconstructing building phases or maintenance within one house. A golden rule of sampling provided by Matthews et al. (1997, 285) is that “in order to understand the complex pre-depositional, depositional, and post-depositional histories of occupation deposits, it is important to adopt research designs which raise these questions during excavation, recording, and sampling, and which enable integration of a range of field and laboratory characterizations, each of which has its own advantages and limitations and potential for different ‘readings’ of macroscopic, microscopic, chemical and physical attributes.”

Given the time-consuming nature of many analytical techniques, the key issues regarding sampling include decisions on the extent and intensiveness of horizontal versus vertical coverage, whether to sample all deposits or only key deposits, and whether to collect bulk samples or in situ, undistributed blocks (e.g., for micromorphology). These decisions have significant impact on the interpretation of archaeological
deposits, which can vary considerably spatially and temporally. While bulk sampling can be conducted expediently in the field, allowing coverage of a large portion of excavated area at a site or across a landscape, micromorphological sampling in the field is very time intensive and more destructive of archaeological sections and deposits (especially notable for small, thin deposits). This means that micromorphological sampling is often restricted to particular locations and deposits and often cannot intensively cover a large area across a site. However, the resolution afforded by micromorphological analysis often outweighs the time constraints and destructiveness of the technique in the field. Sampling strategies will also vary depending on the particular nature of individual deposits and the archaeological questions related to them. Bulk samples (for soil/sediment analyses, pollen or phytoliths) require samples to be taken within individual strata, while micromorphologists are also often interested in capturing and analyzing the transitions between strata. Thus, micromorphological sampling often requires a combination of systematic column sampling and purposive sample selection.

Another key issue faced by geoarchaeologists is ensuring consistency and continuity between field observations and microscale sampling and analyses. There are currently no accepted standards for presenting and publishing geoarchaeological data, but practitioners have over time developed certain guidelines to good practice for reporting and publishing results (Goldberg and MacPhail 2006). Full documentation must be ensured throughout every scale of sampling as it must be possible to clearly link each scale of analysis in order to maximize the interpretative power of each sample (Figure 1). Sometimes the sampling strategies of archaeological excavation must, thus, be modified to accommodate geoarchaeological sampling requirements that usually entail collecting more sediment/soil samples than might otherwise be done. This attention to consistency and continuity is carried into the reporting and publication of geoarchaeological datasets. Work must be presented in a clear and logical manner, readable and interpretable to the nonspecialist, and providing the reader with the possibility of tracing the exact locations of samples and sources of individual data (Goldberg and MacPhail 2006). This includes everything from including full field documentation details to ensuring scales are present on all images. Geoarchaeological interpretations are rarely, if ever, stand-alone datasets and, thus, should be written in consultation with other archaeological specialists (as the original research design was put together). Just as context is important to the geoarchaeologist, contextualizing geoarchaeological datasets with other lines of evidence provides the most robust conclusions and interpretations.

Future directions

With a focus on human–environment interactions over a great time period, and increasing global concerns with climate change, resource sustainability and ecological resilience, geoarchaeology has much to contribute to understanding of the long-term impact of human activities. Burgeoning areas of geoarchaeological research include the incorporation of historical ecology, sustainability studies, and resilience theory approaches to the environment. The interdisciplinary, multiscalar nature of geoarchaeology is well suited to tackling these issues in conjunction with many other environmental scientists. Particular contributions include examining the landscape modification brought on by long-term occupation (Beach et al. 2006; Matthews et al. 1997; Matthews 2010), small-
and large-scale land management practices (Fisher et al. 2003; Glaser 2007; Macphail 1994) and the development of experimental geoarchaeology and ethnoarchaeology (Cabanes, Weiner, and Shahack-Gross 2011; Friesem et al. 2011; Goldberg and Macphail 2006; Goodman-Elgar 2008; Matthews, Hastorf, and Ergenekon 2000).

SEE ALSO: Agricultural environments; Ancient geography; Anthropocene and planetary boundaries; Climate adaptation/mitigation; Climate change and biogeography; Climate and societal impacts; Desertification; Domestic spaces; Environment and urbanization; Global environmental change: human dimensions; Identity; Landscape; Paleosols; Quaternary geomorphology and landscapes; Scale; Soils in archaeological research; Soils as relative-age dating tools; Soils and weathering

References


Geocoding

Daniel W. Goldberg
Texas A&M University, USA

Geocoding is most commonly considered to be the process of converting postal address data into geographic coordinates, for example converting text such as “123 Main Street, AnyTown, Texas 12345” into the latitude/longitude representation “−118.23, 32.01” (Rushton et al. 2006; Goldberg, Wilson, and Knoblock 2007). Geocoding techniques and geocoding systems have been in existence since their earliest invention in support of the US Census Bureau activities in the 1970s (Jaro 1984; O’Reagan and Saalfeld 1987). Today, there are dozens of commercial, open source, and federally supported geocoding services available to serve the multitude of scenarios, applications, and users (Zhan et al. 2006) that need to convert textual information with a locational component into a geographic representation.

Geocoding plays an important role in many business, governmental, and research settings ranging from analyzing patterns of crime to perform predictive policing and planning (Ratcliffe 2004), to analyzing the spread of disease cases to reveal unfolding outbreak scenarios (Higgs et al. 2007; Gao et al. 2008) and understand chronic conditions (Costello et al. 2009; Rushton et al. 2004), to providing the base data necessary for enabling location-based services such as “search nearby” functions in many smartphone and scientific application scenarios (Adams, Ashwell, and Baxter 2003; Boulos 2003; Dao, Rizos, and Wang 2002). In these and many more instances, the process of geocoding serves as the enabling technology that transforms large databases of locationally relevant textual information into digital geographic data usable within mobile, desktop, and enterprise applications, services, and analyses (Goldberg, Wilson, and Knoblock 2007; Hill 2006). Because of this critical role, both the science of geocoding technologies and the science of using geocoded data have undergone in-depth research over the past four decades or more since the technology was first invented (Rushton et al. 2006; Goldberg, Wilson, and Knoblock 2007).

The purpose of this entry is to describe the past, current, and future of geocoding technologies. It begins by describing the internal processes of a geocoding system. It next describes the means by which users can interact with geocoding systems to process their data in secure and convenient manners. It concludes with current and future issues related to geocoding technologies and the use of these tools for scientific, business, and administrative purposes.

Components of a geocoding system

In general, geocoding systems can be divided into six major components (Goldberg 2008). These are: (i) the input data provided to a geocoding system; (ii) the parsing and normalization processes which are applied to those input data; (iii) the reference data repositories which contain the geographic objects that link input data to real geography; (iv) the feature matching algorithms which link input data to reference features; (v) the feature interpolation algorithms which identify the location along or within a reference...
feature where the output should be placed; and (vi) the output data resulting from the process of geocoding.

Input data

The input data provided to a geocoding system are most typically postal address data; however, this is not a requirement (Whitsel 2006; Levine and Kim 1998; Wieczorek, Guo, and Hijmans 2004). Most geocoding systems available today can process many types of textual data that contain some form of location reference within them. These alternative textual inputs can include the names of cities or states alone (without a postal address included), the names of official or colloquial locations, and the names of businesses or buildings. With the rise of social media, and Twitter in particular, the geocoding systems of today must now handle a wider variety of possible input which must now include highly abbreviated and nonpostal address structured references to geographic locations (Davis et al. 2011). In the vast majority of cases today, the data provided to a geocoding system must take the form of a string of text characters. These data can be passed either as a series of parsed address components such as “123,” “Main,” and “Street,” or as a single nonparsed string “123 Main Street.”

Parsing and normalization

Upon receiving an input string of characters to be geocoded, the first step of the geocoding process is to begin to identify what pieces of the input data relate to what types or attributes of geographic features (Boscoe 2008). This process is known as parsing, or recognizing and separating each of the components of the input address into its geographic type and attribute. For a nonparsed address-style input such as “123 Main Street, Anytown, Texas 12345,” the parsing process would use white spaces and other text separators (commas in this case) to separate the input data into a series of tokens each associated with the type of geographic object and feature. This input address would be separated into the following components: Street Number=123, StreetName=Main, StreetType=Street, City=Anytown, State=Texas, ZipCode=12345. To accomplish this task, parsing engines may utilize deterministic approaches where finite state machines process each token of text against known databases of synonym terms with look-back and look-ahead scanning to determine the correct types for each of the components of the input data string (Hutton 1992). Other approaches include probabilistic methods such as neural networks and hidden Markov models (HMMs), which are machine learning techniques that rely on annotated training sets and produce statistically based classifications for geography and attribute types for each of the input data items (Blakely and Salmond 2002; Christen and Churches 2005; Christen, Churches, and Willmore 2004).

Following the assignment of each term in the input data to a geography and attribute type, the next step of the input data massage process is to convert the recognized attribute types into the standard values which are expected by the feature matching and reference data components of the geocoding process. This process, known as address normalization, converts raw input data values into the official standardized versions following an address standard (Fonda-Bonardi 1994). This process results in, for example, the street type component “Street” in “123 Main Street” being converted into the official abbreviation “ST” when following a particular address standard that contains this official abbreviation.
Applicable address standards include the United States Postal Service (USPS) Publication 28 (US Postal Service 2015), which defines the official standardized values for address components for postal delivery in the United States, as well as the data format utilized by the US Census Bureau’s Topologically Integrated Geographically Referenced (TIGER) data files (US Census Bureau 2015), and many others utilized by local addressing authorities. The differences between address standards typically relate to the level of granularity of attributes available. For example, the parcel files maintained by the Los Angeles (LA) County Assessor separate the article of an address where the USPS Publication 28 standard does not. For an address as “123 La Brea Ave,” the LA County Assessor would store the components StreetNumber=123, StreetPreArticle=La, StreetName=Brea, StreetType=Ave, whereas the USPS Publication 28 would include the pre-article and the street name as a single attribute and identify them as StreetNumber=123, StreetName=La Brea, StreetType=Ave.

Reference data

The reference data sources included in a geocoding system represent all of the knowledge that the geocoder has about geographic objects in the real world (Gilboa et al. 2006; Zandbergen 2008). These data sources serve to link input data to real-world geographic objects that have known locations which can be assigned to the output data from the geocoding process. Reference data sources included in modern geocoding systems encompass the full spectrum of administrative geographic data items such as countries, states, cities, streets, buildings, and all forms of named geographic entities. These reference databases include one or more point, line, and polygon geographic representations for each of these administrative geographic data types that can be matched to during the geocoding process.

The most commonly utilized reference data sets include street centerlines, parcel boundaries, address points, and building footprints (polygons) (Goldberg, Wilson, and Knoblock 2007). The accuracy of a geocoding system depends heavily on what types and qualities of geographic reference sets are available to the system, because these define the types of objects that can be matched to in the real world (Zandbergen 2008). For example, if a geocoding system only contains the centroids of cities in the United States, it will only be able to produce output matching to the geographic accuracy of city centroids across the United States. Instead, if a geocoding system contains building footprints for a city, this geocoding system could be capable of producing output to the level of building centroid, which is dramatically better than simply matching to the center of a city for the same input postal address.

In the United States, the US Census TIGER/Line files (US Census Bureau 2015) represent the most accurate freely available reference dataset available for geocoding purposes at no cost. This dataset contains the street centerlines for all streets in the United States and includes the address attributes necessary for geocoding, such as street name, street type, from address, and to address by side of street. A multitude of commercial options also exist which are capable of providing building footprint and address point data for the whole of the world, but in many instances exceed the cost deemed reasonable by users who need to geocode their data. The United States does not maintain a publicly available national parcel dataset such as those maintained by other countries around the globe, for example, the United Kingdom (Ordnance Survey 2015) and Australia (Christen, Churches, and Willmore 2004; Paull 2003). As
such, national parcel-based geocoding in the United States is limited to purchased datasets. However, parcel datasets are rapidly becoming very available at the county level across the United States, allowing geocode users to build regional parcel-based geocoding options where this is the case.

Feature matching

Once processed through the address parsing and normalization components of the geocoder, an input data item will have been broken into several components representing different types and attributes of geographic data objects which can be found in the real world. The purpose of the feature matching component of a geocoding system is to identify possible candidate reference geographic data items, also known as reference features, within the reference datasets available to the geocoding system, score the correctness of the various matches, and determine which of all matches should be treated as the best possible match (Hutton 1992). To accomplish this task, two approaches have been implemented within the published literature on geocoding systems to date: deterministic matching approaches and probabilistic matching.

Both approaches begin by querying the reference data sources available to the geocoding system to identify possible candidate matches. In many geocoding systems, this involves generating and issuing Structured Query Language (SQL) queries to the relational database management systems (RDBMS) which store the reference data layers. Text-based query techniques such as word stemming and phonetic algorithms such as Soundex are utilized by geocoding systems to generate queries to the reference data layers which can account for common misspellings found in address and other location data. These queries are issued to each of the reference data layers in the geocoding system and result in a series of zero, one, or more reference geographic features that have attributes which exactly or partially match the components of the input address (Goldberg 2008).

Deterministic matching approaches will assign a weight to each of the attributes of a geographic feature and penalize candidate matches when the values on the reference features do not exactly match those of input location requested by the user to be geocoded (Goldberg, Wilson, and Knoblock 2007). These weights and resulting penalization factors are used to assign a score, known as a match score, to each of the candidate matches. Probabilistic approaches take into account the distribution of terms in a reference dataset along with the co-occurrence of terms to compute a match between the input data and every reference feature in the reference dataset (Jaro 1984; Christen, and Churches 2005; Christen, Churches, and Willmore 2004). In both cases, the match score associated with a reference data feature describes the likelihood that the candidate is the appropriate and correct match for a given input data item. The candidate matched feature with the highest match score is then used to pick the most likely best matching candidate which is then selected for use to produce the eventual geographic output from the geocoding system. A user-defined minimum match score is typically used within a geocoding system to determine a cutoff point under which a candidate feature will be treated as a nonmatch and not considered as a candidate from which to compute an output location (Boscoe 2008).

Feature interpolation

Once a candidate reference feature has been selected as the most likely candidate, the feature interpolation algorithms within a geocoding system compute the ultimate location at, along, or within the reference feature that is returned
as the final geographic output of the geocoding system. The type of feature interpolation algorithm employed relies primarily upon the type of geographic object stored in a reference data layer, and secondarily upon any ancillary information that is available about the distribution of an attribute of interest at, within, or along the reference feature (Goldberg and Cockburn 2012). For example, if the reference data layer is composed of address points with a latitude and longitude (lat/lon), the feature interpolation process is typically to directly return the lat/lon associated with the reference feature.

If the reference data instead contain polygon representations for ZIP code boundaries, cities, or counties, a feature interpolation algorithm will need to take some approach to compute a lat/lon output from a polygon representation. Several approaches have been documented in the literature for computing the centroid of a polygon including bounding box centroid and geographic centroid, which both have benefits and drawbacks (Goldberg and Cockburn 2010; 2012). However, if, as in most use cases of geocoding, the location data to be geocoded are associated with the addresses of people, a third approach, which uses population distribution within a region to bias the output toward where the majority of the population live, may be a better option.

If the reference data are street centerlines (the most common dataset utilized in free geocoding systems) the address interpolation algorithm must locate appropriate output-based street address attributes associated with the input address and the reference feature (Bakshi, Knoblock, and Thakkar 2004). Street centerline interpolation, also known as linear interpolation and address-range interpolation, utilizes the odd/even, side of street, and start and end address associated with the street segment in order to compute a proportional distance from the start of the street to place the output. This approach suffers from a number of documented shortcomings included an over-estimation of the number of houses on any street segment, non-contiguous address ranges between contiguous street segments, and nonstandard address numbering schemes or odd/even address street-side distributions (Zandbergen 2008; 2011). When additional information about the number and orientation of houses on a street is available from ancillary data sources, address-range interpolation techniques can be augmented to produce more accurate outcomes (Bakshi, Knoblock, and Thakkar 2004).

Output data

Traditionally, the output data from a geocoding system has been limited to simple lat/lon pairs with metadata information describing the quality of the match. In the literature on geocoding, the metadata from geocoding systems has been defined as match rates, match scores, and match types (Boscoe 2008; Rushton 2003). Match rates describe the proportion of input data records that could be geocoded to any level of geographic output of a total number of records in the input dataset (Zhan et al. 2006). This metric is an overall representation of the success rate of the geocoder, on a dataset-specific scale.

As discussed above, a match score is a value that describes the likelihood that a matching candidate feature is in fact the correct match for an input address. These scores are quantitative measures of correctness of a match and may be either deterministically or probabilistically based. In the former, the match score represents a measure of the similarity between the attributes of the input data item and those of the reference feature matched to. In the latter, the match score represents the probability that the correct selection was made from all reference features.
in a reference dataset. In both cases, the weights assigned for computing penalties and the distribution of attribute values in the reference dataset have enormous impacts on the resulting scores generated for candidate matches (Bichler and Balchak 2007). If incorrectly judged, it is easy for a geocode user to misinterpret the match score value and either accept false positive matches (matches that are not correct are included as correct in the output data), or reject false negative matches (matches that should have been included as correct in the output but are not).

The match type metric describes the level of geographic data object that a geocoding system is capable of matching to. The value that this metric takes for a geocode output is primarily dependent upon the type of reference data source that a candidate match is drawn from. If a candidate can be determined in a parcel data layer, the output match type in many systems will be “parcel”; if a candidate comes from a ZIP code layer, the match type will be “ZIP code,” and so on. Although the Open Geospatial Consortium (OGC) has developed many standards for geographic data, there are none that define the types of match types that are possible as output from a geocoding system. This situation has led to a proliferation of service- and community-specific match type schemes, with each geocoding service and community developing their own standards. For example, the North American Association of Central Cancer Registries (NAACCR) has devised a geocode match type scheme used throughout the cancer registry and cancer research communities (Goldberg 2008). This “geocode coordinate quality (GCQ)” includes the most commonly encountered geocode output types ranging from address point to state centroid level geocodes in an ordered hierarchy with a clear ordering of better and worse geocode match types.

The primary drawback of such systems is that they imply that all geocodes of certain match types are better than all geocodes of another match type. For example, the NAACCR GCQ hierarchy implies that all ZIP code level matches are better than all city-centroid level matches. In urban areas where ZIP codes are small and cities are large, this hierarchy arrangement makes sense. However, in rural, less populated areas where ZIP codes are large and cities are small, this assumption breaks down (Goldberg and Cockburn 2010).

A secondary drawback of the match type geocode quality metric is that it assumes that all geographic objects of the same type are of the same quality. Following from the rural/urban example above, it should be clear that a ZIP code in a highly populated urban area will represent a smaller geographic area than a ZIP code in a sparsely populated rural area. Geographic objects of the same type vary in size across and between geographic regions, meaning that one qualitative measure of “ZIP code level accuracy” does not mean the same thing in two different areas (Beyer, Schultz, and Rushton 2008; Drewnowski, Rehm, and Solet 2007; Grubesic 2008; Krieger et al. 2002). In one area it may match to a highly precise, geographically small area, while in another it represents a large area and thus highly imprecise geocode output.

Geocoding use cases and considerations

As of this writing, end users wishing to utilize geocoding technology have an extremely rich set of options available for accomplishing this task. Geocoding tools are now built into classic desktop geographic information system (GIS) packages such as ArcGIS and Google Earth, and are readily accessible via web services provided by commercial firms (Google, Microsoft Bing,
Here.com), academic institutions (Texas A&M University), and governmental agencies (US Census Bureau). The primary differences between these systems are cost (free versus paid), reference data layers included (parcel, address point, building footprint, etc.), and privacy and confidentiality guarantees (Health Insurance Portability and Protection Act (HIPPA) compliant versus non-HIPPA compliant) (Armstrong and Ruggles 2005; Cho 2008). In general, geocoding services with highly accurate, commercially based, or proprietary reference data layers are not free, and many cloud-based commercial options cannot guarantee HIPPA compliance or other requirements such as that data must remain within the borders of the United States during transit and while being processed (Goldberg et al. 2013). However, to a casual user of geocoding with a small number of records to geocode that are not of a confidential nature, many geocoding options exist. The primary decision that these users face is what means they should utilize to geocode their data. These users can choose from an array of available desktop/web and batch/single geocoding options.

Desktop versus web geocoding

The first decision a user must face when choosing a geocoding solution is to determine if they wish to use a desktop or web-based geocoder. A desktop geocoder runs locally on a user’s machine and requires that they have a complete geocoding system available and installed including all necessary reference data. In this scenario, the user opens the geocoding application on their local computer, points the geocoding system to their file containing location data to be geocoded, and clicks a button to begin the geocoding process. The lead up to this process may include the construction of a reference data layer-specific geocoding instance that is built by preprocessing a series of reference data layers to make them accessible and usable by the geocoding system. Processing data on one’s own local desktop ensures that one’s data never leaves one’s computer, thus ensuring confidentiality and security. However, desktop processing can involve high costs, to purchase the geocoding software and a license for reference data, and may require a high degree of technical knowledge and ability to be able to get the geocoding system functioning properly. In most desktop geocoding systems, it is possible for the user to tune the parameters of the geocoding system to optimize processing to meet the specific characteristics of their input data.

Web-based geocoding services operate through a series of web applications and/or web-based application programmer interfaces (APIs) that allow systems running in one’s web browser or on one’s desktop to call a remote geocoding service to process their data. In these scenarios, one’s location data are transmitted over the Internet to the geocoding service provider who processes the raw input data and returns geocoded output data as a result. There are numerous options available for web-based geocoding, ranging from free to very expensive, each of which provides a varying degree of options for tuning the geocoding process options, ranging from no ability to set parameters to complete ability to set all parameters. The primary benefit of utilizing a web-based geocoding solution is that it removes much of the expertise required to build and maintain a geocoding system from the end user, placing this burden instead on the service provider. Other advantages include scalable architectures at the service provider ensuring rapid processing of large amounts of data. As noted, the primary drawbacks of web-based geocoding solutions are cost and security. Many of the most accurate web-based geocoding providers charge nontrivial fees for using these
services and may include additional terms and conditions which limit what the end user can do with these data following their processing. All data sent to web-based geocoding services are transmitted over the Internet, so security may be an issue if proper security procedures are not in place or not followed during data processing. For example, at a minimum, if using a web-based geocoding system operating over the hypertext transfer protocol (HTTP), that is, anything that is a webpage, the HTTPS (secure HTTP) protocol should be used to verify the identity of the sender/recipient and provide a minimum level of encryption while the data are in transit. For health data of any kind, the geocoding service provider should be HIPPA and HITECH (Health Information Technology for Economic and Clinical Health) certified, which means that appropriate security practices, policies, and procedures are in place to protect an end user’s data while in transit and while at rest at the service provider.

### Batch versus single geocoding

The next main consideration for most geocoding system users is whether to use a batch geocoder or to geocode addresses one at a time (Zhan et al. 2006; Goldberg et al. 2013). In the vast majority of cases, users will need to geocode a data file that contains more than one address, such as a list of the addresses of study participants or a list of business names and addresses within a particular industry. As such, batch processing services are typically of more use for end users than single data entry forms that process one address at a time. In batch processing systems, users either upload a database of records to a web-processing service or point a geocoding system locally installed on their computer at a database of records. The next step is to identify the fields in the input database to the geocoding system, so the system knows which fields contain the full address as a single field or which fields represent parsed portions of the input data. It is typical for these systems to require one field to be a unique key which can be used to associate an output geocode with an input record. Geocoding systems will process the file of input record either serially (each data item in a row, one after another) or in parallel to increase processing speed and system throughput. The output of the geocoding process can either be appended as additional attribute fields in the input database or as a separate output file that can be linked back to the original file via the unique key associated with each data record.

### Geocode correction

The final process undertaken in many geocoding endeavors is the manual review of geocoded results which may or may not include manual inspection and correction of geocodes with less than desirable output qualities (Goldberg et al. 2008). These processes are built into many geocoding systems, desktop and web, and automatically prompt an end user to review output results where geocodes could not be found with high match scores or where match types are typically considered of low quality (state centroids, city centroids, ZIP code centroids, etc.). In these systems, a user is presented with an input address, the current output geocode, and the other match candidates which were found but not selected (Boscoe 2008; Abe and Stinchcomb 2008). At this point, the user can choose to accept the current output, choose a different matched candidate as output, or perform additional research to attempt to understand and correct any errors with an input address that prevented it from finding a more suitable match.
If additional data sources are available, such as DMV (Division of Motor Vehicles) records, social security records, or voting records, research has shown that nongeocodable records such as Post Office (PO) Boxes can be transformed into valid postal address for individuals which can be successfully geocoded (Gilboa et al. 2006). Similarly, Internet searches are often capable of translating the PO Box address of a business into a valid mailing address which can be geocoded. These same Internet searches are also quite successful in translating named places such as nursing homes and apartment complexes into valid street addresses which can be geocoded.

Manual geocode correction has been shown in recent years to greatly improve the quality of geocoded data, but also has been shown to be extremely labor-intensive and thus expensive. These expenses are due to the fact that very little of the research process can be automated, and it often takes a bit of detective work to understand the history of a person, address, or place to determine what a correct address should have been. In all cases, it is advisable for geocode users to retain both the raw input address and the corrected address discovered through a research process.

Current issues in geocoding research and practice

Despite the fact that geocoding techniques and technology have been in continuous use for more than four decades, many challenges remain which must be noted by the geocoding developer and user communities. Many of these have been challenging issues since the initial development of the field of geocoding, while others have emerged in recent years as new services and systems have increased the user base and use cases in which geocoding is applied. The following sections provide a discussion of just a sampling of all pertinent issues.

Spatial accuracy

Of primary concern to most geocode end users is the quality of the geocoded data they receive from their geocoding system. Of all aspects of quality possible, these users are primarily interested in spatial accuracy, or the distance between the computed geocode location and the true location of the object on the surface of the Earth (Davis and Fonseca 2007). Large spatial errors in geocoded data mean that all analyses, services, and/or conclusions based on these geocoded locations may be suspect, resulting in poor services and ill-derived conclusions (Jacquez and Rommel 2009). As noted, geocode accuracy primarily depends on the reference data layers available to a geocoding system, that is, a geocoder with better reference data will result in better output geocodes. However, geocoding accuracy has been shown to be the result of many factors beyond this simple technical observation. Most important, research has shown that geocode accuracy is not stationary across space or type of person (Krieger et al. 2002; Karimi, Durcik, and Rasdorf 2004; Oliver et al. 2005). Specifically, data from urban areas will result in better geocodes than data drawn from rural areas. Equally important, research has demonstrated that location data associated with minority populations and those less well-off across any measure of the socioeconomic spectrum are more likely have geocodes of lower quality than those in nonminority and/or more well-off circumstances. Users of geocoded data must be cognizant of these facts when using geocoded data for research or practice, and must be equally aware that simply removing low-quality
GEOCODING

geocodes from one’s dataset is not an appropriate method of dealing with low-quality data.

Security and confidentiality

As noted above, web-based geocoding services in particular may expose researchers and/or practitioners who need geocoding to a host of security and confidentiality concerns. Most important, users of geocoding systems need to be aware that their data are being shipped to a third party for processing when they are using these services. Depending on one’s dataset, this may or may not be an issue, but it does at least warrant some consideration if for no other reason than the terms and conditions of some providers state that they will keep a copy of all data submitted to their services (Curry 1999).

There is wide acceptance that submitting patient medical information (test results, disease stage, etc.) to one of these services would break confidentiality and/or privacy rules unless the service has signed a Business Associates Agreement (BAA) with the user and the service provider has HIPPA and HITECH polices, practices, and procedures in place to guarantee the safety and security of the user’s data in transit to/from the service and at rest while being processed by the service (Armstrong and Ruggles 2005; Gittler 2008). However, there is less clear guidance when no information other than an address is submitted about the patient. In this case, no personally identifiable information has been submitted to the geocoding service, nor has any disease or other medical information about the person. The Institutional Review Boards (IRBs) at academic institutions currently have no clear policies as to whether these data are considered personally identifiable information or protected health information (PHI), since the only data being transmitted are an address with no personal information included. Users of geocoding systems should be advised to check with their local organization’s IRB or IRB equivalent prior to sending any data to web-based geocoding services.

SEE ALSO: Georeferencing: from texts to maps; Interpolation: areal; Spatial database

References


Since the advent of scientific computing, geographers have been engaged in building more sophisticated analytical methods and models. The so-called “quantitative revolution” of the 1960s and early 1970s paved the way for computers to become widely accessible, enabling a range of previously intractable problems to be studied across many research disciplines, including geography. At the start of this period, there were no geographic information systems (GIS) as such and no spatial statistics applications, so the practitioners of the day had to write their own code. The programming language Fortran and punched cards were the technologies of the day, and our forebears rose to the challenge and pioneered many new statistical methods and data structures that explicitly addressed the spatial nature of data – methods that are still used to this day. One result of this research activity was the development and ongoing use of several new spatial association methods. The aftershocks of this computationally-powered revolution can still be felt across the field of academic geography (an interesting account of the somewhat turbulent philosophical debates that followed is given by Macmillan (1997)). Geocomputation continues this tradition of leveraging the latest capabilities afforded by state-of-the-art computing methods and platforms.

Geocomputation, at its heart, is focused around two specific, sometimes overlapping challenges:

1. improving accuracy or predictive power by developing or adapting new approaches to analysis and modelling and
2. increasing the computational scalability of analysis and modelling methods.

The founder of geocomputation, Stan Openshaw (2000), was equally fascinated by both of these challenges.

For some researchers, the analytical inadequacies or size limitations of current GIS are still a challenge that must be overcome. And there are still weaknesses in the established statistical methods currently used, for example their ability to accurately model nonstandard distributions or to work robustly in the presence of noise (Gahegan 2003). Fueling this desire for greater predictive power and better scalability is the exponentially-increasing volume of spatial data at our disposal. Big data is by no means a new obstacle for spatial analysis; ever since the start of the quantitative revolution challenges have been faced based on data size and data complexity when compared to the performance of available computer platforms and analytical methods, and it is likely that they always will. Ultimately, it is the same technologies and needs that power both the supply and the demand side of big data!

The first international conference on geocomputation, hosted by the School of Geography at the University of Leeds in 1996, heralded the
launch of a new research agenda in geographical analysis and modelling (Openshaw 2000). A measure of the interest generated in this field is that the conference quickly became established as an international event (Otago, 1997; Bristol, 1998; Virginia, 1999; Greenwich, 2000; Queensland, 2001; Southamton, 2003; Michigan, 2005; Maynooth, 2007; Sydney, 2009; London 2011; Wuhan, 2013; Texas, 2015). There has been a steady production of new textbooks on geocomputation over this timeframe: from Longley et al. (1998) to Abrahart and See (2014).

As originally coined, geocomputation was not intended as a synonym for GIScience, nor for spatial information theory: these being somewhat orthogonal to the original and ongoing vision and having their own conference series and research focus. In its original sense, it is closer to spatial analysis, with a clear focus on spatial data, as shown in Figure 1. The focus is on computational geography – building on the best that scientific computing, machine learning and statistics have to offer, within the explicit context of geographical problems and data: hence the term “geocomputation.” However, over time, there has been a blurring of meaning, as geocomputational techniques have gained more mainstream appeal, perhaps also because many problems require a combination of analytical approaches to be successful.

**Challenge 1: Improving analytical insights**

Extending analysis and modeling capabilities beyond the established methods of spatial analysis has been a recurrent theme for geocomputation. Researchers have explored many newer analytical approaches, for example based on machine learning, agent-based simulation, and exploratory geovisualization (Brenning 2005;
An excellent summary of these new approaches is provided by Cheng, Haworth, and Manley (2012). In practice, most of the efforts are a translation of progress made in machine learning, simulation, artificial life and information visualization (infoviz) into an explicitly spatial or geographical context. This translation is not necessarily straightforward, given the complexities and subtleties of both the methods themselves and the geographical data to which they are applied (hence, the tongue-in-cheek definition of geocomputation by McNoleg (2008) as: “Geocomputation: the application of impossibly sophisticated methods to fiendishly complex datasets by hopefully optimistic researchers.”). For example, machine learning tools typically place a complex set of “controls” onto the behavior of a learning algorithm that subtly change the inherent bias towards accentuating different kinds of patterns or distributions. Selecting these parameters so that the algorithm performs well on a specific geographical dataset, with all of the subtleties that pertain, is a very complex task. The theory that would “map” the behavior of these new methods onto the understanding of geographic space, patterns, associations, and categories is still largely absent. It is precisely this gap in theory that Couclelis (1998) urges geocomputational researchers to address, and many have (Fischer 2006; O’Sullivan 1999). However, the most common approach to date is still trial-and-error, working methodically through a range of parameter settings whilst evaluating the change in predictive performance. Such an approach is common to many applications of machine learning and is an example of computing power being used as a substitute for theory.

It is worth noting that the focus of geocomputation to date has instead been on improving the predictive power of analysis, though not necessarily improving the explanatory power of any resulting model. Improvements to fitting of distributions typically require the use of more functions, or more complex functions, or both. So the cost of better prediction is often a more complex model that may be more difficult to understand, or to “mine” for useful generalizations. It could perhaps be argued that the spatial domain is complex and it is time we moved on from simple, over-generalizing assumptions.

Progress to date has yielded tangible improvements in the theory and applications of geocomputation: by reducing prior statistical assumptions, improving the tolerance to noisy data, and providing more flexibility in modelling distributions. There have been many published examples of significant improvements in predictive power, or some other form of accuracy gain when moving from established statistical methods to methods based on machine learning (Lasaponara et al. 2014; Beekhuizen and Clarke 2010).

However, it has sometimes been a challenge to gain acceptance for these methodological changes outside of the geocomputational community. Conceptually, researchers must move from a statistical mindset – where distributions are declared and understood in advance and any analysis will always produce exactly the same result – to a stochastic/heuristic mindset where distributions are constructed (learned) from data examples, not assumed, and where results can vary due to the nondeterministic nature of many machine learning approaches (often caused by randomized start conditions employed by the underlying algorithms). (The variation is usually not large (if it is, it usually shows the learning process is not yet configured reliably), but no such variation exists with most parametric statistical approaches. So a different means of reporting is needed, to allow for variation in the results, and a different expectation is needed on behalf of the researcher.)
Challenge 2: Scaling up analysis and modeling

There are many unsolved scientific computing problems embedded within spatial analysis. Often, the price paid for improvements in analysis methods is an increase in the computational complexity, requiring commensurate increases in processing power. Similarly, the drive towards using bigger and more complex datasets places increasing demands on data storage design and efficiency. However, almost all analysis in commercial GIS takes place within laptop- or desktop-sized computers, and much of it runs in a single processing thread, on a single central processing unit core. Even in research settings, most spatial analysis algorithms do not yet scale well to more complicated computing architectures, leading to hard limits on the data volume that can be addressed and the complexity of analysis tasks that can be supported (Wang 2013).

To achieve better performance or scalability, the underlying problem must be broken apart into smaller-sized processes that can be solved concurrently, that is, in parallel. Such decomposition can address either the algorithm itself (task parallelism) – for example, by distributing the algorithm (or parts of it) across a set of processors, or by distributing the data to be processed across a set of processors (data parallelism). Also worthy of note is data distribution design, the fragmentation of data across a set of distributed and fault-tolerant storage devices that can be accessed in parallel (see the entry on CyberGIS).

This approach now plays an essential role in managing large spatial data collections, such as the tile caches to support web mapping applications such as Google Earth (Zhang and You 2012).

Some problems are easier to decompose than others, depending on how tightly coupled the task is, that is, how much communication is needed during execution between the set of processes. Tightly coupled problems require specialized, high-speed interconnections between processing units to support intensive message traffic. At the other end of the spectrum are embarrassingly parallel (EP) algorithms; these are simple to parallelize because they have no internal coupling or coordination needs. A common example is a parameter sweep, where the same process is executed multiple times with a different set of initial values. The calculation of local spatial regression at multiple locations and scales is embarrassingly-parallel, so it can be moved onto a grid or cloud infrastructure with relative ease (Harris et al. 2010). Many problems fall somewhere in between, and may require a mix of decomposition approaches and some refactoring of the original algorithms in order to achieve a radical improvement in performance (Guan, Kyriakidis, and Goodchild 2011).

Even once these decomposition challenges have been solved initially they need to be periodically revisited to align with advances in: (i) computing architecture, (ii) algorithms and scientific computing libraries, and (iii) storage design and optimization.

Summary

Geocomputation is a conscious effort to explore new analytical techniques from the wider computer and statistical sciences – particularly the machine learning and high-performance computing communities – that can be applied to quantitative geography. Its aims are: (i) to enrich geography with a broader toolbox of methods that are more accurate because they are able to adapt to the geographical nuances of the specific data used to train them; and (ii) to address the computational scaling and decomposition needs that will allow geographers to move beyond artificially small and simple problems and work with a more realistic level of detail and complexity.
SEE ALSO: Big data; CyberGIS; Geographic information system; Spatial analysis

References


Geodemographic profiling

Paul A. Longley
University College London, UK

A person’s identity is defined in part by their conception and expression of their individuality, and in part by their affiliations to groups – for example, to nations, cultures, social classes, or even soccer teams. In each respect, geography tells us that a fundamental part of “who you are,” is “where you are.” National, regional, or other cultural affiliations link identities to a single or a small number of unique places on the Earth’s surface, whereas components of identity that arise out of social, economic, or demographic circumstances are likely to recur across multiple different locations. The study and classification of human identities thus corresponds to the two dominant traditions in the history of geography as a discipline: the idiographic tradition that emphasizes the unique shaping characteristics of places, and the nomothetic tradition that emphasizes the observable similarities of process outcomes upon different locations on the Earth’s surface.

Geodemographic profiling (or “geodemographics”) is the neighborhood-scale analysis of people by where they live. Traditionally, it has fallen into the nomothetic approach to geographic inquiry – emphasizing the shared social, economic, and demographic characteristics of different types (or classes) of neighborhoods, independent of their locations relative to unique places. As such, a type of neighborhood may be widely scattered across a territory or jurisdiction. Geodemographic profiles thus provide summary indicators of the commonalities of social structure that link different locations. Yet each neighborhood is embedded, from an idiographic standpoint, in a unique place on the Earth’s surface.

Geodemographic profiling systems are used by businesses organizations to understand consumer purchasing behavior, lifestyle, and social attitudes, with the goals of finding and retaining profitable customers and identifying the best locations from which to serve them. Public service organizations use the same kinds of systems to facilitate resource allocation decisions of public goods, such as law enforcement, schooling, or health care. Figure 1a presents the “super groups” and their constituent groups that make up the UK Output Area Classification (OAC; www.cdrc.ac.uk). This classification is based on 60 variables selected from the 2011 UK Census of Population (a framework data source that is open to reuse by anyone). The typology was derived using k means cluster analysis, which is a method of data reduction and data mining. Figure 1b presents an equivalent classification, derived from 2010 US Census of Population data.

Figure 2 presents two illustrative radial plots that each identify the 60 UK census variables clustered in order to create the classification shown in Figure 1a. The overall classification shown in Figure 1a comprises eight major classes (termed “super groups”), and these plots profile the values of the variables for two of them. It is clear that very different values of the 60 variables characterize these two super groups. Interpretation of plots, standardized cluster scores, and other diagnostics arising from the cluster analysis enables analysts to devise user-friendly summary “pen portraits” of the prevailing characteristics of each class within a
Figure 1  Open geodemographic classifications: (a) the UK 2011 Output Area Classification, showing part of Southampton (courtesy of Chris Gale; contains National Statistics and Ordnance Survey data © Crown copyright and database right 2014); (b) a classification based upon 2010 US census data, showing part of San Francisco (courtesy Alex Singleton).
Figure 2  Radial plots of the 60 census variables making up two of the eight super groups of the UK 2011 Output Area Classification: Multicultural Metropolitans (left) and Rural Residents (right). The plots show the very different profiles of these two super groups with respect to the census variables. Courtesy Chris Gale. Contains National Statistics and Ordnance Survey data © Crown copyright and database right 2014.

geodemographic classification. Two illustrative pen portraits from the London Output Area Classification (LOAC; www.cdrc.ac.uk) are shown in Box 1. Web mapping also makes the dissemination of geodemographic classifications accessible and readily intelligible to users, providing useful, detailed, and generalized summary information about the characteristics of human identity at the neighborhood scale of analysis.

Most geodemographic classifications include many more than the eight classes shown in Figure 1a and the nine shown in Figure 1b. Typically, each of the high level classes (“super groups” in the terminology of the classifications shown in Figure 1a) will be subdivided into lower-level classes (e.g., “groups”), and each of these may in turn be further subdivided (e.g., into “subgroups”), and so a national geodemographic taxonomy will often number 60 or more classes in total. These geodemographic classifications are said to be “open”: the methodology that was used to create the classifications shown in Figure 1 and Box 1 is published, the analysis method is fully documented, and the data that were used to create each of them are readily accessible to all (Gale et al. 2016; Spielman and Singleton 2015). However, many other geodemographic classifications are “closed”; that is, at least some of the data sources used to create them are not in the public domain, or undocumented and commercially sensitive methods were used in the data analysis.

A number of issues of spatial analysis underpin the creation and use of geodemographic classifications. First, it is assumed that the choice of variables used to specify social similarity has some intrinsic merit in representing the distinctive identities (defined in social, economic, and
**Box 1** Names and pen portraits of two of the eight super groups that make up the London output area classification and their constituent groups.

<table>
<thead>
<tr>
<th><strong>D Urban elites</strong></th>
<th><strong>E City vibe</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>This super group comprises young professionals working in the science, technology, finance, and insurance sectors. Additionally, large numbers of students rent rooms in centrally located communal establishments. Most others rent privately owned flats, also predominantly in central locations. White residents are disproportionately drawn from pre-2001 European Union (EU) countries, and there is also high representation of affluent households drawn from emerging markets or leading economies worldwide.</td>
<td>Young, single professionals predominate in this super group, living in Zone 2 of the London travel network. Compared to the London average, few individuals originate from the Indian subcontinent, but a full range of other ethnic groups is represented, particularly those drawn from pre-2001 EU countries. A large number of households comprise full-time students in shared or communal establishments. Individuals rent within the private sector, are well qualified, and are employed in a range of professional, scientific, and technical occupations.</td>
</tr>
<tr>
<td><strong>D1 City central</strong></td>
<td><strong>E1 City student fringe</strong></td>
</tr>
<tr>
<td>Many of the residents of these neighborhoods are employed in financial, insurance, and real estate activities, or are information and communications industry professionals engaged in a range of scientific and technical activities. Those that are non-UK citizens from the EU originate predominantly in pre-2001 EU states.</td>
<td>Many of this group are students living in communal establishments in some of London’s less fashionable central locations. Black British residents are much in evidence, and citizens of post-2001 EU countries are also well represented relative to the London average.</td>
</tr>
<tr>
<td><strong>D2 Educational advantage</strong></td>
<td><strong>E2 Graduation occupation</strong></td>
</tr>
<tr>
<td>Student lifestyles characterize this group, with residents living in centrally located communal establishments at high residential densities. Many individuals originate in emerging economies of the Middle East and China, and post-2001 EU accession countries are also well represented.</td>
<td>The residents of these areas are students, or recent graduates who are employed in the communications, scientific, and technical sectors. Pre-2001 EU countries are well represented, and travel to work on foot or by bicycle is more common than average for London.</td>
</tr>
</tbody>
</table>

Source: Author.
demographic terms) of different neighborhood classes, and that the analysts that created the classification have selected and weighted the variables commensurate with their relative importance to the classification. Second, there is an implied assumption that the pen portrait (Box 1) of any class represents the modal identity of households in the neighborhood and that, by implication, other households are either small in number or similar to the modal subgroup. Third, there is an assumption that the units of analysis, such as UK census output areas or US census tracts, are appropriate for capturing the spatial heterogeneity of neighborhood structure, either alone or through aggregation. Finally, users will assume that the classification is up-to-date when they use it, and thus that there is temporal stationarity between the time of data collection and when the classification is utilized.

History and current state of the art

The conceptual roots to generalized representation of neighborhood characteristics and differentiation between small neighborhood areas can be traced to the work of Park, Burgess, and other members of the “Chicago School of Urban Ecology” from 1916 onwards. The root metaphor to this work was biotic competition and vegetation succession in ecological systems. Seen from this perspective, neighborhood communities were fashioned by subsocial forces, and “society” was a cultural superstructure above the more basic competitive level of the community. By analogy to ecological systems, the dynamics of invasion and succession were community-level forces, as vividly captured in Burgess’s famous (1923) concentric zone model of Chicago.

Much of the enduring appeal of this work originates in Burgess’s ascription of explicitly spatial terms such as “neighborhood” and “district” to areas whose residents shared collective identities and which drove urban expansion and change. Thus, the city came to be envisioned in cross-section as a mosaic of subareas, defined in terms of the patterning of socioeconomic and demographic groups and differences in built form. The root biological metaphor provided the dynamics that underpinned this cross-sectional view, and was deemed to drive the kaleidoscope of change that characterized the Chicago metropolitan area in the 1920s and 1930s. This image is both vivid and enduring, with some writers (e.g., Saunders 2001, 42) suggesting that the ecological analogy retains greater relevance (at least in empirical terms) to our understanding of residential structure than subsequent models of urban differentiation, such as those in the urban managerialist or Marxist traditions.

Powerful and enduringly relevant though the dynamic ecological metaphor may have been, subsequent empirical work focused rather repetitively upon the measurement task, elevating detailed description of the urban mosaics of different cities to an end in itself. This humdrum, unimaginative, and normal science endured as “factorial ecologies” during the quantitative revolution of the 1960s, as advances in computing made possible reduction of large and complex datasets, such as those arising out of censuses of population, for the first time. But geographic science is itself driven by change, and the routine nature of such data reduction exercises meant that interest in factorial ecology waned during the 1970s, although not before the approach was used by the UK Department of the Environment (DoE) to devise a UK neighborhood classification as the basis for prioritizing different inner-city policy initiatives. By 1979 the DoE’s system was used by some fifty local authorities.

It was this applied policy analysis that provided the template for the generalized geodemographic
GEODEMOGRAPHIC PROFILING

neighborhood classifications that were developed in the following 20 years. However, analytical refinement in the specification, estimation, and testing of classifications became the preserve not of urban sociologists or geographers, but of marketeers working in large commercial organizations (Harris, Sleight, and Webber 2005). At first, these commercial applications were created exclusively using population census data, but they subsequently broadened to include data harvested from social surveys and other sources of business data. This development in data very much reflected the preoccupations of the new business applications over those of government policy and planning. The societal mandate for population censuses is very much in response to government needs for relevant indicators of deprivation and hardship, whereas the new commercial classifications were more attuned to discerning differences in consumption behavior within more affluent socioeconomic echelons.

A consequence of this changing locus of activity in geodemographics was that, in a period of rapid and accelerating development in computation, the principal refinements in technique largely took place outside of academia and government. Private sector sources of data, of largely unknown provenance, were used to supplement and partially replace framework population data sources such as censuses. The motivation for this was at least twofold. First, the somewhat stale diet of census variables was not conducive to segmenting increasingly sophisticated consumer-led markets. Second, the decennial nature of census data collection (along with delays in making the resulting data available) meant that classifications based exclusively upon census data were deemed out of date even before they appeared. It also served commercial ends that the new classifications were “black box,” in that the content and lineage of many of the constituent datasets were unknown, and the precise methods used to cluster them were closely guarded commercial secrets.

The archetypal end use of a geodemographics application in the 1980s and 1990s was to improve segmentation in retail markets, for tactical or strategic gain. Within academia, there was little attempt to develop new and innovative applications, and no clear vision of the ways in which neighborhood classifications were integral to linking and summarizing the increasing range of administrative and survey datasets. The few academics who maintained active research interests in business and service planning felt “locked out” of developments that utilized private sector “lifestyle” data to augment census sources. The inevitable consequence was that geodemographic profiling remained rather rudimentary in the academic and public sectors. The archetypical public sector application of the 1980s and 1990s was the deprivation indicator, which standardized and summed census indicators to identify areas that were experiencing hardship.

Whether devised in the public, private, or academic sectors, geodemographic classifications of this period remained focused upon cross-sectional outcomes – good marketing prospects and deprived communities alike were deemed to be static outcomes. There was little interest in representing the dynamics of change in the way the Chicago urban ecologists originally envisioned neighborhood transition and secular change. While geographers were interested in patterns of spatial interaction, these were usually framed using coarse zonal geographies, or did not develop sophisticated taxonomies of human identity or neighborhood type.

The early years of the new millennium saw a renaissance of interest in academia and government in the use of neighborhood geodemographic profiling. This was driven by the turn of the millennium, by a number of Western governments, to “evidence-based”
policies, improved access to and distribution of turn-of-millennium censuses, and the development of a performance culture in local public service delivery. The focus of the archetypical geodemographic application shifted to public service tasks, such as improving preventative communications programs (e.g., in health and policing) and the opportunities to improve efficiency by targeting neighborhoods most at risk from social ills. There was also renewed academic interest in the importance of local context in understanding the responsiveness of local communities to policy initiatives. By this time, however, the analytic skill base of many parts of local and central government, and of academia, had been hollowed out as a consequence of shifts in academic focus. Consequently, many public sector applications and academic studies were, in effect, consumers of the sophisticated yet black box classifications created in the commercial sector – which was fleet-footed in spotting the opportunity to rebadge existing classifications for new public sector clients. This raised a number of issues of research practice and ethics, some of which were discussed in John Pickles’s (1995) edited collection on the social implications of geographic information systems (see also Longley 2005). Commercial applications of geodemographic profiling are very much driven by predictive success in improved targeting of potential customers, and the black box nature of classifications remained of little concern in achieving this end. Yet all citizens are stakeholders of public services, and the motivations for social policy are much more broad-based; thus, any failure of commercial classifications to represent the interests of stakeholder groups in social organization presents important issues of ethics and accountability.

More recently, a number of governments have come to understand the value of public services and government in the creation of quality data. This could provide a much improved context to decision-making but hitherto has been neglected by (or even hidden from) potential users. The open data movements of many countries, worldwide, are leading to the greater dissemination of data collected about the public to organizations that represent the public. Open data from administrative sources are thus beginning to provide important elements of an inclusive socioeconomic data infrastructure, and hence for geodemographic classifications. This is a significant development, particularly given that the response rates to conventional sample social surveys appear to be in inexorable decline in many countries. This brings with it concerns about the nature of response bias in the data that are collected. On a more positive note, data from the most recent censuses suggest that they have not been vulnerable to this attrition effect (completion is a legal requirement in many countries), and they have been made more readily available for download than ever before. Although the limited remit of the short census form in the United States gives cause for concern about the accuracy with which a full range of population characteristics can be ascribed to small areas (using the American Community Survey) (Spielman and Singleton 2015), the future of censuses in the United States and United Kingdom does now seem assured.

Publicly funded data collection nevertheless now accounts for a reduced share of the totality of routinely collected digital data pertaining to household lifestyles – particularly when social media and transactional “big data” sources are taken into account. In the spirit of the private sector developments of the 1980s and 1990s, there is obvious potential for use of new data sources assembled in the course of business-to-customer transactions, and also through reuse of geographically referenced social media data such as Twitter. The provenance and
population coverage of these sources remains a very under-researched issue; much work remains to be done in order to open up access to such sources. Yet the potential benefits to geodemographic profiling are enormous, in terms of the range of variables that might become available, and the temporal granularity with which human activities might be measured.

Taken together, the availability of new and open sources of geographically referenced data is ushering in an era in which the creation of application-specific geodemographic classifications is becoming routinely possible for a broad constituency of spatial analysts. This is making it possible to build geodemographic classifications tailored to the widest range of end uses, using data that are detailed, pertinent, and up to date. Yet for this to make a significant contribution to evidence-based policy, data sources must be of known provenance, with sufficient endeavor to accommodate bias in their coverage.

Prospects: toward geotemporal demographic profiling

Viewed retrospectively, it is clear that the practice of geodemographic profiling of neighborhoods has been guided by two overriding considerations – one substantive, the second procedural. In substantive terms, nighttime residence has always been seen as key to generalized understanding of the relationship between human identity and the spatial organization of society. This assumption is increasingly outmoded given the fission of lifestyles in societies that are manifestly differentiated in many other ways, such as by measures of family structure and kinship, consumption, education, or employment. Residential structure is but one (albeit very important) indicator of social structure that needs to be considered alongside indicators of activity patterns on a daily, weekly, seasonal, or still longer-term basis. The nature and pace of change in the production and reuse of the built environment may make change monitoring important at greater frequencies than decadal census intervals. All of these issues will come to the fore as classifications for emerging economies are developed, and it is also likely that new big data sources (such as social media) will be particularly important in recording activity patterns in these new and dynamic settings.

The second, procedural, issue is the use of small area aggregations in order to preserve respondent confidentiality. This gives rise to the well-rehearsed arguments of ecological fallacy (confounding the characteristics of areas with particular individuals who live within them). As lifestyles become ever more differentiated, so patterns of individual behavior in space become increasingly central to defining what makes any person or group distinctive. From a geodemographic perspective, the zonal geographies of residence are by no means the only ones that shape identity at the level of the human individual, and a more appropriate representation of individual behavior should seek to capture more of the spatial trajectories and exchanges of information that are central to identity.

Some recent censuses have gone some way toward addressing these issues by providing small area workplace statistics, broken down according to a limited range of social, economic, and environmental variables. Although providing important local economic indicators, human identity encapsulates more than the duality of work and residence – many people do not undertake paid work, and for many others the rhythm of day-to-day activities is about more than the journey to and from work. It is thus increasingly the case that applications of geodemographic profiling need more information – for example, in order to measure and monitor the changing
convenience culture of retailing that continues to redefine the mission of retail store location analysts.

Geodemographics can rise to this challenge by incorporating the temporal dimension, at temporal scales from the dynastic to the diurnal. Over the most expansive timescales, biosocial research is providing context to the origins of populations and providing benchmarks for estimating the cumulative genetic effects of migration upon unique places over the entirety of human history (Leslie et al. 2015). While biological genotyping remains a prohibitively expensive means of establishing the biological uniqueness of places, family names have been used as a surrogate indicator of place effects, since many family names were first coined at unique locations (typically between the twelfth and fourteenth centuries in Anglo-Saxon cultures). Thus, Cheshire and Longley (2012) and Mateos, Longley, and O’Sullivan (2011) use Anglo-Saxon family names to analyze distributions of cultural, ethnic, and linguistic groups, present day and historic, at scales from the local to the international. Linkage of this work to biosocial markers provides valuable indicators of the cumulative effects of migration over the last 900 or so years, and hence of unique ancestry profiles of populations in different places. However, the efficacy of this approach is likely to depend upon the efficacy of naming conventions in representing the accumulated effects of language and culture in the period since family names came into common parlance.

This analysis provides context to the more recent and more generic patterning of social similarities that are measured using census and related survey data, and which have been overlain upon these unique historic biosocial landscapes. Bringing together the long-term biosocial place characteristics with the much more recent geodemographic profiling makes it possible to view patterns of social similarity in relation to historic place formation.

At the other end of the temporal spectrum, social media data are being used to capture the formation and sustenance of place effects through the measurement of social interactions. Thus, Birkin, Harland, and Malleson (2013) and Longley, Adnan, and Lansley (2015) use georeferenced Twitter data to capture the movements of individuals and draw conclusions about the sociospatial characteristics of the flows of people and information within cities. This evolving research agenda uses social media data to address issues such as:

- geographic variability in the connectedness of locations, and the creation of geotemporal profiles of (a) the connectedness of neighborhoods to others that are similar in terms of quantifiable characteristics and (b) the timescales on which people at a unique place become aware of, or react to, external events;
- diurnal, weekly or seasonal variation in the connectedness of systems of locations to one another;
- summarizing the locus of daily activity patterns, and the impacts of trip chaining behaviors upon place formation – such as the multipurpose trips of working parents, or the routing behavior of public services such as police patrols;
- the functional interdependencies between places in terms of volumes of interaction generated by different social, economic, and demographic groups.

What is defining about all of these applications is not just the infusion of the temporal dimension into representations but also their recourse to the human individual as the unit of enumeration and analysis. This is becoming central to understanding the short- and long-term dynamics of how
human societies and their settlement structures function. This also brings focus to human identity as shaping, and being shaped by, the information flows and population movements that define both unique and generalizable characteristics of places. It also brings sharp focus to ethical issues of disclosure control and anonymization in an age in which geographic information technologies facilitate the potential linkage of diverse data sources from the administrative, business, and service delivery domains.

The granularity of time-stamping that can be achieved through new sources of social media data is far greater than anything that hitherto has been incorporated into geodemographics. Less immediately apparent is that the coverage is very much less complete than has traditionally been the case with census-based analysis; sometimes heroic assumptions must be invoked at the individual level in order to assign attributes to individuals. The issue of coverage also applies to monitoring the biosocial characteristics of populations; cost issues restrict the size of samples and it remains unclear what constitutes, in practice, a “good” sample of the biosocial structure of any place.

From the perspective of geotemporal demographics, places are social constructs defined by the cumulative effects of highly distinctive interactions between population characteristics and environmental attributes over space and time. The fundamental challenge remains to filter the shared social similarities that can be generalized between different locations – as with the mosaics of urban areas identified by commercial geodemographics – while remaining cognizant of the uniqueness of different places when doing so. Place effects may reveal shared elements over the very long term (as revealed through biosocial analysis or migration history) or the very short term (as revealed through indicators of spatial interactions and activities gleaned from social media).

Geography tells us that widely scattered places (e.g., in Washington, California, Florida, or Illinois) share important social or physical similarities that are manifest in different ways and to differing degrees. Geotemporal demographic profiling offers the prospects of using standardized quantitative measures to measure the repetition of place effects across space, while also posing questions about what is unique, or unquantifiable, about places. The geotemporal profiling of places begs important questions, not just about data science and the potential of big data, but more fundamentally about the hypothesis-driven research designs that will underpin geodemographic profiles that are efficient, effective, and, above all, safe to use.

**SEE ALSO:** Big data; Corporate retailing; Geographic information system; Human geography; Qualitative GIS; Place; Regional geography; Social geography; Spatial context

**References**


Geodesign

Thomas Fisher
University of Minnesota, USA

What is geodesign?

Geodesign combines the analytical capability of geography with the synthesizing characteristic of design, giving those responsible for shaping the built environment the geographical information systems (GIS) they need to make more informed decisions. Geography and design may seem to have little to do with each other, with one rooted in the social and physical sciences and the other based in the arts and humanities. However, these two fields have a lot of overlap.

Consider their names. Geography means to write (graph) on the Earth (geo), studying the traces of the cultures, movements, and flows of humans across the world and their natural and built environments. Meanwhile, design means to mark (sign) out (de) something, which includes the spaces, structures, and paths that people have placed on the planet. Geography, in other words, has long studied the designs of human beings and the designs that they have had on the places we inhabit.

As such, geography and design have many traits and interests in common. Both fields, for example, have a strongly spatial focus, studying how people construct, use, and interact in space. Both also look at the activities and institutions of people around the globe, attending to cultural differences and community coherence. And both use many of the same analytical tools, from maps and plans to databases and documents of all sorts.

The significant distinction between geography and design – and the value of combining them in geodesign – lies in their different temporal focus. A good deal of geographical work helps us understand the world as it was and as it is, while design offers a way of envisioning the world as it could be. Geodesign, in other words, marries our grasp of the past and the present with a desire to imagine a better future. It also ties into related geographical endeavors, such as the development of spatial decisions support systems or multi-criteria evaluation tools, which use geospatial technologies to envision possible futures.

Another distinction has to do with the difference between studying a situation and changing it. That distinction comes out in a recent definition offered by leaders in the field, Carl Steinitz, Stephen Ervin, Michael Flaxman, and Tess Canfield at the Geodesign 2014 Summit: “Geodesign applies systems thinking to the creation of proposals for change and impact simulations in their geographic contexts, usually supported by digital technology.”

Why geodesign?

Several forces have prompted the emergence of geodesign. Technology has evolved to the point where GIS software and 2-D and 3-D visualization tools can now interact in ways not possible in the past. Geodesign tools not only give a data-rich understanding of particular places, but also provide an immersive way of envisioning and depicting the changes we want to make to those places.
At the same time, geodesign reflects the emphasis in recent decades on interdisciplinary interactions and the new knowledge that can come from intersecting fields once thought of as distinct. Technology has helped prompt this trend as the web has become not just a place to seek information, but also a metaphor for reality as increasingly the world is seen in networked ways.

A further motive for geodesign has come from a realization that we can no longer justify or sustain many features of the world that we have designed for ourselves. With the built environments in which most humans live having such a negative effect on global ecosystems, endangering our own species along with millions of others, the design community has recognized the need for a new approach and new tools that enable us to make better decisions based not only on short-term needs but also on long-term impacts, locally and globally.

The history of geodesign

Geodesign, in some sense, has existed ever since people first made decisions about what we do on and with the land. As Steinitz observes, “people have designed and changed the geography of their landscape for thousands of years, often without the participation of design professionals or geographic scientists” (Steinitz 2012, 9). Steinitz traces the more recent history of geodesign to the work of the landscape architect Warren Manning, who first overlaid maps on light tables as the basis for his design decisions. The landscape architect Ian McHarg greatly extended that process, described in his ground-breaking (1992) book Design with Nature, using map overlays to determine where and what type of development should go where. Geography in their hands became an essential design tool.

The intersection of that work with computers occurred in the mid-1960s in the Harvard Laboratory for Computer Graphics founded by Howard Fisher. There Carl Steinitz and Peter Rogers conducted some of the first digitally based geodesign work, using early computer graphics software to develop scenarios for the future of the Boston region. Jack Dangermond, a graduate student in the Harvard laboratory during that period, went on to combine his education in landscape architecture and computer graphics to establish ESRI, which has led the development of geodesign. Other geographers also contributed greatly to GIS and, ultimately, geodesign. They include Roger Tomlinson, who created the first GIS: the Canadian Geographic Information System or CGIS to inventory Canada’s natural resources, and David Dickmore in the United Kingdom, who pioneered computerized mapping. Goodchild (2010) notes that, indeed, one of the original, primary goals of GIS was to support better landscape design and planning that fits the mold of geodesign.

Geodesign’s leadership

Geodesign itself dates from the first decade of the twenty-first century, arising out of Jack Dangermond’s conviction that we needed “a framework for understanding the complex relationships between human-designed settlements and the changing environment” (ESRI 2013, 44). Since then, geodesign has flourished, with many people contributing to its evolution. In addition to Dangermond, several colleagues of his at ESRI, including Bill Miller, Shannon McElvaney, Eric Wittner, and Bernie Szukalski, have all played key roles in the development of geodesign tools.
In the design and computational community, Carl Steinitz remains the primary theorist of geodesign, along with colleagues and former students of his, like Tess Canfield, Stephen Ervin, Michael Flaxman, and Juan Carlos Vargas. Geographers such as Michael Batty, Michael Goodchild, and Diana Sinton have also influenced the development and direction of geodesign.

Architects, landscape architects, and planners have played a major part in the evolution of geodesign, as well. They include academics such as Thomas Fisher, Kelleann Foster, Tom Paradis, Ryan Perkl, David Pitt, Janet Silbernagle, Ron Stolz, and Paul Zwick, as well as practitioners such as Doug Olsen, Keith Besserud, and Elliot Hartley. That group, along with a growing number of regular participants in the annual Geodesign Summit, has helped define the field and expand its reach.

Geodesign Summits

The annual Geodesign Summits that ESRI has hosted, beginning in 2010 at its headquarters in Redlands, California, have served to gather the leading practitioners of geodesign, to enable them to share best practices, and to hear about the latest geodesign tools under development. What started as a relatively small group of geographers and designers has since grown to a much larger and more diverse group of people. By 2014, the Geodesign Summit had grown to over 250 registrants representing 19 countries and 33 states. Geodesign had, by that time, gone international, with the first Geodesign Summit in Europe attracting 250 people to The Netherlands and the first one in China gathering 500 people in Beijing, with an estimated 50,000 watching it online.

During those summits, keynote addresses by some of the leading thinkers of our time have helped shape geodesign as an intellectual inquiry. Bran Ferren of Applied Minds described geodesign as a form of storytelling that can reveal the global challenges faced and raise the level of public discussion about them. And Janine Benyus of Biomimicry 3.8 spoke about the relationship of BIS — biological information systems — to GIS, suggesting that we need to embed our understanding of how the natural world works into geodesign. The intellectual diversity apparent at the summits demonstrates both the breadth and depth of inquiry necessary for successful geodesign research and practice.

Who does geodesign?

While some geodesigners see it becoming its own field, separate from either design or geography, others see it differently, as a method, a process, and a set of tools rather than a separate field. Steinitz has argued that geodesign “cannot and should not become its own full-fledged profession … rather … all the relevant design professions and geographic science should adopt and adapt geodesign ideas and methods, and then collaborate as needed on the world’s most serious geodesign challenges” (Steinitz 2012, 11–12).

Geodesigners have come, initially, from the ranks of geographers and geographic information scientists on one hand and design fields like landscape architecture and urban or regional planning on the other. However, like GIS itself, geodesign has begun to pervade a wide number of fields as a growing number find themselves having to devise creative alternatives to the way they do things now.

At the 2014 Geodesign Summit, for example, attendees represented 16 different disciplines, including architecture, biology, design, ecology, forestry, finance, health care, insurance, landscape architecture, mining, marine science,
planning, parks and recreation, transportation, urban design, and water resources. Whatever their differences, these diverse fields – and many more – have begun to recognize the value of geodesign methods and tools.

How geodesign works

That value comes down to what geodesign has to offer that neither geography nor design can. Unlike much of geography, geodesign offers a powerful way of not just understanding a place, but also imagining something there that does not yet exist. And unlike design, geodesign provides a means not just to shape the spaces we use and inhabit, but also to assess the results of design decisions based on relevant information about a place.

Geodesign, in other words, combines the data-rich capabilities of GIS with the visualization capacity of drawing and modeling software. It also links the critical analysis of geographers with the creative imagination of designers. The intersection of these fields lets us reframe the questions we ask and the assumptions we make about a situation, knowing that, as Albert Einstein observed, we cannot solve our problems with the same thinking we used when we created them.

Geodesign also promises to alter the way in which both geographers and designers will work in the future. It provides a readily accessible and highly visual way of understanding complex information and relationships. This enables everyone from policymakers to community members to see the consequences of our actions and decisions, broadening the ability of the public to participate in planning and providing opportunities for all involved in or affected by a proposed change in the physical environment to test different options and to see where various scenarios might lead.

While both geographers and designers work at multiple scales, examining situations from various perspectives and moving back and forth from one frame to another, the two fields approach scales from different directions. In his book – *A Framework for Geodesign* – Steinitz sees geodesigners operating in the overlap between geography, which generally starts at the scale of the planet (1:1 000 000) and goes down to the scale of infrastructure systems (1:5000) and design, which typically starts at the scale of the person and the building (1:100) and moves up to the scale of the region (1:250 000). “Geodesign activities,” writes Steinitz, “range from a large development project … through to a … city … (or) regional watershed, in the 1:5000–1:250 000 scale range” (Steinitz 2012, 5).

A geodesign framework

The framework that Steinitz lays out in his book offers a rigorous way of approaching geodesign and builds on earlier work (1995). He rightly sees design as an iterative process and he envisions it involving at least three passes through a set of six questions and six models. The first pass entails scoping out the situation: describing it, understanding its operation, evaluating its performance, defining the extent of its possible change, considering the impact of that change, and deciding how to make the change. In many ways, that first pass parallels the stages of a typical GIS analysis.

The second pass, as Steinitz describes it, involves a methodological phase that geographers rarely get to and designers rarely discuss, involving who needs to be participate in decision making, what issues a design needs to address, what strategies a designer should use in making
change, what criteria should be used in evaluating alternative designs, what process a designer should follow, and what scales the design needs to encompass.

The third pass involves the design phase, in which the geodesigner devises alternative scenarios with those who must implement the design and, ideally, those most affected by it giving their feedback along the way. Here, too, the steps involve describing the design, determining its operation, assessing its functionality, deciding what needs to change, why, and how. In a typical design process, this third pass may occur several times and, as Steinitz observes, it may also require looping back to the first two phases, changing the scale, assumptions, or methods in order to arrive at an optimal solution.

How to geodesign

Steinitz identifies nine different ways of designing, which represents a substantial contribution to the design-thinking literature as well as that of geodesign. The first, anticipatory approach involves leaping to a final result and then retracing the route to arrive at that foregone conclusion. Too many designers take this approach, applying their signature style to a problem, often in inappropriate ways, and Steinitz’s analysis shows how this has merit only in situations where the designer has “a sufficient and adequate amount of experience” (Steinitz 2012, 56).

The second, participatory approach incorporates multiple design options from a number of participants, leading to a consensus design that includes elements of several proposals or an agreement about one proposal that best meets the needs of a situation. Here, Steinitz recognizes that this process, like all participatory processes, takes more time and demands a shared sense of what is appropriate to a place in order to recognize the most suitable end result.

The third, sequential way of designing entails a step-by-step process in which decisions get made in order, leading to an optimal end result. This requires what Steinitz calls an “abductive logic” (essentially a reasoned guess based on evidence) in order to have the clarity and confidence needed to ensure the correctness of each decision along the way, and it demands adequate criteria to ensure that it does not go off in the wrong direction.

The fourth, constraining approach works best when a situation lacks adequate criteria. The design effort involves seeking out constraints when too many options exist at every stage of the process, with designers problem-seeking in order to find the best solution in a given situation. This approach often leads to discoveries that no one fully understood at the beginning.

A combinatorial approach constitutes the fifth of Steinitz’s categories. When faced with uncertainty about the relative importance of various criteria or different options, designers will resort to making as many reasonable combinations as possible in order to test them according to their success in meeting the main objectives of a project. Digital tools have greatly enhanced this approach by increasing the pace and complexity of possible combinations and improving the efficiency with which a designer can assess their validity.

The sixth option, which Steinitz calls “rule-based,” entails the application of a set of agreed-upon rules to identify and assess alternatives along the way. In a process like this, many designers employ informal rules-of-thumb or heuristics, based on experience, although geodesign and other digital tools also allow the use of algorithmic rules.
A seventh, optimized approach works best, as Steinitz acknowledges, when everyone agrees on beforehand “the relative importance of each of the desired requirements and also its decision criteria” (Steinitz 2012, 58). This applies to situations in which one measure, like maximizing economic return or ecosystem health, dominates all other criteria.

The eighth, agent-based approach, which Steinitz sees as “necessarily computer-intensive,” applies “independent ... rule-based actions of independent agents” to each stage of the process, with the design adjusting as it proceeds (Steinitz 2012, 59). Geodesign lends itself to this method, given the ability of GIS to handle the complexity of interactions that exist within a particular geography.

A final mixed approach involves various combinations of any of the above strategies. While this may appear to lack rigor, it often characterizes complex design processes in which designers have to switch methods and alter approaches when one proves unworkable or inappropriate.

We tend to judge a design process, as we do the outcomes of a scientific experiment, according to its outcomes. Steinitz’s change models reveal the diverse ways in which we can get to those outcomes, with different approaches more or less relevant to different situations. In geodesign, the choice of the right method can matter as much as the arrival at the best solution.

The process of geodesign

Steinitz’s framework shows how design, rightly understood, involves a process as systematic and rigorous as the scientific method. Because most people go through their entire education without ever learning about the design process, the rigor that it requires may surprise those who think of it as something intuitive and subjective. Knowing when to use what method and how far or fast to use it may depend upon a geodesigner’s experience and the intuitive understanding that comes from that. But as Steinitz shows, the design process itself depends upon the ability of multiple people to engage in it, understand it, and communicate it, and that requires a discipline as demanding as that of science.

At the same time, Steinitz’s framework demonstrates to designers the methodological complexity that often goes undiscussed in the design community. Too many designers often end up using a favorite process or approach, whether or not it fits the requirements of a situation, because it has worked for them before or because they have become comfortable with it. Steinitz’s change models show that every situation requires an assessment of the best approach to the problem, with each method having its place.

Geodesign, in other words, reveals a lot about design to geographers as well as to designers themselves, while affecting the ways in which we think about both geography and design. It can help bring geography more into the center of public- and private-sector conversations by providing not just the critical analysis of a situation but also the alternative scenarios in response to it. Geodesign also enables designers to make more informed choices by giving them access to information about the impacts of their ideas at a rapid enough pace to affect what they explore and what options they ultimately propose.

The greatest benefits, though, accrue to those affected by change. Geodesign provides a method and set of tools that allow people to participate in and understand the consequences of decisions in ways that many do not have available to them now, empowering the public and potentially altering the nature of professional practice.
The practice of geodesign

Both geography and design have long remained within the purview of experts, professionals who, through their extensive training and sometimes-esoteric knowledge, have largely served the needs of those in power or those who have the means to pay for the services of these disciplines. Geodesign promises to disrupt the relationship between professionals and power.

The move of GIS online and to the cloud will, alone, give many more people access to geographical information once inaccessible to the general population. Geodesign promises to provide a means by which the public as well as those in power can assess the pros and cons of particular proposals and arrive at a consensus that involves those most affected by a decision as well as those with the most to gain from it. This democratization of geographical information and the design-review process will potentially mark a shift in the role of GIS and design professionals, moving from an expert model to more of a facilitation model of practice.

Geodesign also brings together the complementary skills of designers and geographers. Steinitz recognizes that most design professionals get “educated as generalists … know(ing) a little about a lot,” in contrast to most geographic scientists, whose “education produces specialists … (who) know a lot about a little” (Steinitz 2012, 5). That complementarity also applies to the way in which both geography and design see situations. Designers, as Steinitz puts it, tend to see “every design … (as) a unique experience for a unique place,” while geographic scientists tend to think that “geography is universal” or see “regional geographical and cultural differences” (Steinitz 2012, 6).

Those differences can make geodesign challenging if either geographers or designers see theirs as the privileged approach to problems. But “collaboration and cooperation among these groups,” writes Steinitz, “is not new or unknown, and there is a long history of successes” (Steinitz 2012, 5). The activity of geodesign, in other words, predates the idea of it.

Geodesign education

One indication of geodesign’s usefulness and potential lies in the number of education programs that have emerged. Since the first Geodesign Summit, geodesign education has remained a central topic, with faculty gathering to share information and ideas about coursework and curriculum. And over the course of a few years, several universities have started programs.

They range greatly in size and type. Just in the United States at the time of writing: the University of Southern California offers an undergraduate BS degree; Penn State a graduate certificate; Philadelphia University an MS degree; Northern Arizona University an undergraduate BS track; the University of Wisconsin, a multicampus post-baccalaureate program; the University of Arizona a graduate concentration; and the University of Georgia an undergraduate or graduate certificate in geodesign.

At the same time, a number of universities have faculty who offer geodesign courses. The rapid expansion of educational opportunities in geodesign shows how much this emerging area of inquiry resonates with faculty and students, capturing the growing interest among so many “millennials” (people born sometime between the early 1980s and the early 2000s) in fields that lead to jobs while also having the potential to do good and to change the world for the better.
Applications of geodesign

The Geodesign Summits have also demonstrated the variety of applications of geodesign. Captured in Shannon McElvaney’s (2012) book, Geodesign: Case Studies in Regional and Urban Planning, the examples of geodesign range from regional responses to climate change to urban designs for sustainable development to state-wide plans for growth to neighborhood-scale urban agriculture.

A similar diversity has occurred in the case studies and lightning talks at the various Geodesign Summits. Speakers have shown how geodesign has proven useful in improving the lives of people in informal settlements in Latin America and small-town Georgia, in preserving habitat corridors around Tucson, in assessing the impact of fracking in Pennsylvania, in engaging communities in development alternatives near the Twin Cities, in creating a new campus in Kuwait, in imagining a more sustainable future for Detroit, in promoting stewardship of the Great Lakes, in using chaos theory to reshape cities, and in inventing new uses for public land in San Francisco. Those examples convey not just the range of geodesign’s applications, but its usefulness in addressing some of our most pressing social, environmental, and economic challenges.

Limitations of geodesign

Geodesign has critics coming from both geography and design. Some geographers, for example, see geodesign as overly simplistic. After a recent talk about geodesign, one geographer in the audience reacted negatively to the focus of geodesign on problem solving, an idea common among the design fields. He argued that most human situations do not lend themselves to “a solution” and that most human activities and geographical ways of knowing involve dilemmas that resist resolution and that lead to unintended consequences if not rightly understood and if viewed too simplistically as a problem to solve. More broadly, geodesign is as open to critique as GIS, including practices such as public participation GIS, and therefore is subject to standard concerns of critical GIS and science and technology studies, such as questions pertaining to positionality and power, the politics of data, and so on (see the entry on Critical GIS).

At the same time, some designers see geodesign as overly constraining. At the same geodesign talk, a designer conveyed a similar skepticism about it, asking if basing every decision to data can inhibit new ideas and limit the ability of people to imagine something new and create something we have not seen before. She saw the potential of geodesign to reduce design to a set of computational tools or to drive decision making in particular directions because of the nature of GIS tools.

Such reactions represent invaluable feedback because they reveal the potential blind spots of those so close to the development of geodesign that they no long see its limitations. Those comments also reflect the inherent tensions that occur in an interdisciplinary activity like geodesign, which combines two established fields and yet does not fit comfortably in either.

Geodesign, though, also suggests a possible response to such challenges: the answer to this criticism depends upon the scale at which we look at the question. At the scale at which geographers typically work, issues rarely lend themselves to a solution, but at the scale at which landscape architects and urban designers work – the site, neighborhood, or regional scale – there do exist solutions that can help make a situation better.
The same applies to the criticism of the designer who saw geodesign as a limitation on a designer’s freedom. At the scale of a specific site, geodesign may represent too blunt a method or tool for decisions about where to place certain elements or what details to use in a particular location. But at the scale of the site and its relationship to larger flows, networks, or systems, geodesign enhances, rather than limits, a designer’s ability to make the right decision.

The future of geodesign

The future of geodesign looks promising, in part because we know that we can no longer afford or sustain many of the development patterns, resource allocations, and public-policy decisions of the past. Geodesign provides a means of allowing us to envision a better future for people and for places, based on what we know about their past and present. Geodesign not only helps us create our future but also helps us ensure that, as one of the most vulnerable species on the planet, we will have a future at all. Embedding in geodesign the ways in both human and natural ecosystems work could make it not just a powerful design tool but also a transformative one.

At the same time, geodesign can also empower us to change our behavior in ways that benefit us as well as our environments. After several centuries of disaggregating the world, chopping up its study into discrete disciplines and isolating us from the full consequences of our actions, geodesign offers us a way of re-aggregating things, revealing the interactions of complex flows and phenomena and helping us see the possible effects, in real time, of doing one thing or another. That suggests that geodesign, in the future, might include the creation of an easily downloadable application for mobile devices that every individual can use in daily decision making.

This, in turn, implies a role for geodesign as a new way of organizing our knowledge about the world. We damage both human and natural communities by acting as if they mirrored the disciplinary way in which we study them rather than as highly interrelated sets of ecosystems that evolved in each place. Geodesign offers a way of re-aligning how we know the world with how we know it actually exists by providing a tool that arranges knowledge according to the cultures and phenomena of a place rather than according to disciplinary categories. Geodesign will not just draw from database libraries; it may become the basis for how we organize the libraries of the future.

**SEE ALSO:** Multicriteria decision-making; Overlay, graphical; Participatory modeling; Public-participation GIS; Regional planning: the resilience of an imperative; Spatial decision-support system

**References**


Geodesy is the field of knowledge that studies the size and shape of the Earth and its gravitational field, and their changes over time. It covers the precise measurement of the Earth and its rotations, the abstraction of these measurements into models that can be operationalized for mapping, navigation and engineering, and gravimetry, or the measurement of the force and directional geometry of the local gravity vector. From the point of view of geography, geodesy is critical in providing accurate and updated Earth models known as datums from which positions can be found and heights measured – the foundation for cartography and geographic information systems. These Earth models are termed “ellipsoids” or “spheroids,” with their size and shape being determined by the lengths of their semiaxes, or the degree of flattening (Li and Götze 2001). Commonly used in contemporary mapping and positioning are the World Geodetic System of 1984 (WGS84) and the International Terrestrial Reference Frame (ITRF). In the United States, the major mapping agencies use the North American equivalent of WGS84 based on the Geodetic Reference System of 1980 (GRS80), termed the North American Datum of 1983 (NAD83), and its vertical equivalent for heights, the North American Vertical Datum of 1988 (NAVD88). In addition to ellipsoids, geodesy is also concerned with more precise Earth models called geoids. A geoid is the equipotential surface that would be created if Earth was completely covered with water. Differences in gravity direction locally are called deflections of the vertical. The deflections are caused by mountains, oceanic crust, and by differences in Earth’s shape. Geoid models are under continuous revision, and are sufficiently sensitive to reflect tectonic activity and changes in Earth’s rotation. A recent geoid is the GEOID12A, known as a hybrid geoid as its purpose it to tie together with a datum. Other geoids are related to precise gravimetry and to an effort to redefine the American Vertical Datum (GRAV-D) by 2022.

Earth size

The first recorded attempt to measure the size of the circumference of the Earth was by Eratosthenes of Cyrene (ca. 276 BC to ca. 195/194 BC), the Chief Librarian of Alexandria and “Father of Geography” (Rawlins 1985). Around the year 240 BC, he used the knowledge that at a well in Syene, Egypt (now Aswan), the sun was directly reflected from the water at the time of the June solstice. To Eratosthenes, this was evidence that Syene lay on the Tropic of Cancer. Using an obelisk as a gnomon at Alexandria, he measured the shadow length and gnomon height to compute the shadow angle as 1/50 of a circle (7.12′).
on the solstice. The distance from Alexandria to Syene was 5000 stades, confirmed by travel times for camel trains. This yielded 700 stades per degree, for an Earth circumference of 252,000 stades. While Pliny noted that the stade was 157.2 m, others give lengths up to 166.7 m. This translates to an Earth circumference of between 39,300 km and 41,675 km, or between 98.15 and 104.08% of the average of today’s estimate of the polar and equatorial circumferences.

Eratosthenes’s measurement was neglected in further writings, and Claudius Ptolemy changed the 252,000 stades in his *Almagest* to 180,000 stades in the later *Geography*. This led to Columbus’s much discussed underestimate of the distance for sailing west from Europe to reach India. The medieval Persian Abu Rayhan Biruni (973–1048) measured the Earth and distances on it using triangulation and calculated an Earth radius of 6339.6 km. However, it was not until the survey conducted in 1669–1670 by French astronomer Jean-Félix Picard (1620–1682) that a more accurate figure was computed. Picard measured a degree of latitude along the Paris Meridian using triangulation involving 13 triangles going from Paris to a clock tower near Amiens. His survey gave a result of 110.46 km for one degree of latitude, with a corresponding Earth radius of 6328.9 km.

During the French revolution, in 1791, the French Academy of Sciences selected a meridional definition of the meter because the force of gravity had been found to vary over the surface of the Earth, which changes the period of a pendulum of a standard length. The meter was originally to be defined as 1/10 000 000 of the distance between the North Pole and the equator, on the Paris meridian. The Academy commissioned an expedition led by Jean Baptiste Joseph Delambre and Pierre Méchain, lasting from 1792 to 1799, which measured the distance between a tower in Dunkirk and a castle in Barcelona to estimate the length of the meridional arc passing through Dunkirk, assumed equal to the length along the Paris meridian. Based on initial results from this expedition, France adopted the meter as the official unit of length in 1793. The first official meter rod, however, neglected the true shape of the Earth and so was short by one-fifth of a millimeter, making the actual Earth circumference on this meridian 40 007 863 m. Further accurate measures of the size of the Earth have necessitated having a model of Earth’s shape as either an ellipsoid or a geoid.

Earth shape

The shape of the Earth had long been assumed a sphere, based on evidence provided by Aristotle (384–322 BC). Other cultures assumed a flat disk and other forms, but the spherical shape seems to have dominated until an experiment conducted by Giovanni Domenico Cassini (1625–1712) while Director of the Paris observatory. Cassini noted that Isaac Newton’s theory of universal gravitation predicted that the Earth’s rotation would lead to its bulging out at the equator, flattening the poles. This implied that the equatorial radius would be larger than the polar, making the Earth an oblate spheroid. Isaac Newton’s theory of gravitation predicted the Earth to be an oblate spheroid with a flattening of 1 part in 230. A consequence would be that a degree of latitude measured near the pole along a meridian would be longer than the same degree at the equator. Cassini re-examined the measurements along the Paris meridian and concluded the opposite, that the length of a degree became longer as the meridian progressed south, implying an Earth model of a prolate ellipsoid. The ensuing intellectual debate over this issue became vigorous, and led to an
ambitious attempt to measure meridional arcs at high and low latitudes.

The dispute was settled by two measurements made by independent expeditions (Hoare 2004). Pierre Louis Maupertuis led an expedition to Finland in 1736–1737 to measure a degree of latitude by triangulation from a baseline on Lake Tornio, reporting back to the French Academy in 1737. According to Maupertuis, a degree at 66.23°N was 111,948.76 m, compared to Cassini's measurement at Paris of 111,212.22 m. Newton had been proven right, and the Earth was indeed shaped like an oblate spheroid, a special case ellipsoid. Maupertuis's result gives an equatorial radius of 6,397,300 m, and a polar radius of 6,363,806.28 m, for a first measure of the flattening of 1/191.

The second expedition (1735–1744) was sent to the equator in Peru, in what is now Ecuador. In May 1735, Louis Godin, Pierre Bouguer, and Charles Marie de la Condamine started near Quito. After 10 years, they had completed a meridional triangulation arc of 3°7′3″ giving a length of a degree as 110,624.82 m, again confirming the oblate ellipsoid model.

The lengths of arcs of both meridians and parallels became better known as triangulations expanded to cover the world, with major contributions in Africa (Clarke), India (Everest), Russia (Krassovsky), and North America (Hayford). Table 1 shows the steady improvement in the values for the standard measures. Starting in the 1920s there was an effort to standardize the ellipsoids globally, rather than have each nation use its own best ellipsoid. The Hayford ellipsoid, introduced in 1910, was renamed the International ellipsoid of 1924 after it had been adopted by the International Union of Geodesy and Geophysics in 1924. Another significant change was the creation of the first Earth-centered datum by the US Department of Defense as the World Geodetic System 1972 (WGS72). Over three years, data from Doppler satellites, surface gravity measurements, and astronomy worldwide were subjected to a large-scale least squares adjustment. In addition, a surface gravity field consisting of 410 10′ × 10′ equal area gravity anomalies were determined from ground measurements. The value of WGS72 ellipsoidal flattening of 1/298.26 is that still in use today.

A geodetic system is used in geodesy, navigation, cartography, and navigation systems to translate positions on a map to their actual positions on the ellipsoid or other Earth model or vice versa. A datum is a set of parameters and their values used to define a specific geodetic system. Among the most commonly used datums is WGS84, a standard for use in cartography, geodesy, and navigation and the preferred system for use with the Global Positioning System (GPS). It defines a standard coordinate system for the Earth, a standard reference ellipsoidal surface for height data, and a gravitational equipotential geoid that defines the nominal sea level. WGS84 was established in 1984 and last revised in 2004. The origin of WGS84 is located at the Earth's center of mass with a computed error of less than 20 mm. The WGS84 prime or zero meridian is the IERS Reference Meridian, which is 5.31 arc seconds or 102.5 m east of the Greenwich meridian at the latitude of London's Royal Observatory. The WGS84 datum surface is an oblate spheroid with an exact major semi-axis radius of 6,378,137 m at the equator and a flattening \( f = 1/298.257223563 \). The polar semi-minor axis is 6,356,752.3142 m. Until recently, the WGS84 Earth model used the EGM96 (Earth Gravitational Model 1996) geoid, revised in 2004, with deviations from the WGS84 reference ellipsoid ranging from about −105 m to about +85 m. The EGM96
Table 1  Historical measurements of the two axes of Earth’s ellipsoid.

<table>
<thead>
<tr>
<th>Reference ellipsoid name</th>
<th>Equatorial radius (m)</th>
<th>Polar radius (m)</th>
<th>Inverse flattening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maupertuis (1738)</td>
<td>6397 300.00</td>
<td>6363 806.28</td>
<td>191.00</td>
</tr>
<tr>
<td>Everest (1830)</td>
<td>6377 299.37</td>
<td>6356 098.36</td>
<td>300.80</td>
</tr>
<tr>
<td>Everest 1830 Modified (1967)</td>
<td>6377 304.06</td>
<td>6356 103.04</td>
<td>300.80</td>
</tr>
<tr>
<td>Everest 1830 (1967 Definition)</td>
<td>6377 298.56</td>
<td>6356 097.55</td>
<td>300.80</td>
</tr>
<tr>
<td>Airy (1830)</td>
<td>6377 563.40</td>
<td>6356 256.91</td>
<td>299.32</td>
</tr>
<tr>
<td>Bessel (1841)</td>
<td>6377 397.16</td>
<td>6356 078.96</td>
<td>299.15</td>
</tr>
<tr>
<td>Clarke (1866)</td>
<td>6378 206.40</td>
<td>6356 583.80</td>
<td>294.98</td>
</tr>
<tr>
<td>Clarke (1878)</td>
<td>6378 190.00</td>
<td>6356 456.00</td>
<td>293.47</td>
</tr>
<tr>
<td>Clarke (1880)</td>
<td>6378 249.15</td>
<td>6356 514.87</td>
<td>293.47</td>
</tr>
<tr>
<td>Helmert (1906)</td>
<td>6378 200.00</td>
<td>6356 818.17</td>
<td>298.30</td>
</tr>
<tr>
<td>Hayford (1910)</td>
<td>6378 388.00</td>
<td>6356 911.95</td>
<td>297.00</td>
</tr>
<tr>
<td>International (1924)</td>
<td>6378 388.00</td>
<td>6356 911.95</td>
<td>297.00</td>
</tr>
<tr>
<td>NAD 27 (1927)</td>
<td>6378 206.40</td>
<td>6356 583.80</td>
<td>294.98</td>
</tr>
<tr>
<td>Krassovsky (1940)</td>
<td>6378 245.00</td>
<td>6356 863.02</td>
<td>298.30</td>
</tr>
<tr>
<td>WGS66 (1966)</td>
<td>6378 145.00</td>
<td>6356 759.77</td>
<td>298.25</td>
</tr>
<tr>
<td>Australian National (1966)</td>
<td>6378 160.00</td>
<td>6356 774.72</td>
<td>298.25</td>
</tr>
<tr>
<td>New International (1967)</td>
<td>6378 157.50</td>
<td>6356 772.20</td>
<td>298.25</td>
</tr>
<tr>
<td>GRS-67 (1967)</td>
<td>6378 160.00</td>
<td>6356 774.52</td>
<td>298.25</td>
</tr>
<tr>
<td>South American (1969)</td>
<td>6378 160.00</td>
<td>6356 774.72</td>
<td>298.25</td>
</tr>
<tr>
<td>WGS-72 (1972)</td>
<td>6378 135.00</td>
<td>6356 750.52</td>
<td>298.26</td>
</tr>
<tr>
<td>GRS-80 (1979)</td>
<td>6378 137.00</td>
<td>6356 752.31</td>
<td>298.26</td>
</tr>
<tr>
<td>NAD 83</td>
<td>6378 137.00</td>
<td>6356 752.30</td>
<td>298.26</td>
</tr>
<tr>
<td>WGS-84 (1984)</td>
<td>6378 137.00</td>
<td>6356 752.31</td>
<td>298.26</td>
</tr>
<tr>
<td>IERS (1989)</td>
<td>6378 136.00</td>
<td>6356 751.30</td>
<td>298.26</td>
</tr>
<tr>
<td>IERS (2003)</td>
<td>6378 136.60</td>
<td>6356 751.90</td>
<td>298.26</td>
</tr>
</tbody>
</table>

The International Earth Rotation and Reference Service (IERS) was established in Paris in 1987 by the International Astronomical Union and the International Union of Geodesy and Geophysics. The primary objectives of the IERS are to “serve the astronomical, geodetic and geophysical communities by providing data and standards related to Earth rotation and reference frames.” Most important, the continued integration of geodetic measurements worldwide has led to complex models that link the geoid to ellipsoids and to nationally supported datums such as NAD83. This is especially important because today’s geodetic models
and geoids are time-sensitive. Due to changes in the location of the Earth's rotational axis (Chandler wobbles), and in the distribution of mass on Earth due to tectonics, land subsidence, constructions, dams, water flow, ice masses, and so on, there is a need to continuously update both geoid and ellipsoidal models. IERS has created the International Terrestrial Reference Frame (ITRF) for this purpose. It distinguished between the fluid models of the geoid and reference systems that are Earth-centered, Earth-fixed (ECEF). These are systems that tie to a particular ellipsoid and datum, which is why the date is included in the datum name. ECEF positions, such as GPS data, must record the GPS epoch in which coordinates were recorded, since over time the positions relative to fixed systems will change.

**Earth gravity**

Earth gravity can be defined as the force exerted on a mass due to the combination of both the gravitational attraction of the planet on the mass in accordance with Newton's universal law of gravitation, plus the rotational force of the planet, which has both a centrifugal acceleration and a force due to the equatorial bulge. The force of gravity is weakest at the equator due to the centrifugal force and increases with latitude toward the poles because the poles are closer to the center of mass. A standard acceleration due to gravity of 9.80665 Newtons (m/s²) was adopted by the International Committee on Weights and Measures in 1901 at 45° latitude. Nevertheless, at any specific position on Earth, this value is not uniform but follows a complex spatial pattern that reflects tectonics and material density as well as Earth’s shape and topography. The net force exerted on an object due to the Earth is termed the “effective gravity” or “apparent gravity.”

During the 1735–1744 French Academy of Sciences expedition to Peru, Bouguer noticed that the down vector, or astronomical vertical, was pulled in the direction of large mountain ranges, due to their higher gravitational attraction. This was demonstrated by Nevil Maskelyne and mathematician Charles Hutton in a classic experiment in Scotland in 1774 on a symmetrical mountain called Schiehallion. The experiment measured the deflection of a pendulum by the mass of the mountain, giving estimates of the mean density of Earth, its mass, and a value for Newton’s gravitational constant. The study of the “undulation of the geoid” and the “deflection of the vertical” became the focus of geodesy during the nineteenth and twentieth centuries. Early geodesists discovered and explained some of the differences as the result of isostatic uplift following the last ice ages.

From the 1950s onward, it became possible to measure the acceleration force due to gravity directly using a gravimeter, essentially an accelerometer that measures forces on a spring. First developed in the 1950s–1960s as a navigation aid for nuclear submarines, by the 1980s, transportable gravimeters were available for use on ships, in the air, and on satellite gravity surveys. These devices were accurate to units of gals (cm/s²), and even milligals, the more commonly used (but non-SI standard) units of acceleration. Gravimeters are of two types. Absolute gravimeters measure local gravity in absolute units, while relative gravimeters compare the value of gravity at a point to another known point.

Apparent gravity on the Earth’s surface varies by around 0.7%, from 976.39 gals in Peru to 983.37 gals in the Arctic Ocean. High-resolution airborne and satellite gravimetry show that the
The gravitation field is highly complex (Figure 1). A highly successful spaceborne gravimetry satellite, the Gravity Recovery And Climate Experiment (GRACE), a joint mission of NASA (National Aeronautics and Space Administration) and the German Aerospace Center, has been making high-resolution measurements of Earth’s gravity field since March 2002 (Tapley et al. 2005). GRACE has conducted worldwide comprehensive mapping at high spatial resolution (Figure 2).

### Geodetic mapping

As the various national mapping programs were launched during the nineteenth and twentieth
centuries, each country chose fixed points at which to conduct triangulations. These triangles were further subdivided into two more depths in a hierarchy of triangles and the main networks of triangles were extended in a mesh pattern (Figure 3). The networks included historical triangulations, such as the United States’ Mason–Dixon line, and the great arcs of the Survey of India. In most cases, the network of triangulation triangles had to be tied to the ground at a single reference point. In the United States, a position at Meades Ranch in Kansas (Figure 4) at 39°13′26.71218″N 98°32′31.74604″W (NAD83) was chosen as the single point for reference for the North American Datum of 1927 (NAD27). Most of the first generation of US topographic mapping used this datum, both horizontally and vertically. The placement of ground markers or benchmarks (Figure 5) by the United States Geological Survey (USGS) and the National Geodetic Survey extended this control system nationwide. Benchmarks and temporary geodetic tie points are shown on US and other topographic maps at detailed scale, and their reference information is usually available through public websites.

Vertical versus horizontal control

The adoption of a particular datum involves assuming the lengths of the ellipsoid semiaxes, but it also assumes an orientation or “fit” of the ECEF coordinates. Once established, this surface will be the model onto which the positions of Earth’s features are projected. This involves elevations both below and above the datum, with mountains above and oceans below. This
mapping means that different datums place features in different locations and at different heights. For this reason it is essential that maps and geographic information system (GIS) data layers carry metadata about which datum they reflect. This becomes problematic when data from different datums need to be converted to a common reference frame for overlay. In general, it is best to convert from a particular coordinate system back to geographic coordinates (latitude and longitude) and then to convert the datum. While this impacts assembling maps across tiling systems (such as the division of topographic maps by equal ranges of latitude and longitude) and for overlay, it most noticeably means changes in elevations. Map spot heights, contours, and elevations within digital elevation and surface models may change significantly.

Early vertical heights were mapped using mean sea level, and the heights were carried across land masses by using horizontal level surveying along transects. This method obviously continuously adjusts to changes in the down vector, but was inaccurate because sea level was arbitrarily set and errors were multiplicative along the transect. Similarly, positions surveyed by astronomical observation and chronometer, then corrected to sea level by altimeters using air pressure, were also inaccurate. It has only been since the advent of Global Navigation Satellite Systems, including the US Department of Defense Global
Positioning System (GPS), that the height above a datum at a point has been accurately measurable. As more global navigation satellite systems (GNSS) come into service, including the Russian GLONASS and European Galileo systems, multisystem GNSS receivers can measure latitude, longitude, and height with precisions in the millimeters. The use of the Continuously Operating Reference Station network (CORS) administered by the National Geodetic Survey, and the ability of advanced GNSS users to contribute public data to geodetic models via the Online Positioning User Service (OPUS), means that maps and GIS data can be corrected to any datum and their accuracy greatly increased. Collectively, these systems form the National Spatial Reference System (NSRS) for the United States, and align with the International Terrestrial Reference Frame globally (Altamimi et al. 2001).

Figure 4 Triangulation station at Meade's Ranch. Control point for the North American Datum 1927 (in 1940). https://en.wikipedia.org/wiki/Meades_Ranch_Triangulation_Station#/media/File:Meade_Ranch_triangulation_station_ca._1940.jpg. Reproduced from NOAA.
Figure 5 A typical US Coast and Geodetic Survey benchmark. https://en.wikipedia.org/wiki/Benchmarking_(geolocating)#/media/File:USCGSE134.jpg. Reproduced from NOAA.

SEE ALSO: Cartography: history; Discrete global grid systems; Geocoding; Geographic information system; Geolocation services; Global navigation satellite systems; Map projections and coordinate systems; Metadata; Routing and navigation; World cities

References


Hoare, M.R. 2004. The Quest for the True Figure of the Earth: Ideas and Expeditions in Four Centuries of Geodesy. Burlington, VT: Ashgate.


Further reading


There has been a recent and rapid growth of interest within the scientific and policy community in exploring a range of techniques, collectively termed “geoengineering” (or alternatively “climate engineering”), for deliberatively intervening in the climate to counteract global warming (see, for example, Royal Society 2009). Within the space of a few years, and with the endorsement of learned societies and governance institutions, geoengineering has become a topic of mainstream respectable scientific and policy debate.

One class of geoengineering techniques, termed solar radiation management (SRM), is intended to counteract the warming effects of climate change by reflecting some of the inbound sunlight back into space and hence reducing global average temperature. Methods include increasing the reflectivity of the Earth by brightening human settlements or deserts to make them more reflective, enhancing marine cloud reflectivity through for example the spraying of salt-rich sea water into the sky, injecting aerosols such as sulfates into the stratosphere to mimic the dimming effects of large volcanoes, or placing space mirrors or sunshades in space to reduce the amount of solar radiation entering the atmosphere. These techniques, particularly stratospheric aerosols, have received particular attention because of the perception that they could be fast acting and relatively cheap, but have also prompted concerns about unintended and potentially harmful side effects. The other main class of techniques comprises carbon dioxide removal (CDR) methods, which seek to slow down or halt anthropogenic climate change by removing and sequestering greenhouse gases from the atmosphere. These include land management practices to enhance carbon sinks such as through large-scale afforestation and peatland restoration, the use of biomass for carbon sequestration such as biochar and dark earths, the direct air capture of ambient carbon dioxide through innovations such as artificial trees, and the increase in the oceanic uptake of carbon dioxide including through iron fertilization to stimulate phytoplankton blooms. These techniques are for the most part seen as safer, but slower and more expensive (Figure 1).

Although geoengineering is largely at a pre-research and development stage, it has received particular attention in recent years in the science policy community. The scientific debate was granted legitimacy following an essay in 2006 by Nobel Prize-winning atmospheric chemist Paul Crutzen, in which he discussed the disconnect between carbon dioxide emission trajectories (which continue to rise despite mitigation efforts) and stabilization requirements (which according to most official estimates require 60–80% emission reductions), leading to fears that rising greenhouse gas concentrations
may push the global climate into a new state, with potentially catastrophic consequences. This provides one explanation for the rise of geoengineering as an emergent science policy discourse: concerns that current mitigation policies may not produce the necessary reductions in emissions that are necessary to avoid dangerous climate change (the “climate emergency” argument). In the context of the lack of progress in meeting Kyoto targets, the lack of strong international agreement at United Nations conferences at Copenhagen and Durban, the rapid and ongoing growth of emerging economies such as China and India, and more general questioning on the United Nations Framework Convention on Climate Change (UNFCC)/Kyoto model as a plausible mechanism to deliver change, geoengineering is presented as at least a prudent option to explore, as a means of avoiding “potentially drastic climate heating” (Crutzen 2006, 216).

The second reason is more pragmatic: proponents of geoengineering argue that solar radiation management techniques (stratospheric aerosol injection techniques in particular) could reduce global temperatures not only relatively quickly, perhaps within a few months of deployment, but also relatively cheaply, relative to the cost of implementing greenhouse gas emissions reductions. Geoengineering is thus cast by advocates, among them fellow eminent scientists David Keith (Harvard) and Ken Caldeira (Stanford), as a potential and cost-effective means of buying time to curb emissions and promote more sustainable practices (the “buying time” argument); or, more evocatively, in the words of Keith, as perhaps “the only human response that can fend off [the] rapid and high-consequence impacts [of climate change].” For both of these reasons, geoengineering is fast emerging as a third policy route for responding to climate change, alongside mitigation and adaptation, and as a necessary option that needs to be researched and developed in case it is ever to be required. As Paul Nurse, President of the Royal Society, stated in a letter to the UK Guardian newspaper in September 2011: “A time may come when mankind will need to consider geoengineering the climate to counteract climatic effects of greenhouse gases. If that time comes, we need to have a good understanding of whether such efforts will work and, just as importantly, whether they will have any negative side effects.”

The policy debate on geoengineering governance and regulation remains in its infancy. There currently exist no international treaties that cover all geoengineering techniques, although it is widely assumed that most techniques could be covered by an extension of existing treaties. Nevertheless, there have been some early statements on geoengineering by policy bodies. In October 2010, for example, the United Nations Convention on Biodiversity declared that there should be no field tests of geoengineering projects that might affect biodiversity; in October 2011 the European Parliament expressed its opposition to proposals for large-scale geoengineering; while in October 2013 the Intergovernmental Panel on Climate Change (IPCC) highlighted the risks of a scenario of SRM termination (following a period of deployment), which could lead to abrupt and rapid increases in global surface temperatures, with significant ramifications. Such statements are not legally binding, but they nevertheless reflect political unease with the prospect of geoengineering taking place without adequate regulatory arrangements. In response, a number of small-scale initiatives have occurred in recent years, aimed at articulating the goals and possible form of geoengineering research governance. These include: the establishment of the “Oxford Principles” for the
GEOENGINEERING/CLIMATE INTERVENTION

responsible conduct of geoengineering research; the development of a framework for responsible innovation aimed at guiding assessment on whether elements of a potentially controversial SRM project should be permitted to proceed (the UK Stratospheric Particle Injection for Climate Engineering – SPICE project); and the Solar Radiation Management Governance Initiative (SRMGI) set up to examine in depth the governance issues raised by research into solar radiation management methods.

There is still considerable diversity of opinion about exactly what form geoengineering governance should take. However, there seems to be an emerging consensus that it should involve a combination of soft law and hard law, be guided by principles such as “the public interest” and transparency, and involve “upstream” engagement with wider stakeholders and the public. It is also argued that governance during the research stage might be relatively “soft” to permit or even encourage “safe,” laboratory or small-scale research (with proposed governance mechanisms ranging through laissez-faire permissiveness, self-regulation, independent national policies, to an informal consortium of countries); however, most argue that governance would have to become “harder” before any large-scale field research or deployment, probably through a multilateral, international body such as the United Nations.

Pathways for critical geographical inquiry

Despite the fact that geoengineering remains at an early stage of development, there remains considerable scope for critical and engaged geographical inquiry. In the remainder of this entry we suggest the forms that this may take. First, there is the argument from nature – that geoengineering is imagining its interaction with the natural world in novel and distinctive ways. Unlike many technoscientific issues, the distinctiveness of geoengineering does not lie in the use of novel technologies with new properties: the actual interventions themselves typically involve mundane technologies such as mirrors, iron dust, sulfate particles, or crumbled rock, albeit deployed at a very large scale. Its novelty rather lies in the intention to use these technologies in a project of bringing planetary systems under human control. In philosophical terms, this encompasses the existential argument that the deployment of geoengineering would represent a hubristic extension of human powers, in that it would establish a radically new relationship between society and nature through the “making” of new climates. This constitutes arguably a distinctly new form of geopolitical formation in which humans are self-consciously considered as geologic agents within the Anthropocene – the characterization of our current geological epoch dominated by the increasing power and effect of humanity on the Earth’s systems. Cultural geographers can help illuminate the ways in which geoengineering rhetoric deploys a highly distanced global perspective and how this can be contrasted to more local and phenomenological approaches that can be used to explore the lifeworld of a geoengineered future.

Second, there is the argument from risk governance – that geoengineering may produce unforeseen and unintended consequences, with serious implications for its governability. The prominence of geoengineering in science policy debates is based largely on the idea that it could be “cheap and effective” compared with other carbon mitigation strategies. Such assessments, however, make the crucial assumption that the impacts of geoengineering interventions on the climate will be in line with those predicted by climate models. However, this remains far from
certain. Indeed, geoengineering has a distinctive relationship with uncertainty that problematizes such assumptions. With most previous technologies that have been subject to regional and global governance and control (e.g., persistent organic pollutants and chlorofluorocarbons) it is the side effects, often after decades of use, that have been of concern. Such side effects have typically been hard to predict or attribute, but research has increased certainty to a sufficient degree as a basis for global action, often underpinned by a degree of precaution, making possible both hard and soft governance arrangements. But with geoengineering technologies the task of governance is significantly different: it is the intended effects (for example, a reduction in global temperature) that are global; that may only become apparent over long timescales; and that are probabilistic and highly mediated, since they involve affecting technological and statistical constructs such as “global average temperature” through intervening into an Earth system which is highly chaotic and always in formation. Any deployment of geoengineering techniques is thus likely to have the unfinished character of research; an important role for geographical inquiry will be to problematize the boundaries and distinctions between research and deployment.

Third, there is the argument from politics – about the likely impact of geoengineering on political systems and on our collective ability to manage such interventions. Part of this concerns the “moral hazard” dilemma: the idea that geoengineering could make emission reductions politically harder once there exists the theoretical possibility for a technological alternative solution. There remains a need for social research on geoengineering’s impact on environmental attitudes and behavioral change, both with policymakers and with the wider public. Another element concerns the uncertainties and difficulties of causal attribution mentioned above, which may lead to increased conflict and strain on international relations and institutions. But another element of this argument concerns geoengineering’s relationship with intent. Questions of intent have been central to the political processes shaping the constitution of geoengineering as a technology. The formal definition of geoengineering is tied to the goal of offsetting anthropogenic climate change. Whether or not a specific action such as releasing particles into the upper atmosphere counts as SRM geoengineering deployment, or as research, or even as mere pollution cannot be determined by a mere technical procedure, but only by reference to intent. This implies that meaningful geoengineering governance would logically require the scrutiny and regulation of the intentions, whether explicit or implicit, of a huge range of research and deployment activities. Furthermore, the intent in SRM geoengineering is likely to be unstable and open to plural interpretations. The framing of SRM as a means of counteracting anthropogenic climate change is likely to be joined or displaced by alternative frames – for example, as a means to achieve humanitarian, environmental, nationalistic, military, or commercial goals. Understanding and mapping the interpretive flexibility of intent in SRM research, its intersections with democratic governance, and its potential for generating new kinds of conflict and controversy is a rich arena for future geographical inquiry.

Fourth, there is the argument from political economy – that geoengineering is likely to be conditioned by economic relations that are in tension with the imperative for democratic control. Such issues can be expected to become significant shaping factors in geoengineering research once deployment becomes more than a theoretical possibility; yet they are currently receiving little attention. The science and politics
of geoengineering have so far developed through a particular and restricted assemblage of actors and ideas, with the result that policy and regulatory treatments of geoengineering have largely failed to acknowledge questions of political economy. There have been accusations that advocacy for geoengineering research is part of a project that aims to protect established political and economic interests by creating a rhetorical defense against more aggressive carbon reduction measures. Such arguments are perhaps overly simplistic, but it is nevertheless true that geoengineering can readily be co-opted by vested interests. Substantial economic opportunities are likely to be created by any plan to deploy geoengineering, including: the patenting of geoengineering techniques; the design of particles for release into the stratosphere; the design of delivery systems; the sourcing and transport of raw materials; the design and implementation of monitoring systems; and the establishment and running of financial schemes of funding and possible compensation. In addition, there are likely to be profound issues of justice: that a geoengineered world may simply perpetuate North–South inequalities, and intensify rather than reverse the “colonization of the sky” represented by historical anthropogenic greenhouse gases from the Global North. Exploring the possible differential spatial effects of different geoengineering interventions and the geopolitics of producing artificial climates, critical geographical scholars can draw on world systems theory and postcolonial and subaltern studies, situating geoengineering in the long view of North–South relations.

Fifth, there is the argument from public engagement – the need for robust research to clarify public attitudes to geoengineering and the conditions, if any, under which geoengineering research and deployment might secure public acceptability. Current research has to date been dominated by public opinion surveys, although there also exist a number of attempts aimed at generating more qualitative public dialogue. Methodologically, there exist a number of challenges for the delivery of public engagement research. A key consideration concerns how to design the dialogue exercise in a way that permits participants to explore ways of thinking about the debate that do not simply reproduce dominant framings. It is important too that discussions are not framed by experts, that moderators seek to ensure a diversity of voice independent of background or experience, and that techniques are devised that support participants in the process of imagining the kinds of world that geoengineering might bring into being. Geographical inquiry, especially at its interface with science and technology studies, can provide examples of good practice and guidance in using deliberative methodology to research the unfamiliar, complex, and ethically challenging issues arising out of scientific and technological innovation.

More generally, it is important that critical geographical approaches – and the interpretative social sciences more generally – move away from the logic of “subordination,” in which they are allocated the task of filling in the gaps within a dominant frame provided by the physical and natural sciences, towards a more reflexive imaginary for geoengineering, that promotes and benefits from the various ways of seeing that different disciplines offer (Barry, Born, and Weszkalnys 2008).

SEE ALSO: Climate policy; Construction of nature; Corporations and the Anthropocene; Democracy; Environmental futures; Environmental governance; Environmental science and society
References


Further reading


Geographic data mining

Chao Zhang
Jiawei Han
University of Illinois at Urbana-Champaign, USA

Geographic data mining (also known as spatial data mining) applies computational methods to massive geospatial datasets to discover novel, interesting, and useful patterns and knowledge. Typical geographic data mining and knowledge discovery tasks include spatial association and autocorrelation analysis, spatial segmentation and clustering, spatial classification and trend analysis, and spatial outlier and anomaly analysis.

Historical background

As early as in the 1960s, the emergence of computer technology nurtured the management and analysis of digitized geographic data. In the following decades, geographic information systems (GIS) and spatial database management systems (SDBMS) grew rapidly, with numerous techniques designed to provide effective management as well as basic analysis of geographic data. For example, to enable fast browsing and querying of large-scale geographic data, various spatial indexing methods have been proposed to organize geographic data into efficient data structures (Rigaux, Scholl, and Voisard 2001).

Geospatial data mining began several decades ago as a confluence of geostatistical data analysis, data mining and machine learning research (Miller and Han 2009; Longley et al. 2010). While sharing a similar spirit of manipulating geographic data, geographic data mining is quite different from SDBMS techniques. The main focus of SDBMS is on organizing, browsing, and querying geographic data, whereas geographic data mining aims at designing effective, efficient, and scalable methods to extract useful patterns and knowledge from georeferenced data. Unfortunately, classic data mining methods developed for discrete data (e.g., transactional data) cannot be directly applied to discover knowledge in geographic data, because geographic data objects usually fall in a continuous space. Therefore, various algorithms tailored for geographic data have been proposed for different mining tasks.

Interest in geographic data mining has grown further in recent years, due to the unprecedented availability of diverse and large-scale geographic data. In addition to conventional geographic data types, such as map and tracking data, new types of georeferenced data are now routinely collected. For example, when a mobile phone user makes a phone call, his or her location can be inferred based on the signals received from multiple cell towers. The location information is typically also attached with abundant metadata (e.g., text message, call duration), which can be fused with geographic data for knowledge discovery. As another example, with the enormous growth of online social networking and social media sites, georeferenced data are being generated rapidly by the users of these services. A user can either specify a GPS location or mention the name of a geographic location in his or her social media post (e.g., tweet, check-in, photo, video). Such georeferenced social media data enable the discovery of interesting georelated knowledge.
Foundations

Geographic data mining tasks can vary from one application to another. Generally, major geographic mining tasks include the following: spatial association and co-location mining, spatial clustering, spatial classification and prediction, spatial outlier detection, mining spatiotemporal data, and mining text-rich spatial data.

Spatial association and co-location pattern mining

Spatial association and co-location patterns involve sets of spatial objects (or types of objects) that are often spatially located close together. For example, cities are often located close to rivers, lakes, and oceans. Finding co-location patterns may help discover associations among spatial objects and facilitate spatial reasoning and prediction. Since there are potentially an exponential number of possible combinations of spatially co-located sets of objects, efficient algorithms have been developed to reduce computational complexity and find meaningful patterns. A useful method is to find co-location patterns in a process of successive refinement. For example, patterns that are geographically close can first be found before searching for more refined spatial relationships (such as touch, intersect, overlay), or first find semantically meaningful high-level patterns before examining more refined spatial patterns. This approach can reduce the number of candidate sets to be examined and can help to find patterns at a refined scale (Koperski and Han 1995; Zhang et al. 2014).

Spatial clustering

Spatial clustering, also called spatial segmentation, is the process of partitioning spatial objects into groups such that the objects within the same group are similar or spatially close, whereas those in different groups are further apart. Given a set of spatial objects, several kinds of clustering algorithms have been developed (Han, Kamber, and Pei 2011).

Partitioning-based methods: given a database of \( n \) spatial objects, partitioning-based methods divide the \( n \) points into \( k \) groups. The well-known partitioning methods include the \( k \)-means clustering algorithm, which iteratively assigns each object to its nearest group center and then recomputes the center of each group until the assignment becomes stable. Another influential algorithm, the expectation-maximization (EM) algorithm, adopts a soft partitioning approach, assigning objects to each group in a probabilistic way and iteratively refining assignments until stability is achieved. It has also been popularly used in spatial clustering.

Hierarchical methods: hierarchical clustering methods create a hierarchical structure for the given set of objects. The hierarchy can be constructed in either an agglomerative (bottom-up) or divisive (top-down) manner. An interesting hierarchical algorithm, called BIRCH (Zhang, Ramakrishnan, and Livny 1996), first clusters spatially close objects into microclusters at a low level and then uses various kinds of effective clustering methods over those obtained microclusters to form flexible and interesting structures.

Density-based methods: partitioning-based methods can only find sphere-shaped clusters. Density-based clustering methods are able to find arbitrarily shaped clusters. One representative density-based clustering method, called DBSCAN (Ester et al. 1996), clusters objects such that a set of spatial objects are defined as core objects if a sufficient number of spatial objects are located within a specified radius, or if they are spatially connected to those core objects. Thus a cluster can be grown into arbitrary shapes by incorporating new objects.
that are density-reachable from other objects in the cluster. Other density-based clustering algorithms are based on the notion of probabilistic data distributions.

Many other clustering methods have been developed in recent data mining research, such as grid-based clustering, constraint-based clustering, stream clustering, subspace clustering, and high-dimensional clustering. In many geographic applications, clustering should not only consider spatial proximity but also proximity on time, text, network, or semantics associated with various kinds of attributes, as well as application constraints and requirements. Further development of advanced clustering methods will benefit geospatial clustering and its broad application.

Spatial classification and prediction

Spatial classification is the process of identifying the category (or class) to which a given spatial object belongs. Compared with clustering, classification relies on training data that contain the geographic instances whose categories (or class labels) are already known. Spatial classification is also called spatial prediction if the value to be predicted is a class label (i.e., category). On the other hand, when it refers to predicting numerical values with given spatial objects, spatial prediction is also called spatial regression or spatial trend analysis if the value to be predicted changes with time.

For spatial classification, each spatial object is described by a set of features, which can contain both spatial and nonspatial attributes of the object. Typical classifiers (e.g., decision trees, Bayesian classifiers, artificial neural networks, and support vector machines) can be used to build classification models based on training data, if the spatial correlations among the objects are not essential to the determination of classification labels. Then, given a new object, the trained model can be used to predict its category according to its features. In addition, methods have also been developed to consider not only the features of target objects but also the features of the nearby neighbors and their relationships to such objects. When spatial correlation is essential in determining classification labels, feature selection should be conducted to select critical features for proper classification.

In addition to spatial classification, it is sometimes useful to predict the location of an object at a specific time. For this location prediction task, the dependencies between the geographic objects need to be modeled along with the features of individual locations. Classic regression models can be applied to fit the time-stamped location sequence for the next location prediction task. Other statistical models, such as the hidden Markov model (HMM) and conditional random field (CRF) for modeling sequence data, can also be applied to location prediction problems.

Outlier detection

An outlier is a data object whose observed value deviates significantly from the values of other data objects. The following types of outliers can be defined over a set of spatial points: (i) a global outlier is an object that deviates significantly from the other objects in the entire dataset; (ii) a contextual outlier is an object that deviates significantly with respect to a specific context of the object; and (iii) a collective outlier is a set of objects that collectively lead to significant deviation to other objects.

Methods for outlier detection can be categorized into: supervised, semi-supervised, and unsupervised. For supervised outlier detection, part of the spatial dataset is examined and labeled by domain experts. Such labeled data are used as
training data to construct a classifier. Based on the trained classification model, outlier detection can be considered as a classification problem, and any new spatial points can be classified as “normal” or “outlier.” Unsupervised outlier detection methods assume the normal data points are somewhat “clustered.” Based on this assumption, this class of techniques first extracts clusters from the given spatial points, and then detects those points that do not belong to any cluster as outliers. Semi-supervised outlier detection is used when there are only a limited number of labeled spatial points, while most of the points are unlabeled. The labeled data can be leveraged to improve outlier detection performance. For example, given some labeled normal objects, nearby unlabeled data are first found and also considered normal. Semi-supervised methods are often used together with unsupervised methods to deal with situations where the number of labeled spatial points is small.

Key applications

Owing to the popularity of various location-based services, georeferenced data are becoming ubiquitous and geographic data mining is becoming increasingly useful in many real-life applications. For example, public transportation vehicles (e.g., taxis, buses) in a city are mostly equipped with GPS modules, enabling them to constantly report their locations to a data center for tracking and scheduling purposes. With the vehicle locations as input, geographic data mining methods can extract hot spots for different time periods in a city. In urban planning, such hot spots are important for improving road network design and transportation scheduling. As another example, in online social networks (e.g., Facebook, Twitter, Flickr), millions of geotagged texts and photos are being generated every minute. The massive amount of georeferenced social content opens doors to revealing regional topics, attractions, and sentiments. Such knowledge can greatly benefit tasks like targeted advertising, trip advising, and localized content delivery.

Future directions

In the past few decades, significant progress has been made in geographic data mining and numerous commercial systems have been developed for different tasks. However, three newly emerging areas have introduced new opportunities as well as challenges to this field, making geographic data mining an increasingly active research field.

Pattern discovery from heterogeneous data: nowadays, geographic data can be collected from multiple sources, including map data, sensor data, mobile data, and social media content. As a result, the geographic data are often associated with other types of data, such as phone-call data, text, photo, and video. Such georeferenced, heterogeneous data provide great opportunities for extracting semantically meaningful and important geographic patterns. For example, research efforts have been paid to mining useful knowledge from geotagged social media content and semantic trajectory data. How to deal with data heterogeneity, and thus discover useful geographic patterns, remains an important research direction.

Interactive and exploratory geographic data mining: in many geographic data mining applications, the users’ mining intents may not be known beforehand. It is important for a geographic mining system to support interactive data mining, such that end users can understand the output patterns and adjust their mining intents accordingly. Data visualization techniques have been employed to provide an intuitive demonstration of mining
results. One critical challenge is efficiency and scalability, as exploratory analysis needs real-time analysis, despite the sheer size of georeferenced data. It is critical to design efficient and scalable algorithms so that effective mining can be conducted within limited time. Many geographic data mining applications may face the dilemma between choosing a computation-intensive model and a less effective, but fast, approach. The design of efficient, effective, and scalable geographic data mining algorithms remains a critical research direction.

**Mining spatiotemporal and moving object data in cyber-physical systems:** with rapid advances in computer and sensor technologies, it is expected that many interconnected objects in the future will produce records that are geocoded and changing dynamically. It is critical to develop sophisticated geospatial data mining technology that takes time, location, text (or other high-dimensional semantic features), and sensor-network features into consideration. Spatiotemporal data mining, moving-object or trajectory data mining, geotext analysis, and physical, information, and social network analysis should be integrated to handle emerging data-mining challenges in cyber-physical systems.

**SEE ALSO:** Big data; Decision analysis; Geographic information science; Geographic information system; Machine learning; Spatial database; Spatial decision-support system

**References**


**Further reading**


Geographic information science

Matt Duckham
RMIT University, Australia

Geographic information is information concerning the spatial location and spatial relationships of phenomena and processes at or near the Earth’s surface. Geographic information science (GIScience) refers to the scientific study of geographic information, its creation, characteristics, structure, analysis, and transformations, as well as the community of experts that specialize in those topics.

On its own, however, this definition of GIScience – as the scientific study of geographic information – is tautologous. Further explanation of the complexities of this definition, and exploration of its implications, are needed to understand more precisely what is, and perhaps equally importantly what is not, GIScience. This exploration is made more challenging because GIScience is inherently multidisciplinary, drawing on expertise from disciplines such as geography, computer science, measurement science, mathematics, statistics, psychology, and design, amongst many others. This diversity has historically made it difficult to identify a common core of knowledge that makes up GIScience, as distinct from the many related areas of inquiry.

In seeking a deeper understanding of GIScience, three complementary perspectives are explored in this entry. First, a historical perspective summarizes some of the established definitions of GIScience, and some of the longstanding debates about the status of GIScience. A second perspective looks at two distinct styles of inquiry common in GIScience: as information science and as data science. Finally, a third perspective looks in more detail at GIScience from the perspective of the shared expertise common to those connected with the field.

History

A good place to start exploring the definition of GIScience is to look at the history of the usage of the term. The term “GIScience” was first coined in 1992 by Michael Goodchild (Goodchild 1992). The term was introduced better to distinguish between the technologies associated with geographic information (primarily geographic information systems, GIS) and the science behind these technologies (i.e., geographic information science, GIScience). At the time, the intellectual validity of a science of geographic information was a matter for polarized debate. Many of the most cogent criticisms of GIS were captured in the 1995 book *Ground Truth: The Social Implications of Geographic Information Systems* (Pickles 1995). Amongst these criticisms were included: the characterization of GIS as a purely technical exercise with no intellectual content; the association of GIS with purely positivist forms of inquiry, marginalizing other forms of knowledge; and the view of GIS as a technical evolution of traditional mapping technologies, a new mechanism for the continued empowerment of military and other elites. At the other extreme, GIS had been portrayed as the savior of geography (e.g., Openshaw 1997).
Its strongest proponents argued GIS provided a mechanism for applying geographical principles to a wide range of real-world problems, at a time when geography was under considerable pressure to demonstrate its import and impact.

Today, the storm of controversy has subsided somewhat. However, the debate left an enduring legacy to the emerging field of GIScience. Even the most technically minded GIScientist would today be expected to have a keen awareness of such criticisms. For example, any GIScientist would usually have a sophisticated understanding of the importance of recognizing uncertainty in any geographic information (see the section on “Expertise”). On the other hand, even the most implacable critic of GIS would typically acknowledge that important intellectual contributions have been made by researchers in the area of GIScience over the past 20 years.

The use of the term “GIScience,” as distinct from “GIS,” grew rapidly, with several of the leading academic journals in the fields changing their names (from the *International Journal of Geographic Information Systems* to the *International Journal of Geographic Information Science* in 1997; and from *Cartography and Geographic Information Systems* to *Cartography and Geographic Information Science* in 1999) and a biennial conference called *GIScience* starting in 2000.

Various subsequent definitions have elaborated on the definition of GIScience as “the science behind the systems,” in turn emphasizing different aspects. For example, Keith Clarke defines GIScience as “the discipline that uses geographic information systems as tools to understand the world” (Clarke 1997). Another often-quoted definition is as “the development and use of theories, methods, technology, and data for understanding geographic processes, relationships, and patterns” (UCGIS bylaws, quoted in Mark 2003). An attempt at a fuller, more detailed definition was the focus of a 1999 National Science Foundation workshop, defining GIScience as “the basic research field that seeks to redefine geographic concepts and their use in the context of geographic information systems. GIScience also examines the impacts of GIS on individuals and society, and the influences of society on GIS. GIScience re-examines some of the most fundamental themes in traditional spatially oriented fields such as geography, cartography, and geodesy, while incorporating more recent developments in cognitive and information science” (Mark 2000).

Mark (2003) provides an in-depth discussion of several of these definitions, while others have considered in similar detail specific aspects of the definition, content, history, focus, and contribution of GIScience (see, for example, Goodchild 2010, 2011; Longley 2000; Schuurman 2000).

Amongst the key themes underlying all such definitions, it is possible to identify two distinct styles of inquiry, two complementary approaches to creating new scientific knowledge connected with geographic information: GIScience as information science, and GIScience as data science.

**GIScience as information science**

*Information science* is concerned with the fundamental properties and behavior of information: how information is created, communicated, transformed, and used in both artificial and natural systems (cf. Mark 2003). Correspondingly, GIScience is often regarded as “that branch of information science that deals with the geographical domain, or as the set of fundamental scientific questions raised by geographical information and the technologies that collect, manipulate, and communicate it” (Goodchild 2011).
GEOGRAPHIC INFORMATION SCIENCE

The “GIScience as information science” paradigm is best exemplified by one of the most highly cited GIScience papers of all time: “Point-Set Topological Spatial Relations” (Egenhofer and Franzosa 1991).

Intuitively, topology concerns those properties of, and relationships between, (spatial) objects that are invariant under rubber-sheet transformations. For example, whether or not a point is inside a polygon is a topological relationship, because it is unaffected by rubber-sheet transformations such as stretching. The area of a polygon is not a topological property, however, since stretching and distorting a polygon will alter its shape and so its area.

The question posed by Egenhofer and Franzosa (1991) is: What are the possible topological relations between two simple spatial regions (i.e., spatial regions in the plane that are connected – all of one piece – and contain no holes)? The answer to this question is found by looking at the set-based intersections between the boundaries and interiors of the two regions. An exhaustive search of all possibilities reveals eight possible distinct relationships, each of which is uniquely defined by the four intersections between the two region boundaries, the two region interiors, and each region boundary with the other region’s interior. Table 1 summarizes the eight possible topological relations based on these four intersections (hence, termed the “four-intersection model”).

The Egenhofer and Franzosa paper highlights several important features in the context of the “GIScience as information science” paradigm. The paper shows mathematically that these eight topological relations are jointly exhaustive: the topological relations between any pair of simple regions can be described using one of these eight possibilities. In other words, it is not possible to draw a pair of simple regions whose topological relations cannot be described by one of these eight possibilities. The relations are also pairwise disjoint: any pair of simple regions must be described by exactly one of these possibilities. Assuming we have complete information about the regions, there can be no ambiguity – only and exactly one of these relations must hold.

Thus, the four-intersection model answers a fundamental question about geographic information, and helps us know what is and is not possible in this specific case of topological relations between simple regions. The question answered is about geographic information independent of specific domains of geographic knowledge. In practice the regions in question might be land parcels in a cadaster; regions of different land cover; or political, administrative, linguistic, or religious regions. As long as the boundaries and interiors of the regions are precisely defined (but see section on “Expertise in uncertainty,” below), we can be certain that the topological relations between any pair of regions can be classified into exactly one of these eight possible relations.

Precisely because such knowledge about the structure of geographic information is application-independent, it can in turn be most valuable in building tools such as spatial databases and geographic information systems (i.e., “the science behind the systems”). Today, with few if any exceptions, every spatial database and GIS will provide functions to compute the Egenhofer and Franzosa topological relations between simple regions. These functions form part of the building blocks of many higher-level spatial analyses and underlie the practical application of geographic information to diverse problem domains.

Following the analysis pioneered by Egenhofer and Franzosa, other studies have followed similar patterns. Subsequent GIScience research has investigated, for example, the topological relations between a region and a line, between directed lines, between regions with holes,
Table 1  Four-intersection topological relations between two simple regions. Adapted from Egenhofer and Franzosa (1991) by permission of Taylor & Francis Ltd., www.informaworld.com.

<table>
<thead>
<tr>
<th>Name</th>
<th>Example</th>
<th>Boundaries of A and B</th>
<th>Boundary of A, interior of B</th>
<th>Interior of A, boundary of B</th>
<th>Interiors of A and B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disjoint</td>
<td></td>
<td>Do not intersect</td>
<td>Do not intersect</td>
<td>Do not intersect</td>
<td>Do not intersect</td>
</tr>
<tr>
<td>Meet</td>
<td></td>
<td>Intersect</td>
<td>Do not intersect</td>
<td>Do not intersect</td>
<td>Do not intersect</td>
</tr>
<tr>
<td>Overlap</td>
<td></td>
<td>Intersect</td>
<td>Intersect</td>
<td>Intersect</td>
<td>Intersect</td>
</tr>
<tr>
<td>Equals</td>
<td></td>
<td>Intersect</td>
<td>Do not intersect</td>
<td>Do not intersect</td>
<td>Intersect</td>
</tr>
<tr>
<td>Contains</td>
<td></td>
<td>Do not intersect</td>
<td>Do not intersect</td>
<td>Intersect</td>
<td>Intersect</td>
</tr>
<tr>
<td>Inside</td>
<td></td>
<td>Do not intersect</td>
<td>Intersect</td>
<td>Do not intersect</td>
<td>Intersect</td>
</tr>
<tr>
<td>Covers</td>
<td></td>
<td>Intersect</td>
<td>Do not intersect</td>
<td>Intersect</td>
<td>Intersect</td>
</tr>
<tr>
<td>Covered by</td>
<td></td>
<td>Intersect</td>
<td>Intersect</td>
<td>Do not intersect</td>
<td>Intersect</td>
</tr>
</tbody>
</table>
between regions on the sphere, and so on. Each of these studies has revealed something new and interesting about the different possible topological relationships between different types of spatial object. However, it would not be true to say that such research could continue indefinitely. As new studies uncover fewer and fewer novel and interesting patterns, the interest of the GIScience community in further such studies diminishes accordingly. It is likely that there are today no new scientific insights to be gained from further similar studies. An investigation of the topological relations between a region with a hole and a directed line, for example, would be unlikely to yield more than a mechanical application of existing knowledge based on Egenhofer and Franzosa’s already established methodology. In other words, the body of knowledge in GIScience has grown, and research has moved on.

GIScience as data science

An alternative and complementary perspective on GIScience is as a data science. Where “information science” is concerned with creating new knowledge about the fundamental properties of information, “data science” is concerned with extracting new and generalizable patterns and knowledge from data about the world. Thus, the paradigm of “GIScience as data science” is most in accord with the definition of GIScience as “the discipline that uses geographic information systems as tools to understand the world” (Clarke 1997).

Data about our geographic world continue to grow in volume, the variety of data sources, and the speed of updates to data. It should perhaps come as no surprise, then, that there is a growing body of contemporary examples of GIScience as data science. One recent exemplar of the “GIScience as data science” paradigm is the paper “Exploring Place through User-Generated Content: Using Flickr Tags to Describe City Cores” (Hollenstein and Purves 2010).

In their paper, the authors examine a large dataset of user-generated content. Specifically, the dataset comprises 8 million images, and the associated human-generated tags for those images, voluntarily posted by human users on a social media website. The questions posed by the paper are twofold. At one level, the paper asks what generalizable knowledge we can extract from this data source about how humans use vernacular place names. At a second level, the paper also aims to increase our understanding of the fundamental nature of user-generated geographic data more broadly, what characteristics and properties one might expect other such datasets to exhibit.

At the first level, Hollenstein and Purves (2010) identify a variety of interesting patterns and regularities in human usage of vernacular place names. For example, the data allow the construction of spatial “footprints” that capture that region of London that is commonly referred to as “Camden,” or the region of Chicago that is commonly referred to as “downtown.” Such vernacular regions have rather different characteristics to the administrative regions that might even bear the same appellation. Unlike the administrative London Borough of Camden, the vernacular region (the places that people typically refer to as “Camden”) does not have a clear or crisp boundary. There tends to be a core of places that most people agree can be referred to as “Camden,” and many places that most would agree cannot be called “Camden.” But there typically exists a gradation between the two: a periphery where opinion is more or less divided, where, as we move away from the core, we encounter places that fewer and fewer people would refer to as “Camden.” Certainly, the
administrative regions appear only infrequently to coincide exactly with the corresponding vernacular region. Maps of vernacular regions are quite different from those of administrative regions of the same name.

At the second level, the paper also adds to our understanding of the nature of user-generated geographic content, such as that generated via social media. For example, the paper shows that while some tags (such as “london”) are commonly used by a broad range of users, some other tags (such as “northlondon”) are strongly biased towards the most prolific users. This phenomenon, called the *Pareto* principle (sometimes the “80:20 principle”), is widely found in user-contributed data: the majority of the content (e.g., 80%) is contributed by a minority of prolific users (e.g., 20%). Even leaving aside those sections of society that are in any event expected to be underrepresented in data generated through social media (such as elderly people who are not as used to the latest technology, rural communities that have limited access to high-speed digital networks, or poorer communities that have reduced opportunities to access the latest mobile computing technology), it is frequently the case that such data contain inherent biases towards a small number of highly prolific users.

Thus, the Hollenstein and Purves paper highlights important features of the “GIScience as data science” paradigm, just as the Egenhofer and Franzosa paper did for the “GIScience as information science” paradigm. The work reported in Hollenstein and Purves (2010) relies on GIS technology and GIScience (as information science) concepts. Kernel density estimation, a technique from spatial analysis and a basic GIS function, is used to generate the vernacular place-name “footprints” from tagged point locations; spatial databases and GIS are required to store and map the large datasets and the analysis results. These results can help generate new and generalizable knowledge about our geographic world (e.g., that vernacular place names are expected to differ systematically from for administrative place names, lacking crisp boundaries or direct coincidence), and about data about our geographic world (e.g., that social geodata is expected to contain substantial biases that need to be accounted for when analyzing and interpreting patterns in such data).

**Expertise in GIScience**

The first perspective of this entry on GIScience was historical: the genesis and existing definitions of the term. The second perspective was methodological: the different mechanisms and focuses by which new knowledge is created in GIScience. The third and final perspective is descriptive: what expertise is common to practitioners of GIScience. A description of this common expertise is itself another form of definition. The written definitions encountered previously attempt to delimit GIScience by *intension* – specifying the content and focus of the field. By contrast, a description of the expertise common to research and researchers across GIScience can be thought of as a definition by *extension* – the topics and knowledge that together make up GIScience.

With a broad and interdisciplinary field such as GIScience, the precise list of topics of expertise is necessarily a somewhat subjective choice. However, it is possible to identify with reasonable objectivity a minimal list of six areas of expertise that together are the most strongly associated with GIScience: expertise in uncertainty, in dynamism, in scale, in spatial cognition, in information design, and most importantly in the structure of spatial information. Arguably, at least one, and usually more than one, of these topics
lie at the core of any research in GIScience or application of knowledge from GIScience.

Expertise in the structure of spatial information

Ask a group of GIScientists what they regard as the most important principle that underpins their expertise and almost as one they are likely to volunteer: “Tobler’s first law.” This principle posits that near things in space are more similar than distal things. This property is manifest in all geographic information. Wherever we look, geographic information obeys this law. Information about one location immediately conditions one’s expectations about what might be observed at other, nearby locations. Indeed, this property of information is measured statistically as “spatial autocorrelation”: the degree of interdependency of observations in geographic space. Autocorrelation is ubiquitous in geographic information, so much so that it is sometimes almost forgotten, regarded as “common sense” rather than “law.”

However, looking at other domains of expertise it is clear that autocorrelation is more than simple common sense. There are innumerable examples of research in other fields that start with assumptions to the contrary: that information in space is random. Experts in computer science and distributed systems, for example, may frequently assume that computing devices (such as sensor nodes) are randomly distributed through space; experts in ecology may assume that movement of organisms through space is random; experts in social networks may assume random graphs as null models of social networks. Most GIScientists would quickly and even unconsciously avoid such assumptions in their own work and identify such assumptions in others’ work. That is not to say that these other fields lack validity; research in other areas such as computer science, ecology, and economics is founded on its own specific areas of expertise. Nor would it be true to represent GIScience as the only field with expertise in autocorrelation; other areas of specialism, including geostatistics and, of course, geography, can equally point to high levels of expertise in this topic. However, the implication is that an awareness of the unavoidable autocorrelation of geographic information is not so obvious as to be common sense; and expertise in autocorrelation is at the core of GIScience.

Indeed, autocorrelation is not the only non-random aspect to the structure of geographic information. Spatial heterogeneity (also referred to as the statistical concept of “nonstationarity”) is an inherent feature of geographic information that has been proposed as a companion to Tobler’s first law (TFL) as one of the fundamental “laws” of geographic information (e.g., Goodchild 2011). Spatial heterogeneity arises because different locations in space may exhibit arbitrarily different properties. In other words, there is no “average” place: phenomena in geographic space are inevitably unevenly distributed, and an individual location cannot reasonably be treated as a sample from a population of idealized locations. Instead, the systematic similarity of nearby places (TFL) itself varies spatially, leading to an expectation of “clumps” of things in geographic space.

Spatial heterogeneity is closely related to the modifiable areal unit problem (MAUP), another area of expertise familiar to any GIScientist. The essence of MAUP is that any observed properties of geographic information that are based on an analysis of arbitrary (i.e., “modifiable”) regions (i.e., “areal units”) are not guaranteed to be independent of those units. A practical example of MAUP can be found in any first-past-the-post democratic election. The boundaries of the electoral regions can themselves have a defining
GEOGRAPHIC INFORMATION SCIENCE

effect upon who wins the election. Indeed, deliberate manipulation of these boundaries in order to gain advantage in elections is common enough that there exists a word to describe it: “gerrymandering.” Gerrymandering works because voters are spatially heterogeneous (i.e., “clumped”). Depending on where electoral boundaries are drawn, they will either split or aggregate votes from voter clusters, to the advantage or disadvantage of one or another political party.

These and other aspects of the inherent nonrandom structure of geographic information together form a core of expertise within GIScience. While GIScientists are not alone in possessing this expertise, any GIScientist can be expected to be knowledgeable on TFL, autocorrelation, spatial heterogeneity, and MAUP, and to apply that expertise in his or her research or work. However, while expertise in the structure of geographic information is a necessary skill for a GIScientist, it is not on its own sufficient expertise to define GIScience. Rather, in GIScience, this expertise is applied in combination with five other core areas of expertise.

Expertise in uncertainty

Uncertainty is an inherent and endemic feature of geographic information. Uncertainty arises because with few, if any, exceptions geographic information can always be safely assumed to be imperfect. Imperfection in geographic information arises in three main ways: inaccuracy, imprecision, and vagueness.

Inaccuracy concerns a lack of correctness in information, where the geographic information deviates in some way from “true” reality, or what is taken to be true. Imprecision concerns a lack of detail in information, where the geographic information is not as specific as it could be.

Imprecision and inaccuracy can occur independently: the statement “London is in Europe” is at once less precise (provides less detail) and more accurate (accords more closely with the actual state of affairs) than the statement “London is in France.” Although imprecision and inaccuracy are distinct concepts that can occur independently, they are related. Generally, being less precise tends to be associated with greater accuracy; conversely being more precise often leads to lesser accuracy.

Vagueness concerns the existence of borderline cases. A vague predicate will lack a crisp boundary between instances where that predicate holds, and where it does not. For example, the vernacular region of “Camden” pictured in Figure 1 is vague: it is not possible to precisely delineate where the vernacular region of Camden begins and ends. However, the administrative London borough of Camden is not vague: it does have a precisely defined fiat boundary.

Today, a considerable body of GIScience research has been devoted to the capture, description, computation with, and communication of information about these three aspects of uncertainty. Any GIScientist would be expected to have a sophisticated understanding of the importance of uncertainty in any analysis of geographic information. To date, the field has not always been successful in embedding this understanding in GIS technology. Today, most common GIS and spatial databases will, if queried, provide the distance between two points to a great many significant figures, usually as a double-precision floating point number irrespective of the precision or accuracy of the query points. This high level of numerical precision frequently equates to a precision of millimeters, micrometers, or even finer detail in the world. This high and unnecessary level of precision gives rise to inaccuracy – a problem
termed “false precision.” The actual distance between the two points will almost certainly not be correct if represented at such an excessively high level of detail (even were such an answer to make sense).

Some practical applications of uncertainty have arisen, such as in the inclusion of metadata on uncertainty in international standards for spatial data transfer. But even though the technology of GIS lags behind, every GIScientist can be
expected to have a sophisticated awareness of
the structure and implications of the unavoid-
able uncertainty associated with geographic
information.

**Expertise in dynamism**

We live in a dynamic, ever-changing world. It
follows that information about our world is nec-
essarily time-dependent. As a result, it will be
no surprise that GIScience research has a long
history of focusing on both spatial and spatiotem-
poral information.

One of the most basic mechanisms for adding
dynamism to static spatial information relies
on the “snapshot” metaphor. In the snapshot
metaphor, dynamic geographic phenomena
are represented as collections of timestamped
states. This representation may be familiar, for
example, in the everyday tracking of moving
objects, such as people. A set of timestamped
\((x, y, t)\) coordinates can be used to represent the
movement of an object in the plane.

An alternative to storing information about
states of the world through time is to store
information about the occurrences in time. Two
frequently distinguished types of occurrence are
events and processes. Events are “countable” hap-
penings, such as a traffic jam, a landslide, or the
arrival of a driver at her destination. Processes are
not countable, and include dynamic phenomena
such as traffic flow, erosion, and wayfinding.

To an extent, it is possible to infer information
about events and processes from information
about the states through time. Conversely, informa-
tion about states of the world through time
can be used to infer information about events
and processes. For example, knowing that a
person was located at home at time \(t_1\) and at
work at time \(t_2\), it is possible to infer that person
has departed home and arrived at work at some
(unknown) time between \(t_1\) and \(t_2\).

As with uncertainty, today’s GIS technology
lags somewhat behind GIScience research and
expertise in supporting dynamism and spatiotemporal information. Many GIS and spatial
databases do offer some support for storing
and analyzing moving point objects. However,
evolving regions, such as changing land-cover
regions, are less well supported. Nevertheless,
the pursuit of better computational models
for storing and processing information about
dynamic geographic phenomena has a long
history in GIScience.

**Expertise in scale**

The term “scale” has a variety of distinct
meanings, which include “spatial and temporal
extents” (e.g., an environmental occurrence
might be said to be a “large-scale event” if
it has large spatial and/or temporal extents);
“levels of measurement” (e.g., “nominal” versus
“ordinal” measurement scales); and “map scale”
(the relationship between distances as depicted
on a map and their corresponding distances in
the world).

However, in the context of expertise in
GIScience, the term “scale” is most strongly
connected to the scale-dependence of geographic
information. Scale-dependent information
reveals different patterns depending on the level
of detail at which it is analyzed. For example,
different levels of detail in a topographic dataset
can result in diametrically opposed landforms
being derived from this data. What appears to
be a peak at one scale can easily appear to be
a pit at another scale. This kind of effect can
be observed across many other types of geo-
graphic information, including in data about the
movement of animals and people; ecological and geophysical region boundaries; and even cartographic information about the structure of maps. In short, the level of detail at which information is captured and represented can have a defining effect upon the geographic phenomena that can be identified from that information (Figure 2).

Expertise in spatial cognition and spatial thinking

GIScience is not only concerned with computational models of geographic information. Over millions of years of evolution, humans (and other animals) have evolved sophisticated mechanisms for thinking and reasoning about objects and movements in geographic space. The structure and characteristics of human spatial reasoning are in stark contrast with our automated spatial reasoning procedures. For example, in a classic experiment, Stevens and Coupe (1978) showed that human subjects tended to judge that the city of Reno, in Nevada, was to the east of San Diego, California, when in actuality it is to the west.

The explanation, supported by other subsequent studies, indicates that humans do not store and reason with the geographic locations of Reno and San Diego directly. Instead, a better explanation is the use of a simple heuristic to remember only that Reno is in Nevada, which is broadly to the east of California, which in turn contains San Diego. The resulting rule (that things in Nevada are generally to the east of things in California) is frequently — though not always — correct. These simplified hierarchical structures may assist humans with efficient memory and reasoning connected with spatial relationships. Numerous subsequent experiments have shown that human reasoning exhibits systematic distortions in connection with a range of spatial reasoning tasks. Indeed, it is argued that such distortions are not “faults”; instead they are necessary features of human spatial cognition.

In this way, human spatial cognition and geographic information systems are in stark contrast. Precise digital information systems, such as GIS, are able to rapidly and reliably compute the cardinal direction between any pair of stored locations. On the other hand, humans routinely reason with and communicate uncertain geographic information, something that, as we have seen, today’s GIS struggle to cope with. A major theme in GIScience is an understanding of the mismatches between human spatial cognition and digital spatial computation. GIScientists typically possess expertise in the structure and types of inherent distortions in human spatial cognition and spatial thinking more generally.

Expertise in cartography and geovisualization

Finally, GIScientists are expected to possess expertise in the design of interfaces to spatial information, whether static cartographic
interfaces or more interactive systems for geovisualization.

GIScience draws upon a long history of expertise in cartographic design. The fundamental principles used in the design of traditional maps (e.g., map projections; the use of color, symbols, text, and type; and the different types of map, whether topographic, thematic, or schematic) are familiar to all GIScientists, evident in every GIS textbook. The problems posed by digital mapping have also been amongst the most active and successful topics for GIScience research.

For example, many ingenious algorithms for cartographic generalization have been developed with the aim of automatically generating maps at a specific level of detail. Cartographic generalization is the process by which the graphical clarity of a map is enhanced through simplifying or otherwise distorting the shapes and geometries depicted. The problem of positioning and orienting text labels on maps has similarly received much attention. In both cases, individual solutions are often straightforward, such as the famous Douglas–Peuker algorithm for line simplification. However, such per-object solutions often lead to adverse interactions between objects (for example, where moving a label to an individually ideal position then obscures another label’s optimal position). It is at least in part due to this research that today most digital mapping systems are able to make as good a job of map generalization and labeling as ever did a moderately experienced cartographer, and in a fraction of a second.

Indeed, the so-called map metaphor is still evident in the design and structure of the interface to most of today’s GIS. Like traditional cartographic maps, GIS interfaces tend to offer primarily static, planimetric views of the world, with lesser support for user interaction and nonvisual, dynamic, or three-dimensional representations of the world. GIScientists are keenly aware of such limitations, and can be expected also to have expertise in the principles of using interactive computer interfaces to geographic information. The topic of geovisualization is concerned with using the dynamic, interactive, and multimedia capabilities of computers to help users explore and gain insight into the spatial and temporal patterns and relationships embedded in geographic information. Today, geovisualization is just as active a topic in GIScience as cartographic generalization ever was, and hopefully will generate just as much benefit in the long term.

In summary, how geographic information is presented to a person is expected to have an important effect on what knowledge that person derives from the information, and even what decisions that person makes based on that information. Knowledge of the principles of presentation of and interaction with geographic information is this entry’s sixth core area of GIScience expertise.

**Expertise summary**

GIScientists can be relied upon to possess a sophisticated understanding of each of our six core areas of GIScience expertise (structure, uncertainty, dynamism, scale, spatial cognition, and design of geographic information), and most GIScientists would consider themselves experts in the majority of these topics. There are naturally, however, many other related areas of expertise that have close connections to an interdisciplinary field such as GIScience, and which many GIScientists will possess. An understanding of the social context of GIS, for example, is an area that we have already seen GIScientists will have at least a basic awareness of. From a more technical perspective, most GIScientists will understand at least the basic principles behind
spatial algorithms and spatial databases. However, while it is true that high levels of expertise in such areas exist within GIScience, it is arguably not common across all GIScientists. Thus, these six areas of expertise represent a summary of the topics that any GIScientist would usually have expert knowledge about, and by extension the topics that make up the core of the field of GIScience.

SEE ALSO: Cartographic design; Modifiable areal unit problem; Scale; Spatial thinking, cognition, and learning; Spatiotemporal analysis; Tobler’s first law of geography; Topological relations; Toponymy; Uncertainty; Visualization

References


Geographic information science and technology: educational directions

Josef Strobl
University of Salzburg, Austria

Geographic information science and technology (GIS&T) became established as a term, describing what initially used to be known as simply “GIS,” with the initial publication of the GIS&T “body of knowledge” (DiBiase et al. 2006). The original concept of geographic information systems advanced towards geographic information science, recognizing the increasing independence from the original parent disciplines and the establishment of a generic spatial sciences discipline.

In parallel, academic education embarked on a move from offering “GIS classes” in a variety of spatial disciplines towards defining the curricula and syllabi for full geospatial education programs. This move went hand in hand with recognition as a professional field and the designation of distinct professional roles and job profiles by governmental bodies and professional organizations.

Due to the multifaceted and multidisciplinary roots of GIS&T, and a variety of labels such as geoinformatics, geomatics, or spatial information management, educational programs evolved across many different academic traditions. The community of GIS&T educators has contributed to strengthening common understandings and objectives among these varied constituencies, particularly through the development of joint reference documents for curricula and qualifications.

Disciplines and traditions engaging in GIS&T education

“GIS education” initially was rooted in only a few types of departments, or more correctly, advanced by individuals from only a few disciplines. The founding institutions of the National Center for Geographic Information and Analysis (NCGIA) (www.ncgia.ucsb.edu) were strong in geography as well as surveying and this was typical of this phase. These efforts were complemented by initiatives from cartography (which in turn is frequently rooted in the above subjects), planning, and computer science.

The focus of GIS&T education (Unwin et al. 2012), of course, varies depending on the respective home discipline, with only relatively few disciplines serving as backdrops for full degree programs in GIS&T.

- Geography (as a discipline, in some academic environments sharing departmental homes with other domains): due to the “quantitative revolution” and contributions from theoretical geography, close links with cartography, and engagement of geographers with numerous other disciplines and application domains (Baerwald 2010), geography worldwide certainly is the most prolific discipline when it comes to graduating geospatial professionals, typically with an applied orientation.

- Geodesy (or surveying engineering) contributes strengths in reference frameworks, spatial data acquisition, and mathematical skills. To enhance their appeal to students and reflect changes towards digital techniques,
many programs have been rebranded as “geomatics.”

- Cartography as a professional and independent academic discipline in many places is struggling to maintain its independent existence, and widely re-establishing itself successfully within a GIS&T framework.
- Planning, with orientations towards landscape, environment, and urban planning, is considered one of the fields where GIS originated. This link currently is emphasized through the emerging field of “geodesign.”
- Computer science (CS), with its core areas of data(base) management, communication protocols, distributed systems, and, of course, software design and development, is an indispensable foundation for GIS&T. While CS plays an important role in geospatial professional qualifications, this discipline rarely leads full academic programs.

These diverse academic backgrounds, traditions, and practices soon led to the insight that academic program orientation, syllabi, and resulting qualifications varied widely depending on discipline, and even on individual department and faculty orientation. It was due to this observation that one of the initial NCGIA initiatives was directed towards establishing a common platform, or at least a joint reference framework for education through the “core curriculum in GIS.”

Curriculum development


Additional early impulses towards including geospatial methods competences in the discipline of geography were provided by initiatives such as the “Geographer’s Craft” and the “Virtual Geography Department” to develop open educational materials. Results offered important insights into constraints and motivations for integrating GIS&T into discipline-specific frameworks, and contributed to the debate of dedicated GIS&T study programs versus modular integration into traditional curricula.

The core curriculum designation turned out to be a less than fitting name for this product, even though the materials had a huge impact worldwide: the distributed documents were not really structured as a curriculum, but rather as materials that could be used in a variety of combinations by instructors. Due to the paucity of learning materials and textbooks at that time, these lesson units frequently were used as lecture handouts and even self-learning materials. Furthermore, this did not turn out to be “core” as a common denominator across disciplines and program orientations, but rather a broad collection from the authors’ different backgrounds and approaches.

One measure of the success of the core curriculum initiative is the fact that it generated several spin-off projects: a “remote sensing core curriculum” was attempted, as well as a 53-unit “GIS core curriculum for technical programs (CCTP)” as a first manifestation recognizing the need to distinguish more technically oriented education in (community) colleges.

Technical advances and a discipline evolving rapidly in its early days led to quick obsolescence of some units, and even more newly emerging themes. The unit on manual digitizing (CCTP unit 13 plus a dedicated tutorial, CC unit 31) is but one example of materials needing updating,
leading to the subsequent “core curriculum in GIScience” initiative.

With the latter not able to replicate the resounding success of the original CC, the University Consortium for Geographic Information Science (UCGIS) with its “model curricula” project now took the lead towards developing a “GIS&T body of knowledge (BoK)” (DiBiase et al. 2006). This reference document is structured as a hierarchy, built from knowledge areas containing units, which in turn are divided into 330 topics. Each topic includes a list of 5–10 educational objectives for a total of 1660 educational objectives.

The BoK is accepted as a major milestone, the critical attention received leading to the BoK-2 initiative launched in 2013. In parallel, a European Commission project provided support for “Geographic Information: Need to Know” (www.gi-n2k.eu), a multipartner network aiming “to improve the way in which future GI professionals are prepared for the labor market so that the GI sector in general can evolve in a dynamic and innovative way.”

A parallel initiative with a primary focus on two-year colleges is the geospatial technology competency model (GTCM) based on an industry competence model established by the US Department of Labor (2016). This approach is based on an analysis of current job profiles, addressing the development of entry level geospatial programs for technicians.

These curricula follow a range of objectives: providing a framework for developing academic programs, offering a baseline for accreditation, orientation for prospective students and employers alike, and a reference for credit transfer or admission into more advanced study programs.

Practical work with curricula, syllabi, and learning objectives has been explored with software tools based on hyperlinks and a strong semantic foundation. These platforms support the checking of curricular consistency, identification of learning pathways, comparison and accreditation of academic programs, and generally a top-down analysis of geoinformatics programs.

**Levels of GIS&T education programs**

GIS education is established on all levels of education. K-12 is part of a separate debate (National Research Council 2006) building on concepts of spatial awareness and spatial thinking, with links to “spatial citizenship.” Only a few curricula explicitly address GIS&T as a methodology or technical skills for primary or secondary schools (Schulze, Gryn, and Kanwischer 2014), while exposure of pupils to geospatial technology interfaces is increasingly promoted (http://connected.esri.com).

The tertiary level addressed through the above listed curricula is differentiated by established undergraduate versus graduate program categories. On an undergraduate level, GIS&T is widely present through classes in a broad range of programs across numerous disciplines, with primarily technical and junior colleges addressing the potential job market for designated undergraduate geospatial qualifications.

The broadest presence and visibility of academic programs is observed on the Master’s level (Lukinbeal and Monk 2015) with more or less recently established offerings worldwide. MSc (or MTech in some regions) is considered a key qualification for geospatial professionals leading organizational departments or major projects or those in technical development, essentially serving as a passport for establishing a career in this industry.

Geoinformatics MSc programs currently are offered worldwide in numerous different curricular flavors and academic backgrounds.
They range between 1 and (a majority) 2 years in duration, and, in combination with admission from Bachelor levels, are rapidly replacing 5-year diploma studies in European and Asian countries. Due to the transdisciplinary character of GIS&T, a majority of MSc students are admitted based on other undergraduate qualifications.

Dedicated doctoral programs in geographic information science are a much more recent development and part of the pathway from “systems” to “science.” PhDs with a GIS&T focus have graduated from the leading departments worldwide since the emergence of the discipline, but the first designated PhD programs were established only recently (see for example the National Science Foundation (NSF) Integrative Graduate Education and Research Traineeship (IGERT) initiative, or Blaschke, Strobl, Donert 2011).

Due to the broad interest in and relevance for a range of disciplines, GIS&T competences increasingly are offered as “minors” or in similar formats complementing a range of majors from a geospatial methodology perspective. Depending on curriculum frameworks, a minor typically carries approximately one semester’s worth of credits and frequently is built from introductory classes plus a range of choices from existing GIS&T classes. Efforts discussing more rigid curricular frameworks for minor programs have not yet taken off.

Much longer experience exists in the closely related field of academic certificates or postgraduate diplomas, offered under different frameworks at tertiary institutions. These add-on qualifications either are taken in parallel to main study programs, or are used as continuing education facilities. Admission to a certificate in GIS or geoinformatics can be open to anyone interested, while admission to a postgraduate certificate requires a prior academic qualification.

Different educational strategies and policies are followed regarding the sequencing of GIS&T qualifications throughout academic qualification pathways. Undergraduate full GIS&T programs primarily aim at technical qualifications; completing this kind of program requires professional embedding into a broader setting – similar to other entry level technology qualifications with little background in application domains.

Many people choose to add GIS&T qualifications after they are already established in an academic discipline, which leads to the prevalence of MSc-level and postgraduate certificate programs. The job market and ultimately career prospects for GIS&T-only graduates effectively are more limited than for mixed domain-and-methodology qualifications.

Finally, essential and recognized GIS&T education occurs outside of curriculum-controlled environments. Short intensive courses and “summer schools” are offered to motivated individuals and aim at leveraging the advantages of interdisciplinary, creative, and less structured approaches. Excellent and very different examples are the “Vespucci Institutes” (www.vespucci.org) and the “Open Source Opportunities in GIS” summer schools (e.g., http://www.sigte.udg.edu/summerschool2014).

**Focus of GIS&T education programs**

Full geospatial programs started to emerge from the early 2000s, with earlier academic offerings predominantly taught as individual classes in a variety of programs. While GIS&T classes integrated in any discipline’s curriculum depend on demand, departmental strategy, and faculty availability, dedicated programs typically are explicitly linked to an academic department. This department discipline’s traditions then
reflect onto the focus and orientation of a GIS&T program’s curriculum.

- A *spin-off* study program emerging from a geography or earth sciences department is expected to offer strong conceptual and methodology foundations, stressing insights into application domains and cartographic communication competences.

- A *converted* program, for example from traditional geodesy to geomatics or from a surveying-based cartography (mapping) focus, will retain a majority of components from the original background discipline, keeping its technical orientation and focus on skills in either data acquisition or geovisualization products.

- *Merger* programs were established at universities where critical mass of faculty could be combined from qualified teachers across different departments, for example, geography, computer science, sociology, forestry, planning, and so on. While this approach potentially could result in very strong programs it tends to have issues of leadership and of compromise by offering alternative tracks instead of a clear focus.

- *New paradigm* study programs aim at developing new qualifications rooted in traditional disciplines. One example is the emergence of “geodesign” studies on undergraduate as well as graduate levels. Core ideas have a background in the (landscape) planning and architecture disciplines, but implementation predominantly occurs in strong GIS&T environments, branching off existing GIS&T study programs. Geomarketing, geohealth, and big Earth data are other examples for establishing paradigms towards GIS&T-linked disciplines.

- *Tracked* programs offer a GIS&T specialization, for example, in computer science (CS). Graduates with this kind of skill set are among the most sought-after professionals, in particular in the software industry. Due to the small footprint of GIS&T from a CS perspective, this option mostly is followed as an individual’s specialization.

- *Technical* programs are mostly established at undergraduate technical colleges (or universities of applied sciences) within a broader portfolio of study programs with a more vocational and thus stronger “training” orientation. GIS&T is identified as a professional field with perceived demand, and, depending on the respective institutional environment curricula, emphasizes software development, data acquisition, or sensor technology skills with a lesser focus on conceptual backgrounds.

While examples for all these “type models” exist, there are obviously further and mixed foci and traditions involved. Disciplinary backgrounds, programmatic positioning, and even regional specifics translate into a variety of nomenclatures for GIS&T study programs. “Geoinformatics” has taken the lead from simple “GIS” designations, with the latter increasingly being re-interpreted as “geographic information science,” although not always in the most suitable contexts: while perfectly suitable for a PhD program, it appears to be less adequate for a “GIS minor.” “Geomatics” derives from a specific disciplinary context (see above), and some newer designations with less ballast from academic traditions, such as “spatial sciences,” are getting traction.

The latter are an indication for increasing decoupling of GIS&T study programs from original parent disciplines, sometimes combining, for example, the “spin-off” with the “new paradigm” types discussed above. Examples such as the “Spatial Sciences Institute” at the University of Southern California (http://spatial.
GEOGRAPHIC INFORMATION SCIENCE AND TECHNOLOGY: EDUCATION

usc.edu), the “Interfaculty” Department of Geoinformatics at the University of Salzburg (www.zgis.at), or the U-Spatial unit at the University of Minnesota (http://uspatial.umn.edu) demonstrate the value of a trans-disciplinary home for GIS&T services, research, and education, offering an academic reference point for the “spatial perspective” increasingly of interest across numerous disciplines.

Distance learning and continuing education online

As an emerging field within a rapidly evolving technology environment, GIS&T could not establish the required quantities and qualities of human resources by regular patterns of academic education alone. The need to address experienced professionals and provide continuing education opportunities to the current workforce was already evident in the 1990s.

This demand has not changed, even after the implementation of regular study programs, due to the rapid growth of the industry, and even more so as GIS&T reaches into numerous other disciplines and professions, where individuals are exposed to the need for an add-on qualification for geospatial methods and technologies. Due to limitations active professionals face when attending full academic graduate programs, distance learning modes of delivery are an obvious choice for many.

As a first initiative, and now sustained for the longest time of any such initiative, the worldwide UNIGIS distance learning network of universities (www.unigis.net; Strobl 2011) in 1993 began to offer postgraduate programs. Its member universities have graduated thousands of students with MSc degrees or postgraduate certificates.

Penn State University’s program is another well-established and highly regarded option for flexible distance learning (DiBiase and Rademacher, 2005), with numerous additional offerings recently emerging primarily in North America, ranging from local colleges to full universities.

Online learning started opening up to a general public and to reach far beyond previous target groups with the arrival of massive open online courses (MOOCs) in the discipline. These courses were pioneered by Penn State University’s “Geospatial Revolution” MOOC (Robinson et al. 2015) in 2003, the success of which was followed and replicated by additional technology oriented courses and has started to branch out into specific application domains (geo-intelligence, geobusiness, geohealth). Clearly, MOOCs have taken an important role as an opportunity for many for a first academically inclined exposure to the field, reaching out to a huge number of potential students who might then opt for seriously studying the discipline.

Similarly, technology-specific online options for acquiring software skills are rapidly expanding. Technology vendors having established “virtual campus,” “online academy,” or online “university” platforms providing online training opportunities, complementing academic programs sometimes even to the degree of credit transfer.

Professional organizations and certification

All major professional organizations in this field have implemented working groups or commissions fostering educational initiatives. Some examples are as follows.

• ISPRS (International Society for Photogrammetry and Remote Sensing) TC6:
Education, Technology Transfer and Capacity Development: http://www2.isprs.org/commissions/comm6.html

- FIG (Fédération Internationale des Géomètres/International Federation of Surveyors) Commission 2: https://www.fig.net/commission2
- OSGeo (Open Source Geospatial Foundation) Education and Curriculum: http://www.osgeo.org/education
- ICA/OSGeo/ISPRS: “Geo for All” www.geoforall.org

These initiatives address the respective institutional membership to promote capacity building and continuing education. A specific mission of these and other organizations such as the Geospatial Information and Technology Association (GITA) (www.gita.org) and the Urban and Regional Information Systems Association (URISA) (www.urisa.org) is the establishment of certification programs aimed at distinguishing individuals, or even providing a framework for granting the right to practice a geospatial profession.

As an initiative launched by URISA in 2004 and with several other organizations as members, the GIS Certification Institute (www.gisci.org) is offering recognition as a “certified GIS professional” (GISP). Candidates need to pass a “GISCI geospatial core technical exam” plus a portfolio-based review based on ethics agreement, education, experience, and professional contributions. Recertification is required after a three-year period, based on continuing education and service to the profession.

Certification is frequently linked to legal codes, and therefore tends to have a national focus. An international example in a related discipline is PLATO, the South African Council for Professional and Technical Surveyors (www.plato.org.za), with a legal mandate reaching far beyond the surveying trade. To qualify, for example, as a “professional GISc practitioner,” candidates need to have obtained a qualification from a program accredited by PLATO, and to maintain this through continuing professional development (CPD).

In addition to professional certification, technically oriented certification programs have been established with a focus on mostly vendor-specific software skills, documenting individuals’ proficiency and familiarity with best practice in technology application. These either build on a collection of training courses assessed by some kind of exam or performance review, or are managed through testing centers. None of these (professional and technical) certifications should be confused, though, with the above-mentioned academic certificate programs.

SEE ALSO: Geographic information science; Geographic information system; Geography in higher education

References


A geographic information system (GIS) is a computer system for capturing, storing, querying, analyzing, and displaying geospatial data (Chang 2014). Geospatial data describe both the location and the attributes of spatial features. For example, to describe a road, its location (i.e., where it is) and its attributes (e.g., length, name, speed limit, and direction) are referred to. A GIS allows the user to manage road data and many other geospatial data, thus distinguishing it from business management systems that deal with nonspatial data.

A GIS comprises hardware, software, people, and organization, in addition to geospatial data. GIS hardware includes computers for data processing, data storage, and input/output; printers and plotters for hard-copy maps; digitizers and scanners for digitization of spatial data; and GPS and mobile devices for fieldwork. GIS software, either commercial or open source, includes programs and applications to be executed by a computer for data management, data analysis, data display, and other tasks. Additional applications may be used in GIS for specific data analyses. Common user interfaces to these programs and applications are menus, icons, and command lines, using an operating system of Windows or Linux. GIS users define the purpose and objectives for using GIS and interpret and present the results. GIS operations exist within an organizational environment; therefore, they must be integrated into the culture and decision-making processes of the organization for such matters as the role and value of GIS, GIS training, data collection and dissemination, and data standards.

History of GIS

The origins of GIS in its present form lie in the application of rapidly developing computing tools, especially computer graphics in a variety of fields such as urban planning, land management, and geocoding in the 1960s and 1970s. The flourishing of GIS activities in the 1980s was in large part prompted by the introduction of personal computers, such as the IBM PC, and graphical user interfaces, such as Microsoft Windows. Unlike mainframes and minicomputers, PCs equipped with a graphical user interface were more user friendly, thus broadening the range of GIS applications and bringing GIS into mainstream use in the 1990s. Also in the 1980s, commercial and free GIS packages appeared in the market. Environmental Systems Research Institute, Inc. (Esri) released ARC/INFO, which combined spatial features of points, lines, and polygons with a database management system (DBMS) for linking attributes to these features. Partnered with Intergraph, Bentley Systems developed Microstation, a CAD software product. Other GIS packages developed during the 1980s include GRASS, MapInfo, TransCAD, and Smallworld.
By definition, geospatial data cover the location of spatial features. To locate spatial features on the Earth’s surface, either a geographic or a projected coordinate system can be used. A geographic coordinate system is defined by its datum (e.g., NAD83, WGS84) and expressed in longitude and latitude, whereas a projected coordinate system is defined by its parameters (e.g., standard parallel, standard meridian, scale factor) and expressed in $x,y$ coordinates. Many projected coordinate systems are available for use in GIS. An example is the universal transverse mercator (UTM) grid system, which divides the Earth’s surface between 84°N and 80°S into 60 zones. A basic principle in GIS is that map layers representing different geospatial data must align spatially; in other words, they are based on the coordinate system. A GIS has tools for projecting or re-projecting coordinate systems.

Based on the representation of spatial features, geospatial data can be grouped into vector and raster data. The vector data model uses the geometric objects of point, line, and polygon to represent spatial features with a clear spatial location and boundary. Each feature is assigned an ID so that it can be associated with its attributes. The raster data model, on the other hand, uses a grid and grid cells to represent spatial features so that points are represented by single cells, lines by sequences of neighboring cells, and polygons by collections of contiguous cells. The cell value corresponds to the attribute of the spatial feature at the cell location. Although the raster representation of spatial features is not precise, it has the distinctive advantage in having fixed cell locations. In computing algorithms, a raster can be treated as a matrix with rows and columns, and its cell values can be stored into a 2-D array. Because of their simple structure, raster data can be easily manipulated, aggregated, and analyzed. In contrast, manipulation and analysis of vector data must consider the geometry of their locations and boundaries, and, when they are overlaid, the intersections of their boundaries must be calculated.

Generally speaking, vector data are ideal for discrete features such as streams, land parcels, and vegetation stands while raster data are ideal for continuous features such as elevation and precipitation. A GIS allows vector layers to be superimposed on raster layers as long as they are based on the same coordinate system. A GIS also has the capability of converting vector to raster data (rasterization), and raster to vector data (vectorization), at a loss of some data accuracy (e.g., smooth lines changed to zigzag lines after rasterization).

**Elements of GIS**

GIS operations usually involve data acquisition, data management, data query, vector data analysis, raster data analysis, and data display. These operations do not have to be sequential; for example, data display may be conducted for exploring data at the beginning of a project and for presenting results at the end of the project.

**Data acquisition**

Data acquisition is usually the first step in conducting a GIS project. The need for geospatial data by GIS users has been linked to the development of data clearinghouses and geoportals. Since the early 1990s, government agencies at different levels in the United States as well as many other countries have set up websites for sharing public data and for directing users to various data sources. As an example, Data.gov is a US government geoportal that allows public access to datasets generated by
GEOGRAPHIC INFORMATION SYSTEM

the Executive Branch of the US Federal Government. As of December 2016, the website shows more than 132,000 geospatial datasets and provides links to various agencies. Another example is the INSPIRE geoportal, which provides the means to search for spatial datasets and services, and to view spatial datasets from the member states of the European Union, including roads, populated places, land cover/use, administrative boundaries, elevation data, and ocean floor. Nongovernmental websites also exist for the distribution of geospatial data. For example, DIVA-GIS offers country-level data and data on global climate, species occurrence, crop collection, elevation, and satellite images. Another example is OpenStreetMap, a nonprofit based in the United Kingdom, which offers free geographic data of global coverage, such as street maps, points of interest, and land use, collected through crowdsourcing.

To use public data, it is important to obtain metadata, which provide information about the data. The International Organization of Standards (ISO) has implemented ISO 19115-1:2014, which defines the standards of metadata for describing geographic information and services. According to the standards, metadata should provide information on the identification, the extent, the quality, the spatial and temporal aspects, the content, the spatial reference, the portrayal, distribution, and other properties of digital geographic data and services.

If public data are not available, new geospatial data can be digitized from paper maps or created from satellite images, GPS data, survey data, street addresses, and text files with x and y coordinates. Geospatial data can be digitized using manual digitizing, scanning, or on-screen digitizing. Satellite images can be interpreted and classified to create land-cover and other maps. GPS and survey data can be converted to point, line, or area features. Likewise, street addresses and text files with x and y coordinates can be converted to point features through geocoding, a process similar to online geocoding services provided by Google, Yahoo, MapQuest, and others.

Data management

A GIS usually employs a DBMS to handle attributes of vector data, which can be large in size. Each polygon in a soil map, for example, can be associated with dozens of attributes on the physical and chemical soil properties and soil interpretations. Data management is much less important for raster data because cell values are bound to fixed cell locations. Attributes of vector data are maintained in different tables in a DBMS. An important type of table is the feature attribute table, which represents spatial features by row and attributes by column; therefore, the intersection of a column and a row lists a particular attribute value of a particular feature.

A DBMS offers two methods for linking attribute tables, such as linking a feature attribute table to an attribute table. A join operation brings together two tables by using a common attribute field (e.g., feature ID). A relate operation connects two tables but keeps the tables physically separate. Thus, using a DBMS, attributes can be linked to spatial features and to each other.

Data query

Data query allows users to select a data subset for close examination. Because the spatial (location) and attribute components of geospatial data are synchronized, data query can be approached from either component. Attribute data query typically uses a query language that is based on SQL (structured query language) developed by IBM in the 1970s. The basic syntax of SQL has three keywords: the select keyword selects
attribute(s) from the table, the _from_ keyword selects table(s) from the database, and the _where_ keyword specifies the condition or criterion for data query. Attribute data query, which can work with vector and raster data in a GIS, is basically the same as database query using a DBMS.

In contrast, spatial data query is unique in GIS because it allows users to select features based on their spatial relationships rather than their attribute values. Spatial data query is particularly useful for vector data because spatial features in vector data are represented with clear locations and boundaries. Three basic types of spatial relationships relate _features for selection_ and _features to be selected_. Containment selects features that fall within features for selection, such as finding schools within a county; intersect selects features that intersect features for selection, such as finding land parcels that intersect a proposed pipeline; and proximity selects features that are within a specified distance of features for selection, such as finding pet shops within one mile of Main Street. Variations of these basic relationships occur. For example, containment can be further classified by whether features to be selected are entirely, or partially, within features for selection.

It is common in GIS that a query combines both attribute and spatial data. For example, “finding schools within a county that has over 500 students” involves spatial data in the first part and attributes data in the second part. In such cases, the start can be with either spatial or attribute data query and the same result obtained.

Spatial join is an extension of spatial data query. Like join in a DBMS, spatial join also joins attribute data from two tables; but, instead of using attributes in the operation, spatial join uses the spatial relationship between features. For example, a spatial join operation can join attribute data of schools to those of counties by first matching schools and counties using the containment relationship.

Figure 1 Buffering around line features (black lines). The buffer zone is shown in gray.

**Vector data analysis**

Two basic tools for vector data analysis are buffering and overlay (Chang 2014). Based on the spatial concept of proximity, buffering creates two areas: one area that is within a specified distance of select features and the other area that is beyond (Figure 1). The specified distance is the buffer distance, and the area within the buffer distance is the buffer zone. Spatial features to be buffered can be points, lines, or polygons. Other options for buffering include creating multiple buffer zones around a feature and dissolving boundaries of buffer zones around features.

Overlay combines the geometries and attributes of two input layers to create the output (Figure 2). Overlay can be point-in-polygon, line-in-polygon, or polygon-on-polygon. In this classification, the first layer contains point, line, or polygon features and the second layer contains polygon features. In a point-in-polygon operation, the same point features are included in the output but each point is assigned with additional attributes of the underlying polygon.
In a line-in-polygon operation, the line features are dissected by the polygon boundaries and each line segment on the output is assigned with additional attributes of the underlying polygon. In a polygon-on-polygon operation, the output combines the polygon boundaries from both layers to create a new set of polygons, each carrying attributes from both layers. Depending on the area extent of the output, an overlay operation can also be union, intersect, or identity (Figure 3). Union preserves all features from both layers; intersect preserves only those features falling within the area common to both layers; and identity preserves only those features falling within the area extent of the input layer.

Besides buffering and overlay, there are tools for managing map layers or spatial features. Using these tools, two or more adjacent layers can be appended to create a single layer, split a single layer into two or more layers, and clip a layer by following the area extent of another layer. It is possible to erase from a layer those features that fall within another layer, and update a layer by replacing features with those in another layer. It is also possible to eliminate or select features in a layer based on an attribute data query, and aggregate adjacent features in a layer if these features have the same attribute value or values.

Two special types of vector data analysis are network analysis and point pattern analysis. Network analysis is based on a system of line features that is properly connected and has the appropriate attributes for measuring flow. The most common type of network analysis is shortest path analysis, which finds the path with the minimum cumulative distance or cost between nodes on a road network. Shortest path analysis can be extended to solve problems of vehicle routing, finding the closest facility (e.g., fire station), analyzing the spatial distribution of public resources (e.g., hospitals), and matching the supply and demand of facilities (e.g., siting of supermarkets). Given a distribution of point features, point pattern analysis can determine if the distribution is random, dispersed, or clustered at the global level or if the distribution contains...
local clusters of high or low values. Point pattern analysis has proved to be useful for analyzing crime locations (LeBeau and Leitner 2011) and public health data (Jacquez and Greiling 2003).

**Raster data analysis**

Four basic tools for raster data analysis are local, neighborhood, zonal, and global operations depending on if the analysis is performed at the level of individual cells, or groups of cells, or cells within an entire raster. A local operation can create a new raster from either a single input raster or multiple input rasters on a cell-by-cell basis (Figure 4). The cell values of the new raster are computed by a mathematical function (arithmetic, logarithmic, trigonometric, or power) relating the input to the output. A neighborhood operation involves a focal cell and its neighborhood (e.g., a $3 \times 3$ area centered at the focal cell) in a computation and then assigns the computed value to the focal cell. A zonal operation involves an input raster and a zonal raster, and produces an output raster, which summarizes the cell values in the input raster for each group of cells with the same values or like features in the zonal raster. A physical distance measure operation is an example of a global operation in which straight-line distances are calculated from cells designated as the source cells on a raster to all other cells within the raster. These four basic tools can be combined in analysis using *map algebra*, an informal language with syntax similar to that of algebra (Tomlin 1990). In addition to these data analysis tools, there are also tools for raster data management, such as clipping or mosaicking raster layers and extracting raster data based on an attribute or spatial data query.

Several special types of data analyses are based on raster data to take advantage of the ease in data manipulation and computation. Terrain mapping and analysis use a digital elevation model, a regular array of elevation points. A GIS has tools for mapping the terrain in contours, profiles, hill shading, and 3-D views, and for analyzing the terrain with slope, aspect, and surface curvature (Wilson and Gallant 2000). Terrain analysis also includes viewshed and watershed: a viewshed analysis determines areas visible from one or more observation points, and a watershed analysis traces water-flow to delineate streams and watersheds. Spatial interpolation uses points with known values to estimate values at other points and, in the process, create continuous surface data from the sample points. For example, spatial interpolation can be used to create a precipitation raster with known precipitation...
readings at weather stations. Least-cost path analysis finds the least accumulative cost path in a raster based on costs (real or relative) and cost distance measures. It is useful for planning roads, pipelines, transmission lines, and trails. For example, least-cost path analysis has been used to locate footpaths in mountainous areas (Rees 2004).

**Data display**

Maps are an interface to GIS (Kraak and Ormel-ing 1996). Map-making can be informal or formal in GIS. It is informal when viewing and querying geospatial data on maps, and formal when producing maps for professional presentations and reports. Here the focus is on the latter case. A professional map combines the title, map body, legend, scale bar, north arrow, and other elements to convey geographic information to the map reader. These elements should be arranged so that the map is balanced, coherent, ordered, and interesting to look at. The focus of the arrangement is usually the map body, which uses map symbols to depict the mapped data. For example, graduated circles may be used to display the distribution of different-sized cities in a state. Besides the graduated circle map, other types of quantitative maps include the dot map, choropleth (area-based) map, chart map, flow map, and isarithmic (isoline) map; these map types employ different visual variables of color (hue, value, chroma), shape, size, pattern, and texture to communicate quantitative information to the map-reader (Slocum et al. 2008).

After a map is composed in a GIS, it can be printed or saved as a graphic file for presentation. A map prepared in a GIS can also be converted to a KML file, imported into Google Earth or Google Maps, and shared publicly on a web server. For time-dependent data, such as population changes by county in the United States from 1900 to 2010, a series of map frames can be prepared and displayed in temporal animation (Cinnamon et al. 2009).

**GIS applications**

Since its inception, GIS has been important in resource management, land-use planning, natural hazards assessment, wildlife habitat analysis, riparian zone monitoring, and timber management. GIS has also been used for emergency planning, crime analysis, public health, land records management, transportation applications, precision farming, and military operations. Four specific examples of GIS applications are discussed here.

**Riparian management**

Riparian buffers are strips of land along stream banks that can filter polluted runoff and provide a transition zone between water and human land use. Riparian buffers are typically defined using the buffering tool for vector data. Zimmerman, Vondracek, and Westra (2003) incorporated riparian buffers of 30 m and 100 m in their simulation study of agricultural land-use effects on suspended sediment concentrations in two watersheds. They reported decreases in sediment concentrations with the presence of riparian buffers, especially the 100 m buffer.

**Site selection**

Site selection of a ski resort, a supermarket, or a landfill must consider a large number of factors. It presents an ideal scenario for applying map overlay, a tool that can combine the locations and attributes from two or more layers in either vector or raster format. In their study of emergency
evacuation shelters, Kar and Hodgson (2008) considered the following nine factors: flood zone; proximity to highways and evacuation routes, hazard sites, and health-care facilities; and total population, total children, total elders, total minority, and total low-income in the neighborhood. They chose a raster-based model because it was more efficient than a vector-based model.

Response time to fires

Many countries have established guidelines for responding to disasters and emergencies such as fires. The response time to fires is the time it takes for the arrival of an engine company at a fire scene. Shortest path analysis is the tool for calculating the response time. A study by Murray and Tong (2009) in Massachusetts (USA) reported that 25% of fires exceeded the response standard of four-minute travel time and, to be able to respond within four minutes to at least 90% of the fires, it would need 180 additional fire stations in Massachusetts.

Visual impact of “greenhouse parks”

A “greenhouse park” refers to the clustering of large-scale greenhouses on a single site. Although this clustering can reduce production costs by sharing infrastructure, such as energy, water, and gas facilities, it can impact the aesthetics of the surrounding landscape. Rogge, Nevens, and Gulinck (2008) applied viewshed analysis to a study of the visual impact of a greenhouse park in Belgium and found that it was visible in 39% of the area within a 1200-meter radius of the greenhouse park. They then examined different scenarios in which viewsheds could be reduced through the planting of trees and hedgerows.

Integration of desktop GIS, web GIS, and mobile technology

The introduction of PCs brought GIS into mainstream use in the 1990s. But the dominance of PCs has been downgraded in recent years and now PCs must coexist, and even compete, with web and mobile technologies. Web GIS or web-mapping services lets users find data (e.g., administrative boundaries, DEMs), access services (e.g., geocoding, routing), and publish maps to share with others. Web mapping received a boost after Google Maps and Google Earth appeared in 2005, followed by “Google Maps mash-ups” in 2006. The idea of mash-ups (i.e., combining a user’s own contents such as text, photos, and videos with web-based maps) quickly found its applications in real estate, vacation rentals, quasi-taxi services, and many other location-based services. Notably, maps are the foundation of these services.

Mobile devices such as smartphones and tablets with GPS can be used to perform many GIS-related tasks. For example, it is possible to download maps from a web server, collect real-time data in the field, edit the maps, and save the maps through a cloud storage service. More sophisticated operations can also be carried out by using apps that have been developed for smartphones. These apps let users conduct attribute and spatial queries; use custom templates for collecting and editing data in the field; and geocode and upload photos, video, and text in real time to map layers on the web.

Web technology’s inherent characteristics of communication and data sharing have helped the recent development of collaborative web mapping and volunteer geographic information (VGI). A good example of such development is OpenStreetMap. Often described as the Free Wiki World Map providing free geographic data such as street maps to anyone, OpenStreetMap
GEOGRAPHIC INFORMATION SYSTEM

is a collaborative project among registered users who voluntarily collect data using GPS, aerial photographs, and other free sources. As of March 2016, OpenStreetMap claimed over one million mappers around the globe. Although VGI provides a new and important data source to GIS users, the main concern about VGI is its accuracy and reliability in cases where data other than images are needed (Haklay 2010; Sui and Goodchild 2011).

From all indications, it is quite clear that desktop GIS, web GIS, and mobile technology will continue to coexist. A number of desktop GIS functions, such as data entry, data editing, map-making, address matching, attribute query, and spatial query, overlap with those of web GIS and mobile devices. Given this condition, GIS users should try to create an integrated, complementary GIS environment in which desktop GIS can play the role of performing “heavy-duty” tasks such as projection, data management, data exploration, vector data analysis, raster data analysis, and GIS modeling.

SEE ALSO: Geographic information science; Geoportals; GIS: history; Map projections and coordinate systems; Spatial database; Spatial social networks; Volunteered geographic information; Web-mapping services

References


The geographical imagination affords ways of thinking about space and place, whether conscious or unconscious, emphasizing how power shapes practices, behaviors, and social structures. Scholars attach a range of related definitions to the term, most of which can be traced back to the work of the geographers Hugh Prince, David Harvey, and Derek Gregory. The geographical imagination is also often confused or used interchangeably with other key concepts, such as “imaginative geographies,” “geographic imaginary,” and “geopolitical imagination.” Gregory recently addressed the theoretical contributions of the geographical imagination in *The Dictionary of Human Geography* (2009). Theoretical and applied uses of the term address how power and knowledge are deployed in and on space on behalf of social justice. The following threads emerge in this work: perception, cognition, and behavior; processes in the urban and in nature; tensions between the arts and sciences; trends in communications and mapping; elements of identity ranging from the individual to the structural; and imagined futures, especially as they relate to policy.

Educational theorist Maxine Greene describes the imagination as that which “makes empathy possible”:

To call for imaginative capacity is to work for the ability to look at things as if they could be otherwise. To ask for intensified realization is to see that each person’s reality must be understood to be interpreted experience – and that the mode of interpretation depends on his or her situation and location in the world. (2000, 3)

While geography appears to focus primarily on the material, the imagination then opens up questions of the abstract, creative, and possible. The geographical imagination affords the user ways to pry open the power in assumptions, stereotypes, and expectations associated with space and place, and to delve into how and why they are linked.

The term “geographical imagination” was first coined by Hugh Prince in a 1962 article of the same name in J.B. Jackson’s journal *Landscape*, which led to much of the ensuing confusion around the term’s meaning, Prince relayed the contradictions inherent within the humanistic field of geography where, at the time, a focus on spatial sciences took precedence. Responding to Carl Sauer who described geography as “a finely representational art,” Prince designated the “real problem of geography” as how to “combine the subjective view … with explanation in which the subjective view has no place” (1962, 25). Only by addressing the imagination through the project of history-making could the subjective and objective be bridged, particularly through works of art and literature.

David Harvey theorized the geographical imagination with a focus on space and place rather than landscape, nature, and the aesthetic. His approach built on C. Wright Mills’s “sociological imagination,” a conceptual tool for use by individuals to compare their personal biographies to larger social structures within their specific historical era. Harvey (1973) conceived of the geographical imagination also to
consider the role of space, place, and the political economy. Developed primarily from studies of the urban political economies, Harvey’s reading laid out the geographical imagination as a “spatial consciousness” that recognized the spaces between situations, scales, and events and their effect on relationships with the individual. He demonstrated the profound effect of space by breaking down the organic, perceptual, and symbolic spaces of everyday life.

Derek Gregory responded to the exoticization of “world-as-exhibition” that continued to permeate the era of the spatial turn, or the academy’s general cross-disciplinary interests in space and place, by elaborating on Harvey’s definition. Operating at the intersection of power–knowledge–space that allows questions of representation to become a focus of inquiry, Gregory defined the geographical imagination, in the words of Cosgrove and della Dora (2005, 388) as “the complex of culturally and historically situated geographical knowledge and understanding that characterizes a certain social group.” In Geographical Imaginations (1994), Gregory describes a feeling beyond our own world that allows us to transcend our particularities and connect to universalisms in order to realize our intertwined existences. His work pluralizes the concept, showing its situatedness and emergence as a distinctive tradition of Western intellectual inquiry. Gregory gives extra attention to the geographical imagination’s role in reproducing colonial norms in map-making, art, travel writings, and scholarly writings in philosophy, mathematics, and sciences.

The imagination’s place in geography is wide-ranging, and multiple terms and related concepts inform the geographical imagination. Studies of perception and cognition as they relate to space are essential to the geographical imagination because, as Yi-Fu Tuan wrote, they deal with the common human psychological experience of imagining. Distinct from the related term “geographical imaginary” – which denotes a more unconscious construction per Jacques Lacan – the geographical imagination can be manifest in the consciousness and therefore includes a reflective construction. The geographical imagination is of key import to the field of environmental psychology, which examines how people relate to and define their space and place, and vice versa. Environmental psychologists define place identity as an accumulation of learned cognitions related to space and developed through selective emotional attachments and meaningful experiences. These connections attach to particular places past, present, and anticipated, through memory and interpretation as well as fantasies and imagination. Gregory (1994) also describes how self-identifications on spaces and places afford processes of fabrication. As a compounded life experience, patterns of stereotyping and expectation are forged in relation to emotional geographies, material and imagined, as well as behaviors and perceptions of access, belonging, and safety.

Beyond concepts of landscape, the geographical imagination plays a key role in the processes that define the production of space, from its role in the production of nature to urbanization. The 1990s shift in studies of political ecology with a keen focus on the “production of nature” embraced a dialectical paradigm that was both social and ecological, material and conceptual. The production of nature describes the development of people’s relation to material nature through a historical context. The assumption of “man’s” dominion over “nature” has required major interventions in popular media and academic writing, such as through studies of political ecology, that is, how our environment is produced through political, economic, and social processes at a variety of scales. Today, increased discussions around climate change and
the Anthropocene indicate yet another shift in the geographical imagination around nature.

In the urban sphere, radically different schools of thought have addressed the geographical imagination, from the Chicago School of the 1930s to recent work on urban assemblages, mobilities, and network geographies. In 1973 David Harvey wrote that the “only adequate conceptual framework for understanding the city is one which encompasses and builds upon both the sociological and geographical imaginations.” The geographical imagination is advantageously deployed by critical urban and Marxist geographers who examine urban processes as they relate to capital. The geographical imagination can serve as an umbrella for ways space is rendered, including the social production of space, jumping scale, time–space compression, and time–space expansion. All of these terms afford new understandings of the city and nature and ways to imagine it.

The geographical imagination speaks to the interdisciplinary nature of geography as a humanistic social science, ranging from its role in the arts to cutting-edge technologies and new media. Researchers suggest that the geographical imagination both fuels and biases scientific research as well, ranging from the energy discourse to biogenetics. Advancements in technology relate heavily to the geographical imagination, including the geographies of traditional media (film, TV, radio) and new media (gaming, social media). Relatedly, the heavy influence code plays on our everyday lives requires a different geographical imagination to recognize that space and code are mutually produced, or what Rob Kitchin and Martin Dodge (2011) call “code/space.” The widespread use of GPS (Global Positioning System) on mobile devices has radically reduced spatial cognition and awareness as individuals become increasingly dependent on a screen map rather than remembering their paths.

The geographical imagination is often taken up as a visual practice of representing space and place. Traditional cartography and spatial images strongly affect our abilities to create and sustain spaces and places; nowhere is this more evident than with geographic information systems (GIS). Other maplike visualizations – such as photos, memories, paintings, and images – shape the mind’s spatial eye as well. Processes of map-making require that both the cartographer and map viewer possess a geographical imagination while also producing that geographical imagination of a place: the popular and erroneous Mercator projection map renders the United States larger than the continent of Africa. As Kevin Lynch’s (1960) work demonstrated, humans possess an “image of the city” that informs their experience and expectation of a space, the image being urban or otherwise. Studies around cognitive and mental mapping reveal how situated and specific the geographical imagination is to a person, and also how it is related to a person’s race, class, gender, sexuality, and sense of embodiment and privilege (Gieseking 2013). Consequently, feminist geographers have called for a re-envisioning of GIS and feminist visualization by disturbing the “God’s-eye view” that we assume to be the “proper” way to see maps and instead working from our own situated knowledges.

The geographical imagination is prominently used in regard to nationalist discourses, and profoundly helpful in critiques of colonialism and imperialism. Edward Said’s (2000) imaginative geographies demonstrate how the geographical imagination of citizens’ minds can be manipulated and exploited to portray a fashioned social political history of the state. Benedict Anderson’s (1983) imagined communities describe how the
media of a nation can create a shared social identity. Frederic Jameson (1990) suggests that the stories of the colonized reside in narratives of population rather than in the conqueror’s version of history. Similarly, the connection with the explorative and imperial means that the geographical imagination is often linked to travel writing and travelogues. Later work on philosophers of the Enlightenment, namely Hegel and Kant, reveals how present-day ideas of race and racism developed from these and other scholars’ reading of erroneous and, more often, fabricated travel accounts of the period. The seemingly self-explanatory “territorial” narratives of popular nationalist geographies override those who came before, erasing indigenity. Geographies of war, such as World War II news cartography and school maps, heavily influence and limit the geopolitical narrative of a nation and its people – what Cosgrove and della Dora (2005) refer to as the geopolitical imagination.

Scholars also examine how global processes are produced within and by more intimate scales through issues of sexism, racism, and heteronormativity. Scholarship in feminist geography is particularly essential in the work of reframing and refuting patriarchal, colonial, and heteronormative assumptions often prevalent in the geographical imagination. Gillian Rose’s (1993) use of the geographical imaginary to point out how a psychoanalytic approach to geography can reveal the masculinism inherent in geographical thinking, such as assumptions of territory as a norm. Tapping into the psyche, Gillian Rose suggests, allows for the freeing of one’s self from patriarchal geographical models by the imagination. In her book on black women’s geographies, Katherine McKittrick critiques the oft used term margins to describe all black women’s geographies because the margin is “emptied out, placeless, just theory, just language” (2006, 57).

Extending work that upends patriarchal and top-down approaches by including bodies and homes, Geraldine Pratt and Victoria Rosner (2012) call for a rejection of the false binary of the global and local, replacing it instead with the global and the intimate which are kept in tension but are actually coproducive. Feminist theoretical conceptions of space that seek to encompass and break up hard lines in binary framings of identity queer the geographical imagination by redefining what is “normal,” probable, and possible. Geographers of sexualities and queer theorists rely heavily on the geographical imagination as well. For example, Kath Weston (1995) writes of the sexual imaginary of LGBTQ people that urban neighborhoods would afford safety and solace as opposed to the seemingly heteronormative settings of rural and suburban environments. By recognizing individual, structural, and territorial geographical imaginations, as well as the intersectionality of the subjectivities that embody them, scholars and individuals alike are able to unpack the production of meaning that is assumed to define spaces and lives.

Since its core conceptualizations, the geographical imagination has broadened into a tool to describe and analyze the power within the literal and metaphorical ways people imagine and render space. The concept plays a central role in envisioning and enacting possible futures that are more just. Each person’s constant production of their geographical imaginations yields fissures which can be broken open to produce social and spatial justice. For example, the legal scholar Elizabeth Brown (2010) describes how many laws and governmental policies require attainable geographical imaginations that fit the political, economic, and cultural contexts to which they are applied. The geographical imagination is therefore a central concept of critical geography, geography that aims to develop theory, methodologies, and research that combat social
exploitation and oppression. Relatedly, Gregory (2009) writes that spaces of hope are alive in the practice of creativity. As Stephen Daniels announced to the 2010 Royal Geographic Society Annual Conference, the geographical imagination encompasses “the condition of both the known world and the horizons of possible worlds” (2011, 183), or what Greene (2000) earlier referred to as “worlds not yet imagined.”

SEE ALSO: Cognition and spatial behavior; Feminist geography; Geography and the study of human−environment relations; Human geography; Humanistic geography; Imaginative geographies; Intersectionality; Landscape; Sexualities; Spatial thinking, cognition, and learning; Visualization

References


Geographies of death

Avril Maddrell
University of the West of England, UK

Mapping death has long been an important preoccupation of epidemiologists and social scientists. Think of John Snow’s 1854 map of incidences of cholera in Soho, London, which led to the identification of the local water pump as the source of the water-borne disease. Durkheim’s study of the social factors affecting suicide might also come to mind. Thus geography and death are intimately related; however, within geography, analysis of that relationship has been largely limited to demographers’ concerns with life expectancy and mortality, for example, Dudley Stamp’s (1964) medical geography The Geography of Life and Death. More recently, others have used demographic maps, based on quantitative data, to show often powerful spatial patterns of mortality and life expectancy in order to identify causal factors and to inform public debate, policy, and service provision (see Dorling 2010).

Grounded in feminist, emotional, and affective geographies, another geographical perspective on death that has emerged and gained currency within the discipline since the 1990s is that of social and cultural understandings of death and mourning (Kong 1999). Initially focusing on roadside memorials, cemeteries and columbaria, sites of tragedy, and the politics of memorialization, the concept of “deathscapes” introduced in this work has been deployed more discursively to explore a variety of spatialities including physical, corporeal-psychological, and virtual spaces, and non- or more-than-representational practices situated in particular landscapes and social-cultural settings (Maddrell and Sidaway 2010).

Drawing on and contributing to interdisciplinary death studies and other cognate fields, recent geographical work has been attentive to the emotional and affective, including other forms of loss such as home, dementia, and missing persons. Recent work deploys concepts of relationality, the absence-presence of continuing bonds, and performativity to interrogate the embodied spatialities of death, mourning, and remembrance, the poetics and politics of care and loss being intimately intertwined in the everyday, universal, but often transformative, experience of bereavement.

Geographical scholarship is also addressing the geopolitics of death in warfare (including genocide and civilian “collateral” deaths), judicial death sentences, and life on death row as well as the politics of death through gendered violence such as domestic violence, misnamed “honor” killings, sexually motivated attacks, and femicide on the US–Mexican border. Human discourses of entitlement to kill other animals for food and other products are a growing area of engagement within geographies of death. The environmental politics of sustainability and carbon-reducing lifestyles have also prompted shifts in attitudes to burial and cremation, influencing post-life consumption and disposal choices in the West, with the growing “natural” or woodland burial movements and the disposal and mourning rites in non-European cultures drawing attention from geographers. Social changes deriving from the religious needs of migrant communities have also prompted work on the political, social, and cultural geographies associated with particular
migrant communities, for example, the growing need for Muslim cemeteries in Western Europe and the campaign for open-air Hindu cremation in the United Kingdom.

Geographies of death is a revitalized field where much-needed work on situated class, gender, cultural and ethnic experiences of death, mourning, and remembrance is emerging, as is Butler-influenced work on grievability and killability, economies of mourning, and posthuman grief.

SEE ALSO: Religion

References


Further reading

Emerging in the mid-1990s, the term “geographies of exclusion” refers to the subtle and not so subtle signals that some social groups are not welcome in urban and rural spaces and thus experience sociospatial exclusion. This builds on ideas around socially constructed structural differences, and in particular the sense that some forms of visible difference from mainstream society can lead to experiences of not belonging in public spaces, or an inability to participate fully in social, community, economic, and cultural institutions.

The phrase “geographies of exclusion” was coined by David Sibley in his 1995 book of the title. Sibley argues that social exclusion within Western society develops from a fear of the other and manifests itself through the social codes that determine how places such as the shopping mall, the street, countryside, and leisure sites are to be used. For instance, he describes how groups of teenagers can be viewed as rowdy by shopping mall security guards and removed from the semipublic space of the mall, and how New Age travelers are depicted as a nuisance in the English countryside because they diverge from mainstream views of how rural spaces should be used and become the focus of social and political pressure to move on. Exclusion is achieved through negative interaction and nuanced forms of social pressure that build on structural exclusions along gender, race, ethnic, or class lines, among others (rather than through physical boundaries and borders); this suggests that an unwillingness or inability to fit into dominant ways of understanding social interaction can have profound repercussions for those who are excluded. David Sibley links these anxieties about different social groups to historic fears of moral pollution and discomfort with ambiguity, and notes that the ways in which such social exclusions are put into practice create patterns (or geographies) of exclusion.

These ideas have been both taken up and challenged by geographers. Parr, Philo, and Burns (2004), for instance, have argued for a need to think about degrees of exclusion and inclusion, and outline how the same group of people can at various times be both welcomed and pushed aside. The result is a more complex pattern of exclusion which makes it difficult to conceptualize groups of people as being excluded and of certain spaces as being entirely exclusionary. This is a point also taken up by Angus Cameron (2006) who warns that, in speaking of a binary of exclusion/inclusion, we may be indirectly affirming these divisions and perpetuating the social and spatial divisions challenged by this analysis. Research around these issues has moved toward documenting not only who is excluded but also who is doing the excluding.

SEE ALSO: Difference; Identity; Public space; Social constructionism

References


Overview of human–environment geography

Human–environment geography, comprised of forms of knowledge that integrate the in-depth analysis of both human–social conditions in their interactions with the environment and the dynamics of the biogeophysical world, is also referred to as environmental, nature–society, and environment and society geography (Castree, Demeritt, and Liverman 2009). By integrating across the discipline’s human–environment divide, these forms of knowledge are distinct relative to other subfields (Zimmerer 2007, 2010). Human–environment geography often combines in-depth empirical research with the synthesis of information through conceptual frameworks and models. One of the leaders of human–environment geography, B.L. Turner II, has referred to this defining characteristic as the “specialist-synthesis merger.” (Turner’s works, and those of others, establish the use of the dash in human–environment geography.) The realm of human–environment studies in geography incorporates a diverse range of actively evolving approaches and interests. Variety and depth of the new knowledge systems in this realm are expanding amid the intensifying interactions and complex new relations of human societies and environments.

The accelerated intensification and potentially novel forms of current human–environment interactions are indicative of pronounced change as well as a sign of the expanded importance of this subfield. These trends are resulting in both a further strengthening of the core nodes of topical and thematic interest and, also, the rise of new issues and approaches. The factors that influence human–environment geography’s current phase include neoliberal globalization, urbanization, global environmental changes (e.g., biodiversity loss, climate change, food security issues), industrial ecology (e.g., energy and mining), population dynamics including the size, movement, and gendering of demographic factors (e.g., migration), the politics of so-called environmental security, and the growth of environment–related citizenship and social movements.

Several of the human–environment approaches responding to current changes are termed blended studies and hybrid sciences. These knowledge systems are synthesizing information, analytics, and interpretation across multiple areas of human–environment study. Such “blended human–environment studies” are characteristic of state-of-the-art knowledge and are a response, in part, to the unprecedented social urgency and complexity of environmental issues being encountered in the Anthropocene.

The Anthropocene is a term under consideration by leading scientific institutions (such as the Geologic Society of London and the Geologic Society of America) to demarcate the new geologic epoch distinguishable through the magnitude of human impacts on the Earth’s environments. The term Anthropocene also evokes
GEOGRAPHY AND THE STUDY OF HUMAN–ENVIRONMENT RELATIONS

the graphic image of planetary environmental boundaries, in which it is estimated that three of the eight most important global environmental systems currently approximate or exceed the limits of planetary sustainability (biodiversity, climate, and the nitrogen cycle) as the result of human activities. At least a couple others, namely water and land use, are significantly nearing the planetary environmental limits. Regardless of the eventual nomenclatural decision – whether or not the term Anthropocene is officially designated to denote the current geologic epoch (i.e., to follow the Holocene epoch) – it is clear that the scope of human–environment interactions has gained importance.

As a result of the aforementioned factors, the existing elements of human–environment inquiry and understanding are varied and dynamic within contemporary geography. Taken together, these human–environment approaches are clustered into nine identifiable nodes. The contours of this expansive human–environment terrain in geography are shaped principally through research and scholarship. Other influences in the coalescence of these nine nodes include the analysis and implementation of “real-world” environmental governance at multiple scales, practical applications to resource management, and environmental activism and public awareness.

This surge of interest, ideas, debate, and practice is transforming the range of human–environment knowledge systems as they seek to respond to ongoing changes beyond geography per se. Major influences currently reshaping various parts of human–environment geography range from the pivotal reconfiguring of post-structural theory in the academy to the amassing impacts of anthropogenic environmental changes leading to major societal challenges and uncertainty. Tracing the outlines of these influences can begin with the significant expansion and diversification of the human–environment sciences and scholarship through the restructuring of the academy during the 1990s and 2000s to the present (Turner 2002; Zimmerer 2010). The academic “restructuring” in these accounts refers to the influential expansion and consolidation of powerful interdisciplinary fields, such as environmental studies, the environmental sciences (including such offshoots as sustainability science), and ecological approaches incorporating social analysis as well as concomitant changes in the structure and organization of research funding. These academic shifts are leading to both new blended approaches that have recently arisen (see below) and the nodes of specialization in human–environment geography.

The contours of current human–environment geography also have significant influences outside the boundaries of academe per se. These influences stem from the increased awareness of abrupt, socially disruptive environmental changes associated with biogeophysical systems (e.g., climate change, ocean acidification, water shortage, and storm intensity). “Biogeophysical” is an appropriate term that reflects a view of the environment made up of biogeographic, geosystem, and physical geographic factors. Powerful influences arise from the policy, politics, and social issues of human–social endeavors concerning the environment. Recent trends of human–environment geography highlight the need to situate the dynamic biogeophysical changes in interacting scales of space and time. Human–environment geography is focused on the biogeophysical elements of change from a perspective centered on the interactions and entanglements with modern human societies, political economies, environmentalism, and environmental movements. Such forces exert major direct influences on human–environment studies.
The dynamism of human–environment studies has contributed to its widespread recognition as vital in the contemporary discipline of geography. In general, two disciplinary structures predominate. First is the view of human–environment geography as a kind of overlapping connective tissue across the skeletal disciplinary structure of human geography and physical geography. This treatment of human–environment geography tends to rely on a framework in which nature serves as backdrop to the analysis of social, economic, and political dynamics (in the case of human geography) or as a kind of triggering release that produces impacts on the dynamics of environmental systems (in the case of physical geography).

In a second predominant mode of disciplinary configuration the treatment of human–environment geography – which is used synonymously with “nature–society geography” (see Zimmerer 2010) – is situated as a principal and discrete core of the four-field or five-field framework of the discipline. The four-field framework, for example, recognizes human–environment geography (or nature–society geography as it is sometimes called) in conjunction with physical geography, human geography, and GIScience, with the last mentioned encompassing cartography and visualization. A five-field approach to the discipline also implemented, spreads human geography into multiple subcategories that include economic geography and regional development.

While the abovementioned modes of disciplinary configurations are distinct, each recognizes and highlights the role of human–environment geography as crucial to the potential of far-reaching intradisciplinarity within geography. Extended also to radical intradisciplinarity, this trend recognizes the major role of human–environment interactions and nature–society relations within diverse undertakings in contemporary geography (Zimmerer 2007). Whether viewed as “borderlands” or “embeddedness” the intellectual space of human–environment studies enables synapses across many highly active areas of current geography (Table 1).

Overall, a trend toward intellectual diversification is characteristic of the current status of human–environment geography. This diversification is reflective of the general pattern of multi-strand branching and transformations in areas of thought in human–environment geography, albeit in the absence presently of a single or small number of dominant paradigms. By contrast, much of the history of human–environment geography was previously distinguished by the existence of preeminent paradigms. The delineations of cultural ecology and human ecology, which became popular in the 1960s and 1970s, have continued to evolve and diversify in the context of geography as well as other disciplines and interdisciplinary fields. “Cultural ecology” is now used to refer to human–environment studies with a significant component of cultural studies, as in the works of Doolittle, Head, Knapp, Mathewson, and others; human ecology is focused on a systems-based view of human–environment interactions. Works by Brush, Butzer, Moran, and others delineate the lineage and current usage of human ecology.

As noted below, the strengths of cultural ecology and human ecology subsequently fueled the rise of such major human–environment approaches as social-ecological, land, sustainability, and coupled natural–human system sciences as well as political ecology and environmental history. These approaches have expanded prodigiously since 1990. The approach of political ecology, also rooted in cultural ecology and human ecology, has particularly burgeoned. It includes a subgroup that is focused on
Table 1  Topical and thematic areas in relation to intra-disciplinarity within geography.

<table>
<thead>
<tr>
<th>Topical and thematic area of human–environment geography</th>
<th>Proximate intellectual relations within human–environment geography</th>
<th>Most important proximate intellectual relations within the discipline of geography (outside human–environment geography)</th>
<th>Generally important intellectual relations within the discipline of geography (outside human–environment geography)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environmental hazards, risk, vulnerability, and resilience</td>
<td>Social-ecological and coupled human–environment systems</td>
<td>Climatology; hydrology; fluvial geo-systems</td>
<td>Physical geography</td>
</tr>
<tr>
<td>Land use, land systems, land change, and biodiversity</td>
<td>Social-ecological and coupled human–environment systems; political ecology; livelihoods and agricultural landscapes</td>
<td>GIScience; remote sensing</td>
<td>Physical geography</td>
</tr>
<tr>
<td>Social-ecological and coupled human–environment systems</td>
<td>Land use, land systems, land change, and biodiversity; livelihoods and agricultural landscapes; political ecology</td>
<td>Highly varied</td>
<td>Physical geography</td>
</tr>
<tr>
<td>Political ecology and environmental governance</td>
<td>Resource political economy, management, and politics</td>
<td>Political economy; economic geography; social geography; cultural geography</td>
<td>Human geography</td>
</tr>
<tr>
<td>Livelihoods and agricultural landscapes</td>
<td>Political ecology; land use, land systems, land change, and biodiversity; environmental landscape history; political ecology</td>
<td>Cultural geography; economic geography</td>
<td>Human geography; physical geography</td>
</tr>
<tr>
<td>Resource political economy, management, and politics</td>
<td>Political ecology; landscape environmental history</td>
<td>Economic geography; social geography</td>
<td>Human geography</td>
</tr>
<tr>
<td>Food, health, and bodies in relation to the environment</td>
<td>Political ecology; livelihoods and agricultural landscapes</td>
<td>Social geography; health geography; economic geography</td>
<td>Human geography</td>
</tr>
<tr>
<td>Environmental landscape history and ideas</td>
<td>Political ecology; knowledge concepts</td>
<td>Historical geography; social geography</td>
<td>Human geography</td>
</tr>
<tr>
<td>Knowledge concepts in environmental management and policy</td>
<td>Political ecology; environmental landscape history; resource management</td>
<td>Social geography; economic geography; historical geography</td>
<td>Human geography</td>
</tr>
</tbody>
</table>
human–environment interactions. This inclusion grew from the founding text of political ecology, *Land Degradation and Society*, that was authored in 1987 by Piers Blaikie and Harold Brookfield. Subsequently, political ecology was advanced in the series of texts that were launched in the late 1990s and early 2000s by authors such as Bryant and Bailey, Forsyth, Hecht, Martínez-Alíer, McCarthy, Neumann, Robbins, Rocheleau and Thomas-Slayter, Schroeder, Sullivan and Stott, Wolford, and Zimmerer and Bassett among others. Each of these works includes an emphasis on human–environment relations.

Subsequent developments of human–environment geography (2005–2015, detailed below) have taken hold as a result of the influences of both fields outside geography as well as disciplinary particulars. Consequently, human–environment studies in geography must be viewed as closely related to state-of-the-art developments in other disciplines and interdisciplinary fields. For example,
the extensive interdisciplinary dimension of human–environment geography is key to the suggested reinvention of US geography as an “interdisciplinary discipline.” This suggestion revolves around the idea of strengthening the discipline-based integrity of geography while emphasizing its interdisciplinary connectivity. Leaving aside the specific pros and cons, as well as the general feasibility, this suggestion is important with regard to underscoring the highly interdisciplinary nature of human–environment studies.

Important parallels to the high level of interdisciplinarity of human–environment geography are also found in the related fields of ecology, economics, sociology, anthropology, archaeology, history, political science, and landscape architecture. Active authors of major works based in these other human–environment fields include Belsky, Besky, Bieling, Blesch, Brondizio, Brosius, Carey, Clark, Collins, Craib, Cronon, Curran, Dove, Durham, Escobar, Fischer, Fischer-Kowalski, Foley, Friedmann, Goldman and Schurman, Grau, Hinrichs, Ingold, Irwin, Kirch, Kloppenburg, Langston, Loos, J. Liu, McMichael, Matson, Mittman, Moran, Morrisson, Nadasdy, Nazarea, Nelson, Padoch, Peluso, Perz, Plieninger, Rhoades, T. Roberts, Rudel, A. Sachs, Saito, Scoones, Stedman, Tucker, and R. White. Some of the major interdisciplinary fields important to human–environment geography are those of ecology and society initiatives, environmental studies, environment science, development studies, food studies, global studies, urban studies and urbanization sciences, and programs focused on climate change, biodiversity conservation and conservation biology, human ecology, social ecology, political ecology, and environmental policy and management (Table 2).

Prominent instances of the successful interdisciplinary of human–environment geography involve a number of the approaches mentioned above – such as the ones focusing on sustainability, social-ecological systems, vulnerability and resilience, land systems, and coupled natural–human system sciences. Geographic authors active and influential in these interdisciplinary human–environment domains include Adger, Aspinall, Bassett, Bebbington, Brannstrom, D. Brown, K. Brown, Chowdhury, Coomes, Cutter, DeFries, Eakin, Evans, Haase, Hostert, Kasperson, Kates, Kreutzmann, Kümmerle, Liv- erman, Lambin, McSweeney, Mertz, Moseley, Munroe, Radel, Reenberg, Seto, Southworth, Tschakert, B.L. Turner II, M. Turner, Vadujenc, Walker, Walsh, Wilbanks, Young, and Zimmerer, among many others. In addition to journals in geography and other fields, these works appear in high-impact science journals such as Nature, Science, and the US Proceedings of the National Academy of Sciences (PNAS) as well as other interdisciplinary journals with central emphasis on human–environment interactions, such as Global Environmental Change, Annual Review of Environment and Resources, Society and Natural Resources, Regional Environmental Change, and Ecological Applications.

In conjunction with these previous developments new insights are now addressing the challenges and potential tradeoffs of pronounced interdisciplinarity as it relates to human–environment geography. One possible tradeoff is the attenuation of cohesiveness of human–environment geography as a center within the discipline. The discussion of such potential tradeoffs has become vital in geography as well as other disciplines, such as anthropology, where interdisciplinarity has become vital and integral to the discipline itself. Internationally the influence of interdisciplinarity on human–environment geography, along with the discipline of geography in general, can vary significantly in extent and degree. In the United
States, for example, where geography is a discipline of small to moderate size, interdisciplinary influences are prevalent in human–environment studies. The influences include the role of the National Research Council and its major reports on geography in 1997 and 2010. A lesser degree of interdisciplinarity tends to occur in Britain where the discipline of geography enjoys significantly greater size and strength, and can therefore draw to a greater degree on intradisciplinarity in human–environment studies. The relative extents of environmental interdisciplinarity and intradisciplinarity have also become varied in the diverse countries of Europe, Asia, Africa, and Latin America.

Continuities and abiding influences also remain important to the understanding of human–environment studies in the discipline of geography and other fields. This theme of continuities can be used as a guide to uncover influential legacies. Viewed historically, a pair of well-defined channels of human–environment endeavors had long shaped the characteristic landscape of geography, especially in the US context of the discipline (Zimmerer 2010). One brings together landscape approaches based upon a cultural-historical framework. The other consists of the studies of recent human and social interactions with environmental changes. It has tended to focus on floodplain and other forms of resource management along with other practical applications.

The former is often referred to as the Berkeley or Sauerian School while the latter is termed the Chicago School. The Sauerian School used ecological science in order to offer cultural and historical interpretations, and thus can be thought of as adopting a cultural-historical ecology. The Chicago School used ecology explicitly in models of human–ecological interaction. As a whole, ecology was both a cornerstone and a chimera in these human–environment approaches since its usage could range from being central, mechanistic, metaphorical, or, at the other extreme, almost entirely implicit and unacknowledged. Nonetheless, this pair of parallel channels defined much of the landscape of human–environment studies in geography, especially between the early twentieth century and the 1960s. The depth and continuity of these approaches can be traced to current major nodes, such as environmental landscape history and ideas (Sauerian School influence) and the focus areas of environmental hazards, risks, vulnerability and resilience (Chicago School influence).

This pair of predominant traditions, the Sauerian and Chicago Schools, functioned for decades as a defining intellectual landscape seemingly comprised of twin gorges incised ever more deeply in geography’s intellectual landscape, each with ample depth and continuity, into the 1980s. Subsequent recontouring occurred in a transition to multiple, diverse approaches to human–environment studies in geography that continue today. This transition evokes the image of a braided stream with multiple channels. Braided-stream topography is an apt metaphor for the combined distinctness and interconnectedness of multiple topical and thematic nodes within current human–environment geography. The braided-stream image also provides the connotations of directionality and crossing-over – anastomosing in stream geomorphology – that reflects the reality of several co-existing current trends, as well as the ample continuity that can be traced to the powerful precedence of earlier flows where upstream intellectual topography continues to exert major influences. The twin channels of the intellectual landscape serve as an important complement to another spatial metaphor of human–environment geography, namely the image of spirals, bridges, and tunnels put forth in the work of B.L. Turner II.
Topics and approaches in human–environment geography

Well-established topical and thematic nodes of research and understanding are characteristic of human–environment geography. These core areas tend to cluster around nine nodes: (i) human–environment interactions in hazards, risk, vulnerability, and resilience; (ii) land use, land systems, land change, and biodiversity; (iii) social–ecological and coupled human–environment systems; (iv) political ecology, environmental governance, and human–environment relations; (v) human–environment relations in livelihoods and agricultural landscapes; (vi) resource political economy, management, and politics; (vii) food, health, and bodies in relation to the environment; (viii) environmental landscape history and ideas; and (ix) knowledge concepts in environmental management and policy.

Each node is distinct yet related and often intersecting significantly, as described in the examples mentioned below and illustrated in Table 2. In addition, the core areas can be loosely grouped into a pair of general thematic areas within human–environment geography, namely “human–environment interactions” and “nature–society relations.” Each individual node of human–environment geography is differentially situated with respect to this pair of thematic areas, and tends to reflect distinct and differentiated intellectual locations with regard to methodological and conceptual domains in human geography, physical geography, and GIScience. As discussed below, for example, the general theme of human–environment interactions tends to ally most fully with environmental hazards, risk, vulnerability, and resilience and, also, to the area of land use, land systems, and land change. In general, it ties to physical and economic geography. On the other hand, the general theme of nature–society relations is associated most closely with political ecology, with defining connections to human geography.

Human–environment relations in hazards, risks, vulnerability, and resilience

One prominent core of human–environment geography is comprised of the studies of environmental hazards, risks, vulnerability, and resilience. Environmental hazards and risks are significant themes that have continued to evolve in the context of accelerated global biogeophysical and socioeconomic changes and their human–environment interactions manifest in response to such disruptive environmental events as drought and floods as well as price shocks and the collapse of institutions. Vulnerability and resilience have become increasingly widespread concepts incorporating human–environment interactions. For example, combined social and environmental dynamics of vulnerability have become a cornerstone recognized within human–environment geography as well as geography in general (Cutter 2003). Such works tend to establish a view of vulnerability, hazards, and risks as processes involving the behaviors and multilayered networks of social actors and institutions.

At the same time, environmental shocks, such as climate–driven effects, can lead to the utilization of adaptive capacities whereby the institutional capabilities mobilized in human–environment interactions can result in significant socioeconomic benefit, such as was the case among certain forest–dwelling smallholder land users in Central America in the wake of Hurricane Mitch in 1998 (McSweeney and Coomes 2011). Equally important, the social conditions and power dynamics underlying vulnerability lead to a prevalent use of political ecology (see critique in Mustafa 2005), with attention to
such issues as social “winners and losers” in these processes and spatial patterns. Additional prominent authors in this topical and thematic area of human–environment geography include Adger, Barnes, Birkenholtz, Cutter, Downing, Eakin, Krueger, Leichenko, Liverman, Montz and Tobin, Mertz, Mortimore, O’Brien, Polsky, Ribot, Smit, Tschakert, and B.L. Turner II. This area shows close connections to various interdisciplinary fields (Table 2).

**Land use, land systems, land change, and biodiversity**

This topical and thematic area is focused to date on human–environment dynamics involving the spatial and temporal properties of vegetative cover, principally forests and urban spaces. It is closely associated with land change science, which has been subject to influential definitions and conceptual framework-building in the interdisciplinary scientific literature by human–environment authors such as Aspinall, D. Brown, Chowdhury, Crews-Meyer, DeFries, Evans, Hostert, Kümmerle, Lambin, Manson, Moran, Munroe, Parker, Reenberg, Rindfuss, Rudel, Seto, B.L. Turner II, Verberg, and Walsh. Methodologically this focus area tends to make extensive use of the combination of remotely sensed imagery, surveys of land users, and census data. The observation of major shifts of vegetative cover is considered in the context of spatial land-use systems (sometimes referred to as “land systems”) and land-cover change. Significant attention and insight thus far has focused on the entwined spatial and socioeconomic processes of deforestation involving conversions to pastureland and agriculture, as well as the regrowth of forest through so-called secondary forest transitions. Linkages to biodiversity are sometimes inferred in these studies, while they hold considerable future promise with increased use of methods incorporating the human–environment, ecological, and taxonomic assessments of biodiversity.

The changes of vegetative cover areas are linked to models of micro- and macrolevel economic and political factors ranging from household labor availability to national and international economic policies (Coomes, Barham, and Takasaki 2004). Several important studies have been located in regions of the world’s major tropical forests (especially the tropical humid forests or “tropical rain forests”), such as the Amazon Basin of Brazil and neighboring countries (e.g., Bolivia, Peru, Ecuador, Colombia) and Mexico and Central America. Forest areas of other tropical regions in Africa and Asia as well as temperate-zone forests (e.g., Europe and the United States), have also been the subject of these studies. Global market integration and postsocialist transitions, to name only a couple, are commonly incorporated as factors on the human–social side. Influential authors that have offered significant advances to this topical and thematic area, in addition to those persons mentioned above, include Aide, Arrima, Caldas, Chowdhury, Evans, Hostert, Klepeis, Kümmerle, Kreutzmann, Millington, Müller, Munroe, Ramankutty, Redo, Southworth and R. Walker. Human decision-making, social movements, governance preferences and disputes, and the active alteration of land use (“agency”) exert feedbacks whereby these modifications and activities contribute to the subsequent reworking of human–social conditions (“structure”). Human–environment interactions are understood explicitly as bidirectional and central to this framing. Here the concept of so-called structuration is being used to understand the bidirectional linkages of land use. Authors of major works on this use of the structuration theme in human–environment
geography and the study of human–environment relations

The frameworks of social-ecological systems (SES) and coupled human–environment systems are conceptual cornerstones in interdisciplinary human–environment studies that connect a thriving interdisciplinary domain to a focus within geography. SES is focused on human–social systems and the management and governance of resources. It owes much to the contributions of Nobel Prize winner Elinor Ostrom and to the legacy of earlier works examining the “tragedy of the commons” not as a demographically and culturally driven fait accompli but rather through social institutional processes. The SES framework often focuses also on interactions across the realms of humans and nonhumans. One chief SES contribution in human–environment geography is the development of ideas and examples of resilience, vulnerability, and adaptive capacity. These are being applied, for example, to the analysis of biodiversity in agroecosystems undergoing intensification and livelihood diversification (Zimmerer 2013; see also Beymer-Farris, Bassett, and Bryceson 2012).

The framework of coupled human–environment systems, also termed coupled natural–human systems (CNHS), is built upon the defining idea of coupled drivers emanating from human–social factors and feedbacks from natural systems. The coupled system CNHS perspective was initially proposed by Liu and others in the early 2000s. In human–environment geography the CNHS framework has been applied principally to the use and management of water resources and forests. Major authors in the topical and thematic areas of SES and CNHS, who come to these frameworks from the perspective of human–environment geography, include Bury, Eakin, Evans, French, Leichenko, López-Carr, and O’Brien, B.L. Turner II, and Wrathall.

Political ecology and human–environment relations

Political ecology is a burgeoning field covering many areas that include human–environment relations and interactions. It is a significant albeit minor portion of political ecology that undertakes the integration of environmental and ecological analysis (Zimmerer 2015). At the same time, the major share of political ecology is focused on nature–society relations, especially social power relations in resource conflict and coordination, environmental representations, and the roles of multiple environmental knowledge systems.

One productive subset of political ecology centered on human–environment interactions is the analysis of global climate change in relation to policy initiatives and political issues associated with sustainability, globalization, and neoliberal management (Liverman 2004). This work combines a perspective on human–environment interactions and the biogeophysical dynamics of climate change in order to better understand multiscale governance amid dominant political economic regimes, in particular neoliberal globalization. Another noteworthy subset is concerned with the multiscale political and environmental dynamics of national and international governance along with the community- and user-based management of fisheries, marine organisms, and forest and range resources. Technological change in these resource systems is often fundamental to the issues of environmental governance and to the forms of social and political coordination and contestation that arise.
GEOGRAPHY AND THE STUDY OF HUMAN–ENVIRONMENT RELATIONS

Human–environment interactions involving fisheries, grazing, and forest resources are often explicitly territorial and thus well-suited to studies that incorporate the geographic themes of territory-making and territoriality.

Still another subset of political ecology studies is concerned with the governance of water resources though social power dynamics involving planning, climate change, agriculture and land use, and international relations. Here too spatial and territorial designs are often explicit in such initiatives as the increased privatization of water supplies occurring in neoliberal globalization. Urban political ecology is notable for its examination of the roles of social and political power of the human–environment interactions involving water resources. Gendered dimensions of social power are often influential in determining control and access to water resources. Not coincidentally, the approach of feminist political ecology (FPE) has been productive and influential in understanding the human–environment geography of water resources. It is also opening new vistas that include major connections to such resource systems as biodiversity management, forest resources, and energy systems. Major geographic authors in the topical and thematic area described thus far in this subsection – in addition to those mentioned in the general introduction – include Baka, Bakker, Bassett, Bell, Birkenholtz, C. Brown, Bridge, Campbell, Carr, Emel, L. Harris, Hecht, Heynen, Huber, Kaika, Mutersbaugh, Mansfield, McCarthy, Neumann, Nightingale, O’Reilly, Perreault, Rocheleau, St Martin, Shapiro, Sneddon, Sultana, Swynge-douw, M.D. Turner, Wolford, and others.

Biodiversity management and environmental conservation are also a focus of human–environment analysis in political ecology. This focus is often centered on environmental management in the design of protected areas (PAs) (see Environmental management). The spatial extent and number of protected areas have increased worldwide in significant ways since the widespread accounting and monitoring of these units began a few decades ago. Diversity of environmental management schemes is also much increased within protected areas. One reason for this additional complexity is that protected areas have become more commonly designated in categories associated with human use and activities (e.g., “buffer zones”), rather than being strict set-asides intended to eliminate or prevent the presence of humans. As a result of these trends, human–environment geography and the recognized role of these interactions are now increasingly important to the environmental management of protected areas globally.

Still another topical and thematic area is concerned with the mix of issues related to agriculture, food security, land tenure, policy-related land change, pesticide use and transgenics, and agrarian reform and policy institutions, including urban and peri-urban food systems and land use. This mix has become part of a significant emphasis on agri-food systems, both expanding global industrial agriculture that is based on biotechnology and that incorporates a growing corporate organic sector, as well as alternative and local systems that include efforts to conserve agricultural landscapes and biota. The topics and themes mentioned in this paragraph and that preceding are being productively investigated through the perspective of human–environment geography by such authors as W. Adams, Bassett, Braun, Bezner-Kerr, Brannstrom, Buck, Campbell, Carney, Freidberg, Galt, Guthman, Graddy, Jarosz, Jepson, B. King, Kosek, McAfee, Medley, Mosely, Naughton, Neumann, Roth, Sayre, Schroeder, Wainwright, Wolford, Young, and Zimmerer.

The topical and thematic area of political ecology, as synopsized here, is illustrative of
the cross-cutting among the nine nodes identified in this section. This particular subsection contains a major emphasis on environmental governance and its influence through the forces and dynamics of political economy. Environmental governance, as it is broadly defined by Agrawal, Bulkeley, Heynen, Lemos, Liverman, McCarthy, Perreault, Prudham, Timmons, O. Young, and others, refers to environment-related institutions, plans, knowledge, management, decision-making, and practices. Yet the focus of political ecology on environmental governance is widely shared and must be appreciated as an example of cross-cutting topical and thematic influences. This theme is addressed as a major focus in several of the other principal topical and thematic areas currently active within human–environment geography. For instance, social-ecological systems (SES), mentioned above, are also focused extensively on environmental governance.

Other topical and thematic nodes identified here also evidence the significant focus on environmental governance. These include environmental hazards, risks, vulnerability, and resilience; land-use systems, land change, and biodiversity; environmental landscape history and ideas; and scientific concepts in environmental management and policy. In sum, each of these areas addresses environmental governance in a distinct and important way that needs to be distinguished in current human–environment studies. Another cross-cutting focus is concerned with the human–environment concepts of adaptation, resilience, and vulnerability. While adopted most extensively within the area of social-ecological and coupled systems, the concepts of adaptation, resilience, and vulnerability are also part of an expanding focus in political ecology by such authors as Bassett and Fogelman (2013), Beymer-Farris, Bassett, and Bryceson (2012), Ribot (2011), M.D. Turner (2014), and Zimmerer (2013). Butzer, Cote, Eakin, Nightingale, M. Taylor, Tschakert, and Watts, among others, have also authored important contributions in this area.

Livelihoods and agricultural landscapes

This topical and thematic area has arisen partly as a result of global-scale shifts to part-time land use and the importance of diversified livelihoods among the 2.0–2.5 billion smallholders who continue to engage in food production amid major multiscale changes in socioeconomic and environmental conditions. Emphasis on agri-food landscapes and agri-food systems owes in part to the fact that these smallholders include many of the world’s most food-insecure populations. Another important focus involves the extensive empirical analysis and understanding of the increasing roles of urbanization, livelihood diversification, and development on land use. These drivers incur complex human–environment relations involving agriculture, food, and the cover of anthropogenic forests. Diversified livelihoods, including the phenomenon of vast peri-urban peripheries and the “new rurality” (defined as the predominance of livelihood diversification, urban connections, and consumptive economic values in the countryside) now impinge directly on large swaths of landscapes and involve the livelihoods of much of the world’s population.

Integration with product and labor markets under neoliberal globalization, including the formation of extensive informal sectors and urban/peri-urban areas, drives many of the human–environment dynamics of development, peri-urban places, and the impacts of the new rurality (Carney 2008; Zimmerer, Carney, and Vanek 2015). The world’s 2.0–2.5 billion smallholder land users are as equally an emphasis as large corporate industrial agriculture. At
the same time, the rise of alternative and local food networks is leading to potentially novel agri-food systems and landscapes in various kinds of places. These networks may exert a sustainability enhancing influence on agriculture and food landscapes in the management of such environmental factors as soil nutrients, food and woodland biodiversity, and farm woodlands. Such human–environment interactions are presently a major priority of new research and understandings being developed and debated in human–environment geography. Authors of major human–environment works on topics and themes covered thus far in this subsection include Baker, Bassett, Bebbington, Carney, Denevan, Doolittle, Galt, L. Gray, Hecht, Hedberg, Lerner, McMichael, Moseley, Price, Radel, Rangan, Rocheleau, Shillington, WinklerPrins, and Zimmerer.

Current human–environment research suggests that the widespread and growing use of migration remittances and other factors linked to part-time land use can exert complex and sometimes favorable influences on environments, forests, and food-growing landscapes. The complex pathways of migration can lead to either the disintensification or intensification of food-producing landscapes, along with either increase or decrease in measures of social equity and environmental quality. In other words, a priori assumptions of determinacy no longer govern the investigation and interpretation of these human–environment dynamics. Instead, research scholars, policymakers, and planners are seeking to understand the conditions under which certain relatively favorable outcomes occur with regard to both migration and the environment.

The array of diverse mechanisms producing these varied human–environment interactions, ranging from labor withdrawal to the intensification of food-producing landscapes, are found to depend on the migration-related factors of remittance and the gendering of decision-making that can impact the specific properties of forest cover and biodiversity. The biodiversity in these food-producing landscapes often incorporates the variation of the biota of food assemblages (“managed agrobiodiversity”) and agroecosystems encompassing below-ground and uncultivated elements (“associated agrobiodiversity”). The geographic range of these food-producing and consuming landscapes, which include gardens, interstitial spaces, and connections among rural, peri-urban, and urban locales, adds further to their complexity (Zimmerer, Carney, and Vanek 2015). Major geographic authors in this area—in addition to those mentioned above—include Coomes, Doolittle, Eakin, Graddy-Lovelace, C. Gray, L. Gray, Hecht, Knapp, Lerner, Momsen, Radel, Schmook, Torres, Vanek, and WinklerPrins.

Human–environment relations in resource political economies, management, and politics

The political economies and politics of many issues of resource governance and politics extend across national and global scales and frequently entail the role of combined urban–rural spaces. The resource systems and the institutions range from minerals and hydrocarbons involving major multinational corporations to urban, village-based, social justice, and industrial issues that are related to water, environmental quality, and waste. Areas of emphasis in human–environment geography include the social dynamics of resource extraction, land and resource grabs, and management as well as related resource and agrarian issues. These social dynamics often entail the role of active social movements with alternative visions of development and social justice. Studies have revealed the complex dynamics and often unanticipated outcomes
of certification programs that appear to foster environmental sustainability outcomes using such resource management as sustainable forestry certification and fair-trade coffee.

Social justice aims may sometimes be incorporated in these programs with such goals as supporting poorer neighborhoods or small-holder farmers. While in certain cases they are well-intentioned, such initiatives may, in fact, not function as hoped, for their desired environmental and social benefits may be either undermined or suffer unintended negative consequences. Contributors of major human–environment research and publications on the range of issues mentioned thus far in this section include Bebbington, Bridge, Bury, Bryant, Calvert, Castree, Emel, Hindery, Huber, Humphreys, Klooster, Labban, Le Billon, Liu, S. Moore, Mutersbaugh, Perreault, Pulido, Silva, Valdivia, Watts, and Wolford.

Trenchant critiques in this area have highlighted the powerful influence of innovative neoliberal policies in prevalent environmental management policies and approaches. Examples include the ecosystem service frameworks that are being applied to reduce deforestation and hopefully invigorate sustainable forest management, to mitigate and manage wetlands while compensating for loss resulting from real estate development, and to restore freshwater stream morphologies and ecological habitats. Market-based environmentalism has become a particularly powerful form of environmental governance under the dominance of neoliberal policies. The spatial design of this governance is one reason among several for the significant increase of human–environment study on urban locales and multiple research sites spanning the Global North and the Global South (Schroeder, St Martin, and Albert 2006). Still other findings are focused on the functioning and environmental impacts of markets as sites of complex activities that may not be distilled entirely into neoliberal logics but rather may be distinguished by embedded relations of trust and knowledge that can support environmentally friendly outcomes. Leading authors in this topical and thematic area include Bakker, Klooster, Keleman and Hellin, Hinrichs, Lave, McAfee, Robertson, and Shapiro.

Human–environment relations: food, health, and bodies

The bodies, health, and nourishment of humans and nonhuman organisms are seen increasingly as being embedded in the matrix of resources and the environment created through human societies. It is an important new topical and thematic area in human–environment geography. Much of the expanding interest in this area is designed to treat issues of health, disease, and the body related to environmental interactions. It is frequently centered on the role of agri-food and fisheries systems that integrate the consumption choices of individuals, communities, and societies together with the transportation networks and agro-ecologies of production systems. The vast majority of consumption, transformation, and production occur through the global corporate, industrial food system. At the same time, growing interest surrounds whether and how the dominant food system is contested and challenged through alternative and local food networks. For example, alternative and local food networks may demonstrate “embeddedness” in which cultural practices and links to ecological landscapes (e.g., conservation agriculture) are considered as simultaneously market-based and potentially involving dimensions that extend beyond economic transactions per se, such as social trust and ethics of care and responsibility.

Hybrid food systems, such as those of peri-urban and urban areas, are also central
to this area of human–environment study. The new direction in human–environment geography described here is also being amplified through new works on the political ecology of health and wellbeing. It includes potential new human–environment insights on bodies in health and disease dynamics linked to the multifaceted role of genetics that is subject to major increases of information and scrutiny. Technological as well as theoretical advances in genetics now highlight the powerful influence of so-called epigenetic effects that result in humans and other organisms being thoroughly embedded in environments in far more extensive and previously unanticipated ways (Guthman and Mansfield 2013). Major geographic authors active in the area covered in this subsection include Bryant, Carter, Durham, Freidberg, Galt, Guthman, Hayes-Conroy, King, Lerner, Mansfield, Morris and Kirwin, Mutersbaugh, Sage, and St Martin.

Environmental landscape history and ideas

The history of environmental landscapes is a major core of human–environment geography that is rooted in a productive past and rapidly evolving productive present state. This topical and thematic area has become transformed in recent years as the result of widespread scientific findings and interpretive shifts, such as postcolonialism. It is being re-invigorated in the light of today’s recognition of global environmental change. This recognition tends to intersect in illuminating ways with influential re-framings of the so-called end of nature and the treatment of the current geologic epoch as the Anthropocene. The history of human–environment interactions in particular landscapes is providing important new insights on the nature of these changes as combined resilience, change, transformation, and, in some cases, the collapse of human societies owing to combined human–environment factors (Butzer and Endfield 2012). The insights extend to a cascade of human–environment works in geography using historical frameworks to demonstrate and detail both the transformation and the resilience of environmental landscapes under anthropogenic change across the time spans of multiple centuries and millennia. They tend to detail chronologies that combine episodes of catastrophic change typically driven through powerful drivers on both sides of human–environment interactions.

The abovementioned findings are often presented as a contrasting and compelling counter-interpretation to purportedly cataclysmic environmental events and narratives of intrinsic marginality that accompanied the earlier interpretations of European colonialism and Euro-American domination of many of the world’s landscapes. Historical resilience of traditional-appearing land use, for example, is shown to also offer potentially innovative contributions to the design of human–environmental sustainability. Leading current authors and established leaders in this topical and thematic area include Beach, Bell, Biehler, Brannstrom, Butzer, Carney, Carter, Colten, D. Davis, Doolittle, Denevan, Dunning, Endfield, Gade, Lightfoot, Lowenthal, Luzzadder-Beach, Myers, Offen, Rangan, Sluyter, B.L. Turner II, Vale, M. Williams, Wilson, and Zimmerer.

Concepts of landscape, including those created through ideas and institutions of environmental governance, are as important as the force of biogeophysical processes on the understandings of this topic. For example, the idea of a “pristine myth,” of nature created in the arts and literature of the Romantic movement in Europe and North America during the nineteenth century, has exerted tremendous albeit misleading influence on environmental understanding. Environmental interpretations steeped in the pristine myth tended to assume the
absence of significant pre-European impacts on the biogeophysical nature of landscapes. While ironically a complete opposite-tending political message, the “empty land” interpretation, an early colonial precursor to the pristine myth (the latter become predominate in the nineteenth century) tended likewise to assume certain ideas about the absence of indigenous presence on the landscape. The legacies of the pristine myth and empty land notions persist to the present, though they are now seen as both more complex and more influential in their spatial and environmental dimensions. For example, they often imply not only the treacherous entwining of the enclosing and erasures of non-Western peoples and their presence but also the creation of relational landscapes comprised of the geographic pairing of unsettled and settled areal expanses suited to European expansions in conjunction with the subjugation of indigenous peoples.

Urban spaces are increasingly understood as having highly dynamic human–environment histories whereby changes, functioning, and morphology of built environments are forged through institutional ideas of resources and landscape. The case of colonial water management in the 1500s in the urban center of Lima, Peru, for example, depended on extensive environmental governance and institutional ideas of resource dynamics and access (Bell 2015). This topical and thematic area can overlap significantly with the approach termed historical political ecology.

The concept of hybrid knowledge systems and spaces has become particularly important to human–environment studies. It is advanced in understandings of such issues as the political ecology of water resources, the production and perception of “invasive” trees in modern forest management, and the territorial designs of networks of protected areas intended for environmental management and conservation use (Zimmerer 2000). Citizen science, based on local knowledge systems, is increasingly promoted in understanding human–environment interactions and the evolving role of the environmental and ecological sciences. Indeed, it appears that citizen science may become a characteristic approach of the Anthropocene as it becomes more widely deployed in myriad projects devoted to understanding present-day environments and human–environment interactions. Important leading authors in the aforementioned topical concepts and ideas has gained new salience amid a broad spectrum of applications in these studies. They have probed and offered insights on such topics as ecological science and concepts related to human–environment dynamics and livelihoods (e.g., the nonequilibrium models of the “new ecology”), the hydrologic cycle, biological conservation corridors, carrying capacity, the science of back-to-nature farming, and scientific forestry management and politics. Technology has also become an area of increased focus in the human–environment analysis of such topics as hydrologic dams and irrigation tube wells that have been spread across large landscape areas (Swyngedouw 1999). These works illustrate the power of scientific ideas and technological tools as deriving from both their geographic dimensions and their usefulness to environmental management as so-called boundary concepts and place-based sites of interaction, negotiation, and dispute.

Human–environment knowledge concepts in environmental management and policy

Knowledge concepts in environmental management and policy comprise an important topic of human–environment geography that spans the diverse domains of advanced science and technology and those of citizen science and local knowledge systems. The history of scientific

New trends in human–environment geography

The accelerating influx of new scientific findings, ideas, and conceptual and theoretical orientations has become a defining characteristic of human–environment geography. Most notable is the increasing rate and scope of anthropogenic environmental change and transformation that now lead human–environment geography, together with other fields, to focus at the intersections of socially vital issues. These topics include the influence and impacts of global climate change on such human–environment systems as water resources, energy, food, health, biodiversity, land systems, and the roles of neoliberal globalization, urbanization, and migration. Framing issues at scales that encompass the global level has become de rigueur in the research design and analysis of human–environment interactions.

Urbanization is also an increasingly predominant theme. For example, new human–environment geography is now focused more extensively on urbanization related to climate change (Leichenko 2011) and also global changes of land systems in a rapidly urbanizing world. Long-distance interactions, referred to as tele-coupling, are increasingly found to link global land and resource use to urbanization impacts and transformations. The powerful influence of these forces of urbanization and industrial expansion has contributed to the important role of cities as sites of much human–environment geography. Leading geographic authors on the kinds of issues mentioned in this paragraph include Braun, Chen, Chowdhury, DeFries, Heynen, Holyfield, Keil, Livermore, Meyers, Munroe, Pelling, Reenberg, Seto, and Swyngedouw.

A second nexus of new intersecting issues is concerned with agri-food systems. This expanded direction in human–environment geography stems in large part from the growing recognition and policy prioritization of the importance of food quality and ecologically sustainable intensification. Major global fora and policy initiatives are being focused on agri-food systems through such high-profile human–environment formulations as sustainable intensification (SI) and ecological intensification (EI) that seek to increase food security while not incurring damage to the environment. Human–environment geography is taking a significant role in addressing these issues through the analysis of converging concerns (Sage 2011; Zimmerer, Carney, and Vanek 2015). One example is the intersecting focus on land systems that, as described above, uses existing methodological designs, tools, and conceptual frameworks to examine land use, its spatial and social organization, and the analysis of land cover and change. Until recently much of the work on land systems dealt principally with deforestation in tropical lowland and temperature areas of tree cover with insights linked to potential biodiversity impacts and climate impacts through influences on carbon stocks and sequestration. Major geographic works addressing these kinds of human–environment issues are authored by Carney, Galt, Moseley, Radel, Sage, Schmook, B.L. Turner II, and Zimmerer, among others.

Agri-food and environmental transformations are also increasingly recognized as integral, often in complex ways, to the human–environment geographies of health and disease (Carter 2014).
Here food and environment mix with political economies and social power dynamics to provide new insights into wellbeing or the lack thereof, since disease dynamics are often wrought through a constellation of similar factors. Agri-food issues are also being expanded in human–environment studies related to the social, spatial, and biogeophysical footprints of energy development, with the emergence of multiple new “energy geographies” highlighting connections to water resources and climate change as some of the latter’s most powerful cause–effect anthropogenic pathways.

Intersections with climate change lead to the most prevalent overarching concern invoked in the current framing of human–environment geography around intersectional issues. Climate change is unleashing growing impacts on such issues as agri-food, health and disease, and energy. These connections are redefining human–environment studies in the discipline of geography as well as more broadly across interdisciplinary endeavors. Impacts of the issue of climate change are also deeply impacting other human–environment disciplines such as anthropology. Major current contributions to the kinds of issues addressed here are authored by Agrawal, Barnes, Brondizio, Calvert, Crate, Dove, Liverman, Moran, Nelson, Orlove, Redman, and Ribot.

The combination of reflexive and instrumental elements is a trend characteristic of works in human–environment geography. Reflexive here refers to the element of critique and broadly defined social analysis that draws on growing social studies of the science, with the latter broadly defined to range from ecology and toxicology to genetics and human biology (Zimmerer 2015). The meaning of instrumental in these examples refers to the practice and application of science and scholarship in human–environment analysis (e.g., identifying the causes and consequences of anthropogenic climate change). Consideration of the reflexive–instrumental integration is a topic of works authored by Forsyth, Goldman, and Turner, among others, that have helped understand and advance this important new trend. Additional significance of this trend stems from the writings of major current social analysts – such as Bruno Latour and Michael Burawoy – who have recently advocated for combined reflexive–instrumental forms of environmental and social knowledge systems.

Examples of the integrated reflexive–instrumental perspective include the studies within human–environment geography that are engaged with the rapidly evolving role and use of ecological and environmental science in political ecology. In addition, the reflexive–instrumental perspective is proving fruitful in developing, using, and critically reflecting on several of the leading current theories and concepts within human–environment geography, especially those of resilience, adaptation, and vulnerability (Bassett and Fogelman 2013; Beymer-Farris, Bassett, and Bryceson 2012; Ribot 2011; Turner 2014). Such studies are able to respond to the increasing social content of scientific concepts amid the accelerating and widening influence of global socioeconomic and environmental changes in such issues as climate change and water resources. Additional leading researchers and authors currently active in this field are Birkenholtz, Lave, McCarthy, Nightingale, Sayre, and Sultana.

Understanding the human–environment impacts and social dynamics of neoliberalism operating across multiple spatial and temporal scales is another trend that defines current human–environment geography. New studies and understandings emphasize the powerful influence on human–environment systems of the political economy, politics, and subjectivities
of constantly evolving neoliberal configurations. The human–environment implications of neoliberalism incorporate factors that range from major global commodity booms (minerals and hydrocarbon-based energy, as well as soy and oil palm) to neoliberal-inflected environmental management and science. The influence of neoliberalism extends also to the protest and social movements among worker, citizen, and peasant groups arising in response to these developments.

Even such alternative movements as those associated with Living Well, which is cast as an alternative to neoliberal development, are unfolding amid the latter's influences. Neoliberalism's vast reach in the realm of resources and the environment thus extends from explicit "market environmentalism" and sustainability payment schemes to extractive industries and far-reaching subjectivities contained in prevailing ideas of citizenship and, more generally, one's sense of community belonging and modernization. Major current contributions to the kinds of issues addressed here are authored by Bakker, Lave, McAfee, McCarthy, Prudham, and Wolford, among others.

Important insights are being revealed in the pronounced influence of neoliberalism – often evidenced through its logics of accounting, metrics, and fungibility – pronounced in such mainstream approaches to environmental management as ecosystem services and territorial designs for nature protection (such as protected areas). Agri-food systems also show the powerful effects of neoliberalism in the dynamic evolution of corporate, industrial systems that effectively utilize global supply chains and consumers' choices on a daily basis. Similar to the examples mentioned in the previous paragraph the neoliberal development of agri-food systems is also unleashing potential alternatives. The human–environment dimension of the international trade in illicit drugs can also be seen as symptomatic of the far-reaching scope of neoliberal market influences. It exerts significant influence on land systems and biodiversity impacts in extensive areas. Major current contributions to the kinds of issues addressed in this pair of paragraphs are authored by Galt, Guthman, McSweeney, and Steinberg.

An additional trend is the productive unfolding of the engagement with concepts of embodiment and perspectival frames. It stems from such diverse influences as technological transformations (e.g., cyborgs) and feminist critiques of science and human–environment studies (e.g., feminist political ecology). These insights are reshaping such foundational concepts as the dualistic assumption underlying human–environment interaction and the "chain of explanation" of interlinked scales of influence in human–environment interactions (Rocheleau and Roth 2007). The existing precepts hinge, respectively, on assumptions of the categorical distinction of humans and the environments as well as non-overlapping scalar processes. Rather than binary-based interactions and neatly nested scales, the new works build on insights resembling human–environment geography's early clarion call to examine the "inner-actions." Rocheleau and her colleagues have written leading publications in this area, which builds on important prior insights by researchers and writers such as Michael Watts.

Theories and concepts of hybridization, socionatural entanglement, and epigenetics have also become central to understanding the multifaceted, reciprocal, and relational intermingling of the natural and social (Guthman and Mansfield 2013; Zimmerer 2000). The process of intermingling among human and nonhuman elements, which is less structured than the earlier image of co-production, bears
broad-brush similarities to the interests and influence of assemblage theory, which likewise strives to incorporate human and nonhuman actors and elements. The new direction in human–environment geography described here is contributing also to rethinking issues of scale and the patterns and processes of scaling that are central to human–environment studies. Important active writers across the areas mentioned in this paragraph include Anderson and McFarlane, Braun, Heynen, McCarthy, Neumann, Rice, Sayre, Shillington, Swyngedouw, Whatmore, and Yeh, among others.

Promise and peril: beyond binaries in human–environment geography

Human–environment geography is currently being defined through a multiplicity of topics and trends. These developments hold both promise and peril for its future. They have been spurred through a remarkable growth of the productivity of research scholarship. While numerical estimates are difficult, one coarse approximation demonstrates a roughly four-fold increase in the 1990s alone with the impression of a similar rate of continued growth in these studies to the present (Zimmerer 2010). To be sure, the arena of human–environment studies was already active in much of the mid-twentieth century. Still it was decidedly thinner and less diverse, being dominated by the two-channel intellectual landscape described above. By the 1970s and 1980s the accelerating developments within human–environment geography foreshadowed a diversification of topics that nowadays ranges widely across global, national, and local levels, urban, peri-urban, and rural spaces, formally regulated and highly informal sectors, from the intensely domesticated to the mostly wild, and multiscale areas of focus from microscale effects at the genome level to the scale of planetary systems.

Mixed methods and matching designs of research methodology have become integral to human–environment geography. Corresponding to its topical and thematic diversification, a swell of research has flooded human–environment geography with an exceptionally broad suite of choices of methods. Both quantitative and qualitative techniques are widely used. Similarly a mix of positivist and varied nonpositivist theoretical orientations has become common, with this critical pluralism emerging through the rapid growth and diversification of human–environment geography. The utility of this “critical pluralism” resonates with the abiding pragmatism of human–environment geography addressed by Wescoat and others. Yet such mixing of methods and concepts does not detract from the importance of generalized distinctions and differentiating theoretical commitments. For example, the large-scale, remote sensing-based human–environment modeling of land use and vegetative cover is often undertaken through the use of knowledge systems quite distinct from that of discourse, whether the latter is global or local in scope (Turner and Robbins 2008). At the same time, there are human–environment works that can and do seek to bridge these gaps, both methodologically and conceptually (Castree, Demeritt, and Liverman 2009; Zimmerer 2015).

In sum, this study’s design has highlighted the intellectual landscape of human–environment geography. A conspicuous feature is the noticeable shift away from the earlier intellectual landscape of binary intellectual spaces. It traces the emergence of the distinctly multistrand configuration of human–environment geography that began in earnest in the 1970s and 1980s and that has flourished since the 1990s. This shift to the post-binary intellectual spaces of human–environment geography has been
entwined with the growth of this area and new efforts to locate its intellectual endeavors within the discipline. This expansion of human–environment geography (or its broad designation as nature–society geography) has helped create the growth of a distinct subfield and intellectual space in the design of so-called four-field or five-field geography. Recent debate and contestation over this distinctness reflects a continued healthy level of intellectual activity in this area of study.

One such central and vigorous debate concerns the degree of distinctness among human–environment and nature–society epistemic centers, and whether and how these knowledge domains may be overlapping in a sense that enables mutual use and possible synergism. For example, a moderate or high level distinctness is illustrated in the depiction of human–environment studies as holding the key to the larger scale of analysis (e.g., global-scale studies) while nature–society is local and based on place-specific case studies (Turner and Robbins 2008). In that research, human–environment studies are represented by sustainability science while the approach of political ecology is used to illustrate a nature–society framework.

It may also be suggested the distinction of these epistemic centers is likely to reside principally in the extensive social and discursive analysis characteristic of nature–society geography. More to the point is the potential value to works that seek to integrate elements of the human–environment epistemic center (e.g., ecological systems, land use, land systems, and remote sensing) and that of nature–society relations (e.g., political ecology). Such integration has been a central, recurring emphasis of various works (e.g., Zimmerer and Bassett 2003). Additional authors of major works that focus on this interface are Bassett, Brannstrom, Beymer-Farris, Chowdhury, B.L. Turner II, M. Turner, and Vadjunec.

Similar ongoing debate concerns the disciplinary niche and openness of human–environment studies. On one hand, this area benefits from active exchanges with interdisciplinary realms as diverse as ecological sciences, urban studies, and environmental studies and philosophy, in addition to development studies, world systems, and political economy. Intellectual borderlands are also vibrant with regard to the other geographic subfields. Related physical geography, for example, incorporates human activities as triggers of disturbance events and management activities that induce the biogeo-physical processes and pathways of vegetative, geomorphic, and climate-driven conditions and changes. Many productive interactions are similarly promised in the interconnections of human–environment studies of land and climate change to human geography in general and economic geography in particular.

A signal of concern does attach to the otherwise positive prospect of future human–environment studies in geography. It suggests the ongoing success, demonstrated through its diversification and expansion amid the recognition of the Anthropocene, may potentially entail tradeoffs. Such developments could lead to a scenario of human–environment issues becoming more widespread as topics of interest while potentially diluting points of specific, concentrated research and understanding. Strengthening coherence and intellectual rigor, while continuing to embrace a characteristic openness and critical pluralism, is both a challenge and opportunity that must be engaged. Doing so is necessary to address the rapidly evolving roles of human–environment geography. Actively engaging such challenges and opportunities must become a priority, both in the discipline of geography and in the broader realms of interdisciplinary environmental studies and sciences.
SEE ALSO: Agricultural environments; Berkeley School; Biodiversity; Chicago School; Climate change adaptation and social transformation; Conservation and capitalism; Development; Energy resources and use; Environment and development; Environment and migration; Environment and urbanization; Environmental degradation; Environmental discourse; Environmental education; Environmental governance; Environmental hazards; Environmental health; Environmental history; Environmental (in)justice; Environmental knowledges and expertise; Environmental management; Environmental policy; Environmental science; Food security; Health and wellbeing; Human ecology; Land change science; Land degradation; Landscape; Livelihoods; Natural resources; Neoliberalism and the environment; Political ecology; Population and natural resources; Resource extraction; Resources and development; Social-ecological transformation; Social resilience and environmental hazards; Social vulnerability and environmental hazards; Sustainability science; Sustainable development; Urban political ecology; Vulnerability; Water resources and hydrological management

References

McSweeney, Kendra, and Oliver T. Coomes. 2011. “Climate-Related Disaster Opens a Window of


Further reading


Internationally, national policies and standards in geography education span a wide range of formats, implementation procedures, and assessments. Some countries engage directly with national policies and standards to create and assess a national curriculum or course of study in geography at both primary and secondary education levels. These national curricula are integrated into the education system and may or may not include teacher training to prepare instructors to teach the required content. Other countries maintain a decentralized or non-federally controlled education system. Efforts made to compare geography education in these various systems are subject to the constraints and limitations of evaluating the different systems in which geography curricula are determined, taught, and assessed.

A comprehensive comparison of geography education curricula at the primary and secondary levels across many countries is not easily produced based on existing literature. Some international efforts provide limited descriptions and some discussion or comparisons of how geography education is included in national education policies and standards. In addition to the policies or political entities influencing geography, there are also differing philosophical approaches to how geography content is presented. Geography is sometimes taught as a natural science, social science, or an integrated study of human–environment interactions. In some cases, a specific geographic approach, such as the spatial and ecological perspectives or space and place, heavily influence the geography content in certain countries. It would be valuable for future publications to include descriptions of the national policies and standards for geography education in more countries to establish more examples to increase the understanding of geography education from an international perspective.

The place and role of geography education varies greatly from country to country. The context of the education system itself provides various settings of distributed, locally controlled systems to nationally controlled or determined curricula. Examples of the standards and national policies in selected countries are provided in this entry. While not intended to be exhaustive, the examples provide a snapshot of the wide scope of geography curricula and national policies in just a few countries with a more detailed discussion of the United States national context. The following countries are highlighted based on recent publications documenting an international effort to examine and compare geography education at the national level.

Case study: Australia

In 2008 the Australian government moved away from geography content determined locally by
GEOGRAPHY EDUCATION: NATIONAL POLICIES AND STANDARDS

states and territories to a national Australian curriculum. The national geography curriculum for primary and secondary schools was approved in 2013. The conceptual and philosophical underpinnings of the new Australian geography curriculum are explained through its definition, concepts, specification of place knowledge, embedding of skills in inquiry, and state or implied perspectives (Maude 2013). Of these concepts, the curriculum tends to differ from some other national curricula in its definition, range and choice of concepts, and specification of place knowledge. The new national curriculum document includes a definition of geography that was accepted as it was inclusive of existing definitions from the states and territories.

The document is built around key concepts that are complex ideas that could not be easily defined, but only described by the ways of thinking they produced (Maude 2013). Twenty-seven points in the curriculum document form a catalog of ways of thinking geographically that are a guide for teachers on the broad geographical understandings students should develop. Previously taught as part of studies in society and environment or human society and its environment at the primary level, the new national curriculum strengthens the position of geography in Australian schools. At the primary and lower secondary levels, there is now a separate and compulsory curriculum for geography. Geography is an elective subject for upper secondary with states or territories having the option to make it a compulsory subject (Maude 2014). The national curriculum board in Australia is called the Australian Curriculum, Assessment and Reporting Authority. This entity established the framework for the curriculum, but did not influence the content which was developed by a group of writers and advisors with oversight and reviews at many levels. The recent transition from state and territory curricula to a national geography curriculum is viewed as a positive change for geography education in Australia.

Case study: Finland

Finland’s geography content has served as a model for other countries in the process of reviewing or updating their national curricula. Finland itself is currently preparing to implement a revised curriculum. The Finnish National Board of Education is the controlling body for a specific core curriculum in Finland. The national curriculum was revised during 2012–2014 and will be implemented in classrooms starting in 2016. The role of the curriculum is highly regarded and importance is placed on the curriculum as a guide for instruction and content throughout the country. There can be some local modifications of the curriculum so long as it still adheres to the national core curriculum. A national education evaluation center was established in 2014 to monitor educational progress in the country.

The curriculum places geography as part of the national core curriculum for grades 1 through 9 or the basic education level. Traditionally, geography is more closely connected with the natural sciences rather than the social sciences. In grades 1–4, geography is taught as a natural science in environmental and natural studies. During grades 5–6, students take a required geography and biology course. And in grades 7–9, a required stand-alone geography course is included in the curriculum (Solem, Lambert, and Tani 2013). In the renewed version, geography will be integrated along with biology, chemistry, physics, and health education as a part of environmental studies for the primary level students.
Case study: Philippines

The Philippines does have a national curriculum, however, geography is not considered a discrete subject in this curriculum. The “K to 12 Basic Education Curriculum” (primary and secondary) was published in 2013 and includes curriculum guides for different subject matter of study. Closer examination of the national “K to 12 Curriculum Guide: Science” reveals an emphasis on inquiry and informed decision-making on issues that involve science, technology, and environment (Republic of the Philippines Department of Education 2013). The earth science components include some physical geography concepts but lack the spatial perspective as a part of study and analysis.

The “K to 12 Basic Education Curriculum Guide: Social Studies” contains seven broad themes that are based on the “US National Council for Social Studies” ten themes for the social studies which is cited as a source. Geography is not identified as a stand-alone discipline but is embedded within the seven themes content. Appearing most often in grades 1 to 4, geography content is very limited in the curriculum. The use of geography skills is referenced at grade 4, but is not used after that in the curriculum sequence. As is the case with most curricula focused around social studies, there is thematic content based on broad concepts or ideas, but a lack of discipline specific content. The diffusion of curriculum ideas between countries seems to be a factor influencing the content of this national curriculum.

Case study: Singapore

The Singapore national curriculum in geography is described by a national syllabus document that is reviewed and subject to change every six years. Geography is included as part of a social studies subject at the primary level and is a compulsory subject in lower secondary levels. At the upper secondary, geography is an elective subject. The national government is very influential in determining the primary and secondary school curriculum. Singapore has been described as using education as an agent for social change (Yip, Eng, and Yap 1997). Government identified issues such as climate change, environmental stewardship, and respect for all cultures are examples of governmental influences on curriculum content. Structured government and Ministry of Education influence guides changes made in the national curriculum for primary and secondary schools. The development of the national syllabus document also includes academic geographers and reflects changes in the discipline heritage as well (Chang 2014).

The geography content at the primary level is expressed through four social studies goals that include the interactions and relationships between humans and physical environment in which they live. The lower secondary school geography is organized according to human or physical geography topics around a central theme of physical–human relationship (Chang 2012). The upper secondary school geography has evolved to a more conceptual approach highlighting the human–environmental and spatial relationships with some recent evidence of changes back to a more regional studies approach (Chang 2014). The geography course at the upper secondary level is elective and results in approximately 50% of students selecting this course. Acknowledging the strong state presence in geography curriculum development and the review process, Chang (2014) describes the necessity for a continuing strong partnership between academic geographers and geography educators in future reviews and discussions of the national syllabus document for Singapore.
Case study: United Kingdom

The geography education standards are part of the national curriculum program of study in the United Kingdom. The current program of study was recently revised and published in 2013 for classroom implementation in 2014. Although it is a national curriculum, individual countries may review the curriculum independently. While England will implement the revised curriculum, Wales is scheduled to conduct a review of the curriculum. Geography is a required subject at primary school level and also in the secondary to 14 years. There is no requirement to offer geography after 14 years, but it is offered in most secondary schools in the country. Geography is not necessarily taught as a discrete subject, but rather it is most often integrated with science or history or in themes such as environmental studies (Solem, Lambert, and Tani 2013).

The geography national curriculum is articulated in three key stages. Each key stage includes a purpose of study, aims, attainment targets, and subject content. The 2014 national program of study is described as being concise and defines core knowledge that students should acquire including locational and place knowledge, human and physical processes, and some technical procedures as well as content in fieldwork, the use of maps, and written communication (Geographical Association 2014). There is also a secondary level geography course that students may take in preparation for the General Certificate for Secondary Education (GCSE) exam. This content recently underwent revision and the updated content will be implemented in classrooms in the 2016 academic year. Specific geography content is included in the national program of study at all levels of education in the United Kingdom.

Case study: United States

Geography education in the United States does have a place in national education policies and standards despite the lack of a national curriculum or a federal education system. Education systems in the United States are primarily controlled at the state or school district levels rather than the federal level. In this distributed education system environment, each state determines its own curricula, assessments, criteria for accreditation, and criteria for teacher preparation and licensure through state or local school boards or legislative mandates. Therefore, the only way to accurately describe the national context of geography education in the United States is to provide a description of standards, curriculum, assessments, and teacher licensure requirements for each of the 50 states, federal district, and territories. The description of geography in the national education context is further complicated by the fact that geography is often taught as a strand in the social studies curriculum which includes the disciplines of history, civics (political science), economics, and occasionally sociology and psychology as well.

Geography’s inclusion in the social studies is most common at the primary level. Middle school curriculum increases the chances of a stand-alone geography course in many states. At the secondary or high school level, curricula vary the most across the United States. At the secondary level, states may have a stand-alone course, an embedded approach, or no geography at all. Most recently, the National Council for Social Studies approved the College, Career, and Civic Life for Social Studies State Standards (C3 Framework), which identifies civics, economics, geography, and history as the focused four disciplines comprising the social sciences. The scope of adoption of the C3 Framework by US states is still unknown among the states.
Finally, enrollment in the Advanced Placement Human Geography course offered through the college board for college credits has increased significantly during the past 10 years. This addition to the K-12 geography curriculum most often appears at year 9 in the US system. Given the context of a distributed education system, national policies are limited and national standards are voluntary and typically serve to inform the independent development of mandatory state level social studies education standards.

United States policies and initiatives

Efforts to improve geography education at the national level included the High School Geography Project during the 1960s, but met with limited success (Association of American Geographers 1966). A second national initiative responded to the publication of a less than glowing report on the state of US education in the early 1980s. The publication of *A Nation at Risk: The Imperative for Educational Reform* issued a call to improve US education (National Commission on Excellence in Education 1983). The geography community responded with a publication, *Guidelines for Geographic Education* in 1984 by the Joint Committee on Geographic Education. Despite a collective effort by the geography education community to establish geography content in the curriculum, the first standards document, *Geography for Life* wasn’t published until 1994 (Geography Education Standards Project 1994).

Despite the lack of a national curriculum or course of study, geography was included as a subject in the National Assessment of Education Progress (NAEP) national testing in 1994. The National Assessment Governing Board (NAGB) oversees the administration of the NAEP testing and determines the test schedule for subject matter examinations. A framework for the assessment was developed as a separate effort from the National Geography Standards which were published the same year as the first NAEP test. The 1994 NAEP test set an initial benchmark for student achievement in geography at grades 4, 8, and 12. Subsequent NAEP assessments of student achievement in geography were conducted in 2001 and 2010.

United States national geography standards

While there are no nationally mandated education standards in the United States, efforts at developing discipline specific voluntary national standards are actively pursued by nonprofit professional organizations, foundations, and with some funding from the US Department of Education. The US Department of Education requirements shaped how the content was presented in the standards document. A large committee of geographers and K-12 geography educators worked to develop an outline of the content necessary to develop a “geographically informed person.” The goals of *Geography for Life: National Geography Standards* 1994 outlined the geographic perspectives, skills, and content knowledge necessary to be geographically informed.

The spatial, ecological, and subperspectives provide the context in which students formulate geographic questions. The geographic skills outline a scientific approach to asking and answering geographic questions using geographic data, content knowledge, and analysis. The geography content knowledge was organized into six “Essential Elements” that included The World in Spatial Terms; Places and Regions; Physical Systems; Human Systems; Environment and Society; and The Uses of Geography (Geography Education Standards Project 1994). Eighteen content standards were organized under the six Essential Elements. The
standards served to describe and explain physical and human geography content for the K-12 curriculum. *Geography for Life* (1994) was a highly successful document that influenced state-based authors of curriculum and state standards documents. Following its publication, standards for social studies and geography education in all US states and the District of Columbia reflected content in *Geography for Life: National Geography Standards 1994*.

In 2007, a committee of geography educators, K-12 teachers, and academic geographers were recruited to work on a revision of the 1994 standards document. This committee agreed the second edition document should follow a similar format due to the number of state standards documents that had used first edition content and maintain the goal of identifying what is needed to develop a geographically informed person. The second edition committee organized all content by standards and for each standard; all three grade levels were presented in a three column layout allowing users to view content across the grade levels. As a part of the revision, the committee focused on the three principles of alignment, scaffolding, and intellectual skills (Heffron and Downs 2012). Alignment was defined as the vertical structure of a standard’s content for each grade level. Scaffolding was defined as the horizontal structure of a standard’s content across or between the grade levels to insure students would be building knowledge and complexity of understanding in the geography content for each standard. Intellectual skills were defined as the ways in which students can think geographically and were represented by a set of verbs selected to vary the intellectual challenge of the geography content.

Published in 2012, the second edition document increasingly focuses on outlining geographic perspectives, content knowledge, and skills, all working together to engage students in geographic inquiry. Revisions to two of the 18 content standards (Standards 1 and 8) and updated content including organizing themes for each content standard were made in the document. Additionally, an expansion of the extensive glossary was deemed valuable due to the limited teacher preparation specific to the discipline. The second edition document went through numerous reviews and public comment periods to inform the revision process prior to its publication.

**Future research directions**

Future research on the nature and content of curricula from an international perspective may serve to help inform the revision and development of future national standards or curriculum efforts. International perspectives would also benefit the discussion of the philosophical underpinnings and the way geography subject matter can contribute to the value of national education systems. There is evidence of the diffusion of ideas and concepts in the curricula between countries. Future research might chart or map international exchanges of ideas that are influencing national curricula. Further analysis and research may also inform areas of collaboration that might result in the areas of teacher training and professional development, instructional materials and resources, and strategies for high-quality student assessments. There are also underlying questions about the use of emerging geospatial technologies and geolocated data in instructional settings as a part of future geography curricula. Identifying potential alignment in curriculum and content may also provide opportunities for collaboration that would develop or enhance international student perspectives about other places in the world. Future research in
the area of national policies and standards holds
great potential for informing and improving
geography education at a global scale.

Summary

These six case studies are presented to highlight
differences not only in national education con-
texts and policies, but also in how geography
content is placed, explained, and taught in both
compulsory and voluntary national standards.
Some geography content at different grade levels
is evident in all countries regardless of whether
there is a national curriculum or federally man-
dated system or not. Several of the case examples
demonstrate a strong collaboration between
academic geographers and geography educators,
while it appears to be absent in other contexts.
In one context, the national curriculum may be
politicized to address perceived national issues
through education. While a few publications
document the state of geography in each coun-
try, additional published accounts for more
countries would help to inform future analysis
and comparisons of geography education from
an international perspective.

SEE ALSO: Geography education: primary
and secondary; Geography education, primary
and secondary: international perspectives;
Geography education, workforce trends,
twenty-first-century skills, and geographical
capabilities; Spatial thinking, cognition, and
learning

References

Association of American Geographers. 1966. High
School Geography Project: Geography in an Urban Age.
New York: Macmillan.

Teaching and Learning School Geography: The
Case of Singapore.” International Research in
Geographical and Environmental Education, 21(4):
1–13.

Chang, Chew-Hung. 2014. “Is Singapore’s School
Geography Becoming Too Responsive to the
Changing Needs of Society?” International Research
in Geographical and Environmental Education, 23(1):

Geographical Association. 2014. “Introduc-
ing the New National Curriculum.” www.
geography.org.uk/news/2014nationalcurriculum/
introducingnc/ (accessed May 16, 2016).

Geography Education Standards Project. 1994. Geog-
Washington, DC: National Geographic Research
and Exploration.

Heffron, S.G., and R.M. Downs, eds. 2012. Geo-
raphy for Life: National Geography Standards, 2 n d
Geographic Education.

Maude, A. 2013. “The Vision of Geography Under-
lying the Australian Geography Curriculum.”
Review of International Geographical Education Online,
3(3): 253–265.

Maude, A.M. 2014. “Developing a National Geog-
raphy Curriculum for Australia.” International
Research in Geographic and Environmental Educa-
858437.

National Commission on Excellence in Education.
1983. A Nation at Risk: The Imperative for Edu-
cation Reform. Washington, DC: US Government
Printing Office.

Republic of the Philippines Department of Edu-
cation. 2013. “K to 12 Curriculum Guide:
Science.” http://www.deped.gov.ph/k-to-12/
About/curriculum-guides/Grade-1–10 (accessed
May 16, 2016).

“Geocapabilities: Toward an International Frame-
work for Researching the Purposes and Values
of Geography Education.” Review of International
GEOGRAPHY EDUCATION: NATIONAL POLICIES AND STANDARDS


Further reading


Community colleges in the United States provide academic, vocational, and professional education for students who have typically finished their secondary education and are seeking additional training or qualifications for professional advancement or intending to transfer into a four-year bachelor’s program. Such institutions typically award associate of arts (AA) or associate of sciences (AS) degrees as well as other types of professional certificates, diplomas, and qualifications. Institutions similar to US community colleges exist in other countries, but this entry focuses on US institutions because they: (i) enroll a large number of students in introductory geography courses, and (ii) serve as a major source of students moving into four-year bachelor’s programs. In the United States, these institutions are sometimes called junior colleges, technical colleges, two-year colleges, and city colleges, although the focus of this entry is on those that award two-year associate degrees in geography.

Community colleges began a century ago as an effort to develop low-cost undergraduate education at convenient locations for a growing US white-collar labor force as well as to provide advanced technical training for those in the blue-collar workforce (Ratcliff 1994). They still serve much the same role over a hundred years later, though often encompassing a broader mission including serving an expanding demand for adult and community education. There are currently over one thousand community colleges in the United States. The majority of these are public, but approximately a hundred are independent and private and about thirty-five are tribal colleges. Approximately 750,000 associates degrees and 460,000 certificates are awarded yearly by community colleges (American Association of Community Colleges 2015a).

Today, almost half (46%) of all the undergraduates in the United States are enrolled in the community college system and of these 41% are first time students. In the fall of 2013–2014, there were a total of 12.4 million students enrolled, with 7.4 million enrolled for credit. Of this total, 4.5 million were part-time students, and 2.9 million were full time (American Association of Community Colleges 2015a). Clearly, community colleges play an integral role in the higher educational system of the United States and tend to serve a more diverse range of students than four-year colleges and universities.

Vocational, technical, and further education colleges in the United States and other countries are able to offer more specialization in certain areas, specifically, in high-demand technology fields. The mission of these institutions is much narrower than community colleges, which place greater emphasis on a traditional liberal arts curriculum, particularly bread-and-butter English and mathematics courses. The relationship between all of these institutions is becoming increasingly more complex, with the “unbundling” of certificate and degree paths (students transferring credits from multiple institutions) that create new pathways toward diplomas and credentials, but also alter traditional structured progressions that are typical of both associate’s and bachelor’s degrees. In some cases...
this relationship becomes even more complex, with community colleges subcontracting with other public and private institutions for some teaching, technological, administrative, financial, and management support.

Community colleges are localized in structure, funding, and mission and the placement of geography in the curriculum is as varied as the institutions themselves. In some cases geography is part of the science, technology, engineering and mathematics (STEM) curriculum when programs feature physical geography and geographic information systems (GIS) classes, while in others geography is part of the liberal arts and social science curriculum with programs that feature world regional geography as well as introductory classes in human and cultural geography. Many programs also offer classes such as economic and political geography, specialized courses with a regional and local focus, and other courses that fit into the liberal arts and social science curriculum. Sometimes geography courses are offered within other academic programs such as global studies, environmental science, and anthropology.

Unlike many four-year colleges and universities, few community colleges are residential and only about 25% of community colleges offer any sort of on-campus housing (American Association of Community Colleges 2015b). Most students live at home and may attend college part time while working. Community colleges are funded through tuition (often underwritten by student loans) as well as local, state, and federal support. In contrast to most four-year colleges and universities which receive little or no local or municipal funding, community colleges receive an average of about 17.8% of their funding from such sources as well as approximately 9.9% of their funding from grants and business partnerships (American Association of Community Colleges 2015a).

Roles and demographics

Community colleges play a key role as pathways to education and better jobs for immigrant communities and, in turn, are heavily influenced by the needs of these communities. Currently, one in four community college students are landed immigrants or children of immigrants (Connell 2008). “The national conversation about immigrant education has only just begun, but community colleges are well positioned to meet the needs of this growing student population. By helping our newest neighbors successfully integrate into US society, community colleges will contribute to the overall health of the economy and fortify the country’s social fabric” (Wisell 2010).

Certainly there is a challenge to be met – while immigrants’ overall educational attainment has remained the same since 1970, the educational attainment these immigrants bring with them has decreased dramatically (Haskins 2007). This would certainly seem to indicate that while immigrating to America today is more accessible than ever, the people that are emigrating from their countries of origin are not as well prepared to meet the educational requirements for employment in the US workforce. These immigrants need access to upgrade skill sets (especially technological) at local community colleges that match the local labor markets and allow them entry into the market. The education offered by community colleges can also act as a stepping stone that allows immigrant students easier access into 4-year institutions, particularly those with what are termed “Direct Transfer Agreements” with community colleges. Also, for many families that struggle after moving to America, community colleges offer the most affordable pathway to higher education.

Community colleges are particularly important for many students because they focus on helping
students succeed in classes needed for their first two years of college. Geography is often included as a “general educational requirement” in these programs and, as a consequence, can sometimes serve to attract students into further study. This is important because community colleges tend to attract a more diverse student body with greater proportions of underrepresented populations. This means that community colleges in the United States can help to promote diversity and inclusion within the discipline.

Community college students, in contrast to many in 4-year institutions, are often balancing other work, family, and personal responsibilities as they pursue their education. Of all community college students, 59% (4016789) attend on a part-time basis compared to 41% (2849203) that attend full time. This seems to reflect a divide among community college students between those seeking an associate’s degree for transfer to a 4-year institution and those in vocational/technical programs or enrolled in fewer classes as they work and balance family commitments. When looking at community college students starting out, 39% transfer to a 4-year institution, while 42% receive a vocational/technical certificate. While the students whose goal is a four-year degree may not realize what academic or vocational area they may choose to enter, students in vocational/technical programs and balancing life commitments often have very specific goals in mind, such as becoming a nurse, welder, or GIS professional.

However, the reasons students attend community college is as diverse as their personal backgrounds (Johnson and Sullivan 2010) as shown in the following list.

- High school graduates completing lower-division courses before transferring to four-year institutions.
- High school students co-enrolling in courses at community colleges or taking courses at colleges as part of homeschooling programs.
- Students pursuing Career Technical Education (vocational) degrees or certificates.
- Adults working toward General Equivalency Diplomas (GEDs) or entry-level workplace competencies (such as Welfare-to-Work2).
- Re-entry students who have been in the workforce for some time and are looking for new careers or skills.
- Students of all ages pursuing personal interests.
- Students who have already earned bachelor’s degrees or higher and want new skills or skill upgrades. The growing trend of these so-called reverse transfer students has led some to refer to community colleges as the new form of graduate school (Arnone 2001).
- Long-term unemployed students and high school graduates who hope to acquire skills needed to enter or re-enter the workforce.

Some facts that stand out are that the average community college student is 28 years of age, 36% of students are the first generation to attend college, and 17% of all students are single parents. This would certainly present a much different picture than comparable demographics at 4-year institutions and, to put it simply, many of these students are getting a chance they would not get elsewhere. With the average community college tuition at $3347, just over a third of the $9139 tuition cost at 4-year institutions (American Association of Community Colleges 2015a).

Challenges facing community colleges

Geography is part of the curriculum at many community colleges, but it is not a major discipline. This means that that 24% of geographer
faculty in 2-year institutions serve alone as the only instructors in their field or what are termed stand-alone geographers (SAGE). This can often place these faculty and part-time instructors in challenging positions in regards to synergy and interaction with colleagues in their field as well as competition for resources for teaching. 

Teaching geospatial classes can be especially challenging at the community college level, with these classes requiring more investment in computer software and hardware, and typically requiring lower faculty to student ratios and smaller class sizes. In effect, these classes require a greater investment on the part of the college and this requires support from administration especially when competing for resources with engineering and other technological fields.

One significant demographic characteristic of community college students is that 61% of all students are female and 39% male. With a higher representation of night-time and online classes than 4-year institutions, community colleges offer greater flexibility for single-parent families, and those where a female partner is attempting to re-enter the workforce after her children begin school (this would be only one example of many different arrangements that would require more flexibility in scheduling attendance in classes). Looking at ethnicity, 68% (3,981,099) of all community college students are Caucasian, 15% (1,003,795) are Hispanic, 13% (865,578) American Indian/Alaskan Native, and 6% (421,097) are Asian/Pacific Islander.

Although community colleges rely less on state funding than four-year colleges and universities, many have still not rebounded from cuts from the great recession less than 10 years ago. These institutions are also subject to localized issues that cut back funding, such as major industries relocating out of an area or shifts in allocation of local revenues.

This is causing many if not most community colleges to “do more with less.” Most community colleges have responded with a variety of efficiency measures from utilizing more part-time instructors to “lean” methods of streamlining routine campus operating costs (such as printing, classroom reservation systems, advising, how class catalogs are offered, and enrollment systems). In some cases, this has also resulted in the offering of fewer geography classes (and liberal arts classes as well) and more “bread-and-butter” math and English classes.

The most common trend is to rely heavily on more affordable part-time teaching staff and it is becoming increasingly important to make the case of why full-time faculty make a difference in the quality of higher education.

More recently, community colleges are increasingly finding their relationships with for-profit institutions becoming more complex. While for-profit institutions played a minor role in the 1990s, they are becoming more of a factor with their emphasis on customer service, degree completion rate, and support for students finding employment after they graduate. This is balanced by community colleges maintaining a greater emphasis on a broad-based curriculum with fewer constraints than for-profit institutions which often have very limited missions.

There are also many cooperative arrangements between the community and for-profit colleges with community colleges preparing students for later upper-division classes at for-profit institutions. It is still the case that community colleges have an advantage over for-profit colleges in regard to articulation agreements or direct transfer agreements with four-year colleges and universities. These agreements specify how course credits can be transferred from institution to institution. Many four-year colleges and universities do not accept credits from for-profit colleges or transfer the credits as less
specialized general study (independent) credits. Many community colleges and statewide 4-year institutions have direct transfer agreements in place, which allow priority for transferring community college students.

**Instructional and learning technologies**

The extent and quality of instructional facilities are as varied as the communities where these colleges are based and owe much to each college’s unique history and funding sources. One thing that most share is the lack of classroom space. This has meant that, for several decades, online programs and classes have provided community colleges with the opportunity to increase enrollment and add new classes and programs, even during challenging times.

Many colleges are now starting to question the efficacy and success rates of online classes: “Poor online performance rates in community colleges are not simply due to the characteristics of students who choose to enroll in those courses. Challenges related to the online format – including technical difficulties, a sense of isolation, a relative lack of structure, and a general lack of support – may contribute to poor performance, particularly among community college students” (Moore 2013).

One of the great challenges over the next several decades will be the balance that institutions will seek to provide in regards to blended learning and teaching. Learning communities, linked classes, and hybrid and flipped classrooms are some strategies being employed. These will continue to evolve as the era of the lecture classroom collides with modern technology and lifestyles. However, this evolution will be tempered by the reality of fiscal economics and providing efficient forms of classroom structures that feature quality and technology that are innovative and cost effective.

**Opportunities and challenges**

Over the next decade increasing opportunities in geographic information science (GIScience), geographic information systems (GIS), and geospatial technologies may play a role in the development of additional courses and programs in community colleges. The development of the Geospatial Technical Competency Model by the National Geospatial Technology Center of Excellence in 2013 has provided community colleges with a model and framework to provide the knowledge and skills needed by today’s geospatial technology professionals. The model framework was developed through a collaborative effort involving the Employment and Training Administration (ETA), the Geotech Center, and industry experts (GeoTech 2015).

In addition, the Geotech Center also provides a wide range of support for geospatial education from skills competitions and mentoring programs for undergraduates to model courses and syllabus repositories for faculty and instructors.

Courses and programs focusing on geospatial technologies, remote sensing, and other GIS applications are growth areas, but geography also has a vital role to play in exposing students to a wide range of important contemporary environmental and geographical topics such as global climate change, the impacts of globalization, the spatial dynamics of American cities, and many other topics. The rise of interest in spatial thinking and learning as a fundamental mode of human cognition that cuts across many domains of STEM education may also offer opportunities for curricular innovation.

In general, one of the greatest issues facing community colleges is the increasing reliance
on part-time rather than full-time faculty. Currently there are often more than twice as many part-time faculty as full-time faculty at most institutions. This means, potentially, that there are fewer full-time faculty available to advise student and to devote energy to curricular innovation and improvement. Some of the more encouraging factors are that female faculty and faculty of color are well represented and increasing in numbers. However, many faculty at community colleges, especially full-time faculty, are approaching retirement age. Addressing this generational changeover and attracting replacements for retiring faculty may prove a challenge; those qualified to teach often have more diverse employment options available (Rifkin 2015).

Community colleges do offer opportunities for successful professionals to “moonlight” and supplement their income with part-time teaching. It might be useful in the future to encourage such moonlighting by full-time educators at four-year colleges and universities to help build, strengthen, and expand community college programs. This is because one of the greatest challenges facing community colleges is the issue of staff turnover and providing some sort of financial security for faculty. While some of these factors can be ameliorated with better benefit packages for full-time faculty, this does not necessarily impact part-time faculty. All faculty – full- and part-time – can benefit from faculty development programs, early career training, support and mentoring, and programs for professional development. Such programs are, unfortunately, short-changed in efforts to cut cost.

**Conclusion: looking to the future**

In the United States, community colleges currently serve almost half the nation’s undergraduates. The forecast is for enrolment to increase by approximately five million students over the next five years. The financial and staffing strain on smaller programs may be daunting without increased support from the federal government. Currently, federal assistance accounts for about 15% of community college funding, with state revenue and tuition providing about 29% each of community college revenue. About 40% of all students work to pay tuition and living expenses (American Association of Community Colleges 2015a). The federal proposal to boost federally supported Pell Grants and to provide “no cost” tuition for the first two years of community college education could have a revolutionary effect on the number of students served by community colleges. Even if these changes do not materialize, community colleges will remain a backbone of higher education for blue-collar, working-class students, a path to accessing technological skills for others, and a point of entry into American society and the US workforce for first-generation immigrants.

In looking forward over the next few decades, it may emerge that community colleges will begin to offer both associate’s and bachelor’s degrees, thus blurring the distinction between two-year and four-year institutions. It may also transpire that consortiums will emerge between private industry and community colleges to create educational “pipelines” for students interested in specialized career opportunities as varied as aircraft manufacturing and remote sensing. The success of companies such as Google and Esri has increased a global demand for geospatial skill sets, and community colleges will continue to be centers of innovation in these and other areas of innovation.

Community colleges will continue to play an important role in debates over affordability and access to higher education in the United States and, perhaps, in other countries. Recently, for example, the community college system in the
United States has been thrust into the national spotlight with a proposal to provide education free of cost to all Americans. The ten-year, $60 billion plan, mentioned above, would pay for the first two years of community college for full-time students that maintain satisfactory grades. The impacts of this new legislation could be profound. In a recent speech, President Obama (2015) noted that jobs requiring at least an associate degree are projected to grow twice as fast as those requiring no college experience. It was suggested that by 2020, America will once again have the highest proportion of college graduates in the world, and that community colleges produce an additional five million graduates. This certainly presents a charge for community colleges in the coming years, placing considerable weight on their ability to serve these numbers.

Addressing this charge may involve changes on many fronts. One of the most important is closer coordination between community colleges and other institutions of higher education. Most articulation agreements between community colleges and universities are between public institutions within a state system or within a given geographic area. More can be done in this area. For example, one effort underway in the State of Washington has been to create “common course” numbering to eliminate a multitude of different course numbers for similar classes offered at different institutions.

A second area for improvement involves having a more meaningful dialogue between not just administrators and instructional councils, but also faculty in how courses align and match between these institutions. It is in everyone’s best interest that students have a clearer understanding of what classes they need to take for transfer from one institution to another and from one degree program to another. Given current financial pressure and growing enrollments, both two-year and four-year colleges have an interest in increasing degree completion rates and lowering the average time-to-degree. These efforts are also in the students’ best interest as the cost of education and the risk of high student loan debt rises.

Professional development is another area of concern for faculty at community colleges. Given the changes underway in higher education, as well as the pace of change in some fields like geospatial technologies, many part-time faculty find themselves too busy or simply are not aware of opportunities for professional development. Certainly organizations on and off campus can play key roles in improving these opportunities.

More geographical and educational research needs to be directed toward community colleges. Although community colleges are a very important part of US higher education, they haven’t received as much research attention as four-year colleges and graduate education, particularly doctoral programs. Most attention has focused on policy plans or on issues regarding administrative streamlining, financial cost-cutting, and topics such as class size. These last issues are important but, given the role community colleges play in US higher education, more attention might focus on the sorts of broader issues alluded to in this entry. These include the impact of community colleges in helping blue-collar and working-class students move into and advance in the workforce, developing student technological skills, and supporting the success of first-generation college students and recent immigrants. In addition to research in these areas, more attention needs to focus on the impact of part-time versus full-time faculty in these institutions as they prepare for yet another era of increased enrollments and reduced budgets (Belfield 2015).

Finally, it is worth considering the role of institutions like community colleges in other countries. This entry focuses on the United States, if only because community colleges offer
so many geography courses to so many students. But geography is an element of vocational, technical, and further education colleges elsewhere around the globe. It is this sector of higher education – positioned midway between secondary schools and four-year colleges where change and growth may occur. This is not only in the United States, but in other countries where increasing demand for students with geospatial and geographical knowledge and skills may be served by increasing numbers of innovative programs.

SEE ALSO: Geographic information science and technology: educational directions; Geography education: digital and online trends; Geography education, workforce trends, twenty-first-century skills, and geographical capabilities; Geography in higher education

References


Geography education: digital and online trends

Michael N. DeMers
New Mexico State University, USA

Traditional lecture-style, face-to-face classrooms are still the norm in geography education, especially in colleges, but this is changing rapidly. The costs of classroom space, the varied schedules and learning styles of today’s learners, and the availability of an increasingly sophisticated set of computer software tools are resulting in a move to digital and online learning as a growing proportion of the total learning environment. Concurrently, new learning paradigms are appearing with increasing frequency. They include models of teaching based on the nature of the learner’s age, background, experience, and motivation such as pedagogy versus andragogy (Knowles, Holton, and Swanson 1998) versus heutagogy (Hase and Kenyon 2000). Some models are more structural, including problem-based learning and the flipped classroom. Still others focus on styles of learning from auditory, to reading-based, to visual, to tactile and experiential learners (Kolb 1984) and even include the concepts of multiple intelligences (Gardner 2000).

While much of this explosion in research regarding learning paradigms may be prompted by pure research stemming from educational and especially cognitive psychology, a major driving force may be found in the very nature of today’s neomillennial learner – generally those born after 1982. Ready access to the World Wide Web encourages the comparison of often incomplete and collectively inconsistent information and knowledge seeking, filtering, and synthesizing rather than relying on information from a single authoritative source (Dede 2005). Multiple sources of digital interaction (e.g., iPods, mobile phones, and Internet access) encourage multitasking which reduces focus and often results in superficial learning. The availability of multiple access points has also allowed people to demand individually structured digital products (e.g., tailored MP3 music as opposed to prepackaged albums) including individualized learning, often called personal learning environments (PLE). Today’s learner demands access to information whenever and as often as he or she wishes.

As digital environments grow and change, as digital content expands, and as accessibility grows, educators are presented with considerably more learning opportunities at the same time as learners demand greater personal relevance. There is also a growing demand for a shift from passive delivery methods to more active and collaborative learning, which involves the development of learning communities, placing the burden of learning on the learner rather than on the instructor. This is particularly important for the online learning environment, where the faculty member is not physically present as in a classroom. A major change in the delivery of online courses is the vastly improved learning management systems (LMS) available today.

Principles of online teaching

While all forms of educational delivery share principles, online delivery has some that are very specific to that domain and some of the shared
principles need to be emphasized for online courses. Pelz’s (2004) first principle derives from the constructivist school of thought in that the learner is responsible for his or her learning and, therefore, for doing most of the work of learning. Examples of such work include:

- student-led discussion;
- locating and discussing web resources (web quests);
- peer assistance, whereby students help each other;
- student-to-student assessment;
- case study analysis.

The second of Pelz’s (2004) principles is that, at the center of asynchronous learning, there are three meaningful interactions between learner, content, and instructor. These interactions should also include interactions with the Internet, with the entire class, within small groups and/or teams, and one-to-one with a learning partner. The interactions should include not just course content but also learning about the nature and parameters of assignments, problem-solving, laboratory activities, case studies, and more.

The final principle is that there must be an emphasis on presence. Pelz (2004) suggests three forms of presence for online learning: social presence, cognitive presence, and teaching presence. Social presence, the feeling of being there, allows for the creation of a community of learners, not unlike a community of practice in the professional work environment. There are three forms of social presence:

- affective (expression of emotion, feelings, and mood);
- interactive (evidence of attending, reading, understanding, and contemplating others’ responses);
- cohesive (a sense of belonging, commitment, shared goals, and objectives).

Cognitive presence is the extent to which a community of inquiry and discourse forms to construct meaning. This can be demonstrated by introducing relevant facts, concepts, and theoretical knowledge into the discussion. The value of this depends on the source, clarity, accuracy, and comprehensiveness of the knowledge acquired.

Teaching presence is defined as the facilitation and direction of the cognitive and social processes for the achievement of meaningful and educationally worthwhile learning outcomes. There are two ways in which the educator and students can add a teaching presence to a discussion. The first is direct instruction, such as presenting content and questions, focusing and summarizing discussions, confirming understanding among the learners, diagnosing misperceptions, infusing relevant content from diverse sources, and responding to technical concerns (e.g., file incompatibility). The second component of teaching presence is the facilitation of discussion by identifying areas of agreement and disagreement; seeking to achieve consensus and/or understanding; providing encouragement and reinforcement; setting a climate of learning; encouraging and prompting discussion; and assessing the effectiveness of the process.

New tools for online learning

The primary tool for managing the many components of online courses is the learning management system. The LMS is able to provide the following functions, among others:

- store and deliver content in many different forms;
- facilitate email and other digital communication between learners and educators;
- deliver assessment materials;
- track assignments, student progress, and attendance;
In its early years the LMS was often used primarily for storage and delivery of content, an approach to online education that has often been considered the norm. This approach does not take advantage of the tools variously available in existing LMS like Blackboard, Canvas, Desire2Learn, Moodle, and many others. It has also often contributed to a negative perception of online courses. A particular complaint of such unidirectional courses is that the instructor is not present and doesn’t interact with the learner.

Fortunately the use of LMS and non-LMS social media, collaboration, and multimedia software is becoming more common as a result of efforts by educational institutions to provide training in both the tools and the theoretical underpinnings of online teaching (Solem 1991). While stored course content can include text-based materials, audio files, and instructional videos (including narrated lectures and other podcasts), the material doesn’t in itself provide interaction with the instructor or the other students. Fortunately, most modern LMSs provide chat rooms for group discussions that some geography instructors use for projects, and discussion channels that enable learners to consider difficult questions posed by the instructor and to share ideas with fellow learners as well as with the instructor. The chat room, and especially the discussion channel, allow for original posts and for responses (replies) to be posted asynchronously so that learners have control over when they participate.

These tools allow the instructor to monitor conversations and to intervene when the students wander off topic or reach conclusions inconsistent with the evidence. Some instructors use wiki (e.g., PBworks) and blog (e.g., Blogger) sites outside of the LMS for convenience and because of the advanced features of these commercial tools. One tool that has been increasingly used for group discussions is VoiceThread, which provides text, voice, and recorded webcam video interactions rather than being limited to text. One advantage of this tool is that it allows learners to ask questions and add comments directly on narrated PowerPoint-style lectures. It also allows participants to actively draw on visuals for emphasis.

Some LMSs incorporate webinar tools directly, while others support external webinar systems such as Adobe Connect, GoToMeeting, BigBlueButton, and Meeting Burner. For the educators and the learners who are more comfortable with synchronous interactions, which are more like face-to-face lectures, the webinar allows for the sharing of visuals, online real-time lectures, and question-and-answer sessions during a preselected time period. Additionally, the entire webinar can be recorded both as a means of study and recall and for those who are unable to attend the scheduled event. Such webinar software is becoming increasingly sophisticated and available through educational site licenses.

Special circumstances of geography education

Impediments

Unlike some subjects, geography is both vast in its content and varied in the methods normally applied both to research and to the teaching. Conceptual material is relatively easy to deliver online via readings, either digital or analog, but laboratory and field-related courses present serious problems for delivery online. Questions arise
about how to gather weather information for a physical geography or meteorology course or how to perform rock and mineral identification in an online environment. How would field course-related interactions or a problem-based learning environment be delivered online? Or, more generally, how can almost any experiential course content be delivered online?

While the solutions to these circumstances may require some creative thinking, there are solutions, especially if one moves from a purely online to a blended course, in which some of the activities are performed offline. To use the meteorological exercise as one example, the learners may be provided with the necessary instruments, either as needed or as part of a course kit, and given a time frame in which to gather the data and record them. In a group project, the data could be shared using the Google Sheets software, which allows for multiple users to enter and delete data from a single spreadsheet. The spreadsheet can also be downloaded and converted as an Excel file if needed.

The Google spreadsheet is part of a suite of online collaborative tools that can enhance group projects. Another common tool provided by Google is Google Docs, which allows collaborative work on a single document online. This is an easy solution to the development of group reports resulting from the collection of any data, whether of weather, soil, or surveys.

Field data can also take advantage of digital technology. For example, most of today’s mobile phones have on-board Global Position System (GPS), which allows for the recording of locational information, alongside photographic evidence of a site via geotagging. This technology can be used in a field trip as part of a blended course to demonstrate tagging in a neighborhood, the condition of buildings or bridges, or the visual appearance of a soil profile or geomorphological feature. These digital data can easily be uploaded to the LMS either as part of a data repository or as an integral component of a final digital report. Beyond the LMS itself there are numerous online repositories such as Dropbox, Google Drive, OneDrive, and iCloud, which provide means of sharing content from learner to learner.

Another difficulty of many geography team projects is that they may consist of multiple tasks such as data collection, data compilation, project write-up, not to mention the need for scheduled meetings, whether they be online via a chat room or face to face in a blended environment. Coordinating these roles is difficult enough in a face-to-face environment, let alone in an online setting where many interactions are asynchronous. Fortunately, technology again comes to the aid of the educator with collaboration tools such as Trello or Bitrix24, which provide methods of scheduling tasks, managing projects, and tracking progress for each individual collaborator in the context of the larger project. These tools, like many others, have professional versions that can be prohibitively expensive for educators, but also have free or low-cost alternatives, most of which are sufficient for the online classroom.

Laboratory exercises such as the rock and mineral identification example pose problems in the purely online environment but have ready solutions in a blended environment. One online approach would involve the use of video-conferencing tools such as Skype, which would allow a laboratory instructor to monitor the work of several learners working with their samples and testing equipment. This can be done in real time and would allow for the student-to-student interaction necessary for a robust online learning experience.

Digital laboratory exercises such as those encountered in geographic information system (GIS) classes can be managed in one of
two ways. Step-by-step laboratory exercises, designed for personal training by the software vendors themselves, can be used in a completely online environment. The assignment due dates are scheduled and the learner returns the content within the LMS for grading and feedback. Alternatively, in a blended environment, the laboratory portion of the course can be delivered through a traditional face-to-face component of the course, while lecture, discussion, and other substantive content can be completely online. For learners who cannot attend the face-to-face laboratory sessions because of scheduling or geographic separation, the laboratory instructor can set aside time for the use of Skype or another video-conferencing tool as a consulting period for the learner. In some cases part of the consulting period takes place during the laboratory time slot itself. This allows learners who are separated by distance but who are otherwise free during that time to work on their laboratories with the lab instructor present. When difficulties arise, the learner is able to share his or her screen with the instructor, thus allowing the instructor to spot mistakes or procedural problems much as they would in a classroom.

Field excursions in which the instructor shows the learners what features look like are invaluable experiences for sensitizing the learner to visualize the landscape. While the face-to-face version of this is preferable, the experience can be simulated through the use of video. A good example is to actually take either teaching assistants or students on a site visit that would normally be visited by the class. The characteristics of the site can easily be recorded using relatively inexpensive digital video recorders, and the assistants and/or student aides can be given appropriate scripted questions to pose to the instructor.

Another approach to such field excursions is the use of live-feed video, which is also quite inexpensive. Action cameras can be set to live-feed mode so that learners can visit the site via an Internet connection. This approach could also include real-time audio connectivity, allowing learners to ask their own questions as if they were actually in the field. The live excursions could also be recorded so that learners have access to the experience at any time. This is an enhancement of the face-to-face field experience, which is often a one-time event and is usually recorded via pictures and notes.

Opportunities

The combination of advances in learning technology, social media, multimedia, and geospatial tools provides not only an opportunity to teach geography effectively online but also experiences that are otherwise not available. One example employs digital globe technology provided by Google Maps and Google Earth. Google has both a normal and an enhanced, professional version of Google Earth product (Patterson 2007). This tool allows educators to develop learning modules based on a bird’s-eye view of anywhere on the Earth. For example, an instructor in geomorphology in an online course offered by a school in California can create an exercise on glacial finger lake processes by using Google Earth to transport the learner to New York. Instructors can develop urban field trips by using Google Street View to take learners along the streets of Chicago, for example. These trips can be recorded, placed in moviemaking software for learners to edit, narrate, and annotate. Not only does this provide a unique opportunity to explore other places, but it also provides an alternative approach for the learner to demonstrate his or her mastery of the material beyond the traditional paper reports.
Another tool developed by Google for educators is Lit Trips. This tool leverages the power of Google Earth to provide a more complete geographical context for a wide variety of literature. Characters’ travels are linked to positions on Google Earth, and the actual terrain they traverse is displayed in planimetric view or as street view, if available. This approach is currently popular among educators of young children. Some educators require their students to create the story maps themselves, giving them an opportunity to use geospatial tools in both geography and nongeography courses.

A similar but more robust tool for linking geographic position to experiences, landforms, and conditions around the Earth is the story map recently developed by Esri as an outgrowth of its ArcGIS Online environment. Using prefabricated templates, users can develop stories based on portions of the Earth in which they are interested. Story maps allow learners and/or educators to leverage real GIS data and combine it with text descriptions, pictures (with or without annotations), movies, and other media to create maps with advanced functionality. The templates themselves provide a wide variety of display types such as tours, journals, shortlists, countdowns, playlists, tabs, side accordion, swipes (sliders), and spyglass. In addition to some minor programming, custom designs for story maps can also be created. Such a wide variety of options provides for creativity both in presentation and in assignment creation.

Today’s handheld devices provide opportunities for more creativity, whether in online, blended, or face-to-face environments. Tools such as Google Goggles provide some amazing opportunities for experiencing enhanced reality of observed landscapes. Ritter (2012) pointed out how the use of this enhanced virtual reality too can be effectively employed in online physical geography courses. Google has already released glasses with this technology, bypassing the need for a mobile phone app to perform such tasks. As these and newer technologies develop, innovative geographic educators will find unique ways to employ them in their classes.

Challenges and research needs

As with any new technology applied to learning, it is important to recognize that the technology can become compelling in and of itself and may not be the best, or even a viable, approach to learning. Little research has addressed whether proof-of-concept studies for the use of online learning environments in education are more a function of the nature of the learners or of the instructor. The educators who attempt to develop virtual learning environments are often either digital natives with gaming experience or digital adaptives, those who have taken the time to adapt to the world to which their learners will be exposed during such educational experiences.

Even more so than in education in general, studies linking the sciences of cognitive psychology and neurophysiology to learning in online and blended classes are lacking. Literature identifying how learners begin to identify with their avatars and how they develop social presence is required to provide the hard evidence needed to make bold statements about the relationship between social presence and learning outcomes. User interfaces for virtual worlds need to be simplified, particularly in the area of in-game building tools which, while useful, are difficult to learn and detract from the focused learning objectives.

There is very little research quantifying the degree to which established learning objectives are obtained via online or blended learning.
experiences. Controlled experiments comparing learners using online and blended versus face-to-face learning environments are needed to establish its ultimate utility. Case studies using different control groups would also be invaluable to the literature. Such studies may prove to be the needed impetus for a renewed focus on the scholarship of geography learning.

**SEE ALSO:** Geographic information science and technology: educational directions; Geography education, workforce trends, twenty-first-century skills, and geographical capabilities; Geography in higher education

**References**


**Further reading**


Geography education: primary and secondary

Joseph P. Stoltman
*Western Michigan University, USA*

Geography in primary, and later in secondary, education became established in many countries and regions of the world during the eighteenth century. Colonists migrating to new places sometimes viewed the study of geography as a means of maintaining an attachment to the home country as well as preparing for the challenges of living in unfamiliar places and environments. Information about distant places was presented as geography lessons in school. Vivid geographic descriptions of new settlements in distant places such as Asia, Africa, and Australia beckoned potential immigrants to land, sunshine, and prosperity. The extension of geography in primary and secondary education coincided with the era of exploration and globalization as distant places became linked by migration, trade, and politics. Immigrant populations were on the move, and the geography of the destination and its environmental conditions often attracted potential settlers. The study of geography continues in the twenty-first century with the exploration of contemporary conditions and issues on Earth. Just as earlier generations learned about the Earth through storytelling, print materials, photographs, and film, the current generation has the opportunity to study geography using the Internet, the World Wide Web, and virtual field trips to learn about the Earth, its people, and its environments.

Geography education and its importance to society

Geography is an academic discipline and a school subject based on established traditions and emerging opportunities to study the Earth and its people. Geographic education includes the perspectives, concepts, and skills that professional geographers use in a form adapted for people studying geography in both formal and informal settings.

Geographic education is both studied and applied by all twenty-first-century societies. For example, when a farmer in West Africa uses a mobile phone to research the market price for the groundnuts that will be sold to a local commodity broker at harvest time, he or she is acquiring geographic information. Geography may also be used to predict the effects of a desert locust outbreak on groundnut production in a nearby country, which may consequently influence groundnut pricing. In each case, the farmer is either using geography to make decisions regarding pricing or collecting geographic information regarding supply and demand. In both cases, geography education is being applied to make personal decisions locally, while using global information about conditions in other places. Geography education also serves to mitigate disasters, such as the deadly outbreak of ebola in West Africa in 2014. Medical treatment and fatality data were converted to spatial information and mapped. The maps enabled health workers and local community leaders to alert people in the affected region regarding the geographical spread of the disease. Procedures were taken...
to establish geographical zones where travel and personal contact were greatly limited, thus stemming the spread of the disease. Geography education is applied each day by every one of us, whether it relates to a local journey to the market or an air journey halfway around the world. Geography education provides the information and skills that individuals and societies must know and apply in a dynamic world. Geography education has played that role for a very long time.

Geography education: the early years

The early years of primary and secondary geography education were similar in many countries during the period of colonial influence. There were mainly two types of colonization that affected education. The first is the one most readily recognized as a colonizing country taking economic and political control of a place and introducing social norms and values. This included the introduction of an educational system. Colonial America was a good example of this process. The second form of colonization was the diffusion of ideas about geography in primary and secondary education but without the exercise of colonial control; this is often described as cultural colonialism. Examples of cultural colonialism include the educational philosophies of Jean-Jacques Rousseau (1712–1778) in Europe, and later John Dewey (1859–1952) in the United States, which were diffused internationally through writing and research. Both philosophers influenced educational practices relative to the view of nature and the role of the environment in education. In his analysis of American geography and how it was influenced by its European roots, William Warntz (1964) reflects on the role of educational colonialism. The writing of textbooks and atlases for school and home became a major educational influence in colonial and later post-colonial countries. Published materials further signaled the diffusion of geography through schools. Geography was viewed by educators as a good preparation for the development of reasoning and for the study of natural and moral philosophy.

The memorization of geographic names and locations has often dominated the study of geography. While the memorization of geographical names may be less demanding intellectually, place names are considered important for students to develop a worldview. Developing a proficiency in geography requires some memorization, map work, and cumulative experiences in the study of regions of the world at different scales. Developing a worldview has become a standard goal of geography in primary and secondary education, which persists globally in the twenty-first century. The discipline serves as a spatial platform for the study of economic, social, and political changes that affect places and regions. Those changes in spatial patterns of trade, population, and territorial claims introduce students to the dynamics of a twenty-first-century worldview.

Early teaching of geography

In its early years geography instruction consisted mainly of the study of maps, agricultural and industrial production, and physical geography, including weather and climate. It was reported to be bland and not very motivating. Methodologies incorporating the educational theories and philosophies of Johann Pestalozzi, a Swiss educator, and Arnold Guyot, a Swiss geographer, found ready acceptance in the teaching of geography. Both became influential and affected
the teaching of geography in Europe and North America. Pestalozzi promoted the philosophy that direct observation and sense perception are essential for meaningful learning by children. What better subject than geography in which to implement that pedagogical approach? Maps and line drawings became essential components of textbooks and other instructional materials within educational systems in many countries. Drawing maps of their home region and completing field studies became widely accepted ways to teach geography (Dahmann 2011). Geography proved to be an ideal subject for using information from photographs, black line graphics, and other visuals to enhance student learning. Graphics in early textbooks were the initial version of geospatial technology used in teaching.

Arnold Guyot was educated as a geographer in Switzerland and arrived in the United States at a time when state-supported secondary schools were being established. In Europe, Guyot had become familiar with the pedagogical practices of Pestalozzi. He had considerable success using field studies to build observational skills among his students and as a means to collect geographical information. He achieved considerable fame in American geography as an author of secondary school and college textbooks. Secondary school physical geography prospered under his influence.

By the start of the twentieth century, primary and secondary geography education had matured in content, scope and topics, teaching methodologies, and philosophical underpinnings. The technology of the time permitted the reproduction of diagrams, maps, textbooks, and wall maps, which were combined with direct observation to engage students in geographic content. Teacher-led recitations of content were enhanced by visuals and increased information about the world, which led to the development among students of a more comprehensive worldview. Physical geography was widely accepted as an important scientific study of the Earth and assumed a prominent position within the curriculum.

The philosophy of education that was to dominate the early part of the twentieth century, the progressive education movement, looked favorably on geography education. Local study, developing a sense of geographic place and space, and discovery learning complemented geographic study. Moreover, Rousseau and Pestalozzi became the dominant influences as, in many parts of the world, the curriculum engaged students in local study in the primary grades, gradually extending to the geography of the world in secondary school. This was the initial indication of a curriculum movement that later became known as the expanding environment.

Academic and professional societies and geography education

The twentieth century was a period when professional and academic organizations assumed a greater role in geography education in the United States and elsewhere. In the United States, the National Education Association (founded in 1857) completed major reviews of school geography in the 1890s. Reviews also took place in other countries, where they were usually sponsored by academic or professional societies. Scholarly societies addressed the role of geography education, the content most suitable for primary and secondary students, and what should be taught to students of different ages and grades. Among the most influential of the professional societies of geographers were Société de Géographie (Geographical Society, 1821), the Royal Geographical Society (1859), the American Geographical Society (1851),
and the National Geographical Society (1888). While the societies were academic in their charters, they were broadly interested in promoting geography and in applying developing technologies in cartography, photography, and publishing in their mission to educate the public. Those missions, however, strongly influenced primary and secondary education, since students represented inquiring minds about the world and were potential future supporters of geography, exploration, and education.

The interest of scientific societies in promoting geography education prompted teachers to become interested in establishing organizations specifically devoted to geography in primary and secondary education. For example, the Association of American Geographers (1904) had among its members geographers who had written textbooks, served on national committees, and greatly influenced the teaching of geography. Educational geographers moved forward to establish the National Council for Geographic Education (1915). The initial mission of the council continues today to support geography teaching through the publication of articles and reports on content, teaching methodologies, and research. Similar organizations of geography teachers began in other countries: Geographical Association (United Kingdom), Associazione Italiana Insegnanti Geografia, Turkish Geography Teachers Association, Korean Association for Geographic and Environmental Education, Australian Geography Teachers Association, Geography Teachers’ Association of Singapore, and the Verband Deutscher Schulgeographen (Association of German Geography Teachers), to name just few. Nearly every country has either a formally or informally organized group of geography teachers in primary and secondary education who represent the voice of geography education.

The International Geographical Union (IGU) is the largest international organization of geographers, representing the discipline in more than 150 countries. Membership in the IGU is largely by national units, with the provision for membership by individuals if membership is not held by a national research organization that represents a country. The IGU sponsors a Commission on Geography Education (IGU-CGE) which communicates information and provides expert advice to national organizations of geography teachers. The commission holds regular conferences in different countries, hosted by a local university or geography teachers’ association. The conferences include research paper presentations devoted to classroom teaching, curriculum development, geographic literacy, assessment in geography, and theoretical treatises on the subject. The commission distributes an electronic newsletter and has regional representatives in many countries responsible for coordinating national participation.

Research, curriculum reform movements, and global impacts

Curriculum is the driving force in determining the importance of and position of geography within a country. The curriculum prescribes the content and skills that are taught and when they are taught in primary and secondary schools. It is perhaps the most important element in the preparation of teachers. Teachers are expected to align the content they learn in their academic preparation as professionals with the curriculum they are expected to teach.

Research in geography education for primary and secondary education has not regularly addressed curriculum questions, which have often been left to educational researchers.
Initially, educational researchers were not specifically interested in geography content, but rather in questions such as: How do children learn? What materials attract the attention of children and promote thinking in specific ways? Is there a sequence of learning that advantages students as they progress from early to later years in school? Those questions remain familiar, for they have persisted into the twenty-first century and continue to be addressed by contemporary researchers. However, the difference is that, while some of the early twentieth-century research was completed while observing students in learning environments, much of the research was philosophical in nature, based on conventional wisdom regarding primary and secondary students. The research was often in the quest for grand ideas about learning. Those philosophical treatises were then transformed into curricula and the learning progressions that were most appropriate in elementary and secondary schools. It was the development of curriculum and sequencing the curriculum for optimum learning that engaged geography education.

Curriculum development and curriculum change have been important topics within educational research for the past half century. Curriculum change is the product of dynamic societal expectations or demands on the educational system. Discussions regarding reforms in geography education and the implementation of reforms began in earnest in the latter part of the twentieth century. There was little research on the need for, or the type of reform that would provide, geography education with a proactive position within the curriculum. The result was that geography education has been a product of reform rather than an advocate with the potential to shape its role within the curriculum through reform.

Curriculum decisions are made at either state or national level, depending on the country being reviewed. For example, England has a national curriculum, whereas in the United States and Germany state and lander authority, respectively, determine the curriculum. In all instances, the curriculum is expected to meet the demands of a country’s state or national ministry, department of education, or another national legislative or governmental body.

Numerous national curriculum reform movements have impacted geography education. Perhaps the most monumental was the transition in the national curriculum that occurred in African countries following independence for Ghana in 1957. The primary and secondary educational systems during the colonial period were replicas of the curriculum in the colonizing country. This was the case not only with European countries but also with Japan’s colonial rule of Korea and with US colonial rule in the Philippines. In the former British colonies, students who were fortunate to complete primary and secondary school were administered either the Cambridge or Oxford Sixth Form Examination. That practice continued well into the late twentieth century, when new African-based examinations were introduced, for example the West African Senior School Certificate Examination (WASSCE), which initiated a transition from a Eurocentric to an African curriculum. Today students across much of Africa take examinations based on their own postcolonial curriculum models. The Kenya National Examinations Council provides countrywide assessments as a means to ensure conformity to Kenya’s national goals and governmental policies relating to the curriculum and examinations. Attention is focused on a national or regional curriculum and examinations are aligned with the curriculum. In the twenty-first century African students examine the urban structure of not just London, as they had done for the Oxford and Cambridge examinations, but also
Lagos, Nairobi, and other African cities. In both West Africa and the Kenyan examples, the examination drives the curriculum. Research on the success of the geography curricula and the adjustments needed for changes in the discipline are continually being pursued by African geography educators in the twenty-first century.

In the United States, earlier in the twentieth century, a major reform in education began that would eventually have considerable influence on the role of geography education, both nationally and internationally. This was the social studies movement, whose influence continues into the twenty-first century. The basis for the movement began in the late nineteenth century in the United States and gained considerable attention from educators in the early years of the twentieth century (James 1969). Those far-reaching effects began with educational policy recommendations by the US National Education Association (NEA) regarding geography in the curriculum. NEA recommendations favored a curriculum organized as social studies. The social studies provided a new structure for adding subjects to what had been principally a geography and history curriculum. Social studies was intended to include geography, history, government, economics, and other subjects that were concerned with the social development of citizens and with content that had societal applications. Social studies represented a major curriculum reform. The geography curriculum that was taught as a stand-alone series of courses beginning in primary school and extending through secondary school would, in time, become integrated within the social studies curriculum.

Two additional circumstances at the time weakened the role of primary and secondary school geography. First, there was disagreement between geographers regarding the roles for physical, human, or regional geography as the curricular focus. Second, geographers did not participate in the committee discussions that formulated the social studies curriculum (James 1969) because physical geographers did not regard geography as a study of social issues. The lack of participation in the policy recommendations disengaged geographers from the subsequent educational decisions that established social studies permanently. Geography education had a place in the social studies curriculum design, but few geographers were willing to help transition the subject into the new curriculum. However, changes were underway and over the six decades following the NEA policy recommendations, geography education gradually became integrated into the social studies curriculum in the United States. In the twenty-first century, students may study geography, but the discipline is a component of social studies. This diminished role of geography in education continued during the remainder of the twentieth century. The discipline assumed a less prominent role in the curriculum, in terms of both disciplinary name and content.

The social studies reform movement of the twentieth century has had international implications for geography education. In many countries geography was initially taught as a stand-alone course in primary and secondary education. However, as the social studies curriculum diffused to other countries, the preference for an integrated series of topics that drew on the disciplines became the norm. Social studies’ major objectives included curricular concepts such as civic development and citizenship. In practice, social studies served a purpose that neither geography nor history provided for newly independent countries. It was the development of citizens through school, with the eventual outcome being an informed, responsible citizenry.
Attention to social studies varied between democratic and totalitarian systems of government, and reflected different priorities between different countries within each category of governmental structure. Over time, social studies has become the preferred curriculum choice among educators and in both primary and secondary schools. Internationally, the curriculum is called social studies in primary school, but in secondary school geography sometimes remains clearly defined as a subject. The 2014 curriculum change in Australia is a case in point. Formerly social studies had been designated for all grades, but the national government reinstated a specialized geography course in the high school.

Internationally, the social studies expanding environment curriculum model, or contextual derivations of it, has achieved widespread acceptance. Few countries depart significantly from its basic structure (Box 1), although the terminology used for the topics studied in the curricular sequence may vary. The expanding horizons model has been the curriculum most consistently applied to primary education. It dominates social studies curricular frameworks in textbooks, national content standards frameworks, and national examination formats.

While the expanding environment has been prominent in primary schools internationally, it has not in secondary schools. In most countries, geographic content is taught in courses referred to as “geography” in the curriculum. When asked, students say that they are studying geography. In the United States, courses with substantial geographic content are often titled “global studies,” “world studies,” “cultural studies,” or “contemporary issues.” The major exception in the twenty-first century is Advanced Placement Human Geography (APHG) and the International Baccalaureate (IB) in geography, both of which are rigorous and content- and skills-intense courses in high schools.

The expanding environment, while widely used, is not without critics. A major criticism is based on the greater opportunities for primary and secondary students to extend their interests to more distant geographic locations and prior historical eras by using twenty-first-century technology. The globalized world is often cited as another reason the expanding environments model is outmoded. These criticisms are difficult to support because the geographical study of nearby spaces precedes and links to that of more distant and abstract spatial topics. While the expanding environment curriculum appeared to offer little flexibility, there was considerable opportunity to link the local environment to national and global topics in the early years of schooling, and to link national and international topics to the local environment in the later years. There is still little research on the alignment of the expanding environments curriculum with child development or the manner in which social studies and geography learning occurs. Within primary education, research is needed on learning progression, concept development, and the understanding of basic and complex content. Such research will provide important information for the design, structure, and content of the curriculum in both geography and social studies.

The curriculum remains a critical component in determining how much time is allocated to geography education during the school year, and what content from the discipline is emphasized. Curriculum provides the access for including geography within the educational experiences of students. For example, there was little attention to curriculum structure as it was applied to geography education in the second half of the twentieth century. In 1994 the United States published the National Geography Standards in response to a national educational reform.
Box 1 The expanding environment/horizons curriculum model.

Topics in the curriculum are quite consistent for grades 1 to 5 in many countries, but they vary in secondary schools. High school courses generally follow social science disciplines, often embedded in the social studies curriculum. Political science content is often labeled “government” or “civics.”

<table>
<thead>
<tr>
<th>Kindergarten</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
<th>Grade 5</th>
<th>Grade 6</th>
<th>Grade 7</th>
<th>Grade 8</th>
<th>Grades 9–12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self and others</td>
<td>Home and family</td>
<td>School and community</td>
<td>Regions and cities</td>
<td>The state</td>
<td>The nation</td>
<td>The Western Hemisphere, world geography, or world history</td>
<td>World geography, Eastern Hemisphere, or world history</td>
<td>Ancient civilizations</td>
<td>Disciplinary content includes history, geography, economics, psychology, anthropology, sociology, and political science. Examples of stand-alone courses are Advanced Placement Human Geography in the United States, A-level geography in the United Kingdom, Abitur-level geography in Germany, and the International Baccalaureate in geography, with each being rigorous.</td>
</tr>
</tbody>
</table>

in many subjects. The intent was to make the subjects studied by students more rigorous (the popular term used was “raising the bar for content proficiency,” comparable to raising the bar for a high jumper in track and field). However, there was little corresponding research-based curriculum information available to inform the standards-building process. Research in curriculum development is necessary in sequencing topics in geography and in social studies, in the learning progressions that students respond to in sequencing content, and in identifying the ways a student should demonstrate proficiency in knowing and applying knowledge. The curriculum must be a well-developed plan for presenting the study of geography for students.

Curriculum research and planning is a complex matter. A continuing issue that faces geography education in the twenty-first century is the discipline’s future role and placement within the curriculum in different countries. For many decades, geography education filled two major functions in the education of primary and secondary students. First, the discipline’s content is diverse, and includes both human and physical study of the Earth. The discipline is dynamic and changes with new technologies, new ways of thinking about the world, and new ways of researching the Earth and its inhabitants from a geographic perspective. Second, different societies have different expectations for geography education, although one seems to dominate. The societal view of geography by policymakers and teachers is largely that of developing a worldview that enables students to become informed about the countries of the world and the ways in which
they are either linked or separated by contemporary issues. In part, that expectation is tied to the geography education of the past, when knowledge of the world was the main concern of geography. At present, modern geography education is dependent on the availability of geospatial information and the tools of mapping and graphing those data to solve problems and make decisions. In the twenty-first century teachers, students, policymakers, and the public regard geography as providing the content, skills, and perspectives essential to make sense of the world. What is needed is a widely recognized research basis for how an understanding of the world develops progressively in primary and secondary students (Lambert and Jones 2013; Butt 2015).

International initiatives in geography education

The United Nations, founded in 1945, took an early interest in geography education. While there were international conferences of geographers prior to 1945, the United Nations claimed the role of geography education was to enhance international understanding. This was partly as a result of the important role geography played in establishing peacetime territorial agreements at the end of World War II. The United Nations also took the view that, if young children learned about the world and its human and physical geography, there would be greater international understanding. The educational task of the United Nations was delegated to the United Nations Educational, Scientific and Cultural Organization (UNESCO) in 1945. UNESCO assumed responsibility for nearly all educational programs on behalf of the United Nations. From the very beginning, geography education was viewed by UNESCO leaders as a necessary subject in primary and secondary education. Collaboration between the International Geographical Union Commission on Geography Education (IGU-CGE) and UNESCO led to the publication of two major international handbooks on the teaching of geography in the 1960s and 1970s. The books were designed to inform teachers in developing countries regarding the most recent and effective practices in geography education.

The interest of UNESCO in the growing international popularity of the social studies curriculum movement had spread from the United States to other countries by the 1990s. The international focus on social studies gained traction with a newly published handbook on the teaching of social studies in primary and secondary schools (Mehlenger 1981). Each chapter of the book was devoted to a subject or method that applied to the social studies. A single chapter was devoted to geography. The sourcebook was intended to provide a menu for interdisciplinary or multidisciplinary contributions to curriculum and teaching. It was widely distributed to developing countries through UNESCO. In developing countries, this often marked a transition from geography as a stand-alone subject to its inclusion as part of the social sciences and humanities content that constituted the social studies curriculum. A central focus of the sourcebook for social studies was the importance of preparing students for citizenship in their countries and the essential content and skills to meet the demands of twentieth- and twenty-first-century societies.

An additional handbook on geography was published in the early 2000s. It was sponsored not by UNESCO but by the IGU-CGE, and distributed mainly through the network of the Commission on Geography Education and direct sales by the publisher. Entitled the International Handbook on Geographical Education (Gerber
2003), it was organized in much the same way as each of the prior handbooks were composed. Geography experts in different regions of the world were invited to contribute to a planned table of contents that represented critical issues in the teaching of geography at the time. While topics such as geographical skills persisted through each of the handbooks, the applications of classroom methodologies changed. For example, the classroom technology made the transition from the use of wall maps and chalkboard visuals in the first sourcebook to the use of computer-based and Internet graphics in the final handbook. These three books represent an important extension of geography education information to teachers primarily in developing countries from the 1960s to the beginning of the twenty-first century.

Similar to handbooks have been the specialized methods books devoted to the teaching of geography that have been published, but they are not common. In the United States there is just one current teaching methods book devoted specifically to geography education (Gersmehl 2014). In the United Kingdom the major geography methods book is published by the Geographical Association, the professional society for geography educators (Balderstone 2006). Similar geography teaching methods are published within the contexts of specific countries, such as India, Australia, Germany, France, and Singapore.

Geospatial technology, World Wide Web, and teaching geography

During the twentieth century, perhaps the most important source for practical classroom suggestions regarding teaching was a printed geography instructional methods book that demonstrated ways to teach content and skills. Methods publications in print form have largely been replaced by materials that may be accessed on the World Wide Web, either as electronic books or as suggestions for teaching that are contributed by other teachers to websites. Both professional societies and commercial publishers have engaged in providing methods materials for preservice geography education students and classroom teachers engaged in professional development. The Annenberg Learner project entitled “Teaching Geography” (www.learner.org/workshops/geography) provides professional development for teachers as well as curriculum materials that can be used in the classroom. National Geographic Education provides classroom materials in human and physical geography for download from its website (http://education.nationalgeographic.com/education/?ar_a=1).

The advantage of the World Wide Web for professional materials in geography education is its global accessibility to teachers and students in primary and secondary schools. Search engines list many hundreds of lessons, reference materials, and scholarly articles for keywords such as “geography methods,” “teaching geography,” “geography skills,” and “teaching with geographic information.” The materials available on the web must be evaluated carefully prior to use, since they vary greatly in content accuracy and teaching methodology. Websites often include basic suggestions for teachers and materials for students, as well as lessons that advance particular methodologies such as inquiry, problem-solving, and decision-making as a component of geography instruction.

The introduction of computers in the latter decades of the twentieth century to geography education marked a major transition in the availability of reference materials and resources for students in primary and secondary schools. Geography through virtual field trips and collections of photographs, and access to data about
the world in electronic format, were early advantages. An early use of computers in primary and secondary geography education was to search and view maps that would normally be available only in specialized map collections. Basic maps could also be printed. Initially printed maps were used mainly for labeling countries, cities, and topographical features. As geography instruction began to rely on statistics that reflected climatic, population, and economic data, the nature of map study as a teaching method changed. Students were able to use electronic maps to show the spatial patterns of population change, soils, climate, and the economic conditions of local communities as well as to compare their country or region with other places. Most of the early maps were reproductions of printed maps in electronic form. In electronic format, the spatial patterns on several different maps could be compared visually to search for similarities and differences that helped explain the geography of the place.

Since maps are the geographer’s most widely used means to display geographic information, the capacity of computers to produce and deliver maps continued to evolve. With the advent of the World Wide Web, high-resolution electronic maps of both historical and contemporary topics became available. They could be viewed electronically and saved as electronic files to use for future work on projects. Lessons were developed to teach students how to use electronic maps. Many of these lessons are available on geography education websites to new generations of teachers and students. Electronic maps provided greater opportunities to interpret and explain the spatial patterns relating to numerous topics in physical and human geography.

In the twenty-first century, geographic information science (GIScience) is one of the most rapidly growing applications of academic geographic research. GIScience has also stimulated an important transition in the teaching of geography, perhaps the most important invention since the printing of the first geography textbook. It permits mapping, the use of photographs, the streaming of video from handheld devices, and the use of geographic data to address geographic issues and problems from the local to the global scales. The availability of web-based GIScience has made it possible for primary and secondary students to access geographic information and mapping software with only an Internet connection and a computer or handheld device. GIScience information may be shared with similarly aged students or academic experts to prepare research reports and data collected in the field. During the twenty-first century, students use web-based mapping in their research and class projects in local schools, as well as participate in research challenges and contests with international groups of students to focus on topics or issues of significance. The continuing potential applications of GIScience for use in primary and secondary education increase with improvements in the software. Whereas considerable expertise in GIScience was once required to produce a map, the web-based mapping programs of the twenty-first century have become available to the general public, including schools, at little or no cost for the right to use the software.

The early twenty-first century also saw initiatives to encourage the use of GIScience in primary and secondary education. The first, Esri, is the internationally recognized leader in the development of GIScience. The Esri leadership viewed the wider use of web-based mapping in schools and by the general public as a corporate responsibility. Esri has developed web-based mapping software for education that does not require a large amount of computing power but, rather, relies on the capacity of the web to process data that can be used by students.
The result is its continued support and provision of software for educational purposes, and its licensing of software use to many thousands of schools. Its web-based mapping software and operations are less complex, are menu-driven, and include datasets that can be selected as mapped layers for the overlaying and analysis of spatial patterns.

The second initiative came from the National Geographic Society (NGS), a major provider of geography teaching materials and teacher professional development. NGS software was developed to introduce students to the use of the web in mapping. Called MapMaker, it applies basic mapping skills, using scale, geographic patterns, and drop-down menus, to place symbols on maps (http://mapmaker.education.nationalgeographic.com/?ar_a=1&b=1&ls=000000000000). The maps can be viewed on screen and printed, and several types of information can be compared and analyzed at a very basic level of complexity. It is web-based mapping specifically designed for use in primary and secondary geography education.

GIScience is one example of how geospatial technology is reshaping primary and secondary geography and social studies teaching in the twenty-first century (Milson, Demirci, and Kerski 2012). While more sophisticated applications are used in secondary schools, there are instances where younger students are using GIScience, mapping, and satellite images on websites. They may use geospatial devices such as computers, tablets, smartphones, and so on to display geographic information: for example, for a bird’s-eye view of the school and nearby street patterns. In secondary schools, such devices are used to analyze data and compare and contrast conditions, and to explain why a change in one geographic pattern, such as land use, can result in a change in a second pattern, such as vehicular, pedestrian, and bicycle traffic. Geography and social studies teachers in the twenty-first century must have a good grasp of geospatial technology to address local and global issues. Students should learn the benefits and advantages of using geospatial technology both inside and outside of school.

Geospatial technologies are common components of modern devices used for communications between individuals. When they are used to address geographic problems and aid decision-making, these devices have considerable relevance to geographic education. Geography educators consider geospatial technology as a means to enhance learning in geography. Students often demonstrate an interest in using technology in many different and creative ways. The challenge for the future is to make the use of geospatial technology and the content of geography as seamless and engaging as possible. Geospatial technology may be applied widely across societies to address problems and to formulate solutions relating to the world in which these students live (Sinton et al. 2014).

Geography education research in the twenty-first century

The major investigation topics in geography education prior to the twenty-first century were on the effectiveness of teaching methodologies and student learning of geography, with some attention to curriculum development (Lambert 2015). Geography education research related to thinking and learning received less attention during the same period. However, it became increasingly apparent that ways of learning and thinking about the world were an important item on the research agenda (Lambert 2010). Prior research in geography education was based largely on research in education and on the educational aspects of geography. What was needed,
and which steadily materialized during the final decade of the twentieth century, was an initiative that would integrate geography education and educational research with theoretical positions supported by both geography and education (Seo and Kim 2012; Segall and Helfenbein 2008).

The research topic that gained prominence among geography educators was spatial thinking. The spatial tradition of geographical science served as the gateway to spatial thinking research and the development of geospatial perspectives by students in primary and secondary education. It also served as the logical extension to the spatial thinking research being pursued in other disciplines. For example, spatial thinking had a long research tradition in psychology, especially among developmental psychologists. Spatial thinking and its application to learning were also of interest to educators in mathematics, science, and cognitive studies. Those researchers had demonstrated that spatial thinking was related to perspective-taking, to hierarchies of inclusion whereby students understand that different geographic features fit within larger categories and that linear and area measurements required spatial transformations. Those topics were deeply embedded in the spatial nature of geography as a discipline but had not been highlighted for research in geography education. By the 1990s, research reports began to address spatial thinking in geography, written by both geographers and psychologists. The publication of *Learning to Think Spatially* (Geographical Sciences Committee 2006) was a milestone that built a case for the relationship between geography and spatial thinking, as well as highlighting the research potential for geography education. By the 1990s, research reports began to address spatial thinking in geography, written by both geographers and psychologists. The publication of *Learning to Think Spatially* (Geographical Sciences Committee 2006) was a milestone that built a case for the relationship between geography and spatial thinking, as well as highlighting the research potential for geography education. The opportunity to pursue spatial thinking and the development of a geographic perspective gained considerable attention as a research topic. This was due, in part, to the concurrent attention to geospatial technology in the curriculum. While geography education had traditionally viewed spatial learning as a goal, there had never been compelling developments that promoted it as a research topic. GIScience, global positioning systems (GPS), and the use of handheld geospatial devices, such as smartphones and tablets, provided the opportunity to connect spatial thinking and geography education research. Using such devices, students both in primary and secondary classrooms and outside of school were immersed in spatial data and spatial thinking. The connections between the geospatial devices and the spatial attributes of geographical thinking were not readily known. This provided a research opportunity.

The transformations resulting from the geospatial revolution that was underway in geography, and the use of GIS in other subjects, became an inviting research agenda. While there had been reviews of research in geographical education, there had not been a projective discussion of what was needed to advance the research. The four major scholarly and professional societies in geography in the United States joined forces to address the future of geographic education. The project produced *A Road Map for 21st Century Geography Education*, which devoted most of its pages to projecting the directions for the future of geography education (Bednarz, Heffron, and Huynh 2013).

The project concentrated on three topics in geography education that were high priority: (i) research; (ii) instructional materials and professional development; (iii) and assessment (Bednarz, Heffron, and Huynh 2013). While the *Road Map* project initially focused on researchers and future research in the United
States, attention was given to international views on geography education. What the Road Map aimed to accomplish for the United States also resonated with geography educators in other countries. The three priority recommendations of the Road Map project are attracting interest in other countries. The Road Map project presents a big picture, leaving individual researchers and their institutions to develop the capabilities to engage in the research.

As a result of the Road Map, projects are being planned and initiated that involve international collaborative research. The GeoCapabilities project is researching two questions about geography education’s role in society. The first asks how geography education can contribute to the development of human capabilities. The goal is to explore the significance of geography education in preparing students for a successful lifetime career and civic engagement. The second question is the obverse of the first: What are the consequences for young people, in terms of their future potential and wellbeing, of not having access to geography education? The project seeks to identify the disadvantages for individuals and societies if young people are deprived of a geography education. The long-term negative effects on each person’s capacity to play a participatory role as a citizen through a career and civic life have particular relevance in countries where geography education has a low profile in the curriculum. These are big questions that arise from the societal belief that education benefits everyone. Does a geographic education benefit societies in ways that other disciplines cannot? The research may identify the particular role that geography education plays in the attainment of those goals, and the negative consequences of not providing geography education. Among the partners and affiliates of the project are Belgium, England, Finland, Greece, the United States, China, the Czech Republic, Germany, the Netherlands, Portugal, Serbia, Singapore, Sweden, and Turkey. While each country will research GeoCapabilities within its particular national context, the research results will be shared and analyzed to provide a research-based global rationale for teaching geography in primary and secondary schools.

The second project recommended by the Road Map project encompassed international research collaboration as a natural outcome. Learning progressions in geography education have long been a research interest. To date, most beliefs about learning progression are based on conventional wisdom, such as the expanding environments curriculum that is dominant in social studies education. The intent of the research is to determine the progress students make in studying geography compared to the desired standards for information acquisition, conceptual understanding, and geospatial skills applications. The theoretical and practical interests in learning progressions in geography education emerged in the first decade of the twenty-first century (Bennetts 2008). Reports on the progress students make in learning various subjects are reported in the research literature, and Bennetts (2008) used a critical analysis methodology to examine progressions and understanding in geography specifically. He made the case that learning progression research is necessary to align the curriculum to national geography standards by way of students in the classroom who are engaged with the subject. Learning progression is a universal concern in geography education and lends itself well to international collaborative research. Initial research on GeoProgressions as a topic within learning was initiated in the United Kingdom (England and Wales) with the support of the Geographical Association, and was later recommended as a main research topic by the
**Road Map** project in the United States. The major emphasis of the most recent learning progressions research initiative is on teaching early career scholars and researchers about the theoretical and practical attributes of learning progressions. The project to achieve capacity building is sponsored by the American Association of Geographers. Its goal is to develop the critical capacity necessary to design and complete sustained investigations about learning progressions in geography content, understanding, and skills.

### Journals in geography education

Geography education research has been widely diffuse. Much research is the result of doctoral dissertations completed at universities. Dissertation research that is synthesized, rewritten to journal publishing standards, and submitted for publication in research journals becomes accessible to other researchers. The first journal devoted specifically to geography education research was sponsored by the Commission on Geography Education of the International Geographical Union (IGU–CGE). Entitled *International Research in Geographical and Environmental Education* (IRGEE), it has been the main source for publishing research articles on geography education since it was started in 1992. The journal features research articles that employ both qualitative and quantitative methodologies, and accepts for peer review all research topics that are representative of geography and/or environmental education.

A second research-oriented journal is *Research in Geography Education* (RGE), published by the Grosvenor Center for Geographic Education. First published in 1999, RGE has research as its principal focus. It also publishes opinion, philosophical, and historical reflections on the state of geography education. The journal’s wide interpretation of research in geography education results in a broad range of topics.

There are general journals that represent a broader cross-section of geographic education than the research journals. They include suggested teaching methods for elementary and secondary classrooms, examples of online satellite images and photographs that can be printed, and lessons designed to address a particular topic or issue. The journals are sponsored by geography teachers’ organizations in different countries, but may be printed by a major publisher and marketed as printed copies and/or as electronic journals or articles.

The following is a list of print and electronic-access journals for geography education: *Geography* (UK), *Journal of Geography* (USA), *Geographical Education* (Australia), *New Zealand Journal of Geography*, *Journal of Korean Association of Geographic and Environmental Education*, *Geography Teacher* (USA). The following are online journals for geography education: *European Journal of Geography* (international), *J-Reading – Journal of Research and Didactics in Geography* (international), and *RIGEO: Review of International Geographical Education Online*.

### Summary

In the United States, geography in primary and secondary education is a component of the social studies curriculum. In other countries the placement of geography varies, but a common pattern is for geography to be included in elementary social studies, and then taught as a geography content course in the secondary school curriculum. In some countries physical geography is an important component of the science curriculum. Within the social studies curriculum, the
challenge for geography is to attain, and then maintain, an equitable role among the other content subjects, namely history, government, economics, and, in some cases, the humanities. The introduction of geospatial technology in the twenty-first century has provided an opportunity for geography to assume a larger role within the growing technology component of education, both in the United States and internationally. In the United States the science, technology, engineering, and mathematics (STEM) initiatives represent an opportunity for geography and geospatial technology to join the privileged curricular components that are given priority. STEM subjects have increasingly become priorities in many countries, since they are viewed as essential for sustained national economic growth.

In order for students to meet STEM content and skills expectations, a major adjustment in primary and secondary geography education began early in the twenty-first century. First, new cohorts of geography teachers are being prepared, either in their initial teacher education or through professional development, to gain proficiency in using geospatial technology. Second, a scientific basis for geography through STEM will need to draw from the earth science tradition of the discipline. Geospatial technology has opened up numerous opportunities for student engagement, advanced education, and careers. Geographic content and geospatial technology are powerful academic and applied components for primary and secondary students to pursue as they prepare for college, careers, and civic life within their communities. Geography education can play a leading role, with skills and technologies that provide access to spatial information. Spatial information is essential for solving problems, making decisions, and thinking critically about the world in which students live.

SEE ALSO: Geography education, informal and for public engagement; Geography education: national policies and standards; Geography education, primary and secondary: international perspectives; Geography education: promoting diversity and broadening participation; Geography in higher education

References


Gersmehl, Phil. 2014. Teaching Geography. New York: Guilford Press.


Lambert, David. 2015. “Research in Geography Education.” In MasterClass in Geography Education:
Diversity is a multifaceted issue within geographic education, one that brings together social, cultural, political, and ethical questions and concerns. How is diversity conceptualized? How do perspectives and initiatives change over time and place? Who teaches and studies geography? Do curricula and pedagogy reflect commitments to social inclusion, justice, and equity? This entry emphasizes American geography, primarily since the 1960s, but also introduces examples from other settings to illustrate some ways in which concepts, priorities, and approaches may cross boundaries but also differ.

Conceptualizing and measuring diversity

How diversity is understood has varied over time and across different constituencies within geography, reflecting national and local contexts. Much, though not all, attention in the literature has focused on higher education. The range of interpretations is diverse, most commonly dealing with “race,” ethnicity, and gender, but also with national and immigrant origins and status, (dis)abilities, sexualities, indigeneity, and social class. Major attention emerged in the United States during the 1960s as the Civil Rights Movement and federal government policies prompted concerns about the low representation or exclusion of “Blacks” or “African Americans” within higher education. By the 1970s, with growing feminist activism, attention to the situation of women also became an issue in the literature and within professional organizations. In subsequent decades, interpretations have widened to incorporate additional categories and to recognize their “intersectionality.” The initial ones remain those most often addressed, however, and accordingly will be discussed in the most detail here.

The first significant organizational endeavor by the Association of American Geographers (AAG) to increase diversity was the establishment of the Commission on Geography and Afro-America in 1964–1965. Its primary goals were to recruit and support African American undergraduate students in graduate studies, emphasizing those from Historically Black Colleges and Universities in the Southern states (AAG Diversity Task Force 2006). Beginning in 1972, additional categories were included in summary statistical profiles of composition of the overall membership of the AAG (published in its Newsletter): “Blacks,” “Spanish Americans,” “Oriental Americans,” and “Native Americans.” Difficulties of making estimates were noted. By 1981 the categories “Not Classified,” “American Indian,” “Asian,” “Black,” “Hispanic, “Native Alaskan,” and “Pacific Islander” were adopted, though for the next 20 years more than 90% of members were reported as “Not Classified.” Longitudinal membership data since 1972 are available from the “Disciplinary Data” section of the AAG website. These should be interpreted with caution, however, since not all geographers are AAG members and not all members reside in
GEOGRAPHY EDUCATION: DIVERSITY AND PARTICIPATION

the United States. Increasingly, with the use of online applications for membership, individuals do not provide these data. What are available do indicate some small growth over time of those identifying as (Black) African American (1.17% in 1972 and 3.15% in 2012), and Hispanic (0.13% in 1972 and 4.38% in 2012), but the other minority categories remained below 1% except for the noticeable increase in “Asian” representation, from 4.02% to 11.53%. Going beyond AAG membership, a 2004 survey of 74 departments (Estaville, Akiwumi, and Montalvo 2009) revealed that two-thirds of them had no minority faculty, that only 1.1% of the faculty was Hispanic and 3.6% Black. Asians accounted for 7.1%, more than their representation in the national population, likely including those who had come as international students and remained as faculty. That survey also sought information on undergraduates. Responses from 66 departments showed Blacks accounting for only 2.2% of geography students, and Asians 4.6%. The changes have been slow and limited and are not confined to geography programs, however.

National scale data present only part of the picture. As a recent theme issue of The Professional Geographer (66(2), 2014) highlighted, the location of institutions has implications for the potential and actual recruitment of students of diverse backgrounds. Large metropolitan areas like New York and those in the Middle Atlantic and Pacific Coast states offer more potential than central and southeastern states. Those community colleges in which geography is taught appear to have a greater likelihood than four-year institutions of enrolling minority students. Also documented is that sustained initiatives at the departmental level make a difference in increasing the representation of minority students on geography programs. Examples from two universities in Illinois illustrate ways in which setting and missions are relevant. Southern Illinois University Carbondale (SIUC) is distant from major metropolitan centers that have substantial minority populations. Normally, its incoming students have little prior knowledge of geography as a discipline. To foster diversity, the geography department adopted several strategies, among them collaboration with other units in a summer program to help underrepresented students make transitions from high school to university studies. Faculty also engaged in outreach to schools to reach potential students of minority or of the first generation in their families to participate in higher education. The faculty also established personal ties with geographers at the nearest Historically Black College and University with whom they shared interests in geographic information systems, resulting in recruitment of its students into SIUC’s geography graduate program. Different strategies were pursued at Southern Illinois University Edwardsville, located close to urban St Louis, Missouri. The institution has an advantage in that it offers the lower in-state tuition rates to students from across the state border. Faculty engaged in outreach to schools in St Louis. The department also used its undergraduate qualitative methods course to interview nonmajor women and minority students to investigate what would make geography attractive to them as a major. Drawing on those findings, a new information brochure was designed for undergraduates which included information on careers of alumni. Additionally, urban studies were reinvigorated within the department’s course offerings. The department has continued to monitor how these approaches impact on the makeup of their student body.

As with the initiatives about racial representation, international feminist activism in the 1970s fostered attention to the position of women in geography, a theme highlighted...
by AAG president Wilbur Zelinsky (1973). As women increasingly entered graduate programs and early-career faculty positions they became increasingly politically active. Through the newly created Committee on the Status of Women in Geography, they devoted themselves to supporting research and teaching on women and promoting the election of women to positions in AAG governance. Research also examined the representation of women across types of institutions and by specialties. Supplementing and challenging mainstream feminist perspectives were the voices of women of color, highlighting their distinctive experiences and priorities for social justice that reflected their personal and communal experiences (e.g., Pulido 2002; Sanders 2006).

The growing representation of women in geography reveals how cultural, political, and economic contexts both in academia and in the larger society have been implicated in the changing gender representations and position of women in the discipline. Historical research has revealed that women had been significant in geographic education from the late nineteenth to the mid-twentieth century in the United States both in the schools and in the institutions designed to prepare teachers for elementary and secondary teaching, though they were barely present on the faculties of universities. In 1917, for example, shortly after its founding, 645 women were members of the National Council on Geography Teachers (later the National Council on Geographic Education) and held leadership positions. This contrasted markedly with the elitist male-dominant membership of the Association of American Geographers that, over many decades was oriented to university faculty. Only 10 women were elected to its membership between its founding in 1904 and 1945, and several of those worked outside educational institutions, for example in federal government agencies (Monk 2004).

Since the 1970s, the increasing enrollment of women in graduate education has been sustained, though unevenly, intersecting with other political and economic changes. Data compiled by the National Center for Educational Statistics of degrees awarded in geography indicate that both the number and percentage of PhD degrees awarded to women have increased markedly over the succeeding decades from 6.0% in 1960–1961 to 46% in 2010–2011, a period in which the total number of doctorates increased almost fivefold. Awards of master’s degrees to women similarly increased from 14% to almost 43%. Gains have been slower at the bachelor’s level, with women earning only about one-third of those awarded in recent years at a time when they make up about 60% of undergraduate students nationally. The latter situation has been largely unexplored, but trajectories of change in gender representation at the graduate level appear connected to aspects of academic labor markets and public policies, as well as to women’s changing aspirations. The number of graduate degrees earned by men peaked relatively and absolutely around 1970, dropping or growing slowly thereafter, whereas increases in graduate degrees earned by women continued. Changes in representation in doctoral awards may be interpreted not only in relation to women’s aspirations or departmental cultures, however, but also as a result of men’s engagement with graduate education having decreased rapidly in the 1970s. The ratios of male to female student memberships in the AAG declined precipitously from 7.2:1 in 1967 to 1.7:1 in 1988. That change likely reflected not only expansion in the numbers of women but also men’s responses to a tight academic labor market as the expansion of undergraduate enrollments due to the baby boom passed, with related decreases in the creation of new faculty
positions. It may also have reflected the cessation of government fellowships offered under the National Defense Education Act (most of which had gone to men), and the end of the Vietnam War during which some men had gained military draft deferments through their status as graduate students. Government policies on affirmative action may also have shaped increasing employment opportunities for women (and minorities), though the extent and patterns of their influence are not clearly evident.

The location of women has, however, been institutionally uneven. Analysis shows that in departments that granted a total of 30 or more PhDs between 1990 and 2000, half or more of the recipients were women whereas in others they received only about 15% of PhDs granted (Monk 2004). Women’s representation also increased more on the faculties of doctoral departments than in institutions in which the highest degrees awarded were the masters or bachelors. This pattern may have reflected the smaller total size of the latter and perhaps also greater responsiveness to federal government affirmative action policies in the former which were more sensitive to external research funding. Still, it was noted that women took longer to achieve promotion in rank than their male peers.

Several studies of the representation of women in geography outside the United States also reveal the effects of geographic, political, and cultural contexts. In the 1970s research by Linda McDowell revealed that 43% of British geography departments lacked women faculty, but they were much less well represented in most English university departments than they were in Scotland and in the London colleges. The proportions of women declined rapidly with the progression from undergraduate to master’s and doctoral studies, and even more at the faculty level. In Spain changing demographics, legal contexts (including policies relating to practices of academic hiring), and shifts in labor markets saw low representation of women faculty in the 1970s, with increases in their hiring from the 1980s when government policies afforded greater local autonomy to universities and when academic curricula in geography were especially associated with preparation for teaching. More recently, the increased masculinization of academic geography departments has been evident as academic labor markets have shrunk while curricula geared to professional and technical work outside education have expanded in Spain (Garcia-Ramon and Pujol 2004).

Widening interpretations: beyond dichotomies

As critical studies have developed on the theme of diversity, it has been recognized that categories related to race and gender do not reflect the varied and layered social and educational experiences of students or faculty. The notion of intersectionality has become recognized, with calls to avoid essentialism. Research with American master’s and doctoral students, for example (Schlemper and Monk 2011), found that they spoke of themselves and their peers in relation to citizenship, country of birth, marital status, social class, language fluency, life stage, and sub-discipline. Less often did they mention physical disabilities or sexual orientation. Compared to students, however, faculty tended to refer to their students in the dichotomous categories of “men” and “women” or sometimes also to “minorities.” The location of institutions played a role in the perspectives expressed – those in large metropolitan areas reported different experiences and daily practices from those in smaller college towns. Married students and faculty also brought up the particular situation of dual career couples in relation to finding employment.
In the United States, foreign-born faculty and students are significant among those teaching and studying geography. They represent a variety of cultural backgrounds, national origins, and choice of specialties. Articles in the Journal of Geography in Higher Education (32(1) and 32(2)(2), 2008) explore a range of issues related to foreign-born faculty and teaching. Some commonly expressed (or assumed) views include students who are native speakers of English reporting problems understanding the accents of those from abroad, and students differing from teachers in their expectations of assumed knowledge or of grading scales, and in the choice of an appropriate language for interpersonal interactions. Yet research has also revealed positive comments by students, for example, that their perspectives have been widened and that they appreciate engagement with teachers who have firsthand experience of another culture or of employment outside academia. In addition to their academic work, international faculty may also simultaneously face legal problems relating to residency and challenges in balancing family and other personal responsibilities that differ from those of their native-born colleagues. Their concerns may include not only pedagogical matters but also aspects of marginalization by colleagues, and expectations that they are identified with technical and quantitative specialties. Beyond these aspects of internationalization in higher education at the individual level, however, are the longer-term implications of increasing internationalization in multiple national settings for the reshaping of knowledge, as recently proposed by Madge, Raghuram, and Noxolo (2015).

Placing diversity in the curriculum

An important issue in geographic education is how to approach teaching about diversity and to diverse groups of students. Attention to these matters is uneven in terms of contexts and approaches. The literature tends to focus on higher education rather than elementary and secondary levels. While it is not possible to make a global assessment, it is evident that initiatives to instigate change have varied over time and place, as explored, for example, in a series of articles on “Feminist Geographies Around the World” in the journal Belgeo (2007/3) and also in International Research in Geographical and Environmental Education (20(3), 2011). Gender themes, for example, emerged in geography teaching in an array of Western countries in the 1970s and 1980s, but have been more recent in Central and Eastern Europe, and continue to be largely absent in geography in some parts of the region. Research and teaching on gender and sexualities have recently expanded in some institutions of higher education in Brazil. In Asia, there is a strong program in Singapore, sustained teaching in Taiwan, and growth of participation in some institutions in India and Japan but apparently not in China. Most articles in International Research in Geographical and Environmental Education address elementary and secondary education but attention to diversity as a curriculum theme or of students is relatively rare within them other than by disaggregation of the quantitative data of male and female student responses on surveys or test scores. These tend not to be interpreted, however. An exception is a study of why British boys do not achieve as well as girls in secondary school examinations in geography, with differences connected to gendered cultural norms and preferences for different styles of teaching and classroom interaction.

Resources for teaching about gender in geography, especially in undergraduate courses, were developed in the United States as early as the 1970s, for example, when a federal government–funded project within the AAG
generated a set of teaching modules under the title “Women and Spatial Change.” An early introductory text by Mazey and Lee followed in 1983. In Britain, the Women and Geography Study Group of the Institute of British Geographers (IBG) published a widely used text, *Geography and Gender*, in 1984 (WGSG 1984), while a group of Spanish geographers published the text *Mujeres, espacio y sociedad* in 1995 (SabateMartínez, Rodríguez Moya, and Díaz Muñoz 1998). The research journal, *Gender, Place & Culture*, initiated in 1994, has periodically included articles related to educational themes and experiences, while the text *Space, Place and Sex* by New Zealanders Johnston and Longhurst (2010) widens the scope of attention to include sexuality, as have other subsequent texts aimed at university level, especially to advanced students and researchers. Joni Seager’s multiple editions of an atlas of women in the world, published between 1986 and 2009 under slightly different titles, some with co-authors, have also been a widely disseminated resource for students in other disciplines as well as geography (e.g., Seager 2009). Introductory undergraduate texts in the United States in human geography and world geography also integrate attention to gender situated in a variety of contexts.

Although a significant research literature in diversity has developed in geography, somewhat less attention has addressed experiences of teaching about race and ethnicity than gender. Two focused issues of *Journal of Geography* (98(4), 1999 and 107(6), 2009) included descriptions of courses, including one taking a historical approach in an introductory-level undergraduate course on the experiences of multiple racial/ethnic groups and another on developing student-led field explorations in multiple types of urban neighborhoods. Other articles reflected on the challenges of dealing with racism in the classroom, noting that approaches that may be well received in one setting may raise problems in another, reflecting the race/ethnicity of the teacher and those of the student body, and who is identified as “other.” A significant project entitled “Finding a Way,” led by Rickie Sanders, sponsored two summer institutes for teachers in 1996 and 1997 (Monk et al. 2000). These took up both content and pedagogy, particularly in relation to women and gender, while also considering multiple forms of diversity and incorporating pedagogies that are responsive to aspects of classroom climate and alert to local school and community contexts. At the secondary level, a workshop of the Australian Geography Teachers Association was followed by the book *Teaching Geography for a Better World*, which incorporated attention to gender (Fien and Gerber 1988). However, publications reporting sustained initiatives focusing on diversity themes in elementary and secondary geographic education appear to be relatively rare, though they may be cast within wider goals as in inclusion in a New Zealand mandate to foster critical thinking in secondary teaching.

Attention to forms of diversity other than gender and race are not as well represented in the geographic education literature, especially in relation to elementary and secondary education, but articles relating to higher education recur. A review of two decades of articles addressing the “other” in the *Journal of Geography in Higher Education* (Monk 2000) revealed articles on fostering sensitivity by the able-bodied to those who have various physical limitations, on dealing with diverse sexual orientations of students on field courses, or ways in which majority-culture students might be brought to appreciate the values and perspectives of indigenous or “other” communities. How students might gain insights into othering presents another pedagogical challenge that some geographers have addressed via exercises that involve experiential approaches to
simulate being in positions of the other or ones that involve tasks framed in terms of analogies. Recently, interest appears to have increased in online communications about the availability of resources to enhance teaching geography to students with impaired vision as, for example, they participate in online courses.

**Place, politics, and traditions**

Addressing multiple forms of diversity within education is a political endeavor. In a valuable examination of the dimensions of diversity and higher education, Sylvia Hurtado and colleagues (2008) argued that, in order to create and support an inclusive climate, it is important to consider community members’ attitudes, perceptions, behaviors, and expectations. She saw these as involving the historical legacy of campus cultures, traditions, and policies, structural diversity in relation to inclusion of previously underrepresented groups, the psychological dimensions of how people perceive diversity, the behavioral dimension of intergroup relations, and the degree to which individuals actively engage with diversity on campus. How these are enacted emerges in various ways across time, space, and place. The complexities are well illustrated in a study of postapartheid experiences in South African universities, in which Higham (2012) presents interpretations of students’ voices in the historically white University of Cape Town and the historically “colored” University of the Western Cape. Higham found that, though the desegregation of schools and universities aimed to promote inclusion, new class cleavages and other social differences emerged, based on group and institutional heritage and practices, complicating students’ experiences and perceptions. Black students at these institutions, for example, were drawn largely from emerging black elites, likely coming from schools that had already been privileged, not from the critically underresourced Bantu schools. Among students in the formerly “colored” university were feelings of loss of their previous solidarity and cultural heritage. Other issues across the campuses related to differential availability of financial support, the persistence of culturally distinctive classroom and recreational traditions and of gendered representations, and the difficulties some women saw of achieving access and status in the face of male stereotypes. Yet, Higham observed, not all students shared the same perspectives.

At another political level, Torres and Wicks-Ashbun (2014) identified the dilemmas and challenges of undocumented youth in North Carolina who had been brought as children to the United States. While state laws enabled them to enroll in elementary and secondary education, they were not subsequently eligible for in-state tuition rates or scholarships in public universities. Such policies place obstacles in the way of youth’s educational aspirations, possibilities, and/or on performance in high school. Other research has also shown that the way students of diverse racial ethnic backgrounds interact in school settings may generate yet other tensions, such as in the creation of spatial territorialities that are expressed in students’ sense of racial and ethnic subjectivity.

Overall, then, although geographers have become more alert to diversity within the profession, interpretations and practices are uneven across temporal and spatial contexts. Challenges remain in structures, in professional participation and engagement, and in the creation and implementation of pedagogies that strive to be inclusive of opportunity and learning. The current scene is that, though geographers are more cognizant of diversity issues within the discipline – as in who participates, what content is taught, and pedagogical approaches – interpretations and practices vary across temporal and spatial contexts.
and between individuals. Initiatives appear to be more extensive and sustained, or at least better documented, within higher education than in elementary and secondary levels. That they have been slow and limited is not confined to geography programs, however. As Patricia Price (2014) has noted, contemporary higher education espouses policy goals that articulate diversity as positive, ethical, and just, but the pressures of neoliberal economies that emphasize individualism, increasing budgetary pressures, and, in some contexts, legal challenges to affirmative action in student admissions and faculty hiring contribute to making the attainment of diversity elusive.

SEE ALSO: Difference; Ethnicity; Gender; Race and racism; Sexualities

References


Informal education is a term applied to learning outside of organized or “formal” school settings. It takes a range of forms including visits to museums, libraries, zoos, and exhibitions; field experiences, tours, and travel; civic engagement projects; as well as festivals, public lectures, community events, and many other types of activities. Informal education also encompasses learning that occurs through the mass media, including movies, newspapers, television, the web, social media, magazines, and even video games.

Informal education is generally contrasted with its formal and nonformal counterparts. Formal learning occurs within educational or training institutions, usually accredited schools and universities, where courses are intentionally designed and structured in terms of learning objectives, learning time, or learning support, whether in-person, in-class or online. Nonformal education can include community or noncredit adult education courses or activities (such as sports, fitness, or health programs), participation in noncredit MOOCs (Massive Open Online Courses), professional conferences, and continuing professional development. These are “organized” in a technical sense but no credential, degree, or certificate is awarded for participation or completion. Learners participate to increase skills and knowledge, or simply for enjoyment.

Informal learning tends to be even less structured and more spontaneous, also with no intention that the activities lead toward any particular certification or degree. Informal education may or may not be organized around specific learning outcomes, but these objectives may be in the background. Learners typically participate because the activities themselves are fun, interesting, and engaging. Informal education may also be embedded in programs of formal programs, such as in community engagement projects, forming a pedagogical link between formal and informal education. In recent years, there is an increasing trend to engage students, even starting in preschool, in various nonformal and informal activities, especially in developed countries (Salmi, Kallunki, and Kaasinen 2012).

Informal education in geography takes many forms and occurs in many different contexts depending upon targeted educational goals. Formal geography education in which students learn about people, place, and environments often involves teacher-centered, school-based, whole-group instruction. Informal learning contexts stand in marked contrast to this. Informal education is generally community based, individual or small-group-focused, and learner-centered (Gerber 2002). Geography education that takes place outside of the classroom – in local communities, field settings, or through the offerings...
of other cultural institutions – allows students to understand how geographic work contributes toward the solution of problems and improvement of society. Geography educators have explored and used these alternative learning methods and contexts to advantage for a variety of purposes.

Informal education has long been an essential part of geographic work, given its tradition as a field-based discipline. The discipline’s central tenet is that location matters for understanding the relationships and interdependencies that shape varied local communities. This spatial perspective provides a cross-cutting way of looking at processes and phenomena that other disciplines tend to treat in isolation. Sensitivity to the ways peoples, places, and environments are connected can help communities address the social and environmental concerns they face. Geography education in informal settings or through public or community engagement can help students connect the classroom subject to the ways in which this knowledge can contribute to society. Informal education that is unconnected to formal purposes may also help people understand their own communities better, although ascertaining whether this occurs is a challenge, given the very nature of informal learning.

With increased globalization, public issues and social problems stretch across spatial scales, affecting large numbers of people and offering potential for individuals and communities to learn together. The concept of engaged learning is based on partnerships between educational institutions and local communities. These opportunities incorporate the advantages of informal learning within the structured, formal setting. The concepts of civic, community, and public are three interchangeable terms that reflect the notion of participatory citizenship. The term “community” is a notion with varying definitions, uses, scales of inclusions, and significances to different constituencies and stakeholders. The “public” comprises an infinite range of potential audiences for whom targeted activities/events are appropriate: families, adults, special-interest groups, the elderly, community groups, schools, teachers, and so on. Indeed the term public is now being more generally used (Nicholson and Wheater 2012). By whatever name, community engagement may be undertaken at a range of geographic scales – from the immediate or local community, to inter-, cross-, multi-, or transnational communities – to communities founded on other diverse characteristics, such as education, occupation, age, ethnicity, or social, sporting, political or environmental affiliations. It can be applicable in local, national, or international settings. Hanson (2004) noted how important it is for geographers to convey to others, particularly students, what we do and why it is important. This can be achieved through the diversity of the questions we ask to show relevance of geography to a variety of problem areas (access to clean water, responses to climate change, sustaining biodiversity), in a range of settings and scales (urban neighborhoods, smallholder farms, national parks), and to an array of audiences (community groups, NGOs, government agencies). Examples are discussed later.

Types of informal education in geography

A wide range of informal learning in geography can be observed outside the classroom and typically outside of academic settings. Important examples include community engagement projects, field trips, museum visits, public lecture, and geography festivals. Many of these
informal opportunities may be a component of a formal education, organized as supplementary learning opportunities for classes. Along with educational institutions, most of these forms of informal education have been organized by different geographic organizations and professional associations like Royal Geographical Society (London), National Geographic Society, Geographic Society of Chicago, American Geographical Society, Geographic Alliance Network, Esri, and various government agencies and community groups.

Community engagement

Generally, community engagement can be defined as “the collaboration between institutions of higher education and their larger communities for the mutually beneficial exchange of knowledge and resources in a context of partnership and reciprocity” (Holland and Ramaley 2008). Community based learning is a process by which people from the community engage with students and faculty to bring about changes that enhance local life. In terms of learning, community engagement is a complex, multifaceted process that involves activity in, for, with, and/or through communities, with potential benefits and challenges for all parties involved (Bednarz et al. 2008).

Engaged learning represents a blend of formal and informal settings and is part of a broader effort in higher education to incorporate community engagement in teaching and research. For faculty members, engaged scholarship incorporates teaching, research, and service responsibilities with the goal of partnering with community groups for mutually beneficial outcomes. Students work informally (within their formal educational experience) with community groups on a variety of projects. For students, engaged learning is a journey through “knowing that” to “knowing how.” Student learning through community engagement is rooted in problem-based, reflective, “deep learning” pedagogies of empowerment, transformation, critical thinking, and social participation. Such engagements can involve learning settings that take the student “outside” to work, research, and learn in partnership with communities – on a variety of civil, social, environmental, economic, and moral issues with local resonance (McEwen 2013). Geography education through community engagement is more evident in the United States, Australasia, and Western Europe than in other parts of the world, but such partnerships are now also being tried in the developing world, such as in Pakistan.

Table 1 presents summaries of the benefits received from geography–community partnerships in informal settings. The engagement benefits vary with the level of involvement and the stakeholders involved.

Community engagement can be used to attain various goals. For example, project goals include reaching cognitive outcomes (community based learning through fieldwork, Liverpool, UK); promoting participatory citizenship (understanding flood histories – understanding risk, UK); and improving student employability in the field of geography (supporting emergency planning using the geographic information system (GIS), USA). Similarly, student volunteers from the geography program at the University of Gloucestershire worked with community members to develop an informed community understanding of flood risk. Together, they examined flood probabilities and the issues associated with living with uncertainty, through development of local knowledge of historic floods and extended awareness of local flood impacts. Students had the opportunity to develop their skills by engaging with environmental and social
Table 1  Selected case studies of geography–community engagement.

<table>
<thead>
<tr>
<th>Geography–community engagement</th>
<th>Objective</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Students</td>
<td>Community</td>
</tr>
<tr>
<td>Advancing geospatial skills in science and social science, USA</td>
<td>Spatial thinking and problem solving through GIS and remote sensing</td>
<td>Developed competencies and interpersonal skills</td>
</tr>
<tr>
<td>Community engagement for experiencing indigenous values, New Zealand</td>
<td>Understanding Maori cultural framework</td>
<td>Understanding of how people interact with their environment</td>
</tr>
<tr>
<td>Understanding flood histories – understanding risk, UK</td>
<td>Risk awareness flood resilience among local communities</td>
<td>Action learning opportunity, real-world experience of flood-prone areas</td>
</tr>
<tr>
<td>Supporting emergency planning using GIS, USA</td>
<td>Develop efficient emergency response system</td>
<td>Learned about emergency response Practice geographic skills, GIS applications</td>
</tr>
<tr>
<td>Community based learning through fieldwork, Liverpool, UK</td>
<td>Fieldwork exercise in development geography</td>
<td>Understanding of global citizenship and development geography Development of personal skills for societal contribution</td>
</tr>
<tr>
<td>Study tour and field survey to northern areas of Pakistan, The Islamia University Bahawalpur, Pakistan</td>
<td>Physical features, socioeconomic and cultural study of people living on mountains</td>
<td>Understanding of geomorphic features and social, economic, and cultural issues of natives</td>
</tr>
</tbody>
</table>

aspects of sustainable flood-risk management and flood-risk science. Similarly, geography students were involved in the Plymouth East End Urban Regeneration project, which offered various community benefits including housing renovation, improvements of local parks, commercial improvements, and development of community villages (Klein et al. 2011; McEwen 2013). Learning through community engagement offers many benefits to all stakeholders, particularly to students and the community. It helps students to develop interpersonal communication skills and improve their understanding of geographic contents. It may also increase student’s motivation, add to their work experience and career development, and heighten their moral and civic responsibility.

Community engagement can also help foster a culture of interaction and mutual respect among community leaders and students. Proponents of community engagement argue that it helps produce a better-skilled and educated workforce that is self-motivated, flexible, and adaptive to change. For institutions of higher education, community engagement can reduce barriers and create strong links between them and the public. It also increases students’ awareness about civic and social issues and promotes active citizenship. The most prominent benefit to the community is that it helps find solutions to problems. For faculty members, the main reward is to develop and improve their personal pedagogy skills. Key barriers for faculty are workload issues, a lack of incentives, confidentiality issues, inexperience with engaged pedagogy, and insufficient financial resources. On the other hand, engaged scholarship can provide a meaningful structure for course content and improve a variety of learning outcomes. Engaged teaching and learning can transform instructors into reflective practitioners as well as provide them with future research opportunities.

Field trips, tours, travel, and field experiences

Fieldwork is at the heart of geography, arguably one of its signature pedagogies. Fieldwork often plays an important role in formal geography education and research. In the informal domain, field experiences, field trips, and travel can also serve to promote geography education. The UK Quality Assurance Agency (QAA) benchmark statement provides a definition of fieldwork as being, “active engagement with the external world.” Defined “fieldwork” may include field teaching, field trips, field research, or field camps or indeed “any arena or zone within a subject where, outside the constraints of the four walls classroom setting, supervised learning can take place via firsthand experience.” Field experiences are recognized as a vital part of the teaching of geography at all levels. Not only is it considered essential, but it is considered by both academics and students to be an extremely effective and enjoyable learning and teaching method (Fuller et al. 2006). Michael Palin, former president of the Royal Geographical Society, said that field trips were a vital component of geography, insisting that students’ interest in the subject was fired by good teaching combined with opportunities to “get out of the school building.” Exposure to television, advertising, and modern travel has extended learners’ awareness of other places and environments, but nevertheless they often remain unobservant, uncritical, and unchallenged. Many learners make little use of their experiences in their local area to really understand why things are as they are. UK universities with an overseas fieldwork component always find places on such trips easy to fill, no matter how far the destination (e.g., University of Plymouth to Australia; University of Wales, Aberystwyth, to New Zealand). Massey University and Victoria University, Wellington, also offers field trips to the South Island of New Zealand, although the
GEOGRAPHY EDUCATION, INFORMAL AND FOR PUBLIC ENGAGEMENT

geographic isolation of New Zealand generally precludes overseas field trips.

Field trips can also be done at the local scale. In formal educational settings at any grade level, geography courses often include local trips to better understand and interact with the community. Students of urban geography at the Department of Geography, The Islamia University Bahawalpur, Pakistan, were assigned a field study task to visit the core city to know and understand the urban morphological character as well as the issues and challenges of the area. This activity facilitated interactions between students with the local community and highlighted the issues faced, such as environmental degradation in these communities. Research students in medical geography from the same department were sent to the urban slums to identify the socioeconomic and environmental issues affecting public health, particularly sanitation, water quality, and housing. After appropriate research findings, they voluntarily communicated with local dwellers about water purification methods, hand-washing and maintaining domestic sanitation utilizing local resources to avoid health problems, such as diarrhea. Involved students and faculty also communicated their findings and suggestions to the Tehsil Municipal Authority Bahawalpur for environmental upgrading in slums.

Research on the outcomes of field trips has pointed not just to the social benefits of spending time working together, but also to the added benefits of the development of cognitive and effective learning. It aids motivation and self-development of students and provides them an opportunity to be creative and to have personalized outcomes which give them a sense of ownership of the learning process. In addition, it enhances students’ understanding of geographical features and concepts and allows them to develop specific as well as general skills. However, significant amount of planning and resources are required to carry out successful field trips. Liability issues relating to field risks and lack of local knowledge can also undermine effective outcomes.

Similarly, variously themed guided tours have been organized by the American Geographical Society, National Geographical Society, New York Times, the Royal Geographical Society, and various other organizations. One example is Discovering Britain, which is an exciting series of geographically themed walks created by the Royal Geographical Society (with the Institute of British Geographers) that aim to bring these stories alive and inspire everyone to explore and learn more about Britain. Similarly, but more informally, guided tourist trips are organized by private travel agents, sometimes working with local musicians or others who want to show fans their home country (e.g., Albannach, a Scottish band, leads tours annually to remote corners of Scotland). This does not even mention the travel options available to tourists at any resort where features of local interest can be explored. Again, by definition, there are no learning objectives specified or assessed, but it is likely that these kinds of experiences broaden people’s geographic awareness.

Museums and exhibitions

Many museums can be used in geography education. School field trips to local museums to study the local history and geography, or to natural history museums, zoos, and other museums, are common in the United States and Western Europe. Museums offer a variety of possible learning experiences for different age groups and learning styles. In the specific context of geography, the National Geographic Museum, Washington DC, is great for all ages and features a wide range of changing exhibitions, from interactive experiences to stunning photography
exhibitions featuring the work of National Geographic explorers, photographers, and scientists. Similarly, The National Museum of Natural History (NMNH) Smithsonian Institution, Washington DC, is dedicated to inspiring curiosity, discovery, and learning about the natural world through its unparalleled research, collections, exhibitions, and education outreach programs. In the past, the exhibition called “The Power of Maps” by Denis Wood, author and a cartographer, at the Cooper-Hewitt National Museum of Design in New York in 1992 was considered as a great exhibition in the context of geography. That show was remounted at the Smithsonian Institution in Washington in 1994.

Many academic institutions also arrange exhibitions on various geographical issues. For example, the Department of Geography and Environmental Management at the University of Western England in Bristol designed an interactive sediment-core exhibit, which helped visitors understand how climate and environmental changes are recorded in fossil sediments spanning 15,000 years. This activity benefited academic staff as they communicated complex science and unfamiliar subjects to a nonacademic visitor audience, enhancing the profile of the department.

Therefore, in terms of strengths, this informal setting offers an extremely effective, interactive and enjoyable learning form. It is an easy way of elaborating knowledge which enhances intellectual curiosity of visitors. Moreover, exhibitions are the most effective way of disseminating geographic research works to the public. Lack of money, accessibility, and mobility are the key barriers for this form of informal education.

Community events and festivals

Informal geography education also occurs in varied community, nationwide, or even international events that are devoted to familiarizing the general public with the content, methods, and relevance of geography. National Geographic Education Programs (NGEP) created Geography Awareness Week more than 25 years ago. This annual event, during the third week in November, focuses on a specific geographic theme; for example, in 2014 the topic was food and nutrition. In most US states, the geographic alliance sponsors a variety of events and resources related to the topic. Students, families, and community members celebrate the importance of geography by using lessons, games, and challenges in schools and the community, often meeting with policymakers and business leaders as part of that year’s activities. The main objective is to raise awareness among young Americans about geographic issues, recognize their impacts as global citizens, and excite people about geography as a discipline, as a career, and as a part of everyday life. This annual public awareness program also encourages citizens, young and old, to think and learn about the significance of place and how they affect and are affected by it. Each year more than 100,000 Americans actively participate in Geography Awareness Week (GeoWeek).

GIS Day has been celebrated every year since 1999. This is an international forum for users of geographic information systems (GIS) technology to demonstrate real-world applications that are making a difference in our society. Esri president and co-founder Jack Dangermond credits Ralph Nader with being the person who inspired the creation of GIS Day. He considered GIS Day a good initiative for people to learn about geography and the uses of GIS. He wanted GIS Day to be a grassroots effort and open to everyone to participate (www.gisday.com).

The International Geography Festival is also one of the important geographic events organized by ADFIG (Association for the Development of FIG (International Federation of Surveyors)) in Saint-Dié-des-Vosges in France.
since 1990. The festival is under the patronage of the International Geographical Union. In the United States, local festivals held in conjunction with Earth Day in April also may include presentations and community resources related to environmental and sustainability issues that tie to geographic content.

Social media and “popular geographics”

There are many other informal settings and a wide range of activities that offer geographic learning opportunities to the public community, for example, online games and simulations, social media, mass media, and community clubs, among others. Sometimes these kinds of informal opportunities are sponsored by institutions of higher education. One of the examples is that the University of Gloucestershire established a cafe society in Cheltenham to share and explore the latest ideas in “science and technology” and “geography and environmental sciences” (i.e., climate change). This informal setting facilitates interaction among members of the public with academic staff and students at a neutral and public venue. Geocaching is an interesting outdoor recreational activity in which participants use a global positioning system (GPS) receiver or mobile device and other navigational techniques to hide and seek containers, called “geocaches” or “caches,” anywhere in the world.

The proliferation of GPS-based mobile apps and widely available mobile mapping technologies has created a revolution of public interest in maps and spatial representations. This is a growing part of informal geography education. Geospatially referenced posts on Twitter and other social media have led to an explosion of interesting maps and other creative ways of visualizing spatial information that now circulate across the Internet. Beyond that, there is an ever-larger array of emerging location-aware technologies, which allow people to connect to local spaces by coordinating and communicating with others without depending on physical proximity to each other. Lacking any official structure or leadership, and manifested in many different types of technologies, this revolution reflects the democratization of spatial information; Jerry Dobson of the American Geographical Society has called this phenomenon “popular geographics” (Dobson 2011/2012). Individuals of all stripes, who are not mapping professionals and may have little formal geography education, are generating an enormous amount of new spatial data. While some of this is perhaps unwitting – the spontaneous and hidden creation of big geodatabases used in marketing research, based on browser searches or other uses of GPS-equipped apps – there is also an intentional side. Basically, people are suddenly mapping for fun. Examples range from widely used desktop and mobile apps such as Google Earth and Google Maps (and its competitors) to open-source GIS software for people seeking to analyze publicly available spatial data.

The public has also itself become, collectively, an agent of spatial data-gathering (“crowdsourcing”). Tens of thousands of individuals voluntarily collect georeferenced information for sharing and creation of crowdsourced maps, such as Open Street Map. These allow users to read and write locations. Social networks and some location-based services enable users to read tips and comments that previous users left attached to that location. This information, now with geographic coordinates, becomes an intrinsic part of the location. All of these (and more, both in existence and yet to be invented) comprise a new form of informal geography education. Popular geographics are built on the foundation of applying local knowledge. Going forward, it is likely that these informal
activities will be incorporated into formal educational experiences, overlapping with community engagement efforts, for example, when students and community groups work together to map issues of local concern. Murmur is an example of such a project. Launched by the Canadian Film Centre’s New Media Lab, it integrates audio interviews into cell phone-based tours. Similarly, the Center for Digital Storytelling created storymapping.org in 2006, with projects in Mendocino (California), Houston, New Orleans, and Tuscaloosa (Alabama) to promote the connection between storytelling and issues of local memory and civic planning. Popular geographics is an emerging area of research in geography education and to what extent these public activities will improve people’s understanding of their own local conditions and spatial connections with others remains to be seen.

Informal education and future of geography

Geography as a discipline has a long history of developing learning settings beyond the traditional parameters of the classroom and for engaging with diverse communities as part of its fieldwork. Effective geographic practice is at its best when engaging across a variety of places, spaces, and scales. In recent years, geography educators have sought to make their pedagogy relevant to pressing social and environmental issues, integrating the approaches of place-based education to make geography education more relevant. Community based research, teaching, and learning provide significant potential for geography’s and geographers’ engagement in societally relevant research and learning. Placing the academic curriculum into a “real-world” context helps students realize geography’s applications. Through informal learning opportunities, geography students can engage more fully with their local communities, “not just as passive observers but as active participants and contributors” (Mohan 1995).

SEE ALSO: Geography education: fieldwork and contemporary pedagogy

References


Geography education, workforce trends, twenty-first-century skills, and geographical capabilities

Michael Solem  
*American Association of Geographers, USA*

This entry focuses on the role of geography in professional and personal life, framing the discussion on international examples from the research literature and broad-scale projects and programs led by geography’s professional associations. The discussion will illustrate the complexity of workforce development issues in geography, touching on the uneven opportunities for students to study geography in schools and universities, the variety of contexts in which geographers find employment in the public and private sectors, and how different countries view what it means to be geographically “literate,” “competent,” and “capable.”

Geography education varies strikingly when considered from an international perspective. There is tremendous diversity in the scope and sequence of geography in schools, both within and between different national jurisdictions. In some countries, geography in schools is affiliated with the natural sciences (Finland, for instance, aligns geography with biology in the curriculum). In other places, such as the United States, school geography is considered to be primarily a social science, albeit one with an increasingly high technological profile in the form of geographic information systems (GIS). Travel to schools in Japan, and geography assumes a greater curricular role as the study of human–environment relationships. The reasons for this variety are political, cultural, and historical. This rich tableau also extends to undergraduate and graduate (postgraduate) settings across the globe.

Unsurprisingly, all of this disciplinary complexity means students leave schools and universities with quite different educational experiences in geography (if they are fortunate enough to have received any at all). And yet, recent work on comparative education in geography finds some commonality in vision as to the discipline’s purposes and values in the education of young people and adult learners. When thinking about what it means to be a professional geographer in a global context, a broader view of the discipline and the significance of its body of knowledge for life and work – a philosophical issue – can reveal insights that may not be evident through a finer-grained analysis of the skills students acquire through their geography classes – a workforce development issue. This dualism is explored further in the following passages, beginning with a discussion of the trends shaping skills-led curricula for workforce development.

**Trends in the twenty-first-century world of work**

One of the benefits of geography’s high degree of interdisciplinarity and cross-cutting analytical technologies is the wider range of potential industries where students might find employment. In the United States, the US Department of Labor projects “faster than average” or “much
faster than average” growth in openings for geographers, geoscientists, cartographers, urban and regional planners, and other geographic professionals, with projected needs of upwards of 15,000 additional employees in each of these career fields between 2008 and 2018 (Solem, Foote, and Monk 2013). Many of these opportunities are in new areas of employment for geographers, and employers often seek novel combinations of geographical and general skills.

Nonetheless, many secondary and undergraduate students are unfamiliar with the extent to which geography is practiced in public and private sector organizations (the same might also be said for their faculty advisors and career counselors). Because few vacancies explicitly call for a “geographer” per se, one of the main challenges that students have in finding work is convincing a potential employer that a geography degree qualifies individuals with a unique set of skills and abilities. In practical terms, preparing students to enter the geography workforce requires consideration of a very broad range of occupations, few of which carry the job title of “geographer” but all of which stand to gain from the geographer’s spatial perspective and repertoire of applied and analytical skills.

At the graduate level, a master’s is now considered by many employers as the entry-level degree in many countries. Employers are increasingly seeking graduates with advanced credentials and prior working experiences that reduce the need for additional training. The emergence of degree models such as the professional science master’s, which blends scientific training with coursework in management and administration, warrants attention from geographers. Constructive relationships can be built through the establishment of internship programs, which are often major recruitment mechanisms. All varieties of internship program can lead to applied experience that students frequently cite as a reason for having entered the profession in the first place. Internships also have value for helping departments understand industry needs and challenges in order to prepare graduates for future careers.

Geographers pursuing a PhD in preparation for faculty careers also face significant challenges. The modern university is undergoing considerable change with the advent of sophisticated learning technologies, rapidly changing demographics, new business models affecting everything from tuition to tenure decisions, and increasing expectations for excellence in teaching and research. Against this backdrop, the importance of faculty development, in the form of mentoring and professional support in areas such as writing, grant proposals, curriculum and assessment, and publishing, will likely continue to grow as new faculty make the transition from doctoral study to academic posts.

As the role of the PhD in society continues to evolve, students with advanced degrees will likely see more opportunities to participate and contribute through nonacademic careers, especially given the demands for a highly skilled, analytical workforce in a knowledge economy. The question remains, however, of whether the “Plan B” stigma (i.e., the belief that a nonacademic career is a “fallback” plan) often associated with PhDs working outside of academia can be overcome. The issue concerns the extent to which postsecondary faculty, whose career experiences are largely shaped by academic culture and contexts, can better prepare PhD students who prefer a different career path.

**Modeling competencies in geography**

A common thread across these varying international contexts and levels of education is a sense of the need to ensure geography students obtain the education and master the skill sets that will prepare them for careers in both core areas where
Geographers have traditionally found employment (such as secondary education, government agencies, and urban and regional planning) and in emerging and nontraditional areas – including geospatial intelligence, medical geography, sustainability management, and operations and logistics analysis. In Eastern Europe, for instance, researchers have documented a disconnect in Poland between academic preparation, career advising, and employer expectations, a situation that has resulted in difficulties with job placements for many geography graduates (Piróg 2014). Hungarian geography has a centuries-old tradition of preparing students for careers in elementary and secondary education, but has only since the early 1990s attended to workforce development and the education of geographers at the undergraduate and postgraduate levels (Mezősi, Mucsi, and Garamhegyi 2001).

Geographers have led multiple studies for the purpose of identifying the ways geography education prepares students for work and life in the twenty-first century. International attention to workforce issues in geography has increasingly focused on connections between employment sectors and academia. In Europe, there is widespread activity attempting to define and measure competencies – that is, collective sets of knowledge, skills, and perspectives required for effective work in different industries. Competency models are now available for a large number of industries, and have become popular tools for preparing future workforces and assessing the abilities of current employees.

Debates over competence-based curricula across Europe are mainly influenced by the competency-based design of educational processes in the wake of the Bologna Process. Within this process the major goal is to connect subject-related knowledge and generic competences among students in order to achieve basic competences to foster multiple problem- and application-oriented solutions. This competence-led approach, with its concrete description of knowledge, skills, and perspectives as educational outputs, has been implemented on the European level in the form of the European Qualifications Framework for Lifelong Learning and on the national level in the form of the German Qualifications Framework for Lifelong Learning. Devised on the basis of the Tuning Project (Donert 2007) and the Dublin Descriptors of the Joint Quality Initiative, a model has been developed by German researchers with the goal of identifying core competences such as spatial thinking, technical skills and problem-solving, and modeling and measurement of spatial data in the geospatial technology sector (Schulze, Kanwischer, and Reudenbach 2011).

In the United States, research with employer organizations and professional geographers led by the Association of American Geographers’ Enhancing Departments and Graduate Education (EDGE) project (http://www.aag.org/edge) found that employers value training in both geographic and more general “soft” skills (Solem, Cheung, and Schlemper 2008; Schlemper, Adams, and Solem 2014) (Figure 1). Employers across the public and private sectors reported that they are especially interested in hiring people with strong skills in areas such as time management, writing, problem-solving, technology, and communication. Employers want employees who are systems thinkers, who are interdisciplinary, and who can locate and integrate information from varied sources, but specialized skills and disciplinary knowledge are also highly valued. Professional geographers are also often expected to collaborate in team settings and, as they progress in an organization, must exercise a number of interpersonal skills to supervise and manage subordinates. Transferable skills such as these are important for practice, irrespective of scientific specialization. The
Figure 1  Skill areas noted by geography alumni (N = 280) as ones they “always or very often” need to perform. Numbers in brackets signify the overall ranking based on the percentage responding.
full range of geographic and general skill areas expected to be performed in professional geography is available in a competency model produced by the EDGE project (Boxes 1 and 2). A competency model has also been developed specially for the geospatial technology industry (DiBiase et al. 2010).

Geography’s graduate programs generally provide students with a diverse and highly valuable set of transferable skills (Solem, Kollasch, and Lee 2013). Graduate students in all types of programs received the most extensive skill preparation in writing, critical thinking, and research planning and design. There are some differences
in skill preparation, depending on the type of department offering a master’s degree. For instance, PhD-granting departments appear to provide master’s students with more preparation in ethical practice and time management. Doctoral training, meanwhile, tends to concentrate more on academic publishing, grant proposals, and professional and organizational culture.

There are many similarities across graduate programs in terms of skill areas that are relatively underdeveloped. Graduate students in all types of programs receive the least training in fiscal management, entrepreneurial skills, and foreign languages. They also tend not to acquire much experience in developing skills that are especially valued outside of academia, such as supervising, relationship-building, project management, visioning, and adaptability.

Improved professional development and career preparation can help support the participation of women and underrepresented ethnic and racial groups in geography. This is an issue that has long been of international concern within the discipline (McEwen et al. 2008; Adams, Solís, and McKendry 2014). Professional development can help enhance the discipline’s diversity through recruitment and retention because there is often a mismatch between the implicit knowledge and skills needed to succeed in professional careers and the topics covered explicitly in curricula, advising, and mentoring. While many teachers and advisors are well equipped to prepare students for careers, they are often unprepared to advise students about the specific ways geography can address the needs and demands of employers in the business, government, and nonprofit sectors where the majority of students will seek careers. For students from economically disadvantaged backgrounds, having a clear understanding of what a discipline such as geography offers them is perhaps the steering criterion when making decisions about future career paths.

**Geography education for human capabilities**

As schools and universities seek better ways of preparing students for the twenty-first-century world of work, students also need help with understanding the broader purposes and values of the geographical knowledge they acquire through their academic studies (Solem, Lambert, and Tani 2013). Although being knowledgeable and proficient in geography opens pathways leading toward careers in a broad array of industries, many students find it difficult to express what their geography education affords them in terms of their future potential and well-being beyond employment. This is largely a consequence of educational systems that often uncritically promote a skills-driven agenda to make curricula more “relevant,” without giving due consideration to the broader purposes and values of education in life.

At the same time, career preparation is increasingly emphasizing considerations of lifelong learning, professional ethics, and the significance of disciplinary knowledge for thinking and reasoning about life in the broadest sense. This perspective on life, education, and careers attempts to strike a balance between education for human capital development, such as the competency approach described in the preceding section, and education for human welfare development, known as the capability approach. In addition to the challenges of addressing workforce development, researchers in several countries have begun to consider questions of how education contributes to human potential and wellbeing. Whereas much of the current reforms shaping curricula internationally in
schools and universities can be traced to neoliberal economic pressures for producing human capital (skills and competencies), capabilities offer a different (yet complementary) view of the role of education in human life. Rooted in the early writings of economist Amartya Sen and philosopher Martha Nussbaum (Nussbaum and Sen 1993), capabilities refer to sets of “functionings” that, once attained, provide people with real opportunities to reach their potential and wellbeing.

The capabilities movement is not intended to discredit efforts to prepare students for the modern global economy; rather, it is simply meant to broaden the dialogue over the role of schools and universities in ways that capture more holistically the benefits of disciplinary knowledge for lifelong learning, citizenship, and social relations. As such, the concept of capability is distinct from that of competency. Capabilities represent more clearly our educational goals and are more broadly related to human welfare development. Graduate attributes are often used by educators in synonymous terms with capabilities. They are difficult to assess as they are more subjective, holistic, and values-based.

In geography, the GeoCapabilities project is exploring these issues in partnership with several universities and scientific organizations in different countries (www.geocapabilities.org). Using comparative methods, researchers are analyzing the ways geography is considered by nations to be a form of knowledge that enables the development of human capabilities. A challenge for geographers is to identify the specific ways that geographical knowledge, skills, and perspectives contribute to the development of graduate attributes, or “geocapabilities,” which enable us as humans to think creatively and critically about ourselves, our communities, and our world. This challenge might be summarized by the question: What unique ways of understanding, thinking, and explaining do geographers introduce to students in schools and universities?

A capabilities study comparing the national geography curricula and standards in the United States, England, and Finland found that, despite the considerable variation in geography content and sequencing across grade levels, all three nations share a view that the role of geography in schools is to prepare students for life in three specific ways (Solem, Lambert, and Tani 2013). First, geography education is seen to promote individual autonomy and freedom by cultivating the ability to use the imagination and to be able to think and reason using geographic information and concepts. A second capability of geography is being able to identify and exercise choices in how to live, based on worthwhile distinctions with regard to citizenship and sustainability. A third contribution of geography human capability development is the way the subject helps people see their potential as a creative and productive citizen in the context of the global economy and culture. Collectively, these are the ways that three countries share a view of how geography education provides students with opportunities to achieve their life’s potential and wellbeing; without it, they are restricted and deprived in ways that have a real impact on their lives and those of others.

Work on the capabilities approach in geography has extended beyond Europe to include educational systems in Singapore, Japan, India, and China, as well as in Australia and New Zealand. In all of these settings, educators are debating how to best design curricula that fulfill societal expectations for the future workforce while not subjugating geography education only to vocational training and an over-reliance on a skills-defined experience. This means geography teachers, whether at the school or university level, must become re-engaged with some of the
most enduring and fundamental questions about educational purpose and practice in their discipline. A capabilities approach asks teachers to consider the specific curricular role of geography in helping young people think about their life in relation to the world, and what may become of their communities as well as people, places, and environments around the world.

Conclusion

It is important to contemplate what the trends and demands discussed in this entry mean for the role of geography in schools and universities worldwide. On the one hand, it is clear that geographers are wrestling with how curricula can be engineered to produce graduates with the skills needed for achieving the workforce goals of different nations. Opportunities abound for international dialogue and collaboration aimed at expressing the many ways geography can be deployed to enhance the operations, efficiency, and effectiveness of businesses and governmental agencies. Given the globalized and highly interdependent nature of the contemporary world economy, geography’s status in education seems tied, at least in part, to its perceived value in building an internationally competitive workforce.

There are, however, other values of a geography education that can help to solidify and expand its presence in school and university curricula. The capability approach to education is meant to orient a discussion about educational aims more toward how subjects such as geography provide knowledge that is needed for attaining one’s full potential and wellbeing in life. The opportunities for geography scholarship and geography careers are exhilarating, and the connection between higher education and workforce needs will remain critically important in the coming years. At the same time, schools and higher education institutions will be challenged to consider how curricula can be shaped to ensure a place for geography in the education and preparation of young people for life and citizenship in the modern world. This is different from a narrow focus on a skills-led curriculum that tends to emphasize learning outputs (competencies) over considerations of why subject knowledge (such as geography) is important for the fulfillment of particular educational goals; a capabilities approach asserts that both are important.

SEE ALSO: Geography education, informal and for public engagement; Geography education, primary and secondary: international perspectives

References


Piróg, Danuta. 2014. “Do Geography Degree Programmes Facilitate a Smooth Transition to the Job


**Further reading**

The magnificent diversity of Earth’s cultures and environments is fully evident in the many expressions of geography in school curricula around the world. Although the fundamental concepts and principles that make geography a spatial and ecological science are generally shared across borders, the world’s nation-states have set their own priorities as to what, when, and how the subject is taught to school children. Consequently, geography is a school subject that shows little conformity from place to place in terms of expectations of what students should know and be able to do as an outcome of their primary and secondary education.

In the United States, for example, geography is strongly aligned with the social sciences (history, economics, civics, and philosophy), whereas students in countries such as Japan and Finland study geography more as a natural and environmental science. In other parts of the world, such as in England, geography has a curricular association with the humanities. Much of this variation can be attributed to the fact that, like other subject areas (and especially in the social sciences), the content and status of geography in the curriculum is susceptible to external influences and vicissitudes in public opinion, economic conditions, political governance, and environmental change. This influence is most often felt in the content of the subject’s national curriculum standards.

In the following sections, selected countries are profiled to illustrate the range of curricular traditions that geography has experienced internationally as a school subject.

**United States**

Since 1994, the United States has had voluntary national standards for geography that set benchmarks for geographic learning at the 4th, 8th and 12th grades. Published in the volume *Geography for Life* and most recently updated in 2012, the 18 US national geography standards specify expectations for student learning in six domains of knowledge: the world in spatial terms, places and regions, human systems, physical systems, environment and society, and the uses of geography. In a nutshell, the purpose of school geography in the United States is to develop a geographically informed person – “someone who sees meaning in the arrangement of things on the earth’s surface, who sees relations between people, places, and environments, who uses geographic skills, and who applies spatial and ecological perspectives to life situations” (Heffron and Downs 2012, 7).

Because of the voluntary nature of national curriculum standards in the United States, states...
determine the goals of the curriculum, the instructional materials used to support implementation of those goals, and the assessments designed to measure student attainment. Forces ranging from economic globalization and global environmental change, to political calls for national educational reform and accountability, and an utter lack of federal funding appropriations for geography, have all contributed to the waxing and waning of the subject in US schools (Bednarz, Heffron, and Solem 2014).

As of 2015, at either the middle school (grades 6–8) or high school (grades 9–12) level, geography may be present as a strand within social studies standards or as a separate set of standards (sometimes paired with history), often linked to a course (McClure and Zadrozny 2015). In the elementary grades (K–5), geography is mostly integrated with the social studies. At the middle school level, nine states require a stand-alone geography course, and four states offer a combined geography/social studies course. At the high school level, four states require a stand-alone geography course, six states offer a combined geography/social studies course, and geography is an optional course for graduation in four states.

Europe

There is tremendous diversity in school geography requirements across Europe. Some of the significant differences and experiences are highlighted here. In England, geography is compulsory in years 5–14 with optional courses of study in subsequent years. As stated in a report to the International Geographical Union (Brooks and Catling 2014), the aim of the national curriculum for geography in primary education carries an affective component (to inspire curiosity and fascination about the world) as well as a content dimension, which calls for students to become knowledgeable about people, places, and resources, as well as physical and human environments and processes. Through their learning of geography and applications of map skills and field methods, students are expected to develop locational knowledge, including the human and physical features and characteristics of places, as well as the spatial interactions and interdependence of places and how these can change over time. In secondary schools the emphasis shifts to the study of where places are, how places and landscapes are formed, how people and their environment interact, and how a diverse range of economies, societies, and environments are interconnected.

Nordic countries offer an interesting comparison with England and the United States, but even there the differences can be remarkable. As in the United States, school geography in Sweden is classified as a social science. In grades 1–3, geography is a subject integrated with the social studies. At the middle school level, nine states require a stand-alone geography course, and four states offer a combined geography/social studies course. At the high school level, four states require a stand-alone geography course, six states offer a combined geography/social studies course, and geography is an optional course for graduation in four states.
two mandatory courses in geography (one in physical geography, one in human geography). In addition to these two courses, every school in Finland offers two extra courses, one on hazards and another on regional studies. The Finnish national curriculum will also be updated and implemented in 2016 in upper secondary schools. It has been decided that geography will have only one mandatory and three voluntary courses in the future.

Continental Europe exhibits even more variation in geography education. Primary geography education in the Netherlands (4–12 years old) is taught alongside history and biology under the guise of a “world orientation” curriculum. In lower secondary education (12–15 years old), geography is taught as a separate subject on par with the hours dedicated for science education (but less than given to languages, arts, and mathematics). The content focuses heavily on human–environmental relationships. In neighboring Germany, geography education bears a resemblance to the situation in the United States, with geography’s position in the curriculum varying significantly between different federal states. Some states within Germany promote a regional geography (most commonly environmental and physical geography), while thematic issues and topics are the focus in others. In Germany’s upper secondary schools, geography is affiliated with the social sciences in all of the federal states.

Asia

The status of geography education in Asia is also highly variable. In some Asian countries, requirements for school geography are on par or exceed those in Western societies. Elsewhere, it is difficult to find even cursory attention given to geography education. Formal geography education in India is generally neglected in schools. Where it is taught, geography is characterized by a highly didactic pedagogy, with little, if any, exploratory learning. In all but the most prestigious international schools and private schools, students are tasked with rote learning and memorizing disparate facts; little emphasis is given to teaching critical thinking and application skills. The state does not provide for teachers to deliver geography education that develops practices of analysis, application, and communication. For these reasons, an informal education sector has emerged to provide alternative educational experiences for youth and professional development opportunities for teachers across India (Solem and Balanchandran 2013).

The geography content in Japan’s primary and secondary schools is influenced by the government’s interest in promoting safety and awareness of natural hazards (Ohnishi and Mitsuhashi 2013). In high school geography, on the topic of natural environment, students learn about the geomorphological features of Japan and the ways Japanese cities and towns have worked to engineer the built environment to reduce the adverse impacts of earthquakes, tsunamis, landslides, and other natural disasters. Students learn the geographic features of the natural environment of Japan and their relationships to natural disasters and human responses to disasters. Students also study how human–environment interactions with regard to natural disasters vary across different regions in Japan, often by completing activities that make use of data, maps, and other geographic representations.

In Singapore, the national curriculum in geography from primary to pre-university level is described by a national syllabus document (Chang 2014). Geography is taught as part of an integrated social studies course at the primary level. At the lower secondary level, geography
is a compulsory subject, while it is an elective subject at the upper secondary and pre-university levels. One distinctive element of Singapore’s national syllabus document is the inclusion of desired outcomes of education (DOE), which are a broad set of attributes that define the aims of school education. Subject teachers are expected to correlate syllabus content and instructional methods with the four-part DOE, which are: (i) a confident person who has a strong sense of right and wrong, is adaptable and resilient, knows himself, is discerning in judgement, thinks independently and critically, and communicates effectively; (ii) a self-directed learner who takes responsibility for his own learning, who questions, reflects, perseveres in the pursuit of learning; (iii) an active contributor who is able to work effectively in teams, exercises initiative, takes calculated risks, is innovative and strives for excellence; and (iv) a concerned citizen who is rooted to Singapore, has a strong civic consciousness, is informed, and takes an active role in bettering the lives of others around him.

At grades 9 and 10, students in Singapore take the Singapore – Cambridge General Certificate of Education (Ordinary Level) Examination. During these grades students take a mandatory combined humanities syllabus, which is organized around two core ideas – “being rooted” and “living global.” The twin core ideas are delivered through six themes reflecting the topics from disciplines such as history, geography, political science, sociology and economics. The O-level geography elective syllabus is structured around three major themes: “our dynamic planet”, “our changing world” and “geographical skills and investigations.” The physical geography topics span plate tectonics, geomorphology, weather and climate, natural vegetation, and rivers and coasts. The human geography topics include agriculture, industry, tourism, and economic development.

Teacher education in geography

United States

Just as the fifty US states set their own school geography standards, so, too, do teacher certification requirements vary from state to state (Heffron 2013). Most colleges and universities offer an accredited teacher education program. At some institutions, pre-service teachers enroll in an academic degree program (subject major) at the same time they complete instructional methods coursework in the teacher education program. Other institutions require students to complete a traditional four-year undergraduate major followed by a fifth-year or master’s teacher preparation course of study. In either format, pre-service teachers are usually expected to complete a teaching practicum in a local school. Only a handful of US teacher education programs offer pedagogy courses specifically for geography.

Chronic teacher shortages and high rates of teacher attrition have led to so-called alternative teacher certification programs. These programs are aimed at practicing professionals who are interested in a career change to K-12 teaching. Alternative certification options are now available in nearly every state.

Europe

In Europe, the Bologna Declaration signed by the ministers of education in 29 countries in 1999 was designed to reform higher education systems in significant ways. Foremost among the changes was the adoption of systems of “easily readable and comparable degrees” at both the undergraduate and graduate levels, with the undergraduate (bachelor) level lasting normally three years and the graduate (master’s) level for two years. In addition, the comparable system of credits (the ECTS system) was established. These
changes resulted in enhancing mobility for students, teachers, researchers, and administrative staff of higher education institutions.

Implementation of the Bologna Declaration has proven especially challenging for teacher education because of the great variation regarding its structure, length, and required qualifications. Major reasons for this can be found from the fact that teachers are employed in national education systems, which can vary not only between the countries but also nationally. Bauer and Prenzel (2012) have shown, however, that despite this variation most of the countries that have signed the Bologna Declaration now have teacher education delivered in tertiary institutions, most often in universities. Some countries (e.g., Denmark, Hungary, Italy, and Luxembourg) have shifted their teacher education systems to the university level during the Bologna process and 38% of all the Bologna countries now require a master’s degree for at least all secondary teachers. Moreover, 20 countries have established new national standards and guidelines for teacher education (Bauer and Prenzel 2012, 1642).

The following examples, adapted from a report prepared for the GeoCapabilities initiative (Tani 2015), illustrate these diverse trends and patterns in European teacher education.

In the Flemish region of Belgium, teacher education for primary and lower secondary school is provided by professional bachelor institutions. At the lower secondary school level, students are required to select two subjects (e.g., geography and mathematics) and study a program that includes general courses on pedagogy, specific courses on subject education (subject didactics), and a third-year internship. For qualification at the upper secondary education level, students generally concentrate in a major subject and enroll in courses and internships leading to a master’s degree.

In England, becoming a qualified teacher once only required a course of study consisting of a three-year bachelor’s degree followed by a one-year postgraduate certificate of education (PGCE). That model has been criticized on grounds that the separation of subject content and pedagogical instruction made it too difficult for teachers to link theory, knowledge, and practice. This led to the introduction of alternative pathways to teacher preparation. As of 2014 in England, there are four major routes toward qualification.

1. The PGCE is one-year postgraduate certificate normally carrying 60 master’s level credits. The program is university-led and organized in partnership with local schools, in which trainee teachers spend around 60% of their time – usually in two blocks of teaching practice in two different schools.

2. “Teach First” is a one-year program starting with an intensive “summer school” before participants are allocated to often very challenging schools where they teach a heavy timetable from the start. Teach First, run by the independent body in partnership with universities, is designed to attract high performing and often very ambitious young graduates who are encouraged to see themselves from the start as “leaders.”

3. “School Direct” was introduced by the coalition government of 2010. From the trainees’ point of view it is almost identical to the PGCE, except he or she applies and is given a training place directly by a school. This is seen as a measure to wrest control from the universities, but it has had a slow start and is seen as a threat to standards.

4. “School Direct” (salaried route) is a variation that diminishes even further the role of the university element in training: the trainee is employed by the school from the start. This
route is an attractive option for those considering a change of career to become a school teacher.

Finland’s teacher education system for primary as well as lower and upper secondary schools is organized by universities. Students seeking to become qualified for these schools need to earn a master’s degree. In primary school teacher education, the students study educational sciences as their major discipline, while in subject education (to be qualified for teaching in lower and upper secondary schools) the students will most often have two subjects. For example, almost all the geography student teachers have a combination of geography and biology in their degree; one of these is studied as a major and another one as a minor subject. Unlike in many parts of the world, the teaching profession carries a good amount of social prestige and respect in Finland. This can be seen for example in the popularity of the studies: the intake in primary teacher education is only 10% of all applicants.

Teacher education for secondary schools in Germany is organized by the federal states; thus, there are a number of different approaches. Notwithstanding the differences, students are usually required to study two school subjects. In some federal states and universities one of these subjects has to be German, mathematics or a foreign language. In other federal states students can combine practically all subjects. This often leads to combined courses of study (e.g., physical education and geography). With the Bologna process teacher education became extremely diversified, making it a lot harder to move from one federal state to another. The reason for this is that some federal states did not change their teacher education system at all, but kept the Staatsexamen (wishing to keep state control over examinations). Others changed completely over to the Bologna system, but introduced different undergraduate and postgraduate degree systems. Still others kept the old system, but introduced modularized studies so as to better combine teacher education courses with courses for academic geographers. Independent of the Bologna process, there is a move to integrate more pedagogy and more practical studies into the first phase of teacher training. This has led to an influx of non-academic educators at universities (thereby lowering the costs of teacher education) and a marked reduction in subject studies. At the moment there is also a major move towards integrating modules on inclusion into all teacher-training courses, which further reduces the time for subject studies.

In the Netherlands, pre-service teacher training programs for the various types of school are part of higher education, some being provided at institutions of higher professional education (HPO), and some at universities. There are full-time, part-time and dual (i.e., work–study) HPO teacher training courses. There are also full-time, part-time and dual university-based training courses leading to a grade one secondary school teaching qualification for all levels of secondary education, including pre-higher education. These courses are open to university students and graduates only. The HPO institutions provide teacher training at both bachelor’s and master’s level. Universities provide training at master’s level only.

All teacher education in Sweden was integrated within the university/university college system in 1977. Two important changes on the structural level have affected the Swedish school system from the early 1990s. In 1991 there was reform which increased local autonomy for the municipalities and, secondly, a very open policy towards independent schools was introduced (1992). The first change meant that all teachers are employed by the municipalities or by independent schools. Government grants are given
but the system gives the municipalities or the independent school owners the right to decide on local priorities.

Since the social democratic government was replaced in 2006 by a center-right alliance, a massive reform program has been launched in the educational sector. Several new changes have been introduced to the school system, for example, a new education act (2011), new curricula (2011), new grading system (2011), and introduction of teacher certification (2013). Grading has been reintroduced in year 6 (2012), and national tests have been introduced in natural and social science subjects in years 6 and 9 (2013). In 2011, a new reformed system of teacher education was launched in Sweden. The previous common degree of Bachelor/Master of Education was replaced by four new professional degrees: pre-school education, primary school education, subject education, and vocational education.

Asia

India’s teacher education system pays only cursory attention to geography, if any at all (Solem and Balanchandran 2013). With the teaching of geography being left to the discretion of teachers, it is becoming more and more an ancillary topic in many schools. Teachers have long felt unprepared to teach the subject due to inadequate training. In recent years, informal approaches to teacher education have emerged in an attempt to promote geography in schools and support schools that wish to include the subject in their curriculum. This work has largely been led by The Indian Institute of Geographical Studies through a regional network of GeoVidyaa Centres located at the Army Public School in Bangalore, Nirmala College in Coimbatore, and the Cascade Montessori Resource Centre in Chennai.

In Japan, certification for primary school teachers is relatively simple: the “primary education license” allows anyone to qualify for teaching practice at the primary level (Ida et al. 2015). In the teacher education curriculum in Japanese universities, almost all students select some specialty subject, such as mathematics, social studies, or comprehensive studies. Student teachers who wish to acquire a social studies teacher’s certification must major in a social studies-related discipline, such as geography, history, and civics. This means that teachers who elect a discipline other than geography may require additional training and preparation to teach the geographical content of the curriculum (a situation which mirrors that of the United States).

The National Institute of Education at Nanyang Technological University in Singapore is that country’s sole provider of a geography teacher education program. Initial teacher preparation in Singapore consists of three possible pathways: primary, secondary, and junior college. Each provides a foundation in pedagogy and specialized subject knowledge in at least one discipline. Programs offered include a two-year diploma in education (for primary education), a four-year bachelor of arts or science in education, and a 1–2 year postgraduate diploma in education. Components of the degrees include coursework in subject areas, education, language and communication, character development, and field experience (school practicum).

Research, curriculum and professional development resources

There are many international examples of resources, curriculum materials, conferences, and professional development programs for geography educators and researchers (Table 1).
Table 1  Selected international programs and resources for primary and secondary geography.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGU-CGE Declaration for Research in Geography Education. A statement advocating a research culture and agenda for improving the quality of geography educational practices internationally.</td>
<td><a href="http://www.igu-cge.org">www.igu-cge.org</a></td>
</tr>
<tr>
<td>GeoCapabilities. A project researching the purposes and values of geography in schools.</td>
<td><a href="http://www.geocapabilities.org">www.geocapabilities.org</a></td>
</tr>
<tr>
<td><strong>Professional development</strong></td>
<td></td>
</tr>
<tr>
<td>iGuess (EU). Training teachers in the use of GIS for teaching multiple subjects.</td>
<td><a href="http://www.iguess.eu">www.iguess.eu</a></td>
</tr>
<tr>
<td>School on the Cloud (EU). Promoting digital citizenship through cloud-based educational applications in schools.</td>
<td>schoolonthecloud.eu/index/</td>
</tr>
<tr>
<td>I-Use (EU). A program supporting the teaching of statistics in different subject areas.</td>
<td><a href="http://www.i-use.eu">www.i-use.eu</a></td>
</tr>
<tr>
<td>Esri-ConnectED (US). An initiative to support the implementation of cloud-based GIS technology in schools.</td>
<td><a href="http://www.esri.com/connected">www.esri.com/connected</a></td>
</tr>
<tr>
<td>Geography: Teaching with the Stars (US). A hybrid approach to preparing geography teachers through face-to-face workshops, videos, and online forums.</td>
<td><a href="http://www.geoteach.org">www.geoteach.org</a></td>
</tr>
<tr>
<td><strong>Publications</strong></td>
<td></td>
</tr>
<tr>
<td>International Research in Geographical and Environmental Education</td>
<td><a href="http://www.igu-cge.org/publications.htm">www.igu-cge.org/publications.htm</a></td>
</tr>
<tr>
<td>Review of International Geographical Education Online</td>
<td><a href="http://www.rigeo.org">www.rigeo.org</a></td>
</tr>
<tr>
<td>Journal of Research and Didactics in Geography</td>
<td>j-reading.org/index.php/geography</td>
</tr>
</tbody>
</table>
One of the leading world organizations for primary and secondary geography education is the International Geographical Union’s Commission on Geographical Education. The IGU-CGE sponsors conferences and symposia in different parts of the world, often in conjunction with the IGU Congress. The Commission is also pursuing a broad range of initiatives aimed at improving research capability and capacity, supporting international curriculum projects, an international geography Olympiad, and charters for geography education and sustainable development and cultural diversity.

National organizations such as the Geographical Association in the United Kingdom, the National Council for Geographic Education in the United States, and the Geographical Education Society of Japan offer professional memberships for primary and secondary teachers, teacher educators, and geography education researchers. Through their annual conferences, publications, and special projects, geography teacher associations such as these provide important opportunities for continuing education and professional development. Similarly, many of the professional associations listed in this Encyclopedia’s appendix have internal specialty groups and task forces that focus on issues of primary and secondary education.

There is also growing international cooperation in the area of geography education research. One influential US report originated in 2013 called for broad-scale improvements in geography education research, including international collaborative research. That report, issued by National Geographic’s Road Map for 21st Century Geography Education project (Bednarz, Heffron, and Huynh 2013), presented a research agenda framed by four key questions.

1. How do geographic knowledge, skills, and practices develop across individuals, settings, and time?
2. How do geographic knowledge, skills, and practices develop across the different elements of geography?
3. What supports or promotes the development of geographic knowledge, skills, and practices?
4. What is necessary to support the effective and broad implementation of the development of geographic knowledge, skills, and practices?

The road map project spawned subsequent research projects in the United States and also had an influence internationally. A National Center for Research in Geography Education, established in 2013 by the Association of American Geographers and Texas State University, is coordinating research activities of networks of geography and STEM education researchers affiliated with over 50 universities in the United States and several international institutions. Some of these initiatives involve research collaborations with universities abroad as well as with the state Geographic Alliances: GeoProgressions, a capacity-building project for learning progressions research, and GeoSTEM (GeoSpatial Teaching Enrichment Modules), a pilot project that is supporting the Esri-ConnectED initiative by creating ArcGIS online-based resources, materials, and tools for STEM teacher education programs.

The IGU Commission on Geographic Education is using the road map project report as a reference for designing an International Declaration on Research in Geography Education. This declaration, which was released in 2016, assesses the international need for more and better quality research in geography education. One multiyear research project, GeoCapabilities, is active in several countries across Europe, Asia, and North America. The project’s research component is gathering evidence of how teachers in different countries understand the value of
geographic knowledge and thinking in terms of its capacity for human development over the lifespan. As a leading example of an international comparative research project in geography education, GeoCapabilities may well offer a model for other disciplines to follow in researching the purposes and values of their subject in the context of primary and secondary education.

Acknowledgements

Multiple authors contributed content on a country-by-country basis for the GeoCapabilities report: David Lambert (England), Tine Béneker (Netherlands), Anke Uhlenwinkel (Germany), Gabriel Bladh (Sweden), and Luc Zwartjes (Belgium).

SEE ALSO: Geography education: primary and secondary

References


Further reading


The fact that everywhere on Earth is different makes geography both a fascinating and challenging field of study for students. The description and analysis of spatial variation in Earth processes and human activities underpinned the lectures and publications of Alexander von Humboldt (1769–1859) and Karl Ritter (1779–1859) in Germany, and laid the foundation for the work of innovative educators such as Arnold Guyot (1807–1884) and William Morris Davis (1854–1935) in the United States.

Geography has been among the disciplines at the forefront of innovation in teaching and learning since the early twentieth century. The first discipline-based academic journal devoted to teaching and learning was the *Journal of Geography*, first published in 1902. Geographers were early adopters of objective assessment in the 1920s, and also of well-defined learning objectives and active learning in the 1990s.

Contemporary teaching and learning in geography is informed by a significant body of research on the ways that people learn and the most effective ways for students to achieve their learning objectives. The scholarship of teaching and learning geography has gained considerable traction in recent years; many instructors now research not only their discipline area but also their classroom practice (Day 2012). The result is a much better understanding of how to design teaching activities to encourage engagement by students. Research findings have become influential through blogs, academic conferences, and postgraduate teaching qualifications. Journals such as the *Journal of Geography*, the *Journal of Geography in Higher Education*, and *International Research in Geographical and Environmental Education* are respected outlets for peer-reviewed research on geographical education. Some mainstream geography journals, such as *Area*, the *New Zealand Geographer*, and *Progress in Physical Geography*, also occasionally publish papers on geography education and/or have a section on it.

Teaching and learning is increasingly recognized as an important activity in its own right, not just an adjunct to discipline-based research. Geographical societies around the world have prizes, awards, and honors that recognize excellence in teaching, such as the AAG Gilbert Grosvenor Geographic Education Honors, and Distinguished Teaching Honors.

Learning and doing geography is an important way for people to interact with the world, to see the world differently, and to structure and think about that information. Thus, education is not just the acquisition of skills and information but, as John Biggs (2003) said, it is about conceptual change. The aim of geography education is to produce graduates who can think, act, and practice as geographers. To develop such graduates it is necessary to have good alignment between the intended learning outcomes, the teaching methods, and the assessment regime using a framework called “constructive alignment”
(Biggs 2003) in which course design begins by considering what students should be able to do by the end of the course. This determines the assessment regime and the teaching and learning methods (so-called outcomes based or backwards design). Ideally, curriculum mapping occurs, with desired outcomes mapped across all courses in the degree program to ensure that key outcomes or attributes are developed and assessed.

A range of teaching approaches can be used, providing there is good alignment with the intended learning outcomes and the assessment regime. However, it is known that some teaching approaches can be particularly powerful in terms of promoting engagement and developing ways of thinking and practicing. Signature pedagogies in geography include fieldwork, visualization, map-making, and map interpretation (Komoto 2009), and integrative learning, learning through inquiry, and service learning (Spronken-Smith 2013). Ten other high-impact education practices identified by George Kuh and colleagues (Kuh 2008) include service learning and undergraduate research, first-year seminars and experiences, common intellectual experiences with required core course, learning communities, writing-intensive courses, collaborative assignments and projects, diversity/global learning, internships, and capstone courses and projects. Many of these practices are commonly used in geography curricula to prepare students for citizenship, work, and life (Spronken-Smith 2013). Increasingly, there has been realignment of the curriculum to focus more on skill development and the needs of employers (see entry Geography education, workforce trends, twenty-first-century skills, and geographical capabilities).

Good geography instructors encourage students to take a deep approach to learning, promote active learning and engagement, use an outcomes-based or backwards-design approach, and incorporate research experiences into the curriculum. Although there are many possible routes to attaining learning outcomes, many demonstrably successful pedagogies and teaching approaches are constrained by infrastructure, institutional norms, and union-management negotiated labor agreements. In practice this means that new approaches commonly have to be incorporated into traditional frameworks of lectures, laboratory/practical work, and fieldwork. Fortunately, good practical advice is available (Solem and Foote 2009) and new pedagogies are supported by new technologies.

### Lectures: technological and pedagogic change

The traditional lecture is rapidly changing in universities and colleges. Mick Healey, Alan Jenkins, Sue Vajoczki, L. Dee Fink, and other geographers have led the way in a refocus of the lecture on active learning, with an influence that has extended beyond the discipline. It is now widely recognized that students cannot focus on a lecturer speaking for long periods. Activities such as group discussions and think–pair–share exercises are widespread. In-class tests can be undertaken as group work, an approach that most students find engaging and effective.

Technology plays a major role in the modern lecture. “Clickers” or “personal response systems” (PRS) are now used in many lectures. Originating from game shows, the clickers (or smartphones) allow students to anonymously key in responses to questions with responses captured and displayed instantaneously on a screen. Clickers can be used to gauge background knowledge on a topic, to immediately test understanding of the lecture, and to gauge viewpoints on polarizing issues. Students enjoy
using them and are more engaged. Some lecturers also have live-streaming of twitter feeds (or other social media) and use digital whiteboards to promote participation and generate questions and comments.

Technology is usually a key tool in the “flipped” or “inverted” classroom. Typically, the lecturer generates a video or audio recording of lecture material, or other digital resources on a learning management system or other digital platform. The expectation is that students will access this material online in advance of their “lecture” session. The lecturer then facilitates discussion of key concepts and skips over easy material that the students have already mastered before the lecture. Some lecturers use a “just-in-time” method, whereby the students’ grasp of the background material is tested by an online quiz or questioned by asking the student to volunteer their uncertainties and connections to the topic. The testing and questioning is usually based on reading of a textbook or paper and is undertaken just prior to the lecture. By uncovering common misconceptions, misunderstandings, and topics of particular interest, the lecturer uses the lecture to address relevant points.

Lectures are also increasingly taking place in an exclusively online environment. Massive open online courses with lectures developed by geographers such as Anthony Robinson and Todd Bacastow (Pennsylvania State University), and Brent Hecht (University of Minnesota) have attracted tens of thousands of participants. Online lectures and webinars have also become important teaching tools in the teaching of geographic information system (GIS) software (see entries on Geographic information science and technology: educational directions and Geography education: digital and online trends).

Many lecturers are now wedded to a constructivist theory of learning. This means that they see their role as a guide on the side – rather than the sage on the stage – or even a doc on the laptop. Instead of just talking, lecturers work to get students engaged in their learning. Although the form of the lecture is changing, the lecture itself is likely to continue its importance for the foreseeable future.

Laboratory/practical work: new technologies for more complex problems

Although understanding geography is valuable for intellectual development, doing geography is the difference between knowing about geography and being a geographer. Geographic skills are taught in the context of laboratory and practical classes, and involve map analysis; statistical methods; calculation of the rates and quantities of environmental, social, and economic properties; and spatial modeling.

Maps have long been synonymous with geography and are still an essential component of all geographers’ education. However, paper maps, protractors, rulers, compasses, and dividers are now virtually obsolescent, and along with that obsolescence there has been a decline in the importance of learning tedious calculations of map scale, area, gradient, and magnetic declination. At the same time, the widespread use of GIS and the availability of global navigation satellite systems, such as GPS on smartphones, has meant that there has been a rise in the importance of geodesy, coordinate systems, map projections, and spatial modeling in student practical work.

Descriptive and inferential statistical methods have been important since the 1960s. First taught with logarithmic and statistical tables and small hypothetical datasets, most statistical exercises now rely on larger, more relevant real-world datasets analyzed with user-friendly software. Spatial statistics are now much more accessible to students in the context of GIS.
Current or even real-time data are now readily accessible on the Internet and are used in laboratory exercises. These datasets include extensive meteorological, hydrologic, and oceanographic data. Other data, such as digital elevation models and socioeconomic data, are also freely available. Students are also able to obtain their own data, which they can quickly download for analysis, for example from a GPS receiver or smartphone. The use of digital data has now largely supplanted air photo interpretation based on stereoscopes. Remote sensing imagery, such as Landsat and European Space Agency data, may be freely downloaded, and many jurisdictions provide GIS and remote sensing data and imagery that are available online. Images can be processed with free software such as Multispec and LEOWorks. Students also generate their own remote sensing imagery with balloon kits and inexpensive cameras.

Fieldwork: opportunity and challenges

In an educational context, fieldwork is any organized teaching and learning activity that takes students out of the lecture hall into the environment that is being studied. The value of geography fieldwork has long been self-evident to participants (Boyle et al. 2007). Practical skills developed and practiced in fieldwork include the ability to work safely and efficiently as part of a team in pursuit of a defined goal. Technical skills, such as sampling, the use and care of scientific instruments, socioeconomic surveys and other methods of data collection, and the accurate recording of results, are developed, along with associated skills such as report writing, the application of statistical techniques and analysis, and the preparation of maps and diagrams. Fieldwork enables students to make sense of the world and to better understand and critically evaluate theories that explain it.

Contact with the field environment promotes more than just a cognitive function; it also provides a sensory experience that associates an environment with long-term learning. Fieldwork promotes positive attitudes and values toward the environment and, therefore, enhances both cognitive and affective learning. Social relations between students and instructors developed during fieldwork enable instructors to better appreciate the challenges faced by students and to mutually develop teaching approaches that enhance learning. Conversations with individual students in the context of fieldwork also enable an instructor to promote and develop higher-level learning by the student.

Fieldwork has traditionally been more important in Europe and countries of the British Commonwealth, than in the United States. Many European, UK, and Australasian universities require one or two residential field trips for all their undergraduate students. In the United States many universities offer residential field courses but they are not always mandatory and may be unattractive to students because of scheduling conflicts with summer employment opportunities. In the United States and elsewhere many faculty members also have concerns about the legal liabilities arising from the risks of fieldwork and, as a consequence, the importance of fieldwork in the geography curriculum has declined in many institutions in the United Kingdom, Australia, and Canada. However, this reduction in fieldwork is not universal and some geography departments have expanded fieldwork within the curriculum.

Concerns about the risks of fieldwork are legitimate but the risks can be mitigated. Safety concerns require that expectations and responsibilities be made clear to everyone, preferably through a code of conduct that is discussed prior
to departure. In some cases, students participate in the development of safety and conduct guidelines. Site-specific issues may be discussed at or adjacent to a site. Every field trip needs a qualified person to deal with major or minor accidents, and appropriate first aid equipment to be carried. The extent of the training and available equipment depends on a prior risk assessment that takes into account the environment, physical abilities of the students, and access to medical assistance. Access to medical assistance in areas of poor telephone coverage may involve the use of radios.

Safety issues may be different for students with disabilities of various types. Issues of hearing, sight, mobility, heart and lung diseases, allergies, agoraphobia, and other conditions all pose different challenges for both the student and instructor. In some cases an initial fieldwork plan that assumed a site was accessible may need to be modified to ensure that the work is accessible to all students. In some cases a student with a disability cannot be accommodated and an alternative program may be required. Arrangements of this type are usually made through the college or university office that specializes in dealing with the learning challenges of students with disabilities.

Although cost and health and safety considerations are considered to be barriers to the implementation of fieldwork, one of the biggest barriers is the way that curricula have been designed. In many cases the curriculum is so content-rich that there is little room for even short local excursions.

Despite the barriers there have also been issues with overuse of certain areas for fieldwork. Some areas of the United Kingdom, such as Dartmoor and Malham in the Yorkshire Dales, have been subject to intense environmental pressure by field classes. However, there is a risk of damage to field sites anywhere. This has included indiscriminate hammering of outcrops, unauthorized collection of samples, damage to fences, and impacts on livestock through disturbance and escape of livestock onto roads. The Geographical Association and Field Studies Council in the United Kingdom have developed a code of conduct for fieldwork, and many universities and colleges have their own guidelines.

Concerns about the risks of fieldwork have inspired interest in self-guided audio tours and have also resulted in some interest in virtual fieldwork, that is, the development of investigative case studies based on virtual worlds. In some cases this virtual fieldwork is a complement to a visit in the real world, but it can also function as a substitute. However, even with immersive technologies, the virtual world fails to provide students with connectedness to nature. In fact, there are broader concerns that students are increasingly alienated from nature. This has been framed as “nature deficit disorder” (a term used by Richard Louv in his 2005 book, *Last Child in the Woods*). Although the term is more of a metaphor than a medical diagnosis, Louv draws attention to numerous health benefits associated with time spent in nature. Edward O. Wilson’s biophilia hypothesis is also a framework for attachment to nature, the place that is truly “home” for people. These societal contexts and the fact that fieldwork is usually a social group-based activity means there are good sustainability arguments for an enhanced role for fieldwork in the geography curriculum. Combined with the role that fieldwork plays in discipline-based learning, the position of fieldwork in the curriculum should be unassailable.

Experiential and service learning: making geography relevant

Experiential learning can include the assignment of students to employers for internships
or cooperative education work terms during their degree. In the United Kingdom, the term “sandwich course” is more or less synonymous with “cooperative education.” The student is given academic credit, a salary, and an opportunity to apply what they have learned in their education to the “real” world of employment. In this experiential learning process the student connects their life as a student to their subsequent work life. Sometimes students succeed admirably and obtain long-term employment from the employer. In other cases there is a mismatch between the student’s abilities and interests, and the needs of the employer. This too is learning.

Student experience of cooperative education and internships is generally positive. Areas of student concern include dissatisfaction with a lack of constructive performance feedback, the need for more time on the job, and a need for better training (Jackson and Jackson 2009); particular attention should be paid to these issues to implement a successful program.

The desire to produce graduates who are relevant to society has also given rise to service learning, where students may assist a community group or project. In contrast to cooperative education, the student is not paid for their work, and the work assignment is always to a not-for-profit organization for some social good. Service learning often involves groups of students, rather than single students. Student service learning benefits the community, assists the student in generic skill development, and also has the aim of producing graduates who are global citizens and agents of social justice. Involvement in service learning not only helps the student develop a host of desirable generic skills and insights into social problems but also has been recognized as a high-impact educational experience leading to higher retention rates (Kuh 2008). The community partner should also benefit, but in some cases the community benefits are apparently modest and have not been investigated as closely as the benefits to students. Oldfield (2008) advocates for service-learning projects that are built around community skill, knowledge, and experience.

Issues to consider with the implementation of service learning include whether it is a compulsory part of the curriculum, whether it involves the entire class or small groups working with agencies or organizations, how best to assess it, evaluation of the impacts of the program, and the nature of the relationship between the university and the wider community.

In general, geography curricula have embraced service learning more than cooperative education. Within many geography departments there may be a quiet tension between those who support a notion of geographic education as a process of intellectual development and those who see geography as a route to career opportunities. Service learning supports both perspectives, but cooperative education has been seen by some as leaning too far toward career development as a fundamental goal of a geographical education (Zhou, Smith, and Spinelli 1999). In contrast, geographers were among the pioneers of service learning in the academic community from the early 1990s.

Undergraduate research and inquiry: moving the subject forward

Undergraduate research reinforces linkages between geographic research and the teaching of geography (Healey 2005). Undergraduate research has been particularly important since the Boyer Commission in the late 1990s noted that many undergraduates complete their degrees unaware of the research culture of the institutions in which they studied. Undergraduate research equips students with transferable skills, better preparing them for future postgraduate study.
Students are participants

<table>
<thead>
<tr>
<th>Research tutored</th>
<th>Research based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engaging in research discussions</td>
<td>Undertaking research and inquiry</td>
</tr>
</tbody>
</table>

| Emphasis on research content | Emphasis on research processes and problems |

<table>
<thead>
<tr>
<th>Research led</th>
<th>Research oriented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning about current research in the discipline</td>
<td>Developing research and inquiry skills and techniques</td>
</tr>
</tbody>
</table>

| Students frequently are an audience |

Figure 1  The links between curriculum design and the research–teaching nexus. Source: Healey and Jenkins (2009). Reproduced by permission of the authors.

and/or the workplace. Undergraduate research shifts the student from audience to participant in the teaching–learning process. Healey and Jenkins (2009) presented a framework showing the link between curriculum design and the research–teaching nexus (Figure 1) and argued that more emphasis should be placed on research-tutored and research-based pedagogies, as these models have the most benefit for student learning.

The discipline of geography has been well known for innovation and engaging pedagogies, and for many years has embraced the idea of students doing research (Figure 2). It is commonplace for geography undergraduates to undertake a range of research and inquiry activities during their degree. Activities may range from a major project in an honors degree program to smaller, shorter activities. For example, Spronken-Smith et al. (2008) describe in-class activities with questions such as “Where might sand-dunes be on Mars?” or fieldwork in groups to assess the impact of hazards on local communities, or international student collaborations working on geographic problems based around topics such as population, the global economy, and nationalism in order to promote understanding of geographic concepts and international perspectives. Paid employment as student research assistants also provides opportunities to participate in research.

Some programs have courses designed around inquiry, with students working in groups over an entire semester or longer on one problem, while others design their whole program to progressively develop inquiry skills in their students. Thus the term “inquiry” can encompass a range of activities but, importantly, all are student-centered, with the students active in their learning and the instructor acting as a facilitator. Students benefit from greater exposure to learning through inquiry (e.g., progressive development of inquiry skills throughout a degree).
and open, discovery-oriented inquiry offers the greatest benefit for developing higher-order learning outcomes and a host of other generic skills.

Textbooks: going online

Although advanced courses in geography are typically based around readings from the research literature, most lower-level undergraduate courses are built on, or are supported by, textbooks. Modern textbooks have their origins in works such as the *Geographia Generalis of Varenius*. The 1672 edition of *Varenius* (edited by Isaac Newton) was particularly influential and was translated into several European languages. Textbooks expanded rapidly in the nineteenth century with the *Kosmos* of Alexander von Humboldt (1845–1862), the *Physical Geography* of Mary Somervell (1848), and works on physiography by Eugene Cortambert (1836) and Thomas Huxley (1877). Human geography textbooks developed in the twentieth century with the *Principles of Human Geography* by Huntington and Cushing (1921) and by Paul Vidal de la Blache (1923). Later textbooks such as Arthur Strahler’s *Physical Geography* (1951), *Spatial Organization* by Abler, Adams, and Gould (1971), and *Geography: A Modern Synthesis* by Peter Haggett (1972) both reflected and supported later developments in the discipline. By the end of the twentieth century the revision cycle for textbooks had shortened, regional editions became more common, chapter structures were increasingly nonlinear as publishers demanded “features” in boxes within the chapters, and instructor support materials, student manuals, and online support material became ubiquitous. Some recent textbooks emphasize the role of learning theory in textbook organization.

Textbooks are increasingly free of the constraints of print formats, with electronic versions of traditional textbooks produced by commercial publishers, together with free online textbooks self-published by several authors, for example the introductory physical geography texts by Michael Pidwirny (http://www.physicalgeography.net/fundamentals/contents.html) and the interactive text of Michael Ritter (http://www.earthonlinemedia.com/ebooks/tpe_3e/title_page.html). The Saylor Foundation free online courses include an introductory undergraduate course in world regional geography with a free textbook available under an attribution-noncommercial-share alike license (http://www.saylor.org/site/textbooks/World%20Regional%20Geography.pdf).

Although the move toward electronic textbooks has been a slow one, the trajectory is clearly moving in that direction. Commercial publishers, such as Pearson Prentice Hall,
McGraw-Hill, and W.H. Freeman/Macmillan offer learning management systems in support of textbook content. The development of new digital content, including videos, animations, interactive components, and the ability of instructors to choose content, means that there has been an inevitable blurring of pedagogy and textbook materials.

Conclusions

The last decade of the twentieth century saw a massive shift in geographic pedagogy. The advent of active learning approaches and technology in the classroom, a greater interest in applied aspects of the subject, and enhanced focus on undergraduate research and experiential learning all took place in a broader context of the growth of social networking and increased government demands for accountability in terms of graduate employability and system efficiency. These changes have impacted the relationship between teaching and research in universities, as well as the relationship between universities and their communities.

The next few decades will likely continue to see innovative, technology-driven pedagogies that will evolve in a way that cannot be predicted with any great degree of confidence. However, there will probably also be consolidation of advances made in the past two decades. A wide variety of research questions need to be answered (Day 2012). Among the more pressing topics are: Do students learn particular knowledge and skills better by inductive or deductive approaches? How can virtual field trips be improved? How can the fieldwork experience of students with disabilities be enhanced? What is the effect of class size on the effectiveness of different teaching methods? Does subject research by faculty enhance student learning, and, if so, in what way?

One likely trend in the coming decade is the development of more rigorous testing of innovative approaches. Broader student diversity means that it is almost impossible to provide universally applicable and definitive conclusions on any intervention; therefore, studies will probably become more focused and less likely to purport to be the final word on any particular pedagogy. The more studies that are undertaken, the more confidence there will be in the results, which may increase the rate at which innovation is diffused. In order to accommodate an increase in the number of research studies, there will be a need to either increase the frequency or size of existing journals and/or introduce new journals.

SEE ALSO: Environmental education; Fieldwork in human geography; Geographic information science and technology: educational directions; Geography education: community colleges; Geography education: digital and online trends; Geography education: promoting diversity and broadening participation; Geography education, workforce trends, twenty-first-century skills, and geographical capabilities; Geography in higher education

References


Further reading


The current position of geography in higher education

Geography is part of the higher education landscape in most countries. Growth in the discipline has taken place through a rich mix of structures including dedicated departments, multidisciplinary units, and combined programs. Geography is also encountered in a variety of institutions. For example, it might be encountered in traditional universities, dedicated geoscience institutions, modern universities (former polytechnics and colleges in the United Kingdom) and in further education colleges/community colleges where selected courses are taught at higher education level or where foundation or associate degrees are taught. Some higher education providers are state run or funded whereas others are private (fee paying).

There has been a rise in the number of universities developing overseas campuses, particularly in Asia. These are universities where English is not the first language of the country but in which courses are taught in English. Additionally, geography is sometimes part of international educational consortia and reciprocal mobility agreements. All these are responses to increased marketization of higher education. It is no wonder then that the study of geography is characterized by a diversity of provision, state and commercial drivers, content and perspectives, and methods and conceptual frameworks.

Within this rich tapestry it is possible to specialize in either human geography or physical geography within subdisciplines of these or to take an integrated degree course.

The breadth of content in physical and human geography has had impacts upon higher education provision: first, in some countries an increasing specialization and separation of the disciplines has taken place over time, although this has been tempered where an integrated geography is still taught; second the breadth of the discipline has created a demand for taught MA courses that allow for more specialist or vocational skills to be developed. However, the evolution of different geographical traditions has a pattern of its own and the content and scope of the discipline remains contested and open to productive dialogue as well as dynamism. For example, standard geography departments in universities in the Soviet Union (USSR) include specializations in physical geography, economic geography, and cartography. Areas such as cultural geography are largely confined to the United States, Europe, Australia, and New Zealand. Entire institutions dedicated to the geosciences have developed in China.

The relationship between geographic research and teaching in institutions providing geography are also varied; some specialist teaching institutions exist and research intensive universities usually also have some teaching activity, although some have limited undergraduate places; however, the majority of providers offer a blend of teaching and scholarly and research activity. Unsurprisingly, topics, theories, and methods vary in both research and teaching
GEOGRAPHY IN HIGHER EDUCATION

internationally as well as the degree to which each informs the other.

During the 1950s to 1970s the number of geography departments around the world grew significantly and in North America GIS provision grew rapidly. In the United States geography developed in close connection with cartography, the links to GIS still being apparent to a greater extent in, for example, the United States than in Europe.

From 2000 onwards there has been a global diversification of geography provision in universities with a huge variety of courses, combinations, interdisciplinary projects, and collaborations. There has been a clear trend toward managing geography within larger multidisciplinary units as seen currently in many British and Australian universities, so although the name geography persists as a “unit,” “division,” “subject area,” or “department,” there has been a merging of units for cost saving purposes or attempts to encourage cross-disciplinary collaboration. One reason for this may be the “spatial turn” in the humanities and social sciences which has allowed other disciplines often closely allied to geography to view space as an essential element in their own research and teaching agendas. Geographical patterns and processes and the significance of space and place have been highlighted by geographers such as Doreen Massey and space–time ideas have been widely adopted in other disciplines, particularly in the social sciences. The question of whether geography’s relationship with other fields of study will become further subsumed within interdisciplinary departmental structures as a result remains to be seen, but this process is taking place in some countries where the imperative to research in interdisciplinary teams in order to address global problems is paramount.

The overall political and economic context of higher education continues to evolve with cuts in some areas and expansion in others, but a positive growth trend for the discipline is currently in evidence and a wide range of courses available.

The range and diversity of geographical courses

The study of geography has moved from traditional single subject departments and schools (generally geography remains a distinct department or school in the universities where it has been established the longest) to departments of environment/earth sciences/society, enterprise, and environmental studies. The move to modular courses in the late twentieth and early twenty-first centuries across many universities has further encouraged the interdisciplinary connection of geography with other subjects (Johnston 2004).

Geography draws from and contributes to a wide variety of other subjects from the natural sciences, social sciences, and humanities. As a result students may choose to combine geography with other subjects and vary the weighting of this combination either as a major/minor or an equal combination. The length of degree courses varies significantly as does the degree of specialization and choice of content. Degrees in geography can be awarded as a Bachelor of Arts (BA) or Bachelor of Science (BSc). The demand for postgraduate degrees has created a diverse market for specialist masters courses. Differing amounts of independent research are integrated into masters and doctoral courses in different institutions. Two-year community (or technical or junior) colleges in the United States offer associate’s degrees, an alternative to the traditional four-year bachelor’s degree programs. These institutions offer geography, among a wide range of subjects, to postsecondary students of all ages and academic levels. Students studying at a community college
may either receive a standalone two-year qualification, the associate’s degree, or transfer to a four-year institution to complete the final two years of the bachelor’s degree. In the United States just over a third of students who completed a degree at a four-year university in the academic year 2012–2013 had previously been enrolled at a two-year institution, however, only selected universities and colleges share articulation agreements allowing these credit transfers to take place. In the United Kingdom foundation degrees in partner colleges are completed as “honors” degrees in universities reflecting an outsourcing of foundation level work which is completed at a lower cost for students. In the United Kingdom most students are working toward an honors degree, completed typically over three or four years, whereas in the rest of Europe degrees can take much longer as students finance themselves part time and students tend to start degrees later. Certificates are also awarded in specific programs that typically require less than two years of study.

At the graduate or postgraduate level, degrees also vary. There is the MA/MS, the PhD/DPhil and a host of advanced degrees and attainments (such as the habilitation in some countries). Many of these advanced degrees are unique to their national setting. In recent years there have been efforts to standardize these qualifications (see discussion of neoliberal trends below), often employing the US model of a two-year taught MA/MS including a research thesis or professional report and 3–5 year PhD degree. It is the MA/MS degree where much growth is occurring at this time. Standardization at this level has been initiated in the European Union via the Bologna process. In the United States institutions offering master’s courses but not doctoral study are termed “master’s only” which has become a defined Carnegie category. The growing number of master’s courses represents a highly dynamic part of higher education (HE). However, this growth and development has not gone on unquestioned. Several articles have proposed the “rethinking” of postgraduate education (e.g., Monk, Foote, and Solem 2012).

Postgraduate provision

As the proportion of people with undergraduate degrees increases, the demand for master’s courses is rising in response to the need to distinguish oneself in the graduate labor market and master’s degrees provision is expanding rapidly. International mobility has made the taught master’s degree (usually one year, sometimes two) a globally competitive marketplace and the degree can serve multiple functions: as a stepping stone to doctoral research; as a vocational course; for interest from lifelong learners; a means to change career track; allowing students to specialize beyond a broad degree; and an opportunity to apply a geographical background to an interdisciplinary research area. Master’s courses are more closely aligned to career opportunities than the degree or PhD. The pedagogic challenge associated with this is the requirement to bring a diverse group of learners together to cover the same material when their backgrounds and previous experience may be significantly different. As master’s courses increasingly become a gateway to doctoral study the increased diversity of learners may help to address concerns with a lack of diversity at doctoral level. Fairly traditional models are still being adopted in master’s courses although some include projects, placements, online and blended elements, and use practitioner teaching staff. The traditional research-based master’s degree is being replaced with a mixture of taught modules that cover soft skills as well as disciplinary content, however, a research project is usually the culminating experience.
Boyle, Foote, and Gilmartin introduced a special issue of GeoJournal in 2014 on “Rethinking the PhD in Geography.” It outlines current challenges and new models for a changed trajectory for this degree type. The key challenges are the lack of diversity of PhD candidates, the narrow training that the degree provides, poor retention and completion rates, extended deadlines, and the dangers of corporate sponsorship reorienting the degree. Most PhD students come from a privileged group and this situation is worsening with the growth of self-funding. One reason for this may be the narrowness of training provided during a PhD, which is often aimed at an academic career for which the supply of doctoral students far outstrips demand. Indeed the casualization of the labor force has created many “teaching only” contracts for which a PhD is a requirement but does not provide time for ongoing research, resulting in a skills mismatch. Many PhD students experience difficulty in gaining teaching experience, although in some institutions it is embedded in their training. There is a general lack of consistency across institutions in this respect. A careful balance needs to be struck between overloading and exploiting PhD students with teaching related work and providing carefully scaffolded opportunities for them to build a portfolio of experiences, for example, in marking student work, being a fieldwork assistant, laboratory demonstrating, leading tutorials, and lecturing. While corporate sponsorship of PhDs may be seen as an opportunity to diversify training and shift it toward more vocational and skills oriented objectives, this could be at the expense of critical reflection. The authors suggest that the “market model” may create more teaching opportunities in the future but this could further contribute to the precariousness of a casualized labor force in HE.

Alongside the supervisory relationship, complementary classes and training opportunities are being provided to enhance doctoral education and student-led activities, such as postgraduate conferences, provide opportunities to present work in a supportive environment as well as to develop skills in event organization and committee work – the RGS postgraduate mid-term conference is one example. Another discipline specific initiative is the inclusion of doctoral students in the writing groups of the INLT so that PhD students in geography get to meet international colleagues and co-author journal articles. The EDGE project has produced research articles and books, conference mentoring, and career planning resources among other support initiatives for graduate students. Other multidiscipline research training initiatives come from foundations such as the US National Science foundation’s research traineeship program and German research foundation’s graduate research training groups as well as foundations encouraging underrepresented groups into science. The United Kingdom’s Higher Education Academy accredits university teaching certificates for graduate students and awards associate fellow status for those who successfully complete this.

Geography, like all disciplines in higher education, is being shaped by a number of broad-based economic, political, social, and technological
changes at the national and international levels. Online education, interdisciplinary imperatives, competition, marketing, and rebranding have all impacted on the nature of courses being provided at undergraduate and postgraduate level.

There is a desire for more students to receive college degrees more rapidly at less cost and online education seems to provide a possible solution (Erickson 2012). MOOCs (Massive Open Online Courses) are one way in which institutions are testing the appetite for online learning, although completion rates are low in some cases.

Movement toward modular teaching and assessment with clearly stated learning outcomes in the late twentieth and early twenty-first century in many institutions enabled the blending of geography with many different subjects at the outset of degree courses (combined degrees) or during a period of study so that students can incorporate geography modules with almost any subject. This has led to geography being branded and rebranded. It can be marketed with geospatial technologies and environmental science, which are seen as more entrepreneurial and more clearly linked to graduate labor markets. Geography in most Nigerian universities is now found within environmental science or environment and regional planning faculties, resulting from national level changes in education. In Australia, there are now no standalone geography programs; all are combined with other disciplines. This blending of geography with subjects including geology, biology, earth sciences, social sciences, history, and economics for an undergraduate degree reflects the integrated nature of twenty-first-century geography departments.

There is clearly a danger of a loss of identity in blending geography with other disciplines just because it has permeable boundaries and can quickly respond to interdisciplinary agendas. In such a situation the gap between human and physical geography is likely to widen further and potentially weaken the discipline. The future of geography would seem to hinge on clear links to employability for geography graduates and to the supply of geography students from schools.

Geography as a discipline of synthesis has generated many innovative pedagogic approaches and is well known at higher education level for providing pedagogic guidance that other disciplines seek to emulate. This reflects epistemic diversity and gives rise to the broad portfolio of geographical courses on offer. The development of geography as a discipline shows how geographical ideas have been established, challenged, and consolidated, shaped by diverse scholarly traditions and manipulated to fit the ideological needs of the time.

Development of the discipline

The rise of geography in higher education was inextricably linked to changes in economy and society in the nineteenth and early twentieth century. Geography had discipline status before, but the need for territorial security, exploration and empire, industry, agricultural development, and developing military interests led to the organization of both professional associations and the development of new programs of study.

The state, in terms of territorial security and empire, coupled with private commercial interests helped launch geography as a discipline through learned societies and universities in late nineteenth and early twentieth-century Europe (Maddrell 2009) and since then geographical education has been used internationally as a strategic tool to acquire applied knowledge for military, urban, industrial, and agricultural development. In France in the early nineteenth century it was recognized that scientific and
engineering research could help the military, followed by the rise of the German research university as a model for education in the service of empire and capitalism. Prior to 1900 geography at higher education level was initiated by the activities of a small number of individuals forming learned societies, established during an era of European and global exploration and by the publication of landmark texts such as Alexander von Humboldt’s *Cosmos: A Sketch of a Physical Description of the Universe* published in 1845, Carl Ritter’s *Die Erdkunde* in 1817, and *Physiography* (1877) by Thomas. H. Huxley (Martin 2005), which had a significant influence on the discipline. The rise of associations such as the Royal Geographical Society, American Geographical Society, Berlin Geographical Society, and Russian Geographical Society helped to lead, and was concurrent with, the transformation of higher education (on the German model) beginning in the late nineteenth century. The combination of texts and society backing led to the eventual recognition of geography as a subject in universities. Relatively soon after, organized graduate programs were established.

The period between 1900 and 1950 was characterized by growth in the number of universities offering geography, particularly in the United Kingdom, the United States, Belgium, Canada, and New Zealand as well as growth in the number of national societies supporting the discipline. With the growth of geography in universities, the discipline gradually attained a higher status. Although geography grew steadily throughout the first half of the twentieth century the growing prestige of “hard” sciences and the division between “physical” and “social” sciences led to geography becoming marginalized in some places.

The concerns of individual states have been central to the development of the discipline over time. Several countries, including Canada, recognized the need for geographers in collating an inventory of resources for development. Booms and slumps in the number of departments offering geography have characterized the development of the discipline in the United Kingdom and United States and by the end of the 1950s geography was a popular subject choice for undergraduates in the United Kingdom, with an intellectually vibrant atmosphere (Johnston 2004). Johnston suggests that human geography was not identified as part of social sciences until the end of the 1960s and did not fit contemporary perceptions of the frontiers of knowledge and that without advocates in the British Academy and Royal Society at that time, geographers had no influence on the structure and future of the academy. Indeed the academy of sciences structures around the world differ greatly, particularly in socialist and postsocialist countries, leading to differences in the way that the subject is organized.

There was further expansion of higher education provision in the United States but geography was dropped as a subject by several high ranking universities in the 1950s and 1960s, a trend that continued on a more dramatic scale into the 1970s with 32 geography departments closed in the United States in the period between 1970 and 1976.

The Open University, founded in 1968 and admitting students from 1971 included geography from the outset. Allowing access to geography courses at a distance via technology meant that students who did not have the opportunity to attend traditional campus universities to access degree courses, could now through distance courses. Through Future Learn, an international university collaboration, online learning now has a global audience.

The “traditions” within human geography have seen prominence at different times during the evolution of the discipline in the twentieth
century. These include a regional approach to geography, the quantitative revolution, spatial turn, and cultural turn, that have impacted on different countries at different rates and in some cases have spawned subdisciplines as geography has evolved.

In the late 1950s in North America the quantitative and theoretical revolutions started a change that had effects across the geographical community demanding a rigorous scientific approach. Emphasis on quantitative data collection and computation soon became embedded in geographical degree programs. Research teams and groups were needed to gather sufficient data to solve large-scale problems. For example, international research collaborations like the Metropolis Project, a multidisciplinary, multisite inquiry into migration and the integration of ethnic and religious minorities in the world’s urban areas. Funding became much more important to sustain multi-institutional teams and research councils, foundations, charities, and societies became crucial to supporting this work. Gradual fragmentation and modularization in some institutions has led to student choice from a range of specialist modules, rather than being able to cover the entire scope of the discipline.

The role of the nation-state

Higher education is often closely tied to national political and economic agendas, particularly in those nations in which the higher education system is funded, for the most part, centrally. A set of articles in a 2007 issue of the Journal of Geography in Higher Education examined the role that national contexts and the state can have on intellectual and educational agendas in geography with examples from China, South Africa, Australia, Canada, Singapore, and the United Kingdom. The Canadian geographical project has been inextricably linked to the nation-state with the government understanding the strategic importance of an inventory of the nation’s resources for planning and management in rural and urban areas. This led to rapid growth of geography departments with significant expertise in quantitative and statistical expertise as well as GIS growing from the government need for maps for land management. In China the state has even more significant, direct influence on the higher education system by allocating student quotas for HE institutions, recommending textbooks and controlling the classification of undergraduate and postgraduate programs. The government also decides whether geography forms part of admission tests for HE which has driven decline and more recent expansion in HE geography in this way. Greater global outreach by Chinese institutions, the use of non-Chinese textbooks, publications in international journals, and collaboration with overseas scholars and partnerships with overseas institutions all reflect greater openness in China today. Despite less direct intervention by the state in Singapore, the national context is evident in the research questions that Singaporean geography considers, for example, national identity, transnational corporations, multiculturalism, migrant workers, and water resources.

Beyond the influence of the state an increased awareness of the subject by prospective students has led to a proliferation of courses (for instance, there are more than 200 undergraduate geography courses in US universities today). Alexander Murphy, writing about the United States in his article for the 2007 collection, suggests that the main reasons for growth include: geography being seen as relevant to issues of the day such as increased public interest in what geographers do as geography is in the news and the geo-technology revolution and forensics are popularized through TV shows; geography being increasingly appreciated among scholars.
in other disciplines; huge growth in the use of GIS; an expanded job market for those with geographic training; and, finally, the emergence of more analytical and sophisticated geography in school curricula. The relationship between school and university geography is strong with schools creating a demand for and feeding undergraduate courses.

While the role of the state is still highly significant, more recently the marketization of higher education provision, created by the neoliberal agenda, has internationalized the availability of access to higher education degrees in geography and this is leading to new developments in terms of competition and divisions of labor.

The neoliberal agenda

In recent decades, many universities’ systems have had to confront neoliberal policies on several fronts. One of the most significant is that students are being asked to cover more and more of the costs of higher education as government funding declines, creating a consumerist culture. Second, universities are under increasing pressure to publish outcomes. World rankings of geography departments now exist and several departments from the United Kingdom as well as those in the United States, Singapore, and Canada feature in the top ten. These are created through a variety of teaching and research metrics. In 1986 the Research Assessment Exercise (RAE) was carried out across UK higher education and this was linked to funding allocations for research based on assessments of measures of quality. The RAE led to a growth of a management and accountability culture in universities. New measures of research quality and impact have become embedded in the more recent incarnation in the United Kingdom, called the Research Excellence Framework (REF), and metrics are likely to continue to be used to make distinctions between types of research output and activity, particularly in the area of demonstrating research impact beyond higher education. Similar research quality assessments are carried out in other countries such as the Research Quality Framework (R.Q.F) in Australia.

A teaching excellence framework (TEF) to measure teaching quality, learning environment and student outcomes, and learning gain is being implemented in England. Teaching quality has also been measured through student survey data (student satisfaction in the United Kingdom and student engagement surveys in the United States and Australia) that have differentiated geography courses. Key information sets such as the number of faculty who are trained teachers and the number of contact hours on courses, are becoming public information indicators that students can use to make choices between courses. In addition, metrics such as retention rates, graduation rates, the proportion of students graduating with “good” degrees (as defined by grade point average scores in the United States or by degree class in the United Kingdom), graduate employment (particularly whether graduates are in employment within six months of graduation and whether graduates have successfully secured graduate level jobs) are also made known. The success of particular groups is also monitored, including gender, non-traditional learners, such as those from families with no previous higher education experience, minority ethnic groups, and people of differing generational status. In some cases metrics show that particular pedagogic practices benefit nontraditional students much more, for example, engaging in undergraduate research with a faculty mentor.

The segmentation of academic labor into research only and teaching only roles has created a contingent labor force of nontenure-track faculty (or those on fixed term and temporary
contracts). For those focused on teaching there is no provision for research or writing time as part of their employment, and both postdoctoral researchers and teaching fellows share precarious employment terms.

Competition between universities as part of this marketization of HE has had benefits in terms of fore-facing learning outcomes and the unique academic offering of particular courses. In addition it is ensuring that universities communicate academic research more widely and with public groups to market their research activity and clearly demonstrate the impact it is having.

**Learning, teaching, and assessment**

**Content, standards, and quality markers**

The wealth of potential subject content at higher education level is vast and an appreciation of the multiple ways in which geographical phenomena can be interpreted and studied is part of the disciplinary perspective. The range of techniques and methods used by geographers in HE courses allows students to gain familiarity with a number of technologies and approaches, for example, cartography, fieldwork, GIS and remote sensing, modeling, and quantitative and qualitative methodologies for gathering and interpreting data. Fieldwork provides an engagement with real places allowing concrete understanding of the complexity and dynamism of geography.

Drivers for curriculum content come in a variety of forms. In the United Kingdom, the Quality Assurance Agency created a first geography subject benchmark statement in 2000, drawn together through consultation with academics from a number of institutions as well as the Royal Geographical Society (with the Institute of British Geographers), hereafter RGS (IBG). The subject benchmark statement is revised through consultation with academics, learned societies, and employer representatives. It describes the nature and characteristics of geography degree programs in the United Kingdom and creates an agreed expectation of the standards students should achieve. The statement contains the aims, subject knowledge, skills/attributes, and the types of learning, teaching, and assessment which are likely to be encountered, recognizing the diversity of educational providers. Further quality assurance is provided by external examining which is used to demonstrate comparability of awards (a practice largely confined to the United Kingdom and Ireland).

**Contemporary pedagogy**

The country in which geography is being studied makes a large difference to the type of material being taught, the variety of assessments being made of student work, and the freedom which students experience in choosing the elements they study as well as the extent to which the discipline mirrors school geography. For example, the idea from geographers Healey and Jenkins that students learn *through* as well as *about* research (Healey and Jenkins 2009) has not reached all institutions, in policy or practice, despite a wealth of inspiring case studies where it is taking place.

The regional approach developed a commitment to fieldwork as a pedagogy and this has been adopted in a range of places including local and exotic field classes. The benefits of fieldwork in terms of participation in communities and experiential learning have created concerns about “othering,” particularly in human geography in some countries. A range of active learning approaches that promote dialogue and reflection has been adopted where there has been significant consideration of the values
communicated through choice of discipline content and geographical pedagogy. These include simulations, such as role play, debates, authentic decision-making exercises, and research-based learning, in order to synthesize reason and factual content with emotion and moral values. This synthesis could be considered a signature pedagogy for the discipline and has given rise to a range of creative forms of assessed work. Students are encouraged to reflect on their learning through assignments, such as reflective diaries, briefing papers, critical reviews, and creative ways of conveying the combination of affect and discipline content, such as exhibitions and video. Greater diversity of assessment types may result from universities as they seek to develop creativity and employability skills in their students and to show how their curriculum is transformational. More information about contemporary pedagogy and trends can be found under Geography education: fieldwork and contemporary pedagogy.

The graduate attributes agenda, to which David Nicol in Scotland and Simon Barrie in Australia have made significant contributions, has been influential recently with universities nominating the graduate capabilities that their students will leave with and geographers are increasingly being asked to demonstrate links to these generic capabilities (such as communication, information literacy, research literacy). This agenda has emphasized self-regulation and development through self and peer learning that is social and reflective. It will become increasingly important for geographers to identify discipline-specific graduate and postgraduate capabilities.

The student experience

The student experience of learning geography in universities can be described on a continuum. At one end there is a high degree of choice in content via specialist modules with emphasis on being rewarded for wide reading, originality, and demonstrating critical thinking and often this is coupled with an individual research project as the culmination of the course. At the other extreme is a heavily prescribed curriculum with low levels of choice and autonomy where core knowledge is tested particularly through examinations. In order to develop transferable skills for employment, students may be expected to engage in a wide variety of assessed tasks beyond just essays and examinations; for example, work placement reports, press releases, briefing documents, podcasts, blog entries, reflective diaries, journal articles, and posters.

Employability and transferable skills

Learning outcomes became a feature of module specifications in the early twenty-first century and the vast majority of university courses are modular. The importance of “employability” is evident from university webpages and the transferable skills that graduates can hope to obtain during a period of study with the particular university feature strongly in the introductory literature that advertises degree courses. Further specialization as a result of the wide choice of modules on degrees plus the need for graduates to differentiate themselves in a competitive job market has led to the development of specialist and vocational master’s degree courses (see also Geography education, workforce trends, twenty-first-century skills, and geographical capabilities).

Learning technologies

There has been a proliferation of learning technologies to support higher education and student expectations are growing in terms of accessibility to online material, online communication, and the use of technology in learning
and teaching. Distance education courses have a long history in some institutions and are now providing access to learners from around the world. Some face-to-face taught courses have transitioned to online provision, whereas others have a blended approach allowing technology to enhance face-to-face contact such as through online submission of assignments, online feedback, institutional virtual learning environments that store course materials, and online links to resources. University academics have also adopted technology as a means of coping with larger numbers of students, such as through the development of multiple choice questions for assessment with automated feedback, and online marking allowing students to receive feedback more quickly. Technology has also had a major transformative effect on geography pedagogy as noted in Geography education, workforce trends, twenty-first-century skills, and geographical capabilities and Geographic information science and technology: educational directions.

**Key challenges**

**Debates on the purpose and value of geography**

There is an ongoing debate about the balance between and importance of providing training for a future workforce with educating students with a passion for the discipline and due consideration of the values of studying geography in terms of developing global citizenship and ethical thinking that extend beyond graduate employability. This is echoed to some extent for academics in relevancy debates about the need to carry out applied research that informs policy. The value of geographers, in contrast to those from other disciplines, suffers from a lack of public visibility demonstrating the impact of geographers in the wider context and on public policy, evidenced by repeated calls in presidential addresses from geographical societies in the 1970s and early 2000s. Very few geographers have entered the public’s consciousness through media interventions, in contrast to archaeologists, historians, and geologists, and as such have had little impact on policy direction.

Engaging with students in the neoliberal university is more challenging than in the past. Students can be highly strategic in their engagement with the discipline if they are only seeking a qualification that results in employment, prioritizing assessments over a deeper engagement with the geographical perspective. One response to this is the new concept of “GeoCapabilities,” the discipline-specific capabilities that a geographical education engenders, allowing graduates and postgraduates to contribute to their own welfare development as humans, that of their community, and the wider world through a lifelong geographical perspective. The approach is informed by the work of Amartya Sen and Martha Nussbaum whose human capability development advocates a capabilities approach to education, resulting from a reaction to the neoliberal conception of the modern university. The capabilities approach is a resource for the lifelong development of human potential and wellbeing. Solem, Lambert, and Tani (2013) have started the process of defining internationally recognized GeoCapabilities. This is a very different educational agenda from the managerialist approaches that characterize neoliberal universities with students being seen as consumers and faculty (teaching and research staff) as “human resources.” Decreased government funding has meant that universities have to find alternative finance, competing to attract income from overseas students, taught postgraduate courses, consultancy, refresher courses, and continuing professional development provision. This
is monitored and evaluated to ensure value for money in a culture of audit and accountability. A recent development in the university sector is the senior manager roles being taken by people with a nonacademic business background. State funding of universities shows great variability globally, giving rise to differing outcomes as a result of diverse funding models. An example of the impact of economics is noted by Xiaojian Li, Yunfeng Kong and Baovu Peng writing about education in China since 1980 in the *Journal of Geography in Higher Education*. China is one of the largest countries in terms of geography student enrolment in higher education and the subject has shared favorably in the country’s recent expansion in higher education. The regional patterns by number of geography programs correlate to the size of population while the quality of geography programs shows a close relationship with regional economic development.

Marketization and growth in HE provision has served to increase access to HE geography, allowing a greater number of students to take degree courses. This has led to a challenge in terms of ensuring participation is accessible to all groups of prospective students and universities are often provided with targets for “widening participation” (nontraditional) students as well as international students who pay significantly higher fees. In order to achieve targets HE providers often create bursaries and scholarships to ensure that support is created for recruitment and retention of particular sectors of the student population.

The transition challenge

Students experience three key transition challenges as they engage in higher education. First is the transition into university from school (or alternative prior experience), second is the transition they go through as they become an autonomous learner and “becoming a geographer,” and, finally, the transition into further study, employment, or other setting beyond the university. These are dealt with in turn.

Translating the critical, often experimental, and tentative material which forms part of university level research in geography into teaching occurs in many departments where the development of critical skills is a central objective. However, this does not easily translate into the school curriculum where subject content needs to be more enduring, certain, and stable. This has led to a declining influence of university geographers on the geographical content of the school curriculum and therefore to a key challenge to transition into higher education. Students arriving have different expectations about the geography they will be learning from the academics involved in teaching and grading their work. In the United Kingdom previously strong links between geography in schools and universities have weakened, creating a transition challenge for staff and students (e.g., Sidaway and Johnston 2007). The motivation to integrate cutting edge university research in the school curriculum has weakened as a result of the national curriculum dictating content and student performance being used to create school league tables, reducing teacher autonomy. The syllabus at advanced level (the level students achieve to gain entry to university) allows school teachers to omit skill and content areas they are less confident to teach, such as statistical techniques and some of the content of physical geography.

In developing autonomous learners, universities often adopt a “students as researchers” pedagogy where students are encouraged to formulate research questions, to collect or use existing data sources, and to present their findings. This approach requires support from the
faculty and with the growing numbers of students on courses the challenge is to provide this support to all students. Where this is not possible on a one-to-one mentoring basis, supervision and group work is used to provide peer mentoring. Embedding research-based learning early in the curriculum is seen as a good way to support this transition (Walkington et al. 2011).

Links to employability for geography graduates

The preconditions for growth in geography as a discipline in higher education relate strongly to the ability of students to perceive a direct future benefit in terms of employment from a qualification in the subject. In South Africa’s post-apartheid era, teaching and research has been local or applied with geography courses being cross-disciplinary and vocational, providing marketable skills. Clear links to employment have been crucial to the growth of the discipline in many other countries too. In Spain a dichotomy is evident between an inability to attract undergraduate geographers and greater visibility of geographers in the public eye, particularly in GIS, remote sensing, and planning. Geography is on the political agenda of the EU through the Infrastructure for Spatial Information in Europe (INSPIRE) project. This aims to build a common geographical space through ensuring that the infrastructure which holds spatial data is compatible across all member states, the objective being to allow sharing of data to solve transboundary sustainability challenges a development that may attract more undergraduates to geography in Spain, particularly through technical approaches like GIS. Geography recruitment is healthy in Singapore, but the subject is not oversubscribed as it is not seen as effective as subjects such as economics in securing employment and social mobility. There is clearly a challenge for geography faculty in making clear to students the links between their academic study, lifelong learning, and employment.

Fragmentation of geography as a discipline

In some countries there has been a tendency over the past twenty years to reduce the link between geography and area studies. Fragmentation of the discipline for marketing purposes has led to new labels for geographical courses. In Australia Chris Gibson reported in a 2007 paper in the Journal of Geography in Higher Education a recent decline in many undergraduate geography programs, a commensurate reduction in faculty numbers, and a reduction in independent departments named “geography.” Department names have been replaced with titles such as “environmental management” and “urban studies.” Reduced government funding and increasing staff/student ratios, the decreasing popularity of high school geography and the strong emphasis of the Australian RQF on interdisciplinarity have all contributed to the need for geography as a discipline to be adaptable. Fragmentation can be conceptualized as adaptability, resilience, and resourcefulness of the discipline, particularly at a time when interdisciplinary research is seen as an approach to solving complex large-scale problems, however, it could lead to marginalization and a loss of what “geography” means as a discipline (Holmes 2002).

Tensions between teaching and research

The teaching–research nexus is an ideal state where mutual benefits are derived from a combination of teaching and research. The relationship between teaching and research varies across institutions and countries, with some institutions
focused primarily on research with teaching being a relatively smaller enterprise, whereas increased participation in higher education in some places has led to larger cohorts in the universities that focus on teaching. In Russia there is a split between largely teaching-focused universities and government research institutes.

Le Heron, Baker, and McEwen (2006) suggest that institutional rules and government strategies, funding, and assessment have divided teaching and research and harmfully continue to suggest that they are unrelated whereas a culture of co-learning should be developed where synergetic relationships between research and teaching contribute to the development of both and the quality of teaching and research are enhanced. Most universities have the philosophy of a department that is “research led” with “research-led innovative teaching,” but what this means is different in different places. The number of students who participate and produce research, rather than simply learning about it, varies significantly from one institution to another. Two national journals of geography dedicated purely to publishing the results of undergraduate research – GEOview in Australia (www.iag.org.au/publications/geoview) and Geoverse in the United Kingdom (http://geoverse.brookes.ac.uk) – demonstrate what is achievable.

Sidaway and Johnston (2007) highlight the positive outcomes of research assessment arguing it has improved productivity and quality of UK research and that dynamic, strategic goals have been incorporated into a previously static system and it has created opportunities for innovative academics. However, they note that it can be to the detriment of teaching quality. Precedence is now given to particular formats of publications such as single-authored journal articles in high ranking journals. Sidaway and Johnston (2007) suggest that part of the reaction to the RAE has been a renewal of focus upon teaching and learning, described by Healey (2003) as “the scholarship of teaching.”

Inclusion

Higher education continues to be a site of privilege and exclusion and this continues to be the case despite years of effort to create a more inclusive discipline. Gender, ethnicity, citizenship, class, disability, life stage, age, sexuality, and international origins are all aspects of the diversity and inclusion agenda in higher education. These challenge the discipline on a number of levels. The content of the curriculum, the nature of the pedagogic approaches employed, the composition of the student body and that of faculty, and whether they are tenured and promoted all highlight huge challenges in undoing sexist, racist, ablest, heteronormative pasts. The challenge is to teach about diversity issues in an inclusive way regardless of the makeup of the student population. Globalization has made higher education much more international in terms of students, faculty, and outlook and in some places a third of geography departments comprise of overseas students and faculty personnel. In other institutions diversity issues may be more hidden. Recognizing students as individuals who intersect with all these diversity variables is essential. Schlemper and Monk (2010) have acknowledged that multiple approaches are needed for an inclusive approach, including a proactive and sustained approach to the recruitment and selection of students and staff, by integrating inclusive perspectives across the curriculum through proactive planning, fostering participation in events that address inclusivity, reflecting on inclusive perspectives at departmental level so that the climate is one that welcomes diversity, raising staff consciousness of
websites, noticeboards, use of space, and so on as the face of its culture and embedded behaviors.

In summary the discipline of geography has been challenged in terms of identity by its interface with other disciplines and inherent tensions between synthesizing or fragmenting into a human–physical divide or more specialist areas. Academics have to balance their time between teaching and research and many are faced with rapidly growing student numbers. Students who might not have previously accessed HE are now attending university and staff have to respond by supporting learners to make three significant transitions, that is, into, through, and out of HE. The integration of an academic’s own research into their teaching and adopting a “students as researchers” pedagogy are positive responses to these challenges.

Opportunities

There are many opportunities within geography for contributing to research and teaching in higher education generally as the discipline lends itself well to connecting with other subjects, being an integral part of international collaborations and addressing global research problems. The pedagogic practices which are distinctive are also of interest to other disciplines and there are a wealth of sharing mechanisms in place through networks, journals, learned societies, and online resources.

There is a wealth of opportunity to address global challenges through higher education research and teaching in geography. The discipline is composed of a range of diverse perspectives as geographers have an ability to combine reason and emotion in decision-making that has resulted in a “signature pedagogy.” Geography can provide students with a discursive environment to scaffold and advance learning and to maintain a reflection on values and opinions as they are developing, all within a global perspective. Field-based learning uses a spatial framework and this can serve to broker dialogue with science, social science, and humanities researchers due to the ability of geography to speak to and through many different disciplines. The language of the discipline is one which values scientific, social science, and humanistic perspectives allowing geographers to coordinate multidisciplinary teams and work on interdisciplinary problems. Furthermore, the public can engage with geographical data in new ways through volunteered geographic information, allowing them to engage in geographical knowledge production and dissemination.

International mobility and collaborations

The European Higher Education Area has specifications for academic qualifications as part of the Bologna declaration, an agreement made by 29 countries in 1999 to ensure comparability of standards and quality of higher education qualifications. It describes European credit transfers for degrees at bachelor’s, master’s, and doctoral level with 60 European credit transfers equivalent to a year of study. This system allows greater ease of movement between countries for both study and employment and is complemented by international exchange schemes such as Socrates/Erasmus, as well as multiple institutional links that are developing rapidly. In the United States, statewide articulation agreements and transfer practices are developing for general education which can include six credit hours for geography. Students are therefore more mobile than ever.

Specialist staff deliver intensive courses in different parts of the world as “flying faculty” and universities with multiple campuses create the opportunity for students in several countries
to share the same curriculum. This type of development has led to new laboratories being set up and laboratory practices being shared, for example, UK universities with campuses in Asia have been developed through staff partnership. The norms and practices of different countries can be challenged and reflected upon as a result of these collaborations. The growing movement of students and faculty internationally, for example, through the Erasmus program and Fulbright Fellowships, allows new practices to be shared as does the growing popularity of online courses which allow ideas to be shared without the need for travel.

UNIGIS is a worldwide network of universities collaborating on the creation and delivery of geographical information science and systems through distance learning. Currently postgraduate courses are offered in six European languages and partner institutions are present in over 40 countries. The initiative acts like a large multi-institution department. UNIGIS also hosts residential workshops and summer schools. This style of learning has provided enhanced accessibility to leading-edge teaching for students in countries all over the world.

Open educational resources (OER)

Making educational resources openly available for use and reuse worldwide is the basis of the open movement. Each university has adapted this in their own particular way with some favoring lecture capture which can be accessed as downloadable online resources, others developing institutional repositories for research and teaching materials with varying levels of public access, and others engaging greater participation through online courses. In the United Kingdom, OpenLearn labspace has collated many geography open resources and JISC (a UK charity which champions digital technologies in research and education) uses a repository called JORUM to make these searchable and available. Some institutions have developed MOOCs allowing the public and prospective students to access teaching materials freely, such as videos of lectures, and to join in interactive exercises and discussions online. A project called C-Change, developed through multi-institutions working in association with learned societies from geography, earth, and environmental sciences on climate change, attempted to create a pedagogic shift to support colleagues to open up their own resources for sharing by creating examples of practice and clarifying copyright law and policies. The American Association of Geographers (AAG) has also created modules to support the teaching of geographical issues based on country specific case studies developed through partnerships negotiated through the Center for Global Geography Education.

The interest in a self-sustaining community of practice through OERs is growing, especially where academics view greater exposure of their work as a positive action and institutions recognize that branding and sharing their intellectual property is a valuable marketing tool.

Networks

Learned societies provide excellent networking opportunities through research groups, teaching groups, support materials, training, journals, and conferences. The International Network for Learning and Teaching (INLT) is a collaboration among geography academics to facilitate collaborative writing on themes of international interest to develop teaching and learning practice. Many of these writing events result in symposia for the Journal of Geography in Higher Education. Other journals include the Journal of Geography and International Research in
Globalization could soon mean that the dominance of English as the medium for educational debate in geography could be diversified or replaced as agendas for the internationalization of committee memberships of journals and conferences are implemented.

**Frontiers in research and practice**

**Technology**

In the last decade social media has become a significant vehicle for engaging with the public and disseminating research, but also for enhancing face-to-face teaching and attracting students. Technology has the power to transform learning and people have envisioned a future without formal classrooms at place-based campuses while still providing directed learning. Students can be taught while in the field by a faculty member who remains within the university, for example, by carrying out field research methods, but gaining feedback on observations they relay via social media to peers and faculty elsewhere. Academic material can be sourced, packaged, and marketed for online consumption and products can be accessed on the move in the field or at a desk through a wide range of electronic devices. In the search for cheaper education “flying faculty” are already delivering intensive short courses around the world, though these are currently only complementing place-based courses rather than replacing them. New technological developments already allow students to engage in the remote control of laboratory equipment, the use of virtual microscopes, and involvement in whole class virtual fieldwork through the use of avatars and even virtual graduation alongside students at a real ceremony.

**Measuring learning gains**

Despite technology assisting with some transactional aspects of the student experience there remains a strong desire for face-to-face “contact time” with faculty, particularly in individual or small group settings. In addition, increased demand for work placements and overseas experiences (as part of a course through reciprocal mobility arrangements or exotic overseas fieldwork and expeditions) have cost and equality of access implications. The learning gains from these types of added-value experience are challenging to measure. Methodologies for doing so need to be developed across different types of institution as contextual factors can be highly influential. There is a strong ethos of collaboration among the learning and teaching community in geography, but pedagogic research and the scholarship of teaching and learning are often poorly funded and as a result the status of this form of inquiry suffers or remains unpublished. If we are to adopt an evidence-based practice approach then further research is essential to explore the most effective strategies for: creating learning communities online; maximizing the impact of overseas experiences on learners’ geographical understanding of the world; inclusive assessment; rapid but effective feedback; internationalizing curricula (Haigh 2010); fieldwork with larger classes; and the development of attributes, such as global citizenship, through discipline teaching.

**Conclusion**

There is clearly potential for radical change if informal learning technologies prove able to successfully create a sense of learning community online and allow authentic engagement in dialogue and decision-making without face-to-face
contact. Chism (2006) suggests that social learning can take place in virtual spaces that transcend the social and academic divide. Yet, significant differences exist around the world in access to computing hardware and software, the Internet, educational resources, and funding within and between countries. In addition, national or commercial monitoring of the web has the potential to widen existing inequalities further.

The trend toward international, national, and regional university consortia is likely to continue as universities seek to stratify and differentiate themselves. Opening a formerly elite HE system to mass educational access does not necessarily reduce inequalities. Michael Young has argued that we have the potential to provide epistemic access to all students through formal study and the transmission of powerful knowledge. What constitutes powerful knowledge in geography will be an interesting area of future debate. The impact of technology in opening up apparently endless opportunities for learning “any time, any place” needs to be clearly differentiated from the unique offering of a university or college experience in gaining epistemic access where a new way of thinking can become the basis for suggesting and practicing realistic alternatives.

Geography in higher education has been characterized by changing fortunes. The strength of the discipline lies in its ability to synthesize a range of perspectives and generate problem-solving and decision-making research areas as well as to create innovative pedagogies to teach a variety of forms of geographical thinking. Signature pedagogies (Schulman 2005) in geography that develop applied skills, ethical thinking, spatial understanding, as well as engagement with more generic graduate attributes through the powerful knowledge of a geographical education can develop students’ geocapabilities to prepare them for living responsibly in a supercomplex world (Barnett 2000).

SEE ALSO: Geographic information science and technology: educational directions; Geography of education and educational systems; Geography education: fieldwork and contemporary pedagogy; Geography education: promoting diversity and broadening participation; Geography education, workforce trends, twenty-first-century skills, and geographical capabilities

References


**Further reading**


At first glance it may not be advisable to divorce the terms “knowledge,” “education,” and “science.” They are all based on learning processes and are strongly interrelated. However, the rapid expansion and specialization of research has sundered the wide field of the geography of knowledge and education as it was defined until the late 1990s (Meusburger 1998) into geographies of education, geographies of knowledge, and geographies of science. The boundaries between these three fields of inquiry may occasionally blur, but scholars working in them have different research interests, use different theoretical approaches and methodologies, and build different citation networks.

The geography of education in the narrow sense – the topic of this entry – deals with the causes and consequences of spatial variations in the provision and consumption of education; the impact of local milieus and social environments on educational attainment; spatial variations of the sociodemographic structure of the teaching profession; the influence that political systems, educational policies, and demographic change have on location patterns of schools and universities; the spatial mobility of scholars, students, and ideas; and other issues where places, spatial relations, and social environments have an impact on educational processes. So far, the vast majority of publications in this field have centered on formal education, but nonformal education and various forms of preschool education and child care also figure as research issues of geographies of education.

“Knowledge” is a much broader term than “education.” It encompasses not only the results of learning processes within the educational system but also professional expertise, abilities, capacities, experience-based intuition, and values. It is represented in individuals, organizations, technologies, research results, and material products. Geographies of knowledge study the places and spaces of knowledge generation, the diffusion and application of various categories of knowledge, and the reasons and consequences of spatial disparities of knowledge. Many interests pursued by geographers of knowledge lie beyond the educational system. Because knowledge constitutes a capacity for action, a knowledge-oriented approach lends itself to the study of many different topics of human geography. Scholars in this field study the relations between knowledge and power, knowledge and economic performance, and knowledge and various societal processes.

Geographies of science (Livingstone 2003) examine the spatiality and situatedness of science and research. They inquire into the sites at which scientific knowledge has been produced, assessed, and consumed; study the travel of explorers, scientific ideas, theories, and methodologies over physical and cultural space; and examine the geography of reading by analyzing why scientific works or authors are interpreted very differently in varying cultural and political contexts.

At first glance, geographies of education seem to have some research topics in common with comparative education. But the theoretical
approaches and methodologies of these two disciplines diverge greatly. Geographers are not as interested in comparing educational systems, policies, and practices of various states as they are in exploring the profound impact that places, local milieus, regional structures, social environments, and cultural contexts have on processes and institutions of education. In their view, organizational and cultural contexts of learning are not abstract concepts, but become manifest in sites, places, and regions.

Early roots and development of the geography of education

Empirical research on the spatial structures and disparities of education harks back to the first half of the nineteenth century. Many social reformers in the 1800s believed that social problems associated with early industrialization, such as poverty, child labor, crime, alcoholism, and a lack of decent housing resulted from ignorance and a lack of moral education. They therefore started investigating the social and spatial variations of illiteracy and educational attainment, spatial inequalities in the provision and quality of various types of schools, the skills and salaries of teachers, the availability of books in households, and other issues to describe the social, cultural, and economic settings of neighborhoods (Marsden 1987).

France was the first country to engage in the comprehensive, nationwide study of regional differences in illiteracy. This work drew primarily on examinations of military recruits and the inspection of marriage registers, both sources yielding data on the ability to read, write, or both. In 1826 Charles Dupin published the “Carte figurative de l'instruction populaire de la France” (Thematic Map of Popular Education in France). This map (reprinted in Meusburger 1998, 193) revealed remarkable regional disparities in educational attainment between northern and southern France. The map’s accompanying tables compared educational attainment, the number of patents for inventions, and membership in the Académie Française with various economic indicators, suggesting a high correlation between the population’s educational achievement and regional economic performance.

In the United Kingdom, a number of statistical societies emerged in the 1830s and 1840s to study the living conditions and social situation of the poor, and to foster social reform, with literacy again being a main focus. One result was a map entitled “Ignorance in England and Wales,” published by Fletcher in the Journal of the Statistical Society of London (1849). When C. Booth, R. W. Rawstron, J. Fletcher, and others analyzed poverty in London, Manchester, and other cities, they, too, used various indicators pertaining to schools, teachers, and educational attainment. A number of studies in different disciplines described inequalities between states with regard to literacy, the introduction of compulsory education, educational attainment, enrollment in universities, the distribution of leading scientists, and the level of technology. But few studies focused on inner urban disparities. One of the exceptions was C.R. Shaw, a member of the Chicago School of social ecology, who in the 1920s investigated spatial disparities of the relationship between truancy and juvenile delinquency in Chicago.

However, these traditions were not taken up in the mainstream human geography of those days. Apart from occasional early studies, education attracted little attention among human geographers until the 1960s. Early exceptions were W. Christaller and H. Bobek, the pioneers of central place studies, who used the hierarchical structure of the educational system (from
elementary school to university) as one of various indicators to construct their system for ranking central places. After World War II, E. Parker and C. Colby in the department of geography at the University of Chicago supervised four doctoral dissertations on educational land use. Three of them were published as research papers of the department of geography at the University of Chicago between 1948 and 1949.

The interest that geographers showed in education grew in the 1960s and 1970s, when many countries strove for unprecedented expansion of their educational systems, including the establishment of many new secondary schools and universities, a steep increase in university enrollment, and the needed reorganization of elementary school location patterns. It was necessary in that period to plan and negotiate locations for new schools and universities, justify the closure of small elementary schools, reorganize catchment areas and student transport, and explore the reasons for intolerable spatial disparities in educational attainment in order to counteract them. These tasks required competence in geography; experience in urban and regional planning; sensitivity to the importance of local identities and social environments; and detailed knowledge of the history, culture, and economy of the towns and regions in question. Geographers soon became involved in applied educational research and regional educational planning. This boom in reforming educational systems was also nourished by emerging discourses about the economic importance of human resources and about brain drain, the knowledge society, and the legitimacy of social and spatial inequalities.

One of the first and internationally most influential pioneers of the geography of education was the German geographer Robert Geipel (b. 1929; Figure 1), whose research focused predominantly on this field from the early 1960s to the late 1990s. He was professor of social geography at Goethe University in Frankfurt am Main, Germany (1961–1969), then held a chair for applied geography at the Technical University in Munich (1969–1994), and served as director of the Bavarian State Institute for Higher Education Research and Planning in Munich (1982–1994).

In 1967 Geipel and scholars from other disciplines founded the working group on spatial educational research at the Academy for Spatial Research and Planning in Hanover, Germany, which until the late 1970s was an important platform of inter- and transdisciplinary research about spatial disparities of education. In these
functions he had close scientific relations to the pioneers of this research field in other countries (particularly the United Kingdom) and disciplines (sociology, psychology, and economics).

In 1983 the Association of German Geographers founded the Geography of Education Study Group, which is still highly active today. Members of this study group produced the sixth volume of the *National Atlas of the Federal Republic of Germany on Education and Culture* (Mayr and Nutz 2002).

In the United Kingdom the early pioneers of the geography of education in the 1970s and 1980s were G. Hones, W. Marsden, R. Ryba, W.T.S. Gould, L. Bondi, and M.H. Matthews. However, the number of education-related publications by geographers did not begin to soar in the United Kingdom until the 1990s. The great majority of the authors restricted their citations to English-language publications and were not aware of the research in other countries (see review articles of Hamnett, Ramsden, and Butler 2007; Hanson Thiem 2009). Sporadic publications about the geography of education also appeared in other countries.

**Main research topics**

In the early period between the 1960s and late 1990s, geographers of education and educational systems investigated a wide sweep of topics, including: the spatial organization of educational systems (size, location criteria, and catchment areas of schools and universities); the reasons and consequences of spatial disparities in educational attainment; relations between towns and their universities; the economic and cultural impacts of universities; the social composition of schools (measured by the educational achievement and occupations of both parents); school segregation and student transport; spatial disparities of vertical social intergenerational mobility (comparison between the educational achievement of parents and of their children); spatial variations of professional skills, sociodemographic structures (age, social origin) and the feminization of the teaching profession; the impact that geographic structures (e.g., population density, size and distribution of settlements, road and rail communications, and local employment structures) have on location patterns of various types of educational institutions; spatial variations of expenditure on education; spatial impacts of various political ideologies on educational policies and school-planning concepts; relations between communities and schools, especially the social, cultural, and economic effects of school closures on communities; the impact of educational achievement on professional careers and migration; the mobility of academics and students; the role of schools in ethnic conflicts and processes of nation building; illiteracy as an obstacle to economic development and societal modernization; the spatial distribution of jobs for high- and low-skilled employees; career mobility and networks of scientists and highly skilled professionals; and the correlation of different levels of educational achievement with unemployment, women’s employment rates, housing quality, and reproduction rates. In this early period, many geographers situated educational systems and educational attainment in a very broad economic, social, cultural, and political context (for details see Freytag 2003; Geipel 1966; Kramer 1993; Mayr and Nutz 2002; Meusburger 1998; Schmude 1988).

Since the late 1990s, new research questions and theoretical approaches have been added. Some of the latter deal with normative understandings of childhood, treat young people as subjects rather than objects of education, deepen the understanding of children’s spaces of learning, attend to gendered identity construction in schools, and focus on various facets of
nonparental childcare provision, to name just a few of the aspects addressed (for details see Holloway et al. 2010; Holloway and Jöns 2012; Matthews and Limb 1999). Other new research activities focus on the neoliberal international education industry. The intention of establishing schools and universities overseas is to attract fee-paying (revenue-generating) students, but it also creates new geographies of social exclusion within societies that send the students (Brooks and Waters 2011; Koch 2014).

Other research focuses on the spatiality of learning processes, alluding to the fact that learning and problem-solving do not take place in a social, political, or economic vacuum but rather are influenced by a multitude of factors whose local interaction results in a spatial context, learning environment, or knowledge milieu.

Because the educational system is an important sounding board of political, societal, and economic structures and processes, various indicators of educational attainment and educational achievement have been used in other fields of human geography as well to study spatial variations of societal structures, to uncover social and economic inequalities, and to display unequal power relations.

History shows that educational systems have an important political and economic function. In the nineteenth century the processes of nation building and democratization were intimately connected with the spread of literacy and the introduction of compulsory schooling. A high degree of literacy is regarded as a precondition for the evolution of large and complex organizations and bureaucratic administration and for the successful adoption of many innovations. In many new nation-states the primary functions of the educational system were not only to eliminate illiteracy and create a well-educated labor force prepared for various innovations but also to consolidate administrative power, support national identity and cultural memories, create heroic national historiography, shape stable interpretations of the world, promote an official language of instruction, construct an ideological and cultural homogeneity, and assimilate ethnic minorities.

The administrative control of schools, the language of instruction, the selection of teachers, the spatial organization of elementary schools, and the location of new universities have been of paramount political importance. In communist and other totalitarian systems basic education is a decisive instrument for controlling the minds of the population. In multiethnic or bilingual areas the language of instruction is a salient political issue. When territories are annexed, one of the first measures taken is to replace teachers and schoolbooks of the former system and to change the language of instruction.

The educational system has functioned as the key component of nationalization and denationalization policy. In order to Italianize the northern and northeastern regions of Italy, where the vast majority of the inhabitants spoke German, Slovenian, or Croatian, the so-called Gentile Reform of Fascist Italy in 1924 forbade the use of non-Italian languages in schools and kindergartens. A decree issued by the prefect of Trent on November 27, 1925, prohibited even private tutoring in German within families if it was given to more than three children at the same time. Police watched diligently for any form of private or clandestine instruction, and intervened frequently. In the United States the Bureau of Indian Affairs strictly forbade the use of Indian languages in schools it controlled, and enforced the rule until the end of World War II.

Varying priority of research questions

The priority of research issues examined by geographers varies not only over time but also
from one country to another. Priorities depend on the answer to a number of questions: What kind of governance does elementary education have – centralized or decentralized? Which administrative authority (local community, district, province, state, nation, or church) establishes, finances, and controls various types of schools or selects and pays teachers? Are elementary school catchment areas defined by school administrations or by parental decisions? What are the mechanisms for distributing pupils of different ability levels across different types of schools? Is the selectivity of schools based on the children’s cognitive abilities or on the financial resources of parents? Is school attendance free of charge? What is the degree of social inequality, ethnic diversity, and vertical social mobility in the society of a given region? Is competition for students between schools open or not? To what extent is the ethnic composition of a population a source of social strife? Are the problematic or socially deprived areas of education located in urban or rather in rural regions? What data sources are available to scholars?

Major recent topics of geographers of education in the United States and the United Kingdom have been school segregation, desegregation busing, the competition between private and public schooling, educational inequalities that arise from residential sorting, the flight from working-class schools, repositioning of inner-city schooling, relations between family income and educational attainment, the neoliberal education market, and the studentification of urban areas. In states that have no tuition fees, a less hierarchic class society, less segregation, or high-quality public schooling, the focus has been on other topics. Accordingly, different research cultures and citation networks have developed within the geography of education. To avoid such national specialties, the following sections of this entry discuss three research topics of general relevance.

**Organization, size, and location patterns of elementary schools**

Siting primary schools in sparsely inhabited rural areas is highly controversial in many countries. It requires more geographical, historical, and cultural expertise than siting institutions of secondary or tertiary education in urban agglomerations. Discord between communities and school administrations frequently arises when educational planners or school administrators show little sensitivity to regional and cultural diversity, the importance of local identities, social cohesion within villages and small towns, and the symbolic meaning of places. Some politicians and administrators have been unaware that planning principles and guidelines concerning the location and organizational structure of schools (e.g., minimum size of schools) should be flexible and adaptable in order to account for regional diversity and local requirements. Sparsely inhabited rural areas suffering from depopulation have had to respond much earlier to declining school enrollment, fiscal constraints, and the difficulty of attracting skilled teachers than more densely populated urban agglomerations. They have therefore had to experiment with specially designed classes, new organizational forms of schools, and new teaching methods in small schools, where children from several grade levels are taught in one classroom.

In rural areas of many states, two different concepts of regional educational planning compete (for details see Kramer 1993). That of large centralized elementary schools to which students are transported over long distances by school buses or parents is claimed to increase financial efficiency, enhance the flexible assignment of teachers, augment the variety of study courses,
broaden social equality, improve the quality of education, and exemplify modernity. Proponents of this concept usually overlook its attendant negative implications, such as transporting 6- to 10-year-old children over large distances into social settings unfamiliar to them, confronting the children with values contrary to those of their family environment, exposing them to emotional stress, and unduly reducing their leisure time.

The alternative concept of regional educational planning promotes small decentralized elementary schools located in or near the villages and towns where the 6- to 10-year-old pupils live. Designed for small classes and schools, this approach requires more teachers, and thus greater financial resources, per 100 students than large urban schools. It is less flexible in the assignment of teachers. However, supporters of this concept assert that its pedagogical advantages and many positive extracurricular effects outweigh the higher financial costs. Its adherents have a different view of childhood, are more sensitive to spatial disparities and local needs than advocates of centralized elementary schooling, acknowledge the negative experience many children have with long-distance busing, and are primarily interested in the needs of communities. They maintain that elementary school teachers in peripheral rural villages play an important extracurricular role and have a greater long-term impact on the community than their counterparts in large cities. If sparsely inhabited rural areas are successful in attracting skilled teachers, their schools will be indispensable elements of the community’s infrastructure. They will be places that form and strengthen local identity, social networks, and social cohesion, and that nurture cultural activities and sensitivity to regional environmental problems. In social milieus of that kind, skilled teachers are still key people in the community and have many functions to perform beyond the school (e.g., church choir director, head of the amateur theater, or leader of the volunteer fire department). In urban areas the professionalization of these functions greatly reduces the need for the extracurricular roles of teachers. The closure of the only elementary school in a sparsely inhabited rural area, and the disruption of children’s living and learning spaces, have far more negative social, cultural, and economic effects than in a large metropolitan area.

Clashes between these two concepts become particularly explosive in multiethnic or multilingual areas. Since the advent of compulsory schooling in the late eighteenth and early nineteenth centuries, administrative control of school systems and the language of instruction in elementary schools have emerged as flash points in bilingual areas. During the nation-building process in Europe and North America, one of the primary functions of public elementary schools was to foster national unity and assimilate ethnic minorities. The language of instruction in the early grades of elementary school was of paramount instrumental value in suppressing or emancipating ethnic minorities. School authorities could easily disadvantage autochthonous ethnic minorities by raising official standards for the minimum school and class size, by forbidding the use of minority languages for instruction, and by assigning teachers who did not speak the language of the school’s minority students and were not familiar with their culture.

Cost-efficiency and “rationalization” of educational infrastructure are powerful arguments in favor of centralizing basic education. They are often used as pretexts for the pursuit of political goals. In Hungary the communist system in the 1950s considered all settlements with fewer than 3000 inhabitants as “uneconomical” to maintain, and planned to close thousands of small schools. This decision was also shaped
by the regime’s negative attitude toward rural society, which it considered a feudal vestige that had no place in the new socialist society it wanted to create. In the 1970s the country’s authorities brought about another historically unprecedented decline of scattered farms, small villages, and their educational infrastructures, a demise much more severe than similar trends in capitalist countries (Gyuris 2014, 538–545).

In alpine regions of western Austria, it was reasoned by educational planners that cost–benefit analysis should not be focused on education alone, that it should also include the unintended consequences that school closures have in other areas of society and the economy, and that educational policies and concepts of educational planning should be coordinated with other spheres of regional planning. The same logic applies to urban areas. Without improvement in the housing quality and the reputation of a given urban area, educational policies will not be able to raise low levels of educational attainment there (Bondi 1987; Gordon and Monastiriotis 2007).

When analyzing the political significance of basic education in bilingual areas, researchers find it revealing to compare the size of elementary schools shortly before they were closed. An instructive example is South Tyrol, which Italy annexed after World War I. From 1971 through 1981, a large number of small elementary schools in the Alpine regions were closed, whereas numerous one-class Italian schools in South Tyrol were maintained as long as possible. Of the 66 elementary schools with Italian as the language of instruction, 71.2% had no more than one to three pupils before they were shut down, and not a single one of the schools had more than 10 pupils. By comparison, only 5.1% of the 78 schools with German as the language of instruction were allowed to continue until they shrank to that size. Indeed, 42.3% of the German schools in South Tyrol had more than 10 students in the year preceding closure, but not a single Italian school did (Höfle and Meusburger 1983, 247).

Spatial disparities of educational achievement and attainment

Indicators of educational achievement describe the final result of an educational process. Indicators of educational attainment are used to describe a current status in an ongoing (unfinished) process of education.

The educational achievement of an adult population is measured by indicators such as the percentage of that group that is illiterate, the percentage of its adult members who have completed a certain level of education (e.g., elementary school, vocational school, college, or university), or the median number of school years completed. Indicators of educational attainment illustrate both the flow of students through the hierarchical structure of an educational system and the students’ performance in various types of schools. Spatial disparities in the educational attainment of students can be measured by indicators such as the percentage of a given age group attending a certain type of school, the transition (enrollment) rates of an age cohort to various institutions of higher education, the school survival expectancy or persistence of school attendance (i.e., how many years children remain in the school system), performance on general certificates or achievement tests (e.g., the United States’ National Assessment of Educational Progress; the United Kingdom’s Pupil Level Annual School Census and General Certificate of Secondary Education (GCSE)), the share of dropouts and truants (rates
of unexcused absences at school), and rates of suspensions from school.

Vast ethnic disparities of educational achievement within a state can mirror existing inequalities and potential political conflicts. An ethnic group with a high proportion of adults who are illiterate or have a very low educational achievement is caught in a vicious circle (Gamerith 2004). Autochthonous minorities who are not allowed to have elementary schools teach in their mother tongue, who are offered underequipped schools and poorly qualified teachers, and whose cultural heritage is devalued in or omitted from schoolbooks may demonstrate low performance, low commitment, low aspirations, and high resentment. If the value system of an ethnic minority (e.g., First Nations in Canada, Native Americans in the United States, and Aborigines in Australia) does not meet the expectations and attitudes of a modern society, many students in this group will develop strategies for resisting the rules and goals of the school. This response will create an antischool culture marked by truancy, refusal to study, and academic failure. This behavior then becomes justification for further discrimination, and reduces minority people’s power to compete on equal terms for jobs in the labor market.

Table 1 ranks the ethnic groups in the former Yugoslavia according to their degree of illiteracy shortly after World War II. At that time the illiteracy rates of ethnic groups such as the Gypsies (the term “Roma” was not used then), Albanians, Turks, and autonomous Muslims were still extreme, ranging between 54% and 74%, compared to less than 10% among the Slovenes, Czechs, Slovaks, Germans, Italians, and Hungarians in Yugoslavia. In other words, the ethnic groups at the bottom of the ranking had illiteracy rates that had been surpassed by those at the top 150 to 200 years earlier. Such large disparities of literacy within a nation almost inevitably result in large social and economic inequalities, social distance, cultural prejudice, and political tensions, especially if the social disparities are coupled with religion or the mother tongue, or if the politically dominant group has a higher percentage of illiterate adults than the groups that feel politically disadvantaged.

It goes without saying that the so-called collective status of an ethnic group is just a rough concept. Within each ethnic group, too, there are major regional and social disparities. As shown by Table 2, the educational achievement of the cigány (Gypsies) in Hungary varies with the size of their place of residence: the larger the size of their villages or town, the lower the percentage of Gypsies without a basic education.

The usefulness or explanatory power of indicators of educational attainment and achievement is not stable over time and space. First, it is influenced by the differentiation and selectivity of the school system. A highly differentiated and selective school system provides indicators of educational attainment that are of greater significance than those from a less differentiated and less selective one, where almost all children achieve high levels of education.

One of the greatest challenges in this research field is to explain spatial and social disparities of educational attainment. They are influenced by at least eight groups of factors, and explaining them requires knowledge of a broad range of theoretical concepts from various disciplines.

The foremost factor of influence is the sociocultural milieu in the family. The family background includes the social status of the parents, their educational achievements and professions, their cultural values and attitudes toward education, the time they spend with their children, and the encouragement they give their children. Families as educational settings
The impact of teachers – for example, their qualification and interaction with the community – on the educational performance of children can hardly be overestimated. However, not all teachers have the same training and skills, not all teachers will provide the same inducement to students, not all teachers are willing to commit themselves to meeting the needs of the local community, and not all teachers are socially accepted by the community. High turnover rates among teachers usually indicate that they are not satisfied with the situation at their school. High rates of extracurricular activities that teachers

Table 1  The proportion of illiterate adults (aged 15 years and above) in the Yugoslavian population by nationality, 1948.

<table>
<thead>
<tr>
<th>Nationality</th>
<th>Proportion of total population</th>
<th>Proportion of illiterate adults</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Male</td>
</tr>
<tr>
<td>Slovenian</td>
<td>9.0</td>
<td>2.3</td>
</tr>
<tr>
<td>Czech</td>
<td>0.2</td>
<td>3.0</td>
</tr>
<tr>
<td>Slovak</td>
<td>0.5</td>
<td>4.8</td>
</tr>
<tr>
<td>German</td>
<td>0.4</td>
<td>6.2</td>
</tr>
<tr>
<td>Italian</td>
<td>0.5</td>
<td>6.9</td>
</tr>
<tr>
<td>Hungarian</td>
<td>3.1</td>
<td>8.0</td>
</tr>
<tr>
<td>Romanian</td>
<td>0.4</td>
<td>17.6</td>
</tr>
<tr>
<td>Croatian</td>
<td>24.0</td>
<td>18.1</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>0.4</td>
<td>19.9</td>
</tr>
<tr>
<td>Montenegrin</td>
<td>2.7</td>
<td>24.1</td>
</tr>
<tr>
<td>Serbian</td>
<td>41.5</td>
<td>27.7</td>
</tr>
<tr>
<td>Macedonian</td>
<td>5.1</td>
<td>30.2</td>
</tr>
<tr>
<td>Walachian</td>
<td>0.7</td>
<td>38.2</td>
</tr>
<tr>
<td>Autonomous Muslims</td>
<td>5.1</td>
<td>54.6</td>
</tr>
<tr>
<td>Turkish</td>
<td>0.6</td>
<td>63.9</td>
</tr>
<tr>
<td>Albanian</td>
<td>4.8</td>
<td>73.7</td>
</tr>
<tr>
<td>Gypsies</td>
<td>0.5</td>
<td>74.0</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>25.4</td>
</tr>
</tbody>
</table>

Source: UNESCO 1953, 166. Reproduced with permission.

contribute to the development of the cognitive abilities, motivations, and aspirations of children. However, the sociocultural milieu of a family should not be equated with its financial circumstances. True, in many cases – especially where high school fees are charged – there is a close positive correlation between the affluence of parents and the educational attainment of their children. But affluence is only a proxy variable for other factors, and there are many examples of rich communities whose students have low enrollment rates in institutions of higher education.
Table 2  Percentage of Hungarian Gypsies (Cigány) aged 15 years and above who have not attended school or completed eight grades of primary school, by size of place of residence, 1990

<table>
<thead>
<tr>
<th>Population of town or village</th>
<th>Cigány* older than 15 years who did not attend school or complete 8 grades of primary school (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>Up to 500</td>
<td>63.3</td>
</tr>
<tr>
<td>501–1000</td>
<td>59.7</td>
</tr>
<tr>
<td>1001–2000</td>
<td>56.8</td>
</tr>
<tr>
<td>2001–5000</td>
<td>57.9</td>
</tr>
<tr>
<td>5001–10 000</td>
<td>55.3</td>
</tr>
<tr>
<td>10 001–20 000</td>
<td>57.4</td>
</tr>
<tr>
<td>20 001–100 000</td>
<td>52.3</td>
</tr>
<tr>
<td>100 001–1 million</td>
<td>44.2</td>
</tr>
<tr>
<td>Budapest</td>
<td>33.5</td>
</tr>
<tr>
<td>Hungary total</td>
<td>55.0</td>
</tr>
<tr>
<td>Absolute numbers</td>
<td>45 189</td>
</tr>
</tbody>
</table>

*The Hungarian Census of 1990 used the category “Cigány” for both “nationality” and “mother tongue.” The Census of 2001 offered three alternatives, “Cigány (Roma)” (plural cigányok, romák), “Beás” (plural beások), and “Romani” (plural romanik). The Census of 2011 used one category, “Cigány (Roma),” for nationality, and one, “Cigány (Romani, Beás),” for mother tongue. In the 1990 Census, 142 683 Hungarians declared their nationality as “Cigány.” The total number of Gypsies in that year was estimated to be about 600 000. This terminological muddle reflects the change in political discourse. Until the early 1990s the word “cigány” was used in both scientific and political discourse. It thereafter became politically incorrect and was largely replaced by the term “Roma,” especially in official debates. Since about 2006, people have been using “cigány” as the designation, for Hungary is host to several groups of Gypsies who vehemently reject being called Roma. Because the data are based on self-declaration as cigány, it is scientifically unacceptable to call these people Roma or Sinti.


offer their students or community signal a high degree of commitment. In some countries the impact of teachers on educational attainment is especially strong because teachers are entitled to recommend the type of school a child should attend after completing the fourth grade of elementary school.

The third factor of influence is the school composition. Schools differ strongly with regard to the social background and abilities of their students. Because high achievers and low achievers are sorted into different schools after they pass elementary school, the composition of the student body shapes the values and attitudes prevailing among peer groups. Student culture and peer approval may establish positive or negative stereotypes for ambitious students.

The fourth group of influences can be labeled neighborhood effect. The type of neighborhood in which a child grows up is another very effective predictor of her or his educational performance (Webber and Butler 2007). The neighborhood effect is partly based on the social composition of the resident population. However, it also consists of close social or cultural relations, shared values, a sense of belonging to a local community, and social control within the community. Emotional attachment to a place, the feeling of safety among
their “own” people, the support of strong social networks, and the fear of becoming an outsider in schools located in other areas seem to be strong motives for parents from ethnic minorities and working-class families to keep their children in low-performing schools in their own neighborhood even if they are free to opt for a school in another area (Bunar 2010, 142).

Some authors mention a fifth group of factors of influence, namely organizational structure (e.g., comprehensive school vs grammar school, pupil–teacher ratio, and size of schools). However, the importance that the size of a school has for educational attainment is overestimated in many publications. A large school is not necessarily a better one. Wide curricular and extracurricular choice does not automatically imply that the quality of education or student performance improves. Combined with other factors such as the professional skills and motivation of teachers, or the governance of a school, however, organizational structures and size of classes may have an influence.

The sixth group of factors of influence is the spatial allocation of educational opportunities, the locally offered density and variety of secondary schools and institutions of higher education. Urban agglomerations provide a greater density and variety of schools than peripheral rural areas. In rural areas the distance to secondary schools on average is larger than in urban areas. Among upper-class parents long distances to school have little or no impact on the choice of secondary schools and institutions of higher education, but long distances may deter parents from lower social strata from sending their children to these schools, especially in ethnically segregated areas and areas where free school transport is not available.

It has been shown that the quality of the regional labor market also has a certain influence on educational attainment. A highly differentiated economy, the availability of many highly skilled jobs, and high vertical social mobility influence the occupational and educational aspirations of children by offering them more attractive role models and more economic prospects than a lower-skilled and less differentiated labor market does. Children from working-class and peasant families seem to be more influenced in their educational aspirations by the diversity and qualification structure of jobs offered in a region than children of academics.

Some countries influence educational attainment by harsh political intervention. Highly feudal European territories, where the abolition of serfdom was delayed in the nineteenth century, experienced much lower rates of school attendance and became literate much later than areas with an early introduction of democratic elements and prospects of vertical social mobility. Between 1740 (South Carolina) and 1838 (Ohio), a number of US states adopted compulsory ignorance laws, which excluded “Negro” children from public schools or even declared that anyone who taught slaves to write or employed slaves as scribes would be fined or imprisoned for up to 12 months. Under some communist regimes (e.g., the German Democratic Republic), children from ideologically “unreliable” families were prohibited from attending grammar schools or studying at universities, and states governed by a fascist regime openly restricted the proportion of Jewish students at their universities.

The relations between the various factors influencing educational attainment are so complex that it would be naive or presumptuous to attempt to formulate explanatory models that are equally applicable to all societies, cultures, and spaces. Regression models using individual variables lead to equivocal and inconsistent results. It is the local combination of multiple
deprivations or multiple privileges – represented by a local milieu, knowledge environment, or context – that entails low or high educational attainment. Because high or low educational attainment tends to occur in many different contexts and for quite different reasons (Webber and Butler 2007, 1241), an investigator interpreting causal relationships needs deep knowledge of the society, culture, economy, and value systems of the neighborhoods, places, or regions in question.

Institutions of higher education

Some of the research issues discussed so far are also relevant to universities and other institutions of tertiary education. However, universities pose additional challenges for the geography of education by raising a number of other interesting questions: What factors influence the catchment areas of universities and the spatial mobility of scholars and students? What factors are responsible for the intellectual ups and downs of universities? Why do Nobel Prize winners or other outstanding scholars tend to cluster at such a small number of universities? What impact does a university have on the economy, society, and culture of its town or wider region? How do transnational knowledge networks of scientists evolve, and what long-term consequences do they have? These questions indicate that talent, motivation, and wealth of ideas are not the only characteristics determining how successfully a scholar’s research and academic career develop. A number of external factors come to bear as well. The interaction of these external components yields a spatially localizable societal macrophenomenon called a knowledge environment. A university’s knowledge environment is definable as the result of systematic interdependencies and causal interactions relevant for the generation and diffusion of scientific knowledge. Knowledge environments have an impact not only on the generation, evaluation, and legitimation of new scientific knowledge but also on the reading and interpretation of texts. Universities are not simply locations but social spaces, epistemic venues, and knots of scientific networks. Studying local knowledge environments and worldwide communication spaces of academics opens up additional horizons for explaining academic careers and scientific achievements.

What does a scientific milieu or knowledge environment consist of and how can one explain its impact on the research processes and careers of scholars and students in a nondeterministic way? Academics depend on financial resources and research infrastructure (e.g., libraries, laboratories). They are affected by formal and informal organizational structures, institutional regulations, and expectations of their social environment. A university’s knowledge environment can have a bearing on whether and how soon new scientific concepts, practices, or technical innovations are accepted and acted upon. It also plays a part in determining the possibility of conducting expensive experiments, becoming part of important international networks, hearing promptly about crucial new developments, and meeting with agreement (or criticism) when new ideas are aired. The research interests and fields of interaction of young scholars are influenced by the experiences, scientific stimuli, and critique of other scholars active at the same site, and by the encouragement received from key people and critical audiences at various places. Spatial proximity is pivotal because it facilitates spontaneous face-to-face contact, but it does not guarantee interaction. Yet, for all the importance of clustering and proximity, the key words are actually communicative interaction and ways to transcend the silo principle.
A knowledge environment should not be thought of as an independent variable that directly influences all actors in the same way. It is a locally available potential or local range of resources. It stands for challenges, stimulation, opportunities, and support networks that can be used, overlooked, or ignored. A knowledge environment also consists of resistance and obstacles. Some scholars will be paralyzed by local restrictions and opposition to their ideas; others will overcome them. The local potential of a knowledge environment should be clearly distinguished from the scientific results arrived at within it. No one can reliably predict whether and how someone will make use of a milieu’s potential. Knowledge environments vary from place to place. The discrete elements of a knowledge environment are never stable or immutable; they are constantly changing. Nevertheless, the scientific reputation of a knowledge environment may be remarkably consistent because of a university’s high recruitment standards and scientific attractiveness.

The stronger the scientific reputation of a university and the greater its resources, the more it acts as a magnet for talent and brains, the more it can draw scholars from prominent other universities, the more international its catchment area of students is, the more it disseminates ideas and scholars throughout the world, the more professorships other universities offer its young researchers, and the more impressive its global networks are. However, the spatial expanse of scientific relations and the areas from which a university gains its professors and students can also be drastically affected by political and religious conflicts, language barriers, and the economic and political situation in the home countries of scholars and students.

Figure 2 illustrates the regional provenance of Heidelberg’s students in the Faculty of Theology between 1560 and 1620, a period when Heidelberg University stood as a European center of Calvinism and enjoyed its first golden age, which was abruptly ended by the outbreak of the Thirty Years War.

Since medieval times, scholars and students have been expected to be geographically mobile for various reasons. First, mobility can inspire new ways of thinking and bring about new networks and cooperation in the world of academia. Second, in order to remain creative and competitive, a university’s knowledge environment must continually be replenished by the best available candidates bringing in new research questions, scientific methods, and networks. Faculties that systematically favor their own former students for professorships run the risk of narrowing their scientific perspectives and eventually diminishing the quality of their scientific performance.

Third, the career mobility of a scholar can serve as one of many indicators of his or her scientific potential. If two or more faculties confirm that the scientific achievements, competence, and other merits of an academic warrant that individual’s appointment to a position as a lecturer or professor, then the uncertainty about the person’s scientific potential is much lower than if confirmation comes only from the faculty where the candidate studied and graduated. An academic who has worked successfully at various universities generally has a greater range of experience, more extensive personal networks, and a better academic reputation than one who has worked in only one institutional setting.

Fourth, revolutionary scientific concepts and findings that contradict common paradigms of mainstream science are not immediately accepted elsewhere. Despite the Internet, it can take years or even decades before new theories or methods are recognized within one’s discipline or in adjacent ones. The surest way to transmit contested
epochal scientific findings and methods from place A to place B is the geographic mobility of scholars who possess the relevant knowledge.

Other variables describing scientific relations and the migration of ideas are networks of correspondence, collaboration in joint research projects, co-authorships in joint publications, the destinations of sabbatical leaves, the distribution of fieldwork sites, and the distribution of honorary doctorates.
Further perspectives

Further perspectives include the impact of political ideologies on educational systems and educational priorities; the varying role of the meritocratic principle in different societies; the relevance of educational inequalities in political discourses; the increasing internationalization and marketization of higher education and the attendant consequences thereof; the development of international student flows and the impact that alumni networks in foreign countries have on cultural and economic policy; the schooling of underprivileged ethnic minorities and of children with special needs; and spatial differences in the philanthropic support of education.

A highly challenging research issue is the description and analysis of knowledge environments and their impact on learning and research processes. Various indicators help describe and analyze a university’s staffing, financial resources, research equipment, intellectual attractiveness, scientific achievements, and reputation. It is more difficult to reconstruct interactions, networks, and knowledge exchange retrospectively by conducting interviews and analyzing documents, biographies, and protocols. The impact a knowledge environment has on the learning and research processes and scientific achievements of scholars becomes apparent only after years or even decades.

Structuration theory, actor-network theory, environmental psychology, environmental phenomenology, gestalt psychology, interaction analysis, and communicative constructivism are able to help explain the interactions between knowledge environments and the learning processes of individuals to a certain extent. But further research and theoretical concepts dealing with social macrophenomena such as knowledge environments are needed. Students of knowledge environments are less concerned with the discovery of causalities in the sense of universal laws than with the causal reconstruction of the network of interactions. In their empirical analysis of social macrophenomena, they do not seek abstraction and maximal simplification as much as specification and an adequate explanatory complexity.

SEE ALSO: Children and youth; Culture; Indigenous knowledge; Inequality; Infrastructure and regional development; Knowledge-based economy; Neighborhood; Place; Regional development policies; Regional inequalities; Uneven regional development

References


GEOGRAPHY OF EDUCATION AND EDUCATIONAL SYSTEMS

Geographische Arbeiten 87. Heidelberg: Department of Geography, Heidelberg University.


Further reading


Geography of evolution

Amanda J. Chunco
Elon University, USA

Geography has been central to the study of evolution since the idea of natural selection was formalized by Charles Darwin and Alfred Russel Wallace in the 1850s. Observations of patterns of species distributions provided critical evidence underlying early work in the field of evolution. Today, the availability of modern genetic and geospatial techniques is furthering our understanding of the myriad ways in which the environment shapes the organisms that live there.

For both Darwin and Wallace, the idea of gradual change in a population or species over time was heavily influenced by their observations of spatial patterns in the distribution of current and fossil species. In particular, both men observed: (i) differences in biological communities in different geographic regions which suggested that the local environment shaped the species that lived there, and (ii) similarities between fossil and extant species in the same region which suggested that species had changed over time. Both men also began to speculate on the importance of geographic isolation to divergence and speciation — a theme that is still critical to the study of evolution. Darwin noted that island fauna were morphologically similar to, yet distinct from, fauna from the nearest continent, while Wallace observed that geographic barriers, such as large rivers, frequently separated morphologically similar species. Many of the ideas raised by Darwin and Wallace are still important in modern evolutionary biogeography.

Following the foundation laid by Darwin and Wallace, a second era of advancements in the study of evolution occurred during the first half of the twentieth century when ecologists began tackling the question of how similar species coexist. This was a difficult theoretical issue given Darwin’s arguments for the importance of competition (which would likely be greatest between similar species) and Geory Guase’s experiments in the 1930s that demonstrated extinction of one of the competing species. Work by Joseph Grinell, Charles Elton, and G. Evelyn Hutchinson led to the formalization of the concept of the niche as a way of describing how species partition time, space, and resources to minimize competition. Although the meaning of niche continues to be debated, Hutchinson’s definition of the niche as the multidimensional set of environmental parameters that allow the long-term existence of a species remains commonly used (Hutchinson 1957). This work furthered our understanding of how the abiotic and biotic environments shape the distribution of individual species and entire biological communities. This era also saw the advent of the modern synthesis, a flurry of research activity aimed at integrating new advances in genetics with natural selection during which the field of speciation developed in earnest. Ernst Mayr (1942) extended earlier work by Darwin, Wallace, and Grinell in identifying a fundamental role for geography in the formation of new species. In Mayr’s paradigm, speciation could occur only with geographic isolation between two populations. Whether speciation could occur in the absence of isolation remained highly controversial and was a leading area of theoretical and experimental research throughout the twentieth century (Coyne and Orr 2004).
The study of geographic patterns in population divergence and speciation remains an important focus of research. When describing the geographic relationship between different populations or species, biologists typically recognize three specific distribution patterns categorized by the relative ease of migration and gene flow between the populations: allopatry, sympatry, and parapatry. At one extreme, allopatric populations are physically isolated from one another via a physical barrier or a long distance resulting in no (or very low) gene flow between the populations. This physical and genetic isolation leads to the accumulation of genetic differences as a result of local adaptation and random genetic drift. Over time, sufficient genetic divergence can result in speciation.

In contrast, sympatric populations co-occur within the same geographic area. Because the potential for gene flow is high without physical barriers to prevent interbreeding, sympatric speciation was considered a theoretical impossibility, or at least exceedingly rare, by many evolutionary biologists for much of the twentieth century. This view was not, however, universal, and several theorists (including John Maynard Smith) outlined conditions that could potentially result in speciation, even in the absence of physical barriers (Coyne and Orr 2004). Recent work has shown that sympatric divergence and speciation is theoretically plausible under specific circumstances, and several empirical cases of sympatric speciation have now been documented (Coyne and Orr 2004). Outlining the environmental and genetic conditions that can promote sympatric speciation and documenting empirical examples remains an active area of research.

Whether two populations are allopatric or sympatric can have profound implications for the evolutionary trajectory for both groups. If two sympatric populations compete for the same resources, competition may lead to the extinction of one population. For example, competition with invasive species can result in declines in population size and even local extinction for native species, which is a major current conservation concern. Alternatively, both populations may persist, but diverge in terms of morphology, resource use, and behavior. This will result in heightened differences between the two populations in areas where they are sympatric, but not in areas where they are allopatric, a pattern (and process) referred to as character displacement. Allopatric populations, however, will diverge as a result of geographic separation (i.e., isolation by distance) and adaptation to local conditions. Even populations that are widely separated environmentally and geographically, however, may retain similarities in their basic requirements. The degree to which populations retain similar niche requirements (i.e., niche conservatism) versus adapting to new environments (i.e., niche evolution) remains a critical topic in evolution.

Allopatry and sympatry have been the most commonly studied geographic patterns of distribution and represent the extreme ends of the spectrum of physical isolation, from complete isolation (i.e., allopatry) to complete contact (i.e., sympatry). In between these extremes lie parapatric distributions. Parapatric populations are not isolated by any physical barrier, but populations occupy different geographic space and only overlap at a relatively narrow contact zone. Typically, the distribution of parapatric populations spans a gradient of environmental conditions resulting in divergence as populations adapt to local conditions. Because gene flow is hindered both by environmental divergence (due to local adaptation), and by distance, a parapatric distribution can result in speciation. Several other geographic patterns of speciation have been studied, although less prominently than the three mentioned above (for a review, see Coyne and Orr 2004).
Identifying a species’ distribution as allopatric, sympatric, or parapatric depends heavily on the likelihood of contact and gene flow between populations. In practice, it can be very challenging to classify a given population into a single discrete geographic category (i.e., allopatric, sympatric, or parapatric) for several reasons. First, the relationship between distribution and degree of gene flow is scale-dependent and relative to particular organisms. Populations of poorly dispersing organisms will be considered allopatric at distances much closer than well-dispersing organisms. Thus, even two populations that appear to co-occur at a coarse geographic resolution, and would therefore be considered sympatric, may actually show isolation at a finer geographic scale (i.e., microallopatry). Second, populations may be isolated by habitat in a way that limits the potential for gene flow even in the absence of any physical barrier. For example, parasites that reproduce in different host species may have very low potential for gene flow, while still being in relatively close physical proximity. Finally, it has been argued that the very use of terms that demarcate discrete physical categories to describe phenomena that, in reality, span a continuum from complete isolation to complete overlap should be abandoned in favor of specific measures of genetic isolation (Butlin, Galindo, and Grahame 2008). Given the heterogeneous nature of both habitats and species’ distributions, a single species may exhibit the entire spectrum of conditions, from complete sympathy to complete allopatry across its range, making it difficult to place any given set of species or populations into one discrete category.

Determining whether a speciation event occurred under a specific geographic pattern is further complicated by the fact that species distributions are not static, but shift in space over time. Two populations may move repeatedly from sympathy to allopatry and back repeatedly over geological time frames. Because current species result from speciation events in the distant past, the current distribution of a species cannot definitively represent its distribution when speciation occurred. If two related species are currently isolated, speciation in allopatry is supported. This observation cannot, however, eliminate the possibility that divergence occurred during a period of sympathy, followed by a period of geographic isolation. This shift between allopatry and sympathy may actually occur during the speciation process. For example, two populations may start to diverge in allopatry, but come back into secondary contact before the speciation process is complete. If relatively little genetic divergence has occurred, the two populations may merge back into a single interbreeding population. If divergence has occurred to an extent that interbreeding is deleterious (e.g., because it produces less fertile offspring), then there will be strong selection pressure on individuals to mate only within their population. This process can strengthen the degree of isolation between the now sympatric populations through a process termed reinforcement (Coyne and Orr 2004).

Although research on patterns of distribution remains important in evolution, the rapid advances in GIS (geographic information systems) technologies, genetic tools, and widely available high resolution environmental data have led to intense research activity at the interaction of geography and evolution over the last decade. Three areas of research in particular are worth noting.

First, the field of phylogeography, which was launched in the late 1980s (reviewed in Hickerson et al. 2010), has provided significant insights into the historical origins of a growing diversity of taxa. Phylogeography uses genetic tools and models to elucidate the spatial and temporal patterns of genetic lineages. This work can provide insight into the ecological
GEOGRAPHY OF EVOLUTION

and evolutionary forces shaping current distributions. Currently, phylogeography is used in conservation by highlighting geographic areas with high phylogenetic diversity and by elucidating how past periods of climate change influenced biodiversity in order to predict future responses to current anthropogenic climate change (Hickerson et al. 2010). In addition, advances in computer software have resulted in phylogeographic information systems that can visualize geographically rooted phylogenies thereby increasing the ability to explicitly link genetic changes to a physical space.

Second, ecological niche models that integrate geographic and environmental data to predict the distribution of species are increasingly being used to study evolutionary processes. New software packages (including DIVA-GIS, Maxent, and GARP) along with an increasing availability of free, high resolution, geographically referenced environmental (including temperature, rainfall, and land use) and species locality data has greatly simplified the process of creating niche models. The increasing interest in niche models has led to a resurgence of interest in the theoretical definitions of the niche initially defined decades ago. Ecological niche models have shown great potential for investigating how the abiotic and biotic environment contributes to selection, divergence, and speciation (Hickerson et al. 2010). Improvements in paleo-climate models are allowing niche models to explicitly test hypotheses about both past distributions and the environmental and evolutionary drivers of current species distributions. This work can provide insight into patterns of niche conservatism. Finally, the combination of phylogeographic tools and ecological niche models can provide a particularly powerful approach for testing how species respond over evolutionary timescales to changing environments and for predicting how species will respond to ongoing anthropogenic global change.

Third, the field of landscape genetics is providing insight into how current human modification of the landscape alters the genetic makeup of resident populations. By geographically linking land use with genetic diversity, landscape genetics can address questions related to demography, gene flow, and local adaptation in human altered environments. This approach can be used to prioritize conservation efforts for habitats and populations that will maximize genetic diversity and evolutionary potential.

SEE ALSO: Biodiversity; Biogeography: history; Ecogeography/macroecology (range and body size); Endemism; Global environmental change: human dimensions; Island biogeography; Niche theory and models; Paleoecology

References


Further reading


While the two parts of the term “geohumanities” – “geo” and “humanities” – have a long and diverse history in and of themselves, their merging is a very recent phenomenon. In 2011 the edited book GeoHumanities: Art, History, Text at the Edge of Place (Dear et al. 2011) was a key product of a series of events designed to explore the intersection of geography (mostly cultural and historical geography but including elements of geographic information science (GIScience)) and other humanities disciplines. These included a Geography and Humanities Symposium at the University of Virginia in 2007 organized by the Association of American Geographers (AAG). Subsequent meetings of the AAG included highlighted sessions on “Geography and the Humanities” that, in turn, led to the edited volume GeoHumanities (Dear et al. 2011), as well as the edited collection Envisioning Landscapes, Making Worlds: Geography and the Humanities (Daniels et al. 2011). In 2015 a new AAG journal, GeoHumanities, was launched. In a parallel development, in 2013 the Alliance of Digital Humanities Organizations (ADHO) recognized a GeoHumanities Special Interest Group, intended to address the spatial, spatial-temporal, and place-related aspects of the digital humanities.

These two formalized articulations of geohumanities indicate not only a growing interest in and commitment to cross-disciplinary thought and practice, but also something of the differing aims and objectives, emphases, and expertise brought to bear in the making of this loosely contained field. The involvement of the AAG, for example, speaks to the continuing evolution of a modern-day discipline thoroughly engaged with the broad range of humanist thought, manifest, for example, in an emphasis on the “geo” as a human making of worlds, but also, and more recently, of a more-than-human worlding that both draws on and shapes the Earth. By contrast, the input of the digital humanities is more targeted; it indicates a desire within humanities disciplines such as English literature and history to bring to light the role of increasingly everyday geolocation technologies (such as Global Positioning System, or GPS) in shaping people’s spatial imaginaries, but also to use technologies such as geographic information systems (GIS) in turn to visualize imaginaries that are otherwise glossed in cultural texts, from diaries to poetry. The “geo” here thus refers to the location and distribution of phenomena over Cartesian space, but also to relational accounts of people and place.

Clearly, there are many moments when these two articulations of a geohumanities overlap: GIS, for example, is very much indebted to the spatial science turn in the discipline of geography, while an emphasis on cultural texts as “readable” has helped shape a cultural geography concerned with understanding landscape as just such a text. What is also clear, however, is that this twofold emergence of a geohumanities undercuts any easy notion of shared objectives, concerns, approaches, and methodological
apparatus. Their moments of intersection and difference owe much to the development of disciplinary specialisms, but also to the particular framings of the “geo” that are brought to bear. This highlighting of “geo” is important precisely because it is such a complex, and fecund, term that has been afforded meaning and import from the classics through to the (European) Reformation, the Renaissance, and the Enlightenment. According to the Oxford English Dictionary, it derives from the ancient Greek γεω (Earth) that appeared in a number of words including γεωγραφία (geography). The career of “geo” really took off in the fourteenth century with the arrival of familiar terms such as “geometry” and (now) less familiar ones such as “geomancy.” “Geography” as an English word did not appear until the fifteenth century, while entirely English inventions that utilize the meanings that “geo” brings with it did not appear until the early nineteenth century when adjectives such as “geohistorical” and “geomorphic” began to appear.

“Geo,” then, brings with it a dense association of meanings related to the Earth, while also having the capacity to add to and transform the meanings of other words, usually working in symbiotic combination with those that are of Greek or Latin origin. “Earth” itself is a remarkably complicated term that brings to mind both the planet we inhabit and some of the stuff it is made of – particularly the earthiness of soil. Somewhere between these two, “Earth” also refers to the ground in general, the lowest, basest and dullest of the four elements (earth, air, water, fire) in classical philosophy, and the world as a dwelling place for humans (contrasted with heaven and hell). Adding “geo” to a word clearly brings a lot of baggage, and a lot of possibilities.

To bring “geo” into proximity with the term “humanities” is to set in train a number of possibilities. The term “humanities” itself has its origin in European scholastic traditions of the fifteenth century, and referred to a branch of learning that was opposed to theology. Its focus was on ancient Latin and Greek, and the various associated traditions of philosophy, rhetoric, history, and grammar. By the early seventeenth century, in the work of Francis Bacon, the universe of knowledge had been divided up into three fields – divinity (theology), natural philosophy, and the humanities. Today, the humanities are generally thought of as those disciplines that engage with human culture and meaning, and that utilize interpretive, qualitative methods rather than scientific, quantitative ones. These typically include literature, history, philosophy, art, and music as well as modern languages. The discipline of geography has historically been included in the humanities but less so in recent years when it has been demarcated more as a combination of social and physical sciences.

At a very broad level, then, geohumanities indicates an interest in the human condition, certainly, but a condition that is somehow connected with the materiality and processes of the Earth itself. It also indicates an interest in how proximities and distances, whether absolute or relational, and the spatial imaginaries that underpin these, such as Cartesian, topographic, and topological, are fundamental to such a condition. The full richness of the term, however, only emerges with further archaeological digging.

Geography/humanism/the humanities

Importantly for us, the use of the prefix “geo” became widespread at about the same time as the term “humanities” became meaningful. Both arose in the European Renaissance, and both had their origins in classical Greek and Latin literatures and thought. Indeed, the sundering of humanities as a distinct sphere of learning, apart
from divinity, was an outcome of the broader emergence of Renaissance humanism, with its insistence on a human-centered view of the world wherein “Man” became “the measure of all things.” Centering of the human brought with it a whole new range of learning – both the humanities, as we now understand them, and natural philosophy (later to become natural history, and then the natural sciences). Central to Renaissance humanism was a strong belief in the very human power of reason and rationality, a belief that, of course, also underpinned a natural philosophy that was capable of investigating and understanding the Earth and the cosmos, despite our more modern tendency to worry at a humanities/sciences epistemic divide (Cosgrove 1984). As the notion of humanness became centered in learning, so too did attention turn from the heavens to the Earth and the human relationship to it. Geography, geology, geometry, and other “geo” subjects all experienced a revival thanks in part to the translations of classical texts by Strabo, Ptolemy, and others that began to appear in centers of learning (Glacken 1967). It is this turn toward the Earth – toward the “geo” – that arguably enabled the flourishing of humanistic thought. Importantly, it also set in train a way of thinking about phenomena – a geological perspective, one might say – that organized social as well as physical phenomena. That is, the visualization techniques developed by natural scientists, such as the geological cross-section and its delineation of formations and boundaries, produced a stratified ontology wherein layer built on layer. In turn, these were to indicate “deep time” – a historicism necessary for such sedimentation processes to occur. This idea of a stratification was to percolate through the Romantic literature of the eighteenth century, and its presentation of social groups.

To be sure, the reorganization of geography in the late Enlightenment as a synthesizing discipline that sought to uncover general Earth processes did much to sunder such a utilitarian endeavor from what became regarded as an unscientific, aesthetic preoccupation with articulating a distinct sense of place; indeed, the current institutional organizing of geography into either physical or human shows the force of such a division. Despite intermittent efforts through the nineteenth and twentieth centuries to work across such a divide, it was not until the 1970s that a concerted effort emerged that sought to ground philosophical concerns with “a life worth living” in an inquiry into how we live with, as well as on, the Earth. With the rise of “humanistic geography,” geographers such as David Lowenthal (also a historian), Yi-Fu Tuan, Edward Relph, and Anne Buttimer combined a knowledge of the field’s older tradition as a humanities discipline with newer interests in philosophies of meaning (such as transcendentalism, existentialism, and interactionism) to create a mode of inquiry that challenged the then hegemonic versions of geography as a spatial science. This new humanistic geography centered the formally marginalized questions of what it was to inhabit the Earth, and how individual and shared meanings arise from human–environment interactions. Traditional humanities values were very much part of this scholarship, including: centering the human as the basis for experience, knowledge, and understanding; looking to interpretive rather than scientific modes of inquiry; and pioneering the use of texts and images as evidence for the human relationship with the Earth. The notion of “geo” as of the “Earth” (with all its attendant meanings) was thus at the center of this field of inquiry. Indeed, Tuan (1991) attempted to provide a humanistic definition of the whole discipline by focusing on what he considered to be a key intellectual
problematic but also an everyday preoccupation for people, that is, how humans make the Earth into home. For Tuan, spatial science risked a reductionist view of humanity, but also of the Earth itself. Geography, he wrote, for all the technical sophistication of its specialized subfields, is not remote or esoteric knowledge but rather a basic human concern. Humans everywhere seek to understand the nature of their home ... if in every instance geographers insist on precision and quantifiability at the expense of a roundedness of view and resonance, they can never hope to understand the earth as home. (Tuan 1991, 99, 106)

The strong sense of a humanistic geography informed by philosophies of meaning and opposed to more quantitative and scientific forms of inquiry gradually percolated across the general field of human geography and became less distinct as an approach, despite several rounds of reinvigoration. Nevertheless, a deeper sense of geography as a discipline rooted in the humanities continues to exert its influence and it is this sense of “humanistic” inquiry that helps shape the emergent field of geohumanities.

One aspect of this deeper and broader version of humanistic scholarship is its interdisciplinarity, or even transdisciplinarity. Geohumanities is not a disciplinary project, despite the immediate efforts of the AAG to bring both the concept and the journal with that name into being. The subtitle of the journal GeoHumanities is Space, Place, and the Humanities. This collection of key words points toward a serious effort to have what might be thought of as disciplinary concerns with space and place become the subject of both conversations between disciplines (interdisciplinarity) and emergent themes in the spaces between disciplines (transdisciplinarity). This conversation, in turn, recognizes a wider interest in space and place across humanistic disciplines such as literary studies, history, and philosophy.

In literary studies, perhaps emerging with the groundbreaking work of Raymond Williams and Edward Said, geographical imaginations have become active components of the ways literature and other representational forms are interpreted. Space and place are no longer regarded as settings or backdrops for literary action, but are themselves actors in the construction of literature (Harding 2014). Similarly, history has seen an enthusiastic embrace of what has been called an “environmental history,” exemplified by William Cronon’s Changes in the Land: Indians, Colonists, and the Ecology of New England (1983). More recently, a “spatial history” has been identified, with a landmark contribution from Paul Carter, The Road to Botany Bay (1987), followed by initiatives such as the “digital history” emerging from Stanford’s Spatial History Project.

Philosophy, meanwhile, has seen something of a rebirth of interest in space and, particularly, place, as key elements of what it is to be a human on the Earth. There is no doubt that Deleuze and Guattari’s “geophilosophy” has had a profound impact on academia, insofar as it poses questions not as to what the Earth and its inhabitants are, but how they work. It is a Deleuzian-inflected feminist philosophy, however, that has perhaps delved most deeply into the nature of the “geo” as a means of getting to grips with the human condition as an embodied one. A key preoccupation here is the notion of an “earthly body,” that is, one that is no longer constrained by the model of a sovereign corporeal subject, but is immersed within, draws sustenance from, engages, and negotiates a material realm that cannot always be reduced to or even connected with an imaginative and discursive realm. Such bodies are certainly open to cultural intervention – they can be inscribed on as a text, one might say – but this does not thereby negate their earthy dimensions, which provide for a body’s form and capacities (Grosz 2008). Importantly
for this reconfiguration of the “geo,” such an approach provides a sense not only of the complex material proximities and distantiations that continually rework the corporeal geographies of the body but also of other temporalities, such as the evolutionary tempo of species and the deep history, or “Earth history,” within which all life, as well as the materials that enable and aver this combination, is embedded. It can reference the capacity of corporeal bodies to undertake work, a capacity enabled and limited by their interaction with other bodies, some human, most not. And it can prompt explorations of the plurivitality of the body, taking into account, for example, the work of microbial viruses and bacteria, as well as the earthly components of air, water, earth, and fire, all of which not only subvert the idea of a sovereign subject in a fleshy container, but prompt an ethics of responsibility to the these “nonhuman” presences. Such a feminist philosophy has permeated numerous academic spaces, particularly in the arts, where notions of a “dark ecology,” for example, and “becoming animal” have taken hold.

These kinds of developments have sometimes been referred to as a “spatial turn” in the humanities, reflecting the belief that space and place were previously neglected or reduced to passive roles. Certainly, the importance of space and place as analytics has become lauded across academia, helped in large part by the translation into English of key texts by prominent social theorists such as Henri Lefebvre and Michel Foucault. These developments have also been prompted by what has been termed a “material turn,” wherein emphasis is placed on interrogating and moving beyond the art/science divide in writings by Deleuze and Guattari, Elizabeth Grosz, Rosi Braidotti, Isabelle Stengers, and Karen Barad. Taken together, what these turns make clear is that space, place, and earthliness – all of which have become associated with

the term “geo” – are no longer seen as the property of the discipline of geography; there is a distributed sense of the importance of these terms, enabling in turn more creative forms of transdisciplinary conversation.

Geohumanities/spatial humanities/digital humanities

Bearing in mind the broad and long-term pervasiveness of the influences noted, the more immediate cause of the increased “visibility” of geohumanities is, undoubtedly, rooted in the widespread and relatively sudden uptake of geographically coded data across humanities disciplines thanks to a combination of GIS, GPS, and various forms of locative media attached to GPS-enabled mobile devices. For this reason the prefix “geo” has had something of a second renaissance since the turn of the twenty-first century. In addition to the long-standing uses of “geo” in terms such as “geostationary” and “geochemistry,” we now have “geo” entering popular culture thanks to the idea of being “geolocated,” of data being “geocoded,” and of the ability to partake in “geotagging.”

What is not immediately apparent is what these terms have to do with the older idea of “geo” as referring to the Earth and all of the humanistic implications that come with that association. Rather, underpinning all of these technologies is a framing of the Earth as Cartesian – that is, composed of four dimensions – such that an absolute location can be discerned. This mode of understanding and engaging with the Earth has its own historical trajectory, being implicated in classic and Renaissance cartographic efforts, the rendering of the Earth and its inhabitants as machine-like entities in the Enlightenment, and especially the emergence of a military science that relied on a universal cartographic language
to render territories knowable, and a precise mathematical knowledge of how projectiles moved through the air (Gregory 1994). Small wonder, then, that geotagging, which refers to the addition of locational metadata to almost any form of media, has been facilitated by a twentieth-century network of military satellites that provide precise locations. These systems were only completely freed from military use in 2000, since which date the prefix “geo” has exploded as a way of denoting that something is precisely locatable.

The term “spatial humanities” tends to refer to a particularly close partnering of GIS and the geospatial web with traditional humanities scholarship, and is perhaps best illustrated by work undertaken by historians and historical geographers who use GIS alongside other modes of visualization. According to Bodenhamer, Corrigan, and Harris, there are pitfalls as well as promises here in that

Spatial humanities, especially with a humanities-friendly GIS at its center, can be a tool with revolutionary potential for scholarship, but as such, it faces significant obstacles to knowledge that, at first glance, rest on different epistemological footings. Humanities scholars speak often of conceptual and cognitive mapping, but view geographic mapping, the stock in trade of GIS, as an elementary or primitive approach to complexity at best or environmental determinism at worst. Experts in spatial technologies, conversely, have found it difficult to wrestle slippery humanities notions into software that demands precise locations and closed polygons. At times, applying GIS to the humanities appears only to prove C.P. Snow’s now-classic formulation of science and the humanities as two separate worlds. (Bodenhamer, Corrigan, and Harris 2010, viii)

Bodenhamer, Corrigan, and Harris correctly identify a deep humanist distrust of GIS as a tool that arose out of the sciences, with all its privileging of certainty and authority. To be sure, GIScience practitioners have themselves acknowledged such a disciplinary lineage and its epistemic implications. And yet, some have also pointed to the need to interrogate such presumptions critically and, in practice, to augment the visualizations of GIS with other forms of knowledge-making, such that its conceptual thresholds are demarcated.

Perhaps a key moment in the negotiation of how GIS was to be brought into alignment with the notion of a geohumanities came with the 2011 publication of GeoHumanities: Art, History, Text at the Edge of Place (Dear et al. 2011). GeoHumanities represents a broader interdisciplinary and transdisciplinary convergence of geography and the humanities that centers the concepts of space and place rather than the techniques and epistemologies of GIS. Indeed, the book privileges creativity, with contributions from architecture, the arts, English, history, and curatorial practice as well as geography; and it is in the midst of these discussions that the potential of GIS for creative and historical research is situated. What is more, GIS is augmented by other forms of digital geovisualization, including the kinds of virtual reality environments most often associated with the world of gaming. If we extend our technological imaginations, Harris, Rouse, and Bergeron argue in this collection, we might “contribute to a Humanities GIS that is sensuous, reflexive, and sensitive to issues of humanities place as well as geographical space” (2011, 238). Such a call resonates with other discussions within the collection of the place of media within geohumanities, including a photographic essay on Broadway in Los Angeles by Robbert Flick (2011), and a critical essay on the Sophie Ristelhueber’s artistic installation, Fait, by Caren Kaplan (2011).

It has become apparent in the early years of the twenty-first century that framings, media, and interfaces provided by GIS alongside GPS,
mobile digital devices, locative media, and so on have produced new ways of thinking about classically humanistic problems. One of the most prominent of these is the practice of "distanced reading" proposed by Franco Moretti and the Stanford Literary Lab. Moretti and his colleagues have used the resources of the web to analyze published literature as a source of "big data." That is, rather than reading one or two books in depth (for a close reading), Moretti argues for the analysis of thousands of texts at once. Large patterns over long periods of time are often presented digitally, hence the title of his book *Graphs, Maps, Trees: Abstract Models for a Literary History* (Moretti 2007). One of the pamphlets produced by the Stanford Literary Lab, for example, analyzes the texts of 2958 British novels. It finds that, for instance, the spaces of the novel over time expand from relatively small domestic and rural spaces, through the spaces of the city, to the expansive realms of adventure and science fiction novels. In a century that featured the invention and triumph of the railway as well as the emergence of Britain as the world’s first urban nation, this is hardly surprising (Hueser and Le-Khac 2012).

This kind of banal outcome, articulated by an often visually stunning image, is not unique to the Moretti et al.’s distant reading projects. Mike Crang has noted similar outcomes across the digital humanities, suggesting that

Our banal social lives become digitally mediated and can be subject to quantitative encapsulation through lexical analysis. For instance, Alan Mislove and colleagues applied a word-rating system — scoring positive and negative connotations — to US-based geolocated tweets to produce stunning time lapse maps of the "mood of the nation." Similar approaches have correlated postings with stock market movements, and yet so far the conclusions have been banal. The poetics and affective power of the visualisation have often been more powerful than the supposed “result.” (2015, 354)

Despite these analytic shortcomings of the still young digital humanities, it is clear that aspects of the new field are key to the even younger field of geohumanities. Geohumanities would not exist as a word or interdisciplinary field without the emergence of a spatial web of technologies that produces massive amounts of located data, some of which can be broadly “humanistic.”

**Geo/geo/geo**

Geohumanities, then, can be traced through a number of genealogies including forms of classical scholarship, the long history of humanism, the specific history of “humanistic geography,” the spatial and material turns across academia, and the more recent embrace of the digital by humanistic disciplines.

In regard to the future, consider the many meanings of “geo” and the contrast between the older “geo” of the Earth as dwelling place and the newer “geo” of location. Understanding something of the polysemous nature of “geo” helps us to fabricate a future for the only just born, yet very old, project of the geohumanities. The ancient Greek origins of “geo” take us back to a time of radical predisciplinarity. Scholars, explorers, and others such as Strabo, Ptolemy, and Herodotus engaged with the Earth as both a planet (Earth) and the site of human dwelling separate from the heavens. Well before the specialist landscape of the disciplines, there was a general attempt to understand the Earth, and concerns with space, place, and location were at the center of this endeavor. Greek philosophers and historians were also geographers (insofar as these labels were meaningful). Herodotus, claimed as the father of both anthropology and history, spent much of his time trying to find the source of
the Nile and might reasonably be claimed as the father of geography too. His project was one of constant accounting for, and description of, the heterogeneity of the inhabited Earth. Meanwhile, the librarian of Alexandria, Eratosthenes, was busy measuring the Earth and developing the system we now know as latitude and longitude, and which locates our every thought and move through our mobile phones. He was a mathematician, geographer, poet, astronomer, and music theorist. Eratosthenes (and later Ptolemy) thus provided a geometry within which the position of humanity could be assessed.

These are all examples of a radical predisciplinarity that reveals entwined goals and desires: these exceed the expertise of what now constitute scientific and humanistic approaches as to what it is to live on and with the Earth. As such, they present an alternative form of knowledge-making to that which now dominates academia, one that can move laterally across conceptualizations and mediums, as well as dig down into specific problematics. What they also help bring into clarity, however, is a tension that haunts the birth of the interdisciplinary or transdisciplinary project of the geohumanities. On the one hand, we can see a desire to root the project in a critical conception of the roles of space and place in what it is to be more than human in the twenty-first century; on the other hand, we can discern a desire to harness the analytical potential of the digital geospatial web. This tension is itself a fecund one that needs be acknowledged but does not require a “bridging.” To congeal geohumanities as a distinct field of inquiry – a field with common concerns, goals, and an all-purpose methodological toolkit – would be to bring a halt to this lateral movement. As geohumanities continues to unfold, it is its transdisciplinarity that emerges as key.

The other “geo” is the “geo” of earth as soil/ground – the “geo” in a term such as “geophagy,” which is the act of eating soil. To be sure, over the course of the nineteenth and twentieth centuries the humanities, including human geography, addressed themselves in large part to the rapid course of urbanization and “city life.” A direct contemplation of the physical world was left to the physical scientists such as geologists and geochemists. This distinction has also, however, become blurred, particularly with the rise of the term “Anthropocene” and its enthusiastic take-up across academia. Environmental concerns have become increasingly visible across the humanities as evidenced by environmental history, eco-poetics and eco-philosophy – all dealing with forms of meaning-making in relation to the Earth. There has even been a “geological turn” as humanistic scholars have turned toward the very matter of earthly inhabitation (Yusoff 2015). A robust geohumanities project, then, will take this notion of “geo” as soil/earth seriously, and become a venue for transdisciplinary explorations that incorporate the work and aspirations of physical scientists.

SEE ALSO: Anthropocene and planetary boundaries; Art; Critical GIScience; Cultural geography; Feminist geography; Geocoding; Geographic information science; Geographic information system; Geophilosophy; Humanistic geography; Interdisciplinarity and geography; Place; Space; Spatiality

References


Geolocation services

Chin-Te Jung
San Diego State University, USA

Geolocation is the identification of the geographic location of an object. Every object that exists in the physical world, even Internet-connected computer devices (e.g., computers, notebooks, and mobile phones) has a geolocation to describe its position. The most common and precise way to represent the geolocation of an object is by using geographic coordinates (i.e., a set of latitude and longitude coordinates). For example, San Diego, California, is located at 32.73°N and 117.17°E. Several locating engines and technologies can be used to retrieve geolocations in coordinates, such as the global position system (GPS), assisted GPS (A-GPS), or the Wi-Fi positioning system. More details about these geolocation technologies can be found in Djuknic and Richton (2001).

However, a physical address, ZIP code, or place name is often used as a meaningful way to express geolocation. For instance, the geolocation of San Diego State University (SDSU) is 5500 Campanile Drive, San Diego, CA, 92182 (a physical address); or a basketball game will be held at SDSU (a place name). In addition, for network-connected computer devices, an Internet protocol (IP) or a media access control (MAC) address is used to identify the geolocation of the devices on the Internet. These alternative ways to describe a geolocation can eventually be converted to a set of geographic coordinates by a mapping database or a web service. For example, someone can look up an IP address on a WHOIS (http://tools.whois.net/whoisbyip/) service to retrieve the registrant’s physical address and its coordinates can be mapped using MapQuest’s open geocoding web service (http://developer.mapquest.com).

This entry mainly focuses on geolocation web services (geolocation services) which (i) can reveal the geolocation information of an object using geolocation technologies, such as GPS, A-GPS, or the Wi-Fi locating system, and (ii) can geocode an address, a ZIP code, a place name, or an IP address into geographic coordinates. A geolocation can be simply retrieved by sending a request to a geolocation service in a web browser (e.g., Internet Explorer), rather than by installing any professional GIS software. Additionally, a geolocation service can be a mashup (i.e., integrated with other web services, e.g., searching points of interest) to provide a location-based service (LBS), such as finding points of interest near a user’s location or automatically pushing relevant location-based information to a user’s mobile device (Schiller and Voisard 2004).

This entry lists several common geolocation services, as shown in Table 1. For a more complete list of geolocation services refer to Santos (2012). Geolocation services are classified by type (i.e., revealing geolocation information or geocoding), input (i.e., what geographic data can be geocoded), and output (i.e., geographic coordinates) with their data type format, such as Extensible Markup Language (XML) or JavaScript Object Notation (JSON).

The development of GIS has been greatly influenced by information technology (IT) and the Internet, which stimulated the evolution of GIS from desktop-based systems to web-based GIS services. The geolocation service belongs
<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Inputs</th>
<th>Outputs (data type)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Maps: Geocoder&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Geocoding</td>
<td>Address/place name</td>
<td>Latitude and longitude (JSON)</td>
<td>Google Maps provides a JavaScript application protocol interface (API) to access various basemaps (e.g., street map, satellite images, terrain map, and traffic map) and geospatial functions (e.g., geocoding, best routing, and terrain profile functions). Geocoder is its geocoding service which can geocode addresses and place names into geographic coordinates (latitude and longitude) in JSON format.</td>
</tr>
<tr>
<td>Bing Maps: find a location by address&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Geocoding</td>
<td>Address/place name</td>
<td>Latitude and longitude (XML/JSON)</td>
<td>Bing Maps supplies a JavaScript API to access basemaps (e.g., street map and satellite images) and geospatial functions (e.g., geocoding and best routing). Bing Maps provides a geocoding service, called “Find a Location by Address,” to geocode addresses and place names into latitudes and longitudes in XML or JSON format.</td>
</tr>
<tr>
<td>Yahoo BOSS Placefinder&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Geocoding</td>
<td>Address/place name</td>
<td>Latitude and longitude (XML/JSON)</td>
<td>Yahoo BOSS Placefinder is a geocoding web service which can geocode addresses and place names into latitudes and longitudes in XML or JSON format.</td>
</tr>
<tr>
<td>MapQuest open geocoding API&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Geocoding</td>
<td>Address/place name</td>
<td>Latitude and longitude (XML/JSON)</td>
<td>MapQuest open geocoding API is an open (i.e., free) geocoding web service which matches addresses and place names with geographic coordinates (i.e., latitudes and longitudes) in XML or JSON format.</td>
</tr>
<tr>
<td>GeoNames&lt;sup&gt;e&lt;/sup&gt;</td>
<td>Geocoding</td>
<td>Place name/ZIP code</td>
<td>Latitude and longitude with additional information (XML/JSON)</td>
<td>GeoNames supplies multiple geocoding services, which mainly geocode place names and ZIP codes into latitudes and longitudes in XML or JSON format. Additional information, such as country and population, is also attached in the outputs.</td>
</tr>
</tbody>
</table>

<sup>a</sup> See also: https://developers.google.com/maps/documentation/geocoding/

<sup>b</sup> See also: https://docs.microsoft.com/en-us/previous-versions/dt.105/bb527617(v=vs.85)

<sup>c</sup> See also: https://developer.yahoo.com/boss/placefinder/

<sup>d</sup> See also: https://developer.mapquest.com/documentation/geocoding-api/

<sup>e</sup> See also: https://www.geonames.org/developers.html
<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Inputs</th>
<th>Outputs (data type)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>W3C Geolocation APIf</td>
<td>Revealing geolocation using technologies</td>
<td>IP address/ Wi-Fi/GSM/ CDMA cell IDs</td>
<td>Latitude and longitude</td>
<td>The World Wide Web consortium (W3C) Geolocation API defines a standard interface to retrieve only geolocation information (i.e., latitude and longitude) associated with the devices hosting the implementation (Popescu 2013). The geolocation information source can be retrieved from GPS or inferred from network signals, such as IP address, Wi-Fi, and GSM (Global System for Mobiles)/CDMA (Code Division Multiple Access) cell IDs (Popescu 2013). For example, if a web browser supports the W3C geolocation API (as Google Chrome does), a developer can use the geolocation API to retrieve a client’s geolocation, as long as the client is willing to share. The geolocation source could be the IP address if the client uses a network-connected computer or from GPS if the client uses mobile devices.</td>
</tr>
<tr>
<td>MaxMind GeoIPg</td>
<td>Geocoding</td>
<td>IP</td>
<td>Latitude and longitude/ZIP code (JSON)</td>
<td>MaxMind GeoIP is a web service which first searches for a registrant’s information using their IP address and then geocodes their address into latitude and longitude in JSON format.</td>
</tr>
<tr>
<td>FreeGeoIPh</td>
<td>Geocoding</td>
<td>IP</td>
<td>ZIP code/place name/latitude and longitude (XML/JSON)</td>
<td>FreeGeoIP is a public geocoding service which searches for the geolocation of an IP address. The geolocation includes ZIP codes, place names, and latitudes and longitudes in XML or JSON format.</td>
</tr>
</tbody>
</table>

*a* [https://developers.google.com/maps/documentation/javascript/reference#Geocoder](https://developers.google.com/maps/documentation/javascript/reference#Geocoder)  
*c* [https://developer.yahoo.com/boss/geo/](https://developer.yahoo.com/boss/geo/)  
*e* [http://www.geonames.org/export/web-services.html](http://www.geonames.org/export/web-services.html)  
*f* [http://www.w3.org/TR/geolocation-API/](http://www.w3.org/TR/geolocation-API/)  
*h* [http://freegeoip.net/](http://freegeoip.net/)
GEOLOCATION SERVICES

to one of the GIS services that allow groups of GIS users and organizations to access clients’ geographic coordinates more easily. Additionally, the geolocation service can be integrated with various standards of GIS services defined by the Open Geospatial Consortium (OGC), such as a web map service (WMS) (OGC 2006), a web feature service (WFS) (OGC 2005), or a web processing service (WPS) (OGC 2007).

By integrating GIS services, geolocation services lower the difficulties of learning and manipulating GIS software, and increase the usability of GIS with various domains and purposes (e.g., to mashup Google Maps with real estate data). Moreover, they have emerged as the service-oriented approach that allows users to access, assemble, and process geospatial content and functionality distributed across a network via a standard Internet browser (Tao 2001).

SEE ALSO: Geographic information system; Open Geospatial Consortium standards; Web-mapping services

References


Geomancy, or fengshui (meaning “wind and water” in Chinese), has played a key role in the site selection and landscape design of cities and rural settlements, temples, monuments, graves and houses in East Asia. It has been a difficult and confusing subject to grasp for those in the West, as there is no equivalent concept in Europe to geomancy in China. A contemporary Western scholar declared, “if there is a subject which should have captivated Western Sinologists, it is Chinese geomancy” (Lemoine 1974, 1), and nineteenth-century Western scholars labeled fengshui the rudiments of natural science in China (Eitel 1873, title p.) or a quasi-religious and pseudoscientific system (de Groot 1897, 935). Geomancy cannot be simply classified as a superstition, religion, or science, but combines elements of all three into a “unique and highly systemised ancient Chinese art of selecting auspicious sites and arranging harmonious structures on them by evaluating the surrounding landscape and cosmological directions” (Yoon 2006, 4). The idea of fengshui is also known as dili (the principles of land), which also means “geography,” or Kanyu (the carriage loaded with a myriad of things, meaning “heaven and earth”). Historically, geomancy influenced East Asian culture so significantly that a proper understanding of East Asian landscapes requires some knowledge of the art.

Traditionally in China and Korea, people were more interested in locating auspicious gravesites than house sites, which led some scholars to argue that geomancy originated from the art of finding auspicious gravesites. A careful analysis of the fact that geomantic principles regarding gravesites are essentially the same as those of house sites contradicts this earlier view and suggests that ancient geomancy seemingly first developed as an art for choosing ideal house sites, especially for the preparation of a comfortable cave dwelling in Loess Plateau, North China (Yoon 2006, 15–32). Geomancy later developed into two schools from the time of the Han dynasty (206 BCE–220 CE) through to the Song dynasty (960–1279): (i) The Jangxi School that emphasized the methods of landform observation in site selection, and (ii) the Fujian School that emphasized the methods of divining cosmological directions by compass. The ancient Chinese art of geomancy diffused into Korea, Japan, and other neighboring regions along with an early wave of Chinese cultural diffusion and influenced East Asian cartography through its distinctive method of portraying the relief of landforms in maps for documenting auspicious places.

The principal criteria for examining an auspicious site are the conditions of the landform, availability of water, and whether the site faces a sunny auspicious direction. An ideal landform condition is a location surrounded by hills on three sides, like a horseshoe or an armchair shape with an open front (Yoon 2006, 76). Thus, a basin is often accepted as an auspicious site and its surrounding hills are named; the one behind, the main mountain (or black warrior); on the left, the azure dragon; and on the right, the white tiger. If a place with these background hills also faced southward, it would enjoy maximum sunlight, while being protected from the cold wind blowing from the northwest during winter.
Thus, a southward direction is considered auspicious and its quality is examined with a geomantic compass that reflects the complicated Chinese cosmology based on the combination of the Yin-Yang and the Five Element Theory, the 12 zodiac animals, and the Eight Trigrams from the *I-Ching* (Book of Changes). To be an auspicious place, a site with an appropriate landform and an auspicious direction also requires a watercourse in front of it. If such a site (basin) is formed on a big scale, it can become a site for a major city, while smaller scale sites are suitable places for smaller towns or villages.

Geomancers believe that a place that meets such geomantic requirements is auspicious and blesses the people who occupy it, because such a site can conserve vital energy (shengqi) which is mysteriously developed from Yin-Yang energy. An ancient classic on geomancy elaborates the function and mechanism of the energy cycle: the energy of Yin-Yang belches and becomes wind; ascends and becomes cloud; descends and becomes rain; flows underground and becomes vital energy (Guo 1875, 1). This is a crude but ancient Chinese concept of the environmental energy cycle. This vital energy mysteriously ingenerates and supports a myriad of things and flows only through underground veins dispatched like blood vessels. It is accumulated in an auspicious site and blesses the occupier of the place. That is why the believers of geomancy spent much energy and resources in finding and occupying an auspicious site. This attitude toward geomancy helps explain why the rulers of China, Japan, and Korea attempted to shift capitals so eagerly.

Traditional East Asian capital cities such as Nanjing in China, Seoul in Korea, and Kyoto in Japan are all located in a basin with protective hills in three directions, with the southward face open and with watercourses nearby. Beijing is exceptional as it is located on a plain and the required protective hills are somewhat distant from the city. That is why an artificial hill of about 45 m high was created behind the palace as a gesture of making up the geomantic deficiencies of the Chinese capital city.

Like major cities, many smaller towns and villages in East Asia also have geomantically important characteristics, being at the end of foothills, with an open front and background hills. Numerous Buddhist temples and other religious shrines in China, Japan, and Korea are also located in geomantically considered places.

Professional geomancers in East Asia chose auspicious sites for their clients who paid fees, and were also location (place) diviners as well as traditional folk geomorphologists and landscape consultants. In the practice of geomancy the following three images of the environment (attitudes toward the landscape or nature) are important:

First, the environment is treated as a magical being that can bless people auspiciously or affect them inauspiciously in a magical manner (Yoon 1980). A place that accumulates vital energy can make the occupiers happy and prosperous, while a place with no such energy can cause the people to suffer misfortunes.

Second, the environment is perceived as a personified being (Yoon 1980). The auspicious spot and its surrounding environment are treated as a functioning system which brings in vital energy, stores it, and makes it available to the occupier. For example, a local landscape system is identified as being akin to a sleeping cow, a sailing boat, and so forth. An entire local landscape or a portion of it can be named after and treated as any personified beings, depending on how they were perceived by a geomancer. A landscape which was beautiful and peaceful looking was considered auspicious, while ugly and dangerous ones were considered inauspicious.
Third, the environment is recognized as a vulnerable entity that can be easily damaged or repaired by human actions (Yoon 1980). Vital energy in an auspicious spot is vulnerable, because it is drawn from an unstable phase of a continuously transforming cycle of the Yin-Yang energy and can always be lost from an auspicious site if its landscape harmony is disturbed. A mining shaft drilled by a miner or the placement of a highway through the landscape of a sleeping cow can ruin the place, but the damaged landscape can be restored as an auspicious site by removing or moderating human disturbances.

Original Chinese geomancy was concerned with the relationships between a human-made structure (a house) in an auspicious spot and its surrounding landscape, especially its landforms. However, as time passed, and when the art of geomancy was applied on flat land with no protective hills on the right (white tiger) or on the left (azure dragon), a stream or a road was justified to substitute the necessary hills. Geomancy has been adapted into modern built-up urban areas. In urban situations, the characteristics of the neighboring house, road pattern, and geomantic directions are important in divining the auspiciousness of a house. Geomantic principles are applied now in the interior design of buildings and are even used in arranging office furniture. However, these later applications do not appear in ancient Chinese geomantic texts. Numerous geomantic textbooks written during different periods of Chinese history exist today, and the art of geomancy, or fengshui, is still practiced for site selection, city planning, environmental management, architectural theory, and interior design in China, Korea, Japan, and other East Asian countries.

**SEE ALSO:** Cross-cultural research; Cultural diffusion; Cultural heritage; Culture; Environmental management; Historical geography; Historical settlement: Asia; Indigenous knowledge

**References**


Eitel, E.J. 1873. *Feng Shui: Or the Rudiments of Natural Science in China*. Hong Kong: Lane, Crawford & Co.


A hazard, generally speaking, is an element of the environment that is potentially harmful to humans in the form of loss of life, property damage, and socioeconomic disruption. Hazards can be “natural” in origin and only become a hazard because of the juxtaposition of geomorphic forms, processes, and materials with human activities. Hazards can also be caused or exacerbated to some degree by human activities. When an event occurs that actually does result in impacts to humans, the hazard becomes a disaster. The degree to which humans contribute to a hazard/disaster is often difficult to discern. A new field of inquiry termed “attribute science” has developed to focus on this challenge.

Geomorphic hazards/disasters are those that originate from exogenic processes – processes that originate at or near the Earth’s surface from the action of water, wind, ice, and gravity. Geomorphic hazards are distinguished from geological hazards that originate deep within the Earth from the endogenic processes of volcanism or earthquakes. The distinction is not always clear between geomorphic hazards and other main categories of hazards: meteorological, hydrological, biological, technological, and social. For instance, floods are usually classified as hydrological hazards but river channel changes are considered a geomorphic hazard.

Geomorphic hazards considered here (Table 1) include:

- expansive soils,
- soil erosion, including dust storms,
- slope failures,
- ground subsidence and karst,
- periglacial,
- river channel changes,
- glaciers,
- coastal erosion,
- climate change.

Burton, Kates, and White (1978) listed the critical information needs regarding hazards:

- magnitude: high to low,
- frequency: often to rare,
- duration: long to short,
- areal extent: widespread to limited,
- speed of onset: rapid to slow,
- spatial dispersion: diffuse to concentrated,
- temporal interval: regular (e.g., annual, seasonal, diurnal) to random.

Expansive soils cause billions of dollars of damage each year to homes in the United States alone (Alexander 1993). The swelling is caused by chemical attraction of water to plates in certain clay minerals. Pure montmorillonite (or smectite) clay, for example, can expand 15 times when water is added. Soils with high gypsum
### GEOMORPHIC HAZARDS

**Table 1** General characteristics of geomorphic hazards.

<table>
<thead>
<tr>
<th>Type of hazard/disaster</th>
<th>Frequency</th>
<th>Duration</th>
<th>Speed of onset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expansive soils</td>
<td>Seasonal/irregular</td>
<td>Months to years</td>
<td>Months to years</td>
</tr>
<tr>
<td>Soil erosion</td>
<td>Progressive</td>
<td>Hours to millennia</td>
<td>Minutes to years</td>
</tr>
<tr>
<td>Wind/dust storms</td>
<td>Seasonal</td>
<td>Hours</td>
<td>Hours</td>
</tr>
<tr>
<td>Slope failures</td>
<td>Seasonal/irregular</td>
<td>Seconds to decades</td>
<td>Seconds to years</td>
</tr>
<tr>
<td>Subsidence</td>
<td>Sudden/progressive</td>
<td>Minutes to decades</td>
<td>Seconds to years</td>
</tr>
<tr>
<td>Periglacial</td>
<td>Seasonal/progressive</td>
<td>Months to millennia</td>
<td>Hours</td>
</tr>
<tr>
<td>River channel change</td>
<td>Seasonal/irregular</td>
<td>Months to centuries</td>
<td>Hours to years</td>
</tr>
<tr>
<td><strong>Glaciers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calving ice</td>
<td>Seasonal</td>
<td>Months to decades</td>
<td>Seasonal</td>
</tr>
<tr>
<td>Ice advances</td>
<td>Seasonal/progressive</td>
<td>Decades to millennia</td>
<td>Decades</td>
</tr>
<tr>
<td>Surging glaciers</td>
<td>Seasonal to years</td>
<td>Years</td>
<td>Months</td>
</tr>
<tr>
<td>Ice avalanches</td>
<td>Seasonal/irregular</td>
<td>Seconds to years</td>
<td>Seconds to years</td>
</tr>
<tr>
<td>GLOFs</td>
<td>Seasonal/irregular</td>
<td>Hours to days</td>
<td>Seconds to years</td>
</tr>
<tr>
<td>Coastal erosion</td>
<td>Seasonal to years</td>
<td>Days/seasonal</td>
<td>Hours to decades</td>
</tr>
</tbody>
</table>


Content can also experience shrink-and-swell to the point where structural damage occurs. Most soils have a portion of clay that would cause 25–50% expansion. Damage to structures is likely if the soil expands more than 3%. In the US Department of Agriculture (USDA) *Soil Taxonomy*, expansive soils fall into the order of Vertisols, which cover about 2% of the ice-free land surface on Earth. It is critical to consult soil surveys when trying to avoid expansive soils or to plan for mitigation.

Soil erosion has been increasing steadily since the advent of cropland agriculture about 3500 years ago, primarily from fluvial processes of rainsplash, sheetwash, rills, gullies, and piping. Wind erosion is important in flat terrain, and slope failures become important in mountainous terrain. Perhaps as much as one-half of the topsoil on Earth has been lost in the last 150 years through erosion or degraded soils quality. Soil erosion becomes a hazard when rates of erosion exceed rates of new soil formation, progressively destroying soils structure, nutrient content, and fertility. Rates of soil erosion are naturally high in semiarid climates, where vegetation cover is sporadic and storm-period rainfall intensity is high. Land-use/land-cover change (from deforestation, agriculture, overgrazing, climate change, desertification, wildfires, etc.) can greatly accelerate soil erosion by water and wind. Soil erosion causes millions of hectares of land to be removed from cultivation each year. Soil erosion differentially affects poor countries such as Ethiopia and El Salvador, where arable land is limited, and farmers lack modern farming technology and training in soil conservation methods. Trimble’s (2012) award-winning narrative about historical agriculture and soil erosion in the upper Mississippi Valley hill country documented how agriculture initially triggered landscape destruction, impoverishment and instability. However, farmers eventually adapted...
their land use and settlement practices to recover and enrich the landscape.

Dust storms pose a hazard to health and property by blocking roads and causing highway accidents, filling irrigation ditches and canals, burying crops and ruining machinery, dispersing weeds and insects, depleting soils, and sandblasting plant roots. Dust storms play a role in human health, animal health, climate change, and ocean sedimentation. Dust storms can reduce visibility to less than 1 km and can reach altitudes greater than 10 km. They can be traced for thousands of kilometers on remotely sensed imagery. Dust from World War II tank battlegrounds in North Africa has been found in the Caribbean. The magnitude, frequency, and duration of wind exert key controls on dust production, especially when wind acts on a surface with an abundance of silt-sized sediment that is dry and low in organic material. The generation of dust is further enhanced on a surface with low vegetation density, low litter cover, low surface roughness, and a long fetch. In a geospatial analysis of dust storm potential in the Chihuahuan Desert, Marston (1986) found these variables combined to create highest dust storm potential in the broad basins of northern Mexico, where irrigated cropland had been abandoned or is seasonal, and in deteriorated rangeland and regions of desert shrub vegetation (Figure 1).

Slope failures cause an annual average of 25 deaths and losses of US$ 1 billion in the United States. One of the worst natural disasters occurred in Hsian, China, because of widespread slope failures triggered by an earthquake in 1556. A 1963 landslide of 238 million cubic meters into the reservoir behind Vaiont Dam in Italy created a wave 90 m high over the dam. Amazingly, the dam survived but 2600 people lost their lives downstream. The landslide was composed of fractured limestone blocks that slid on beds of clay and soft marl. The landslide had been monitored for three years prior to the disaster; it was moving between 1 and 30 cm per week. Rising groundwater and heavy rains caused a threshold to be exceeded for the massive movement.

Slope failures are pervasive in the Nepal Himalaya but linking cause and effect has been made difficult by the multiplicity effect in geomorphology—any one variable can explain some of the variation but the interactions between variables remain poorly understood. Marston, Miller, and Devkota (1998) showed that forest cover did not affect the occurrence of slope failures, but position above and below the main central thrust was shown to be important, as was slope aspect. Dahal and Hasegawa (2008) demonstrated that a threshold of rainfall intensity and duration initiated landslides. In April 2015, a magnitude 7.8 earthquake in Gorkha, Nepal, caused thousands of fatalities, destroyed entire villages, and triggered thousands of landslides, some of which blocked rivers to create yet another hazard (landslide dam outburst floods) downstream (Collins and Jibson 2015). Variables

![Figure 1 Dust storm from SE winds entraining dust from a broad basin in northern Chihuahua. Note that the borders of Texas, New Mexico, and Mexico have been added. Source: Aqua MODIS, 2003. Reproduced from NASA. Image dated April 4, 2003.](image)
that control slope failures in the Himalayas change from one time to another and contrast with the controlling variables in other mountainous areas of the world.

The direct hazard posed by slope failures on humans is illustrated by the work of Butler (2013). Slope failures (rockfalls, debris flows, rockslides, snow avalanche tracks) were mapped in relation to hiking/skiing trails and campgrounds that are heavily used each year by backcountry visitors to the national park (Figure 2). The National Park Service could use this information to educate backcountry tourists and to modify their location of trails and campgrounds.

As defined by Alexander (1993, 276), “subsidence is a type of mass movement that involves the downward displacement of surface material caused by natural or artificial removal of underlying support (collapse) or by compression of soils (consolidation).” The extraction of oil, gas, and groundwater has been demonstrated to cause subsidence in Long Beach, CA, in Venice, Italy, and in Niigata, Japan, as just three examples. The construction of the tower in Pisa, Italy, on soils of contrasting compressibility caused the tower to lean over the course of centuries. Soils on deltas consolidate over time as organic matter is oxidized and soils are compressed under their own weight. This enhances flooding hazards, in the Mississippi River delta region, for example. An interesting example is provided by the Dead Sea, which has been affected by river inflow being diverted for irrigation agriculture and other purposes. The water level has dropped over 7 m in the last decade and the lake shore has receded. Fresh groundwater has replaced the saline groundwater, dissolving the buried salt deposits and causing the ground to collapse.

Sudden collapse of the ground, forming sinkholes, is more common in karst terrain. The Florida peninsula is a structural arch of fractured limestone. Groundwater has dissolved the limestone, particularly where fractures intersect, to create caverns. When groundwater drops during extended droughts, the support provided by the water in caverns is lost and sinkholes appear. In the rapidly urbanizing Rio Grande Valley downstream from El Paso, TX, former agricultural fields are covered by impermeable asphalt and concrete. Gophers abandon the urban area for the few remaining fields that are surrounded by pavement. When farmers dump a meter of water on their field to leach salts from the soils, the gopher tunnels experience erosion by piping, sometimes leading to formation of a sinkhole in alluvium and ruining the farm.

Periglacial environments are dominated by the action of ground ice (ephemeral, seasonal, permanent) and mass movement. The active zone is affected by freeze–thaw cycles over diurnal and seasonal intervals of time. Permafrost underlies about 26% of the Earth’s land surface. Freeze–thaw action displaces larger particles more than smaller sediment, resulting in an upfreezing of stones and sorting the mix of sizes. Mass movement occurs by the process of frost creep on slopes, whereby repeated freeze–thaw cycles lead to a slow movement of material downslope. Solifluction involves the slow downslope movement of saturated soils during the summer thaw. An intriguing variety of periglacial landforms is created by these processes, including elongated thaw lakes, ice wedges, pingos, patterned ground, blockfields, and rock glaciers. Hazards are created when human activities and settlements encroach on periglacial landscapes. During thaw periods, the bearing strength of soils is lost, which leads to differential settling and frost heaving, and structural damage to buildings, roads, airfields, railroads, and utilities. In the cases where felsenmeer and rock glaciers are situated on a cliff, the melting of interstitial ice can lead to rockfalls, which
GEOMORPHIC HAZARDS

Figure 2  Map of existing slope failures (rockfalls, rockslides, debris flows, snow avalanches) for Death Canyon in the Grand Teton National Park compared to locations of campgrounds and popular trails for hiking and cross-country skiing. Source: Butler (2013). Reproduced by permission of William Butler.

become a hazard for downslope human activities and settlements.

River channels change location by meander cutoffs, meander migration, and by avulsion. These changes can occur in the absence of human activities but can be exacerbated by humans through direct manipulation of the channel or through land-use/land-cover changes upstream. When rivers change course, land on the floodplain that would otherwise be available for farming and other human activities can be lost. Stream channel change is driven by the frequency and magnitude of peak flows on the one hand, and the resistance of stream banks (controlled by alluvial materials, large woody debris, and streamside vegetation) on the other hand. A landowner situated on the outside (cut bank) of a meander bend stands to lose a significant amount of land if erosion progresses as is normal. Often a river marks a property boundary or a political border. According to the law in many jurisdictions, if the river moves
by progressive meander migration, the border moves with it, causing the landowner on one side to gain land by accretion while the landowner on the other side loses land. The landowner who gained land may further rub salt into the wound of his or her neighbor by acquiring a permit to mine the river for sand and gravel. In the case where a river changes course by a sudden avulsion, the property line does not move with it, creating another set of problems and disputes.

Glaciers can create hazards for humans in at least five ways. First, glaciers and ice sheets that extend to the coast will experience calving of ice at the terminus. This generates icebergs, which then become a hazard for shipping. The Greenland and Antarctic ice sheets are notorious for this hazard, as are coastal glaciers along the southern coast of Alaska. Second, ice advances, although slow, can overrun villages and farms. This was known to have occurred in the Alps, Norway, and Greenland during the Little Ice Age. Third, surging glaciers can block water bodies, leading to potential outburst floods. This has occurred multiple times with the Hubbard Glacier in Yakatat Bay, AK, blocking the entrance to Russell Fjord. Fourth, ice avalanches from tall peaks can create a rock-ice avalanche and demolish towns that are located down valley. The collapse of the ice summit of Mount Huascaran in Peru in 1970 resulted in the deaths of 25,000 people. A smaller scale event in 1962 claimed 4000 lives. Fifth, glacier changes can lead to a glacial lake outburst flood (GLOF), or jokulhlaup. When valley glaciers retreat, they leave behind a moraine, which may dam water between the retreating glacier and the moraine. If the moraine experiences piping, it may collapse. If a rock-ice avalanche lands in the lake, it can create a wave that overtops the moraine dam, causing it to collapse. If an advancing valley glacier cuts off a tributary, it can create a lake behind the ice. The ice dam may collapse or a subglacial tunnel may enlarge to suddenly release the water, endangering life and property downstream. GLOFs are common in glaciated mountains and cases of imminent danger have been identified in the Himalayas, central Asia, the Andes, and the Alps.

Coastal hazards include erosion and sedimentation by waves and longshore currents. Coastlines are dramatically affected by tropical cyclones, hurricanes, and tidal flooding, but these are usually considered meteorological hazards. Tsunamis are triggered by earthquakes or large slope failures into a water body, so are not considered here. Alexander (1993) stated that five factors influence coastal erosion: exposure of the rocks and sediments to waves and currents; the supply of sediments (beach starvation or nourishment); the topography of the coast and neighboring continental shelf; the tidal range and intensity of currents; and the climate of the coastal area. Once again, it is worth pointing out that natural processes along the coast become hazards because of the intersection with human activities. The hazards depend on the geomorphic type of coastline.

Sandy beaches experience seasonal transfer of sand, offshore in winter, onshore in summer. Thus, in winter, beachfront homes lose a buffer against attack by storm-period waves, especially during high tides. Littoral movement of sand, another important process affecting the beach sand budget, is responsible for a variety of coastal landforms and can close the mouths to rivers, estuaries, and harbors if left unchecked. Sand spits and barrier islands are commonly breached by storm-period events, leading to a call for artificial beach sand nourishment. On the Arctic coast, the longer fetch for wind over ice-free ocean already appears to be creating higher waves and greater coastal erosion, especially where permafrost is melting and sea level rising.
Cliff coastlines experience direct wave attack throughout the year, with progressive cliff erosion and slope failures as a hazard. Wave-cut abrasion platforms, if present, serve to dissipate wave energy. The rate of cliff retreat needs to be studied by coastal geomorphologists, so coastal planning authorities can establish appropriate setback regulations. Along the central Oregon coast, cliff retreat from 1935 to 1975 was measured at 23 cm per year in marine sandstones/shales versus 5.3 cm per year in basalt over the same period.

Earlier mention was made about consolidation of delta deposits. Deltas will be subject to erosion if the rivers that originally deposited the sediment are not allowed to deposit new material on top of the consolidating sediment. This is the situation being faced along the Gulf coast shores of Louisiana, where about 4900 km² of coastal land was lost between 1970 and 2000. Some of this loss is attributed to saline water intruding into freshwater wetlands, killing the freshwater marsh grasses that hold the sediment in place.

Coral reefs protect the shore from the direct impact of waves, but coral are threatened in several ways. They are being attacked by a crown-of-thorns starfish, the population of which is exploding because their predators have been overfished. The coral polyps that create reefs are supported by algae. The coral die within weeks if the algae die, and algae are dying from warmer ocean temperatures and sedimentation.

Climate change will have an impact on all of the geomorphic hazards discussed here by affecting one or more of the seven dimensions of hazards defined by Burton, Kates, and White (1978) listed previously. However, global circulation models must continue to be refined to the point where predicting change is possible for geomorphic hazards. Geospatial analysis provides many tools for understanding and visualization of the results. It is clear that the geomorphic hazards discussed here are sensitive to climate, so the degree to which these controlling factors change deserves further study. Critical research is underway to identify the key indicator variables in geomorphic systems and their response to external driving forces.

SEE ALSO: Anthropogeomorphology; Applied geomorphology; Coastal depositional processes and landforms; Coastal erosion processes and landforms; Eolian erosional processes and landforms; Fluvial depositional processes and landforms; Fluvial erosional processes and landforms; Geomorphic thresholds; Geomorphological mapping and geospatial technology; Glacial depositional processes and landforms; Glacial erosional processes and landforms; Glacier lake outburst floods; Global climate change; Karst processes and landforms; Land-use/cover change and climate; Mass movement processes and landforms; Natural hazards and disasters; Periglacial processes and landforms; Soil erosion and conservation

References

GEOMORPHIC HAZARDS


tion. Cambridge: Cambridge University Press.


Further reading

Alcántara-Ayala, Irasema, and Andrew Goudie, eds. 2010. Geomorphological Hazards and Disaster Preven-
Geomorphology deals with highly complex assemblages of topographic elements or “geomorphic units” that exist in the real world. The units include different types of slopes in mountains and hilly lands (e.g., rocky cliffs, taluses, and other types of valley-side slopes), depositional and erosional landforms in piedmonts and floors of intermontane basins (e.g., river terraces, alluvial fans, and floodplains), and those landforms in coastal areas (e.g., deltas, sand dunes, and rocky benches). An assemblage of such units can be regarded as a single mega-unit, typically a drainage basin, which is sometimes called a “fluvial system” (Schumm 1977). Like this usage of the term “system,” an assemblage of geomorphic units is often referred to as a “geomorphic system.”

However, “geomorphic system” means much more than a mere assemblage of geomorphic units. It is also related to “general system(s) theory” (Bertalanffy 1951), which describes the structure of a system including the hierarchy and cascading of units and subunits; the behavior of a system, particularly interactions between units and subunits; and the temporal evolution of a system with responses to change in external conditions. Chorley (1962) pointed to the close connection between general system theory and geomorphology, and this way of thinking exerted a strong influence on subsequent geomorphological studies. Although recent geomorphological publications seem to put less emphasis on general system theory, this may not mean that the application of the theory has become obsolete. In a sense, geomorphologists well understand the ideas from general system theory and their usefulness as background knowledge for research, and they do not have to re-emphasize them. At the same time, the classic system approach in geomorphology has received some criticism in recent years and has been re-evaluated.

Development of the system approach in geomorphology

Alongside Bertalanffy’s general system theory, Chorley (1962) stated that some geomorphological papers published in the 1950s had already applied general system theory. These studies were strongly influenced by the morphometric and statistical analyses of drainage basins and stream nets established by Horton (1945). Since then geomorphologists have been considering drainage basins as geomorphological units with structural hierarchy, and their quantitative characteristics have sometimes been interpreted from the viewpoint of general system theory, especially whether the form of basins reflects or has adjusted to external environmental conditions.

Chorley (1962) emphasized the difference between a closed system and an open system. The former changes without a supply of energy or mass from outside, whereas the latter receives them. Although a system of either type tends to attain a time-independent steady state or
GEOMORPHIC SYSTEMS

equilibrium, the ultimate equilibrium condition of a closed system can be predicted only from the initial condition of the system. By contrast, that of an open system is the result of dynamic interaction between the system and the surrounding environment. The latter is more realistic in geomorphology because landforms are influenced by various external factors such as climate, tectonics, and lithology. Strahler (1952) put forward similar ideas concerning dynamic relationships between forms and processes associated with transformation of mass and energy. Equilibrium in geomorphic systems was interpreted using principles of thermodynamics (Chorley 1962), as in preceding applications of general system theory in other scientific fields. Chorley (1962) correlated closed systems with some classic geomorphological concepts or approaches, such as the cyclic model of erosion by Davis and the slope retreat model by Penck, but correlated open systems with newer geomorphological ideas, with more attention to geomorphic processes and interactions with the surrounding environment. It should be noted that G.K. Gilbert already had an idea comparable to open systems that evolve toward equilibrium in relation to his innovative approach of process geomorphology (Huggett 2007).

After the recognition of open systems in geomorphology, the adjustment of systems to external variables and resultant equilibrium or steady-state conditions received increased attention. As noted, the morphometry of drainage basins was interpreted from this viewpoint. Another example discussed early on was the adjustment of channel cross-section geometry and river longitudinal profiles to environmental conditions such as flow discharge (Langbein and Leopold 1964), which followed not only general system theory but also a concept of “grade” proposed by Mackin (1948). Other more complex landform units, such as ridges and hollows on hillslopes, were also considered to be interacted systematically and to change toward equilibrium (Hack and Goodlett 1960). Adjustment of geomorphic systems to tectonic conditions in terms of the balance between uplift rate and erosion rate, and resultant landforms in dynamic equilibrium, were also discussed (Howard 1965), providing a basis for subsequent investigations.

At the same time, complex relationships between landforms and the external environment were indicated. Geomorphic systems are often affected by episodic large events such as volcanic eruptions, glaciation, and megafloods which lead to a disturbance of system adjustment, sometimes in a catastrophic way. Also, responses of geomorphic systems to external changes tend to be nonlinear, including lag time, relaxation time, and thresholds. Schumm (1973) focused on this kind of threshold based on examples of both field studies and laboratory experiments. There is also feedback between a geomorphic system and outside; for example, topographic form affects the surrounding microclimate and the resultant microclimate affects future topographic change. In spite of such complexity, Howard (1965, 302) provided a general viewpoint for understanding geomorphic systems in relation to the history of geomorphic processes: “the influence of a past process having acted upon a geomorphic system is proportional to the intensity and duration of its action but inversely proportional to the elapsed time since it action.” Phillips (1988) further examined the role of spatial scales in geomorphic systems based on a review of literature and a case study in a desert stream.

The idea of landforms as open systems was combined with the concept of allometry or allometric growth. The concept was originally developed in biology, and Woldenberg (1966) was one of the first to introduce it to geomorphology. In allometric growth, the growth
rate of a part of a system is proportional to
that of another part or to that of the whole
of the system. The relation is expressed by a
power function or a straight line on a log–log
plot. Although Horton’s law is expressed as a
straight line on a log-normal plot, the integer
stream order represents a logarithmic increase
of absolute magnitude to the base of the stream
bifurcation ratio (Woldenberg 1966). Therefore,
Horton’s law and the regime theory of channel
gometry represented by power functions were
regarded as examples of allometry.

Chorley and Kennedy wrote a unique book,
Physical Geography: A System Approach (1971),
with strong emphasis on the system approach in
physical geography, particularly geomorphology
and hydrology. Another book by Bull, Geomor-
phic Responses to Climatic Change (1991), focuses
on responses of geomorphic systems to climatic
change. It deals with case studies from different
parts of the world, with a consistent emphasis
on system concepts such as relaxation time in
response to climatic change and sensitivity of
landforms. The viewpoint of the book simulated
additional discussion concerning relationships
between climate and landforms.

Re-evaluation of the system approach
and future perspectives

Although the complexity in the structure and
behavior of geomorphic systems has been rec-
ognized for a long time, studies conducted until
the end of the 1980s tended to assume that
general system theory in a classic sense is widely
applicable in geomorphology. This situation
clearly changed in the 1990s, and the applica-
bility of the theory was re-evaluated. This shift
reflected the introduction of nonlinear concepts
to geomorphology and related research fields.
The concepts include catastrophe, chaos, fractals,
self-organization, and dissipative systems.

Phillips (1992) noted that geomorphic systems
tend to exhibit complex and even random
behaviors and spatial patterns, mainly because
numerous individual process–response mecha-
nisms give rise to cumulative impacts in a system.
To solve this complexity, he introduced the the-
ory of nonlinear dynamical systems (NDS).
NDS theory deals with the complexity due to
the coupling of relatively simple systems using
specific concepts including chaos, dissipative
structures, and fractals, as well as bifurcation
and catastrophe theory. Phillips showed that the
application of NDS theory to mass flux provides
results corresponding to the classic concepts of
g geomorphic processes, aggradation, degradation,
and steady states, although the viewpoints of
NDS differ from classical thinking including
general system theory.

Another trend was criticism of equilibrium as
a key concept in geomorphological applications
of general system theory. Although general
system theory does not assume that all systems
are in equilibrium, it indicates that many systems
are approaching equilibrium and often close
to it. However, some subsequent publications
indicated that discrepancy from equilibrium
is much larger than expected, and terms like
“nonequilibrium” and “disequilibrium” were
emphasized (Renwick 1992). In addition, a
geomorphic system may change toward multiple
equilibria rather than a single equilibrium, which
also differs from the classic assumption of general
system theory or the concept of equifinality
(Phillips 1992). In extreme cases, systems may
have no particular normative state at all (Phillips
2009).

Newer concepts concerning system nonlinear-
ity have also been introduced to geomorphol-
ogy. One is self-organized criticality, originally
developed in physics as a property of dynamic
systems that have a critical point as an attractor.
Another concept introduced is the synchrony and
asynchrony of a system in relation to the number and connectivity of system components (Phillips 2012).

Although further new concepts useful for understanding complex systems will doubtless be introduced to geomorphology, researchers also clearly recognize extreme complexity in the behavior and structure of geomorphic systems. Elverfeldt (2012) indicated five major problems in system theory in geomorphology: (i) coherence of basic assumptions and concepts; (ii) openness and determinacy; (iii) the physical basis; (iv) equilibria; and (v) complexity and nonlinearity. In spite of such multiple problems, general system theory still seems to be one of the most important usable ideas in geomorphology. Nevertheless, only a small number of actual cases can be well explained by the theory even if newer related ideas are additionally introduced, because of the high complexity of landforms and environments in the real world. An important aspect is that some recent papers indicate the relatively close relationships between the current heterogeneous recognition of geomorphic systems and classic ways of geomorphological thinking.

In spite of recent progresses in the system approach in geomorphology, including applications to cases in the real world, the concept of geomorphic systems still seems to be a perspective that is more useful for general theoretical or philosophical discussion rather than for practical application. For the further development of geomorphology, it would be fruitful to find more concrete linkages between theoretical ideas of geomorphic systems and the actual development and structure of landforms. With knowledge accumulated through such efforts, the factors controlling the applicability of system theory in geomorphology will be more clearly and objectively understood.

**SEE ALSO:** Fractal analysis; Geomorphic thresholds; Hillslopes; Geomorphology: history; Models in geomorphology

**References**


Further reading
Geomorphic thresholds

Michael Church
The University of British Columbia, Canada

The Oxford English Dictionary defines “threshold” (of a relevant sort; there are also doorsteps) as “a level or point at which something would start or cease to happen or come into effect; a limit below which a stimulus causes no reaction.” There are two definitions here, and both are relevant to geomorphology. The first might apply to some discontinuity in space or time – to a morphological transition such as a slope base or stream bank, or to a process change such as the onset or cessation of glaciation (a “threshold of the first kind”); the second definition implies a transformation in the response to an ongoing process (a “threshold of the second kind”), such as the much-studied threshold for (significant) sediment motion in streams. It is immediately perceived that “thresholds” in geomorphology may be considered in various ways depending upon context. The concept of thresholds was prominently introduced into geomorphology by the American geomorphologist Stanley Schumm (1973, 1979, 1980) as a means to explain the episodic and conditional changes he observed in landscape. He viewed thresholds and the accompanying idea of “complex response” as important correctives to earlier paradigms of continuous evolution of landscapes.

A 1980 volume on “thresholds in geomorphology” (Coates and Vitek 1980) signaled widespread acceptance in the discipline of the importance of thresholds. It contains a range of definitions, including:

- “... self-explanatory terms to denote existence of zones or critical conditions at which change in topographic form or process occurs” (Ford 1980, 345, quoting Frederking and Vitek from a conference presentation, cf. threshold of the first kind);
- “the point at which a stimulus begins to produce a response” (McKerchar 1980, 171, cf. threshold of the second kind);
- “... a boundary condition that separates two distinct but interconnected processes, fed by an identical energy source” (Fairbridge 1980, 43, cf. Chappell 1983);
- “a point or period of time that separates different modes of operation within part of a landscape system” (Bull 1980, 259);
- “boundary conditions ... that when exceeded can cause sudden and vast changes” (Coates and Vitek 1980, 5).

Schumm was largely concerned with landscape morphology, and one of his early examples of a geomorphological threshold was the distinction between ungullied and gullied valley floors in the semiarid American west. He showed a discriminant relation that depended on valley gradient (Figure 1a). He used this diagram to elaborate the concept by distinguishing “intrinsic” and “extrinsic” thresholds. An “intrinsic threshold” is one that occurs as the result of changes endogenous to the geomorphic system; an “extrinsic” threshold is one that is imposed by changes external to the geomorphic system. In the example of valley-floor gullying, an intrinsic threshold would be crossed if sediment accumulation on the valley floor increased the gradient to the point that gullying commenced; an extrinsic threshold would be crossed if an
increase in stream runoff (that is, a climate change) initiated gullyling. The gully threshold might be viewed as a threshold of the first kind if the morphological relations of gradient and landform are emphasized, or of the second kind if the underlying physical process – the point at which runoff becomes sufficiently powerful to initiate erosion – is the focus of attention. Extrinsic threshold crossings characteristically entail some processes intrinsic to the system, but intrinsic threshold crossings may occur independently of extrinsic forcing.

Begin and Schumm (1984) further developed this example and the concept of thresholds by pointing out that the valley-floor threshold is gradational or transitional (Figure 1b). They plotted relative shear stress in the discriminant diagram, showing that the gullied valley floors were distributed over a range of stress values. Intervening effects that would promote such a range might include the state of vegetation cover or character of the sediments on the valley floor, or simply time elapsed since a critical runoff event. A closely related threshold concept is that of the origin of channeled drainage, which has been studied by Horton (1945) with his analysis of the critical distance from a drainage divide to the inception of rills, and by Montgomery and Dietrich (1988) (Figure 2) with their analysis of the relation between source area and gradient at the inception of channeled drainage.

Transitional thresholds might often be modeled using logistic regression, a form of regression that quantifies the probability for a system to be in one
Figure 2  Correlation between drainage area at the point of channel initiation and local gradient at that place for three contrasting terrains on the American west coast. In this diagram, the data are plotted in terms of the critical point at which the threshold occurs. The data scatter shows that the threshold is a transitional one (Montgomery and Dietrich 1988, Figure 1d. Reproduced with permission from Nature Publishing Group.).

state or another, or in some transitional condition, given values of some controlling conditions. Figure 3 shows an example, the transitional probability for tributaries to Sukunka River (in northeastern British Columbia, Canada) to be significant sediment sources (as classified by the presence or absence of a tributary-mouth alluvial fan). Most, but not all, tributaries can be classified as either significant ($p = 1.0$) or nonsignificant ($p = 0.0$), but some do not fit this dichotomous distinction. The discriminating variables in the regression are the ratio of tributary to mainstem area (i.e., relative sediment source area) and the product of tributary gradient and area, a surrogate for tributary stream power.

Ferguson (1992) made a systematic classification of threshold concepts (Figure 4). His “sharp thresholds” are probably mainly thresholds of the first kind (such as a stream bank or glacier margin – a boundary where dominant process and landform change abruptly), whereas his “transitions” and “rate discontinuities” characterize thresholds of the second kind – that is, ones involving physical processes. Indeed, almost all physical processes in geomorphology will give rise to transitional thresholds because thresholds are almost never controlled by a single, simply characterized condition. Examples do exist, however: a glacier may be classified as either “polar” (base of the ice frozen to the substrate) or “temperate” (base of the ice at the pressure-melt temperature and lubricated). But this condition might be spatially variable.

An important characteristic of thresholds, introduced by Chappell (1983), is whether or not the threshold crossing results from persistent or transitory changes in the controlling boundary or forcing condition. Chappell defined cases in which the new state is the consequence of a persistent change in the boundary or forcing condition as “transitive,” a transition that could, presumably, be reversed if the controlling condition subsequently regressed toward its original state. Transitions in which the new state is the consequence of a temporary perturbation in the controlling condition, but where the response does not relax to its original state, he nominated as “intransitive.” The terminology is mathematical: a process is transitive if it is reversible, otherwise it is intransitive. Chappell’s concept of threshold is also closely analogous to the
concept of a “critical point” in physics or chemistry wherein a phenomenon (physical state; landform) may remain stable under a substantial range of variability in the controlling parameter, but then change radically under a small additional change (his opening example is that of an ice cube melting in a beaker, hence crossing the solid–liquid critical point of 0°C at atmospheric standard pressure). The analogy is useful when considering what may be a meaningful threshold in geomorphology.

Ferguson (1992) completed a characterization of threshold crossings by describing the nature of the response (Figure 5). He classified responses as “simple” or “complex,” according to whether the response element relaxes monotonically toward its new state, or oscillates. With these definitions Ferguson tied threshold concepts back to Schumm’s (1973) second conceptual contribution – that of complex response. An additional mode of response might be added in which the response variable partially recovers after the initial perturbation: a damped complex response that would imply either an initial overshoot of the new resultant condition or the influence of compensating elements in the environmental system.

The most prominent instances of thresholds of the second kind in geomorphology are those related to incipient land surface instability. The stability of surficial material on slopes is governed by material strength, given by the
Mohr–Coulomb equation, the full form of which is shown in equation (1),

\[ s = \epsilon' + \epsilon_r + [(1 - m)\rho_bgd_m + \rho_dgd_m\cos^2\theta \tan \phi'] \]  

(1)

in which \( \epsilon' \) is the effective material cohesion, \( \epsilon_r \) is pseudocohesion provided by root strength, \( \rho_b \) is material bulk density, \( \rho_{sat} \) is the density of saturated material, \( d_m \) is the depth of surface material above the potential failure plane, \( \theta \) is the hillslope angle, \( m \) is fractional saturation (in effect, \( d/d_m \)), where \( d \) is the depth of the saturated zone, \( g \) is the acceleration of gravity, and \( \phi' \) is the effective angle of the material’s shear resistance. Material strength is compared with the shear stress \( (\tau) \) on a candidate failure plane equation (2).

\[ \tau = [(1 - m)\rho_b + m\rho_{sat}gd_m\sin \theta \cos \phi'] \]  

(2)

\( F = s/\tau \), known as the “factor of safety,” indicates the likelihood of failure. The interesting aspect of this stability criterion is that it contains several variable quantities, including \( \epsilon' \), \( \epsilon_r \), and \( d \); hence, in terms of parameters that are fixed at a site, notably \( \theta \), or possibly \( d_m \) or \( \phi' \), the threshold will appear to be transitional. If failure occurs, a debris slide or debris flow, or a slump or block slide, happens depending upon the water content of
the material and the cohesion of the failed mass. On steep slopes and in hollows, debris slides or flows are most common.

In flowing water, the equivalent stability criterion is the Shields relation:

$$\tau^* = \frac{\rho gdS}{g(\rho_w - \rho)}D$$  \hspace{1cm} (3)

wherein $$\tau^*$$ is a dimensionless shear stress known as the “Shields number,” $$\rho$$ and $$\rho_w$$ are the material densities of sediment and water, $$d$$ is flow depth, $$S$$ is the energy gradient of the stream, and $$D$$ is the size of the grain to be moved (often assumed to be $$D_{50}$$ in mixtures). This is, again, a ratio of mobilizing and resisting forces (the numerator is just $$\tau$$, the shear force of the flowing water imposed upon a grain while the denominator is grain inertia), so $$\tau^*$$ plays the role of $$1/F$$ (mobilizing force and resisting force are inverted in comparison with $$F$$). In the original experimental investigations, Shields (1936) found $$\tau^* = 0.06$$, approximately, for narrowly graded sediments coarser than about 2 mm, but a substantial range of values, varying from 0.01 to greater than 0.1, has been recorded since and it has been established that the value of $$\tau^*$$ depends, amongst other things, on the value of $$S$$ (Lamb, Dietrich, and Venditti 2008).

In nature, a major factor in the variation of the Shields number is the structural arrangement of the sediment grains on the streambed (Church, Hassan, and Wolcott 1998), and this depends on flow history. Here, then, is an example of a variable threshold.

The actual critical value of both the thresholds just discussed depends on the recent history of the hillslope or stream. Accordingly, the critical state has been analogized to self-organized criticality. The ideal representation of this state is a dry granular pile, an example of which is a talus slope. Since it is dry, the critical limit value of $$\Theta$$ is about 35° (and the threshold is, exceptionally, relatively sharp), while repose angles after failure are somewhat lower. A similar threshold is the limit to relief represented by rock strength – in fact, by the weakest member of a rock formation – in mountains (Schmidt and Montgomery 1995). In streams, however, the critical value of $$\tau^*$$ depends on the state of the streambed at the end of the last event competent to move grains; hence it is a self-organized state of conditional criticality, and this is the case with many process thresholds in nature.

A different kind of threshold – which can perhaps be called a “state threshold” and is a threshold or series of thresholds of the first kind – is represented by changes in channel planform morphology in rivers. The classical empirical analysis gives $$S \propto 1/Q^{1/2}$$ as a discriminator between single and multithread channels, in particular between meandered and braided channels. This relation may be interpreted as a limit condition of stream power that can be accommodated by a single-thread channel system (i.e., $$SQ^{1/2} \sim$$ constant $$\rightarrow SQ/w \sim$$ constant, $$Q$$ being stream discharge and $$w$$ channel width, the combination being directly proportional to specific stream power). The implication is that, beyond some particular magnitude, the sediment load is too great to be moved through a single, increasingly wide channel, so that a division occurs into two or more channels, each of which can convey a part of the load. The discrimination was introduced by the American river engineer Emory Lane (1957) and, in its most famous form by Leopold and Wolman (1957) (which is a remarkably similar precursor of Schumm’s valley floor incision case). Figure 6 gives a developed form of the relation in which it is seen that the “downward” transition (from multithread to single-thread channels) is different than the “upward” transition. The reason for this is probably related to stream bank strength, which might be greater on the upward transition from a relatively stable single-thread channel than on the downward transition, when a single
channel is to be defined within a broader, multithread channel zone.

Many investigators have subsequently presented diagrams similar to Figure 6 with varying datasets, while others have investigated different discriminant functions. Xu (2004) investigated a range of simple, mainly geometric, indices using a large sample of rivers (150–196), both meandered and braided, and flowing in sand or gravel. Remarkably, he found that a fixed value of width/depth ratio, $w/d = 110$, misclassified only 8.7% of his sample. Fredsoe (1978), in an early stability analysis of alluvial channels, had found a result that suggested a discriminant value of about 50. Similarly, Eaton, Millar, and Davidson (2010) found a simple discriminant value of 60. None of these results has any physical meaning, but the range of numerical values implies that there is regional variability in the threshold, probably related to the character of the alluvial material. More strictly physically based criteria have been suggested but have no greater success in discrimination and need not be pursued here. They do, however, imply that the discrimination depends most strongly on $w/d$, which explains the success of the simple discriminant values given above.

Finally, thresholds play an important role in processes that control even the largest-scale features of the planetary surface. Relief production
GEOMORPHIC THRESHOLDS

occurs when rock material is displaced vertically by forces within the crust. In abrupt movements, a threshold is represented by the breaking or yield strength of the rocks associated with the movement. And this introduces yet another aspect of thresholds; they may be either “soft,” presenting such low yield strength that rock creep may occur essentially continuously and aseismically, or “hard,” having strength that yields only to great force and with great release of seismic energy.

SEE ALSO: Eolian erosional processes and landforms; Fluvial depositional processes and landforms; Fluvial erosional processes and landforms; Geomorphic hazards; Glacial depositional processes and landforms; Glacial erosional processes and landforms; Hillslopes

References

An asterisk denotes basic reading.


Fieldwork and mapping have been, and continue to be, the foundation of geomorphological research to study surface processes, process–form relationships and feedbacks, and system couplings to understand landscape evolution. The field tradition in geomorphology has resulted in the ability of geomorphologists to make relevant assessments and develop maps of landscape change based on their conceptual knowledge developed through “experience, intuition and tacit knowledge” (Phillips 1999). The majority of field scientists, however, do not have the ability to quantitatively formalize and evaluate their conceptual knowledge and predict complex spatial patterns based on the spatial distribution of climatological, geological, biological, pedological, and morphometric information. Qualitative interpretation has been accomplished through the analysis of spatial patterns and environmental relationships, along with conceptual understandings of process and geomorphological systems. Reproducibility is then dependent on local geographic and geomorphological knowledge and experience. As argued by Church (1996, 147), field scientists have a “highly censored view of geomorphologically significant events” that is specific to a particular location at a single spatial and temporal scale.

The rapid proliferation of geospatial technologies has transformed the field of geomorphology. As a result, geomorphological assessment and mapping focuses on quantitative characterization of landscape properties and mathematical formalization and modeling of concepts, processes, and system dynamics. Specifically, mapping capabilities have been revolutionized by developments in geodesy, remote sensing, geographic information technology (GIT), geophysics, and high-performance computing and geocomputation (Bishop 2013). It is now possible to collect high-resolution locational, surface biophysical, morphometric, and near-surface spatiotemporal data that can be used to study and map the scale-dependent complexities associated with geomorphological systems and landforms. Similarly, new developments in geostatistics, geocomputation, geovisualization, and spatiotemporal models offer new ways to explore the variance structure in data, extract and synthesize information, and predict and formalize understanding. Given these capabilities, geomorphological research and digital geomorphological mapping (DGM) are increasingly dependent upon the syntheses of knowledge from geomorphology and geographic information science and technology (GIST).

Bishop and Shroder (2004) discuss the reciprocal nature of geomorphological and GIST issues that arise from complex spatial and temporal problems in earth science. These include spatiotemporal data collection, space–time representation, semantic modeling, scale, indeterminate boundaries, information extraction and integration, and geovisualization.
They conclude that geospatial technologies are dramatically improving the ability to study landforms, process mechanics and regimes, process–form relationships, the scale dependencies of surface processes and feedback mechanisms, and the polygenetic evolution of landscapes. The use of DGM, however, tends to be limited to relatively simple and empirical investigations that use satellite imagery and digital elevation models (DEMs) for assessing land-surface properties, topographic structure, erosion and deposition, and process regimes. The resulting interpretation of the Earth surface is used to address a variety of application problems related to volcanism, tectonics, hazards, hydrology, ecology, geology, and pedology. DGM and geomorphometry also facilitate assessment and mapping of specific features/landforms, including hillslopes, ridges, peaks, valleys, volcanoes, glaciers, dunes, floodplains, drumlins, and assemblages of landforms. Smith, Paron, and Griffiths (2011) provide a review of developments in geomorphological mapping, including types of data, methods, and mapping applications.

DGM efforts have been constrained by geographic area and mapping objectives, and various techniques and approaches do not necessarily have universal applicability in geomorphology. Many constraints are caused by the need to account for conceptual and technical issues that are not addressed in the mathematical underpinnings of techniques and analytical approaches. An attempt to better formalize understanding of mapping geomorphological systems involves research on semantic modeling, landform elemental object/unit segmentation, and assessment of the morphographic structure of the landscape. The mapping objectives, and therefore the difficulty of DGM applications, can range from relatively simplistic (e.g., land cover or topographic feature mapping) to extremely difficult (e.g., process regimes and polygenetic overprinting). The landscape can be mapped using a particular theme or combination of themes (geobotanical, morphometric, hydrologic, geologic, etc.), although these are notoriously difficult to accomplish given that software tools have not been designed to specifically address a multitude of geomorphological concepts and issues. Unfortunately, analytical success is often dependent on the scope of the mapping objectives and the degree of landscape complexity, meaning that issues of mapping accuracy and reproducibility remain unanswered. An improved conceptual and analytical framework to extract and integrate diagnostic landscape information is warranted, to ensure that DGM becomes fundamental to integrative science.

Geospatial technologies

Fundamental geospatial technologies utilized in DGM include global navigation satellite systems (GNSS), remote sensing (including geophysical sensors), and GIT. These technologies are used to obtain spatiotemporal data that contain detailed locational, biophysical, and structural information about the landscape and near-surface environment. Spatial analysis and pattern recognition algorithms can then be used to characterize landscape properties and produce thematic maps related to various geomorphological and environmental science themes.

Multispectral and multitemporal remotely sensed data are critical for geomorphological mapping, as surface biophysical properties inherently characterize the composition and land-cover structure of the landscape and permit change detection (Figure 1). The biophysical properties of the surface are related to surface processes and landscape stability. High-resolution and hyperspectral image data,
Figure 1 Multitemporal satellite imagery can be used to detect changes on the landscape. Spatial and temporal variations in spectral reflectance are detected using feature-tracking software to map the ice-flow velocity fields of alpine glaciers around the world (Ngojumba Glacier in Nepal). This information is required to assess climate change and geomorphological system response to climate forcing. Other applications include detecting mass movements and the migration of dunes and dune fields.

Along with digital image processing software, permit the enhancement of landscape features to differentiate complexities of the landscape. It is widely known, however, that surface compositional variations and land-cover spatial patterns may not directly characterize or delineate landform entities or landscape concepts/processes such as instability, erosion, deformation, and uplift. Nevertheless, geomorphologists routinely use aerial photography and satellite imagery, and interpret spectral and textural variations in reflectance to map landforms, particularly when working with data acquired over remote mountain landscapes and extraterrestrial planetary surfaces.

Topographic information from terrestrial and airborne LiDAR, off-nadir optical imaging, and microwave synthetic aperture and interferometric radar systems has dramatically increased the ability to address a variety of geomorphological mapping problems. Morphometric information about the landscape and changes in surface altitude permit detailed examination of the shape of landforms and directly enable estimates of rates of processes including erosion, deposition, ablation, and energy distribution and flux (Figure 2). The associated field of geomorphometry and the use of DEMs have dramatically influenced geomorphological research in a variety of environments by quantifying the shape and contextual characteristics of the topography. Key parameters such as slope angle, slope azimuth, and curvature metrics govern a variety of surface processes, including water and sediment transport, erosion, deposition, surface temperature, ablation, surface irradiance, precipitation, evapotranspiration, and physical and chemical weathering. Geomorphometry involves the principles of sampling the land surface, generating a digital terrain model, correcting errors and artifacts in the model, deriving land-surface parameters and terrain objects, and applications using parameters and objects. Geomorphometry incorporates the geosciences, computer science, and engineering, and is rapidly becoming a central theme in geomorphological research and DGM.
Digital elevation models and topographic solar radiation transfer models can be used to simulate the surface irradiance over the landscape: total direct irradiance (diurnal integration) variations over the Nanga Parbat Massif in Pakistan on August 15, 2013 (left); total diffuse-skylight irradiance (diurnal integration) variations over the Nanga Parbat Massif for the same day (right). Gray-scale values depict the magnitude of the irradiance component, with white signifying relatively high irradiance and black depicting relatively low irradiance values. Multiscale topographic effects dramatically influence the shortwave energy distribution over mountain environments. Surface irradiance is important for assessing evapotranspiration, drought, surface energy budget, and other aspects of geomorphological systems.

Geomorphometry also serves as a basis for land-surface segmentation and object characterization, which permit the identification and mapping of fundamental elementary forms and landform entities that evolve into and out of existence as the landscape evolves. This can be particularly useful to the understanding of the critical zone, where surfaces that characterize vegetation canopies, land surface, bare surface, soil horizons, stratigraphic layers, bedrock, and subsurface seismic structures can be analyzed spatially and information synthesized to more accurately map various aspects of geomorphological and complex critical zone systems. More detailed information about geomorphometry can be found elsewhere (Hengl and Reuter 2009).

The data derived from these evolving technologies are exceedingly rich and have driven significant advancements in data storage, management, analysis, and visualization capabilities. Geographic information systems (GIS) are now routinely utilized for DGM; however, users must develop complex workflows to manipulate, analyze, and combine information to generate geomorphological maps. Considerable energy is spent on learning GIS software and writing scripts/applications to produce parameters/indices and accomplish mapping tasks, as individual programs to map specific landforms or perform a specific type of geomorphological assessment do not usually exist. Currently, there are only a small handful of programs to produce
the most basic geomorphometric parameters, compared to the wide variety of parameters described in the literature.

The empirical nature of GIS analysis is extremely exploratory and dependent on traditional statistical techniques including statistical indices/metrics and the subjective use of weightings or sensitivity parameters to permit flexibility in obtaining results. It is important to understand that such empirical metrics do not necessarily characterize process or landform boundaries, but are recognized as proxies for fundamental landform properties. Most approaches to mapping attempt to decompose the altitude field based upon local criteria (i.e., homogeneity, heterogeneity, land-surface properties) in order to construct larger landforms or landscape spatial entities. Unfortunately, numerous geomorphological concepts are disconnected from analytical solutions, and a synthesis approach that accounts for knowledge and analytical reasoning is sorely needed. Nevertheless, GIT has been successfully used to address a variety of issues related to geomorphological concepts such as landform features, landscape and landform spatial organization, scale dependency, erosion mapping, deformation, uplift and neo-tectonics, landscape stability, hazards, process regimes, and relief production.

Mapping perspectives

Modern-day geospatial technologies permit the mapping of the functional and structural aspects of the land surface but different requirements dictate which mapping perspective should be used to address a DGM problem. For example, hydrological mapping is required to assess drainage networks and basins and the nature of surface runoff and stream/river flow. In the same way, assessing and mapping landscape evolution requires multiple perspectives that incorporate hydrological, geological, and morphometric information to characterize fluvial erosion, sediment transport, denudation, and relief production.

A common mapping perspective that has been widely utilized is the land systems approach. Mapping involves identifying spatial units that have a reoccurring set of land cover, topography, soils, and vegetation that characterize the climate, geology, and geomorphology. This approach tends to be limited to regional mapping because it depends on subjective criteria. This approach does not characterize the complex 3-D nature of landforms, although the inclusion of more quantitative information about the near-surface environment should allow for improved land systems differentiation.

Mapping from a geological perspective requires the integration of surface composition, mineralogy, and lithology with respect to geologic structure and tectonics. Hyperspectral remote sensing can be used to assess surface mineralogy and lithological variation. Image analysis and geomorphometry can be used for structural mapping efforts. Rock strength, deformation, and uplift can govern the topographic structure, as the interplay between erosion and tectonics may influence basin morphology, drainage density, relief production, and many other properties of the topography. There is considerable potential in using existing land-surface parameter algorithms to characterize subsurface data, as seismic, electromagnetic, and ground-penetrating radar data highlight subsurface properties, stratigraphy, and structure. Examining the substructure of landforms can provide a wealth of information for studying landforms and understanding the nature of polygenetic overprinting, as well as the role of the antecedent topography on landform evolution.
DGM can also be achieved from a morphometry and morphography perspective. As noted, the morphometry of the landscape characterizes the shape of landforms and the properties of the topography, while the morphography is the mapped description of the configuration structure. Maps of this type are rather rare, but maps of the fundamental first-order and second-order derivatives of the altitude field are common (Figure 3). These objects may be spatially aggregated and structured into a morphographic map. Producing multiscale morphometric parameters that characterize mesoscale properties of the landscape can greatly improve our understanding of processes and natural hazards.

Sediment transfer-cascade systems represent spatial structures that are interconnected by flows of mass and energy at various operational scales within the geomorphological system. For example, curvature can be used to delineate those portions of the landscape that depict mass movements within sub-basins and deposits of rock and debris at the base of valley floors. Such mapping provides the basis for development of models of sediment budgets.

Geomorphic genesis is a critical theme in mapping, as it serves as the basis for landform taxonomy. Process domain mapping is, therefore, an important goal of DGM. The ability to reliably depict a process domain, however, is notoriously difficult, as numerous surface processes and exogenic and endogenic forcing factors need to be accounted for. Furthermore, the complex nature of polygenetic evolution must be addressed, such that process–form relationships that are diagnostic in nature serve as the basis for mapping. For example, it is

Figure 3  Slope angle map (left) and unsphericity curvature map (right) over the western Himalayas in Pakistan. The morphometric properties provide insight into the distribution of geomorphological features and the processes responsible for their formation. Mountain valley floors and large alpine glacier surfaces exhibit relatively low slope angles (dark gray tones) and high unsphericity (white gray tones) values depict more planar surfaces. Steep slopes and high altitudes also depict relief production, while changes in the spatial variation of morphometric patterns also depict lithological variations (northeastern area versus the NW-SE central region).
very difficult to map a tectonic uplift zone, since the morphometry, morphography (the description of the morphology), and geological conditions must be differentiated from the surrounding landscape that is not undergoing uplift. Object-oriented mapping can be used to assess deformation and uplift by examining relief production, basin geometry, and drainage density and organization.

Mapping the age of landforms and surfaces is very difficult to accomplish using traditional geospatial technologies. There is a paucity of high-quality temporal data, so spatial information is often used as a proxy for time. For example, if the surface roughness of a particular landform is relatively high compared to another surface, that surface may be considered to be older if surface roughness is assumed to increase with time. This is complicated by the lack of spatial uniformity and (temporal) steadiness of the processes acting on the landscape. In many complex environments, rates of erosion and deposition are not accurately known; this has led to controversy in the literature about the role of different processes in the evolution of mountain systems and aeolian landscapes on Earth and on Mars.

The complexity of the landscape has required geomorphologists to adopt an empirical approach versus a more formal mapping framework or system based upon landform taxonomy, semantic modeling, knowledge, process–form relationships, and landscape evolution modeling (LEM). An argument can be made for both approaches based upon findings to date; however, the complexity of DGM seems to warrant a more systematic approach for addressing conceptual and technical issues. A variety of algorithms and approaches to DGM have emerged in recent years providing for new capabilities that should significantly advance understanding of landscape development and change. At the same time, advancements in geomorphology have identified a need and opportunity for geospatial technology solutions.

**Mapping issues**

Rapid progress in geospatial technologies has raised a number of questions concerning the conceptual and technical aspects of DGM. Bishop et al. (2012) discuss a variety of concepts, issues, and research directions for DGM, and highlighted the capabilities and limitations associated with geospatial technologies. In general, very few researchers have attempted to formalize DGM and account for the plethora of concepts that are important in advancing this part of geomorphology. There does not appear to be a consistent conceptual or analytical framework by which to map the landscape at multiple scales, although there is a need for new geospatial technology solutions to generate standardized information.

**Landscape complexity**

An ongoing and fundamental issue in mapping involves representing spatiotemporal data, as data models and structures do not currently account for the attribute and dimensional complexity of the landscape including the near-surface environment and critical zone. Static cartographic snapshots in time (i.e., GIS layers) do not adequately address spatial and temporal variation in terms of operational scale dependencies. Furthermore, entity models do not adequately account for environmental variation and the issue of indeterminate boundaries of landforms and phenomena. Object-oriented data models have been proposed, while others have indicated that a hybrid approach to representing the landscape with fields of objects enables flexibility and better integration of data and information. Choice of data structures is an
important consideration, as attempts are made to represent complex properties of geospatial data and permit efficient information extraction and storage. New representation schemes offer many new possibilities for DGM that can address boundary, geometry, topology, process, scale, and information integration issues.

**Semantic modeling**

Landform semantic modeling is another important topic. The terminology that is used to characterize and map landforms/landscapes can be different in a variety of disciplines and even within the same subdiscipline of geomorphology (e.g., ridge versus foredune in coastal environments). Formalization and repeatability in DGM is, therefore, dependent upon landform definitions that determine semantic representation of landforms. The main objective of semantic modeling is to provide for a correct semantic definition of landform that is useful to all disciplines, and to incorporate, composition, geometry, and spatial topological information, such that a semantic definition is consistent with taxonomic schemes and system dynamics. It is challenging to formulate basic definitions that account for human perception, process, scale-dependence, geometry, spatial topology, indeterminate boundaries (Figure 4), and a multitude of geomorphological concepts related to a specific environment. For example, mountain peaks (Figure 5) can be mapped with different spatial distributions depending upon the use of concepts such as summit, prominence, symmetry, scale, relief production, and total rock mass. Semantic modeling can be thought of as a way to bring together data and information, concepts, perceptions, and functionality that can be adhered to by different disciplines and applications. Semantic modeling and representation requires that error and uncertainty are explicitly addressed.

**Scale**

Another fundamental issue is the complex topic of scale. Numerous perspectives exist and geomorphologists should be aware of geographic, cartographic, measurement, operational, and computational scale definitions (Bishop et al. 2012). Inherent in any treatment of scale is the recognition of the complex linkages of spatiotemporal concepts and the multiplicity of spatiotemporal scales associated with phenomena, processes, and systems. Therefore, research into process–form relationships and the spatiotemporal organization of the landscape is fundamental to DGM. In particular, special emphasis must be placed on identifying the operational scale of processes and ensuring that this scale is linked to the computational scale at which data are analyzed and information is integrated. Another important scale concept is the
Figure 5  Field photograph of K2 located in Pakistan (8612 m). Many mountain peaks exhibit different geometric shapes, degrees of prominence, and scale-dependent relief. Many mountains are part of a ridge system that exhibits multiple summits. Humans perceive mountains in different ways depending upon their location and application interests.

anisotropic nature of spatiotemporal variation on the landscape, as the spatial structure of the topography can be highly directionally dependent through the influence of geologic structure and deformation, and the spatial variability and nature of surface processes (Figure 6).

Polygenetic evolution

Since dynamic simulations can be used to study the nature of polygenetic evolution, LEM is another critical issue related to DGM. Parameterization schemes that account for the conservation of mass and energy characterize the process mechanics of weathering, regolith production, erosion, deposition, and uplift by a variety of agents including the wind, water, ice, and gravity. Landscape simulations can be programmed to produce maps of landscape properties, the magnitude of processes, and depict the spatial distribution of process regimes. This is in contrast to simplistic spatial relationships that
Figure 6  Three graphs characterizing the anisotropic variation in relief for three points on the Texas Sand Sheet near Padre Island National Seashore.

are used to assess and map the landscape using GIS. Given that geospatial technologies and GIS are not yet well suited for addressing temporal concepts and spatiotemporal dynamics, LEM links process, morphometry, spatial topology, and additional information in a holistic way, such that landscape system characterization and mapping are possible. Such dynamic mapping capabilities are important to go beyond traditional space-dominated DGM.
Analytical reasoning

An important series of related issues that has not been adequately addressed includes knowledge representation, information integration, and analytical reasoning. Numerous developments in sensor technologies now provide more information about the landscape. Nevertheless, a multitude of challenges are faced regarding information integration and visualization of spatiotemporal data. A rigorous framework has yet to be established by which it is possible to associate or use data and information as proxies for important geomorphological concepts. Theoretical geomorphology is rich with spatiotemporal concepts and conceptual models that attempt to formalize our understanding of geomorphological processes and landform evolution. Recent developments in artificial intelligence and fuzzy systems allow current knowledge to be represented, multiscale data/information to be integrated, and analytical reasoning to be employed to assess and model landscape evolution based on direct and indirect causative relationships. This is required, as some aspects of geomorphological assessment and problem-solving do not appear to be mathematically tractable, or cannot be derived from first principles. Geocomputational approaches to DGM can be used to test the utility and significance of concepts and variables in LEM. This is demonstrated in the context of another issue related to validating conceptual models and field observations.

Analytical reasoning permits formalization of process–form relationships and information integration analogous to human cognition, and therefore can be used to examine internal and external geomorphic forcing functions that can be highly variable across a range of spatial and temporal scales. Geocomputational approaches can be quite powerful when integrated with new spatial and temporal data and various algorithms that make it possible to quantify landscape morphology, assess surface biophysical interactions, link process with form, and characterize the scale-dependence and polygenetic nature of the landscape. There is the potential for significant advancements in geomorphology if knowledge and conceptual understandings of landscape change can be formalized and tested through analytical reasoning. For example, aeolian instability can be predicted using an analytical reasoning approach through the formalization and conceptualization of understanding of how prolonged drought conditions limit the ability of vegetation and surface moisture to stabilize the surface (Figure 7).

Future directions

Geomorphological maps are an important source of information for studying geomorphological systems and the critical zone. They are also essential for assessing and managing natural resources and planning for sustainability. Their significance and utility is based upon information fusion and a systems integration perspective, as surface biophysical properties, landscape morphometry, spatial topology, and near-surface properties and structure can be represented, mapped, and visualized uniquely using geospatial technologies. Rapid advancements in sensor technologies, data collection, and analysis and modeling have brought about many new developments and capabilities. Access to data and information has dramatically increased, thereby increasing DEM availability and the impact of geomorphometry on geomorphology. There are, however, many data and representational issues with incorporating multispectral, multi-temporal, and near-surface information into a geomorphological mapping framework. This
Figure 7 South Texas Sand Sheet showing northern and southern lobes within the King and Kenedy Ranches (a). A simple analytical reasoning model was used to predict aeolian instability (Houser, Bishop, and Barrineau, 2015). The instability concept (aeolian transport or no aeolian transport) is the output of the model and predicted aeolian instability (dark speckling) is presented (b).

will require a more rigorous conceptual, representational, and information synthesis approach that includes advancements in geomorphology and GIST.

Advancements in science are driven by theoretical and conceptual developments from a variety of disciplines including GIST and geomorphology. Scientific progress has not kept pace with new and rapidly evolving geospatial technologies. Similarly, the scientific significance of various geospatial technology solutions has yet to be discovered in geomorphology and DGM. Consequently, research directions should attempt to focus on science–technology solutions for characterizing and understanding landform evolution and complex geomorphological systems. Geographers can contribute important solutions that enable the prediction of spatial landforms that result from forcing factors, processes, and polygenetic landscape evolution. This requires that knowledge and conceptual understandings are transformed into models, representation schemes, and analysis and modeling approaches that facilitate problem-solving and decision-making.

Geospatial technologies and numerical modeling will allow geomorphologists to formalize process–form and process–pattern relationships in the very near future. However, research involving new geostatistical approaches to analysis of spatiotemporal variation is sorely needed in order to characterize the complex evolution of landforms and landscapes over space and time. Furthermore, it is necessary to test model parameterization schemes to determine if external forcings, process mechanics, multiscale topographic forcings, and feedbacks are appropriately characterized and produce results that can predict landform evolution. Research involving the integration of LEM and DGM represents a new direction.

Progress is also required on formalizing geomorphological knowledge, evaluating and determining the utility of taxonomic schemes and semantic modeling to better formalize
DGM, and to make sure that information synthesis can be integrated into a spatiotemporal topological framework that permits consistent and repeatable results. While it is recognized that there is tremendous diversity in methodological approaches for addressing mapping applications, and that progress in mapping soils, vegetation, terrain, landforms, lithology, and other applications will continue, an integrative framework and standardized information products appear to be warranted. As Bishop et al. (2012) have indicated, progress in this direction will elevate the status of DGM in the science community and permit routine use of geomorphological maps in integrative science and practical problem-solving.

SEE ALSO: Biogeomorphology; Climate change and land ice; Coastal depositional processes and landforms; Coastal erosion processes and landforms; Earth system science; Geographic information science; Geographic information system; Geomorphic hazards; Geomorphic systems; Glacial depositional processes and landforms; Glacial erosional processes and landforms; Information synthesis; Land systems science; Landforms and physiography; Landscape; LiDAR; Mass movement processes and landforms; Representation; Scale; Spatial analysis; Spatial concepts; Tectonic geomorphology

References


Further reading

The discipline of geomorphology has a long history that has been expertly reviewed by Chorley, Dunn, and Beckinsale (1964), Chorley, Beckinsale, and Dunn (1973), Beckinsale and Chorley (1991), Burt et al. (2008), and Orme and Sack (2013). The role of various national schools has been recounted by Walker and Grabau (1993).

Prior to the nineteenth century, geomorphology was hugely influenced by the belief that the Earth was created by divine intervention only 6000 years ago and had been molded subsequently by catastrophes like Noah’s flood. The time span for geomorphological processes to operate and for forms to develop was very brief. However, towards the end of the eighteenth century ideas began to change, and scientists such as Hutton, Playfair, and Lyell argued for the importance of gradual subaerial denudation over millennia. Gradualist and uniformitarian ideas took hold, the conception that the Earth was old and had a long history was appreciated, and the fluvialists argued for the dominance of rivers in denuding the landscape through slow, long continued action (Kennedy 2006). However, these radical ideas did not go unchallenged. There were some structuralists who believed that valleys were essentially clefts or rents in the ground surface rather than the product of stream erosion as had been maintained by Playfair. There were also the diluvialists who invoked the biblical Flood to explain many geomorphological phenomena.

The glacial theory

In the 1820s and 1830s some scientists started to suggest that glaciers and ice caps had once been much more extensive and could account for much of what was then called “drift.” The most famous exponent of the glacial theory was Louis Agassiz, who developed this concept on the basis of work in the Alps in his Discours de Neuchâtel (an address to the Société helvétique des sciences naturelles of 1837). The glacial theory was not well received by some members of the geological establishment. Lyell found that it was incompatible with his uniformitarian ideas, as in a sense it was, and attributed many of the allegedly glacial phenomena to marine submergence and wave action. Moreover, well into the twentieth century some British geomorphologists continued to argue that glaciers protected rather than eroded the landscape. The significance of the Ice Age beyond Europe was soon recognized and Agassiz’s views were adopted in the USA. A further major development in glacial ideas occurred in the 1870s when it became recognized that there had been more than one glacial advance and that these had been separated by warm phases, called interglacials. In addition, Croll recognized that orbital fluctuations could have caused multiple alternations of glacial and interglacial periods. These trends led to the work of James Geikie, who in The Great Ice Age (1874) appreciated the importance of interglacial periods, and to the
classic and durable Penck and Brückner model of glacial chronology in the Alps. Scientists also started to be intrigued by other sorts of climatic change that might have occurred in nonglaciated regions. Lake basins, of the type that abound in the Basin and Range Province of the American West, with their spectacular abandoned shorelines, gave particularly clear evidence of hydrological change. By the time of World War I a picture was emerging of the scale of climatic change that had taken place in lower latitudes and of the very substantial alterations that had taken place in climatic belts as made evident not only by desiccated or shrunken lakes, but also by old river systems and ancient sand dunes.

River valleys and the power of fluvial denudation

Although Hutton, Playfair, and Lyell had made clear the role of rivers in landscape development, the acceptance of fluvialism was not a straightforward matter as there were many phenomena that appeared to cast doubt upon whether valleys had actually been produced by rivers (Kennedy 2006). In the 1860s, however, geomorphologists began to appreciate once again that rivers molded valleys and were capable of achieving a great deal of geomorphological work and planation. There were various reasons for this. First, increasing acceptance of the power of former glaciers to cause wholesale transformation of the landscape and to produce features such as lake basins and hanging valleys explained away some drainage anomalies. Second, catastrophic/structural views on valley development were viewed with less favor. Third, when geomorphologists moved away from the relatively stable, sod-covered landscape of the British Isles to places like the Pacific islands or the mountains of Ethiopia, they encountered strong evidence of the power of rivers. Fourth, data on sediment loads of rivers demonstrated that they could indeed achieve a great deal of work. The power of fluvialism, however, became sealed as a fundamental concept in geomorphology because in the American West, including the Grand Canyon, there was abundant and dramatic evidence for the power of rivers. This had a great impact on American geomorphologists such as J.W. Powell, C. Dutton, G.K. Gilbert, and W.J. McGee. This work, together with that from French hydraulic engineers, was used in France to good effect by La Noë and Margerie. Their espousal of fluvialism transformed French geomorphology at the end of the nineteenth century.

Rock decay

During the nineteenth century great strides were made in the understanding of physical, chemical, and biological weathering processes (Goudie and Viles 2008). Knowledge of weathering phenomena owed a great deal to the development of an independent science of pedology, or soil science, most notably by scholars like Dokuchaev in Russia. One particular aspect of weathering-related studies was the science of limestone (karstic) relief and solution processes. Prime importance must be accorded to work on the Dinaric Karst and in particular to the extensive studies of Jovan Cvijić. His Das Karst-phanomen (1893), and many subsequent works, laid the theoretical foundations of many of our current ideas, though Serbian scholars had made some important studies before him.

Mountain building

During the nineteenth century there was considerable interest in how mountains formed and
in motions of the Earth’s crust. Suess in Austria and Dana in the USA proposed that mountains formed through compressive stresses generated by a gradual thermal contraction of the whole Earth. Suess argued that on a contracting Earth, mountains resulted from a wrinkling of the crust to accommodate a diminishing surface area. The contraction theory was the dominant paradigm for most of the nineteenth century (Oldroyd 1996, 171).

Another important concept was isostasy. This was made evident by studies in formerly glaciated terrains which would have been affected by downwarping and upheaval in response to ice cap advance and recession, respectively. This was the birth of the theory of glacio-isostasy. During his classic study of pluvial Lake Bonneville in the western USA, G.K. Gilbert found a dome-like pattern of uplift of former shorelines and inferred that this indicated hydro-isostatic recovery following lake desiccation. Using his experience from the American West, where many mountains appeared to be composed of igneous rocks intruded into sedimentary sequences, Dutton, who invented the term “isostasy” in 1882, argued that crustal deformation could be understood as a response to isostatic compensation. He postulated that uplifted portions of the continent are eroded, that material is transported to coastal regions, and that the weight of this material causes subsidence along the continental margins, which causes displacement of materials at depth, with this material moving laterally and producing igneous intrusions and further uplift of the continent. Gilbert built upon Dutton’s ideas and noted that in the Basin and Range Province mountain building was associated with many faults and with crustal extension rather than crustal contraction. The significance of crustal tension was also recognized by Gregory who, working in the context of East Africa, was the first to use the term “rift valley.” The hypothesis that mountain building could result from continental drift developed in the early twentieth century through the work of Taylor and Wegener.

Davis and the cycle

William Morris Davis was the leading American geomorphologist of the late nineteenth and early twentieth centuries. His great contribution was to produce in the 1880s and 1890s a deductive model of landscape evolution, called the cycle of erosion or the geographical cycle. Davis believed that landscapes were the product of three factors: structure (geological setting, rock character, etc.), process (weathering, erosion, etc.), and time (stage) in an evolutionary sequence. Stage was what most interested him. He suggested that the starting point of the cycle was the uplift of a broadly flat, low-lying surface. This was followed by a phase he termed “youth,” when streams become established and started to cut down and to develop networks. Much of the original flat surface remained. In the phase he termed “maturity” the valleys widened so that the original flat surface had been largely eroded away and streams drained the entire landscape. The streams began to meander across wide floodplains and the hill slopes became gradually less steep. In “old age” the landscape became so denuded that a low-relief surface close to sea level (a peneplain) developed, with only low hills (monadnocks) rising above it.

Initially Davis’s model was postulated in the context of humid temperate (“normal”) conditions, but it was then extended by Davis and successors to other environments, including arid, glacial, coastal, savanna, karstic, and periglacial landscapes. Chorley et al. (1973) argued that Davis’s cyclic model was not very successful in Germany. Here, W. Penck’s model of slope evolution, often seen as the antithesis of Davis,
involved more complex tectonic changes than that of Davis, and he regarded slopes as evolving in a different manner (by slope replacement rather than slope decline) through time. An alternative model of slope development by parallel retreat leading to pediplanation was put forward by L.C. King. His model, developed in South Africa, represents an amalgam of the views of Davis and Penck: episodic uplift resulting in both downwearing and backwearing, with the parallel retreat of slopes leading to the formation of low-angle rock-cut surfaces (pediments) which coalesced to form pediplains. By the mid-twentieth century the Davisian model was becoming less dominant. This was partly because there was a growing awareness of crustal mobility that could not sustain notions of initial uplift followed by prolonged structural quiescence.

Denudation chronology and long-term evolution

For geomorphologists, the explanation of how landscapes came to attain their present form has always been a major objective. Up to the 1960s, many workers adopted an historical approach to landscape evolution, their aim being to identify the sequence of stages of erosional development that demonstrated how contemporary landscapes had been sculpted from hypothetical initial fairly uniform and featureless topographies. This sequential approach came to be known as “denudation chronology” (Jones 2004). Classical denudation chronology sought to identify evidence of past planation surfaces and erosional levels in a landscape, in whatever way they formed, and to place them in a time sequence. During the first half of the twentieth century, this became a major preoccupation of geomorphologists in America, under the influence of D.W. Johnson, in Britain, where S.W. Wooldridge was a dominant figure, and in France, where H. Baulig’s study of the Massif Central established a blueprint for subsequent work. Since the 1960s there has been less interest in classical denudation chronology, for that decade witnessed the onset of radical changes to prevailing views of the past arising from growing knowledge about global tectonics and Quaternary climate change. Moreover, many geomorphologists concentrated on understanding the role of present-day processes rather than trying to establish a long-term evolutionary history based on often small fragments of ancient landscapes preserved in the present landscape. However, more recently still there has been an expansion of studies of long-term landform evolution based on such techniques as cosmogenic nuclide and apatite fission analyses. These can give us data on the ages of ancient landforms and on the rates at which they have been denuded over millions of years.

Climatic geomorphology

In the twentieth century, particularly in Germany and France, climatic geomorphology was a major approach. One strand of the development of climatic geomorphology was the study of periglacial and permafrost processes by European explorers of the vast subarctic regions of North America and Eurasia, though it was Lozinski who provided the first unifying concepts of periglacial geomorphology just before World War I. Other distinctive cold climate phenomena were also recognized, including nivation and solifluction. Among the phenomena that scientists studied in lower latitudes were loess, desert dunes, desert weathering, coral reefs, deep weathering, laterites, and inselbergs. Von Richthofen, working in China, cogently argued that loess, a largely nonstratified and nonconsolidated silt,
probably had an eolian origin and was produced by dust storms transporting silts from deserts and depositing them on desert margins. The colonization of the Sahara by the French from the 1880s onwards led to some of the first serious work on desert dunes. However, dunes were not the only field of interest of desert travelers, for the exploration of deserts in the nineteenth century gradually led to the emergence of studies that established the nature of desert processes and their differences from those in other environments. French scientists were very active in the Western Sahara and accumulated a great deal of vital information on the full range of desert landforms. Also notable was the work of Walther, who worked in the deserts of North Africa, Sinai, the USA, and Australia. His *Das Gesetz der Wüstenbildung in Gegenwart und Vorzeit* (1900) was the first full-scale book devoted to desert geomorphological processes and he championed the role of such mechanisms as thermal fatigue (insolation) weathering, salt weathering, and deflation. The development of ideas on the role of wind in drylands is discussed by Goudie (2008). Geomorphologists gradually came to see the distinctive nature of humid tropical landforms and processes and appreciated the extent of deep weathering and laterite formation in the tropics in comparison with higher latitudes. In the USA, Davis recognized “accidents,” whereby nontemperate and nonhumid climatic regions were seen as deviants from his normal cycle of erosion and, inspired by Passarge’s work in the Kalahari, he introduced his arid cycle. Some regard Davis as one of the founders of climatic geomorphology, although the leading French climatic geomorphologists, Tricart and Cailleux, criticized him for his neglect of the climatic factor in landform development. Much important work was undertaken on dividing the world into morphoclimatic regions with distinctive landform assemblages in France (e.g., Birot), Germany (e.g., Büdel), and New Zealand (e.g., Cotton).

**G.K. Gilbert and dynamic equilibrium**

G.K. Gilbert was a remarkable and innovative American geomorphologist whose name is most often associated with what is often termed “dynamic geomorphology.” This blossomed in the second half of the twentieth century, and was defined by Strahler as an approach which treats geomorphic processes as “gravitational or molecular shear stresses, acting on elastic, plastic or fluid Earth materials to produce the characteristic varieties of strain or failure which we recognize as the processes of weathering, erosion, transportation and deposition.” This is exemplified in Gilbert’s *Report on the Geology of the Henry Mountains* (1877), his study of the convexity of hill tops, and his work on the transportation of debris by running water. In some areas of geomorphology, studies based on an analysis of force and resistance occurred earlier than in others (e.g., Terzaghi’s work on slopes and rock mechanics in the 1920s, Bagnold’s work on eolian forms and processes in the 1930s, Hjulström’s studies of processes in gravel rivers in the 1930s, the work of various physicists, such as Nye, Glen, and Perutz, on glacier dynamics in the 1950s, and Johnson’s study of shore processes and shoreline development).

**The latest phase**

Since the 1960s there have been at least four major developments – plate tectonics, the revolution in Quaternary science, quantitative process-oriented geomorphology, and the use of remote sensing – that have taken place. These
have been reviewed by Burt et al. (2008) and Orme and Sack (2013).

Particularly since the 1960s, geomorphology has been concerned with the study of exogenic processes, often at a detailed and Newtonian reductionist level. Indeed, there are two major types of Earth surface process, generally labeled as exogenic and endogenic. This produces a paradox in terms of temporal scales. The former refers to processes that are ultimately fueled by the sun’s energy and usually operating through the climate system, such as weathering of rocks and erosion and sediment transport by water, ice, wind, waves, and organisms. This has often involved studies with a short timescale and at a local spatial scale. Classic and influential examples of this genre are that by Leopold and coworkers on fluvial processes (1964) and that by Yatsu on weathering (1988). Endogenic, on the other hand, refers to those processes powered by energy from the inside of the Earth, such as volcanic and tectonic processes. These operate over long timescales and large regional extents. They have received increased attention as a consequence of the plate tectonics paradigm that has developed since the 1960s, and have contributed to a greater understanding of the global pattern of phenomena such as volcanoes, rifts, mountains, and seamounts. Remote sensing has enabled the mapping and identification of landforms at a regional or global scale. Endogenic processes have also contributed to our knowledge of rates of denudation and channel incision in areas of orogeny. Some of the most interesting current work has shown that climate, as well as the erosional development of the landscape, feeds back into the ongoing tectonic processes. The study of rates of chemical weathering and physical denudation under different climatic and tectonic conditions has been boosted by a concern with how these processes relate to the global carbon cycle. The reductionist exogenic process geomorphology and the new models of long-term landscape evolution, associated with new ideas on plate tectonics and novel geochronometric techniques, need to be combined more effectively than in the past.

The revolution in Quaternary science, much of it resulting from studies of ocean cores, has revealed the duration, frequency, magnitude, and timing of climatic changes in the Late Cenozoic. New dating technologies and methods for reconstructing past environments have been fundamental (Anderson, Goudie, and Parker 2013).

Processes have been studied in detail using new methods of instrumentation, including data loggers, flumes, wind tunnels and the like, but above all geomorphologists adopted quantitative methods (including morphometry), statistical analysis, systems thinking, and model building. Among the processes that have received major attention are those associated with the role of plants and animals, as shown by the emergence of biogeomorphology and zoogeomorphology as major subdisciplines.

Geomorphologists have become much involved with the application of their science and with the investigation of a range of natural hazards. The roles of the geomorphologist in hazard research are many. Of great importance are: the mapping of hazard-prone areas; constructing the history of occurrence of past hazardous events; establishing their frequency and magnitude; predicting the occurrence and location of future events; identifying sources of sediment in catchments; monitoring geomorphological change; and using knowledge of the dynamics of geomorphological processes to advise on appropriate mitigation strategies. In recent years the capabilities of geomorphologists in these roles have increased and the application of geomorphology to the solution of environmental problems has developed. Techniques such as remote sensing
and geographic information systems (GIS), dating (by means, for example, of lichenometry, optical luminescence, and dendrochronology), instrumentation of slopes and other phenomena with data loggers, and computer modeling, have all made major contributions. Geomorphologists are involved with designating and advising on the conservation of geomorphological sites that are of especial value as World Heritage sites, Geoparks, and Geosites. They have also developed a belated interest in the geomorphological consequences of future global changes. Climate change will work in tandem with various anthropogenic activities, including changes in land cover and land use. Geomorphologists have also recognized that submarine geomorphology is an important field, not least for assessing the development of offshore hydrocarbon reserves, as has occurred, for example, on the Nile delta and the Angolan and Congo fans. Landslides, turbidity currents, salt dissolution features, and the like pose problems for pipelines and rigs. Another burgeoning area of international research is into the geomorphology of other bodies in our solar system, such as Mars, Titan, and Venus.

SEE ALSO: Anthropogeomorphology; Applied geomorphology; Biogeomorphology; Fluvial erosional processes and landforms; Geomorphic hazards; Glaciations; Karst processes and landforms; Landforms of other planets; Pre-Quaternary landforms and landscapes; Tectonic geomorphology; Weathering processes and landforms

References


Geophilosophy

Keith Woodward
University of Wisconsin–Madison, USA

Geophilosophy employs a range of spatial concepts to reimagine the discipline of philosophy. Invented by French philosophers Gilles Deleuze (1925–1995) and Félix Guattari (1930–1992), it does not constitute a formal intervention in or contribution to the discipline of geography. Rather, the “geo” prefix offers several challenges to widely held views that philosophy: (i) concerns the historical progress or development of reason and (ii) entails the search for “transcendent” (top-down) truths. Recognizing that the history of philosophy “is marked by detours and contingency,” Deleuze and Guattari (1994, 88) create a system of spatially distributed concepts, each of which is “immanent” to (bottom-up, emergent from) its complex situations (for example, its relations to its own “internal” components and to other “external” concepts). “[T]hinking,” in other words, “takes place in the relationship of territory and the earth” (Deleuze and Guattari 1994, 85). Geophilosophy remains concerned with abstract philosophical ideas (“concepts”), but does not view these as generalizations detached from site-specific, messy materialities. Instead, geophilosophical concepts are “differential”: they emerge with the contingencies of sociospatial differences and encounters, of proximity and entanglement, and of dynamic material relations.

Although geophilosophy is not formally named until What is Philosophy? (1994), it arguably runs throughout Deleuze and Guattari’s collaborations. The geographical lexicon provides Anti-Oedipus (1969) and A Thousand Plateaus (1980) with several key concepts that reconfigure the stakes of philosophy, most notably de/re/territorialization, landscapity, and milieu. Spatial abstractions (lines, planes, smooth and striated space) also play prominent roles in these works. Further, in A Thousand Plateaus, Deleuze and Guattari are careful to identify the practice of mapping (as opposed to tracing) with several elements associated with difference and openness. All of these are implemented primarily to address and transform philosophical problems. Still, geophilosophy’s theoretical innovations have since spurred new developments in geographic theory and practice.

Geography was slow to embrace Deleuze and Guattari during its “poststructuralist” heyday. It was not until the 2000s that they began to figure regularly in critical geographic interventions. One of the benefits of “differential” theories such as geophilosophy is that they lend themselves to seemingly infinite theoretical variations. For this reason, this entry does not exhaust geography’s many interrogations, redeployments, and applications of Deleuze and Guattari. After assessing geophilosophy’s critical tradition and key concepts, it reviews several major incorporations by geography: the role of “minor theory”; the spaces of the body examined by nonrepresentational and affect theory; the “ontological turn” that reexamined
GEOPHILOSOPHY

the nature of space from the perspective of difference; and new geographic approaches to human–environment relations inspired by “new materialism,” “vital materialism,” and the “zones of indiscernibility” between the human and the nonhuman.

Philosophy and geography

The name “geophilosophy” takes its cue from several earlier twentieth century experiments with the prefix “geo” by nongeographic disciplines. In addition to the well-known concept of “geopolitics,” the most prominent among these are French Annales School historian Ferdinand Braudel’s “geohistory” and adaptations of Portuguese psychiatrist François Tosquelles’s milieu therapy known as “geopsychiatry.”

Geohistory concerns broad sociospatial structures (such as the conditions of capitalism) that repeat through everyday social lives in disparate spaces and across long durations of historic time. As Braudel puts it, geography “helps us to rediscover the slow unfolding of structural realities, to see things in the perspective of the very long term,” which he describes as “the almost imperceptible movement of history” (Braudel 1972, 23). This structural perspective guides his history of capitalism, which asks why it emerged successfully in Europe when similar favorable conditions had not engendered it elsewhere during previous centuries. Incorporating social and physical geography helps Braudel avoid teleological narratives in favor of complex spatial conditions that accompany lengthy historical periods. Thus, geohistory uses geography as “no longer an end in itself but a means to an end” (Braudel 1972, 23).

Geopsychiatry adapts sociospatial connections in local communities for psychotherapeutic ends. During World War II, Portuguese psychiatrist François Tosquelles brought “milieu therapy” to the French psychiatric hospital in Saint-Alban en Lozère. He argued that, rather than hiding patients behind the walls of institutions, they should be encouraged to immerse themselves in the surrounding communities. Due in part to the presence of the resistance fighters it sheltered during World War II, the hospital became “an open place integrated into local life and involved in what the Société du Gévaudan called ‘geopsychiatry,’ which coordinated psychiatric and traditional local activities” (Dosse 2010, 42). The French therapist Jean Oury brought geopsychiatric practice from Saint-Alban to the experimental psychiatric clinic La Borde, where it would be enlisted in Guattari’s own therapeutic practice.

“Philosophy is a geophilosophy,” Deleuze and Guattari explain, “in precisely the same way that history is a geohistory from Braudel’s point of view” (Deleuze and Guattari 1994, 95). In beginning to grapple with this proposition, consider Deleuze and Guattari’s rejection of Western philosophy’s historic privileging of reason. In response, they identify philosophy with the supreme geographic concept: contingency. Following Braudel, they describe the birth of philosophy as a lucky accident that arose from many contingent sociogeographic relations. The victory at the Battle of Salamis in 480 BCE, they explain, freed Greece – for the first time – from the constant threat of Persian conquest. The subsequent period of newfound political autonomy and, for a time, cooperation among the city-states of the Delian League introduced transformations to Greek social, cultural, and political life. This “milieu of immanence” (Deleuze and Guattari 1994, 87) was friendly to philosophy, but the Greek sociopolitical and educational environments were hostile to philosophers. Consequently, the pre-Socratics took their education in Egypt and subsequently brought that knowledge back to Greece. Thus
while the Greek deterritorialization (political autonomy and cooperation) fostered new social and political ideals that philosophy would embrace—“immanence, friendship, and opinion” (p. 88)—it was the arrival of minor outsiders (Egyptian philosophers) into that milieu that gave rise to the Western philosophical system. In other words, Deleuze and Guattari suggest, the hallowed questions that have guided Western philosophy’s search for truth may well be the products of historical accidents and contingent geographic proximites.

While the history of philosophy will matter to few geographers, many attach importance to the broader implications of this geophilosophical reading. It suggests that philosophy and social theory are not the products of pure abstractions, historical structures, or “armchair thinking.” Rather, Deleuze and Guattari insist that theory arises from complex, situated, and contingent relations in sociospatial life. This liberates thought and theory from Western pretenses about necessity and truth. Further, it places geography in a pivotal—if differential—position with respect to the creation of theory and the work of theorizing. Geography, Deleuze and Guattari suggest, is more than just “physical and human, but mental, like the landscape. Geography wrests history from the cult of necessity in order to stress the irreducibility of contingency” (1994, 96). In aligning geography with “thought” (normally the domain of philosophy), geophilosophy presents the Earth as a plane upon which concepts and materials arrange themselves, create systems, solidify, create new arrangements, or simply explode.

Geophilosophy

The centrality of contingency likewise plays a role in the way that geophilosophy is performed. When Deleuze and Guattari identify the philosopher’s task with the creation of concepts, they are once again describing a spatial process. Each concept, they argue, consists of multiple internal elements (for example, “geo” and “philosophy” or “geo” and “psychiatry”) that operate in proximate relation to one another and create a novel concept whose meaning, as we saw above, is constituted by those interacting elements. Thus, on the one hand, a concept is a “multiplicity,” a system composed of complex entries, exits, and intersections (much like rhizomes or maps). On the other hand, a concept is a “singularity”: its meaning is specific to its particular constitutive elements.

Concepts also have external relations: they sit in proximity to or distance from other concepts. For example, Deleuze and Guattari do not identify geophilosophy with geohistory, but rather place the former in proximate relation to the latter. Geophilosophy operates in a similar way to Braudel’s concept, but with very different internal elements. Finally, concepts always operate within a field of problems that Deleuze and Guattari call a “plane of immanence”; that is, the “image of thought” (a way of thinking) that serves as the “prephilosophical intuition” (or, nonphilosophical given) upon which each philosopher constructs concepts. Occasionally, a philosopher creates a new problem that gives rise to a new plane of immanence. Descartes’s famous formulation “I think, therefore I am,” for example, created a new way of thinking about the relationship between thought and existence that happened to take for granted (as a prephilosophical intuition) the status of the “I.” Many others would add further concepts to Descartes’s new plane of immanence until Kant introduced a new plane through the creation of the “I” as a philosophical concept (Deleuze and Guattari 1994, 52–56).
A Thousand Plateaus is a treasure trove of concepts and planes (which Deleuze and Guattari sometimes call “plateaus” or “planes of consistency”) drawn from a range of different disciplines – anthropology, history, linguistics, literature, politics, political economy, and sociology, among others. There, the “multiplicity” offers a corrective of sorts to social theory’s tendency to represent the world’s complexity through a relatively small, relatively simple set of structures. Deleuze and Guattari turn to these many disciplines to exhume their differential thinkers and ideas: particularly those who developed strategies for describing multiplicity and variation in opposition to overarching structures. Matching these “minor” thinkers, Deleuze and Guattari develop an approach that discovers small similarities (or “repetitions”) in lines of thought from multiple disciplines. Each chapter in the book proceeds like a transversal line intersecting several other disparate lines. For example, their geometrical concept “numbering number” aligns, among other things, the politics of antistatist social formations with a character’s irregular steps in one of Frank Herbert’s Dune novels and with Julia Kristeva’s feminist treatment of semiotics in art and science. Thus intersecting proximate and contingent fields, disciplines, and problems, A Thousand Plateaus models several strategies and methods for creating concepts.

Geographic concepts guide many of these multidisciplinary encounters. However, contemporary geographers have been better served where they have been cautious with respect to geophilosophy’s geographic elements. For much of the twentieth century, French geography remained intoxicated by the work of Vidal de la Blanche, whose geographical approaches tended to be static and literary (Dosse 1997, 314–315). Although it is not clear what degree of exposure Deleuze and Guattari had to the discipline, it was almost certainly geography in the Vidalian tradition. It was not until the mid-1970s that French political geographer Yves Lacoste founded the critical geography journal Hérodote, whose first issue featured “Questions on Geography,” the famous interview between its editors and Michel Foucault. A Thousand Plateaus’ concept of “landscapity” draws upon “Paysages,” Maurice Ronai’s article on landscape from the same issue of Hérodote. Ronai links landscape to “a very particular semiotic system and very particular apparatuses of power” and argues that landscape is “one of the sources of geography, as well as one of the sources of its political subordination (the landscape as ‘the face of the fatherland or nation’)” (Deleuze and Guattari 1987, 533). However, such engagements with critical geography are rare. The majority of Deleuze and Guattari’s geophilosophical concepts arise from transversal encounters with the more traditional, conservative lines of French regional geography. Among these, their adaptation of “territoriality” (along with deterritorialization and reterritorialization) is the best known. For geophilosophy, this set of spatial concepts describe the process by which gathering of bodies (organisms, materials, or even utterances or concepts) constitute a territory by occupying a space. The concepts are differential because they incorporate the dynamic ways that those spatial relations also “deterritorialize” – depart to “reterritorialize” on something else. Deleuze and Guattari’s great transversal innovation is that this process characterizes not only political geographic relations, but also those of affects on areas of the body, capital flows across the Earth, ideological investments in a social field, and even thought itself. Indeed, they claim that the last of these processes holds revolutionary potential for geophilosophy and fuels the creation of a “new Earth.”
Deleuze and Guattari did not receive broad, sustained critical attention in geography until the 2000s. Throughout much of geography’s “poststructuralist” period, they remained overshadowed by their contemporaries – especially Michel Foucault and Jacques Derrida – due in part to critical geography’s excitement over semiotics and epistemology (the study of knowledge) in the wake of the cultural turn. Added to this, many thinkers across the social sciences and humanities during the last quarter of the twentieth century were deeply critical of the Western ontological tradition. Ontology was dominated by a small group of privileged white males who, despite the narrowness of their demographic, tended to insist upon the universality of their experience and their claims about the nature of existence. Consequently, many critically oriented human geographers rejected ontology in favor of a more dynamic set of epistemological problems that explored the circulation of power relations within language, signs, and discourse. Unlike Foucault, Derrida, and many others, Deleuze remained a metaphysician. While his collaborations with Guattari covered a range of concerns from politics to ethics and semiotics to psychology, they are ultimately contributions to ontology – albeit, an ontology that differs greatly from the Western tradition.

Because Deleuze and Guattari figured only marginally in geography’s early treatments of postmodernism and poststructuralism (1980s–1990s), they were sometimes conflated with the projects of very different thinkers, such as Derrida’s deconstruction. By the mid-1990s, Marcus Doel published a series of papers that singled out Deleuze and Guattari’s project, and celebrated certain thematic elements that he claimed shattered contemporary geographic spatiality – particularly the in-between-ness of “becoming” (as opposed to “Being”) and the inclusivity of the “and” (as opposed to the “or” of dialectics). This offered helpful first steps, but such treatments also deepened misunderstandings by emulating Deleuze and Guattari’s quirky prose and fetishizing the liberatory dimensions of their thought at the cost of ignoring their cautiousness. A more careful and faithful early reading appeared in Cindi Katz’s (Katz 1996) analysis of “minor theory.” If “major” theories tend to deal in titanic dichotomies such as “nature/culture,” minor theory concerns the many entry points that such abstractions overlook. The “major” attempts to identify and capture the entirety of the world within its two categories. The “minor” seeks to describe the rich variety of differences that arise through grounded relations. Appearing at a moment when geographic debates over Marxism and feminism had reached an impasse, Katz’s contribution helped introduce geography to an alternative language that embraces difference as a positive element in political theories and struggles. The appeal of minor theory can still be glimpsed today, for example, in efforts by the Zapatistas and the World Social Forum to create modes of global solidarity that connect disparate social movements.

During the late 1990s, interest in acute variations in bodily affections and perceptions, along with the openness of the body to its outsides – the geophilosophy of the body, if you will – gave rise to “nonrepresentational theory.” This area, spearheaded by Nigel Thrift’s explorations of the embodiment and affect, studies of bodily relations that occur without (or before) presenting themselves to conscious, intentional thought. It draws extensively upon Deleuze and Guattari by challenging the sovereignty of subjectivity and emphasizing instead bodies’ constitutive engagements with externalities and environments. Nonrepresentational geographers such as Mitch Rose and John Wylie, for example, employed
GEOPHILOSOPHY

geraphical problems to build rich studies
that re-center landscape as a formative element
in embodiment, subjectivity, and politics.

In 1996, Jane Jacobs wrote a brief introductory
commentary in *Environment and Planning D*
that gathered many of Deleuze and Guattari’s
geographical borrowings under the umbrella
title “geophilosophy.” Nearly a decade later, in
2004, Mark Bonta and John Protevi’s *Deleuze
e Geophilosophy* offered an extensive glossary
of Deleuzoguattarian concepts drawn from
geographic discourse. Inspired by the work of
philosopher Manuel DeLanda, who situates
Deleuze and Guattari’s project in the context of
science, Bonta and Protevi’s treatment empha-
sizes the real and potential relationships between
concepts and the complex environments and
ecosystems. *Deleuze and Geophilosophy’s* focus
upon natural systems and physical science antici-
\(\ldots\)pated geography’s “ontological turn” during
the mid-2000s – the period during which
Deleuze and Guattari achieved greater critical
prominence. During this time, for example,
Sallie Marston, John Paul Jones III, and Keith
Woodward’s (2005) rejection of geography’s
epistemological darling, scale, in favor of “site
ontology” – immanent, site-specific accounts
of complex spaces – was largely inspired by the
“differential ontology” of Deleuze and Guattari.

By the late 2000s, geographers increasingly
invoked geophilosophy to articulate spatial com-
plexity in a variety of contexts. In his critique
of capital and agrarian development in India, for
example, Vinay Gidwani (2008) contrasts the
political complexities of geophilosophy with the
almost Manichean dichotomies of traditional
development discourse. Yet, Gidwani avoids
privileging minor or major tendencies, recog-
nizing that each works off the other, producing
complex situations on the ground. Thus, for
example, the limited connections between privi-
leged and marginal social bodies and the rigid
protocols that further curtail their interaction can
be momentarily overturned by the contingencies
of policy decisions. Development becomes “an
abstract machine that reorganizes the condi-
tions – or ecology – of human life for its better-
\(\ldots\)ment” (Gidwani 2008, 70) but that remains
problematically inconsistent. Other thinkers bring
geophilosophy to bear upon complexities engen-
dered by social difference and localized bodily
movement. Arun Saldanha (2007), for example,
examines the arrangements, arrivals, and depar-
tures of different racialized groups within
shared public spaces (de/re/territorialization) as
dynamic components in localized dimensions of
sociospatial viscosity.

Extending beyond the sociospatial dimensions
of Deleuzoguattarian approaches to geography,
the early 2010s have been characterized by an
interrogation of “new materialism” or “vital
materialism,” which draws upon the ontolog-
ical turn to trouble the difference in lingering
dichotomies such as human/nonhuman and
life/environment. For thinkers, such as Bruce
Braun (2015), this has meant wrestling with
the entanglements between differential, non-
linear treatments of nature and neoliberalism’s
efforts to capture these emerging perspectives
in its governance apparatus. Others employ the
Deleuzoguattarian concept “assemblages” (or
*agencements*: a collection of disparate materials
that displays active organizational characteris-
tics) to describe complex aggregates of organic
and inorganic material. This has inspired some
geographers to consider socioenvironmental
implications of, for instance, corporate and
governmental experiments in “biomimicry”
(the robotics engineering that mimics biological
organisms). Elizabeth Johnson (2010) explores
techo-science’s current efforts to “‘reverse engi-
\(\ldots\)neer …’ the observable behaviors expressed in
\(\ldots\)ological life” (2010, 179) and their implications
for new regimes of technology, environments,
and “biomimetic capitalism.” Johnson is wary to make too much of what appear to be strong Deleuzian currents of affectivity in such experimentation. Suffice it to say that biomimetic techno-science involving in/organic physiological movement and its potential capitalization intersects multiple lines from diverse planes (biological, technical, social). Finally, projects such as Johnson’s introduce important questions about the potential gap between, on the one hand, geophilosophy’s countless exciting theoretical possibilities and, on the other, its evacuation of many of the moral and political guardrails found in traditional social theory.

A new Earth?

This is not to say that geophilosophy does not offer nuanced approaches for engaging political questions of oppression and revolt. However, these have definite limits. Because it is grounded in contingency and immanence, geophilosophy is incapable of offering prescriptive transcendent or general statements: “Not being a power, philosophy can’t battle with the powers that be, but it fights a war without battles, a guerrilla campaign against them” (Deleuze 1995, vii). What does such a campaign mean in an era of “global,” deterritorializing capital and environmental crisis? Near the close of each of their major collaborations, Deleuze and Guattari describe geophilosophy as a system capable of thinking a “new Earth” (Deleuze and Guattari 1987, 510). This is not an appeal to utopianism, a concept that remains subject to history (Deleuze and Guattari 1994, 110). Rather, their new Earth is associated with “absolute deterritorializations” and “becomings” that are “more geographical than historical” (p. 110). In short, a new Earth is rendered possible by the ability to experiment with new modes of thinking and to invent a “new people” (p. 101).

Geophilosophy offers, in other words, a revolutionary dimension of a philosophy capable of thinking in terms of contingency and difference: a differential political thinking.

SEE ALSO: Affect; Annales School; Difference; Nature; Nonrepresentational theory; Ontology: theoretical perspectives; Poststructuralism/poststructural geographies; Spatial concepts; Subjectivity; Territory and territoriality

References


GEOPHILOSOPHY


Further reading

In 1943, German sociologist, Werner Cahnman (1943, 56), offered a definitive conspectus of geopolitics:

The features of significant parts of the earth’s surface as explored by geographers provide Geopolitics with a framework into which political events must fit themselves if they are to lead to success in the long run. To be sure, the actors upon the political scene are free to depart from such a framework now and then, but the close relationship of political events to the earth’s surface will certainly make itself felt sooner or later. (Cahnman 1943, 56; emphasis by entry author)

According to Cahnman, the influence of environment over the actions of humans is considered to be a natural rather than a historical relationship in geopolitical thought and practice. Essentially, Cahnman is engaging in a form of environmental determinism insofar as he upholds that the environment, conceptualized as the “natural world” that exists “out there,” is the primary cause or determining factor in geopolitical interactions and activities.

The intricate relationship between geopolitics and the (physical) environment, as highlighted by Cahnman, is not an intuitive one. In the field of international relations (IR), for example, it has been argued that political realism (often simply known as realism), which focuses on the military security of states to the neglect of other factors (including environmental ones) in shaping world politics, has been the dominant approach/perspective (Walker 1993). Realism is often associated with the “Westphalian model” of world politics that involves a world allegedly characterized by the territorial sovereignty of states, an anarchic international arena, legal and political equality between states, the inherent right of states to use force to settle disputes, and limited interstate cooperation. For many IR realist scholars, political life embodies a fundamental contrast between life inside and life outside of the state. Within the state, it is possible to live the “good life” and to become part of society characterized by citizen, community, and culture. Outside the state, the notion of an international community of people is effectively abandoned and replaced by an interpretation that judges IR to be populated by self-interested states motivated by their own agendas rather than collective improvement. The principal political force here remains the nation-state, with war, violence, and uncertainty constituting global norms as struggles for power ensue among different state actors.

Those who believe that the Westphalian model does not adequately capture the complexities of modern political life have, however, hotly disputed realist views of politics. One line of criticism is leveled at the approach favored by IR realist scholars, who seem to be more concerned with the role and behavior of states, to the detriment of other agents such as firms, non-governmental organizations, and international
bodies, whose contributions are often significant in shaping distinctive political agendas. But, more crucially, there have been serious concerns raised about the tendency of realists working in the IR tradition to view the international system as an anarchical domain occupied by states (regardless of their social and cultural backgrounds) that function in an undifferentiated manner. According to Singer (1981), interactions between states do not operate in a vacuum and IR research will be severely impoverished if the role of the environment is being ignored. In his words, “most of the variations in the behavior of nations in conflict will be accounted for, not by their internal attributes, but by variation in their environment” (Singer 1981, 7; emphasis by entry author). Interestingly, Singer’s notion of a geographically sensitive analysis of IR can be located in the early works by (classical) geopolitical writers. Traditional geopolitics has similarly been underwritten by many of the assumptions of realism concerning the nature of the international arena and the significance of state sovereignty and national interests. However, in contrast to realist conceptions of world politics, political geography and geopolitics have focused on the power of the (physical) environment (e.g., land and sea) in shaping IR. Classical geopolitical scholars such as Mackinder (1904) endowed the “Heartland” of Eurasia with the potential to influence world politics at the expense of the so-called rimlands and surrounding outer crescents (Figure 1). Geographical features, characteristics, and divisions in this sense were considered timeless, and thus immune to human alteration. Geography was, thereby, reduced to the role of simply providing a territorial stage on which the interactions of states unfolded.

Mackinder’s work, alongside other classical geopolitical writings, is representative of what O’Lear (2013, 305) terms as “realist-inspired geopolitics of the environment.” What this essentially means is that an objective characterization of the world’s natural environmental features underpins the study of statecraft and the divination of patterns of global politics. In such a formulation, geopolitics has an intellectual value given its capacity to uncover the environmental challenges facing different states so that their vital interests can be protected through force if necessary. But as this entry will illuminate, the rise of critical approaches in geopolitics has raised important questions about the problematic assumptions that belie the seemingly neutral relationship drawn between the environment and military geostrategizing at a global level. Specifically, by adopting a critical geopolitical perspective in looking at how two contemporary environmental concerns – climate change and resource scarcity – are bounded up with geopolitical calculations and praxis, there is a wish to underscore the ways in which power is fully implicated in environmental overtures of geopolitics. This entry simultaneously embodies an effort to go beyond criticism in order to rethink environmental geopolitics for ushering in modes of living that are not only socially just but also ecologically less destructive.

Critical geopolitics, power, and the environment

As a broad set of writings that first emerged in the early 1990s, critical geopolitics came under the radar of the geographical discipline through the seminal works of O’Tuathail and Agnew (1992) and O’Tuathail (1996). This pioneering research quickly proliferated into an extensive array of publications that helped to shape the intellectual boundaries and trajectories of the now established field of critical geopolitics. Influenced by poststructuralist philosophies
particularly emanating from Michel Foucault and Jacques Derrida’s works, the key concern for critical geopolitics is to expose and challenge the various mappings of dangers, risks, and threats used to legitimate political power. If such a perspective is applied to look at the interface between geopolitics and the environment, epistemological pathways are opened up to “decode narratives about how environmental features pose risks to society and how environmental resources empower society and narratives about dangerous or desired human-environment interactions” (O’Lear 2013, 305). This is reminiscent of Castree’s reminder that although environmental problems may be real, there is no one singular, “best” solution to tackle such problems. Instead, he urges us to examine the geopolitics behind representations of environmental problems – the dominant discourses that define and establish the norms for managing environmental challenges – so as to “contest existing geopolitical arrangements, uncovering power relations inherent in them and the possibilities for more just interstate relations” (Castree 2003, 437; original emphasis). In other words, it is imperative to delve deeper into whose perspectives on environmental geopolitics are being privileged and whose interests are being marginalized in the process.

This is not to suggest that classical geopolitical engagements with the environment are no longer relevant and are replaced by these critical theoretical impetuses. Contrary to that, the project of critical geopolitics exists precisely to cast watchful scrutiny on the continued proliferation of classical variants of geopolitics with their echoes of environmental determinism. The latest revival of classical geopolitics is what Megoran (2010, 187) calls “neoclassical geopolitics,” which refers...
to “ways of thinking about the effects of geography on IR that explicitly locate themselves within the Mackinder–Haushofer–Spykman tradition, but which creatively rework it with reference to changed social, economic, political and cultural factors.” One widely cited example of neoclassical geopolitics is Robert Kaplan’s 2012 book, *The Revenge of Geography*, where he adapts Mackinder’s compelling slogan “Man [sic] and not nature initiates, but nature in large measure controls” to account for geopolitical patterns and developments in current times. Geography, according to Kaplan, remains the biggest shaper of geopolitical forces and the map will provide good indications of where the major geopolitical fault lines lie. Within this formulation, Kaplan (as with Mackinder) uses “geography” and “the map” as synonymous with the physical environment. He goes on to gather up select trends and presents them as ominous threats to what is implicitly held to be the benign world order secured by American military power. In particular, he points out that population growth threatens to fill out the empty spaces on the map and create a conflict over scarce resources as states compete for territory, food, and water. These compounded problems of economic destitution, environmental degradation, and violence would be most gravely felt in the poorer parts of the world, which will turn them against the economically advanced nations such as the United States. Kaplan’s message is thus clear: America needs to recognize that these circumstances have destabilizing effects to world order and should thus take proactive steps to alleviate such threats.

In contrast, geographers such as Simon Dalby, Juliet Fall, and Gerald Toal have engaged in critical geopolitical readings of Kaplan’s work. They argue that Kaplan’s thesis hinges on sweeping generalizations and is limited in explaining the actual situation on the ground. Dalby (2007) highlights how Kaplan conveniently ignores larger structural factors, such as the geohistorical context of violence and environmental degradation (e.g., colonial politics and neocolonial arrangements), in shaping geopolitical outcomes of those “less developed” and “threatening” places that he talks about. Toal, on the other hand, points out that by tenuously linking environmental scarcity with geopolitical violence, Kaplan’s agenda appears to be closely tied to America’s hegemonic interests on the world stage. In sum, these disparaging voices draw on a critical geopolitical perspective to question a realist geopolitical representation of the environment.

This example clearly exemplifies how different geopolitical traditions have diverging conceptions of the environment. While classical geopolitics equates the environment with the “natural” world that has circumscribing effects on human behaviors (see entry on Nature), critical geopolitics has refuted the idea of an ontologically separate environment that is outside of human existence by focusing on how the environment is represented, used, and impacted by different actors in the pursuit of specific geopolitical agendas. In recognizing how human activities and wellbeing are entangled and intertwined with the environment in multiple and complex ways, critical geopolitics share similar concerns with recent discourses about the Anthropocene. Indeed, there is increasing consensus that human activities and technologies have produced alterations to the biosphere at an unprecedented scale, prompting some Earth system scientists to conclude that a new geopolitical era known as the Anthropocene is being ushered in (Crutzen and Stoermer 2000). Although placing humans at the heart of environmental change, the Anthropocene does not, however, imply that everyone is equally exposed to both environmental benefits and impacts. As Dalby (2007) has pointed out, acknowledgments of
the Anthropocene must be accompanied by critical investigations into the power geometries that have led to disparities between groups of people and between the environments they inhabit. In other words, attuned attention must be bequeathed to understanding the dominant economic and political structures that manipulate environmental systems to create staggering gaps between “haves” and “have-nots” in terms of economic wealth and environmental health (O’Lear 2013). Such is the role of critical geopolitics in the epoch of the Anthropocene, to question such inequalities and rethink our collective role and responsibility as humans become key agents of geological change. This critical agenda, which Dalby (2007, 2013) has championed for as “Anthropocene geopolitics,” will be adopted in the following sections to investigate two environmental concerns that have been the subject of intense debates and interest in recent years: climate change and resource scarcity.

The critical geopolitics of climate change

On May 10, 2013, the New York Times announced that “the level of the most important heat-trapping gas in the atmosphere, carbon dioxide, has passed a long-feared milestone” (Gillis 2013). Reaching a daily level of 400 parts per million (ppm), atmospheric concentrations of CO₂, scientists estimate, are higher than at any time in human history since the Pliocene. This report is a microcosm of the larger concerns within the policy and academic arena about the significant repercussions to human societies brought about by anthropogenic climate change. Indeed, climate change is not only the issue at stake here, but it is also a symbolic marker of the larger transformations underway, processes that now frequently invoke discussions in terms of the putative new geological epoch of the Anthropocene.

In lieu of the recognition that fossil-fueled industrial capitalist humanity is driving climate change in era of the Anthropocene, responses have thus far been state-centered, relying on the actions of different national governments in reducing the emission of greenhouse gases. For instance, the Kyoto Protocol under the United Nations Framework Convention on Climate Change aims to control the amount of carbon emissions by slotting states into the categories of Annex 1 or Annex 2, depending on their contributions and capacities to reduce emissions. Other international meetings/conventions, such as the World Climate conferences, also involve state delegates as key participants, which once again imply the legitimacy of state actors in deciding on agreements to govern the world’s climate future. What this means is that states have the absolute authority in formulating and implementing domestic laws and regulations regarding emissions levels to bring about the greater good of positive environmental change at the global scale.

But criticisms abound of these seemingly benign state initiatives to alleviate the impacts of climate change. O’Lear (2013) explains that many of such criticisms are concerned with the “territorial trap” fallacy that is inherent in statist accounts of climate change. In privileging the state’s role in combating climate change, critics have asserted that there is a willful ignorance of the ways problems and solutions associated with climate change cut across national boundaries and cannot be contained by states working independently of one another. Furthermore, it has been argued that there are unequal power relations among states and, by maintaining a state-level focus, those powerful countries can then have the liberty to decide on the “best” approach to climate change that suits their
interests, often resulting in the marginalization of “less developed” nations. For instance, Paterson and Strivelope (2007) have examined state-centered articulations of climate change. They uphold that by naturalizing an understanding of climate change based on territorial states, certain perspectives and solutions appear legitimate. Citing the case of ex-US President Bush’s refusal to participate in the 1997 Kyoto Protocol that aims to reduce greenhouse gases on a country-by-country basis, Paterson and Strivelope show how the United States appealed to the notion of parity to defend America’s action. Bush’s reasoning was that since developing countries were not required to cut their emissions as drastically as their richer counterparts, they would have an unfair economic advantage and threaten US security. Yet, as others have pointed out, it is only just that the United States takes on the bulk of the responsibility given that it is the world’s biggest producer of greenhouse gases, contributing 22% of total emissions. Dalby (2013) takes further issue with what he sees as the increasing securitization of climate change. He notes that the formal military apparatuses of a number of “Northern” states are beginning to take concrete steps to “secure” their states against the risks of climate change. These include discussions of climate and security in key military planning documents of states such as the United States and United Kingdom, and scenario exercises on extreme contingencies in these same states are beginning to incorporate environmental dimensions. Dalby is, however, highly critical of climate change functioning as a factor in states’ calculations of grand strategy. He warns that such negative formulations of security will ultimately lead to spatial exclusion — discourses related to the necessity of (military) pre-emption of climate threats by powerful states are foreboding signs of the global extension of imperial power, domination, and violence. Besides highlighting the negativities arising from the securitization of climate change, critical voices have also emerged to contest the increasing marketization of climate change as part of states’ risk adaptation strategy. There is a keen interest in drawing attention to the undesirable consequences as a result of climate change solutions being embedded within circuits of capital. For instance, Elliott (2013) focuses on the case of the United Nations reducing emissions from deforestation and forest degradation (REDD+) program, which essentially embodies the view that countries, companies, communities, or individuals could be paid to manage forests and reduce deforestation in recognition of the role forestry has in carbon sequestration and the significance of land-use change in annual carbon emissions on a global scale. In this sense, REDD+ acknowledges that powerful economic imperatives are needed to discourage forest clearance and envisages a North–South flow of finances in order to fund its multiple initiatives. Elliott (2013), however, points out such a top-down policy driven primarily by global, Northern agencies and governments could have serious negative impacts on forest-dependent communities and indigenous peoples in the Global South, whose cultures and livelihoods are closely linked to those forests. There is also an ethical question to be raised with regard to carbon markets, whereby state (or corporate) interests pay for carbon credits to be used within schemes such as REDD+ while continuing to pollute and extract fossil fuels elsewhere (potentially at a cost to other people). More fundamentally, Elliott argues that REDD+ does not address the underlying drivers of deforestation (e.g., reducing demand for agricultural or timber products). Hence, as illustrated from the example of REDD+, mitigating climate change through capitalist logics and rationalities raises concerns about diverting attention away from
tackling the root causes of climate change while, at the same time, perpetuating the pattern of dispossession for marginalized nation-states and their communities.

The critical geopolitics of resource scarcity

A second environmental issue that has been given heightened prominence in the so-called era of the Anthropocene is resource scarcity (see entry on Natural resources). Rees (2013), a scholar with the Cambridge University’s Centre for the Study of Existential Risks, has crucially argued that climate change and resource scarcity are two of the biggest ecological shocks that humans’ incessant demands have inflicted on the biosphere. It is hardly surprising that such a perspective exists, given the long-standing worry that accelerated population growth and the advent of fossil-fueled capitalist societies have put immense stress on the planet’s resources.

The rising concern of resource scarcity in the Anthropocene has often been bounded up with discussions of “environmental security” (Dalby 2007). The establishment of such a relationship is not something novel given that, back in the 1980s, Cold War environmental security discourse (i.e., instabilities and disruptions to society as a consequence of environmental problems) was already a dominant theme. The assumption here is that resource scarcity might cause conflict and, on the largest scale, warfare between states. According to these advocates of the scarce resource wars hypothesis, people or nations will fight one another so as to secure access to resources necessary for their survival. The underlying implication is that the more scarce the resource, the more intense the conflicts. While there has been anecdotal evidence advanced to corroborate this view, several counterarguments exist to question whether the scarce resource war perspective can indeed be generalizable. First, resource scarcity and population pressures can lead to socioeconomic innovation, including a diversification of the economy, which often results in a more equitable distribution of power across society. Second, the scarce resource wars theory fails to take into account the possibility of international trade and market mechanisms in counterbalancing localized scarcities and shifts in resources. Finally, resource-poor countries can also invest in human capital, rather than simply protecting and relying on existing resources. In such cases, “the likelihood of violent conflict decreases as human capital develops (e.g., through education, trading, and manufacturing skills) and the economy diversifies, and governance becomes more representative and accountable” (Le Billon 2001, 564). Japan and Singapore are two classic examples that are said to embody the essence of these counterarguments— their resource-impoverished condition did not prevent the peaceful development of these two economically high-achieving states. Hence, in sum, there are limitations to the resource scarcity perspective, especially with regard to its inability to explain why a scarcity of valuable resources is not a necessary or sufficient precondition for conflicts/warfare.

In lieu of the aforementioned weaknesses, scholars working in the critical geopolitics tradition have proposed alternative frameworks to create more nuanced understandings of the “problem” of resource scarcity (Le Billon 2001; 2013). These works do not deny that natural resources such as timber, water, and diamonds may indeed shape how and where conflict unfolds. However, they are concerned that accounts related to resource conflicts often imbue natural resources with an agency that they do not actually have. They are also united in their view that context is important in teasing
out the diverse range of processes and dynamics at work, rather than assuming a straightforward relationship between resource scarcity and conflicts. Le Billon (2001; 2013) is a good reference point for this body of literature. Drawing inspiration from Zimmerman, who argues for the need to look at Earth’s natural endowments as socially constructed, Le Billon brings to the fore that, whether or not nature is transformed into a resource, it is related to human desires, needs, and perspectives. He goes on to suggest that resources and a place’s vulnerability to conflicts are intimately connected to different factors such as their susceptibility to being looted and the geographical concentration/location of resources, whether the resource in question involves production or extraction, and the extent to which a country’s resources are heavily tied to the global economy through colonial powers, private/commercial interests, and domestic elites. It is hard to do full justice to Le Billon’s nuanced expositions here but, in general, the more dependent a country is on its resource sales to the global markets, the more opportunities there are for different forces to manipulate such lucrative networks for their own benefits, thereby increasing the risks of violence. Additionally, if the resource involves extraction, violence is most likely to take the form of territorial or state control. Lastly, the more accessible a resource is to various parties (i.e., its susceptibility to being looted), coupled with its concentration in a certain geographical area, enhances a place’s susceptibility to conflictual relations.

With similarities to Le Billon’s arguments, Dalby (2003) offers a critique of the ill-defined nature of the scales and contexts through which the complex geopolitics of resources has been analyzed. To quote him at length:

Resources are not just resources; renewables and non-renewables have different qualities. Contexts matter too; the exploitation of oil reserves is different in many ways from use of marginal farmland. Subsistence activities are likely to have different social dynamics from the large-scale commercial exploitation of mineral materials. (Dalby 2003, 5075)

Hence, according to Dalby, the different scales and contexts of resources need to be clearly specified, especially given that these geographical specifications of security, when left unquestioned, often aid in the production of narratives of resource scarcity to form cartographies of risk and danger in the interest of promoting political agendas. A clear example of this is the ways in which some states in the Global North have drawn on fears of conflict over scarce resources to formulate interventionist policies. US foreign policy, in particular, has long focused on maintaining the “stability” of authoritarian regimes in the wider interests of regional security. The task of critical geopolitics, then, as Dalby (2003) upholds, is to expose the geographical assumptions that are inherent in abstract generalizations of resource conflicts, thereby providing possibilities for thinking differently about environmental threats and insecurities.

Towards “planetary” thinking?

The discussion thus far has centered on harnessing a critical geopolitical perspective to look at issues to do with the environment. Regardless of the different conceptions of the environment (from thinking of it as the natural world to seeing it as part of human systems in the epoch of the Anthropocene), this entry contends that critical geopolitics has enabled the consideration of the role of power in human–environmental relationships. More specifically, critical geopolitics contributes to the challenging of the specifications of dangers, threats, and risks that underpin the representation of the environment for certain
political agendas/interests. But following recent calls for critical geopolitics to go beyond critique and provide alternative ways forward for the betterment of society, this entry concludes by delving briefly into recent “planetary” thinking in order to reflect on more socially and ecologically just geopolitical responses to the environment in the current era of the Anthropocene. According to Spivak (2003, 72), the “planet is the species of alterity, belonging to another system; and yet we inhabit it, on loan.” If this perspective is linked to the environment, it can similarly be argued that through our collective global agency of human impact on the environment, we are all united by our identity as planetary subjects. We are part of a biosphere that we are actively remaking, rather than separate species on a planet or separate nations competing for space. The geopolitics of this is especially important: interconnections and cooperative ventures are key to governing the Earth system (Dalby 2013). In this sense, the Anthropocene is not the terminal phase, an apocalyptic end to all things as some have suggested. Rather, it is the next phase that requires collective effort as planetary beings to find ways/avenues to evolve, adapt, and change in response to the environment (Dalby 2013).

SEE ALSO: Natural resources; Nature

References

GEOPOLITICS OF THE ENVIRONMENT


Geopolitics

Paul Reuber

University of Münster, Germany

Geopolitics is a practice of knowledge production that focuses on linguistic, cartographic, and pictorial constructions of geopolitical representations at regional and global scales. “Irrespective of whether the word geopolitics is used or not, the conventional understanding today is that geopolitics is discourse about world politics, with a particular emphasis on state competition and the geographical dimensions of power,” Ó Tuathail (2006/1998, 1) has observed. Since its biodeterminist and geodeterminist framing in the late nineteenth and early twentieth centuries, the term “geopolitics” has expanded to include a wide range of explanations and reputations (Dodds, Kuus, and Sharp 2013). Therefore, an exact definition is difficult, if not impossible, because of its semantic ambiguity and – given the variety of approaches and conceptual positionings – its embodiment of contrasting and contradictory meanings (Reuber 2009). Geopolitics deals with:

- the geopolitical representations of the “self” and the “other” in politics, media, and science;
- the geopolitical discourse itself as a linguistic formation of space and power;
- the “doing geopolitics” of political actors who draw on hegemonic geopolitical imaginations in order to legitimize their political reasoning and practices;
- the academic deconstruction of geopolitical discourses and practices from a critical geopolitical perspective, or rather a post-structuralist political geography.

These four facets not only illustrate the wide range of perspectives represented by the term “geopolitics”; they identify it – following the conceptualizations of Laclau and Mouffe’s (1985) discourse theory – as one of the most powerful “empty signifiers” of the political geographies of modernity. With regard to the content, or meaning, associated with particular signifiers, the term “empty” does not imply a sense of meaninglessness – in fact, quite the opposite. Through its overdeterminism, it acquires and balances a variety of meanings within different fields of reference. Thus, “geopolitics” assumes an indistinctness which gives it a propensity for immersion within various segments of society, ranging from politics – through political think tanks and policy-related disciplines – to the media, and even bar room clichés. In short, “Geopolitics … is an essential part of everyday life” (Dodds 2007, 21).

Within the academic community, this plurality is reflected through a considerable heterogeneity of underlying approaches, which may move toward quite different forms of theoretical concepts and empirical analysis. What scholars from the fields of international relations and political geography publish today under the label “geopolitics” can range from implicit natural
determinist contributions to practice-oriented forms of spatial analysis, and even on to structuralistic and poststructuralistic approaches. Therefore, a critical reading of related publications is strongly recommended. To organize the field, it may be helpful to point out the deep epistemological gap which separates two opposing approaches.

1 A traditional, “realist” form of geopolitics attempts to identify the “natural” and “cultural” driving forces of geopolitical developments at regional and global scales. It portrays geopolitics as a “reliable guide of global landscape using geographical descriptions, metaphors, and templates such as ‘iron curtain,’ ‘Third World,’ and/or ‘rogue state.’ Each of these terms is inherently geographical because places are identified and labeled as such. It then helps to generate a simple model of the world, which can be used to advise and inform foreign and security policy making” (Dodds 2007, 4). The examples mentioned earlier fall into this category – especially those actively engaging in the construction of geopolitical representations of the “self” and the “other.” In the event that they develop into hegemonic concepts, they could be immediately relevant, for instance, within the frameworks of geopolitical identities, political practices, armed conflicts, and so on.

2 A critical, constructivist form of geopolitics (critical geopolitics and poststructuralist approaches) recognizes geopolitical imaginations and their putatively “natural” geographies of “us” and “them” as discursive formations, or rather, as ever-changing constructions of global spatial orders of power on the world political map. Such concepts for analyzing geopolitical representations are discussed in further detail in other entries in this encyclopedia. Each of these contributions reconstructs historic geopolitical narratives by drawing on the theoretical framework of constructivist approaches, especially discourse analysis and semiotic analysis. “So rather than simply assume that labels such as ‘iron curtain’ and ‘axis of evil’ have a certain heuristic value, we proceed to question how they generate particular understandings of places, communities and accompanying identities” (Dodds 2007, 5). The deconstruction of geopolitical representations, in particular, argues against the re-emergence of strongly traditional (and, sometimes, natural deterministic) ideas for which geopolitics is the doctrine of the influence of physical space on the politics of a state. Constructivist reasoning in political geography indicates how such a conceptualization of geopolitics leads to an ideological, as well as an epistemological, cul-de-sac. They take the geopolitical discourses themselves as the objects of their analysis. The issues to address from this point of view are the construction of discursive and cartographic representations, and how these representations intersect with other political imaginations and practices.

Even though variant 1 provides the framework for a wide array of “strategic studies” located, today, within a gray zone between applied political science and political consulting, from a discourse analytical perspective these analyses are not considered to be “real” depictions of (geo)political constellations of power. Rather, they are a form of “doing geopolitics” that legitimizes political scenarios and practices through the postulation of innate geopolitical regionalizations and antagonisms, while supporting them with ostensibly “objective” academic arguments.
Box 1 The nation-state as the fundamental “territorial trap” of the geopolitical imagination

Since the modern period, the nation-state was considered to be the territorial framework for countless geopolitical strategies of containment and exclusion. The powerful social construction of nation-states as “imagined communities” (Anderson 1983) has made them persistently successful as a kind of “natural” entity through which struggles over political identity, sovereignty, and power have been organized over the last two centuries. Even in recent times of increasing globalization, the role of the nation-state appears to have held its ground despite the undeniable loss of power to transnational networks of “global governance.” John Agnew (1994) calls the rigid fixation on the spatial container of the nation-state a “territorial trap.” This expression characterizes the virtual ontologization of the territorial principle as a crucial framing of social, and especially political, structuring. According to Agnew, the territorial trap is based on three main assumptions that must be analytically separated, but that overlap in social practice. The territorial “logic” of modernity is based on the premises that (i) the sovereignty of the modern state requires a territorial organization of power; (ii) internal and foreign policy are strictly separated; and (iii) the territorial state serves as a kind of spatial container of national societies.

(cf., e.g., Samuel Huntington’s clash of civilizations approach and Thomas Barnett’s geopolitical construction of core and gap regions).

In this mode of representation, the nexus between society and space becomes the leitmotif of the geopolitical imagination. Framing social differences in terms of spatial representations leads to a “purification of space” (Sibley 1988). By this kind of discursive practice, the “self” and the “other” are organized along spatial categories and ascribed with boundaries (e.g., the nation-state, see Box 1). Such constructions can be found as powerful geopolitical representations in all historical epochs (see next section). Their logics and geographies may vary according to the hegemonic discourses of their time (e.g., if they exploit expressions such as race, class, or cultural diversity when constructing putative homogeneous spaces), but their basic logic of construction stays the same. Based on these assumptions, geopolitical discourses can be regarded in a Foucauldian sense as politically contested formations of power and space. They serve as historically alterable narratives and, apart from stable phases in which they may appear to be virtually objective formations, are subject to constant negotiation and renegotiation.

Geopolitics and geopolitical representations in different historical periods

Following the arguments outlined, it is apparent that geopolitics and its geopolitical imaginations are situated within their social and/or discursive contexts. This means that geopolitical discourse is subject to change. In order to illustrate this more clearly, concise – albeit
Box 2 The state as an organism: Ratzel’s biodeterministic construction of German geopolitics

The basic element of Ratzel’s mode of construction consisted of the nation-state’s dependency on the biological dynamics of natural selection. Friedrich Ratzel (1844–1904), who had shifted his scientific career from zoology to geography, built his political geography on a positivistic, natural-scientific conception of the world. His construction of the relationship between politics and space is primarily determined by the physical-geographical characteristics of the nation-state. Another crucial point in his construction is the biological framing of the nation-state as an “organism.” The nation-state, in this respect, is ascribed the traits of a living being and is considered to be “healthy” and “strong” only when capable of growth or, in other words, territorial expansion. Thus, this Darwinian representation of geopolitics is used to legitimate imperialism and expansionism. Following Ratzel’s reasoning, as its civilization continues to develop, the nation-state needs more and more space in order to sustain a healthy, growing population. In this manner, the struggle for more space automatically leads to conflicts between nation-states.

Throughout his writings, Ratzel (repeatedly) makes a close connection between population growth and territorial expansion, putting this claim into practice with regard to the German empire as it approached World War I:

Wherever you look space is won and space is lost. Setback and progress everywhere; there will always be reigning and serving peoples. Even the peoples must be anvil and hammer. In no case Germany must limit itself to Europe; as a world power among world powers it can only hope to protect the ground for its people that it needs to grow.

(Ratzel 1906, 377; author’s translation)

In this manner and through supposed “scientific” arguments, Ratzel provided the geopolitical basis for the German empire’s colonial policy and Flottenpolitik. Even after World War I, the connection between politics and scientific geopolitical representations led to a significant expansion of political geography at the universities. With his discursive logic of growing space, Ratzel laid the foundation for the development of the ideology of living space (Lebensraum).

exemplary – accounts of geopolitical framings throughout different historical eras are discussed in the following sections: (i) geopolitics in the historical context of Darwinism, nationalism, imperialism, and colonialism; (ii) geopolitics in the broader historical context of National Socialism; (iii) geopolitical representations during the Cold War; and (iv) new geopolitical representations after the end of the Cold War.

“Classical” geopolitics: geopolitics in the historical context of social Darwinism, nationalism, imperialism, and colonialism

The search for interrelationships between space and politics is much older than the term “geopolitics.” Ó Tuathail (2006/1998, 1) explains how the birth of classical geopolitics dates back to the ages of imperialism and colonialism which developed and peaked toward the
end of the nineteenth century: “All words have histories and geographies and the term ‘geopolitics’ is no exception. Coined by a Swedish political scientist named Rudolf Kjellén, the word ‘geopolitics’ had a twentieth century history that was intimately connected with the belligerent dramas of that century” (Ó Tuathail 2006/1998, 1). Behind this terminology resides much of the conceptual framework of classical geopolitics developed by Friedrich Ratzel; the central component of his approach to geopolitics was a biological reconceptualization of the nation-state (Staatsorganismus). Ratzel, considered by many today to be the founder of political geography, infused some of the popular biological theories of his time with concepts of “the social” and “the political” (see Box 2). Thus, his organic state theory provided a discursive formation rooted in environmental determinism that would serve as the basis for subsequent theoretical developments within international geopolitics.

Historical reconstructions help to clarify how these theses, which today sound rather absurd and illogical, could have been accepted during their time as plausible rationalizations. This was primarily the result of the development of classical geopolitics into a hegemonic discursive field including several popular “isms” characterizing the nineteenth century such as natural determinism, environmental determinism, (social) Darwinism, nationalism, imperialism, and colonialism. In this powerful mode of construction, geopolitics appears as an expression of (assumed) natural spatial order. These particular discourses can be found in the Geopolitik from intellectuals of statecraft such as Alfreed Mahan and Nicholas Spykman in the United States, Friedrich Ratzel and Karl Haushofer in Germany, Rudolf Kjellén in Sweden, and Halford Mackinder (Box 3) in Great Britain. Their geopolitical imaginations supported the self-representation of the modern Western world as the “superior” culture, and subtly suggested the colonization of “premodern” others to be a necessary, if not “humane,” process.

**Box 3  Mackinder’s geopolitics**

Some of these nineteenth-century binary imaginations achieved a level of prominence that can, to some degree, still be found in present geopolitical discourses. For example, the construction of an antagonism between land and sea – or rather land power and sea power – formed the basis for Sir Halford Mackinder’s (1861–1947) construction, which became an important part of Anglo-American geopolitics. In 1904 Mackinder presented his theories at the Royal Geographical Society. His classification and division of the world political map into states of sea and land power formed the basis of many geopolitical imaginations. Russia is seen as the power center of the continental empire – the “heartland” (also known as the “pivot area”) – while the rest of the world, focused on the importance of oceanic influence, is arranged around this arc. This area is referred to as the “rimland” (including nations like Japan, Great Britain, and the United States). The balance of power throughout this dual-world system is the main focus of any developments within the field of foreign policy. Mackinder considered Eurasia to be the pivot of world politics since conflicts occur mostly in the rupture zones, and also argued that, by controlling this area, one could control the world, including dominating the sea power nations.
The purpose of geopolitics at that time had little to do with measuring and mapping the physical world; that was more along the lines of the newly developed and expanding sciences of applied geography and cartography. Geopolitics’ focus was much more toward constructing geopolitical representations of the self and the other on a global scale. This approach outlines the “seductive qualities of geopolitical discourse … [It] transforms the opaqueness of world affairs into an apparently clear picture” (Ó Tuathail 2006/1998, 2). In this way, geopolitics created and/or reproduced powerful geographies of difference on the world political map: the Orient versus the Occident, the Southern Hemisphere versus the Northern Hemisphere, the East versus the West, sea powers versus land powers, and so on.

During the nineteenth century, geopolitical representations began increasingly to influence the foreign policy zeitgeist of European politics. They included political practices that gave rise to countless wars in Europe from 1870 to 1945 as rival empires competed for power. In a similar way, they influenced the global order by serving as discourses of legitimization to support and validate imperialist, colonialist, and National Socialist politics of their times (see Schultz 1987).

Karl Haushofer and the geopolitics of National Socialism

Karl Haushofer (1869–1946) was one of the most influential scholars responsible for integrating the geopolitical imaginations of his time with National Socialist ideology. Haushofer, who considered Ratzel his mentor, shaped the German geopolitical discourse more than any other individual during that era. During World War I he was an officer who rose to the rank of major general. After World War I, he began his career as a professor of geography at the University of Munich. He published many books and papers, and was editor of the Zeitschrift für Geopolitik from 1924 to 1944. Haushofer’s position in the Third Reich was shaped by his close relationship with Rudolf Hess and the political standing of his sons, Heinz and Albrecht, in the Nazi regime. Through these contacts, Haushofer achieved a reputation that enabled him to directly influence the policy of the Third Reich. Thus, in the years between World War I and World War II, geopolitical representations achieved a hegemonic position in discourse, in Germany at least.

Rudolf Hess, one of Haushofer’s students, played a significant role in this development when Hess and Hitler were imprisoned together in Landsberg. Haushofer and Hess influenced Hitler and helped to create important parts of the initial ideological framework of National Socialism (e.g., the demand for more Lebensraum (living space)). Curiously, Haushofer did not become a prominent public figure within the National Socialist government. According to many historians, this was because he was satisfied with his important role of operations in the background.

Historical evaluation of Haushofer’s role will no doubt continue. There are interesting contrasts between the different representations of German and Anglo-American historical reconstructions. According to Anglo-American reconstructions, all geopolitics within the German-speaking sphere were developed by Haushofer, who is represented as the scientific brain behind Hitler. American perception of his importance was further encouraged by a Hollywood-produced propaganda film which portrayed Haushofer and his Geopolitical Institute as the headquarters of National Socialism. As Ó Tuathail (2006/1998) notes, Haushofer and his imaginary Geopolitical Institute in Munich
supported the growing public awareness in America of the role of geopolitics. The American geopolitician Edmund A. Walsh, who participated in the Nuremberg trials, could not conceal his admiration for Haushofer. In his interrogation of Haushofer Walsh did not criticize his general mode of geopolitical representation itself but simply accused him of pursuing the wrong type of geopolitics. Walsh’s demand for a “true” – as opposed to a “false” – geopolitics (re)locates the discipline within the democratic structure of Western society.

**Cold War geopolitics**

The years after 1945 were characterized by new forms of geopolitical representations that developed, to a much greater extent than before, in the scientific and political think tanks of the United States. Their goal was mainly a reframing of the American relation to the Eurasian continent. From a discourse theoretical perspective, it is interesting to note the extent to which these new representations relied on long-standing geopolitical imaginations, such as those of Mackinder (see Box 2) and Spykman. They incorporated the old argument of a “natural” line of conflict between a Western sea power (now the United States) and a central land power (now the Soviet Union). The practical consequence of such a geopolitical representation was the assumption that the political situation after 1945 did not allow for a stable order on the Eurasian continent without a permanent American presence. This was a strategy used to prevent the expansion of the Soviet land power. The geopolitical construction of Mackinder’s heartland theory is more than apparent here; it was the core element of the new geopolitical doctrine of the Cold War.

In 1947 Walter Lippmann, an American journalist, was the first to formulate the geopolitical imagination of an upcoming “cold war.” This was the beginning of the discursive construction of ideological blocs on the world political map. From the American perspective, it was a fight for liberty, democracy, and a market-based economy against communism, repression, and a command-based economy. From the Soviet Union’s perspective, “the good” and “the evil” were framed more or less the other way around. On March 12, 1947, the American president Harry S. Truman announced a similar, new form of geopolitical representation to the American Congress. With its dualistic approach and universalistic claim (liberty vs totalitarianism), it polarized parts of the world that had so far not been influenced by postwar geopolitics. After the Truman Doctrine, a key component of US foreign policy was to deliver active support to all people and states whose putative liberty appeared to be endangered by communism, socialism, or internal left-wing militant minorities. The first examples were Turkey and Greece, but – with hindsight – also (Western) Germany. For decades, the Truman Doctrine was utilized as the discursive legitimation of the American policy of containment and as the basic leitmotif of the Cold War. Another key component of this geopolitical representation was the “domino theory” proposed by US president Eisenhower on April 7, 1954, according to which, if one state falls (i.e., becomes communist) its neighboring states would fall in succession. In this manner, the geopolitical representation of the Cold War served the political goals of the superpowers.

After two decades, the 1970s, in many ways, marked a turning point in international relations. The ideological discourses surrounding the American–Soviet power struggle decreased, while other political framings grew in significance.
**GEOPOLITICS**

The pluralization of Western geopolitical representation after the end of the Cold War

As the Cold War came to an end, the geopolitical representation of a world composed of ideological blocs and iron curtains fell apart. Against this backdrop, it is not surprising that, in the early 1990s, a series of new geopolitical narratives had already begun to appear. Each of them claimed to offer plausible discursive framings for the changing global constellations of power (Ó Tuathail, Dalby, and Routledge 2006/1998).

Among them were, for example, geoecological narratives. Early representations originating from discussions in the 1970s and 1980s focused on the limitation, exploitation, and degradation of environmental resources (e.g., debates about the destruction of tropical rainforests, desertification, deforestation, nuclear power, greenhouse gases, etc.). While these discussions were, at first, confined to a niche group of people from affluent Western societies, they have, in recent times, become pertinent to wide-ranging scenarios in international security debates fueled, among others, by some of their well-known spin doctors. Since the Intergovernmental Panel on Climate Change (IPCC) reports, there has been a notable increase in more specific discourses on the “geopolitics of climate change” (for an analysis of this debate see Mattissek and Wiertz 2011), including risk scenarios such as “climate wars.” In this particular area, the discursive logic once again falls back on deterministic constructions which are on their way to becoming integral parts of political consulting and policymaking.

A similarly implicit form of global geopolitics arose at the beginning of the 1990s as an effort to substitute geoeconomics for geopolitics. According to Luttwak, political conflicts were increasingly being fought with economic weapons in the era of expanding (mainly economic) globalization. Visions of a network society, including new forms of global governance, were challenging the traditional relationship of politics and space – especially the role of the nation-state in international relations. Within most geoeconomical imaginations, the nation-state remains a relatively significant factor in the system as a whole. However, there are also conceptualizations, such as Manuel Castells’s network society, which predicts a less powerful role for the nation-state. Castells’s view represents a shift from a “space of places,” in which the institutions of the nation-state are the centers of (geo)political power, toward a “space of flows,” in which globally expanding networks are the most powerful economic and political structures.

The establishment and discursive power of such representations have changed greatly as a result of the conflicts and wars of the last two decades. Since their development at the beginning of the 1990s, new and alternative geo-“political” framings have emerged in both politics and the media. To bring the variety of these “politics of geopolitics” into some kind of perspective, one might work out two lines of geopolitical representation: (i) the geopolitics of fragmentation and conflict and (ii) the geopolitics of universalism and hegemonic superpower.

The geopolitics of fragmentation and conflict

This geopolitical narrative arose soon after the end of the Cold War and climaxed in the aftermath of the terrorist attacks of September 11, 2001. During this period, the discourse of a clash of civilizations, which had been promulgated by Samuel Huntington and immediately used by him to discursively frame the Balkan Wars, was ubiquitous in the media and in public opinion.
This kind of geopolitical imagination gave new life to an underlying discourse of global cultural difference that had developed in Western modernity over more than two centuries. Seen from this perspective, “Huntington’s thesis is not about the clash of civilizations. It is about making global politics a clash of civilizations” (Ó Tuathail 1996, 149).

Geopolitics of universalism and hegemonic superpower

The second major geopolitical narrative of the post-Cold War era serves the imagination of a new hegemon – the singular superpower, the United States. In his concept of “the end of history?” Francis Fukuyama represents the United States as the ultimate democratic state. His ideas are informed by a reductionist interpretation of Hegel’s idealistic philosophy of the end of history. This construction, in all its simplicity, marks the beginning of a geopolitical narrative that helps to pave the way for a new discourse of American hegemony in a post-Cold War world. Following a teleological and universalistic path of representation, the end of history thesis contains an underlying geopolitical narrative. This narrative is only a short step away from George Bush Sr’s doctrine from the First Gulf War, and it has informed important parts of the geopolitical representations and mappings of US foreign policy since 9/11.

The 9/11 attacks on New York and Washington formed the “discursive event” in the face of which the emerging geopolitical representations rose to prominence. At the same time, these events may be regarded as a shift in the geopolitical representations of power. For the first time, the traditional order of “modernity” (the Westphalian system of territorially defined nation-states) was challenged by an enemy organized (and acting) according to the principles of network societies. In the aftermath of 9/11, politics and the media have impressively shown how difficult (if not impossible) it was to comment on and to react to the attacks beyond the spatial framings and representations established in traditional geopolitical discourses. Thus, the narration about the conflict had to be (and was) partly reframed as a “war between nation-states.” Following Dalby (2003), this telling was indispensable because the geopolitical reasoning of modernity operates within an ontological framework of nation-states. It became, once again, the most important symbol framing political reactions and media reports.

The demand to “localize” the al-Qaeda terrorist network necessitated their attachment to a particular place; thus, Afghanistan was established as a target because of the enemy presence located there. This framing rests on geopolitical imaginations which have been developed in the previous decade, for example, by Michael Klare’s “rogue state” doctrine, by George H.W. Bush’s “axis of evil,” and – more generally speaking – by “failed state” discourses. All in all, the new storyline of a global “war against terrorism” turned very quickly into one of the most powerful doctrines of international geopolitics.

The subtle interplay between the two aforementioned geopolitical representations of hegemony and fragmentation were particularly helpful for this development. Both served as framings for numerous political speeches, as well as media reports and commentary. The majority of these geopolitical representations reified the underlying concept of a cultural fragmentation of the world, from which the necessity of strong US leadership supported by a “coalition of the willing” could be derived. Based on such an imagination, the wars against Afghanistan and Iraq could be justified more easily for the world public. Additionally, Huntington’s rhetoric of
conflict fits not only into the discourses of Western military strategists and politicians but also into the saga of the Islamic jihad. Thus, the clash of civilizations narrative produced a discourse that even antagonists such as Osama bin Laden could easily adopt (Agnew 2001).

Geopolitical imaginations like these are, of course, not the only narratives found within international relations academia. As a result of changing circumstances at the beginning of the new millennium, other variations entered the scene. For example, Thomas Barnett’s geopolitics of core and gap outlined a new world map for the Pentagon. In this representation, the war against Iraq marks a kind of geopolitical and historical turning point – the moment when Washington “really” took over its role as a guardian of strategic security. Barnett’s simplistic geopolitical world order entails regions with stable governments, high standards of living, and good connections to and throughout the globalized network society where democracy, liberal media, civil security, and human rights protection are part of the basic normative assumptions of society, which he calls the “functioning core.” At the same time, regions less touched by globalization and which suffer from repressive regimes, widespread poverty, disease, and – most importantly, in Barnett’s opinion – chronic conflict, in which coming generations of global terrorists are brought up, are denoted as the “arc of instability.”

Many of these geopolitical imaginations show how parts of the currently circulating representations of global and regional fragmentation and/or hegemony draw on key arguments coming from the discursive “archives of geopolitics” (Reuber 2011). For example, representations of global cultural differences had already been developed in the West since the beginnings of modernity (Gregory 1994). Zbigniew Brzezinski’s geopolitical imagination of Eurasia as a grand chessboard includes elements of geopolitics developed by Spykman and Mackinder. Tensions between Russia and the Western European states are often represented as a kind of relaunch of Cold War discourses and/or as long-standing geopolitical power constellations of “the East” versus “the West.” Together, they show, once again, how powerful geopolitical discourses are in framing the world political map and its ongoing clashes, conflicts, and wars.

SEE ALSO: Critical geography; Critical spatial thinking; Discourse; Imaginative geographies; Imperialism; Political geography; Poststructuralism/poststructural geographies; Power; Social constructionism; Territory and territoriality; War

References


Further reading

A geoportal is a type of web portal that is used to find and access geographic information and associated geographic services (e.g., display, editing, analysis) via the Internet. Geoportals are important for the effective use of geographic information systems (GIS) and are a key element of Spatial Data Infrastructure (SDI).

A web portal is most often a specially designed website that brings information together from diverse sources in a uniform way. Usually, each information source is assigned a dedicated area on the site for displaying information (a portlet); often the user can configure which ones to display. A portal may use a search engine to permit users to discover content that fits their demand. In addition, web portals may offer other services such as discovering, viewing, and downloading data. In the 1990s, many governments committed to creating web portal sites for their citizens. These included primary portals to the respective governments as well as portals developed for specific audiences. Examples of government web portals include: Australia.gov.au for Australia, USA.gov (in English) and GobiernoUSA.gov (in Spanish) for the United States, Gov.uk for citizens and businesslink.gov.uk for businesses in the United Kingdom, and Health-EU Portal (http://ec.europa.eu/health/index_en.htm) for issues related to relevant health topics from across the European Union (EU).

The main technical concept behind a web portal is to present the user with a single web page that brings together or aggregates content from a number of other systems or servers. The application server or architecture performs most of the crucial functions of the application. This application server is in turn connected to database servers, and may be part of a clustered server environment. For portals that present application functionality to the user, the portal server is in reality the front piece of a server configuration that includes some connectivity to the application server.

Over the last two decades, many governments and private companies have invested tens of billions of US dollars in the development of geographic information, largely to serve specific communities (e.g., agriculture, urban/rural planning, and mining) within local, state, national, international, and even global contexts. The focus has increasingly shifted towards a platform for integrating geographic information by means of SDIs. SDIs facilitate access to existing geospatial data and services necessary to successfully use GIS. Moreover, SDIs facilitate the exchange and sharing of geospatial data between stakeholders within the geographic information community. This community mainly includes mapping agencies, universities, governmental and nongovernmental organizations, and private companies.

Geoportals can be considered as gateways to SDI. They are not a repository where data are simply stored, but can be seen as a one-stop shop for geospatial data, sourced from numerous agencies. The performance of geoportals can vary enormously depending on numerous factors, such as the functionalities offered, the quality of the information offered, and a user's capacity.
Geoportals are named differently across the geospatial community. Examples of other names used are “geo registry,” “geospatial one-stop,” “geospatial platform,” “geospatial data service,” and “spatial portal.” Although different names are used, it is obvious that the goals for searching and accessing geospatial data and services remain the same.

Geographic information providers, including government agencies and commercial companies, use geoportals to publish descriptions (in terms of geospatial metadata) of their geographic information and/or supply access to a wide range of geospatial data, such as topography, soil, administrative units, addresses, and land use. Geographic information consumers, professional or casual, use geoportals to search and access the information they need. Thus geoportals play an increasingly important role in the sharing of geographic information, and help to avoid duplicate efforts, inconsistencies, delays, confusion, and wasted resources.

Geoportals often include, among others, discovery, web mapping and/or download services. Discovery services allow users to search for geospatial datasets and services on the basis of the content of the corresponding metadata and to display the content of the metadata. The web mapping services make it possible, at a minimum, to display, navigate, zoom in/out, pan, or overlay viewable geospatial datasets, and to display legend information. The download services enable copies of geospatial datasets, or parts of such sets, to be downloaded and, where practicable, accessed directly. Examples of other web services that are sometimes included are gazetteer, geoprocessing, (coordinate) transformation, authentication, ordering, and e-commerce services.

In order to ensure interoperability among the geospatial datasets and services provided by geoportals, standards are essential. These geospatial standards are required to store, transfer, and visualize the datasets in a systematic and compliant way, or to describe them in terms of metadata. Moreover, standards are also needed to apply web services and make them compliant with relevant standards. These standards are developed by standardization bodies, such as the International Organization for Standardization (ISO), Open Geospatial Consortium (OGC), and World Wide Web Consortium (W3C).

In 1994, the US Federal Geospatial Data Committee (FGDC) established the National Geospatial Data Clearinghouse (NGDC; http://www.fgdc.gov/dataandservices), aimed at facilitating efficient access to the overwhelming quantity of existing geospatial data (from federal agencies) and coordinating its exchange, with the objective of minimizing duplication (in the collection of expensive geospatial data) and assisting partnerships where common needs exist. The NGDC is considered the earliest implementation of a geoportal.

Since 1994, the number of countries implementing national geoportals has steadily grown. As of February 2014, around 120 countries have an operational national geoportal in place and 12 countries initiated projects to launch a geoportal in the short-term. Most countries in Asia, Europe, the Middle-East, Oceania, North America, and South America have established a geoportal for their nation, whereas most countries in Africa still have not established such a portal. However, several African initiatives to launch national geoportals appear promising. These national geoportals are evolving worldwide in tandem with national SDIs. A body of literature published in scientific/popular journals and conference proceedings describe the existing experiences (e.g., see conference papers of the Global Spatial Data Infrastructure Association).

First released in 2003, the Geospatial One-Stop (GOS) geoportal was developed as part of a US e-government initiative. Unlike the
National Spatial Data Infrastructure (NSDI) Clearinghouse Network, GOS was built around a centralized metadata catalog database, with an architecture that links users to data providers through a web-based geoportal. The user of GOS may employ a simple web browser (thin client) or may interface directly with a GIS (thick client).

In September 2011, GOS was retired and its content became part of the broader open data site (Geo.)Data.gov (http://catalog.data.gov/dataset). At the same time, the United States federal government launched the Geospatial Platform (http://www.geoplatform.gov/), which represents a shift from focusing on cataloging references and resources, to providing shared web services for nationally significant datasets, an application program interface (API) for developers, and end-user applications (built on those web services and the API).


Geoportals are also implemented at administrative levels other than the national one. These geoportals mainly differ in the size of their spatial coverage, their governance, and the organizations involved in their management. Global geoportals cover the whole world, are based on strong cooperation, and are managed by international organizations, while local geoportals cover a relatively small area (e.g., size of a city, municipality, or county), are more based on clear governance rules and working structures, and are usually managed by local authorities. Each of these geoportals is briefly introduced, followed by some examples.

Global geoportals are electronic facilities that provide access and mechanisms to search geospatial data from all over the world. An example is the GEOSS Portal (http://www.geoportal.org) which is operated by the European Space Agency and the Food and Agriculture Organization of the United Nations. It provides a web-based interface for searching and accessing Earth observation datasets, information, imagery, services, and applications. As such, it aims to connect users around the world to a variety of datasets, services, and portals that provide reliable, up-to-date, integrated, and user-friendly information – vital for the work of decision-makers, managers, and other users of Earth observations.

More recently, there has been a proliferation of geoportals for sharing of geographic information based on an international region. An important example of a regional geoportal is the INSPIRE
Geoportals (http://inspire-geoportal.ec.europa.eu/). This is Europe’s access point to the Infrastructure for Spatial Information in Europe (INSPIRE). The INSPIRE Geoportal has been established using standards and specifications from European, international, and industry consensus-building processes. Where appropriate and possible, this geoportal links to national geoportals and to relevant theme/sector-specific data and services. This geoportal does not store or maintain the data; these are distributed in the many national and thematic/sectorial servers across Europe, and are maintained by the organization responsible for the data. The INSPIRE Geoportal publishes and provides access to metadata and data, and facilitates the delivery, display, and analysis of geographic information. The user can discover and view maps of his/her choice letting the geoportal contact the necessary servers to combine the data.

Other examples of international geoportals are the GeoSUR Portal, Mountain Geoportal, and Central Asian Countries Geoportal. The GeoSUR Portal (http://www.geosur.info/) provides an entry point to geospatial data published by Latin American and Caribbean public agencies. The data may be consulted directly by means of several web services or through metadata services. The Mountain GeoPortal (http://geoportal.icimod.org) is a gateway for geographic information and knowledge resources to support mountain development policies and practices in the Hindu-Kush Himalayan region (which includes: Afghanistan, Bangladesh, Bhutan, China, India, Myanmar, Nepal, and Pakistan). The Central Asian Countries Geoportal (http://www.cac-geoportal.org) mainly aims to distribute geoscientific information to support increased investments in the mining sector of Kazakhstan, Kyrgyzstan, and Tajikistan, and to assist those countries in their efforts related to climate risk management.

Geoportals are also implemented at the level of a state, province, or a similar level of government. An excellent example of this category of geoportals is the IDEC Geoportal – Catalan abbreviation for SDI Catalonia (http://www.geoportal.cat/geoportal/cat/index.jsp, Spain). This geoportal is the gateway to access geospatial information related to the Catalanian state. The Geoportal of IDEC gives access (in Catalan, Spanish, and English) to metadata and web mapping services and provides large amounts of geospatial data and services to the public. For example, almost a billion service requests were made between 2010 and 2012. These included 244 972 requests for discovery services, 732 082 496 for viewing services, 261 878 706 for download services, and 7923 for transformation services. There are also thematic applications addressing the following themes: Planning, Land and Property; Leisure and Entertainment; and Geology. The IDEC Geoportal is accessed by ± 50 000 visitors monthly.

Other interesting State/Provincial geoportals are: Geopunt.be Flanders (geopunt.be, Belgium), Geoportal.rlp – Rhineland Palatinate (http://www.geoportal.rlp.de, Germany), Bayern Geoportal – Bavaria (http://geoportal.bayern.de/georealbayern, Germany) Geoportal Kyzylorda (Baikonyr) region (http://map.e-kyzylorda.gov.kz, Kazakhstan), GeoStor – Arkansas State (http://www.geostor.arkansas.gov, USA), Abu Dhabi Geospatial Geoportal (http://geoportal.abudhabi.ae/geoportal, United Arab Emirates), Geoportal of Province Drenthe (http://www.drenthe.info/kaarten/website/geoportaal, The Netherlands), and New Mexico Resource Geographic Information System (http://rgis.unm.edu, USA). In Switzerland, most cantons manage excellent geoportals, such as Geoportal Canton Bern (http://www.apps.be.ch/geo/de), Geoportal Canton Luzern (https://rawi.lu.ch/themen/geoportal), and
During the past several years, an increased number of local geoportals have been developed. These geoportals are mainly implemented for areas such as a city, municipality, or county. Excellent local geoportals are: Geoportal Kreis Pinneberg (http://www.geoportal.kreis-pinneberg.de, Germany), Geoportal Münsterland (http://www.geoportal-muensterland.de, Germany), Geoportal Waiblingen (http://geoportal.waiblingen.de, Germany), Geoportal BCN–Barcelona (http://www.bcn.cat/geoportal, Spain), Geoportal Campiña de Jerez (http://campinajerez.geoportales.com, Spain), Madrid Geoportales (http://madrid.geoportales.com, Spain), Town in Map ’s-Hertogenbosch (http://geoproxy.s-hertogenbosch.nl/apps1/geoportal.html, The Netherlands), and City of Raleigh Geoportal (http://maps.raleighnc.gov/geoportal, USA). An interesting example of a local geoportal is the one for Eeyou Istchee (http://www.creegeoportal.ca, Canada), which delivers geospatial information to Cree communities to support research, discussion, planning, and decision-making related to resource management and tourism development. This geoportal incorporates Cree traditional knowledge datasets in combination with information and base maps (vector and satellite images) from other distributed data sources.

More recently, there has also been a proliferation of geoportals for sharing geographic information based on themes such as environment, geology, culture, tourism, archeology, and statistics. The following geoportals focus on a specific theme. The NatCarb Geoportal (http://www.netl.doe.gov/research/coal/carbon-storage/carbon-storage-natcarb) provides geographic information concerning carbon sequestration in the USA. The EPA Geoportal of the Irish Environmental Protection Agency (http://gis.epa.ie) is designed to make data about the environment easier to find, browse, and understand. The OneGeology-Europe Geoportal (http://onegeology-europe.brgm.fr/geoportal/viewer.jsp) allows users to make digital geological data more easily accessible and useable. The Spain is Culture Geoportal (http://geo.spainsculture.com/index.php/es) from the Spanish Ministry of Education, Culture, and Sport provides information about cultural sites in Spain; meanwhile GeoVisitGuatemala (http://www.geovisitguatemala.com) provides users access to relevant tourist information about Guatemala. The Archeological Atlas of the 2 Seas (http://www.a2s-geoportal.eu) is a gateway to underwater heritage. It aggregates information about archaeological sites that lie beneath the Channel and the North Sea (Europe), as well as sites situated on the foreshore, dating from prehistoric to the present time. Finally, the national statistics agency of Colombia developed the SIGE Geoportal (http://www.dane.gov.co/geoportal), which provides facilities to access statistical information for the territory of the Republic of Colombia.

Other geoportals are an initiative of a specific sector. A good example of a sectorial geoportal is the Scholars Geoportal (http://geo1.scholarsportal.info), which is a web portal of the Ontario Council of University Libraries consisting of a consortium of 21 academic libraries in the province of Ontario (Canada) acting under the authority of the Council of Ontario Universities. It provides access to millions of articles and electronic books, as well as numeric and geospatial data files for the use of researchers and students of its member libraries.

Finally, some geoportals concentrate on just one specific type of geospatial data. For example, the Russian Federal Space Agency (Roscosmos) launched a geoportal (http://geoportal.ntsomz.ru) that provides access to remote sensing images.
of the Earth only. The images are acquired by Russian remote sensing satellites and updated on a daily basis. It has around 900,000 monthly visitors.

Two examples that will likely set the future direction of geoportals are the OpenGeoportal initiative and the EnerGEO Knowledge Geoportal. OpenGeoportal (http://opengeoportal.org) is a collaboratively developed, open source, federated geoportal used to rapidly discover, preview, and retrieve geospatial data. It is also a collaborative effort to share resources and best practices in the areas of application development, metadata, data sharing, data licensing, and data sources in support of geospatial data repositories. The OpenGeoportal brings together geospatial professionals, developers, metadata specialists, and librarians in order to develop facilities to discover and acquire geospatial data across many organizations and platforms. The initiative also focuses on the development of The Open Geoportal Cloud Federation for sharing innovative ideas and best practices around open source SDI. The EnerGEO Knowledge Geoportal (https://energeo.researchstudio.at/) is an EU EnerGeo project that aims to provide a versatile modeling platform that will enable planners, environmentalists, and governments to calculate, forecast, and monitor the environmental impact of changes in energy mixes on local, regional, and global scales. This gateway to web-based geospatial resources enables users to discover, view, and access geospatial data, services, and applications made available by several providing organizations. Likewise, data providers use the Geoportal to make their geospatial resources discoverable, viewable, and accessible to others. These resources can be metadata records, geospatial datasets, web services, videos, SharePoint® documents, RSS feeds, KML (formerly Keyhole Markup Language) documents, metadata catalogs, and more.

**SEE ALSO:** Geographic information system; Geospatial metadata; Metadata; Spatial data infrastructures; Spatial database; Web-mapping services

**Further reading**


Georeferencing is the process of linking between an entity and a spatial footprint. For example, a linguistic sign, such as a piece of text like the mention of a place name (e.g., Paris) or postal address (e.g., 7 Cosin Court, Cambridge, CB2 1QU, England, UK), can be associated with a particular place on Earth referred to by such an expression, which may be a polygon or a centroid (i.e., a geospatial “center of gravity”) given as latitude and longitude, for instance.

This entry discusses georeferencing as a concept, and furthermore, it describes the nature of georeferenced datasets, the motivation for and history of creating online georeferenced datasets, means to automatically obtain them using toponym resolution, and some applications enabled by it.

Automatic georeferencing as a method (as applied to textual document collections in particular) is described below as a combination of a named entity recognition and classification (NERC) process that recognizes and classifies mentions of places (NERC specifically for geographic named entity types only is also known as “geoparsing”) combined with a toponym resolution process that disambiguates, if necessary, using linguistic and spatial context, that is, it selects one single, most likely intended, spatial interpretation among a set of candidate locations looked up in a gazetteer (a database comprising toponyms, geographic feature types, and spatial footprints). A candidate location (or candidate for short) is a potential referent, that is, the set of all theoretically possible interpretations of places with the same name.

Motivation

A text collection (or any other collection of things, for that matter) is said to be “georeferenced” if particular elements of that collection are linked to a spatial footprint (if we limit our investigation to terrestrial geography, we may speak of a “geospatial” footprint). Researchers have attempted to georeference collections of videos, collections of photos uploaded to photo-sharing websites (digital photographs often contain location information by making use of GPS, the Global Positioning System), and collections of mostly textual documents, including digitized historic manuscripts and webpages from the World Wide Web.

Any georeferenced collection has the advantage that it is capable of being processed by computer with regards to carrying out spatial operations on it (plotting maps, filtering by country, etc.); for example, using the set of words as the representation of a story does not permit us to compute the geographic distance between two stories, one about Mountain View, CA, the other one about San Francisco, CA. But once the stories are connected with a spatial footprint, we can calculate the distance to be 52.32 km (or 32.51 miles) from the numeric latitude/longitude representation of the cities’ respective centroids. Furthermore, there are numerous applications in the natural sciences and medicine: for example, in botany we may want to georeference a document collection.
describing sighting of particular plants, in order to determine the geographic distribution of species. Or an epidemiologist may want to visualize the spread of a disease based on aggregating georeferenced reports about disease outbreaks.

Where spatial metadata such as GPS coordinates are not available, spatial information may be annotated manually, but due to the time and cost required, manual processing is not a viable solution for larger datasets; automatic methods for adding spatial metadata to textual collections have thus been investigated.

History

After some initial, isolated system development efforts in the areas of digital libraries (Smith and Crane 2001), web search, (Ding, Gravano, and Shivakumar 2000), and the defense sector in the 1990s, academic research interest began in the natural language processing community marked by a workshop held at the Joint Conference on Human Language Technology and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL) in Edmonton, Alberta, in 2003 (Kornai and Sundheim 2003), where new algorithms, systems, and applications were presented. An early landmark publication was the monograph by Hill (2006) on georeferencing, the publication of which, however, occurred while research on the topic, especially on algorithms, was still nascent.

Around the same time the Cross-Language Evaluation Forum (CLEF), a European forum for researchers interested in information retrieval in different languages, carried out a shared task (i.e., an international evaluation campaign for benchmarking the state of the art under controlled conditions), called “GeoCLEF” on geographic information retrieval evaluation (Gey et al. 2006). From the campaign emerged a sense that the problem was harder than anticipated, and the evaluation itself posed some challenges (such as geographic diversity of queries and documents in the evaluation collection). Independently, a series of workshops on geographic information retrieval (GIR) have also been held since 2004 as satellite events at geographic and natural language processing conferences (Purves and Jones 2004) to bring the information retrieval and geographic information systems (GIS) communities together.

As a result of the HLT-NAACL, CLEF, and GIR workshops, interest was stimulated in various research communities, which spurred the development of novel methods for toponym resolution and geography-aware search. However, emerging approaches that applied “toponym resolution” (Leidner 2007) to create metadata tags in order to transform implicitly georeferenced collections into explicitly georeferenced collections suffered from the lack of evaluation datasets, a shortcoming which was eventually (at least partially) addressed by the TR-CoNLL and TR-MUC4 corpora (Leidner 2006) and the TRML (Toponym Resolution Markup Language) and SpatialML mark-up languages (Mani et al. 2008).

More recently, toponym resolvers have made use of knowledge-based resources like Wikipedia for inducing statistical classifiers (Speriosu 2013), and adaptive systems for processing document collections performing stream-oriented toponym resolution in feeds of news or social media (i.e., user-generated content) have been developed (Lieberman and Samet 2012).

Terminology

Looking at the linguistic expression of spatial signs, we can distinguish between a toponym, which denotes a single sign (e.g., Long Island) versus a geographic expression, which is a complex
sign (e.g., 10 km north of Paris), for which linguistic rules of semantic compositionality apply. If *spatial grounding* is any association of an entity (such as a linguistic sign) with spatial footprints, then we can distinguish between (address) geocoding, the mapping from postal address records (such as those that can be found in relational databases used for advertising; governments are increasingly under pressure, due to the Linked Open Data movement, to make postal address data and ZIP code information curated from taxpayer income available free of charge) to geospatial footprints and *toponym resolution* (TR), a term coined in Leidner (2006) to denote the mapping from place names mentioned in “unstructured” text to geospatial footprints.

Toponym resolution as a task is challenging for two reasons: (i) *geo/non-geo ambiguity* refers to the potential for confusing toponyms with other words that are not associated with a location – for example, the Turkish city of Of can be confused with the English particle of (which is capitalized at least at the beginning of sentences) – and (ii) *geo/geo ambiguity* is the potential for confusing different referents for a toponym (Paris can refer to Paris, France, or to Paris, Texas, USA, for example). There are millions of named places on Earth, many of which share the same name: for instance, there are over 1600 attested places called Santa Ana. The same geographic names have historically been used by emigrant settlers to denote their new homes, leading to referential ambiguity of place names for machines and sometimes for humans as well. Occasionally, the original name was modified (as in York versus New York).

Toponym resolution requires a number of steps: (i) identifying place names in texts via geoparsing, (ii) retrieving candidate location information from a gazetteer, and (iii) resolving toponyms by selecting one particular candidate from the retrieved set. A toponym resolver performs all three tasks, either sequentially or in an integrated fashion.

A *gazetteer* is a database comprising toponyms, geographic feature types, and spatial footprints, and it provides a mapping toponym $\rightarrow$ feature $\times$ footprint (Hill 2006). Toponym resolution requires the use of gazetteers, since the candidate places to choose from are external world knowledge that cannot be derived from linguistic principles alone (i.e., pragmatic rather than semantic in nature). Geoparsing is the task of identifying the names of places in textual documents (introduced above as NERC). A *toponym resolver* is a program that can carry out the toponym resolution task automatically; it typically contains a geoparser to identify mentions of places and a gazetteer to look up candidates to be chosen from. By analogy, a geocoder carries out the geocoding task automatically on address record data instead of unstructured (prose) text.

A georeferenced document collection contains metadata that links postal addresses, toponyms, geographic expressions, or other location-related textual signs with spatial footprints that refer to the locations in the world external to the text. The addition of such metadata annotation transforms implicitly georeferenced data (e.g., a mention of Paris) into explicitly georeferenced data (e.g., a mention of Paris marked up with XML (Extensible Markup Language) tags referring to the intended Paris’s latitude and longitude). So geocoders and toponym resolvers both turn implicitly georeferenced datasets into explicitly georeferenced datasets.

**Methods**

Methods for toponym resolution can be grouped into two broad categories: (i) heuristic-based methods, which utilize rules informed by human intuition or observations, and (ii) learning-based
methods, which are informed by statistical associations (correlations) acquired from the mentions of toponyms in a training corpus (a ground truth reference dataset annotated by humans) and the associated features of the contexts in which they occur.

Heuristic-based methods

A detailed survey of the literature is given in Leidner (2007). Hauptmann and Olligschlager, as cited in Leidner (2007), propagate readings from explicitly disambiguated contexts (London, Ontario) to other mentions of the same token without any contextual cues (London), thus implicitly applying the common heuristic “one sense per discourse”: it assumes humans never mix different Londons in the same document without giving explicit guidance for the human reader.

Others, such as Smith and Crane (2001), rely on more spatial heuristics, such as the iterative computation of a geographic region of interest: a centroid for a document is first calculated by creating a bitmap of frequency-weighted mentions of toponym candidates (all candidates considered equally likely initially); then, candidates further away from the centroid than two standard deviations are discarded, which leads to a revised centroid. A sliding window containing four mentions of ambiguous toponyms is then used to incrementally resolve each of them, using local rules such as disambiguation patterns, and the per-document centroid as background knowledge.

Perhaps the single most effective heuristic is using population knowledge to favor candidate toponym interpretations that have a higher population, a technique that works particularly well for news documents.

Another type of evidence is the spatial minimality heuristic, which forms the basis of the “minimum bounding polygon” assumption (Leidner, Sinclair, and Webber 2003). The latter can be stated as: “For a set of toponyms in a document or paragraph, always assume the interpretation that minimizes the area of the smallest polygon that contains all candidate referents.” In other words, as can be seen in Table 1, this heuristic exploits the fact that toponyms mutually disambiguate each other because of the higher likelihood that places in geographic proximity to each other are also more likely to get mentioned together. The spatial minimality heuristic also works with superordinate and subordinate toponyms represented as centroids, but should be restricted to application within one event or story to be most effective.

In an attempt to organize how the various approaches relate to each other, Leidner (2007)

| { Berlin; Potsdam; Babelsberg }  | →  | Berlin, Germany |
| { Fairburn; Berlin; Green Lake } | →  | Berlin, WI, USA |
| { West Berlin; Bishops; Dicktown } | →  | Berlin, NJ, USA |
| { Kensington; Berlin; New Britain } | →  | Berlin, CT, USA |
| { Copperville; Berlin; Gorham } | →  | Berlin, NH, USA |
| { Atlanta; Moultrie; Berlin } | →  | Berlin, GA, USA |
| { Berlin; Prouty; New Berlin } | →  | Berlin, IL, USA |
presents a synthetic typology comprising 16 types of spatial and linguistic knowledge (heuristics, pieces of evidence) used in systems up to that time, and an analysis of the various systems in which they are applied.

**Learning-based methods**

An early method of machine learning-based approaches to toponym resolution is the generation of synthetic training data in nearly limitless supply by taking identified toponym mentions that are already uniquely resolved due to subsequent mentions of geographic cues, and by treating these as class labels for training (Smith and Mann, as described in Leidner 2007). After training, the classifier is capable of tagging mentions of “Paris” with the tag “TX,” effectively recovering the most likely candidate seen in the training data, given the particular linguistic context, that is, we apply the class labels to disambiguate toponyms from a test text based on the class characteristics associated with the identified toponyms. For example, “Paris, TX” is treated as a training data point “Paris” with the tag “TX,” and a naïve Bayes classifier using a sliding window of context words as input feature vectors can be trained so as to be able to correctly tag “Paris” with its “TX” pseudo-tag in a version of the data with the pseudo-tag removed. Given enough training material, this reportedly worked over 87% of the time. The downside of the method is that not all toponyms are necessarily annotated by easy-to-identify cues often enough to train one classifier per toponym successfully.

Speriosu (2013) describes a set of learning-based classifiers. The strongest presented approach uses machine learning to construct one model per toponym using so-called indirect supervision: his method WISTR (Wikipedia Indirectly Supervised Toponym Resolver) uses a document centroid and generates training instances at the toponym level using a lookup of toponyms in a separate training corpus (based on the body of all Wikipedia articles which contain manually crowdsourced geographic references) with candidate locations close to location labels on the documents containing them. For example, the Wikipedia articles “Portland Youth Philharmonic” and “Widgery Wharf” have both been manually georeferenced by Wikipedia’s authors (Speriosu 2013). From this information we can infer, for instance, that $P(\text{Portland, Oregon} | \text{“music”}) > P(\text{Portland, Maine – “music”})$ and $P(\text{Portland, Oregon} | \text{“wharf”}) < P(\text{Portland, Maine – “wharf”})$, where $P$ represents probability. Any “Portland” mentions yet to be disambiguated are more likely to be in Oregon if “music” occurs in the textual environment, but are more likely to be in Maine if “wharf” occurs in the textual environment. The reason is that there exists the Wikipedia entry for Widgery Wharf in Portland, Maine, so the word “wharf” bears a stronger statistical association with Maine than with Oregon. Likewise, the georeferenced article “Portland Youth Philharmonic” is about music, hence the word “music” has stronger statistical ties with Portland, Oregon, than with its counterpart in Maine. The WISTR method is a supervised resolver that incorporates more knowledge than heuristic-based resolvers, while not requiring any hand-annotated training instances of the same type as the evaluation instances used (beyond the already-existing manual georeferences in Wikipedia).

Furthermore, Speriosu (2013) also frames toponym resolution as an instance of the traveling purchaser problem (a generalization of the NP-complete traveling salesman problem), where he explicitly models the individual resolution power of the WISTR classifier and spatial minimality as two components of a cost function to be optimized for each document. Ant colony optimization techniques are then used to find
approximately best solutions, but the method is not reported to surpass results of the simpler WISTR method.

**Evaluation**

**Evaluation methodology**

When building any system, it is important to be clear and explicit about how to evaluate success. A complete experimental protocol requires (i) controlled (comparable) conditions in which alternative methods are compared to each other, which in the case of toponym resolution means using both the same gazetteer and the same evaluation dataset, (ii) specifying a metric to quantify system quality, and (iii) specifying the methods compared in a way that can be replicated, at least in principle.

There are many different ways in which a system comprising different subtasks (identifying toponym mentions, disambiguating them) may be evaluated. Leidner (2007) distinguishes between “in vitro” (component) evaluations, where a perfect (oracle) ability of recognizing place names is assumed, to measure a toponym resolver’s ideal-case behavior, and “in vivo” (system) evaluations, which, especially when contrasted with in vivo scores, attest to a methods robustness in the face of systematic errors caused by upstream name taggers.

For example, Amitay et al. (2004) evaluated their Web-a-Where system by running it over three web collections, and using human post hoc judgments to measure the correctness of tagging place names while using the Open Directory Project’s (ODP) location categories as reference answers against which to compare their system’s geofocus algorithm (which has a reported accuracy of 38% for the latter task).

Leidner (2007) was the first controlled study that compared the performance of various heuristics, previously proposed as well as novel, as well as various system combinations, using both the same gazetteer and the same, manually georeferenced, test data collection. The motivation behind it was that prior work had either not reported quantitative performance metrics, or quantitative results were of limited use, as they could not be compared due to differences in the task or even task difficulty.

**Evaluation datasets**

A core prerequisite for building georeferencing systems such as toponym resolvers is the availability of ground truth datasets that can be used to compare system-generated spatial metadata with that ground truth to evaluate the accuracy of the systems. Reference data comprises the original data (such as historic or news documents) and spatial metadata in some standardized annotation format, which may be an application of XML.

Leidner (2006) presents the annotation format TRML, an XML application, and Mani et al. (2008) present SpatialML, the design of which was informed by TRML and for which translations from and to TRML have been defined. Leidner (2006) also describes TR-CoNLL, a dataset that was dedicated specifically for the evaluation of toponym resolution by adding a spatial metadata layer on top of an existing news corpus already commonly used in named entity tagging research. TAME, a web-based editor for generating such datasets, is also presented. Leidner (2007) uses TR-CoNLL and TR-MUC4, another corpus of intelligence analyst reports marked up with spatial metadata. The rationale of using the two datasets in particular was to be able to compare general news stories, which have a rather global geofocus, with stories that describe specific events, focusing on a particular region (e.g., Central America). Both corpora contain annotations
for all toponyms with latitude and longitude information. TR-CoNLL has been used by many research groups in academia, industry, and governments to evaluate different systems.

An alternative evaluation corpus is presented by Overell and Rüger (2006), who describe an evaluation based on a random sample of 1000 Wikipedia pages manually annotated with place names and latitude/longitude information against the Getty Thesaurus of Geographic Names (TGN). The advantage of Wikipedia is its convenient license, which avoids copyright limitations of the data. However, the use of the Getty Thesaurus, a commercial resource, reintroduces these issues. The encyclopedic style of Wikipedia is also more distinct from “general prose” than news articles.

State of the art

Achieving explicit georeferencing for textual data automatically through applying toponym resolution remains a challenging task. Precision (P), Recall (R), and their harmonic mean, F-Score, are the most common metrics used to report the quality of toponym resolution methods.

Leidner (2007) finds that minimality-based heuristics such as spatial minimality (implemented as “assign the interpretation of all toponyms that minimizes area of the bounding rectangle”) and a variant of Yarowsky’s rule from word sense disambiguation (“assume one referent per discourse”) are strong pieces of evidence, together with the maximum population heuristic and contextual patterns. In combination, these rules outperform earlier methods like Smith and Crane (2001). However, the maximum population heuristic alone often beats more complex methods and systems.

Speriosu (2013) reports 62.2% F-Score (P = 74.0%, R = 53.6%) for his learning method WISTR on a held-out test-set portion on TR-CoNLL using automatic NERC to recognize place names. For the same portion of the data, a simple “maximum population” heuristic achieves an F-Score of 60.2% (P = 71.7%, R = 52.0%) with less computational effort. This means that nearly three out of four toponym mentions processed can be mapped to the right coordinates, but that only about half of the toponyms could be resolved. Findings by Leidner (2007), Speriosu (2013), and others also suggest that population is a very strong predictor for the place intended by a toponym mention, and it appears that combining methods, as well as making use of external knowledge, is required to perform significantly better. However, population statistics are not always available, an issue that holds especially for smaller places. Complementarily, local textual evidence (such as contextual patterns) has been found, in general, to be more helpful with global-scope news corpora, whereas spatial evidence (such as the spatial minimality heuristic) have been found to be more valuable with corpora that have a more narrow, “zoomed-in” geofocus, such as intelligence reports.

To achieve significantly higher results, larger training corpora and more investigations into the combination of sources of evidence will be required, and better ways to make toponym resolution automatically adaptive to the text type and specifics of the corpus under consideration should be studied. A broader scope for improvement of robustness of existing techniques could be provided by the exploitation of the confidence level of interim resolution decisions.

Applications

Once a document or document collection has been processed by a toponym resolver, this opens up many possibilities for applications (Leidner
We can distinguish between applications that directly deliver value to the user and such applications that feed into further automatic processing to serve another end.

The most common applications of georeferenced collections are map visualization, spatial summarization, and interactive browsers. Each toponym grounded by resolving it to a centroid given as a latitude and longitude coordinate pair can be placed on a geographic map. A web-based system can retain the link between text and space by providing a hyperlink from the textual occurrence to a point of interest (POI) on the map, which permits navigation back and forth between the textual and the geospatial views.

Spatial summaries are visual surrogates that permit an at-a-glance overview, aggregating the locations mentioned in georeference collections. While tag clouds often serve that purpose for nonlocation key words, georeferenced collections can provide better spatial surrogates for toponyms in the form of polygons or point clouds, indicating the frequency of mention of locations in a story or collection by representing increased density via rendering them as darker colors (spatial heatmaps), for example.

Going beyond static maps, georeferenced collections permit interactive exploration using special browsers or applications and web browser plugins such as Google Earth (http://www.google.com/earth/; formerly known as Keyhole), which uses KML (Keyhole Markup Language), an XML application, for marking up locations, which can be rendered in custom-purpose layers on a three-dimensional model of the globe, which can be rotated, looked at, zoomed in and out, and hyperlinked with other documents (Speriosu 2013).

All of these mapping applications can serve a range of purposes, such as (i) the mapping of crime “hot spots”; (ii) intelligence analysis to spot patterns in terrorist attacks in order to prevent them; (iii) to inform political and commercial decisions such as taking unstructured textual sources (e.g., complaints on social media) into account when making decisions about where to place a new school so that it can serve the largest number of households, or where to place an additional branch of a chain so as to serve their respective constituents best – these latter problems could traditionally only consider structured household or census data, whereas georeferencing additionally permits the use of data sources made up of textual (“unstructured”) prose to supplement the evidence; (iv) the georeferencing of digitized historic text collection in order to support historic research (Smith and Crane 2001), an important part of digital humanities scholarship – geographic boundaries are changing due to changes in physical geography or administrative reorganizations, and to map between historically accurate boundaries remains a challenge.

A fundamentally important function of computer processing is to support a user’s search for information. Georeferenced datasets are more searchable because they support spatial filtering as an advanced search operation, for example to retrieve and show on a map only those documents that mention a particular city, region, or country. Since documents may contain toponyms of various feature types and levels of granularity (from continent to village), keyword matching is not enough to identify location (a text about the political unit the United Kingdom of Great Britain and Northern Ireland or about the geographic unit of the British Isles may just mention London, but is relevant to an information need expressed in a query as “UK” by implication). While it is possible to model geographic or geopolitical part-of relationships as taxonomies or ontologies and to use a purely term-based approach, query expansion with all subterms is likely to retrieve...
many irrelevant documents. In contrast, a spatial filtering approach can be informed by the actual polygons associated with explicitly mentioned toponyms, as was attempted in the SPIRIT geographic search engine at the University of Sheffield and some participants at GeoCLEF mentioned above. Web search engines can benefit from assigning a geographic scope to a webpage at indexing time (Ding, Gravano, and Shivakumar 2000), although different notions of place can (and should) be distinguished (location of the web server, location of the author of a story, location of where an event happens, etc.). Given a particular area of interest, expressed as a pair of a centroid and a diameter or as a polygon, incoming news stories from a feed can be georeferenced, and matched against such a so-called geofence. In case the story mentions an event inside the area of interest, an automatic alert can be initiated.

Once both large georeferenced collections of documents and comprehensive gazetteers are available, spatial reasoning opens up further possibilities for automatic processing. If a city is mentioned to have been the target of two terrorist attacks in a year, and if we further know via gazetteer lookup what country the city is in, then an information need regarding the number of terrorist attacks in that country can be calculated bottom-up from all such individual events in cities in the country under consideration.

Answering geospatial questions is a natural application that can benefit from the reasoning capabilities mentioned above. Once toponyms in documents are georeferenced, questions like, Which town north of Springfield did Peter travel to? can be answered correctly and automatically, given a document, even if Peter traveled to two different towns (if the second town was mentioned as being south of the Springfield referred to). In a similar way, distance questions can be answered numerically, that is, based on an actual distance calculation rather than by extracting a text fragment stating the distance in sentence form, as such textual fragments are unlikely to exist for all pairs of any two toponyms (Leidner 2007). The increased use of mobile devices suggests spatial technologies like georeferencing will grow in importance in the future.

SEE ALSO: Geocoding; Geographic information system; Point-of-interest databases and gazetteers; Toponymy

References


Further reading


Geospatial metadata

David J. Cowen
University of South Carolina, USA

In the general context, metadata is data about data. For example, the metadata for a phonecall would include information about the sender, receiver, location, date, time, and duration. The library card catalog is the most common example of a metadata collection and the Dewey Decimal System is a familiar classification system for describing specific items. Nutritional labels on packaged food products are mandatory metadata descriptions. In a digital Internet environment metadata provide the basis for search, interchange, and even mapping. Therefore standardized metadata entries have emerged in almost every industry. For example, the Exchangeable image file format (Exif) adopted by the photography industry provides fields for latitude and longitude that can be added from a GPS unit in a camera or with GIS software.

The widespread emergence of geographic information system (GIS) files, geospatial databases, and Earth imagery, and the desire to exchange them, prompted the development of standardized geospatial metadata. According to the Federal Geographic Data Committee (FGDC), geospatial metadata represent the who, what, when, where, why, and how of a geospatial data set (FGDC n.d.a). Generally, the metadata for a resource should enable a potential user to determine its fitness for use for a specific application. This would include the following types of information:

- means of creation of the data;
- purpose of the data;
- time and date of creation;
- creator or author of the data;
- location on a computer network where the data were created;
- standards used.

Beyond high-level characterizations, the metadata should provide information about the technical and legal details pertaining to acquiring and using the data. In a similar way to the library patron performing a search on a digital card catalog, the GIS user can use metadata about a geospatial data asset to determine whether to acquire it.

The clearest legal statement regarding geospatial metadata was issued by the US Office of Management and Budget as part of the 2002 Circular No. A-16 Revised, “Coordination of Geographic Information and Related Spatial Data Activities”:

Metadata are information about data and/or geospatial services, such as content, source, vintage, spatial scale, accuracy, projection, responsible party, contact phone number, method of collection, and other descriptions. Metadata are critical to document, preserve and protect agencies’ spatial data assets. Reliable metadata, structured in a standardized manner, are essential to ensuring that geospatial data are used appropriately, and that any resulting analysis is credible. Metadata also can be used to facilitate the search and access of data sets or geospatial services within a Clearinghouse or data library. All spatial data collected or derived directly or indirectly using federal funds will have FGDC metadata.

(Office of Management and Budget 2002)
GEOSPATIAL METADATA

During the 1980s the growth of GIS technology was supported by the large-scale conversion of traditional maps by federal agencies such as the US Geological Survey and the Bureau of the Census. At the same time, state and local governments began to develop their data. Most of this data creation was intended for internal use, with little consideration for data exchange. In fact, data exchange was hindered by complex vendor-specific proprietary data formats. Furthermore, there was little consideration for spatial data quality or documentation. When GIS implementation migrated from mainframe computers to minicomputers, workstations, and ultimately personal computers, new users often wished to incorporate existing GIS data. This expansion encouraged partnerships based on shared approaches to data creation and distribution. The notion of a spatial data infrastructure gained important legal status in the United States with President Clinton’s 1994 Executive Order 12906 to create the National Spatial Data Infrastructure (NSDI) (Federal Register 1994). This executive order defined the NSDI as “the technologies, policies, and people necessary to promote sharing of geospatial data throughout all levels of government, the private and non-profit sectors, and the academic community.” The goal of the infrastructure was to reduce duplication of effort between agencies, to improve quality and reduce costs related to geographic information, to make geographic data more accessible to the public, to increase the benefits of using available data, and to establish key partnerships with states, counties, cities, tribal nations, academia, and the private sector to increase data availability. From the beginning, metadata was viewed a critical part of the NSDI. In fact, Executive Order 12906 included a timetable for implementation:

Beginning nine months from the date of this order, each agency shall document all new geospatial data it collects or produces, either directly or indirectly, using the standard under development by the FGDC, and make that standardized documentation electronically accessible to the Clearinghouse network.

(Federal Register 1994)

The FGDC Standards Working Group established the first Content Standard for Digital Geospatial Metadata in 1994 and the current standard in 1998 (FGDC 1998). The standard describes the data elements and provides details about how they were produced. The general information contained in the metadata standard fall into the following categories.

- **Identification**: What is the name of the data set? Who developed the data set? What geographic area does it cover? What themes of information does it include? How current are the data? Are there restrictions on accessing or using the data?

- **Data quality**: How good are the data? Is information available that allows a user to decide if the data are suitable for his or her purpose? What is the positional and attribute accuracy? Are the data complete? Were the consistency of the data verified? What data were used to create the data set, and what processes were applied to these sources?

- **Spatial data organization**: What spatial data model was used to encode the spatial data? How many spatial objects are there? Are methods other than coordinates, such as street addresses, used to encode locations?

- **Spatial reference**: Are coordinate locations encoded using longitude and latitude? Is a map projection or grid system, such as the State Plane Coordinate System, used? What horizontal and vertical datums are used? What parameters should be used to convert the data to another coordinate system?

- **Entity and attribute information**: What geographic information (roads, houses,
The creation of XML-structured metadata for geospatial data and services has facilitated discovery across the Internet. Metadata linked to servers provide hyperlinks to online resources such as map services, data download locations, data access services, and even applications. Web-crawling tools can utilize these XML files to search for data, harvest them, and create organized collections of data themes. These collections or clearinghouses provide a larger “card catalog” for data publication and discovery. Since all clearinghouse servers are peers, it is possible for communities to form common metadata services. These servers may be installed at local, regional, or central offices. In some cases, these clearinghouses serve as access points for data download. In others, the metadata provides a link to the owner who controls access. While many organizations freely expose their data and encourage download, others may require payment and licensing agreements. For example, the GIS Inventory System established by the National States Geographic Information Council (NSGIC) organizes the metadata provided by state and local governments to build statewide spatial data infrastructures (NSGIC n.d.). As part of its mission, the FGDC created many of these distributed clearinghouses that conformed to specified formats. Now, using the Geospatial Platform, users can search the metadata for US federal geospatial assets. From this cloud-based portal, a user not only has access to the description of the data but can often download it or incorporate the data into a project as a service without needing to duplicate the data (FGDC n.d.b).

The Content Standard for Digital Geospatial Metadata (CSDGM) of 1993 became a model for international standards efforts. In 1996 the International Standards Organization (ISO) began an effort to create an international metadata standard. The ISO process took several years of
GEOSPATIAL METADATA

consensus building by representatives of several nations. This effort resulted in the 2003 release of ISO 19115 Geographic Information – Metadata (ISO 2003). The emergence and adoption of an international standard for metadata was a major step forward in the maturity of geospatial disciplines. The international metadata standard incorporated the basic objectives for data discovery, fitness for use, data access, and transfer. It retained the critical information about spatial extent, quality, temporal dimension, and distribution for individual data features, as well as datasets and series. The developers of the ISO standard sought to expand metadata to a wide range of geographic products, such as web services, data catalogs, and clearinghouse activities. It incorporated open domains and optional elements that enabled greater flexibility than the United States standard individual dataset descriptions. While the CSDGM was written as a text document, ISO metadata was created using the Unified Modeling Language (UML), which can be represented in a consistent XML format. This effort has been fostered by the Open Geospatial Consortium (OGC), which released Geography Markup Language (GML), an XML grammar for geospatial features (OGC n.d.).

In 2010 the FGDC adopted the following components of ISO 19115.

- ISO 19115 – 2: 2009 Geographic information – Metadata – Part 2: Extensions for imagery and gridded data, including the base ISO 19115 metadata standard plus extended elements for the description of imagery, gridded data and data collected using instruments, e.g. monitoring stations and measurement devices …
- ISO 19110: 2005 Geographic information – Methodology for Feature Cataloging: An affiliate standard that supports the detailed description of feature types (e.g., roads, rivers, classes) in a manner similar to the CSDGM Entity/Attribute Section. The standard can be used in conjunction with ISO 19115 to document geospatial data set feature types or independently to document data models or other feature class representations.
- ISO 19119: 2005 Geographic information – Services – Amendment 1: Extends the service metadata model and serves as an affiliate standard that supports the detailed description of digital geospatial services including geospatial data portals, web mapping applications, data models and online data processing services. The standard can be used in conjunction with ISO 19115 to document services associated with a specific data set/series or independently to document a service.
- ISO 19139: 2007 Geographic information – Metadata – XML schema implementation: An XML document that specifies the format and general content of an ISO 19115 the metadata record. (FGDC n.d.c)

Recently, representatives from the United States and Canada developed the North American Profile (NAP) for the ISO metadata standard. The creation of the NAP represents a cooperative effort to establish a common geographic metadata standard that meets the needs of both nations and supports the North American Free
GEOSPATIAL METADATA

Trade Agreements (NAFTA). The NAP conforms to the ISO standard but is considered to have a more complex and robust structure than ISO 19115. Specifically, it has fewer mandatory and more optional elements, and also incorporates extended as well as new elements to capture more specific information. It uses a hierarchical structure that creates packages of metadata that can be reused and combined to form new metadata records. It also provides support for the documentation of new geospatial data topologies and technologies including geodatabases, web-mapping applications, data models, and data portals. Finally, it also provides opportunities to include suggested best practices for populating metadata elements in a manner that enhances the quality and usefulness of the metadata.

One of the strong points of the NAP is the inclusion of topic categories that are not part of the CSDGM metadata records. The following set of 19 high-level subject categories provides a standardized means of quickly sorting and accessing thematic information: farming; biota; boundaries; climatology meteorology atmosphere; economy; elevation; environment; geoscientific information; health; imagery base maps earthcover; intelligence military; inland waters; location; oceans; planning cadastre; society; structure; transportation; and utilities communication.

In the long term, the NAP of ISO 19115 will become the stable metadata standard for North America. The FGDC is currently working to educate users about the NAP and to accelerate its adoption. A major component of this effort includes the establishment of a Metadata Transform Working Group that is led by the National Oceanic and Atmospheric Administration’s (NOAA) National Coastal Data Development Center (NOAA n.d.). This group is developing software tools that ease the transformation of existing metadata files into the NAP. These developments are based on Extensible Stylesheet Language Transformations (XSLT) which is a language for transforming the XML documents. The GIS industry has matured since the mid-1990s. Both producers and consumers have recognized the benefit of documenting geospatial data and services. Prompted by executive-level mandates, the FGDC developed and promulgated the use of standardized metadata. Backed by software developers, the completion of metadata has become a critical part of GIS best practices. XML-based metadata has enabled the creation of web-based clearinghouses that are used to discover and acquire a wide range of geospatial resources. The structure of geospatial metadata is now an international standard that facilitates the development and exchange of data. The adoption of a specific multinational profile for North America is a further sign of the increased importance of geospatial data and services for commerce, logistics, and policy.

SEE ALSO: Data quality standards; Geographic data mining; Geographic information science; Geographic information system; GIS: history; Metadata; Spatial data infrastructures

References


GEOSPATIAL METADATA


The Semantic Web was brought to life in May 2001 following the publication of a highly influential article in *Scientific American* by Tim Berners-Lee, James Hendler, and Ora Lassila (2001). The authors were certainly well known. After all, Sir Tim Berners-Lee was the inventor of the World Wide Web, Hendler was a professor of artificial intelligence who pioneered web-scale knowledge representation languages and was managing the DARPA (Defense Advanced Research Projects Agency) Agent Markup Language (DAML) program, and Lassila was serving on the World Wide Web Consortium’s (W3C) advisory board and had just finished the specifications for the new machine-understandable Resource Description Framework (RDF). Their article outlined a vision for a next-generation web-based infrastructure in which content would be described semantically. This would make it accessible to humans and machines alike in a way that would empower intelligent personal assistant systems to draw conclusions from existing data and to connect multiple data sources on-the-fly to perform complex tasks.

Interestingly, the first example given in the *Scientific American* article is a geospatial one. The authors envisioned a Semantic Web agent that is able to check for local health-care providers within a 20-mile radius of a home by knowing and understanding a user’s health coverage plan. More than a decade later, semantically enabled intelligent personal assistants such as Apple’s Siri are available on most modern smartphones, millions of webpages contain semantic markup, and Google maintains a knowledge graph to enrich query results by directly answering questions instead of merely returning a ranking of webpages.

To understand the value proposition of the Semantic Web as well as design decisions made for its various technologies and standards, it is important to briefly revisit the history of the field. According to Berners-Lee, the reason for the success of the World Wide Web lies in its decentralized nature, its focus on being a social medium, not merely a technological one, and its unique perspective on robustness and reliability. The early web had no built-in mechanism to rate the quality or veracity of content, no hierarchical structure that would favor one site over the other, and no means to ensure that contents were readily available, no matter how important they were. In other words, information on the web can be contributed by any person and about any topic; unavailable sites and servers are a necessary evil but do not affect the overall infrastructure.

Instead of the design-oriented Hypertext Markup Language (HTML) used to create webpages, the vision of a Semantic Web required novel content-oriented languages that would not only describe the resources embedded in webpages, but also make those descriptions understandable by machines in a way that would allow them to infer new information, to combine statements about multiple resources, and to communicate across software agents and humans. Such languages would need to represent entity
data (e.g., the locations of specific health-care providers), as well as model knowledge about types, properties, and relations between entities (e.g., that insurance plans have different coverages with respect to health-care providers). It became clear early on that the principles that made the web successful would also have to guide the Semantic Web and its knowledge representation languages.

For example, an intelligent personal assistant must be able to deal with incomplete and even contradictory information. Similarly, and in clear contrast to querying databases, an agent should not infer that a certain statement does not hold true due to a lack of information. To give a concrete example, a database of US state capitals will answer a query for whether Santa Barbara, California, is a state capital with no as this record is not present in the database. On the Semantic Web, however, we cannot assume that all resources are currently present and accessible. Thus, an agent would only infer that Santa Barbara is not a state capital if (i) this negated statement is explicitly present in some data source, or (ii) if the definition of the type CapitalCity specifies that there can be only one capital per US state and Sacramento is already defined as California’s capital, and if the agent can exclude that Santa Barbara is an alternative name for Sacramento. This assumption is often referred to as the open world assumption (OWA), while the alternative approach typically used in databases is known as the closed world assumption (CWA). Similarly, the assumption that different names, such as Santa Barbara and Sacramento, always also point to different entities is known as the unique name assumption (UNA). As argued above, knowledge representation languages for the Semantic Web do not make this assumption.

With respect to veracity and handling contradictory information, the Semantic Web again makes use of the lessons learned from the World Wide Web. On the web, anyone can say anything about any topic, at any time and from anywhere. This is often called the AAA (or AAAAA) slogan. Therefore, the Semantic Web describes resources in terms of statements about them, not in terms of (true) facts. Additionally, reasoning services focus on drawing inferences instead of checking constraints. For example, a Semantic Web reasoner can infer the new statement that Santa Barbara is part of Germany from the two statements that Santa Barbara is part of California and that California is part of Germany, as well as from the terminological knowledge that partOf is a transitive relation. Whether the initial statements and, thus, the newly inferred statement are actually true is out of scope for Semantic Web reasoners. If necessary, the trustworthiness of a statement can be assessed by examining additional provenance information, restricting the query to certain established datasets, or by comparing the statement to other statements about the same entities from other datasets.

This raises the question of how the Semantic Web ensures interoperability, that is, how statements from different sources can be combined in a meaningful way. Syntactic interoperability, that is, enabling the exchange of data via common data formats, is handled by a set of open and freely available standards for knowledge representation languages, exchange formats, and serializations defined by the W3C. The Semantic Web’s approach to semantic interoperability follows a well-known paradigm from artificial intelligence research, namely to ensure consistency on the local knowledge-base level while allowing for inconsistencies on the global level. The aforementioned knowledge representation languages enable data providers, (research) communities, and individuals to formally specify the meaning of the used vocabulary, thereby restricting the possible interpretation of terms toward their intended meaning (semantics).
These specifications of conceptualizations are called ontologies.

For example, one can specify that the resource Santa Barbara is of type City and that cities are defined as permanent human settlements that are larger than towns. While such descriptions are meaningful to humans, the used terms do not mean anything to computers. Therefore, ontology languages are equipped with a model-theoretic formal semantics that defines in mathematical terms how to draw valid logical consequences from an ontology. For example, to encode that all cities are human settlements in a machine understandable fashion, one would make use of the subClassOf relation that is formally defined as a reflexive and transitive relation which implies that the extension of one class is a subset of the extension of another class. As being a city now logically entails being a settlement, the two axioms (i) Santa Barbara is of type City, and (ii) City subClassOf Settlement, imply a third axiom as a logical consequence, namely (iii) that Santa Barbara is of type Settlement.

Using such axioms, data providers can annotate their data and thereby enable humans and machines to interpret the used terms. While each (local) ontology should be free of contradictions, there is no need for universal agreement. For instance, in the state of California, the concepts of City and Town are synonyms and thus the above definition of cities being larger than towns does not hold, while it is true for Utah. In a web-scale environment, the role of Semantic Web technologies is not to enforce a single view on the world but to make different views and their underlying conceptualizations explicit.

Knowledge representation languages

The Resource Description Framework (RDF) (Brickley and Guha 2014; Cyganiak, Wood, and Lanthaler 2014) is a language for representing statements in the form of so-called triples of the form subject–predicate–object, where each of the entries is identified by a Uniform Resource Identifier (URI). RDF predicates are also called properties. In many cases, the object of one triple is the subject of another triple and, as a consequence, multiple triples form a structure that is often referred to as an RDF graph, namely a network of interconnected statements. An example for such triples would be adl:SantaBarbara adl:partOf adl:California, where adl is a so-called namespace identifier which expands to a proper URI such as http://adl-gazetteer.geog.ucsb.edu/. RDF consists of basic language constructs such as rdf:type which declares that a given resource is of a certain type, rdfs:subClassOf and rdfs:subPropertyOf with the meanings specified above, as well as rdfs:domain or rdfs:range to define the domain and range of relationships (RDF predicates). RDFS is the RDF schema language. With respect to the previous example, three statements could be encoded as follows: adl:SantaBarbara rdf:type adl:City, adl:City rdfs:subClassOf adl:Settlement, and adl:SantaBarbara rdf:type adl:Settlement. The resulting RDF graphs often consist of millions of triples that represent facts about information resources, such as webpages, as well as non-information resources, such as cities. The W3C SPARQL (SPARQL Protocol and RDF Query Language) standard serves as a query language over RDF graphs stored in a so-called triple store. Simplifying, SPARQL and triple stores can be thought of as the Semantic Web counterparts of the Structured Query Language (SQL) for databases. However, there are substantial differences which cannot be discussed here in detail. For instance query federation, that is, searching over multiple triple stores, is out of scope for SQL but key to querying the distributed Semantic Web.
While RDF makes it possible to express relatively simple statements, such as the one above, it is often necessary to deal with more complex knowledge and inferences which cannot be expressed satisfactorily in RDF. For this purpose, the W3C has standardized the Web Ontology Language (OWL) (Hitzler et al. 2012). The formal basis of OWL is mathematical logic, more precisely a subset of first-order predicate logic with counting quantifiers commonly known as a description logic (Hitzler, Krötzsch, and Rudolph 2010). OWL is partly compatible with RDF, and its capabilities are exemplified here by extending the example presented above. It could, for example, be formally represented that \text{SB} and \text{adl:SantaBarbara} are two names for the same thing, which would yield the inference that \text{SB} is a settlement (because \text{adl:SantaBarbara} is). As another example, consider the sentence “Santa Barbara and Sacramento are different,” together with “Each state has exactly one capital,” “California is a state,” and “Sacramento is the capital of California” – the last two can in fact be expressed in RDF as well as in OWL, while the first two can be expressed in OWL but not satisfactorily in RDF. In this case, we would be able to infer that Santa Barbara is not the capital of California. Inferencing (or reasoning) can be automated, although in general it is a computationally very intensive task. OWL provides a significant number of logical language constructs, including conjunction (logical \text{and}), disjunction (logical \text{or}), and logical negation (\text{not}), as well as quantifiers and counting quantifiers and a few more constructs from the realm of logical knowledge representation.

Linked data

The ecosystem created by hundreds of interlinked, semantically annotated, and openly available RDF-based data hubs is called the linked data cloud (Bizer, Heath, and Berners-Lee 2009) and contains billions of triples from a variety of domains including prominent geographic datasets, for example, digital gazetteers. At the same time, the term “linked data” also refers to design principles and specific technologies used to create this cloud. Essentially, and in contrast to today’s document web, the web of linked data employs named and directed links between typed data in the form of statements. Instead of being stored in closed data silos, linked data are globally referenceable by URIs, and, thus, triples can be constructed over multiple hubs, for example, to state that a certain resource called Brussels in the ADL dataset is the same (\text{owl:sameAs}) resource referred to as Bruxelles in a French gazetteer. The combination of URIs, formal ontologies, and open data, also enables federated queries for all relevant statements about Brussels, no matter whether they originate from a historical gazetteer, an administrative source from a government, or volunteered information, for example, from LinkedGeoData, which offers linked data access to OpenStreetMap.

The four main design principles of linked data are as follows.

1. URIs should be used to denote things, including noninformation resources, such as cities or mountains.
2. HTTP URIs should be used so that things can be referred to (e.g., linked) and dereferenced (looked up) by humans and machines alike. An appropriate form of representation should be displayed based on content negotiation, for example, to present an HTML file to a human user that is visiting a URI, while RDF should be served to a machine.
3. W3C-based standards such as RDF or OWL should be employed to provide information about things when their URIs are dereferenced.
4 Data should link to other data, by using their URIs to develop a densely interconnected global knowledge graph.

The successful publication of proper linked data, however, requires a collection of further technologies and strategies (e.g., for minting URIs) that cannot be discussed here in detail. In general, the quality and usefulness of the data will largely depend on the quality of the defined ontology (vocabulary) as well as on additional information that provides a context for its proper usage, such as provenance information and other metadata about RDF datasets.

The value proposition of the Semantic Web

While the previous paragraphs outlined the history of the Semantic Web and illustrated selected technologies, they do not explicitly state the value proposition of the Semantic Web and linked data. The main goal of the Semantic Web is to foster the publication, retrieval, reuse, and integration of data without restricting its heterogeneity. More specifically, Semantic Web technologies break up data silos and enable federated queries over multiple data sources, ease data conflation and co-referencing by providing globally unique and dereferenceable identifiers in the form of URIs, provide human- and machine-readable knowledge representation languages to encode statements and conceptual models, thereby enabling machine reasoning, improve interoperability beyond syntax, support the development and deployment of sustainable cyberinfrastructures via an open standardization process, and provide semantically rich querying capabilities beyond simple keywords.

A major novelty introduced by the Semantic Web is the focus on making data smart instead of developing smart applications. The rationale behind this paradigm shift is the insight that smart data will make future applications more powerful, interoperable, robust, and reusable, while smarter applications do not improve data along the same dimensions (Janowicz et al. 2015). By empowering data and by moving the so-called business logics out of the source code, Semantic Web applications and infrastructures can be developed and deployed by using off-the-shelf software such as reasoners, query engines, triple stores, faceted browsing frameworks, and link discovery tools. As most Semantic Web technologies rely on open W3C standards, exchanging data between specific systems is usually possible with minimal effort. In combination with a variety of open source tools, this makes the Semantic Web (and linked data) sustainable for the future and is one reason for its increasing popularity among data providers and consumers.

Challenges in realizing the Geospatial Semantic Web vision

While Semantic Web technologies, linked data, and ontologies have become widely used and recognized in research, governments, and the industry, a variety of challenges remain to be addressed before the Semantic Web can truly become the next generation of the web. Some of these challenges are listed here, focusing particularly on those that are of interest to geographic information science (GIScience) and geography.

While the Semantic Web was envisioned with heterogeneity in mind, semantic heterogeneity remains one of the most challenging aspects of interoperability and cyberinfrastructure research. Currently, the geospatial semantics community is moving toward using so-called ontology design patterns as a strategy to develop a wide range of purpose-driven ontologies that can handle
different viewpoints and levels of granularity, yet still remain interoperable due to a commonly used set of patterns.

A related problem is the so-called knowledge engineering bottleneck: it is difficult to arrive at suitable formal definitions for domain vocabularies. A wide variety of methods and tools have been developed during the past few decades to help knowledge engineers and subject matter experts reach agreement and formalize scientific workflows, nomenclature, and so forth. Today, machine learning techniques and crowdsourcing are combined with established methods to scale knowledge engineering. Additionally, work on areas such as ontology alignment and matching reduces the need to arrive at common definitions.

With respect to linked data, the discovery of links between resources, co-reference resolution, federated queries, and learning and extracting linked data from semistructured and unstructured resources (e.g., textual description of places) are among the most pressing research areas. The handling of scale and temporal and spatial scopes of statements remain open research challenges as well.

Finally, from a more conceptual perspective, an overall framework for semantic reference systems that enables the precise and reproducible interpretation of all components of geospatial information in terms of their measurement scales and observation procedures is still missing.

SEE ALSO: Big data; Data structure: spatial data on the web; Spatial data infrastructures

References


Geostatistics
Phaedon Kyriakidis
Cyprus University of Technology

Geostatistics is a branch of spatial statistics that deals with the analysis of spatially distributed data. The formal origins of geostatistics can be traced in mining engineering applications in the 1960s, as well as in related work in forestry and meteorology (Cressie 1990). Since then, many theoretical developments have taken place and geostatistical applications have spread across numerous and diverse scientific disciplines, ranging from earth sciences to environmental and atmospheric sciences to socioeconomic applications. Important points in time in the history of geostatistics include the publication (in French) of the original work of Matheron in the 1960s (Matheron 1962; 1963) and its English follow-up (Matheron 1971), the more practical publication Mining Geostatistics (Journel and Huijbregts 1978), geared towards the mining industry, the best-selling and widely accessible book An Introduction to Applied Geostatistics (Isaaks and Srivastava 1989), Cressie’s classic treatise Statistics for Spatial Data (Cressie 1993), which disseminated geostatistics to more mainstream statistics, and the comprehensive geostatistical textbooks Geostatistics for Natural Resources Evaluation (Goovaerts 1997) and Geostatistics: Modeling Spatial Uncertainty (Chilès and Delfiner 1999).

The three major and broad thematic areas of geostatistics include: (i) the analysis and interpretation of spatial patterns in geographical data, particularly the analysis and modeling of spatial association; (ii) the integration of attribute data from different sources and resolutions towards the objective of spatial interpolation (or spatial prediction); and (iii) the assessment of uncertainty in spatial prediction and its propagation to outputs/results of spatial operations or of models with spatially distributed inputs and/or parameters. This entry provides an overview of the main concepts involved in these three thematic areas of geostatistics, highlights relevant applications, and discusses current trends and future directions.

Quantifying spatial association

Geography deals with the analysis and interpretation of spatial pattern, an integral aspect of which is the notion of spatial association, also known as spatial autocorrelation through Tobler’s first law of geography. In quantitative terms, spatial association can be expressed in its most basic form in terms of either similarity (correlation) or dissimilarity (variation) between attribute values measured at different geographical units (also termed supports), for example, rain gauges (points) or pixels of a remotely sensed image (regular tessellation). In geostatistics, and for the case of attributes of interval or ratio measurement scale, spatial association is typically quantified by the sample or empirical semivariogram (Isaaks and Srivastava 1989), formally defined as the expected value of squared semi-difference between pairs of attribute values. To compute such an expected value, that is, an average quantity, pairwise squared semi-difference values are grouped into distance classes based on the (Euclidean) distances between their corresponding measurement locations. The average squared attribute semi-difference for each group is then
Figure 1 Example of a sample or empirical semivariogram (blue asterisks) of precipitation data (measured in mm) obtained from 77 rain gauge stations in Northern California; the number of data pairs used to compute the average semi-variance within each distance class is displayed on top of the asterisks. The red curve corresponds to a fitted semivariogram model with the following parameters: total sill = 8.5 mm², range = 120 km, and nugget contribution = 0.5 mm²; see text for details.

plotted against the average pairwise distance for that group; the resulting plot is termed the “sample semivariogram plot” (see Figure 1). For simplicity, the term variogram, that is, twice the semivariogram, is also widely used in the geostatistical literature.

Alternatively, the plot of average similarity between attribute value pairs in different distance classes versus average pairwise distance is termed a “sample covariogram” or “sample correlogram,” depending on whether pairwise attribute similarity is expressed as (i) the product of the differences between two attribute values from the attribute mean, or (ii) its standardized version dividing the former by the attribute variance.

For attributes exhibiting positive spatial autocorrelation, the sample (semi)variogram is expected to reveal an increasing pattern with distance; the reverse pattern is expected for the sample covariogram and correlogram. It is the particular shape of the spatial association plot, that is, how quickly (in terms of distance) the correlation reaches the zero value or what the shape of the plot is at small distances, that provides a first step towards a quantitative version of Tobler’s first law of geography. Evidently, different attributes and different data sets will result in different sample spatial association plots; it is the expected overall shape that Tobler’s first law postulates, not the nuances of each particular plot.

It is often desirable to also consider the orientation of the vectors defined between pairs of data locations in addition to the length of such vectors, that is, in addition to distance alone. In this case, orientation classes are also used for grouping pairwise (dis)similarity values; the resulting plots are termed “directional sample” variograms or correlograms and are used to elucidate different patterns in attribute variation in different directions, that is, to reveal anisotropic (as opposed to isotropic) spatial variation. The term “omnidirectional” variogram or correlogram is often used to indicate that the grouping of pairwise (dis)similarity values is only based on distance and not on orientation. Other types of spatial information, for example, altitude or land-use type, regarding the data locations and their surroundings or environmental conditions, can also be used to produce more meaningful groupings of pairwise (dis)similarity values depending on the particular application at hand.

In the case of two attributes, for example, temperature and elevation, the above spatial association measures can be extended to the bivariate case, termed “sample cross-variograms” or “sample cross-correlograms”; the variograms or correlograms of temperature data and of elevation
data then receive the prefix “auto” in their respective terminology. A cross-correlogram essentially involves a series of correlation coefficients computed from lagged data of two different attributes, for example, from measurements of elevation and temperature considered at different locations. The classical correlation coefficient between collocated measurements of, say, temperature and elevation is then regarded as the value of the cross-correlogram at zero distance. The case of more than two attributes involves multiple combinations of auto- and cross-measures of spatial association, and often becomes cumbersome to handle, particularly in the case of many (>3) attributes. Alternatively, multivariate spatial association can be quantified by noting that a squared difference between attribute values constitutes a squared Euclidean distance metric in attribute space. The multivariate sample variogram is then expressed as the sum of all sample auto-variograms of different attributes, and quantifies overall dissimilarity across multiple attributes as function distance. As a simple summation of auto-variograms ignores the correlation across attributes, a more representative measure of multivariate spatial association involves the Mahalanobis distance metric for describing pairwise dissimilarity. In this case, the resulting multivariate measure of spatial association involves a weighted average of all sample auto- and cross-variograms. Such measures of multivariate spatial association are widely used in ecology, under the umbrella term “Mantel correlogram,” to quantify spatial association in, say, abundance of different species measured at different sites; here each species is regarded as a different attribute and the corresponding abundance values are considered data of multiple attributes (Legendre and Legendre 2012).

In the case of nominal (categorical) data, for example, soil types or land cover classes, spatial association is quantified using a binary (0/1) indicator transform of the categories observed at the data locations. In other words, each category corresponds to a set of binary indicator data, and there are as many such sets as categories. In geostatistics, indicator variograms and indicator correlograms have been traditionally used to quantify spatial association for each category (Goovaerts 1997). The analysis of categorical data, however, is an inherently multivariate problem, since the indicator data for each category are essentially data on a different binary attribute. Consequently, the quantification of spatial association in categorical data also calls for the analysis of bivariate indicator cross-variograms and cross-correlograms for each pair of categories. A more natural alternative for describing spatial variation in categorical data is transition probabilities, quantifying the probability of transitioning from a location with one category to another location with the same or a different category as a function of distance (Chilès and Delfiner 1999). The set of all auto- and cross-transition probabilities for a particular distance class is functionally linked to the indicator auto- and cross-covariance values for that distance, and provides a coherent means for describing joint (between the same category and across different categories) spatial variation in nominal data.

All measures of spatial association described above are characterized as global measures, in that they involve averages of pairwise (dis)similarity values across all data pairs within a particular distance class. Local measures of spatial association can be defined as average pairwise (dis)similarity values involving each data value and its neighbors; in this case, the computation of pairwise association is limited to only those location pairs within some distance class from each data location. The result is a map of local association measures, for example, local variogram values, whereby a local association value corresponds to each data location. In the multivariate case,
local multivariate variograms have been used, particularly in remote sensing, for quantifying multivariate image texture. In this case, the reflectance values recorded at different spectral bands constitute data of different attributes, and the objective is to furnish a model of image texture across multiple spectral bands; such texture information is often used for enhanced classification of remotely sensed data.

Modeling spatial association

The sample semivariogram or correlogram is often computed from a limited set of sample data, and represents averages of pairwise (dis)similarity values for each distance class; such empirical measures are therefore subject to sampling fluctuations or error. In addition, semivariogram or correlogram values are only available for a limited set of distance classes; one cannot simply consider more classes, because this leads to smaller classes, hence to fewer pairs of attribute values within each class and consequently more fluctuations and less reliable (dis)similarity averages. What is required is a model of spatial association as a continuous function of distance; this model actually pertains to the continuous spatial distribution of attribute values of the population from which the particular sample data set was derived (Isaaks and Srivastava 1989). However, not any function of distance can be used as a model of spatial association. Permissible or valid covariance models, for example, ensure that any weighted linear combination of covariance values is non-negative, no matter the weights involved. This requirement stems from the fact that such a weighted combination expresses variance, and variance cannot be negative; functions that satisfy this requirement are called positive definite functions.

Several parametric functions of distance, such as the spherical, exponential, or Gaussian functions, are known to be positive definite and are widely used as correlogram or semivariogram models in practical applications (Chilès and Delfiner 1999). The shape of the function at small distances, for example, linear or parabolic, is linked to the smoothness of the postulated attribute surface; a Gaussian model, for example, implies a very smooth attribute surface such as atmospheric temperature or pressure. Apart from function shape, typical model parameters in the isotropic case include: (a) the sill, linked to the variance of the sample data, and (b) the range, that is, the distance at which the sill is reached and beyond which spatial correlation is expected to be nil. For models reaching their sill asymptotically, such as the exponential and Gaussian functions, the range is often defined as the practical range, that is, the distance at which 95% of the sill is reached (see Figure 1). More flexible (and more complex) models are available, for example, the Matérn model, which also include a third parameter that modulates the shape of the semivariogram function at short distances; the Matérn model, for example, yields as particular cases the Gaussian, exponential, and power law semivariogram models.

Anisotropic semivariogram models also exist for furnishing different patterns of attribute spatial variation in different directions in a consistent manner (Isaaks and Srivastava 1989). The type of anisotropy whereby the sill of the semivariogram model attains the same value in all directions, but the locus of equal range values in different directions forms an ellipse in two dimensions (2D) or an ellipsoid in three dimensions (3D), is termed “geometric anisotropy.” In this case, an affine transformation (rotation and scaling) is first applied to the vectors defined between any two locations so that isotropic functions can subsequently be employed; the parameters of
the affine transformation are linked to the range of the model in different directions and to the orientation of the major and minor axes of the ellipse in 2D.

More complex parametric models of spatial association can be built as weighted linear combinations of permissible (and possibly anisotropic) nested models (Goovaerts 1997). Such composite models correspond to a postulated attribute surface comprised by an additive superposition of independent nested component surfaces. Each constituent surface is characterized by its own semivariogram model and the (partial) sill of that model expresses the contribution of the variance of that particular nested surface to the overall variance of the composite attribute surface. The composite model used most often in practice corresponds to an arbitrary semivariogram model plus a nugget effect component (see Figure 1). The nugget effect appears as a discontinuity of the semivariogram model at very small distances and corresponds to random attribute spatial variation, measurement error, or spatial variation occurring at scales smaller than the smallest distance between data locations. Composite semivariogram models are also used as zonal anisotropy models, a type of anisotropy whereby an extra component of spatial variation appears only in a particular direction; zonal anisotropy is handled in practice by including in the composite semivariogram model a nested model with geometric anisotropy with very large (infinite) range in one direction.

In the multivariate case, joint models of spatial association are needed to provide consistent modeling of all auto- and cross-variograms defined between multiple attributes. One cannot simply postulate a permissible auto- or cross-variogram model for each sample auto- or cross-variogram, as the entire set of all such univariate and bivariate models must be jointly permissible; the intrinsic coregionalization model (ICM) and the linear model of coregionalization (LMC) provide two solutions to this problem, the latter being the most widely used model in practice (Goovaerts 1997). Last, nonparametric spatial association models can also be constructed, but those are less frequently used in practice as they are rather difficult to generalize, particularly in the multivariate case.

In many practical applications, a parametric variogram model is first selected for the postulated attribute surface, and model parameters are subsequently estimated so that the discrepancy between the model and sample variogram values is minimized. The fitting procedure is typically carried out through iterative numerical optimization algorithms, such as weighted nonlinear least squares with more importance (larger weight) given to the fit at small distances. Other candidate variogram models can be considered and the fitting procedure can be repeated for selecting the best model based on standard goodness-of-fit criteria, such as the root mean squared error or the Akaike information criterion (Chilès and Delfiner 1999). Such a semiautomatic procedure is often encouraged, yet caution has to be exercised to avoid erroneous and physically unjustified results, particularly in data-poor situations. The variogram model fitting step is a key component of any geostatistical study, but is also considered the “art” of geostatistical practice. Although automatic and semiautomatic fitting procedures are also available for the multivariate case, fitting a spatial association model characterizing jointly univariate and bivariate (dis)similarity across data of multiple attributes quickly becomes cumbersome as the number of attributes increases. This complexity in multivariate spatial association model fitting often leads to spatial regression models based on collocated (not lagged) data,
described in the section “Spatial interpolation and kriging,” to be preferred in practice in the multivariate case.

**Random fields and stationarity**

A variogram model fitted to a sample variogram pertains to the population of attribute values from which the sample data are thought to originate. In geostatistics, that unknown population is conceptualized as a field with a continuous spatial distribution, whereby distances between point locations (as opposed to objects such as polygons) are readily defined. The unknown spatial distribution of the population is regarded as a realization (an outcome) of a random function or a random field, that is, an infinite collection of georeferenced and spatially correlated random variables, one random variable defined per location (Isaaks and Srivastava 1989). In the multivariate case, postulated attribute fields of different attributes, for example, temperature, elevation, and solar radiation, are regarded as realizations of a multivariate random field. Multiple constituent random variables are now defined for each location, one variable per attribute; random variables of the same attribute at different locations are spatially auto-correlated, whereas random variables of different attributes at the same or different locations are spatially cross-correlated.

The random field concept is perhaps more easily conveyed via its discrete counterpart, the random vector, a finite set of georeferenced and spatially correlated random variables, through a spatiotemporal distribution analogy. Consider, for example, the case of a limited set of rain gauges recording daily precipitation values over some time period. In other words, consider the set of precipitation time series at the corresponding rain gauge locations, and assume for the moment that all time series are equally informed, that is, there are no missing values for any day. Invoking the random vector construct, for each rain gauge location there is a corresponding random variable and the rainfall value at each rain gauge is regarded as a realization (outcome) of that random variable for the particular day. The univariate statistics, for example, mean, of precipitation could in principle be different from one location to another, for example, due to orographic effects or distance from the coast, and consequently, one could define a location-specific (local) mean for each random variable. Likewise, the covariance or the correlation coefficient between two precipitation time series could vary across location pairs, depending on the environmental characteristics prevailing at any two rain gauges. In other words, the correlation coefficient between any two random variables could be different depending on the particular pair of locations corresponding to those variables.

Geostatistics, however, is typically employed for the analysis of spatial data; hence the spatiotemporal analogy described above is difficult, actually impossible, to operationalize in a purely spatial context. In the case of spatial data, there is only one available realization of the random variables at the data locations, this corresponding to the sample data values. Consequently, one cannot estimate a local attribute mean, or a pairwise correlation coefficient between attribute values at any two data locations. In the first case, there is only one attribute value at each data location, and in the second case, there is only one attribute pair at each pair of data locations. To circumvent this deadlock due to the lack of time replicates of attribute values, working assumptions are typically invoked, allowing the use of spatial replicates of attribute values from different locations for the purpose of computing
meaningful statistics (Journel and Huijbregts 1978).

First-order stationarity is a model assumption stating that all constituent random variables of a random field have the same mean (a first-order moment of a distribution); in other words, the attribute mean is location invariant. Second-order stationarity includes the assumption of first-order stationarity, and in addition states that: (i) the variance (a second-order moment of a distribution) of any random variable exists (is finite) and is constant - this variance value is termed the “a priori variance” of the random field; and (ii) the covariance and the correlation coefficient between any pair of random variables are only a function of the magnitude (and possibly orientation) of the vector linking corresponding locations at which those two random variables are defined. A weaker assumption than second-order stationarity is that of intrinsic stationarity, whereby: (i) the attribute mean is assumed to be locally constant - this implies that random variables at locations not too distant from each other are assumed to have the same mean; and (ii) the attribute variance need not be finite but could increase as the study region becomes bigger and more data become available. In practical terms, intrinsic stationarity is the reason that the sample variogram has emerged as the preferred measure of spatial association in geostatistics: it can still be defined for cases where the covariogram and the correlogram cannot, and it does not require knowledge of the attribute mean, which is difficult to obtain if it is deemed spatially varying. Last, the assumption of stationarity in one form or another is often invoked when performing basic geostatistical spatial interpolation under the umbrella term “kriging,” although several variants of kriging have been developed to relax that assumption (see below).

Spatial interpolation and kriging

Spatial interpolation is the procedure of estimating or predicting unknown values of an attribute of interest at nonsampled locations from known measurements of the same (in its simplest form) attribute at sample locations. Spatial interpolation is widely used for attribute surface creation in multiple fields, as it is often the case that the locations where interpolation is performed coincide with the nodes of a regular grid. The terms “source location” and “target location” are often used for the sample locations and the locations where interpolation is performed, to emphasize the transfer of information from the former to the latter; the corresponding attribute data and unknown attribute values at those two sets of locations are termed “source data” and “target values,” respectively. A plethora of spatial interpolation methods have been developed to date, each involving its own assumptions and procedures (Isaaks and Srivastava 1989). Two main families of spatial interpolation methods, however, can be distinguished: (i) deterministic interpolation methods, in which a unique attribute value is predicted at each target location, once the parameters of the particular method have been selected; and (ii) statistical or stochastic interpolation methods, in which a distribution of possible attribute values is furnished at each target location, given, again, all the parameters adopted from the particular method.

Deterministic spatial interpolation methods include, among others, nearest neighbor interpolation, bilinear interpolation, triangulated irregular network interpolation, inverse distance weighted (IDW) interpolation, and spline interpolation and its variants (cubic splines or thin-plate splines). Of these methods, perhaps the most widely used is IDW interpolation, as it is simple to explain, easy to implement, and widely available in commercial as well as open
source software packages. IDW is, in principle, a parametric spatial interpolation method, as it allows for different importance to be placed on the distance between the source and target locations via an exponent parameter; the most widely used values being 1 and 2, corresponding to simple inverse distance weighted interpolation and inverse distance squared interpolation, respectively.

Kriging is a geostatistical, hence stochastic, spatial interpolation method, which qualifies as the best linear unbiased predictor (BLUP), that is, under the least squared error or minimum error variance optimality criterion, the best predictor of a target value as a weighted linear combination of source data, given a known or previously estimated variogram or covariogram model (Cressie 1993). The term “kriging” was coined by Professor Georges Matheron as a tribute to the pioneering work of Dr Daniel Krige in the early 1960s. Similar to many deterministic spatial interpolation methods, kriging can be performed using all the source data available (global interpolation) or only a limited set of source data within a search neighborhood around the particular target location where interpolation is performed (local interpolation). In the latter case, the search neighborhood is defined based on the range of the (semi)variogram model; for anisotropic models, that neighborhood becomes an ellipse in 2D and an ellipsoid in 3D. In addition, and similar to other spatial interpolation methods, kriging is an exact interpolation method, in that the predicted target value at a target location coinciding with a source location yields the known source value at that location. Depending on the assumptions made regarding expected attribute values at the source and target locations, different variants of kriging have been developed in the literature; such variants furnish similar predictions in regions where source data are abundant and different predictions in regions away from the source locations (extrapolation regions).

In its most basic form, simple kriging (SK) assumes a known expected attribute value (mean) both at the source and at the target locations; in other words, SK requires knowledge of a location-specific (local) attribute mean (Chilès and Delfiner 1999). The target prediction is then expressed as a weighted linear combination of source data, plus a weight attributed to the known mean at the target location. The SK weights are obtained by solving a linear system of equations, known as the SK system, which involves source-to-source covariance values between source data locations and source-to-target covariance values between source and target locations. All these covariance values are derived from the covariance model, based on the length (and possibly orientation) of the vectors defined between location pairs. Similar to most methods of spatial interpolation, the SK weights account for the proximity of the source to the target locations. Contrary to other interpolation methods, however, the SK weights also account for the configuration of the source data locations. In addition, and through the particular covariance model adopted, the SK weights also account for the expected smoothness of the attribute surface, as well as for the redundancy of the source data – the contribution of source data located in clusters is reduced in the computed target value. The weight attributed to the known local attribute mean at the target location is the complement to 1 of the sum of the SK weights given to the source data. That known local attribute mean in SK need not be constant, it just needs to be known (hence certain) or previously estimated without considering the uncertainty inherent in such an estimated quantity.

Ordinary kriging (OK) is typically performed in a local interpolation mode and is the most
widely used kriging variant in practice when no auxiliary data on other attributes are considered in the interpolation procedure. In OK, a constant but unknown local attribute mean is assumed for the target location and all source locations within the search neighborhood; this amounts to the assumption of local first-order stationarity. That unknown mean is implicitly estimated in conjunction with the kriging weights using an augmented linear system, called the OK system. This system includes a unit sum constraint for the resulting weights, since the local attribute mean is not known as it is in the case of SK. Universal kriging (UK) is another variant of kriging, whereby the local attribute mean within a search neighborhood is expressed as a parametric linear function of coordinates (or polynomials thereof), as in the case of trend surface analysis. Additional linearity constraints are appended to the kriging system, resulting in a further augmented system, called the UK system, for estimating simultaneously the parameters of that linear function as well as the interpolation weights given to the source data within each search neighborhood. Under certain choices of the (semi)variogram model adopted, for example, the linear or power model, universal kriging predictions are equivalent to those derived via spatial interpolation with splines. Last, when the attribute data are expressed as a parametric linear regression model on data of other attributes within each search neighborhood, the corresponding kriging procedure is termed “kriging with external drift” (KED), and furnishes local estimates of the regression model parameters accounting for spatial autocorrelation in the regression error term (using generalized least squares), as well as SK weights for the spatial interpolation of the residuals derived at the source locations from that regression model.

In the local regression approach for spatial interpolation furnished by KED, the data of the auxiliary variables should be collocated with those of the attribute of interest at the source locations (for estimating the parameters of the regression model) and at the target locations (for predicting the corresponding target values). Such an isotopic or equally sampled design, however, is often not realized in practice; in this case, data on multiple attributes can alternatively be integrated in spatial interpolation via cokriging. The cokriging weights furnish the contribution of each source data value, be it of the same attribute as the primary attribute of interest or of a different attribute, to the target prediction as a function of auto- and cross-variogram values. The solution of the corresponding cokriging system of equations calls for a permissible joint model (such as the LMC) for all auto- and cross-variograms defined between all pairs of attributes involved. Augmented cokriging systems can also be defined, as in the univariate case for relaxing the assumption of a constant attribute mean within the search neighborhood for each of the variables involved in spatial prediction.

Several variants of kriging can be distinguished depending on whether: (i) the source data contain measurement error – in factorial kriging spatial interpolation that error component is assumed to be additive and independent of the signal (a measurement’s meaningful component) and is filtered out from (not added to) the predicted target values; in this case, kriging is not an interpolation method but a smoothing method, since source data values are not reproduced due to their measurement error characteristics; or (ii) a non-Gaussian assumption is justified for the attribute data – lognormal kriging, Poisson kriging, indicator kriging, and disjunctive kriging are some of the different extensions of kriging available for dealing with non-Gaussian data.

Perhaps one of the most important reasons behind the success of geostatistics and its wide
adoption in numerous and diverse scientific disciplines is its ability to cope with data of different spatial resolutions, in other words, its ability to integrate and utilize measurements pertaining to different supports (Journel and Huijbregts 1978). In punctual or point kriging, the most widely used mode of kriging in practical applications, the source data and target values pertain to the same point support. In block kriging, source data are typically of point support whereas target values pertain to pixels or polygons or blocks in 3D; in this case, target values are typically defined as (possibly weighted) averages of point values within their corresponding supports. In downscaling kriging, source data pertain to pixels or polygons whereas target values are defined on point supports. Last, in geostatistical areal interpolation, both source data and target values pertain to supports of arbitrary shape and orientation. In all the cases above, the form of the kriging system remains the same, but the different covariance values used in that system, as well as the corresponding linear equality constraints, are modified to account for the data resolution differences. It is worth noting that, in the downscaling case, kriging predictions when upscaled reproduce by construction the sample areal data used for downscaling, no matter the particular variogram model used; this implies that kriging furnishes resolution-consistent predictions.

Spatial uncertainty

Being a stochastic interpolation method, kriging provides an estimate of the prediction error variance at each target location. That prediction error variance is zero at the source data locations (except in the case of factorial kriging where that variance reverts to the measurement error variance) and increases away from the source data locations. In the second-order stationary case, the simple kriging prediction error variance at a target location further away than the variogram range from source locations tends towards the sill of the variogram model used. In general, that variance is larger for the case of ordinary kriging due to the additional uncertainty in the estimation of the local mean within each search neighborhood, and further increases for the case of UK and KED due to the additional uncertainty in estimating the parameters of a local regression model linking the attribute of interest to coordinates or other auxiliary variables.

The kriging error variance is homoscedastic in its simplest form, that is, it does not depend on the source data values but only on the relative position of the target location with respect to the source locations, on the configuration of the source locations, and on the particular variogram model used (Goovaerts 1997). This property renders the kriging error variance useful in sampling design applications: different locations can be selected as candidate locations for placing additional source data, the kriging error variance values are computed for at all target locations using each of the candidate locations as additional source location, and the best configuration of new source locations is selected so as to minimize, for example, the average kriging error variance or the maximum kriging error variance.

In the multivariate Gaussian case, the prediction error variance at a target location along with the corresponding kriging prediction at that location furnishes the parameters of a local Gaussian probability distribution of possible attribute values given the known source data, possibly including data on relevant auxiliary variables used for spatial prediction, as well as the particular variogram model adopted. That probability distribution is often used for propagating (either analytically or numerically through statistical simulation) the local uncertainty in
interpolated attribute values, for example, pollution levels, to quantify uncertainty in the results of local geographic information system (GIS) operations involving one grid node at a time, for example, the local probability for grid nodes in residential land cover type to exceed a particular pollution threshold.

Interpolated attribute surfaces, however, often either (i) undergo spatial operations involving more than two locations at a time, for example, slope computation from elevation or other focal or zonal operations in a GIS, or (ii) serve as inputs to environmental or socioeconomic models with spatially distributed inputs or parameters; for example, the spatial distribution of rainfall is a critical parameter in models of overland flow and runoff computations. In such cases, knowledge of the local kriging attribute prediction and variance at a set of target locations, considered one at a time, is not adequate for uncertainty propagation purposes. The preferred means for uncertainty propagation in this case is geostatistical Monte Carlo simulation (Goovaerts 1997). For example, in 3-D hydrogeological investigations involving flow and transport in heterogeneous porous media, the spatial distribution of hydraulic conductivity (a property indicating how fast solute can flow through the pores of a rock formation) is often parameterized in terms of a lognormal random field model. Realizations from such a random field constitute alternative numerical models of the 3-D spatial distribution of hydraulic conductivity, all of which reproduce: (i) any measurements of conductivity at their sample locations, (ii) the variogram model of that attribute, and (iii) any relationships with relevant auxiliary data, such as those provided by geophysical surveys. These alternative realizations are then used, along with physically based simulators of flow and transport, in a Monte Carlo framework for evaluating, for example, the uncertainty in the spatial distribution of solute concentration due to the uncertainty in the spatial distribution of hydraulic conductivity and possibly other relevant variables.

Trends and future directions

Geostatistics has spread through the years from its mining engineering origins to a wide spectrum of applications in the earth and environmental sciences – actually, in almost any field where spatial analysis is applied. Since its conception, geostatistics has been associated with practical engineering applications, and it is often the case that developments in geostatistical theory and applications are driven by engineering problem-solving. A major recent thrust of geostatistical development is driven by the need to characterize more realistically the complexity and heterogeneity of the spatial distribution of attributes, for example, rock type and porosity, involved in reservoir characterization in the oil and gas industry. This need has led to the advent of multiple-point geostatistics (Mariethoz and Caers 2014), whereby spatial patterns involving more than two points at a time (a variogram is a two-point statistic) are “learned” from training images; such images are numerical models, for example, scanned cross-sections of outcrops, remotely sensed images, even hand-drawn images by geologists, that represent prior (before data acquisition) repositories of spatial patterns. These learned spatial patterns, not falling in any parametric or nonparametric statistical model, are then “exported” to the real reservoir via simulation and fused with actual data to provide realistic models of spatial heterogeneity and complexity. Besides oil and gas applications, multiple-point geostatistical applications can be found in mining, hydrogeology, geomorphology, and remote sensing.
Recent developments in geostatistics have also addressed the need to include in the spatial prediction endeavor data on models that operate on (use as input) the spatial distribution of the attribute that is being predicted; for example, when interpolating hydraulic conductivity, data of pressure tests might be available. The task of accounting for model output data when predicting the spatial distribution of model inputs falls in the realm of inverse problems (Chilès and Delfiner 1999). Such problems have been traditionally studied in mathematics and physics, but fruitful recent developments have emerged from a geostatistical perspective, with applications in reservoir engineering, hydrology and hydrogeology, remote sensing, and atmospheric sciences.

The inclusion of time as an additional data and modeling component has also been one of the major areas of development in geostatistics (Cressie and Wikle 2011). Several space–time variogram and covariance functions have been proposed in the literature for modeling joint attribute variation in a spatiotemporal context. Dynamic space–time models linking partial differential equations with geostatistics have also been developed to account for the dynamic evolution of spatiotemporal processes. It is expected that more of these developments will sooner rather than later find their way into commercial, for example, GIS, software and consequently enable the even wider application of geostatistics.

Last, the advent of model-based geostatistics (Diggle and Ribeiro 2007) to address spatial data with non–Gaussian distributions is another notable theoretical development. Although non–Gaussian geostatistical models have been proposed in the past through the work of Mathéron on bivariate isofactorial models, this new general framework has been widely adopted particularly in spatial statistics and represents a major breakthrough. Along the same lines, Bayesian inference in a hierarchical framework has also been a major theoretical development, along with computational algorithms for rendering inference in such a framework feasible in practical applications.

Geostatistics has been dealing with large data sets from its conception; theoretical and practical developments throughout the years have been traditionally tailored to address issues pertaining to large data sets. New technologies and advances in computing and computational statistics, however, have been instrumental in bringing the field of geostatistics (and spatial statistics in general) closer to that of computer science. This convergence trend is expected to continue as new methodologies for big data analysis and modeling become available. Last, as is the case with many scientific fields, the dissemination of geostatistics in different disciplines has been immensely facilitated by the development of computer hardware and relevant software. As both open source and commercial geostatistical software become increasingly available, it is expected that geostatistical methodologies will find their way to an even wider spectrum of applications where geographical methods of spatial analysis and modeling are involved.

SEE ALSO: Geocomputation; Geographic information science; Interpolation: inverse–distance weighting; Interpolation: kriging; Measuring spatial dependence; Modifiable areal unit problem; Quantitative methodologies; Scale; Spatial analysis; Spatial resolution; Spatial sampling; Spatiotemporal analysis; Tobler’s first law of geography; Uncertainty

References


Further reading


Geotargeted alerts and warnings

Michele M. Wood  
*California State University, Fullerton, USA*

Historically, publics have been warned of imminent threat using outdoor sirens, electronic media, route alerts (wherein officials travel through the streets with loudspeakers), and tone alert radio (TAR). Recent technological advances have made it possible to deliver geotargeted alert and warning messages to mobile devices. In the context of risk communication alerts and warnings, geotargeting refers to efforts to transmit risk communication messages only to those individuals physically located within the geographic area at risk (National Research Council 2013). This new technology, in effect, can place a warning “siren” in the pocket of every individual who owns and carries a cellular phone or other mobile device. How this new technology may influence the manner in which people receive and respond to public warnings has a bearing on the further development of risk communication theory and also on emerging warning practice.

The need to warn the public about imminent threat has long been a challenge for human populations. Researchers have studied the effects of alerts and warnings in a disaster context for more than half a century. “Risk communication” refers to real-time information exchange between experts or officials and those who face a hazard or threat to their health or wellbeing; its purpose is to enable those at risk to make informed decisions to mitigate the effects of the threat and take appropriate protective action (WHO 2015).

The emergency alert system (EAS) is a national warning system in the United States that was approved by the Federal Communications Commission in 1994 and implemented in 1997, replacing the former emergency broadcast system (EBS, in place from 1963 to 1997), which, in turn, replaced the control of electromagnetic radiation system (CONELRAD, in place from 1951 to 1963). EAS is coordinated at the federal level by the US Federal Communications Commission (FCC), the US Federal Emergency Management Agency (FEMA), and the US National Oceanic and Atmospheric Administration’s National Weather Service (NOAA’s NWS). EAS is now one part of the US integrated public alert and warning system (IPAWS).

IPAWS is the nation’s infrastructure for delivering emergency alerts and warnings to the public. It represents a modernization and integration of the nation’s alert and warning infrastructure, allowing public safety officials to alert and warn the public about imminent threats using EAS, wireless emergency alerts (WEA), the NOAA weather radio, and other public alerting systems, using a single interface. IPAWS is managed by FEMA. In the event of a national emergency, the President may use IPAWS to communicate to the public with merely 10 minutes notice.

The WEA service allows customers who own “WEA capable” wireless telephones and other enabled mobile devices to receive geographically targeted alerts about imminent threats to safety in their geographic area. The WEA system was established pursuant to the Warning, Alert and Response Network (WARN) Act and delivers alerts using a special broadcast channel in cellular
GEOTARGETED ALERTS AND WARNINGS

telephone systems. WEA was originally called CMAS, the commercial mobile alert system, or PLAN, personalized local alerting network. Rollout of the WEA service began in April of 2012. Wireless service providers participate in WEA voluntarily as part of a public–private partnership between the FCC, FEMA, and the wireless industry. Alerts from authenticated public safety officials are sent through IPAWS to participating wireless carriers and are then pushed by the providers from cell towers to mobile devices that are physically located in the designated or affected warning area. The WEA service provides three types of messages: presidential alerts, imminent threats to life and safety, and AMBER alerts (child abduction bulletins by law enforcement). The messages appear like text messages and are limited to 90-characters. Initially, the NWS created and used templates for nine different types of WEA warnings. These were based on EAS messages, but edited to adhere to the limited WEA message length. WEA capable device users are enrolled in the WEA service as a default setting and can opt out of all but the presidential alerts. One of several components of IPAWS, the WEA service is designed to complement, not replace, other IPAWS supported warning communication channels.

The common alerting protocol (CAP) is a digital format for exchanging information that allows a consistent alert message to be disseminated simultaneously over multiple communications systems. With multiple systems using CAP, a single alert can trigger a wide variety of public warning systems, increasing the likelihood that members of the intended audience will receive the alert by one or more communication channels. CAP provides the capability to geographically target alerts to a defined area using federal information processing standard (FIPS) codes, which designate counties and subdivisions within counties to define the targeted warning area, or by using a list of polygon vertex coordinates to specify the boundaries of the targeted area. The NWS and wireless carriers have developed ways to define even smaller polygons.

A brief discussion of what is known about designing effective warning messages will provide useful context for understanding issues associated with geotargeted alerts and warnings.

Warning messages

Modern social science research on the topic of warnings began in the 1950s with the funding of two studies by the National Academy of Sciences. The first study (Wallace 1956) was an exploratory investigation of individual and community behavior in response to the devastating and unprecedented 1953 tornado in Worcester, MA. That study documented the absence of any official warning to the public about the tornado, as well as an associated 54 deaths and 438 major injuries. The investigator concluded that if an official warning had been issued to the public, up to 90% of the at-risk population might have taken cover, thereby greatly reducing morbidity and mortality caused by the tornado. The second study (Mack and Baker 1961) examined how people respond to air raid sirens, and concluded that, generally speaking, the general public did not understand what air raid sirens meant. This early social science research on warnings, as well as the research that followed, was brought together by Mileti and Sorensen (1990) in a synthesizing state-of-the-art report on risk warning communication. This work included a detailed warning bibliography, which later yielded two annotated bibliographies – one on warnings and the other on public education about disaster preparedness. The effort documented five key elements of imminent threat warnings as well
as five critical style elements. The five message elements documented by Milteti and Sorensen were incorporated in the design of CAP. Their work provided the foundation for the structure of the nation’s WEA messages.

The warning literature distinguishes the terms “alert” and “warning.” Providing an alert is akin to “ringing a bell” – such messages are designed to gain the attention of those potentially at risk, and motivate them to seek additional information. A warning, on the other hand, is designed to provide more complete information about what event has occurred and what actions should be taken to reduce risk. WEA messages, because of their constrained 90-character length limitation, function more as alerts than as true warnings. It is likely, however, that as length constraints initially imposed on WEA messages expand over time, longer WEA messages may be able to provide members of the public with sufficient information about what has happened and what guidance is recommended to enable them to make informed decisions about the protective actions they should take, thereby functioning as true warning messages.

**Sense-making**

After receiving a warning message, individuals engage in efforts to make personal sense out of the situation. This process of “sense-making” occurs during what is sometimes referred to as the “response gap” – the time between receiving a message and responding to it. Mileti and Sorensen (1990) documented five activities that occur as part of the warning message sense-making process: understanding, believing, personalizing, deciding, and searching and confirming. In order to make sense of a warning message, receivers of the message must first understand and attach personal meaning to the message. In addition to understanding the message, receivers must also determine whether or not they believe the message. That is, they must conclude whether the risk, warning, and contents of the message are accurate. Personalization refers to the process of conclusion that one is, in fact, the intended target of the message and that one, indeed, is no longer safe. Deciding refers to forming an idea about an appropriate course of action. Searching and confirming refer to searching for information from other sources to confirm what one understands, believes, personalizes, and what one decides to do or not do in response to the warning message.

**Message content**

The social science research record provides evidence that warning messages will be more effective, that is, the messages will have a higher probability of motivating appropriate and timely protective actions taken by the general public, if they contain the five key warning message topics documented by Mileti and Sorensen in their 1990 report. These five key topics are a description of the hazard and its consequences; protective action guidance; the location and population at risk; the time by which the public should begin taking the protective action as well as the time by which taking the protective action should be completed; and, finally, the message sender or source. These key warning topics provide information about what, when, where, why, and who. The social science warning literature provides evidence that warning messages are more effective at motivating appropriate protective action when they provide information to the public about the impending hazard by describing the event, the consequence of the hazard’s impact, the threat posed, and how taking the recommended action will reduce the hazard’s consequences. Specific guidance about
recommended protective actions should be provided to maximize health and safety, and explicit information about exactly how to take the action should be included. Protective actions linked to basic human values (e.g., “evacuate to keep your family safe”) more effectively motivate warning response than recommendations communicated without connection to such values. To be effective, warning messages must also specify the location of the event, indicating exactly who will and who will not be affected, using physical geographical boundaries to explain the location where people who need to take protective action are located. The use of maps to add visual information about who should and who should not take protective action may help people readily determine whether or not they are at risk (i.e., personalize the message). Warning messages also are more effective when they provide explicit information about time, including a cutoff time indicating by when recipients should begin taking the recommended protective action, as well as a cutoff time indicating by when they should complete taking the action. Finally, messages are more effective when they state the message source, or who the message is from, based on what constitutes the most credible/believable source for the overall population at risk. Given the diverse nature of the multiple publics served by a single alert or warning message, a panel of mixed sources typically is more effective than any single source. Messages that include both official and familiar sources are most effective.

The templates initially used by the NWS in the WEA service included these five key warning topics with one exception. Time in the research literature refers to when people at risk should begin and complete taking a protective action, whereas time in a WEA message refers to when the message expires or is no longer in effect.

Message style

The social science warning literature, as documented by Mileti and Sorensen (1990), provides evidence that five style elements contribute to effective warning messages. These are clarity, specificity, accuracy, certainty, and consistency. Warning messages that are simply worded, free of jargon, and use words that average members of the general public can understand have greater clarity and are more effective than other messages. Specificity refers to precise, nonambiguous information about the area at risk, the actions people should take, the nature and potential impact of the hazard, the amount of time people have to engage in protective action before hazard impact, and the source of the message. Message accuracy refers to the extent to which a message contains information that is correct, complete, current, and error free. Certainty refers to warning messages in which information is stated with authority and confidence, including during those situations in which there is ambiguity. Certainty is especially important in the language chosen for communicating recommended protective actions. Consistency refers to a message’s coherence. Warning messages are more effective when they have external consistency and internal consistency. Messages can have increased external consistency by explaining changes from prior messages, and can have increased internal consistency by not containing information that conflicts or is contradictory, such as, “radiation is in the air, but do not worry.”

Hazard type

Different types of hazards are perceived by the public in different ways. For example, some hazards are more familiar, and some hazards are associated with higher pre-event levels of perceived risk and dread. The historical research, however, indicates that the message content
elements described above are pertinent to all hazards. Thus, it is possible to standardize the types of information included in warning messages, including information that can help increase message personalization, such as geolocation information about the area affected and who is at risk.

Nonmessage factors

Social science research includes the effect of nonmessage factors that are associated with message response. These factors include status (e.g., age, gender, ethnicity, and acculturation); roles of responsibility within and outside of the family (e.g., family size, number of children, number of pets, having a united family, and greater community involvement); hazard experience (i.e., personal experience with prior disasters); pre-event knowledge about the hazard and protective action; environmental cues or indicators in one’s environment that reinforce the presence of the hazard (e.g., smoke and ash indicating the presence of fire); and social cues or indicators in one’s social environment that reinforce the presence of the hazard (e.g., seeing and hearing about others who are taking the recommended protective actions).

Geotargeted alerts and warnings

Within the context of alerts and warnings, geotargeting is a two-step process that involves first defining the targeted area and then delivering a message to individuals located within that area. Lack of precision is a concern during both steps. There may be lack of precision in defining the region and in delivering the message. For example, when the most granular level allowable for a given system is an entire county, and only a portion of the county is at risk, the defined area at risk may be imprecise. Likewise, given current technology, there may be lack of precision in delivering the message to only the affected region.

Systems that provide geotargeting

Several systems currently provide geotargeted alerts. These include the “reverse-911” systems that can dial groups of landline telephone subscribers, wireless-based systems such as EAS, the WEA system in which alerts are transmitted to cellular phones and mobile devices using cellular broadcast technology (only the cell towers that correspond to the defined geographic area broadcast the message), and the NOAA weather radio, which uses dedicated radio frequencies and special purpose receivers to deliver alerts, allowing users to subscribe to geographic areas of interest.

Using geotargeting to communicate risk

The benefit of geotargeted alerts and warnings is that they allow messages to be tailored based on the nature of the particular hazard at a given geographic location and the protective actions that are appropriate for people who are physically present in the given area. Receiving a geotargeted message can strengthen message personalization, or belief that one is indeed at risk and, therefore, should take protective action. Given adequate training and system knowledge, geotargeted messages can take the place of lengthy descriptions of which areas are affected by the hazard, allowing for greater focus on descriptions of what happened (i.e., the hazard) and what should be done about it (i.e., recommended guidance); this is of particular importance for systems with message length constraints. Research involving a food recall experiment has shown that, when a message is
GEOTARGETED ALERTS AND WARNINGS

geotargeted, not only does it attract more attention but message recipients better understand the potential risk they face (Frisby, Veil, and Sellnow 2013).

Communicating geotargeted information

There are three main methods by which geotargeted information about what geographic area is affected and who is at risk can be communicated to message receivers (National Research Council 2013). The first method for communicating geotargeted information in an alert or warning message is to provide text-defined geolocation information. This approach is appropriate for systems that can transmit only text information. Text may be used to define geolocation by using a general statement such as, “Tsunami warning in this area until 12:00 Pacific Standard Time, take shelter now. – NWS.” A benefit of this approach is that the message is short and will work for formats where message length is constrained, such as the WEA service. A potential concern is that without an investment in public outreach and education, recipients may wonder whether or not the message actually does, in fact, apply to them. Another text-based approach to defining geolocation is to specifically name the geotargeted areas, as in the following example message: “Radiological hazard warning in Los Angeles, Orange, and Riverside counties until 12:00 AM PST. Take shelter now. – DHS.” In this example, the areas affected are specifically named, but messages constructed in this way may be longer than possible on some systems, and the areas named may be geographically larger (and therefore may be defined with lower precision) than the actual geographic areas affected. A more precise text-based approach is to name the boundaries of the area affected, as in the following example: “Fire warning in the area bounded by North West Parkway/E-470 on the north, Highway C-470 on the south, Highway 285 on the west, and the eastern boundaries of Adams and Arapahoe counties until 12:00 AM PST. Take shelter now. – NWS.” While this approach lends more precision to the definition of the targeted area, it also adds more length to the message. Furthermore, naming the boundaries can be problematic in terms of identifying boundaries that appropriately define the area at risk and that are readily understood by message recipients. This approach can be particularly challenging to visitors of the area, who may be less familiar with the locality.

A second approach to communicating geotargeted information in an alert or warning message is to use maps. Shading and coloring can be used to distinguish areas affected from those that are not. Inclusion of a location marker indicating the message recipient’s location relative to the area at risk can provide further personalization. This approach is dependent on having sufficient technical capabilities in the system. Maps can be difficult to read on small screens, such as cellular phones, and when resolution is low. Moreover, not all message recipients will be skilled at interpreting maps, which may lead to confusion about what has happened and what areas are affected. This can cause message receivers to spend additional time seeking further information to confirm the initial message, thereby increasing response delay, and possibly increasing subsequent death and injury.

A third approach to communicating geotargeted information in an alert or warning message is to use geodefined message delivery while indicating to the recipient that the message was, in fact, targeted. An example of this approach might be to say, “If you receive this message, you are at risk.” This approach is most appropriate when the warning system allows for fine-grained location definition.
Concerns regarding geoprivacy

In the context of geotargeted alerts and warnings, security involves ensuring that only authorized persons are able to send alerts, message cannot be modified to misinform the public, and messages cannot be suppressed through damage to (or overload of) the delivery system. Privacy refers to an individual having control of what information is exposed to whom, when, and for what purpose. Geoprivacy refers to an individual being secure from unwanted observations and tracking. Techniques exist to preserve geoprivacy while still providing location-based services. In areal masking, for example, location information is only sent when a person enters a particular region; it would not be necessary to track travel within the region. iGeoLoqi has developed tools incorporating such techniques. In 1992, Darrell Ernst began developing technology to geotarget alerts while preserving privacy. Using this technology, an alert would be sent to a wide area, and a location-aware device would determine whether or not the alert was relevant to the device owner. Because the device assesses and filters messages itself, the ability of authorities who send messages to track where end users are located would not be necessary.

The Federal Trade Commission has provided a list of “Best Practices for Mobile Privacy.” These include provision of timely privacy disclosures to consumers and obtaining explicit consent before allowing apps to access and collect sensitive data; the development and implementation of a visual dashboard displaying types of data accessed by various apps; the design, testing, and implementation of a set of simple icons to communicate privacy practices; and consideration of offering a “Do Not Track (DNT)” option, among other recommendations (Federal Trade Commission 2013).

Technologies for geotargeted alerts and warnings

Several recently developed capabilities have facilitated the future development of a geotargeted alert and warning system (National Research Council 2013). These include the CAP standard for formatting alerts, which includes geographical locations by FIPS code or vertices of polygons to define the geographic regions affected, information about the source and nature of the alert, and the recommended action to be taken. In addition, cellular phones and other mobile devices “know” where they are physically located based on mandated Enhanced-911 location capabilities, global positioning system (GPS) receivers, and nearby wireless access (Wi-Fi) sites. Computing devices, such as laptops and desktops, can be outfitted to establish their location and wired devices can use the physical location of the networks to which they are attached to do the same. Applications on any of these devices can be used to receive and communicate geotargeted alerts and warnings. Because devices “know” where they are located, they can use that information in combination with the geographic information coded in alerts to deliver the message only to those individuals physically located within the specified at risk location. Moreover, increased geotargeting resolution is becoming available.

A variety of traditional technologies can be harnessed to provide geotargeted alerts and warnings (National Research Council 2013). Telephone alerting using reverse-dialing alerts can be used by officials to auto-dial landlines or mobile telephone numbers of registered users in a given geographic area using a prerecorded message. Drawbacks include the fact that dialing large sets of numbers can overwhelm local phone switches, the significant decrease in the number of households with landlines limits system reach, reverse dialing systems do little to increase access
GEOTARGETED ALERTS AND WARNINGS

to those with disabilities, and such systems are expensive. Radio broadcast technologies also provide opportunities for transmitting geotargeted alerts and warnings. National Public Radio (NPR) is currently investigating the use of broadcast repeaters to extend the reach of radio alerts and to enable geotargeting of alert and warning messages and also the use of radio broadcast systems (RBDS) to evaluate the efficacy of using household receivers to send text information to people with hearing impairments. The NOAA weather radio (NWR), originally developed in the 1950s and 1960s, uses the specific area message encoding (SAME) standard to geotarget alerts about weather observations to those in flight or at sea. SAME was adopted in 1988 as the first geotargeted alerting standard. It allows NWR to target at the FIPS code level (i.e., generally at the county level). Cable television also can be used to deliver geotargeted alerts and warnings. Alerts received by EAS currently are distributed to all of a cable system’s subscribers, even to individuals who may live outside of the area affected. It may be possible to modify cable boxes to that they “know” their location and can filter messages to provide more precise geotargeting.

Technologies for geotargeting alerts over the Internet also can be used (National Research Council 2013). Geotargeting based on Internet protocol (IP) address currently exists but is limited and relies on privately managed databases that match IP addresses to physical addresses. An alternative approach would be to incorporate alert information, including geography, in an information source that a user frequently monitors, such as Facebook, which would then send the alert to all subscribers in a given state, for example. Services that stream video over the Internet, such as Netflix, YouTube, and Hulu Plus, are increasingly replacing traditional broadcast and cable television for which alerting systems already exist. These services could be modified to receive and display geotargeted alerts delivered by the video stream and displayed by the content viewing application.

The location of mobile device users may be determined in a variety of ways (National Research Council 2013). This traditionally has been accomplished using GPS or cellular tower triangulation. More recently, Wi-Fi access point signals have been used by Skyhook Wireless to match access point and cell tower signals to a proprietary location database. Mobile device location also can be determined based on over-the-air television broadcast signals using TrueFix TV Positioning, developed by TruePosition. Here, location is determined using signal arrival times, which are directly proportional to the distance to the transmitter, with the benefit that the technology can be used in indoor and urban environments, where GPS signals are unable to penetrate building structures. Similarly, UpLink Time Difference of Arrival (U-TDOA, also developed by TruePosition) uses the time of signal arrival at multiple cellular towers. Finally, indoor geolocation of mobile devices can be determined using additional data sources. GPS and other signals used to determine location do not readily penetrate buildings, making accurate indoor geolocation more difficult. Qualcomm has worked on developing a technique that uses the additional data sources of Wi-Fi measurements, building maps, and phone sensors such as accelerometers, gyroscopes, and compasses, to determine indoor positioning.

Current and future technologies for geotargeting alerts to mobile devices include geotargeting of short message service (SMS) messages, third-party application capabilities, and carrier geotargeting of WEA messages. Almost universally supported and widely used to send and receive text messages, SMS can be used to deliver geotargeted alert and warning messages
GEOTARGETED ALERTS AND WARNINGS

The Open Geospatial Consortium has been working to develop a standard (Open GeoSMS) for communicating location information in SMS messages. Location can be displayed in a mapping tool or used to retrieve satellite images. Using third-party applications on a mobile device in combination with the device’s location information also can be used to deliver geotargeted alerts and warnings. For example, iMap Weather Radio, developed by Weather Division Technologies, communicates NWS alerts to the public, providing some of the key features of a NWR (e.g., alerts cause the phone to wake up automatically). Finally, carrier-specific geotargeting methods to deliver WEA messages can be used. Challenges and opportunities for improving geotargeting include the use of the long-term evolution (LTE) standard, in which cellular IDs are assigned to individual antennas on each tower, rather than the tower as a whole, potentially allowing for greater precision. GPS could be used to determine location but this approach poses many challenges, including network congestion and poor indoor and underground performance. It may be the case that the alerting system does not need to accomplish the geotargeting. A phone’s position may instead be determined by a combination of technologies previously described, and the phone can make the determination as to whether or not an alert applies to its location. The greatest opportunity for improving the ability of geotargeted messages to motivate public response with the least impact on networks and devices may be to expand the amount of information provided by allowing pagination, or a series of messages that together provide the full warning message text.

Future research

A report of the National Academy of Sciences 2013 workshop on geotargeted alerts and warnings lists key areas for ongoing and future research (National Research Council 2013). These key research areas include discovering how to facilitate and improve public warning response to geotargeted alert and warning messages, the value of geotargeted information, the best ways to develop and deploy new technology, how to best respect privacy and meet security needs, and effective strategies for facilitating and encouraging practitioner use of geotargeted alert and warning systems.

Much research on ways to facilitate and improve public response is currently underway. Key topics include opportunities for optimizing message content, structure, and delivery; the extent to which more precise geotargeting may reduce morbidity, mortality, human suffering, and other disaster related costs; the best method of communicating locations at risk and the recommended protective actions for each location in order to reduce the amount of time individuals spend seeking additional information (i.e., reduce response delay); system opt-out rates, their causes, and strategies for encouraging people to opt back in; ideal message repetition frequency; and strategies for determining message effectiveness during a given event.

Given the variety of ways in which geotargeted information can be presented (i.e., general text-based information indicating one is at risk, text including a place name for the area at risk, a map, or some combination of these), research regarding the optimal display of such information is needed. This includes discovering what combinations of text and maps are most effective in encouraging message recipients to take action, what is the most effective size of geotargeting boundaries, what is the most effective way to display maps, how does map literacy affect public warning response, what design and visualization principles are most effective for geotargeted alerts and warnings, and how can
GEOTARGETED ALERTS AND WARNINGS

targeted message be made as accessible as possible, particularly for those with disabilities?

Several research gaps exist concerning the development and deployment of new geotargeted alert and warning systems. These gaps include how to quickly adapt technologies developed by the private sector for delivering geotargeted alerts and warnings, as well as understanding what legal, regulatory, technology standard, or other barriers might hinder the rapid deployment of new geotargeted alert and warning technologies. Another research gap involves understanding the manner in which special purpose alert and warning systems (e.g., WEA) and more general purpose messaging systems (e.g., Twitter and other social media) may best work together to provide effective alerts and warnings to the public, including discovering how these systems might complement or interfere with each other. How to design new technologies that allow the inclusion of maps, images, and links that allow the forwarding of alerts and warnings to others, and that incorporate the needs of at-risk populations with physical or mental disabilities, and what strategies can be used to decrease bandwidth requirements, are additional research needs. Finally, future research should address how to best close the gap between what is known about how publics respond to alert warning messages and the technology available for delivering geotargeted alerts and warnings.

The use of geolocation information related to mobile devices raises concerns about user privacy. One important future research need is determining whether or not users have the requisite knowledge to provide explicit consent for using geotargeted mobile alert and warning systems. Another important research area is how to design future alerting systems so that they incorporate privacy and security concerns “up front” as opposed to attempting to address these concerns after the system is designed or deployed.

A final privacy-related research question is, to what extent might concerns about government regulation or negative public relations inhibit developers from incorporating geolocation information into new technologies, and discourage private providers from participating in new systems for alerting the public?

In terms of real-world application, guidance to help practitioners in government roles make decisions about how and when to issue alerts and warnings is needed. Understanding the barriers and facilitators to experimenting with, adopting, and increasing use of new alerting systems and geotargeted alerts, as well as the best roles for involving practitioners in system design, can help facilitate more appropriate use of emerging systems by practitioners.

SEE ALSO: Built environments; Environmental hazards; Health geography; Information and communications technology; Information technology and mobility; Natural hazards and disasters; Social vulnerability and environmental hazards

References

GEOTARGETED ALERTS AND WARNINGS


Further reading


Geotargeting protocol standards

Tien-Yin Chou
Yi-Min Huang
Wei-Yen Lin
Feng-Cheng Lin
Chen-Yu Hao
Feng Chia University, Taiwan

With the advance of information technology, mobile devices such as smart phones, navigators, and tablet PCs have become the major platform for various services. Location-based information has become the major stream of applications. However, there exists an interoperability issue when different systems or services try to communicate. Therefore, standards and formats are important for exchanging location-based data. Geotargeting protocols are the protocols used for transmitting location-based information from sensors, the Global Positioning System (GPS), and mobile devices that can provide their own location coordinates. Among the standards, GeoSMS (Geo Short Message Service), WaterML, and SWE (Sensor Web Enablement) are introduced in this entry.

GeoSMS

Mobile devices such as mobile phones, vehicle navigators, and tablet PCs can be aware of current location through embedded GPS or other global navigation satellite system (GNSS) positioning services. With this location-aware feature, all kinds of location-based services (LBS) appear to be the most popular services and applications in daily life. However, there exists an interoperability issue among those LBS applications when it comes to “talking to each other” because different devices or applications may represent the location information in different formats. Standards for location reference are essential for service developers to create more powerful applications that traverse different platforms and systems.

The Open Geospatial Consortium (OGC) Open GeoSMS could be the most simple but useful geospatial standard for LBS, providing developers with an extended Short Message Service (SMS) encoding and interface to facilitate communication of location content between different LBS devices or applications. SMS is the open text communication service standard most used in phone, web, and mobile communication systems for the exchange of short text messages between fixed line or mobile phone devices. Open GeoSMS uses this existing, widely supported protocol to define an extended format with simple location content in geographic coordinates for achieving interoperable communications; meanwhile, there is still human readability of the content. No extra hardware or communication infrastructure is required for implementing this standard. It was first proposed by the Industrial Technology Research Institute (ITRI) of Taiwan in 2008, and became an OGC standard in 2012.

Format and usages

The Open GeoSMS expression uses the following structure: HTTP/HTTPS (Hypertext Transfer Protocol/Secure) URI (Uniform Resource
GEOTARGETING PROTOCOL STANDARDS

Identifier) followed by payload in a new line. The first line of an Open GeoSMS message is an HTTP/HTTPS that indicates the server that will process the SMS string. It is mandatory and contains location information in coordinates and a postfix in the query string as “&GeoSMS,” which ends the first line of the message. The URI is supported by most web services whose main purpose is to provide receivers with access to web services, such as map service, advertisement services, or other services from telecom providers.

The location information should be shown in the query string of the URI. The parameter element name “geo” is recommended but it is not mandatory. However, the coordinates should be expressed as latitude and longitude in the WGS (World Geodetic System) 84 datum as defined in EPSG (European Petroleum Survey Group) 4326. Coordinates should be in the decimal degree format without quotation symbols, and should be bounded by $\pm 90^\circ$ and $\pm 180^\circ$, respectively. Positive latitudes are north of the equator while negative latitudes are south of the equator. Positive longitudes are east of the Prime Meridian while negative longitudes are west of the Prime Meridian. Latitude value goes before longitude, separated by a comma.

The postfix string “&GeoSMS” in the query string of the URI is also mandatory and can provide extra information indicating the type of message in order to be pre-processed by the server. For example, “&GeoSMS=OP” means that the message is optional and represents an operation description to identify a specific operation for a specific application. “&GeoSMS=P” means that the message is simply a point of interest (POI).

The following is an example of geo-messages by Open GeoSMS.

http://maps.geosms.cc/showmap?location=23.9572,120.6860&GeoSMS

Debris flow alert! Use emergency evacuation route now!

where “http://maps.geosms.cc/showmap?” is the URI of a map service, “location=23.9572,120.6860&GeoSMS” is the location parameters and postfix string, and “Debris flow alert! Use emergency evacuation route now!” is the text content.

Then, the encoder of this geo-message can parse it and display the message in a digital map according to the location it specifies.

Applications of GeoSMS

The Open GeoSMS was first successfully adopted to facilitate disaster management around the time that it became an OGC standard due to its light weight and easy-to-use features for communication during disasters. Many experiences of huge disaster rescues have shown that coordination among rescue forces and humanitarian organizations from different nations is essential but difficult to achieve, mainly because of the lack of a platform and mechanism to collect precise and up-to-date information relating to those incidents and the survivors.

Among those platforms for rescue coordination and information dissemination, the most famous one may be Ushahidi, the open source software for information collection, visualization, and interactive mapping. Each website powered by Ushahidi is created based on its establisher’s need for data or information in a particular geographical area. They can define their own data categories to fit their needs as well as the media by which the public can submit data based on their observations, such as email, SMS, Twitter, or online forms on the website. In the case of the Haiti earthquake of 2010, data categories included deaths, emergencies, threats, responses, missing persons, and so on. Obviously, GeoSMS messages are one of the most convenient media
for reporting emergency information due to the nature of their standardized location reference and string messages.

The Sahana project, initiated by volunteers in Sri Lanka to help people suffering in the aftermath of the Asian tsunami of December 2004, is another free and open source software disaster management system. The system, developed in Phase II as a generic disaster management tool, was further developed with the sponsorship of the Swedish International Development Agency (SIDA), IBM, and the US National Science Foundation (NSF). It has been used by governments and nongovernmental organizations (NGOs) in the Philippines, Pakistan, Indonesia, and Sri Lanka. Sahana is also adopting OGC Open GeoSMS (Figure 1) and used the Android application contributed by its founder, ITRI, as an official app. This open source app supports incident reporting, task dispatch, and peer-to-peer team communications.

Besides its applications for disaster management, OGC Open GeoSMS has also been used in commercial products. Applications for
roadside assistance and taxi booking services have been implemented in Taiwan using this standard. OGC Open GeoSMS messages are sent like regular SMS without Internet connection, tagged with the sender’s location, and travel across heterogeneous platforms of different service providers. Unlike other service providers, Taiwan’s solution does not require users’ location by call center, which means users do not have to struggle using natural language to describe where they are. Figure 2 shows some use cases of sending OGC Open GeoSMS messages between taxi drivers and passengers. Figure 2a is the GeoSMS message sent to a passenger after his/her request is addressed. Figure 2b is a special case of GeoSMS message sent by the passenger to the taxi driver when the passenger needs to change the pick-up point during the waiting period. In both cases, GeoSMS messages can minimize the potential problems of imprecision of location reference and uncertainty, and thus make the business of taxi services more secure.
Figure 3  Milk River basin. http://www.umt.edu/watershedclinic/images/clip_image002.jpg. Reproduced by permission of John Lhotak.

WaterML

WaterML 2.0 is a standard information model for the representation of in situ water observations data, with the intent of allowing the exchange of such datasets across information systems. It is based on Observations and Measurements (O&M) 2.0 and implemented as an application schema according to the rules of Geography Markup Language (GML) 3.2. Through the use of existing OGC standards, it aims at being an interoperable exchange format that may be reused to address a range of hydrological data exchange requirements.

The core aspect of the model is in the correct and precise description of hydrological time series. Interpretation of time series relies on understanding the nature of the process that generated them. Thus, WaterML 2.0 customizes the more generic O&M 2.0 specification to the specific requirements of the hydrology domain.

The basic element of WaterML 2.0 documents is the wml2:Collection, which features one or more wml2:observationMember elements. These members each equal an actual time series and include information about the measuring site and location as well as the parameter, the process, and the actual result.

The feature of interest is defined as a sampling feature which is equivalent to a station or site that provides the data. Metadata include location and time zone information as well as gauge datum and monitoring type. Several other parameters are available with the specification and more can be added by self-definable optional parameters.

The observation data can be stored as either categorical or measurement time series in the om:result element. The short description here refers to the more common measurement time series. Metadata are divided into time series-based metadata which is valid for the whole time series (e.g., if the series is cumulative,
GEOTARGETING PROTOCOL STANDARDS

Figure 4  Souris River basin. http://nd.water.usgs.gov/floodinfo/souris.html. Reproduced from USGS.

equidistant, etc.) and metadata which is valid on a per value basis. This allows the unit, quality, and interpolation information as well as several other fields for each value individually (or defining a default value) to be stored. The last part of the result is the list of values with according timestamps as a sequence of wml2:point elements.

Time series observation data with various conditions can be encoded as WaterML format; however, wml2:metadata that describe metadata of an entire dataset and wml2:point elements that describe each measurement are the most important elements in WaterML.

WaterML application

Recently, the GIS (geographic information system) Center of Feng Chia University was involved in the project CHISP-1 (Climatology and Hydrology Information Sharing Pilot Phase 1), which was initiated by the OGC. This project aims to develop a highly standardized system in cooperation with other contributors to solve the integration problem of two case study watersheds at the border of Canada and the United States (Milk River basin and Souris River basin). Even though both countries used OGC standards when designing their own national systems,
gathering and analyzing the data is complicated because of the differences in realization of the standards when expanded and also because some services aren’t provided. Because of the lack of cross-border data, flood warning systems in each country were unable to use data provided by the other for the same watershed, which resulted in a lack of early warning services.

Several water catchment basins cross this border. The pilot region considered two basins with portions in both the United States and Canada. The Milk River (Figure 3) basin includes parts of Alberta, Saskatchewan, and Montana and the Souris River basin (Figure 4) includes parts of Manitoba, North Dakota, and Saskatchewan.

Not only in North America, but also in Europe, regulations stimulate transboundary cooperation. The European Flood Directive (EFD) facilitates data sharing between member states in order to reduce risk for affected populations and to increase effectiveness by reducing spending on data collection. The EFD also allows for common flood responses to deal with urban areas situated on either side of the border. Transboundary operations may also strengthen cooperation between responsible authorities and increase confidence.

The CHISP project was planned to use the solutions of standard services and to demonstrate how to build up a system using software components as building blocks, which would also mean reusable solutions and interoperability. Currently, data from both the Canadian and the US side appear at their national datacenter, which is harvested by a CHISP Harvester module. Harvester constantly gathers data and stores the data in the database, updating existing data where necessary. This method does not need any user interaction to run. This Harvester service is also able to convert data between different data formats (e.g., WaterML 1.1 and WaterML 2.0). Harvester also notifies the Broker module about changed records.

The user can subscribe to different areas of interest by using email addresses and defining threshold levels of a parameter of interest. This notification service accepts multiple subscriptions at the same time with different thresholds, and also provides an upstream service which means users will be notified whenever the water level exceeds the threshold at any point in the upstream catchment.

The Broker process can be scheduled to compare the latest stored data with user presets. When a new recorded value is over the threshold, Broker automatically notifies the user by email and also via CAP (Common Alerting Protocol) alert from the MASAS (Multi-Agency Situational Awareness System) hub. Using the link provided, the user can log in to the service and check the diagram of historical data for each area of interest, announce a flood warning, and order intervention.

The operability of the system was demonstrated by a historical flood event which occurred in 2011 on Milk River between April and May. The CHISP-1 project exemplifies how cooperative developers can set up a system by following standards. Additionally, established standardized outputs can easily serve data for other purposes; inputs ensure flexible expandability for further developments.

First, significant achievements in this implementation have been made by an interoperable system framework that provides a capability for an EM (emergency management) analyst (or anyone) to view transboundary upstream hydrometric (and groundwater) data via the web in near real-time.

Second, this study implemented a Harvester capability to automatically retrieve and store time-series WaterML2 data for streamflow and water level, and to identify if an identified
threshold value had been reached on a near real-time basis.

Last but not least, this study demonstrated interoperability through integration of OGC SOS (Sensor Observation Service), WPS (Wi-Fi Protected Setup), WNS (Web Notification Services), and CSW (Catalog Service for the Web) services along with the OASIS (Organization for the Advancement of Structured Information Standards) CAP standard to provide near real-time threshold monitoring and notification to support alert mechanisms across international boundaries and jurisdictions.

**SWE**

SWE architecture is developed and maintained by OGC members with the aim of communicating between sensors and sensor-like elements. The SWE framework is designed to create network access through a common interface.

---

**Figure 5** SWE application. [http://www.opengeospatial.org/ogc/markets-technologies/swe](http://www.opengeospatial.org/ogc/markets-technologies/swe). Reproduced by permission of OGC.
GEOTARGETING PROTOCOL STANDARDS

Applications

Data communications

Operations

SOS

(Sensor observation service)

SWE server

Sensors data of debris flow monitoring station

CD camera
Geophone
Rain gauge
Wire sensor
Water level meter

Figure 6  The debris flow monitoring system architecture.

SWE technology mainly includes SensorML (Sensor Model Language), O&M, SOS, Sensor Planning Service (SPS), and Sensor Alert Service (SAS). The goal of SensorML and O&M is to provide a standard format for data and information communication. Data publication (or broadcast) is achieved by the web services, the concept of Service Oriented Architecture (SOA), namely the services of SOS, SPS, and SAS. XML syntax is used in SensorML for sensor modes, and in the O&M data exchange interface to describe the observation or measurement of sensors. SOS provides the users (or other programs) with access to one or more sensors and their measured values. SPS provides users (or other programs)

Figure 7  The user interface of the debris flow monitoring system.
GEOTARGETING PROTOCOL STANDARDS

Figure 8 A request of and response to the “GetObservation” operation.

with the feasibility of collecting data from one or more sensors and submitting collection requests.

SWE has been developed for years and is quite mature on sensorwise applications. Sensors and sensor systems usually have exclusive access to their management, control, and data exchange, and typically allow no online services. SWE with OGC web services, however, provides the solutions to the demands of open data exchange on sensors and systems (see Figure 5).

Figure 9 The framework of the SWE cloud.

SWE offers integrators the following:

- open interfaces for sensor web applications,
- “hooks” for IEEE (Institute of Electrical and Electronics Engineers) 1451, TML (Transducer Markup Language), CAP, WS-N (Web Services Notification), ASAP (Asynchronous Service Access Protocol),
- imaging device interface support,
- opportunity to participate in an open process to shape standards,
- sensor location tied to geospatial standards,
- fusion of sensor data with other spatial data,
- ties to IEEE and other standards organizations.

The main adopted or pending OGC standards in the SWE framework include:

O&M – the general models and XML encodings for observations and measurements (http://www.opengeospatial.org/standards/om).
SensorML – standard models and XML schema for describing the processes within sensor and observation processing systems (http://www.opengeospatial.org/standards/sensorml).
PUCK – a protocol to retrieve a SensorML description, sensor “driver” code, and other information from the device itself, thus
Figure 10  The observation of a debris flow event in Taiwan, 2012.
Figure 11  The data publication of the debris flow event in Taiwan, 2012, using the SWE protocol.
enabling automatic sensor installation, configuration, and operation (http://www.opengeospatial.org/standards/puck).

**SOS** – an open interface for a web service to obtain observations and sensor and platform descriptions from one or more sensors (http://www.opengeospatial.org/standards/sos).

**SPS** – an open interface for a web service by which a client can (i) determine the feasibility of collecting data from one or more sensors or models, and (ii) submit collection requests (http://www.opengeospatial.org/standards/sps).

**SWE application: debris flow monitoring in Taiwan**

There are at least 400 sensors used in debris flow stations in Taiwan. One of the important issues of the debris flow monitoring system is the sensor data integration on a platform. To solve the issue, the SWE framework was implemented in the monitoring system.

The debris flow monitoring system has implemented the “GetCapabilities,” “DescribeSensor,” and “GetObservation” components of SOS operations. Figure 6 is the architecture of the debris flow monitoring system in the SWE framework. The sensors installed in the monitoring station call the operation “InsertObservation” and send the observation to the SWE server through Internet communication. Other applications can connect to the SWE server and use the “GetObservation,” “GetCapability,” or “DescribeSensor” operations to get observation information, service capabilities, or sensor capabilities from SOS operations. Figure 7 is the user interface of the debris flow monitoring system based on Google Earth. Figure 8 illustrates a demonstration of SOS.

**SEE ALSO:** Big data; Decision analysis; Environmental hazards; Environmental management; Geographic information system; Monitoring and evaluation; Natural hazards and disasters; Representation; Rivers and river basin management; Sensor networks, the sensor web, and the Internet of Things; Volunteered geographic information

**Further reading**


Social media and ambient geographic information

The emergence of web 2.0-enabled technologies has impacted human interaction and participation by allowing the general public to publish and disseminate information, and to establish connections. Social media have emerged over the past decade as the prototypical paradigm of this newfound capability, with the term used as a reference to a wide spectrum of digital interaction and information exchange platforms. Broadly, this includes blogs and microblogs (e.g., Blogger, Twitter and Weibo), social networking services (e.g., Facebook and LinkedIn), and multimedia content sharing services (e.g., Flickr and YouTube). Driven by massive participation, the content contributed and disseminated through such applications has emerged as a new big data paradigm. Here we define big data with respect to three properties, that of volume, velocity, and variety. Social media data combine high data volumes with rapid streaming rates (velocity) and content variability. For example, at the time of writing Facebook is managing petabyte-scale data as it processes 2.5 billion content elements and over 500 terabytes of data daily. Twitter traffic also sees on average 5700 tweets posted per second, for an average traffic of nearly 500 million tweets per day. Even at 140 characters per tweet, this can result in terabytes of new data contributed daily when one considers the accompanying metadata and added multimedia content. Similarly, Flickr hosts billions of photos, with over half a billion uploaded in 2013 alone. Ingesting, visualizing, and analyzing such massive amounts of data is a substantial challenge. Harnessing the increasingly geographical nature of these contributions can provide a unique prism through which to comprehend and analyze their content, and a means to support their visualization.

Social media content is often geo-tagged, either in the form of precise coordinates of the location from where these feeds were contributed, or as toponyms of these locations. Prior work on harvesting and analyzing social media such as Twitter has indicated that, on average, the percentage of precisely geolocated (at the level of exact coordinates) tweets ranges typically between 0.5 and 3%. Depending on the area of study and underlying conditions, this rate may occasionally go higher. For example, a dataset collected from Japan following the Fukushima disaster reflected a data corpus where 16% of the tweets were precisely geolocated (Stefanidis, Crooks, and Radzikowski 2013). This spike is attributed to the fact that the dataset from Japan reflected a technologically-advanced community that was on the move, in which case users were tweeting using primarily their mobile devices. Both of these situations, namely the proliferation of technology in a society and an increased use of mobile (and other location-aware) devices to post tweets, tend to produce higher rates...
of geolocated content. In addition to precisely geolocated tweets, studies have reported that approximately 40–70% of tweets come with a descriptive toponym related to the self-described user location. Regarding imagery and video contributed as part of social media, a recent study has indicated that approximately 4.5% of Flickr and 3% of YouTube content is geolocated (Friedland and Sommer 2010). However, geotagging is not the only geographical content of such information, as references to events and corresponding locations reflect a substantial portion of social media traffic.

The geographic content of social media feeds represents a new type of geographic information. It transcends the early definitions of crowdsourcing or volunteered geographic information (VGI) (Goodchild 2007), as it is not the product of a process through which citizens explicitly and purposefully contribute geographic information to update or expand geographic databases. Instead, the type of geographic information that can be harvested from social media feeds can be referred to as ambient geographic information (AGI) (Stefanidis, Crooks, and Radzikowski 2013); it is embedded in the content of these feeds, often across the content of numerous entries rather than within a single one, and has to be somehow extracted. Nevertheless, it is of great importance, as it communicates instantaneously information about emerging issues. At the same time, it provides an unparalleled view of the complex social networking and cultural dynamics within a society, and captures the temporal evolution of the human landscape.

Driven by these trends, social media feeds are becoming increasingly geosocial in the sense that they often have a substantial geographical content, thus enabling the exploration and analysis of social media in the geographic space. For example, it is possible to identify the trail of a tweet as it is retweeted within the user community (Figure 1), or construct a social network describing the follow connections among numerous users. For the first time we are able to explore the physical presence of people together with their on-line activities, and to link the cyber and physical activities on a massive scale.

Before discussing the geovisualization of social media, it is important to define what is meant by the term. Geovisualization is short for geographic visualization, which utilizes advances from a variety of fields including visualization in scientific computing, exploratory data analysis, and geographical information systems to provide new ways of presenting geographical data for theory building and understanding (MacEachren and Kraak 2001). In many cases, the geovisualization of social media feeds predominantly takes the appearance of web map mash-ups, in essence portraying the location of social media usage on a map. Such an early attempt to visualize social media is shown Figure 2. As can be seen, the focus in this type of geovisualization is on communicating where user contributions are being made from, and to show their content without any further analysis. While this approach is informative, it often falls short of capturing the depth, richness, and complexity of the information that can be gleaned from social data. As a result, a need for more advanced geovisualization approaches that are capable of better capturing and communicating the complexity and multidimensionality of social media arises. It is worth noting that this need is not unique to social media content alone – geovisualization at large went through a similar evolution from web map mash-ups at the dawn of the web 2.0 era to more interactive visual analytics in order to support advanced knowledge discovery.

The challenge of geovisualizing social media stems from the very nature of social media data. Unlike more traditional spatial data sources, social media is comprised of an intertwined
tapestry of links and associations between users weaved with both explicit and implicit content and context. This complex structure is the result of several key constructs that shape public participation in social media (Kietzmann et al. 2011).

- Identity: Social media services are built around the notion of user identities, as well as the extent and means through which users can reveal and share this identity within a social media venue. Some social media services enforce a certain degree of identity sharing, while others permit users to maintain high levels of anonymity. Identity mechanisms in social media often allow users to publish their location as part of their profile, either in the form of their actual coordinates, or as a self-declared location (e.g., city of residence or zip code). Such information is essential for geovisualizing social media contributors or contributions.

- Sharing: Social media platforms enable users to distribute, receive, and exchange information. Sharing in social media often takes place around user-generated content objects, such as text, audio, and video – which can be georeferenced at the time of creation or at the
time of sharing. Social media services often provide users with the ability to control the degree of sharing these content objects, from making them accessible to only a small group of selected users, to releasing them widely to all users in a given social media setting. Sharing activities can be based on established links and relationship to other users, or can be created through shared interests or goals.

- Communication/conversation: While sharing relates to the ability to publish information, it is communication that makes social media what it is. The ability to communicate through social media is central to its enablement, while the ability to engage other users in conversation is instrumental in establishing associations and links between users. These links are not simply binary, but rather may span various levels of importance, ranging from active and purposeful links (e.g., retweet or mention in Twitter, or “like” in Facebook) to more passive ones (e.g., follows in Twitter). These complex communication mechanisms often result in a hierarchy of connection groupings (e.g., communities).

- Relationships: Through communication, social media users can establish various types of associations among them, to form larger user communities. Just like communication
links, these relations too may vary in terms of importance. While some social media services rely heavily on well-defined relationships (e.g., LinkedIn), others provide a much looser environment for establishing such relations (e.g., blogs). These relationships may be driven by physical space relationships (e.g., finding work colleagues on Facebook and expanding an existing relationship into cyberspace), they may reside entirely in cyberspace (e.g., initiated entirely in cyberspace, linking two distant users), or may be created in cyberspace and become mirrored into the physical space.

Generally, social media services incorporate all of these constructs at various degrees of sophistication and service levels, resulting in different emphasis areas and target audiences (Hanna, Rohm, and Crittenden 2011). While this results in what appears to be a highly heterogeneous social media landscape, there are three distinct commonalities that can be found across the different constructs, and provide meaning and content to social media data, namely:

- locations of nodes, as they allow us to project cyberspace interactions back to the physical spaces occupied by their participants;
- links that are formed among various network nodes through interaction, expressing the formation of relationships and larger communities;
- content that is shared between nodes to establish links (e.g., two nodes may be connected through a discussion on a specific topic).

Indeed location is of particular importance as it allows the embedding of information about users, links, and content in relation to the physical world, thus providing the ability to view, explore, and analyze social media in a well-structured domain that supports reasoning. Geovisualization, therefore, has a pivotal role in leveraging these advantages and enabling the analysis of social media data for knowledge production.

Geovisualizing social media

The depiction of various types of links between individuals, such as family ties and kinship relations, has a long history that can be traced back to medieval times. However, the visualization of social ties as graphs has a relatively short history, which was driven primarily by the development of mathematical graph theory, as well as computational graph drawing and analysis algorithms. The premise behind this type of visualization is that social networks can be represented as graphs, in which nodes represent individuals and edges represent relationships or interactions. While this visualization paradigm has been explored extensively in particular in the context of social network analysis (Freeman 2000), the geovisualization of social network graphs and social media has remained largely unexplored until recently. In order to better understand the different approaches to the geovisualization of social media, it is beneficial to classify them according to the type of information they aim to highlight: structure, dynamics, and content. Each of these approaches is reviewed here.

Geovisualization of network structure

The geovisualization of network structure focuses on highlighting the underlying social network that can be derived from social media data. The visualization of social networks is typically carried out by embedding the social network graph in an abstract network space that...
GEOVISUALIZATION OF SOCIAL MEDIA

is not related to physical space. As these graphs tend to be large, a graph layout algorithm is used to organize nodes and edges according to some optimization criteria (e.g., organized nodes with respect to the number of edges they are connected to or with respect to their importance to the network's structure). There exists a multitude of algorithms for such layouts, from random layouts to force-based or tree-based layouts (Bender-deMoll and McFarland 2006). As a result of applying such algorithms, each node is then assigned a set of coordinates in the abstract network space and is drawn accordingly.

An example of such visualization is shown in Figure 3a, which utilizes a force-based layout to visualize a social network of a shared interest group as derived from Twitter. The same network can also be embedded in geographical space if the geographic coordinates of the graph nodes are known. An example of such visualization is shown in Figure 3b, in which graph nodes from Figure 3a are redrawn using the geographical coordinates that were derived from social media data as the layout scheme.

This visualization shows the social network spread throughout the United States, including Alaska, Hawaii, and Puerto Rico. As can be seen, embedding the social network in the geographical space provides immediate insights into where members of the interest group are located as well as how members in different regions of the country are connected. It is interesting to note that due to the density and coverage of the group members, this visualization depicts quite vividly the shape of the contiguous United States even without a background continental map.

Geovisualization of network structure dynamics

Within geography, there has been a long tradition in studying and visualizing networks (e.g., roads, railway lines, and utility systems) (Haggett and Chorley 1969). However, such networks tend to remain static and not change their topology and characteristics frequently. On the other hand, social media, and the social networks that are derived from them, tend to

Figure 3 An example of a network structure of social media users (nodes) and their interactions (edges) (a) and the same network laid out in a geographical space (b).
be much more dynamic. This dynamic nature originates from the rapid change of content, users, and links over time that often occurs in social media. For example, during a political election period social media users can respond to ongoing social media campaign by decreasing or increasing their communication activities on specific topics, or by creating, terminating or modifying their relationship to other users. In such cases, capturing the dynamic changes that occur in the underlying social network across electoral districts is of particular importance.

Visualizing such network dynamics poses a particular visualization challenge, which has been tackled in three primary ways (Ahn et al. 2011): the visualization of network summary statistics of the network over time, the visualization of a series of network “snapshots,” and the use of animation. These approaches can also be applied for geovisualizing the dynamics of social networks. For example, Figure 4 shows the geovisualization of interactions within an online community in the aftermath of the Boston Marathon bombing of 2013. The figure shows how users from across the world are interacting as they exchange information about the event. At the first instance, 10 minutes after the event, the discussion is mainly Boston-centric, in the sense that, therein lies the highest concentration of nodes. However, over time, the global community becomes involved in this discussion, marking a transition from a

![Figure 4](image-url)
local to a global event. Moreover, as can be seen from Figure 4, the network (i.e., the connections between people) is continuously changing over space and time. This “snapshot” approach can be further expanded through the use of animated maps, in which change in the network topology and node characteristics can be viewed dynamically over time.

In addition to the visualization of the network structure over time, it is possible to visualize various social network metrics, for example the in-degree or out-degree of a node, using visual variables such as color and size. An example of this technique is shown in Figure 4, where the size of discs representing each network node is proportional to the node degree, and the width of each arc connecting two nodes is proportional to the volume of traffic between them.

Geovisualization of social media content

Moving beyond the spatial and temporal geovisualization of social media networks, visualizing content derived from social media also poses a substantial challenge. Similarly to the importance of links, content in social media is essential as it provides both thematic and contextual information from which knowledge can be derived. One commonly used method to visualize thematic information is the use of tag-clouds (or word-clouds), in which tags are extracted from a text corpus and organized in a graphic layout. In this layout, the significance of tags (for example, their relative frequency or rank in the corpus) is typically expressed through font size or color. A distinct advantage of tag clouds is their ability to effectively summarize large amounts of data and present it in a qualitative manner. In doing so, tag clouds enable the rapid synthesis of textual information towards knowledge generation.

Tag clouds can be coupled with geographic information visualization in order to communicate the associations between a geographic space and a thematic space. An example of the use of tag clouds in social media geovisualization is depicted in Figure 5a, which shows a social media harvesting system (GeoSocial Gauge) (Croitoru et al. 2013) that couples a map display and the related tag cloud of tweets about the conflict in Syria. Here, the density of tweets is shown as a density map over a world map. Other systems, such as SensePlace2 (MacEachren et al. 2011) and Ushahidi (Norheim-Hagtun and Meier 2010), provide similar capabilities. It is worth noting that in addition to density maps, social media activity can be visualized using a tessellation scheme, such as a square-based tessellation or a hexagon-based tessellation (Figure 5b). Similarly, Slingby et al. (2007) presented an applet prototype for overlaying tag clouds on geographic maps in order to explore large spatiotemporal datasets.

Much like social network layouts, tag cloud layouts can take various shapes, from a simple paragraph style to various geometric shapes such as a circle or an ellipse. Recently, as the interest and use of tag clouds in social media visualization has increased, geographic layouts have also emerged. For example, Nguyen and Schumann (2010) proposed the Taggram – a technique for exploring textual geotagged information through a tag cloud visualization. In this technique textual information is embedded inside a map region (e.g., a boundary polygon of a country) by iteratively placing tags according to their rank order and the remaining available space in the region. Similarly, De Chiara et al. (2012) proposed a Tag@Map approach, in which tag clouds are georeferenced and laid out in a simplified boundary of a given territory polygon. It is important to note that the guiding principles of combining tag clouds with a
geographic layout for summarizing and exploring social media content can also be applied to other types of media, such as sound, images or video. For example, Jaffe et al. (2006) presented the tag map generation approach, in which a collection of Flickr geotagged photos is first summarized through a clustering and ranking process, followed by laying out representative photos on a map.

Similar to the dynamics in the topology of social networks discussed above, content in social media also tends to be highly volatile. Themes, topics, and keywords, often rise and diminish at varying temporal scales that can range from minutes to hours, days, and weeks. Visualizing such temporal and spatial variations is of great interest, as it provides an important first step towards the identification of patterns of topic propagation and variation, which can support the prediction of topic and message spreading, as well as the evaluation of effects or outcomes of such activities. The dynamics of social media content can be visualized through the coupling of a dynamic activity that shows the location of social media contributions as a function of given time granularity, and content (or user) rankings that correspond to each time interval. An example of such visualization using Twitter data is shown in Figure 6. Here, changes in Twitter contributions density are shown in the geographic domain (as shown in region B), while changes in content (hashtags) and user activity are shown as a series of ranked lists. In the case of hashtags (or keywords) such lists correspond to a tag cloud (as shown in region C), as they communicate both the relative frequency of the terms within the dataset. Using this approach, users are able to explore spatial and thematic changes using an interactive time slider (as shown in region D). Additionally, it is possible to visualize the change in particular content elements (in this case hashtags) through a set of corresponding time series plots (as shown in region F).
Figure 6  Visualizing social media content dynamics by coupling a Twitter stream viewer (A), a Twitter activity density map (B), and a ranked list top hashtags (C) and top authors (E), a time slider (D), and author/hashtags time-series graphs (F).
Visualization of social media analysis

Social media content is multifaceted, as it expresses a combination of location, connections between people, opinions, reactions, and their temporal variations. Accordingly, the deeper analysis of social media can reveal many levels of knowledge, some of which can be better perceived when viewed spatially, and therefore becomes relevant to geovisualization. This section presents some representative examples and a discussion about the visualization of such products of analysis.

Cluster analysis

Cluster analysis has become a popular tool in the study of social media due to its ability to aggregate individual user locations into groups and identify activity hotspots or event epicenters. There exists a multitude of spatial clustering methods to detect such epicenters, ranging from geostatistical approaches to spatially-aware and pure probabilistic approaches (Christakos 2000). An example of a spatially-aware clustering method, which has become popular in recent years, is the density based spatial clustering of application with noise (DBSCAN) (Ester et al. 1996). DBSCAN builds on density as a measure for defining and detecting clusters. While algorithms such as DBSCAN offer several distinct advantages, including the ability to distinguish noise in the data, accommodate arbitrary cluster shapes, and perform clustering without prior knowledge about the number of clusters, they fall short of accounting for the highly dynamic nature of social media. This temporal characteristic of the data is particularly important as it allows monitoring activities and events as they unfold and provides further insight into the nature of the underlying activities. To address this issue stream-based clustering algorithms can be used, in which clusters are created, updated, and dissolved as a stream of geolocated points is ingested. An example of using this approach to analyze and geovisualize social media activity is shown in Figure 7. In this example, clusters

Figure 7  Visualizing spatiotemporal clusters of tweets following the 2013 Boston bombing. Red circles indicate the approximate radius of each cluster, and color is used to indicate time.
were generated using the DenStream algorithm (Cao et al. 2006) to show the different hotspots of Twitter activity related to the 2013 Boston bombing as they evolved several hours after the event. Circles are used to delineate the different clusters, where the radius of each circle is derived from the spatial span of each cluster, and color is used to communicate the time at which each cluster was created.

**Community detection**

The interest in community detection has increased dramatically over the last few years due to the proliferation of social media. Communities are of particular importance in social media analysis as they convey the underlying organization and structure of social media users, which often leads to a better understanding of the role groups of users have in the social space, as well as to insights into how information propagates between user groups. The incorporation of geographic information, such as user location, brings a new dimension to the detection and study of communities, as it allows exploring how the formation and activities of communities relate to the physical space. Various approaches have been proposed for detecting communities in social networks. Papadopoulos et al. (2012) proposed a typology of community detection approaches consisting of cohesive subgraph discovery, vertex clustering, community quality optimization, network division, or community structure modeling. Recently, Plantié and Crampes (2013) presented an extensive survey on social community detection and proposed a classification method of the different approaches to the problem based on the inputs and outputs of each approach. Their survey indicated that most community detection algorithms use a graph (network) as input and provide a partition of the graph as output, where the partition is formed through a top-down process, a bottom-up process, or a clustering process. Once detected, it is possible to geovisualize communities by plotting them in a geographic space. An example of this approach is shown in Figure 8, in which the three largest communities in the social network of an interest group (Figure 8a) are plotted over the contiguous United States (Figures 8b, 8c, and 8d). In this example the Louvain method for detecting communities (Blondel et al. 2008) was utilized.

Another example of interest groups could be those related to a geopolitical issue, for example the Syrian civil war (Figure 9). In this figure countries are clustered together based on a mentions network from geolocated Twitter data. The mention network is defined by users directly referring to other users within their tweet. These networks are aggregated to the country level (e.g., user A from country A mentions user B in country B) and through the Louvain-clustering method, distinct and significant communities are identified. While these communities could be represented in a table, their spatial relationships would be lost. By simply shading in the choropleth map based on membership, spatial patterns between communities become noticeable. For example, the Arab League forms two distinct clusters while Canada and the United States are in a different community (Crooks et al. 2014).

**Sentiment analysis**

Social media data are opinion-rich as people express their sentiment and views about specific topics and events. Accordingly, it is possible to mine from such data not only references to events or topics but also the public’s disposition towards them. This can be accomplished by utilizing tools for sentiment analysis from the computational linguistics community. For
example, Figure 10 shows the keywords, locations, and sentiment assessment of Twitter traffic in support of (Figure 10a, green), or against (Figure 10b, red) President Obama at a given moment in time. The figure is also communicating the complex need behind visualizing social media content. While some information exists at the level of individual user contributions, it is their aggregation that reveals more useful insight. For example, in this case one can easily identify stronger positive clusters on the West Coast compared to more negative clusters in the South, which reflect the established geographical political leanings in the United States. Furthermore, by coupling the visualization of the spatial and thematic analysis (e.g., in terms of keywords, sentiment), the bigger picture that is emerging from social media contributions can be better communicated. This joint visualization of content and place is becoming a standard approach in today’s social media geovisualization systems (e.g., SensePlace2, GeoSocial Gauge).
Figure 9  Visualization of communities derived from a Twitter mentions network relating to the Syrian civil war.

Figure 10  Visualizing positive (a) and negative (b) sentiment in a discussion pertaining to President Obama. Each figure shows clockwise from top right, individual messages, color coded into positive (green) and negative (red) sentiment, the overall sentiment of the messages, a word cloud of the top 50 key words, and a map of locations of the originating messages.

Emerging challenges

With an estimated 50 million tweets and 1.5 million Flickr images posted every day, social media content is rapidly emerging as a new opportunity and challenge for the geospatial community. The challenge is related to the substantial computational constraints imposed
by processing and analyzing such massive data that comprise diverse content and are streaming at very rapid rates, and as such it is a big data challenge. Additionally, this also presents a complexity challenge as the content of these data is highly interconnected and dynamic, establishing networks across multiple themes, locations, and time instances. Analyzing and visualizing this content in order to communicate its multifaceted nature and support the discovery of valuable knowledge is, therefore, a substantial challenge. This entry has provided examples of typical geovisualization capabilities for social media data, and showed how their visualization in a geographic context supports such knowledge discovery. In addition, it highlighted the tight coupling of spatial and social analysis that is needed to fully exploit the geosocial content of social media.

This charge to geovisualize social media content is pushing the boundaries of traditional mapping. Focusing only on the visualization of buildings and infrastructure no longer sufficient, as new social media sources require the visualization of abstract concepts like the flow of information in a society, contextual information associated with places, and the emergence of communities. Such challenges offer a unique opportunity to better link the quantitative and qualitative analysis in geography. For the first time, social media data allow us to observe and study the human social system as it is affected by and affects the world around it. Unprecedented insight is gained on this constantly evolving complex organism, where people’s roles and activities are adapting to changing conditions, and effect events in space and time. As this insight is multidimensional and highly complex, the geovisualization capabilities presented here are just a first foray into this challenge, which will undoubtedly evolve as they are driven by advancements in our analysis capabilities.

While pursuing these newfound opportunities it is necessary to remain cognizant of the associated privacy issues, in order to ensure the proper use of this public domain information. This challenge exceeds the simple anonymization of such data. A variety of private attributes can easily be revealed through the integrative analysis of multiple datasets. At the same time, the availability of geotagged information further complicates these concerns, as studies have shown that the analysis of human mobility data (e.g., cell phone tracks), allows the unique identification of individuals by using as few as four spatiotemporal points in these trajectories, even when coarse geolocation information is made available (de Montjoye et al. 2013). Accordingly, the broad range of information that is communicated through social media, an aggregate of location, social connections, and personal views is accentuating the need to re-conceptualize the concept of privacy, as suggested by Elwood and Leszczynski (2011).

SEE ALSO: Geocomputation; Geographic information science; Mapping cyberspace; Spatial social networks; Visualization; Volunteered geographic information

References


GEOVISUALIZATION OF SOCIAL MEDIA

Theory and Experiment, 10(P10008). DOI:10.1088/1742-5468/2008/10/P10008.


Germany: Deutsche Gesellschaft für Geographie (DGfG) (German Society for Geography)

Founded: 1995
Location of headquarters: currently Passau
Website: http://dgfg.geography-in-germany.de
Membership: 25,000 (as of 2014)
President: Werner Gamerith
Contact: werner.gamerith@uni-passau.de; office@geographie.de

Description and purpose

The Deutsche Gesellschaft für Geographie (DGfG) is the umbrella organization for the geographical associations and societies in Germany. The DGfG represents the interests of geographers who work in schools, colleges, and in practice. It is committed to providing the content and significance of geography as a school subject, as a science, and as a practical discipline in the community. The DGfG coordinates the activities of the professional geographical associations and represents their common goals in the general public.

Journals or major publication series


Current activities or projects

There are more than thirty working groups within the DGfG, all making an important contribution to research and the development of geographical theories and methods. The DGfG organizes the German Congress for Geography (DKG), which takes place every two years at a university in Germany, Austria, or Switzerland. The DGfG awards outstanding scientific and journalistic performances. The national committee of the International Geographical Union (IGU) in Germany fulfills, in cooperation with the DGfG, the international tasks of the association’s work.

Submitted by Florian Philipp Stelzer
Germany: Gesellschaft für Erdkunde zu Berlin (Geographical Society of Berlin)

Founded: 1828
Location of headquarters: Berlin
Website: www.gfe-berlin.de
Membership: 300 (as of 2014)
Chairman: Hartmut Asche
Contact: mail@gfe-berlin.de

Description and purpose

Originally, the Geographical Society of Berlin was founded to promote research in geography, geosciences, and neighboring disciplines and communicate the results to the public. This last aspect is of prime importance so as to provide a platform for the intellectual exchange between the scientific community in the geosciences and the wider public — through lectures, discussion, field trips, and so forth. Additionally, the society publishes a scientific journal, Die Erde (The Earth) — first published in 1949, however, a forerunner to the journal “Annual Report of the Activities of the Geographical Society of Berlin” was first published in 1833.

Journals or major publication series

Die Erde. www.die-erde.org

Current activities or projects

Lectures and other events are organized where a wide range of geographical and geoscientific issues are discussed, often in cooperation with other institutions, for example, the German-Korean Society, the German Association for Applied Geography, and so forth.

With the help of the Humboldt-Ritter-Penck Foundation the society finances scholarships for student research projects.

Brief history

The society was founded in 1828 by Heinrich Berghaus and Carl Ritter in association with Alexander von Humboldt and other eminent Prussian scientists of the time. This makes the society the second oldest of its kind in the world after its sister society in Paris. All through the nineteenth century the Berlin Geographical Society took a leading part in the exploration of Africa, Asia, and the Polar regions and in disseminating geographical knowledge in Germany and all over the world. Among the society’s
GERMANY

presidents were Carl Ritter, Heinrich Barth, Gustav Nachtigal, Adolf Bastian, Ferdinand von Richthofen, and Albrecht Penck. The society’s building at “Palais Fürstenberg,” Wilhelmstraße 23 in the center of Berlin, served as an important intellectual and societal meeting place since 1899. In the interwar years of the twentieth century, Friedrich Schmidt-Ott served as president and Albrecht Haushofer – geographer, expert in geopolitics, author and a member of the opposition against Hitler – was general secretary and journal editor. After World War II, the society was reestablished in West Berlin, in close proximity to the Freie Universität Berlin. It won back its eminent role under the presidency of Ferdinand von Friedensburg, former mayor of Berlin, and in 1966 the society moved into its own building in Berlin-Steglitz, with lecture hall, library, and offices. After 1990 the society was able to re-obtain its historic library, which was in East Berlin, and regained importance for the geographical communication of unified Berlin and its hinterland.

Submitted by Christof Ellger
The term “ghetto” has been historically used to reference an urban form that reflects intense poverty, crime, family disruption, and reliance on government assistance. This conception has deep roots in Oscar Lewis’s famous culture of poverty theory. First developed in the 1960s within the field of anthropology, this theory asserts that the ghetto was created and sustained by a cycle of intergenerational poverty rooted in human pathology and defective culture. Currently, the name ghetto is also popularly used to describe the spatial confinement of any social group (immigrant ghetto, student ghetto, artistic ghetto).

More recently, however, efforts have been made to reclaim and infuse this term with more progressive meanings and to examine its nature as spaces of neglect that are complex manifestations of societal processes (Wilson 2007). Following this line, geographer Tom Slater (2009) identifies three factors to distinguish a ghetto from an ethnic enclave (e.g., Italian, Irish, Polish) and to distinguish this urban form from other marginal spaces that emerge from different societal processes. First, ethno-racial control and containment by external overarching powers explain it. Second, social and spatial segregation is extreme and rigid. Third, these urban forms become permanent and structural. In short, a ghetto here is “a space deployed by discriminatory authorities to isolate, contain, and exploit a single ethnoracial group; a place to cast out (permanently) a group outcast from society” (Slater 2009, 492).

Many scholars have argued that the historical roots of the term can be traced to Roman persecutions of Jews. The subordination and expulsion of Jews from mainstream society was initially fueled by the spread of Christianity. As a result, Jews were placed under strict regulations throughout many European cities, for example, Jews were forbidden from owning land outside the ghetto. During the Middle Ages, these ghettos consisted of Jewish quarters in overwhelmingly Christian cities (e.g., the famous Warsaw Ghetto). In some cases, the ghetto was a Jewish quarter with a relatively affluent population (for instance, the Jewish ghetto in Venice). In other cases, ghettos were places of deprivation and during periods of population growth, ghettos (as that of Rome) had narrow streets and tall, crowded houses. Around the ghetto stood walls that, during pogroms (i.e., violent massacres or persecution of an ethnic or religious group), were closed from inside to protect the community, but from the outside during religious holidays to prevent the Jews from leaving during those times. The Jewish ghetto in Venice of the fourteenth century constituted the blueprint for cities throughout Europe (e.g., Rome, Frankfurt, Prague, and Amsterdam) as fears about disease and moral upheaval attributed to Jews spread across the continent and resulted in the involuntary segregation. Some scholars have noted that while intense overcrowding, disease, and deprivation characterized Jewish ghettos, they were also terrains of collective consciousness, solidarity, and social organization. This particular duality has given way to the emergence of its own
institutions and administrative priorities behind ghetto walls (e.g., synagogues, schools, markets, and their own justice system). This also resulted in reaching for more or strengthening existing solidarities among Jewish dwellers, even between Jews that experienced strong divisions (e.g., Ashkenazim and Sephardim).

The character of ghettos has varied through times. As Jewish ghettos were gradually dispersed during the nineteenth century, the term served as a model for all urban ghettos that have formed since, such as the African American ghetto created during the industrialization era in the United States (see later, Chicago’s South Side), the buraku of Japan, and the myriads of townships of apartheid South Africa. In American cities, including Chicago, the changing dynamics of the process known as ghettoization have paralleled shifts in racial-ethnic composition and underscored the effects of major public policy failures.

Ghettoization in the United States

According to race theorists, US ghettos, such as Chicago’s South Side, have been the product of a series of strategic decisions by white people aimed at relegating minorities, in particular black people, to substandard residential areas. Studies have shown that white people have used racial violence, binding legal documents, and other strategic means of preventing black people from moving into predominantly white communities. More contemporary research has shown that whites have undertaken a range of methods to preserve race-based residential inequities. These included the physical improvement of particular urban areas at the expense of other spaces.

To elaborate from a historical perspective, modern African American ghettos began in the 1920s and 1930s with the Great Black Migration, escaping from intense racism and poverty, and seeking employment in the booming industries of the north. Unskilled labor was desperately needed to feed the vibrant Fordist capitalism and the drive of urban elites to amass and control cheap labor. Initial magnets for these migrants were Detroit, Cleveland, New York, Baltimore, and Chicago. The most studied example is Chicago’s South Side, and a cluster of adjoining neighborhoods that by 1920 were congealing into a Black Belt. This area housed 92% of Chicago’s entire black population by 1930. By 1940, with growing industrialization and immigration, black ghettos had grown in these cities (for example, Chicago’s Black Belt housed 190,000 people).

Following World War II, segregation in these cities profoundly advanced due to the influence of federal government initiatives in encouraging rapid suburban expansion. Mortgage loans for white families and the construction of highways enabled transportation between downtown employment centers and the suburbs that ultimately led to what became known as “white flight” to the suburbs. Paralleling this movement, “urban renewal” emerged nationally that helped sculpt both the content of these African American ghettos and the economic viability of other housing submarkets. By the 1950s, housing authorities, under the rhetoric of urban renewal, further contributed to concentrating African Americans. These authorities’ urban renewal programs were thought to help boost the economy through the destruction of “blighted” neighborhoods and downtown blocks, replacing them with new and vibrant downtowns including new office buildings and industrial campuses. Disproportionately, these supposedly blighted neighborhoods were
low- and moderate-income black. Due to the construction of highways and demolition of disinvested areas, people were relocated in overcrowded and blighted ghettos and massive public housing projects built alongside the highways. Public housing was ultimately turned into warehouses for the black urban poor, widely termed “second ghettos.” Invariably, projects were monstrous in scale, usually over ten acres in size and housing thousands of people.

The effects on black ghettos were devastating. These projects profoundly isolated and stigmatized black residents. Between the early 1950s and early 1960s, for example, the Chicago Housing Authority built around 35 housing projects, constituting 38,600 units, 98% of which were built in predominantly black neighborhoods. Thus, Chicago’s Robert Taylor Homes and Stateway Gardens housed 26,000 and 6,900 people in 28 buildings and 8 buildings, respectively. Urban renewal policies along with discriminatory real-estate practices (e.g., redlining and zoning) and the threat of violence in adjoining white neighborhoods consolidated mechanisms of racial exclusion to ensure that ghettos expanded and persisted. In fact, ghettos persisted through the 1970s due to the withdrawal of major industries and other employers from Chicago’s inner-city neighborhoods.

Since 1970, the massive economic restructuring experienced in many Rust Belt cities, known as deindustrialization, resulted in a degree of economic deprivation and racial segregation that has yielded the term “hyperghettos,” referring to poor, inner-city African American neighborhoods. As an outcome of recession and deepened deprivation, the 1960s and 1970s became tumultuous years for Rust Belt cities and their black ghettos. It is widely chronicled that the riots brought a mix of hope and despair. Physically devastated neighborhoods, buildings, and storefronts did not soon recover in most Rust Belt cities. On the other hand, the riots brought attention to these marginalized areas. Soon, the riots brought revamped policies, for example, new social spending on black ghettos. Medicare and Medicaid were created and the food stamp program funding more than doubled. However, this social spending on ghetto conditions was short lived with the advent of structural changes in local, regional, and national economies.

By the early 1980s, the introduction of neoliberal policies across urban America profoundly impacted the already declining industrial base. In Chicago, for example, from 1972 to 2000, manufacturing employment plunged by nearly 260,000 jobs (and thus decent blue-collar wages). Soon the industrial economy was replaced with a service economy that consisted of low-wage service jobs with no social protection, unionization, or opportunities for promotion. In this context, the neoliberal (or post-Fordist) ghetto has become differently constituted and socially organized from its Fordist predecessor. To many scholars, the neoliberal state has systematically dismantled the welfare state, and has replaced it with a punitive “workfare” state which forces people, especially young black men, into the service economy sustained by low-wage and unprotected labor. In this vein, citizens rejected by the service economy often enter the punitive and penal neoliberal state and face police, courts, jails, and prisons. Thus, citizens failing to exercise the core neoliberal principles of self-responsiveness, self-industriousness, or self-entrepreneurship became subjects to the impulses of mass incarceration. This has led scholars to argue that ghettos and prisons have become interconnected with the punitive spatial entrapment and management of the black urban poor.

Scholar Loïc Wacquant has become one of the major proponents of this perspective in his work *Urban Outcasts*. Although he received major criticism for the methodology used to compare
the *banlieues* or peripheral working-class areas in Paris, France, and the south side of Chicago, his work posits critical questions and points. He notes that today’s US hyperghettos are constituted by a blending of racial discrimination, class inequality, and state inaction, where the retrenchment of the welfare state and the reduction of employment opportunities along with the precarization of labor, are supported by intensive police activity and the constant threat of the penal apparatus. To Wacquant (2007), and contrary to the legacy of the Chicago School of Sociology, the contemporary ghetto does not suffer from social disorganization, but is organized differently in response to aggressive policies and forces that seek to marginalize and exploit this terrain. Yet, studies conducted by other sociologists suggest that in such areas, the licit and illicit meld together via the economic arena and political sphere. Other researchers focus on the constant movement of individuals who simultaneously inhabit licit and illicit markets and find in both opportunities and income (Portes 2007).

Countering Wacquant’s (2007) notion of black ghettos, Wilson (2007) argues that in neoliberal times, we are experiencing a new type of black ghetto, what he terms, a “glocal black ghetto.” This type of ghetto has become more impoverished and more stigmatized and marginalized than before. Wilson suggests these characteristics have deepened and could not have been possible without an elaborate enabling rhetoric of fear, the “global trope.” This rhetoric, serves up ghetto containment strategies as logical and important in a new supposed global reality. This containment, offered as one part of concentrating public and private resources in downtowns, is identified as the best survival tactic for cities in this new era, making the black ghettos a common “city enemy,” therefore deepening its isolation and exclusion.

Public policy debates

Social scientists and policymakers alike continue to be concerned with improving living conditions and opportunities for ghetto residents. Nevertheless, liberal, moderate, and conservative authorities remain in disagreement about how to alleviate social problems in these distressed communities. Some of the more popular policy prescriptions include the enforcement of federal fair housing legislation and the prosecution of discriminatory housing market practices such as redlining and unfair lending practices. Some less popular strategies include community reinvestment through inner-city “enterprise zones” and gentrification of ghetto areas with special provisions for neighborhood residents to return following restoration. Nevertheless, many scholars have sought to reclaim the term from a growing sense of inappropriate usage and have revisited it in order to frame policies more accurately and sensitively.

**SEE ALSO:** Chicago School; Fordism/post-Fordism; Globalization; Neoliberalism; Race and racism; Suburbanization; Underclass theories

**References**


**Further reading**


Gini coefficient

Frank Witlox
Ghent University, Belgium

The Gini coefficient, also called Gini index when expressed as a percentage, is a measure for inequality. It expresses the inequality on a scale from zero to one (the Gini index ranges from zero to 100), where “0” represents perfect equality and “1” perfect inequality. The Gini coefficient, named after Italian statistician Corrado Gini (1884–1965), is commonly used to express the income inequality of a nation. However, it is also used on different scales or to measure different variables. Corrado Gini developed this coefficient (and index) as a response to the index measuring concentration of income distribution proposed by Vilfredo Pareto in 1895 and 1897. The index now known as the Gini index was first written down in 1914 in a research paper titled “Sulla misura della concentrazione e della variabilità dei caratteri.” The indices described in the paper were based on the so-called Lorenz curve, that is, a graphical way to display and calculate inequality.

Lorenz curves give a graphical representation of inequality. They plot the cumulative share of income earned (y-axis), whereas the x-axis shows the cumulative share of people ranked from lowest to highest income. The closer the curve gets to the 45° line of equality (everybody earns the same amount) the more equal a society is. The more the curve bends towards the axes (one person earns all the money) the more unequal a society is (Figure 1). The Gini coefficient can be calculated by dividing the area A (area between the line of equality and the curve) by the sum of A and B (total area beneath the line of equality). Therefore, the greater the area A, the higher the Gini coefficient, and thus the higher the inequality.

Mathematically, this can be represented by the formula:

\[
R = \frac{\sum_{i=1}^{n-1} p_i - q_i}{\sum_{i=1}^{n-1} p_i} 
\]

with the cumulative proportion of the population

\[
p_i = \frac{i}{n}
\]

and the income

\[
q_i = \frac{A_i}{A_n}
\]

of each individual i. The individuals are ranked from poor to rich according to their income, numbering them 1, 2, 3, … , n; n thus being equal to the total amount of individuals. A is the cumulative proportion of income and calculated as follows:

\[
A_i = \sum_{k=1}^{i} x_k
\]

According to this formula of Gini, R is always a number between zero and one; zero standing for perfect equality and one representing perfect inequality. This formula can be applied to the Lorenz curve. The numerator, when divided by n, equals the concentration area between the Lorenz curve and the 45° line of equality.
Figure 1 Lorenz curve.

The denominator – again when divided by \( n \) – equals the triangular area underneath the line of equality, with vertices \((0, 0), (\frac{n-1}{n}, 0)\), and \((1, 1)\). When used for relatively large populations

\[
\frac{n-1}{n} \approx 1.
\]

Other formulas can be found in the literature. For example, the Gini coefficient can also be calculated by using the formula (Van Ourti and Clarke 2008):

\[
G' = \frac{2 \sum_{i=1}^{n} y_i \cdot R_i}{n \cdot \bar{y}} - 1
\]

with \( i \) the “individual” 1, 2, 3, …, \( n \) (ranked from poor to rich); \( y \) the income; \( \bar{y} \) the average income; and \( R \) the ranked fractional income, calculated as:

\[
R = \frac{1}{n} \cdot (i - \frac{1}{2})
\]

This formula still needs correcting because now the Gini coefficient can never reach the upper bound of one. Therefore, the obtained result \((G')\) needs to be multiplied by \( \frac{n}{n-1} \) before having the actual Gini coefficient \((G)\):

\[
G = G' \cdot \frac{n}{n-1}
\]

Jasso (1979) puts forward an even different formula:

\[
G = \frac{n+1}{n-1} - \frac{2 \sum_{i=1}^{n} i \cdot y_i}{n(n-1) \cdot y}
\]

where \( y \) is the income; \( \bar{y} \) the average income; \( i \) the rank of the individual 1, 2, 3, …, \( n \), yet ranked from rich to poor, as opposed to the previous formulas. Thus, a higher weight is given to poorer people in this formula. It is imperative that the ranking happens in the correct order (rich to poor, or poor to rich) so that the Gini index is a positive number.

The above formulas are all based on the Lorenz curve. There is, however, also a more direct way to express the Gini coefficient:

\[
G = Cov(y, F(y)) \cdot \frac{2}{\bar{y}}
\]

Here the Gini index expresses the covariance \((Cov)\) between income \((y)\) and the cumulative distribution of the same income \(F(y)\). Again, \( \bar{y} \) is the average income.

Although the Gini index is most frequently linked to the distribution of income or wealth, it has also been applied in the medical domain, ecology, education, astronomy, social development, agriculture, and so on.

In transport geography, the Gini coefficient is used to measure traffic concentration at, for example, seaports, airports, intermodal terminals. The coefficient offers information on these various transportation nodes only in a descriptive manner. Notteboom (2006), for example, uses
the Gini coefficient in a seaport context not only to assess (spatial) inequality but also spatial concentration. It can thus provide insight in the spatial dynamics of seaport systems. The following formula of the Gini coefficient can be easily applied to ports:

\[
G = 1 - \sum_{i=0}^{n} (\sigma y_{i-1} + \sigma y_i)(\sigma X_{i-1} + \sigma X_i)
\]

with \( \sigma X \) the cumulative percentage of number of ports, \( \sigma y \) the cumulative percentage of cargo throughput; and \( n \) the number of ports.

A Gini coefficient with the value “0,” in this case, means that one port accounts for the total volume of cargo throughput, while the value “1” represents a port system where all ports are equal in size.

A different example is the Gini coefficient applied to education. To measure relative inequality in education (whether all groups in society have equal access to education or not) a new index, the education Gini \( (E_L) \), was put forward. Thomas, Wang, and Fan (2000) use the following formula to calculate the education Gini based on education attainment distribution:

\[
E_L = \left( \frac{1}{\mu} \right) \sum_{i=2}^{n} \sum_{j=1}^{i-1} p_i |y_i - y_j| p_j
\]

with \( \mu \) the average years of schooling of the population; \( p_i \) and \( p_j \) the proportions of population with certain levels of schooling; \( y_i \) and \( y_j \) the years of schooling at different education attainment levels; and \( n \) the number of levels/categories in attainment data.

The Gini coefficient, and its various derivatives, has multiple advantages. Ray (1998) emphasizes that the Gini index meets all four criteria for inequality measurement. Ray named these criteria: (i) the anonymity principle (names are irrelevant); (ii) the population principle (size of the population does not matter); (iii) the relative income principle (absolute income of a person is irrelevant); and (iv) the Dalton principle (when there is a change in the distribution of income and a part of the income of a richer person is redistributed to a poorer person – when this does not affect the ranking of the individuals – this will result in a greater equity). Moreover, it is possible to work with negative incomes, as long as there are not several large negative incomes. Although the Gini index is scale invariant (i.e., a relative index), it is not translation invariant. When an equal amount of money is subtracted or added to all incomes, the Gini index will increase or decrease accordingly.

The Gini coefficient, however, also has various shortcomings. It is a relative measurement of inequality, so care must be taken when comparing various Gini indices. Although there is a similar income distribution, the absolute wealth may be decisive in the impact the inequality has on people (poor people in a rich country versus poor people in a poor country). When data are grouped by categories or into ranges (e.g., when there is no information available on an individual level) a bias will occur. This also complicates comparisons between multiple Gini coefficients. Furthermore, the same Gini index can be obtained by different income distributions.

Table 1 shows two different distributions

<table>
<thead>
<tr>
<th>Individual</th>
<th>IncomeA</th>
<th>IncomeB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2000</td>
<td>900</td>
</tr>
<tr>
<td>2</td>
<td>3000</td>
<td>4000</td>
</tr>
<tr>
<td>3</td>
<td>4000</td>
<td>4800</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>4800</td>
</tr>
<tr>
<td>5</td>
<td>6000</td>
<td>5500</td>
</tr>
<tr>
<td>Total income</td>
<td>20000</td>
<td>20000</td>
</tr>
<tr>
<td>Gini</td>
<td>0.200</td>
<td>0.200</td>
</tr>
</tbody>
</table>
GINI COEFFICIENT

(and thus two different Lorenz curves) with the same Gini index.

It must be kept in mind that the Gini coefficient, as such, measures the inequality in income, but there are other useful indicators that complement this coefficient (e.g., absolute income, income mobility) to further examine the inequality in income of a society. These can help compare various Gini coefficients, counteracting the discussed disadvantages of the Gini index noted above.

SEE ALSO: Inequality; Quantitative methodologies

References


GIS for transportation

Shih-Lung Shaw

University of Tennessee, USA

Transportation, which serves a critical role in facilitating economic, social, and other activities in modern societies, faces the challenges of efficiency, safety, and sustainability while accommodating increasing demand for travel around the world. Geographic information systems (GIS), which are designed for the management, analysis, and visualization of geographically referenced data, have been used to address problems in a wide range of application domains. When geographic information technology is used to manage and resolve issues related to transportation planning, design, operation, maintenance, and decision-making, it is known as geographic information systems for transportation (GIS-T).

Because of the uneven spatial distributions of natural resources and population, it is necessary to move resources and goods between places to meet varying demands. People also need to travel to conduct economic, social, and cultural activities, and to fulfill other needs that take place at different locations. Travel is therefore considered a derived demand. Transportation, which facilitates the movement of people and goods, is consequently a necessity in the daily lives of the modern world. Both human mobility and accessibility in most parts of the world have improved significantly as a result of advancements in transportation technology. However, increasing travel demands also lead to new challenges. Traffic congestion is now a common challenge in most major cities around the world. Globally, over 1 million people are killed in traffic accidents each year. Air pollution and environmental impacts related to transportation are important concerns in some parts of the world. Meanwhile, globalization requires transportation and logistics systems that can move goods and people around the world in a timely and efficient manner. To address all of these challenges, it is necessary to develop transportation systems that are efficient, safe, and sustainable. Litman (2013) suggests that transportation planning is experiencing a paradigm shift from an emphasis on mobility (i.e., physical travel) to accessibility (i.e., ability to reach services and activities). The use of information technologies, including geographic information technology, to improve people’s ability to reach activities and services is viewed as one effective way to address the transportation challenges in today’s world.

Geographic nature of transportation

Transportation systems encompass transportation infrastructures as well as traffic flows. Transportation infrastructures consist of networks of highways, railroads, airports, bus terminals, and train stations as well as other support facilities such as traffic signs, signals, and railroad crossings. They are built to facilitate flows of people and goods between different places. From a transportation planning perspective, traffic flow volumes among different places need to be estimated and forecast in order to know where, when, and what types of transportation infrastructures are needed to connect different locations. Decisions about the placement of transportation infrastructure are important...
and have major implications for the rise or fall of different places and the flows between them. It is evident that transportation is geographic in nature and GIS can help manage existing transportation systems and investigate how current transportation systems can be adjusted to become efficient, safe, and sustainable transportation systems.

Location plays an important role in transportation operations and management. Transportation agencies must know where highways, railroads, bus routes, airports, ports, bridges, traffic signs, and other transportation facilities are located in order to manage and maintain them. Traffic flows take place between different origin-destination pairs (e.g., air passenger flows from London to Paris) through transportation network links and nodes. Transportation agencies must have adequate tools to manage and analyze geographic data relating to transportation facilities and traffic flows. Geographic information systems are well received by the transportation community because they offer useful functions to manage, analyze, and visualize transportation data in a geographic context that are intuitive and relevant to transportation operations and management.

In addition to knowing the location of transportation features, the connectivity between different links and nodes in a transportation network is another piece of critical information for many transportation operations. For example, passengers only need to know that a flight can take them from Beijing to New York City; they do not need to know the exact location of the flight path. In this case, only the connectivity between network links and nodes as topological relationships between different locations needs to be recorded. Many subway and bus route maps use schematic diagrams to illustrate the connectivity between different routes rather than their correct geometric representations. Geographic information systems for transportation are capable of representing both geographic locations and topological relationships to facilitate mapping as well as analysis functions required by various transportation applications.

**Past development of GIS-T**

Because transportation is closely related to locations, places, flows, and interactions, geographic information systems, which are designed for managing geographic data, performing geographic analyses, and visualizing geographic relationships and patterns, are well suited to transportation applications. Although the use of GIS to address transportation problems can be traced back to its early days in the 1960s, GIS-T began to gain momentum in the 1970s partly as a result of the development of the Geographic Base Files/Dual Independent Map Encoding (GBF/DIME) program by the US Census Bureau to support the 1970 and 1980 population censuses. GBF/DIME files represent transportation features such as roads and railroads plus administrative and census boundary lines as vector features (points, lines, and polygons). The concept of topology, which describes spatial relationships between the vector features, is used in GBF/DIME files to manage the connectivity of network links and nodes as well as the adjacency of polygons. With maps encoded as a set of topologically connected points, lines and polygons make it possible to develop useful transportation applications such as network analysis. GBF/DIME files also included street address range data on road segments that enable geocoding functions in GIS to quickly locate any given street address and to develop applications such as using a shortest-path algorithm to provide navigation guidance. A national coverage of street networks in GBF/DIME files, along with
the demographic and socioeconomic data collected from population census surveys, offered valuable data sources to the transportation community. Many US transportation agencies from the federal level to the local level began to implement transportation applications with GIS. An increasing number of GIS-T papers and panel sessions also appeared at conferences in the 1980s, including annual meetings of the Association of American Geographers (AAG), Geographic Information Systems and Land Information Systems (GIS/LIS) conferences, and GIS-T symposia sponsored by the Association of American State Highway and Transportation officials (AASHTO).

GIS-T experienced a rapid growth period in the 1980s and 1990s as GIS software products began to offer data management and analysis functions designed specifically for transportation applications. For example, transportation agencies often record the locations of transportation features and events using linear referencing systems (e.g., a traffic accident occurring at milepost 62.5 on Interstate Highway 40 in the State of Tennessee, USA). A dynamic segmentation data model was developed to manage linearly referenced transportation data, and this data model was quickly incorporated into commercial GIS software products. Many transportation analysis functions (e.g., shortest-path algorithm, vehicle routing problem) and transportation models (e.g., facility location models, travel demand models) became available in GIS software to solve different types of transportation problems. Consulting firms and GIS software products focused on transportation and logistics applications also appeared on the market. TransCAD, a commercial GIS software product designed specifically for transportation and logistics applications, was released in 1985. Furthermore, a three-phase project sponsored by the National Cooperative Highway Research Program (NCHRP) of the US Transportation Research Board (TRB), which included NCHRP 20–27: Adaptation of Geographic Information Systems for Transportation (1990–1993), NCHRP 20–27(2): Development of System and Application Architectures for Geographic Information Systems in Transportation (1994–1997), and NCHRP 20–27(3): Guidelines for the Implementation of Multimodal Transportation Location Referencing Systems (1998–2001), along with several other NCHRP GIS-T reports, highlighted the attention that GIS-T received during this period.

In preparation for the 1990 population census, the US Census Bureau improved GBF/DIME files to release Topologically Integrated Geographic Encoding and Referencing (TIGER) files which have been widely used as the base map of many GIS-T applications.

Goodchild (2000) offered an assessment of GIS-T status and challenges at the turn of the twenty-first century. He characterized GIS-T evolution into three stages: map view, navigation view, and behavioral view. The map view focused on the inventory and description of transportation systems and often represented transportation features as centerlines similar to most print copies of street maps. GBF/DIME files and TIGER files basically followed the map view with a topological data model to ensure logical consistency of links and nodes in a network. Although GIS databases created under the map view allowed for the representation and inventory of transportation features, the one-dimensional centerline representation suffered inherent limitations in accurately representing roads and intersections for transportation applications that are two- or three-dimensional in nature. At the same time, the centerline representation of links and nodes in a network is well suited for navigation purposes such as finding the best route to travel from point A to point B on a street network. The navigation view,
however, has additional requirements beyond the map view to deal with various navigation situations. For example, it is important to know if a road segment is a one-way street or if a left turn is permissible at a street intersection. One solution developed by the GIS community is to include *turntables* in GIS databases to represent various allowed turn movements at network intersections. The centerline representation under the map view also cannot properly handle relationships between different traffic lanes (e.g., turn lanes vs through lanes) or off-network traffic (e.g., vehicles in a parking lot or in an open field). Furthermore, navigation sometimes involves transfers between different travel modes (e.g., walk, transit, drive). Alternative network representations of traffic lanes, off-network traffic, and multimodal transfers have been proposed to address these requirements for navigation applications. The behavioral view focuses on the behavior of discrete objects (e.g., person or vehicle) moving between different places on and off networks. Hägerstrand’s (1970) time geography suggests a useful framework to study individual activities under different types of constraints in a space–time context, which includes a concept of the space–time path that tracks individual trajectories in a three-dimensional space consisting of a two-dimensional space plus a third dimension of time. The concept of space–time path fits well with the activity-based approach in transportation that is considered a more realistic way of understanding travel behavior than the conventional trip-based approach. The behavioral view employing time geographic concepts has received significant attention since the 1990s. Recent studies under the behavioral view have also employed computer simulation models to study travel behaviors and extended the classical time geography to better reflect changes in human activities due to advancements in information and communications technology such as the widespread use of mobile phones and the Internet.

**Current status of GIS-T**

GIS-T has now become such a part of daily life that we benefit from it without noticing that we are using it. Many transportation agencies and companies across the world are utilizing GIS to help them manage transportation applications ranging from asset management, planning and design, operations, emergency response, environmental assessment, and transportation modeling, to marketing and logistics. Transportation agencies in many countries have also implemented intelligent transportation systems (ITS) to monitor traffic conditions and to facilitate transportation operations. Data collected by ITS have enabled many useful GIS-T applications. For example, an increasing number of cities now offer online applications that allow people to monitor real-time bus locations and to set an alarm to alert them when a bus is due to arrive at a particular bus stop within a user-specified time interval. This service allows people to better manage their time and avoid long waits at bus stops. ITS data also facilitate GIS-T applications in accessing real-time traffic conditions, parking availability, road closures, and other related information via computers or smart mobile devices. A worldwide map of transportation systems can now be easily accessed through online services such as Google Maps or MapQuest. Many people use navigation systems in their vehicles or online navigation services to find the location of a street address or a place of interest (e.g., restaurant, bank, school), as well as to find the best route for traveling between different locations on a transportation network. It can be argued that GIS-T has become an important part of the infrastructure that enables
many useful transportation-related services in the modern world.

The current status of GIS-T can be assessed through perspectives of GIS-T as spatial databases, as analysis and modeling platforms, and as communication tools (Shaw 2010). GIS-T databases are now able to handle data from various sources and in different formats. Vector representations of transportation networks for one-way streets, turn movements, underpasses/overpasses, multimodal transportation systems, traffic flows, and other relevant data are widely used today. Remote sensing data such as satellite imageries and LiDAR (Light Detection and Ranging) data are also frequently used with vector data in GIS-T databases. In addition to the data collected and published by government agencies, an increasing amount of data is now created and contributed by private companies and the general public. With the widespread adoption of information and communications technology (e.g., the Internet and smartphones) and location-aware technology (e.g., Global Positioning System (GPS), Wi-Fi positioning system (WPS)) in today’s world, private firms and the public not only are users of GIS-T services provided by government agencies but can also become active contributors to GIS-T data collection and applications. For example, OpenStreetMap (www.openstreetmap.org) is a collective effort of volunteers to create a worldwide street database based on their local knowledge, GPS tracks, and other public data sources. Waze (www.waze.com), on the other hand, is a community-based traffic and navigation application that allows users to share real-time traffic and road information. Such volunteered geographic information (VGI) and crowdsourced data contributed by the user community enables GIS-T databases to be rapidly and inexpensively updated.

It is useful to have rich GIS-T spatial databases for the inventory and maintenance of current transportation systems and services. However, a major goal of GIS-T is to address and solve the different transportation problems and challenges facing the modern world. It is therefore important for GIS-T to offer analysis and modeling functions in support of transportation applications. There have been a large number of analysis methods and models that have been developed independently from GIS to address a wide variety of transportation problems such as network analysis, travel demand analysis, environmental impact assessment, traffic engineering design, and transportation logistics. Each analysis method or model may have its own data requirements and data structure, which may not be compatible with the data model used to build GIS-T databases. As a consequence, integration of these analysis methods and models with GIS-T databases can be a challenge.

Several approaches can be used to integrate analysis methods and models with GIS-T databases. It is unrealistic to expect the inclusion of all transportation analysis and modeling functions in general-purpose GIS software designed to support a wide range of GIS applications. In this case, data conversion programs can be developed to handle data exchanges between an external transportation analysis/modeling system and GIS-T databases. This approach, which is known as loose coupling, provides a way to benefit from both GIS and other external software. A tight coupling approach creates closer interconnections between GIS-T databases and an analysis/modeling system. With an open development environment offered in some GIS software packages, a transportation analysis or modeling procedure can be implemented using a GIS programming language supported by the open development environment to deliver problem-solving applications. Finally, there are
GIS FOR TRANSPORTATION

GIS software systems developed specifically for transportation and logistics applications. In such systems, GIS-T databases are fully integrated with built-in transportation/logistics analysis and modeling functions. Nevertheless, no single GIS software product is likely to cover all transportation analysis/modeling procedures. Each application developer should choose an approach that best fits their needs.

Most transportation professionals and policymakers and the public have limited knowledge of GIS-T principles. It is important to present complex analysis procedures to them in an intuitive and easy-to-understand manner, as Google has done by keeping the platform simple for users. Web-based GIS-T applications have significantly improved information access and allowed users to interact with online applications to answer specific questions. A government agency can share GIS-T databases and applications with other units in the agency and the public. GIS-T user interfaces can be customized to meet the knowledge level and needs of different user groups. Internet-based public participation GIS (PPGIS) applications can also be implemented to facilitate communications in transportation decision-making processes such as soliciting public participation in transportation improvement programs. We may not hear about GIS-T as often today as we did in the past, but this is not an indication that it has become irrelevant. On the contrary, it suggests that GIS-T has matured and become a part of our daily lives.

Future directions for GIS-T

The main objectives of developing geographic information systems for transportation are to facilitate the current operations of transportation systems and to tackle future transportation challenges. Our goal is to achieve efficient, equitable, safe, and sustainable transportation systems and services. However, this goal cannot be achieved without considering future GIS-T trends. Although there are many ongoing research topics that could influence future directions of GIS-T, here the focus will be on web GIS/cloud computing/CyberGIS, big data, three-dimensional and dynamic GIS, and a hybrid physical-virtual world. As discussed, GIS-T should provide communication tools to make information and services accessible to transportation professionals, policymakers, and the public. Web GIS has not only made GIS-T data accessible online, but also made applications available online for monitoring transportation operations (e.g., tracking current locations of buses or delivery trucks), improving transportation efficiency (e.g., adjusting traffic signals based on real-time and projected traffic), and promoting sustainable transportation (e.g., developing web applications in support of a new model of shared rides such as Lyft). The recent move toward cloud computing, which delivers services on demand via the Internet, can free users from maintaining GIS hardware, software, data, and applications on site, and can be scaled to accommodate future expansion.

“Big data,” a term now in wide use, will have major impacts on the future direction of GIS-T. Volume, velocity, and variety are frequently used to describe the major characteristics of big data. Big data volume is not new to the transportation community, which has long worked with a large number of transportation links, nodes and features, large volumes of remote sensing data, and data collected continuously from various sensors used in intelligent transportation systems. Velocity in the big data era refers to the speed of data production as well as the processing speed needed to handle big data volumes. Traditional computing approaches have
limitations in processing and analyzing big data. High-performance computing (HPC) presents an alternative to speed up the processing of big geospatial datasets by using multiple computing units to process the same task in a coordinated manner. Many transportation problems can be partitioned either by geographic areas or by tasks and then distributed to different computing units for processing. For example, we can find the best route for different users by simultaneously running their requests on different computing units. GIS-T also incorporates a wide variety of datasets in support of different transportation tasks and applications. For example, data collected from intelligent transportation systems can include continuously streamed video and other real-time traffic data. Mobile phone tracking data has become a valuable data source for gaining insights on human mobility at a scale that conventional surveys cannot deliver. Smart transit fare cards now provide detailed origin–destination data for individual travelers. Transportation researchers have also used online social network data to study travel and activity patterns. It is important for GIS-T to manage these large and diverse datasets in an integrated environment to support geographic query, analysis, and visualization functions for transportation studies. Future GIS-T needs to support multiple data models with an open data/science approach and to take advantage of advances in CyberGIS to address challenges in the big data era.

GIS-T developments are also moving toward three-dimensional and dynamic databases and applications. The conventional GIS topological data model has served GIS-T well in the past. However, its planar enforcement requirement is designed for a two-dimensional space, whereas transportation systems and operations are basically three-dimensional in nature. Furthermore, pedestrian navigation systems, which involve off-network trips and navigation in three-dimensional buildings, have attracted significant attention in recent years. Recent GIS data models have relaxed the planar enforcement requirement to permit some three-dimensional analysis functions (e.g., a three-dimensional navigation network for multifloor buildings), but they do not yet support a complete set of 3-D operations. We also begin to see photos and images displayed with 3-D vector features to create augmented reality environments that are useful in navigation and other transportation applications. Time is another dimension that the GIS community has worked on for several decades. Time may appear to be a simple concept in our daily lives, yet it is very challenging to develop a temporal GIS data model that can properly manage, analyze, and visualize historical as well as dynamic data (e.g., real-time tracking of moving vehicles or people, continuously changing traffic flows). The map layer approach employed in the conventional GIS data model is static in nature since each map layer represents the state of a theme at a particular time (e.g., a land-use map in 2010). We must either update an existing map layer or create a new map layer to reflect changes over time. This approach may be sufficient to handle changes between different time points but it is inadequate to represent and analyze processes that take place at various spatial and temporal scales, especially if we want to treat time as the fourth dimension to investigate possible causal relationships between different processes. With the widespread adoption of sensor and mobile technologies, an increasing amount of real-time data is being collected, and the public also increasingly expects to access real-time or near real-time data to assist with travel activities. This is why we have seen a surge in online applications such as Hail a Cab, which allows users to locate and summon a taxi in real time (or schedule a future pickup), or Uber and Lyft, which offer online services for
users to find vehicles for hire or ride sharing. To accommodate these changes, there is no doubt that GIS-T must move from static to dynamic approaches in the future.

Finally, it is important to note that technology advancements are changing the way that economic, social, and transportation systems operate in the modern world. Batty (2013) suggests cities should be understood as a system of networks and flows beyond the simple view of places in space. In particular, the growth of online shopping has changed people’s shopping behaviors and products’ shipment patterns. Online social network services and the Internet in general have changed individual time-use patterns, while smartphones and other mobile devices allow us to constantly stay connected and be more spontaneous in planning activities and trips. Many of the activities we carry out in virtual space using information and communications technology can influence our activities in physical space, and vice versa. These changes are likely to lead to different economic and social systems, which in turn suggests that we may need to design and develop very different transportation systems to meet future societal needs. GIS-T, which now represents and analyzes entities, activities, and interactions taking place in physical space, may become inadequate for this changing world. Future GIS-T must offer an integrated physical–virtual environment that allows us to manage, analyze, and visualize activities in both physical and virtual space as well as interactions between what happens in physical space and what occurs in virtual space.

Geographic information systems for transportation have evolved into a mature application field of GIS over the past few decades. As the world continues to change, our needs for transportation will change accordingly. GIS-T, along with other technology advancements, can assist the transportation community in tackling the challenges and to design better transportation systems and services to meet future needs.

**SEE ALSO:** Accessibility, in transportation planning; Data model, moving objects; Economic geography: spatial interaction; Geographic information system; Graph theory; Information and communications technology; Information technology and mobility; Logistics; Public-participation GIS; Representation: time; Representation: trajectories; Routing and navigation; Sensor networks, the sensor web, and the Internet of Things; Spatial database; Spatiotemporal analysis; Sustainable transport; Time–space convergence; Transport networks; Transport technology; Transportation and land use; Transportation planning; Urban planning: human dynamics; Urban transit; Visualization

---

**References**


---

**Further reading**


Before geographic information systems (GIS) were introduced to the world, and to the world of academic geography, place and the uniqueness of place was considered all-important. In the 1950s and early 1960s the uniqueness of place was challenged by the rise of quantitative and theoretical approaches to geography. In the 1960s the rise of geographic information systems gradually led to an ascendancy of a spatial and scientific approach to geography. By the late 1980s, GIS were providing jobs for students who were educated and trained to use and develop these new software tools. However, throughout the decades since GIS were developed there were those in the discipline who celebrated the uniqueness of place and the value of qualitative geography and a plethora of nonquantitative and aspatial approaches to geography. This entry will show how, during the last two decades, GIS have attempted to accommodate and assimilate these approaches and how those working in GIS have come to respect both the uniqueness and similarities of places.
Steinetz, Parker, and Jordan also recognized the implied use of overlay analysis in the results of a competition for the design of a new plan for the city of Düsseldorf in Germany that were published in 1912. Steinetz and his colleagues described other studies that used overlay procedures throughout the first half of the twentieth century, but it was not until the work of Jacqueline Tyrwhitt and Jack Whittle appeared in the *Town and Country Planning Textbook; An Indispensable Book for Town Planners, Architects, and Students* (cited and discussed by Steinetz, Parker, and Jordan 1976) that detailed and explicit accounts of the overlay process and ensuing spatial analysis were explained. The technique was quickly popularized by many others including, most famously, Ian McHarg (1969). Because of the popularity of McHarg’s book, *Design with Nature* (now cited well over 3000 times in the academic literature), he is regarded as having had a prodigious influence on the development of GIS, but *Design with Nature* did not provide a description of a methodology for computerizing the overlay process, although such a methodology is indeed articulated in the article by Steinetz, Parker, and Jordan (1976).

It is important to note that before software and hardware systems were developed that were formally recognized as “geographic information systems” there were numerous developments that occurred within the classified military arena. Clarke and Cloud (2000), for example, have provided a detailed assessment of the importance to the development of the associated field of analytical cartography of the US CORONA program of reconnaissance satellites that was operated by the Central Intelligence Agency from 1959 to 1972 along with the associated SAGE program for processing the imagery.

**The era of geographic information systems**

In their account of the early history of GIS, published as part of the first major review of the field, Coppock and Rhind (1991) divided their retrospective into four periods: (i) the pioneer period from the mid-1950s to approximately 1975; (ii) the government-supported, experimental period beginning in the mid-1970s and ending in the early 1980s; (iii) the commercial period from the early 1980s to 1990; and (iv) the user dominance era starting in 1990. For all four periods Coppock and Rhind discuss the significance of conceptual developments, progress and improvements in software and hardware, and the contributions of academia, commercial enterprises, and governments. This discussion follows Copock and Rhind’s schema. For an alternative view of the early days of GIS see Foresman (1998). Foresman’s edited book is noteworthy for providing a more operational and raster-oriented coverage of the origins of GIS, and for including many chapters written by the “pioneers” of GIS. His chronology of the evolution of GIS includes a number of overlapping “ages”: the pioneer age, mid-1950s to the early 1970s; the research and development age, the very early 1970s to the mid-1980s; the implementation and vendor age, the early 1980s to the the mid-1990s; the client applications age, the early 1990s to the early 2000s; the local and global network age, the late 1990s onwards into the 2000s. These developments are then correlated with developments in academic geography, computing, and environmental awareness.

**The pioneer period: the mid-1950s to the mid-1970s**

Conceptual developments during this period included the overlay or “layer-cake” model for
the organization of geographical data discussed previously and the popularization of the geographical data matrix by Brian Berry. Originally, the data matrix was organized with the columns representing places and the rows representing their attributes. When Berry’s model was incorporated into early GIS software this arrangement was transposed to make the geographical matrix more compatible with standard database technology. Waters (1998) discusses contributions from the field of operations research, where work on decision support systems by IBM researchers was being incorporated into planning-based software.

During the 1960s the first fully functional vector-based GIS, the Canada Geographic Information System (CGIS), was developed. The CGIS, initially a collaboration between Roger Tomlinson’s company, Spartan Air Services of Ottawa, and the Canadian Government’s Canada Land Inventory, produced a series of innovations, including hardware for laser scanning technology and software for vectorizing the resulting images (Waters 1998) and for storing raster layers efficiently, with such developments as Morton ordering. Because of Roger Tomlinson’s seminal contribution to the origins and development of GIS he has often being credited with being the “father of GIS”. In these early years, those working in the field of GIS were not simply taking advantage of developments in the field of computer science, as computer science did not offer all that was needed for the development of GIS. As the field matured, GIS would indeed take advantage of computer hardware developments and the move away from mainframe computing. External developments in database technology also proved beneficial to the development of GIS.

During this seminal period the Harvard Laboratory for Computer Graphics and Spatial Analysis, established in 1965 by Howard Fisher, laid the foundations for subsequent developments. Thus, the Harvard Laboratory made major algorithmic contributions and produced widely adopted computer mapping packages, such as SYMAP, CALFORM, SYMVU, GRID, POLYVRT, and ODYSSEY. The first three packages for producing line printer maps and 2-D and 3-D plots, respectively, were adopted throughout North American universities. William Warntz became the Director of the Harvard Laboratory in 1969 and made further conceptual contributions, including a recognition that the critical features of surfaces could be used to produce triangulated irregular network (TIN) models to provide more compact storage of surface features. Associated with the Harvard Laboratory at various times was Jack Dangermond, founder (in 1969) and co-owner of the world’s leading GIS company, Esri, which now produces the industry-dominant ArcGIS software. For a complete history of the early days of the Harvard Laboratory see Chrisman (2006). In the United Kingdom, the experimental cartography unit (ECU) was founded in 1967 by David Bickmore, head of the cartography unit at the Clarendon Press. Like the Harvard Laboratory, the ECU stimulated and championed the possibilities of computer-based mapping, providing the incentive for the British Ordnance Survey to move into automated, computer-based mapping.

In 1970 the US Census Bureau produced the first geocoded census. The topological structure of street segments was coded by identifying the IDs of right and left blocks and “from” and “to” nodes. Also recorded, using X,Y coordinates, were the address ranges of street segments. These files were known as DIME (dual independent map encoding) files and were the forerunner of the US Census Bureau’s more sophisticated TIGER (topologically integrated geographic encoding and referencing) files (Mark et al. 1997).
GIS: HISTORY

One of the primary ways in which new knowledge about GIS has been disseminated is through the organizations that have catered to academia, government, and industry. The Urban and Regional Information Systems Association (URISA) is one of the oldest such organizations. URISA held its first conference in 1963 and has continued to hold annual conferences up to the present. Over the decades the proceedings of the URISA annual conferences have been a major source of information concerning new developments in the evolution of GIS.

The government-funded experimental research period: the mid-1970s to the early 1980s

Conceptual and software developments during this period were taking place within academia, government agencies, and industry. For example, algorithms to solve location-allocation problems that had been developed in the mid-1960s were now available in stand-alone programs and were also being integrated into software systems, such as GADS (geodata analysis and software display system) developed by IBM’s research division (Waters 1998). In Europe, government-sponsored research led to the development of the Swedish road data bank and other computerized spatial databases.

Developments in GIS reflected advances in the field of computer science. During this period, mainframe systems had given way to minicomputers based on time sharing and, eventually, to desktop microcomputers, enabling the gradual movement of GIS software to these new computing platforms. One of the most prominent and earliest of the government-supported, mainframe GIS of this period was the Minnesota land management information system (MLMIS). This was a raster-based (as opposed to the vector-based CGIS) resource inventory, where the pixel cells had a resolution of 40 acres. By the end of this period, such mainframe systems were gradually becoming obsolete, due to high maintenance costs, the problem of data currency, access issues, and nonuser-friendly command line interfaces. The introduction of powerful workstations in the early 1980s led to the gradual demise of large mainframe systems.

In 1974 the first Auto-Carto conference was held. These conferences, although not held annually, are still held on a regular basis and are now sponsored by the Cartography and Geographic Information Society. Similarly, in Europe, a series of fourteen EuroCarto meetings were held from 1981 to 1997 (every year except for 1982, 1988, and 1996). In 2015, these conferences re-emerged under the sponsorship of the International Cartographic Association (ICA) as the 1st ICA European Symposium on Cartography (EuroCarto 2015).

The commercial period: the early 1980s to the late 1980s

Tomlinson (1987), writing in the first volume of what was to become a flagship journal of GIS, provided a state-of-the-art review of this newly emerging subdiscipline of geography. He noted that significant progress had been made in adopting GIS software in government and commercial organizations, including in the transport and facility planning and management, cadastral systems, agriculture and the environment, and the forestry and civil engineering sectors. He argued that in the future new innovations would come from academia and from government rather than the commercial sector. This was only to be partially true for, in 1982, Esri released ARC/INFO, the first commercial GIS. ARC/INFO adapted the CGIS model of handling the spatial and attribute data separately. The former used Esri’s topological
ARClstructure while the latter were stored in the INFO relational database. In 1986, Esri released PC ARClINFO due to the popularity of the IBM PC desktop computer. By 1988, Esri had become a $40 million a year company with clients in forestry applications and other government departments.

Tomlinson (1987) had been correct in arguing that the expansion of commercial GIS was being curtailed by the lack of educational opportunities to produce personnel who could run and maintain GIS facilities, conduct basic research, stimulate new innovations, and staff university departments. In 1988, to motivate fundamental research into the development of GIS, the US National Science Foundation (NSF) made a grant to a consortium of universities that included the departments of geography at the University of Santa Barbara (the lead institution where Mike Goodchild was the Executive Director), the State University of New York at Buffalo, and the surveying engineering department at the University of Maine. This consortium is now funded with a grant from the NSF of $5 million annually. In addition to its primary focus on research, which manifested itself in a series of research initiatives, the NCGIA (National Center for Geographic Information and Analysis) also developed a core curriculum in GIS comprised of 75 lectures, grouped into three semesters each of 25 lectures. The lectures were initially written by 35 different authors and contained additional laboratory material. They were remarkably successful and by January 1995 over 1300 copies had been distributed to over 70 countries. Eventually, the original core curriculum was made available in Chinese, French, Hungarian, Japanese, Korean, Polish, Portuguese, and Russian. The core curriculum was intended to provide a core set of knowledge that would allow faculty teaching GIS courses to cover a broad set of topics to students new to the discipline.

In February 1987, the UK’s Economic and Social Research Council established four regional research laboratories (RRLs), in London, Edinburgh, Cardiff, and Newcastle. Their mandate involved four primary functions: data management (the provision of a spatial data archive); software development; spatial analysis; and research training together with professional development. Concurrent with the development of the RRLs, the UK’s Lord Roger Chorley chaired the Committee of Inquiry into the Handling of Geographic Information. The Committee’s Report was made public in 1987 (Waters 1998) and made recommendations (subsequently acted upon) that the British Ordnance Survey, the primary supplier of maps to the British public, should move to a fully digital environment.

The period of user dominance: the end of the 1980s to the mid-1990s

Coppock and Rhind state that during the last of their four periods there was intense competition among GIS software manufacturing companies due to “user dominance” and an associated thinning out of the market to a small number of major vendors. This competition was fostered by the numerous GIS conferences and by the publication of the annual GIS World Sourcebooks (Waters 1998). These volumes were noteworthy for their annual survey of GIS companies and their software, which included detailed tables assessing the capabilities and attributes of the listed GIS software. This ongoing competition resulted in the emergence of a few dominant companies among the GIS software vendors, including Intergraph and Esri. During this period, vendors moved away from the complexity of command line interfaces to
graphical user interfaces (GUIs), once again tracking ongoing developments in computer hardware and commercial operating systems (the so-called WIMP interface of windows, icons, menus, and pointers). It was in 1990 that the US Census Bureau introduced its TIGER file system: this, together with the fact that Census Bureau data was freely available, further boosted the US geodemographics industry. Other countries including Canada and the United Kingdom made census data only available on a so-called “cost recovery” basis, arguably imposing a continuing chill on what might have been a more robust market for GIS data. In recent years many countries have permitted more open access to geospatial data under the Aarhus Convention, which became law in October 2001. However, the only states that signed the agreement were from Europe and Asia.

A number of companies supplied GIS software that specialized in niche markets. For example, Caliper Corporation produced TransCAD, a software package focused on transportation planning and transportation GIS (GIS-T). Others, such as Idrisi, developed by Clark University in Worcester, MA, successfully catered to the academic market, providing detailed tutorials and accompanying datasets. It was during this period that the movement to introduce GIS instruction into the K-12 curriculum in the United States gathered momentum. These efforts were summarized and resource materials and advice for teaching GIS in primary and secondary schools were provided at the first national conference on the educational application of geographic information systems held in 1995 and sponsored by Technical Educational Research Centers (TERC). TERC has continued to provide educational resources to schools for all grade levels (and for university GIS courses as well) and now maintains an active website.

The era of geographic information science (GIScience)

The mid-1990s saw a sea-change in the development of the academic discipline of GIS. In 1992, Goodchild published a major theoretical contribution when he argued that the discipline should move from a concern with the technology of geographic information systems to developing answers to questions that might more properly be considered to be part of a geographic information science (Goodchild 1992). In the opening arguments of his paper, Goodchild stated that much of the early history of GIS was technology driven. It was concerned with how to get geographical data into an information system and with new technologies that were developed largely within government agencies, such as the CGIS and the US Census Bureau, and by remote sensing companies that were developing new technologies to acquire satellite imagery. Now, he argued, it was more appropriate to concentrate on the task of how to “handle” and exploit the data held in these GIS databases. The way forward was prepared by research that was being reported at a newly developed series of conferences, the International Geographical Union’s spatial data handling symposia (the first of these conferences was held in Zurich in 1984 and the 17th of these biennial meetings in Beijing in 2016). Goodchild presented a review of the nascent discipline of GIScience, including a synopsis of three decades of algorithmic research, into a series of topics: spatial data handling; data collection and management; data capture; spatial statistics; data modeling and theories of spatial data; data structures, algorithms and processes; data display (visualization); analytical tools; and, finally, institutional, managerial, and ethical issues. While some of these developments were to benefit from progress in computing (e.g., data display from a new emphasis on scientific
GIS: HISTORY

visualization, often referred to as the second revolution in computer science), most launched new subdisciplines in geography (e.g., spatial statistics, which now produced new stand-alone software packages such as GeoDa, and new functionality in commercial software, such as Esri’s ArcGIS ToolBox, Spatial Analysis and GeoStatistical Analysis modules). Some of these new fields were now located in other academic departments such as information science and geomatics engineering, mirroring developments that had occurred decades before when fields such as geomorphology and climatology moved out of academic geography.

Within a few years a number of journals had changed their names to reflect this new emphasis on science as opposed to systems. For example, in 1997 the International Journal of Geographic Information Systems had replaced the word Systems in its name with Science. The new name was explained in the introduction to that volume by then editor, Peter Fisher, who acknowledged the debt to Goodchild’s seminal article in providing a rationale for the new focus to the journal and the discipline. The American Cartographer, which had been launched in 1974, had changed its name in 1990 to Cartography and Geographic Information Systems. This lasted until 1999, when it also changed the Systems in its title to Science. Other journals, such as Transactions in GIS, circumvented the problem by leaving the acronym in their title undefined. When the 2nd edition of Geographical Information Systems: Principles, Techniques, Applications and Management appeared in 1999, the word Systems was retained, but when the same four authors published the first edition of a textbook two years later, they included both terms, Geographic Information Systems and Science (Longley et al. 2001). The Annals of the Association of American Geographers requests submissions in four broad areas, one of which is Methods, Models and Geographic Information Sciences but its Table of Contents retains the now more ambiguous acronym, GIS, and its specialty group in the field prevaricates, as well, with the title Geographic Information Science and Systems. The new NCGIA core curriculum, which gathered lectures for its website until 2000, also used geographic information science in its title. The lectures and associated material were now all online and no print copies were created. A note posted on the website in August 2000 stated that there would be no further updates due to new educational resources becoming available, including the digital library of Earth system education, the NCGIA’s own center for spatially integrated social science, and Esri’s virtual campus (Waters 2013). A review of progress in the field of GIScience, twenty years after the initial discussions of the concept, has been given by Goodchild (2010).

New educational initiatives also characterized this period. In 1990, the NCGIA board of directors recommended that a new organization be developed to assist researchers in GIScience. The UCGIS (University Consortium for Geographic Information Science) website, in its brief history of the organization, states that an ad hoc steering committee was formed by the NCGIA in 1991 containing 16 members from seven different disciplines. Unfortunately, this brief historical note does not list either the members of the committee nor the seven disciplines represented. In 1994, a founding meeting was held in Boulder, CO, and in 1995 the UCGIS was formally incorporated. The UCGIS had two primary objectives listed on its website: “Advance research in the field of geographic information science” and to “Build scholarly communities and networks to foster multidisciplinary GIS research and education.” The membership of UCGIS has fluctuated over the years, reaching a high of perhaps 70 US universities in about 2005 but declining to 53 US universities at the time of writing. In
GIS: HISTORY

the initial meetings it was made quite clear that this was a US organization and non-US universities would only be granted affiliate status. No Canadian university has ever been a member and at present there are only two non-US universities that have been awarded and retained affiliate status, CentroGeo in Mexico and the University of Salzburg’s department of geoinformatics.

The UCGIS has promoted annual meetings and symposia and, in partnership with the Association of American Geographers, the development of a body of knowledge (BoK) for teaching GIS. This was seen as a natural successor to the core curricula developed by the NCGIA. However, in conceptualization the BoK was vastly different. Waters (2013) provides a critical review of the development process that led to the publication of the BoK in 2006. The BoK was structured into ten knowledge areas that were then divided into 73 units that, in turn, were further subdivided into 330 topics. Contributors to the BoK had suggested that one of the primary weaknesses of the core curricula had been the lack of an explicit structure for continuous updating. Unfortunately, this was equally true of the BoK and it was not until the UCGIS held a special workshop at its summer symposium in Pasadena in 2014 that a new BoK was proposed with a formal timeline. In the field of computer science there is a process for updating the curriculum approximately every six years. This has yet to be achieved by GIScience.

Nevertheless, by the end of the 1990s Mike Phoenix, former higher education specialist for Esri, was able to claim that their GIS software products were used in more than 60 different kinds of university academic and administrative departments (Waters 1998). By this date, GIS had influenced the teaching of many of the subdisciplines of geography and would, in the following decade, go on to influence the operations of many other disciplines that used spatial data.

Reactions to a formal geographic information science approach to GIS

Some academic geographers reacted with dismay to the new emphasis on geographic information science, seeing it as a reassertion of the aridity of the spatial paradigm of the so-called “quantitative revolution,” which had been disavowed by some members of the discipline in the 1970s. The most widely cited critique of the spatial paradigm within GIS came in the form of John Pickles’ edited volume, Ground Truth (Pickles 1995). Although some of the leading proponents of GIScience contributed essays to the book, most of the papers were written by academics unsympathetic to the achievements of those working in GIS.

Shortly after the publication of Ground Truth, the public participation in GIS (PPGIS) movement was launched at an NCGIA sponsored workshop held at the University of Maine. PPGIS may be defined as an approach to community planning that privileges the traditionally marginalized segments of society (the poor, the old, the young, the physically challenged, and members of ethnic minorities, among many others). It leverages GIS software to allow these individuals to participate in the planning of their communities and their environments. To achieve these goals PPGIS broadened traditional approaches to GIS by including a wide range of social science methodologies and theories. It is closely aligned with participatory GIS, a cognate set of methodologies and approaches and low-tech solutions, such as sketch maps, to the acquisition of geographic information. Typically PPGIS and PGIS involve the interaction and collaboration of academics with GIS expertise.
with members of the community wishing to participate in the planning of their communities. These approaches had antecedents, paradoxically in the later work of Bill Bunge, one of the pioneers of the quantitative revolution (Miller and Goodchild 2015). PPGIS has also produced new lines of research, such as the current concern with the design of age-friendly communities.

In the twenty-first century, stimulated by the popularity of PPGIS, which had spawned a series of conferences in conjunction with the annual URISA meetings, GIS began to embrace a host of traditionally nonscientific approaches to geography, including qualitative, feminist, and critical GIS to name a few. PPGIS itself had rarely spelled out whether it was concerned with geographic information systems or science, the latter being somewhat of an anathema to those practicing PPGIS. The more important of these new approaches to GIS included formal adoptions of critical GIS and feminist approaches to the use of the technology. The possibility of the former was introduced in an article by Nadine Schuurman (2000) who identified three waves of criticism of the GIS community. She argued that the first wave had lasted from 1990 to 1994 and the debate had centered on the merits of the positivist approach adopted by GIS practitioners. The second wave occurred in the mid-1990s and represented the beginning of a dialogue between GIS specialists and those concerned with both the social effects and implications of GIS technology. The catalyst for this discourse was the NCGIA’s “Research Initiative 19, GIS and Society: The Social Implications of How People, Space, and Environment Are Represented in GIS.” According to Schuurman, in the third wave, at the end of the 1990s, there was a greater commitment to the technology of GIS. This led to the rise of PPGIS in that decade and ever since. Quite separately, there was a call for a “critical GIS” or “GIS2” that would incorporate the various perspectives of social theorists (see Sheppard et al. (1999) for a discussion of the societal implications of the NCGIA’s Varenius Project, and also the papers in the special issue of Cartographica, 2009, 44(1)).

GIS developments since 2005: the world of volunteered GIS, web-based mapping and mobile GIS, and cloud computing and big data

Since 2005, developments in GIS have again mirrored the importance of progress in computing technology. Crowdsourcing and social networks have been hugely influential developments in the social application and individual use of computing technology, but these changes have been reflected in the move away from the use of authoritative data in GIS and the rise in influence and use of volunteered geographic information (VGI). The attractions of VGI are well described in Goodchild’s (2007) seminal paper on the topic: VGI is inexpensive, timely and available to all on the right side of the digital divide. Goodchild (2010), in reviewing progress in the field of GIScience, notes that the breaking down of the barriers between expert and nonexpert is also referred to as neogeography and that VGI itself has been called crowdsourced geographic information and community mapping, and thus has strong links to PPGIS. VGI has been aided by the move to web-based mapping and the availability of online GIS technology such as ArcGIS Online and the almost universal availability of map-based apps on smartphones and in-car GPS navigation systems (see Miller (2007) for a review of these technologies that have produced a time–space convergence and aided the processes of globalization). GIS and computing hardware are becoming invisible
technologies whose benefits can be leveraged and enjoyed seamlessly by wide sectors of the population in industrialized and developing countries.

Other software developments that have led to the almost universal use of geographic information and related technologies include Google Earth, OpenStreetMap, and software for finding directions, such as MapQuest. The familiarity of large sections of affluent populations with this type of software has led to the rise of location-based services and to new developments such as OpenImageMap and web mapping services.

Advances in computer technology, including cyberinfrastructure (leading to CyberGIS), cloud computing, and the rise in importance of “big data,” are almost immediately integrated into GIS software, with leading GIS companies such as Esri providing online guides to how big data may be used in such GIS applications as predictive modeling and social media. A review of the implications for the discipline of geography in general, and GIS in particular, specifically a move to a “data driven” geography, is provided by Miller and Goodchild (2015). Traditionally, big data was concerned with the three “Vs”: volume, velocity, and variety. The terms are self-explanatory and geographic data often qualify on all three counts. More recently, computer scientists have added additional “Vs” to the list, including veracity (relating to the provenance of the data: thus, for example, VGI might have lower veracity than authoritative data) (Miller and Goodchild 2015), value, validation (whether the data conforms to a priori model expectations), and voracity (the rate of ingestion, which is important in the use of virtual reality simulations with a GIS framework). Vagueness (or uncertainty) might also be added to the list as it is endemic to much GIS data.

GIS returns to its roots in geography and GIS

It can be argued that GIS has returned to its origins, from earlier concerns common to many geographers, at least on the human geography side of the discipline. Miller and Goodchild (2015) summarize the history of geography’s concerns with idiographic (description-seeking) and nomothetic (law-seeking) approaches. At various times over the centuries one approach has been dominant and the other subordinate, disparaged, or overlooked. Miller and Goodchild argue that Strabo and Ptolemy were concerned with both approaches, for they provided detailed descriptions of specific places as well as generalizations that applied to the Earth or large regions. However, Miller and Goodchild also state that such an integration has rarely been adopted by researchers in the two millennia since Strabo and Ptolemy wrote. GIS, with its emphasis on software and algorithms and its concern with creating spatial databases, is seen by Miller and Goodchild as an intellectual activity that permits a dialogue between the idiographic and nomothetic world views. Perhaps this is best demonstrated in the various spatial interpolation algorithms, some of which produce global models (trend surface analysis, for example) while others emphasize local differences (e.g., geographically weighted regression). Miller’s (2007) call for a people-based approach to GIS also emphasizes individual differences.

A recent trend in GIS has been the recognition of the importance of geodesign. This is now the focus of a new series of conferences being held annually in Redlands and also in Europe and China. It has led to new undergraduate and graduate degree programs. Geodesign has been championed by Esri cofounder, Jack Dangermond, and reflects his graduate education in landscape architecture. Dangermond (2014)
has defined geodesign as taking geographic information and linking “it to the design, decision-making, and planning process using collaboration … by building the power of GIS into the process, allowing alternative plans to be visualized, compared, and evaluated.” He credits his former Harvard University professor Carl Steinitz as introducing him to “this methodology” and so the history of GIS returns to its roots.

In 2007, the NSF began supporting research into “transformative science.” There is little doubt that over the past fifty years GIS has been a transformative science for both the academic discipline of geography and the world we inhabit. The impact of GIS on the discipline of geography has been compared often to the impact of the telescope on astronomy but, as with the telescope, people choose what they want to see through the GIS lens. The NSF is interested in how transformative science has changed the world. It would behoove the discipline of geography to do as Sheppard et al. (1999) suggested and study how GIS has transformed the world.

**SEE ALSO:** Cloud computing; Critical geography; CyberGIS; Digital divide; Feminist geography; Geodemographic profiling; Geodesign; Geographic information science; Geographic information system; Geostatistics; GIS for transportation; Local statistics and place-based analysis; Location-allocation models; Network analysis; Overlay, graphical; Participatory geographies; Public-participation GIS; Qualitative GIS; Spatial analysis; Spatial crowdsourcing; Spatial social networks; Volunteered geographic information; Volunteered geographic information: quality assurance; Web-mapping services

---

**References**


Glacial depositional processes and landforms

Paepin Goff
Texas State University, USA

Glacial landforms are primarily the result of erosion during glacial advances and deposition during glacial retreat. The glacial landforms that exist today are the product of both current glaciers and past glaciations. Climatic conditions allowing for the existence of glaciers are found in polar and alpine regions where low temperatures and high levels of precipitation prevail.

In polar and alpine regions, the amount of snow accumulation outpaces ablation, or snow loss from melting, evaporation, wind erosion, and calving (see Figure 1). The difference between accumulation and ablation is the mass balance of a glacier, which is used as a measurement of glacier health. According to the 2011 United Nations Educational, Scientific, and Cultural Organization (UNESCO) Glossary of Mass Balance and Related Terms, annual mass balance is monitored on more than 100 glaciers (UNESCO 2011), and monitoring efforts are expanding.

A glacier has a positive mass balance when it experiences more accumulation than ablation, which means the glacier will grow. In contrast, a glacier has a negative mass balance when it experiences less accumulation than ablation, meaning the glacier will recede (see Figure 2). A glacier that experiences equal amounts of accumulation and precipitation is said to be in equilibrium.

Glaciers with mass balances that average positive over time are said to be actively advancing. As net accumulation builds annually, the snow compresses under its own weight. Fluffy surface snow will compress into a dense layer of snow called firm, which has small air pockets. Increased compression creates an even denser snowpack layer called glacial ice, which forces air into tiny bubbles trapped in the ice.

The same mechanisms that allow for the creation of a glacier contribute to glacial flow, or movement, including:

- ice properties (weight, thickness, steepness, temperature, density),
- topographic relief,
- bedrock properties (yielding potential, temperature, mineral compositions),
- subglacial hydrology,
- terminal environment,
- mass balance.

These factors contribute to glacial flow at the internal, basal, and subglacial levels.

Glacial spreading is known as internal deformation, or creep, and is the result of ice crystal changes as a glacier compacts. Flow at the bottom of a glacier is called basal sliding, which occurs when water collects between the ice and the bedrock layer. Bedrock deformation contributes to flow at the subglacial level, which happens when soft sediments yield to glacial weight and movement. Subsurface hydrology also plays a role in glacial flow by weakening the bedrock.

Multiple types of glaciers contribute to glacial landforms including alpine, or mountain valley glaciers, and continental glaciers, or ice sheets. In general, features created by alpine glaciers are smaller and more distinct than those created by continental glaciers owing to differences in...
Figure 1 Columbia Glacier calving, Prince William Sound, AK, USA, 2005. A dramatic iceberg calving from Columbia Glacier in Prince William Sound, Alaska. The iceberg has just broken free from under the water and shot to the surface, spinning towards the ice face. The ice cliff here is about 70 m (229.7 feet) tall. Icebergs are calved as stress fractures in the glacier merge, eventually resulting in a piece of ice cracking off and falling into the water. The stress fractures are caused by the glacier flowing over uneven stretches of ground on its way to the water. Photo by Shad O’Neel, June 17, 2005. Reproduced from USGS.

glacial flow behaviors. Alpine glaciers, subject to the natural features of mountain regions, flow down valleys from high to low elevation. In contrast, continental glaciers flow outward from their centers and extend over large areas of flat land from the poles.

Depositional landforms

During periods of glacial retreat, when ablation outpaces snow accumulation, glacial drift is left behind, resulting in depositional landforms. Glacial deposition occurs when debris is released from a glacier during transport. The debris, called glacial drift, can range in size from fine sediment particles to large boulders and can be deposited during glacial advance or retreat and during periods of stagnation. Debris deposited directly by glacial ice is called glacial till, which ranges in size from particles of clay to large boulders.

Moraines are a distinctive landform made up of sediment deposited by a glacier. Several types of moraines exist in glaciated and previously
glaciated landscapes: terminal moraines, recessional moraines, lateral moraines, medial moraines, ground moraines, and supraglacial moraines.

At a glacier’s maximum extent, or terminus, are terminal moraines, which vary in size depending on how long a glacier stays in one place. These formations, also known as end moraines, are directly related to glacial speed and the surrounding environment. The longer a glacier remains at its furthest extent, the more sediment builds at the snout, resulting in larger moraines. If a glacier pauses during retreat, it creates recessional moraines, which are located behind the terminal moraine. If the glacier pauses several times, it will leave behind multiple recessional moraines.

On the sides of a glacier, lateral moraines form perpendicular to the terminus as debris is scraped away from the surrounding path and deposited on the margins. Glaciers typically have two lateral moraines parallel to one another, which are similar in size. When two lateral moraines are pushed together by glaciers, the debris combines to form one line between the two ice streams. This process creates a debris ridge between the two glaciers, called a medial moraine (see Figure 3). As the glacier progresses and joins to form one large glacier, the material that forms a medial moraine is carried to its eventual resting place at the terminus, where it is combined with the terminal moraine. If a medial moraine persists through the life span of a glacier, it will remain as a rocky ridge after the glacier melts.

Ground moraines, also known as till plains, are landscapes of glacial deposits that resemble anything from rolling hills to flat plains. In contrast to recessional moraines, which form as a glacier pauses during recession, these landforms are the product of a quickly retreating glacier that does not pause long enough to deposit organized ridges of material. Instead, a fast-moving glacier in retreat deposits material haphazardly across the ground. The resulting landscape is an unsorted mixture of unconsolidated sediments and rocks that covers the area where a glacier previously existed.

The nature of moraine creation lends itself to the scientific study of glacial history. Investigations of terminal moraines give clues about a glacier’s origin, path of movement, speed, and
Figure 3  Medial moraine, Alaska. In this photo, merging glaciers form a medial moraine between two ice streams. Lateral moraines can be seen on the outer sides of the glaciers. Photo by Don Becker, August 26, 2008. Reproduced from USGS.

life span. Multiple terminal moraines in the same location provide evidence for multiple glaciations, which plays a large role in paleoclimate reconstructions. Recessional moraines offer insight regarding environmental conditions, which is helpful in understanding the life span of a glacier and its retreat. Lateral moraines lend themselves to the study of an alpine glacier’s path from the high elevation of the valley head to the lower elevation below. Ground moraines, which often contain other features such as kettles and kames, illustrate processes that may have played a role in glacier flow and retreat. All of the combined information provides scientists with data that can be used to model past, present, and future glacier growth and recession.

Glacial retreat leaves behind another type of plain called a sandur, or outwash plain. These glaciofluvial features are the result of sediment overflow at the terminus of a glacier carried beyond the end moraine by meltwater. The drift in an outwash plain is typically stratified, or sorted by size, with the largest materials remaining close to the terminus of the glacier and the finest sediments deposited farther out. When outwash material is deposited in a ridge-like path, rather than a plain, it is called a valley train.

In the outwash plains, tributary meltwater streams break off into smaller branches, which then split and rejoin in weaving patterns called braided streams. These streams are the result of sediment overload combined with a steep stream
slope, variations in stream flow, erosional potential of substrate material, and lateral pressure gradients. Braided channels are the result of pressurized meltwater flow – pressurized flow and higher amounts of discharge result in a high braiding intensity (Catania and Paola 2001). Braided streams contribute to distinct features in an outwash plain called *outwash fans*, which are the poorly sorted outspreading of sediments.

As a glacier recedes, ice calves off the terminus and is left in the outwash plains. These blocks of ice remain frozen in the ground and are covered by outwash sediment. When the ice finally melts, it can leave behind one of two formations. The first likely formation is called a *kettle*, which is a pond-like hole in the outwash plain, often filled with water (see Figure 4). The second likely formation is called a *kame*, which is an irregular hill or mound of sediment left behind in an outwash plain.

When streams flow through glacial ice and deposit material in a lake or pond at the terminus of a glacier, the water flow creates a landform called a *kame delta*. Although this landform exists while a glacier is still active, glacial melt can act to either collapse the delta through melt or deposit additional material. The material in a kame delta is characterized by a distinct stratigraphy owing to the sorting action of the stream during transport.

While outwash material exists beyond the terminus of a glacier, similar features exist within and beneath the glacier itself. *Eskers*, or ridges of glaciofluvial sediment, form near the terminus. Some eskers form underneath a glacier from deposits in meltwater-carved tunnels. Other eskers form on glacial surfaces or in crevasses in the ice. The main esker channel is referred to as the *tributary esker*, which allows branching eskers to form in the direction of glacial movement. These landforms often wind across several kilometers, but some extend across hundreds of miles. In the United States, esker systems in Maine extend for over 160 km. In Canada, the Thelon Esker stretches across 800 km.

Beneath a glacier, meltwater also contributes to the creation of *tunnel valleys*, which resemble steep-sided U-shaped subglacial pathways running parallel to one another in the direction of glacial flow. These long subglacial drainage pathways near the margins of large ice sheets serve as drainage pathways for large meltwater streams, which have a large impact on glacial flow at the subglacial level.

At the supraglacial level, meltwater drains through the ice, creating waterfalls within a glacier called *moulins*. These channels transport water to the basal level, affecting subglacial
GLACIAL DEPOSITIONAL PROCESSES AND LANDFORMS

water pressure and velocity and thereby affecting glacial flow. On glaciers with moulins, meltwater is delivered to the basal level at discrete points, which increases the glacial sliding speed across a surface (Gulley et al. 2012). However, because access to the basal level of a glacier is naturally restricted by ice, a limited number of direct observations of subglacial water pressure leave much to be studied in the way of glacial hydrology.

If meltwater cannot drain, or if it accumulates faster than it drains, it will collect in pools. Surface water from ablation, or surface melt, will collect in crevasses and melt ponds, while subglacial streams will collect in basal lakes. If these pools of water overflow, the result will be a glacial outburst flood, also known as a jökulhlaup, an Icelandic term that translates to “glacier run.” These floods release large amounts of meltwater and debris across a large area of land. Given their potential for destruction, international monitoring efforts focus on the prevention of harm to people who live in communities where glacial outburst floods are possible.

Meltwater also deposits sedimentary material into tunnel valleys, which carry and store large amounts of meltwater, making them effective aquifers. Similar to U-shaped valleys, tunnel valleys are wide and steep with flat bottoms. The sediment layers of a tunnel valley are made up of glaciofluvial material deposited during the creation of the original formation and subsequent sedimentary infill. The geological composition of tunnel valleys allows for the study of glacial history through investigations of sediment, valley size, and terminal location.

During the life span of a glacier, debris can be transported across a large area. Alpine glaciers can transport plucked debris in addition to material delivered to the glacier’s surface through rockslides. Large boulders are often left behind as a glacier moves and melts. These rocks, known as erratics, are distinct from surrounding native rocks in size and type.

Drumlins, or streamlined teardrop-shaped till deposits, often exist in clusters following the direction of ice movement (see Figure 5). The origin of these landforms is the subject of scientific debate, but most research points to their formation as a result of subglacial meltwater transporting sediment into ice cavities. Erosional hypotheses suggest that meltwater and outwash from fast-moving glaciers carve away weak subglacial material, leaving behind strong sediment deposits in the form of drumlins. Depositional hypotheses, on the other hand, suggest that the increased deposition evidenced by drumlins occurs when changes in ice pressure

Figure 5  Drumlin field in Northern Canada, 2014. This image shows a drumlin field in the Nunavut Territory of Canada, about 27 km (17 miles) southwest from the Amundsen Gulf. The drumlins and surrounding area seen here are thought to have formed below the Laurentide Ice Sheet during the last Ice Age. Photo by Jesse Allen and Robert Simmon, NASA Earth Observatory, June 21, 2014. Reproduced from USGS.
build mounds under the ice, creating a positive feedback loop of deposition and ice deformation.

Today, there is only one active drumlin field in the world, which is the subject of multiple investigations. The Múlajökull glacier in central Ireland, which left behind over 100 drumlins, was the subject of geomorphological studies in 2014 using aerial photographs and remote sensing data (Jónsson et al. 2014). The researchers’ data suggested that the drumlins formed mostly by subglacial erosion and bed deformation, in alignment with depositional hypotheses.

**SEE ALSO:** Glacial erosional processes and landforms; Glaciations; Glacier changes; Glacier lake outburst floods; Glacier mass balance; Glaciers; Mountain geomorphology

**References**


**Further reading**


Glacial erosional processes and landforms

Robin Blomdin
Stockholm University, Sweden

Jonathan Harbor
Purdue University, USA

Glaciers are one of the most effective agents of erosion in the Earth-surface system, and produce landforms spanning scales of millimeters to kilometers in size. Over geological time, glacial erosion wears down topography and plays an active role in the evolution of mountain ranges and continental landscapes in alpine and high-latitude settings. Glacial erosional landforms vary widely in both their size (Figure 1) and the time taken for their formation (Figure 2). On timescales of one or more glacial cycles (several hundred thousand years), erosion by glaciers in mountainous areas produces classic features of large-scale glacial relief, such as glacial valleys, troughs, fjords, overdeepenings, and cirques. On timescales of thousands of years a glacier can produce medium-scale features such as roches moutonnées and rock steps. Glaciers also erode a bedrock surface on a timescale of hours to days to produce small-scale landforms such as striations, grooves, and chatter marks.

Glacial erosion is the loosening of rock, sediment, or soil by glacial processes, and the entrainment and subsequent transportation of this material by ice or glacial meltwater. The spatial and temporal pattern of glacial erosion is complex and reflects both glaciological and topographical factors as well as characteristics of the bed material. Glaciological factors include the basal temperature (warm-based glaciers erode, cold-based ice is largely ineffective in erosion), meltwater drainage, and the amount of debris entrained in the ice, while topographic factors include glacier catchment morphology, existing slope angles, and bed roughness. Bed characteristics include lithology, rock resistance to erosion (hard beds (bedrock) and soft beds (unconsolidated sediment)), and the rheology of deformable bed materials.

Glacial erosional processes and rates

Abrasion

Abrasion is the general wearing down of rock material by clasts (rock particles) embedded in the basal layer of glacier ice. The effectiveness of abrasion is controlled by: (i) the relative hardness of the clast and bedrock, (ii) the force applied at the point of contact and the contact area, (iii) the relative velocity of the clast and bed, and (iv) the concentration of clasts in the basal ice. Clasts can both striate (scour/scratch) and polish bedrock surfaces, producing fine rock particles that are transported away with meltwater (glacial flour), and leaving behind erosional traces. Striations are formed as moving ice drags clasts across a bedrock surface. A concentration of transient stress below the point where the clast contacts the bedrock promotes crack growth and brittle failure. At the micro scale, striations are made up of multiple crescent-shaped cracks, each representing a failure event. Pervasive striation can produce glacial polish, which is a very smooth rock surface (Benn and Evans 2010; Figure 3a).
Several models have been proposed to describe abrasion, which is somewhat analogous to the action of sandpaper on wood. The harder you press down sandpaper against wood and the faster you move the sandpaper, the greater the sanding effect. Similarly, abrasion will be most effective where the force applied on a clast at the bed is large and the speed of the glacier at the bed is fast.

**Plucking**

Plucking (also called quarrying) includes both the fracturing of bedrock under a glacier and the subsequent entrainment (pulling away) of rock fragments at the bed. The rock fragments can be the result of glacial action or may have been isolated by pre-existing cracks and joints. In plucking, rock fragments first have to be...
GLACIAL EROSIONAL PROCESSES AND LANDFORMS

Meltwater landforms
Morainic landforms
Erosional bedrock landforms

Landforms that typically form over multiple glacial events. Suitable for reconstruction of average glacial conditions.

U-valleys**
Cirques**
Rock drumlin*
Roches moutonnées*
Mega lineations*
Drumlins*
Fluting
Ribbed moraines =>
Eskers =>
End moraines
Ice dammed lake features
Marginal meltwater channels

Point data that record “instant” events. Suitable for reconstruction of regional and short-lived glacial events, and for relative age control of ice flow events.

Figure 2 Temporal scales for the formation of glacial landforms, and the value of different landforms in reconstructing past glacial processes and events. Unpublished figure reproduced with permission from Clas Hättestrand, Stockholm University.

Separated from the parent material. As glacier ice overrides a bedrock surface, clasts embedded in the ice cause temporary stress concentrations in the bedrock that enlarge pre-existing cracks and joints leading to isolation of rock fragments. The frequency of existing joints and cracks, and the effective pressure of the ice-bed contact (overburden and basal water pressure) control the efficiency of plucking. Bedrock lithology also plays a role since different rocks have different strengths, resistances, and weathering histories. This fracturing of bedrock can result in rock fragments ranging in size from rock flour to boulders. Once rock fragments have been isolated, weakened, and/or broken apart from the parent material, the primary mechanism controlling the removal and incorporation of these into the basal ice is freezing.

The bed topography underneath a glacier is not an even surface but consists of highs and lows (bumps and steps). When the ice is close to the pressure melting point, meltwater will flow around bumps and refreeze on the lee side of a bump or step, forming a frozen regelation layer on the down-glacier side. This refreezing meltwater incorporates fractured rock fragments into the ice, and the movement of the ice down-glacier then removes these fragments.
Meltwater erosion

Glacial meltwater can be produced in two ways: (i) as surface melt, resulting from ice contact with atmosphere and precipitation at temperatures above the freezing point, or (ii) by melting of basal ice due to geothermal heat and heat from friction and internal deformation of ice. Meltwater at the bed of a glacier erodes underlying bedrock both mechanically, by evacuating rock fragments produced by abrasion and plucking, and chemically, by dissolution, carbonation, and cation exchange. Mechanical meltwater erosion is most efficient in subglacial meltwater channels. The meltwater carries clasts in suspension and as bedload and as these clasts interact they cause fracturing and failure to both the bed and the clasts. The hydraulic potential of channelized subglacial...
meltwater is high enough to erode large bedrock gorges underneath glaciers and ice sheets (Jansen et al. 2014).

Erosion rates

Rates of glacial erosion have been assessed using a range of different approaches (summarized in Hallet, Hunter, and Bogen 1996) including measurements of contemporary basal processes as well as measurements of watershed averaged erosion rates using both volumes of glacial deposits and characteristics that vary with erosion rate such as concentrations of cosmogenic radionuclides in sediments. These approaches yield estimates ranging between $10^{-4}$ and $10^1$ mm year$^{-1}$, reflecting differences in types and scales of glaciation and in the measurement methods.

Glacial erosional landforms

Small-scale erosional landforms

Erosional landforms resulting from abrasion range in size from a few millimeters to a few tens of meters and include striations, chatter marks, and grooves (Figure 3a). Their location and appearance reflect pressure conditions and the availability of clasts at the base of a glacier. Striations are elongated scratches consisting of numerous accurate fracture surfaces. Grooves are deeper tracks that might be straight, curved, or winding in plan form, while chatter marks are sets of fracture marks that record the removal of rock flakes by plucking. The orientation of these features reflects the dominant ice flow direction at their time of formation and thus these forms are used to reconstruct former ice flow directions.

Bedrock depressions that look as if they have been sculpted from plastic (Figure 3b, c, and d) were first labelled “P-forms” and attributed solely to glacial erosion; however, the term “S-forms” (sculpted forms) was added later to reflect the involvement of both glacial and fluvial processes in the development of some forms. P and S forms occur in a wide variety of sizes and shapes, and three main groups: transverse forms (perpendicular to ice flow), longitudinal forms (parallel to ice flow), and nondirectional forms (Benn and Evans 2010; Figure 3b). P and S forms result from erosion by: (i) debris-rich basal ice, (ii) saturated till flowing between the basal ice and bedrock surface, (iii) subglacial meltwater under high pressure, and (iv) ice-water mixtures (Munro-Stasiuk, Fisher, and Nitzsche 2005).

Intermediate-scale erosional landforms

Intermediate-scale erosional landforms reflect complex interactions between bedrock, topography, and ice flow patterns. Roches moutonnées are a classic glacial landform first described by the Swiss naturalist de Saussure at the end of the eighteenth century. They are asymmetric bedrock humps or hills with abraded stoss (up-glacier) faces and plucked lee (down-glacier) faces (Figure 3e) that range in size from a meter to several hundred meters across. Striations are common on the abraded stoss side, while polished surfaces usually occur on the flanks and gently sloping surfaces of the lee side, and plucking dominates along joints and cracks occur on the lee side. The formation of roches moutonnées is controlled by (i) the distribution of stress underneath a sliding glacier as it moves over bedrock humps and (ii) the pattern of rock failures. A high stress on the stoss sides of pre-existing bedrock humps results in large forces where clasts embedded into the basal ice contact the bed surface, and thus high rates of abrasion. On the lee side, lower stresses encourage the formation of cavities and enhance crack
formation in bedrock, while refreezing increases the effectiveness of plucking.

Whalebacks (symmetrical in plan view) and rock drumlins (asymmetrical in plan view; Figure 4a) are elongated and smoothed bedrock bumps that have been streamlined by glacial action but lack the plucked lee side of roches moutonnées. For whaleback and rock drumlin formation to occur there must be no regions of low pressure (no cavity formation) along the bedrock hump so that there is no plucking on the lee side. This can happen underneath thick and fast-flowing ice where normal stresses over the lee side are high enough to inhibit cavity formation. Whalebacks and drumlins appear frequently as swarms in areas where basal stress conditions were suitable for their formation, such as south of Lake Ontario where drumlins were formed by ice draining the Laurentide Ice Sheet at the end of the Pleistocene (Figure 3f).

Crag and tails are streamlined, erosional features, consisting of a resistant bedrock crag at the up-ice end, and a tail of less resistant rocks or sediment extending in the down-ice direction (Figure 4b). They are commonly produced by ice streamlining around an obstacle and protecting the “tail” of the feature from erosion. The classic crag and tail example is in Edinburgh, Scotland, where a castle is built on the crag, which is volcanic intrusive rock, and the tail consist of sedimentary rocks (Figure 4c).

Meltwater channels and tunnel valleys result from meltwater erosion underneath, on the sides, or in front of glaciers and ice sheets. Ranging in length from meters to several hundreds of kilometers (Figure 5), they can be eroded either in bedrock or in consolidated or unconsolidated sediment and can have various degrees of infill. Meltwater erosion under ice is a function of the water pressure gradient, and so is controlled by patterns of high to low pressure under the ice; in some cases this results in water flow that goes down pressure but “uphill” on the underlying bed topography. The regional distribution of meltwater features is dominated by deglacial phases of glaciers and ice sheets. Meltwater channels formed at or close to former ice margins are especially important since they can be used to infer past ice margin geometries and slopes. Well-developed meltwater channels are found in southern Patagonia (Figure 5a), the interior plateau of the Coastal Mountains in British Columbia, and northeastern Quebec-Labrador.

Tunnel valleys are large overdeepened meltwater channels, sometimes forming channel networks hundreds of kilometers long and several kilometers wide (Figure 5b). Tunnel valleys have undulating bed-long profiles, overdeepened valley floors, hanging tributary valleys, and after deglaciation are often filled by thick sedimentary deposits (Benn and Evans 2010). The large size of tunnel valleys may reflect (i) progressive excavation of sediment by normal meltwater discharges, in combination with subglacial sediment deformation, and/or (ii) excavation by extreme meltwater discharges associated with catastrophic drainage events. Well-developed tunnel valleys occur along the margins of the former northern hemisphere ice sheets (e.g., Figure 5b).

Large-scale erosional landforms

Glacial valleys are large-scale erosional landforms, typically with U-shaped cross-sections and long-profile overdeepenings, which result from erosion by valley glaciers in alpine settings (Figure 6a and b and Figure 7; Li et al. 2005). Glacial troughs are generally larger than glacial valleys, with smoothed and steeper valley sides, asymmetric cross-sections, and lacking Spurs. Fjords are glacial troughs in coastal areas with part of the trough below sea level (Figure 6c). The characteristic cross-section shape of glacial valleys, troughs, and fjords, with flat valley floors
Figure 4  (a) Rock drumlin shown in profile and in plan view, (b) crag and tail shown in profile, and (c) Edinburgh Castle crag and tail in Scotland where the crag is a volcanic intrusion and the tail consists of sedimentary rocks. Map data: Google, Infoterra Ltd, and Bluesky.
and steep valley walls, occurs because glacial erosion rates increase with basal sliding velocities, and so erosion is small at the edges of a glacier and increases rapidly towards the center (Harbor 1992). Glacial valleys develop on timescales of one or more glacial cycles and are best developed in areas that have been glaciated multiple times (Li et al. 2005). Glacial valleys, troughs, and fjords range from several hundreds of meters to hundreds of kilometers in length. Glacial valleys and troughs exist in all glaciated alpine regions of the world.

Overdeepenings are deep erosional basins in glaciated areas, and the formation, size, and shape of these basins are controlled by glaciological variables such as thermal regimes at the bed (warm-based versus cold-based ice), stress conditions at the glacier bed, and characteristics of the substratum such as bedrock structure and lithology (Benn and Evans 2010). Basins can range in size from small hollows between roche moutonnées to trough-sized larger overdeepenings. Large overdeepened valleys and basins exist in most alpine and foreland settings.

Cirques are amphitheater-shaped valley heads that are common in alpine settings (Figure 7). They have been cut into mountainsides by smaller ice patches or glaciers. The typical cirque profile has a flat floor/overdeepened basin connected to a steep back wall by a concave slope. Cirques evolve from hollows by retreat of the back wall and flattening of the floor leading

---

**Figure 5** (a) Meltwater channels formed along the margins of the former Magellan ice lobe in southern Patagonia, and (b) examples of tunnel valleys, the Finger Lakes, south of Lake Ontario in northwestern New York, USA, both as expressed in a digital elevation model.
Figure 6  Examples of glacial valleys and fjord landscapes as shown in digital elevation and slope models. (a) Slope model showing glacial valleys in the Tsambagarav Massif, eastern Mongolian Altai and (b) elevation model showing contemporary glaciation in the Tsambagarav Massif. (c) Fjord landscape in central Norway.

Figure 7  (a) U-shaped valley and cirques, and (b) cirques and a tarn, in the northern Swedish mountains.
Glacial erosional landscapes

The individual types of glacial erosional landforms described so far are organized into landscape-scale patterns of glacial erosion controlled primarily by the subglacial thermal conditions of a glacier system. The efficiency of a glacier system to erode depends primarily on basal sliding velocities. Warm-based glaciers have high sliding velocities and erosion rates because the ice-bed contact is at or below the pressure melting point. A cold-based glacier, on the other hand, has very low sliding velocities or is completely frozen to the bed, and has little or no erosion. Thus warm-based glaciers leave behind well-developed glacial landscapes, whereas cold-based glaciers in continental settings leave poorly developed glacial landforms. Underneath ice sheets and ice caps, variations in the bed topography cause differences in basal thermal regimes to occur across the bed. For example,
a patchwork of glacial and nonglacial landforms in northern Scandinavia has been used to infer complex patterns of warm-bed and cold-bed patches underneath the Fennoscandian Ice Sheet (Kleman, Stroeven, and Lundqvist 2008).

Ice flow patterns are linked to basal velocities and erosion rates, and thus also play a role in determining landscape types (Figure 1). There is little or no erosion under ice domes because of low sliding velocities, whereas under larger areas of ice sheets relatively uniform movement causes scouring over large areas (Figure 8). Areas of concentrated ice flow where ice converges (controlled by topography), as in ice streams or outlet glaciers, generally have high-velocity warm-based glacier ice and produce landscapes of selective linear erosion, which include glacial valleys, troughs, and fjord landscapes. Strong gradients in glacial erosion are created as the fast-flowing ice continues to carve the bed where ice flow is concentrated while rates of erosion in adjacent areas are much lower.

The effect of glacial erosion on the evolution of mountain landscapes has received a lot of attention because of interest in the links between tectonics, climate, and glacial activity (Spotila 2013). The observation that glaciers are capable of eroding more material than rivers and at rates that are faster than rock uplift has led to the idea that glacial erosion plays an important role in causing mountain uplift and can also limit the height of mountains.

**SEE ALSO**: Glacial depositional processes and landforms; Glaciers; Ice caps; Ice sheets; Quaternary geomorphology and landscapes

**References**


**Further reading**


Earth periodically experienced glacial conditions throughout the second half of its geological history. During a “glacial age” or “ice age” the mean global temperatures of the Earth’s surface and atmosphere are reduced for a significantly longer than usual time. As a result, existing continental and polar ice sheets and alpine glaciers extend, or new ones develop. The designation “ice age” requires the presence of extensive ice sheets in both the Northern and Southern Hemispheres. Hence, the Holocene or “Recent” Epoch (≈11,700 BP–present) is part of the present ice age, because ice sheets are covering the Arctic, Antarctica, and Greenland.

During an ice age, longer individual pulses of cooler and warmer climate alternate. A “glacial period” or “glacial” or “glaciation” is a stage that has cooler and drier climates over most of the Earth resulting in the formation of new or the growth of existing large land and sea ice masses extending outwards from the poles and of mountain glaciers often reaching into the forelands. Since significant amounts of ocean water were removed for building up the ice masses, sea levels are lower and ocean circulation patterns might have changed. Individual glacials within a specific ice age are separated by “interglacials,” which are periods of warmer and moister climates. The Holocene is such an interglacial of the present ice age.

Glacials are not uniformly cold and interglacials are not uniformly warm. In fact, glacials include “interstadials,” phases of warmer temperatures and minor ice retreat that are not long or intense enough to be considered interglacials, while interglacials include “stadials,” phases of lower temperatures and minor ice advance that are not long or intense enough to be considered glacials. The “Little Ice Age” (≈1250–1650) until the latter half of the nineteenth century or early in the twentieth century) was a stadial within the present Holocene interglacial.

Research results suggest that ice ages and glacials typically develop slowly, whereas they end more abruptly.

As explained above, while the term “glaciation” is a synonym for “glacial period” and “glacial,” it is commonly (and here) used to describe (i) the condition of being covered with glaciers or masses of ice, or (ii) the result of glacial action.

Research history

The documented history of research on glaciations starts in the eighteenth century; once the first ideas about former larger ice masses had been proposed, a chain of discoveries and theories evolved.
GLACIATIONS

1742 Daniel Tilas suggests that drifting sea ice brought erratic boulders to Scandinavia and the Baltic region.

1744 Pierre Martel, after he had visited the valley of Chamonix in the Alps of Savoy in 1742, reports in his publication that locals had explained to him the dispersal of erratic boulders with formerly larger glaciers that must have been receded.

1795 James Hutton describes that erratic boulders were transported from the Alps to the Jura Mountains by a great mass of ice.

1804 Erhard Wrede speculates that Earth’s axis had tilted, raising sea levels in the Baltic region and leading to floods that transported huge blocks of rock on ice floes to northern Germany.

1815 Jean-Pierre Perraudin proposes the idea of extensive glacier advances, and Johann Wolfgang von Goethe discusses similar explanations in his work.

1818 Goran Wahlenberg publishes his theory of a former regional glaciation of the Scandinavian peninsula.

1819 Leopold von Buch explains erratics as having been transported by violent floods or mudflows.

1821 Ignaz Venetz presents his idea of a former larger Alpine glaciation and a glaciation in the plain of northern Germany.

1824 Jens Esmarck argues that a sequence of global ice ages was caused by climate change, which itself was caused by Earth’s orbital changes.

1824 Jean de Charpentier supports Venetz’s idea of a former much-enlarged Alpine glaciation at the annual Helvetic Society meeting.

1832 Albrecht Bernhardi speculates that former polar ice caps reached as far as the temperate zone, and recognizes a former glaciation in Germany.

1833 Charles Lyell announces his “drift theory” that explained the existence of “diluvium” or “drift” deposits in Britain with a transport by icebergs.

1835/1836 Karl Schimper lectures about times of obliteration (German: “Verödungszeiten”) with cold climate and frozen water.

1837 Karl Schimper coins the term “ice age” (German: “Eiszeit”). Louis Agassiz announces the theory of a “great ice period.”

1839 Charles Lyell introduces the term “Pleistocene” (“Most New” or “Newest”) to describe strata in Sicily.

1840 Louis Agassiz publishes his Etudes sur les glaciers in which he describes a vast ice sheet that had extended from the North Pole to the Alps. He mentions neither de Charpentier nor Schimper, with whom he had developed the theory, in this book.

1841 Edward Hitchcock speculatively applies the glacial theory to “flood” features in North America.

1842 Roderick Murchinson points to the fact that “the greater numbers of practical geologists of Europe are opposed to the wide extension of a terrestrial glacial theory.”
Joseph Adhemar suggests that the primary instigator of ice ages might be Earth’s orbital path around the sun.

1846 Edward Forbes redefines the Pleistocene and refers to it as “Glacial Epoch.”

1847 Edward Collomb replaces the early mono-glacial hypothesis with a concept of two glaciations.

1853 Andrew Ramsay supports the concept of two glaciations.

1849/1950 Ernst von Bibra visits the Chilean Andes and learns that natives attributed fossil moraines to the former action of glaciers.

1864 James Croll presents his theory of the effects of variations of Earth’s orbit on climate cycles.

1870 Friedrich von Richthofen identifies loess as glacio-eolian sediment.

1875 James Croll publishes his *Climate and Time, in Their Geological Relations*, which led to international acceptance of the ice age theory.

1883 Rollin Chamberlin presents a “till” classification.

1906 Grove Gilbert publishes his major contribution on glacial erosion.

1909 Albrecht Penck and Eduard Bruckner present their Alpine glacial chronology in *Die Alpen im Eiszeitalter* [The Alps in the Ice Age].

1914 Ralph Tarr and Lawrence Martin recognize that glaciers surge.

1929 Motonori Matuyama discovers the polar reversals.

1930 Milutin Milankovich suggests three orbital causes for glaciations: eccentricity (100,000-year cycle), obliquity (41,000-year cycle), and precession (21,000-year cycle). These are later called “Milankovich cycles.”

1941 Milutin Milankovich suggests the effect of variable solar insolation on climate.

1947 Richard Flint publishes his *Glacial Geology and the Pleistocene Epoch*.

1957 J.K. Charlesworth publishes his *Quaternary Era*.

1964 W. Brian Harland argues for an ice age that was so extreme that it resulted in the deposition of marine glacial rocks in the tropics.

1969 The International Geological Correlation Programme – Project 38 on “Pre-Pleistocene Tillites” is initiated.

1970 Friedrich von Richthofen identifies loess as glacio-eolian sediment.

1987 V. Sibrava, D.Q. Bowen, and G.M. Richmond publish their *Quaternary Glaciations in the Northern Hemisphere*.

1992 Joseph Kirschvink proposes his “Snowball Earth” hypothesis that explains that Earth was nearly totally frozen more than 650 million years ago. His theory receives support from Paul Hoffman in 1998.

2004 Jürgen Ehlers and Philip Gibbard publish their *Quaternary Glaciations – Extent and Chronology*. Francis Macdonald and colleagues support the “Snowball Earth” theory by arguing that the Sturtian glaciation (720–660 Ma) during the Cryogenian (850–635 Ma) was global in extent.

GLACIATIONS

Evidence

Three main types of testimony help in the reconstruction of former glaciations: geological, chemical, and paleontological.

Geological evidence is given with numerous sediments, landforms, and landform features such as rock scouring and scratching, valley cutting, glacial moraines, drumlins, till or tillites, glacial erratics, loess, sand dunes, and glacial lakes. The two main problems in using these testimonies when reconstructing glaciations are their exact dating and that successive glaciations tend to distort and erase the geological evidence.

Chemical evidence exists in the form of variations in ratios of isotopes in fossils present in sediments and sedimentary rocks, particularly from the oceans. Furthermore, atmospheric gases trapped in bubbles of air in ice tell about the changes in the composition of the atmosphere through time. Hence, with the extraction and analysis of ocean sediment and ice cores from drilling projects, the understanding of glaciation chronologies radically improved.

Paleontological evidence is provided with the changes in the geographical distribution of fossils and with pollen data from lakes and bogs.

Causes

Several causes for the formation of glaciations have been identified: positions of the continents; supercontinental cycles; orogeny; and changes of ocean currents, astronomical conditions, and the atmospheric composition. While none of these causes alone can induce a glaciation, the occurrence of a multiple of them can.

The reconstruction of former patterns of Earth’s landmasses revealed that glaciations existed whenever specific configurations of the continents occurred that resulted in the reduction of energy exchange between lower and higher latitudes, and further to long-term variations in global climate at the scale of tens of millions of years. In the first case, a particular continent is located either directly on top of a pole such as Antarctica today or at high paleolatitudes. This repeatedly happened during the Paleozoic era (541–252 Ma) with glaciations ranging from Devonian (419–359 Ma) to Permo-Carboniferous (359–252 Ma) times, when glacial sediments were deposited across vast regions at high latitudes of southern Gondwana. However, paleomagnetic measurements of the very extensive glacial deposits from the earlier Neoproterozoic era (1.0 Ga–541 Ma) put several formerly glaciated areas at low latitudes. A possible explanation could be that Earth’s obliquity probably was much greater in these times, which allows for glaciated continents in low latitudes. A second scenario describes a configuration with an almost land-locked polar sea such as the Arctic Ocean today. In a third case, a supercontinent covers most of the equator such as Rodinia did during the Cryogenian period (850–635 Ma).

Earth’s history tells of the formation and disintegration of six supercontinents, an assembly of almost all of the cratons to form a single large landmass. The significant additions to the continental crust initiate a negative feedback mechanism. During periods of supercontinentiality, the continental lithosphere is elevated, which in combination with enhanced weathering on the now larger continents decreases the amount of carbon dioxide (CO₂) in the atmosphere. The reduced “greenhouse effect” results in significant cooling and, eventually, glaciation. In addition, changes in relative sea levels lead to supercycles of flooding and exposure of the continents, which produces either “greenhouse gas” or “icehouse” conditions at the scale of about 400 million years. However, at the same time the glaciation itself lowers weathering rates, which
in turn increases first CO₂ concentrations and then temperatures again and the ice masses melt.

During periods of extensive orogeny, global precipitation is increased resulting in a removal of CO₂ from the atmosphere and, eventually, cooling and glaciation. The formation of the Himalaya, for example, broadly matches the long-term decrease in mean global temperature in the last 40 million years.

The described tectonic activities result in constant modifications of ocean currents and long-term climate variations. For example, for most of geologic time, the Arctic Ocean was broad and open which allowed for an unabated circulation of the major ocean currents and, hence, a mixing of temperate and polar waters. As a result, climates were generally milder and more uniform reducing global ice cover. However, the Arctic Ocean was repeatedly partly enclosed by continents, which blocked the inflow of warmer waters leading to a cooling of the Arctic and glaciation.

While the described tectonic causes correlated to long-term (millions of years) climate changes, astronomical causes can affect global climate over shorter periods of between a few years to tens of thousands of years. The three “Milankovich cycles” summarize these effects: the eccentricity is a measure of the departure of Earth’s elliptic orbit from circularity and describes a cycle of about 100000 years. The obliquity is the variation between 22.1° and 24.4° of the Earth’s axial tilt against the plane of the Earth’s orbit and represents an approximately 41 000-year cycle. The precession is the trend in the direction of the Earth’s axis of rotation relative to the fixed stars, with a cycle of roughly 26 000 years. While all three cycles themselves and particularly in combination match periods of glaciation and deglaciation, some interpretative problems do exist: for example, while in theory the eccentricity should have the weakest impact on solar forcing, the strongest climate signal match the 100 000 year cycle, at least during the last one million years. Another problem is that the penultimate interglacial during marine isotopic stage 5 (MIS 5; 130–71 Ka) appears to have begun much earlier than hypothesized, that is, “out of cycle.” Milankovich also proposed the idea that the variable solar insolation affects climate. Indeed, short-term climate changes correlate to solar magnetic activity cycles (“solar cycles” or “sunspot cycles”) at the scale of about eleven years and to longer episodes such as the Maunder Minimum that describes the period from about 1645–1715 when sunspots became exceedingly rare; the latter coincided with the middle part of the Little Ice Age. Since the 1970s, the astronomical theory of paleoclimates is accepted as the major explanation for the occurrence of glacial and interglacial periods during the Late Cenozoic era.

Also, the variation of the atmospheric composition, particularly the concentration of greenhouse gases, can help explain climate changes. Records show that CO₂ and methane levels fell with the onset of glaciations and rose during deglaciations. However, it is not clear if these concentration variations are the cause or effect of climate changes. Without doubt, an increase in solar insolation increases the concentration of greenhouse gases. It is believed that the deforestation and farming practices during the Neolithic Revolution (about 10 000 years BCE) increased the greenhouse gas concentration and, hence, stopped following the periodic pattern of the Milankovich cycles with an expected onset of an already overdue glaciation.

Effects

Although glaciations are relatively rare and short phenomena in Earth’s geologic history, they had manifold effects on landscapes, sea levels, ocean
currents, weather patterns, tectonics, and even astronomical parameters.

While glacial landscapes were shaped in a relatively short time, they cover wide regions of the continents and include sediments, landforms, and landform features such as eskers, roche moutonnees, pluvial lakes, and fjords.

Besides this, the heavy ice masses also press parts of continental lithosphere into the asthenosphere with the result that with retreating ice sheets the landmasses rebound. Initially, this isostatic adjustment is a rapid elastic uplift as the ice is being unloaded, and, eventually, it is followed by slow viscous flow. The rebound can be as much as 1 cm per year near the center of the rebound such as in the cases of Scandinavia and the North Sea or the Great Lakes region, where it will continue for at least another 10,000 years.

During glaciations, the redistribution of ice and water and flow of mantle rocks might affect the gravitational field and the distribution of the moment of inertia of the Earth. As a result, changes in the angular velocity, axis, and wobble of the Earth are possible.

Glaciations also lower global sea levels, for example, during the Last Glacial Cold Stage (26,500–17,000 years) by up to 110 m. The retreat of the sea exposes continental shelves and might lead to the formation of land bridges such as the Bering Street between the North American and Asian continents. The availability of new land affects biogeographic patterns, for example, through changed animal migrations. At the end of the last glacial around 12,000 years ago, Mongoloid people from northeast Asia walked over the Bering Street into America.

The disintegration of ice sheets and the unloading during deglaciation can result in earthquakes near the ice margin, and the earthquakes themselves reinforce the ice breakup, which accelerates the calving of icebergs.

---

### Precambrian glaciations

While for much of Earth's history geologic records and datings of material are relatively rare, more data are available for the more recent past. In the last 2.5 Ga of the Earth's history, at least seven major ice ages occurred – four during the Precambrian Supereon (4.7 Ga–541 Ma) and three during the following Phanerozoic Eon (541 Ma–present) (see Table 1): Huronian, Sturtian–Varangian, Marinoan/Natuo, Gaskiers, Andean–Saharan, Karoo, and Quaternary. Since there is insecurity about its dating, it is not clear if the Kaigas can be identified as an individual (eighth) glaciation. (For explanations on glaciations of the following Phanerozoic Eon see Quaternary glaciations)

During the Hadean (4.7–4.0 Ga) and the Archean (4.0–2.5 Ga) eons, the total landmass is relatively small and atmospheric CO2 concentrations are high leading to much warmer global temperatures. Hence, evidence for glaciation in the stratigraphic record is nearly completely absent. However, smaller regional glaciations are documented for the Pongola Supergroup in Swaziland (Pongola Glaciation, 2.9 Ga), the Witwatersrand Succession in South Africa, and the Stillwater Complex of Montana, USA.

Several ice ages occur during the following Proterozoic Eon (2.5 Ga–541 Ma). Evidence for the Huronian ice age (2.4–2.1 Ga) during the Siderian period (2.5–2.3 Ga) and the Rhyacian period (2.3–2.05 Ga) of the Paleoproterozoic era (2.5–1.6 Ga) exist from the Huronian Supergroup of Ohio, Snowy Pass Supergroup of Wyoming, and other locations in North America, and from Australia, Finland, India, and South Africa. In contrast, from 2.0 Ga on until the end of the Paleoproterozoic era, no evidence for further glaciations is known.

The greatest proliferation of ice sheets ever happens during the Neoproterozoic
GLACIATIONS

**Table 1** Documented ice ages on Earth (Ga: billion years; Ma: million years).

<table>
<thead>
<tr>
<th>Ice Age</th>
<th>Age</th>
<th>Epoch</th>
<th>Period</th>
<th>Era</th>
<th>Eon</th>
<th>Supereon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pongola</td>
<td>2.9 Ga</td>
<td>—</td>
<td>—</td>
<td>Mesoarchean</td>
<td>Archean</td>
<td>Precambria</td>
</tr>
<tr>
<td>Huronian</td>
<td>2.4–2.1 Ga</td>
<td>—</td>
<td>Siderian/Rhyacien</td>
<td>Paleoproterozoic</td>
<td>Proterozoic</td>
<td></td>
</tr>
<tr>
<td>Kaigas (?)</td>
<td>750 Ma</td>
<td>—</td>
<td>—</td>
<td>Neoproterozoic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sturtian</td>
<td>720–660 Ma</td>
<td>—</td>
<td>—</td>
<td>Cryogenian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marinoan</td>
<td>650–635 Ma</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaskiers</td>
<td>582–580 Ma</td>
<td>—</td>
<td>—</td>
<td>Ediacaran</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andean-Saharan</td>
<td>460–420 Ma</td>
<td>—</td>
<td>—</td>
<td>Paleozoic</td>
<td>Phanerozoic</td>
<td>—</td>
</tr>
<tr>
<td>Karoo</td>
<td>359–260 Ma</td>
<td>—</td>
<td>—</td>
<td>Carboniferous/Permian</td>
<td>Cenozoic</td>
<td></td>
</tr>
<tr>
<td>Quaternary</td>
<td>2.6 Ma–Present</td>
<td>Pleistocene/Holocene</td>
<td>Quaternary</td>
<td>Cenozoic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

era (1.0 Ga–541 Ma), when widespread and long-ranging glaciations on all continents peak in the three ice ages Kaigas (750 Ma (?), Sturtian (720–660 Ma), and Marinoan (650–635 Ma) during the Cryogenian period (850–635 Ma; from Greek *cryos* “cold” and *genesis* “birth”), and in the Gaskiers ice age (582–580 Ma) of the Ediacaran period (635–541 Ma). Deposits are well documented from central and southwestern Africa, Australia, western North America, the North Atlantic Basin, and Brazil. One of the reasons for this “ice age era” is the greatly (>54°) increased obliquity of the Earth. However, it is not clear, if many glaciogenic rocks form in tropical latitudes, or if – during a second “Snowball Earth” – ice sheets reach from the poles down to the equator.

The future

Uncertainties exist about the future of the Holocene (11 700–present, “The Recent”), which is the current interstadial of the present Quaternary and follows the preceding Pleistocene. Depending on the chosen scenarios for the development of Earth’s climate and atmospheric greenhouse gas concentrations, predictions vary widely. The Holocene might last for another 50 000 years, if CO₂ levels increase to 750 parts per million (ppm), or it might end in 15 000 years, if CO₂ levels drop to 210 ppm. The reason for this uncertainty is that there is still an unsatisfying understanding of some individual input parameters of climate change scenarios, for example, the influence of the orbital forcing on Earth’s climate.

SEE ALSO: Climate change and land ice; Climate change and sea ice; Cryosphere studies: history; Glacial depositional processes and landforms; Glacial erosional processes and landforms; Glacier changes; Glacier inventory; Glacier mass balance; Glacier modeling; Glaciers; Global climate change; Global climate
GLACIATIONS

models; Holocene; Ice caps; Ice sheets; Ice shelves; Oceanic circulation; Oceans and climate; Paleoecology; Paleofloods; Polar climates; Pre-Quaternary landforms and landscapes; Quaternary glaciations; Quaternary science

Further reading


Glaciers are important fresh water resources releasing the water mainly in the summer months when the water demand for irrigation is usually highest. Moreover, glacier meltwater contributes significantly to global sea level rise. Receding glaciers increase the risk of hazardous mass movements such as slope failures and glacial lake outburst floods. Glaciers react sensitively to climate and are therefore recognized as important climatic indicators in remote regions where information from weather stations is rare or nonexistent. Hence, glaciers are naturally subject to changes in line with the Earth’s climate variability. Glaciers develop at locations where mass gain (e.g., through snowfall and avalanches) exceeds mass loss (e.g., through melting and calving). Favorable climatic conditions (e.g., lower temperature, higher snowfall) will lead to a positive mass balance and typically advancing glaciers while unfavorable conditions (higher temperatures, less snowfall) will lead to negative mass balance and retreating glaciers. The position of the glacier snout depends primarily on the glacier flow and the mass loss at the tongue. Higher temperature can cause melt and hence, an immediate retreat of the snout. Increased accumulation, however, results in a much delayed signal at the terminus as the snow needs time to transform into ice and to flow through the entire glacier. The time the glacier needs to adjust to a different climatic signal ("response time") varies between years to decades and is related to the thermal conditions, size, and topography. Hence, the same climatic forcing does not need to result in the same reaction of the terminus. Thick debris cover on the snout can further delay the reaction while calving into a lake and the sea will increase the mass loss. In addition, sudden advances of glacier tongues ("surges") cannot be linked directly to climate but more a sign of dynamic instabilities. Consequently, glacier length and area changes are in general difficult to interpret as direct climatic signals. Increase and decrease of glacier surface height can be more directly linked to the mass balance and, hence, to the climate forcing.

Late Pleistocene and Holocene glacier fluctuations

Glacier sedimentation and deposition forms such as moraines allow the reconstruction of former glacier extents. Further indicators can be sediments in proglacial lakes or trees and shrubs grown since deglaciation. Various dating techniques such as dendrochronological analysis, radio carbon dating of organic material within moraines, or dating of morainic boulders using terrestrial cosmogenic nuclides allow temporal classification. About one-third of the terrestrial land was ice covered during the last ice age. This Last Glacial Maximum (LGM) is usually dated approximately 24 000 to 18 000 years ago (24–18ka). However, maximum extents of mountain glaciation were reported from earlier periods, for example, approximately 35 ka for central Chile and approximately 29 ka.
GLACIER CHANGES

for northern Alaska (Clark et al. 2009). A global warming led, on average, to a strong reduction of the ice cover which started on average at about 19 ka with the earliest date of about 24 ka for the tropical glaciers until the early Holocene (starting at 12–10 ka). Glacier extents in most mountain regions were more likely of comparable size at the beginning of the Holocene than at the end of the twentieth century. Information regarding early Holocene glacier advances is sparse but there is evidence for several periods with relatively synchronous advances throughout the Holocene (Koch 2011). The Little Ice Age (LIA) moraines indicate the maximum glacier coverage of the Holocene. These moraines are usually clearly visible in the landscape and can be identified from high resolution satellite images. The LIA began around 1100 AD and lasted until the early twentieth century. The maximum glacier extents occurred in central Europe in the mid-nineteenth century but in other regions around the globe this date varies between the seventeenth to the late nineteenth centuries (Zemp et al. 2008).

Glacier variations since Little Ice Age

Field measurements of glacier length changes starting in the late nineteenth century and increasingly length change measurements based on satellite data allows a comparatively good overview of glacier changes around the globe for this period. Glaciers retreated on average worldwide since the maximum extent of the LIA until now. This can be well correlated with the global temperature increase since the middle of the nineteenth century. The overall recession is in the order of some kilometers for the larger glaciers and hundreds of meters for smaller ones (Zemp et al. 2008). Advancing or stable conditions were especially common around 1920 and 1980, with strong recession measured in the 1920s and 1940s and especially after the mid-1980s until today in most other parts of the world. However, local variability exists: for example, many glacier advances occurred in Scandinavia in the 1990s. Small glaciers show typically stronger fluctuations while large glaciers with a long response time have continued to retreat unabatedly since their maximum extent at the LIA (Zemp et al. 2008). Measurements of glacier area changes became numerous due to the availability of suitable remote sensing data for free or at low cost. Former glacier extents can be derived from topographic maps, aerial imagery, or declassified reconnaissance satellite images. Hence the earlier period of many area change assessments date usually only back to the 1950s, 1960s, or 1970s. Glacier area loss occurred in almost all regions of the globe in line with the length reduction (Barry 2006). Typical values vary between −0.2% per year for continental glaciers, for example, in central Tien Shan, and 0.7% per year for temperate glaciers in the Alps. However, values reporting absolute changes have to be interpreted with care as the uncertainty may be high. One major problem is different snow conditions which frequently occur especially around the higher parts of the glacier. Debris-covered glaciers, for example, common in many parts of the Himalaya show low or no identifiable area loss despite significant mass loss, because such glaciers tend to thin and waste vertically downwards. Surging or pulsating glaciers which show regular rapid advances (and area gain) followed by a strong retreat (and area loss) can be found in many mountains of the globe but are especially common in Svalbard in the Arctic, and for the Karakoram and Pamir in the western High Asian Mountains. Glacier-wide field-based mass balance measurements which show a more direct climate signal started around the 1950s with the highest observation density in central Europe and Scandinavia (Zemp et al. 2008). Only
few and short-term data series exist for South America and the Himalaya. Available data show on average a strong mass loss at the beginning of the measurements followed by a slowdown around the 1960s and a moderate ice loss until the mid-1980s. An increased mass loss was measured since then until now. The cumulative ice loss since the beginning of the mass balance measurements is more than 20 m water equivalent (w.e.) or about 0.34 m w.e. per year (Zemp et al. 2008).

Variation in the first decade of the twenty-first century and future glacier changes

The availability of more precise remote sensing data with global coverage since the beginning of the twenty-first century, such as from the gravity mission GRACE and the ICESat GLAS laser altimetry data and a detailed glacier inventory, allows now relatively precise estimates of glacier mass changes for the entire Earth. In addition, several satellite missions offer stereo data to derive digital elevation models (DEMs) which can be utilized to calculate changes in glacier volume by differencing of multitemporal DEMs. A consensus estimate for the period 2003–2009 revealed that the glaciers lost on average mass in all regions while the largest absolute ice losses occurred in the most glacierized regions, such as from the local glaciers in Greenland, Alaska, Arctic Canada, high-mountain Asia, and the southern Andes (Gardner et al. 2013). However, the specific mass loss (the value per area unit) was highest for the tropical glaciers, the glaciers in the European Alps, in the Caucasus, in Western Canada and the United States, and in the southern Andes, with values around −1.0 m w.e. per year. The lowest specific mass losses were measured for the glaciers and ice caps in the Antarctic outside the ice sheet, several arctic regions such as Svalbard and the Russian Arctic and high-mountain Asia with values varying between approximately −0.05 and approximately 0.3 m w.e. per year. The mass of all glaciers outside the two ice sheets reduced by 259 ± 28 Gt per year or 0.71 ± 0.08 mm of sea level equivalent per year which would account for 29 ± 13% of the measured sea level rise over this period. However, not all glacier meltwater will contribute to sea level rise because a small portion will be stored in inland drainage basins without outflow or in new lakes developed due to glacier recession. Projections of the future indicate on average further glacier wastage throughout the twenty-first century. The expected contribution to sea level rise until 2100 based on the global glacier inventory varies between 0.12 and 0.23 m on average depending on the applied method and the utilized climate scenario (Church et al. 2013). Highest specific mass losses are projected for the tropical glaciers, western Canada and the United States, Svalbard, Scandinavia, central Europe, and New Zealand, where likely less than 30% of the glacier volume from the year 2000 will remain. The availability of satellite information with higher spatial and temporal resolution combined with field-based studies will allow more precise assessments of the current glacier changes and will help to improve model projections for the future.

SEE ALSO: Climate change and land ice; Cryosphere: remote sensing; Glacier hydrology and runoff; Glacier inventory; Glacier lake outburst floods; Glacier mass balance; Glacier modeling; Glaciers; Ice caps

References

GLACIER CHANGES

Glacier hydrology broadly concerns the storage and transit of liquid water associated with snow and glacier ice. This includes water present at the surface of, in the interior of, at the basal interface of, and emerging from bodies of snow and ice. Today, glacier ice covers around 11% of Earth’s surface area, accounts for approximately 70% of the planet’s freshwater resources, and underpins the water security of over a billion people. For millennia, glacier-fed rivers have been critical for irrigation and agriculture, but have always posed a threat of flooding. Meltwater runoff from glacierized catchments is an important resource for hydroelectric power, accounting for 50% or more of the electricity produced in mountainous countries such as New Zealand, Peru, Pakistan, Switzerland, and Norway. Recently, focus has turned to sporadic, catastrophic, and destructive releases of meltwater from glacierized terrain, the mechanisms for and prediction of which remain poorly constrained. Therefore, the hydrology of glaciers is an important, topical subject demanding continued investigation, particularly in the context of the recent trends observed in glacier extents and the Earth’s atmosphere–climate system.

Glaciers represent a frozen environmental asset: natural reservoirs storing freshwater over a wide range of timescales, from hours to millennia. Long-term freshwater storage, at timescales greater than a year, typically involves glacial ice; intermediate and shorter-term storage more usually involves seasonal snow and liquid water and are best exemplified by the seasonality of glacier runoff. The hydrology and runoff from glaciers is fundamentally controlled by the surface energy balance (Hock 2005). For the vast majority of Earth’s glaciers, outside the tropics, mass is accumulated as snow during the winter, when the energy balance is at a minimum due to low air temperatures and reduced solar radiation receipt. During summer months, increased air temperatures and solar irradiance provide a highly positive surface energy budget, resulting in the melting of snow and glacier ice (termed ablation). Thus, glaciers modulate streamflow by releasing the most runoff during the warmest, driest periods in the year when all other sources of water are minimal. The volume of meltwater runoff that a glacier provides is a function of its surface area and ablation rate. The melt process ensures runoff from glacierized catchments correlates positively with air temperature and exhibits a strong diurnal signal in the ablation season hydrograph. Consequently, glaciers exert a strong influence over a catchment’s runoff, even if glacier ice only occupies a small proportion of the catchment area.

Within a glacierized catchment, the drainage system that routes meltwater from the glacier surface to the ice margin can be conceptualized as a cascade (Figure 1). Meltwater can travel along flowpaths at the surface (supraglacial), within the glacier’s interior (englacial), and at the ice–substrate interface (subglacial). The structure, function, and interaction of these systems determines the efficiency and transit time by which meltwater travels from the point...
GLACIER HYDROLOGY AND RUNOFF

Figure 1 Conceptual model of glacier hydrology. Note the drainage takes an arborescent pattern in cross-profile, which is also apparent in plan view with a dendritic structure leading to the main subglacial conduit(s) or channel(s) at the glacier snout. Water-flow is represented by blue arrows whose size infer discharge. Source: Irvine-Fynn et al. (2011). Reproduced with permission from Reviews of Geophysics.

of melting to the glacier terminus. Accordingly, runoff from a glacier is controlled by the characteristics and hydraulic properties of these three drainage system components, which modulate both meltwater volumes and water quality through solute acquisition and sediment evacuation. Hydrological coupling between the supra-, en-, and subglacial environments is fundamental in mediating the manner in which a glacier responds to environmental variability, primarily due to the role water plays in controlling ice motion. Here, focus is on mid-latitude, temperate valley glaciers, but for details on the hydrology of nontemperate glaciers and ice sheets, see Irvine-Fynn et al. (2011) and Chu (2013), respectively.

Supraglacial hydrology

The character of the supraglacial environment can be categorized into three key states, each with its own distinctive hydrology: snowpack, firn, and bare ice.

The snowpack

The hydrology of snow in the supraglacial environment is generally the same as for other terrestrial snowpacks (DeWalle and Rango 2008). Snowpacks on glaciers are the result of multiple snowfall events; in mid and high latitudes, these typically happen during winter.
months when little or no melt occurs, while for tropical glaciers, snow accumulates during warmer summer months. Consequently, glacier snowpacks are highly heterogeneous, exhibiting variations in crystal grain size and density from each snowfall event, leading to a complex, layered stratigraphy. After deposition, the snow structure is transformed by the pressure of overlying snow, by wind redistribution and compaction, and by internal processes of melting, refreezing, sublimation, and condensation. Such snowpack metamorphosis is driven by temperature gradients and snowpack age, and typically enlarges the crystal size or joins grains together (sintering), further influencing snowpack density. Combined, these processes result in local variations in snow density, ranging from 50 to 400 kg m\(^{-3}\), and lead to contrasts in snowpack porosity and hydraulic conductivity, not only with depth but also over space.

As the melt season commences, snow surface meltwater begins to percolate into and warm the snowpack. The infiltration of liquid water and the latent heat released by refreezing begins to overcome the cold content of the snow, raising the snowpack temperature toward the melting point. Densification of the snowpack occurs as diurnal melt–freeze cycles cause the refreezing of percolating meltwater either within the pore spaces of the snow or as more massive ice lenses. Continued melt ripens the snowpack as gravity-driven percolation causes the wetting front to descend away from the surface into colder snow below. The percolation rate of meltwater can be approximated by Darcy’s Law, with velocities of \(10^{-4}–10^{-5}\) m s\(^{-1}\). Critically, due to the structure and density variations within the snowpack, the wetting front is neither uniform nor horizontal, and preferential flowpaths or flow fingers develop. The velocity of meltwater through a ripe snowpack is dependent on the physical structure of the snow and the water flux: greater volumes of melt travel more rapidly. As the snowpack becomes isothermal (at the melting point) the increased water content raises its homogeneity and hydraulic permeability. Meltwater content and flow will be highest (and most concentrated) where melt rates are greatest. Drainage through a saturated snowpack is driven by the local slope, and transport velocities can increase to \(10^{-2}\) m s\(^{-1}\).

Over the accumulation area of a glacier, at the peak of the summer melt cycle, the varied physical and thermal properties of the snowpack may be classified as three distinct snow facies. The first is dry snow, which is typically only found at very high elevations (or in the interior of ice sheets), where no melting occurs, the water content remains at 0%, and the snow makes no contribution to runoff. Second is; percolation snow, which occurs at intermediate elevations, where some summer melting occurs, meltwater (about 3% water content by volume) penetrates the snowpack or drains into the firn (see next section), but refreezing and storage occurs, ensuring melt does not contribute significantly to runoff. Third comes wet snow, which arises where summer seasonal melt raises the snowpack to 0°C and vertical percolation results in saturated snow (8–15% water content) and lateral flow occurs, particularly at the glacier ice interface, or where ice lenses or layers occur. In contrast to cold or polythermal glaciers, the summer melt season warms the entire snowpack on temperate glaciers and the dry and percolation snow facies are not observed.

At lower elevations, in the ablation zone, the increasingly thin and saturated snowpack may become slush (>15% water content). Slush flows themselves may contribute to the removal of snow from the glacier surface. These slush zones commonly follow the seasonal retreat of the transient snowline, and in this zone supraglacial lakes
GLACIER HYDROLOGY AND RUNOFF

may form where glacier surface gradients are low or topographic depressions exist. Undulations in glacier surfaces can be formed by ice flow itself and by the topography of the glacier bed. Meltwater may also pond in crevasses during the early part of the melt season. Lake volumes of $10^4$ m$^3$ have been described on valley glaciers, while on the Greenland ice sheet, supraglacial meltwater lake volumes of $10^7$ m$^3$, with extents of up to about 9 km$^2$, have been reported. Supraglacial lakes are transient features usually forming early in melt season as supraglacial drainage is initiated, when saturated snow- or slush-plugs prevent flow through supraglacial stream courses. However, on ice sheets, lakes may persist between melt seasons.

At the end of the summer, the snowline position broadly corresponds to the equilibrium line altitude (ELA), which is the average elevation contour at which accumulation equals ablation over a one-year period. At the ELA, slush and supraglacial lakes may persist during summer, but superimposed ice is also evident. Superimposed ice forms where meltwater refreezes at the glacier ice interface, most typically in early summer as meltwater percolates through the snowpack, but also in early autumn when residual melt draining from the accumulation area and precipitation refreeze as air temperature decreases. Although superimposed ice can form across the entire ablation zone, the high ablation rate at lower elevations removes evidence of its formation. Superimposed ice can reduce the permeability of surface glacier ice, so that accumulation of meltwater on the glacier surface is pronounced early in the ablation season.

The firn

Firn represents the transition between snow and glacier ice, and is formed through the processes of firnification. This process represents the compaction and reorganization of snow and ice crystalline structure, which in turn increases its density to >400 kg m$^{-3}$. Snow that survives over a full annual cycle becomes firn. The rate at which this snow transforms into firn (and ultimately into glacier ice) is dependent on the accumulation rate of snow, the temperature of the snow, and the presence of meltwater in the evolving snowpack. These controls impose a gradual reduction in the air or void space within the snowpack, and firn becomes glacier ice upon the occlusion of intercrystalline air passages to bubbles, at a density of about 830 kg m$^{-3}$. The density of firn increases with vertical distance from the snow surface and also with proximity to the ELA, where production and refreezing of meltwater is commonplace. These spatial trends toward an increasingly tight packed crystal structure result in a porosity gradient. Characteristically, firn is found above the ELA and its hydrology is important for runoff sourced in a glacier’s accumulation area.

Due to the presence of firn, the accumulation area of a glacier represents a confined aquifer. Meltwater derived from the surface snow can percolate through the snowpack and into the firn layers. This meltwater drains through the higher porosity, unsaturated firn to form a saturated firn layer at depth, where the reduced permeability impedes further drainage. This saturated firn layer progressively develops following the onset of the melt season; the volume of water retained in this aquifer is dependent on, among other things, the snow melt rate.

The saturated firn layer on valley glaciers has been reported at depths of up to 15 m below the firn surface, with thicknesses ranging from 1 to 7 m above the firn–ice transition. The meltwater transit velocities of $10^{-4}$–$10^{-5}$ m s$^{-1}$ in the saturated layer emphasize the firn’s ability to delay snowmelt runoff at timescales of days to weeks. The slowed transport of meltwater
in accumulation area firn reservoirs results in the damping of any diurnal melt signal during summer. However, as the melt season ends, the saturated firn drains down-glacier or meltwater refreezes in situ (aiding firmification), at least partially reducing the firn aquifer’s liquid water content.

The bare ice

During the melt season, the decrease in snowpack extent, and up-glacier recession of the transient snowline toward the ELA, results in the exposure of glacier ice and initiates rapid supraglacial runoff from the ablation area. If unimpeded, this meltwater flows across bare surface ice at much higher velocities (typically $10^{-1}$ m s$^{-1}$) than those seen in the snowpack or firn.

On all glaciers during the melt season, the ablation area represents a transient thermal layer: in winter, the glacier ice surface is cooled below the pressure melting point (PMP), while during spring and summer the increasingly positive energy balance driving ablation ensures the seasonally cold surface layer initially approaches and then remains at the PMP. Because glacier ice is not a pure medium, the energy balance that drives melt is not uniform over space. Preferential radiation absorption at crystal boundaries or the presence of dust and/or impurities on the ice surface generates micro- (micrometer) to small- (centimeter) scale topographic variations. Meltwater may percolate through or along these flowpaths and, by increasing flow volumes, create rills on the ice surface.

Incident solar radiation also penetrates the uppermost glacier ice, typically to a depth of about 2 m. Through melting at subsurface crystal boundaries, the surface ice density can be reduced to about 550 kg m$^{-3}$ creating a porous weathering crust through which water can be transported. Observations have suggested this shallow weathering crust zone behaves like a perched aquifer, with the capacity to store and release meltwater; however, this weathering crust aquifer is typically saturated, with its water table just a few centimeters below the ice surface. Local differences in ice structure (e.g., foliation) result in contrasts in hydraulic permeability in the ice, which may contribute to the formation of preferential flowpaths within the weathering crust ice. Transport rates in the weathering crust are typically between $10^{-4}$ and $10^{-6}$ m s$^{-1}$.

By following the steepest surface gradient, surface rills coalesce to form supraglacial stream channels. Akin to terrestrial river catchments, supraglacial stream channels commonly develop dendritic or arborescent drainage patterns, with drainage densities ranging from $<10$ to $>20$ km km$^{-2}$, which generally decrease up-glacier where ablation is reduced. Due to the convexity of glacier ablation areas, elongated, subparallel drainage patterns are often observed. Similarly, the structural control imposed by down-glacier longitudinal foliation and annealed crevasse traces can result in rectilinear surface drainage systems. Delays of between 1 and 11 hours have been observed between peak melt production and peak supraglacial stream discharge; these delays relate to catchment size and geometry, but also the hydraulic properties of the weathering crust. The form of supraglacial flowpaths and stream catchment areas varies glacier to glacier: drainage networks may be disrupted by crevasses (see next section) occurring in response to site-specific glacier motion and the associated stresses in the ice.

The geometry of supraglacial streams is controlled by, among other things, stream catchment area and ablation rate. To accommodate increasing meltwater discharge, supraglacial streamflow velocity and water depth increases while channel width remains relatively constant.
GLACIER HYDROLOGY AND RUNOFF

Here, flow velocities of $>1\text{ m s}^{-1}$ are common, even in small streams. Channel roughness is very low; the dimensionless Manning’s “$n$” for supraglacial channels is typically $<0.03$. As is seen in terrestrial river systems, both meandering and step-pool morphologies can develop in supraglacial streams. Meander wavelength relates to discharge and stream width, while the sinuosity of meanders (typically $1.0–2.5$) is inversely associated with reach-length channel slope. Step-pool sequences appear to develop where channel slope exceeds about $15\degree$.

The incision rate of supraglacial streams is principally related to channel slope and discharge. Viscous fluid friction within the turbulent flow can accentuate channel incision. Supraglacial stream water equilibrates to around $+0.1\degree\text{C}$, meaning there is also sufficient thermal energy available to induce downcutting of ice-walled channels at rates of about $2\text{ mm h}^{-1}$. On glaciers in temperate latitudes, typically the ice surface ablation broadly matches stream incision rates (about $0.1\text{ m day}^{-1}$), and streams reform each melt season. Conversely, in colder climates, stream incision may be an order of magnitude greater than glacier ablation, resulting in the development of progressively more deeply incised “canyons” $>2\text{ m}$ deep and persistent over many years. The process of meandering will continue at depth and the planform surface expression of a supraglacial stream may not reflect its active channel.

As a result of the low permeability of surface and near-surface ice, the ablation area of a glacier exhibits a rapid hydrological response to changing melt rates. Here, rapid transport of meltwater via supraglacial streams ensures that the diurnal melt signal is apparent in the bulk meltwater discharge. However, the configuration of drainage structure in the supraglacial environment is, in part, dependent on the occurrence of features that provide hydrological access to the englacial environment.

**Englacial hydrology**

Massive glacier ice is permeable, allowing englacial water to permeate through it slowly. With a density of $830–910\text{ kg m}^{-3}$, at the microscale, glacier ice crystals are surrounded by a film of liquid water; these films form a network of veins at the junction between three or more individual ice crystals. Typically $0.1–10\mu\text{m}$ in diameter, the size of the veins depends on the ice crystal size, meltwater solute content, pressure, and temperature. Ice crystals commonly range from $<1\text{ mm}$ to $\approx10\text{ cm}$ in diameter and the size of the vein junction correlates positively with crystal size. The process of ice crystal formation itself rejects solutes and impurities, resulting in solute-rich (or saline) water within the intercrystalline or interstitial vein network. The lower freezing point of saline water compared to that of pure water allows for the existence of larger veins. Increases in overburden pressure reduce the PMP of ice, potentially enlarging the vein network in ice nearer the glacier bed; the hydraulic permeability of the ice in the lowermost several tens of meters in a glacier increases. Englacial strain heating resulting from the flow of the ice can also contribute to interstitial meltwater production ($\approx0.01\%$ by volume). However, irrespective of crystal size and solute content, as ice temperatures decrease, the vein network contracts; the porosity and permeability of ice at the PMP is considerably higher than for cold ice. Liquid water can be present in ice substantially below the PMP, but usually $<0.1\%$ by volume, while the water content of glacier ice at the PMP may rise to several percent by volume. The flow in the vein network is small, however, with laboratory studies suggesting velocities of between $10^{-8}$ and $10^{-13}\text{ m s}^{-1}$. The presence of air bubbles and debris, capillary forces, the deformation and recrystallization of ice, and
high confining pressures, are all likely to impede this flow.

Despite the low rate of water transport, the interstitial vein network is prevalent throughout a glacier and water will generally flow toward the glacier bed. Assuming ice is a porous medium, water moves from high to low hydraulic potential (Φ). Hydraulic potential is the sum of the gravity–defined potential energy and the water pressure defined by the force imposed by the overlying ice. Theoretically, water flows perpendicular to the 3-D plane with equal hydraulic potential – the equipotential surface – and follows the steepest potential gradient. On valley glaciers, equipotentials typically dip up-glacier at 11 times the ice surface slope. Using calculations based on glacier geometry alone (e.g., ice extent, topography, and thickness) it is possible to calculate an informative overview of likely flow-paths in the englacial environment.

Water-flow, directed by the hydraulic potential, generates heat proportional to the water flux, further enlarging already large veins in the ice. As such a vein enlarges, its water pressure drops compared to smaller veins. Water-flow will be directed toward the larger evolving conduit and, consequently, large conduits develop and grow at the expense of smaller flow-paths. This results in the development of an arborescent drainage pattern following the gradient of equipotentials through the englacial environment. However, studies of flow-paths in the englacial environment from direct observations (glacier speleology), ice-penetrating radar, and video imagery from boreholes do not appear to fully support the theoretical englacial drainage system as driven by hydraulic potential. Englacial conduits following subhorizontal pathways, vertical descent depths of up to 60 m interrupted by small steps and horizontal galleries, or small (<0.1 m) englacial drainage structures with shallow dipping orientations have contributed to a growing realization that structural controls (foliations, fractures, variations in ice rheology, etc.) may provide a greater influence on the geometry of drainage features, and local hydraulic potential fields, than was previously appreciated.

Crucially, in view of the limited water transfer capable by the primary permeability of englacial ice itself, it is the larger-scale structures (e.g., crevasses, fractures, and moulins) that provide the pathways for the bulk of seasonal meltwater to access the englacial environment (Figure 2), even if hydraulic potential governs the general direction of water-flow. These macroscale flowpaths that represent secondary permeability may allow transit velocities of between 10^{-1} and 1 m s^{-1}. Most commonly, crevasses exist as direct pathways to the glacier interior. Such crevasses form in response to high tensile strain rates and have depths of up to about 30 m in temperate ice (ice at the PMP), and greater in colder, more brittle ice or where they can be kept open by water flow (see section below). Crevasses truncate any supraglacial streams they intersect and sequester water from them.

---

*Figure 2 (opposite) Illustration of hypothetical evolution of englacial conduits from crevasses and meltwater streams. (a) Classical moulin formation from truncation of a supraglacial stream; (b) crevasse hydrofracture promotes englacial drainage path; (c) crevasse-bottom incision and glacier motion isolate an active englacial channel; (d) cut-and-closure formation of englacial channel, where meandering at depth ensures englacial conduit geometry does not reflect original supraglacial stream path. Water-flow is represented by blue arrows, and forces acting on ice shown are by red arrows. Ice flow is left to right for (a) to (c) and from back to front in (d). Source: Irvine-Fynn et al. (2011). Reproduced with permission from Reviews of Geophysics.*
GLACIER HYDROLOGY AND RUNOFF

(a)

(b)

(c)

(d)
Energy dissipation by flowing water at the crevasse base may develop a vertically descending conduit at depth. The stress field within the ice causes a lateral deviation of such evolving conduits, with observations suggesting a down-glacier orientation of about 25° from vertical. Where ice flow causes crevasses to close, but supraglacial drainage continues to incise a notch into the up-glacier edge of the former crevasse, a vertical shaft (moulin) forms. Moulins from <1 m to 10 m in diameter may be long-lived features, persisting for multiple melt seasons.

Where crevasses or moulins become water-filled, the process of hydrofracture may arise. Here, the mass of meltwater promotes fracture extension at the base of the crevasse or moulin. The depth of the hydrofracture depends on the volume of stored water and/or the persistence of meltwater delivery to the fracture locus. In the absence of a continuous water flux to the fracture base, refreezing will reduce the likelihood of the fracture descending to greater depths. Under optimal circumstances, crevasse hydrofracture has been observed in glaciers and ice sheets to propagate to depths of $10^2$–$10^3$ m in short (hours, days) timescales. As with crevasse and moulin drainage path formation, the changes in effective stress field with increasing depth can result in the hydrofracture flowpath becoming inclined at progressively shallower angles. Hydrofracture is also not limited to a vertical orientation and appropriate water pressures or structural weaknesses can facilitate shallowly dipping drainage path formation.

It is possible for supraglacial streams themselves to develop into englacial conduits, especially where stream incision exceeds general ice ablation rates: snow bridges and metamorphosed and deforming ice can cover these incised flowpaths, resulting in enclosed, high-roofed englacial channels called cut-and-closure channels. The geometry of these channels resemble supraglacial streams, with flow generally being at atmospheric pressure. Processes of vertical incision and knick-point retreat rates of 0.1–0.5 m day$^{-1}$ in these channels can lead to the formation of moulin-like forms at the channel head. Similarly, subhorizontal channel forms have been suggested to arise from the entombment of conduits formed at the base of the saturated snow zone in the accumulation area of ice masses. Crevasses may also lead to subhorizontal drainage structures: water-flow in crevasse bottoms may initiate a drainage route, and subsequent crevasse closure can isolate the drainage structure from the supraglacial environment. Speleological investigations have also revealed that ice structures – such as refrozen or annealed crevasse and/or fracture traces, debris bands, thrust faults, and highly permeable ice – all provide locations of high hydraulic transmissivity, supporting conduit inception. Studies have reported that englacial fractures can be found throughout 96% of a glacier’s depth. Fractures in ice can also result from frictional drag at the glacier bed, with basal crevasses propagating vertically upwards into the englacial environment.

Current wisdom suggests englacial drainage flowpaths close up over winter months and reopen each summer, when water pressures decrease with channel enlargement. The closure rate of water-free ice-walled channels is a function of ice overburden pressure and viscosity, largely controlled by temperature. In cold ice, even small channels can persist for over a year. In contrast, temperate ice deforms rapidly and, once empty of water, conduits close rapidly: theoretically, at a depth of 100 m, a conduit’s diameter may halve in <20 days. However, englacial conduit closure rates are substantially reduced if they are water-filled, as the stored
GLACIER HYDROLOGY AND RUNOFF

water impedes ice deformation. The englacial system, therefore, can represent a reservoir with capacity to store considerable volumes of water, particularly over winter months. Empirical water balance studies have demonstrated disparities between input and output meltwater volumes that suggest such processes of storage and release occur within the englacial hydrological system.

Subglacial hydrology

Assuming the glacier bed is at the PMP, water reaching the ice–bed interface flows toward the glacier terminus, orthogonal to the equipotential contours formed by the intersection of the equipotential surfaces with the bed. Water can exist at the ice–bed interface beneath cold ice areas due to the hypersalinity of mineralized waters resulting from solute rejection during melting and refreezing. Meltwater at the glacier bed may also be sourced from basal and frictional heating in areas of temperate ice, although annually this rarely exceeds about $10^{-1}$ m of melt. Once present at the ice–bed interface, though, meltwater flow is complicated by variability in both supply and bed conditions.

Where ice is thin, channel enlargement by viscous heat dissipation may offset channel closure, and such channels would be open with meltwater flowing at atmospheric pressure. In this situation, bed geometry is the primary factor in defining the orientation of subglacial channels. However, observed channel locations have often appeared to mirror those reconstructed from hydraulic potential assuming water at ice-overburden pressures. To reconcile this dichotomy, it is quite possible that drainage pathways close by deformation in winter months and subsequently reopen each summer. Since conduits are water-filled, and therefore at (or temporally above) ice-overburden pressure during re-formation in the spring, their initial location is controlled by hydraulic equipotential pressures. Subsequently, and despite reduced water pressures, melt rate is insufficient to allow lateral migration to the location dictated by theoretical open-channel flow (i.e., the low point of the glacier bed). Nonetheless, there is a broad agreement that water-flow at the bed occurs in one or both of two qualitatively and hydraulically different conceptual flow systems: distributed or “slow” flow and discrete or “fast” flow.

Distributed (slow) subglacial drainage

Distributed subglacial drainage occurs through spatially extensive, nonarborescent (or anastomosing) flow pathways at relatively slow velocities of $<10^{-2}–10^{-3} \text{ m s}^{-1}$. For a glacier with a hard bedrock sole at the PMP, microscale pressure-driven melting and refreezing (regelation) occurs around bedrock protuberances up to about one meter in length. The meltwater generated by this process flows around bedrock bumps and clasts as a millimeter-thick film. Sedimentological analyses reasoned that the lack of particles $<0.2 \text{ mm}$ at the rock–ice interface demonstrated the presence of water films capable of eroding finer particles, while numerical analyses indicated an upper limit to film stability of about 4 mm. In reality, small-scale bed roughness probably reduces stable film thickness to substantially less than this. Raised viscous heat dissipation in thicker films, or local increases in water discharge due to supply or bedrock variability, result in the development of preferential drainage paths and protochannels. Where ice flow is sufficiently fast, ice may separate from the bed in the lee of a bedrock protuberance, leading to the development of cavities. Subglacial meltwater may collect in, and thermally enlarge, cavities. The initial cavity size is dependent on
bed topography, ice velocity, and basal shear stresses. Cavities may be either autonomous or interconnected by narrow channels (orifices) where water-flow is sufficient. A network of interlinked cavities forms an anabranching linked cavity system. For small discharges, ice melt is unimportant for maintenance of the cavity system, so small increases in water pressure lead to a greater carrying capacity of the system. However, if meltwater discharge through the inefficient linked cavity system increases further, preferential channel-like flowpaths can develop.

Many glaciers are not entirely in direct contact with consolidated bedrock. Instead, subglacial investigations and observations of recently deglaciated surfaces commonly indicate the presence of unconsolidated sedimentary deposits. With a hydraulic conductivity typically in the range $10^{-6}–10^{-12}$ m s$^{-1}$, Darcian porewater flow is likely through subglacial sediments, although extremely difficult to measure directly. However, macroporous or preferential flowpaths may develop in subglacial till, and porewater flow is likely to occur in conjunction with other more efficient drainage.

**Discrete (fast) subglacial drainage**

Discrete subglacial drainage occurs through channels or conduits, typically forming a stable arborescent network along which meltwater flows rapidly at velocities of $10^{-1}–10^{-2}$ m s$^{-1}$. The relative instability of distributed drainage systems to increased water flux means that they break down as water discharge accumulates either gradually down-flow or locally due to point inputs. Here, preferential drainage paths develop by melting the overlying ice to form semicircular subglacial channels, Röthlisberger- or R-channels, which follow the hydraulic gradient. However, water pressures measured in subglacial channels have been found to be higher than those calculated for ideal R-channels, indicating unusually rough or sinuous conduits, or a cross-sectional shape that is broad and low rather than semicircular. In these Hooke- or H-channels, melting is concentrated on the conduit walls, since channels are not continually water-filled and lateral closure is limited by friction with the bedrock. Where bedrock may be readily eroded or where major subglacial channels persist for sufficient time, such channels may incise into the rock rather than the ice, forming Nye- or N-channels.

In the case of a poorly consolidated sediment bed, due to the instability of films and porewater flow to externally driven variations in water flux, it is thought that shallow anabranching canals incise into the deformable bed material. Canals are enlarged by water-flow removing sediments, with sediment creep and failure of the canal wall counteracting the tendency for growth.

Given the spatially restricted nature of discrete subglacial drainage and the spatially extensive nature of distributed subglacial drainage, it is anticipated that the two systems likely coexist side by side. Discrete channels are likely to form at point sources, such as beneath moulin delivery points, and to extend down-flow of them, irrespective of their location. Channels will also generally increase in representation down-glacier, where accumulated subglacial meltwater flux is largest. In between these channels, distributed drainage will dominate. Indeed, at temperate glacier beds at the PMP, the distributed component will be ubiquitous between discrete channel pathways. Critically, evidence suggests not only that such a combination of subglacial drainage systems do coexist, but that they interact at a variety of scales. For example, coupling between porewater flow and channelized flow has been observed whereby high daytime water pressures within channels drives water from them into the surrounding distributed...
system. At night, when the channel water pressures are markedly lower, that water returns, transferring solute and suspended sediment from the surrounding locality to the channel.

**Temporal evolution of glacier hydrology and runoff**

Conceptually, spatial and temporal changes in temperate glacier hydrology are thought to involve, on an annual cycle, a glacial drainage system that closes over in winter and reopens during summer months.

The relatively rapid rate of ice creep in temperate glaciers, with ice at the PMP, makes it unlikely that englacial and/or subglacial conduits or fast drainage structures remain open once devoid of liquid phase meltwater during winter. However, once closed, meltwater can become trapped through the winter in cavities at the glacier bed. At the commencement of the melt season, the volume of water stored within the glacier initially increases due to poorly interconnected drainage structures (e.g., snowpack, isolated crevasses, and flooded moulins) and the presence of a slow drainage system at the glacier bed. As surface warming continues through the spring and early summer, the surface snowpack becomes isothermal, and snowmelt floods (spring events) occur, which flush a large volume of meltwater through the glacier drainage system, resulting in an early melt season peak in runoff volumes (and commensurate ice velocity). This spring event may also release previously stored meltwater. As the summer progresses, ablation is heightened across the glacier surface and increasing volumes of water are routed into the glacier interior. This increases water pressures and destabilizes slow drainage structures such that a channelized, fast flow network develops at the expense of distributed drainage (Figures 1 and 3).

The hydraulically efficient system then probably persists for the remainder of the melt season with sufficient water flux to maintain fast drainage structures. Once ablation and meltwater volumes decline, the discrete, fast flow drainage structures begin to close through ice deformation.

Importantly, these temporal changes do not occur uniformly across the entire glacier bed but have a systematic spatial expression that is driven by the up-glacier expansion of the area of intense surface melting. This area corresponds to the expanding area of bare ice exposed by the melting of the (hydraulically buffered) supraglacial snowpack. The retreat of the snowline toward the ELA progressively exposes moulins and crevasses that become hydraulically active, delivering melt to englacial or subglacial drainage flowpaths. Consequently, water pressures within and at the bed of a glacier are not spatially uniform, and the growth of a fast, discrete drainage system evolving from a slow system follows the retreat of the snowline up-glacier (Figure 3b). The growth in extent of an efficient drainage system is also likely to reduce storage within the glacier, with more direct coupling between melt processes and runoff volumes following the reduction in both the snow/firn aquifer and the impedance of subglacial flow. Meltwater discharge as the snowline retreats to higher elevations has been demonstrated to result in an increase in bulk meltwater runoff volumes as well as a reduction in the time lag between peak temperatures and peak runoff. As the season progresses, the diurnal runoff signal becomes increasingly peaked due to the loss of hydrological characteristics capable of damping the transport of meltwaters. Later in the melt season, with falling rates of ablation, the reduced supply of meltwater is insufficient to maintain the fast drainage structures, which begin to close and decrease in extent as they transition back
Figure 3  Schematic illustration of the spatial and temporal variation in meltwater travel to the proglacial environment. (a) Systematic change in tracer transit over the melt season, showing decreasing dispersion and increasing velocity; (b) the seasonal, up-glacier extension of fast meltwater transit times mirroring the retreat of the snowline; (c) the associated seasonal runoff hydrograph. The temporal positions of idealized tracer breakthrough curves (A–D) are shown. Source: Richards et al. (1996) and Irvine-Fynn et al. (2011). Reproduced with permission from Reviews of Geophysics.
GLACIER HYDROLOGY AND RUNOFF

to, and integrate with, the distributed flow structures.

The temporal and spatial progression of a temperate glacier’s hydrological system imparts systematic changes in the runoff hydrograph. Runoff may flow from a glacier throughout the year, with a base flow associated with slowly draining water, volumes of which only change at longer (most typically seasonal to annual) timescales. From the onset of melt, and following any snowmelt flood events, a glacier’s runoff hydrograph typically exhibits rising discharge volumes and increases in the amplitude of diurnal discharge variations as the drainage system progressively becomes more evolved and efficient (Figure 3c). The time lag between peak melt and peak discharge at the glacier terminus is progressively reduced. If melt season snowfall occurs or synoptic conditions reduce melt rates, discharges (and diurnal variations thereof) are commonly diminished. As the melt season ends, the hydrograph typically shows a rapid decline in runoff volumes.

Quality of glacial runoff

For most glaciers, water emerges from the glacier terminus in a few discrete proglacial streams associated with the dominant fast subglacial flowpaths. The water quality in these streams is dependent on the flowpaths meltwater has followed; a glacier’s hydrological structure is critical to the acquisition of solutes (ions dissolved in water) and entrainment of suspended sediment by waters emerging at the ice margin.

Solute

The hydrochemistry of meltwaters within glacierized catchments has been studied since the 1970s. The primary hydrochemical constituents in glacial meltwaters are solutes derived from atmospheric deposition and the weathering of catchment bedrock and glacial sediments, predominantly in subglacial or ice-marginal environments. The complex system of solute acquisition by glacial meltwaters may be conveniently simplified by a two-component model whereby bulk meltwater is composed of “quick” and “delayed” flow. Conceptually, the former can be considered as rapid supraglacial and englacial flow, along with channelized subglacial flow, which is largely devoid of solutes, while the latter relates to slower subglacial transport pathways, enriched in solute. In practice, however, these components do not exist as discrete entities, and solutes are acquired to varying degrees by glacial meltwaters that follow composite pathways, seriously undermining mixing model approach to determine solute provenance. However, it is well ascertained that solute acquisition is influenced by the residence time of waters at the ice–bed interface. In response to the variation in meltwater production during the summer melt seasons, typically high solute concentrations are observed at times of low discharge, with low flux and long contact times, and low concentrations during times of elevated meltwater discharge, with high flux and shorter contact times, both at the diurnal and seasonal timescales.

Precipitation in glacierized catchments is the primary source of base levels of solutes observed in runoff. Seasonal snowfall and snowpacks are, therefore, important as sources and stores of solutes within glacierized catchments. As snow begins to melt, the percolation of meltwater results in changes in that snowpack’s chemical composition. Field and laboratory studies have shown that 80% of the snowpack solute load may be released within about the initial 25% of meltwater runoff. This leaching or elution of
solute may be complicated by snowpack heterogeneity. Seasonal elution has also been observed in runoff from firn. This removal of solutes from snowpack and firn, the sources of glacial ice, means the ice itself is relatively dilute. As the snowline retreats up-glacier during the ablation season, dilute ice-melt comes increasingly to dominate supraglacial runoff.

The relatively high solute concentrations measured in proglacial meltwaters must, therefore, be acquired by chemical weathering of sediments in ice-marginal or subglacial environments. It is now widely recognized that hydrolysis and carbonation reactions dominate rock–water interaction and chemical weathering within glacierized basins. Calcium ions (Ca\(^{2+}\)) are the dominant cation in glacial runoff, which reflects its relatively rapid rate of dissolution from silicate and carbonate rocks, although silica concentrations tend to be lower than in nonglacial runoff because silicate weathering is depressed in cold subglacial and ice-marginal environments. The acquisition of other base cations depends on a catchment’s specific geology and geochemical susceptibility. However, observations of persistently elevated sulfate ion (SO\(_4^{2-}\)) concentrations suggest alternative weathering reactions occur along subglacial flowpaths. Anaerobic nitrate-reducing or sulfate-reducing microbes and oxidizing chemotrophs have been shown (or inferred) to exist in a wide range of glacier environments, and subglacial chemical weathering can be increased up to eightfold through microbially catalyzed reactions.

Sediment

The runoff from glacierized catchments characteristically exhibits high suspended sediment loads that influence the use and management of that water. Consequently, substantial research attention has been devoted to glaciofluvial sediment transport, which is typically confined to a limited melt season. Analyses have shown that annual specific glaciofluvial sediment yields, which routinely exceed \(10^2 \text{ t km}^{-2}\) \text{ year}^{-1}, are substantially greater than global averages for other terrestrial catchments.

In view of the conceptual model of valley glacier hydrology during the ablation season (Figure 1), the critical distinction between the quick flowpaths of the largely debris-free supraglacial and englacial systems, with low or intermediate suspended sediment concentration (SSC), and the slower, delayed flow through the subglacial drainage system, exhibiting high SSC through entrainment of rock flour and subglacial debris, is important. The transfer of fine-grained rock flour in suspension dominates sediment evacuation from most glaciers, and particularly from temperate-based glaciers. The concentrations of sediment transported via supraglacial drainage paths, although typically <0.5 g l\(^{-1}\), are highly dependent on the nature and extent of debris at the glacier surface and within the ice body as well as delivery of sediment derived from extraglacial locations, such as lateral moraines, to the supraglacial environment. Debris may be released from englacial and basal ice through the melting and enlargement of conduits by viscous heat dissipation. However, where the glacier bed is composed of unconsolidated materials, the entrainment and removal of sedimentary products will be the dominant source of suspended sediment entrained in meltwaters passing through subglacial drainage paths, with SSC commonly ranging between 1 and 10\(^2\) g l\(^{-1}\). Entrained sediment may itself increase SSC through mechanical erosion (abrasion) of bedrock or basal sediments. However, the relationship between SSC and discharge is neither linear nor stable in proglacial rivers and
sediment yields can vary markedly at diurnal, seasonal, and annual timescales.

Changes in the rate of meltwater production at the glacier surface and the hydrological configuration within a glacier control the temporal variability in suspended sediment transport by influencing the processes of erosion, entrainment, and transfer. The volume of sediment evacuated by runoff is typically viewed as a function of the availability of debris at the subglacial ice–bed interface. At a seasonal timescale, the sediment yield relates to the evolution of the subglacial drainage system: SSC (sediment evacuation) can be high during the spring melt, when snowmelt flushes the glacier’s subglacial drainage network and destabilizes the slow drainage system. Subsequently, up-glacier retreat of the snowline and corresponding growth of the fast subglacial drainage system may cause short-term increases in sediment delivery as the basal area accessed by fast-flowing subglacial meltwater increases. As the melt season progresses and diurnal melt cycles become more pronounced, sediment concentrations may be elevated as subglacial sediments are eroded by enlarging fast drainage structures. However, decreases in suspended sediment concentration may arise on a seasonal basis, arising from the exhaustion of available fines, as the subglacial drainage system spatially stabilizes and then begins to close during the latter portion of the melt season.

Short-term variability in sediment transport by runoff is usually dominated by a marked diurnal cycle, typically characterized by a strongly positive relationship between discharge and SSC. However, the direct, positive association between SSC and meltwater discharge commonly breaks down. On the diurnal timescale, clockwise hysteresis is frequently observed, whereby the flushing and exhaustion of sediment on the rising limb of the diurnal hydrograph occurs, and there is significantly less sediment available for transport when discharge declines. Similarly, anticlockwise hysteresis may occur in cases where subglacial channel water pressures reduce on the falling hydrograph limb and return flow from the distributed system elevates the SSC entrained at lowered discharges. Shorter-term subdiurnal variations in SSC, so-called sediment pulses, have been related to sudden changes in drainage system connectivity over the glacier bed, to glacier dynamics that result in reconfiguration of the drainage system, and to snowmelt or rainfall events where the drainage system becomes inundated. Additional influences on glaciofluvial sediment yields from glacierized catchments may involve the reworking and erosion of typically unconsolidated materials associated with moraines, and braid plains, or sandurs that typify ice-marginal environments and contemporary proglacial areas.

Summary and conclusion

The nature and dynamics of glacier hydrology exert critical controls on the volume and quality of meltwater runoff. Such runoff may be of critical importance as a valuable resource but, occasionally, as an unwelcomed hazard. Our understanding of glacier hydrology has been progressively built up over recent decades; conceptually, the model of drainage pathways through mid-latitude, temperate valley glaciers remains dominant. Here, the principal controls over meltwater generation, routing, and delivery have been summarized, categorizing central flowpaths as supraglacial, englacial, or subglacial. Meltwater produced in the supraglacial environment drains impeded through the snowpack and/or firn, but more rapidly over bare glacier ice. While the interstitial, primary permeability of the englacial environment cannot transport
meltwater at velocities $>1 \text{ m year}^{-1}$, secondary macroscale drainage structures represent key flowpaths delivering meltwater to the glacier interior. Once meltwater reaches the bed, it follows either slow, distributed forms of drainage, or discrete, hydraulically efficient, fast channels, although these systems likely coexist and interact. Thermal and mechanical erosion can enlarge or restructure the englacial and subglacial flowpaths while ice deformation causes their contraction. Although traditionally viewed as a top-down cascade, the supraglacial, englacial, and subglacial components of a glacier’s hydrology are interconnected and meltwater can readily transition between these environments. Seasonally, the entire drainage system of a glacier opens, develops, becomes increasingly efficient, and then begins to close once meltwater production declines. Processes of liquid meltwater storage and release occur within these three subsystems, at subseasonal, seasonal, and annual timescales, and it is these processes that strongly influence the runoff volumes and water quality draining from a glacierized catchment as well as representing a strong influence on the basal motion of a glacier.

Despite the advances in research techniques that have provisioned our current knowledge of glacier hydrology, research challenges remain, particularly to forecast drainage system and hydrograph adjustments under predicted climatic forcing and environmental change scenarios. Research priorities include (i) refining spatial and temporal variations in the generation and routing of supraglacial meltwater and the associated role of the weathering crust; (ii) improving understanding of the architecture, function, spatial extent, and temporal development of englacial drainage systems, perhaps utilizing geophysical techniques; (iii) constraining the reciprocal relationship between seasonal (and longer-term) evolution of subglacial drainage and ice mass motion. In addition to these flow-path-specific research avenues, recent awareness of microbial interaction with meltwaters warrants investigation into the potential role microbes may have on hydrological processes, water quality, and ecological habitat modification throughout the glacier system. Crucially, these questions remain not only for temperate valley glaciers but also for polythermal and cold glaciers, where ice temperatures impose complexity on the manner by which meltwater is transferred through the ice body. For ice sheets, uncertainties remain over the transferability of the conceptual, temperate valley glacier hydrological model as well as the response time of hydrological structures to changes in meltwater provision in a changing climate.

Acknowledgments

Phil Porter, Tom Holt, and Hannah Clarke for proofreading.

SEE ALSO: Fluvial erosional processes and landforms; Glacier lake outburst floods; Glacier mass balance; Glaciers; Hydroclimatology and hydrometeorology; Mountain hydrology; Snow; Water quality; Water resources and hydrological management

References

GLACIER HYDROLOGY AND RUNOFF


Further reading


A glacier inventory is a cadastre of glaciers that is, in general, regionally restricted to a country. It contains for each glacier an identifier (ID), its name (if available), location (coordinates), size, and other data. The information is either collected from topographic maps or, more recently, directly derived from satellite data and a digital elevation model (DEM). While in the early days of compilation all data were stored in a tabular text format, advances in computer technology have allowed the storage of digitized glacier outlines in a vector format with the required inventory information being available in an attribute table. A key requirement for generating an accurate glacier inventory is data acquisition at the end of the ablation period, that is, when the glacier perimeter is free of seasonal snow. This might be difficult to achieve in regions with a maritime climate and/or in high mountains, both having snow cover throughout the year. Additionally, the map or satellite image must have a high geometric quality and a scale or spatial resolution that is good enough to identify glaciers. Otherwise the quality of the inventory would be degraded; for example, only the largest glaciers would be included and details would be missing. Hence, glacier inventories were first created in the nineteenth century in regions with good topographic maps, such as those available for Austria and Switzerland. However, despite the constantly increasing quality of topographic maps, glacier inventories were only created in a few countries at that time and not continued systematically.

This changed in the international hydrological decade (1965–1975) when a working group led by Fritz Müller was appointed by the International Commission on Snow and Ice (ICSI) to prepare a “guide to world inventory of perennial ice and snow masses on and beneath land surfaces” (Hoinkes 1968). “The necessity of a new world inventory of glaciers, based on recently surveyed maps” was also emphasized by Hoinkes (1968), arguing that there was a disturbing 20% difference in the best estimates of the regions covered by ice outside of the polar regions. In the following years, glacier inventories were compiled in many regions of the world according to standardized guidelines from UNESCO (see web links) and a status report was published by the World Glacier Monitoring Service in 1989 (WGMS 1989). This so-called World Glacier Inventory (WGI) contained only preliminary information for some countries but provided the first global overview of the perennial ice masses in a standardized tabular format (Figure 1). Although the WGI served a wide range of purposes and was used for several regional-to-global scale calculations, it became clear that tracking glacier changes through time would be very difficult with the point information stored in the WGI. The emerging technological capabilities of geographic information systems (GIS) in the 1980s along with satellite data having sufficient spatial resolution resulted in numerous activities for the compilation of a global glacier inventory in a vector format from satellite data.
Source data and methods

Creation of a glacier inventory requires that glacier boundaries are available for the intended sample of glaciers. Such boundaries can be drawn (e.g., on a map) once the physical and optical characteristics of a glacier are defined. It is then a matter of interpretation by the respective surveyor, cartographer, or remote sensing specialist where to place the boundaries. Unfortunately, there is no unique interpretation of glacier extents. Glaciers exhibit a wide range of surface characteristics that can be interpreted differently without being wrong. For example, assuming that there are no restrictions in regard to visibility (e.g., due to clouds, shadow, or seasonal snow), the change from debris-covered ice to a lateral moraine – that might still have an ice core but is not moving anymore – is continuous and a precise boundary is difficult to determine in such cases (Figure 2). There are also topographic features like couloirs and depressions that have snow cover throughout the year (and maybe some ice underneath) that might be considered as part of the glacier area or not. In all cases there is variability in the resulting glacier area with an impact on the reported data in the inventory. If only the data table is available and not the source information showing where the outline was drawn, there is no way to decide whether a different area at a later point in time is due to a different interpretation or a real change. Such differences in interpretation apply to all source data used, largely independent of their spatial resolution (Figure 2).

While the earliest maps of glaciers were compiled by topographic surveying, modern topographic maps are based on aerial photography and photogrammetric techniques. The particular challenges for glaciers are related to the variable elevation of the terrain in mountain regions – resulting in a panoramic distortion (that needs to be corrected) – and the poor contrast over snow-covered regions – resulting in a reduced quality of the derived elevation information. If seasonal snow and clouds are not a problem, glacier outlines can be directly transferred from the aerial photograph to the map using a distinct color to reveal them (e.g., white with blue or green outlines and elevation contours). For a glacier inventory these outlines have to be separated and extracted from the rest of the map. For the WGI this was generally performed using planimetric techniques. Using GIS technology, the outlines are either digitized directly from the map or from a scanned version using on-screen digitizing. With the vector information being digitally available, it was for the first time possible to perform change assessment for exactly the same glacier entities. As noted above,
not all countries with glaciers have high-quality and large-scale (e.g., 1:25,000) topographic maps available. With decreasing map scale (e.g., toward 1:100,000), glacier outlines become more generalized and their geographic location more uncertain. Glacier inventories derived from such data can still be useful for a first-order assessment of the overall area covered by glaciers, but tracking glacier changes might not be useful anymore.

Finally, satellite data with sufficient spatial resolution (e.g., the 30 m from Landsat) have become publicly and freely available (e.g., on http://glovis.usgs.gov), and simple multispectral classification methods (such as a ratio of two image bands) were developed and used for automated mapping of debris-free glaciers (Paul et al. 2002). The resulting grid of classified image pixels are converted to glacier outlines (vector format) that can be uploaded in a GIS for further manual corrections (e.g., adding the unmapped debris-covered glacier parts) (Figure 3). A set of guidelines for glacier mapping (see web links and Racoviteanu et al. 2009) and calculation of inventory data (Paul et al. 2009) was created for the community to support consistency of the analysis.

The already orthorectified (i.e., terrain effects being removed and the image being transformed into a map projection) satellite scenes have several advantages compared to the also used aerial photography, among others: the large regions covered at the same time, the coverage of regions not accessible for airborne surveys, and the possibility of an automated multispectral classification based on the spectral properties of ice and snow. While the first advantage allows mapping of an entire...
GLACIER INVENTORY

Figure 3  Glacier outlines from 2003 as derived from satellite data (yellow). While glaciers in shadow are included in the raw classification, the analyst has to add the missing debris cover (red) and remove misclassified lakes (white). The final outlines are composed of the outer yellow and red lines. The satellite image in the background is a false color composite from 2009 showing glaciers in light blue. Some of the meta-information available in the attribute table for Oberaar glacier is shown in the inset.

mountain range within a day, the last one allows for a quick update of inventory information and change assessment.

The disadvantage of satellite images is their lower spatial resolution (around 15–30 m) that results in more generalized outlines and maybe problems in correctly identifying parts belonging to a glacier (Figure 3). However, the latter can also be a problem for high-resolution data (1 m or better) as Figure 2 and several tests have shown (Paul et al. 2013). More recently, high-resolution satellite data and aerial photography have also become available for viewing in tools such as Google Earth. These sources might help in the interpretation of glacier boundaries derived from lower-resolution satellite data and have been used widely for this purpose.

Existing glacier inventories

As mentioned above, basically two types of inventories can be distinguished: (i) tabular listings and (ii) vector outlines with attribute data. A third type is (iii) gridded data that indicates on an aggregated level how much of a larger grid cell is covered by glaciers. The last type is particular useful for models working on a global scale that do not resolve individual glaciers. When information about individual glaciers can
Table 1  Comparison of the different glacier inventories available.

<table>
<thead>
<tr>
<th></th>
<th>WGI</th>
<th>WGI-XF</th>
<th>GGHydro</th>
<th>DCW</th>
<th>GLIMS</th>
<th>RGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Tabular</td>
<td>Tabular</td>
<td>Grid</td>
<td>Vector</td>
<td>Vector</td>
<td>Vector</td>
</tr>
<tr>
<td>Format</td>
<td>ASCII</td>
<td>ASCII</td>
<td>csv</td>
<td>netCDF</td>
<td>shapefile</td>
<td>shapefile</td>
</tr>
<tr>
<td>Glacier count*</td>
<td>131 000</td>
<td>131 000</td>
<td>n/a</td>
<td>n/a</td>
<td>142 000</td>
<td>200 000</td>
</tr>
<tr>
<td>Sources</td>
<td>Aerial photography, maps, satellite</td>
<td>WGI, maps, publications</td>
<td>Maps</td>
<td>Maps</td>
<td>Satellite, maps</td>
<td>Satellite, maps, WGI</td>
</tr>
<tr>
<td>Scale</td>
<td>variable</td>
<td>variable</td>
<td>1:1 000 000</td>
<td>1:1 000 000</td>
<td>variable</td>
<td>variable</td>
</tr>
<tr>
<td>Resolution</td>
<td>n/a</td>
<td>n/a</td>
<td>1°</td>
<td>vector</td>
<td>mostly 30 m</td>
<td>mostly 30 m</td>
</tr>
<tr>
<td>Metadata</td>
<td>&gt;20</td>
<td>&gt;20</td>
<td>none</td>
<td>none</td>
<td>&gt;20</td>
<td>about 10</td>
</tr>
<tr>
<td>Remarks</td>
<td>point data</td>
<td>now part of the WGI</td>
<td>only used for modeling</td>
<td>used as background</td>
<td>multitemporal</td>
<td>merged product</td>
</tr>
</tbody>
</table>

*Numbers are subject to change due to ongoing work.

n/a: not applicable.

be considered in a model, type (i) inventories are more useful. Direct use has been made of the vector information stored in inventories of type (ii) only recently on a global scale (Marzeion, Jarosch, and Hofer 2012; Radić et al. 2014). Overall, the datasets available differ in the above types, their accuracy or degree of generalization, global completeness, epoch, included metadata, and file format, among others (Table 1). The overlay depicted in Figure 4 illustrates some of these datasets for Disko Island in western Greenland.

The WGI and WGI-XF

The WGI was compiled during the 1970s and 1980s mostly from topographic maps following the UNESCO guidelines (see web links at the end of this entry). The entries were organized in a fixed machine readable syntax using number codes with look-up tables to convey descriptive information. The entries listed for each glacier include a unique code (based on the hydrologic basin), name, location (longitude, latitude, country), year of map/photo, size, length, aspect, various elevations (e.g., minimum, maximum, mean), a classification (type, front, profile, nourishment, activity, moraine), and others. In some regions only preliminary information was obtained and for others data were created but not forwarded to the data centers, resulting in detailed information being available only for about one-third of the world’s 200,000 glaciers. The WGI was later extended to the WGI-XF (see web links) with newly assimilated data that are now also part of the WGI dataset available at the National Snow and Ice Data Center (NSIDC) (see web links). The rich meta-information for each glacier has been and still is used for a wide range of applications,
but the information is only available for point locations (Figure 4).

**GGHydro**

The “GLAC” layer in the GGHydro dataset provides the percentage of glacier cover for each 1 x 1 degree grid cell of the Earth’s surface (Figure 4). It has been compiled manually from 1:1 000 000 scale topographic maps and is available online (see web links). Due to the sampling of the information and the large scale of the maps used, small glaciers are underrepresented and glacier outlines are generalized. A distinction
between the two ice sheets of Greenland and Antarctica and glaciers is not made in this vector layer but available from the CRYO layer. The dataset might not be considered as a glacier inventory (e.g., individual glaciers are not resolved, meta-information is missing), but due to its spatial completeness it has been applied in several global-scale studies (e.g., for calculating the contribution of glaciers to sea level change).

The Digital Chart of the World (DCW)

The DCW has been digitized from 1:1 000 000 scale, partly military topographic maps where hydrology-related polygons are available in a separate layer. The outlines of glaciers are regionally highly generalized and locally also incomplete. A distinction between glaciers and ice sheets is not made (Pfeffer et al. 2014). The dataset is also not really a glacier inventory as only outlines are provided (i.e., glacier complexes are not further separated into individual glaciers). The DCW has been used to show where glaciers are located (Figure 4) and also to define regions with glaciers for data acquisition with the Terra sensor ASTER. The dataset is available online in an improved version called vmap0 (see web links).

The GLIMS database

The GLIMS database is the first systematic attempt at compiling glacier inventory information in a vector format using mainly satellite data. Initiated in 1995, the idea was to use freely available images from the ASTER sensor to map glaciers and provide the outlines to the database. Due to the small area covered (1/9 of a Landsat full scene) and the missing orthorectification of the ASTER scenes, glacier inventory work was demanding and progress slow. This changed when orthorectified Landsat scenes were made publicly available by the Global Landcover Facility, allowing a much faster data processing. The glacier outlines in the GLIMS database are made available in the shapefile format (a de facto standard for vector data), and have a very flexible data model for storing attribute and meta-information (Raup et al. 2007). The database is designed to store multitemporal information and a wide range of additional data (e.g., hypsographies, snow lines, debris-covered sections, calving termini, velocity fields). Glacier complexes are generally intersected by drainage divides and each resulting glacier entity has a specific code (GLIMS ID).

The Randolph glacier inventory (RGI)

The RGI was compiled in a special effort for global-scale glacier studies excluding the two ice sheets in support of IPCC AR5 (Vaughan et al. 2013). Its main purpose was to achieve global completeness based on the best available vector datasets in each region (Pfeffer et al. 2014). The RGI is thus based on glacier outlines that had already been available in the GLIMS database and was extended by datasets that were provided by the community (Figure 4), partly prior to citable publications. In some of the still uncovered regions, outlines from the DCW and nominal outlines (circles of the correct size at the respective location) from the WGI were integrated also. The RGI has a more simple data model than the GLIMS database and a reduced number of attributes. It was and still is frequently updated with improved datasets by the community (e.g., replacing the DCW and nominal WGI outlines). Apart from this, a main future goal is to increase the temporal consistency of the dataset (i.e., the epoch covered), which currently ranges from the 1950s to 2014. The RGI dataset is currently integrated into the GLIMS database to allow extraction of an RGI-like dataset from there.
The global inventory

The currently best assessment of the area covered by glaciers globally is provided by the RGI. Although there are still local-to-regional scale quality issues that need to be corrected (e.g., too much seasonal snow has been mapped as glacier ice in South America), a wide range of global-scale applications could be performed for the first time with this dataset. Examples include calculation of glacier volume or projections of twenty-first-century glacier evolution and their contribution to sea level change (Radić et al. 2014). For practical purposes, the dataset was divided into 19 first-order regions (Figure 5), following the suggestion of an earlier study. In each of these regions different data sources have been used to create the outlines, but most of them were derived from satellite data acquired in the period 1999–2011 (Pfeffer et al. 2014). Some key numbers for each region (count, area, mean size, mass, mean thickness, tidewater fraction) are summarized in Table 2. The RGI includes all glaciers and ice caps peripheral to the Greenland ice sheet and on the islands surrounding Antarctica, but it excludes those on the Antarctic mainland. A note of caution has to be made for glacier count: the numbers are highly variable and strongly dependent on several factors (e.g., the smallest size considered, the way of separating glaciers, snow patches being included or not) and will thus be different in various publications as well as subject to change in future compilations.

Referring to the study by Pfeffer et al. (2014), there are about 200,000 glaciers globally covering 727,000 km², of which 39% (by area) are tidewater glaciers (which are in contact with the ocean). When excluding region 19 (islands surrounding Antarctica), the tidewater fraction is reduced to 25%. There is an approximately two orders of magnitude variability of glacier number and areas covered among the regions (Table 2).
Table 2  Summary of glacier inventory data for the 19 main regions. For uncertainties and further details on data sources see Pfeffer et al. (2014).

<table>
<thead>
<tr>
<th>No.</th>
<th>Region</th>
<th>Count</th>
<th>Area (km²)</th>
<th>Mean size (km²)</th>
<th>Mass (Gt)</th>
<th>Mean thickness (m)</th>
<th>Tide water (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alaska</td>
<td>26 944</td>
<td>86 715</td>
<td>3.22</td>
<td>18 379</td>
<td>212</td>
<td>13.6</td>
</tr>
<tr>
<td>2</td>
<td>Western Canada and USA</td>
<td>15 215</td>
<td>14 559</td>
<td>0.96</td>
<td>906</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Arctic Canada North</td>
<td>4538</td>
<td>104 873</td>
<td>23.11</td>
<td>30 958</td>
<td>295</td>
<td>46.8</td>
</tr>
<tr>
<td>4</td>
<td>Arctic Canada South</td>
<td>7347</td>
<td>40 894</td>
<td>5.57</td>
<td>8845</td>
<td>216</td>
<td>7.4</td>
</tr>
<tr>
<td>5</td>
<td>Greenland Periphery</td>
<td>19 323</td>
<td>89 721</td>
<td>4.64</td>
<td>17 146</td>
<td>191</td>
<td>34.7</td>
</tr>
<tr>
<td>6</td>
<td>Iceland</td>
<td>568</td>
<td>11 060</td>
<td>19.47</td>
<td>3988</td>
<td>361</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>Svalbard and Jan Mayen</td>
<td>16 15</td>
<td>33 922</td>
<td>21.00</td>
<td>8700</td>
<td>256</td>
<td>43.9</td>
</tr>
<tr>
<td>8</td>
<td>Scandinavia</td>
<td>2668</td>
<td>2851</td>
<td>1.07</td>
<td>217</td>
<td>76</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>Russian Arctic</td>
<td>1069</td>
<td>51 592</td>
<td>48.26</td>
<td>15 152</td>
<td>294</td>
<td>64.8</td>
</tr>
<tr>
<td>10</td>
<td>North Asia</td>
<td>4403</td>
<td>3430</td>
<td>0.78</td>
<td>109</td>
<td>32</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>Central Europe</td>
<td>3920</td>
<td>2063</td>
<td>0.53</td>
<td>109</td>
<td>53</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>Caucasus and Middle East</td>
<td>1386</td>
<td>1139</td>
<td>0.82</td>
<td>72</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>Central Asia</td>
<td>46 543</td>
<td>62 606</td>
<td>1.35</td>
<td>4531</td>
<td>72</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>South Asia West</td>
<td>22 822</td>
<td>33 859</td>
<td>1.48</td>
<td>2900</td>
<td>86</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>South Asia East</td>
<td>16 046</td>
<td>21 799</td>
<td>1.55</td>
<td>1196</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>Low latitudes</td>
<td>2863</td>
<td>2346</td>
<td>0.82</td>
<td>145</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>Southern Andes</td>
<td>3537</td>
<td>29 333</td>
<td>1.83</td>
<td>6018</td>
<td>205</td>
<td>23.9</td>
</tr>
<tr>
<td>18</td>
<td>New Zealand</td>
<td>2752</td>
<td>1162</td>
<td>0.33</td>
<td>72</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>Antarctic and sub-Antarctic</td>
<td>14 095</td>
<td>132 867</td>
<td>48.28</td>
<td>33 749</td>
<td>254</td>
<td>98.7</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>197 654</td>
<td>726 791</td>
<td>9.74</td>
<td>153 192</td>
<td>153</td>
<td>38.7</td>
</tr>
</tbody>
</table>

Source: Pfeffer et al. 2014. Adapted from the International Glaciological Society.

This can be broken down to three or four classes when dividing the area covered by the number of glaciers, that is, when looking at the mean size of the glaciers in each region. While eleven regions are dominated by comparatively small alpine glaciers (mean size around 1 km²), three regions have mean glacier sizes of around 5 km², three are in the 20 km² range, and two have a mean size of about 50 km². The five latter regions are dominated by large ice fields or ice caps that might have not been separated into individual glaciers, and RGI regions 1, 4, and 5 have many large valley glaciers and ice caps. About three-quarters of the total glacier mass is found in five regions (1, 3, 5, 9, and 19). If all glaciers were to melt, sea level would rise by approximately 0.4 m. The ice mass calculated for region 9 is three times larger than for region 15, although the area covered by glaciers is more than 10 000 km² larger for the latter region. This indicates the different types of glaciers in both regions (large ice caps versus valley glaciers) and their very different mean thickness.
A closer look at the size–class distribution reveals that in nearly all of the 19 regions most glaciers are found in the 0.1–1 km$^2$ class, that is, small glaciers are dominating by number nearly everywhere (Pfeffer et al. 2014). Due to their large number they have also some impact on mean glacier thickness in most regions, but often a few very large glaciers result in a bias toward larger mean thickness values. For the 19 regions in the RGI, glaciers thicker than about 200 m can be distinguished from regions with mean thickness values around 60 m (Table 2), the former being dominated by large ice caps and ice fields. When looking at the area–elevation distribution (hypsometry) of glaciers, the distribution is roughly bimodal (Pfeffer et al. 2014). Most of the ice is located between 500 and 1500 m above sea level (a.s.l.) for polar regions and around 5500 m a.s.l. for glaciers in equatorial regions and in High Mountain Asia (regions 13, 14, 15). The comparatively small glaciers in the Alps, Pyrenees, and Caucasus are an exception with mean elevations around 3000 m a.s.l. Hence, despite their limited areal coverage they are situated in a climatically unique location. When considering all glaciers individually, mean elevation is a continuum with decreasing elevations pole–wards (Pfeffer et al. 2014).

**Summary and conclusions**

With glaciers impacting on hydrology from local (runoff, hydroelectric power) to global (sea level rise) scales and being considered as key indicators of climate change, accurate assessment of these resources and their changes through time requires basic information about their characteristics. This information is provided by glacier inventories in the form of tabular data (e.g., the WGI) and more recently vector outlines with attribute information attached (e.g., the GLIMS database). Compiling this information for a larger region with often thousands of glaciers is tedious work, even when latest technologies and automated classification of satellite images are used. The workload further increases when data from high-resolution sensors or aerial photography are used, as these have to be digitized manually, sometimes without providing better evidence of where the glacier boundary is located (e.g., due to a lack of contrast). As interpretation of glacier boundaries can vary without being wrong, change assessment based on outlines provided by another source might result in wrong conclusions.

With the only recently established globally complete inventory (RGI) it became possible for the first time to provide global numbers on glacier coverage and volume estimates or to model past and future glacier development and contribution to sea level change. Results of these studies might need to be modified in the future as some still existing regions with low-quality outlines in the RGI need to be replaced with more accurate data. With data from the recently launched satellite missions (e.g., Landsat 8 and Sentinel 2) being made freely available at no cost in high geometric quality (orthorectified) and with shorter revisit intervals, update and improvement of existing inventories will be more straightforward in the future. However, with experts being required to obtain a high-quality product, creation of a glacier inventory will always be time-consuming work.

**Acknowledgments**

This contribution was funded by the ESA project Glaciers_cci (4000109873/14/1-NB). Thanks go to J.G. Cogley for his careful review and P. Rastner for help with Figure 4.
Websites

WGI and WGI-XF: http://nsidc.org/data/glacier_inventory
GGHYDRO: http://people.trentu.ca/~gcogley/glaciology/gglggghy.htm
GLIMS and RGI: http://glims.org and http://glims.org/RGI
GLIMS Tutorial: http://www.glims.org/MapsAndDocs/assets/GLIMS_Analysis_Tutorial_a4.pdf

SEE ALSO: Cryosphere: remote sensing; Geocoding; Geographic information system; Glacier changes; Glacier modeling; Glaciers; Ice caps; Metadata; Scale; Sea level rise; Spatial database

References


Further reading

GLACIER INVENTORY


Glacier lake outburst floods

Holger Frey
University of Zurich, Switzerland

Glacier lakes: characteristics and processes

Occurrence of glacier lakes

The formation of glacier lakes in the course of continued glacier retreat is observed in all glacierized regions worldwide. Different processes can lead to the formation of glacier lakes, resulting in different lake types. Distinctions can be made according to the location of the lake relative to the glacier (proglacial, supraglacial, ice-marginal, periglacial, and en- or subglacial) and regarding the properties of the damming material. The latter can be rock, morainic material, ice (glacier ice, dead ice from glaciers of ground ice in permafrost zones, or highly fractured ice from ice avalanche deposits or glacier surges), or other mass movement deposits (rock avalanches, landslides, or rock glaciers).

Glacier retreat has a major influence on the landscape and people’s perception of it mostly conceived as a loss in beauty and attractiveness. Glacier lakes often can replace this loss in attraction to some degree, and constitute new attractive elements in a changing landscape. However, such lakes at the same time have a potential for rapid and catastrophic drainage, so-called glacier lake outburst floods (GLOFs). GLOFs are the most far-reaching glacial hazard and often interact in chain reactions with other hazardous processes.

Mechanisms and triggers of glacier lake outburst

Depending on the dam type, different outburst mechanisms of glacier lakes can be distinguished. Solid rock dams can be spilled over, but normally are not damaged and persist for a long time. Moraine dams can collapse in cases of high water level and the resulting low freeboard of the dam or extreme discharges. Related processes are retrogressive erosion and breach formation, slope failure, or progressive groundwater flow through the dam. The latter, also called piping, is usually caused by very large hydraulic gradients or when new cavities within the dam are formed, for instance in the case of melting dead ice or ice cores. Since all these processes result in the permanent damage of the dam, outbursts of moraine-dammed lakes are often unique events. A typical process in the outburst of lakes dammed by solid glacier ice (supra-, en-, subglacial, and glacier marginal lakes) is the progressive enlargement of subglacial drainage channels. In such cases, discharge increases continuously over hours to days or weeks, until the drainage channels are closed by ice deformation when the water pressure becomes low and falls below the ice pressure, and the outflow ceases suddenly. This process, sometimes also referred to as the Icelandic term jökulhlaup, is often reoccurring. Dams of fractured ice from ice avalanches or surging glaciers can collapse suddenly, leading to complete lake drainage in a short time and related extreme discharges.

Mass movements impacting a lake can cause very high discharges, irrespective of the dam material. In high mountain terrain potential lake outburst triggers involve mass movements like ice, rock, or combined avalanches, rockfalls,
GLACIER LAKE OUTBURST FLOODS

landsides, and debris flows or GLOFs from upstream lakes. By moving the entire water column, such impacts on lakes lead to tsunami-like impact waves, which can overtop the dam and produce a GLOF. These impact waves can swash back and forth in the lake, leading to several subsequent overtoppings with diminishing intensities. Depending on the dam material and properties, the spillover can initiate breaching or failure of the dam.

Properties of the outburst floods are strongly affected by the hydrograph at the lake outlet as well as by the down-valley topography and characteristics. Possible flow types range from debris flows, in the case of steep terrain with loose material available for transportation, mud flows and hyperconcentrated flows, to water floods. In many cases, flow transformations are observed, depending on the slope and availability of material. For instance, loose material from the dam is entrained, leading to a granular debris flow. Further down, when the terrain flattens, most of the material is deposited and the GLOF continues as a hyperconcentrated flow, which then might transform to a debris flow again, if loose material is incorporated in subsequent steeper terrain.

Hence, for an assessment of the hazard situation of a glacier lake, potential mass movements that can trigger displacement waves, lake and dam properties, and the down-valley topography and sediment availability need to be considered. This illustrates that glacier lake outbursts often are linked in process chains, starting from a mass movement impacting a lake, followed by a displacement wave and a resulting GLOF, which then in turn can act as a trigger of an outburst of a subjacent lake. Eventually, depositions from a GLOF (or any other mass movement) can dam a river in the main valley, causing the formation of a temporal lake with a potential for sudden rupture and extreme discharges. In such process chains, glacier lakes thus have the potential to multiply potential reaches of hazardous processes of glaciated regions.

Glacier lakes, GLOFs, and global warming

In a period of atmospheric warming, glaciated high mountain regions undergo changes on different spatial and temporal scales, several of them affecting the hazard posed by glacier lakes. A comprehensive assessment of the hazard situation should thus not only analyze the current situation, but also consider potential future changes of different processes, in order to anticipate potential future developments. In the following, several processes influencing the hazard of glacier lakes are described.

Glacier fluctuations change the disposition of potential starting zones of ice avalanches, that is, if a glacier tongue retreats into a steep part of the bedrock, the potential for ice avalanches increases. If the tongue then recedes from this terrain step (or re-advances back into flatter terrain below), the susceptibility for avalanches decreases again. On a longer timescale, stress fields within steep rock parts or moraines can be modified due to de-buttressing caused by glacier retreat and thinning since the Little Ice Age, thus enhancing the susceptibility for failure. Furthermore, new terrain exposed after glacier retreat often consists of unconsolidated material that, if situated in steep terrain, constitutes starting zones of debris flows and landslides.

Changes in the thermal conditions of the underground influence the hazard disposition of ice and rock avalanches in different ways. Transitions from cold to polythermal or temperate beds of hanging glaciers and the occurrence of liquid water at the ice–rock interface reduces the basal friction and can thus destabilize hanging glaciers.
that are known to be stable. Permafrost degradation and related deep and long-term warming of perenni-ally frozen rock walls can increase the probability of rockfall. Finally, degradation of ground or dead ice in moraines can reduce the stability of moraine dams and increase the probability of piping within the dam (see above).

Glaciers in all mountain regions have experienced retreat since the Little Ice Age and today in most cases have an extent that is smaller than ever before in historic times. Not only are new lakes forming at locations that have been covered by glacier ice before, but also the disposition of potential triggers for lake outbursts are different than in the past. This implies that the situation of glacier lakes and potential triggers for lake outbursts are beyond historical experience and require assessment methods other than consulting known events from the past.

Hazard assessment and prevention

Assessments of high mountain hazard include gaining an overview of the situation, involved processes, and potential interactions (cf. Huggel et al. 2004). Based on these findings, a monitoring system should be established, the acceptable risk defined, responsibilities clarified, and further steps prioritized. After gaining a first overview of involved processes, potential magnitudes and probabilities need to be estimated. This can be achieved by applying rules of thumb developed empirically from past events, by consulting simple models that use such empirical relations, or by modeling potential processes with physically based models.

GLOF modeling

A large variety of models to simulate or estimate mass movements exist, having different complexities and different input requirements. They can be distinguished as either flow routing models or physically based models. The latter are able to simulate, for instance, kinetic energy and temporal evolutions of the flow, whereas the former are used to determine the direction and extension of the mass movement. Such approaches can be applied to large regions and normally meet the requirements for a first-order hazard assessment.

Physically based models typically are applied to more specific case studies, since they require more computational resources and also diverse input and calibration data that often are not available. Most such models require the initial volume of the mass movement that is being simulated, and there can be high uncertainties in cases of high mountain mass movements. The flow propagation is then modeled by solving equations of mass conservation and momentum, integrated over flow depth, for discrete time steps. Flow rheology is often approximated by using frictional parameters. Recent developments aim at integrating the entrainment of material along the flow trajectory, a process that is particularly important for GLOFs. Worni et al. (2012) were able to realistically simulate the discharge hydrograph, breach formation, sediment depositions, and hence also flow-type transformations of an outburst of a moraine-dammed lake in Patagonia.

Estimating values of input parameters for complex models is even more problematic for potential future events. Concepts from other natural hazards, such as flash floods for instance, are not applicable to glacier hazards, since for the latter typically no records exist, as these are new, unknown situations and related events often are unique. For mass movements, this shortcoming is often circumvented by using magnitude–frequency relations. However, data are typically scarce for establishing such relations,
GLACIER LAKE OUTBURST FLOODS

also because small events often are not recorded. Schneider et al. (2014) presented a remarkable work on modeling different outburst scenarios for Laguna 513, a glacier lake in the Cordillera Blanca, Peru. They used a chain of interacting, physical models to simulate different rock/ice avalanche scenarios, the development of the subsequent displacement wave in the lake and the overtopping of the dam, and finally the resulting GLOF, including different flow types. A reconstruction of a real event from 2010 was used for model calibration, and they had a high-quality digital elevation model, lake bathymetry, and records of the wave height available, conditions that are normally not fulfilled for most remote glacier lakes. Nevertheless, this study demonstrates the potential of linking such physically based models for providing reliable and reproducible information on potential future events.

Structural risk prevention measures

If the risk posed by a glacier lake is considered to be unacceptably high, the probability of a potential outburst can be lowered by either lowering the hazard or lowering the vulnerability of potentially affected people. The most rapid way to lower the hazard potential of a glacier lake is to lower the lake level. This is sometimes done as an urgent measure by constructing siphons that can lower the lake level rapidly. Another, more sustainable way to lower the lake level constantly is the excavation of tunnels through bedrock dams. For moraine dams a typical method is to make a cut into the freeboard of the dam and reinforce the dam and the outlet, or to place a culvert and to restore the freeboard. Such works have been successfully applied in different mountain regions, namely the Andes (mainly in the Cordillera Blanca, Peru), the Himalayas (Nepal), and the Alps. Carey et al. (2012) give an overview of such mitigation measures performed at Laguna 513 (see also above). Other mitigation measures involve construction works not at the lake but related to protective measures for potentially threatened infrastructure. Examples are diversion dams to protect exposed infrastructure or constructions with the potential to retain certain amounts of potential outburst floods. Such constructions have the potential for multipurpose use, for instance in combination with hydropower production.

Nonstructural risk prevention measures

Measures to increase the preparedness of the population provide an alternative to physical mitigation measures. Exact forecasting of the moment of a lake outburst or a potential lake outburst trigger remains problematic, but systematic monitoring of critical situations, using both on-site observations and remote sensing, offers the possibility to establish different alert levels and to alarm and evacuate the population in case of an event. Such monitoring can be either limited in time or permanent in the form of an early warning system (EWS). Only a few EWSs exist so far for glacier lakes, but they have a large potential for complementing or replacing mitigation measures. Understanding of the entire system of involved processes and a well-designed monitoring and warning service, including redundancy, are fundamental for EWSs. It is, however, important to recognize that such systems are highly complex and that, besides technical aspects, they include institutional and social components: responsibilities need to be clarified with authorities, ways of communication need to be established and standardized, and permanent contact with the population is necessary to achieve the required response.
Future perspectives

In the future, the formation of numerous new glacier lakes is expected in all glacierized mountain ranges. Such lakes will compensate the loss in landscape attractiveness caused by the vanishing glaciers to some degree, but they need to be carefully and repeatedly observed since they hold the potential for catastrophic drainage, in particular with the increased potential of mass movements associated with atmospheric warming. Conditions in high mountain regions will undergo constant and rapid changes beyond historical precedence, and hence will require innovative approaches to assessing complex and interacting processes influencing the hazard potential of glacier lakes. Establishing standardized assessment procedures will help in defining monitoring strategies and setting priorities for mitigation and adaptation measures. Socioeconomic developments such as changing settlement densities in high mountain valleys will also influence the risk posed by glacier lakes.

SEE ALSO: Climate change and permafrost; Glacier changes; Glaciers; Mass movements in periglacial environments; Natural hazards and disasters; Snow and ice avalanches

References


Further reading


Glacier mass balance

J. Graham Cogley
Trent University, Canada

Glacier mass balance is the sum over the extent of the glacier and over a stated span of time of mass gains, treated as positive, and mass losses, treated as negative (Figure 1). Mass gain is referred to generically as accumulation and mass loss as ablation. Most accumulation is by snowfall, but avalanching of snow from valley walls can be significant. On most glaciers ablation is mainly by runoff of meltwater, but ice discharge – the calving of icebergs or, equivalently, the flow of ice across a grounding line – is significant on many.

Of the global glacier area of 726,800 km², 39% is tributary to tidewater termini (Pfeffer et al. 2014). This area excludes the Greenland ice sheet (1.7 million km²) and the Antarctic ice sheet (12.3 million km²). Thus the ice sheets dwarf the other glaciers, of which there are 200,000 or more.

The word “glacier” has a deep history, traceable to an Indo-European root with the general sense of “cold, freezing,” but its first appearance in English was in 1744, when William Windham reported from Chamonix in Savoy that “the whole Glaciere has a kind of Motion … it slides continually towards the Outlets into the Valley.” This observation influenced later thinking. For example, the Oxford English Dictionary’s definition of “glacier” begins as “A large accumulation or river of ice.”

However, in mass-balance studies the volume under study must be demarcated so that horizontal mass transfer can be quantified. The glacier is separated from its neighbors, if any, at ice divides where this transfer is zero. At grounding lines or calving fronts, the balance must include the outward transfer. Within the boundary there is an accumulation zone at higher elevations where low temperature favors accumulation and an ablation zone at lower elevations where warmth favors melting. The two zones are separated by the equilibrium line at which the local mass balance is zero. It follows from this definition and elementary physics that Windham was right: the ice must flow. If accumulation exceeds ablation the glacier grows, if the converse is true it shrinks, and if the two are equal, and remain so, it attains steady state.

The concept of mass balance had to await the development of this modern definition of the glacier. The first recognizably modern measurement was published in 1935, but the term itself did not become current until the 1960s, when measurements became widespread.

Mass balance has the dimension [M], mass, and the gigatonne (Gt) is often a convenient unit for calculations. The specific mass balance, which is the mass balance per unit area, has the dimension [ML⁻²]; common units are kg m⁻² or mm water equivalent, the two units being numerically identical. The meter of sea level equivalent (m SLE) is the total mass balance (kg) divided by the product of the density of water (1000 kg m⁻³) and the area of the ocean (362.5 × 10¹² m²), such that 1 mm SLE = 362.5 Gt. When regarded as a rate, the mass balance has dimension [MT⁻¹] or [ML⁻² T⁻¹]. Over periods of a year or longer the natural unit of time is the year.
Measurement of mass balance

Measurements of mass balance and other glacier changes are archived at the World Glacier Monitoring Service (UNEP/WGMS 2008).

Glaciological method

In the glaciological method stakes are inserted into the surface in a spatially representative network. The local change of mass is the change of surface elevation, relative to the top of the stake, multiplied by the density of the matter gained or lost. In the accumulation zone the density is measured in snow pits, while in the ablation zone it is usually assumed to be 900 kg m$^{-3}$, near the density of pure ice. The glacier-wide mass balance is extrapolated from the stake network, often by assuming that the local balance depends only on elevation. With this assumption, the glacier’s hypsometry must be known so that each stake can be given its appropriate weight, multiplying its specific balance by the area it represents and summing the products.

Measurements at stakes yield only the surface mass balance. There are also internal and basal components, which can often be assumed to be small. More often they are estimated indirectly or even neglected, possibly leading to large biases. Meltwater that percolates to depth instead of running off, for example, may reduce mass loss significantly.
The frontal mass balance, due to ice discharge or frontal melting, is never positive and may be known confidently to be zero, but if not it must not be neglected. This is particularly true of the Antarctic ice sheet, whose mass loss is overwhelmingly by frontal ablation. Here the ice discharge is measured by measuring its components, ice thickness multiplied by velocity, near the grounding line, with accumulation up-glacier usually obtained from regional-scale climate modeling.

On most glaciers there is a seasonal as well as a spatial separation between accumulation, typical of winter, and ablation, typical of summer. Winter and summer may be defined by fixed calendar dates, but when the glacier cannot be reached on those dates the field survey must be corrected somehow for unmeasured changes before or after the survey date. Alternatively the seasons can be defined in terms of the epochs of annual minimum (end-of-summer) and maximum (end-of-winter) glacier mass, in which case the durations of the seasons will vary and may be unknown.

Geodetic method

Geodetic methods, which have long been used as checks on the accuracy of glaciological measurements, record the mean change of the glacier’s surface elevation relative to a fixed external datum. The resulting volume balance is converted to a mass balance by assuming a density. In the absence of in situ measurements, this assumption introduces uncertainty. It often takes the form of Sorge’s Law, which asserts that the glacier has an unchanging profile of density as a function of depth beneath the surface, with the corollary that the matter gained or lost has the density of ice. Departures from Sorge’s Law become less serious as the relative extent of the accumulation zone becomes smaller.

Terrestrial and airborne surveys have largely been replaced by orbital surveys, greatly increasing the number and coverage of geodetic measurements, which now, when reckoned in mass-balance years, far outnumber glaciological measurements (Figure 2). However, the time between surveys is typically years or decades, so that geodetic methods contribute little to our understanding of year-to-year variability.

Gravimetric method

The GRACE satellite mission, launched in 2002, tracks changes in the Earth’s gravitational field with high temporal resolution (1 month.
or shorter) and coarse spatial resolution (of the order of 350 km). These changes are due to movements of water in the near-surface environment and of rock in the deep interior. For glaciological purposes the tectonic, oceanographic, and purely hydrological mass changes are biases and must be removed. Where the residual mass-balance signal is weak, as in mountain ranges with few glaciers or with regional mass balance near zero, GRACE has little to offer. Nevertheless gravimetry has become a standard tool for measurement of the mass balance of the ice sheets and of heavily glacierized mountainous regions and has contributed to assessments of the mass balance of all glaciers other than the ice sheets (Gardner et al. 2013).

Modeling of mass balance

Models of the surface mass balance rely on precipitation and temperature information, sometimes from in situ or nearby measurements, but more often from climate models of global or regional scope. The climate model cannot resolve the size of most glaciers and its topography is smooth, so it cannot simulate the strong but poorly quantified tendencies for precipitation to increase with elevation and to be greater on the windward than the leeward sides of mountains. The vertical gradient of temperature is better understood, but even so the correction of climate-model temperatures to those responsible for surface melting over the elevation range of the glacier is an uncertain procedure.

Some models estimate surface melting by solving the surface energy balance, but they are more demanding of input information, and therefore less versatile, than temperature-index models. The latter rely on field measurements to calibrate the conversion of climate-model temperatures to ablation, sometimes including additional forcing variables when they are available. The temperature index is usually the positive degree-day sum, which is the mean temperature excess above the freezing point multiplied by the duration of the excess. All temperature-index models acknowledge the difference in reflectivity between snow and exposed ice, the latter being darker and therefore yielding significantly more melt per positive degree-day.

For completeness mass-balance models must also treat the internal and basal balances, requiring knowledge of the subsurface temperature and hydraulic conductivity.

Models of frontal ablation are less advanced than models of surface ablation. Progress has been made in understanding the physics of how crevasses propagate through the entire thickness of ice, so as to detach the iceberg from the parent glacier, but individual calving events remain unpredictable. Measurements of ice discharge have little or no predictive power. Progress has also been made in simulating the behavior of grounding lines, but this remains a difficult subject.

For simulations over more than a few years, the slow dynamic response of the glacier to climatic forcing must be modeled. The glacier always seeks equilibrium with the climate, and both advance and retreat are stabilizing feedbacks. However, the rate of response is limited by flow speed, which is typically only tens of meters per year while the glacier is typically thousands of meters long.

History and future of mass balance

Dated terminal moraines document a maximum in glacier extent as the response of glaciers to the cooling known as the Little Ice Age. The implied dates of maximum mass lie between the late seventeenth and early twentieth centuries.
in different regions, with the global average near to 1850 according to analysis of long time series of glacier length changes calibrated against mass-balance observations made during 1950–2005. This analysis, updated in Vaughan et al. (2014), yields an estimate of approximately 80 mm SLE of loss since 1850. Estimates of total glacier mass in the early twenty-first century range from 350 to 470 mm SLE (Pfeffer et al. 2014), so the length changes suggest a loss to the present of perhaps 15–20% of the uncertain total mass in 1850.

The loss has not been steady. For example, loss rates were greater than average during the 1920s. A twentieth-century mass-balance simulation, driven by reconstructions of temperature and precipitation, exhibits even larger rates during the 1930s which can be localized to glaciers peripheral to the Greenland ice sheet (Vaughan et al. 2014). Without mass-balance measurements at these early times it is not possible to decide whether the increased losses are observational or modeling artifacts, or whether they represent real historical events. There is, however, indirect evidence of accelerated losses in Greenland between 1920 and 1940.

In mass-balance measurements made since the 1940s there is evidence of a minimum in the rate of loss centered near the 1960s (Figure 3), consistent with slight cooling from the 1940s to the 1970s followed by warming. (Northern Hemisphere temperatures are shown because most of the measurements of both temperature and balance are from there.) The first decades of the twenty-first century have seen the most negative balances ever measured.

The agreement in Figure 3 between mass balance and temperature is good to fair, partly reflecting measurement errors in and undersampling of both variables. However, it is also consistent with the well-established finding that over the past 70 years trends have been much clearer and more globally coherent in temperature than in precipitation.

There is substantial year-to-year variability in mass balance and also hints (via the height of the confidence envelope) of strong spatial variability within each year. More importantly, a correction for the spatial bias of the measurements affects the estimate of the global mean noticeably. The corrected red “staircase” is not always close to the center of the blue confidence envelope, suggesting that the glaciological bias toward well-studied mountain ranges is significant. There is a doubt about the reliability of the spatial correction, however. In regions with few or no glaciological measurements, recent gravimetric and geodetic measurements show noticeably less loss than the spatially corrected estimates. Moreover, longer-term variability in the corrected estimate appears less than in the uncorrected annual averages, due in part to the inclusion of geodetic measurements of which most are several to many years long.

Measurements of ice-sheet mass balance have been made only since the 1990s. Most mass loss from the Antarctic ice sheet is by ice discharge, while on the Greenland ice sheet the contributions of melting and ice discharge to ablation are roughly equal. There remain major gaps in understanding of such matters as the role of meltwater in promoting basal sliding (accelerating the delivery of ice to the grounding line), the stability of the grounding line (especially when the bed of the outlet glacier slopes downwards inland), the contribution of subaqueous frontal melting to the balance (water being much more effective than air at transferring heat to the ice), and the role of floating ice in buttressing the grounded ice (thus retarding ice discharge).

Nevertheless all glaciers have shown accelerating losses since the early 1990s (Figure 4). The acceleration has been more marked for the ice sheets (as shown by the upward concavity of their
Figure 3  (a) 90% confidence region of annual-mean Northern Hemisphere near-surface air temperatures, expressed as deviations from the 1961–1990 mean (http://www.cru.uea.ac.uk/cru/data/temperature/CRUTEM4-nh.dat). (b) Blue: global mass-balance rate from the dataset of Figure 2, shown as a 90% confidence region centered on the annual means (not shown) of measurements by glaciological methods; red: global pentadal-mean mass-balance rate (thick line with thin lines demarcating the 90% confidence region), including geodetic measurements and with the spatially biased sample corrected by an interpolation algorithm; measurements are too few before 1950 for reliable interpolation. (c) Number of measurements annually by glaciological methods.
GLACIER MASS BALANCE

Figure 4  Cumulative mass balance of the Antarctic and Greenland ice sheets and of smaller glaciers excluding those peripheral to the ice sheets. Reproduced with permission from IPCC (2013) Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Figure TS.3 (lower panel). Cambridge University Press.

Table 1  Global glacier mass-balance rates,* 1991–2012 (mm SLE a\(^{-1}\)).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Antarctic ice sheet</td>
<td>−0.10:±0.27</td>
<td>+0.20:±0.61</td>
<td>Vaughan et al. (2014)</td>
</tr>
<tr>
<td>Greenland ice sheet</td>
<td>−0.02:±0.20</td>
<td>+0.43:±0.76</td>
<td>Vaughan et al. (2014)</td>
</tr>
<tr>
<td>Other glaciers†</td>
<td>+0.43:±0.76</td>
<td>+0.50:±0.80</td>
<td>Figure 3b (red)</td>
</tr>
</tbody>
</table>

* Given as 90% confidence intervals. Positive numbers represent loss from the glaciers.
† Excluding those peripheral to the Antarctic ice sheet.

curves) than for the smaller glaciers (Table 1). In the first decade of the twenty-first century, these three contributions to sea level rise have been of the same order, and their sum, about 1.6 mm SLE a\(^{-1}\), is slightly greater than the other leading contribution, 0.8–1.4 mm SLE a\(^{-1}\) from thermal expansion of ocean water.

The sizes of glacier accumulation zones are well correlated with their mass balances, accounting for a varying fraction of the total area. The average fraction is approximately 0.44 on glaciers with balance measurements during 1997–2006 (Bahr, Dyurgerov, and Meier 2009), but a reasonable estimate for glaciers when they are at equilibrium is 0.57. That is, the glaciers are too large for the present-day climate. The mass loss needed for all glaciers to reach the equilibrium fraction, and therefore the right size, for the climate of 2000–2010 is 184 ± 33 mm SLE. Adding the loss from 1850 to present, this represents a loss of 50–60% of the ice mass at the culmination of the Little Ice Age.
Figure 5  Simulations of cumulative global glacier mass change relative to the 1986–2005 mean (upper panel; lower panel: smoothed annual rates). Glacier model does not simulate calving. Some cumulative losses exceed 100% because of upscaling to allow for Antarctic peripheral glaciers. Thin lines: model forced by individual climate models. Thick lines: means of thin lines. RCP<sub><i>n</i></sub>: representative concentration pathway <i>n</i>; human radiative forcing of the climate system increases with <i>n</i>. Reproduced from Marzeion, Jarosch, and Hofer (2012) under the Creative Commons Attribution 3.0 license.

Table 2  Projections of ice-sheet mass balance,* 1986/2005 to 2081/2100 (mm SLE).

<table>
<thead>
<tr>
<th></th>
<th>Low emissions</th>
<th>High emissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenland ice sheet</td>
<td>40:100</td>
<td>70:210</td>
</tr>
<tr>
<td>Antarctic ice sheet</td>
<td>−30:140</td>
<td>−60:120</td>
</tr>
<tr>
<td>Rapid ice-sheet dynamics</td>
<td>30:190</td>
<td>30:200</td>
</tr>
<tr>
<td>Sum</td>
<td>40:430</td>
<td>40:530</td>
</tr>
</tbody>
</table>

* Given as 90% confidence intervals. Positive numbers represent loss from the ice sheets.

It also represents committed change. That is, the glaciers would continue shedding mass even if warming were to stop immediately. Thus the glaciers are insensitive indicators of climatic change, integrating the impact of one to many past decades of forcing.

More detailed projections relying on glacier–climate models all show continued mass loss, but the choice of climate scenario influences results strongly, as in Figure 5. Aggressive reduction of human emissions of greenhouse gases (RCP26) limits glacier mass loss up to 2100 to approximately 40% of the late twentieth-century total, while if society remains on the current high-emission pathway (RCP85) the loss is approximately 60%. The average loss by 2300 for RCP26 is more than 60% and RCP85 yields losses nominally exceeding 100%. These averages conceal the largest uncertainty, which is due to the climate models. For example, the six simulations extending to 2300 have dates of complete disappearance under RCP85 between 2120 and 3000.

Figure 5 also illustrates the concept of “peak meltwater.” Total meltwater yield is expected to rise as radiative forcing increases, but then to diminish as more and more glaciers disappear. Peak meltwater is reached before 2050 with aggressive mitigation and late in the twenty-first century with continued high emissions.

For projecting the future of ice-sheet mass balance (Table 2), the scenarios of Figure 5 are of limited value because surface temperature over the Antarctic ice sheet will remain mostly below freezing. In a warmer world, increased Antarctic snowfall is likely to outweigh increased melting, although the balance of the Greenland ice sheet will become dominated by losses from the surface as tidewater outlets retreat onto land. The role of “rapid dynamics” is very difficult to predict given current understanding of the observed and possible phenomena, which include unchecked retreat of grounding lines, especially where outlet glaciers are grounded below sea level. Church et al. (2014) assign only a moderate role to rapid dynamics over the twenty-first century. They do not attempt to quantify the likelihood of rapid-dynamic losses reaching catastrophic proportions.

SEE ALSO: Climate change and land ice; Glacier changes; Glacier modeling; Glaciers

References


GLACIER MASS BALANCE


Further reading


Glacier modeling

Daniel Farinotti
Swiss Federal Institute for Forest, Snow and Landscape Research, Switzerland

The term “glacier modeling” generically describes the activity of representing glacier processes through numerical simulations. Four main branches can be identified in glacier modeling activities, reflecting the interactions of glaciers with other compartments of the Earth system. These branches include the modeling of (i) the internal processes of a glacier, (ii) the interaction between glaciers and the atmosphere, (iii) the processes at the glacier base, and (iv) the interactions between glaciers and open water. The four branches are briefly discussed hereafter.

The most developed part in the modeling of glacier-internal processes is the description of glacier flow. The aim of these so-called ice flow models is to simulate glacier motion and the associated processes. Traditionally, glacier ice is considered as a heat-conducting viscous fluid, whose motion is due to the action of gravity. Only recently have individual models included considerations about elastic properties, especially when addressing ice deformation at short timescales (i.e., hours to days). Two main components contribute to glacier motion on seasonal to secular time scales: basal sliding, that is, the relative movement of the ice body over the subsurface topography supporting it, and creep deformation, that is, the ice deformation caused by mechanical stress. A widely used description of the relation between stresses and strain rates is given by Glen’s flow law for glacier ice, which is implemented in most numerical models.

In this respect, state-of-the-art models consider a three-dimensional full Stokes formulation of the flow and are capable of representing arbitrary evolutions of the glacier geometry. In comparison to creep deformation, the knowledge about the processes controlling basal motion is rather limited. Models often describe basal sliding through empirical parameterizations that include the computed basal shear stress. Another focus of ongoing research is the description of crevasse formation with continuum damage mechanics, which is of particular importance for calving glaciers (see below). Other internal processes sometimes included in models are the internal mass balance, that is, the difference between processes that increase glacier mass and processes that result in the opposite, and the modeling of internal water flow and storage through conduits, cracks, and lakes. Internal mass balance is often taken care of by models dealing with glacier–atmosphere interactions (see below), and well-established theoretical concepts exist for the growth, evolution, and collapse of englacial and subglacial water conduits. The mechanisms contributing to the formation of internal water bodies, so-called water pockets, are poorly understood and, thus, seldom represented in models. For a general introduction in the modeling of glacier ice dynamics, refer to Greve and Blatter (2009).

The modeling of glacier–atmosphere interactions focuses mainly on the description of the surface mass balance of a glacier. In this case, the aim is to model all processes that contribute to glacier mass gain (accumulation) and loss (ablation) at the surface. Processes that contribute to surface accumulation are mainly the deposition of meteoric or wind-drifted snow and
avalanching. Depending on the local climatic conditions, significant additional accumulation can be due to condensation (hoar deposition), freezing rain, hail, or refreezing liquid rain. Processes that induce surface ablation, on the other hand, include snow- and ice-melt, wind erosion, and sublimation. A wide range of models exist for the representation of the surface mass balance. The degree of complexity of the model used is mainly determined by the application context: while studies aimed at calculating the surface mass balance over a large (spatial or time-) scale are forced to use rather simple models because of computational cost and availability of input data, models that are designed for reproducing the mass balance at the small scale are aimed at representing the various processes individually and in detail. Among the simpler models, many are based on so-called temperature-index formulations for the ablation part, that is, an empirical relation between air temperature and melt, and a distinction between snow and rain based on a temperature criterion for the accumulation part. More complex models, on the other hand, are based on energy-balance formulations and compute the actual melt as the residual of the energy balance equation which considers the solar radiation (often divided into various components according to the spectra) and the sensible and latent heat fluxes, as well as the ground heat flux. Similarly, these more sophisticated models compute total accumulation from the sum of various components (solid precipitation, wind drift, avalanching, etc.) that are modeled individually. When snow- and ice-melt occur, part of the generated water will run off at the surface, part of it will infiltrate into the glacier, and part of it will evaporate. The ratio between these three components is mainly determined by the atmospheric conditions and the ice temperature, and the determination of these individual shares is a task usually taken care of by the same models as those considering surface mass balance. A currently open question in the modeling of glacier surface mass balance is the influence of the deposition of black carbon and other aerosols. While it is known that such deposition can significantly lower the albedo of the surface, the magnitude of the effect on surface mass balance has not yet been quantified. Similarly, current models are inadequate in representing the effect of supraglacial debris (i.e., sediments and rocks deposited on the glacier surface): it is known that thin (typically <5 cm) layers of debris enhance glacier melt through a lowering of the albedo, and that thick (typically >10 cm) layers decrease melt through an insulation effect, but correctly accounting for these effects within models is still an open challenge. For an overview of glacier–atmosphere interactions and their modeling, refer to Oerlemans (2009). For a complete definition of concepts related to the mass balance of glaciers, see Cogley et al. (2011).

The modeling of the processes at the glacier base, that is, at the interface between ice and lithosphere, includes the modeling of glacier sliding, which is an important component of glacier motion, the modeling of subglacial water flow, and the modeling of basal accumulation and melt. The last of the three mentioned processes, that is, basal mass balance, is generally considered to be of minor importance, especially when compared to the magnitude of the surface mass balance. The processes of pressure-induced melting, melting through friction or geothermal heat, and ice accretion are, thus, often neglected in models, or mimicked through empirical functions. Considerable effort is currently spent in the representation of subglacial water flow. Particular interest in the topic is given by the close connection of the process to basal sliding, which can significantly affect overall glacier motion. Only recently has the general understanding been moving away from the simplistic “more
subglacial water equals more sliding because of lubrication” to a more complex pattern of interactions. To this end, recent models have been aimed at representing the water flow in subglacial conduits and distributed drainage systems separately. Conduits melted into the base of the ice are known as Röthlisberberger channels (or R-channels), develop during periods of high surface melt (i.e., summer), and form a high-capacity, highly efficient discharge network. On the other hand, distributed drainage systems are typical for winter situations, and consist of a low-capacity, low-efficiency water sheet including flow between so-called linked cavities, thin uniform water sheets, or permeable subglacial till layers. Physically based theories for the evolution of R-channels were proposed early on, while the mechanisms steering the evolution of distributed drainage systems have been proposed more recently. Models that successfully account for both processes are emerging, but a large open question is their validation. The main reason that explains the limited knowledge of these processes and the uncertainties in the relative models is the obvious difficulty in monitoring and measuring any of the relevant variables and processes. Providing direct access to the glacier base is both laborious and costly, while the instrumentation of the glacier base itself can be an even more challenging task because of the equipment requirements. For a review of different subglacial processes and their modeling, refer to Clarke (2005).

Compared to the above-mentioned topics, the importance of modeling the interactions between glaciers and open water has been recognized only more recently. Traditionally, ice–water interactions have been a topic in oceanographic and climatologic sciences, and have focused on the interactions between sea ice and the oceans because of their fundamental role in both the climate and the oceanic system. More recently, studies have highlighted the importance of considering the interactions involving glacier ice as well. Important processes in this respect include so-called glacier calving, that is, the mechanism that induces the breakoff of large ice blocks at the front of a glacier that is in contact with a water body, as well as subaqueous mass balance components, and the dynamics of ice shelves. Both ice calving and subaqueous mass balance can substantially contribute to overall glacier ablation. The importance of these phenomena is given by the large fraction of water-terminating glaciers, especially when considering ice caps and the two ice sheets Antarctica and Greenland. Recent studies have, moreover, repeatedly shown the effect that the removal of ice shelves (that is the floating part of glaciers draining an ice cap or an ice sheet) has on the dynamics of inland glaciers: after the collapse of the so-called Larsen B Ice Shelf, Antarctic Peninsula, in 2002, for example, an increase in ice-flow speed by up to 600% has been observed for the adjacent inland glaciers. This is alarming since the additional ice discharge directly contributes to sea-level change, and according effort is currently promoted for correctly including such effects in modeling studies. In models, calving processes have traditionally been described through empirical relations, mainly based on parameterizations of readily available variables such as the water depth at the calving front. Only recently have approaches that allow an explicit representation of the mechanics involved in glacier calving been emerging. Such approaches are, however, computationally expensive, and are therefore seldom included in large-scale models that represent glacier–ocean interactions. Such kinds of models, on the other hand, increasingly make use of the knowledge available in ocean sciences, in particular to better represent the processes that steer subaqueous glacier melt. Recent studies, in fact, point to a
hitherto underestimated importance of subaqueous glacier mass balance driven by various ocean and marine currents, especially underneath ice shelves. Much as for the buttressing effect of ice shelves, correctly modeling such interactions is of paramount importance, especially in studies addressing long-term sea-level change. Concerning the correct representation of ice shelves as such, particular effort is currently spent in including the elastic properties of ice in the modeling framework. While the importance of considering the elastic components of ice deformation at short time scales had been acknowledged earlier, only recent studies have suggested that this behavior can explain the formation of a hydrological barrier near the grounding line of ice shelves that controls, among other processes, subglacial hydrology. For an introduction to the modeling of ice–ocean interactions, refer to McPhee (2008).

SEE ALSO: Glacier changes; Glacier hydrology and runoff; Glacier mass balance

References


Further reading

Glaciers and the global environment

In programs of global climate observation, glaciers are sometimes called “unique demonstration objects of climate change.” There are two basic reasons for this: (i) everybody can observe within years to decades, from field visits or publicly available satellite imagery, the often striking changes in glacier extent, and (ii) the melting of ice under the influence of warm conditions is an experience of common life for many and, hence, easily understandable.

The continued shrinkage of glaciers in most mountain ranges is indeed a key indication of ongoing global warming (Figure 1) and as such even part of political considerations at national to international levels. Exact numbers, deeper understanding of the processes involved, and model projections of their future development are provided by more than a century of scientific research and monitoring, which recently experienced rapid growth and made important profit from new technologies. Model projections also help to assess impacts from global warming and glacier vanishing on environmental conditions at various scales of space and time.

With the retreat and disappearance of glaciers, the landscape of cold mountain chains loses an element, which is seen by most people as a treasure and beauty. The corresponding reduction of cold-mountain landscape diversity is striking and profoundly changes the public perception of glaciers. The romantic enthusiasm of the eighteenth, nineteenth, and early twentieth centuries still saw the “pure” and “eternal” ice of glaciers as a strong symbol of an intact relation between humans and their environment (Figure 2). As a consequence, glacier pictures were intensively used as postcards and for touristic advertising, books about landscape fascination, and governmental brochures on environmental programs. During the past decades, accelerated loss of glacier ice in many icy mountain ranges and the strong effect on landscape appearance of the exposed bedrock and debris fields caused increasing concern. For many, retreating glaciers today have become strong symbols of a heavily disturbed human–environment relation, perhaps even the “writing on the wall” with respect to possible future environmental disturbances from global warming of even (much) larger dimensions.

As frozen bodies, glaciers play an important role in the water cycle. They store precipitation when it falls as snow during the cold or wet season and release it as meltwater during the warm or dry season. This is essential for livelihood and agriculture in large regions up to continental scales surrounding cold mountains as the glaciers guarantee water supply during the growth period of vegetation and/or dry periods, when no water comes from the sky. Probably more than a billion people, especially in parts of central Asia, South America, or the North American Prairies, to some degree depend on this effect. The meltwater from shrinking glaciers is also contributing to rising sea levels. The total volume of all glaciers in the world is estimated at only about 4 dm of sea level equivalent but will primarily affect
the rate of sea level rise in the coming decades. Together with the effects of thermal expansion of the ocean and mass loss from the large ice sheets, the glacier contribution to sea level rise represents a serious threat to humanity.

Local to regional hazards caused by glaciers in densely populated mountain ranges change with changes of climate and glaciers. While glacier advances destroying farmed land or damming dangerous lakes caused the main problems during the past colder centuries of the so-called Little Ice Age, shrinking ice since then alleviated these threats but now leads to new concerns. Increasing human activities related to traffic,
tourism, or hydropower production followed
the retreating ice, introducing new exposures
and risks. Numerous new lakes also started to
form and will continue to grow in deglaciat-
ing mountain landscapes. The probability of
far-reaching floods from impact waves caused by
large rock/ice avalanches is slowly but steadily
increasing with every new lake forming at the
foot of steep icy peaks.

**Definition and processes**

In simplest terms, glaciers are perennial masses
of ice, combined with layers of snow and firn,
originating on land and exhibiting evidence of
down-slope flow. Under equilibrium conditions,
glacier flow evacuates mass added by *accumulation*
(mainly snowfall) at higher/colder elevations
to lower/warmer areas, where *ablation* (mainly

---

**Figure 2** Advancing Mer de Glace (temperate valley glacier, left) and Mont Blanc (4810 m a.s.l.) with steep Glacier des Bossons (polythermal mountain glacier, right) during the Little Ice Age (probably around 1820) near Chamonix, French Alps. Undated painting by Jean-Antoine Linck ("La chaîne du Mont-Blanc vue de la Flégère"); signed at bottom center "Jn Ante Linck fec."); gouache; 62.0 × 85.5 cm; Musée d’art et d’histoire, Genève, Inv. Nr. 1915–1975). Photo H.J. Zumbühl. Reproduced with permission.
GLACIERS

melting) removes this added mass again and produces meltwater to leave the glacier. The *equilibrium line* represents points on a glacier where accumulation and ablation exactly balance each other. Changing climatic conditions – especially air temperature and precipitation – disturb this equilibrium and cause glaciers to grow or shrink.

The *mass balance* of a glacier is primarily determined by the relation between solid precipitation (snow accumulation) and the *energy balance* (ablation). Snow remaining at the surface during more than one year is called “fri.” Compression of fri under the weight of subsequently deposited snow and fri layers leads to gas exhaustion and densification until the air is trapped in closed bubbles when it reaches the close-off density (about 830–850 kg m$^{-3}$ in cold/dry fri and about 900 kg m$^{-3}$ in wet fri) at the transition to ice. Fri layers on mountain glaciers are usually a few tens of meters thick and the *snow-fri-ice metamorphosis* takes some decades to be completed.

The primary energy source for melting the ice is shortwave solar radiation. The amount of this energy absorbed by glaciers depends on the surface reflectivity (albedo), which is much higher for bright snow than for darker ice. Albedo reduction by shrinking snow cover or dust input from local to regional sources is a strong positive feedback of ongoing glacier vanishing trends. On the other hand, longwave radiation depends on the temperature of the emitting body and cannot change with a melting snow/ice surface at constant temperature. As a consequence, glaciers react to changes in incoming longwave radiation from a warming atmosphere by enhanced melting rather than by rising surface temperatures. Air temperature, with its strong vertical and temporal variability, plays a dominant role in the energy balance of snow and ice surfaces. As air temperature influences not only ablation but also accumulation through the solid/liquid precipitation limit, it has an overall dominant influence on glacier mass balance and is often used in a pragmatic/simplified form (positive degree days) in numerical glacier models (Ohmura 2001). The change of mass balance with altitude on a glacier, the mass balance gradient, determines the atmosphere/glacier coupling and drives long-term glacier movement.

*Glacier flow* involves internal ice deformation and basal sliding. Correspondingly, surface velocity ($u_s$) is composed of the basal sliding velocity ($u_b$) and the velocity due to internal ice deformation ($u_d$):

$$u_s = u_b + u_d$$

(1)

The ratio $u_b/u_s$ is called “velocity ratio.” For internal ice deformation, a power law is commonly applied:

$$\dot{e} = A\sigma^n$$

(2)

where $\dot{e} =$ deformation rate, $\sigma =$ applied stress, and $A$ and $n =$ flow law parameters.

The essential stress related to glacier flow is the shear stress at the glacier bed ($\tau_b$):

$$\tau_b = f\rho gh\sin\alpha$$

(3)

where $f =$ shape factor accounting for friction at valley walls, $\rho =$ ice density, $g =$ acceleration due to gravity, $\alpha =$ surface inclination, and $h =$ ice thickness. Ice thickness is coupled via the basal shear stress to surface inclination, a fact which can be used for ice-depth estimates for unmeasured glaciers. Using digital terrain information and relations between basal shear stresses and surface mass turnover as governed by elevation ranges and mass balance gradients, local ice thicknesses and detailed glacier-bed topographies can now be rapidly calculated for entire mountain ranges (Clarke et al. 2013; Linsbauer et al. 2012).
GLACIERS

No generally accepted “sliding law” exists – a fundamental and still unsolved problem in glacier mechanics. Numerical model simulations often use approaches like:

\[ u_b = \frac{C\tau_b^m}{(P_i - P_w)^p} \]  

where \( P_i \) = ice overburden pressure, \( P_w \) = water pressure at the glacier bed, \( C \) = frictional parameter, and \( m \) and \( p \) = constants. This reflects that basal sliding increases with increasing water pressure, a well-documented effect, which can relate to striking flow instabilities.

A spectacular but still not fully understood flow instability is the phenomenon of glacier surges, sudden acceleration of flow velocity often by an order of magnitude or more and sometimes rapid advances over a distance of several kilometers and at rates up to many tens of meters per day. Why many glaciers in some mountain ranges (Karakoram, Alaska Range, etc.) show this behavior but not glaciers in other mountain chains and why surging glaciers tend to repeat their surges after an intermittent quiescent time interval remain open questions even today. A phenomenon sharply contrasting with the surge advances is the calving instability of glaciers with rapid retreat in deep water after a much slower advance phase on top of a frontal moraine. High water pressures in such cases are due to the calving front being submersed in the sea (or a deep lake) and, during rapid retreat through deep water, can even reach flotation conditions, inducing massive iceberg discharge.

Distribution and characteristics of glaciers

First attempts to compile a global inventory of glaciers were undertaken in the second half of the twentieth century. Detailed quantitative information, however, could only be provided where aerial photographs were available. High-resolution satellite images made it possible to complete a new glacier inventory (Pfeffer et al. 2014) for the Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC) (IPCC 2014). In combination with the digital elevation models from the Shuttle Radar Topography Mission (SRTM)/Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Digital Elevation Model (ASTER GDEM) and a model approach coupling ice depth with surface slope via mass flux and shear stress, detailed and realistic information on numbers, areas, thicknesses, and volumes of all glaciers worldwide were derived (Table 1; Huss and Farinotti 2012). The total number of glaciers depends to some degree on the definition of the lower size limit – small perennial ice patches or glacierets without clear features of flow are often considered separately – but can be estimated reasonably well at some 170,000. The total glacier area as documented by the new global inventory is about 735,000 km\(^2\) and the best estimate for the total volume about 170,000 km\(^3\) or around 40 cm of sea level equivalent. Roughly 10% of this ice volume is, however, already below sea level or below the level of lakes forming with continued glacier vanishing and will not contribute to sea level rise (Haeberli and Linsbauer 2013). The largest volumes of glacier ice are found in the periphery of the Antarctic and Greenland ice sheets, the Canadian Arctic, Alaska, central Asia, and Arctic Russia. Glacier volumes in low-latitude high-mountain chains are generally smaller but can be especially important for people living there and in surrounding lowlands. Behind overall numbers of areas and volumes there is a great variability of glacier characteristics.

Local topographic conditions determine the morphological types of glaciers (Figures 3 and 4).
GLACIERS

Table 1 Glacier data for 19 regions worldwide. \( n \) = number of evaluated glaciers, \( S \) = total areas, \( h \) = average thicknesses, \( V \) = calculated volumes, \( SLE \) = sea level equivalent. Source: Reproduced with permission from Huss and Farinotti (2012). John Wiley & Sons, Ltd.

<table>
<thead>
<tr>
<th>Region</th>
<th>( n )</th>
<th>( S ) (km(^2))</th>
<th>( h ) (m)</th>
<th>( V ) (km(^3))</th>
<th>( SLE ) (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>22916</td>
<td>89 901</td>
<td>226</td>
<td>20 402</td>
<td>50.7</td>
</tr>
<tr>
<td>Antarctic and Subantarctic</td>
<td>3318</td>
<td>133 173</td>
<td>281</td>
<td>37 517</td>
<td>93.1</td>
</tr>
<tr>
<td>Arctic Canada, north</td>
<td>3205</td>
<td>105 139</td>
<td>327</td>
<td>34 399</td>
<td>85.4</td>
</tr>
<tr>
<td>Arctic Canada, south</td>
<td>6679</td>
<td>40 893</td>
<td>240</td>
<td>98 14</td>
<td>24.4</td>
</tr>
<tr>
<td>Caucasus</td>
<td>1335</td>
<td>1121</td>
<td>55</td>
<td>61</td>
<td>0.2</td>
</tr>
<tr>
<td>Central Asia</td>
<td>30 131</td>
<td>64 448</td>
<td>77</td>
<td>5026</td>
<td>12.5</td>
</tr>
<tr>
<td>Central Europe</td>
<td>3888</td>
<td>2060</td>
<td>56</td>
<td>117</td>
<td>0.3</td>
</tr>
<tr>
<td>Greenland periphery</td>
<td>13 860</td>
<td>87 765</td>
<td>216</td>
<td>19 042</td>
<td>47.3</td>
</tr>
<tr>
<td>Iceland</td>
<td>289</td>
<td>11 055</td>
<td>401</td>
<td>4441</td>
<td>11</td>
</tr>
<tr>
<td>Low latitudes</td>
<td>4979</td>
<td>4074</td>
<td>35</td>
<td>144</td>
<td>0.4</td>
</tr>
<tr>
<td>New Zealand</td>
<td>3002</td>
<td>1160</td>
<td>60</td>
<td>70</td>
<td>0.2</td>
</tr>
<tr>
<td>North Asia</td>
<td>3455</td>
<td>2816</td>
<td>49</td>
<td>140</td>
<td>0.3</td>
</tr>
<tr>
<td>Russian Arctic</td>
<td>353</td>
<td>51 665</td>
<td>325</td>
<td>16 839</td>
<td>41.8</td>
</tr>
<tr>
<td>Scandinavia</td>
<td>1795</td>
<td>2846</td>
<td>90</td>
<td>256</td>
<td>0.6</td>
</tr>
<tr>
<td>South Asia, east</td>
<td>13 615</td>
<td>21 699</td>
<td>60</td>
<td>1312</td>
<td>3.3</td>
</tr>
<tr>
<td>South Asia, west</td>
<td>22 563</td>
<td>33 961</td>
<td>95</td>
<td>3241</td>
<td>8</td>
</tr>
<tr>
<td>Southern Andes</td>
<td>19 089</td>
<td>32 521</td>
<td>205</td>
<td>6674</td>
<td>16.6</td>
</tr>
<tr>
<td>Svalbard</td>
<td>2058</td>
<td>33 932</td>
<td>285</td>
<td>9685</td>
<td>24</td>
</tr>
<tr>
<td>Western Canada</td>
<td>14516</td>
<td>14 615</td>
<td>70</td>
<td>1025</td>
<td>2.5</td>
</tr>
<tr>
<td>Total</td>
<td>171 046</td>
<td>734 856</td>
<td>231</td>
<td>170 214</td>
<td>422.6</td>
</tr>
</tbody>
</table>

Glacier ice can cover flat terrain as ice caps, occupy small cirques as cirque glaciers, cover slopes as mountain glaciers, or fill valleys as valley glaciers. Small hanging glaciers can be frozen onto slopes steeper than about 45°, in strong contrast to sometimes very large and flat glaciers reaching the ocean as tidewater glaciers.

Thermal glacier types are a function of average air temperature conditions at the surface, but temperatures within glaciers can be heavily influenced by ice flow (Wilson and Flowers 2013). Where mean annual air temperatures are below about −20°C, snow and firn of the accumulation area will remain cold and dry. With mean annual air temperatures increasing toward −10°C, firn will still remain cold, and small amounts of meltwater forming at the surface will refreeze and not leave the glaciers. Where mean annual air temperatures are higher than about −10°C, meltwater will completely warm up the firn to melting temperatures where solid, liquid, and gaseous phases are in equilibrium and can, therefore, leave the glaciers. The ice of ablation areas is more or less impermeable to
meltwater and therefore below melting temperature already with mean annual air temperatures around −3 to −4°C. As a consequence, the firn of accumulation areas may be warmer than the ice of ablation areas – a characteristic feature of many glaciers such as those on Svalbard. Glaciers with ice which remains everywhere below melting temperature are “cold glaciers,” glaciers with all ice at melting temperature are “temperate glaciers,” and glaciers containing both, cold as well as temperate firn/ice, are named “polythermal glaciers.”

In combination, regimes of mean annual air temperature and precipitation cause pronounced differences in the characteristics of maritime and continental-type glaciers. Where precipitation is abundant (meters per year), large amounts of snow must be melted. Mean annual air temperature at the equilibrium line of such maritime-type glaciers is therefore close to 0°C, and glaciers are mostly temperate, have a large mass turnover, and flow under relatively high shear stresses rapidly down to regions with seasonal frost only and rich vegetation, often
Figure 4  Glaciers in the Central Karakoram National Park, Pakistan. The almost 50-km-long, heavily debris-covered valley glacier in the center is Hispar Glacier which terminates (arrow) less than 2 km up-valley of the village of Hispar Nagar with its irrigated fields (see also similar irrigated fields at the village of Shimshal). Landsat ETM+ of 13 August 1998, RGB (4,3,2) channels, combined with void-filled SRTM version 4 (year 2000, resolution 90 m, 1.2 vertical exaggeration for oblique view). Vegetation is shown in red. Reproduced from GLOVIS-server of NASA and CGIAR-CSI website. © US Government.

forests. In regions of low precipitation (tenths of a meter per year), much less snow must melt, and the equilibrium line on such continental-type glaciers is therefore at altitudes with low mean annual air temperatures (< −5°C); glaciers are polythermal to cold, have small mass turnover, flow under lower shear stresses at lower speeds, and often end far above timberline in areas of widespread permafrost.

Glaciers at high latitudes are exposed to strong seasonal variations of atmospheric temperatures, causing distinct parts of the year with accumulation/freezing in winter and ablation/thawing in summer. Most of them are cold to polythermal and far away from human infrastructure. Extensive areas with high-latitude glaciers are in the Canadian and Russian Arctic. For mid-latitude glaciers, daily and seasonal temperature variations are both strong. Where precipitation falls during the entire year, as in the European or New Zealand Alps, accumulation mainly takes place in the cold season and ablation in the warm season. Under Monsoon influence, as, for instance, in the Tian Shan or Sikkim Himalaya mountains of central Asia, dry conditions coincide with the cold season, and accumulation and ablation both take place during the warm/wet time of the year. Large maritime-type glaciers exist in
Patagonia and Canada, while the dry regions of Tibet and the dry Chilean Andes contain smaller continental-type glaciers. Many high mountains in central Asia have transitional climatic conditions. Daily temperature variations predominate for glaciers in the central and outer tropics such as those on Kilimanjaro or in the Peruvian Cordilleras. There, ablation takes place during the entire year but accumulation only in the wet season; the equilibrium line is close to the \(0^\circ\text{C}\) air temperature isotherm and glaciers are temperate with the exception of the highest peaks with cold hanging glaciers frozen to steep permafrost rock walls.

**Internationally coordinated glacier observation**

In contrast to continental ice sheets, the enormous masses of which are *drivers in the climate system*, and in contrast to snow, sea ice, and permafrost, which, via their influence on surface albedo, ocean circulation, or greenhouse gases, are involved in multiple and important *feedback processes within the climate system*, glaciers and seasonal ice on lakes and rivers with their limited areas and volumes are *passive indicators of climate change*. They both primarily reflect atmospheric temperature changes but are complementary in that lake and river ice mostly relate to winter and lowland conditions while glaciers are more closely connected to summer and high-altitude effects. Striking traces left in the landscape by former glaciers such as moraines, polished rocks, or erratic boulders had indeed paved the way in the nineteenth century to the detection of past ice ages and to a modern understanding of natural climate changes. It was a logical step then to start monitoring fluctuations of existing glaciers in order to better understand past and possible future developments.

Following the example of the Swiss glacier-monitoring network initiated in 1893, *internationally coordinated glacier monitoring* formally started in 1894 with the foundation of the Commission Internationale des Glaciers (International Glacier Commission) and has built up since then one of the longest global environmental observation series. The early establishment of a systematic program of data collection and free dissemination is exemplary despite intermittent crises in the evolution over time. Two main scientific goals envisaged at the beginning of long-term observations were (i) to find physical causes for the dramatic climate excursion of the just detected Ice Age and (ii) to document whether climate and glacier changes were caused by terrestrial or extraterrestrial influences. A first phase of the activity was, therefore, characterized by the search for periodical oscillations (solar cycle) in the reported glacier length changes (advance and retreat; Figure 5). The two world wars and the economic crisis between them strongly reduced the program in a second phase of the 1930s and 1940s with the result that the scientific literature hardly noticed the drastic retreat of glaciers during these years. In a third phase beginning with the late 1940s, systematic glacier mass balance observations were begun in Scandinavia and soon also in other countries (Figure 6). These field measurements with ablation stakes and snow pits to determine annual or even seasonal mass gain/loss calibrated by precision mapping repeated at about 10-year intervals now captured the central element in the process chain linking changes in atmospheric conditions with the mass balance, thickness change, flow dynamics, and, finally, the easily observed advance/retreat of glacier margins (Figure 7). A fourth phase, starting in the 1970s, saw the compilation of the first worldwide glacier inventory and the foundation of the World Glacier Monitoring Service.
Figure 5  Cumulative length change of glaciers between high northern latitudes (Svalbard) and mid-latitudes of the Southern Hemisphere (Argentina) compiled by Leclercq and Oerlemans (2012) using long series of observations (mostly WGMS data) and reconstructions from various authors and data sources. The glacier retreat tendency since the end of the Little Ice Age around the mid-nineteenth century is strong and global. Reproduced under the CC-BY-NC license.

(WGMS). The fifth phase is characterized by the integration of glaciers as an Essential Climate Variable (ECV) within the Global Terrestrial Network for Glaciers (GTN-G) as part of the Global Terrestrial Observing System/Global Climate Observing System (GTOS/GCOS) in support of the United Nations Framework Convention on Climate Change (UNFCCC). This most recent phase of internationally coordinated long-term glacier monitoring was characterized by a rapid development of new technologies, especially high-resolution satellite measurements and digital terrain information.

A tiered observing strategy helps with bridging the gap between detailed local measurements and global coverage. On a small number of sites (about 10), detailed measurements about climate, mass/energy balance, flow, length change, and
Figure 6  Annual (a) and cumulative (b) mass balance of glaciers. The average rate of mass loss, expressed as thickness change in water equivalent, accelerated from 0.2 m per year for 1980–1990 to 0.4 m per year for 1990–2000 and to 0.8 m per year for 2000–2010. Reproduced from World Glacier Monitoring Service (WGMS), www.wgms.ch/.
Figure 7 Sketch illustrating characteristic features of mountain glaciers and the process chain linking climate and glacier response. C = accumulation area covered with snow and firn, A = ablation area (bare ice), EL = equilibrium line, LIA = moraines from glacier advances during the Holocene and the Little Ice Age, O = bed overdeepening beneath a flat, crevasse-free part of the glacier with compressing flow and possible site of a future new lake forming with continued glacier retreat.

runoff are carried out for process understanding and for the development of numerical models. Mass balance measurements at somewhat more than 100 glaciers provide a reference time series, primarily for the Northern Hemisphere. The question of how representative these measurements are at regional scales can be analyzed using length change measurements (especially concerning the past) and differences between digital terrain elevation models covering different times (especially for recent decades and the future). Glacier inventories repeated at time intervals of years to a few decades then enable coverage and change assessments at global scale.

Latest gravimetric and laser altimetry technologies such as the Gravity Recovery and Climate Experiment (GRACE) or the Ice, Cloud, and land Elevation Satellite (ICESat) make it possible to detect significant mass changes of great numbers of glaciers over time periods as short as a few years (Figure 8; Gardner et al. 2013). Automated image processing and glacier parameterization have the potential to provide millions of mass and length change data, requiring new formats for dimensions of data far beyond past requirements. However, long time series of mass balance have been carried out on logistically treatable small glaciers, which are shrinking or even disintegrating, may completely disappear within coming decades, and must be replaced in time by larger glaciers with a longer potential lifetime. Most glaciers are far out of equilibrium, and their length and area now reflect climatic conditions of the past decades.
Figure 8  Regional glacier mass budgets and areas 2003–2009 from satellite gravimetry and altimetry and from local glaciological records. Red circles show 2003–2009 regional glacier mass budgets, and pale blue/green circles show regional glacier areas with tidewater basin fractions (the extent of ice flowing to termini in the ocean) in blue shading. Peach-colored halos surrounding red circles show the 95% confidence intervals in mass change estimates, but can only be seen in regions that have large uncertainties. From Gardner et al. (2013). Reprinted by permission of AAAS.

rather than the present (Mernild et al. 2013). Besides such decoupling in time of glacier length change from changing atmospheric conditions, increasing debris cover, downwasting rather than retreat, lake formation, disintegration, and even collapse phenomena make definitions of ice margins difficult. The GTN-G not only combines field measurements (mainly WGMS), satellite observations (the Global Land Ice Measurement from Space (GLIMS) initiative), and data processing/management (the National Snow and Ice Data Center (NSIDC), Boulder, Colorado), but also continuously keeps scientific reflection with experts alive to deal with such challenges.

A future world without glaciers?

Since the beginning of coordinated worldwide observations at the end of the nineteenth century, strong overall glacier retreat and mass loss have been documented, providing strong evidence that climate change takes place at a rapid rate and at a global scale. The reduction in glacier length still remains key information concerning past developments. With few regional exceptions (Kääb et al. 2012), and despite small readvances around 1920 and 1960/70, glaciers worldwide retreated from extended positions reached during the Little Ice Age, which in many regions had come to an end around the middle of the nineteenth century (Figure 5). Rising
global temperatures are certainly the primary cause of this evolution. Accelerated warming and melting since about 1980 are reflected in the growing rates of glacier mass losses beyond historical precedents (Figure 6; Zemp et al. 2015). Glaciers in many mountain ranges may now be as small as, if not smaller than, during the warmest periods of the Holocene – the warm/high-energy limit of preindustrial climate variability since the end of the last ice age appears to have been reached or already passed.

Analysis by WGMS of glacier fluctuations and inventory data with a view to possible future glacier changes already in the 1990s showed that entire mountain ranges like the European Alps may lose most of their glacier volumes even in the case of moderate scenarios of global warming. More sophisticated approaches (Zemp et al. 2006 and various later studies) confirmed this and provided more detailed assessments. While some of the highest mountain peaks may keep their ice cover well into the future, many glaciers, especially the large, flat, and thick valley glaciers with their low-elevation beds, will disappear. With them, the largest glacier areas and volumes, and, hence, the most important ice storage part in the water cycle, are likely to vanish over the coming decades. Besides problems with the parameterization of surface effects (mass balance) and flow (deformation and sliding) of glaciers, uncertainties of corresponding numerical model simulations primarily relate to: (i) different possible climate scenarios, (ii) the still existing ice volumes, (iii) the evolution of the albedo under the influence of dust input, (iv) lake formation, and (v) a number of other minor effects. The results from simple as well as complex model calculations are nevertheless robust and show that presently existing glacier landscapes will rapidly turn into new landscapes of rocks, debris, sparse vegetation, and numerous lakes. The new lakes forming in characteristic overdendments of glacier beds (Cook and Swift 2012) are of special interest in view of their hazard potential but also for hydropower production, water supplies, and tourist attractions. A large number of lakes have indeed already formed and some of them have caused outbursts with severe damage to humans and infrastructure (Carey et al. 2012).

Fundamental changes have taken place in our perception of glaciers and their changes. Concern is growing that the ongoing trend of worldwide and fast, if not accelerating, glacier shrinkage at the century timescale is of noncyclic nature – clearly different from the originally envisaged periodical oscillations. Under the growing influence of human impacts on the climate system (enhanced greenhouse effect), dramatic scenarios of future developments – including complete deglaciation of entire mountain ranges – must be taken into consideration. A broad and worldwide public today recognizes glacier changes as a key indication of regional and global climate and environment change. Observational strategies established by expert groups within international monitoring programs build on advanced process understanding and include extreme perspectives. These strategies make use of the fast development of new technologies and relate them to traditional approaches in order to apply integrated, multilevel concepts (in situ measurements to remote sensing, local-process oriented to regional and global coverage), within which individual observational components (length, area, volume/mass change) fit together, enabling comprehensive views. Such comprehensive views are especially important with respect to future developments, which may lead far beyond the range of historical/Holocene variability and most likely introduce processes (extent and rate of glacier vanishing, distance to equilibrium conditions) without precedent in the Holocene.
Assessing the impacts of vanishing glaciers and preparing adaptation/mitigation strategies constitute increasingly urgent tasks in view of a world in which little remains of those glaciers that still exist.

SEE ALSO: Climate change and land ice; Glacial depositional processes and landforms; Glacial erosional processes and landforms; Glaciations; Glacier changes; Glacier hydrology and runoff; Glacier inventory; Glacier lake outburst floods; Glacier mass balance; Glacier modeling

References


Haeberli, W., and A. Linsbauer. 2013. “Global Glacier Volumes and Sea Level – Small but Systematic Effects of Ice Below the Surface of the Ocean and of New Local Lakes on Land.” The Cryosphere, 7: 817–821. DOI:10.5194/tc-7-817-2013


Further reading


Websites

Glacier and Permafrost Hazards in Mountains (GAP-HAZ): http://gaphaz.org/

Global Land Ice Measurements from Space (GLIMS): http://www.glims.org/

Global Terrestrial Network for Glaciers (GTN-G): http://www.gtn-g.org/

World Glacier Monitoring Service (WGMS): http://www.wgms.ch/
Global cities

John Rennie Short
University of Maryland, Baltimore County, USA

In 1915 the Scottish polymath Patrick Geddes noted that much of the world's business is conducted in what he termed “world cities.” Rather than countries or regions, it was cities, he implied, that were the connecting points of a globalizing economy. His remarks did not inaugurate a body of work and the idea languished, to be resuscitated only in the 1960s by Peter Hall. He used the term “world cities” to describe London, Paris, Randstad, Rhine-Ruhr, Moscow, New York, and Tokyo (Hall 1966).

The idea of Geddes and Hall did not immediately become the basis for scholarly research because economic geography was concerned with regional complexes and national economies. The main emphasis of urban geography was on urban hierarchies at the regional and national level or on studies of individual cities. The connection between the global economy and the city was not made until the 1980s when the idea of globalization came to the fore. The terms “world cities” and “global cities” were used interchangeably although, more recently, the latter designation is more commonly employed. These were identified as cities that were important hubs in a global urban network and important command centers for the organization and management of a globalizing economy.

It was Friedmann and Wolff (1982) who first connected patterns of urbanization to the internationalization of capital. They identified a global urban hierarchy with world cities at the apex, characterizing them as the control centers of the global economy with a concentration of producer services, housing a highly mobile, transnational elite, and the sites of massive economic, social, and physical reorganization. They identified the following as world cities: Tokyo, Los Angeles, San Francisco, New York, London, Paris, Randstad, Frankfurt, Zurich, Cairo, Bangkok, Singapore, Hong Kong, Mexico City, and São Paulo. These world cities were asserted rather than demonstrated. In a later paper, Friedmann (1986) identified Tokyo, Los Angeles, Chicago, New York, London, Paris, Zurich, Rotterdam, Zurich, and São Paulo as the first-order centers of a global urban hierarchy, a similar but not identical list to their previous one. The authors drew attention to the global city as command center and to the global network of cities, but their loose definition of these cities, based on simple assertion rather than careful documentation, became a hallmark of many subsequent studies in what has been termed the “dirty little secret” of early global cities research.

Global urban networks

Waves of globalization are often associated with the creation of global urban networks. Merchant trading cities were nodes in an emerging global trade. We have a visual record of this network in the compilations of city maps and prospects in the first city atlas, Civitates Orbis Terrarum, by Georg Braun and Frans Hogenberg. One volume was published in 1572 but it became so popular that by 1617 the work consisted of six volumes with over 363 urban views. Collectively, the images indicate the increasingly global reach.
of mercantile capitalism and European colonization, forging a world economy in such urban centers as Aden, Peking, Cuzco, Goa, Mombasa, and Tangiers, as well as many other cities around the world. While the cities in the Civitates are depicted separately, the effect of the compilation is to reveal a global economy of urban nodes and a trading world of connected cities.

Today we can identify a global urban network connected by flows of goods, people, capital, and ideas. Some of the flows have been identified and measured. A substantial body of material has emerged from the work of Peter Taylor and his colleagues at the Globalization and World Cities (GaWC) research network. The website (http://www.lboro.ac.uk/gawc/) lists data sets as well as more than 450 research papers and is an indispensable guide to the metageography of urban networks. In 2000, they collected data on the distribution of 100 global advanced producer service firms, which includes accountancy, advertising, banking/finance, insurance, law, and management consultancy, across 315 cities. They analyzed the resultant data matrix to identify a global urban hierarchy. In 2012, they extended the analysis to 176 firms in 525 cities. The result was a fivefold hierarchy that identified cities as Alpha, Beta, Gamma, High sufficiency, and Sufficiency. New York (NY) and London (LON) dominate with a classification of Alpha ++ (Table 1). NYLON is an important pivot in the global networks, and its dominance reflects the historical legacy of London as the center of British Empire and the continuing importance of New York as the financial center of America’s more informal empire. Their data also hint at the rapid rise of new cities into the top tier as national economies make their way from periphery to core. In 2000, Shanghai was in the Alpha − category while Beijing was only a Beta city. In 2012, both Shanghai and Beijing were classified as Alpha +, only one step below NYLON. Both cities are moving into the top tier as China’s economic growth, both absolute and relative to the rest of the world, continues apace.

While the aggregate analyses are immensely useful, they also have their biases. The cities of the Global North tend to be privileged sites for investigation. Intensive fieldwork in less privileged cities reveals a different story, as in the detailed case study of Accra by Richard Grant (2009). As the capital of the West African nation of Ghana, Accra does not figure highly in the standard measures of globalization. On the GaWC classification it is listed Gamma −, well down the list. At first blush then, just another city in remote Africa, largely cut off from global flows. Grant presents a very different picture. Foreign companies now play an important role in the urban economy, constituting one third of all firms. The urban form itself was transformed. A new foreign-orientated central business district was established close to the airport. The city’s economy is fueled by remittances from Ghanaians abroad. Returning home, these repatriates often play key roles in transnational operations and flows. But even those stuck in the city also have global connections. Grant shows that shantytown dwellers are also connected to transnational nongovernmental organizations.

### Table 1 At the top of the GaWC global hierarchy.

<table>
<thead>
<tr>
<th>Alpha ++</th>
<th>London, New York</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha +</td>
<td>Hong Kong, Paris, Singapore, Shanghai, Tokyo, Beijing, Sydney, Dubai</td>
</tr>
<tr>
<td>Alpha</td>
<td>Chicago, Mumbai, Milan, Moscow, São Paulo, Frankfurt, Toronto, Los Angeles, Madrid, Mexico City, Amsterdam, Kuala Lumpur, Brussels</td>
</tr>
</tbody>
</table>
working with them to frame their demands and need. This detailed case study reveals the pervasive and multiple connections with the global space of flows of people, money, discourses, practices, and ideas. Accra is revealed as a networked city with globalization being imposed as well as embraced, transforming from above as well as changing from below. No African cities make the top tier of GaWC cities and yet, as the Grant study shows, Accra is a globalizing city firmly connected to global flows.

Lisa Benton-Short and colleagues also sought to identify a global network, but their work was based on flows of people. They looked at immigration into cities around the world and established an index based on the percentage of foreign-born, the total number of foreign-born, the percentage of foreign-born not from a neighboring country, and the diversity of immigrant home. The result was a threefold division into Alpha, Beta, and Gamma cities. The Alpha immigrant cities are New York, Toronto, Dubai, Los Angeles, London, Sydney, Miami, Melbourne, Amsterdam, and Vancouver. The Beta-ranked cities include Riyadh, Geneva, Paris, Tel Aviv, Montreal, Washington, DC, The Hague, Kiev, San Francisco, and Perth in Australia. They identify established gateway cities, unrecognized gateway cities, accidental gateway cities, and bypassed gateway cities. Examples of different types are listed in Table 2. Unrecognized gateways are particular strong in the former Eastern Bloc, and bypassed gateways, globalized cities with relatively low levels of foreign immigration, are common in Asian cities. There are similarities and differences between the GaWC and the Benton-Short et al. results. Confining comments to the top level and looking first at similarities, some cities appear in the same category in both analyses. New York and London are Alpha cities and sit atop the apex of both hierarchies. Other Alpha cities in both studies include Toronto, Los Angeles, Sydney, and Amsterdam. Looking at differences, while Miami, Melbourne, Vancouver, and Dubai are considered pivotal points of global migration reaching Alpha status, in the advanced producer service category they only make it to Beta status.

Some of the differences relate to issues of data and data availability, but they also refer to the nature of the global urban network. Networks vary according to the flow. Some flows “pool” in some cities rather than others. The flow of people, while matching the connections in command functions, also has slight differences. National regulations concerning immigration, the demand for labor, and the relative openness of societies to foreign migrants all play a part. The sheer need for labor, whether the global talent pool of specialized knowledge experts or cheap unskilled labor, varies throughout the urban network. The result is a similar but not
There is no one fixed global urban hierarchy but more of a “space of transformative flows.” Money, people, ideas, goods, and practices do not just flow through the global urban network; they are transformed in the process. Consider flows of people. The movement through the urban network is not a simple geographical movement; it involves cultural exchanges. These can refer to the new work habits and job practices of the temporary worker as well as the complex cultural transformations of long-term migrants as they adapt to new milieu and in turn transform their surroundings. Rather than mere transfers, flows along the urban networks are transformative experiences, even in the flows of inanimate things such as money, goods, and services. A small amount of money in London or New York when remitted back to Ghana or El Salvador can become the source for land and house purchase, enabling a new business, or paying for better schooling. Flows through the global urban network change the medium and the networks; as people adapt, ideas are tweaked, practices mutate, money is reimagined, and commodities are reappropriated.

To show the complexity of global city networks we can consider briefly three other types of flows. First, there is the general movement of knowledges and expertise. Contemporary urban polices such as neoliberalism, for example, move quickly as they are transferred around the cities of the world, recreated as they move, adapting to and undermining national and local practices. Urban policies are shaped and shifted as they travel and mutate through the global urban network of policy adoption and adoptions. Second, there is the burgeoning interurban circuit of contemporary art, comprising transnational biennial exhibitions and art fairs. Cities barely recognized as major urban, much less art, centers a decade ago – Luanda, Sharjha, Guangzhou, for example – ring increasingly familiar as sites of contemporary art display and debate (and art-making), gaining degrees of cosmopolitanism in the process. Now major exhibitions of contemporary art (biennials and triennials geometrically increase every year) similarly function as dazzling vehicles for global aspirations. Third, there is the franchising of culture and education through the global urban network. The Guggenheim Museum now has “branch” museums in Abu Dhabi, Bilbao, and Venice. Higher education institutes are selling their name and creating linkages and feeder institutions across the globe. Yale has established a branch campus in Singapore while New York University (NYU) has perhaps the most ambitious global presence with degree-granting institutions such as NYU Shanghai and NYU Abu Dhabi and centers in Accra, Berlin, Buenos Aires, Florence, London, Madrid, Paris, Prague, Sydney, and Tel Aviv. The liberal values of a traditional US university have to be tweaked to operate in such places as Abu Dhabi, Shanghai, and Singapore. Ideas and practices mutate rather than simply diffuse through the global urban networks.

While most studies concentrate on identifying the central hubs of global urban networks, a few have begun to consider the idea of black holes and loose connections in the global urban network. One study, by Short (2004), identified black holes, 11 cities which met three criteria: they had a population of over 3 million, were not identified by GaWC as a world city, and did not share their national territory with a world city (Table 3). They ranged from Tehran, with a population of 10.7 million, to Chittagong, with a population of 3.1 million. There are a number of reasons behind these very large cities’ non-world city status. Some cities, despite their size, are so poor that they do not represent a market for advanced producer services. They are the
black holes of advanced global capitalism, with many people but not enough affluent consumers or complex industries to support sophisticated producer services. There are not only cases of endemic poverty, but also cases of catastrophic decline where there has been an almost complete collapse of civil society. In recent years, Khartoum and Kinshasa, for example, have witnessed the decline of the rule of law and descent into social anarchy. War and social unrest have been the norm rather than the exception. Sustained social disruption reinforces the global disconnect. Cities are bypassed as forms of risk aversion by capital investors. The perception of risk can be self-fulfilling. High risk puts corporations off from establishing connections, which in turn increases the risk rating.

Command centers

Global cities are centers for coordination and control, sites of production of specialized services, especially producer services such as accountancy, financial services, and consultancies. A central reason for the concentration of command functions in selected global cities is the need for social interaction in global financial business deals. Trust, contact networks, and social relations play pivotal roles in the smooth functioning of global business. Spatial propinquity allows these relations to be easily maintained, lubricated, and sustained. Global cities are the sites of dense networks of interpersonal contact and centers of the important business/social capital vital to the successful operation of international finance. Global service corporations are adept at producing their own commodities, including new financial products, new advertising packages, and new forms of multijurisdictional law. All of these depend upon specialized knowledge. The coalescence of a range of expertise produces state-of-the-art commodities to meet the specific needs of clients. In order to be able to put together such packages, firms need to be in knowledge-rich environments. Face-to-face contacts between experts are facilitated by the clustering of knowledge-rich individuals in cities like New York and London. Cities at the apex of this particular network are “privileged sites” housing the “knowledge elite” that enact the economic reflexivity crucial for economic success. Reflexivity and networking are at the heart of understanding global cities as places where people, institutions, and epistemic communities work to establish and maintain contacts. More importantly, these communities act as crucial mediators and translators of the flows of knowledge, capital, people, and goods that circulate in the world.

The globalizing economy creates lots of information, narrative uncertainty, and economic risk that must be produced, managed, narrated, explained, and acted upon. Global cities are epistemic communities of surveillance, knowledge production, and storytelling. Studies of cities as

---

**Table 3** Black holes.

<table>
<thead>
<tr>
<th>City</th>
<th>Country</th>
<th>Population</th>
<th>Risk rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tehran</td>
<td>Iran</td>
<td>10,700,000</td>
<td>High</td>
</tr>
<tr>
<td>Dhaka</td>
<td>Bangladesh</td>
<td>9,950,000</td>
<td>Significant</td>
</tr>
<tr>
<td>Khartoum</td>
<td>Sudan</td>
<td>7,300,000</td>
<td>Very high</td>
</tr>
<tr>
<td>Kinshasa</td>
<td>Congo</td>
<td>6,500,000</td>
<td>Extreme</td>
</tr>
<tr>
<td>Lahore</td>
<td>Pakistan</td>
<td>5,500,000</td>
<td>High</td>
</tr>
<tr>
<td>Baghdad</td>
<td>Iraq</td>
<td>4,950,000</td>
<td>Extreme</td>
</tr>
<tr>
<td>Rangoon</td>
<td>Myanmar</td>
<td>4,650,000</td>
<td>High</td>
</tr>
<tr>
<td>Algiers</td>
<td>Algeria</td>
<td>3,950,000</td>
<td>High</td>
</tr>
<tr>
<td>Abidjan</td>
<td>Ivory Coast</td>
<td>3,850,000</td>
<td>Significant</td>
</tr>
<tr>
<td>Pyongyang</td>
<td>North Korea</td>
<td>3,550,000</td>
<td>Very high</td>
</tr>
<tr>
<td>Chittagong</td>
<td>Bangladesh</td>
<td>3,100,000</td>
<td>Significant</td>
</tr>
</tbody>
</table>

*Risk rating from World Markets Country Analysis. Source: Adapted from Short (2004).*
economic command centers identify a number of key characteristics: they have articulated infrastructures of market trading involving an agglomeration of demand and supply, an environment of innovation, business support services, pools of highly skilled labor, a proximity of business organizations allowing information to be generated, analyzed, and disseminated, and deals to be struck, all set within shared cultures of expertise and of contacts set within an even broader context of a disciplined market, multicurrency trading, and a responsive central bank and governmental system of economic regulation.

Global cities operate as major nodes of reflexivity in global networks through the migration of business elites. As the contemporary international service economy requires specialist professionals to be globally mobile to deliver intelligence, skills, and knowledge to the point of demand, the development of a cross-border transnational migrant elite contributes to the production and consumption of the global city. In the world of high finance and international corporate law it is essential for international companies providing these services to maintain a tailored personal relationship, built on mutual trust, with clients around the world. And this requires global firms to maintain offices in cities around the world staffed with highly skilled professionals. These elite workers move through the distribution of corporate offices to pass on the corporate culture, to gain knowledge of local markets and work with specific clients; their career paths trace routeways through the global urban hierarchy.

A large proportion of command functions are still concentrated in just a few cities, but rather than a model of competition between global cities for global dominance, it is more accurate to consider a network of global cities with economic control concentrated in only a few cities but with the operation of this power dispersed in the second and third tiers of the global city hierarchy through branch plants and back offices. The global urban hierarchy is a network of flows that transmits global command functions across the world.

**Polarization**

Increasing social and spatial polarization is a feature of the large global city. Chris Paris (2013) describes the process of decoupling where the housing markets of global cities, because they attract the rich and superrich, are more linked to prime real estate neighborhoods in other global cities than to the national housing market. The result is a decoupling of housing prices from national markets effect. Lower-income households are squeezed out by rising house prices and gentrification is common. Rising inequality is a by-product of a globalizing city. Social and spatial polarization is not accidental but integral to the making of the global city as they house both the super-rich and the wealthy elites, as well as the marginal and the poor.

**Becoming a global city**

One important feature of global city research has been the search for global cityness. Many studies have been devoted to identifying whether this or that city is a global city. While the work is interesting up to a point, it tends to focus on a narrow range of cities at the top end of the urban hierarchy. This focus tends to ignore how globalization is acting in and through all cities. A more useful term is “globalizing city” to refer to the fact that almost all cities can act as a gateway for the transmission of economic, political, and cultural globalization.

The focus on globalizing city as opposed to global city shifts the attention away from the
question of which cities dominate the global urban network to show all cities in the network are affected by globalization. It also focuses attention on the making of global cities and the promotion of global city status.

Global city status requires the attributes of a global city. Such prerequisites include an international airport, signature buildings of big name architects, impressive buildings (the most recently constructed tallest buildings in the world are in Kuala Lumpur, Shanghai, Hong Kong, and Taipei), and cultural complexes such as art spaces and symphony halls. Combining these elements is always a useful strategy: hiring a famous architect to design a cultural complex, as in James Stirling’s Neue Staatsgalerie in Stuttgart, Isozaki’s Museum of Contemporary Art in Los Angeles, Pei’s glass pyramid in front of the Louvre in Paris, or, one of the oldest yet still impressive, Jorn Utzon’s Opera House in Sydney. The construction of these complexes involves urban redevelopment schemes through which local developers, landowners, and politicians amass substantial fortunes. Chang, Huang, and Savage (2004) consider the case of Singapore’s attempts to create urban landscapes that project the city-state’s global aspirations. The Singapore River Development Zone Project is intended to project global city status and successfully connect with global flows of capital movement and tourist flows. Both global and local forces are at work and the nature of the development is a negotiated outcome between the perceived need to make global connections with local empowerment and community rights all mediated through debates about national identity.

The city of Shanghai, for example, has embarked on an ambitious goal of projecting global city status. Since the early 1990s, improvements have been made to the physical infrastructure, including the building of a shiny new downtown of high-rise towers and the construction of a new international airport, land-use planning measures have been introduced to encourage the dispersal of industry from central sites and the development of mixed commercial and residential use of prime land, and organizational streamlining has been inaugurated to improve urban management. The project is to make a global city that sits alongside New York, London, and Tokyo. A self-consciously global city is being created and represented.

An important part of seeking and maintaining global city status is the hosting of global events such as the Olympics, World Cup, or other high-profile event that involves an opportunity not only to remake the city as a global city but to represent it as a global city. Hosting these events is not just an opportunity to be the site of a global spectacle, and hence to receive international name recognition, but also an opportunity for business and real estate deals, acting as giant urban redevelopment projects. Hosting the Summer Olympics, for example, involves a massive urban transformation including the provision of venues and major infrastructural investment. And it allows urban makeovers designed to create a modern global city defined in terms of a firmer insertion into international circuits of capital flows and the construction of modern urban forms. Hosting the Games provides an opportunity for city reimaginings as well as remakings. The changes often come with increased costs for city residents unless there is a specific commitment to redistributional outcomes.

The globalizing project varies in detail by individual city but overall there are recurring features across the world, including the reimagining of the city, the rewriting of the city for both internal and external audiences, the construction of new spaces, and the hosting of new events. A cosmopolitan lifestyle is also promoted as part of the project, complete with settings and performance that synergize the four c’s of culture,
GLOBAL CITIES

consumption, cool, and cosmopolitan. The globalizing project also involves a spatial reorientation of the city, the spectacularization of settings, the creation of specifically global sites, and the encouragement of transnational locations.

Global imagineering

In a globalizing world, the marketing of a city as a global city is now an important strategy. Because of the growing competition, there is a constant need to upgrade and improve a city’s image. From 1997 to 2007, Hong Kong was specifically marketed as “Asia’s World City,” a response in part to growing competition from Shanghai. Darel Paul (2004, 2005) explores the politics of “global imagineering” in Montreal, Canada, where business interests triumphed and Minneapolis–St Paul, USA, where similar interests were trumped by populist politics. In St Petersburg, Russia, not only was the old Soviet name of Leningrad dropped, but the city was actively promoted as an international hub of circulatory capital, host to corporate power, and stage for globalist megaprojects. St Petersburg was rebranded as an entrepreneurial, economically competitive, globally connected city (Golubchikov 2010). Kim and Short (2008) note that the rhetoric of global city branding is often associated with attempts to justify controversial and costly urban redevelopment projects, to justify tax incentives and business-friendly economic policies, and to heighten multiculturalism and cosmopolitanism in the city.

Cosmopolitan global cities

Global cities tend to have more foreign-born than their national territory. Table 4 lists the number of foreign-born in selected cities from the Benton et al. study that highlights the reliance of Arab cities on foreign labor, the lack of foreign-born in Asian cities, and the important immigration role of Canadian cities. The cozy and reassuring notion that immigration leads to greater cosmopolitanism, for example, is undermined by the experience of many cities in the Middle East where very high rates of immigration into cities are marked sometimes by cultural xenophobia and religious fundamentalism (not cosmopolitanism). There is no one-to-one relationship between foreign-born in the city and the cosmopolitanism of the city, and the disparity is especially marked even in cities such as Singapore where there are high levels of foreign-born female domestic workers.

Global cities with large numbers of foreign-born are hubs of diasporic convergence held together by shared identity but separated by thousands of miles. These new diasporic communities can enrich and enliven global city cultures. Some global cities are emerging as both points of cosmopolitanism where new hybrid identities and cultures are emerging and

<table>
<thead>
<tr>
<th>City</th>
<th>Foreign-born as percent of total city population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dubai</td>
<td>82.0</td>
</tr>
<tr>
<td>Miami</td>
<td>50.9</td>
</tr>
<tr>
<td>Amsterdam</td>
<td>47.2</td>
</tr>
<tr>
<td>Toronto</td>
<td>46.9</td>
</tr>
<tr>
<td>Muscat</td>
<td>44.6</td>
</tr>
<tr>
<td>Vancouver</td>
<td>39.0</td>
</tr>
<tr>
<td>Auckland</td>
<td>39.0</td>
</tr>
<tr>
<td>Geneva</td>
<td>38.3</td>
</tr>
<tr>
<td>Mecca</td>
<td>37.5</td>
</tr>
<tr>
<td>The Hague</td>
<td>36.5</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>36.2</td>
</tr>
<tr>
<td>Tel Aviv</td>
<td>36.0</td>
</tr>
</tbody>
</table>

Source: Adapted from Benton-Short et al. (2005).
as places of resistant local diasporic identities. These diasporic communities are also conduits of foreign trade as family connections merge into business connections, and small import/export companies arise from family and ethnic ties. Foreign trade is not just between giant multinational corporations; it also takes the form of smaller-scale linkages between communities. Immigrants play a key role in the economic development of the city in circulating money from rich to poor places and in identifying possibilities for international trade and exchange.

SEE ALSO: Globalization; Urban elites; Urban geography; Urban redevelopment

References


Further reading


Global climate change

David F. Porinchu
University of Georgia, USA

Global climate change, which is driven by human activity and natural variability, refers to the change in the mean state and/or variability of the climate system at decadal timescales and includes changes in temperature, precipitation, and atmospheric circulation patterns (IPCC 2013). Observations of recent global-scale changes in the climate system can be placed into a broader temporal context through comparison with the geologic record, which in turn can provide insight into sensitivity of the climate system to various natural and anthropogenic forcings and serve to provide baselines against which projections of future climates can be compared. Knowledge of global climate change requires consideration and associated impacts. Documenting the causes and consequences of global climate change requires knowledge of the spatial and temporal patterns and the biophysical processes responsible for driving past, present, and potential future climate change and climate change-related impacts.

Observed changes in the global climate system

Global-scale observations of the climate system, based on instrumental records and remote sensing, provide detailed records of changes in the climate system that extend from the mid-nineteenth century to the present. These direct measurements document the response of the atmosphere, oceans, sea level, cryosphere, and land surface to recent climate change at seasonal, annual, and decadal timescales. The observed globally averaged combined land surface and ocean surface temperature anomaly for the interval between 1880 CE and 2012 is 0.85 °C, with nearly the entire surface of the Earth experiencing warming during the twentieth and early twenty-first centuries. Fourteen of the fifteen warmest years on record have occurred during the twenty-first century (see Figure 1), with a strong likelihood that 2014 is the warmest year since 1880 CE. In addition, the diurnal temperature range (defined as the difference between daily maximum and daily minimum temperatures) has decreased during the latter half of the twentieth century and the incidence of extreme weather events (heatwaves, heavy precipitation, droughts) has increased for the majority of Earth’s surface during recent decades. Approximately 60% of the net energy increase in the climate system between 1971 and 2010 CE has been stored in the upper 700 m of the ocean, resulting in the temperature of the upper 75 m of ocean increasing at a rate of 0.11 °C/decade during this interval. Ice sheets, alpine glaciers, and sea ice have experienced notable reductions in extent and/or volume since the 1970s. The average rate of ice loss from Greenland and Antarctic ice sheets has increased from 34 to 215 gigatons (Gt) year⁻¹ and from 30 to 147 Gt year⁻¹, respectively, between 1992 and 2001 CE and 2002 and 2011 CE. The extent of the summer sea ice minimum in the Arctic Ocean decreased at a rate of between 9.4% and 13.6% per decade between 1979 and 2012 CE. Satellite, tide gauge, and documentary data
Figure 1 The annual trend in average global air temperature through December 2013. For each year, the red vertical bars indicate the range of uncertainty. The blue line tracks the changes in the trend over time (NOAA 2014).

indicate that the mean rate of globally average sea level rise was 1.7 mm year\(^{-1}\) between 1901 and 2010 CE. The rate of sea level rise has been increasing in recent decades, with the rate nearly doubling to 3.2 mm year\(^{-1}\) between 1993 and 2010 CE. The observed rise in sea level is largely attributable to glacier mass loss and the thermal expansion of ocean water. Other components of the Earth’s biophysical environment that have been influenced by global climate change include river discharge, permafrost temperature, and biogeochemical cycles. Some of these changes have the potential to further amplify warming. For example, elevated temperatures in high northern latitudes have accelerated rates of thaw and aerobic and anaerobic decomposition in permafrost, facilitating the increased release of carbon dioxide and methane, which in turn acts to further augment radiative forcing (RF) (Hodgkins et al. 2014). It is important to note that methane is approximately 20 times more potent than CO\(_2\) in terms of radiative forcing over a 100-year period, and therefore has the potential to greatly accelerate warming.

Drivers and attribution of recent climate change

Changes in the concentrations of greenhouse gases, aerosols, land cover, and solar radiation influence the Earth’s energy balance and are considered to be the major drivers of observed climate change. Total anthropogenic radiative forcing in 2011 CE is 2.29 W m\(^{-2}\) relative to preindustrial (1750 CE) conditions. Currently, the most significant driver of recent climate change, in terms of radiative forcing, is the increase in the concentration of greenhouse gases, the most important of which is carbon dioxide (CO\(_2\)). The concentration of CO\(_2\), which at the end of 2014 CE stood at 396 parts per million (ppm),
is approximately 40% greater than its preindustrial concentration of 280 ppm. The current concentrations of the other major greenhouse gases, methane (CH₄) and nitrous oxide (N₂O), are 1893 parts per billion (ppb) and 326 ppb, and 150% and 20% greater than preindustrial levels, respectively. The well-mixed greenhouse gases (CO₂, CH₄, N₂O, and halocarbons) are responsible for an increase in radiative forcing of 2.83 W m⁻² relative to 1750 CE (see Figure 2). It is important to note that ice core records indicate that the current concentrations of CO₂, CH₄, and N₂O are higher than at any time during the last 800,000 years, and exceed natural variability. The rate of increase in these greenhouse gases in recent decades is unmatched in the last 22 ka (IPCC 2013). An increase in planetary albedo, due to altered land use, increased emission of aerosols, and aerosol–cloud interactions, which act to counteract the influence of greenhouse gases on global mean radiative forcing, have limited the rate and magnitude of recent
GLOBAL CLIMATE CHANGE

warming. Attribution studies have identified that anthropogenic activities are responsible for more than 50% of the observed increase in global mean surface temperature and have substantially contributed to the observed increase in global mean sea level, ocean acidification, and a reduction in Arctic sea-ice extent since the 1970s.

Past climates: placing recent climate change in context

The instrumental record (<150 years) is too short to adequately document the full range of climatic behavior and variability. Extending the observed record further back in time through the use of natural archives such as ice cores, marine and lake sediments, corals, tree rings, speleothems, and glacial deposits places observed changes in the global climate system during the twentieth and twenty-first centuries into a broader temporal context and provides evidence of longer-term climate change and variability at centennial to millennial scales. In doing so, the paleoclimate record enables an assessment of how unusual the rate of recent warming is. For example, analyses of multiple proxies extracted from marine sediment and ice core records indicate that the past 2.6 million years of Earth’s climate history have been characterized by repeated cycles of glacial advance and retreat. These cycles, which are driven by changes in Earth’s orbital elements, result in the spread of continental ice sheets in the Northern Hemisphere, the expansion of alpine glaciers and sea ice globally, and significant changes in sea level. During the Last Glacial Maximum (LGM), which occurred at 21 ka BP, global sea level and mean temperature were approximately 120 m and 4 °C lower than at present, respectively. The geologic record can also be used to document rates of change associated with past climate change events. For example, the Paleocene–Eocene Thermal Maximum (PETM) involved an initial increase in global temperature of 5 °C that occurred over approximately 10,000 years (Diffenbaugh and Field 2013); this rate of warming is 1–2 orders of magnitude slower than the rate of warming projected to occur this coming century. Rates of change associated with the most recent glacial–interglacial transition, the Medieval Climate Anomaly and the Little Ice Age are all lower than the current and projected rates of change. Increasingly, climate scientists are concerned about the potential of abrupt climate change, which occurs when the climate system crosses a threshold or “tipping point,” and responds in a nonlinear fashion at rates that exceed the initial forcing (NRC 2013). Importantly, the paleoclimate record documents the existence of abrupt changes in global climate during the most recent glacial–interglacial transition. These events, the largest of which is the Younger Dryas, a millennial-length cold period that resulted from an abrupt change in Atlantic Meridional Overturning Circulation, led to a return to near-glacial conditions between 12.9 ka and 11.6 ka through much of the high latitudes of the Northern Hemisphere. Proxy–based paleoclimate data can also be used to test and verify the output of coupled ocean–atmosphere global climate models, simulating the response of the climate system to large changes in boundary conditions and forcings. Proxy and model-based studies documenting the expression of these abrupt climate change events provide valuable insight into the response of ocean and atmospheric circulation to climate forcing, and the resulting impacts of these changes on the structure, function, and composition of terrestrial and aquatic ecosystems. Other areas of concern related to abrupt change include the behavior of the ocean (acidification, oxygen concentration); atmospheric circulation (El Niño Southern
GLOBAL CLIMATE CHANGE

Impacts of climate change are widespread and include changes in the biophysical environment, phenology, species distribution and abundance, habitat availability, agricultural productivity, and the spread of infectious disease. Reductions in the volume and extent of alpine glaciers have influenced downstream freshwater availability, especially in mid- and low latitude regions, where the need for fresh water is increasing. Populations in mid- and low latitude regions are highly dependent on glacier meltwater for consumption, irrigation, and hydropower generation during the dry season, and as a result are susceptible to global climate change induced alterations of mountain hydroclimate. Earlier snowmelt also leads to reduced late summer stream flow, which negatively affects aquatic ecosystems, lowers the groundwater tables, and lengthens the fire season. However, it is important to recognize that the societal impacts associated with climate change are expressed differentially due to the varying exposure and vulnerabilities of populations to climatic and nonclimatic factors. For example, multidimensional inequalities in socioeconomic status and income, resulting from uneven development, influence the magnitude, risk, exposure, and vulnerability of various groups to current and future climate change (Richardson, Steffen, and Liverman 2011). The increase in the atmospheric concentration of CO₂ has resulted in ocean acidification. The decrease of ocean surface water pH of 0.1, which has occurred since the Industrial Revolution, negatively influences calcifying marine organisms. Changes in phenology, or the timing of life-history events, have been attributed to recent climate change. For example, in response to elevated temperatures, the growing season has lengthened by approximately two weeks in the mid- to high northern latitudes during the late twentieth and early twenty-first centuries. A lengthening of the growing season can influence trophic relations and result in trophic mismatches, such as the timing of migration, breeding, and/or food availability. Shifts in regional climate and the varying velocity of these climate shifts can also influence biotic interactions, species distribution, and community structure through competitive displacement, increased predation, or altered predator–prey relationships (Blois et al. 2013). Global climate change also influences the provision of ecosystem services (Nelson et al. 2013), with the provision, timing, and location of ecosystem services such as nutrient cycling, wildfire regulation, fisheries productivity, coastal flood protection, and water supply altered both negatively and positively by global climate change. Agriculture will be heavily affected by global climate change, with crop growth and quality, livestock health, and pests all potentially influenced by changing climate (temperature and precipitation) and extreme weather events, which in turn may threaten global food security. For example, increases in growing season temperatures may lead to decreases in crop and livestock production and yields through increasing pressures from pests, weeds, and pathogens. Changes in precipitation patterns may lead to periodic crop failures and contribute to long-lasting declines in crop yields, requiring shifts in the crops and livestock raised in particular regions (Thornton et al. 2014). Changes in climate and the resultant shifts in ecological conditions may facilitate the spread
of pathogens, parasites, and food-, water-, and animal-borne diseases. For example, the spread of vector-borne infectious disease as a result of the direct and indirect effects of climate change is an area of increasing concern. As temperature increases, the latitudinal and elevational range of the vectors responsible for transmitting malaria (Anopheles mosquito) and Dengue fever (Aedes aegypti mosquito) will expand and increase the human population exposed to the parasites responsible for these diseases. It is also expected that the incidence of waterborne infections, such as cholera, that result in diarrheal distress will increase as untreated runoff and/or sewage, associated with extreme precipitation events, enters the water supply (Altizer et al. 2013).

Projections of future climate

Coupled ocean–atmosphere global climate models are used to simulate the Earth’s climate system, investigate the sensitivity of the climate system to various forcings, and project changes in temperature and precipitation regimes in the future under various scenarios. The scenarios used to guide the climate models that form the basis for the global and regional climate projections presented in the most recent Intergovernmental Panel on Climate Change assessment report (IPCC 2013) are referred to as representative concentration pathways (RCPs). The four RCPs (RCP2.6, RCP4.5, RCP6.0, and RCP8.5) which reflect the radiative forcings (in W/m²) associated with four potential greenhouse gas concentration trajectories consistent with physical, demographic, and socioeconomic constraints, describe four possible climate futures. The RCP2.6 scenario characterizes a climate future in which strong reductions in greenhouse gas emissions occur and radiative forcing peaks in the mid-twenty-first century and declines by 2100 CE. The RCP8.5 scenario, which is most similar to the trajectory we are currently following, characterizes a climate future where greenhouse gas emissions continue to increase throughout the twenty-first century. It is important to note that the choice of RCP scenario greatly influences the magnitude of the projected changes in global temperature as early as the middle of the twenty-first century. By the end of the twenty-first century, global mean surface temperatures are projected to exceed 1.5 ºC relative to the 1850–1900 CE mean global surface temperature for all RCP scenarios except RCP2.6. Under all scenarios high latitude and high elevation regions are expected to warm at a considerably faster rate than low latitude regions (see Figure 3). For example, the RCP8.5 climate model simulations indicate that the warming experienced at high northern latitudes will exceed 4 ºC relative to the 1985–2005 CE midway through the twenty-first century, and that the warming in these same regions will exceed 6 ºC by 2100 CE (Diffenbaugh and Field 2013). Precipitation patterns are expected to show greater spatial variability relative to temperature, with the contrast between wet and dry regions and wet and dry seasons increasing during the twenty-first century. Ocean circulation and acidity will be influenced by the increase in ocean heat content and the partial pressure of atmospheric CO₂, respectively. Continued reductions in global glacial ice volume, sea ice, and snow cover, together with the thermal expansion of ocean water, will contribute to a projected sea level rise of between 28 cm and 98 cm by 2100 CE. The occurrence, duration, and timing of extreme weather and climate events and variability in weather patterns are all projected to increase as the planet warms. For example, most land areas will experience notable increases in the frequency of extreme hot seasons for over 80% of years, with mean summer temperatures above
Figure 3  Map of CMIP5 multimodel mean results for the scenarios RCP2.6 and RCP8.5 in 2081–2100 of annual mean surface temperature (IPCC 2013).
the late-twentieth-century maximum by 2100 CE (Diffenbaugh and Field 2013). It is expected that improvements in the parameterization of clouds, ground hydrology, ocean–atmosphere interactions, and the inclusion of models that incorporate the dynamic behavior of ice sheets, and refinements to ocean and terrestrial biogeochemical cycles, will serve to better constrain climate projections in the future.

SEE ALSO: Climate change and biogeography; Global climate change; Global climate models; Intergovernmental Panel on Climate Change (IPCC); Oceans and climate; Paleoclimatology; Paleoecology

References


Global climate models

H. Damon Matthews
Cassandra Lamontagne
Concordia University, Canada

Defining a climate model

Climate model basics

A climate model is a mathematical representation of the global climate system and its interacting components (Figure 1). The climate system is a complex entity, encompassing the atmosphere, the ocean, ocean sediments, sea ice, the land surface, land vegetation, and continental ice sheets. All of these components of the climate system interact with each other, exchanging energy, water, and atmospheric gases such as carbon dioxide (CO₂). In addition, all of the components of the climate system have the capacity to change over time in response to changes in the flows of energy entering or leaving the system.

The aim of a global climate model is to simulate the most important components of the climate system, how they interact with each other, and how they change over time. The basic foundation of climate models is the physical laws that govern the motion of air and water in the atmosphere and ocean and the transfer of energy by radiation and convection. Other mathematical equations are used to represent key processes such as the formation of clouds, as well as key interactions such as the transformation of atmospheric CO₂ into organic carbon by photosynthesis.

Climate models vary widely in character and complexity, from a few equations to hundreds of thousands of lines of programming code. State-of-the-art global climate models are based around what are called general circulation models (GCMs) of the atmosphere and ocean. GCMs are three-dimensional models of atmospheric and oceanic circulation, where the atmosphere and ocean are divided into thousands of grid boxes. Within each of these discrete grid boxes, the model solves the three-dimensional equations of motion to simulate the horizontal and vertical motion of air in the atmosphere or water in the ocean. Energy and water are allowed to move between grid boxes, as determined by equations representing the transfer of energy by radiation, and the transfer of both water and energy by convection, advection, and diffusion.

Evolution of climate models

Climate modeling is a relatively young scientific discipline, and climate models have developed enormously in scope and complexity over the past few decades. The earliest climate models, developed in the 1970s, grew out of weather prediction models, and were very much focused on the atmospheric component of the climate system. Additional components were added to models over time, beginning with the land surface, clouds, and ice in the 1980s. In the early 1990s, climate models began to include a simplified representation of ocean processes, and by the mid-1990s had grown to include three-dimensional representations of both atmospheric and oceanic circulation. These were the first generation of atmosphere–ocean general circulation models (AOGCMs).

The 1990s also saw the recognition of the importance of aerosols (both natural and...
Figure 1 Model representation of the components of the global climate system, showing the basic components of the climate system, as well as the primary exchanges of energy (red arrows), water (blue arrows), and carbon (green arrows).

anthropogenic) as drivers of climate change, which led to an expansion of the representation of human climate forcings in global models. The turn of the century was marked by the first introduction of the carbon cycle as a dynamic component of climate models, as well as a continued focus on aerosols in the atmosphere and deep ocean circulation. The current generation of global climate models includes interactive chemistry in the atmosphere, as well as dynamic vegetation distributions that are able to change in response to changing climate patterns.

This increase in the complexity of models over time has represented an increase in the number of processes and interactions included within models. As a result, components of the climate system have shifted over time from being variables prescribed as inputs to climate models, to being included as interactive components simulated by the models themselves.

Accompanying this growth over time in the complexity of climate models was an increase in the spatial resolution of models. The resolution of a climate model refers to the size of grid boxes, or, correspondingly, the number of grid boxes within a given area. Over time, the resolution of models has increased, which has resulted in more and more smaller grid boxes. This, as much as the increase in complexity described above, has resulted in a large increase in the computer power required to run global climate models.

Modeling observed climate change

Before a model can be used for future projections, it is important to establish the model’s ability to simulate climate changes that have already occurred. The ability to hindcast (or reproduce observed climate changes) is an
important test of a model’s ability to simulate the climate response to human greenhouse gas emissions, and builds confidence in the model’s ability to project these changes into the future.

Model validation

Hindcasting is an example of model validation, in which a climate model simulation is compared to observed climate data to establish the model’s accuracy and validity. For example, a model simulation of present-day temperatures can be compared to temperature data to establish how well the model is able to simulate current climate conditions. The left panel in Figure 2 shows the spatial pattern of temperature, as simulated by a global climate model, and the right panel shows the observed spatial pattern of temperature according to the data provided by the Climate Research Unit (CRU) of the University of East Anglia. This side-by-side comparison shows that, in general, the model simulation is true to the temperature data, though close inspection does also reveal some regional differences. Given that all climate models are a simplification of the real world, it follows that any model validation will reveal differences between the model simulation and the real-world data, so-called model biases. There is no objective rule for what constitutes an acceptable difference; deciding which model biases are tolerably small, and which need to be the subject of further model development and improvement, is part of the art of climate modeling.

A key benchmark that is often used for model validation is the global temperature record since 1850, the period of time over which thermometer temperature measurements are widespread enough to estimate changes in globally averaged temperature. Over this time period, we also have good records for the factors that have driven climate since industrialization. If these climate drivers (which can be expressed as changes in radiative forcing over time) are provided as input to a model simulation, the climate model should be able to simulate the observed pattern of temperature change over the past century and a half.

An example of this comparison of model-simulated temperature change to historical temperature data is shown in Figure 3. The model used for the simulation is in a class of climate model known as an Earth model of intermediate

![Model surface air temperature](image)

**Figure 2**  Comparison of model-simulated temperatures (left) to observed temperature data (right) provided by the Climatic Research Unit of the University of East Anglia.
complexity (EMIC), which is a global climate model in which some aspects of the climate system are simplified in the model to decrease the amount of computer processing time required to run the simulations. This particular model has a simplified atmospheric component, with the result that year-to-year (interannual) temperature variability of the model (red line) is greatly reduced in comparison with the temperature data (gray line). This known limitation does not prevent comparison of the longer-term trend. As can be seen in the figure, the model captures the warming that was observed in the early part of the twentieth century, followed by a slowdown of the warming trend in the 1950s and 1960s, and finally a rapid warming between 1960 and 2010. Overall, the model is able to simulate the approximately 0.8 °C of global warming that has been observed since 1860.

Sensitivity studies

We can also use climate models to understand why temperatures have increased in the manner that has been observed. We can use models simulations, driven by a range of different input simulations, to construct any number of different time series of temperature changes over the past 150 years, and in so doing determine which factors contributed to the observed change and how.

This process of repeat simulations of the same interval of time is an example of a sensitivity study. In general, a sensitivity study is a series of simulations that are performed to determine a model’s sensitivity to either variations in internal model parameters, or variations in external input data. In either case, the aim of a sensitivity study is to determine why a model gives the results it does.

In the case of the simulation of global temperatures over the past century and a half, a sensitivity study can be performed to determine how much of the model-simulated warming is the result of human climate drivers (CO₂ and other greenhouse gas emissions, aerosol emissions, and land-use change), and how much can be attributed to naturally driven climate change (solar insolation changes and volcanic eruptions). This requires two additional simulations: one where the model is given information only about natural climate drivers, and a second where the model is given information only about anthropogenic climate drivers (Figure 4). The blue line (natural only) shows that some of the warming prior to 1960 can be attributed to natural variation in solar and volcanic activity, but that since 1960, natural climate variability by itself would have caused global temperatures to decrease. By contrast, the green line (anthropogenic only) shows that human greenhouse gas and aerosol emissions caused some amount of warming in the early part of the twentieth century, and are responsible for all of the warming that has been observed over the past half-decade.
Projecting future climate change

When using a climate model to simulate future climate change in response to a scenario of future emissions, it is very difficult to establish the actual likelihood of the results given by the particular simulation that is performed. The future climate changes produced by the model will always be contingent on the choice of emissions scenario and on the choice of the climate model that is used. Climate modelers often prefer to use the term “projection,” rather than “prediction,” to describe the scenario of future climate change generated by a particular model in response to a particular future emissions scenario.

Emissions scenarios

Efforts to forecast future emissions generally fall into the category of emissions scenarios, which are projections based on plausible storylines that reflect a range of different possible future societal decisions with respect to energy technologies, globalization, and environmental concern.

The first widely used set of emissions scenarios was produced by the Intergovernmental Panel on Climate Change (IPCC) in association with its First Assessment Report in 1990 (IPCC 1990). These were the so-called IS92 scenarios, which included the widely used IS92a business-as-usual emissions scenario (a scenario that assumes continued economic growth and no climate policy aimed at emissions reduction) with continually increasing greenhouse gas emissions throughout the twenty-first century. The next major round of greenhouse gas emissions scenarios came in the year 2000 with the publication of the IPCC Special Report on Emissions Scenarios (SRES) (IPCC 2000). This report contained details of a total of 40 different business-as-usual emissions scenarios, each reflecting different storylines of economic development, technological change, and population growth. None of the SRES scenarios included explicit climate policies to decrease emissions, and therefore they represent a range of possible futures that human societies could follow in the absence of specific efforts to decrease our impact on the global climate system.

In contrast to business-as-usual emissions scenarios, stabilization scenarios do include climate mitigation efforts, whereby future emissions are projected to decrease to a level that is consistent with the stabilization of atmospheric concentrations of greenhouse gases. Stabilization scenarios are typically not based on projections of economic growth, but are rather set based on a desired target for atmospheric CO₂ concentrations, such as 550 ppm.

The most recent development of future scenarios are the “Representative Concentration Pathway” (RCP) scenarios, which include some scenarios with storylines that assume business-as-usual emissions (similar to the SRES scenarios) in addition to other scenarios with storylines that assume plausible policy efforts to
GLOBAL CLIMATE MODELS

curb global greenhouse gas emissions (stabilization scenarios). There are four categories of RCP scenarios, named according to the target level of radiative forcing – the amount by which greenhouse gases in the atmosphere change the balance of incoming and outgoing radiation at the top of the atmosphere – at the year 2100. At the high end, the RCP 8.5 is a scenario of increasing emissions throughout most of this century, resulting in a radiative forcing of 8.5 W m\(^{-2}\) in the year 2100 (representing atmospheric CO\(_2\) levels of about 800 ppm). At the low end, the RCP 2.6 scenario has radiative forcing increasing until mid-century and then decreasing to 2.6 W m\(^{-2}\) at the end of the century. This change in forcing results from emissions that start to decrease at the year 2020, and reach zero by the last quarter of this century. In this scenario, net human greenhouse gas emissions actually become negative (implying intentional human greenhouse gas sequestration) by the end of the century. The RCP 4.5 and RCP 6 scenarios represent intermediate scenarios with emissions that increase for the next several decades and then decrease in the latter half of this century.

Climate forcing and response

To illustrate the process of generating a climate projection, consider the following example. Figure 5 shows a time series of radiative forcing which results from future emissions given in the RCP 4.5 scenario. In this figure, radiative forcing is broken down into its individual components, showing atmospheric CO\(_2\) concentrations increasing considerably throughout the twenty-first century in response to continued CO\(_2\) emissions. Nitrous oxide concentrations (and hence forcing) also increase somewhat, though the forcing due to methane and halocarbon both decrease slightly in the latter part of the century. Finally, the negative radiative forcing from aerosols also becomes smaller over time, reflecting a decrease in aerosol emissions between now and the year 2100.

This time series of radiative forcing can be provided as input to a global climate model. The model will then simulate the temperature changes that would be expected to occur if future generations produce emissions that are similar to the RCP 4.5 emissions scenario. According to one model simulation, global temperature would increase by about 1.5 °C in response to the overall increase in radiative forcing (which is mostly driven by CO\(_2\) emissions) shown in Figure 6; therefore, the RCP 4.5 scenario would result in 1.5 °C of warming by 2100 (left panel of Figure 6).

The right panel of Figure 6 shows the spatial pattern of warming over this same time period. As can be seen here, the entire globe does not warm at the same rate. Warming over high-latitude northern continents is about double the global average warming, whereas the tropical oceans warm somewhat less than the global average. This regional pattern is partly

![Figure 5 Radiative forcing from 2005 to 2100 associated with the RCP 4.5 emissions scenario.](image-url)
because land warms faster than oceans, and partly because of a powerful positive climate feedback that amplifies warming at high latitudes: the ice–albedo feedback. As temperatures over snow-covered regions increase, this leads to fast melting (or less accumulation) of snow and ice. This in turn leads to a less reflective surface, since snow has a much higher albedo than either soil or vegetated land surfaces. Consequently, more solar radiation is absorbed, and the warming in these regions is amplified.

**Sources of uncertainty in model projections**

The projection of future climate warming shown above is contingent on both the particular climate model that was used for the simulation, and the particular emissions scenario that was chosen as input to the model. Clearly, a different scenario of future emissions (e.g., the RCP 2.6 or RCP 6 scenario) would result in a different amount of simulated future warming. However, using the *same* emissions scenario as input to a *different* climate model would also give a different amount of simulated warming. In fact, the range of simulated warming produced by different models could be as large as, or larger than, the range of simulated warming in a single model in response to different emissions scenarios.

This difference in the climate change projections produced by different models, or intermodel uncertainty, reflects fundamental uncertainties in our understanding of which components of the climate system are important to include in models, as well as how to model certain interactions and feedbacks which occur in the real world. For example, there is considerable uncertainty in our understanding of how the global carbon cycle might change in response to human CO₂ emissions. The choices that modelers make in building the carbon cycle component of a particular climate model will result in a range of representations of the carbon cycle and carbon sinks in different climate models. Consequently, there will be carbon cycle uncertainty in future warming simulations because different models will simulate different trajectories of atmospheric CO₂ increase in response to a given scenario of future emissions.

Similarly, there are many different climate feedbacks which affect how much global temperatures will respond to a given increase in atmospheric greenhouse gas concentrations. As in the case of carbon cycle processes, the
GLOBAL CLIMATE MODELS

atmosphere, ocean, and ice components of different climate models vary considerably in terms of which feedbacks are included in the model, and in how a particular feedback is represented. This is true of the ice-albedo feedback described above, which means that the projected pattern of high-latitude warming will vary from model to model.

Another particularly important (and uncertain) feedback is the cloud feedback. Climate warming will likely lead to changes in the distribution and amounts of different types of clouds, which would in turn affect both the amount of incoming solar radiation and the amount of longwave radiation that is absorbed by clouds and re-emitted back to the surface. The formation of clouds is a difficult process to model correctly, with the result that climate models vary widely in how clouds are represented. Furthermore, different types of clouds have different effects on radiation transfer in the atmosphere. An increase in high clouds due to warmer temperatures will tend to increase the strength of the greenhouse effect, resulting in a faster rate of warming (a positive feedback). By contrast, an increase in low clouds will lead to more reflected sunlight, which will slow the rate of warming (a negative feedback). Because of the widely varying representation of clouds in different models, some models will simulate a negative cloud feedback (and hence less warming), whereas some other models will simulate a positive cloud feedback (and hence more warming).

The net effect of all feedbacks in a climate model will determine its climate sensitivity: the amount by which the climate changes in response to a given change in radiative forcing. There are two specific measures of climate sensitivity that are commonly used to compare the performance of different climate models and to try to quantify some of the intermodel uncertainty associated with future climate change projections.

The transient climate response (TCR) measures the global temperature change in a model simulation at the time that atmospheric CO₂ concentrations reach 560 ppm, or double their preindustrial value. Equilibrium climate sensitivity, on the other hand, measures the long-term temperature change that would occur if atmospheric CO₂ concentrations were to be stabilized indefinitely at 560 ppm. Because of the slow response of the climate system to changes in forcing, climate will continue warming for several centuries after the time that CO₂ concentrations reach and are stabilized at 560 ppm. Consequently, the equilibrium climate sensitivity of a given model will always be higher than its TCR.

The progression of uncertainty, from future emissions scenarios, to changes in atmospheric concentrations of greenhouse gases, to changes in climate and associated impacts, is sometimes called the cascade of uncertainty (see Figure 7). Human choices are subject to uncertainty, given the wide range of future decisions we could make regarding energy sources and the resulting greenhouse gas emissions. But even given some scenario of future CO₂ emissions, predicting the eventual impact of climate change on human societies depends on uncertainty.

Figure 7 Progression from emissions to impacts, showing the increase in total uncertainty associated with climate change prediction.
associated with the carbon cycle response to emissions, followed by additional uncertainty surrounding the climate response to an increase in atmospheric CO₂ (climate sensitivity being one of the most important unknown quantities in climate science), followed by even more uncertainty associated with understanding how much a given amount of climate change will impact a given sector of human society.

Predicting the future is never easy, and predictions of future climate change are no exception. It is important to understand, however, that some things are more uncertain than others. For example, we are confident that continued greenhouse gas emissions will lead to additional climate warming. How much warming actually occurs, and how this warming will vary regionally across the globe, is ambiguous because of the many contributors to uncertainty discussed above. Nevertheless, even given uncertainty, we can draw some general conclusions, which represent common trends that emerge from the many different model projections that have been performed.

Global changes

Temperature

Model-simulated changes in global mean temperature for a given future year are often depicted as a range of modeled temperature deviations from some baseline (such as the average of 1986–2005 mean temperatures) for a range of scenarios of future greenhouse gas emissions and across different global climate models. Across all RCP scenarios and all models used in the IPCC’s Fifth Assessment Report (IPCC AR5), global temperatures are projected to increase throughout the twenty-first century, with a likely range (which excludes some model outliers) of 0.3–4.8 °C above 1986–2005 levels (Collins et al. 2013).

This large range of model-simulated warming is due mainly to different emissions scenarios; that is to say, human choices regarding future emissions are the largest single determinant of how much climate will change by the end of the century. There is also uncertainty within results associated with a given emissions scenario, however, which reflects differences between models (intermodel uncertainty). For example, for RCP 6, the simulated likely range of warming across all models is 1.4–3.1 °C above 1986–2005 levels by 2100 (Collins et al. 2013).

Global warming does not mean that the entire globe warms by the same amount. Figure 8 shows the average spatial pattern of warming simulated by a range of climate model projections, for each degree of globally averaged temperature increase. As already discussed, temperatures warm more over continental regions and at high northern latitudes. The figure shows that some regions of the Canadian Arctic will warm by as much as 2.5 or 3 °C per degree of global warming.

Precipitation

In general, precipitation is expected to increase in a warmer world as increased temperatures drive more evaporation from the oceans, and warmer air is able to hold more water vapor. However, the expected pattern of precipitation change is highly variable, with some areas seeing a decrease in precipitation, and other areas seeing an increase. Figure 9 shows the percentage change in precipitation associated with each degree of global temperature increase. The United States is an example of a nation that will experience a trend toward more precipitation in some of its regions and less precipitation in others. The most intense increase (with as much as 35% more precipitation per degree
Figure 8  Change in regional temperature per 1 °C change in global mean temperature, based on the range of model projections from current global climate models (Collins et al. 2013).

of global warming) is expected to occur over the Pacific equator, in a northern subtropical portion of the African continent, and over a portion of the Arabian Sea and coast. Other regions of the world will see significantly less precipitation, particularly subtropical areas such as the Mediterranean and the western tip of Australia. In these areas, precipitation is expected to decrease by as much as 10% for every degree of global temperature increase.

A common pattern that emerges from climate model simulations of precipitation changes is that regions that are already wet will tend to get wetter, and regions that are dry will get drier. This is consistent with an overall enhancement of the global hydrological cycle, which will tend to enhance precipitation at higher latitudes and along the equator, and decrease precipitation in dry subtropical regions. However, evidence also suggests that areas which are projected to receive more precipitation may receive that increased precipitation in the form of more intense individual precipitation events, rather than as a simple overall increase (Collins et al. 2013). This means that regions such as Canada or Russia, though expected to receive more precipitation in total, could also see an increase in periods of drought, especially in areas of the country that already experience dry weather conditions at certain times of the year.

Extreme events

Many extreme events are expected to change in frequency and duration in a warmer climate. As previously discussed, climate change will likely lead to an increase in precipitation extremes, which means that unusually intense storms will bring more precipitation as a result
of warmer air temperatures. There is also some evidence that tropical cyclones and hurricanes could increase in intensity (while decreasing or remaining unchanged in frequency) as tropical ocean temperatures increase and are able to provide more energy to fuel the development of these intense storms (Christensen et al. 2013). Tornados are another type of extreme weather event that derive their energy from high surface temperatures, and which could therefore increase in intensity in a warmer climate.

Perhaps the most intuitive change in extreme events that we might expect is in the frequency and severity of temperature extremes. As expected, a warmer climate, and warmer winter temperatures in particular, could lead to fewer very cold days and nights. Additionally, warmer temperatures will lead to an increase in very hot summers. This in turn would likely increase the frequency and intensity of heat waves, such as that seen in Europe in the summer of 2003. Recent projections indicate that for many regions, a 2°C increase in global temperatures will result in every summer exceeding the temperature of the hottest summer of the past 20 years (National Research Council 2011).

Glaciers and ice sheets

Some of the most striking manifestations of climate change are occurring in the cryosphere: the frozen glaciers and ice sheets of the planet. The Canadian Arctic has warmed more than almost any other location on the planet, and this pattern of warming is expected to continue into the future. Mountain glaciers and ice caps the world over are receding in response to warmer temperatures, and this trend is expected to
GLOBAL CLIMATE MODELS

continue with additional warming. By the end of the century, glaciers around the world could decrease in volume by more than 75%, which would mean the disappearance of many individual mountain glaciers in the latter part of this century. With every degree Celsius increase in global temperatures, glacial volume will reduce by 0.32–0.40 m per year (Meehl et al. 2007).

Equally striking are the changes occurring in the world’s two major continental ice sheets: Antarctica and Greenland. Many Antarctic glaciers have accelerated the rate at which they discharge ice to the Southern Ocean, particularly along the margins of the Antarctic Peninsula (the portion of Antarctica that juts out northward toward South America). There is also concern among glaciologists that warming ocean temperatures could dramatically increase melting of the West Antarctic Ice Sheet adjacent to the Antarctic Peninsula, and could decrease the stability of this large mass of ice that is grounded below sea level.

In Greenland, the seasonal melt area has also increased dramatically in recent years, and is expected to continue to increase as arctic temperatures continue warming. As in the case of the West Antarctic Ice Sheet, scientists are increasingly concerned about the stability of the Greenland ice sheet. While simple melting of the Greenland ice sheet is a relatively slow process (it would take close to 2000 years to fully melt Greenland, even if CO₂ levels rise to more than 1000 ppm (Meehl et al. 2007)), there is the potential for dramatic increases in the flow of Greenland’s outlet glaciers, which could accelerate the decrease in ice volume that is expected in the coming decades.

Sea level rise

Many factors contribute to the potential for climate change to lead to rising sea levels. As discussed above, mountain glaciers, ice caps, and continental glaciers are all decreasing in volume, which is resulting in the transfer of this volume of water to the ocean. If all of the mountain glaciers and ice caps on the planet were to melt completely, this would increase global sea levels by about 70 cm.

The major concern for rising sea levels, however, comes from the potential melting of continental ice sheets. Greenland alone contains enough frozen water to raise sea levels by 6.6 m, were it to melt completely (Allison et al. 2009). An additional potential for sea level rise of 3.3 m comes from the volume of water stored in the vulnerable parts of the West Antarctic Ice Sheet.

In addition to these potential increases in the total amount of water in the ocean, the volume of ocean water will increase simply by virtue of increased temperatures. Similar to the increase in volume which occurs when air is heated, warmer water also increases slightly in volume. When applied to the entire volume of the ocean, this thermal expansion has a notable contribution to global sea level rise; an estimated 75% of observed sea level rise is attributed to thermal expansion (Church et al. 2013).

There is a huge amount of uncertainty associated with future sea level rise projections, particularly in regard to the contribution of Greenland and Antarctic melting to global sea levels. The Copenhagen Diagnosis report (Allison et al. 2009) summarizes some of the recent estimates of future sea level rise, showing continued greenhouse gas emissions will likely increase global sea level by between 50 and 150 cm by the year 2100. As with other climate changes, however, the regional pattern of sea level rise will not be the same everywhere; both local ocean temperatures and the patterns of ocean circulation will lead to regional differences in sea level relative to the global average. For example, sea levels along the arctic and eastern coasts of Canada will likely increase by 2–12 cm.
more than the global average by the year 2100. In addition, sea level rise will continue long after the year 2100, potentially increasing by as much as 1–3 m by the year 2300 (Church et al. 2013). The uncertainty associated with such longer-term estimates is even larger than for near-term sea level rise, and sea level increases beyond the year 2100 will also depend strongly on how quickly human societies are able to decrease our emissions of greenhouse gases and slow the rate of climate warming.

SEE ALSO: Climate policy; Climatology

References


Global commodity/value chains

Deborah Leslie
_University of Toronto, Canada_

A global commodity chain is composed of all of the processes associated with the production of one good or service, including manufacturing, consumption, design, retailing, marketing, advertising, and disposal. Global commodity chain analyses trace these activities, analyzing the distribution of power across a chain and the flow of material resources, value, money, knowledge, and symbols between sites. A key concern is to investigate the way in which the production and distribution of goods and services reflects but also tends to perpetuate patterns of uneven development in the global economy.

There has been a growing interest in tracing the linkages between sites in a chain, partly because an increasing number of products consumed are produced in other countries (and often multiple countries), and yet consumers often know very little about how or where they are made. Commodity fetishism means that consumers are focused on the price of a product, as well as its physical and symbolic characteristics. This emphasis masks the underlying relations of production and distribution associated with goods. Commodity chain analyses allow us to trace different moments in the life of a commodity to uncover these meanings.

A variety of different commodities have been studied using a commodity chain approach, including clothing, footwear, furniture, gold, diamonds, electronics, flowers, coffee, tropical fruits, and vegetables. Recent research also explores shifts in the commodity chains associated with energy and many services, including call centers, accounting, and research and development. While there are a number of related concepts (such as supply chain, commodity circuit, actor network, _filière_, system of provision and global production network), this entry will focus on the global commodity chain (GCC) approach, and more recent versions of it known as global value chain (GVC) research. It will begin by tracing the development of these approaches and the main arguments in these literatures. Strengths and criticisms of the GCC and GVC frameworks will also be summarized.

The global commodity chain approach

The concept of the commodity chain originates in the work of world-systems theorists. Hopkins and Wallerstein (1977) first used the concept of the commodity chain to refer to the raw materials, transportation networks, and labor processes involved in the production of a commodity, as well as the food inputs into the labor involved in making it. They define a commodity chain as “a network of labour and production processes whose end result is a finished commodity” (Hopkins and Wallerstein 1986, 159). The concept of the commodity chain is thus meant to denote the transformation of raw materials into final products, but also the connections between production and the social reproduction of human labor at each site in the chain. A particular focus is on the historical globalization of capitalist commodity chains, and how this internationalization is tied to patterns of uneven development.
Another key figure in articulating a commodity chain approach is Gary Gereffi. Along with a number of colleagues, Gereffi (1996) uses the term “global commodity chain” to refer to the more recent interfirm relationships that constitute global production networks. Gereffi argues that, in the current period, commodity chains are characterized by increasing geographical complexity and international scope. Unlike the more macro-level and holistic approach of world-systems theory, which examines the logic and structure of world capitalism, emphasis is placed on the microscale. In particular, the focus is on the linkages between firms involved in the production and distribution of a product. The aim is to identify how, where, and who is adding value to the product, and the implications of these patterns for economic development.

Global commodity chains have four important characteristics, according to Gereffi (1996). The first is an input–output structure, which refers to a group of products and services linked together in a sequence of value-adding activities. Second, each chain has a distinct geography. Some chains are spatially concentrated, while others are more globally dispersed. High value-added activities are typically located in key cities in the Global North, while manufacturing takes place in a range of geographical locations that are subject to constant change. The third characteristic is that individual GCCs have unique forms of power relations or governance. This refers to how chains are organized and controlled by “lead” firms. Fourth is the institutional framework: how local, national, and international conditions and policies shape chain dynamics. For example, trade and tax policies, labor and environmental certification schemes, health and safety rules, environmental regulations, and government policies relating to employment and minimum wages all impact commodity chains. Commodity chains also often span several institutional contexts, raising considerable challenges for firms that must navigate different rules and regulations.

A number of themes animate contemporary GCC research. One issue receiving a lot of attention is the expanding geography of chains, partly as a result of innovations in transportation and communication technologies that make it easier to communicate at a distance and to move commodities around the world rapidly. The international division of labor shifted during the mid- to late twentieth century, with developing countries becoming important sites of manufacturing. This is especially the case with labor-intensive industries that have low barriers to entry. In contrast, more capital-intensive industries remain clustered in developed economies. A central preoccupation in GCC analysis has been to map and explain these patterns.

As one example of the growing role of developing countries in manufacturing, China now accounts for 67% of the world’s processing exports (WTO and IDE-JETRO 2011, cited in Gereffi and Lee 2012, 26). Manufacturing is concentrated in a number of large contract manufacturing companies such as Hon Hai/Foxconn, Flextronics, and Quanta. While China has benefited tremendously from this production, it does not create or capture a high level of value in the commodity chain. As Gereffi and Lee (2012) point out, there is often a discrepancy between where final goods are produced and where value is created and captured. They use the example of the Apple iPhone, manufactured in China by a Taiwan-based company, Foxconn, and then exported to the United States. China has a trade deficit with the United States, however, because it is charged for the inputs that it imports from the United States. In addition, most of the value for the iPhone is generated
in Korea, which produces the display panel and the memory chips – the most expensive parts of the phone (Gereffi and Lee 2012). This example illuminates the importance of tracing where value is being added in chains stretching across global space.

As is evident in the case of the iPhone, the current era is also associated with changing patterns of ownership. Globalization no longer encompasses transnational corporations setting up branch offices abroad. In many countries in the Global South, manufacturing supply firms are domestically owned, or owned by companies from other developing countries. GCCs also increasingly involve trade in components and subassemblies rather than finished goods. The growing international trade in intermediate goods is referred to as a shift from a trade in goods to a trade in “value added” or “tasks” (OECD 2011, cited in Gereffi and Lee 2012, 25).

Of the four features of global commodity chains identified by Gereffi (1996) governance structures have received the most attention in the literature. A particular focus has been on factors internal to a chain, rather than external governance. GCC scholarship examines the power relationships between firms, identifying powerful or lead firms, known as “chain drivers.” The early literature on GCCs identified two main ideal types of governance. First are the producer-driven chains. Here, transnational manufacturing firms play a pivotal role in coordinating the chain, including backward and forward linkages. Producer-driven chains are common in capital-intensive industries such as aerospace, computers, and automobiles, where suppliers make components rather than the final product. Producer-driven chains tend to be characterized by high wages and highly skilled labor. Profits stem from advantages related to scale and technological innovation. The second type of governance structure is the buyer-driven chain, where retailers and brand name marketers play the central role in coordinating production networks and dictating conditions along the chain. Buyer-driven chains are typical in labor-intensive industries, such as garments, footwear, toys, and electronics. Production involves multiple tiers of subcontractors, with suppliers making products according to the detailed specifications of buyers. In these industries, subcontractors make the finished product rather than components; labor tends to be unskilled and low wage, and profits derive from design and marketing.

Over time, there has been a shift from producer-driven to buyer-driven chains in a number of sectors, including food, footwear, and clothing. Powerful retailers such as Walmart, and brand name companies such as Nike and Apple, are increasingly able to exert pressure on their suppliers to conform to their rigid specifications. They dictate how chains operate and require suppliers to conform to their exacting standards. The role played by these powerful buyers has been augmented by concentration in the retail sector. For example, Walmart now operates 11,000 retail stores, under 71 banners in 27 countries. It is China’s seventh largest trading partner, larger than the United Kingdom (Gereffi and Lee 2012, 28).

While this conceptualization of producer-driven and buyer-driven chains is useful, there is also an acknowledgment that different governance structures may coexist within an industry, particularly within different segments (such as mass market vs specialty products, or organic vs conventional agriculture). Governance may change over time and across space. Others argue that the growing import-driven chains, such as automobiles and computers, are now outsourcing not only the production of parts but also final assembly. Researchers also increasingly
GLOBAL COMMODITY/VALUE CHAINS

recognize that different types of power may exist at different sites in the chain, and that chains may be controlled by actors other than manufacturers or retailers, such as trading and distribution companies or online intermediaries. The role of external actors, such as nongovernmental organizations (NGOs), certification bodies, and other experts, is also significant. These shortcomings have led to the search for alternative conceptualizations of governance within commodity chains.

The transition from global commodity chains to global value chains

In the 2000s Gereffi and colleagues shifted to a new terminology of “global value chain.” Beyond the conceptual weaknesses identified in GCC approaches to governance, researchers were using a variety of terms to refer to a similar concept, and there was a desire for a common vocabulary. Another motivation for developing the concept of the global value chain was that the notion of “commodity” was perceived to refer narrowly to agricultural staples.

As Gereffi and Lee (2012) note, the GVC concept is now widely accepted as a label for the literature on commodity chains. The notion of a global value chain has been widely adopted not only in academic discourse but also within international organizations interested in economic development and poverty alleviation, such as the World Bank, World Trade Organization (WTO), Organisation for Economic Co-operation and Development (OECD), and International Labour Organization (ILO).

While it is similar to the global commodity chain concept, which emerged in sociology, the GVC approach is more heavily influenced by international business school literatures. Less emphasis is placed on chain drivers and greater importance is attached to the coordination of chains. There is a strong focus in GVC approaches on the technical and organizational dynamics of chains and on transactions costs. There is also more of a focus on the relationship between a lead firm and its first-tier suppliers. As Gibbon, Bair and Ponte (2008, 323) point out, in GVC analysis the analytical scope is narrowed from the entire length of the chain to one specific site. This takes away from one of the main advantages of a global commodity chain analysis, which is examining the effects that governance structures have across the chain. There is considerable debate about whether GCC and GVC approaches are complementary or competing (see Bair 2005; Gibbon, Bair and Ponte 2008). For many external analysts, GVC approaches mark a distinct break with the theoretical traditions that informed GCC research. In this view, GCC/GVC analysis is better conceived of as a methodology for studying globalization rather than as a coherent theoretical framework seeking to explain it.

Moving beyond the binary of producer- versus buyer-driven chains, Gereffi, Humphrey and Sturgeon (2005) present a more complicated map of commodity chain governance based on the complexity of information involved in a transaction, the extent to which knowledge can be codified, and the technical capabilities of suppliers. These three variables are rated as either high or low. Using these criteria, they create a matrix identifying five ideal types of governance in global value chains. Each form of governance involves different levels of outsourcing, different relationships between buyers and suppliers, and contrasting forms of power. Market governance refers to situations where the complexity of information is low and knowledge is easily codified. Product specifications are usually straightforward and transactions can be completed with little coordination. Relationships are at arm’s
length and price is the central factor in negotiated price. Modular governance describes cases where technical standards are typically high: product specifications are complex but also uniform and easy to codify, which simplifies the process of interaction and reduces the need for monitoring suppliers. Buyers and suppliers lower coordination costs by exchanging information. Relational governance refers to situations where exchanges rely on tacit knowledge, which is not easy to codify. Exchange of tacit knowledge requires frequent face-to-face interactions. Relations of mutual dependence thus develop between buyers and suppliers, so there is a high cost associated with switching partners. Captive governance refers to situations where suppliers are dependent on a small number of buyers. There is a high level of product specification and control by the lead firm. As a result, supplier capabilities are low, with firms often completing a narrow range of tasks. Under these conditions, suppliers face considerable switching costs and are thus “captive.” The final type of governance is hierarchical governance, which occurs when there is a high level of vertical integration. Manufacturers make products in-house because capable suppliers cannot be found. Products are complex and knowledge cannot be codified. Lead firms often seek to control the resources, the exchange of tacit knowledge and intellectual property. Each of these forms of governance involves different degrees of power asymmetry, ranging from low in the case of market governance to high in the case of hierarchical governance.

A key theme in both the GCC and GVC literatures concerns possibilities for industrial upgrading among firms in developing countries. Upgrading can be defined as the ability of a firm or group of firms to improve their position within a chain. It refers to the increasing competitiveness of a firm and its ability to capture greater value added in the production process (Bair 2005, 165). The potential for upgrading is related to the degree to which buyers transfer knowledge to their suppliers.

Four types of upgrading have been identified: product; process; function; and intersectoral upgrading (Humphrey and Schmitz 2001). Product upgrading refers to making more sophisticated products with higher unit prices. Process upgrading refers to enhancing efficiency, either through restructuring the production process or by implementing new technologies. Functional upgrading refers to the ability to take on new roles or functions, such as making the product rather than the components, or carrying out logistics management, distribution, or design functions. Intersectoral (or interchain) upgrading signals an ability to take knowledge learned from one sector (e.g., clothing) and expand into another sector (such as housewares). A key argument in the GVC literature is that chains organized along more hierarchical lines lend themselves to product and process upgrading, but are unlikely to allow firms to expand into marketing or design.

In general, the literature suggests that there are increasing barriers to upgrading in developing countries. Marketing, branding, and design are increasingly important in all commodity chains, especially in buyer-driven chains, and developed countries have significant advantages in these capabilities. As these “intangible” aspects become critical, it is increasingly difficult for developing countries to engage in upgrading (Gereffi 2001).

Another theme in the literature on GCCs and GVCs concerns the shifting geography of end markets. With a slow recovery from the economic crisis of 2008–2009 in developed nations, emerging economies such as China, India, and Brazil have become increasingly important markets. This has led to the development of regional South–South supply chains that
GLOBAL COMMODITY/VALUE CHAINS

alter the conditions of possibility for upgrading (Gereffi and Lee 2012). The demand for lower quality and less sophisticated products in developing countries presents opportunities for developing country firms to produce products for these markets. These firms can engage in higher value-added functions, such as product development and design, roles they have less opportunity to play in global supply chains. This enables them to develop innovations suited to the local context. However, focusing on low-income developing country consumers has the disadvantage that it could restrict developing country suppliers to more competitive and cut-throat markets with slim profit margins. Nonetheless, the growth of markets in the Global South points to the growing complexity of global value chains, and the variety of upgrading opportunities within these networks.

Contributions of a GCC/GVC approach

In her review of the literature on global commodity chains, Bair (2005) identifies a number of important contributions of the GCC/GVC literature. Tracing spatially diffuse and complex production networks allows researchers to “explore globalization in situ,” analyzing particular locales while also illuminating how these places are connected through global commodity chains (Bair 2005, 159). A GCC (and GVC) approach thus facilitates an understanding of an abstract concept, globalization, in a concrete way. GCC/GVC studies also illuminate how power is exercised in global industries, which has important policy implications for developing countries. Understanding how chains are organized is useful for countries in determining how to gain access to the skills, technologies, and know-how necessary to function in global value chains (Bair 2005).

GCC research in particular has also influenced ethical campaigns surrounding global production. The ability to trace complex supply chains has led to antisweatshop campaigns targeting large retailers, holding them accountable for labor conditions in the factories of their suppliers and subcontractors. This has led to a number of initiatives designed to alter the organization of global commodity chains, such as fairtrade movements, corporate codes of conduct, and a variety of certification schemes targeting human rights, labor, and environmental practices.

Criticisms of the GCC/GVC approach

While many valuable insights have been gained through the focus on GCCs and GVCs, a number of criticisms have been made of both approaches. These can be divided into six main categories.

The first, and perhaps the most common, critique of the GCC/GVC approach centers on its linear conceptualization of the chain. It is presumed that there is a sequence of activities—a chain, starting with raw materials and proceeding through the stages of design, production, marketing, distribution, and retailing. In this conception, consumption is viewed as the end point of the chain. Alternative approaches draw on the less linear vocabulary of “circuits,” “actor networks,” or “global production networks.” Rather than searching for a singular or determining logic of a chain, and tracing relationships of causality (or identifying chain drivers), these accounts explore the multiple webs of interdependence between sites and the multistranded relationships between nodes in the chain, and between firms, states, organizations, and raw materials.

In actor-network theory, for example, power is conceptualized as an effect of relations between actors in the chain. Agency is generated through
a network of heterogeneous, interacting materials, including not only retailers, suppliers, designers, and other human actors but also technologies, material natures, and the commodity itself. The value of an actor-network approach stems from its more fluid conception of networks, and its open conception of spatiality. While production and other activities are localized, they are constituted by influences from different time periods and locations. Practices are therefore both local and nonlocal. This topological conceptualization allows for a consideration of overlapping spaces.

In economic geography, the concept of global production networks (GPN) has also been forwarded as an alternative to GCC/GVC approaches (see, e.g., Coe, Dicken, and Hess 2008). Like circuits and actor networks, the GPN approach is a more open, relational framework that includes a range of actors and relationships, and is more sensitive to the role of geography. Avoiding an emphasis on one site or scale, it also examines how relationships stretch across space.

Bair and Werner (2011) raise a second criticism of the literature on global commodity and value chains. They argue that GCC/GVC approaches adopt a narrow firm-centered perspective and are not attentive enough to the historical and geographical context in which chains operate. Greater examination is needed of the “coupled processes of linking and delinking that underlie the shifting geography of global industries and condition the position of regions within transnational production arrangements” (Bair and Werner 2011, 1000). They propose the concept of disarticulation as a way to unravel the histories of dispossession, accumulation, and disinvestment that shape a region’s position in global commodity chains, and to examine how these histories are both constitutive of and constituted by networks of global production. The focus on lead firms also tends to distance GCC analysis from the way chains are embedded in particular places, histories, and ecologies.

In a similar vein, Bair (2005) argues that the GCC/GVC focus on individual firms and sectors leads to the neglect of a country or region’s position within the global economy. This firm-centered approach poses a dilemma of how one understands the implications of firm upgrading for larger units such as regions and nations. Bair (2005) suggests that there needs to be greater attention to who benefits from upgrading. Process upgrading, for example, does not necessarily translate into increased security and profits for subcontractors. In addition, larger producers within a country may be able to upgrade, but smaller players may be excluded from upgrading. Finally, an ability to upgrade may not necessarily benefit workers, who continue to struggle with low wages and job insecurity. Thus the global value chain literature needs to pay greater attention to the winners and losers in processes of upgrading and to historical and geographical context (Bair 2005).

A third criticism of GCC/GVC approaches centers on their economistic approach. Bair (2005) calls for greater attention to social reproduction (something foregrounded in earlier world-systems research), as well as the social, cultural, political, and institutional context in which chains operate. While the GCC/GVC approach has proven useful in understanding the internal power dynamics between firms and their suppliers, greater attention needs to be paid to factors external to chains, such as institutions, government policies, and other regulatory measures. GCC/GVC research emphasizes the processes of production and value-adding sequences, but the role of discourses could be given greater weight. Chains produce not only goods and services but also meanings and identities. GCC/GVC approaches neglect the mutually constitutive link between the material and the symbolic. Research
GLOBAL COMMODITY/VALUE CHAINS

needs to emphasize the way material qualities are transformed and culturally valued, and how this is related to the capture and realization of value.

A fourth criticism centers on the neglect of particular sites in the chain. Very little GCC/GVC work has examined the dynamics of consumption, the role of consumers in shaping chains, or consumer spaces such as the home (Coe, Dicken, and Hess 2008). Consumption can shape the nature of chains through ethical campaigning and the development of standards and codes of conduct. Research in geography on user-led innovation also illustrates the growing role that consumers play in driving innovation in some chains, especially sporting goods, medical equipment, and video games. Beyond consumption, there is also a lack of attention in the GCC/GVC literature to the disposal, recycling, and reuse of commodities, and on the natural environment, both as a source of raw materials and as a site of waste (Coe, Dicken, and Hess, 2008).

A fifth criticism is that GCC/GVC approaches neglect labor, which is often viewed as a passive victim of globalization. Labor needs to be integrated into the heart of a global commodity chain analysis because commodity chains are networks of embodied labor. Labor is also a key actor, engaging in struggles over outsourcing, restructuring, and other actions that influence the dynamics of commodity chains.

Echoing many of these criticisms, a final critique of the literature on global commodity and value chains surrounds its lack of engagement with the gendered and racialized nature of global commodity chains. Women constitute the majority of workers in many overseas subcontractors. Contemporary commodity chains often exploit differences in gender, race, ethnicity, nationality, religion, sexuality, age, and citizenship. They rely on superexploitation: exploitation that depends on so-called noneconomic factors, such as gender, race, ethnicity, and nationality (Tsing 2009). Despite the fact that the exploitation of women has been a key component of global commodity chains, Palpacuer (2008, 402) argues that “mainstream GCC/GVC research has barely ventured into the world of women workers.” Ramamurthy (2004, 741) similarly criticizes what she views as a strongly masculinist tone characterizing many GCC accounts that refer to chain “drivers.” She argues that a feminist commodity chain analysis would begin “with the possibility that global commodity chains – as connections across times and places – are neither linear nor unidirectional nor closed” (Ramamurthy 2004, 741). In parallel with earlier criticisms, she advocates an approach to commodity chains that recognizes their open-endedness, contingency, and rupture (Ramamurthy 2004, 743).

The future of GCC/GVC research and politics

The global commodity/value chain approach has proven tremendously useful in the last several decades as a vehicle for understanding processes of globalization and the growing international division of labor. Using this approach, researchers now have a better sense of the unique characteristics of different types of chains, and how power is exercised within them. GCC/GVC approaches have given us an understanding of what different types of governance mean for the possibilities of upgrading in developing economies, and provided valuable insights for policymakers and activists tackling the growing problems of inequality. Despite the many criticisms that have been advanced, there are important contributions in GCC/GVC analysis that can be built on and enriched. Greater
attention to culture, regulation, institutions, and the structural dynamics of capitalism, combined with a more fluid, open-ended approach to power and space, can help to illuminate the complexity in global commodity chains.

Tsing (2009) reminds us that commodity chains are constantly contested and shifting. This makes them unpredictable. She suggests that this new form of capitalism – based on globally extensive supply chains – is not as all-powerful as it seems. Almost by definition, it is dispersed and disorganized. Although lead firms claim to be in control of the flow of products and profits through their complex logistics systems, “no one can fully keep track of the activities of every firm in the chain” (Tsing 2009, 172). Commodity chains are characterized by constant scandals which, while not destroying the chain, provide openings for critique and mobilization (Tsing 2009, 172). These opportunities need to be better exploited. As Tsing (2009) points out, this requires greater sensitivity to the diversity of practices within global commodity chains. It also requires attention to the intersecting dynamics of gender, race, nationality, and class through which commodity chain exploitation is possible.

SEE ALSO: Brands and branding; Corporations and global trade; Global production networks; Industrial upgrading; International division of labor; Internationalization; Labor geography; Outsourcing

References


GLOBAL COMMODITY/VALUE CHAINS


Further reading


Global dimming/brightening

Martin Wild  
*ETH Zurich, Institute for Atmospheric and Climate Science, Switzerland*

Solar radiation (sunlight) incident at the Earth’s surface is the ultimate energy source for life on the planet, and largely determines the climatic conditions of our habitats. The amount of solar energy reaching the surface is a major component of the surface energy balance and governs a large number of diverse surface processes, such as evaporation and the associated hydrological cycle, snow and glacier melt, plant photosynthesis, and terrestrial carbon uptake, as well as the evolution of surface temperatures on diurnal, seasonal, and decadal timescales. It has also major practical implications, for example, for the planning and management of the rapidly growing number of solar power plants in support of the world’s pressing demands on nuclear- and carbon-free energy sources, as well as for agricultural production. Changes in the amount of solar energy reaching the Earth’s surface can therefore have profound environmental, societal, and economic implications (Wild 2012).

Observational and modeling studies emerging in the past two decades indeed suggest that surface solar radiation (in the following referred to as SSR) is not necessarily constant on decadal timescales, as often assumed for simplicity and lack of better knowledge, but shows substantial decadal variations. Largely unnoticed over a decade or more, this evidence recently gained a rapid growth of attention under the popular expressions “global dimming” and “brightening,” which refer to a decadal decrease and increase in SSR, respectively.

Evidence and causes of dimming and brightening

Monitoring of SSR began in the early twentieth century at selected locations and since the mid-century on a more widespread basis. Many of these historic radiation measurements have been collected in the Global Energy Balance Archive (GEBA) at Eidgenössische Technische Hochschule Zurich (ETH Zurich) and in the World Radiation Data Centre (WRDC) of the Main Geophysical Observatory in St Petersburg. In addition, more recently, high-quality surface radiation measurements, such as those from the Baseline Surface Radiation Network (BSRN) and from the Atmospheric Radiation Measurement Program (ARM), have become available. These latter networks measure surface radiative fluxes at the highest possible accuracy with well-defined and calibrated state-of-the-art instrumentation at selected sites worldwide.

Changes in SSR from the beginning of widespread measurements in the 1950s up to the 1980s have been analyzed in numerous studies (see Wild 2009a for a review). These studies report a general decrease of SSR at widespread locations over land surfaces between the 1950s and 1980s. This phenomenon is now popularly known as “global dimming” (see Figure 1, left-hand panel, for a schematic illustration).

Increasing air pollution and an associated increase in aerosol concentrations are considered a major cause of the observed decline of SSR...
Changes in cloud amount and optical properties, which may or may not have been linked to the aerosol changes, have also been proposed to contribute to the dimming. An attempt has been made in Norris and Wild (2007) to differentiate between aerosol and cloud impacts on radiative changes over Europe. They show that changes in cloud amount cannot explain the changes in SSR, pointing to direct and indirect aerosol effects as potential major causes of these variations. Alpert et al. (2005) found that the decline in SSR in the period from the 1950s to the 1980s is particularly large in areas with dense population, which also suggests a significant anthropogenic influence through air pollution and aerosols. Several studies (Wild 2009a and references therein) also noted a dimming over the same period at remote sites, suggesting that the phenomenon is not of a purely local nature and that air pollution may have far-reaching effects (a concept on how SSR in remote areas may be modulated by subtle changes in background aerosol levels is introduced below).

More recent studies using SSR records updated to the year 2000 found, however, a trend reversal and partial recovery at many of the observation sites since the 1980s. The term “brightening” was thereby coined to emphasize that the decline in SSR and associated global dimming did not continue after the 1980s (Wild et al. 2005) (Figure 1, right). Particularly in industrialized areas, the majority of the sites showed some recovery from prior dimming, or at least a
leveling off, between the 1980s and 2000. The brightening has been somewhat less coherent than the preceding dimming, with trend reversals at widespread locations, but still some regions showed continued decrease, such as in India (see Wild 2009a, 2012 for an overview). Brightening is not just found under all sky conditions, but often also under clear skies, pointing once more to aerosols as important contributors to this trend reversal (e.g., Norris and Wild 2007; Wild et al. 2005). The transition from decreasing to increasing SSR is in line with a similar shift in atmospheric clear sky transmission determined from pyrheliometer measurements at a number of sites (Figure 2). This transition is also in line with changes in aerosol and aerosol precursor emissions derived from historic emission inventories, which also show a distinct trend reversal during the 1980s, particularly in the industrialized regions. The trend reversal in aerosol emission towards a reduction and the associated increasing atmospheric transmission since the mid-1980s may be related to air pollution regulations and the collapse of the economy in Eastern European countries. A reduction of aerosol optical depth over the world oceans since 1990, which may be indicative of the global background aerosol level, was inferred from satellite data by Mishchenko et al. (2007). This fits well with the general picture of a widespread transition from dimming to brightening seen in the surface radiation observations at the same time.

Updates on the SSR evolution beyond the year 2000 show mixed tendencies. Overall, observed brightening is less distinct after 2000 compared to the 1990s at many sites. Brightening continues beyond 2000 at sites in Europe and the United States, but levels off at Japanese sites, and shows some indications for a renewed dimming in China after a phase of stabilization during the 1990s, while dimming persists throughout in India (Wild et al. 2009). On the other hand, the longest observational records, which go back to the 1920s and 1930s at a few sites in Europe, further indicate some brightening tendencies during the first half of the twentieth century, known as “early brightening” (Ohmura 2009; Wild 2009a). This is illustrated in Figure 3 with the longest available record of SSR, from Stockholm.

Wild (2009a, 2012) proposed a conceptual framework for the explanation of dimming and brightening, suggesting that aerosol-induced dimming and brightening can be amplified or dampened by aerosol–cloud interactions, depending on the prevailing pollution levels. In pristine regions, small changes in cloud condensation nuclei (CCNs) can have a much bigger impact on cloud characteristics than in polluted environments, because clouds show a nonlinear

Figure 2  Time series of annual mean atmospheric transmission under cloud-free conditions determined from pyrheliometer measurements at various sites in Russia (Moscow), Estonia (Tartu-Toravere and Tiirikoja), Switzerland (Payerne), and Japan (average of 14 sites). Wild et al. 2005. Reprinted with permission from AAAS.
(logarithmic) sensitivity to CCNs. Additional CCNs due to air pollution in pristine regions may therefore be particularly effective in increasing the formation, lifetime, and albedo of clouds, which all act towards a reduction of SSR through enhanced cloud shading. Thus, aerosol–cloud interactions in pristine environments may cause a strong amplification of dimming (brightening) trends induced by small increases (decreases) in aerosols. This implies that dimming/brightening could be substantial even in areas far away from pollution sources, where small changes in background aerosol levels induced by long-range transports can effectively alter SSR through cloud modifications. This mechanism potentially could also be responsible for the brightening over oceans, with decreasing aerosol background levels (Mishchenko et al. 2007) between the mid-1980s and 2000 consistently seen in satellite-derived SSR records (Wild 2009a and references therein).

In polluted regions, on the other hand, cloud microphysics effects tend to saturate with the logarithmic sensitivity to CCNs, whereas the direct extinction of SSR by aerosols, which increases proportionally to the aerosol loadings, becomes more relevant. Absorbing pollution layers further heat and stabilize the atmosphere, and attenuate SSR and related surface evaporation. This generally leads to a suppression of convective cloud formation, and dissolves clouds in layers heated by absorbing aerosol (known as the semidirect aerosol effect). The associated reduced cloud shading may partly counteract the aerosol-induced reduction of SSR in heavily polluted areas. Thus, in contrast to pristine areas, aerosol–cloud interactions may tend to dampen...
dimming/brightening trends induced by direct aerosol effects (Wild 2009a, 2012).

Environmental implications of dimming and brightening

A growing number of studies provide evidence that the variations in SSR have a considerable impact on climate and environmental change (see Wild 2009a and 2012 for a review). Wild, Ohmura, and Makowski (2007) investigated the impact of dimming and brightening on global warming. They present evidence that SSR dimming was effective in masking and suppressing greenhouse warming, but only up to the mid-1980s, when dimming gradually transformed into brightening. Since then, the uncovered greenhouse effect has revealed its full dimension, as manifested in a rapid temperature rise (+0.38 °C/decade over land since the mid-1980s). Wild, Ohmura, and Makowski (2007) provided further evidence for a strong impact of dimming and brightening on global warming, based on analyses of decadal changes in observed daily maximum and minimum temperatures and related diurnal temperature range (DTR). They found a slight decrease in the daily maximum temperatures over the “dimming” period from the 1960s to the 1980s, which is in line with the reduction of SSR over this period. Daily minimum temperatures, on the other hand, have been increasing over the same period, as expected from an increasing greenhouse effect, causing a decline in DTR. In contrast, since the mid-1980s, observed minimum and maximum temperatures have increased at a very similar pace. This is indicative of a major change in the surface radiative forcing regime, where the maximum temperature rise was no longer suppressed as in the previous decades, due to the transition from dimming to brightening (Wild, Ohmura, and Makowski 2007).

The climate models used in recent assessments of the Intergovernmental Panel on Climate Change (IPCC) do not seem to fully account for the observed decadal modulations of global warming and DTR through solar dimming and brightening, even if all known forcings are included in the models (Wild 2009b). They underestimate the suppression of warming through dimming between the 1950s and 1980s over global land surfaces, and simulate a higher rate of temperature rise than observed during this period. On the other hand, for the more recent decades, the models underestimate the rate of global warming, indicative of a lack of brightening. Also, the trend reversal in DTR in the 1980s is not reproduced by the climate models (Wild 2009b). Indeed, numerous studies show that climate models seem to underestimate SSR dimming and brightening when compared to the direct surface observations. The potential explanations that have been put forward for these modeling issues include the large uncertainties in historic emission inventories and associated aerosol burdens in the atmosphere, which may not consider the full extent of decadal variations, the inadequate representation of the direct and indirect aerosol forcings, and a general underestimation of the unforced natural variability in the models.

More recently, Wild (2012) pointed out that the absence of global warming from the 1950s to 1980s, and the subsequent reversal into rapid warming, was most prominently seen in the Northern Hemisphere, while in the Southern Hemisphere a steady gradual warming was observed from the 1950s onward (Figure 4). This fits well with the asymmetric hemispheric evolution of anthropogenic air pollution, which strongly increased from the 1950s to the 1980s and declined thereafter in
the Northern Hemisphere with increasing air quality measures, while pollution levels in the Southern Hemisphere were an order of magnitude lower and steadily increased with no trend reversal (Wild 2012). This again points to a possible large-scale influence of aerosol-induced SSR dimming and brightening on global warming. Climate models, on the other hand, do not reproduce these characteristic differences in inter-hemispheric warming (Wild 2012). In the Northern Hemisphere, with considerable aerosol loadings, the average warming simulated by the models in twentieth-century “all forcings” experiments was too strong during dimming, and too weak during subsequent brightening as noted above for the global scale. This points again to an insufficient representation of the processes causing dimming and brightening in the models to dampen and enhance greenhouse warming, respectively. In contrast, in the Southern Hemisphere, where aerosol pollution is much smaller and greenhouse-gas forcing dominates, the models perform very well, with warming rates close to those observed (Wild 2012). This suggests that climate models simulate decadal warming trends adequately when greenhouse gases act as sole major anthropogenic forcing, as in the Southern Hemisphere, but may have difficulties when in addition strong decadal aerosol variations come into play, as in the Northern Hemisphere.

Interestingly enough, the observed suppression of warming during the dimming period in the Northern Hemisphere was even slightly stronger over ocean than over land areas (Figure 5), with a slight cooling of $-0.03\, ^\circ\text{C}$ per decade over oceans between 1958 and 1985, compared to a slight warming over land with $+0.04\, ^\circ\text{C}$ per decade over the same period, based on data from the Climate Research Unit, Norwich, and the Hadley Centre, Exeter, United Kingdom (Wild 2013). Even though anthropogenic air pollution sources are located over land, subtle changes in background aerosol levels over the relatively pristine oceans could have amplified SSR trends through effective cloud–aerosol interactions, as outlined.
Figure 5  Annual 2-m temperature anomalies observed over Northern Hemispheric oceans (left) and land surfaces (right). Observations from CRUTEM3 (land) and HADSST2 (oceans), anomalies with respect to 1960–1990. Linear trends over the dimming phase (1950s–1980s) in blue, over the brightening phase (1980s–2000) in red. Reproduced from Wild 2013 with the permission of AIP Publishing LLC.

in the conceptual framework above. This may explain the lack of warming particularly over oceans during this period (Wild 2012, 2013).

It is interesting to note that the indication for a slowdown of the brightening since 2000 noted earlier (Wild et al. 2009), and the possible reversal into a renewed dimming, coincides with a lack of warming since the turn of the millennium (“global warming hiatus”). Such a renewed reversal in SSR tendencies may be favored by factors such as the renewed increase in global sulfur emissions after 2000 (particularly originating from Asia) (Streets et al. 2009) or enhanced volcanic activities during this period (e.g., Santer et al. 2014).

SSR is also a major determinant of surface evaporation and thereby the main driver of the global water cycle. Wild et al. (2004) suggested that surface solar energy reductions outweighed the increasing thermal energy from the greenhouse effect from the 1960s to the 1980s, resulting in a reduction of surface net radiation and associated evaporation over land surfaces, causing an attenuation of the intensity of the associated water cycle (Figures 1 and 6). In contrast, for the more recent period of the 1980s to the 2000s, Wild, Grieser, and Schaer (2008) pointed out that SSR brightening adds to the increasing energy from the enhanced greenhouse effect, leading to higher evaporation and an intensification of the global terrestrial water cycle since the 1980s (Figures 1 and 6). Impacts of the transition from dimming to brightening can further be seen in the more rapid retreats of glaciers and snow cover, which became evident after the 1980s as soon as the dimming disappeared (Wild 2009a and references therein).

Modeling studies further suggest that SSR dimming/brightening may also impact the terrestrial carbon cycle and plant growth (Mercado et al. 2009). During dimming, plant photosynthesis and associated terrestrial carbon uptake might have been enhanced despite a reduction in SSR, since the stronger aerosol and cloud scattering enlarged the diffuse radiative fraction in this period. Diffuse light penetrates deeper into the vegetation canopies than direct sunbeams and can therefore be more effectively used by plants for photosynthesis.

Further research will be required to establish the full dimension of the impacts of SSR.
GLOBAL DIMMING/BRIGHTENING


Dimming and brightening on climate and environmental change. Research on the impacts of dimming and brightening may also provide additional insight into the potential consequences of geoengineering, as the proposed actions predominantly target the modulation of incoming solar radiation.

SEE ALSO: Atmospheric aerosols; Climate adaptation/mitigation; Climate change, concept of; Climatology; Environmental science

References


Global Dimming/Brightening


Global environmental change: human dimensions

Robin Leichenko
David C. Eisenhauer
Rutgers University, USA

Research on the human dimensions of global environmental change (HDGEC) encompasses a broad array of social science research on processes and outcomes of human-induced environmental changes. The changes include, for example, climate change resulting from rising levels of atmospheric greenhouse gases, stratospheric ozone depletion, urbanization and land-use change, and loss of biodiversity. Local-scale changes that occur on a worldwide scale, such as consumptive water usage and degradation of water quality, are also recognized as global environmental changes because of the scope and magnitude of their impact. The concept of the “Anthropocene,” which suggests that human activity has become a dominant factor changing natural, life-supporting processes, is also central to research on global environmental change.

As an interdisciplinary area of study, HDGEC research draws upon fields including climate change impacts, land-use change, natural hazards, political ecology, energy, and livelihoods, among many others, to explore societal causes, impacts, and responses to large-scale environmental changes. Human dimensions work has long found a home in major geography journals, such as the Annals of the Association of American Geographers, as well as in established interdisciplinary journals, such as Global Environmental Change, Mitigation and Adaptation Strategies for Global Change, Ambio, Environment, Climatic Change, Land Use Policy, and Natural Hazards. Over the past decade, a plethora of new journals with space devoted to the human and policy dimensions of global environmental change, often with a primary emphasis on the topic of climate change, have emerged. Some of the many examples of recently or newly established journals in this area include Weather Climate and Society, Climate and Development, Nature Climate Change, WIREs Climate Change, Urban Climate, Change and Adaptation in Socio-ecological Systems, Journal of Extreme Events, and The Anthropocene Review.

Along with a growing number of journal outlets for HDGEC research, numerous international research networks and programs have been formed in order to connect HDGEC researchers and to promote and carry out collaborative HDGEC projects. These networks take many forms and are sponsored or cosponsored by numerous international scientific organizations and research programs, such as the International Council for Science (ICSU) and the International Social Science Council (ISSC). Of particular importance for the establishment and promotion of HDGEC work over the past two decades has been the International Human Dimensions Research Programme (IHDP) (www.ihdp.unu.edu). Founded in 1996, IHDP was instrumental in the formation and establishment of an interdisciplinary HDGEC research community through hosting of international open meetings and through the many research projects that it sponsored. These projects were generally 10-year efforts designed by leading HDGEC scientists to facilitate and promote research on cutting-edge human dimensions.

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0649
topics such as human security, urbanization, land-use change, governance, the carbon cycle, and water systems. IHDP closed in 2014 and all of its functions were incorporated into the newly formed Future Earth Initiative (www.futureearth.info). Future Earth also incorporated a number of other major international programs on global environmental change, including DIVERSITAS (www.diversitas-international.org), the International Geosphere-Biosphere Programme (IGBP) (www.igbp.net), and the World Climate Research Programme (WCRP) (wcrp-climate.org). While IHDP had an explicit focus on social science, Future Earth also includes research within the physical sciences, engineering, humanities, law, and other areas. This integration across many fields of science reflects wider acceptance of the value and prominence of the “human” in global environmental change.

One of the major goals of IHDP and other international scientific networks over the past two decades has been the diversification of the pool of scientists involved in HDGEC work. The Intergovernmental Panel on Climate Change (IPCC), for example, has worked to include scientists from all regions of the world. By incorporating a variety of voices across geographic regions, the generation of scientific knowledge reflects a broader array of cultural, political, and institutional viewpoints and includes more information gathered from disparate geographic locations.

The diversity of research outlets, networks, and participants that support human dimensions work is matched or perhaps exceeded by the theoretical and epistemological diversity of human dimensions research. As discussed next, the scope of the field can be broadly captured through examination of some of the major discourses that are implicit in the work. Following the discussion of HDGEC discourses, key issues and topics addressed in HDGEC work are discussed, including: (i) social, economic, political, and cultural drivers of global environmental change; (ii) impacts of and vulnerability to these changes within different sectors, regions, communities, and social groups; and (iii) societal responses, including adaptation, mitigation, and social transformation. In addition to shared discourses, research across these areas is also connected via a number of unifying and cross-cutting themes. Several of these themes are highlighted in the final section, including governance, equity and social justice, and the social production and communication of scientific information.

Discourses on global environmental change

Definitions and approaches to research on the human dimensions of global environmental change are embedded in a number of competing discourses. As described in Leichenko and O’Brien (2008), discourses entail underlying assumptions, values, and judgments, which provide the basic terms for analyses and debates, influencing both agreements and disagreements. Discourses also carry political weight such that proponents of some discourses exert a greater influence over how specific issues are framed and on which responses are prioritized (Liverman 2009). Discourses influence definitions of global environmental change and the approaches used to evaluate and address these changes. Leichenko and O’Brien (2008) organize the discourses on global environmental change into three broad and sometimes overlapping categories: biophysical discourses, human-environment discourses, and critical discourses. Recognition of these multiple and competing discourses helps to explain the breadth and diversity of HDGEC research approaches.
Biophysical approaches tend to frame global environmental changes as biophysically based processes that pose challenges for sustainability of human systems and need to be addressed through a combination of technological innovations and international commitments to reduce, limit, or suspend activities that contribute to these changes. Research within the biophysical discourse focuses on empirical documentation of physical changes that may influence biogeochemical cycles, atmosphere–biosphere–ocean interactions, and other natural and physical processes. Physical science–based mathematical models are often used to develop future projections of temperature change, sea level rise, acidification of precipitation, land-use change, and so forth. Within this discourse, better understanding of Earth system processes and how these processes are influenced by human activities, in combination with better communication of scientific findings, is seen as vital to addressing global environmental change. The biophysical discourse is based on an underlying belief that advancement of scientific knowledge about environmental processes will reduce technical uncertainties such that a social consensus on actions and responses will follow (Leichenko and O’Brien 2008).

Researchers working within human–environment discourses emphasize the integral connections between human and natural systems and stress the importance of connecting social and biophysical processes in order to understand drivers, impacts, and vulnerabilities. Research approaches emphasize the importance of integrating both biophysical and social processes in order to understand how outcomes of environmental change are distributed within society. Quantitative, qualitative, and mixed-method approaches are used in this research in order to explore topics such as social and economic vulnerabilities, livelihoods systems, resilience of coupled socioecological systems, adaptation, and social learning. Like the biophysical discourse, the human–environment discourse also subscribes to the idea that better scientific understanding of the drivers, processes, outcomes, and responses associated with global environmental change is needed (Leichenko and O’Brien 2008). However, the human–environment discourse increasingly emphasizes the importance of coproduction of scientific information about global environmental changes through collaborative engagement between scientists, governments, private sector actors, and members of civil society (Cornell et al. 2013).

Critical scholars often go further than the human–environment discourse by seeking to locate the root causes of environmental vulnerability and marginalization. A core tenet of this work is that both local and broader inequalities must be addressed in order to reduce environmental degradation (Peet, Robbins, and Watts 2011). Traditionally, critical researchers framed their work within “neo-Marxist” perspectives. In doing so, they sought to explicate causal chains of explanation located within structural forces, such as the global capitalist economy, coercive power relations, and unequal access to resources and institutions. Recently, critical approaches have broadened to include fields such as poststructuralism, third-wave feminism, and science and technology studies. This expansion has contributed to questioning of many of the causal assumptions of previous critical research and has opened up new avenues of research exploring the ways in which individuals change the way they see themselves (i.e., their subjectivity) in relation to environmental change and how the circulation of power contributes to this process. Critical environmental scholarship often unfolds within field research that seeks to detail the ways in which unequal power dynamics contribute to environmental change on the
ground. Through empirical case studies critical scholars highlight broad forces and the frequently technical framing of environmental change and they demonstrate how local capacity to manage resources can, under some circumstances, be more sustainable and equitable than reliance on centralized governments or market forces.

Societal drivers of global environmental change

Human actions are generally understood to be a driving force behind global environmental change, whether through population growth, resource extraction, energy consumption, urbanization, technological change, changes in consumer demands, or shifts in lifestyles, attitudes, and beliefs (Steffen et al. 2004).

Research on the drivers of global environmental change documents how human activities contribute to systemic changes that affect the entire Earth system, such as increasing atmospheric concentrations of greenhouse gases and stratospheric ozone depletion. Examples of systemic-focused research include studies of projected future carbon emissions associated with continued consumption of fossil fuels including oil, coal, and natural gas, and studies of the implementation of international regulations of chlorofluorocarbons and the consequences for stratospheric ozone levels. This work also entails study of the role of human activities in contributing to local-scale changes that cumulatively influence the global environment through the magnitude and distribution of their effects. Research that emphasizes the cumulative effects of local-scale changes draws attention to activities that are occurring on a worldwide basis, such as rising levels of meat consumption, rapid urban spatial expansion, and overfishing of large predatory species, and explores how these activities affect global land and water usage patterns, species survival rates, and so forth.

Biophysically based studies of the human drivers sometimes involve empirical documentation of global-scale changes based on remote sensing analysis of changing land-use patterns or statistical analysis of temperature records. The work may also entail development of future projections of environmental change based on social factors such as population growth, changing energy consumption, and growth of per capita income, as well as various policy responses. The economic or social causes of cumulative local environmental changes are also emphasized, such as the legal, policy, and economic causes of over-exploitation of groundwater for irrigated agriculture. Research also explores topics such as the influence of individual attitudes and beliefs on energy consumption and lifestyle choices, and the consequences of these decisions for energy use patterns. Critical work on this topic examines the connections between environmental change and longer-term expansion of the capitalist economic system. This work questions the implicit logic of continuous economic growth and capital accumulation and emphasizes the unequal political and economic power relations that produce environmental degradation.

Societal impacts of global environmental change

Human dimensions researchers also investigate the consequences of global environmental change for human and coupled social-ecological systems at a variety of scales. This work emphasizes topics including impacts, risks, vulnerability, and resilience. Impact and risk-based research typically focuses on how global environmental changes, such as higher temperatures, reduced precipitation or how more frequent extreme
events associated with climate change may affect specific sectors of an economy, such as real estate, agriculture, energy, transportation, health care, or tourism. Examples include documentation of populations and property at risk to sea level rise, impacts of water supply changes on crop production, effects of extreme events on infrastructure, costs of air pollution on asthma treatment, and impacts of coral bleaching on numbers of visitors to a tourist destination. Impact and risk research may also adopt a regional focus, looking at how a political, economic, or ecological region, such as a coastal city or agricultural area, may be affected by global environmental changes, again with an emphasis on the prediction and quantification of impacts and risks. In the case of a transportation system in an urban coastal region, for example, a study of climate change impacts and risks would identify key types of climatic impacts likely to affect the system, assign probabilities to these impacts (e.g., probability of a certain amount of sea level rise by 2050), and estimate potential economic costs.

Vulnerability and resilience studies explore factors that explain why some regions, systems, communities, and individuals are more likely than others to be harmed by global environmental change or are less able to recover and bounce back from associated shocks and stresses. Vulnerability studies, which developed partly from the hazards tradition in geography, explore underlying social, political, institutional, and biophysical factors that make some more likely than others to be harmed by climate and environmental change. The factors seen as most important typically depend upon the discourse that the researcher is working within (Leichenko and O’Brien 2008). For example, critical researchers will often point toward economic, political, and social inequalities that create susceptibility to harm; whereas, biophysical-oriented researchers will emphasize the degree of physical exposure to climatic or environmental change. Resilience work is more directly concerned with how communities and regions respond to shocks and stresses and the identification of actions that enhance the ability to recover. The concept of resilience developed separately within the fields of ecology, hazards studies, engineering, and psychology. While HDGEC works has drawn largely from ecological and hazard-based understandings of resiliency, these other fields are increasingly influencing research in this area.

A common point of departure for both vulnerability and resilience studies is emphasis on the importance of local and regional contextual factors in shaping options for response. These contextual factors include economic, social, cultural, political, and institutional conditions that either constrain or enable effective responses to various types of environmental stresses. Wealth, education levels, social capital, access to political power, and worldviews are among the many attributes that affect these responses. Both approaches focus on how society anticipates and plans for future changes including the factors that constrain or enable people to pursue outcomes that they value.

In addition to emphases on the capacity to respond to single processes of global environmental change, researchers also recognize that environmental stresses and shocks typically originate from multiple sources. Multiple stressor approaches recognize that individuals, communities, and regions are often subject to more than one ongoing process of global change. For a single region, the effects of climate change, rapid urbanization, globalization, and the spread of infectious disease may be felt simultaneously. Acknowledgment of multiple stressors is seen as key to understanding differential exposure, and multiple stressor approaches typically combine investigation of vulnerability to changing environmental conditions with examination of other
stressors such as migration, warfare or conflict, and economic recession. Expanding the analyses beyond “single stressor” approaches has also provided important insights into social processes that underlie the drivers of global environmental change and shape vulnerability and responses (Leichenko and O’Brien 2008). Multiple stressor frameworks are especially prevalent in research on global environmental change and human security (Sygna, O’Brien, and Wolf 2013). This work draws attention to ways in which environmental change acts as a threat multiplier for many factors that contribute to human insecurity including worsening poverty, unplanned migration, civil conflict, and loss of cultural identity.

Societal responses to global environmental change

Societal responses to global environmental change are also a major area of work for the HDGEC community. Adaptation research explores approaches and strategies to respond to stresses and hazards associated with global environmental change. The potential actors involved in adaptation include individuals, households, governments, organizations, and businesses. These actors operate at different political levels – ranging from local to international – and within a wide variety of contexts and geographic locations. The differences between and within these groups make adaptation a significant political challenge (Pelling 2011). Further, individuals and groups do not experience environmental change in isolation; rather, they experience the effects of global environmental change along with other changes in their social and economic lives. Efforts to adapt to environmental change are classified along a variety of dimensions. Adaptations can occur proactively to reduce foreseen vulnerability. Farmers can diversify their crops or purchase insurance to lessen their vulnerability to drought and/or flooding, for instance. Adaptation also can occur retroactively based on experience of past hazards. For example, city planners may build levees after a significant flood event. Central authorities or governments can plan adaptations or individuals and communities can autonomously develop them. Adaptation efforts can aim to spread or reduce risk to environmental change. Additionally, adaptation can aim to build resilience within a system so that it can maintain its structure and function, transition to a slightly different state in which the basic structure is maintained, or transform the system into a new form (Pelling 2011).

Researchers also recognize that adaptation requires resources and capabilities, which are often encompassed within the notion of adaptive capacity. Adaptive capacity refers to the store of material, social, and ecological capacities available to respond to climate change. This capacity is latent – meaning it is difficult to measure. Further, not all capacity is useful in every potential scenario. For this reason, HDGEC research distinguishes between specific adaptive capacity and generic adaptive capacity (Eakin, Lemos, and Nelson 2014). Individuals and groups can mobilize specific adaptive capacity to deal with a narrow range of hazards and stresses. For instance, investing in emergency centers can increase a community’s specific adaptive capacity to hurricanes or typhoons. Generic adaptive capacity is useful for a broader spectrum of situations. Economic development or education, for example, can aid people in adapting to many forms of environmental change. However, there are also recognized barriers and limits to each type of adaptation. A barrier to adaptation presents an obstacle that likely must be overcome. For example, laws, tax policies, insurance subsidies, and cultural preferences all contribute to building homes in risky coastal areas. While
legal and social barriers can potentially be overcome, limits to adaptation are more immutable. If sea levels increase too much, some coastal areas will become uninhabitable. Thus, sea level may act as a strict limit to adaptation in some cases.

Within the HDGEC research community, a “pathways” framing of adaptation action has become increasingly common. Across venues ranging from the IPCC to local governments to academic research centers, adaptation pathways are seen as flexible policy frameworks in the face of uncertain and rapid global environmental change. Broadly, pathways are alternative avenues to possible futures. Designing and implementing pathways can take a variety of forms. Back-casting involves describing desirable futures and designing pathways that could get society there. Alternatively, pathways can be designed to respond to potential environmental changes. That is, crossing a threshold or tipping point triggers a shift to a different pre-planned regime. Overall, pathways aim to be flexible, open to new information, and applicable within diverse contexts (Leach, Scoones, and Stirling 2010). Climate adaptation pathways entail designing alternative paths to maintaining a desirable state within a changing climate. Within this design process, it is necessary to define what is desirable and constantly monitor change and results. When the state of the system passes a threshold or moves into an undesirable condition, such as experiencing repeated damage from flooding, there is need for system transition to a new path. Knowing when and how to choose a new path depends on the social tolerance of risk and the ability to monitor changes in hazards and vulnerability (Pelling 2011).

Mitigation research emphasizes behavioral changes, policy measures, and technological responses that are intended to slow or halt processes of global environmental change. This differs from the usage of the term “mitigation” in hazards research where it is associated with local hazard and risk reduction. Mitigation work is often tied directly to the issue of sustainability with the notion that mitigation is needed in order to ensure that the activities of the present generation do not undermine the ability of future generations to meet their needs. Mitigation research typically focuses on ameliorating the causes of climate change through reduction of greenhouse gas emissions from energy production, transportation, buildings, construction, agriculture, and other sectors. Mitigation research is also directed to land-use issues that relate to climate change, particularly slowing the pace of tropical deforestation. Prominent topics for policy-focused mitigation research include carbon pricing, emissions trading, carbon taxes, and green development mechanisms. Mitigation research also investigates the feasibility of alternative energy technologies and the implications of shifts to alternative energy for job creation and destruction. Other types of technological solutions to climate change, including geo-engineering measures, are also considered in the HDGEC literature, though the primary emphasis of that work is more typically on political and ethical concerns surrounding deployment of such technologies. Mitigation research also explores local political organizing around global environmental issues such as urban-based efforts to reduce carbon emissions, which have been adopted by many hundreds of cities worldwide.

Mitigation pathways share many common features with adaptation pathways as they aim to describe alternative ways to reduce the impacts of climate and environmental change. However, they do so by seeking to preclude negative change from occurring by reducing greenhouse gas emissions. Mitigation pathways describe
possible carbon emission trajectories and usually are compared to a “business as usual” scenario, which assumes that no mitigation occurs. Targets, or future atmospheric greenhouse gas concentrations, typically seek to stabilize the global climate system at a state in which both the economic and social costs and benefits of decreasing emissions are balanced. Once targets are identified, pathways to achieve these goals are built. These pathways include a variety of mechanisms to achieve emissions reductions, such as energy efficiency, carbon trading, reforestation, and technology development. Mitigation pathways seek to be flexible so as to be able to change in light of new information.

Transformation is an emergent concept within HDGEC research. Though there is no universal definition, transformation generally refers to a significant change in a system that alters its basic functioning. In other words, transformation entails moving a system to a new state. Whereas coping, adaptation, and mitigation efforts tend to focus on dealing with the effects of environmental change, transformation aims to address the root causes of negative change and vulnerability found within social, political, economic, and ecological systems (Pelling 2011). By doing so, transformation incorporates aspects of both mitigation and adaptation. For a growing number of HDGEC researchers, the inevitability of and necessity for transformation is clear. Transformation is inevitable because society faces significant environmental changes. It is necessary because people will need to reconceptualize and alter the ways that they manage and plan in relation to the environment. Transformation ultimately requires challenging basic understandings of the state of the world and humans’ place within it. Transformation not only entails making significant alterations to the physical world but also includes fundamental, societal change (Pelling 2011).

Cross-cutting themes

There are many topics of cross-cutting interest that are relevant to all facets of HDGEC research. Some of the most prominent of these cross-cutting themes include governance, equity and social justice, and the social production and communication of scientific information.

Research on governance focuses on institutions, policies, and regulatory frameworks that influence how resources are managed. Within HDGEC research, much attention has been directed to examination of how neoliberal (i.e., free market) approaches, such as privatization of common lands, influence patterns and rates of environmental changes, such as deforestation and loss of biodiversity. This work also explores how different types of institutional arrangements affect the vulnerability and resilience of local communities and examines the potential of various governance mechanisms to promote adaptation. Human dimensions researchers explicitly recognize that efforts to both adapt and mitigate the causes of global environmental change may require new types of governance structures or new types of cooperative agreements. Some of the many characteristics of governing institutions that are identified as promoting effective responses to environmental change include: polycentricity, transparency, accountability, flexibility, and inclusiveness. But rather than prescribing a single, “best practice” arrangement, the environmental governance literature advocates a diversity of approaches, suggesting that effective institutional arrangements take many different forms (Ostrom 2010).

Issues of equity and justice are another prominent theme in HDGEC research. Questions of equity and justice frequently arise regarding mitigation of climate change, particularly with respect to who will pay the costs for efforts to
reduce emissions and who will bear the economic burdens associated with these reductions. Issues of temporal equity also come into play in these discussions. Most of the greenhouse gases above preindustrial levels that are currently present in the atmosphere are the result of energy usage and other economic activities that have occurred within developed countries, such as the United States, Australia, and the countries of the European Union. However, a subset of developing countries, including China and India, now account for a growing share of total global greenhouse emissions. While per capita energy usage within developed countries remains substantially higher than in most developing countries, per capita usage is rising within developing countries, as growing numbers of middle-class residents adopt energy consumptive lifestyles. Issues of intergenerational equity are also relevant, including obligations to maintain a wide range of options and ensure quality of life for future generations. HDGEC researchers have also long recognized that the effects of environmental change are highly uneven and create winners and losers (O’Brien and Leichenko 2003). Equity and justice issues arise, in part, because the differential effects of environmental changes and of adaptations to those changes often harm groups that are already politically and economically marginalized.

HDGEC research increasingly acknowledges that science is not removed from social relations. Scientists work within cultural, institutional, political, and interpersonal arrangements that influence their work and its dissemination. This is not to say that science cannot be objective, but rather that scientists always have a situated perspective. The norms and rules of science, such as measurement standards, replication, and peer review, aim to create universal knowledge valid in all contexts. The transformation of situated knowledge into universal knowledge has become a topic of interest within the HDGEC research community. Following from this interest, several subtopics have recently emerged in the literature: uncertainty and the limits of scientific knowledge; participation in science; and communication of scientific knowledge.

Uncertainty and ambiguity permeate many global environmental changes – this is especially true when predicting future changes (Hulme 2009). For example, many uncertainties exist within the study of climate change, including the sensitivity of the climate to increased emissions, how deep sea currents will respond to warmer temperatures, and the ways weather patterns will change. Additionally, there is uncertainty about how individual components of the complex and chaotic atmosphere, ocean, and terrestrial systems will interact as the climate changes. Finally, the role of society in mitigating, adapting, and further altering each system is uncertain. All of these uncertainties compound each other, limiting what scientists can definitively conclude about Earth systems and global environmental change (Hulme 2009). Scientists and society at large can, however, ameliorate some effects of this uncertainty through opening up the knowledge creation process and clearly communicating uncertainty.

HDGEC research is increasingly committed to productive engagement between scientists, policymakers, and the public in order to assess what these changes mean for society and to identify feasible solutions. Particular effort has been made to include nonscientists in the production of scientific knowledge (Whatmore 2013). This approach, known as the coproduction of knowledge, recognizes that the broader public has valuable insights into topics in which scientists are concerned. Within this model, all forms of knowledge are treated equally and a shared understanding of the problem is not assumed to exist. Rather, both scientists and the public
interact, debate, and collaborate to enrich each other’s knowledge about the world. This does not eliminate uncertainty. Rather, it treats uncertainty as fertile ground for exploring a problem. This approach is especially useful within controversial situations (Whatmore 2013). During and after a controversy, previous beliefs and assumptions regarding the world are challenged. For instance, a surprising flood might upset people’s beliefs in the behavior of the weather, settlement patterns, and flood control measures (Whatmore 2013), leading to a rapid increase in uncertainty. Coproducing a solution to an environmental controversy has the potential to combine both scientific knowledge and local knowledge in unique and unexpected ways that can be more effective than purely technical solutions.

Additionally, a great deal of recent effort has been spent on communicating both scientific knowledge and uncertainty of global environmental change (Boykoff 2011; Hulme 2009; Liverman 2009). This especially holds for climate change science. It is often assumed that scientists should communicate what they are certain about and couch their language within scientific models and terminology. On the one hand, this approach brings to the fore the valuable information scientists provide for understanding a problem and weakens skeptical claims of scientists’ biases. On the other hand, it also tends to decontextualize global environmental change and detach it from people’s experience and concerns (Boykoff 2011; Hulme 2009). Alternative and complementary approaches to communicating scientific knowledge seek to connect abstract science with local concerns and understandings. By using symbols and examples that the general public understands and relates to, scientists, scientific organizations, and others can contextualize global environmental change in ways that strengthen their ability to increase awareness of scientific knowledge (Boykoff 2011). Nevertheless, communicating what science does and does not know about global environmental change remains a challenging endeavor.

Beyond communicating the science and uncertainty, many researchers in the HDGEC have begun interrogating the persistent gap between knowledge about the dynamics of global environmental change and action by governments at all scales (Castree et al. 2014). Two emerging and related explanations for this gap are the limited conceptualization and integration of environmental social sciences and humanities within global environmental change science and knowledge and the overwhelming tendency to frame climate and environmental changes as physical, rather than social, problems (Castree et al. 2014; Hackmann, Moser, and St Clair 2014). Therefore, there have been increasing calls for giving the social sciences and humanities more prominence within the production of global environmental change science and knowledge. These scholars argue that the environmental social sciences and humanities can help elucidate the various and alternative values, technical and political means, and normative ends at stake – or what Castree et al. (2014, 766) call “values–means–ends’ packages.” By doing so, HDGEC research would redefine climate and environmental change as inherently social problems rather than physical ones (Hackmann, Moser, and St Clair 2014), thus shifting the dominant framing of global environmental change from biophysical to critical and/or human–environment discourses.

SEE ALSO: Anthropocene and planetary boundaries; Climate change adaptation and social transformation; Climate and societal impacts; Intergovernmental Panel on Climate Change (IPCC); Land-use/cover change and climate; Political ecology; Sustainability science;
References


Further reading

A brave new borderless world?

The global factory emerged as a significant phenomenon in the last decades of the twentieth century, corresponding to the resurgence of free-market ideologies and neoliberal approaches to economic development. Characterized by a post-Fordist, “flexible” form of production in an ostensibly borderless world, its rapid rise to prominence marked a critical conjuncture in the global political economy.

By the 1980s, rising wages and labor-protective legislation prompted many Northern corporations to relocate manufacturing units out of the United States and Europe to Asia and Central and Latin America in search of cheap labor. Around the same time, international financial institutions began to actively endorse structural adjustment policies, “free trade,” and export-oriented industrialization as pathways to economic development. Southern nations were compelled to liberalize trade policies in order to compete in the new regime of globalization. Most industrializing countries set up tax-exempt special economic zones, also shielded from national labor laws, in order to attract foreign capital investment. In nationalist and development narratives, export-based industrial employment signified progress for the nation as well as for the individual worker.

The production of labor-intensive low-end products such as apparel, electronics, and toys, which require minimal investment and allow for maximum capital mobility, flourished in the newly restructured global economy. “Fly-by-night” factories emerged and folded, moving to wherever labor costs were lowest. This relentless search for increasing returns effectively cheapened labor, as Southern economies vied with one another to attract multinational corporations in what has been called a race to the bottom. States desperate to retain foreign direct investment actively, often brutally, suppressed worker resistance and protest.

In line with global trends, labor precarity – high turnover, job insecurity, and casualization – marked work in export factories. The global factory also brought about major shifts in the international division of labor. A predominantly female, Southern migrant labor force characterized the new global assembly line. As such, the emergent proletariat was raced, classed, and gendered in ways that reflected older colonial and imperial divisions of consumption and production, as well as of wealth and inequality.

Transnational production and the “disposable” woman from the Global South

The long hours, low wages, and precarious working conditions typical of export factories prompted some commentators to classify the latter as sweatshops. Indeed, in popular and left academic circles, the “third-world” female factory worker became synonymous with
GLOBAL FACTORY

sweatshop labor and the depredations of global capital (Siddiqi 2009).

Women workers have figured centrally in debates around the sweatshop, an early version of which grappled with the relationship between capitalism and patriarchy, and whether insertion into the transnational chain of production liberated women or enchained them further. Anchored in the valorization of the export-led development model, one view embraced factory work as a harbinger of modernity and emancipation. The expectation was that industrial employment would undermine patriarchal structures as well as improve everyday living conditions.

Other feminist scholars, especially those working within a Marxist-historical tradition, were less sanguine. Several key anthologies in the 1980s set the tone for subsequent debates on the feminization of the international division of labor and the merits of multinational production (Ehrenreich and Fuentes 1983; Nash and Fernandez-Kelly 1984). These authors were unequivocal in their view of the hyperexploitation of women factory workers from the Global South in the service of capital.

This body of research carefully unpacked the tropes of “nimble fingers” and “docility” used by multinationals and others to justify hiring women at low wages for industrial work. An inherent capacity for agility, patience, diligence, and compliance, the argument went, made women naturally suited for repetitive monotonous assembly-line work. Considered to be culturally docile, “Oriental” women made especially desirable employees. Biological and cultural discourses were layered with a sociological one that positioned women as secondary earners, who worked either to supplement the wages of a male head of household or to satisfy their consumer desires (“working for lipstick”). Put differently, the naturalization of women’s labor through tropes of docility and agility simultaneously cheapened it, masking exploitative working conditions, low wages, and, ultimately, the suppression of organized labor movements.

By the 1990s feminist scholars, especially those working with postcolonial, poststructural, or transnational theories, moved away from the binary framing of exploitation versus emancipation. They also challenged the depiction of all workers as voiceless, passive victims of capital. Building on Ong’s now classic ethnography of Malaysian electronics workers, scholars foregrounded the contradictory and inconsistent outcomes of factory work, the messiness or complexity of everyday social relations, and unconventional modes of resistance on the shop floor (Ong 2010/1987).

Localized ethnographies often yielded contradictory insights. Fernandez-Kelly’s (1984) ethnography of Ciudad Juarez contends that factory employment increases women’s dependence on patriarchal family relations. In contrast, in a compilation of ethnographic work from a variety of localities (Rothstein and Blim 1991), Jaffee (1991) shows that mothers in South Africa may be able to delay childbirth and obtain a measure of autonomy. Kim (1991) documents the incorporation of young women into the militant labor movement following their entry into the factories of South Korea’s export processing zones. In short, despite the standardization of the global production chain, the “local” meanings of the global factory are not readily generalizable. The impact on women and on gender relations is intimately related to prevailing social and economic contexts.

A host of anthropological studies complicate received notions of culture in relation to political economy. Family and kinship ties, for instance, are critical not only to the survival of households in the new economy but also in the mobilization
GLOBAL FACTORY

of capital and the recruitment of workers (Rothstein and Blim 1991).

Working on the assumption that the material and ideological are co-constitutive rather than separate spheres, scholars have mapped the ways kinship and other social relations reconstitute the production process even as the structures of transnational production shape— in highly contingent ways—gendered meanings, processes, and subjectivities. Other ethnographies trace the close entanglements of localized regimes of sexuality and the regulation of labor, as well as young migrant workers’ negotiations of kinship ties, desires, and new practices of consumption and self-making.

Recent ethnographies also revisit questions of agency exercised in extremely constrained circumstances. Drawing on Michel Foucault’s work on technologies of the self, Pun Ngai (2005) explores “minor genres” of resistance among Chinese workers producing goods for Walmart. Ngai argues that the agency of workers effectively limits the domination of managers and creates spaces of mutual support and even hope.

The proliferation of maquiladoras (factories on the United States–Mexican border) following the ratification of the North American Free Trade Agreement, and the subsequent violence enacted on women’s bodies, have generated some of the most theoretically provocative work on transnational production. Drawing on fieldwork in China and Mexico, Melissa Wright (2006) unpacks the myth of the disposable “third-world” woman as it circulates across the globe. This myth, she argues, consolidates the present regime of transnational capital even as the myth itself is remade across different scales and registers. Wright powerfully connects the wearing out of workers’ bodies or gendered “waste in the making” within the maquiladora to the murder of women outside factory walls. She also highlights the significance of worker resistance in nontraditional registers. Her analysis echoes Pun Ngai’s (2005, 11) call to reactivate social struggle “by rooting it in class experience from below – that is, in the everyday infrapolitics of the Chinese workers themselves.”

The twenty-first century sweatshop

The collapse of Rana Plaza on April 24, 2013, in Savar, Bangladesh, in which over 1100 workers were killed, numerous others maimed for life, and the livelihoods of several thousand workers destroyed, catapulted the sweatshop into global headlines. The worst industrial disaster in the history of the global garment industry, Rana Plaza has reopened debates on the merits of sweatshop employment.

Some liberal commentators see the sweatshop as an inevitable stepping stone in the trajectory of economic development. Others have condemned it as hyperexploitation that benefits Northern capital at the expense of Southern labor and have compared the incident to the Triangle Shirtwaist Factory fire of 1911. The many layers of globalized subcontracting, the incorporation of local petty commodity production and home-based workers as well as factory workers across national boundaries, and new networks between producers and consumers, render such comparisons moot.

The Rana Plaza incident and its aftermath recalls and complicates Wright’s analysis of the essentially disposable nature of Southern women’s bodies. The stakes involved are enormous for the multibillion-dollar apparel industry, as well as for the workers whose bodies are rendered into “waste” in the process of contemporary capital accumulation.

Questions of accountability and transparency in the production chain are now at the forefront
GLOBAL FACTORY

of public debates. In the emergent discourse of ethical production and consumption, there is no place for visibly disposable and disposed-of bodies. Scholars must attend to the mediations, elisions, and accommodations at work to render invisible once more violence that was momentarily made hypervisible by the Rana Plaza collapse. It could be argued that this violence is foundational to the global apparel industry.

The global factory has become a rejuvenated site of transnational activism, alliances, and solidarity among groups as diverse as university students, fashion models, and consumers as well as workers’ rights organizations. The fraught nature of building solidarity in a multilayered globalized system is an urgent issue with which researchers and activists must engage (see Brooks 2007). Collaboration between activists and scholars is essential in this context. It may be time for researchers to take a step back and revisit the processes through which workers in places like Bangladesh have few alternative forms of livelihood and have been made desperate enough to choose work that is badly paid and often dangerous, sometimes even deadly.

SEE ALSO: Economic development zones; Gender and development

References


Overview of GPS

The global positioning system (GPS) developed and maintained by the US Department of Defense (Xu 2007) was the first and is still the most utilized satellite navigation system. The first GPS satellite was launched in 1978 and the system became fully operational in 1995. The term global navigation satellite system (GNSS) refers to a constellation of Earth-orbiting satellites transmitting signals encoded with their positioning and timing data. A GNSS can provide autonomous geospatial positioning in all weather conditions with global coverage.

The GPS consists of three main components: the space segment, the control segment, and the user segment. The space segment is composed of 24 satellites in six orbital planes with approximately 55° inclination with respect to the Earth’s equator (Figure 1). In each plane, there are nominally four satellites orbiting at an altitude of approximately 20,200 km. Each satellite is in a nearly circular orbit with a period of about 12 hours so that the satellite passes over the same locations every day. For such a constellation design, at least six GPS satellites are visible almost everywhere on the Earth’s surface.

The control segment is composed of twelve monitor stations, a master control station (MCS, located at Schriever Air Force Base in Colorado), an alternate MCS (located at Vandenberg Air Force Base in California), four dedicated ground antennae (located at Kwajalein, Ascension Island, Diego Garcia, and Cape Canaveral), and six dedicated monitor stations (located at Hawai’i, Kwajalein Atoll, Ascension Island, Diego Garcia, Colorado Springs, and Cape Canaveral). All the monitor stations are equipped with GPS receivers to track all visible GPS satellites and collect observation data. The MCS then processes the data to determine broadcast GPS satellite ephemerides and to model the satellite clock errors. The ephemerides and clock error corrections are transmitted to each satellite via the ground antennae. The satellites, in turn, transmit the updated ephemerides and clock error corrections to GPS receivers.

The user segment consists of GPS receivers and the user community. In general, a GPS receiver refers to a combination of an antenna, receiver-processors, and a highly stable clock. These can process GPS signals and perform calculations to output the position, velocity, and time estimates referenced to the GPS antenna.

How GPS works

GPS satellites transmit three low-power radio signals (also called carriers), designated L1 (1575.42 MHz), L2 (1227.60 MHz), and L5 (1176.45 MHz, only available on Block IIF
satellites). Ranging codes (including L1 C/A, L1 P(Y), L1M, L1C, L2C, L2 P(Y), L2M) and navigation messages are modulated to the carriers under the control of onboard atomic clocks (Misra and Enge 2006). A GPS signal modulated to the carrier wave contains three different types of information: a pseudorandom code, ephemeris data, and almanac data. The pseudorandom code can be used to determine the range distance between the GPS satellite and GPS receiver, while both ephemeris and almanac data can be used to calculate the GPS satellite position and velocity in space. Compared to almanac data, ephemeris data are more accurate but have a shorter validity period. The US government’s current policy is to make GPS available in two services: precise positioning service (PPS) and standard positioning service (SPS). The PPS is available to the military and other authorized users only, while the SPS is available to any user worldwide. PPS users have access to the encrypted P(Y)-codes and M-codes (starting with Block IIR-M) on the L1 and L2 carriers, while the SPS users can observe the public codes L1 C/A, L1C, L2C, and L5. The encryption of the P-codes began on January 31, 1994 through the so-called anti-spoofing (AS) measures (Hofmann-Wellenhof, Lichtenegger, and Wasle 2008). Before May 1, 2000, the SPS users were affected by the GPS selective availability (SA) measures that degraded the GPS signal accuracy by intentionally introducing artificial errors to the GPS satellite clock and ephemeris data. The SA policy was formally implemented on March 25, 1990 and was turned off on May 1, 2000.

A GPS receiver calculates its position by a technique involving the intersection of at least three range measurements, as illustrated in Figure 2, Figure 3, and Figure 4. When one GPS receiver tracks signals from at least three GPS satellites, three ranges can be obtained. Each range measurement is actually calculated from the elapsed transmission time (the time GPS signals have taken to travel from satellite to receiver) multiplied by the speed of light.

Each GPS satellite has a unique pseudorandom noise (PRN) sequence, or Gold code. The good property of the PRN sequences is that the correlation coefficient has the maximum value 1 when two PRN sequences are exactly the same. The correlation coefficient is significantly low when the two PRN sequences are different. Each GPS receiver has the capability to generate PRN sequences for all the GPS satellites. GPS receivers generate PRN sequences to correlate with the incoming GPS satellite signals so that they can detect which GPS satellites are observed through the analysis of correlation coefficients, as shown in Figure 5.

Ideally, the GPS receiver clock is exactly synchronized with the atomic clock onboard the GPS satellite such that they generate the same
One measurement narrows down position to the surface of a sphere with a radius of 20,000 km.

**Figure 2** Spherical surface defined by a satellite position and a distance measurement.

digital code at the same time. When a GPS satellite transmits a ranging code (code 1) that arrives at the receiver after traveling the amount of time $\Delta t$, the receiver duplicates the same ranging code (code 2), as shown in Figure 5. Code 2 is a duplication of code 1, but it is delayed by $\Delta t$. The correlation coefficient between code 2 and code 1 is very low. The local duplication is shifted in order to get a higher correlation coefficient. The maximum correlation can be achieved when the shifted amount is exactly $\Delta t$, which is the time the GPS signal takes to travel from the GPS satellite to receiver. By multiplying $\Delta t$ by the speed of light, the receiver produces the pseudorange observations.

In addition to the pseudorange code measurements, the carrier wave can also be used as a GPS measurement. To generate carrier phase measurements, the receiver first reconstructs the carrier phase by removing the modulated ranging codes and navigation messages. The carrier phase measurement represents the range distance between a satellite and receiver expressed in units of cycles of the carrier frequency. Carrier phase measurements have very high precision (of the order of millimeters) but they also have unknown ambiguities, which are in the number of whole carrier cycles. The accuracy of carrier phase measurements is improved by resolving the carrier phase ambiguities, which is also known as ambiguity resolution. This is typically done by using multiple GPS satellites and taking advantage of the fact that the phase ambiguities are constant for a given satellite and time period. The accuracy of carrier phase measurements is further improved by using integer ambiguity resolution techniques, which involve solving for the integer carrier phase ambiguities.

Three measurements narrow down position to just two points.

**Figure 4** Three spherical surfaces intersect to generate two points. Usually the receiver can discard one of the two points because it is nowhere near the Earth. This leaves one point that is the location of the GPS receiver.

Two measurements narrow down position to the intersection of two spheres.

**Figure 3** Two spherical surfaces intersect to generate a circle.
phase measurements has a measuring accuracy of 1 mm or higher, but the ranging code has an accuracy of approximately $\pm 3$ m for C/A code and $\pm 0.3$ m for P code.

Once the receiver has the distance measurements, GPS positioning and navigation are now reduced to a problem of geometry. If the positions of GPS satellites in the space are known, it is possible to compute the GPS receiver location through trilateration. In theory, only three measurements from three GPS satellites are needed to determine the 3-D position of a GPS receiver. However, generation of GPS measurements closely depends on the GPS clock. For GPS measurements that contain a large clock error, the estimated 3-D positioning solution contains a large error too. If the GPS receiver clock error is estimated as an unknown parameter together with the 3-D coordinates of the receiver, the GPS positioning solution will become more accurate. Thus, in practice, at least four measurements from four satellites are needed in the GPS SPS.

The mathematical equations expressing the GPS distance measurements can be written as shown in equations 1 and 2.

\[
\rho_i = \sqrt{(X^i - X)^2 + (Y^i - Y)^2 + (Z^i - Z)^2} - \epsilon \cdot V_{Ri} + \epsilon \cdot V_{Si} + (V_{ion})_i + (V_{trop})_i + \epsilon_i \tag{1}
\]

\[
\phi_i \lambda = \sqrt{(X^i - X)^2 + (Y^i - Y)^2 + (Z^i - Z)^2} - N_i \lambda - \epsilon \cdot V_{Ri} + \epsilon \cdot V_{Si} - (V_{ion})_i + (V_{trop})_i + \epsilon_i \tag{2}
\]

where $\rho_i$ is the pseudorange measurement from satellite $i$ to the receiver; $(X_i, Y_i, Z_i)$ are the coordinates of satellite $i$; $(X, Y, Z)$ are the coordinates of the receiver; $\epsilon$ is the speed of light; $V_{Ri}$ is the receiver clock error; $V_{Si}$ is the clock error of satellite $i$; $V_{ion}$ is the ionospheric delay; $V_{trop}$ is the tropospheric delay; $\epsilon_i$ is the noise; $\phi_i$ is the carrier phase observation from satellite $i$ to the receiver; $\lambda$ is the wavelength of the signal; and $N_i$ is the integer ambiguity.

In pseudorange positioning mode, the tropospheric delay is usually corrected by models or eliminated by the combination of different frequency signals; the satellite clock error is corrected by using the clock correction information in the navigation message. The rest of the four parameters (three receiver coordinates parameters and one receiver clock delay parameter) need to be estimated by using at least four satellites’ pseudorange observations. In carrier phase positioning mode, the interambiguity should be estimated in addition to the coordinate parameters and the receiver clock error parameter. As the carrier phase measurements have a high accuracy, they are usually used for high-precision positioning. In addition, the satellite position error and satellite clock error, ionospheric delay, tropospheric delay, and other corrections should be carefully dealt with in high-precision GPS positioning.

GPS signals undergo two effects when they move through the atmosphere, bending and...
time delay, due to the atmospheric refraction. There are two major atmosphere layers: the troposphere and the ionosphere. The geometric bending of GPS signals with elevation angles above 5° is negligible. The delay in the ionosphere is dependent on signal frequency and, thus, can be eliminated by the combination of signals of different frequencies. The troposphere is essentially a nondispersive medium for frequencies below 30GHz (GPS signals frequency: below 2GHz). So, the tropospheric delay cannot be eliminated by frequency combinations; it is usually estimated in positioning models or predicted by tropospheric delay models.

In addition, errors in satellite orbits and clocks, relativistic effect, multipath, and receiver measuring noise also have an impact on the positioning accuracy. The receiver clock error is usually estimated as an unknown parameter in observation equations. The satellite orbit and clock errors can be reduced by using International GNSS Service (IGS) precise orbit and clock products. The relativistic effect can be calculated by physical equations. Multipath error can be weakened by using advanced receiver antennae or placing the antenna in an open-view environment or extending the observation time. The receiver measurement noise can be weakened by using advanced antennae and receivers.

During GPS positioning, geometrical dilution of the precision (GDOP) is usually used as an important index of evaluating satellites’ geometry. The wider separation among the satellites, the lower the GDOP value and the better positioning accuracy can be obtained.

In positioning and navigation, the accuracy of the GPS SPS is limited to a few meters. This is due to the large residual of atmospheric errors as well as satellite orbit and clock errors. To improve the SPS accuracy, a method called differential GPS (DGPS) has been developed. In DGPS, the differences in GPS observations from two receivers and two satellites can be used to eliminate common errors. Differential observations between two receiver stations can eliminate the satellite-related errors and significantly reduce spatially correlated errors (mainly tropospheric and ionospheric delays). With differential observations between two satellites, the station-related errors are eliminated and spatially correlated errors are reduced. Double differential observations eliminate the satellite-related and station-related errors at the same time and spatially correlated errors are further reduced. Because of this, DGPS is widely used around the world. In recent years, a new method called precise point positioning (PPP) has been developed, in which carrier phase measurements from a single GPS receiver are used to achieve positioning and navigation accuracy much higher than the SPS (to a few centimeters in static mode and to the decimeter level in kinematic mode). However, the PPP method requires the use of precise orbit and clock data from GPS satellites. Meanwhile, the GPS receiver takes tens of minutes to fix the carrier phase ambiguities in order to get the expected accuracy.

Global Navigation Satellite Systems

Integrated GPS and GIS systems: data collecting and mapping

GPS is an advanced tool for collecting accurate positions for point, line, and polygon features while the geographic information system (GIS) has developed to be a very useful cartographic tool for storing, manipulating, and visualizing all kinds of geographic data and attribute information with useful spatial analysis functions. By incorporating GPS and GIS, location data could be quickly and accurately collected and imported into GIS, which will be very helpful for creating, validating, updating, and maintaining GIS spatial information and attributes/features.
The progress in GPS/GIS integration in geography and geographic information science (GIScience) has led, during the past several decades, to the creation of real-time interactive (RTI) GPS/GIS systems. RTI GPS/GIS has created new possibilities to interact with, represent, and make decisions about the world in real time (Richardson 2013). It enables individuals to interact in real time with both their actual and digital environments simultaneously in daily life, and also allows researchers to better approximate and spatially model how individuals and populations interact with the world around them (Kwan 2009, McQuoid and Dijst 2012).

RTI GPS/GIS functionality has enabled the collection and use of geospatial data and attribute information to be achieved in a more detailed, timely, and accurate way than ever before, generating massive quantities of new, fused spatiotemporal data (including sensor data attributes). Location-based services (LBS) are among the many industries where RTI GPS/GIS technology plays an important role. In LBS, the GPS provides accurate location information while the GIS displays and updates in real time the street maps, attributes, and users’ locations. Besides the location service, LBS can also provide other location-related services that users are interested in, such as restaurants or shops nearby and how to get there. Miniaturized RTI GPS/GIS functionality has now been embedded also in consumer electronic devices such as cell phones and navigation devices, which further enriches the application of LBS and many other applications.

As technology advances, mobile devices become common in professional GIS and survey work. Mobile devices with embedded RTI GPS/GIS functionality, such as smartphones and tablets, have become an essential part of how people communicate and get their information. These devices usually have a position accuracy of 5–6 m under ideal conditions, which has long prevented them from being a viable solution for GIS survey work in the field. Modern professional-grade GNSS receivers are becoming smaller and more portable, and address the accuracy limitations of an older smartphone or tablet device. The small, compact professional-grade receiver provides submeter location data directly to the smart device via Bluetooth®. An example of a Trimble R1 GNSS receiver is shown in Figure 6. By simply pairing the device with their smartphone or tablet, GIS professionals can get the accuracy needed to quickly and reliably locate their asset, collect data, and move on to the next task. Other models of professional GNSS receivers for GIS applications have been placed on market, too. For instance, the Trimble R2 GNSS receiver is built for GIS professionals and surveyors alike and is capable of delivering between submeter and centimeter positioning accuracy in real time to a Trimble handheld or a smart device. Such high-accuracy handheld GNSS receivers are often combined with rugged cases for carrying out GIS field data collection and industrial or environmental inspection activities.

**Other GNSS systems**

**GLONASS**

The global orbiting navigation satellite system (GLONASS), maintained by the Ministry of Defense of the Russian Federation, is the second operational satellite navigation system with global coverage. The first GLONASS satellite was launched in 1982 and it became a fully functional navigation constellation in 1995. The GLONASS constellation consists of 24 satellites equally distributed on three orbital planes, with eight evenly spaced satellites on each plane (RISDE 2008). The three orbital planes are separated by 120° in the equatorial
Figure 6  The Trimble R1 receiver connects wirelessly to a smart device via Bluetooth connectivity.

plane. Satellite orbital altitude is about 19 130 km above the ground surface with an inclination of about 64.8° and a period of 11 hours and 15 minutes (RISDE 2008). The GLONASS orbit was designed especially to be suitable for use in high latitudes, where receiving a GPS signal can be problematic. Unlike the GPS that uses the code division multiple access (CDMA) technique to discriminate the signals of different satellites, GLONASS uses frequency division multiple access (FDMA) to distinguish between the satellites. All GLONASS satellites transmit the same C/A and P codes, but each satellite is allocated a slightly different frequency. Thus, users’ receivers can identify the satellite according to these different frequencies. However, modernized GLONASS satellites are gradually adopting the CDMA technique. The launch of the first GLONASS-K satellite in February 2011 marked the introduction of CDMA technique to the GLONASS signal broadcasting.

Galileo

Galileo is the European navigation satellite system currently under development by the European Union and European Space Agency. The first experimental satellite was launched on December 28, 2005 and sixteen satellites have been launched as of July 2016. Currently, nine satellites are in operation and four others are in the commissioning stage. The Galileo system aims to provide its initial operational capability (IOC) by 2017–2018 and reach full operational capability (FOC) in 2019. The constellation will consist of 30 medium Earth orbiting (MEO) satellites (24 in full service and 6 spare satellites), distributed evenly over three orbit planes (EU 2010). The orbital altitude is 23 222 km and the inclination is 56° with ascending nodes separated by 120° longitude (eight operational satellites and two active spares per orbital plane). The ground segment includes the Galileo control centers, telemetry tracking and control stations, uplink stations for transmitting navigation signals, and up to 40 Galileo sensor stations for receiving the navigation signals from the constellation and transmitting them to the Galileo control centers. The responsibility of the ground segment is to manage navigation of the satellite constellation and control the core functions of the navigation mission, such as satellite orbit determination and clock synchronization.

BeiDou

The BeiDou navigation satellite system (formerly known as COMPASS) is the Chinese version of a GNSS and is being developed rapidly. It began to offer services in China and neighboring areas in December 2011, with 10 satellites in use; a regional version including 16 satellites covering Asia and the Pacific area was completed by December 2012. As of December 2015, 24 satellites for the BeiDou system have been launched, 18 of which are in service. All the BeiDou satellites use the CDMA technique and share the same nominal frequency. The
BeiDou navigation satellite system is scheduled to become globally operational upon its completion in 2020. By then, the space segment of BeiDou system will consist of five geostationary Earth orbit (GEO) satellites, 27 MEO satellites, and three inclined geosynchronous satellite orbit (IGSO) satellites (CNSO 2013). The GEO satellites are operating in orbit at an altitude of 35,786 km and positioned at 58.75° E, 80.2° E, 110.6° E, 140° E, and 160° E, respectively. The orbital altitude of MEO satellites is 21,528 km with an inclination of 55° to the equatorial plane (CNSO 2013). The IGSO satellite orbit altitude is 35,786 km above ground surface and its inclination is 55°.

Regional augmentation systems

IRNSS: The Indian regional navigational satellite system (IRNSS) is an autonomous regional satellite navigation system being developed by the Indian Space Research Organization, with an aim to provide accurate positioning and timing services throughout India. The Indian government approved the project in May 2006 and completed the system in April 2016. On April 28, 2016, the IRNSS was officially renamed as Navigation with Indian Constellation (NAVIC) by the Indian Prime Minister. The constellation of seven NAVIC navigational satellites (orbital altitude is 36,000 km) includes three GEO satellites, four GSO satellites, and a support ground segment. The three GEO satellites are located at 32.5° E, 83° E, and 131.5° E. Two of the GSO satellites cross the equator at 55° E and two cross at 111.75° E. This orbit design means that all seven satellites are always visible to the Indian region.

QZSS: The Japanese quasi-zenith satellite system (QZSS) is a regional time transfer system and satellite-based augmentation system (SBAS) for the GPS. The first satellite was launched on September 11, 2010. The QZSS is expected to have four satellites by 2018 and up to seven satellites after 2018. The signals transmitted by QZSS satellites are compatible with the GPS L1 C/A signal, as well as the modernized GPS L1C, L2C, and L5 signals. The QZSS is designed primarily to increase the GPS availability in Japan, with the secondary function to increase the accuracy and reliability of GPS-derived positioning solutions.

WAAS: The wide area augmentation system (WAAS) is a satellite-based SBAS developed by the US Federal Aviation Administration with the goal to enhance the accuracy, integrity, and availability of the GPS. WAAS models the errors (mainly satellite orbit error, satellite clock error, and ionospheric error) in the GPS satellite signals using a network of ground-based reference stations in North America. Master stations process the measurements from the reference stations and send the correction messages to geosynchronous communication WAAS satellites. The satellites then broadcast the correction messages to users who are equipped with WAAS-enabled GPS receivers to compute their positions with improved accuracy.

EGNOS: The European geostationary navigation overlay service (EGNOS) is an air navigation aid developed by the European Space Agency, the European Commission, and EUROCONTROL to augment the GPS, GLONASS, and Galileo systems. Consisting of three geostationary satellites and a network of ground stations, the EGNOS system aims to improve the GPS positioning and navigation accuracy to within 3 m. The initial operations of the EGNOS system started in July 2005, with accuracy better than 2 m and availability above 99%. The EGNOS Open Service has been available since October 1, 2009; its Safety of Life service was officially declared available for aviation on March 2, 2011.
Other transformational applications

In addition to the traditional applications of global navigation satellite systems in navigation and positioning, the potential of global navigation satellite systems has been demonstrated in various applications, ranging from civilian to military, from general public to scientific research (Gleason and Gebre-Egziabher 2009; Someswar, Rao, and Chigurukota 2013).

**Precision agriculture:** In precision agriculture, global navigation satellite systems can be applied in farm planning, field mapping, tractor guidance, and crop scouting (Gleason and Gebre-Egziabher 2009). Global navigation satellite systems allow agricultural equipment to steer automatically. Farmers may use global navigation satellite systems to work during conditions such as rain, dust, fog, and darkness when low visibility occurs.

**Timing and synchronization:** The high precision of GNSS timing can be used for a variety of applications for timing and synchronization purposes. For time division multiple access (TDMA) communication networks, precise timing is often adopted to synchronize radio frequency generating equipment, network equipment, and multiplexers.

**Emergency services:** Emergency services can make use of GNSS functionality to locate the person in danger. The rescue coordination center can thus act quickly to get to the site.

**Weather prediction:** In the propagation from the satellite to ground receiver, GNSS signals will be delayed and bent by the atmosphere. Global navigation satellite systems can be adopted to determine atmospheric conditions, such as moisture, that can be used for weather prediction.

**Social networking:** Cell phones equipped with GNSS technology can provide geographic location for social networking functions (Someswar, Rao, and Chigurukota 2013). Based on the location information, many apps offer the ability to search friends or to pinpoint nearby facilities on custom-created maps.

**Transportation and logistics:** In many countries, traffic congestion and road safety have become serious social issues. Global navigation satellite systems can offer meter- to submeter-level service to land vehicle lane-level navigation, intelligent transportation, and management. Global navigation satellite systems can also be used for e-commerce or online shopping to improve the logistics and transportation of commodities for reaching clients. The instantaneous service of GNSS positioning could be used by clients to track their commodities online.

**Future trends**

Global navigation satellite systems are now widely used throughout our daily lives and this trend will continue in the foreseeable future. With the rapid development of GNSS technology during the past several decades, global navigation satellite systems have been applied in many scientific research and engineering applications, such as earthquake monitoring, meteorological weather service, ionosphere and space weather monitoring, geohazard monitoring, natural resource monitoring, LBS, and surveying and mapping, land, water, and air navigation systems. GNSS applications are also found in many fields of scientific study, including space weather, geophysics, geography, geology, ecology, and biology (Gleason and Gebre–Egziabher 2009). In the future, the development of global navigation satellite systems will be toward multisystem and multi-application, with goals to ultimately provide better performance and broader services for mankind. The majority of GNSS applications are currently based on GPS and GLONASS performances. Once the Galileo and the BeiDou systems are
fully operational, the ability to process signals from multiple systems will further improve availability and accuracy for users. At present, many countries and governments are concerned with substantial socioeconomic issues, including environmental pollution, transportation safety, extreme poverty, and natural disasters. Global navigation satellite systems are increasingly integrated with other technologies and systems to help address these problems. For example, applying a GNSS in emergency situations creates an efficient tool for numerous emergency services, such as the fire brigade, police, and sea and mountain rescue, allowing them to respond more rapidly to those in danger. The future of global navigation satellite systems is bright, with the progressively increasing number of constellations from multiple global navigation satellite systems, as well as new applications that will benefit from the increased robustness and availability of high-accuracy GNSS solutions.

SEE ALSO: Digital Earth; Discrete global grid systems; Geocoding; Geodesy; Geolocation services; Location-based services; Map projections and coordinate systems; Surveying

References


Global production networks

Martin Hess
University of Manchester, UK

In the twenty-first century, the world economy has seen substantive challenges and changes, not least the global financial crisis, the ramifications of which are still being felt around the globe. Contemporary economic globalization can be characterized by the increased functional and geographical fragmentation and reconfiguration of production processes, deepening outsourcing and offshoring, changing geographies of production and consumption, and associated labor market dynamics including the ascent of temporary and migrant work. These dynamics have become increasingly prevalent with the rise of neoliberalism and the end of the Cold War, triggering new lines of social sciences inquiry into globalization and economic development. Moving beyond more state-centric approaches to economic development studies, approaches such as commodity chain research and global value chain analysis have been developed to better understand the social and developmental consequences of contemporary capitalism (Bair 2005). It is within this context, sharing such a research agenda, that the global production network (GPN) concept emerged as an analytical framework and heuristic tool to understand the changing nature and dynamics of economic globalization and regional development. The following discussion will first present the conceptual foundations of the GPN framework and its intellectual influences. Subsequent sections will focus in turn on power relations between actors and their impact on governance structures in and of GPNs; the increasing fragmentation of GPNs through outsourcing and offshoring, driven by, among other factors, corporate as well as other forms of financialization; the role of labor and labor agency in GPNs; and the dynamic relationships between GPNs and regional development. The entry will conclude with brief reflections on GPNs as politically contested fields and their “discovery” by policymakers and international organizations as arenas of international governance and developmental “tools.”

GPNs: conceptual cornerstones

As an analytical framework, the GPN concept has been developed since the late 1990s by a group of scholars in economic geography and international economic sociology, mostly based at the University of Manchester at the time. It emerged from a growing dissatisfaction with existing theories of economic development that operated either at macro-levels or micro-levels of abstraction and thus failed to capture the increasingly complex, networked nature of economic activities under neoliberal globalization and their impacts on uneven development at various scales. The construction of the GPN framework rests on a number of historical precursors. Hess and Yeung (2006) identified four strands of literature informing this approach: (i) the 1980s value chain literature associated with the work of Michael Porter; (ii) work on networks and the social embeddedness of economic activities as developed in economic

The International Encyclopedia of Geography.
sociology; (iii) actor-network theory (ANT), which emerged in the context of science and technology studies; and (iv) most notably, the literature on global commodity chains and global value chains, originating in world-systems theory and developed since the 1990s by Gary Gereffi and his colleagues.

Porter’s value chain analysis has been influential in both academic and policy circles, highlighting the various activities a firm performs and the resulting systems of inputs, transformations, and outputs making up the production process. Influential in economic geography for the analysis of industrial clustering, it has also informed the GPN framework with regard to the centrality of value creation and the spatial organization of production processes and service provision. To better understand how economic activities are organized within and between firms, the role of social networks and the embeddedness of economic action in ongoing social relations – in contrast to the methodological individualism of transaction cost economics – have to be integrated as crucial elements in GPN analysis. Economic sociology and an emerging literature on relational economic geography therefore had a major impact on the development of the GPN framework. This was complemented by insights from ANT and its emphasis on a nonessentialist approach to studying networks and actors. ANT reinforces a relational view, avoiding artificial dualisms like structure/agency and global/local, and thus opening up analytical space for investigating multiple actors and their heterogeneous relations in GPNs.

Finally, global commodity chains (GCCs) and global value chains (GVCs) have provided a major impetus for the GPN framework. GCC analysis addresses four different dimensions (Gereffi and Korzeniewicz 1994): an input–output structure describing the production process and associated activities along the chain, from raw materials to the final product; the territoriality of GCC, which represents its geographical configuration; the governance dimension, which denotes the power relations of actors (firms) along the commodity chain; and an institutional dimension, the background of state regulation and other institutional rule-setting against which firms in GCCs operate. The GVC framework built on this conceptualization with a view to refining the forms of governance found in inter-firm value chains, and – echoing some insights from Porter’s value chain approach – paying more attention to the local and regional dimension of clusters within GVCs.

Based on a critical engagement with these conceptual antecedents, and following Henderson et al. (2002), a GPN can be defined as the nexus of interconnected functions and operations through which goods and services are produced, distributed, and consumed. Over the years, GPNs have become organizationally and geographically more complex, increasingly blurring traditional organizational boundaries. They integrate regional and national economies, cutting through state boundaries in highly differentiated ways to create discontinuously territorial structures that are shaped by regulatory and nonregulatory barriers as well as variegated sociocultural conditions of the places connected by GPNs.

The GPN framework draws on three analytical registers: value, power, and embeddedness (Henderson et al. 2002). Two notions of value are important for GPN research. First, value is to be understood as surplus value created in transnational production systems through the labor process. This brings into focus issues of employment, working conditions, and productivity at various points in the network, with far-reaching consequences for socioeconomic development. Second, value refers to the notion of different forms of rents that firms can realize
within GPNs. These include technological rents, through access to advanced product and process technologies; brand rents, realized through a strong market presence and consumer preferences; organizational rents, achieved by optimizing managerial and organizational skills and production processes; and relational rents, through strategic links with other firms in the wider GPN. Other forms of rent may also occur. For example, in sectors where global trade is highly regulated or restricted, preferential access may generate "trade policy rents," as in the now defunct global trade regime in textiles and garments known as Multi Fibre Arrangement.

In relation to economic development, these two vectors of value creation are complemented by processes of value enhancement (increasing the value added at various stages of production) and, crucially, value capture: that is, the process of retaining economic benefits. The latter is of particular importance for both firms in GPNs and the places and societies they connect.

How and where value is created, enhanced, and captured depends on the power relations between the multiple actors in GPNs. Firms exercise corporate power based on their position within GPNs, their different capabilities, and the resources available to them. State and civil society organizations exercise power that influences firms’ operations; rather than forming an external environment within which firms act, these nonfirm actors are conceptualized as integral parts of GPNs. Thus state organizations assume power through authority as regulators and facilitators of economic activity, but also exercise buyer power as major consumers (public sector purchasing), and producer power (through state-owned enterprises). Civil society actors produce various forms of collective power, for instance through labor unions, nongovernmental organizations (NGOs), and consumer initiatives, which can also have considerable impact on GPNs by putting pressure on lead firms.

Finally, GPN analysis takes into account the embeddedness of economic activity, along three dimensions (Hess 2004). Societal embeddedness – a notion that draws on the work of Karl Polanyi – refers to the importance of an actor’s institutional and cultural background in shaping their actions, rather than deploying a universalistic notion of rational economic agency. Network embeddedness – from Mark Granovetter’s work in economic sociology – puts an emphasis on the relevance of social ties that shape the relations between actors in GPNs. Finally, territorial embeddedness takes into account the geographical dimension of GPNs and the varying degrees to which firms, nonfirm organizations, and institutions are “anchored” in particular places. Together, embeddedness, power, and value constitute the three analytical lenses of the GPN framework, guiding research into GPNs’ organizational, geographical, and developmental dynamics.

**Governing GPNs**

From early GCC analysis through to the GVC and GPN frameworks, a central concern has been with the governance of interfirm relations and the asymmetrical distribution of power between firms. The initial taxonomy developed in this context distinguishes between buyer-driven and producer-driven commodity chains. In industrial sectors where high capital requirements are the main barrier of entry for firms, as in the automotive and aircraft industries, the value chain is driven by large manufacturers (producers). Buyer-driven value chains are found in sectors that are labor-intensive and where entry barriers are primarily the design, marketing, and branding capabilities of lead
GLOBAL PRODUCTION NETWORKS

firms. Examples include the garment and light electronics industries (Gibbon, Bair, and Ponte 2008). While this taxonomy has proven useful to identify and analyze one important aspect of the distribution of power within GPNs, the increasing complexity of contemporary production systems does not always conform to this. Other forms of governance have subsequently been identified, such as technology-driven value chains characteristic of information and communication industries (particularly software), where entry barriers lie with the control of technological standards and related intellectual property rights.

While providing insight into some fundamental power configurations within GPNs, these types of chain-driving governance structures are quite crude, obscuring the multiplicity of power relations between firms (and nonfirm actors) within a GPN. To better capture the various power relations between firms, a typology of value chain coordination has been developed that describes the nature of exchange between firms as five possible forms of governance: market-based governance, where transactions are easily codified around simple products; modular value chains, where codification of transactions extends to complex products (e.g., between automotive manufacturers and first-tier module and component suppliers); relational value chains, where governance of complex products and exchange requires frequent communication; captive value chains, where suppliers are highly dependent on and monitored by their buyers; and hierarchical value chains, where exchange is internalized within the firm (in-house production).

Even this taxonomy, however, does not include an appreciation of power relations in GPNs extending beyond interfirm exchange, and actors which are also crucial for the configuration, dynamics, and governance of GPNs: state and nongovernmental organizations, civil society and consumers. Economic processes of production, distribution, and consumption are not simply driven by lead firms in GPNs and coordinated between firms along the value chain (Coe, Dicken, and Hess 2008). They are embedded in wider systems of sociospatial relations and shaped by nonfirm actors operating with their own spatial logic and according to their own specific goals and priorities. The GPN framework therefore explicitly recognizes nonfirm actors as integral parts of the production network. Consequently, state–capital–society relations in the places linked through GPNs are fundamental in shaping economic and social outcomes. Thus an explicitly geographical perspective is critical to understanding the ways in which GPNs are governed and different stakeholders struggle over the creation and capture of value.

This requires GPN research to conceptualize power more explicitly, rather than only in terms of the individual or collective actors exercising it. Following the work of John Allen (2003) on geographies of power, three different forms of power can be distinguished in investigating governance structures of/in GPNs (Hess 2008). The most commonly used view is a realist conceptualization of power as a capacity, something which individuals or organizations (firms, states) possess, enabling them to dominate others by virtue of social relationships. For firms, this potential to dominate derives from various resources and firm capabilities, and enables them to control or direct the actions of others, as in the case of lead firms driving value chains. Governments, at various scales, derive their power from the authority and sovereignty accorded to them as political institutions.

While conceiving of power as a capacity to influence others, whether or not it is exercised, is certainly important to understand the dynamics of GPNs, it is not sufficient. Actors can also
mobilize resources that are not all of their own making, through collective action and cooperative network relationships. In such a relational, networked view, shared resources become the medium through which power is exercised. This opens up possibilities for "powerless" actors (firms lacking individual resources, workers, consumers) to work together for mutual benefit and achieve their respective goals. Examples for this range from the labor union movement to consumer and NGO campaigning, from firms’ strategic alliances to state-negotiated trade agreements, none of which can be understood purely through power as a capacity.

Finally, and more recently, GVC and GPN analysis have both developed a third lens of power relations by drawing on Foucauldian notions of power as knowledge and practice, moving from a governance perspective to a governmentality approach. Of interest here is how power is mobilized and practiced, and the specific techniques and discourses used and “normalized” to direct the behavior of other actors and achieve specific outcomes. These practices and techniques, including supply chain management to orchestrate interfirm relations and corporate social responsibility codes of conduct, often become institutionalized in the form of standards as an important element of GPN governance. Some types of standards, for instance environmental or social, may be coproduced through the collaboration of various firm and nonfirm actors; others, like some proprietary technological standards such as Apple Inc.’s mobile phone operating system, are produced in the context of the competitive struggle for market control and value capture through brand rents (Coe, Dicken, and Hess 2008).

The variety of governance forms and power relations found in GPNs, within and across different territories, has become increasingly complex since the mid-1980s, as global production systems have continued to expand their global footprint and become increasingly fragmented in organizational terms. The following section examines this “new wave” of globalization (Milberg and Winkler 2013) in more detail.

GPNs, fragmented production, and financialization

The second half of the twentieth century has seen a growing tendency of transnational corporations (TNCs) from the Global North shifting production activities to the Global South, especially in labor-intensive industries. This process has been described in the literature as a new international division of labor (NIDL), taking advantage of a large and growing labor force in developing economies, the transfer of standardized production processes that require only low-skilled workers, and the reduction of transport and communication costs based on new technologies. While still relevant today, as in the garment and consumer electronics industries, the NIDL thesis does not fully capture the drivers for the reorganization of GPNs that have emerged since the 1980s. Contemporary capitalism is characterized by a global division of labor that has experienced major geographical restructuring among world regions, the growing interpenetration of global processes, regional dynamics and local conditions, and massive transfers of people through migration, a workforce that GPNs seek to attract. Another striking feature of these transformations is the increasing fragmentation and vertical disintegration of production, through outsourcing, subcontracting, and offshoring.

Outsourcing refers to a firm’s strategic decision to purchase goods or services from other companies, rather than producing them in-house. The reasons for outsourcing usually lie in attempts to
save cost and enhance profitability by focusing on what a firm sees as its core competences and where it has a competitive advantage. Interfirm divisions of labor and outsourcing are by no means new, described in such classic texts as Alfred Marshall’s account of the Sheffield cutlery industry in *The Economics of Industry* (Marshall and Paley Marshall 1881). Alfred Marshall’s work highlighted the role of spatial proximity for the organization and profitability of interfirm divisions of labor, and was drawn on in the emerging literature on geographical clusters of economic activity. He could not foresee, however, the wave of outsourcing across national borders that became increasingly global in nature, referred to as offshoring. This has become a central empirical and conceptual feature of GPN and value chain research.

The transnational relocation of manufacturing and service activities – whether through international outsourcing or offshoring in-house activities – is pursued for a variety of reasons, including the pursuit of greater flexibility, the avoidance of risks through location diversification, and of course reducing cost to maintain or increase profitability. The combined effects of outsourcing and offshoring are manifest not only in the complex geographies of contemporary GPNs, but also in substantial changes in international trade. The “slicing up” of value chains has led to a growing trade in intermediary goods and components, along with finished goods. Therefore, for firms in GPNs, supply chain management has become an important element of competitiveness and a crucial strategic asset. At the same time, the offshoring of service tasks (for instance, call centers, back-office functions) has further increased what is called “trade in tasks.” These developments provide economic opportunities for some firms in developing and emerging economies, but other players, failing to realize the potential of new global divisions of labor, have lost their competitiveness.

Financialization has become much more important for GPNs since the mid-1990s (Coe, Lai and Wójcik 2014; Milberg and Winkler 2013). In the context of GPN research, financialization can be analytically broken down into three interrelated forms. First, it denotes the increased significance of the financial sector vis-à-vis other economic sectors. Financial markets have always been important for the functioning of GPNs, but their powers have continually grown since the liberalization and deregulation of the sector from the 1980s onward. By the early twenty-first century, the value of global financial markets is estimated at about three times the global gross domestic product, with approximately US$15 trillion traded annually. Accelerated financial integration, with a growing propensity for financial decisions in one place to influence conditions in others, has substantial consequences for GPNs. The financial system incorporates all actors and territories in GPNs (states, firms, individuals, cities, regions, and nations) via investments and borrowing, but also through taxation and public expenditure. This leads to the second aspect of financialization, the formation of global financial networks. Such financial networks, global in nature but centered on a small number of global cities like New York and London, intersect with GPNs in various ways and are crucial for their working. Networked financial institutions over the last decades have designed, produced, and sold increasingly sophisticated financial products, with implications for the third form: corporate financialization.

Firms in GPNs have shifted their focus of profitability away from traditional sources like production and toward profits gained through financial activities. Corporate financialization is linked to the fragmentation of production
Laboring in GPNs

While emphasizing how transnational production systems are orchestrated by a variety of firm and nonfirm actors, arguably the role of labor and the agency of workers in shaping GPN have been undertheorized. This is problematic, not least because the process of value creation and enhancement under capitalism is not only a question of generating different forms of economic rent from the production process or of generating profit in financial markets, but is also reliant on the labor process by which labor power is transformed into surplus value. Labor therefore needs to be considered explicitly in GPN research. Such an analysis should recognize the outcomes for workers, for example in terms of wages, working conditions, and labor rights, but also the potential for workers to transform and shape GPNs based on their individual and collective power and agency. Much of the literature addressing labor has focused on public- and private-sector governance of labor relations and on developing labor standards, but it often conceptualizes workers as rather passive and at the receiving end of capital and state power. In order to eliminate this blind spot, GPN literature has started to engage with the work of labor geography and the industrial relations literature (Coe and Hess 2013).

Since the formation of the labor movement in Europe during the Industrial Revolution and the subsequent legalization of trade unions, collective labor agency in the form of labor unions has played an important role in ensuring labor rights and improving working conditions. Through collective bargaining, workers exercise power vis-à-vis capital to negotiate a fair share of the value created and generally improve their lot. As production in GPNs has extended in scale and scope, however, the global fragmentation of value chains has posed additional problems for organized labor, as it tries to match the capability of capital to organize across national and international space. Unlike capital, labor is socially and, to a large extent, territorially embedded, and therefore in danger of being played off against labor in other locations. Yet trade unions have developed strategies to upscale their activities, establishing their own global networks. Two forms can be distinguished: labor networks centered on a specific TNC, where bargaining takes place, for instance, through international framework agreements between unions and a single employer; and Global Union Federation-centered networks, which usually cover a specific industrial sector and negotiate with various firms in GPNs.

While collective agency through trade unionism clearly has empowering effects for labor in GPNs, this has been gradually eroded in this era of neoliberal globalization. Labor markets around the world have been increasingly deregulated, leading to growing numbers of “flexible,” casualized, and often precarious jobs in manufacturing as well as services. In the wake
of this, new labor market intermediaries such as temporary staffing industries have emerged to play an important part in many countries, fueling GPNs with domestic and migrant labor. Under these circumstances, traditional collective action by organized labor has a declining potential to galvanize workers’ power. Yet individuals and groups still have agency in shaping GPNs and the conditions of work within them. Following Cindy Katz (2004), worker agency can take the form of strategies of resilience, where people develop everyday coping mechanisms. It can result in reworking strategies, where individuals or communities work to actively improve the material conditions of their existence, and it can lead to strategies of resistance, through which the status quo of capitalist social relations is directly challenged. Thus labor is most certainly not a passive element in the formation of GPNs.

Yet working conditions and enabling rights in GPNs continue to be areas of serious concern. The collapse in 2013 of the Rana Plaza building in Dhaka, Bangladesh, in which more than 1100 garment workers died, was a horrendous incident that threw into sharp relief the dark side of fragmented production in GPNs in the pursuit of profits, and reopened academic and public debates about the responsibilities of governments, firms, and consumers in GPNs to safeguard workers’ lives and livelihoods, especially in labor-intensive industries in the Global South.

One way in which recent GPN research has taken up the challenge of conceptualizing labor, as socially embedded agents and human beings rather than simply factors of production, is the notion of social upgrading in GPNs (Barrientos, Gereffi, and Rossi 2011). Social upgrading can be defined as the process of improving the rights and entitlements of workers as human beings and social actors, enhancing the quality of their employment, and improving the living standards of workers, their families, and communities. In most of the literature, this is broken down into two components: (i) measurable labor standards, including wage levels, working hours, social protection entitlements, and type of employment – regular or irregular; and (ii) enabling rights, including freedom of association and the right to collective bargaining. The concept of social upgrading is usually deployed together with the notion of economic upgrading which describes improvements in technology, skills, and productivity that generate enhanced profits derived from participation in GPNs. It is often assumed that economic upgrading is positively associated with social upgrading, but empirical studies have shown that this is not necessarily the case. How the economic and social dimensions of labor in GPNs play out in specific places to generate positive outcomes for workers depends on how labor agency is shaped by the governance structures of particular GPNs (the vertical dimension) and by the local social and economic conditions of employment and work (the horizontal dimension). This returns us to wider questions of socioeconomic development at the global–local nexus.

**GPNs and regional development**

To understand regional development in an era of global, networked capitalism, it is clearly not enough to focus on local, endogenous factors alone. At the same time, economic globalization cannot be adequately explained without paying attention to the specific assets and socioeconomic conditions in different countries, regions, and localities. In a relational view, both the growth factors within a given region and the strategic priorities of transregional and transnational firms that orchestrate GPNs need to be taken into consideration in order to investigate and explain development outcomes. In the GPN
literature, this has been labeled the globalization of regional development, a process whose outcomes are far from certain, depending on the strategic coupling of regional assets with GPNs (Coe et al. 2004). Regional assets include the size and composition of the workforce (human capital), education and research facilities, and natural resource endowments, among others. Three aspects are characteristic for the strategic coupling process between regions and GPNs: first, it is time- and space-contingent, subject to change but also path-dependent; second, it is multiscalar and transcends territorial boundaries; third, it is strategic because the coupling process comes about through the intentional and active intervention of multiple actors.

Mediating the strategic coupling process is a range of institutions operating within and beyond the region, in particular government, labor, and business agencies, with the aim of generating beneficial outcomes in terms of value generation, enhancement, and capture, and hence economic development through the integration of regions into wider GPNs. Positive outcomes from globalizing regional development through strategic coupling are not guaranteed, however; insertion into GPNs may be detrimental in specific circumstances, depending on the power configurations and asymmetries in the relevant networks. Coe and Hess (2011) term this the “dark side” of development, which can manifest itself in the form of ruptures within regions (e.g., political exclusion, displacement or eviction of people) and between regions and GPNs (e.g., disinvestment, exit of foreign firms). It can also result in friction when such connections cause social and economic tension and conflict within and beyond the region, such as social and class conflict, struggles over uneven value capture, labor exploitation, or growing gender inequality. In such cases, institutions will often aim for strategic de- or recoupling (Horner 2014) to improve regional developmental outcomes.

GPNs are part of development as the geographically uneven and historically contingent expansion and extension of capitalist systems of production, exchange, and regulation. They are also an arena for organized intervention by private and public actors with different implicit and explicit goals (Hess 2009). As firms in GPNs develop their networks of capital accumulation, they are always operating in the context of (potential) societal and social resistance and protective social movements, a process that has been described by Karl Polanyi (1944) in his institutionalist view of development as the “double movement.” This process plays out through organizational networks and coalitions across different scales, and GPN research grounded in relational concepts of space needs to be conscious of this if it is to produce non-reductionist understandings of globalization and regional development, in both their material and discursive dimensions.

Arguably, GPN analysis has much to offer in terms of research into regional development and the experiences of different places enmeshed to various degrees in global networks, but of course it cannot claim to be fully adequate to capture all dimensions of development, let alone be the only heuristic to guide empirical research. In a sympathetic critique, Kelly (2013) argues that some crucial elements for fully understanding development in place are obscured by exclusively utilizing the GPN concept. For example, he highlights the importance of environmental and landscape changes that usually escape the gaze of GPN research unless they are immediately relevant for the transnational and regional firms being studied. A similar “blind spot,” in Kelly’s view, is the realm of households and families. He attributes these shortcomings to a tension between the network ontology underlying a
relational GPN framework and a place-based ontology found in other work on (regional) development that assumes localities and regions as “complete” – albeit not bounded – entities and spaces of lived experience. A dialogue and exchange between the GPN concept and cognate approaches, recognizing the possibilities and limitations of one another, needs to continue as part of a critical cultural political economy of GPNs and regional development.

GPNs: a politically contested field

To conclude, the GPN framework as developed since the turn of the century provides a heuristic device with which to analyze the complex realities of economic globalization and economic development. On the ground, GPNs are not only networked forms of organization and value creation in increasingly global markets, but also political in character – making them highly contested fields (Levy 2008). They connect multiple sites of struggle for the creation, enhancement, and capture of value, shot through with power relations that ultimately determine the implications of GPNs for territorial development. One of the antecedents to the GPN framework, global value chains, has recently gained much prominence in international policy circles, from the International Monetary Fund and the World Bank to the World Trade Organization and the International Labour Organization. Indeed, some are concerned that GVC analysis is running the danger of becoming another neoliberal tool, despite its rootedness in critical social science, as its central ideas are “translated” into the international development policy realm that, critics argue, still follows orthodox economic development ideas and practices (e.g., Fernández 2014; Neilson 2014). Thus it is imperative for global production network analysis and cognate approaches to maintain their critical focus, investigating GPNs as both economic and institutionalized, highly political, and discursive structures with far-reaching and often contradictory developmental outcomes.

SEE ALSO: Corporate financialization; Corporations and the nation-state; Development; Global commodity/value chains; Globalization; Governance and development; International division of labor; Labor geographies and the corporation; Relational assets; Vertical integration

References

Global Production Networks Perspective.” *Transactions of the Institute of British Geographers*, 29: 468–484.


Further reading


Globalization and rural areas

Michael Woods  
Aberystwyth University, UK

Globalization is a major driver of change in rural areas around the world, impacting on economic activities, structures, and relations; migration patterns and the composition of rural populations; and cultural traditions and practices of environmental management. Globalization is a complex and multifaceted phenomenon, involving the multiplication, stretching, and intensification of social, economic, political, and cultural relations over space which leads to places becoming more tightly integrated in transnational networks of inter-dependence. As such, globalization challenges popular perceptions of rural areas as places that are isolated from mainstream economies and culture and that retain greater economic independence and cultural distinctiveness than cities. In practice, many rural areas have long been integrated into global networks of trade and migration, notably through colonialism. However, late twentieth-century and early twenty-first-century globalization is distinguished by the reach and intensity of transnational connections, and by the immediacy of communications and transactions between distant places.

In practice, many rural areas have long been integrated into global networks of trade and migration, notably through colonialism. However, late twentieth-century and early twenty-first-century globalization is distinguished by the reach and intensity of transnational connections, and by the immediacy of communications and transactions between distant places.

Research on globalization in geography has tended to focus on urban areas and particularly on the concept of the “global city,” to the relative neglect of rural areas. This tendency has contributed to the conflation of globalization and urbanization in analysis of rural change, with the erosion of rural cultural and economic distinctiveness attributed to integration into urban systems, which in turn have become increasingly interconnected through globalization. As such, rural experiences of globalization were positioned as mediated through the concentration of power in global cities. This analysis was particularly associated with research conducted from a political economy perspective, which focused primarily on structures, processes, and institutions rather than on communities, and thus emphasized dependency and convergence over local difference.

More recent rural research, however, has drawn on the relational approach to globalization developed by geographers including Ash Amin and Doreen Massey, which sees places as unique entanglements of social, economic, and political relations that are reconfigured in globalization. The relational approach has a number of implications for the analysis of globalization in a rural context. First, by highlighting the complexity and contingency of the relations that link places, and the uniqueness of their intersection in a specified place, it suggests that globalization produces cultural and economic hybridity, not homogeneity. Second, by revealing relations that can be rural-to-rural as well as rural-to-urban, it shows that globalization is not necessarily always mediated through global cities. Third, by arguing that globalization is reproduced through the reshaping and substitution of relations within place, it demonstrates that the outcomes of globalization are not inevitably positive or negative. And, fourth, by pointing to the diffusion of power in globalization, not its concentration, it indicates that rural actors have agency to affect the outcomes of globalization in their localities. These features are presented by Woods (2007)
as characteristics of an emergent “global countryside” in which rural places are changed by globalization but continue to be different from each other (see also Aguayo 2008).

**Rural areas and the global economy**

One of the most significant expressions of globalization in rural areas has been the incorporation of rural economies into longer and more complex commodity chains, linking sites of production and consumption. The development of global commodity chains (or global value chains) has been facilitated by technological advances, market deregulation, and trade liberalization affecting all economic sectors, but notably in a rural context in relation to agriculture and food (see, e.g., Challies and Murray 2011). Between 1970 and 2010, the value of global agri-food exports increased from less than US$100 billion to over US$1100 billion per year, with a greater share of food being produced for export, and the distance traveled by produce bought by consumers (known as “food miles”) increasing. The lengthening of these supply chains has been enabled by improvements in transport and preservation technologies that have changed the principles of agricultural geography by allowing perishable food to be sold over longer distances, and also by trade agreements that have dismantled controls on agri-food imports designed to protect domestic farmers. Free markets for agri-food products have been created within regional economic blocs such as the European Union and the North American Free Trade Agreement (NAFTA) area, while negotiations at the World Trade Organization (WTO) have attempted to remove trade barriers and abolish production subsidies in agriculture under pressure from the Cairns Group of agri-food-exporting nations.

Global agri-food commodity chains do not only connect sites of production and consumption in different countries, but can also involve commodities being sent for processing or packaging in third or fourth countries, as companies seek cost efficiencies. Accordingly, the spatial division of labor in agriculture has become increasingly fragmented, with impacts for rural localities. The weakening of local supply chains has altered social and cultural relations between farmers and local communities; food-processing industries that had located close to sites of production have closed down as their functions are concentrated in lower-cost locations; the incomes of individual small farmers have been squeezed by competition from cheaper imports, and their financial power reduced; and landscapes have been transformed as production has been switched to new commodities with higher market value or export potential. Globalization has consequently promoted increased regional concentration and specialization in agricultural production, including the relocation of industries such as intensive livestock farming to regions that offer not only lower production costs but also looser environmental and animal welfare regulations.

Footloose economics is evident in rural areas in respect to not only agriculture, but also other traditional industries including mining and forestry, with smaller mines and paper and lumber mills closed as production is concentrated in fewer, larger, sites. The search for new resources to supply global markets has led mining companies to increasingly remote locations (often staffed by “fly in, fly out” shift workers), and spurred both deforestation in regions such as the Amazon and new commercial forestry plantations in areas such as Australia. In manufacturing, some regions such as rural Ireland benefited in the late twentieth century from new branch plants attracted by land availability and low labor costs;
yet some of these factories have subsequently been closed – along with longer-established factories in industries such as textiles – as production has been relocated to more cost-effective locations in Central America and Southeast Asia. In some rural areas, the closure of a mine, mill, or factory has meant the loss of the dominant employer in a single-industry town, prompting community decline and depopulation.

Economic globalization is also associated with the increased influence of transnational corporations that dominate markets in traditional rural industries, having squeezed out or taken over local firms. Mining and forestry are dominated globally by a handful of companies that each operate on a transnational scale. In agriculture, while the majority of the world's farmers are still small-scale individual farmers, agribusiness accounts for an increasingly substantial share of global production, and key parts of agri-food commodity chains are controlled by limited numbers of transnational corporations. These include biotechnology firms with control over inputs such as seeds and fertilizers, and supermarkets, food-processing companies, and fast-food brands, which are the major purchasers of agricultural produce. In some cases, companies have formed strategic alliances to achieve “vertical integration,” with interests at each step of the commodity chain, from “seed to shelf” (Hendrickson and Heffernan 2002). In this way, relations between transnational corporations and individual farmers are unequally weighted. In India, for example, transnational biotechnology firms have been accused of pushing genetically modified seed that would make farmers dependent on their supplies and chemicals. Similarly, supermarkets and fast-food corporations can use their market power to depress prices paid to markets, standardize requirements for the appearance and variety of produce, and determine practices in animal welfare, pesticide use, and labor conditions.

Corporate power in the global agri-food system is exercised not only by agribusiness and supermarkets, but also by banks, pension funds, and traders, who supply the financial capital for investment and set the prices of agricultural commodities. Financial speculation has been an important part of capitalist agriculture since the nineteenth century, permitting farmers to off-load risk by selling future crops to speculators at guaranteed prices. However, deregulation of trading in food futures in 2000 sparked a surge in investment in food derivatives from US$3 billion in 2003 to US$126 billion in 2011, which has been accused of contributing to the sharp rise in global food prices in 2008, but which has had little effect on farm incomes. Over the same period, international investments in rural land have also increased, colloquially referred to as “land grabbing.” These have in part been driven by concerns about food security following the 2008 crisis, with sovereign wealth funds in affluent but land-poor countries such as Qatar, Saudi Arabia, and the United Arab Emirates, and corporations in Japan and South Korea, purchasing land in Africa, Central and South Asia, and Australia, in order to secure the supply of food for domestic markets. Yet, private investment brokers and funds have also purchased farmland in Africa and parts of the former Soviet Union as an opportunity for Western investors. Although some of these projects are presented as assisting agricultural development, international land investments have been criticized for replacing staple crops with cash crops, diverting resources such as water, extracting profit from poor rural communities, and in some cases displacing peasant farmers.

Indeed, across much of the Global South, and notably in South America, globalization is associated with depeasantization, as the
livelihoods of peasant farmers have been threatened by the opening of local markets to imported food and the diversion of resources to support export-oriented industrial agriculture. To survive, farmers have been forced to become migrant workers, or paid farmworkers on large corporate farms, in some cases selling their land to agribusinesses (see, e.g., Echánove 2005; Rigg 2006).

International migration and rural areas

Alongside economic globalization, rural areas have also been affected by the increasing global mobility of people as migrants and tourists, acting as both a source of and a destination for international migrants. As described, the impacts of economic globalization have contributed to a flow of labor migrants from disadvantaged rural areas, especially in the Global South, to more prosperous regions both nationally and internationally. In many cases, migrant workers leave families at home, at least initially, and remittances from labor migrants have become a major source of income in many rural areas. Furthermore, as labor migrants from the Global South tend to be young men, the demographic character of rural communities is changed, as are roles within the community, including the feminization of agricultural work in some parts of Africa and Asia.

Although most labor migrants head to cities, there has been a growth in the number of labor migrants living and working in rural areas of Europe, North America, and Australia, commonly employed in marginal jobs in agriculture and food processing, with limited pay and conditions. As such, migrant workers have filled a void created by efforts to achieve global competitiveness by minimizing labor costs that have deterred recruitment of local workers and made industries such as horticulture, meat-processing, and salad packing dependent on immigrant labor. Some of the highest growth rates in the Latino population in the United States since the mid-1990s have been in rural counties in the South and West, especially areas with large-scale poultry production, while migration from Eastern Europe to Britain since the enlargement of the European Union in 2004 has produced new concentrations of Polish residents in areas of intensive agriculture such as Lincolnshire. Many of these localities have little previous experience of large-scale immigration, and tensions can develop over cultural differences and competition for jobs and housing. Yet, Torres, Popke, and Hapke (2006) point to a “silent bargain” in which rural communities accept labor migrants for the economic contribution they make, and migrant workers accept low wages and poor employment conditions for the perceived safety and tranquility of living in a rural area.

Rural communities can also be transformed by international amenity migration, an extension of domestic counterurbanization with individuals moving to rural areas abroad for a combination of economic and lifestyle reasons. Amenity migration is associated with rural localities that offer recreational amenities (mountain and coastal resorts) or idyllic landscapes and historic villages, but also cheap, empty property left by depopulation or a low cost of living, and access by budget air travel, including parts of southern and Eastern Europe and Central America. In particular, amenity migrants are commonly attracted to areas that are already popular with international tourists, and there is a blurred line between amenity migration and residential tourism, with some migrants moving permanently but others occupying foreign properties only seasonally or for vacations. Both international tourism and amenity migration can boost rural economies and help revive declining villages, but they can also generate tensions around infrastructure.
Globalization and rural areas

developments, the inflation of property prices, and the commodification of local cultures for tourist consumption (see, e.g., Woods 2011).

Global consciousness and rural environments

In addition to the multiplication, stretching, and intensification of social and economic relations, some commentators also describe globalization as including the growth of global consciousness, as improved communications, increased travel, and the growth of the global media have made people better informed about other parts of the world and more likely to think in global terms. This has affected how people perceive rural areas, with understanding of rural life informed by globalized media representations that blur geographical differences. The growth of global consciousness has also involved a globalization of values, in which there is an expectation that universal standards apply across the world, regardless of historic cultural differences. The globalization of values is promoted by transnational nongovernmental organizations (NGOs) and campaign groups, which can elevate localized issues to international causes, and is encoded in international treaties and agreements. In rural areas, conflicts have arisen when globalized values clash with traditional cultural practices, such as seal hunting in Canada or bullfighting in Spain, or farming activities such as the live export of livestock. Similarly, the transnational conservation movement has promoted standardized approaches to environmental protection, including the designation of protected areas (such as national parks) on international models that introduce restrictions on the traditional use of natural resources by rural communities or, in some cases, the eviction of people from designated zones. There are estimated to be up to 136 million “conservation refugees” globally, who have been displaced from protected areas (see Brockington and Igoe 2006).

Rural responses to globalization

Globalization extends to all parts of the rural world, but different rural regions experience globalization in different ways. As globalization includes a multitude of different processes, some will have greater impact in some regions than others, and the balance of negative and positive impacts will also vary (Woods 2007). The impact and outcomes of globalization in a particular region are shaped by a number of factors including geographical location, economic structure, natural resources, transport infrastructure, and human capital, as well as by the policies and actions of governments and other local agents. In many regions, rural development strategies have attempted to engage proactively with globalization by seeking to attract foreign direct investment (FDI), encouraging international tourism, or supporting entrepreneurship to develop export markets for local products (see, e.g., Cheshire and Woods 2013). At the same time, revived interest in regional cultures and efforts to promote local food systems can also be seen as responses to globalization, as people react to a perceived loss of identity or economic independence.

Responses to globalization can also be more confrontational, as activists attempt to resist changes that are perceived to threaten local economies, cultures, and environments. Protests by farmers against food imports or the influence of transnational corporations have occurred in several countries, including Britain, France, Australia, India, and South Korea, with actions
including blockades of ports and attacks on symbols of globalization such as fast-food outlets. In Western countries such actions have not generally been coordinated internationally or been framed as antiglobalization protests. More direct links to the counterglobalization movement have been made, however, by rural movements in the Global South, especially in Latin America and Asia, and notably through the international farmers’ movement La Via Campesia (The Peasants’ Way). Rural activists have joined counterglobalization protests at major international summits, including WTO meetings, and regional concerns have been highlighted by transgressing space to reach international audiences, with, for example, Indian farmers mounting a series of demonstrations in Europe to protest against trade agreements and the introduction of genetically modified (GM) crops by transnational corporations (Featherstone 2003).

SEE ALSO: Agricultural geography; Agri-food multinational enterprises; Colonialism, decolonization, and neocolonialism; Conservation and capitalism; Corporations and global trade; Food security; Global cities; Global commodity/value chains; Globalization; Governance and development; Labor migration; Migrant labor; Migration: international; Mining and mineral resources; Neoliberalism; Tourism

References


Globalization

Matthew Sparke
University of Washington, USA

A specter haunting the world with both promise and peril, globalization names a process of increasing but uneven global integration that is bringing market-mediated interdependency along with enormous inequality, asymmetry, and volatility to life across the globe (Sparke 2013). As a promotional and promissory Discourse, globalization has been increasingly politicized in ways that associate observations about global market integration with repeated calls for more market freedom. Thomas Friedman, for example, is a columnist for the New York Times who has made a name for himself globally by recommending pro-market reforms everywhere on the basis of arguing that the world is becoming flat, and that we must adapt or just get flattened. Globalization, he says, “involves the inexorable integration of markets, nation-states, and technologies to a degree never witnessed before – in a way that is enabling individuals, corporations and nation-states to reach around the world farther, faster, deeper and cheaper than ever before. The driving idea behind globalization is free-market capitalism – the more you let market forces rule and the more you open your economy to free trade and competition, the more efficient and flourishing your economy will be” (Friedman 1999, 7–8).

The sudden uptake in the use of the actual word “globalization” in the last two decades of the twentieth century has clearly been linked to such advocacy of market rule, or what critics call Neoliberalism (Harvey 2005). For the same reason, its emergence as a key term of political debate needs to be examined in relation to at least three intersecting global imperatives: first, the economic shifts of the post-Fordist era (from roughly the 1970s onwards) in which corporations shifted from the Fordist balancing of mass production and mass consumption nationally to pursue market opportunities globally; second, the ideational achievements of pro-market academics, advocates, and think-tanks in promulgating neoliberal orthodoxy as a “business-knows-best” consensus; and third, the global political and economic dominance of the United States, which has led to a series of complex associations between globalization, Americanization, and the ups and downs of US hegemony.

Between militarized shock therapy for some and the commercial advance of McDonaldization, CocaColanization, and Wal-Martization on others, the variegated triangulations of globalization-neoliberalization-Americanization have been best theorized as a form of informal, market-mediated Imperialism (Gindin and Panitch 2012). Some of the associations with more overt forms of US militarism in the Bush administration’s “global war on terror” may also at least partially account for the fall-off in the use of the non-US spelling “globalisation” (with an “s”) after 2003 (see Figure 1).

But these terminological tendencies noted, globalization should by no means simply be viewed as American, unidimensional, or unipolar, especially now that Asian economic growth, the new triangles of Asian investment in Africa and Latin America, and the emergent terms of “Asianization,” “Easternization,” and “ChIndia”
Moreover, as feminist scholarship further underlines, the ties and tensions of global integration are also embodied in intimate power relations that are as uneven and violent as they are intertwined with the asymmetries associated with changing configurations of global hegemony (Mountz and Hyndman 2006). Christa Wichterich, the author of *The Globalized Woman*, puts it like this: “For women around the world … globalization is not an abstract process unfolding on an elevated stage. It is concrete and actual. Female textile workers from … Eastern Germany are losing their jobs to women in Bangladesh; Filipinas clean vegetables and kitchens in Kuwait; Brazilian prostitutes offer their services around Frankfurt’s main railway station; and Polish women look after old people at rock-bottom prices in … Germany” (Wichterich 2000, 2).

It is amidst such embodied experiences of globalization, at once global and personal, that diverse articulations of precariousness are being critically reinterpreted as the basis for more emancipatory kinds of global consciousness and collaboration. Global precarity may in this sense inspire global solidarity (Butler and Athanasiou 2013). The globalization and adaptation of the *Indignados* and *Occupy* movements following the 2008 financial crises can certainly be interpreted this way. But, by the same token, the sad fate of the Arab Spring that co-inspired the *Indignados* and *Occupy*, indicates that cruel winters of precariousness can return with a vengeance, and, in Egypt at least, new rounds of neoliberalization too (Atia 2013).

Reimagined in terms of planetary precariousness, ideas about living in a fragile global ecumene may also create the basis of transnational communities that can globalize both human health rights and (more-than-human) environmental security in the face of market-induced insecurities. But for global precarity to inspire solidarity in this way, it clearly needs to be distinguished from the personalized, marketized, and, as such, increasingly neoliberalized practices of “resilient life” – practices that invoke the
inevitability of catastrophes of global finance, global health, and global climate change only to advocate for individualized and enclaved efforts at market-mediated adaptability. Against such apocalyptic-turned-adaptive accounts of enduring the “Anthropocene” (see Anthropocene and planetary boundaries), it is possible instead to explore how the interdependencies of the market and its degradation of the global commons nevertheless open opportunities for the active creation of the world as a common community beyond the market. Such alter-globalization efforts consciously politicize the political unconscious of the Anthropocene, rendering the inevitability invoked in pro-market globalization discourse open to resistance in the name of global climate justice, and thereby also opening globalized pathways out of the Kyoto–Copenhagen–Cancún–Durban dead end (Bond 2012).

Before contemporary climate justice struggles, the globalization of anti-neoliberal resistance in other areas has long been discussed by critical theorists. David Harvey’s revision of Marx’s notion of “primitive accumulation” (generally thought of as a pre-capitalist precursor moment of dispossession in which workers were deprived of their capacity for subsistence) as ongoing “accumulation by dispossession” (rethought as a form of ongoing destruction or removal or privatization of the commons and the personally shared) works well to further highlight how we have been witnessing a vast global accumulation of grievances against market-mediated dispossession (Harvey 2005). Conceptualizing all these responses as a singular global response to globalized capitalism, however, risks romanticizing resistance as well as obscuring how exactly ongoing value extraction and accumulation is tied to diverse forms of extra-economic domination. Most notably, for example, this is the danger in suggesting that a so-called Empire of networked global capitalism is co-creating its own nemesis in the form of a globalized and media-enabled “Multitude”: a multitude that is becoming interconnected and experimental enough to break out from within the matrix of atomized consumerism and contingent labor control. In order to be more materialist about such emancipatory possibilities, feminist geographers suggest that we need first to denaturalize the disabling “impact model” of globalization as an inevitable, unstoppable, and leveling juggernaut of market-led integration. And key to this denaturalization of dispossession is the challenge of replacing the flat world visions used by pro-market promoters to naturalize neoliberalism with more accurate geographies of the inequalities, asymmetries, and unsustainability in the actual real-world experience of uneven global integration.

The discipline of modern geography, with its origins in enabling early imperial rounds of globalized dispossession, is by no means a “natural” place from which to denaturalize dispossession. Indeed, as critically informed as the field of geography now is, the asymmetries in its globalized patterns of publication still point to Anglo-American dominance and the tendency towards exclusion of non-English accounts from publications – including this Encyclopaedia – that aspire to move to more inclusive and collaborative kinds of intellectual globalization. Reciprocally, some of the most powerful geographical critiques of flat-world globalization discourse have come from outside the Global North and outside the discipline. One of the best textbooks on globalization available, for example, is a brilliantly geo-graphic (as well as graphic) account by the Mexican cartoonist El Fisgón (Rafael Barajas Durán), published under the subversively subaltern title of How to Succeed at Globalization: A Primer for the Roadside Vendor. “Location, location, location,” is what El Fisgón’s main character – a roadside
vendor from Mexico City – is told when he asks how he might succeed at globalization; the point being that he really needed to have started out in another household, in another city, in another wealthier part of the world in order to benefit from the skewed arrangements of opportunity and exploitation that shape global integration (Fisgón 2004). This kind of geographical critique is also echoed in Vandana Shiva’s powerful retort to Thomas Friedman himself. “Friedman,” she writes in a scathing review of The World is Flat, “has reduced the world to the friends he visits, the CEOs he knows, and the golf courses he plays at. From this microcosm of privilege, exclusion, blindness, he shuts out both the beauty of diversity and the brutality of exploitation and inequality, he shuts out the social and ecological externalities of economic globalization and free trade, he shuts out the walls that globalization is building – walls of insecurity, hatred and fear – walls of ‘intellectual property’, walls of privatization” (Shiva 2005).

Overcoming the epistemic exclusions and blinders of its own imperial roots and routes, research within the formal discipline of geography is also now mapping the uneven political geographies of dominance, governance, and resistance shaping actually existing globalization. Attuned to the powerful presentism through which neoliberal globalization discourse asserts its own novelty, such work has further underlined how flat-world ontologies are inevitably ahistorical too, despite having historical-geographies. To be sure, some poststructuralist geographers (see Poststructuralism/poststructural geographies) further suggest that we need another kind of flat ontology in order to challenge fantasies about equalized access through market flattening. But, inspired by thinkers such as Badiou, Derrida, and Deleuze and Guattari, they do so deconstructively with a view to shaking off preconceived axioms about what sorts of spatial connection and hierarchy get to count as globalized in the first place. This in turn makes it possible to go from debunking globalization as discourse to examining its consequential framing effects in ways that include attention to the uneven economic developments that the dominant flat-world discourse itself tends to obscure.

One area where the assertive framing power of globalization discourse has been especially obvious (and obscuring) is in international relations and associated diplomatic, military, and foreign policy discussions that frame war, peace, and security in the terms of, on the one side, disconnection and geopolitical danger, and, on the other, global integration and geoeconomic opportunity. The actual term “geoeconomics” is still not widely used, and the work of the French economic geographer Jacques Boudeville, who first developed it, remains largely unknown. Nevertheless, as market-led globalization has intensified, geoeconomic arguments and visions have become increasingly popular and have even been formalized by Edward Luttwak as a new grammar for foreign policy. In place of orthodox geopolitics and its concerns for soldiers and citizens, this geoeconomic grammar tends to elevate the entrepreneurial interests of investors and customers; in contrast to a geopolitical focus on national borders and place, it privileges networks and pace; and instead of concentrating international politics on building alliances for “security” against supposed “evil empires,” geoeconomics is primarily concerned with building international partnerships that advance “growth,” “integration,” “harmonization,” and “efficiency” against the threats of “traditionalism,” “isolationism,” “anachronism,” and “anarchy.” In this dualistic discursive system, globalist claims about the “borderless world” and ideas about geoeconomics eclipsing Geopolitics have become influential scripts accompanying claims about globalization’s supposed historical novelty – even
though historical-geographical work documents that earlier rounds of global integration themselves came with similar talk of transnational capitalist expansion and opportunity.

While historical work points up the dangers of too quickly periodizing and separating a geopolitical past from a geoeconomic present, ongoing geographical research points in turn to the need to avoid partitioning the globe into geopolitical spaces defined by disconnection and geoeconomic spaces of globalized connection. Postcolonial geographies of the colonial present instead make clear that the supposedly “disconnected places” – places that are depicted by partisans of geoeconomics as mired in geopolitical strife because of a lack of globalization – have in fact long been connected through predatory forms of imperial connection and biopower. This point is also ironically confirmed by today’s apologists for empire who map the borderlands of intervention in terms of making the world safe for globalization and minorities at the very same time. In another way, political geographies of actual real-world borders have made clear that border regions are sites where, despite all the attempts to envision borderless geoeconomic futures, geopolitical ideas, affects, and imperatives keep coming back to shape practices of (b)ordering transnational flows on the ground. So rather than locate geopolitics in the past or in regions wrongly represented as disconnected, it seems far more useful to see geopolitics and geoeconomics as entangled, geostategic discourses that encode underlying tensions of global uneven development (Sparks 2013, Chapter 8). These tensions can in turn be at least partly explained in terms of the ongoing tensions between spatial fixity and spatial expansion in capitalism itself (Harvey 2005). And while all sorts of other imperatives ranging from the personal and emotional to the national and territorial to the global and ecological also overdetermine uneven development, much of the ongoing oscillation between geopolitical and geoeconomic discourses can thus be parsed in terms of the tensions between capitalist fixity and expansion that have made the long-term development of globalization so episodic, asymmetric, and geographically transformative.

SEE ALSO: Anthropocene and planetary boundaries; Discourse; Empire; Fordism/post-Fordism; Geopolitics; Imperialism; Neoliberalism; Poststructuralism/poststructural geographies

References


GLOBALIZATION


Glocalization refers to the interaction of the spatial scales of global and local influences, with universalizing influences and models modified to fit local preferences and practices. In this regard, “glocalization” is less global and more a local response to perceptions of modernization requirements. Studies of glocalization appear in diverse fields. Originally a Japanese business concept for taking initially local products to a wider market, the term was later adopted by academics in the social sciences (Castree, Kitchin, and Rogers 2003). A popular jingo exhorts individuals to “think globally, act locally.” Glocalization is an issue for developing regions seeking to modernize by combining traditional systems with global practices for acceptance into or by a desired larger-scale group. Glocalization asserts the necessity for rejecting homogenization, given the need to retain some forms of locally specific traditional values or practices.

Concepts of time and distance are frequently involved when practices are intermixed from different locales and cultures, and the Western/developed becomes “global” versus the local “other” (Robertson 1996). Interaction between scales means that systems such as Japanese corporate-level “just-in-time” and “lean manufacturing” can permeate a global set of production practices that result in the subnational shaping the supranational (Sharma 2009). Rather than a polarity of opposite or differentiated parts, the term refers to a blend of complementary practices whose territorial identification occurs across political borders, interacting at a different cultural or intercommunication level. Transborder and within-border ethnicities with supra-border extensions also exemplify glocalization processes that display the effects of transboundary ties. Rather than referring to the border-transcending quality of the multi- or transnational, glocal emphasizes the incorporation of nonglobal specificities. The increase in physical and virtual mobility underlies the use of this term in a number of fields, signifying accelerating hybridity and an awareness that to understand this phenomenon requires a sensitivity to embedded values. Corporations selling products in varied countries and markets need to “glocalize” by adapting their portrayal to segmented market attitudes regarding cultural portrayals, for example.

SEE ALSO: Development; Economic development zones; Local/global production systems; Modernity

References


Governance and development

Richard Ballard
University of Witwatersrand, South Africa
University of Johannesburg, South Africa

The origins and locus of governance

As a heuristic tool, governance draws attention to the fact that authority is plural and diffuse rather than singular and central. Governance is accumulated in many different spheres of authority (Rosenau 1997). Government itself is not a single site of authority but rather a kaleidoscope of intersecting political affiliations, ministries, jurisdictions, and scales. Authority occurs in many spheres beyond government, as can be seen in relationships between employers and employees, customers and suppliers, nongovernmental organizations (NGOs) and beneficiaries, donors and NGOs, leaders of a political party and members, and religious leaders and their followers. Furthermore, governance can occur in the absence of any authority insofar as processes of socialization guide everyday practices through informal rules and conventions.

These many spheres of authority are, in turn, shaped and contained by one another. Within a national context, the conduct of the government is itself governed by the law, the results of elections, oversight committees, and watchdog organizations. Meanwhile, national government is informed by the international context. For example, the way in which individual nations govern trade across their borders is shaped by agreements reached at the World Trade Organization. Therefore authorities do not govern in isolation but are themselves governed and circumscribed by superiors, peers, and citizens, and by intersecting and competing spheres of authority.

Governance is strongly defined by the legacy of systems established in the past. Contemporary authorities do not start from a blank slate but inherit an established bureaucracy which cannot easily be dismantled or changed. Authorities often commit themselves to incrementally improving existing systems or to introducing fixes for limitations. Major new bureaucratic capacities require enormous investment and political will. In 2011, for example, India began a process of registering hundreds of millions of people according to biometric information such as iris and fingerprint scans in order to enable direct welfare distribution (Szreter and Breckenridge 2012).

In addition to temporal considerations, governance has spatial qualities. Some authorities have spatially defined jurisdictions, as in national governments over a national territory, a province or state over a subnational region, or a municipal government over a city. This can have important material outcomes. For example, water delivery may be entirely different between two nearby cities as a result of the divergent approaches of the respective municipal governments. However, territorial boundaries do not always demarcate governance in a simple manner (Agnew 1994). Authorities do not have complete sovereignty over everything within their borders, and powerful influences can originate elsewhere. Furthermore, authorities govern their territories...
and populations in uneven ways. Chatterjee (2004) argues, for instance, that the way in which middle classes are governed in India is very different to the way in which poor populations are governed. Moreover, authorities do not simply “govern” people en masse by virtue of their happening to be within a particular territory, but relate to people through specific social interactions and diverse modalities (Allen 2003).

The scale at which populations and processes are, and should be, governed is an important strategic consideration. The foundational unit of governance in the modern period is the Westphalian nation-state which has enormous capacity to set and police legal parameters, and to acquire and allocate tax revenue. Yet nations may enter into cooperation in order to achieve social objectives that transcend their territories. For example, the efforts of one nation to reduce the emission of greenhouse gases are ultimately futile if emissions are not reduced globally, and so cooperative efforts have been underway to secure international governance systems on carbon emissions. Many more multilateral institutions such as the United Nations, the European Union, and the World Bank have resulted in governance parameters being defined beyond individual nation-states.

While the governance of some issues occurs at the supranational level, other responsibilities have been shifted toward local scales (Batterbury and Fernando 2006). Some advocates of reform have argued that some responsibilities should be devolved to subnational regional government units and local government in order to achieve greater efficiency and accountability.

These shifts in scale have enormous implications for the raising of revenue, enforceability, and the way in which different interests try to assert their own advantage by moving between places in order to find favorable governance environments. For example, the flight of wealth to tax havens presents an important governance challenge.

Shifts in scale are evident in other spheres of governance, such as the way manufacturers govern themselves. Under Fordist conditions, manufacturing was vertically integrated. The trend toward post-Fordism in the late twentieth century was characterized by the outsourcing of some, or even all, production processes to sometimes distant suppliers. Outsourcing has replaced one kind of governance, of employees by employers within a factory, with another kind, of suppliers by customers within a value chain. Suppliers comply with price, quality, and other pressures for fear of losing their contracts. Manufacturers stand to benefit from reducing the complexity of their direct responsibilities and the option of locating production in regions where labor is cheaper.

Steering people, processes, and organizations

As Antonio Gramsci (1971) argued, authorities use a combination of coercion and consent in order to attempt to bring about specific or general processes and outcomes. Coercion refers to the way in which authorities assume, or are granted, the ability to force people or organizations to conduct themselves against their inclinations. Disciplinary measures are a normalized feature of modern society. For example, traffic fines deter drivers from speeding, the threat of disconnection compels consumers of electricity and water to pay their bills, and the police may repress protests. In contexts where political rights are not recognized, authorities can resort to draconian methods. In the late 1800s, for example, colonial authorities in southern Africa attempted to force Africans to
enter the workforce by imposing taxes so that they would have to earn a wage.

While coercion is ubiquitous, and authorities at particular junctures try to dominate society, there are limits to this form of power. For example, many city governments in the Global South have attempted to ban street trading but eventually given up because people trade on the streets irrespective of the illegality of doing so. Neither can a goal such as gender equality be fully achieved through decrees and sanctions from above. Large social processes cannot be controlled through force in any absolute sense and there is a limit to coercion in the absence of popular buy-in.

Within liberal democratic contexts, authorities ostensibly relate to free citizens. Rather than micromanaging everyday life, liberal authorities regard citizens as free agents who make their own choices (Miller and Rose 2008). The role of authorities within liberal contexts is to establish governance frameworks which are akin to the rules of a game. This is not to say that authorities cannot direct people. Behavioral economists argue that people can be steered through incentives and disincentives. For example, taxes on alcohol and cigarettes are intended to discourage consumers from buying them without forbidding them from doing so. Meanwhile incentives can be provided to encourage certain behavior. In Brazil and Mexico, grants are awarded to poor families on condition that they keep their children in school.

The power of modern governments tends to derive from the alignments between those who govern with the majority of those who are governed. Where such alignment exists, most people and organizations conform to rules voluntarily, and authorities only need to police a defiant minority. Those in authority therefore govern with the consent of most subjects. As Michel Foucault (2008) has shown through his work on governmentality, this consent is itself manipulated. The “art of government,” he argues, is to shape subjects’ values, desires, and habits so that people internalize what is required of them, thereby largely negating the need for discipline.

The role of bureaucracy

Centralized bureaucratic authority has occupied an ambivalent position within development thinking. The decades following World War II were characterized by Weberian optimism that bureaucracies were the foundation of an ordered and modern society. Experts and professionals would solve social problems by collecting and scientifically analyzing data, and arriving at rational and impartial interventions.

Over time this optimism has given way to concerns that technocratic approaches are not necessarily benign. Top-down programs run the risk of prioritizing the vision of experts and technocrats over the welfare of vulnerable groups. The goal of slum-free cities is, at face value, a well-meaning ambition in that slums are poor living environments. Yet overzealous governments who remove people from slums in an attempt to create slum-free cities leave those evicted in a much worse off position (Huchzermeyer 2011). Also, Marxist approaches argue that in capitalist societies, states are not neutral arbitrators but tend to side with privileged interests and to support the process of capitalist accumulation (Glassman and Samatar 1997). Likewise, experts may present themselves as neutral, but sometimes represent elite economic interests over the needs of the poor (Ferguson 1990).

Technocratic approaches also necessarily flatten a complex world through their attempts to apprehend it through data and categorizations. These simplifications are made in order to render
complex society more “legible” (Scott 1998, 2),
but they result in a single bureaucratic response

to a great deal of individual variety. These
abstractions can, in turn, have disabling conse-
quences for those they classify. Cash transfers in
Mexico have been criticized for adding to the
burden of women by casting them in a maternal
role (Molyneux 2008). Furthermore, processes
of categorization can divide people, resulting in
fundamentally different treatment. The Chinese
*hukou* system of household registration was used
in the twentieth century to discourage urbaniza-
tion. However, large populations have migrated
to cities for work anyway but, as a result of their
rural registration, they are not covered by urban
social support systems.

Influential concerns about bureaucracies were
also raised by free-market, or *neoliberal*, thinkers
that the “ambitious and impatient reformer”
(Hayek 2011/1960, 376) does more harm than
good, and central authorities are not benevolent
(Lal 1985). These critics feared that bureau-
cracies became self-justifying, commandeering
ever more resources and imposing themselves
on parts of society that would be more likely to
succeed on their own. By the 1970s and 1980s,
various economic crises were attributed to
bureaucracies overreaching themselves, resulting
in price distortions and the crowding out of the
private sector.

More generally, the once positive image of
bureaucracy has unraveled as a result of the fact
that, in practice, it has not always been impartial,
rational, or progressive. Those in authority
might use their positions for their own finan-
cial or political interest. Rather than enabling
development, then, structures of authority have
the capacity to undermine it, not least because
vulnerable people in need of support may be
unable to access it unless they are able to posi-
tion themselves to benefit from favoritism or
to pay a bribe. Furthermore, funds intended
for the benefit of the public are lost to private
accumulation.

These growing doubts resulted in greater
amounts of development work being done
through NGOs and private corporations, and a
desire to diminish state intervention in order to
allow markets to meet needs and grow. How-
ever, by the 1990s, sentiment was shifting away
from “sterile debates about ‘how much’ states
intervene” toward an interest in how states do
intervene and the impacts of these interventions
(Evans 1995, 10). Evans argued that newly
industrializing countries were not explicable by
the free market alone but by states that were
sufficiently autonomous to act independently of
rent-seeking interests, yet sufficiently embed-
ded to facilitate development. The varieties
of capitalism literature explored the way in
which institutions have shaped the behavior of
firms and the long-term trajectory of economic
growth (Hall and Soskice 2001). Meanwhile,
new institutional economics has highlighted the
role that the state can play in reducing transaction
costs and encouraging investment by protecting
property.

From the 1990s, mainstream development
targeted the reform of bureaucracy under the
rubric of good governance to ensure that it was
enabling development and to reduce its coun-
terproductive tendencies. An entity is taken to
be well governed if it demonstrates good control
over its budgets and executes its functions in
an impartial, fair, transparent, procedural, and
accountable way. Performance management
systems have been widely introduced in order to
attempt to govern the effectiveness of officials.
Many definitions of good governance incorpo-
rate the need for democracy to ensure that vested
interests do not parasitically entrench themselves.

Another important outcome of the critique
of top-down bureaucratic governance has been
pressure to expose governance decisions
and systems to deliberation and participation. Experiments in institutionalized participatory democracy in Kerala (India) and Porto Alegre (Brazil) show that ordinary people can be incorporated into the operational aspects of governance. Radical democrats draw attention to the governing role of popular pressure through which ordinary people can assert their wishes through resistance, protests, and lobbying. Meanwhile, autonomists, anarchists, and those who subscribe to postdevelopment thinking, argue that greater social justice is fostered when people govern themselves without the interference of government, and when they are able to escape the free market through self-sufficiency.

**Governing production and social reproduction**

Marxist analysis argues that there is an important relationship between *production* – the way in which society produces goods and services – and *social reproduction* – or the way in which society invests in the welfare of people over time (Katz 2004). This is so, first, because systems of production generate the revenue that enable people to meet their needs, and, second, because economic systems require investments in people to create a labor force needed in the economy. Where there is a high demand for productive labor, capital and the state have reason to ensure that labor has good education, health-care, and living environments. Fordist social compacts associated with welfare states in the Global North were the quintessential example of improving access to the means of life through market (wages and occupational welfare), state (in-kind services and cash-based welfare), and family (the division of labor within families so that those not earning wages are materially supported). In the Global South, the focus of many development strategies has been to expand the productive base in order to increase the size of the economic “pie,” and therefore the possibility of improving welfare. These strategies have taken divergent forms.

In the decades following World War II, some nations allied themselves with the USSR (e.g., China, North Korea, Vietnam, Cuba) or pursued socialism independently (e.g., India, Ghana, Egypt, Tanzania, Argentina). Governance arrangements were based on state planning, direct state control of production, nationalization of key assets, communalization, price control, subsidies, and social spending. In some cases – given that industrial sectors were not developed – these incorporated experiments in agrarian socialism.

In the context of the Cold War, officials in the West were concerned about what they regarded as the threat of communism, and so promoted a governance strategy based on a virtuous cycle of industrial investment, improving productivity, improving wages, and improving savings. Expansions in the industrial sector would draw in populations who were “underemployed” in the agricultural sector, causing a general process of modernization. As countries developed their comparative advantage, they could engage in trade with others.

A great deal of the industrialization that did take place in the postwar period was inspired not by anticommunist neoclassical approaches but by a desire to reduce dependence on the industrial countries of the North. Structuralists and dependency theorists were concerned that periphery nations that traded with core nations would always be disadvantaged since, they believed, the primary goods they exported would buy them fewer secondary goods over time. Many nations, particularly in Latin America, deployed protectionist governance systems known as import substitution industrialization, through which
high import tariffs discouraged importation and encouraged local production.

This strategy was ultimately limited by the size of the domestic market. This, along with a wider crisis of inflation and debt, laid the ground for a radical shift in the 1980s toward trade liberalization. Economic growth was to be unleashed by enabling producers to supply to the global market. However, the concomitant reduction of protection devastated many industries oriented to domestic markets. Meanwhile East Asian nations were well placed to benefit from trade liberalization since industrialization efforts were export-oriented. The ensuing “East Asian Miracle” was initially touted as a vindication of neoclassical approaches to trade, but subsequent analysis has shown that interventionist states played a crucial role in this success.

Widespread debt crises in the early 1980s resulted in bailouts from the International Monetary Fund (IMF) and the World Bank. These structural adjustment programs carried the condition that states would restructure themselves to become market-friendly, not only by reducing import tariffs but also through other state interventions regarded as price distorting such as price controls and subsidies. They were urged to pursue various forms of deregulation, for example of financial controls. Further, states were enjoined to reduce responsibility for social expenditure and to privatize the provision of some services and public enterprises.

Proponents of liberalized economic growth argue that they have been proved correct by the dynamism of middle-income countries, the appearance of large middle classes in the Global South, and the improvement of workers’ conditions in places such as China. Critics argue that gains have been highly uneven and have very often benefited privileged groups at the expense of others (Smith 2008). Economic growth has, at times, been achieved, but at the same time labor supply increased dramatically as a result of the entry of women into the workforce and the introduction of previously unavailable labor pools from China and the USSR into global production (Harvey 2010). The result has been that large populations who would like to be employed cannot find work, or, given the competition for work, any available jobs are highly exploitative. Some development perspectives have valorized the informal economy, arguing that entrepreneurialism provides the livelihoods that are not provided by the labor market. One mechanism which has received a great deal of attention in this regard has been microfinance, through which loans or savings schemes are intended to support the establishment of small enterprises.

From the 1990s, there was widespread recognition of the need to more actively manage the fallout from liberalization strategies which both failed to absorb all those seeking labor and hobbled the ability of states to provide social support. Poverty alleviation and poverty reduction efforts from the 1990s culminated in the United Nations-led Millennium Development Goals, which set out to achieve a series of targets by 2015. One of the most striking interventions on poverty has been the introduction of small monthly cash transfers, often in the form of grants targeting children and pensioners. In some cases the coverage is extensive: a quarter of Brazil’s population was receiving the Bolsa Família grant within a decade of its introduction. This approach has been accommodated within neoliberal thinking because neoclassical economic theory has long acknowledged the possibility of market failure and the need for states to provide safety nets. Cash is easier and cheaper to distribute than “in kind” support such as food, and beneficiaries are often expected to “earn” their grants by meeting certain conditions like keeping their children in school. While
grants are undoubtedly used for consumption by poor households, there is also some expectation that grants will improve the capacity of households to earn their own money by helping them secure jobs or start enterprises. Advocates of market-based growth tend to support grants only on condition that they are not generous enough to remove the incentive to work.

Since the 2000s, some governments have announced a desire to move to a post-neoliberal governance regime. A radical few, notably Bolivia and Venezuela, are pursuing what they call “socialism in the twenty-first century.” Many others have not abandoned their liberalized economic orientation but are combining market growth strategies with increasingly comprehensive social protection. Rather than containing state support within deliberately minimal safety nets, the United Nations and International Labour Organization have proposed the introduction of a global social protection floor which will progressively increase the extent (coverage) and depth (value) of social protection. This protection would include services and supplementary incomes funded from taxes, as well as labor market regulation such as minimum wages.

SEE ALSO: Development; Developmentalism; Finance and development; Governmentality; Industrialization; Modernization theory; Neoliberalism; Newly industrializing economies (NIEs); Postdevelopment; Power and development; State-owned enterprise; States and development; Trade, FDI, and industrial development

References


**Further reading**


Governmentality

Matthew G. Hannah

Universität Bayreuth, Germany

Definition

“Governmentality” is a term brought into circulation by the French philosopher Michel Foucault in 1978 to refer to a modern political problematic centered on the question of how to govern socioeconomic life as “economically” as possible. This entry first offers a detailed definition of the term, then turns to the context in which Foucault coined it and subsequent inflections in its use, and, finally, surveys geographical engagements with governmentality.

To govern “economically” means, roughly speaking, to impinge to a minimal extent on the free unfolding of all the social, economic, and demographic processes of social life, or on the individual choices and decisions that make up its fine grain. The techniques and programs of governmentality are thus often aimed, most clearly in wealthy liberal societies, at indirectly guiding, framing, or eliciting human activities, often on the basis of expert knowledge about social life. Direct regulation or coercion are to be avoided as much as possible, in this way of thinking. Some illustrative examples of governmental techniques in liberal contexts are the use of demographic data in infrastructure planning, training, and incentive systems to promote safe driving, or health insurance schemes.

It is important to note up front, however, that, although Western liberalism and neoliberalism afford the clearest historical instances of this economic logic, they do not in any way exhaust its possibilities: governing economically is not merely another term for liberalism. Governmentality can also embrace measures serving more clearly coercive or unjust aims, such as control of colonized populations or the perpetuation of patriarchal relations between men and women in domestic life. These last examples make clear that instances of governmentality must always be understood within their specific political and geohistorical contexts. For this reason many authors speak, in the plural, of “governmentalities.”

Even with reference to privileged sectors of liberal societies, governmentality should not be understood merely as the name for a set of “neutral” political tools. It is a critical concept. However neutral, harmless, or even beneficial they might seem, techniques of governmentality nevertheless always involve power relations, that is, they are potentially contestable projects of ordering. Governmental measures may be employed by state institutions but also by other actors in a range of social settings including families, workplaces, and so on. Across these different settings, governmental techniques operate alongside and in connection with other forms of power, most notably sovereignty and disciplinary power. In turn, these variably combined techniques often make up the inconspicuous substratum of practices that perpetuate patriarchy, racism, economic exploitation, and other forms of injustice. They may also be brought into play as “countergovernmentalities” to resist or subvert such projects.

This points to another important sense in which governmentality is best thought of as a “problematic.” Governmental programs and techniques are products and relays of ongoing
governmentality

struggles, and also often end up becoming problems themselves. This may be because they fail outright, because they create unanticipated difficulties that discredit them, because they don’t succeed in blocking competing approaches to a problem, or because the successes they do have are only short-lived (Rose and Miller 1992). In this sense, governmentality is a name for the shifting problematic of government through which the ongoing series of specific reforms in a given context are in turn problematized and subjected to further reform, both from within and from outside the state.

There are two related but distinct aims implicit in many governmental rationalities and practices: self-government and governability. Promoting self-government, whether of individuals, groups, organizations, regions, or local state entities, represents the ultimate in “economy of government.” The more actors freely and spontaneously behave in ways that promote social order and prosperity (however these are defined), the less effort need be expended by public or private regulatory authorities. Reforms or practices that promote self-government are often designated in the governmentality literature with the term “responsibilization.”

But, of course, promoting self-government is rarely if ever sufficient to maintain social order. Thus a second, supplementary aim of governmental techniques is to instill at least a minimal level of governability. If people and organizations cannot be counted on to live ordered and productive lives when left entirely undisturbed, a second-best outcome, from the perspective of governing authorities, is that the governed will still respond to incentives and disincentives (whether moral, financial, or penal). Thus education and training in the broadest sense can be seen as part of the sphere of governmentality, because they instill not only specific knowledge but also habits of responsiveness to systems of rewards, incentives, and punishments. This dual cultivation of self-government and governability is referred to in the governmentality literature, following Foucault’s formulation, as “the conduct of conduct” and lies at the heart of Mitchell Dean’s (2013, 215) very useful definition of governmentality: “Governmentality is . . . not simply orderability itself but the forms of making human beings and things orderable, [a project] . . . which always allows [for] multiplicity, polyvalence and resistance, for problematization and counter-problematization, conduct and counter-conduct.”

The ideals of self-government and governability also highlight a final, crucial characteristic of governmental measures: their effect is not merely to influence the behavior of fully formed “subjects” who exist outside of or previous to relations of government, but rather to shape the very formation and re-formation of subjectivities. In their dual orientation toward constructing subjects and thereby establishing or stabilizing larger social orders, techniques of governmentality are sometimes described as both “individualizing” and “totalizing.”

To review, then, the term “governmentality” designates a modern problematic, and associated governing programs and practices, that:

- aim at an economy of government, characterized by a minimum of direct intrusion into individual or collective activities and processes;
- are always context-specific;
- are not merely neutral tools but are always political, that is, understandable as contestable projects of ordering;
- are articulated with other forms and techniques of power (sovereignty, discipline);
- do not necessarily originate with the state, though they may come to invest and shape state activities, and in fact have increasingly done so over the past two centuries;
are often unsuccessful or only temporarily or locally effective, provoking further governmental reforms;

- are typically aimed at promoting self-government, governability, or both;

- construct and reconstruct subjectivities, rather than simply influencing the actions of fully formed subjects.

A genealogy of governmentality

The general definition given in the first section can serve as a basic sketch of what is meant by “governmentality.” However, to grasp the concept of governmentality adequately, it is also necessary to place it, as Foucault himself would have done, within its contexts of emergence and subsequent adoptions and transformations. Foucault’s major writings on governmentality were not books intended for publication, but rather transcripts of provisional lecture courses from 1977–1978 and 1978–1979. These quite accessible courses trace a broad arc through a number of key precursors and sources of modern economic government (Foucault 2007a), and then delve more deeply into the varieties of neoliberalism that have underpinned many governmental programs developed since the late 1970s (Foucault 2007b). One of his central historical claims is that governmentality has not always been the province of states, but rather has a wide array of institutional and discursive origins that long predate the “governmentalization of the state” which has taken place over the past two centuries. His second major historical assertion is that techniques of governmentality have to some extent come to assume a pre-eminent importance as against sovereignty or disciplinary power.

When Foucault first introduced the concept in a lecture in early 1978, it was a product of his thinking about two intersecting contexts: first, his ongoing academic project of understanding the specific forms taken by power relations in the modern era, and, second, his engagements with domestic political struggles and shifts taking place in a number of Western European countries in the narrower time frame of the late 1970s. Broadly speaking, the first context helps explain why Foucault’s initial definitions of the term distinguish it sharply from definitions of sovereignty and disciplinary power. By contrast, the more immediate geohistorical context of European domestic politics in the late 1970s helps explain why governmentality is so closely associated in his empirical explorations with the emergence of neoliberalism. Both contexts, but particularly the latter, led him to emphasize the independence of governmentality from the category of “the state.”

The 1977–1978 lecture course, and the one given the following year, were originally conceived by Foucault as comprising a detailed genealogy of “biopower,” an umbrella term for the modern government of social life. In his 1975–1976 course, and then in the first volume of his History of Sexuality, Foucault had broadly defined biopower as a power aimed not simply at dominating social life but at fostering and cultivating it. This set it off, in Foucault’s reading, against the older and more familiar political logic of sovereignty.

At the core of Foucault’s account of biopolitics and then also of governmentality is the conviction that the eighteenth-century “discovery” of the concept of the population as an entity with its own economic and demographic processes and self-regulating mechanisms marked a decisive watershed in the history of modern power relations. This was because the government of populations shifted the focus of state power from death or the threat of death to social life, and the ends of state power from self-maintenance to the
GOVERNMENTALITY

cultivation of an entity outside the state. Second, the fact that populations and economies were discovered to have their own animating laws and processes meant that the *means* of government could not remain centered, as they had been up to that point, on coercion.

Foucault’s introduction of the term “governmentality” signaled, on the one hand, a shift in his ongoing genealogy from a focus on the *ends* of biopolitics toward a focus on the appropriate *means* of achieving those biopolitical ends, and, concretely, a move away from analyses of regulation aimed at collectivities toward analyses of the construction and management of individual freedom. The resulting “detour,” as he called it, transformed the bulk of his 1977–1978 course, and the entirety of the subsequent course, into a genealogy of governmental techniques and rationalities stretching far back before the modern period. The topics covered, starting with the watershed fourth lecture range from early Christian pastoral techniques, through early modern critiques of Machiavellian discourses, the rise and decline of what was called in the eighteenth century “police,” through to German *Ordo-liberalism* of the mid-twentieth century and the self-help technologies of late twentieth-century American neoliberalism. The 1978–1979 lecture course, in particular, was largely centered on the exploration of the emergence of the neoliberal programs of government just then on the verge of gaining major influence in the United States, the United Kingdom, and other wealthy countries.

There is a good case to be made that Foucault’s shift of focus in early 1978 was also quite specifically linked to political events in France, West Germany, and other European democracies during the late 1970s. In Italy and in West Germany, violent domestic extremists were able to provoke quite repressive responses from state authorities, prompting many engaged left-wing intellectuals in these countries, in France and elsewhere, to reflect on the nature of “the security state.” A widely held thesis (shared by Deleuze and Guattari, for example) was that such repression exposed the fundamentally fascist character of so-called liberal democracies. The problem of fascism, as understood at the time, was that of explaining how populations could be led freely to choose their own subordination to patently unjust or even violent systems of government. For a number of reasons, Foucault found the notion of fascism deeply flawed and far too simple. The problematic of governmentality can be seen in part as his much more complex and historically sensitive response to the problem many of his contemporaries sought to explain with the blunt term “fascism.”

The academic discourses on governmentality that have emerged in the years since Foucault’s death span many different disciplines and an extremely wide array of concrete empirics. Probably the most interesting general development has been a searching reconsideration of the supposed pre-eminence of governmentality vis-à-vis sovereignty in modern political assemblages. This reconsideration was prompted in part by the seeming reassertion of sovereign power in the wake of September 11, 2001, and in part by the writings of the Italian philosopher Giorgio Agamben, who in an intriguing series of historical studies has sought to demonstrate primordial links between sovereign rule and the complex of biopolitics and governmentality that Foucault had long sought to distinguish from older forms of power.

Accompanying this shift has been a stronger focus on exceptional apparatuses and techniques of “security” capable of dealing with extraordinary risks, as well as the kind of everyday problems typical of liberal governance in the Global North. At the same time the notion of governmentality as economical has given way to
a more diffuse construction of governmentalities as variable “economies of government,” that is, intelligible orderings of techniques, positions, discourses, and practices.

Geographies of governmentality

Geographical studies have thus far sought to uncover the ways in which modern geographical imaginations and constructions of territory and place have played a role in programs and practices of governmentality, and how governmental measures have in turn had implications for constructions of spaces and places. Geographical research along these lines has proliferated so quickly in part because so many programs of societal reform aimed at economic government of populations and individuals have had an explicitly territorial or place-related component.

This might seem counterintuitive if one were to base one’s expectations on Foucault’s discussion of territory in his watershed governmentality lecture of 1978. There he argues that, with the historical shift to a concern for cultivating the life of populations, modern government to some extent turned away from the more clearly territorial projects he associates with traditional forms of sovereign rule, projects such as state control over territory and borders. However, as Stuart Elden argued in an important paper, territorial and geographical issues in fact necessarily remain central to programs of governmentality (Elden 2007). The dispositions of people, activities, and resources through which populations are governed are often in part pursued through explicitly spatial ordering projects.

A good deal of the early work done by geographers under the heading of governmentality can be seen as a corpus of fascinating elaborations on the link between spatial (re)orderings and the wellbeing or productive activity of populations. Only a highly incomplete sampling of themes, mostly taken from the anglophone literature, can be mentioned here. The work of Margo Huxley on reformist urban visions, of Reuben Rose-Redwood on the complex emergence of street address systems and city business directories, of Peter Merriman on the government of driving, and of Stephen Legg on urban design and reform in colonial India are all excellent illustrations of how to operationalize a governmentality perspective.

In general, the anglophone geographical engagement with governmentality has moved from the initial focus on the construction of “calculable territory” (Hannah 2009) animating the aforementioned works to embrace an ever wider set of additional issues having to do with place-based and corporeal enactments of subjectivity, as well as with the renewed issue of coercive sovereign power raised by Agamben and others. Stephen Legg provides a useful way of keeping an analytical handle on this proliferation of themes, distinguishing different types or manifestations of governmentality according to the dimensions of episteme, identities, visibility, techne, and ethos (Legg 2007, 12).

One important new direction in geographical research concerns ways of spatially framing human behavior that do not revolve around attempts to construct governability by relying on rationality but instead by seeking to govern, and to govern through, irrationality and affect. Indicative here is, for example, the research into “soft paternalism” or “nudge” by Mark Whitehead, Rhys Jones, and Jessica Pykett (Jones, Pykett, and Whitehead 2010). Here we see the basic problematic of governmentality, namely, how to promote efficient governability, addressed through new methods taking as their point of departure new evidence of the nonrational sources of human behavior.
Like governmental rationalities, academic geographical engagements with governmentality have varied across linguistic and national academic contexts. The reception and development of geographical aspects of Foucault’s analyses in France, for example, were for a long time much more attenuated (see Fall 2007). In Germany, as Füller and Michel (2012) explain, critical appropriations of governmentality have tended to focus strongly on discursive at the expense of material or nondiscursive constructions of spatial ordering projects. The German discourse is arguably also less strongly oriented around historical geographic studies than in the United Kingdom and North America, instead paying closer attention to current urban and immigration politics in a variety of settings.

Finally, it is worth singling out recent work on subaltern countergovernmentalities in the Global South by Ananya Roy, Arjun Appadurai, and others. This area is fascinating because it highlights bottom-up dynamics of activism for inclusion in and alteration of governmental systems. Focusing in many cases on South Asia, but also Africa and other parts of the Global South, this new strand of research explores the ways in which community groups have mobilized for projects such as self-enumeration and inventories of local infrastructure, particularly in the informal settlements surrounding megacities. Such studies open up interesting affinities with the long tradition of anarchist projects of local self-organization. Among other benefits, a countergovernmentality perspective could move scholarship beyond the lingering tendency to assume that governmental ordering projects (whether or not they originate with states) are essentially top-down affairs.

**SEE ALSO:** Biopolitics; Neoliberalism; Political geography; Power; Security; Sovereignty; Space of exception; State, the; Territory and territoriality

**References**


**Further reading**

GPU computing

Wenwu Tang
University of North Carolina at Charlotte, USA

Graphics processing units (GPUs) are built using a many-core architecture to provide acceleration for general-purpose computation (see Owens et al. 2008; Kirk and Hwu 2010). The use of GPUs for accelerating computing can be traced back to the 1980s (Kirk and Hwu 2010). GPUs were originally designed and used exclusively for graphics operations (i.e., as a video accelerator). As modern computing evolved from single to multi and many cores, the use of programmable GPUs for general-purpose scientific computation has been extensively recognized. The development of GPU hardware in terms of its floating-point performance remains compliant with Moore’s law, that computing performance doubles about every 18 months (see Nickolls et al. 2008), even though this law has been claimed to be inapplicable to CPUs. For example, the latest NVIDIA Tesla K40 GPU, based on the Kepler architecture (see NVIDIA 2013) can deliver a peak double-precision floating point performance of 1.43 tera flops (floating-point operations per second), much higher than a current era CPU. GPUs are widely available on alternative computing devices, including smartphones, desktop computers, servers, and supercomputers. This widespread availability has stimulated the application of modern GPU technology for accelerating general-purpose computing in different scientific domains.

GPU computing falls within the paradigm of stream processing (Buck et al. 2004). A GPU is composed of a group of streaming multiprocessors (SM) each of which includes many (say, 32) streaming processors (SP). As an example, the latest NVIDIA Tesla K40 GPU has 2,880 computing cores (90 SM × 32 SP per SM). To efficiently employ this many-core stream-processing hardware, parallel programming standards and platforms have been developed and released, including CUDA (Compute Unique Device Architecture; see NVIDIA n.d.), OpenCL (Khronos Group 2013), and OpenACC (OpenACC n.d.). These standards and platforms allow users to harness the many-core supercomputing power in GPUs using an SIMD (single instruction multiple data; see Wilkinson and Allen 2004) parallel approach. GPUs provide support for massively parallel computing because of the large number of cores available (see Armstrong and Marciano 1995). Fundamentally, GPU computing uses fine-grained thread and shared-memory parallelism for data-intensive parallel computing (Kirk and Hwu 2010). Data of interest are split into a number of smaller sub-datasets that can be processed concurrently by the massive number of available threads.

The use of modern GPUs for general-purpose computing took off in 2006 when NVIDIA released CUDA to provide parallel computing support for GPU-accelerated general-purpose computation. In CUDA, a significant number (millions) of lightweight threads are launched to execute the same program for processing partitioned sub-datasets. These threads are organized into blocks and further grids to harness the many-core computing power in GPUs. A grid of threads corresponds to a kernel function, which encapsulates the algorithm to be executed by each thread in the grid (Kirk and Hwu 2010).
GPU COMPUTING

2010). The set of thread blocks is scheduled (automatically handled by CUDA) to streaming multiprocessors in the GPU. Thread blocks are executed independently to ensure the high throughput performance of GPUs. In CUDA, data can be maintained in GPU memories at different levels: global, shared, and register. The entire grid of threads can access data stored in the global memory, but with a relatively slow access speed. Each block of threads has fast access to shared memory, and each thread has its own (fast) register memory. Data should be maintained in the shared or register memory as much as possible because of the overhead latencies associated with access to the global memory. Thus, the combined use of thread and memory hierarchies enables the exploitation of supercomputing power in GPUs.

GPU coprocessors can be used to accelerate computation associated with spatial algorithms. Using GPUs to accelerate spatial algorithms normally requires support from a conventional CPU-based system which serves as a host. Generally, data associated with spatial algorithms must be read and maintained in memory within a CPU environment. Then these data are transferred and reconstructed in the GPU memory. Through the invocation of kernel functions, threads handle the computation associated with these data. Once the parallel computation of all threads is complete, the results are aggregated and transferred back to the CPU.

The acceleration performance of a spatial algorithm is mainly affected by the following aspects: its internal characteristics (e.g., data structures, procedures), characteristics of the data, parallel strategies (e.g., domain decomposition or synchronization; see Wilkinson and Allen 2004), CPU/GPU hardware features, and data transfer between CPU and GPU. The quantitative evaluation of GPU acceleration performance relies mainly on the use of the acceleration factor, a performance metric that is a ratio of the computing time of a CPU-based sequential algorithm in relation to a GPU-based parallel algorithm. The acceleration factor is similar to the speedup factor for CPU-based parallel algorithms (see Wilkinson and Allen 2004). Usually, a fast, modern CPU is chosen to support the derivation of acceleration factors. For instance, an acceleration factor of 100 means that the performance of the GPU of interest is equivalent to the simultaneous use of 100 CPUs. In practice, the acceleration performance of GPUs for spatial algorithms can be several orders of magnitude higher than that of a single CPU (Kirk and Hwu 2010; Nickolls et al. 2008). This level of performance acceleration often enables a single GPU to outperform a small CPU cluster, and a small multi-GPU cluster has a performance equivalent to a large CPU-based supercomputer cluster. Further, due to the cost-effective characteristics of GPUs, many supercomputers employ GPU acceleration (see www.top500.org).

Since the emergence of general-purpose GPU technologies in 2006, a series of spatial algorithms has been parallelized using GPUs (Zhang 2011; Zhao, Padmanabhan, and Wang 2012; Tang and Bennett 2011; Shi, Kindratenko, and Yang 2013; Tang 2013). The data-parallel nature of GPU programming makes GPUs most suited to data-intensive spatial algorithms (both vector- and raster-based). Conversely, the concurrent data characteristics of spatial algorithms (e.g., local, neighborhood, and global) make GPUs a suitable parallel computing platform for acceleration. In particular, as researchers are paying increased attention to big data analytics (Manyika et al. 2011), GPUs are now playing a pivotal role in the acceleration of spatial algorithms for big spatial data. The use of GPUs for acceleration, however, requires that data structures and spatial algorithms be transformed for accommodation in GPU environments (Tang 2013; Tang and
This represents a potential bottleneck because, without an appropriate transformation, the computing performance of a GPU-enabled parallel spatial algorithm may be even slower than its CPU-based sequential counterpart. As the development of GPU technologies continues, the barrier related to GPU-accelerated spatial algorithms may be lowered. In sum, GPUs, as an evolving technology in computational science and geocomputation (Armstrong 2000), hold enormous potential for accelerating computationally intensive spatial analysis and modeling.

SEE ALSO: CyberGIS; Geocomputation; Parallel computing; Spatial analysis

References


Graph theory

César Ducruet
Centre National de la Recherche Scientifique (CNRS), France

The origins of graph theory are often attributed to the works of Leonhard Euler. In a classic study of the city of Koenigsberg in 1735, he designed a problem where a set of nodes (or vertices) connected by links (or edges) had to be traveled without using by the same road or bridge twice. Such an exercise proved to be mathematically impossible and motivated further efforts to define the topological properties of graphs during the subsequent decades and centuries. Today still, the application of graph theory to geography remains largely focused on technical networks, using traditional mathematical tools (Arlinghaus, Arlinghaus, and Harary 2002). This entry describes graph theory in general as well as its use by geographers, notwithstanding parallel evolutions of the field in other social and natural sciences, and their mutual influences.

Network types and structures

A wide and specific vocabulary describes various sorts of graphs having distinct internal structures, as well as subgraphs (subset of connected nodes) and individual nodes by the architecture of their mutual links. Graphs may be visualized either through matrices or by the node/link view, and many graph visualization software programs exist nowadays that are capable of applying numerous types of visualizations and measures, even on very large networks composed of thousands of links and nodes. Visualizing large graphs is often an exercise with limited outputs, due to poor readability, so that many different operations of network simplification should be applied prior to visualization, such as filtering links or nodes according to certain quantitative thresholds or qualities. Furthermore, visualizing a network on a map, such as in geography (although this has become so popular nowadays that any spatially embedded dataset made of nodes and links can be instantaneously projected on the screen regardless of its size, etc.), brings in additional issues of readability. The matrix itself (or adjacency matrix) is represented by a table where nodes are rows and columns, and links are, in their simplest understanding, represented in a binary fashion by the number 0 or 1 where lines and columns intersect. There are specific mathematical calculations based on matrices that are not detailed in this review. Visualizing graphs as sets of nodes connected by links proved to be more useful to illustrate the existence of various network morphologies, or fundamental components of networks, such as cycles, cliques, chains, motifs, and also various situations characterizing individual nodes, such as isthmus, bridge, ego network, star, and so on. Graph theory is based on a specific vocabulary to describe networks. Nodes and links are labeled vertices (vertex) and edges, respectively. Vertices may be junctions or end points in the graph, with edges between them. Connectedness refers to the property of a graph to have all its vertices interconnected, which otherwise contains isolates (i.e., vertices without edges) or at least two connected components. A complete graph is defined by the maximum number of edges.
among vertices. The graph is directed (digraph) when two edges exist between two same vertices, referring to two directions (AB vs BA). When both directions exist the graph is said to be symmetrical, otherwise it is asymmetrical. Other dimensions are also important, such as when vertices and edges have certain values, in a weighted graph (e.g., passenger flow, number of vehicles), when vertices are connected by only one type of edge (simple graph) or at least two types (multigraph, multiplex graph), such as in a social network, friendship, and professional ties. Finally, a bipartite graph exists when two or more types of vertices are connected, such as conferences and speakers, journals and scholars, and so on with the condition that vertices of the same nature are not connected with each other.

Numerous internal structures have also been identified within graphs at different levels of their organization. For instance, chains and cycles are both continuous alignment of edges, while the cycle starts and ends at the same vertex. A path between two nodes is a sequence of edges travelled in the same direction. Several paths may exist between the same two nodes, thus motivating the search for the shortest path and the minimum cost path, the latter being only possible in the case of a weighted graph where edges are valued by a distance, time, or other metric (e.g., days, hours, kilometers, monetary value). For the shortest path, topological distance refers to the number of edges to be crossed in order to reach the final destination of the path. Topological distance is often referred to as path (or edge) length but may also be depth. The clique is one fundamental structure of graph theory, defined as a maximal complete subgraph, that is, a subset of nodes being fully connected with each other, the smallest example being the triangle of “3-clique.” In social network analysis (SNA), the clique is an essential component of social organization and is often opposed to the “star” configuration, where several vertices are connected to only one vertex. The ego network comprises the connected neighbors of a given vertex as well as the edges among these neighbors. More complex internal structures have been the focus of several studies on graphs, such as the so-called “clusters” or “communities,” whose definition may vary according to several criteria. Overall, a cluster or community is defined either by a higher density of connections or by the similarity of their connections. In other words, tightly connected vertices may form a community with a high probability to be relatively close to each other in the graph. Vertices having similar connectivity to others, conversely, may be close to or far away from each other. Vertices and edges lying in between two or more communities are often labeled isthmuses or bridges, due to their high importance. In social network analysis, it is also called a weak tie or structural hole in the case of edges in-between different groups. Finally, a graph is said to be assortative when the number of edges of a given vertex is positively and significantly correlated with the average number of edges of its direct neighbors in the graph; otherwise it is said to be disassortative.

Based on the aforementioned characteristics, graphs may be labeled differently in terms of their overall structural properties. The regular graph (or k-regular) possesses exactly the same number of edges per vertex. The random graph exhibits a normal or Gaussian statistical distribution of the number of edges per vertex. Scale-free and small-world graphs are defined below when it comes to the more recent field of complex networks. The tree graph refers to a situation where no clique, cluster, or cycle can be found, as in the case of river networks for instance, which function as dendrograms. More conventional measures help reveal the overall structure
of the graph and define its main character. For instance, the diameter is the length of the longest of the shortest paths among all pairs of vertices in the graph, measured by the number of edges needed to be travelled. The average shortest path length is similar but refers to the average number of edges composing the shortest paths across the graph, while high values often imply more difficulty to circulate in the network. Planar graphs often have high values for both diameter and average shortest path length. The average clustering coefficient (or transitivity) measures on average the proportion of closed triangles in the maximum possible number of triangles in the graph. In that sense it is close to another classic measure of graph theory, the alpha index, which looks at the proportion of existing cycles in the maximum possible number of cycles, with the difference that a cycle can have a length of three edges (triangle, 3-clique) but also of four, five, and so on. High values of the average clustering coefficient mean that the graph is tightly connected, while tree graphs will always have zero.

Networks in geography and natural sciences

One of the first imports from mathematics into geography was made by Kansky (1963), incorporating previous studies on telecommunication flows and highway networks among US cities. Such works notably reviewed the various measures and indices applicable to mostly planar graphs, that is, graphs where two edges cannot cross without creating a vertex, as it is often the case for road, rail, and river networks; the main goal being to describe the morphology, connectivity, and density of various graphs (Haggett and Chorley 1969). Without reviewing them entirely, most common indices include the diameter (number of links between the two most distant nodes), the alpha (share of observed cycles in the total possible number of cycles – the cycle being a closed path of various length in terms of its number of links), beta (number of links per node, or density), and gamma (share of links in the total possible number of links, or completeness) indices. Three limitations of such general indices should be addressed, however. First, one major drawback is that two networks of identical size (number of nodes and links) but with drastically different topology (e.g., star-shaped network where one node connects all others versus chain-like network formed by an alignment of links) will be characterized by the same results. Second, and related to the first, general indices are not robust to network size, so that their level will vary according to the size of the network. Third, graph theory and its application to geography long remained binary (presence or absence of a link between two nodes), thereby completely ignoring the possible weight of links (e.g., traffic intensity or capacity, kilometric length) and their quality (e.g., links and nodes of different types).

Other measures include those at the local level, such as calculating, for each node, its distance from all other nodes by the number of links needed to connect the most distant node (Koenig number) or the total length of shortest paths connecting all other nodes (Shimbel accessibility index) (Figure 1). Such studies focusing on small, planar networks thus largely neglected the analysis of nonplanar networks (i.e., allowing edge crossings without creating nodes). Due to limited interactions with social network analysis (SNA) that developed from the 1920s onwards, no imports have been realized from sociology, although the latter had developed many tools for analyzing nonplanar networks. For instance, betweenness centrality (number of times a node lies on a shortest path in the graph) as well as transitivity (or clustering coefficient) came from
SNA, and later on were integrated in studies of complex networks. They consider that for each node, there is a share of observed triangles – or cliques – in the maximum possible number of triangles among its neighbors. Transitivity had no use for studying planar networks, as rectangles rather than triangles are often found in road or railway networks. It is only in the early 1990s that geographers started to analyze networks of a different nature, or nonplanar networks, such as those formed by multinational corporations, airlines, and shipping lines, thereby stepping out of classic graph theory (Ducruet and Beauguite 2014), and from a more urban perspective. Other important refinements of graph theory have been to consider the weight and spatial dimension of networks. The cited measures have been adapted to incorporate the weight of links and nodes so as to better reflect inequality in, for instance, traffic distribution, compared with the sole topological (binary) aspects. Thus, all measures can be weighted according to various metrics at global (whole network) and local (individual nodes) levels. When it comes to the spatial dimension of networks, physicists were keen to demonstrate the influence of spatial constraints on network growth and topology, whereas such aspects were often taken for granted by geographers, since all their studies focused on spatially embedded networks in contrast with studies of nonspatial networks such as social networks, collaborative and citations networks, the World Wide Web, and so on. Geographers focused more on detailed analyses of binary graphs in abstract spaces and of real networks using geographical information systems but often

Figure 1 Common measures of node centrality in complex networks.
without reference to graph theory. Physicists studying spatial networks notably underlined that preferential attachment occurs over shorter distances, as there is a cost of creating new links to overcome the friction of space, thus resulting in higher average clustering coefficients than in nonspatial networks (Barthelemy 2011).

Efforts made by geographers to improve their use of graph-theoretical methods occurred in relative isolation, though, from the rapid evolution that graph theory went through in the late 1990s and early 2000s. Natural scientists, such as physicists, have been eager to rejuvenate graph theory, and notably the aforementioned limitations as well as its approach to graph dynamics. They notably proposed to define a so-called “complex network” by nontrivial topological attributes, which are neither random nor regular, but are found in most real-world networks. Another dimension of complex networks is their large number of nodes and links, and their analysis clearly benefited from the growing availability of large datasets and computation power. New measures were proposed, although they were, in fact, a rediscovery of existing ones, such as transitivity (clustering coefficient) originating from SNA and the gamma index (density) proposed by Kansky (1963). Still, the merit of natural scientists has been to provide clearer definitions of network models and their evolutionary mechanisms, with new applications on real-world networks that were traditionally analyzed by social sciences in more limited ways. One of their major achievements has been to complement the random graph model, which was dominant until then (a random graph is characterized by a normal or Gaussian distribution of the nodes’ degree), by two new models: scale-free and small-world.

The small-world network is defined by a higher average clustering coefficient and smaller diameter than the random graph of equal size (Watts and Strogatz 1998). It means that small-world networks are more likely to contain cliques (subgroups of at least three connected neighbors forming a closed path) and to provide more efficient ways to circulate in the network, in terms of the length needed to connect its distant parts. Scale-free networks are similar to small-world networks but their distribution of degrees follows a power-law (the slope exponent ranging between two and three), which means that very few nodes have plentiful connections, and a vast majority of nodes have very few connections. Another dimension of scale-free networks is the process of preferential attachment by which newly added nodes primarily connect larger nodes, thereby reinforcing the hierarchical structure of the network (Barabasi and Albert 1999). Despite the spread of such models across almost all scientific disciplines, geography has remained somewhat reluctant in adopting and testing these innovations. In fact, geographers carrying out network analysis have continuously relied on the same indices and measures over the last five decades. Just like the innovative character of complex network research is questioned, the benefits of applying such methods to test geographic questions also remain unclear. Should this be indebted to the models or to the way they were applied? One part of the answer lies in the fact that geography as a whole went through a “behavioral” turn after the so-called quantitative revolution of the 1960s, graph theory and network analysis – and spatial analysis as a whole – remaining “somewhat of a backwater” since then (Waters 2006). Networks continued to be analyzed by regional scientists but not necessarily using graph theory, so in even more abstract spaces while focusing on costs, and engineers focusing on performance and optimization. Yet, a small number of geographers have promoted the rediscovery of graph theory through...
the adoption of certain methods and models rejuvenated by complex networks research.

Key directions of current and future research on networks

One primary and key issue raised by geographers has been the relationship between node centrality and socioeconomic features. As such, various centralities (i.e., degree, betweenness, and closeness) were statistically confronted with local urban/regional data (e.g., population or gross domestic product) for various kinds of network, such as Internet, airline, telecommunications, and scientific collaborations among cities (see Ducruet and Beauguitté 2014 for a synthesis).

The second dimension explored by geographers is the influence of distance on network topology as mentioned above, although this influence is often taken for granted by geographers. One original approach has been to apply space syntax to urban streets so as to consider the network from a more cognitive point of view. Space syntax notably proposed to consider crossroads as links between streets, the latter being defined by their name or number instead of as alignments of different segments (Hillier and Hanson 1984). Such an approach made it possible to have a completely different approach to centralities and topologies in a planar network, and to extract meaningful information on the small-world and scale-free dimension of many cities (i.e., dual graph). In the same vein, and as a third aspect of renewed graph theory applied in geography, several geographers have proposed a combination of methods rather than a strict, direct application of complex networks. It is interesting to mention that, for instance, spatial interaction models (SIM) such as the gravity model (Schintler et al. 2007), geographic information systems (GIS), multiplicative growth models, and social network analysis (SNA) tools such as structural equivalence and blockmodeling were applied, sometimes even jointly, to various networks, such as Internet backbone networks, commuting flows, road and railway networks, airline and shipping networks. This is mostly the case for regional science, which had already been using such methods for decades. It allowed, for instance, a better appreciation of the cost factor in network organization.

A fourth dimension is the existence of subgroups of densely connected nodes in the network, often referred to as clusters or communities of nodes. Geographers were keen to detect and define such subgroups with reference to the concept of region, sometimes also mentioning the multilevel organization of networks, whereby the different subgroups are linked to each other in various ways. Methods of partitioning (or clustering) were thus applied to a number of networks – such as the worldwide airline network of intercity passenger flows, container shipping networks, and mobile phone communication networks to verify the influence of geographic proximity in the emergence of clusters – and commuter flows between localities in order either to redefine functional urban regions or to identify secondary poles inside metropolitan areas. Such approaches were already far more advanced than the traditional single (or multiple) linkage analysis proposed in the 1960s by Nystuen and Dacey (1961) and further applied in numerous contexts afterwards. Other approaches also looked at the emergence of clusters within large cities based on the cartography of vehicle trajectories.

Fifth, geographers have attempted to go beyond the analysis of single (or simple) networks by proposing an analysis of coupled or interdependent networks of different nature, labeled multigraphs (or multiplex graphs) in the classic graph theory literature. Due to the
absence of a formal mathematical definition of multigraphs and, therefore, of a precise methodology to analyze them, scholars had to create innovative ways to extract meaningful information dedicated to such graphs. Sometimes the network is simply a combination of various networks (or network of networks) in which the nature of the links is ignored, but the analysis compares the effect of adding (or removing) one or more networks at the level of nodes or the entire network. This was applied to the analysis of world regions and global cities based on various types of links (e.g., trade, diplomatic, airlines, and shipping lines). Other examples include the combination of airline and Internet networks, airline and corporate (multinational) networks, and so on. Yet, geographers have poorly investigated to what extent certain places are more specialized than others in terms of combined centrality, how certain networks and nodes are more vulnerable than others according to their physical embedding and distribution across space. In particular, the local and global impacts of random or targeted attacks that are well investigated by physicists (e.g., cascading failures) have not yet received much attention from geographers so far. This is rather surprising given the fact that such approaches would well apply to the analysis of cities and urban systems, which are connected by a variety of networks (e.g., power grid, water, telecommunications, transport, etc.).

Finally, and as a sixth aspect of recent research based on graph theory, geographers have paid particular attention to network dynamics. Very few works have approached directly the dynamic evolution of real-world networks. Existing studies approach dynamics more through an analysis of successive static states of the network or phases, mostly due to a lack of continuity in available time series datasets of any kind. Other approaches to network dynamics are also those applying simulation methods to systems of cities, starting from an initial stage and testing the role of different parameters on the evolution of urban size in a relational perspective (Bretagnolle and Pumain 2010), but often without actual information on intercity linkages.

Graph theory has had an important influence on the way geographers studied spatial phenomena, but this influence was stronger in the early phase of its adoption, and has remained mostly true in the specific field of transportation networks. Natural scientists have been faster in adopting (and even fostering) innovations that appeared in the late 1990s through what is known as complex network research. While geographers became skeptical about the benefits of graph-theoretical methods applied to spatial issues, physicists and computer scientists had no doubt that their models, concepts, and methods would improve our understanding of the uneven distribution of flows and networks, especially across space. Nevertheless, one major limitation of the complex network research carried out by natural scientists is the absence of discussion about the link between their results and the wider social, economic, institutional, and political context in which networks evolve. Stronger interactions between natural and social sciences, among which is geography, are thus needed in order to better bridge abstract space and concrete space. In the past decades, geographers have been able to explain in detail the strategies of specific actors involved in building and expanding networks of all kinds, but without a firm engagement into renewed methods of network analysis. On the contrary, natural scientists keep on refining their models and tools to analyze graphs, but without questioning the possible link with existing geographical research. Issues of scales, borders, territories, urbanization, innovation, diffusion, globalization, regional planning, environmental, and socioeconomic impacts are still to be explored in relation to what graph theory may
provide to the concerned actors and institutions. Yet, an area well covered by geographers in the last decade, in collaboration with computer scientists, has been the search for better methods to visualize spatial networks and flows (Rozenblat and Melançon 2013). This constitutes, perhaps, the main contribution of geography so far to graph theory and network analysis as a whole.

**SEE ALSO:** Network analysis; Transport networks; Visualization

### References


### Further reading

Grassland ecosystems

Dwight A. Brown
University of Minnesota, USA

Grasses can thrive from tropical to arctic latitudes and in wet to dry environments. The major species have a common trait of a perennial growth habit. They often produce seeds the first year and, effectively use wind, water, and animal transport to disperse their seeds. Some species tolerate shallow burial by wind and water borne soil. As a result they can colonize large areas quickly and persist for centuries. However, the truly vast expanses of waving grasslands of two centuries ago are now substantially reduced. The areas dominated by grass species are naturally limited by competing vegetation types that rob grasses of sunlight, nutrients or soil water. Historically, the greatest cause of grassland loss has been the conversion of immense areas to cropland. High crop prices coupled with relatively low energy prices that enabled groundwater pumping or the river diversion have stimulated grain or fiber cropping even into relative dry environments.

We cannot treat the world grasslands as anything close to a homogenous unit. There are considerable variations in grass species and the fauna that use them. The details of the world grasslands information vary a great deal by continent and country as is evident in a recent FAO publication (Suttie, Reynolds, and Batello 2005). For most countries, only lists of species present are available. Rarely is species abundance information available. In the United States, where grazing of native ranges is still important, county soil publications often list species composition and abundance.

Grasslands occur on every continent except Antarctica, although hair grass (Deschampsia antarctica) is present. Extensive grasslands are confined between 55° north and south latitude (Figure 1). The species composition of the world’s grasslands differs by geographic location, landscape position, and soil traits. Some physiological traits help different species thrive under different environmental constraints. Most large areas of grasslands are dominated by relatively few species. They reflect episodes of dispersal that favored particular species from the nearby gene pool at times when disturbances created opportunities for new plants to establish. The dominant species or grasslands are generally tolerant of defoliation by grazers, drought, and fire. The lack of these tolerances by woody plants retards their invasion. The dominant grass species differ in their resource requirements and degree of tolerance. Genetics governs the ease with which seedlings become established plants. Photosynthetic chemistry differences give species different temperature tolerances and water use efficiency.

Grass photosynthetic chemistry falls into two broad types. On an evolutionary time scale, the oldest form among grasses involves three carbon atoms to convert atmospheric carbon dioxide into plant carbon. These are commonly known as C3 or cool-season grasses. Poa, Festuca, Agropyron, Pascopyrum, Elymus, Phleum, Dicanthelium, Hesperostipa, Bromus, Agrostis, Achnatherum, Puccinellia, and Nassella are major native grass genera of C3-dominated grasslands. The term cool-season grass comes from their early greening, maturity, and maximum
productivity at about 20°C (70°F). At higher temperatures productivity falls sharply. Common grain crops of wheat, oats, rye, and barley are C3 grasses, as are many lawn grasses.

A later evolution in grass physiology was the development of four-carbon photosynthesis in grasses (C4 or warm-season grasses). These grasses peak in productivity at about 35°C (85°F). This evolution probably occurred during distant geologic times when atmospheric carbon dioxide was much lower than the present. The C4 processes near the leaf surface enrich carbon dioxide fed to C3 processes deeper in the leaf structure. They are much more efficient water users and therefore more tolerant of arid conditions than C3 grasses. *Andropogon, Astrebla, Calamovilfa, Sorgastrum, Spartina, Schizachyrium, Panicum, Plectrachne, Triodia, and Bouteloua, Buchloe, and Bothriochloa* are major native grass genera of C4-dominated grasslands. Corn, sugarcane, and sorghum are common domestic grasses with C4 photosynthesis.

C3 grasses tend to dominate higher latitudes, but do occur in lower latitudes where winter rainfall enables them to mature before the high summer heat (Figure 2). They also are common at higher elevations or in high to mid-latitudes. Winter precipitation climates tend to favor C3 grasses because of their greater growth at cooler temperatures and their ability to germinate at lower soil temperatures than C4 grasses. They also grow in tropical grasslands, growing in the cooler winter season.

**Importance of photosynthetic differences**

C4 grasses occur in three primary variants, PCK, NADP, and NADP-ME in order of increasing photosynthetic productivity (Figure 3). Some species are transitional between two photosynthetic types. C4 grasses are more efficient water users than C3 grasses.
In Australia C3 grasses are dominant in the south and NADP-ME species are the most abundant variant of C4 grasses in the northern deserts and savannas (Figure 4). In North America, NAD-ME C4 species are abundant. In the southwestern plains and deserts, NADP-ME species dominate the eastern grasslands and C3 species dominate the northwestern grasslands (Figure 5). The Russian and Mongolian Steppes are dominated with many of the same genera of C3 grasses found in the United States and Canada although few of the species are the same.

Fauna of the grasslands

Grazers are a major element in maintenance of grasslands. The dominant grasses are tolerant of defoliation by grazers. Large hoofed mammals are the best known grazers, but burrowing rodents and insects are also important grazers. All grasslands support predators, carrion eaters, and decomposers to control populations of grazers and clean up the refuse.

Maintenance of grasslands

By far the greatest factor in the demise and continuing threat to grasslands is cultivation. High crop prices naturally result from a growing population’s demand for food, from wars that disrupt food and fiber production in some parts of the world, and from some government policies that stimulate farmers to grow more crops and bring grasslands under cultivation. Where natural rainfall is sufficient or irrigation water is readily available, grasslands have succumbed to the plow. When crop prices fall or irrigation is no longer economically viable, fields are left bare. The resulting buildup of salt and the isolation

Figure 2  World grass species abundance by photosynthetic types. Sources: Dwight Brown, data from Hartley 1957; Hartley 1958a, 1958b, 1973; Hartley and Slater 1960.
of these fields from seed sources of the previous grassland species limit grassland regeneration.

The almost complete annual fallout of plant material in grasslands, coupled with relative low decomposition rates allows a buildup of soil carbon that is great relative to forest environments where the plant carbon is maintained in standing trees and shrubs. As a result, the grassland soils can hold large stores of plant nutrients, which make them ideal for crop agriculture.

In some former grasslands with low rainfall, the chances of reestablishing grasslands to their previous composition are very low because the climatic conditions that enabled the previous grass species to be established no longer prevail. This is especially true for various warm-season species of grama grasses (Bouteloua) which have seedling root structures that require a period of surface soil moisture that has rarely occurred in the historic weather record. Some cool-season grasses are more easily established, but they are not as useful to grazers in late summer.

Erosion by running water, and to a lesser degree by wind, and animal burrowing have been major forces opening land for the dispersal of grass species into new spaces. Particularly with water erosion and subsequent deposition, there is often considerable connectivity of bare land, which provides a convenient corridor for nearby dispersing species that are able to colonize new sites.

The persistence of grasslands through tens of thousands of years is substantially due to their tolerance of defoliation. Most grasses have similar responses to grazing, fire, and drought. Because these defoliation agents are common throughout the world grasslands, the surviving and now dominant species are tolerant of these agents, except for a few species that are intolerant of defoliation early in the growing season. The intolerant species are normally lower in abundance and are not widely distributed.

The future of native grass populations will depend partly on the rising CO₂ levels in the atmosphere. C₃ grasses get a production boost from higher CO₂. C₄ grasses do not enjoy the same benefit because the C₄ process already enriches the CO₂ that is fed to the C₃ cells deeper in the leaf structure.

Grasslands of continents

North American grasslands

The American midcontinent is a converging ground for C₃ cool-season grasses that migrated from the northwest, short, warm-season grasses (C₄, NAD-ME) that spread from Sonora and Chihuahua in Mexico, and tallgrass warm-season species (C₄, NADP-ME) that probably found their way from Africa via the Caribbean (Figure 5). These tall grasses once dominated Indiana, Illinois, Iowa, parts of Minnesota and Missouri, and the eastern parts of Oklahoma, Kansas, Nebraska, and the Dakotas. Most of these grasslands have given way to the plow. Occasional tiny remnants of these native grasses can be found between cultivated fields. The species of these three major areas are listed in order of their abundance.


*Pascopyrum smithii* (Western wheatgrass), *Hesperostipa comata* (needle and thread), *Nassella viridula* (green needlegrass), *Sporobolus cryptandrus* (sand dropseed), *Elymus canadensis* (Canada wildrye), and *Koleriamacrantha* (Junegrass) are the most abundant C3 species of the northwestern plains and the Pacific northwest.

*Bouteloua gracilis* (blue grama), *Bouteloua curtipendula* (sideoats grama), *Buchlöedactyloides* (buffalograss), *Bouteloua eriopoda* (black grama), and *Bouteloua hirsute* (hairy grama) are the most abundant species of the southwestern plains. All are C4 NAD-ME species, except *Bouteloua curtipendula*, which is a PCK/NAD-ME intermediate.

The tallgrass prairies of the eastern plains are dominated by *Schizachyrium scoparium* (little bluestem), *Andropogon gerardii* + *hallii* (big + sand bluestem), *Panicum virgatum* (switchgrass), *Sorghastrum nutans* (Indiangrass), and...
GRASSLAND ECOSYSTEMS

Bothriochloa saccharoides (silver bluestem). All are C4 NADP-ME photosynthetic type.

North American grasslands were home to bison, mammoth (now extinct), deer, and antelope. Ten million years ago, now extinct camel, horse, and rhinos grazed grasslands of the midcontinent. Prairie dogs and ground squirrels also played important roles in disturbance and change in grass cover. Coyote, wolf, bear, and cougar were common predators before the rise of agriculture.

South American grasslands

Major grasslands of South America include the Llanos of Venezuela and the Pampas of east-central Argentina, Uruguay and southern Brazil. The Llanos of Venezuela are dominated by C4 NADP-ME grasses Andropogon spp., Axonopus spp., Elyonurus adustus, Leptocoryphium lanatum, Panicum olyroides, Paspalum spp., and Trachypogon spp. (Figure 2).

The natural grasslands of southern Brazil include some forests and herbaceous vegetation. Grasses dominate, with some legumes. Latitude, altitude and soil conditions enable the dominance of C4 grasses to grow in the warm season, and C3 species to grow in the winter. The area is species rich with more than 800 grass species. Major species on drier soils include Paspalum spp., Andropogon lateralis, Axonopus affinis, and Aristida filifolia (all C4, NADP-ME species).

The Uruguay grasslands contain both C3 and C4 grasses. Trees are rare. The genera with most species are Paspalum, Axonopus, Setaria, Digitaria, Andropogon, Bothriochloa, and Schizachyrium (all C4, NADP-ME species except Panicum, which contains a few C3 species); the genera Eragrostis, Distichlis, and Chloris, Eleusine, and Bouteloua (all C4, NAD–ME species). C3 genera include Calamagrostis, Agrostis, with only a few species present.

The Pampas in Argentina is dominated by C4 grasses at low elevations. Andropogon lateralis is the most common species, with some grasslands dominated by Sorghastrum agrostoides, Paspalum quadrifarium, and P. intermedium. All are C4, NADP-ME species.

Pampas deer, armadillo, guanaco, and rhea are the primary native grazers of the Pampas and capybara in the Llanos.

The steppes of Russia and Mongolia

The vast grasslands Inner Asia and Central Asia lie between taiga forests to the north and desert
to the south. The grass cover is dominated by *Stipa* spp., *Festuca* spp., *Elymus chinensis*, *Agropyron* spp., and *Koeleria macrantha*. All are C3 grasses and these genera are common throughout the Northern Hemisphere (Figure 2). *Koeleria macrantha* is also native to North America. A minor grass species that is native to both Eurasia and North America is *Hierochloe odorata*. In Eurasia it is known as bison grass and is used in some vodka as a flavoring. In North America it is known as sweetgrass and is used in Native American rituals.

The major grazers are or have been horses and asses, camel, antelope, and gazelles. Steppe polecat, black-footed ferret, gray wolf, and Siberian tiger represent some of the carnivores that once roamed in the steppes in greater number.

**Africa**

The grasslands of Africa are diverse, with many hundreds of grass species represented. Most are C4 grasses, with genera of *Andropogon*, *Setaria*, *Panicum*, *Aristida*, *Chloris*, *Digitaria*, *Eragrostis*, *Paspalum*, *Schizachyrium*, and *Sporobolus*, all represented by numerous species. *Festuca*, a C3 genus, is also present and other C3 grass species are numerous near the Mediterranean where winter rainfall supports them (Figure 2). There are also numerous C3 species in South Africa.

Elephant, antelopes, kudu, ostrich, giraffe, wildebeest, gnu, and zebra are but a few of the large mammals of African grasslands, which are richer in large animal species than any other grasslands. Lion, cheetah, leopard, wild dogs, and hyena are major predators of the grazers.

**Australian grasslands**

All photosynthetic types occur in Australia, with C3 grasses dominating south of 30° south latitude (Figure 4). NADP–ME grass species are the most abundant of the C4 grasses, but they are rare south of 30° south latitude. Xerophytic tussock grasses, dominated by Mitchell grass (*Astrebla* spp., C4 NAD–ME biochemical type), are widespread (40 million ha) on clay plains of inland Queensland.

Prior to settlement by humans, the major Australian grazers were wallabies, kangaroos, wombats, emus, rat–kangaroos, and lizards. Lacking were the hoofed animals universal in the grasslands of the other continents, but excluded from Australia by a deep ocean barrier. Europeans introduced cattle, sheep, goats, horses, and rabbits as today’s major grazers.

**SEE ALSO:** Biogeography; Soil fertility and management; Tropical savanna ecosystems; Zoogeography

**References**


GRASSLAND ECOSYSTEMS


Further reading


The basic principles of green capitalism

Green capitalism is a form of environmentalism that emphasizes the economic value of ecosystems and biological diversity and attempts to reduce human environmental impacts by ensuring that the importance of environmental services is reflected in the way that markets operate. It starts from the recognition that ecosystems perform a wide range of services that societies depend on. This includes provisioning services (e.g., the supply of water, food, and energy), regulating services (e.g., carbon sequestration and water purification), and also cultural services (e.g., recreational services such as ecotourism and outdoor sports). Green capitalism extends the economic concept of capital (assets used to produce goods and services) to include “natural capital.” The United Nations Natural Capital Declaration (UNEP 2012, 3) defines natural capital as “the stock of ecosystems that yields a renewable flow of goods and services that underpin the economy and provide inputs and direct and indirect benefits to businesses and society.”

Proponents of green capitalism see pollution, loss of biodiversity, and the unsustainable use of natural resources as a form of “market failure.” In other words, environmental degradation is the result of the failure of capitalist systems to account for the financial value of environmental services:

Capitalism, as practiced, is a financially profitable, nonsustainable aberration in human development. What might be called “industrial capitalism” does not fully conform to its accounting principles. It liquidates its capital and calls it income. It neglects to assign any value to the largest stocks of capital it employs – the natural resources and living systems, as well as the social and cultural systems that are the basis of human capital (Hawken, Lovins, and Lovins 1999, 5).

As a result of the inability to value natural capital, the profit motive at the heart of capitalist societies tends to drive environmental degradation, since it is cheaper to pollute than to control emissions and more profitable to use resources now than to save them for the future.

The solution from a green capitalist perspective is to factor the value of nature into the way markets operate to encourage producers to become more efficient and innovative in the way they use natural resources. Rather than relying on state or international regulation (so called “command and control” strategies) or demanding radical cultural, political, and economic changes, green capitalism is based on the premise that private property, entrepreneurial business, and economic growth can be good for the environment (Beckerman 1974). Green capitalism is also referred to as “natural capitalism” (Hawken, Lovins, and Lovins 1999), “free-market environmentalism” (Anderson and Leal 1991), “blue-green environmentalism,” or “eco-capitalism.”
Green capitalism’s roots

Belief in the power of markets to solve socioenvironmental issues is grounded in the ideas of classical and neoclassical economics where, in a world of finite resources and infinite human wants, markets are seen as the most efficient way of allocating scarce resources. According to free-market thinking, individuals and firms will rationally pursue their own wealth. Competition in a free market will therefore stimulate an entrepreneurial spirit of hard work, innovation, and efficiency. This produces more and better goods and services for everyone. Furthermore, resources are allocated to those who need them most (measured by the willingness to pay).

Green capitalism takes the idea of the “invisible hand” (i.e., the idea that markets are inherently self-regulating systems that should not be interfered with, especially by governments) and extends it to the environment.

Although green capitalism is a relatively recent form of environmentalism, it draws on a long history of Western environmental ideas. Its emphasis on the rational use of natural resources is reminiscent of early twentieth-century “wise use” environmental philosophies in the United States of America. These ideas emerged from the development of scientific forestry and the principle that forests could be managed and optimized to satisfy human needs. Foresters such as Gifford Pinchot rejected the romantic and preservationist views of nature promoted by early conservationists such as John Muir and Henry Thoreau. Instead, the wise use movement argued for efficiency, waste reduction, and the management of forest resources for multiple activities (e.g., logging, recreation, and wildlife conservation).

One of the most important ideas at the heart of green capitalism is that nature provides financial benefits to societies and that any damage to ecosystem function has an economic cost. The problem is that these costs are not factored into market exchanges. In other words, the price of goods and services does not reflect their environmental impacts. Environmental economists describe these as negative externalities. In *The Economics of Welfare* (1920) Arthur Pigou defined an externality as an economic activity whose cost or benefit affects someone who did not choose to incur it. For example, a factory produces consumer goods that are mostly consumed by people far away from the factory. The company owning the factory gets the financial benefit of selling the goods, while consumers get the benefit of purchasing and using the goods. However, the factory also emits pollution that has numerous economic costs. Industrial wastewater released into a river impacts on other river users (for example, recreational fishers) and also other businesses (such as fish farms). Atmospheric pollution can affect the health of nearby residents. These third parties gain none of the benefits of production (unless they buy the goods produced or are employed by the factory) and suffer all the costs of pollution.

A possible solution to the problem of negative externalities is to impose a tax on polluting activities to “internalize” the cost of pollution. These Pigovian taxes make polluting a costly activity. There is thus a strong incentive for companies to reduce pollution in order to reduce costs and stay competitive. This in turn changes consumer behavior, since they rationally seek to purchase the cheapest (and therefore less environmentally damaging) products.

As well as Pigovian taxes there are also Pigovian subsidies. These are designed to stimulate positive externalities. Examples include government subsidies for installing solar panels and feed-in tariffs for energy generated from renewable sources. Such measures are designed to encourage investment in green technologies.
GREEN CAPITALISM

that might not otherwise happen due to the high start-up costs.

The imposition of Pigovian taxes and subsidies faces a number of challenges. The biggest challenge is calculating the correct level of taxation necessary to counterbalance negative externalities. For example, carbon taxes have often been too low to stimulate changes in production and consumption. This is compounded by the fact that economic sectors that generate sizable externalities (the fossil fuel energy sector, for example) have often lobbied governments to keep environmental taxes low or block proposals for new taxes. Environmental taxes are also regressive, since poorer households spend a large proportion of their income on acquiring basic resources such as water and energy, and are thus disproportionately affected. Free-market environmentalists also argue that subsidies do not necessarily support the best or cheapest environmental solutions and are thus economically inefficient.

Alternative solutions have been proposed based on the belief that market-based solutions are more efficient ways to reduce externalities than regulation or taxes (Anderson and Leal 1991). The economist Ronald Coase (1960) argued that as long as there were clear property rights over natural resources and transaction costs were sufficiently low, negative externalities could be dealt with through negotiations between those creating the externalities and those affected by them. Returning to the example of a factory releasing pollutants into a river that affect a fish farm, the Coasian solution would be to assign property rights to the river. If the fish farm owns the rights to the river, the factory will have to compensate it for any impacts, thereby creating a strong incentive to reduce pollution. Conversely, if it is the factory that owns the property rights, it may accept payments from the fish farm in return for a reduction in pollution. These payments would compensate it for the opportunity for costs of reduced production or the costs of developing solutions to deal with the pollution. From a Coasian perspective, this delivers the most economically efficient solution without the need for expensive regulation or monitoring. Coase’s basic idea has since been expanded on and interpreted in various ways, including mathematical models of the most efficient and therefore “socially optimum” level of pollution and its correct price.

Examples of free-market environmental policies influenced by Coasian theory include pollution permits and emissions trading. In a “cap and trade” system, a central authority sets an overall limit on the amount of emissions of a particular pollutant. The total cap is then divided into individual units and either allocated or sold to polluters. This gives the polluter the right to emit a given quantity of pollutants. Surplus pollution permits can be sold. The advantage of this system is that polluters who can find quick, easy, and cheap ways to reduce pollution are rewarded by being able to sell surplus permits to those sectors and industries that find it more difficult. The ability to sell surplus permits creates a continual incentive to reduce emissions. Coasian policies move away from traditional “command and control” models where states and international organizations simply set regulatory limits on pollution at the point of emission (for example, vehicle exhaust emission standards for pollutants such as nitrogen dioxide and sulfur dioxide). While “command and control” solutions have the advantage of simplicity, once minimum pollution standards are met there are no incentives to go any further.

Although the various ideas underpinning green capitalism have a long history, it is only since the emergence of the concept of sustainable development in the 1980s that they have
GREEN CAPITALISM

become mainstream. This represents a significant shift away from the “limits to growth” and “zero growth” environmentalism of the 1960s and 1970s, which argued for radical political, economic, and cultural changes to drastically cut production and consumption. Green capitalism also fits well with neoliberal economic thinking, which places an emphasis on individual liberty, minimal involvement of the state, and free markets as the most efficient way to coordinate the diverse needs of people.

Emerging forms of green capitalism

Accounting for nature and paying for ecosystem services

There have been growing efforts to include the value of natural capital into business activities and government policy. These depend on (i) being able to calculate the value of various ecosystem services and (ii) creating mechanisms whereby those who benefit from ecosystem services pay those who maintain those services. Calculating the economic value of ecosystem services remains technically challenging. With regard to creating financial flows to pay for natural capital, the most advanced attempts to establish such schemes have been under the banner of Payments for Ecosystem Services (PES). These are defined as voluntary transactions that involve the purchase of a well-defined ecosystem service from a service provider, who is paid if (and only if) the provision of that ecosystem service is secured (Wunder 2005). From a green capitalist perspective the advantage of PES is that they conform to Coase’s view that environmental issues are best left to negotiations between individuals or groups with clear ownership rights over natural resources.

A green industrial revolution: harnessing the competitive and innovative aspects of market forces to improve manufacturing processes

Internalizing the value of natural capital into the operation of markets is the first step toward green capitalism. Once pollution has a cost and natural capital has financial value it is expected that the logic of capitalism will drive innovation and efficiency to reduce costs and maximize income. In the same way that the first industrial revolution harnessed machinery and new forms of fossil fuel energy to dramatically increase productivity, a green industrial revolution would harness technology to deliver radical improvements in efficiency.

Hawken, Lovins, and Lovins (1999) propose a range of ways that a green industrial revolution might occur. First, radical improvements in productivity and efficiency could allow societies to produce more from fewer resources. Looking at the automobile industry, for example, ultra-light “hypercars” with fuel-efficient engines could dramatically reduce fuel consumption. In order to generate radical leaps in productivity, Hawken, Lovins, and Lovins (1999) advocate biomimicry, a design principle that imitates biological processes and structures in order to improve manufacturing and create new materials. For example, the physical properties of spider silk could be mimicked to make ultra-strong and ultra-light materials.

The concept of biomimicry can also be extended to industrial processes through “cradle to cradle” manufacturing, which models processes on ecosystem nutrient flows and metabolic pathways. Cradle to cradle manufacturing involves two types of materials: (i) synthetic technical materials, which must be non-toxic and able to be used continuously in production cycles without losing their integrity (i.e., there is no “downcycling” to lower quality products);
and (ii) biological materials, which can be put back into ecosystems and nutrient cycles. There is also the closely related concept of closed-loop manufacturing, which involves planning manufacturing processes around the life cycle of materials so that they are reused with minimum waste.

Green consumerism and ecolabeling

As well as changes in the production of consumer goods, green capitalism also needs to be able to change consumer behavior. Green consumerism emphasizes the ability of “consumer power” to deliver more sustainable patterns of resource use. It is part of a broader trend of “ethical consumption.” This is based on the belief that consumer choices are driven by more than just price and are often based on the moral attributes of goods and services. “Eco” labeling works on the assumption that consumers can be provided with information about the conditions of production through the use of labels. This enables them to make informed decisions to reduce the environmental impacts of their consumptive patterns. The expectation is that producers will then react to changing demands in the marketplace by changing production to reduce environmental impacts.

In addition to changes in the types of goods purchased, green consumerism also needs to deliver changes in the amount of goods purchased, owned, and consumed. Green capitalists envisage a move away from economies based on the ownership of goods to “service and flow” economies where resource goods are primarily rented (Hawken, Lovins, and Lovins 1999).

Critiques of green capitalism

Green capitalism is based on the premise that market forces and profit motives can drive more sustainable resource use patterns. While green capitalist thinking has had considerable influence on environmental policy-making, it has given rise to a large body of critical work. The strongest critiques of green capitalism have emerged from a broadly Marxian perspective. Ultimately, these critiques argue that any attempts to reduce environmental impacts will need radical economic and cultural changes that are not possible within a capitalist framework.

Capitalism and the metabolic rift between societies and the ecosystems that support them

According to Green Marxist theory, the underlying drivers of environmental degradation in capitalist systems are: (i) capitalism’s logic of competition, economic growth, and a relentless increase in the productive and consumptive capacities of society; combined with (ii) capitalism’s social relations, which are based on an unequal distribution of wealth, property, and power. Capitalism thereby allows the resources that everyone depends on to be owned and exploited for profit by a wealthy elite. This results in a “metabolic rift” whereby people are increasingly separated – both spatially and socially – from the ecosystems that support them.

For example, in eighteenth- and nineteenth-century Britain, wealthy elites claimed that peasant agricultural practices were backward and unproductive and argued that land could be put to more profitable use. As a result, parliamentary land enclosures moved large areas of land from communal management into the private sphere. Peasants lost their rights to carry out subsistence activities such as grazing animals and collecting firewood. The process was often violent with peasants forcibly evicted. By seizing control over the means of agricultural production and subsistence, land enclosure forced peasants to seek
work in order to support themselves. However, given the changes in agricultural production and increased mechanization during the eighteenth and nineteenth centuries, work was increasingly unlikely to be found in rural areas. Workers were forced to move to growing industrial towns to seek employment in factories.

The changes brought about by land enclosure and urbanization not only led to social upheaval but also had serious environmental consequences. While industrialization led to obvious environmental impacts such as air and water pollution from factories, Foster (2000) argues that there were more profound changes in human–environment relations through disruptions in the exchanges of nutrients between society and the environment. The nineteenth-century chemist Justus von Liebig was the first to point out that urbanization led to a physical separation of food production and consumption. In agrarian societies, food production and consumption are carried out in close proximity and nutrient cycles can easily be maintained by returning organic waste (and therefore nutrients) to the land. However, with urbanization, food is produced in rural areas and moved to cities to be consumed. Waste, rather than being returned to the soil as fertilizer, is disposed of as sewage and the nutrient cycle broken. Marx drew on this idea to argue that capitalism not only robbed people of control over their livelihoods but also robbed the soil of its nutrients. So for Marx, the exploitation of natural resources and the exploitation of people were two sides of the same coin.

The concept of metabolic rifting has since been expanded spatially and ecologically by Marxist scholars, most notably by Foster (2000). They argue that processes such as agro-industrialization, industrial fishing, and aquaculture have focused on modifying biological processes to maximize yields and generate economies of scale at the expense of biological diversity and ecosystem function. Together with the globalization of food systems they have created vast commodity chains. For example, prawn farms in Asia not only clear mangroves to create their ponds, but often use wild-caught species as feedstock, leading to problems of over-fishing. Their produce is then shipped halfway across the world, using fossil fuels for refrigeration and transport, to be consumed by affluent Western consumers. Industrial food systems and global commodity chains mean that consumers have very little knowledge or interest in the environmental impacts of their consumptive actions. The metabolic rift is therefore not just spatial and physical but also social and cultural.

The treadmill of production and the continuous expansion of capital

Schnaiberg’s (1980) concept of the “treadmill of production” combines insights from environmental sociology and technology studies to understand how and why capitalism tends to push production and innovation down certain paths. According to the treadmill of production, the competitive logic of capitalism means individuals and firms must always invest, innovate, and grow or risk being outcompeted. In other words, they must always be running to stay in the same place.

Capitalism’s competitive logic has major implications for the way firms and individuals behave, especially with regards to what they do with profits and gains in productivity. In theory, profits can be used in a variety of ways. For example, they might be used to raise worker wages and improve working conditions. The productive gains provided by innovation also open up various possibilities, including the option of producing the same amount of goods with fewer resources. The latter is the basis of the
green capitalist model for a “green” industrial revolution.

Treadmill theory argues that in reality, capitalism’s competitive logic means that profits will always be reinvested back into increasing production. Any productive gains provided by innovation will be used to produce more with the same resources rather than producing the same with fewer resources. This has important environmental implications, since every “turn” of the treadmill involves the extraction of resources and the addition of pollutants. Furthermore, the treadmill of production in turn fuels a treadmill of consumption. By stimulating competition between firms, capitalism drives ever-increasing efforts to sell us goods and services through branding and advertising. This inevitably creates mass consumption societies where more and more goods are manufactured, purchased, and disposed. The cultural values of mass consumer societies, where individuals define themselves primarily by the goods that they purchase and own, creates a significant barrier for the implementation of a “service and flow” economy in which individuals mostly rent goods.

“Green grabbing” and the neoliberalization of environmentalism

Green capitalism’s emphasis on natural capital involves commodifying material substances and processes that have never previously been commodities, for example, carbon dioxide and carbon sequestration. This leads some critics to consider green capitalism as just one part of a much broader neoliberal encroachment of market relations into various spheres of human life. Critics argue that green capitalism is leading to fundamental changes in environmentalism. This is especially the case with international environmental nongovernmental organizations (NGOs), which play a key role in green capitalism by providing expert knowledge about the functioning and value of ecosystem services. They have thereby gone from being at the margins of global capitalism (and often critiquing the practices of trans-national corporations) to acting as facilitators in the expansion of capitalism.

There are concerns that attempts to commodify natural capital will lead to individuals and groups being dispossessed of their land by more powerful state and corporate actors. This is because market-based environmental solutions are based on establishing property rights over the environment. Any scheme that seeks to create payments for environmental services must be based on a clear agreement of the precise ecosystem service being purchased, as well as who will be paid for delivering it. However, many ecosystems have multiple, complex, and often contested values as well as ownership rights. For example, many forests in sub-Saharan Africa are in theory owned by the state but in practice managed by rural communities according to customary rules. There are worries that the financial value suddenly created from ecosystem services will lead to strong incentives for powerful interests to take control of these ecosystems. Market-based schemes could see rural communities dispossessed of their rights to land so that ecosystem services can be secured. Critics argue that much like the “land grabs” that occur through the enclosure of agricultural land, green capitalism will lead to “green grabs,” where land is acquired to benefit from the revenue streams created by payments for environmental services (Fairhead, Leach, and Scoones 2012).

Ecolabeling and the fetishization of commodities

Green capitalism’s emphasis on consumer power has also come under considerable criticism. Scholars have drawn on the concept of
commodity fetishism to argue that green consumerism will only ever be a niche activity and may also act as a barrier to meaningful changes in consumption. In the common usage of the word, the term “fetish” is used to describe a strong and also unusual desire for an object or activity. It is also used in anthropology to describe objects believed to have magical powers, for example, religious idols. According to Marxist theory, capitalism encourages both forms of fetishism with regard to commodities. This occurs because commodities are primarily produced for the purpose of making profit. This means that their exchange values are prioritized over their use values or the value of the labor that went into creating them. Furthermore, individuals in mass consumer societies tend to define themselves through conspicuous consumption, so that consumer goods are purchased less for the functions they perform and more for their status-enhancing properties.

Capitalism is particularly good at fetishizing objects because of its social relations, which separate workers from the objects that they produce and separate consumers from workers. This means that consumers do not know where the goods they consume are produced, by whom, and under which social and environmental conditions. In theory, solutions such as ecolabeling help to defetishize commodities by providing consumers with information about the environmental conditions of production. However, Green Marxists argue that by reducing complex socioecological problems to simple labels and brands, ecolabeling in fact fetishizes commodities in new ways by selling the idea to consumers that they are “saving the environment” simply through their consumptive choices. Critics argue that green consumerism’s emphasis on consumer power in fact hides the need for more radical changes, especially in terms of how much individuals consume.

SEE ALSO: Commodification of nature; Conservation and capitalism; Ecological modernization; Ecosystem services; Environment and consumption; Environmental certification and eco-labeling; Environmental valuation; Environmentalism; Neoliberalism and the environment; Property and environment; Sustainable development

References


Green infrastructure is increasingly used as a way of bundling a host of components of the natural environment into a single heading or category. The term refers to a wide range of things, including parks, rivers, street trees, natural floodways, nutrient recycling schemes, and protected habitats such as forests and moorland. The term is mobilized in policy documents to assert the value of natural environmental processes and materials in enhancing healthy living, maintaining urban environments, addressing sustainability questions from local to planetary scales, even addressing threats to property values in a context of industrial decline. Thus, high levels of quality green infrastructure are argued for in order to combat anthropogenic climate change, manage storm events and flooding, moderate urban heat and coldness events, enrich agricultural assets including in urban areas, and enhance the quantity and quality of sub-surface water quality through better absorption processes.

These sorts of arguments have made green infrastructure a popular policy term in development and planning policies in both advanced and developing nations. But there are implications of such use, especially in terms of how the natural environment and nonhuman species are valued.

**Definition**

Because it is a pragmatic term, it is difficult to nominate a single correct use of the term “green infrastructure.” In many ways the term has emerged as a way of tidying up the growing number of ways that the natural environment is referred to in planning and policy documents. A handy article explaining the evolution and, at times, contested meanings for the term is Wright (2011). Because it is a term in common use, however, it is handy to have a working definition for it. Hence, for here at least, green infrastructure refers to those components of the natural environment that are assembled or mobilized to underpin human settlement. This underpinning includes the provision of the basic elements of life and then the supply of environmental services and materials in ways that enhance human productivity and the quality of human existence.

Underpinning the idea of green infrastructure is the assertion that the environment is more than a passive bystander or context in the day-to-day operations of cities and regions. Drawing on the meaning of the word “infrastructure” – and thereby a universal acceptance that society has to build and operate a wide range of infrastructure assets in order for its members to live successfully, especially in urban conglomerations – the idea of green infrastructure asserts that human settlement draws by necessity on things like green spaces, resources-rich ecosystems, and cohabiting nonhuman life. Most importantly, by attaching the simple adjective “green” to the
powerful economic word “infrastructure” the phrase implants an inseparable economic logic to the presence and acceptance of a cared-for natural environment in and around human settlement. This appropriation parallels the successful development of the idea of “social infrastructure” in the 1960s when social development theorists and advocates argued successfully that state expenditure on education, health, and social welfare was vital to enduring economic growth and prosperity (Rankin 2009). Likewise, the idea of green infrastructure is very much an opportunistic and politically pragmatic view of the natural world.

As an infrastructure asset, green infrastructure can be seen to take on various characteristics of infrastructure assets in general (O’Neill 2013). These characteristics are fourfold: that infrastructure mobilizes factors of production, expands markets, enhances the productivity of both labor and capital, and supplies essential services not capable of being supplied by private capital on a for-profit basis or without unfair monopolization practices. Talking about the environment in infrastructure terms, then, says that the environment is a key supplier of factors of production, particularly resources, such that the protection and nourishment of ecological systems are as vital to running a successful economy and a prosperous society as are the provision of reliable energy, water, and transport systems. Second, this conceptualization consciously removes the divide between urban life and the natural environment and between the economy and the natural environment by accepting the vital role of the environment in not just generating the essential conditions of life (air, water, sunlight, food) but also maintaining the stocks and flows of energy and materials essential for the formation of both the means of production and the final consumption goods and services. Green infrastructure, then, can be seen to be a prime determinant of market growth and innovation in capitalist economies. Third, the idea of there being “green” infrastructure says that both capital and labor will be more productive in the presence of quality green assets. Better quality natural environments in cities, cleaner work environments, and superior natural materials enhance the work process, generate higher incomes, and fund better employment conditions. Likewise, plant and equipment (capital) operate more efficiently with lower maintenance costs and fewer delays (e.g., due to congested supply chains) when their host cities are pollution-free with environmentally sensitive transport and energy systems, and so on. Fourth, inherent to the idea of green infrastructure is the assumption that the components of green infrastructure are best treated as societal “commons,” meaning their operation would be distorted by commercialization and marketization, and by their control and supply by for-profit private producers. In this sense, green infrastructure assets are seen as natural monopolies with essential public goods characteristics. Therefore, like public health and education systems, for example, the supply of green infrastructure assets is best organized by some sort of public delivery mechanism, or through private delivery under a regime of heavy public regulation or contractual conditions.

Major dimensions of green infrastructure

A useful discussion of the components and dimension of green infrastructure can be found in Tzoulas et al. (2007). Drawing on this work, a number of dimensions of green infrastructure can be nominated: its green space role; its ecosystem services role; its complementary contribution to human health; its functioning as an economic resource; and its place in human thought and planning. These are dealt with in turn here.
Green spaces

Green infrastructures evolution as an integral part of a city or a region can be traced to the creation of large city parks during Victorian times, places that went beyond carefully cultivated display gardens or civic gathering spaces. Large parks like Central Park in New York and Hyde Park in London included parts which purposely mimicked natural forests and woodlands. There were multiple motivations for such extensions to the view of what was acceptable inside a city. Some were scientific, others were recreational, aesthetic, and spiritual, and there were fledgling arguments that natural spaces enhanced city living, hence such parks carried descriptors such as the “lungs of a city.” By the twentieth century, green spaces were accorded their own land-use classifications, and then became tools for planners to curb, guide, and shape the growth of urban areas. Green spaces as urban planning tools were thus manifest in greenways, greenbelts, garden cities, and new towns (Wright 2011). This evolution from a relatively idle urban space into a purposeful planning action was important in enlisting green spaces as mainstream components of cities, and establishing an argument for their preservation and maintenance, not to be obliterated by economic or residential functions as the opportunity presented.

Ecosystems and their health

Beyond green spaces as land-use zonings and buffers, and protected habitats and conservation zones, green infrastructure also refers to the presence of regenerating, operative ecosystems as active participants in cities and regions. Sometimes these aspects of green infrastructure are referred to collectively as “ecosystem services.” Thus, green infrastructure can be seen as capable of addressing obstinate environmental concerns, such as air pollution and smog events, generated by the capture of mass automotive and industrial pollution within inverted layers of the atmosphere; and urban heat islands where the heat output of urban buildings, especially high-rise office buildings, is exacerbated by direct heat reflection and absorption by common road, pavement and building materials, and urban surfaces, especially concrete and asphalt.

At one level this means an accepted presence in highly populated areas of things like habitat trails and corridors, acknowledging the rights of nonhuman species to have formally constituted transport and breeding spaces, and thus accepting nonhuman species as normal occupants of a city or region. At another level, there is the utilitarian notion that cities and regions that are host to thriving ecosystems generate better economic and social conditions for human life. This could involve, on one hand, the active preservation of pre-existing thriving natural ecosystems or, on the other hand, a conscious program of detoxification, rehabilitation, and species regeneration on lands degraded by, say, decades of industrialization, land clearing, waterway channeling, and straight-out neglect. It might also involve reconsideration of hard infrastructure items, such as water collection and discharge systems, and their reincorporation into water catchment and wetland rehabilitation schemes in ways that allow high levels of natural ecosystem functioning (Faucette 2012; Benedict and McMahon 2006).

At a regional scale, a green infrastructure approach might involve the integration of urban areas with natural environmental zones and sympathetic agricultural land uses in ways that are beneficial for all users. Such integration offers prospects for efficiencies across multiple economic sectors as well as in civil and other infrastructure investments, notably flood mitigation systems. This integrated provision is highlighted in the European Commission’s 2013 report into green infrastructure, a document...
GREEN INFRASTRUCTURE

which mandates green infrastructure as a core concern in urban and regional planning.

For human health

Green infrastructure is seen as a major contributor to the conditions for good human health, especially in urban settings. Urban forests, woodlands, and wetlands are seen to improve local air quality, moderate temperature extremes, and contribute to better groundwater quality, all of which have positive human health implications. Greenspaces are also important spaces for formal and informal recreation and sporting activity, as well as being desirable spaces for meditation, reflection, and intimate human exchanges.

For aggregated urban and economic resources

Green infrastructure is increasingly seen as a prime economic resource for a number of reasons. Properly designed and managed green spaces enhance other urban infrastructure assets. The ways that green spaces can be deployed as relatively inexpensive assets in storm water and flood management have already been discussed. Properly managed, large belts of green space are also useful as passageways and thoroughfares for the operation of other infrastructure assets, including roads, electricity wires, gas and water pipes, telecommunications cables and towers, and so on. Green infrastructure also gives rise to improved productivity of natural resources. Water, soils, timber supply, even breeding conditions for domesticated animals for food are all enhanced by well-functioning ecosystems.

For thinking and planning

Given the dimensions of green infrastructure explored above, it is not surprising to find that green infrastructure is accepted increasingly as a legitimate budget item for urban and regional development spending in both advanced and developing nations. This progression is now actively encouraged by international agencies such as the World Bank (Baietti and Shlyakhtenko 2012). At the same time, thinking about the nature and desirability of coexistence with functioning ecosystems gives rise to non-instrumentalist ways of valuing green assets. On the one hand this involves respect for the presence of nature for its own sake; and, then, the understanding that the presence of wild things in and around us is mutually beneficial, for example, in major questions like solving urban social issues and addressing climate change challenges.

Elevating green spaces and natural things to having a legitimate copresence in cities and regions also raises the question of the priority status of green infrastructure assets. Is green infrastructure necessarily accorded the status of a “commons” in property terms; or, like any other infrastructure asset, can it be held in a variety of property forms, ranging from discrete private property units through to publicly owned and operated commons title? Beyond ownership, there is also the question of the governance of green infrastructure, including the extent to which the green infrastructure assets of a city, be they in public or private ownership, can be regulated in ways that control their use, performance, and ongoing availability. These questions are more readily resolved with traditional environmental assets such as waterways and parklands, but require complex negotiation when green assets are held in private hands – such as a pristine natural habitat on a private estate – while delivering considerable public benefit.
Future directions in research, theory, and methodology

Although never mentioning the phrase directly, Jennifer Wolch (2007) provides an important exposition of the options for deployment of green infrastructure in creating cities that hold more promise for the future, especially regarding better social and ecological justice outcomes. Wolch notes that it has become common, even in rich cities, for urban residents to be forced to tolerate toxic soils, airborne pollutants, poor water quality, and species eradication. A source of the problem, says Wolch, were twentieth-century ideas that constructed dualisms between city and wilderness and between nature and culture. It has already been seen how green spaces were enlisted in early twentieth-century planning not as spaces for integrated, sustainable human use but as separation or buffer zones to human activity, as refuges for escape from the intense nature of urban activity, or as some sort of environmental recharge device to deal with urban excesses, especially in regard to air and water but also for material waste, such as landfill. As Wolch notes, it was rare for urban scholars in the twentieth century to think about “rivers, or oak trees or red-legged frogs,” and beyond issues of economic growth, class, gender and racial equality, human behavior (voting, commuting, shopping, parenting), and political movements and processes (Wolch 2007, 373). But global and local awareness of environmental devastation and unsustainability have dramatically changed the ways our urban world is seen, with growing cognizance of the inseparability of human and nonhuman worlds and of the need for significant and urgent responses. Nature and urbanism, says Wolch, are increasingly spoken of as integrated parts of a new approach to environmentalism.

Wolch illustrates her arguments with a case study of the Los Angeles River – and it is here that Wolch presents an agenda for interpretation and analysis, and for policy change, that can be easily tagged as belonging to a contemporary green infrastructure approach to urban and regional management. The Los Angeles River runs through Los Angeles from the outer affluent suburbs at the foot of the San Gabriel Mountains, through the movie-making center of San Fernando, through downtown LA and its adjoining industrial districts and poor inner neighborhoods, and then across a vast military industrial complex to the river mouth at Long Beach, with much of its urban journey steered by concrete channels and pipes.

Concrete passageways and diversions were built into the Los Angeles River from the 1930s as a way of controlling its waywardness during flooding events. The river needed to be tamed in order for urban activity to proceed. A by-product of flood engineering, however, was the destruction of major riparian habitats, the loss of infiltration and groundwater replenishment systems, and the removal of fish breeding spaces. The river as a key asset in the Californian water cycle was broken. It had ceased functioning, if you like, as a core piece of the city’s green infrastructure.

Replenishing and fixing the Los Angeles River required major engineering, political, and behavioral elements, which Wolch documents. These are encapsulated in the state strategy “Green Visions Plan for 21st-Century Southern California.” This strategy operates at a number of levels. Importantly, it builds a regional vision that integrates the river into urban life by promoting its restoration as essential to watershed health, habitat conservation, and recreational open space. Then the plan assembles parcels of data and analysis tools to assess, monitor, and guide changes to the way the river exists as an urban entity. The document argues for a better understanding of what will happen as water flows are changed in speed and direction, as
GREEN INFRASTRUCTURE

Local flooding events are accepted as normal, and as nonhuman species are invited back into the valley to live.

Beyond a direct focus on the river itself, Wolch shows how the emergence of a planning approach that values and nurtures green infrastructure requires also a reappraisal of the human behaviors that lead to stress on local water flows. These include consumption behaviors responsible for the waste that requires landfill, energy systems whose waste products end up in parts of the water cycle, and the worrisome ownership and investment practices that impede not-for-profit uses of riparian lands, including sustainable small-scale agriculture and wetland regeneration schemes. Central to a new planning approach, says Wolch, is the development of new types of “urban ecological citizenship” (Wolch 2007, 380) capable of devising or insisting on the governance processes necessary to power-up the claims for a renewed Los Angeles River environment. In this sense, the claim is for the river to eventually need no justification for riparian and water health other than for being a river. Perhaps this is why Wolch avoids the term “green infrastructure”; perhaps she sees green infrastructure as creating a view of nature as being a servant of human occupancy of the planet, with its worth measured by the contribution it can make in terms of services to human life. Perhaps she sees the river as having its own integrity and rights.

SEE ALSO: Built environments; Green capitalism

References


Ground-based LiDAR

Ruisheng Wang  
*University of Calgary, Canada*

Ground-based LiDAR (light detection and ranging), also called terrestrial laser scanning, is an active imaging system that uses LiDAR measurement technology at ground level. Each LiDAR measurement contains range and intensity values. The range values combined with the LiDAR sensor’s position and orientation produce X, Y, Z values; the intensity values are determined by the laser property and reflectivity of the surface. Ground-based LiDAR is used for close-range applications, such as heritage documentation, façade modeling, mapping and monitoring of road infrastructure, and precision forestry. In comparison with aerial LiDAR data, which are normally sparse with low laser ranging accuracy, the point clouds from ground-based LiDAR are relatively dense and have higher accuracy.

Ground-based LiDAR systems can be mounted on a stationary platform, such as a tripod, or on a moving platform, such as a vehicle, trolley, or boat. A lightweight LiDAR scanner can be portable and used for a handheld 3-D mapping system (CSIRO 2016). The choice of the platforms depends on the application. For instance, a tripod platform is suited for detailed documentation of cultural heritage, while the use of a moving vehicular platform is suitable for large-scale corridor mapping tasks. A hand-held device may be fit for indoor mapping applications. In general, ground-based LiDAR systems are often used for producing detailed 3-D models of buildings, trees, bridges, and other artificially constructed or natural objects.

Airborne platforms are frequently used for topographic mapping and digital terrain model (DTM) generation applications.

The laser scanner is an essential component of ground-based LiDAR systems. A laser scanner can be composed of three subcomponents: the optomechanical scanner, the ranging unit, and the control processing unit (Yang et al. 2015). The optomechanical scanner generates a laser pulse that is fed to the ranging unit. These pulses are reflected to the target via a rotating mirror within the ranging unit. There is also an electro-optical receiver within the ranging unit that records the total laser transmission and reception travel time. This value is passed to the control processing unit to further calculate the distance between the laser scanner and the target. This process occurs at a rate of 100,000–200,000 times per second (Yang et al. 2015); this timing accuracy is an important factor in determining the accuracy of LiDAR systems.

Unlike aerial LiDAR, which is normally georeferenced via GPS (global positioning system)/IMU (inertial measurement unit), terrestrial laser scanning on stationary platform georeferenced data cannot be provided by the laser scanner itself. The coordinates of the resultant point clouds are only related to arbitrary local coordinate systems. When multiple scanning at different locations is required to produce a consistent dataset for a complete scene, registration among these scanning data is needed. Using iterative closest point algorithms (Besl and McKay 1992) is a common practice to align multiple scanning data into a consistent point cloud. To relate point clouds to geodetic data, either GCPs (ground control points) or a direct
A georeferencing component is required. A GCP can be defined as a point on the surface of the Earth with known geographic location that is used to georeference data collected from sensors, such as remotely sensed images or point clouds. A direct georeferencing system relates data collected to the Earth by measuring the geographic position and orientation of the sensor, rather than using GCPs.

Ground-based LiDAR on a stationary platform provides point clouds with good accuracy due to its stable scanning. However, this scanning mode is not suitable for large-scale applications, such as scanning an entire city or mapping long distance corridors.

Ground-based LiDAR mounted on a moving platform, also known as mobile LiDAR, is a newly emerging technology for large-scale mapping applications. A mobile LiDAR system is defined as an integrated platform consisting of one or more short-range laser scanners (e.g., 100–200 m range), a high-precision positional system, and, optionally, one or more electro-optical cameras (Graham 2010).

In comparison with an airborne platform, mobile LiDAR systems are often operated at street level in cities, where GPS signals can be intermittently lost or affected by induced errors due to tall buildings. It is known that the IMU aids GPS positioning when the satellite signals are intermittently lost for short periods. However, over longer periods of GPS outage, the positioning information will start to degrade due to the problem of IMU drift. To overcome this problem, mobile LiDAR systems employ additional motion-detecting sensors, DMIs (distance measurement indicators), to provide corrections to GPS/IMU solutions. Since the data have been georeferenced to geodetic data, the resultant point clouds compose a scene without the need of registration.

Mobile LiDAR systems collect 3-D and visual data while vehicles drive at a posted speed; it is an extremely efficient data acquisition technology. In corridor mapping applications, such as railway or highway corridor mapping, mobile LiDAR systems are becoming a major technique for direct 3-D data collection (Graham 2010). In particular, companies such as Microsoft and NAVTEQ have adopted this technology and jointly deployed more than 40 mobile LiDAR mapping vehicles for their mapping applications. Google also introduced a significant number of mobile mapping vehicles for their street view data collection.

There are two types of mobile LiDAR systems in terms of the data accuracy provided: survey grade and geographic information system (GIS) grade. Survey-grade mobile LiDAR systems are designed to collect engineering/survey grade LiDAR data. The data resolution can be as good as 1 cm and the absolute accuracy can be 1–5 cm. Such accuracy is suitable for certain engineering applications where accuracy is an important factor, such as road crack detection and compliance certification. GIS-grade mobile LiDAR systems provide quality data that requires lower accuracy than design engineering projects, which suits applications where accuracy is not primarily concerned, such as visualization and animation.

Ground-based LiDAR has found wide applications in many disciplines. Terrestrial laser scanning can be used to determine parameters of trees, such as height or crown width in forestry, and can also be used to render detailed tree structure in computer graphics. In geomatics, it is often used for documenting heritage or pipelines and so on. Obviously, the scales of these applications are limited.

Mobile LiDAR is becoming an important tool for road infrastructure mapping and modeling. It has been used to obtain the geometry of roads.
and road attributes such as curbs, lane markings, traffic signs, and poles. Furthermore, it is also being used for detecting road cracks for roadway asset management. Research has indicated that the accuracy of crack detection and width measurement as well as the accuracy and repeatability of rut depth measurement can be improved using LiDAR technology. A mobile LiDAR system is a promising and cost-effective solution to assess pavement surface conditions, such as rutting and cracking, and for detecting potholes. Besides road pavement assessment, mobile LiDAR systems can be utilized for detecting and assessing conditions for traffic signs.

Mobile LiDAR can be also used for detailed modeling of building façades. Building models produced from airborne LiDAR lack detailed building façade information and are not suitable for applications where street-level representation of cities is required. Mobile LiDAR systems can produce accurate georeferenced 3-D points with sufficient detail in a short time and provide great potential for the generation of detailed 3-D models. However, due to the line-of-sight problems of any ground-based data acquisition system, it is difficult to always obtain complete and sufficient data from all the building surfaces, such as rooftops. The combination of mobile and aerial LiDAR will contain information from both a top-down perspective and a street-level view and will, therefore, be a valuable data source for large-scale photorealistic urban modeling.

SEE ALSO: Geodesy; Geographic information system; Global navigation satellite systems; LiDAR

References


Further reading


Ground-based radar

Liming Jiang

Institute of Geodesy and Geophysics, Chinese Academy of Sciences, China

Ground-based radar (RAdio Detection and Ranging) is an active instrument on board a terrestrial platform; it measures the strength and round-trip time of microwave signals that are emitted by a radar antenna and reflected off the scattering surfaces or objects. Radars have a wide variety of working forms in terms of their various applications in military and civilian fields. This entry, however, focuses on an advanced imaging radar called synthetic aperture radar (SAR), which can produce high-resolution remote sensing imagery. Specifically, ground-based SAR (GBSAR) is a relatively new imaging radar that has been developed to measure surface topography and deformation by exploiting the interferometric capability of the received microwave signal, which is based on the same principles as the well-established spaceborne SAR technique (Henderson and Lewis 1998).

Typically, it is equipped with two antennas and a radar sensor that moves along a linear rail of a certain length, usually 1–3 m (Figure 1). This configuration allows the acquisition of 2-D SAR images by synthesizing an antenna aperture electronically.

As a terrestrial radar imaging technique, GBSAR is characterized by the following unique features: (i) imaging capacity with high spatial resolution (submeters to meters) and nearly real-time sampling (10 s to minutes), independent of weather conditions and sunlight illumination; (ii) a long measurement range, up to several kilometers; (iii) high flexibility in terms of data acquisition over a specific area of interest, at the required time and with the needed viewing angle; (iv) high precision and accuracy of deformation measurements (submillimeters to millimeters); and (v) a wide applicability range, for monitoring deformation phenomena from a few millimeters per year up to meters per day. Consequently, these enhanced characteristics of the GBSAR represent a fundamental advantage with respect to airborne or spaceborne SAR systems, and also make it complementary to other remote sensing techniques (e.g., terrestrial LiDAR scanner, digital photogrammetry and topographic total station) for topography and deformation measurements.

There are two different acquisition modes to obtain GBSAR data: (i) continuous acquisition mode (C-GBSAR) and (ii) revisit acquisition model (R-GBSAR). In the continuous mode, the GBSAR is mounted on a regular ground base and acquires SAR images with high temporal sampling, for example, every a few minutes. In the revisit mode, the instrument is installed and dismounted at each campaign by revisiting a site to acquire data with a given time interval, for example, weeks, months or years, depending on the deformation kinematics of interest. The C-GBSAR allows a near-real-time monitoring of ground deformation and can be used to measure fast deformation phenomena with displacement in the order of mm day$^{-1}$ to m day$^{-1}$ (Rödelsperger 2011). The R-GBSAR, on the other hand, is appropriate for measuring slow ground deformations and can be used to monitor several survey sites of interest with the same instrument; in this case the C-GBSAR is not recommended due to either logistic or cost...
**GROUND-BASED RADAR**

![Diagram of ground-based radar](image)

**Figure 1** Working principle of ground-based SAR.

The radar sensor transmits, step-by-step, continuous waves at discrete frequencies over a frequency range. Furthermore, GBSAR measures the in-phase and quadrature components of the received signal at each frequency and position on the rail.

**Scattering surface**

Tx is a transmitting antenna and Rx a receiving antenna. The radar sensor transmits, step-by-step, continuous waves at discrete frequencies over a frequency range. Furthermore, GBSAR measures the in-phase and quadrature components of the received signal at each frequency and position on the rail.

Reasons (Monserrat, Crosetto, and Luzi 2014).

The phase information of the GBSAR complex data can be exploited for digital elevation model (DEM) generation and deformation measurement using interferometric techniques. It is worth noting that the deformation measurement is usually performed using a zero-baseline GBSAR configuration where all images are taken from the same position, whereas a nonzero baseline configuration is needed when GBSAR measurements are aimed at DEM generation (Monserrat, Crosetto, and Luzi 2014). Generally, the working principle of ground-based interferometric SAR (GBInSAR) is similar to that of satellite SAR interferometry (Hanssen 2001), and key interferometric processing steps consist of image coregistration, interferogram generation, phase unwrapping, atmospheric effect correction, displacement calculation, and geocoding. In this entry, detailed descriptions on the InSAR basic principle are not discussed; readers are referred to the literature (Hanssen 2001; Rödelsperger 2011).

Since the first prototype GBSAR system operating with the radar sensor based on a vectorial network analyzer (VNA) technology, a number of GBSAR systems have been developed by companies and research teams from the European Commission (EC), Spain (ES), Italy (IT), Ukraine (UA), Netherlands (NL), Switzerland (SW), China (CN), Japan (JP), United Kingdom (UK), and Korea (KR). Table 1 lists the main GBSAR systems presented in the literature; they are generally classified by the two major characteristics of (i) radar type – VNA, SFCW (step frequency continuous wave), FMCW (frequency modulation continuous wave), and others, for example MIMO (multiple-input and multiple-output) – and (ii) scanning model – linear scanning and angular scanning.
Table 1  Available main GBSAR systems and their major characteristics.

<table>
<thead>
<tr>
<th>Owner (country)</th>
<th>Name</th>
<th>Radar type</th>
<th>Band</th>
<th>Scanning model</th>
<th>Range resolution (m)</th>
<th>Azimuth resolution (mrad m(^{-1})) at 1 km</th>
<th>Nominal precision (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Research Centre (EC)</td>
<td>Lisa</td>
<td>VNA</td>
<td>C/Ku</td>
<td>Linear</td>
<td>0.5</td>
<td>3</td>
<td>0.02–4</td>
</tr>
<tr>
<td>Technical University of Catalonia (ES)</td>
<td>RiskSAR</td>
<td>FMCW</td>
<td>X</td>
<td>Linear</td>
<td>1.25</td>
<td>4</td>
<td>1.6</td>
</tr>
<tr>
<td>IDS SpA (IT)</td>
<td>IBIS-L/M</td>
<td>SFCW</td>
<td>Ku</td>
<td>Linear</td>
<td>0.5/0.75</td>
<td>4.4</td>
<td>0.03–4</td>
</tr>
<tr>
<td>Institute for Radiophysics and Electronics (UA)</td>
<td>GB NM-SAR Noise radar</td>
<td>Noise radar</td>
<td>Ka</td>
<td>Angular</td>
<td>1</td>
<td>12</td>
<td>NA</td>
</tr>
<tr>
<td>MetaSensing (NL)</td>
<td>FastGBSAR</td>
<td>FMCW</td>
<td>Ku</td>
<td>Linear</td>
<td>0.5</td>
<td>4.5</td>
<td>0.1</td>
</tr>
<tr>
<td>Nonstrictly SAR GB systems GAMMA remote sensing AG (SW)</td>
<td>GPRI</td>
<td>FMCW</td>
<td>Ku</td>
<td>Angular</td>
<td>0.75</td>
<td>7</td>
<td>0.02–4</td>
</tr>
<tr>
<td>JRC (EC)</td>
<td>Melissa</td>
<td>MIMO</td>
<td>Ku</td>
<td>No motion</td>
<td>0.89</td>
<td>1.2</td>
<td>0.01</td>
</tr>
<tr>
<td>National Key Laboratory of Microwave Imaging Technology (CN)</td>
<td>ASTRO</td>
<td>MIMO</td>
<td>C</td>
<td>Linear</td>
<td>NA</td>
<td>5</td>
<td>0.1–1.5</td>
</tr>
<tr>
<td>Tohoku University (JP)</td>
<td>—</td>
<td>VNA</td>
<td>S/C/X</td>
<td>Linear</td>
<td>0.4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Centre for Earth Observation Science, Sheffield University/Cranfield University (UK)</td>
<td>—</td>
<td>VNA</td>
<td>C/X</td>
<td>Linear</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>KIGAM Korean Institute of Geoscience and Mineral Resources (KR)</td>
<td>—</td>
<td>VNA</td>
<td>C</td>
<td>Arc</td>
<td>0.25</td>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>National University of Defense Technology (CN)</td>
<td>—</td>
<td>SFCW</td>
<td>L</td>
<td>Linear</td>
<td>0.25</td>
<td>1.9</td>
<td>NA</td>
</tr>
</tbody>
</table>
GROUND-BASED RADAR

In the last decade, the GBSAR has gained increasing interest as a terrain mapping and early warning tool for use in a wide range of environmental and geological applications (Monserrat, Crosetto, and Luzi 2014). These applications of GBSAR, especially of GBInSAR, take advantage of the unique capabilities already discussed to measure surface displacement and topography. The GBSAR with a nonzero baseline configuration has been proven to be effective in generating high-resolution DEMs of restricted areas (e.g., within a ranging distance of few kilometers) with an accuracy of about 5 m. The resultant DEMs can be subsequently utilized to geocode the GBSAR data and mitigate the atmospheric influences, hence, correctly interpreting and exploiting the displacement measurements.

The main GBSAR application filed is deformation monitoring (Monserrat, Crosetto, and Luzi 2014). In recent years, monitoring of natural and artificial instable slopes related to different phenomena, such as rockslides, open pit slopes, landslides, volcanic movement, and snow covered slopes, has become the most consolidated deformation-related application of the GBInSAR technique. In fact, great developments at both radar systems and data processing levels have been achieved in the last few years, which enables this technique to be a real-time monitoring and operational early warning tool to assess the risk of rapid landslides. Another group of important applications is the displacement monitoring of human-made structures, such as urban infrastructures, historical buildings, bridges, dams, and dikes. A major advantage for urban and structural monitoring is the capability to detect remotely small displacements with high spatial resolution and deformation sensitivity. Finally, recently the GBInSAR has gained increasing interest for investigating glaciers and snow, although some are not exactly related to deformations. Several studies demonstrated the potential of this technique for monitoring glacier dynamics (e.g., ice flow and elevation change) and snow avalanches at long range and high resolution. In addition, this technique can be used to retrieve snow depth and snow water equivalent, but these are still at an early research stage.

SEE ALSO: Ground-based LiDAR; Photogrammetry: 3-D from imagery; Radar remote sensing; Synthetic aperture radar

References


Grounded theory

LaDona G. Knigge
California State University, Chico, USA

Grounded theory is an inductive approach to research discovered by two sociologists in the late 1960s as a way of systematically collecting and analyzing qualitative data (Glaser and Strauss 1967). The purpose of grounded theory is to generate or discover theory inductively from data that is free from a priori assumptions. The process of generating theory from data requires that concepts, categories, and most hypotheses emerge from the systematic process of data collection and analysis in order to create theory that is grounded in empirical data and thereby bridging the gap between theory and practice. While grounded theory has its roots in the collection and analysis of qualitative data, its application was never limited to qualitative data. It is flexible enough and often requires the collection and analysis of quantitative data.

Grounded theory is an inductive, recursive research method that involves successive iterations of data collection, analysis, and verification in order to reveal categories and their properties as well as evidence to be collected and analyzed in the next round of research. Comparative analysis is central to the grounded theory methodology (Glaser and Strauss 1967). Generating theory through comparative analysis involves coding of data for categories and their properties and recording the findings in memos. Comparative analysis of additional similar or dissimilar instances allows conceptual categories to emerge and inform the next round of data collection.

Grounded theory’s process of comparative analysis serves several purposes: to verify evidence, to broaden the scope of the theory through empirical generalization, to specify a particular concept, and to generate and verify theory. Conceptual categories and their properties are derived from evidence obtained through comparative studies. By comparison with other similar groups, evidence can be verified for accuracy and be used to illustrate the conceptual category (Glaser and Strauss 1967). Constant comparison reveals similarities and differences in the data and generates conceptual categories and their properties. This empirical generalization both limits the extent of the theory and broadens its applicability and explanatory power as well as specifying the properties of a particular concept. As the research process proceeds and theoretical hypotheses emerge, comparative analysis allows the researcher to test the relevance and extent of the categories and their properties, to explore the applicability to different areas, and to search for similarities and differences, all the while systematically building and modifying theory (Glaser and Strauss 1967).

The sampling framework associated with Glaser and Strauss’s method of comparative analysis involves theoretical (not statistical) sampling to generate conceptual categories from the data. According to Glaser and Strauss “[t]heoretical sampling is the process of data collection for generating theory whereby the analyst jointly collects, codes, and analyzes his data and decides what data to collect next and where to find them, in order to develop his theory as it emerges” (1967, 45). Theoretical sampling involves the collection of diverse forms of data that may consist of interviews, conversations, field notes,
existing reports, historical records, or other secondary data that provide different angles from which the researcher understands the category and its properties. Theoretical sampling is intent upon discovering categories and their properties in order to make determinations about their relationships in order to generate theory. In contrast, statistical sampling involves the use of a scientifically developed framework in order to determine statistical distributions of the group or groups sampled. While statistical sampling requires that all instances of the sample of the predefined population be considered, the groups and categories in theoretical sampling arise from the coding and analysis of data, with the addition of new groups and categories that are identified as the research process proceeds.

Theoretical sampling of additional instances that are similar or dissimilar to the conceptual categories continues with the addition of new categories and properties derived from the coding process which informs the next round of data collection. As the research process proceeds, the direction and scope of the theoretical sampling is shaped by the emerging categories and their properties which are coded and analyzed. The researcher is constantly coding, analyzing, and writing up the findings in field notes or memos. In the course of the research, some categories become higher-level, integrative conceptualizations that are explored until further sampling does not reveal any new data about the properties of the particular category. When the researcher starts seeing the same instances again and again the category is said to have reached theoretical saturation, and further sampling does not reveal any new data to develop the properties of the category (Glaser and Strauss 1967). The theory may be for a substantive area or several substantive comparisons may generate formal theory in a formal, conceptual area of social life (Glaser and Strauss 1967).

The history of grounded theory

The development or “discovery” of grounded theory by sociologists Barney G. Glaser and Anselm L. Strauss was the result of the methods used in their study of death and dying in public institutions in conjunction with Public Health Service Research Grant NU-00047 from the Division of Nursing, Bureau of State Services—Community Health that led to the coauthored book Awareness of Dying in 1965 (Glaser and Strauss 1967). While grounded theory was discovered by two sociologists, they came from quite different backgrounds. Barney G. Glaser was trained in Columbia University’s Department of Sociology and influenced by “Merton’s middle-range theory and Lazarsfeld’s quantitative methodology” (Glaser and Strauss 1967, vii). Anselm L. Strauss (1916–1996) obtained his BA and PhD at the University of Chicago, studying under Herbert Blume and was influenced by the “Chicago tradition” in qualitative research (Glaser and Strauss 1967, vii). Their collaboration was the result of perceived inadequacies in the approaches of both in their respective backgrounds.

Grounded theory has diversified over the years as it gained popularity in nursing, anthropology, information technology, geography, and many other disciplines. After their publication of The Discovery of Grounded Theory: Strategies for Qualitative Research (1967), Barney G. Glaser and Anselm L. Strauss pursued separate scholarly endeavors and grounded theory diverged into two schools. This split between Glaser and Strauss became apparent when Strauss published Qualitative Analysis for Social Scientists (1987) and coauthored The Basics of Qualitative Research with Juliet Corbin in 1990, which contain detailed steps for the grounded theory process. Strauss and Corbin’s expanded coding approach included open coding, axial coding, selective coding, and
coding for process. Open coding is first level, line by line coding to identify concepts and their properties. This includes in vivo codes taken from actual wording of the respondents. Axial coding is a second level form of coding that finds relationships and common dimensions to create linkages between categories, subcategories, and their properties. Selective coding integrates categories and their properties into interrelated, theoretically saturated central explanatory concepts in order to build theory. Coding for process is a way of relating the interplay between structure and process (Strauss and Corbin 1998).

Strauss and Corbin’s approach sometimes referred to as the Strausarian School of grounded theory has been cited widely and used extensively by researchers in many disciplines. The first edition, published in 1990, was intended to supplement prior work on grounded theory and serve as a prescriptive guide for inexperienced researchers (Strauss and Corbin 1998). After Anselm Strauss died in 1996, Juliet Corbin published the third edition of Basics of Qualitative Research, coauthored by Strauss posthumously (2008). It is interesting to note that this edition no longer contains the detailed coding strategies of the previous versions. Additionally Corbin reflexively addresses some of the critiques of grounded theory’s claims of researcher objectivity and the idea that theory exists and can be discovered. Corbin acknowledges the importance of reflexivity and her own positionality and biases in this volume. Noting the changes over time in her own approach to research, she identifies with the constructivist approach as well as feminist self-reflection and commitment to social justice (Corbin and Strauss 2008). The third edition also notes that while it is intended for inexperienced researchers, it “can be used by anyone regardless of whether their research aim is theory building; rich, thick descriptions; or case study analysis” (Corbin and Strauss 2008, 17).

The approach to grounded theory developed by Strauss and Corbin was the subject of criticism by Barney Glaser (1992) in the Basics of Grounded Theory Analysis: Emergence vs Forcing where he provided a correction of various subsections and book chapters of what he perceived as the misconceptions and inadequacies of The Discovery of Grounded Theory. Glaser was particularly critical of the detailed coding structure of Strauss and Corbin’s work that he suggested forced data rather than allowing theory to emerge. Barney Glaser stated numerous times that he felt that Strauss and Corbin’s work produces “a forced, preconceived, full conceptual description” that is descriptive rather than facilitating the discovery of theory (Glaser 1992, 3).

Barney Glaser has remained committed to the original tenets of grounded theory from the collaboration with Anselm Strauss. In 1978 he published Theoretical Sensitivity which further detailed the method. Much of his subsequent writing has been critical of Strauss and Corbins’s prescriptive approach which he contends forces data into preconceived categories rather than allowing the categories to emerge through the comparative approach. He continues to publish his work mainly through Sociology Press, which is referred to as “the foremost publishers of classic grounded theory methodology” (www.sociologypress.com) and other venues, including the open access journal The Grounded Theory Review (http://groundedtheoryreview.com) that are dedicated to furthering research using the Glassarian approach to grounded theory.

Grounded theory has also been adapted by other researchers. Kathy Charmaz developed a constructivist approach to grounded theory (2006). Starting from Glaser and Strauss’s invitation extending the flexible use of grounded theory to other users, Charmaz focuses on flexible guidelines for the principles and practices.
of grounded theory (Charmaz 2006). While Glaser and Strauss emphasize the discovery of theory from data that is separate from the researcher, referred to as objectivist grounded theory, Charmaz assumes an interpretive role for the researcher who constructs grounded theories that are admittedly subject to the researchers, past experiences, interactions, and perspectives (2006). While Charmaz’s constructivist approach to grounded theory acknowledges the social construction of data and analysis and the role of subjectivity in theorizing, it also adheres to many of the methods of earlier versions of grounded theory, including simultaneous data collection, coding, comparative analysis, and memo writing in order to construct theory grounded in empirical reality (2006).

Grounded theory in geography

Geography, with its focus on place, the interplay of human and physical processes, identity, regions, and visual representation, has a history of using mixed-methods approaches that includes grounded theory in combination with other methods, both quantitative and qualitative. While the discovery of grounded theory by Glaser and Strauss was in opposition to what both perceived as inadequacies in their own epistemological grounding, in the 1990s grounded theory was associated with positivistic assumptions by many sociologists who questioned the claims of researcher objectivity and that theory existed and was discovered by the researcher. Geography experienced its own quantitative revolution, followed by the cultural turn that questioned positivist epistemologies and the emphasis of spatial science. However, in the 1990s, geographers’ emphasis on mixed-methods research uncoupled epistemology from methodology and suggested that rather than having the choice of method dictated by the epistemology, mixed-methods research could be informed by a number of different epistemological stances. For geographers, mixed methods allow for diverse research techniques that cannot only triangulate or confirm results, but may also contradict findings, confirming that knowledge is situated, partial, and contextual. Feminist geographers, in particular, have used both qualitative and quantitative methods and data to integrate multiple forms of knowledge and that allows an understanding of both the general and the particular.

The mixed-methods approach in geography has been extended to include the use of GIS technologies. Recent developments in critical and feminist geography led to the emergence of new areas of study that include critical GIS, feminist GIS, and qualitative GIS that frequently combine grounded theory with other methods in a mixed-methods approach. Rather than viewing GIS as a positivist tool, research by critical and feminist geographers has sought to decouple GIS from its association with quantitative methods and claims of positivism and integrate it into studies from a variety of epistemological stances. For example, Marianna Pavlovskaya used grounded theory to analyze interviews and ethnographic data along with GIS maps to understand patterns of urban change in a study of post-Soviet Moscow (Pavlovskaya 2002). This mixed-methods approach facilitated the discovery of categories, patterns, and processes that would not have been detected using only one method or the other. Another example of the combined use of grounded theory and GIS in flexible, recursive rounds of data collection and analysis is the study of community gardens in Buffalo, NY. LaDona Knigge and Meghan Cope developed an iterative, reflexive method of analysis that combines the use of grounded theory with GIS, exploratory data analysis and visualization techniques in an approach that
they call grounded visualization (Knigge and Cope 2006). This approach iteratively and recursively integrates multiple forms of data (both qualitative and quantitative) and analysis in order to build theory that is grounded in empirical data.

Several other examples of integrative approaches using grounded theory and GIS can be found in Qualitative GIS: A Mixed Methods Approach (Cope and Elwood 2009) including further work by Knigge and Cope, Pavlovskaya, and others. In this edited book, Jin-Kyu Jung (2009) develops a novel approach to qualitative GIS that combines the coding strategies of grounded theory operating within a computer-aided qualitative data analysis software (CAQ-DAS) platform that is integrated with GIS. Jung creates an imagined grid that contains qualitative data (photos) within GIS with a hybrid relational database to link GIS and CAQ-DAS into what he refers to as a computer-aided qualitative GIS (CAQ-GIS). These examples demonstrate how grounded theory can be integrated into mixed-methods approaches by geographers and the flexibility and adaptability of the grounded theory method which will no doubt continue to be developed within geography and other disciplines in the future.

SEE ALSO: Critical GIScience; Ethnography; Feminist methodologies; Interviews; Mixed-method approaches; Positionality; Qualitative data; Qualitative GIS; Reflexivity; Visualization

References


Groundwater is water that exists mainly in subsurface pore spaces but also in defined channels, such as those found in karst formations, which are created by dissolution of soluble rocks such as limestone. After the polar ice caps, groundwater is the next largest reservoir of freshwater on Earth, containing more than 100 times the volume of streams and freshwater lakes (Shiklomanov 1993). Groundwater plays an important role in the hydrologic cycle, in plant growth and soil formation, and in providing water for human activities. Since groundwater is difficult to observe and track directly, monitoring, mapping, and modeling efforts are crucial to the understanding of its storage, distribution, and patterns in flow through the subsurface, how it may be sustainably used, and how it may become contaminated. Humans have used groundwater throughout history, but the demand for it and the societal ability to consume and contaminate it have all increased exponentially over the past century. Uneven distribution of human populations combined with variations in availability and accessibility to groundwater have resulted in overexploitation of several major aquifers serving population and agricultural production centers in China, India, and the United States. Significant current and future threats to groundwater quantity and quality that require attention from policy and planning perspectives include climate change, fossil fuel exploration and extraction, natural contaminants such as radium and arsenic, underground storage tanks and industrial contaminants, and agricultural use and pollution.

Groundwater: natural and physical science perspectives

Groundwater’s place in the water cycle

More than 68% of fresh water is found in polar ice caps and glaciers and is thus largely unavailable for societal use (Shiklomanov 1993). Approximately 30% of global fresh water is groundwater, whereas only 1.2% is found in streams and lakes. Geologic formations that yield a significant amount of water to wells or springs are called aquifers. An aquifer consists of two or more permeable layers in the subsurface separated at least locally by intervening layers that impede groundwater movement but do not significantly affect the regional hydraulic connectivity of the system. With the exception of fossil aquifers, groundwater is replenished mainly by precipitation falling within the recharge area of an aquifer. Precipitation on land flows horizontally across the surface or infiltrates into soil and moves either horizontally or vertically within the subsurface. Groundwater consists both of water that remains in the unsaturated or vadose zone (also often termed “soil water”) and
of water that reaches the saturated zone (aquifer) where pore spaces are completely filled.

Global distribution of groundwater

Groundwater may be found almost anywhere on Earth if one digs deep enough, but most accessible groundwater is generally found within 1 km of Earth’s surface (Hess 2014). Beyond this depth, groundwater availability decreases gradually and quality is often poor due to high salinity and mineral concentrations. Worldwide, approximately 36% of the land surface (excluding Antarctica) is underlain by major aquifers, primarily composed of sedimentary rocks, while another 18% of the subsurface is dominated by aquifers of more complex and heterogeneous geology (Richts, Struckmeier, and Zaepke 2011, WHYMAP, Table 1). The remaining approximately 47% of the land surface is underlain by local and shallow aquifers along stream valleys and lowlands. There are significant regional variations in the distribution of aquifers. For example, shallow local aquifers dominate much of North America, whereas major regional aquifers spanning thousands of square kilometers are the predominant aquifer type throughout much of continental Europe (Table 1).

Currently, approximately 24% of global aquifers are highly overexploited; the major driving factor is irrigated agriculture for feeding human and livestock populations (Gleeson et al. 2012). Pumping from the most exploited aquifers (Upper Ganges, North and South Arabian, Persian, Western Mexico, and High Plains or Ogallala) is on the order of tens of times the natural recharge rate, and may also act as a limiting factor in future agricultural output (Figure 1). Although 76% of aquifers are not being significantly exploited, overexploitation of a few large aquifers has led to a total groundwater footprint of more than 3.4 times the global area of aquifers (Gleeson et al. 2012, Table 2). If 1 and 2% of the largest pristine or unexploited aquifers (7 and 15 observations from 748 considered in Table 2) are removed from consideration, the composite average of the groundwater footprint rises to 4.81 and 6.34 (39 and 84% increase, respectively) indicating that a few outlying pristine aquifers have undue influence on the composite average of the global groundwater footprint. Similarly, if 1 and 2% of

<table>
<thead>
<tr>
<th>Continent</th>
<th>Major groundwater basins (million km²)</th>
<th>%</th>
<th>Complex hydrogeological structures (million km²)</th>
<th>%</th>
<th>Local and shallow aquifers (million km²)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>13.5</td>
<td>44.9</td>
<td>3.3</td>
<td>11.0</td>
<td>13.2</td>
<td>44.1</td>
</tr>
<tr>
<td>Asia</td>
<td>14.5</td>
<td>32.0</td>
<td>7.8</td>
<td>17.3</td>
<td>23.0</td>
<td>50.7</td>
</tr>
<tr>
<td>Australia, New Zealand</td>
<td>2.6</td>
<td>32.5</td>
<td>2.9</td>
<td>36.3</td>
<td>2.5</td>
<td>31.1</td>
</tr>
<tr>
<td>Europe</td>
<td>5.2</td>
<td>53.0</td>
<td>1.8</td>
<td>18.8</td>
<td>2.7</td>
<td>28.2</td>
</tr>
<tr>
<td>Central/South America</td>
<td>8.4</td>
<td>45.0</td>
<td>2.0</td>
<td>10.9</td>
<td>8.2</td>
<td>44.1</td>
</tr>
<tr>
<td>North America</td>
<td>3.2</td>
<td>15.0</td>
<td>5.8</td>
<td>26.9</td>
<td>12.4</td>
<td>58.1</td>
</tr>
<tr>
<td>World (excl. Antarctica)</td>
<td>47.3</td>
<td>35.6</td>
<td>23.6</td>
<td>17.8</td>
<td>62.0</td>
<td>46.6</td>
</tr>
</tbody>
</table>

Source: Richts, Struckmeier, and Zaepke 2011. Date of the source data is 2008. With permission from WHYMAP, BGR, Stilleweg 2, 30655 Hannover, Germany.
the aquifers representing the largest footprints (7 and 15 observations from 748 considered in Table 2) are removed from consideration, the composite average of the groundwater footprint falls to 1.52 and 0.93 (56 and 73% decrease, respectively) indicating that a few outlying aquifers with large footprints also have undue influence on the composite average of the global groundwater footprint. In contrast, the median of the global groundwater footprint of 0.11 is robust and is not significantly impacted by the extreme outlying values of aquifer areas or groundwater footprints.

Exploitation of aquifers also varies significantly within aquifers. For example, the average water level in the High Plains (or Ogallala) aquifer of the central United States has declined by more than 4 m since before World War II. However, this includes some areas where water levels have increased by up to 20 m and others that have decreased more than 70 m (McGuire 2013). Total water storage in the High Plains aquifer has decreased by approximately 8%, but storage in 5% of the aquifer area decreased by over 50% (McGuire 2013).

Types of aquifers
The unsaturated zone of an aquifer is immediately below the land surface where the soil/rock pores contain both water and air, but are not totally saturated with water. It differs from the zone below which is saturated with water. The top of the saturated zone is referred to as the water table (USGS 2014a). Aquifers may be classified by whether they are confined or
unconfined and by their geologic composition. Confined aquifers exist between two impermeable layers, often comprised of clay or clay-derived rock. Unconfined aquifers lack a confining layer above the water table, and are thus vulnerable to pollution from the surface. Unconsolidated aquifers consist of sand, gravel, and other materials that have not been cemented together, and where water fills up spaces between the particles. In contrast, water-bearing formations that have been cemented together are termed consolidated aquifers. Relatively large pore spaces in unconsolidated aquifers and solution channels in carbonate aquifers can hold significant amounts of groundwater with high hydraulic conductivity. In contrast, only the presence of fractures and joints in igneous and metamorphic rocks allow water to enter and move through the formations. Sand and gravel aquifers are generally unconfined, and can be several hundred feet thick. They are comprised of alluvial or glacial deposits and found in valleys, depressions, and lowlands. Alluvial aquifers were formed over hundreds to thousands of years by water deposition of sand, silt, and clay particles. These alluvial deposits can contain large groundwater reserves that are fairly accessible and easy to withdraw. Sand and gravel aquifers of glacial origin may also have large reserves of groundwater that are fairly easy to withdraw, but hydraulic conductivity in such aquifers may be quite variable due to lack of sorting in glacial till.
Groundwater

Carbonate rock aquifers are comprised primarily of limestone or dolostone formed in ancient marine environments. Weak carbonic acid in rainwater dissolves carbonate rock and over thousands of years has in many areas formed karst landscapes characterized by numerous solution cavities, caverns, and sinkholes through which water can move rapidly. Where carbonate layers are exposed to the surface, runoff from precipitation or entire streams may connect directly with groundwater, resulting in significant contamination potential. Sandstone aquifers and igneous and metamorphic rock aquifers typically store and transmit water only along bedding planes, joints, and other cracks and fractures. Hydraulic conductivity in such aquifers is low, but large regional aquifers can yield high total volumes of groundwater.

Groundwater flow and wells

The geology of the aquifer influences water storage capacity as well as hydraulic conductivity, or the ability of aquifer materials to transmit water. Factors affecting hydraulic conductivity and storage capacity include porosity (proportion of pore space) and permeability (connectedness) of the pore spaces. Groundwater flow is down-gradient from high to low pressure, which often corresponds with moving from high to low elevation. An exception to this occurs in the case of confined aquifers contained within tilted rock formations where considerable pressure builds up in the lower reaches of the aquifer. In such cases, a well or a spring located at a point where the potentiometric surface (or water table) is higher than the land surface will flow freely without the need for pumping; such wells or springs are termed artesian.

Wells drilled into unconfined aquifers or confined aquifers with insufficient pressure to flow naturally require pumping to bring water to the surface. Some level of drawdown (or reduction in the height of the potentiometric surface) occurs with pumping of water from aquifers. During pumping, a cone of depression (Figure 2) forms around wells – the shape, size, and depth of the cone of depression as well as the rate at which the well water level recovers depend upon the pumping rate and duration, aquifer characteristics such as hydraulic conductivity, and the presence or absence of nearby wells. Pumping tests are done to determine whether the aquifer will provide an adequate yield (measured in gallons or liters per minute) for the intended purpose of the well.

Soil formation

As it moves through soil, water dissolves chemicals and sometimes deposits them at lower levels. This leaching process can deplete the topsoil of essential nutrients (Hess 2014). Fine particles such as clay are picked up by water and carried in suspension and deposited elsewhere at greater depths. Water, thus, constantly changes the physical, chemical, and biological makeup of the soil as it moves downward due to gravity, or sideways depending on available openings. When silica is removed by dissolution leaving behind...
the more insoluble iron compounds, reddish colored soils such as latosols are produced in tropical and subtropical regions of the world. In mid- and high latitudes, water removes minerals like oxides of aluminum and clays and forms shallow acidic soils like podzols. Gley soils are acidic and oxygen-poor and develop in waterlogged conditions in cool climates. In the more arid and semiarid regions, capillary action brings up soil moisture, which evaporates leaving behind salts like chlorides and sulfates. Soil salinization is exacerbated by inadequate drainage and low precipitation rates, resulting in calcic and other salt hardpans.

Groundwater chemistry

Groundwater quality may be affected by contaminants of both natural and human origin. Factors that determine groundwater quality include the thickness of the aquifer, aquifer structure, mineral composition of the aquifer, presence or absence of a confining layer, presence of direct conduits (e.g., abandoned wells or sinkholes) to the aquifer, and in the case of unconfined aquifers, presence of pollutants on the overlying land surface. Unconfined aquifers are more susceptible to contamination from human activities because of their proximity to pollution sources. The age or residence time of water in the aquifer is also a significant factor affecting groundwater quality as longer residence times provide more opportunity for dissolving minerals.

Many substances may be dissolved by groundwater as it moves through the subsurface. The most common chemical constituents in groundwater include sodium, calcium, magnesium, potassium, chloride, and sulfate. However, depending on mineral composition of the subsurface and on length of exposure, high levels of arsenic, fluoride, iron, manganese, or radionuclides may also be found.

Modeling and visualizing groundwater

The importance of modeling groundwater

Because of its intractability, understanding the processes associated with groundwater is more complicated relative to those associated with surface water. Hence, monitoring, mapping, and modeling efforts are crucial for this purpose. As in any area of inquiry, however, the applicability and accuracy of groundwater models is highly dependent upon (i) the quality of the input data (i.e., knowledge of initial conditions and the existence of adequate hydrogeological data) and (ii) the “fit” between the modeling approach (assumptions and equations) and the scope and purpose of the problem being explored.

Many groundwater models in one, two, or three dimensions rely at least to some degree upon Darcy’s Law (Henry Philibert Gaspard Darcy, 1803–1858), which among other things allows for estimating water velocity in an aquifer as well as time required for water to travel between points within the aquifer, and which holds for nearly all hydrogeological conditions. In brief, Darcy’s Law provides that water flows only when there is a gradient, that water flows down-gradient from high to low head (pressure), that flow velocity is proportional to head loss, and that velocity is modified by the hydraulic conductivity of the aquifer materials.

Modeling approaches and data needs

There are three types of groundwater models: analytical element, finite element, and finite difference. The latter two are data driven and are calibrated or fitted with extensive field data. The simplicity of the analytical element method is in
its assumptions. It uses points, lines, and polygons to spatially represent contaminant sources, rivers, flow barriers, and groundwater zones. Assumptions such as flat water tables, uniform aquifer thickness, and constant conductivity are typical in such modeling efforts. However, such models may not accurately represent the complex geometry and heterogeneity of the real world of subsurface groundwater (Bennett 2014).

Depending on the scope and purpose of modeling efforts, data of varying quality at compatible scales may be required. Generally, the data required includes hydrologic measurements (e.g., precipitation, infiltration, evapotranspiration), initial and boundary conditions (e.g., potentiometric surface height, presence of confining layers), subsurface characteristics (e.g., thickness, composition, hydraulic conductivity), and human conditions (e.g., presence and pumping rate of wells, surface modifications). There is much more detailed hydrologic and subsurface data available in developed nations than in developing nations. For example, in the United States, vast amounts of hydrologic data may be obtained from federal government agencies such as the US Geological Survey and the National Oceanic and Atmospheric Administration and most states require submission of well boring logs to permitting authorities for every well that is drilled, resulting in a slowly increasing store of knowledge of the subsurface. In recent decades, satellite technology has also been employed to search for signs of groundwater (e.g., landforms such as ancient lake beds) or overexploitation of groundwater (e.g., changes in surface elevation) over time.

Software use in groundwater modeling

Many commercial and open-source software packages exist for purposes of modeling and displaying various aspects of groundwater quantity and quality, including flow, transport, and geochemical reactions. A selected list of open-source software for groundwater models is provided by the US Geological Survey (USGS 2014b).

MODFLOW is the USGS’s three-dimensional (3-D) finite-difference groundwater model (USGS 2013) that is considered an international standard for simulating and predicting groundwater conditions and the interactions that take place between ground and surface water. Originally developed and released in 1984 as a computer code for simulating groundwater flow, its modular structure provides a robust framework for integrating additional features. Current capabilities include simulations of coupled groundwater and surface-water systems, solute transport, variable-density flow (including saltwater), aquifer-system compaction and land subsidence, parameter estimation, and groundwater management (USGS 2013).

To enable users to make appropriate software selections, Kumar (2012) has provided a review of the capabilities and limitations of available groundwater modeling software. 3-D groundwater computer software uses borehole stratigraphy, geologic layers, cross sections, and terrain data (for example, digital elevation models) to construct 3-D models of the subsurface (Strassberg, Maidment, and Jones 2014). With its ability to provide preprocessing data input to MODFLOW, and store results from MODFLOW, the Arc Hydro Groundwater data model of the flagship software company ESRI of California (Strassberg, Maidment, and Jones 2014) has created a very rich and comprehensive environment for groundwater modeling and visualization.
Groundwater models and geovisualization, selected examples

Hydrogeologists often use diagrammatic representations (Figure 3) to characterize and visualize the subsurface flow systems and the influence of geology and climate on such systems over time. The movement of groundwater in the subsurface of the Death Valley Regional Flow System (DVRFS) is captured and depicted in Figure 3 (Faunt, D’Agnese, and O’Brien 2010). It shows the flow paths of groundwater movement under the influence of hydraulic gradient through varying zones of permeability, from recharge to discharge areas in a regional context.

In addition to testing local aquifers to determine well or aquifer yields, and using well data to describe subsurface flow patterns, models are also used to understand groundwater both at spatial (local to regional aquifers) and temporal (e.g., tracing historical contamination to its source) scales. Two examples of 3-D groundwater modeling efforts are provided below and include predicting the migration of a uranium plume (Figure 4) and assessing risk of encountering high arsenic concentrations at various depths (Figure 5).

To construct 3-D images for visualizing concentrated plumes of contaminants such as uranium in subsurface sediments of waste sites, as shown in Figure 4, scientists at the US Department of Energy (2009) have developed groundwater models. Such models enable the prediction of movement of contaminants such as uranium over tens to hundreds of meters on temporal scales ranging from hours to years.

Arsenic contamination of shallow groundwater and its impact on the health of affected populations is one of the most serious environmental...
problems in developing countries. In Figure 5, the authors of a study (Winkel et al. 2011) on groundwater contamination of the Red River Delta in Vietnam show the probability of Arsenic (As) concentration exceeding 10 μg/l at 10 m depth intervals. Such analyses resulting from models enable us to study the connections between large-scale pumping of groundwater and the resulting vertical migration of arsenic in the subsurface.

Social, political, and planning aspects of groundwater

Historical access to and use of groundwater

Throughout history, humans have accessed groundwater in several different ways, likely beginning with locations where seeps or artesian upwellings occurred, progressing to hand-dug wells and wells accessed via windlass and bucket,
and finally tube wells with electric or fossil fuel-powered pumps.

Qanats, or hand-dug tunnels with numerous vertical shafts that access groundwater and deliver water by gravity for irrigation or domestic use, have been in use for more than eleven centuries throughout much of the Middle East, western Asia, southern Europe, northern Africa, and Central America. This technology is believed to have originated in Persia and spread via trade routes and expansion of Arab and Roman empires. Water from qanats was historically viewed as a community resource to be shared, and the maintenance and cleaning of qanats was also a laborious and challenging exercise. Thus the use of qanats for accessing water in effect demanded a cohesive community to successfully maintain channels. Flow in qanats varies with the level of the water table in the aquifer, which ensured sustainable use of the aquifer but also may have limited use for crops (and therefore crop yields) or other uses during periods of low recharge. Although qanats are still in use in many regions, their importance for domestic supply and irrigation has decreased due to increasing use of the same aquifers by powered pumps or other hydrological modifications that may have reduced recharge (e.g., as in the case of Morocco’s Tafilalet oasis and aquifer system).

Rapidly growing urban populations with increasing water use and consumption during the twentieth and twenty-first centuries have
stressed aquifers with relatively high withdrawal rates. In the case of coastal megalopolises such as Houston, New York, and Chennai, the increasing water use rates have outstripped recharge rates, thus enabling saltwater intrusion into coastal aquifers. Human population and agricultural centers have not all been developed in close proximity to locations with ample surface or groundwater resources. Major river systems or areas of significant precipitation often give rise to the highest yielding alluvial freshwater aquifers. Large human settlements in arid and semiarid regions and areas overlying saline groundwater have long been water stressed.

Groundwater rights

Modern groundwater rights are most often determined by ownership of overlying land, or, where legally permissible, by leasing or purchasing the rights to groundwater residing beneath lands owned by others. The pumping or abstraction rate may be determined by principles such as absolute ownership (pumping without regard to depletion or conflict with others), correlative rights (considering the rights of other users), or a permit that specifies a maximum pumping rate. Historically, groundwater pumping rates have not been of significant concern, except in water scarce regions.

In recent decades, with increasing groundwater consumption for irrigation and other purposes, and with more complete information about aquifer extents, capacities, and depletion rates, pumping has become increasingly regulated in semiarid regions. In some situations (e.g., Texas, United States), groundwater rights may be separated from the overlying lands and may be purchased or leased and exploited by individuals or entities who do not own the land. Groundwater rights in some locations are also affected by the distinction between percolating groundwater (e.g., as in a sandstone aquifer) and groundwater flowing as a subterranean stream in a defined channel (e.g., as in a karst formation).

Groundwater, conflict, and transboundary aquifers

The destruction, deliberate draining, or poisoning of wells has been a feature of many conflicts throughout history; instances include Assyria’s use of rye ergot to poison enemy wells in the sixth century BCE, Saladin’s sanding of wells in the defeat of the Crusaders in the twelfth century CE, poisoning of wells in German Southwest Africa by retreating German troops in 1915, Japan’s use of biological weapons to contaminate wells in China in 1939–1942, Serbian cut-off of well water to Bosnians in Sarajevo in 1992, Botswana troops’ destruction of Khoisan (Bushmen) wells in 2002, and intentional contamination of Darfur wells in 2004 (Gleick et al. 2012).

There are numerous transboundary aquifers worldwide, some of which are shared by more than two nations. The nature of aquifer sharing includes several different types of situations, for example:

- two or more nations overlie an unconfined aquifer;
- two or more nations overlie a fossil aquifer with no meaningful recharge;
- one nation has the recharge zone and part of an unconfined aquifer within its territory, whereas down-basin nations have access only to the water in the aquifer (Eckstein and Eckstein 2005).

With some exceptions, and despite the catastrophic potential impacts of aquifer contamination and depletion, there has been relatively little progress on the international management of groundwater. In 2008, the United Nations adopted Resolution A/RES/63/124 on the Law
of Transboundary Aquifers, which provides a set of guidelines and recommendations to assist nations in establishing bilateral and multilateral agreements. More recently, the UN has discussed establishing a more formalized convention on transboundary aquifers. Key issues in the management of transboundary aquifers include prior use patterns, military conflict, economic growth, population growth, irrigated agriculture in arid or semiarid lands, land use, climate change, and the proportion of overlying land or recharge area controlled.

There are several notable regional conflicts in which groundwater plays an important role. Groundwater is interwoven into the Israeli–Palestine conflict because the Mountain Aquifer provides a significant proportion of the domestic and agricultural water to Israeli territory, but the recharge zone lies nearly entirely within the West Bank. The lack of stable and durable peace in the region makes it difficult to manage and use the resource efficiently and also construct and operate water and wastewater treatment facilities in the West Bank so as to protect the highly vulnerable aquifer.

The Nubian sandstone aquifer is a fossil aquifer in northern Africa shared by Libya, Egypt, Sudan, and Chad. In 2013, these four nations adopted the Regional Strategic Action Plan on the Nubian Sandstone Aquifer in response to declining water levels, increased salinization in the Post-Nubian System (north of the Nubian System), and the declining health of dependent ecosystems.

Key issues related to planning and future sustainability of groundwater resources

Water quantity: climate change

Climate change is already impacting groundwater and will continue to impact it in at least two important ways: (i) groundwater may become a more valuable resource in a changing climate, where precipitation becomes less predictable and more “flashy,” and (ii) aquifer recharge may also be altered by changes in precipitation patterns and evaporation rates. Climate change impacts on groundwater resources may be most significant in areas that have scarce precipitation or which already are water stressed, as well as in monsoon regions.

Water quantity: over abstraction

In addition to the 15 highly overexploited aquifers of the world (shown in Table 2), over-abstraction or mining of groundwater for agriculture (e.g., San Joaquin Valley, California, United States), industrial use, and drinking water supply (e.g., Osaka City, Japan) has occurred in numerous other aquifers worldwide. In many locations, such overuse has caused permanent compaction and aquifer capacity reduction and accelerated sinkhole formation in karst regions. Resulting land subsidence of several tens of meters in some cases (e.g., Houston, Texas, United States) have increased the vulnerability of such areas to flooding. Due to approximately 8 m of subsidence in the last century, numerous buildings in Mexico City have been damaged and significant repairs have been necessary to much of the subsurface infrastructure, including subways, sewers, and potable water pipes.

Groundwater quality issues

Arsenic (arsenicosis), fluoride (fluorosis), and radionuclides (mutations, cancer) are all found in groundwater in some regions and have significant human health impacts. For example, relatively low concentrations of arsenic can be found in groundwater in many regions; however, the highest concentrations consumed as
potable water appear to be found in the southern Asian nations of Bangladesh, India, and Vietnam (Mukherjee et al. 2006; Winkel et al. 2011). Other groundwater constituents tend to have impacts that are less health-related but significant nonetheless. For example, the main impacts of calcium and magnesium are the formation of scaly deposits on pipes and reduced performance of soaps and detergents. Sulfur compounds typically affect the smell and taste of water, whereas high levels of iron can leave reddish stains on fabrics washed with groundwater.

Groundwater contaminants of human origin include industrial discharge, landfills and legacy waste disposal sites, agriculture, toxic chemical spills, leaking underground storage tanks (LUSTs), and domestic sewage discharge. The behavior and toxicity of contaminants varies widely, as does the ability to remediate and restore groundwater quality. Nitrate, the most common agricultural groundwater contaminant, is highly soluble in water, but its presence in affected aquifers may be significantly reduced by source reduction or prevention. Other contaminants that are less soluble and less mobile may be extremely difficult and expensive to remove from aquifers. Dense nonaqueous phase liquids (DNAPLs) such as trichloroethylene (TCE) tend to sink toward the bottom of aquifers where they may slowly degrade and release toxic by-products. Gasoline from leaking underground storage tanks is perhaps the most common light nonaqueous phase liquid (LNAPL). Benzene, toluene, ethylbenzene, and xylene (BTEX, four of the major hydrocarbons found in petroleum products) spread in plumes at different rates throughout the top of the aquifer. Microbial contaminants may be introduced to aquifers by improperly sited septic tanks as well as by untreated or inadequately treated municipal discharges above unconfined surficial aquifers.

Significant and growing threats to groundwater: agriculture

Agricultural productivity has increased significantly over the past several decades. Yields have risen due to increased use of pesticides, fertilizers, and irrigation (often from groundwater sources), as well as to expansion of cultivated areas. But gains in productivity have been accompanied by increasing groundwater degradation and depletion. For example, herbicides such as atrazine and alachlor (used respectively on cereals and fruit trees) have been extensively detected in groundwater in Europe and North America; and in South Asia, commonly used insecticides including carbofuran and lindane have been widely detected in groundwater (Mateo-Sagasta and Burke 2011). Elevated nitrate levels, sometimes exceeding the US Environmental Protection Agency (EPA) maximum contaminant level of 10 mg/L are nearly ubiquitous in unconfined shallow aquifers in major breadbaskets such as the Midwestern United States.

Significant and growing threats to groundwater: fossil fuel exploration and extraction

Hydraulic fracturing or “fracking” is a drilling technique used to obtain natural gas and oil from shale rock and other nonconventional sources (Bigham 2013). For every fracking well, between 8 and 20 million liters of water along with sand and hundreds of compounds (e.g., for lubricating, anticorrosion) are injected to fracture deep tight rock formations, releasing gas and oil from small pores within them (Figure 6). Groundwater contamination associated with fracking operations appears to be associated mainly with (i) the disposal of “produced” water and infiltration into shallow unconfined aquifers, and (ii) well barrier and integrity failures, which
Fracking for fuel
Hydraulic fracturing is used to access oil and gas resources that are locked in nonporous rocks.

Water recovery tanks
Polluted flowback water may be injected into a deep storage well, recycled, or sent to a treatment plant.

Leakage of fracking fluid from the fracture zone is highly unlikely.

Figure 6  Hydraulic fracturing. Source: Howarth, Ingraffea, and Engelder 2011. Reproduced by permission of Nature.
may allow migration of methane and fracking fluids to aquifers used for drinking and irrigation. Davies et al. (2014) indicate well failure rates are highly variable (1.9–75%) depending upon when and where they have been drilled. Other recent studies indicate that groundwater in close proximity to fracking operations is much more likely to be contaminated with produced water and methane, and that in some cases the chemical composition of methane found in drinking water aquifers was quite similar to the gas found in the much lower shale formations (Jackson et al. 2013; Gordalla, Ewers, and Frimmel 2013).

To facilitate rapid exploration and recovery of crude oil and natural gas, changes have been made to existing legislation such as the United States’ Safe Drinking Water Act to exclude hydraulic fracturing from the definition of underground injection, a practice that was until the early 2000s largely prohibited or tightly regulated. Within the United States, fossil fuel exploration is mainly regulated at the state level, but the US EPA has recently undertaken a nationwide study on fracking and may issue nationwide regulations in the future. Hydraulic fracturing is already being employed or planned in Germany, the United Kingdom, India, and China (Gleick et al. 2012), and numerous suitable formations for exploration and extraction of resources via hydraulic fracturing exist on every populated continent.

Strategies for protecting and replenishing aquifers

At least in part due to difficulties in directly observing groundwater, efforts to conserve and protect the quality of groundwater resources have lagged behind such measures for surface water. In addition to international cases noted above, this is also true at national scales. For example, whereas the Clean Water Act is aimed at protecting surface water, there are no federal environmental laws focused on groundwater in the United States. Much of the regulation of groundwater is then left to individual states. But since aquifer and state boundaries rarely coincide, this leads or has the potential to lead to problems similar to those faced with international aquifers. One major policy need, therefore, is to “scale up” policy related to protecting and conserving aquifers.

Numerous strategies have been proposed for situations in which aquifers are stressed due to over abstraction from irrigated agriculture, including substituting dryland-adapted crops (e.g., wheat, sorghum, millet) in place of rice or maize, developing drought-tolerant strains of rice or maize, reducing tillage, reestablishing native grasslands, limiting drilling of new wells, and using moisture sensors and ultra-efficient irrigation technologies. In urban areas where aquifers are depleted, strategies such as aquifer recharge via infiltration ponds or galleries and injection of surface water or treated municipal wastewater are being increasingly employed.

SEE ALSO: Aquifers; Environmental uncertainty; Hydrologic cycle; Soil water; Waste and waste management; Water conflicts; Water: drinking; Water and human rights; Water quality; Water rights

References


Further reading


The concept of the growth machine refers to systemic properties operating within capitalist cities which generate growth around the intensification of land use. This often happens despite pressures and political interests opposed to such growth. The term “growth machine” can also refer to coalitions of business and local government interests seeking to attract inward investment, promote redevelopment, and market the city. The growth machine concept was first proposed by the American urban sociologist Harvey Molotch in the 1970s (Molotch 1976). It was subsequently refined into a thesis about the tensions between use and exchange value arising from unregulated capitalist forms of urban growth (Logan and Molotch 1987). By connecting urban political agency to certain systematic properties of capitalism, the growth machine thesis was able to transcend some of the shortcomings of social scientific analysis of urban power structures in the USA.

Logan and Molotch (1987) suggested that the activities of rentiers – land development interests – lie at the crux of understanding the behavior of growth machines. Rentiers prepare the ground for land-use intensification and profit-making in the city. Often this is to the benefit of a wide range of business interests and organizations, such as corporations, contractors, the local media, universities, utilities, sports franchises, chambers of commerce, trades unions, and so forth. Local government is not a disinterested party in this process as it strives to influence land-use outcomes in ways that enhance the local tax base and foster a sense of community solidarity. Growth coalitions tend to converge around those levels of government having fiscal resources and land-use powers necessary for delivering growth and jobs.

For urban geographers, the growth machine thesis has provided fertile territory for critique and further refinement (Jonas and Wilson 1999). Some critiques have highlighted how its focus on the tension between use and exchange value subverts knowledge of class tensions and progressive urban politics. Others examine the discursive tools deployed by growth machines and their effects on marginalizing the voices of minorities in the city. Then there are those who have sought to apply the thesis to settings outside the USA, finding that the national state is a more powerful driver of urban development than local growth coalitions. However, the jury is still out in respect of whether the growth machine thesis is conducive to the international comparison of urban development politics.

Urban scholars today argue that knowledge of power structures operating in cities requires a shift in focus away from territorial growth coalitions and towards relational processes and networks operating beyond the city limits. Nevertheless, the growth machine remains an attractive thesis for critical urban geographers. National and local states in many countries are incentivizing local public–private partnerships in a manner which suggests that growth machine-like processes are compatible with neoliberal urban development. Moreover, the growth machine offers a seductive argument about how the infrastructural conditions of
GROWTH MACHINES

urban growth can be sustained even under conditions of fiscal austerity.

SEE ALSO: Urban elites; Urban politics; Urban regimes

References

The idea of growth poles

For a long time, economic development had been regarded as an expansion of a balanced system within which resources were mobilized and allocated proportionately in various sectors. Such an equilibrium understanding of economic growth is inconsistent with real development processes and rules out structural changes of economies. Starting from a different point, Perroux interpreted economies as being composed of units, firms, and industries with asymmetrical powers and relations. For economic or institutional reasons, some firms and industries would take a dominant position over other units in market transactions. These asymmetrical interrelations of firms and industries result from their big size, input/output structures, technical advances, or some institutional arrangements. For Perroux, what is crucial for economic development is that these dominating firms and industries can stimulate structural change for the whole economy (Perroux 1950a). Inspired by Schumpeter’s emphasis on innovation and the role of large corporations, Perroux argued that propulsive industries (industries motrices) with high rates of growth, advanced technologies, and innovation capabilities can drive the growth of an economic system through their intensive linkages with other industries (industries mues). From this, he developed the idea of growth poles and defended it convincingly as “growth does not appear everywhere at the same time; it becomes manifest at points or poles of growth, with variable intensity; it spreads through different channels, with variable terminal effects on the whole of the economy” (Perroux 1970/1955, 94).

An important point for interpreting Perroux’s idea of growth poles is that although not denying an understanding of growth poles in a geographical setting, Perroux originally defined the concept in an abstract economic space constructed by transactions and relations of different units and activities (Perroux 1950b). As original as the idea of growth poles was, it was initially quite elusively defined, far from the clear development strategy that was so popular later on. It is safe to say that the original concept of growth poles did not explicitly address the problem of regional development, but was merely a dynamic understanding of unbalanced economic growth. What is central to the idea is how new and propulsive firms and industries can induce the development of other units through material exchange and information flows.

The concept of growth poles was later transformed from an idea framed in an abstract space into an entity in a geographical setting through a process in which new elements were included in the concept (Boudeville 1966; Hansen 1967). From a spatial perspective, it may be natural to expect that propulsive firms and industries are located in some regions, which makes these areas develop at an extraordinary rate and
become growth centers of regional and national economies. But from Perroux’s idea of how propulsive units affect the growth of other units, it cannot be taken for granted that induced activities would occur in surrounding areas. In other words, an argument of regional development supported by development poles or centers has to be established on some further rationale besides Perroux’s original conceptualization. Another important question behind a spatial version of growth poles concerns why, if propulsive units are a set of firms or industries, rather than a single unit, these firms agglomerated in growth centers. From here, growth poles meet with location theories and urban and agglomeration economics with similar questions. In response to these questions, external economies, broadly interpreted as including availability of public and business services, development of specialized suppliers, attraction of consumers, labor market sharing, and exchange of ideas and technologies, were later given greater emphasis to back up growth poles in geographical settings, which may be titled differently as growth centers or development poles (Cameron 1970).

Growth poles in a spatial reading for regional development are further articulated by two prominent economists in Europe and North America. Also unsatisfied with an equilibrium understanding of economic growth, Myrdal (1957) and Hirschman (1958) developed similar ideas of unbalanced development processes, but in concrete regional contexts, which endowed the concept of growth poles with more practical meanings. Myrdal (1957) and Hirschman (1958) clarified and enriched the relations between growth poles and hinterland areas with “backwash” or “polarization” effects and “spread” or “trickling down” effects. Unlike Perroux, who viewed growth poles only in a positive way, Myrdal and Hirschman proposed that at early stages of development, core areas would develop at the expense of periphery areas by attracting talents, investment, and resources (“backwash” or “polarization” effects). Besides consuming skills and inputs from hinterlands, competition from industries in growth centers will restrict the development of similar activities in backward areas. However, on the other hand, the growth of poles or centers also provides positive influences for surrounding areas in relation to demand for goods from hinterlands, investment opportunities, and diffusion of technologies, together described as “spread” or “trickling down” effects. It is argued that positive effects would predominate in later stages of development since the concentration of resources and activities in growth poles will finally drive up production costs and bring about external diseconomies. In this period, regional disparity would be expected to decrease.

With a large body of literature in the 1960s and 1970s, growth poles were widely accepted as a general understanding of how regions grow in an interactive way and turned into a popular idea of economic development in policy arenas. An important reason for the adoption of the concept of growth poles by policymakers is that it describes a regional development process, which is not only flexible enough to be consistent with different policy goals in various settings but also quite specific, thereby helping governments direct resources and investment spatially.

**Growth poles as regional development strategies**

It is worth mentioning that the wide acceptance of growth pole strategies in many countries in the 1960s and 1970s should be interpreted in the historical context (Polenske 1988). For both developing and developed countries, after World War II, economic recovery and restructuring...
was a major concern for governments of various levels. It was a period when strong government intervention in economic activities was in high demand. Growth pole theory thus provided a theoretical reference point to justify many economic projects and plans in that period. As regional development strategy, the concept was implemented in different settings with different rationales targeting various goals. Within cities, large plants were built as propulsive units to stimulate urban economic development (Ryder 1990). In metropolitan contexts, growth poles were established in suburban areas to alleviate an over-concentration of population in large cities (King 1974). In some regional backgrounds, large industrial complexes based on heavy industries were planted to create employment in surrounding areas (Roberts 1995; Florio 1996). In some planned high-tech industrial complexes, such as Sophia Antipolis, several large firms were established and a pole of propulsion was expected to promote manufacturing growth in nearby regions (Hansen 1990). In broader regional contexts, some towns and cities were designated as growth centers to induce development in rural areas (Shah 1974; Katzman 1975; Widner 1990). In national contexts, growth pole strategies were even adopted to balance uneven patterns of economic development (Semple, Gauthier, and Youngmann 1972; Richardson and Richardson 1975; Kinsey 1978). To a large extent, growth pole strategies were regarded as a panacea to deal with almost any issue of economic development.

As development strategies, the meaning of growth poles or centers is interpreted by policymakers quite loosely; in many ways such interpretations do not comply with the theoretical content. In theory, although with different emphases, growth poles are generally discussed on the subject of units of economic activities, such as firms and industries. In reality, in many contexts, cities and towns, regardless of their industrial structures, are treated as being equivalent to growth centers for regional and national economies. In the concept of growth poles, propulsive units are linked with others in economic, technological, and social ways. As a policy instrument, propulsive firms and industries are narrowly selected on a basis of static input/output structures. Of course, there are practical concerns for relying on simple and observable parameters to operationalize such a theoretical concept in designing development policies. However, a pitfall of this is that, in doing so, the concept may be misunderstood and its originality may be lost in operationalization. Unfortunately, this is the case for growth poles. In degenerating into an input/output analysis, growth poles lost their core content of emphasizing innovation and structural changes in the course of economic development.

When cities and towns were carelessly included in development poles, a different argument developed in response to how to implement growth center strategies. For policymakers, a key question of the strategies is the selection of growth poles for public investment. Besides some planned poles that are created from scratch, the choice of specific areas as growth centers requires another rationale that the theoretical concept of growth poles could not offer. The rationale develops from decreasing returns of investment. To maximize the welfare of the whole economy, it is argued that public investment be invested in places with highest returns. In large metropolitan areas, investment cost is high and congestion and external diseconomies make such an investment less profitable. On the other hand, infrastructure development in rural regions and small towns may not pay off because railroads, highways, and other facilities would not be efficiently used. From an economic perspective, growth center strategies are believed to be best targeted at medium-sized cities (Cameron 1970).
GROWTH POLES AND GROWTH CENTERS

Clearly, a growth center policy in this sense distracts from the original idea of growth poles, which rests on relations of economic units. In some cases infrastructure projects are directed to connect poles and surrounding areas, with little consideration of how local economies grow. By blindly assuming induced development will occur automatically, growth poles in these contexts are changed into a tool to justify investment in certain areas. For noncenter regions, it becomes an excuse for place discrimination in allocating public resources, a point which is used in criticism of growth pole strategies.

In some other settings, growth pole strategies were practiced in a way that seems to be closer to its theoretical meaning. Large firms are carefully selected on the basis of intensive input/output linkages and implanted in industrial complexes in order to solicit the development of related industries in local and nearby areas. The rationale behind these planned growth poles is traded linkages of economic activities. This may be partly related to the idea of the dominating relations of propulsive units over others. But it is not in line with the core of Perroux’s original idea in two fundamental ways.

First, with calculated input/output analysis in these policies, growth poles are interpreted in a static manner. Connections between economic units are restricted to transactions, which excludes dynamic dimensions of interfirm relations, such as social interaction and innovation spillovers. Without considering these nontraded linkages of economic units, these growth pole strategies in fact presume a mechanical and banal picture of economic development. In many industrial complexes, it is believed that the structure of local economies can be planned. This is in sharp contrast to the original meaning of growth poles, which emphasized innovation and structural changes of economic development.

Second, even if unbalanced relations of economic units are only understood in the sense of traded linkages, these policies confuse industrial complexes in a concrete geographical setting with growth poles defined in an abstract space. In an abstract economic space delineated by various degrees of influence of propulsive units, related firms or industries may be induced to grow. However, there is no strong reason in growth pole theory to expect that suppliers and related industries would develop in the vicinity of propulsive units, which is usually the goal of many industrial complexes. In Hirschman’s (1958) theoretical model, there are only two places, the developed North and the developing South. In theory, congestion effects and external diseconomies in the North in the end will force economic activities to move to the South because there is no other locational choice. In reality, there are so many places for firms to choose. It is untenable for policymakers to expect that related activities have to develop in nearby regions.

As illustrated by the above, when the concept of growth poles was widely accepted in policy arenas as a development strategy, intentionally or unintentionally, its meaning was abridged and modified for practical purposes. In many ways, the concept lost the richness of its content in empirical analysis. With hindsight, it is not surprising that most growth pole strategies were later abandoned by policymakers for failing to fulfill their initial goals. Although in some contexts growth pole policies were targeted at goals that were too ambitious, such as balancing spatial patterns of national economies, in many cases these strategies were also not successful in bringing about promised changes that were less demanding in local economies (Ryder 1990; Widner 1990; Florio 1996). There are many reasons for the failure of growth pole strategies, such as insufficient resources to sustain these policies, low
commitment of new government, wrong selection of growth centers or propulsive industries, and so on (Higgins and Savoie 1988). However, besides these individualistic factors, there is no doubt that limitations of the concept per se are also responsible for its failure in practice.

First, growth pole theory does not explain the development of innovative activities in pole areas, which is a fundamental issue for economic development. Propulsive units, characterized by advanced technologies, are taken for granted in the analysis of growth poles. In a simple understanding, economic growth can be described in two stages, (i) the creation of new products and solutions, and (ii) the diffusion of innovation to the rest of an economy. Although Perroux may initially intend to emphasize the whole process, growth pole theory unfortunately only captures some features in the second stage. Theoretically, Perroux and his disciples emphasized the significance of innovative units for economic development. Empirically, the focus of growth pole strategies is usually the development of nonpole areas. Unable to explain how propulsive activities are established, growth pole theory is naturally read as suggesting that less developed regions cannot grow by themselves, but have to rely on the inducement of development poles or centers. An endogenous growth process, which is crucial not only for less developed regions but also for those pole areas, is excluded in growth pole theory. By focusing on input/output relations, large firms are often in policymakers' favor in growth pole strategies. Small and medium-sized firms and entrepreneurship, which are crucial for innovation and employment creation, are thus little considered because they are not significant in growth pole theory.

Second, it may also be admitted that one remarkable weakness of the concept of growth poles is that it is defined very elusively and inconsistently. The concept of growth poles is built on Perroux's broad thinking of economic development, especially on his works of different types of space and the domination effect of economic units. Although there are insightful ideas in these terms, they are described in a very abstract way by Perroux. For example, in his explanation of propulsive industries, Perroux initially thought that key industries could not be specified by their technical characteristics, but should be defined relatively on a basis of the activities influenced by them (Perroux 1970/1955). It becomes a chicken-and-egg definition because propulsive industries delimit their affected activities which are used to define propulsive industries themselves. This generally denies a clear identification of propulsive units, and makes Perroux's original idea of growth poles very hard, if not impossible, to be properly operationalized.

The most confusing point of the concept of growth poles is its geographical dimension. Although Perroux initially did not define growth poles in a geographical setting, the concept was widely interpreted in a concrete context, especially in policy arenas. It should be mentioned that even Perroux defined growth poles in different ways in his writings. Unlike in his first paper on growth poles in 1955, he later accepted a geographical understanding of the concept and defined growth poles as “the concentration of productive agents, organized resources and technological and economic capacity within a non-geographical matrix, or a structured group of individuals in a geographical entity” (Perroux 1983, 98–99). As analyzed before, growth poles in nongeographical and geographical settings involve very different theoretical arguments. These inconsistent definitions of growth poles impair the concept and give it less clarity and rigor.
GROWTH POLES AND GROWTH CENTERS

Third, a mixed spirit of being both explanatory and normative puts growth pole theory into an irrefutable position. Fundamentally, growth pole theory is built on the basic fact that growth is geographically uneven. As a large part of growth pole theory, the concepts of propulsive activities and polarization effects are developed to describe and explain this unbalanced spatial pattern of economic development. Beyond this explanatory content, growth pole theory further foretells that growth will spread out to nonpole areas, an argument that is scarcely substantiated with evidence. By combining the two development stages, growth pole theory presents an economic development process with an unusual quality of being both explanatory and predictive. Broadly interpreted, it provides an answer to many different questions, such as why some regions grow, why others do not grow, how regional economies will grow in the future, and how regional growth can be promoted. For many students and policymakers, there is no need to test growth pole theory before it can be accepted and implemented because growth pole theory tells a “truth” of regional uneven development. In many economic development plans, the theory becomes a belief of how regional growth should be. To some extent, it is hard to disagree with growth pole theory since it could be interpreted in many ways, being simultaneously an argument, a belief, or even a “fact.” As a result, it may not be surprising that growth poles were wrongly treated as a panacea for regional development issues.

In many ways, a loose definition and explanation of growth poles facilitated the acceptance of this concept in various settings. However, the popularity of growth poles in policy arenas did not help consolidate the concept, but instead ruined it in the end. From this, growth pole theory gains nothing but failure in its operationalization process. An important lesson that can be learned from growth poles is that it could be very dangerous for a theory or concept to be widely accepted and implemented before it has been scrupulously examined, critically debated, and solidly evidenced.

SEE ALSO: Cores and peripheries; Industrial agglomeration; Industrial complex; Regional development policies; Uneven regional development

References


Further reading


Growth theory

Simon Xiaobin Zhao
University of Hong Kong, China

Growth theory is an umbrella term for a number of different theories concerning the economic development of nations. It is a loose school of thought that emerged in the post–World War II era from a number of different economists with differing theories. The basic aim of these theories was to explain how so-called third world countries might achieve economic development and convergence with the First World (developed countries). Growth theorists first sought to understand the varying factors that contributed to economic growth in the developed world; they then attempted to apply these theories in order to create a universal model for economic development. Growth theories were also shaped by the political climate of the 1940s and 1950s, particularly the Cold War politics that was emanating from the United States of America and Western powers. The political context for growth theory is most pertinently shown by Walt Whitman Rostow’s The Stages of Economic Growth: A Non-Communist Manifesto (1960); a reaction to economic growth theories emanating from the Soviet Union (USSR) at the time.

Modern growth theory economics is in part a reaction to the previously existing neoclassical models, namely the Harrod–Domar model (Harrod 1939; Domar 1946) – an example of post-Keynesianism – which later led to the development of the Solow–Swan model (Solow 1956; Swan 1956). The Harrod–Domar model, the earliest example of growth economics, explains the long run economic growth of an economy in terms of the level of national savings within the economy and the productivity of the capital available, with an increase in either resulting in economic growth. The model does not subscribe to the idea that economics should experience balanced growth; according to the Harrod–Domar model there are three distinct types of growth: warranted growth, actual growth, and the natural rate of growth. The Solow–Swan model instead focuses on the rate of technical progress as the key focus for economic growth and development and was the first growth model to distinguish between the vintages of the capital. The model shows that growth can be achieved by increasing investment or reducing taxes that directly relate to new capital accumulation, thus increasing the productivity of the economy. The Solow–Swan model growth is only short term in nature, as in the long run the natural growth rate of an economy converges towards a steady state. Both models are criticized for assuming too much within their framework: the Harrod–Domar model assumes that economic growth will result in full employment; the Solow–Swan model assumes that too many of the factors of production are fixed and that capital is subject to diminishing returns in a closed economy.

Rostow’s (1960) linear stages of growth model is a simple stage theory that explains the process of economic development based on the historical experience of the industrialized powers. Rostow’s model builds off of Harrod–Domar’s growth theory, emphasizing the role of savings (investment) in stimulating economic expansion. The premise of this structuralist theory was that economic growth for all countries followed a set
GROWTH THEORY

pattern, and therefore economic growth could be shown to progress through five distinct stages of varying lengths. The five stages are traditional, preconditions for takeoff, takeoff into self-sustaining growth, drive to maturity, and age of high mass consumption. The traditional stage is characterized as an almost completely undeveloped primary sector economy. The preconditions to takeoff involve the development of excess production of raw materials, mainly cash crops, in order to utilize the surplus coupled with an expansion of the technological frontier. The takeoff into self-sustaining growth involves the expansion of the secondary sector, mainly in the form of a few key leading industries. It is then followed by the drive to maturity in which the industries of the secondary sector are grown, diversified, and developed away from capital goods towards consumer-driven goods (e.g. apparel, household appliances). The final stage is the age of high mass consumption, characterized as an economy that is heavily reliant on the secondary and tertiary sectors, with the primary sector greatly diminished, and where the consumption of high-value goods is widespread.

According to Rostow, the onus for an economy to emerge from being traditional to the takeoff stage is the mobilization of internal savings in order to fuel investment endogenously; however, this can be replaced by the exogenous injection of foreign aid or capital. Rostow’s model assumes that the developing countries will follow the same patterns of growth that had been observed in the developed countries of the period, which is the major criticism of this theory.

As a result of the rigid nature of Rostow’s theory it has been considered less analytical and less historically sophisticated than other models. A similar but contrasting model was developed by Alexander Gerschenkron – the “backwardness” model. Gerschenkron (1962) theorizes that some economies, particularly those in Eastern Europe, are more “backward”; although he never provides a specific definition for “backwardness,” he relates it to low income per capita, savings rates, human capital, and levels of technology. The “backward” nature of an economy creates a greater chance that the country will lack economic prerequisites and therefore the industrialization process will involve substitutions to achieve these – a focus on the production of producer goods instead of consumer goods, the reliance on imported technology rather than indigenous technology, and the use of specialist institutions (the state in the case of Russia) to facilitate the use of human and physical capital in designated industries. Gerschenkron’s model, like Rostow’s, is more of an observation on development rather than a fully functioning model.

A successor to Rostow’s theory, which came to the forefront of the discussion in the 1970s, was Lewis’s (1954) structural change model. Lewis’s model is an analysis of the internal process of change that a developing economy will undergo in order to generate economic growth. The Lewis model, also Keynesian in inspiration, theorizes that a developing country’s economy has two sectors, primary (agriculture) and secondary (industry). In the model the factors of production, predominantly labor, are underutilized in the primary sector, which results in a marginal productivity of effectively zero; this means labor can be extracted from the primary sector without reducing its output. In order for the economy to grow and develop, the factors of production, normally labor in the case of agricultural societies, need to be reallocated towards the secondary, urban, industrial sector enabling the economy to grow off the back of cheap, plentiful labor. To encourage labor to migrate, Lewis assumed that the urban wage would have to be approximately 30% higher than the wage, or equivalent, available in the rural
sector. The profits created within the economy from exploiting cheap labor can be reinvested towards capital accumulation, which can lead to self-sustaining economic development. However, this ignores the potential for capital to replace the need for labor – as happened during the Industrial Revolution in England, leading to the Luddite movement that emerged as a reaction to the resulting unemployment. The model also assumes a perfect, competitive labor market, which is unlikely in a developing economy.

Following the focus on Keynesian economics that the previous growth theories had been based around, the neoliberal counterrevolution, which occurred primarily under the Reagan and Thatcher governments, was a monetarist approach. The Chicago School of Economics is credited as being the intellectual source of the neoliberal movement, with Milton Friedman developing the *laissez-faire* economics and Friedrich Hayek developing the model’s political roots. Friedman sought to remove as much of the state’s influence on markets as possible in order to maximize their efficiency. The change of policy within key Western governments towards the neoclassical approach was both a political and an economic shift, and was heavily influenced by the Cold War geopolitical situation. The privatization of many nationalized companies in the United States and the United Kingdom, for example British Telecom, resulted from a notable shift in government policy. Neoliberal policymakers chose to focus on supply-side macroeconomics, privatization of public companies, and reducing, or eliminating, the controlling role of the state over economies. The main argument of the neoliberal counterrevolution is that underdevelopment and stagnation occur due to poor resource allocation by the state due to the rational expectations of the market being usurped, and instead free markets should be trusted to optimize resource allocation. The fundamentals of neoliberal economic thought are reliant on a strong legal grounding to create the ideal environment for the *laissez-faire* economics to be effective. This requires a transparent legal system with strong property rights enforcement and market-friendly institutions to support economic policies. On a global level, neoliberal policies are promoted by organizations such as the International Monetary Fund (IMF) and the World Trade Organization (WTO), which are heavily influenced by Western powers. Developing countries’ ability to receive foreign aid was, in part, based on the adoption of neoliberal economic policies.

Prior to the 1980s, most growth models were based on the assumption of exogenously driven long run growth, the savings rate for the Harrod–Domar model, and the rate of technological progress in the Solow–Swan model. In contrast, most contemporary growth theories are based on endogenous growth, even though many of these still exist within a Keynesian framework. The late 1980s saw the emergence of a neoclassical growth model that allowed for endogenous growth, called a new growth theory (Romer 2011). The impetus behind the development of the new model was the disenchantment with other neoclassical models’ ability to explain the increasing disparity between the developed and less developed economies. The major difference is that the new growth theory regards knowledge and technological development as, at least partly, endogenous as the economy will internally develop technology through research and development. This also applies to human capital, which can be improved endogenously through increasing investment in schools and training. Another difference between the new growth model and the older neoclassical models is that in the newer model the traditional neoclassical assumptions of diminishing marginal returns on capital and
increasing returns to scale in aggregate production are negated. However, the theory still utilizes many of the assumptions of the original neoclassical models that may not apply to the real world, such as the assumptions of perfect information for the labor market and the free mobility of capital. The natural inefficiencies are not accounted for by the model and may reduce the potential of the less developed economies to grow.

The evolution of growth theory is clearly a reflection of the prevailing political and academic thought paradigms, which are often dictated to the developing world from the developed world. One of the difficulties faced by growth theorists is attempting to develop a nomothetic theory when faced with an idiographic reality; this is reflected in the different developmental paths observed in different parts of the world (see Dependency theory and World-systems theory). Growth theory continues to be a key global political and economic issue driven forward by both international organizations and countries.

SEE ALSO: Comparative advantage; Dependency theory; Neoliberalism

References


Further reading


Habitat destruction and fragmentation

Jing Sun
Michigan State University, USA

Habitat destruction and degradation, as a result of land-use and land-cover changes (LULCCs), and consequent fragmentation are becoming central themes in understanding environmental and socioeconomic changes at scales from local to global. The global map produced by Ellis and Ramankutty (2008), and the derivations included here, are perhaps the most interesting geographical representation of the extent of LULCC. The area of habitat destroyed loses most of its plants and animals, and so research has focused on remnants, often as islands, with particular attention to their edges. These fragments are one part of a continuum that has been quantified and mapped globally by Riitters et al. (2000). The role of the converted land, however, is still important. Effects are at species level, community level, and broader (Malanson 2014).

Processes and stages of habitat destruction and fragmentation

Given the recognition that LULCCs are a matter of spatial pattern and process, steps of spatial change have been identified: dissection, perforation, fragmentation (the cutting into pieces), shrinkage, and attrition (Forman 1995) (sometimes referred to in sum as “fragmentation”).

Perforation is the creation of holes in a pre-existing habitat landscape. Perforation in the absence of roads is not likely, but it could be the process that initially decreases the amount of interior habitat while creating edges. In this stage, connectivity of habitat remains generally high. Perforation occurs with low-density settlement.

Dissection is primarily the result of roads. It has a small effect on area, but can create isolation and edge effects. Dissection can be complete, as in the case of roads cutting through a habitat or placing a grid on one, or less so, as with a network partially penetrating a region.

Fragmentation occurs when multiple pieces of habitat are isolated by the expansion of roads and new LULCCs. Fragmentation as a process occurs through the expansion of LULCC after the initial perforation and dissection stages.

Shrinkage is the simple reduction in area of the fragments of habitat created by the above processes. Area declines while isolation increases and edges may never stabilize. Extensive agriculture or forestry and high-density settlement lead to shrinkage.

Attrition is the culmination of these LULCC processes, when individual remnants disappear. Only new land uses remain. This stage is most likely with very extensive agriculture or more intensive development.

Another stage is recovery. New land uses in fragmented landscapes are not static and some degree of recovery toward the pre-existing habitat is possible. Recovery will often depend on the degree of initial destruction or modification within the new land use as well as on the continuing use or nonuse, but it will also depend on the characteristics of a given ecosystem. These characteristics are often summarized as resilience, a term with alternate definitions, which summarizes how close recovery can bring
HABITAT DESTRUCTION AND FRAGMENTATION

an area to its pre-existing landscape and how fast and by what pathways it changes.

Conceptual basis for the effects of habitat destruction

The effects of habitat destruction are greatest in the area of the LULCC, where species and their communities are at least locally eliminated, but what happens in the remnants is of interest. LULCC processes usually disconnect conterminous habitat areas and thus change the spatial properties of ecosystems, in particular, landscape connectivity. The reduced connectivity limits food sources and behaviors of wild animals, as well as the flow of material, energy, and plant propagules. The consequences of the resulting spatial patterns are often the focus of studies of the changes in species composition or ecosystem function.

Habitat destruction has been identified as having five spatial aspects important for surviving biota: sample effects, forest area, isolation of forest remnants, edge effects, and the effect of the surrounding “matrix” on all of these (and their interactions) (Kupfer, Malanson, and Franklin 2006).

A sample effect exists because the organisms in the remnants are possibly a biased sample of those in the pre-existing habitat. This effect will have consequences for the next four conditions.

Reduced area of habitat has three primary consequences: reduction in resources, metapopulation dynamics, and altered disturbance regimes. Area is the primary basis for the resources available for organisms and thus affects population sizes and the probabilities of local extinction – which then affect metapopulations. In some cases, such as for species at higher trophic levels that must have an extensive hunting area, even individual survival is area dependent. Reduced area also results in a reduction of the range of habitats and thus in lower diversity and lower probabilities for species interactions. Disturbance regimes are also area dependent on the probability of events. The notable secondary consequences, and the driver of many studies of fragmentation, are that the diversity of species depends on area because it depends on the probability of extinction (MacArthur and Wilson 1967).

Fragmentation, by definition, increases isolation and so the division of populations into metapopulations and the probabilities of immigration versus local extinction (and again, diversity). Isolation also affects the disturbance regime for those that spread. The degree of isolation, the very meaning of distance, will vary with the motility of the species or the disturbance as well as with the characteristics of the surrounding landscape (i.e., corridors or stepping stones).

Fragmentation creates edges, which are habitats themselves, often because of a different microclimate, for different species – and sometimes invasive species in particular. Edges also change the function of the forest in relation to the flux of energy and matter, and so exports and imports tend to increase, but they can also be filters and sinks. Edges also elevate further human influences, such as noise, simply by being the location for activities.

All of these can interact, and so fragmentation becomes a multidimensional spatial process. The area of converted land can affect all the above interactions, but the processes will depend on the nature of the new land use and land cover. The matrix may provide usable resources to the organisms in the remnant; it may provide connectivity among populations; it can be a source, conduit, or barrier for disturbances; it will alter the flows of energy and matter to and
away from the edges; and its use will determine the continuity of human pressure on the remnants. The new land use can also be long- or short-term. And single land covers may have different types or intensities of land use.

The details of spatial patterns matter in the processes of LULCC and fragmentation, and these are best seen in the methods applied in research.

The major methodological area of LULCC is remote sensing. Remote sensing has quantified the extent and pattern of LULCC at most scales. Global- and regional-scale remote sensing provides new views of the planet. At local scale, remote sensing is a primary source of data for quantifying the spatial pattern that affects the processes. A variety of spatial metrics have been developed in association with habitat destruction and fragmentation. Simulation modeling is another important method because it can include the details of spatial patterns (in area, isolation, edge, and matrix) that are not possible in real landscapes – and with replication.

Remote sensing is covered elsewhere in this encyclopedia. Here, it provides the primary data for spatial metrics. Too many spatial metrics exist: they are redundant and confusing. Riitters et al. (1995) improved this situation by identify six metrics that represented the most important factors of fragmented landscapes: average patch compactness, overall image texture, average patch shape, patch perimeter-area scaling, diversity, and fractal dimension. Understanding LULCC at multiple spatial scales is a challenge. Newer metrics address changing spatial and temporal dimensions with spectral and wavelet analyses to capture modifiable areas and time series. Spatial information that captures the processes of fragmentation with more nuance can be seen in morphological spatial pattern analysis, a metric based on image morphology that can be used to identify spatial patterns from a single-layer binary map (e.g., forest and nonforest maps) and describes the size, shape, and other geometric characteristics of spatial entities (Soille and Vogt 2009). In addition, some metrics from graph theory are also for quantifying connectivity.

Locations of LULCC

The most interpretable view of LULCCs is that of the areas mapped as wildlands or as “remote” forests and rangelands (Figure 1). Among these, the remote forests are the frontiers of LULCC and habitat destruction, and the remote rangelands can be subject to overgrazing. Wildlands may exist to some degree in Amazonia, the Arctic, and the barren deserts. The changes, even from 2000 to 2009, are among the most significant global changes created and faced by humans.

As an example of slightly less extensive fragmentation, Riitters et al. (2002) analyzed land-cover maps for the continental United States, which used 1992 Landsat Thematic Mapper imagery. They summarized indicators of four aspects of forest fragmentation – patch size, edge amount, interpatch distance, and patch contrast – and found that while most forestland in the United States was reasonably well connected over large regions, especially in places where forest was generally dominant, much of it was highly fragmented and in close proximity to roads (Figure 2).

The future of fragmentation

Kupfer, Malanson, and Franklin (2006) identified the most important areas where information on fragmentation is lacking:
Figure 1  Anthropogenic biomes of the world. (a–d) Ellis et al. (2013). © 2013, The Trustees of Columbia University in the City of New York; (e) Reproduced from Center for International Earth Science Information Network (CIESIN). © 2009, The Trustees of Columbia University in the City of New York.
Figure 1 (Continued overleaf).
corridors and connectivity,
• natural heterogeneity,
• edge-mediated dynamics,
• species-specificity,
• landscape metrics.

And they noted that the approaches to these areas needed to address variability in habitat quality in fragments, feedbacks and thresholds, and the methodologies for quantifying heterogeneity. They further made three recommendations for addressing habitat destruction and fragmentation directly: (i) promote connectivity using a variety of approaches; (ii) maintain stand structural complexity and landscape heterogeneity; and (iii) embrace risk-spreading and recognize the importance of different conservation strategies at different spatial scales.

Current work on habitat destruction is linked to climate change and global trade. In the world beyond academics and research, the most important effort in this regard is the United Nations collaborative initiative on Reducing Emissions from Deforestation and forest Degradation (UN-REDD) in developing countries (www.un-redd.org). This program aims to reduce forest destruction and degradation, and provides guidance to partner countries in their efforts.

**SEE ALSO:** Disturbance in biogeography; Island biogeography; Land-use/cover change and climate; Patch metrics
HABITAT DESTRUCTION AND FRAGMENTATION

Figure 2  Average number of forest edges per unit forest area ($n = 127,012$). Adapted from Riitters et al. (2002). © With permission of Springer.

References


Further reading


Habitus

Tom Lusis
University of Guelph, Canada

At the time of his death on January 23, 2002, Pierre Bourdieu was considered by many to be the most influential intellectual in France. Born on August 1, 1930, Bourdieu began his academic career studying philosophy but later changed disciplines to anthropology and sociology, where he investigated French cultural, educational, and class systems. One of his key theoretical contributions was habitus, the origins of which are found in Aristotle’s concept of *hexis* – an acquired set of moral characteristics that guide an individual’s feelings and emotions, and, by extension, their actions. Bourdieu built on this idea, and habitus became central to his theory of an economy of practices that recognized how social conditions shape, influence, and inhibit individuals and social structures in society.

Bourdieu first introduced habitus in *Outline of a Theory of Practice*. He defined it as a system of lasting, transposable dispositions that integrate past experiences and function as a matrix of perceptions, appreciations, and actions (Bourdieu 1977). Dispositions are shaped by social conditions (e.g., family background, education) and are socially reproduced as habitus generates practices that reflect the structures that originally created them. Although they may be long-lasting, dispositions and the habitus that formed them are not unchangeable. They can be eroded, modified, or even reversed through pedagogic effort, new experiences, education, or training.

Arguably Bourdieu’s most famous application of habitus has been in his book *Distinction: A Social Critique of the Judgement of Taste*. Here Bourdieu (1984) illustrates how “taste” is the product of distinct social experiences and can be common to those with comparable social conditions. The bourgeoisie habitus, for example, is characterized by food, art, music, and clothing associated with aesthetic indulgence. In contrast, the working-class habitus and consumption patterns are more functional in nature. Thus habitus is key to understanding how classes set themselves apart from one another in the field of cultural consumption.

Bourdieu’s ideas have been employed by researchers in feminist thought, women’s studies, organizational research, literary history, anthropology, educational research, urban studies, and medical/health studies. The concept of habitus has caused some debate in geographic circles. In 2001 Edward S. Casey initiated a discussion in the *Annals of the Association of American Geographers* when he suggested the concept could be used to answer the question “what ties place and self together?” He envisioned habitus as a mediating third term, a way to better understand the relationship between lived space and the geographic self. Others countered that habitus had limited geographic value due to its implicit reductionism of place as a theoretical concept. Critiques notwithstanding, habitus has been used by geographers in a range of research areas. Select examples include: religious geographies and the role social institutions play in regulating behavior outside the church; economic geographies such as the entrepreneurial activities of youth in India or the labor market integration of immigrants in Canada; rural geographies where moral...
landscapes influence agricultural activities; urban geographies and how consumption and “taste” delineate city neighborhoods; and immigration studies that explore the transformative nature of transnational migrations.

Bourdieu once argued that habitus is important and indispensable to understand human action and is a tool that can be used in relation to geographic and social space (Bourdieu 2002). Judging by the impact habitus has had in many different disciplines, it seems that many academics and researchers would agree.

SEE ALSO: Class; Cultural capital; Social capital

References

Health and development

Robert Huish
Dalhousie University, Canada

Health may not always be the goal of international development, but it will always be the outcome. For geographers the goal is to understand how the organization of space and the nature of place matter in influencing these outcomes. The connection between international development initiatives and health outcomes runs deep. As Phillips and Verhasselt (1994, 3) argue, “it implies complex interaction between humans and their environments, more particularly between social and economic factors, physical environment and biological environment.” Disease ecology, a subfield of medical geography, influenced by Jacque May, acknowledged this complexity as early as the 1940s. Today, it is well understood that the lived health outcomes of these complex interactions are strongly governed by social inequities within societies, rich or poor (Cockerham 2013). Development influences the health of place, as access to resources, economic opportunities, human rights, and environmental stewardship all matter greatly for the maintenance of good health (Meade and Earickson 2000; Phillips and Verhasselt 1994). Development is a contributing factor to good health, and yet good health is also a contributing factor to development. Education levels, exposure to conflict, work opportunities, safe working environments, and local autonomy are easier to guarantee to a healthy society, and at the same time these elements are needed for a healthy society. These are all priorities for the study and practice of international development.

Despite such an intimate relationship between health and the outcomes of development, the overwhelming amount of funding for research and practice of health within international development is currently focused on developing innovative curative practices, vaccination routines, and public health interventions through new technology. Private foundations, such as the Gates Foundation and the Clinton Foundation, handle the majority of funding for health projects within the development sphere. Organizations like Grand Challenges give enormous sums of money for research that employs technology to address disease-specific health problems. Government interventions, such as the President’s Emergency Plan for AIDS Relief (PEPFAR), are invariably disease focused. In this sense, the interaction between health and development has been generally labeled as “Global Health.”

Medical practitioners seldom engage in practices of development in the course of their efforts, even though global health, a field intent on connecting health to development, is a subject of increasing popularity. Within medical schools global health is a loosely defined field that deals with improving quality of care in marginalized communities throughout the Global South. Often, global health programs focus entirely on curative care, short-term health interventions by Northern practitioners in the South, and provision of human and material resources to underfunded clinics. Little attention is given to the geographical connection of health to place. On the medical side, the approach to development is overwhelmingly focused on costly and reactionary interventions rather than laying foundations for improved
HEALTH AND DEVELOPMENT

social and environmental conditions of a particular place that promote good health (Farmer 2005, 200).

The relationship between health and development is deeply intertwined, and even though both development practitioners and medical professionals have interests in improving health outcomes in marginalized communities, the formal connection between health, place, and development remains narrow. Very little development practice is done, either on the medical or on the development side, to specifically improve health outcomes by addressing the social determinants of health. For geographers, the lack of practical and scholarly attention to connect health to place within development practice should be taken as an opportunity to fill a looming knowledge gap. What is more, geographers have an important role to play in understanding spatial and temporal spheres that interact to create healthy spaces (Lewis and Kieffer 1994). The disconnect between the well-established knowledge of how inequities lead to poor health and the necessary action for improved health outcomes shows the need for research that connects health and place to improved practices of development work. In particular, greater attention is warranted to understand how inequitable health landscapes emerge through the uneven consumption of health-care services (Peet and Hartwick 2009), the interaction between individuals and the environment, and social inequities within place (Harvey 2006), and to then connect this knowledge to pressing global health challenges.

A brief history of global health

Acknowledging the connection between place and health dates back centuries; however, the dedicated effort to improve health conditions in another country, with seemingly little benefit to one’s national compatriots, begins in the early twentieth century. According to Palmer (2010), global health begins in 1913 with the Rockefeller Foundation initiative to rid hookworm disease in Central America and the Caribbean, where workers were falling ill on plantations and in infrastructure projects such as the Panama Canal. Before then, it was rare for one nation to take a vested interest in another nation’s health status. In the years leading up to World War I, a great deal of American capital was invested in the region, mostly for fruit production. The Rockefeller Foundation took it upon itself to deliver hookworm medicine in the region, building on a 1910 investment to eradicate the disease in the Southern United States. In a sense it was an effort to extend an already tested method into a region deeply impacted by US capital. The intervention was purely pharmaceutical and it did not involve the establishment of long-term medical facilities or Foundation-run services.

In contrast, the British government also engaged in a form of global health in its colonies through the establishment of public health infrastructure that had a much greater focus on the role of place. Great Britain attributed good health to be the result of sound planning that addressed hygiene and proper sanitation. As a result, the British military encouraged the construction of public health infrastructure in its colonies. To the Rockefeller Foundation, as well as the US government, the idea of constructing public health infrastructure smacked of old world colonialism. The pharmaceutical approach expressed sentiments of modernity and compassion without having to physically design a country’s infrastructure (Palmer 2010). This approach, from eradicating smallpox to combating HIV/AIDS (human immunodeficiency virus/acquired immune deficiency syndrome),
set the tone for the relationship between health and development for the twentieth century as focusing on curative initiatives with intentionally minimal intervention in the organization of a nation’s health system.

In international development circles, attention to health dealt with eradicating certain diseases through the assistance of foundations and nongovernmental organizations well after the rise of the post-World War II development era. The United Nations established the World Health Organization (WHO) in 1948 and its constitution laid out one of the most progressive definitions of health to date, as “a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity” (WHO 1948). Despite a constitution that grounds the understanding of health in a social determinants approach, the mandate of the WHO has been mostly for coordination of international health initiatives that are often disease focused. Its role is to coordinate campaigns and to provide national governments with technical assistance if requested in order to fulfill long-term health objectives. The coordination efforts have paid off in the eradication of smallpox and the near elimination of polio, as well as for campaigns to quell tuberculosis and malaria. National-level campaigns in the Global South, coordinated with the assistance of the WHO, have typically involved vaccine strategies, pharmaceutical therapies, and prophylactic regimens.

Rarely have there been effective WHO-coordinated national-level campaigns that address health lifestyle challenges related to social or economic status. In part, this lack is due to the paltry resources that nations in the Global South are able to contribute towards health systems. While the WHO, despite its holistic definition of health, approaches health needs through an overly medicalized lens, global development institutions such as the World Bank and the International Monetary Fund engage in development in the South with an overtly monetary view of progress. Through decades of loans for infrastructure development funded by the World Bank, many poor countries were put in debt and were then forced to satisfy their balance of payments by reducing public spending on public health and education initiatives. In the post-World War II era until the end of the twentieth century, health and development were widely divided between a medicalized disease-focused framework and that of a monetary-focused approach to development. In an era when some of the wealthiest nations struggled to secure their own national health systems, it was assumed that poor nations needed to overcome poverty before securing quality health and health care.

In 1984, Wilbert Gesler showed that regardless of the economic status of a nation, all societies strive to be healthy. Even if conditions of poverty lack adequate health resources, it does not quell the desire to be healthy in any way. When modern medical resources are absent, poorer populations turn to traditional practitioners, green medicine, and local knowledge. Gesler correctly pointed out that it is the goal of national governments “to increase the availability of modern medicine. Thus in the late twentieth century the challenge remains – to deliver health care to a largely poor, often rural, population in the context of a sometimes bewildering variety of local beliefs and traditions about illness and its treatment” (Gesler 1984, 4). The drive to bring modern medicine to the poor established a link between health and development scholars through formal evaluations of medical systems, and on a smaller scale, program evaluations of disease-specific interventions. This line of research triggered a more critical
HEALTH AND DEVELOPMENT

...turn in the social sciences to suggest that pushing the knowledge of modern medicine over traditional knowledges of health and wellbeing serves as a destructive form of neocolonialism. This is to say that medical knowledge is loaded with discourses of power and hierarchy that can enforce, and reinforce, colonial power structures in a development landscape. While medicine brings about important health improvements, the manner in which medical knowledge is delivered can subjugate vulnerable populations to power relationships that are seemingly inherent to medical practice. The “critical turn” in geographical thinking raised important questions as to the relationship between health and development, particularly in how health needed to be taken into consideration in its own right as something more than the consequence of economic development, or the receipt of modern medical treatment (Phillips and Verhasselt 1994).

The critical turn in geography echoed through the social sciences and created a tense relationship between medical practitioners and critical development scholars who believed that processes of colonialism through medicalized interventions created ethical challenges, as well as harm to marginalized communities. Paul Farmer, a physician and anthropologist, negotiated this tension well and called for an approach to health and development that embraced local autonomy and encouraged access to modern medicine. Working in Haiti, Farmer (2005) argued that all communities desire good health and often pursue it through both modern and traditional methods. It becomes an ethical problem when knowledge exists to help communities in need, but society fails to deliver assistance. Farmer called for the strengthening of public health resources in the Global South in order to prevent suffering, as well as to relieve it. He further suggested that it is no accident that most of the global health burden impacts the poor. Diseases like tuberculosis, malaria, HIV, and cholera systematically impact the marginalized over the affluent. An individual who succumbs to tuberculosis is not just suffering from a bacterial infection, but from the embodied effects of poverty, inequity, and harmful social relations.

Even though the majority of funding for health and development remains limited to disease-specific interventions and innovations, Farmer’s approach demonstrates how deeply connected development, place, and health are. It suggests that health is not the result of development, but rather it is implicated in it. Failure to approach health in this way works to disconnect understanding of how health and development are congruent rather than consequential to each other. Many development practitioners continue to approach health as an inevitable long-term outcome of economic development through modernization and urbanization. This situation often results in needless suffering, as immediate attention to health concerns can be neglected in favor of pursuing economic goals. This result can be seen in how health-care workers are employed, how environmental determinants impact health, and how many diseases of poverty abound.

Health-care workers

The majority of global mortality occurs in nations where human resources for health are the weakest, creating a global landscape of massive spatial inequality. Health workers in Southern Africa, for instance, are limited to begin with. There is one physician for every 1300 people in South Africa, one for every 8300 people in Zambia, but in Malawi, a country with scarcely 300 doctors for over 12 million people, there is one physician for every 50 000.
HEALTH AND DEVELOPMENT

people. These numbers are a stark contrast to Europe where the widest ratio is 1 doctor for every 500 people in Albania, and the smallest 1 physician for every 200 people in Greece. To makes matters worse, Southern Africa experiences a “brain drain” of qualified persons out of their home countries to serve in wealthier health systems. The brain drain is considered a significant global health problem, with many nations struggling to develop retention strategies for their health workers. The Philippines government and the World Bank, however, embrace outmigration of health workers as an economic development strategy. The Philippines Off Shore Worker program encourages the training of nurses, among other professions, to head overseas and send remittances home. In 2004, the Philippines received the fourth largest volume of remittances in the world, with over $11 billion being sent from workers abroad. In 2003, over 160,000 Filipino nurses worked in over 80 countries; however, only 29,466 nurses remained in the country to meet the needs of roughly 94 million people. The push for outmigration, while bringing in much-needed hard currency, has not allowed for the development of a robust health system to attract and retain health workers. The result is that entire islands are without health workers, and those who are present in rural areas of the country are underpaid compared to their offshore colleagues, and work in challenging circumstances. The 2013 Typhoon Haiyan exposed the fragility of health-care services in the Philippines as millions were left to cope in the storm’s aftermath without access to quality care. In sum, development objectives of increasing remittance income through the outmigration of health workers has increased the spatial vulnerability of the poor and demonstrates a noticeable disconnect between development objectives and health needs of marginalized places in the Global South.

Environmental determinants of health

The environmental determinants of health are broad, complex, and deeply associated with the interaction of land and life. Agents that alter the state of health in individuals can be found in any society. In poor communities, difficulties arise in securing the resources to identify, measure, and react to environmental agents that negatively impact health. Vulnerable populations have increased chances of being exposed to harmful agents and chemicals associated with heavy industry, agriculture, and extractive industries. When national-level regulations are compromised due to austerity, it becomes incredibly difficult for countries in the Global South to address environmental health concerns. Furthermore, the sorts of health conditions that develop from exposure to harmful chemicals and agents over the long term tend to be chronic illness, rather than vector-borne illness.

The African continent is facing an alarming increase in cancers in the past decade that are thought to be associated to increased exposure to harmful environmental agents (Livingston 2012). In both men and women, young and old, cancers are increasing on the African continent at rates higher than ever before seen, creating tremendous geographical inequality to the point where illness is experienced in radically different manners depending on place. This trend is true too in the Global North; however, in the North, health systems have made impressive efforts to respond to cancer care through advanced research, treatment, and hospice services. African health-care systems are completely unprepared to handle advanced-stage cancers. The mortality rate of breast cancer in Nigeria is 22 per 100,000 while it is only 17 per 100,000 in the United States. Prostate cancer deaths in Uganda are 33 per 100,000, which is exponentially higher than Canada’s rate of 13 per
100 000. Even though cancers are pervasive on the African continent and contribute to a growing burden of suffering and mortality, oncology aid to the continent is incredibly limited. The United States contributed over $7.7 billion to HIV in Africa in 2009. Close to $4.4 billion went to tuberculosis in 2012, and $32 million went to Guinea worm treatment in West Africa. But US donors raise not even a million dollars for oncology treatment in the African continent. Most of it is earmarked for the human papillomavirus (HPV) vaccines associated with the PEPFAR initiative. It is a telling example of how health and development issues in the African continent are assumed to be vector-based infections, rather than complex chronic illnesses that require advanced medical care. This assumption stems from a lack of understanding of emerging disease trends and the connections vulnerable populations have to them.

Social determinants of health

Two examples of illness that could be totally eradicated with improved social determinants are tuberculosis and cholera. Tuberculosis and cholera systematically target the poor, and both require poverty in order to flourish. These illnesses are preventable and avoidable, yet they have a devastating impact almost entirely in poor places. In international development circles the response to these illnesses comes in the form of coordinating reactionary care. The medicines are very effective. They can be costly, however, and do nothing to prevent the onset of the illness. Tuberculosis and cholera are entirely preventable. In order for these diseases to thrive they require poverty, just as much as a human host. Poor housing, poor ventilation, lowered immune systems, and poor diet will all give tuberculosis the advantage over the poor. Cholera requires inequitable geographical landscapes that involve poor water and sanitation in order to survive. These factors are entirely dependent upon development policy that maintains an unhealthy place and yet the funding and political attention to these illnesses is almost exclusively directed towards reactionary care, rather than dealing with the structural inequities of the landscape in the first place.

Addressing the social and economic conditions that lead to poor housing, unsafe water, and bad sanitation requires a policy package that deeply appreciates the structural barriers to health. Yet, with these illnesses the development funders pursue a reactionary course of action of drug therapy. The question remains as to whether the reactionary method persists because it is easier or if it is due to a lack of understanding of how to effectively address the social determinants within a particular place.

Concluding remarks

The Millennium Development Goals are all health related, and in 2015 the world failed to achieve each and every one. This defeat may well be the result of the loose epistemological connection between health, place, and development, and it may well signal the need for a deeper association and paradigm shift towards a social determinants approach to health and development. Rising from the ashes of the Millennium Development Goals are the new United Nations Sustainable Development Goals. With a deeper commitment to studying and acting upon the links between health and development it can only be hoped that the new goals will produce more fitting actions based on sound knowledge of what links health to place.

SEE ALSO: Development; Health geography; Health inequalities; Poverty
References

Health and wellbeing

Sarah Atkinson
Durham University, UK

The last two decades have witnessed a shift in policy goals from economic security, growth, and material prosperity to a more generalized attention to wellbeing. The growth of interest in wellbeing reflects serious effort to shift the assumptions of governments and their constituent populations away from the idea that doing well, whether as a nation or as an individual, is primarily evaluated by material prosperity to the idea that doing well should be assessed by considering some variant of wellbeing. Thus the decision to emphasize wellbeing as the ultimate goal for policy intervention repositions concrete policy goals, such as economic growth and material prosperity, as merely the means to human flourishing rather than as the ends themselves.

While an interest in assessing the elements that constitute a good life is not new per se, more recently governments and policy communities across several countries, including the United Kingdom, France, Canada, and Australia, have given increased attention to assessing personal or subjective wellbeing, following the lead of Bhutan and its long-established national Gross National Happiness Index. This move toward the personal and subjective in wellbeing has been reflected across the social sciences, including human geography: however, the term “wellbeing” has had a high-profile use in the domain of health since the World Health Organization (WHO) mobilized it in its 1948 constitution and defined health as complete physical, mental, and social wellbeing and not merely the absence of disease or infirmity. There is therefore a widespread assumption that wellbeing, and particularly subjective wellbeing, is primarily an aspect of health; within geography, subjective wellbeing to date has been primarily the remit of health geographers.

The study of medicine, health, and wellbeing within geography has a history of research trajectories that is similar to that of many other social sciences, and has been strongly influenced by the support given to social medicine, disease prevention, and health promotion by the WHO. Early studies focused on the distribution of diseases, associated factors, and availability and accessibility of appropriate care, thereby developing a distinctly geographic spatial epidemiology. In the early 1990s, however, geography spawned an almost parallel subdiscipline mobilized through a positive understanding of health and health care, rather than only disease and clinical medicine, and concerned with the domains of social, cultural, and political life. For this newly emergent health geography, wellbeing afforded an invaluable construct which not only enabled a model of health that went beyond the constraints of a biomedical science, but which was additionally legitimated through the WHO’s seminal definition of health. Nonetheless, the subdiscipline has, until now, self-identified as the geographies of health rather than the geographies of wellbeing, and it has been a broad notion of health that has been the primary research focus.

Since the mid-2000s, as wellbeing has been given prominence within the public policy agenda, geographers have returned to considering the notion of wellbeing itself as the primary research focus. This return is reflected
in the publication over recent years of collections of essays by geographers on wellbeing (see Schwanen and Ziegler 2011; Atkinson, Fuller, and Painter 2012; Schwanen and Atkinson 2015).

Geographies of wellbeing

Defining wellbeing for the purposes of measurement, evaluation, research, and policy is commonly identified as a major problem. Debates on what the key aspects of a good life might be range across objective and subjective approaches. Different positions on what constitutes wellbeing are closely connected to different positions on the kinds of actions and interventions appropriate to improve wellbeing. Such positions are informed in turn by a mix of political, cultural, and social values related not only to what constitutes a good life, but also to what constitutes personhood, society, the relationships between them, and with whom and where responsibility is placed for enabling wellbeing. Wellbeing, then, can offer a broad and encompassing concept through which to engage and question many debates that are central to the social and spatial sciences and to public policy. Moreover, regardless of how wellbeing is defined, it is always understood as embedded within the when, where, and how of everyday lives and as such is first and foremost inherently geographical.

Geographical engagements highlight several important features of contemporary concerns with wellbeing. First, the definitions of wellbeing, of where we understand wellbeing as residing and the contingent influences on wellbeing, may differ greatly across time, place, culture, gender, age, and so forth. Second, not only can wellbeing be examined at different scales of analysis, but it is also inherently multiscalar in production and expression. Third, a tendency to understand wellbeing as always located as the endpoint or outcome of policy is a misrepresentation of dominant usages. Last, wellbeing is almost always positioned as an individual attribute that is either an outcome or a determinant of other factors. Relational and emotional geographies offer a different engagement in which wellbeing becomes understood not as an attribute but also as an effect.

Dominant definitions and influences

The majority of contemporary approaches to wellbeing, whether objective or subjective, locate wellbeing at the level of the individual. This is an important point to reflect on as engagements with wellbeing have not always focused on this more personal or subjective aspect. Indeed, wellbeing as a concept has not even always been understood as an individual quality. In an analysis of the roots and development of the concept of wellbeing into its current usage through a review of British newspapers, sociologist Eeva Sointu has described an understanding of wellbeing in the 1960s and 1970s as a collective concept, most commonly used in the context of the economy. This more collective understanding of the term was elaborated within geography by David Smith, who examined inequalities in wellbeing in the 1970s through developing multidimensional social and objective indicators to compare territorial units, and used the concept in relation to the moral attitudes to social and environmental inequalities (Conradson 2012).

As an individual attribute, the acknowledged complexity of the concept of wellbeing is broken down into component elements that can be measured for assessment and for the identification of determinants. There are various lists of factors, both objectively and subjectively assessed, which are posited as either constituting wellbeing or influencing wellbeing (see, e.g., the work of the
Commission on the Measurement of Economic Performance and Social Progress: Stiglitz, Sen, and Fitoussi 2009). Lists may include participatory consultations of what communities or populations themselves consider their priorities for wellbeing, as took place in the United Kingdom through a national consultation by the Office for National Statistics in 2011–2012. A 2008 report for the Department for Children, Schools and Families in the United Kingdom usefully spelled out series of binaries on which various definitions of wellbeing may differ, and on some of which they found understanding did differ between different government departments: individual versus collective; subjective versus objective; permanent versus temporary; general versus specific; reducible to components versus an irreducible holistic totality; responsibility through structure versus agency; a neutral state (nothing wrong) versus a positive state (better than neutral); a state or place versus a process or journey; an end in itself versus a means to another end. In reality, most approaches treat wellbeing as individual, in which any collective assessment comprises the aggregated individual measures of socially or territorially defined groups. Wellbeing is taken to show some permanency, at least in the short to medium term, and almost all approaches handle the abstract and complex nature of wellbeing by breaking it down into reducible parts, in what has been called the components approach (Atkinson 2013).

The impetus for interest and the growth of investment in measuring wellbeing has emerged in the high-income countries of Europe, Australia and New Zealand, and Canada. Some attention to wellbeing has been given to other locations beyond the initiatives in Bhutan. The research program on Wellbeing in Developing Countries (WeD) at the University of Bath explored wellbeing in low-income countries through both objective and subjective measures that drew conceptually on the capabilities approach of Amartya Sen and needs-based approaches in development studies. The research in this program emphasized the importance of the collectivities within which everyday lives are shaped and the tradeoffs that have to be made between different components of wellbeing. The WeD researchers also promote a “messy” definition of wellbeing that allows for both local specificity and universal rights. Despite evident differences in detail and expression of the components constituting wellbeing or the components influencing wellbeing, most lists include what can be seen as universal components of at least income, social relationships, and health.

In relation to the growth of interest in subjective wellbeing, these factors become influences or resources for subjective wellbeing to flourish. Subjective wellbeing is typically captured through either an economic, utilitarian approach and associated with the idea of happiness and the work of Richard Layard, or a psychological approach associated with ideas of meaning, purpose, and self-worth exemplified by the work of Carole Ryff and colleagues. This distinction is described as not only disciplinary (economic or psychological), but also as philosophical, hedonic compared with eudaimonic, or in popular terms as happiness- and pleasure-based compared with fulfilment- and meaning-based. Examples of different sets of components in circulation are summarized in Atkinson, Fuller, and Painter (2012), and associated factors and methodological questions are summarized in Schwanen and Wang (2014). Tensions between the hedonic and eudaimonic variants of wellbeing have been drawn out across many different settings (see various examples in Atkinson, Fuller, and Painter 2012). Hedonic wellbeing may be nurtured through physical gratifications, through individual gain, and through unsustainable consumerism, gains that may be in explicit
conflict with gains for eudaimonic wellbeing, which are nurtured through social and ethical purpose. Moreover, whether hedonic or eudaimonic wellbeing is privileged will vary depending on both setting and scale. Although we might expect the gains for eudaimonic wellbeing in terms of purpose and meaning in life ultimately to trump the conflicting hedonic goals, the hedonic wellbeing from contemporary and individualized consumerism has been cast as a threat not only to individual wellbeing but also to societal and even global wellbeing.

Conceptualizing wellbeing as a set of entities that can be individually acquired brings a number of advantages for both research and policy practice. First, it enables us to ascribe stability to subjective wellbeing, at least in the medium term, and ascribing stability is essential if the measurement of wellbeing is to be meaningful. Thus, second, the measurement through which to identify and monitor trends and to assess interventions at different levels of governance can be based on individual assessments, which are then aggregated into appropriate or comparative social units. Third, changes in subjective wellbeing can be related to changes in those opportunities and resources, whether external or internal to the person, which may facilitate the acquisition of the various components. These components can then be evaluated through the standard designs of “before and after” assessment or cross-sectional comparisons.

Multiscale geographies of wellbeing

As the geographies of wellbeing as a field of research grows, one of the most important insights is that “spatial differences in subjective well-being are multiscalar” (Schwanen and Wang 2014, 835). The intimate scale of subjective wellbeing, personalized and individualized, provokes a tendency to examine determinants of wellbeing operating at a similarly individual or local scale. This tendency may be exacerbated by participatory approaches that aim to identify people’s own criteria for subjective wellbeing in terms of what is important to them. This approach, while respectful and cognizant of both contextual and personal variations, is most likely to elicit the identification of “intentional” aspects of life, those aspects that participants are most aware of and confront on a daily basis, while determinants that operate beyond the everyday encounters are less likely to be identified. Moreover, the observable change in the dominant understanding and usage of the term “wellbeing” toward an individual and subjective entity has been explicitly located as part of those changes associated with the emergence of contemporary forms of neoliberalism. An individualized responsibility for self-actualization has been positioned as a significant feature of Western society within a Foucauldian analysis of contemporary forms of governance. Indeed, the only trace of a more collective understanding to be found in current uses of wellbeing remains in relation to environmental sustainability. The broad political critique leveled at this approach to wellbeing is twofold: first, that subjective wellbeing is predicated on predefined qualities, attitudes, and ways of being that constitute the desirable citizen; and, second, that achieving these qualities is presented largely as an individual and internalized task of self-management, such that a failure of subjective wellbeing can be positioned as a failure of responsible citizenship. Seen through this lens, subjective wellbeing is an outcome of individual agency rather than societal structures, inequalities, or context.

Spatial differences in wellbeing have been described at a range of scales commonly investigated in geographical inquiries – international,
national, interurban, intraurban, neighborhoods—and are shown to involve both physical and social dimensions of geographical space. Subjective wellbeing tends to be lower in more densely populated, urban locations although not consistently, and this is countered by a tendency for wellbeing to be higher with easier access to shops, schools, transport, health facilities, and so forth. More particularly, geographic studies have demonstrated the tendency for area-level and nation-level deprivation, prosperity, and resource availability to influence local and individual levels of subjective wellbeing (summarized in Schwanen and Wang 2014). At a global scale, studies in peri-urban districts of the new economies in Latin America and Asia illustrate how global trends toward an individualized aspiration and consumption are in tension with traditional values of community cohesion and unity (see examples in Atkinson, Fuller, and Painter 2012).

The importance of geographical work on spatial variations and inequalities in wellbeing across different scales can be illustrated through the body of work on the relationships between income and subjective wellbeing. Income has attracted disproportionate attention in association with subjective wellbeing because of its centrality in policy debates. The policy move toward wellbeing is founded on a line of argument that proposes that economic growth has not been matched by comparable growth in some notion of subjective wellbeing, such as satisfaction with life or happiness. The argument traces its lineage back to a foundational and influential study by Easterlin, generating the so-called Easterlin Paradox. Easterlin examined life satisfaction and GDP per capita for the American population during the postwar decades. His striking results demonstrate a steady rise in the country’s economic growth, but by contrast the levels of self-reported life satisfaction flatlined. This result thus underpins the argument that there is no link between a society’s economic development and its average level of happiness once what is a surprisingly low level of average per capita income (indicated through GDP per capita) has been reached. Moreover, the result dovetails with arguments that it is the extent of relative inequalities, rather than absolute values in material aspects of life, that drives associated inequalities in quality-of-life outcomes such as health status. In Wilkinson and Pickett’s “Spirit Level” hypothesis (the title of their 2010 book) the importance of relative inequality in income for health is likely mediated by some kind of dissatisfaction through a comparison resonant with a notion of subjective wellbeing.

This argument not only fuels the growing attention to subjective wellbeing as the endpoint of policy intervention, but also indicates income redistribution through progressive taxation as the appropriate policy to increase average levels of subjective wellbeing. The primary argument of the Easterlin Paradox has not passed without contestation; for example, Betsy Stevenson and Justin Wolfers have presented several analyses that show both a clear positive association with average levels of subjective wellbeing and GDP per capita across countries, and the absence of evidence of a satiation point from which no further gains are made in subjective wellbeing in the richer countries. They find the same association within countries between individual income and individual subjective wellbeing over time, and argue for a renewed attention to absolute income as a determinant of subjective wellbeing.

Geographers have made important and critical contributions in exploring the determinants of subjective wellbeing by drawing attention to the complexities of scale and place for subjective wellbeing. First, comparison across very different settings confronts cultural factors that can affect
HEALTH AND WELLBEING

the extent to which subjective wellbeing or happiness is expressed. Second, given the key policy debate about the importance of relative inequality and social comparison versus absolute levels in measuring determinants of subjective wellbeing, far greater understanding is needed of with whom and at what scale we compare ourselves. While the majority of researchers have tended to examine relative inequality locally within the neighborhood, Dmitris Ballas and Danny Dorling (2013) suggest it is comparison across the nation-state that is crucial. Third, the importance of determinants for subjective wellbeing varies across space and across time. Spatially informed research using large European datasets has shown important influences on subjective wellbeing for a range of macroscale factors. Furthermore, the debate over whether the absolute or the relative value in income and other indicators is key for subjective wellbeing gets a new direction with evidence that it varies across different regions and countries (Aslam and Corrado 2012). This evidence highlights the importance to understanding subjective wellbeing of considering the particular macropolitical, economic, and historical trajectories of a given setting. These kinds of studies, working with large datasets, both emphasize the complexities involved in developing and evaluating policy interventions and expose a range of issues that need consideration across different scales of governance.

Subjective wellbeing as process

In the rhetoric and practice of policy-facing research and discussion, subjective wellbeing is located as the outcome, the desirable endpoint of other aspects of daily life and of governance, of policy and of intervention. Susan Smith and Donna Easterlow (2005) argued, in relation to the geographies of health, that positioning health as always and only the outcome in research produced a “strange geographies of health inequalities” that was unbalanced in not examining how differential health experiences may influence other aspects of life. Similarly, research across geography and development studies has begun to revise the one-sided approach to subjective wellbeing as only the outcome of a variety of factors including policy measures, to understand subjective wellbeing as also an important process factor for other outcomes, and to complicate this relationship in ways that challenge the political implications of an individualized, subjective wellbeing and that question the best routes for policy intervention. The implications of positioning subjective wellbeing as not only a possibly desirable endpoint but also a significant process factor have been drawn out through several strands of geographical research.

The first of these research strands is a challenge to the calls (and indeed large amount of funding that is being invested) to identify a tighter definition of the concept. The justification for this follows an argument that an ill-defined understanding of subjective wellbeing not only makes it difficult to evaluate interventions, but also creates barriers to communication across different sectors. In a counterargument, research with local governments in the United Kingdom has demonstrated how an imperative to develop strategies and indicators for wellbeing opened up a space in which competing discourses of wellbeing could be aired and deliberated, and thereby generated discursive shifts in the underpinnings of longer-term policy formulation (Scott and Bell 2013). Thus, wellbeing, exactly because of its vague and all-embracing character, offers a conceptual unifier across different sectors through acting as part of the process of policymaking as much as being an outcome of policymaking. From this perspective, establishing a tight and measurable definition of wellbeing in
order to evaluate outcomes would seriously limit
the scope of the policy work that the concept
has the potential to achieve (Atkinson 2013).

Second, thinking of wellbeing as process
challenges the separation of subjective wellbeing
from the objective material aspects of living and
from health status. Increasingly, both in research
and in popular imagination, wellbeing is equated
at an individual scale with various terms of pos-
itive affect. Happiness has tended to be the term
to capture this, partly because of its populist
currency and partly because it fits well with
an economic approach to flourishing through
utility: that is, subjective wellbeing as happiness
can be positioned as the utility that people want
to maximize in their lives, as the driver of their
choices in life, and as the entity to be enhanced
through policy. The argument of Richard
Wilkinson and colleagues that inequalities in
health relate strongly to relative inequalities in
the distribution of material and developmental
resources, however, is strongly suggestive of a
mediating role for a experiential happiness with
one’s circumstances that is first and foremost situ-
ated in comparison with the norms of one’s social
settings and that then impacts other outcomes
such as health status. It is difficult to distinguish
which of happiness, health, and wealth is the
process and which the outcome. Research that
only considers the local scale, rather than a mul-
tiscalar perspective, may gloss over problematic
aspects of how subjective wellbeing is expressed.
In the context of Wilkinson’s relative inequalities
thesis, those who benefit from advantages in
life may express contentment and happiness
without regard for the global and local relations
of inequality underpinning such advantage. This
may be informed by naivety, ignorance, or com-
placency and a lack of care for the circumstances
of others. On the other hand, those living with
deprivation and disadvantage may express con-
tentment and happiness that is based on a lack of
aspiration and an acceptance of social injustices.
In all of these scenarios, subjective wellbeing is
founded on various cognitive and affective states
that can be considered deficient or undesirable.
Further, in these scenarios, subjective wellbeing
is not only the outcome; it not only reflects but
also feeds a lack of critical reflection on society’s
values or a lack of imagination with respect to
alternative possibilities (Atkinson 2013).

Third, certain psychological approaches to
subjective wellbeing, defined as some variant
of positive affect, increasingly position this
positive affect as under internal and individual
control. The growing body of research on the
determinants of subjective wellbeing, including
spatial analyses of inequalities, identifies deter-
minants that are both internal and external to
the individual. The internalized experiential
happiness that mediates outcomes in Wilkinson’s
relative inequalities thesis is clearly shaped by
the comparison of external material resource
differentials; however, in the literature informed
by the positive psychology movement, subjective
wellbeing is very differently positioned as no
longer primarily an outcome of external factors,
the proper concern of local and national gov-
ernments, but a process of internal management
and the object of personal responsibility. This
emphasis on personal action effectively positions
many of the components that typically make up
psychological definitions of subjective wellbeing
as influences on various other personal out-
comes and socially recognized criteria of success.
These outcomes include health conditions and
health-related behaviors, employment and earn-
ing capacity, and productivity at work. Although
geographers have not engaged greatly with this
literature to date, there are clear spatialities to
be explored in relation to where wellbeing is
located, where responsibility is located, and
where policy intervention is located.
HEALTH AND WELLBEING

The promise of the power of working on one’s own internal wellbeing, happiness, or positive affect logically leads to policy responses that similarly focus primarily on individual deficits in fostering and sustaining positive wellbeing. For example, cognitive behavioral therapy has become very popular across a range of conditions and social contexts, while Buddhist-inspired practices of mindfulness are also attracting increasing policy interest. Geographical critiques of the responsibilization of the individual for his or her own health outcomes are being extended to the internalized and universalized presentation of wellbeing; subjective wellbeing as personal responsibility becomes no longer an outcome but rather part of the processes through which our conduct is directed according to the requirements of the political or economic imperatives of others. This is exemplified in the rise of the modern spa as a site explicitly dedicated to enable wellbeing through self-care, in which a series of power relations mediated through gender and identity enable the regulation of the body (Little 2012). Further critique indicates that viewing the acquisition of the objective material resources commonly understood as part of a flourishing life is largely determined by internal attitudes and assumes a meritocratic society in which inequality can be dismissed as a result of personal deficiency (Atkinson 2013). A focus on personal deficiency, in what Lynne Friedli has called a shift from welfare to wellbeing, takes concrete form in current policies related to unemployment benefits in the United Kingdom, whereby those who are unemployed have to perform approved attitudes and behaviors or attend positive attitude training to stay in the benefit system. Although such policies appear benign or even helpful, Friedli argues that they expose a growing governance of personality in which only quite a constrained definition of attitudes and behaviors is considered acceptable.

Subjective wellbeing as effect

Subjective wellbeing is most commonly located in or with the individual: it is measured as individually experienced and expressed, as determined by or as a determinant for other factors, and collective wellbeing is captured as the aggregate of individual experiences. The movement within geography toward greater engagement with the emotional, the relational, and the nonrepresentational domains of social life has been expressed within the geographies of health through very different conceptualizations of wellbeing in which emotional and imagined connections to place are central. Research on housing and wellbeing has demonstrated how affective bonds shape housing choices and thereby shape neighborhood health status, and how affective tensions between investments in housing as an asset and as a home are strongly reflected in differential wellbeing. Owners who are primarily preoccupied with financial returns consistently express low wellbeing scores. These housing studies describe the inseparability of emotion, investment, place, and time. This investment in an idea of home takes on particular poignancy for wellbeing as people age. The emotional and embodied connections with the home can be vital to the wellbeing of older people, through a mixture of a sense of belonging, emotional attachments from memories, and a sense of familiarity with the immediate surroundings. Research thus indicates the need for social support to enable continued residence at home rather than a move to a residential care facility. On the other hand, when older people have become frail and unable to travel independently, continued residence in their own home may bring increasing risks of social isolation, with associated deterioration of wellbeing through feelings of loneliness and depression. New technologies to support older
people in their own homes tend to focus on surveillance and physical risk; however, the value to wellbeing of caring for another has generated interesting experiments with robotic pets for the elderly that can provide a companion needing emotional care.

How the experience of being in a given landscape or environment interacts with subjective wellbeing has given rise to a substantial body of work shaped through the concept of therapeutic landscapes or spaces. Wil Gesler first elaborated the concept of therapeutic landscapes and emphasized the multiplicity and interrelatedness of the factors that can contribute to what makes a place beneficial for wellbeing. Research on therapeutic landscapes has expanded from the original application to designated exceptional sites to examine a range of different places and spaces, including the everyday and the imaginary (Williams 2007). In relation to wellbeing, geographers have given particular attention to the benefits that come from being in natural spaces, from urban parks and ponds, the greenspace and bluespace of plants and water, and from the activities of nurturing these through gardening or spiritual investment. Peter Groenewegen has dubbed the benefits of greenspace vitamin G, suggesting that improved access to existing urban parks and investment in new ones may be a cost-effective strategy for public health and health promotion. These engagements from a European or North American health geography tend to position the so-called natural environment predominantly as an instrumental factor for wellbeing and health. In some settings, such as among Aboriginal communities in Australia, wellbeing is incomprehensible if thought of as separate from the community’s relation to the natural environment; the wellbeing of people is the wellbeing of the land, the rivers, and all coexisting life forms (Gibson 2012). Health geographers have also challenged the dominant presentation of natural environments as beneficial, re-emphasizing the complexities of therapeutic landscapes and exposing the alienation of some from natural environments or the therapeutic value of unvalued, risky, or spoiled settings.

Therapeutic landscapes, then, involve a complex web of interactions between what qualities they present and what meanings are attributed to them in any given time and social setting. Human geographers have promoted an understanding that place and people mutually constitute one another, a relational understanding of geography that enables complexity and plurality as well as temporal and spatial specificity. This relational approach has been drawn on within health geographies to advance our understandings of therapeutic landscapes and spaces. In a relatively early attempt to capture the complexities and relationalities of both objective and subjective wellbeing through a geographical lens, Fleuret and Atkinson (2007) proposed what has become known as the spaces of wellbeing approach. The approach explores how different social and spatial contexts may facilitate wellbeing, and proposes that wellbeing is emergent through four interrelated spaces of resource mobilization: capabilities, social integration, security, and therapeutic processes. This framework served to expose the emphasis within studies on wellbeing, in both geography and other disciplines, on either the components of wellbeing or the factors associated with it. The framework demands a greater attention both to the relationalities between the different spaces of wellbeing and to the processes through which wellbeing, and particularly subjective wellbeing, may be enhanced. Subsequent work on the geographies of wellbeing have both built on the spaces of wellbeing approach and drawn on other trends within the discipline as a whole.
to offer several new directions and expose new underresearched topics and fields. A relational geography demands that we think of wellbeing not as an entity that can be acquired or attained but as an emergent effect of any space and time. The spaces of wellbeing approach offers a first framework to shape research on the processes, resources, and relations through which wellbeing may be emergent as an effect rather than an acquisition. But, similarly to components approaches or sets of objective and subjective variables, how such different resource spaces relate to one another is underresearched beyond preliminary suggestions that such spaces are not hierarchical but mutually reinforcing.

The implication of thinking of wellbeing as effect is that wellbeing is rendered potentially unstable, always emergent and becoming. At the same time, wellbeing appears to manifest relative stability, at least in the short to medium term, without major disruptive life events, allowing wellbeing to be meaningfully captured through measurement. Geographies of wellbeing have begun to explore theoretical and conceptual resources not only from a relational geography, but also from Deleuze and Guattari’s work on assemblage and on striated and smooth space, from attention to embodied experience and from nonrepresentational theory (Andrews, Chen, and Myers 2014; Atkinson and Scott 2015). These conversations allow wellbeing to be an effect that is emergent and becoming, but also to be simultaneously stable and unstable, fixed and changeable. Research on how behaviors, norms, social structures, and so forth become fixed can provide insights for understanding how an emergent and becoming wellbeing may similarly be fixed. Similarly, research on how assemblages and atmospheres may be disrupted, on flows and stickiness of affects, may offer insights for understanding how a fixed wellbeing may be destabilized and amenable to change.

New directions in the geographies of wellbeing

The last few years have witnessed a rapid rise in geographical research that is engaging with wellbeing from a range of important and critical directions. Two research elements of an agenda for a geography of wellbeing to take forward are implicated from existing recent endeavors.

First, several different approaches to wellbeing have emphasized the importance of time, whether through the variations in wellbeing across a life-course, through transitional moments that destabilize wellbeing, or through the trajectories of different nation-states. Geographers have barely engaged time into experiential studies of wellbeing. A recent study by Tim Schwanen and Donggen Wang (2014) draws on the various strands of time-geography and demonstrates how the relations between social contacts and subjective wellbeing extend across multiple timescales. Geographical engagements with time–space have the potential to offer an important corrective to more static understandings of wellbeing and to contribute to elaborating the interactions of time with wellbeing in those areas where time is already identified as significant. Second, the idea of wellbeing as a collective entity has had limited traction within contemporary policy agenda. The attention to the emergence and becoming of wellbeing decenters the individual person as the primary focus of research; instead, the primary focus is on the assemblage, the atmosphere, the movement of affects, or the tensions between the habitual and disruptive and destabilizing embodied experiences. This decentering, from a geography that emphasizes emotionality, relationality, assemblage, and nonrepresentational theory, may afford the concepts and theories through which to comprehend wellbeing as collective, beyond the aggregation of individuals.
SEE ALSO: Aging; Biopolitics; Bodies and embodiment; Emotional geographies; Governmentality; Health geography; Public policy; Spatial epidemiology; Therapeutic landscapes

References


Gibson, Lorraine. 2012. “‘We Are the River’: Place, Wellbeing and Aboriginal Identity.” In Wellbeing and Place, edited by Sarah Atkinson, Sara Fuller, and Joe Painter, 201–216. Farnham, UK: Ashgate.


Williams, Allison, ed. 2007. Therapeutic Landscapes. Farnham, UK: Ashgate.

Further reading


Geographers have long-standing interests in the ecology of disease and the spatial organization of medical services. From the early 1990s, a new approach emerged, centered on the dyad of health and place. The focus on health entailed looking beyond geographers’ traditional concerns for pathological matters to encompass positive conceptualizations of mental and physical wellbeing. The concern for place involved a shift away from spatial-analytic approaches centered on distribution and diffusion toward engagement with everyday contexts that are sites of meaning and familiarity. Critically, health and place were not envisioned as separate categories, but as mutually constituted: the health status of the population and localized provision of care shape the character of place; at the same time, the physical, social, and cultural qualities of place influence individual and collective health.

These developments within geography emerged in the context of broader transformations in global public health. At a conceptual level, it was increasingly shifting away from a biomedical model centered on the causation and treatment of disease, and toward a socio-ecological model recognizing the multiple (and overlapping) social, institutional, and environmental influences on health. Related to this shift were changes in practice, including an increased emphasis on prevention and a stronger focus on primary care, often provided at the local level. Helping to underpin these changes was the growing influence of the World Health Organization’s positive definition of health as “a state of complete mental, physical and social wellbeing and not merely the absence of disease or infirmity.”

The specific case linking these changes to the need for a new approach within geography was articulated by Kearns, who argued that the established subdiscipline of medical geography was preoccupied with “spatial relationships between individuals, place and institutions, rather than with the health-related characteristics of places themselves” (Kearns 1993, 140). Specifically, medical geography envisioned “places” as locations and as regions, but not as sites of meaning that influence health experiences and outcomes. Related to this constrained understanding was a lack of engagement with humanist and structural theories of place that were influential elsewhere in human geography, and a limited ability to conceptualize the increased importance of place in the philosophy and practice of public health.

In light of these limitations, Kearns argued that the concerns of medical geography should be supplemented by a “geography of health” centered on “the dynamic relationship between health and place and the impacts of both health services and the health of population groups on the vitality of places” (1993, 145). This new approach was envisioned as having strong connections to social and cultural geography, and the developments that were animating those fields such as concerns for identity, postmodernism, and structure–agency. Two case studies were pre-
sented to sketch out potential approaches within the geography of health. The first centered on women’s experiences of planned home birth, which entailed a conscious rejection of the institutional environments so often central to medical geographers’ analyses. The second offered a new perspective on primary health-care centers, by focusing on their contribution to the collective health of community rather than on the treatments received by individual patients. Both studies were informed by qualitative methods (interviews and partial participant observation, respectively), in contrast to the quantitative methods dominant within medical geography. As Kearns argued, interpretive and qualitative approaches were especially well suited to exploring “the subtle contribution of health practices to the social geography of places” (1993, 144).

This call for change – and diversification – in geographers’ approaches to issues of wellness, illness, and care prompted considerable debate. Critically, a new body of research – and a new subdiscipline of geographical inquiry – rapidly emerged along the lines suggested by Kearns (Andrews et al. 2012). This new geography of health can be characterized in a variety of ways. For Kearns and Moon (2002), it was defined by a focus on place, adoption of theory, and critical approaches – played out across diverse thematic interests. Arguably, foremost among these themes was the idea of therapeutic landscapes, first articulated by Gesler (1992), and considered in more depth later in the entry. A review by Kearns and Collins in 2010 added concerns for individual experience, and the sociospatial constitution and negotiation of risk, to this set of characteristics. Most recently, Andrews et al. (2012) highlighted health geography’s contributions across three substantive areas of inquiry: health-care settings (encompassing political and economic pressures for reform, professional organization, and individual experience); public health (encompassing socioecological influences on population health and the role of place in health promotion); and environmental health (encompassing spatial correlations between exposures and ill-health, as well as concerns for environmental inequalities).

This tripartite conceptualization highlights the ways in which health geography retains connections to the medical geography field from which it emerged, consistent with Kearns’s original proposal for “two inter-related streams … within the medicine/health/geography nexus” (1993, 144). With regard to health-care settings, medical geographers have a long-standing interest in the spatial organization of such facilities at local, regional, national, and international scales – and in related issues of accessibility and utilization. Health geographers share a concern for medical systems and sites of treatment, but generally approach them from a different perspective – emphasizing policy frameworks, and experiences of the production and consumption of care – and are more likely to use qualitative methods and social theory in their analyses. With respect to public health, the subdisciplines have a common concern for the prevention of disease and injury, although health geographers’ contributions again tend to be distinguished by more theorized and politicized perspectives. In the third field, environmental health, subdisciplinary boundaries are particularly blurred, as quantitative approaches predominate and there is a mutual concern for identifying inequalities (Andrews et al. 2012).

These areas of overlap complicate efforts to map out a clear distinction between medical and health geography, something that is reflected in ongoing slippage in the use of these subdisciplinary labels (Kearns and Collins 2010). Nevertheless, they are typically held to differ in overall emphasis and approach. One point of contrast is the distinction between “applied”
and “theoretical” epistemologies. Medical geography has been seen to have a more applied focus, producing knowledge of both disease and health-care organization that is directly relevant to practices of disease control and systems of treatment. By contrast, work in health geography has been more likely to be overtly theorized and to incorporate a critical stance toward biomedicine. This stance has extended to criticism of geographical work perceived as being in the service of medicine (Andrews and Evans 2008). However, the tension between these two approaches should not be overstated. Health geographers have emphasized that medicine is neither monolithic nor unchanging, and that it is increasingly oriented toward health (Kearns and Moon 2002). Moreover, the broad field of health care is not reducible to medicine (Andrews and Evans 2008). In addition, the field of medical geography has continued to evolve, including via the adoption of new approaches (e.g., political ecology) and thematic concerns (e.g., socioeconomic inequality) that overlap with the concerns of health geography, as well as new methods (e.g., multilevel modeling) that help to explicate the relationship between health and place (Mayer 2010).

Diverse interests

One of the distinguishing features of health geography is its remarkable breadth, encompassing an array of substantive concerns, theoretical positions, and methodological approaches. This diversity stems from a number of factors: connections to other social science perspectives on health, which have also proliferated since the mid-1990s; the magpie-like tendencies within the parent discipline, manifest in a willingness to adopt and adapt ideas from other fields; and the flexibility of health geographers themselves in selecting projects and approaches. It follows that health geography is characterized not only by enduring links to medical geography but also by connections to myriad other fields of inquiry that speak (empirically and/or theoretically) to the interconnection of health, care, and context. A useful conceptualization of the diversity within health geography is provided by Andrews et al. (2012), who identify three categories of difference. They note it has contributed to specific subjects of in-depth inquiry (e.g., complementary medicine, intoxicants, sport and fitness), to established fields of knowledge (e.g., aging, mental health, disability), and to overarching research themes within the social sciences (e.g., gender, the body, globalization). In making these diverse contributions, health geographers have necessarily engaged with work located in other academic disciplines (e.g., disability studies, sociology, gerontology). To illustrate these points, one subject area receiving increasing attention within health geography is smoking, with studies exploring spatial variations in prevalence and levels of restriction, as well as place-based meanings of smoking and exposure. This research has brought health geographers into contact with particular fields (e.g., health promotion, social marketing, and tobacco control) and with larger questions about power, rights, and inequality. It has also involved engagements with other disciplines, particularly via theories of stigmatization and denormalization adopted from sociology and social psychology, respectively (Collins and Procter 2011; Pearce, Barnett, and Moon 2012). The variety of theoretical approaches utilized in health geography (discussed in further detail later in the entry) speaks to the flexibility of researchers as well as the porous nature of subdisciplinary boundaries.

This porosity is linked, in part, to the positioning of health geography within the domain of public health. As Andrews et al. (2012, 358)
HEALTH GEOGRAPHY

outline, public health encompasses “a rather monstrous assemblage of laws, bureaucratic institutions, health professionals, materials and practices which aim to prevent disease and injury, create healthy environments, all for the broader purpose of securing and promoting the general health of the population.” Health geographers have engaged with this “monstrous assemblage” in two broad ways: first, as participants seeking to contribute to its overarching purpose; and, second, as critics seeking to elucidate its assumptions, silences, and consequences. In the former role, health geographers have explored the ways in which health-related resources, behaviors, and socioeconomic structures vary across diverse physical and social contexts – and how these contexts might be reworked to support public health goals. Neighborhoods have been a prominent area of interest, with studies connecting their socioeconomic characteristics and built form to variations in levels of alcohol use, active travel, outdoor recreation, obesity, smoking, and so on. In the latter role, health geographers have considered public health as an object of critical inquiry, exploring the medical, social, and political dimensions of its historical development, as well as its contemporary functions across population surveillance, health promotion, risk reduction, and policy intervention. Much of this work draws on theoretical insights originating with Michel Foucault (1926–1984) to examine the role of public health in organizing and governing bodies, space, and power.

In both of these capacities, health geographers contribute to themes and concerns developed across many fields of academic inquiry, spanning the humanities, social sciences, and natural sciences. In particular, they have engaged with contemporary public health research, which since the adoption of the socioecological model in the 1970s has been attentive to diverse contextual factors (e.g., relationships, institutions, communities, policies) shaping individual and collective experiences of health. One consequence of a heightened concern for context is a “spatial turn” manifest in converging concerns for the health implications of home, workplace, neighborhood, city, and region (Andrews et al. 2012). Moreover, opportunities for – and expectations of – cross-disciplinary engagement exist at many stages of the research process (e.g., in developing funding proposals, presenting at conferences, and submitting work for peer review). It follows that health geography is by no means an insular endeavor. The same is broadly true for medical geography, with Mayer (2010) emphasizing interconnection with fields as varied as bacteriology, social epidemiology, and development studies.

A central theme

In light of these observations about diversity, it is reasonable to ask whether there are points of commonality that give health geography a sense of identity and coherence. The defining feature of the subdiscipline is its emphasis on “the centrality of place in the study of health” (Kearns and Collins 2010, 16), which provides a shared focal point for inquiry, although there are multiple interpretations of both key terms. In first proposing a focus on health and place, Kearns (1993) conceptualized a two-way interconnection: place both influences health and is partly constituted by the provision of local health services and the health status of the local population. Over time, the former relationship has received substantially more attention. Indeed, perhaps the most common question asked in health geography scholarship is “How does (this) place affect health?” While this work is often qualitative in design, there is a sense in which it is concerned with causation and envisions health as a dependent
variable – if only implicitly. This approach confers some benefits in that it connects health geography to broader concerns for the determinants of health. It also allows for place to be envisioned as a site of therapy as well as a site of risk (Andrew et al. 2012). However, there is a sense in which concern for the connections from health to place has been overshadowed, notwithstanding some studies considering the ways in which health-care institutions seek to create place identity and contribute to a broader sense of community.

In terms of how place is conceptualized, three distinct understandings have been identified (Kearns and Moon 2002; Kearns and Collins 2010). First, a large body of work examines particular localities, both urban and rural, for the ways in which they support or undermine health. Within these locality studies, approaches vary from humanistic explorations of the meaning of place to positivist examinations of exposure to localized risk factors, such as air pollution. Andrews et al. (2012) note that studies in low- and middle-income countries often focus on environmental issues and policy interventions in places, whereas those in high-income countries tend to involve more theoretically oriented investigations of place effects on health. Another characteristic of this work is that a great deal is known about certain places and very little about others. This selective focus is evident across multiple scales. Internationally, health geography research has been concentrated in high-income, predominantly English-speaking countries – the subdiscipline’s “Atlanto–Antipodean hearth” (Kearns and Moon 2002, 608). Moreover, within these countries, places proximate to university-based health geography research programs have received the most attention.

Second, landscape analysis has been a hallmark of health geography. The term “landscape” is generally interpreted as referring to the converging layers of history, social structure, and built environment at particular sites. The most prominent approach follows Gesler’s (1992) identification of therapeutic landscapes – a concept that has contributed substantially to the profile and purchase of health geography. Originally an overtly theoretical approach that drew on humanist, structural, and cultural materialist interpretations of landscape, it was subsequently applied in empirically oriented investigations of places as varied as “natural habitats, towns, beaches, retreats, clinics and homes” (Andrews et al. 2012, 354). This proliferation of case studies has prompted some criticism, including that the “therapeutic landscapes” label may be applied in descriptive and uncritical ways. Another concern is the lack of specificity in identifying the qualities of place that support restorative experiences and enable health (Duff 2011). Health geographers have also conceptualized landscapes that have a more ambiguous relationship with health. This approach has included adopting Dear and Wolch’s (1987) understanding of landscapes of despair to provide insights into geographies of mental health and addiction that emerged after the widespread deinstitutionalization of the 1970s and 1980s.

A third distinctive perspective on place is evident within multilevel modeling studies. The expansion of work in this area reflects the refinement of statistical methods, increasing availability of large datasets, as well as “an enduring quantitative tradition based around the application of the generalized linear model to questions of health equality/inequality” (Kearns and Moon 2002, 611). It conceptualizes place as the context (often understood in principally socioeconomic terms) in which individuals are situated, and seeks to attribute variations in health status and experience to these different levels. This work is often conducted at the
neighborhood scale, where unpacking the relative importance of context (the characteristics of place) and composition (the characteristics of residents) is of central analytical importance (Macintyre and Ellaway 2003). While the understandings of place produced by this multilevel modeling are in part a function of data availability, the relative complexity of its statistical methods contributes to its purchase in a broader research environment that privileges “robust” quantitative approaches (Kearns and Moon 2002). This environment has been shaped by the rise of evidence-based health care, which has privileged certain quantitative methodologies.

In terms of conceptualizations of health, there has been a broad shift away from classical biomedical preoccupations (e.g., pathogens and patients) toward a more contextual, collective, and positively oriented understanding of the term (Kearns and Collins 2010). This shift is reflected in attentiveness to the social/cultural, political, built, and natural components of place-based communities; in concern for identifying and promoting health-enabling places and policies at multiple scales; and in an overarching interest in protecting the wellbeing of the entire population (a preventive rather than curative focus). Widespread adoption of the term “wellbeing” within health geography signals a desire to frame health as more than a state of absences, for example as an empowering resource or capacity. This is not an idiosyncratic approach, as more holistic and positive conceptualizations of health have accompanied the adoption of the socioecological model and an increasing emphasis on population health perspectives within both public health and medicine.

A focus on health and wellbeing helps to connect health geography to important empirical developments and a changing landscape of health research, but it also raises questions as to whether health geographers can or should continue to engage with pathological matters and places of (conventional) health-care provision. In practice, there has been considerable emphasis on individual experiences of illness and wellness (encompassing not only physical health issues but also mental and social wellbeing, as well as disability), and individual accounts of finding and receiving health care. Recent examinations of the accounts of medical tourists extend the latter theme in an international direction. Work in these areas is interpretive and experiential, speaking to the enduring influence of humanism within health geography; at the same time, the focus remains “‘downstream’ of health care work” and thus it seldom engages directly with health-care providers or clinical practices (Andrews and Evans 2008, 762). Indeed, a subdisciplinary desire to be “postmedical” risks a disconnection from (and even dismissal of) the domains of biomedicine and health care. This tendency is problematic, because it narrows research opportunities, potentially blinding geographers to the ways in which these domains continue to evolve to promote health across multiple settings.

Theoretical perspectives

As described, the emergence of health geography implied a shift in the theorization of place and health by geographers interested in culturally sensitive, “postmedical” perspectives. In this vein, health geographers have conceptualized both “place” and “health” as constructed and experienced phenomena rather than understanding place as a neutral backdrop against which events unfold, and health as something that can be measured as a state of absences (in terms of the incidence of disease or mortality, etc.). These shifts reflected a wider “theoretical turn”
that began decades ago in the parent discipline of human geography as well as more recent theoretical reformulations in subdisciplines such as cultural geography that placed greater emphasis on process, power, resistance, and identity (Gesler and Kearns 2002).

What has distinguished health geography since its inception is a reluctance to take place and health for granted: both cultures of place and cultures of health matter when it comes to health experiences and health-care provision. This orientation, paired with a good deal of openness to philosophy and theory from outside the discipline of geography, has stimulated the theoretical diversification that health geography is known for today. The self-conscious application of social and cultural theory to place and health, and an attraction to “critical” theory in particular, has helped to distinguish health geography from the primarily positivistic subfield of medical geography. This theoretical “equipment” has provided a greater sensitivity to place-making, its intersection with multiple axes of difference (such as class, ethnicity, race, gender, and ability), and its role in shaping personal experiences of illness and disease, social exclusion, and health inequalities.

Health geography’s theoretical diversity can be organized into five general perspectives that reflect selective engagements over the past 20 years. First, the most enduring engagement with social theory by health geographers has been with perspectives that conceptualize social differences such as class, gender, and age. In particular, structural perspectives (e.g., political economy) have provided the theoretical basis for scholarship linking social inequalities in health to place. For example, Barnett, Pearce, and Moon (2009) show how in New Zealand, a country experiencing rapid restructuring, smoking cessation is more evident among European-background populations than Maori populations, and that cessation may be mediated by levels of relative deprivation in local environments. Second, theoretical engagements with humanistic perspectives that draw attention to personal experience and symbolic meaning of place have also been foundational. Taking cues from philosophies of meaning, such as phenomenology and existentialism, health geographers have highlighted the role of human agency in the creation of “lifeworlds” relevant to health and wellbeing.

Third, attempts to integrate these two perspectives represent another important theoretical perspective. This endeavor has entailed a search for theoretical balance between the constraints laid down by social forces and the freedoms individuals exercise in making decisions and taking action in the world. The potential value of such an approach was initially signaled by Kearns (1993), who called for engagement with the structure–agency dynamic that was animating many areas of social science inquiry. Important applications of this theory included Dear and Wolch’s (1987) examination of the ghettoization of discharged psychiatric patients within deprived inner city neighborhoods in North America. Subsequently, health geographers analyzed health-related behaviors and experiences as the product of the interaction of social structures and individual agency in particular locales. Today, the structure–agency debate is less prominent but its tenets still influence health geographers interested in landscapes of care, the dynamics of marginalization, and health promotion.

Fourth, health geography has been heavily influenced by postmodern and poststructural perspectives. Postmodern theories have been particularly significant given their emphasis on diversity in social and cultural life and their skeptical stance toward universal narratives that leave little room for difference. Postmodernism’s predilection for local knowledge has resonated
strongly with health geography’s quest for place-sensitive accounts of health and health care. Poststructural theories have been important to health geographers insofar as they problematize social categories such as gender, race, and class by showing how they are dependent on processes of differencing (distinguishing between like and different) and negation (privileging like over different) that are grounded in – and proceed through – space and place. Imbued with power relations, these systems of sociospatial exclusion provide individuals with sources of meaning and identity. This particular application is well executed by health geographers exploring “postasylum” geographies. Parr’s (2008) work demonstrates the strengths of a poststructural perspective. It explores how mental health clients experience community-based care; particularly its emphasis on sociospatial othering of people with mental differences and clients’ related attempts to rescript their lives in the community, demonstrates the strengths of a poststructural perspective. Here, the ideas of Foucault have been highly influential. Concern with discourse, as a medium of knowledge and power, is common within health geography, especially in analyses of landscapes and health policy. Foucault’s conceptualization of power, and its links to notions of discipline and governmentality, have received particular attention. Governmentality theory has found application in critical examinations of public health where it has been used to illuminate the political rationalities and power modalities invested in both run-of-the-mill and exemplary health policies.

Fifth, health geography has come to be influenced, more recently, by what can be called “relational perspectives” (Gatrell 2005). These perspectives are connected, by way of extension or critique, to the poststructural theories introduced earlier. They view individuals as embedded in multiple networks across time and space, and draw attention to the processes and interactions between people and social and material resources in the environment over time. One example is “complexity theory,” which in the context of health geography can be used to understand disease ecologies, health inequalities, health care, and health policy as complex systems characterized by emergent, nonlinear, and self-organizing interactions between a large number of elements, both human and non-human. Another example is actor-network theory, which treats phenomena such as “health” and “place” as enacted and relational effects of particular webs of association between people, material objects, and discourses (Duff 2011).

While each of these five perspectives brings strengths to health geography, various gaps remain. The status of the body, particularly its materiality, has maintained an uneasy and underexamined presence within health geography. In the early 1990s the notion of “body politics” (addressing the body as a site of representation, identity, and experience) was part of the debates that launched health geography; yet, a fully embodied health geography has been slow to develop. Writings on the body by health geographers have focused primarily on issues of representation and the body’s role as a site of medical inscription. The importance of the real and material physicality of the body has in some ways been overlooked, existing only in the shadows of humanistic and poststructural approaches.

A final dimension of theory in health geography pertains to the idea of “criticality.” On the one hand, invocation of critical theory has signified opposition to health inequality and an affinity with social justice and progressive social change. For example, critical geographies of health have included scholarship contextualizing health inequalities in specific localities, often linking these health outcomes to state restructuring (Barnett, Pearce, and Moon...
2009). On the other hand, the invocation of critical theory points toward interrogations of the very relationship between health and society (Parr 2004). For example, critical geographies of health have included Foucauldian-inspired scholarship on new public health, which illustrates how populations and individuals are managed and regulated through health discourses and related practices (Brown and Duncan 2002). While matters of criticality are often associated with cutting edge theories from outside the discipline, “applied” quantitative approaches are, in many ways, equally critical and instrumental to social change (Kwan and Schwanen 2009). Finally, many health geographers have come to practice criticality beyond academia via community engagement and, on occasion, overt political activism. In this regard, Parr (2004) distinguishes between praxis and critical thinking, the former referring to forms of activism and the latter referring to ideas and theoretical possibilities that open doors for new forms of praxis.

Methods

Health geography encompasses a wide range of research approaches, techniques, and materials. Census data and population health data gathered through randomized surveys are commonly used, as are quantitative techniques such as statistical and mathematical modeling approaches. These methods are typically combined with structural perspectives to quantify the relationship between place and health. In addition, maps and geographic information systems (GIS) have been instrumental in visualizing spatial variations in health and disease. As far as methods are concerned, however, health geography is renowned for the utilization of interpretive approaches that rely on qualitative techniques to gather together a wide range of texts and images for analysis. Given their sensitivity to context and meaning, qualitative approaches are well suited to a cultural focus on place and health. They are also aligned with health geography’s interest in the everyday contexts of health within and outside of formal spaces of medical care.

Common methods include interview and focus group techniques that allow people to speak in their own words. For instance, they have been used to decipher people’s views of environmental risks in the context of particular locales, and to decenter medicalized approaches to illness, health, and care, making room for subjugated knowledge and resistance. Whereas these talk-based methods offer insights into the meaning of participants’ experiences, more immersive techniques, such as observation, involve direct engagement with health-relevant settings. Observational techniques allow health geographers to construct interpretations of particular contexts through physical presence, and varying levels of participation in activities and events (Gesler and Kearns 2002). Bodily differences in terms of race, age, or ability can present barriers as far as interpersonal relations are concerned, and participating in field settings can require researchers to present themselves or to perform their bodies in challenging ways. Health geography has remained particularly sensitive to the complexity and politics associated with “fieldwork.”

Textual methods represent another common approach utilized by health geographers, and include a wide range of specific theoretical approaches including content analysis, semiotic analysis, and critical discourse analysis. They have also engaged with a diverse array of “texts.” Everyday texts from popular culture (e.g., film, literature, music) have served as rich data for unpacking health-place relationships. Health policy, both past and present, has also been an important data source. For example, health
geographers have analyzed health policies for the ways in which they represent the body in relation to broader ideological systems and associated power relations. Finally, landscapes themselves have served as important sources for textual analysis. Here landscapes are understood as sociocultural repositories which can be read as texts using methods such as semiotic analysis or critical discourse analysis. Such readings have been particularly influential in the development of the therapeutic landscapes concept, as well as in work interrogating the consumer landscapes of contemporary health care.

While health geography has a special affiliation with qualitative methods, quantitative approaches remain important. Moreover, as health geographers have engaged with specific theoretical debates, such as structure–agency or the complexity of health phenomena, they have been more inclined to utilize mixed method approaches that combine quantitative and qualitative data and analytical techniques. When combined, quantitative and qualitative methods may offer an especially comprehensive approach to understanding the richness of place and its relationship to health; however, the (in)compatibility of their underlying epistemologies – positivism and postpositivism – raises challenges, particularly in research seeking a “postmedical” orientation.

Finally, connected to health geography’s kinship with critical geography has been engagement with research approaches that seek to empower participants (Andrews et al. 2012). It is becoming more common for health geographers today to see their own vocation as being partly academic and partly activist, leading more health geographers to engage directly with institutions in civil society using participatory action research frameworks. These frameworks democratize the research process by involving, even empowering, those typically regarded as “subjects” in conventional research designs. Moreover, approaches such as “arts-based” methodologies involving film and photography have opened new doors to collaboration between geographers and research participants.

Key foci

Health geography encompasses an array of substantive concerns, theoretical positions, and methodological approaches, which give it remarkable breadth. To help give this diversity and pluralism some order, it is helpful to consider two key foci in health geography studies. First, health geography has advanced understandings of environment and health relationships, generally reflecting a fundamental concern about “unhealthy places.” Second, health geography has advanced understandings of therapeutic settings, generally reflecting a fundamental interest in “healing places.” In some ways, these key foci speak to the traditional concerns of medical geography – disease ecology and access to health care – and their continued relevance. They could also be interpreted as reflecting a broader tendency to see the environment itself in terms of either risk or care (Andrews et al. 2012). The story of health geography is very much about the myriad ways these original foci, as well as themes such as risk and care, have been modified through a reformulated understanding of place (as a complex cultural and symbolic phenomenon constructed through relationships between people and their everyday worlds) and expanded understandings of health (understood, for example, using socioecological frameworks or other theoretical perspectives).

The first focus, environment and health, encompasses fundamental interests in the various pathways by which environments deleteriously
affect health. This scholarship can be organized into three categories: environment and disease, environmental health, and environment and health inequality. Gathered together under environment and disease is scholarship dealing most directly with the relationship between contagious disease diffusion and natural and social environments. Environmental health scholarship, on the other hand, draws on perspectives such as environmental epidemiology to focus attention on the susceptibility and exposure of individuals to environmental pollutants in the air and water. Here one also finds a concern about the inequitable distribution of environmental risks, and political commitments to environmental justice. Finally, environment and health inequality refers to scholarship interested in the linkages between the kinds of places people occupy (e.g., deprived neighborhoods), health-related behaviors, and social inequalities in health. Cutting across all three categories is the important theme of risk. Whether it be neighborhoods, environmental contaminants, or behaviors, ascribing risk is an important practice in the “new” public health movement by which populations are governed.

The second focus, therapeutic settings, encompasses a wide array of research concerned with the healing process and how it unfolds in particular situations, settings, and milieux. Smyth (2005) has organized this scholarship into three categories: therapeutic places, therapeutic spaces, and therapeutic networks. Gathered under therapeutic places is the burgeoning scholarship undertaken in the therapeutic landscapes tradition, which seeks to understand how certain places acquire and sustain their reputation for healing. Therapeutic spaces encompasses varied scholarship on spaces of health care concerned with issues of accessibility, utilization, and more culturally sensitive work interested in the influence of consumerism on health-care landscapes. Work on therapeutic networks includes scholarship exploring the informal arrangements of care that exist outside biomedicine and draw on actors in civil society. A prominent theme cutting across therapeutic places, spaces, and networks is the notion of care. Acts of service, guidance, or protection engender diverse landscapes of compassion at multiple scales. Hence, for geographers care is something that simultaneously invokes moral feelings, interpersonal connections, and relationships between people, spaces, and places. In other words, health and wellbeing are always shaped by, and situated within, broader landscapes of care.

Increasing engagement with notions of care speaks, once again, to health geography’s breadth. Yet, as with any subdiscipline, certain limitations and silences remain. First, the field’s origins in high-income, predominantly English-speaking countries continue to delimit inquiry in certain ways; engagement with themes of global health and development is limited and conversations with work situated in other language traditions (e.g., French health geography) remain relatively rare. Second, it has arguably been slow to incorporate the insights offered by recent developments in social theory, such as those related to mobilities and posthumanism (the latter offers a potentially valuable counterpoint to the anthropocentrism inherent in health geography). Third, in terms of methodology, historical approaches remain scarce, and much could be gained from situating the health–place dyad in historical perspective. This said, the rapid development and diversification of health geography since the mid-1990s speaks to an inherent flexibility and an ability to respond to such challenges and opportunities.

SEE ALSO: Caregiving; Environment and health; Health inequalities; Medical tourism;
HEALTH GEOGRAPHY

Mental health geographies; Neighborhoods and health; Therapeutic landscapes

References


Health inequalities are differences among human populations in physical or mental health that are associated with sociogeographical position and environment. Geographical inequalities in health exist both between and within countries at multiple spatial scales, ranging from differences between world regions and countries at the global scale to the level of local neighborhoods and individual people. For example, at the global scale, the World Health Organization (WHO 2014) publishes, in its Global Health Observatory Data Repository, information on the health of populations in different countries around the world. Data for 2011 showed that age standardized death rates per 100,000 population (which control for differences in the age structure of populations) varied from less than 350 in relatively highly developed countries, such as Japan, Switzerland, and Italy, to values over 1950 in countries with low levels of human development, such as Sierra Leone, Lesotho, and Swaziland. Even within regions of the world, international health inequalities are clearly evident. For example variations among European countries are reported (Richardson et al. 2013). Furthermore a study by Mackenbach et al. (2008) demonstrated that inequalities in mortality within national borders were relatively small in some southern and Scandinavian European nations but large in most countries in the eastern and Baltic regions, suggesting that the processes contributing to health inequalities vary among countries in Europe.

Health inequalities among countries are associated with differences in socioeconomic conditions and levels of human development. Evidence of this type points to the existence of a social gradient in health such that social groups (and geographical areas) that are more deprived typically have worse health than those that are more advantaged, and health outcomes worsen progressively in association with worsening social conditions.

Health inequalities within populations of countries are, in many cases, seen to be persistent or widening, even in situations where average health of the population is improving. Such persistence of health inequalities occurs despite knowledge that they are, to a large extent, avoidable. Health inequalities can be reduced if societies, communities, and individuals invest effort and resources to address health inequalities by tackling the causes of ill health, especially for those who are most disadvantaged. Health inequalities therefore represent a significant challenge for societies globally and often constitute a form of injustice, which is why differences in health in human populations are sometimes referred to in the literature as health inequities, rather than as health disparities.

Geographical perspectives on the “determinants” of health inequality

In geography, as in other disciplines focusing on health inequalities, research focuses on the “determinants” of health: processes which contribute to health variation. Some of these
determinants are associated with geographical factors that interact with health-related practices of individuals and population groups. This association is illustrated, for example, in research on whether unequal access to environments offering opportunities for healthy lifestyle choices may contribute to differences in healthy behavior; however, individual health-related behaviors are not the only determinants of health inequality.

To some extent health inequality is also influenced by varying access to, and use of, health services relative to need for care (e.g., see Gatrell and Elliot 2009, Chapter 5). Some of the earlier work in countries such as the United Kingdom drew upon the idea of an inverse care law, which argued that provision and access to health care is worst in areas where people have poorer health and need to use health care most. There is a long tradition in health geography of research focused on sociogeographical inequality in access to health services within countries (e.g., Joseph and Phillips 1984; Meade and Earickson 2000). Furthermore, differences in health-care systems also have important implications for inequalities in access to health care. There is an international debate, for example, over the extent to which universal health-care coverage can help to reduce health inequality.

More fundamentally, health inequalities depend on social and geographical variation in nonmedical determinants of health that are not limited to variations in individual lifestyle practices and use of health care. These determinants encompass dimensions of the natural environment (e.g., air, water, and soil quality, access to green space) and of the built environments (e.g., housing quality and built infrastructures such as roads or public spaces in cities). They also include social conditions such as social cohesion, social capital, income, and education. Wider health determinants arise from broader processes such as global climate change, ecosystems, and economic processes, as well as finer-scale differences in living conditions and social relations in local communities. Curtis (2004) emphasizes, from a geographical standpoint, how these different aspects of health determinants create interconnected "landscapes" of inequality in physical and mental health, associated with processes such as unequal power relations, conditions of poverty and wealth, disparities in consumption and commodification of health care, geographical disparities in the quality of natural environments, and varying psychological and symbolic values attributed to geographical settings by different social groups.

There is a growing literature on how to define geographical indicators of inequality in health determinants, as well as measures of population health outcomes, and how these might be used to assess levels of inequality in healthy environments for human populations. Such indicators may be used to assess varying “need” for medical care in the population and can be used to help determine a fair allocation of resources in relation to need for services. In addition to the choice of components for such indicators, related methodological issues from a geographical perspective include the significance of the geographical scale of analysis and the units used to assess health variation and associated health determinants.

Interpretations of health inequalities

The literature on health inequality among geographical areas has also given rise to debates about how to interpret these differences and the complexity of the processes that may lead to health inequality. For example, arguments have focused on how far we can interpret geographical health inequalities in terms of: vulnerability versus resilience; context or composition; health selection; or socially, culturally, or individually contingent and
relational interpretations of what constitutes health and its determinants.

Much of the literature is consistent with the idea that greater disadvantage in terms of health determinants is associated with greater health disadvantage; however, this relationship is not universal and there is a growing body of geographical research exploring “health resilience” among deprived populations. This research illustrates how populations in some socioeconomically deprived areas enjoy somewhat better health than might be expected, given the level of socioeconomic deprivation that they suffer (e.g., Cairns, Curtis, and Bambra 2012). Work in this field offers important new insights into strategies that may help to mitigate the impact of challenges to health for disadvantaged groups. It also underlines that it is not necessarily helpful for health geography to represent deprived individuals and communities only as “weak” or “vulnerable,” since they often display remarkable strengths as well as disadvantages.

Geographers have explored the value of compositional and contextual explanations as ways to make sense of health inequalities. Compositional interpretations consider that health inequalities arise from uneven geographical distribution of individuals with better or worse health. Contextual explanations argue that characteristics of the physical or social environment in an area affect the health of all individuals in the area to some extent, and that health determinants can be attributed to characteristics of places as well as people. There has been an ongoing debate over the relative importance of compositional (individual) and contextual (area) factors for explaining place effects on health and health inequalities. More recent interpretations of these arguments suggest that the idea of “contextual effects” is valuable in that it provides a way to draw attention to the significance of places for health; however, it is theoretically difficult to separate out independent “compositional” and “contextual” processes and it is probably more helpful to think in terms of various interactions between individuals and their environment that are associated with health inequality.

Another stream of international literature has focused on the idea that health selection may help to account for social and geographical health variation (health inequalities among socioeconomic groups may be partly explained by unequal social and economic chances over the life course for people with health problems, resulting in selection into disadvantaged groups of those in worse health). Also, it is argued that, in part, the geographical variation in population health may arise because those in worse health tend to be “trapped” in disadvantaged areas, or to move into these places, while those initially in better health tend to move to wealthier areas.

The significance of social and cultural differences in perception and experience of health, and of social and physical environments, also leads to interpretations of health inequalities and their determinants which invoke relational interpretations of space and place (Cummins et al. 2007). Geographers have emphasized the contingent and variable nature of our experience of landscapes that may be variably therapeutic or damaging for health for different individuals (Conradson 2005). Viewed from these perspectives we can understand that health inequalities need to be understood in terms of very diverse individual lived experiences, as well as at the aggregate population health level.

The research on health inequalities is important for health policies designed to address health inequality. It draws attention to the need to address inequalities in (ill) health, and access to and use of remedial services, and also informs policies designed to “reshape” the physical and social environment in ways that will help to
HEALTH INEQUALITIES

reduce the geographical and social inequalities of health determinants.

SEE ALSO: Accessibility, in transportation planning; Environment and health; Health geography; Health and wellbeing; Human geography; Neighborhoods and health; Population geography; Social justice; Spatial epidemiology; Vulnerability

References


Cummins, Steven, Sarah Curtis, Ana V. Diez-Roux, and Sally Macintyre. 2007. “Understanding and Representing ‘Place’ in Health Research: Relational Approaches.” Social Science and Medicine, 65(9): 1825–1838. DOI:10.1016/j.socscimed.2007.05.036.


Health systems

Mark W. Rosenberg  
Queen’s University, Canada

Health systems are generally among the most complex public, private, or not-for-profit systems in a country. Their complexity stems from competing theoretical perspectives, the goals of the health system, the legislative and regulatory frameworks in which they are constituted, the people who provide the services within them, where those services are provided, and the totality of the services. An alternative way of conceptualizing health systems is to think about them in terms of the underlying values on which they are based. From the definition of legislative frameworks to how and where services are provided, to the values that underlie them, health systems are manifestly geographically expressed and infused with geographic issues.

The components of a health system

Two competing theoretical perspectives dominate thinking in the discussion of any health system. On the one hand, there is the desire to design health systems that maximize some notion of equity, usually framed as providing all necessary medical services to everyone within a society, regardless of their socioeconomic status. What is meant by necessary and for whom will vary from jurisdiction to jurisdiction. Competing with this theoretical perspective is the desire to design health systems that are efficient. In those health systems that are mainly constituted around private provision, efficiency usually means some notion of maximizing revenues and minimizing costs to the providers of services. In those health systems that are mainly constituted around public provision, efficiency criteria usually focus on minimizing costs but often ignore or treat the maximization of revenue as a secondary outcome. Minimizing the fees that patients pay is also a goal in some health systems, but mainly in those that stress equity over efficiency and/or where public provision is emphasized. The efficiency/equity debate is one well known to geographers, especially in research on health systems that examine the location and allocation of health services (Tanser, Gething, and Atkinson 2010). For example, if the goal is to emphasize equity in the location and allocation of hospital services, then the solutions generally focus on covering the total population and minimizing the average distance traveled by patients. If the goal is to emphasize efficiency, then the solutions generally focus on maximizing the number of people covered within a set distance of a facility, even if this is at the expense of some part of the population being beyond the set distance. The efficiency/equity debate is also part of larger debates within geography about what constitutes social and spatial justice (Rosenberg 2014).

The second level of complexity comes from the definition of the goals of a health system. While at a global level all member countries of the World Health Organization (WHO) share an agreed definition of health based on the Alma Ata Declaration, operationally no such shared definition exists when national and lower-level (state, provincial, or regional) governments articulate the goals of their health systems. Operational goals depend on whether the focus is on curative,
Health systems are constituted through legislative and regulatory frameworks. Legislative and regulatory frameworks depend on whether the health system is centrally planned or planned at various levels of government. For example, in the United Kingdom, the National Health Service (NHS) is defined through national legislation to provide health care throughout the United Kingdom on a uniform basis. In contrast, the Canadian constitution devolves the responsibility for health care to the provincial and territorial governments, except for health care for the military and Aboriginal peoples under some circumstances. The federal government, however, plays a role in setting some standards through the Canada Health Act, but mainly the provincial and territorial governments define their health systems through their own individual legislation. In essence, the NHS can be thought of as one geographically based health system, whereas in Canada there are ten provincial, three territorial, and two federal health systems, each with its own internal geographies.

Equity versus efficiency, the mix of goals, and central planning versus regional/local planning are layered within the triangle of how people pay for their care. Generally, health systems are distinguished by whether a person pays privately for care, the government covers health costs through a public health insurance system, or a person’s health costs are paid for through a not-for-profit entity (e.g., a religious organization or a union). In reality, it is rare to find a health system which is purely private, public, or not-for-profit, even if health systems are often labeled as such. For example, the health system of the United States is often referenced as a private health system and, indeed, private-sector principles dominate the thinking about how the US health system is organized. The US health system is, however, not purely private in that millions of US citizens are insured through Medicare and Medicaid (federally funded public health insurance systems for the older and poor populations, respectively), veterans often receive their care through Veterans Administration Hospitals (i.e., publicly funded hospitals), and many people using either private or public health insurance receive their care from not-for-profit hospitals (often operated by religious-based organizations). In contrast, the Canadian health system is generally described as a public, universal, comprehensive health system but there are some parts of it that are not publicly insured (e.g., cosmetic surgery or dental care for most people). As developing countries switch from private to public health insurance schemes, either as substitutes or as complementary systems, health geographers are contributing valuable research to explain why some people and areas are enrolling in the new public health insurance schemes and why other people and areas are generally retaining the old
HEALTH SYSTEMS

systems of paying and choosing traditional health systems over “Western” medicine.

The private/public/not-for-profit triangle is a fourth level of complexity which manifests itself spatially in how geographers think about economic access, social cultural access, and geographic access to care (Ricketts 2010). The direct barriers to economic access are generally related to the inability to pay for a health service, prescription drugs, or medical devices (e.g., a properly fitted wheelchair). Sociocultural access barriers are often measured through ethnicity, race, immigration status, cultural and religious values, and the ability to communicate in the official language(s) of the health system. Geographic access has mainly been viewed in terms of time or distance to the nearest health service, and/or the density of health services within geographically defined neighborhoods, spatial statistical units (e.g., census tracts), or administrative regions (e.g., health regions). Health geographers have focused on whether geographic areas where people can be characterized as having the least ability to pay for health care are the areas where people have the lowest utilization rates, and poorer health status, and whether those areas have fewer health services within or near them. Social and feminist geographers have focused on issues such as the concentration of women in the poorest-paying parts of the service sector and how being in the low-waged service sector constrains their everyday lives. For example, the inability to get paid for time off to see a physician (i.e., an indirect economic barrier) has many ramifications for access to health services. Understanding indirect economic barriers to health remains a research area in which health, social, and feminist geographers might do well to collaborate to understand the complexity of health systems.

A fifth level of complexity stems from how nations define their health systems’ labor forces. These have evolved from relatively simple conceptions based on physicians and nurses to conceptions which now include general practitioners (or family practitioners), medical specialists, various types of nurse specialist, nurse practitioners, midwives, dentists, various types of technician, physiotherapists, occupational therapists, mental health therapists, health promotion workers, emergency first responders, and various types of alternative health care and traditional medical practitioner. The geographic variations in practitioners within health systems have been subjects of two types of study by health geographers. The most common type has been to determine the ratio of practitioners to a population for a given set of geographic areas (e.g., the number of physicians per population per county) to determine whether geographic areas are over- or underserved, and what factors explain the geographic distribution of practitioners. Early studies seeking to explain the geographic distribution of medical practitioners in particular were mainly carried out using ecological modeling. More recently, weighted spatial regression and spatial autocorrelation models have been employed to explain the underlying factors leading to over- and underserved areas. Underpinning the interpretation of underserved areas is the assumption that a spatially equitable health system is one where there are no under- or over-served areas.

Closely linked to how the labor force is defined is the question of how services are defined and where those services are located within a health system. Health systems are often conceptualized as being inherently hierarchical, both organizationally and geographically. Conceptually, the most medically specialized care is found only in tertiary care hospitals. These are usually the largest hospitals found mainly in the largest cities within a country, and are often associated with university medical schools. Following the
HEALTH SYSTEMS

hierarchy of the urban system, large cities have secondary care hospitals equipped and staffed to carry out many forms of specialist care, and then smaller cities and towns have secondary care hospitals only equipped and staffed to carry out basic emergency and surgical care. The smallest towns are likely to only have primary care health centers, or sometimes even only primary care physicians practicing in solo or small group practices. In developing countries, or even in remote locations within developed countries, there might not even be primary care physicians but only health care centers staffed by nurse practitioners or health care workers specially trained to provide basic services in rural and remote communities. Health geographers often contribute to the debates about what are the appropriate services and levels of service for a community, given its size and place in the urban system or as a rural or remote community.

The other way health services and labor forces are organized is through regionalization. For example, from its inception, the NHS was organized based on regional health authorities deciding on how resources allocated from the central government should be deployed. In many countries, health geographers have participated in debates on whether regional health authorities should be implemented, in contrast to central planning, on the basis of both efficiency and equity criteria. Those geographers who favor regionalization often argue that regional decision-makers are in a better position to understand the needs within their regions than central decision-makers, who tend to favor the largest urban places where they are based. These types of argument are often framed in terms of local equity versus central inefficiencies. Those geographers who favor centralization often argue that only through centralized decision-making can a government ensure that everyone in its jurisdiction receives the same benefits from the health system, regardless of where they live. They also argue that rural and remote regions often do not have the expertise or the resources to deliver higher-order health services, and that small population sizes make it inefficient to fund higher-order services which will rarely be used. Health geographers who frame their arguments in these terms are often arguing for a notion of overall equity and against local inefficiencies. More recently, whether health services and labor forces should be organized or reorganized has been discussed in terms of primary care reform, hospital restructuring, and, more generally, health care restructuring (Barnett and Copeland 2010).

Health systems without boundaries

An alternative way of conceptualizing health systems is to think about them in terms of the underlying values on which they are based. Globally, the most common health system is what is often referred to as the biomedical system of health or the Western medical health system. It is the dominant view of health and health systems thinking in virtually all developed countries, and it is rapidly overtaking other traditional health systems thinking in developing countries. The biomedical system has its origins in the ideas that transformed natural science and industry at the end of the nineteenth century and the beginning of the twentieth century. It is based on a positivist view of science, treating the body in terms of its parts and subsystems, specialization, and emphasizing curative medicine, technology, and the use of pharmaceuticals created in laboratories to overcome specific illnesses. In contrast, there are much older health systems which tend to be holistic in their philosophies and emphasize natural remedies. Throughout the Middle East and South Asia, there remain millions of people
who follow Ayurvedic or Unanic health systems. In China and throughout Southeast Asia millions follow health systems based on Confucian principles. Ideas and practices from these major health systems can also be found in various other forms of traditional medicine in the developing world and alternative or complementary health systems (e.g., homeopathy, chiropractic, etc.) in the developed world (Andrews, Adams, and Sargent 2010).

In the developed and developing world, many people now utilize services that reflect a mixture of health systems. They might seek help through the biomedical system and a complementary health system for what ails them, or seek help for one health issue through the biomedical system and see a complementary health practitioner for other health issues. In many countries, some (or all) of the biomedical and complementary health systems are licensed by the government and covered under the prevailing health insurance systems, while others are not and must be purchased out-of-pocket. Beyond what is provided by governments or can be bought out-of-pocket, people are also their own care providers and providers of care for their families and friends (informal care). Formal and informal care systems together are yet another way of thinking about health systems (Milligan and Power 2010). Health geographers are uniquely qualified to analyze the cultural values and social and economic institutions that people must navigate to seek help from the health systems that operate in their local communities.

The very fact that no two health systems are alike means that health systems create their own unique sets of border issues, and in some cases have become globalized health systems. For example, some US citizens cross the border to Canada or Mexico to get basic health care because health care in the United States is often more expensive. Canadian and Mexican citizens cross the border to the United States to take advantage of the relatively shorter waiting times for some health services, for services not offered in their respective countries, or the perception that the quality of health care is superior in the United States.

At the global level, “medical tourism” can be viewed as a product of the broader forces of globalization. The private health sector is taking advantage of the processes of globalization to create global health systems which mainly cater to persons in the developed world seeking health services at lower costs or health services unavailable in their countries but available in hospitals and clinics in the developing world. What distinguishes these global health systems from local health systems is that they are developed almost exclusively to cater to international medical tourists and the local level elites at the expense of spending on health systems which cater to the needs of the local population. The private health sector also takes advantage of the tremendous social and economic inequalities in some developing countries to market their services in the developed world to the elites of the developing world. Border issues among developed countries, medical tourism, and the marketing of private health to the elites of the developing world are creating new global health system geographies. Health geographers are already contributing much to the understanding of these new systems of health, and there is much more research to be done.

Whether health systems are defined through their composition, the values that underpin them as local or global health systems, or as a product of borders and globalization, what is apparent in virtually all countries is the pluralism that now characterizes health systems. Whether sanctioned by a government or covered under existing health insurance systems, people seek their health care from several health systems they can
HEALTH SYSTEMS

access, given their own values and socioeconomic conditions. Health systems defined in these ways are manifestly geographical and deserve the attention of geographical thinking.

SEE ALSO: Globalization; Health geography; Health inequalities; Location-allocation analysis; Social justice; Welfare geography

References


Further reading


**Health-related behaviors**

Jamie Pearce  
*University of Edinburgh, UK*

**What are health-related behaviors?**

Health-related behaviors refer to the “actions of individuals, groups and organizations, as well as their determinants, correlates, and consequences, including social change, policy development and implementation, improved coping skills, and enhanced quality of life” (Glanz, Rimer, and Viswanath 2008, 66, 12). Health-related behaviors including smoking, diet, alcohol, and physical activity are considered to be major public health intervention points for improving population health and reduce inequalities. Health promotion work has traditionally tended to focus on individual-level interventions such as smoking cessation services. This approach can be criticized for “victim blaming” and failing to recognize that health-related behavioral “decisions” are shaped by the political, social, cultural, and physical contexts in which they are made.

**Geographers and health-related behaviors**

Geographers have been prominent in shifting attention from individualistic accounts of health behaviors to recognizing a wider set of factors. This work has emphasized that space and place are integral to understanding societal processes affecting health behaviors, societal processes operating at multiple scales from the global to the local (Pearce, Barnett, and Moon 2012; Pearce and Witten 2010). Although these geographical categories are not mutually exclusive, the framing emphasizes the ways in which places constrain and enable different forms of behavior at various spatial levels.

**Global**

At the global level, transnational corporations have been implicated in the disease burden of health behaviors through the direct promotion of products that are damaging to health. Diet, for example, is influenced by global food production systems and international trade agreements. The efficiencies associated with the industrialization of agricultural production have resulted in a wider range of food options. But these changes have prioritized commercial interests over nutritional priorities resulting in various agricultural incentives (e.g., food subsidies) to promote increased consumption. “Less healthy” products such as high fructose corn syrup and other processed foods have received higher government subsidies and therefore tend to be less costly than “healthy” options (e.g., fresh fruit and vegetables). The result is an apparent “policy paradox” whereby governments are on the one hand advocating for reduced consumption of high fat, salt, and sugar items but at the same time supporting the overproduction of these items.

The tactics of large corporate interests in subverting various global public health goals have also been scrutinized. Transnational tobacco corporations have developed strategies to counter...
HEALTH-RELATED BEHAVIORS

global and national tobacco control measures. Ploys that have been applied include: claims of damaging economic development and harming trade agreements; portraying tobacco control as an agenda promoted by high-income countries; questioning institutional mandates; and suggesting that precedent would be set beyond tobacco (Weishaar et al. 2012). For instance, an analysis of tobacco industry documents demonstrates how the tobacco industry has successfully lobbied and undermined the German Health Ministry’s position on the Framework Convention on Tobacco Control. Similar corporate tactics aimed at influencing the regulatory environment and increasing consumption have been adopted by the alcohol industry despite the public health implications of excessive drinking.

In addition to affecting favorable public policy outcomes, global factors falling outside of the public health sphere have been implicated in understanding health behaviors. For instance, levels of physical activity are affected by patterns of trade associated with globalization and urbanization. These structural adjustments have been associated with a shift from manual (and often physically active) occupations to more sedentary employment opportunities. Other macro-level changes including the rise in social and economic inequalities which have accompanied the global emergence of neoliberalism have influenced health behaviors such as smoking prevalence and cessation. A sharp rise in social and economic inequalities can result in a social polarization of smoking and smoking cessation. Similarly, macro-level factors that account for high levels of residential segregation, particularly in North America, may influence health behaviors through several interrelated material (e.g., racism and healthcare provision) and psychosocial (local social capital and community norms) pathways with more segregated areas tending to have higher smoking rates.

National and regional

While global-level factors are increasingly important in understanding health behaviors, many aspects of the national context are also pertinent. Public health policies, such as reducing the prevalence of smoking, harmful alcohol consumption, and poor diet, are often implemented at the national level. Geographers have explored the health implications of legislation including the unintended wider spatial effects. For instance, national regulation or codes of conduct for advertising food, particularly to children, varies between countries. Different levels of advertising and other marketing strategies between countries may partially explain international variations in childhood obesity levels. Similarly, the urban design strategies of central and local governments are likely to influence the “walkability” of urban areas. Mixed land use and the variety of destinations and community resources that are accessible locally, such as places of work, parks, and retail outlets, increase opportunities for local residents to walk or cycle around their neighborhood.

Other examples include “smoke free” legislation, which is widely recognized as a public health success. In jurisdictions where smoking in public places has been restricted, the prevalence of smoking and exposure to secondhand tobacco smoke has reduced with health benefits for smokers and nonsmokers. At the same time, the social acceptability of smoking – a key strategy of tobacco control – has continued to decline. Yet as geographical accounts have emphasized, smoke free legislation has also had unintended consequences. Smokers are increasingly displaced to marginal places such as outside of public areas or smoking areas in hospital grounds (Collins and Procter 2011). Smokers are often socioeconomically disadvantaged in a number of ways and being banished from public and private social spaces may lead to “spoiled” identities and
associated feelings of low self-efficacy, powerless-lessness, and hopelessness. The stigmatization of smokers has led to active resistance to cessation efforts, ultimately undermining tobacco control initiatives.

Local

Geographers have paid a great deal of attention to the role of local context (e.g., neighborhoods, schools, and workplaces) in understanding health behaviors. Local context can facilitate or constrain behaviors through two interrelated pathways: the local “practices” which affect what is a normal activity; and the place-based regulation which implicitly or explicitly polices behaviors. For instance, residing in a neighborhood without access to nutritious, high quality, and inexpensive food (a “food desert”) has been identified as a significant environmental barrier to healthy eating. Similarly, supermarket offers on energy dense food, or the local availability of fast food options, may influence purchasing decisions, and ultimately nutritional intake. Smoking and drinking behaviors and patterns are also affected by the local retail environment, especially among younger people. Other work has indicated that local green spaces may encourage people to engage in physical activity by increasing the opportunities for walking and cycling.

Geographers interested in health behaviors have offered epistemological and methodological insights such as emphasizing the complexities of everyday life, and incorporating a wider set of issues including attending to the concerns of those with impairments, disabilities, or chronic illness. Others have emphasized that places are nonstatic and dynamic and that richer accounts of place and health behaviors require an appreciation of the reciprocal relationships between place and behaviors. Behaviors can be conceptualized as a set of social practices embedded in place, which in turn is represented and mediated by local populations. The concept of “collective lifestyles” explains how individual behavior, collective behavior, and series of resources can be brought together in local settings to explain the uneven geographies of health behaviors (Frohlich et al. 2002). The approach is useful because it considers the role of wider structural and societal forces at the same time as the mediating effect of local neighborhood particularities.

Conclusion

Geographical understanding of health behaviors is not only an academic concern. Policy responses that target encouraging individuals to change behavior are likely to be inadequate. Successful interventions to affect behavioral change require concerted efforts at a range of scales, but with a particular emphasis on modifying the social, physical, and cultural context in which individual-level decisions are made.

SEE ALSO: Environment and health; Environmental health; Health geography; Health inequalities; Health and wellbeing; Neighborhoods and health; Public health: human dynamics

References


HEALTH-RELATED BEHAVIORS


Heteronormativity

Linda J. Peake
York University, Canada

A shorthand term for normative heterosexuality, “heteronormativity” is a powerful global ideology, deeply embedded in the ways we know ourselves and the ways everyday life is lived. Normative heterosexuality refers to the encompassing regime in which an individual, sexually attracted only to a person of the opposite sex, is assumed to be a natural and universal norm or way of being human. Hence, the sexed bodies of male and female, the gendered identities of masculinity and femininity, and the gendered desire for the other are assumed to align with each other, and to prioritize the male and the masculine in a gendered hierarchy, forming what is referred to as the “heterosexual matrix” (Butler 1993). The power of heteronormativity is evident in the privileging it lends to other supposedly natural signifiers with which it is most associated – monogamy and marriage. In other words, heteronormativity is understood as most completely institutionalized when performed in a family setting in which a man and a woman, or more than one woman (though rarely more than one man), are legally bound to remain with each other for life. Given the dominance, albeit a diminishing one, of this nuclear family form, it can be argued that heteronormativity undergirds the constitution of the social, so much so that its signifiers are also desired by some nonheteronormative individuals. The assimilation of heteronormative ideals into queer lives, referred to as homonormativity, is revealed, for example, in desires to be married, to form long-term monogamous couples, and to have children. The practices and ideologies that normalize heterosexuality also inscribe difference, leaving other forms of sexuality as inferior and deviant such that those whose desires are neither permanent nor for the opposite sex are considered to fall outside of what is deemed natural, as are their sexual desires. Their failure to reproduce the social group causes them to be valued differently and to be subject to a range of (homophobic) disciplinary practices (Bell 2009). But hegemonic control over non-normative sexualities is constantly in a state of slippage as those who fail to subscribe to the regulatory discourses of heteronormativity rework gendered and sexualized boundaries.

It was in the 1990s that the study of heteronormativity entered the already established geographical field of sexuality and space, which explores the complex relationships between bodies – sexed, gendered, sexualized – and space. Initially, this field focused on the sociospatial construction of (white) gay and, to a lesser extent, lesbian identities, the attention paid to the fluidity of nonheteronormative sexualities and sexual practices speaking to the predominantly queer sexual orientations of researchers in this field and their desire to highlight alternative sexual identities to the unspoken heterosexual norm underlying studies of gender relations pioneered by feminist geographers. Another identifying characteristic of this early period
HETERO-NORMATIVITY

was the conceptualization of identities and the spaces associated with them – for example, lesbian neighborhoods, gay gentrified neighborhoods – as “always already” gay or lesbian. Later, however, the understanding of space and identity formation as fluid, negotiated, and contextual meant that attention turned to the quotidian performative practices through which spaces became heterosexualized (or cisgendered). Space is brought into being as heterosexualized through the repetitive performativity of bodies in gendered and sexualized practices and norms – sex in the bedroom, holding hands in the street, the man putting out the garbage, the woman doing the housework, and so on. So naturalized is the day in, day out, repetition of heterosexuality that it becomes taken for granted and invisibilized. Norms, for example, regarding the form of family life – husband, wife, children – and its immediate spatial expression of the home, arguably the most concentrated space of heteronormativity (Bell 2009), saturate the spatialization of everyday life. In this sense the sociospatial order is heteronormative (albeit contested), with both public and private spaces regulated through disciplinary discourses that normalize and fix heterosexual gendered and sexual behaviors and practices while rendering invisible nonheteronormative bodies.

Social and cultural geographers especially have engaged with queer theory, which is crucial for developing an analytical framing that explores the ways in which various spatialities beyond the bodily are reproduced (and contested) through heteronormative practices and identities – the home, the urban, the national, and the global being the sites of most interest. Social geographers have theorized not only the street, the home, and housing as sites of sexual regulation, but also the spaces of sex education, workplace, leisure and retail spaces, landscapes of sex tourism and of commercial sex, and spaces of love, caring, and intimacy. Historical geographers have also highlighted the links between sex, intimacy, white supremacy, and governance in studies of empire formation; deciding who could or could not marry and who could have sex with whom, where, and when, was an important preoccupation of colonial administrators and companies. This sociosexual logic of maintaining and reproducing territorializations, boundaries, identities, and interests is also now being interrogated in critical geopolitics, which is just learning to see with a queer eye. It is at the urban level that research has engaged in showing how heteronormativity is further reinforced by the values attributed to spatialities, such that attempts may be made to keep “good” forms of heterosexuality spatially separate from “bad” or “immoral” forms, either overtly through planning that may, for example, exclude areas of commercial sex work from residential areas, or more informally through unspoken rules of exclusion (Hubbard 2008). It has been shown that such rules also serve to make those whose sexual identities do not conform to the heteronormative matrix feel out of place and as not belonging. And, while some queer public spaces may have a formality and longevity, such as suburbs or gay villages, others are fleeting and ephemeral in their temporality, either by necessity, such as queer venues in countries where homosexuality is still illegal or gay pride parades, or simply by the passage of time, as particular ways of being sexualized and the meanings attached to their associated spaces fall out of favor.

Geographers have also outlined how the nation-state is deeply implicated in regulatory regimes that define the contours of heteronormativity. It is through the family that the state regulates sexuality via the heteronormative matrix, including such factors as age of marriage and social welfare and taxation policies
that influence the kind of families that are state-supported — usually those headed by a heterosexual couple — and those that are not. Nonheterosexual individuals and unions are being increasingly, if unevenly, incorporated into the national social body in terms of the decriminalization of certain sexual practices (such as sodomy) and inclusion in the nation’s military, and through changing contours of sexual citizenship, with, in some places, same-sex civil partnerships and marriages being increasingly recognized. These demands of the “new homonormativity” (to be considered “normal”) can be seen as converging with the now dominant neoliberal agenda, with its emphasis on individual rights and freedoms and self-surveillance. Puar argues that, just as nationalism has increasingly infused LGBTIQ movements and identities, contemporary US nationalism relies on claims to progressive sexuality. Building on the concept of homonormativity, she outlines its geopolitical equivalent of homonationalism, a regulatory script through which collusion is generated between some homosexual subjects and racial norms (read: predominantly white and middle-class gays and lesbians) and American nationalism through the “normalizing impulses of patriotism after September 11th, 2001” (Puar 2007, xxiv–xxv).

Current work on heteronormativity, having shifted from a focus on identity to questions of practice and performance (Hubbard 2008), is now challenging the universalizing notions of heterosexual domination and queer resistance, and investigating how all sexual orientations are far from solidified and fixed but are fragmentary, defying a categorization that names one component at the expense of all others, and instead celebrating a range of nonheteronormative ways of being as well as considering the multiple desires and bodies that can be accommodated within the category of heterosexual. The Eurocentric and white nature of queer theory has also come under question with the realization that modern sexual identity politics does not necessarily travel well. As urban queer identities in the Global North have and still are gaining acceptance and visibility, they are now defined by some as constricting, as where homonormativity is most advanced. And in the Global South, while some are proudly proclaiming their queerness, in other places they exist under the heteronormative radar as the political space for queerness does not exist and may still be punishable by death. The study of heteronormativity is also entering a “postqueer” stage as critiques are emerging of queer as neither necessarily liberating nor progressive, its contingency resulting in it being utilized in the service of scripts both radical — when it critiques sexualities in relation to class, race, and gender — and hegemonic — when it is tied to white supremacy. This malleability of heteronormativity will ensure that it remains deeply implicated in the organization and reproduction of the social at diverse scales from the bodily to the global.

SEE ALSO: Bodies and embodiment; Citizenship; Difference; Feminist geography; Gender; Patriarchy; Sexualities

References


HETERONORMATIVITY


Further reading

High-tech industry

Yu Zhou  
Vassar College, USA

What is the high-tech industry?

While the high-tech industry has attracted considerable attention from geographers since 1970, a clear consensus on what constitutes such industries hasn’t formed. One definition is based on the characteristics of production. Sectors spurred by new innovation or enabled by highly complex and dynamic technology – such as the information, telecommunications, biotechnology, medical, pharmaceutical, aerospace, and digital media industries – are deemed high-tech (DeVol 1999). However, distinctions between high-tech and traditional or low-tech industries are often ambiguous, as new technologies also transform traditional industries. Another way to define high-tech industries is through characteristics of the workforce. For example, industries with a high share of highly skilled professionals – particularly in STEM (science, technology, engineering, and math) fields – and research and development (R&D) personnel are considered high-tech (Bay-Area Council Economic Institute Report 2012). According to this definition, traditional extraction industries such as oil may also qualify. Beyond manufacturing, services such as engineering and design are also the core of the high-tech industry. It is also worth noting that what constitutes the high-tech industry is context specific. Sectors that are traditional in advanced countries, such as the automobile industry, are high-tech in developing countries.

Why study the high-tech industry in developed countries?

The high-tech industry emerged as a distinct analytical category during the 1970s and 1980s. Since the 1960s, technological advances in microelectronics, digital telecommunication, robotics, biotechnology, and other sectors have spurred the rise of new industries, such as information technology (IT), and transformed existing ones by reinventing organizational and production protocols. Facilitated by technological developments, labor-intensive manufacturing sectors, such as textile and low-end electronics production, gradually relocated to developing nations. Certain original manufacturing cores in industrialized countries suffered from steep decline and depression as a result of this shift while emerging high-tech industries experienced rapid growth. Labor and industrial statistics in the United States show that the growth of high-tech sectors outpaces that of others by a ratio of 3–4:1, and is the strongest contributor to regional growth. High-tech workers earn a substantial wage premium and enjoy far lower unemployment rates than workers in other sectors (DeVol 1999; Bay-Area Council Economic Institute Report 2012).

Global commodity chain and global value chain analyses by Gereffi (1994; 1996) suggest that the highest value-generating segments of production – usually the most high-tech and knowledge intensive – are concentrated in developed countries while low-wage
manufacturing activities are outsourced to developing countries. Global production network (GPN) theory argues that production processes are organized hierarchically. Top tier, large transnational corporations (TNCs) assume the strategic, technological, and organizational leadership of global production, and use developing countries as low-cost manufacturing or resources bases (Coe et al. 2004; Ernst and Kim 2002). Intellectual capacity, rather than physical input or natural resources, thus drives growth and expansion of new industries (Powell and Snellman 2004). The emergence and expansion of the intellectual property rights regime, especially the TRIPS (Agreement on Trade Related Aspects of Intellectual Property Rights) regime under the World Trade Organization, provides institutionalized political, economic, and ethical support for the knowledge-based high-tech industry. In short, high-tech industries have become the harbinger of industrial, employment, regional, and institutional growth in developed countries.

High-tech industry in developing countries

Interests in high-tech industry spread to developing countries in the 1980s. This is in part because some of the most successful newly industrialized economies – South Korea, Singapore, and Taiwan – have become prominent players in the global high-tech sectors. Their experiences demonstrated that technological catchup could occur through integration into global markets. As large countries such as China and India embraced globalization, they also experienced rapid development of their high-tech industries through integration into GPNs.

However, being a part of Western-led GPNs and intellectual property regimes also creates anxiety in developing countries. Political and industrial leaders worry that being locked into Western orbits may limit opportunities for indigenous technological development, permanently trap their economies in low-end manufacturing, and create dependency on imported intellectual property rights. Large developing countries such as China, India, and Brazil have vast human resources and large domestic markets, thus have long aspired to achieve technological autonomy from the West.

China is the most aggressive in promoting R&D investment from domestic sources (Zhou 2008). The technological content of China’s exports rose steadily as more multinational firms outsourced their upstream production to China from 2000, and as China's indigenous firms and research institutes increased their R&D spending. Patent applications and other scientific indexes also rose sharply in China (Crookes 2012). However, a study on Apple products showed that even though final assembly of the most innovative products takes place in China, the contribution of China to value-added is only a tiny fraction of the total (Xing 2014). This suggests that while China is making high-tech products, technological upgrades are still largely driven by external sources with limited indigenous inputs.

In India, the debate on the high-tech industry revolves around its contribution to the national economy. India is the world’s largest exporter of computer software and tech-services (Gregory, Nollen, and Tenev 2009). As the most prominent and dynamic sector of India’s economy, IT has transformed the lives of many middle-class Indians and inspired a new optimism and confidence in the country. Yet, IT growth has been overshadowed by the continued poverty of the vast majority of the Indian population, and by slower change in the rest of the Indian economy. Some argue that dependence on a relatively
small proportion of the labor force has failed to engender broader social development in India (Barnes 2013; Gregory, Nollen, and Tenev 2009; Xiang 2007).

Brazil has long had some of the most technologically sophisticated manufacturing industries in the developing world, including its IT, automobile manufacturing, and aerospace sectors. However, commodities and raw material extraction dominate its economy (Schoonmaker 2002). Brazil’s dependency on large foreign TNCs also raises similar questions about indigenous capacities as those raised in China and India.

Examples of high-tech development in developing countries suggest that, while the industry can enjoy spectacular growth in these countries, it does not necessarily lead to transformations of indigenous capacities or higher levels of economic or social development. Despite the difficulties and occasional disappointing results, what is certain is that the desire, expertise, and supporting institutions for high-tech development are here to stay.

Where does most high-tech industry locate?

The rise of IT makes it tempting to think that location no longer matters for the high-tech industry. But this is not the case. Scholars have found that the global diffusion of high-tech industry coexists with regional clustering. Several scholars explain the existence of innovation hubs through the provision of human resources. Florida (2002) argues that “the creative class” – highly educated and innovative populations – is instrumental in creating an environment conducive to the high-tech industry. Others stress the roles of unique regional institutions that foster the social, cultural, and organizational synergy underlying sustained technological dynamism (Storper 1997; Scott 1988; Saxenian 1994). Still others focus on the role of public policies and investment in inducing high-tech development (Markusen et al. 1991; Castells and Hall 1994; Lazonick 2007).

Governments around the world encourage high-tech growth through infrastructure improvement, preferential land and tax policies, grants to universities, targeted industrial policies, tariff reduction, environmental regulation, and so on. Not all have succeeded. However, Silicon Alley in New York City, Silicon Glen in Scotland, Tel Aviv in Israel, Hsinchu in Taiwan, Zhongguancun and Shenzhen in China, and Bangalore in India are examples of a global network of innovative high-tech centers.

SEE ALSO: Global commodity/value chains; Global production networks; Industrial districts; Industrial restructuring; Industrial upgrading; Information and communications technology; Local/global production systems; Newly industrializing economies (NIEs); Technology and development; Technology spillover

References


HIGH-TECH INDUSTRY


Hillslopes

Carol P. Harden
University of Tennessee, USA

A hillslope is the flank of a hill—a topographic feature of a size that would require minutes to hours to ascend on foot. Hillslopes are sloping land surfaces mantled with soil or other unconsolidated materials and, in most environments, covered with vegetation. In studies of hydrological processes, hillslopes may refer to all portions of the land surface except surface water bodies. Hillslopes are local places that affect the human experience in a landscape. The downslope movement of Earth materials and water on hillslopes creates localized differences in the physical characteristics of a place, thus presenting a range of opportunities and challenges for people, as well as for other species of plants and animals.

Hillslopes may appear everlasting to human beings, but they are dynamic environments that, like entire mountain ranges, change over the span of geologic time. Rates of change on hillslopes vary, from slow, almost imperceptible changes brought about by soil creep and soil erosion, to rapid, dramatic changes caused by landslides and debris avalanches. Gravity pulls everything downward on hillslopes, including raindrops, rocks, soil, skis, and skateboards. Geomorphologists (who study landforms and land-forming processes) are drawn to hillslopes as places to examine landscape change, while city planners, engineers, skiers, and homeowners need to know whether hillslopes are stable so that they can avoid hazards that slopes may pose to people and structures.

Hillslope form

A hillslope can be viewed as the inclined hypotenuse of a right triangle (Figure 1). The angle (\(\alpha\)), the slope gradient, can be measured directly with a protractor or clinometer, or calculated when the height (\(y\), the difference in elevation between top and bottom) and length (\(x\), horizontal distance) are known. Mathematically, this \(y/x\) ratio (called “rise/run” by carpenters) is the tangent of the slope angle. Multiplying the \(y/x\) ratio by 100 expresses this gradient as a percentage; alternatively, the arctangent function expresses the gradient in degrees (e.g., a \(y/x\) ratio of 0.25 (25%) represents an angle of 14°).

Human perception frequently exaggerates the steepness of hillslopes. A 100% gradient is the equivalent of a 45° angle. People rarely venture onto such a steep slope. The seemingly vertical slopes of ski jumps in international competitions rarely exceed 36°: the steepest portion of the in-run has an angle of between 28 and 36°, and the steepest part of the landing is between 35 and 36° (Ski Jumping 2016). Even indoor stairways rarely have gradients greater than 30°. In the United States, the maximum allowable grade on interstate highways is 6% (3.4°), with some exceptions in mountainous terrain. Gradients of roads that merit special warning signs for vehicles (Figure 2) may appear insignificant when drawn on paper.

The parts of hillslopes are named anatomically, with the top of the hill called the head of
HILLSLOPES

Figure 1 The gradient of a hillslope (α) is the angle between the slope and an imaginary horizontal line. The length of line x represents the length of the horizontal projection of the slope; the length of line y represents the difference in elevation between the top and bottom of the slope.

the slope, the lowest part called the toe, and the lower portion above the toe called the footslope. Gravity moves materials from higher to lower slope positions over time; thus, the footslope contains material derived from upslope sources. Soils tend to be thinner on upper slopes, which are sites of erosion, and thicker on footslopes, which are sites of deposition. These characteristic, hillslope-related differences in soils are known to soil scientists as a soil catena or topo-sequence. Hillslope position is thus an important component of the classification and mapping of hillslope soils.

The shape of the longitudinal profile of a hillslope is affected by its geologic history, structure, and materials, and by the influences of climate and vegetation on hillslope-shaping processes. Where hillslope materials are cohesive and do not readily break down into smaller particles, the processes that move slope materials are considered to be weathering limited. In this case, a scarcity of movable particles, relative to the capacities of gravity and flowing water to cause movement, causes the hillslope shape to be controlled by the form of the underlying rocks or cohesive materials. In the opposite case, on transport-limited segments of hillslopes, unconsolidated materials accumulate and reshape the hillslope. Often, hillslope processes at the head of a slope are weathering limited (where sediment transport exceeds sediment production), while those at the foot are transport limited (sites of accumulation). In between, where weathering and deposition are in sync with erosional processes, a hillslope can maintain a constant profile. On slopes where erosion by flowing water is a dominant process, the differences between the erosional and depositional regimes of upper and lower slope segments produce a concave profile, with a steeper, weathering-limited, upper portion and a less steep, transport-limited, portion at the toe. Some hillslopes, undercut at the toe by a river or by anthropogenic excavation, are convex in profile, as are hillslopes on which soil creep is a dominant process.

Hillslope dynamics

In the physical system of a hillslope, the angle of the slope determines the proportion of the force of gravity that can move water and materials downhill. Gravity can be illustrated as a vector of force aimed at the Earth’s core. Conceptually, that vector can be subdivided into two components: one component of the force of gravity that pulls materials toward the Earth and one pulling materials (e.g., soil particles, rocks) downhill, parallel to the slope (Figure 3). If the land surface has no slope, the entire force of gravity acts to pull materials toward the Earth’s core, thereby holding them securely in place. When a hillslope is steep, the downslope component of the force of gravity can exceed the combined resistance to motion provided by friction, cohesion, and the component of gravity that pulls materials toward the slope. In this situation, gravity pulls the materials downslope. People become acutely aware of the
downslope component of gravity when frictional resistance is low, as when a sloping surface (e.g., a parking lot) is covered with ice.

On hillslopes, the potential energy of materials at rest is determined by the height of the slope, the mass of the material, and the downslope force exerted by gravity. This means that longer, steeper hillslopes store more potential energy. When materials (e.g., rocks, mud, soccer balls, or sleds) begin to move downslope, potential energy becomes kinetic energy. Kinetic energy \((KE)\), the energy of the mass \((m)\) in motion, is proportional to the square of the velocity \((v)\) of the moving mass \((KE = \frac{1}{2} mv^2)\). Small increases in velocity can substantially increase the kinetic energy of the moving material. Like the energy transferred to other masses by a cue ball on a pool table, the kinetic energy of moving snow or mud on a hillslope can transfer to other masses, such as more snow, more mud, or trees, houses, and cars, and set them, too, on a downslope course of motion. The transfer of kinetic energy by moving water on a hillslope dislodges and moves soil particles, causing soil erosion. Where slopes are steep and long, the speed and energy of gravity-driven mass movements of materials can be stunning. We recognize these rapid, dramatic movements as avalanches, landslides, and torrents.

**Hillslope processes**

Precipitation interacts with physical and chemical attributes of soil, bedrock, and vegetation on
HILLSLOPES

Figure 3  The force of gravity (g) on a mass (M) can be represented by two vectors of force, one (Mg sin α), which is parallel to the slope, and one (Mg cos α), which is perpendicular to it. As the slope angle (α) increases, the parallel (downslope) component of the force of gravity increases and the perpendicular component decreases.

hillslopes to determine the volume, timing, and chemistry of water that flows to streams. Water moves downslope over the land surface (overland flow) and through the soil (throughflow), with connections between these two pathways where throughflow emerges from or overland flow infiltrates into the soil. Water on hillslopes plays a key role in moving sediment and altering the shape of a landscape.

Downslope movements of materials can be categorized as soil erosion, the entrainment and movement of individual soil particles by flowing water, or mass movement, the downslope movement of masses of slope materials caused by gravity.

Soil erosion

Water flowing downslope can dislodge soil particles and push or carry them to a lower position on the hillslope or to another location beyond the hillslope. Rates of soil erosion depend on the exposure and erodibility of soil particles and the energy of the moving water. In the 1960s, researchers from the US Department of Agriculture published the Universal Soil Loss Equation (USLE) to predict annual soil losses (A) from hillslope plots in croplands by unchannelized rainfall runoff (Wischmeier 1976). An empirical relationship based on 10,000 plot-years of experimental data, the USLE has since been refined and extended. The outline of the equation, $A = RKLSCP$, serves as a checklist of factors contributing to soil erosion on hillslopes. The $R$ factor represents the erosivity of rainfall, recognizing that some locations receive more intense rain than others. Soil erodibility ($K$) represents characteristics of the soil that promote or resist detachment by rainfall. Some components of soil erodibility, such as soil texture, remain constant over periods of years, while others may change frequently. Soil moisture, for example, varies spatially and changes within hours, days, and seasons. The length ($L$) and gradient ($S$) of a hillslope affect the kinetic energy of flowing water, with longer, steeper slopes generating faster flow velocities. The $C$ (cropping) and $P$ (practices) factors reflect the nature of the vegetative cover on the hillslope and allow users of the equation to model the effects of different management practices on soil losses.

Hillslope soil erosion is a natural process that acts over scales of geologic time to redistribute materials and wear away the high points of a landscape, but human activities often increase erosion rates. Human interventions, including agriculture, construction, and mining, remove vegetation, thereby increasing the exposure of bare soil to the impact of raindrops and the effects of rainfall runoff. Practices that degrade the health of a soil, such as unsustainable farming or industrial pollution, reduce the ability of plants to protect the soil surface. Heavy vehicle use and overgrazing promote soil erosion both
by removing vegetative cover and by compacting the soil surface so that rainfall is more likely to run off than be absorbed and stored in the soil.

Some human interventions are designed to reduce soil erosion. These include decreasing slope gradient and length by building terraces, plowing along the contour, or interrupting flow with vegetated strips. Such practices reduce the occurrence of high-energy, erosive flow across the hillslope and facilitate infiltration. Farmers can reduce the erodibility of soil in agricultural areas by using no-tillage agricultural practices to preserve soil structure and promote the storage of organic matter in the soil, or they can reduce the duration of exposure of bare soil to rainfall by planting cover crops.

The amount of soil eroded is reported in linear units of the depth of land lost (millimeters per year) or as a mass of material per area (tons per hectare). Although soil losses in the range of millimeters per year are nearly imperceptible over a year or two, they add up over decades and centuries. Seemingly minor erosional losses may be important in the short term because the uppermost layer of soil, the first to be eroded, contains more nutrients and organic matter, and captures and retains moisture more effectively than underlying layers. Moreover, soil erosion on agricultural lands can also remove seeds, fertilizers, and other valuable inputs, making eroded hillslopes less productive.

On hillslopes, flowing water finds the path of least resistance. Once channels form, the flow of water becomes deeper, hence less affected by friction and therefore faster, with more energy to erode and transport soil particles. Rills and gullies accelerate the removal of soil and water from hillslopes and link hillslopes to stream channels. Away from their hillslope origins, soil particles alter aquatic habitats and occupy space in channels and reservoirs.

Mass movements

Whether, or when, hillslope materials move downhill depends on the balance of forces acting on them. Driving movement is the downslope component of gravity. Resisting movement are the component of gravity that pulls materials toward the Earth and characteristics of slope materials – cohesion, friction, and structural stability – that hold them in place. Over human timescales, the dynamic nature of hillslopes can be hazardous, or at least inconvenient, for human occupants and the infrastructure of human society. With the ever-present force of gravity working to move materials downslope, it is important to be aware of the conditions that promote the stability or instability of hillslopes. Armed with this awareness, people can avoid or alter places where slope instability would create danger or cause problems.

Engineers calculate a factor of safety when they determine the slope angle to build for a highway embankment or other constructed slope. In its simplest form, the factor of safety is the ratio of forces that resist movement (based on mass, strength, gradient, and cohesion of the materials, minus the reduction of cohesion by porewater pressure) to the gravitational force that drives movement (a function of mass and gradient). The factor of safety of a stable slope is well above 1.0. A factor of safety that approaches 1.0 for a constructed or natural slope indicates a threshold condition in which a very small change, such as added weight or an increase in porewater pressure, could tip the balance and cause a mass movement. Types of mass movement, ranging from rockfall to the failure and motion of entire hillslopes, are briefly described in the following paragraphs.

The simplest case of mass movement is rockfall. A rock falls from an outcrop onto a lower position on a hillslope when the force of gravity acting on the rock exceeds the forces that hold
the rock to the outcrop. Exposure of a rock to the effects of weathering processes creates cracks and chemically alters minerals within the rock. Both mechanical and chemical weathering processes reduce the resistance of the rock to the force of gravity. How far downslope a falling rock travels depends on factors that include the mass and shape of the rock, the slope gradient and length, the roughness (frictional resistance) of the slope, and whether any barriers are encountered. Rockfall is a common hazard in places where people have excavated rocky hillslopes to construct roads or have occupied lands below cliffs. Gravity causes the rock to fall, but additional factors, such as the expansion of freezing water in cracks or vibrations from human activity or earthquake, can cause a rock to fall sooner rather than later. A dramatic example of rockfall was documented in 2014 in Ronchi di Termeno (northern Italy), where two unusually large boulders tumbled from a cliff, rolled through a barn, and came to rest in a vineyard (Woo 2014).

Soil-mantled hillslopes frequently show evidence of the type of mass movement known as soil creep, a process in which the effect of gravity on soil particles causes them to slowly migrate downslope. Evidence of soil creep appears as tilted fence posts, tilted headstones, or bulging retaining walls. One cause of soil creep is that soils containing certain clay minerals swell when wet and shrink when dry. Swelling pushes particles outward, perpendicular to the slope surface, but gravity affects the direction of shrinking, pulling the soil vertically downward. The impact of raindrops landing on bare soil has a similar effect, in that most soil particles splashed up into the air will land downhill of their initial position. The net effect is the gradual movement of soil particles down the hillslope.

Entire hillsides can slide downslope, gradually or catastrophically. The release and movement of all or part of a hillslope is referred to as a landslide. Specialists subdivide landslides into categories based on the geometry and magnitude of the mass movement. In a planar landslide, a mass of the land surface slides over an inclined surface. Planar landslides occur where the substrate contains layers of rock or regolith, with contrasting characteristics of strength and permeability, which are tilted at approximately the same orientation as the hillslope surface. The surface material, possibly including forests and buildings, detaches and slides downhill over a subsurface layer. Gravity causes the movement, but the presence of water can contribute to instability by reducing cohesion and friction.

Some of the most dramatic mass movements in human history have been planar landslides. One was the Gros Ventre slide in Wyoming on June 23, 1925, in which 38,000,000 cubic yards of material, including a forest, slid down into a valley with so much momentum that some of the material traveled 90 m up the opposite hillslope. Eyewitnesses estimated that the entire movement took less than three minutes (US Forest Service 2016).

Other landslides, small to large in magnitude, rapid or gradual in speed, have curved surfaces of rupture. These are called rotational landslides because the head of the slide rotates before moving downslope (Figure 4). Slow-moving rotational landslides may be called slumps, especially if they are relatively small (e.g., on a roadside or streambank). The outward rotation of a mass of hillslope leaves a steep escarpment with one or more low-angle steps near the head of the slide. Water can accumulate on these steps, forming sag ponds. Seepage of ponded water can further destabilize the subsurface material and cause further movement at the same site.

The range of processes categorized as mass movements grades from gravity-only processes (e.g., rockfall) to downslope movements that...
involve increasing amounts of water. Water plays a key role in debris flows, which carry a mixture of rock-particle debris, water, and air downslope as a thick slurry. Debris flows require an accumulation of rocks and fine sediment and enough moisture available to initiate movement. More than 40–50% of the mass of a debris flow is sediment; the remainder is water. A debris flow can occur on an unchannelized hillslope, but more commonly follows a pre-existing channel. It can move quickly, ripping out trees and built structures. Acting like wet cement, the moving mass of debris can carry much larger rocks than could be moved by water alone (Figure 5).

Debris flows in steep channels can achieve velocities of meters per second and travel tens of kilometers. The size of the debris in debris flows ranges from clay to large boulders. When the flow stops, the material stops, characteristically as a deposit of fine sediment supporting a collection of larger particles. In wooded regions, debris flow deposits may contain woody as well as rocky debris. A mudflow is a special case of debris flow in which most of the grains are smaller than sand. Some of the most spectacular and devastating mudflows have occurred on the flanks of volcanoes, where abundant, movable volcanic ash combines with water from rain or melted ice to form a mobile slurry of debris. Examples of volcanic mudflows, called lahars, include the one initiated by the eruption of Nevado del Ruiz, which buried a portion of the city of Armero,
Figure 5  This debris flow in Iceland transported large boulders down the hillslope. Several cows near the fence in the foreground provide a sense of scale. Photo C. Harden.

Colombia, and the one caused by the eruption of Mount St Helens (Washington state, USA), which left deep deposits in the Toutle River valley.

The term earthflow is applied to slow-moving landslides, typically on hillslopes of less than 20°. Earthflows may remain active for years, with movement occurring when sufficient moisture is present. Earthflows in California and New Zealand move during periods of winter rain. They may move a few meters per year, leaving a lumpy or cracked topography that extends over hectares. On hillslopes underlain by permafrost, thawed surface soils slip downslope over the frozen layer in a mass movement process called solifluction. Solifluction is aided by the lubricating effect of liquid water on the impermeable frozen layer of soil.

Hillslopes as landscape subsystems

Hillslopes are essential units of analysis in hydrologic systems and watershed models. The surface characteristics of a hillslope determine whether precipitation will infiltrate into the ground or flow across and through the soil to join streams. In turn, this division of hydrologic pathways controls the rate, volume, and quality of water reaching streams, and it affects the
moisture available to plants, the rate of recharge of groundwater, the likelihood of a flood, the chemistry of the water, and the energy available to cause soil erosion. These factors are interrelated, with vegetation playing a key role in site-specific feedbacks involving moisture, soil condition, hillslope stability, soil erosion, and rainfall runoff (Marston 2010). At broader scales, the hydrologic characteristics of hillslopes affect the volume of water stored in aquifers and river systems, and the chemistry and sediment loads of surface waters.

Among the systems of feedbacks affecting hillslope systems are the actions and responses of people. People alter drainage, vegetation, and surface characteristics of hillslopes through engineering, agriculture, trampling, and paving. In the processes of rainfall interception and transpiration, trees can use 40–80% of the precipitation that falls on a hillslope. Therefore, removing trees from a hillslope creates, in the short term, a much wetter hillslope, and planting trees on hillslopes is a well-known strategy for removing moisture. By altering the water budget through deforestation or afforestation, people affect the stability of a hillslope. Sometimes this is deliberate, as when trees are planted to dewater a slope to increase stability; at other times, the consequences are unintended, as when deforestation, without regard to hillslope hydrology, is followed by landslides.

Research on hillslopes

Recent research on hillslopes by geographers and others focuses on multiple themes: (i) better understanding hillslope hydrology and linking it to broader-scale patterns of water and sediment movement, (ii) improving the ability to predict slope instability and mitigate hazardous situations, (iii) understanding hillslope-scale change to improve models of landscape evolution, (iv) using hillslopes as examples of systems in broader studies of the organization and feedbacks of complex systems (e.g., Marston 2010; Poulos et al. 2012), and (v) deducing characteristics of environments and materials on other planetary bodies from imagery of hillslope form. New techniques have stimulated new avenues of research. High-resolution imagery and greater computing capacity have enabled researchers to link spatial scales and numerically model complex systems, and stable isotope analyses have provided new views of the pathways and residence times of water on hillslopes.

Conclusion

Hillslopes are important units of study. From a theoretical perspective, hillslopes are landscape subsystems in which organisms (plant, animal, microbial, human) interact with their environments and physical and hydrological processes interact with materials to alter landforms. From a practical perspective, the dynamic nature of hillslopes will continue to pose challenges to societies that expect permanence in landscapes and invest in rigid infrastructure. As places that integrate abiotic, hydrologic, and biotic (including anthropogenic) processes and feedbacks, hillslopes are essential building blocks of landscapes and landscape analyses at all scales.

SEE ALSO: Geomorphic systems; Geomorphic thresholds; Hydrologic cycle; Infiltration; Mass movement processes and landforms; Mountain geomorphology; Soil erosion and conservation; Soil water; Soils on slopes: catenas; Watersheds
References


Further reading

Historical geography

Akihiro Kinda
Kyoto University, Japan

Historical geography studies phenomena occurring over time while simultaneously analyzing the spatial matrix of those phenomena. It maintains an attitude to research that embraces both time and space, for the most part focusing on various spatial phenomena that have undergone changes over the passage of time. If we think of it as a subdiscipline that brings time and space into its purview at the same time, it is the study of geography into which time is incorporated; it can also be, as exemplified in the French Annales school, the study of history that incorporates space, but this entry centers on historical geography from the vantage point of geography. In 1975 the first issue of the Journal of Historical Geography proposed a widened discipline: “The writings of scholars of any disciplinary provenance who have something to say about matters of geographical interest relating to past time.”

The questions that historical geography asks reveal curiosity about places of the past. Think of the Histories of Herodotus, written in the fifth century BCE, with their abundant geographical information on ancient Greece and much of the ancient Oriental world. Many people of later generations – not only historians but also those interested in the eastern Mediterranean region of ancient times – were eager to identify where those places were that Herodotus described in his description of historical occurrences. A geographical work called Shui jing [Water Classic] had the same effect in China. This account of China’s waterways is thought to have been completed during the Early Han dynasty (second and first centuries BCE). Such information, relating to water control and use, was extremely important in China, where flood control was vital for rivers like the Huanghe (Yellow River), which regularly overflowed, and where from time to time the courses of rivers were changed by flooding and by the building of artificial levees or dikes. This treatise was considered highly valuable, and in the early sixth century 40 volumes of commentary were completed, called the Shui jing zhu [Commentary on the Water Classic]. By around the tenth century, however, a part of the commentary was lost and some of it was said to have been mixed in with the original work, which jeopardized its integrity. During the Ming and Qing dynasties (fourteenth to early twentieth century), specialists in historical artifacts and documents struggled to replicate the whole commentary. The Shui jing zhushu [Annotation of the Commentary on the Water Classic], completed in 1615, is one of the most famous of these attempts, and in the eighteenth century the entire 40 volumes of the original commentary were restored and republished. These are just two examples, but a desire for knowledge about places linked with bygone phenomena in politics, history, geography, or literature was the wellspring of a general interest in the themes of historical geography, the chief aim of which was to identify the present-day location of places named in the past.

In his lectures on physical geography, Immanuel Kant (1724–1804) articulated the idea that there is a causal relationship between human phenomena and natural phenomena, and that the concept of the whole derives from
various phenomena within the framework of space. Space-oriented accounts are geography, he proposed, and time-oriented accounts are history. History, therefore, deals with phenomena in temporal sequence (zeitliche Nacheinander), while geography considers phenomena in their spatial proximity to one another (räumliche Nebeneinander). That is to say, Kant’s categories make geography the study of space without the time dimension, and history the study of time without the space dimension. Alfred Hettner (1859–1941) carried on Kant’s basic ideas in the substantive works he published on topography. Hettner’s topography (Länderkunde) argued that places big and small should be studied in conformity with their essence and that should be the core of geography. Richard Hartshorne (1899–1992), continuing in the direction taken by Hettner and other German scholars, made efforts to organize the subjects and approaches of regional geography and posited time and space as separate entities with no logical mutual involvement nor any direct link between them. He argued that geography is the study of changes occurring in the same space, not of the process of change itself; geography investigates only what is related to the present region.

Theoretically, space and time can certainly be separated, but looking at all kinds of real human activity, they are clearly intertwined, if not inseparable. Time geography, for example, examines the patterns of movement and activity that individuals undertake over the course of a day; it looks at the simultaneous consumption of time and space by human movement and activity as people move through time and act in space. The behavioral approach deals with human movement in units of days and months; social geography, using longer units of time than behavioral geography, looks for tendencies in social groups. What all these approaches have in common is the concept of the passage of time. Often historical geography looks at far longer units of time, but the research approach that takes into account both time and space in studying how humans live and act is the same.

Henry Clifford Darby (1909–1992), a giant in the field, who organized the subjects and approaches of historical geography, set out four categories of research that are generally accepted as basic to the underlying ideas of historical geography: past geographies, changing landscapes, past in the present, and geographical change (Darby 1962).

Hugh C. Prince (1971) also accepted the four-category paradigm, but he moved in somewhat new directions by developing a phenomenological and perceptual approach, on the basis of which he proposed three spheres in the world of the past that historical geography should address. Prince called them the “real,” “imagined,” and “abstract” worlds. He says that (i) “real” worlds of the past are close to Darby’s four subjects of historical geography and researchers find them useful, but this approach is not entirely successful in explaining the process of change or the reasons for change. (ii) Different people at different times all have their own ways of imagining past worlds; their intent and ways of thinking, prodded by cultural preferences and biases, are reflected in the images they form. When historical geographers living in the present seek to get closer to a past world, they do so positioned within a given cultural milieu. Thus the testimony to past times and places that they are trying to make sense of is itself colored by present necessity and present-day critical awareness. The task of understanding the past, therefore, is far from easy, but at the very least, argues Prince, it demands impartial comparisons of different observers. (iii) For historical geographers, it is important to look for order and regularity, and to construct logically consistent models for diachronic change.
Alan R. Baker (1972) basically agreed with Prince's ideas, but adjusted them in certain ways. Baker reframed Prince's category of “imagined” realms as “perceived” realms, and labeled Prince's “abstract” approach “theoretical.” In Baker's view, historical geography is the study of (i) past geographies – horizontal cross-sections; (ii) changing landscapes – vertical themes of landscape transformation; (iii) past in the present – historical explanations of the present-day geography; and (iv) geographical change – investigation of the influence of geographical conditions, and the (a) real, (b) perceptual, and (c) theoretical approaches that make it possible to pursue those inquiries. Twelve discourses have become the “new orthodoxy” for historical geography as follows: past geographies in the real, perceptual, and theoretical realms, and past geographies, changing landscapes, and geographical changes in those three realms. Landscape continues to receive a great deal of attention, especially ongoing evolution of landscape and present-day remains of landscapes from the past.

Landscape studies emerged in Germany and became a particularly strong and somewhat idiosyncratic current. August Meitzen (1822–1910), in working out how villages and farmland were formed, developed a theory of origins of the different landscape patterns shaped by different folk or cultural groups. He proposed that the different types of rural community – Dorf (village), Platzdorf (green village), Weiler (hamlet), and so on – and patterns of agriculture observed in Germany and neighboring countries originated from the cultural traditions of German, Slav, Celtic, Roman, and Finnish peoples. Otto Schlüter (1872–1959) did a careful and detailed review of that theory and used it to study villages in Thuringia and other parts of central Germany. His research and theoretical work became the basis for the discipline of Landeskunde (landscape studies), which focuses geographical inquiry on regional landscape. Diverging from historical geography, where space and time are studied simultaneously, Schlüter endeavored rather to foreground spatial elements and discount the influence of elements of time, and he tried to explain surface phenomena by giving causality to geographical factors through time. His work had a strong impact on the direction taken by, and the growth of, landscape studies in Germany.

Carl Sauer (1889–1975), citing in particular the work of Siegfried Passarge (1866–1958), introduced German landscape theory to America. Sauer's doctrine of landscape morphology (Sauer 1925) bears some marks of anthropological/sociological theory of culture, but it includes the temporal dimension, change over time, and thus locates his approach basically at the same vantage point as that of historical geography. Both of Sauer's diagrams outlining the process of morphological change in the natural landscape and the cultural landscape (for the cultural landscape, see Figure 1) incorporate the passage of time. In his model, historical geography includes tracing changes in the landscape and interpreting changes in regions, and that necessitates the study of cultural landscapes: developments and movements within human communities, changes born of new relationships, and improvements occurring within culture and cultural landscapes have to be considered in order to account for change. Sauer’s model was pursued by his research groups through the investigation of indigenous societies in Mexico and California. Before long it was taken up at symposiums organized around geography, anthropology, and related disciplines, and a collection of papers derived from one such symposium (Thomas Jr 1956) showed that perspective expanding to its most inclusive scope. Then, cultural geography, with its focus on elements of anthropology and
culture, gradually became the center of interest in the field and moved away from the direction of historical geography. According to Richard Hartshorne, an orthodox scholar of regional geography, Sauer’s version of historical geography was already discredited as an area of geography. The German school of landscape studies that had so influenced Sauer itself changed course. One reason was that, like Georg Niemeier (1903–1984) who studied the origin of rural settlements in central Germany and asserted his *Eshkentheorie* (Esch core theory) to describe the prototype of the rural landscape in the region, German historical geographers were beginning to use new and different kinds of increasingly specific data to document historical change in village communities. Another reason was that academics were turning toward *Sozialgeographie* (social geography), which puts social groups—the main factors in the formation of landscape—front and center as research subjects. On the other hand, landscapes, especially landscapes of the past, are studied as much as ever in the field of historical geography, and present-day remains of past landscapes continue to furnish productive research materials.

The materials for historical geography are not simply landscapes per se. Henry Clifford Darby, who in 1962 set out four subjects of historical geography, earned his doctoral degree in 1932 after completing research in the Fenland region of eastern England. Relying on many different kinds of historical sources to reconstruct and describe the historical geography of a part of the Fens, he structured his thesis in six parts: (i) the prelude to the draining, 1500–1600; (ii) the Fen project, 1600–1663; (iii) the consequences of the draining; (iv) the eighteenth century; (v) the nineteenth century; and (vi) conditions after 1900 (Darby 1940). He dealt with the draining, the heath, the “changing arable,” windmills, town, industry, and other topics to reconstruct the geographical conditions of a past time, and he used successive cross-sections to analyze and describe the change. Considering his use of historical materials, we can describe this approach as the historical source method. The distinctive feature of Darby’s completed thesis is a strong ecological emphasis in his treatment of the changes resulting from draining and banking in the heath and

---

**Figure 1** Sauer’s diagram showing processes of morphological change in the cultural landscape. Source: Sauer 1925, 343.
peat region, and the special characteristics of the Fenland’s natural environment.

Darby’s application of the historical source method is even more evident in his reconstruction of locations and distribution of various items making primary use of the Domesday Book, a comprehensive and detailed survey of population, land, livestock, and other assets ordered by William I of England and completed in 1085 and 1086. Darby collaborated with other scholars in the field to produce a series of studies called the “Domesday geographies,” the first of which appeared in 1952. Together they provide detailed reconstructions of geographical patterns found in several regions in Britain, accompanied by a rigorous critique of historical sources (Darby et al. 1952–1977). Darby took the lead in developing and using the historical source method, which prioritizes historical materials. Although other scholars have continued to use it in different ways, it has become a standard approach to research in historical geography.

Among the most widely used and indispensable primary source materials for scholars interested in space and time when following the historical source method are old maps. A map is, by its nature, a way of expressing space. If language is a mode of conveying some idea, relationship, or information, we can think of maps as the language of space. Maps vary greatly—from small-scale maps of the world, continents, and countries wherein the representative fraction is small, to large-scale, detailed maps of regions, cities, or villages, whose representative fraction is larger. What features are described on a map depend on the scale: in modern scientific maps, the most common small-scale maps show lines of longitude and latitude, landforms, locations of major cities, the main geographic features of the world, continents, and countries, whereas a thematic map contains abundant data of a particular kind. Large-scale maps showing survey results, land-use planning, the shape of land lots, or conditions of land use provide very useful data for analysis. Cartographic techniques and the types of features shown in maps have changed over time, but the study of maps is central to geography in general, certainly to historical geography, and cartography has become another subdiscipline of its own. What can be called “old maps” are basically premodern maps with some features that appear in early modern maps, but they differ from scientific modern maps. For example, some small-scale old maps of the world, continents, or countries show not only the way the world or a country was portrayed in the past but what it was actually like then. For historical geography, such a map can be an object of research itself as well as a source for research on something else. Also, old maps commonly reflect the enormous expansion of geographic knowledge acquired during periods of exploration of little-known places, and they clearly reflect the perception of space that pertained at the time when the map was made.

While old maps serve as prime sources for historical geography, there are other useful sources as well, such as historical and archaeological materials from that period (e.g., Baker and Butlin 1973). Many elements of physical geography, such as the climatological, floral, faunal, and topographical features of a place, and other elements like location and distance between prominent features and various conditions, including cultural and historical environments, are also important in historical geography. It is notable that historical geographical research often stressed the significance of the natural environment in the historical events of places (e.g., Hassinger 1931; East 1967/1938).

Land allotment also provides valuable information for historical geographers. Patterns of land allotment generally take the form of partition into blocks, divisions, subdivisions,
parcels, sections, or lots for purposes of land-use planning, land ownership, and land utilization. Land ownership units are usually given the most weight. Allotment patterns can be ascertained by a variety of sources, but manorial and cadastral maps give the most information about the state of affairs when the map was made. When lots are identified by minor place name, number, or via some other identifier, it is often possible to reconstruct contemporaneous conditions even if subdivisions occurred after the numbers were designated. In addition to what is found in cadastral and other old maps, there is much to be learned from land division patterns visible in aerial photos. Maurice Beresford (1920–2005), who started out as an economic historian, developed an abiding interest in historical geography over the course of finding and exploring abandoned remains, especially of medieval villages in England, which led him also to become involved in archaeological excavations. Using aerial photography to look for correlations and contrasts between visible evidence in photos and information in old maps, he charted new pathways in research. Using that approach to study the interaction between the patterns of land allotment and documents including old maps, he was able to confirm the morphology of many long-forgotten villages and the layout of their fields. In 1957 he published a report, together with an aerial photographer, on the remains of 118 medieval villages in England (Beresford and St. Joseph 1958). Beresford was drawn to medieval cities as well. He applied the same technique of investigating remains on the ground and using aerial photos to establish or suggest, through analysis of land divisions and uses, the form, structure, and other data about the so-called new towns in medieval England, Wales, and Gascony (Beresford 1967). To this day Beresford’s technique of combining aerial photos with documentary or cartographical evidence remains a standard method in historical geography.

About the time when Beresford was working on medieval settlements, William George Hoskins (1908–1992), who also began as an economic historian and was an advocate of studies in local history, asserted the need for research methods focused especially on the effect of human activity on the history of landscape. He believed that the most effective research approach was to look first at landscape elements still remaining on the ground. He studied, for example, the remains of granite walls built by the Celts and by later groups, remains of Roman roads, traces of long strip fields, and the ridges running between fields deriving from medieval farming; he also examined field patterns and farmsteads that resulted from the onset of the enclosure movement, central greens of nucleated villages and meadows in the vicinity, and the type and structure of farmhouses in different periods. His purview of a landscape extended from the rural village to the factories and streets of the urban (Hoskins 1955). Both Beresford and Hoskins took a different approach from documentary historical research. They prioritized land division patterns seen on the ground, old maps, aerial photographs, and the present-day landscape – precisely the kinds of data central to historical geography studies.

Beresford, for one, had already been looking underground for evidence in archaeological remains and structures. Such remains, besides being immediately useful materials for historical geography, are effective in identifying the period of the object, and they enable a broader perspective. Karl W. Butzer (1934–) has done paleoanthropological and archaeological fieldwork in places on the northern edge of the steppe extending from North Africa to Europe and West Asia, where he also did a physiographical analysis of the geomorphology,
pedology, and climatology. He used the results to present a broad argument about the life of Paleolithic people and the origination and diffusion of farming, settlements, and towns (Butzer 1964). Since artifacts and structures lying underground can be analyzed together with sedimentary strata, archaeological remains and antiquarian materials have come to occupy an extremely significant role in research on culture, civilization, and environment. In Japan, Saneshige Komaki (1898–1990) cited Halford John MacKinder’s (1861–1947) unique concept of a cross-section to propose a relationship between shifts in desert sand dunes and the life of an essentially Stone Age people during the period 10 BCE to 4 BCE (Komaki 1937, in Kinda 1997). Kenjiro Fujioka (1914–1985) worked that concept of a cross-section into an archetypal thick cross-section (to reconstruct a landscape over a comparatively long time span, such as prehistory, ancient, medieval, early modern, and modern) that has stood the test of time up to the present. However, it might lead to another problem (as shown in Figure 3, below) (Kinda 1997).

In sum, historical geography research utilizes standard basic resources including landscape morphology, land divisions, old maps, and so on, but in many cases historical documents, structural remains, artifacts, and other archaeological sources of evidence are equally important. The three types of source – geographical, historical, and archaeological – overlap to some extent. As illustrated in Figure 2, from the viewpoint of historical geography, whatever is being studied from whatever angle, materials from all three – geography, history, and archaeology – will be relevant to part of it.

The methods of studying historical geography and producing an account of that study are inextricably bound to the type of source materials being used. Darby’s four types of research subjects mentioned earlier are related to the way the results of each are recorded. The first one, “past geographies,” is described by using horizontal cross-sections, and the second, “changing landscapes,” represents conclusions in successive cross-sections. A classic example of this method of description is found in An Historical Geography of England before 1800 (Darby 1936). The fourteen cross-sections in that book present various methods, such as views of England by a single individual at one given time, and views over longer spans of centuries or periods. For the revised edition (Darby 1973), in response to criticism that it lacked coherence, Darby rethought and reorganized the narrative description between the years of the cross-sections and the cross-sectional analysis of 1085 – the year of the Domesday Book – and created cross-sections for 1334, 1600, 1800, 1850, and 1900 (Box 1).

Darby’s successive cross-sections were methodologically adjusted so that thinner cross-sections covering narrow bands of time were integrated with narrative explaining the changes between
them. Among the more enthusiastic proponents of archaeological source materials, there were some who were working with well-defined thick cross-sections that covered broad swaths of time, and later in his life Darby reported that he, too, could have used thick cross-sections as an alternative (Darby 1983).

Since thin cross-sections are strongly dependent on what source materials could be used, they may tend to miss significant landscape elements within a given year, and if they are thick enough to contain landscape elements of different times, they may end up trying to explain a band of landscape elements that in fact belong to different periods. It is so easy to get trapped by problems of logic that some researchers have sought different methods. The one that Alan Baker and John B. Harley (1973) applied was to divide the diverse ways the landscape morphology had evolved from era by era into 16 themes, according to the special properties of the place, and to describe typical landscape or landscape elements for each theme. Their concept aimed less at methodological perfection than at providing more substantial explanation. This method, which approaches features separately, carries the risk of fragmentary treatment, whether of landscape or landscape elements, but this can be overcome to a certain degree by emphasizing the contextual approach. An actual landscape is made up of countless factors, both natural and cultural. Any element of the natural landscape, such as a forest, may become part of the cultural landscape. Those elements are established at different times, and while some of them remain the same over time, others change, and still others disappear. Furthermore, some elements, after disappearing, may be reconstructed. From time to time, an element will disappear, only to re-emerge later, but overall, new elements are continuously being added. It is an irrevocable process of change wherein social, economic, technological, ideological, and cultural factors have a certain impact at any given time. What determines whether they survive or not are government policies and planning, and the social value people give to them. When we study whole landscapes, we must trace as accurately as possible the appearance and disappearance of particular elements and identify the interactions between and among them, and at the same time we must expand our field of vision to encompass the whole context when examining and analyzing them (Figure 3; Kinda 2010).

Darby’s work on the Fenland and Carl Troll’s (1899–1975) perspective on landscape ecology (ökologische Bodenforschung), as well as Butzer’s aforementioned work, along with many other ecological approaches, have generally proceeded in that way.

**Box 1** Darby’s method in *A New Historical Geography of England* (1973).

1. (Before Domesday)
2. Domesday England
3. Changes in the early Middle Ages
4. England circa 1334
5. Changes in the later Middle Ages
6. England circa 1600
7. The ages of the improver: 1600–1800
8. England circa 1800
10. England circa 1850
11. The changing face of England: 1850–circa 1900
12. England circa 1900
Another type of research aims at identifying and describing the changes that have occurred over an entire region without using cross-sections but looking at a wide range of objects. Considerable work along these lines has been done, taking states, colonies, and countries as the unit, due to availability of historical sources and researchers’ interest in them. More work has been done in this way by states/provinces or colonies in North America and in Australia than by countries in Europe. In North America, a good example of this approach is the work by Andrew H. Clark on an island in the Maritime provinces in Canada, and another is the huge endeavor by Donald W. Meinig that culminated in four volumes of The Shaping of America (Meinig 1986–2004).

The practice of historical geography is contingent upon the availability of data and depth of analysis. Today, increasingly sophisticated ways to parse information in historical documents and techniques for component analysis are enabling ever more exact estimates of time and place, while global positioning systems give us fast and accurate specifications on localities, all of which raise the possibility of valid inference. Advances in the digital humanities and geographic information systems have brought greater diversity to analysis techniques, while making possible the storage and processing of huge amounts of data.
Historical geography, like other disciplines, is exploiting the possibilities these tools offer.

The vast amounts and greater specificity of data available today, along with ongoing progress in methods of sorting and examining it, are yielding improved research results, and at the same time facilitating the transmission of the results to the public. By digitizing their holdings, regional history archives, for example, anywhere in the world can be accessed by anyone with a computer connected to the Internet. Also, being able to bring more minute detail into large-scale maps and digitizing them will make it easier to enter and compare the location and configuration of archaeological remains, and this will bring more archaeological materials within easy reach of historical geography researchers.

Being able to integrate finer and more abundant detail and a wider assortment of data into reconstituted landscapes has already had a huge impact. In museums around the world it is common to find large-scale dioramas and assorted computer graphics showing the past of cities and explaining regional change and development. But with such dioramas and computer graphics we still come up against the same problem that Darby confronted as he grappled with (vertical) narrative and (horizontal) cross-section, and that is the problem of time in the cross-section of a reconstructed landscape: is it a fixed, specified time or does it represent a span of time? This question highlights the importance of dealing with cross-section in research.

Making thin cross-sections depends on whether there are adequate materials relating to the year in question that allow us to reconstruct the year and to make a description. Had the Domesday Book not existed, it would have been impossible to make a thin cross-section for the year 1085 or 1086 in England. When there are no adequate materials directly spotlighting the targeted year, the only recourse is to search around for some other sources and methods on which to base an estimate. Here, archaeological structures, with information coming from the strata where they are found, are invaluable. Steadily improving analysis techniques have made it easier to get more accurate chronological dating of those structures and to reconstruct the conditions of their natural environment; that allows one to examine different kinds of natural environment and to derive a picture of the way things were ecologically, thus facilitating an intersection with the focus of the historical geographical study. Once the year and character of structural remnants from the past estimated from aerial photos have been confirmed by archaeological methods, then aerial photos can be used to study such remnants in a much wider area. Further, when using large-scale old maps like cadastral maps to estimate conditions at a time before the map was made, archaeological and historical materials can also play a significant role in establishing the year and illuminating the state of the landscape.

With greater precision and richer substance, historical geography is contributing to a much more nuanced understanding of space in historical time.

**SEE ALSO:** Annales School; Behavioral geography; Climatology; Cultural geography; Geographic information system; Humanistic geography; Landscape; Marxist geography; Phenomenology; Regional geography; Social geography; Time geography and space–time prism

**References**

Historical settlement

Jeffrey S. Smith  
Kansas State University, USA

Prior to the Neolithic age hunting and gathering were the primary means of securing food for early humans. Small bands of less than fifty people would stalk wild game and migrate between locations as dictated by growing seasons. Banding together offered mutual protection, but settlements remained small because larger groups could easily outstrip the food supply available within walking distance. Between 12,000 and 10,000 years ago the Neolithic revolution (also known as the agricultural revolution) began. Owing in large part to the ending of the ice age, over the course of thousands of years and at differing rates over disparate locations, humanity made the gradual transformation from a nomadic food gathering lifestyle to one based on food stuffs raised by a more sedentary populace.

The agricultural revolution brought about many changes for humanity. First, reliable sources of food made life more secure with fewer people dying of starvation. This led to a dramatic increase in the world’s total population. Second, as humans perfected irrigation techniques and larger scale agriculture, select people could be called upon to raise enough food to be distributed among the remaining population or stored for future use. Because not everyone was required to share in the food gathering duties, individuals could devote their time to other specialized activities. Some people focused on developing trade and commerce while others became governing leaders of the community. At the same time, arts, culture, and architectural diversity also emerged. This diversity led to a division of labor and social stratification within society. Third, as food supplies became more reliable and the capacity to store food ensured that it was available throughout the year, the previously small dispersed bands of humans slowly began to congregate into organized, sedentary population centers. As time passed, some of these settlements evolved into highly organized and noteworthy places.

Neolithic era settlements (10,000 BCE to ca. 1500 BCE)

Levant (near present-day Jericho, West Bank) is now considered to be the oldest organized human settlement. Marking the beginning of the Neolithic period at about 10,000 BCE, the diet of the local inhabitants became more dependent upon grains than wild game. As these people experimented with various agricultural techniques, the settlement’s population began to grow. Between 10,000 BCE and 8000 BCE, the farming techniques developed in the Levant area spread to Asia Minor (Turkey), Northeast Africa (Egypt) and Mesopotamia (Iraq, northeast Syria, and southeast Turkey). Two of the earliest settlements to adopt this new lifestyle include Çatalhöyük (present-day south-central Turkey) and Eridu (present-day Iraq). Archaeological evidence suggests that Çatalhöyük thrived between 7500 BCE and 5000 BCE with its heyday around 6000 BCE. Eridu first developed about 1500 years later around a ziggurat (temple and administrative building) that was surrounded by a residential space and a protective wall. As time
HISTORICAL SETTLEMENT

passed, concentrations of human settlements developed in four main areas. Most of the earliest settlements developed along river valleys because they afforded the local population a reliable supply of water, soils rich in nutrients, and efficient routes for trade and transportation. Cultural geographers refer to these locations as cultural hearths (or cradles of civilization).

Fertile Crescent/Mesopotamia

Believed to be among the oldest settlements in the world, Eridu (5400 BCE) and Ur (3800 BCE) were situated on the southern banks of the Euphrates River in southeastern Mesopotamia. The region was settled by three separate cultures (Sumer, Ubaidian, and Semitic) who all became integrated. As the local population honed their pastoral, fishing, and agricultural techniques, the ample food supply enabled the population to grow and to become denser. At the top of the division of labor was a strong ruling class who offered religious guidance. A merchant class traded surplus food and crafts with peoples in other parts of the Fertile Crescent and distant lands. Eridu and Ur entered a period of decay when the center of Mesopotamia’s population shifted north and the Euphrates River changed its course inundating many of the early settlements.

Around 2000 BCE the settlement of Babylon emerged farther upstream along the Euphrates River in a particularly fertile agricultural zone. What started out as a minor population center grew into a city-state eventually, becoming one of the most important cities in Mesopotamia. For a short time, it even served as the capital city of Babylonia, a center of learning and the arts that featured the Tower of Babel and the wondrous hanging gardens. Babylon was plundered and destroyed repeatedly with its ultimate demise coming at the hands of the Roman Empire.

Nile River valley

Within the Nile River valley two settlements are of noteworthy importance: Thebes and Memphis. Thebes was first established in ca. 3200 BCE on the eastern banks of the Nile River (in present-day Egypt). Not only was it an important religious and political settlement, but it became an important trade city. Merchants exchanged goods up and down the Nile River as well as across the Red Sea between various parts of the Fertile Crescent. Scholars believe that in 2000 BCE Thebes had 40,000 residents and by 1000 BCE its population had grown to 80,000 people. Its main rival along the Nile River was Memphis.

Pharaoh Menes established Memphis (ca. 3000 BCE) at the mouth of the Nile River delta. The site was ideally located not only because the surrounding fertile lands enabled farmers to raise an abundance of food, but its location facilitated an extensive trade and transportation network that extended throughout the Mediterranean region. Memphis became the capital of Egypt during the end of the First Egyptian Dynasty (ca. 2800 BCE) and grew in importance thereafter reaching its apogee during the Sixth Egyptian Dynasty in 2300 BCE. The waning of Memphis coincides with the rise of Alexandria (Egypt) which remained an important political center during the First Persian (550 BCE), Greek (323 BCE), and Roman (31 BCE) empires.

Indus River valley

Early humans began concentrating in settlements in a third cultural hearth located along the Indus River in present-day Pakistan. The Indus River valley attracted a mix of people including hunters, gatherers, and herders who traded among themselves. Over time the various populations concentrated in settlements along most of the Indus River valley with the two most noteworthy being Mohenjo-Daro and Harappa. Both
were established just after 3000 BCE and despite the considerable distance separating the two places, they were constructed of similar earthen building materials, giving scholars confidence to assert that they were ruled by a strong central government whose jurisdiction stretched up and down the river valley. Mohenjo-Daro, located at the mouth of the Indus River, was the most prominent settlement in the region and the local culture invested considerable effort into developing a system of irrigation, animal husbandry, and civil engineering. Around 1800 BCE the Indus civilization entered a period of protracted decline due to changes in the local environment and many of the early settlements including Mohenjo-Daro and Harappa were abandoned.

Wei/Huang He (Yellow) River valley

The fourth cultural hearth is found along the Huang He River valley and its tributary the Wei River. Located in present-day central China, humans first inhabited the region before 5000 BCE and established settlements around 2200 BCE in areas with highly structured irrigated farming. Over centuries the population grew and spread throughout the river lowlands. By 1750 BCE the Shang Dynasty consolidated power and the Huang He River valley emerged as the cradle of Chinese civilization. Key settlements in the region include Zhengzhou, Anyang, and Lanzhou. Each served as important administrative and cultural centers and the area surrounding these settlements has been continuously inhabited ever since. Over time city-states emerged and by 221 BCE the Qin Dynasty unified the independent principalities into a centralized empire focused on the Huang He River region.

Western Hemisphere

As settlements were being formed in the Eastern Hemisphere, humans began clustering together in various parts of the Western Hemisphere as well. Approximately 16 000 years ago the Earth’s climate began to warm and the glaciers ringing the Arctic Circle retreated creating passageways for one of humanity’s greatest diasporas. Early humans migrated across the Bering Strait and moved into parts of the Western Hemisphere. Although most of the hemisphere was thinly populated, humans became concentrated in two main areas (Mesoamerica and the Andes).

Mesoamerica encompasses much of present-day central and southern Mexico, the Yucatan Peninsula, and most of northern Central America. Within central Mexico the Aztecs built upon previous settlements first established by the Olmec, Toltec, and Teotihuacan cultures. They founded Tenochtitlan (present-day Mexico City) as their capital city on an island in the middle of Lake Texcoco. Tenochtitlan fell to the Spanish when Hernando Cortez invaded and founded Mexico City on the same site.

The Maya established numerous religious and political settlements throughout the Yucatan Peninsula and extending into northern Central America (especially Belize and Guatemala). Some of the most noteworthy include Uxmal, Chichen Itza, Tikal, and Copan. All of these high-density settlements were abandoned when the Maya population diffused into smaller, low-density settlements spread throughout the greater Yucatan region where they still live today.

Farther to the south were the ancient settlements of the Inca. Unlike most other Neolithic settlements, the Inca built many of their population centers high in the Andes Mountains. Cuzco (3352 m (11 000 ft) above sea level) became the Inca’s most important administrative settlement. An extensive series of roads (highland and coastal) connected a vast empire that extended from present-day southern Colombia to central Chile. Descendants of the Inca still inhabit the mountainous region of South America.
HISTORICAL SETTLEMENT

Humanity spread from each of these four cultural hearths, establishing settlements over much of the Earth’s surface. As settlements in these four regions continued to grow in size and population density, there were numerous challenges that never seemed to be properly addressed including reliable supplies of potable water, the management of waste (trash and fecal material), efficient transportation systems, and security from invading forces. If leaders were going to address these chronic issues the form and land-use functions of settlements would have to change. Spontaneous growth based on agriculture and trade needed to give way to planned cities. The next major era of human settlement was the Iron Age where Greek and Roman settlements became most noteworthy.

Ancient era settlements (1500 BCE to 476 CE)

As the fifteenth century BCE came to a close, the centers of human settlement began to shift. The hearths of civilization in Mesopotamia and along the Nile and Indus River valleys entered a period of decline. In their place the sun began rising on a new empire. The Greeks emerged from the Dark Ages around 800 BCE and over the course of 950 years the civilization expanded its influence throughout the Mediterranean and Black Sea regions. During the reign of Alexander the Great (356–323 BCE), the Hellenic Empire covered a vast territory from the Indus River valley and the foothills of the Himalayan Mountains on the east, through the Fertile Crescent and Nile River valley to present-day Greece on the west.

The Greek authorities founded an untold number of small settlements (polis, aka city-state) throughout their domain. A typical polis was comprised of a walled settlement surrounded by agricultural lands. Within the city walls was the acropolis – an area reserved for religious worship as well as a defensive site. Near the center of the settlement was the agora (a space for the public to gather to exchange goods and merchandise). In larger settlements (e.g., Athens and Sparta) the agora became a dedicated market place and entertainment district. Most of the residential space in early Greek settlements (especially on the western frontier in France and Spain) lacked formal planning and featured irregular, curvy streets with narrow walkways. Over time, however, Greek authorities imposed city planning mandates that redesigned the transportation network and improved the flow of traffic through the settlements. Streets then adhered to a rigid grid pattern and land use became specialized among different professions within the larger population (e.g., artists and craftsmen, farmers, military).

In 146 BCE Greek society fell under Roman rule and settlements began taking on a different character. Over the next 600 years, the Romans expanded their zone of influence to include all shores around the Mediterranean Sea and as far north and west as present-day Western Europe (including parts of the British Isles). A network of trade routes between the far-flung settlements combined with security and protection provided by the Roman Legions were key to holding the vast empire together.

In the early years Roman settlements adapted and built upon the foundations left by the Greeks. Under Roman authority however, water, sewer and sanitation, and transportation systems were all improved (including aqueducts and paved roads). Gradually, settlements evolved to reflect the rigid social stratification that characterized Roman society. For example, special administrative and religious complexes were set aside and segregated from space afforded to common citizens.

Most Roman settlements were laid out around a large central square or rectangular space with
its focal point being the forum. The forum was similar to the agora found in Greek settlements in that it provided a space for public gatherings and a market where merchants could sell their wares. The Roman forum also hosted political functions, sporting events, state funerals, and other public interest activities. In settlements that grew to become central places for trade (e.g., Rome, which had a population between 300,000 and 500,000 people), the Roman authorities built a basilica (a permanent, covered structure where vendors could sell their merchandise). On the frontier of the Roman Empire settlements took on a more specialized form serving as either administrative centers or military outposts. Yet, in all cases the layout, design, and function of Roman settlements reflected the empire’s strong sense of order and rigid social stratification. When the Western Roman Empire fell in 476 CE, the Roman Legions no longer provided safe passage for trade caravans. People were left to their own devices and many of them abandoned the larger settlements in favor of new, more secure places to live. Urbanized settlements were transformed once again.

Feudal era settlements (476 CE to 1492 CE)

Following the collapse of the Roman Empire, Western civilization entered the Middle Ages (aka Medieval period) where Europe made the transformation from being an exploited region on the margins of the Roman Empire to the center of population growth and innovative thinking. For the better part of those 600 years, however, the Middle Ages were more commonly characterized as a time of social unrest and civil war with countless invasions by marauding barbarians (e.g., Vikings, Germanics, Huns) from northern lands. In the absence of a central government, regional kings constructed walled settlements that surrounded a fortified castle. Seeing the obvious benefit of living within a secure settlement, residents in the region pledged loyalty to the king in exchange for protection. By the ninth century a feudalistic society emerged with an entrenched social hierarchy whereby everyone ultimately fell under the jurisdiction of the king. Serfs and landed peasants worked plots of land giving a portion of their annual harvest to the landlords. Knights served as the regional army providing protection throughout the kingdom. Lords, nobles, and barons (with their own smaller fortified castles) provided loyalty and tax support to the king, which in turn funded the activities of the knights.

In the eleventh century identifiable regions began to emerge based on specialized agricultural practices. For example, the lowland countries (present-day Belgium and Netherlands) became known for sheep production and a strong weaving industry developed. Slowly trade between these specialized regions expanded with some settlements assuming a primary role in hosting trade fairs. As the reputation of these important regional settlements spread, merchants, craftsmen, and laborers gravitated toward them (e.g., Ghent, Belgium and Vienna, Austria). What started out as small, walled settlements evolved into major European urban centers with sizable populations. Although religious conflict (e.g., Protestant Reformation) and economic stagnation limited urban growth in other parts of Europe, merchant activity throughout Italy (e.g., Venice, Genoa, Florence) continued to grow and expand. The Medieval Period came to an end with the rise of the Italian Renaissance and the Age of Discovery in 1492 CE.

Laying a foundation for modern settlements

The founding of settlements in the Americas did not occur in a vacuum. The general design of
human settlements can be traced back centuries to Greek, Roman, and Medieval influences. Europe’s discovery of the New World was largely the culmination of a series of fortuitous events both within and outside Europe. First, as the spirit of the Renaissance and the Age of Discovery swept across the European continent, royal leaders funded explorations to uncharted lands hoping to find abundant riches. As a handful of European monarchs gradually consolidated authority from local feudal lords and modern nation-states emerged, a new economic mindset developed in Europe. Mercantilism (a political economic system based on the principles of accumulating bullion, creating a merchant marine, establishing colonies, and exploiting natural resources to attain a favorable balance of trade) became the guiding force that steered the decisions of Europe’s rulers. As life under the modern nation-states became more secure, landed elites also looked to expand their wealth by investing their riches in joint-stock companies (precursors to modern-day corporations) which they hoped would return handsome dividends.

Geographers James Vance Jr (1970) and Donald Meinig (1986) explain that European settlement of the New World progressed in three main stages. Stage I (exploration) involved establishing settlements that served as a base from which the search for information and new knowledge could proceed. The explorations of Christopher Columbus (1492–1506) are an excellent example of this stage. Each voyage funded by European governments to the New World produced an increasing awareness of what the region had to offer. Stage II (fixation) refers to the establishment of a coastal settlement (entrepôt) where imports and exports are brought for collection and redistribution. In the third stage (colonial expansion), European governments began populating the Americas. A variety of settlement types radiated out from the port-of-entry along trade or transportation routes. During colonial expansion distinct differences became apparent in how the European governments settled the New World.

By the second decade of the sixteenth century, Spain established anchor settlements in strategic locations throughout the Caribbean, central Mexico, and the highlands of South America. From these administrative outposts Spain explored their newly claimed lands and searched for new sources of wealth. For example, from Mexico City and Cuba various Spanish explorers headed north to what is now the American Southwest (New Mexico, Texas, Arizona, and California) and Florida. With the aim of securing these lands for the Spanish crown, the Spanish established permanent settlements in coastal Florida, along the upper and lower reaches of the Rio Grande, and along coastal California. Some of the most noteworthy settlements include St Augustine (1565), Santa Fe (1609), Albuquerque (1706), San Antonio de Bexar (1718), San Diego (1769), San Francisco (1776), Santa Barbara (1786), and San Jose (1797). Most Spanish settlements followed the guiding spirit of the Laws of the Indies, but there were three main types: mission (settlements that focused on religious services), plaza (civil and administrative settlements), and presidio (fortress/military complexes). Nearly all of the major settlements were linked to Mexico City via various branches of the Camino Real (royal road). The multipronged “road” served as the principle trade and transportation corridors similar to the trade network that existed during the height of the Roman Empire. The best example of the ribbon settlements established along the Camino Real was located in coastal California where the Spanish built 21 settlements (primarily Franciscan missions) that stretched over 800 km from San Diego to San Francisco. Like the Spanish, the French colonial government’s primary concern with respect to
the New World was resource extraction and wealth accumulation. Therefore, most of the settlements France established were port or trade route communities. Quebec (1608), Trois Rivieres (1642), and Montreal (1642) were three of the earliest and most successful. Although each served its immediate hinterland, their main focus was to facilitate the shipping of precious goods (especially furs) back to France. As France’s influence in the New World expanded into the interior, additional settlements were established along water transportation routes (e.g., the Great Lakes, Mississippi River). These settlements (e.g., Niagara (1678), Detroit (1701), Memphis (1739), and Pittsburgh (1754)) tended to be fortified outposts where goods could be traded and stored for shipment to the larger coastal port facilities including Quebec, Montreal, and New Orleans (founded 1718). Most French settlements in the New World were little more than thinly populated outposts and did not move far beyond the fixation stage. Because France was more committed to the acquisition of wealth than establishing permanent settlements, other than in the Province of Quebec, it had a fleeting impact on settlement patterns in the New World.

British settlement in the New World was dramatically different. Instead of focusing on finding stores of wealth, their primary objective was to establish permanent settlements (e.g., Boston (1628), Hartford (1635), New York (1664), Philadelphia (1681)) that would relieve some of the population pressure and patterns of religious persecution in England. Over time coastal America became divided into three main regions (northern, middle, and southern colonies), yet the settlement pattern was remarkably similar following the three major stages explained above. After immigrants established an initial port colony, the population grew and expanded into the interior along trade and transportation routes. British frontier settlements were self-supportive but maintained well-developed trade connections with the coastal settlements. Most of the early settlers originated from Britain, yet thousands of people also came from other parts of Europe including Holland, Germany, Switzerland, and Sweden. Each brought with them cultural baggage that would have an influence on the form and function of New World settlements.

SEE ALSO: Cities and development; Cultural diffusion; Cultural geography; Diaspora

References


Further reading


HISTORICAL SETTLEMENT


The oldest settlements in Asia in the sense of permanently inhabited sites may have appeared around 12,000 years ago, before the development of agriculture. These settlements were created by people who depended on wild sources of food obtained from aquatic environments, which provided dependable sources of sustenance throughout the year. The number of people in these early settlements probably totaled 30 or 40. Eventually some of these sedentary hunters and gatherers began to domesticate plants and animals. Population grew, and some settlements reached a total of approximately 400 people. This number seems to represent the limit of population size that societies can sustain without some mechanism of dispute resolution. Once new forms of government were developed, the size of the largest settlements expanded by to perhaps 5000 people.

In English-language terms, these three settlement types can be equated with villages, towns, and cities. These words, however, have implications for the types of activities carried out in settlements of each type, which are commonly found in Europe but not in all other regions of the world. Historians and archaeologists prefer to think in terms of hierarchies of settlement. A region with settlements of one size ("villages") would be termed a region with a single-tiered settlement hierarchy. A region with settlements of two sizes ("villages" and "towns") would be termed a two-tiered hierarchy. The number of tiers in a regional settlement hierarchy is, in theory, unlimited. The use of this system enables scholars to compare the historical development of settlements in different parts of the world while avoiding culturally specific terminology.

Settlement patterns can also be studied from the perspective of the number of settlements of different levels of a hierarchy. In some regions, the proportion of settlements in the third level of a hierarchy ("cities") relative to the number of first- and second-level settlements may be higher than in other areas. Differences between regions in terms of the proportion of settlements of different sizes may be seen by plotting them on a log-normal graph. This method has been used in Mesopotamia to compare different trajectories of settlement formation (Adams 1981). Settlement patterns can be correlated with economic and political institutions.

In southwest Asia, important prehistoric settlements include Catal Huyuk and Jericho. At the beginning of the historic period approximately 5000 years ago, the largest settlements were Eridug, Ur, Uruk, Girshu, Lagash, Nina, and Umma, all located near modern Baghdad. Research on these sites has focused on a search for the factors which led to their appearance. Among the most popular are water management, formation of religious cult centers, warfare, and trade. Scholars differ on the relative importance of these factors and the sequence in which they became significant. The same types of arguments developed for Mesopotamia recur in the study of historic settlements in other parts of Asia.

The concept of the “Islamic city” is another subject of debate, which arose in southwest Asia and spread to other Islamic areas. Not all scholars agree that Islam can be correlated with
HISTORICAL SETTLEMENT: ASIA

the formation of a type of settlement (Eldem, Goffman, and Masters 1999). The phenomenon of urbanization was widespread in the countries where Islam spread, from the Near East to North Africa, South Asia, and Southeast Asia. During a period when most Europeans lived in agrarian villages and only a few cities in Europe boasted 100,000 people, Cairo and Constantinople each had a population of several hundred thousands. The most prominent centers of activity in Muslim Asia and Africa, as well as the Arab countries, were the palaces, markets, and mosques. Cities were divided into quarters according to occupation and ethnicity. Whether Islam was the main independent variable in this development or one of a number of interrelated variables is a subject of complex debate.

In Central Asia, the area between the Caspian Sea and China, important settlements of the historic period began to form around 2000 years ago, when the Roman and Han empires provided two large and sophisticated markets for exotic commodities. Oases and caravan resting places such as Bukhara and Samarkand (now in Uzbekistan) existed as early as 300 BCE. Many of them became centers of Buddhism and later Islam.

In East Asia, important early settlements include Zhengzhou, Anyang, and Loyang in northern China. These appeared around 3500 years ago, when the first Chinese historical sources become available. Wheatley (1971) argued that the earliest large settlements in China coalesced around ceremonial centers, and that the same factor was responsible for the formation of settlement hierarchies elsewhere in Asia.

An important feature of early Chinese capitals is that they consisted of clusters of discrete activities, separated by open space. Administration, religious activity, various crafts, and settlements were conducted at separate locations.

In South Asia, the oldest large settlements appeared in the Indus Valley region of Pakistan and northwest India around 4500 years ago. The major sites of this period are Harappa and Mohenjo-daro. The Indus Valley phase came to an end around 3700 years ago; it was followed by a gap of about 1000 years until a second phase of large-scale settlement formation began, with sites such as Taxila and Ujjain. This phase too ended in a period of urban “decay.” In the subsequent phase, between 600 and 1300, important cities arose in places like Tanjavur in south India (Champakalakshmi 1996) and in Bengal.

Around 2000 years ago, important settlements appeared at seaports along the southern coast of Asia, from the Persian Gulf through the Indian Ocean to the South China Sea. The settlements of the Asian littoral were quite different from those of the agricultural hinterlands. The hinterland settlements were characterized by an emphasis on monument building, warfare, irrigation, and rigid social stratification. In coastal societies, it was easier to improve one’s status through becoming wealthy. Few resources were invested in large monuments or fortifications (Miksic 2013). Hinterland empires seem to have been well disposed to the development of autonomous ports on their fringes; they provided exotic status symbols desired by the elite and acted as a buffer against foreign cultural influences (Fox 1971). As long as taxes and tribute were paid, hinterland capitals did not concern themselves greatly with the administration of the ports. By the beginning of the Common Era, Greek and Latin authors became aware that a standard system of international trade linked ports in the Red Sea with those in the Malay Peninsula.

The growth of maritime trade provided a major stimulus to the growth of settlements in Southeast Asia, which lies at the midpoint of the Asian coastal sea routes. The oldest large Southeast Asian settlement, Oc-èo, in
the Mekong Delta of Vietnam, was probably known to the Alexandrian Greek cosmographer Klaudios Ptolemaeus. By 1450, Southeast Asia had become one of the most highly urbanized parts of the world. This maritime trade linking the Indian Ocean, South China Sea, and Java Sea also encouraged the growth of cities along the southeast coast of China. Cities such as Melaka in the Malay Peninsula and Bantam on the island of Java formed during the Islamic period, which some scholars classify as “Islamic cities.”

Southeast Asian lowland settlements are difficult to study. Many were built over water, both on stilt houses and rafts, and as a result their remains are submerged beneath water and sediment. Archaeologists have been unable to recover much data from important early settlements such as Palembang in Sumatra. That they were impressive is clear from the reports of the first Europeans who visited them in the sixteenth and seventeenth centuries.

Much more archaeological research is necessary to enable scholars to verify and expand on the historical descriptions of historical Asian settlements. The subject of settlement archaeology has only recently been introduced to Asian research. Most attention has been focused on the largest settlements and the most impressive buildings within them. Attention is now shifting toward the study of settlements of different sizes, and to the areas where people of the lower classes lived and worked. More research is being designed to recover patterns of settlement rather than concentrating only on individual settlements. This development will eventually make it possible to apply more sophisticated tools of geographical analysis to examine such subjects as demographic and economic change.

**SEE ALSO:** Ancient geography; Cosmopolitanism; Cultural heritage; Historical geography; Urban geography

## References


## Further reading

Writing on “Extraordinary Cities: Early ‘City-ness’ and the Origins of Agriculture and States,” Taylor (2012), supports Jane Jacob’s controversial thesis, which used Sahlin’s “stone age economics” of cities inventing agriculture. Using Mesopotamian studies, Taylor contends that states were invented in cities, and argues for “central flow theory” as an alternative to traditional central place theory. In identifying 12 possible early cities beyond western Asia, Taylor lists Great Zimbabwe (1250–1450 CE) and the Middle Niger urban complexes (first millennium CE, possibly earlier) as African examples. Taylor further identifies Upper Egypt (3200 BCE) and the Guinea and Sahel riverine areas of West Africa (900–1800 CE) as two of the six cities and city-states in the main “civilization” zones (the other regions being the Indus Valley, China (Erlitou culture and Shang empire), Mesoamerica (pre-Maya to Aztec), and “Greater Peru” (coastal and Andes)). The Upper Egypt cities included Hierakonpolis, Nagada, and This (or Thinite, near Abydos), while the West African Hausaland city-states were based on seven cities: Kano, Katsina, Zaria, Biram, Daura, Rano, and Gobir. In short, Taylor draws attention to some of the earliest African settlements.

This entry focuses on selected early African settlements in the following areas: Egypt; south central Africa (ca. 500 to 1450 BCE); the Pre-Aksumite and Aksumite settlements of northeast Tigrai (Ethiopia) (800 BCE–700 CE); the Iron Age (Early and Later) settlements in South Africa (third to nineteenth century CE); early medieval North Africa (Morocco, Libya, and Tunisia) (650–800 CE); and West Africa (the Hausaland city-state of Kano) (post-twelfth century CE) (Figure 1).
Figure 1  Historical settlements in Africa (courtesy of Dr Rajendra Pillay and Ms Susan Abraham).
sporadic, and not well integrated, and there was a lack of reciprocal Egyptian influence on Uruk sites or in the intervening regions; by the Terminal Predynastic period (ca. 3300–3050 BCE), however, there was an extensive network of contacts between these two regions, but it dissipated very quickly thereafter.

Savage contends that in the Early Predynastic period (ca. 5000–3900 BCE) the settlements in the Hierakonpolis region (from Hierakonpolis to Asyut in the Upper Nile) appeared concentrated close to the river, and showed evidence of remnant foraging, hunting, and industry (e.g., ceramic manufacturing). By the middle of the Late Predynastic period (ca. 3650–3300 BCE), settlements were found on the Nile levees, wadi fans, and the low desert terraces at the edge of the floodplain, and were more than 4–5 km apart in the most densely populated areas. This pattern contrasted with the early settlements in the Nagada and Thin(ite) regions (downstream from Hierakonpolis), with small sites at Nagada (3 ha or less) on low terraces overlooking the floodplain and about 2 km apart, while those in Thin on the east bank of the Nile were 1–2 km apart; however, during the Late Predynastic (ca. 3650–3300 BCE) and the Terminal Predynastic (ca. 3300–3050 BCE) periods the settlements in Thin – especially at Naga-ed-Der and the Abydos villages – were more nucleated, with fewer than 100 people each. In consequence, affiliated settlements contracted, and were curtailed or abandoned.

Smaller outlying sites of various functions in the Hierakonpolis area flanked the larger, more nucleated regional centers, and included clustered agricultural settlements of simple structures with hearths and pits. Agricultural production during the Late Predynastic period in Hierakonpolis and Nagada involved the cultivation of wheat and barley, and the raising of cattle, sheep, goats, and pigs, and fishing in Nagada. Surplus production in Hierakonpolis helped facilitate trade with Lower Egypt and the Southern Levant.

The Hierakonpolis villages seemed to have developed organically, with the spaces between houses irregular in shape and size, and with larger industrial-scale production facilities up the wadis. In the regional centers of Hierakonpolis, there were larger structures that probably fulfilled religious and administrative functions. Savage (2001) notes that the houses were circular and small; or slightly larger, rectangular, and semisubterranean. In contrast, the houses in the Nagada region often contained hearths and storage pits, and sometimes subfloor burials. The sites in Nagada, which were sometimes occupied for several generations, housed 50 to 250 persons, and trash pits and animal pens were scattered among the dwellings. Furthermore, the larger communities in the Nagada region seemed to have been planned, and Nagada South Town had rectangular, mudbrick buildings inside a 2 m wall thought to be a fortification system.

**Early settlements in south central Africa (ca. 500–1450 CE)**

Kathryn de Luna (2012) writes on the early settlement in south central Africa with reference to linguistic and archaeological records that go back some three millennia, using these records to produce a regional settlement chronology in south central Africa. Her work focuses on the Botatwe language in the Kafue–Machili–Zambezi–Linyanti river basins. She argues that the Proto-Botatwe-speaking communities diverged from the ancestral Bantu-speaking societies located to the north and east of the extant Botatwe languages. Moreover, the archaeological records and linguistic evidence correspond to identify the Greater Kafue region and the area immediately to the north as the homeland.
HISTORICAL SETTLEMENT: AFRICA

of the Proto-Botatwe speech communities for about 15 centuries. Over time the speakers of Proto-Botatwe slowly shifted from their original location to establish new settlements to the west and east of the Kafue River basin (ca. 500–1700 CE).

The archaeological and linguistic records show that several activities were undertaken in these early African settlements, including shifts in settlement patterns and developments in hunting, cattle keeping, cereal agriculture, and metal work. With the spread of food production and developments in the food economy in the Greater Kafue and Batoka Plateau, smaller localized changes and sometimes delays accompanied the adoption of cereals and livestock. While the Early Iron Age settlement in the Kafue region only occurred in the good agricultural lands of the grasslands, during the Middle Iron Age (i.e., from the last quarter of the first millennium through the early second millennium) settlements were extended into new ecological niches along the rivers and in the woodlands. Also, during the Middle Iron Age the Kafue settlements extended into a greater diversity of microenvironments. For example, the Botatwe settlers in the Batoka and Falls regions experimented with farming in previously untried ecological zones.

The Botatwe community engaged in hunting, and their hunting styles changed over time. The adoption of spears (and spearcraft) (eighth to thirteenth centuries) marked a change from archery to large-enterprise communal spear hunting involving the entrapment and encirclement of the hunt. Later (during the first few centuries of the second millennium), communal hunting gave way to group hunting, especially among the Proto-Kafue group. Proto-Kafue societies had adopted a largely sedentary lifestyle and hunted antelope and wetland antelope species in the Kafue floodplain. De Luna (2012) notes, however, the high mortality rates of species (e.g., buffalo, hippopotamus, elephant, zebra, and antelope) typically hunted by communal spear-hunting groups within the Botatwe communities. Hunting was accompanied by iron tool-making, particularly in spear parts and arrow forms in the Kafue region and the Batoka Plateau.

Cattle keeping was important among the Botatwe communities in an area now known as the Central Cattle Complex (de Luna 2012). The earliest evidence of cattle in the region dates back to the last two or three centuries before the Common Era. Cattle keeping was gradually adopted into the Botatwe subsistence economy. While small numbers of cattle were kept at most settlements in south central Africa by the close of the first millennium, cattle keeping had intensified across the western and southern regions during the second half of the first millennium – thereby becoming an important element of the economy as in the emergence of intensive pastoralism in the grasslands of western Uganda in the early second millennium.

Pre-Aksumite and Aksumite settlements of northeast Tigrai (Ethiopia) (800 BCE–700 CE)

D’Andrea et al. (2008) and Fattovich (2009) present windows on the world of the early settlements in Ethiopia with special reference to the Pre-Aksumite and Aksumite settlements of Northeast Tigrai, and of Yeha (northern Tigrai), respectively. D’Andrea et al. report on 17 settlements in the Gulo-Makeda region 15 km north of the present-day town of Adigrat, while Fattovich’s work is on Yeha (west of Gulo-Makeda). The Tigrai province of Ethiopia shares a common border with Eritrea to the north, while the Pre-Aksumite and the Aksumite periods extended from 800 BCE to 400 CE and
HISTORICAL SETTLEMENT: AFRICA

400 BCE to 700 CE, respectively, with the period 450–400 BCE identified as the Proto-Aksumite. Fattovich’s focus on Yeha coincides with the Pre-Aksumite period (although he suggests that a “Pre-Aksumite culture” did not exist). D’Andrea et al. contend that, while western Tigrai (e.g., Aksum and Yeha) has been extensively researched, the eastern Tigrai had not, prior to their study, been systematically surveyed. Thus, within the context of developments at the urban center of Aksum, their principal objective was to examine the nature and role of rural economies in the development of ancient Ethiopian complex societies. The focus here, however, is on the settlements in ancient Ethiopia.

D’Andrea et al. (2008) note that the Pre-Aksumite peoples were largely an endogenous community with cultural links limited to the Gash Delta region near Kassala in the Sudan, and to the Tihama region along the South Arabian coast. Politically, the Pre-Aksumite period reflects a peripheral role for Gulo-Makeda. Fattovich (2009) contends that it was during the eighth or seventh centuries BCE that contacts between the people of central Tigrai, Agame, and Akkele Guzay with the populations of southwestern South Arabia may have existed. In contrast, the Aksumite period was characterized by a dynamic empire that dominated the southern Red Sea region, had economic and political contacts with Egypt (fourth to second millennia BCE) and South Arabia (first millennium BCE), and links via Adulis on the Red Sea to the expansion of Roman commercial interests into the Indian Ocean during the early first millennium CE. In addition, the town of Aksum had developed a monopoly of trade routes leading to central Africa, from where various products (e.g., ivory, rhinoceros horn, hippopotamus teeth, gold, and incense) were sourced for export via Adulis. Products imported into the Aksum sphere of influence in Africa included iron, glass, ceramics, textiles, wine, aromatic substances, and spices, with several settlements serving as important transfer points in the salt trade. In light of the trading patterns, settlements were mainly located along the Adulis–Aksum belt and along the North–South trade route from Kohaito to Mekelle. The Aksumite period was therefore characterized by strong cultural, social, and economic influences along several fronts, with the settlement at Gulo-Makeda serving as an important regional center that connected the Aksum–Yeha region to the Red Sea and beyond.

D’Andrea et al. (2008) note that the settlements in eastern Tigrai were more numerous, but smaller, than those in western Tigrai, and that pastoralists and farmers occupied Gulo-Makeda for at least 3000 years. Also, the inhabitants of Gulo-Makeda experienced and participated in the trade and movement of persons from Yeha and Aksum in the west, across Akkele Guzai in Eritrea, and via the port of Adulis, with South Arabia. Gulo-Makeda was characterized by continuous occupation of the larger settlements throughout the Pre-Aksumite and Aksumite periods. In addition, water availability, irrigation, and agriculturally favorable locations were critical factors in the establishment and growth of settlements. Gulo-Makeda was marked by an increasing number of settlements over time, and with a long-term trend toward increasing population density. The larger settlements in Gulo-Makeda also had monumental buildings and sometimes tombs, as was the case, according to Fattovich (2009), in Yeha (ca. seventh and sixth centuries BCE) – thereby marking the existence of a social hierarchy at the time. Also, D’Andrea et al. (2008) note that workshops and factories for the production of primitive tools (e.g., scrapers), pottery, and beads were situated close to the town centers and not near raw material sources.

In Gulo-Makeda, smaller settlements were in close proximity to the big centers to
facilitate exchange of household goods and produce. Some rural settlements were located on hillslopes, so that grazing and vegetable cultivation based on water from wells and small dams occurred in the valleys. Elsewhere, dryland cultivation took place in areas where runoff naturally concentrated and where water diversion enabled cultivation on terraced land and sediment slopes. At the same time, Fattovich (2009) notes that cattle herders who were migrating across central Tigrai, Agame and Akkele Guzay, and Hamasien in the early first millennium BCE contributed to the development of an exchange network among these people, and between the communities on the highlands and those in the lowlands and coastal regions.

Settlement in post-Byzantine (early medieval) North Africa (650–800 CE)

The political hold of the Roman empire over North Africa (i.e., classical Africa) ended with the fall of Byzantine Carthage (Roman Africa) toward the end of the seventh century (ca. 698 CE). North Africa (Tunisia, Libya, Algeria, and Morocco), according to Fenwick (2013), officially became the Umayyad province of Ifriqiya, part of Dar al-Islam, under Arab control from the eighth to the nineteenth century. Fenwick contends that the major centers and settlements in North Africa during Arab control (medieval North Africa) (e.g., Kairouan, Beja (Roman Vaga), Sousse (Roman Hadrumentum), Tunis (Roman Thuna), and Tripoli (Roman Oea)) show instances of continuity (e.g., at Cyrenaica and inland Baqqa) and difference (e.g., the decline of Carthage) from the Roman period. In short, Byzantine towns were not necessarily destroyed or abandoned following the Arab conquests (with about 30% of them still occupied until the eleventh century), and pre-existing communication networks and trade routes linking towns to one another endured. Fenwick challenges the earlier traditional model that the transition from Byzantine Carthage to Arab North Africa was catastrophic and a moment of crisis for North Africa. He adds that several cities gained new significance as administrative, military, or trade hubs, for example Beja (in northern Tunisia), Tunis, Tripolis (Libya), and Kairouan (the new inland capital of Ifriqiya). In the next section, the focus shifts to the settlements of medieval North Africa.

Fenwick (2013) notes a decline in the number of rural settlements (e.g., at Jerba on the Tunisia–Libyan border) in the Late Antique II period (500–700 CE), followed by an increase during the Early Medieval I period (700–1050 CE). During Arab dominance, rural settlements became increasingly egalitarian and nucleated in several areas, as at Kasserine during the sixth century, and in Tripolitania and central Tunisia. This phase marked the progressive abandonment of farms, villas, and estates of the Roman period; however, hierarchical differences emerged between persons living in fortified structures, and those in simple huts and buildings. Elsewhere, settlement differences persisted. For example, in northern Tunisia small and large farms, villages, and towns continued to exist (e.g., around Carthage), representing continuity from the Roman era, while in the eighth century there were also large landed estates in the north as a result of land grants to the Arab conquerors and their Berber allies. Over time (700–1050 CE), the number and density of rural settlements increased, especially in Tripolitania and along the coastal areas near Leptis Magna, while further west at Jerba there emerged fortified villages and tower–granaries. Fenwick (2013) argued that the prevalence of large estates in North Africa probably represented a throwback from the Roman times where the church, state, and local elites
owned such property. The Arabs confiscated this property as conquest booty. In general, the nature of the rural landscape in North Africa during the early medieval period was indicative of agrarian wealth.

Following the Arab conquest of Byzantium, urban settlements were not abandoned, with urban diversity an important feature during the early medieval period. Fenwick (2013) notes that in Tocra (Taucheira) – a coastal Cyrenaican town with an irregular street plan – houses of the eighth to tenth centuries were humble structures comprising mudbrick walls and earthen floors, with rectangular rooms organized around courtyards that were also used for commercial activities. Sbeitla (Sufetula) in central Tunisia had been reduced by the seventh century to a series of small, thriving settlement clusters with single-storied fortified complexes, each with an internal well, cistern, and stabling, a church, and a production site – a pattern that survived to the ninth century. In some instances, former religious and public spaces gave way to residential or industrial uses. In contrast, the urban settlement of Volubilis (renamed Walila) (in Morocco) had a Roman settlement on the western third of the city, and within the sixth century wall to the east were Christian burial and industrial (e.g., pottery kilns) zones, while further east and south were Muslim graves. Volubilis was a thriving city and had an Arab quarter with a central tower built out of spoliated Roman blocks, an area of dense housing, a public building (probably a mosque), and an eighth- to ninth-century complex comprising a bathhouse, a granary, and three large – presumably public and administrative – buildings.

Fenwick (2013) contends that, in general, a defining characteristic of the early medieval townscapes in North Africa were fortifications and intramural fortresses – with the Arab army building new forts in the Byzantine style. In addition, there were monumental buildings that served as mosques and for civic purposes, while in some areas churches remained a feature of the townscape with some new ones built (e.g., at Kairouan). Elsewhere churches were repurposed for other uses including olive oil pressing, residences, and light industry (kilns and artisan workshops), while housing became simpler (and smaller) from the mid-sixth century. The port cities also engaged in trade with the Middle East and Spain, exporting oil and wine. In general, urban settlements showed continuity and differences from the Roman to the Arab-controlled periods, while rural settlements increasingly became nucleated.

Early Iron Age settlements in south Africa

Feely and Bell-Cross (2011) and Maggs (1976; 1980) have written on the (Early) Iron Age settlements in south Africa, including the Eastern Cape region, the southern Highveld, and the area to the south of the Vaal and the Pongola rivers, respectively, with the southern Highveld corresponding largely with present-day (i.e., postapartheid) Free State province.

The Iron Age settlements in the southern Highveld were built mainly of stone in the northeast, and of reed and mud huts in the south, and occurred in large numbers. The settlements varied in size from a few units to over 100, and comprised a primary circular structure. Each settlement consisted of a number of units that included huts and livestock pens. The large primary enclosures – built as stock pens – formed the nucleus of the settlement unit, though often several smaller enclosures were clustered at the center. The huts and other domestic structures were built in a ring around the nucleus, highlighting the economic
importance of stock, especially cattle, among the southern Bantu societies. Secondary enclosures were built against and around the primary structures, and separated the living areas from the cattle pens; however, in the southwest the nucleus of the settlements comprised a group of large primary enclosures, usually three to eight in number, with thickened walls on either side of their entrances and arranged in a loose circle around the stock pens.

The settlements to the northeast of the southern Highveld date back to the fifteenth century CE, although Maggs (1976) contends that settlements dating back to the Early Iron Age – some 1000 years earlier – could have existed in the area. These settlements were generally located on the tops or slopes of volcanic (dolerite) intrusions which were quarried on-site to produce the stone for dry wall construction. The settlements along the ridge or scarp crest were not contiguous, with cultivation and grazing in the valleys and along the slopes. Settlement density increased over time as new settlements were built in between older ones. Also, the drier conditions in the western part of the southern Highveld favored stock farming rather than cultivation.

Maggs (1980) contends that the earliest “expression” of inhabitants in the area to the south of the Vaal and Pongola rivers in southern Africa dates back to the Early Iron Age (EIA) (third to fourth centuries CE). The inhabitants were farmers, and excavated pottery in these areas was similar to those found in Matola, near Maputo in Mozambique, in eastern Kenya and Tanzania. Expressions of occupation south of the Vaal and Pongola rivers are lacking from 400 CE to 600 CE, however, suggesting a break in occupation during these years. Settlements reappear in the post-600 CE period, with patterns in the Late Iron Age (LIA) (tenth century onward) suggesting the presence of a different type of community in the area.

The settlements found in the area to the south of the Pongola River exhibit expressions of livestock farming, hunting, food gathering, and iron smelting in the EIA period. Evidence of cultivation is scattered and drawn from expressions of large, semipermanent villages with pits that served as storage for grain. Crops cultivated included millet, sorghum, cow peas, and cucurbits. EIA settlements were confined to sites below 1000 m, especially along the narrow coastline and further inland along the incised river valleys, and in areas of high rainfall (>1000 mm per annum). Inland settlements were also confined to the deeply incised river valleys with deep alluvial or colluvial soils, moderate rainfall (700–800 mm per annum), sweet grazing, and the availability of forests (for firewood and timber). Maggs (1980) notes that valley settlements during the EIA comprised large villages, often around 8 ha in extent, while those along the coast were smaller and more dispersed.

The Eastern Cape EIA farmers also visited the coastal areas to harvest marine shellfish.

Settlements during the LIA were located on hilltops and in the valley slopes in southern Africa, were small and dispersed, and comprised small groups of people who engaged in hunting and gathering. Later (1400 CE onward), settlement expanded into the grassland areas of the southern Highveld, western Lesotho, and KwaZulu-Natal province to secure livestock grazing opportunities, with transhumance a feature of the time. Given the absence of timber in
Settlements in West Africa (ca. 1500 CE)

Nast’s (1996) article, grounded in Foucault’s notion of archaeology, focuses on Islam, gender, and slavery in West Africa around 1500, with particular reference to Kano palace in northern Nigeria. Nast incorporates important elements on settlements in pre-Islamic and Islamic West Africa. Kano City is located in the southern part of Hausaland in present-day northern Nigeria and southern Niger, with radiocarbon dating of iron-smelting furnaces in the Dalla Hills region highlighting that the area was used in the seventh century CE. While Kano palace represented, perhaps, the oldest, largest, and most important palace ever built in sub-Saharan Africa, Kano city ranked with Fez and Cairo among the most important commercial centers of Africa around 1500. The prominence of Kano (formerly Dalla; renamed in the late fifteenth century) as a major economic, political, and cultural center was attributed to its Islamization, the rise of long-distance trade across the Sahara and North and West Africa, and the capture and consolidation of new territories.

The Kano city region was in the Late Stone Age part of a settlement cluster associated with six hills to the north of the Challawa River, with farming, hunting, and fishing the principal human activities. The Fangwai and Santolo settlement groups (located some 25 km to the southeast) initially controlled the Dalla Hill settlements, but were later defeated and scattered by the Dalla Hill groups. Subsequently, the walled city-state of Kano was established, with its Muslim cavalry-based aristocracy and merchant class, and slaves drawn from conquered territories (e.g., Nupe in the south). The establishment of Kano city was also linked to the introduction of a (grain) tax system of one-eighth of production during the reign of King Naguji at the end of the twelfth century, the development of a large and intricate state bureaucracy, and the strengthening of territorial alliances through marriages as a political tool in the 1400s.

Unlike the rest of Africa, matters of production within Kano’s sphere of influence were largely gendered, cultural, and class-based, with distinct separation of responsibilities between the royal household, men, women, eunuchs, and slaves; between adherents to the Islamic faith and those who were non-Muslims (pagan or Maguzawa); and on the basis of age. For example, while Muslim women were largely secluded and engaged their prepuberty daughters as trade intermediaries hawking their crafts, Maguzawa women undertook agricultural production on family and private fields; engaged in spinning, weaving, and beer brewing; were less stringently restrained in their sexuality and daily movements; and intermingled spatially with men both within and outside the household. Essentially, the Islamization of West Africa – especially during the reign of Muhammadu Rumfa (1463–1499) – affected gender relations along sociospatial lines in all spheres of life.

Rumfa’s reign also marked the construction of Kano’s central (Kurmi) market and the massive Kano palace. The palace was built in a separate walled suburb outside the walled city and, for the first time, it rooted the state in one place. The suburb and the palace – as in (Muslim) North Africa and Europe at the time – were rectangular in shape, with the southern portions of the palace aligned to point to Mecca, the Muslim holy city. The palace was also characterized by a strict spatial division along gender and class lines that included royal wife seclusion and massive concubinage, the formal exclusion of women from matters of state, the rise of a state slave
bureaucracy, eunuchs performing state treasury functions, and a special cadre of underclass slave women who tended to the needs of a leisure class of royal women in the inner palace and undertook other menial chores (e.g., laundry). In sum, Kano emerged as a major economic and cultural center, and became a political force that dominated the surrounding hill settlement systems. In addition, Kano exerted considerable influence over northern Nigeria, southern Niger, and West Africa. The Islamic dominance over West (and North) Africa from about the fifteenth century continued somewhat unabated until European states (Holland, England, and France) pursued their separate colonial interests in West (and North) Africa in the seventeenth and eighteenth centuries (Newton 2013).

Conclusion

From North Africa (Tunisia, Libya, Algeria, and Morocco) (650–800 CE), through the Nile Region in Egypt, across the Pre-Aksumite and Aksumite settlements of northeast Tigray (Ethiopia) (800 BCE–700 CE), into West Africa (Kano city), and then southward into south central Africa (the Kafue region), and south Africa, the nature, designs, structures, layouts, locations, functions, and the socio-politico-economic and cultural attributes of the historic settlements of Africa reflect the rich diversity and multifaceted characteristics of the different communities that settled these areas in earlier times.

Acknowledgments

The author thanks Dr Rajendra Pillay of the Department of Geography and Environmental Science, University of Fort Hare (UFH), South Africa, and Ms Susan Abraham for their kind assistance in producing the figure for this entry; and Mr Z. Ntaka, graduate student at UFH in 2014, for his kind assistance with the research for this entry.

SEE ALSO: Historical settlement; Historical settlement: Asia

References


Further reading

The Holocene is the most recent geological epoch and is still ongoing. Charles Lyell introduced the term “Holocene” in 1839 to describe strata in Sicily that had all their molluscan fauna still alive today. This distinguished it from the older Pleistocene epoch with the deposits containing about 30% of extinct molluscan taxa. Lyell constructed the name Holocene (“wholly new” or “entirely new”) from the Greek ὅλος, hólos, “whole,” and καινός, kainós (latinized as cænus). Both tradition and widespread current usage have accorded the Holocene a high geochronological rank of epoch, as is also the case with the Pleistocene. Together they are placed within the Quaternary period, the most recent interval of the Cenozoic era.

The first classification of the climatic phases of the Holocene was based on the study of peat bogs in northern Europe. In 1876 Axel Blytt (Norway) hypothesized that the darker oxidized peat layers with buried pine stumps were deposited in drier conditions, while the lighter layers accumulated in moister conditions. Blytt applied his term “Boreal” to drier (continental) times and “Atlantic” to wetter (oceanic) times. In 1908 Rutger Sernander (Sweden) further developed Blytt’s classification, taking into consideration changes in temperature. He also defined Subboreal and Subatlantic periods within the Holocene and added the subdivisions of the Late Glacial (the Older Dryas, Allerød, and the Younger Dryas) to the sequence. The classification was later incorporated into a sequence of regional pollen zones defined by Lenart von Post (Sweden), one of the founders of palynology. Other paleobotanists, palynologists, and Quaternary geologists have since added information to further develop and improve the Blytt–Sernander climatostratigraphic periodization of the Holocene.

At the beginning of the twentieth century Gerard Jacob De Geer (Sweden) worked out the first timescale for the Holocene and the preceding Late Glacial based on the studies of varve (annual laminated) deposits in southeastern Sweden. The lower boundary of the Holocene was then placed at approximately 8700 years ago, at the time when the retreating Scandinavian ice sheet separated into two parts. Later the Blytt–Sernander sequence has been substantiated by a wide variety of dating methods, first of all, by radiocarbon dates obtained from organic remains and sediments. Although the climatic fluctuations during the Holocene were more complex than Blytt–Sernander classification can identify, it still has been widely used until lately as a temporal framework for the large-scale interregional correlations and comparisons. The Holocene is currently subdivided, in chronological order, into Preboreal, Boreal, Atlantic, Subboreal, and Subatlantic stages of the Blytt–Sernander scheme. According to the modern view, the lower boundary of the Holocene corresponds to the beginning of the rapid warming at about 11700 years before 2000 CE, which corresponds to approximately 10.3$^{14}$C kyr BP (Walker et al. 2009).

Worldwide practical application of radiocarbon dating of organic remains made it
possible to determine the age of the major climatic events and boundaries within the Holocene and provided an opportunity for spatial climatic reconstructions of definite time intervals (so-called time slices) for large areas, up to a hemispheric or global scale. The use of the accelerator mass spectrometry (AMS) further expanded the capabilities of the radiocarbon method and has significantly clarified the Holocene chronology, as AMS $^{14}$C dating can be used for such deposits and regions where the use of conventional radiocarbon analysis was not possible due to low organic contents. Apart from the radiocarbon analysis, a broad variety of methods is currently used for creating the timescale for the paleogeographic reconstructions of the Holocene, including varve counting, calculations of annual ice layers, measurements of cosmogenic isotopes accumulation, tephra- and dendrochronology, lichenometric analysis, and archaeological and historical data. For a correct comparison of the dates obtained with the radiocarbon method and chronologies in calendar (astronomical) years built using other methods, calibration of the radiocarbon dates is required, that is, making corrections for changes in $^{14}$C content in the atmosphere with time.

The Holocene is perhaps the most intensively studied interval of recent geological time. The paleogeographical records of the Holocene contain a wealth of detail on such diverse phenomena as climate change, geomorphological processes, glaciers and permafrost dynamics, sea level changes, vegetation developments, faunal migrations, and human evolution. Studies of peat, lake and marine deposits, soil and loess profiles, speleothems, ice cores, and so forth by a broad variety of methods show a general correspondence of the major landscape and climatic changes during the Holocene across Eurasia and North America and to some extent also globally. At the regional level, changes in the water balance are estimated from variations in the lake levels, in the morphology of meandering river channels, in frontal positions of mountain glaciers, and other natural “archives.” In addition to a variety of high-precision analyses of ice isotopic composition and that of inclusions in the ice, ice cores drilled on the Greenland and Antarctic ice sheets provide a high-resolution timescale for the Holocene, which is widely used as a basis for the interregional correlations of the Holocene landscape and climate reconstructions.

The present-day research of the Holocene is characterized by growing interest in quantitative reconstructions of climate, increasing their accuracy and time resolution, improving methods of reconstruction based on pollen data, broadening the variety of the paleoclimate indicators, and comparing the reconstructions obtained by independent methods. A large variety of proxy data is applied to the reconstructions of the climatic and environmental changes during the Holocene on the continents, including the detailed studies of the sediment’s grain size, chemical, mineral, and isotopic composition, magnetic properties, and so forth. Changes in the composition and geographical distribution of various groups of organisms, for example, large and small mammals, mollusks, and insects (especially Coleoptera), are also widely used for paleoecological reconstructions. In the studies of marine and lacustrine deposits paleoclimatic information is derived from changes in the composition of various groups of aquatic organisms (Foraminifera, Ostracoda, Diatomeae, Chironomidae) as well as from the isotopic composition of foraminifera tests and mollusk shells, and so forth.

Despite a relatively short duration of the Holocene, significant changes in climate and landscape occurred during this interglacial epoch. A rapid warming developed at the transition from the Late Glacial to the Holocene in all latitudinal zones of both hemispheres.
Data on the vertical growth of corals indicate that sea level rise caused by the melting of continental ice sheets continued with variable speed up to about 6 \(^{14}\)C kyr BP (\(\approx 7000\) cal year ago). In the early Holocene (11 700–8000 cal year ago) the sea level rise slowed down, which reflects a reduction of the meltwater discharge from the Antarctic and Laurentide ice sheets.

A large number of reconstructions of the climatic changes during the Holocene were based on various proxy data from specific terrestrial or marine sites. Based on such reconstructions, generalized paleotemperature curves were created to show the mean changes in time for large regions and, ultimately, for the global scale. Among the earliest generalizations of the kind, one can mention the paleotemperature curve worked out by N.-A. Mörner on the basis of changes in the isotopic composition of lacustrine carbonates in southern Sweden, the reconstruction of paleotemperatures in the eastern Mediterranean made by Nesteroff with co-authors, and the curve of the mean July temperature compiled by Borzenkova on the basis of palynological data for the latitudinal belt of 60–75\(^\circ\)N in Europe. These reconstructions showed the approximate synchronism of the most significant warming and cooling phases in different regions and latitudinal zones of Europe during the Holocene.

Based on pollen data from northern Eurasia, N.A. Khotinsky reconstructed three main thermal maxima within the Holocene: Boreal (8.9–8.3 \(^{14}\)C kyr BP), Atlantic (6–4.7 \(^{14}\)C kyr BP), and Subboreal (4.2–3.2 \(^{14}\)C kyr BP). The Boreal maximum was most clearly manifested in Siberia and the Russian far east, while the Atlantic and Subboreal ones were more conspicuous on the Russian Plain. Despite these regional differences, only the Late Atlantic phase marked by an optimal ratio of heat and moisture supplies can be considered the Holocene Climatic Optimum (HCO) across the whole of northern Eurasia. According to the reconstruction by I.I. Borzenkova, in mid-latitudes of the Northern Hemisphere (NH) the most significant warming phases corresponded to the early Boreal (9.0–8.6 \(^{14}\)C kyr BP) and early and Late Atlantic (7.9–7.5 and 6.2–5.3 \(^{14}\)C kyr BP). The Late Atlantic’s warming corresponded to the greatest sea level rise during the Holocene (\(\approx 2\) m higher than at present) and to the maximum expansion of coral habitats in the Pacific Ocean, which required the surface water temperature to be not lower than 20.5\(^\circ\)C. The peak of this warming (about 6–5.5 \(^{14}\)C kyr BP) is considered the HCO in a global context.

Spatial reconstructions of the main climatic indexes, based mainly on palynological data, were conducted for the HCO (6–5.5 \(^{14}\)C kyr BP) in the Northern Hemisphere by A.A. Velichko and colleagues. The reconstructions indicate a general decrease in positive temperature deviations from present-day values from high to mid-latitudes. South of approximately 40\(^\circ\)N the mean annual temperatures at the HCO were close to the present-day values. North of approximately 65\(^\circ\)N annual precipitation exceeded present-day values by 50–100 mm. In Eurasia, within the latitudinal belt between 60\(^\circ\) and 40\(^\circ\)N the rainfall was similar to present-day one, except for the areas with small negative deviations (\(-25\) mm per year) in the Russian and West Siberian plains. A substantial reduction of precipitation (over \(-100\) mm per year) occurred within the same latitudinal belt in eastern North America. The greatest positive precipitation anomalies (+100 mm per year and more) were reconstructed for Central Asia and North Africa on the basis of paleobotanic and archaeological data, as well as the data on the lake levels.

During the last decade several research papers provided reconstructions of macro-regional or global temperature anomalies for the entire
Holocene or for a large part of it, derived from comparative analyses of proxy-based reconstructions from various sites (e.g., Marcott et al. 2013; Mayewski et al. 2004; Wanner et al. 2008). In these studies, the preference was given to the longest and continuous data series of the Holocene with the highest time resolution. Based on the data from 73 globally distributed records for the past 11,300 years, Marcott et al. (2013) showed that the early Holocene (10,000 to 5,000 years ago) warming was followed by approximately 0.7°C cooling through the Middle to Late Holocene (after 5,000 years ago), culminating in the coolest temperatures of the Holocene during the Little Ice Age about 200 years ago.

Although the Holocene climate is conventionally thought to have been relatively stable, evidence from North Atlantic deep sea cores revealed that it was punctuated with episodes when cool, ice-bearing waters from north of Iceland penetrated as far south as the latitude of Britain and the atmospheric circulation above Greenland changed abruptly. These abrupt shifts (so-called Bond cycles) appear to be the most recent manifestation of a pervasive millennial-scale climate cycle operating independently of the glacial–interglacial climate state, apparently due to solar variability and the response of the global climate system to these oscillations.

Mayewski et al. (2004) compared about 50 globally distributed high-resolution climate proxy records to estimate the degree of climate variability in the Holocene. This study revealed six periods of rapid climate change (RCC) during the time periods 9000–8000, 6000–5000, 4200–3800, 3500–2500, 1200–1000, and 600–150 cal year ago. Most of these RCCs were characterized by polar cooling, tropical aridity, and major atmospheric circulation changes, although in the most recent interval (600–150 cal year ago) polar cooling was accompanied by increased moisture in some parts of the tropics. Comparison of paleoclimatic records with climate forcing time series suggests that changes in insolation related both to orbital variations of the Earth and to solar variability played a central role in the global-scale changes in the climate of the Holocene.

Bradley (2003) examined the role of the main forcing factors in the Holocene climate change. These factors operate on different timescales: lower frequency (millennial-scale) climate changes associated with orbital variations, century-scale variability associated with solar activity, and annual- to decadal-scale variability associated with volcanic forcing. Feedbacks within the climate system may involve nonlinear responses to forcing, especially if critical thresholds are exceeded. In addition, there may be distinct regional climate anomaly patterns that result from certain types of forcing. Some anomalies that appear in Holocene paleoclimatic records may be unrelated to external forcing factors, but reflect conditions entirely within the climate system.

Wanner et al. (2008) used selected proxy-based reconstructions of different climate variables together with time series of natural forcings (orbital and solar variations, large tropical volcanic eruptions, land cover, and greenhouse gases) to explain the climate changes from the mid-Holocene (≈6000 years ago) to preindustrial time. Their comparison showed that the redistribution of solar energy, due to orbital forcing on a millennial timescale, was the cause of a progressive southward shift of the summer position of the Intertropical Convergence Zone in the Northern Hemisphere, accompanied by weakening of the monsoon systems in Africa and Asia and increasing dryness on both continents. The shift was associated with summer cooling of the NH, combined with changing temperature gradients.
in the oceans. Although Wanner et al. (2008) did not find a worldwide coincidence between solar irradiance minima, tropical volcanic eruptions, and decadal to multi-century cooling events, they showed that such cooling events, accompanied by advances of mountain glaciers, occurred in the NH (e.g., in Scandinavia and the European Alps) during the Late Holocene. The most significant of them, the Little Ice Age, occurred between 1350 and 1850 CE when the lower summer insolation in the NH due to orbital forcing coincided with solar activity minimum and several strong tropical volcanic eruptions.

Nevertheless, it is impossible to explain some of the abrupt climate fluctuations that occurred during the Holocene by the influence of the forces external to the global ecosystem. The most abrupt and widespread of such events occurred between 8400 and 8000 cal year ago, when the temperature dropped by 4–8°C in central Greenland and 1.5–3°C at marine and terrestrial sites around the North Atlantic Ocean. It also appears to have been generally cool over much of the Northern Hemisphere throughout this interval, as evidenced by increased amounts of ice rafted debris, strengthened atmospheric circulation over the North Atlantic and Siberia, and mountain glacier advances in northwestern North America and Scandinavia (Mayewski et al. 2004). D.C. Barber and colleagues hypothesized that the so-called 8.2-kyr cold event was forced by a massive outflow of fresh water from glacial lakes, originally dammed by a remnant of the Laurentide ice sheet, into the Labrador Sea. A sudden increase in freshwater flux reduced sea surface salinity and altered ocean thermohaline circulation, resulting in dramatic reduction of heat transfer from the ocean to the atmosphere in the North Atlantic and adjacent regions. Thus, the 8.2-kyr event is likely to be similar in its origin to the younger Dryas cold stage and other abrupt climate fluctuations of the Late Glacial linked to changes in thermohaline circulation of the ocean.

A detailed global temperature reconstruction for the past 1500 years carried out by Marcott et al. (2013) shows a cooling trend from a warm interval from approximately 1500 to 1000 years ago (so-called Medieval Climatic Optimum) to a cold interval (≈500 to 100 years ago) that is approximately equivalent to the Little Ice Age. Global temperatures of the past decade have not yet exceeded peak interglacial (that is, the Holocene optimum) values but are warmer than during approximately 75% of the Holocene temperature history. Therefore, the mean global temperature has risen from near the coldest to the warmest level of the Holocene within the past century, reversing the long-term cooling trend that began after the HCO, about 5000 years ago.

Recently, the Global Stratotype Section and Point (GSSP) for the base of the Holocene Series/Epoch has been ratified by the International Union of Geological Sciences. The base of the Holocene is defined in the North Greenland Ice Core Project’s ice core drilled in the year 2003 in the central Greenland ice sheet at 75°06’N and 42°18’W (Walker et al. 2009). In the NorthGRIP ice core record, the horizon at a depth of 1492.45 m shows the clearest signal of rapid climatic warming, an event that marks the end of the last cold episode (Younger Dryas Stadial/Greenland Stadial 1) of the Pleistocene and the beginning of the Holocene. This climatic event is most clearly reflected in an abrupt shift in deuterium excess values, followed by more gradual changes in δ¹⁸O, dust concentration, and annual layer thickness. A timescale based on multi-parameter annual layer counting and constrained by bracketing tephra layers provides an age of 11 700 year before 2000 CE for the base of the Holocene.
SEE ALSO: Ice sheets; Sea level rise

References


Further reading


Home

Alison Blunt
Queen Mary University of London, UK

From the domestic scale to broader ideas about belonging, privacy, and security, the home is a rich and diverse subject for geographical research. Often connected to, but distinct from, work on housing, household, and family, geographical research on home investigates imaginative and material spaces of home and the relationships between them, the significance of home from the domestic dwelling to the nation and beyond, and the ways in which home is bound up with identity, power, place, and mobility. For Blunt and Dowling (2006, 10): “home is a series of feelings and attachments, some of which, some of the time, and in some places, become connected to a physical structure that provides shelter. Conversely, one can live in a house and yet not feel ‘at home.’”

Humanistic geographers writing in the 1970s and 1980s provided the first sustained geographical engagement with ideas about home. In contrast to the abstractions of spatial science and structural Marxism, humanistic geographers valued the relationships between people and place, often expressed most eloquently through a love of home. For many humanistic geographers, home was the central anchoring point for human life, the heart and hearth of human creativity and self-expression. Home was a unique place – and a uniquely human place – to be celebrated as authentic and meaningful. Humanistic geographers trace the poetics of home in relation to both intimate and wider worlds, writing evocative accounts that bind individuals and their dwellings to a sense of home in the wider cosmos. Other geographers have studied the home as a “more-than-human” place, whether through the entanglements of nature and culture within and beyond the built fabric of a dwelling or in the space of the domestic garden.

The positive meanings of home celebrated by many humanistic geographers have been subject to significant critique by feminist geographers for ignoring the power relations and inequalities that shape experiences of home, particularly for women. As many feminists working in geography and beyond have shown, the home is a gendered place and is often a site of fear and violence rather than security and belonging (Brickell 2012). The home is an important and yet contested site in feminist debates about power and identity, often articulated through life stories in a variety of forms (Varley 2008). Some feminists have written about the political limits of home, implying a bounded and stable identity, underpinned by the exclusion of difference and a nostalgic appeal to unity and privilege. Others, in contrast, explore the political possibilities of home and critique ideas about home as a necessarily oppressive and apolitical space, including bell hooks’s (1991) influential work on “homeplace as a site of resistance” for black women.

Inspired by such research, Blunt and Dowling (2006) developed a “critical geography of home,” which argues that material and imaginative geographies of home are closely intertwined; that home, power, and identity are intimately linked; and that geographies of home are multi-scalar. Home takes material form in a wide range of ways, including through domestic architecture, interior design, domestic material culture, and homemaking practices. And yet, as Blunt
and Dowling argue, the materialities of home are closely shaped by and understood in relation to imaginative geographies of home. As they write, “the material form of home is dependent on what home is imagined to be, and imaginaries of home are influenced by the physical forms of dwelling” (2006, 22).

The relationships between home, power, and identity are also closely tied to the ways in which “[w]hat home means and how it is materially manifest are continually created and re-created through everyday home-making practices, which are themselves tied to spatial imaginaries of home” (Blunt and Dowling 2006, 254). The home is an important site for studying the politics of identity in relation to inclusion, exclusion, belonging, and inequality, particularly in terms of gender, class, age, sexuality, race, and religion. Feminist and postcolonial research on home as a site of both power and resistance has been important in this context, as shown by work on the relationships between home, nation, and empire; the inclusions and exclusions that underpin the geopolitics of home and homeland; and the politics of dwelling and belonging in relation to indigeneity, migration, and diaspora.

Connected to this interest in home within but also beyond a domestic dwelling, the third part of Blunt and Dowling’s “critical geography of home” concerns the multiscalarity of home, whereby “[T]he relations of domesticity, intimacy and belonging that construct home not only extend beyond, but also help to produce scales far beyond the household” (2006, 257). Research in this context spans the political mobilization of home in imperial politics, anti-imperial nationalism, and the contemporary politics of homeland (in)security; the transnational built form of domestic dwellings such as the bungalow and high-rise; the transnational and/or diasporic importance of homemaking practices and domestic cultures alongside wider ideas of home, belonging, and attachment; and the employment of domestic workers in the global economy.

Geographical research on home therefore spans past and present homes, home on domestic to global scales, and positive meanings of home alongside the home as a site of violence, fear, and oppression. Such research is part of a wider focus on home across the humanities and social sciences, and has made a distinctive contribution through its focus on the spatiality of home, the relationships between home and place, and home across multiple and coexisting scales. Examples of the wider cross-disciplinary interest in home include the journal *Home Cultures*, which was launched in 2004 and is dedicated “to the critical understanding of the domestic sphere, its artifacts, spaces and relations, across timeframes and cultures;” and the Centre for Studies of Home, a partnership between Queen Mary University of London and The Geffrye Museum of the Home, which spans “research on the home and domesticity (including everyday domestic life, domestic architecture, interior design, and domestic material cultures) to the significance of home beyond the domestic (including broader ideas about dwelling, belonging, and security).”

**SEE ALSO:** Domestic spaces; Domestic workers; Feminist geography; Housing; Humanistic geography

**References**


Brickell, Katherine. 2012. “‘Mapping’ and ‘Doing’ Critical Geographies of Home.” *Progress in*

Homelessness

Erin DeMuynck

University of Wisconsin–Fox Valley, USA

The notion of home is associated not only with housing or a material structure but also with social relations and feelings of security and attachment. From this perspective, it is possible to be homeless but not unhoused or, perhaps, to be unhoused but not homeless. Homelessness is a complex concept that can be understood as composed of a number of experiences that lie along a continuum. At one end of this continuum, the visible homeless sleep and carry out most of their daily activities in public. At the other end of the spectrum, the hidden homeless have access to shelter but may live in unsafe or highly insecure conditions. The visible homeless tend to be more easily identified visually, while those who are part of the hidden homeless population may blend in to varying degrees with the nonhomeless population. During a lifetime, individuals may move back and forth along the continuum between visible and hidden homelessness. They may also move in and out of homelessness one or multiple times over the course of their lives.

Across types, homelessness can be understood as both an effect of severe social and economic marginalization and subsequently as a cause of further marginalization. While homelessness has been a feature of many societies throughout space and time, it tends to broaden and deepen or to diminish depending on social, economic, and political factors. Contemporary homelessness can be understood as an enduring outcome of the deepening poverty and inequality that have accompanied globalization, economic restructuring, and the dismantling of social safety nets. Deindustrialization and the increased mobility of companies initiated an increasingly asymmetrical system of income distribution in which many stable middle-class jobs have been replaced with contractual and part-time precarious employment opportunities at a lower pay scale and without employment benefits such as health insurance. This trend led to a growth in poverty, which was met by a retreat of the state at the national scale. The obligation for human service and welfare provision was then transferred to state and municipal levels. Neither state and local government nor the private sector has been able (or willing?) to meet the growing demand for human services. Today there continue to be calls for cutbacks in welfare spending, tightened eligibility requirements, and renewed efforts at welfare state reorganization at all levels of government.

As the numbers of people in need increased and national funding for human services was severely retrenched, local governments acted. However, rather than addressing causes of the growing social need or confronting it directly, municipal policy responses served to further marginalize people who are homeless and those who are vulnerable to homelessness. As a mode of adaptation to severe federal funding reductions and a postindustrial reality, available resources were redirected away from human welfare and toward pro-growth and pro-business strategies such as corporate tax breaks, privatization of public services, the creation of public–private ventures, and cultural development and image-building projects. The goal was to attract

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0991
an economically active population back to the city at a time when funds for services and infrastructure had been drastically reduced, unemployment was rising, and social need was high. As a result, cities have been reimaged as cleaned up, safe, and welcoming places for business owners, developers, and wealthy tourists and residents. This urban policy focus remains ascendant today.

People experiencing homelessness have no place in this kind of city. Revanchism – taking back the spaces of the city from those who make them undesirable to capital – is achieved through surveillance, policing measures, and regulations aimed at unwanted populations, and it keeps revitalized parts of the city free from the homeless and poor. These efforts criminalize homelessness and prevent individuals who are homeless from entering and staying in public spaces, where they are regarded as out of place. Their rights to the city are subordinated in the name of making spaces more attractive to middle- and upper-class residents and consumers. The results of this policy approach can be visibly seen in cities in the form of a deeply uneven landscape in which vibrant, highly designed, and maintained spaces are produced in showcase urban districts while the poor are shunted into increasingly neglected and deprived parts of cities. This kind of gentrification in privileged areas and the concomitant concentration of poverty elsewhere put pressure on housing markets and have the effect of driving up prices, most intensely at the low end of the market.

In the midst of converging challenges involving reduced wages, reduced resources for the poor, increased housing costs, and revanchist policies, people are actively blamed for their homelessness. Moreover, they are blamed for broader urban problems of unemployment, underemployment, and poverty. In this neoliberal era of hyperindividualism, all problems are considered the responsibility of the very individuals who are victimized by larger systemic and structural forces. As a result, people experiencing homelessness are constructed as not only out of place but also undeserving of assistance or care. In this climate of exclusion and disdain for the other, the impacts of homelessness on the physical and psychological health of individuals can be particularly harsh. While physical and mental health conditions may have predated the homelessness of some, this is not necessarily the case for all. Homelessness itself and its logic of exclusion are often the causes of illness and pain rather than homelessness being an outcome of pre-existing physical or mental illness.

The intensification of inequality and exclusion has also served to broaden the reach of homelessness and extend it to new spaces and populations. Young African American men were hardest hit by deindustrialization and became homeless at high rates during the 1980s, and racial inequality continues to be a factor in access to jobs and compensation today. Women and children have also joined the homeless population at higher rates as a result of deepening inequality and poverty, and decreased access to services and affordable housing. Gender identity and immigrant status are two additional factors associated with homelessness. LGBTQ youth are disproportionately represented among the homeless. Like homeless youth in general, LGBTQ youth may run away or be forced out of the home as a consequence of severe family conflict, abuse, or neglect; for LGBTQ youth, this happens at a higher rate. Immigrant youth are another overrepresented population. Those who arrive in the United States without a place to go or family to take care of them are at risk of becoming homeless.

Personal factors can increase vulnerability to homelessness for anyone. For example, limited access to education or training, mental or physical disability, lack of family to rely on, and
alcohol or drug addiction increase vulnerability. Yet, it should be noted that, while structural constraints compounded by individual vulnerabilities impact the everyday realities of people who are homeless on a deep level, individuals experiencing homelessness also exercise agency every day in their actions and decisions. They actively establish social networks, develop complex relations with people and places in the city, find moments of enjoyment, and create a life for themselves. These kinds of adaptive survival patterns may or may not make it easier to move out of homelessness in the long term; however, they do make it clear that individuals who are homeless should not be seen only as victims but also as people who have the power to develop alternative ways of understanding and being in urban space.

**SEE ALSO:** Gentrification; Housing; Neoliberalism; Poverty; Public space

**Further reading**


Housing

Daniel J. Hammel
University of Toledo, USA

Housing is an extraordinarily broad topic that touches on a vast range of human activities. It is also a topic that spreads across numerous subfields in geography and has significant cultural, economic, and environmental dimensions. Despite the breadth of the topic, there are relatively few geographers who claim a research specialization in housing. Yet, many geographers focus on related areas that draw from and contribute to a larger literature on housing in the social sciences.

The meaning of housing

Housing has both economic and cultural significance. In addition to fulfilling a basic human need for shelter (included in Article 27 of the Universal Declaration of Human Rights), housing also has legal and political ramifications, including the right to privacy.

The economic component of housing is perhaps the most commonly discussed and studied, particularly in the more industrialized world. Housing, like any economic good, has both use- and exchange-values. The use-value relates to the role of housing as a basic human right and also as a provision that structures people’s quality of life. The exchange-value of a house, referencing its economic worth, has become substantially more complex in advanced capitalist economies. Intricate procedures and laws have been developed that allow houses to be bought and sold with relative ease. Housing markets, which are often specific to individual cities, have developed with particular sets of characteristics relating to the processes of exchange. This formalization of the market allows housing to be treated as an investment. There is some argument that the investment component of housing is as important as the shelter component, especially for middle- and upper-class homeowners (Adams 1984). The investment component has enabled wealthier homebuyers in strong markets to purchase houses, extract a use-value from them for years, sell their houses, and realize substantially more than the original purchase price even when adjusted for inflation. The investment potential of housing is unevenly distributed at the global and national scales, and even within submarkets at the metropolitan scale. In general, moderate- and lower-income homebuyers and renters do not often receive a positive return on investment in their houses. Housing markets are closely linked to the finance industry through mortgages and a wide array of investment instruments brought about after deregulation in the financial and housing finance industries. These issues will be discussed in more detail in the “Housing Finance” section.

The cultural and social components of housing have also been the subject of significant attention by scholars. Housing can be understood in a number of different ways. One of the most important aspects of purchasing or renting a home is that it provides individuals with a physical space in which they may claim a level of privacy for conducting intimate human activities. Housing has also been an instrument of personal expression and identity formation, and some of these functions have been expressed...
in larger society shifts. David Ley (1980) suggests that cultural shifts throughout the 1970s created a new class of urban dwellers who defined themselves through the consumption of architecturally appropriate inner-city houses, which he claims became a driving force of gentrification. Housing has also been understood as a site of consumption. In Marxian thought consumption is a key facet of capitalist economies and the private residence is an important locus of consumption activities. While houses clearly have an exchange-value, the culture of consumption located in the home is an essential part of the capitalist economy.

The importance of the economic and cultural value of housing clearly varies over time and space, but it also varies across class. A traditional neoclassical view of housing has implicitly suggested that middle- and upper-class homeowners place a greater emphasis on the exchange-value and the investment potential of housing than lower-income individuals. This emphasis guides decisions about which house to purchase and what improvements to make to the house. These assumptions, even if once true, have not held up well since the 1990s. Housing is increasingly marketed and sold as an investment, despite the fact that gains in housing value adjusted for inflation appear to be short-lived and are only substantial in a few locales. In addition, most working-class individuals have very limited housing choices, confined to areas where price increases lag well behind those in wealthier areas.

The geography of housing

Geographers have paid a great deal of attention to the location and distribution of housing in cities. While the classic models of urban structure were not specifically housing-based models, they clearly identified strong patterns of housing. The earliest of these, the concentric ring model, used the terms “housing,” “house,” or “home” frequently, but not consistently, in the description of the ecological areas in Chicago. The sector model focused on neighborhoods and housing (Homer Hoyt trained as a real estate economist). While later models included other activities and some focused on commercial land uses, housing has always been seen as a key component of the geography of cities.

The theoretical basis of the models, particularly the social ecology of the Chicago School, has long been the subject of substantial critique. Yet the models, or at least the classic three – Burgess’s concentric ring model, Hoyt’s sector model, and Harris and Ullman’s multiple nuclei model – almost always appear in human geography and urban geography textbooks. Their inclusion is in part because these models are historically important, but also because they identify empirical regularities that are present in many American cities (and the American cities of nearly a century ago). There is a general pattern of lower-income housing around the central business district, despite decades of gentrification in many cities, and high-income housing in far-flung suburbs; although there are certainly poor and declining suburbs as well. The sectoral pattern of wealth and poverty that Hoyt observed is remarkably consistent in eastern and Midwestern American cities, although the complicated geographies of large metropolises resist simple description.

In the latter decades of the twentieth century several models were created that attempted to describe the structures of Latin American cities and Asian port cities, and clearly showed that the classic American models did not fit the rest of the world. Historical patterns of development, politics, culture, and a range of other forces help determine the geography of housing described
in these models, but this entry will turn first to the housing market to help understand the basics of housing geography.

The geography of housing depends a great deal on the workings of housing markets. There is frequent reference in the popular press and even in some economics literature to a housing market. Markets operate at the national level and are linked tightly to the performance of the national economy. While there are clearly national policies that influence and enable housing markets, in a country of any size there are no real national housing markets. A house in Amsterdam, for example, is not in the same market as a house in Rotterdam. Even though there are national-level forces that affect house prices and house sales in both cities, an individual wishing to purchase or rent a house will focus his or her search on a small area and rarely consider looking for a house “somewhere in the Netherlands.”

Housing economists frequently focus their analyses on a metropolitan housing market, and there are substantial differences between metropolitan markets. Part of this difference is explainable by the size of the metropolitan area, areas with larger populations often being more expensive. In 2012, for example, all but one of the five most expensive housing markets in the United States were part of the New York, Los Angeles, or San Francisco Bay metropolitan areas. Honolulu, the most expensive market in the United States, is a special case and provides clear evidence that population is not the only factor in determining house prices (NAR 2014). Detroit, perhaps another special case, had median house prices of about 9% of Honolulu. The general conditions of the metropolitan economy play a role in determining house prices, but so do the level of amenities present and factors that can physically limit the size of the built-up area. Finally, political factors can have a substantial effect on house prices.

There are also substantial variations within metropolitan areas that geographers have termed “housing submarkets.” The idea that metropolitan areas had submarkets that were interrelated but that functioned with some independence has a long history. This principle was implicit in some early urban models, such as the sector model, but it was formalized by geographers in the 1970s doing work on urban mobility (e.g., Adams 1969).

There is some disagreement as to why these submarkets existed and what they looked like. At the most basic level, geographers have pointed to the sheer size of some metropolitan areas and the tendency for the agglomeration of different kinds of economic activities in different parts of the city. Others have suggested that the patterns of industrial activities identified by Hoyt in many American cities guided daily activity patterns in such a way that strong housing sectoral patterns developed. Even a basic understanding of housing patterns in some US cities, particularly in the Midwest, would suggest that such patterns exist; however, there are also clearly divisions between submarkets created by race, ethnicity, and class. In the United States the early studies on filtering and vacancy chains identified clear sectoral patterns, but also strong inner city/suburban divisions driven by race. David Harvey and Lata Chatterjee’s work in Baltimore identified strong divisions of this sort. Harvey and Chatterjee (1974) focused on the workings of the capitalist housing market but, empirically, the existence of class and race divisions had been implicitly understood for years. The clearly inscribed borders of the Black Belt on the South Side of Chicago are perhaps the best example, and were identified in the early 1900s at the very beginning of the Great Migration.

Although the discussion of submarkets is set in the US context, they are present in every city and are particularly easy to see in large
HOUSING

cities. They are often long-standing and slow to change, although change in the characteristic of the markets does occur (the East End/West End divide in London in an excellent example). Although limits to human mobility and the sheer size of many cities create submarkets, many, perhaps most, submarkets are also features of race or class divisions present in the society.

Housing and economic rent

Different land uses have different location or site and situation preferences. They also produce different economic rents, with commercial land uses (office space and retail) producing the highest rents. The ability to produce high rents enables commercial developments to pay a higher price for land. Residential land uses are generally unable to pay as much for land as commercial uses. Of course, high-rise or high-income apartment buildings represent a high intensity of land use and enable very high rents to be paid for land; however, commercial uses at the same intensity will usually pay higher rents.

William Alonso suggested a bid rent function to determine land use, where the land most desirable for particular uses was put to the use that could pay (or bid) the most for it. Thus, areas of high accessibility, whether in the central business district or at a suburban crossroads, will be dominated by commercial uses. Residential uses will typically be found in areas surrounding the commercial uses on somewhat less accessible and cheaper land. This general pattern holds at several scales. Main thoroughfares in suburban areas are often lined with commercial land uses with high-intensity residential uses just beyond the commercial areas, and then lower-density uses beyond the higher-density ones.

The bid rent function becomes more complicated when different land uses are seen to have different attributes beyond accessibility, or when accessibility is more complex. Retail activities, for example, are often highly sensitive to accessibility at various scales from availability to large population centers to easy access to building entrances. Residential land uses, however, typically have lower accessibility requirements. Various physical and cultural amenities play a more significant role than access in the geography of housing. Rolling landscapes, water fronts, and public open space all attract higher-income residential development while heavy transit corridors, industrial development, and areas of high pollution discourage it. In short, the bid rent function suggests that the wealthy will be able to outbid the poor for residential land and typically will choose accessible locations with high amenities and a lack of disamenities.

While economic factors play a significant role in the geography of housing, social and political factors are just as important, especially in political systems with a high degree of local control and variation in the provision of public services, particularly public education. In the United States, sociopolitical factors are so important that they overwhelm or substantially alter the economic influences. In most countries, the central city is home to the wealthiest residents because of a high degree of accessibility to employment, and substantial amenities in shopping and entertainment. In many US cities, this situation is reversed. While some US cities do have significant numbers of wealthy residents in the core, in many others even the middle class has substantially abandoned the city. The pattern is most evident in the United States, but occurs to an extent in other countries.

In the United States, local governments are largely reliant on their own tax bases, typically property tax bases, to fund city services and, most crucially, public education. In most suburbs, these services can be supported at a substantially higher level than in the central
city through higher property taxes, which have the added effect of raising the actual cost of homeownership to a level that is unaffordable to potential moderate-income homebuyers. Higher taxes also raise contract rents to levels high enough to be a barrier to renting – if there are even rental properties available.

In addition, some local governments have established blatantly exclusionary policies through land-use planning. Requirements for lots of a particular size or houses of a particular size effectively prohibit moderate-income people from living in the area. This type of exclusionary zoning has been regulated for nearly 40 years, beginning with a series of legal cases involving the Philadelphia suburb of Mount Laurel, New Jersey. The Mount Laurel cases created a generally accepted “fair-share” doctrine, which noted that any municipality allowing for growth must provide housing for moderate- and low-income residents. This doctrine, however, is easily circumvented, in part because municipalities can actually buy out of it, by subsidizing housing in other municipalities. More recently, the pre-2007 housing bubble significantly inflated prices, pushing upward the price of “moderate-income” housing, which has thereby become unaffordable for even middle-class buyers.

While income and wealth are important variables in creating the geography of housing, in the United States race is the underlying factor in understanding much of the geography of housing. Through a variety of means, African Americans have been excluded from white areas. During a massive suburbanization in the postwar era, race, either openly or in latent form, was a driving factor, and in some areas it was the deciding factor in the decision by white families to move to the suburbs. The hypersegregation of many American cities was maintained through the housing market and a variety of other means, sometimes aggressive and violent, and was perhaps the most striking feature of the geography of housing in American cities. It remains so despite 40 years of legislation and policy prescriptions. Most recently, racial and ethnic segregation in US cities has been maintained through mortgage lending practices that will be discussed further in the “Housing Policy” and “Housing Finance” sections.

Much of the research on race and segregation has focused on the United States, where there are long-standing issues of racial discrimination and racial and ethnic segregation. Other traditional immigration-based countries (e.g., Canada, Australia, New Zealand, South Africa, and some South American countries) share some of these issues, particularly as they relate to indigenous residents, and Canada has some issues similar to the United States given its recent immigration history. With the exception of South Africa, discrimination and segregation are not as intensive and extensive in cities in these countries as they are in US cities. Many of these countries are large federal states, and the fragmentation of local government does make segregation easier to create and maintain, but it is certainly present in countries with centralized governments as well.

The United Kingdom was among the first European countries to experience high levels of immigrant segregation. Large numbers of South Asian and Afro-Caribbean immigrants began to arrive in the 1950s, and by the early 1980s a significant literature on segregation in the immigrant experience in the United Kingdom had begun to develop (see Peach 1996 for an overview). While there were clearly some similarities to the US experience, there were also significant differences – in particular, the experience of Muslim and non-Muslim South Asian immigrants – with a significantly higher and enduring level of segregation among Muslims, pointed to a religious dimension that was not
present in the United States. Despite some very high levels of segregation among ethnic and racial groups in UK cities, segregation is still not at the level of most US cities.

Other European countries have increasing issues of segregation and discrimination, perhaps most evident in the uprisings in Paris and other French cities in the fall of 2005. The geography of European segregation is profoundly different from that in the United States, and even the United Kingdom, with most immigrants segregated in housing projects outside the center of the city and in the suburbs.

Housing and the economy

The connections between housing and the economy are important at a much more theoretical level, and are developed most clearly in a Marxist analysis of housing and uneven development. As a significant portion of the built environment, housing plays a key role in the circuits of capital. Traditionally, the primary circuit, made up of investments in production of all types, is the most important area for capital investment. But, in times of crisis, defined as periods when a glut of capital occurs and rates of profit fall, the secondary circuit becomes an outlet for the excess capital and a target for increased capital investment.

The secondary circuit of capital is composed of investment in the built environment. While capital always flows into this circuit, when it receives comparably greater investment capital flows to areas that have the highest return on investment. These are often areas of traditionally lower capital investment with lower levels of development (Harvey 1978; 1985). In US cities, the suburban fringes or declining areas of the inner city are often the areas of investment. Not all of the investment is in housing (some is in commercial development), but much is.

This theory has been used to explain both suburbanization (Walker 1981) and gentrification (Smith 1982), and at its most essential level it suggests that such phenomena are the result of the need for capital to find a higher rate of return. Cultural factors, consumer preference, transportation innovations, all of which are often seen as important drivers of suburbanization, are thus moved to the role of enabling factors, but they are not the root cause. This view of housing expands on its economic importance as an exchangeable good, and makes it a key facet of the maintenance of global capitalism.

Housing policy

Because of the basic importance of housing and its important role in national economies, it is the focus of significant policy. Most housing policy is created at the national level, but some cities have policies that play a significant role in local housing provision. Thus, policy has profound impacts on the provision of housing and on how people acquire and use it.

Public and private housing provision

The most basic division in housing policy relates to homeownership and the provision of housing. At the two extremes are countries where almost all housing is owned by the government and all new construction is contracted by the government and often built by government laborers, and countries where almost all housing is provided through the private market, construction is carried out privately, and all housing is privately owned, although much of it may then be rented.

Public housing has been a hallmark of socialist countries. As late as the early 1990s, the vast majority of the world’s population lived in public
housing, or in housing highly regulated by the government. In China and the USSR all housing was public. In urban areas in India housing was highly regulated. Significant percentages of the populations of many European countries lived in public housing. One of the few areas of the world with significant amounts of privately owned housing was North America. Historically, in the United States, homeownership was not the norm in cities, but all housing was owned by private firms or individuals and much of it rented to residents.

It is difficult to generalize about the characteristics of these two systems, but typically public housing provision has been faulted for chronic housing shortages. Shortages are often attributed to poor central planning on the supply side, but they also have some roots in the demand for housing when it carries no specific cost. Private housing systems often create housing affordability issues, particularly in high-growth areas, and also have the potential to create high private debt levels for those attempting to buy housing. Some economists have argued that high housing costs can drain investment from other key growth areas of the economy and, certainly in poorer countries, limit the capital available to young people who might want to develop their own businesses.

Traditionally many capitalist countries have adopted a mix of housing provision, attempting to meet basic housing needs for lower- and moderate-income residents by providing the option of public housing, while higher-income people chose private housing. The mix of public and private housing has varied greatly. For example, since 1970 the homeownership rate in the United States has been reasonably consistent at around 63% to 65%. In the United Kingdom the rate increased from 55% to 67% in 2012 (currently higher than in the United States), and in Sweden the rate has gone from around 35% to over 70% during the same time period. While not all renters lived in publicly owned housing in the United Kingdom and Sweden, many did. In the United States in 1970 about 1.7% of all households lived in public housing. By 2010 fewer than 1% of all households lived in public housing, and about 4% received housing assistance (discounting the mortgage interest deduction).

In some countries the balance of public and private housing has changed dramatically. Of course, the fall of the USSR and the conversion of China to a market-based economy had the greatest effect on the global balance of public and private housing, but shifts were also evident in a number of other countries.

The neoliberalization of housing policy

Neoliberalism, the recent worldwide political movement that reduces government size and prioritizes market strength over government responsibility, is often thought of in two phases, rollback and rollout. These two phases affected housing in very different ways. Rollback neoliberalism refers to the movement to privatize the provision of goods and services that were traditionally publicly provided, and housing is clearly one of these goods. Rollout neoliberalism refers to government policies that, in one form or another, augment the free market in such a way as to provide significant advantages to particular private concerns. Although rollout neoliberalism has had some direct effect on housing policy, its most significant effect has been through housing finance (which will be discussed in the next section – the focus here is on neoliberalism and the privatization of housing).

The shift toward private housing can create some advantages. In good market conditions, it can aid in individual wealth creation through the appreciation of house prices, but just as often...
it results in substantial housing affordability and personal debt issues. It also exposes individuals, families, communities, and ultimately cities directly to the vagaries of the global capital that controls housing finance.

The United States is often thought as one of the central points of origin for neoliberal policy, but in larger housing policy, rollback neoliberalism wrought only minor changes in the United States. To a great extent, this is due to the lack of a coherent national housing policy, but primarily it is a function of the already dominant private provision of housing in the United States. Nonetheless, the very little housing in the United States that was government-owned was dramatically changed—principally through the HOPE VI program. HOPE VI worked to push local public housing agencies to become self-supporting, largely by encouraging them to develop market-rate housing in projects that had advantageous locations.

Neoliberal policies contributed to the substantial decrease in publicly provided housing in European countries (particularly the United Kingdom and Sweden), across sub-Saharan African (Jones 2012), and in Latin American countries, with Chile being one the starkest examples anywhere (Posner 2012). It is difficult to categorize Chinese housing policy as neoliberal, with its limitations on permanent residence (hukou) and the still substantial amount of corporate worker housing (danwei), but the dramatic shift toward market-based housing solutions in most Chinese cities has had a similar effect to neoliberal policies in other nations.

Subnational housing policy

It is common for special housing policies and programs to be developed for particular regions of countries, most commonly for large cities. Some of the funding for such programs comes from the national level, but there are significant housing affordability programs that are essentially completely local. Housing affordability is a major issue in global cities and many of them, particularly in Europe, have policies that attempt to provide housing at affordable rates. Most commonly, many cities control development rights and require the creation of a certain number of affordable units when they permit large residential developments. While such schemes can be helpful, there are a number of issues. In some cases, developers are allowed to buy out the requirements for affordable units—presumably with these funds going to construct affordable units elsewhere. In addition, given the extraordinary cost of housing in global cities, the definition of affordable is often quite malleable. In the United States, for example, the affordability definition is often set (by legal principle) at a percentage of the area median income (AMI). In many cities, the rate is set at 30% of the AMI, but in places with a particularly strong market, the rate is 60% of the AMI and there is a chronic shortage of housing even at this rate.

Rent control is a somewhat less common but perhaps more effective way of providing affordable housing. New York City provides an interesting example. Although it was initially a World War II era federal program, the program was taken over and continued by New York State in 1950. The program includes a complex mix of regulations that attempt to keep rents somewhat affordable. Actual rent control is the most stringent requirement and less than 2% of all units in the city are formally controlled; however, a series of rent stabilization programs limit rent increases in moderate-priced units (currently those with rent under $2500 per month). An increasing percentage of units in the city are not regulated at all. In 2011 the amount had risen to about 40%. Some municipalities have downpayment assistance programs...
for moderate-income buyers that are funded through various local fees and housing taxes, but even the most aggressive of these programs is limited in its effect in strong housing markets.

Many housing affordability programs are at the subnational level because, particularly in large countries, there are often many areas where housing is quite affordable. As of yet, there has been little attempt to restrict these programs at the national level akin to the efforts to weaken local predatory lending laws by passing superseding legislation at the state or national level. So, while local affordable housing efforts have not been helped much by national-level legislation, national governments have not acted to restrict local efforts.

**Housing finance**

Because few people can afford to buy a house outright, complicated financial systems have been developed to allow people to borrow money to purchase houses. These systems vary greatly from one country to the next, even in advanced capitalist countries with similar urban systems and homeownership cultures (e.g., the United States and Canada); however, government involvement in these systems is a constant. Almost all systems have or have had a fair degree of government regulation, and many governments have taken a proactive role to advocate for different types of finance instruments or to make financing more available to a wider spectrum of citizens. In the most extreme cases, government agencies often guarantee loans or even make mortgage loans themselves. The US housing finance system was one of the first systems to make homeownership available to a broad swath of the public. Many of its qualities, both positive and negative, have been emulated in other countries. And its failings helped trigger one of the largest global recessions in recent history, starting in 2007.

The birth of modern US housing finance dates back to the Great Depression and the series of acts passed by Congress in the 1930s. The creation of the Federal Housing Administration (FHA) in 1934 was a key moment. While the FHA involved itself in the provision of housing, its most important activities involved creating the means for Americans to be able to buy homes, and in a real sense it was a housing finance agency. The FHA began to guarantee long-term, fixed interest loans from private lenders, which allowed those without significant collateral to gain access to mortgages at reasonable rates with affordable payments. These types of mortgages had been in existence in the US housing market for some time and were becoming more common, but, in the midst of the economic downturn, the willingness of the federal government to guarantee loans spurred lending by many financial institutions and the long-term fixed interest mortgage loan became the backbone of American housing policy, and the key mechanism to provide US workers of modest means with the ability to own homes.

The creation of the Federal National Mortgage Association (FNMA) in 1938 also accelerated this trend. The FNMA created a secondary market to purchase mortgage loans from private institutions which allowed those institutions to escape the long-term commitment of a 20- or 30-year mortgage and to make more capital available for funding more new loans. The creation of a secondary market was a fundamental component of the long-term, fixed interest loan. It meant that private institutions were able to rid themselves of the perceived risk of these loans by selling them on the secondary market. It could have also broadened the range of people to whom financial institutions would be willing to lend, and probably did encourage lending to
HOUSING

middle- and working-class whites. In the Great Depression, when money was tight, it also increased the capital available for lending. In the 1930s, the FNMA only purchased FHA products. Its role gradually increased after World War II. FNMA, of course, is now officially known by an expansion of its acronym, as Fannie Mae.

In general, the federal programs in the United States were successful in expanding access to home loans for white families. Both institutional and individual racism worked to deny African Americans and other minorities access to the same mortgage finance. While strides were made throughout the 1980s and early 1990s, continued racism, coupled with several decades of deregulation of the housing finance industry, led to the targeting of poor and minority borrowers for high-interest loans. Through a complex series of financialization schemes, these risky loans eventually triggered the Great Recession in 2007.

Housing, development, human rights, and the environment

Housing is the space in which a number of significant human activities occur. It is also a tremendously important and emotionally charged subject in many societies. The exchange-value of housing links it to global economic systems. There are few issues that housing does not touch on; the following brief discussion highlights the most significant ones.

Housing is often a key component of the economies of both developing and highly developed countries. Decent, safe housing is a significant concern almost everywhere and is often used as one measure to gauge a country’s level of development. Housing is also a significant avenue for wealth creation. In developed countries, long-term mortgage lending is thought to allow people to purchase housing, pay off the mortgage, and own a house that has appreciated and will continue to appreciate in value. Take the example of a home purchased in 1965. If it was purchased at an average interest rate on a mainstream mortgage and located in a stable neighborhood, even accounting for the cost of basic repairs and upkeep and the real cost of property taxes (after accounting for the value of the tax deduction), the house could have been sold at a price in 1990 that would have likely covered most of the associated costs. Thus, a person could have lived in the home for 35 years and walked away with a cash payment that covered all the costs of living in the home during that time. This scenario typically would only be the case for white homebuyers in certain suburban areas, however; in other areas, homeowners may still walk away with a cash payment, but the costs of interest, upkeep, and so on would probably exceed the sale price. In purely economic terms, it would have been better for these homeowners to have used some of the money they paid for the home on rent, and saved or invested the rest. Despite these issues, however, homeownership is still advocated by many in the housing industry and by many government policies as an effective means of wealth creation.

In countries with highly unstable economic situations, housing can also be a safe investment for those with few resources. An additional room or a stronger roof always has use-value and provides a good option to saving when the currency is inflating. In addition, quite a lot of work is done by the homeowner and many materials involved can be purchased in small quantities and stored. Many houses in the developing world are in a constant state of construction, which is an effective use of spare capital, assuming the homeowners can stake a reliable legal claim to ownership.

The question of ownership is an important one in many developing countries. Significant portions of housing in cities in Latin America,
Africa, and South Asia have been and continue to be built on land that residents had no legal right to occupy. The living conditions in these squatter settlements are often deplorable, but policies toward the settlements are highly variable. Often, squatters are viewed as a nuisance and significant private and public efforts are made to move them and demolish the houses. People must then move and repeat the process of constructing some sort of shelter. In situations where eviction and demolition are common, individuals will not expend much effort on improving their houses, and in many situations they have little time and no resources to make improvements. There are examples of somewhat more enlightened policies toward squatter settlements. One of the more effective government policies has been to recognize their existence; provide water, sewer, and electrical services; and grant residents the legal right to the land. In such instances, which are rare, homeownership has been a path to modest wealth creation.

Housing also has significant legal and civil rights components. Claiming a right to privacy or a right to control the space within the home can be a key facet of developing broader human and civil rights. Conversely, in highly patriarchal societies, the home can also be the site of abuse and violence toward family members that is tacitly or even explicitly sanctioned by the state.

In countries with well-established human rights, housing is still a point of contention and sometimes a key indicator of civil rights. Chester Hartman (1984) proffered the idea of the “right to stay put” in the context of the urban renewal and highway building programs of the 1960s and 1970s. The significant slum clearance for highway construction throughout the postwar era in the United States destroyed neighborhoods, particularly poor and predominately black areas. Hartman suggested that the right to stay put should act as a significant counterbalance to market forces and the state’s power of eminent domain. Eminent domain, stemming from English common law, provides the state with the right to take land (and housing) if it needs to be acquired for a community’s health, safety, and welfare. Large public works projects may have come under these principles, but later eminent domain was used (often by quasi-public economic development entities) in contexts where the need to increase property tax revenues was couched in terms of a health, safety and public welfare issue.

The right to stay put is not an issue only in English common law countries, nor only in democracies; it is also an ongoing concern in rapidly developing countries such as China, where extraordinary levels of development in urban areas have created a crisis of displacement. For example, land rights protests are reasonably common and have given rise to the term “nail houses” to describe the houses and owners who would not sell out to development concerns. Often the house stays while the new development proceeds around it. In some ways this is less an issue in both China and India because of the huge gulf in wealth between the average urban dweller and the developers. Developers can often easily afford to pay the residents to be displaced a sum of money that seems extraordinarily high to the residents, but the issue still arises. The Chinese central government has recognized, in a manner, the right to stay put, while pressuring local governments to increase revenues through development. When the displaced appeal to the central government, it simply notes that the issue of displacement is a local one.

Although “sustainability” is a nebulous term, it is clear that there has been a long-term concern on the part of geographers about human impacts on the local and global environment and
HOUSING

the effects these may have on longer-term health of life on the planet. In this sense, housing is a key issue for anyone concerned with environmental issues from habitat protection to global climate change. There are three basic concerns with housing: where it is located, how large it is, and how it is made.

The location of housing is an important component of its environmental impact. Is the housing located near other locations to which household members need to travel? Is it located near effective mass transit with lower environmental impacts? Commuting long distances to work or other daily activities in a private automobile represents one of the most significant environmental impacts in many highly developed economies. It burns significant amounts of fossil fuel, contributing to global carbon output and is a significant contributor to local air pollution problems.

House sizes in much of the world have increased dramatically over the past century, even as family sizes have declined in many places. In 2009 Australia had the largest average new house size at 2303 ft² (214 m²) followed by the United States, at 2164 ft² (201 m²), which was down slightly from 2006 when new house sizes in the United States reached their peak. In the United Kingdom the average size of a new house was 818 ft² (76 m²), which was among the smallest in Europe. China’s new housing units averaged 616 ft² (60 m²).

House sizes matter for several reasons. Most new construction uses significant resources, few of which are renewable or environmentally benign. In the United States in 2005, half of all the wood cut domestically or imported went into housing construction. Even in countries where high-rise or higher-density houses are the norm, concrete, steel, aluminum, and glass all require significant natural resources to produce, and the larger the unit the greater the impact.

The size of a house also determines the amount of fuel used to heat or cool it, and single-family detached homes are the least efficient. Finally, there is also a correlation between larger houses and greater consumption of consumer goods, even when accounting for the influence of wealth and income.

Construction practices can also have an impact on the level of resource use for new housing units. There is a movement in some countries to use green methods for new construction by utilizing renewable resources and low carbon emission products. In the United States, the Green Building Certification Institute established and certifies green construction standards through the Leadership in Energy and Environmental Design (LEED) program. While these standards can also be applied to homes, they are most commonly seen in larger commercial projects. Green construction also emphasizes reducing other toxins that are common in the production of building materials. Highly insulated houses with efficient means of climate control are also important. Of course, renovation and reuse are typically much more environmentally sound and can substantially reduce the carbon footprint of older homes.

Directions for future research

Much research on housing and housing-related topics in a wide variety of disciplines precludes a comprehensive discussion of future research, but several areas do seem to be of importance and of interest to geographers. Urban policy in neoliberal market-based economies has increasingly been reduced to encouraging private homeownership. It has also served as a substitute for the lack of other government policies concerning broad issues of health, safety, and community welfare. The drive to homeownership has long
been a facet of US urban policy, and, while the emphasis has not been as strong in other industrialized countries, the rates of homeownership have still risen dramatically.

Homeownership certainly has benefits, but it also poses significant risks, as the most recent recession has shown. For many years, academics concerned with housing have questioned the benefits of ownership, particularly for moderate-income people, and the legitimacy of renting has been gaining more credibility. Renting, however, faces numerous institutional challenges, which are highlighted by Donald Krueckeberg in his aptly titled paper “The Grapes of Rent” (1999). Many attitudes about renting versus ownership are highly spatial in nature and include issues of place attachment and place-making as well as a range of other community development-based questions.

Geographers have also begun to investigate alternative forms of homeownership, including some that are intended to provide community members with a greater stake in the governance and management of communities. Community land trusts are one example of this kind of arrangement. They are more common in rural areas but are increasingly used in cities to mitigate some of the risks of homeownership while providing residents with a greater stake in their communities and an increased voice in neighborhood governance. They are also used to provide “perpetual affordability,” which guarantees that, even in the most overheated housing markets, affordable housing can be provided. While community land trusts have been used for over half a century, there are still significant questions about their effect on homeownership, governance, and affordability that become more significant as the instrument is more widely used.

Segregation has also received much attention from urban scholars. Yet this issue is still insufficiently understood as the nature of segregation evolves. The United States has been primarily concerned with racial and ethnic segregation, while the larger issue of sociospatial segregation has been the focus of less racially diverse countries such as China. In multicultural countries, however, there has been little change in levels of racial segregation since the 1950s. Even in postapartheid South Africa there has been little change in levels of residential segregation (Christopher 2005), despite often strong rhetoric from the national government (Turok 2001). This is not to say that changes in policies and laws have had no effect in cities across the globe; they have, and racial segregation levels have dropped in many places. Many US cities, however, are still hypersegregated, which suggests that the topic needs further attention. Politically, there have been calls from both the far left and the far right to allow segregation and to work toward (at least from the left’s perspective) a separate but equal society. There is sophisticated research ongoing about the effect of segregation on children – their worldview, their access to a range of opportunities – and geographers could play a larger role in bringing these discussions to the fore.

Inequality is currently a buzzword in discussions of socioeconomic conditions across the developed world, and much empirical evidence suggests increasing rates of inequality in North America, Europe, and East Asia. Inequality is deeply related to housing and homeownership policies, and is perhaps the strongest challenge to date to the use of homeownership as a wealth creation vehicle, particularly in a neoliberal society. Inequality is also inherently geographical. In the larger social science community there has, as yet, been little discussion of the spatial effects of inequality, which are related strongly to a range of housing issues. More importantly there is no discussion of the spatial origins of inequality by policymakers largely working out
of a tradition of neoclassical economics. These two broad topics have been the focus of a large body of research by geographers. The increasing importance of socioeconomic inequality provides geographers with an opportunity to highlight what is already known on these issues and to work to increase our understanding. Housing-related research is one important avenue for achieving these goals.

SEE ALSO: Cities and development; Finance and development; Gentrification; Geographies of exclusion; Home; Homelessness; Neighborhood; Race and racism; Slum; Suburbanization; Urban geography; Urban redevelopment; Urban uneven development

References


Further reading


Broadly defined, human capital refers to useful skills, knowledge, competencies, and other attributes that enhance an individual’s productivity in the labor market and facilitate other forms of personal and social wellbeing. This capital arises from investments in areas such as formal and informal education, health, on-the-job training, and migration. Notions of what constitutes human capital have evolved over time. While earlier definitions more narrowly focused on the skills and knowledge that increase the economic productivity of individuals as workers, more recent understandings also include the sociopolitical and personal benefits that may arise from higher levels of personal development. Other elements, such as the ability of individuals to access labor markets or benefit from enabling institutional environments, have also been incorporated into more recent definitions of human capital. Measures of human capital have also become more sophisticated in capturing the complexity of personal development and valuable attributes. The characteristics now understood to comprise human capital are acknowledged to have both instrumental value in bringing about certain outcomes (e.g., economic growth, higher wages) as well as intrinsic value to individuals (e.g., enhanced self-respect, greater job satisfaction). As such, higher levels of human capital can be viewed as both a means to and a goal of human development.

Early theorizations of the economic production process considered three basic factors in the creation of economically valuable goods and services: land, labor, and capital (i.e., fixed non-financial assets used as components in the creation of goods or services that do not get significantly depleted in the production process). However, the view that not all laborers contribute equally to the production process and that workers can also have fixed assets has a long history. The notion that workers’ abilities (i.e., what is now referred to as human capital) shape economic productivity is grounded in the early work of classical economics, particularly the writings of Adam Smith and John Stuart Mill. For example, Smith (1981) wrote that workers acquire talents through education and experience that brings them and their society economic benefits via higher profits both in terms of personal wages and enhanced efficiency in the production process.

In the 1950s and 1960s, Jacob Mincer, T.C. Schultz, and Gary Becker pioneered the theoretical and empirical articulation of human capital as an important component in production within neoclassical economics, emphasizing that investments in people lead to individual and national economic growth. In particular, differences in human capital investments were theorized as the result of individual free choice in regards to the time allocated to education and training (Mincer 1958). Education as a human capital investment is predicted to incur near-term costs but the resulting knowledge base of individuals is projected to bring substantial
returns to these investments in the future. In the US context, economic returns to investments in human capital were found to be higher than those of physical capital (Schultz 1961). Moreover, human capital was shown not only to increase the productivity of both time and goods used in the production process but also to produce additional human capital (Becker 1993). Thus, human capital theory suggests that individuals and their respective societies derive economic benefits from investments in people, particularly with regards to formal education and on-the-job training. The long-term benefits of human capital investment are predicted to accrue to individuals and societies rather than to particular firms because human capital is not owned by the firm which employs it, but rather resides in individual workers who can change their place of employment.

The early theorists acknowledged non-economic benefits of human capital, including higher rates of voter participation, increased awareness of family planning options, and greater appreciation of classical music and literature (see Becker 1993, 21). However, their empirical studies and theoretical formulations focused on the relationship between human capital and monetary outcomes such as income inequality, economic growth, unemployment, and competitiveness in international trade. The traditional proxy for human capital was, and largely remains, measures of educational attainment, although health, nutrition, and ability to migrate have long been identified as important in improving worker productivity (Schultz 1961). Early concepts of human capital were generally one-dimensional, viewing human labor as a commodity in the production process and human capital as assets which increase worker quality and lead to greater economic output.

**Key critiques of neoclassical theories of human capital**

The neoclassical theory of human capital has been critiqued on a number of grounds. Within Marxian perspectives, critiques have centered on the failure of human capital theory to acknowledge the relevance of power, class, and class conflict in explaining labor market phenomena. Marxists and other heterodox economists often explicitly reject the view that earnings reflect the market rate of return on human capital (Bowles and Gintis 1974; Folbre 2012). Rather they present the theory as a problematic extension of neoclassical economic thought which promotes the commodification of people according to their knowledge base while ignoring structural imbalances of power that distort returns to education. Marxists reject the theory’s almost exclusive focus on attributes of human capital that arise from individual choice and investments while neglecting to account for unchosen attributes, such as race or gender, which lower returns to schooling for particular individuals. Some Marxists also maintain that the theory lacks explanatory power because it only considers the beneficial aspects of schooling for individuals and does not account for the role of educational institutions in legitimizing structural inequities of power (and wealth) that they argue are inherent in and produced by capitalist labor markets. However, some Marxist economists have also praised human capital theory for rejecting simplistic notions of a homogenous labor force and calling attention to the importance of basic social institutions – schools and the family – in economic analysis.

Within the development literatures, the human capital approach to promoting economic development in less advanced economies has also received substantial criticism. For example, the World Bank’s lending programs have been
heavily influenced by human capital theory since the early 1970s, and financing education remains a key component in the majority of its poverty alleviation initiatives. However, critics have argued that the World Bank’s application of the theory has overemphasized financing educational infrastructure and largely neglects the role of teacher training and the necessary incentives (e.g., salaries) to improve educational quality. Easterly (2002) notes that many teachers in rural areas of the developing world often fail to report to work or teach with very little oversight, often resulting in low quality education – and thus limited human capital production – for the poor. Other scholars and policymakers have noted that, in theory, human capital’s impact on wage outcomes and economic growth depend on assumptions of perfectly competitive labor markets and high levels of employment. However, the acquisition of human capital is unlikely to substantially improve economic wellbeing in areas where formal jobs are very scarce and excess labor capacity leads to downward pressure on wages (Mehrotra 2005).

Placing human capital within broader perspectives of human development

Contemporary concepts of human development view progress in improving the human condition not just in terms of increasing personal income or national economic growth, but as the creation of an enabling environment for people to enjoy long, healthy, and worthwhile lives. For example, Amartya Sen (1999) argues that economic growth is useful only in so far as it increases the opportunities and abilities of people to live lives they have reason to value. Development can thus be evaluated according to the opportunities people have to achieve personal development rather than by focusing on economic indicators. Within this view, human capital accumulation involves increasing intrinsically valuable elements of human wellbeing in addition to those having instrumental value in bringing about better economic outcomes. Thus Sen expands on previous notions of human capital’s value for people and societies.

Sen (1997) notes that the neoclassical view of human capital can easily fit into more inclusive conceptualizations of human development. For example, human capital, such as education and good health, can make workers more efficient in production processes, thus increasing the value of economic output and individual wages, while at the same time providing nonmonetary benefits to people that improve other dimensions of human wellbeing. Several influential studies of poverty find that many traditional components of human capital including literacy, numeracy, and freedom from disease are widely perceived by the poor as important in alleviating nonmonetary types of deprivation, such as social exclusion and the ability to appear in public without shame, even if they do not enhance income (cf., Narayan et al. 2000). Many international organizations and national governments now promote education, health, and mobility as important developmental outcomes independent of their potential contributions to economic growth.

Measuring human capital

Empirical studies of human capital’s influence on economic development commonly employ educational measures as proxies. However, the indicators used often differ in studies of developed countries (which tend to focus on human capital’s role in regional economic competitiveness and development) and less developed countries (which tend to focus on the ways in which human capital impacts poverty
HUMAN CAPITAL

alleviation). In developed countries, studies often employ measures of growth in university-level qualifications, graduation and enrolment rates, time invested in education, and education expenditures. Largely due to data constraints, studies of less developed countries tend to use more simplistic measures including school enrolment figures, years of schooling, literacy rates, and ratio of skilled adults to total adults. However, few of these measures account for possible barriers faced by individuals or groups in making human capital investments in education.

Some studies also include measures of human capital investments in health and nutrition. Common health indicators include health status and life expectancy (often broken down by socioeconomic groups), infant mortality rates, quality and accessibility of the health systems, and public infrastructure affecting health (e.g., sources of potable water, transportation networks, and sanitation systems). More recent work, particularly within advanced economies, has attempted to measure the complex cognitive and cultural characteristics of human capital. In addition, Richard Florida (2002) has argued for measures based on the occupational composition of regions, particularly employment in high skill sectors including science, finance, arts, and entertainment.

Perhaps the best illustration of the shift in viewing human capital accumulation as a goal of development rather than simply a means to expand production and economic growth is the Human Development Index (HDI), which was introduced by the United Nations Development Programme (UNDP) in 1990. Sen’s views on the intrinsic value of certain types of personal development, notably education and health, profoundly influenced the development of the HDI which replaced purely monetary measures of development, such as national gross domestic product (GDP), with a composite statistic comprised of indicators measuring educational attainment (i.e., mean years of schooling and expected years of schooling), life expectancy at birth, and gross national income at purchasing power parity per capita. Thus, traditional indicators of human capital take on enhanced significance as development outcomes when evaluating progress according to the HDI. In 2010, an inequality-adjusted HDI (IHDI) was introduced which accounts for the national-level distribution of health, education, and income measures, and thus better captures the potential costs to human development arising from structural inequities within societies. The HDI is now one of the most widely used measures of international quality of life comparisons and the IHDI is growing in popularity.

The growing realization that other indicators are necessary to more effectively capture the multidimensionality of human capital and its relationship to human wellbeing led to the recent development of the Human Capital Index (HCI), first published in 2013 (WEF 2013). The HCI includes multiple measures of human capital at the national level, as well as the social, physical, and economic characteristics of societies that shape the ability of individuals to benefit from personal development. Educational indicators incorporated into the HCI include both quantitative and qualitative aspects of learning. Health and wellness measures include aspects of physical wellbeing, cognitive functioning, and mental health. Workforce and employment indicators include measures of experience, talents, knowledge, and training for working age adults. The HCI also measures the extent to which human capital can be employed effectively (i.e., the degree to which an enabling environment exists) by incorporating indicators of the quality of legal frameworks, institutions, and infrastructure as well as the prospects of social mobility within a particular county. The index also takes a long-term approach to measuring human capital
by including measures on the differential access to education according to gender and the quality of early childhood development, both of which shape the potential human capital acquisition of individuals in the future.

**Current and emerging research themes**

Within economic geography, mainstream economics, and regional science, much of the theoretical and empirical work on human capital continues to focus on the effects of individual skill sets and knowledge in enhancing the economic competitiveness of firms, cities, regions, and countries. A general consensus exists on the importance of human capital to regional development, particularly with regards to employment growth, wages and income, entrepreneurship, and innovation. Migration, cultural diversity, and the presence of colleges and universities have been found to be a significant factor in the geographic distribution of human capital at the subnational level. For example, studies find that cities remain highly competitive even in theoretically mobile industries such as advanced business services, in part because urban areas attract highly skilled workers. Research also indicates that the level of human capital within localities positively effects the labor market participation of both men and women. However, a common theme in the literature is the need for better proxies that more effectively capture the full dimension of human capital including social skills and complex cognitive abilities such as analytical thinking.

The interest in human capital and its impacts on various elements of wellbeing has spread beyond the economic development literatures to a number of other academic disciplines and policymaking arenas. In particular, human capital has emerged as an increasingly important element in shaping differential vulnerability and adaptation to climate change and the processes driving violent conflict. Some recent research suggests greater human capital investments enhance climate change adaptation and improve risk management, especially among the most vulnerable to negative climate shocks. However, other studies have identified numerous channels through which climate change can actually diminish human capital formation, particularly among individuals living in poverty. For example, climate stresses and extreme weather can undermine health via contributing to food shortages and poor sanitary conditions, decreasing access to potable water, and destroying human shelters and other infrastructure. When they diminish income, climate-related disasters can also reduce the ability of families to send their children to school and intensify pressures on families to transfer children into labor markets. The relationship between human capital and violent conflict also appears to be multidirectional. Some studies find that higher levels of human capital enhance conflict resolution skills and participation in democratic processes. However, research also indicates that exposure to violent conflict negatively affects human capital accumulation by destabilizing educational and health care institutions, decreasing household assets, and contributing to involuntary migration. Taken together, the research to date suggests that exposure to violence, particularly in early childhood, has long-term consequences for educational attainment and human welfare.

**Important areas for future research**

A large body of theoretical and empirical work has produced detailed knowledge of the ways in which human capital investments and a skilled workforce can enhance economic growth, influence regional competitiveness, and contribute to the geographic clustering of
economic activity in advanced economies and China. However, much less is known about mechanisms that can promote and enhance human capital development in countries and regions where people face high barriers to both building and effectively using their human capital. Additional research on the characteristics and conditions that allow poor communities and individuals to accumulate and leverage their human capital to mitigate violent conflict or adapt to and recover from climate change also remain important areas for further study.

Researchers have also articulated many direct channels through which human capital can contribute to improvements in wellbeing; yet most empirical work, particularly within economic geography, has continued to view human capital in the traditional neoclassical sense as assets of workers that improve their economic performance and their leverage in the workforce. Other avenues of impact between human capital and human development, such as its potential role in bringing about social change and expanding the capabilities that allow people to live freer and more worthwhile lives, receive very limited attention in the literature. The potential social, political, and cultural benefits of human capital accumulation in both developed and developing world contexts merit further investigation.

**SEE ALSO:** Factors of production; Labor market segmentation; Livelihoods; Population mobility and regional development; Social capital

**References**


**Further reading**

Human ecology

Gregory Knapp
*University of Texas at Austin, USA*

Human ecology encompasses a variety of methods and theories for understanding the relationships between people and their environments at all scales, times, and places. The term has been used by various disciplines, including geography, but is usually understood today to be interdisciplinary. The reference to “ecology” implies that humans and their environment have mutual relationships. Although the term “human ecology” is historically new, dating from the early twentieth century, the underlying subject area, the study of human–environment relationships, is much older.

Every human society has studied and reflected on its relationship with its environment. The oral traditions of “ethnoscience” typically include empirical observations about natural resources, hazards, and management strategies. Spiritual and religious traditions, including shamanism, encompass stories and beliefs about the relationship between people, plants, animals, landscapes, water, and weather. As literacy has expanded in human societies, oral traditions have been incorporated into official discourses of knowledge and power. Although ethnoscience may be maligned in the contexts of colonialism and modernization, local knowledges have proven important and useful for understanding resource practices and even for identifying resource management options for the future.

In European civilization, the ancient Greeks and Romans wrote about agronomy, the impact of the environment on human health, and the relative habitability of various parts of the world, but had a limited grasp of the impact of humans on the environment. In medieval times, St Francis’s teachings urged the valuing nature as part of God’s creation. Many other cultural traditions suggested a close connection between human consciousness, society, and nature, including Taoism and Buddhism in Asia. However, the systematic empirical study of the relationship between people and the environment came later, especially with European exploration of the world and the development of modern science.

Beginning with Marco Polo and accelerating after the voyages of Columbus, European chroniclers published detailed descriptions of new landscapes, ways of life, and resource management practices; colonial and commercial authorities sponsored local studies which provided even more data about different coping strategies and challenges to health and livelihoods in different environments. Encounters with foragers and pastoralists prompted questions about the merits, drawbacks, and historical explanations of agriculture and civilization. By the late eighteenth century, Humboldt was mapping altitudinal zonation in the Andes, and Malthus was writing about the population–resource ratio and warning about the tendency of societies to outgrow their resource base.

Europeans also began to become aware of modernization itself as a new transformation in human history, comparable in significance to earlier transformations such as the rise of agrarian civilization. Modernization involved a shift toward promoting efficiency of production, through division of labor and increased mechanisms of oversight and control within
bureaucracies and firms; the process was facilitated by entrepreneurs in the marketplace and by managers in organizations, and was promoted or impeded by governments and popular movements. Initially, writers addressed (i) the desirability of the goal of increased efficiency of production; (ii) the degree to which this goal could be achieved by modernization methods such as free market institutions; and (iii) the degree to which human nature itself required specific institutional controls (checks and balances, competition) or, alternatively, could be trusted to undertake dramatic revolutionary changes in social structures (Sowell 1987). The ideological struggle over the degree to which human nature is constrained has continued to influence human ecology, separating sociobiologists and neo-Malthusians from political ecologists, for example.

During the nineteenth century, Darwin took up Humboldt’s challenge to perform fieldwork in South America; together with the work of Wallace, his field observations led to the formulation of a theory of natural selection in diverse environments to explain the origin and extinction of species. The theory suggested that humans, having evolved from primate ancestry, were also the products of natural selection. Humans have continued to evolve, developing such traits as lactose tolerance; plants, animals, and disease organisms have coevolved with human innovations in agriculture and civilization. Human cultural habits could also be viewed as subject to environmental selective pressures, based on their relative efficiency in survival and reproduction. The Darwinian tradition also includes sociobiology and its assertions that modern human behavior may be influenced by genetic traits that evolved in a foraging lifestyle in a savanna setting.

Karl Marx is important in the development of human ecology because of his focus on the central role of social arrangements for resource management and production in human history. He argued that the social arrangements for exploiting natural resources also had critical impacts on other aspects of society, and explained modernization as the product of the logic of capital accumulation as expressed by a capitalist class. The environmental implications of this perspective are still being worked out by political ecologists, and the spatial implications for the organization of the world-systems by scholars in development studies.

Many geographers in the late nineteenth and early twentieth centuries focused on potential environmental impacts on human affairs, a tendency that has been referred to as “environmental determinism.” Some of these writers attempted to explain the dominance of European civilization in ecological terms. Others, such as American geographer Ellen Churchill Semple, combined aspects of determinism with more careful empirical studies of human–environment relationships. In opposition to environmental determinism, the French geographer Vidal de la Blache helped develop the concept of “possibilism,” the assertion that the environment presents humans with possibilities and constraints, but that the ultimate outcomes reflect human agency. The work of his students focused on the careful study of ways of life (genres de vie) in particular landscapes (pays) that were also affected by long-distance processes. His students helped found the Annales School of history, which was devoted to the careful empirical study of environmental, demographic, economic, and other processual factors in the history of places and regions. In the United States, Berkeley geographer Carl Ortwin Sauer played a similar role to the Annales School in encouraging detailed long-term analyses of the relationship between people and places. He deployed the concept of ecology at times in his writings but
did not embrace a systems framework; instead he focused on cultural practices as affected by inventions and diffusions. In 1955 Sauer, along with the biologist Marston Bates and the urbanist Lewis Mumford, co-chaired an important symposium on “Man’s Role in Changing the Face of the Earth.” This helped launch the modern concern for human impacts on the planet and further reduced the importance of the determinist perspective.

The term “ecology” dates from the late nineteenth century, and there were a number of discussions of the relevance of biological relationships to human society, but the term “human ecology” was first used by geographers in the early twentieth century. Members of the department of geography at the University of Chicago, J. Paul Goode and Harlan Barrows, were pioneers in defining the term, and eventually argued for its deployment as synonymous with geography itself (Gross 2004). Barrows’s 1922 presidential address to the Association of American Geographers was titled “Geography as Human Ecology” (Barrows 1923). The Chicago geography program was important for the training of Carl Sauer and other influential scholars, including Gilbert F. White, who pioneered the systematic study of adaptation to natural hazards.

Sociologists at Chicago also embraced human ecological frameworks for field research on human society in urban environments. Chicago sociologists Robert E. Park and Ernest W. Burgess studied the concentric zonation of activities in cities, using analogies from biology including the ecological concept of community. By the 1940s and 1950s, sociology had turned away from ecological explanations toward more purely social explanations of social facts. In the late 1970s sociologists William R. Catton and Riley E. Dunlap were arguing for a “new human ecology” or environmental sociology, continuing the tradition of Park and Burgess (Gross 2004).

In Germany, the geographer Carl Troll coined the term “landscape ecology” in 1939 to describe detailed interactions of climate, soil, and plants (including crop plants) in particular places. He helped found the journal *Erdkunde*, which has published many important papers in human ecology, and influenced Karl Butzer, who built on Troll’s focus to include long-term analyses of demography, agriculture, and landscape impacts (Butzer 1982; Butzer and Endfield 2012).

Anthropologists also deployed ecological approaches for the study of cultures. British anthropologist C. Daryll Forde related cultures to habitats in the 1930s. In the 1940s Julian Steward deployed an environmental framework in editing a multidisciplinary handbook on South American Indians, and helped create the subfield of “cultural ecology,” defined by him as a study of the processes by which a society adapts to its environment. He urged careful study of the cultural “core,” practices directly related to making a living in a particular setting. By the 1960s and 1970s, cultural ecology was a flourishing paradigm in anthropology and archaeology and was influencing geography, particularly as deployed by William Denevan and his students at Wisconsin (B.L. Turner, Bernard Nietschmann, and others). Although some studies stressed environmental limitations, more commonly scholars placed an emphasis on the role of adaptation. Chicago-trained anthropologist John W. Bennett urged the study of human–environment relations in terms of adaptive dynamics, with full attention to questions of society, identity, and long-term change. Emilio Moran promoted the study of adaptations to the conditions of such environments as arid zones, high altitudes, grasslands, and the Arctic (Moran 2006).
As in biological ecology, systems approaches have been deployed in human ecology to organize data, guide research agendas, and integrate research teams. The Man and Biosphere Program (MAB) was launched by the United Nations in 1971 and initially attempted to include humans in systems analyses of energy flows, including feedback processes. More recently, the study of coupled human–environment systems has been promoted under the rubric of land change science. Although the human factor is considered to be a subsystem, actual human behaviors may pose challenges for systems analysis owing to their dynamism, local specificity, and the role of culture, governance, and politics. That said, certain regularities do seem to govern relationships between population, transportation, settlement patterns, and land use, and models of these relationships continue to be developed and tested using quantitative data (Turner, Lambin, and Reenberg 2007).

Since the 1980s, “political ecologists” have studied the roles of colonialism, globalization, and neoliberalism as leading to the disempowerment of local people and the despoilment of natural resources; some political ecologists have also deployed activist and/or participatory agendas to achieve emancipatory and conservationist goals (Robbins 2012). In addition, there has been a growing focus on the analysis of discourse and interpretive communities, with the aim of critiquing dominant narratives of progress and dominant depreciation of the validity of ethnoscience. The integration of political ecology, land change science, and cultural ecology approaches continues to be a challenge.

Ongoing research in human ecology across time and space has focused on the major transformations in human–environment relationships since the emergence of the human species, and on the resources, risks, and adaptive strategies deployed in varied environments (Bates 2005). Transformations and episodes which have been studied include the human revolution and emergence of humans as a species; the foraging way of life; domestication; the emergence of permanent settlements and agriculture; the rise of cities and civilization; and modernization.

Work on human origins has involved the investigation of early dietary and behavioral patterns and the geographical characteristics of early human habitats. Chimpanzees have been considered the closest living relatives, and their behavior has been examined for clues to early human behaviors. Recently, bonobos have received a great deal of attention as possibly being even more closely related to humans, and as possibly exhibiting less violence and more gender equality. Early bipedalism and nakedness have been considered hominin adaptations to conditions of emerging grassland and savanna environments (the “savanna hypothesis”). Management of fire enhanced hunting possibilities, and allowed for the cooking of food for more efficient acquisition of nutrients, especially to support large brains. Campfires could have been early centers of social life and sharing. Subsequent migrations to Australia and the Americas were made possible in part by the early use of watercraft.

Foraging (hunting and gathering) has been considered the earliest human adaptation. Studies of modern foragers such as the !Kung by Richard Lee and others have demonstrated that foraging can be labor-efficient; it usually involves gender specialization (women gather, men hunt), food sharing, and opportunistic migration by small bands. Specialization and sharing would make humans more efficient at using resources than other primates, but also required heightened communication skills for planning and coordinating, and mechanisms for rewarding cooperation and punishing cheating. Larger brains, language, and culture – the “human
revolution” – thus may have been driven more by the social than the technological aspects of the foraging adaptation. Forager mobility and perceived resource limitations probably helped place a limit on fertility and birth rates. There is an ongoing debate over the level of violence and warfare associated with foraging societies. Some such as Douglas Fry argue that warfare was rare, while others such as Richard Wrangham and Samuel Bowles argue that warfare was common. The debate depends in part on the definition of warfare, and in part on the inclusion or exclusion of sedentary foragers from the databases.

Since foraging seems an attractive lifestyle, the emergence of agriculture is something of a puzzle. Certain kinds of domestication (including dogs from wolves) are consistent with migratory foraging, and many early plant domesticates were probably cultivated in conjunction with a dominant foraging lifestyle. Jared Diamond has suggested that only some wild plants and animals were suitable for domestication, and their geographic distribution helps account for the subsequent competitive advantages of certain cultures and geographic regions. A shift to agriculture may have been spurred by “Pleistocene overkill” of large prey animals, by climate change at the end of the last glaciation, and perhaps by religious, funerary, and other cultural practices promoting sedentary settlements.

Early forms of farming could have included shifting cultivation in forested environments, which in modern times has been shown to be labor-efficient and consistent with sustainable productivity. However, Denevan has pointed out that shifting cultivation is much less efficient without metal tools, and has argued that intensive agriculture may have emerged much earlier than is sometimes thought. Much recent research has focused on “dark Earths,” modified soils high in organic matter, archaeologically found in such locales as the Amazon basin. Irrigation was probably also labor-efficient under early conditions, especially where the water was rich in nutrients (floodwater farming).

Sedentary life with agriculture encouraged higher birth rates, which resulted in population growth and, in some cases, conflict. Napoleon Chagnon documented high levels of violence among agricultural societies in the Amazon associated with struggles over resources and women. Robert Carneiro hypothesized that cities and state-level social organizations first emerged in areas of circumscribed resources (theory of environmental circumscription), and that conflict and war were the key to forcing members of chiefdoms to join larger civilizational units. Such circumscribed resources would be characteristic of deserts (Egypt, Mesopotamia, Indus river valley, coastal Peru) and mountains (central Mexico, Andes). In some regions, high population densities might create the “social circumscription” that was conducive to state formation (Maya lowlands). Other scholars have argued that large-scale states rely as much on ideological and religious means to gain and maintain power.

Jared Diamond’s popular books, including Collapse (Diamond 2005), have redirected attention to the many historical cases of population decline and societal simplification. Easter Island, the classic Maya, Viking settlements in Greenland, and others provide examples for analysis. His interpretations continue to stimulate critique and debate (Butzer and Endfield 2012). Ongoing research on climate change, volcano activity, earthquakes, hurricanes, and El Niño events contributes to these debates, as does continued archaeological and archival work on reconstructing population change, processes of colonialism, agricultural options and impacts, vegetation change, and soil erosion. A special case is the demographic catastrophe associated with the European conquest of the Americas; the
epidemiology of depopulation remains a topic of interest, as does the subsequent “Columbian exchange” of organisms and technology.

Modernization is the most recent transformative process studied by human ecologists. Involving the division of labor and maximization of efficiency for production of goods and services, modernization involves markets, firms, and organizations including state bureaucracies. Modernizing societies have tended to succeed through the same combination of improved efficiency, population growth, violence, and ideological appeal as marked the earlier triumphs of agricultural civilizations over chiefdoms and of agricultural villagers over foragers. Frameworks of analysis of modernization have varied: classical and neoclassical economics, Marxism (and Gramscian variants), Weberian sociology, and Foucauldian attention to discourse and power, for example.

The rise and development of the modern world has also been a topic of interest in human ecology. In addition to the Columbian exchange and other impacts of long-distance trade, environmental historian Gregory Cushman (2013) has called attention to the end of the ecological “old regime” as fertilization has been decoupled from local nutrient sources. First, guano and nitrate mines provided ready sources of transportable fertilizer, and, second, chemical fertilizers allowed for industrial agriculture. In addition to capitalists, international technocrats have been important in pushing this agenda. Innovations in technology, fertilization, pesticide and herbicide use, and genetically modified organisms (GMOs) have enabled increased agricultural productivity. Sometimes these have been managed by smallholders; Robert Netting (1990) argued that, even under conditions of modernization, family farms have certain advantages and can persist. Large-scale farms have become dominant in many areas, however, and the negative social and ecological impacts of the modern food system has attracted research attention.

Modernization has been associated with fossil fuel use, and increased food production (associated with meat eating) with deforestation; both have added greenhouse gases (carbon dioxide and methane) to the atmosphere, with impacts on climate change. Human ecologists have increasingly paid attention to the challenges of human-induced climate change, in terms of analyzing the impact of different practices, the array of alternatives for adapting, and policy alternatives to enhance resilience, adaptation, and the reduction of greenhouse gases (including sequestration and payments for ecosystem services). The social and environmental impacts of modernization have sometimes been analyzed in the context of Kuznets curves, suggesting that initial increased stresses and inequalities will eventually be followed by reductions in problematic variables. Kuznets himself looked at income inequality, but others have examined pollution, deforestation (forest transition theory), and other impacts.

Demography is another crucial variable in modernization. By improving public health, modernization initially triggered a population explosion even in regions far from the industrializing core. In the 1960s neo-Malthusian biologists were warning about the dire prospects of starvation (Paul Ehrlich) and overuse of common resources (Garrett Hardin). However, research has shown that most societies have possessed the knowledge and means to control their fertility. Research conducted at the School of Forestry at Yale, the Land Tenure Center at Wisconsin, and the International Association for the Study of Common Property has also demonstrated the widespread success of local groups in managing common resources.
In 1945 the demographer Frank W. Notestein suggested that, as the cost–benefit ratio of children goes up, birth rates tend to go down, leading to a “demographic transition” even in the absence of modern birth control or antinatalist policy. This theory has tended to be verified in almost every country experiencing modernization. The education and employment of women has increased the maintenance and opportunity costs of children, reducing birth rates to the point that even many poorer countries are now on a trajectory to population stability or even population reduction. The overcoming of the Malthusian dilemma is making possible new human ecological adaptations, although research on the implications of these trajectories is in its early stages. In theory, environmental impacts should decrease as populations decline, but much depends on consumption practices (per capita resource use), which in turn are connected to resource use.

Political ecologists have focused on power imbalances and resulting environmental impacts in colonial and modernizing contexts. Much effort has been devoted to critiquing official narratives of science, conservation, and progress. Some political ecologists have supported non-market forms of modernization as pursued by states and other institutions, while others have critiqued modernization itself, arguing for a postdevelopment world. Feminist political ecology has promoted research on gendered spaces, counter-mapping, and resistance to hegemonic discourses. Activist scholars engage with historically marginalized groups such as indigenous peoples or women to help them to organize to combat issues such as environmental racism. Others engage in ethnographic research, which sees local people as the keys to both understanding and solving environmental problems.

Much human ecological research has looked at particular regions and places in terms of their specific resources (including ecosystem services), hazards, and possibilities for adaptation. Many studies have combined ethnography, field research, and archival research in a process of “progressive contextualization” (Andrew Vayda), others have assembled teams to create and calibrate models using systems frameworks, while still others have focused on the role of political actors as implementing and resisting processes of colonialism and modernization.

Polar research has looked at arctic and subarctic adaptations to severe winters, short summers, and limited plant productivity. Climate change may lead to the enhanced importance of these lands. Arid land research continues to document the wide range of social and technological practices to tap, distribute, apply, and drain water supplies, as well as the actual and potential impacts of salinization, climate change, and conflict over water. Grassland research has focused on pastoralism and ranching, and also on changing patterns of grain and soybean production in the contexts of technological and climate change.

Mountain research has included detailed studies of management, including agricultural biodiversity, agricultural terracing, mountain irrigation practices, mountain wetland utilization, traditional rotational practices, adaptations to low air pressure, and use of multiple niches (verticality). Climate change, including glacier melt, may lead to upward movement of altitudinal temperature zones, impacts on pathogens including malaria, and changing patterns of precipitation.

Temperate forested areas have been the scene of a variety of agricultural technologies, including shifting cultivation, dairy farming, rice paddy agriculture, truck farming for cities, and animal raising, as well as efforts at forest conservation. These regions in North America, Europe, Japan, and elsewhere have been the scenes of experimentation in organic and other alternative
forms of agriculture. Tropical forests have a similarly wide range of agricultural possibilities; research has focused on traditional (dark Earth, raised field) and modern (ranching, soybean) cultivation practices.

Many other environments (coastal areas, oceans, Mediterranean climates) have attracted human ecological attention. Human ecologists are increasingly addressing concerns for sustainability in local contexts, where culture, demography, politics, economics, soils, climate, and water are involved in trajectories with unique characteristics. Maps, remote sensing, and geographic information systems (GIS), including participatory mapping and participatory action research, provide important tools for visualizing relationships. Fieldwork, including field ethnography, can document practices and alternatives, as well as all-important perceptions and narratives that influence adaptation and other behavior. Arenas of research and policy include land change studies, sustainability studies (combining human rights, poverty reduction, and environmental protection), up through and including education (including service learning and cooperative international education), geo-tourism, and activism. One ambition of human ecology has always been to suggest norms for behavior, including models, if not for some future equilibrium state, at least for institutional relationships and environmental practices that can promote adaptation for the future.

Since many disciplines have claimed proprietorship of the term “human ecology,” its usefulness as a common term for the study of human–environment relationships has been compromised (see Gross 2004). However, its use continues to be promoted by Human Ecology: An Interdisciplinary Journal, which began publication in March 1972. Other organizations promoting human ecology include the Society for Human Ecology (which publishes the Human Ecology Review) and various specialty groups such as the Cultural and Political Ecology Specialty Group of the Association of American Geographers.

SEE ALSO: Agriculture; Annales School; Climate and societal impacts; Culture; Environmental determinism; Environmental history; Geoarchaeology; Geography and the study of human–environment relations; Land change science; Political ecology; Sustainability science

References


Human geography

Audrey Kobayashi
Queen’s University, Canada

Human geography has ancient roots in many parts of the world. It became a formal discipline in Europe during the mid-eighteenth century. The term “geography” derives from the Greek geo (Earth) and graphia (writing), and is often defined in the broadest terms as Earth writing, creating a large scope for the Geographical imagination, which has not diminished since ancient times. Geographers’ imaginations, however, have always been guided by what is defined as knowledge of the world, as well as by the political, economic, and cultural interests embedded in human knowledge. Above all, therefore, human geography can be understood within the contexts of the interrelated processes – including imperialism and colonialism, capitalism, globalization, and recently neoliberalism – that have written the script of global development.

Geographers often bemoan the gap between public understanding of the discipline and what geographers actually do, yet among human geographers there is a very broad range of interests, methods, and approaches. Human geographers value the flexibility of interests and the lack of disciplinary dogma, however, and recognize that any approach that includes concepts of space, place, location, landscape, nature, or region is human geography. But all of these terms also represent concepts that are utilized in daily life. Their historical usage has varied, and they have been enthusiastically contested by geographers over time.

Geography in historical context

Most historians of geography locate Ancient geography in ancient Greek and Roman scholarship and even earlier (Burke and Lowenthal 2011; Mayhew 2011). It also has strong roots in early Chinese and Arab scholarship. It encompassed virtually everything that people knew about the Earth, from speculation about faraway places and how to get to them (often with the aim of conquering them), to understanding human interactions with their environments at a local scale. Arab and Chinese scholars had maps that informed flourishing trade networks long before the Europeans began their “age of exploration.” Human geography today, however, is closely associated with the process of Globalization, from its European roots in the fifteenth century up to the present. Early modern European geography was a practical affair which depended on exploration and cartography to support not only the geographical imagination, but also the projects of imperialism and colonialism. As the world became a vast trading network that allowed movement of commodities – and increasingly people – from one part of the world to another, relations between center and periphery became the foundation of uneven development centuries later. The scale of global expansion increased rapidly as a result of European exploration and colonialism, providing knowledge about lands distant from Europe, and about their flora and fauna. Geography as the study of the Earth became the basis not only for formal studies in universities, but also for the massive power of the European commercial network, attended by the movement not only of natural resources (gold, furs, timber) but also of people, including...
HUMAN GEOGRAPHY

colonists, slaves, indentured laborers, and in many cases unwanted humans such as convicts and religious outcasts. To obtain geographical knowledge by mapping the world was often quite literally to take possession of the world. Human geography is thus associated not only with imagination, innovation, and wonder but also with some of the most brutal practices, military and nonmilitary, of humans toward one another.

Geography as a formal discipline was established in European universities during the Enlightenment period of the mid-eighteenth century. At that time what we call geography today ranged across the fields of history, biology, zoology, geology, and more – anything that involved Earth writing. Immanuel Kant (1724–1804) held a chair at the University of Königsberg where he taught “physical geography.” Kant’s influence on the discipline is twofold and profound. First, the published notes from his lectures include wild speculations about racial differences, which reinforce the superiority of European culture and human bodies, and thus support the dominant discourse about colonial expansion (Kant 1997/1775). It is alleged that Kant garnered at least some of his knowledge by interviewing sailors returned from distant sea voyages. In any case, he began formal scholarship on the geography of races, based on an idea of the other as savage and inferior, which permeated human geography – and science as a whole – for two centuries thereafter (Livingstone 1992; 2003).

Kant’s second, and arguably more enduring, contribution was to write extensively on the concept of Space, which he viewed as a form of cognitive hardwiring that allowed the human mind to perceive and organize the external world. Understanding space for Kant was fundamental to ontology and epistemology. Spatiality, therefore, is not only a foundational concept in geography but is important for all disciplines.

Ironically, over the next two centuries geographers discussed extensively Kant’s notion of chorography, the spatial arrangement or clustering of things on the surface of the Earth, but his understanding of space was largely taken for granted. Only much later did it become a foundation of the discipline, and then with a very different ontology, based on the nature of space rather than on the human capacity for spatial experience.

During the nineteenth century, geography grew in stature as a scientific discipline. University offerings increased substantially, albeit still available only to an elite, almost completely male, student body. Especially in the United Kingdom and its Commonwealth system, the curriculum was strongly geared to Eurocentric colonialism. Geographical associations were devoted to collecting information about distant places brought gloriously into the imperial fold. Reports to these societies – in the case of the Royal Geographical Society (RGS) from the likes of Livingstone, Stanley, Scott, Shackleton, Hunt, and Hillary – became an important feature of public intellectual life. Publications, including eventually the National Geographic as the largest purveyor of popular geographical knowledge in the United States, had a significant impact on the anglophone worldview (Johnston 2011).

Meanwhile, scholarly developments in Germany and France were the most influential in establishing the formal discipline of human geography in the nineteenth century, as universities in Germany grew in conjunction with expansion of the imperial state. Carrying on the Kantian tradition of defining geography (along with history) as a foundational and overarching discipline, Alexander von Humboldt (1769–1859), Karl Ritter (1779–1859), and Friedrich Ratzel (1844–1904) were pre-eminent German geographers who most profoundly influenced the future of the discipline. Von
Humboldt traveled extensively, and his huge corpus of writing gave substance to a discipline that was seen as feeding not only imperial impulses but also a growing public and scientific interest in understanding all aspects of the world. Ritter “brought long-term depth to the explanation of human interaction with the various aspects of the Earth” (Berdoulay 2011, 76). Ratzel coined the term “anthropogeography,” which is close to the notion of human geography today, although many have identified Ratzel as inspiring, if not espousing, environmental determinism. In France, in contrast, Vidal de la Blache (1845–1918) and the French School practiced la geographie humaine as explicitly non-deterministic, focusing on human beings as agents of environmental change.

**Geography in the twentieth century**

By the twentieth century, according to Berdoulay’s review (2011), human geography had grown along five trajectories, all with deep Kantian roots: Environmental determinism, Regional geography, Landscape studies, methodological debates over the nomothetic–ideographic distinction, and the conflictual question of human geography’s political role.

The most overt environmental determinism is associated with American geographers such as Ellsworth Huntington (1876–1947), Griffith Taylor (1880–1963), and, in a much more moderate form, Ellen Churchill Semple (1863–1932), who as a student in Germany was influenced by Ratzel’s lectures. The question of environmental determinism – the belief that both human physical and mental traits and their capacity for civilization are determined by environmental conditions – generated vigorous debate. From the time of the Enlightenment, racial science, including the classification of separate “races,” was considered by most scholars to be a legitimate approach to understanding human difference. Geographers were an intellectual force in arguing, as had Kant, that racial characteristics were determined by climate, implying that one’s place on Earth was a significant measure of one’s place in the human hierarchy. Although no geographers challenged the existence of race on scientific grounds, much less addressed the process of racialization, determinism was gradually outweighed by possibilist ideas by mid-century. In the process, there developed a widening gulf between human and physical geography, mainly as a result of the process of scholarly specialization but also – and ironically – as a reaction to environmentalism. Physical geographers drew away from climatic determinism in order to study climate and its effects without reference to humans. Human geographers eschewed climatic determinism but continued to consider climate, and its classification, as an important aspect of regional geography.

By the mid-twentieth century, chorological thinking had given rise to the notion of the region and the landscape, as sections of the Earth’s surface where certain contiguities, or arrangements, of things and, of course, people are located. Controversy raged, however, over the processes at work and over what processes informed geographical science. For Carl Sauer (1889–1975), it was the process of culture that gave rise to the cultural landscape (Sauer 1925). Richard Hartshorne (1899–1992) named spatial variation, or Areal differentiation (or chorology), as the fundamental but taken-for-granted principle that justified his definition of geography as a taxonomic attempt to account for all differences across the surface of the Earth (Hartshorne 1939; 1959). The understanding of regions and of cultural landscapes represented fundamentally
different interpretations of geography’s subject matter and of the ontological status of sections of the Earth’s surface. According to Nicholas Howe (2011), Sauer and Hartshorne were influenced by different strains of German scholarship, the former by the antipositivist cultural approach of scholars such as Otto Schlüter (1872–1959) and Siegfried Passarge (1866–1958), and Hartshorne by the urging of scholars such as Alfred Hettner (1859–1941) to enshrine chorology as a scientific approach. For the landscape advocates, chorology represented a narrow classificatory approach to science rather than an understanding of process; for the chorologists, landscape was vague and failed to give geography a suitable, unique disciplinary base (see Mathewson 2011).

The ideographic–nomothetic distinction was just as hotly debated. Fred Schaefer (1953) caused a stir when he suggested that geography should abandon the ideographic approach (referring to regions or areas as unique) based on German idealism from Kant to Hettner and Humboldt, and instead should adopt a nomothetic approach, based on the positivist science of spatial laws. Hartshorne (1955), among others, responded vociferously that Schaefer’s work involved the “falsification” of science and facts about the discipline, and reiterated his long-standing belief in the concept of areal differentiation (the study of things in relationship to one another), rather than the study of particular phenomena, as the basis for the discipline (Hartshorne 1958). These were opening volleys in debates that continue today, but they are deeply rooted in the Eurocentric philosophies of the eighteenth century.

Geography flourished following World War II. Trained geographers, with their ability to create maps and their knowledge of different parts of the world, had been instrumental to the war effort in every country involved, and many of these geographers went on to productive and influential scholarly careers. Universities grew; cities grew; so too did gross national products, at least in the developed world. By the 1960s, human geography was becoming more specialized. Urban geography came into its own, especially in relation to the profession of planning, as cities with vast suburbs grew at unprecedented rates throughout North America and Europe. Economic geography also grew, animated by new methods of modeling economic systems, newly available datasets, and new developments in computers and regional science. Cultural and historical geography remained predominantly focused on landscapes and the methods of areal differentiation, but international areal studies had received a large boost as a result of the development of international expertise, especially in the United States, during wartime.

Barnes and Farish (2006, 807) argue that after World War II the “hallmark of change was an altered idea of science, made manifest in a different conception and treatment of ‘region,’ long a cornerstone of geographical inquiry.” According to this account, the heavily funded technoscience that developed during World War II and burgeoned during the Cold War was “formalized as a spatial science of society” (Barnes and Farish 2006, 808), founded on a political concept of region as a structural phenomenon deployed in the interests of political power. This interpretation takes the concept of region far from its seemingly innocent roots in cultural geography. Neil Smith (2003) also lays out the political implications of geographical scholarship in his sweeping analysis of the discipline based on the work of Isaiah Bowman (1878–1950), whom he depicts as “Roosevelt’s geographer” for his role in the redrawing of national boundaries after World War II. Bowman advocated a form of US-centered globalism and progress. He also caught the imagination of geographers with his notion of the frontier as being made up of those
diminishing sites still subject to the process of civilization.

The quantitative revolution

Whether the term “revolution” is apt could be debated, but from the 1960s geographers began to benefit from both the increasing power of technology and the sophistication of mathematical models. Major developments occurred, especially in urban and economic geography, which thrived at a few leading geography departments at the time, in particular at the University of Washington in the United States and at Bristol University in the United Kingdom. Perhaps at no other time has the discipline been invested with so much sheer intellectual power, leading not only to a proliferation of scholarship but also to scientific claims that offered disciplinary status as well as promise for geography’s contribution to knowledge.

To map spatial relationships between urban places, geographers turned to Central place theory, a conceptual model developed by the German geographer Walter Christaller to address optimum distances between cities and towns of different orders according to the goods and services they provided. During the 1960s and 1970s, central place theory was perhaps the single most heavily adopted concept in the discipline, not only applied by geographers to analyze spatial distributions but also utilized by planners and governments to address practical problems of urban growth. During this period of the most rapid urban and economic growth in world history, the emphasis was on rationality, optimalization, and an at times utopian assumption of universal progress.

The quantitative turn was more than sophisticated methodology. In the preface to *Explanation in Geography* (1969), arguably the most influential book to emerge from this period, David Harvey claimed that the quantitative revolution involved not only learning technical methods but also adopting a fundamental philosophical shift. It involved deep questions about the existence of a systematic, external structural order that could be discovered and through which scholars could interpret the world. The famous line “by our theories you shall know us” (Harvey 1969, 486) concludes the book that became required reading in most university human geography courses.

The Swedish geographer Torsten Hägerstrand (1916–2004) made a monumental contribution to the discipline, initially by developing the notion of diffusion as a quantifiable process of distribution of things over the Earth’s surface. Diffusion modeling became a standard classroom subject across all branches of geography. Hägerstrand was most interested in human behavior, however, and developed the field of Time geography and space–time prism, based on a concept of space and time as finite resources that people manage, thus creating patterns of daily movement. Notwithstanding later critiques that his theories likened human beings to physical particles, Hägerstrand influenced a vast number of detailed empirical studies based on a fascination with spatial pattern and process, which seem to be exhibited in all human life.

Quantitative approaches are often conflated with “positivism.” Geographers of the quantitative revolution tended to assume a number of aspects of positive science, including a belief in external, law-like structures, an assumption of neutrality (which discouraged overt ethical positions), faith in scientific method, and an assumption of a homologous relationship between empirical evidence and truth; but few adopted a more formal philosophy of logical positivism. These research tenets tended to discourage human geographers from engaging in or opining about politics, although, of course,
many scholars today would claim that their work was nonetheless implicitly political.

Methodologically, two historical precedents were especially important. The Chicago School had advanced the concept of urban ecology for modeling the spatial development of the city. Geographers took advantage of new quantitative methods to map and model what became known as the “internal structure of the city,” a term applied widely in both university courses and geographical publications (see Bourne 1971; Berry and Kasarda 1977). Urban ecology was inspired by the work of the Chicago School founder, Robert Park, who used biological, Darwinian metaphors to interpret the spatial arrangement of people, as well as what he called their “moral economy,” across the city. Urban ecology thus presented a powerful means of locating and describing the process of city formation.

The assumptions of quantitative geography shifted substantially over the years. Reg Golledge (1937–2009) pioneered the subdiscipline of Behavioral geography, focused on cognitive processes and spatial reasoning. Gilbert White and others established a new field of environmental perception studies. Such approaches pushed geographers to consider human beings as environmental actors. The geographical literature of the 1970s and 1980s also displays what Nayak and Jeffrey (2011, 54–57) term “extension,” early examples of mixed-method approaches that extend a quantitative analysis by adding qualitative methods to a quantitative foundation. Nayak and Jeffrey distinguish extension and revision, the latter referring to the complete rejection of external laws or structures. A hybrid approach derived from a philosophy of critical realism that received considerable attention starting in the 1980s was that of “structure and agency,” premised on the assumption that some things are structured while others are the result of human agency. There was considerable debate about what things are structured and what agential, and about the circumstances in which one or the other comes into play.

Quantitative methodologies remain a very important part of the discipline today, and have grown ever more sophisticated. The theoretical assumptions behind the methods that Harvey had valued most highly, however, came under increasing critique, starting in the 1970s. The positive science of space as an external category or a container has largely fallen away, although it is still not uncommon to read geographical studies that seem to assume an ontology of space. Many geographers felt, however, that the emphasis on space as a neutral and externally structured force was reductionist, and that it failed to capture the full meaning of humanity. Others questioned the ethical implications. Central place theory, for example, which had been the topic of Christaller’s PhD thesis in 1938, was a central tenet in the Nazi plan to repopulate the plains of central Europe after they were cleansed by the Holocaust. Christaller was actually hired by the Nazi regime to work on the plan (Minca and Barnes 2010). But for a discipline in which, at least on the surface, it was the veracity of the theory that counted rather than the ways in which it was applied, Christaller had contributed to what was hailed as the most important theory in human geography, and he was given the Association of American Geographers’ (AAG) Outstanding Achievement Award in 1964 (Kobayashi 2014).

Urban ecology does not have such infamous roots, but it represented a putatively neutral reading of the ways in which people are distributed throughout the city, treating such attributes as race, class, and gender as variables that could be mapped to reveal social geographical patterns. It did not address the processes by which race, class, and gender came to be the major factors
differentiating human beings; nor did it address the experiences of those who were racialized, classed, and gendered. Linked to Robert Park’s notion of a Darwinian process of survival of the fittest across urban terrain, urban ecology is therefore one example of the ways in which the assumptions of scientific process, and the uncritical use of scientific categories, can lead to and reinforce popular notions of difference.

Examined retrospectively, these two concepts, central place theory and urban ecology, are examples of some of the far from innocent ways in which human geographers have attempted to understand the human condition. The point that would be made by critical geographers today, however, is not that human geography was – at least in any simple way – unethical, but rather that it reflected both the overt ideologies and the implicit assumptions of its times, and that it has done so throughout its long history. Only occasionally can we identify a direct connection between geographical scholarship and political power, although Walter Christaller provides one example, and Isaiah Bowman another, of geographers who were literally men of particular places and times. It was in the context of the times that some of the brilliant geographers who had advanced the quantitative revolution began to shift their ideas again.

The rise of radical geography

While the roots of radical geographical opposition to the dominant political powers can be traced at least to the nineteenth century when, for example, Russian geographer Peter Kropotkin (1842–1921) advocated anarcho-communism, it was during the 1960s in North America that radical geography gained visibility. The 1960s were a period of turbulent social change. The world was caught up in the tensions of the Cold War, which mobilized governments and citizens alike to express or confront civil fears, entrenched ideologies, and the implications of global systems of power. Many geographers were part of the so-called student movement, and were concerned about the state of the world. They participated in movements to overthrow colonialism and in the antiwar movement, the Civil Rights Movement, the environmental movement, and a rapidly growing feminist movement. In 1969 the AAG was scheduled to hold its annual meeting in Chicago but, in response to police violence against demonstrators at the Democratic Convention in that city the year before, protests mounted within the discipline to the extent that the meeting was moved to Ann Arbor. There, a growing group of geographers opposed both to the dominant social trends and to the scholarly trajectory within geography began to coalesce around what would become known as Radical geography (Peake and Sheppard 2014).

The impetus behind radical geography was to change a world in which structural inequalities define human existence, and radical geographers undertook to rewrite spatial theory. At first, radical geography was broadly socialist, organized through a number of coalitions. The largest of these was the Union of Socialist Geographers, which met periodically through the 1970s and was made up of geographers from a range of national backgrounds studying or teaching at universities mainly in the United States and Canada. In 1969 Antipode: A Journal of Radical Geography was established at Clark University. At first the journal published a range of anti-establishment articles on issues such as poverty, war, colonialism, racism, and gender, many of them passionate about the need for social change and written by scholars who were activists in every sense of the word.
Some geographers pushed for a grassroots movement to bring about change, inspired by Bill Bunge’s work on the streets of Detroit through the Detroit Geographical Expedition and Institute, founded in 1968 with Gwendolyn Warren, an African American community leader. Several “expeditions” mounted in other cities by radical geographers connected academics and local activists in what is probably the first instance of Participatory action research (PAR) in the discipline (Heyman 2007).

According to Peake and Sheppard (2014; also Peet 1977), who interviewed many of the early proponents of radical geography, it took only a couple of years to coalesce into a Marxist undertaking. David Harvey published the first explicitly Marxist paper in anglophone geography in *Antipode* in 1972. Weighing in on an ongoing conversation over the segregation of African Americans, Harvey (1972) opined that it was not more empirical work documenting the geographical concentration of race and poverty that was needed, but a Marxist theory that would account for the production of inequality as a result of capitalism. The strong early interest in race, stemming from the Civil Rights Movement, therefore, was quickly occluded as interest shifted to questions of class.

Notwithstanding the activist passion of the 1960s and 1970s on the part of Bunge and others, the enduring contribution of radical geography during that era is theoretical. Throughout the 1970s, as theoretical debates became more sophisticated, radical thought was consolidated in a Marxist framework that viewed capitalism as the major engine of historical change, and class difference as the result in the spatial ordering of society. Some, such as David Harvey and Dick Peet, addressed the big questions of capital formation and power at a global scale, although Harvey also dedicated a great deal of writing to connecting the local and the global, a topic that became very important in radical geography by the 1990s. Geographers such as Milton Santos and James Blaut contributed to understanding colonial or imperialist relations between the developed and developing world as both economic and racialized. Still others focused on poverty at the local scale. Neil Smith’s (1984) *Uneven Development* became one of the most influential books of the twentieth century.

Marxist geography was heavily critiqued, however, by those who viewed the structuralist version that dominated (although by no means completely) as substituting one deterministic structure for another. The so-called relevance debate occurred through a disparate set of introspective publications through the 1970s, instigated by a 1974 presidential address to the AAG by Wilbur Zelinsky, in which he opined that the “demigod” of science, “In grappling with the complicated, abstracted, fluid, and intersubjective data of the social realm … has been applying a fatally inappropriate model to the world of human beings” (Zelinsky 1975, 139). Zelinsky was among a growing group of human geographers who shared the Marxist concern with issues such as alienation, but he advocated a nonstructuralist approach that could move beyond scientism – that is, belief in any system of causal, universal, final, or perfect knowledge – as an article of faith. Zelinsky was also one of the first in human geography to cite Michel Foucault (1926–1984), whose work would prove to be a dominant influence on human geography some two decades later, and would play a role in bridging the gap between Marxist and humanistic geography.

Humanistic geography

It is difficult to date the origins of Humanistic geography. The Enlightenment roots of the modern discipline were broadly humanist, and
HUMAN GEOGRAPHY

included a number of the German philosopher-geographers who addressed landscape change through human or cultural systems. Carl Sauer’s interest in cultural landscapes was certainly humanist in the broad definition of that term. Clarence Glacken’s (1967) *Traces on a Rhodian Shore*, with its emphasis on the human yearning to impose order on nature, epitomizes a humanist perspective and the power of human agency to transform the world. During the 1970s, many geographers who were disillusioned with positivist perspectives “humanized” quantitative geography by focusing on the psychological factors that influence human experience of their surroundings.

But the explicit use of the term was established in 1978 by David Ley and Marwyn Samuels whose edited collection, *Humanistic Geography: Prospects and Problems*, brought between two covers a range of authors loosely described as humanistic. The collection was informed by “a reawakened awareness of humanist principles and aims over and against a preoccupation with the techniques of scientific rationality” (Ley and Samuels 1978, 1). The influences here were from early twentieth-century philosophies, including American pragmatism, symbolic interactionism (which claimed a very different lineage from the Chicago School than the urban ecologists), and continental philosophies that included phenomenology and, in a limited way, existentialism. Humanistic geographers such as Ley and Samuels, Anne Buttimer, Donald Meinig, Edward Relph, and Yi-Fu Tuan opened up new and exciting avenues of study that linked geography to literature, to symbolic landscapes, and to the world of ideas, impressions, and representations that include the emotional, aesthetic, and generally experiential. Methodologically, humanistic geography shunned quantitative techniques for qualitative, ethnographic, and textual methods. Since that time, ethnogeographic methods have come to dominate field research in human geography.

For many radical geographers, however, humanistic geography, having arisen from within the relevance debate, was just as irrelevant as earlier regional and quantitative approaches. Radical geographers still sought an understanding of structural conditions, even if understanding of the exterior and a priori status of such structures was beginning to soften. Feminist geographers such as Doreen Massey and Linda McDowell emphasized that humanistic geography was still pretty much a male domain. Much of humanistic geography also came under scrutiny for its idealism, emphasizing geographies of the mind seemingly abstracted from its material surroundings. The abstraction of external structures and internal structures thus presented two sides of a single problematic coin.

The walls come down: poststructuralism as the new paradigm

In 1989, as the destruction of the Berlin Wall symbolized a new turn in world history, human geography underwent a significant paradigm shift. Many geographers who, through the 1980s, had identified themselves either as Marxist or humanist engaged one another much more under the general umbrella of poststructuralism, influenced mainly by the post-World War II convergence of those categories in continental philosophy (Kobayashi and Mackenzie 1989). Poststructuralism is a large umbrella of approaches with common roots in anti-essentialist historical materialism: a rejection of independent external structures in favor of recognizing all human history as socially constructed. During the 1990s, debates over structure versus agency, therefore, were largely superseded, and replaced with discussion of the relative power involved in the historical
structuration of human conditions. These influences were expressed across the subdisciplines of human geography, including in growing areas such as Health geography, Tourism and Recreation geography, Political geography, Social geography, and Urban geography.

The development of poststructuralism has led to the increasing use of the term Critical geography to describe geographies with radical roots and a strong influence of both nonstructural Marxism and a range of humanisms. The term “poststructuralism” has no inherent meaning, but generally reflects the growing interest in the humanistic Marxism that dominated post-World War II continental philosophy. Although geographers interacted with many social theorists across the social sciences and humanities, two stand out: Henri Lefebvre for his theories of the construction of “space”; and Michel Foucault for his illustration of the emergence of social institutions and the role of discourse as a form of power/ideology in institution building.

Postmodernity came to the attention of geographers during the 1990s as a collection of global characteristics that coalesced especially after the end of the Cold War and the demise of most socialist states. For some, it was an aesthetic movement that involved new forms of art, architecture, and fashion – and therefore new landscape features – and that rejected or mocked the more rigid and ideologically legible forms that characterized modernism. For others, it was a potentially emancipatory political movement that celebrated differences and emphasized the rights of minorities, including indigenous peoples, around the globe. For members of the Los Angeles School of urban geography, it was a set of characteristics that marked Los Angeles and other emerging world cities as a distinctive late twentieth-century form based on a shift from productive to postindustrial finance, an expansion of technological and creative industries, increasing inequalities in a highly differentiated labor market, as well as a certain style of living that Soja (1996) characterized as constituting a “thirstspace” of heterotopic lifestyles.

As with all “isms,” it is important to distinguish between postmodernist critique and a critique of postmodernism. David Harvey made this point clearly with The Condition of Postmodernity (1989), a direct response to Jean-François Lyotard’s The Postmodern Condition (based on a study commissioned by the government of the Canadian province of Quebec, which saw itself as an emerging postmodern society). Postmodernism as a theoretical perspective strongly associated with deconstructionist theory in the humanities has been widely critiqued by geographers as failing to provide a viable alternative political economy, and, with its wide open, anything-goes approach to methods, whose irreverence is often also irrelevant and, at worst, crosses over into disdain for the actual struggles to survive that face most of the Earth’s populations.

The turn to postmodernism also encouraged a Cultural turn in geography that flourished during the 1990s. Encompassing a wide range of perspectives and topics, the cultural turn refers to the recognition of the importance of cultural systems in influencing every aspect of human relations. Geographers have been less interested in the formal aspects of culture such as film, music, literature, and art – although certainly geographical studies of those formal activities have burgeoned as well since the 1990s – than in the cultural constructions of everyday life, including work, politics, the economy, and every aspect of human differentiation along lines of class, race, gender, or any culturally defined identity. Cultural geographers (e.g., Jackson 1989) were influenced by social theorists such as Stuart Hall, Paul Gilroy, and members of the Centre for Contemporary Cultural Studies (known as the Birmingham School), whose work explores the
ways in which contemporary life is structured through processes of representation, as well as by Michel Foucault, whose concept of discourse emphasizes the structured and power-laden ways in which human relations are forged and solidify into institutions. Geographers have been particularly concerned with Colonialism, decolonization, and neocolonialism as cultural processes of creating racialized difference and inequality at a global scale.

Poststructuralism profoundly influenced the emergence of Feminist geography and Antiracist geography, both of which depend on the anti-essentialist position that gender and race are historical constructions based on uneven social relations. In other words, gender is historically made and there is no essential sex; and race is a product of the historical process of racialization. Both of these perspectives therefore challenge uncritical assumptions of “race” and “gender” as variables that can be mapped without questioning the processes by which they are historically produced, and the lived conditions that result.

Although anti-essentialism is now overwhelmingly accepted throughout the social sciences and humanities, human geographers have been especially interested in its implications for understanding the spatial construction of human relations, the discursive construction of landscapes, and the meaning of distance and positionality in establishing difference, power, and social justice. In the process, interest has also widened to recognize that dominant identities are also socially constructed, with attention to the ways in which, for example, masculinity, whiteness, and heteronormativity came to represent normative human traits and values.

With its emphasis on the dialectical relationship of humans to the entire Earth, poststructuralism has also brought renewed focus on ethics as part of geography and indeed, for many, as a necessary part of geography. Certainly both feminist and antiracist geographies are by definition ethical, since they treat questions of social justice that result from unequal relations of power. This point extends to questions of environmental justice, relations between the human and the nonhuman, and the spatialities of political economy and geopolitics. Geographers addressing such questions continue, of course, to debate ethical needs and implications, but seldom do they deny that ethics underlie the geographical project.

The result has been a shift from the concept of space – even as relational – to spatiality, as the human condition. It is fair to say that there exists a great deal of confusion in the geographical literature today on the concept of space, and it is often used uncritically, thus unintentionally ontologizing external space. At the same time, however, the theoretical literature in human geography overwhelmingly supports an anti-essentialist perspective which, following Lefebvre’s influential book, *The Production of Space* (1991), emphasizes human interaction with and human production of environment, and the dialectical relationship that results. The active formulation of spatiality is the single most enduring legacy of poststructuralism in geography.

**Human geography in the twenty-first century**

Human geography has always been contested, never more so than in the twentieth century. If four decades ago human geographers were engaged in a relevance debate in the attempt to bring greater clarity to the ways in which geographers might contribute to make a better world, there is little question today that the majority of human geographers see relevance as their mandate. Of course, the question of
relevance to what remains at issue, as does the question of how to go about understanding the world and its potential. Geographers have always engaged in Earth writing, however, and, as the Earth has become more populated, more complex, more differentiated in some respects and more standardized in others, larger in its possibilities but smaller in the ease with which humans command distance; as the environment has been increasingly modified, not always in ways that are beneficial to humans, their progeny, or the Earth itself; and as technology has placed at our disposal increasingly powerful means of writing the Earth, the discipline of human geography has become more diverse, more interesting, and more challenging.

The new age of big data

Human geography has always involved a fascination with the technical, from the earliest maps, highly speculative over the extent of the Earth and its contents, to today’s techno-world where the volume of information at our disposal, and the means with which to portray that information, are far beyond the most speculative imaginations of our academic forebears. In no other discipline is there greater scope for harnessing technology to analysis. We have entered the era of Big data, and are mapping more and more of the Earth using the power of Geographic information science (GIScience) and the Global Positioning System (GPS), which open up immense possibilities of observing the Earth, using not only satellites and drones but also thousands and thousands of closed-circuit cameras. There is no corner of the Earth that is not potentially subject to intense scrutiny. Perhaps most powerfully of all, the vast Internet coagulates and channels information about everything, including individual people. Because they are so closely entwined with governance, military power, state regulation, and policing, new technologies are of interest to virtually every branch of human geography.

The age of big data has, of course, allowed much expanded observational and analytical power in areas where technical expertise is important, including observing climatic trends, modeling traffic and trade systems, disaster planning, and much more. It has also opened up new areas of geographical interest, such as the cultural impact of new forms of communication, especially through social media. Recent research addresses topics such as crowdsourcing, grassroots organizing, and changing consumer patterns. New forms of communication also mean new, or at least altered, forms of human relations, in which distance takes on new significance. At the same time, an increasing number of geographers have expressed ethical concerns about the range of ways in which data technology can be used by the state, enhancing its power to survey its citizens; by groups intent on bringing harm to others; and by private corporations which use information powerfully to advance the interests of capital and to intrude on the lives of consumers. All of these issues represent huge methodological scope and challenges for geographical understanding.

Methodological challenges

Technology has opened up huge vistas of geographical possibility, based on ever more sophisticated methods. But questions of methodology are much bigger, and involve tethering methodological techniques to questions of ontology and epistemology. Geographers have struggled for many decades over the critical task of going beyond the map to address just what it is they are mapping, how the mappable came to be, and what difference it makes that they choose certain things over others as the subjects of analysis.
Against the illusion of analytical power that attends big databases stand questions of what is assumed and what can be assumed.

Geographers have become more eclectic in their methodologies. Many studies use a mixed-methods approach of combining descriptive quantitative analysis with in-depth qualitative analysis. The legitimacy of mixed methods is seldom questioned, but questions do arise of representation, sample sizes, and the wisdom of making generalizable conclusions based either on large-scale numbers or on small-scale experiences. Issues of meaning rise to the surface in a context where discourse remains the dominant analytical concept for understanding both the production of the world and the production of the discipline.

A case in point is the extensive discussion over ontologies of Scale. While scale is a technical concept that represents the ratio of a map to the surface of the Earth, there is a debate whether scale itself, that is, the local, regional, national, or global, actually affects social, political, and economic processes. Antiscale theorists argue that ontologizing scale, giving it agency, has political as well as theoretical consequences. The same might be said of space, region, or any other product of the geographical imagination. Geographical concepts are not innocent. Their power is wielded to change the world.

The looming question is not only how geographers study what they do, but how they choose the objects of their concern, from people to all the other things that exist on Earth. Given the constant becoming of all things, perhaps the greatest methodological challenge is how to capture things in flux. Dynamic and interactive maps and other computer programs help but only go so far. If the world is ordered by vast, complex, fungible assemblages of things, how can the geographer present both the more durable things such as text, photographs, and numbers while still acknowledging that, although nothing is fixed, some things are more transient, harder to represent, but not thereby less important or powerful.

**Beyond – and around – the cultural turn**

The last decade of the twentieth century involved a new and theoretically sophisticated fascination with culture, discourse, and representation. The first decade of the twentieth century has seen increasing skepticism over the cultural turn. There has been a growing emphasis on things – commodities, buildings, bodies – and on the ways in which they are connected, drawn into webs of significance, given differential power to influence human experience and action, as well as on the inter-relationship between humans and things. Diverse theorists inform this trend, although it is often linked to the anti-dualist Dutch philosopher Baruch Spinoza (1632–1677). Gilles Deleuze (1925–1995) is often credited (among other things) with influencing a geography of difference that builds on geography’s legacy of understanding human relations. Bruno Latour has influenced geographers to think about the complex relationships between things (including people) on the Earth, especially in an age of advanced technology.

These influential intellectuals, along with a number of poststructuralists already mentioned, have in common a dialectical approach that emphasizes change, power, and the historical production of global circumstances. Within human geography, Actor-network theory, based on the thought of Latour and others, has been particularly well suited to analyzing the complexity of assemblages of things, the distribution of human actions and their results across a global landscape, and the systematic but often ineffable structure of those distributions.
Questions about ineffable structures have in large part underlain recent interest in Nonrepresentational theory (NRT) (e.g., Anderson and Harrison 2010) as a response to what many geographers see as the fascination with too fixed representations of geography’s cultural turn. Nonrepresentational theorists build on poststructuralist insights on the dialectical, always-becoming world to emphasize things in flux and flow, and the problems of attempting to fix meanings in place. Empirically, this approach has encouraged new research forays into such areas as dance, theater, and music, and into understanding affect as the emergent, immanent qualities between things in the landscape that influence embodied experience, but also lead interest toward the nonhuman or more-than-human world where affect is also produced.

Controversies over NRT include claims that it is universalistic and pays insufficient attention to questions of power (Tolia-Kelly 2006). Growing numbers of proponents of Emotional geographies emphasize the need to situate emotional experiences in a lived world where norms and habits are always and already structured as well as emergent, and in the same vein feminist geographers have warned of the dangers of moving beyond the everyday relations that regulate emotional lives (Thien 2005). Others, including Popke (2008) have expressed optimism over the potential of NRT for understanding embodied materialism, but skepticism over the ways in which its incorporation into geography has addressed questions of ethics. Two ethical issues arise. One is the tendency on the part of some NRT scholarship to refuse any distinction between the human and nonhuman, which both erases the power – and the responsibility – of the human to create the conditions of being, and privileges things that do not in themselves have any ethical intentions. The other issue concerns a tendency to cleave the affective from the political, with the attendant perils of failing to understand the potential for changing world conditions, for better or worse.

What human geographers do

There is a line, attributed to Peter Gould (1985) but repeated too many times to count, that “geography is what geographers do.” Recognizing that all boundaries Borders, boundaries, and borderlands are socially constructed, it may seem that geography has none. Certainly geography is an interdiscipline, but it is also increasingly interdisciplinary, as geographers engage with colleagues across the physical and social sciences and the humanities, as well as with a range of practitioners that may include policymakers, people in business, activists, or artists. Geographers also continue to import ideas from other disciplines, probably more so than the other way around. For many, the lack of discipline is a strength that allows a huge scope for the practice of Earth writing. It also creates institutional problems for many geography departments, however, as university administrations like to downsize or merge units, especially those they do not fully understand. There are increasing numbers of departments of “geography and …” (planning, environmental studies, development studies, etc.), which may be only the cutting edge of what will become the dominant institutional framework.

The question of what geographers do, therefore, is not one of simple capricious choice or of indulging an unfettered curiosity. It is a question of the geopolitics of knowledge. If at individual institutions that means negotiating administrative politics in an increasingly neoliberal environment, on the world stage it also means that anglophone geography has come to dominate
the field, a fact that has much to do with pro-
cesses that were set in motion (at least) from the
time of fifteenth-century European explorations.
Advances in communication technology mean
that it is now easier than ever before to commu-
nicate and collaborate across greater distances,
but it is also easier to impose epistemological
hegemony. The term “globalization” there-
fore needs to be understood in complex ways
(Derudder 2015), and it needs to be recognized
that epistemology and ethics are intertwined.

Questions of ethics range across the discipline
of human geography. Some of the concern has
been institutionally driven, as universities and
granting agencies apply a neoliberal lens through
which ethical precautions are strongly tied to
systems of audit and accounting, researcher/
institutional liabilities, and emerging definitions
of privacy. A broader set of ethical questions
continues the line of thinking of the relevance
debate to address issues of responsibility toward
communities and toward the Earth itself. Femi-
nist and antiracist geographies question the role,
identity, and positionality of the researcher, as
well as the difference that the scholar can make
in not only studying, but actually changing the
world (Castree et al. 2010).

Geography has seen a surge in both the number
and the analytical power of studies of Geopolitics
in the twenty-first century. The journal Political
Geography has become one of the most com-
pelling specialized journals in the discipline, as
the purview of political geography has expanded
from a much narrower focus on the spatial
patterns of formal politics to encompass the
range of discursive processes through which the
world is organized and power created, wielded,
and distributed. Political geographers have a full
slate addressing the geopolitics of world power,
warfare and violence, possibilities for peace, mass
migration, and the effects of climate change,
to name some of the most important areas of
concern. In the process, nearly every aspect of
human geography has also incorporated recogni-
tion of the political. Feminist geographers have
been instrumental in stressing that all human
relations, from the personal to the global, are
political. Antiracist geographers address the
geopolitics of colonialism and imperialism in
creating a world in which a putative science
of race was actually a means of constructing
racialized political hierarchies. Political economy
remains a powerful means of understanding the
relationship between formal political structures
and the globalization of resources. Cultural
and historical geographers today seldom fail to
recognize that all cultural systems are politically
forged and are fundamental to understanding
national, diasporic, or local identities.

Climate change, which has absorbed the
interests of more and more physical and environ-
mental geographers, is of immense importance
to human geographers, who are well positioned
to understand just what climate change means
for the quality of human life, in terms of health,
food production, and availability and in some
cases for the continued existence of places that
face inundation by rising water levels or sinking
land levels. Climate change has implications for
understanding shifting resources, human migra-
tion, changes in livelihoods, and the potential of
violent confrontations over resources and power.

Human migration has long been a main-
stay of geographical research. Going back to
the days of Chicago School influence and
urban ecology, migration – whether inter- or
intranational – was the process that underlay
urban development, as the residential spatial
patterns that ensued. In the twenty-first century
these processes continue, but geographers have
expanded their view. Questions of Race and
racism have never been more important as racial
violence, the racialization of poverty, and the
uneven distribution of the world’s resources
Human geography according to racialized patterns continues apace. Warfare has created millions of new refugees, whose circumstances cannot be separated from the geopolitics of border control. Larger and larger numbers of people are moving from one part of the world to another in search of a better life, or in some cases in search of bare life; and shifting economic activities are leading to shifting populations, causing some places to grow rapidly and others to decline. The treatment of migrants, including those subjected to human trafficking or to the bare life conditions of refugee camps, is of increasing ethical concern for human geographers.

Human geographers’ long history of speculation on the processes of spatial differentiation is evident in the abiding concentration of research on the diversification of humans living in a spatial relationship. Feminist geographers have addressed, among other things, the importance of studying gender relations in the home or the neighborhood. Geographers studying post-colonialism and decolonialism show that these processes have instigated significantly different human – economic, political, cultural – relations between core and periphery. Issues of multiculturalism as cities become more diverse encompass more than just cultural practices involving food, religion, and festivals; they also signal questions of labor markets, changing class divisions, and access to services such as health care – all of which are important ongoing challenges for human geography.

Climate change, human migration, ongoing urbanization, and the diversification of humans living in proximity to one another represent but three important research agendas for geographers in the twenty-first century, all with major geopolitical and ethical dimensions. Geographers have greater methodological power to address these issues, and greater capacity to write the Earth in ways that will make a difference.

Geographers continue to struggle to make their work known to the world, however, and perhaps that challenge is one of the most significant if geographers are to be known by the end of this century as the world changes. If the hallmark of human geography a half century ago was to be known by our theories, then perhaps an appropriate slogan for the twenty-first century is “By our actions you shall know us.”

References


Minca, Claudio, and Trevor J. Barnes. 2010. “Nazi Spatial Theory: The Dark Geographies of Carl


Further reading


Hubbard, Phil, and Rob Kitchin, eds. 2011. Key Thinkers on Space and Place, 2nd edn. London: SAGE.


Human rights

Mark Boyle
National University of Ireland, Maynooth

Human rights thought is based on the claim that human beings, precisely because they are members of the human family, are endowed with a number of rights which pertain to all, irrespective of race, age, nationality, gender, class, sexuality, disability, religion, and so on. While emerging and fermenting over a long time period, the idea of human rights arguably rose to particular prominence in 1948 with the ratification of the United Nations Universal Declaration of Human Rights (UDHR). Since then, states and peoples have consistently invoked, mobilized, and worked to enforce the UDHR, and in so doing have afforded it a degree of binding authority. This entry explores, first, the genesis of the idea of human rights and the substantive content with which the UDHR populated it; second, the various ways in which human geographical research has engaged human rights thought; and, third, recent neoconservative critiques on the one hand, and feminist and postcolonial critiques on the other, which call for declarations such as the UDHR to be historicized and provincialized.

The idea of human rights and the UDHR

Although the idea that human rights exist as a special or separate category of human entitlement has featured in many prior civilizations, a particular story about the historical and geographical emergence of this idea has settled into conventional wisdom (Nice 2014). According to this account, human rights thought began in earnest in Mesopotamia in the sixth century BCE under the progressive leadership of Cyrus, the first king of Persia. It then developed and persisted throughout the rise and fall of ancient Greece and then Rome. Strangely, and against all odds, it found life in Europe during the “dark” or middle ages (fifth to fifteenth century CE), evidenced most clearly in the English Magna Carta (1215). It re-emerged with renewed vigor with the rise of Renaissance humanism and the European Enlightenment (from the sixteenth century onward). It was developed and given shape in the thought of Enlightenment philosophers and political theorists such as Locke, Berkeley, and Hume in England; Montesquieu, Voltaire, and Rousseau in France; and Jefferson, Franklin, and Paine in the United States. It came to full fruition during the age of revolutions (1775–1848) when it served as a progenitor of the rise of constitutional republics. Here the two most oft-cited examples are the American and French revolutions and their attendant credos, the American Declaration of Independence (1776) and Bill of Rights (1791), and the French Declaration of the Rights of Man and of the Citizen (1789).

The horrors of the first half of the twentieth century, and in particular the traumas and violence of World War I and World War II, announced another pivotal moment in the history of the idea. Under the stewardship of Eleanor Roosevelt and French jurist René Cassin, and following only two short years of work, the United Nations Assembly drafted the
HUMAN RIGHTS

watershed United Nations Universal Declaration of Human Rights. The UDHR was endorsed by the General Assembly in 1948, with 48 of the then 58 member states voting for its acceptance. Fifteen of the signatories were Western states, the remaining 33 non-Western. Abstainers included Saudi Arabia, Soviet Bloc countries, and South Africa. The UDHR is not a treaty and cannot by itself impose legally enforceable obligations and sanctions on member states. Subsequently, and to give substantive effect to the declaration, the United Nations Commission on Human Rights issued two further covenants: the International Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights, both of which were formally ratified in 1976. Together with the UDHR, these covenants comprise the International Bill of Human Rights. The UDHR and the covenants it bequeathed have inspired more than 80 international human rights treaties and declarations, a great number of regional human rights conventions, domestic human rights bills, and constitutional provisions, which together have worked to put in place a (yet to be completed) legally binding system for the protection of human rights.

The UDHR comprises a short preamble followed by 30 articles (United Nations 1948). The preamble confirms the importance of the violent twentieth-century backdrop and the aspiration never to return to a world where “disregard and contempt for human rights” results in “barbarous acts” which “outrage the conscience of mankind” and where violent “rebellion against tyranny and oppression” is mankind’s only and final option. “Recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family,” the preamble continues, “is the foundation of freedom, justice and peace in the world.” The United Nations General Assembly, it declares, undertook to ensure “universal and effective recognition and observance” of the declaration, “both among the peoples of Member States themselves and among the peoples of territories under their jurisdiction.” Member states were asked to “pledge themselves” to the “promotion of universal respect for and observance of human rights and fundamental freedoms.”

Articles 1 and 2 assert that all human beings are born already endowed with basic human rights, irrespective of their race, age, nationality, gender, class, sexuality, disability, religion, and so on. These rights are both inalienable (they cannot be renounced) and indivisible (they cannot be parceled and passed on).

Articles 3 through 21 address civil and political rights. Article 3 confirms the “right to life, liberty and security of person,” article 4 prohibits “slavery and servitude,” and article 5 renders illegal “torture and cruel, inhuman or degrading treatment or punishment.” Articles 6 to 8 assert equal protection under the law, equal status in the eyes of the law, and the right to seek redress should these two prior rights be violated. Article 9 states that “no one shall be subjected to arbitrary arrest, detention or exile” and article 10 confirms the right to a “fair and public hearing by an independent and impartial tribunal,” while Article 11 declares that those charged with committing a crime should be “presumed innocent until proven guilty.” Article 12 outlaws “arbitrary interference with privacy, family, home or correspondence.” Article 13 confirms the right to mobility within states and the right to leave and return to one’s homeland at any time, while article 14 states that “everyone has the right to seek and to enjoy in other countries asylum from persecution.” Article 15 confirms the right to nationality, while Article 16 proclaims that “men and women of full age, without any limitation due to race, nationality or religion, have the right to marry and to
found a family.” Article 17 asserts the rights of people to own property, article 18 the right to “freedom of thought, conscience and religion,” and article 19 the right to “freedom of opinion and expression.” Article 20 upholds the right to “peaceful assembly and association” and article 21 the right to take part in the government of a country, “directly or through freely chosen representatives.”

Articles 22 through 27 set out a number of basic economic, social, and cultural rights. Article 22 asserts “the right to social security,” article 23 “the right to work, to free choice of employment, to just and favorable conditions of work, to protection against unemployment … to equal pay for equal work … to just and favorable remuneration and to form and to join trade unions.” Article 24 asserts the “right to rest and leisure, including reasonable limitation of working hours and periodic holidays with pay,” while Article 25 insists on the right to an adequate standard of living and states that “motherhood and childhood are entitled to special care and assistance. All children, whether born in or out of wedlock, shall enjoy the same social protection.” Article 26 upholds the right to education that supports the “full development of the human personality” and confirms parents’ “right to choose the kind of education that shall be given to their children.” Article 27 confirms the “right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits.”

Article 28 announces the conclusion to the declaration and asserts that “everyone is entitled to a social and international order in which the rights and freedoms set forth in this declaration can be fully realized.” Article 29 confirms that the rights and freedoms asserted in the UDHR may in no instance be exercised contrary to the purposes and principles of the United Nations. Finally, article 30 concludes: “nothing in this Declaration may be interpreted as implying for any state, group or person any right to engage in any activity or to perform any act aimed at the destruction of any of the rights and freedoms set forth herein.”

Human geography and human rights

Recently, the Association of American Geographers (AAG) engaged in a series of collaborative projects with the American Association for the Advancement of Science (AAAS) designed to promote the contributions that human geographical research and teaching might make to the protection of human rights around the world (Richardson and Bromley 2008). This collaboration has bequeathed the AAG with a “clearinghouse” for human geographical research with a human rights focus, otherwise known as the Geography & Human Rights Bibliographic Database (AAG n.d.); a Human Rights Coalition; and a joint project exploring “geospatial technologies and human rights.” The database lists over 700 human geographical research papers which, even if they do not deploy the lexicon of human rights explicitly, in one way or another place human rights concerns at their core. An examination of this dataset points to at least five ways in which human geographers have engaged with human rights thought: mapping and ruminating on the geographies of human rights law, enforcement, and enactment; tracking instances of human rights abuses; using geospatial technologies to document human rights abuses that might otherwise be concealed from the world; harnessing the idea of human rights to frame social movements concerned with human geographic topics; and working to uphold the human rights of human geographers and other scholars themselves.
HUMAN RIGHTS

To enforce the protection of human rights, the United Nations turns first to its own Committee on Economic, Social and Cultural Rights (CESCR), the Office of the United Nations High Commissioner for Human Rights (OHCHR), the Human Rights Council (HRC), and treaty-monitoring bodies like the Committee on the Elimination of Discrimination against Women (CEDAW) and the Committee on the Rights of the Child (CRC). Responsibility for responding to particularly heinous breaches of human rights (such as crimes against humanity, genocide, crimes of aggression, and war crimes) falls to the United Nations Security Council and to the International Criminal Court (ICC). Meanwhile, the work of translating the International Bill of Human Rights into law in particular jurisdictions is ongoing. As this bill has cascaded down through many geographical scales, it has become filtered, reworked, and transformed. Moreover, its “local” enforcement has given birth to many legal tensions, struggles, and contestations. Human geographers are interested in mapping these concrete groundings of human rights thought and in understanding the ways in which local context mediates and conditions supposedly universal rights. A good example can be seen in the European Convention on Human Rights (ECHR), drafted in 1950 as a European response to the UDHR, coming into force in 1953, binding on all member states of the European Union and overseen by the European Court of Human Rights. The ECHR has in turn given birth to human rights legislation in member states. For example, the Human Rights Act 1998 incorporates the convention into UK law. British citizens who claim human rights violations must first take their claim through the British judicial system; if this claim is rejected, they then have the option of taking their case to Europe for a determination. It is here that a politics of scale becomes important, as global, European, and British legal systems compete and jostle for sovereignty.

Human geographers are also interested in exploring geographies of human rights breaches and abuses. For example, on the basis of an analysis of the geography of 720 Amnesty International freedom writers’ letters, Selya (2012) tracked spatial and temporal variations in human rights abuses across the world between 1986 and 2006. The most common abuses cited were “prisoner of conscience,” “disappearances,” “detention without charge or trial,” “threats,” “torture,” “extrajudicial killing,” and “unfair trials.” Selya argues that, at best, the number of human rights abuses has remained stable over the period; certainly there is no sign that progress toward eliminating abuses has been made. Based on the total number of citations, the worst offenders were Syria, China, Myanmar, South Korea, Peru, Turkey, Colombia, Cuba, the Czech Republic, Sri Lanka, and India. Based on the highest abuse rates per 1 million of the population, the chief culprits over the period were Equatorial Guinea, Brunei, Western Sahara, Bahrain, Kuwait, Cyprus, Israel, Mauritania, Syria, and Bosnia and Herzegovina. Of course these geographies are dynamic and would look different again if the period from 2006 to the present were examined. Selya identifies 10 factors that together help to explain the lack of progress these data imply, including flawed conceptions of the democracy–development–respect for human rights cycle; the rise of authoritarian capitalism; the spread of neoliberalism; globalization and the demand for efficiencies; chaos and anarchy resulting from endeavors to democratize countries; deviant cultures and personalities that deem themselves to be above the law; the hypocrisy of Western democracies that cite human abuses inconsistently; the hypocrisy of signatories that promptly ignore their own commitments; the persistence of corruption and weak institutions
in new democracies; and the use of irregular paramilitaries in civil conflicts.

Geographers have also developed a raft of new geospatial technologies which enable new types of data to be collected, archived, scrutinized, stored, and visualized, and which organizations such as Amnesty International and Human Rights Watch are now finding useful when documenting human rights abuses in obscure, hidden, and difficult-to-reach places. Remote sensing and global positioning system (GPS) technologies are paving the way for a new era of big data. Big data comprise datasets that are huge in volume, consisting of terabytes or petabytes of data, and rapid in velocity, being produced in real time. Meanwhile the emerging field of geographical information science (GIScience), or geocomputation, is radically expanding the capacity of analysts to digest and to model big data. These new capacities allow evidence of past human rights abuses to be uncovered, evidence which might otherwise have gone unregistered and unrecorded. Moreover, because these developments allow for real-time data collection and analysis, it is now theoretically possible to monitor human rights abuses as they unfold, to forewarn at-risk groups and law enforcement agencies, and to prevent violations from taking place in the first instance. Geospatial technologies have been used, for example, to draw attention to attacks on civilians in Darfur, Burma, and the Ogaden region of Ethiopia; to support indigenous land rights in Guatemala; to document aerial defoliation in Colombia; to identify mass graves in Cambodia; to locate secret prisons in North Korea; and to document human right violations in Syria and track and anticipate Syrian refugee movements (Richardson and Bromley 2008).

Some critical and activist geographers are also drawing on human rights thought to inform and support the real-world political and legal struggles they themselves are actively engaged in or align with and contribute to. Framing topics as varied as climate change, housing, food security, access to public space, child protection, sexual equality, and refugee flows and asylum claims as human rights issues opens up powerful new possibilities for activist groups and social movements. Rights-based approaches have advantages over approaches that likewise seek to defend impoverished and excluded communities – for example, charity-based approaches (where powerful groups are asked to act benevolently for the sake of conscience) and needs-based approaches (where individuals are protected on the basis of what they lack) – because they insist that progressive interventions are mandated by juridical imperative and merely work to uphold basic inalienable rights. Human rights thought de facto pervaded the entire tradition of welfare geography, which came to the fore in the 1970s and 1980s and which sought to explore, as its practitioners put it, who gets what, where, why, and how. Welfare geography, in turn, has now been augmented by a potent new language which speaks of “spatial justice” and whose concern is the ways in which sociospatial processes conspire to render some social groups marginalized, impoverished, and minoritized. Political action to correct injustices may then follow. The “right to the city,” in particular, has enjoyed a meteoric rise to prominence within the human geographical activist community and has played a central role in social movements such as the Occupy movement (Harvey 2013).

Finally, human geographers are actively working today to protect their own rights as scholars. Higher education institutions worldwide are under attack, it seems, from political systems that value certain kinds of knowledge and pedagogy at the expense of others. Everywhere the concept of the public university is under threat, from fascist and despotic political regimes on the
one hand and from neoliberal funding models and interinstitutional competition for resources on the other. In this context academic freedom is being compromised and the health of all of basic, applied, critical, and public scholarship impaired (Burawoy 2011). In association with faculty from other disciplinary traditions, and in part through its learned societies, human geographers are actively confronting these threats and pressures. Given that many human geographers are interested in caring and taking responsibility for, and in acting ethically toward, others at a distance, human geographers are arguably themselves both especially vulnerable and especially well placed to contribute. For example, many geographers engage with the Scholars at Risk Network, an international network of higher education institutions established to protect scholars under attack by those in power who feel threatened by ideas that counter their own and who act to restrict academic freedom and to repress research, publication, teaching, and learning. The Scholars at Risk Network provides “positions of sanctuary” for scholars suffering serious threats to their lives, liberty, and wellbeing, and supports scholars who are effectively imprisoned in their home countries as a result of threats, restrictions, or wrongful imprisonment. This network also works to prevent more systemic attacks on higher education systems and sets standards and legal protocols to uphold academic freedoms.

The limits of human rights thought

While respecting the work of human rights thought in the world and while embracing, harnessing, and engaging the UDHR in various ways in their research and teaching, human geographers have also been at the forefront of recent calls for a critical discourse on the subject of human rights. The first section of the entry drew attention to the very particular history and geography of the idea of human rights. Critics claim that all formulations of human rights are spatially and temporally situated, embodying the cultures from which they originate, and that therefore claims to universality are questionable. Focusing on the politics of universality, as opposed to the fact of universality, critics question which institutions put the UDHR into the world, what work the declaration does in the world, who benefits from it, and how it survives and prospers over time. Such ruminations have led to at least two competing critical readings of the UDHR: a neoconservative reading which laments what it considers to be the appropriation of the idea of human rights by liberal and leftist camps; and a critical feminist and postcolonial reading that casts suspicion on the idea of human rights by dint of its imbrications in European colonial history and the work it is now being called to do as a “civilizing instrument” in our colonial present.

Neoconservative critics argue that, while the intention of human rights law is to protect and safeguard basic human rights, in reality it is being viewed by lawbreakers on the one hand, and by countercultures on the other, as a weapon to evade prosecution and/or to attack and relativize traditional and conservative values. Invoking the concept of “lawfare” (the use or abuse of law as a weapon of war), these critics argue that violent, despotic, and unsavory political regimes routinely exploit international human rights law with impunity, ensnaring lawyers in fruitless, endless, and spurious exchanges in a bid to delay, avoid, and frustrate justice. These claims take on a local hue too. In the British case, clashes between popular right politicians and ECHR law over, for example, the right of prisoners to vote and aspects of immigration, deportation, and extradition law and practice have fueled right-wing political commitments to temper the
ECHRI, to relegate it to a mere advisory body, and to reassert the sovereignty of the Supreme Court of the United Kingdom. Meanwhile the Christian right has criticized the (in its view) excessive extent to which human rights law has generated a climate of “political correctness,” which has resulted in the rights of secular groupings, Islamic militants, and LGBTQ peoples being privileged over the rights of those who hold “conservative,” “Christian,” and “British” values. Arguably, few human geographers support neoconservative critiques of the UDHR; certainly few declare their opposition publicly. But, as the far right gains a degree of popular authority in political systems across Europe and beyond, many are interested in studying the implications, especially for those who may be rendered more vulnerable should human rights legislation be repealed or reformed.

In contrast, recent writings in the field of critical legal geographies, and feminist, postcolonial, and subaltern geographies, contend that, in spite of its claims to embody timeless and universal concepts of justice, international human rights law, too betrays historical and geographical roots in the European Enlightenment and needs to be historicized and provincialized. The international legal system provides a critical infrastructure through which the international community seeks to care for (often distant) others by taking responsibility for their right to protection, justice, and legal redress in circumstances in which those rights are violated. But, given that it has been colonial and neocolonial powers that have conceived, authored, built, and now govern international law, it is pertinent to inquire into the limitations of routing caring (from a distance) through seemingly universal ethical, legal, and juridical frameworks. Specifically, critical scholars question the extent to which the UDHR has been appropriated by the West and put to work as the latest in a series of civilizing narratives that conceal Western geopolitical aggression. Some claim that the international legal system is systemically compromised because it is pro-Western by geopolitical fiat; for the West, the rule of human rights law has simply meant rule through human rights law (Boyle and Kobayashi 2015).

Meanwhile, deeper critiques again argue that international human rights law is systemically contaminated because it is conceptually Western and masculine. They point to the nexus between social power, epistemology, and human rights law. This critique has led feminist, postcolonial, and subaltern geographers to propose alternative ethical paradigms, care ethics being a case in point. Care ethics explicitly seeks to decenter established Enlightenment ethical frameworks, on the basis that these approaches to ethics are profoundly Cartesian, Western, masculinist, and limiting. Enlightenment ethics searches for justice based on universal principles that are capable of being articulated abstractly by autonomous subjects, prioritizes cool detachment, valorizes human reason, and advocates objective, neutral, and rational decision-making. The UDHR is viewed as embodying such justice through reason. In contrast, care ethics begins with a social ontology of connection, foregrounding social relationships of mutuality and trust rather than dependence, where ethical judgments need to be made in the context of caring relationships. Care ethics prompts reflection on the limitations of framing human rights discourse in and through Western ethical, legal, and juridical frameworks, and underscores the value of locating questions of justice in the context of complex and charged relationships.

Conclusion

Historically ingrained and tenacious power-laden and asymmetrical relationships and connectivities between peoples and places will ensure
that human rights thought remains of central importance in human geography. But recent feminist and postcolonial interventions in this literature insist that these very same power geometries also situate, condition, and color many endeavors (however well meaning) to frame and uphold the human rights of others (Noxolo, Raghuram, and Madge 2011). Supporters of the UDHR should not be overly disheartened by recent critical reflections on the origins of this particular declaration, the potential existence of different conceptions of human rights in differing cultural contexts, and the possibility or impossibility of ciphering and codifying a universal bill of human rights. The UDHR has served as a template and source of encouragement for progressive agendas for over 60 years, and it is essential to remember what has been achieved by international human rights law. But, to care that justice is served to others, it is necessary to question whether our justice is seen by others as a just justice and to entertain the possibility that potentially there exist other traditions of human rights thought.

**SEE ALSO:** Feminist geography; Geographic information science; Imperialism; Legal geography; Moral geography; Political geography; Postcolonial geographies; Social justice

**References**


Further reading


Most prominent in the 1970s and 1980s, humanistic geography is a conceptual perspective claiming that a comprehensive understanding of human–environment relationships must consider individual and group experiences and meanings of space, place, landscape, region, mobility, and related geographic phenomena. Partly propelled by 1960s research in behavioral geography and environmental perception, humanistic geography incorporated a wide range of philosophical approaches that included phenomenology, existentialism, idealism, pragmatism, grounded theory, and symbolic interactionism (Ley and Samuels 1978). Geographers most commonly associated with humanistic geography included Edmunds Bunkse, Anne Buttimer, James Duncan, J. Nicholas Entrikin, David Ley, David Lowenthal, Douglas C.D. Pocock, J. Douglas Porceous, Edward Relph, Graham Rowles, Robert David Sack, Marwyn Samuels, David Seamon, Yi-Fu Tuan, and John Western. The first geographer to describe humanistic geography formally as a disciplinary subfield was Yi-Fu Tuan (1976). He defined the approach as the geographic study of human beings’ experiences and understandings of space, place, and the natural world.

The development of humanistic geography

Though humanistic geography became an explicit subfield of the discipline only in the 1970s, there were several earlier geographers who, at least implicitly, pointed toward humanistic approaches, methods, and themes. Examples include Alexander von Humbolt’s interest in how landscape painting could contribute to the public’s awareness of the Earth’s interest in the Earth’s natural regions; Johannes Gabriel Granö’s efforts to develop an experientially grounded cartography that could map sensory and perceptual aspects of natural and human-made landscapes; and Paul Vidal de la Blache’s field studies of genre de vie, a term encompassing the idea that the “way of life” of a region reflected its inhabitants’ psychological, social, and economic identities imprinted on the landscape. Though not a geographer, another significant representative was French historian Eric Dardel, who examined geographicalité (géographicit), the experiential linkages that supported human worlds environmentally and geographically, including ties to places, landscapes, and regions.

In spite of these early researchers’ efforts, it was not until the mid-twentieth century that geographical thinking pointed toward a formal explication of humanistic geography. In a 1947
HUMANISTIC GEOGRAPHY

article in the *Annals of the Association of American Geographers*, J.K. Wright called for geographers to include a humanistic perspective in their studies. He advocated a subfield of geographical research that would study peoples’ subjective geographical understandings and values. He labeled this field of inquiry “geosophy,” which he defined as the examination of geographical knowledge in all manner of human forms. In a 1952 article in the *Indian Geographical Journal*, William Kirk, working independently from Wright, extended his call for study of geographical knowledge by developing the concept of behavioral environment, which Kirk defined as the environment not as it is known objectively but as it is perceived and understood by individuals and groups. Wright and Kirk’s efforts were significant for humanistic geography because both thinkers realized that geographers needed to expand their research horizon to incorporate human consciousness and cognition, since the ways in which individuals and groups structure and make sense of their world play a primary role in how they act in and make use of that world.

In the 1960s, Wright and Kirk’s ideas would help spawn a new disciplinary subfield of behavioral geography and environmental perception, which largely focused on the cognitive dimensions of environmental behavior. In the 1970s, the development of humanistic geography greatly benefited from this behavioral research, which had shifted the study focus from measurable aggregate analyses of spatial and environmental behaviors to individuals’ environmental images, attitudes, preferences, and worldviews. Though much of this research remained quantitative and focused on the consciously grounded dimensions of geographical actions, experiences, and meanings, this work was crucial for the development of humanistic geography because it helped to justify the study of human beings’ lived relationships with the places, spaces, and environments comprising their geographic worlds.

Why “humanistic geography”?

There are at least two reasons for the label of “humanistic geography” rather than “experiential geography,” “lived geography,” “existential geography,” or some similar term. First, the 1970s marked a time when psychology and sociology had already drawn on the “humanistic” label to identify new subfields in their disciplines. Unsettled by the behaviorist and Freudian perspectives that dominated psychology, Abraham Maslow advocated an alternative approach he called “humanistic psychology,” which emphasized free will, creativity, human potential, and self-exploration. Similarly, sociologist Peter Berger called for a humanistic approach in his discipline, suggesting that the societal dimensions of human life could be more thoroughly examined not primarily via social structures, networks, and institutions but via the experiences, actions, and understandings of the individuals and groups involved.

A second reason for the label “humanistic geography” related to links with “humanism,” a philosophical, ideological, and ethical perspective with a complex intellectual history often incorporating conflicting understandings (Relph 1981). Most broadly, humanism refers to a belief in the unity of humankind and in human beings’ potential to improve their own lives and worlds, making careful, critical use of accurate intellectual knowledge and relevant life experiences. Humanist hallmarks include reason, tolerance, individual responsibility, and understanding and action grounded in mature personal experience. “Humanist” originated from the fifteenth-century Italian umanista, a scholar of classical Greek and Latin literature. Originally,
these scholars used “humanism” to spotlight the core of the Italian Renaissance, which, in seeking to revive classical learning, emphasized that human beings themselves, rather than divine power, play an instrumental role in who they are and what they become. Over the centuries, many different modes of humanism arose, often contradictory philosophically and ideologically. In the twentieth century, humanism continued to incorporate a wide range of meanings, though one can argue that its primary philosophical and ethical tenets included: (i) the emancipatory potential of human reason; (ii) the significance of free, open inquiry; (iii) the understanding of things and events mostly as they offer value for human beings and human life; and (iv) the wish to make life better for all people, particularly the less able or less fortunate.

As it developed in the 1970s, humanistic geography generally hewed to these central humanist tenets but reinterpreted them in innovative ways that assimilated shifting philosophical and practical concerns, including Earth’s ecological crisis. Humanistic geographers accepted the constructive possibilities of human reason but contended that intellectual knowledge grounded only in scientific method too often misinterpreted phenomena and reduced them to inaccurate, piecemeal counterfeits. Humanistic geographers appreciated the possibilities of earnest, open-ended inquiry but looked toward conceptual perspectives like phenomenology and hermeneutics that respected the phenomena being studied and provided descriptive and interpretive methods whereby researchers could more accurately and comprehensively locate and understand those phenomena. Like social scientists, humanistic geographers were keen to use their knowledge to contribute to human betterment, but they emphasized that any practical plans or policies should be grounded in the experiences, needs, and wishes of affected parties rather than unilaterally dictated by outside governmental or corporate decisions and demands. In relation to environmental and ecological deterioration, humanistic geographers argued that, because humans are Earth’s most conscious and environmentally exploitive beings, their efforts at betterment must extend beyond the human world to protecting and strengthening the welfare of other sentient beings as well as ecosystems, places, landscapes, natural regions, and the planet as a living whole.

Key themes in humanistic geography

Broadly, one can identify four central conceptual and methodological themes relating to humanistic geography as it developed in the 1970s and 1980s.

1 Humanistic geographers understood human life and experience to be a dynamic, multivalent structure that incorporates bodily, sensory, emotional, attitudinal, cognitive, and transpersonal dimensions. Humanistic researchers argued that a comprehensive human geography must describe these many dimensions; understand what they contribute to environmental experience, action, and meaning; and seek out integrated frameworks identifying how these many dimensions relate and interact in supportive and undermining ways. For example, Edward Relph (1976) delineated a spectrum of spatial experience that ranged from the instinctive, bodily, and immediate to the cerebral, ideal, and intangible. He probed how the experience of space differs from the experience of place and contended that space becomes place when it gathers human meanings, actions, and identity environmentally and temporally. Similarly, Yi-Fu
Tuan (1974) delineated a conceptual structure of environmental attitudes and values by consolidating similarities and differences in the ways that human beings respond to their geographical worlds physiologically, psychologically, socially, and culturally. He concluded that every person is, simultaneously, a biological being, a social being, and a unique individual. He demonstrated how environmental perceptions, attitudes, and values arise from and contribute to all three aspects of human being.

Humanistic geographers emphasized that much of human experience is opaque, ineffable, or beyond taken-for-granted awareness. To identify and describe these less accessible aspects of human life, humanistic geographers largely turned away from conventional scientific method that required tangible, measurable phenomena explicated and correlated mathematically and statistically. Instead, humanistic geographers turned toward ontological perspectives that accepted a much wider range of experience and presence. They drew on epistemological perspectives that sought to be open to phenomena and to accept all aspects of their constitution. The aim was an empathetic, wider-ranging mode of discovery whereby the phenomenon was given time and space to present itself. The emphasis was on “methodologies of engagement” that allowed researchers to encounter and understand the worlds and experiences of their subjects carefully, accurately, and comprehensively. In working toward a more intimate encounter with the phenomenon under study, some humanistic geographers used directed intuition and self-reflective explication; others carefully studied real-world situations, for example, a specific urban neighborhood or a small number of individuals asked to describe their environmental experiences and actions as accurately and as thoroughly as possible.

Many humanistic geographers argued that, as much as possible, the evidence, general principles, and understandings of humanistic geography should arise from self-knowledge grounded in researchers’ firsthand experiences. Research should work toward a forthright engagement with the experiences of others, whether those “others” are people, places, landscapes, elements of nature, aspects of the human-made environment, or other sentient beings. Humanistic geographers called into question conventional empirical research that defined the topic of research in objectivist fashion as a thing or situation separate from and unrelated to the life or experience of the researcher. Humanistic geographers argued that, by understanding the significance of environmental and geographical experiences in their own lives, individuals might act more responsibly and generously toward other human beings and toward the places and environments that one inhabits or knows (Tuan 1976). In this regard, Edward Relph (1981) advocated for an environmental humility—a way of engaging with the world whereby things, places, landscapes, people, and other living beings are all respected just for being what they are and, therefore, are thoughtfully cared for and intentionally protected.

Broadly, humanistic geographers grounded their work in two complementary research models, the first of which can be identified as explications of experience; and the second, as interpretations of social worlds. Explications of experience were most often associated with “place studies” and represented
by such geographers as Anne Buttmer, Douglas C.D. Pocock, Edward Relph, David Seamon, and Yi-Fu Tuan. Much of this work was grounded in phenomenology and, for its place interpretations, drew on a wide range of descriptive sources that included first-person experience, philosophical argument, archival reports, accounts from imaginative literature, and experiential evidence from photography, film, and other artistic media. Typically, this work emphasized lived commonalities in relation to environmental and place experience, though these humanistic researchers also asked how those commonalities varied in terms of individual and group differences. In the 1980s and 1990s, this work would be criticized as *essentialist* – claiming generalizable, universal structures such as “place” and “home” and largely ignoring lived variations grounded in social, cultural, and historical factors (Cresswell 2013; see “criticisms” below).

The second research model for humanistic geography – interpretations of social worlds – was represented by the work of such geographers as James Duncan, David Ley, Marwyn Samuels, Susan Smith, Graham Rowles, and John Western. This work incorporated a wider range of philosophical traditions than experiential explication and included pragmatism, grounded theory, symbolic interactionism, poststructuralism, and Marxist perspectives. Typically, this research was grounded empirically in a specific place or social situation – for example, David Ley’s work on inner-city subcultures, housing, and gentrification; John Western’s documentation of the impact of apartheid on Cape Town, South Africa; or Graham Rowles’s research on the everyday environmental and place experiences of American elderly populations. These researchers interpreted place and related geographical phenomena as a “social construction” arising from purposeful actions of people-in-place. Place was interpreted as a negotiated reality via which people facilitated places, which in turn facilitated the lives of people associated with those places. In the 1980s and 1990s, this “social-constructionist” approach to place became one significant bridge to poststructuralist thinking and the “new cultural geography” (Adams, Hoelscher, and Till 2001; Cloke, Philo, and Sadler 1991).

Humanistic geography, 1970–1978

Though interest in humanistic geography still holds sway today, the most significant work was accomplished in the period 1970–1978. During this time, humanistic geographers produced important substantive research and explored broader conceptual and methodological concerns. Though humanistic research incorporated a wide range of philosophical traditions, phenomenology was most often used because it emphasized the elucidation of everyday human experience and could be readily applied to taken-for-granted geographic phenomena such as place, home, lived space, and environmental experience. The first explicit discussion of phenomenology and geography was a 1970 article in the *Canadian Geographer* by Edward Relph, who gave examples of how the phenomenological approach was appropriate for probing the relationships between human beings and their natural and fabricated environments. A year later in the same journal, Yi-Fu Tuan also considered the geographical value of phenomenology and concluded that the perspective was potentially helpful because it considered neither the world nor human beings in the abstract but, rather, emphasized “human-being-in-the-world” as it
incorporated environmental, geographical, and place aspects.

The next productive year in humanistic research was 1974, marked by four significant works. First, David Ley published *The Black Inner City as Frontier Outpost*, an ethnographic study examining an African American neighborhood in Philadelphia. Second, Anne Buttimer published *Values in Geography*, a work that considered how taken-for-granted personal and professional understandings and values ground scholarly knowledge, often in ways of which researchers were not self-consciously aware. Yi-Fu Tuan published two notable works in 1974, the first of which was an article in *Progress in Human Geography* in which Tuan described two different modes of place: public symbols, places of prominence, like New York City’s Time Square, that yield their meaning to the eye; and fields of care, places like a well-liked tavern or neighborhood only known through prolonged experience and typically undistinguished architecturally or visually. Tuan’s second significant work in 1974 was *Topophilia* (Tuan 1974), which delineated an outline for a phenomenology of environmental and place experience. This book became one of the best known humanistic-geographic works for researchers outside the discipline, partly because Tuan introduced the term “topophilia,” referring to attachment to and love of place.

The year 1976 marked a number of significant advances in humanistic research, including an explicit formulation of the subfield and two penetrating works that further clarified the relationship between humanistic geography and phenomenology. In a special June issue of the *Annals of the Association of American Geographers* devoted to the philosophy of geography, two important articles appeared, the first of which, by Anne Buttimer, examined how the phenomenological concept of *lifeworld* – the taken-for-granted world of everyday living – might offer insights for research on place, social space, and time–space rhythms. In the same *Annals* issue, Yi-Fu Tuan (1976) provided the first formal conceptualization of humanistic geography, which he described as a branch of the discipline that leads to a more thorough understanding of the human condition in relation to environmental and geographic concerns.

The most significant humanistic work in 1976 was Edward Relph’s *Place and Placelessness* (Relph 1976), a phenomenological study that interpreted place experience in terms of insideness and outsideness. Relph argued that the most intimate experience of place could be described by existential insideness, the lived situation in which a place is experienced and understood without self-conscious awareness yet is permeated with cognitive, sensory, and affective meaning usually unnoticed unless the place is changed in some way – for example, one’s home and neighborhood is destroyed by storm. Also in this work, Relph formulated the concept of *placelessness*, which he defined as the fragmentation and elimination of distinct places in the world. Of all the 1970s work in humanistic geography, *Place and Placelessness* perhaps had the most lasting impact because it provided a lucid, applicable presentation of why places are important in human life, what their constitution is experientially, and how they have been undermined in modernist and postmodernist times.

The year 1978 marked the high point of humanistic research in that David Ley and Marwyn Samuels (1978) published *Humanistic Geography: Prospects and Problems*, an edited collection illustrating the broad conceptual and thematic range that humanist perspectives could provide geography. In their introduction, the editors argued that the humanistic tradition was important for geographers because it offered one
conceptual and applied pathway for reconciling such dualisms as objectivity and subjectivity; materialism and idealism; agency and structure; and knowledge and wisdom. Chapters focused on such diverse topics as existential geography, alternative cartographies, a humanized economic geography, links between imaginative literature and geography, words for places, landscapes as experienced by tourists, and the phenomenological studies of the natural world produced by the eminent late eighteenth-century German poet and dramatist Johann Wolfgang von Goethe. In spite of their eclecticism, the 20 chapters of the volume effectively contributed to the editors’ main aim: to reconcile the “science and art of geography” (Ley and Samuels 1978, 10).

After 1978 and into the 2000s, important humanistic work continued to appear, including David Seamon’s *A Geography of the Lifeworld* (1979); Anne Buttimer and David Seamon’s *Human Experience of Space and Place* (1980); Douglas C.D. Pocock’s *Humanistic Geography and Literature* (1981); Edward Relph’s *Rational Landscapes and Humanistic Geography* (Relph 1981); Yi-Fu Tuan’s *Segmented Worlds and Self* (1982); David Seamon and Robert Mugerauer’s * Dwelling, Place and Environment* (1985); Edward Relph’s *The Modern Urban Landscape* (1987); J. Douglas Porteous’s *Planned to Death* (1989); J. Nicholas Entrikin’s *The Betweenness of Place* (1991); Robert David Sack’s *Place, Consumption and Modernity* (1992); Anne Buttimer’s * Geography and the Human Spirit* (1993); David Seamon’s *Dwelling, Seeing, and Designing* (1993); Paul Rodaway’s *Sensuous Geographies* (1994); Yi-Fu Tuan’s *Cosmos and Hearth* (1996); Robert David Sack’s *Homo Geographicus* (1997); David Seamon and Arthur Zajonc’s *Goethe’s Way of Science* (1998); Anne Buttimer’s *Sustainable Landscapes and Life-ways* (2001); Robert David Sack’s *A Geographical Guide to the Real and the Good* (2003); Edmunds Bunkse’s *Geography and the Art of Life* (2004); and Yi-Fu Tuan’s *Humanist Geography* (2012).

Most generally, however, the perspective of humanistic geography largely fell from sight or metamorphosed into the “new cultural geography” molded from poststructuralist, feminist, and critical perspectives. In this regard, many human geographers shifted their attention to the cutting-edge work of philosophers Michel Foucault, Jacques Derrida, Gilles Deleuze, Félix Guattari, Bruno Latour, and other poststructuralist, critical, and relationalist theorists (Cresswell 2013). One example of how humanistic themes shifted in the new millennium is *Textures of Place* (Adams, Hoelscher, and Till 2001), an edited collection dedicated to Yi-Fu Tuan and the humanistic tradition. Overall, the volume’s 27 chapters demonstrated how an engagement with critical social theory worked to transform earlier humanistic understandings of place, environmental experience, and geographical meaning. The editors of the volume called for a reconsideration of humanistic geography in the context of “revised assumptions about human subjectivity, the transparency of language, and the use of descriptive categories based upon Western traditions of understanding” (Adams, Hoelscher, and Till 2001, xvii).

**Criticisms of humanistic geography**

Beginning in the 1980s, humanistic research faced increasing criticism from quantitative-analytic geographers, on the one hand, and Marxist, feminist, and poststructural geographers, on the other hand (Cloke, Philo, and Sadler 1991; Cresswell 2013). Quantitative geographers largely criticized humanistic work in relation to research method: In turning away from deductive theory, predefined concepts, and measurable validation, how could humanistic
geographers be certain that their interpretive conclusions were accurate, comprehensive, and trustworthy? In response, humanistic geographers emphasized that their approach was generally inductive in that it drew on the richness and complexity of human situations and events to locate generalizable descriptions and theories. Humanistic geographers pointed out that the conclusions of any humanistic study were no more or no less than interpretive possibilities open to the public scrutiny of other interested parties. Humanistic geographers emphasized that their interpretive sources were wide-ranging and included field notes, focus groups, autobiographical descriptions, accounts from participant observation, and material texts like photographs, films, buildings, landscapes, imaginative literature, and archival documents. One methodological device used by humanistic geographers to better assure accuracy and trustworthiness was triangulation, whereby researchers drew on multiple modes of evidence-gathering methods to identify different lived perspectives and to corroborate different information sources.

The criticisms of feminist, Marxist, and post-structural geographers emphasized conceptual, ideological, and ethical concerns. Feminist geographers claimed that humanistic research was essentialist in uncritically assuming an unchanging, universal human condition that ignored individual and group diversity, including gender, social, cultural, and economic differences. These feminist geographers argued that humanistic work was authoritative in that it appeared to privilege the interpretive powers of scholarly experts who arbitrarily claimed the status to identify and describe the geographical situations of “more ordinary” people. Feminist critics contended that humanistic work presupposed an implicit masculinist bias that assumed academically trained men (mostly) could understand all others’ situations – for example, the experiences of women, the less able, gays and lesbians, ethnic and racial communities, and so forth. Marxist geographers criticized humanistic geography because they saw it as voluntarist in that it uncritically interpreted social life as a function of intentional, willed plans and actions of individuals. The Marxist claim was that humanistic thinking gave too much weight to autonomous human agency at the expense of entrenched, transparent social structures and power relations. Marxist critics pointed out that humanistic geographers gave little attention to the underlying economic and political dynamics shaping places and peoples’ everyday lives.

Humanistic geographers responded to the essentialist, authoritative, and masculinist charges by arguing that, in fact, humanistic work recognized human differences and sought conceptual and methodological ways for thoroughly engaging with the uniqueness of individuals and groups. They pointed to studies that used participant observation and other qualitative methods to understand particular geographical situations – for example, David Ley’s work on how African Americans negotiated their lives in the place context of Philadelphia’s inner city. In regard to the Marxist charge that they neglected the role of societal structures in constraining human freedom, humanistic geographers responded that their perspective could examine phenomena such as power, exclusion, resistance, and conflict, though little work was done in this direction, partly perhaps because most humanistic geographers instinctively favored freedom, creativity, and personal and group autonomy. Humanistic geographers accepted the Marxist claim that structural conditions are critical for the understanding of human action but, equally important, they argued, was the role of people’s values, beliefs, worldviews, intentions, and taken-for-granted ways of coping.
with the world. Humanistic geographers focusing on interpretations of social worlds probed the structural constraints of places and social worlds directly but gave equal weight to human agents being aware of and being able to change their lifeways in relation to limiting social and economic structures.

Poststructural geographers questioned humanistic work in yet other ways. Some poststructural critics claimed that humanistic geographers ethically favored place, insideness, and rootedness over non-place, outsideness, and mobility; place itself was assumed to be centered, static, bounded, and exclusionary. Instead, poststructural critics spoke of a “progressive sense of place” and focused on how places relate and respond to their wider social and environmental contexts. For these critics, places held their importance geographically, but the crucial theoretical and practical aim was finding ways whereby places could better incorporate diversity and partake in constructive interconnections and exchanges with other places. Another group of poststructural critics questioned whether “place” even existed in the postmodern world, claiming that real-world places were becoming marginal and obsolete because of trends toward globalization, non-places, and hyperspace. Some poststructural critics went so far as to suggest that, in our proliferating “hyper-real” world of digital environments and virtual realities, the lived distinctions between “real” and “imagined” places should be critically called into question. These critics challenged the rigid, unchanging stasis of physical places and environments that they claimed were encompassed by humanistic accounts. These critics spoke instead of provisional, shifting connections and flows among people, spaces, places, nation-states, information, worldviews, and digital representations. Key themes were mobility, flux, hybridity, relativity, relationality, discontinuities, rhizomes, assemblages, hyper-worlds, virtual places, and smooth and striated spaces.

Humanistic geographers responded to these poststructural criticisms by suggesting that, even as globalization eroded some places, it strengthened other places and contributed to new kinds of places. Humanistic geographers pointed out that, even with the growing importance of digital communication, hyperspace, and virtual realities, real places retain their importance because people are bodily beings who always unavoidably live a life in some physical place. This inescapable embodiment-in-place was often ignored by the poststructural critics who aimed for a more progressive sense of place grounded in a dynamic, ever-shifting network of intertwined, porous places. Humanistic geographers contended that a good portion of such dynamic exchange remains grounded in the habitual regularity of emplaced bodies. Humanistic geographers also emphasized that any dynamic interchange among places presupposes a robust integrity of each place itself; this robust integrity is at least partly founded in the habitual regularity of lived bodies inescapably bound to physical place (Seamon 2013).

Humanistic geography today

Though humanistic geography as an explicit subfield largely disappeared by the early 1990s, interest in humanistic themes continued inside and outside the discipline, particularly on the part of phenomenological philosophers concerned with the phenomenon of place. Humanistic geographers’ interpretations of place in the 1970s were largely subjectivist in that place was understood as a cognitive or affective representation inside the human being and ontologically separate from the objective environment outside. As phenomenological
philosophers Edward Casey (2009) and Jeff Malpas (1999) probed the topic in the 1990s and 2000s, they argued that place is a primary ontological structure that encompasses both human experience and the physical world in which that experience unfolds. This argument that a human being is always human-being-in-place highlighted an important new way of geographical thinking because it claimed that place is necessarily an integral, inescapable contributor to human existence and life. This understanding meant that places are not material environments existentially apart from the people associated with them but, rather, the holistic unit of human-beings-experiencing-place. Sometimes called *lived emplacement* or *embodied place*, this phenomenon was understood to be complex and dynamic, and to incorporate generative processes via which a place and its experiences and meanings shift or remain the same (Seamon 2013).

Partly because of Casey and Malpas’s writings, researchers inside and outside geography brought renewed scholarly attention to the lived qualities of place and to other topics associated with the humanistic tradition. For example, geographers Soren Larsen and Jay Johnson (2012) worked to link a place-grounded ontology with affinity politics, and geographer Sara Johansson (2013) developed a method of “rhythm analysis” to understand how the “lived body” encompasses and is encompassed by the urban environment as experienced. Echoing earlier claims on lived embodiment by French phenomenologist Maurice Merleau-Ponty, Johansson argued that the bodily dimensions of environmental experience are as meaningful and as important in understanding place as environmental cognition and intellectual geographic knowledge.

In research by non-geographers, one also finds a continuing body of work involving a humanistic approach to geographical and environmental topics. One example is philosopher Ingrid Stefanovic’s efforts (2000) toward a phenomenology of sustainability via an examination of how place and lived emplacement provide a foothold for grounding environmental responsibilities and actions in relation to particular individuals, groups, and localities. A second example is the research of literary scholar Anna Westerståhl Stenport (2004), who drew largely on Swedish writer August Strindberg’s works relating to Paris and Stockholm to examine how the nineteenth-century city shaped imaginative literature and how, in turn, that literature shaped perceptions of the nineteenth-century city. A third example is ethnographer Urzula Woźniak’s examination (2009) of at-homeness and placelessness in the context of current global migration. Drawing on Ukrainian, Turkish, and Vietnamese examples, she used the concept of community attachment to understand the contrasting degree of identification that different immigrant groups feel for their place of relocation; she demonstrated how mental associations with immigrants’ original home place play a significant role in their understanding of and feelings toward their new place of residence.

These studies and others exemplify a new generation of researchers who continue to be interested in such humanistic topics as place experience, at-homeness, community involvement and identity, out-of-placeness, environmental personhood, lived emplacement, mobility and place, supportive or undermining processes shaping place, and the lived similarities and differences between real places and virtual places (Seamon 2013). All of this work remains grounded in a central humanistic aim: to bring “human beings in all of their complexity to the centre-stage of human geography” (Cloke, Philo, and Sadler 1991, 58).
SEE ALSO: Bodies and embodiment; Cognition and spatial behavior; Emotional geographies; Feminist political ecology; Home; Marxist geography; Nature, art, and aesthetics; Phenomenology; Place; Space

References


Hungary: Magyar Földrajzi Társaság (MFT) (Hungarian Geographical Society)

Founded: 1872
Location of headquarters: Budapest
Website: www.foldrajzitarsasag.hu
Membership: 499 (as of December 1, 2014)
President: Gyula Gábris
Contact: info@foldrajzitarsasag.hu

Description and purpose

The main purpose of the Magyar Földrajzi Társaság (Hungarian Geographical Society) is to generate more interest in geography by promoting and disseminating geographical knowledge and to encourage research, especially on the geography of Hungary. The society also serves as a forum for geography teachers active at all levels of the educational system.

The society builds good relationships with nonprofit associations operating in the field of geosciences. MFT organizes national and international conferences, lectures, study tours, exhibitions, educational competitions, and family programs. MFT also publishes two journals and has a website to further knowledge and arouse interest in geography.

Journals or major publication series

Földrajzi Közlemények (Geographical Review).
http://www.foldrajzitarsasag.hu/kiadvanyok/foldrajzi-kozlemenyek
A Földgömb. www.afoldgomb.hu

Current activities or projects

Activities of the society includes co-organizing scientific conferences (e.g., http://eugeo2015.com) and lectures all around the country, while the annual meeting offers guided study tours in Hungary as well as abroad. The society is committed to assist talented pupils and university students interested in geography by offering them grants and co-organizing educational competitions nationwide. Journals published by the society target both scientific community and the general public. The society operates a library with the most significant geographical collection in Hungary.

Brief history

The MFT was founded in 1872 to promote the science of geography. Membership and interest grew, mainly because of the work of Lajos Lóczy, a determinative geographer who is credited with establishing modern geography in Hungary. The Földrajzi Közlemények, the scientific journal of the society, was first published in 1873. The Magyar Földrajzi Társaság Könyvtára book series had been published since 1880 and all 67 volumes became very popular due to its exclusive appearance and high quality scientific content.

Before World War II members of the society were accomplished researchers, some of them influential in politics, for example, Count Pál Teleki, who became prime minister between 1920–1921 and 1939–1941. This contributed to the abolishment of the society under communist rule between 1949 and 1952. The society gradually became more open towards geography education from 1952 onward when the membership mainly consisted of primary and
high school teachers and researchers. After the regime change and with the approaching twenty-first century and its challenges, the society renewed its efforts to recover its popularity and regain its scientific reputation.

Submitted by Ágnes Erőss
Hybridity

Clemens Driessen
Wageningen University, Netherlands

Hybridity in the social sciences is the idea that the world does not consist of pure things. Hybrid entities are considered to be combinations of two or more types of substances, mechanisms, processes, and so forth. In everyday language one can think of hybrid cars that have both an electrical and a combustion engine, or hybrid fast food such as the cheeseburger pizza. In (human) geography and other social sciences the term is used to break down dualisms that structure our (Western) thinking and research. A dualism is a combination of two terms that are thought of as opposed and which are used to categorize and order the world into mutually exclusive domains. Examples of these dualisms are subject/object, mind/body, social/natural, material/symbolic, and human/nonhuman.

Hybridity involves an ontological claim that the world really is a much more heterogeneous and messy place than many of our common binary “either/or” conceptual schemes imply. In its strong form, a hybrid world is not to be thought of in terms of separate objects, spheres, or domains that interact or interface. Instead the world is to be understood in terms of relations and processes only. Objects or subjects “in themselves” in this view have no essential qualities such as being powerful or weak, active or passive. Everything is part of assemblages, networks, or mesh-works whose characteristics emerge depending on the ways in which elements are (inter)related.

This intricately connected world can be characterized using hyphenated terms like “nature–cultures” and “social–technical” (the term sociotechnical in fact no longer even requires a hyphen). Processes of change in these intertwined domains are analyzed in terms of mutual constitution, co-production, co-shaping, and co-evolution: notions that emphasize dynamic interaction without privileging a one-way directed causality between separate entities. Ultimately, hybridity generates an interest in processes of becoming rather than states of being as the primary mode in which things exist; a view that has been informed by the work of philosophers such as Whitehead, Heidegger, and Deleuze.

Symmetrical descriptions of science and technology

The main opposition that is analyzed in terms of hybridity is that between the natural and the social. The French philosopher/anthropologist Bruno Latour is a key author associated with “actor–network theory” in which a particular view of hybridity has been developed. His ethno-graphic fieldwork in a scientific research laboratory and his studies of the development of new technologies resulted in empirical claims on how science actually operates and how new technologies come to be successful. His description of the ways in which truth is constructed in laboratories highlight how both material and social operations are necessary for making facts emerge and stick. In order for facts to be true and scientific findings to be real, Latour outlined a process of purification and translation: in a lab phenomena...
are isolated and purified, after which—in order to be found true—the conditions of the lab need to be translated and spread out over the world, connected with various types of humans and nonhumans alike. For example, the microbes that Pasteur isolated in his lab only got accepted as real after a variety of sites such as farms and hospitals were changed in such a way to resemble the conditions of the laboratory in crucial respects.

In Latour’s account, the ways in which facts come to be held as true, and technologies come to adequately function, cannot be meaningfully explained in terms of these being true representations of nature or inherently feasible technologies. Only in hindsight may technical factors such as the quality of scientific instruments and the purity of substances be distinguished from social forces such as economic factors, powerful institutions, and the interests of important individuals. But even then trying to keep these factors apart is not the most interesting way of understanding efforts to find the facts of nature or to invent new technologies.

What is needed instead are “symmetrical” accounts of scientific and technological change, which means a failed technology or disproven scientific theory should be explained and described in the same terms as successful ones: so without referring to a nature that sooner or later had to be discovered or to a technology that had always been viable but merely needed to be invented. Technological, or rather “sociotechnical,” innovation in this view requires “heterogeneous engineering,” a process that in particular ways manages to link very diverse things—material artifacts, economic structures, political institutions, organisms, and groups—so as to “align” these with each other. Alignment is the term used for the activity of generating networks of humans and nonhumans that together produce a certain effect.

Latour expanded his early work on laboratories and technology into a wider theory of hybridity, according to which our world is increasingly populated by hybrids. Environmental crises, genetically modified organisms, and climate change are issues in which political and social debates are hard to disentangle from scientific and technological uncertainties. Trying to nevertheless separate society and nature means missing out on the mixed, constructed, and intertwined character of these issues. By continuing to try to sort out political decisions from scientific uncertainties, we actually muddle things further and allow for the unacknowledged proliferation of hybrids. Latour claims the “modern constitution,” which strictly separates society from nature, and values from facts, has always kept from view how science and technology actually thrive by connecting and mixing these. Therefore he argues “we have never been modern” (Latour 1993) in the way that we think we have.

In symmetrical accounts of science and technology, the material and natural sides of modernist dualisms are actively co-shaping our world, rather than merely being a mute and static background to human actions. Agency, the ability to act and to influence events, then cannot be accorded to human beings alone. Actors can only become powerful when they come to be aligned within networks of various kinds of beings, including material objects. Therefore, in Latour’s “flat ontology” there is no (a priori) distinction to be made between humans and “nonhumans,” so we can discern “nonhuman agency” just as well as “human agency.”

Biologist/science historian/philosopher/anthropologist Donna Haraway offers a similar symmetrical and relational take on science and technology. She combines this with ideas springing from ecofeminism in which feminist understandings of social justice and alternative
forms of knowing are linked to a critique of environmental exploitation and domination. She calls for accounts of the world conceived as coming with active “subjectivity,” rather than merely a passive resource to be mapped and appropriated.

In *Simians, Cyborgs, and Women* (Haraway 1991) she proposes a further set of dualisms to be broken down, including male/female, mind/body, organic/inorganic, and subject/object. This feminist-inspired understanding of science emphasizes embodied and immersive “situated knowledges” rather than an abstract universalist ideal of science as inherently timeless and placeless. It would of course be too simplistic and dualist to merely map all these dualisms onto each other: projecting an essentially feminine relation to the world that is embodied rather than distant, closer to nature instead of abstracting from it, subjective rather than objective. Instead, this whole array of binary hierarchies are all to be muddled as part of an emancipatory process that extends to and beyond women, people of different ethnicities, and a variety of animals.

Haraway’s cyborg is both organism and machine, though without constituting a merging of these two elements as if they were pure ingredients. Both already are infused with (information) technologies, desires, and projections, making it pointless to try and fully disentangle natural origins from social and cultural influences. This model figure of the cyborg is thus not merely the Robocop-like man-machine found in popular culture, but embodies the way that all of us are nowadays mixtures of chemicals and flesh, hooked up to gadgets and functioning as part of various technological systems.

Without claiming or seeking a perspective outside of hybridizing practices and situations, this theorizing of hybridity is affirmative and seems to welcome the liberating potential of “posthuman” futures. This promotes the idea of us “never having been human” in the first place, that is, human as in independent from our environments and sociomaterial relations.

An argument was extended in later work where Haraway (2008) proposed the “companion animal” as another model subjectivity to embody the breakdown of distinctions that structure (ideas of) the self and its relation to the world. She recounts how in practices such as dog agility training human selves are actively transformed in close interaction with animal counterparts as part of cross-species learning processes.

### Hybrid geographies

Geography can be considered to be conducive to appreciating hybridity due to its spatial and situated focus on practices such as food consumption and production, nature conservation, and resource extraction that are inherently both social and natural.

After discarding binary divisions such as natural/social that conventionally guide research practices, hybrid geographers aim to trace complexities that cannot or should not be dealt with by sorting out the elements of belonging to either nature or culture. They describe movements of goods and information, focusing also on the material practices and spatially situated character of an issue to help establish interconnections and discern multifaceted processes. Thereby, a seemingly mundane artifact such as a cup of coffee can be found to be part of global flows of substances as well as concepts, involving regional climates, water scarcity and distribution, migrant labor exploitation, free trade negotiations and their legal as well as economic ramifications, historical coffee house culture and early experiences of democracy, effects of coffee consumption on human bodies, moods, sleeping patterns, and organizational social dynamics as well as processes of urban gentrification around American coffee
HYBRIDITY

chain shops competing with local translations of Italian espresso culture. Hybrid research aims to unravel and foreground the intertwine ment of these variegated aspects, offering alternative ways of conceptualizing and understanding relations of humans and their environments.

The hybrid geographical mappings of processes and movements lead to topological rather than geometrical modes of analysis. This means that in a world defined by relations rather than more universalizing abstractions, such as conventional maps, understandings of space can be geared to the issue at hand. This involves analyzing how disparate places become closely connected through flows generated by industrial production and transportation as well as ecological migration and geological processes. Natural processes in turn may be inextricably bound up with human practices, for example, climate change, erosion, and deforestation.

Understanding hybridity involves embracing everyday knowledge-practices that are embodied, lived, active, and place bound. This form of knowing is explicitly partial and incomplete and challenges the conventional ideal of knowledge as universal, purely cognitive, formalized, distanced, and abstract. Understanding the world as hybrid means studying it is also a performative activity that involves enacting a particular reality rather than merely trying to represent one. Research then is intervening in the midst of things, ideally a lively and affective affair in which researchers develop and promote a sensibility for experiencing the fluid and relational character of the world (see e.g., Bennett 2009).

Mostly, research into hybridity is done by foregrounding the relations between humans and nature through cases in which these relations are unsettled or complicated. Thus hybrid geographers such as Sarah Whatmore have followed global linkages and the ways in which things like genetically modified soybeans travel across the globe – constantly changing in the process as they move between patent offices, corporate headquarters, research laboratories, and the fields of farmers (Whatmore 2002). Other poststructuralist theorizing tends to focus on agency and meaning coming solely from humans and mainly expressed in discourse. But hybridity has led researchers to move beyond a strict focus on language and symbolic meaning only, to analyze also material things and concrete places as sources of meaning and change. Understanding soy then means studying discursive practices that produce soy across texts of various kinds, including legal and policy documents as well as interpreting imagery such as advertisements. But tracing the movements and transformations of soy also means analyzing field sites that offer insights into the ways in which nongovernmental organization (NGO) campaigns, supermarket executives, national politicians, and EU bureaucrats come to define and “enact” particular types of soy and their production practices.

The fields of food consumption and agricultural production lend themselves well to hybrid analysis, being inherently mixings of nature and culture. Hybrid views of agriculture complicate analysis in terms of naturalness versus artificiality. Instead they offer detailed accounts of the ways in which modern high-tech agricultural production generates particular relations and experiences.

Relations with animals and nature comprise another topic on which the implications of hybridity have been widely noted. When nature and culture are not considered as distinct realms, conservation can no longer be guided by ideals
of pristine and untrammeled nature. Attention is drawn instead to how the practices of wildlife conservation may involve intricate breeding programs and close monitoring of genetic variety of remaining specimens, whereby animals and their genetics are taken up and translated into information systems. In relation to this development, the variety of motives of conservationists and associated ideals of wildness and authenticity can be scrutinized together with assumptions in conservation biology such as the concept of species and models of population viability.

Science and lay knowledge

Work under the banner of hybridity can be viewed as a particular response to an ongoing debate since at least the 1980s on how to deal with an apparent crisis in the public role of scientific expertise. In public contestations of (technological) risks in the face of pervasive uncertainties, science is deemed unable to provide clear and unambiguous descriptions of what is the matter in an issue. Controversies over nuclear energy and radioactivity, HIV/AIDS drug development, ecological and health effects of pesticides, farm animal disease outbreaks, and the genetic modification of food crops led to a mainstream questioning of existing institutional relations between science and society.

The risks and uncertainties embedded in these issues were no longer accepted to be dealt with in terms of an “information deficit model” via one-way efforts to educate a lay public to understand and accept scientific findings. Not only were the sciences found to be incapable of providing clear answers in response to public concerns, but these concerns were actually produced or attributed to technological risks emerging from within these scientific disciplines themselves, such as biochemistry, nuclear physics, and molecular biology. Moreover, the techno-scientific products of these fields of research and development – such as the cancer research animal “Oncomouse®” which is genetically engineered to develop tumors – often unsettle the basic dualisms in which the public used to make sense of these concerns: natural/artificial, living/nonliving.

Working to expose and detail hybridity tends to come with an implicit or explicit aim to embrace particular forms of knowing and involvement: as situated, local, embodied, and immersive. This is the type of knowledge that can be intimately related to communal practices, everyday language, and local environmental stewardship. At the same time, analyzing in terms of hybridity means opening up the workings of science in practice, to reveal how universal truth claims are in particular ways local too and consist of an array of translations and purifications. Some natural scientists have felt attacked by these hybrid analyses which questioned their privileged access to the truth. In what in the 1990s was called the “science wars” they responded by dismissing the hybrid accounts of their day-to-day practices as merely the fashionable product of ignorant philosophers and social scientists.

However, science and lay knowledge can often be found to be mixed, for example, where one is harnessed to inform the other: when biodiversity prospecting involves using local guides; when pharmaceutical companies try and utilize indigenous knowledge of medicinal herbs handed down over generations; or with nomadic herders adding GPS technology to guide their herding practices.

Projects have been set up to do flood modeling whereby the complex mathematical/computer models of a river system are developed and calibrated by input of local “lay” publics who have more direct experience with flooding in a particular catchment area (Lane et al. 2011). In
HYBRIDITY

this way the modeling exercise can be understood as integrating dispersed and informal ways of knowing a river with more formal, disembodied computer modeling to generate knowledge that transcends the lay/expert dichotomy. Hereby the natural (and social) science work itself also provides a platform for conflict resolution and dealing with uncertainties and risks. In debating the impacts of particular mitigation strategies in one site, these may be found to be connected to effects elsewhere. This can be considered an example of a variety of “hybrid fora” – that have sprung up or were explicitly created as such – in which various forms of expertise come together to deliberate knowledge controversies.

Geography seems uniquely positioned to engage and experiment with these hybrid modes of knowledge production and combinations of physical and social forms of representing and intervening; though this has not meant the split between human and physical geography is easily overcome. Most geographical forms of expertise, subdisciplinary institutional divisions, and research practices still seem firmly rooted in separate social and physical domains. It would be too simple to see the challenge of hybridity for geography solely in terms of an opposition between social theory oriented toward critique of power relations and spatial analysis aiming for objective representation. The point of what has come to be labeled as “more-than-human geography” is not merely to combine qualitative and quantitative and/or social and natural science methods and approaches in a single research project. Nor does it mean that there can be only one proper way of doing geography: an idealized integration of multiple approaches into a single coherent unified science. The aim rather is to interweave the reductions and translations of natural science approaches with other forms of knowing and of deliberating matters of concern.

This of course is difficult, as it involves practical, institutional, and theoretical challenges.

Methods and approaches of hybridity

Ethnography is an important method for producing hybrid accounts. Taking a cue from anthropology, researchers do not seek to discern a social world separate from its understanding of nature, but to describe a society and its nature as part of a single coherent (or conflict ridden) cosmology – thus studying how knowledge, social relations, and symbolic and material cultures interrelate in a web of practical meanings.

Hybrid research involves a tension between, on the one hand, an emphasis on lived everyday experience and, on the other, being informed by poststructuralist ideas on the subject as an effect of discourses and practices. Experience in posthumanist thought, however, does not occur in isolated bodies, but these are always immersed, emergent, and relational – thus building on, for example, Heidegger’s understanding of “being in the world,” in which the environment is thought not as something to deal with “from the outside” as a space “out there” from which humans can be considered separate.

Also, the affective character of hybrid research may be in tension with the impetus to chart and describe traces and patterns across the globe. The status of the knowledge claims produced in terms of hybridity can be questioned: who is doing the tracing of flows and patterns? Emulating ethnographic work by anthropologists, researchers often render themselves present in the “thick” accounts of situations they offer; this in order to prevent the hybrid view of the world becoming another “view from nowhere” that merely discerns processes of mixing, purification, and translation everywhere. Authors thereby self-consciously reflect on the type of
HYBRIDITY

intervention that particular methods of analyzing and reporting bring, meanwhile negotiating the question of who it is that is self-conscious here: the author him- or herself (or itself?), as an assemblage of methods, theories, biography, and cultural identities? This offers a way not to hide but to actually make present the partialities and situatedness of hybrid geographers, who of course bring their own positionalities while traversing multiple sites and getting themselves entangled in associated webs of meaning and material practices.

For instance, ethnogeomorphology is an approach to environmental research that incorporates the sentient and cultural relations of local inhabitants of an area. Thereby it reflects multiple ways of relating to a landscape that is conceptualized as emergent (see Wilcock, Brierley, and Howitt 2013). It will be interesting to learn whether it is possible to move beyond aboriginal types of indigenous people that have been the main focus of this type of research to probe other perhaps heterogeneous groups that after all “have never been modern” too.

The experience of a dissolving nature/culture divide also seems to motivate and inform new experimental modes of living with (formerly wild) animals and ecologies. This includes acknowledging “invasive species,” appreciating “novel ecosystems,” and developing modes of conservation involving “rewilding” and “dedomestication,” which in practice do not refer to pure wild nature as conventional conservation does. For these and other sites of human–animal interactions, “multispecies ethnography” has been proposed as a way to symmetrically study both human and animal (or even plant) behaviors and interactions as part of more-than-human nature–cultures.

Also, new combinations of science and the arts are explored as a mode of generating forms of knowing, intervening, debating, or otherwise engaging with the world as both a physical and social place. Visual as well as performed arts and creative design are not just inherently material and symbolic, but can involve experimenting with forms in which to integrate these dualisms.

Politics and ethics in a hybrid world

As opposed to common theorizations of politics that leave establishing “the facts” to science, work starting from hybridity has proposed alternative modes of doing politics. Hybridity has been found to be a productive way to look at all types of societal issues. Since scientific facts do not have to be assumed as neutral or given, what may seem natural conditions and factors can also be described as riddled with societal choices and interests, identities, and communal meanings. Likewise, technological change can be redescribed to be as much “social” as it is “physical” in character, with particular ways of thinking and idiosyncratic choices hidden in material devices and practices.

Seeing science as a material and social process can be thought of as democratizing or at least opening up scientific research to scrutiny by more lay publics. However, hybrid geographies tend not to be easily translated into policymaking, as policymaking is mostly still (formally) operating within modernist dualisms separating facts and values, science and politics, human decisions, and environmental impacts.

Annemarie Mol (2002) has described various ways of enacting a particular disease, by doctors, patients, and others, emphasizing how the body and its disease are “multiple.” She highlights how the messiness of, for instance, medical practices is not something that should be solved by rationalization, or be used to criticize medical professionals or policymakers, but is a sign of complexity that we need to live with. She
HYBRIDITY

proposes an “ontological politics” that promotes doubt and the cultivation of multiple modes of improving practices rather than arguments for optimization in terms of a singular understanding of what is the matter. Likewise, Isabelle Stengers (2010) has proposed a “cosmopolitics” that emphasizes contingency and diversity in the way sciences create particular truths as they shape practices of intervening and debate. This is contrasted with conventional ideals of “decision-making” based on facts and technological choices that already make for particular (self-)understandings of political subjects and the nature of the situation they find themselves in.

The hybrid understanding of subjects and agency has of course implications for how to conceive of politics and ethics. No longer can we assume or idealize the liberal subject as a rational, autonomous, atomistic individual. This means traditional understandings of morality as personal ethical dilemmas and calls of conscience to redirect our will are misguided. At best we can see relational characters immersed in practices or, as Whatmore identifies them, “intimate assemblages of corporeal becoming.” Approaches to ethics that start from a relational and fluid understanding of humans, animals, and technologies call for learning to be open, to tune in, and to seek to encounter “mindful others.” For instance, in relation to animals this means not extending humanist models of subjectivity and rights to a select group of higher mammals, but becoming aware of the various ways in which our language and material practices contain species hierarchies, make us overestimate our human specialness, and prevent us from more truly encountering and imaginatively responding to a variety of living beings.

Hybridity entails moving away from oppositional thinking when confronted with “risks” and uncontrollable processes. Risk can then be appreciated as not just a technical challenge but also a social and cultural challenge of acknowledging and appreciating that human intentions are not the sole source of agency. For instance, new designs of living with river flooding, wildlife, and bacteria are being experimented with that acknowledge the agency of nonhumans such as animals and even water. Viewing these processes in terms of hybridity suggests that agency is distributed across many sites, not located in one place or body; that we stop trying to command and control, but allow for nonhuman agency to partly shape how we live.

Critiques of hybridity

Hybridity as an ontological position and research agenda has drawn critique, especially for giving up valued terms. For instance, if hybrid approaches see pristine wild nature as merely a romantic dream and embrace how we are surrounded by successions of novel ecosystems, urban wilds, and formerly invasive species, there seems to be no way to conceptualize what we should still try and protect because we cannot separate out a “natural” state. When domesticated farm animals have no natural equivalent that could function as benchmarks of behavior and welfare, there seems to be no holds barred to their adaptation and industrial use. When genetic modification is seen as merely an extension of interweavings of the organic and the mechanical and of nature and society associated with ordinary breeding, it seems difficult to argue against more meddling in a gene base to further (particular) human interests.

So, although hybrid approaches may have noble anti-authoritarian and democratizing aims, hybrid theorizing and its joyful dissolution of boundaries may inadvertently leave things to powerful actors who get to strengthen their take on what is left of nature.
One way of dealing with this loss of critical capacity would be to take hybridity much further and purge it from the “residual humanism” it tends to retain as commonly only humans are considered central in the muddling of culture and nature. Lulka (2009) has proposed to symmetrically grant this hybridizing ability to nonhuman agents, too. For instance, beaver burrows can be the site and the expression of consecutive generations of beaver communities responding to a flooding river in a process of co-evolution of groups of beavers with particular material cultures and abiotic processes in their environment. Perhaps a truly hybrid mode of knowledge production and environmental management would then be to actively involve beaver communities in participatory flood modeling and flood control experiments.

Beyond hybridity?

The idea of hybridity seems to become more and more mainstream. For example, formerly “natural events” such as floods are increasingly blamed not only on bad weather but also on governmental authorities, shifts in land-use patterns, and climate change. Or obesity is viewed as not just the result of consumer choices, lack of exercise, or genetic disposition, but also associated with health policies, the invention of high-fructose corn syrup, fast-food culture, and the spread of “food deserts” in which affordable healthy food is difficult to come by.

Nowadays it is for many (human) geographers and other social scientists evident that all agency, human or otherwise, is the outcome of particular interweavings or assemblages of things that formerly may have been thought of as belonging to either a natural or a social order. However, it is one thing to state that we live in a hybrid world and that subject/object distinctions should be disbanded. But everything, from the grammar of writing to common ideas of doing scientific research, still seems predicated on this distinction, making it a stubborn idea in our minds (and bodies).

To emphasize the hybrid character of our world a hyphenated vocabulary has emerged. But after more than a decade of research into social-material (or sociomaterial) nature-cultures, eventually this may be perpetuating obsolete dualisms. In practice, dualisms can become so muddled that they lose meaning as a distinction or even as a gradual difference. One could think of virtual/real as no longer helpful to describe how our lives have become embedded in information technology networks.

While hybridity is ostensibly meant to oppose thinking in terms of purity, the notion can be seen as ill-chosen, since it (historically) actually presupposes two purified organisms with homogenous characteristics. Hybrid breeding of plants and animals is predicated on creating “pure lines” of breeding stock which carry particular hereditary traits. These are then to be combined into a single hybrid animal that comes with “hybrid vigor.” Since these entities before they are combined into a hybrid are first made more pure, beyond what can be found in freely procreating and germinating animals and plants, hybridity may implicitly promote thinking in terms of pure ingredients, substances, domains, and so forth. In view of Latour’s (1993) warning against conceiving hybrids as a mixture of two pure forms, it can be argued that for furthering hybrid geographies it would eventually be better to leave the notion of hybridity behind.

**SEE ALSO:** Actor-network theory; Animal geographies; Construction of nature; Feminist political ecology; Modernity;
HYBRIDITY

Nonrepresentational theory; Ontology: theoretical perspectives; Posthumanism; Postmodernity; Poststructuralism/poststructural geographies; Social constructionism; Socio–nature; Technology

References


Further reading


Hydroclimatology and hydrometeorology

Paul Houser
George Mason University, USA

The water cycle describes the circulation of water, a vital and dynamic substance, in its liquid, solid, and vapor phases as it moves through the atmosphere, oceans, and land. Water is an integrating component of the climate-energy-geochemical cycles, regulating biological and ecological activities at all spatial and temporal scales. There is an important nexus between water and energy, this being the relationship between how much water is evaporated to generate and transmit energy, and how much energy it takes to collect, clean, move, store, and dispose of water. Life in its many forms exists because of water, and modern civilization must continuously learn how to live within the constraints and extremes imposed by water availability.

North America is among the greatest consumers of water worldwide: “Americans use water even more wastefully than oil. The U.S. relies on non-renewable groundwater for 50 percent of its daily use, and 36 states now face serious water shortages, some verging on crisis” (Barlow 2008). The High Plains aquifer and Central Valley aquifers are being depleted, there is oversubscription of surface waters based on outdated hydrographs, and a lack of reservoir, snow, and soil moisture data observation coordination and centralization prevents informed management. Looming climate change complicates North America’s water balance, but projected increases in water demand from increasing population, industrial, energy, and agricultural needs may have four times the impact on the water supply–demand imbalance that climate change does (Kummu et al. 2010).

Problems that are directly associated with water can be simply classified as too much water, too little water, and/or poor water quality. This seemingly simplistic classification is key for resolving current water-related issues and challenges. The fact is we do not even know how much water is stored in North America’s lakes, reservoirs, streams, groundwater systems, or snowpacks (Famiglietti 2012), which is fundamental knowledge needed to manage any resource.

Reliable prediction of hydrologic change and extremes is of critical importance for policy- and decision-makers to adapt to future water challenges (Schiffer and Unninayar 2011). However, our water availability, flooding, and drought models are not up to the task of addressing our most pressing societal issues of food, energy, water, and national security. State-of-the-art, comprehensive computer models are not able to seamlessly ingest satellite observations and measurements to help monitor and forecast snowpack, river flow, soil water, and groundwater levels (Famiglietti 2012).

There are important gaps in knowledge of where water is stored, where it is going, and how fast it is moving. Our skill in predicting the water cycle is woefully inadequate to reliably inform critical societal decision-making. Figure 1 shows our low precipitation forecast skill in comparison to temperature. We need a decisive and coordinated effort to systematically quantify water storages and fluxes and improve water cycle prediction skill, and reliable
methodologies to translate those predictions into enlightened water resource management.

Definitions

Meteorology is the science that deals with the state of the atmosphere with respect to wind, temperature, cloudiness, precipitation, moisture, pressure, and so on. Meteorology describes the processes involved in atmospheric phenomena, and as such provides a basis for weather forecasting. Climate is simply the averaged weather conditions over a relatively long period of time as exhibited by temperature, wind, and precipitation. It follows then that hydrometeorology and hydroclimatology are branches of meteorology and climate that specifically deal with the hydrologic cycle: the study of the occurrence, motion, and changes of state of atmospheric, terrestrial, and ocean water (Figure 2). Hydrometeorologic and hydroclimatologic studies address questions regarding land use, precipitation, river flow, groundwater, snowpack, floods, droughts, and the long-term effects of climate change on water resources.

Hydroclimatology was first defined by Langbein (1967) as the “study of the influence of climate upon the waters of the land.” The boundaries of hydrometeorology and hydroclimatology are not clear cut and overlap with many other scientific disciplines (such as climatology, oceanography, hydrology, meteorology, geology, geography, civil engineering). There is an important and practical connection between hydroclimatology and engineering design that uses relationships between meteorological variables, precipitation intensity, and duration, and the resulting runoff dynamics, as the basis for the design of flood-control and water-use structures. Determining rainfall probabilities, the statistics of storm occurrence, snowmelt dynamics, and ocean tides, surges, and rise are critical to inform and mitigate natural disasters and adapt to climate changes. Water supply, sanitation, and environmental quality are also of growing importance as populations increase and concentrate in urban centers, and agricultural/industrial activities intensify.
As a basis for this discussion of hydrometeorology and hydroclimatology, we must introduce the concept of hydrologic balance based on a series of linked mass conservation equations between the land, ocean, and atmosphere. The water balance equation for the terrestrial branch of the hydrological cycle can be written at a point as equation 1, where $S$ is terrestrial water storage term, $P$ is precipitation, $E$ is evaporation, and $R$ is runoff. $S$ can include changes in groundwater, canopy interception, soil moisture, surface water storage, and snowpack; $R$ can include both surface and subsurface horizontal water flow; $P$ can include liquid or solid precipitation, condensation, and irrigation; and $E$ can include evaporation, transpiration, and sublimation.

\[
\frac{\partial S}{\partial t} = P - E - R \tag{1}
\]

The water balance equation for the oceanic branch of the water cycle is given by equation 2, where $B$ is the ocean water mass (storage), and $F$ is the horizontal input or removal of fresh water.
HYDROCLIMATOLOGY AND HYDROMETEOROLOGY

\[
\frac{\partial B}{\partial t} = F + P - E \tag{2}
\]

The atmospheric branch of the hydrological cycle can be expressed in the form of a simple equation of vertically integrated terms, as in equation 3, where \( W \) is the water vapor content in the atmosphere, which is defined as equation 4, where \( q \) is the specific humidity, \( p \) is pressure, \( P_s \) is surface pressure, and \( g \) is constant. In equation 3, \( C \) is the vertically integrated moisture convergence flux, defined in equation 5, where \( V \) is wind speed.

\[
\frac{\partial W}{\partial t} = C + E - P \tag{3}
\]

\[
W = \frac{1}{g} \int_{0}^{P_s} q dp \tag{4}
\]

\[
C = -\nabla \cdot \left( \frac{1}{g} \int_{0}^{P_s} q V \right) \tag{5}
\]

When equations 1–5 are evaluated over a long time or large area basis, some smaller terms can be dropped through a scale analysis.

Components of the hydroclimate system

Water is an essential for life; it is a critical resource for humans and ecosystems. Due to the abundance of water, Earth as viewed from space looks like a blue marble. Despite this appearance, only a small percentage is fresh and available for ready use. This small percentage of fresh water is also unevenly distributed across the Earth, with large populations living in relatively dry places. Water withdrawals to support food and energy production are outstripping available supplies all around the globe, resulting in significant groundwater and surface storage depletions. To understand these disparities between water supply and demand, we must investigate the underlying physical processes that account for these spatial and temporal differences. A combination of both natural and anthropogenic processes collectively recognized as the hydrologic cycle provides the mechanism for the natural redistribution of water among the land, oceans, and atmosphere. The major components of the hydrological cycle are summarized below (adapted from Shuttleworth 2012).

Atmosphere

Water occurs in solid, liquid, and gas form at the range of temperatures and pressures found in the Earth’s atmosphere. Atmospheric water can be visible when it is in the form of clouds and precipitation, while water vapor is transparent and serves as a critical greenhouse gas. Water plays a crucial role in weather and climate because its phase transitions absorb or release huge amounts of energy, and the formation of clouds greatly increases the albedo of the planet. Clouds are among the most complex and difficult phenomena in the Earth’s system to predict, but they also teach us a great deal about the movement of the atmosphere. The atmospheric water cycle interacts with the land and ocean water cycles principally through evaporation and precipitation, but also can have profound impacts on surface temperatures through the presence of clouds and atmospheric circulations. The atmosphere also circulates and is responsible for the general transport of water from the oceans into the continents, and from the tropics toward the poles.

It is important to recognize that the amount of water vapor the atmosphere can hold is strongly dependent on its temperature. According to the Clausius–Clapeyron equation, the vapor pressure of water increases significantly more rapidly than the temperature of the air (approximately logarithmic). Therefore, warmer air can hold more water. The atmospheric relative humidity
is the amount of water vapor in the air, expressed as a percentage of the maximum amount that the air could hold at a given temperature. This explains a large number of things that we observe about the behavior of water in the atmosphere. The most fundamental is that when the atmosphere cools it can become oversaturated with water vapor, causing condensation, clouds, and precipitation. Another common phenomenon we observe in the winter is when we take air from outside that is cold and heat it up in our homes; this causes a large drop in relative humidity resulting in very dry indoor air conditions.

Next, it is important to understand what drives atmospheric temperature changes. Pressure in the atmosphere decreases with height according to the hydrostatic equation, which describes the weight of the overlying fluid. Therefore, when a parcel of air moves higher in the atmosphere its pressure must decrease, and according to the ideal gas law, its temperature also decreases causing it to have less ability to hold water vapor. So when a parcel of air rises over a mountain, is elevated on a column of hot air (i.e., convective thunderstorm), or when two weather fronts collide causing one to rise over the other, air temperatures can decrease causing relative humidity increases and cloud formation.

We can extend these concepts to understand the effect of climate change on the water cycle; a warmer atmosphere would hold more water, and would be expected to lead to a general intensification of water cycling. However, an intensified water cycle would have both positive feedbacks on climate (water vapor is an important greenhouse gas) and negative feedbacks through increased cloudiness. Cloud feedbacks are the greatest uncertainty in modeling the general circulation of the atmosphere.

**Figure 3** The distribution of the Earth’s water. Numbers are rounded, so percentage summations may not add to 100. Gleick (1993). Reproduced by permission of Oxford University Press, New York.

**Hydrosphere**

Here the hydrosphere is defined as the liquid water component of the Earth, including the oceans, seas, lakes, ponds, rivers, streams, and groundwater. The hydrosphere covers over 70% of the Earth’s surface and accounts for over 97% of the total volume of the Earth’s water. The vast majority of the hydrosphere is encompassed by the oceans, with only a fraction of a percent on the land (Figure 3).

The ocean plays a central role in the Earth’s water cycle, and is the principle regulator of both water mass and energy flows. The ocean is also the primary reservoir for much of the energy and carbon that drives the Earth’s climate system. The flow of mass and energy between the ocean and atmosphere expands the ocean’s impact on global weather and climate. The ocean is also impacted by the atmosphere and land through precipitation, evaporation, and runoff exchanges. The ocean is divided by the land into basins which regulate ocean currents and heat balance.
Ocean circulation is the large-scale movement of waters in and between the ocean basins. Winds and rotational forces drive surface circulation, and the cooling and sinking of higher-density water in the polar regions drives deep thermohaline circulation. Surface circulation carries the warm upper waters poleward from the tropics, which warms the atmosphere along the way. At the poles, the water is further cooled during winter, sinks to the deep ocean, and gradually rises virtually everywhere else in the ocean. The Earth’s rotation deflects ocean currents proportional to their speed, a force also at play in the atmosphere known as the Coriolis force. Ocean circulation changes can lead to variations in heat transport and weather patterns. For example, the El Niño Southern Oscillation (ENSO) is a periodic ocean circulation anomaly that has a profound impact on global weather patterns. The North Atlantic Oscillation and the Indian Ocean Dipole are other well-known ocean circulation variations. Knowledge of these oscillations helps us to better predict seasonal to interannual weather, and mitigate potential hardships.

Fundamental thermodynamics and climate models suggest that the water cycle should intensify in response to a warmer climate. A clearly anticipated result of an intensified water cycle would be increases in evaporation, which over the ocean would result in saltier ocean water surfaces. The trend for increasing evaporation has been confirmed from enhanced balance between precipitation and runoff (Milly and Dunne 2001; Walter et al. 2004), sea temperature warming (Curry, Dickson, and Yashayaev 2003), increased freshwater discharge (Syed et al. 2010), and saltier ocean surface water (Durack, Wijffels, and Matear 2012). A recent study of observed global ocean surface salinity changes shows evidence of an intensified global water cycle at a rate of $8 \pm 5\%$ per degree of surface warming (Durack, Wijffels, and Matear 2012). This rate is double the response projected by current-generation climate models and suggests that a substantial (16–24%) intensification of the global water cycle will occur in a future 2–3°C warmer world.

Lakes, rivers, and groundwater make up the balance of the hydrosphere (see Figure 3). These water bodies can have significant influence on regional weather and climate, by such processes as lake-effect snow or land-water temperature breezes.

**Cryosphere**

Water stored in solid form, in particular in the Antarctic and Greenland ice sheets, accounts for 2% of total water on Earth, and about two-thirds of fresh water. Frozen water is also stored in snowpacks, glaciers, permafrost, seasonally frozen ground, and sea ice, all known as the Earth’s cryosphere. Large populations and agricultural areas rely on melt runoff from mountain glaciers and snowfields for their water supply.

The cryosphere has a profound impact on climate because of its very high reflectance. Snow can change the albedo significantly, thereby reflecting much more radiation back into space. This is why the melting ice sheets and seasonal snow cover changes are of significant concern; the lower albedo of ice/snow-free areas captures more solar radiation and increases heating. Another profound impact of cryospheric change is sea level rise, which can amplify the impact of storm surge, tides, and coastal flooding. It should be noted that changes in sea ice do not have a sea level impact, because it already displaces ocean water.

The cryosphere generally interacts with the atmosphere through both precipitation and evaporation, and with the hydrosphere through runoff. Freezing ocean water often rejects salt,
thereby increasing ocean salinity, whereas melting ice usually results in ocean salinity decreases through additions of fresh water. Both of these salinity changes can impact ocean circulations by changing seawater density.

As temperatures rise, the possibility of precipitation falling as rain rather than snow increases, especially in fall and spring at the beginning and end of the snow season, and in areas where temperatures are near freezing. Such changes are already observed in many places, especially over land in middle and high latitudes of the Northern Hemisphere, leading to increased rains but reduced snowpacks (Mote 2003; Knowles, Dettinger, and Cayan 2006).

Lithosphere

The Earth’s lithosphere includes the crust and the uppermost mantle, which constitute the hard and rigid outer layer of the Earth. The lithosphere can interact with the water cycle in a number of ways, including water vapor release through volcanic activity, subduction through plate tectonic activity, or chemical reactions with the atmosphere, hydrosphere, and biosphere through soil-forming processes. The Earth’s crust and mantle contain a vast but unknown amount of water. This water takes two forms: “free” water in cracks and pores, and chemically bound water – hydrated minerals formed in the presence of water and that contain hydrogen and oxygen. The lithosphere also largely determines the topography of the land and ocean, thereby resulting in critical boundary conditions that govern the movement of ocean and atmosphere circulation.

Biosphere

The Earth’s biosphere is defined as the regions of the surface, atmosphere, and hydrosphere of the Earth occupied by living organisms, and what supports them, including soil, groundwater, water bodies, and the atmosphere. The biosphere generally extends from the ocean depths to about 6 km above sea level. The biosphere has a profound impact on the water cycle, because water is used, stored, and regulated by living things. Plants can take water from the soils, and animals can drink water from rivers and lakes or eat plants. Even microbes deep in the ground live in tiny films of water surrounding rocks. The water will then stay in the biosphere until released through evaporation, transpiration, excretion, decay, respiration, and combustion and the whole process begins again.

Plants in particular have a significant impact on the water cycle through transpiration. Transpiration is the process by which moisture is carried through plants from roots to small pores on the underside of leaves, called stomata, where it changes to vapor and is released to the atmosphere. Transpiration is essentially evaporation of water from plant leaves, but it can be much more efficient because the plant can access deeper storages of water via its roots. Many plants can actually access groundwater reservoirs, providing another connection between disparate parts of the water cycle. About 10% of the atmospheric moisture is a result of transpiration, which is remarkable considering that plants cover less than 30% of the Earth’s surface.

There is a common myth that states that water in the Earth’s hydrologic cycle is never created or destroyed, and that we have roughly the same amount of water as we did when the Earth was formed (plus the water supplied by intercepted comets). In reality water is being constantly destroyed and re-created, and one very common mechanism is plant photosynthesis which liberates oxygen with the photolysis of water. There are many other similar chemical reactions that
consume or re-create water from its hydrogen and oxygen components.

**Anthroposphere**

The anthroposphere is a relatively new term that refers to the part of the environment that is made or modified by humans for use in human activities and human habitats. The anthroposphere is the sphere of the Earth system where human activities constitute a significant source of change through the use and subsequent transformation of natural resources, as well as through the deposition of waste and emissions.

Humans now occupy 8% of the global land surface with dwellings and agriculture, and 75% of the Earth’s land surfaces have been altered by people in some way. Humans have made profound changes to the water cycle, from diverting entire rivers for use in agriculture, to draining wetlands, depleting aquifers, and creating large lakes and reservoirs, and human activities have even resulted in the loss of inland seas. Excessive groundwater extraction has led to significant land subsidence and regional climate changes. Deforestation, agriculture, irrigation, hydropower production, and urbanization all have profound impacts on water pathways and storage. On larger scales, the burning of fossil fuels and large-scale agriculture are leading to climate changes that are melting ice sheets and glaciers, reducing snowpack, changing precipitation patterns, increasing evaporation demands, and raising sea levels.

**Challenges**

Water is an integrating component of the meteorological and climatological water–energy–geochemical cycles, regulating biological and ecological activities at all spatial and temporal scales. It is well recognized that human demands from a growing population for food, energy, shelter, and fiber has impacted the landscape (Foley et al. 2005) and threatens biodiversity and water security (Vörösmarty et al. 2010). Anthropogenic climate change can exacerbate these impacts and threats (Karl et al. 2009).

In order to deal with climate, population, and environmental change and its uncertainty, we must extend the current basis of the science with modern observations, and models and decision tools, to provide guidance to water planners and engineers. A decisive and coordinated effort to systematically quantify water storages and fluxes, improve water cycle prediction skill, and develop reliable methodologies to translate those predictions into enlightened, informed water resource management is needed. By entraining, integrating, and coordinating the vast array of interdisciplinary observational and prediction resources available, we can significantly advance skill in assessing, predicting, and managing variability and changes in water resources to meet ever-increasing demands and climate change complexities.

Outlined here are three hydroclimatology challenges that must be addressed to meet the water cycle demands of the future. The first deals with developing a scientific basis and tools for adapting to changes in the water supply–demand balance. Adaptation refers to our ability to anticipate and adjust to changes in water supply and demand, to take advantage of opportunities, and to cope with the consequences. The second challenge is benchmarking: to use incomplete and uncertain observations to assess water storage and quality dynamics, and to characterize the information content of water cycle predictions in a way that allows for model improvement. The third challenge is to establish clear pathways to inform water managers, practitioners, and
decision-makers about newly developed tools, observations, and research results.

Challenge 1: Adaptation

It is clear that human activities are modifying the Earth’s environment – this is usually referred to as “environmental change.” Anthropogenic climate change from increased greenhouse gases in the atmosphere is one driver of this change, whose effects are directly related to changes in climate system, and in particular to those affecting water and energy cycling. These changes, and projection of future change, have been documented in the Intergovernmental Panel on Climate Change (IPCC) reports, resulting in calls to mitigate future changes by reducing the release of carbon-based greenhouse gases. It is unclear that sufficient reductions will take place to mitigate the projected changes to the water cycle, or that other drivers of change (population increases, land conversion, reservoir construction, river modifications, urbanization, irrigated agriculture, and so forth) will significantly impact the mean and variability of the water cycle. The end result is that any performance evaluation of water resource systems must adapt to a wide variety of current and projected water cycle changes.

Hydrological time series are no longer stationary due to water cycling influences from a variety of sources (Milly et al. 2008) – changes brought about due to human activities to secure clean and reliable water supplies for drinking water, irrigated agriculture, energy production, and manufacturing as well as to climate change. Engineering hydrology is concerned with developing water resource designs that alleviate the situation of too much water (floodings), too little water (drought), and coping with natural hydrologic variability (reliable water supplies) – basically, extremes. Hydrologists use long-term data on which to base water resources designs, implying that hydrologic design by its nature is “risk-based design” and needs statistical models to represent the data. Are the underlying statistical distributions stationary or nonstationary? If the latter, how are the moments changing, can they be “predicted,” and what are the implications for the reliability of current and future water structures?

Subject to current and projected environmental, population, and climate changes, there is evidence that hydrologic and water cycle time series are becoming nonstationary. This is a significant challenge to the hydrology and water resources scientific and management community, and it is important to determine the best methods to model nonstationary processes and to develop procedures for incorporating nonstationarity into hydrological and water resources designs. The development of these new methods and tools needed for water resource systems to adapt to climate and environmental change in a nonstationary world is referred to as hydrologic adaptation science. Developing this science is a major challenge for hydrology.

Actions needed to develop a scientific basis and tools to adapt to climate and environmental change include not only addressing nonstationarity, but also establishing the scientific basis of water sustainability. This will require enhanced investments in hydroclimate process science, land change science, precipitation prediction, hydrologic ensemble generation, model building and calibration, Earth system model development, advancing land–atmosphere coupled models, including human dimensions in models, and developing risk-based uncertainty metrics.

In the water resources context, adaptation also involves achieving a balance between water supply and demand and easing water quality issues, thereby making more water available for human and environmental use. Improvements
HYDROCLIMATOLOGY AND HYDROMETEOROLOGY

in water supply or availability could take many forms, as in the following examples.

- Enhancing groundwater recharge by slowing runoff, using pervious paving methods, and so on.
- Reducing evaporation through proper forest management, improved irrigation practices, reductions in open water area, canal coverings, or land-use optimization. Reduced evaporation may also improve water quality.
- Rainwater harvesting from rooftops.
- Water treatment and reuse.
- Weather modification.
- Enhanced below- and aboveground water storage.
- Water conveyance from wet to dry areas.
- Desalination and water treatment science and technology.
- Alternative energy development that does not require thermal water cooling, such as wind, solar, and air-cooling.
- High-efficiency appliances, toilets, showers, and so on.
- Drought-tolerant landscaping and agriculture.
- Water pricing and economic incentives.
- Improved water distribution system efficiency (reduced pipe leaks and canal losses).

Challenge 2: Benchmarking

Benchmarking is the process of improving performance by identifying, understanding, adapting, and implementing best practices and processes. Benchmarking involves the creation of partnerships to exchange information on processes and measurements, bridging observational and prediction gaps, and setting realistic water cycle knowledge and prediction improvement goals. Effective benchmarking is a process framework within which indicators, best practices, and effective environmental management are continuously improved.

Models provide a key resource in our understanding of weather and climate, allowing/providing predictions that are used by national sectors on seasonal to climate timescales. Numerical simulation and prediction of water, in all its phases, presents challenges at many time and space scales. Improving model capabilities relies on consistent high-quality benchmark data and methods.

While there is a great need for analyzed precipitation to close water and energy budgets, it is a difficult parameter to predict, even on short timescales. Surface evaporation is not directly observed and relies on theory and calculations. There exist many similar gaps in our observational knowledge of the water cycle over North America. Higher spatial and temporal field studies can provide much-needed independent observations of these critical parameters, thereby exposing weaknesses in the numerical approaches that can be corrected to improve predictions. The current observing network has the capacity to provide certain measures of meteorology and climate, but the current lack of availability, integration, and quality control between disparate observational networks limits their usefulness for model benchmark investigations.

Single observations, even with uncertainty values, are often not a reliable benchmark. Multiple estimates from varying approaches can provide a range and better sense of the uncertainty for a parameter. However, it is unrealistic for a single organization to accurately develop multiple approaches to measuring all the water quantities needed, given the specialized knowledge required for each. As such, a community collaborative effort would be the most reliable
approach to gain success in reducing the uncertainties of the current analyzed regional water cycles and their prediction.

Actions needed to understand hydroclimate sensitivity and benchmark prediction models and decision tools include reanalyzing hydroclimate change and documenting its uncertainty, conducting field verification across regions and hydroclimate gradients, leading field campaigns focused on regions of high prediction uncertainty, developing prediction and operational support capabilities, and establishing water and energy budget closure methods across scales.

Challenge 3: Science informing decisions

Water cycle variability and extremes affect all aspects of society and the environment. Decisions are made every day to help communities, farmers, industries, and the environment effectively minimize (mitigate) and prepare for (adapt to) changing water availability and demand. The water science community has created a strong scientific foundation for informing decision-makers who need science to understand and envision a range of potential water-sector impacts, risks, vulnerabilities, opportunities, and tradeoffs. The challenge is to emphasize the decision-maker focus to better inform decisions — a focus that will conduct fundamental, user-inspired research while delivering credible, relevant, timely, and accessible information.

Central to the success of better-informing water decisions is strengthening the dialogue and engagement between the science and decision-making communities. This collaboration and coordination at the interface of science and decision-making requires new methods and a framework for multidirectional information exchange, including the following.

• Facilitating meaningful partnerships between science and decision-making to assess decision needs, science capabilities, and requirements; identifying knowledge gaps and inspiring a use-focused research agenda; and establishing sustained dialogues including development of communities of practice, training and outreach forums, and technology transfer.

• Providing easy access to science knowledge and operational practices through integrated water information, decision tools, and operational status; and developing simulators, libraries, and forums to make research available to the larger community.

• Guiding and coordinating water science efforts to ensure they are relevant to informing water management decisions; benchmarking uncertainties of the observational record, model predictions, and operational needs; developing new methods to transfer climate predictions into decision space (downscaling, application sector modeling); developing operational observation capacity to validate and improve operational products; and identifying and transitioning research products into operational applications.

In developing a strategy for informed decisions, it is recognized that scientific knowledge is only one part of a much broader decision process. For instance, information may be scientifically relevant without being decision relevant. Therefore, we must define a framework for informing decisions that connects to the broader decision process. We must develop processes that enable better “science-based” information that moves water management decision processes forward while also developing and transitioning science solutions. The desired framework will include approaches to assess the value of proposed decision support information, provide support for
understanding risk management options, and communicate uncertainties associated with data and projections. It is envisioned that science can inform decision-making, along with other critical information, in an adaptive management framework.

SEE ALSO: Climatology; Earth system science; Hydrologic cycle

References


Hydrologic cycle

Taikan Oki
Hyungjun Kim
The University of Tokyo, Japan

Water on Earth

The Earth system is unique in that water exists in all three phases, that is, water vapor, liquid water, and solid ice, compared to the situations on other planets. All organisms on the Earth, including humans, require water for their survival. The total volume of water on the Earth does not change on shorter than geological timescales (Oki 1999); it is estimated as approximately $1.4 \times 10^{18} \text{ m}^3$, which corresponds to a mass of $1.4 \times 10^{21} \text{ kg}$ (Figure 1).

The proportion of water in the ocean is large (96.5%). The global hydrologic cycle cannot be considered without including oceanic circulation. The ocean circulations carry large amounts of energy and water. The surface currents of oceans are driven by surface wind stress. The atmosphere itself is sensitive to the surface temperature of the sea. Temperature and salinity determine the density of ocean water, and both factors contribute to the turnover and general deep ocean circulation.

Other major reserves are the solid waters on land masses (glaciers and permanent snow cover) and groundwater. Glaciers are the accumulation of ice originating in the atmosphere, and they generally move slowly on land over a long period. Glaciers form discriminative U-shaped valleys over land and leave moraine deposits when they retreat. If a glacier “flows” into an ocean, its terminated end often forms an iceberg. Glaciers react over a comparatively longer timescale climatic change, and they also induce isostatic responses of continental-scale upheavals or subsidence over even longer timescales. Even though it is predicted that the thermal expansion of oceanic water will dominate the anticipated sea level rise due to global warming, glaciers over land are also a major concern as a cause of sea level rise associated with global warming.

Groundwater is the subsurface water occupying the saturated zone. It contributes to runoff in the low-flow regime between storm events, that is, during dry spells. Deep groundwater may also reflect the long-term climatological situation. The groundwater quantity shown in Figure 1 includes both gravitational and capillary water, but groundwater in the Antarctica (roughly estimated as $2 \times 10^6 \text{ km}^3$) is excluded. Gravitational water is the water in the unsaturated zone (vadose zone) that moves under the influence of gravity. Capillary water is water found in the soil above the water table by capillary diffusion, a phenomenon associated with the surface tension of water in soils acting as porous media. In terms of groundwater recharge, Döll and Fiedler (2008) estimated the global recharge flux to be $12,666 \text{ km}^3 \text{ year}^{-1}$ and approximately $1000 \text{ mm year}^{-1}$ in the Amazon region. They assumed the recharge flux is a fraction of total runoff. Koirala (2010) estimated groundwater recharge flux by coupling a land surface scheme with a macroscale groundwater representation. The global distribution of model-simulated groundwater recharge is illustrated in Figure 2a. Total groundwater recharge flux is estimated as $31,789 \text{ km}^3 \text{ year}^{-1}$ and the value is close to
The terrestrial water balance does not include Antarctica.

Figure 1 Global hydrological fluxes (1000 km³ year⁻¹) and storages (1000 km³) with natural and anthropogenic cycles are synthesized from various sources. Vertical arrows show annual precipitation and evapotranspiration over ocean and land with major landscapes (1000 km³ year⁻¹). Parentheses indicate the area (million km²). Modified from Oki and Kanae (2006).

the flux of subsurface runoff in Figure 1 (30,200 km³ year⁻¹).

Soil moisture is the water being held above the groundwater table. It influences the energy balance at the land surface in the way that a lack of available moisture suppresses evapotranspiration (which consists of soil evaporation, plant transpiration, and interception loss), and changes surface albedo. Soil moisture also alters the fraction of precipitation partitioned into direct runoff and infiltration. The precipitation water becoming direct runoff cannot be evaporated from the same place, while the water infiltrating into soil may be taken up by hydraulic suction and evaporated back into the atmosphere. The global distribution of model-estimated mean soil wetness index is shown in Figure 2b. Generally, the distribution is correlated with precipitation distribution (Figure 2c) as well as with runoff distribution (Figure 2d), but the global distribution of river discharge (Figure 2e) accumulates the total runoff generated in upper watersheds, and the shape of river channels can be seen in the distribution.

The atmosphere carries water vapor, which influences the heat budget via latent heat exchange processes. Condensation of water vapor releases latent heat, which warms the atmosphere and affects the general atmospheric
circulation. Liquid water in the atmosphere, such as cloud droplets and raindrops, is another result of condensation. Clouds significantly change the radiation in the atmosphere and at the Earth’s surface. However, the volume of liquid (and solid) water contained in the atmosphere is relatively small: most of the water in the atmosphere exists as water vapor. Water vapor is also the major absorber of both shortwave and longwave radiation in the atmosphere. Precipitable water is the total water vapor in the atmospheric column from the land surface to the top of the atmosphere.

The amount of water stored in rivers is rather tiny compared to other reserves at any time. However, the recycling speed of river water (river discharge), which can be estimated as the inverse of the mean residence time, is relatively high, which is important because most societal applications ultimately depend on river water as a renewable and sustainable resource.

The amount of water stored transiently in a soil layer, in the atmosphere, and in the river channels is relatively minute, and the time spent moving through these subsystems is relatively short. However, they play a dominant role in the global hydrological cycle.

The water cycle plays many important roles in the climate system. The various flow paths of water in the global hydrologic system are
HYDROLOGIC CYCLE

illustrated schematically in Figure 1 (Oki and Kanae 2006).

Fluxes

Precipitation is the water flux from atmosphere to land or ocean surface. It drives the hydrological cycle over the land surface; it also changes surface salinity (and temperature) over the ocean and affects its thermohaline circulation. Rainfall refers to the liquid phase of precipitation. Part of it is intercepted by the canopy over the vegetated area; the remaining part reaches the Earth surface as throughfall. The highly variable, intermittent, and concentrated behavior of precipitation in the time and space domain compared to other major hydrological fluxes mentioned below makes the observation of this quantity and the aggregation of the process complex and difficult. Global distribution of precipitation is presented in Figure 2c. Currently, satellite-based estimates merged with in situ observational data have been produced and are available to the public.

Snow has special characteristics compared to rainfall. When snow accumulates, the surface temperature will not rise above 0°C until the completion of snow melt. The albedo of new snow can be as high as cloud albedo and, being aged (dithered by dust), varies between 0.6 and 0.9. Consequently, the existence of snow changes the surface energy and water budget enormously. A snow surface typically reduces the aerodynamic roughness, so that it may also have a dynamical effect on the atmospheric circulation and hydrologic cycle.

Evaporation is the return flow of water from the surface to the atmosphere and the latent heat flux from the surface. The amount of evaporation is determined by both atmospheric and hydrological conditions. From an atmospheric point of view, the partition of incoming solar energy to the surface between latent and sensible heat flux is important. Wetness at the surface influences this partition significantly because the ratio of actual evapotranspiration to the potential evaporation is reduced due to drying stress. The stress is sometimes formulated as a resistance, under which evaporation is classified as hydrology-driven (soil-controlled). If the land surface is wet enough compared to the available energy for evaporation, the condition is classified as radiation-driven (atmosphere-controlled).

Transpiration is the evaporation of water through the stomata of leaves. It has two special characteristics different from evaporation from soil surfaces. One is that the resistance of stomata is related not only to the dryness of the soil but also to the physiological conditions of vegetation through the opening and closing of stomata. Another is that roots can transfer water from deeper soil layers than is the case of evaporation from bare soil. Vegetation also modifies the surface energy and water balance by altering surface albedo and by intercepting some precipitation and evaporating this rainwater. The global distribution of total evapotranspiration is shown in Figure 2f, estimated by a land surface model.

Runoff returns water from the land to the ocean; it may otherwise be transported in vapor phase by evaporation and atmospheric advection. The runoff into the ocean is also important for the freshwater balance and the salinity of the ocean. Rivers carry not only water mass but also sediment, chemicals, and various nutrient matter from continents to seas. Without rivers, global hydrologic cycles on the Earth are not closed. Runoff at a hillslope scale is a nonlinear and complex process. Surface runoff can be generated when the intensity of rainfall or snow melt exceeds the infiltration rate into the soil, or when precipitation falls on a saturated land surface. Saturation at the land surface can be
formed mostly by a topographic concentration mechanism along the hillslopes. Infiltrated water in the upper part of the hillslope flows down the slope and discharges at the bottom of the hillslope. Because of the highly variable heterogeneity of topography, soil properties (such as hydraulic conductivity and porosity), and precipitation, basic equations such as Richards’s equation, which can be fairly valid at a point scale or hillslope scale, cannot be directly applied on the macroscale because of the nonlinearity involved. The global distributions of runoff and river discharge are illustrated in Figure 2d and Figure 2e, respectively.

**Water balance**

In the field of hydrology, river basins have commonly been selected for study, and water balance has been estimated using ground observations, such as precipitation, runoff, and storage in lakes and/or groundwater. The water balance over the land is described as:

$$\frac{\partial S}{\partial t} = P - E - R_0 - R_u$$  \hspace{1cm} (1)

where $S$ represents the water storage within the area, $t$ is time, $(\partial S/\partial t)$ is the change of total water storage with time, $P$ is precipitation, $E$ is evapotranspiration, $R_0$ is surface runoff, and $R_u$ is groundwater movement (all fluxes above are given in the unit volume per time step). $S$ includes snow accumulation in addition to soil moisture, groundwater, and surface water storage, including retention water within the control volume. The control volume is defined by the area of interest over the land, with its base generally at the impermeable bedrock. These terms are shown in Figure 3a (Oki 1999). Equation 1 means water storage over land is increased by precipitation and decreased by evapotranspiration, surface runoff, and groundwater movement.

If the considered area of water balance is set within an arbitrary boundary, $R_0$ represents the net outflow of water from this area (i.e., the total outflow minus total inflow from surrounding areas). Although, in general, it is not easy to estimate groundwater movement $R_u$, the net flux per unit area within a large area is expected to be comparatively small. If all groundwater movement is considered to be that

![Figure 3](image-url)  
(a) Terrestrial water balance, (b) atmospheric water balance, and (c) combined atmosphere–land surface water balance corresponding to equations 1, 2, and 3, respectively (Oki, Kanae, and Musiake 1996).
HYDROLOGIC CYCLE

observed at the gauging point of a river \((R_u = 0)\), then equation 1 becomes:

\[
\frac{\partial S}{\partial t} = P - E - R_o \quad (2)
\]

This assumption is generally valid at the outlet of a catchment. In most cases, surface runoff \(R_o\) becomes river discharge through the river channel network. River discharge is an integrated quantity over the whole catchment and can be observed at a downstream point, in contrast to other fluxes, such as \(P\) and \(E\), which have to be spatially measured.

Atmospheric water vapor flux convergence contains water balance information that can complement the traditional hydrological elements such as precipitation, evapotranspiration, and discharge. The basic concepts, as well as the application of atmospheric data to estimate terrestrial water balance, were first presented by Starr and Peixoto (1958). The atmospheric water balance for a column of atmosphere from the bottom at the land surface to the top of the atmosphere is described by the equation:

\[
\frac{\partial W}{\partial t} = Q + (E - P) \quad (3)
\]

where \(W\) represents the precipitable water (i.e., the column storage of water vapor), \(Q\) is the convergence of water vapor flux in the atmosphere (all fluxes given in the unit volume per time-step). Since the atmospheric water content in both solid and liquid phases is generally small, only the water vapor is considered in equation 3. The balance is schematically illustrated in Figure 3b, which indicates that water storage in an atmospheric column is increased by the lateral convergence of water vapor and evapotranspiration through the bottom of the column (i.e., land surface) and decreases by the precipitation falling out from the bottom of the atmosphere column to the land.

Since there are common terms in equation 2 and equation 3, they can be combined:

\[
-Q + \frac{\partial W}{\partial t} + Q = (P - E) = \frac{\partial S}{\partial t} + R_o \quad (4)
\]

Figure 3c illustrates the balance in this equation, and that the difference between precipitation and evapotranspiration is equal to the sum of the decrease in atmospheric water vapor storage and lateral (horizontal) convergence, and also to the sum of the increase in water storage over the land and runoff. Theoretically, equation 4 can be applied for any control volume of land area combined with the atmosphere above, even though the practical applicability depends on the accuracy and availability of atmospheric and hydrologic information.

Further assumptions are often employed in annual water balance computations: the annual change of atmospheric water vapor storage is negligible \((\partial W/\partial t) = 0)\) and the annual change of water storage of the land is negligible \((\partial S/\partial t) = 0)\). With these assumptions, equation 4 simplifies to:

\[
Q = (P - E) = R_o \quad (5)
\]

If a river basin is considered as the water balance region, \(R_o\) is simply the discharge from the basin. The simplified equation 5 demands that the water vapor convergence, “precipitation minus evaporation,” and net runoff should balance over the annual period when the temporal change of all storage terms can be neglected.

Generally, it is not an easy task to obtain large-scale evapotranspiration \(E\) based on observations except for the annual timescale, where \(E\) can be estimated as the residual of \(P\) and \(R_o\). However, the combined water balance can help estimate \(E\) at a shorter timescale, for example, monthly. Note that equation 3 can be rewritten as:
which can be applicable over a period shorter than a year, unlike the assumption in equation 5. If atmospheric and precipitation data are available over a short timescale, such as a month or a day, evapotranspiration can be estimated at the corresponding timescales, but it is also subject to severe limitations imposed by the data accuracy. The region over which evapotranspiration is estimated is not limited to a river basin; rather it depends on the scale and the associated accuracy of the available atmospheric and precipitation data.

Total terrestrial water storage (TWS), as the sum of surface water (such as river water, snow water, and water in lakes), soil moisture, and groundwater, is generally difficult to estimate on the global scale. Combining equation 2 and equation 3 yields:

$$\frac{\partial S}{\partial t} = -\frac{\partial W}{\partial t} + Q - Ro$$  \hspace{1cm} (7)$$

which indicates that the change in water storage in the control volume over the land can, in principle, be estimated from the atmospheric and runoff data. Although an initial value is required to obtain the absolute value of storage, the atmospheric water balance can be useful in estimating the seasonal change in total water storage in large river basins.

The values quoted in Figure 1 were estimated based on various observations, with some assumptions in order to obtain global perspectives. These values are sometimes different in other references, probably because the source of observed data, methodology to estimate, and assumptions are different. In some cases, global water balances are estimated using the empirical relationship of evapotranspiration to precipitation in each latitude belt (Baumgartner and Reichel 1975).

Global distribution of hydrological quantities

A 35-year (1979–2013) land surface simulation (Kim and Oki 2014) was conducted to estimate global water and energy balances as a pilot study of an international research project. The physical quantities in this entry are mainly based on this simulation, and the project is the third phase of the global soil wetness project (GSPF). The original goal in the precedent phase (Dirmeyer et al. 2006) was to estimate global distribution of surface soil moisture, which is not easy to observe globally but is relevant for an understanding of the land–atmosphere interactions using multimodel ensemble simulations. In the third phase, a comprehensive global dataset of land surface fluxes, state variables, and related hydrologic quantities will be produced. It will extend the original science investigating long-term (1850–2100) interactions between energy, water, and carbon cycles over 0.5° terrestrial grids, excluding Antarctica. Three-hourly meteorological forcing data will be prepared as a hybrid product of a dynamically downscaled atmospheric reanalysis product and various observational data, based on in situ and satellite monitoring, to achieve further realistic representation with reduced bias in higher spatiotemporal resolution.

The advantage of using models to estimate global water balance is the capability to have more detailed insights than using observations. For example, snow over the land, excluding ice and glacier areas, is approximately 10% of total precipitation, and the ratio of surface runoff and subsurface runoff is approximately 1:2 in Figure 1. Even though, at present, it is difficult to assess the validity of these breakdowns due to the lack of global observations on the partitions between snow and rain or between surface and subsurface runoff, such model-based
HYDROLOGIC CYCLE

estimates will stimulate scientific interest to collect and compile global information on these important hydrological quantities in the future.

Kim et al. (2009) underscored the importance of the river component in TWS variation over global river basins. To reduce simulation uncertainty, ensemble simulations were performed with multiple precipitation data and a localized Bayesian model averaging technique was applied to the global river runoff simulation using a river routine model named TRIP (total runoff integrating pathways) (Oki and Sud 1998).

It was found that river storage not only explains different portions of total TWS variations, but also plays different roles in different climatic regions. It is the most dominant water storage component in wet basins (e.g., the Amazon) in terms of amplitude, and it acts as a “buffer” that smooths the seasonal variation of total TWS especially in snow-dominated basins (e.g., the Amur). It signifies that model simulation of TWS may not be able to properly reproduce the amplitude and seasonal pattern of observed TWS variation by GRACE (the gravity recovery and climate experiment) (Tapley et al. 2004) without an appropriate representation of the river storage component. The dominant role of river storage was already indicated in a pilot study that compared total TWS changes estimated by the atmospheric water balance method and a general circulation model (GCM) simulation coupled with TRIP in the Amazon River basin (Oki, Kanae, and, Musiake 1996). However, the message was not convincing until several years ago when satellite-observed GRACE data became available. Using a geodesy approach, Han et al. (2009) employed a fixed velocity version of TRIP in the Amazon River basin and its vicinity, and compared the model simulations to the residual of GRACE raw measurements derived from removing all the gravity-influencing factors except for the horizontally moving water. They demonstrated that the optimal flow velocity of TRIP in the Amazon varies between rising and falling water levels.

Water and climate

Based on observed precipitation (Xie and Arkin 1996) and river discharge records archived at the Global Runoff Data Centre, annual water balance for most river basins worldwide was estimated (Oki, Nishimura, and Dirmeyer 1999); this is presented in Figure 4. The ordinate in this figure is the residual of long-term mean annual precipitation and runoff; this annual loss should correspond to long-term mean annual evapotranspiration. Different symbols are used for plotting: red stars indicate the water balance of the river basins where gauging stations are located between 20°S and 20°N. The plus symbols indicate river basins where gauging stations are located between 20° and 40° in both hemispheres, and the blue circles are 40° or higher. Lines connect the mean precipitation and annual loss for each 5° latitudinal belt. As seen, even though the scatter is large, approximately 70% of precipitation undergoes evapotranspiration in high-latitude river basins. On the other hand, mean evapotranspiration in tropical river basins is approximately 1000 mm year−1 with less dependency on annual precipitation. Such analyses on the relationship between P and E have long been used for estimations of global water balance, for example, in Baumgartner and Reichel (1975). It is also clear from Figure 4 that river basins with an annual precipitation of less than 800 mm year−1 have a marginal amount of river runoff, as most precipitation is used for evapotranspiration. In these river basins, evapotranspiration
is mainly controlled by the availability of water (water-controlled); this is in contrast to tropical river basins with precipitation higher than 1000 mm year\(^{-1}\), where annual evapotranspiration is limited by the available energy (energy- or radiation-controlled).

Budyko (1974) proposed an equation:

\[
\frac{E}{P} = \left[ \zeta \left( \tanh \frac{1}{\zeta} \right) \left( 1 - \cosh \zeta + \sinh \zeta \right) \right]^{1/2}
\]

where \( E \) and \( P \) are annual evapotranspiration and precipitation, respectively, and Budyko’s dryness index \( \zeta \) is defined as:

\[
\zeta = \frac{Rn}{lP}
\]

where \( Rn \) and \( l \) are net radiation and the coefficient of latent heat, respectively. This equation is derived by considering that \( E/P \) should be asymptotic to 1.0 for dry regions (large \( \zeta \)) as \( E \) should be less than \( P \), and \( E/P \) should be asymptotic to \( Rn/IP \) for wet regions (small \( \zeta \)) as \( E \) should be less than \( Rn/IP \).
Budyko’s equation 8 is conceptual but can provide a realistic water balance, as shown in the inset in Figure 5a. Mean water balance averaged for each $0.2\zeta$ ($= Rn/IP$) bin estimated by a land surface model corresponds fairly well with the curve according to Budyko’s equation 8, even though large scatters are found in the plots of each $1^\circ$ longitudinal and latitudinal grid box. Yang et al. (2009) analyzed annual water balance in 99 river basins in China and concluded that this scatter can, at least partially, be explained by the vegetation cover in the river basin.

Annual water balances estimated by the pilot phase of GSWP3 were analyzed and each $1^\circ \times 1^\circ$ grid box was classified into one of six
The six classified regions are illustrated in Figure 5a along with the ice-covered region. The differentiation is difficult between the tropical humid region and the very humid region using only Budyko’s dryness index ζ. Therefore, annual precipitation is considered in the classification. It is interesting to see ζ is similar in both tropical and high-latitude regions in addition to the Asian monsoon region. Perhaps it is necessary to also consider the seasonal change of major water balance terms for better differentiation of these regions.

Long-term mean water balance for each climatic region classified by ζ is presented in Figure 5b. Separation of mean annual precipitation into evapotranspiration, surface runoff, and subsurface runoff is also illustrated. The sum of these corresponds to annual precipitation and the ratio of annual evapotranspiration to precipitation is close to the mean ζ of each region. Slightly negative evapotranspiration in the ice region indicates net sublimation in the region, that is, the land surface obtains energy from the atmosphere through sublimation.

Further, evapotranspiration can be divided into four components: bare soil evaporation (Es), evaporation from intercepted water on leaves (Ei), evaporation from open water (Eiw), and transpiration from vegetation (Et); these are illustrated in Figure 5c averaged over classified climatic regions. Transpiration and evaporation from intercepted water are proportional to the vegetation biomass in each region, and it is interesting to note that the magnitude of bare soil evaporation is more uniform globally than other components, except for arid, very humid, and ice regions.

Even though the values in Figure 5c are not definitive, it is interesting to see that bare soil evaporation and transpiration from vegetation compare closely, and interception loss is approximately 10% of the total evapotranspiration. It would be interesting if these estimates could be revised and validated by observation-based measures.

**SEE ALSO:** Aquifers; Arid climates and desertification; Atmospheric/general circulation; Climate change and biogeography; Climatology; Clouds; Cryosphere studies: history; Dams; Droughts and water shortages; Earth’s energy balance; Freshwater resources: past, present, future; Geodesy; Geoengineering/climate intervention; Glacier hydrology and runoff; Global climate change; Global climate models; Groundwater; Hydrology: history; Hydroclimatology and hydrometeorology; Hydrologic flow models; Infiltration; Irrigation; Lakes and limnology; Land-use/cover change and climate; Monsoons; Mountain climatology; Mountain hydrology; Oceanic circulation; Permafrost: definition and extent; Plant–water interactions; Polar climates; Precipitation; River basin management and development; River ice and ice jams; Rivers and river basin management; Rivers and streams; Snow; Snow cover; Snow cover changes; Soil water; Water budget; Water and climate change; Water quality; Water resources and hydrological management; Watersheds; Weather, extreme; Wetlands hydrology
HYDROLOGIC CYCLE

References


Further reading

Hydrologic flow models

Douglas A. Howard
Georgetown University, USA

Hydrologic flow models are conceptual or mathematical representations or simulations of real-world physical hydrologic systems. Although this definition includes all of the hydrologic cycle, we will focus on the hydrologic flow of fluid on or beneath the surface of a planetary body. In general terms, a flow model is an accurate simulation of a process that cannot be directly observed, which applies to both surface water and subsurface, aquifer, flow regimes. Hydrologic flow models are used to understand why a flow system behaves in the observed way and to predict or forecast how it will behave in the future. They are also used to model hypothetical flow conditions. Typical hydrologic problems that models address are surface water flow (flood routing), groundwater flow (well-pumping and developing intersecting cones of depression), and solute transport (tracing chemical paths and their attenuation along pathways).

Types of hydrologic flow models

There are four basic types of hydrologic flow models, conceptual, physical, analog, and numerical, that can be used together or individually depending on a modeling approach.

Conceptual hydrologic flow models are broad general constructs of a hydrologic flow system. They are typically used for initial planning stages of more sophisticated hydrologic flow models. Conceptual models tend to be static and describe a specific condition at the highest level and do not fully describe the real-world hydrologic system.

Physical hydrologic flow models are commonly tangible large-scale representations of real-world hydrologic systems used to understand an engineering structure or natural formation utilizing simple hydrologic conditions. They are useful for understanding a specific hydrologic condition; however, they tend to be costly and inflexible in design. Henry Darcy created a physical sand-column model where he used dye-tracers to show flow paths and contaminant or pollutant distribution and subsequently established Darcy’s Law, the basis for modern groundwater analysis (Darcy 1856).

Analog models use the properties of one phenomenon to model another. One hydrologic example is the use of Ohm’s law governing the flow of electricity to model Darcy’s law of groundwater flow. It was found that the flow of electricity through circuits mimics the flow of water through a pipe. Although this particular example is useful, the disadvantage of analog models is that they tend to be very specific to hydrologic conditions and constraints, and are inflexible to adaptation for general hydrologic flow application.

Numerical or mathematical hydrologic flow models are now the preferred approach to hydrologic modeling due to their flexibility and wide range of applications. They are the basis for modern hydrologic modeling software programs. They are dynamic models that can be manipulated as needed to make predictions of future flow behavior. Numerical hydrologic flow models are generally sequences of equations that convert numerical inputs representing
HYDROLOGIC FLOW MODELS

Hydrologic parameters to numerical outputs such as flow velocity and discharge. These models are used when working with distributed parameters that vary across the modeled area, such as transmissivity and storativity, or when there are complex or moving boundary conditions. Numerical models recast partial differential equations that represent the flow conditions into an algebraic (matrix) form that is easily solved by digital computers. These numerical representations approximate the natural conditions to be modeled; as a result the output is also a close approximation of real-world conditions.

The most common numerical methods used in hydrologic modeling are finite difference and finite element techniques. These gridding techniques allow the modeled study area to be subdivided into smaller units where numerical analysis is accomplished more accurately.

The finite difference technique is the most commonly used for problems involving distributed parameters where algebraic difference approximations replace partial differential equations to compute, by discretization, approximate parameter solutions. These solutions represent the parameters at each cell node. Using the finite difference technique, parameters are not represented continuously across grid units but are represented at a node associated with each polygonal cell. Each cell has a node which is the location (reference point) of input or output parameters and can be either block-centered or mesh-centered. Block centered nodes are located at the center of the polygon while mesh-centered nodes are located at the corners of each polygon. Polygons can be either rectangular grids for two-dimensional (2D) models or prisms for three-dimensional (3D) models. The selection of the appropriate grid depends upon the boundary conditions of the modeled area.

The finite element technique uses a numerical approach to solve differential equations through integrals and, most commonly, the Galerkin interpolation method. This interpolation technique enables computation of output parameters that are obtained at any point within the problem area and not only at cell nodes as with the finite difference technique. The grid subdivisions are “elements” in the shape of triangles for 2D models and tetrahedrons for 3D models. The finite element technique is more accurate than finite difference for problems that have a moving boundary condition such as a transient water table.

Simplified forms of numerical models are known as analytical models. Analytical models are very often “back-of-the-envelope” first approximations of unknown parameters, typically in one dimension. Darcy’s law is based on an analytical flow model where the initial and boundary conditions are known.

With the use of desktop computers, hydrologists now have the ability to use or develop computer numerical hydrologic models to simulate more complex hydrologic systems than physical or analog models. These computer models are more flexible, dynamic, and reusable, and can mimic the field conditions more accurately, especially with the use of digital terrain models and geographic information systems (GIS).

Components of a hydrologic flow model

There are generally four key components of a hydrologic flow model process for developing reliable output parameters. The first is conceptualization of the problem. Although it is difficult to simulate the entire system accurately, this step sets up the problem to be modeled. The next step is to select or develop an appropriate model. As discussed, the numerical model has become the modeling method of choice; however, physical modeling is still used for understanding engineering structures. The third step is to
calibrate the model where iterative adjustments to the model parameters enable best-fit estimates to real-world parameters. Finally, the model is validated by historical matching of the model output to a long record of observed data.

The general approach to hydrologic modeling begins with building a conceptual model, followed by selecting a mathematical equation, or series of equations, that approximates the physical behavior of the hydrologic system. It is important that the hydrologic system, and how it pertains to the problem, is well understood. Additionally, selecting the appropriate equation relies on an appropriate set of simplifying assumptions – what does the equation require and assume to be true. For example, stable boundary conditions should be well defined for a finite difference numerical method; otherwise the finite element technique may be used. Simplifying assumptions can be tested to understand the tradeoffs for a less-than-optimal solution.

Model calibration is as important as any other modeling component and is accomplished by comparing computer model-generated predictions with measured known variables or parameters. The calibration processes is iterative, by trial and error, adjusting the model until it agrees with the real-world data within an acceptable margin of error or tolerance. For example, a groundwater model may compare its output to the known groundwater table or potentiometric surface condition. Another component of calibration is sensitivity analysis to determine the uncertainty of the output model parameters against known field values of the most sensitive parameters. The model is tested to determine which terms or parameters are most sensitive. For example, open channel flow may be very sensitive to Manning’s $n$, surface roughness coefficient. Therefore, this parameter may be adjusted within a range of roughness coefficients to test the uncertainty of the output of the model. Knowing the sensitivity of each parameter within the flow model enables a better understanding of the model outcome.

Model verification following calibration uses historical matching where the model output is compared to a long record of observations, which is typical of surface water flood predictions. The model may be considered validated if it consistently reproduces historical known output. If the model does not achieve the desired accuracy, model parameters may be changed to recalibrate the model.

An additional component of the modeling process to be aware of is the large amount of data required by the model (input) and the output of the model. Selection of the appropriate data to support and solve the problem is important. It is equally important to collect and compile the data into a database that the hydrologic model can readily use. As we have seen, each numerical flow model grid consists of a node where data are collected, computed, or interpolated. Each node may have many parameters assigned to it, such as flow velocity, discharge, transmissivity, storativity, or dispersion coefficients. Additional input data may consist of project-specific data (locational, temporal, or design characteristics) and existing data from external sources.

**Uses of hydrologic flow models**

Hydrologists use hydrologic flow models to better understand how changes in land use and climate will impact groundwater and surface water quantity and quality on a local to regional scale. Hydrologic flow models are also used to assist water-resource management and design approaches, for prediction and forecasting, and as tools to understand scientific questions. A thorough understanding of the physics of hydrologic processes allows the practitioner the foundation to develop and apply these hydrologic models.
HYDROLOGIC FLOW MODELS

There are many computer models for hydrologic modeling on the market today, many of which use existing computer programming code or application programming interfaces (API) due to well-validated and stable code, better quality control, and lower costs. Likewise, many hydrologists use published, well-established models because they have been validated and widely used and accepted as a standard, especially if used for regulatory processes or litigation. The most widely accepted and used hydrologic flow model is MODFLOW (McDonald and Harbaugh model; US Geological Survey 2013) and has been adopted by numerous commercial software packages.

With the wide use and accessibility of GIS, both hydrologic and hydrogeologic flow-modeling tools are included in or available as add-on modules to the software. The ease of use of GIS and its ability to visualize model output in 2D and 3D representations provides hydrologists a valuable tool for hydrologic flow modeling and presentation of results.

The abundance of hydrologic flow models and software from commercial vendors or freely available in the public domain makes the selection of the appropriate model difficult. The following are considerations when selecting a hydrologic flow model or software. First, does the model have the ability to solve the problem? What are the computer hardware requirements to run the software, data storage, and output? Is the user interface user-friendly based on the user’s abilities to input and visualize the data? Historically, many hydrologic models were based on command-line data entry requiring well-formatted databases or text files. In the modern age of computing, many, but not all, of these command-line models have been converted to user-friendly graphical user interfaces (GUI), increasing the usability to accommodate a larger audience. Another consideration is the acceptance of the model by the industry or institution in which it is to be used. Hydrologic models came into use mostly due to regulatory action, where a standard on which to measure hydrologic processes was needed. Subsequently, certain models have been accepted as standards by the industry. For example, standards for groundwater modeling were established by the American Society for Testing and Materials (ASTM). Finally, the availability and cost of the software can be a limiting factor for some users of hydrologic flow models. On the one hand, there are many highly robust hydrologic flow models free to the public, but they tend to be less user-friendly. On the other hand, there are many very user-friendly proprietary software products that are available at considerable cost. Two resources that provide access to hydrologic models and software are the Integrated GroundWater Modeling Center (IGWMC) clearinghouse for groundwater modeling software and the US Geological Survey.

SEE ALSO: Groundwater; Hydrologic cycle; Models in geomorphology

References


Further reading

Hydrology is a relatively new science and a truly ancient one (Biswas 1970). It is fascinating to consider that the word hydrology, which refers to the study of water flows from precipitation to infiltration, groundwater, evapotranspiration, and runoff, only developed in the eighteenth century (Perrault and Perrault 1721). Even then it was regarded by some as the “wisdom of God,” as well as an empirical science (Tuan 1968). The scope of hydrologic science has been shaped by a combination of direct observation, remote sensing, practical action, and cultural beliefs. It is interesting to try to push the origins of hydrologic science back to include relationships between hydroclimatic phenomena and prehistoric human settlement (e.g., Magill, Ashley, and Freeman 2013). One may even wonder whether other species evolved sophisticated water finding, navigation, and regulatory capacities as hydrologic agents (Finlayson 2014). In each of these respects, the history of hydrology has far greater historical depth than is commonly perceived.

The benchmark study on this topic is Asit Biswas’s (1970) History of Hydrology which drew together a rich array of sources to survey the field. His volume begins with ancient archaeological and historical innovations from antiquity to 600 BCE and then proceeds to the classical world of Greece, Rome, and the Hellenistic empire. After an understandably short chapter on the medieval period (though see Magnusson 2001), the greater part of the volume moves chronologically from sixteenth-to nineteenth-century Europe. It offers brief descriptions of inventors, inventions, experiments, and scientific formulas.

Notwithstanding the fascinating contents of Biswas’s volume in 1970, a search of the literature indicates that it did not presage much advanced historical research, which remains limited in quantity, fragmented across subdisciplines, and muddied by terminology that sometimes conflates time series analysis with historical research. Recent assessments of hydrology rarely treat historical research as a source of insight for the field. For example, a 2012 US National Research Council report on Challenges and Opportunities in the Hydrologic Sciences discusses water processes over time, but it does not mention how historical scholarship can help interpret or explain them. Even so, the glass is also half full, as this entry hopes to show.

We may begin with a brief perspective on how the hydrologic sciences are currently envisioned. The United Nations International Hydrological Decade (1965–1975) and UNESCO International Hydrologic Programmes (1975–present) have included essays on the history of their programs and the topics they covered (see Linton 2010, 166–172). A key source for contemporary hydrologic thought is the International Association of the Hydrological Sciences (IAHS), established in 1922, which has nine commissions, four working groups, and a decadal vision statement (Table 1) (Montanari et al. 2013).

These topical groups indicate the diversity of the hydrological sciences and their association with water resources engineering (for a critical view see Klemes 1986). As Biswas noted in 1971, hydrology remains broadly distinct from
This approach is complicated by two factors. First, the sources cited still encompass a wide range of historiographies (i.e., modes of historical writing). The language of archaeological research differs from research based on archival documents, instrumental records, biographical archives, and ethnohistorical accounts. These historiographical approaches have developed more in parallel than in dialogue with one another (e.g., in Landa and Ince 1987; Patton 1987; and Rodda and Ubertini 2004). Second and more problematic is that much of the literature consists of perspectives by hydrologists who do not have training in historical theory or methods. While valuable in terms of firsthand knowledge, they rarely employ conceptual approaches that shed light on the social context of how and why the field developed in some directions rather than others, or what impact those trends have had on society and environment as well as science. Some social scientists interpret this neglect of history as part of a broader depoliticizing strategy in modern science and technology (e.g., Linton 2010; Schmidt 2014).

Before taking up these debates, it is useful to reflect upon how the field has developed since Biswas’s pioneering work. The History of Hydrology remains a valuable reference work, particularly for European hydrology in the sixteenth through nineteenth centuries. It strives for a global perspective through vignettes on ancient hydrologic sites, instruments, and records up to the sixth century BCE. Subsequent chapters are descriptive and largely chronological. Because these historical advances are not widely known among water scholars, the book has continuing value as a survey of innovations and inventors through the nineteenth century.

After the passage of over forty years, it is timely to revisit the history of hydrologic research, extend the account through the twentieth century, and consider the growing role of

---

**Table 1**

<table>
<thead>
<tr>
<th>Table 1 International Association of the Hydrological Sciences (IAHS) commissions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snow and ice hydrology</td>
</tr>
<tr>
<td>Surface water</td>
</tr>
<tr>
<td>Water quality</td>
</tr>
<tr>
<td>Continental erosion</td>
</tr>
<tr>
<td>Coupled land-atmosphere systems</td>
</tr>
<tr>
<td>Tracers</td>
</tr>
<tr>
<td>Remote sensing</td>
</tr>
<tr>
<td>Statistical hydrology</td>
</tr>
<tr>
<td>Water resource systems</td>
</tr>
</tbody>
</table>

national and international hydrologic research institutions such as the International Hydrological Programme. Proceedings of the American Geophysical Union also include historical topics (e.g., debates in paleoflood research by Landa and Ince 1987; Patton 1987; and others). A small but significant literature follows in Biswas’s footsteps with short essays that offer historical perspectives on the field at large, with increasing emphasis on the twentieth century (e.g., Back 1988; Dooge 2004; Nace 1974, 1975; Reilly 2004; Rodda 2006).

Ancient civilizations and their hydrologies – Mesopotamia, Indus, Rome, China, and beyond

One of the strengths of Biswas’s volume was its inclusion of water development in early civilizations. Over the past four decades several lines of water heritage have received detailed archaeological and historical attention. For example, Biswas’s brief discussion of Joseph Needham’s magisterial Science and Civilization in China invites consideration of the wealth of additional material in that source, as well as in subsequent research. In Egypt, Dooge (2004) has underscored the social as well as hydrological significance of Nilometers. Historian Terje Tvedt (2011) and colleagues have undertaken extensive historical studies of Nile River basin flows and management and have curated a multivolume History of Water series that includes papers on the history of hydrologic science. Ancient Mesopotamian irrigation systems have been the subject of archaeological surveys, excavations, dating methods, and remote sensing analyses that test hypotheses about the links between water systems, agricultural domestication, urbanization, environmental degradation, and decline. Research on ancient Roman water systems, hydraulics, and flooding continues to flourish, stimulating the formation of aqueduct studies groups and the Frontinus society named after the famous Roman curator aquarum.

Medieval Europe is not generally known for advances in hydrologic science, but locally adapted water technologies were developed in many places. Conversely, Islamic scholars of the ninth through twelfth centuries achieved scientific renown particularly for their contributions to hydraulic and agronomic sciences (vis-à-vis hydrology in the modern sense of the term; Hill 1994).

The transition to modern hydrology occurred in colonial regimes as well as European centers, which examined ancient water systems while generating new forms of water science and management. Gilmartin (1994) frames this shift as a dynamic tension between “imperial science” and the “science of empire.” Imperial science linked canal hydraulics with irrigation engineering, and developed a canal regime theory to maintain stable flows, channel banks, and sediment transport processes. The science of empire used water development to sedentarize the population, extract revenue, loyalty, and raw materials for European manufactures. D’Souza (2006) offers a similar critique of colonial hydrologic research and practice related to flood control in eastern India.

Histories of ideas about the hydrologic cycle, climatic water budget, and water molecule

Just before Biswas’s volume was published, geographer Yi-Fu Tuan (1968) wrote a fascinating monograph on The Hydrologic Cycle and the Wisdom of God, which reappraised the contributions of seventeenth- and eighteenth-century natural theologians. These were scientists such as John
HYDROLOGY: HISTORY

Ray (d. 1705 CE) and Thomas Burnet (d. 1717 CE) who strove in varying degrees to reason jointly from physical observation, experimentation, and theology. Tuan retraced their writings about the hydrologic cycle as a study in the history of ideas. They treated rainfall–runoff relationships, groundwater, springs, sea level, and natural disasters as matters of cosmological design as well as geophysical order.

The history of ideas approach has been eclipsed in recent decades by materialist and realist historical methods. The latter approaches also address the hydrologic cycle (Karterakis et al. 2007) and increasingly the hydrosocial cycle, which encompasses water use, degradation, and reuse (Schmidt 2014). Other scholars have reviewed the development of hydrology in specific regions and sectors, for example, arid hydrology, forest hydrology, and watershed hydrology (Gordon 1985; McCulloch and Robinson 1993; Todini 2011). Experimental watersheds established in the early twentieth century to test forest and grassland cover effects on streamflow, sediment, and erosion have become subjects of historical research (e.g., experimental watersheds at Wagon Wheel Gap, Colorado; Coweeta, North Carolina, and so forth; in Ice and Stednick 2004; Schumann and Herrmann 2010). A similar sequence is unfolding in urban hydrology in which early studies such as Wolman’s research on urban runoff and sedimentation in the 1960s have expanded to sustained urban hydrology research programs at the Long-Term Ecological Research (LTER) sites in Baltimore and Phoenix (Purcell 1998).

Water budget and water balance analysis have been sustained long enough to stimulate historical research, which includes a biographical focus on the work of Warren Thornthwaite, Russell Mather, and others (Mather and Sanderson 1996). Water budget analysis combines basic biophysical research with research on measurement, modeling, and socioeconomic application in irrigation, power, and other sectors (Wescoat 2014).

At the global scale, research on hydrologic processes has evolved from Soviet-led water budget estimates to research on the potential impacts and feedback effects of global climate change. Determining the relative importance of climate variability, extreme hydrologic events, channel change, social stratification, and conflict are central to these studies. As the field of climate history expands, so too will historical studies of hydrologic processes and interventions (e.g., Wigley, Ingram, and Farmer 1981). They are emulated in planetary studies of water evidence and landform evolution, particularly on Mars (Detay and Gaujous 2002). They will increasingly be linked with historical reconstructions at regional and local scales on Earth (e.g., Pastore et al. 2010).

At the smallest scale, several important papers are published every year on the physics of the water molecule in leading journals such as Nature and Science. Its properties and anomalies are endlessly fascinating. A recent monograph on the history and philosophy of water physics and chemistry offers a provocative argument for scientific pluralism, going so far as to suggest that even the long-rejected phlogiston concept may have continuing value for its way of thinking about water (Chang 2012). It is interesting to reflect upon the potential implications of these molecular scale studies for larger scale hydrologic processes, from topics in atmospheric physics to heat transfer, turbulence, sediment transport, and water chemistry.

Hydrologic processes and measurement

Historical research has concentrated on surface water hydrology and fluvial landforms
Ground water was the subject of ancient speculation about storage in a vast underground cavern. In a modern legal case, ground water was deemed, “so secret, occult, and concealed that an attempt to administer any set of legal rules in respect to them would be involved in hopeless uncertainty and would be, therefore, practically impossible” (Frazier v. Brown 1861). Karst hydrology is perhaps the most mysterious of all but has drawn attention of spelunkers as well as geologists (Clendenon 2009). Spring flows, formerly known as fountains, were one of the enigmatic phenomena to be scientifically explained as pluvial rather than subterranean in origin by Perrault in the eighteenth century, though this account has been shown to have modern rather than historic recognition in Perrault’s times (Linton 2010, 144–147). Histories of hot springs, mineral waters, and coastal and island aquifers have likewise drawn perennial attention (Epstein 1987; Weisz 2001).

In future histories, scientific research on groundwater dynamics will stand out as a major emphasis of late twentieth-century hydrology, due in part to rapid advances in groundwater pumping technologies, withdrawals, and contamination in all water use sectors. Contaminant hydrology is an ancient as well as modern topic. Historians have analyzed urban sewers from Roman to modern times. Reconstructing soil water contamination has become a forensic topic in environmental remediation research (Chapelle 1999; Colten 1985).

Historic contributions to fluvial geomorphology have been compiled in a four-volume collection, the first three volumes of which were organized around works leading up to and following from studies of landform evolution by William Morris Davis (Chorley, Dunn, and Beckinsale 2009). The field is much older, as Leonardo da Vinci’s studies of runoff, sedimentation, and channel change attest (Tokaty
HYDROLOGY: HISTORY

Indeed, floods constitute the most intensively studied hydrologic phenomenon in historical terms (Brazdil, Kundzewicz, and Benito 2006; Costa 1987; Cyberski et al. 2006). Their catastrophic effects, dramatic narratives of heroism and destruction, and generally well-recorded characteristics help account for the attention of both hydrologists and historians. Perhaps the largest research frontier concerns the history of global hydroclimatic change, which draws upon paleoclimatic, paleobotanical, and archaeological methods along with historical research.

Toward an historical sociohydrology?

Interestingly, late twentieth-century efforts to define hydrology as a distinct physical sciences discipline, in contrast to the caricature of “dillettantism” of applied water resources research (Klemes 1986), has been supplanted by renewed emphasis on scientific synthesis (National Research Council 2012). The extent to which the former view of hydrology as a reductionist science accurately characterizes modern water research has been debated (Linton 2010; Schmidt 2014). For example, Schmidt (2014) associates early modern social research on water with the vitalist philosophy of Henri Bergson. What seems clear is that there is renewed emphasis on linking social and hydrologic research in order to understand what Sivapalan, Savenije, and Blöschl (2012) present as the coevolution of social and hydrologic processes.

Historical sociohydrology is one of three main lines of research in this emerging field. Zlinszky and Timar (2013) have shown how historic maps can help reconstruct the history of both wetland hydrology and modification. Another recent study retraces the “duty of water” standards of water use from eighteenth-century steam engines in Britain to nineteenth-century canal irrigation in India, and twentieth-century water rights in the western United States (Wescot 2014). The development of a sociohydrology working group in the IAHS Panta Rhei (“Everything Flows”) decade may help coalesce individual studies into broader contributions.

Whatever the future of particular investigations, it seems clear from this review that the history of hydrology is an underdeveloped topic with considerable prospects. Ancient in its roots, fascinating in its modern developments, the history of hydrology includes many topics that beckon for future inquiry to better understand transformations of the field and better imagine its future paths.

SEE ALSO: Freshwater resources: past, present, future; Groundwater; Plant–water interactions; Soil water; Surface water; Water budget; Water and climate change; Water engineering; Water resources and hydrological management

References


Frazier v. Brown. 1861. 12, Ohio Supreme Court, 294.


HYDROLOGY: HISTORY


Tuan, Yi-Fu. 1968. The Hydrologic Cycle and the Wisdom of God. Toronto: University of Toronto.


**Further reading**


Remote sensing is the acquisition of information about an object or phenomenon without making physical contact with the object. In modern usage, the term generally refers to the use of aerial sensor technologies to detect and classify objects on Earth (both on the surface and in the atmosphere and oceans) by means of electromagnetic radiation. It may be split into active remote sensing (when a signal is first emitted from aircraft or satellites) or passive (when information is merely recorded) (Schowengerdt 2006).

As one special and effective way to acquire more information from a single observation, imaging spectrometry can simultaneously obtain a set of images in many relatively narrow, contiguous and/or noncontiguous spectral bands throughout different ranges of the electromagnetic spectrum (Chang 2003; Jensen 2005); this is also known as hyperspectral imaging. The concept originated in the 1980s, when A.E.H. Goetz and his colleagues at NASA’s (National Aeronautics and Space Administration) Jet Propulsion Laboratory began a revolution in remote sensing by developing new instruments such as the Airborne Imaging Spectrometer (AIS), then called AVIRIS, for Airborne Visible Infra-Red Imaging Spectrometers (Goetz et al. 1985). The basic workflow design of the image spectrometer has been vividly illustrated in detail (Shaw and Manolakis 2002). A sensor loaded in the satellite makes the acquisition from reflected sunlight passing through a spectrophotometer apparatus that splits the energy into different channels thus forming a continuous display of the spectrum of some material, such as soil, water, or vegetation, as shown in Figure 1. This kind of remote sensing activity using hyperspectral imaging equipment is called hyperspectral remote sensing (HSRS).

HSRS began a revolution in remote sensing by combining traditional two-dimensional imaging technology and spectroscopy, allowing for the synchronous acquisition of both images and spectra of objects (Tong et al. 2014). HSRS is able to describe the specific spectral characteristics of some materials in great detail. The reflective features of some materials can be extended to a continuous spectrum signature instead of several sample values, unlike single or multispectral remote sensing. As shown in Figure 1, data from multispectral sensors such as SPOT and Landsat can only acquire spectral signatures for several channels, which may not be able to best embody the spectral features of some similar materials, while HSRS sensors are greatly superior. For example, in bands 5 and 7, the grass with different water content cannot be separated in multispectral imageries. For material with extremely similar reflective spectral features, such as two subclasses of aspen leaf, they can be well separated using hyperspectral data.

Many commercial and government-funded hyperspectral imagers operate in the reflective range of the electromagnetic spectrum, from approximately 0.4 to 2.5 μm (400–2500 nm). This portion of the spectrum covers the visible through shortwave infrared (SWIR) wavelengths. Hyperspectral imaging systems designed...
### Table 1  Some representative hyperspectral image spectrometers.

<table>
<thead>
<tr>
<th>Hyperspectral sensors</th>
<th>Institution</th>
<th>Nationality</th>
<th>Band number</th>
<th>Wavelength (μm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVIRIS</td>
<td>NASA</td>
<td>USA</td>
<td>224</td>
<td>0.4–2.5</td>
</tr>
<tr>
<td>AISA</td>
<td>Spectral Imaging Ltd</td>
<td>Finland</td>
<td>286</td>
<td>0.45–0.9</td>
</tr>
<tr>
<td>CASI</td>
<td>Itres Research</td>
<td>Canada</td>
<td>288</td>
<td>0.43–0.87</td>
</tr>
<tr>
<td>DAIS 2115</td>
<td>GER Corp.</td>
<td>USA</td>
<td>211</td>
<td>0.4–12.0</td>
</tr>
<tr>
<td>HYMAP</td>
<td>Integrated Spectronics Pty Ltd</td>
<td>Australia</td>
<td>128</td>
<td>0.4–2.45</td>
</tr>
<tr>
<td>PROBE-1</td>
<td>Earth Search Sciences Pty Ltd</td>
<td>USA</td>
<td>128</td>
<td>0.4–2.45</td>
</tr>
<tr>
<td>PHI</td>
<td>SITP (Shanghai Institute of Technical Physics)</td>
<td>China</td>
<td>224</td>
<td>0.46–0.85</td>
</tr>
<tr>
<td>Hyperion</td>
<td>NASA</td>
<td>USA</td>
<td>242</td>
<td>0.35–2.6</td>
</tr>
<tr>
<td>HYDICE</td>
<td>HDOS</td>
<td>USA</td>
<td>210</td>
<td>0.4–2.5</td>
</tr>
<tr>
<td>CHRIS</td>
<td>ESA (European Space Agency)</td>
<td>EU</td>
<td>62</td>
<td>0.415–1.05</td>
</tr>
<tr>
<td>ROSIS</td>
<td>DLR German aerospace center</td>
<td>Germany</td>
<td>102</td>
<td>0.43–0.834</td>
</tr>
<tr>
<td>HJ-1</td>
<td>China Center for Resources Satellite Data and Application</td>
<td>China</td>
<td>110–120</td>
<td>0.4–0.95</td>
</tr>
</tbody>
</table>

for the emissive range, also called the thermal or longwave IR region, typically operate in the span from 7 to 14 μm. These sensors are less prevalent in commercial and academic research communities than in the government sector since they require more sensitive optics and more costly and complicated electronics. The AVIRIS system is able to cover the wavelength region from 0.4 to 2.5 μm, splitting it into more than 200 spectral channels, at nominal spectral resolution of 10 nm. Several other famous imaging spectrometers have been designed and put into practice, including the MODIS (Moderate Resolution Imaging Spectroradiometer) and the HIRIS (High Resolution Imaging Spectrometer). Since the initial development of hyperspectral sensors by NASA, data from these have been successfully applied to both qualitatively and quantitatively analyze various phenomena on the surface of the planet, prompting many other countries to start developing their own hyperspectral sensors. Canada, Australia, and China have all researched and developed their own image spectrometers, some of which are listed in Table 1.

Taking advantage of the unique capacity of HSRS to identify different materials, in the early 1970s, NASA established a preliminary Earth resources information system (ERSIS) gathering together more than 2000 kinds of reflectance laboratory spectrum data for vegetation, soil, mineral, and water bodies. In the process of creating this database for specific application, NASA also accumulated data and technology
for the rapid development of such spectrum databases. These include the Jet Propulsion Laboratory (JPL), United States Geological Survey (USGS), and Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) spectra databases as shown in Table 2.

Different ranges of wavelengths are supposed to excel at different application fields in HSRS. Development of advanced satellite platforms and sensor hardware was a boost to HSRS. Computer science, data storage, and information processing technology enable HSRS to play a significant role in promoting economic growth, environmental protection, and homeland security (Grahn and Geladi 2007). HSRS has been widely used in applications worldwide such as food safety, pharmaceutical process monitoring, and quality control in biomedical, industrial, biometric, and forensic processes at laboratory scale (Bioucas-Dias et al. 2013). A breakdown of common spectral data exploitation tasks and their correspondent spectrum regions are summarized in Farrell (2005).

HSRS applications meet Chinese demands for resource exploration and land utilization. Since the 1980s, the economy of China has grown rapidly; subsequently, exploration for mineral resources has become a vital concern. HSRS is also used in agriculture and forestry. Rapid economic growth has created serious problems; HSRS plays an important role in environmental management, in particular in monitoring water and atmospheric pollution and the effects of urban heat islands.

HSRS images have features that are different from those of traditional remotely sensed data. The data volume of a specific scene enormously increases with the refinements in the spectral channel. Information from adjacent bands, however, is analogous with great redundancy. Therefore, methods and algorithms such as PCA (principal component analysis), ML (maximum likelihood), and NDVI (normalized difference vegetation index) aimed at exploiting traditional remotely sensed data are no longer applicable. In

<table>
<thead>
<tr>
<th>Spectra library</th>
<th>Material</th>
<th>Number</th>
<th>Wavelength (μm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPL</td>
<td>Mineral</td>
<td>160</td>
<td>0.4–2.5 (0.4–0.8 1 nm) (0.8–2.5 4 nm)</td>
</tr>
<tr>
<td>USGS</td>
<td>Mineral+vegetation</td>
<td>498</td>
<td>0.4–2.5 (0.4–0.8 4 nm) (0.8–2.35 10 nm)</td>
</tr>
<tr>
<td>ASTER</td>
<td>Natural+manmade object</td>
<td>2000</td>
<td>0.4–25</td>
</tr>
</tbody>
</table>
Hughes phenomenon of HSRS data.

HSRS image processing, scientists have identified some digital image processing methods that may be applied in a relatively straightforward sequence to extract useful information (Jensen 2005). The general steps are shown in Box 1.

The main processing techniques applicable to HSRS data cover dimensionality reduction, endmember extraction, unmixing, classification, target detection, change detection, and physical parameter retrieval. A brief introduction to these techniques is given in the following section.

**Dimensionality reduction**

HSRS data often have hundreds of bands. Enormous dimensionality will lead to high correlation among bands and therefore yield highly redundant information. HSRS data in their original form suffer from the Hughes phenomenon, the “curse of dimensionality,” referring to the loss of classification accuracy as data dimensionality increases, as shown in Figure 2. Data with higher dimensionality are not expected to yield a better classification accuracy. Moreover, data without dimensionality reduction are difficult to process without incurring huge computational costs, which impedes efficient image analysis.

The two dimensionality reduction methods for HSRS data are feature selection and feature extraction. Feature selection methods permit users to directly choose significant band subsets from the original bands according to their various needs, keeping the original information found in HSRS data in a partial form. Feature extraction methods are used to generate new features by transforming the original bands following certain rules. Once defined accuracy requirements are satisfied, feature extraction methods can save on the monetary and computational costs of measurement and calculation, with fewer feature observations and less analysis of the data.
Box 1 General steps used to extract information from hyperspectral data.

**State the nature of the information extraction problem**
Specify the geographic region of interest.
Define the classes or biophysical materials of interests.

**Acquire appropriate remote sensing and initial ground reference data**
Select remotely sensed data based on the following criteria
Remote sensing system considerations:
spatial, spectral, temporal, and radiometric resolution
Environmental considerations: atmospheric, soil moisture, phonological cycle, etc.
Obtain initial ground reference data based on: a priori knowledge of the study area

**Process hyperspectral data to extract thematic information**
Subset the study area from the hyperspectral data fight line(s).
Visual individual band examination
Visual examination of color composite images
Animation
Statistical individual band examination; signal-to-noise ratio
Radiometric correction:
Collect necessary *in situ* spectroradiometer data (if possible)
Collect *in situ* or environment data (e.g., using radiosondes)
Perform pixel-by-pixel correction (e.g., ACORN)
Perform band-by-band spectral polishing
Empirical line calibration
Geometric correction/rectification:
Use onboard navigation and engineering data (GPS and Inertial Navigation System information)
Nearest-neighbor resampling
Reduce the dimensionality of hyperspectral dataset:
Minimum Noise Fraction (MNF) transformation
End-member determination — locate pixels with relatively pure spectral characteristics
Pixel Purity Index (PPI) mapping
N-dimensional end-member visualization
Methods of mapping and matching using hyperspectral data:
Spectral Angle Mapper (SAM)
Subpixel classification (linear spectral unmixing)
Spectroscopic library matching techniques
Matched filter or mixture-tuned matched filter
Indices developed for use with hyperspectral data
Derivative spectroscopy

**Perform accuracy assessment**
Select method:
Qualitative confidence-building
Statistical measurement
Determine number of observations required by class.
Select sampling scheme.
Obtain ground reference test information.
Create and analyze error matrix: Univariate and multivariate statistical analysis

**Accept or reject previously stated hypothesis**
Distribute results if accuracy is acceptable
HYPERSPECTRAL REMOTE SENSING

Unmixing

Three types of mixtures exist in HSRS images: (i) linear mixing on a checkerboard type surface, (ii) nonlinear (linear plus bilinear) mixing in a two-layer medium, and (iii) nonlinear mixing in an intimate (particulate) medium (Dobigeon et al. 2014). As a result of the complexity of pixels in HSRS images, when mixing occurs, it is not possible to determine what materials are present directly from the respective measured spectral vectors. The unmixing problem addresses how materials can be quantitatively separated in every mixing vector; the objective is to precisely identify materials contained in a pixel.

HSRS unmixing provides a comprehensive quantitative mapping of elementary materials present in the acquired data. More specifically, unmixing can identify the spectral signatures of these materials (usually called endmembers) and can estimate their relative contributions (known as abundances) to the measured spectra. The main steps of HSRS unmixing cover endmember number estimation, endmember extraction, and spectra abundance resolution.

After carefully determining the amount of endmembers existing on the image, there is an array of methods which can be applied to extract the pure endmember spectrum from the data. These methods can be classified into two categories: sequential endmember extraction algorithms (SQEEAs) and simultaneous endmember extraction algorithms (SMEEAs). SQEEAs acquire endmembers separately, one by one, after each iterative procedure, while SMEEAs obtain all the spectra of endmembers at one time after the processing. The former methods suffer from accumulative errors caused by the iterations, while the latter approaches suffer from high computational costs. Having obtained the number of endmembers and their corresponding spectral signatures, estimation of their abundance in every pixel still remains a challenge due to the ill-posed inverse problem because of model inaccuracies, observation noise, environmental conditions, endmember variability, and data set size making the unmixing results nonunique (Zare and Ho 2014). Researchers have devised and investigated many models searching for robust, stable, tractable, and accurate unmixing algorithms that are related to the signal-subspace model, the geometrical model, the statistical model, the sparsity-based model, and the spatial-contextual model.

Classification

The goal of classification is to assign a unique label to each pixel vector so that it is well defined by a given class. Different from the classification problem in other remotely sensed imageries, HSRS classification has its own pros and cons. The availability of HSRS data with high spectral resolution makes even similar materials display with different spectral signatures, both those with more abundant detailed information and those at a more precise wavelength range. Take the image shown in Figure 3 as an example; only HSRS data can precisely divide the whole farm area into dozens of kinds of crops. However, accurate classification demands a high level of data preprocessing for atmospheric correction, geometric correction, and spectral calibration. Moreover, the number of available training samples tends to be insufficient due to the extremely high dimensionality of HSRS data, which decreases classification accuracy.

HSRS data classification can be conducted by either unsupervised or supervised methods. In an unsupervised classification, the identities of land cover types within a scene to be specified as classes are not generally known a
priori because ground reference information is lacking or surface features within the scene are not well defined. The computer must group pixels with similar spectral characteristics into unique clusters according to some statistically determined criteria. Therefore the unsupervised classification is also called clustering in the pattern recognition field.

In a supervised classification, analysts utilize some area of known spectral characteristics to train the classifier for eventual land-cover mapping of the remainder of the image with limited training samples. In order to address the above issues, feature mining, subspace-based approaches, and semisupervised learning techniques have been developed. In feature mining and subspace approaches, the goal is to reduce the dimensionality of the input space in order to better exploit the limited training samples. In semisupervised learning, the idea is to exploit the information conveyed through additional unlabeled samples, which can complement the available labeled samples with a certain degree of confidence.

**Change detection**

Biophysical materials and human-made features on the surface of Earth are dynamic and frequently changing. It is important that such changes be inventoried accurately so that the physical and human processes behind such changes can be more fully understood. Change detection is the process of identifying differences in the state of an object or phenomenon by observing it at different times. Remote sensing technology can observe the surface of a research area periodically for a long time at larger scales. It has been widely applied in the monitoring of land-use/land-cover research, city expansion, wetland vegetation monitoring, ice detection,
protection zone management, and disaster prevention.

Sometimes, anomaly change detection is of more significance, as shown in Figure 4. This term refers in particular to real changes other than the normal changes (seasonal changes of vegetation) found in the scene. Change detection results from HSRS data are well suited to detect subtle changes and contain detailed information. HSRS can identify not only the changed area but also “from–to” analysis information (what features were there before and after the change).

Target detection

HSRS can provide crucial spectral details of similar materials due to the high spectral resolution and continuous spectral curve in these images; therefore, it has exceptional advantages when detecting certain materials. Target detection is one of the most important applications of HSRS image processing. Typical examples include: the detection of specific terrain features and vegetation, mineral, or soil types for resource management; detecting and characterizing materials, surfaces, or paints; the detection of man-made materials in natural backgrounds for the purposes of search and rescue; the detection of specific plant species for the purposes of countering narcotics; and the detection of military vehicles for the purposes of defense and intelligence (Manolakis et al. 2014). Target detection can be also considered a classification problem as it assigns each pixel to either a target class or a nontarget class. However, these two classes are always unbalanced in number, with nontarget class pixels superior to target pixels.

Figure 4  Generalized change detection and anomaly change detection.
The general criteria for minimizing total misclassification error rates are no longer suitable for target detection tasks, as shown in Figure 5. When a target detection task is being conducted, a projection method is required to maximally separate target class pixels from other classes. HSRS data are often mapped into a hyperplane where target pixels and nontarget pixels can be best separated according to certain rules. Then, a threshold mechanism is adopted to tear them apart as shown in Figure 5. This projection operation is called designing a target detector. The state-of-the-art target detectors include, but are not limited to: detectors based on spectra unmixing, which determines a pixel by the target endmember abundance it contains; detectors based on signal matching, which determines a pixel by the correspondence to a given target referential signal; detectors based on hypothesis testing, which determines a pixel by the probability it belongs to the two opposite hypotheses with different statistical features; detectors based on feature space projection, which maps the original data cloud into a new feature space, and determines a pixel by its distance to the a priori information in this new space.

Despite the success that has been enjoyed in different HSRS techniques described in the above section, there still remain a lot of problems and challenges in HSRS information extraction processing. This entry could not cover all the interesting and relevant aspects exhaustively; however, HSRS is a mature and promising field and contributes actively with frontier cross-disciplinary research activities.

**SEE ALSO:** Microwave remote sensing; Optical remote sensing; Supervised classification; Unsupervised classification

**References**


HYPERSPECTRAL REMOTE SENSING


Island biogeography

Rosemary G. Gillespie  
*University of California, Berkeley, USA*

The flora and fauna of oceanic islands have inspired scientific interest from the moment explorers first began sailing the oceans. In the history of scientific understanding, islands played a fundamental role in providing insights into the operation of biological evolution. The Galápagos Islands played a key role in the development of Charles Darwin’s theory of evolution by means of natural selection. Likewise, Alfred Russell Wallace contemporaneously developed similar theories based on studies in the Spice Islands of Indonesia. More recently, the closed nature of insular systems, their relative simplicity, and often replicated pattern within a known temporal (geological or climatological) framework, has allowed them to serve as microcosms for understanding fundamental processes in both ecology and evolution. Thus, studies exploiting the discrete nature of islands have allowed the advancement of pervasive organizing theories of community ecology, particularly involving island biogeography and related ideas of species abundance and diversity. More recently, with the advent of accessible molecular genetic tools, research on islands has allowed unique insights into the processes that generate biotic diversity, especially the mechanisms of speciation.

Unfortunately, islands are also prime targets for biological invasions, mediated largely by anthropogenic disturbance. The severity of such impacts on island biotas may be the result of their evolution in isolation, but it is certainly compounded by their characteristically small population sizes. Moreover, while the extraordinary rates of extinction of vertebrates on islands have been well documented, extinction among arthropods is largely unknown. Indeed, the profound lack of knowledge of many arthropod groups on islands means that it is likely that whole radiations of species of insects, spiders, and likely other invertebrate groups will go extinct before they are ever even recognized.

What is an island?

An island is generally defined as a landmass, smaller than a continent, that is surrounded by water. However, from a biological perspective, it is more useful to consider an island as any area in which a set of biota is isolated as a result of being surrounded by a dramatically different habitat matrix (Figure 1). Thus, the feature that make islands biologically interesting is their isolation, and the barriers that impose isolation vary according to the organism in question. A lake in the middle of a continent may be isolated for a fish but not so much for a lizard; an oceanic island may be isolated for a gecko or an insect but not for a whale; and a rock in the middle of a forest may be isolated for an ant, though not for a pig. Thus, the level of isolation is dictated by the impermeability and extent of the isolating matrix relative to the organism in question. But, whatever the context, the well-defined nature of islands results in properties such as a microcosmal nature and a uniquely assembled biota. In essence, islands can be considered nature’s test tubes. Each island represents a trial in an
experiment and each new island is the repeat of one of these experiments.

**Types of insular environments**

Figure 1 shows a diversity of insular systems. (a) The thousands of mangrove islands in the Florida Keys have been used to test the equilibrium model of island biogeography. (b) The Hawaiian Islands, the most remote archipelago in the world, are home to numerous adaptive radiations of terrestrial organisms. (c) The Galápagos Islands off the coast of Ecuador provide textbook examples of adaptive radiation. (d) The Canary Islands have also provided many examples of adaptive radiation among plants and animals. (e) Multiple pairs of sticklebacks have been found in the glacial lakes on Texada Island, illustrating the early stages of adaptive radiation (Schluter 2000); and numerous islands in the Barkley Sound have formed the basis for many studies on island plants (Cody 2006). (f) The African Rift Lakes have provided isolated habitat for aquatic organisms, with particularly spectacular examples of adaptive radiation in cichlid fish. (g) Uluru (Ayers Rock), Australia, where inselbergs, or “island mountains,” provide isolated habitat for organisms that are limited to the unique environments provided by the higher-elevation habitats. (h) The summits of the ancient tepuis or table-top mountains in the Guiana Highlands of South America, especially in Venezuela and western Guyana, have been isolated for approximately 180 million years. (i) The Panama Canal, in which Lake Gatun was created by the building a dam across the Chagres River in 1913, with the subsequent formation of Barro Colorado Island, provides a system for many studies on the response of organisms to insularization. (j) Theories based on island biogeography have been applied to the management of fragmented forest “islands” in Sumatra, Indonesia.

**Figure 1** The diversity of insular systems: (a) Florida Keys, USA; (b) Hawaiian Islands; (c) Galápagos Islands, Ecuador; (d) western Canary Islands; (e) glacial lakes on Texada Island, Canada; (f) African Rift Lakes; (g) Uluru (Ayers Rock), Australia; (h) tepuis in the Guiana Highlands of South America; (i) Panama Canal and Barro Colorado Island; (j) forest fragmentation of the remaining tropical forests in Sumatra, Indonesia (patches of deep green) can be seen as isolated by agricultural land (patches of lighter green). All images from NASA.
“True” islands – surrounded by water

The traditional view of an island is of a landmass surrounded by water. However, a huge range of habitats can be considered within this context, from intermittent tidal islands such as Lindisfarne in the United Kingdom to the remote islands of Hawai’i, and tiny rocks and islets off Greenland, which itself is considered the largest island in the world. While the size and isolation of an island is integral to understanding its biogeography, so also is its geological history. Since the 1960s, geological understanding of islands has advanced enormously. In many oceanic archipelagoes, the age of each island is often known with some level of precision, and information of various levels of precision is generally available for changes in island size, elevation, and associated climatic factors over millennia (Gillespie and Clague 2009). At the same time, information on islands that have long since disappeared provides insights into historical patterns of connectedness, and work on the complexities of the splitting, sinking, and uplifting of the ancient landmass of Gondwana has highlighted the potential impact of the dynamic nature of island formation on patterns of biodiversity. This has given biologists a temporal framework with which to examine ecological and evolutionary processes and how biodiversity has formed over time.

Mountaintops – sky islands

The environments of mountaintops are often quite different from the surrounding slopes, and can include cloud forests, alpine grasslands, herbfields, and páramo; these sites tend to be relictual fragments of previously more widespread habitats, isolated by the changing climate. Perhaps the best-known sky islands, at least from a biological perspective, are the American Madrean “sky islands” of southeastern Arizona and New Mexico. Here, striking phenotypic divergence has been found among populations of the jumping spider Habronattus pugillis. Given that these habitats are only around 10,000 years old, diversification has been rapid.

Caves

Caves can be considered rock cavity islands, with the taxa adapted to these systems isolated both by the impenetrable rock surrounding them and by the surface environment. Among cave spiders, beta diversity tends to be higher for troglobiont assemblages, often with minimal gene flow between caves, while alpha diversity and phylogenetic and functional diversities tend to be lower. Cave-adapted species are found among multiple different kinds of organisms, and many cave species have evolved suites of morphological traits associated with cave life, including reduced or missing eyes, reduced or missing pigment, elaboration of appendages, and hypertrophy of extrasensory structures (Culver and Pipan 2009). Moreover, cave-modified species tend to show the recurrent evolution of similar forms, controlled by the same genetic pathways that can span many taxonomic levels; for example, adaptation to cave environments has been found to involve a similar mutation at the first step of melanin synthesis in both planthoppers (from Hawai’i and Croatia) and Mexican cavefish.

Habitat fragments

The process of fragmentation has resulted in insularization at many different scales across space and time. In the Australian landmass, refugial rainforests dating from the Miocene are found in small pockets along the eastern and southwestern coasts. The diversity of taxa that are confined to these habitats is extraordinarily high, with species often restricted to single mountains, deep divergence between mountains
being due to ancient fragmentation of forest habitats and subsequent diversification.

On a somewhat more recent timescale, ongoing volcanic activity on the young island of Hawai‘i has created a dynamic mosaic of habitats of different ages as large areas of forest are destroyed or fragmented when lava flows around them. Fragmentation on scales of hundreds of meters and over several hundred years can be sufficient for the genetic differentiation of arthropod populations, though – because of its transitory nature – this fragmentation is likely to be more important in maintaining genetic diversity than in the separation of populations.

Considering habitat islands that have been created within recorded history, many of the ideas originally developed for islands in the sea have been extended to these systems. Most such islands are fragments of habitats that were historically connected, such as remnant trees and forest patches. For habitat islands, as for islands in the sea, ecological and evolutionary processes are governed largely by isolation, time, and the nature of the matrix relative to the dispersal abilities of the organisms in question. Because of their discrete nature and ease of manipulation compared to islands in the sea, habitat islands have been exploited in the development of many ecological principles including those related to metapopulation dynamics and the physical design of nature reserves.

**Fragment versus de novo islands**

From a biological perspective, there are in general two types of islands (Figure 2), depending on whether they are formed by fragmentation or de novo emergence. *Fragments*, which include continental and most habitat islands, are formed by separation from a larger entity, and become fragments in which various habitats have usually been occupied prior to isolation; thus, the number of species, at least initially, will tend to decline as a result of stochastic events. *De novo islands*, which include oceanic islands and certain lakes within a terrestrial landmass, are formed without life, so colonization inevitably occurs from outside. If the islands are not extremely isolated, an equilibrium will be reached between the immigration of colonists and extinction. With greater isolation (and lower immigration rates), local endemism will increase. For a given species, there is a level of isolation at which the frequency of colonization is insufficient to allow for the occupation of available ecological niche space through migration, and which can provide the opportunity for adaptive radiation for individuals that can establish; this area has been called the “radiation zone.” Here, species diversity will arise through the formation of multiple new species adapted to exploit the available ecological space.
Neo-endemics typically form on isolated islands that have been created de novo, and have abundant empty ecological space into which those few colonists can diversify. Besides Hawai‘i, other volcanic archipelagoes, including the Marquesas, Societies, and Galápagos in the Pacific, the Canaries in the Atlantic, and the Mascarenes in the Indian Ocean, have formed de novo and provided conditions for the formation of neo-endemics. However, neo-endemics can also form on fragment islands including the Caribbean islands and the islands of New Zealand and Madagascar. After these islands formed, genetic connection with the source populations was reduced, and over time this allowed organisms to diverge genetically. Unlike islands that form de novo without any species and tend to accumulate species through time, fragment islands are usually ecologically saturated at the time of separation and tend to lose species through ecological time, a phenomenon termed relaxation. Over evolutionary time, the species on these islands may change through a process termed relictualization, with the formation of paleoendemics, usually without adaptive radiation.

Adaptive radiation can then accentuate this effect when multiple species arise from the few that successfully colonize, the result being very few families and genera compared to the overall diversity. This pattern, whereby the species diversity is represented by few higher taxonomic groups (genera, families) and disproportionate numbers of species within a few genera, has been noted in many remote islands.

Endemism

On islands formed de novo, the pattern of species accumulation will depend on the rate of formation of new species relative to the frequency of island colonization, the extreme cases leading to extensive adaptive radiation as a result of in situ evolution, with associated adaptation to occupy the available ecological space. On remote islands, the frequency of colonization becomes vanishingly rare. As a result, the biodiversity of remote islands is made up in large part by the diversification of the few initial colonists, and levels of endemism are consequently very high.

Loss of dispersal ability and flightlessness

Loss of dispersal ability is a common feature of organisms on more remote islands. The phenomenon was first documented by Darwin, who suggested that “powers of flight would be injurious to insects inhabiting a confined locality, and expose them to be blown to the sea” (Carlquist 2010). However, not all species lose their dispersal ability on islands. Indeed, loss of dispersal requires that (i) there be some dispersal behavior on which selection can act, and (ii) there be no opposing selection to maintain dispersal ability. Thus, some birds and strong flying insects can still maintain flightedness even on remote islands. Nevertheless, flightlessness in birds has evolved repeatedly on islands where
predators are absent; there are about 40 species of flightless birds known today, including the kiwi, moa, emu, cassowary, and dodo. Many of these species have been driven to extinction following the arrival of humans; notable extinctions include the dodo of Mauritius, the elephant bird of Madagascar, and the moa of New Zealand.

Innovations

The known disharmony of islands appears to stimulate the development of some unusual traits in island groups. For example, Hawaiian birds have developed a remarkable diversity of elongated and decurved bill shapes, presumably because of the availability of nectar resources. The Hawaiian silversword plant alliance includes trees, shrubs, subshrubs, rosette plants, cushion plants, and a vine that occur in some of the wettest sites recorded on Earth to extreme desert-like environments (Bramwell and Caujapé-Castells 2011). Among insects, the absence of native social insects in some of the more remote islands of the Pacific may have been largely responsible for the adaptive shift to predation in some terrestrial insect groups and the diversity of other predatory arthropods on these islands. One of the most striking innovations is the development of predatory behavior in a lineage of Hawaiian caterpillar moths in the genus *Eupithecia*. Another innovation in Hawai’i has been the development of odonates with terrestrial larvae, associated with the paucity of lakes in the upland yet moist forests of the islands. Similarly, the unusual habitat of cryptogram herbivory in beetles in the subantarctic islands is considered to be a result of the disharmony of the flora.

Size

Change in body size is also commonly observed on islands, with species showing a tendency toward size extremes, both gigantism and dwarfism, a phenomenon that has been called the “island rule” (Meiri 2009). For example, large mammal species that colonize more remote islands tend to get smaller, while small species tend to get larger. Thus, gigantism is well known among rodents, as well as many birds (often associated with flightlessness, as mentioned earlier). It has also been documented among lizards (e.g., giant tortoises of the Galápagos, Seychelles, and Mascarenes; Komodo dragon of Timor), though in some cases the effect of larger species on islands is likely due to the extinction of their mainland relatives. Among arthropods, gigantism is best known among crickets (e.g., wetas in New Zealand and tree crickets in the northwest Hawaiian Islands); however, it is also known among beetles (long-horned beetles in Fiji), earwigs (St Helena Island), stick insects (Lord Howe Island), millipedes (Madagascar), and spiders (giant sheet web spiders in Hawai’i and the Juan Fernández Islands).

Island dwarfism is best known among mammals, including elephants (e.g., different species from the islands of the Mediterranean), hippopotamids (e.g., Cyprus dwarf hippo), sloths, and hominins, as well as foxes (e.g., Channel Island fox in California), felids (e.g., the Javan tiger, the Zanzibar leopard), procyonids (the Cozumel raccoon) and bocids (e.g., Balearic Islands cave goat), and cervids (e.g., Corsican red deer). It has also been documented in various dinosaurs, with some additional examples from lizards, snakes, and birds.

Plant reproductive shifts and arborescence

In plants, reproductive changes are common and include the development of dioecy, monocarpy, developmental heterophyll, sexual dimorphism, asexual reproduction, and low rates of reproduction. Stem woodiness and related differences in growth form, habitat, and floral morphology are
also recurrent and well-known adaptations of island plants.

Species diversity and the equilibrium theory of island biogeography

Equilibrium theory of island biogeography

Larger islands contain more species. MacArthur and Wilson formalized this idea in the 1960s with the development of the equilibrium theory of island biogeography (ETIB) (MacArthur and Wilson 1967). This theory relates species and area by the formula, \( S = cA^z \), where \( S \) is species number, \( A \) is area, \( c \) is a constant measuring overall species richness, and \( z \) measures the extent to which increases in area have diminishing returns in terms of species number. Values of \( z \) tend to vary between 0.18 and 0.35; that is, doubling the species number requires increasing the area by a factor of between 7 and 100. The premise of the theory is that species diversity on an island is a balance between immigration and extinction. The rate of immigration decreases with increasing distance from a mainland source, whereas the rate of extinction decreases with increasing island size (Figure 3). As the number of resident species on an island increases, the chance of an unrepresented species arriving on that island decreases and the likelihood of extinction of any one resident species increases.

When the curve trajectories of immigration and extinction rates are plotted (as ordinates) against species richness, the species diversity of a given island can be assessed by projecting to the abscissa of the curves’ intersection point, while the rate of turnover is reflected by the projection to the ordinates. Thus, species richness, together with immigration, extinction, and turnover rates, are all island-specific parameters that change according to the area of the island and its isolation, with species richness remaining roughly constant over time, while species composition is continually changing (Fernández-Palacios 2009). Thus, the predictions of the model are that (i) the number of species on an island should change little once the equilibrium is reached; (ii) there should be continual turnover of species, with some becoming extinct and others immigrating; (iii) small islands should support fewer species than large islands; and (iv) species richness should decline with remoteness of the island because islands farther from the source will have lower rates of immigration.

Figure 3  MacArthur and Wilson’s equilibrium theory of island biogeography, showing how (i) immigration, which is dictated by distance from a source of propagules, declines with the number of species present; and (ii) extinction increases with the number of taxa present, the rate declining with island area. \( S \) (species number at equilibrium) is the point where immigration and extinction curves intersect. Species numbers for a small far island (\( S_1 \)) are lowest; those for a large near island (\( S_4 \)) are largest. The turnover of species (\( T \)) can be determined from the y-axis: the turnover is lowest for a large far island (\( T_1 \)) and highest for a small near island (\( T_4 \)).
The equilibrium theory marked a substantial turning point in biogeographic understanding in general, and has played a key role in many different aspects of ecology and evolutionary biology since its publication (Losos and Ricklefs 2010). While there is widespread support for the basic tenets of the theory, in some cases it has been found that species numbers are affected by both area and isolation, whereas others have shown that an equilibrium does not exist, or that parameters other than area per se may dictate species richness. Such factors include habitat diversity, climatic conditions, island age, and even the status of knowledge concerning the presence of resident species. However, the theory has proved to be remarkably useful and, although it was developed for islands, it has had relevance for the study of many kinds of ecological communities.

Island colonization in the context of the ETIB

Given a suitable abiotic environment, the success of any given colonization event will depend largely on the availability of ecological space. Detailed studies of immigration patterns have been conducted on mangrove islands in the Florida Keys following artificial sterilization. Species of insects and spiders accumulated to an equilibrium number, though turnover of species was not randomly distributed between species – particular types of species were more likely to colonize or to go extinct. These studies supported theoretical work showing that the initial successful set of colonists is somewhat predictable based on the niche space available, and reaches a noninteractive equilibrium; however, over time, continued immigration and extinction can produce sets of species that are increasingly co-adapted, reaching interactive equilibria, each successive equilibrium having more species because the more co-adapted the assemblage the more species can co-occur. Recently formed volcanic islands, such as Krakatau, following the 1883 eruption (Thornton 1997), and Surtsey in 1963, have been used to calculate immigration rates on ecological timescales.

Long-distance dispersal, which is well known to play a role in the colonization of remote islands, is more difficult to study because of its infrequent and unpredictable nature, and therefore precludes development of testable hypotheses (Gillespie and Baldwin 2009). However, while a single rare long-distance dispersal event may indeed be impossible to predict, an understanding of the mechanisms involved in long-distance dispersal over extended time periods can lend predictability to the process. In particular, assuming that the three primary vectors for long-distance dispersal are wind, birds, and ocean currents, the likelihood of a propagule to be vectored will depend on its mechanism of association with the vector, coupled with its ability to withstand the environment to which it is exposed during transit, attributes that impose strong filters to the kind of organisms involved.

Factors facilitating establishment

Features of the new environment that affect the ability to establish will impose additional filters, with successful colonization and establishment being more likely in environments that approximately match the source environment. For example, organisms from higher latitudes are most likely to become established at lower latitudes in higher-elevation habitats with climatic conditions matching their original environment. Similarly, lower elevations at low latitudes should be more easily colonized by taxa from tropical sites. In addition, for taxa that live inside other organisms or are associated with particular substrates, it seems likely that establishment will
occur more readily if they arrive with (or after the establishment of) their substrate, symbiont, or host. Surprisingly, even some highly specialized mutualisms are maintained (or recovered) following the independent colonization of remote islands by the different partners. An additional factor that may contribute to successful establishment is escape from predators, parasites, and competitors, a phenomenon well documented through studies of successful invasions where predators are lacking.

Niche pre-emption

In a given area, once a niche has been filled, it appears to be more difficult for ecologically similar individuals to enter (MacArthur and Wilson 1967). Lineages of plants that are diverse in the Canary Islands are monophyletic, while those that are less diverse have colonized the islands repeatedly, suggesting the possible importance of niche pre-emption. Niche pre-emption also may contribute to the “progression rule” of successive dispersal from older to younger islands in the Hawaiian Islands (Wagner and Funk 1995).

Ecological release

What happens to a population subsequent to successful colonization? Subsequent to the initial establishment of species in new habitats, a common outcome is ecological (or competitive) release, with colonizers expanding their range to fill the available ecological space. Clearly, the extent and duration of any ecological release will depend on the isolation of the habitat. On islands that are close to a source of colonists, the niche space will be filled quickly by diverse taxa; on isolated islands, the infrequency of colonization allows the initial colonists to “explore” the ecological space much more broadly.

Regular cycles of distributional change following colonization of islands have been proposed several times in the literature, starting with the idea suggested by E.O. Wilson for a “taxon cycle” in Melanesian ants in which widespread, dispersive populations give rise to many more restricted and specialized populations or species. While the tendency is for species to expand their range to fill the available ecological space on colonization of a new island, the consistency of subsequent distributional changes associated with the taxon cycle has been found to be less predictable.

Evolution and the ETIB

The role of evolution within the dynamic of the equilibrium theory has only recently been examined, and shows that larger islands tend to have higher diversification rates (i.e., in situ speciation minus extinction) compared to smaller islands, and species diversity changes in concert with the geological ontogeny of an island (Whittaker and Fernández-Palacios 2007). Following on from these concepts, theoretical expectations for phylogenetic diversity–area curves using data from ecological community assembly models, together with macroevolutionary models, have revealed strong relationships between regional phylogenetic diversity and area, provided that species were derived only from within-region speciation. Moreover, there is support for the notion of an equilibrium diversity over evolutionary time, as species richness has been shown to be determined by island-specific limits on total diversification (Moen and Morlon 2014). However, the nature of equilibrium diversity – how it differs between lineages, and how it is attained and maintained – is still not clear. Studies using island chronology to look at changes in diversity over the course of adaptive radiation have indicated that diversity on islands...
does not change linearly over time; indeed, some lineages appear to reach their highest diversity early during the course of adaptive radiation, while others accumulate species more slowly, the result being that the oldest islands have the highest diversity (Gillespie and Baldwin 2009).

Evolution on islands

Anagenesis versus cladogenesis

On islands that are within the geographic distance to which a species is likely to disperse, the genetic coherence between populations from the source and those on the island can be maintained. In such circumstances, new species will tend not to form on the island (unless there is very strong selection driving divergence). On islands that are beyond the range within which populations can maintain genetic contact with source populations, then the island population will diverge to form a new species through the process of anagenesis (Whittaker and Fernández-Palacios 2007).

Cladogenesis and adaptive radiation

Where islands are farther from a source of colonists, fewer representatives from the entire source community will establish within the same time period. As a result, there is sufficient time for diversification; thus, cladogenesis within the island can lead to the formation of new species. The phenomenon whereby single (or few) colonists, isolated genetically from their source population, give rise to a series of ecologically disparate species, often through a rapid burst of speciation, is termed adaptive radiation. Adaptive radiation tends to be limited to islands that are beyond the so-called radiation zone, or normal range of dispersal, of a given organism. Species that form through adaptive radiation are typically neo-endemics, as described above, formed in situ and found nowhere else.

Adaptive radiation tends to be associated with the formation de novo of extremely isolated habitats, with subsequent colonization being very infrequent, giving the few successful colonists sufficient time to “explore” the ecological space available, diverge, and diversify into multiple species. Well-known examples of adaptive radiation include that of finches in the Galápagos Islands, which was likely triggered by the appearance of land through volcanic activity over the last 3 million years. In the same way, adaptive radiations in the Hawaiian Islands are found to be mostly associated with the volcanic activity that resulted in the initial formation of the current high islands approximately 5 million years ago. And the origin of the adaptive radiation of the African rift lake cichlid fish approximately 5 million years ago coincided with a period when rivers in the area became increasingly swampy.

For less dispersive organisms, adaptive radiation can occur on island archipelagoes that are relatively large and not extremely remote. Spectacular radiations among mammals are known among primates in Madagascar and rodents in the islands of the Philippines; among frogs in the islands of Southeast Asia, Madagascar, and the Caribbean; and among lizards in the islands of the Caribbean (e.g., Losos 2009). Each of these lineages has provided impressive evidence of adaptive diversification. For example, cloud rats in the Philippines provide one of the most spectacular cases of adaptive radiation by mammals anywhere in the world, with species ranging in size from 2.6 kg to 15 kg (Heaney and Goodman 2009).

For more dispersive groups, such as many arthropods and birds, adaptive radiation is largely limited to the most isolated islands. The most thoroughly documented instance of adaptive radiation is that of the finches in the Galápagos.
However, the Hawaiian honeycreepers are another radiation of finches that shows a good deal more ecological diversity with an associated larger number of species, although many have already suffered extinction. For arthropods, particularly noted examples of adaptive radiation are found in the Hawaiian Islands and include *Drosophila* flies, which are well known for their diversity of mating behaviors; lineages of crickets that have diversified explosively in song repertoire; sap-feeding planthoppers and psyllids that have proliferated by tracking and switching between plant hosts; and beetles that have formed new species on different substrates. Diversification may follow a predictable pattern, at least in some groups; for example, among *Tetragenatha* spiders similar ecological sets of species have evolved over and over again on each of the different Hawaiian Islands.

Among plants, the largest known island radiation is that of the Hawaiian lobeliods (Campanulaceae), with the >100 species now thought to have arisen from a single colonization event (Stuessy and Ono 2007). The radiation exhibits extraordinary diversity in vegetative and flower morphology, with species inhabiting a huge array of habitats. The diversity is considered to have arisen in concert with that of the Hawaiian honeycreepers, the lobeliods displaying a suite of morphological characteristics associated with bird pollination, including deep, tubular, long-lived inflorescences, with an abundance of nectar and no odor. Another spectacular example of adaptive radiation in plants is the Hawaiian silversword alliance (Asteraceae–Madiinae), in which 28 species are known that display a huge diversity in life form, from trees to shrubs, mats, rosettes, cushions, and vines, occurring across habitats from rainforests and bogs to desert-like situations. An analogous radiation of 23 species in the genus *Argyranthemum* (Asteraceae–Anthemidioideae) has occurred in the Macaronesian islands, although the largest radiation of plants in the Canaries is that of the succulent, rosette-forming species of the genus *Aeonium* (Crassulaceae).

Some of these radiations provide spectacular examples of the rapidity of adaptive diversification. For example, the estimated rate of diversification for Hawaiian crickets in the genus *Laupala* is 4.17 species per million years; the rate estimated for cichlid fish, which have undergone adaptive radiation in the African rift lakes, is 2.02–2.09 species per million years in Lake Malawi and Lake Victoria.

Most island adaptive radiations occur within an archipelago setting, with allopatry between islands implicated in providing sufficient isolation for adaptive radiation. Interesting in this regard is a radiation of small flightless weevils in the genus *Miocalles* (Coleoptera: Curculionidae: Cryptorhynchinae) on the single small island of Rapa in the southern Australs of French Polynesia. Rapa is home to almost half of the 140 species that occur across the western Pacific and Australia. Here, the beetles collectively feed on 24 genera of native plants, and show varying degrees of host specificity, thus utilizing almost all genera of native plants found on Rapa.

Some radiations have been studied because they are readily accessible to scrutiny of the process of adaptive radiation, in part because they have occurred very recently. In this context, Dolph Schluter’s (2000) research has focused on stickleback fish of the deglaciated lakes of coastal British Columbia, Canada. These lakes harbor a number of sibling species of fish, and repeated co-occurrence of pairs of species has been attributed to novel ecological opportunity provided by deglaciation and recolonization. In particular, the threespine stickleback, *Gasterosteus aculeatus*, is a species complex which has diversified in each lake such that no more than two species occur in any one lake. Interestingly, it
appears that pairs of species in different lakes have evolved independently of other pairs, and species have diverged as a result of parallel episodes of selection for alternative feeding environments. Research in this system has highlighted the role of divergent natural selection as a mechanism underlying adaptive radiation.

Isolation, hybridization and admixture

In addition to rapid genetic and ecological differentiation, there are elements that have become increasingly associated with adaptive radiation – in particular, the potential role of multiple colonizations and admixture in enhancing variability. While a break in genetic connectance is necessary for adaptive differentiation, hybridization and genetic admixture are key in the generation of adaptive variation and functional novelty (Seehausen et al. 2014). There have been a number of studies demonstrating how the negative effects of genetic founder effects may be offset if different colonization events result in multiple genotypes within the introduced population, highlighting the potential role of admixture among successively introduced populations in providing the genetic variation to allow adaptive evolution.

Parallel evolution and convergence

Many of the best-known island radiations are characterized by repeated shifts in phenotype and speciation, which are often associated with marked morphological and ecological convergence. This phenomenon, in which similar sets of ecological forms have evolved largely independently, was first documented in Anolis lizards (Losos 2009), but has now been shown in a number of adaptive radiations, including the sticklebacks mentioned earlier, as well as spiders in the islands of Hawai’i (Figure 4). It has been suggested that such examples provide powerful evidence that Darwinian natural selection acting on organisms exposed to similar adaptive landscapes can sometimes overwhelm historical and ecological contingencies (Losos 2009). Therefore, diversification during adaptive radiation may occur along predictable ecological axes and in a predictable sequence.

Conservation

Although islands have long proven themselves as extraordinary laboratories for studying processes associated with the generation of diversity, they are now contributing to our understanding of processes leading to the loss of diversity. The vulnerability of islands to extinction is now widely recognized. Islands have suffered much higher rates of extinction than continents (Steadman 2006). Given the information on species diversification that scientists have been able to determine from oceanic islands, the question is whether we can also use the islands as tools to understand extinction. Extinction is clearly more gradual in continents than on many islands, but the trajectory is similar. Accordingly, understanding of the processes involved in species extinction in the islands, and the replacement of largely native environments by species from elsewhere, may allow the development of effective conservation strategies in mainland habitats.

Major issues in island conservation include the lack of knowledge of the biota, the generally small size of the populations making them more vulnerable to extinction, and the large numbers of invasive species on islands. These are basically the same conservation issues as confront continental biotas; they are simply more acute on islands because of their small size and microcosmal nature.
Figure 4  Ecomorphs for Hawaiian *Tetragnatha* spiders, illustrating the repeated evolution of ecomorphological similarity. (a) Three of four known ecomorphs of the spiny leg clade: species shown from different ecomorphs on the same island are more closely related to each other than to the same ecomorph on different islands. (b) Representatives of different web morphologies of the web-building lineage where different web types on the same island are more closely related than the same web type on different islands (inset photos show the species that build the webs). Photos: (a) Todd Blackledge; insets R. Gillespie; (b) R. Gillespie except *T. kamakou*, Darko Cotoras.
ISLAND BIOGEOGRAPHY

Taxonomic impediment

The biota of islands is often unique; for example, the islands of the Pacific have been designated a biodiversity hotspot. Assessing this diversity, particularly for arthropods, is problematic. The major impediment is a lack of taxonomic understanding of arthropods on many islands, particularly those that are more remote. New species are being collected at a remarkable rate in areas such as French Polynesia, Madagascar, and even the relatively well-studied Canary Islands, New Zealand, Hawai’i, and the Galápagos, yet the training of arthropod systematists has lagged behind.

Size of distribution and related impacts

Species on islands tend to have a limited geographic distribution, and the naturally small population sizes means that most of these populations will not suffer high levels of inbreeding depression. However, they are vulnerable to extinction as a result of demographic accidents. In the tropical islands of the Pacific, extinction has been much higher than on the islands of the North Atlantic, where ranges are much broader. Species with naturally small population sizes are similarly more vulnerable to habitat modification, simply because loss of even a small amount of habitat for a geographically restricted species could easily reduce numbers below sustainable levels.

Islands can also be highly climate-sensitive, and a major concern is that climate change will compress ranges further, leading to increased extinction rates of mountain species that have limited opportunities for migration: the high-elevation cloud forests that characterize many oceanic islands have a narrow geographical and climatological niche that may be eliminated with even a slight climatological change. Climate change may also effect declines in forests due to floods, droughts, or increased incidence of pests, pathogens, or fire. In addition, increases in the frequency or intensity of hurricanes may cause disturbance that favors invasive species. Sea level rise is also a concern for oceanic islands, resulting in coastal erosion and salt water intrusion into freshwater lenses. However, it is the low-lying islands and atolls that are the most vulnerable to these effects.

Invasive species

The isolation of islands has makes them more vulnerable to invasive species and other stresses, with extinctions attributed to alien species invasion having been noted in several archipelagoes. However, alien species in the Hawaiian Islands, perhaps because it is the most remote archipelago, appear to have taken a much greater toll on the native biota. The numbers now estimated for animals purposely or accidentally introduced into Hawai’i and established are 3046 arthropods, 20 reptiles, 46 land birds, 19 mammals, and 927 plants. Multiple parameters may interact to dictate invasion success. The most important of these are (i) species diversity itself, which can serve as an ecological barrier to invasion – this effect has been used to explain the higher frequency and impact of invasions on islands; (ii) disturbance, and the opening of ecological space; (iii) propagule pressure, in particular the relative abundance of native versus non-native propagules; (iv) characteristics of propagules (more generalist in habitat requirements); and (v) novelty of perturbations and the “naivety” of native biota.

It is widely known that the worst kind of introduction to an isolated archipelago is that of a generalist predator or competitor capable of exploiting a broad array of habitats and causing secondary impacts. This can sometimes result in an effect known as “invasional meltdown.” For example, the crazy ant Anoplolepis gracilipes has
invaded Christmas Island, resulting in a complete modification of the ecosystem in just two years.

Devastating introductions of generalist predators or competitors such as pigs and goats mostly happened before protocols were established to mitigate the introduction of non-native species. However, abundant, small human commensals, such as ants and mosquitoes, are continually expanding their range to oceanic islands.

Conclusion

Islands have long provided intriguing insights into almost every aspect of biogeography, and indeed have spawned foundational understanding of ecological and evolutionary principles. The current rapid change affecting these systems is of immediate relevance, alerting us to the need to develop a more complete understanding of the biodiversity within these systems, to make use of the opportunity for understanding the shifting dynamic, and, most critically, to explore ways of mitigating impacts and potentially restoring ecosystems before they go beyond a state from which they will no longer be able to return.

SEE ALSO: Alien and native species; Biodiversity; Endemism; Habitat destruction and fragmentation; Mountain biogeography; Zoogeography

References


Israel: Ha’Aguda Ha’Geographit Ha’Yisraelit (Israeli Geographical Association (IGA))

Founded: 1961
Location of headquarters: Beersheba
Website: www.geog.bgu.ac.il/iga/index.html
Membership: 153 (as of 2014)
President: Tal Svoray
Contact: tsvoray@bgu.ac.il

Description and purpose

Ha’Aguda Ha’Geographit Ha’Yisraelit (IGA) is the professional association for Israeli geographers. Members of the association deal with all areas of human and physical geography. The association has been operating since 1961 and its aims are as follows: to promote knowledge and research of geography in Israel, to increase awareness of the theoretic and practical achievements of geography in Israel, and to present these achievements to geographic organizations and bodies. The IGA aims to promote its activities through knowledge and research of geography; in addition, it strives to further the professional position of geographers and to increase public awareness of their professional contributions, to promote the study of geography in Israel, to strengthen the connections and exchange of knowledge between organizations that deal with geography, and to increase cooperation between geographic departments at universities in Israel and around the world. In addition, members of the organization present Israeli geography in international forums.

Journals or major publication series

Ofakim Be’Geographia (Horizons in Geography)
HaReshet Ha’Geographit (Geographic Network)

Current activities or projects

The IGA conducts professional seminars throughout the year which promote the discussion and study of theoretical issues regarding geographic space and the factors that affect it. The association holds an annual conference that is attended by geographers from across Israel and around the world. The conference is held at a different university every year and includes the conferring of awards for distinct contributions to geography and geographic education. The IGA also oversees two geographic journals and maintains a website and mailing list.

Brief history

The association was founded in 1961. Its first president was Professor David Amiran, who later received the Israel Prize for his contribution to research in Israeli geography.

Submitted by Oren Ackermann
Description and purpose

The AGeI aims to stimulate, promote, and coordinate geographical research, initiatives for the training of geographers, and for the reproduction of a culture of the territory in Italy. Its members are geographers who work or have worked in Italian universities.

The association achieves these goals through its journal *Geotema* and the organization of the Italian Geographical Congress, the Days of Geography, and the Interuniversity Geographical Excursion. The AGeI coordinates 16 national research groups and seven commissions and is a member of the Association of Geographical Societies in Europe (EUGEO).

Journals or major publication series

*Geotema.* www.associazionegeografitaliani.it/geotema.html

Current activities or projects

The AGeI sets research programs and promotes 16 research groups in collaboration with national and international institutions, coordinates Italian participation in the International Geographical Congress, supports the action of the other national associations engaged in spreading the culture of the territory, sustains the activities of the Italian professional geographers and the work of young researchers also through research grants and awards, improves the conditions of teaching geography in schools at all levels with special focus on universities, and publishes the directory of Italian geographers and the Italian Department of Geography.

Brief history

In 1978, under the leadership of the first President Giacomo Cora Pellegrini and faced with the growing need for skills and geographical knowledge, the association of the Italian academic geographers (AGeI) was founded in Rome. One of its first initiatives was the organization of the congress Geographic Research in Italy 1960–1980, a critical analysis of the contribution of Italian geographers in research.

The AGeI replaces the CoGeI which included only full professors in its membership. The new association fostered the spread of innovative theories and methodologies in Italian geography. Previous presidents of the AGeI included Adalberto Vallega, Berardo Cori, Marcello Zunica Luciano Lago, Alberto Di Blasi, and current president Franco Farinelli.

Submitted by Andrea Riggio
Italy: Società di Studi Geografici (Society for Geographical Studies)

Founded: 1895
Location of headquarters: Florence
Website: www.societastudigeografici.it
Membership: 500 (as of 2014)
President: Lidia Scarpelli
Contact: info@societastudigeografici.it

Description and purpose

The main aim of the Società di Studi Geografici is the promotion and support of initiatives which will serve the progress and spread of the geographical disciplines. To this end the society publishes Rivista Geografica Italiana (Italian Geographical Review) and Memorie Geografiche (Geographical Memoirs), organizes study meetings, and participates in national and international research programs. Most members are university lecturers and school teachers.

Journals or major publication series

Rivista Geografica Italiana (Italian Geographical Review). www.rivistageograficaitaliana.it
Memorie Geografiche (Geographical Memoirs) – 1907 to 1919, 1995 to present. www.societastudigeografici.it/italiano/pubblicazioni.html

Current activities or projects

The society organizes and participates in a range of meetings, seminars, and conferences in Italy and beyond. The society publishes Rivista Geografica Italiana (Italian Geographical Review) and Memorie Geografiche (Geographical Memoirs). The society has a library with more than 20,000 books, over 500 scientific reviews, and a rich selection of maps.

Brief history

The society’s origins go back to the establishment of the Florentine section of the Italian Africa Society in 1884 which encouraged research in regions of Africa, such as Ethiopia where a peaceful settlement movement was underway, subsequently transformed into military conquest. From 1885 to 1895 the Florentine section of the Italian Africa Society published a bulletin (Bullettino della Sezione Fiorentina della Società Africana d’Italia) dealing with questions regarding Africa, its geographical description, and the civil, political, and economic activities that Italy was called upon to carry out there.

The organization became the Società di Studi Geografici e Coloniali (Society for Geographical and Colonial Studies) in 1895 and the bulletin was replaced by the Rivista Geografica Italiana (Italian Geographical Review). The association’s name was changed to the Società di Studi Geografici (Society for Geographical Studies) in 1936. Between 1907 and 1919 the society published a series of extempore studies known as Memorie Geografiche (Geographical Memoirs). A new series started in 1995.

Submitted by Lidia Scarpelli
Italy: Società Geografica Italiana (Italian Geographical Society)

Founded: 12 May 1867  
Location of headquarters: Rome  
Website: www.societageografica.it  
Membership: 924 (as of 2014)  
President: Sergio Conti  
Contact: segreteria@societageografica.it

Description and purpose

The society is a nonprofit organization, protected and supervised by the state, whose fundamental purpose is the advancement of science and geographical knowledge. To this end, the Società Geografica Italiana promotes and favors the progress of geographical studies with particular focus on knowledge of the territory, scenery, and environment and the safeguard of cultural, environmental, and scenic heritage. Moreover, it promotes the advancement of geographic culture in Italy, initiating public meetings, conferences, excursions, educational trips, films, conventions, roundtables, and so forth.

Journals or major publication series

The society's house organ is the *Bollettino della Società Geografica Italiana*, the oldest Italian geographic periodical, published every three months since 1868.

In recent years, it has been issuing an annual report, *Rapporto annuale della Società Geografica Italiana*, which is an important event aimed at a wide and qualified spreading of geographic culture. The 2003 edition focused on migration; the 2004 edition on transport in Italy; the 2005 edition on relations between Italy and the Mediterranean region with an in-depth view of environmental, economic, and political topics; the 2007 edition on tourism, cultural tourism in particular; the 2008 edition on cities; the 2009 edition on landscapes; the 2010 edition on the "northern question"; the 2011 edition on the "southern question." The annual report is always presented to the Italian parliament.

The book *Riflessi Italiani* (Italian reflections), published for the 2004 International Congress of the International Geographical Union of Glasgow, has so far been published in three languages.

Other publications: *Memorie della Società Geografica Italiana*, *Ricerche e studi della Società Geografica Italiana*, and *Geo Italy*.

Current activities or projects

In 2007, to celebrate its 140th anniversary, the Società Geografica Italiana organized the opening edition of *Esplorazioni*, a film series on travel. The events in 2008 included an exhibition on Chinese items (February–May), in particular of the Miao minority group, hosted in the Rome venue of the Museum of Oriental Arts, owned by the society.

Several events were held in 2009, for example, the Travel Literature Festival, and many other initiatives in line with the focus of the association. In 2010, the events included the third annual event of the Travel Literature Festival and the fifth forum on geographical books about the Piedmont region. In 2011, the events were focused on the celebration of the 150th year of the Constitution of the Unification of Italy through displays, conferences, and shows. The fourth annual event of the Travel Literature Festival was dedicated to the theme of Italy and the
Italian travelers to foreign countries, while the sixth forum on geographical books was about the Latium region. In 2012, the fifth annual event of the Travel Literature Festival covered several countries, among which Afghanistan and Jordan, and the seventh forum on geographical books covered Tuscany. In 2013, the sixth annual event of the Travel Literature Festival covered the Americas, having as hosting countries Brazil and United States.

Activities have increased with five events held every month, while about 12 volumes are published annually.

**Brief history**

The society promotes and supports all studies aimed particularly at knowledge of the national territory and all other regions on Earth. The society maintains a highly specialized library, making it the most important in Italy and among the most prominent in Europe. At the moment it houses over 400,000 books, which include various ancient texts and numerous valuable printed works. The map collection is of considerable historic and scientific value and equally impressive. A complete reordering and computerized filing of all maps have recently been undertaken. Among the most valuable maps is the famous nautical map of the Genoese school, drawn up by Albino de Canepa in 1480.

The society’s archive, which contains original documents of considerable importance, also has an important photographic collection, which has recently been reorganized and currently numbers some 120,000 photographs divided into two sections, namely historic and contemporary.

Submitted by Sergio Conti
Ice caps

Frank Paul
University of Zurich, Switzerland
A.L. Ramanathan
Arindan Mandal
Jawaharlal Nehru University, India

Definition and terminology

According to the classification scheme of the United Nations Educational, Scientific and Cultural Organization (UNESCO), an ice cap is a dome-shaped ice mass with radial flow (UNESCO 1970). It covers the surface topography completely and is smaller than 50,000 km². In particular, an ice cap has to be distinguished from the much larger ice sheets in Greenland and Antarctica (ice masses of continental size) and so-called ice fields that might have a similar shape and size but neither show radial flow nor obscure surface topography completely (i.e., their (lateral) extent is limited by topography). Unfortunately, in nature we find a continuous transition between ice caps and ice fields, making a clear assignment sometimes difficult (Figure 1). Moreover, both types can have outlet glaciers that might not have a clearly defined accumulation area but are confined in their lower parts by surface topography (e.g., glaciers of the Southern Patagonian Icefield).

On a more semantic level, the terminology is confused with popular terms such as “polar ice caps,” which generally refers to the two ice sheets but sometimes also includes arctic sea ice cover and excludes the Greenland ice sheet. In this regard, the use of the term “glaciers and ice caps” for the related essential climate variable (ECV) by the Global Climate Observing System (GCOS 2003) to distinguish these ice masses from the two ice sheets caused confusion for those not familiar with the UNESCO classification and will be changed.

Ultimately, ice sheets also originate from compressed snow (i.e., consist of glacier ice), have outlet glaciers, and are surrounded by so-called local or peripheral glaciers and ice caps (GIC), which might merge with the ice sheet (in the accumulation region) or its outlet glaciers (in the ablation region) to a variable degree (Figure 2). As a consequence, the recently compiled inventory for all local GIC on Greenland (Rastner et al. 2012) introduced three different connectivity levels with the ice sheet (CL0: none, CL1: weak, and CL2: strong) to be used by the community for reference (e.g., to avoid double counting of calculated mass changes). Such a separation is not yet in place for the GIC in Antarctica (in particular those located on the Antarctic Peninsula) and the last Intergovernmental Panel on Climate Change (IPCC) report (Vaughan et al. 2013) only considered the completely disconnected ice masses on the surrounding islands (Bliss, Hock, and Cogley 2013). In order to avoid the confusion in terminology mentioned above, the most recent IPCC...
ICE CAPS

Figure 1  This satellite image shows an ice cap (cold ice) on Ellesmere Island (Canada) with a complex shape and several outlet glaciers. While the ice dome on the right is rather well defined, in the left part some topography is visible constraining the extent of the ice flow. The ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) sensor acquired the false color infrared image on July 29, 2000. Vegetation is red, bare terrain is brown, ice is grey, snow is white, water is dark blue, and water-saturated snow (slush zone) is light blue; the image extent is 40.4 by 27.4 km. National Aeronautics and Space Administration.

Figure 2  The Greenland ice sheet (in the background) with some of its outlet glaciers and various local ice caps of differing complexity that are connected to the ice sheet to a variable degree (the large ice cap in the foreground is not connected). Photo taken by the author on July 9, 2012.

assessment report (AR5) only distinguishes between glaciers (including ice caps, ice fields, and other glacier types) and the two ice sheets (Greenland and Antarctica). As ice caps are not yet indexed consistently in the global glacier inventory, they have not been counted and their global distribution cannot be shown on a map.
Forms and shapes

Ice caps can have nearly any form and size, from circular (Figure 3a) to completely irregular (Figure 3b), with sizes ranging between 0.01 and about 20,000 km². Thereby, the required radial flow sometimes allows discrimination of distinct ice domes that strictly speaking are individual ice caps despite merging. A well-known ice cap with at least three major separate domes is the Jostedalsbreen Ice Cap in southern Norway (Figure 3b). While most ice caps are rather flat, pancake-shaped ice masses with limited variability in surface slope, a considerable number also have distinct outlet glaciers that can be rather steep. In this case, the flat area of the ice cap provides a comparably huge accumulation area for the emerging outlet glaciers. Ice caps covering volcanoes are somewhat different in this regard, as the central dome-shaped part is often small (only filling a central crater) and the outlet glaciers covering the flanks of the volcano have their own accumulation areas (Figure 4). According to the UNESCO classification, small ice aprons and corniche-type structures can also be listed here as they might exhibit radial flow (with a nonflowing central part). Compared to the other shapes, they have generally steep surface slopes and might even exhibit calving ice cliffs. For the purpose of this overview, the above shapes are assigned the letters A (only flat, no outlet glaciers, circular shape), B (flat with steep outlet glaciers, irregular shape), C (covering volcanoes), and D (ice aprons).

At least one of the above four ice cap shapes can be found in all glacierized regions. While type D is present in all mountain ranges, types A and B are found in Arctic Canada (e.g., Devon, Barnes, and Penny ice caps), Greenland (e.g., Sukertoppen, Flade Isblink), Iceland (e.g., Vatnajökull), Svalbard (e.g., Austfonna and Vestfonna), the Russian Arctic (e.g., Novaya and Severnaja Zemlja, Franz Josef Land), the Tibetan Plateau (e.g., Purogangri), the tropical Andes (e.g., Quelcaya), Southern Patagonia (e.g., Gran Campo Nevado), and several sub-Antarctic (e.g., Kerguelen, Bouvet Island) and Antarctic islands (e.g., King George and James Ross islands). The volcano type C is common in the Rocky Mountains of North America (e.g., Cascade Range) and the Andes in South America, but can also be found in Iceland (e.g., Vatnajökull), New Zealand (e.g., Ruapehu), the Caucasus (e.g., Mt Elbrus), and on top of Kilimanjaro in Africa. The wide range of geographic locations also has an impact on thermal conditions of the ice. As for glaciers, ice caps in the high Arctic or at very high locations are cold (e.g., Flade Isblink, Devon, Colle Gnifetti Saddle), while at lower elevations or latitudes the ice could also be polythermal (e.g., Quelcaya) or temperate (e.g., Vatnajökull).

The example images shown in Figures 1 to 5 illustrate the above points and the various shape types. While the ice cap in Figure 3a (located southwest of Tsast Ula in Mongolia) shows a simple type A shape clearly revealing the radial flow pattern, the example from Liverpool Land (eastern Greenland) in Figure 5a also shows the radial flow pattern but has at least one distinct outlet glacier with its extent constrained by topography (type B). The examples in Figures 3b (Jostedalsbreen in Norway) and 5b (on Baffin Island) have a much more complex shape with several ice domes and outlet glaciers but also some topography (rock outcrops) visible. Both images also show several smaller and sometimes less complex shaped ice caps of type A. Separating these ice caps into distinct hydrologic drainage divides (as required for a glacier inventory) is rather challenging and might not even be appropriate for type A ice caps. The volcano in Figure 4, Cotopaxi, is topped by a typical type C ice cap with several steep outlet glaciers covering...
Figure 3  (a) A small ice cap (approximate diameter is 2.4 km) located to the southwest of Tsast Ula in the Altai Mountains (Mongolia) that is largely snow free thus revealing its radial flow pattern. The image is a screenshot from Google Maps and was acquired by the Quickbird satellite on August 5, 2006; (b) The southern and central domes of the Jostedalsbreen Ice Cap (Norway) with its several outlet glaciers and surrounding smaller ice caps. The image is 36 by 39 km in size and was acquired by the Landsat TM sensor on September 16, 2006. United States Geological Survey.
Figure 4 The ice-capped volcano Cotopaxi (5897 m) is located in Ecuador near the equator. Several individual glaciers emerge from the central ice cap. Reproduced by permission of X-ploreGroup (http://x-ploregroup.com/wp-content/uploads/2014/05/cotopaxi-22.jpg).

the flanks of the volcano. Very common to all ice caps and their outlet glaciers is the missing debris cover on the surface due to the missing ice-free rock walls that would provide debris when located above the ice cap.

Climate sensitivity

Ice caps (types A and B) have some special characteristics in a climatic sense, making them particularly vulnerable to melting when temperatures increase. The problem is that they receive accumulation in the form of snow only due to the elevation of their highest area rather than from much higher rock walls extracting extra amounts of snow when it falls. As the elevation range of ice caps (type A) is small, temperatures that are only slightly higher can remove the entire snowpack from the previous winter (and maybe further firn layers from previous years). The decreased albedo causes increased absorption of solar radiation and thus stronger melt. The resulting enhanced mass loss causes a further elevation reduction of the entire ice cap. The two ice caps shown in Figures 3a and 5a are in this situation, as well as the smaller ice caps shown in Figures 1, 3b, and 5b. With the melting of their highest points, a reinforcement feedback starts, as temperatures at lower elevations are higher causing stronger melt in summer and a reduction of solid precipitation in winter. At the same time, direct radiation is also melting more ice due to the lower albedo of bare ice compared to snow. As long as a complete loss of snow cover only occurs occasionally, the ice cap might still recover and persist. But once the reinforcement feedback is in effect, the ice cap will inevitably disappear. On Baffin Island in arctic Canada such an acceleration of mass loss might have already started for some ice caps (Gardner et al. 2012). The larger ice caps in Figures 3b and 5b are still snow covered at their highest elevations, so they will be more stable in the long term.

A further point to consider is the ice thickness distribution (Figure 6). Ice caps (type A) are very flat over large parts of their area, resulting in comparably thick ice, while for the steep outlet glaciers (type B) thickness is much lower. These outlet glaciers are thus reacting very strongly and quickly to small changes in climate (Winkler, Elverhøy, and Nesje 2009). On the other hand, the much thicker ice found at the central and highest parts of an ice cap will result in a limited change of ice cap extent during meltdown over decades (e.g., Giesen and Oerlemans 2010). Accordingly, the major contribution of ice caps (and also ice fields) to global sea level rise will come later than that of large valley glaciers, where most of the volume is concentrated in the thick flat tongues at comparably low elevations (Linsbauer, Paul, and Haeberli 2012). However, in a case where the ice cap was resting in a closed overdeepening (i.e., limiting the possibility for
Figure 5  (a) A small ice cap (diameter up to 2.5 km) on Liverpool Land (East Greenland) with only a few remaining snow and firn patches (white), clearly revealing the radial flow pattern. The ice cap has several outlet glaciers differing in development. The image is a screenshot from Google Earth and was acquired by the Quickbird satellite on July 19, 2005. Reproduced from Google, Quickbird satellite © 2015 Google; (b) A larger ice cap with a rather complex shape (with several domes and outlet glaciers) on Baffin Island (Canada). The Landsat ETM+ satellite image (path-row 21-20) was acquired on July 30, 2002, and covers an area of 35 by 25 km. United States Geological Survey.
Figure 6  A schematic diagram illustrating the general ice thickness distribution for different glacier types. Most of the ice volume (grey) is found where the surface (black) is flat, that is, in the accumulation region of ice caps (i.e., they are “top heavy”) and in the ablation region of valley glaciers (“bottom heavy”). Mountain glaciers are comparably thin.

meltwater to leave), the ice would quickly melt away in a surrounding water body.

Paleoclimate

The radial flow of ice caps results in a point within the ice cap that should not flow, either at the surface, or at the ground. This is to a large extent also valid for ice masses of type D or those located at topographic saddle points (Haeberli et al. 2004). Such locations are ideal for retrieving ice cores (with annual layers undisturbed by ice flow) that reveal the climatic history at that location, given that the ice is cold (i.e., the snow metamorphosis is dry and not disturbed by liquid water). Nearly independent of the ice cap size, the missing flow at these locations results in potentially very old basal ice that can be found here (much older than in the largest glaciers). Accordingly, ice cores have been drilled into ice caps all around the world revealing the climatic history in the respective regions over the past millennia from the isotopic analysis of enclosed gases and particles. The variability derived from the respective components allows us to see currently observed changes (in climate and glaciers) from a much longer perspective, that is, to confirm evidence that current changes are extreme and exceptional over a timescale of 1000+ years (e.g., Fisher et al. 2012; Thompson et al. 2006).

Regarding the limited basal movement of the ice for the central part of an ice cap, it has also to be mentioned that human artifacts or even entire bodies can be preserved more or less unchanged for thousands of years at these locations, opening up a possibility for archaeological studies (e.g., Grosjean et al. 2007). The Ötztal Iceman (“Ötzi”) is likely the most prominent and well-known example, revealing that current ice mass loss is unprecedented for >5000 years in this region of the Alps (Baroni and Orombelli 1996).

Hazards

In several regions of the world, ice caps cover still active volcanoes. In the case of an eruption, parts of the ice cover can melt and cause enormous flood waves. For the ice-clad stratovolcanoes in North and South America, the mixture of melt water with the loose clastic material covering the surface of such volcanoes often results in huge and devastating debris flows called lahars (a Javanes term). They can reach velocities of more than 100 km$^{-1}$ (depending on the slope of the volcano) and travel over distances of >100 km. Their generally high sediment load results in characteristics similar to liquid concrete. While the well-known eruption of the Mount St Helens volcano in the Cascade mountain range caused a lahar that devastated large regions surrounding it, the 1985 eruption of the volcano Nevado del Ruiz in Colombia created a lahar that destroyed the town of Armero causing an estimated 25,000 fatalities. That lahar, known as the deadliest lahar in history, created an awareness that had long been missing of this widespread problem in the Andes. It has to be noted that lahars can
also be created on volcanoes without an ice cap and long after an eruption, for example due to high amounts of rain. However, an ice-capped volcano has an increased hazard potential.

For ice-capped shield volcanoes (or basaltic domes), such as the Grimsvötn volcano located under the Vatnajökull ice cap in Iceland, the situation is very different, as the ice covering the crater is much thicker and surrounding steep slopes of the bedrock are missing. This results in a possible accumulation of huge amounts of water underneath the ice that may after some days create sufficient pressure to lift the ice and break out catastrophically. The related event is called a jökulhlaup (an Icelandic term) and is also well known for its devastating nature affecting large areas. As these events are to some extent predictable, safety measures can be taken and fatalities are limited. This is not the case, however, for the also englacial but much smaller water pockets that can form inside any glacier and break out suddenly, thus creating catastrophic debris flows, sometimes with many fatalities (e.g., Vincent et al. 2010). Ice caps can also be related to a further destructive type of water release, so-called glacier lake outburst floods (GLOFs), which are related to a sudden drainage of a proglacial or supraglacial lake. The use of the terms jökulhlaup and GLOF in the literature is often interchangeable.

Inventory

In principle, a glacier inventory provides information on water resources for distinct hydrologic catchments. This often requires a separation of an ice cap into different parts according to the ice flow directions on the glacier surface (i.e., along drainage divides). Though this is in general possible using a digital elevation model and watershed calculations, it is unwanted for some glaciological calculations that require consideration of the respective ice cap in one part, for example for volume estimates (Grinsted 2013). The situation gets more difficult for ice caps with visible topographic undulations (Figures 1 and 2) or those with distinct outlet glaciers (Figure 3b). In these cases, there should be a possibility to analyse them separately (i.e., have drainage divides applied). When an ice cap has only one clear outlet glacier that can be separated by drainage divides, the other parts of the ice cap should also be separated. This is also the rule that has been applied for the inventory of the local GIC on Greenland (Rastner et al. 2012). The number of glaciers in an inventory and their size distribution varies depending on the decision whether or not to separate ice caps (Racoviteanu et al. 2009; Winsvold, Andreassen, and Kienholz 2014).

Conclusions

Despite the fact that ice caps are only a morphologically different type of glacier, they have some specific characteristics deserving of an in-depth description. The consequences of their generally flat topography with radial flow are their high climate sensitivity, the rather different ice thickness distribution (top heavy), the possibility to extract ice cores at their flow center and find undisturbed human or natural artifacts revealing paleoclimatic and archaeological information, and their association with natural hazards when they cover active volcanoes. They are also special in a more methodological sense as they can be difficult to distinguish from other glacier types (such as ice fields), and the separation along drainage divides that is required for a glacier inventory is limiting glaciological applications (e.g., volume calculation and change assessment). They are thus not (yet) properly indexed in glacier inventories and we do not exactly know how many ice caps we have on Earth or what area they cover.
Acknowledgments

The first author received funding from the ESA project Glaciers_cci (4000109873/14/I-NB). I would like to thank W. Haeberli and Y. Arnaud for helpful comments.

SEE ALSO: Climate change and land ice; Cryosphere: remote sensing; Glacier changes; Glacier lake outburst floods; Glaciers; Ice sheets; Landforms and physiography; Metadata; Natural hazards and disasters

References

ICE CAPS


Further reading


Ice sheets

Marie-Françoise André
Blaise Pascal University, France

Definition and genesis

Ice sheets are huge dome-shaped continental glaciers up to several kilometers thick and covering at least 50,000 km². They form in vast areas where winter snow accumulation exceeds summer snow melting: over thousands of years, the layers of snow pile up into thick masses of ice. The growth and decay cycle of ice sheets is forced by orbital variations of the Earth around the sun. Ice sheets grow as a response to climate cooling and the lowering of equilibrium line altitudes. When they form, ice sheets store massive quantities of water as solid inland ice, inducing sea level lowering and continental crust subsidence. Conversely, ice sheet decay is accompanied and followed by rapid sea level rise and more gradual continental uplift called postglacial isostatic rebound. Two main models of ice sheet formation have been proposed: (i) highland origin, in which ice sheets start to form as alpine glaciers in mountain areas and flow out to lower regions before thickening; and (ii) instant glacierization, in which extensive flat areas become snow accumulation zones as a result of cold conditions and continue to grow as a result of high albedo and positive feedback processes. These two models should be considered as more complementary than antagonistic because most ice sheets develop or have developed in vast and flat shield areas fringed with mountains.

Current and past ice sheets

The two current ice sheets, which contain more than 99% of the freshwater ice on Earth, are located in Greenland and Antarctica. The Greenland ice sheet, which formed 3 million years ago, is up to 3 km thick, covers 1.7 million km², and displays on its margins mountain peaks called nunataks and outlet glaciers called ice streams. The Antarctic ice sheet, which started to form 35 million years ago, is more than 4 km thick and covers 14 million km². It is fringed by floating ice shelves that break into giant flat icebergs (e.g., the 11,000 km² B-15 iceberg which broke away from the Ross Ice Shelf in 2000). The Antarctic ice sheet is divided by the Transantarctic Mountains into two unequal sections: the East Antarctic Ice Sheet (EAIS) rests on a major landmass, retains 85% of the ice volume, and reaches almost 5 km in thickness; the West Antarctic Ice Sheet (WAIS) retains 15% of the Antarctic ice volume, is thinner (1–2 km), and is partly marine-based because it rests on an archipelago. During the Last Glacial Maximum (20,000 years ago), two additional ice masses existed: the Laurentide Ice Sheet which covered most of Canada and the northern part of the United States, and the Fennoscandian Ice Sheet which covered much of Europe from Scandinavia to the British Islands and the Germano-Polish Plain. At that time, ice sheets contained more than twice as much ice as the present ice sheets and sea level had dropped by 120 m.
Ice sheet motion and thermal regime

Ice sheets flow downhill and outward from their dome-like centers and much of ice flow terminates at the ocean as ice tongues or ice shelves. Ice sheet flow is a function of surface slope and ice thickness. In the central part of the ice sheets, where most of the glacier bed is subhorizontal, flow speeds are a few centimeters to a few meters per year, whereas along outlet ice streams, ice speeds can reach hundreds of meters or even kilometers per year. Comparison of landform patterns beneath the former Laurentide and Fennoscandian ice sheets to the flow organization beneath parts of the contemporary Antarctic ice sheet by Kleman and Glasser (2007) indicates a complex thermal regime of ice sheets associating frozen-bed patches, ice streams, and lateral shear zones. In some areas (e.g., near the ice divide in continental areas), frozen-bed zones are extensive, whereas in other areas they constitute isolated “islands” in regions dominated by thawed-bed conditions. Sharp lateral boundaries mark the location of shear zones between thawed-bed ice streams and intervening frozen-bed areas.

Ice sheet landforms

The classical zonation of ice sheet landforms identifies three landscape associations (Bennett and Glasser 2009): (i) interior tabular shield areas glacially scoured in a subdued “knob and basin” topography characterized by polished and striated roches moutonnées (e.g., the Ungava lowlands east of Hudson Bay); (ii) mountainous margins displaying glacial troughs and fjords dominated by alpine peaks and cirques (e.g., the Narvik region in Norway); and (iii) marginal southern plains with terminal moraines damming lakes (e.g., Lake Michigan). Beyond the sharp contrast between the glacial imprints of ice sheets in their inner and outer parts, the variations in space and time of the thermal regime of ice sheets have induced a complex landscape mosaic including widespread remnants of preglacial topographies and weathering mantles preserved during the Quaternary glaciations below cold-based ice. Low sedimentation rates inferred from the off-shore record indicate an overall limited erosional impact of ice sheets except in their mountainous margins where fast-flowing thawed-bed ice streams operate.

Ice sheet climate record

Ice sheets contain a unique record of Earth’s climate history, which is provided by the air bubbles trapped in the ice layers accumulated over thousands of years. Pioneer ice-coring operations were carried out by the Americans in the 1960s at Camp Century and Byrd stations in Greenland and Antarctica, respectively, and this allowed them to reconstruct the climate history for the last 80,000 years. In the 1980s and 1990s, the French–Russian Vostok ice cores provided a 420,000-year record of past temperatures and atmospheric carbon dioxide (CO$_2$). To date, the longest climate record is held by the European Project for Ice Coring in Antarctica (EPICA) team, which in the 2000s extracted an ice core from the Dome C area in East Antarctica. This 3300-meter-long ice core provides information on the natural climate variations over the last 800,000 years. It also confirms that the current level of atmospheric CO$_2$ is unprecedented and should therefore be partly attributed to human emissions (EPICA Community Members 2004). In the Arctic, the international North Greenland Ice Core Project (NorthGRIP) operation provided, in the 2000s, a high-resolution ice core covering the last 123,000 years.
Ice sheet mass balance

As complete melting of the Greenland and Antarctic ice sheets has the potential to raise the sea level dramatically, scientists try and measure the ice sheet mass balance, that is, the difference between ice gains and losses (Shepherd et al. 2012). These estimates are based on two approaches: (i) the component approach, which compares mass input by snowfall and mass output by meltwater runoff and iceberg calving; and (ii) the integrated approach, which measures changes in the surface height (hence volume) or gravitational attraction (hence mass) of the ice sheets using instruments mounted in satellites, including radar and laser altimeters and high-precision gravity-measuring systems (Rémy and Parouty 2009). Mass balance estimates for Greenland show thickening at high elevations, but this mass gain by winter snow accumulation is far exceeded by losses due to increased melting and iceberg discharge from fast-moving outlet glaciers (up to 14 km per year). Dynamic thinning (ice loss resulting from accelerated glacier flow) now reaches all latitudes in Greenland, and the mass balance of the whole ice sheet is clearly negative, resulting in a current contribution to a sea level rise of about 0.20 mm per year. In Antarctica, dynamic thinning has intensified at key areas of West Antarctica such as the Amundsen Sea sector, with special reference to the Pine Island Glacier (Flament and Rémy 2012). Recent breakups of the Larsen, Wordie, and Wilkins ice shelves in the Antarctic peninsula have also induced important ice losses. But, as most of the EAIS is gaining mass slightly as a result of increased winter snowfall, the mass balance might be close to 0, inducing a negligible sea level rise.

Projections and uncertainties

According to pessimistic mass balance estimations and emission scenarios, some models suggest that accelerated ice loss from Greenland and West Antarctica will continue and cause sea level rises of 12 cm and 5 cm respectively by 2100. But predictions are uncertain because of high natural variability that occurs on a range of timescales. Separating long-term trends from the effects of this variability requires observation over long time periods. Another open question deals with the significance of the current accelerated ice flow of outlet glaciers on the ice sheet margins (Figure 1): are they indicators of destabilization or normal modes of the dynamic readjustment of ice sheets? In any case, the improved knowledge of the subglacial thermal organization of ice sheets indicates that widespread frozen-bed

Figure 1  Map of ice velocities (m per year) derived from satellite radar altimetry data showing the numerous ice streams draining the Antarctic ice sheet. © Frédérique Rémy (Rémy and Parouty 2009, 1226, fig. 8).
patches contribute to the stability of ice sheets by laterally constraining and isolating peripheral drainage basins and their ice streams (Kleman and Glasser 2007). In Antarctica, if the future of the WAIS is uncertain, the long-term view indicates an overall stability of the EAIS, which is the main ice mass on Earth.

**Controversy over the long-term stability of the East Antarctic Ice Sheet**

The interpretation of the sediment marine record had long promoted the view of a stable Antarctic climate and ice sheet over the last 14 million years, but the discovery of puzzling deposits in the Transantarctic Mountains challenged this view and triggered a 10-year-long controversy (Sugden 1996). The Sirius till deposits contained fossil wood, indicating a glacial advance into a temperate forest, and marine diatoms entrained by glaciers from seabed dating back to the Pliocene warm period. It was inferred from this finding that the EAIS had experienced massive deglaciation and caused dramatic sea level rise 3 million years ago. This deglaciation hypothesis was contradicted by the preservation of fragile Miocene deposits on 30° slopes in the Dry Valleys area, which indicate that the landscape had been unmodified for the last 14 million years. But geomorphological evidence was not considered at that time and, as the instability theory had major implications for the potential risk of future ice sheet collapse in the context of global warming, it achieved some success until diatoms were found in cracks and surface layers at various places in Antarctica including the South Pole. It meant that the diatoms had been airborne instead of glacially entrained from the seabed and that the Pliocene dating of the Sirius deposits was wrong, which ruled out the hypothesis of a Pliocene collapse of the Antarctic ice sheet. The addition of geomorphological evidence, the marine sediment record, and glaciological modeling strongly supports the hypothesis of the 14-million-year stability of the EAIS, the main ice mass on Earth.

**The marine ice sheet instability hypothesis and the threat of collapse of the West Antarctic Ice Sheet**

The WAIS is the world’s most vulnerable ice sheet. This is because it is marine-based, and therefore prone to rapid melting due to the upwelling of warm, deep ocean water along the coast of Antarctica. This warm water has increased the melting of floating ice shelves at the edge of the WAIS. Ice streams draining the WAIS into the Amundsen Sea, which have a grounding line on a reverse bed slope (becoming deeper inland), have started to thin and accelerate. The marine ice sheet instability hypothesis is that atmospheric and oceanic warming could result in increased melting and recession at the grounding line on such a reverse slope gradient. This would activate flotation, basal melting, increased iceberg production, and further retreat within a positive feedback loop. Ultimately, it could result in the massive collapse of the WAIS, triggering rapid sea level rise. If the ice shelves that fringe the WAIS are known to have a stabilizing or buttressing effect on the ice sheet, the recent speedup of their outlets is spreading far inland, which is considered by some scientists as the start of marine ice sheet instability. Some numerical models suggest that a full-scale collapse of this sector of the WAIS may be inevitable, but leave a large uncertainty in the timing.

**SEE ALSO:** Antarctica; Climate change and land ice; Glacier mass balance; Ice shelf buttressing; Quaternary glaciations
References


Ice shelf buttressing

Daniel N. Goldberg
University of Edinburgh, UK

The Antarctic and Greenland ice sheets are the most massive bodies of ice in the world, containing about 30 million and 3 million km$^3$ of ice, respectively. Input of mass to the ice sheets is exclusively through snowfall on the surface densifying into glacial ice. Mass is lost through melting, either at the surface or at the base of the sheets, and through icebergs calving off the margins of the ice sheets into the ocean. The two ice sheets differ greatly in the importance of these outputs, however: in Greenland, surface melting is quite important and a large part of the ice sheet margin is land-terminating. In Antarctica, surface melting is negligible and ice flows into the ocean around its entire perimeter. These inputs and outputs control the volume of ice in the sheets, a quantity of great societal importance because it affects global sea levels.

In Antarctica, snow accumulation has increased in recent decades, possibly due to warming air masses. Loss of ice by flow across the ice sheet margin, however, is poorly characterized and its future behavior is almost completely unknown. Around most of Antarctica, the ice sheet goes afloat in the form of large ice shelves, which vary in area from the size of Greater London to roughly that of France. Since the ice shelves are in hydrostatic balance with the ocean, their direct sea level contribution would be negligible were they to melt entirely. On the other hand, the indirect effect of such an occurrence would have far-reaching effects, and in some cases could threaten the stability of large parts of the Antarctic ice sheet. This is due to “ice shelf buttressing,” the term used to describe the resistive forces imparted to the ice sheet by its ice shelves.

Physical setting and theory

The transport of ice toward the ocean is not uniform around the continental margin; rather, it is concentrated in relatively narrow, fast-flowing ice streams. These ice streams empty into ice shelves, and the point at which the ice goes afloat is called the grounding line. The position of the grounding line is determined by a floatation condition: the weight of the ice column must equal the weight of the water column it displaces. In other words,

$$\rho g H = \rho_w g D$$

where $\rho$ and $\rho_w$ are average ice and ocean densities, respectively, $H$ is the ice thickness (distance from top to bottom), and $D$ is the bedrock depth. Thus the grounding line is dynamic; if ice thickness changes, the grounding line will adjust to a position where equation 1 is satisfied. As ice shelves do not contribute to sea level, any changes in grounding line position, or in the flux of ice across the grounding line, translate to sea level change.

To appreciate the role that ice shelves play in grounding line position and ice flux, one should first examine the dynamics of a grounding line with no ice shelf attached, that is, where there is simply an ice cliff at the grounding line, and all ice that crosses it calves off instantly as icebergs (Figure 1). In this case there is an imbalance between the pressure within the ice and the
ocean pressure along the ice face, and this results in a tensile stress acting at the grounding line in the seaward direction, pulling the ice forward. Averaged over the depth of the ice sheet this equals

\[ \sigma = \frac{1}{4} \Delta \rho g D \]  

(2)

where \( \Delta \rho = (\rho - \rho_w) \). (Note: some sources quote a \( \frac{1}{2} \) factor on the right-hand side; this is because they are discussing membrane stress, which in this context is twice the tensile stress.) This tensile stress induces a flow of ice across the grounding line; however, the relationship between equation 2 and ice velocity depends on a number of factors. Weertman (1974) derived an approximate relationship between the two, and deduced the result that ice flux across the grounding line increases quite strongly with depth. More recently, Schoof (2007) derived the analytical result that ice volumetric flux per unit width is proportional to \( \sigma^\alpha D^\beta \), where \( \alpha \) varies from approximately 1.5 to 2.25 and \( \beta \) from approximately 2 to 2.5, depending on the nature of the ice-bed interface.

That ice flux increases with grounding line depth has far-reaching implications for ice sheet behavior and stability. A large sector of Antarctica, the West Antarctic Ice Sheet, is a marine ice sheet: it rests on a bed that is far below sea level, in part due to the weight of the ice and in part due to millennia of erosion. Moreover, it is deeper, on average, toward the center than at the margins. For such an ice sheet to be in steady state, input from snow accumulation, which is roughly

---

**Figure 1** A visualization of many of the features and concepts discussed in this article. Hannes Grobe, Alfred Wegener Institute for Polar and Marine Research, Germany. Reproduced from Wikipedia (http://en.wikipedia.org/wiki/File:Antarctic_shelf_ice_hg.png).
ICE SHELF BUTTRESSING

proportional to its area, needs to be balanced by the flux of its ice streams across its grounding line. If stress and flux are related as described above, such an ice sheet cannot be stable.

To see this, consider a small inland retreat of part of its grounding line: this decreases the input to the sheet by decreasing its surface area. At the same time, the grounding line has retreated onto a deeper bed, increasing the flux. Thus, in this new configuration, less ice is delivered to the grounding line than moves across it, and ice thins at, and upstream of, the grounding line. Since total accumulation input has decreased, this imbalance will not correct itself; rather, ice at the grounding line will go afloat and subsequently calve into icebergs or melt. Similarly, a small advance of the grounding line would result in unchecked advance to the edge of the continental shelf. This instability is often referred to as marine ice sheet instability (MISI) (Mercer 1978). The instability has significant implications: if the entire West Antarctic Ice Sheet were to collapse, sea level would rise by 3–5 m globally.

An important point to make is that the above arguments hold not only for a grounding line with no ice shelf, but also for a grounding line with an ice shelf that is completely unconfined, that is, it does not exist in an embayment and there are no points of contact between the ice and the ocean bottom. However, such is rarely the case for Antarctic ice shelves. Most ice shelves are embayed, flanked by slow-moving grounded ice that limits the oceanward flow of the ice shelves; and the larger ice shelves – the Ross Ice Shelf and the Filchner-Ronne Ice Shelf – are pinned at large ice rises in their interior. These embayment walls and ice rises exert a resistive force on the ice shelf, which is carried through the ice shelf and exerted on the grounding line, lowering the tensile stress relative to equation 1, and in turn decreasing ice velocity relative to the unconfined case.

This ability of ice shelves to exert resistive force and decrease velocity of outlet streams is commonly known as buttressing: the ice shelf buttresses against the tendency of ice to flow under its own weight. Strictly speaking, the force is not originating from the ice shelf itself; the ice shelf is simply allowing forces along different parts of the grounding line to be felt nonlocally. If a part of the grounding line, for example a ridge along the side of the shelf, experiences strong friction at its base, the ice shelf allows this frictional force to slow other parts of the grounding line, such as the outlet of a fast-moving ice stream. In the following, different types of buttressing will be examined.

Buttressing by rigid sidewalls

Consider an ice shelf in a rectangular channel of width $W$ and length $L$; at one end ($x = 0$), an ice stream flows into the shelf; and the other end ($x = L$) is open to the ocean. On either side, the ice is slow-moving, and at every point along these rigid sidewalls a shear stress $\tau$ resists flow in the ice shelf. It can be shown (Thomas 1973) that the tensile stress felt at the grounding line at $x = 0$ is lessened by $F$, where $F$ is the total amount of force arising from this stress. If $\tau$ is assumed uniform, then tensile stress at $x = 0$ is given by

$$\sigma = \frac{1}{4} \Delta \rho g D - \tau \left( \frac{H_r}{H} \right) \left( \frac{L}{W} \right)$$  \hspace{1cm} (3)$$

where $H_r$ is a representative ice shelf thickness (in general smaller than $H$; see Figure 2). Let us consider how this might change grounding line velocity. If we consider a value of 750 m for $D$, and $10^5$ Pa for $\tau$, then $\sigma$ will decrease by a factor of approximately
ICE SHELF BUTTRESSING

![Diagram of ice shelf buttressing](image)

Figure 2 Two simplified cases of an unconfined and buttressed grounding line: (top) an unconfined shelf with depth \( D \) at the grounding line; (bottom) a confined ice shelf. There is stress \( \tau \) along the sides of the ice shelf which effectively slows ice velocity.

\[
0.5 \times \left( \frac{L}{W} \right) \times \left( \frac{H_s}{H} \right) \tag{4}
\]

relative to its unconfined state. From the arguments of Schoof (2007), this can be translated to velocity change. For instance, if the unconfined grounding line velocity (the velocity if \( \tau \) is zero) is 1000 m per year, an ice shelf of aspect 1 \( (L/W \approx 1) \) and thickness \( H_s/H \approx 0.5 \) will reduce this speed to approximately 600–650 meters per year. A narrower, longer, or thicker ice shelf will reduce this speed even more.

It should be pointed out, though, that the above analysis is oversimplified: equilibrium ice shelf thickness would change with width and length. Additionally, \( \tau \) would not be constant, but would decrease as the shelf widens, and \( \sigma \) is unlikely to drop to zero. To truly determine the effect of buttressing, the differential equations governing stresses within the ice must be solved.

In general, though, the narrower, longer, and thicker an ice shelf, the more effective it will be at buttressing the flow of the ice stream that feeds it.

This has implications concerning MISO: if a grounding line retreats, the length of its attached ice shelf will increase, unless its calving front retreats at the same rate (which is unlikely because of the different physics determining the positions of the two). Thus, grounding line tensile stress could actually decrease as grounding line moves inland, even though the bed deepens, and the grounding line may stabilize. It should be pointed out that whether or not grounding line retreat is unstable depends on a number of factors, including the specific geometry and mechanical properties of the bed, and any factors influencing the ice shelf, such as crevassing and melting by ocean currents.

**Buttressing by ice rises and ice rumbles**

Theoretical studies suggest that a relatively narrow ice shelf is capable of reversing MISO, but a wide ice shelf is not. The Ross and Filchner-Ronne ice shelves are 700–800 km wide at their calving fronts, and it is unlikely that these ice shelves could effectively restrain ice stream flow through sidewall buttressing alone. However, these ice shelves come in contact with the ocean bed in several places in the form of massive ice rises – regions of grounded ice that are not connected to the main ice sheet. Ice flows away from the rise in all directions, as if it were an isolated ice cap or ice sheet, implying a strong (likely frozen) bed. Ice rises influence ice shelves in much the same way as slow-moving ice at ice shelf boundaries, and in
ICE SHELF BUTTRESSING

doing so effectively narrow a wide ice shelf. In Ross Ice Shelf, Roosevelt Island and Steershead and Crary ice rises are spaced more closely than the width of the overall ice shelf, and so the sides of the rises exert a stronger buttressing force on the Siple Coast ice streams than the side margins of the ice shelf could. Roosevelt Island may have had a strong influence on the location of the grounding line of MacAyeal Ice Stream since the Last Glacial Maximum. Similarly, it has been suggested that ungrounding of these ice rises may have had a dramatic impact on the flow of Whillans and Kamb ice streams; See Figure 3.

Ice rumples are isolated regions of grounding as well, but ice slides over the contact; the ice shelf flow slows but does not stagnate. In general, the surface elevation contrast between rumples and the surrounding shelf is less than that of ice rises. Ice rumples provide buttressing as well, by transmitting frictional force from the bed back to the grounding line. While the bed is weaker than under ice rises, ice rumples can be as important as ice rises in providing buttressing. There are several locations in which ice rumples are thought to provide ice stream stability, or have in the recent past. As recently as a few decades ago, Pine Island Ice Shelf, West Antarctica, was in contact with an undersea ridge; ungrounding of the shelf from this ridge is thought to have led to acceleration of Pine Island Glacier (Jenkins et al. 2010). On the other hand, the adjacent Thwaites Ice Shelf is in contact with a seamount at its base, and while this local grounding has a noticeable effect on the flow of the ice shelf, recent studies indicate it does not play a strong role in the stability of Thwaites Glacier.

Buttressing by unconfined ice shelves

A frequently made statement is that an unconfined ice shelf does not exert any buttressing force on ice flow at the grounding line. This is not completely true, however. Ice shelf buttressing refers to any transfer of force across distances due to the presence of an ice shelf. If a section of the grounding line overlies a stronger bed than the rest, this stronger bed effectively provides a restraining force to weaker-bedded parts of the grounding line; the strong-bedded section is essentially pulled forward by the weaker-bedded sections. If the ice shelf were to be removed, the latter would accelerate (and the former would decelerate somewhat). In general, buttressing by unconfined ice shelves is likely less important than that of sidewalls and ice rises, but it may be of importance to short, narrow ice tongues, such as those in Greenland fjords whose margins are too crevassed and broken up to provide any support.

Climate impacts on ice shelf buttressing

As in the simple example of sidewall buttressing, thinning of an ice shelf will limit its buttressing ability. However, the cause of thinning is important. If the shelf thins because it is “starved” by the ice sheet, it may be due to changes in the ice sheet interior, such as long-term changes in precipitation. But if the thinning is due to melting at its lower surface by heat from the ocean, the impacts on the ice sheet can be very important. As the ice shelf thins due to melting, buttressing is lost and velocities at the grounding line increase. The thinning of grounded ice propagates inland because the lowering of the ice surface increases the basal slope, which increases velocities. If the ice sheet bed deepens inland, extensive grounding line retreat occurs as well. The velocity increase results in more mass being fed to the shelf; but if the ocean contains sufficient heat to remove this ice, the ice shelf does not thicken and the retreat does not reverse itself.
Figure 3  Plan view (top) and section (bottom) of an ice shelf buttressed by rigid sidewalls and ice rises. Arrows represent ice velocity magnitude and speckling density represents the level of resistive force carried by the ice shelf. Thomas 1979. Reproduced by permission of International Glaciological Society (IGS).
This process is currently occurring on several ice shelves in Antarctica, most notably in those of the Amundsen Sea Embayment in West Antarctica. Almost all of the ice shelves around Antarctica are exposed to some degree of melting in their deepest parts, since ocean waters can be no colder than the surface melting point (about $-2^\circ$C) due to sea ice formation, and the melting point decreases with depth. However, due to deep underwater troughs in the Amundsen Sea, along with the fact that waters tend to be warmer at depth at high latitudes, these ice shelves are exposed to waters of 1.2°C or greater (Jenkins et al. 2010). Exposure to these warm waters results in large thinning rates; ice shelf thinning of 6–8 m per year has been inferred, which has caused extensive speedup and thinning of the Pine Island, Thwaites, and Smith glaciers (Shepherd, Wingham, and Rignot 2004). On Pine Island, thinning and speedup are seen hundreds of kilometers upstream from the ice shelf (Wingham, Wallis, and Shepherd 2009; See Figure 4). This thinning could have far-reaching implications: these ice streams (Pine Island, Thwaites, and Smith) drain a portion of the ice sheet that could raise global sea level by 1.3 m, and they rest on inland-deepening beds, making them unstable.

Ice loss due to melting ice shelves is not limited to the Southern Hemisphere. The floating tongue of Jakobshavn Isbrae, a fast-flowing outlet glacier in southwest Greenland, rapidly thinned and broke up when a shift in large-scale ocean circulation brought relatively warm waters from the Atlantic into its fjord. The loss of this small ice shelf led to a doubling of the glacier speed (Holland et al. 2008). Elsewhere in Greenland, marine-terminating glaciers without floating extensions can be exposed to high melt rates along their calving cliffs. It is important to realize, however, that this is different from ice shelf melting: in the latter case, grounded ice is

![Figure 4](image-url)
affected through the loss of buttressing; in the former, melting affects grounded ice directly.

Loss of buttressing due to melting has been observed to be a slow process, operating over years to decades. However, far more rapid collapse of ice shelves has been observed. In the past two decades, several ice shelves along the Antarctic Peninsula have “disintegrated”: seemingly intact, they are transformed into a mass of small icebergs in a matter of days and then drift away. The process is thought to be due to melting at the upper surface by solar radiation and heat from the atmosphere, as the Peninsula lies farther north than the rest of Antarctica. Extensive melt ponds form, seeping into surface crevasses; the crevasses then quickly deepen, wedged open by the weight of the water. The breakups may have been preconditioned through the removal (by iceberg calving) of parts of the ice shelves vital to their structural stability. At any rate, after the disintegration of one of the largest of these ice shelves, Larsen B, large speedups were observed on the ice streams that fed it, indicating that the Larsen B provided buttressing for these streams (Scambos et al. 2004; see Figure 5).

Finally, it should be stated that the mechanisms through which the ocean and atmosphere affect ice shelves are still poorly characterized. An improved understanding of the processes involved is necessary before the effects of climate change on the Antarctic and Greenland ice sheets can be assessed.

SEE ALSO: Antarctica; Climate change and land ice; Glacial erosional processes and
landforms; Ice caps; Ice sheets; Ice shelves; Oceans and climate; Sea level rise

References

Scambos, T.A., J.A. Bohlander, C.A. Shuman, and P. Skvarca. 2004. “Glacier Acceleration and Thinning after Ice Shelf Collapse in the Larsen B Embay-

Ice shelves

Adrian Jenkins

British Antarctic Survey, Natural Environment Research Council, UK

Ice shelves are the parts of an ice sheet that are afloat in the ocean, forming at the marine margins where the ice is no longer thick enough to maintain contact with a bed that lies below sea level. Although floating, they remain an integral part of the ice sheet, providing a critical control on the outflow of ice from inland and hence on the size of the grounded ice sheet. Understanding the behavior of ice shelves, and in particular how they interact with the ocean waters that circulate beneath them, is a prerequisite for understanding how the transport of ocean heat toward the Earth’s great ice sheets controls their size.

Ice shelves range in scale from the simple extension of a single tidewater glacier, rarely more than approximately 10–100 km long and often not accorded a geographical name distinct from that of the parent glacier, to vast amalgamations of many tributary glaciers that can extend approximately 1000 km from the grounding line, where the ice goes afloat, to the ice front, and individually cover areas of up to $5 \times 10^5$ km$^2$. They are nourished by the flow of ice across the grounding line and surface and, less commonly, basal accumulation, while they lose mass by calving of icebergs from the seaward edge and basal and, less commonly, surface ablation. Their state of free flotation on a frictionless substrate lends them a distinctive stress regime, in which vertical shear stresses are almost entirely absent. The dominant longitudinal stresses drive horizontal spreading and vertical thinning of the ice. Since the rate of thinning is strongly dependent on the ice thickness itself, the process acts to smooth variations in ice thickness. The resultant near absence of surface topography and crevassing made the larger Antarctic ice shelves important highways to the interior of the ice sheet for the early explorers of the continent.

The vast majority of currently extant ice shelves are to be found in Antarctica, especially west Antarctica, where they almost completely surround the marine-based West Antarctic ice sheet. Antarctic ice shelves range in thickness from several hundred meters at the calving face to one to two kilometers, or even more in a few places, at the grounding line. Although they represent only a small percentage of the Antarctic Ice Sheet by area ($\approx$11%) or volume ($\approx$2.5%), their coastal location means that they receive around 20% of the overall accumulation directly and around 80% of the remainder indirectly as drainage from the grounded ice sheet. Thus the majority of mass input to Antarctica is returned to the oceans via the ice shelves, from where about 50% is lost by iceberg calving and about 50% by basal melting (Rignot et al. 2013). In contrast, only a few Greenland outlet glaciers retain a floating tongue, while the few small ice shelves that historically fringed some of the smaller ice caps of the Arctic have all disintegrated.

Systematic scientific studies of the Antarctic ice shelves began in the second half of the twentieth century. Initial efforts involved mapping of ice thickness and underlying seabed topography, with measurements of the former being revolutionized in the 1960s with the development of ice-penetrating radar techniques. In the 1970s the simple stress regime
of ice shelves was exploited as they became large-scale samples on which to test flow laws for ice in a natural setting, while in the late 1970s and early 1980s major international campaigns were initiated to study the vast Ross and Filchner-Ronne ice shelves. In the latter half of the twentieth century, interest in ice shelves waned somewhat, but was reinvigorated as developments in satellite remote sensing allowed continent-wide monitoring of ice extent, surface elevation, and ice flow. It soon became apparent that dynamical changes were taking place in both the Greenland and Antarctic ice sheets on timescales that could be observed. The most rapid changes were taking place on the ice shelves and floating glacier tongues, from where the signals extended inland to the outlet glaciers that were feeding them. This was a major stimulus for recent research work focusing on the ice shelves and how they interact with the atmosphere and especially the ocean circulating beneath them.

The loss of ice shelf cover along the Antarctic Peninsula has been one of the more dramatic indicators of recent climate change in the polar regions. Ice shelves appear to be subject to a strict climatic threshold, in that once surface temperatures are high enough for significant melt to occur in the summer, they are no longer viable. The lack of significant surface topography means that surface melt does not run off, but accumulates in minor depressions and surface fractures. When meltwater fills a crevasse, the pressure within the crevasse exceeds that in the surrounding ice and the fracture can penetrate the full depth of the ice shelf. Opening an access route to the ocean beneath allows the meltwater to drain, but full-depth fracturing of the ice shelf eventually leads to its disintegration. Warming of the Antarctic Peninsula has driven the climatic threshold southward, leading to progressive loss of ice shelves along both coasts.

Below the climatic threshold, ice shelves show little sensitivity to surface temperature. Although the viscosity of ice, and hence its flow rate in response to a given stress, is a function of temperature, snow and ice are poor conductors of heat, so that the core of an ice shelf remains cold even as the surface warms. Thus for most ice shelves, the primary control on the mass balance is set by the ocean. Although the polar oceans freeze at the surface during winter, at depth they remain above the freezing point year round, and the enormous specific heat capacity of water means that even a small temperature increment above freezing represents a substantial reservoir of thermal energy. Furthermore, the lowering of the freezing point with depth means that even water at the surface freezing point has the potential to melt ice at depth. Technically the process is one of dissolution, rather than melting, since, because of the presence of salt in the water, the phase change occurs at temperatures below the melting point of pure ice. Thermodynamically the process is equivalent to that by which ice can be cleared from roadways by spreading salt over the surface. The resultant heat loss from water that is already at the surface freezing point results in a distinctive water mass that would be supercooled at atmospheric pressure and remains liquid only because of the elevated pressure. When such a water mass rises back toward the sea surface, the reduction in pressure results in supercooling and freezing. Thus, melting of thick ice near the ice shelf grounding line tends to be partially offset by freezing beneath thinner ice nearer the ice front, and the net melt rates of the ice shelves floating in such cold waters are relatively low.

Although the surface waters are cooled to freezing by heat loss to the atmosphere, warmer waters are found at depth within the polar oceans, and where these can reach the ice shelves
the basal melting rates can be much higher. The warm waters have their origins in the subtropics, from where they take a relatively direct route to high northern latitudes. There the surface layer is cooled by the atmosphere, a process that makes the water more dense, but surface freshening, particularly at the coasts where runoff supplements precipitation, lowers the density and keeps the cold layer at the surface. Thus the floating ice tongues that extend into the Greenland fjords experience a two-layer water column, with a near-freezing layer typically around 200 m thick overlying a warmer, saltier layer with temperatures up to four or five degrees above freezing. Nowhere does it appear that the surface layer is thick enough to block the access of the warm water to the ice, even where the inflow of the deeper layer is partially impeded by seafloor sills. Many of the recent changes in Greenland’s floating ice tongues appear to be linked with changes in the quantity and temperature of this warm water layer in the fjords.

In the centers of the subpolar seas to the east and west of Greenland, freshening of the subtropical waters is insufficient to offset the impact of cooling, and denser water formed at the surface sinks to great depths to form new deep water, which is subsequently carried at depth to southern high latitudes where it feeds into the Antarctic Circumpolar Current. Core temperatures are typically around 1–2 °C where the deep waters approach the continent. As in the north, cooling and freshening at the surface produces an overlying layer of near-freezing water that is typically a few hundred meters or more thick near the coast. Coastal runoff is practically nonexistent around Antarctica, so the density difference between the cold surface layer and the underlying warm water is not large. Consequently, deepening of the surface layer isolates most of the ice shelves from the warm waters, reducing the basal melt rates. However, along the Pacific coast of west Antarctica the warm layer of deep water reaches to the coast and basal melting of the ice shelves is one to two orders of magnitude greater than elsewhere around the continent. The temperature of both the surface layer and the deep water is relatively constant and largely unaffected by contemporary atmospheric temperatures. However, winds affect the ocean circulation and control the relative thickness of the two layers. Thus, changes in wind forcing can change the amount of warm water beneath the ice shelves and have a profound impact on the basal mass balance. Thinning of the ice shelves in west Antarctica that has led to acceleration of outflow glaciers and thinning in the interior of the ice sheet appears to be linked to wind-forced changes in the delivery of warm water to the ice shelves.

When ice shelves change in thickness or extent, they affect sea level in two ways. The direct impact is commonly, although mistakenly, considered to be zero. Since ice shelves are freely floating, they do displace their own weight in seawater. However, once they have melted the equivalent volume of freshwater is greater than that of the displaced seawater, because it is lower in density. As the freshwater mixes with seawater the total volume is conserved, because the equation of state for seawater is approximately linear in salinity. There is thus a small steric rise in sea level, equal to about 2.7% of the ice volume, resulting from the net dilution of the ocean, that occurs when floating ice shelves melt. The indirect effect, whereby changes in ice shelf geometry influence the flow of inland ice into the ocean, is more significant because changes initiated by ice shelf thinning and retreat can propagate far into the interior of an ice sheet. Current thinning of some of Antarctica’s major drainage basins has been initiated by either
ICE SHELVES

ocean-driven thinning or atmospheric-driven breakup of the ice shelves. Similarly, some of the most rapid thinning rates seen in Greenland have been linked to the thinning and breakup of floating ice tongues.

SEE ALSO: Ice sheets; Ice shelf buttressing; Oceans and climate; Polar climates

Reference


Further reading


Idealism

Alex Jeffrey
University of Cambridge, UK

The concept of idealism is open to multiple interpretations. In general terms, idealism refers to a broad range of philosophical positions within the social sciences that have sought to understand how human interpretation shapes understandings of reality. These approaches can be roughly organized into two philosophical stances: a form of metaphysical idealism, where reality is understood to be primarily in the mind, to forms of epistemological idealism, where understanding is limited to perception of external objects. In this framework, idealism stands as an antonym for either materialism (where matter is privileged as the primary concern) or realism (which assumes an external and independent world operating beyond human beliefs or ideals).

In its philosophical traditions, idealism orientates attention to the systems of interpretation and categorization through which an external world is apprehended and understood.

Such general characteristics of idealism are both reflected, but also somewhat reworked, in the adoption of the term within geographical scholarship. While geographers had focused on issues of perception for much of the twentieth century, most significantly through the work of Richard Hartshorne, the challenges to logical positivism in the 1970s brought refreshed interest in the philosophical and methodological possibilities of idealism. Theoretical perspectives such as humanism, phenomenology, existentialism, and feminism each sought in turn to better incorporate the experiential elements of being human into geographical scholarship. Where logical positivism had sought to accentuate the objectivity and distance of geographical scientists, these diverse perspectives emphasized embodied practices of experience, thought, and practice.

At first glance idealism fits within this postpositivist mindset, and certainly its early adopters in the 1970s saw idealism as a means of challenging the supposed objectivity of logical positivism. This desire is apparent in the work of Leonard Guelke, who was one of the first geographers to adopt a language of idealism, if not its full philosophical implications. For Guelke (1974, 193), idealism offered a means of illuminating how human geography should be “largely concerned with rational actions and products of human minds.” In his wide-ranging explanation (from witchcraft through to the voyage of Columbus), Guelke views idealism as a means of explaining the wide range of human actions in different geographical environments. For Guelke, actions cannot be unified within a single theory, but rather the rational decisions of individuals can be understood within the frameworks of interpretation and meaning in a given context.

Guelke’s account was also a critique of emerging humanist and Marxist theoretical positions within geography. Where such accounts had sought to construct wide-ranging theoretical explanations for geographical phenomena, advocates of idealism were more circumspect on the possibility of grand theoretical propositions. By foregrounding the decisions and reflections of the individual, Guelke sought to posit an “atheoretical” account, since the character of human thought “precludes the possibility of
human geographers developing general theory capable of meeting minimum criteria of scientific criteria” (Guelke and Chappell 1976, 168). Thus for a geographical idealist, and in contrast to more structured accounts of human practice, an action was explained when the thought behind it had been understood (Guelke 1974).

Taken together, these attributes suggest a more ambiguous position toward logical positivism, Marxist geography, or humanist forms of theorization. In the intervening years these ambiguities have prompted scholarly criticisms, focusing in four main areas. First, scholars have questioned the extent to which Guelke grounded his conception of idealism in the appropriate philosophical canon. Reflecting on the relationship between cognition and the material world has its roots as far back as the seventeenth-century meditations of René Descartes and has been a constant feature of political and philosophical writings since. Immanuel Kant, for example, espoused a form of transcendental idealism which posited the unknowability of the “thing in itself,” pointing to the epistemological limitation that an individual cannot step outside his or her own mind (Guelke and Chappell 1976). Rather than contextualizing the term in this long intellectual tradition, scholars criticized geographical idealism in the 1970s for failing to employ standard conceptions of the term and therefore limiting the possibilities of interdisciplinary debate regarding idealism.

Second, commentators have questioned the extent to which the geographical interpretation of idealism represented a genuine break from the logical positivism which it sought to supersede. The continued attachment to the rationality of the individual suggested an enduring sense in which certain frameworks structure human behavior, even as individuals conceive of the world in different ways. In tandem, Guelke’s work does not refute the verification principle, since his accounts continue to espouse the necessity of grounding knowledge in experience, but through the reconstruction of the thoughts behind actions. Supporting this positivist attachment, Guelke was eager to distance his concept of idealism from more phenomenological and existential accounts, keen as he was to stress the focus on human rationality and lack of interest in psychological and emotional concerns. There is a sense, then, that idealist accounts wanted to break from positivism while retaining key elements that provided the possibility of verification and corroborations.

As a third criticism, the methodological possibilities of idealism were considered to be rather ambiguous. While Guelke sought to enter the mindset of the subjects of his research, he simultaneously advocated a form of theoretical detachment. Guelke’s assumption that one’s own theories could be “put to the back of one’s mind” within an atheoretical idealism seemed to reproduce the imagined detachment and neutrality of the positivist researcher. As Harrison and Livingston (1979, 78) suggest, this position imagines the researcher could simply “shake off the shackles of their own presuppositions in complete detachment from the matter under study.” There is a confusing construct here of an idealist scholar who understands the significance of human cognition, but only in those under investigation and not in themselves. Feminist work on positionality, psychoanalytic work on the subconscious, and postcolonial scholarship exploring the production of academic knowledge have subsequently challenged this enduring imagination of the privileged and neutral viewpoint of the social scientist.

Finally, idealism has been criticized for its avowedly humanist focus. Recent posthuman geography has sought to incorporate the agency of nonhuman actors and objects within
geographical theorizing. In doing so, this work orientates attention on the limits of an epistemology that focuses solely on human rationality, since it brackets the wide range of nonhuman forces and intentions that shape a given set of circumstances. This does not discount the significance of human cognition, but it does direct scholarly inquiries to the ways in which the material and discursive boundary between the human and the nonhuman is drawn, and with what political and intellectual consequences.

Taken together, these criticisms have limited the theoretical and methodological application of idealism within geography. But we must be clear that the fundamental claims of idealist philosophy – that human interpretations shape understandings of reality – are crucial to many strands of contemporary human geography. For example, the increased use of ethnographic methods and concerns of positionality have brought analysis of interpretation to the forefront of studying understandings of space and place. In these terms, we can perhaps say that idealist concerns have become a starting point for much scholarship within human geography, rather than a body of thought that marks a distinctive philosophical position.

SEE ALSO: Humanistic geography; Marxist geography; Ontology: theoretical perspectives; Phenomenology

References


Further reading


Identity

Elaine Lynn-Ee Ho
National University of Singapore

Identity belongs to a realm of experience that is intimately felt at an individual and community level, yet it can be slippery to pin down conceptually. Identity refers to a person’s sense of self and how it shapes an individual’s interaction with others in the social world. Identity claims can be invoked to reinforce or challenge systems of domination; hence it is a powerful tool in the hands of social actors and the social structures they represent. Social constructivists, influenced by the poststructural and postmodern turn in the social sciences, do not view identity as a biologically determined condition or fixed across time. Rather identity is amenable to change in response to social processes, relationships, and culture. Geographers further argue that identity is spatially situated and in turn shapes space. Analyses of identity should also be situated in wider structural systems associated with globalization and transnational migration as well as macro-ideologies such as postcolonialism and cosmopolitanism.

Identity is usually categorized according to a person’s social and cultural affiliations, such as socioeconomic status (or class), “race” or ethnicity, religion, gender, sexuality, or nationality. On the one hand, such affiliations serve as a bonding mechanism that helps define social membership. On the other hand, they also serve as markers of difference that delineate social groups from one another. A person is likely to identify with more than a single social categorization at any one time, experiencing not only multiple identities but also intersectional identities. Intersectionality refers to the interrelationship of different axes of identity, such as gender with ethnicity, making the experiences of certain social groups distinct from others. For example, critical race feminists (e.g., Hooks 1983) argue that women of ethnic minority status experience an additional layer of oppression compared to white women, which in turn affects their social positioning, access to resources, and life opportunities.

The process of identity-making is known as subjectification, and identity is also sometimes referred to as subjectivity (or subjectivities in plural). Powerful social actors can forge identities through the use of discourse, ideology, social interaction, and material means. Political elites regularly deploy policies and discourses that promote ideologies corresponding to their political beliefs. Such policies in turn affect social relationships and resource allocation. Together, such mechanisms reify and reproduce the desired identity among the wider population. Notwithstanding the hegemonic practices of powerful social actors, individuals and the social groups to which they belong are also capable of enacting agency to contest and resist dominant ideologies and identities.

Research within the social sciences also foregrounds the emotional aspects of subjectification that contribute towards how identity is shaped. Studies on the way emotions guide identity-making are gaining traction, led by disciplines such as cultural studies, sociology, and geography. In these readings of identity, emotions act as a type of logic that guides decisions and actions, thus contributing to subjectification. But emotions are amenable to regulation
by individuals or other social actors, and emotions circulate in society to shape individual and collective human relations (Hochschild 1983; Ahmed 2009). Studying emotions reveals the emotional logics by which identities are formed intersubjectively and in turn constitute social structures (Ho 2009).

Although subjectification is often thought to be spontaneous, individuals and social groups regularly embody and perform the identities they aspire towards or wish to preserve (Butler 1988). Class identities, for instance, are differentiated in terms of access to income, educational levels, and occupational status. But class identities are also distinguished by way of moral and behavioral attributes associated with the “rough” working class compared to the “respectable” middle and upper classes. Such class identities are performed regularly to maintain class distinctions, but class identities are fluid in other ways. Class identities can be improved by engaging in consumption that represents upward social mobility or efforts to practice the mannerisms associated with a particular class status; equally, class identities can be degraded as a result of deskilling experiences. The fluidity of identity suggests that certain features of identity can be invoked situationally to assert or downplay social affiliations depending on the circumstances.

Geography and identity

Geographers have considered how identity marks out the body as a type of spatial representation. When markers of identity are inscribed onto bodies, these bodies become set apart from others. “Race” or ethnic identity is manifested on the body not only by phenotype but also by attire or comportment. Bodies bearing racial differences that are considered incompatible with the mainstream society become vulnerable to abjection and spatial exclusion in their everyday lives. Whereas earlier geographical research focused on mapping ethnic segregation in urban spaces, geographers today are more likely to be attentive to how identity-driven discrimination contributes to spatial exclusion and multiple layers of social disadvantage. “Race” also cuts across wider national, cosmopolitan, and postcolonial identities and spatial referents.

The politicization of identity by different social groups can result in place contestations as competing groups stake claims to material and immaterial spaces. Such spaces can bear symbolic meaning for the social groups in question or contain resources coveted by those groups. Geographical research further argues that identities are not only situated in space but also constituted through space. For example, divisions in gender identities maintain a public and private divide that entrenches women’s reproductive role in the home by relegating them to domestic spaces. Gendered ideologies are also more likely to associate the productive roles of men with public spaces. Such spatial differences privilege masculinity and the earning power of men whereas the work performed by women tends to be undervalued, thus reinforcing unequal social relations and positioning women’s identities as subordinate to men.

Other geographical concepts also contribute to identity-making or trigger identity change. Scale refers to levels of social organization, such as at the home, community, or nation-state. They influence social relations contributing to one’s personal or communal identity. An influential set of debates on identity from the late 1990s onwards concerns the scalar relationship between the local, national, and global. Economic and cultural globalization, particularly that representing “Western” values, are thought to transform and challenge “local” or national identities. Under siege, scalar units at
the local and national levels reassert themselves through community-based activism, ideologies promoted by the state, or claims to pan-regional identities.

Inasmuch as the global has a bearing on the national and local scales, identities at certain sub-national levels can also cut across scale to leverage resources transnationally. Transnational religious movements are an example of how locally based religious identity politics can escalate to the global scale in order to advance certain parochial visions of national identity (van de Veer 2002). Geographers are attentive to how globalization and transnational migration prompt greater awareness of global interconnectedness across spaces and relational identities (Massey 2004). As flows of people, information, and resources become concentrated in key nodes, greater interconnectivity between such nodes results in distinct networks taking shape. In this rendering, a topology of flows and networks characterizes the geography of globalization even as scale remains useful as an analytical tool to examine the power relations that mediate identity-making (Leitner, Sheppard, and Szairto 2008).

Identity and membership in citizenship spaces

Identity connotes a personal dimension that defines selfhood, but it also has wider implications for social and spatial inclusion or exclusion from community and national membership. Belonging or nonbelonging to a collective body is premised on perceptions of difference, which is in turn defined by axes of identity. Identities function as a reference point for the social location of attachment and belonging. This can act as both a centripetal and centrifugal force, simultaneously drawing in those considered as belonging to a community while expelling others that do not fulfill the criteria for belonging.

At the national level, identity is considered a key component in the institution of citizenship, which in turn defines the extent to which a person can claim the rights associated with citizenship as well as the type of citizenly responsibilities ascribed to that person. Citizenship delineates territorial nation-states externally while internally it determines belonging, membership, and rights. The social processes defining citizenship spaces can have symbolic and emotional effects on identity, as demonstrated by nationalism claims, and even material effects in terms of resource allocation. This raises critical questions about who defines identity and how identity is shaped in citizenship spaces.

“Race” or ethnicity, for example, is a controversial topic in immigration debates on identity, belonging, and citizenship. Criteria such as phenotype, language, and cultural practices are used to classify new immigrant arrivals and differentiate them from the citizen population. Racial categorizations can contribute to social prejudices that feed into perceptions of group membership or nonmembership. Rhetorical devices invoking emotional logics are used to advance the superiority of the self versus the abjection of the “Other,” portray immigrant cultures as a threat to national identity, and create binary framings delineating winners and losers. These representations are regularly mobilized for legitimizing policies that exclude or subordinate social groups considered different from the mainstream society.

National governments play a key role in setting the tone for the extent to which plural cultural identities receive official recognition and equal treatment. Multicultural policies, for example, represent the nation-state’s willingness to accommodate identity differences while stressing the need for harmonious intergroup relations within
the nation. The ideology of multiculturalism is advanced discursively through government pronouncements as well as socially and materially by way of policy guidelines over schooling, housing, or workplace relations. Such policies influence the degree of social interaction or segregation between plural cultural groups. While the model of multiculturalism encompasses diverse types of identity rights (e.g., gender or sexuality rights), it is more often used to refer to ethnic or religious difference associated with immigration. State-sponsored multiculturalism has been critiqued, however, for failing to cultivate immigrant integration fully. This gives rise to competing representations of the nation to be advanced by chauvinist agendas, oftentimes provoking emotional responses to promote indigene identity and limit ethnic minority groups to partial citizenship rights.

Nonetheless, intersecting axes of identities contribute to a variety of immigration and citizenship experiences. Markers of identity to do with class, gender, sexuality, or religion can contribute to multiple layers of abjection experienced by certain ethnic minority groups. Yet it can also privilege the social status of other groups and concentrate resources in their hands. Although identity experiences and negotiations are anchored in citizenship spaces they can also transcend the national to span wider referent points, particularly in postcolonial contexts.

Identity politics writ large in postcolonial nation-states

Identity negotiations at the personal, community, and national levels are inflected by wider structural systems and macro-ideological beliefs arising from past or present colonial relationships that exist alongside contemporary globalization and transnational migration trends. Postcolonial readings of identity underline the colonial histories of violence and dispossession that continue to shape people’s social worlds today, reworked through globalization processes and neoliberal market forces. Postcolonial identities are constituted by Anglo-Eurocentric cultures of the past as well as in the present. They remain significant in postcolonial nation-states that grapple with colonial legacies even as they aspire towards building a national identity and citizenship. Writing about postcolonial cities, Yeoh (2001) argues that they represent sites where claims of an identity different from the colonial past are expressed. Likewise, accounts of postcolonial identities in citizenship spaces are attentive to the colonial discourses and practices that constitute belonging and membership. But they also assert the agency of colonized peoples, especially when defining what citizenship means in postcolonial contexts.

Globalization and transnational migration accentuate old and new divisions of identity that are borne by the figure of the ex-colonial subject and in the postcolonial nation-state. Former colonies, often bearing migration histories, have become sites of new migration as the demands of global economic competition drive transnational migration (Fechter and Walsh 2010). In postcolonial immigration contexts this development could mean inflows of expatriates from former colonial metropolitan centers because they are seen to embody cosmopolitan dispositions and skills coveted by industry employers, policymakers, and political elites. Also converging in these postcolonial settings are other inflows of less privileged migrants from poorer former colonies. They undertake menial work considered undesirable to the local labor pool. This meeting of postcolonial migrations also crosscuts with other types of racialized, classed, and gendered identities. In the ensuing
Identity politics, some differences are bridged while other social divisions become exacerbated in postcolonial citizenship spaces.

On one level, markers of white Anglo-European identity associated with colonial metropolitan centers set apart “Western” expatriates from “nonwhite” migrants who come from former colonies. On another level, there are internal stratifications within both of these referents as well. The category of Western expatriates is characterized not only by nationality or gender distinctions but also by classed identities that span a range of occupational and skills profiles. In this respect, class identity is more likely to forge commonalities regardless of racial difference, differentiating the privileged transnational capitalist class from not only transnational labor migrants, but also less privileged citizens of postcolonial nation-states.

On another level, notwithstanding such bridging axes of identity, racialized social hierarchies still matter in how social groups make sense of their identity relationally. “Western” expatriates may mobilize the legacy of colonialism to portray the identity of ex-colonial subjects, even the citizen population of postcolonial nation-states, as inferior to self. But ex-colonial subjects are just as likely to invoke situational identities advantageously. In the popular imagination, Western norms have become associated with cosmopolitan attributes and modernity. At times, ex-colonial subjects perform what they perceive to be cosmopolitan identities in order to accumulate cultural and social capital for advancing their upward social mobility (Ho, 2011). Yet when confronted with aspects of Western identities that they resist, ex-colonial subjects lay claim to other place identities, such as pan-regional “Asian” or “African” values, to affirm their difference and claim their rightful place. The key point is globalization and transnational migration processes in postcolonial citizenship spaces complicate identity politics but also demonstrate sharply how place histories and representations matter in the making of relational and situational identities.

Identities have the power to enforce social and material inclusion or exclusion in the places that shape them. A person can embody multiple identities and perform certain identities situationally. Identities are thus fluid in nature and relationally formed or performed. Subjectification is also underpinned by the emotions and emotional logics that contribute towards shaping social relations, social structures, and intersubjectivity. The processes of identity-making are spatially situated and further constitute place meaning. Identity politics grow in complexity when people draw in wider social referents to make sense of their social worlds. Among them are structural systems and macro-ideologies that cut across the geographical spaces and scales in which identities are formed.

SEE ALSO: Citizenship; Class; Cosmopolitanism; Emotional geographies; Gender; Intersectionality; Migration: international; Postcolonial geographies; Race and racism; Religion

References

IDENTITY


Imaginative geographies are representations of peoples and places that express the perceptions, desires, fantasies, fears, and projections of their authors, who are generally external observers. Power is an integral part of imaginative geographies, which are both a tool and an outcome of colonialism and imperialism. Within these broad social fields and political histories, imaginative geographies are indicative of the power dialectic that exists between the observers and those who are observed, the colonizers and the colonized. Reification of self and other is an integral part of imaginative geographies and, as these systems of knowledge develop and become entrenched, the reification can go both ways.

The concept of imaginative geographies was formulated by the Palestinian American post-colonial critic Edward Said (1935–2003) in his analysis of Orientalism, originally published in 1978. A professor of English and comparative literature at Columbia University from 1963 to 2003, Said worked primarily with text, visual representations, and other cultural productions to build his critique of Orientalism, which, he stated, emerged in the Christian West in the fourteenth century as a field of learned study based on a geographical, cultural, linguistic, and ethnic unit called the Orient. Although his approach was rooted in discourse analysis, Said always emphasized the geographical underpinnings of his concept, presenting Orientalism as a “more or less total geographical position toward a wide variety of social, linguistic, political and historical realities” (Said 1978, 50). Therefore, as a structured field for the production, acquisition, and dissemination of knowledge about peoples and places, Orientalism is at its core an imaginative geography. One of its key characteristics is that it constructs and circulates a binary order of meaning where East and West function as opposite categories. Said’s notion of imaginative geographies is anchored in Michel Foucault’s (1970, 1977) analysis of the creation of truth regimes via discourse and episteme. As colonial truth regimes, imaginative geographies are profoundly polarizing, producing and solidifying self and other as inevitably separate and opposite rather than co-constitutive: because they render what is foreign familiar to the observer, the other — that is, the one who is being observed — can only derive his or her identity negatively (Said 1978, 64). Such ordering serves to secure a specific place for each thing so that it can be located again and again, “therefore giving things some role to play in the economy of objects and identities that make up an environment” (Said 1978, 360).

In the light of these elements, imaginative geographies can be understood as particular forms of power that are constituted not only through discursive formations but, more specifically, through the production of authority by way of discourse. As systems of knowledge, imaginative geographies facilitate the circulation, citation, reiteration, and performance of various meanings, with the result that these meanings become dominant, eventually acquiring the status of truth; this despite the fact that they are neither chosen nor shared by the people
to whom they refer. Thus, Said’s concept of imaginative geographies intersects with some aspects of behavioral geography, which posits that what individuals believe about a place or environment will largely determine their spatial behavior independently of (purported) factual data. One of the chief consequences of this iterative process is that objects that are made by the mind will end up appearing to exist as objective realities even though they have only a fictional identity (Said 1978, 54). Said gives the following example:

A group of people living on a few acres of land will set up boundaries between their land and its immediate surroundings and the territory beyond, which they call “the land of the barbarians.” In other words, this universal practice of designating in one’s mind a familiar space which is “ours” and an unfamiliar space beyond “ours” which is “theirs” is a way of making geographical distinctions that can be entirely arbitrary. (Said 1978, 54)

Again, the assignation of a people and their space to the negative side of the binary system does not require their acknowledgment for the colonial regime of truth to operate. Therefore, as a form of generalized ethnocentrism, imaginative geographies set up a universe of meaning and practice where the “subaltern cannot speak” (see Spivak 2010).

And, yet, the concept of imaginative geographies problematizes the relationship between knowledge and geography, showing that no simple dichotomy can exist between “real” and “perceived” worlds, and underscoring the ways in which discursive formations become geographical formations. Although Said elaborated his theory from literary and other cultural productions, the key point of Orientalism and its attendant imaginative geographies is that they enter fully into the creation of material worlds, dramatizing as they do the distance – cultural, political, epistemological – between what is close and what is far away, notably through notions of “home” and “abroad,” “our space” versus “their space” (Said 1978, 55). Said vividly captured this issue when he stated that “Racism is spatially as well as socially constituted” (1978, 3): from the Great Wall of China, to Native American reserves, or the South African apartheid system, worlds are built on the basis of ideas. Consequently, Said does not propose to rectify the error of ethnocentrism by countering its mistakes with another truth, that of the insider. Such an approach would contradict the Foucauldian notion of power that informs his theory, as well as the discourse–materialism dialectic on which his concept of imaginative geography is founded. The political scientist Timothy Mitchell (1989) has critically engaged this question in his own work: drawing on Said as well as on Walter Benjamin, he has referred to the play of colonial representations that enter into the constitution of the modern world as a “labyrinth without exits.”

In the early 1990s Derek Gregory enjoined geographers to engage more fully with Edward Said’s work and to chart the evolution of his theory from a discursive analysis of cultural productions to what Said termed “the struggle over geography,” to which his inquiry had led him (Gregory 1995, 447). Gregory’s chief interest was to develop more critical approaches to understand human geographies, approaches that allowed for more integrated analyses of cultural, historical, and political landscapes. As Gregory states, the effects of cultural imperialism and its geo-imaginings were not limited to epistemological violence, as is strongly indicated by the fact that Said’s analysis of Orientalism is “framed by a series of wars” (Gregory 2012, 151). Against that background, Gregory’s own book, The Colonial Present: Afghanistan, Palestine, and Iraq (2004), is a rigorous engagement with the material effects of the cultural-political performances of
imaginative geographies. In it, he traces one of Said’s key notions about geographical imaginations – namely the dramatization, and reification, of the distance between home and abroad – into the political geographies of the “war on terror.” Analyzing the cultural and political responses to September 11, 2001, Gregory underscores how, as they “circulate in material form,” imaginative geographies render some of the most extreme forms of violence onto the lives of ordinary people. Gregory’s use of the expression “colonial present” signals both the limits of and the continuing need for postcolonial studies. Offering a meticulous illustration of what Said meant by “the struggle over geography,” The Colonial Present traces the spatiotemporal reach of British colonial imaginings into the contemporary chaotic landscape of a particular Orient, namely the “Middle East” (Gregory 2012, 151). Through his engagement with Said, Gregory has contributed to revivifying the tools and approaches of historical geography, chiefly by building stronger bridges with the disciplines of critical geography and political geography. For him, the ordering of the Orient through a binary system of meaning translates into the political “disorder” of modern war, which he describes as a widening net of ocularcentric surveillance via increasingly sophisticated military means (Gregory 2012).

The question of how, through imaginative geographies, space relays and reproduces ethnocentrism has led to new avenues of inquiry in geopolitics, and brought closer attention to the ways in which the dialectics of violence produce their own geographies, thus deepening the interest in human geography for studying militarization (Cowen and Gilbert 2008) and geographies of violence (Gregory and Pred 2007; Barkawi and Stansky 2012). The dynamics of scale and globalization are key aspects of these analyses. In the current political context, the deep colonial roots of Orientalism’s imaginative geographies continue to activate different forms of polarization across global space, a process that was powerfully illustrated by George Bush’s unilateral placement of Iran, Iraq, and North Korea into an “axis of evil” in his 2002 State of the Union address. Against this background, critical geopolitics (Ó Tuathail 2003) and military studies in geography have analyzed how – on the international political stage – notions of “terror,” “security,” “homeland,” or “foreign land” are constituted within the paradigm of otherness and othering that is inherent to imaginative geographies. Interrogating how colonial histories of power continue to produce – and indeed multiply – violent geographies across the Earth, these conceptual developments have underscored how the ramifications of geographical imaginations across global space have a direct impact on peace and security, at home, abroad, and across the full spectrum of unevenly interlocking geographical scales.

With the overarching reach of the “war on terror” since the early 2000s, these debates have brought attention to the various technologies that are implicated in the production and circulation of imaginative geographies. In 1983 Benedict Anderson published an important book that resonated strongly with Edward Said’s concept of imaginative geographies, Imagined Communities: Reflections on the Origin and Spread of Nationalism, where he analyzed how print capitalism enabled the imagination of a shared “communion” between members of a nation, despite the fact that they will never meet each other face to face. As Anderson put it, “the members of even the smallest nation will never know most of their fellow-members, meet them, or even hear of them, yet in the minds of each lives the image of their communion” (Anderson 2006, 6). Through their sedimentation over time, imaginative geographies may
very well have produced the opposite effect where, as it were, the West will never meet the East and yet in the minds of each will live the image of their separation. Like Said, Anderson never lost sight of how the production of the imagined community of the nation simultaneously generates the outside of that political community: as the collective “us” of the nation is imagined, so is the “them” of those who do not belong to it. Imagined Communities was reprinted in 1991 to include chapters on the census and the map, which, in addition to mass media, also contributed to the construction of a shared national space. As it relates to geographical imaginings, Anderson’s development of his theory reminds us of the importance of technologies in sustaining and renewing colonial (dis)orderings of global space, and in relaying distanciation through means not only of the dramatization, but indeed the materialization, of distance. If segregation and exclusion (Sibley 1995) have always contributed to the shaping of human geographies, modern military technology has exacerbated the production of distance between self and other. From the armored ground vehicles and aircraft of World Wars I and II, to distance surveillance systems, unmanned air and ground combat vehicles, and the recent development of drones, the modernization of warfare can be interpreted as an attempt to increase the distance and, ultimately, to eliminate all forms of contact, between human combatants. Indubitably, this process contributes to the upsurge in the externalities of war and their increasing reach into the everyday spaces of civilian lives. While Edward Said stressed the importance of discourse in the production of imaginative geographies and its many forms of violence, geographers have contributed to outline and analyze the role of spatial tools and technologies in materializing these realms.

Lastly, imaginative geographies also reach into the politics of nature (Braun and Castree 1998; Castree and Braun 2001) and there is a growing body of work in political ecology and environmental history looking at the role of imaginative geographies in producing dominant understandings of nature and the environment. Colonialism produced various normative representations of nature, chiefly through representations of a “normal” — meaning temperate — nature, as opposed to an “unnatural” nature, generally referring to environments that correspond to tropical, desert, or polar regions of the Earth. In The Road to Botany Bay: An Essay in Spatial History (1987), Paul Carter brought together history, discourse analysis, and cultural geography to examine how mapping, narrative, urban planning, and aboriginal mythologies participated in the imperial conquest of Australia. As Carter demonstrated, spatial knowledge contributed to bring nature within the colonizers’ realm of intelligibility, a process he viewed as being key to colonial dispossession.

Stemming from the fine-grained analysis of culture that was the hallmark of Edward Said’s work, the concept of imaginative geography allows for a deeper understanding of the work of culture in bringing places, resources, and environments into view in and through the ontology of the state (Willems-Braun 1997), the market, the military, and the various extractive industries that are related to them. In Canada, Caroline Desbiens has examined how — during the rise of the Quebec nationalist movement in the 1970s — hydroelectric development in Northern Quebec relied on what could be described as “Norientalism” (2013, 12), that is, an imaginative geography of the north that denied its aboriginal histories and subjectivities to envision its development narrowly, according to the economic and political priorities of the south, in
this context the urbanized core of the province. The geographical coordinates of this Canadian dialectic between north and south are reversed in the international arena but the imaginative geographies that sustain it are similar: indeed, critical studies of development demonstrate how frequently peripheral spaces function like a “theatrical stage” (Said 1978, 363) that is seen to exist solely as an extension to the center, whatever and wherever that center may be, depending on scalar shifts and dynamics. Whether they are activated by racism, colonialism, terror wars, resource extraction, or performances of national sovereignty, imaginative geographies are ultimately power moves. As such, we cannot forget that they necessarily generate various forms of resistance and activism, which in turn aspire to the creation of alternative geographies that are hopefully built on more hybrid, dynamic, and relational versions of the world.

SEE ALSO: Behavioral geography; Colonialism, decolonization, and neocolonialism; Critical geography; Data model, event-oriented; Discourse; Environmental history; Geopolitics; Globalization; Historical geography; Imperialism; Ontology: theoretical perspectives; Orientalism/Occidentalism; Political ecology; Political geography; Postcolonial geographies; Scale; Socio–nature; Violence

References

IMAGINATIVE GEOGRAPHIES


Further reading

The concept of imperialism and related ideas of empire were restored and rejuvenated by the United States’ invasion of Iraq in 2003. Previously, people had preferred the more neutral term “globalization.” Imperialism was avoided as outdated, pejorative, or left-wing, and its pre-World War I reference to investing capital in the colonies (Lenin 1964/1917) appeared to be especially outdated as transnational movements of capital are now mostly between advanced countries, with ex-colonies relatively neglected. When referring to the United States’ dominance over other countries, some geographers preferred the more consensual term “hegemony” (Agnew 2005, 12–36). It was often unclear what “empire” and “imperialism” could mean in today’s world, when they had apparently been superseded by national states and nationalism. Where it was discussed at all, imperialism was usually confined to an “age of empires” that was safely located in the past.

These views were either unnecessarily constraining or simply mistaken. The Iraq invasion showed that imperialism and empire were still very much alive (if not necessarily very well). Multiple publications appeared with these terms in the title (e.g., Harvey 2003; Smith 2003; Colás 2007; Münkler 2007), as supposedly left-wing terminology was “legitimized” by neoconservatives advising President Bush to invade. Vice-President Cheney, Secretary of Defense Rumsfeld, and Deputy Secretary of Defense Wolfowitz and others effectively declared that the United States was an empire – indeed, the empire – and should actively assert its global hegemony. A British neoconservative historian weighed in with the suggestion that global disorder was due to America not behaving like an empire (Ferguson 2004): it should learn from British hegemony and police an unruly world instead of being reluctant imperialists, castigating the British empire as the 1776 butt of their own anti-imperialist credentials (forgetting, it might be added, their own Manifest Destiny imperialism which created the continental United States on Indian territories).

For some the Iraq invasion signaled a “new imperialism,” although in truth it was a lot like the “old” historical imperialisms which stretch back to Sargon of Akkad, history’s earliest named person, who around 2300 BCE created the Sumerian empire in what is present-day Iraq (Mann 1986, 131–150). Imperialism traditionally meant a state taking over other people’s territories by force and theft; sometimes ruling more indirectly on the empire’s outer frontiers through puppet governments, local elites, or client states; establishing colonies, sometimes for economic reasons (oil loomed large in Iraq); and all typically with the moralistic self-justification of a “civilizing” mission (to democratize Iraq), for beyond the expanding frontiers of empire people have traditionally been seen as “barbarians,” with rival empires adding to the threat. Indeed, it was these historically familiar features of empire that quickly popularized the notion that imperialism had returned.

The reality was quite different but less obvious: imperialism had never gone away. When Europe’s overseas territorial empires were...
dismantled after World War II, supposedly ex-imperial powers were occasionally accused of neo-imperialism or neocolonialism in former colonies; imperialism continued in more hidden forms, pursuing similar objectives by different means. It was adapted to fit the new era of politically independent national states. Instead of occupying or annexing territory and making formal claims to rule, imperial powers switched to working more subtly and indirectly within and through independent states. It was now an informal and largely nonterritorial imperialism, shaping developments in its own interests by a “carrot and stick” approach, a variable mix of economic, cultural, and political inducements and pressures. Instead of formally claiming territory, it generally went along with the conventional assumption—or useful fiction—that all states have equal sovereign independence. This more indirect or informal type of dominance, and not the ill-fated Iraq adventure, was the real new imperialism.

Its better fit with independent national states—where nationalism can effectively mobilize against obvious “foreign interference”—was confirmed, ironically enough, by the costly debacle of the old-fashioned and unsubtle imperialism of the Iraq invasion; likewise by the failed Russian and US occupations of Afghanistan (1979–1989 and 2001–2014, respectively). Furthermore, informal imperialism is not really new, only much more generalized. It emerged with the rise of capitalism and, more specifically, in the late nineteenth century when the “gunboat diplomacy” of the British empire’s navy supported the profitable operations of private British capital in the already independent states of Latin America. The overseas empire of the United States, almost entirely informal, emerged shortly afterward (Smith 2003, 9–17).

We cannot do justice to imperialism’s complex history over several millennia. Instead, with contemporary relevance as the touchstone, the focus in this entry is on the general transformation of imperialism since the French and Industrial Revolutions. The first ushered in nationalism as a major force in politics; the second clearly transformed the material capacities and infrastructural (as distinct from despotic) power of states (Mann 1986, 27) and the logistics of ruling over extensive territories (e.g., using railways, telephonic communications, and modern bureaucracy). Less obviously, the spread of capitalism revolutionized the relations between political and economic power that made possible the transformation to the new imperialism.

While for some geographers imperialism is political and military, others recognize important political economy dimensions. Some emphasize the exploitation of less developed countries, but a broader view holds that imperialism and imperial rivalries constitute the global geopolitical-geoeconomic system (Callinicos 2009, 3–14). This entry first characterizes old imperialism and formal, territorial empires as an ideal type, contrasted with national states. But ideal-typical simplifications can mislead unless combined with more concrete analysis. Second, the entry turns briefly to how imperialism and nationalism have now been interacting for over two centuries, rather than imperialism being replaced by nationalism. Third, it argues that the partial separation or “contradictory unity” of political and economic forms of power, and their territorial implications, explain the eventual dominance of the new, informal, imperialism. Crucially, they allow foreign direct investment to escape being seen as foreign interference, and thus facilitate the coexistence of a single global economy with multiple independent states despite the apparent contradiction. However, new imperialism is not the whole story. Fourth and finally, there are also very significant continuities from
the old imperialism, especially the continuing importance of territoriality, theft, and physical force, particularly for the hegemony of leading powers. The invasion of Iraq was not entirely an aberration or a throwback. The indirect power of informal imperialism can mean less power, which occasionally calls for old-fashioned imperialist interventions despite the often costly consequences (especially if there is the subsequent aberration of prolonged occupation).

**Old imperialism and its heterogeneous territories**

Historically the quickest way to accumulate wealth was to steal other people’s possessions – “primitive (first) accumulation” or “accumulation by dispossession” (Harvey 2003, 137–182). For state elites (e.g., monarchs, aristocratic hangers-on) it meant stealing other people’s territories. Theft was central in the territorial formation of states and of empires: initially, the distinction was largely one of geographical scale, which changed through history, with early state formation effectively the imperialist expansion of empire (e.g., the continental United States; or “Wessex imperialism” amalgamating separate territories into the Anglo-Saxon kingdom of England). Interestingly, up to the mid-nineteenth century the term “British empire” meant amalgams of territory within the British Isles, while imperial possessions beyond the core UK state were referred to as “the British empire in North America,” “the British empire in India,” and so forth.

Imperialist accumulation used a diversity of violent and nonviolent means, including territorial invasion and military occupation, colonization and settlement, dynastic marriage, barter or purchase from other powers. Piecemeal expansions mostly resulted in a heterogeneous hodgepodge of territories, very different from each other and from the imperialist or metropolitan core state, in terms of ethnicity, culture, economy, and political rule (the multination United Kingdom of England, Scotland, Wales, and Northern Ireland is itself an example).

Giovanni Arrighi (1994, 33–36) identifies two logics behind imperial strategy. A territorial logic of accumulating wealth and power through accumulating territory – its natural resources and agricultural producers the main source of wealth – was the dominant logic historically. State elites also followed a capitalist logic, where merchant capitalists amassed wealth by economic trading. But trading was usually subordinated to the logic of territoriality associated with the ruling landowning classes: capital accumulation was a means of acquiring territory rather than a priority in its own right. In the merchant-led city-states of medieval Italy, however, the capitalist logic became more central, and it dominated in the city-empire of Venice. Here the priorities were reversed and territorial acquisition, rather than being an end in itself, was subordinate to accumulating wealth through trade. Venetian wealth and power were based on seaborne trade and determined a network form of empire, not extensive territories but trading posts and ports around the eastern Mediterranean, with the trade routes between them protected by the large Venetian navy. This system, however, was exceptional. Most empires were primarily territorial up to the mid-twentieth century, although the capitalist logic also featured strongly in the hegemonic empires of the Dutch in the seventeenth century, and even more the British in the nineteenth century – by which time territoriality was combined with the much more powerful logic of
industrial, as distinct from merchant, capital. And now the capitalist logic clearly predominates in the largely nonterritorial new imperialism.

The two logics are associated with the dual character of capital accumulation: through capitalist production and “a transaction between the capitalist and the wage laborer in apparent peace, property and equality,” but also through the “other aspect of accumulation” involving “force, fraud, oppression, looting … political violence” — in the words of Rosa Luxemburg (1968/1913), as quoted by David Harvey (2003, 137–149). The latter, overtly violent, aspect clearly predominated in the old imperialism, and very obviously so in its colonial contexts, but it still plays a role in contemporary imperialism.

Key features of the old imperialism’s territorial empire as an ideal type are highlighted by contrasts with the more familiar nation-state. Where the idealized space of the nation-state is uniform, at least theoretically homogeneous, with all parts of the state territory effectively the same (e.g., in citizenship rights), the territorial empire was heterogeneous, with multiple, differentiated spaces and diverse interrelations between them. The citizens of the nation–state belong directly to the nation and the state as individuals, whereas people belonged indirectly to the empire, by virtue of first belonging to one of its many constituent communities and territories. On the face of it, pragmatic territorial empire and nation–state could hardly be more different, the latter reflecting doctrinal principles of nationalism, its idealized objectives making it more than simply an analytical ideal type.

According to the doctrine, the nation and its territory should be culturally homogeneous within clear state borders; all nations have an equal democratic right to self-determination (self-expression, self-protection) in their own independent states; and the borders of nation and state should coincide geographically. Such a happy coincidence rarely, if ever, occurs in reality, however, and attempts to make reality fit the nation–state ideal typically result in serious ethno-national conflict.

In contrast, the old imperialism was arguably more grounded in reality. It was generally guided by pragmatic power politics and unburdened by (often unachievable) principles — its justificatory civilizing missions were perhaps the nearest it got to principle. Instead of attempting to culturally homogenize, it was generally more flexible, inclusive, and accepting of cultural and ethnic difference; however, it was often more overtly hierarchical and more likely to accentuate differences between ethnic groups in order to divide and rule diverse populations. With its typical commitment to expansion and to hierarchy (especially of the metropolitan core over the peripheries), imperialism did not presume any sovereign equality between separate national or other entities, and it was much less concerned with (even lip service to) democratic principles. It depended largely on the military and political power of the core state to capture and control peripheral regions, and it ruled or exerted influence within often fuzzy boundaries or indeterminate frontier zones rather than within precisely demarcated borders (Colás 2007, 62–63).

Toward the frontier, its rule generally became more indirect and informal, sometimes operating through client states, with central power declining in unstable spheres of influence, which often overlapped with the spheres of rival empires. Whereas state power and cultural identity tend to be relatively uniform throughout the territory of the modern national state (or so it is hoped), and usually end in a sharp inside/outside dichotomy at the border, power in territorial empires generally declined on a gradient with increasing distance from the metropolitan core, and petered out more gradually on the frontier. This was
often the case de jure and nearly always the case de facto. Even where the introduction of modern technological infrastructures (e.g., railways, telephones) greatly facilitated rule from a distance, the more peripheral territories furthest from the core were still more likely to have greater ethnic, cultural, ideological, or economic differences, and be more susceptible to influence from rival empires. The core–periphery/distance–decline power gradient is a key distinguishing feature of the old imperialism.

As a very oversimplified ideal type, territorial empires comprised three concentric circles: the inner core state; a broad middle zone of heterogeneous territories and colonies ruled from the core in a variety of more or less direct ways; and an outer frontier of more informal, indirect, and often unstable rule, or merely influence, through local elites or client states.

The transition from empires to national states

Ideal–typical contrasts initially enlighten but can mislead. There were complex interweavings and mutually supportive, as well as conflictual, relations between imperialism and nationalism. There was no simple opposition or linear replacement of imperialism, contrary to the perspective of some nationalists and academics who see nationalism as only “an anti-imperial ideology” (Münkler 2007, 123). History is a lot more complicated. Imperial powers fostered nationalism; they themselves adopted nationalist strategies; some nationalisms allied with rival imperialisms; and some developed their own imperial pretensions (Anderson and O’Dowd 2007).

Some of the earliest nationalisms developed in the core states of the imperial powers (e.g., in France and England), and they spread nationalist ideals at home and abroad, and not least to their own colonies. Links between the metropolitan core and cities in the peripheries were often crucial in the promotion of nationalist and democratic ideals of self-determination, obviously appealing to opponents of authoritarian colonial rule. Resistance usually intensified when the empires modernized and, to varying degrees, adopted nationalistic policies toward their subject peoples, essentially attempting to extend the nation-building of the empire’s core state, hoping thereby to enlarge the core and/or its control over peripheries. Initially, the combined forces of capitalism and nationalism probably boosted traditional imperial power as much or more than they boosted oppositional national movements. Capitalism, by transforming the logistics of rule from a distance, gave traditional empires a new lease of life, and central controls generally became more pervasive and effective. Compulsory state education from the late nineteenth century was an opportunity for some imperial powers to make schools in the periphery use the language of the core (e.g., German rather than the local Polish in parts of the Prussian empire; the attempted Russification of the czar’s Finnish territory). But such nationalistic policies were a significant departure from traditional pragmatic imperialism (contradicting its ideal type), and they would generally prove counterproductive, galvanizing nationalist opposition.

Ultimately, nationalist opposition would prove disastrous for the old imperialism. But it is important to grasp the latter’s resilience. Most of Europe’s territorial empires were still firmly in place up to World War I, and its overseas empires in Africa and Asia outlasted World War II. Their dismantling was often due to national anticolonial movements, but also to US opposition to territorial empires: the American empire was very largely of the new, informal type, and Europe’s formal empires had often marginalized or excluded US business interests.
IMPERIALISM

It has only been in the last 50 years or so that the very long era of territorial empires (such as the USSR up to 1989) eventually fizzled out and national states became the global norm, and only then that the new imperialism became fully and firmly established. Traditional empires incorporated and benefited from nationalism and capitalism before eventually being undone by them.

But imperialism adapted: the three concentric circles of the ideal-typical old territorial empire have effectively been reduced to two: the broad middle zone of heterogeneous territories ruled more or less directly from the core has been removed, leaving the inner core state and a greatly expanded surrounding zone of more informal, indirect, but now predominantly economic power over independent national states, sometimes with geopolitical influence exercised through client states (e.g., Israel for the United States, Cuba for the old USSR).

Politics and economics in capitalism and the new imperialism

Nationalism necessitated the transition to the new imperialism, while capitalism enabled it. Intercapitalist competition ensured the continuation of a multiplicity of states, while the contingent struggles of nationalism and imperialism decided which ones. There seem to be contradictions between the politics of multiple states and the economics of a single global market, between the independence of national states and the interdependencies of economic globalization. However, the new imperialism and inter-imperial rivalries now constitute a combined geopolitical-geoeconomic system unified by a predominant capitalist logic.

These related propositions are best explained by the theory of capitalism’s partial separation or “contradictory unity” of political and economic forms of power, their mostly separate institutionalizations, and the territorial consequences. Capitalism not only greatly increased state power, but it also revolutionized the nature of power. Economic power, now associated with the capitalist market, became more distinct from political, including military, power associated with the state; and capitalist economics are now arguably more important than national states as sources of social power. But the “new,” and to a large extent economic, imperialism is far from transparent, and theory is needed to reveal its inner workings.

As the late Ellen Wood (1995) explains, the partial separation of politics and economics arises uniquely from within the capitalist mode of production, from the way in which economic surplus is extracted from the producers. In pre-capitalist production systems the surpluses had typically been extracted by extra-economic or political means, often relying on the direct use of physical force at the point of production (e.g., armed overseers on slave plantations, the private armies of feudal landlord). In sharp contrast, surpluses in capitalism are mainly created by the “free” labor of waged and salaried workers through the economic operations of the free market. Here, however, the absence of force is more apparent than real. In the partial separation of politics and economics, physical force is only displaced from the economic sphere of production to the political sphere of the state, less obvious than in slavery or feudalism but still necessary. In capitalism, state force maintains property rights and the rule of law, without which production and the so-called free market could not operate. Ultimately, force is necessary to exclude workers from possessing or using the means of production (e.g., land, machinery, raw materials) to produce their own necessities of life, generally leaving them no alternative but to
sell their labor power to employers who possess the means of production.

The degree of partial separation varies and is politically contested. It is clearest in liberal or neoliberal regimes, where the market is clearly dominant. But even in the state capitalism of state-owned enterprises there is partial separation because these enterprises usually follow the dictates of the market more than political criteria. Furthermore, as in privately owned firms, the partial exclusion of politics not only applies to physical force, but it is also the exclusion of democracy from the major decisions about what to produce and where to invest — representative democracy stops at the gates of the workplace (points further developed in Anderson 2005).

The partial separation is expressed in the institutionalized division of labor between business people and state personnel, including politicians. It has several profound consequences for territoriality and is spatialized, as in the division between Wall Street, New York, and Washington, DC, or between the City of London and the City of Westminster. It transformed state–city relations, with cities generally becoming politically subordinated (de jure and de facto) within the more homogenized and centrally controlled space of the national state, and (with some exceptions) becoming more one-sidedly economic and cultural centers, where previously cities also had greater political roles or political autonomy. Crucially for the new imperialism, the separation transformed state territoriality, with borders differentiating more sharply between political and economic processes.

The extraction of economic surplus in precapitalist systems was inseparable from politics and often required force at the point of production; and hence extracting surplus in foreign territories readily called for direct political control through incorporation in formal territorial empire. For instance, Europe’s overseas empires combined slave labor with territorial possession: the Spanish operated mines in Spanish-held territories, the British sugar and cotton plantations were in British colonies, the Dutch plantations in Dutch colonies, the French in French, and so on (Rosenberg 1994). But such direct territorial possession was rendered unnecessary by capitalist production’s separation of politics and economics.

This separation enables foreign direct investment (FDI) to take place in politically independent sovereign territories; and, conversely, it enables states to make plausible claims to be independent, despite the interdependencies of economic globalization. The economic and political objectives are facilitated as two sides of the same coin: considerations of political sovereignty are in effect removed or forgotten where FDI is concerned; and considerations of economic interdependency are forgotten when political sovereignty is being claimed. The claims to sovereignty and national self-determination apply only in the sphere of politics, and they are usually recognized because economics is only marginally involved in this sphere and is treated as separate.

FDI, the bedrock of contemporary (mainly economic) imperialism, is generally accepted despite national states and populations being imbued with nationalist ideals of sovereign independence, because FDI is deemed to be economics rather than politics. Furthermore, democracy stops at the border as well as at the workplace gate, so in a sense FDI is doubly insulated from the sphere of politics. Market criteria could be enforced by independent states for foreign-owned as for indigenous capital, both having the same state-protected rights (at least in principle). Exploiting foreign labor could now be negotiated as a purely economic transaction between foreign employer and indigenous
worker – the foreignness of the exploiter being effectively depoliticized. Political control of the territory was no longer required. Hence, far from rejecting FDI as foreign interference, independent states fall over themselves trying to attract it – sovereignty is sacrosanct in politics, but financial investment is merely, well, economic. Political problems do arise from foreign decisions to disinvest, but the separation of politics and economics is crucial in defusing potentially explosive contradictions between sovereign states and economic imperialism, and it mostly ensures a relatively harmonious combination of many states and one global economy. Less benignly, and reminiscent of Lenin’s association of economic imperialism with colonies, it is also the basis of the “superexploitation” of very cheap labor in the Global South.

The global geopolitical-geoeconomic system is best seen as a single contradictory unity rather than as two ontologically separate systems which somehow “come together.” Following Arrighi’s (1994) formulation, the system is unified by a logic of capital accumulation, which now predominates over a territorial logic for state elites as well as it obviously does for business people. David Harvey (2003, 26–36, 183, 204), however, has reformulated this understanding, confining the capitalist logic to business people, with state elites allegedly following a different territorial logic of geopolitical competition to augment state power. The dualism of two different sorts of actors with two different motivations could, arguably, help to avoid an economic reductionist or functionalist approach to imperialism (Callinicos 2009, 15), but it is not necessary for avoiding reductionism and it has its own pitfalls and inconsistencies (see further Anderson 2005). Arrighi’s formulation of both logics as state strategies – alternative, complementary, or contradictory – seems clearly preferable. If the elites running the Venetian city-empire predominantly followed a capitalist logic, with capital accumulation as their main goal and state territory seen as only “a means and a by-product of the accumulation of capital” (Arrighi 1994, 33), that is surely at least equally, if not more, true of contemporary states in global capitalism. There is, however, more to the story.

Hegemony and continuities from the old imperialism

We have seen that capital accumulation has two aspects: violence-free economic production, backed at a distance by physical force, but also the fraud, theft, invasion, and violence of “accumulation by dispossession” (Harvey 2003, 137). The latter is now generally more muted or secondary than it was formerly in the old imperialism. It is not the dominant aspect, but it continues and surfaces strongly in questions of hegemony over the global system and whether states follow hegemonic powers voluntarily or are persuaded by economic coercion or by political violence (threatened or actual). John Agnew (2005) prefers “hegemony” to “empire” in analyzing the United States’ dominance, whereas Michael Mann (2013) sees the United States as the world’s “sole remaining empire.” But, while by far the leading military superpower and still hegemonic, the United States is far from being the only empire or practitioner of imperialism. Lesser powers also practice it (e.g., Serbia’s murderous aggression in the former Yugoslavia, Israel’s illegal colonial occupation of the West Bank and Gaza). Other major powers – Russia, the European Union, and rising powers such as India and Brazil – increasingly challenge the United States politically or economically, and China may potentially be a future contender for hegemony. Largely because of growing competition between imperialisms big and
small, US hegemony is starting to decline in an increasingly multipolar world. And all three concepts – imperialism, empire, and hegemony – are needed to make sense of the complexities.

The new imperialism can usefully be described as less territorial, less direct, more informal, and more economic than the old imperialism. But ideal types aside, real-world imperialism is now a mix of both. Territoriality and direct non-economic power and violence are still crucial, are sometimes used against independent states in the outer zone of informal empire, and are present in the empire’s inner core or all-important home territory. For instance, in the continental United States, the Russian Federation, or the Chinese Republic, power is still direct, territorial, and based on political sovereignty, whereas the rest of their empires are non-territorial and involve mainly economic power (e.g., China’s increasing economic involvement in Africa) – at least most of the time.

But core–periphery relations have a variety of aspects. First, operations in the periphery can produce problems of “blowback” in the core, such as the Home Rule struggle in Ireland which produced a constitutional crisis in Britain in 1912 (Anderson and O’Dowd 2007), while US support for Israel and military interventions in the Middle East arguably contributed to the terrorist attacks of 9/11. Second, success in the periphery depends on support in the core, at minimum its pacification, and here recruitment to the empire’s military forces played a key role, as did incorporating potential opponents in the periphery. Third, imperial success and leverage depend on the relative size and weight of the core state (e.g., Portugal was too small and weak to develop its vast and potentially rich territorial empire, so a self-serving British empire helped out by treating it as a junior partner). Fourth, the historical pattern has been for hegemony to require ever larger core states: the global hegemony of the small seventeenth-century Dutch Republic was followed by the hegemony of the medium-sized British national state in the nineteenth century, and it in turn was followed by the hegemony of the vastly larger United States in the twentieth century. Fifth, it can be inferred that the global hegemon or an eventual contender (perhaps China) now requires a home territory of continental proportions, partly because of the extra leverage needed to exercise hegemony in the more difficult conditions of informal empires and politically independent states. Sixth, practices of territoriality in the home territory are especially evident where the territory itself is being expanded and/or threatened. For example, China’s home territory in the far northwest, resource-rich, province of Xinjiang (meaning “new frontier” or “new territory”) is being threatened by the national separatist movement of Muslim Uyghurs, as China aggressively populates the region with Han Chinese settlers in an old-fashioned imperial expansion.

This process follows Arrighi’s “logic of capitalist rulers,” for whom territory is a secondary means toward accumulating capital, rather than an end in itself, but it nevertheless involves territoriality. Similarly, the eastward expansion of the European Union since the collapse of the USSR involves a different sort of capitalist empire enlarging its home territory and increasing its informal leverage. Here analysts mostly concentrated on the economic processes of securing former Soviet satellites as investment opportunities rather than treating them only as a source of cheap migrant labor, but typically they failed to see its military aspect (Armstrong and Anderson 2007, 125–141, for example), blinded by the partial but institutionalized separation of economics from politics, in this case the
European Union from the North Atlantic Treaty Organization (NATO). This failure was exposed in print by Perry Anderson (2007), who pointed out that all the former satellites admitted to the European Union had first been incorporated into NATO under US command (the Clinton administration reneging on US nonexpansionist undertakings given to the USSR’s reforming President Gorbachev). And it was exposed in practice in 2014 when EU/NATO expansion resulted in a confrontation with Russia and armed conflict in a Ukraine caught between the two. It looked very like classical territorial competition between empires in their overlapping spheres of influence. While the European Union’s external operations are typical of contemporary informal empires, the novel construction of its expanding home territory with a “ring of friends” on its complex outer frontiers (see Armstrong and Anderson 2007, 142–159) is paradoxically more like a traditional territorial empire combining heterogeneous units with a gradient of power declining toward the edges. Furthermore, the institutionalized division of labor between the “economic” European Union and the “political” NATO indicates that the European Union, rather than constituting itself a rival empire to the United States, is acting as its junior imperial partner – as advocated, for example, by Germany’s government adviser Herfried Münkler (2007).

Using military force against independent states is particularly problematic in outer zones where indirect power generally means less power – the Achilles’ heel of informal empire. Hegemony works best where states decide for themselves that following the hegemon’s lead is in their own best interests, or where economic carrots and sticks suffice to persuade them. But, for more recalcitrant states, imperial powers intervene with stronger measures, usually starting with economic sanctions, then threatening military force (recently dressed in the clothes of human rights or democracy), then actually using force with drones or conventional air strikes, and finally engaging in armed invasion as a last resort. Iraq was not an aberration because of invasion per se – after all, US forces have been bombing or invading other (mostly small) countries for decades. The problem was that invasion seemed a first resort rather than the last, and the prolonged occupation was an overambitious and failed attempt to prolong US hegemony (partly by controlling the oil supplies of rival powers; see Harvey 2003). That failure was unusually spectacular (as was Vietnam), but interventions against independent states usually bring substantial risks. In typically requiring interstate cooperation, they are vulnerable when that is withheld or is widely opposed, especially by leading imperialist rivals. Other vulnerabilities include client states acting in their own independent interests which may contradict those of the empire (some American analysts now see Israel as a wayward example).

US military power is unrivalled. Its spending in the military sphere exceeds the combined spending of the next seven or eight countries, accounts for about 40% of global arms expenditure, and is over five times greater than China’s. Its networks of surveillance and detention, and its military archipelago or “empire of bases” – territorial bases, mobile aircraft carriers, and nuclear submarines – is unparalleled. Yet even here there are challenges, and militarism brings its own problems.

Military power is the mailed fist in the velvet glove of foreign direct investments. Here the physical force is largely displaced to external network space, an extension of the displacement of armed force from the economic point of production to the political state, for the force is generally hidden outside the particular state that the United States deems is misbehaving. The United States
polices its own external interests, especially its FDI, and at the same time it inevitably protects the FDI of private capitalists from other imperial powers. This extended protection goes a long way to explain the appeal of US hegemony, but also the growing US complaints that other leading powers are not contributing enough to global security, for it is indeed an economic burden on a declining US economy (declining relative to its post-World War II zenith, and to the rise of China and of the European Union as rival economies of roughly similar size). While obviously greater in scale, the United States military archipelago is reminiscent of Britain’s global network of naval bases in the nineteenth century, but, whereas those were directly owned colonies of Britain, most of the United States territorial bases are within the sovereign territory of other independent states and are subject to formal contractual agreements which may circumscribe the purposes for which the bases can be used.

More problematically, military power is inappropriate or counterproductive for many hegemonic purposes, but there are nevertheless temptations to use it simply because the United States has an overwhelming lead in this sphere. Furthermore, to be credible, threats of force sometimes have to be carried out. If Arrighi’s (1994) historical analysis of past hegemonic transitions between the Dutch, British, and Americans is any guide to the future, the political turbulence in periods of hegemonic decline or transition was mainly due to belligerent attempts by the declining hegemon to prolong its dominance (and Bush’s Iraq adventure is easily seen as such). The United States might need to be saved from itself by its friends in the European Union showing some independence, rather than simply following along as junior partners.

In short, exercising and maintaining hegemony is more complicated in today’s new imperialism and in an unruly, multipolar world of independent states; and it is likely to become even more difficult. The Dutch and British arguably had an easier job policing the simpler world dominated by formal empires (which vitiates any British pretentions to give the Americans lessons in how the British did it). The relative economic decline of the United States, increasingly reliant on its dollar being the global reserve currency, has reduced its leverage on recalcitrant states. Even more challenging for the future, there is the global turbulence and recurring economic crisis tendencies of capitalism, and the looming threats of climate change, resource scarcity, and even resource wars. Believing that only “colonial possession” guaranteed control over raw materials, Lenin (1964/1916, 22, 260) warned that “The more capitalism is developed, the more strongly the shortage of raw materials is felt, the more intense the competition and the hunt for sources of raw materials throughout the whole world, the more desperate is the struggle for the acquisition of colonies.” Future territorial control may not literally take the form of colonies, but the likely logic could be very similar. The geopolitical and geoeconomic implications are unpredictable, but it is not too difficult to foresee more of the old imperialism reappearing.

Capitalism (and its partial separation of politics and economics) underpins the present nexus of a global economy, multiple national states, and the informal empires of the “new,” less territorial, imperialism. But territoriality, along with theft, force, and invasion, are still integral to the system. Imperialism now is a variable mix of old and new imperialisms.

SEE ALSO: Borders, boundaries, and borderlands; Colonialism, decolonization, and neocolonialism; Cross-border, transnational, and interregional cooperation; Ecological imperialism; Empire; Frontiers; Geopolitics;
Globalization; International division of labor; Local/global production systems; Marxist geography; Nationalism and geography; Nation-state; Newly industrializing economies (NIEs); Territory and territoriality; Trade, FDI, and industrial development; Transnationalism

References

Indexing

Changxiu Cheng
Beijing Normal University, China
Chinese Academy of Sciences

Indexing is the process of assigning and maintaining the organizational or storage structure of data to provide efficient data access. In most cases, indexing is a core database management system (DBMS) function that automatically updates and manages data indices.

Traditional indexing approaches (B+ trees, etc.) are often inadequate for spatial data handling because of distinctive characteristics of spatial data. Spatial data come in various forms, the most common being points, lines, and regions. In geographical information systems (GIS) applications, typical spatial queries are selection, join, and aggregation (Shekhar and Chawla 2003; Cheng, Song, and Zhou 2013). “Selection” queries include the spatial range query (“find all objects that lie within a given query region”) and the spatial point query (“find all objects that contain a given query point”). In addition, spatial relationship functions (e.g., overlap and disjoint) and operators (e.g., nearest neighbor) are essential to spatial data handling. Spatial relationships serve as the foundation for spatial join queries (e.g., “find all pairs of objects that intersect each other”) and nearest neighbor queries (e.g., “find the five objects nearest to a given query point”).

Several spatial data indexing approaches were developed over the years to accelerate spatial data access. In the early days, spatial data indexing focused on spatial ordering, such as the Z curve, Gray curve, and Hilbert curve. All of the indexing strategies map two-dimensional spatial data onto one-dimensional linear orders while preserving the relative locality of the data points. Once the data are sorted into one of the ordering schemes, spatial data can then utilize one-dimensional data indexing methods, such as B-trees, hash tables, and so on, to facilitate query and data access. Another multidimensional access example is the k-d tree (short for k-dimensional tree), which is a special type of binary space partitioning tree. Attempts to apply multidimensional indexing schemes proved challenging due to the lack of total ordering in a multidimensional space. Therefore, it was necessary to develop some specialized data structures for spatial data. Grid indexing, quad-tree indexing, and R-tree indexing are the three most common spatial indexing structures. These methods divide data space into several partitions and record the local relations between partitions and the minimum bounding rectangles (MBR) of spatial objects, while the actual spatial object is stored separately.

- **Grid index:** Grid-based spatial indexing registers objects to their position or positions in a grid then creates indices to relate object identifiers with grid cell identifiers for rapid access. Figure 1a shows an example of some features (F1–F3) and two levels of grid partitions: four quadrants of A, B, C, and D, and 16 cells (1–16). Figure 1b shows two grid index strategies corresponding to each of the two levels. Grid indexing is a space-driven or data-independent method; the procedures first create the index structure and then register data on an ongoing basis within the index structure. As a result, a common grid
indexing scheme can facilitate data merging from multiple sources or from individual data collections and indexing activities.

- **Quad-tree index**: The quad-tree was named by Finkel and Bentley (1974) and is now perhaps the most common indexing scheme for spatial data. Quad-tree indexing partitions a two-dimensional space by recursively subdividing it into four quadrants or regions. The regions may be squared, rectangular, or in arbitrary shapes. Figure 2a shows an example of some features, those features’
MBRs (F₁–F₁₉), and their recursive quad divisions. Figure 2b shows the corresponding quad-tree indices, in which each internal node has exactly four children quadrants: northeast, northwest, southeast, and southwest. In Figure 2b, a rectangle node symbolizes the MBR of a feature, whereas a circle node represents a quadrant divided by the quad-tree. Both the grid index and the quad-tree index are space partitioning indexing techniques; they divide the overall region using split operations. The difference is that the quad-tree index is a recursive quadrant division that is based on spatial distributions of data objects rather than seeking a height-balanced structure of the space partitions taken by the grid index approach. Thus the quad-tree indexing scheme is more efficient for spatially skewed data than the grid indexing scheme. In addition, the quad-tree index is a “data-driven” or data-dependent method. If the maximum capacity of a cell is reached as new data are coming, the cell should be further divided to register spatial objects at a finer resolution.

R-tree index: The R-tree was proposed by Guttman (1984). R-trees are commonly used to index multidimensional information, such as geographical coordinates, rectangles, or polygons. Central to the R-tree indexing scheme is the idea of grouping nearby objects and representing them with their MBR to form a node in the tree. Every nonleaf node in an R-tree contains a set of pointers to the corresponding child nodes and the MBRs of their child nodes. Furthermore, every leaf node contains a pointer to an actual spatial object and the object’s MBR. For example, the R-tree in Figure 3b has the “root” node that covers the maximal rectangle region in Figure 3a. The root node in Figure 3b has two child nodes that separately cover region I and II in Figure 3a, and leaf node 8 contains a pointer to a spatial object and the object’s MBR. Like quad-tree indexing, R-tree indexing is also a “data-driven” or data-dependent method. Being data-driven, quad-trees and R-trees are generally efficient for spatial data storage and spatial search. Many database management systems embed quad-tree or R-tree indexing schemes in

Figure 3 Example of R-tree indexing: (a) labeled rectangles in different levels and a dash-line query window; (b) R-tree index of (a).
INDEXING

their internal data structures to take full advantage of their efficient storage and searching capabilities. R*-tree, R+-tree, Hilbert R-tree, and X-tree are the different variants of R-tree indexing.

Data indexing helps sort candidates in response to spatial queries, especially for complex spatial queries. A spatial database management system (SDBMS) breaks down complex spatial queries into a series of basic spatial operations (such as spatial selection and spatial join) and executes these basic operations with efficient two-step query processing (Figure 4). First is the filter step that uses indices to locate possible candidate objects whose MBRs satisfy given query conditions. For example, a dash-line rectangle, as a window query, in Figure 3a will start with the entire extent (i.e., the root node of the tree). For each rectangle in a node, a decision needs to be made whether the rectangle intersects with the query window. If yes, the search process checks corresponding child nodes recursively until the search has traversed all intersecting nodes. The objects in leaf nodes, intersecting with the query window, will be put into a set of candidates, such as objects 4, 8, and 9. Then the refinement step performs exact comparisons among the candidate objects to select those that should be in the final query result. The filter step involves only an index search and is considered a low-cost step. In contrast, the refinement step is computationally expensive, but the computations are applied only to a smaller set of objects resolved from the filter step.

In recent years, spatial data have been collected from diverse sources and used as an ensemble to derive information referred to as big data (Dasgupta 2013). Volume, variety, and velocity are three characteristics of big data application. Big data are massive and can have a volume ranging from tera- to petabytes, even contiously increasing, and can come from multiple sources. Due to the large volume, big data repositories do not undergo rigorous preprocessing as commonly performed by operations via Structure Query Language (SQL) in a DBMS; instead, computational demands are shifted to data access and delivery. In addition to large volume and pronounced variety, big data come from multiple data streams at high velocity. An SQL DBMS quickly reaches its limits when handling the vast volume, velocity, and variety of big data. NoSQL databases play a very important role in spatial big data applications.

Geohash is an example of spatial indexing methods commonly adopted in NoSQL database systems. Geohash applies a hierarchical coding scheme and a Z ordering scheme for spatial indexing at multiple spatial resolutions. Geohash coding schemes map latitude and longitude coordinates into one single location index, namely geohash code. Furthermore, a geohash code comprises a hierarchical string that can be used to conveniently group objects at nearby locations. Hence, the geohash code serves as a key in a NoSQL database, such as Google Bigtable or MongoDB, to facilitate spatial search. For example, Bigtable, in a nutshell, is a NoSQL
massively parallel table indexed in a three-level B+ tree scheme. Bigtable employs a map (an associative array) that is indexed by a row key, a column family, and a timestamp. A common row key used by Google is reversed URL (such as com.google.ww) to facilitate webpage searches. Geohash codes can serve as a row key to support spatial search on web map applications. As semistructured or unstructured data indexing schemes adopted in NoSQL databases are widely used in geospatial semantic web applications, Geohash and other new spatial indexing schemes will continue to grow in popularity in spatial big data applications. Nevertheless, SQL database technology remains important to organize data and assure data integrity and quality in structured and long-transaction database applications (Dasgupta 2013).

SEE ALSO: Data structure: spatial data on the web; Spatial database; Topological relations

References


Further reading

India: National Association of Geographers, India (NAGI)

Founded: 1978
Location of headquarters: Delhi
Website: www.nagi.org.in
Membership: 2226 (as of January 20, 2015)
Secretary General: S.C. Rai
Contact: raisc1958@rediffmail.com

Description and purpose

The main objectives of the organization are to foster geographical research in the country; cooperate with professional organizations of cognate disciplines and to promote academic interaction within an interdisciplinary frame; cooperate with international organizations with similar objectives; improve the teaching of geography at all levels; foster cooperation between government and quasi-government organizations, research institutions and educational institutions for the advancement of geography; hold an Indian Geography Congress, ordinarily once a year; to take all such measures needed to secure geography's legitimate place in national life; and to strengthen among geographers a sense of national responsibility and professional efficiency.

Journals or major publication series

Annals of the National Association of Geographers, India. www.nagi.org.in/publications.html

Current activities or projects

The association runs various projects determined by a yearly schedule. The 36th Indian Geography Congress was held on November 17–19, 2014 in Jaipur, Rajasthan. The main theme was “urban systems, rural livelihoods security, and resource management”.

Brief history

Although the discipline of geography was institutionalized in India long before the establishment of the association, the organization came into existence only in 1978. Before that, individual universities and research institutes had their own geographical societies, but in 1978 an organization covering all of India was founded by eminent geographers to make the discipline more research oriented and interdisciplinary. The association played a significant role from its beginning in bringing young geographers and scholars together into a single organization and gave an opportunity to interact and discuss the emerging issues in geography. It has a long history of productive cooperation with the leading educational institutes in the country and runs a variety of projects of national importance.

Submitted by Jan Monk
Indigeneity

Sarah Hunt
University of British Columbia, Canada

Indigeneity and indigenous geographies are relatively recent areas of geographic scholarship. Yet, beneath and before the conceptualization of nation-states and continents within Western European thought, indigeneity provided the originary modes of being in relation to diverse territories across the globe. Indigeneity – that is, the original peoples and their ways of being, knowing, and acting in relation to the territories on which they lived – existed prior to the formation of the discipline of geography and its subdisciplinary modes of inquiry. In settler-colonial countries such as Canada, the United States, New Zealand, and Australia, place-based modes of thinking and knowing began within the context of indigenous epistemologies before being represented within the colonial geographic imaginary. It is from this perspective of indigenous peoples’ diverse lived experiences, both precolonial and neocolonial, that this entry on indigeneity in the field of geography arises.

Indigeneity on its own terms

Conceptualizations of indigeneity are shifting and contested, and emerge within specific sociopolitical and historic contexts. The diversity of indigenous political movements, cultural practices, and critical scholarship cautions against representing distinct peoples as essentialized versions of indigeneity. In a literal sense, indigeneity is taken to mean anyone who is born in a place; thus anyone who has not migrated to that place may be regarded as indigenous. Yet, within the political context of ongoing neocolonial relations in places like North America, Australia, and New Zealand, indigeneity has a specific meaning intended to distinguish between colonized groups and colonizers or settlers. Thus, in this context, indigenous specifically means peoples with ancestral and spiritual ties to particular territories, whose ancestors lived in relation to that land prior to colonization and settlement by others. In colonial and neocolonial contexts, indigenous is also taken to mean nations that, in large part, remain under the power of the settler or colonizing governments.

To be precise in identifying indigenous peoples, the preference is generally to use indigenous peoples’ own names for themselves, in their own languages. Indeed, the process of identifying, categorizing, and naming indigenous peoples is one that has been inherent in colonialism the world over, and is thus also a key site of decolonial efforts. In the United States, indigenous people were named “American Indians” within national legal frameworks, which have shifted in recent years to the terminology of “Native Americans” and “Amerindians.” In the Canadian context, indigenous peoples have been legislatively categorized as “Indians” or as “Aboriginals,” with the limitations and definitions of this categorization determined by the state rather than by terms set by indigenous peoples themselves. This terminology has, over time, shifted to include “First Nations,” “Métis,” and “Inuit” categories, each with its own state-delineated definitions, rights, and status.
has been argued that the legal categorization of American Indians and Indian, Métis, and Inuit peoples has been designed within frameworks of extermination aimed at producing fewer and fewer natives over time. While legal definitions and formulations designed to count indigenous peoples have changed over the years, many state formulas continue to deny recognition of indigenous status to large numbers of people claiming indigenous heritage, particularly women. This gendered inequity of indigenous status is due to the patrilineal lines of descent written into current or past government policies. Indigenous people thus continue actively to resist and contest such state-defined formulations of indigeneity while asserting their own self-determined individual and collective identities.

Although definitions of indigeneity differ greatly across historic and current global sociopolitical geographies, indigenous peoples have identified some shared cultural markers. Indigeneity is generally rooted in living in a reciprocal relationship with the places that comprise indigenous peoples’ territories. Indigenous worldviews often recognize the active agency of the land and nonhuman actors, such as animals and spirits, and regard the relationships between humans and their environment as ones of reciprocity. Although the ability to honor these relationships has been deeply impacted by colonization, indigeneity is recognized as being alive in the bodies, teachings, and knowledge of indigenous peoples, despite the constraints placed on their ability to enact them as they did prior to colonialism. Displacement from their territories has also impacted indigenous peoples’ ongoing relationships with their lands. Numerous indigenous knowledge-keepers and scholars have described their orientation to these territories as mobile, whereby, although they may travel away from their lands by choice or by force, their orientation to these places remains active. While some indigenous peoples link indigeneity with an individual’s blood quantum of indigenous ancestry, others define indigeneity according to an individual’s orientation to indigenous principles, teachings, and ways of life. The ability of such markers of indigeneity to change over time is a point of difference between indigenous groups, yet there is increasing recognition that indigenous practices, beliefs, and traditions have always been adaptive and fluid rather than static. Indeed, Western scholarship has often portrayed indigeneity as static in order to justify imagining indigenous peoples as vanishing, as development, displacement, and conquest have impinged on their precontact ways of life. It is around the shared markers of indigeneity, as well as global struggles for indigenous rights and recognition, that the vitality of indigenous sociopolitical geographies has emerged in recent years.

While indigenous peoples may share certain commonalities, diversity and heterogeneity are a central aspect of indigeneity the world over. Indigenous peoples are necessarily different from one another, as indigenous political, cultural, and social practices and worldviews are lived in deep relationship to place. Despite the aforementioned state-determined processes for grouping indigenous peoples together under certain legal categories (e.g., “Indians,” “Métis,” “Inuit”) with attendant rights and status, indigenous people insist on the distinctness of each indigenous group, while acknowledging long-standing relationships across and between indigenous peoples in areas where territories are shared.

The long-standing place-based relationships of indigenous peoples to their territories are evident in place names and stories told about the land. Indigenous place names also mark reciprocal responsibilities to territories. Names, stories, and
meanings relating to place emerge within relationships with traditionally used lands, with many indigenous peoples seeing the land itself as integrally involved in meaning-making rather than humans as mapping meaning on top of the land. Processes of renaming previously held indigenous territories have been central to asserting settler presence while simultaneously erasing indigenous claims. Yet indigenous peoples continue to assert territorial naming rights. The stories and names that have been connected to specific places contain knowledge that is vital to their ecological wellbeing – something that has more recently been acknowledged within sustainability science and other environmental management fields.

Within many indigenous epistemologies, as mentioned above, animals, landscapes, oceans, and nonhuman elements are imbued with their own spirit and are also understood as having agency. Within long-standing relationships with such nonhuman elements, indigenous peoples often assert a responsibility to be in active communication with these elements of their cultural landscape. These understandings emerge within deeply relational and collectively formed place-based epistemologies, in which people, animals, and the natural world have a responsibility to one another extending back through millennia.

While the inclusion of indigeneity within the field of geography will be explored later in the entry, it should be noted that many indigenous scholars and scholars working with indigenous people internationally have used place-based analysis outside the discipline of geography and outside formal scholarly disciplines altogether. Indeed, indigenous peoples have long contended that indigeneity does not require formalization or recognition by scholars in order to gain legitimacy, but that indigenous ways of knowing and being are valid on their own terms. The reframing of indigeneity within various disciplinary frameworks has indeed been central to colonialism, and reclaiming indigenous control over representations of indigeneity is an important aspect of decolonization.

Some geographers working in the area of indigeneity have cautioned against creating a dichotomy between indigenous and Western, as there are many distinct cultural worldviews within each of these realms. Western epistemologies are broadly understood to mean modern systems of thought based in the Cartesian split between discourse and materiality, mind and body. This split is generally not found in indigenous worldviews, yet it has historically been used by many scholarly disciplines to represent indigenous epistemologies, reimagining them in this Western framework. For example, indigenous worldviews may understand meaning to be immanent within land, territory, and nature themselves; yet, when represented by Western academics, indigenous meanings are understood to be mapped onto the landscape rather than emerging from it. Indigenous peoples live in a great variety of urban and nonurban settings, negotiating hybrid identities, sometimes described as “thirstspace,” “fourth world,” or “dancing between worlds.” This hybridity of Western and indigenous epistemologies is arguably congruent with the fluidity of indigenous cultures, which were never static but changed and adapted over time, despite their representation within colonial discourses as static, backward, and temporally bound to premodern times. Thus, although indigenous and Western geographies are largely talked about as distinct schools of thought emerging within neocolonial power relations, this call for hybrid and nonessentialized indigenous geographic knowledge should be kept in mind.
Representations of indigeneity within geography

It has been argued that it is only since the mid- to late 1960s that the subdiscipline of indigenous geographies could be said to have existed. Indigenous geographies grew largely within critical and radical geographic scholarship, as geographers began to rethink commonly held views of indigenous peoples. Though geography’s engagement with indigeneity has been limited, the impact of Western geographic representations of indigenous peoples and their territories has been felt widely among indigenous communities the world over, as those representations were central to colonial processes and power relations.

In recent years, geographic representations of indigeneity and of traditionally held indigenous territories have themselves become subjects of critique as part of broad efforts to reflect critically on the discipline’s role in historic and ongoing colonialism. Historical approaches to indigeneity have been examined by critical geographers in order to denaturalize the racialized portrayal of indigenous people and knowledge as less evolved or as rooted in antiquated notions of “tradition” (see Cameron, de Leeuw, and Greenwood 2009 for a historical overview). Some of the more critical engagements with these representations have pointed out the inherently violent function and devastating outcomes of these representations, as well as their role in erasing indigenous peoples’ presence altogether, in some cases. Critical geographers in Canada and the United States have demonstrated that the role of geography and geographic discourses and practices in empire building must be acknowledged in order to move the discipline beyond its conventions.

Imagining empty lands

The Western geolegal concept of *terra nullius* – or the empty lands doctrine – is at the heart of colonial power relations. In order to imagine the lands of North America, and other parts of the world, as available for settlement, they needed to first be cleared of any legal “owners.” Racialized discourses about “Indians,” which categorized indigenous peoples as less than human, were embedded in Western legal discourse beginning with the doctrine of discovery, which served to neutralize the settlement of North America, as indigenous people could not be recognized as having legitimate territorial claims.

Further, the Western geographic imaginary imposed a Cartesian worldview on colonized lands, through which the lands could be imagined as blank spaces on which to map new areas of progress, extraction, and settlement. Combined with the authoritative legal power of the newly formed states, geographic tools such as grids, maps, and surveys effectively rendered invisible indigenous territorial relations and jurisdictions. Geographers have critically evaluated the role of these spatial tactics in naturalizing indigenous deterritorialization, arguing that they have worked to render the epistemic and material violence of settlement invisible.

One of the key outcomes of indigenous spatial erasure was the formation of Indian reservations or reserves. As geographers such as Cole Harris have shown, Indian reservations in Canada and the United States can be understood as a manifestation of colonial ideologies, produced in order to realize the ongoing dispossession of indigenous peoples of their territories, and to establish the sovereignty of the Canadian state. The creation of colonies entailed not only the mental work of imagining “Indians”
as without legal tenure, but also the practical work of individuals mapping out and setting aside lands to which indigenous people and their claims could be confined. Through this process, indigenous people became trespassers in their own territories whenever they traveled beyond these newly created borders. Although the reservation geography of settler-colonial countries is often taken as fixed or natural spaces of indigenous occupation, geographers, particularly those working in areas of colonial law and policy, have demonstrated the fraught nature of their construction and realization.

This geographic imaginary is not only historical, however, as indigenous peoples continue to contest the division of land within settler-colonial worldviews, thus undermining the naturalized character of terra nullius imaginaries. Reserve geographies are produced within the rationale of terra nullius, and in relation to the imagined frontier, which are all at work under the spatialization of contemporary sociolegal relations. The geographic examination of imagined and material colonial geographies is useful in understanding contemporary indigenous–state relations, as we can see how reserves continue to be produced as spaces of assumed neglect and violence, while “Indians,” inherently “of the reserve,” are produced as subjects of this naturalized marginalization. Representations of indigeneity today thus continue to be conditioned through neocolonial power relations, but they are also produced within the resistance and vitality of underlying indigenous territorialities.

While the mapping and dividing up of land, and the designation of certain spaces for natives, has clearly been a source of marginalization and oppression for indigenous peoples, it has also been a site of resistance. In the 1990s indigenous geographers and geographers working collaboratively with indigenous peoples began using counter-mapping to bring new form to indigenous knowledge using the technology of geographic information science (GIScience). Geographers have also studied the disruptive potential of blockades and other strategies of making indigenous jurisdiction visible, particularly in places such as British Columbia, Canada, where the legal status of lands is highly contentious. In these and other ways, geographers have contributed to making underlying indigenous geographies visible, responding to calls to decolonize the discipline of geography and, more broadly, to decolonize dominant settler–indigenous relations.

**Urban indigenous geographies**

In the processes of settlement described earlier, cities emerged in settler-colonial contexts as set apart from areas legally recognized as belonging to indigenous peoples. Sites of indigeneity were imagined as either temporally or physically at a distance from sites of civilization and progress. We see this imaginary at work when indigenous people are said to be “migrating” to cities from reservations. This representation naturalizes reserves as “Indian spaces” while reaffirming the whiteness of urban spaces and rendering underlying indigenous territories invisible. Urbanization has been facilitated by the material displacement of indigenous peoples from lands that were desired by settlers and, in the process, the prior relationships to those lands were frequently forgotten or erased. Indigenous people, in the meantime, have been imagined as belonging to “Indian spaces” such as rural areas, Indian reservations, and, in the context of urban centers, racial ghettos.

Since the mid-1990s, geographers have contributed to understanding how urban areas have been constructed through racial and gendered discourse and policies such that indigenous
peoples are “out of place” in these white settler spaces. Attention has been given to the ways in which certain racialized bodies, such as those of indigenous women, have had violence against them normalized both through social norms that render them as apart from the rest of society and through legal processes that are inherently colonial. In decolonizing these dominant relations, indigenous geographers have created accounts of indigenous women’s identities and resistance that unsettle dominant conceptions of race and gender in urban spaces. The memorialization of certain kinds of violence and certain kinds of bodies within urban centers has also emerged as an area of study among urban geographers, helping to contextualize the unwillingness to pay tribute to indigenous women who have faced widespread violence in urban spaces. In more recent processes of gentrification within poor and industrial areas of North American cities, geographers have demonstrated how frontier imaginaries are being mobilized to render those spaces as empty and thus available for development.

Another important theme within urban indigenous geographies is the reclamation of indigeneity, which is active and alive underneath urban geographies in settler-colonial countries. Indigenous people are reasserting indigenous place names within cities, using billboards, street signs, and other representational tools to make their long-standing presence known. The mobilization of urban indigenous people for various social justice causes, such as the disappearance of hundreds of indigenous girls and women, has also been a venue through which indigenous presence in cities has been asserted. These urban indigenous geographies remain an emergent area of study for geographers interested in decolonizing the discipline.

**Strengthening the presence of indigeneity in geography**

Since its emergence in the 1960s, indigenous geography has clearly been an area of innovation and critical research and of theoretical development within the discipline. Yet, more broadly within geography, indigenous issues and people continue to be identified as a gap in the field. The discipline’s historic lack of engagement with indigeneity is evident in the absence of indigenous knowledge, people, and ways of being in most geography textbooks and course offerings. This absence has been said to suggest that the dominant geographic imagination is not generally organized by a recognition of the indigenous title and territories underlying settler-colonial geographies. Perhaps most tellingly, it is only recently that the discipline has seen a growth in the presence of geographers of indigenous heritage.

Scholars in diverse areas of geography have pointed out that Aboriginal, Indigenous, First Nations, and Native American people and knowledge constitute significant gaps in the field of geographic research. This lack is made apparent in the evaluation that in the years 1997 to 2004 publications with the term “indigenous/indigeneity” represented less than 2% and “aboriginal/aboriginality/aborigines” represented 3.5% of international geography journal content (Shaw, Herman, and Dobbs 2006). Indigenous geographies have remained peripheral to broad geographic theory, as “indigenous geographies are somewhat removed from the rigors of disciplinary debate and remain out there, on the post-neocolonial edges of the disciplinary orbit” (Shaw, Herman, and Dobbs 2006, 269; emphasis original). Scholars have thus called for members of indigenous community movements to be more fully engaged as geographers, so that their expertise can be sought and learned
 Others have noted that creative effort is needed to improve First Nations and Native American studies in geography, as not only are indigenous people underrepresented within academic geography departments, but there is comparatively little work by geographers on the political struggles of indigenous groups in North America and internationally.

However, since the mid-1990s there has been a growth of spaces dedicated to indigeneity within the discipline, due in large part to the innovations of indigenous scholars working in geography, as well as their allies. These include the formation of the Indigenous Peoples Specialty Group of the Association of American Geographers (AAG) and the Indigenous Peoples Working Group of the Canadian Association of Geographers (CAG), as well as work to institutionalize indigenous issues within the International Geographical Union (IGU). International professional associations like the Institute of Australian Geographers have also seen the formation of such spaces dedicated to indigenous issues. These groups have developed funding support to assist in bringing indigenous students and community members to regional and national meetings, to encourage the growth of their voices within the discipline. The growth of these specialty groups has facilitated the development of sponsored conference sessions and keynote lectures by indigenous political leaders, experts, and knowledge keepers at annual meetings. Additionally, special issues of international geography journals, including Cultural Geographies and The Canadian Geographer, have been dedicated to themes of indigeneity, and Geografiska Annaler B: Progress in Human Geography has also recently included progress reports on indigenous geographies, indicating the increasing impact of this area broadly within the field.

In response to this growing interest in indigeneity, along with a recognition of the exploitative research practices that have typified scholarly engagements with indigenous peoples historically, power dynamics within knowledge production and legitimation have become a key area of critique. The foundations of knowledge production have been destabilized through the increasingly vocal mobilization of indigenous peoples’ self-determination struggles, which have included an insistence on developing their own methodologies. The reassertion of indigenous knowledge paradigms through indigenous-led research is integral to dismantling colonial power relations: “It has involved a ‘knowing of the colonizer’ and a recovery of ourselves, an analysis of colonialism, and a struggle for self-determination” (Smith 1999, 7). While differing widely, indigenous methodologies generally draw on culturally specific knowledge creation and mobilization principles, while taking colonial power relations into account. In a recent move away from its colonizing project, critical geography has seen renewed engagement with issues of indigeneity through careful, inclusive, and collaborative research initiatives. In these processes, indigenous people have sought control over representations of themselves and their knowledge, and have asserted that recognition of indigeneity is not required by formal disciplinary standards but is valid on its own terms.

Such assertions of self-determined representation have challenged geography, as well as other disciplines, to recognize methodologies that account for diverse expressions of indigenous knowledges, including those that are not written down but are presented orally. These methods continue to challenge the standards of many scholars across diverse fields of study. Geographers have been compelled to ask how research serves indigenous community concerns and how it might support, or detract from, self-determination at a personal and community level. Engaging in deeply relational,
place-based forms of indigenous inquiry, scholars are challenged to balance their accountability to institutions and communities which are culturally and epistemologically divergent (Kovach 2009). Ethical concerns about the long-term impact of research on indigenous peoples serve as a reminder that scholars cannot always know how knowledge will be used once it is made public. For example, the mapping of traditional uses of indigenous territories may benefit their legal land claims but risks being misused by people or industries that want to exploit that knowledge. Collaborative working relationships and obtaining prior informed consent for conducting research are thus key to allowing indigenous peoples to be actively involved in shaping research about them.

Alongside these developments in indigenous methodologies, some geographers have questioned the potentially harmful effects of achieving inclusivity of indigenous people and knowledge. For example, while community-based research (CBR) is often lauded as inherently decolonial, critical geographers have been careful to point out how the language of CBR can gloss over power dynamics and replicate some of the very colonial relations researchers are trying to avoid. Thus, geographers must remain attuned to the ways their work is shaped by institutional norms, including the push to publish solo authored papers, for example, and to adhere to institutional research ethics which may not match up with indigenous concerns. Additionally, institutions are still challenged to recognize CBR as legitimate scholarship, as so much of the relationship-building that goes into truly collaborative work with indigenous peoples cannot be captured by neoliberal methods of measuring research outputs. Decolonizing the discipline is thus a much larger project which must extend beyond geography associations and journals to the culture of the institutions in which geographers are situated.

One area of indigenous leadership over research has been in the development of decolonial principles to guide research. Indigenous peoples have developed the framework of ownership, control, access, and possession (OCAP) in shaping a self-determined agenda for scholarly research involving indigenous knowledge and communities. These principles are designed to allow indigenous people to shift the power over control of their knowledge into their own hands, with academic partners, researchers, and students conducting their work in response to these self-determined approaches. While institutional ethics protocols and legal ownership procedures are shifting slowly, challenges still exist as to how to ensure that OCAP principles do not conflict with institutional norms. Increasingly, indigenous peoples are initiating their own research agendas, writing their own stories, and capturing their own knowledge with geographers and other specialists as partners or collaborators, again ensuring that indigenous epistemologies are at the center of geographic knowledge production.

**Future directions**

As the presence of indigenous people as scholars and collaborators increases in the field of geography, the discipline’s engagement with indigeneity is bound to shift significantly. In particular, there has been a movement away from viewing indigenous peoples as victims of neocolonial relations, with increased attention on indigenous political resurgence, success, and survival. This is broadly true within academia, as indigenous scholarship has seen extensive growth since the 1990s, with the emergence of indigenous studies departments, indigenous-focused courses, indigenous research streams, and
departmental recruitment of scholars working specifically in areas related to indigeneity.

Calls have been made for geographers to acknowledge that some indigenous concepts can only be understood within indigenous spatiotemporal ontologies, which are distinct and locally situated. These calls, which have emerged within the recent “ontological turn” within the discipline (Cameron, de Leeuw, and Desbiens 2014), will undoubtedly continue to fuel discussion about how indigeneity is represented and mobilized within institutional contexts that remain fundamentally rooted in colonial power relations. Geography is challenged to allow for shifts in ontologies of space, place, temporality, and landscape, among other foundational geographic concepts, as indigenous epistemologies push at the limitations of the English language and challenge the terms in which geographic knowledge has developed from its Western roots. Critical geographers have examined the potential for encounters with indigenous and other non-European geographies in exposing the cultural specificity of modern geographic knowledge and providing challenges to geography’s reliance on Cartesian ontologies. At the same time, indigenous peoples have also called for geographers and other scholars to continue to be mindful of the capacity for academic appropriation and misrepresentation of indigeneity in the pursuit of unsettling disciplinary norms.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Indigenous health; Indigenous knowledge; Indigenous people and regional development; Indigenous technical knowledge

References


Further reading


Indigenous health

Chantelle A.M. Richmond
Western University, Canada

Globally, there are approximately 370 million indigenous people (or 5% of the world’s population) living in 70 countries including the Circumpolar North; South, Central, and North America; Africa; Asia; the Pacific Islands; New Zealand; and Australia. The term “indigenous” is used to refer to the descendants of original populations who lived in an area before colonization. There are a number of characteristics that are generally included in the meaning of the term “indigenous,” including having a common ancestry, being accepted by one’s community, possessing a cultural distinctiveness from the majority population, and a common experience of dispossession from traditional lands and territories.

The term “indigenous” is generally used in a global sense, whereas indigenous groups and communities typically have geographically contextualized terms that relate to the places they inhabit, as well as the languages they speak. For example, in central Canada, the Ojibway people are often referred to as the Anishinabe, a term that reflects the language spoken by this group, Anishinabemowen. Around the globe, there are various other terms used to reflect one’s indigeneity, including “native,” “aboriginal,” “tribal peoples,” “original,” and “first.”

Indigenous people tend to speak languages and to practice traditions that are distinct from the dominant society, reflecting rich and diverse cultural knowledges, spiritual practices, values, traditions, customs, food preferences and practices, forms of livelihood, trade and economy, and the necessities of everyday living. There invariably is a significant and interdependent relationship between indigenous peoples and their local ecosystems, one that is central to the health of the people and vital for a way of life based fundamentally on the resources of their traditional environments. Prior to colonization, indigenous societies could be described as subsistence cultures, meaning that their daily nourishment was provided by the physical resources of the air, land, and water. This total reliance has fostered a deep respect on the part of indigenous people for the natural environment, but for many this relationship has also led to a deeper cognitive and spiritual relatedness wherein the environment is seen as conceptually interrelated with humankind. Here lies an important and often overlooked interrelatedness: the health of the environment is inseparable from the health of those whose existence has so emphatically relied on it. This relationship has been fostered, practiced, and celebrated in many ways by indigenous societies around the world. In contemporary times, many indigenous groups continue to live a way of life sustained by resources of the lands, air, and waters (e.g., harvesting and consuming “traditional foods” or “country foods,” as they are often referred to), but processes of environmental dispossession are changing the ways in which indigenous people can safely and reliably sustain themselves fully from a subsistence diet (Ford 2012).
Indigenous knowledge systems: meaning for health and ties to the land

In all indigenous cultures, daily life and continued wellbeing have historically depended on a deep knowledge and intimate relatedness to one’s local ecosystem, which can be referred to as an indigenous knowledge (IK) system. At the most basic level, it is this knowledge system that has enabled indigenous people not only to survive but to thrive, in some cases for millennia, in some of the harshest, most inhospitable environments on Earth, such as the Australian Outback, with its low annual rainfall, or the Canadian Arctic, where some of the Earth’s coldest temperatures have been recorded.

While there are no commonly accepted definitions of what an indigenous knowledge system is, it generally refers to the culturally and spiritually based ways indigenous people relate to their local ecosystems and to one another (LaDuke 1994). These relationships are centered on a number of key characteristics. First, an IK system is based on localized knowledge and understandings to utilize the resources of one’s ecosystem sustainably for medicines, food, shelter, and many other purposes related to everyday life. Second, it is a dynamic, applied knowledge. From season to season, for example, the patterns of flora and fauna (e.g., the migration and hibernation patterns of local birds and mammals), as well as weather patterns, vary. In order to learn how best to utilize these resources, indigenous people draw on their strength of observation of these constantly changing patterns. Third, an IK system is an integrated knowledge that consists of many parts and is thus best learned by doing or practicing. Because the interaction with the local ecosystem contains so many biological and physical processes, and also because most IK systems are maintained within an oral tradition, learning is best done by practicing, being out in the environment, and through firsthand observation.

Indigenous concepts of health

One of the consistencies in indigenous worldviews across the globe is the strikingly similar conceptions of health and healing, which may nevertheless be represented in diverse ways. Among Canada’s First Nations population, the medicine wheel embodies a common understanding of health: a circle divided into four sections representing the four elements of life – the physical, emotional, mental, and spiritual. If one element of the wheel is neglected or receives too much attention, health suffers in all four areas. Among New Zealand Māori, “health as a house” is one of many conceptual frameworks used to illustrate the multidimensionality of Māori health and wellbeing. Like the medicine wheel, this conception is represented by the four walls of a house: each wall is a different health dimension, and if one wall should fall the whole house (one’s health) will collapse. Among Aboriginal Australians, the three dimensions of healthy living are thought to be land, spirituality, and people; these three interrelated dimensions provide a blueprint for living.

Another key dimension of the IK system is its relational context – meaning that individual wellbeing reflects social and moral rules – and the inherent obligations one has not only to oneself, but also to the wider community and ecosystem. Inuit Qaujimajatuqangit (IQ) is the term used to describe Inuit epistemology or the indigenous knowledge among Inuit in Nunavut. This place-based worldview explicitly recognizes humans’ connection to the land, to the greater ecosystem, and to one another, and the term translates directly as “that which Inuit have always known to be true.” At the individual
level, a good or moral person is someone who is *inummarik*, or “a most genuine person.” From the earliest times, the sharing of food and resources was a central concept in Inuit life and social structure. The ways in which an individual’s wellbeing is bound by social values and his or her connection to wider community are also illustrated in the Native Hawai’ian context. Among Native Hawai’ians, a good or healthy person is thought to have the following values: *lokahi* (living in balance), *pono* (doing the right thing), and *kokua* (working without expecting reward). Here, the ultimate goal is not constrained to individual wellness, but rather seeks to strengthen and protect family and the wider community.

IK systems are not only based on an individual’s connection to physical and social dimensions, but are also fundamentally supported by a spiritual dimension. One of the most challenging parts of conceptualizing the IK system results from the fact that it is bound within a holistic philosophy that differs quite significantly from a biomedical, industrial, or capitalist philosophy that tends to frame Western notions of life, learning, or, for example, what it means to be healthy (Berkes 2012). A holistic philosophy is rooted in the understanding that life, learning, and wellbeing are part of a larger system that cannot be understood as separable, or as components, but rather as fluid, connected pieces. The complexity of this philosophy is embedded in many indigenous languages, including Anishinabemowen, which does not differentiate between male and female. Rather, in Anishinabemowen philosophy, all parts of our world are differentiated into animate and inanimate objects – meaning those things that have spirit or life and those that do not, respectively. The distinction between these categories is not straightforward in a way that sees objects in the natural world as having spirit or life (e.g., a tree) or not having spirit or life (e.g., a rock). Rather, the distinctions are deeply rooted in Anishinabe spiritual beliefs, including a relatedness that generally does not differentiate between the physical and spiritual worlds. These conceptual ideas stem from Anishinabe ideas of creation, many of which persist only in the oral tradition, and conceptualize human life as deeply connected to various parts of the physical world, including nonliving objects. Several Anishinabe artists – most prominently Norval Morrisseau – have used their craft to depict this complexity and the interrelated nature of the physical, biological, and spiritual worlds.

Probably the most important dimension of the IK system is that it is shared almost exclusively through an oral tradition. To learn and to be part of the IK system means practicing and living the knowledge. In practical terms, accumulating IK is based fundamentally on relating to the environment – through being out on the land and sea – and also through access to those who can teach the knowledge. In many indigenous cultures, there are varying levels of access to IK. Some are rooted strongly in daily life, in the basic struggle for survival, in the maintenance of cultural identity, and in interaction with the local ecosystem, but there is also applied or spiritual knowledge that may be accessible only through ceremony or role obligation, for example puberty rights or the practice of traditional medicines and healing.

### Environmental dispossession and indigenous knowledge systems

Given the special connection between indigenous peoples and their ecosystems, including the social, spiritual, and economic ties it supports, indigenous peoples are more vulnerable to the processes of environmental change and
environmental dispossession (Richmond and Ross 2009; Ford 2012), and with strikingly similar consequences for health and wellbeing. In the greater global context, IK systems are based fundamentally on local experiences and a connection with and ongoing observation of the local ecosystem. Because of the vast and often rapid changes brought on by global industrial development, including various types of contamination and climate change, indigenous peoples are particularly susceptible to these changes. In fact, environmental dispossession has become a common unifying experience among indigenous peoples, and could be framed as a second wave of colonization. Environmental dispossession refers to the processes by which indigenous peoples’ access to their traditional lands and territories is reduced or eliminated (Richmond and Ross 2009). These processes take various physical and political forms, were generated historically, mainly through processes of colonialism, and continue in contemporary times. The outcomes of these processes vary across place and culture.

Direct forms of environmental dispossession involve processes that physically disable use of the land, such as industrial activities or contamination events that may sever ties to traditional foods or resources required for sustaining daily activities. The effects of climate change are particularly prominent among indigenous peoples and populations (Ford 2012), many of whom inhabit the most fragile ecosystems globally. Indirect forms of dispossession occur as a result of policies or regulation that lead to the severance of indigenous peoples’ link to the land, and the indigenous knowledge it fosters. For example, indigenous peoples have been subject to massive relocation projects, as well as various assimilationist policies that have included the forced attendance of indigenous children at church- and state-run residential/boarding schools in Canada, the United States, and Australia. These processes have undermined the ability of indigenous peoples to practice their land-based cultures and IK systems. In contemporary times the processes of environmental dispossession have significantly disrupted indigenous peoples’ relationships to land, and thus to the cultural knowledge and sociocultural practices that are central to their maintenance of health and wellbeing (Big-Canoe and Richmond 2014).

Patterns of indigenous health

On a global level, the processes of environmental dispossession – whether operating in direct or indirect forms – have had significant impacts on the health and wellbeing of indigenous people because they work to destabilize IK systems. They do so by reducing indigenous people’s confidence in environmental conditions, for example, as climate change makes it more difficult to predict the weather (e.g., with extended dry or wet periods), or as migratory animals return later in the season than usual. While IK systems are dynamic, fluid, and constantly changing in relation to the natural rhythms of local ecosystems, the processes of environmental dispossession can accelerate these changes in a way that makes normal interaction with the land risky for human health (Ford 2012), for example, when local fish stocks or other country food has been contaminated by heavy metals. These changing conditions can lead to people spending less time on the land because the risk of doing so can become too great. Less time spent out on the land means that opportunities for sharing IK and for practicing skills may also become limited over time, leading to the weakening of cultural and social ties that fundamentally support indigenous cultural identity, social roles, moral obligations,
and self-esteem (Big-Canoe and Richmond 2014).

On a global scale, the consequences of environmental dispossession for indigenous health are remarkably similar. The reduced access of indigenous peoples to their traditional lands and territories has had many detrimental impacts for the preservation, practice, and intergenerational transmission of IK, leading to significant changes in their way of life and resulting in a contemporary health and cultural profile characterized by a greater burden of morbidity and early mortality than that of the nonindigenous population. While the last half century has witnessed significant gains in life expectancy for indigenous populations, and a considerable reduction in infant mortality, several other patterns of health are troubling, including a high prevalence of cardiovascular disease, type 2 diabetes, and cancers; mental illness, addictions, and violence; and food insecurity. While most indigenous populations have undergone the epidemiologic transition, where the main causes of indigenous morbidity and mortality have shifted from infectious to long-term chronic disease, infectious disease is common in the health profile of indigenous populations. The persistence of infectious disease among indigenous peoples includes many ailments that have been all but eradicated in the general population, such as active tuberculosis and rheumatic fever. International research indicates that these health disparities are linked to alienation from the land (Richmond and Ross 2009), to limited opportunities to practice activities that foster indigenous cultural and social connectedness, and traditional health and healing, and to a number of social determinants of health, including poverty.

Another significant outcome of environmental dispossession is an unprecedented shift to cities by indigenous populations. Over the course of the twentieth century, indigenous people in Canada, the United States, Australia, and New Zealand have moved, in considerable numbers, into urban areas and away from their ancestral lands. This pattern of urbanization has led to significant shifts in the cultural and social identities of indigenous people, but it is important to note that indigenous cultural identities and the urban environment are not mutually exclusive. Rather, the small base of literature on indigenous urbanization illustrates that urban indigenous people tend to migrate rather fluidly between the city and their traditional homelands. In the city, indigenous identity is sustained and exercised through participation in various social or political organizations, among other institutions such as university and college campuses where indigenous services and programming have been created to nurture indigenous students in the academic setting. Despite the growing trend of urbanization among indigenous populations worldwide, very little is known about the drivers and consequences of this migration; this is a decidedly underresearched topic in the geographical literature. An important recent text by Peters and Anderson (2013) details international indigenous experiences of urbanization and the ways in which indigenous cultural identities are constructed and reconstructed in cities.

Researching indigenous health and environmental repossession

In spite of the troubling patterns of health evident in indigenous populations, there are reasons to be hopeful. As noted, the wide gap in the health and social profile of indigenous people is slowly closing over time. There are many factors leading to the reduction in health disparities, including increased access to culturally safe and high-quality health care; the development of health and social policy that reflects
INDIGENOUS HEALTH

indigenous ideals and philosophies; the ongoing resolution of indigenous land claims; increased funding for indigenous social development and programming; and steady increases in socio-economic wellbeing (e.g., increases in high school completion and postsecondary attendance).

Perhaps one of the most important contributions to the closing gap between indigenous health status and that of the general population relates to a growing and hopeful movement of indigenous community participation in health and social research, that is, by indigenizing the way research is done. Building on Linda Smith’s seminal work on decolonizing methodologies, a wide body of indigenous research has been undertaken, led by a common goal of “centering our own concerns and world-views, and coming to know and understand theory and research from our own perspectives and for our own purposes” (1999, 39). Such methods require meaningful research partnerships between indigenous communities and researchers from universities and various governments and organizations, with the ultimate goal being to design and carry out research that will lead to improved local conditions. These partnerships are a method of decolonization, as the indigenous communities involved are placed at the center of the research, thereby empowering local people to influence research topics, debate, and research design, and prioritizing local influence in the development of community programs. Based on research with Anishinabe youth in Ontario, Big-Canoe and Richmond (2014) recently proposed the term “environmental repossession” to refer to the social, cultural, and political processes by which indigenous peoples and communities are reclaiming their traditional lands and ways of life. The concept of environmental repossession is rooted in the idea that indigenous people’s health, ways of living, and indigenous knowledge systems are highly dependent on access to their traditional lands and territories. As a response to the ongoing effects of colonization and dispossession, environmental repossession demonstrates a way of thinking and an applied practice that are in broad opposition to the hopelessness that is too often used to characterize patterns of indigenous ill-health and social problems, particularly in the North American context. Working within a decolonizing methodological framework, environmental repossession may provide a promising approach for researchers and communities to collaboratively document the unique and varied processes within which indigenous communities from around the globe are asserting their indigenous rights and improving their health and wellbeing.

SEE ALSO: Cultures of nature; Environment and health; Health and wellbeing; Indigeneity; Indigenous knowledge

References


**Further reading**


Indigenous knowledge

David Barker
University of the West Indies, Jamaica

Indigenous knowledge (IK) is the body of traditional knowledge of people who have lived as a community in a local area over a long period of time. In the vernacular, indigenous knowledge is often referred to as traditional knowledge, local knowledge, or traditional wisdom. It can refer to a community’s rich knowledge of plants, trees, animals, and other facets of their local environment which are utilized in their livelihood activities. This includes knowledge of local soils, topography, water resources, and weather. It can refer to more specialized knowledge and use of specific natural resources, as in folk medicine, health, and nutrition. Indigenous knowledge is recognized as playing a role in community disaster management and conservation of biodiversity, and in contributing to our understanding of climate change. Respect for and legal recognition of the intellectual property rights of indigenous peoples and their indigenous knowledge systems was formalized in the United Nations Declaration on the Rights of Indigenous Peoples in 2007.

A precise definition of indigenous knowledge is elusive. Paul Richards, an early influential writer, tentatively suggested community environmental knowledge, folk ecology, and peoples’ science as useful descriptors. Warren’s (1991) definition is widely accepted: “Indigenous knowledge (IK) is the local knowledge – knowledge that is unique to a given culture or society. IK contrasts with the international knowledge system generated by universities, research institutions and private firms. It is the basis for local-level decision-making in agriculture, health care, food preparation, education, natural resource management, and a host of other activities in rural communities.” A cognate term is “ethnoscience,” defined as the study of the shared patterns of belief and knowledge within an ethnic group or community, and is widely used in anthropology.

“Indigenous peoples” are defined as ethnic groups with economic, social, cultural, and political characteristics distinct from the societies in which they live, and which have historical continuity with precolonial or presettler societies. However, the concept of indigenous can be less rigorously interpreted. In the insular Caribbean, Amerindians were annihilated as a result of colonialism, and a peasant class of small farmers emerged after the abolition of slavery, well after the onset of European colonization. Yet research among Caribbean small-scale farming communities also has produced valuable insights into traditional knowledge as a critical component of rural livelihoods and food security.

The main characteristics of indigenous knowledge are that it is community-based and unique, being indigenous to a specific culture group and particular to a geographical area. It is dynamic rather than static, and so enables a community to try to cope with and adapt to environmental change. The body of knowledge accumulates or is edited over many generations, as local circumstances dictate, and is communicated through interpersonal and intergenerational transmission. Thus, it is constructed, acquired, and articulated within an informal milieu, in contrast to knowledge acquired through formal educational institutional settings (based on Western science).
INDIGENOUS KNOWLEDGE

The corollary is that, in the past, IK was rarely written down. Following the United Nations General Assembly’s Declaration of the Rights of Indigenous Peoples (UNDRIP), indigenous peoples have been encouraged to document their knowledge and assert intellectual property rights.

An indigenous knowledge system is predicated on the identification and classification of locally relevant environmental and ecological phenomena, and interpretation of salient dynamic relationships between their cognitive categories. In this sense, the content of IK comprises detailed empirical local knowledge (using local names) of plants and trees, animals and insects, soils and topography, weather patterns, and so on. This empirical database forms an intellectual template within which practical livelihood skills, best practices, and traditional technologies are construed, learned, and employed in decision-making. The latter practical skills are sometimes termed “indigenous technical knowledge” (ITK), and may include cottage industry and food processing activities. The problem-solving functionality of indigenous knowledge is integral to successful livelihood strategies at the household level in resource-poor rural communities. While empirical knowledge and its content in theory can be distinguished from the practical applications of the knowledge, indigenous knowledge and indigenous technical knowledge are used interchangeably.

Epistemologically, for many ethnic indigenous peoples, IK is embedded within their local cultural and spiritual belief systems. Thus, indigenous knowledge systems reflect, concomitantly, a utilitarian classification of environmental phenomena and associated skills, and a broader cosmography of the community’s relations with the natural world. Anthropologists have documented the rich and sophisticated intricacies of indigenous peoples and their belief systems in elaborate detail. This permanent record is extremely important in the context of human history and world heritage because such ways of life and ideas are rapidly disappearing as globalization incorporates more and more of the world’s more isolated communities into the mainstream.

In geography, some of the first scholars to advocate the importance of traditional knowledge were Harold Brookfield (Pacific Islands), Paul Richards (Africa), and Theo Hills (Caribbean). Their research clearly demonstrated the utility of traditional knowledge and the inherent rationality of decision-making by tropical small-scale farmers in managing their land resources, especially with respect to risk minimization. Local knowledge underpins farmers’ choice of crops and crop combinations, techniques to manage soil fertility, methods of pest control, and many other aspects of land, crop, and animal management. Traditional knowledge is fine-tuned to specific, local sociocultural conditions and specialized agroecological regions. Early research juxtaposed the wisdom of indigenous knowledge against the negative stigmatizing and stereotyping of small-scale tropical farmers during colonial times, and lamented the discriminatory policy bias in favor of large-scale, export agriculture. Field research in geography drew inspiration from ethnography and cultural ecology. Appropriate research methodologies tapped into the emerging field of environmental perception, to explore novel ways of eliciting cognitive environmental data.

Holistic versus specialized studies of IK

A substantial amount of empirical academic research on indigenous knowledge has accumulated since the 1980s. Much of it has focused on identifying its sophisticated content and
demonstrating clever and innovative practical applications across many different cultures and agroecological conditions, and on all continents of human habitation. Since the livelihoods of indigenous peoples depend critically on the use of natural resources, studies of agriculture (crop cultivation and animal herding), artisanal fishing (coastal and riparian), and rainforest-based activities feature prominently in the literature. However, the bulk of this research relates to agriculture. Agricultural research can take a holistic view of a particular cultivation or pastoral system or focus on a specialized facet of indigenous knowledge within a local livelihood system or agroecological region.

A good example of a holistic approach is tropical home gardens. These cultivation systems occur throughout Africa, Asia, Central and South America, and regions of tropical islands such as the Caribbean and Pacific Ocean. They are known also as “compound gardens,” “kitchen gardens,” and “food forests,” and have numerous local names in local languages. Indigenous knowledge guides the spatial and temporal organization of these sophisticated agronomic cultivation systems. Tropical home gardens have dozens of different species packed together in a small geographical space, with symbiotic relationships between many plant species. Typically, there are three or four tiers of useful trees, bushes, and plants, ranging from tall trees to ground-level plants and tubers. Plant censuses of valuable species present in a single garden are as high as 150–250 in places like Java and Sri Lanka. They are highly productive and play a critical role in household food security. They provide a nutritionally well-balanced diet because small animals such as chickens, goats, and pigs are often incorporated into the assemblage. Many nonfood items, such as plants for folk medicines, dyes, cosmetics, herbal teas, craft and household items, and trees for construction, furniture, and roofing material are usually present in these gardens. Tropical home gardens help cement social capital within extended families and rural communities through the sharing and exchange of food, they play a role in children’s botanical education, and are mobilized as a form of community-based, disaster relief after a tropical storm or other natural disaster. Tropical home gardens are an example of traditional agroforestry, a more geographically extensive system of farming which provides households with additional economic opportunities through commercial tree crop production. It is particularly well documented for the Pacific Islands.

An example of research focusing on specialized facets of indigenous knowledge is the study of traditional methods of soil classification. Soil color and texture are universal practical methods used by farmers to assess local soil fertility. Ingenious, more ad hoc techniques include soil taste, the presence of earthworm casts, and depth of footprints. Africa is strongly represented in the ethnopedology literature. Research on indigenous pest control methods is another example of more specialized research. Its focus, not surprisingly, tends to be on visible insect and animal pests rather than on microscopic organisms. An important theme in traditional pest management is the biodiversity of cropping systems, which creates conditions in microhabitats for natural pest control through predation, and lower insect herbivore populations than in monocropping.

**IK and development planning**

In the 1980s, a new dimension to indigenous knowledge research emerged: its use as a potential tool in development planning. The new
direction resulted from the seminal work of Robert Chambers on rural development, which brought IK to center stage in the development discourse. He argued that local indigenous knowledge should play a more important role in development planning, in partnership with conventional scientific knowledge, to help empower rural communities in developing countries. Chambers (1983) explored connections between IK and participatory development, and reported on innovative community-based local methodologies such as participatory mapping and natural resource inventories. Later, Chambers was involved in developing the pivotal concept of rural sustainable livelihoods, which again highlighted indigenous knowledge and traditional sustainable farming practices as key components of successful livelihood strategies.

As indigenous knowledge was integrated into development thinking and development practice, it became part of the institutional conventional wisdom through endorsement by international agencies. For example, Agenda 21 and the Rio Summit in 1992 highlighted its potential role for sustainable development. Many traditional preindustrial communities whose ways of life were connected intimately to indigenous knowledge had once lived in harmony with the natural world. So, low environmental impact solutions to sustainable food production, for example, might be found in the best practices of livelihood systems based on indigenous technical knowledge. The 1992 UN Convention on Biological Diversity recognized the value of traditional knowledge in the protection of species, ecosystems, and landscapes while the World Bank’s endorsement of the value of indigenous knowledge in 1998 consolidated its place in international development policy debates.

An ongoing debate revolves around the perceived need to demonstrate the veracity of IK to a seemingly skeptical world of formal science-based knowledge. The classic example of the scientific validity of traditional farming practices is the use of intercropping or inter-tillage. Intercropping is ubiquitous in the tropics. But it was derided and held in low esteem by colonial agricultural officers trained in European farming techniques, until scientific research clearly showed it had many advantages over monoculture, and could be more productive and profitable. Often, Western science and indigenous knowledge are presented as two competing knowledge systems and there may be tension and conflict between their proponents regarding merits and demerits, often played out in relation to the adoption or nonadoption of agricultural innovations. Western science tends to be equated with modernism, objectivity, openness, transferability, and objective and detached rationality, whereas indigenous knowledge is seen as primitive, parochial, nonrigorous and therefore nonintellectual, emotive, and associated with “backward” ways of life. The situation is often characterized as “us versus them.” While there is some truth in these characterizations and conflicts, the widespread acceptance of integrated pest management, whereby traditional methods of pest control are combined with low-cost, low-environmental impact, science-based methods, illustrates that successful practical “partnerships” between science and IK are possible.

Briggs (2005) has argued that indigenous knowledge has not lived up to earlier expectations in promoting sustainable development in resource-poor rural communities. While acknowledging the literature successfully demonstrates its importance to local communities, such research has had little impact on development practice. He suggests this represents a “false dawn,” and that indigenous knowledge has become “the Cinderella of development.” Others, more trenchantly, label IK as a new rhetoric for development professionals. Briggs
argues a more comprehensive understanding of the epistemology of indigenous knowledge is required with a greater focus on process rather than content. This implies a shift toward a better understanding of how local people observe, discuss, question, and analyze their problems. It requires the interrogation of indigenous knowledge and externally referenced, science-based knowledge by local people and outsiders, and offers a better route for developing synergies between the two knowledge systems. By way of example, Jamaican farmers were insightful and reflective in their thinking about a government-sponsored innovation to introduce minissett yam. But ultimately they rejected the innovation, often with sound reasoning. In contrast, they were positive and constructive in their reactions to another proposed innovation, a plastic yam stick, and engaged researchers with their own suggestions as to how to improve the technology. Such reasoning underscores another important theme in the indigenous knowledge literature; small-scale farmers often conduct their own informal experiments which draw on their indigenous knowledge and experience, and that indigenous innovation is more common than generally acknowledged.

Two other important issues concern the relevance of IK to development planning. First, indigenous knowledge tends to be distributed unevenly within a group. Not all farmers or fishermen in a community are recognized as local experts by their peers. Further, the body of community knowledge may be differentiated and compartmentalized reflecting specialization and the division of labor within the group, or the distribution of age, wealth, and gender, or power and influence. All these variables present challenges in understanding resource management and production orientation, and are possible sources of conflict in mobilizing effective development initiatives. The second issue relates to the place-specific nature of IK, and the fundamental problem of transferability and scaling up. How can local knowledge that provides great insight into livelihood strategies within a particular location, with unique agroecological and socio-cultural conditions, be applied in another community or region, which is likely to have different conditions? Similarly, how can local knowledge be transferred to the national level or from one country to another country, again, where prevailing conditions are likely to be different? Strong reservations have been expressed by some experts about the usefulness of local knowledge to development planning outside the spatial locales in which it was contextualized.

**IK and climate change**

More recently, climate change has given a new impetus to research on indigenous knowledge. The detailed knowledge of farmers in identifying and monitoring changing weather patterns has long been recognized in the literature. Resource-poor farmers, whose decision-making and livelihoods critically depend on an acute understanding of weather patterns, are often the first to detect subtle changes in weather patterns. Many studies among arable farmers and pastoralists throughout Africa, Asia, and Latin America have documented the sensitivity of farmers to more extreme weather conditions, including increasing aridity and seasonal shifts or variations in the rainy season. Nigerian research, for example, illustrates how farmers factor traditional knowledge of rainfall, thunderstorms, wind storms and the harmattan wind into their decision-making. In the polar latitudes, where early signs of climate change are already evident, detailed ethnographic studies of indigenous people in Canada and the Russian boreal forest have
INDIGENOUS KNOWLEDGE

reported how they have been tracking climate change since the 1990s.

The Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC) highlighted indigenous knowledge as an invaluable tool to help develop adaptation strategies for natural resource management in the face of environmental change. Not surprisingly, therefore, an active research field has emerged incorporating indigenous knowledge into a climate change framework. The IPCC, hitherto anxious to establish the scientific accuracy of its predictions of climate change, more recently acknowledged the potential role of the indigenous knowledge in corroborating scientific evidence through oral history. Indigenous knowledge, based on long-term observation of weather patterns by local people, can supplement scientific knowledge about climate change. An example is Jamaican research on small farmers’ knowledge of weather patterns, where their reports of changing weather patterns corroborated scientific findings based on satellite imagery, showing that the short rainy period May/June has become increasingly unpredictable in timing, duration, and the amount of precipitation over the last 30 years.

Recognition of the value of indigenous knowledge by the IPCC represents a new avenue that might reduce the bipolar tensions between Western science and indigenous knowledge, alluded to by Briggs. Oral history can complement scientific knowledge. But identifying cost-effective adaptive strategies among resource-poor communities is an urgent priority for the developing world. Indigenous knowledge and traditional best practices can be a foundation to help build capacity to mobilize participatory, community-based mitigation strategies in collaboration with external interventions. Current research on climate change and food security exemplifies such an approach; this burgeoning research field is replete with projects focused on how to use traditional practices to strengthen community resilience to climate change for a more sustainable future.

SEE ALSO: Agricultural geography; Agroforestry; Climate change adaptation and social transformation; Cognition and spatial behavior; Community-based natural resource management; Food security; Human ecology; Participatory development

References


Further reading

American Indian reservations experienced a remarkable turnaround in economic development following the passage of the 1988 Indian Gaming Regulatory Act, which sanctioned the legal recognition of tribal rights to operate gaming enterprises. By 2008 tribal gaming revenues topped $26.7 billion, creating more than 636,000 jobs for the 233 Indian tribes and two Alaska Native Villages that operated 411 casinos, bingo halls, and pull-tab operations in 28 states. Tribally owned and operated casinos reached annual revenues and employment that exceeded all of the non-Indian gaming in Las Vegas (Schaap 2010, 369). The success of tribal casino gaming has also expanded into other aspects of economic, community, and regional development. This dramatic success of tribal casino development does not affect all reservations equally, however, and many continue to reflect persistent poverty and lack of economic opportunity. Research has shown that it is not only economic factors that impact tribal success, but that success is linked to the effective evolution of tribal governments as tribes have taken control of their own programs, lands, and resources and seek to reinvest to create sustainable, sovereign tribal nations.

Indigenous economies: from success to dependency

American Indian tribes established successful Indigenous economies prior to European
contact that evolved over time, linked to the unique resources available within their Indigenous lands and surrounding areas. These successful economies were greatly disrupted as a result of European contact, and during the following centuries the expansion of European colonies sought to ignore or eliminate those cultures through isolation, removal, and acts of genocide that denied political and economic, rights, despite the official “recognition” documented in treaties and the establishment of reservations. American Indian nations were categorized as “wards” of the federal government and became dependent upon federal rations, resources, and allocations through a special “trust” relationship established in those treaties.

A central factor of tribal economic transformation has been the ability of tribes to overcome the dependent status and isolation of many reservations through the concept of sovereignty: each American Indian reservation retained powers, generally expressed in treaties, executive orders, or congressional action, that identified American Indian reservations as empowered through federal relations with each individual sovereign tribal nation. Each tribe also retains inherent sovereignty from the powers understood in the meaning of words within each tribal language and culture when treaties were formed regarding their lands, resources, and culture.

A number of books have documented the resiliency of American Indian culture (see McNickle 1968) and the perseverance of American Indian peoples and cultures, despite forced assimilation and acts of genocide. Despite isolation on the periphery of non-Indian economies for many reservations, and woeful records of the lack of support for the education of American Indian peoples and their rights to support and develop their own tribal governments, specifically through the 1934 Indian Reorganization Act and the 1975 American Indian Self-Determination and Education Act, tribal governments have emerged since the 1990s as effective decision-makers and agents of good government for each American Indian reservation. The expansion and success of tribal governments has been a catalyst to overall reservation economic development, with tribal governments emerging in the twenty-first century as a major employer on the reservation, and a major employer within their region.

Understanding tribal economic development

A major force in understanding tribal economic development has been the long-term engagement of the Harvard Project on American Indian Economic Development, started in 1987 “to understand and foster the conditions under which sustained, self-determined social and economic development is achieved among American Indian nations through applied research and service” (http://www.hks.harvard.edu/hpaied/). Working directly with individual tribes in ongoing research and recognition of best practices and successful economic development, four major findings emerged through this research as central to tribal economic development success: sovereignty matters, institutions matter, culture matters, and leadership matters (Harvard Project on American Indian Economic Development 2011).

The persistence of poverty on reservations remains, however, and Leichenko (2003, 367) found that increasing regional disparities in income seem to be getting worse for rural areas – and potentially for American Indian reservations – as a result of globalization, and that “tribal areas may be among those that are most likely to be left behind because of lower levels of human capital, poor access to
markets, poor infrastructure, and other constraints.” Leichenko’s research also recognized the complexity of tribal economies and the continued role of noncash exchange and obligations, which remain important aspects of reservation economic analysis.

**Barriers to American Indian reservation regional development**

As a legacy of economic dependency on the federal government and the structures that created and restricted full tribal control of decisions, many barriers and restrictions to tribal economic development remain. A major issue is land ownership on reservations which were initially established on tribal lands held “in trust” for the whole tribe, with no fee-simple ownership. Tribal trust lands were subsequently subject to allotment through the Dawes Act in 1887, which allocated tribal lands to individual tribal members and their families, and allowed any surplus lands to be transferred to non-Indians. As a result, over 90 million acres were removed from Indian ownership (Indian Land Tenure Foundation 2014). Not all reservations were impacted by allotment, so a complex system of land ownership exists in terms of tribal trust lands, tribal lands assigned to individuals, tribal member ownership of reservation lands, and non-Indian ownership of reservation lands.

In addition, under the General Allotment Act the Bureau of Indian Affairs (BIA) established a complex structure for allotted lands so that when the allottee died, title ownership was divided among all of the heirs, but the land itself was not physically divided (Indian Land Tenure Foundation 2014), resulting in fractionated ownership where reservation lands owned by American Indians could be owned in part by hundreds of descendants (heirs), each with a vote for a percentage of “shares” so that no leases or transfer of lands could occur without a majority vote of all heirs. Only recently have tribal codes and changes in federal law sought to resolve this issue as a result of tribal initiatives. These two factors create uncertainty in land ownership and questions about transfer or sale of land on reservations, a major barrier to standard US development and investment patterns, where land can be used as collateral for loans.

A third factor related to land ownership is the overall development patterns on reservations. As a result of the complex land-ownership patterns, many reservations located tribal offices, housing projects, schools, health care, and other facilities on isolated tribally owned lands, often separated by great distances. As non-Indian communities embrace “smart growth” as a means to overcome auto dependence and the failure of suburban land-use patterns to create livable communities, American Indian reservations are left with some of the worst, least sustainable land-use patterns in the nation, which result in dysfunctional, auto-dependent settlements that fail to provide access for many within each reservation, and restrict the opportunities for health and livability.

Transportation is often associated with economic development in terms of access to markets, and reservations were intentionally located away from urban non-Indian communities. Many reservations developed with limited or no funding for transportation, so that non-Indian regional interests and priorities demanded federal and state roads through reservation lands. One solution that tribes have actively implemented since 2008 is the creation of tribal transit programs to create access and resolve tribal transportation needs, along with the active assumption of responsibility for tribal transportation programs.
Tribal development beyond the casino

Tribal governments are making changes toward the future that incorporate strategic planning, comprehensive plans, and seven-generation (150-year) plans as mechanisms for tribal leadership to address the future and to engage tribal communities in effective decision-making. A major response to new revenues from casinos has been the purchase of critical Indigenous tribal lands that should remain part of the reservation “trust” lands, or that have potential for development. Many tribes have made major investments in tribal language, history, and culture programs, and in tribal wellness and educational-cultural facilities. Sustainability, including local foods/cultural foods and initiatives to create more compact settlements and sidewalks, paths and trails separated from highways to promote walking, access and exercise, are also part of long-range planning. Investment and governance development strategies that impact over seven generations and that frame sustainable sovereign tribal nations and tribal economies are a priority for many reservations.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Dependency theory; Development; Economic development zones; Environmental (in)justice; Globalization and rural areas; Governance and development; Indigeneity; Local development; Postcolonial geographies; Poverty; Rural geography; Social justice; Sovereignty; Sustainable development; Territory and territoriality

References

Indigenous technical knowledge

Mara J. Goldman
Eric Lovell
University of Colorado Boulder, USA

“Indigenous technical knowledge” (ITK) is a term broadly used to refer to knowledge about the local environment that is produced, held, and used by indigenous communities. The term is often used interchangeably with the following terms: “traditional ecological knowledge” (TEK), “indigenous knowledge” (IK), “local ecological knowledge” (LEK), “local knowledge” (LK), and “indigenous technology,” which are all commonly defined in contrast to scientific knowledge – epistemologically and ontologically. ITK is assumed to be handed down from one generation to the next, and to be culturally and geographically situated. Scholars and activists often refer to ITK as a “way of life,” suggesting a knowledge–belief–practice system, embedded in everyday performances. While the value of ITK has been heralded by groups as diverse as the World Bank, the United Nations, and indigenous advocacy organizations, the problematic nature of the terminology and its application have been actively debated within the academy.

Discussions about ITK emerged alongside the “farmer first” movement within development, which called for more participatory forms of development and a focus on farmers’ pre-existing knowledge about their local environment. The argument was made that rural development should be based on indigenous and local knowledge about local environmental conditions, since such knowledge is more reflective of and adapted to local conditions, does not rely on heavy technical inputs, and is more culturally appropriate than Western interventions. Practically, the farmer-first movement sought to proactively involve farmers in development, through participatory processes and by challenging power dynamics and hierarchies of knowledge between Western scientists, extension officers, development workers, and local communities. Theoretically, this challenged the superiority and universality of Western scientific knowledge, raising the possibility of multiple ways of knowing. Such calls were taken up in allied fields (academic and applied), with claims of the importance of ITK for biodiversity conservation and ecosystem management. Geographers entered the debate both as proponents and as critics, discussing the use of ITK as it related to rural development, conservation, and mapping.

The acknowledgment of the existence of ITK was an important step in providing indigenous people with a voice concerning the management of land and resources, while challenging conventional wisdom regarding environmental processes and knowledge production more broadly. Recognizing and naming ITK enabled more robust discussions of ecological management as well as of the politics of mapping in indigenous territories. It also raised complex questions and critiques related to what ITK means, how it should be studied, and how it can be used, including problems with the “indigenous,” “technical,” and “knowledge” components of the name itself. This is evident from the use of multiple acronyms to refer to ITK.
INDIGENOUS TECHNICAL KNOWLEDGE

To start with, ITK must, by virtue of its name, refer to knowledge held by "indigenous" people or communities. This raises important practical and political questions regarding who is indigenous and how this gets decided. For many, the indigenous aspect of ITK is strategically important, drawing attention to specific knowledge and ways of life of indigenous peoples, while highlighting the politics of knowledge suppression and hierarchy that corresponded with the spread of imperialism and the growth of modern science. This is particularly important as indigenous peoples struggle to maintain rights to their own land and resources around the world. However, defining who is indigenous is not always a straightforward process, reflecting political motives and questions of origin. The process of designating certain people as "indigenous" often excludes others, such as more recent migrants to particular areas who are also vested in management outcomes. Additionally, a focus on "indigenous" knowledge as somehow isolated and different risks essentializing and romanticizing the knowledge and the communities as homogeneous and isolated from larger political economic contexts.

The T part of the ITK acronym suggests a focus on technical knowledge which can be incorporated into modernist projects and scientific studies with relative ease. While initial proponents argued that this focus put ITK on a similar footing with science in providing particular techniques for development and conservation projects, others have suggested the focus is misguided. Scholars and activists alike have argued that a focus on "technical" knowledge not only disempowers the very users of ITK but also dilutes the value of the knowledge at hand. If the embedded and active nature of ITK is acknowledged, then it must be seen as linked to entire belief systems from which particular "techniques" cannot and should not be extracted. Doing so ignores the cultural processes of knowledge production and risks turning complex multifaceted knowledge production processes into discrete pieces of data. Extracting particular techniques, separate from the culture and beliefs to which they are tied, risks losing the meaning and value altogether. Some have argued that this process disempowers indigenous communities by recognizing only certain aspects of their knowledge as valuable, and only when it can be adequately incorporated into Western science.

This gets to the final part of the ITK acronym, "knowledge." ITK is often discussed as an altogether different type of knowledge than science. ITK is often viewed as practical (utilitarian) knowledge based in craft (i.e., techne) and intuition. Science, on the other hand, is viewed as grounded in theory, empirical analysis, and justified belief (i.e., episteme). While some scholars argue that ITK is also based in experimental knowledge, it is often still understood in relation to (and often opposed to) Western science. ITK is holistic, situated (i.e., local), embedded in a worldview, fluid, experiential, relational, and closed (to new ideas/innovations). Science, on the other hand, is understood as reductionist, rational, abstract, universalizing (i.e., global), rigid, predominately quantitative, and open (to paradigm shifts). Consequently, in order for ITK to be useful for science and management, it must be “integrated” into science. Some have critiqued this process as furthering the dominance of science while simultaneously diluting ITK and disempowering ITK holders (Nadasdy 1999; Louis 2007). Others have opted to talk about bridging knowledge types so as to avoid upholding knowledge hierarchies while still bringing indigenous knowledge into conversation with Western science (Rocheleau 2003). Yet others have argued against the focus on difference altogether, arguing that such a focus upholds
knowledge hierarchies and ignores the historical situatedness of Western science itself (Agrawal 1995). However, some scholars and activists insist that there is a need to recognize differences – in methods, epistemology, and ontology – as based in the colonial and imperial relations and highly contested politics associated with the production of Western science (Goldman 2008).

Recent work within geography and allied fields have taken discussions of ITK further, drawing from the traditions of political ecology, science studies, posthumanist theory, and participatory mapping. Political ecology has highlighted the historical links between indigenous communities and larger political economic processes that have impacted processes of knowledge production, transfer, and use over time. This has often included a focus on the creation of “hybrid” knowledge across indigenous and nonindigenous groups, including scientific communities. Political ecologists have also stressed differences that exist within communities that impact knowledge production and use (i.e., gender, class, caste, ethnicity). This work has promoted the importance of recognizing ITK, while simultaneously problematizing how it is perceived, studied, and discussed. Much recent political ecology work has also drawn on science studies for a more robust treatment of knowledge production that recognizes the situated, embodied, and performative nature of all knowledge – science and that which is called ITK.

These insights are also reflected in the cartographic wing of geography, particularly in relation to participatory geographic information systems (GIS) and indigenous mapping. Here, scholars and activists have been actively engaged in the political and theoretical questions raised earlier. There has, for instance, been a large focus within the Association of American Geographers (AAG) Indigenous Peoples Specialty Group on the importance of indigenous voices in geographical research (Louis 2007). Scholars have emphasized ethical concerns that arise when working with indigenous peoples and on ITK, and the methodological, ontological, and political issues related to the interpretation and representation of ITK with Western scientific tools (GIS and cartography). Specifically avoiding the pitfalls raised earlier about extracting data, such work calls into question how maps can be created with the tools we have and challenges us to be more open to other ways of knowing. This reflects ontological as well as epistemological questions and pushes our understanding and use of ITK in new and critical directions. All of these new directions focus on the importance of seeing knowledge as process, which expands how ITK (and scientific knowledge) can be viewed and understood.

SEE ALSO: Environmental knowledges and expertise; Indigenous knowledge; Indigenous people and regional development; Livelihoods; Participatory action research; Participatory development; Political ecology; Posthumanism; Public-participation GIS

References


INDIGENOUS TECHNICAL KNOWLEDGE


Further reading

Industrial agglomeration

Ke Chen
East Tennessee State University, USA

Industrial activities have never spread evenly over a featureless plane. While some industry development is driven by natural resources, such explanation cannot fully illuminate industrial agglomeration in most cases. Instead, the very nature of being close to each other may be a fundamental mechanism that accounts for the spatial formation of urban areas and industrial development. From early scholars such as Alfred Marshall and Alfred Weber to modern ones like Paul Krugman and Michael Porter, the phenomenon of industrial agglomeration has attracted keen attention from economists, planners, sociologists, and historians, as well as geographers. Numerous studies have been conducted to unravel the myth of the power of geographic proximity in economic activities, various terms have been created to highlight some distinct natures of industrial agglomeration, and a wide variety of methodologies have been proposed to scrutinize the formation of business concentration. Despite the centrifugal force of globalization to re-organize industrial activities on a global scale, the centripetal power that draws businesses in spatial proximity endures.

Advantages of industrial agglomeration are first mentioned in Alfred Marshall’s prominent book Principles of Economics in 1890, where he explicitly describes how localized small businesses of similar characteristics can benefit from external returns to scale. According to Marshall, industrial agglomeration can originate from natural conditions, including favorable soil and climate characteristics, proximity of mines and quarries in the neighborhood, and easy accessibility by land or water. It can also develop from social and historical reasons, such as family groups expanding into villages, the patronage of a court, and the systematic summoning of artisans by rulers to build a town. Sometimes, the emergence of a localized industry may appear by chance. No matter for what reason industries start to emerge and concentrate, all of the businesses in the locality can take advantage of being close to each other. Although Marshall does not develop a fully-fledged theory of industrial agglomeration, he lays the foundation for the investigation of the subject.

Twenty years after Marshall published his original book, another economist, Alfred Weber, developed a theory of industry location based on minimum cost in production (Weber 1909). Weber defines agglomeration as the concentration of economic activities or entities in particular localities. Agglomeration allows individual firms to enjoy both internal and external economies, where benefits come from reduced cost in transportation and infrastructure, improvement in labor organization, and enhanced efficiency. Weber’s theory on industrial agglomeration is regarded as classical and his insights on the role of transportation costs have been inspirational for geographers in both theory formulation and empirical studies.

While stressing the importance of spatial concentration, Weber’s theory does not make a distinct difference when explaining internal scale economy and external scale economy. It is Hoover (1948) who fills this gap and clarifies
three types of agglomeration: (i) internal scale economy, (ii) localization economy, and (iii) urbanization economy. Internal scale economy refers to the return to increased scale of operation for individual firms. Localization economy benefits from the clustering of firms in the same industry, similar to that described by Marshall (this text uses Marshallian-type agglomeration to refer to the localization economy). Urbanization economy denotes the concentration of industries in urban areas to take advantage of lower infrastructure cost, large market, extensive knowledge spillovers, and large labor pool.

Although it may be a difficult task to quantify the importance of internal and external agglomeration effect in the formation of business activities, some analyses of employment growth across American metropolitan systems support the claim that external economies of scale are replacing internal scale economy in industrial locations (Ó hUallacháin 1990). In this entry, terms that are closely related to localization economy, for example, industrial complex, industrial district, and industrial cluster, are first introduced. The cause and impact of urbanization economy are discussed next, and then recent studies on the key factor of knowledge creation in industrial agglomeration in the era of knowledge economy are discussed.

Localization economy: industrial complex, industrial district, and industrial cluster

In the 1950s, Isard developed a concept of industrial complex to describe a set of activities at a specific location that are linked by certain technical and production interrelations. In Isard’s (1959) study of regional economies of Puerto Rico, he found that traditional economic theories on individual industries failed to develop appropriate spatial concepts to explain the economic growth of the region in a historical perspective, since any basic industry is always interrelated with other sectors. To fill the gap, Isard developed a model of industrial complexes, where businesses benefit from agglomeration by reduced transportation and transaction costs within the input–output network, a large labor pool of professionals, and technology spillover (Gordon and McCann 2000), all of which resonated with earlier literature in industrial agglomeration.

Isard’s ideas and methods in industrial complex theory are followed and improved by some later scholars (Czamanski and Ablas 1979; Roepke, Adams, and Wiseman 1974; Ó hUallacháin 1984). Nevertheless, reliance of input–output tables for identifying industrial complexes becomes more and more subject to criticism. For instance, Latham (1976) pointed out that input–output tables are purely aspatial and concludes that they are unsuitable to help identifying industrial complexes. Ó hUallacháin (1984) acknowledged the weakness of principal components analysis for detection of vertical value chain linkages. Moreover, too much focus on technical issues has limited the application of the industrial complex as a policy tool to boost regional economies in the 1970s, when the world was going through substantial changes in a more globalized world: international capitals were moving more easily into developing countries and manufacturing companies were shifting their location of production to cheap-labor abundant parts of the world. Studies of industrial complexes fail to address new patterns of production and international division of labor, and therefore start to lose popularity in both academics and public policy.
When the western world was hit severely by the oil crisis in the late 1970s and early 1980s and fell into economic recession, some peripheral areas in northern Italy surprised the industrialized nations with remarkable resilience and growth, driven by small-scale businesses in labor-intensive industries. Under such a background, the Italian type “industrial district” became a popular phrase, and the model motivated economic development organizations across the world as an approach to stimulate economic growth and job creation.

Becattini (1990a) defined an industrial district as a socioterritorial entity that is characterized by the active presence of both a community of people and a population of firms in one naturally and historically bounded area. In a district, unlike in other environments such as manufacturing towns, community and firms tend to merge, sharing knowledge on a specific industry originated in history and tradition. Some industrial districts in northern Italy, such as textile businesses in Carpi and Prato, and the furniture industry in Brianza and Cascina, have a coherent location and a narrow specialization profile. Locating in such a district enables many Italian small businesses to become competitive against large enterprises in the international market.

An important characteristic of industrial districts is that such local milieus are occupied with small or medium-sized firms, a production system totally different from the vertically integrated Fordist model. In addition to the classical driving forces in Marshallian-type localization economy, another key for the success of industrial districts in previously peripheral places is an environment of cooperation and competition in the social structures and practices. Such a setting supports dense information exchange, knowledge creation, and technology innovation in both traded and untraded forms (Piore and Sabel 1984; Becattini 1990b; Scott 1994). Success of industrial districts can also be attributed to a strong sense of belonging, where a local community of people shares similar values and views, as traditional Italian culture is based on the idea that social relationships of kinship and friendship coexist with market relations. Some other essential factors for success include an ethic of labor and activity, credit system for small firms and so on (Becattini 1990a). Overall, innovation associated with social ties and network is pervasive in industrial districts.

Unlike investigations on industrial complexes that emphasize the use of quantitative methods to identify interindustrial trade relations, studies of industrial districts are more focused on places, historical context, social structures of production, and the creative environment in the local milieu. Industrial district models show that, with limited resources to invest in research and development, firms are capable of remaining competitive in a new international production and distribution system. Therefore, “industrial districts” became a much more popular term than “industrial complexes” as a policy tool for regional development and industrial strategy in the 1980s.

However, while scholars have applauded the use of industrial districts as a potential model for economic growth or recovery, they faced an even more challenging economic picture in the late twentieth century. With a more deepened level of globalization, advanced economies continue to lose comparative advantages in manufacturing jobs and capital flows. Although there are a few successful industrial districts, the social, economic, and historical peculiarity of such districts makes it hard to easily replicate and provide a general formula for similar success in the industrialized world.
INDUSTRIAL AGGLOMERATION

Against such a background, Harvard Business School scholar Michael Porter (1998), in his renowned book *The Competitive Advantage of Nations*, discusses the innovative nature of competition of industrial clusters. Porter defines clusters as geographic concentrations of interconnected companies and institutions in a particular field. Clusters encompass an array of interlinked industries and other entities important to competition, including suppliers of materials, services, and specialized infrastructure. Clusters also often extend downstream to include customers, manufacturers of complementary products, and companies related by skills, technologies, or common inputs. Furthermore, many clusters include governments and other institutions, such as universities, standards-setting agencies, think tanks, vocational training providers, and trade associations.

Porter’s cluster theory particularly addresses the difference from the classical theory of comparative advantage, which is based on factor endowment. In comparison, Porter argues that prosperity is created, not inherited. It does not grow out of natural endowments, labor pool, or current economic situations such as interest rates or exchange rates. Instead, a locality, region, or nation can create new advanced factor endowments such as skilled labor, a strong technology and knowledge base, government support, and culture.

In particular, Porter stresses the active role of governments to grow industrial competitiveness, where the government may act as a catalyst and challenger to encourage or even push companies to raise their performance. Governments may stimulate early demand for advanced products, focus on specialized factor creation, and encourage local rivalry by limiting direct cooperation and enforcing antitrust regulations. Meanwhile, these government actions are different from traditional industrial policies, where government targets desirable industries and intervenes in competition to tilt market outcomes in a nation’s favor. In cluster theory, all clusters can improve productivity and deserve attention. The focus is not on distorting competition but removing obstacles and constraints to grow productivity.

Compared to the concepts of the industrial complex and industrial district where success rests on trade linkages, culture, tradition, history, and community sense, the model of industrial clusters is more encompassing by offering a more universal and practical framework for local businesses and governments to improve competitiveness in a global economy. As a result, cluster theories have been warmly received among regional planners and industrial strategists since the 1990s. It is considered a canonical model for economic development in recent years. The idea itself is also related to the knowledge economy or applied to knowledge-based clusters (Maskell 2001).

Urbanization economy

A second type of external industrial agglomeration effect comes from urbanization economy, which denotes the concentration of industries in urban areas to take advantage of reduced transportation and transaction costs, lower infrastructure cost, a large urban market, the availability of specialized inputs, large labor pool, and extensive knowledge spillovers (Mills 1967). Overall, there is a positive connection between city size and productivity, where increases in scale are likely to generate greater specialization, higher levels of human capital, and other positive externalities.

One set of theories about external effect emphasizes the gains that come from reduced transportation costs of goods and people (Krugman 1991). For firms of all kinds, locating in urban areas means proximity to both suppliers
and customers. In the study of manufacturing belts in the United States, Krugman addresses the importance of proximity to market based on increasing returns to scale. It is cheaper to produce one single product in large quantity than a variety of products in smaller quantities. With fixed production costs, transportation costs become an important factor in the firm’s location decision, which can be reduced by locating in proximity to large markets. In Krugman’s model, urbanization economy, together with the effect of internal scale economy, helps to increase profits at the firm level.

In addition to reduced transportation and transaction costs that come with city size, Chinitz (1961) provides some more insights on the effect of urbanization economy. Based on his observations of New York and Pittsburgh, the former being a large urban area and the latter a localized economy, Chinitz argues that a large urban economy tends to have a more diversified industry structure. This implies more entrepreneurial supply and capital supply, which is particularly essential for small businesses. Overall, a diversified urbanization economy has a higher prospect of success than localization economy. Jacobs (1969) also stresses that diversity is the source of urban economic growth. Her thesis is that cities grow because of innovation they stimulate and urban areas that contain many independent firms performing a multitude of tasks will grow rapidly.

A more recent argument on the positive externality of urbanization economy can be related to Richard Florida’s (2002) thesis of the creative class. Florida views creativity as the precursor of competitiveness. He keeps on arguing that, in the era of globalization, US cities with diversified economic activities and human capital have developed a creative class while others have witnessed the erosion of creativity. Comprised mainly of scientists, engineers, artists, and other professionals, the creative class generates or advances technology to produce commodities and services, and constitutes a significant part of the population of cities that are characterized by diversity and tolerance in terms of ethnic, cultural, religious, and sexual orientation. Therefore, urban areas that can provide creative classes with amenities that nurture their creativity will enjoy prosperity and lead economic growth.

To summarize, there is a growing shift in the study of urbanization economy from transportation costs and market size to urban diversity and innovation. Urban economists like Glaeser and Gottlieb (2009) and Fujita and Thisse (2002) agree that earlier manufacturing firms cluster to reduce the costs of moving goods, but this force no longer appears to be important in driving urban success in the contemporary world. Instead, modern cities are far more dependent on the role that urban size and diversity can play in speeding the flow of ideas and generating new technology.

Agglomeration economy and knowledge creation

In the more recent literature on industrial agglomeration, there seems to be a tendency to investigate the social nature rather than the pecuniary nature of the benefits from agglomeration, whether it is localization or urbanization economy. A careful examination of industrial cluster and creative class accounts shows that both agree on the importance that geographic proximity plays in innovation. But why is it so? Some scholars propose that the key to understanding industrial agglomeration lies in its superior capability of creativity and innovations due to face-to-face contact, networks, and social interaction (Fujita and Thisse 2002). Storper and Venables (2004) develop some foundations for
face-to-face contact amongst economic agents and show how these improve coordination and collaboration, and subsequently increase productivity in close geographic proximity. They view the understanding of industrial agglomeration as incomplete unless grounded in the most fundamental aspect of face-to-face contact, which is an efficient communication technology, which helps solve incentive problems, is capable of facilitating socialization and learning, and provides psychological motivation.

While the diffusion of knowledge within the local milieu is rapid, it can be quite slow outside the industrial cluster. Although an increasing proportion of skilled human capital such as top management and experts has become internationally mobile, an important part of human capital resides in a multitude of local formal and informal networks that have been built over a long time, and therefore cannot be taken out of context without losing much of their value. Furthermore, the large groups of middle and lower level managers and workers, who make up an important part of the formal and informal relationships between firms, are typically much less mobile than top management and experts. As a result, a large literature has demonstrated the existence of localized, industry-specific knowledge spillovers within the science and technology communities.

Geographic proximity, characterized by face-to-face contact may also facilitate cross-sectorial specialized networks and social interaction. Networks of firms and industries clustered within regions interact more heavily with university-based scientists in proximity than with those further way. For example, Washington DC has become a major high-technology region in biotechnology due to close interactions with government agencies such as the National Institutes of Health. The accomplishments of Silicon Valley and the biotechnology cluster in San Diego can also be attributed to the close communication among scientists, entrepreneurs, and venture capitalists. Co-location is especially important to these processes because it provides a low-cost means for new ideas and talent to make their way into existing activities, by facilitating access for newcomers and by lowering the costs of evaluation on the part of those already in the relevant loops. New relationships are hence made easier, cheaper, and much more effective than they would be without close geographic proximity and the opportunities it provides in terms of face-to-face contact and social networks through formal and informal ways.

These explanations offer some essential accounts on the localized nature of technology spillover and knowledge creation in industrial agglomerations. In today’s global economy, the advancement of transportation and communication and the integrated global market have led to a predictable surge of outsourcing, with companies relocating many facilities to areas of low wage, taxes, or other input costs. However, competitiveness in the long run will rest on a locality’s knowledge pool, human capital, social interaction, and the associated innovative capabilities, all of which are enhanced by geographic proximity.

Conclusion

To summarize, over a century has passed since the original work on industrial agglomeration was published, and the concept has evolved from emphasis on economic cost reduction and benefits from trade relations to social structures and environment that support dense information exchange and constant innovation. Even though the forces of globalization have modified the world production system and spread industrial
INDUSTRIAL AGGLOMERATION

activities from advanced economies to developing countries, the centripetal agglomeration force that attracts businesses together still remains, and proves to be even more important for knowledge generation.

SEE ALSO: Industrial complex; Industrial geography; Industrial linkage

References


FURTHER READING


Industrial complex

Ke Chen
East Tennessee State University, USA

Walter Isard (1960) defined an industrial complex as a combination of industrial activities occurring at a given location that are subject to important production, marketing, or other trade relations. By locating in an industrial complex, individual firms may reduce transportation and transaction costs within the particular input–output production network (Isard and Vietorisz 1955). Firms in an industrial complex may also enjoy a large labor pool, and take advantage of potential technology spillover, as a result of social structures and practices that support information exchange and innovative activities (Gordon and McCann 2000). Therefore, a region with an industrial complex has comparative advantages over other places and may be more likely to achieve economic prosperity.

Isard, Schooler, and Vietorisz (1959) developed this idea of the industrial complex in their study of the regional economy of Puerto Rico in the 1950s. The purpose of the study was to identify specific combinations of industrial activities in which Puerto Rico had an advantage, and to estimate the magnitude of the locational advantage of such combinations. Isard and his colleagues first examined some conventional approaches in regional economics, including broad economic development methods, comparative-cost analysis for individual industries, the location quotient, the coefficient of localization, the labor coefficient, and input–output approaches. They found that these techniques of regional analysis were valid, with each being pertinent for certain regional situations but not others. However, none of these traditional economic methods was entirely satisfactory for explaining the economic growth of the region in a historical perspective. Any basic industry is always interrelated with other economic sectors; the connection grows even more intense over time; and the development in an area comes to influence the location pattern within and outside the region through direct and indirect linkages.

To address these issues, Isard, Schooler, and Vietorisz (1959) devised conventional analytical techniques to deal with various aspects of regional development. The major method of analysis was an extension of input–output or interindustry analysis called activity analysis, in which a process may produce more than one output and there may be more than one process that generates a given output. In the study of Puerto Rico, 73 production processes were described in terms of their physical outputs and inputs. The input and output levels of the 76 commodities other than labor and capital were assumed to increase in proportion to process levels, while labor and capital inputs were presumed to obey the fractional power law and thereby increase less than proportionately to process levels. Altogether, 19 preliminary industrial complexes covering operations from refinery through synthetic fiber production were selected for further analysis. Then, for each complex, total inputs by type and outputs by type were calculated. Cost and revenue differentials between Puerto Rico and alternative mainland locations were estimated for each input and output. In the next step, Puerto Rico’s preliminary locational advantages or disadvantages were
obtained for each complex through multiplying input and output levels by the appropriate price differentials and summing all inputs and outputs of the refinery processes. A region is assumed to have an advantage if the sum is positive and a disadvantage if the sum is negative. Regions should focus on those of its industrial complexes that possess locational advantages.

From a viewpoint of economic development, there are several advantages of developing an industrial complex. First, by locating in an industrial complex, individual firms may reduce transportation and transaction costs within the input–output network (Isard and Vietorisz 1955). This is in line with traditional Weberian industrial location theory which emphasizes cost minimization. Second, industrial complexes generate a large pool of trained labor and professionals. This effect is well recognized in the literature on industrial agglomeration, particularly the Marshallian-type localization economy. Another important benefit of being in an industrial complex is the potential technology spillover, resulting from dense information exchange and constant innovation among adjacent companies. These advantages are essential for understanding major industrial developments not only in the fiber industries in Puerto Rico, but also in the Ruhr manufacturing region, with its heavy orientation to steel-fabricating complexes; the Gulf Coast agglomeration, with its primary focus on hydrocarbon complexes; the American Midwest region, with its automobile complexes; and the more recent spatial formation of Silicon Valley, with its competitiveness in electronics complexes.

In the 1960s the concept of the industrial complex refreshed the fields of economic geography, regional economics, space economy, and economic planning. Its use of quantitative approaches has also advanced the study of industries and localities by filling an important gap in the relevant fields, demonstrating how the investigation of individual industries and regional economies may be supplemented and complemented with analysis of the interrelations of these industries in a region. In an era when quantitative revolution prevails in academia, these mathematical techniques that combine an input–output table with an analysis of the factors leading to regional differences in costs and revenues are particularly appealing to scholars who are seeking to make regional studies more scientific.

Following Isard’s initial publications, scholars attempted to study the spatial groupings of industries and to identify complexes that enjoy locational advantages. In general, they relied heavily on input–output tables to reveal vertical trade relationships. For instance, Czamanski (1971) used multivariate analysis on purchases and sales data in an input–output table to classify industrial complexes. Using a slightly different technique, Roepke, Adams, and Wiseman (1974) employed factor analysis to group industries for industrial complex identification. In a later study, O hUallachain (1984) evaluated the use of principal component analysis in the industrial complex studies and suggested that this technique may provide more information on the industrial structure. Hewings and Jensen’s (1986) review of regional and interregional models also placed much reliance on the input–output table to discuss the interconnectedness and interrelatedness between industries and regions.

While methodological concerns have dominated the research in the literature of industrial complexes, studies in regional economics in the 1970s and 1980s failed to apply the concept and the technique as an effective policy tool to address economic recession and to boost regional economies in a global market. In addition, the compound quantitative nature of the approach in industrial complex analysis may have deterred many regional planners who did not possess the
necessary mathematical skills to use the concept. As a result, passion among scholars started to fade, and use of the term began to decline from the 1980s. Instead, a similar but new term, “industrial district,” gradually began to gain popularity both among scholars and in public policies, when a few manufacturing regions in northern Italy exhibited remarkable resilience and growth during a time of economic stagnation and recession in many advanced economies (Pyke and Sengenberger 1992). In contrast to the concept of the industrial complex, studies on industrial districts argued that the high economic growth rates and significant improvement of social conditions in these industrial districts were causally related to the properties of Marshallian agglomeration economies, where small and medium-sized businesses specialize in particular tasks (Piore and Sabel 1984; Pyke and Sengenberger 1992). Another major difference is that, in contrast to the highly quantitative analysis in the industrial complex approach, studies of industrial districts do not place much emphasis on the input–output table but more on the historical and social context of the region. Gradually, the idea of industrial districts gained prominence among regional planners and industrial strategists, who advanced the concept as a policy tool for industrial restructuring, regional development, and economic renewal.

However, the concept of the industrial complex has not totally disappeared in more recent literature. Instead, techniques used in the industrial complex approach have seen some rejuvenation since the 1990s with the emergence of the industrial cluster model. Industrial clusters refer to the geographic concentrations of interconnected companies, specialized suppliers, service providers, and associated institutions in a particular field (Porter 1990; 1998; 2000). A major difference from earlier theories on industrial agglomeration is that the cluster model focuses on competitive advantages instead of comparative advantages, arguing that the innovative nature of competition in geographic clusters can be nurtured. This new model has gained acceptance among scholars and practitioners, and has been considered a canonical model for economic development and revitalization in recent years. An important task in the industrial cluster is to identify industries that hold competitive advantages, and some scholars resort to the techniques used in industrial complex analysis – the analysis of input–output tables – to reveal interindustry linkages and regional advantages.

In retrospect, industrial complex analysis remains an important approach in regional economic development that has general significance beyond its pathbreaking application to Puerto Rico. It outlines the importance of spatial proximity between firms, emphasizes the industrial growth of a region in a historical perspective, and advances the uses of regional and interregional input–output tables in economic growth analysis. It is an important concept in the literature of agglomeration economy, location theory, regional studies, and space economy. Its concept and methods can be used by regional economists, city and regional planners, business executives concerned with location decisions, and community and state industrial development groups.

SEE ALSO: Industrial agglomeration; Industrial districts; Industrial location theory; Input–output analysis; Localization/delocalization

References

Industrial Complexes Using Input–Output Data.”

**Further reading**


Industrial conversion

Ray Hudson
Durham University, UK

Places develop distinctive industrial structures, a response to specific combinations of circumstances at particular times. For example, the conjunction of particular raw materials and technological developments that enable them to be used in commodity production can provide the basis for mono-industrial towns. The subsequent development of economies of scope and scale in production can then lead to the creation of major industrial complexes in particular cities and regions. In response to this, people migrate in large numbers to such places in search of work, often from rural and agricultural areas that have lacked an industrial past, sometimes from places in which industry had developed but has subsequently declined and jobs disappeared. In time such new industrial spaces may form their own distinctive place-based communities, grounded materially and institutionally in the industries that form the rationale for the place.

However, a characteristic of capitalist development is that it tends to erode the conditions that originally made an industry successful, so that capital is devalorized, written off in some places and reinvested in others. As such, devalorization is always of necessity place specific (Harvey 1982), so that the decline of an industry inevitably has place-specific effects and consequences. The question to be explored here – essentially that posed by Doreen Massey (1979) as “in what sense a regional problem?” – or put slightly differently, “who is a regional problem a problem for?” – is what happens when, for whatever reason, that favorable conjunction of circumstances changes and the economic viability of those industries and places is compromised and eroded and people’s livelihoods are imperiled as employment disappears? This is fundamentally a political question: to borrow a phrase from the striking steel workers of Longwy as they and their families and friends marched through Paris in 1979 (Carney, Hudson, and Lewis 1980), do people have the right to live, learn, and work in a place even though the capitalist interests that originally invested to give rise to that place no longer have an interest in it?

Declining profitability does not necessarily mean that capital abandons a location, however. Companies may respond to growing pressures on profitability by seeking to restructure in situ, changing production processes and/or products, adopting strategies of innovation, and learning to change business models and move onto new Schumpeterian trajectories of growth. They have considerable place-specific sunk costs that outweigh any benefits in terms of cost reduction as a result of relocation – for example, because of pools of skilled labor that cannot be re-created elsewhere, ties into supplier chains that are critical to success, or brands that are strongly identified with production in a particular location. For instance, in the United Kingdom high quality ceramics production has remained and revived in “the Potteries” as companies have adopted new business models and production strategies as mass produced products became no longer economically viable there. There is a downside for workers in such strategic responses,
however, for they may – and typically do – result in considerable job losses. While they may be efficacious for capital, they impose considerable costs upon labor. Even so, trades unions and other labor organizations may acquiesce in such strategies as part of place-based cross-class alliances, defending industry in place on the basis that to retain some jobs is better than to retain none.

The alternative to in situ restructuring for capital is to disinvest from some places, possibly devalorizing capital there, and/or moving capital to invest in new locations that offer greater prospect of profit. This does not mean that capital has been destroyed and industry come to an end but rather that as capital disinvests from some places and moves production elsewhere in search of greater profits, initially intranationally, increasingly internationally, as evidenced most strikingly in the rise of China as a center of industrial production, it helps establish the material and social bases for creating new production structures in “new” places. This could involve producing existing commodities in these new locations with new or existing processes or moving into new products and sectors there.

Capitalist development is always inherently uneven, with growth in some places inextricably linked to decline in others. Important though the dynamics of that process of uneven development are, though, they are not the main concern here. The focus here is on the consequences in those places that suffer decline, upon what happens in those places that one set of capitalist interests has abandoned as their inhabitants, social groups, and political forces seek to construct a new economic rationale for them, typically by enticing new capitalist interests to invest and create employment there. What happens to those places that once were centers of industrial production – workshops of the world – but no longer are as capital is devalorized and/or moved to other parts of the world?

State policies and industrial conversion

While corporate self-interest alone may be sufficient to stimulate in situ restructuring, it often needs positive intervention by the state (at national, subnational, or in some instances, such as the European Union, supranational levels) to bring it about. State involvement can be particularly visible when the state has been involved in the destruction of the pre-existing industrial base. Industries that were once nationalized as part of social-democratic industrial strategies, such as coal and steel, became subject to, first, rationalization, capacity closure, and job loss, and then terminal decline and complete closure as part of a neoliberal turn in economic policy that began in the United Kingdom with the Thatcherite governments of the 1980s (Beynon, Hudson, and Sadler 1991; Hudson 1989) and quickly spread elsewhere (Hudson and Sadler 1983). This helped create pressures for state action to mitigate the effects of the resultant job losses, via various incentives to new industry to invest in such places – for example, in the form of branch plants or new small and medium-sized enterprises (SMEs), discussed further below.

In other circumstances, rather than seek to eliminate particular industries, state policies seek to effect a transition from “old” to “new” industries, preserving and adopting skills and production capacity from the former to the latter. Examples include the shift from coal and steel to environmental engineering in the Ruhr in the 1980s (Refeld 1995) and the transition from shipbuilding to the production of offshore electricity turbines in northern Germany in the 2000s (Fornahl et al. 2012). Such restructuring and reconversion programs often lead to job losses but are justified on the basis that they adapt existing skills and competences and create “better,” more skilled and remunerated jobs as a result. However, such restructuring in search
of profitability can also lead to deskill ing and “worse” jobs, justified on the grounds that some jobs of whatever quality are better than none.

In all these cases, however, one key effect of state involvement either to remove some industries from particular places or to preserve others is that it can lead to substantial net reductions in employment and become the proximate cause of pressures for the state to intervene in the place market to seek to attract new forms of employment and work to those places. The threat of capacity closures and job losses in existing industries has led to political pressures to entice new industries to replace them via incentives of various sorts – infrastructure investment, grants to cut the cost of fixed capital, wage subsidies or grants to retrain and upskill workers. Through such incentives public authorities seek to move to a new industrial structure, manufacturing, and/or services, involving some or all of the following: branch plants, SMEs, call and contact centers, new retail parks, or different forms of tourism (including industrial heritage). More recently there have been attempts to encourage and promote the emergence of so-called learning regions, knowledge-based economies, and smart specialization, linked to learning and knowledge transfer from universities as well as between firms and other place-based civil society organizations. These options are explored below.

Branch plant economies: from global outposts to embedded branch plants?

A common policy response to the decline of one industry or related complex of industries is to seek to attract inward investment in companies and industries that are new to that place. While this may involve the relocation of entire companies, more often it involves the attraction of branch plants that are part of wider company structures. As the scale of production has increased, the intracompany organization of production has become more complex, enabled by advances in transport and communication technologies. Companies have increasingly created intracompany spatial divisions of labor, (re)locating particular production functions in those locations most appropriate for them. This has offered opportunities for areas that previously were industrialized, but which have become deindustrialized, to seek to attract new activities, typically branch plants producing or assembling components (Massey 1984). Such plants are typically attracted by combination financial incentives and other subsidies to production costs as well as by the availability of large amounts of labor – both workers with experience of industrial work who have lost their former industrial jobs and others, often female, drawn into the wage labor force for the first time.

While such branch plant economies often contain a wider variety of industries than the original industrial economies of such places, and in that sense can be seen as more diversified, in another sense they are more homogeneous as they tend to comprise routine assembly or production functions employing predominantly unskilled or semiskilled labor. Moreover, such branch plant economies, “global outposts” located at the extremities of global chains of corporate command and control, are vulnerable, susceptible to decisions about investment and disinvestment taken in distant boardrooms. They can be vulnerable to closure, especially in circumstances in which companies have adopted multiple sourcing strategies that enable them to play off one branch plant against another. Consequently, such economies can be subject to further deindustrialization, as branch plants reach the end of their product life cycle, or are closed prior to that point because of falling
demand or because branch plants in other places are more profitable. As spatial divisions of labor have increasingly been recast from the intra-national to the international scale, amid claims of a globalizing economy and “global shifts” (Dicken 2011) to the BRICs (Brazil, Russia, India, and above all China) and the MINTs (Mexico, Indonesia, Nigeria, Turkey), branch plant economies in deindustrialized regions in the historic capitalist cores of Europe and North America have become increasingly vulnerable to devalorization and disinvestment of capital.

At the same time, however, partly enabled and necessitated by a shift to new methods of flexible high volume production, incorporating principles of “just-in-time” and “lean” production, there has been a shift from “global outposts” to “embedded” branch plants, with a mandate for limited process-related research and development (R&D) and with some parts of supply chains co-located adjacent to – or in extreme cases, actually within – assembly plants. This reliance upon localized supply chains and the pools of appropriately skilled and socialized labor and support services that develop around the assembly plants (allegedly) ties the latter more effectively to place. An iconic example of such an embedded branch plant is Nissan’s assembly plant at Sunderland in the United Kingdom, established in 1986 and now employing some 7000 workers with perhaps another 4000 employed in component suppliers in the local supply chain as an automobile rolls off the line every 60 seconds. With the recent decision to locate production of the new Juke model at Sunderland, employment could increase further. As well as this Nissan plant, there are other well-known examples from the automobile industry (Hudson and Schamp 1995). Such localized supply chains typically involve components that are bulky and difficult to transport and/or of low unit value. However, it is often unclear as to the extent to which this is more of a discursive shift rather than one reflecting a generalized shift in corporate practices, and in so far as it is a matter of practice, the extent to which it is one that is limited to a narrow range of products – for example, in consumer electronics as well as automobiles – while many branch plants remain as classic global outposts.

Enterprise cultures and new SMEs?

As a result of the changes outlined above, and recognition of both established industries and branch plants relocating and those that remained restructuring and shedding labor in an effort to remain competitive, attention came to focus upon the size of firms and plants. Over much of the capitalist world, “big” became seen as the problem, “small” as the solution to fill the space and solve problems of industrial decline and unemployment. Drawing on perceptions of the experiences of places in which development seemed to be based upon small firms (for example, parts of southern Europe: Garofoli 1992), this led to the widespread emergence of policy responses that focused on endogenous development and encouraging the formation and subsequent growth of new enterprises, often linked to ideas about the (re)birth of an entrepreneurial culture, as the route to regional economic regeneration. The reality, however, is that in old deindustrialized places such developments were more often a response to and an alternative to the reality or perceived threat of unemployment as people established one-person firms, often in various services (such as taxi driving or window cleaning) rather than in manufacturing, using redundancy payments from their previous job and/or grants from the state initially to establish their firm. The rationale of such firms is not, however, enterprise and
innovative growth but the provision of work and an income stream to the owner, self-employment as a tacit recognition of the absence of secure and well-paid waged work in the local labor market.

To some extent, the growing emphasis on small firms reflected a recognition that parts of the world that historically were not seen as industrial powerhouses nevertheless had become successful centers of industrial growth, based on collaborative networks of small firms, such as those of the “Third Italy” (Bagnasco 1975; see also Local development). Equally influential were perceptions of previously unindustrialized places that grew rapidly around clusters of small firms in new high-technology areas – with Orange County perhaps the emblematic new industrial space (Scott 1988). What policy initiatives that sought to reproduce such development in deindustrialized places largely ignored, however, were the very specific political-economic and cultural conditions in these regions that enabled them to develop successful industrial economies, conditions that could not be replicated in old industrial regions that had declined.

Typically, industrial districts such as those of southern Europe were specialized in the production of mature consumer goods, with relatively low-technology production methods but dense collaborative networks of economic and social relations between participants in the district. In time, however, these places too became vulnerable to competition from lower wage cost places, as geopolitical changes opened the economies of eastern and central Europe and China to capitalist investment and routine production functions were decentralized from established industrial districts to new and cheaper locations (Hudson 2003). Ironically, just as such models of industrial districts were being lauded as the way forward in deindustrialized areas in northern Europe, they were entering crisis and collapsing in the “model regions” of southern Europe in which they had formerly flourished (Hadjimichalis and Hudson 2014).

In other deindustrialized areas, the emphasis was placed much more upon “high technology” SMEs, often seeking to capitalize on innovation based in scientific research in local universities, and these are discussed below in the context of new knowledge-based economies. To some extent this reflected the influence of research in the United States, especially California with Orange County perhaps the emblematic “new industrial space,” which investigated the new industrial districts based on information and communications technology (ICT) and new technologies and their links to leading research universities (Scott 1988). This encouraged public authorities in very different parts of the world to seek to encourage “knowledge transfer” links from cutting-edge research in universities there in electronics, ICT, biotechnology, and advanced materials into new small firms that would both grow and form links with other firms to create vibrant new industrial districts. Typically, however, such policies were attempted in places that lacked both an entrepreneurial culture and the infrastructure (hard and soft, but especially the latter) needed for such policies to succeed. Typically, too, they lacked the level of state support that came from links to defense industries and the military-industrial complex in the United States. As a result, the effects in terms of net job creation have typically been disappointing, although in other respects there has been some limited diversification of economic structures.

From manufacturing to services?

While industrial conversion can involve seeking to replace one type of manufacturing activity with another, it often involves a shift from manufacturing to service sector activities,
reflecting broad sectoral shifts in the structure of capitalist economies and the growing recognition that manufacturing was more likely to be a sector of employment loss than employment growth. The service sector is, however, loosely conceptualized and embraces a plethora of different types of activities. As might be expected, given the diversity of activities within the categorization of “services,” the following discussion will of necessity be selective. First, consider the attempts to attract back office functions such as call and contact centers to replace jobs lost in manufacturing and/or mining. This approach has been followed in many deindustrialized places, such as northeast England, parts of the former German Democratic Republic, and in Cape Breton in Canada. Such activities do have the capacity to bring relatively large numbers of new jobs to places which are desperate to attract employment, albeit often part-time and poorly paid and requiring different types of skills and competences to those found in previously industrialized places. They often seek to recruit female workers, rather than men who have lost their jobs in manufacturing, as one element in a changing gender division of labor there – men who in any case may be resistant to such forms of work. In many ways these are the service sector equivalents of the old manufacturing branch plants, the latest variant on the theme of “global outpost,” vulnerable to closure as firms both outsource and offshore such activities, to locations such as India, with a resultant loss of employment. A second strategy in many deindustrialized places is to seek to commodify their past industrial glories and heritage and turn to industrial heritage as a basis for tourism. In the process, such places undergo a transformation from places of production to places of nostalgia – no longer places of material production but of an often rose-tinted nostalgia for an industrial past. Examples include Big Pit in south Wales and Beamish in northeast England, but there are many more. This is but one facet of a broader turn towards tourism of various sorts in the search for new bases of economic development in deindustrialized places (Hudson and Townsend 1992).

Third, there has been an emphasis on retailing and consumption as bases for economic regeneration, but of course, like tourism, employment in such activities is dependent upon people in other places having jobs and enough money to spend. However, the growth of out-of-town shopping centers and retail parks in many former industrial areas attests to the fact that this has been a common element in many attempts to regenerate these areas. Moreover – as with heritage tourism – this has increasingly involved a turn to the “experience economy,” in which the rationale is not simply purchasing and then consuming commodities but the experience of purchasing itself, whether of material goods, or a holiday, or entry to a museum and, as a result, increasingly creating or tailoring built environments to that end (Lorentzen and Jeannerat 2013).

While involving very different sorts of activities, these service sector jobs have certain characteristics in common. They tend to be linked to significant shifts in gender divisions of paid labor, to be poorly paid, to be regarded as unskilled, and often to be part-time or casual, often seasonal. Thus they offer very different forms of paid work to those formerly found in now deindustrialized places and provide a rather fragile and precarious base around which to seek to reconstruct a local or regional economy.

Knowledge and learning regions: old wine in new bottles

To some extent the growing interest in place-specific conditions as critical to suc-
cess cuts across the manufacturing/services dichotomy, recognizing that many functions once based in manufacturing companies have now been outsourced to separate service sector companies, that product and not just process innovation has become increasingly important in relation to competitiveness, and that increasingly the knowledge and symbolic content of many commodities is of growing significance. From the late 1980s there was a growing recognition that industrial success and the sources of competitiveness lie not simply within the boundaries of the firm but also in the relationships between firms, public authorities, community groups, and diverse organizations in civil society in a specific territorial domain, recalling the nineteenth-century economist Alfred Marshall’s concept of industrial districts and the importance of “something in the air” to the economies of successful places. In short, this was a recognition that proximity and links to others and “being there” (Gertler 1995) matter to the success of individual firms and so to regional economies. This led to new ways of thinking about how to regenerate regional economies that had experienced severe economic decline and introduce new areas of growth to them.

As a result, there was growing interest in and emphasis on regional innovations systems (Braczyk, Cooke, and Heidenreich 1998), on knowledge spillovers and the significance of embedded and often embodied tacit as opposed to codified knowledge, on pools of appropriately skilled labor and input–output relations between firms, on the social relations of civil society beyond the workplace, and on the complementarities of activities between firms in an intraregional territorial division of labor that combined both “traded and untraded interdependencies” (Storper 1995). In response to this body of research, public policies in many parts of the world sought to create the conditions in which such knowledge-based economies and learning regions could emerge and flourish, although others were more skeptical as to what was new in these approaches and how successful they had been and can be in regenerating deindustrialized places (Hudson 1999, 2011).

A later variant on a related theme is the emergent interest in “smart specialization” as a strategy to build on key competences that remain in a region but to adapt these to the production of new commodities for which there are perceived to be growth opportunities and in this way to convert and reconstruct regional economies. Smart specialization has been defined as a strategic approach to economic development through targeted support to research and innovation, involving a process of developing a vision, identifying competitive advantage, setting strategic priorities, and making use of smart policies to maximize the knowledge-based development potential of any region, strong or weak, high-tech or low-tech (European Commission n.d.). Related to the emphasis on knowledge and learning, there has been growing interest in creativity and culture as a route to reconstructing the economies of cities and regions (Cooke and Lazzeretti 2008). Whether such approaches will be any more successful than earlier ones remains to be seen as they require quite precise social and material conditions to have a chance of success. Finally, we cannot eth that there is also recognition that the authenticity of products in some cases crucially depends upon where they are produced and that this identification of product with place offers an opportunity to reconstruct regional economies by moving up the value chain. One example of this is the resurgence of the mechanical Swiss watchmaking industry and the capacity of such watches to command premium prices precisely because they are known to be produced in Switzerland, indeed in particular parts of
Industrial Conversion

Switzerland, with the result that both firms and regional economies have prospered (Jeannerat 2013). Authenticity is not, however, something that can be transplanted to other places and so appeals to authenticity as a strategy for regional economic revival are strictly limited.

Beyond the formal mainstream: social enterprises, informality, and illegality

With the increasing difficulties of providing replacement jobs via any form of mainstream private sector development, in many areas that suffered severe industrial decline there has been growing interest in a range of social economy alternatives, such as social enterprises, community businesses, and cooperatives. This is a very different approach to industrial conversion, concerned at least as much with the social relations of the economy as with products and processes. While diverse, these approaches share a common characteristic in not being driven by concerns to maximize profit and shareholder value (Gibson-Graham, Cameron, and Healy 2011; Hudson 2009). Social enterprises and cooperatives do seek to generate a surplus but are committed to using this for collective social and economic development purposes rather than private gain. Other forms of social economy organizations do not seek to generate a monetary surplus but rather focus on providing services that help lead to environmental and health improvements or training to enable those who have been unemployed, especially the long-term unemployed, to acquire new skills as a first step onto the ladder back into the formal labor market. Often such initiatives have involved learning from experiences in peripheral parts of the global economy in the Global South (Amin 2009).

In some places that have suffered the effects of deindustrialization and economic decline, and in which policies of reconversion and regeneration fail and state support for those without work is meagre, there may be industrial conversion of a very different sort as the void in the economy may be filled by illegal activities of various sorts as people seek to construct individual and household strategies to “get by” and survive. This can involve activities that are legal but pursued by people who are not authorized to perform them (often referred to as activities in the “informal” sector). It can also involve activities that are simply illegal. This could encompass both the production of things that are illegal or more typically revolve around activities that involve trade and trafficking in a variety of goods, both those that are illegal and those that are legal but illegally acquired, often the proceeds of criminal activities, but also often linked into the formal legal economy (Hudson 2014).

Conclusions

This entry has summarized some of the main ways in which, via state policies and the actions of various civil society organizations, attempts have been made to address the problems that result as one set of industries decline in a place. This can involve restructuring those industries in situ, attracting inward investment, or developing a variety of endogenous local alternatives. The form and content of these responses has changed over time, though there is no simple linear sequence of policy responses that is universally applicable. Moreover, what has been successful in one place and time is by no means guaranteed to be so in other times and places. Given the character of capitalist economies, with the restless movement of capital in search of profits, industrial change and conversion is an ongoing process for many places. Some achieve this relatively unproblematically – for example, the
considerable loss of manufacturing employment in London was offset by a compensating growth in banking, financial services, and associated service sector activities, albeit at the cost of much more deeply segmented labor markets and income inequality (Buck et al. 2002).

For most other places, however, the process of policy-steered industrial conversion is much more challenging and difficult, not least because of structural and spatial change in the economy itself. Often these places have been the site of a succession of policies seeking to attract manufacturing branch plants and service sector back offices, encourage the formation and growth of indigenous small enterprises across the whole spectrum of the economy, and foster the development of a tourism economy based around industrial heritage. Typically, too, such policies have all failed to have their intended effects in terms of the emergence of a new economy and new sources of formal sector employment. As a result, such places remain blighted by unemployment, worklessness, and poverty and so may become places in which the process of conversion is a more radical one, from legal to illegal economy.

SEE ALSO: Branch plant economy; Brands and branding; Deindustrialization; Industrial districts; Industrial restructuring; Local development; New industrial space; Political economy and regional development; Regional development policies; Regional unemployment and regional labor markets; Uneven regional development

References


Further reading


Industrial districts

Liang-Chih Chen
Shiuh-Shen Chien
National Taiwan University

An industrial district (ID) is a spatially defined production system characterized by the concentration of firms specializing in certain industrial sectors. The emergence and development of such a system is emphasized to be deeply linked to the local contexts and social relations of the area which it embeds. The concept of IDs originates from Alfred Marshall's work (1961/1890) identifying the external economies associated with agglomerated small- and medium-sized enterprises (SMEs) in regions like Lancashire and Sheffield in England during the nineteenth century. In the 1970s and 1980s, the term “industrial district” was coined by a group of scholars to describe SME-based regional economies, notably the Third Italy (i.e., the central and northeastern regions), that demonstrated higher competitive and economic resilience. Being able to secure productivity and competitiveness advantages through the creation of external economies of scale and by promoting flexible specialization for the agglomerated SMEs, IDs were further considered as an alternative model to the industrial organization of large firms. Examples of well-known IDs include ones such as Emilia-Romagna in Italy, Silicon Valley in the United States, Baden-Württemberg in Germany, Ishikawa in Japan, and Hsinchu in Taiwan.

Two general features of IDs are often stressed (Rinaldi 2005). The first emphasis focuses on the industrial structure in which the production system is composed mainly of small, independent firms which are geographically agglomerated and are specialized in one or a few phases of a complete production process. As these ID firms often have a low degree of vertical integration and only a small portion of them produce finished products, the production process has to be carried out through the collaboration of local firms tied together by subcontracting arrangements.

The second key feature of IDs focuses on its governance mechanism. As opposed to the traditional agglomeration theory that conceptualizes local economies as collections of atomistic competitors regulated mainly by price/cost signals, the ID concept emphasizes the greater reliance of firms on locally embedded social relationships in governing their interfirm transaction (Harrison 1992). The institutional environments of ID have produced an atmosphere of trustful behavior, allowing for the emergence of cooperative competition and interactive learning among ID firms.

Factors accounting for ID dynamism

Aside from the typical notion of agglomeration economies, the ID concept specifically addresses key industrial and social settings that facilitate the economic vibrancy of such a territorial industrial system. First, from the perspective of industrial organization, owing to spatial concentration, ID firms are able to build up networked production systems, giving rise to an extensive division of labor between local firms, allowing these individual firms and the
industry as a whole to increase their competitiveness through flexible specialization in production. In addition, the ID environment promotes the development of SMEs as smaller firms can rely on external sources to grow rather than engaging in expensive vertical integration. Expanded entrepreneurial opportunities then induce a constant influx of entrepreneurs who not only sustain the dynamism of the production systems, but also inject innovative ideas and vitality.

Second, one should understand that the most important advantage of ID is the development of local economic-social communities based on non-economic factors such as local cultures, norms, and institutions. In such a kind of community, people share common culture, values, or beliefs, and so on, which help to lubricate social relations, build consensus, support group loyalty between entrepreneurs and employees, and provide a common language to speed innovation and information exchange (Rinaldi 2005). The prevailing social and economic interactions further nurture the development or reinforcement of mutual trust (or social capital) between ID firms, allowing them to develop reciprocal relationships and engage in joint action, thereby enabling the smooth and sustained operation of the ID’s production networks.

Third, the geographical and institutional settings of IDs are particularly instrumental to promoting innovation. The industrial atmosphere of IDs can enhance the potential of firms for acquiring knowledge and skills to support the development, adoption, and diffusion of innovation. On the one hand, the existence of close social relations or high degrees of trust among ID firms not only facilitates the broad and fast circulation of information and knowledge, but also fosters the formation of innovative cooperation. On the other hand, geographical proximity also helps to provide good opportunities for frequent interpersonal face-to-face contact to ensure efficient exchanges of tacit knowledge and interactive learning among local actors. In addition to the abovementioned institutions, which are often informal in nature, one should note that the interaction and cooperation between ID firms in terms of production and innovation might also be encouraged by the existence of formal support institutions, such as industrial associations or research institutes, whose establishments have often been the results of public and private joint efforts.

**The changing characteristics of IDs**

The current global economy is characterized by the inevitable and intensified interconnectedness of different localities, including IDs. In this regard, the conventional ID model that explains a district’s global competitiveness by looking solely at the structure and dynamics of its local industrial system has been unable to provide a satisfactory explanation in regards to the development and evolution of IDs. On the one hand, economic actors within the IDs might require closer and more formalized collaboration allowing them to enhance or sustain competitive edge vis-à-vis those foreign larger firms in the world market. On the other hand, it is also observed that ID firms actively acquire new knowledge or capabilities from multiple geographical sources in addition to those situated within their IDs. As a result of these trends some prototypical characteristics of IDs have undergone notable changes. Two phenomena – the greater external openness of the ID industrial systems and the rise of large firms, particularly business groups within local production networks, have been addressed in literature concerning the recent evolution of IDs.
The increasing openness of IDs

To understand the current dynamics of IDs, the self-contained character of industrial organizations of IDs should not be overemphasized. Recent studies have noted a trend of increasing openness of IDs in the sense that districts’ production systems have not only been penetrated by external actors, but have also extended their spatial boundaries to other localities within or outside their home countries. On the one hand, seeking opportunities to exploit the advantages provided by IDs, some outside firms, including multinational corporations, have been attracted to invest in districts and consequently have emerged as important members of these districts’ production communities. On the other hand, some larger ID firms have expanded to a transnational and global scale with the capacity and motivation to establish extensive production and business networks beyond their local territory. Moreover, facing increased competition in the global market, other ID firms have sought to subcontract to extra-district suppliers so as to obtain specialized or cheaper supplies. Studies of Italian IDs, for instance, have pointed out that the majority of subcontracting work now relies on extra-district suppliers. The destination of outsourcing has not only been these Italian IDs’ neighboring regions in Central and Eastern Europe, but also some distant Asian countries, such as China.

The opening up of IDs is generally viewed as a critical development for sustaining the vitality of IDs by allowing previously closed systems to access external capability and knowledge and thus avoiding lock-in effects. Acknowledging the fact that ID firms have significantly increased the volume of their operations outside their home region and the IDs’ territorialized production systems seem to be undergoing a de-location process, one might wonder if such developments would impair interfirm production and social relationships within the districts and whether the characteristic local embeddedness of the ID model may become obsolete. However, some current ID studies have shown that, while the opening of IDs has substantially influenced the nature of interaction among district firms, it has not necessarily led to the complete abandonment of local relationships. For instance, ID firms are observed to rely mainly on extra-district suppliers only for standardized production activities and still work closely with their local partners on more important tasks such as product design and development. Scholars also maintained that the development of IDs’ local and nonlocal relationships might actually be complementary. Although the spatial boundaries of production networks in the IDs are indeed not confined to the locality, ID firms depend greatly on the districts’ pre-existing locally rooted relationships to manage their non-local networks better. To be sure, many of these critical nonlocal networks might actually fail to emerge or develop without the production advantages derived from the IDs’ industrial networks. Furthermore, it is also due to the existence of knowledge sharing and learning mechanisms within the IDs, one firm’s improved competence, as a result of their internationalization, would be diffused to others, leading to a synchronized improvement of all ID firms.

The emergence of business groups within IDs

In the recent evolution of organizational structures of IDs, there has been a strong tendency toward an increase in firms’ concentration. More specifically, the escalation of leader firms with the rise of business groups established by the former has become an emerging trend in the growth process of IDs. Owing to their business success, a handful of local SMEs have expanded in size and, consequently, gained predominance
in affecting the changes of the districts. In addition, some of these leading firms have sought to reorganize themselves into business groups, that is, a set of firms which are legally independent but are established under the same control and ownership.

As suggested by some scholars, the establishment of business groups is especially suited to facilitate the growth process of SMEs. The emergence of such groups in IDs have further received wide discussions since it represents one specific organizational form adopted by ID firms to cope with new conditions of market and technology (Cainelli and Iacobucci 2007). Studies on Italian IDs, for example, have discovered that, rather than increasing the internal sizes of their original firms, district firms tend to acquire other local firms to form business groups in response to the need to expand in size, as related research have found that business groups tend to be more common in IDs than in non-district areas.

Generally, the increasing relevance of business groups in IDs is seen as resulting from a process of concentrating ID outputs and the ability of district firms to achieve better governance over the often loosely coupled subcontracting relations within their networks. It is both the defensive and aggressive strategies employed by leading firms that stimulate the formation of business groups. To protect core knowledge from leaking out to their neighboring counterparts through the district's embedded social and production networks, leading firms would buy out subcontractors or competitors so as to enclose and control externalities within the same corporate structure. Besides, with their superior access to markets, information, knowledge, and finance, some leading firms would seek to increase their dominance over their districts through promoting industrial consolidation.

In addition, the specific characteristics of ID firm executives, and the unique production and institutional environments in the districts are also stressed as critical factors for stimulating the emergence of business groups as the preferred modes of organization and governance.

First, ID firms are renowned for obtaining competitive advantage through specialization, and they might wish to extend such practices to their new ventures. By establishing business groups, the leading firms control overall strategy and achieve economies of scale in some activities (e.g., marketing, finance, R&D, etc.), while simultaneously maintaining an organizational structure in which each group firm specializes in specific products or markets. Such arrangements also offer contractual autonomy to firms in the group, since each firm has the autonomy needed to adapt its contractual relationships with customers and suppliers to its specific and changing needs.

Second, the stronger entrepreneurial culture found in IDs also influences the development of business groups, particularly the internal dynamics within the formation process. Business groups enable each legal unit to retain higher degrees of autonomy thus preserving the entrepreneurial style of management that characterizes many ID firms. Moreover, the formation of business groups could be the result of the entrepreneurial reproduction process within the IDs. In some cases, the setting up of a new firm within the group might aim to provide entrepreneurial opportunities for their more talented and ambitious people from the leading firms while keeping that talent within the existing group.

Third, the geographical and social proximities of ID firms also help to nurture the creation of business groups by facilitating the matching processes of consolidating firms. Collocation enables the rapid circulation and sharing of information among local firms, allowing those who are to form business groups to efficiently learn and
screen suitable candidates for acquisition. By exploiting their social networks within the IDs, these entrepreneurial firms further gain additional knowledge about potential buyout targets, providing them with additional confidence in their expansion strategies.

Future of IDs

IDs have continued to be considered as critical engines of regional, national, or even world economies. The competitiveness of IDs has still been argued to be largely derived from the efficiency and innovativeness of their local production networks. Nevertheless, such perspectives have been challenged by scholars who warned that an ID’s local networks that are too closed and exclusive might actually harm its sustained development due to over-embeddedness. Furthermore, it is also noted that the development of information and communication technologies and advanced transport has greatly relaxed the spatial constraints of distant relationship building and knowledge exchanges. As a result, spatial co-location might therefore become less relevant or even disadvantageous for ID firms’ attempts to maintain or advance their competitiveness, consequently leading to these firms’ dis-embeddedness from their districts. However, in our discussions regarding the changing characteristics of IDs, it is noted that local embeddedness still plays an essential role in stimulating the restructuring and reconfiguration of industrial organizations of IDs. While other factors, such as firm capabilities or external shocks, may indeed affect the evolutionary processes and outcomes of a particular region, it is still believed that the geographical concentration of actors within IDs remains a preferable arrangement to support individual firms and the industry as a whole in coping with volatile business environments.

SEE ALSO: Flexible specialization; Growth poles and growth centers; Industrial agglomeration; Institutions and development; Networks, social capital, and development; New industrial space; Regional development models; Regional development policies; Relational assets

References


Further reading


INDUSTRIAL DISTRICTS


Industrial geography

Canfei He
Peking University, China

Shengjun Zhu
Swansea University, UK

Industrial geography has traditionally been defined as the study of the spaces, places, and geographical circulation of industry. It is a sub-branch of economic geography and deals with the spatial arrangement of manufacturing or secondary activities. Due to the transformations in the production system and the emergence of new kinds of output, organization forms, distributive functions, and new modes of competition, it is increasingly more difficult and less meaningful to distinguish clearly between manufacturing and other forms of economic activities. It has been widely acknowledged that, in order to thoroughly understand spatial arrangement of industrial activities, attention should be paid to the assemblage of a wide range of factors, in particular the broader societal, economic, political, institutional, and technological context in which industrial activities situate. In this sense, industrial geography encompasses not only the location of manufacturing or secondary activities but also the broader processes of structural change in the production of goods and services and various ways in which space, place, and geographical circulation are part of these processes.

The discipline was established in the early twentieth century in both Western Europe and the United States, driven partly by theoretical arguments developed in economics, and remains an important subfield within Anglo-American geography. It has been energized and replenished in a decentralized manner by researchers asking different questions within the same subdisciplinary space. Change has been incessant, and the history of industrial geography has been punctuated by numerous moments of disciplinary reinvention and reconfiguration. Industrial geographers thus never settle on one single methodology, theoretical framework, set of techniques, or perspective. For instance, a persistent theme in industrial geography has been examinations of the spatial distributions of industrial sectors, based on a perspective of industries. Industries are composed of companies, which make location decisions to open and close facilities of different production capabilities. These create the “geography of enterprise” and a firm perspective. Complementing the systematic approaches by sector and by company is the regional perspective. Another reason why industrial geography has been subject to so much change is that the discipline has been highly empirically grounded and closely tied to its empirical object of study, which over time has constantly undergone fundamental alteration. This is especially true, for example, at the end of the twentieth century and the beginning of the twenty-first century, resulting from technological advances in terms of transportation and communication, on the one hand, and a process of globalization and marketization in countries and regions that had not been fully exposed to market economy before, on the other hand (e.g., China, India, and former communist countries in central and eastern Europe).

This entry is organized along a rough timeline. The next section seeks to review briefly the main phases of the discipline's historical development.
from the early twentieth century when industrial geography was institutionalized in both Western Europe and the United States to its most recent situation. The third section outlines the contemporary research in industrial geography by listing some of the primary groups of theoretical and empirical studies. The last section concludes this entry and discusses the future direction of industrial geography.

Historical development of industrial geography

Before the formal institutionalization of the discipline, commercial geography – the precursor of economic geography – which was formally established in 1882 by German geographer Götz, had conducted some rudimentary industrial geography studies and researched the causes and reasons of economic geographical specialization, production of commodities, and movements of goods. However, commercial geography had been less an academic discipline than a tool serving the European imperial project and providing practical knowledge and information about where places were located and the specific commodities they produced, to colonist states (Barnes 2009). Works on industrial geography conducted in this period, such as George Chisholm’s Handbook of Commercial Geography, were mainly written in a commercial geography style, filled with maps, tables, and industrial geographical facts. Chisholm (1889) argued in his Handbook that the environment made each place uniquely suitable to undertake a specific type of industrial activity.

Location theory and spatial science

While pre-war geographers focused on global commodity production and trade, with environmental determinist and colonialist overtones, other competing academic disciplines had proceeded with conceptual and statistical advances. Specifically, a number of social sciences and even some humanities, especially in the United States, shifted towards an approach that stressed rigorous statistical techniques of description and analysis, and turned on the widespread use of scientific methods and philosophy. This gave rise to a new generation of geographers with a shared view of developing a thorough, mathematically justifiable geographical science that could be taken seriously by other competing academic disciplines. The subsequent shift to spatial science, also known as the “quantitative revolution” fundamentally altered the discipline, emphasizing the systematic application of scientific forms of generalization, inferential statistical methods, and abstract modeling – often inspired by other disciplines such as physics, mathematics, sociology, and particularly economics.

As the quantitative revolution proceeded, industrial geography started to massively import theories and methods from economics, especially those on industrial location patterns. The birth of modern industrial location theory dated back to 1909, when the German economist and location theorist Alfred Weber published his seminal book, Theory of the Location of Industries. This formed the basis for variable cost analysis, which was further extended significantly by Tord Palander and Edgar M. Hoover. At the same time, development of theory on locational interdependence provided the foundations for the variable revenue analysis. In much of the 1960s, the variable cost analysis dominated in industrial location research, whereas the variable revenue approach, mainly in market area analysis, found its most obvious expression in the context of central place theory. The fusion of these two approaches had proved to be very difficult theoretically, and the most influential attempts in this respect were those made by
August Lösch, Melvin L. Greenhut, and Walter Isard.

However, theoretical contributions by geographers at this stage were limited. In 1910, Chisholm introduced Weber’s work into geography, to interpret “the seats of industry.” Throughout the late 1950s and most of the 1960s, industrial geographers embraced the work of Weber and other members of the German Location School (e.g., Johann Heinrich von Thünen, August Lösch, Walter Christaller, etc.), and predicated the interpretation of industrial location patterns on a type of neoclassical economic theory of rational choice and maximizing behavior. This theoretical and conceptual shift resulted from a combination of the quantitative revolution unfolded not only in geography but also throughout other academic disciplines, which paid specific attention to statistical techniques and scientific abstraction, and the influence of regional science that applied neoclassical economic theories to the measurement of industrial location patterns and to the identification of optimal location of industry. Common themes at this stage included the locational decision-making behavior and distribution of industrial activities, geographical diffusion over time, and the geography of land rent in cities and its impact over the distribution of industrial activities.

The behavioral approach

As neoclassical formalism and traditional location theory were increasingly criticized during the ascendancy of the behavioral approach, geographical contributions became more prominent by the end of the 1960s. The transformation from abstract modeling and scientific generalization based on assumptions of rational choice and maximizing behavior, to observation and analysis of actual location practice accorded with the traditionally more empirical predilections of the geographers. Because the data demands of variable cost and variable revenue models could not be easily met, it was often difficult to apply location theory to spatial economic analysis. Formal location theory was also inadequate to totally interpret real-world locations of industries for some other reasons. For instance, it was unable to deal appropriately with spatial disequilibria and to incorporate the wide range of factors, other than economic ones, which played significant roles in shaping locations of industrial activities. Rather, the empirical identification of locational decision-making was much easier to undertake. The behavioral approach thus relies on in-depth interviews with corporate locators, ranging from asking about the opening or closing of a specific plant to considerations of the working of the local network and of the influence of corporate culture on the structure of the firm.

The 1970s had witnessed two distinct but interconnected themes emerging in the behavioral industrial geography. One focused on the location decision-making process and stressed suboptimal outcomes rather than maximization, and the other, also known as “geography of enterprise” or “corporate geography,” emphasized the role of industrial organization in the location decision-making process and in the spatial organization of industrial activities in general. Unlike neoclassical regional science that often treated firms as undifferentiated and rational decision-makers and privileged external spatial patterns over internal corporate structures, behavioral industrial geography thus considered the firm as a complex entity, like an organism, that evolved, adapted and struggled in a larger competitive setting, was often not rational, and whose inner workings defied formalization. Studies predicated on the behavioral approach gradually broadened, from preoccupation with the location of single-plant firms, through the
INDUSTRIAL GEOGRAPHY

complexities of giant multiplant, multisite, and multiproduct firms, to concern with entire spatial industrial systems.

Marxism, the radical approach, and regulation theory

From the late 1970s to the end of the 1980s, developments in industrial geography reflected the applications of Marxian economics to location issues, in lieu of the neoclassical economics. Dissatisfaction with the abstract, socially irrelevant and apolitical conceptualization of industrial geography that neoclassical location theory provided surfaced in the 1970s when David Harvey started to reject spatial science that he had earlier championed in his Explanation in Geography. The intellectual shift towards the radical approach, colored especially by Marxism, accorded with shifts in the real world. The restructuring of industrial activities, as well as deindustrialization, especially in the seemingly entrenched Western industrial regions attracted increasing attention, within the broader context of change in the capitalist economy. The traditional industries of Western countries, such as textiles, iron, and steel, were subject to severe economic trauma as firms shut down, moved offshore or turned lean and mean, resulting in massive manufacturing job losses. The radical approach and political economy, with its emphasis on economic crisis and uneven development, thus offered an ideal theoretical framework. In the United Kingdom, Doreen Massey’s watershed book, Spatial Divisions of Labour, offered original theoretical interpretation and trenchant empirical analysis. Her work was influenced by political economy as well as a new methodological approach, critical realism, codified and circulated primarily by the economic geographer Andrew Sayer, which linked industrial geography methodologically with intensive, on-the-ground case studies of specific industries and their geographical sites.

A group of radical geographers in California also conducted significant theoretical and empirical research, not on cases of decline in capitalism, but on the flip side of capitalism: its innovativeness, creativity, and power of revival. During the 1980s, drawing on both political economy and institutional economics, the California School, including Anna Lee Saxenian, Allen Scott, Michael Storper, and Richard Walker, carried out a series of theoretical and empirical studies based primarily in California, especially on the vibrant high-tech, garment, and film industries. They emphasized capitalism’s vitality and capability to reorganize its industrial geography and recuperate, though the United States was also suffering severe economic decline, manufacturing job loss, and deindustrialization, especially in its Midwest, Rust Belt region.

The underlying clash between the UK and US versions was reconciled in the late 1980s through regulation theory, which set the intellectual agenda for industrial geographical work for the next decade. Developed by the French economists Alain Lipietz and Robert Boyer to explain why capitalism survived in spite of Marx’s best prediction of its demise, regulation theory argued that capitalism endured because historically an appropriate conjunction formed between the regime of accumulation and the mode of production. In this sense, UK and US versions were just two sides of the same coin, because while UK industrial geographers were documenting the disintegration of an older Fordist regime of accumulation (based on mass production and mass consumption) and mode of regulation (Keynesian welfare state), their US counterparts were examining the formation of a new regime of accumulation (post-Fordism or flexible production) and a new mode of regulation (neoliberal, deregulated state).
Industrial district, embeddedness, and cultural turn

In analyzing the geographical character of post-Fordism and flexible production during the 1990s, industrial geographers rediscovered the idea of industrial district that underscored the predilection of firms in the same sector to cluster and co-locate geographically, and to become closely tied and interconnected. This idea was initially formulated by the English economist Alfred Marshall, and significantly extended by Michael Piore and Charles Sabel in their influential book *The Second Industrial Divide*, which contended that industrial district was central in the shift from mass production to flexible production. As post-Fordism and flexible production had come to characterize the economy, small firms reemerged as a focus of interest during the 1990s. Increasing attention was also directed towards networks of information exchange, which might have been partly replacing traditional concerns such as input costs as factors influencing the location of industrial activities. Economic and industrial activities remained central, but softened by recognition of the social institutions, and even the culture, in which they were embedded. The idea of industrial district and flexible production was quickly related to “embeddedness,” a concept associated with the economic anthropologist Karl Polanyi and sociologist Mark Granovetter, implying an inseparable relationship between the economic and the sociocultural. Proponents argued that it was this inseparability that characterized and contributed to the rise of post-Fordist industrial districts such as the Third Italy, Silicon Valley, and Baden-Württemberg, which flourished because firms in these industrial districts were tightly embedded in the cultural institutions, social relations, and forms of life cultivated over time in those places.

Increasing emphasis on industrial district and embeddedness in industrial geography also reflected the intellectual shift in the larger discipline of economic geography or even human geography, where “cultural turn” gradually became keywords. Rather than treating the economy as a single inviolable object, geographers needed to reject the idea of a monolithic, purified economy and to conceive it as a culturally and socially constructed product, fragile, performed, and capable of realization in various forms. Similarly, “relational turn” drew attention to the ways in which sociospatial relations of economic actors are intertwined with processes of economic change at various geographical scales; “institutional turn” stressed the role of institutions in influencing firm’s behavior, particularly in relation to inter-firm networking and industrial relations. These “turns” represented an attempt to rethink the very nature of the discipline, its empiricist epistemology, its entrenched masculinism, and its economistic logic, and to use case studies of particular substantive topics to work through a cultural, institutional, and relational methodology. These turns are far from dominant, however, and criticism quickly mounted. For instance, some argued that these turns might result in a loss of important knowledge within industrial geography, such as literature on branch plant location decisions, because this type of research was perceived as outdated. The most striking attribute of the history of industrial geography is the absence of overall disciplinary progress, and there is no standard paradigm within the field. Instead of new paradigms emerging and completely erasing the old ones, industrial geography is more a multilayered field where past versions coexist with the new ones, and old approaches continue contributing to the discipline side by side with the emerging ones. For instance, despite these
and other contemporary developments, neo-classical location theory in the space economics tradition still has its adherents, and the traditional models still have some utility as both predictive and prescriptive devices in a competitive capitalist economy. While there is no single leading approach, the discipline may look messy but vibrant and lively as a variety of approaches complement and/or conflict with each other.

Contemporary research in industrial geography

Globalization

An interest in globalization has recently defined the field, with research on both new geographical patterns of production and new forms of industrial organization that make the former possible. After a long period of being overlooked by industrial geographers as the discipline became the study of only Western industrial economies, globalization regained its position in the 1970s. The emergence of a new international division of labor, the rise of multinational corporations (MNCs), the development of new forms of transportation and communication technology, and the growth of international financial capital and foreign direct investment (FDI) reminded industrial geographers that they needed to have a more holistic geographical view and to take into account the whole world. Furthermore, it forced industrial geographers to realize that exclusively focusing on the West and separating it from the rest of the world was not just problematic but obstructive, as places around the globe were increasingly connected by globalization. Peter Dicken’s *Global Shift* was one of the first to focus on this point through a study of FDI, which was mainly carried out by large MNCs around the world. Partially driven by the increasing size, number, and dominance of MNCs and FDI was the drastic industrialization in the Global South, including not only the so-called newly industrialized economies in East and Southeast Asia, but also the emerging economies such as China, India, Russia, and Brazil – the reverse side of deindustrialization in the Global North. A novel conceptual vocabulary and theoretical framework are being forged, as well as new methods, such as global networks, global commodity chain analysis, and the analytics of MNCs. All have contributed to studies on spatial patterns of FDI, multinational corporations and their temporal changes, interregional shift in production, and subsequent rise and fall of industrial regions around the world.

On the other hand, Dicken critiqued the idea championed by economists in some instances that globalization represented “the end of geography,” and he emphasized the process of globalization was not seamless, eliminating all geographical differences. Geography was still of central importance but in a different form, and geographical differentiations were preconditions for globalization. Similar conclusions have been reached in subsequent work in industrial geography over the last decade. Geography matters fundamentally in the process of globalization and it determines the spatial patterns of FDI, MNCs and their temporal evolution, the ways in which firms are managed and the spatial organizations they forge, interregional shift in production, how global commodity chains connect distant places to one another, and industrialization and deindustrialization around the globe. Recent works on reindustrialization in the Global North, as well as the emergence of new considerations in global sourcing, such as the importance of short-cycle replenishment, lean retailing, or just-in-time production, have added further stickiness to the flow of global financial capital and investment, reinforcing the importance of geography and spatial proximity. In this sense,
Globalization is thus not smooth and homogeneous, flattening geographical distinctions; rather it is uneven, occurring in very particular places and strongly determined by spatial variation, differentiation, and diversification (Barnes 2009).

**Firms, networks, institutions, and poststructuralism**

The focus on firms in industrial geography reflects a deep-seated belief that “the firm” is a fundamental category to define the arbitrary boundary of industrial geography. The firm became central in industrial geography when location theory was introduced into industrial geography to interpret “the seats of industry.” Each new generation of industrial geographers has thereafter provided its own interpretation of the firm. In the 1930s, Michael Wise focused on small firms and the idea of “industrial district,” formulated by Alfred Marshall, which stressed the propensity of firms to co-locate and cluster geographically. The 1960s and 1970s witnessed the emergence of “geography of enterprise” or “corporate geography” in the behavioral industrial geography that was preoccupied with complexities of giant multiplant, multisite, and multiproduction firms. In the 1980s and 1990s, small firms and innovative firms regained their position and became the key focus of empirical research for many industrial geographers. These firms were deemed dynamic and able to make crucial contributions to regional development. However, the most recent studies have pointed out that the key is not the size of the firm per se, but rather the ways in which firms of different sizes cluster together in the form of a network, and the social, technical, and economic relations that make the clustering possible. There can be many types of networks, such as business networks to supplier commodity chains, production networks, and innovative networks, and they function as both a governance structure and a process of socialization through which disparate actors and organizations are interconnected in a coherent manner for mutual benefits and synergies (Yeung 2000).

In addition to the analysis of firms and networks in recent industrial geography, attention has also been drawn towards the relationship between institutions and firms, exemplified by key concepts such as “institutional thickness” (Amin and Thrift 1994). On the one hand, strong institutional presence may help to embed relatively footloose industrial firms in specific localities and regions, and to reduce their propensity for relocation. Strong institutional presence may also reinforce firms’ growth potential and innovative capacities, induce new firm formation, and enhance the competitiveness of existing firms. On the other hand, strong institutional thickness and dense network may be a positive hindrance to development and growth if they are anti-growth (Yeung 2000).

Paralleling the work on firms, and drawing partly on cultural turn and poststructuralism, was allied research on firms and the sociocultural, which sought to rethink the “firm” as one of the most contested analytical categories in industrial geography and to point out that it is not a self-contained and homogeneous black box driven purely by profit-maximization motives, as argued by the neoclassical industrial geography. Instead, it is internally heterogeneous and contested by different interest groups and corporate actors that possess different levels of power and varying knowledge and resources, and these actors are also subject to multiple discourses (Schoenberger 1997). The “monolithic” understanding of the firm as well as the economy has been thereafter deconstructed to show the ways in which firms and their geography specificity are socially and discursively constructed (Gibson-Graham 1996).
Industrial district, embeddedness, cluster, and agglomeration

Ideas of both industrial district and “embeddedness” have been at the forefront of an outpouring of work on industrial cluster, agglomeration, innovation system, and innovative milieu, much of which is also more or less interconnected with studies listed in the previous section, on the firm and networks. During the 1980s and 1990s, Allen Scott was one of the first to rediscover Alfred Marshall’s idea of industrial district and constructed a new theory of the dynamic firm based on transaction costs, external economies, and spatial linkages. The industrial district, defined as a tight-knit agglomeration of small and dynamic firms with a high degree of specialization and production interconnection, became a focal point for a large amount of work primarily carried out by Allen Scott and other members of the California School. Furthermore, Michael Piore and Charles Sabel’s influential book *The Second Industrial Divide* took this strand of literature to new heights, by claiming that industrial districts were central in the shift from mass production to flexible production. Later work within this field became increasingly diverse: some of these studies tested the idea of “innovative milieu” by examining the formation and transformation of industrial districts; some were concerned with defining different types and forms of industrial districts; some emphasized the importance of an appropriate institutional framework (see above “institutional thickness”), in particular the presence of thick social relations among firms and the larger communities in which they embedded (Karl Polanyi and Mark Granovetter’s idea of “embeddedness”); some focused on the coherence, spatial boundaries, and openness of clusters, and analyzed clusters along multiple dimensions, such as horizontal, vertical, institutional, external, and power dimensions (Bathelt 2005). Concepts of social embeddedness and industrial district have also been applied to analyze innovation and high-tech development. For instance, Michael Storper, member of the California School, coined the term “untraded interdependencies” in his work on high-tech industrial districts to better understand the role of cultural and social relations.

The most recent studies are becoming highly critical of the industrial district literature as a success story and as a universal model for policy initiatives. In many of these studies, the flexible specialization on which industrial district is based has been challenged by an alternative interpretation that there is no inevitable transition from one mode of production systems (mass production) to another (flexible production). Often observed is the simultaneous existence of both mass production for scale economies and flexible specialization for scope economies even in the same sector located in similar regions (Hullacháin 1992). Contrary to predictions derived from the industrial district model, horizontally and vertically integrated firms may be no less competitive than their specialized counterparts. Other studies have also stressed that traditional industrial district model and previous studies on cluster tend to focus on the local network of social relations (“local buzz”) at the expense of extra-local linkage (“translocal pipeline”) (Bathelt 2005). The central question is thus at what spatial scales industrial districts and clusters are constructed and analyzed.

Global production networks, global commodity chains, and global value chains

The above-mentioned contemporary research in industrial geography has all influenced what could be called the Manchester school of global production networks (GPNs) around Peter Dicken, Neil Coe, Martin Hess, Henry Yeung, and others. The GPN approach, defined as...
the globally organized nexus of interconnected functions and operations by firms and nonfirm institutions through which goods and services are produced and distributed, provides a broad relational framework for the study of economic globalization. It offers an open and geographically sensitive perspective that goes beyond the more restricted and linear framework proposed by the related concepts of global commodity chains (GCCs) and global value chains (GVCs), which are defined as a network of labor and production processes whose end result is a finished commodity. Specifically, on the one hand, these three approaches share something in common. They all focus on the globally coordinated inter-organizational relationships that underpin the production of goods and services, and the power and value dynamics therein. On the other hand, the GPN approach has five distinctive features that are absent in GCC/GVC analysis: first, by explicitly taking into account the extra-firm networks, GPN necessarily incorporates the broad range of nonfirm organizations that may shape industrial activities; second, GPN analysis is designed to be multiscalar and to stress the interactions and mutual constitution of all spatial scales from the local all through to the global; third, the adoption of the metaphor of networks over chains helps to avoid deterministic linear interpretations of how production systems operate and generate value; fourth, governance structures of GPNs are conceived as more complex, contingent, and variable over time than is suggested in GCC/GVC analysis; finally, GPN analysis is not to consider the networks in an abstracted manner for their own sake, but to reveal the dynamic developmental impacts that result for both the firms and territories that they interconnect (Coe 2011). More recently, GPN and GVC/GCC analysis have been gradually converging as the differences between the two approaches are narrowing.

The GPN approach is based on three conceptual categories (MacKinnon 2012). The first one is value, which aims to incorporate both Marxian notions of surplus value and orthodox definitions of economic rents. One subsequent question is about which actors and locations in the networks are able to capture and retain value. The second conceptual category is power, defined primarily as a practice in terms of the capacity to exercise power, including corporate, institutional, and collective power. The last one is embeddedness, which exists in three forms in GPN analysis – societal, network, and territorial embeddedness. In this sense, regional development is considered as a product of the dynamic process through which actors in cities and/or regions coordinate, mediate, and arbitrage strategic interests between local actors and their counterparts in the global economy, that is, “strategic coupling.” While recent GPN analysis has started to examine the so-called dark side of strategic coupling, it still tends to overlook the tensions that arise from the differential powers of key agents, such as the often unequal nature of the relations between transnational corporations and local residents and communities (MacKinnon 2012; Coe 2011).

Labor and labor market

Given the emphasis on production by industrial geographers, it is odd that labor and the labor market until recently received so little attention. For instance, by focusing on trust-based ties and dense social networks, industrial district literature in some instances tended to underplay the role of other important factors, such as power and inequalities within industrial districts, the limitations of networks, the informal economy, gender and ethnicity, the generation gap, wages and working conditions, and so on, and portrayal of networks as webs of trustful
ties led to a tendency to overemphasize their more “benevolent” attributes and to overlook the dark side of industrial districts and networks. Recently the attitude to labor and the labor market has changed and it has become a fertile topic of research. In part, the new interest stems from globalization, which has stimulated research at both the high and low end of the labor market. On the one hand, there is work examining the international movement of various high-earning, highly skilled international business elites, such as bankers, managers, and entrepreneurs, and highlighting their high level of mobility around the world. On the other hand, increasing attention has been directed towards the unequal and highly segmented nature of the contemporary labor market, contingent and precarious work, “docile” and “submissive” workers in developing countries, and the immigrant workforce.

Furthermore, there is a significant movement away from geographies of labor to labor geographies. While the former treats labor as the passive victim and conceives geographies of labor as a passive territorial outcome of industrial restructuring, labor geographies reconceptualize labor as an active constituent of the global economy, and ascribe causal power to workers as active geographical agents and to labor relations as a key set of forces transforming contemporary industrial landscapes. The viability of labor geography does not hinge on arguing that labor is powerful, but hinges on underscoring that labor can be a distinct and autonomous force in specific temporal and spatial circumstances. Four major strands in the literature on labor geographies can be identified: restructuring of labor markets as a result of increasing numbers of new entrants (such as women and immigrants); social regulation of labor markets and their distinctive geographies of governance; labor organization and unionism; labor market segmentation (Yeung 2002).

The evolutionary approach

The use of the evolutionary approach in industrial geography is far from new. Evolution was central to behavioral understandings of the “geography of enterprise” or “corporate geography” in the 1960s. Radical approach and Marxist theorization of industrial geography in the 1970s and 1980s were centrally concerned with the long-term development and transformation of the capitalist space economy. The use of evolutionary terminology such as “path dependence” and “lock-in” in research on regional economic change by industrial geographers has really taken off since the 1990s, as part of a wider institutional turn driven by fertile engagements with economic sociology and the embeddedness-network paradigm (Coe 2010). For instance, a range of work on changing geographies of production in old industrial regions has tended to focus on regional development trajectories, and analyses have been predominantly preoccupied with regional path dependence and lock-in. The last few years have even seen a conscious effort to develop a more systematic and rigorous theoretical framework for the evolutionary approach. However, the various deployments of evolutionary terminology in industrial geography still fall short of offering a coherent analytical framework. Some concepts are deployed in an isolated manner, and the entire evolutionary project remains pluralistic, incorporating a range of theoretical approaches, such as evolutionary theory of the firm, Darwinian notions of variety, selection, and heredity, and complexity science. Coe (2010) has identified three main branches of evolutionary research: work on path dependence and lock-in in different geographical contexts; research pertaining to clusters, localized learning, and related variety; and analyses of the spatial evolution of industries across the economic landscape.
Conclusion

When most people talk about industrial geography, they refer to an intellectually lively, open, eclectic, versatile, dynamic, and pluralistic discipline, particularly from the latter half of the twentieth century onwards, which encompasses a diversity of ways of describing and examining the spaces, places, and geographical circulation of industry. Due to this, industrial geography is necessarily an interdisciplinary pursuit, one that evades precise definition and intellectual closure. Contemporary industrial geographies are hybrid formations in which multiple paradigmatic viewpoints, drawn from multiple historical and spatial contexts, coexist and jostle within the same institutional and subdisciplinary spaces. The last decade has witnessed increasing intrusions by other social science disciplines into industrial geography’s traditional intellectual territory. While it is tempting to turn one’s back on such intrusions, an alternative strategy is to embrace them, learn from them, and explore industrial geography’s connections with studies of regional development, labor markets, and industrial change in development studies, economic sociology, international political economy, business and management studies, gender studies, and so on. There are also ongoing reflexive debates about the relationships between industrial geography and economics. This strategy not only gets the discipline of industrial geography better recognized in the postdisciplinary era, but also inherently benefits its research and public reach.

The final question for this academic discipline concerns what industrial geographers can do to change the real world. Since the introduction of location theory into industrial geography there has always been an interest in problems of planning industrial development and the policy relevance of the discipline. This policy relevance ranges from the strong impetus provided by industrial geographers working with local economic development authorities, to the involvement of industrial geographers in shaping business strategies and practices at the firm level. There is clearly a need for industrial geographers to be much more proactive in influencing public policy and setting policy agenda. In other words, industrial geographers have to reach out to the actors who are not just at the receiving end of economic changes but, equally important, also at the beginning of the decision chain that produces economic changes, and to widen the audience that industrial geographers are speaking to (Yeung 2002). In short, industrial geography remains a work in progress, as are its object of inquiry, theoretical framework, methodology, set of techniques, perspective, and policy relevance.

SEE ALSO: Behavioral geography; Cultural turn; Global commodity/value chains; Global production networks; Globalization; Industrial agglomeration; Industrial districts; Industrial location theory; Network analysis; Trade, FDI, and industrial development

References


Industrial linkage can be broadly defined as the contacts and flows of information and/or materials between two or more industrial sectors or firms. The concept is widely used in industrial and economic geography to portray interfirm interdependence. A firm’s linkages can be divided into: (i) **backward linkage**, which provides goods and services for its production activities; (ii) **forward linkage**, which refers to links with customers purchasing its products; and (iii) **sideways linkage**, which refers to interactions with other firms involved in the same processes. Here information linkages are as important as material linkages. There is no universal consensus among geographers as to what industrial linkages are. Geographers sometimes deliberately define “industrial linkage” differently to make their studies manageable and/or to plug some perceived gaps in the literature. For instance, some take a tight manufacturer-to-manufacturer view of linkages while others incorporate all economic activities. Micro-level research focuses on interfir or interplant linkages, whereas macro-level research may examine intersectoral linkages owing to the lack of relevant data.

Geographers who have rediscovered Alfred Marshall’s (1920/1890) idea of the “industrial district,” which underscores the predilection of firms to cluster and co-locate geographically tend to place emphasis on the spatial binding forces exercised by industrial linkages. They argue that industrial linkages are still a powerful sustainer of industrial agglomerations, and thus have fundamental impacts on firms’ location/relocation decision-making. The interactions between linkage partners represent one of the routes through which the economic multiplier operates, to further enhance some initial change in the economy through its interdependencies with others. This argument was extended by François Perroux’s (1955) idea of “growth centers” or “growth poles” to establish, in a confined geographical area, a nexus of linked economic processes to the mutual convenience and economic benefit of all concerned. Albert Hirschman (1958) further elaborated the ways in which diffusion of economic growth would be propelled through forward and backward linkages. The last type of practical importance of industrial linkages is that material and information linkages also serve to widen a firm’s knowledge of space, and this in turn encourages and facilitates further spatial expansion when needed. Industrial linkage does not necessarily occur in a “short distance,” as depicted in the industrial district and agglomeration research. Instead, its geographical patterning is dependent on transportation costs, development of telecommunication technology, standardization of production, as well as political, institutional, and societal factors. Studies on industrial linkages are often based on: (i) macro-level studies employing an input–output framework, industrial employment analysis, an agglomeration approach, or spatial linkage analysis; (ii) static micro-level studies of national linkages using cross-section data; (iii)
INDUSTRIAL LINKAGE

dynamic and quasi-dynamic micro-level studies of material linkages; and (iv) information flow studies.

However, it has been argued that a traditional and neoclassical understanding of industrial linkages tends to rob industrial geography of the ability to perceive and treat plants and business organizations as unequal participants in webs of interactions, that is, in power networks. Power in the studies of industrial linkages can be defined as the capability of one organization to control the resources necessary for the functioning of another, such as capital, finance, materials, land, and labor. In this sense, industrial linkage is reconceptualized as the operational manifestations of power relationships, and as the spatial dimension inherent in all power networks.

SEE ALSO: Growth poles and growth centers; Industrial agglomeration; Industrial districts; Input–output analysis

References

Industrial location theory

Dean M. Hanink
University of Connecticut, USA

Industrial location theory concerns the determination of the optimum location of commercial establishments (typically factories, but farm production and service production are also included), although many of its essential constructs are also often applicable to non-commercial facilities. It falls within the general class of normative theory, so it primarily concerns ideal locations rather than actual ones. It does, however, generate empirically testable propositions. Fundamentally, industrial location theory for market economies can be reduced to a simple tenet: firms or individuals locate production facilities so that profits are maximized. While the tenet is simple, it hides a very complicated reality consisting of variations in costs, prices, geographical markets, uneven spatial distributions of resources, and so on that require attention in the interest of gaining insight into specific principles of location.

Foundations

Transportation costs are perhaps the most common complicating factor considered in industrial location theory and models. The impact of transportation costs on the location of production is considered essential in two fundamental early formulations of location theory: von Thünen's theory of the location of agricultural production and Weber's theory of the location of manufacturing. Both Thünen's and Weber's theories assume producers are price-takers in competitive markets, and in their simplest forms concern the location of a single producer.

Johann Heinrich von Thünen's work, initially published in German in 1826, extended David Ricardo's theory of land rent by placing it in a spatial context (Dunn 1955). Ricardo's theory essentially states that land use sequence resulted from variations in fertility. The most fertile agricultural land is used first; the least fertile land is used last. That result is essentially a response to the relation between market prices and production costs. Highly fertile land requires little additional cost in production while less fertile land requires fertilizer, for example, which is costly. Less fertile land will only be put into production when market prices are high. The difference between market price and production cost (including normal profit) is called economic rent, and land is only used when that rent is non-negative. Thünen's work is concerned with location rent, the expected profitability of agricultural production at a particular location.

Thünen made simplifying assumptions in focusing on the impact of location on agricultural production. One is that all commercial agricultural output is consumed locally at a town – a point location. Another is that the town and its surrounding land are located on a homogeneous plain. There is no variation in soil fertility, which is everywhere sufficient for crop production, and the plain is isotropic, so that movement is possible everywhere with uniform cost per unit of distance regardless of direction. There are variations in market prices for individual crops, and variations in production costs based on their characteristics rather than
soil fertility, but the really significant variation in Thünen’s theory is in the transportation costs of individual crop varieties: wheat, for example, has one transportation cost, corn another, and so on. From the general propositions of Thünen’s theory a location rent (LR) identity can be defined:

$$LR_{ij} = Q_i(p_i - c_i) - Q_f k_j$$  \(1\)

where \(Q_i\) is the quantity produced of the \(i\)th crop, \(f_i\) is its cost of transportation per unit of distance (e.g., kilometers), and \(k_j\) is the distance from some production point, \(k\), to the marketplace point. (This rent identity was actually provided by Dunn (1955); see next section.) Production of a crop can occur anywhere that \(LR \geq 0\) (Figure 1). In Thunen’s system the expansion, or contraction, of agricultural land use does not vary with environmental quality as in Ricardo’s, but rather varies with transportation costs. Cheaply transported products can be grown at relatively great distances from the marketplace, while those that are expensive to transport must be grown relatively close to where they will be consumed. Further, a decrease in transportation costs in general, due to technological improvement, for example, extends the total area over which agriculture is profitable, while an increase in general transportation costs decreases the area of feasible production.

The crop that can earn the highest location rent at any particular point is expected to be grown at that location in a market economy. Thünen’s theory leads to a rationalized landscape of concentric zones of homogenous production when extended to multiple crops. Crops of a single type, for example wheat, are grown within certain radii of the market where they enjoy, as a group, higher location rents than other crop types, while crops of other types have their own zones, with again each defined by a commonality of relatively high location rents within types.

Thünen’s theory was developed for its time in the early nineteenth century so it concerned agriculture. Working early in the twentieth century, Alfred Weber’s (1909) industrial location theory concerned manufacturing. Like Thünen, Weber assumed punctiform demand, with the simplest system based on a single location of consumption of the manufactured good. Raw materials required for the production of the manufactured good are not evenly distributed across a homogenous plain as in Thünen, however, but normally, like the market, located at points. Weber’s theory in its basic form defines the optimum location for manufacturing a product to be at the point where the total transportation cost of shipping the raw materials from their sources to the factory plus the cost of shipping

![Figure 1](image-url)
the finished product to market is minimized. Finding the point where such transportation cost is minimized is often called the Weber problem, and underlies a class of mathematical problems in location analysis and operations research that have objective functions of distance-related cost minimization.

Weber used the product of weight × distance as an equivalent to the cost of transportation, so raw materials with lower weights have less impact on the selection of a factory’s optimum location than do raw materials that are heavier. In addition, the weight of finished products relative to the weights of their raw material inputs plays an important role. A comparison of the raw material and finished product transportation cost is embodied in a material index (MI): $MI = \frac{\text{weight of localized raw material}}{\text{weight of finished product}}$. When $MI > 1$ then manufacture of the product in question has a raw material orientation; that is, its likely best point of production is nearer the raw material source(s) than it is to the marketplace. When $MI < 1$, then manufacture of the product has a market orientation; its optimum production location is likely nearer the marketplace than it is to the raw material source(s). The value of $MI$ can only be greater than one when a nonlocalized (non-point-specific location) raw material, or ubiquity, is used in production. If only ubiquitous raw materials are used, the optimum plant location is at the marketplace (Figure 2).

Weber extended his plant location calculus to include the impacts of labor availability and also the achievement of external economies of scale (decreased average production costs induced by proximity to other producers or large populations) through the agglomeration, or spatial concentration, of manufacturing plants. As with raw materials, low-cost labor sites are assumed to exist at particular points that are not coincident with either the locations of raw materials or the marketplace. If labor costs at one of those points are low enough to offset the increase of transportation cost incurred by moving from the point that minimizes them to the labor source, then those low labor costs would have an effect on a plant’s location. The availability of external economies of scale is treated in the same way. If a decrease in costs associated with agglomeration is greater than the increase in transportation costs associated with the relevant change in the plant’s location, then the site of the agglomeration is the optimum location.

Both Thünen’s and Weber’s approach to location theory postulated markets at fixed points, and concentrated on minimizing the transportation costs incurred in supplying those locations where demand occurs. An alternative to that geography of demand is to assume it occurs over a continuous surface, and that is the supposition followed in location theories.

Figure 2 The basic Weber model. The best location for the plant ($P$) is determined by the pull-effects of raw material sources ($M_1$ and $M_2$) and the market ($C$). Source: Hanink DM 1997. Fig. 7.9, p. 254. Reproduced by permission of John Wiley & Sons, Inc.
that focus on revenue maximization, rather than cost minimization, as their means of maximizing profits. The basic format for revenue maximization location models was developed initially by the German geographer Walter Christaller (1933) in his central place theory – a theory of the spatial hierarchy of settlement. Christaller’s theory is essentially top-down. Given a uniform distribution of population across a uniform plain, a central place that is a distribution point for all goods and services can be expected to form in the actual spatial center of the population. Given that place’s location, a series of lower-order central places can be located that minimize goods- or service-price to transportation:cost ratios for consumers and, at the same time, provide sufficient numbers of consumers to make the distribution of goods and services from particular locations profitable (Figure 3).

The German economist August Lösch (1954), working at about the same time as Christaller, formalized the relationship between cost minimization for consumers and revenue maximization for producers as the foundation of the spatial economy. Lösch also developed a central place system but, unlike Christaller, took a bottom-up approach that provided a comprehensive location theory of the firm. Again, the initial assumption is that consumers are evenly distributed over a large plain, which is boundless for all purposes. The plain is also isotropic, a simplifying assumption with respect to making transportation costs uniform. Given that simple landscape, where should a producer locate? The answer to that question depends upon what is being produced. Production costs are not an explicit issue in the location model, but price in relation to transportation costs is the characteristic of interest.

As in Christaller’s system, location on the Löschian landscape depends upon the joint evaluation of consumer and producer needs. The distance over which a good or service can be sold is called its range. The range of a good or service is dependent upon the ratio of its purchase price to its transportation price (as in Christaller’s system), which is assumed to be paid by consumers. At extremes, expensive goods and services that have low transportation costs can have very long ranges, while inexpensive goods that have relatively high transportation charges have short ranges, as consumers are unlikely to pay more for a good’s or service’s transportation than for the good or service itself. There are a variety of intermediate relationships between purchase price and transportation costs, each defining the critical characteristic of a good or service.

Figure 3 The basic geometry of central places. High order places (H) include low order places (L) in a nested spatial hierarchy. The markets have hexagonal forms to ensure full coverage without any overlap. Source: Hanink DM 1997. Fig. 8.3, p. 286. Reproduced by permission of John Wiley & Sons, Inc.
with respect to its location of production. Every good and service has an *outer range*, which is the greatest distance it can be sold from its place of production. Expensive goods and services are assumed to be bought infrequently by consumers, while inexpensive ones are purchased with relatively high frequency. The volume of sales a producer needs to exist, or minimum operating scale of production, is called the *market threshold*. Market thresholds can be defined by spatial extents, because consumers are uniformly distributed across the plain. Market thresholds define market areas that are necessary to support the minimum operating scale of the producer of any particular goods or services. In the absence of competition, each area that just contains the market threshold is a circle with a radius called the *threshold range*.

In the simplest case, the Löschian landscape would have a uniform distribution of a series of producers with circular market areas, but that shape is unsustainable. If the circular market areas do not overlap then some consumers would not be in range of a particular good or service, and if the circles overlap then producers may not have market threshold demand. Rather than circular market areas, a developed Löschian landscape contains a series of producers serving hexagonal markets, with shape being the regular tessellation that minimizes the aggregate transportation cost to consumers while ensuring that market areas do not overlap, thus ensuring that threshold demand is met for each producer in an effective spatial monopoly. In its basic form, the necessity of spatial monopoly to ensure producer survival means that concentrations of producers of the same good or service cannot exist. Spatial concentrations of producers do emerge, however, but only of those producing different goods and services. For example, a concentration of producers of goods A, B, and C could emerge, but two producers of service A or of good B would not be found at the same location.

The monopolistic competition in Löschian location theory runs counter to a spatial extension of Harold Hotelling’s (1929) principle of minimum differentiation. That theory actually concerned the tendency of competitors to minimize the differences of their products so as not to risk losing market share to each other. One company’s cornflakes are very similar to another’s, and so on. Hotelling used a locational analogy of competing firms locating along a main street in a town to illustrate the principle. As in Lösch, consumers are uniformly distributed, so barring differences in the firms’ products they are likely to simply shop at the firm nearest to them. If there is one firm, it could locate anywhere on the street, but if there are two firms, their best strategy is to co-locate so that neither has a locational advantage. Just as locations are the same for the competing firms, in an aspatial context their products would be very similar as well. In Hotelling’s model, if there is more than one producer then a spatial monopoly would likely lead to diminished profits rather than the assurance of survival as in Lösch’s location model.

**Extensions**

Thünen’s original land use theory is part of the body of classical economics that also includes work by notable early economists such Adam Smith and David Ricardo. A more contemporary, neoclassical, version of Thünen’s theory was provided by the economist Edgar Dunn (1955). Dunn relies on the basic assumptions of Thünen’s model, a homogeneous plain, a central market, and so on, but expands the analysis by developing the marginal rent function, described previously, that explicitly ties land/location rent to distance. By formalizing the concept...
of distance decay, Dunn was able to specify the locational limits for crop production singly, but more importantly was able to specify the equilibrium spatial pattern of a series of crops that are essentially treated as competitors for locations in a regional economy. Dunn also introduced a dynamic analysis in which changes in transportation prices induced changes in the locational patterns of agricultural production.

William Alonso (1964) extended the Thünen/Dunn analysis to urban land. In doing so, Alonso developed the concept of the bid–rent curve, which allowed for a more direct analysis of the market – supply and demand – for locations within a metropolitan area. Alonso’s city is assumed to have essentially the same simplifying background conditions as the agricultural region of Thünen and Dunn. Instead of a central marketplace where agricultural goods are sold, Alonso’s city has a central business district (CBD) that typified many American cities at the time. The CBD was just that because it is a transportation sink, the point in the urban region where aggregate transportation costs are minimized. As in the Thünen/Dunn analysis, the region has an isotropic surface that allows transportation costs to be uniform across the region without the complicating factor of transportation arterials channeling flows.

The effect of transportation costs in the Alonso model is not, however, the same as in the agricultural land use model. Businesses in the urban region do not transport their goods or services to a central place of consumption, thus lowering their transportation costs if centrally located themselves. Instead, their costs of doing business are increased with distance from the central point for two reasons. One is that their selling cost is raised as their distance from the transportation sink is increased, because the transportation cost for local consumers increases, on average, with distance from the center. That increased transportation cost requires the business to provide some sort of offset, such as a price discount, to attract sales. The second reason that business costs increase away from the center is that the business’ employees have, on average, increased commuting costs when working away from the transportation sink. Businesses must provide higher wages with distance from the center. The transportation-related factors attract business toward the center in the region, but the higher land costs at the center due to competition for low business-cost locations raise the price of land. Bid–rent curves trace the tradeoff between higher (lower) land prices and lower (higher) selling/labor costs for each business in the urban area. A business’ location is determined by the tangency of its bid–rent curve with the land price gradient that peaks at the center and declines with distance from that point. Businesses with the need for frequent interaction with customers and that are labor intensive are likely to be found near the urban center, while those that interact with customers infrequently and use less labor are likely to locate in less central locations.

The relationship between selling and labor costs and distance from the urban area’s transportation sink provides a theoretical foundation for the spatial concentration of businesses in urban places. An alternative theoretical foundation for such concentrations was developed as an extension of the Hotelling principle of minimum differentiation in its spatial context by Yorgo Papageorgiou and Jacques-François Thisse (1985) and by Masahisa Fujita (1988). These models concern the interaction between households, which prefer to be near a large number of businesses so that they have access to a wide variety of goods and products, and business, which, in turn, are attracted to locations with a large number of consumers. Complete spatial concentration of business and households
is limited by two factors. One factor is that consumers must pay high rents when located in proximity to business. The other factor is that businesses are faced with significant competition in such concentrations. Under such conditions, significant spatial concentrations only occur when consumers place very high values on access to a wide variety of goods and services.

The interaction of consumers and producers is also the focus of numerical models that are focused upon retail business’ location and market areas. At the aggregate scale, Reilly’s Law of Retail Gravitation (Reilly 1931) concerns the spatial extent of a city’s retail area and can be used to define a breakpoint between cities in terms of their spatial retail market share domination. The model ultimately has its foundation in Löschian analysis, but is probabilistic rather than deterministic in defining spatial markets. The geographer David Huff (1963) provided a microscale retail location model that also relies on probabilities of consumer attraction to particular places to delineate spatial markets, and also allows for potential competition among retailers to play a role in their selection of a location.

The economist Walter Isard (1956) extended many strands of location theory. Some of his work fuses aspects of the work of Thünen, Lösch, and Weber into a general theory of the space economy. He also tried to integrate location theory into broader economic analysis by treating transportation cost as an input in the same way that capital (including land) and labor are considered. In a standard production function, the volume of output is a response to capital and labor inputs. Those capital and labor inputs may be substituted for each other within certain limits, so labor will be substituted for capital if it is efficient to do so, and vice versa. In Isard’s analysis, transportation inputs may be substituted for each other, so that raw materials from less costly sites are used in preference to those from more costly sites. Isard also extended location theory by considering the effects on site selection and market areas of transportation costs that are not simply linear functions of distance. Another of his extensions considered the impacts of internal economies of scale, which are decreases in average costs associated with increasing volumes of production, on the size and location of competing businesses in a spatial economy.

While the foundations of location theory and its early extensions were established before game theory was developed, that analytical approach was later used in location theory that implicitly or explicitly requires interaction among producers or producers and a government entity. For example, the spatial version of the Hotelling model yields a Nash equilibrium, a basic construct of game theory, if two businesses locate in the same place, albeit under some restrictive assumptions including constant prices and a uniform market of consumers. Isard (1966) used game theory in analyzing the location of producers and the development of agglomerations, as considered earlier by Weber. Weber’s agglomerations are established where conditions beneficial to producers with respect to increasing output and lowering transportation costs occur. Isard showed that benefits to economic growth in a region could induce its government to provide a side payment to producers to establish geographical concentrations within its jurisdiction. The payments are made in a game that has producers and government as players. Robert Cromley and Dean M. Hanink (Hanink and Cromley 2008) extended the analysis and related the interaction between producers and government to the cluster policies used by a large number of governments in the interest of regional development.

In addition to the increasing sophistication of its principles and relaxation of its more restrictive assumptions, location theory was also extended
by its operationalization in various forms of mathematical optimization. The economists Martin Beckman and Thomas Marschak (1955) described the locational decision of a business as a type of activity analysis in which each of its actions has an implication for its location, or locations, of production. The objective function of the analysis is typically profit maximization, and its input requirements, distribution costs, and so on can be treated as constraints that limit not only profitability but locational choices as well, due largely to transportation costs. The regional scientist Benjamin Stevens (1961) formulated the Thünen model as the transportation problem of linear programming. In this case, the minimization of transportation costs is the objective function and constraints concern volumes of production and demand.

Criticisms and redevelopment

Location theory has received criticism on several counts. One of the most telling is that its predictions are not consistently empirically observed; locations of production often do not appear in expected places or patterns. The geographer Glen Norcliffe (1975) cited several factors that apparently undermined the applicability of location theory’s basic forms. One was that the declining relative importance of transportation costs in the entire production process undermined the usefulness of theory in which such costs play a major role. The expansion of light industry (especially compared to the situation when Weber developed his least-cost model), an increasing efficiency of material use and substitution, improved transportation technologies, and economies of scale in freight movement have all contributed to a declining impact of transportation costs on industrial location. Instead of transportation costs playing a pivotal role in the location of a producer, Norcliffe and other theorists argued that place characteristics, such as market size and land costs, should be considered the foundation for a new location theory that would provide better insight into location decisions and resulting location patterns than can theory based on transfer costs.

Criticisms such as Norcliffe’s are based on changes in production and transportation technology that seem to make the critical assumptions of location theory, and even its objectives, moot. The geographer Doreen Massey (1973) criticized industrial location theory on more profound grounds. Massey criticized the entire notion of location theory as an entity unto itself for three reasons. One is that the location of production cannot often be disintegrated from the other economic decisions of a business. Another is that the location of production is rightfully considered as the result of factors specific not to an individual business, but rather to the interaction of the business with its sociopolitical environment. The third is that the aspatial characteristics of an economic system have a spatial manifestation that can override the locational preferences of any individual business. Massey also argued that the partial equilibrium approach in location theory, which treats business locations as independent of other location decisions, would be better formulated in a general equilibrium context that considers the effects of those location decisions on other agents in the economy. Massey had specific criticisms in addition to her more general ones. With respect to individual approaches, she noted that Weber assumed perfect competition among businesses within an industry, but perfect competition is unlikely in a spatial economy. Another criticism, leveled at much of economic theory, is that the location decision is made by a completely rational actor with complete information, leaving no behavioral variation. Massey also criticized
the Lösch model, noting, as have others, that its location decisions are based on spatial monopoly but yield a system of places with varying levels of spatial agglomeration as the unlikely result of profit maximizing behavior by businesses that leads to fulfilling welfare criteria such as minimizing consumer travel.

The economist Paul Krugman (1995) has been more recently critical of industrial location theory. He noted the criticisms of its unlikely assumptions, but was sympathetic in their usefulness as simplifying constructs that allow a focus on the most important factors in the location decision. More tellingly, Krugman criticized received location theory as being more geometric than economic in its analysis. The locational patterns expected in location theory are derived largely with respect to principles of space, but the economic content is thin. Krugman questioned, for example, the structure of Weber’s theory, almost in a way that mirrors Massey. He points out that Weber’s theory does not consider the prices charged by the producer, or whether there are competitors, or whether multiple sites of production are better than a single one. There is no context for the location decision, and without prices the theory does not even meet the requirements for providing a true partial equilibrium model. Krugman also has criticisms of central place theory. The Lösch model, for example, has an outcome based on spatial optimization rather than (aspatial) market processes and while Christaller’s model has outcomes based on market processes, it is unclear about market structure.

The simplifying, but restrictive, assumptions of industrial location often lead to challenging tests of empirical verification. The geographer E.M. Rawstron (1958) provided three principles of industrial location that make location theory more flexible with respect to its predictions and also provide guidance with respect to the addition of controls in empirical testing of its models. The principles are physical, economic, and technical restrictions to locational choice that effectively elaborate upon the context of a theoretical model and define the possibility of a range of potential locations within the restrictions rather than focus on one particular site. The geographer David Smith (1971) elaborated upon Rawstron’s work and described spatial limits to production profitability that combined elements of revenue maximization and cost minimization in defining a geographical range of suitable locations that extended beyond a single optimum site. The approach is in line with the principle of satisficing as a decision-making criterion introduced by Herbert Simon (1947) in which choices are made based upon the achievement of an acceptability threshold rather than a maximizing criterion.

Behavioral location theory was developed by the geographer Alan Pred (1967). In his approach, location decisions were no longer implicitly made by completely rational agents with perfect knowledge and perfect information, but rather made by agents with varying levels of ability and information. While particular locational patterns are still expected in the behavioral models, they are much more accommodating of exceptions and conform well to Smith’s spatial limits model.

Industrial location models have also been modified to incorporate uncertainty and risk. Work on locational behavior of businesses recognized that modern corporations have multiple objectives that are beyond the scope of classical location theory. In an uncertain business environment, firms must plan an overall investment policy accounting for risk. The geographers Robert Cromley and Dean Hanink (1985) used portfolio theory from financial economics to model location strategy. In their approach, as
INDUSTRIAL LOCATION THEORY

in the behavioral models, location is not deterministic but rather is related to the firm’s utility function. As management varies in its goals from total risk aversion to full profit maximization, the optimal pattern of location is altered. Additionally, their model incorporates existing patterns of locations into the business’ investment planning by finding a new pattern of investment that is complementary to the existing distribution.

Paul Krugman (1995) developed the currently dominant variant of industrial location theory in the 1990s in what is now called the new economic geography. The model is largely consistent with many earlier ones in that the location of a business’s production faces two often contradictory forces. One is a centripetal force caused by the benefits of economies of scale that make the business want to have all its production in one place. The other is a centrifugal force caused by transportation costs incurred by consumers, which makes the business want to have a dispersed number of locations. If transportation costs are very low and economies of scale are significant, then businesses are large in size and located in one or a very few locations. If transportation costs are very high, and economies of scale are minor, then businesses are small in size and spatially dispersed. Krugman’s contributions were to provide a general equilibrium location model and to use on a Dixit–Stiglitz utility function for consumers that ultimately allows imperfect competition among businesses. The resulting model is very similar to that of Lösch in its potential development of a landscape of hierarchical agglomerations. It also accounts, however, for the prices of goods and services, and the wages of the labor force that comprises the set of consumers, thereby providing a complete picture of a spatial economy not developed in other theoretical approaches.

While aspects of location theory have been subject to criticism, it is widely held that the principles of location theory, even in the most basic forms, are useful in providing insights into current and future spatial patterns of economic activity (Kilkenny and Thisse 1999; McCann and Sheppard 2003). For example, the flow of foreign direct investment from wealthier countries to poorer ones in the interest of securing low cost labor is easily explicable in the context of Weberian location theory. In addition, the Alonso model provides a rationale for the spatial concentration of business in city centers, but can easily be extended to more contemporary circumstances in which metropolitan areas have several places with high levels of accessibility, and to the problem of urban sprawl. Even the early Thünen model is useful in analyzing contemporary patterns of urban-centered agricultural production for local markets, while the recent theory of the new economic geography has been useful in predicting the impacts of enlargement of the European Union.

SEE ALSO: Agricultural geography; Central place theory; Industrial agglomeration; Industrial districts; Industrial geography; Location-allocation analysis; Location and multiplant firms; Manufacturing industry; Regional science; Spatial analysis; Transportation and land use

References

INDUSTRIAL LOCATION THEORY


Industrial restructuring

Chun Yang
Hong Kong Baptist University, China

Industrial restructuring refers to structural changes of territorial economies from the manufacturing to services sector at a range of spatial scales involving national, regional or city levels. It contains a quantitative dimension, typified by the loss of manufacturing jobs and the growth of services; a spatial dimension, usually associated with the geographic redistribution of manufacturing jobs at national and international scale; and a qualitative one, suggested by the greater incidence of both low-wage, low-skill jobs and high-level professional jobs in service industries, a decline in wages and unionization rates in manufacturing jobs, and a feminization of the job supply. Since the Industrial Revolution, industrial restructuring in the advanced economy has experienced two major stages. Firstly, in the early part of the nineteenth century, the advanced economy completed the shift from an agrarian economy to an industrial and service economy, with the change of population shifting from a predominantly rural one to an urban one. Secondly, by the end of the twentieth century, the advanced economy shifted from an industrial economy to a services economy, with employment in manufacturing declining and employment in the service industries dramatically increasing. Such changes may represent a response to changed conditions induced, for instance, by time–space compression, technical change, or conflicts between labor and capital in the workplace, or transmitted through the competitive conditions endemic to capitalism. Restructuring may involve one or more of a number of transformations, including structural adjustment, which is defined as adaptation to sudden or large, often unexpected changes in the context of economic geography.

The decline of manufacturing and shift to services raise questions about changes in the economic base of cities and the impact of these changes on employment and income distribution. The shift to services and the spatial redistribution of manufacturing received considerable attention throughout the 1970s and this continues today. The locational patterns of the new leading industries, particularly producer services, and their impact on urban economies and the urban hierarchy are significant outcomes of industrial restructuring. Moreover, industrial restructuring has resulted in uneven geographical impacts, such as the rise of world cities, spatial mismatch, and metropolitan growth.

Sectoral and spatial dimensions of industrial restructuring

Sectoral transformation of industrial restructuring

Industrial restructuring is marked by a process of structural transformation as reflected in employment and value added in primary, secondary, and tertiary industries. According to Simon Kuznets, structural transformation occurs when the share of nonagricultural sectors in an economy increases. A reconstruction of national accounts from a variety of sources for western countries shows that such a transformation...
INDUSTRIAL RESTRUCTURING

raised overall productivity and increased the returns to workers and capital. For instance, from 1800–1849 to 1951–1960, agriculture as a share of gross domestic product (GDP) declined from 30% to only 5% in the United Kingdom and from 20% to 4% in the United States. The share of industry inclusive of manufacturing meanwhile increased from 23 to 56% in the United Kingdom and from 33 to 43% in the United States (Kuznets 1966). The sectoral shift is mirrored in the structure of employment. In consequence, the population is redistributed from rural to urban areas and the capital:labor ratio in the nonagricultural sectors of the economy rises. History indicates the critical role of industrialization as an engine of sustained growth in the long term. Since productivity growth is associated with technological change and industrial upgrading, continuous structural change in technologies and industries is the main feature of industrial restructuring in modern economy. Manufacturing has contributed to the modernization of the agricultural and mining sectors, which provide raw materials through backward linkages, and spawning services through forward linkages. Since there is a tight nexus with the services sector, technological progress and growth in manufacturing lead to a larger demand for services, propelling overall economic growth.

Long-term economic trends from the pre-industrial stage of development in a large number of developed countries confirm that at the end of each episode of catch-up, the fast grower’s economy had a structure closer to that of a developed country as opposed to a low-income one. Industrial Revolution started in the United Kingdom in the 1700s but for about 50 years it did not spread to other countries because the British government forbade the export of machinery, manufacturing techniques, and skilled workers. Eventually, in the nineteenth century, it gradually spread to other countries in Western Europe. The earliest center of industrial production in continental Europe was Belgium, where production of coal, iron, textile, glass, and armaments flourished. By 1830, French firms had employed many skilled British workers to help establish the textile industry, and railroad lines began to appear across Western Europe. Germany was a latecomer in developing industry, mainly because no centralized government existed there. Relative to the United Kingdom, industrialization was delayed in the United States because the country at that time lacked the basic factor endowments – labor and capital – to invest in business. When it finally picked up in the 1820s, its growth was explosive. Laborers and capital came from Europe, where political revolutions sent immigrants to the United States. Rapid industrialization and structural transformation then followed. In 1800, farmers comprised 85% of the US population, but the proportion declined to 50% in 1860. Starting with an income level that was only one-third of that in the West during the 1850s, Japan achieved rapid development in fifty years to become the first industrial country in Asia in 1904. After opening up trade in 1854, its government encouraged learning from Western technology and institutions by sending high-level missions including about half of the ministers to the United States and Europe for nearly two years.

The economies of Western Europe and Japan have enjoyed unprecedented growth and technological upgrading in the decades since WWII in the so-called “golden age of capitalism” (1950–1974). During this period, nearly all developing countries pursued dirigiste capitalism, except for Japan, South Korea, and other Asian NIEs. The crux of Japan’s and East Asia’s success was that their development followed closely their comparative advantage and their governments played the role of facilitators.
Japan’s historical labor statistics record that a rising share of labor in Japan’s manufacturing sector coincided with a declining share of labor in the manufacturing sectors of the United States. In the 1960s–1970s, Japan supported its heavy manufacturing sectors, including machinery and automobiles. In the 1980s–1990s, just as the United States was upgrading its industrial base, Japan expanded its shares in the home appliance, electronics, and computer markets (Figure 1). Restructuring of Japanese economy during the last three decades had involved two distinctive transformations: in the secondary sector, the emphasis has shifted from heavy to light industry, particularly consumer industry and high-technology industry; contraction of the primary sector has been offset by expansion in the tertiary sector. The first change was made because of Japan’s declining competitiveness on the world markets for heavy industrial goods (e.g. steel, ships, chemicals) and the rising costs of labor. The second change was triggered by a massive shakeout of agricultural labor and by the advent of the affluent society in Japan with its greatly increased expenditure on consumer goods and services.

Figure 2 shows the employment shares in the United States between 1958 and 2005 for five subsectors selected from 99 manufacturing industries, ranked from most labor intensive to most capital intensive. As the capital:labor ratio increases over time, industrial and employment structures change dramatically; specifically, the share of labor employed in the most labor-intensive sectors, such as fabrics, declined monotonically. In sectors such as computer manufacturing, the share of labor employed first increased and then declined, showing a hump or inverse shape. In industries such as aircraft and automobile manufacturing, which are capital intensive but subject to labor-saving scale economies, the share of labor showed a slow and declining trend. In the most technology-intensive sectors, such as plastics, including fiber optics and lens, the share of employment shows a monotonic increase, indicating that the United States still maintains a comparative advantage in these industries.

Spatial shifts of industrial restructuring

Studies on industrial restructuring were initially grounded in a detailed empirical analysis of intranational spatial restructuring in the UK economy in the 1960s and 1970s (Massey 1984), in which the emergence of a new spatial division of labor in manufacturing was based upon a separation of skilled and managerial tasks (performed in many cases in London and the Southeast) from unskilled manufacturing work done in so-called branch plants (located in areas such as northeast England and southeast Wales). The industrial restructuring has expanded from the 1960s onwards, with the emergence of a new international division of labor (NIDL), in which developing countries took on new roles within the global economy through the offshoring strategies of transnational corporations (TNCs).
Since the 1970s, industrial restructuring has been extended to involve the establishment by European, North American, and Japanese TNCs of a global manufacturing systems based on developing labor-intensive export platforms in so-called “newly industrialized economies” (NIEs) in East Asia, namely, Hong Kong, Singapore, South Korea, and Taiwan. In response, TNCs used their global reach to relocate production from the industrial core to low-cost production sites in the NIEs. The manufactured goods were then, in turn, exported back to core markets in the West from the offshore branch plants.

With the global shifts of industrial activities from North America and Western Europe to Japan and later on to the first and second generation of NIEs, such as China, India, Malaysia, the Philippines, and Thailand in Asia, the process of industrial restructuring has extended over time.
and across space to Asian NIEs and recently to emerging economies such as China. From 1965 to 1990, Japan emerged as the world’s biggest exporter of manufactured goods, increasing its share of the world market from 8 to 12%. Japan’s success was followed by a second generation of newly industrializing economies in the 1970s (Hong Kong, South Korea, Singapore, and Taiwan), a third generation in the 1980s (Indonesia, Malaysia, Philippines, and Thailand or ASEAN 4), and a fourth generation in the 1990s (China and later Vietnam). Taking South Korea as an example, the share of manufacturers in GDP rose from merely 9% in 1953 to 30.1% in 1988, while that of the agriculture and mining sector shrunk to single digit figures in the 1990s. During the phase of industrial restructuring guided by export-oriented industrialization, the benefits of economic backwardness were exploited with sequential structural transformation from labor-intensive industries (e.g., clothing) to capital-intensive industries (e.g., machinery and transport equipment). Since 1983, capital-intensive machinery and transport equipment products have accounted for the majority of total exports. After the mid-1990s, their share exceeded half of total exports.

Since the 1970s, Asian NIEs have been facing labor shortage as a result of their demographic transition. Because of rising incomes, attitudes towards manual jobs and work ethics are changing. In association with labor shortages, wage rates have been increasing rapidly. The unionization of labor also contributes to wage increases and to strained labor-management relations. International forces, including growing competition from other Asian economies, particularly Southeast Asian countries and China (in labor-intensive low-skill products), increasingly competitive markets for low-technology and medium-technology manufactured goods, pressure from the United States and the European Community for currency revaluation and open trade, and a growing protectionist tendency in the North American and European markets, put considerable pressure on the Asian NIEs to restructure their industrial economies. Research has demonstrated how the rapid appreciation of the Japanese currency, the yen, during the 1980s – the “endaka” phenomenon – undermined the export competitiveness of Japanese manufacturers and, by cheapening imports, triggered decentralization of their production networks more broadly within East Asia. In later waves, deindustrialization has hit both South Korea in places such as Busan and Ulsan in the south, and Taiwan within and beyond Taipei. Industrial restructuring has created unprecedented challenges for these developmental states and their export-led industrialization strategies that they are only now struggling to come to terms with. In addition, Hong Kong had experienced deindustrialization from the 1980s through the 1990s, where manufacturing collapsed from 17% to less than 2% of GDP and decline of manufacturing in employment composition (Figure 3) in the context of an export-dependent economic structure and a noninterventionist state (various authors 1993).

More recently, the collapse of Asia’s exports to the West during the 2008–2009 global financial crisis caused concern about the sustainability of external demand as an engine for growth. It is argued that Asia needs to rebalance away from dependence on exports to the United States and the European markets. Growth instead should be driven by domestic and regional demand. Asian economies need to reform their industrial structure to mitigate the negative impacts of the global financial crisis and to support long-term growth and development (Thorbecke and Hsieh 2013). Notably, the Chinese central government designated the transformation from export-led to domestic consumption-driven...
INDUSTRIAL RESTRUCTURING

Figure 3 Employment transformation in Hong Kong, 1994–2006. Reproduced with permission from the Hong Kong Monetary Authority.

development in the national 12th Five-year Plan (2011–2015) in the wake of the 2008 global financial crisis. Over the past three decades since China initiated its opening and reform policy in late 1970s, the export-led industrialization has been designated as a model for industrialization and economic growth, and contributed to the rise of China as the “world factory” (Table 1). Since 2000, the institutional environment for the export-oriented industrialization in the 1980s and 1990s has changed dramatically. The industrial restructuring of the export-led regional development, particularly the market rebalancing of TNCs from external to China’s domestic market, has turned out to be a viable strategy but daunting task in the context of the changing global–local dynamics (Yang 2012).

Consequences of industrial restructuring

Demographic and employment change

Traditionally, industrial geography has been primarily concerned with the location of industrial firms. Since 2000, a growing attention has been paid to the interrelationship between industrial restructuring and labor markets (Yeung 2002). For instance, it is demonstrated that the general employment decline and the rise of “flexible” employment in British retail banks in the early 1990s resulted from financial liberalization and competitive restructuring in the British banking industry. Empirical studies are devoted to analyze the impact of industrial restructuring on labor processes at the national, metropolitan, and community levels.

The peculiar outcome of industrial restructuring is deindustrialization, which refers to the

Table 1 Manufacturing’s share of GDP in the top 15 countries (2010).

<table>
<thead>
<tr>
<th>Top 15 countries</th>
<th>Manufacturing’s share of GDP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>33</td>
</tr>
<tr>
<td>South Korea</td>
<td>28</td>
</tr>
<tr>
<td>Indonesia</td>
<td>25</td>
</tr>
<tr>
<td>Japan</td>
<td>20</td>
</tr>
<tr>
<td>Germany</td>
<td>19</td>
</tr>
<tr>
<td>Mexico</td>
<td>17</td>
</tr>
<tr>
<td>Italy</td>
<td>15</td>
</tr>
<tr>
<td>Russia</td>
<td>14</td>
</tr>
<tr>
<td>Brazil</td>
<td>13</td>
</tr>
<tr>
<td>India</td>
<td>13</td>
</tr>
<tr>
<td>Spain</td>
<td>12</td>
</tr>
<tr>
<td>United States</td>
<td>12</td>
</tr>
<tr>
<td>Canada</td>
<td>11</td>
</tr>
<tr>
<td>France</td>
<td>10</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>10</td>
</tr>
</tbody>
</table>

decline (relative or absolute) in the importance of manufacturing experienced by most developed countries since the mid 1960s, as measured by its contribution to GDP and its percentage of the total labor force. It is explained by declining productivity and strong foreign competition, particularly from countries where labor is less costly and unionized. The United States and United Kingdom have been among those countries most affected to date, and their areas of heavy (iron and steel, ship-building) and traditional (textile) industry have been hardest hit. A significant amount of the manufacturing lost by developed countries has been relocated in less developed countries, with their attractions of cheaper land and labor and less stringent controls on pollution. The deindustrialization of manufacturing employment has exacerbated joblessness in the poor black community, correlating with a rise in single-mother households, high premature mortality rates, and increasing incarceration rates among black males (Wilson 1980). This creates a growing economic class division within the black community accentuated by global economic restructuring without government response to the disadvantaged. Furthermore, Wilson asserts that as the black middle class leaves the predominantly black inner city neighborhoods, informal employment information networks are eroded. This isolates poor, inner city residents from the labor market, compounding the concentration of poverty, welfare dependency, rise of unemployment, and physical isolation in these areas.

The declining manufacturing sector leaves behind strained blue-collar workers who endure chronic unemployment, economic insecurity, and stagnation (Logan and Swantrom 1990). Meanwhile, widening of the social hierarchy occurs when high-level, high-income, salaried professional jobs expand in the service industries alongside a greater incidence of low-wage, low-skilled jobs, usually filled by immigrants and minorities (Sassen 1990). The effects of social polarization include the increasing concentration of the poor, blacks, and Hispanics in large US cities and distinct social forms, such as the underclass, informal economy, and entrepreneurial immigrant communities. In addition, a notable qualitative dimension involves the feminization of the job supply, as more and more women enter the labor force, usually in the service sector.

### Producer services as leading industries in post-industrial cities

Since the 1970s, industrial restructuring has been used to remedy the situation through the geographical redistribution of production and consumption. City economies across the globe moved from goods-producing to service-producing outlets. Breakthroughs in transportation and communications made industrial capital much more mobile. In consequence, producer services emerged as a fourth basic economic sector, where routine low-wage service employment moved to low-cost sites and advanced corporate services centralized in cities (Sassen 1990). These technological upheavals brought about changes in institutional arrangements with the prominence of large corporations, allied business and financial services, nonprofit, and public sector enterprises. Global cities, such as New York and London, become centers for international finance and headquarters for multinational corporations offering cross-currency exchange services as well as the buildup of foreign banking and trading. Other cities become regional headquarters of low-wage manufacturing. In all these urban areas, the corporate complex grows, offering banking,
INDUSTRIAL RESTRUCTURING

insurance, advertising, legal council, and other service functions. Global industrial restructuring allows markets to expand in size and capacity from regional to national to international scopes.

The institutional arrangements buttressed by improved technology reflect the interconnectedness and internationalization of firms and economic processes. Consequently, capital, goods, and people rapidly flow across borders. Global economic changes and technological improvements in communications and information systems encourage competitive organizations to specialize in production easily and assemble temporary workers quickly for specific projects. Thus, the norm of standard, steady employment unravels; this began in the mid-1970s. As industrial restructuring encourages high-technology service and knowledge-based economies, massive public de-investment results. Across many parts of the United States and the industrialized Western nations, steep declines in public outlays occur in housing, schools, social welfare, education, job training, job creation, child care, recreation, and open space. To remedy these cutbacks, privatization is installed as a suitable measure. Though it leads to some improvements in service production, privatization leads to less public accountability and greater unevenness in the distribution of resources. With this reform in privatizing public services, neoliberalism has become the ideological platform of industrial restructuring. Free market economic theory has dismantled Keynesian and collectivists’ strategies and promoted the Reagan and Thatcher politics of the 1980s. Moreover, industrial restructuring requires decentralization as states hand down power to local governments. Where the federal government focuses on mainly warfare–welfare concerns, local governments focus on productivity.

Spatial effects on economic landscapes

Industrial restructuring has brought about a significant change of urban landscapes, especially in the United States. Cities such as Baltimore, Detroit, St. Louis, and others face population losses that result in thousands of abandoned homes, unused buildings, and vacant lots, contributing to urban decay. Such transformations frustrate urban planning and revitalization, fostering deviance in the forms of drug-related activity and vagrancy. Older, compact, industrial US cities have been rendered obsolete. Urban spaces become playgrounds for the urban gentry, wastelands for low-paid service workers, and locations for the underground economy. In some areas, gentrification projects have caused displacement of poverty-stricken residents. Sunbelt cities, such as Miami and Atlanta, rise to become key business centers while snowbelt cities, such as Buffalo and Youngstown, decline. Even housing markets respond to economic restructuring, with decaying housing stocks, escalating housing prices, depleting tax base, changes in financing, and reduction in federal support for housing. Soon, spatial divisions among wealthy and poor households are exacerbated. Moreover, with the movement of blue-collar employment from central cities, geographically entrenched housing discrimination, and suburban land use policy, African American youths in inner cities become victims of spatial mismatch, where their residences provide only weak and negative employment growth and they usually lack access to intrametropolitan mobility.

High-order services, an expanding sector in the industrialized world, become spatially concentrated in a relatively small number of large metropolitan areas, particularly in suburban office agglomerations.
New tendencies of industrial restructuring in the globalizing world

In the wake of the 2008 global financial crisis and economic downturn, a new round of industrial restructuring, characterized by re-industrialization in advanced Western countries, has been witnessed. Some developed countries have been encouraging reinvesting or “reshoring” back to their own markets, as a boost to their own struggling sectors. For instance, the United States launched its official “reshoring initiative” in 2010. According to a recent survey, as many as 39% of companies operating abroad were reported to be considering moving some manufacturing activities back to the United States (BCG 2011). Executives indicated that their companies are “likely” or “very likely” to bring back the various categories of work, including IT services (42% of executives), components/products (37%), services (e.g., customer services, call centers) (35%), and material (34%). In all, 37% of businesses are likely or very likely to bring goods and services work back to the United States – and/or a similar percentage are moving work closer to home (e.g., Mexico, Latin America). This means that as much as 5% of overall US procurement may come home. The huge numbers could dramatically impact US trade balances – and should provide an enormous boost to domestic manufacturers, retailers, wholesalers/distributors, and service providers. There is no doubt that growing efforts by the United States and the European Union to boost their own domestic manufacturing will certainly impact China as a “world factory.” It will, however, be limited, owing to the differences in industrial structures between China and the United States, and countries in the European Union.

Furthermore, the severe economic downturn has led to declines in manufacturing wages in some regions of advanced economies. As noted, in the United States, for instance, real wages in manufacturing have declined by 2.2% since 2005 (BCG 2011). This argues that China’s overwhelming manufacturing cost advantage over the United States is shrinking fast. In 2005, the average cost of Chinese labor was 22% of that of US labor. By 2010, average labor costs had risen to 31% of the US level. It is forecast that rising Chinese wages, higher US productivity, a weaker dollar, and other factors will virtually close the cost gap between the United States and China for many goods consumed in North America. Companies will undertake a rigorous product-by-product analysis of total costs, rather than just factory wages. For many products sold in North America, the United States will become a more attractive manufacturing option. For products that have a high labor content and are destined for Asian markets, manufacturing in China will remain the best choice because of technological leadership or economies of scale. China is certainly no longer the default option. However, these likely trends do not suggest that Chinese manufacturing will decline or that transnational corporations will shut down their mainland plants. More Chinese production capacity will be devoted to supplying the country’s enormous domestic market, which is gaining millions of new middle-class households each year, as well as other growing economies in Asia.

According to research by McKinsey Global Institute (2012), when offshoring entered the popular lexicon in the 1990s it became shorthand for efforts to arbitrage labor costs by using lower-wage workers in developing nations. As a “new normal” is settled in to – catalyzed by the global financial crisis, the ensuing recession, and an uneven global recovery – traditional arbitrage models seem increasingly outmoded. For some products, low labor costs still furnish a decisive
INDUSTRIAL RESTRUCTURING

competitive edge, of course. But as wages and purchasing power rise in emerging markets, their relative importance as centers of demand, not just supply, is growing. It is postulated that a next-shoring perspective emphasizes proximity to demand and innovation. Both are crucial in a world where evolving demand from new markets places a premium on the ability to adapt products to different regions and where emerging technologies that could disrupt costs and processes are making new supply ecosystems a differentiator.

The witnessed recent increase in manufacturing output and employment has taken place not only in North America but also in emerging markets, for example, China, since demand bottomed out during the recession following the financial crisis of 2008. Meanwhile, surging local demand helps explain why rapid wage growth in China has not choked off manufacturing expansion. The narrowing labor-cost gap between China and advanced economies reinforces the importance of local demand factors in driving manufacturing employment. Relocating manufacturing close to demand makes it easier to identify and meet local needs. Certainly, next-shoring will look at different locales. Near-shoring and re-industrialization have emerged and no doubt will evolve in unexpected ways that deserve close observation and investigation.

SEE ALSO: Deindustrialization; Industrial upgrading; Industrialization; International division of labor; Manufacturing industry

References


Further reading

Industrial upgrading

Shengjun Zhu
Swansea University, UK

The concept of industrial upgrading

One of the feasible responses of firms to maintain or increase their competitiveness in the increasingly globalized economy is to upgrade their production. The concept of industrial upgrading – making better products, making them more efficiently, or moving into more skilled activities – has often been used in the literature on competitiveness (Porter 1990; Kaplinsky 2000). Similarly, in the global commodity chain and global value chain approach, the concept of industrial upgrading refers to the process by which economic actors – nations, firms, and workers – move from low-value-added to relatively high-value-added activities in global production networks (Gereffi 2005).

Gereffi (1999) has argued that upgrading could take place at different geographical scales: first, at the factory level, upgrading could be shifting from production of simple and cheap to complex and expensive items, as well as from small to large orders; second, within interfirm networks, upgrading involves shifting from mass production of standardized items to flexible production of differential items; third, upgrading, within local or national economies, also refers to moving from simple assembly and processing with supplied materials to higher-value-added and more complicated original equipment manufacturing (OEM) and original brand manufacturing (OBM) that involve a higher degree of integration of forward and backward linkages at the local or national level; finally, at the regional level, upgrading involves moving from interregional trade flows to a more fully developed intraregional cooperation incorporating all parts of the commodity chain from supply of raw materials, through production and distribution, to consumption. Different mixes of government policies, institutions, corporate strategies, technologies, and worker skills are associated with upgrading success. The implicit normative expectation in this work has been that firms and countries that upgrade produce higher-value-added products and/or take on more sophisticated functions over time and this captures a higher proportion of value in the chain, while firms and countries that are unable to add functions, change products, or change their relationship to others in the chain may see orders decline, their costs squeezed, and their employment prospects diminish.

Different types of industrial upgrading

Industrial upgrading is vital for creating possibilities to enhance value and thus for creating possibilities for economic development. Industrial upgrading debates have framed this issue in terms of, first, the structure of production, where upgrading is considered as an increase in the complexity of production activities and a movement from assembly to OEM, original design manufacturing (ODM), and OBM; second, intrasectoral upgrading, mainly involving a greater use of forward and back linkages along the commodity chain; third, moving from labor-intensive production activities to
more capital-, technology-, and skill-intensive economic activities, based on organizational learning in order to improve the position of firms and nations in global production networks; fourth, value capture, particularly as firms, regional networks, and local and national economies are moving from low-value-added to high-value-added activities along the value chain (Gereffi 1999).

Humphrey and Schmitz (2002a) have attempted to systematize these forms of upgrading in four ways: process, product, functional, and chain upgrading.

1. **Process upgrading**: incorporation of more sophisticated technologies into production and/or re-engineering production lines, such as in lean manufacturing.
2. **Product upgrading**: producing higher-value products.
3. **Functional upgrading**: moving to higher-value functions.
4. **Chain upgrading**: leveraging expertise gained in one industrial sector to enter a new sector.

Although helpful as a starting point, there are some difficulties in working with the four-type classification of upgrading highlighted above (Ponte and Ewert 2009). First, it is sometimes difficult to distinguish between product and process upgrading, especially as, in some industrial sectors, the introduction of new processes also leads to new products. Second, in some value chains, the capability to provide a wider range of related products with different specifications is a more important aspect of upgrading than the capability to produce higher-value-added products. Third, to operate profitably in some value chains, some other factors are also of central importance, such as scale economies, which require firms not only to achieve process upgrading, but also to move from small to large orders. However, moving from small to large orders often entails product downgrading since low-value-added products sold in larger amounts may easily enable some firms to aggregate orders to increase the volume of sales and thereafter to become more profitable. Finally, the status of “intersectoral” upgrading is unclear, as it relates to a trajectory of upgrading, while the other three categories describe what aspect of a given business is being upgraded.

Industrial upgrading in clusters and global value chains

This section examines the impact of governance over industrial upgrading by drawing on two sets of literature. On the one hand, the industrial district literature emphasizes the advantages that come from local-level cooperation; on the other hand, global value chain theory emphasizes the importance of cooperation between firms in developed and developing countries (Humphrey and Schmitz 2002b). These two lines of work both highlight the importance of upgrading so as to increase competitiveness in the globally distributed production networks. Likewise, they both pay special attention to the ways in which industrial upgrading is shaped by governance, which is defined as coordination of economic activities through nonmarket relationships. Governance plays a critical role in the processes of knowledge transmission, diffusion, and spillover, as well as in the processes of innovation and firm learning. However, these two lines of work see governance operating at different scales, and therefore with different implications for upgrading opportunities, mechanisms, and prospects for firms in developing economies.

Governance and upgrading in clusters

The importance of governance and inter-firm cooperation at the local level has been
emphasized by cluster studies, since actors are likely to enhance ties with partners that they have the most interactions with and are geographically close to, especially if those ties involve complicated interactions and trust. Specialized labor markets and exchanges of tacit knowledge are especially dense, efficient, and vibrant when it is possible for agents to have face-to-face interaction. Competitiveness of firms partly hinges on their embeddedness in local clusters. Cluster studies therefore focus on the local sources of industrial upgrading driven by local economic agents and interfirm synergies which generate collective efficiency and knowledge spillovers, embeddedness, and traded and untraded interdependencies. The cluster literature in general and the industrial district literature in particular claim that knowledge generation, transfer, and diffusion among firms within the same cluster are based not only on interfirm synergies – the so-called industrial atmosphere – but also on policy networks and public/private governance (Humphrey and Schmitz 2002b). The cluster literature sees governance mainly as operating at the local level, and local public and private governance designed to support industrial upgrading are understood as an essential complement to the incidental interfirm synergies arising from agglomeration.

This set of literature recognizes the importance of firms’ and clusters’ external linkages, particularly as some firms and clusters are export-oriented. However, the nature of these external linkages is characterized explicitly or implicitly as arm’s length, and these linkages are therefore often treated as exogenous factors in the analysis of upgrading. For instance, conventional cluster studies tend to portray industrial districts as incorporating all phases of the commodity chain or at least the full range of activities to produce finished products. Even when some production activities are outsourced to lower-cost locations outside the district, core functions remain in the district. In addition, these districts are capable of acquiring and absorbing codified knowledge from external sources if needed and therefore able to innovate and improve their performance. In short, to achieve industrial upgrading, local enterprises need to rely to a large extent on local sources of innovations, which are not only the result of knowledge spillovers and interfirm synergies within a local cluster but are also fostered by policy networks of public and private actors. The sources for upgrading come from within the cluster at the local level, while firms’ and clusters’ external linkages are considered as less important and remain the least developed dimension in such analyses.

**Governance and upgrading in global value chains**

The literature on global value chains, which believes that the design, production, and marketing of products involves a chain of activities divided between different enterprises often located in different places, takes a very different view of knowledge transmission and interfirm linkages. It also focuses on upgrading and its relationship with governance, but the upgrading process has been attributed to the knowledge and information flow within global value chains from lead firms to their suppliers (Gereffi 1999). In other words, upgrading refers to acquiring new capabilities and increasing competencies through participating in particular chains. Gereffi (1999) has defined governance structures in global value chains as authority and power relationships that determine how financial, material, and human resources are allocated and flow within a chain, in order to stress the important role played by global lead firms that undertake the functional integration and coordination of internationally dispersed activities. In this sense, the global value chain literature often argues that the actions and
motivations of global buyers are the key causal forces in the organization of global contracting systems, and further determine the upgrading opportunities and prospects of local producers. Interfirm relationships within the value chain are of central importance no matter where the firms are located, while interfirm relationships within the locality are given much less attention.

First of all, by highlighting the asymmetric power relationship between local producers and global lead firms, the global value chain literature calls attention to the fact that the organization of global contracting and sourcing systems is heavily affected by the actions and motivations of global lead firms, as the latter firmly control the access to the end market in the North. Second, local producers could learn from global lead firms about how to transform inputs into outputs more efficiently and quickly (process upgrading). Third, product upgrading is also facilitated, particularly in the process of “organizational succession” through which local producers begin supplying low-end global lead firms and then move up to lead firms catering for the high end of the market (Gereffi 1999). In other words, this succession enables local producers to achieve product upgrading as they meet global lead firms’ demands for higher-end, more complicated and sophisticated products. Finally, global value chain analysis also emphasizes a network of labor and production activities that are often globally distributed and interconnected. Firms and localities may specialize in a narrow range of functions. The process of moving from low-value-added to high-value-added functions and acquiring new functions that generate higher value is a critical part of functional upgrading strategy. In all types of upgrading, learning from global buyers is especially crucial for local producers in the developing countries that are new to the global economy and have no prior experience in supplying end market in the North. Although the knowledge and information flow with global value chains has high value for local firms, the role played by the global buyers in fostering and supporting local firms’ upgrading process is less clear. Gereffi (1999), mainly focusing on East Asia, makes a rather optimistic conclusion, stressing the role of global lead firms that almost simultaneously promote process upgrading, product upgrading, and functional upgrading among local producers. However, since the seminal work of Humphrey and Schmitz (2002b), it has been increasingly accepted that, sometimes, global buyers both contribute to the process and product upgrading of local suppliers and place limits on functional upgrading and market diversifications. Functional upgrading has been viewed as encroaching on the core competence of the lead firms, while process/product upgrading remains in the interests of global buyers. In other words, global value chain governance can create barriers to local upgrading, and power relations may inhibit upgrading and limit knowledge transfer within the chain. While the conclusions differ, they have a common emphasis on the global/external powers that influence upgrading in the cluster.

Industrial upgrading in cluster and global value chain studies

The cluster literature highlights the need to improve interfirm cooperation and governance at the local level, and sources for innovation and upgrading are seen largely to derive from within the locality. Links with the wider world are frequently acknowledged, but they are often underestimated. Overall, the external world is characterized as a market presenting competitive challenges that must be met through improved organization and effort within the cluster.
Risk-coping mechanisms within the cluster make these challenges easier to meet.

In contrast, the global value chain approach highlights links with the external world, leaving the locality largely unexplored. While it does not exclude possibilities for local institutions to affect outcomes, the role of business associations and local interfirm cooperation in competitiveness and industrial upgrading has been underestimated. Upgrading takes place as a result of learning by exporting and by participating in global value chains. Interfirm knowledge transfer from global lead firms to local producers also helps to attenuate the risks for the latter of entering global markets. Interfirm cooperation within the chain rather than within the locality is considered as the source of competitive advantage and industrial upgrading.

How value chain governance influences upgrading in the clusters

Studies on clusters and industrial districts focus on the significance of linkages within the locality while the global value chain analysis stresses the impact of interfirm cooperation and governance at the global level over industrial upgrading. There is a need to reconcile these two perspectives, particularly as local clusters are increasingly integrated into globally organized production systems. In order to answer this question, Humphrey and Schmitz (2002a, 2002b) have distinguished between four types of upgrading (see above) and four types of chain.

1 *Arm’s length market relations*: because the product is standard (or easily customized) and buyers’ requirements could be easily met by a wide range of suppliers, buyers and suppliers do not need to collaborate closely.

2 *Networks*: the relationships between buyers and suppliers are more or less equal and often characterized by reciprocal interdependence. Buyers and suppliers collaborate in product definition and innovation, and combine complementary competencies. Buyers may specify certain standards (e.g., quality and speed), but the risk to buyers is low because suppliers are capable of meeting them.

3 *Quasi-hierarchy*: buyers play a dominant role in product definition and exercise a high degree of control over suppliers. They have to face losses due to the suppliers’ performance failures. In other words, the high level of control of buyers over suppliers results not only from the buyers’ dominance in product definition, but also from buyers’ doubts about the competence of suppliers. Where high competence of the supply chain is not generalized, buyers tend to invest in specific suppliers and exercise control over the first-tier suppliers and, in some instances, indirect suppliers along the chain.

4 *Hierarchy*: suppliers are directly owned by buyers in the chain.

Different forms of chain governance have different upgrading implications: (i) In a quasi-hierarchical chain, global buyers both contribute to process and product upgrading of local suppliers and place limits on functional upgrading and market diversifications. Buyers are responsible for product definition and logistics, therefore investment and risks involved in entering export markets are considerably reduced. However, functional upgrading is often hindered, because functional upgrading of developing country producers encroaches on the core competence of their main buyers. (ii) In a market-based chain, process and product upgrading tend to be slower, because, when entering export markets, developing country producers do not receive information about market requirements nor
assistance to meet these requirements in the way that firms in quasi-hierarchical chains do. Functional upgrading is not necessarily easy but there is no barrier purposefully created by buyers. Compared with quasi-hierarchical chains, market-based chains provide more space for functional upgrading. In some instances, local governance (public, private, and public–private) acts as a substitute for some of the governance processes seen in quasi-hierarchical chains. While this probably slows down process and product upgrading, in the longer run it might create local capabilities that would enable functional upgrading and market diversification strategies.

(iii) Although insertion into a network-based chain offers the most favorable conditions for upgrading, suppliers in developing countries are often unlikely to participate in this type of chain because of their low level of competencies. It is therefore particularly important to take into account the interrelationship between interfirm learning within global value chains and within clusters. Insertion into global value chains does offer favorable conditions for suppliers’ upgrading and competency improvement. However, entry into network-based global value chains also requires investment in upgrading and related services within the cluster at the local level. Rather than substituting local-level investment, the entry into chains characterized by network governance is reliant on intensive investment in upgrading and competency improvement within the clusters at the local level. Finally, there are other determinants of local upgrading, such as the strategic intent of local firms, that is, their conscious effort to learn and their willingness to make the required intra- and interfirm investment, and the supportive policy environment created by various levels of government.

Further development of the concept of industrial upgrading

While conventional upgrading studies focus on manufacturing and use the concept of “industrial upgrading,” recent work on upgrading has moved beyond manufacturing and increasingly sought to include sectors related to agriculture and services where the term “industrial upgrading” is less appropriate. Industrial upgrading is therefore replaced by a more generic concept – “economic upgrading” – which applies across sectors (Barrientos, Gereffi, and Rossi 2011). Whereas most studies on upgrading, in particular from the global value chain perspective, emphasize promoting economic development and growth, few have focused on what such upgrading means for social wellbeing and living standards, including not only wages and working conditions but also rights and entitlements of workers. A growing body of evidence calls for more attention to a more extensive definition or conceptualization of upgrading where criteria beyond economic development are encompassed. This gives rise to work on social upgrading, which is the process of improvement in the rights and entitlements of workers as social actors, and of enhancement of the quality of their employment (Barrientos, Gereffi, and Rossi 2011). It can be subdivided into two components: measurable standards (e.g., type of employment, wage level, social protection, and working hours) and enabling rights (e.g., freedom of association, the right to collective bargaining, nondiscrimination, voice, and empowerment). Recent upgrading studies have increasingly stressed that economic upgrading in value chains does not automatically translate into social upgrading through better wages and working conditions, and the links
between economic and social upgrading are often complex.

**SEE ALSO:** Global commodity/value chains; Global production networks; Globalization; Governance and development; Industrial agglomeration; Industrial districts; International division of labor

**References**


Industrialization is the process by which the industrial sectors come to play a dominant role in a national economy and an agrarian society transforms into an industrial society. The most pervasive feature of industrialization is the transformation of the structure of production, whereby the industrial sectors typically grow more rapidly than agriculture (Chenery, Robinson, and Syrquin 1986). This transformation is accompanied by similar changes in the structure of demand, international trade, and types of occupation in the industrial sectors. The process of industrialization is closely related to historical, technological, and social-economic factors.

The process of industrialization

The Industrial Revolution that occurred in Great Britain in the eighteenth century is considered to have been the beginning of the process of industrialization (Mantoux 2013). In developed countries, industrialization has a history of more than 200 years. Generally, there are three main stages in the process of industrialization.

In the early stages, light industries and labor-intensive industries (e.g., the textile industry) grew at high speed and became the leading industries in the national economy. Technical improvements were the driving force behind the process, rather than the result of it (Mantoux 2013). The invention of the steam engine was the symbol of the first Industrial Revolution, and it was widely used in various industries, such as cotton spinning and weaving, and paper-making. Therefore, autarkic production in small volumes was gradually replaced by mass production in factories that could produce goods for larger and more distantly located markets. The productivity of the society was hugely improved, specialized industrial regions grew up, and internal and international trades emerged (Smith 1863).

In the middle stage of the process of industrialization, heavy industries such as raw materials, energy, electronic, chemical, and mechanical industries became important in the economy. The dominant industries at this stage were capital- and technology-intensive. Electric power was widely used and the production process was highly mechanized. Rapid progress in industrial organization and management meant that enterprises could take advantage of the large volume of production and enjoy economies of scale. Therefore, multi-factory, multilocalational, and multioperation enterprises emerged, and these large-scale enterprises become influential and had a great impact on society. Moreover, industrialization became significantly concentrated in urban areas (Kim 2006).

With the advancement of technology after World War II, industrialization entered a new phase in which production conditions and productivity greatly improved. The knowledge-intensive industries, such as information and communication technology,
biotechnology, nuclear technology, space technology, new materials, micro-electronics, and so on, became leading industries. The high-tech industries grew rapidly, replacing traditional industries and becoming dominant in the national economy. With the continuing growth of productivity, fewer workers were needed in some industries. In addition, with the increase in income, the proportion of it spent on food in relation to total consumption (the Engel coefficient) declined sharply. Therefore, consumption patterns changed and service sectors boomed. Industrialization entered its final stage.

Bell (1973) called the new stage postindustrialization, where the main industries shifted from the industrial sector to the service and advanced knowledge and technological sectors. From the 1970s, some developed countries began to enter the era of postindustrialization (Keinath 1985). Lever (1991) described the shift from the manufacturing to the service sector as deindustrialization. In this stage, manufacturing sectors have a lower share of the total output or employment, while the service industry becomes increasingly more important in the national economy.

Some studies have attempted to measure and evaluate the stages of industrialization according to the evolution of industrialization in developed countries. Hoffman (1969) measured a country's stage of industrialization using an index known as the Hoffman index, which is the difference between the net output of consumer goods and that of the capital goods industries. Kuznets and Murphy (1966) and Chenery, Robinson, and Syrquin (1986) defined the stage of industrialization according to the industrial structure: that is, the proportions of primary industry, secondary industry, and tertiary industry in the national economic output.

Pathways to industrialization

There are different pathways to industrialization for each country as a result of different political, economic, and social factors. Given the importance of industrialization to the national economy, many countries, especially developing countries, have adopted strong policies to encourage or enforce industrialization. There is no single successful model of industrialization with unique characteristics. However, the pathways of industrialization in different countries can be divided into categories according to some general factors.

Market-oriented industrialization: Europe and North America

Great Britain was the cradle of industrialism (Smelser 2013). In the eighteenth century the process of industrialization via innovation in manufacturing began with the Industrial Revolution in the northwest and Midlands of England. The commercial expansion, redistribution of land, and growth of new industries helped facilitate the market economy in Britain in the preindustrialization period (Mantoux 2013). In the eighteenth and nineteenth centuries Britain experienced a massive increase in agricultural productivity, and unprecedented population growth. Thus, the preconditions for industrialization, including the accumulation of capital, a commercial economy, and a large workforce, were in place. The invention of the steam engine accelerated the process of industrialization. From then on, Britain entered the era of industrialization and, by the 1830s, it was the first country in the world to achieve a high level of industrialization. Later, the industrialization process spread to other parts of Europe, for example France, Germany, and Russia.
Industrialization in the United States can be seen as a further development of the model of Britain. The availability of capital, free market, plenty of land, an educated labor force, and so on all contributed to the United States’ rapid industrialization. Industrialization in the United States also started in the textile industry with innovations in machinery. By using technologies imported from Britain, and later relying on domestic technology and management innovation, the United States realized its potential for large-volume production in a relatively short period and became a powerful industrial county.

Import substitution: industrialization in Latin America

The less developed economies of countries in Latin America sought to develop their industrial capacity by producing as many sorts of products as possible, so that they could reduce the import of goods from other countries. This strategy is called import substitution. Since the mid-to late twentieth century, import substitution (inner-directed) industrialization has become the main development model in most Latin American countries, such as Brazil, Argentina, and Mexico (Baer 1972). In order to clear the obstacles to industrialization in Latin America (James 1950), states took measures to improve and protect domestic industrial capacity by subsidizing the vital industries, imposing high taxes on imports, and using other trade protection policies. These countries developed their economies by relying on their own labor, raw materials, and technology as much as possible to produce what was needed for the domestic market, and tried their best to avoid importing manufactured goods. This pattern of industrialization proved to be a failure in fostering social and economic development in Latin America in the long term.

Export-oriented industrialization: East Asia

Japan started its path to industrialization after the Meiji Restoration and achieved a high level of industrialization, similar to that in Western countries. After World War II, Japan and other Asian economies adopted an export-oriented strategy to catch up with the developed countries and to achieve industrialization. In this model, firms are encouraged to export goods in which their country has a comparative advantage.

The late twentieth century saw the phenomenon of the rise of the Four Asian Tigers (Hong Kong, Singapore, South Korea, and Taiwan), all of which had adopted an export-oriented strategy and successfully achieved industrialization. The success of these economies also relied on other factors, including a stable political environment; a well-structured society; a strategic location; a large amount of foreign investment; a skilled, motivated, and cheap workforce; a competitive exchange rate; and low custom duties, among others. These economies took advantage of developed countries shifting their labor-intensive industries to developing countries. They adjusted their economic development strategy, given that they were able to offer sufficient cheap and high-quality labor (Brohman 1996). They encouraged export and also welcomed direct foreign investments: the two reinforced each other. Their success in attracting foreign investment further facilitated the export economy. All in all, the Four Asian Tigers succeeded in rapid industrialization and economic growth by adopting an export-oriented strategy.

The unevenness of industrialization across the world

Despite the rapid industrialization of some less developed countries in the past several decades, the global unevenness of industrialization has
INDUSTRIALIZATION

persisted. Most developed countries have been in the stage of postindustrialization for a long while. However, some countries in the Global South are still in the early stages of industrialization. The existing global governance of economy and politics has impeded the industrialization of some less developed countries. The industrial sectors in these countries have faced competition and exploitation as a result of the dominance of multinational corporations, which are mostly based in the rich countries of the Global North and increasingly in some emerging economies.

Entering the twenty-first century, emerging countries, including Brazil, Russia, India, China, and South Africa (BRICS), are becoming increasingly influential in the world economy because of their rapid and successful industrialization (Mudambi 2008; Winters and Yusuf 2007). Both India and China have achieved a high economic growth rate and their high-tech sectors have also made rapid progress (Arora and Gambardella 2006); these sectors have also started to make significant investments in developing countries as well as developed countries. China has become the world’s factory. Industrial clusters have grown up in China, and a large number of cities have seen significant growth of industry (Fan and Scott 2003; He and Pan 2010; He, Wei, and Xie 2008). India has been famous for its competitive high-tech sectors such as bioengineering, pharmaceuticals, and information and communication industries.

In addition, the geography of industrialization within some countries is also significant, especially in some large developing countries such as China and India. The achievement of a more evenly distributed geography of industrialization within these countries depends very much on the convergence policies and their implementation. However, unlike global unevenness, the variation of industrialization within countries is easier to overcome.

Consequences of industrialization

Industrialization has completely changed the world. On the one hand, it has brought great changes for society. One of the most significant consequences is urbanization. Urbanization is the fundamental transformation of society caused by industrialization (Kim 2006). In the process of industrialization, the major resources are allocated to the industrial rather than agricultural sector (Kuznets and Murphy 1966). The system of mass production in industrial activities is also associated with a large volume of the workforce migrating from rural areas to cities to work in the industrial sectors. The concentration of labor in cities has given rise to large towns, resulting in large-scale urbanization (Scott and Storper 1992). The spatial pattern of the metropolis is created by the intricate ramifications of the social division of labor, the transactional structure of production, and the dynamics of local labor market formation (Scott 1986). During the process of urbanization, family structure (Wong 1975) and the rural landscape (Claval 1988) also undergo significant changes.

Another consequence of industrialization is economic growth and regional development. Industrialization has greatly improved productivity and can lead to significant economic growth and regional development. Manufacturing industries can provide a large number of jobs and generate a huge amount of tax and wealth for an economy. Creating industrial agglomerations is one way to boost industrialization and improve economic growth and regional development (Scott and Storper 1992). The process of industrialization is also closely related to technological innovation. For those who benefit from it, rapid industrialization has improved the standard of living enormously, and it underpins further advancements in science and engineering.
Industrialization has greatly transformed human civilization. In the transformation from an agricultural-based economy to an industrial-based economy, traditional, small, family-run workshops have been replaced by large-scale production in factories, and an agricultural civilization has changed to an industrial civilization. With the significant rise of their capacity to change the world, people’s attitudes to nature have also changed with the process of industrialization.

On the other hand, large-scale industrialization has also caused many problems. The first is population explosion. With great improvements in productivity, the population also grows very quickly. A larger population needs a greater volume of industrial products for consumption, as well as other natural resources, which may threaten the sustainability of the planet. In addition, the concentration of the population in urban areas places great pressures on a society. A shortage of housing, expensive health care and education, insufficient clean water and food, and high crime rates are all serious problems that face large cities.

Industrialization also leads to environmental problems. Environmental pollution such as air pollution, water pollution, noise pollution, and so on is widespread. Most pollution is released from industrial facilities or is related to industrial production. Environmental pollution may harm human health as well as jeopardize the survival of other species and their habitats. Other environmental problems caused by industrialization include deforestation, land degradation, extinction of species, and global warming.

Moreover, rapid industrialization exacerbates the shortage of natural resources. Fossil fuels and other mineral resources are basic inputs for industrial production, but stocks of these natural resources are limited. The price of these raw materials will eventually increase to very high levels, which may lead to severe problems for industries. For example, an oil crisis that originates locally could lead to serious consequences for the global economy (Akins 1973).

It has become a key challenge for humans to achieve sustainable development in the future. The benefits of industrialization are obvious, but many countries are still suffering from poverty and a lack of industrialization. The pursuit of rapid industrialization in these countries is reasonable and should be encouraged, but the side effects of industrialization are also alarming. All countries need to work together to solve or at least alleviate the problems caused by the process of industrialization.

SEE ALSO: Deindustrialization; Environment and urbanization; Fordism/post-Fordism; Industrial agglomeration; Industrial districts; Industrial restructuring; Industrial upgrading; Rural/urban divide

References


**INDUSTRIALIZATION**


Inequality

Michael Webber  
University of Melbourne, Australia

Inequality, according to the Oxford English Dictionary, is the difference in size, degree, or circumstances (www.oxforddictionaries.com) or more simply, the condition of being unequal (http://dictionary.reference.com). The latter adds the useful information that equality is correspondence in quantity, degree, rank, or ability. In geography, references to equality and inequality are usually concerned with qualities such as wellbeing, health, income, wealth, or social status. These qualities are commonly measured at the level of individuals or groups within an area (thus, interpersonal inequality or household inequality in China or in Rome) or as regional averages (e.g., interregional inequality or international inequality). Of itself, the concept of inequality carries no judgment about the fairness of the differences that are observed nor about the origins of those differences. Equity, by contrast, does involve such judgments, referring as it does to the quality of being fair or impartial (www.oxforddictionaries.com). Inequality is about difference; equity is about fairness.

Inequality bears some relation to another important concept, poverty. Dearth (sometimes called absolute poverty) is being unable to satisfy one’s basic human needs, including for food, safe water, sanitation, health, shelter, education, and information. It is commonly measured by a set standard that is constant over time and between countries: the World Bank currently sets this standard at $1.25 per person per day (in US prices). Absolute poverty means the inability to participate effectively in society. In rich countries, such as Australia, effective participation in society means having sufficient clothes to present for school or work, having access to news and information through computers and television, having money to pay for transport, housing, and so forth, all of which cost a single person about $400 per week in 2010. In principle, both dearth and absolute poverty are unrelated to inequality: high-income countries, with few poor people, may have high inequality; low-income countries, with low inequality, may have many poor people. Relative poverty, however, is closely related to inequality. Understood as the daily perceived contrast between the lives of the poor and the lives of their peers around them, relative poverty is often measured as the proportion of the population with incomes less than some proportion (such as 60%) of the median income. Such a measure is independent of the average income of a society: both high- and low-income countries can exhibit high rates of relative poverty.

Both words, equality and inequality, first appeared in late Middle English: that is, the late fourteenth century, but before that the word, and so the concept, were available to the middle and upper classes through Latin. However, it was not really until the eighteenth century, and the Enlightenment in Europe, that the tools with which to think about inequality were first developed. These tools included the notion of social reform or progress and Locke’s epistemology of
INEQUALITY

empirical observation, which in turn gave rise to the emergence – as a program, if not yet as a set of results – of the human sciences. Especially important among these new human scientific works for the study of inequality was Rousseau’s Discourse on Inequality (Rousseau 1755). In this Discourse, Rousseau argued that in “primitive” human societies there were small differences in wealth and status between families; however, once the division of labor emerged as a way of raising agricultural and metallurgical output, so private property, social class, and, therefore, inequality were introduced too. Apart from the original sources, a good introduction to Rousseau’s thought is provided by the Internet Encyclopedia of Philosophy (Delaney, n.d.).

Measurement of inequality

To say anything useful about trends in inequality or about differences in inequality between different places, we must be able to measure it. This means in the first place that the quality of interest – health status, for example – must be measured as a scalar variable; ranking measures will not do, for we are interested in the magnitude of the differences between individuals, not simply in the fact of the differences. It means in the second place that we must agree on the variables that will be used to measure the quality of interest. For complex qualities, such as wellbeing, health, or social status, these measures are highly disputed. Even for apparently simple qualities, such as income, consumption, or wealth, alternative ways of measuring them have a tremendous impact on apparent inequality. One specific form of this requirement is the need to agree on whether input or output measures of the quality are more desirable. In political debates, for example, those on the right tend to favor measures that reflect “equality of opportunity” whereas those on the left tend to favor measures that reflect “equality of income or wealth.” Box 1, which lists some of the variables that are used to measure income, reveals just how disparate measures of even a single quality can be. And third, it means that we must agree on how to transform these measures at an individual level into an aggregate measure of inequality.

There are a huge number of ways of measuring inequality in a variable (Coulter 1989 identifies about 50), but four are common: the Gini index, a ratio such as the 20:20 ratio, the Theil index, and the Hoover index (Fleurbaey 2012 provides a thorough introduction to the measurement of inequality). That is, so long as we can provide a number to measure an individual’s, group’s, or region’s wellbeing, health, income, wealth, or social status, then we can calculate the inequality of that variable. Consider the example of income measured at the level of individuals. The Gini index ranges between 0 (complete equality) to 1 (all income is concentrated in the hands of one person). It is perhaps the most commonly used measure of inequality, but it does have problematic properties (for example, the Gini coefficients of different regions within a country cannot be added to derive the Gini coefficient for the country as a whole) and quite different income distributions can generate the same coefficient value. A whole body of literature is devoted to identifying means of discriminating between income distributions that are more equal (or less) at low incomes and less (or more) equal at high incomes. The income ratios are easy to comprehend: How much do the richest 20% have compared to the bottom 20%? The ratio starts at 1 (complete equality) and has no upper bound. But information about the majority of the population (the 60% in the middle) is lost. The Hoover index is simple to calculate and understand: What
Box 1 Variables used to measure inequality of income

1 Equality of opportunity versus equality of outcome: Does everyone have the same chance of earning a high income versus does everyone in fact have the same income?

2 Household income versus individual income: Do we simply calculate inequality from total household incomes (which implicitly assumes that everyone in the household has the same income) versus do we calculate the actual incomes of each person in a household separately? Do we include children and others still in education? Do we include retired people? Do we include non-employed homemakers? (Note: measuring the income of everyone in a household poses great conceptual and practical difficulties. Is a standard of living shared equally? And whose account of income sharing within a household should be believed?)

3 Total income versus income after taxes and transfers: Do we include the imputed value of owner-occupied housing? Do we include the imputed value of subsistence production and directly exchanged goods? (Note: inequality of earnings is usually computed from data on individual employment-related income, before tax, excluding transfers and imputed incomes.) These differences are important because taxes and transfers form different shares of total income at different income levels; they are also different between countries. In different countries, housing, subsistence production, and direct exchange form different shares of total income.

4 When making international comparisons (and some within-country ones too), do we measure incomes at official exchange rates or at PPP rates? (Note: PPP = purchasing power parity. If a Chinese woman (say) goes to the United States and purchases an average standard of living, it costs her RMB 6.06 to buy each US dollar (USD) worth of commodities. This is the official exchange rate. If she bought an equivalent standard of living inside China, each USD of purchases would cost less than this, because prices in China are on average lower than in the USA, relative to exchange rates. According to the World Bank’s development indicators (http://data.worldbank.org/?display=default), in 2012, the PPP exchange rate between USD and the RMB was 4.06. Roughly, incomes in China are one-third higher than official exchange rates suggest. The number, 4.06, is the PPP exchange rate between RMB and USD.)

Proportion of the total income would have to be moved from the richest people to the poorer people in order to equalize incomes? It ranges from 0 (no income has to be shifted: that is, all incomes are equal) to 1 (all income has to be shifted: that is, one person earns all the income).
INEQUALITY

The *Theil index*, though less easy to calculate than the others, has a desirable property, notably decomposability, which is extremely useful in geographic analyses of inequality. This property means that the Thiel index of inequality within a country equals the index of inequality between provinces plus the average of inequalities within each province (weighted by provincial shares of national income). Here is how to calculate and interpret these four measures of inequality.

Suppose that there are $N$ groups of people, ranked according to their income. Group 1 consists of people with the least income; group 2 has those with the next lowest incomes; ...; and group $N$ consists of the people with the highest incomes. The groups are not necessarily of the same size, so let $P_i$ denote the number of people in group $i$, for $i = 1, 2, \ldots, N$. The total population is

$$P = \sum_{i=1}^{N} P_i$$

The proportion of the population in group $i$ is then

$$p_i = \frac{P_i}{P}$$

Let $V_i$ denote the total income earned by all the people in group $i$. Again, the total income earned by all the people is

$$V = \sum_{i=1}^{N} V_i$$

The proportion of the total income earned by group $i$ is then

$$v_i = \frac{V_i}{V}$$

Finally, let $s_i$ denote the cumulative shares of total income earned by the groups up to and including group $i$: $s_0 = 0$; $s_1 = v_1$; $s_2 = v_1 + v_2$; and so on. With these definitions, the four measures are:

$$G = 1 - \sum_{i=1}^{N} p_i [s^i - 1 + s_i]$$

$$H = \frac{1}{2} \sum_{i=1}^{N} |v_i - p_i|.$$

$$T = \frac{1}{N \cdot \ln(N)} \sum_{i=1}^{N} \left[ \frac{v_i}{p_i} \cdot \ln \left( \frac{v_i}{p_i} \right) \right].$$

$$z : \text{z ratio} = \frac{V_N}{V_1}$$

provided that group 1 and group $N$ contain the same proportions of the population (e.g., $z\%$).

Measures of inequalities in health or wellbeing can, in principle, use the same indexes. However, the literature about health inequalities is dominated by a concern with the sources of the observed differences in health status; thus many measures of and studies about health inequality rely on indexes of health status that are calculated for separate groups, classified for example by gender, age, ethnicity, or income. A good introduction to this literature is provided by Asada (2010).

All of these differences of opinion about good measures of a quality and about good indexes of inequality in those measures mean that trends in inequality or differences in inequality between places are hardly robust. Different variables and different indexes can provide quite different results. It is desirable, therefore, that studies of trends in inequality or spatial differences in inequality use different measures and different indexes to identify the extent to which the results depend on the variables and the indexes used. In understanding inequality as a quantitative variable, robustness demands a variety of measures and indexes, and caution must be used.
in interpreting arguments that rely on a single variable or index.

Furthermore, all these measures are independent of the mean. That is, inequality depends on the dispersion of an index, not on its average value. Inequality can be high in high-income and in low-income societies, and low in both, too. Thus, some argue that in high-income countries, inequality (and relative poverty) is less important than in low-income countries: in rich countries, everyone has enough, it is said, so being relatively poor is not so important. This argument, though, faces two problems. First, even if everyone has enough, it is not clear why some should have so much more (this is the equity question).

Second, there is evidence that happiness within society depends on one’s relative position as well as on one’s absolute position, at least for those who have adequate incomes – the striving for more material goods is driven by relative scarcity as well as simply absolute poverty.

Global inequality

Figure 1 suggests that levels of income inequality are highest in countries in southern Africa (South Africa, Lesotho, Botswana, and Namibia) and some other countries in sub-Saharan Africa (Sierra Leone and Central African Republic).
INEQUALITY

High levels of inequality (Gini over 0.44) are also recorded by the United States, Mexico, most countries in South America (but not Venezuela), southeast Africa, China, Iran, Malaysia, Singapore, and the Philippines. At the other extreme, the lowest levels of inequality are recorded by Sweden, Slovenia, Montenegro, Hungary, Denmark, and Norway. Most European countries apart from the United Kingdom have Gini indexes of less than 0.36, as do Australia, Canada, several North African countries, Pakistan and Bangladesh. However, remember the caveats about data, measures, and indexes. The website http://commons.wikimedia.org/wiki/File:Wealth_gini_2013.svg also contains a map of Gini coefficients for wealth rather than income. Inequalities of wealth are far higher than inequalities of income. There are also discrepancies in the relative ranking of countries by inequalities of wealth and of income: Russia and the United States both record wealth coefficients over 0.85; the Scandinavian countries and Germany also exhibit high wealth coefficients (in contrast to their low coefficients of income). But southern America and southern Africa both exhibit high inequalities of wealth, as of income.

There is also inequality between countries. According to Davies et al. (2008), North America, with 6% of the world’s adult population, controls 34% of the world’s wealth; the rich countries of the Asia-Pacific (principally Japan, South Korea, Taiwan, Australia, and New Zealand) have 5% the world’s population, but 24% of the world’s wealth; while Europe contains 15% of the world’s population, but 30% of the world’s wealth. In contrast are Latin America (8% of the adult population, 4% of the wealth), Africa (10% and 1%), China, India, and other Asia-Pacific countries (50% and 6.5%). These data are at official exchange rates, which do slightly overstate concentrations of wealth in the richest countries. Trends in inequality of incomes between countries are confusing (Milanovic 2012).

1 If we consider countries as the units of observation and calculate the Gini coefficient of their average household incomes, then we see that this rose sharply after World War II (from about 0.47 to 0.52), was stable in the 1960s and 1970s, and then began to rise again with the onset of neoliberal globalization after 1980 (reaching 0.58), before declining since 2000 to 0.55.

2 If we consider countries as the units of observation and calculate the Gini coefficient of their average household incomes, weighted by their share of the world’s population, then the Gini coefficient has fallen slowly from about 0.66 in the early 1950s to about 0.63 in 1990 and then more sharply to 0.53 in 2010.

3 The sharp difference in these trends is driven by the fact that average household incomes in two of the poorest countries in 1950 (China and India) have grown relatively quickly, especially since the mid-1990s.

However, these relatively small changes in between–country inequality since the 1950s come after very rapid and much larger increases during the nineteenth century and the first half of the twentieth century.

Another popular method of measuring the global distribution of wellbeing is the Human Development Index (HDI), produced by the United Nations Development Programme and published in its annual Human Development Report (UNDP 2013). This is a measure that combines measures of life expectancy at birth, average years of schooling, and gross national income per capita. Until 2009, these measures were normalized (so that each had a range between 0 and 1) and then the average taken,
Figure 2 Human Development Index by major world regions, 1975–2004.

Note: To develop a sense of scale, HDI indexes in 2013 have the following components, on average:

<table>
<thead>
<tr>
<th>HDI</th>
<th>Life expectancy at birth (years)</th>
<th>Mean years of school (years)</th>
<th>Expected years of schooling</th>
<th>GNI/capita (2005 $PPP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.91</td>
<td>80</td>
<td>11.5</td>
<td>16.3</td>
<td>33 391</td>
</tr>
<tr>
<td>0.76</td>
<td>73</td>
<td>8.8</td>
<td>13.9</td>
<td>11 501</td>
</tr>
<tr>
<td>0.64</td>
<td>70</td>
<td>6.3</td>
<td>11.4</td>
<td>5428</td>
</tr>
<tr>
<td>0.47</td>
<td>59</td>
<td>4.2</td>
<td>8.5</td>
<td>1633</td>
</tr>
</tbody>
</table>


so that each index had the same weight in the final index. Since 2010, a slightly different method of normalizing and averaging has been used. Figure 2 illustrates trends in this index, by region. In all regions except sub-Saharan Africa the HDI has been rising since the mid-1970s, at roughly the same rate, though Eastern Europe and the Commonwealth of Independent States (CIS) saw a sharp decline in HDI at the time when the communist regimes were overthrown, and East Asia has been growing slightly more quickly than everywhere else. The real tragedy is in sub-Saharan Africa, where average levels of HDI have hardly changed over 30 years. The data for some individual countries are presented in Table 1 (the full data are available in the annual World Development Reports).

Also important is inequality at a global, individual level. This is difficult to measure in the long run because data on income or
Table 1  Trends in HDI for selected countries, 1980–2012.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Norway</td>
<td>0.804</td>
<td>0.852</td>
<td>0.922</td>
<td>0.955</td>
<td>0.59</td>
<td>0.79</td>
<td>0.32</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Australia</td>
<td>0.857</td>
<td>0.880</td>
<td>0.914</td>
<td>0.938</td>
<td>0.27</td>
<td>0.37</td>
<td>0.23</td>
<td>0.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 United States</td>
<td>0.843</td>
<td>0.878</td>
<td>0.907</td>
<td>0.937</td>
<td>0.40</td>
<td>0.33</td>
<td>0.29</td>
<td>0.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Netherlands</td>
<td>0.799</td>
<td>0.842</td>
<td>0.891</td>
<td>0.921</td>
<td>0.52</td>
<td>0.56</td>
<td>0.31</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Germany</td>
<td>0.738</td>
<td>0.803</td>
<td>0.870</td>
<td>0.920</td>
<td>0.85</td>
<td>0.81</td>
<td>0.53</td>
<td>0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 New Zealand</td>
<td>0.807</td>
<td>0.835</td>
<td>0.887</td>
<td>0.919</td>
<td>0.34</td>
<td>0.60</td>
<td>0.33</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48 Bahrain</td>
<td>0.644</td>
<td>0.713</td>
<td>0.781</td>
<td>0.796</td>
<td>1.02</td>
<td>0.92</td>
<td>0.16</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49 Bahamas</td>
<td>—</td>
<td>—</td>
<td>0.794</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 Belarus</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.793</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51 Uruguay</td>
<td>0.664</td>
<td>0.693</td>
<td>0.741</td>
<td>0.792</td>
<td>0.42</td>
<td>0.68</td>
<td>0.58</td>
<td>0.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52 Montenegro</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.791</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85 Brazil</td>
<td>0.522</td>
<td>0.590</td>
<td>0.669</td>
<td>0.728</td>
<td>1.23</td>
<td>1.26</td>
<td>0.82</td>
<td>0.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95 Tonga</td>
<td>—</td>
<td>0.656</td>
<td>0.689</td>
<td>0.710</td>
<td>—</td>
<td>0.49</td>
<td>0.28</td>
<td>0.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>96 Belize</td>
<td>0.621</td>
<td>0.653</td>
<td>0.672</td>
<td>0.702</td>
<td>0.51</td>
<td>0.29</td>
<td>0.40</td>
<td>0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>96 Dominican Rep.</td>
<td>0.525</td>
<td>0.584</td>
<td>0.641</td>
<td>0.702</td>
<td>1.07</td>
<td>0.93</td>
<td>0.85</td>
<td>0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>96 Fiji</td>
<td>0.572</td>
<td>0.614</td>
<td>0.670</td>
<td>0.702</td>
<td>0.71</td>
<td>0.87</td>
<td>0.43</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>96 Samoa</td>
<td>—</td>
<td>—</td>
<td>0.663</td>
<td>0.702</td>
<td>—</td>
<td>—</td>
<td>0.52</td>
<td>0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>101 China</td>
<td>0.407</td>
<td>0.495</td>
<td>0.590</td>
<td>0.695</td>
<td>1.96</td>
<td>1.78</td>
<td>1.55</td>
<td>1.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>136 India</td>
<td>0.345</td>
<td>0.410</td>
<td>0.463</td>
<td>0.554</td>
<td>1.75</td>
<td>1.23</td>
<td>1.67</td>
<td>1.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>143 Solomon Islands</td>
<td>—</td>
<td>0.486</td>
<td>0.530</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.70</td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>144 São Tomé</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.525</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>145 Kenya</td>
<td>0.424</td>
<td>0.463</td>
<td>0.447</td>
<td>0.519</td>
<td>0.88</td>
<td>−0.33</td>
<td>1.34</td>
<td>1.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>146 Bangladesh</td>
<td>0.312</td>
<td>0.361</td>
<td>0.433</td>
<td>0.515</td>
<td>1.49</td>
<td>1.83</td>
<td>1.61</td>
<td>1.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>146 Pakistan</td>
<td>0.337</td>
<td>0.383</td>
<td>0.419</td>
<td>0.515</td>
<td>1.29</td>
<td>0.89</td>
<td>2.03</td>
<td>1.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>148 Angola</td>
<td>—</td>
<td>—</td>
<td>0.375</td>
<td>0.508</td>
<td>—</td>
<td>—</td>
<td>2.97</td>
<td>2.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>182 Mali</td>
<td>0.176</td>
<td>0.204</td>
<td>0.270</td>
<td>0.344</td>
<td>1.50</td>
<td>2.86</td>
<td>2.45</td>
<td>2.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>183 Burkina Faso</td>
<td>—</td>
<td>—</td>
<td>0.343</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>184 Chad</td>
<td>—</td>
<td>—</td>
<td>0.290</td>
<td>0.340</td>
<td>—</td>
<td>—</td>
<td>1.47</td>
<td>1.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>185 Mozambique</td>
<td>0.217</td>
<td>0.202</td>
<td>0.247</td>
<td>0.327</td>
<td>−0.70</td>
<td>2.00</td>
<td>2.57</td>
<td>2.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>186 Congo, DR</td>
<td>0.286</td>
<td>0.297</td>
<td>0.234</td>
<td>0.304</td>
<td>0.37</td>
<td>−2.34</td>
<td>2.35</td>
<td>2.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>186 Niger</td>
<td>0.179</td>
<td>0.198</td>
<td>0.234</td>
<td>0.304</td>
<td>0.98</td>
<td>1.72</td>
<td>2.42</td>
<td>2.20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

wealth inequalities within many countries are only available for the past 25 years or so. Furthermore, the definitions and meanings of wealth and income differ between countries. So the following history is only tentative. The Gini coefficient of household income (measured in PPP dollars) across all the people in the world, considered as individuals, is about 0.7; it rose slightly in the 1990s and fell equally slightly in the 2000s (Milanovic 2012). This is roughly 0.1 higher than the Gini coefficient for Brazil, itself one of the most unequal countries in the world. Milanovic (2012) also presents a fascinating graph of changes in income according to income – that is, which people saw their income grow during the 1990s and 2000s? His graph is reproduced as Figure 3. Milanovic’s calculations indicate that two groups of people have done well in the 20-year period and two groups have fallen behind.

The people who were the poorest in the world in 1988 have hardly seen their incomes change. These are people of average and below average incomes in poor countries, and people of very low incomes in richer countries.

2 The people who were between the poorest 10% and the wealthiest 30% received large increases in income in this period – 40% or more. This group contains most of the population of China, many Indians, and large numbers of wealthy people in moderately well-off and poor countries. Milanovic calls these people the global middle class.

3 People whose incomes put them in the top 75th–95th percentile witnessed either a drop in income or only a small increase. These include many of the moderately well-off people in Eastern Europe, the former USSR, and Latin America, as well as poorer persons in the rich countries.

4 The people who were in the top 5% of global income earners in 1988 saw their incomes increase by 20–60% (and the more they started with, the greater their increase). This group contains the middle classes of rich countries and the richest groups of people in moderately well-off and poorer countries.

According to Milanovic, about two-thirds of between-person inequality these days is due to differences between countries, rather than differences within countries.

Credit Suisse (2013) has estimated the global distribution of wealth in the early 2000s. Their data suggest that the bottom 58.7% of the world’s adults who have a net wealth of less than US$10,000 own only 3% of the world’s wealth. By contrast, the 0.7% of the world’s adults who have a net worth of more than US$1 million control 41% of the world’s wealth. Half the world’s wealth is owned by just 2% of the world’s adults. The global wealth Gini of adults is estimated at 0.892 (Davies et al. 2008), the same level as would be recorded by a population of 100 individuals in which one person owns $900 and the other 99 persons own $1.

Figure 3 Change in real income (PPP dollars) for people at different percentiles of the global income distribution, 1988–2008. Source: Milanovic (2012).
Why inequality?

Summarizing these data, we might say that at a global level, income inequality between individuals is high, but has been stable in the past 20 years or so; within most countries, though, income inequality has been rising in the last 20–30 years. There are generally larger inequalities of wealth than of income within countries and globally.

The principal source of this inequality is unequal ownership of assets. In agrarian economies, such as rural areas of India, Indonesia, and Latin America, the principal asset is land, and inequalities in ownership of land drive inequalities in income and wealth. Some people own large farms, others have smallholdings, and yet others own no land and so are forced to work as landless laborers. Data on inequality of land holdings are to be found at developmentdata.org (http://www.developmentdata.org/inequality.htm) and a review of the relations between inequality of land ownership and income inequality is provided by Carter (2000). The interplay between these inequalities and the emergence of such new agricultural technologies as the green revolution provides a fertile ground for understanding the dynamics of inequality (Patel 2013).

The changeover from communist rule in Eastern Europe, the former USSR, and China provides direct evidence of the impacts of asset ownership on inequality. In those countries, between the 1950s and the 1980s, most forms of industrial and agricultural production were organized through communally owned assets – state-owned enterprises and communal farms. There certainly existed inequality in those societies – most notably between urban and rural areas, between managers and workers, and between the politically connected and everyone else; the evidence suggests that this level of inequality was rather low, compared to levels that currently exist in those countries. There were important differences between those societies and the manners in which they removed the communist organizational forms after 1989, but the privatization of these formerly state- or communally owned assets led to new inequalities in asset ownership and an increase in income inequality. The privatization of productive assets did not occur through the equal division of those assets to all households, but rather through a process in which some individuals or groups acquired majority interests in those assets. In China, for instance, this meant that many state-owned enterprises were effectively converted to private enterprises when their managers or bureaucratic superiors simply assumed ownership of them; land has remained in communal ownership, and equally distributed to farm households, but is increasingly being commandeered by local arms of the state for conversion to urban and industrial purposes. In effect, bureaucratic position, state power, and personal connection have been converted into unequal ownership, which in turn has spurred differences in income and wealth between the new class of owner-managers and the workers in those farms. This process is called primitive accumulation.

Political economic geographers begin their analysis of inequality in societies where capitalist production is already generalized as the outcome of a class society in which ownership of assets is unequally distributed, production is controlled by a small group of owners and managers, and the actual work of producing goods and services is performed by ordinary people, often working for relatively low rates of pay. The level of inequality in such a society is then understood to broadly reflect the share of society’s net product that goes to the owners and managers as compared to that which goes to workers, mediated by the degree to which governments tax and redistribute. This division of the social product is a matter of
struggle between classes, and one of the direct indexes of that struggle is the share of net product that is profits, compared to the share that is income. In Europe, North America, Australia, Japan, and India, the share of wages in national net income has fallen consistently since the early 1970s, and in China since the mid-1990s.

As a first approximation, then, we can understand inequality of income as reflecting inequality in the distribution of assets, which generates differences in the tasks performed by those who own and those who do not own those assets, and which is the site of struggle over the division of the net social product between classes. It is also clear that in many rich and moderately rich countries, the redistributive system (taxes on the side of governments’ income; social expenditures on the side of governments’ expenditures) has increasingly failed to remedy these inequalities. Reductions in the rates of tax paid by the highest income earners, relocation of assets offshore, income splitting schemes, and other allowable deductions from income have produced the same effect; goods and services taxes (or value added taxes, or consumption taxes) mean that the poor are on average taxed more heavily than the rich, since poorer people spend a higher proportion of their income on consumption than do richer people.

One of the richest men in the United States, Warren Buffett, made the point:

Actually, there’s been class warfare going on for the last 20 years, and my class has won. We’re the ones that have gotten our tax rates reduced dramatically.

If you look at the 400 highest taxpayers in the United States in 1992, the first year for figures, they averaged about $40 million of [income] per person. In the most recent year, they were $227 million per person – five for one. During that period, their taxes went down from 29 percent to 21 percent of income. So, if there’s class warfare, the rich class has won. (Buffett 2011)

Most billionaires are less honest than Buffett: an Australian government proposal to impose a special tax on the super-profits enjoyed by mining companies during booms led to outrage in the mining industry, prompting the CEO of one of the biggest mining companies, Fortescue Metals’ Andrew Forrest, to declare that the proposal represented class warfare. The term “class warfare” has been used to delegitimate proposals to rein in the incomes and entitlements of the rich, while moves against the poorest are “fiscal responsibility.” Inequality, apparently, is seen as the product of a class war, at least in the richer countries of the world.

Of course, there are then questions about the reasons why some workers earn more than others. Two general approaches to this question have developed. One, orthodox within economics, is that such differences in income reflect differences in the “human capital” brought to the labor market by workers: those with high skills can demand higher wages than those with lower skills. (It also turns out that gender, age, ethnicity, and such other variables also influence wage rates.) In such theories, the incomes of owners and managers are then interpreted as just returns for their efforts in initiating and managing production. As a first approximation, in this approach, incomes reflect relative productivity. A somewhat looser sociological statement of the same principle is that social inequality is thus an unconsciously evolved device by which societies insure that the most important positions are conscientiously filled by the most qualified persons (Davis and Moore 1945, 243).

The second approach argues that such differences in income reflect the segmentation of the labor market into separate sections with different requirements for membership and separate reward structures. Entry into these sections of the labor market is mediated by a variety of institutions, including corporate managers, job
INEQUALITY

allocation agencies, school counseling, unions, and the like, all of which bring their prejudices about the skills and attitudes of different kinds of people (i.e., of men and women, of younger and older people, of this ethnic group and that). For descriptions of segmented labor markets in Europe see Pratschke and Morlicchio (2012).

Why should we care about inequality?

There are two groups of answers to this question. The first centers on fairness. This can be cast in two different forms. The first argument about fairness is that there is a lot of evidence that incomes do not reflect simply human capital, that people’s earnings do not simply reflect their productivity. Differences in earnings by gender and ethnicity, for example, are mostly about prejudice rather than productivity – that is to say, these differences persist even among people of comparable “human capital” (Karlsen 2012). Furthermore, there is no evidence at all that a manager who earns 1000 times the wage of one of his (usually) ordinary workers is responsible for 1000 times the product (though there is evidence that good managers are worth a lot in terms of stock values). Such differences in income, unrelated to productivity, are unfair on anyone’s terms. The second argument about fairness goes deeper: From whence come those differences in human capital that are supposed to be responsible for differences in incomes? A typical girl who grows up in a white, middle-class, liberal household in a rich country might expect high educational standards, encouragement to learn and acquire skills, self-confidence, and a range of social contacts and skills that are not directly productive but useful in speaking the language of managers. A typical boy who grows up in a poor district of a country like Ghana might expect none of these advantages and has no opportunity to develop similar skills (unless he happens to be good at football). The girl has high “human capital,” the boy little; their earnings are correspondingly different, but it is clearly unfair that they have these differences in “human capital,” and so it is unfair that their earnings are different.

The second argument about inequality concerns its effects. Again, there are two general forms to this argument. Those in favor of inequality argue that differences in income and wealth provide an incentive to those who are not rich – an incentive to poorer people to improve themselves by becoming an entrepreneur, learning a new skill, or moving to a place or industry where incomes are high. Inequality, it is said, is a cause of high rates of growth of national output. To evaluate this argument, we would want to know the following:

1. Is inequality actually associated with high rates of growth of output? The evidence does not support the argument that inequality is good for growth (for a review, see Gallo 2002). For example, data about many of the G20 countries indicate that an increase in the share of income going as profits reduces private consumption and has a positive but smaller effect on investment; higher profit shares tend to reduce domestic demand and tend to reduce rates of gross domestic product (GDP) growth.

2. To what degree do we want to value high rates of growth of output in comparison with other social goals? Those who make the incentive argument usually assume that asserting it is sufficient (see Barro 2000 on higher-income countries). However, the economy (output) is not a thing separate from society; it is merely one particular evaluation of society. Thus, even if the incentive argument is correct, it is still necessary to demonstrate that the incentive effect outweighs other social considerations.
Sridaran (2012) makes this argument well: “[the incentive argument] cannot prevail because the reasons for moderating income inequality are dictated by societal wellbeing and political democracy, reasons which are not subordinate to economic policy.”

And this leads to the second argument about the effects of inequality: is social life better in societies with higher rates of inequality than in societies with lesser rates? One fundamental reason for preferring low rates of inequality was provided by Blum and Kalven (1953), who argued that genuine democracy requires that all citizens have similar political power, and that in turn requires that all citizens have similar levels of resources at their command. A remarkable book by Wilkinson and Pickett (2009) demonstrates more concrete relations between inequality and social dysfunction. Compared to countries in which levels of inequality are low, countries with high levels of inequality:

- have higher infant death rates (comparisons between Europe, North America, Australia, New Zealand, Israel, and Japan);
- have higher rates of mental illness (comparisons between Europe, North America, Australia, New Zealand, and Japan);
- have a higher incidence of drug use (comparisons between Europe, North America, Australia, New Zealand, Israel, and Japan);
- have higher high school dropout rates (comparisons between states of the USA);
- have higher rates of prison incarceration (comparisons between Europe, North America, Australia, New Zealand, Israel, Japan, and Singapore);
- have higher rates of obesity (comparisons between Europe, North America, Australia, New Zealand, and Japan);
- have lower rates of social mobility (comparisons between Scandinavia, the United Kingdom, North America);
- exhibit lower levels of interpersonal trust (comparisons between Europe, North America, Australia, New Zealand, Israel, Japan, and Singapore);
- have higher rates of homicide (comparisons between Europe, North America, Australia, New Zealand, Israel, Japan, and Singapore);
- give less foreign aid, as a proportion of national income (comparisons between Europe, North America, Australia, New Zealand, and Japan).

SEE ALSO: Class; Development; Difference; Geographies of death; Geographies of exclusion; Gini coefficient; Human capital; Labor market segmentation; Living wage; Social justice; Water security

References


INEQUALITY


Further reading


Websites

http://www.developmentdata.org/inequality.htm
http://www.equalitytrust.org.uk/
http://inequality.org/income-inequality/
http://wider.unu.edu/
Infiltration

Carol P. Harden
University of Tennessee, USA

In the hydrologic cycle, the process of infiltration – the downward entry of water into the soil – divides surface from subsurface water. Soils with high infiltration rates absorb most or all of the water that reaches them, but soils with low infiltration rates absorb water very slowly, causing water to remain at the surface in puddles or to become surface runoff. By controlling this separation of water pathways, infiltration is a key component of the hydrologic cycle.

The proportion of snowmelt and rain water that infiltrates the soil controls the amount of moisture available for plant growth and groundwater recharge, the delivery of water to streams and rivers, the heights and depths of floodwaters and low flows in rivers, the likelihood of flooding, and the composition of particulate and dissolved materials carried by surface waters. Anthropogenic changes to infiltration rates are widespread, important ways in which human activities alter the hydrologic cycle. People alter natural rates of infiltration by paving, plowing, and trampling the land.

During the process of infiltration, water molecules on the soil surface enter pore spaces in the soil and are pulled into the soil matrix by the force of gravity or by forces of adhesion and capillarity. The infiltration rate depends on characteristics of the soil. Some characteristics, such as soil texture, vary over space, while some, such as soil-moisture content, also vary over time. Infiltration occurs at the soil surface, but the infiltration rate is partly controlled by the capacity of the subsurface soil to store and transmit moisture. Thus, the rate of infiltration depends on characteristics of the soil surface and also on the availability of space within the soil, which, in turn, depends on the size, distribution, and status of soil pore spaces.

Infiltration rates are affected by a suite of physical, chemical, and biological factors. These include soil texture; antecedent soil moisture; site conditions, including irregularities of terrain, the depth to the water table, and soil hydrophobicity caused by forest fire; landscape position; and precipitation characteristics. Soil texture is a major control. In sandy soils and soils with stable aggregates, spaces between soil particles are relatively large and well interconnected, thereby facilitating water storage and movement. Where soil pores connect directly to the atmosphere, and especially in larger pores (macropores), such as those created by burrowing organisms, cracks in the soil, or the decay of plant roots, gravity is the dominant force that moves water downward into the soil. Fine-textured soils, even those with a high proportion of their volume as void spaces may contain tiny pores in which the forces of adhesion and capillarity (called “matric potential”) can exceed the force of gravity. Fine-textured soils, are slow to absorb moisture and very slow to release it. Infiltration rates in clay soils are on the order of 1–5 mm hr⁻¹, compared to rates of >20 mm hr⁻¹ for sandy soils.

The amount of moisture already present in soil determines the amount of space available for additional water molecules. No new water can enter the soil when all of the pore spaces are full. When rain falls on a dry soil, infiltration proceeds rapidly at first and then decreases with time as soil
pores become water-filled (Hillel 1971). Within minutes to hours, under conditions of steady, uniform rainfall, the infiltration rate stabilizes at a quasi-steady state. The United Nations calls this the “basic” infiltration rate (Brouwer et al. 1998). In a homogeneous, structurally stable soil, the basic infiltration rate essentially equals the saturated hydraulic conductivity of the soil (the rate at which water moves through saturated soil). In other words, once the pore spaces become largely occupied by water, a column of soil can only take on new water as fast as its older water can move away.

The slope of the land affects soil infiltration rates by controlling the amount of time water spends at a given place on the land surface. Gentle slopes, irregular surfaces, and depressions hold water, giving it time to infiltrate even where the infiltration capacity is low. In contrast, steep, smooth slopes do not allow water to linger and present fewer, briefer opportunities for infiltration. Vegetative litter or mulch can increase infiltration by slowing the movement of overland flow and by collecting and holding moisture on the surface of a slope.

When a soil nears saturation, conditions that slow the movement of soil water hinder infiltration. Infiltration is slowed by clay particles that swell when they become wet and by ice that blocks subsurface pathways of water movement. The presence of impermeable layers or bedrock at shallow depths in the soil restricts the capacity of soils to store and move water. For example, shallow soils can quickly become saturated during a storm if they are not well drained. Some soils form thin surface crusts after the physical impacts of raindrops break weak soil aggregates apart, redistributing fine particles, such as clays, which clog soil pores. Soil crusts can also form where clay particles, carried into a puddle during a rainstorm, remain as a thin layer after the puddle has disappeared.

The maximum rate at which soil can absorb moisture is frequently called the infiltration capacity (Dunne and Leopold 1978). The word “capacity” may be misleading, as infiltration capacity is a flux (rate of change) rather than a volume, but the term has persisted among hydrologists and soil scientists despite efforts to rename it (e.g., “soil infiltrability” proposed by Hillel 1971). By comparing the maximum infiltration rate to measured and predicted rainfall intensities, researchers can identify threshold rainfall conditions at which overland flow will begin and flooding may become a concern.

Because infiltration is strongly affected by the conditions of the soil surface, human uses of the land surface change infiltration rates and patterns. Infiltration capacities become reduced when soils are trampled by people or animals, or compacted by vehicles and heavy equipment; and infiltration is reduced to zero when the land is paved with an impervious material. Such alterations are typically the unintended consequences of actions designed to achieve other objectives, such as creating a parking lot and constructing buildings. Human activities that inadvertently promote soil erosion or deliberately remove the uppermost layers of soil, as during construction, decrease infiltration rates by removing the most absorbent (O and A) soil horizons and exposing denser soil horizons with lower infiltration capacities.

Like rain and other components of the hydrologic cycle, infiltration is reported in units of depth per time (e.g., mm hr\(^{-1}\)). Infiltration rates are most effectively measured on the flat surfaces of deep, homogeneous soils; thus, measurements made on agricultural soils dominate the scientific literature of infiltration rates. To measure infiltration rates, water is applied to the soil surface and the amount that disappears into the ground is recorded over intervals of time. Water can be applied with a sprinkling system, but most infiltration measurements are made with...
INFILTRATION

Water ponded within a metal cylinder that has been driven about 15 cm into the soil. Lateral movement of water through the soil is avoided by using a double-ring system. Water in the outer ring moistens the surrounding soil and reduces the matric potential so that water in the inner cylinder can move downward into the soil. The constant depth of ponded water in the rings exerts a positive pressure on the soil surface. Assuming that evaporation has been minimized, the amount of water added to maintain a constant water depth equals the amount that has infiltrated. Quantities of water applied typically reach or exceed amounts that would occur in a high intensity rainfall in the area, but such large quantities are intentionally applied to exceed, and thus help identify, the infiltration capacity or maximum infiltration rate of the soil.

Because of the spatial and temporal variability of soil characteristics, measurements of rainfall infiltration have more value as relative rather than absolute rates. Nonetheless, general trends, which are well known from comparable measurements, have supported the development of mathematical models that are used to predict infiltration rates under different conditions. Hydrologic models that predict rainfall runoff for applications ranging from flood hazard zone mapping to irrigation management must include sections that calculate infiltration rates.

Infiltration is difficult to measure in irregular terrain, on steep slopes, or where roots interfere with the process of inserting an infiltration ring and the disturbance created by cutting roots and forcing the rings into the ground alters the infiltration process. Despite the challenges of making measurements in forests, data indicate that infiltration rates in forests exceed those of grasslands by an order of magnitude, largely due to the greater presence of macropores in the forest (Taylor, Mulholland, and Thornburrow 2009). The need to keep infiltrometer rings level to maintain a constant head essentially prohibits the use of double-ring infiltrometers on sloping lands. Sprinkling infiltrometers can apply water to slopes, but water that does not infiltrate flows downhill and pools in the downslope portion of the plot.

When the multiple controls of infiltration are understood, land managers can intervene to increase infiltration rates. In rural landscapes, infiltration rates can be raised by increasing the organic matter content of soils to promote aggregate stability, improving soil drainage, and altering land cover to facilitate the development of root channels and other macropores. In steep-lands, terraces slow the movement of overland flow and give water more time to infiltrate.

As the footprint of urban areas increases, more attention is being paid to increasing infiltration rates in places where they had been reduced by impervious surfaces. Cities and states in the United States are now making efforts to capture the first few centimeters of storm water runoff by creating new opportunities for infiltration and storm water retention. Where surfaces have been paved, impervious pavement can be replaced with semi-pervious surface materials, and areas of impervious surface can be punctuated with structures, such as retention basins and rain gardens, that collect and hold rain water. Although the rate of input of storm water to those structures may greatly exceed their infiltration rates, the water they capture has the chance to infiltrate slowly over time rather than pass by as overland flow. The benefits of infiltration-promoting structures in urban areas include the removal of pollutants as infiltrated water passes through soil systems, the reduction of flood-producing storm runoff, and the local recharge of groundwater resources.

SEE ALSO: Hydrologic cycle; Land degradation; Mountain hydrology; Soil water;
INfiltration

Soils of urban and human-impacted landscapes; Water resources and hydrological management

References


Further reading


The term “informal sector,” coined by Keith Hart, was popularized by the International Labor Office (ILO) World Employment Program (see in particular ILO 1972). Economic activities recognized and fostered by the state were designated as the formal sector, and those ignored by the state constituted the informal sector. Although the activities in different sectors may often be similar, the risks are magnified in the informal sector due to lack of state support. The ILO frames the “problem” of the informal sector as one of inequality, to which it recommends greater incorporation of the informal sector within the formal one.

Incorporated into international development discourse since the 1970s, the original definition of the informal sector was critiqued for creating an untenable binary, which led to single policy prescriptions covering a wide range of activities because of a failure to appreciate how the fractured nature of labor markets creates a continuum of degrees of formality which are not independent but continually interacting. Furthermore, applications of the concept have been deemed urban-centric and criticized for defining a present with no future (since advocates promote formalization) and eliding neighborhoods, households, people, and activities with enterprises (by failing to recognize, for example, that people may work in both formal and informal sectors at different times of day, year, or life course). Critics also charge that the urban poor are often conflated with the informal sector, ignoring waged workers who have low paying jobs and little job security. By 2002, the ILO itself endorsed a move to using the term “informal economy” as a means of encompassing the expanding and increasingly diverse group of workers and enterprises in both rural and urban areas that are not recognized by legal and regulatory frameworks, and therefore face heightened vulnerability. The ILO currently promotes “decent work” along an informal/formal continuum, while continuing to promote a transition to formality.

By most estimates, the number and proportion of people working in informal economies has grown considerably over the past 30 years, due to rural-urban migration, structural adjustment policies, political crises, and globalization, particularly the flexibilization of production and employment relationships. The related concept of informality has also been developed to describe modes of governance through which differential spatial value is produced and managed. Thus urban informality is a “heuristic device that uncovers the ever-shifting urban relationship between the legal and illegal, legitimate and illegitimate, authorized and unauthorized” (Roy 2011, 233). Recent studies have also examined how different modes of informality describe and prescribe housing, citizenship, and the lives of rich and poor in both the Global North and Global South.
INFORMAL SECTOR

SEE ALSO: Development; Inequality; Outsourcing; Precarious work; Vulnerability

References


Information and communications technology

Barney Warf
University of Kansas, USA

Information and communications technologies (ICTs), which include various forms of telecommunications such as the telegraph, telephone, satellites, fiber optics, and the Internet, play a central role in the contemporary global economy and our daily lives, shuttling massive quantities of information between firms, individuals, and places, and contributing to the significant waves of time–space compression that characterize the contemporary world. Telecommunications performs a pivotal role in the organization and transformation of economic activity. As production systems have become stretched over ever larger distances, and have simultaneously become more information-intensive, the need and ability to transmit vast quantities of data have grown accordingly. For more than 150 years, telecommunications has been crucial to accelerating the long-distance flow of information and bringing places closer to one another in relative space.

The origins of telecommunications can be traced to the invention of the telegraph in the early nineteenth century. This innovation initiated the process by which communications became detached from transportation, a process that was advanced in 1876 when the telephone allowed aural information to be included. Telecommunications was essentially synonymous with the simple telephone service, which became critical to the growth of industrial city systems, allowing multiepablishment firms to centralize their headquarter functions in the cores of large cities while they spun off branch plants to smaller towns. Even today, the telephone remains by far the most commonly used form of telecommunications for businesses and households.

In the late twentieth century, as the cost of computing capacity dropped and the microelectronics revolution rapidly increased the capacity of electronic equipment, fiber optics and satellites drastically increased the capacity of telecommunications. With the digitization of information, telecommunications companies formed worldwide integrated information networks.

This entry considers ICTs from several angles. It looks at fiber-optic lines, which now comprise the backbone of the global telecommunications industry, and at satellites, then turns to the Internet, now used by 46% of the people on the planet, which has essentially dissolved the boundaries between the virtual and the physical worlds, so that dichotomies such as online/offline fail to do justice to how closely interwoven cyberspace and real space are. Social and spatial discrepancies in Internet access, loosely termed “the digital divide,” are examined. The entry also explores how ICTs have reshaped the world of global finance, and their effects on global back-office relocations; how digital networks have contributed to an ongoing reconstruction of urban space; how mobile (or cellular) phones have put more people in touch with one another than at any other time in world history; and how social media are having enormous impacts on how individuals view themselves and communicate with one another.
The two primary technologies deployed by the global telecommunications industry are satellites and fiber-optic lines. Firms that provide these services serve overlapping but slightly different markets: satellites overwhelmingly dominate mass media transmission, although fiber carriers have recently begun to invade this market (e.g., cable television). Fiber carriers are heavily favored by large corporations for data transmission and by financial institutions such as banks for electronic funds transfer systems (EFTs), partly because of the higher degrees of security and redundancy they offer.

Since the late 1950s, more than 5500 satellites have been launched worldwide, the vast majority having been sent into orbit by the United States and the Soviet Union/Russia. In addition to military applications, satellites are used extensively by telecommunications companies, multinational corporations, and the global media to link far-flung operations, including international data transmissions, telephone networks, teleconferencing, and television and radio programs. Satellites in orbit appear in a variety of sizes and degrees of technological sophistication. Large satellites capable of handling international transmissions sit 35,700 km (22,300 miles) high in geostationary orbits. From its vantage point, a broad-beam geostationary satellite can transmit to (i.e., leave a “footprint” over) roughly 40% of the Earth’s surface, creating instantaneous time–space convergence. Because the cost of satellite transmission is not related to distance, it is commercially competitive in rural or low-density areas (e.g., remote islands), where high marginal costs often dissuade other types of providers, particularly in fiber optics (Warf 2006; 2007).

Increasingly satellites have been marginalized by the growth of fiber optics. Fiber optics are long, thin, flexible, highly transparent rods of glass (or, less commonly, plastic) about the thickness of a human hair, which can transmit light signals. Because light oscillates much more rapidly than other wavelengths (200 trillion times per second in fiber cables vs 2 billion times per second in a mobile phone), such lines can carry much more information than other types of telecommunications. Modern fiber cables contain up to 1000 fibers each and are ideal for high-capacity, point-to-point transmissions. The transmission capacities of fiber optics grew rapidly in the late twentieth century in parallel with the microelectronics revolution. Financial and producer services firms led the way in the construction of a vast, seamless integrated network of fiber cables which allowed the deployment of EFTs. These constitute the nervous system of the international financial economy, largely because of the higher degrees of security and redundancy that this medium offers. Banks can move capital around at a moment’s notice, manage arbitrage interest rate differentials, take advantage of favorable exchange rates, and avoid political unrest. Although their transmission costs have declined, satellites have failed to match the latest leaps in fiber optics’ capacity and compete with transoceanic submarine cables only with great difficulty; today 94% of all international telecommunications traffic is transmitted via cables (Warf 2006).

The worldwide network of fiber lines links cities, markets, suppliers, and clients, and constitutes the backbone of Internet traffic (Figure 1). The geography of global fiber networks centers primarily on two distinct markets crossing the Atlantic and Pacific oceans, connecting the major engines of the world economy, Europe, North America, and East Asia. In 1988 AT&T initiated the world’s first transoceanic fiber-optic cable, which carried 40,000 telephone calls simultaneously. Because large corporate users are the primary clients of such networks, it is
no accident that the original and densest web of fiber lines connects London and New York, a pattern that resembles that of the telegraph and telephone. In the 1990s a growing web of transpacific lines mirrored the rise of East Asian trade with North America, including the surging economies of the newly industrialized countries.

The complex interplay of deregulation, globalization, and technological change increased the international transmission capacities and traffic volume for fiber-optics carriers exponentially. Between 1988 and 2003, for example, transatlantic fiber-optic cable capacity increased from 43,750 voice paths to 45.1 billion (an increase of 103,000%), while across the Pacific Ocean cable carriers’ capacity rose from 1800 to 1.87 billion voice paths (an astonishing 1.6 billion percent).

In addition to these two major markets, fiber lines have extended into several newer ones. In 1997 several firms opened the Fibeeroptic Link Around the Globe (FLAG), the world’s longest submarine network, which connected Europe, the Middle East, and Southeast Asia via routes across the Indian Ocean.

The Internet

The largest, by far, and perhaps most important ICT network is the Internet. In November 2015 it connected an estimated 3.36 billion people in more than 200 countries, roughly 46% of the planet’s population (Figure 2). From its military origins in the United States in the 1960s, the Internet emerged on a global scale through the integration of existing telephone, fiber-optic, and satellite systems, a process made possible by the technological innovation of packet switching, in which individual messages may be decomposed, the constituent parts transmitted by various channels, and these parts then reassembled instantaneously at the destination. Spurred by declining prices of telecommunications, services, and equipment, the Internet diffused rapidly, allowing any individual with a microcomputer and modem access to cyberspace. In poorer countries, cybercafes often provide a major point of access.

The Internet has had significant economic and social impacts. For well-educated elites, it offers a means of connection to the global economy even
as they become increasingly disconnected from the local environments of their own cities and countries. Online retail shopping, or “e-tailing,” has grown exponentially, particularly for large firms, and given birth to entirely virtual stores such as Amazon.com. The Internet is central to reservation and ticketing systems for airlines and hotels. The higher levels of competition facilitated by the Internet have reduced costs for consumers. Many schools and universities use it for online courses, allowing students who may otherwise not have access to their services to participate in education. Most newspapers offer online versions, greatly facilitating access to the news. The downloading and streaming of music files and films has become a big business, raising concerns over intellectual property rights. Internet-based telephony (e.g., Skype) has grown rapidly. E-business allows firms to advertise, recruit, submit contracts and bills, monitor inventories, and perform other functions quickly and cheaply, reducing transaction costs and raising productivity. Weblogs, or blogs, have allowed vast numbers of people to share their opinions publicly. Graphical user interfaces have enabled the Internet to become a major source of recreation, including online games, while others browse for pornography, raising widespread concerns about its impact on traditional mores, and especially its effects on children. For many users, cyberspace allows for the creation of “communities without propinquity”: that is, groups of users who share common interests but not physical proximity. Political groups of various types and other groups, such as gay people, have used the technology to great advantage. The extent to which digital relations can substitute for real-world face-to-face interactions, however, is open to debate.

The Internet is also used for illegal, unethical, or illegitimate purposes. Spam, or unsolicited email advertising, makes up a large share of the traffic on the Internet. Online gambling has grown rapidly. Electronic systems are used to monitor everyday life, including credit cards, visas and passports, tax records, medical data, police reports, telephone calls, utility records, automobile registration, crime statistics, and sales receipts. These, together with the growth of Internet fraud and identity theft, raise serious concerns about privacy. Thus, telecommunications can be used against people as well as for them.

**Digital divides**

There remain significant discrepancies in access to the Internet worldwide. In all countries, access to the Internet is highly correlated with income, educational level, and employment in professional occupations. Rural areas often have slow or substandard connections. Difficulties in procuring the skills, equipment, and software necessary to access the Internet threaten to exclude economically and socially marginalized groups from the benefits of cyberspace. Thus, the microelectronics revolution has accentuated the digital divide between groups that are proficient in computer technology and those that neither understand nor trust it.
At the international level, inequalities in access to the Internet reflect the long-standing bifurcation between developed and developing countries (Figure 3). The best-connected nations are the Scandinavian countries, Canada, Australia, and Singapore, where more than 90% of the population is connected. In the United States, the Internet penetration rate was 87% in 2015 (with substantial internal variations). Outside of the economically advanced countries, the vast bulk of the world’s people – that is, the developing world – have comparatively little access, particularly in Africa, although mobile phone use has made great advances. Low incomes, inadequate infrastructures, and unhelpful or oppressive governments play major roles in this context.

The digital divide, or social and spatial differentials in Internet access, has been the subject of a growing body of literature (Howard, Busch, and Sheets 2010; Korupp and Szydlik 2005). Fundamentally, this question is about who has access to and can use the Internet and who does not. Admittedly, “access” and “use” are vague terms and embrace a range of meanings, including the ability to log on at home, in school, at a cybercafe, or at work. Rather than a simple access/nonaccess dichotomy, it is more useful to think of a gradation of levels of access, although data at this level of subtlety rarely exist. Thus, it is increasingly common to speak of digital differentiation rather than a divide (Selwyn 2004).

The digital divide is a complex, changing, and multidimensional phenomenon that reflects the diverse channels through which social inequalities are reinscribed in cyberspace. Everywhere, age plays a key role: the elderly are inevitably the least likely to adopt the Internet. In some places, gender is important too: in North America the gendered divide has disappeared, but in Europe it persists, and in the developing world it is pronounced. The digital divide is also a geographical phenomenon. Everywhere, large urban centers tend to exhibit higher rates of connectivity than rural areas (Mills and Whitacre 2003).

As the uses and applications of the Internet have multiplied, the costs sustained by those who are denied access rise accordingly. At the historical moment that professional success came to rely on digital technologies, large pools of the economically disenfranchised have been shut out from cyberspace. As the upgrading of required
skill levels steadily renders information technology skills necessary even for low-wage service jobs, the lack of access to cyberspace becomes increasingly detrimental to social mobility.

In addition to international discrepancies in access, Internet usage is also uneven within countries: invariably, the poor and elderly, ethnic minorities, and those in rural areas enjoy markedly less access (Chakraborty and Bosman 2005). Unfortunately, relatively little is known about this issue, largely because of lack of data. However, the enormous growth of the Internet means that the digital divide is also changing rapidly and, as access improves for many hitherto marginalized groups, it may slowly decline over time.

The latest frontier in the digital divide is the arena of broadband delivery services, which varies widely in availability between the world’s countries. Broadband applications include digital television, business-to-business linkages, Internet gaming, telemedicine, videoconferencing, and Internet telephony. With large, graphics-intensive files at the heart of most Internet use today (e.g., downloading forms, reading online newspapers), broadband has become imperative for web browsing. However, the geography of broadband access replicates the globe’s geographies of wealth and power – it is largely confined to the economically developed world. Thus, rather than eliminating the digital divide, broadband reproduces, and in some cases accentuates, it.

ICTs and global cities

As the Internet unfolded across the world stage, globalization generated a resurgence of “global cities.” Such places are by definition connected via vast tentacles of investment, trade, migration, and telecommunications to clients and markets, suppliers, competitors, consumers, and producers around the world. Not surprisingly, global cities, and indeed most large metropolitan areas, are well-connected hubs with extensive fiber-optic linkages, within which constellations of Internet-related businesses are generated.

Face-to-face contact is essential to the performance of actors engaged in nonroutine functions in global cities, which is critical to the creation and communication of expertise. Despite the enormous ability of ICTs to transmit information instantaneously over vast distances, face-to-face contact remains the most efficient and effective means of obtaining and conveying irregular forms of information, particularly when it is highly sensitive (or even illegal, as the current wave of corporate malfeasance and insider trading demonstrates). In the context of face-to-face meetings, actors can monitor one another’s intentions and behavior through observation of body language, which is essential to establishing relations of trust and mutual understanding. Digital interactions are not a viable substitute for face-to-face communication because they do not convey the subtleties that the latter can.

The continued importance of global cities in the face of a networked global economy is worth emphasizing. Repeated claims that advances in telecommunications would allow everyone to engage in telecommuting, dispersing all functions and spelling the obsolescence of cities, the death of distance, a “flat world,” and a “borderless world” have fallen flat in the face of the persistent growth of dense urbanized places. Indeed, financial and business services firms not only pay high rents to be near city centers, and endure the congestion of such locations, but spend lavishly to fly their executives around the world to meet with their counterparts in person. Even electronic conferencing, such as that offered by Skype, has been unable to replace direct personal
contact. For this reason, a century of technological change, from the telephone to fiber optics, has left most high-wage, white-collar, administrative command and control functions clustered in downtown areas. In contrast, telecommunications are ideally suited for the transmission of routinized, standardized forms of data, facilitating the dispersal of functions involved with their processing (i.e., back-offices) to low-wage regions.

In short, despite the near-ubiquity of the Internet in economically advanced countries, global cities retain their importance and continue to exert hegemony, precisely because they facilitate the sorts of interactions that cannot be conducted easily over the web.

**ICTs and global financial markets**

Few industries have been more profoundly reshaped by ICTs than finance, notably the shift into digital money that began in the late twentieth century. Electronic funds transfer systems constitute the architecture of foreign exchange markets and transaction payment systems. Using the worldwide network of fiber optics, international banks and speculators can shift vast sums around the world at a moment’s notice, creating havoc with national monetary controls. As a result, the velocity of global capital has accelerated to unprecedented speeds: freed from many technological and political barriers to movement, capital has become not merely mobile, but hypermobile.

Capital markets worldwide were profoundly affected by the digital revolution. Electronic funds transfer systems reduced transaction costs, increased the velocity of money, improved capital market efficiency, and generated economies of scale in finance. They eliminated transaction and transmission costs for the movement of capital much in the same way that deregulation and the abolition of capital controls decreased regulatory barriers (Batiz and Woods 2002).

One of the primary forms of EFTs is real-time gross settlement (RTGS) systems (O’Mahony, Peirce, and Tewari 2001), which handle money flows between financial institutions and governments. The largest of these is the US Federal Reserve Bank’s Fedwire system, which allows depository institutions with a Federal Reserve account to transfer funds to the account of any other depository institution. In 2007 Fedwire traffic amounted to $2.1 trillion per day (Federal Reserve 2009). The other major US payments mechanism is the privately owned Clearing House Interbank Payments System (CHIPS) in New York, which clears about $1 trillion in daily transactions. In Europe, the Belgian-based Society for Worldwide Interbank Financial Telecommunications (SWIFT) plays a comparable role. Formed in 1973, SWIFT extends into 208 countries and handles €2.6 trillion in daily transactions. In the United Kingdom settlements are made through the Clearing House Association Payments System (CHAPS), run by the Bank of England since 1984, while in the European Union the Trans-European Automated Real-Time Gross Settlement Express Transfer (TARGET) is used to settle transactions involving the euro. In Japan the Bank of Japan Financial Network System (BOJNET) fulfills a similar function.

Private firms have similar systems. Citicorp has a global information network that allows it to trade $200 billion daily in foreign exchange markets around the world. MasterCard has its Banknet, which links all its users to a centralized database and payment clearing system. Reuters accounts for 40% of the world’s financial trades each day. Other systems include the London Stock Exchange Automated Quotation system (SEAQ), the Swiss Options and Financial Futures Exchange (SOFFEX), and the
The rise of EFTs fundamentally undermined the traditional role of national monetary policy. National borders and monetary policies mean little in the context of massive movements of money around the globe, rendering central bank interventions increasingly ineffective. EFTs present the global system of states with unprecedented difficulties when they attempt to reap the benefits of international finance while simultaneously attempting to avoid its risks. The power of electronic money is evident when currency speculators mount an attack on a given national currency, such as that launched against the Thai baht in 1997, which initiated the disastrous Asian financial crisis.

Offshore banking

ICTs have also accelerated the rise of offshore banking (Warf 2002). Roberts (1994) identified five major world clusters of offshore finance, including the Caribbean (e.g., the Cayman Islands, Bahamas, Panama); Europe (the Isle of Man, Jersey, and microstates on the continent); the Middle East (Cyprus, Lebanon, and particularly Bahrain); Southeast Asia (Hong Kong, Singapore); and the South Pacific (Vanuatu, Nauru). New ones have emerged in Africa and the Indian Ocean. Such places provide commercial investment services, foreign currency trades, asset protection (insurance), investment consulting, international tax planning, and trade finance (e.g., letters of credit). Offshore banking is relatively capital-intensive compared to the labor-intensive headquarters in global cities, and generates relatively few jobs.

Offshore centers are the black holes in the global topography of financial regulation, a status that emanates directly from the enhanced ability of large financial institutions to shift funds electronically to take advantage of lax regulations and
freedom from taxes and currency control restrictions. As the technological barriers to moving money have fallen, legal and regulatory barriers have increased in importance. Even relatively minor differences in regulations concerning corporate taxes or repatriated profits may attract large quantities of capital to particular places, or repel it.

Offshore banking centers have long suffered from suspicions that they constitute little more than havens for tax evasion and money laundering of illicitly obtained funds. As electronic money has come to dominate global finance, the use of offshore banking centers for illegitimate purposes has grown apace, including tax evaders, drug cartels, arms traffickers, terrorists, and corrupt government officials. Given that they often straddle the boundary between legitimate and illegitimate financial activity, a key issue in the success or failure of offshore banking centers is the degree of confidentiality that investors feel they offer. Indeed, the quality of offshore banking centers is often judged by the quality of the laws protecting the privacy of investors.

Back-offices, call centers, and the offshoring of services

Another domain heavily affected by ICTs is low-level, routinized back-office functions. Back-offices perform clerical functions such as data entry of office records, library catalogs, payroll or billing information, bank checks, insurance claims, and magazine subscriptions. These tasks involve unskilled or semiskilled labor, primarily women, and operate on a 24-hour basis.

Historically, back-offices were located adjacent to headquarters in downtown areas to insure close management supervision. However, with ICTs, many service firms uncoupled their headquarters and back-office functions, and moved the latter to cheaper locations on the urban periphery. Recently, back-offices have relocated on a much broader, continental scale. Many clerical tasks have become footloose and susceptible to spatial variations in production costs. Digital call distribution systems have made possible the relocation of phone services that were once confined to large cities.

Internationally, this trend has taken the form of the offshore office. The primary motivation for offshore relocation is low labor costs, although other considerations include worker productivity, skills, turnover, and benefits. Notably, many firms with offshore back-offices are in industries, including insurance, publishing, and airlines, which face strong competitive pressure to enhance productivity. Offshore back-office operations remained insignificant until transoceanic fiber-optic lines made relocations on an international scale possible. Inputs, usually documents or magnetic tapes, are sent by air to offshore processing facilities. After processing, the results are returned via satellite or fiber-optic line.

For example, several New York-based life insurance companies established back-office facilities in Ireland. Often situated near Shannon Airport, they move documents by FedEx and the final product back via satellite or one of the numerous fiber-optic lines that connects New York and London. Likewise, the Caribbean, particularly Jamaica and Barbados, has become a particularly important locus for American back-offices. In Asia, India has an enormous number of foreign back-offices and Manila has emerged as a back-office center for British firms, with wages 20% of those in the United Kingdom.

Offshore data entry functions have been supplemented by other types of job. The animation
of television programs such as *The Simpsons*, for example, has been outsourced to India. American high school students can email Indian tutors for help with their homework (Lohr 2007). Some US video game players, seeking to skip the easy and boring early stages of online role-playing video games, outsource this stage to hired, often impoverished, Chinese players known as “gold farmers,” allowing them to advance rapidly to the later, more challenging, stages (Barboza 2005). Offshoring also includes higher-skilled jobs such as radiological readings of x-rays and PET scans and tax preparation. Increasingly, chip design and software debugging are moving to India too, as have editing and proofreading functions.

A related form of low-wage, low value-added service involves call centers, which include sales, marketing, customer service, telemarketing, and technical support or other specialized business activities (Richardson and Belt 2002). Like back-offices, call centers are primarily screen-based and do not require proximity to clients. The major cost consideration is labor, although the workforce consists primarily of low-skilled women, and high turnover rates are common. They are thus the epitome of a foot-loose industry. There are an estimated 80,000 to 100,000 call centers in the United States, which employ between 3% and 5% of the national labor force (Uchitelle 2002), the majority of which are located in urban or suburban locations.

Like back-offices, call centers have become increasingly globalized. India has attracted a significant number of customer service centers near its software capital of Bangalore, where workers are trained to speak American English (Waldman 2003). Wages there, which average $2000 per year, are higher than average Indian salaries but are only 10% of what equivalent jobs pay in the United States.

### ICTs and urban space

ICTs have had profound effects on the organization of urban space. Although large cities typically have much better developed telecommunications infrastructures than small ones, the technology has rapidly diffused through most national urban hierarchies.

One example is “teleworking” or “telecommuting,” in which workers substitute some or all of their working day at a remote location (almost always home) for time usually spent at the office. Teleworking is most appropriate for jobs involving mobile activities or routine information-handling such as data entry or directory assistance. Proponents of teleworking claim that it enhances productivity and morale, reduces employee turnover and office space, and leads to reductions in traffic congestion, air pollution, energy use, and accidents.

However, the growth of teleworking may, ironically, increase rather than decrease the demand for urban transportation services, as was widely expected. While telecommuters travel to their workplace less frequently than most workers do, many have longer weekly commutes overall, leading to a rise in aggregate distances traveled. In addition, the time released from commuting may be used in traveling for nonwork-related purposes. Moreover, even if telecommuting did reduce some trips, the associated reductions in congestion could simply induce other travelers onto the roads.

Another potential impact of information systems on urban form relates to transportation informatics, including smart metering, electronic road pricing, automated toll payments and turnpikes, automated navigation and travel advisory systems, remote traffic monitoring and displays, and improved traffic management and control systems. Such systems are intended to minimize
congestion and optimize traffic flow, especially during the rush hour peak travel times.

Mobile telephones and Internet telephony

Since the mid-1990s, the mobile (or cellular) telephone has become the most widely used form of telecommunications in the world (Campbell and Park 2008; Comer and Wikle 2008; Ling and Donner 2009). Rapid decreases in the cost of mobile phones have made them affordable for vast numbers of people, including in the developing world. In 2014 90% of the world's inhabitants — 6.5 billion people — used mobile phones, more than twice as many as used the Internet. Because mobile phones are convenient and require much less infrastructure than landlines, it is not surprising that mobile phone usage has leapfrogged over landlines in the developing world, particularly in sub-Saharan Africa and South Asia, which generally lacked investment in the copper cable telephone systems that formed the basis of landline telephony during the twentieth century.

The extent of mobile telephony worldwide is evident in Figure 4. Average penetration rates exceed that of landlines in every country, and in many there are more mobile phones than people. China has the world's largest population of mobile phone users (Ding, Haynes, and Li 2010). Even in sub-Saharan Africa, half of the population uses this technology. Moreover, mobile phone growth is explosive: between 2005 and 2015, the world's number of mobile phone users jumped by 280%. This growth was unevenly distributed and was related to inequalities in income and literacy, gender barriers, and other dimensions of social life. While the rates of increase were modest throughout the economically advanced world, in much of Africa and Central Asia they exceeded 100% in 7 years.

At the same time, the number of the world's landlines fell roughly 2% annually, including reductions in the number of public pay phones and home landlines. Even in the United States, less than half of the population used a landline in 2010. Thus, mobile telephones are largely substitutes for, not complements of, landlines.

Mobile telephony has had important repercussions for the contours and rhythms of daily life, including a growing porosity between private and public realms as once private conversations are held in public spaces (Ling 2004; Ling and Campbell 2009). It has, for example, allowed for sharply improved coordination of mobility, better use of time, and security in case of emergencies, and its enhanced connectivity erodes the disadvantages of physical isolation.

One of the prime applications of mobile phones is text messaging. In 2014 81% of the world's mobile phone users engaged in texting, sending a total of 6.3 trillion messages, or 200,000 per second. Everywhere, teenagers took the lead in adopting texting at far higher rates than their elders (Bolin and Westlund 2009). In poorer countries, texting is a popular substitute for voice traffic because it is considerably lower in cost. Instant messaging services such as Twitter have also grown explosively. Starting in 2006, Twitter grew to include 316 million users worldwide in 2015, who send more than 510 million tweets per day.

Mobile Internet access, still largely in its infancy, may well help to mitigate the digital divide. In many developing countries, wireless telephony allows increasingly large numbers of users to bypass the high access costs of fixed-line technologies. Indeed, it is not unrealistic to claim that the growth of mobile broadband services is
erasing the boundaries between telephony and the Internet altogether.

Another impact of the Internet is Voice over Internet Protocol (VoIP), or telephone traffic conducted entirely through cyberspace, which allows users to bypass public switched networks. VoIP is not a new product per se, but a means of utilizing the existing Internet infrastructure more efficiently. It has grown rapidly at the expense of public switched telephone network (PSTN) traffic; from a minuscule 4% of the world’s telephony traffic in 2000, VoIP has expanded to roughly 28% in 2011.

The world’s most popular VoIP application by far is Skype. In 2009 more than 580 million Skype users (see Figure 8 below) accounted for 12% of the world’s international telephony minutes, or half the global total of VoIP call minutes. Skype is particularly important in international telephony, as it is now the world’s largest international provider of telephony services. Half of Skype calls are between countries, and account for one-fourth of all calls that cross national borders. Video calls have also surged in popularity among Skype users and now comprise 40% of its call volume.

Social media

Social media include ICTs used for personal communications. Facebook is by far the most popular networking site in the world (Kirkpatrick 2010), with more than 1.5 billion users in 2015 (Figure 5), or 21% of the planet’s population. Its users are unevenly distributed across the planet (Figure 6). The United States, with 157 million users (or half of its population), forms the largest single national group of Facebook users. Within the United States, the
highest Facebook penetration rates are found in relatively wealthy, educated states such as Massachusetts, New Jersey, and Washington, and those with a large metropolitan area, such as Illinois and Georgia (Figure 7).

While it is the largest, Facebook is only one of several social networking sites worldwide (Figure 8). In Russia and neighboring states, the Kontakte network is popular. China has promoted its home-grown Qzone system, while in Brazil Zing reigns supreme. Dozens of other smaller sites also exist, such as Maktoob in the Arab world, hi5.com in Mongolia, and Habbo in Finland.

Social media have given rise to new forms of individual identity and behavior in which interpersonal actions are increasingly telemediated (Valentine and Holloway 2002). However, the extent to which digital interactions can replicate or substitute for close interpersonal connections has been widely debated (Nie 2001). Some hold that social media can successfully replicate the emotional depth of face-to-face contacts (Hampton and Wellman 2001), that Internet users have more offline social interactions than nonusers, and that digital interactions are displacing place-based communities as the primary vehicles of sociability. For example, the enormous popularity of massively multiplayer online role-playing games (MMORPGs) such as World of Warcraft, with 8 million players worldwide, reflects like-minded people defined by shared interests rather than spatial proximity (Li, Papagiannidis, and Bourlakis 2010).

However, others maintain that, precisely because digital interactions allow anonymity and disallow nonverbal cues, they discourage the formation of intimate ties. From this perspective, the Internet and mobile phones complement, but do not substitute for, face-to-face interactions. Haythornthwaite (2005) maintains that meaningful interactions via social media are invariably preceded or accompanied by face-to-face ties. As a result, digitally mediated communities are typically ephemeral and lack significant levels of trust or emotional depth. For example, Hampton et al. (2011) argue that the average Facebook user has 229 “friends,” although an overabundance of such friends may actually undermine the capacity to sustain close, strong ties. Telemediated ties have difficulty negotiating social difference: thus few Facebook users cultivate friends with
Figure 7  Facebook penetration rates in the United States, December 2012. Based on Internet World Stats (2015).

Figure 8  Dominant social networking services, December 2012. Based on Vincos (2014).
people who are dissimilar to themselves (Vasalou, Joinson, and Courvoisier 2010).

Digital social media may actually make people less, not more, sociable. Turkle (2011) asserts that Internet and mobile phone use isolate users from one another and deprive them of intimate ties. In her view, telemediated ties lead people to feel alone even when they are together, and she blames texting in particular for an ostensible decline in conversational capacity. Because they confuse being connected with intimacy, heavy users of social media become unable to be alone, and thus often feel lonely. Not surprisingly, therefore, the growth of ersatz intimacy on Facebook has been accompanied by rising levels of loneliness (Marche 2012). What is equally disturbing is that the Internet may generate subtle but pervasive changes in brain structure, including shortened attention spans (Jackson 2008; Carr 2010).

Forging close personal ties necessitates the exchange of tacit knowledge and the cultivation of trust. Thus, the creation of close interpersonal relations is never completely free from the need for face-to-face contacts. Social media are a poor substitute for face-to-face connections, which are fundamental to the creation and transmission of tacit knowledge. The reasons that face-to-face contact is so essential is that it allows for the use of body language and emotional cues, eliminates the possibility of the anonymity that haunts digital encounters, and creates a common feeling of belonging to a community, which are prerequisites for the creation of trust and mutual understanding in a dense network of individuals in close contact with one another.

At the scale of the individual, digital media have changed “what a person can be” (Lanier 2010). The enormous enhancement of human extensibility offered by telemediated ties (Adams 2005) has arguably led to a far-reaching redefinition of the self as social networks increasingly shift from a series of one-to-one ties to webs of one-to-many connections. Digital social media blur the boundaries of the autonomous individual.

SEE ALSO: Global cities; Globalization; Internet and global capitalism

References

Hampton, K., L. Goulet, L. Rainie, and K. Purcell. 2011. Social Networking Sites and our Lives: How People’s Trust, Personal Relationships, and Civic and Political Involvement are Connected to their Use of Social


Information synthesis

Daniel Sui
Ohio State University, USA

Information is an elusive concept as James Gleick (2012) described so eloquently in his best seller “The information – a history, a theory, and a flood.” In the age of big data and knowledge discovery, we seem to have lost sight of the complex meanings of information. In general, information is understood along the data–information–knowledge–wisdom hierarchy, although in recent years the boundaries among the four terms have been increasingly blurred. Nonetheless, data generally refers to the raw facts or descriptions about the world; information is regarded as the processed/structured data that gives the user a special purpose and meaning, and both data and information are needed to produce knowledge – a logical, organized, or structured description about the world. Wisdom generally refers to the application of knowledge at the right place and right time for the right reasons.

Synthesis literally refers to a combination of two or more entities together to form something new. Synthesis (“syn” – putting together) is often understood in the context of analysis (“ana” – breaking up), and the precise meaning of synthesis can be wildly different across the disciplines in science, engineering, social sciences, humanities, and the arts. As a methodology, information synthesis is generally about “organizing the different pieces to create a mosaic, a meaning, and a beauty greater than the sum of each shining piece” (Keene and Zimmermann 2007, 5). Although the precise meaning may be different technically, multiple other terminologies have been used in the geographic literature to refer to synthesis, such as fusion, integration, combination, amalgamation, assimilation, and so on. In essence, synthesis is essentially an abductive sense-making process (Couclelis 2009). Different from induction (something may be true) and deduction (something must be true), abduction only alleges something may be true – typical after vast amount of information is synthesized.

Information synthesis: why (necessity and importance)

In the age of big data, we have continued to witness the exponential explosion of information, increasing rapidly every minute (if not second). To make sense of the latest data and information deluge, we need to develop better ways to carry out both analysis and synthesis, and yet, despite the widely claimed importance of synthesis for both research and education, scholarly coverage on synthesis is far less than that of analysis. Meaningful synthesis is urgently needed more than ever for both intrinsic/cognitive reasons and extrinsic/contextual reasons. Intrinsically, humans are inherently meaning-seeking species, which by default requires us to constantly connect the dots and integrate vast amount of information. It is primarily this advanced cognitive capability that differentiates humans from other species. Meaning-seeking underlies almost all human activities, ranging from basic economic activities for simple survival to all kinds of other social, cultural, political, scientific activities for meeting a hierarchy of human
INFORMATION SYNTHESIS

needs. Extrinsic, humans have had to deal with various threats from either the environment or other hostile forces in the environment since the dawn of civilization. To alleviate these threats, humans have realized the necessity of synthesizing information, thinking outside the box, and making connections for things that seem to be independent and separate at the surface.

Both analysis and synthesis play important roles in geography as an intellectual enterprise, but different research paradigms in geography may have emphasized analysis and synthesis in different ways. For example, among geography’s four traditions as defined by Pattison (1964), synthesis plays the leading role in the area studies (also known as regional geography) tradition whereas the spatial analysis tradition favors more analytical approaches. The man–land relationship (or human–environment interaction) tradition uses a more hybrid approach of both analysis and synthesis, depending on the research questions involved. Analytical approaches used to dominate the Earth science tradition (otherwise known as physical geography) but, increasingly, synthesis has also become more important in recent years with the incorporation of the concepts of place and storytelling into physical geography and Earth sciences (Phillips 2012). In recent years, the discipline of geography as a whole has made enormous methodological efforts to cross the quantitative–qualitative chasm with the goal of developing a new level of synergy of both analysis and synthesis in geographic inquiry (Sui and DeLyser 2012).

Geography as a discipline has oscillated between analytical and synthetic paradigms. During the first decade of the twenty-first century, along with geographers’ current and recent turns to specific domains and approaches (e.g., the critical turn, the cultural turn, the relational turn, the computational turn, the communicational turn, the mobilities turn, the performative turn, etc.), one turn deserves particular attention here: the turn to synthesis and to holistic approaches. At least three trends over the past ten years provide broader context for this synthetic and holistic turn.

First, calls for a unified geography as the new disciplinary identity and synthesis will be the driving methodology for studying our planet not bit by bit but all at once. As a conceptual framework, hybrid geographies are proposed as the first step towards a unified geography. Rose (2000, 364) observes that hybrids “transgress and displace boundaries between binary divisions and in so doing produce something ontologically new.” Through a higher level of synthesis, hybrid geographies aim to end two major divisions within geography: the partition between physical and human geography, of nature from society; and the separation of spatial-analytical geographies, which attempt to create a mode of disembodied geographical analysis, from social, cultural, and political geographies. Also sometimes known as the boundary projects, hybrid geographies seek to integrate elements thought to be incompatible or conflicting in grounded practices. Hybrid geographies challenge existing boundaries and forge creative connections within geographies – physical and human, critical and analytical, qualitative and quantitative – aiming to integrate perspectives on space, place, flow, and connection.

Second, the first decade of twenty-first century has witnessed a spatial turn across the physical sciences, social sciences, and humanities, with scholars from different disciplines using their domain expertise to address challenging issues from a spatial perspective. Space has become an integrating theme across the social sciences – evidenced by emerging spatially integrated social sciences (http://csiss.org) – and scholars across the humanities have made GIS
and spatial analysis integral parts of their research methodologies (Warf and Arias 2008; Scholten, van de Velde, and van Manen 2009).

The science of complex systems has advanced dramatically in recent years, as research has shifted from an older model of the single investigator to today’s multidisciplinary collaborations. Future scientific advances are likely to involve mining of multidimensional datasets and to require the kinds of data synthesis that can only be achieved if systems are to a large degree interoperable. Early calls for a new emphasis on synthesis in geography (Gober 2000) are also reflected in recent efforts in adjacent disciplines to accelerate synthesis in and between ecology and the environmental sciences (Carpenter et al. 2009).

Neither has this spatial turn been confined to the Ivory Tower. Policymakers have realized the crucial importance of space and place in understanding the complexity of the world’s problems, and have sought solutions to these problems that will work well under diverse local circumstances. The World Bank framed its 2008 world development report entirely from a geographical perspective, concluding that alleviating, and eventually eliminating, poverty problems must start with reshaping the world’s economic geography; and the White House under President Obama urged all US federal agencies to develop place-based policies to address the pressing problems facing American society. The revenge of geography in global economic and political lives (Kaplan 2012) will further elevate the stature of synthesis as a dominant mode thinking about the global issues facing the world today.

Information synthesis: what to synthesize and the general process

The dearth of information on synthesis is paralleled by a paucity of literature on how to conduct synthesis. Although the technical descriptions on synthesis may be abundant in various specialized fields, general descriptions on synthesis as a way of thinking or as a method of research are lacking. To be successful in synthesis, certain philosophical predispositions are almost a prerequisite in order to encourage synthesis. Gardner (2009) argues that the adoption of multiperspectivalism can enhance one’s proclivity to synthesize. The so-called multiperspectivalism recognizes that each different analytical perspective can only contribute partially to the phenomenon being studied. To gain a more complete, or holistic, elucidation of an issue at hand, one by necessity must synthesize those partial results from analysis in order to avoid synecdoche – as it is revealed in the six blind men and the elephant allegory. Multiperspectivalism enables us to appreciate the complementary strength of different perspectives, each contributing to the understanding of the organic whole.

Information synthesis is an integral component of geographic scholarship of all kinds. Synthesis can take place at different levels of granularity or scale covering varying breadth and depth of a topic, depending on the goals, objectives and audience. For example, synthesis can be done for a particular topic in a subfield, between two subfields, or even between/among several different disciplines – more popularly known as interdisciplinarity or transdisciplinary synthesis. All synthesis work needs to balance breadth and depth, and there is typically a tradeoff between the two. Synthesis with both breadth and depth is possible, but extremely difficult to achieve.

Generally speaking, synthesis can be done at the data, methodological, and conceptual levels. Data can exist in different formats and media, including text, image, audio, video, and maps. This entry focuses on information/data synthesis, but issues related to methodological
and theoretical synthesis are also be discussed for hybridizing physical with human geography.

In general, any efforts to synthesize entail the following four loosely ordered steps.

1 Setting the goal: a simple statement or concept of what the synthesizer is trying to achieve; define the topic and relevant information about the topic, the purpose of the synthesis, and the target audience.

2 Gathering information: with a starting point – an idea, image, or any previous work on which to build, systematically gather the relevant information, and assess the validity of the information.

3 Selecting a strategy and method: during this stage, the synthesizer’s disciplinary training comes into play. The synthesizer must choose a qualitative, quantitative, or a combination of qualitative and quantitative approaches to achieve the goals of synthesis; some of the tools may be discipline specific, with predictable fit for the goal.

4 Presenting results and obtaining feedback: take an initial crack at a synthesis by presenting validated information in a way useful to the target audience; the first cut, provisional synthesis. The first drafts are often crude and primitive but often contain the crucial elements of final product, abduction.

One area that best depicts the broader synthesis efforts is geographers’ recent attempt to synthesize across different domains by hybridizing physical and human geography. Although geographers following the cultural/political ecology tradition have been working on topics that link physical and human geography for decades, the growth of hybrid geographies has drawn an increasing number of human and physical geographers (who normally would practice either human or physical geography exclusively) to cross the physical–human divide. Neil Smith (1998) speculates on the emergence of what he called the El Niño capitalism – both literally and metaphorically – noticing the striking similarities in the rhythms of El Niño Southern Oscillation (ENSO) and capitalism’s periodic crises, and striving to link them in a causal chain. Although intriguing, El Niño capitalism remains controversial, and others have disputed the extent ENSO has affected US macroeconomic performance. Using data for ENSO fluctuations and the rates of US inflation and economic growth over the 1894–1999 timespan, Berry and Okulicz-Kozaryn (2008) explored whether there has been any co-cyclicality between the two phenomena and whether aperiodic ENSO shocks have had any impact on these macroeconomic parameters. They discerned no co-cyclicality or aperiodic shocks, concluding that while ENSO may briefly influence the performance of particular sectors of the economy in particular regions, such locally important effects “vanish into the noise surrounding macroeconomic trends in an economy as large and complex as that of the US” (Berry and Okulicz-Kozaryn 2008, 625).

It is not only human geographers addressing topics typically thought of as physical in new ways; physical geographers have also embraced topics more traditionally human – hybridizing physical and human geography is a two-way street. Inspired by Allan Pred’s work (1984) on the formation of place, Phillips (2001) highlights the primacy of place in human impacts on the environment, and contingency of place has been a dominant theme in his subsequent publications (Phillips 2009). In parallel to human geographers’ efforts to examine economic impacts of global climate change, physical geographers have studied the environmental consequences of the rising divorce rate. Yu and Liu (2007) tracked carbon footprints of couples in twelve countries who married and then divorced, stayed married,
or divorced and remarried. Because divorced couples typically require two residences, Yu and Liu found that energy consumption, water usage, and the amount of space occupied per person all increased dramatically with divorce. While perhaps unsurprising in its results, their study points to the real significance of social and cultural issues affecting individuals on large-scale processes (typically thought of as purely physical) such as global climate change.

More broadly, Clifford (2009) suggested that the study of globalization need not be the exclusive domain of human geographers or social scientists. Instead, he argued that physical geography has always been global at heart, and that globalization must be seen historically in the global export of western science – including physical geography – that underpinned colonial resource exploitation. And Mistry, Berardi, and Simpson (2009), drawing from their field experiences for a water-quality project in Guyana, traced their journey as three physical geographers from identifying themselves as top-down experts to becoming participatory facilitators – addressing issues of reflexivity and positionality often (mistakenly) thought only of concern to human geography.

Continental philosophy, long of interest to human geographers, has become a source of methodological reflection for physical geographers. Using Nietzsche, Comrie (2010) aimed to engage physical geographers and fellow physical scientists to reconsider their roles as scientists. Debunking the mystique of science and the misconception of seeing science as independent of people and society, Comrie showed that science gains its power by the way meaning is attached to it and its findings, so we should act on our ability to bestow that power. Through Nietzsche, Comrie challenged physical geographers to overcome their trained tendency toward detached environmental science and, instead, fashion a new physical geography that includes meaning and action. Recent calls for physical geographers to have more intellectual interactions with critical human geography (Lave et al. 2014) further reflect this synthesis trend in the discipline to hybridize human and physical geography.

**Information synthesis: methods and tools for synthesis**

Multiperspectivalism is necessary but not sufficient for meaningful synthesis to take place. A variety of methods and tools are needed in order to achieve the goal of synthesis. According to Gardner’s (2006) theory of multiple intelligences, humans possess seven high-level intelligence capabilities other species lack in one way or the other: musical–rhythmic, visual–spatial, verbal–linguistic, logical–mathematical, bodily–kinesthetic, interpersonal–intrapersonal, and naturalistic. As far as the specific synthesis methods or tools are concerned, people with different aptitudes tend to rely on different strategies for synthesis. For example, those who are strong in verbal–linguistic capabilities may rely on stories and narratives for synthesis whereas those logical–mathematical types always crave new concepts and theories to achieve the goal of synthesis. Of course, synthesis could occur across multiple intelligence domains. An Oscar-winning movie is usually a successful synthesis of verbal, visual, logical, and musical components.

Generally speaking, approaches for synthesis can be grouped into the three broad categories identified here, although a higher level of synthesis could happen between or among these three types of syntheses:
INFORMATION SYNTHESIS

Synthesis through qualitative/narrative approaches

This is perhaps the most frequently used approach for information synthesis. Vast amounts of information can be synthesized through some simple taxonomy and classification, such as Linnaeus taxonomy for plants or Köppen’s classification for climates. These taxonomies and classifications are usually based upon scientific concepts or models or rules that are commonly accepted by the scientific community. Developing models or theories is also sometimes the most effective way to achieve the goal of synthesizing vast amount of information.

Although reductionism has dominated scientific practice since the scientific revolution in the sixteenth century, scientists in various fields have never abandoned their search for a more holistic understanding of how the universe works. From the unified theory of fields in physics to the Gaia hypothesis in ecology, from evolutionary economics to Gestalt psychology and holistic health in medicine, scholars have never ceased their quest for a more comprehensive understanding of the world and ourselves. Although much of the success of our scientific enterprises to date has come from a reductionist analysis of a system’s parts, understanding and seeing the whole can be fundamentally more gratifying. Higher level synthesis often leads to major breakthroughs in science. Ample examples can be found of successful synthesis stories throughout the history of science. A good example is James Maxwell’s synthesis work that unified electricity and magnetism in the nineteenth century. A more recent effort is David Deutsch’s (1998) attempt to synthesize quantum theory, critical epistemology, computation theory, and evolution theory to better understand the fabric of reality.

Perhaps there is no better word to capture the spirit of synthesis than consilience, a term originally coined by British philosopher William Whewell in the mid-nineteenth century, but popularized by E.O. Wilson (1998) in his best-seller Consilience: The Unity of Human Knowledge. Consilience literally means “jumping together of human knowledge” and is used widely to refer to the unity of human knowledge in the conceptual realm.

One such consilience effort was made by Anne Buttimer (2002), who argued, with compelling evidence, that most models or theories in geography are conceptually motivated by four metaphors or analogies: the world as forms, the world as machines, the world as organisms, and the world as spontaneous events. Inspired by Buttimer’s work, Sui (2011) further argued that most urban models developed so far are drawing from these four metaphors as well – cities as forms, cities as machines, cities as organisms, and cities as spontaneous events.

Narratives and metaphors are certainly not confined to synthesis in human geography only. In fact, physical geography (and Earth sciences more broadly) is full of stories as well. Phillips (2012) found that recent Earth science literature followed eight predominant story lines (cause/effect, genesis, emergence, metamorphoses, destruction, convergence, divergence, and oscillation) in explaining the Earth surface evolutionary processes. Physical geographers may strictly follow a scientific approach but the scientific questions they pose are more or less dictated, consciously or subconsciously, by those metanarratives (or archetype story lines).

Synthesis through quantitative and mapping approaches: from overlay to mash-up

Besides the qualitative/narrative approaches, quantitative/modeling approaches are also important means for synthesizing information, ranging from regression models to multicriteria evaluation to Bayesian modeling. The growing
literature on space–time integration represents geographers’ systematic effort to synthesize information across both space and time using a variety of quantitative methods and techniques (Dijst 2013).

Also, with growing importance, synthesizing vast amounts of information is taking place through mapping vast amounts of digital geocoded information. The first decade of the twenty-first century saw the explosive development of Web 2.0 technologies along with major advances in geoweb and geospatial technologies, including traditional geographic information systems (GIS), remote sensing (RS), global positioning systems (GPS), and location-based services (LBS) to reach an unprecedented moment in human history: we can now know where nearly everything, from genetic to global levels, is at all times. These technological advances can be brought to bear on the corresponding data avalanche – the vast amounts of user-generated content and volunteered geographic information (VGI) that pours out as individuals become sensors, gathering and disseminating data about their environments and themselves with increasing spatiotemporal granularities. These vast amounts of big spatial data are also increasingly linked, otherwise known as the vinculation, which has greatly facilitated a more sophisticated level of synthesis work.

Many forms of synthesis in the context of online digital information can be described as mash-ups. Borrowed from the music industry, the term originally refers to a song or composition created by blending two or more songs. Yet in the context of web-based applications, a mash-up may have multiple meanings. At the functional level, a mash-up may be a web page or application that combines data or functionality from two or more external sources to create a new service. In terms of actual content, a mash-up can be a digital media file containing a combination of text, map, audio, video, and animation, which recombines and modifies existing digital works to create a derivative work. The term implies easy, fast integration, frequently using open application program interfaces (APIs) and data sources to produce something new. A growing number of industry leaders, such as Google, Microsoft, Yahoo, Mapbox, and Cartodb, have developed APIs that users are adapting to develop their own creative applications. But mash-up is much more than a technical advance, and we believe that the true significance of mash-up lies in its potential promotion of a new habit of mind towards synthesis.

When mash-up is used to integrate multiple sources of data based on shared references to the same geographic locations, the operation is conceptually related to the traditional GIS function of overlay, a type of spatial join, although the technical processes are quite different. Pioneers such as Manning, Dusseldorf, Lewis, and McHarg used manual overlay methods to integrate multiple layers of information, in order to develop a comprehensive understanding of spatial patterns and relationships (Elwood, Goodchild, and Sui 2012). Better integration and synthesis of diverse sources of georeferenced information were a top priority during the early days of GIS development in the 1960s and 1970s. However, spatial analysis has been at the forefront in GIS during the past 20 years, with a primary focus on improving spatial-analytic functions. The spatial analysis tradition often took a reductionist approach, focusing on individual layers to identify spatial patterns, rather than on the synthesis of multiple layers. We contend that the growth of mash-up practices in the age of Web 2.0 is revitalizing an interest in synthesis in GIS. Mash-up, as both a concept and a practice, resonates well with the traditional spirit of geography in its quest...
INFORMATION SYNTHESIS

to understand the multidimensional nature of
the Earth’s surface. Although often couched in
different terms (e.g., data conflation, data fusion,
or data integration), geographers, in general,
and GIScientists, in particular, have developed
a considerable number of useful techniques for
synthesizing data from multiple sources. Soft-
ware tools, such as Microsoft’s Photosynth, IBM’s
ManyEyes, HistoryFlow, and TouchGraph, can
further facilitate the synthesis of data in diverse
media.

Admittedly, synthesis is much more chal-
lenging than analysis from a methodological
perspective, due to its involvement with data
in multiple formats and media (number, text,
oral story, photo, video, simulation, etc.). Geog-
raphers’ current efforts to develop a more
eclectic approach by linking diverse quantitative
and qualitative methods may be tapped for
VGI applications. Unlike the spatial analysis
of a previous era that was often conducted by
a lone analyst, synthesis of VGI tends to be
much more participatory, through a mixing
and remixing of multiple data and methods,
as demonstrated in VGI applications in several
recent disaster-relief efforts. Instead of seeking
truth, the new mash-up efforts focus more on
developing narratives about various locales.

Synthesis by crossing
the qualitative–quantitative chasm

The interdisciplinary nature of geography itself
can foster a methodological synthesis in the
mixing of qualitative and quantitative meth-
ods, and recent work features sophisticated
mixed-methods approaches that cross the divide
between spatial-analytical and social-critical
approaches in human-geography research (Sui
and DeLyser 2012), revealing the binary between
approaches as pseudo rather than real. Critical
geography need not be qualitative and can use
numbers – after all, Karl Marx used quantitative
methods extensively in his work. As Elwood
(2010) points out, thoughtful mixed-methods
research must bridge not only methodologi-
cal but also epistemological and philosophical
divides. “Cherished theoretical principles”
may “become renegotiated” – and Bergmann,
Sheppard, and Plummer (2009, 265) seek noth-
ing short of a “methodological reinterpretation
of what employing mathematical arguments
could mean within larger, postpositivist theoret-
ic projects in critical human geography.”

Mixed-methods research offers human geog-
raphers the opportunity to identify the appropriate
role for different methods and, in transport
geography, to explore how context affects
human travel behavior. Zolnik (2009) devel-
oped a multilevel, mixed-methods approach to
address some of the criticisms of quantitative
methods in transport geography, revealing that
quantitative modeling (at a single level) can
complement qualitative analysis (across multiple
levels). Goetz, Vowles, and Tierney (2009)
examined recent transport-oriented research in
highly cited geography journals, revealing that
geographical research on transport topics is much
more prevalent and (though often influenced by
civil engineering approaches) reflects a wider
range of epistemological and methodological
approaches than frequently assumed, propos-
ing a critical-transport-geography research
agenda that calls for a more seamless integration
of qualitative analysis into the predominant
quantitative-modeling approach.

Perhaps most dramatic in mixed-methods
research in human geography is the development
of qualitative GIS (Aitken and Kwan 2010),
which, along with participatory GIS, feminist
GIS, and critical GIS, works to reconceptualize
GIS as more than only quantitative in terms
of data, analysis, and representation. Emerging
in response to critiques that characterized GIS
as rooted in positivist epistemologies and most suited for quantitative techniques associated with the discredited spatial science, qualitative GIS reveals that GIS, from its inception, “has been more than quantitative” (Cope and Elwood 2009, 171). Using mixed, or hybrid, methods in representation, mode of analysis, and conceptual engagement, qualitative GIS embraces noncartographic forms of data, qualitative analysis, and multiple modes of representation (Cope and Elwood 2009).

Knigge and Cope (2009) show how the inductive, iterative analysis practices of grounded visualization can engage scale in GIS as both a cartographic representation and a sociopolitical construction. Elwood (2009), drawing from her work related to grassroots GIS practices, demonstrates how cartographic representations generated in a GIS might be engaged to produce multiple and different understandings of neighborhood, negotiating the meanings and characteristics associated with neighborhood as flexible and fixed, and engaging them as both material and imagined space. Aitken and Craine (2009) show how a nonrepresentational reading of GIS-based representations can illuminate greater insights into affective and emotive politics than more traditionally technical readings. Indeed, the boundaries between qualitative and quantitative are not so clear-cut and are often crossed and synthesized these days. Most recent work in the broadly defined spatially integrated social sciences and humanities embodies the spirit of synthesis at data and methodological level. Kraak’s (2014) book is a remarkable success for location-based storytelling, in which he artfully synthesized historical and geographic analysis with cartographic visualizations and fascinating historical narratives about Napoleon’s Russian campaign of 1812.

Information synthesis: assessment and educational challenges

Cognitive psychologists have suggested that synthesis, rather than analysis, is the defining characteristic of human creativity (Wallace and Gruber 1989; Gardner 2009), and there has been a turn towards synthesis in multiple disciplines in the age of big data. Geography is uniquely positioned to lead this new wave of synthesis with respect to geographic information due to its traditional emphasis on synthesis.

The goal of information synthesis is to bring together disparate information in order to develop a more complete and comprehensive understanding of the situation or problem at hand. Synthesis can be transdisciplinary, but also is domain specific. Contingent upon the topic and mission, what is considered as a good synthesis in one area may be inappropriate in another area. Synthesis normally involves creative leaps, and creativity typically entails synthesis – one type or another (sometimes colloquially known as “connecting the dots”). For analytical results, multiple qualitative and quantitative measures have been developed to assess their effectiveness and validity. Unlike analysis, assessment on the quality of synthesis can be tricky. Although synthesis is normally desired, there is no a standard set of criteria and protocol to assess the quality of synthesis. It is still an ambivalent business to decide when a productive synthesis has been accomplished, as opposed to when the proposed synthesis is premature, misguided, or even fundamentally “wrong-headed” (Gardner 2009).

Generally speaking, quantitative synthesis can be evaluated and replicated better than qualitative synthesis (assuming all the data used in the synthesis work are made available). The assessment of qualitative synthesis is a bit more challenging.
INFORMATION SYNTHESIS

Sometimes, superficial synthesis by simple juxtaposition and lumping can be easily spotted, but problems for qualitative synthesis can run much deeper because, during the synthesis, concepts can be misleading, metaphors can be deceptive, and theories can be misguided. On top of all of these, a variety of technologies and tools are used these days in synthesis related to big data. The enframing nature of technologies has further exacerbated the hidden biases embedded in the process. For example, quite a bit of synthesis work using social media data (such as Twitter) has been reported in recent years to infer social, political, health, and consumption trends, and yet social media data such as Twitter are normally very biased to begin with (e.g., only a very unique subset of the population tweet on a regular basis), and currently we do not have a good set of tools to handle such hidden biases, which will ultimately affect the quality of the synthesis work.

The difficulty in assessing the quality of synthesis work poses serious educational challenges in terms of how and what our students should be taught to do a better job in synthesis, with the ultimate goal of achieving creativity in whatever they do. The coverage on synthesis in the literature is rare because synthesis as a method is often incorporated subconsciously into the process of an intellectual endeavor. What was tortuously achieved by hard work can only be shared by intuition, so obvious and omnipresent that its provenance is hard to track. Indeed, scholars rarely write about synthesis because most of them write with it.

As Isaiah Berlin (1993) described so eloquently in his classic The Hedgehog and The Fox, all writers and thinkers throughout human history can be divided into two categories: hedgehogs, who view the world through the lens of a single defining idea (examples given include Aristotle, Shakespeare, Montaigne, Goethe, Pushkin, Balzac, and Joyce). Cognitively, Berlin’s hedgehogs and foxes can be related to Gardner’s (2009) laser intelligence versus searchlight intelligence. Laser intelligence probes deeply into a topic but ignores opportunities to cross-pollinate, which is often best suited for work requiring discipline. On the other hand, searchlight intelligence may not probe as deeply but is always scanning the environment and may, therefore, more readily discern connections (and identify differences) across spheres. Both hedgehogs with laser intelligence and foxes with searchlight intelligence may have the proclivity to synthesize, but the contents that they synthesize and the criteria for success may differ widely. It remains a challenge for educators to develop effective strategies and methods of teaching synthesis for both hedgehogs and foxes. Most pedagogic materials developed in geography and GIS are still focusing on analytical skills despite the importance of synthesis skills.

Summary and conclusions

Like spatial analysis, information synthesis is an integral part of geographical scholarship that transcends subdiscipline specializations. Indeed, analysis and synthesis are two inseparable aspects of a holistic geographic methodology. Analysis and synthesis always go hand in hand; in most cases, they complement one another – there are important situations in which one method can be regarded as more suitable than the other. Meaningful synthesis is built upon the results of preceding analysis, and meticulous analysis...
makes more sense only when a subsequent synthesis is achieved. Successful geographic research typically rests upon an artful combination of analysis and synthesis.

In this entry, a broad survey on why, what, and how to conduct synthesis in geographic research has been presented. Hybridizing, remixing, and mashing up conceptual frameworks, data sources, and modes of analysis in the spirit of synthesis, as some of the most recent works have revealed, may provide the best hope to cross the methodological and epistemological chasms that have divided geographic and GIScience research. Although the adoption of multiperspectivalism can enhance one’s proclivity to synthesize, the difficulties of synthesis should not be underestimated. Meaningful synthesis is harder to achieve than analysis and the quality of synthesis is also often difficult to assess. As Wyly put it, “how can we ever find the time to master the dizzying array of traditions and techniques required to create truly hybrid geographies, without giving up the depth that comes with specialization in social theory or spatial econometrics or feminist ethnography or participant observation or policy analysis or the list goes on?” (Wyly 2009, 319). Even institutional culture often discourages transdisciplinary research. And yet, as Turner (1989) argued, neither has a single-minded analysis nor synthesis approach served the discipline well. Instead, a balanced analysis–synthesis approach adapted in each new situation may prove to be the most effective methodological framework in the age of big data.

Acknowledgment

Part of the material in section “Information synthesis: methods and tools for synthesis” was drawn from Elwood, Goodchild, and Sui (2012) and Sui and DeLyser (2012), which contain more details and references on recent work related to synthesis.

SEE ALSO: Big data; Qualitative information: representation; Quantitative methodologies; Spatial analysis; Spatiotemporal analysis

References

INFORMATION SYNTHESIS


Further reading

Information technology and mobility

Tim Schwanen
University of Oxford, UK

Transportation and information technologies (ITs) are two sides of the same coin: both have been developed and are deployed to overcome the friction of distance. When drawing and writing were invented, transportation and communication began to be disjointed but they only become separate domains with the diffusion of telegraph, telephone, radio, and television. The advent of the information age has accelerated this process of separation. At the same time, digitalization and computerization have also multiplied the intensity and complexity of linkages between ITs and transportation.

Geographers have examined the spatiality and geographical implications of those linkages since at least the 1960s. Analytically inclined transport and urban geographers have undertaken the bulk of research, but geographers working on “mobilities” have also made important contributions. Spatial-analytical research has, from the start, been preoccupied with the promise of “substitution” of physical movements of people and goods by interaction-at-a-distance via landline telephones, wired computers, and wireless devices. This is unsurprising, given that substitution has long been considered the most desired outcome by transportation planners seeking to reduce congestion, productivity and monetary losses, adverse environmental effects, and transport poverty. Since the mid-1980s, however, geographers have argued that a wider range of relations among transportation and ITs should be considered. Salomon (1986) argued that direct impacts “resulting from a change in the relative use of the two technologies” (page 223) should be distinguished from long-term indirect impacts of IT use on transportation via changes in the spatial organization of activities and land use. Among direct impacts substitution is certainly not the only possibility. Generation may also occur: more IT use for communication purposes can induce more transport activity and vice versa. Moreover, ITs can be used to improve the operational efficiency of transportation systems, and ITs and computer code are now indispensable to the functioning and optimization of any system for motorized transportation. This is perhaps most evident for freight transport and logistics, where information sharing is widely used in supply chain management and IT-enabled decentralized intelligence – the localization of decision making on routing – is believed to create further efficiency enhancements in the near future. Whether efficiency increases reduce transport volumes is a moot point. Evidence is scant to date, but rebound effects in the form of more travel across longer distances may be quite common.

Salomon’s taxonomy of interactions has proven influential; numerous studies in passenger transport have sought to establish which type of impact prevails, when, where, and in what circumstances. The effects of teleworking on commuting and non-work travel have been examined most widely, although the past decade has witnessed greater attention to the transport implications of e-commerce, online leisure, and videoconferencing. Results vary across studies and are difficult to generalize; whether IT
use substitutes, modifies, or generates transport activity varies spatially and depends on the specifics of place. It nonetheless appears that (modest) substitution of passenger transport is the most common net effect. The traditional focus on substitution has diminished somewhat over the past decade; however, the need to reduce greenhouse gas emissions from transportation at a time of unabated growth of global transport volumes – particularly long-distance travel – keeps fueling the hope for significant substitution of physical transport by IT use. Resilient optimism regarding IT’s substitution potential can also be observed among freight transport researchers and stakeholders.

Analytically inclined geographers have also moved beyond questions of substitution of passenger transportation. Drawing on time-geography, they have examined how individuals’ IT use displaces time for offline activities (and vice versa) and relaxes space-time constraints on activity/travel practices. Discarding the “death of distance” hypothesis and the idea that ITs allow everything to be done everywhere, they have shown that space, time and activity are only partially decoupled and that the constraint-relaxing effects of ITs are dependent on place-specific social and cultural contexts. Mei-Po Kwan and colleagues, for instance, have shown that differences between men and women in activity/travel practices and space-time constraints are hardly reduced with the advent of the internet and mobile phone because the effects of IT use are mediated by place-specific gender regimes (Schwanen and Kwan 2008).

The geographies of everyday activity/travel practices are nonetheless changing in many ways. First, activities are being “fragmented” (Couclelis 2004): ITs afford people the opportunity to disaggregate activities that previously were conducted in specific places and times (e.g., shopping) such that they are disassembled into smaller bundles that can be carried out in other places (at home, on public transport, etc.), at other moments (outside store hours), and in alternative sequences (e.g., online information search, experience the product in a store, buy it online, receive at home or at a collection point). Consequently, the “contact set” of locations that are contacted physically or via electronic mediation will expand; more agents and stakeholders may participate in the activity; and the timing, destination, and length of physical trips may change. Second, ITs have greatly facilitated multitasking, allowing people to do multiple things simultaneously (e.g., work aboard a train) or switch continuously between activities (e.g., gaming on a mobile device and watching TV). According to many commentators, IT-enabled multitasking may increase people’s willingness to travel in general and as a passenger in particular. This can increase the attractiveness of sustainable transportation modes (e.g., public transportation) but may also increase travel distances, congestion, and greenhouse gas emissions. To date, however, there is only limited empirical evidence and further empirical research is clearly required. Third, IT technologies have boosted the rise of “personalized networking” (Wellman 2001): social relationships are now increasingly person based rather than territorially based, with social media allowing individuals to create their own unique networks for informational and emotional support, sociability, identity formation, and to satisfy belongingness. This may generate more social travel – particularly over long distances – but robust empirical evidence is again scant.

The effects of IT technologies on social networks have also been studied by mobilities scholars. Part of their argument is that IT devices and competencies contribute to individuals’ “network capital” (Urry 2007) to support “mobile lives” – everyday activities
and identities organized around modern and fast transportation and communication systems and characterized by individuality, flexibility, adaptability, and reflexivity (Elliott and Urry 2010). “Network capital” refers to individuals’ capacities to enter into and maintain beneficial social relations at-a-distance through digital and physical mobilities. For Urry this capital produces new social hierarchies alongside, for instance, gender, class, and race/ethnicity. Like other mobility scholars, he draws attention to the ways in which the physical and digital mobilities of certain individuals “immobilize” others. This is because the mobility of the former is made possible by place-based workers who ensure the smooth functioning of mobility systems (from airport baggage handlers to mobile phone company helpdesk workers in countries and regions with ample cheap labor), cleaning and hospitality staff, and other service sector employees.

The work by Urry and other mobility scholars also shows how physical mobility and IT use coalesce into complex hybrids so that mobility practices are intensely and often iteratively coordinated with others using mobile devices; rescheduled on-the-fly; and adapted in response to real-time information from “formal” sources, such as transportation system managers and service providers, and informal expertise provided by fellow users via Twitter and specialized smartphone applications. The experience of being on the move is changing accordingly. There is some evidence to suggest that the coalescence of physical and digital mobilities into hybrids increases a sense of security and comfort but further research on this topic is required.

In short, the interrelations of IT use and physical mobility are changing rapidly. Further research that critically examines the conjectures of academics, the media, and others about those interrelations is therefore desired, and working with young people as frontrunners in the adoption of new ITs and with stakeholders in the freight sector is advisable. However, given that IT use and physical mobility are refracted through place-specific social and cultural contexts and (re)shape sociospatial inequalities, it is equally important to examine empirically who – which individuals, social groups, and economic activities – and which places gain and lose because of changing relations between transportation and IT. The political potentials of the changing inter-relations of transportation and IT also deserve critical attention: developments in smartphone technology are expanding the opportunities for public participation in the governance of transportation systems through crowd sourcing, citizen science, and other initiatives. Perhaps this offers greater potential to make transportation systems more efficient and sustainable than the decades-long quest for substitution of physical transportation by digital mobility.

**SEE ALSO:** Daily mobility; Digital divide; Information and communications technology; Road transport; Time geography and space–time prism; Transport geography; Transport policy; Transportation history; Transportation planning; Urban geography

**References**


Infrastructure and everyday life

Heather Chappells
University of British Columbia, Canada

Background

Infrastructure can be understood as the physical and organizational structures and facilities (e.g., buildings, roads, power supplies, water systems, telecommunication networks) needed for the operation of a society. More than systems of technical hardware, infrastructures embody the rules, standards, norms, and protocols of their structuring institutions and societies. Access to infrastructures enables personal and societal mobility through the circulation of goods, services, knowledge, people, and power. At a macroscale, infrastructures are of key importance for the security of national and global economies and for societal functioning and wellbeing. Critical infrastructures include those that provide services vital to national security, such as energy, water, transportation, and telecommunication systems. Such large technical systems dynamically evolve in accordance with shifting societal and political priorities. Calls for global, national, and regional integration have seen the scale, reach, and interdependency of infrastructures increase. Integration has improved network coordination and service efficiencies but also increased the potential for economic and social disruption from cascading failures. As more infrastructures come under the control of private companies rather than being managed as national assets, the meaning of strategic and sustainable management has shifted, bringing new social and environmental challenges. New flexible and mobile configurations of infrastructure, as seen in the digital convergence of telephony and other media, can both empower and marginalize consumers.

Infrastructure networks support everyday life in the realms of home, work, and leisure through facilitating an array of differentiated service needs. The power we use in our homes, the water we drink, the transportation that moves us, and the communications that connect us form the backbone to everyday life. Central as they are, the criticality of these background networks may rarely be reflected on by those living in industrialized nations. The extent of everyday human reliance on critical infrastructures is highlighted primarily at times of disruption and breakdown, including power cuts, gridlock, and drought. At such moments the sociotechnical resilience embodied within different infrastructural arrangements is exemplified through differentiated opportunities for individual and institutional adaptation, such as utilizing backup systems or modifying demand to meet changing conditions. In the context of climate change and resource scarcity, understanding the scale of potential future disruption and the potential for adaptation in everyday life takes on a renewed sense of urgency.

Scholarly interest in infrastructure and everyday life transcends many disciplines, including geography, sociology, science and technology studies, anthropology, political economy, history, and urban planning. Within these diverse fields, infrastructures have been variously represented as simple technical networks for the delivery of...
INFRASTRUCTURE AND EVERYDAY LIFE

resources, politically defined conduits for the structuring of power relations, and sociomaterial assemblages through which citizen and provider roles in resource management are reproduced. Specific topics and themes reflect macro- to microscales of interest, from understanding the role of large systems in constituting the underlying fabric and flow of urban life, to exploring the local–global interfaces between infrastructure networks or their role in the microstructuring of household routines. How infrastructure shapes institutional dependencies, social inequalities in resource access, and differentiated capacity to respond to crises have been other recurrent themes.

Multiple configurations of infrastructure

Different social, political, and cultural factors have shaped the development of infrastructure networks. Bearing in mind that historical timescales and geographical reference points are fluid, three predominant modes of urban infrastructure development can be discerned. Each mode is associated with distinctive technological, spatial, and social arrangements that reflect shifting political ideologies and changing relations between consumers and providers.

The establishment of early infrastructure networks has been described in terms of a deliberate manufacturing of demand, with the construction of infrastructures to meet emergent new service needs. For example, lighting was one of the first services developed in the electricity sector but this was only needed after dark, so utility providers sought to design other services to spread demand out over the entire day and justify the cost of infrastructural investments. In the context of Britain, early twentieth-century utility networks have been described as localized and piecemeal, with many small networks emerging to meet different user needs for services such as lighting, heating, drinking, bathing, sewage, and solid waste disposal (Graham and Marvin 1995). This patchwork of local systems also produced differentiated user experiences in terms of standards of supply and prices, with service disruption a common occurrence.

A second major configuration of infrastructural development refers to a large-scale expansion, consolidation, and standardization of networks and services, usually at a national scale. Early examples are the expansion of British and American railroads in the nineteenth century to facilitate trade and promote national unity across regions and territories. This required both technological standardization of equipment (e.g., engines, track gauges, signaling) and sociopolitical innovation (e.g., federal regulation, standard time systems, and operating protocols) (Edwards 2003). Later developments during the mid-twentieth century saw energy and water planners in Britain and other industrialized nations use demand models and projections to determine future infrastructural requirements for national economic growth. This “predict and provide” approach to infrastructure planning justified the construction of large-scale reservoirs, networks of power stations, and integrated grids of cables, pipes, and wires sufficient to meet universal demand. In the context of Britain, Graham and Marvin (1995) describe an era of Fordist consolidation with the development of integrated networks operated by large centralized organizations, including nationalized public utility companies. Within this integrated model consumers were positioned as the mass recipients of undifferentiated resources, standardized services, and unlimited supply. Large-scale integration and uniformity of service generally have less saliency in the Global South, where many consumers remain off-grid even today (Edwards 2003).
A third mode of infrastructural development reflects new priorities of political, economic, and environmental restructuring, within what has been termed a market-environmentalist regime (Bakker 2005). Infrastructural unbundling is a key feature of this developmental mode, with previously integrated elements and functions being separated and reorganized to facilitate privatization and competition. In contrast to the centralized, homogeneous, and unified character of large nationalized networks, infrastructures today are heterogeneous and diversified. This heterogeneity may be manifested in new physical grid arrangements, but more often involves restructuring and coordination at an institutional or virtual level. Advances in communication technologies, including smart grids and meters, enable new service configurations that are laid over existing infrastructural arrangements. The supplementation of traditional landline telephony services with satellite and cellular services offered by multiple providers to meet the needs of diverse customer segments is one example of this reordering.

Within this third infrastructural orientation the focus is on managing differentiated consumer service needs rather than meeting universal requirements. The rolling back of universal service provision under market-based regimes has been associated with the intensification of social and spatial inequalities in access to utility services. As national monopolies and public service providers have been replaced by multinational and privatized utility companies, consumer rights to resources and services have been renegotiated. Concurrent processes of social dumping of marginal customers and cherry-picking of lucrative customers have been highlighted in England and Wales in the context of the energy and water privatizations of the 1980s and 1990s, with prepayment meters implicated as a utility-based tool for policing of access to basic services (Guy et al. 2011). Elsewhere in the world, violent disputes over the takeover of local public utility services by transnational corporations have been well documented, as in the struggles of Bolivian water users against water concessions in the city of Cochabamba, or resistance to the use of prepayment meters for electricity and water services in the townships of South Africa.

Infrastructures in interaction with everyday life

Infrastructures are macroscale constructions but also exist at the microscale of the household or workplace where they are implicit in the regulation of everyday life (Shove, Trentmann, and Wilk 2009). Networks enable spatial and temporal ordering and the synchronization of daily interactions that may serve to regulate the lives of citizens through supporting or prohibiting particular activities (Shove, Trentmann, and Wilk 2009). The scheduling of national railways in nineteenth-century Britain and United States required the coordination of regions and activities to standard rather than local time. Later, electrification programs enabled the reordering of activities, which were no longer confined to daylight hours as a result of advances in lighting. Global accounts of drinking-water access further reveal the mediating power of institutions through specific configurations of infrastructure, as in the West Bank where Israeli control of aquifers and piped-water networks has historically served to regulate the flow of water to Palestine. Such situations exemplify how infrastructures are sociopolitical entities supporting power relations through physical entrenchment in material arrangements.

Consumers are not passive recipients of infrastructures but actively shape them. Consumer resistance to emergent infrastructural...
INFRASTRUCTURE AND EVERYDAY LIFE

arrangements has been well documented, as in the aforementioned cases of Bolivia and South Africa. The coordinating role of infrastructures in relation to the modifying role of consumers engages with a fundamental debate in the social sciences concerning the relative role of agency (individual choices) and structure (social and institutional rules and resources) in defining everyday life. Under a universalizing mode of infrastructural development, users have been depicted as passive recipients of undifferentiated goods and services. The dominant market-environmentalist discourse, which underpins contemporary infrastructure arrangements, points to a more active role for consumers in differentiating between service options and providers. Sociologists and geographers have debated this point, demonstrating that collective sociomaterial influences can be just as important as individual choices in shaping consumer practices. This argument is particularly relevant with respect to forms of consumption that are delivered directly via physical infrastructures, such as energy and water grids, which render the home both a private and a public space (Gronow and Warde 2001). This structuring role of institutions is evident in situations where utility providers have exerted control over resource use at times of energy crisis or drought. Recent contributions from sociology, anthropology, and cultural geography offer deeper insights into the engagement of infrastructures within the cultural and material geographies of the home. It has been observed, using the example of energy, that resources are not directly consumed but are used to provide a range of domestic services, such as lighting or comfort, the meaning of which is culturally and materially defined (Shove 2003). Decisions about how to heat or cool a home, or provide a particular ambience, are thus influenced by cultural and collective ideas about domesticity and coziness, as much as by functionality or efficiency. They are also assisted or thwarted by existing material infrastructure arrangements. This can help to explain why seemingly irrational household behaviors persist, as where people continue to use open fires for coziness or retain incandescent lights for a softer glow, when more energy-efficient options are available.

Implications for sustainable infrastructures and the reordering of everyday life

Understanding the way in which infrastructures and everyday life interact takes on a new urgency in the context of contemporary concerns about promoting environmental sustainability and adapting to climate change. Predominant policy approaches to the greening of household consumption have tended to prioritize the agency of individuals as the key to managing resource-intensive lifestyles, as exemplified in the multitude of programs to encourage people to make environmentally conscious lifestyle choices. This approach has had some impact on the intensity of resource use, but it has not tackled trends supporting less sustainable living arrangements, such as increases in house size or more single-person households. The development and diffusion of new kinds of consumer goods, and the global convergence of Westernized comforts and conveniences such as air-conditioning, have further contributed to increasing resource demand. Given the ongoing reproduction and escalation of resource-intensive ways of life, it has been observed that making consumers aware of environmental concerns is not always enough to redefine expectations or break unsustainable habits (Shove 2003). What may be required is a totally new way of thinking about the services that infrastructures provide, people’s cultural attachments to them, and the
alternative sociomaterial arrangements through which they might be met.

The concept of *coprovision* has been utilized to evaluate the relative agency of individuals in relation to the environmental modernization of infrastructures (Vliet, Chappells, and Shove 2005). Within this framework, citizen-consumers are viewed as active partners in reconfiguring utility services, while they are, at the same time, structurally constrained by the different collective sociomaterial arrangements they encounter. An example of new coprovision arrangements is the trend toward the decentralization of infrastructure networks observed in the context of microgrid development. These new infrastructural arrangements can radically reconfigure established relations between consumers and service providers, as in cases where households construct their own minigrids to harvest rainwater or generate renewable energy on-site. Such local infrastructure arrangements further transform space and the ordering of everyday life through establishing proximity to resources and imposing local limits on resource availability that are not necessarily apparent when services are delivered through large-scale centralized grids. These new sociomaterial arrangements have reopened questions first raised in the 1960s about the appropriate scale for infrastructure as a means to promote environmental sustainability and facilitate sociotechnical resilience. However, rather than being framed in terms of finding the optimal system scale, new infrastructural arrangements are seen to offer coexisting sociotechnical trajectories that can produce multiple, flexible, and sustainable reorderings of everyday life.

In countries where centralized access has never been universally available or accessible, localized infrastructural arrangements at a household or community level are often the norm. In the context of intensifying resource shortages and health concerns in developing regions, there has been renewed interest in providing infrastructural solutions that are both sustainable and adapted to local needs. A key message of studies of technological diffusion is the need for culturally adapted innovations that fit within the distinct material geographies of home and utilize local distribution networks (Akrich 1992). For example, in India attempts to diffuse new types of cookstoves based on more efficient and healthier fuel sources than biomass, such as kerosene, have only been partially successful. It has been observed that, if a modern distribution system is not in place, households cannot obtain access to alternative fuels, even if they can afford them. Furthermore, the availability of energy-using equipment is just as important as the affordability of fuels in the switch away from biomass. Moreover, cultural traditions are found to determine household fuel choices regardless of fuel availability and income, as where some affluent Indian households keep a biomass stove to prepare bread in a traditional way. Difficulties in adapting new technologies to incumbent sociotechnical arrangements may thwart the adoption and uptake of infrastructural innovation, but there is also evidence that more flexible arrangements can act to empower citizens. The rapid rise of mobile telephony in Africa since the early years of the twenty-first century serves as an example of how new configurations of infrastructure can act to empower citizens and enable social change. In a region with limited landline network coverage, mobile phones, cellular service, and prepaid financing options have proved to be a flexible and popular communication option. While service inequalities still exist, mobile telephony has improved communication between those living in rural and urban areas, reduced the need for travel, and provided access to services such as banking from which many citizens were previously excluded.
Critical infrastructures, disruption, and resilience

Disruption can provide further insights into the tenuous and dynamic relationship between infrastructures and the organization of everyday life and reveal options for restructuring in the context of calls for sustainability and resilience. It has been asserted that greater material and technological dependence has increased the risk of breakdowns in everyday life and that the increased sophistication and complexity of networks reveal the possibility of systematic and cascading failure in critical infrastructure (Trentmann 2009). In this respect, it is important to consider where the points of criticality lie within current infrastructural arrangements and how institutions and governments might reconfigure systems to be adaptive to these challenges. As recent blackouts have shown, relying on air-conditioning for comfort is not a robust solution during heat waves when electrical networks are already overburdened. In such situations more creative and innovative social and cultural adjustments, such as the Cool Biz campaign in Japan, focused on adaptations in clothing, or “forced” vacations and homeworking may offer solutions that are both socially and technically resilient to changing conditions.

Drawing on the lessons of coprovision, more fluid and flexible arrangements, including more decentralized grids and local backup systems, may offer further adaptive opportunities where large-scale interconnected networks fail. In this sense, lessons might be learned from the history of infrastructure development and the relative flexibility that different sociotechnical arrangements have embodied for providers and consumers at past points of criticality. Solutions that are adapted to the local context and to the needs and expectations of different consumers are also highlighted from a sociopolitical perspective. For example, appeals to citizens to curb energy use, as successfully utilized in the 2001 California energy crisis, have been resisted in situations where citizens have viewed continual access to resources as a right. In this context, renegotiating the entitlement to universal service such that consumers are willing to practice demand-side management at times of acute need might be required, while ensuring that those most in need of essential services maintain access. From a geographical perspective, certain places and people are also more likely to be affected by infrastructural breakdown, as illustrated in the Chicago heat wave of 1995, when social infrastructure and neighborhood environment dramatically shaped adaptive capacity, with the most isolated members of society struggling the most to adapt. Such cases have highlighted the need to address inequalities in access to service provision and the importance of building community-scale resilience that is inclusive of all sectors of society.

Finally, it has been observed that no one mode of infrastructural configuration is likely to guarantee smooth flow and universal access at all times, and that the rupture of routines is a perfectly normal feature of human society (Trentmann 2009). From this perspective it is instructive to reflect on the purpose of infrastructure within the realm of everyday life, and on what resilience and sustainability actually mean from a sociocultural as well as an engineering vantage point. This approach might reveal alternative system configurations that respond to variable user requirements for continuity or discontinuity of service, as where industry is asked to shut down systems during times of energy crisis, rather than follow a more traditional approach of building more robust but ultimately unsustainable systems to meet every need and eventuality. This line of argument supports a view of infrastructures not as providing resistance to floods, droughts, or power outages through
material reinforcement, but as embodying fluid resilience that encompasses socially and culturally appropriate modes of adaptation.

SEE ALSO: Consumption; Energy resources and use; Environment and everyday life; Infrastructure

References


Further reading


Infrastructure and regional development

Kingsley E. Haynes
Zhenhua Chen
George Mason University, USA

Definitions

First, infrastructure is defined as the basic physical and organizational structures that provide support for the economy and society to function, and is divided into two categories: (i) physical or economic infrastructure and (ii) soft infrastructure, also referred to as social overhead capital (Haynes and Nijkamp 2006). The first includes all kinds of capital facilities such as buildings, water and sewer systems, road networks, railway and air transport, and telecommunications. The second includes parks and recreation areas, open spaces, social welfare support, education facilities, and health and hospital systems. Infrastructure contributes to regional development by facilitating mobility and communication and providing the physical and social underpinnings that improve the quality of life.

Regional development is, however, a different concept, which encompasses a range of objectives and concerns, centered around goals of enhancing the base of jobs, income, and business activity in a region. Its scope includes the process and policies by which a region improves the economic, political, and social wellbeing of its people (O’Sullivan and Sheffrin 2003). The term includes institutional shifts that vary between short-term concerns and long-term strategies to improve regional function and social and economic wellbeing.

Infrastructure is central to the support of high-density environments and, hence, underpins regional agglomeration and urbanization. Everything from water and wastewater services to transportation are central to infrastructure investments. Infrastructural investments are strongly concentrated geographically in urban areas with interregional infrastructure connecting cities and regions and connecting urban centers to labor, energy, water, materials, and so on. These considerations link the discussion of agglomeration economies and productivity with the drivers of urbanization.

Infrastructure plays a vital role in promoting regional economic development. All levels of government and international organizations invest heavily to improve infrastructure to facilitate economic activities. Every year the World Bank provides a large amount of financial support to developing countries in order to help them build infrastructure and promote regional development. In the past decade, China has experienced a rapid expansion of massive transportation infrastructure with strong central government support. A national high-speed rail network with a total distance of over 17,000 km has been built to boost regional economies. In the United States, infrastructure investment is one of the major policy instruments being examined to stimulate economic growth and development. During the great recession in 2009, US$30 billion was dedicated to infrastructure investment in the American Recovery and Reinvestment Act of 2009 with the purpose of stimulating economic recovery and job creation.
INFRASTRUCTURE AND REGIONAL DEVELOPMENT

Mechanisms

How does infrastructure facilitate regional development, and what are the linkage mechanisms between infrastructure and regional development? To answer these questions, one should note that the influence of infrastructure on regional development varies during different periods in its life cycle. During the construction period, its impact on regional development is achieved through capital expenditure and resource accumulation. The need for labor and capital input drives the expansion of consumption which promotes the growth of gross regional product. Since economic output depends heavily on factor inputs, the impact of infrastructure investment on regional development is relatively immediate and short-term through reducing the costs of factor input accumulation and agglomeration support to increase productivity.

After completion of an infrastructure project, its contribution to regional development is more likely to be achieved through network and spatial spillover effects. For most infrastructures, such as road, railway, airport, and broadband, regional connectivity and accessibility are normally achieved after network completion. Regional development is thus expected to occur at multiple locations along the network through spatial spillover effects. The decline of mobility and communication costs further facilitates economic activities through the increase of both demand and supply and the support of specialization. Gross regional product and the employment rate are thus expected to increase.

Impact assessment

To empirically assess infrastructure’s impact on regional development is not an easy task. The assessment has to be based on economic and geographic analysis and is subject to data availability. While the impact assessment of infrastructure can be conducted in various ways, the following economic theories have been most widely applied: neoclassical economic theory, spatial economic theory, and general equilibrium theory. For the ease of presentation, only transportation infrastructure is referred to in the following discussion, but it is just one of a set of infrastructure investments that can be assessed in this way.

Neoclassical economic theory

Impact assessments of infrastructure on regional development were not well developed until the emergence of a series of papers starting in 1989 by Aschauer. The fundamental argument Aschauer made is that the enhancement of infrastructure provision through public expenditure facilitates a region achieving its economic potential. Since then, a large number of studies evaluating the regional impact of public infrastructure investment have been conducted by following neoclassical economic theory through some form of aggregated production function approach (Gramlich 1994; Fernald 1999; Boarnet 1998). Because of different evaluation methods, time periods, measures of economic outcomes, and control variables used, the findings of these studies are not consistent. For instance, some find that the US highway infrastructure has a positive and significant effect on productivity (Fernald 1999), although such an effect diminished after the completion of the systems (Fernald 1999). Others cast doubt on even the positive effects of public infrastructure by adopting different data and evaluation methods (Boarnet 1998).

Specifically, Aschauer (1989) and Munnell and Cook (1990) analyzed the relationship between public infrastructure capital and economic performance for the period 1970–1986 at the US national and state level, respectively.
The elasticity of output (economic returns) with respect to infrastructure was found to range between 0.38 and 0.56 in Aschauer’s study and was 0.15 in Munnell’s study, with highway alone contributing over a third of that benefit. The elasticities of output (economic returns) with respect to infrastructure indicate that the amount of output would rise and returns would increase for each percentage increase in the nation’s infrastructure stock.

Scholars also argue that the scale of analysis matters in this kind of assessment. The rate of return may decline in significance as the scale of analysis is altered. By using a partial equilibrium model as well as state-level public capital data, Holtz-Eakin and Lovely (1996) found that public capital did not affect output significantly. Many of these studies have been subject to a range of criticisms. Gramlich (1994) provides the following summary on the defects of these studies.

- Unclear causal relationship between infrastructure provision and economic performance.
- Vague definition of “infrastructure” making the quantitative analysis speculative.
- Policy variables that are too short-term to be consistent with relevant infrastructure variables.
- Isolation of factors influencing macroeconomic performance: where transport to soft infrastructure operates in the context of legal, educational (human capital), and business institutions, and in terms of national defense priorities.
- Different methodologies applied to different datasets, resulting in implications that are attributed to imprecise quantitative estimation.

Further, earlier studies did not consider the spatial interactions among units across different geographic locations. These early analyses assumed the existence of spatial homogeneity. Further, as the geographic scale of the research area changes, estimated impacts change as well. Based on a meta-analysis of a large number of studies discussing highway infrastructure and the economy, Shatz, Kitchens, and Rosenbloom (2011) indicate that the effects of highway infrastructure on economic output vary when different levels of data are applied. They concluded that studies tended to find higher rates of return and strong productivity effects of highway infrastructure at the national scale than at the state and substate scale.

Spatial economic theory

The assessment of infrastructure impact on regional development has involved the application of spatial economic theory, which considers the nature of spatial dependence and heterogeneity. Munnell and Cook (1990) pointed out that the impact of infrastructure on regional development became smaller as the geographic focus narrowed. She believed that this was due to the effects of leakages from infrastructure investment that could not be captured at a small geographic scale. Although such a hypothesis may not be fully accurate, as indicated by Boarnet (1998), it does suggest that the spatial dimension has an influence on the estimated impact and should not be ignored.

LeSage and Pace (2009) emphasize that traditional econometrics have largely neglected the spatial/geographic dimension of sample data. When data has geographic information, two issues arise due to the violation of the Gauss–Markov assumptions. The first one is spatial dependence between observations, and the second is spatial heterogeneity. Without considering these spatial issues, any estimation may be statistically biased.
Thanks to the development of spatial econometric techniques by Jean Paelinck, Luc Anselin, James LeSage, Paul Elhorst, and many other regional scientists, a number of empirical analyses with spatial considerations have been carried out. One of the compelling functions of spatial econometrics is to allow for measuring spatial spillover effects. This refers to the situation in which the input in one sector or region influences changes in neighboring economies through trade linkages and market relationships. Infrastructure is likely to have a spillover effect on regional development because the benefits generated from infrastructure improvement would not be confined to the specific region where the improvement occurred. This hypothesis has been empirically tested in various spatial models (Holtz-Eakin and Schwartz 1996; Kelejian and Robinson 1997; Cohen and Morrison Paul 2003). As yet, there is no consistent conclusion on whether or the degree to which spillover effects of infrastructure are positive and significant.

Boarnet (1998) constructed a spatial lag model to evaluate the spatial effects of public transport infrastructure in Californian counties. He found a negative spatial effect for the Californian road system, due to migration. Holtz-Eakin and Schwartz (1996) found that highway stocks do not have substantial spillover effects on private productivity. Kelejian and Robinson’s study (1997) found that the results are quite sensitive to model specification.

On the other hand, positive spillover effects of transportation infrastructure have been found (Cohen and Morrison Paul 2003). Cohen and Morrison Paul conducted a series of studies aimed at the benefits of airport, highway, and port investments accruing to the US manufacturing sector. They apply cost functions with considerations of spatial autocorrelation adjustments to data from 1982 through 1996. In their analysis, positive and significant spatial autocorrelation parameters were obtained, which they concluded are indications of positive spillover effects. However, in terms of ports, they found that the elasticity of the shadow value of a neighbor’s port improvement with respect to their own state’s port infrastructure was negative and significant.

A brief overview of the literature on infrastructure and regional development suggests that the conclusions are not consistent, given that different data, methods, regions, and periods are employed in each analysis.

General equilibrium theory

According to the neoclassical and spatial economic theories, impact assessment of infrastructure on regional development is conducted under the assumption of partial equilibrium. The linkages between economic output and infrastructure input are assessed only from the supply side by assuming the demand of infrastructure remains constant during any specific period of investigation. Clearly, the outcome of economic impact is only partial since the impact caused by the change of demand is not captured under such a theoretical framework. Practically, the impact of infrastructure on travelers’ welfare measured by levels of utility cannot be adequately measured under a partial equilibrium analysis alone. In order to achieve a comprehensive regional development impact evaluation of infrastructure investment on both the supply and demand sides, a general equilibrium theoretical framework is required.

A computable general equilibrium (CGE) model developed by Johansen in the 1960s enables impact analysis with a consideration of both demand and supply. The theoretical framework relies on the Walras–Arrow–Debreu theory of general equilibrium with modern modifications and extensions allowing for imperfect
markets. Because CGE provides a clear linkage between the microeconomic structure and the macroeconomic environment, the model can be used to not only describe the interrelationship among multiple industrial sectors and markets, but also to assess direct and indirect effects from the change of public policy on various kinds of economic variable, such as output, employment, prices, income, and welfare (Chen and Haynes 2014).

The applications of CGE in evaluating the impact of infrastructure on regional development are abundant. Depending on the stages of infrastructure provision, impacts can be evaluated differently. For instance, the short-term impact of infrastructure investment (such as job creation and demand for raw materials) can be measured by examining linkages of the infrastructure sector with other economic agents, including consumers, producers, and government. On the other hand, the long-term impact of infrastructure (such as a reduction of transportation costs due to network improvements) can be evaluated by examining the variation of trade margins among different regions. Because the impact can be measured under a general equilibrium among multiple geographic regions, this modeling structure is often called spatial CGE or SCGE, but is in reality an interregional CGE model.

In recent decades, different types of SCGE models were established to evaluate the impact of infrastructure investment policies. For instance, Miyagi (2006) evaluates the impact of infrastructure in Japan in relation to the accessibility change using SCGE. In his model, the impact on regional development was measured through a reduction of congestion due to a specialized infrastructure investment (Miyagi 2006). Another SCGE analysis applied to transportation infrastructure evaluation was conducted by Goce-Dakila and Mizokami (2007) with a focus on the Philippines. They established an SCGE model with consideration of five subregions and found that technological improvement in land transport has the highest impact on output.

Haddad and Hewings (2005) assess the regional impact of changes in Brazilian road transportation infrastructure by applying a multiregional CGE model. By introducing non-constant returns and non-iceberg transportation costs, they find asymmetric impacts of transportation investment on the spatial economy of Brazil.

In addition, there are a few different types of SCGE models developed for the regional impact assessment of infrastructure. For instance, CGEurope is another SCGE model developed by Bröcker (1998) with a primary use of spatial analysis in the distribution of welfare effects linked to changes in accessibility within and between regions. The MONASH model developed by Dixon and Rimmer (2000) is another widely used multiregional and multisectoral dynamic CGE model that can be used to assess the impact of infrastructure on regional development. Their model allows for different choices of levels of sectoral and regional disaggregation, and infrastructure such as transportation can be treated as a marginal sector where costs are imposed on the purchase price of goods in trade and services.

Major issues

One should note that the understanding of infrastructure and regional development is still limited despite the development of modern techniques of economic modeling. Given the existence of issues of equity, causality, geographic scale, and different types of spillover effect, further research is needed.

Equity of investment

One major concern of infrastructure policy is the equity impact of investment both in terms
of geographic locality and types of infrastructure. Despite the fact that the primary goal for many national governments is to promote equal opportunities for economic development among subnational regions through improved infrastructure systems, some studies have found that sometimes the effects of infrastructure investment can produce the opposite outcomes. For example, the original objective of the Japanese Shinkansen high-speed rail network was to reduce regional disparity. The actual outcome was not as expected. Factors such as labor forces were found to be more moveable than expected. One consequence was that urban metropolitan areas continued to agglomerate, leaving other regions behind and underdeveloped. Therefore, how to invest in infrastructure wisely so as to reduce regional disparity continues to be a challenge for decision-makers.

The issue of equity also exists when decisions have to be made on how to allocate funding between different types of infrastructure. For instance, roads are the dominant mode of infrastructure in the United States. Studies confirm that highway infrastructure contributes the most to regional economic development from all modes of public infrastructure, including airports, ports, and public railways. Given the economic impact of highways and the level of existing capital stock, the question is whether new public financial resource should still be allocated equally to different modes, and how that relates to the efficiency consequences of infrastructure investment.

Causality

The nature of the causal relationship between infrastructure and regional development is another critical issue for impact assessment. This issue is sometimes known as the endogeneity problem. Infrastructure investment enhances the connectivity of regional transportation networks, which subsequently facilitates both freight and passenger movement through the reduction of transportation costs. On the other hand, the improvement of economic performance may, as a result, lead to an increase in demand for both freight and passenger transport and thus require more investment for infrastructure improvement. Such results are also referred to as a reflection of latent demand. Failure to appropriately address the endogeneity issue may jeopardize the outcome and interpretation of impact investigation and lead to erroneous policy implications.

Geographic scale

Since an impact assessment of infrastructure on regional development needs to be conducted at a specific geographic scale depending on the research objectives, how to choose the right geographic scale becomes a sensitive issue. As noted above, the variation in impact assessment with respect to scale has been discussed extensively by Shatz, Kitchens, and Rosenbloom (2011). Studies using national rather than state-level or substate-level data were more likely to find a positive and significant relationship between infrastructure and economic outcomes. One reason for these differences is the observation unit’s spatial stability, as there is a tendency for highway infrastructure to reallocate economic activity. Therefore, negative spillover effects are more likely to be found at a smaller geographic scale. On the other hand, a national-level assessment may be more likely to capture geographically distant (and often positive) spillovers (Shatz, Kitchens, and Rosenbloom 2011).

The sensitivity of the scale of analysis also relates to methodologies for spillover estimation. To effectively account for the spillover effects of infrastructure, spatial econometric techniques are usually adopted. Boarnet (1998) found a negative
spatial lag effect for the Californian road system on economic output at the county level. However, based on a different analytical scale, Jiwat-tanakulpaisarn et al. (2009) found a completely different answer across the interstate highway system: that growth in roadway density had positive spillover effects on state employment growth.

Overall, the tradeoff between geographic scale of analysis and interpretation of results are issues that deserve significant attention in the assessment of infrastructure impacts on regional economic development. While the assessment may be easier to implement at the aggregated level (such as the national level or state level) given data availability, the assessment becomes challenging at the disaggregated level (especially at the Metropolitan Statistical Area (MSA) or county level) as data is usually not available and capital stock has to be estimated. As a result, it should be noted that as the data being estimated move away from the “ideal,” so does the level of confidence in the results. Hence it is important for any user to carefully review and assess such geographic disaggregation for the level of confidence needed.

Global versus local spillover effects

The spatial spillover effects on regional development from infrastructure facilities have been widely observed in empirical studies. But how to measure these effects appropriately and accurately is still challenging. One of the main reasons for this is due to lack of sufficient information to determine the impact scale of infrastructure. This directly relates to the robustness of impact estimation given the sensitivity of spatial modeling specifications. Some studies consider only local spillover effects but, as infrastructure networks provide accessibility and connectivity from one region to adjacent regions, therefore the impact on regional development involves not only neighbors, but high-order neighbors (neighbors to the neighbors, neighbors to the neighbors of the neighbors, and so on).

However, in other cases, the influence of infrastructure on high-order neighbors’ regional development may exist and/or may decline with distance: for instance, when highway congestion occurs, the behavior of commuters in different regions is affected, given that feedback or endogenous interaction effects are present (LeSage 2014). As a result, the congestion of highway infrastructure may have global spillover effects on regional development despite the fact that the magnitude of the effect decays as the distance to the core increases.

How to choose the appropriate spatial spillover effects for an infrastructure’s regional impact assessment is fundamental. Although spatial modeling techniques and application principles have been extensively discussed (LeSage 2014), given the uncertainty of data availability and different theoretical foundations being applied, challenges confronting assessment of the impact of infrastructure on regional development remain.

SEE ALSO: Monitoring and evaluation; Regional development policies; Regional economic integration; Spatial organization and structure

References


Infrastructure plays a vital role in structuring the relative spatial connectivity of place. Infrastructure systems tend to be immobile but facilitate numerous mobilities which, at their core, provide the mechanisms and context through which modern life functions. They enable the process of time-space compression and shape the relational networks through which localities are articulated within broader social, economic, political, and environmental systems. They expedite technological transformations and sociospatial change, and foster new spatial imaginaries. The Internet, for instance, is celebrated as the backbone of a new era of connectivity and progress. Yet, while the “information age” is premised on the idea of uninterrupted digital flows and circulation, such discourses render other infrastructure systems mundane and overlook inequalities in access and processes of uneven development.

Infrastructure is usually understood in physical terms. Technical, or hard, infrastructures are physical systems consisting of material elements—highways, pipelines, power stations, cables, energy grids, airports, fiber optics, sewers—that mediate resource flows. Transportation, water, energy, trade, and telecommunications networks materially connect more or less distant places by facilitating various social processes and relations across space. The development of technical infrastructure systems plays a vital role in the development of territorial units from cities to nation-states. At the same time, by supporting processes of globalization via telecommunications and trade innovations (such as containerization and intermodalism), technical infrastructure provides an important means to challenge the construction of places as bounded, internally organized, territorial units.

Variations in the density and concentration of technical infrastructures spur distinct patterns of uneven geographical development. Infrastructure systems can invoke distinct environmental and social problems, from local events (oil spills, water pollution) to global crises (climate change); however, the transformative potential of engineered systems also provides the potential to realize new spatial fixes. Infrastructure megaprojects, as material and symbolic spaces, have been closely associated with programs of modernization. They often function as a means, and context, for sustainable development while, in the wake of the 2008–2009 global financial crisis, infrastructure investment has emerged as a key policy tool to reinvigorate local and national economies.

Infrastructure may also be understood through the production and operation of social relations. Social, or soft, infrastructure consists of the formal institutions and informal practices employed by various actors (individuals, households, etc.) that structure the capacities of people in place. The relative “thickness” of social networks strongly influences the practices and experiences of everyday life. Social infrastructures, for example, public services or utilities, may be provided by governmental agencies. They may also be forged by a diverse array of actors operating at multiple scales—from small-scale cooperatives to transnational organizations—when the state is unable or unwilling to provide them. Simone (2004) directly ties the concept of infrastructure to
people’s activities in cities by foregrounding the economic collaborations between marginalized residents in Johannesburg. Examining the case of urban markets, he frames practices of “cooking, reciting, selling, loading and unloading, fighting, praying, relaxing, pounding, and buying … on stages too cramped, too deteriorated, too clogged with waste, history, energy, and sweat to sustain all of them” as providing the concrete acts and context through which the city is reproduced (Simone 2004, 426). In this context, people’s networks and rules, which are dynamically invoked and reinforced, form the infrastructure conditioning social practice.

Entering the twenty-first century, the nature and focus of geographic inquiry shifted alongside a broader reappraisal of infrastructure studies across the social sciences. Geographic engagement with infrastructure, particularly those in the sphere of transportation geography, has strong roots in the discipline’s quantitative turn. Systematic approaches to the mapping and measurement of spatial processes developed through the 1950s and 1960s helped establish geography as a science and enabled geographers to inform public policy and investments decisions. Despite a long-standing interest in social justice issues within geography, infrastructure systems consequently tended to be relegated to the apolitical domain of engineers and technocrats (see Furlong 2010; Graham 2010). Over the past decade, a groundswell of critical analyses – drawing on a diverse amalgam of theoretical frameworks – has focused on examining how infrastructure profoundly shapes the production, experience, governance, and transformation of social relations. Work conducted under the rubric of the politics of infrastructure, techno-natures, urban metabolism, and the sociotechnical city have encouraged a reconsideration of geographic engagements, spurred, in no small part, by a reinvigorated commitment to “[open] up the ‘black box’ of urban infrastructure to explore the ways in which infrastructures, cities and nation states are produced and transformed together” (McFarlane and Rutherford 2008, 364).

In contrast to scholars in other disciplines – for example, science and technology studies, which concentrates on the technological or engineered aspects of infrastructure systems – geographic scholarship has tended to foreground social, political, and economic factors to demonstrate, empirically and conceptually, that infrastructure systems are not isolated, apolitical, static, or stable entities. Inequalities in access and mobility produce distinct power relations and articulations of uneven development that position them as central objects of social struggle. Developments like urban growth and shrinkage or public budgetary crises can challenge traditional forms of infrastructure provision and require new technical or social solutions. Moreover, scholarship influenced by the new mobilities paradigm challenges normative understandings of infrastructure stability by drawing attention to how, at a microscale, infrastructure may be characterized by dynamism and change. Seemingly fixed, material objects are constantly being modified and refashioned in subtle ways: streets are repaved, buildings painted, grass mowed.

Poststructural approaches, notably drawing from actor-network theory and cyborg studies, have attempted to conceptually collapse the distinction between the human body and technological networks. Infrastructures are conceived as a series of interconnecting unconscious life-support systems that make urban life possible. For instance, the modern home – with its provisions of light, heat, and water and telecommunications networks – is a normalized yet essential exoskeleton that blurs the distinction between the organic and the technological. Shifting scale, the hybrid urbanization embodied within the cyborg city produces urban space as
an inseparable fusion of the social and technical. Marxist scholars have engaged such normalization as a mode of fetishism that obfuscates the social relations that underpin the production of infrastructure systems. Scholars utilizing assemblage theory have further problematized notions of agency by focusing analytical attention on sociomaterial interaction. Assemblage theory constructs infrastructure systems as bringing together and organizing multiple human and nonhuman relationships in a manner that distributes agency beyond the human actants involved. Bennett (2005), for example, interprets the 2003 North American blackout as a moment of crisis rooted in the specific arrangement of flows, users, commodities, production processes, lifestyles, profit motives, and electron streams bundled in the specific infrastructural constellation of the East Coast electric grid.

Splintering urbanism

The form, function, and governance of infrastructure networks vary across geographical and historical contexts. Graham and Marvin’s (2001) “splintering urbanism” thesis has proved a highly influential analytical framework to conceptualize broad transformations in the production and management of sociotechnical systems. Between 1850 and 1960, urbanization, especially in advanced capitalist countries, catalyzed a movement from piecemeal and fragmented infrastructure provision toward an emphasis on the centralized and standardized systems that underpinned the modern networked city. The intersection of modernist aesthetics and technology promoted rationality and order in the production of urban space. Infrastructure systems served as both functional, material networks and symbolic representational spaces that spurred dreams of mobility, modernity, and circulation. The modern infrastructural ideal, as a decidedly Western construct, was buttressed by an ideological belief in the positive social transformative capacity of networked technologies. The ascension of modern theories and practices of urban planning helped codify new ways of thinking about and shaping cities and their sociotechnical relations. Government support for near-universal access to infrastructure networks across urban, regional, and national space was vital to the extension of the networked technologies that facilitated new forms of mass production and consumption. The Fordist New Deal is commonly accepted to be the nadir of technological modernism in the United States, with the machine emerging as a motif for both industrial production and social organization.

Across the globe, postwar modernization programs were characterized, and defined, by standardized modern infrastructures; from highways and high-rise residential tower blocks to vast electric, water, and sewage grids. However, the ascension of the modern infrastructural ideal fostered a concomitant critique of dehumanizing and alienating impacts of technological modernism. By the early 1970s, social critiques regarding the lived experience of high modernism, perhaps most influentially in the writings and activism of Jane Jacobs, undermined the development of infrastructural networks as idealized technological-engineered systems. As the long postwar capitalist boom subsided, governments struggled to invest the constant inputs of capital and labor required to maintain modern infrastructure networks which consequently became vulnerable to protracted fiscal crises and physical decay into the 1980s.

Processes of political and economic restructuring following the crisis of Fordism directly impacted the planning, management, and governance of infrastructure systems. Planning rationales that legitimized the construction of modern integrated infrastructure systems were undermined by increased technical specialization.
and a gradual shift in attention from concerns regarding built form and mechanic metaphors to administrative, legal, and social issues. Under the auspices of neoliberalism, the logics of infrastructure provision have shifted from the modern ideal of public provision and universal access to collectively distributed services, toward the valorization of individual choice and atomized mobility, in a manner that obscures the continued reliance on public infrastructures that enables such mobility. In lieu of nationally scaled spatial fixes, local governance units have taken on increasing responsibility for developing the urban infrastructures necessary to support growth in their own territorial jurisdictions. Infrastructure restructuring, in terms of both material networks and their governance regimes, has provided a lens to reveal both the processes of deterritorialization associated with globalization and the rise of the “network society,” and the modes of reterritorialization through which new scalar relations are produced.

A key mechanism here is the cleaving, or secession, of infrastructure elements from collective public systems. Publicly managed infrastructures have been increasingly splintered through processes of deregulation and privatization (Graham and Marvin 2001). The unbundling of existing infrastructure networks establishes premium network spaces (e.g., toll roads, privatized express rail links) that are integrated into selective global political economic frameworks through specialized development funds, financial tools, and public–private (P3) partnerships. Material, political, and economic relations foster new topological geographies that tie together a privileged archipelago of elite global nodes reformulated in a manner that constructs and reinforces sociospatial relations. In an era of free trade, just-in-spatial production, and globalized supply-chain networks, the production of premium infrastructure networks enables localities to create competitive advantages while erecting barriers to entry for their competitors. City-regions that are able to construct world-class infrastructure, develop multimodal transportation centers, and lower transportation costs greatly strengthen their competitive position in the international economy. Consequently, several planning scholars now consider strategic investment in infrastructure as presenting a new spatial planning paradigm, with urban infrastructure planning held as a potentially visionary yet pragmatic tool for planners.

Local units of governments have subsequently adjusted how they perceive and utilize their infrastructure assets. As austerity regimes limit the public capital available to invest in public infrastructure, a major trend in local urban policy has been the financialization of infrastructure networks; either through engaging in (P3) partnerships or through selling them off outright. Technical infrastructure has acted as an experimental testing ground for P3 funding arrangements. A common form of P3 arrangement enables governments to lease the operation of such infrastructure to a private company over a limited time frame for a lump sum payment while retaining ownership of the physical systems. The City of Chicago has exemplified such strategic unbundling through landmark leasing arrangements for the Chicago Skyway and municipal parking meters. A broad global trend is emerging in which the public ownership and local management of technical infrastructure is usurped by supranational governance regimes, whereby infrastructure systems are privately owned by global companies and regulated by local actors.

Critics of P3 arrangements point to the dangers associated with local governments’ reliance on increasingly risky financial arrangements and security-backed speculation. The production, financing, and regulation of urban infrastructures
produced through supranational governance deepens the multifaceted and multiscalar connectivity of places, but in doing so opens local struggles over collective consumption amenities to the disciplinary logic of private capital. Moreover, unbundling has profound implications for class struggle and environmental justice by engendering differential access to infrastructure networks. New articulations of uneven geographic development intensify sociospatial polarization with metropolitan space. Places that are physically bypassed by globally privileged networks suffer from limited material and social connectivity and are often discursively framed as corridors that require traversing rather than as spaces of habitation (see Young, Wood, and Keil 2011).

Disruptions, crises, and consequences

Investigations into the political production and transformative capacity of infrastructure systems have pointed to a central paradox. While infrastructure systems are essential to our everyday lives, their ubiquity renders them invisible, as normalized, and taken for granted. Oftentimes it is only once systems break down, fail, or are disrupted that their materialities, roles, geographies, and social functions are revealed (see Graham 2010). This is the case for both technical and social infrastructure. Differing infrastructure systems are themselves fused together in complex and inter-dependent relationships. Consequently, a crisis arising at a specific point may quickly spread through other infrastructures and networked places. Crises can be place-based. Natural disasters may strike specific locales, as when Hurricane Katrina hit New Orleans in 2005, or localized fat deposits in city sewers may cripple basic sanitations systems. They can also be distinctly reticulated. Disruptions to transnational mobility caused by the spread of the SARS (severe acute respiratory syndrome) and H1N1 (swine flu) viruses through global air networks revealed how infrastructure-based crises cascade between places. Increases in the visibility of infrastructural failures mean infrastructure disruptions – from the challenges of climate change to the threat of Internet worms and identity theft – can become ingrained as a normalized expectation within modern society. Moreover, infrastructure networks have emerged as a mechanism for political insurgency, whether in the form of infrastructural terrorism (most notably in the case of the September 11, 2001, attacks in New York and Washington, DC), Anonymous attacking governmental websites, or First Nations protesters in Ontario blocking major highways to gain visibility for their cause. Many recent state interventions around infrastructure networks have therefore been marked by a concern with securitization and surveillance. Infrastructure disruptions are experienced differentially across geographical and social contexts. In advanced capitalist countries, infrastructure tends to be normalized until large-scale crises viscerally insert them into political and economic mechanisms. By contrast, infrastructure disruptions tend to be foregrounded for precarious social groups whose lack of access or relative disconnectivity leads everyday life to revolve around a constant struggle to obtain adequate water, food, sanitation, and mobility. Several scholars have illustrated that the epochal shift between the modern networked city and the unbundling neoliberal city region has unfurled in uneven geographically and historically unstable patterns, and continues to do so (see Coutard 2008). Critics drawing on evidence from cities in the developing world demonstrate that the construction and implementation of the modern infrastructural ideal was far from universal. Rather, standardized networks developed unevenly and exhibited
significant geographic variations within and across national contexts. Historical analyses also complicate simple narratives of the bundling and unbundling logic of infrastructure networks by disclosing an ambivalent relationship between standardized service provision and increased levels of urban integration and the complex relationship between publicly owned networks and the driving role of private interests in infrastructure construction (see Coutard 2008).

Despite, or even because of, their technical and fiscal vulnerabilities, infrastructures are not only vital in demarcating the practical possibilities of governance regimes but also crucial in defining the ideological parameters of political discourse (Gandy 2005). For instance, as splintering urbanism is the product of strategic coalitions within multiscalar governance regimes, unbundling processes are open to social contestation and political intervention. Flexible networks and creative investment strategies can open possibilities for future urban growth and development. As an unstable and multistage process, infrastructure splintering fosters fissures in which new modes of social and spatial justice, as well as collective action, can emerge. Network splintering may cleave off premium network space, but differentiated service provision within public networks enables institutional and financial capacity to better serve marginalized users and urban inhabitants. Contestation over infrastructure production and a rescaled territorial politics of collective provision can animate political movements centered on class struggle at broader spatial scales, as seen in the mobilization of the Los Angeles Bus Riders Union. The struggle between global forces controlling commodified networks and attempts to democratize infrastructure systems will likely form an increasingly central component of urban, national, and international politics.

There is considerable scope for conceptual and applied geographic research to probe the limits and possibilities of political movements around a politics of infrastructure, particularly at the interdisciplinary nexus of political economy and ecology, critical urban studies, and security studies. While much of this research may center on major societal shifts and moments of crisis, significant conceptual and political insights can also be revealed by uncovering the everyday adaptability and transformations of infrastructure, including stressing the role of mediating technologies in influencing infrastructure provision and adaptability. The elaboration of multiple scalar perspectives offers a productive avenue to further examine infrastructure’s role in shaping the governance practices, progressive development frameworks, and the spatial processes conditioning contemporary social relations and everyday life (see Furlong 2010; McFarlane and Rutherford 2008).

SEE ALSO: Actor-network theory; Built environments; Mobility gaps; Neoliberalism; Socio-nature; Technology; Topological relations; Urban political ecology; Urban politics

References


Gandy, Matthew. 2005. “Cyborg Urbanization: Complexity and Monstrosity in the Contemporary


Innovation and regional development

Chun Yang
Hong Kong Baptist University, China

Over the past decades, innovation-driven growth has become a major driver of development dynamics in both advanced and emerging economies. Innovation is considered a core growth driver in knowledge-based economies and a subject for wide-ranging policy efforts. It is also regarded as a socially and spatially embedded interactive learning process that cannot be understood independently of its specific national/regional institutional and cultural contexts. Since the 1990s, the geography of innovation has become an important field of research in economic geography. The renewed interest in the relationship between innovation and regional development has arisen due to the performance of a few highly innovative industries such as the Silicon Valley electronics cluster in southern California, Emilia-Romagna in Italy, the science-based industrial cluster around Cambridge, England, the Hsinchu Science Park in Taiwan, and Zhongguancun technology hub in Beijing, China.

Regions have become key actors shaping the generation and diffusion of new knowledge in the global economy. Policymakers increasingly recognize the role that regions play in cultivating, attracting, and retaining innovative people and firms. The main challenge for regional innovation policies is to ensure a favorable environment for the diffusion of innovations and for entrepreneurship and business growth. In order to understand the relationships between innovation and regional development, this entry will start by defining and measuring innovation. Then the scales and spaces of innovation over the last few decades will be explored. Particular emphasis is put on the evolution of innovation systems approaches. An investigation of empirical experiences of innovation and regional development in developed countries and emerging economies, particularly China, will follow. A summary and agenda for future research is presented in the conclusion.

Defining and measuring innovation

The earliest conceptual exploration of the role of innovation in economic development is the Schumpeterian theory of entrepreneurial innovation. To Schumpeterians, innovation can be regarded as the creative destruction of the old economic system and its replacement by the successful introduction to the market of new combinations of materials and forces. Five types of innovation were identified, including new products, new methods of production, new sources of supply, the exploitation of new markets, and new ways to organize business. Following Schumpeter, Porter put more emphasis on the first two types of innovation, calling them “technology innovation.” Technology innovation is regarded as an effort “to create competitive advantage by perceiving or discovering new and better ways of competing in an industry and bringing them to market” (Porter 1990, 45).

According to the OECD innovation strategy (OECD 2010), innovation is defined as
the implementation of a new or significantly improved product (goods or service) or process, a new marketing method, or a new organizational method in business practices, workplace organization, or external relations. All innovation must contain a degree of novelty. Innovation rarely occurs in isolation. It is a highly interactive process of collaboration that is increasingly international, involving a growing and diverse network of stakeholders, institutions, and users. Innovation involves the successful implementation of a new product, service, or process. Distinguishing between product and process innovations is a problem because one often leads to the other. One firm’s new product is another firm’s new process, and vice versa. New processes can allow new products to be developed, while the mass production of successful new products often requires process innovation. Also, new products, when they are consumed either as factor inputs or intermediate goods, can contribute to changes in the way other products are produced.

We may make a distinction between narrow and broad dimensions of a system of innovation. The narrow definition includes organizations and institutions involved in searching and exploring, such as Research and Development (R&D) departments, technological institutes, and universities. The broad definition includes all parts and aspects of the economic structure and institutional setup affecting learning as well as searching and exploring. To this end, innovation capability can be assessed using the input and output measures of the innovation system. The input measures are usually represented by indicators such as the amount of R&D investment and the number of researchers in R&D, whereas output measures are reflected by indicators such as patents, high-tech/service exports, and academic output like publications in scientific and technical journals. Despite skepticism, the number of patents granted has been accepted as the most appropriate output measure. Other output indicators, such as those tracking R&D inputs, paper citations, and new product announcements, can be adopted.

Evolution of innovation systems approaches

The way innovation yields “new” developments takes two forms. The first form is most manifest in nascent, high-tech (sub)sectors and knowledge-intensive services. It is often associated with a more linear and technologically oriented view of innovation. In the second form, innovation as new to the region refers to the process in which “local” firms absorb new knowledge and practices from elsewhere, to tailor their products and services better to a wider market area. Such a broader, more entrepreneurial definition of innovation tends toward a system perspective, emphasizing the importance of the mutual interaction of many actors. Over the past decades, the “innovation systems” approach and its three perspectives, namely national innovation systems (NIS), sectoral innovation systems (SIS), and regional innovation systems (RIS), have been developed and adopted by many researchers as conceptual frameworks to analyze innovation in different contexts. Systems are defined as a set of interrelated components which share a common boundary and work toward a common purpose. The systemic approach to innovation highlights the fact that innovation is an interactive process that requires intensive cooperation between firms and agents (such as universities and public research facilities, technology centers, educational establishments, financing institutions, industry associations, and government agencies and bodies).
National innovation systems (NIS)
The national innovation system (NIS) was the first application of the concept to gain widespread recognition. Christopher Freeman and Bengt-Ake Lundvall trace the roots of the approach back to Friedrich List’s concept of national systems of production which took into account the importance of a wide range of national institutions (Freeman 1987; Lundvall 1992). The first contemporary use of the term is found in Freeman’s study of technological change in the Japanese economy. Freeman defined national systems of innovation as “the network of institutions in the public and private sectors whose interactions initiate, import, modify and diffuse new technologies” (Freeman 1987). He underlined the roles of four key components in driving the innovative performance of Japanese firms: government policy, corporate R&D, the education and training system, and the industrial structure of the Japanese economy. Lundvall used the concept to analyze interactions between firms and knowledge institutions in the innovation process. Refining the concept, Lundvall pointed to the interactive nature of the learning process that exists between producers and users of new technology and defined it as being “constituted by elements and relationships which interact in the production, diffusion and use of new, and economically useful, knowledge and that a national system encompasses either located within or rooted inside the borders of a nation state” (Lundvall 1992, 2). A third application of the national innovation system concept is provided in the collaborative study of innovation across 15 nations edited by Richard Nelson (1993). Nelson argued that the use of the concept of “system” to analyze the innovative performance of firms and nations drew attention to the role of a “set of institutional actors that, together, play the major role in influencing innovative performance” (Nelson 1993, 349).

Technological and sectoral innovation systems (SIS)
In a series of case studies in different sectors of the Swedish economy – factory automation, electronics, materials technology, and pharmaceuticals – researchers led by Bo Carlsson identified a distinctive technological dimension that grounded the firms involved in their national or regional contexts. A concept of a technological system was developed and defined as a network of agents interacting in a specific technology area under a particular institutional infrastructure to generate, diffuse, and utilize technology. Technological systems are defined in terms of knowledge or competence flows rather than flows of ordinary goods and services (Carlsson and Stankiewicz 1991). A closely related approach is the idea of a sectoral system of innovation developed by Franco Malerba and others. A sectoral system of innovation is defined as “that system (group) of firms active in developing and making a sector’s product and in generating and utilizing a sector’s technologies; such a system of firms is related in two different ways: through processes of interaction and cooperation in artifact-technology development and through processes of competition and selection in innovative and market actions” (Breschi and Malerba 1997, 131). The primary focus of the SIS approach is on the competitive relationships among firms, with specific reference to the selection environment for firms and their products. The SIS approach is open about the spatial boundaries of the innovation system; the actual boundaries of the system emerge from the specific conditions of the sector under investigation.
Regional innovation systems (RIS)

An extension of the NIS concept, the regional innovation systems (RIS) perspective has been recognized as the most appropriate scale for understanding the dynamics and organizing policy interventions for regional catch-up in the global economy (Cooke, Uranga, and Etxenbarria 1997). This approach identified institutions at the regional level that contribute to industrial innovation as the “regional innovation system,” partly in response to the spread of globalization and increased attention to the importance of geographical proximity for innovation. The governance of regional innovation systems involves a mix of public and private organizations, and their norms, rules, routines, and conventions. These organizations can include research institutes, industry associations and chambers of commerce, financial services providers, technology transfer agencies, public sector development agencies, and regional government, especially in more decentralized systems (Asheim and Gertler 2005).

The systemic dimension of RIS results from the coupling of three subsystems (Cooke, Uranga, and Etxenbarria 1997) leading to the synergy effects of enhanced regional innovation capacities. The first subsystem of finance refers to the availability of regional budgets and capacities to control and manage regional infrastructures. The cultural setting of regions constitutes the second subsystem and defines the milieu within which knowledge networks are embedded. Interactive learning is identified as the third subsystem and represents the core element of RIS as new knowledge is created and exploited. By defining more or less favorable conditions for these subsystems, the RIS approach became popular in the formulation of regional innovation policies. Two key dimensions of innovative activities have been identified in the RIS approach, namely the governance structure and the business structure.

The governance dimension focused on the modes of regional technology transfer as well as modes of governance: grassroots, network, and dirigiste. The business innovative dimension was characterized by the prevailing set of relations among firms in the global economy. This dimension concerns the role of leading firms in the regional economy, the tendency of firms to obtain R&D from in-house activities as opposed to public research institutions, and the innovative milieu within which firms operate. There is no best-practice or one-size-fits-all model of RIS. Instead, tailor-made policy measures adapted to specific regional arrangements are required (Todtling and Trippl 2005).

Despite the proliferation of the RIS approach in the literature, Asheim, Lawton, and Oughton (2011) argued that much of the empirical work dealt with well-functioning, successful regional economies and innovation in high-tech sectors, and should be supplemented with further theoretical and empirical analysis of “less successful” systems and of innovation in more traditional industries. Moreover, most studies are conducted from “a static perspective, while questions in relation to where RIS initially emerge, and why and how RIS and the advantages associated to them change over time are largely ignored” (Tel Wal and Boschma 2011, 923). The extent of knowledge transfer does not depend exclusively on geographic proximity but also on firms’ capabilities, absorptive capacity, and ability to renew their capabilities over time.

Global innovation networks

Although the system approach is very helpful in understanding the internal structure and the evolution of innovation in a specific country/region, it has been criticized for lacking consideration of relationships with key actors outside of the system and for a poor understanding of the
dynamics of the innovation system (Humphrey and Schmitz 2002). Despite a rich stream of research covering OECD and developing countries, the innovation systems approach has been criticized for its weakness in understanding the role of external linkages, its depiction of only a part of the story, and its weakness in dealing with the dynamics of innovation systems. As Hotz-Hart (2000, 444) put it, “systems of innovation are increasingly complex and intertwined, with regional, national, and international levels of integration of innovating activities.” The study of spaces of innovation needs to be more oriented toward exploring the linkages and interrelationships between and across these various spatial levels of scales from the “regional/local” through to the “global” (Bunnell and Coe 2001, 577). Fromhold-Eisebith (2007) called for a “master of scales” to integrate global, national, and regional innovation processes.

To address these weaknesses, researchers have proposed to integrate trans-local linkages into territorially bounded innovation systems and have analyzed catch-up in innovation capability, taking into account changes in technological regimes and the global economic environment. The global value chain (GVC) (Gereffi, Humphrey, and Sturgeon 2005) and global production network (GPN) (Coe et al. 2004) approaches have emphasized how firms have been instrumental in transferring technology, particularly tacit knowledge, to local companies. However, studies have indicated that interfirm networks have a limited impact on innovation capability as measured either by the number of patents, the product novelty, or impact. The relationship between GVCs and innovation systems is nonlinear. The 2000s has witnessed the emergence of global innovation networks (GINs). The traditional process of innovation in centralized corporate R&D departments has given way to GINs (complex and multilayered “networks of networks” that involve both global corporations and “local” companies), which integrate geographically dispersed engineering and product development activities.

Changing spaces and scales of innovation

Learning regions

The notion of “learning regions” is one of a family of concepts known as territorial innovation models (TIMs) that evolved from the late 1970s onwards. Although there are several definitions and perspectives, most scholars consider learning regions as a regional innovation strategy in which a broad set of innovation-related regional actors are strongly but flexibly connected with each other. The founding fathers of the concept of learning regions as distinctive TIMs in their own right were Storper (1993), Florida (1995), Asheim (1996), and Morgan (1997). Richard Florida (1995) is credited with the term “learning region.” He argued that in these new circumstances, networks would be the dominant organizational mode of production, and that the world economy would see the rise of global networks of companies but these companies would be dependent on the key knowledge assets of their home regions. These home regions (especially as the home base of headquarters and R&D functions of transnational corporations (TNCs)) would provide the underlying environment and infrastructure for the generation of ideas and learning and knowledge transfer. Also in the United States, Michael Storper (1993) sought to explain the emergence of new “worlds of production” or new forms of production organization emerging with the growth of a knowledge-based economy. Storper considered three empirical cases: design-intensive or craft-based industries in northeast and central Italy; the Paris fashion clothing industry and
the Île-de-France high-technology aerospace, electronics, and nuclear energy industries; and California high-tech districts concentrating on electronics and aerospace. He found that within the industrial sectors in these particular places technological learning could be seen as a social process that was heavily influenced by the conventions, culture, and informal institutions of local networks, which he called “untraded interdependencies.” Asheim’s (1996) study in Norway focused on the nature of successful learning in particular places. Like Florida (1995) and Storper (1993), he argued that in a knowledge-based economy networks are the key organizational form facilitating technological learning, and that agglomeration provides advantages with regard both to the formation of local networks and consequential learning.

In addition to informal arrangements, Asheim (1996) argued that learning regions required a formal institutional setup. These formal institutions for learning and innovation were what distinguished a learning region from an industrial district. In a learning region geographically concentrated learning networks, territorially embedded conventions, and formal institutions, including a regional innovation policy and support organizations, work together as part of a coherent system. However, the introduction of a requirement for formal institutions to stimulate learning made Asheim’s (1996) view much more akin to the concept of a regional innovation system than a learning region. This more policy-oriented approach to learning regions was developed in Wales by Morgan (1997). In Wales, one of the United Kingdom’s less-favored regions (LFRs), Morgan (1997) confronted the failure of the United Kingdom and Europe to close the gap between rich and poor regions on the basis of traditional regional economic policy. Morgan (1997) linked innovation to economic development and regarded it as an instrument to further social welfare. This gave the learning region a policy agenda in addition to its theoretical one. He argued that innovation was an interactive process influenced by a variety of institutions and conventions, which constitutes its “social capital.”

Recently, critical voices have started to challenge the notion of learning regions (Hassink 2007; Cooke 2002). In particular, its “fuzziness,” normative character, strong overlap with other concepts, and squeezed position between national innovation systems and global production networks have been criticized (Hassink 2007). In a study of innovative firms in the south-east of England, Simmie (1999) found that they learned as much from national and international networks as they did from their local regional networks. Cooke (2002) argues that the learning region is likely to last as a policy framework because of asymmetries rather than equality in local collaborations using different forms of economically valuable knowledge. Hudson (1999) argues that the social and structural constraints imposed by the capitalist system of production set the limits of what is possible. He argues that the prospect of generating convergence between different regions on the basis of a policy focus on learning is unlikely to succeed in the face of the forces driving uneven economic development.

Open innovation and proximity

Since 2000, literature on innovation has turned to consider the boundaries of the regional innovation system, and, in particular, has examined the issue of linkages to organizations outside the region. For instance, Bathelt, Malmberg, and Maskell (2004) argue that local and global linkages can coexist in a context where local interactions create a “buzz” that adds to local innovation processes, while channels of communications or “pipelines” to actors outside the region
are developed in order to access knowledge that is external to the region. New regional advantage in a global economy is generated by combining local buzz with global pipelines for knowledge circulation, within as well as beyond local territories (Bathelt, Malmberg, and Maskell 2004). Pipelines could be constructed either by bringing embedded knowledge in the form of mobile professionals into cluster communities or by sending local entrepreneurs outside to learn in global knowledge networks. Learning from foreign direct investments, transnational entrepreneurs, and global buyers in value chains is in line with the former type of pipeline formation. The concept of open innovation defined as “the use of purposive inflows and outflows of knowledge to accelerate internal innovation and expand the markets for external use of innovation” (Chesbrough 2003) and its impacts on regional development have aroused wide attention among researchers and practitioners.

Cooke (2002) argues that open innovation is one of the challenges that underline the importance for global competitiveness of regional innovation systems and clusters within them. Open innovation in networks of universities, research institutes, industry, government, and financiers leads to the transformation of dynamic knowledge networking capabilities into a regional capability. So, regional knowledge capabilities are rooted in open innovation. Cooke uses biotechnology as an example, but finds evidence that “open innovation” from knowledge-capable regional clusters can be seen in other sectors as well: the automotive sector, information and communication technology (ICT), and electronics and media. Open innovation is still a relatively new model and there are many areas for future research. The focus of open innovation research has been on global networks, but the potential impact on regional development is enormous. In parallel with the emergence of “open innovation,” the concept of a global innovation system/network is proposed, indicating the inadequacy of traditional notions of the innovation system, which focused solely on local or national processes and tended to ignore the international dimension of innovation systems.

Along with the popularity of open innovation, there has been increasing interest in the notion of proximity in the context of economic growth in general and innovation in particular (Boschma 2005). When analyzing the logic and dynamics of innovation, four functions of proximity have been identified. First, being close to each other helps companies to develop an efficient division of labor and coordinate their actions with a core of specialized suppliers and partners. Second, there are externalities of proximity available to all within a region. Third, there is evidence that when companies in the same industry are located close to each other it forces them to innovate. Fourth, and more importantly, proximity is associated with knowledge spillovers and mutual learning. Boschma (2005) has argued, however, that geographical proximity per se is neither a necessary nor a sufficient condition for learning. Therefore, when discussing the roles of proximity and distance in innovation policy, there is a critical dilemma. On the one hand, as most modern innovation policies and technology centers and science parks recognize, there is a need for a mechanism for enhancing physical, social, and cognitive proximity between the relevant actors in the innovation process. On the other hand, innovation policy must include mechanisms for enhancing social and cognitive diversity, the openness of innovation networks, and the ability of an innovation network to connect itself to the wider national and global knowledge base: that is, mechanisms for ensuring sufficient distance between the actors. Obviously, the central challenge for regional innovation policies is not to maximize proximity, but to create an efficient
Table 1  Dimensions of proximity and distance in innovation networks (Cooke et al. 2011, 559).

<table>
<thead>
<tr>
<th>Distance</th>
<th>Source</th>
<th>Innovation potential</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographic</td>
<td>Physical proximity/distance between actors.</td>
<td>Does not automatically lead to innovations but may facilitate social proximity.</td>
</tr>
<tr>
<td>Communicative</td>
<td>Closeness/differences in concepts and professional languages.</td>
<td>When making a new idea understandable, concepts from other fields or sciences may be utilized.</td>
</tr>
<tr>
<td>Organizational</td>
<td>Arrangements to coordinate transactions and enable exchange of information within and between organizations.</td>
<td>Contributes to efficient organization of utilization of knowledge and other resources. An organization should have both strong and weak links in its networks.</td>
</tr>
<tr>
<td>Functional</td>
<td>Differences or closeness in expertise in different industries/clusters.</td>
<td>Functional distance brings novel information from outside of one’s own field of operations. Information often needs to be adapted.</td>
</tr>
<tr>
<td>Cultural</td>
<td>Differences or closeness of cultural habits and values: “rules of the game.”</td>
<td>The challenge is to get people working in different organizational cultures to collaborate.</td>
</tr>
<tr>
<td>Social</td>
<td>Intensity of trust-based social relations, e.g., friendship or kinship.</td>
<td>Innovation requires interaction among different kind of actors. Trust helps in the creation of radical ideas.</td>
</tr>
</tbody>
</table>

balance between the contradictory purposes of enhancing proximity and distance.

Seven central dimensions of distance and proximity, as well as their relevance to innovation, are summarized in Table 1. The extent to which manufacturing companies have moved production activities from advanced economies to offshore locations has raised concerns that R&D and manufacturing know-how will follow and be “lost.” Even for industries in which innovation and production are tightly linked, companies do not automatically relocate R&D to their offshore production sites. For instance, makers of mobile phones and other consumer electronics products need to engage with dozens of parts suppliers that cluster around the Asian production sites, but Apple continues to come up with its innovative iPhone designs in California.

Innovation and regional development: divergent trajectories

Innovation and regional development in developed countries

Innovation is not uniformly distributed across geographical landscapes. Innovation clusters have emerged as places with dense webs of interconnected technology companies, customers, and suppliers. Improving a cluster’s chance of flourishing are factors such as liberal immigration laws and venture capital financing.
Since the end of the 1970s, industrial districts have been studied in Italy (Dunford 1996), more widely in Europe, and in North America (Saxenian 1994), not only because of their capacity to compete in international markets, but also because they have turned out to be significant wellsprings of product and process innovations. This issue has been studied by numerous scholars and schools through various perspectives, among which are the innovative milieu and “collective learning” approach, the regional innovation systems approach, the Italian school of industrial districts, the California school of economic geography, and the regional economists’ tradition of localized knowledge spillovers. It is argued that all these different streams overlap with respect to the mechanisms associated with the formation of innovation networks, including market relationships, social ties, and policy-driven relations. First, networks have been considered a significant channel for the diffusion of knowledge and for the generation of inter-organizational innovations in regions and subregional areas. The diffusion of knowledge through local market linkages is facilitated by the geographical proximity of firms and other organizations, which eases the transfer of tacit knowledge and promotes innovation accordingly. Second, what really seems to make a difference in regions and subregional areas is networks originating from nonmarket relationships and stemming from the regional, cultural, and social proximity of entrepreneurs and workers. Storper (1993) considers “social ties” a dimension of “untraded interdependencies,” while Saxenian (1994) considers them a driver of the way engineers interact with each other in the high-tech cluster of Silicon Valley. Third, besides spontaneous processes of collective innovation related to the presence of market and/or social networks, scholars have emphasized the importance of public institutions and policies in fostering the formation of innovation networks at the regional and subregional levels. All these studies have helped generate an incredibly rich and descriptive account of regional and subregional innovation networks – although rarely named as such.

Silicon Valley has been widely regarded as a successful model of innovation-driven regional development. Saxenian (1994), in her comparative study of Silicon Valley and the Boston area, analyzed the role of regional inter-organizational networks in making regions innovative, emphasizing the role of regional culture, identity, and human capital. She argues that the Silicon Valley pioneers created a decentralized system characterized by open information flows and learning between firms and between firms and regional universities (for example, Stanford University) with the ability to respond rapidly to changing markets by very quickly putting together people with different backgrounds and starting new companies. The cultural dimension also shows up in a willingness to experiment and take risks in the relatively flat organizational hierarchies, in the informality of work styles, and in occupational mobility. Although Saxenian did not use the term “open innovation,” she describes how Silicon Valley firms were successful by competing intensely while collaborating in informal and formal ways with one another and with regional universities to learn about fast-changing markets and technologies. As Saxenian (1994, 156) said, “the openness of this partnership ensured that the design and manufacturing innovations that it produced would diffuse rapidly throughout the region and the industry.” Interfirm supplier networks reinforced the regional advantage of Silicon Valley as an attractive location for innovative companies. Another benefit of open innovation derives from the maintenance of open boundaries to benefit from global connections and cultural diversity. Silicon Valley’s immigrant entrepreneurs and engineers also contribute to
maintaining open boundaries. They not only bring cultural diversity, but also connect Silicon Valley to Taiwan and India through informal social and professional ties and formal corporate alliances and partnerships, thereby facilitating flows of capital, skills, and technology.

Innovation and regional development in emerging economies: insights from China

Cities all over the world have tried to replicate the success of Silicon Valley. Of them, Beijing has emerged as a serious contender. Beijing, China’s political and cultural capital, has seen dramatic growth in recent years. In 2011, Chinese venture capital firms invested US$13 billion, half as much as their US counterparts, 30% of this in Beijing (Figure 1). Beijing hosts rare concentrations of wealth and some 70 institutions of higher learning, including China’s best computer science departments. Beijing produces what few other places can – giant, fast-growing tech companies, such as Baidu (worth $31 billion in 2014), Lenovo, and smartphone maker Xiaomi, which sold handsets worth $2 billion in 2013. China’s government has set in motion ambitious plans to break into high-tech and electronics manufacturing. In the 1980s it opened the Zhongguancun technology hub in Beijing, taking Silicon Valley as its inspiration (Zhou 2008).

Prior to the 1980s, China had an NIS similar to other centrally planned economies such as the Soviet Union and India, characterized by the complete separation of research activities, education, and manufacturing activities, with the first
two in public research institutes (PRIs) and universities, and the last in state-owned enterprises (SOEs), respectively. Since the start of reform and opening up in 1978, the Chinese government has started to actively transform this Soviet model, especially by emphasizing the co-location of research and manufacturing. The government has used both carrots and stick to pull and push R&D institutes to adapt to a market environment and to conduct R&D that has direct or indirect industrial applications. For instance, while it reduced institutional funding for PRIs in order to push them to conduct more market-oriented R&D through industries and universities, it also offered financial incentives to commercialize R&D results, especially through the 1988 Torch Program, a national science and technology program specifically targeting high-tech industrialization. The government also promoted organizational transformation by advocating the merger of some R&D institutes with enterprises in 1987 and reforming established R&D institutes into entities with economic functions, such as production and consultancy centers, from the 1990s.

Continuous reform has transformed the pre-reform government-led model of innovation. Firms have been reorganized and acquired R&D and marketing functions, and their innovation capability has increased notably. An enterprise-led and market-based technological innovation system has not yet been well established because SOEs, rather than private enterprises, dominate the national R&D landscape in China. The technology landscape in China is certainly highly uneven, with considerable differences in endowments, cultures, and institutions, varying positions in state strategies, and a variety of linkages with the global economy, resulting in distinctive patterns of innovative regions and cities.

In Zhongguancun in Beijing, China’s Silicon Valley and the most innovative region, the influence of actors in the regional systems of innovation (that is, the central and local state, TNCs, local firms, and research institutions), wax and wane, and each is challenged by others and by the changing political and institutional environment. Shanghai’s regional innovation systems have benefited from global linkages and recent TNC R&D centers. In Shenzhen, the contribution of universities and research institutes to technological innovation is relatively minimal compared to Beijing, which is a national scientific hub with a number of prestigious universities and institutes. Accessible venture capital from Hong Kong, the completeness of the ICT industrial value chain, and active private entrepreneurship have been cited as the basis of the prosperity of domestic high-tech firms in Shenzhen.

The East Asian model based on the postwar experiences of Japan and the four “tigers,” that is, South Korea, Taiwan, Hong Kong, and Singapore, has found limited applicability in post-reform China. In its simplest form, in the East Asian model, export-driven industries are the chief and superior vehicle for industrial competitiveness and technological upgrading. Over the last two decades, however, empirical experience in China has challenged this model. There is a growing recognition that internal dynamics such as the roles of the state, domestic markets, and the resources and strategies of indigenous companies are indispensable determinants of the success of technological upgrading.

The major findings of studies of innovation and regional development in China can be summarized as follows. First, FDI and export-oriented industry have an indirect effect on innovation development in China. Firms or regions with strong internal R&D capacity can absorb external knowledge and benefit from FDI-related technological innovation.
Second, China’s hierarchical science and technology system has a heavier imprint on the map of knowledge creation than the location of industry or foreign investment. However, China’s public institutions play only a limited role in creating and disseminating knowledge to industry, although a knowledge transfer network is gradually emerging. Third, China’s heavily export-oriented regions have been pursuing policies or strategies to move up the value chain since the 1990s, both at the enterprise and local state level. So far only a few successful indigenous companies have established dominant roles in regional innovative networks (e.g., Huawei, ZTE, BYD, and Tencent). While the innovativeness of these companies is not widely shared by other enterprises in the region, the trajectory toward more technology-driven and international collaboration is definitely on the way.

A comparison of the importance of indigenous and foreign innovation efforts in China indicates that the major drivers of the technological frontier are domestic firms in low- and medium-technology industries, and foreign-invested firms in the high-technology sector. The R&D activities of foreign-invested firms have not had significant positive effects on the technical transformation of local firms. Similarly, subcontracting will not necessarily improve the ICT suppliers’ technology innovation in China. Moreover, overdependence on subcontracting for the global market and global contractors will impede technological innovation in China (Sun et al. 2013). Negative correlations between technological investment and regional specialization in export-oriented industries are found for ICT firms in Beijing, Shanghai-Suzhou, and Shenzhen-Dongguan regions. ICT firms in Beijing, the most domestic-oriented region, perform better than those in the other three regions. The empirical experience of Shanghai’s Zhangjiang high-tech park shows that both a developmental state and entrepreneurship play key roles in the governance of innovation in China. In the initial stage, the municipal government is more crucial, but integration into global knowledge flows is more important in the later stage. Interestingly, in Shenzhen, heavy investment in R&D seems to belong exclusively to Chinese indigenous companies, while multinational companies, especially those from Hong Kong or Taiwan, have hardly participated. Despite the massive and intensive production linkages, firms in the Shenzhen ICT industrial cluster show weak innovative performance. Most local firms gain technological knowledge from internal R&D activities instead of external technology. Companies such as Huawei started with import substitution practices, but as they build up their internal R&D capability and invest heavily abroad, they are turning themselves into truly China-based transnational corporations and global lead firms in their respective GPNs and GINs.

In post-reform China, the state’s role in innovation systems has changed dramatically over time and space. Since the mid-2000s, Chinese leaders have realized that core technologies failed to come along with foreign equipment purchases or foreign investment flows. This recognition is reflected in the seminal document entitled “Guidelines for the Implementation of the National Medium- and Long-term Programme for Science and Technology Development (2006–2020)” (MLP hereafter) (China Science and Technology Newsletter 2006). As a grand blueprint for science and technology development designed to bring about the renaissance of the Chinese nation, the MLP defines indigenous innovation as enhancing original innovation through co-innovation and re-innovation based on the assimilation of imported technologies. Indigenous innovation
is planned to turn the Chinese economy into a technology powerhouse by 2020 and a global leader by 2050. Fu and Gong (2011) argued that the term “indigenous” (zizhu), which literally means self-direction, is problematic. As a matter of fact, it does not mean that technology has to originate in China or that the participation of TNCs is excluded. Policies are designed to boost indigenous R&D efforts and encourage domestic enterprises to exercise more strategic control in their technological interactions with foreign parties. The ongoing campaign for technological upgrading in response to the paradigm shift in innovation policies in China has two aspects. One is to foster technological innovation in the export-oriented production networks of TNCs, which have relocated their production activities, but with limited engagements in R&D in China. The other is to develop indigenous innovation through the technological upgrading of domestic firms, fostered by a paradigm shift in innovation strategy envisaging the state-initiated coupling of domestic firms and global leading TNCs (Yang 2014). The indigenous innovation strategy has been implemented since 2006. Prior to the onset of the 2008 global financial crisis, the Chinese state had already pushed Chinese enterprises to alter the nature of their relationships/positions in GPNs, with the explicit goals of technological upgrading and greater independence from TNCs, but full implementation of this new strategy did not start until 2009 in the wake of the global financial crisis.

The financial crisis provided stronger reasons for the Chinese government to adopt an aggressive innovation strategy. R&D funding in China reached a historically high level of 1.7% of gross domestic product (GDP) in 2009, surging from 0.8% in 2001. Efforts have been taken to assess the impact of changing state innovation policies on the transformation of NISs. In practice, indigenous innovation has been recognized as a viable strategy to foster technological upgrading in selected Chinese cities and sectors. In 2010, seven strategic emerging industries (SEIs) were selected by the central government to be technology intensive, and to achieve technological breakthroughs in a short period of time. These sectors are energy-saving and environmental protection, new information technology, biology, advanced equipment manufacturing, new energy, advanced materials, and new energy vehicles (Table 2). These SEIs, regarded as knowledge- and technology-intensive industries with less resource consumption, are expected to foster the technological upgrading of domestic firms. Local government authorities, such as the Guangdong provincial and Shenzhen municipal governments, announced similar SEIs to replace the low value-added, labor-intensive, environment-polluting and high energy consumption industries primarily invested in by TNCs from Hong Kong and Taiwan (Yang 2014). “Transition from production to innovation” through developing state-designated SEIs has been listed on policy agendas at various levels of government in China.

Innovation policies and regional development in the globalizing world: the research agenda

Innovation is nowadays considered a highly interactive and systemic process. In the globalized, post-Fordist knowledge economy, the relationship between innovation and regional development has become a central avenue of inquiry for scholars in a range of disciplines, particularly economic geography. The increasing mobility of production factors from 1990 to 2010 has radically changed the scale and space of innovation. The quality and performance of RISs depends to a high degree on
the knowledge interactions within and beyond their regions. The regional dimension is essential for innovation. This is partly due to its content but also to the social embeddedness of regional economic relations. But regional systems or networks are not self-contained and self-sufficient entities. The relational view has to be extended. Trans-regional innovation is an essential element of innovation potential. Many empirical studies show that RISs are highly open systems and that international or global knowledge links are often vital. Regional systems that have been and are still developing along trajectories of cumulative innovation and knowledge find themselves part of a multiscalar system.

Scholars dissatisfied with existing approaches to the study of innovation networks have in recent years proposed a methodological shift. Key findings of the literature and directions for future research include the following.

1 Understanding innovation networks. Scholars have investigated the structural properties of innovation networks within and across regional clusters and have found that different types of network structure may also lead to different types of reward (and constraint) for its members. Innovation-related knowledge is diffused in a rather structured, selective, and uneven way within industrial clusters and is thus not pervasive or collective, as presumed by conventional theories – leaving a number of firms isolated from the local knowledge network, even when they are located nearby or have established other types of business linkages with nearby firms.

2 Analyzing the relationship between network position and firm characteristics. The analysis of the structural properties of networks reveals that actors are not equally positioned within a network. The position of actors in a network is associated with the innovative performance of firms or with other firm-level characteristics.

3 Comparing different networks. Social network analysis permits the comparison of different types of network formed by the same set of actors. It helps to illuminate what type of network (for example market, friendship, professional) is most likely to

Table 2  Strategic emerging industries designated by the central, Guangdong provincial, and Shenzhen municipal governments (Yang 2014).

<table>
<thead>
<tr>
<th>Seq.</th>
<th>National</th>
<th>Guangdong</th>
<th>Shenzhen</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Energy saving and environment protection</td>
<td>High-end electronics and IT</td>
<td>Biological industry</td>
</tr>
<tr>
<td>2</td>
<td>New generation information technology</td>
<td>New energy vehicle</td>
<td>Internet industry</td>
</tr>
<tr>
<td>3</td>
<td>Biotechnology</td>
<td>LED</td>
<td>New energy industry</td>
</tr>
<tr>
<td>4</td>
<td>High-end equipment manufacturing</td>
<td>Biotechnology</td>
<td>Cultural and creative industry</td>
</tr>
<tr>
<td>5</td>
<td>New energy</td>
<td>High-end equipment manufacturing</td>
<td>New material industry</td>
</tr>
<tr>
<td>6</td>
<td>New materials</td>
<td>Energy conservation and environmental protection</td>
<td>New generation industry</td>
</tr>
<tr>
<td>7</td>
<td>New energy vehicles</td>
<td>New energy and new materials</td>
<td></td>
</tr>
</tbody>
</table>
transmit knowledge relevant to innovative results.

4 Analyzing network evolution over time. A key innovative step has to do with the study of innovation network evolution over time. While previous studies have taken a historical or descriptive perspective, recent research has explored whether network analysis could offer an alternative view. Because of the lack of longitudinal network data at the regional level, scholars have so far primarily used co-inventor networks based on patent data. Very few data exist, however, on the evolution of informal networks. This line of research is certainly one of the most promising.

Policy actors at various levels (local, regional, national, and supranational) are integral elements of innovation systems. Increasingly, we observe a system of multilevel governance. Three main and interdependent policy foci are needed to help enhance and improve a region’s innovation capability: policies aimed at tackling weaknesses and inadequacies in regional “fundamentals”; those aimed at enhancing external economies associated with the region’s existing and potential industries and clusters; and those aimed at improving the adaptive capability of a region’s economic asset base and fundamentals (firms, industries, workers, institutions, infrastructures, governance arrangements, and so on). Different regional authorities may be expected to assign different relative emphases to these groups of policy initiatives, but, in almost all regions, policies will be needed on all three fronts.

The global recession that began in 2008 suggests that the innovations of the second half of the twentieth century may not be a model for the twenty-first century. Technology innovation will not end, but both are likely to be subject to a more “managed” policy environment, with government support.

SEE ALSO: Industrial districts

References

INNOVATION AND REGIONAL DEVELOPMENT


Further reading


Input–output analysis

Dazhong Cheng  
*Fudan University, China*

Peter W. Daniels  
*University of Birmingham, UK*

Input–output analysis is one of the most widely applied analytical frameworks in economics. Its fundamental purpose is to analyze the interdependence of various industries or sectors in an economy, such as agriculture, manufacturing, or services. The original idea for input–output analysis can be traced as far back as the early work of a British physician and Oxford professor, William Petty, in the mid-seventeenth century. The first historic breakthrough was made by a Russian-born US economist, Wassily W. Leontief (1906–1999), in the late 1930s, for which he received the Nobel Prize in Economic Science in 1973 (Miller and Blair 2009).

Two pillars of input–output analysis

Over the past eight decades, the evolution of input–output analysis has been based upon two major pillars: theoretical foundation and data construction. Therefore, some studies may be purely theoretical, exploring the formal relationships that can be derived under various assumptions from sets of equations; other studies may be purely descriptive and dependent upon the construction of input–output tables; others will incorporate a mixture of both empirical data and theoretical relationships in an attempt to explain or predict actual interdependence between industries or sectors (Christ 1955).

The theoretical foundation of input–output analysis, or the input–output model, is viewed by Leontief as a general equilibrium theory. In its most basic form, an input–output model consists of a system of \( n \) linear equations with \( n \) unknowns which form a matrix that can be readily used to represent the mathematical structure of the model. An essential condition for empirical input–output analysis is the input–output table, which was first used by the French economist François Quesnay in *Tableau Économique* (Kuczynski and Meek 1972). Input–output analysis became a widely used tool of economic analysis after a standardized system of economic accounts built around input–output concepts was developed, initially by Leontief (1941) and then by Richard Stone (1961), who was awarded the Nobel Prize in Economic Science in 1984 (Miller and Blair 2009).

The basic structure of an input–output table records the flows of goods or services from each sector (as a producer or seller) to each of the other sectors or itself (as consumers or purchasers) (Table 1). The distinction commonly made in input–output analysis is between the production of goods and services and their utilization, as reflected in the four quadrants of the input–output table. Quadrant I shows the intermediate transactions; that is, the flows of goods and services that are both produced and consumed in the process of current production. Quadrant II shows the sales by the producing industries to final uses such as private consumption, government consumption, fixed capital formation, and net exports. Quadrants I and

---

*The International Encyclopedia of Geography.*  
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.  
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.  
DOI: 10.1002/9781118786352.wbieg0071
**Table 1** Basic structure of a national input–output table.

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
<th>Intermediate use (producers as consumers)</th>
<th>Final use (final demand)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sector 1</td>
<td>Sector 2</td>
</tr>
<tr>
<td>Producers (intermediate inputs)</td>
<td>Sector 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sector 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sector n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary inputs (value added)</td>
<td>Employees</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Business owners and capital</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Government</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Based on Miller and Blair (2009, 3). Reproduced from Cambridge University Press.
II together allocate the total output of each industry in the economy. Quadrant III shows the primary (factor) inputs that are required by each producing industry; these constitute the gross value added (including employee compensation, depreciation of capital, and indirect business taxes) in each industry. Quadrants I and III combined show the total inputs used for production by each industry in the economy. Quadrant IV records two equivalent measures of gross domestic product (GDP), one derived from expenditure components, and the other from value-added components.

Recent major developments in input–output analysis

The widespread availability of high-speed modern computers has made input–output analysis more comprehensive and practical. Among the most important advances are the improved implementation of computable general equilibrium (CGE) models and the ability to continually update and expand input–output tables. CGE models use the input–output model as a benchmark for working with actual economic data to estimate how an economy might react to changes in policy, advances in technology, or other external factors. So far, there are two versions of mathematical programming and optimization software suitable for CGE modeling and analysis: the general algebraic modeling system (GAMS) (see www.gams.com/) and the general equilibrium modeling package (GEMPACK) (see www.copsmodels.com/gempack.htm).

There have been two important developments in the use of input–output tables in recent decades. First, since 1936, when Leontief first presented US interindustry transactions tables for 1919 and 1929, an increasing number of countries and regions around the world have been routinely constructing input–output tables. Two crucial advances can be identified in the evolution of these national-level input–output tables. One is the compilation of regional and multiregional input–output (or MRIO) tables, which provide the basis for impact analysis and examining the interconnections between different regions within a country. The other is the construction of noncompetitive import type input–output (or noncompetitive IO) tables. In these tables, domestically produced intermediate inputs and imported intermediate inputs are treated separately, as they are presumed to be poor substitutes for each other. If the inputs imported by a country are differentiated further by country of origin, it is possible to trace its external economic and trade relationship with other countries (of which more below).

Second, as more and more countries have become engaged with the process of globalization, industrial linkages now reach well beyond national borders and it has become more difficult for the conventional national-level input–output tables to capture the details of global industrial and value chains. Therefore, various attempts to harmonize input–output tables for different countries and to construct international input–output tables are ongoing. Examples include the Asian IDE-JETRO tables compiled by the Institute of Developing Economies and Japan External Trade Organization (Inomata and Okamoto 2015), the GTAP tables constructed by Purdue University (see https://www.gtap.agecon.purdue.edu/), and the recently released World Input–Output Database (WIOD), which covers 27 countries of the European Union and 13 other major countries in the world for the period from 1995 to 2011. The latter attempts to analyze the effects of globalization on trade patterns, environmental pressures, and socioeconomic development (see www.wiod.org/new_site/data.htm) (Timmer 2012). In contrast to a national input–output table, a world
INPUT-OUTPUT ANALYSIS

input–output table traces flows of products or services for both intermediate and final use that are divided into those that are produced domestically and those that are imported (see Table 2).

An application of input–output analysis to producer services

In contrast to consumer services, producer services enter the production process of other manufacturing and services firms as an input (Grubel and Walker 1989; Stibora and de Vaal 1995). Firms can obtain producer services via two channels: in-house provision and/or market transactions. The former reflects the specialized division of labor inside the producers or firms, and thus the internal resource allocation and industrial linkages directed by firms’ decisions. Unfortunately, these types of producer services cannot be captured by input–output tables in the System of National Accounts (SNA) and will frequently be grouped according to firms’ main business (e.g., manufacturing). The latter, market-driven group reflects the specialization of, and division of labor between, different firms in the market, and hence the resource allocation and industrial linkages based on market competition. These types of producer services are captured in the SNA input–output tables, with the number of such records increasing as the outsourcing and marketization of producer services, a natural evolution of specialized division of labor and resource allocation from inside the firm to the market, continues.

However, from a statistical perspective it is difficult to distinguish between producer services and consumer services; activities such as banking or transport not only fulfill intermediate demand but also meet the needs of final consumers, even though individual firms may emphasize the provision of services to one group or the other. As a result, research that uses an arbitrary classification of producer services cannot accurately detail their status, their role, and their contribution to a national economy. In order to circumvent such problems, some researchers have turned to input–output analysis (Khayum 1995; Windrum and Tomlinson 1999; Cheng and Daniels 2014).

By way of illustration, two examples are presented here. First, the output of any one specific service sector must be classified into two parts: output consumed by final users and output that will be used by firms as an input. The ratio of the former part in the output is the consumer services ratio, while the latter part (\( SI/SO = \text{services input/services output} \)) is the producer services ratio. This identifies the extent to which a specific service-sector category is a producer service. A large producer services ratio means that more of the output of a specific service sector is being used as input to other parts of the economy; this suggests that the sector is more like a producer service. If the ratio is low for a specific service sector, it is considered to be functioning more like a consumer service. As a general rule, a ratio of 50% or above is considered the threshold for judging whether a service activity can be characterized as a producer service. Data for the USA (Table 3) show that the sectors with a producer services ratio averaging 50% or more are research and development (R&D) (about 90%), other business activities (about 80%, including knowledge-intensive business services (KIBS) such as consulting, design, legal services, etc.), transport and storage (70%), post and telecommunications (over 60%), and finance and insurance (above 55%). The remaining sectors with values below 50% are more closely aligned with consumer services. For China in the mid-2000s (see Table 3) six activities have more than 60% of their output used as producer services: wholesale and retail trade and repairs, hotels and restaurants,
Table 2  Basic structure of world input–output table with three economies.

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
<th>Intermediate use</th>
<th>Final use/consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Country A</td>
<td>Country B</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Country A</td>
<td>Sectors: 1 ... n</td>
<td></td>
</tr>
<tr>
<td>inputs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Country B</td>
<td>Sectors: 1 ... n</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RoW</td>
<td>Sectors: 1 ... n</td>
<td></td>
</tr>
<tr>
<td>Primary inputs</td>
<td>Value-added in A</td>
<td>Value-added in B</td>
<td>Value-added in RoW</td>
</tr>
</tbody>
</table>

Source: Based on Timmer (2012).
transport and storage, post and telecommunications, finance and insurance, and other business activities. The comparison suggests that although R&D behaves more as a producer service in the United States, in China it is more akin to a consumer service, along with real estate activities, education, and health and social work. On the other hand, wholesale and retail trade and repairs and hotels and restaurants in China act more as producer than consumer services.

Input–output analysis is also used to compute two linkage coefficients: the backward linkage coefficient ($BL_j$)

$$BL_j = \left( \frac{1}{n} \sum_{i=1}^{n} b_{ij} \right) / \left[ \frac{1}{n} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \right) \right]$$ (1)

and the forward linkage coefficient ($FL_i$)

$$FL_i = \left( \frac{1}{n} \sum_{j=1}^{n} b_{ij} \right) / \left[ \frac{1}{n} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \right) \right]$$ (2)

(from the Leontief inversion matrix $B = (b_{ij})_{n \times n} = (I - A)^{-1}$, the Leontief complete consumption coefficients $b_{ij}$ can be obtained, where $A$ is an input coefficient matrix based on Quadrant I in Table 1 and $I$ is an identity matrix). The first coefficient measures the backward economic linkage of one specific sector $j$ to the rest of the economy; that is, when the output of sector $j$ (e.g. transport services) increases by 1 unit, how much of the increased demand from sector $j$ (as a purchaser) will rely on (upstream) sectors (e.g. transport vehicles) whose outputs are used as inputs to production in sector $j$? The greater the value of this coefficient, the stronger the pulling power of sector $j$ on the rest of the economy. The second coefficient measures the forward economic linkage of one specific sector $i$ to other sectors; that is, when the output in every sector of the economy increases by 1 unit, how much of the increased demand from these (downstream) sectors will depend on sector $i$ (as a seller) (e.g. transport services)? The greater this coefficient, the higher the pressure of demand experienced by sector $i$.

Indeed, there has been considerable literature arguing about the linkage measures and offering numerous suggestions for differing definitions and refinements of them (Beyers 1976; Miller and Blair 2009). Our aim here is just to introduce the simplest forms of these measures and in particular to show how they are derived from information in the IO table.

It is common for the same kind of services (e.g. transport services) to be used as both upstream and downstream services. Referring again to Table 3, in the mid-2000s, R&D, other business activities, education, and health and social work have larger backward linkage coefficients in China than in the USA. However, the backward linkage coefficients of real estate activities and finance and insurance are obviously much smaller (only 0.57 and 0.74, respectively) in China. The values of the forward linkage coefficients for wholesale and retail trade and repairs, transport and storage, and other business activities in China are greater than 1, while the values for the other sectors are much lower; five scored less than 0.5 in the mid-2000s. In the USA, the forward linkage coefficient for R&D was well above 1.7 from the early 2000s, approximately four times the equivalent value for China. Indeed, almost all the service sectors in China, except hotels and restaurants and transport and storage, have much smaller forward linkage coefficients.

The fact that every sector of the economy in China depends to a significant degree on hotels and restaurants and on transport and storage again confirms that these services are utilized more as producer services and have therefore performed a key role in industrialization. Once again, the forward linkage coefficient for real estate activities
Table 3  Application of input–output analysis: China and the USA.

<table>
<thead>
<tr>
<th>Service sector</th>
<th>Producer services ratios (SI/JO)</th>
<th>Backward linkage coefficients (BL)</th>
<th>Forward linkage coefficients (PL)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Early 2000s China</td>
<td>USA</td>
<td>Mid-2000s China</td>
</tr>
<tr>
<td>Wholesale and retail trade and repairs</td>
<td>66.42</td>
<td>30.21</td>
<td>61.96</td>
</tr>
<tr>
<td>Hotels and restaurants</td>
<td>62.17</td>
<td>22.11</td>
<td>67.82</td>
</tr>
<tr>
<td>Transport and storage</td>
<td>80.46</td>
<td>70.17</td>
<td>77.96</td>
</tr>
<tr>
<td>Post and telecommunication</td>
<td>81.14</td>
<td>61.88</td>
<td>75.29</td>
</tr>
<tr>
<td>Finance and insurance</td>
<td>76.1</td>
<td>56.63</td>
<td>74.59</td>
</tr>
<tr>
<td>Real estate activities</td>
<td>31.14</td>
<td>29.87</td>
<td>20.13</td>
</tr>
<tr>
<td>Research and development (R&amp;D)</td>
<td>–</td>
<td>89.22</td>
<td>37.29</td>
</tr>
<tr>
<td>Other business activities</td>
<td>54.77</td>
<td>79.9</td>
<td>76.22</td>
</tr>
<tr>
<td>Education</td>
<td>12.55</td>
<td>16.52</td>
<td>9.46</td>
</tr>
<tr>
<td>Health and social work</td>
<td>4.49</td>
<td>2.25</td>
<td>13.18</td>
</tr>
</tbody>
</table>

Source: Calculation based on input–output tables of China and the USA. “—” denotes data vacancy. Both coefficients have in their denominator the average value of a coefficient in the inverse matrix.
INPUT-OUTPUT ANALYSIS

in China is less than 0.5 and is just one-third of the equivalent USA value. It is clear that Chinese real estate activities cannot exert strong pulling or pushing powers on the rest of the economy: this sector functions independently from (or is not strongly connected to) the rest of the economy since both linkage coefficients are well below 1.

To sum up, subject to the availability of suitable and reliable data, an input–output analysis can greatly improve our understanding of the development and contribution of producer services to economic change and restructuring, especially at the national scale but also at the global scale as the appropriate models for analysis begin to be introduced.

SEE ALSO: Deindustrialization; Demand and supply for producer services; Externalization; Factors of production; Global commodity/value chains; Global production networks; Globalization; Industrial linkage; International division of labor; Producer services: definition and classification; Value added

References


Institutions and development

Anthony Bebbington
Clark University, USA
University of Manchester, UK

Institutions, institutions everywhere – and we’re still not satisfied

A city in South America, early 2015, and the author is interviewing a senior international economic journalist about their take on extractive industries and development. The journalist, who is simultaneously of a social democratic and neoliberal inclination, says something along the lines of: “Of course, the social scientists are right when they say it’s all a problem of institutions. But this doesn’t help very much. I can’t help feeling that this is their trade and that they should be able to say rather more than just that.”

In a short sentence, the journalist summarizes the problem of institutionalist thinking about development: that a typical response is “Yes, of course institutions matter, but so what?”

The idea of an institution is difficult to pin down. The term is used to refer to formal organizations at scales ranging from the World Trade Organization through to panchayats; recognized sets of rules that guide practices in such a way that the rules are reproduced, for instance the institution of marriage; and sets of practices that become so routinized that they imply the existence of sets of rules that actors follow, for instance the institution of the gift. In these meanings – albeit very different – the constant is a reference to stability, reproduction, and acceptance, and to rules that govern behavior whether through the exercise of force or the command of legitimacy. But institutions also change, often profoundly: think again of the institution of marriage, for instance, which has changed profoundly in modern history and continues to do so. So institutions are stable, but they are also not stable.

Another binary that institutions seem able to bridge is the formal and the informal. Laws, constitutions, tax codes, property – these can all be highly formal, tangible institutions, the rules underlying which can be pulled off a shelf and perused (at least in modern economies). However, systems of reciprocity, moneylending, or “corruption” can equally be based on rules of which participants are quite aware, but these institutions are in no way formal: there is nothing written down or formally adjudicated, and there are no rule books on the shelf. And then, in some sort of liminal space between the formal and informal are institutions such as systems of indigenous or religious law which may be only partially formalized and only partially recognized by the state, and which are completely binding for those who recognize them and adamantly rejected by others who do not.

This expansiveness of the term “institution” leads some to find relatively little use in the concept, but it also reflects a broad-based recognition that explanations of social phenomena cannot wholly rely either on the purported motivations of individual actors or on the power of overwhelming structures. The concept of institution gets at the idea that there is something
between structures (such as the logic of capitalist development) and human actions (such as those of rational microentrepreneurs or motivated development workers) that influences outcomes. The recurring interest in institutions also reflects the desire to intervene in development processes in pursuit of different outcomes. Whether authors call for tighter financial regulation, state oversight of enterprises, more stringent tax codes, government decentralization, more opportunities for popular participation, ways of making budgets transparent, indigenous autonomy in local government, or the liberalization of capital markets – these are all calls for forms of intervention to change the rules and make these new sets of rules binding. In short, they are calls to change institutions and reflect a belief, stated or not, that a change of institutions might just change development outcomes.

Institutions and development

The importance of institutions in explaining development outcomes and processes became one of the more important themes in development theory and policy from the 1990s onward. A milestone text in this regard was Douglass North’s *Institutions, Institutional Change and Economic Performance* which understood institutions as “the rules of the game or more formally, the humanly devised constraints that shape human interaction” (North 1990, 3) and argued that these rules of the game are the primary determinant of the degree and quality of economic performance. The reasons for the rapid rise of these institutionalist arguments are many. While one reason might, of course, be the general power and persuasiveness of the argument itself, Ben Fine (1999) suggested instead that the idea of institutions became so popular because it was a way of being able to recognize that markets do not function perfectly without having to return to discussions of the developmental state. Rather than rebuilding states reduced by structural adjustment and other neoliberalizing reforms, the institutional argument led to an emphasis on property, contracts, and information asymmetries. Whether this was the reason or not, the institutional turn became apparent in many development agencies as well as in the literature – and it is certainly the case that development economists and rational choice political scientists were at the fore of this intellectual innovation.

One ongoing debate among macro-economists has been whether institutions (understood primarily as enforceable contracts) or geography (understood primarily as location and resource endowments) are the prime explainer of overall development performance. While authors such as Dani Rodrik emphasized that “institutions rule” (Rodrik, Subramanian, and Trebbi 2004), others in the camp of Jeffrey Sachs (2003) argued instead that location, distance from the sea, and other basic locational characteristics were ultimately more important. A variant in this discussion has been the turn to historical institutions on the part of some development economists and political scientists who have sought an explanation of divergent forms of development in the nature of colonial institutions of property and exclusion (Acemoglu, Johnson, and Robinson 2001).

Another avenue toward the institutional question has been the long historical debate on the relationships between state, market, and civil society. This debate was fueled by, inter alia, the increasing interest in nongovernmental and community-based organizations in certain aspects of development (in particular local development and natural resource management). Typically such discussions have come to the view that the quality of development hinges not on the absolute prioritization of one or other of these
sectors, but instead on the nature of the relations between them, or in other terms how they “govern” each other, be this through relationships of regulation or of complementarity. The focus on governance always tracks back to “getting the institutions right” and also establishing relationships that will bring such “good” institutions into being and sustain them. One particular favorite has been to establish institutions that enhance transparency: of the many venues in which this can be seen, the Extractive Industries Transparency Initiative (EITI) and the Open Government Partnership are just two examples.

Another domain in which these questions of complementarity, regulation, and institutional design have been explored has been that of financial services – both formal and microfinancial. The debate on institutions for regulation of the formal sector has taken on renewed vigor since the financial crises of the last two decades, while the discussion of microfinance has been a cornerstone in development thinking and intervention for at least three decades now. Early microfinance discussions focused rather more on the establishment of alternative institutional forms for delivering credit to very poor people who otherwise did not have access to such financial services. These discussions tried to understand the formal and informal rules that made Grameen such a success, as well as the formal and informal rules that underpinned emerging models such as the increasingly popular “village banks.” These ideas were then taken up by an increasing number of nongovernmental organizations (NGOs) and donor agencies as they each rolled out models for the delivery of microfinance as a vehicle of poverty reduction, economic empowerment, and enterprise development. As time went on, however, discussions began to consider the types of institutional arrangement that would be needed to provide additional financial services, such as savings accounts, insurance, and money transfers; and, at the same time, the formal banking sector began to look at these experiences, and to see the lower parts of the pyramid, if not the bottom, as a potential market. The effect of these developments was that the microfinance discussion was steadily taken toward the state–market–civil society framework. The questions became: What sort of rules would need to be in place to allow formal banks and NGOs to collaborate, to ensure that village banks or NGOs that began deposit schemes would be good guardians of people’s savings (and did not run off with them), and to bring the microfinance sector into the sphere of regulation like the formal financial services?

A further strand of institutionalism that has been especially influential in development theory and practice is that stemming from Elinor Ostrom’s (1990) work on institutions for collective action. Originating in an effort to understand the institutional arrangements under which common pool resources could be managed collectively in ways that avoided the “tragedy of the commons,” this work led to a far broader framework for understanding the institutional arrangements that would lead to more sustainable and inclusive management of resources such as forests or irrigation systems, that could facilitate the joint provision of services through the combined action of government and civil society, and that could promote what became known in World Bank parlance as “community driven development.” If the line of analysis reflected by the work of Rodrik, Acemoglu, Johnson, Robinson, and others was concerned above all to discern the institutional explanations of variation in macro-economic development, Ostrom’s approach (since continued by scholars such as Arun Agrawal, Krister Andersson, and others) was instead interested in explaining institutional sources of variation in local and resource specific development as well as in
identifying the “design principles” under which such development might be more “successful.” Within development geography, this work on institutions for local natural resource governance has been perhaps most influential in efforts to understand the dynamics surrounding protected area management and co-management, as well as use and management of forest and forest margins.

For many geographers these approaches to institutional explanations of trajectories of development have felt excessively formalistic, requiring a degree of simplification (in order to define variables and indicators that can be compared across cases) and emphasis on design principles that sit awkwardly with concerns for context and place specificity. Indeed, the discipline has made only limited contributions to theories of the relationships between institutions and development. It is probably economic geography, and in particular those parts of the subdiscipline most closely related to economics and regional science, where most has been contributed to more theoretical debates about institutions (Storper 2013). The primary concern of this work has been to understand the institutional conditions underlying the emergence of vibrant regional economies and industrial clusters, with a particular emphasis on OECD (Organisation for Economic Co-operation and Development) economies.

Institutions and development geographies

Institutional theory has been dominated by the work of economists, political scientists, management theorists, and some social theorists. While this literature has, by and large, treated institutions aspatially, there are many senses in which the relationships between institutions and development need to be understood geographically.

First, while the rules constituting some institutions may exist aspatially (e.g., tax rules, or rules governing access to social protection programs), their effects are very much spatially differentiated. Second, to the extent that institutions only exist when they are practiced, the geography of their practice may well be highly uneven across space. Rules governing environmental standards may exist nationally, but their enforcement will vary across space as a consequence of the spatially uneven presence of the state or other bodies that oversee the application of such rules. Third, one institution may interact with other institutions in ways that vary across space. Continuing with the earlier example of environmental standards, enforcement of formal governance institutions may vary depending on the geographies of corporate responsibility (leading some companies to take up environmental rules more willingly than others) or the geographies of corruption (in which some subnational authorities are more resistant to corruption, and thus more likely to enforce standards, than others).

A fourth reason for being spatially explicit in the discussion of institutions relates to the very ways in which the nation-state is understood. Thus, while sets of rules may exist at a national level, the state (or indeed other actors) may establish areas of exception in which these rules do not apply, because of state policy (which may, for different reasons, suspend human rights and environmental or tax institutions in certain spaces in the name of national interest, security, or the like), because of the geographies of insurgency (determining areas in which some state rules cannot be applied but others can), or because of geographies of crime which can also displace the rule of law as well as a swath of other institutions. Perhaps the more general observation here is that the presence of the state (as a bundle of institutions) varies across space to the extent that the rules that underlie the state are practiced
unevenly across space. Indeed, in some sense the project of nation-state building is one of trying to extend its institutions across territory in ways that are increasingly similar. The converse situation is one of institutional patchiness and, at an extreme (which characterizes very many countries), the coexistence across “national” territory of differentially governed fiefdoms (as, for instance, where parts of “national” territory are instead governed by insurgent groups, companies, mafias, paramilitaries, or large property owners).

**Explaining institutions and their change**

Let us return to the lament of the economic journalist in the introduction. While existing institutional arrangements might be critiqued, and alternatives called for, such commentaries beg two questions. First, if existing arrangements are undesirable or suboptimal, why do they exist in the first place? Second, what does this explanation of current institutional arrangements imply for the processes through which they might change? There are no generic answers to such questions, and answers may likely vary depending on the type and scale of institutions in question. However, some general observations can be made.

While existing institutional arrangements might be understood as reflecting dominant interests, such an observation neither identifies what those dominant interests are, nor the indicators that prove that they are indeed “dominant.” Perhaps most importantly, the observation itself does not establish the mechanisms through which interest groups were able to secure such arrangements and render them sufficiently stable and respected such that they could be deemed “institutions.” Two relatively recent and influential efforts to develop such a line of analysis have come from the team of Daron Acemoglu and James Robinson (2012), and from the political economist Mushtaq Khan (2010). Acemoglu and Robinson pose the macro-developmental question of “why nations fail” and argue that those nations that do indeed fail tend to be characterized by “exclusive” economic and political institutions that allow wealth accumulation through the control of rents, exclude many in society from property rights and market access, limit political participation, lead to discretionary application of the rule of law, and prevent access for the majority to authority and government. This argument is very similar to North, Wallis, and Weingast’s (2009) who distinguish between limited-access and open-access orders, the former being much more exclusive in how access to political power is organized. Acemoglu and Robinson (2012) explain the general structure of institutional arrangements in a society in terms of “political equilibria” that are understood as the set of political and economic institutions that are compatible with the balance of power between groups. In a similar vein, Khan (2010, 20) argues that the determinants of institutional arrangements have to be found in the “political settlement” on which society is founded, understood as “the interdependent combination of a structure of power and institutions at the level of a society that is mutually ‘compatible’ and also ‘sustainable’ in terms of economic and political viability.” These settlements, according to Khan, constitute a sort of agreement regarding how power will be exercised to enforce rules governing access to political authority and the distribution of benefits in society.

These approaches imply that actually existing institutions can only be understood in terms of the way in which political and social power are exercised in a given country. (Indeed, the focus of these approaches is on the national level, though the arguments could easily be taken to a subnational level, raising the interesting question
as to how subnational and national settlements interact with and constitute each other.) This is an important counterpoint to what can often be design- and blueprint-oriented approaches to institutional change, because the implication is that, in addition to design (because, of course, design still matters), sociopolitical relationships and incentives must also be such as to favor a given institutional change. This might seem like an obvious observation: however, it is an important one, not least because much development intervention has paid relatively little attention to understanding the real world of political power when devising institutional changes — helping to explain, perhaps, why so much development “fails.” Moreover, while this emphasis on political power may, again, seem obvious to geographers, these contributions force the very difficult questions of how the nature, structure, distribution, and exercise of sociopolitical power is actually organized. This is not an idle task, and indeed contributions such as these have elicited significant work in some development organizations seeking to address the “political drivers” of development and state formation (see, e.g., www.effectivestates.org).

While these frameworks offer important insights, they say much less about the cultural politics that underpin particular institutional arrangements. Different institutions become dominant not only because of the political interests to which Khan, Acemoglu, and Robinson draw attention, but also because of the ideas and narratives that make those institutions seem acceptable and appropriate. The emergence of the institutions that underlay the neoliberal economic reforms characterizing the period since the 1980s was itself facilitated by the rise of certain ideas and arguments, themselves elaborated and broadcast by networks of intellectuals, think tanks, and journalists. Similarly the institutional changes fostered by progressive governments in Latin America since the 2000s have also depended on the crafting of arguments regarding resource nationalism, sovereignty, and the role of the state. The importance of such narratives in stabilizing social order is, among other things, one of the lessons of poststructural approaches to government and Gramscian approaches to hegemony. While such literatures are not typically associated with thinking about institutions, they contribute a lot to understanding how certain rules of the game become dominant. More familiar analysts of institutional change say something similar: “the politics of ideas is intrinsic rather than epiphenomenal to the processes of coalition formation that underpin institutional change” (Hall 2009, 212).

These insights into how institutions become dominant also help explain how they might change, with consequences for patterns of development. First, institutions can change as a result of the emergence of actors with what Khan would call “holding power,” the ability to upset existing political settlements. This can occur at different scales. At a national level, Boix showed how the steady emergence and consolidation of a professionalized middle class in Franco’s Spain slowly but surely asserted its growing influence on the development trajectory of that country, ultimately helping to take it toward democratic transition and economic development. Subnationally, institutions governing forest control and access (for instance) can change over the course of time as movements of forest dwellers (indigenous or otherwise) have become stronger; examples range from Chipko in South Asia, through to rubber tappers and indigenous organizations in Latin America. That said, these institutions can also change in ways that are less than inclusive or sustainable as small- or large-scale loggers become more organized, as drug-funded networks become consolidated, or as new industrial interests emerge (e.g.,
hydrocarbon industries concerned to access oil deposits located in forested protected areas).

Second, institutions can change under the influence of external pressures, albeit generally when also aligned with particular domestic interests. The institutional transformations in Mexico and Central America that have derived from free trade agreements with the United States clearly reflected the exercise of US power, albeit with the clear support of certain domestic interest groups also. These institutional changes in everything from environmental governance to trade tariffs and laws governing capital flows rewrote the rules of the game in ways that have left profound imprints on patterns of development in the region. As another example, the World Bank and the Canadian government have clearly influenced the rewriting of mineral codes in many countries. And, in a more progressive sense, the United Nations and other external bodies have also influenced human rights codes.

Third, institutions can change under the influence of new narratives. The importance of narrative and ideas in regulating social life and social organization has been an insight of post-structural and Gramscian approaches in social movement theory and political ecology, as well as of work on the role of epistemic communities in policy processes. There are many examples of this within development studies and practice. To return to an earlier example, the general spread of microfinance institutions was assisted by a whole narrative-building machine that assembled academics, development observers, Muhammad Yunus (the spokespeople for NGOs such as BRAC), and others to craft an argument that finance should be thought of differently, and that there was no reason that poor people could not be disciplined subjects of banking services. The emergence of joint and community forest management institutions has been in some sense an effect of the combined demands of forest user organizations and an international epistemic community working on community forestry. In a similar sense, the growth of participatory institutions, while clearly owing much to the demands of strong and party politically articulated social movements (as in the case of participatory budgeting in Brazil) also has to be understood as an effect of the emergence of a global movement around participation and development combining ideologues (ranging from Robert Chambers to Orlando Fals Borda). Experiments in gender-based budgeting (to name but one) similarly reflect the combination of domestic feminisms and international epistemic communities in gender and development. And, finally, the slow emergence of institution regulating for prior consultation of large-scale investment projects with indigenous peoples reflects the crafting of domestic arguments about indigeneity, race, and rights, as well as the global pressure of movements organizing around the same issue and of dedicated bodies and activists within the UN system. The emergence and consolidation of certain sets of institutions over others depends, then, on the circulation and persuasiveness of narratives endorsing these institutions.

One of the implications of such observations is that institutional change will rarely occur just because a donor agency or a government wishes to introduce new institutions. Instead the change will occur when these underlying conditions are favorable to change. Of course, donors (say) may still introduce change – but the success of such change depends on how far proposed changes align with the broader political context, and the degree to which the changes are assumed by other social actors. Agencies that take a far longer approach to institutional change would first work on strengthening the political influence and legitimacy of certain actors such that they have the political power to subsequently be “insurrectionary” drivers of institutional change.
(to use Mahoney and Thelen’s (2009) term). There is no doubt that some actors have indeed followed such a strategy in the past – building first the political infrastructure for institutional changes before then introducing these changes and determining their details. Reminding us that there is nothing necessarily progressive about such a strategy, this approach was, for instance, the idea that lay behind US foreign policy during the 1980s and 1990s to support conservative forces in Central America (incumbent governments, business groups, factions of the military, counterrevolutionary groups). Only by enhancing the power of such groups were certain institutional changes in the organization of production, trade, ownership, taxation, and so on possible. However, supporting such groups alone would not itself have been sufficient to secure the institutional changes that the United States wished to promote. The detailed work on institutional reform also had to be done. The support offered by other actors to pro-democratic forces during periods of dictatorship in Chile, for instance, was a similar strategy – a strategy of building the political foundations of institutional change as a necessary but not sufficient precursor of such change – just pulling in an ideologically different direction.

Even though the definition of “institution” can often remain slippery, there is now much convergence around the idea that the nature of institutions is a critical determinant of the quality of development. This institutional turn then begs the questions implied by the journalist at the beginning of this entry: Where will “better” institutions come from? What will drive their emergence? And what will they look like? These questions take the discussion of development back to social and political terrains that are familiar to geographers – but which require very careful and disaggregated forms of analysis.

SEE ALSO: Corporations and the nation-state; Governance and development; Networks, social capital, and development; Participatory development; Property and environment; States and development

References


Interdisciplinarity and geography

Louise J. Bracken
Durham University, UK

Calls for interdisciplinary approaches to real-world, complex problems are becoming increasingly common, especially in research concerning the natural environment. The rationale behind such calls is to bring quality research from different disciplines together to overcome increasing specialization. Geography is regarded by many as a discipline, but it involves very different trainings for practitioners in human and physical aspects. Thus the issues surrounding geographers with different backgrounds working together are analogous to those around interdisciplinary working between the physical and social sciences. This entry initially reflects on what interdisciplinary working is and why researchers enjoy undertaking such research. It then considers a number of themes that are central to undertaking interdisciplinary research: the role of language, the importance of framing, and the inclusion of stakeholders in interdisciplinary research.

What is interdisciplinarity?

The term “interdisciplinary” has multiple meanings, “from simple borrowings and methodological thickening to theoretical enrichment” (Klein 1996, 153). Some have therefore suggested that “interdisciplinarity is best understood not as one thing but as a variety of different ways of bridging and confronting the prevailing disciplinary approaches” (Huutoniemi et al. 2010, 80). The appropriateness of interdisciplinary research is based on the premise that its associated collaboration and networking will produce innovative concepts and methods to answer complex research questions that are beyond the expertise of individual disciplines (Bruce et al. 2004). At the outset, it is important to clarify what is meant by the terms “discipline” and “interdisciplinarity.” A discipline is a branch of learning or scholarly instruction which is defined by institutional boundaries constructed by the needs of teaching, funding, administration, and professional development. The perceived boundaries of a discipline may change over time. Students are prepared for work within a discipline by an appropriate and specific training. As Dewsbury and Naylor state, “disciplines themselves are the outcome of ‘a variety of practices whose conceptual identities were the outcomes of local patterns of training and socialization’” (2002, 255).

Interdisciplinarity is a collective term encompassing all forms of scientific collaboration where the field of a single discipline is transgressed (Balsiger 2004). Interdisciplinarity is therefore about creating something new by crossing boundaries between academic disciplines or schools of thought, as new needs and thinking across them emerge (Bracken and Oughton 2006). Figure 1 presents a simple representation of the differences between discipline-based research, multidisciplinary, interdisciplinary, and transdisciplinary research. During discipline-based research disciplines bring their own trainings, understandings, and perspectives to a research issue/topic. There is no interaction between the
disciplines (Figure 1a). In a multidisciplinary approach, research continues to be undertaken within distinct disciplines around a certain theme or topic, but ideas and outputs are then brought together to relate to the research topic (Figure 1b). Outputs are thus joined and thought about together, but methods and perspectives are not integrated. An interdisciplinary research approach demands interaction between the trainings, expertise, and perspectives between disciplines. Thus the research topic sits at the intersection of a range of disciplines (Figure 1c).

Thus an integrated understanding and approach to the research topic is established. Transdisciplinary research is similar to interdisciplinary research but with the addition that those who are affected by the issues with which it is concerned, or who have a stake in the issue, are involved in the research (Figure 1d). Transdisciplinary research has come to symbolize a broad shift in research agendas toward an emphasis on utility (Bracken, Bulkeley, and Whitman 2015). This approach is based on the understanding that there are many equally legitimate sources of

Figure 1 Differences between (a) discipline-based research, (b) multidisciplinary research; (c) interdisciplinary research; and (d) transdisciplinary research.
knowledge and evidence that need to be drawn on to inform the management of environmental problems (Wynne 2003; Petts 2007).

Achieving good interdisciplinary research is not necessarily straightforward. It is well recognized that interdisciplinary practice is challenging because of differences in epistemologies, knowledges, and methods (Evans and Marvin 2004); different ways of formulating research questions (Oughton and Bracken 2009); differences in the nature of communication (oral and written) (Bracken and Oughton 2006); and differing attitudes within and between disciplines as to the importance of interdisciplinarity (Lowe and Phillipson 2009). These challenges can be exacerbated by expert knowledge becoming segmented and bureaucratized and compounded by the fact that science and technology are often grounded in the philosophy of reductionism, which tries to simplify and prise apart complex, real-world problems into manageable chunks of discipline-specific research (Balsiger 2004). With the increasing appreciation of difficulties and barriers to undertaking interdisciplinary research, publications have recognized and highlighted practice for successful collaboration. Key suggestions are: (i) longer start-up phases of projects to promote cohesion and to learn to value contributions from other disciplines; (ii) development of projects which satisfy complex societal problems; (iii) effective and experienced management of interdisciplinary teams; (iv) flexibility of researchers involved in conducting the research; and (v) understanding the ways of thought and language of others (Bracken and Oughton 2006). Of central importance is learning to work with the messiness of complex real-world problems rather than trying to simplify them (Donaldson, Ward, and Bradley 2010).

Boundaries can be both intellectual and administrative, and exist at a variety of scales. Often tension exists around these boundaries.

For instance, geography as a discipline is composed of two very different approaches, social science and physical science, which map onto very different ways of working. Interdisciplinarity can thus unfold within a single department of geography as well as between other departments (Bracken and Oughton 2006). Administrative practices can also influence the ways in which disciplines are encouraged to come together to undertake research. For instance, the Research Exercise Framework in the United Kingdom has tended to encourage departments to have a strong disciplinary core. Against such a backdrop, publishing and reaching out beyond high-profile geography journals in talking to interdisciplinary audiences becomes challenging.

Why do interdisciplinary research?

Oughton and Bracken (2009) found that academics leading funded interdisciplinary projects were keen to undertake this type of research to “give something back” and/or to “make a difference” (emphasis original). In their paper Oughton and Bracken (2009) were referring to applied research rather than coproduced knowledge. Justifications for undertaking interdisciplinary research were implicitly grounded in the notion that, by including the needs of others, it is somehow more valuable. Successful projects were able to identify and support the processes that allowed the communication and negotiation that is necessary, not just for the initial framing of a research funding proposal, but were able to maintain negotiation throughout the research (Oughton and Bracken 2009). Self-awareness, continual reflexivity, and a willingness to be questioned by others were essential to this process. It is vital to acknowledge the continued role of reframing. This is especially important for projects collecting data across the social and
natural sciences because it is difficult to map different epistemologies and methods successfully. Although hard, this is rewarding and may lead to new and exciting knowledges. Oughton and Bracken (2009) found three different, although not mutually exclusive, routes to undertaking interdisciplinary research: first, through collaboration with people from different disciplines, where researchers remained centered within their own area of expertise but were willing to engage with and to trust the expertise of others; second, through reading adventurously and developing understandings that allowed researchers to work critically with others in different disciplines, but without more formal collaboration; and, finally, by undertaking training in a completely new area or field, with some of those interviewed taking high-level qualifications in multiple fields of expertise. These three routes were supported by specific practices, such as working with physical objects (e.g., models), visiting places to experience how interdisciplinary research is practiced, and serendipity.

The role of language in conducting interdisciplinary research

Language may determine the positionality of the researcher, the way in which the research question is framed, the translation of the “field” to the academy, and the development of the theoretical context (e.g., Mirowski 1994). To try to improve practice in interdisciplinary research, previous work has tried to understand the ways in which we use language and how it may help or hinder our understanding of each other and of our respective disciplines. This work has been concerned with the contemporary, multiple meanings and uses of words in practice. Bracken and Oughton (2006) identified three distinct issues: registers of speech termed “dialects” and “metaphor” and the process of articulation (see Bracken and Oughton (2006) for case studies of instances where different perspectives on the meaning of words can lead to problems for common understanding).

_Dialects_ represent the difference between everyday use of a word and specialist, disciplinary use of a word (Bracken and Oughton 2006). Dialects are also produced by the same word having slightly different meanings within different disciplines (Bruce _et al._ 2004), which may also be different from their everyday meaning. Words which are in everyday use tend to be those that cause the most difficulty for the unwary interdisciplinary scientist. These misunderstandings may be exacerbated by the fact that academics are unlikely to question the meaning of a word with which they are already familiar. The conversation may be well developed before it becomes apparent that a particular word has a specific disciplinary interpretation as well as its everyday use. This situation is bound to lead to frustration for those coming together from different academic specialisms to undertake interdisciplinary research.

The second aspect of language is _metaphor_. The use of metaphors is common in everyday discourse and is embedded in our language; we rarely think about them or are aware that we use them. When working within disciplinary teams we tend to have common metaphors on which we draw to express and explain ideas. For good interdisciplinary practice we need to be aware of the times at which we move into separated speech communities and when the form of metaphor being used may be misinterpreted (Bracken and Oughton 2006).

The final aspect of language highlighted by Bracken and Oughton (2006) differs from the first two in that it is a process rather than a register of speech. _Articulation_ involves deconstructing
one’s own disciplinary knowledge in conjunction with those of other disciplines in order to understand the building blocks and thereby reconstruct a common understanding. The idea of articulation is particularly stimulating and an accurate description of the very active discussions involved in framing and undertaking interdisciplinary research. Slightly different definitions of the same word can highlight alternative starting points in terms of thinking about the research, related to disciplinary backgrounds. Breaking these down and developing a shared, common understanding creates a more powerful platform from which to develop a project. One aspect of this way of working is to define a whole that is greater than the sum of its parts. A second aspect can also emerge through the process. Each member of the team is constantly tested in their assumptions and perceptions, and while the process is difficult and time-consuming it can be rewarding and result in a much stronger basis from which to develop research (Bracken and Oughton 2006).

Framing

The use and role of language can be important in developing the frame by making sure people from different disciplinary backgrounds understand each other, but also by enabling new knowledges to be produced from which to develop a frame. Framing encompasses the processes of identifying and bounding the area of research. Few studies have explored practices of framing, although there are some guidelines for good practice (e.g., Mostert and Raadgever 2008). Distinctions have been made between framing and sense-making with the term “sense-making” being used to stress the internal, self-conscious process of developing a coherent account of what is going on, while framing emphasizes the external, strategic process of creating specific meaning in line with political interest (Fiss and Hirsch 2005). Others do not distinguish between the two because they see both as being an integral part of the framing process (Oughton and Bracken 2009).

There is a considerable amount to be learned about the negotiations behind successful framings and the factors that affect them that may be of use to the researcher embarking on interdisciplinary research. Within traditional disciplinary ways of working, the framing of the research question often arises directly from previous research or is left as implicit. The process of framing the research is not explicitly defined because it is accepted, or habitual, within a school of thinking which shares an ontology and epistemology. Hence the research process tends to follow an established pattern and is concerned with limiting a distinct experiment or investigation rather than placing it in a context (Oughton and Bracken 2009). Explicit negotiations associated with framing in this context need not be extensive and may be minimal. This is not to say that framing undermines the integrity of the research, but it may be that it is just not necessary. In the case of an interdisciplinary project none of this process can be taken for granted. Extra negotiation and uncertainties surround working in a research project that takes place outside and across disciplinary boundaries. By exploring the activity of framing across a number of different projects, it is possible to identify a range of features that support and shape the successful initiation and continuity of the research process (Oughton and Bracken 2009).

Many of the ways and processes of developing interdisciplinary research projects are no different from the initiation of ideas in disciplinary projects, but what is very different is the negotiation that is involved in developing the frame. Oughton and Bracken (2009) proposed two core
INTERDISCIPLINARITY AND GEOGRAPHY

elements to framing: identifying the research focus and determining the common objects with respect to scale, place, and time, and identifying the research team. The team may exist before the idea or may be brought together in line with the research question. Hence, researchers affect the research question and the research question determines who is drawn in as a researcher. One aspect of the process that interviewees commented on explicitly was that the iterative and negotiated natures of developing the research project which may well mean withdrawing a particular contribution, perspective, or person from the project should their input no longer seem relevant – in this case, the frame.

There is also a third element to developing the frame of research that is especially relevant to undertaking research around risk: the problem itself and letting the problem exist as a frame (Bracken 2012). This element of framing is therefore about letting those who live with an issue challenge the ways in which it is framed in our research. Allowing those who matter to speak back may not simply involve “normal conversations” with, for example, those who live with flooding, but conversations with a much wider set of “things,” both animate and inanimate, that constitute and formulate the problem that is of interest. Gaining input from others to help frame and shape research can occur in a range of ways. These range from stakeholders raising an issue which sparks an idea with researchers, to knowledge being coproduced with stakeholders (see the discussion on transdisciplinary research in the next section for further discussion of the role of coproduction in interdisciplinary research).

What came out strongly from the work of Oughton and Bracken (2009) was not just the amount and intricacy of debate that was necessary at the outset of establishing an interdisciplinary research project, but the continued negotiation that was demanded throughout the lifetime of the research. This is because the act of researching, or coming to know, a particular framing becomes, in itself, something that unsettles that framing in a similar way; this is termed “knowledge shifts.” It also emerged that researchers found it important to recognize that interdisciplinary projects are iterative because they are drawing on a number of different disciplines and that the research may be particularly complex and liable to fail in some aspect. One therefore has to be ready to change to ensure an outcome, especially with the involvement of stakeholders who are often looking to the research to solve immediate problems (Oughton and Bracken 2009). Researchers found it important to understand what others were saying and meaning to be able to negotiate the developing frame; thus the explicit use of language was important.

**Transdisciplinary research: the inclusion of stakeholders**

Few studies have tried to understand what stakeholders feel about being involved in research and how academics could work differently to develop more successful transdisciplinary practices. Such research is vital to developing more sustainable transdisciplinary research through greater reflections on the relationships between “science” and the “public,” as suggested by Irwin and Wynne (2003, 7). Bracken, Bulkeley, and Whitman (2015) initiated a study with a variety of different types of stakeholder (nonacademic expert and the wider public) involved in academic research in a range of ways: as subjects in focus groups and members of competency groups through to inclusion in participatory
action research, advisory panels, and work shadowing schemes. The research demonstrated that stakeholders have multiple interests and reasons for getting involved in academic research, although a desire to acquire new knowledge, either for its own sake or around a particular issue, was pertinent to all participants. A common thread that united stakeholders was that they all advocated a research process that is open, based on dialogue, flexible and where there are multiple measures of success.

A number of authors have raised the issue of “stakeholder fatigue” in participatory research and suggested that it can be due in part to the ephemeral nature of knowledge and poor communication within research projects (Du Toit and Pollard 2008; Holman 2013). Bracken, Bulkeley, and Whitman (2015) found that participants in their study confirmed that they would continue to take part in academic research if they found it interesting and relevant to their needs. Thus stakeholder fatigue may be lessened, and more sustainable participation maintained, if academic projects engage and communicate more effectively with stakeholders. To this end it is essential that there is flexibility in research project design and/or the research process to enable thoughts, ideas, and suggestions from stakeholders to be incorporated from conception to outputs within projects (Bracken, Bulkeley, and Whitman 2015). It was also clear that stakeholders valued and appreciated the debated and contested nature of the knowledges produced in their projects. This led to their gaining a better understanding of other people’s perspectives (both from each other and from academics), which were then applied, either through professional and/or personal routes, in their own contexts. Consequently, transdisciplinary working can counteract the “exclusionary and socially disengaged policy tradition characterized by invocation of the objective authority of scientific expertise” (Irwin, Elgaard, and Jones 2012, 128). Nonacademic experts and the wider public had similar experiences of being involved in academic research and both gained substantial knowledge and enjoyment. This is an important finding because the wider public is not often included in transdisciplinary research, only scientists and practitioners (e.g., Lang et al. 2012; Jolibert and Wesselink 2012).

The aspect that varied most between stakeholders was the benefits from being involved in academic research. Phillipson et al. (2011) noted that there was a positive impact on research relevance and quality, but the impact on stakeholder practices or knowledge seemed slight. In contrast, Bracken, Bulkeley, and Whitman (2015) demonstrate that the impacts on stakeholders can be significant. All stakeholders valued the increased stock of brokered knowledge that was gained through the coproduction process and suggested that this had made a difference to them even without giving an explicit impact or change to their daily ways of working (practice). In other words, the process itself was of value to them independent of any specific outcomes – a finding that questions those who argue for the ephemeral nature of coproduced knowledges (i.e., Holman 2013). Participants felt a personal pride at being involved in their respective research projects, even if something concrete had not been attained, although as mentioned earlier, the new networks and brokered knowledge were judged to be incredibly valuable to participants. Another important finding was the value of the “intangible impacts” on stakeholders of being involved in participatory research. These ranged from increased personal confidence, learning the languages of other stakeholders, to their continued engagement once researchers had completed their studies (Bracken, Bulkeley, and Whitman 2015).
INTERDISCIPLINARITY AND GEOGRAPHY

Both nonacademic experts and the public feel strongly that their judgments and values should be regarded as an integral part of the research process, as proposed by Macnaghten and Chilvers (2012). Furthermore, flexibility, openness, and the iterative development of projects enable stakeholders to contribute to the coproduction of knowledge and to enjoy being part of the research process (Bracken, Bulkeley, and Whitman 2015). This, they argue, may challenge previously published research where legitimacy has been used in reference to determining whether a stakeholder holds the “right” type of knowledge and experience to bring to transdisciplinary research (Lang et al. 2012).

While such approaches support the inclusion of stakeholders at the outset of the research process to shape projects, it remains the role of academics to determine which stakeholders hold “relevant” knowledges or expertises to be invited into the research (Lang et al. 2012). Such approaches maintain the authority of scientific knowledges and expertise while continuing to marginalize or even ignore the judgments and values of the wider public, based as these are in what might be termed more “local knowledges.”

Summary

The defining characteristic of interdisciplinary research is that it involves the integration of two or more academic disciplines into one activity such as a research project. Interdisciplinarity is thus about creating something new by thinking and working across boundaries. This can include working with stakeholders to coproduce knowledge, as in transdisciplinary research, or remaining within academia, as in interdisciplinary research. Interdisciplinary and transdisciplinary working are often necessary to explore real-world issues that do not fall within the bounds of a single discipline. There are many pitfalls to be avoided, but these can be pre-empted and sensible project planning can be used to develop successful interdisciplinary ways of working. Micropolitics plays a significant role in the ability to practice good interdisciplinary work and this is reflected in the ways in which certain kinds of language become used and in terms of how projects become framed. The notional hierarchies of disciplines, the personal ambitions and competitiveness of colleagues, not to mention the implicit and long-standing issues of power surrounding gender relations all play an important part in determining interactions within a research team. Understanding the role of language in interdisciplinarity is not going to solve these problems—it does, however, offer a route to making them visible. Sharing and exploring a speech community involves transparency. There is nothing to hide behind if one is sincerely translating one’s work for others and is simultaneously engaged in actively listening to their contributions. This process involves becoming vulnerable and thus open to the misuse of power. Good interdisciplinary work, therefore, is not possible without mutual trust and respect. Recognizing the roles of dialects, metaphor, and articulation is a move toward creating a new space for interdisciplinary intellectual engagement. Related to this crucial use of language is the way in which interdisciplinary research is framed.

Individuals tend to have a strong predisposition to work across disciplinary boundaries and an openness to collaborate with others. Successful projects are able to identify and support the processes that allow the communication and negotiation that are necessary, not just for the initial framing of a research funding proposal but as the project progresses. Self-awareness and continual reflexivity and a willingness to be questioned by others are essential to this process.
It is vital to acknowledge the continued role of reframing. This is especially important for projects collecting data across the social and natural sciences because it is difficult to map different epistemologies and methods successfully. Subprojects which may encounter problems when underway should, therefore, not be written off but negotiated so that they can be reframed. Although hard, this is a rewarding exercise and may lead to new and exciting knowledges. What is required for the successful framing of a research project is the coordination of the different perspectives to reveal the problem thus making it open to research.

Critical to the coproduction mode of transdisciplinary research is maintaining flexibility within the research process both in framing how the research is conducted and in the meaning of what constitutes “success.” Negotiation and working iteratively can help both stakeholders and researchers expand their knowledges about an issue, understand each other’s views, and hence reach a common understanding (although not necessarily agreement). Embracing the idea that knowledge is multiple and as likely to be held by community-based stakeholders or professionals as by academics is critical to the development of such dialogues. Transdisciplinary research is thus highly relevant, and should be a foundation of impact assessments of academic research.

Geography and geographers are well placed to undertake inter- and transdisciplinary research. A wealth of good practice and successful interdisciplinary research projects involving geographers already exists. We should learn from these success stories, take stock of studies that can help us refine interdisciplinary research practice, open our minds to being challenged from other disciplines, and embrace the multiple sources of knowledge that can assist us in developing high-quality, novel, and useful research. Most of all, we should have fun and enjoy challenging boundaries – undertaking inter- and transdisciplinary research is one way of doing this.

**SEE ALSO:** Critical geography; Cross-cultural research; Environmental science and society; Ontology: theoretical perspectives

**References**


Maps produced on paper present their information visually. Cartographers designing for static maps communicate by making marks on a page, arranged in some way that the spatial relationships among the marks correspond to the spatial relationships of the objects in the real (or imagined) world. Apart from the mental processes of knowledge construction, readers interpret a paper map passively, able to interact with it only very little, possibly rotating it in their hands in order to aid in navigation, or folding it in order to limit the view (or fit it on a table).

Digital maps, on the other hand, can vary significantly in the ways that information is presented to users and in the ways users may interact with the displays. Digital map-makers now must consider not only the design of the marks on the page (e.g., the screen, still very important) but also the design of the myriad ways that the map can be manipulated and used. In this sense, cartographers are now user interface designers – while in the past their primary concern was an appropriate and intuitive graphical display of geographic information, today, cartographers must examine the diverse uses, and diverse users, of the maps they create, and design appropriate and intuitive interfaces that connect the user and the map (and thus the information contained within the map).

The term “mapping” is used in a noncartographic sense in human–computer interaction to refer to any encoding of user input (like a mouse-click) to an action on the computer (like moving from one web page to another), or any encoding of a piece of information (like a news item) to a stimulus (like a hyperlink on the screen). Now that digital maps are much more than passive visual representations of geographic data, with capabilities for both dynamic, customizable display and detailed, multifaceted interaction, it is useful to think of the cartographic process in this broader sense as well – actions of the user can be mapped to a reaction by the computer (a double-click will zoom the map’s scale), and geoinformation can be mapped to a mark on the screen (a church’s location is represented as a point in the pixel space of the display). It should thus not be surprising that cartographers are keenly interested in interface design as a means of efficiently connecting users of digital geoinformation to that information and the meanings and interpretation of it.

However, despite the possibilities afforded by digital geographic representation, the user interfaces for digital mapping environments have changed very little in the last thirty or so years. Digital maps owe much to the so-called WIMP (windows, icons, menus, pointers) interface, popularized in computers like the Apple Macintosh in the 1980s. As is the case with most computing environments in use today, digital maps rely on input from point-and-click computer mice (or, more recently, touchpads) and display their output via graphical windows that display visual plan–views of the Earth. This does not have to be the case, however. Research efforts in human–computer interaction (HCI), virtualization, and geographic visualization have all experimented with alternative and multiple modalities.
of digital map user interfaces that employ senses and cognitive channels that complement or even replace traditional interface modalities.

Vision and mapping conventions

In many ways, static maps themselves can be considered user interfaces: they are the intermediaries between a user and information, enabling users to understand a complex world through graphical symbols, conventions, and metaphors. Computer user interfaces (like the Windows desktop or the “home screen” of a smartphone) are abstractions of complex computational information: rather than typing commands and reading machine languages, users are allowed to interact with computers graphically and rely on metaphors such as folders, hierarchies, and slider bars to make sense of, and influence, the inner workings of the computer. Maps – even those created on clay tablets – do the same thing: maps are an abstraction of a complex world, simplified and generalized to their “essences,” at least according to cultural cartographic conventions.

While conventions of representation are useful to create and conform to a common language of information communication, they can also serve to restrict information to that which is expected or traditional. A classic example of a cartographic convention is that north is “up” on static maps; more subtle is the convention that elements of the environment that are culturally significant (such as courthouses or shopping malls) are shown at the expense of those that are deemed less important. But an even more fundamental convention on maps is that the Earth is represented visually.

This convention is very logical for the purposes of maps when they were simply static representations of Earth. Vision is the dominant mode of processing spatial information, as our vision is spatial – though we focus on only one point as we use our eyes, we can sense both breadth (using our peripheral vision) and depth (using our binocular vision). The transformation of the 2-D (or “2½-D,” if surface undulations are considered) surface of the Earth onto a 2-D sheet of paper is natural and intuitive, and represents a way of organizing the world that appears to be culturally universal (Blaut et al. 2003).

However, the dominance of the use of vision can lead to both cognitive mismatches (if a user is not trained in those conventions, or if the conventions are context-dependent) and constraints on the types of data that can be represented. Spatial descriptions that can be measured and quantified, for example, are privileged on a map, relative to a nonspatial story passed down through generations, that is also keenly important in the understanding of a place. In HCI, this has been called “blinkering” (after the “blinkers” on a racehorse that inhibit its view to only what is straight ahead (Nelson 1990)).

Additionally, vision is not always the most appropriate sense for representation of the use or the user of the map. Of course, those that have various forms of vision impairment must use alternate modalities to understand information but, perhaps less obviously, those whose eyes are engaged elsewhere – like soldiers on the battlefield, surgeons in the operating room, or even drivers of a car – would likely benefit from a “mapping” that allows for communication using modes other than vision.

Nontraditional modalities for cartographic interfaces

Humans naturally use multiple senses, or modalities, to communicate with each other and interact with the world. As we communicate, we speak, listen, watch our audience, often gesticulate, raise our eyebrows and scrunch our
noses, supplement our speech with sketches, and so on. When we observe phenomena in our environment to make sense of the world, we make connections among our senses to recognize or learn (looks like a duck, walks like a duck, quacks like a duck, etc.).

However, many of the aspects of the environment that are traditionally difficult to map are those that are experienced with senses other than vision. Most maps, for example, are relatively devoid of emotion, and it has always been challenging to represent (or emphasize) place over space on maps. Emotion is influenced by not only sights but also sounds, tastes, smells, and (literally) feelings, and a natural “mapping” of emotion and place should probably include at least some of these alternate modalities. Fortunately, the technology associated with digital mapping has opened possibilities for doing just this.

Interface input modalities

Map interface modalities (as with most types of interfaces) can be separated into different general kinds of uses: input and output. In terms of paper maps, output modalities consist of the map itself and the visual encoding of information of the marks on the page, whereas input modalities of paper maps are essentially nonexistent. As maps have become more interactive, it is these input modalities that have been borrowed and adapted from the realm of human–computer interaction, and that represent a great deal of innovation in the representation of geographic information.

Most digital maps receive input from the manipulation of a computer mouse and, to a lesser extent, through keyboard entry (typing in an address in an online map, for example). The operation of a computer mouse is a remnant of the popular graphical user interfaces (GUIs) of the 1980s. Prior to the mouse, one of the most notable experiments in user interface design was, coincidentally, demonstrated using a map. The “Put–that–there” system, developed by a team at MIT (Bolt 1980), included an entire “media room” with a wall-sized color map panel. The user would sit in a chair and point to the panel and issue verbal commands, to which the computer would respond. This is an early example of both gestural and voice modalities for interface input. Using a limited set of recognized active verbs (e.g., “create,” “move,” “put”) and pronouns (“that,” “this”) along with natural pointing gestures, the user could interact with shapes on a base map.

As there are multiple modalities (speech and gesture) working in concert with one another in this scenario, this style of interface is known as multimodal (Oviatt 2002). Speech input requires software processing of natural language, and for the computer to recognize and parse the semantic meaning of the commands. As we know when speaking on the telephone, this task is made difficult without communication clues given by visual cues of gesturing, gaze, and other natural behavior. Because, as is true in natural language, context and semantic clues to human verbal communication can be better understood when multiple modes are utilized, multimodal interfaces can accept input from parallel streams (for example, both voice and gesture).

Multimodal fusion like this was demonstrated in a crisis-management mapping scenario through the use of an environment called DAVE_G (Sharma et al. 2003). In that project, designers created a system in which multiple individuals would collaborate with a large-scale display that would respond intuitively and quickly to gesturing and verbal commands. The use of multimodal interactions here was considered important because the gesture and voice recognition supported the time-sensitive
and multi-user nature of the problem-solving scenarios.

Input to visualization systems through gestures and voice requires hardware as well as software to map the inputs to virtual interactions. Microphones (and headsets) are the most obvious of these input devices that would supplement or supplant the computer mouse for voice commands. For gestural input, there are several varieties. The wired (or “data”) glove can capture physical data such as finger bending or motion tracking. These are, not surprisingly, used in virtual reality systems where the user is immersed in a virtual environment (and not seated at a desk). Head and motion tracking devices can also use a video camera (mounted near the actual visual display) to observe infrared light emitted by markers worn by the user. These have become very popular in the gaming community. Other game system devices can also be used for gesture input, such as “Wiimotes” or joysticks that combine kinetic input with optical sensors and accelerometers. The Wiimote, for example, can be used to control applications with keyboard shortcuts, like Google Earth, quite easily.

Appearing at about the same time as gesture-based game controllers and systems were mobile devices such as smartphones and tablet computers that represented the first significant departure in about thirty years from the mouse point-and-click computer interface in wide use. Early mobile devices such as the Palm required a stylus to substitute for the mouse, with the point-and-click interaction very similar to that of a desktop computer. Now, the mobile device accepts gestural input such as swiping, shaking, rotating, pinching, and spreading. Additionally, the devices feature speech recognition and interpretation (through Siri, for example) and, in the case of wearable devices like Google Glass, require speech interaction. The rules for these interactions are still being developed (and have led HCI designers to call for standardization (Norman and Nielsen 2010)) but these devices surely will continue to demonstrate the utility of alternative modes of interface input and manipulation.

Interface output modalities

Prior to the advent of digital technology, nonvisual forms of geographic representation primarily included auditory abstractions, such as oral storytelling through language and song. Composers of so-called “tone poem” music popular in the nineteenth century intentionally represented place through nonverbal music; Bedrich Smetana’s “Die Moldau/Ma Vlast,” for example, evokes a soundscape of folk music and whitewater rapids to represent a river in Bohemia.

In addition to representations of geography that are designed for the ears, haptic representations, which rely on the sense of touch, have been developed, primarily for the visually impaired. So-called “tactile diagrams” are vacuum-formed raised surfaces similar to (and in some cases using) Braille. A significant impetus for nonvisual means of cartographic presentation is accessibility for the visually impaired; there is a long-standing subgroup of the International Cartographic Association (ICA) on Maps and Graphics for the Blind and Partially Sighted. Rice et al. (2005) and Griffin (2001) provide comprehensive reviews of haptic maps; efforts to represent geographic information to the disabled highlight the importance of alternative interface modalities for diverse users of geoinformation.

Cartographers have long used (and debated the use of) the “visual variables” presented by Bertin (1983) to guide the encoding of information to graphic marks on a map. Griffin (2001) draws parallels for encoding types of geographic data to
haptic variables, first classifying haptic sensations into “tactile,” perceived when skin comes in contact with an object, and “kinesthetic,” which are those stimulated by motion of the skin or body. A virtualized map displaying weather could feel physically colder in colder regions, and this would be represented as the tactile variable of temperature. Alternatively, a map of slope could encode steepness by the relative resistance provided by a computer mouse as it is (kinesthetically) moved around the map.

Computers can afford the opportunity to represent data sonically as well as visually. This is probably the most common alternate modality employed by map-makers, primarily because the hardware for creating sound (loudspeakers and an audio card) is now ubiquitous. Krygier (1994) categorized abstract sounds that might be produced by a computer into their appropriate cartographic use. The abstract sound variables, as he called them, included the basic building blocks of music: pitch, loudness, timbre, attack and decay, and register, among others. Such musical primitives are shown to adequately represent geographic variables of various levels of measurement (timbre for nominal data, pitch and loudness for ratio data, for example) and, as such, they are useful for bivariate or multivariate maps, as Fisher (1994) used them in representing uncertainty with sound. Rice et al. (2005) used sound in connection with haptic devices to create a “haptic soundscape” environment for the visually impaired.

Music – comprising higher-order structuring of the primitives Kyrgier described – holds potential for conveying emotion (as in the score of a film) as well as quantitative information. As such, sound and other modalities may be appropriate for conveying aspects of geographic reality that visual displays struggle to represent. The dynamic maps of Caquard et al. (2008) explore the use of such cinematic devices, supplementing maps with nonabstract sounds such as political speeches, airplane engines, and natural sounds. Sound plays a significant role in their studies, as spoken language, scores, songs, and sound effects play significant roles in film.

These alternative and multiple modalities are of course important for accessibility of geographic information to diverse users, such as those with visual impairment. But alternate modalities – in particular sound – can be important and effective for diverse uses of maps as well. User interface designers have, over the last two decades or so, explored the use of sound to make systems useful to individuals whose “eyes and hands are busy.” GPS receivers – which are being used in cars with the same purpose as road maps and atlases in the past – now feature verbal turn-by-turn instructions for easily distracted drivers.

Multimodal interfaces open new cognitive channels for humans to readily absorb information, as we do in our everyday environments. Digital maps are not unique in their potential for improvement using representation and interaction methods that more closely replicate the way we communicate with each other and understand the world. However, because cartographers have for centuries been tasked with abstracting complex information into useable and useful forms, they both contribute to and are informed by research and innovation in human–computer interaction, including the use of nontraditional interface modalities.

SEE ALSO: Representation; Virtual geographic environments; Visual variables; Visualization; Web-mapping services

References


**INTERFACE MODALITIES**


**Further reading**

The Intergovernmental Panel on Climate Change (IPCC) was officially constituted during its first meeting in November 1988, convened in Geneva by its two sponsoring bodies, the World Meteorological Organization (WMO) and the United Nations Environment Programme (UNEP). It received its United Nations mandate through a Special Resolution of the 70th Plenary Meeting of the UN General Assembly, passed on December 6, 1988. The IPCC has published five full-scale assessments to date: in 1990 (the First Assessment Report: AR1), 1996 (AR2), 2001 (AR3), 2007 (AR4), and 2014 (AR5).

On each occasion the assessment of scientific knowledge has been organized around three Working Groups (WGs), which have retained broadly similar scope: the physical science of the climate system (WG1); the impacts of climate change on ecosystems and society and options for adaptation (WG2); and the economics and technological options for mitigation (WG3). In addition to these periodic assessments, nine “special reports” on more narrowly defined issues have been produced by the IPCC, the most recent being in 2012 on managing the risks of extreme events and disasters. IPCC reports have been influential for national and international policy development and in 2007 the Panel received, jointly, the Nobel Peace Prize for “its effort to build up and disseminate greater knowledge about man-made climate change and to lay the foundations for the measures that are needed to counteract such change.” During 2010 the governance of the IPCC came under significant public and political scrutiny owing to a small number of errors in AR4 (WG2), most notably about the future melt rate of Himalayan glaciers. The United Nations secretary-general commissioned the Inter-Academy Council (IAC) to conduct an independent investigation into the errors, an investigation which made wide-ranging recommendations for adjusting the governance and assessment procedures of the IPCC (IAC 2010).

Origins and mandate

The origins of the IPCC can be traced back to the growing number of scientific and political discussions in the 1980s about the extent of human influence on the climate system (Agrawala 1998). In 1985, an international meeting of climate and environmental scientists was convened by WMO/UNEP/ICSU (International Council of Scientific Unions) at Villach in Austria, which presented the case for more comprehensive international knowledge assessments of climate change. Rather than concede too much influence in this process to environmental nongovernmental organizations (NGOs), the US government in particular argued for
such assessments to operate under an intergovernmental mandate. Several events in the late 1980s converged to catalyze the formation of the IPCC as a new institution for global knowledge assessment about climate change. These included the growing capabilities of climate modeling and Earth system science through which climate was conceived as a global system, the rise of global environmental politics throughout the 1980s, and the success of the 1987 Montreal Protocol on Substances that Deplete the Ozone Layer. The end of the Cold War in 1989/1990 was also influential in spurring a new optimistic mood of political internationalism.

The IPCC was officially constituted during its first meeting in Geneva in November 1988. The UN Mandate to the IPCC was to “coordinate scientific assessments of the magnitude, timing and potential environmental and socio-economic impact of climate change and realistic response strategies.” Scientific understanding of climate change was to embrace both natural and human causes of change, but to distinguish between their effects where possible. It was to accomplish this task by recruiting international experts who would assess and synthesize peer reviewed literature, generating knowledge which would fall short of offering policy recommendations or advice. The distinction between science and policy thus implied has been an ongoing source of tension and reflection for the IPCC.

**Governance and learning**

The IPCC has evolved its own rules of governance and its own protocols for knowledge assessment, partly in response to external events and criticisms. The Panel is governed by a Bureau of around 30 officials and experts, elected by the IPCC Plenary in which each member state of the United Nations has a vote. The Plenary also elects the overall chair (and a co-chair) of the IPCC and the chairs (and co-chairs) of the three Working Groups. To date there have been four IPCC chairs: Bert Bolin (a Swedish biogeochemist who chaired AR1 and AR2), Robert Watson (a British atmospheric chemist who chaired AR3), Rajendra Pachauri (an Indian economist who chaired AR4 and AR5), and Hoesung Lee (a South Korean economist who is chairing AR6). In association with chapter lead authors, the IPCC Plenary also approves, line-by-line, the Summary for Policymakers of each WG Report. This approval process of the headline messages of each assessment cycle affords the IPCC’s conclusions the distinctive status of joint ownership between governments and scientists. The Bureau has responsibility for agreeing to the Working Group structures and the outline scope of each chapter. Together with the WG chairs and co-chairs, the Bureau filters nominations for lead authors and agrees to the final list of experts to be invited to convene and coordinate each chapter writing team.

The formal knowledge assessment work of the IPCC is governed by strict rules of procedure. These have undergone three major revisions, in 1993 (after AR1), in 1999 (after AR2), and again in 2010 (after AR4). The changes adopted in 1999 were made to accommodate more diverse regional sources of knowledge for the regionally focused chapters of WG2, but also were partly a response to a controversy which erupted in 1996 around the wording of Chapter 8 (“Detection of climate change and attribution of causes”) in WG1 of AR2 (see Lahsen 1999). The changes made in 1999 included: the introduction of independent review editors to oversee the peer review process; the adoption of formal rules for the making of an IPCC Synthesis Report (a single, shorter report drawn from across the three Working Groups); and the clarification of
circumstances under which non-peer reviewed (“grey”) literature would be acceptable as a primary source of knowledge.

It was the use of some of these latter sources in the WG2 Report of AR4 which provoked another major controversy and which led to a further round of procedural and governance revisions. In January 2010, shortly after the inconclusive meeting of the United Nations Framework Convention on Climate Change (UNFCCC) at Copenhagen (COP15) and after the release of controversial emails originating from the University of East Anglia, it emerged that a statement in the 2007 Report about the imminent melting of Himalayan glaciers was erroneous. This was traced back to a non-peer reviewed source and to poor review editing. The IPCC had no mechanism in place for the correction of errors and the IPCC leadership handled poorly the ensuing crisis of authority. The entire AR4 Report was placed under critical scrutiny and other minor errors and ambiguities were discovered. The United Nations secretary-general and the chair of the IPCC invited the IAC, a multinational organization representing the world’s science academies, to conduct an independent review of the IPCC processes and procedures.

These recommendations included, inter alia, establishing an Executive Committee and director; making the selection of experts a more transparent and rigorous process; clarifying the use of non-peer reviewed material; strengthening the role of review editors; and developing an effective communication strategy. The episode was a test of the leadership and transparency of the IPCC and of its peer review system for making reliable public knowledge about climate change. The credibility of many climate change policies continues to rely on the scientific authority of the IPCC, which in turn is grounded as much in trust as in truth (Beck 2011).

**Disciplinary and geographical expertise**

The mandate given to the IPCC by the United Nations was wide-ranging: to assess global knowledge about climate change and its potential implications for society. The scope and significance of this mandate means that who is selected to serve as an “expert” on the Panel becomes a matter of public interest: Which knowledge disciplines are favored by the IPCC? What is the geographical representation of the Panel’s experts?

Reflecting its origins in the meteorological and environmental sciences, the IPCC has been much more heavily influenced by natural science disciplines than by the social sciences or the humanities. In AR3 (2001) for example, only about 12% of the 8500 citations to peer reviewed journal articles were from social science literature (and of these about one-third were from economics). Although there was a modest broadening of disciplinary experts in AR5, the IPCC remains heavily skewed towards assessing knowledge of the physical climate system and ecological impacts, rather than knowledge about human beliefs and values with respect to causes, consequences, and responses to climate change. This has been referred to as the “IPCC fallacy” and social scientists have sought to understand the epistemological, institutional, and political reasons why the social science and humanities disciplines have been marginalized by the IPCC. Yearley (2009) argues that climate science is currently constructed through assigning the (interpretative) social sciences a specific role – a subsidiary one. “The institutional assumption of the IPCC is that the most relevant social science is economics” (401), thus marginalizing knowledge about climate change which emerges from disciplines such as anthropology, social psychology, communication science, philosophy, and history.
The IPCC has also faced criticism that its expert authors are drawn disproportionately (about 75%) from North America and Europe and disproportionately (about 60%) from Anglophone cultures. Up to and including AR4, 45% of the 193 nations of the world had never had an expert author selected for any IPCC author team in one of the four major assessments (Ho-Lem, Zerriffi, and Kandlikar 2011). The consequences of this uneven geography of IPCC expertise are significant, politically as much as epistemically. The IPCC experts who have constructed development scenarios, who have differentiated the value of human lives, and who have conceived of tropical forests as “empty spaces” for carbon sequestration – all examples of controversies from IPCC assessments – have therefore been drawn from a relatively small cohort of nations. In 1991, the then chairman of the IPCC, Bert Bolin, remarked: “Right now, many countries, especially developing countries, simply do not trust assessments in which their scientists and policymakers have not participated. Don’t you think credibility demands global representation?” (cited in Schneider 1991, 25).

Consensus, risk, and uncertainty

From its first assessment report onward the IPCC has been explicit about seeking to generate a “scientific consensus” about how climate is changing and especially about the extent of human influence on the climate system. Consensus-building in fact serves several different goals. For example, seeking consensus can be as much about building a community identity – what scholars refer to as an “epistemic community” – as it is about seeking the “truth.” Nevertheless, the IPCC’s drive for consensual knowledge claims has been a source of both strength and vulnerability. On the one hand, IPCC consensus-making has been largely driven by the desire to communicate climate science coherently to a wide spectrum of policy users – to construct authoritative globally agreed knowledge for use by governments in their development of policy. On the other hand, without a careful explanation of what consensus means, and without a full and careful exploration of uncertainties and disagreements among experts, the IPCC is vulnerable to criticisms of premature foreclosure of scientific discovery or of silencing dissent.

Consensus-making can also lead to the criticism of the IPCC being too conservative and of encouraging “scientific reticence.” Some have argued, for example, that the IPCC AR4 has been too conservative in reaching its consensus estimates about the range of future possible sea level rise. Given the broad acceptance that humans do now exert a significant influence on the climate system, it may be that governments would find it more useful were the IPCC to focus more on areas of disagreement and uncertainty than on consensus.

Within the IPCC, consensus-making is often an exercise in collective judgment drawing upon subjective (or Bayesian) probabilities in areas of uncertain or incomplete knowledge. To increase the rigor with which the IPCC reached and communicated its consensus, official uncertainty guidelines were introduced in 2000 as part of AR3. Three uncertainty schemes were adopted, all of which relied upon Bayesian forms of reasoning and judgment: a likelihood scale (from “virtually certain” to “exceptionally unlikely”), a confidence scale (from “very low” to “very high”), and a level-of-understanding matrix (“level of agreement” versus “amount of evidence”). The aim of these guidelines was to introduce a consistent methodology for uncertainty communication across the different IPCC Working Groups, yet the guidelines have
been difficult to police. Different Working Groups, familiar and comfortable with different epistemic traditions, construct and communicate uncertainty in different ways. This creates possibilities for confusion and misunderstanding, not just among policymakers and the public, but among IPCC experts themselves. One of the recommendations made by the IAC in 2010 was for the IPCC to be much clearer in explaining the processes whereby it attaches different levels of confidence to its headline conclusions.

Equally important is how the uncertainties reflected in the IPCC’s knowledge claims are understood in nonexpert public and political settings. Expressions such as “likely” or “extremely likely,” or claims of “high” or “low confidence” in certain knowledge statements, mean something very specific in expert IPCC terms. Yet the vernacular interpretation of such phrasing varies across audiences and cultures. Studies have shown how different forms of uncertainty communication used by the IPCC – for example uncertainties deriving from model differences versus disagreements between experts – alter the level of trust among lay audiences of the veracity of the IPCC’s claims. Work has also shown the effect of cultural differences on message reception; for example, Chinese and British citizens attach different intuitive probabilities to expressions such as “likely” or “unlikely.”

Wider influence and impact

There is little doubt that the IPCC has had a significant influence on climate change knowledge, on public discourse about climate change, and on climate policy development. On the other hand, scholars disagree about the exact reasons for this influence and whether it has always been for the best. Knowledge that is claimed by its producers to have universal authority is received and interpreted very differently in different political and cultural settings (Jasanoff 2010). Revealing the local and situated characteristics of climate change knowledge thus becomes central for understanding both the acceptance and the resistance that is shown towards the knowledge claims of the IPCC. There is also concern that the sheer size of full IPCC assessments (over 5000 pages in AR5) makes its product too unwieldy and cumbersome; an alternative arrangement of much shorter and more regular assessments is favored by some.

The IPCC has helped to fashion and consolidate a global climate change epistemic community. Thousands of scientists from across a hundred countries have served as lead authors and tens of thousands of experts from most countries in the world have participated in the reviewing process. The impact of this IPCC epistemic community has been examined from a number of different regional perspectives – for example, Brazil, France, China – with some concluding that IPCC knowledge “travels well.” Others, however, have identified some of the problems with the circulation of IPCC knowledge and warn against the “epistemological hegemony” of the IPCC. There remains considerable detailed empirical work to be done around the world about exactly where, how, and why the practices of climate change knowledge production developed by the IPCC have altered scientific practice, in the biogeophysical sciences and social sciences as well as in the design of interdisciplinary research.

The IPCC has also gained visibility in public spaces as the authoritative voice of climate change knowledge – “the privileged speaker and discursive leader” – a visibility enhanced through being awarded the 2007 Nobel Peace Prize. The “boundary work” between science and policy that the IPCC performs has also
legitimized the scientific vocabulary that businesses and NGOs have been able to deploy in public spaces. With regard to the impact of the IPCC on policy development, opinions become more polarized. The IPCC has operated under the rubric of being “policy relevant but not policy prescriptive” (i.e., policy neutral). Its work has underpinned the adoption by many political actors of “two degrees” of warming as being the definition of dangerous climate change and therefore the ultimate goal of international climate policy. On the other hand, Pielke (2007) argues that the IPCC has failed in its role as an “honest broker” and has moved towards being an “issue advocate,” or even on some occasions a “stealth issue advocate.”

The authority, credibility, and reflexivity of the IPCC have become matters of much debate. Some point to the successive adjustments of the IPCC governance system as a sign of institutional learning. They point to how the IPCC has become a benchmark and role model for newer forms of global knowledge assessment in domains such as biodiversity and desertification. Other critics, however, have pointed out the need to remain vigilant of the ways in which international knowledge institutions like the IPCC gain power and influence in international policy deliberations, while being less than open or democratically accountable in their own modes of operation. Writing in 1997 about the future status and impact of the IPCC, Shackley concluded his commentary by noting: “Of particular concern is whether the IPCC can make its knowledge more socially relevant and trusted by bridging the gulf which exists between scientific experts and on-the-ground decision-makers and members of the public” (Shackley 1997, 174). This question of the IPCC – relating to participation, trust, governance, and policy advocacy – remains as important today as it did in the late 1990s.

SEE ALSO: Climate adaptation/mitigation; Environmental knowledges and expertise; Environmental science and society; Geopolitics of the environment; Global climate change; Sustainability science; Wicked problems

References

Further reading


Intermediaries facilitate the transmission of services between two parties, acting as a conduit between supplier and consumer. They play a dynamic role in shaping the geography and regulation of the economies and labor markets in which they are embedded. Product market intermediaries include actors such as accounting or advertising firms, mortgage or insurance brokers, or financial or legal consultants that offer producer services. Labor market intermediaries (LMIs), the focus of this entry, are similar to product market intermediaries in their function and geography, but because their “product” is labor they play a unique role in mediating the relationship between capital and labor.

Labor market intermediaries facilitate and regulate the matching of workers to employers, acting as brokers that influence and sometimes control the negotiation process between the two parties. LMIs comprise a diverse array of entities including labor recruiters and temporary staffing agencies, executive search firms or headhunters, public employment offices, labor unions, state regulatory bodies, criminal records providers, some training/educational institutions, and online job boards or search engines. They can be public or private, require voluntary or compulsory membership, and fall along a spectrum from formal and established to informal and less “legitimate” operations. Research has shown that, while private, for-profit intermediaries increasingly drive market development and shape employment norms, public and membership-based intermediaries tend to arise more often in response to changing employment conditions.

Intermediaries redress and capitalize on information asymmetries. Where there is a cultural, linguistic, or significant geographic distance between employers and workers, LMIs use the competitive advantage of their information about and access to employers or labor pools to reduce the friction and costs associated with job matching. For employers, they provide a supply of prescreened, just-in-time workers. For job-seekers, they offer immediate, often short-term, employment with no investment in a job search. Beyond job matching, LMIs shape labor market dynamics in fundamental ways by mobilizing workers and impacting labor agency, shaping compensation levels, mitigating and displacing risk (generally onto workers), and contributing to production and social network building (see Benner, Leete, and Pastor 2007). The triangular relationship between intermediaries and the employers or workers who contract them is not merely aimed at cost reduction but is also shaped by, and in turns shapes, nonmarket social relations.

The role of intermediaries is a relatively under-studied area in economic and labor geographies. However, since the 1990s scholars have begun to place greater emphasis on labor as an agent actively creating and shaping economic geographies – and concomitantly have begun exploring the impact of LMIs on labor mobility and agency. Furthermore, since the mid-1970s, economic restructuring and intensified global economic
integration have created the conditions for significant expansion of LMIs, which have thus received increased attention from geographers in recent years.

Economic restructuring and labor flexibility

The recent expansion of labor market intermediaries should be considered within the context of economic restructuring and globalization. The period since the 1970s has been characterized by massive changes in labor markets and employment relations. Neoliberal deregulation and public and corporate policies promoting labor flexibility have made labor markets increasingly competitive and volatile in recent decades. As the demand for more flexible, contractual, individualized, and temporary labor has increased, intermediaries have emerged as a significant industry.

Most economic geography research on the industry has focused on temporary employment agencies, and particularly those servicing low-skilled industry sectors, where the largest growth has been. Though initially associated closely with clerical and sales work, these agencies have also become important employers for blue-collar workers in production and transportation. The temporary employment industry accounted for an increasingly large share of job creation in the 1990s, especially in the United States, the United Kingdom, Canada, countries in continental Europe, and Japan. Scholars tracing the evolution of temporary staffing agencies in many “developed” countries have noted a shift as public institutions that traditionally performed job-matching functions have been replaced by private, for-profit agents. In fact, the International Labor Organization now recognizes the temporary staffing industry as a necessary fixture in the twenty-first-century labor market.

Research has highlighted how the temporary staffing industry not only benefits from an increased demand for temporary employment but is also a promoter and orchestrator of it. By delivering just-in-time labor on a contingent basis, the industry reinforces the demand for it. As Peck and Theodore (2001) note, temporary employment agencies “actively shape the growth in contingent labor through their role in labor-market intermediation, by selling employers the cost-cutting and labor-controlling virtues of workforce flexibility while mobilizing contingent workers and brokering connections to employers” (p. 477). LMIs are active agents in marketing their labor pools, and some scholars have argued that the availability of workers through temporary employment agencies in fact encourages employers to pursue high turnover, low skill-investment human resources strategies. Though LMIs often represent themselves as neutral intermediators, researchers argue that their very presence contributes to the formalization of contingent work as a feature of flexible labor markets.

Labor market regulation and segmentation

Temporary employment agencies typically operate on extremely tight profit margins and absorb much of the risk (on behalf of employers) associated with fluctuations in labor demand and the uncertainties associated with hiring and managing a more flexible and contingent workforce. While there are organizational strategies for mitigating such risk (such as geographical expansion or diversification of services), the most common strategy is to pass it on to workers. In so doing, LMIs play an increasingly significant role in the regulation and segmentation of labor markets.

Some scholars studying LMIs discuss labor flexibility as a post-Fordist strategy for increasing
capital accumulation and labor control. There are many reasons companies may choose to hire an intermediary – for cost reduction, as part of internal restructuring, or because of changes in the nature of production or the regulatory environment. Regardless of the reason, though, by contracting LMIs employers externalize and minimize their costs, risks, and responsibilities (including obligations to workers). Researchers have shown that outsourcing to temporary employment agencies directly regulates workers by subjecting them to a dual layer of management – to overlapping systems of sanctions imposed by both the intermediary and the contracting employer.

Temporary employment agencies also indirectly regulate labor through implied worker substitution. While they are able to mobilize unemployed and underemployed sectors of the population, more often than not it has been for precarious positions with lower wages and fewer benefits. The absence of regular, secure jobs has rendered many workers dependent on temporary employment and the LMIs that provide it. This has led to intensified competition within the contingent labor sector, especially in low-skilled occupational sectors, where workers are often viewed as substitutable. LMI practices that contribute to the individualization and precarity of work also inhibit collective action and weaken trade unions. Anecdotal evidence suggests that some firms use temporary help agencies to illegally screen out potential union organizers. In addition to engendering a dependent, highly substitutable, and reliably contingent and compliant labor supply, the possibility of employee substitution implicitly threatens relatively more secure “core” labor as well.

LMIs such as temporary employment agencies also shape assumptions about which sectors of the labor pool constitute the contingent workforce. They act as gatekeepers, regulating entry to the labor market through worker recruitment, screening, and placement. Research has shown that these activities have consequences for labor market segmentation. Employment agencies have a financial incentive to comply with stated, implicit, and anticipated employer demands and preferences, and as such they tend to reproduce and entrench discriminatory hiring practices. The activities of LMIs route workers into particular kinds of work – in terms of wage rates, industry sectors, and occupations (for example, women into clerical work). This in turn naturalizes and institutionalizes gendered and racialized ideas about who belongs in what kind of work.

In fact, some LMIs differentiate themselves by associating with particular segments of the labor market, defining their market niche in terms of an industry sector or the ethnic makeup of their labor pool. Geographers have also noted that the location of an agency’s primary labor pool – especially at the local, urban scale – is often seen as implicit coding for certain social characteristics, such as socioeconomic status, presumed employability, and race. Peck and Theodore (2001), for example, found that employment agencies specializing in ethnic segments of the labor market, tied to particular neighborhoods in Chicago, reinforced historic race-structured hiring regimes that effectively excluded many African American workers, even from low-wage temporary work. The presence and practices of LMIs in many cases contributes to intensified patterns of exclusion and inequalities, both socially and spatially, in already segmented labor markets.

Globalization, labor mobility and agency

As transnational companies have expanded as part of global economic integration, producer services have followed, and LMIs have expanded
INTERMEDIARIES

and diversified. The viability of LMIs is to some extent predicated on a spatial mismatch between employers and jobseekers, one that has only intensified through globalization. While the demand for more flexible labor and the increased presence of LMIs has meant increased precarity and vulnerability for some, it has resulted in enhanced mobility for others. Knowledge-based economies compete for increasingly mobile global talent. LMIs such as headhunters and executive search firms can directly contribute to career and salary advancement, particularly for workers in specialized or high-skilled, elite labor markets. Some larger employment firms differentiate themselves within the LMI industry by catering specifically to these workers.

Geographers have noted a polarization within the industry itself in recent years, made evident, for example, in the locational strategies of major agencies. The spatial agglomeration of agencies can be indicative of their echelon in the industry. Whereas employment agencies placing low-skilled or “lower-end” labor will often co-locate with their primary labor pools, research has found that those catering to more high-skilled, elite workers or companies in search of such workers tend to locate themselves in the same corporate milieu as their clients – distancing themselves physically and perceptually from less established or “shadier” parts of the industry.

The LMI industry has not only pushed upward into new occupational sectors but also outward into new markets. Globalization of the industry has occurred for two primary reasons. For one, more countries have opened up their economies and deregulated their labor markets, which has created the conditions for LMIs to expand. Second, scholars point out that the very presence of established LMIs acts as an indicator to the multinational companies that might contract them that an economy is stable and worthy of investment. Thus the industry both results from and contributes to the ongoing restructuring and (neo)liberalization of labor markets and the wider globalization process. Research has documented how European and US-based multinational LMIs, in particular, have been active agents in the liberalization of labor markets in countries such as Italy, Spain, Germany, and Japan and are increasingly capitalizing on emerging markets in places such as Mexico, China, India, and Brazil (Ward 2004; Peck, Theodore, and Ward 2005). As these agencies expand transnationally, their staff, offices, and labor pools are dispersed geographically and they are run in a decentralized fashion. Local markets limit the extent to which agencies can simply implement business models developed in their home markets. While the industry itself may be globalizing, the information and services offered by LMIs remain highly territorially embedded and local in nature.

As LMIs have expanded transnationally, so too have their services, and profit-seeking intermediaries increasingly facilitate and organize the transnational movement of workers. As migratory movements become established, the need for specialized information and services arises and a “migration industry” emerges, especially in the context of weak social networks. This industry encompasses a range of actors including LMIs such as labor recruiters, brokers, and smugglers. LMIs specializing in migration help solidify transnational migrant networks, organize migrant emigration and subsequent return, and facilitate job placement and local integration. Some are formal businesses with established reputations, but there are a myriad of other intermediaries who operate informally, and the industry is frequently characterized by fraudulent, exploitative, and illegal practices.

Most migrant workers who find work through intermediaries use informal agents, and are thus more likely to be charged exorbitant
recruitment-related fees and to experience coercion and forced labor. Indeed, for-profit employment agencies are in a position to exploit precisely the information asymmetry they are hired to resolve – to use their informational advantage to exploit rather than assist their clients. As a consequence, governments in both sending and receiving countries are increasingly instituting measures to regulate the migration industry, either by establishing highly regulated licensing and monitoring regimes for the private recruitment industry or by offering alternative forms of public assistance.

Some countries are becoming more directly involved in the management of labor migration – to the point that some scholars argue that states themselves are acting as labor brokers. The Philippines, for example, actively encourages and regulates the emigration of its nationals as foreign workers to reap the benefits of remittances and transnational connections abroad. Though criticized by some, this approach is often held as a model for other sending countries that wish to develop their overseas employment potential. States that are hesitant to become involved in the organization and regulation of labor migration may indirectly encourage the proliferation of commercial recruitment. In the absence of governmental regulation, employers or workers must rely on the integrity of the LMI they contract, and, more often than not, vulnerable migrant workers are the ones who suffer abuses at the hands of unscrupulous intermediaries.

Intermediaries are playing an increasingly significant role in assisting migrants crossing national borders, often in the context of restrictive immigration policy regimes, and this has important implications for labor mobility and agency. LMIs figure as gatekeepers, granting or limiting access to workers and routing them to particular destinations and positions based on factors such as their race or ethnicity, gender, language ability, and nationality. They offer some workers the flexibility and enhanced mobility required to compete globally, and they can facilitate opportunities that improve the wellbeing of workers with few assets other than their labor. At the same time, a reliance on intermediaries may introduce additional barriers for some workers, limiting their access to global labor markets and constraining their mobility and employment opportunities. While the impact of LMI use on individual workers may be varied, however, research has shown that their effect on labor organizing and trade unions is undoubtedly deleterious. Temporary employment arrangements inhibit collective action, and at the transnational scale workers are further fragmented, arguably making political efforts to organize and build solidarity more difficult. In this vein, some geographers have called for more in-depth research on the impact of LMIs on labor agency, with specific attention to how transnational connections might in fact open up new opportunities for labor organizing and worker rights and protection.

Though intermediaries represent a growing institutional presence in the organization of labor markets at a range of geographic scales, research on them is still limited. The currently fragmented and partial view of intermediaries results in part from challenges associated with their oftentimes elusive and transitory nature, and future studies may need to adopt innovative data-gathering and methodological approaches to capture a more comprehensive picture. Within economic and labor geographies, attention has tended to focus on LMIs in North America and Western Europe, and on temporary staffing agencies in particular. In migration research, the commercialization of migration and involvement of LMIs in transnational labor movements remains one of the most understudied topics. Future studies
on intermediaries should draw on both these literatures, and might examine the role of LMIs in the restructuring of labor markets, in the changing nature of employment relations, and in reinforcing or challenging existing inequalities of labor market access. As a topic of study, the role of labor market intermediaries is not just of theoretical import, but also has significant implications both for policymaking and for the development of new and alternative strategies of labor organization.

SEE ALSO: Global production networks; Labor geography; Labor market segmentation; Labor migration; Neoliberalism; Precarious work; Regulation/deregulation

References


Further reading


International: Aljografiyah fi Duall Majlis Altaaon Alkhleejy (Geographical Society Gulf Cooperation Council)

Founded: 2000  
Location of headquarters: Riyadh, Saudi Arabia  
Website: http://gccgs.org.sa  
Membership: 400 (as of 2015)  
President: Obaid S. AlOtaibi  
Contact: obaidal@hotmail.com, obaidal58@gmail.com

Description and purpose

The society furthers geographical studies, organizes public activities to develop theoretical and applied knowledge, and initiate consultation and applied scientific studies for public and private sectors. It is a scientific and educational nonprofit association that aims to promote cooperation and understanding among its members, working with geographers from Gulf Arab countries, and associates from other countries.

To achieve these goals the society establishes beneficial partnerships with community-based organizations active in the Gulf society as well as with government institutions and non-profit organizations, the private sector, and geological survey and environmental preservation bodies, among others.

Journals or major publication series

Almajalah Aljografiyah Alkhleejiah (The Gulf Geographical Journal)

Current activities or projects

Regular scientific meetings are held – the next (fifth) meeting will be held in Salalah in the Sultanate of Oman in 2016. Scientific trips are organized. Previous trips were to the State of Qatar and Oman. The society publishes The Gulf Geographical Journal as well as a book, The Geography of GCC Countries.

Brief history

Geographical departments have existed in Gulf universities for almost 40 years. The first was created at Kuwait University with additional departments established since in other Gulf universities. Hundreds of students have graduated from these programs. Some graduates work in scientific, governmental, and private institutions; a number hold prominent positions in their countries; and some have become well known for their work, for example, Professor Abdullah Al-Ghunaim from Kuwait. One of the most important goals of the association is to bring together specialists and interested geographers in the GCC countries within an institutional framework. It has had a prominent role in the reunification of geographers in scientific meetings of academic and nonacademic activities. The association also realized its objective of writing a book on the “Geographical Cooperation Council (GCC)” from a unity and common fate perspective. The association is involved with geography departments’ students in universities by introducing and training them on academic and social activities as well as participating in academic and cultural cooperation agreements with the Saudi Commission for Tourism and Antiquities.

Submitted by Obaid S. AlOtaibi
International: Association of Geographical Societies in Europe (EUGEO)

Founded: 1997
Location of headquarters: Utrecht, Netherlands
Website: www.eugeo.eu
Membership: 24 (as of 2014)
President: Henk Ottens
Contact: hfl.ottens@gmail.com

Description and purpose

The aim of EUGEO is to represent its members at European level and to coordinate and initiate joint activities of the members to advance research and education on the geography of Europe and to promote the discipline of geography of Europe. EUGEO has members from 21 countries. EUGEO also functions as a network and a forum for strengthening the position and operations of its member organizations, in particular for activities with a European dimension.

Current activities or projects

EUGEO organizes a biannual congress entitled The Geography of Europe and Geography in Europe. Congresses have been held in Amsterdam (2007), Bratislava (2009), London (2011), and Rome (2013). Coming congresses are planned in Budapest (2015) and Brussels (2017). The association organizes a yearly seminar on “Improving the State of Geography in Europe” in combination with the yearly General Assembly meeting. EUGEO cooperates with EUROGEO and IGU in a roadmap project to strengthen the position of geography education in Europe (and worldwide). EUGEO participates in European projects in which geography education and/or research is a theme.

Brief history

EUGEO was formed on the initiative of the Italian Geographical Society (Società Geographica Italiana, SGI). In 1994 representatives of geographical societies in the European Union gathered in Rome at the headquarters of SGI, Villa Celimontana (the “home of geography”) and decided to establish an association. The idea was to encourage and enhance greater collaboration between the independent scholarly European geographical societies, associations, and institutes, each of them having a community of geographers from a nation or region in a EU member state as their membership.

The first president of EUGEO was Henri Nicolaï of the Société Royale Belge de Géographie (Royal Belgian Geographical Society) and the first secretary-general was Armando Montanari of the Italian Geographical Society. Henri Nicolaï was followed by Christian Vandermotten, also from the Belgian Geographical Society, and Armando Montanari was followed by Rita Gardner of the Royal Geographical Society (with IBG). In December 2009, Henk Ottens of the Royal Dutch Geographical Society was elected president. Massimiliano Tabusi (SGI) became secretary-general in 2012. Christian Vandermotten and Zoltán Kovács of the Hungarian Geographical Society MFT were elected as executive committee members. EUGEO now welcomes societies from all nations in Europe as members.

Submitted by Henk Ottens
International division of labor

Kean Fan Lim  
University of Nottingham, UK

For much of the eighteenth to mid-twentieth centuries, commodity production within the global economy took on distinct geographic patterns. Manufactured goods and producer services were largely built and assembled in industrialized economies such as the United Kingdom, France, and the Netherlands. Many of these products were traded with other industrialized economies and surpluses were sold to nonindustrialized economies, many of them colonies. Raw materials and agricultural products were in turn extracted from these nonindustrialized economies, giving rise to a phenomenon known as the “old” international division of labor. The very notion of international division exemplifies an apriori conception of economic production and consumption as intrinsically transnational. It exemplifies, specifically, an apriori conception of labor power interdependence between different parts of the world.

This interdependence changed during the mid- to late 1970s as firms in industrialized economies embarked on vertical disintegration and relocated production processes – particularly labor-intensive manufacturing – to developing economies. In a highly influential book, The New International Division of Labour, the German social scientists Folker Fröbel, Jürgen Heinrichs, and Otto Kreye (1980) outline key tenets of this process, now widely known as the “new international division of labor” (hereafter NIDL). To Fröbel, Heinrichs, and Kreye (1980, 45), NIDL is a tendency which:

1. undermines the traditional bisection of the world into a few industrialized countries on one hand, and a great majority of developing countries integrated into the world economy solely as raw material producers on the other; and
2. compels the increasing subdivision of manufacturing processes into a number of partial operations at different industrial sites throughout the world.

From a longue durée perspective of world economic history, NIDL is not exactly novel. As the renowned economic historian Andre Gunder Frank (1977, 2096) puts it, the accelerated and profound transformation in the international division of labor (and of production and work processes) during the 1970s re-expresses a cyclical crisis of profit realization in the same way that earlier crises precipitated changes in the production process. However, there were at least three new features that distinguished NIDL from earlier historical changes in the international division of labor:

1. The geographical extent of the “division” was much wider after the mid-1970s, with manufacturing activities spreading to Central and Latin America, East and Southeast Asia, and parts of Africa.
2. The intensity and range of transnational economic activities increased as more multinational corporations (MNCs) were formed in the 1970s compared to a century earlier.
INTERNATIONAL DIVISION OF LABOR

This phenomenon can be largely attributed to improvements in “space-shrinking” transport and communications technologies, which enhanced the coordination of production and distribution processes by corporate headquarters, as well as MNCs’ ability to raise financial capital through expanding capital markets in the industrialized world following the collapse of the Bretton Woods monetary system in the early 1970s (Murray 1971; Aglietta 1982). As Emmanuel (1975, 36) puts it, there existed by the mid-1970s “a relative mobility of [financial] capital, sufficient to give rise to a tendency for the world-wide equalization of the rate of profit, and a relative immobility of labour allowing considerable predetermined disparities in the wage rates of various countries.”

A massive and growing pool of low-skilled or unskilled laborers became available for employment in many parts of the world, in part as an outcome of improving agricultural technologies, in part as a result of weak or failed domestic industrialization. To Harvey (1999/1982, 436), “Beneath all of the nuanced shifts in the international division of labour, in technology and organization and in the distribution of productive forces, lies a basic Marxian proposition: the accumulation of capital is increase of the proletariat” (emphasis original). Put another way, NIDL was correlated positively with the expansion of wage laborers at the global scale.

NIDL was a combination of these three features. Crucially, its emergence was enabled in large part by the “crisis” of Fordism. Fordism refers to a regime of accumulation first implemented in the United States and then expanded in varying degrees to Western Europe, Canada, and Australia. It is characterized by the mass production and consumption of standardized goods, high wages, and relatively fixed currency exchange rates which facilitated international trade. To keep the Fordist regime running, labor unions fought to keep wages high and growing, which would putatively translate to expanded consumption as well as stable fiscal revenues for the state, while the state took responsibility for providing collective goods such as domestic infrastructure, health care, and education and for overseeing stable fiscal and monetary policies. By the mid-1970s, it became apparent that the Fordist regime was becoming increasingly strained as profit rates declined. Large corporations in those economies consequently began exploring production and market opportunities elsewhere around the world. This phenomenon corresponded with FDI-friendly or import substitution policies in “host destinations,” which in turn encouraged offshoring and produced modern-day transnational corporations (TNCs) and global production networks.

Economic geographers were quick to analyze the causes and differentiated geographical impacts of NIDL. Initial research probed these causes by examining firms’ strategies vis-à-vis the Fordist structural crisis in the so-called core group of industrialized economies (see Wallerstein’s (1974) discussion of core–periphery dynamics). Quite what triggered such systemic distress remains moot: the primary issue is whether demand- or supply-side factors, or a combination of both, contributed to the dismantlement of Fordism. To Doreen Massey (1978), market contraction was a major trigger of growing deindustrialization within several major industrial sectors in the United Kingdom. This in turn highlighted the limits of trade and labor market protectionism that undergirded the Fordist system. Once product markets began to open, the onslaught of competition – much of
which emerged from firms within other “core” economies, notably Germany and Japan – caused many domestic firms to explore alternative production methods and strategies to remain in business. Alain Lipietz construes this supply-side factor as a major cause of NIDL:

Unlike the thirties crisis, however, it is essentially due not to the tendency of supply to exceed popular demand, but to the fact that insufficient surplus-value results from a growing mass of invested capital. For capital in general (not, of course, for each individual capitalist), the problem is not so much to find markets as to drive up the rate of exploitation. And that has been the case since the late sixties. (1982, 36; emphasis original)

The underlying issue, however, was more intricate than simple cost reduction or increased exploitation. To Schoenberger (1988, 261), the nature of interfirm competition during the Fordist era planted the seeds for NIDL: “uncontrolled product competition is incompatible with Fordist production techniques,” which were astute at churning out standardized models in huge numbers to capture economies of scale. Similarly,

excessive price competition is incompatible with maintaining the mass-consumption base. In general, the system relies on the maintenance of an orderly oligopoly that will support stability in terms both of products and of prices while competition is channelled into such areas as superficial product differentiation, advertising, and distribution. (Schoenberger 1988, 261)

The effect, as Schoenberger shows, was in fact so significant as to trigger a “cultural crisis of the firm,” wherein an “entire ensemble of material practices, social relations, and ways of thinking, and the material and social assets tied up in them, were invalidated” (1997, 224; emphasis added). From this perspective, NIDL precipitated industrialized economies’ variegated adoption of industrial deregulation: with more open markets and increasing competition, the Fordist accumulation system could not cope. While it was not a prescribed policy panacea, NIDL was the result of this nascent shift in economic governance.

Perhaps more crucially, firms’ strategies to develop transnational production networks are intimately connected to concrete state policies within multiple economies that created fresh conditions for capital to flow internationally. Empirical research has demonstrated that developing countries were not passive actors in shaping MNCs’ location strategies, nor were their industrialization strategies predicated on an ecumenical blueprint sourced from the “center”; the wholesale imitation of Fordism, notably the establishment of high wages for semi- or low-skilled workers, did not occur at all. New markets also emerged outside the “center,” first in so-called newly industrializing economies like South Korea and Taiwan, and then in emerging economies such as China, India, Russia, Brazil, and Mexico. Even within the “core,” common markets were formed in what is now known as the European Union and North American Free Trade Area (NAFTA). These findings provided a much-needed refinement to the NIDL thesis, which is often critiqued as describing a unidirectional and mechanistic transfer of labor-intensive, lower-order economic activities from industrialized countries to developing countries in order to achieve cost savings; they show how the international division of labor is driven as much by place-specific political objectives as by the profit motive.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Comparative advantage; Cores and peripheries; Fordism/post-Fordism;
INTERNATIONAL DIVISION OF LABOR

Fragmentation of production; Global production networks; Labor geographies and the corporation; States and development

References


Further reading

International:
Environmental Information
Systems Africa
(EIS-AFRICA)

Founded: 2000
Location of headquarters: Pretoria, South Africa
Website: www.eis-africa.org
Membership: 93 (as of October 2014)
Executive Director: Sives Govender
Contact: sgovender@eis-africa.org,
sives.govender@gmail.com

Description and purpose

EIS-AFRICA is a membership-based organization created by African experts to advance Africa’s development goals through more effective access and use of geospatial information, science, and technology. EIS-AFRICA has been incorporated as an association under Section 21 of the South African Companies Act (Act 61, 1973) and registered as a nonprofit association in the Republic of South Africa since 2000. The organization is governed by an elected board of directors (from its members) and is audited annually by Grant Thornton. EIS-AFRICA’s daily activities are undertaken by a secretariat which is managed by a board-elected executive director.

EIS-AFRICA’s vision is an African society where geospatial information for sustainable development is readily available and easily accessible and its mission is to be recognized by its strategic partners, clients, and the public as a legitimate and leading organization developing African capacity to generate, use, manage, and disseminate geospatial information, science, and technology to enrich policy debate and support decision-making for the wellbeing of Africa’s people.

Journals or major publication series


Current activities or projects

EIS-AFRICA publishes a monthly electronic newsletter which is received by over 5000 contacts from 108 countries in the world and hosts the biennial AfricaGIS conference and exhibition. EIS-AFRICA has contributed to publishing of selected papers from its conferences to a special IJAGR–AfricaGIS journal edition.

EIS-AFRICA contributed to the studies on the determination of fundamental geospatial datasets for Africa which was published by the UNECA. EIS-AFRICA is key contributor to GSDI, ISO TC 211, and UN–GGIM.

EIS-AFRICA in collaboration with the American Association of Geographers (AAG) hosted a Global Connections and Exchange 2014 Youth TechCamp: My Community, Our Earth Program (MyCoE) during July 2014 where 10 American and 40 South African high school students learnt about geography, mapping, and environment management (http://www.aag.org/cs/my_community_our_earth/mycoe_
Brief history

EIS-AFRICA has been promoting the use of geospatial information, science, and technology for sustainability on the continent since the early 1990s (first as the EIS program for sub-Saharan Africa, then EIS-AFRICA).

The organization was a key partner in developing the National Environmental Action Plans which are known today as State of the Environment Reports and was an early adopter of GIS technologies. EIS-AFRICA has acted on behalf of the Africa geospatial community to engage with global, multinational agencies to access spatial information, training and software to increase its capacity, and competency to better manage our environment.

EIS-AFRICA sits on various UN, global associations, continental, and intergovernmental meetings advising on how geography and geographic information is a key infrastructural requirement in better managing our environment for a sustainable future. EIS-AFRICA has also been involved in hosting many capacity building activities ranging from the AfricaGIS conference and exhibition (first held in 1993) to many national training events.

EIS-AFRICA’s network includes participants from all 54 African nations as well as all the leading regional and subregional organizations, national mapping agencies, and many universities. EIS-AFRICA’s membership program allows its members free membership to the American Association of Geographers and the organization have also just established a facility to host the newly formed South African Geography Teachers Association which will be expanded to a continent-wide group.

Submitted by Sives Govender
EUROGEO is a European scientific society, which aims to develop, support, and promote policies and actions designed to advance the status of geography; establish and promote cross-border cooperation; promote education and training in geography from a European perspective; and represent nationally and internationally the views of its members.

The association supports geographers in their jobs and careers, promoting good practice and cooperating with agencies such as the European Union, Council of Europe, European Commission, UNEP, and other relevant organizations.

To meet these aims, the association operates in relevant European policy areas and organizes events and activities for its members, the most significant of which is the EUROGEO annual conference.

The main research and development projects of the association are: iGuess 2 (www.iguess.eu), a project for the creation and implementation of European training courses on the use of geographic information systems; SPACIT (www.spatialcitizenship.org/), a teacher training project on issues of citizenship and spatial thinking; PiriReis (http://pirireis.dicle.edu.tr), a transfer of innovation project to create and develop an association of geographers in Turkey; I-Use: Statistics in Education (http://i-use.eu), a project promoting the use of statistics in education; Geocapabilities (www.geocapabilities.org), examining powerful knowledge and teachers as curriculum leaders; and Geoskills Plus (www.geoskillsplus.eu), a transfer of innovation project bridging the gap between universities and vocational training with employers.

EUROGEO is a consortium partner in a new “Science in Society” research project called “Sea Change,” which aims to raise awareness of the importance of understanding the oceans.

EUROGEO has helped establish, as founder members, a number of innovative networks.
INTERNATIONAL: EUROPEAN ASSOCIATION OF GEOGRAPHERS (EUROGEO)

in Europe, such as the European Citizen Science Association and the European Alliance of Social Science and Humanities Research (EASSH). EUROGEO is also a stakeholder in the UNEP EyeonEarth, “Geo for All,” and Smart City Learning initiatives. The association has developed active working links, not only with national and international associations in Europe, but all over the world.

Brief history

EUROGEO was established in 1979 as an association of national associations, with strong support from the European Commission. National bodies in Europe were networked together to form the European Standing Conference of Geography Teacher’s Associations (ESCGTA) and to represent the views and opinions of geographers in different European countries. They met every 2 years and published a bulletin called EUROGEO.

As Europe grew, organizations from other countries became members and the association took the name of its bulletin: EUROGEO: the European Network of Geography Teachers’ Associations.

In 1987, in recognition of its activities, the association was granted official consultative status to the Council of Europe, making it one of the first to be awarded this position. Since 2003 EUROGEO has been a full participant of the Council of Europe, being the voice of geography and geographers there, actively lobbying politicians and their advisors on themes such as climate change, landscape convention, migration, borders, energy, human rights, and citizenship education.

Following reform in 2009, EUROGEO became a membership organization allowing individuals and institutions with an interest in geography, its theory, and methods to become involved.

The association is an international nonprofit organization (INGO) incorporated under the name EUROGEO: the European Association of Geographers by the provisions of Title III of the Belgian law of June 27, 1921.

Submitted by Karl Donert
International: Instituto Panamericano de Geografía e Historia (IPGH) (Pan American Institute of Geography and History)

Founded: 1928  
Location of headquarters: Mexico City  
Website: www.ipgh.org  
Membership: 21 member states and 4 permanent observers  
Secretary General: Rodrigo Barriga-Vargas  
Contact: secretariageneral@ipgh.org

Description and purpose

The main purposes of the IPGH are to encourage, coordinate, and publicize cartographical, geographical, historical, and geophysical studies, as well as other related scientific studies of interest to the Americas; to promote and coordinate scientific and technical development, research, relations among institutions and specialists, studies and training in cartography, geography, geophysics, and history; and to promote and stimulate cooperation among the specialized institutions of the Americas and international organizations in its four fields of activity.

Journals or major publication series

Revista Cartográfica  
Revista Geográfica  
Revista Geofísica  
Revista de Historia de América  
Revista de Arqueología Americana  
Boletín de Antropología Americana

Current activities or projects

Through its commissions, IPGH offers a wide range of courses, workshops, and conferences directed at specialists and professionals in the areas of interest of the institute, and in general to other professionals interested in those fields.

Through its Commission on Cartography the institute develops important projects in the areas of: fundamental geospatial data (SIRGAS, fundamental data, geographic names, aeronautic and hydrographic charts); institutional strengthening and technical cooperation (geospatial data infrastructure); thematic applications (geospatial information for the prevention and mitigation of disasters, information for planning and decision-making); standards (positioning, digital cartography, data exchange).

In the field of geography, the Commission on Geography is developing projects on urbanization and environmental impacts, geographic information systems, regional development, intermediate cities and quality of life, and teaching and professional training.

The Commission on History develops projects on archaeology, anthropology, folklore, bibliography and archives, economic, political, social history, and history of the ideas.

The Commission on Geophysics supports projects on seismology, vulcanology, meteorology, atmospheric physics, geomagnetism, aeronomy, physical oceanography, gravimetry, and environmental geophysics.
Brief history

The General Assembly is the supreme organ of the IPGH and determines the institute’s scientific, administrative, and financial policies. The Directing Council is the Pan American organ of the IPGH that is entrusted with the functions of the General Assembly during the intervals between meetings of the latter. Meeting of Officers is the organ that directs and coordinates the institute’s activities between the meetings of the Directing Council. The General Secretariat is the IPGH’s central and permanent organ responsible for administration, coordination of organ activities, providing the necessary assistance for the proper functioning of said organs, execution of the tasks entrusted to it, and ensuring compliance of the agreements adopted for the smooth functioning of the IPGH. The secretary general represents the IPGH. The Commissions are the organs entrusted with promoting the scientific and technical development of their respective fields of action in the member states. They are also responsible for coordinating, encouraging, and supervising projects and other agreements involving research approved by the General Assembly or the Directing Council. There are four commissions: Cartography, Geography, History, and Geophysics, which are subdivided into committees and working groups. The National Sections are entities established by each member state, whose role is to fulfill the objectives of the IPGH in the sphere of their respective countries.

Submitted by Santiago Borrero
International: International Geographical Union (IGU)/Union Géographique International (UGI)

Founded: 1922
Location of headquarters: Cape Town, South Africa
Website: www.igu-online.org
Membership: 75 (as of 2014)
President: Vladimir Kolossov
Contact: vladimir.kolossov@gmail.com

Description and purpose

The International Geographical Union (IGU) is an international, nongovernmental, professional organization devoted to the development of the discipline of geography. The purpose of the IGU are primarily to promote geography through initiating and coordinating geographical research and teaching in all countries of the world. Its work is conducted through the instruments of its national committees, commissions, and task forces. The IGU hosts the International Geographical Congress every four years and also promotes regional conferences and other meetings that further the objectives of the union. The IGU also facilitates the participation of geographers in the global community of scientists through its formal affiliation as a member union within both the International Council for Science (ICSU) and the International Social Science Council (ISSC).

Journals or major publication series


Current activities or projects

International Year of Global Understanding: Thinking globally and acting appropriately locally presupposes global understanding. This initiative by the International Geographical Union aims to bridge the awareness gap between local acts and global effects. It will achieve this aim through research, education, and information. Ultimately, the International Year of Global Understanding will encourage everyone to make daily decisions in the light of global challenges.

IGU “Journals Project”: There are many journals globally that deal with the broader discipline of geography. In recent years the domination of major publishing houses in the scientific journal market has become much stronger and it is clear that many geographical journals are published that have a much lower profile and yet could provide a very valuable resource for geographers – researchers and teachers alike – in particular national or regional contexts. It was for this reason that the IGU embarked on a project to establish a searchable global database of geographical periodicals. Follow the link from the IGU homepage at www.igu-online.org.

Our Sustainable Cities (“OurSus”): This project is based on the idea that, despite their different sizes, location, and functions, all cities are facing similar challenges in terms of sustainable
features, such as green R&D, environment-friendly consumption, green campaigns, challenges, and education, and so forth. In 2013 the project had a full double Internet infrastructure, in English and Chinese (www.oursus.org and zh.oursus.org).

**Brief history**

The IGU was formally established in Brussels in 1922. However, the history of international meetings of geographers is much longer. The first of a series of congresses met in 1871 in Antwerp. Since its early days the union has consisted of three major components: a general assembly of the delegates appointed by the member countries which meets at the time of the congress and is the highest authority of the union; an executive committee which consists of a president, eight vice-presidents and a secretary-general and treasurer; commissions and study groups which continue their work between general assembly meetings. The working languages of the union are English and French. Recent International Geographical Congresses have been held in Tunis (2008) and Cologne (2012) and the next one is scheduled for Beijing in 2016.

Submitted by Mike Meadows
International: International Cartographic Association (ICA)

Founded: 1959
Location of headquarters: Vienna, Austria
Website: www.icaci.org
Membership: 78 national members; 37 affiliate members (as of 2013)
President: Georg Gartner
Contact: georg.gartner@tuwien.ac.at

Description and purpose

ICA is the world authority body for cartography and GIScience. Through national membership and affiliate members the ICA focuses on the following aims.

- Advancing the study of cartographic or geographic information (GI) science issues. In particular it is concerned with the processing, storage, and analysis of source material and the design, construction, reproduction, and display techniques of maps and associated forms of graphic communications.
- Initiating, fostering and coordinating research in cartography and GI science.
- Organising international and regional conferences, meetings, exhibitions, and outreach programmes.
- Establishing commissions and working groups to work on issues of particular interest to cartography and GI science.

Journals or major publication series

Book series


Affiliate journals

Cartography and Geographic Information Science, CaGIS. www.cartogis.org/publications/journal.php

Newsletter


Proceedings


Current activities or projects

ICA supports and recognizes cartographers and GIScientists around the globe through its commissions and working groups (www.
icaci.org/commissions), the International Cartographic Conferences (www.icaci.org/icc), several publication channels, cartographic exhibitions and displays, children map drawing competition, research agenda and research grants, awards, outreach programs, and the close cooperation with sister societies in the Joint Board of Geoinformation Societies (JBGIS) as well as with international bodies such as the United Nations.

**Brief history**

ICA was founded on June 9, 1959, in Bern, Switzerland. The first general assembly was held in Paris in 1961, that is also when the statutes were accepted. Later conferences have been held around the world from India (Delhi, 1968) to South Africa (Durban, 2003), from Italy (Stresa, 1970) to Mexico (Morelia, 1987), from Australia (Perth, 1984) to China (Beijing, 2001) and in many other locations.

The international nature of ICA activity has also been reflected by the work of its commissions and working groups over many decades, under the leadership of many different people, and in many different places. These organizations have addressed the full range of scientific, technical, and social research which is the mark of ICA activity.

Throughout its 50-year history, ICA has brought together researchers, government mapping agencies, commercial cartographic publishers, software developers, educators, earth and environmental scientists, and those with a passion for maps. The cartographic world has changed significantly since 1959, while the role and impact of ICA has been steadfast.

Submitted by Georg Gartner
International: Southeast Asian Geography Association (SEAGA)

Founded: 1990
Location of headquarters: Singapore
Website: www.seaga.info
President: Chew Hung Chang
Contact: seaga@nie.edu.sg

Description and purpose

The Southeast Asian Geography Association (SEAGA) is a scientific and educational society set up in 1990 by a group of geographers. Its members share interests in the theory, methods, and practice of geography and geographic education. SEAGA’s vision is to be an association that will raise the profile of Southeast Asian geography and geographers. The key mission is to promote and disseminate geographic research, promote geographic education, and develop a global network of geographers interested in the Southeast Asian region.

Journals or major publication series

A list of publications can be found on the association’s website at www.seaga.info/publications

Current activities or projects

The association organizes a biennial conference, regular international forums, and other workshops hosted in different ASEAN countries which provide a platform for sharing research findings in geography and related area with a focus on Southeast Asia.

SEAGA places a strong focus on geography education within the conferences, often supported by a sizable contingent of geography teachers, as well as university lecturers and professors who have a keen interest in geography education.

SEAGA Publishes an eNewsletter twice a year as well as proceedings of the meetings. Occasional special issues of journals, book chapters, and books arising from meetings are also encouraged.

Brief history

The founding president of the association was Professor Goh Kim Chuan. From 1990 to 2005 he was instrumental in encouraging and assisting geography departments in the region to host SEAGA’s biennial international conferences. To date they have been held in Brunei, Indonesia, Malaysia, Thailand, Singapore, Philippines, Vietnam, and Cambodia. The events also allow geographers to connect with each other and enhance cooperation in joint research, student fieldwork, or even staff exchange.

The association has been formally registered with Singapore’s Registrar of Societies since 2006. The secretariat for the executive committee of SEAGA is hosted at the Humanities and Social Studies Education Academic Group of the National Institute of Education, Nanyang Technological University, Singapore.

One aspect SEAGA has been encouraging is to make the conferences accessible to geography
teachers and graduate students in each of the host countries either through their direct participation in sharing their research findings or having a complementary one-day workshop on some aspects of the teaching of topics in geography. In this way, the association hopes to improve and revitalize the status of geography in the region.

Submitted by Chew Hung Chang
International: The International Association of Chinese Professionals in Geographic Information Science (CPGIS)

Founded: 1992
Location of headquarters: Ann Arbor, MI
Website: https://cpgis.org
Membership: 1450 (as of September 21, 2015)
President: Mei-Po Kwan
Contact: http://meipokwan.org

Description and purpose

The CPGIS is a non-profit organization committed to promote research, education, and entrepreneurship in geographic information science and technology. Its members include scholars, researchers, students, and employees of businesses and government from over 30 countries. The CPGIS seeks to: promote members’ professional development through collaboration; promote intellectual exchange in GIS and related sciences and technologies between Chinese GIS professionals abroad and those in China; provide a channel for interaction and collaboration between its members and other GIS professionals; and promote education of geographic information sciences at all levels around the world.

Journals or major publication series

- Annals of GIS, www.tandfonline.com/loi/tagi20#.VgAY55eguDk
- The CPGIS Book Series

Current activities or projects

The CPGIS organizes an international conference on GIS (Geoinformatics) annually in various countries. It edits a peer-reviewed academic journal (Annals of GIS). It maintains a website (www.cpgis.org) and a listserv (CPGIS-L@listserv.buffalo.edu) with over 1400 subscribers. The CPGIS also organizes a wide range of GIS-related workshops, initiates various projects to promote GIS in both the public and private sectors, particularly in China, and encourages young scholars to be involved in its various activities (e.g., Health GIS Workshops, Go Home Professional Workshops, and Young Scholar Summer Workshops). It recognizes the accomplishments of Chinese professionals in GIS through various awards: Excellence in Service, Distinguished Scholar, and student paper competition awards.

Brief history

In the summer of 1992, a group of students led by Hui Lin and Yuemin Ding organized an international conference at SUNY-Buffalo and founded The Association of Chinese Professionals in Geographic Information Systems - Abroad (CPGIS). To reflect its growing visibility in the international community and the rapid expansion of the field, the CPGIS changed its name to The International Association of Chinese Professionals in Geographic Information Science (CPGIS) in October 2001. Since its establishment in 1992, the CPGIS has grown into a mature and vibrant association with active members from over thirty countries around the world. It edits its official journal (Annals of GIS), organizes annually the Geoinformatics international conference, the annual Young Scholars Summer Workshops, Go Home Professional Workshops, and many other
INTERNATIONAL

activities. The CPGIS is a unique organization through which Chinese students, scholars, and professionals in GIS can develop their social networks, leadership, and professional careers. The CPGIS is governed by an elected president and an elected board of directors. Its many programs and activities are organized with the assistance of other appointed officers and committees. A group of regional officers and directors help coordinate its members in the United Kingdom, Germany, Japan, Australia, and other countries.

Submitted by Mei-Po Kwan
Internationalization

Patrik Ström
University of Gothenburg, Sweden

Internationalization of business activities occurs when a firm starts to trade or make investments in new host markets abroad. Before, internationalization often meant sending goods to markets overseas or establishing a production plant in another country. Today, technological advances enable firms within advanced production and producer services to supply distant markets through the Internet and telecommunication. A physical presence in the markets abroad is no longer the only way of internationalizing business activities. A producer service firm can supply a market overseas with services without having a physical base. It can receive and supply services through secure Internet channels. The firm may alternatively send a consultant abroad to supply the service or establish an office abroad. This new form of internationalization has also enabled service firms from emerging markets to supply their services over long distances to customers in mature markets.

The outcome of internationalization is evident at many levels of geographical and locational analysis. Its effects are evident in the study of the macro-economy, where trade and foreign direct investment connect places to each other, in the micro-level study of firms and in studies of how the internationalization process has developed and been implemented over time. The fundamental aspects of internationalization will first be considered in relation to the approaches used in economics. The analysis is then deepened by incorporating a more spatial perspective. This will illustrate the important connection between economic geography and international business in an attempt to uncover and explain the reasons why firms engage in internationalization. It also illustrates the complexity that exists in the internationalization of producer services. The entry concludes with a summary of new ways of exploring internationalization from the viewpoint of the speed with which it can occur and the growing involvement of companies from emerging markets. This is an aspect of the study of internationalization that requires further theoretical development.

Classic approaches

The more modern explanations of international trade require an appreciation of the concept of comparative advantage which states that a country should specialize in producing and exporting only those goods and services that it can produce more efficiently than other goods and services, which should in turn be imported. Using the fundamental parts of the Heckscher–Ohlin theory, Krugman (1994) adds other explanatory reasons for international trade that are overlooked in traditional trade theory. Trade arises not only from differences in factor endowments between countries but also because of the inherent advantages of specialization. The Krugman approach also puts more stress on the important factors of arbitrariness and increasing returns. The more dissimilar the factor endowments of two countries, the more important are the comparative advantages. If factor endowments are similar, there is a good potential for...
INTERNATIONALIZATION

intraindustry trade. Thus, interindustry trade reflects natural comparative advantage while intraindustry trade reflects acquired comparative advantage. The parts and components required by large companies for the assembly of final products such as vehicles or laptops are usually produced at diverse locations in order to minimize production costs. It is product complexity together with the increased production of intermediate goods that opens the way for intraindustry trade.

Technological development itself can generate temporary comparative advantages. In the United States, transistors and integrated circuits gave manufacturers a valuable headstart but this soon disappeared when other countries such as Japan imitated the technology. Typically, the most industrialized countries export the most technologically intensive products and import more standardized products. This follows the principles of the so-called product life cycle theory, in which location and internationalization are related to the cost of production (Vernon 1966). Finally, transportation costs and environmental standards are also important determinants of international trade; for example, decreasing transport costs for automobiles have facilitated long-distance international trade.

Global shifts and spatial division of labor

On the basis of the transformational power of the economy, Dicken (2011) states that earlier theories of the changing economic situation in the world have been too static. He argues that it is organizational and institutional forces that create shifts in the international economy. Comparative advantage is not simply a given. Rather, it is a process created and recreated as a consequence of human action that is reflected in such things as the variations in trade policy between countries or in levels of innovation and technological development. Countries must try to sustain their competitive advantage and firms must keep up with technology in order not to lose ground. These are the considerations that largely explain why firms tend to reorganize their international operations. They have mainly been used to understand the internationalization of manufacturing, but may also be applicable to the changes taking place within the service economy and its tendency to outperform the manufacturing sector in terms of contributions to GDP and employment. The recent rapid restructuring of economies from manufacturing to services in both mature and emerging markets has been characterized as the second global shift (Bryson 2007). It has been spurred by the second unbundling of globalization, that is, the radical lowering of data and other transmission costs as a result of information and communications technology (ICT) (Baldwin 2014). This has resulted, for example, in changes to trade patterns and to global value chains in East Asia whereby trade in goods has been displaced by trade in tasks. Competition in low-value-added services that can also be traded is fierce, but competition is increasing in high-value-added subsectors of the service industry, where the content of the services provided in the international market has a strong impact on the competitive advantage of a product or service. As the combination of manufacturing and services becomes even more important for how and where the value added in production is generated, it has further impacts on the internationalization of trade and foreign direct investment (FDI) (Daniels 2012).

Cultural considerations are evident for all kinds of international transactions, but in the case of service industries they may be an absolute barrier. Such barriers have their greatest impact on verbal and media-based services, but standardized and highly technological services may
be exceptions to this general rule. These services can be found in various back-office activities handling, for example, financial transactions. Specialized services that are essential in global value chains are more often delivered through face-to-face interactions; such internationalization is thus more dependent on people. A further factor is risk in general, such as that associated with political situations, insurance requirements, financial probity, or adaptation to unfamiliar legal systems. It is argued that these barriers, or risks, may be greater for service firms than for goods-producing firms involved in internationalization. The sheer complexity of undertaking international operations that span several countries or continents also poses new challenges for operational management. In general, firms located in major metropolitan areas are more likely to be confronted with various internationalization opportunities. Firms in more peripheral areas also show an interest in working internationally, but their opportunities are more limited.

In relation to the classification and conceptual difficulties associated with service trade, empirical findings show that the direct contribution of services to national exports is growing slowly and that the pattern of trade is highly concentrated. Instead, it may be useful to acknowledge and nurture the indirect contribution of services to overall national export activity. Service support for many of the manufacturing sectors enhances total export competitiveness. Specialized services then become embodied in the goods and services exported to a specific country or out into the world market. In the case of services, because of their intangible character, FDI in the form of mergers and acquisitions seems to be more important than greenfield investments and there is a need to locate at specific places in order to be recognized as a serious service provider. Investment shifts in the economy due to variations in labor costs are not a new phenomenon. Much of the evolutionary character of international economic development has been a result of increasing wage levels. As wage levels rise in the developed world, more production is located in the developing countries. This trend is said by some to have raised the capital and knowledge content of products produced in emerging markets (Alvstam, Dolles, and Ström 2014). Since the mid-1990s the rise of emerging market multinationals has started the repositioning of competition in the global market. The impact is clearly visible in the changing pattern of trade and FDI flows, and it has also helped to refine conceptual understanding of how these firms internationalize.

Internationalization theory and the firm

The study of company internationalization can be divided into two main research approaches: economic theory and behavioral theory (Dunning and Lundan 2008; Vahlne and Johanson 2013). The purely economic viewpoint originates from classical trade theory which is based on comparative advantage. Over time, theories such as transaction cost economics have been developed to interpret and explain the complex international business environment. Most of the research has focused on the manufacturing sector, but more recently there have been attempts to explain the growing trade and internationalization activity within the service sector.

There has also been a tradition within internationalization theory of adopting a more behavioral approach. After decades of analyzing internationalization, mainly from the economic viewpoint (see Vernon 1966), international business research entered a new era following the work of Johanson and Vahlne (1977). The so-called stage theories of the Uppsala School,
which are based on the role of proximity, psychic distance, and experiential knowledge, have greatly helped to advance the theoretical development of the behavioral approach. Its significance has also been tested, analyzed, and further extended. However, a basic problem with stage theory is the nature of the conceptual framework and its relatively modest empirical foundations. The early studies were based on only a few case studies of multinational Swedish manufacturing companies. These have since been extended in terms of geographical coverage and through the broader coverage of different industries. These concerns have also been alleviated to some extent by incorporating network theory into the research (Vahlne and Johanson 2013) so that the dynamic dimension and speed of the internationalization process are examined in relation to the difficulties arising from foreignness and outsidership in distant markets. This implies that it is not the distance and cultural aspects that necessarily have a strong influence on the possibility of internationalizing operations, but rather that a firm is able to access business networks at new international locations. These aspects are highly important for the internationalization of producer services.

International business and location

Research on international business (IB) takes into account many different aspects of the firm and its environment. The focus is on the firm and its international operations, with special reference to the organization, strategic aspects, and knowledge transfer. The relationship between strategic aspects and location is seen as a key issue for a satisfactory explanation of internationalization (Buckley and Gauri 2004).

In early IB research, Hymer’s (1976) objective was to explain the reasons for rather than the operations of a multinational corporation (MNC). However, the contributions made by Vernon (1966) regarding cycles and trade/antitrade FDI seem to equate with a more normative explanation of the internationalization process. Since these early studies, the complexity of the global economy has created a need for theories that develop an understanding of competition between firms. The technological environment or technological accumulation possibilities are important for shaping the pattern of international operations, where technological abilities and research and development (R&D) constitute competitive advantage to a greater extent in complex global value chains. These are aspects that clearly relate to the discussions in economic geography about firm clusters and national innovation systems. This can be seen as an increasing interest for including location in management-related sciences since, traditionally, location factors have received limited attention in the IB literature. However, Dunning was a forerunner among those placing more emphasis on the role of location and location-bound assets. This connects to the debate within economic geography about the development of clusters with special characteristics and the way in which locational advantages can be created by regional innovation systems. That locational configuration in itself can be an ownership-specific advantage adds another perspective to IB research and the value of the ownership, location, and internalization (OLI) configuration.

Locational configuration, such as regional innovation systems, lies within the range of governmental influence, which means that a business-oriented approach is necessary to create and sustain innovative clusters that can both stimulate the growth of new companies and attract existing MNCs. This issue is of great importance in the field of economic geography because of the changing landscape of
industrial production and the way in which high knowledge content services are put forward as new engines of regional development and where regional innovation systems help to facilitate the attraction of regions or cities for international firms. These are very important locational aspects in driving the internationalization of producer services. The most essential resource is the competence and experience of existing and potential employees; they are most likely to be found in urban areas with a strong service economy orientation in different places in the international market.

Combining research on the managerial aspects of internationalization and localization with the more descriptively or conceptually oriented research undertaken by economic geographers is a useful way forward (Ström and Wahlqvist 2010). The internationalization of manufacturing and service firms needs to be explained with reference to both the locational and the managerial perspectives. The strategic perspective seems to miss the subnational locational dimension that relates to the advantages of agglomeration, proximity, and global connectedness. These are described as pipelines of communication. The possibility for firms to utilize the specific advantages of regional innovation systems to enhance their internationalization is essential in explaining why some regions and not others have a strong position within global production chains (Ström and Wahlqvist 2010). These complex value-adding networks have become the cornerstone of internationalization for the production of both goods and services (Beyers 2012).

Even if firms today are operating on a global stage, the aspects of location that facilitate internationalization have not diminished. Derived from economic geography theory, research on innovation shows at least three characteristics: (i) there is a need for a reduction of technical and economic uncertainty over time; (ii) interactions with outside parties are necessary; and (iii) face-to-face contact is still important which means there remains a need for business travel even if information technology has reduced the friction of distance. These three characteristics underpin the ongoing importance of industrial clusters or agglomerations of various kinds. If the diffusion of knowledge, ideas, or good business practice, for example, is easy, globalization is said to override locally confined innovation processes. If the diffusion process is difficult and involves expensive transaction costs, localized innovation processes are still of great importance and will eventually have an impact on the internationalization processes of firms, not least their geography.

Rate of internationalization and emerging market multinationals

Since the mid-1990s the rate of internationalization has changed drastically. In the early studies of how firms internationalized, the process usually extended over several years. Today firms can become international in an instant, thanks mainly to the possibilities offered by a combination of increasingly sophisticated telecommunications and the Internet. Rather than being a later step in firm development, some firms have internationalization on the agenda right at the start of operations. This “born global” phenomenon enables firms to become international using connections they have established with partners around the world. It often coincides with the entrepreneurial aspects of how to expand a firm over extended geographies and to markets with large cultural differences.

The last two decades have also witnessed another phenomenon in relation to internationalization. Whereas firms from mature and the
most advanced economies were the first to take advantage of their competitive strengths to venture into foreign markets, firms from emerging market economies in Asia and Latin America are also now rapidly expanding their geographical reach to both developing and developed markets (Ramamurti and Singh 2009). These emerging market multinationals have become serious competitors to many of the most established firms in the mature economies. Technological possibilities have also enabled firms from emerging markets to compete in high-end producer services. Information technology services are one example where firms from India have established a strong presence on the international market, by using back-office functions in the domestic market for supplying services to customers in Europe and North America.

Several emerging market firms have also established themselves as the largest players in their respective industries. From the perspective of traditional internationalization theory, they have become internationalized from a position of disadvantage. Their only advantage was seen as the low cost structure for the production of goods and services, but recently these firms have moved in to high-value-added activities such as advanced production networks and high knowledge content services. Some have also tried to leapfrog into internationalization by acquiring well-known and respected corporate brands. Examples include the takeover by Lenovo of IBM’s laptop business and the takeover by Geely of Sweden’s iconic Volvo Car Corporation.

Internationalization has become the backbone for connecting business locations around the world. Economic integration has enabled regions to increase their value-adding possibilities. Complex regional and global production networks have managed to reduce costs and allowed more complex divisions of labor, which now makes up a larger share of service activities. Internationalization is difficult to understand unless studies of the spatial dynamics that drive location patterns are combined with strategy-oriented studies of how and why firms internationalize in a service-dominated economy. As global markets become ever more intertwined, the study of internationalization will develop further. Apart from examining how firms can now find their way into international markets in an instant, it is also necessary to better understand how firms from emerging markets are capturing distant markets in both mature and developing economies so effectively. Technology has helped to shrink the world and to increase the speed of service delivery across markets. It has enabled new forms of internationalization. This calls for further cross-disciplinary studies that combine the economics, political economy, spatial dynamics and configuration, and firm-level location strategies that together shape the internationalization of services and manufacturing.

SEE ALSO: Global production networks; Industrial geography; Knowledge-based economy; Producer services and trade liberalization

References


INTERNATIONALIZATION


“The Internet changes everything” was the over-the-top mantra of the 1990s dot-com boom, easily critiqued in light of the subsequent bust in the early 2000s. Twenty years later, a similar refrain can be heard from the top executives of technology corporations like Google and Facebook and seen in initiatives to apply information and communication technologies to everything from the complex problem of international development to more mundane matters like catching a cab or finding a parking spot. While not wishing to characterize such simplistic sloganeering as truly visionary, it is important to recognize that there are now few corners of our social, economic, and political structures (at least in the industrialized world) that remain untouched by the Internet and its associated technologies. Computer networks, powerful digital devices, and integrated databases are now deeply embedded in the products and daily practices of the global economy.

This entry reviews this complex intertwining of information technologies and the more conventional material geographies of the global economy. We seek to provide some answers to the questions of how the geographies of global capitalism have shaped the Internet and how the Internet has reshaped the geographies of global capitalism. Clearly it is not enough to simply assert that everything (or nothing) has changed; instead we must understand the nature of changes in the space economy, how these changes came about, where and to whom power and wealth has shifted, and how new products, organizations, and spaces in the global economy are emerging thanks to the availability of the Internet and digital technologies.

The informational infrastructures of the global economy

During the commercialization and rapid diffusion of the Internet in the early 1990s, techno-utopian visions abounded: workforces freed from the constraints of co-presence by the power of telecommuting; the decline of cities as agglomeration economies became less relevant; and, most famously, the “death of distance” as the Internet provided constant connectivity (cf. Cairncross 1997). These visions were, at their core, technologically determinist, predicting uniform and massive upheaval in long-standing social and spatial structures with little regard to the complex ways the Internet could be put to work in the economy. Not surprisingly, the reality that emerged was much different and considerably more complex than these simplistic visions.

A key reason behind this is that the Internet is grounded in particular places through the massive, albeit often unseen, network of material infrastructure necessary to support its existence, from fiber-optic cables spanning the continents.
and oceans to so-called server farms located in peripheral locations, which house all of the data we generate and store “in the cloud.” Despite widespread visions of ubiquitous connectivity, access to the Internet has remained stubbornly uneven, from the local scale to the national and supernational scales, due to the “splintering” effect engendered by Internet infrastructures (Graham and Marvin 2001; Malecki 2002). This splintering has meant that rather than the predicted universal access to the Internet, some places (particularly urban areas) have become quite well connected, while other nearby places remain practically off the map.

The tendency for informational infrastructures to cluster together in relatively dense places has ultimately meant that the geography of the Internet has often reinforced existing patterns of urbanization and agglomeration, rather than overturning them. Using a range of metrics – from bandwidth measures, points of presence, and domain name registrations, among other things – early geographers of the Internet demonstrated that the then-emerging information economy was not somehow separate from other sectors of the economy and their geographic influences (Townsend 2001; Zook 2000). This work highlighted the importance both of a handful of key “global cities” like New York which were at the center of the global network, and of long-standing national powers, such as the United States. Although the United States retains a powerful place within the landscape of Internet connectivity and governance, its centrality has diminished somewhat in more recent years. For example, it was recently announced that the US Commerce Department would cede its control over the Internet Corporation for Assigned Names and Numbers (ICANN), the body responsible for governing the domain name registration system, to the international community.

As has been the case with the United States’ role, these patterns of continuity, however, have also been accompanied by reconfiguration of urban hierarchies tied to Internet infrastructure. Some cities, such as Washington, DC, San Francisco, and Seattle, have benefited economically during the last two decades due to their positioning as transoceanic cable landings (Townsend 2001). While New York remains the most significant city in this respect, due largely to its direct connections with Europe, these other cities have encroached on the status of America’s more conventional global cities such as Chicago and Los Angeles. The groundedness of this infrastructure in particular places, however, also leads to potential for disruptions. While targeted terrorist attacks on critical infrastructure (Roberts, Secor, and Zook 2012) or disruptions to global financial markets caused by earthquakes or other natural disasters tend to preoccupy some, much more straightforward and mundane disruptions are also possible. For example, the ability of the Egyptian government to simply “switch off” the country’s Internet during the Arab Spring protests in 2011 due to the centralization of the country’s global connections in a single government-owned building, or the case of a Georgian woman scavenging for copper who sliced through a cable and disconnected 90% of Armenia’s Internet, highlight the importance of these material manifestations to this supposedly virtual domain or global capitalism.

Infrastructure alone is rarely enough as the immaterial knowledge infrastructures present in urban centers – often built around existing industries or capabilities – have also proven key in the reordering of these economic hierarchies. For example, Zook (2000) has shown that places like San Francisco and the Bay Area in California, Provo, Utah, Portland, Oregon, Austin, Texas, and Las Vegas, Nevada, emerged as relatively specialized centers of Internet content.
production, while a handful of cities in Florida have become well known for their specialization in the Internet adult industry. While the reasons behind this clustering are complex and many – indeed, not all places can develop such clusters – these particular urban nodes have succeeded by leveraging existing local economic activities such as software development and gambling onto the Internet. More contemporary data show that specific slices of informational activity are concentrated unevenly and sometimes in somewhat unexpected ways. A case in point is that the Tokyo metropolitan region retains its centrality in terms of user-generated Google Maps content, but has been surpassed, along with New York and London, by Jakarta, Indonesia, in tweeting activity. So while the particular places privileged by the Internet’s geography may not accord precisely with rankings of global cities or other expectations about the distribution of economic power and resources, larger scale social and economic processes, such as geographically uneven development, have been intensified by the Internet, rather than eliminated.

Digital commodities and consumption channels

While the Internet has contributed to the restructuring of global urban hierarchies, perhaps the greatest impact has been on the (re)organization of the economy. This is most evident in creating space for new kinds of products and services which are either wholly digital or which rely upon the Internet to facilitate the consumption of particular offline goods or services. This includes digital versions of previously material items such as music or movies, and inherently digital commodities for use in gaming or virtual worlds like Second Life and World of Warcraft, as well as novel services like eBay or AirBnB, which redirect consumption patterns, or social media platforms focused on the production, capture, and exchange of user-generated information.

A commonality across new digital products is that the Internet has made it ever more difficult to exclude access to them; perfect copies can easily be made and distributed for simultaneous use by multiple individuals across great distances. While more conventional business models of the twentieth century were predicated on the excludability and rivalrous nature of consumer products, the Internet has largely turned digitized products, such as music, into both a nonexclusive and nonrivalrous good (cf. Leyshon 2001). This has led to significant disputes over intellectual property rights with a spiraling set of technologies and counter-technologies competing to permit or prevent copyright holders from making digital commodities into excludable and rivalrous goods. In parallel, new businesses, such as Google and Spotify, have emerged in which consumers no longer pay directly for digital products with money, but do so by forfeiting the right to data about themselves and their online interactions, which is then used to serve targeted advertisements to these potential consumers. The growing importance of social data as a product for resale has led to a growth in user-generated content through websites like Facebook, Twitter, Flickr, and Instagram, which all are free for users, but which for commercial viability also rely on the sale of data generated by these users to third parties.

In addition to new digital products, the Internet has engendered novel ways of channeling consumption through a range of digital services. Though the aforementioned social media platforms have been important commercial actors in their own right, other platforms attempt to more directly tie the virtual world of the Internet and social media to the material world, especially
INTERNET AND GLOBAL CAPITALISM

through consumption. While Google Maps provides spatial search and wayfinding services, its algorithms also have the potential to “re-route” users to particular kinds of locations based on a number of more-or-less opaque factors, perhaps leaving some places undiscovered by anyone without deep, place-based knowledge (Zook and Graham 2007). Similarly, Foursquare provides social and monetary incentives for individuals to record and annotate their daily spatial practices of consumption, such as the offer of a discount for habitually “checking-in” at one’s favorite coffee shop. One important recent development is Facebook’s purchase of Oculus VR, a maker of virtual reality headsets most commonly used for gaming. This investment of $2 billion is indicative of the importance of linking these virtual experiences and interactions to the material world and everyday life. As Facebook notes in their press release about acquiring Oculus, “virtual reality technology is a strong candidate to emerge as the next social and communications platform,” highlighting the centrality of digital mediation to consumption in the economy.

The impact of the Internet on consumption is not limited to these highly visible examples, however. Indeed, online retail, or “e-tailing” as it was referred to in its early manifestations, is transformational precisely in so far as it has receded from our view and seamlessly melded into the daily routines of so many individuals. From online travel websites such as Expedia and Priceline to multipurpose retailing websites such as Amazon, consumers now have easy access to a wide variety of products via quick shipping options. Brick and mortar stores remain, but online consumption channels are solidly established in our economies. Moreover, some of these channels have introduced new ways of widely distributing material products at reduced volumes making smaller scale operations commercially viable. This can be seen in the growth of new sales channels such as the craft website Etsy, in which artists and artisans can access new and distant markets, or AirBnB, which allows one to rent out one’s home as if it were a hotel room. In short, the Internet has enabled a range of new products, services, and pathways for consumption to emerge within the economy.

Internet-enabled changes in labor geographies

Alongside these new products and services have come important changes in the practices and geographies of labor. Initially, the preoccupation with the Internet’s effects on work and labor tended to focus on two key ideas. First was the notion that the introduction of new ICTs would necessarily lead to the deskilling of labor and loss of jobs, as computers automated many routine tasks performed by unskilled information processing workers. The second important idea was the now somewhat mundane notion of “telecommuting” or the ability of workers to work from home due to the remote accessibility enabled by an Internet connection. In other words, the Internet was expected to render co-presence obsolete, as companies and workers alike could save time and money, avoiding lengthy commutes, unnecessary capital outlays for office space, or other things by communicating online. While there has been deskilling and telecommuting, the Internet’s effects on labor and its geographies are much more complex. New forms and definitions of labor have emerged alongside deskilling, while spatial practices of labor have evolved to take advantage of these emerging forms of work and organization.

One of the key shifts associated with Internet-enabled labor practices has been the emergence of crowdsourcing, or a kind of
“volunteered” labor associated with the production of user-generated content without financial compensation. Ritzer and Jurgenson (2010) identify this process as the rise of “prosumption” on the Internet, or the increasing tendency for production and consumption of online information to be indistinguishable, occurring simultaneously. They argue that this process is characterized not only by free labor and free products, but also by an economy of abundance, rather than scarcity. They point to the potential for prosumption to act as an alternative to conventional capitalist social relations, especially around the control of exploitation of labor by capital. Rey (2012) furthers this argument, likening Facebook to Marx’s factories as the key site of exploitation under informational capitalism, whereby capitalists extract surplus value from the unpaid labor of Facebook users.

While the prevalence of unpaid labor is increasingly important for an economy where information is a primary commodity, the Internet also has allowed for the expansion of systems of paid informational labor outside of more conventional employment relationships. One of the key examples of this is the rise of “microwork,” such as that offered by Amazon’s Mechanical Turk service. This platform connects employers with workers for short-term information processing tasks, such as identifying typos in a sentence or comparing the objects in a set of images, which tend to pay just one or two cents per task. Despite the low average pay of $2.30 an hour – enabled by Amazon’s classification of workers as independent contractors – 15% of the Mechanical Turk workers living in the United States (which accounted for 56% of the global Mechanical Turk workforce in 2009) rely upon this income “sometimes or always” in order “to make basic ends meet” (Ross et al. 2010).

Alongside informational microwork, there is the process of offshoring, or the outsourcing of jobs from developed economies to developing countries in order to profit from decreased labor costs and lower burdens of government regulation. While often discussed in terms of manufacturing employment, a number of informational occupations, including call centers and customer service, also have been affected (Gereffi 2006). While offshoring of routine and even skilled informational jobs is ongoing, Zook and Samers (2010) argue that the geographies of outsourcing are also taking the form of “nearshoring,” “onshoring,” or “homeshoring,” indications that although the Internet can allow for coordination and communication across great distances, social relations remain incredibly place bound, whether by language or cultural similarities, affecting a business’s wellbeing. For instance, countries such as India, the Philippines, and South Africa have all emerged as key localities for call center outsourcing catering to English-speaking countries, while Spanish-speaking call centers tend to locate in Latin America, and German-language call centers in Turkey.

In short, it is important to embed these changes within broader histories and cultural specificities that create a significant amount of internal differentiation between places. While the reduction in labor costs remains an important consideration for firms wishing to outsource any type of job, there is no single set of criteria by which these decisions are made, with different locations offering a variety of advantages depending on their particular relationships with other spatially distant locations. Similarly, just as no places are affected in the same way by these structural changes, the kinds of changes wrought by the Internet do not apply equally to all kinds of labor. While some jobs have been entirely automated by computers, outsourced to other countries or continents, or broken into a series of microtasks, other opportunities have emerged, particularly for what Castells (1996)
characterizes as self-programmable labor, where workers are able to quickly adapt to volatile employment opportunities.

New forms of manufacturing in the Internet era

The influence of the Internet on labor practices is not confined to the use and dissemination of information. The Internet also has important impacts on how material goods are produced. From robotics to computer-aided design (CAD), enabling designs to be produced in one location and easily transferred to computer numerical control (CNC) tools in another, the Internet has altered the kinds of work performed in manufacturing material goods, providing an important role for information processing and transfer even in this sector. CNC systems are central to flexible manufacturing, permitting the creation of “reconfigurable manufacturing systems” which “allow adding, removing, or modifying specific process capabilities, controls, software, or machine structure to adjust production capacity in response to changing market demands or technologies” (Mehrabi, Ulsoy, and Koren 2000, 404).

These techniques of flexible manufacturing have further evolved with the growth of 3-D printing technologies in which solid objects based on CAD specifications are created by precisely depositing material – often a plastic but potentially metals or other materials – in a series of layers, a feature that is particularly useful in prototyping. While traditional mass production techniques produce lower cost (and generally higher quality) products, 3-D printing is beginning to be used by a range of corporations and likely will become even more useful as the technology advances to include the ability to print electrical circuits directly into objects.

While the high costs of 3-D printing remain its greatest challenge, some business scholars highlight its potential to support a range of “mobile manufacturing” strategies (Stillström and Jackson 2007, 188). For example, the start-up company Mebotics has developed a desktop system which is marketed as “the world’s first machine shop in a box.” One of the founders of the company sketches out a scenario of being in a remote location but needing a spare part; access to a 3-D printing system and the Internet would allow the user to acquire printing specifications and create a replacement part. Proponents of 3-D printing, moreover, argue that normal consumers, even with access to standard retail outlets, will eventually find it economical to print out a range of simple consumer products using open-source CAD files for objects at online repositories.

The necessity of CAD files (open source or proprietary) highlights ongoing concerns about intellectual property that are associated with the Internet. Just as the music industry lost the ability to monetize product through physical media (records, CDs) as digital files enabled perfect copies of music, manufactured goods could face the same threat. Once CAD source files are available on the Internet, then anyone with a 3-D printer could create a material copy. One could easily imagine groups of activists creating high quality CAD files of desirable (yet proprietary) goods – jewelry, silverware, and so on – and making them available via bit torrent systems – much in the same way that copyrighted videos are currently shared – to create open (albeit illegal) production of material goods. This potential strikes directly at a range of design and creativity-based economies and economic development policies; if design-intensive items can be scanned and available to anyone to download, this potentially threatens the value
that designers and corporations can capture (as in the recording industry).

**Transformation of finance and logistics**

The Internet has been equally important in introducing new spatial and organizational dynamics into service industries such as finance and logistics. While one focuses on money flows and the other on material goods, both are fundamentally dependent on the flows of information through the material infrastructures of the Internet.

**Finance**

In many ways, the financial sector has been the area of the economy most fundamentally impacted by the growth of the Internet and other digital technologies. Indeed, it was the intertwining of the Internet and finance that led to early pronouncements of the “end of geography,” (O’Brien 1992). Not only has finance grown to become one of the most important parts of the global economy, but it also has begun to shift away from its historical function of providing capital investments for new business or trade ventures. The financial sector has increasingly focused on the trading of a range of complex financial instruments or derivatives that are loosely, if at all, connected to material assets or the competitive prospects of a given firm.

The Internet and associated information technologies underlie many of these new products and practices, as they could simply not exist to any meaningful extent without computers and standardized global communication networks. Like the Internet’s intertwining in other processes, the importance of these technologies is often hidden from view for all but those experts who experience these practices directly. For example, one of the key mechanisms contributing to the 2008 financial crisis were instruments known as collateralized debt obligations (CDOs), which combined hundreds of income streams deriving from hundreds or thousands of home mortgages. Each income stream did not represent an entire mortgage but instead was a slice or tranche of the payment prorated in various ways. More expensive tranches were drawn from lower risk mortgages and/or were guaranteed first payment rights. The goal and claim of makers of these CDOs was that this provided a reliable way to mitigate risk, but events quickly proved that complexity was no barrier to financial hazard. While dividing, bundling, and sharing risk is a long-standing and important practice within the financial sector, the sheer magnitude of these practices associated with the mortgage crisis was only possible with digital technologies that are seemingly able to make risk disappear.

The financial industry has also developed new practices that fundamentally rely on the Internet. For example, during the past 20 years trading at stock exchanges has shifted from actual trading floors to server farms in which computer algorithms conduct tens of thousands of trades every second. This phenomenon, known as high-frequency trading (HFT), works to compress time–space at very small scales (milliseconds and microns) to construct informational advantages vis-à-vis other traders. One of the key techniques used in HFT is to identify small price differences in pricing for the same or similar stocks in different (and physically distant) trading venues. With this information, trading algorithms execute buy and sell orders across hundreds of kilometers via an increasingly sophisticated array of networks. This has involved the laying of new fiber–optic networks between major trading centers, such as between Chicago and New York, as well as setting up a series of microwave and laser transmitters, simply in order...
to ensure that an HFT buy or sell order arrives a few thousandths of a second before those of one’s competitors (MacKenzie et al. 2012).

While HFT tends to be out of view for most individuals, a more common way that the Internet has influenced finance is through the emergence of virtual currencies associated with virtual worlds and online social networks, such as Linden Dollars in Second Life or the attempt to create a completely decentralized currency, Bitcoin. While these currencies currently suffer from issues of stability, liquidity, and uncertainty of legal obligations, they offer an intriguing glimpse of what may be possible in a global information economy in which software and networks allow global collections of users (versus states or large financial firms) to literally create money. Equally interesting are the political values associated with these currencies; in the case of Bitcoin a type of algorithmic libertarianism poses behind a stance of apolitical objectivism. Whether the specific Bitcoin project succeeds or not is less relevant than the individualization of empowerment – the ability to design or manipulate software (either directly or via proxy) – that allows some individual actors to undertake political and economic projects at scales and scopes that previously were not possible. In short, the power of the networked individual is a fundamental characteristic of global capitalism in the age of the Internet.

Logistics and global production networks

The virtual spaces of the Internet have become more and more intertwined with the flow of material goods around the globe. Just as the near-instantaneous global information flow has allowed for high-frequency trading and other new financial products, these same informational infrastructures are put to use in tracking and changing the way material products move through space on a global scale (Coe, Dicken, and Hess 2008).

While spatial structures in the economy have long been driven by transportation networks and population, new IT-enabled logistics has allowed for new forms of competitive advantage for firms and localities across the globe, with firms increasingly outsourcing their logistics operations to specialized third-party logistics providers who concentrate on the collection and analysis of information about products, rather than on their design and manufacture. Rather than focusing on building competitive advantage solely within the production process, new logistics services allow firms to reduce the cost of moving goods and maintaining inventory, as pioneered by the just-in-time manufacturing systems of Toyota, or the cross-docking restocking of the American retail chain Walmart.

With easily accessible information and the ability to reroute flows of goods relatively easily, inputs for products or final goods can be “stored in transit” rather than warehoused, thus reducing the necessary capital outlays for production or retail stocks. The augmentation of these material products and the vehicles on which they move with radio-frequency identification (RFID) chips and global positioning systems (GPS) receivers allow for individualized tracking of products, even across great distances. This facilitates the further spatial extension of supply chains, as important factors such as trust and tacit knowledge become codified through digital information. It has been argued that contemporary Internet-enabled logistics services represent an unprecedented merging of transportation and information industries, making the movement of material goods through space a virtual phenomenon as much as a material one (Aoyama and Ratick 2007).

The digital augmentation of supply chains can also be produced through new forms of
infrastructural systems. Space-constrained ports can use their knowledge of distanced supply chains to better manage the flow of goods in and out of the port, increasing the efficiency of transportation. As one IBM promotional video on the baggage handling system at Amsterdam Airport states, “We can’t make the airport bigger. The physical footprint is limited by highways, cities and villages. We must make more capacity available by making the system smarter … the intelligent software creates space where there was no space before.” Although couched in advertising rhetoric, this quote is indicative of the broader trend of using information flows to overcome the constraints of existing physical infrastructures.

While the use of the Internet in managing supply chains has stretched economic relations across greater social and spatial distances, it has not eliminated the massive inequalities that exist within the global economy. For example, Graham (2011) has shown that despite the expectations for the Internet disintermediating supply chains, that is, allowing for more direct relationships between producers and consumers, evidence suggests that there is little change in the organization of many sectors, at least not in such a way as to alter more macro-level development outcomes. Moreover, examining small-scale manufacturers in Tanzania (which enjoy widespread use of information and communication technologies), Murphy (2013) shows that the benefits of these technologies tend to accrue primarily to extra-local actors who are already situated in dominant positions relative to local-scale businesspeople. That is, the Internet may offer the means for economic exploitation of marginal places and places in the information age, even as it opens up significant opportunities for others (Ya’u 2004). Even the production of commodities emblematic of the information economy, such as the iPhone, tends to reinforce these patterns.

While the iPhone’s supply chain is truly global, with manufacture of the phone alone stretching across China, South Korea, and Germany, most of the value produced in the process of creating the several-hundred-dollar device is captured by designers located at Apple’s headquarters in California, not by workers assembling the devices. So while Internet-enabled logistics are reshaping parts of the sociospatial system, this seems to have done little to alter the fundamental economic structures that influence everyday life in the Global South (Carmody 2012).

Conclusion

The Internet has undoubtedly played an important role in the evolution of global capitalism over the last two decades. While the changes are significant, they are best characterized as a reconfiguration or intensification of existing structures and processes, rather than a wholesale creation of new forms of economic organization that are somehow qualitatively distinct from previous eras. The examples outlined here highlight the complex and contradictory role of the Internet in reconfiguring flows of products, labor, and capital in the global economy. New products and consumption channels have emerged just as labor and production practices have evolved in response to the availability of the Internet, bridging the seeming divide between the virtual and material.

The Internet has facilitated new uses for old spaces and enabled new forms of capital investment, from fiber-optic cable infrastructure to high-frequency trading to the commodification of personal data. Ultimately, the sociospatial reconfigurations of the global economy wrought by global networks of information technologies are remolding the terrain on which capitalism operates. While this has brought many changes
INTERNET AND GLOBAL CAPITALISM

to our everyday lives and to the processes through which the economy is constructed, capitalism remains the dominant form of economic organization worldwide, even in the age of the Internet and notwithstanding the notion that the Internet supposedly changes “everything.”

SEE ALSO: Digital divide; Information and communications technology; Information technology and mobility; Local/global production systems; Logistics

References


Interoperability of representations

Boyan Brodaric
Geological Survey of Canada

Interoperability is an essential requirement in complex digital environments where multiple geographical information sources must work together. For example, an online hotel booking system might require a mapping component to work with payment and reservation systems, and the mapping component itself might access a variety of online sources to contextualize the location, such as nearby tourist attractions, restaurants, or other amenities. Often these components are distributed geographically, are heterogeneous in their deployed technologies, data structures, and data content, and are managed autonomously by distinct agents with diverse goals, policies, rules, and regulations. These differences must be overcome in order for the systems to work together cohesively and become interoperable.

Interoperability defined

Interoperability is defined variously. All definitions include the notion of diverse autonomous entities working together to achieve some common goal in a particular setting. Working together in this sense requires some or all of the following: cooperation, compatibility, unification, coupling, sharing, exchange, negotiation, and communication (Manso and Wachowicz 2009). It also usually carries a quality connotation, as in working together smoothly rather than laboriously. The definitions also vary as to whether interoperability is (i) a measure of how well entities work together in the attainment of a common goal (e.g., less or more interoperable), (ii) a state-of-affairs describing a working functional relationship between entities (e.g., several components continuously interacting to read and analyze environmental conditions), or (iii) a similar ability hosted by participating entities (e.g., being able to understand a specific, often standardized, language or to exchange specific data). The challenge common to these various orientations involves overcoming differences, that is, heterogeneities, among the interacting entities to ensure their unified functioning. Indeed, heterogeneity is an essential characteristic of the ecosystem composed of the interacting entities: without heterogeneity they could clearly interoperate without further effort. The interacting entities could be natural or artificial, such as humans or machines, respectively, making interoperability per se a very broad notion. An interoperability example among humans is a teleconference, which could prove challenging due to the use of different languages, and to the need to avoid speaking at the same time. An example of machine interoperability involves the integration of multiple distributed heterogeneous datasets, or the synchronized operation of diverse software running on different machines.

Geographical interoperability occurs when the attainment of a common goal requires interaction with geographical entities — these are entities that occupy some space, or unfold in it, such as landforms, buildings, or earthquakes. The occupied space need not be physical — it
might be, for example, abstract, emotional, social, or representational (e.g., digital) – but it is necessarily structured such that pertinent entities can occupy locations in the space and partake in relations or actions within it. Geographical interoperability in this sense can be more complex than interoperability between nonspatial entities, say financial accounts, due to the wider range of heterogeneity resulting from the additional spatial aspects. Examples of geographical interoperability are numerous, and include such things as the use of multiple blueprints to construct an edifice, the coupling of air traffic or road systems, or any other environment in which spatial aspects influence the interaction of the entities. By far the most prevalent form is information interoperability: the working together of digital representations of geographical entities that typically reside in various kinds of geospatial databases, geographical information systems (GIS), or web pages.

Frameworks for interoperability
(i) – general

Frameworks for interoperability, including geographical information interoperability, are numerous and follow from the various definitions of interoperability (Manso and Wachowicz 2009). Each is typically portrayed as a stack of stratified and nested levels. Thus measurement, or “maturity,” stacks distinguish levels according to how well entities, typically organizations, function together or adapt to changes in computational environments. States-of-affairs stacks delineate levels according to relationships between types of components, such as the existence or degree of connectivity between databases or software. Ability stacks differentiate levels according to the type of transformation required to align entities, such as syntactic or semantic transformation. However, an apparent shortcoming of these and other stacks is their lack of a comprehensive theoretical framework. In contrast, a broader approach depicts interoperability as a communication process, in which messages are necessarily exchanged between entities (Brodeur et al. 2003). This reliance on message passing is not limited to information interoperability singularly, as it applies to all forms of interoperability: fundamentally, in this paradigm, entities could not interoperate without communicating, and this implies the transmission of messages and the shared use of some common language.

Why representations?

Representations are at the core of interoperability in the communication paradigm, not only due to the messages passed between interoperating entities, which are in fact representations external to the entities, but also due to the internal representations of these messages maintained by senders and receivers: it is the sender’s internal representation that is encoded into an external message, which is then decoded into the receiver’s internal representation. In general, two types of messages are exchanged: (i) control messages that influence how entities interoperate, such as error messages or requests for data or software, and (ii) content messages containing actual subject matter, such as geographical data or software. For example, a teleconference participant can signal a desire to speak by coughing or asking permission, and can then supply relevant discourse in one language or another. Analogously, an information system in a federated data network might use Representational State Transfer (RESTful) principles for exchanging messages, and might require data content to be structured as objects and vectors.
As becomes quickly evident, the main obstacle to interoperability in this paradigm is the meaning challenge: whether receivers can decode a message to understand the sender’s intended meaning, given the almost inevitable variability among representations. Achieving understanding requires, at a minimum, communicators to share an external representation, which enables them to encode and decode a common language that is rich enough to transmit the intended meaning. This becomes more complex when the number of communicators increases, as it almost invariably amplifies the number of distinct internal and external representations within the communication community. The effect of this increase is significant: it can exponentially increase the total number of decodings within a community, in order to enable a member to dialog with any other member. It can also increase the chances of having severe differences between representations, such as different languages deployed by telecom participants, or different communication interfaces and data structures deployed by interacting information systems. Overcoming the meaning challenge, in general, then devolves to the problem of overcoming differences, or heterogeneity, between representations by translating one representation to another, ideally with minimum meaning loss.

Frameworks for interoperability (ii) – theoretical

A significant advantage of the communication paradigm for interoperability is its theoretical foundation, namely semiotics. Semiotics provides not only a theory for describing the representation and interaction of messages – called signs – but also enables description of the heterogeneities inherent in sign interaction. At its core is the structure of a sign, which is dominated by Peirce’s formulation in the context of information interoperability. A sign has four components: (i) an object that is being represented, (ii) a symbol, which is an external representation of the object, (iii) a concept, that is an internal representation, and interpretation, of the object and symbol, and (iv) an agent that hosts the internal representation, and that either produced or consumed (interpreted) the external representation. A unit of meaning then consists of the four-way relationship between these components, such that a particular symbol stands for a particular object as produced/consumed via a certain concept by a specific agent. It is often depicted as a triangle consisting of object, symbol, and concept nodes, with the agent placed nearby.

The sign, as a semiotic unit of meaning, is further delineated by Morris (1938) into the three semiotic dimensions of syntax, semantics, and pragmatics. Syntax refers to the relation between a sign and its symbol, and encompasses the formal structure of a sign, such as the grammar of a language or structure of a message. Semantics refers to the relation between a sign and the object being represented, and encompasses the conceptual and referential structure of a sign, for example, as famously noted by Frege, that “morning star” and “evening star” are different signs, hence meanings, referring to the planet Venus. Pragmatics refers to the relation between sign and agent, and encompasses contextual aspects such as how signs are created and used, and their related effects – for example, the varied interpretations of some body gestures among cultures, or the context-dependent interpretations of natural language indexical terms such as “it” or “them.” Understanding, that is, decoding, a message can subsequently be achieved by the progressive understanding of its syntax, semantics, and pragmatics, as widely noted by linguists – though some cognitive linguists blur the line between semantics and pragmatics,
particularly for human communication. As a side effect, the three semiotic dimensions are then also often viewed as broad categories for associated heterogeneities.

For example, consider an information integration use-case in which an online system, for example, for hotel reservation, is merging information about a specific geographic location from the DBpedia and GeoNames web databases. Syntax differences occur when the structure of messages vary for the same information, due to inherent differences in the external languages, that is, HTML (Hypertext Markup Language) and RDF/XML (Resource Description Framework/Extensible Markup Language). Semantic differences occur when the content of the messages varies – for example, when each has unique properties not included in the other, such as a ‘photo collection’ (DBpedia) or a list of ‘nearby features’ (GeoNames). Pragmatic differences are also apparent, as the HTML format is primarily intended for use by web browsers, but the RDF/XML format is largely geared for use by reasoning engines. To overcome these differences, the online system must create its own internal sign, typically a composite of the external signs normalized to a preferred option along each semiotic dimension.

More generally, while the example just described above illustrates a blending scenario, in which an agent consumes and amalgamates multiple signs, it equally applies to communication scenarios in which an agent consumes or produces a single sign, or acts as a producer of many signs for external consumption. Heterogeneities in all three semiotic dimensions can, and regularly do, then exist for the sign blending and communication scenarios, because in both scenarios the agent is transforming signs either to or from internal representations.

The three semiotic dimensions (syntax, semantics, and pragmatics) and two sign interaction scenarios (blending and communication) also apply to the other major information interoperability use-case of software orchestration. This use-case includes any processing environment in which software components are coupled to attain a specific goal. It requires agents, such as GIS, to coordinate actions by exchanging control messages, the most significant of which conform to declarative templates, called signatures, describing interfaces to the actions. Heterogeneity exists between the internal representation of the message-producing agent, who is requesting the action, and the external signature of the message-consuming agent, who is carrying out the action. Syntax heterogeneity is then related to the language and structure of the signature, semantic heterogeneity is related to the concepts inherent in the signature, and pragmatic heterogeneity is related to the information required to enable the software components to function together, such as the sequence and validity of connections and the reactions to monitored events, as well as the criteria for overall fitness for use. The orchestration use-case clearly exemplifies the communication scenario, as control messages are sent to and from a producer, and can also exemplify the blending scenario: though control messages are rarely integrated into one composite message, even if arriving simultaneously, signatures often do refer to information inputs that might in fact require blending from multiple sources prior to processing by the receiving agent.

Frameworks for interoperability (iii) – geographical information

Prominent frameworks for information interoperability, or related information system design,
include the Semantic Web (Horrocks et al. 2005), database interoperability (Sheth 1999), simulation interoperability (Tolk, Diallo, and Turnitsa 2007), and computational semiotics (Stamper et al. 2000). As shown in Figure 1, these frameworks are organized into nested ability stacks, in which the semiotic dimensions are further subdivided within the syntax and pragmatic layers, and a new bottom layer is universally added, here called systems following the database interoperability stack. While domain-neutral, all these stacks can be applied to geographical information interoperability – for example, the database interoperability stack has been applied to GIS interoperability (Bishr 1998; Sheth 1999), the simulation interoperability stack has been adapted as a framework for Spatial Data Infrastructures (SDIs) (Manso, Wachowicz, and Bernabé 2009), and the Semantic Web stack has been implicitly applied to geospatial Linked Data (Kuhn, Kauppinen, and Janowicz 2014).

**Systems**

Semiotically, systems can be seen as the relationship between a message (sign) and the media carrying it, though this distinction is an artifact of later information theory and not part of original semiotic formulations. In information interoperability, media encompasses both the hardware and the software infrastructure required for message passing. For instance, in the Semantic Web stack it generally refers to all aspects of web infrastructure, and significantly to the universal resource identifier (URI), which is required to identify and possibly navigate to a web-based entity called a resource; it also includes the alphabet used to encode URIs, and the hyper-text transfer protocol (HTTP) for transmitting messages. In the database and simulation stacks it refers to computing foundations, such as operating systems or database management systems, including specific GIS. In computational semiotics the physical and empirical nature of the media is emphasized: the physics subdivision
INTEROPERABILITY OF REPRESENTATIONS

refers to the physical aspect of the media, such as hardware infrastructure, including connectivity, while empirics refers to how signals are organized within the physical media, including tendencies such as entropy. As in the Semantic Web, the latter includes the organization of signal patterns into alphabets, though it is possible to argue that the abstract identity of a character belongs to the syntax layer, while its digital encoding as a specific character set, or as a digital pulse in a communication channel, belongs to the system level. Regardless, alignment of systems between entities is required for their interoperation – for example, different operating systems and character encodings, or perhaps even the different size constraints imposed by GIS on the representation of geometry types, such as the maximum number of nodes or storage limits for a line or polygon.

Syntax

Syntax refers to components of the language used to encode a message, including the words, the grammar, and possibly the alphabet. Semiotically, this is understood as the relation between a message (sign) and its symbol. Syntaxes can be abstract or concrete; for example, RDF has an abstract grammar consisting of subject–predicate–object triples arranged in nested graphs, and concrete syntaxes such as RDF/XML, N3 (Notation3), or others, that variously encode the abstract grammar. Syntax heterogeneities then concern differences in abstract or concrete language aspects. In geographical information interoperability, they further encompass iconic languages, for example, to overcome differences in cartographic presentations.

In the geographical domain, iconic standards have been developed by various communities, such as in transportation, facilities, geology, ecology, or soils, to normalize the presentation of specific geographical entities. Likewise, the Open Geospatial Consortium (OGC) and International Standards Organization (ISO) standards bodies have developed the textual Geographical Markup Language (GML), an application of XML for the standard encoding of geographical entities. Similarly, the Semantic Web stack includes standard XML applications such as RDF/XML and Web Ontology Language (OWL). For query languages, the Semantic Web community widely uses SPARQL (SPARQL protocol and RDF query language), while its geospatial extension is GeoSPARQL. In terms of interfaces for online software components – that is, web services – the Semantic Web community is increasingly migrating to RESTful principles, which reuse a simple syntax for web operations, while the OGC and ISO provide a fixed terminology for all its web service interfaces, and is selectively adding RESTful versions. Note the Semantic Web and data interoperability stacks exclude configuration aspects, which are found instead at the structure level.

Structure

Structure refers to the configuration of the parts of a message. Semiotically, message structure is known as the syntagmatic relation between the elements of a message, associated primarily with the semiotic tradition of Saussure. It is explicitly differentiated in the database interoperability and Semantic Web stacks, and merged with syntax in the remaining stacks. The database interoperability stack differentiates meta, conceptual, logical, and physical schemas, following database design principles, such that interoperability requires alignment over each schema type. Meta-schemas define the components in a schema, such as entities and relations, conceptual schemas adapt a meta-schema and describe a domain in a technology-neutral manner, logical schemas
adapt a conceptual model according to a specific technologic style, and physical schemas adapt a logical model to a specific technologic system. Structural heterogeneity evident in physical encodings can thus be the result of differences in any of the related schemas – this can be seen as heterogeneity in propositional structure. Another category of structural heterogeneity consists of differences in spatial structure, such as varied representations for the geometric shape or geographic location of a geographical entity.

In support of geographic data interoperability, the OGC and ISO have developed a meta-schema as well as frameworks for logical and physical schema, but lack formal mechanisms for conceptual schemas. The meta-schema describes the structure of a geographic feature type (General Feature Model, ISO 19110), which can be implemented in an ad hoc conceptual schema and/or in GML to enable development of a logical and physical schema within a domain, such as in geology (GeoSciML), hydrology (WaterML), and city infrastructure (CityML), and across domains, such as Observations & Measurements (O&M) and metadata. In contrast, the Semantic Web relinquishes predefined and fixed schemas altogether. Feature descriptions in the Semantic Web can evolve over time and vary across different manifestations at the same time, as well as be linked to a normative-type definition. This flexibility is particularly useful in open environments, such as the web, where structure is inherently unpredictable. It is less desirable in closed environments, where a predictable and consistent structure is often preferred, such as in some scientific data transfer scenarios, or in urgent care or emergency response situations, where, for example, missing information could be crucial and sequence could play a role. These drawbacks are overcome in the Semantic Web via the semantic and pragmatic levels, where ontologies can regulate message content, and practice conventions can guide usage and other structural aspects such as sequence.

The OGC and ISO also provide a standard fixed structure for each web service interface. Key web services include those that access data such as maps (web mapping service), features (web feature service), fields (web coverage service), observations (sensor observation service), and metadata (catalog service), and that invoke geospatial processing software (web processing service). Each web service is self-descriptive, allowing the discovery of supported functional components and representational schema. The structure of these OGC web service interfaces diverges from the Web Service Description Language (WDSL) industry standard, which has been used by many non-geospatial web services, and from the recently popular RESTful principles, used particularly for simple applications where the structure of response data can be relatively easily determined without formal self-description. However, a lack of canonical mechanisms for self-description in the response can be a deterrent to interoperability when the response structure is complex and not readily discernible.

Semantics

Semantics refers to the meaning inherent in the content of a message. In terms of semiotics, it can be understood as the relation between a message (sign) and an object. Classical semantics is divided into extensional and intensional varieties. In extensional semantics, meaning is referential, and focuses on the relation between symbol and object: an agent interprets a message by directly referring to the objects it represents. For example, an extensional interpretation of “lake” (a symbol) is the collection of all lakes in a domain (some objects), where a domain consists
INTEROPERABILITY OF REPRESENTATIONS

of the entities of interest, such as the geographical features found in a geographic area or in a particular GIS. Extensional collections are called models, or extensions. In classical extensional semantics, the extension of a name, or literal constant, such as “Lake Ontario,” is the collection of all lakes with that name. The extension of an n-ary relation predicate is the collection of actual n-ary relations in the domain; for example, Lake Ontario might be in the extension of the unary “lake” predicate, and the pair Lake Ontario and Niagara River might be in the extension for the “flows-into” binary predicate. Lastly, the extension of a propositional sentence is a truth value: the extension of “Ontario is a lake” is true if Lake Ontario exists in the domain. However, associating semantic meaning with extensions can be problematic in several respects, for example, with nonextensional symbols that are difficult to match to a reference, such as those concerned with necessity, agent beliefs, and hypothetical and abstract entities. Also, most notably, an extension can be shared by messages with different meanings; for example, “cold lake” and “deep lake” could refer to the same collection of lakes in a domain, causing them to be interpreted as having identical meaning in extensional semantics.

To overcome these and other problems, intensional semantics focus on properties rather than extensions: “cold lake” and “deep lake” then differ because they carry different properties, which comprise their intensions. Intensional semantics can in this way be seen to focus on the semiotic relation between concept (i.e., intension) and object. Classical intensional semantics accomplishes this by making extensions relative to possible states-of-affairs (worlds) or their fragments (situations) in some domain. Thus if “cold lake” and “deep lake” have different extensions in some possible world or situation, such as in a distinct GIS as is conceivably the case, then their intensions are different. However, this says very little about the nature of an intension, particularly regarding its metaphysical status and the properties that comprise its content.

- Metaphysically, intensions in cognitive semantics are viewed as agent-dependent internal constructs that can be shared to a point, in the sense that very similar constructs can be built internally by agents when they possess similar cognitive hardware and undergo similar experiences. In contrast, in realist semantics, intensions are agent-independent invariants in reality, so-called universals, unmediated by agents but perceived similarly because they exist invariably in reality: for example, scientific laws that are discovered and verified independently by scientists.

- In terms of content, classical intensions are functions that return an extension for some symbol, relative to a specific world or situation. In effect, they serve as a meaning waypoint in the path from symbol to object, and this mediated path serves as an intensional interpretation of the symbol. They are typically fleshed out, in this tradition, using necessary and sufficient truth conditions, and associated reasoning mechanisms. Necessary conditions are true for every member of the extension, while sufficient conditions allow an entity to be placed into an extension. For example, an Aristotelean approach to intensions involves specifying a genus and differentiating conditions – for example, ‘lake’ and ‘river’ intensions might possess a necessary and sufficient condition consisting of the ‘water body’ intension as their genus, and different ranges of ratios for surface width to length as their differentiating conditions. Thus, the symbols ‘Lake Ontario’ and ‘Niagara River’ can be placed into the...
‘lake’ and ‘river’ extensions, respectively, due to their widths and lengths and because they are members of the water body extension.

Following Guarino, a computational ontology (henceforth ontology) is then a digital artifact that externally expresses a collection of intensions in some machine-readable language (Guarino, Oberle, and Staab 2009). Its contents are often influenced by ideas from philosophical ontology, which is the study of what exists. Note the two are very different: the former is a digital artifact used for computing, and the latter is a discipline of study. An ontology is also distinguished from a conceptualization by their diverse expressions: the former is externally expressed, while the latter is not externally expressed, but is an artifact of some metaphysical commitment such as an internal agent representation or invariant in reality (possibly internally represented by an agent). However, the two are linked when an ontology is defined as the external expression of such a conceptualization. It is then possible for an ontology to vary from the conceptualization that it is expressing: a good ontology will have an extension proximal to that of the conceptualization, and a bad one will have one that is distant. Factors such as the selection of the domain, expressivity of the language, and detail of expression will all govern this proximity. Ontologies in practice can then take many forms; for example, conceptual models for various schema, as well as glossaries, data dictionaries, vocabularies, and thesauri, might be considered ontologies in some sense, in addition to the logical theories so prominent in the Semantic Web. Opinions vary as to which of these forms are, and are not, an ontology, and how they differ from an ontology, with the criteria typically including the degree of language expressivity, the generality of the intensions, and their run-time availability.

Ontologies also vary in generality, prominently stratified by Guarino into foundational, domain, and application variants (Guarino 1998): foundational ontologies apply across all domains and include concepts such as physical object or process; domain ontologies apply within a specific domain (e.g., the “lake” concept could be applied in a geographical or hydrological domain, but perhaps not in the domain of financial management); finally, application ontologies apply in specific situations, such as the concepts referenced in a particular GIS (e.g., “deep lakes in Ontario”). Somewhat orthogonal to this stratification are ontology design patterns (ODP); these are essentially self-contained ontology fragments or replicable patterns, from any level of generality. ODP are primarily concerned with ontology modularity to minimize the complexity of the fragment or pattern, as well as minimize their dependence on other ontologies while maximizing re-use.

Interoperability with intensional semantics clearly exemplifies the communication paradigm for interoperability, because understanding the meaning of a message involves translating the ontologies between the sender or receiver of a message (their internal representation) and the message itself (an external representation). Translation of ontologies typically requires the following steps.

- **Semantic enrichment** involves attaching an ontology to an artifact, such as to some data, a GIS database schema, or a web service signature. This can involve discovering the ontology, if unspecified, using a variety of data-driven or knowledge-driven methods.
- **Semantic matching** involves determining the similarity between concepts in ontologies, to identify equivalences and differences, and expressing these as a mapping in some language. As with semantic enrichment, there
exist many methods for semantic matching in the geographical domain, including lexical, statistical, and knowledge-driven, as well as a variety of mapping languages to capture similarities and differences.

- **Semantic integration** involves combining ontologies, once matched, into a unified conceptualization and representation, again using a variety of approaches, such as the use of a bridge ontology to connect diverse ontologies, or the use of a common ontology that is specialized by the diverse ontologies.

Geographical information interoperability benefits from both intensional and extensional aspects, primarily because geographical location can provide a useful extensional reference for semantic matching. For example, while distinguishing a cold lake from a deep lake primarily requires knowledge of intensions, matching the “lake” concept in one data source with a “lake” concept in another could be aided by their extensions: if similarly located objects with some shared properties exist in both extensions, then the intensions are likely similar in comparable contexts. More generally, extensions can provide facts that help to assess the logical consistency of an ontology and its implied inferences.

Intensional semantics, as realized in logic-based ontologies, does however have shortcomings related to meaning representation, and thus interoperability. In particular, the necessary and sufficient conditions used to represent concepts in logic-based ontologies crisply determine whether an entity is a member of an extension, or not. In contrast, prototypical concepts do not have crisp boundaries as extension membership could be graded: some entities might be better examples of a concept than others. For example, a robin is a better example of a “bird” than a penguin, because flying is central, though not necessary nor sufficient, to the bird concept: not all birds fly, and not all that flies is a bird. Many geographical concepts have been shown to be prototypical in this sense, such as “mountain” or “lake.” This has motivated a variety of representations of prototypical concepts in the geographical domain, most notably building on Gardenfors’s (2000) theory of conceptual spaces, in which concepts are represented as portions of a space constructed from quality dimensions such as height or size, and where similarity to a prototype (a location in the space) can be measured. The key role of spatiality in such representations is further reflected in semantic reference systems (SRS), which use spatiality as a metaphor for an overarching theory for geographical semantics: effectively, if concepts are prototypical and thus spatial constructs, then a theory for their representation and manipulation can be formulated analogous to SRS (Kuhn and Raubal 2003). Various connections have been proposed between concepts as spatial constructs and traditional logic-based semantics, and these constitute a promising direction for semantics in geographical information interoperability.

In open geospatial standards, the OGC and ISO provide basic support for semantics. Geospatial feature types, such as definitions for road or river, comprise concepts that are formed by instantiating the OGC meta-schema. Each in turn can then serve as a coarse definition for a feature-type manifestation in a conceptual, logical, or physical schema, without, however, being linked to these schemas in any standard way. As a result, features in data streams, such as specific roads, cannot canonically refer to their type definitions within the data stream, which hinders interoperability. Likewise, categorical values in a data stream cannot also canonically link to a definition, though various physical schema are adopting best practices to achieve this, such as linking to simple knowledge organization system (SKOS) thesauri. Geospatial standards are
increasingly exploring Semantic Web approaches to overcome this obstacle, mainly because types and instances are tightly coupled therein, and moderately expressive languages such as RDF and OWL enable richer type definitions.

Semantic heterogeneities take many forms. As semantics is the relation between sign and object, semantic heterogeneities can be characterized as variations in the sign relations along the path to an object. Because intensional semantics emphasize the role of concepts along this path, then related heterogeneities involve concept relations within a sign, specifically the concept–symbol, concept–object, and concept–concept relations. Variations in the concept–symbol relation include polysemy and synonymy: polysemy occurs when a symbol has diverse intensions, such as “wood” meaning either ‘a piece of tree’ or ‘a collection of trees,’ and synonymy occurs when diverse symbols share an intension, such as when the English “lake” and French “lac” terms both refer to a common ‘lake’ concept. This type of heterogeneity is predominantly the focus of ethnophysiographic studies, which examine the relation between concepts and symbols in different human cultures and languages. Variations in the concept–object relation involve classification heterogeneity: an object can be diversely classified, such as a water body classified as a ‘lake’ or ‘pond,’ while an intension can variously classify objects in diverse databases, for example, a specific water body might be included in the extension for ‘lake’ in one domain, but not in another, due to differences in the description of the water body in the domains. Variations in the concept–concept relation occur when the same meaning has different intensions, such as “lake” defined with different width–length conditions in different ontologies. This also encompasses intensional differences due to variation in the underlying spatial structures associated with prototypes, conceptual spaces, and SRS. Lastly, extensional semantic heterogeneities involve the symbol–object, or reference, relation: a symbol might refer to diverse objects, such as “London” referring to a city either in Canada or in the United Kingdom, and an object might be a reference for various symbols, such as London referred to as “the location of the 2012 Summer Olympics” or “the capital city of England.”

Pragmatics

Pragmatics is defined semiotically as the relation between a message and an agent: how an agent creates or uses a message, or the effects of the message on the agent or others. These aspects are seen as contextual, because they are not part of the message itself, but they control action and reaction related to the message. For example, the concept of ‘recreational lake’ might not vary intensionally among agents, that is, its semantics are uniform, but whether a lake can be used recreationally, and thus be interoperable within that context, varies according to things such as agents’ pollution policies, seasonal restrictions, or even the occurrence of specific obstacles such as low water levels or debris congestion. In other cases, intensions might vary due to experience, such as when varying evidence causes scientists to create different intensions about the same geographical entity (Brodaric 2007) – as a result, experientialists see the semantics and pragmatics distinction as vague, because intensions and experience are so intermingled. Notwithstanding this ambiguity, pragmatic context is generally seen as helping determine, for example, whether a certain lake is included, or not, in response to a query for recreation locations near a hotel, by taking into account contextual details.

Pragmatic heterogeneities include differences in the creation, use, and effect dimensions. Creation differences include various provenances for a geographical entity, such as the diverse intensions developed by distinct geoscientists.
Use differences recognize, for example, that a lake might have various purposes in different databases, such as recreation, water supply, or transportation, and that some uses are more fitting than others for certain states-of-affairs. Effect differences include potential or actual impacts associated with an entity, such as differences in the flood plain associated with a lake.

Pragmatics is categorized variously in the interoperability stacks. One categorization delineates pragmatic context into intensional or extensional types: prescriptive constraints are intensional, such as policies and regulations, while restrictive states-of-affairs are extensional. The simulation interoperability stack uses this distinction to delineate pragmatic and dynamic levels: the former is about awareness of possible actions, such as workflows, and the latter is about adjusting actions dynamically in reaction to a running state-of-affairs. Note this stack is capped with a conceptual level that seems more aligned with semantics as described above. Another categorization delineates pragmatics according to agent action or interaction: action constraints guide agent behavior, such as a workflow plan, while social interchange among agents guides interaction constraints, such as the creation of policies or regulations. The computational semiotics stack uses this distinction to delineate the pragmatic and social levels, while the data interoperability stack focuses on agent interaction, primarily in terms of reaching consensus about data access policies such as security. In contrast, the Semantic Web stack uses a mix of categories: the proof level is extensional, providing provenance about some resource or action, while the trust level involves agent interaction as it concerns evaluation of one agent’s resource or action by another agent.

Pragmatic constraints in geographical information interoperability exemplify many of these distinctions. The more prominent examples include trust evaluation of OGC web services used to display maps online (extensional); the specification of domain-specific best practices for the use of certain standard geospatial web services, for example, sensor observation services for hydrology, and standards to describe the creation, supply, and use of specific datasets (intensional and interaction); critical investigations on the effect that institutional policy and behavior have on interoperability within SDIs (intensional and interaction); and the effect of individual observers on the development of shared geoscientific concepts (extensional and interaction).

Architectures for geographical information interoperability

An interoperability architecture is a plan that describes how computational components function cooperatively to achieve an intended goal. The plan includes a strategy and technical design that both constrains and guides the selection of components, as well as their arrangements. In this sense, it is a specialized version of the broader notion of enterprise architecture, used to describe the design of any major information system. Frameworks for an enterprise architecture are usually composed of multiple dimensions, which denote multiple perspectives on the system (e.g., the Zachman framework). A frequently included dimension is level of technicality, which usually grades from detailed technical details to inherent abstract concepts and their motivating requirements. A conceptual architecture then refers to a system design at the level of broad abstractions – it is general and applicable to a wide variety of situations. In contrast, other technical levels increasingly become more specific, ultimately ending at an implementation plan for a particular system. While large data sharing initiatives, such as national...
and international SDIs, have variously developed multi-perspective architectures for their specific environments, research efforts in geospatial information interoperability have been largely focused on developing conceptual architectures. In these efforts, conceptual architectures can be described using a mixture of approaches, including component-based, which specifies major components and their roles; layer-based, which categorizes components according to the type of usage (Heitmann et al. 2014); and structure-based, which describes the various approaches to connecting components.

**Component-based architectures**

Typical constituents of the component-based approach are shown in Table 1. These consist of functional components and knowledge artifacts. Functional components are system elements that undertake specific actions related to data and metadata, queries, and workflows. In general, requests to the interoperability system are structured as queries, which can retrieve data from multiple sources and possibly select and invoke workflows, using metadata to help find, evaluate, and select the required components. Knowledge artifacts are declarative representations from various levels of interoperability, and consist of (i) schemas for data, queries, or interfaces to executable codes, (ii) ontologies, (iii) mappings involving schemas and ontologies, and (iv) workflows, which document a plan for the coordinated running of executable codes. It is noteworthy that these components are universally applicable to the three major approaches to geospatial information interoperability: database interoperability, SDIs, and Linked Data.

**Layer-based architectures**

Unlike component-based architectures, which are oriented around specific technical resources, layer-based approaches organize resources into strata containing components with similar roles or functions, and which interact with adjacent strata in similar ways. This is most evident in online geospatial computing environments, where prominent client–server, service-oriented, and Web-Oriented Architectures (WOAs) have led to two-layer and three-layer information interoperability architectures.

The client–server architecture is a distributed computing model in which computing tasks are variously segmented between resource requesters (clients) and resource providers (servers). The server layer mainly provides resources such as geospatial data or executable codes, and the client layer requests these resources and further processes and displays them, typically using web browsers controlled by humans. This architecture is considered tightly coupled, because access to server resources is largely platform-specific (including vendor-specific), causing clients to be primarily constructed for specific platforms. In terms of interoperability, functional components for storage, cache, access, and update are server responsibilities, while display components are uniquely client responsibilities; the remaining functional components are then variously carried out by either server or client, resulting in an exponential number of potential solutions as each platform creates a solution for every other platform.

In contrast, Service-Oriented Architectures (SOAs) are loosely coupled. They introduce middleware between server and client that is platform-neutral and often standardized, thus easing communications across platforms. This leads to a three-layer model in which data and code resources comprise the lowest layer, connecting middleware comprises the middle layer, and presentation mechanisms comprise the top layer. Often the middle layer is further subdivided according to specific roles, for
## Interoperability of Representations

Table 1 Components of geospatial interoperability architectures.

<table>
<thead>
<tr>
<th>Functional component</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data or metadata</strong></td>
<td></td>
</tr>
<tr>
<td>Store</td>
<td>Persistently stores data or metadata (e.g., in a geospatial database). Metadata can include descriptions of data, services, and knowledge artifacts, locations of all functional components, and data summaries such as aggregated statistics, value ranges, most recent additions, etc.</td>
</tr>
<tr>
<td>Cache</td>
<td>Optimized copy of one or more stores, typically maintained locally near some of the remaining functional components to increase efficiency of access.</td>
</tr>
<tr>
<td>Access</td>
<td>External machine interface to data stores and caches, for manipulating data.</td>
</tr>
<tr>
<td>Update</td>
<td>Refreshes data caches by harvesting from data stores.</td>
</tr>
<tr>
<td>Discovery</td>
<td>Finds components in the network, most significantly stores and knowledge artifacts, using pre-existing metadata or via dynamic exploration of a network.</td>
</tr>
<tr>
<td>Translation</td>
<td>Translates data described by schemas and ontologies, using mappings and metadata; performs semantic enrichment, matching, and integration as required.</td>
</tr>
<tr>
<td>Integration</td>
<td>Unites data from multiple sources into a single whole; includes uniting geographical features and relations that share identity.</td>
</tr>
<tr>
<td>Distribution</td>
<td>Returns translated and integrated data to a requester.</td>
</tr>
<tr>
<td>Mediator</td>
<td>Brokers data queries (requests) and responses, invoking other functions.</td>
</tr>
<tr>
<td>Display</td>
<td>Displays returned data, e.g., in a web browser, typically to a human user.</td>
</tr>
<tr>
<td><strong>Query</strong></td>
<td></td>
</tr>
<tr>
<td>Display</td>
<td>User interface for constructing queries, often using a common public ontology and query language.</td>
</tr>
<tr>
<td>Segregation</td>
<td>Determines which responders (data stores and/or service compositions) are appropriate.</td>
</tr>
<tr>
<td>Translation</td>
<td>Translates a query from the requesters’ language to the responders’ language.</td>
</tr>
<tr>
<td>Distribution</td>
<td>Distributes the requested query to the identified responders.</td>
</tr>
<tr>
<td><strong>Workflow</strong></td>
<td></td>
</tr>
<tr>
<td>Store</td>
<td>Stores executable codes and their arrangements as workflows; these can split into distinct stores for codes and workflows.</td>
</tr>
<tr>
<td>Compose</td>
<td>Creates a workflow in which components are functionally linked.</td>
</tr>
<tr>
<td>Execute</td>
<td>Executes a workflow.</td>
</tr>
<tr>
<td><strong>Knowledge artifact</strong></td>
<td></td>
</tr>
<tr>
<td>Schema</td>
<td>Structure describing data, a query, or an executable code such as a web service.</td>
</tr>
<tr>
<td>Ontology</td>
<td>Specification of a conceptualization of some domain.</td>
</tr>
<tr>
<td>Mapping</td>
<td>Mappings within and between other knowledge artifacts.</td>
</tr>
<tr>
<td>Workflow</td>
<td>Specific arrangement of some executable codes, such as web services or other software, intended to carry out a specific task.</td>
</tr>
</tbody>
</table>
example, to differentiate functions supporting data translation and search. Component interactions follow the publish–find–bind strategy, in which server resources are published as metadata, then discovered by clients who request resources from servers. In the geospatial community, the OGC has developed standard web-based service definitions for the access functional components, enabling platform-neutral pathways to online geospatial data and executable code. While web services are often deployed for the remaining applicable functional components, they have not yet been standardized and can vary by system.

WOAs extend SOAs, to a large degree by following RESTful principles. Key aspects include the adoption of standard web functions in the interface to a web service, for example, HTTP GET, PUT, POST, and DELETE; a standardized identifier for a resource that can be dereferenced, that is, the URI; and a standard web transport protocol, that is, HTTP. In essence, WOA web resources are accessed over HTTP via a standard web function operating on a URI. WOA is thus resource-centric in contrast to SOA’s function-centric approach, though WOA does not alter the three-layer interoperability model, as services are still prominent. In terms of geospatial interoperability, the main advantage of the WOA approach is mainly pragmatic, as the relative simplicity and uniformity of the WOA approach typically make services easier to implement in a web environment, particularly for uncomplicated applications, such as the simple retrieval of a geographic feature from a GIS. WOA can thus be seen as primarily advancing pragmatic interoperability. However, this pragmatic gain might be questionable when complex interfaces to resources are required, for example, in some scientific computing environments.

Structure-based architectures

In central/distributed architectures, functional components can be variously arranged within an interoperability network. The storage components usually comprise the resources being networked, and as such they differ variously in their distribution, heterogeneity, and autonomy. The remaining functional components then support interoperability among such resources. These remaining components can be arranged in various degrees of centrality and aggregation. A key component is the mediator, which effectively acts as a virtual node in the network, providing an interface to a subnetwork of resources by brokering requests and responses targeted at these resources. Mediators can be arranged centrally, in clusters over distinct subnetworks, or uniquely for each resource in peer-to-peer networks, and they can also be nested. These variations can result in complex arrangements whereby mediators are parts of one or more subnetworks. The remaining components can optionally be logically co-located (aggregated) with a mediator. For example, metadata, cache, update, discovery, segregation, distribution, and integration components are typically logically co-located with a mediator as their contents and functions usually span its subnetwork, while access and translation components are often logically, and physically, co-located with each resource in the subnetwork, as they are resource-specific.

Monistic/pluralistic architectures refer to the multiplicity of knowledge artifacts present in a network or subnetwork, specifically ontologies and/or schemas. This, in effect, reflects the degree of artifact standardization imposed on a (sub)network and the number of mappings required within it, which influences the placement of mediators and translators. In a monistic approach, a single normative ontology and/or schema is exposed by a (sub)network,
such that the ontologies and schemas of its resources are translated to and from these norms, requiring a mapping for each translation. In contrast, a pluralistic approach exposes multiple knowledge artifacts for a (sub)network, thus requiring an increased number of mappings for each combination of resource and artifact. Monistic approaches are favored in static knowledge environments, where ontologies and schemas are well known and do not change, while pluralistic approaches are favored in dynamic environments where such artifacts are not well known or are subject to change, due to the addition of heterogeneous resources to the network, or due to knowledge evolution within the network.

Static/dynamic architectures refer to a style of network governed by the degree of internal change, that is, by how frequently and intensely resources are added, deleted, or modified. For example, some SDIs are relatively static in practice, in that major resources are typically known beforehand and do not join or leave the network regularly – this requires significant a priori network design to characterize such resources and allow fixed components to be built around them. In contrast, some sensor networks can be highly dynamic, as sensors move in and out of the system frequently, so the presence of specific resources cannot be predicted. This architectural aspect also refers to the degree of change in knowledge artifacts. Linked Data is a good example of a dynamic environment, inasmuch as knowledge can change rapidly within its web-based resources. Consequently, they need not be annotated beforehand with schemas and ontologies, and indeed rapid or severe change might preclude this, requiring these artifacts to be discovered on the fly to the degree possible. Architectures in highly dynamic environments thus cannot anticipate the presence of specific resources, must adapt to changes, and might require substantial exploratory discovery mechanisms.

Approaches to geographical information interoperability

Significant approaches to geographical information interoperability follow major advances in computing, most notably the early rise of database systems and their geospatial manifestations, GIS, followed by the rise of the web with SOA fueling SDIs and, lately, WOA powering Linked Data.

Spatial database interoperability

Early, largely pre-web, geographical information interoperability was file-based and exhibited structure interoperability: specific GIS transformed foreign schema, in various syntaxes, into their own proprietary structure and syntax. Semantics and pragmatics were largely not supported. In essence, a static, monistic, distributed peer-to-peer architecture was deployed, using file-based media transfer. Subsequent research into schemas for geographical features enabled a centralized approach with an intermediary common schema that could be deployed in stand-alone software, for example FME (Feature Manipulation Engine). Database techniques that deployed common schema via dynamic mapping of views were largely absent among GIS, or played a minor role, as they became overtaken by SOA approaches.

Spatial data infrastructures

A SDI is a network of online geospatial data resources, as well as associated policies, agreements, standards, and people, that interoperate to share data. SDI resources are typically
INTEROPERABILITY OF REPRESENTATIONS

distributed across one or more jurisdictions, such as municipalities, states, or countries. Notable SDIs include NSDI (USA), GeoConnections (Canada), INSPIRE (Europe), and GEOSS (Global). SDIs are founded on the SOA, and related geospatial standards developed by the OGC and ISO, with support for all interoperability levels, albeit with ongoing developments in semantics. Functional components defined by the OGC are restricted to geospatial data and metadata storage, discovery, and access, as well as executable code execution, and schemas are by far the primary knowledge artifact. Although OGC/ISO standards enable either central or distributed approaches, SDIs typically adopt a central approach to metadata functional components, and a distributed approach to data components. Likewise, a monistic approach is favored for knowledge artifacts, via the development of normative schemas for particular domains. SDIs vary widely in terms of a static versus dynamic approach to overall network structure, and they are slowly following the gradual migration of geospatial standards to WOAs and related technologies.

Linked Data

Linked Data is an approach to connect data on the web, such that the data, its definition in the form of a type description, and its connection to related data are made explicit and easily accessible using web-based mechanisms. For example, data about “Lake Ontario” can be obtained by following links within and between the online DBpedia and GeoNames databases, each of which are further linked to a type definition for “lake.” Traversal of such links is in contrast to the navigation of web pages, which is a means of viewing representations of the data, such as a map of a specific lake. Linked Data is thus explicitly aligned with semiotic principles by inherently distinguishing an object (data about a geospatial entity), its representation (a web page), and its conceptualization (a type definition). It is also founded on four technical principles (Bizer, Heath, and Berners-Lee 2009): (i) use of URIs to identify any web entity, (ii) URIs able to be dereferenced via web mechanisms, (iii) dereferencing supplies information using Semantic Web approaches, such as RDF, OWL, and SPARQL, and (iv) connectivity between resources is promoted. As Linked Data adopts the Semantic Web interoperability stack, it addresses all interoperability levels, though omits structure, shifting structural issues to the semantic and pragmatic layers (Kuhn, Kauppinen, and Janowicz 2014). It also adopts the WOA, with its RESTful principles, and leverages the full range of architectural components, minus schemas. Data and metadata components were initially typically centralized in RDF databases, but are becoming more distributed with an increasing number of RDF data providers. Linked Data also takes strongly pluralistic and dynamic architectural approaches, to reflect the diversity and rapid change of web resources, including the geospatial. Outstanding challenges in the context of geographical information interoperability include things such as adaptation to structurally rigid data, and raster data, appropriate level of granularity for linking (e.g., a raster pixel or a sensor measurement?), integration with legacy systems and architectures, and linking software, not just data.

SEE ALSO: Data structure: spatial data on the web; Geographic information science; Geographic information system; Geospatial metadata; Geospatial Semantic Web; Metadata; Ontology: domain applications; Ontology: theoretical perspectives; Open Geospatial Consortium standards; Representation and presentation; Service-oriented architecture;
INTEROPERABILITY OF REPRESENTATIONS

Spatial data infrastructures; Spatial database; Spatial feature classes; Web-mapping services

References


The term “areal interpolation” was first coined in Goodchild and Lam (1980) to denote the problem of transforming data defined in one set of areal units to another, where the two sets of boundaries do not coincide (Lam 1980, 1983, 2009). The areal interpolation problem occurs when there is a need to incorporate different types of areal data defined in different administrative or political boundaries into a single platform for analysis. An example commonly found in geospatial analysis is to interpolate the socioeconomic variables (such as number of people or housing units) defined originally in census units into other administrative units, such as traffic zones or school districts, for integrated modeling. In areal interpolation terminology, the original areal units where data are available are called source zones, whereas the units into which the data are transformed are called target zones.

The ability to integrate data from one set of areal units to another set is crucial to many types of inquiry where disparate data defined in different forms are required. With the rapid development of geospatial technology and increased need for multidisciplinary integrated analysis, areal interpolation has become an indispensable tool in making many applications possible. Since 1980, there have been steady theoretical and methodological developments in the topic, and new forms of applications of areal interpolation are multiple. Areal interpolation is now a core body of knowledge and technology in geographic information science. This entry first explains the concept of areal interpolation and the basic methods for areal interpolation, then is followed by a summary of its latest developments and an illustration of two very different applications of areal interpolation. The entry concludes with remarks on future development.

Classification of areal interpolation methods

In a broader sense, the areal interpolation problem is a variant of the spatial interpolation problem. Spatial interpolation refers to the procedure of using a set of sample locations with known values to estimate values at other locations. Spatial interpolation includes two main forms: point and areal interpolation. Point interpolation deals with data collectable at a point, such as elevation or temperature readings, whereas areal interpolation deals with data aggregated according to some boundary delineations, such as census tracts or police precincts. Point interpolation has been frequently used to create isolines (lines of equal values), such as contours in topographic maps, and many point interpolation algorithms had already been developed prior to 1980. Point interpolation methods can further be categorized into two groups: exact or approximate (Table 1). Exact methods will retain the original sample point values after interpolation, whereas approximate methods do not guarantee the preservation of the original sample point values.
values and generally include a certain degree of smoothing. Some well-known examples in the exact method category are distance-weighting and kriging, whereas trend-surface and Fourier models are representatives in the category of approximate methods. The choice of an appropriate point interpolation method depends on the type of attribute to be interpolated, the accuracy of the sample points, the type of the underlying surface, and the computational resources required. In all cases, the fundamental issue associated with every interpolation method is that an assumption is made about the underlying surface for interpolation and that assumption may or may not be true.

The discussion of point interpolation and the general concept of spatial interpolation is very relevant to understanding the principle of areal interpolation and its historical, intellectual, and technical developments. Before 1980, the areal interpolation procedure was seldom applied and the only approach available was to mimic point interpolation by reducing the areal data into points. Hence, the “pre-1980” approach would generally start by overlaying a grid mesh on the source zones and assigning a control point (usually the centroid of the areal unit) to represent each source zone. A point interpolation method is applied to estimate the value at each cell. Then, the estimates of the cells are summed or averaged according to the new target-zone boundaries to produce the final target-zone estimates. In this approach, the major variations among the various methods are the different point interpolation methods used in estimating the grid point values. This approach was termed the “point-based areal interpolation approach” because it utilizes points in the interpolation process. A major problem of this point-based approach is that no matter how accurate the point interpolation method is, the approach does not guarantee the preservation of the original source-zone values. This property, called the volume-preserving property, was considered a vital property for accurate areal interpolation. Subsequent empirical tests demonstrated unequivocally that areal interpolation methods that have the volume-preserving property (preserving the original source-zone

---

**Table 1** Areal interpolation methods in the context of spatial interpolation.

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4 – Example algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial interpolation</td>
<td>Point</td>
<td>Exact</td>
<td>Distance-weighting; kriging; splines; interpolating polynomials</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approximate</td>
<td>Trend models; Fourier models; least-square splines</td>
</tr>
<tr>
<td>Areal</td>
<td>Point-based</td>
<td></td>
<td>Various point interpolation algorithms</td>
</tr>
<tr>
<td>(non-volume-preserving)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area-based</td>
<td>(volume-preserving)</td>
<td></td>
<td>Without ancillary data: areal weighting; pycnophylactic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>With ancillary data: binary dasymetric; EM (expectation/maximization); regression</td>
</tr>
</tbody>
</table>

---

INTERPOLATION: AREAL
values) are far more accurate than those generated by non-volume-preserving methods (Lam 1980, 1983).

The volume-preserving property

The first two methods developed for volume-preserving areal interpolation are the areal weighting and pycnophylactic methods (Goodchild and Lam 1980; Lam 1980; Tobler 1979). The areal weighting method superimposes the target zones on the source zones, and their areas of intersection are used to derive weights to distribute source zone values proportionally into target zones. In equation form, an estimate of target zone \( t \) is obtained by:

\[
V_t = \sum_s U_s a_{ts} / \sigma_s
\]  

(1)

where \( V_t \) is the target zone \( t \) estimate, \( U_s \) is the value of source zone \( s \) intersecting target zone \( t \), \( \sigma_s \) is the source zone area, and \( a_{ts} \) is the area of intersection between target zone \( t \) and source zone \( s \). The areal weighting method is intuitively simple. The method uses the source zone itself as unit of operation instead of reducing them into points; hence, it is considered a type of area-based areal interpolation. The method by nature preserves the original source zone values, and therefore it is a volume-preserving method. The method assumes spatial homogeneity within each source zone. The higher the spatial homogeneity of the attribute (i.e., evenly distributed) within each source zone, the more accurate the resultant target zone estimates. Unlike the pycnophylactic method discussed next, no smoothing across source-zone boundaries is enforced.

The pycnophylactic interpolation method was originally developed by Tobler (1979) for isopleth mapping. Lam (1980) incorporated the method for areal interpolation by adding the last step of re-aggregating the grid cells into new target-zone estimates. In Tobler’s terminology, pycnophylactic means volume-preserving. The method starts by overlaying a mesh of grids on the source zones and assigning the same mean value to those grid cells belonging to the same source zone. The cell values are then modified iteratively according to two criteria: smoothing and volume-preserving. The smoothing condition is applied by replacing the cell value with the average of its four neighbors. The volume-preserving condition is enforced by comparing the sum of all cell values within a source zone with its original source-zone value in each iteration and adjusting the cell values by either increasing or decreasing them. Let \( U_s \) be the value of source zone \( s \), \( \sigma_s \) the area of source zone \( s \), \( Z_{ij} \) the value in cell \( ij \), and \( \alpha \) the area of a cell. Also, set \( q_{ij}^s \) as 1 if \( ij \) is in zone \( s \); otherwise set it at 0. The pycnophylactic conditions include:

\[
\sum_{ij} (\alpha Z_{ij} q_{ij}^s) = U_s, \quad \sum_{ij} \alpha q_{ij}^s = \sigma_s, \quad \text{and} \quad \sum_s q_{ij}^s = 1
\]  

(2)

Unlike the areal weighting method, the spatial homogeneity assumption is no longer required in the pycnophylactic method. Instead, the method assumes the existence of a smooth surface that takes into account the effects of adjacent source zones. This smoothing assumption, however, may or may not hold true in real-world applications. An empirical test conducted using hypothetical surfaces generated from a fractal algorithm showed that the performance of both volume-preserving methods was far better than that of the non-volume-preserving methods, and that within the group of volume-preserving methods, the areal weighting method performed slightly better than the pycnophylactic method (Lam 1980). In a nutshell, these two volume-preserving methods can be considered as the two ends of a continuum between a
very discrete and a maximally smooth surface. There should be some real-world cases where reliable interpolation occurs somewhere along the continuum for the whole or parts of the study area. Areal interpolation with ancillary data, subsequently developed after the 1980s, is designed to provide this additional information to produce better estimates.

Areal interpolation with ancillary data

With ancillary data, the accuracy of areal interpolation can be substantially improved, because the incorporation of ancillary data can effectively remove the need for an assumption about the underlying surface or reduce the error resulted from imposing a faulty assumption. The most commonly used ancillary data are land-use–land-cover data derived from remote sensing images. Other types of ancillary data, such as geological layers, road networks, and land parcels, have also been used. Since 1980, improvements of areal interpolation methods have been centered on how to incorporate additional data layers to make the target-zone estimates more accurate.

Areal interpolation with ancillary data is also called “intelligent” areal interpolation (Cromley, Hanink, and Bentley 2012). Ancillary data such as land cover categories are called “control” variables. Areal interpolation with additional control variables is very similar to the principle of dasymetric mapping, a mapping technique designed to reflect within-zone variations. Dasymetric maps use line symbols to depict the spatial distribution of attributes that are defined in areal units. Unlike choropleth mapping, where variables are mapped following the areal unit boundaries, dasymetric mapping uses additional control variables to identify variations within zones and depicts the spatial variation using a set of isolines on top of the areal unit boundaries.

Three groups of intelligent areal interpolation methods can be distinguished. The simplest one is the binary dasymetric method. In this method, control variables, such as land-cover data, are used to identify locations where certain attribute would not exist. For example, in estimating the population number for target zones, if a source zone has a large body of water, it can be assumed that the water area does not have population. Therefore, instead of using the total area of the source zone to derive the weights in areal weighting interpolation, only the area minus the water body is used.

The expectation/maximization (EM) algorithm is used to refine the basic dasymetric method by producing better estimates of subsurface zone values given the ancillary data. The EM method was developed originally to solve the problem of missing data. The method uses two steps. The expectation (E) step calculates the conditional expectation of the subsurface zone value given the total source zone value and the ancillary variable, whereas the maximization (M) step fits the model by maximum likelihood. These two steps are repeated until the algorithm converges. The method can be generalized for different types of ancillary data, including both continuous and categorical types.

The third group of intelligent areal interpolation methods utilizes statistical techniques such as regression to derive functional relationships between the ancillary data and the attribute to be estimated. For example, different land-cover categories are likely to have different population density ratios. Thus, a large part of subsequent research has been on perfecting the regression method to achieve higher accuracy in target-zone estimates. The standard ordinary least squares (OLS) method was first experimented with. However, it was soon discovered
that the ordinary least squares method has several problems: (i) ordinary regression models normally have an intercept, which means that population could still exist even when all the independent variables are zero; (ii) ordinary regression models can lead to negative regression coefficients, which could lead to negative target-zone estimates; (iii) regression coefficients developed for the entire study region may not apply to different parts of the study region.

Several approaches have been developed to address part or all of the three issues related to the ordinary least squares method. The first problem of intercept can be corrected by forcing the regression line through the zero origin, an option that is available in most statistical software. To address the second problem of ensuring no negative values, a number of methods can be used, ranging from simply adding a scalar to offset the negative values to the use of the spatial error model, Poisson regression, and quadratic programming optimization techniques to ensure no negative values. The third problem of heterogeneous substudy area estimates can be overcome by using hybrid models for different parts of the study area. Recently, the quantile regression model has been suggested as an efficient method that can resolve all three issues. Quantile regression minimizes the sum of absolute deviations about the regression line, as opposed to minimizing the sum of squared deviations in ordinary least squares regression. The technique is complex, as it requires linear programming. Furthermore, an empirical study comparing quantile regression with other areal interpolation methods showed that not much accuracy was gained by using this method, making the second-highest accurate but much simpler method, the binary dasymetric method, a viable choice for most areal interpolation (Cromley, Hanink, and Bentley 2012).

Areal interpolation in practice

Two diverse applications of areal interpolation are selected for illustration here: the national historical geographic information system project (Schroeder and Van Riper 2013) and the carbon dioxide emissions estimation project (Shu, Lam, and Reams 2010).

The national historical geographic information system (NHGIS) project, together with a follow-up project called the integrated spatio-temporal aggregate data series (ISTADS), aims to produce a web-based access system that provides summary statistics and boundary data for US census from 1790 to the present (Schroeder and Van Riper 2013). A major task of the project was to standardize the geographic units across censuses, which changed from time to time. Areal interpolation was the main technique to link the historical census data to the present.

The project first used a method called target-density weighting (TDW) for areal interpolation, which is more accurate than simple areal weighting. The target-density weighting method assumes the same proportional distribution of an attribute between two time periods and utilizes the proportional information from one time period to allocate source-zone values to target zones for another time period. For example, to allocate a 1980s tract population to two intersecting target 1970 tracts, if one of the 1970 tracts had twice the population density as the other in 1970, then the same 2:1 ratio was used to allocate the population in 1980 (Schroeder and Van Riper 2013).

However, such method may produce large errors, especially when the assumption of constant data distribution ratio through time is not true. To improve interpolation accuracy, researchers in the project have recently developed a new method, called the geographically weighted expectation-maximization (GWEM).
INTERPOLATION: AREAL

The new method combines both the EM method and the geographically weighted (GW) regression. The EM step provides a framework for incorporating proper constraints on data distribution, whereas the GW step allows an estimation of data distribution to vary spatially. An assessment of the GWEM method based on a study using a land-use–land-cover dataset to interpolate 1980 US census tract pairs showed that the target-density weighting method generally outperforms GWEM, but a combined GWEM-TDW method improves estimates substantially (Schroeder and Van Riper 2013). Given the importance of producing reliable historical census data that will be used by many, it warrants the use of a complex interpolation procedure such as GWEM-TDW to gain higher interpolation accuracy.

A relatively new application of areal interpolation is the downscaling of broad model estimates into fine-scale cell estimates for a variety of environmental applications. For example, detailed estimates of carbon dioxide (CO₂) emissions at fine spatial scales are useful to both modelers and decision-makers who are faced with the problem of global warming and climate change. Shu, Lam, and Reams (2010) demonstrated the use of the volume-preserving principle in downscaling carbon dioxide (CO₂) emissions from transportation from the county-scale to a 1-km² grid cell for the entire state of Louisiana, USA. The volume-preserving interpolation (VPI) method, together with the distance-decay principle, were used to derive emission weights for each grid based on its proximity to highways, roads, railroads, waterways, and airports. The total CO₂ emission value summed from the grids within a county was made to be equal to the original county-level estimate, thus enforcing the volume-preserving property. The distance-decay principle is similar to the use of ancillary data for intelligent areal interpolation. The results revealed a more realistic spatial pattern of CO₂ emission from transportation, which can be used to identify the emission “hot spots.” The fine-scale grid values can be used to integrate with other attributes for further analysis and modeling.

Future directions

The concept and techniques of areal interpolation, and its underlying volume-preserving property, is considered a contribution originated from geography, cartography, and geographic information science, as opposed to many other methods that are borrowed from other disciplines. Although there are always errors involved in the interpolation estimates, recent methodological advances in areal interpolation have substantially reduced the errors, making it an indispensable tool for many applications. However, unlike kriging or other point interpolation techniques, areal interpolation remains an “elite” technology that is difficult to access and use by the general scientific community. In the short run, two future research directions are obvious. First, building a software module for the various options of areal interpolation and making it accessible would be very useful to many different applications. Second, an error estimation module such as Monte Carlo simulation is needed to help generate confidence limits of target-zone estimates. These two steps will help propel the field, not just areal interpolation, from academic scientific investigations to practical decision-making.

SEE ALSO: Geographic information science; Interpolation: inverse-distance weighting; Interpolation: kriging; Overlay, topological
References


Interpolation: kriging

Changshan Wu
University of Wisconsin–Milwaukee, USA

Kriging represents a group of geostatistics-based interpolation techniques that attempt to give an optimal estimate of the value of a variable on a surface. Kriging was initially developed in the 1960s by Matheron (1963, 1973) based on the theory of regionalized variables, which was, in turn, an extension of the methods employed by David Krige in the mining industry of South Africa (Cressie 1993). Similar to other spatial interpolation methods, such as spatial average and inverse distance weighting, kriging is a distance weighting technique. The major difference between kriging and other distance weighting techniques lies in how the weights are decided. Techniques such as spatial average and inverse distance weighting employ expert knowledge to decide the range of spatial autocorrelation and the rate of distance decay, while kriging examines this through analyzing the spatial structure of the control point data.

Steps of kriging

The basis of kriging is to estimate the weights assigned to each control point, and this is implemented using three consecutive steps: (i) deriving the experimental variogram through examining the spatial variation in the control point data, (ii) fitting the experimental variogram into a mathematical model using a regular mathematical function, and (iii) interpolating the variable value of a particular location through solving a least squares function. These three steps can be summarized as follows.

Step 1: Deriving the experimental variogram through examining the spatial variation in the control point data. The objective of this step is to examine the patterns of spatial variation (spatial dependence) within the control points. That is, for every two points, it calculates the squares of the differences between them, and examines whether two nearer points have similar values, and two further points have significantly different values. The formula of estimating the experimental variogram is shown in equation 1, where \( \hat{\gamma}(d) \) is the experimental variogram with a lag of \( d \), \( \Delta \) is the lag width, \( z_i \) and \( z_j \) are the values of points \( i \) and \( j \), and \( d_{ij} \) is the distance between points \( i \) and \( j \).

\[
\hat{\gamma}(d) = \frac{1}{2n(d)} \sum_{d_{ij} = d - \Delta/2}^{d_{ij} = d + \Delta/2} (z_i - z_j)^2
\]

Step 2: Fitting the experimental variogram into a mathematical model. With the experimental variogram, numerous theoretical mathematical models, such as linear, spherical, exponential, Gaussian, and circular, and so on, can be employed to fit it. A number of variogram fitting techniques have been developed in the literature, including ordinary least squares, weighted least squares, maximum likelihood, and so on; the weighted least squares technique detailed by Cressie (1985) has been widely employed. With the fitted mathematical function, several essential parameters, including nugget, range, and...
INTERPOLATION: KRIGING

sill, can be obtained (see Figure 1). Specifically, nugget represents the variance at zero distance, which is also considered as measurement error or the uncertainty of the attribute values. Range indicates the distance with which the variogram begins to level off; therefore, it is also considered as the range of spatial autocorrelation, and for the purpose of interpolation, any points beyond the distance of range should not be considered. Lastly, sill is the constant value of semivariance beyond the range.

Step 3: Interpolating the value of a variable at a particular location through solving a least squares function. With the variogram function obtained, the next and final step is to determine the weight of each sampling point. The interpolation can be expressed as in equation 2.

\[ \hat{z}_s = \sum_{i=1}^{n} w_i z_i \quad (2) \]

As an unbiased estimation, kriging requires that the mean of the estimated value is equal to the mean of the “true” variable, that is, \( E(\hat{z}_s) = (z) \), and the variance of the estimated error should be minimized (see equation 3).

\[
\begin{align*}
\text{Minimize } & \ E\{[\hat{z}_s - z_s]^2\} \\
= & \ 2 \sum_{i=1}^{n} w_i \gamma(d_{is}) - \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \gamma(d_{ij}) \\
\text{Subject to } & \sum_{i=1}^{n} w_i = 1 \quad (3)
\end{align*}
\]

As a constrained linear optimization problem, it can be solved using the Lagrangian relaxation technique, and the resultant weights, as well as a Lagrangian multiplier \( \lambda \), form a system with \( n + 1 \) linear equations with \( n + 1 \) variables (equation 4), where \( w_i \) represents the weight of each sampling point \( i \), \( \gamma(d_{ij}) \) is the variogram value with the distance from point \( i \) to point \( j \), and \( \lambda \) is the Lagrangian multiplier. With the derived \( w_i \), the attribute value at point \( s \) can be estimated through applying equation 2.

\[
\begin{align*}
& w_1 \gamma(d_{11}) + w_2 \gamma(d_{12}) + \ldots + w_n \gamma(d_{1n}) + \lambda = \gamma(d_{1s}) \\
& \vdots \\
& w_1 \gamma(d_{ns}) + w_2 \gamma(d_{ns}) + \ldots + w_n \gamma(d_{nn}) + \lambda = \gamma(d_{ns}) \\
& w_1 + w_2 + \ldots + w_n = 1 \quad (4)
\end{align*}
\]

Variations of kriging

Ordinary kriging, as introduced above, assumes that a constant mean exists for the underlying surface, and an interpolation can be implemented through only considering spatial dependence.

Variations of kriging

**Figure 1** An example of fitting an experimental variogram using a spherical model.
If there is drift in the underlying surface, however, universal kriging, which includes a trend surface analysis and ordinary kriging, needs to be applied. Other variations of kriging include indicator kriging (interpolate a binary variable) and cokriging (with two or more variables) (Cressie 1993).

Advantages and disadvantages of kriging for interpolation

An obvious advantage of kriging is that it employs the control point data to estimate the range and extent of spatial autocorrelation. That is, through constructing the variogram, it chooses the function, weights, and neighborhood for spatial interpolation. It is a statistically rigorous method, and the estimate is considered as unbiased. Moreover, kriging allows the estimation errors to be quantified using the derived variogram, the spatial pattern of the samples, and the weights.

Major disadvantages of kriging include the computation cost. Kriging is an especially complicated interpolation technique, and its every step is computationally intensive. With a large number of control points, the inversion of a large matrix is essentially burdensome, even with state-of-the-art computation powers. Second, as a statistical technique, kriging requires a large number of control points to be statistically rigorous. For an accurate estimation and modeling of the variogram, in particular, a large number of point pairs are necessary. Finally, although kriging is considered an automatic technique, human interventions are still necessary, and some arbitrary choices (e.g., distance band number and lag interval, model choices, etc.) might have a significant impact on the interpolation results.

SEE ALSO: Geostatistics; Interpolation: inverse-distance weighting; Spatial weights

References


The inverse-distance weighting (IDW) method is one of many spatial interpolation methods found in the literature (Bailey and Gatrell 1995), and it is very popular among geographic information system (GIS) users as the tool is easy to use, and the concept is relatively straightforward. The general concept of spatial interpolation can be explained by referring to Figure 1, in which elevations (in meters) of selected locations (part of the Shenandoah National Park in Virginia, USA) are available and shown, but the elevations of most part of the covered region are unknown. The basic objective of spatial interpolation is to estimate the value (in this case, the elevation) of a location where there are no data available using data that are available in other locations. In other words, the method interpolates a value for a location from values found around that location. By extension, if the method can estimate the value of a given location, it can estimate values for all locations in the study region. Thus, the ultimate result from spatial interpolation is often a surface in raster format (grid cells), providing interpolated values for all locations covered in the study area.

However, spatial interpolation should be explained in the context of spatial sampling (Stehman and Overton 1996). Most phenomena are spatially continuous; that is, theoretically, a measurement of these phenomena can be taken at any given location. However, different types of spatially continuous phenomena can vary over space in dramatically different manners. Some vary relatively continuous without big jumps and many physical phenomena can serve as such examples: temperature, barometric pressure, elevation, and ultraviolet light intensity. Measurements of this type of phenomenon usually change gradually in natural landscapes (with some exceptions), but may have abrupt or sharp changes in certain human-modified landscapes, such as the drop in elevation by several hundred meters from the edge of a skyscraper to just a few centimeters away from the building. However, many human-oriented phenomena vary more dramatically over space, partly due to the spatially discrete nature of human populations, and partly due to the artifact of how measurements are reported over space. Theoretically, we can measure population density around a “neighborhood” at every location, but the density may change relatively discretely as people are counted or not counted within the “neighborhood,” regardless of how a neighborhood is defined. Whether the phenomenon changes relatively gradually or discretely over space, the first law of geography is still valid in most cases: all things are related but closer things are more related (Tobler 1970). Thus, locations within the vicinity should have similar values. Therefore, taking a measurement from every location, which is impossible theoretically and practically, may not be necessary. We can just take measurements from selected locations (spatial sampling), hoping that these measurements from selected locations can provide an accurate overall description of the spatial distribution or pattern of the phenomenon.
INTERPOLATION: INVERSE-DISTANCE WEIGHTING

Although remote-sensing technology provides a means to capture spatially continuous phenomena in the form of a raster surface, measurements from selected locations are still quite common. Such sample data are often not sufficient to provide a visual depiction of the spatial variation of the phenomenon, such as the elevation example depicted in Figure 1. On the other hand, a surface describing the variation of the phenomenon is often preferred for visualization, analysis, and modeling. Then spatial interpolation techniques are employed to “fill in the holes (values)” between sampled locations. Sometimes, the sampled locations or points may be representative locations of regions rather than the actual sampled locations. That is, measurements may be taken within each region, and the summary of the measurements is assigned to the representative location, such as the geometric centroid of the polygon representing the region. In general, inputs of spatial interpolation are measurements in interval-ratio scales sampled from point locations, and estimates in interval-ratio scales are computed for unsampled locations. However, specific types of spatial interpolation, such as indicator kriging, can be used to provide estimates in categorical scale.

The general objective of spatial interpolation is to estimate the values in locations that have not been sampled based upon values in sampled locations. Let \( a(s_i) \) be the value of a random variable \( a \) in location \( s_i \), which may be represented by a Cartesian coordinate pair of \((x, y)\). In other words, values of \( a \) are location-dependent. Then, formally, spatial interpolation is to estimate the value of \( a \) in unsampled location \( s_0 \) (e.g., Figure 1) such that the predicted value:

\[
\hat{a}(s_0) = \sum_{i=1}^{n} \lambda_i a(s_i)
\]

where \( \lambda_i \) is the weight for location \( s_i \), and \( n \) points have been sampled for the value of \( a \). The weight \( \lambda_i \) should be smaller than or at most equal to 1, indicating how values in sampled locations \( s_i \) are used to predict the value of \( a(s_0) \). Different spatial interpolation methods vary in the way in which the weights are determined, and IDW provides a specific way to determine the weights in equation 1.

Two general approaches have been used in the literature to determine the weights. The statistical or geostatistical approach operates under the premise that the weights should be determined according to how the sampled values are spatially autocorrelated. To evaluate the magnitude of spatial autocorrelation in the sampled data, a variogram needs to be constructed showing the variation between observed values (semivariances) relative to their spatial separation (distance or spatial lag). If the phenomenon follows the first law of geography well, then semivariances of sampled values for closer pairs of locations
should be smaller than the semivariances for farther pairs. A mathematical function fitting the empirical variogram will provide a set of weights to be applied to values for sampled locations given their distances from the unsampled location. Kriging is a spatial interpolation method using the geostatistical approach. Many types of kriging have been developed and different kriging methods are needed for various situations and different assumptions (Bailey and Gatrell 1995; Lloyd 2010).

Different from the geostatistical approach where the weights are determined empirically by the spatial structure of the sampled values, weights of the deterministic approach are determined independent of the characteristics of the sampled data, but based upon certain premises about the relationships between values in sampled locations and estimates in unsampled locations. For instance, when estimating the value \( a \) in location \( s_0 \) (i.e., \( a(s_0) \)) in Figure 1, a logical estimate will be 461, the value from the nearest sampled location, if no other information is provided or used. In other words, unsampled locations will be assigned the values of the nearest sampled location, and the resultant value surface will consist of stepped flat surfaces with a blocky structure, exhibiting a configuration similar to the Thiessen polygons, in which boundaries assign locations to the nearest points or neighbors (Lloyd 2010; O’Sullivan and Unwin 2010). In this case, the weights \( \lambda_i \) for all sampled locations are set to zero, except that the weight for the nearest sampled location is set to 1.

Using the nearest sampled value as the estimate is a “winner-takes-it-all” strategy, as values of other sampled locations are completely ignored. On the other hand, IDW takes all sample values into the estimation but, logically, values in sampled locations nearer should be considered more while values in farther away locations should be considered less. In other words, IDW considers only the distance between the sampled and unsampled locations, and weights are determined under the premise that they should be inversely related to distance. If \( s_i \) is farther away from an unsampled location \( s_0 \) than another sampled location \( s_j \), then the sampled value \( a(s_i) \) should be weighted less than \( a(s_j) \) merely because \( s_i \) is farther away from \( s_0 \) than from \( s_j \). Therefore, a simple specification of the weight can be \( 1/d_{i0} \) for the sampled location \( s_i \), where \( d_{i0} \) is the distance between \( s_0 \) and \( s_i \), assuming that the inverse relationship is strictly proportional to distance.

Two additional conditions should be considered in determining the weights. While the general notion that the weight should be inversely related to distance is understandable, the exact relationship still needs to be determined: whether it is directly proportional or not. To accommodate this flexibility, the weight can be specified as \( 1/d_{i0}^\alpha \) or simply \( d_{i0}^{-\alpha} \) where \( \alpha \geq 0 \). The inclusion of \( \alpha \) as a power function can accommodate various degrees of inverse distance relationship. Figure 2 shows the change in weight with increasing distance when \( \alpha \) takes on different reasonable values. In general, weights decrease with increasing distances (i.e., sampled locations farther away from the unsampled location \( s_0 \)), but such distance-decay effect is more substantial when \( \alpha \) becomes larger. When \( \alpha = 0 \), increasing distance from the unsampled location has no impact on the weight. According to the spatial interaction literature, a distance-decay parameter of 2 is applicable to many phenomena based upon many empirical studies (Fotheringham and O’Kelly 1989).

However, if \( \alpha = 1 \), then the estimated value of the unsampled location will be the sum of all sampled values: not a desirable outcome. In fact, the estimate will still be extremely large even if \( \alpha \) does not equal 1. The problem is that we have not imposed a condition on how much we should take in total across all sampled values to
derive the estimate. Ideally, the sum of all weights for an estimate should equal 1 such that a fraction of each sampled value is used to derive the estimate. In other words, \( \sum_{i}^{n} \lambda_i = 1 \). This is the second condition to be considered in modifying the simple \( 1/d_{i0} \) weight. IDW is essentially a spatial averaging method through which an estimate draws on a fraction of each neighboring value. Thus, to ensure that weights sum to 1, we have to standardize the weights such that:

\[
\lambda_i = \frac{d_{i0}^{-\alpha}}{\sum_{j=1}^{n} d_{j0}^{-\alpha}}
\]

(2)

where all terms were defined above. However, a special condition is when \( s_i = s_0, \lambda_i = 0 \) for all \( i \neq 0 \), but \( \lambda_i = 1 \) when \( i = 0 \). This condition means that when estimating values for location \( s_i \), the sampled or observed value for location \( s_i \) will be used. If this special treatment is not invoked, then \( d_{00} = 0 \), creating a weight of 0 and thus the sampled value will be discarded in deriving the estimate for location \( s_i \). That will not be a logical way to handle the situation, as the estimated value should take on the observed value. Therefore, the estimated value for \( s_i \) is \( a(s_i) \), the observed value, and thus IDW is an exact interpolator.

Besides determining the value of the distance-decay parameter \( \alpha \), another parameter to be determined in performing an IDW interpolation is to determine the size of neighborhoods within which sampled values will be used to estimate the value of an unsampled location. Theoretically, the \( n \) in equation 1 can be the number of all sampled locations. But if the study area is relatively large, it may be pointless to include sampled values that are far away as their contributions to the estimate may be negligible. Therefore, it is also an efficiency consideration by limiting the number of sampled values to be included in determining an estimate. At least two approaches can be used to make such a decision. A spatial approach is to define a neighborhood of a given size around each unsampled location. Typically, a distance from the unsampled location is chosen to define a circular neighborhood within which all sampled values will be included in the interpolation process. All unsampled locations will use the same size to define their neighborhoods. Another approach is simply to determine the number of sampled values nearest to the unsampled location to be included in the interpolation. In this case, areas of neighborhoods for different estimates will not be the same because the distributions of sampled locations usually vary across unsampled locations.

Both the choice of neighborhood size, regardless of how it is defined, and the distance-decay parameter will affect the results of the IDW procedure. In general, the smaller the neighborhoods (by area or number of sampled locations included) the more sensitively the estimates will respond to local variations of sampled values, as only sampled values closer to unsampled locations are included, and closer values will have more weighting. Using larger neighborhoods leads to more observed values.
The four maps in Figure 3 are interpolated elevation surfaces using data reported in Figure 1. Figure 3a uses a relatively small $\alpha$ (0.5) and the 15 nearest sampled values (out of 24). As a result, the elevation surface is relatively smooth with the two peaks around the sampled locations of 507 and 596 m. In general, the slopes are relatively smooth mainly because of the small distance-decay parameter. The map in Figure 3b uses the same small $\alpha$ (0.5), but only the eight nearest sampled values. As a result, boundaries defining areas with similar elevation levels have quite irregular shapes as the distributions of nearest neighbors are also quite irregular. Figure 3c uses a larger $\alpha$ (2.0) and also the 15 nearest sampled values. Compared to the two maps using a smaller $\alpha$, the larger
distance-decay factor here increased the influence from the local values, and thus quite a few local peaks emerged. The final map (Figure 3d) also uses a small $\alpha$ (0.5), but the search for sampled values used an approximately 1 km circular neighborhood around each location. Thus, boundaries of some elevation regions are artificially circular.

The maps in Figure 3 show one important characteristic of the IDW spatial interpolation method. As the term *interpolate* implies inserting something in between, the IDW method, as for most interpolation methods, cannot estimate values beyond the edges of the sampled region. The spatial extent of the four maps in Figure 3 is essentially the minimum bounding rectangle of all the sampled locations. Values beyond the rectangle have to be estimated through extrapolation, which IDW cannot offer. In other words, in the context of using spatial sampling to generate a continuous surface, the extent of the sampled locations determines the spatial limits of the interpolated surface.

Figure 3 also shows a potentially alarming issue, not so much about the conceptual formulation of IDW but more about the algorithmic implementation of the method. In addition to being a weighted spatial averaging method, IDW is also an exact interpolator, as discussed above. While checking how well all observed values match up with the predicted values may take some effort, checking the predicted values against the overall observed minimum (218 m) and maximum (596 m) elevations is relatively straightforward. No maps except Figure 3c include the minimum and maximum elevations on the elevation classes, contrary to the exact interpolator claim that estimated values at the sampled locations should equal the observed values. This seeming inconsistency is the artifact of the rendering or data model adopted in GIS, as the raster cells have relatively low resolutions such that the minimum and maximum values, like all observed values, are smoothed or averaged by the neighboring predicted values when generating the raster maps.

IDW is only one of many deterministic methods. A popular alternative to IDW among deterministic methods is using spline functions. There are many types of spline and related functions (Lloyd 2010). The general idea is that these are relatively smooth mathematical functions that can be used to fit smooth surfaces between sampled values. Such spline function fittings may estimate values beyond the range of the sampled values, overcoming one of the limitations of IDW. Another limitation of the general IDW interpolation concept is the use of one distance-decay parameter for the entire region, assuming that the magnitude of the distance-decay effect is uniform spatially. However, Lu and Wong (2008) suggested that in areas where sampled values are relatively dense $\alpha$ should be smaller, drawing sampled values from farther away, while for less-dense areas $\alpha$ should be larger, putting more weight on close-by sampled values. Such local adjustments of the distance-decay parameter are implemented in the adaptive IDW method, borrowing ideas from the adaptive kernel density estimation method. However, implementing such adaptive methods in GIS is still challenging. As IDW is a popular choice among geographers, Lu and Wong (2008) offer some guidelines about when IDW or kriging should be used.

**SEE ALSO:** Distance decay; Geostatistics; Interpolation: areal; Interpolation: kriging; Measuring spatial dependence; Spatial sampling; Spatial weights
References


Intersectionality recognizes the multiple axes of social structural difference in individuals, and groups, and the ways in which these differences combine to advantage or disadvantage individuals and groups in different contexts. Intersectionality is a significant advance over one-dimensional identity politics. It has transformed theory, methodology, and teaching around social and spatial issues of human difference and disadvantage in women’s studies and other feminist endeavors, including feminist geography.

Juridical activist and scholar Kimberlé Crenshaw recognizes Anna Julia Cooper’s work in her own development of the idea of intersectionality. Cooper was a leading intellectual of the late nineteenth century, a period during which questions concerning who could speak for whom around issues of race and class were important. A widely respected black man named Martin R. Delany was known to claim that when he entered any gathering, the entire race [sic] entered with him. He intended that, in his person, he stood for all black Americans. Cooper disagreed: “Only the Black Woman [sic] can say ’when and where I enter, in the quiet, undisputed dignity of my womanhood, without violence …., then and there the whole Negro race enters with me’” (Cooper 1886). Cooper’s counter-statement illustrates her understanding of black women’s intersectional position as more disadvantaged, or “below” that of black men.

Cooper’s commitment to the most oppressed Americans informed her view that effective solutions to human problems must elevate the most disadvantaged. Cooper’s concern for black girls, in particular, birthed her realization that women and – under reconstruction – girls were more vulnerable than boys, even within what she described as the Negro [sic] community. In her analysis of the structural oppression of black women, Cooper recognized intersectionality (although she did not name it as such): black and female, which is both more disadvantaged under many conditions than either black or female (alone), yet which is not additive or separable under most circumstances.

The Combahee River Collective was next to formally articulate an intersectional ethos (see Patricia Hill Collins (2000) The Social Construction of Black Feminist Thought for many others who did so). Based in New York, this group of black women coalesced in the mid-1970s around political, ethical, and social concerns that arose from their lives. Chief among these were members’ experiences in civil rights groups, where they were frequently discriminated against as women by black men, and in feminist groups, where they were frequently discriminated against as blacks by white women. Both Cooper and the Collective recognized that work relations, or class, were a significant dimension of their oppression.

Kimberlé Crenshaw is currently the theorist most associated with intersectionality. She
explored the mistreatment of black women complainants in legal cases concerning workplace discrimination and rape, and in the inadequate remedies for battered women of color. Her investigations show why remedies for “women” must explicitly seek redress for those at the intersections of structural oppression such as race and class, or race and gender. Among her findings were that when solutions for women are developed based on the needs of women who are white and middle class, only they will be well served. For example, Crenshaw learned that rape crisis center clients in underserved areas may have needs for food and shelter that, in the clients’ eyes, must be met before they will accept assistance with the aftermath of having been raped. White, middle-class clients are less likely to have these needs.

The concept of intersectionality has been developed from American black women’s experiences of neglect in politics, law, social policy, and life, and these experiences and subsequent analyses arise from a variety of times and places. What all have in common is that black women have labored in them, disadvantaged by being female within groups of black women and men, by being black within groups of white and black women, and often through their relationship to the means of production as workers.

Crenshaw’s work launched intersectionality into the mainstream. Almost inevitably, a concept’s move from scholarly deployments to writing for a general audience is accompanied by simplification. In addition, intersectionality’s grounding in standpoint theory clashed with some poststructural theories that, among other things, declared the death of categories. So, while Cooper’s, Crenshaw’s, and the Combahee River Collective’s formulations were grounded in their experiences and those of women whose lives they sought to improve, in 2009 Ms. Magazine published an article which said, in essence, that by applying intersectionality to their own lives, all students in women’s studies classes could understand it (Thornton Dill 2012, 27). By 2012, this statement had found its way into a popular textbook.

To be fair, the same essay (written by a prominent black feminist) presented a nuanced, if only weakly political, view of intersectionality. But it stripped intersectionality of its roots, and thus its power, by suggesting that students’ immediate grasp of the concept was adequate. Intersectionality was born from and is intended to improve the lives of women who, in Crenshaw’s words, are “on the bottom” (Crenshaw 2000, 218). She offers this analogy:

Imagine a basement which contains all people who are disadvantaged on the basis of race, sex, class, sexual preference, age and/or physical ability. These people are stacked – feet standing on shoulders – with those on the bottom being disadvantaged by the full array of factors, up to the very top, where the heads of all those disadvantaged by a singular factor brush up against the ceiling. Their ceiling is actually the floor above which only those who are not disadvantaged in any way reside. In efforts to correct some aspects of domination, those above the ceiling admit from the basement only those who can say that “but for” the ceiling, they too would be in the upper room. A hatch is developed through which those placed immediately below can crawl. Yet this hatch is generally available only to those who – due to the singularity of their burden and their otherwise privileged position relative to those below – are in the position to crawl through. Those who are multiply burdened are generally left below (Crenshaw 2000, 218–219).

In other words, while intersectionality can illuminate the life of each student in the women’s studies classroom most American classrooms are not filled with black women. Those with “singular” burdens, including straight white women, white gay men, and black men, are not
who intersectionality is for, in an analytical or political sense. Nonetheless, all can benefit from analyses and politics that build on intersectional understandings of oppression.

Those for whom intersectionality is intended have been the subjects of research in a variety of fields. In geography, for example, the concept has been taken up largely by feminist geographers to illuminate investigations into water access in Bangladesh, the experiences of deaf people in Britain, and gender and work studies in Canada. Recent and exciting scholarship concerns immigration, transnationalism, and nation-states. Indeed, there has been an explosion in empirical work informed by intersectionality in the past decade and case studies may well be the methodology for which intersectionality offers the most. Intersectionality is also informing theory as well as offering important new empirical knowledge. Scholars have pursued theoretical questions including: What happens when intersectionality clashes with poststructuralism? And, how can intersectionality’s power be realized through a conversation among its strongest proponents (Roshanravan 2013)? Answers to these and other questions highlight how crucial it is to ground intersectional analyses in the histories and geographies through which feminist standpoints are achieved.

Over a century after its appearance on the American scene, intersectionality has become ubiquitous within women’s studies and within feminist analyses in other disciplines, including geography. A large measure of its appeal is its power to shed light on differences within groups such as women, and its potential to aid in coalition building. Its ongoing salience lies in its utility for case-study research, especially where case studies can inform larger quantitative analyses that may usefully inform policy. Keeping intersectionality focused on those most disadvantaged, as called for by the women of color who developed it, will be key.

SEE ALSO: Difference; Feminist geography; Feminist methodologies; Identity

References


Further reading

INTERSECTIONALITY


Interviews

Fernando J. Bosco
San Diego State University, USA

Interviewing is one of the most commonly used methods for gathering information about people, their lives, and the places, landscapes, and environments they occupy. In an interview, the collection of data occurs by inviting individuals at a set time and place to answer questions and/or have a conversation with a researcher, with a specific research goal in mind. As commonly used in contemporary geographic research, the goal of interviewing is to probe an issue in-depth to gain access to the meanings that people attribute to different experiences in different contexts and to understand the reasons people do what they do. Interviewing is a people-oriented research methodology; the focus is often on social relationships and human actions in specific contexts. Geographers have adopted interviewing as a method for research in a range of settings and scales, from the urban to the rural, the city to the neighborhood, the village to the agricultural field, the mundane to the exotic, the home to the workplace. The outcomes of an interview are often rich, detailed, and multilayered data (Valentine 2005). Depending on the research project, the outcomes of the interview might be preserved in textual, sound, or visual forms – or a combination of these. During an interview, researchers record, in different forms, either the entire conversation or key points and ideas discussed. It is not uncommon for participants to raise issues during an interview that researchers may not have anticipated. Thus researchers often write additional notes and reflections immediately after an interview to provide an additional layer of context and content. The objective is to have access to all the information collected at a later time so that it can be analyzed as primary data to help answer specific questions in a research project.

Interviews have generally been associated with several different systematic branches of human geography – such as cultural, urban, and feminist geographies. However, this method has been increasingly used by geographers pursuing all kinds of investigations in almost all branches of the discipline. For example, interviews play a central role in more recently emerged areas of specialization, such as in children and family geographies or in participatory GIScience where they are used together with other people-oriented methods of data collection. It is also a method increasingly used in geographic studies of human–environment interactions. Regardless of the topic or area of investigation, many geographers today see interviews as belonging to a shared set of fieldwork and data collection practices that involve questioning and having purposeful conversations with people. The nature, format, and outcomes of such questions and conversations are dependent on the research design guiding the investigation. Thus, interviews range from the formal to the informal; they include conversations of varying length that are based on structured or semi-structured questions and prompts or can follow an unstructured format that leads to more participatory and intimate contacts between researchers and research subjects.
Interviews and geographic thought

Interviews have a relatively long history as a geographic method of data collection – an even longer history if one interprets interviewing in a broader sense. Some scholars have made connections between geography and interviewing all the way back to classic Greek travel accounts, arguing that certain descriptions of places and people likely were informed by conversations rather than by direct field observations. Similarly, vestiges of early forms of interviewing can be traced back to the European ages of navigation and exploration during the sixteenth, seventeenth, and eighteenth centuries. However, what is clear is that this early form of interviewing was supplemental to fieldwork and not a direct, purposeful, and planned method of data collection as it is today (Cloke et al. 2004).

The practice of interviewing in geography has evolved in relation to broader philosophical changes in the discipline and together with a widening of the scope of geographic inquiry, reflecting changing scientific, sociopolitical, technical, and ethical issues. In the context of contemporary North American geography, some have connected early forms of interviewing to the Sauerian tradition of landscape and regional investigations of the early twentieth century. Sauer encouraged using questionnaires as a way to collect information on how “local people” lived and moved around an area, for example. Subsequent critiques of the landscape and regional traditions in human geography led some scholars to push for more refined and systematic forms of interviewing. With the rise of positivistic geography and spatial science in the 1950s and 1960s, geographers relied more on non-field and secondary data sources that could be analyzed through statistical methods. Those still involved in the collection of primary data through interviews placed more emphasis on ensuring that the contents of conversations with people remained independent of their social characteristics, suggesting a particular distanced or remote relationship to research subjects (McDowell 2010). The main attempt was to obtain objective data – to the extent possible. There was little attention paid to personal characteristics of the interviewees or the specificities of place, time, or topics under investigation. Interviews during this time were more similar to structured questionnaires that lend themselves more easily to data quantification or tabulation. This aim for objectivity was an attempt to introduce a more rigorous system for talking to people, but some have argued that such methodological choices further delayed the development and adoption of interviewing as a methodology for obtaining rich, in-depth accounts of the experiences of people in places.

Beginning in the 1970s, new interpretive and critical approaches made headway into the discipline and resulted in significant new developments in the area of qualitative methodologies, in particular in human geography. Key influences also came from other disciplines such as anthropology. The combined effect of these influences impacted the ways in which geographers conceptualized and conducted interviews. For example, with the advent of humanism and Marxism in geography, a participatory, interpretive, and political approach to collecting data began to play a more prominent role. More attention was being paid to the ways in which interviews were shaped by the social context within which they were conducted. Interviews were seen as a more appropriate way to understand how people assign meaning to the social relations and places that make up their lives. Humanists believed that in order to understand the world, researchers needed to understand how people perceive it in the first place. Marxists, on the other hand, saw talking to people not just as
More importantly, feminist geographers provided the discipline with a set of conceptual tools and ethical concerns that allowed for a complete rethinking of the procedures and goals for interviewing. In fact, feminism has proven to be one of the most important philosophies from which geographers have rethought the practice of interviewing. Feminist geographers emphasized the need to be critically aware of the conditions that enable the production of knowledge. In relation to interviews, this means paying attention to how gender relations and other differences such as race, ethnicity, and sexuality shape the interview process and how they are implicated in the research methodology itself. Equally important is paying attention to how gendered assumptions (or racialized, sexualized, and other difference-based assumptions) affect the type of information that is generated through interviews, as well as its interpretation (Herod 1993). Attention to the effects of gender and other differences as fluid, contextual, and dynamic in turn openly challenged notions of objectivity in the interview process, dispelling the belief that objective truths and data existed in the worlds and lives of the people being interviewed to be uncovered by designing more “scientific” interview procedures. Some feminist geographers saw more traditional or standard forms of interviewing as masculinist practices that kept a problematic hierarchy and distance between researchers and research subjects. Instead of searching for objectivity, feminists and other critical geographers emphasized that power dynamics were always present in the interview process and recognized that geographers conducting interviews were as much part of the construction of meanings as those being interviewed. Thus, geographers’ engagement with different philosophies changed the practice of interviewing, with the emphasis changing from objectivity to subjectivities and to paying more attention to the ways in which power relations and differences, place and social context, and emotions and affect mattered in the context of interviews. Today, the questions and issues that arise out of the uneven relations of power that exist between researchers and research subjects occupy a central place in the construction of knowledge that occurs through the interview process. Regardless of the way in which interviews are conducted or the perspective that guides the research, the majority of geographers today acknowledge that interviews are subjective and that it is not possible to collect unmediated, unbiased, or completely objective data. Recognizing issues of positionality, reflexivity, and situated knowledges and embracing (rather than avoiding or denying) the politics of interviewing have become crucial elements in contemporary practices of interviewing.

Types of interviews

In contemporary geographic research interviews can be grouped along a continuum ranging from collecting data based on predesigned research questionnaires and structured questions to uncovering information in more open-ended, interpretive, and intimate manners. Some in fact suggest distinguishing between questionnairing, a methodology more adequate for social surveys or extensive research strategies where the face-to-face interactions are more structured and consistent across the population being surveyed; and interviewing, more adequate in intensive research strategies where the goal is to have conversations to give a voice to research subjects (Cloke et al. 2004; see also Sayer 1993). All practices along this continuum involve asking people questions, so the lines between questionnaires and interviews are fuzzy.
INTERVIEWS

Questionnaires and more structured interviews given to particular populations still tend to be associated with the practices of quantitative and/or positivistic human geography, in particular when the goal is to collect information in a way in which it can be quantified or consistently coded into specific variables or nodes during the analysis phase. A good example of this type is the short answer interview typically conducted over the phone. This is a more detached approach to interviewing in which the researchers remain in control of the questions asked and where there are fewer opportunities for interaction and negotiation between the researchers and the people being interviewed. In-depth interviews with open-ended questions, on the other hand, resemble conversations among two or more people. In this type of interviewing, investigators do not shy away from engaging with participants and probing issues in-depth, sometimes even generating spirited discussions and debating or contesting some of the answers. This more participatory approach to interviewing places the researchers in intense contact with the research subjects and is typically associated with critical perspectives in human geography and with qualitative methodologies. In-depth interviews do not, in general, aim to be representative of a whole population; rather they often respond to a theoretically motivated research question that seeks to illustrate how people make sense of their own experiences (Valentine 2005). However, no philosophical tradition is necessarily linked to a particular way of conducting interviews since, for example, a feminist geographer could collect interview data in a fairly structured way and yet provide a feminist and critical interpretation of the data in the analysis. In general, however, there is a consensus that some forms of questioning or different types of interviewing are more adequate for certain types of research. It is true that questions of embodiment and power relations are now taken into account much more often. However, not all researchers pay the same amount of attention to the politics of interviewing. Interviews as a method could be better described as including a range of ways of asking questions that themselves reflect a range of different relationships to research subjects as well as a range of philosophical and theoretical concerns.

Geographers have produced significant scholarship explaining the importance of carefully thinking about interviews’ style, format, and context in relation to the different types of people involved in the interview process. For example, in the context of interviewing families, one might need to pay attention to the power relations between household members, in particular if interview questions are designed to uncover tensions or highlight differential power relations. In this context, a geographer might choose to conduct separate interviews with parents and children, or with husbands, wives, domestic partners, or other adults in a household, both for ethical reasons and to achieve fuller disclosure from participants (Aitken 2001). Variations from place to place are also important. Some advocate talking to people in their own territory because participants might feel more comfortable and more can be learned about their own place-specific contexts and spatial routines. Others advocate conducting interviews in neutral spaces. Regardless of the place chosen, the safety of the participants and the researchers should always be considered. Particular attention should be paid when interviewing people in different cultural contexts since linguistic and cultural misunderstanding can easily emerge, even if working with an interpreter.

Other forms of interviewing employed by geographers require a different style of engaging
with participants and careful attention to interview dynamics. A good example is life history interviews, a method that aims to capture an individual’s biography and life history in his or her own words. This method of interviewing is often utilized by geographers interested in the role that personal history, memory, and tradition play in the social construction of place and identities. Life history interviews share some commonalities with regular interviews; for example, power dynamics are always present and assumptions about gender and other differences are similarly embedded in interviews. Yet there are significant differences, including the intensity of personal relation between the interviewer and interviewee, which can lead to a deeper exploration of emotions and affect. In life history interviewing, different dynamics may develop over the course of the interview, thus reflexivity and interpretation may occur more than in a regular interview (Jackson and Russell 2010).

Geographers’ engagement with new philosophies, theoretical frameworks, and new spatial technologies has also helped produce new hybrid forms of interviewing, where different types of data can be obtained at once. The walking interview, for example, involves walking alongside participants while asking questions as a way to capture attitudes, emotions, and knowledge about the surrounding environment. This type of interview can provide a more intimate methodology for learning about both individuals and places. Qualitative data from the walking interview can be enriched by spatial quantitative data collected through a global positioning system (GPS) in the interview process, providing an additional layer of information that can contextualize qualitative data and generate new insights about the connections of the participants’ stories and meanings in place (Evans and Jones 2011).

**Conducting interviews and analyzing interview data**

There is significant scholarship that suggests what to consider when incorporating interviews in geographic research design and setting out to collect data. There is also voluminous scholarship providing advice on how to conduct interviews. Overall, there is general agreement that the interview process is similar to an exchange rather than an interrogation. While interviews are typically arranged with a specific research goal in mind, the focus of an interview sometimes identifies itself in the course of conducting the research. The context of the interview itself can alter such focus because interview participants lead complex lives and power relations and identities are negotiated in the process (McDowell 2010; Aitken 2001). Researchers have to consider the influences of all these variables when trying to establish contact and rapport with interview participants, keeping in mind that unforeseen ethical issues might surface as one becomes more involved in the process. Because of the fluidity of interviewing, it is not uncommon for research projects to deviate from carefully planned schedules. Gaining the trust of participants often depends on how researchers position themselves politically in relation to them. As several geographers with significant experience with this methodology have attested, persistence and open-mindedness are needed in order to conduct successful interviews (McDowell 2010; Aitken 2001).

Geographers pay increasing attention to the politics of representation in relation to the data collected through the interview process, acknowledging that researchers play a defining role in the act of recording, transcribing, interpreting, and representing interview data. Who sponsors the research and who is selected for the interview, for example, will likely affect the type
Interviews

of data obtained. Transcription is central to the process of analyzing interview data, but until recently that process had received little attention in methodological discussions. Transcription of verbal information collected through interviews used to be considered mostly a technical chore performed by the researchers themselves or by third parties, such as research assistants or others providing professional transcription services. More recently, some geographers have begun advocating including research participants in the transcription process as a way to further democratize the interview methodology. Presenting participants with transcripts of interviews to ascertain their own opinions about the geographers’ interpretations can work as a mechanism to achieve more clarity regarding the information collected as well as a way to find a more equitable balance of power between the researcher and participants. Through this practice, further opportunities for reflection on the data can emerge, though giving transcripts to research participants generates new and unexpected ethical dilemmas (Mero-Jaffe 2011). For example, some have found that interview participants can be embarrassed by verbatim extracts of their own words (Jackson and Russell 2010). Thus, geographers working with interviews grapple with complex decisions about representation at all stages of the process, indicating that interviewing is far from being an easy or simple methodological choice but rather one that has to be approached mindful of its methodological complexities and implications.

Geographers have also moved beyond only analyzing the verbal and textual output of interviews, and are now paying attention to the nonrepresentational dimensions of interviewing as a way to gain further insights into how knowledge is constructed during the interview process. Lessons from a range of approaches, such as nonrepresentational theory and psychoanalytic methodologies indicate that a critical and more comprehensive analysis of interview data should include attention to the influences of silences and other embodied experiences that take place during an interview, such as laughter, visceral reactions, and body movements and gestures. This is particularly useful in instances of less successful interviews, which tend to be perceived as failed instances of research. In these cases, a researcher’s first inclination might be to discard data that does not tell much or appears not to have enough content and thus is seen as less useful. However, a reconsideration of the nonrepresentational might provide useful insights into the embodied politics of interviewing and say something about fluid negotiations of identity in the interview process (Nairn, Munro, and Smith 2005). This in turn can help clarify why certain questions were answered or left unanswered and how the construction of knowledge and data took place over the course of the interview.

See Also: Feminist methodologies; Fieldwork in human geography; Oral history and narrative; Positionality; Qualitative data; Quantitative methodologies

References


Intrametropolitan location

Anne Aguiléra
French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR), France

Producer services are major players in today’s productive system. Over the last decades, these activities have expanded exponentially in industrialized countries and, more recently, in developing countries. Geographically speaking, producer services companies tend to be concentrated in large urban areas, where there is a wide variety of clients and where they have special access to transportation and communications networks.

At the intra-urban scale, producer services initially located downtown, that is, in the historic centers of European cities and the central business districts (CBD) of North American cities. This configuration was primarily explained using land rent theory (Alonso 1960 and Figure 1). This was originally used to account for the location of agricultural activities before it was widely adopted for analyzing the structure of activities and populations from the 1960s. Assuming a monocentric city, namely a city with only one center concentrating all the customers, according to land rent theory, the price of land and real estate decreases proportionally as the distance from the center of the city increases. The center is the most expensive location, because transportation costs there are minimized. According to this theory, only activities with a substantial need for accessibility and a need for limited space can bear the high costs of a central area location. Producer services usually meet both of these criteria. First, many are office-based activities which typically do not require extensive floor space. Second, they require frequent interactions with their clients, most of which involve complex information that necessitates the presence of all of the parties (Goe et al. 2000) and, therefore, requires that at least some of the participants will need to travel. Indeed, the producer service production process is characterized by the close interdependence between producer and user (client). Services, which are by definition intangible, require some degree (depending on the type of service) of client participation, not only during the definition of the content of the service (co-definition), but also during production (coproduction). Due to the often complex nature of the information exchanged with clients, face-to-face meetings are often preferred and the service provider will travel to the client, or vice versa. Geographic proximity to clients can simplify these interactions and reduce costs (Clapp, Pollakowski, and Lynford 1992). In addition to facilitating supplier/client interactions and reducing the costs of face-to-face interactions, clients see geographical proximity as a way of reducing uncertainties, particularly for service transactions that may be perceived to be risky. Thus, land rent theory explains the central location of producer services on the basis of their need for proximity to clients. This reasoning appears to be of universal importance for explaining the strategies used by producer service firms for identifying locations, especially in the early stages of firm formation and growth. Land rent theory overlooks two additional influences on the location
decisions of producer service firms: the prestige associated with an address in a central location, and the role of the property developers who create the supply of office buildings, many of them state-of-the-art, landmark structures, in downtown areas that command the highest rents and impart the prestige that retains the attraction of city centers for producer services.

The development of more polycentric urban spaces profoundly upset the conceptual frameworks for analyzing the intra-urban locations of economic activity, including producer services. Indeed, since the 1990s, first in industrialized countries and more recently in developing ones, the spatial structure of people and jobs location within metropolitan areas has undergone major transformations. The tendency is a shift towards more polycentric forms with the development of suburban centers concentrating jobs outside the CBD (McMillen and Smith 2003). Outlying employment suburban centers do not usually replicate all of the morphological or functional attributes of the primary center but do compete in the hosting of economic activities. Many of these suburban centers are, for example, located strategically along major highway routes into and around the center of metropolitan areas, like the New Towns in the Paris region or Orange County in the Los Angeles region. At the outset, this pattern of polycentric development was encouraged by the decentralization tendencies of industrial activities, and retail and household services escaping the high costs of a central city location. Relocating industrial activities were able to find new, larger, and less expensive sites that offered easier access and parking for freight services and employees. Furthermore, public authorities were anxious to move high-pollution industrial activities out of city centers. For their part, retailers and household services were following the dispersal of the population. The latter increasingly had access to personal automobiles and wanted to live in single-family housing at locations away from the center of the city where prices were lower and there was less crowding. Following this first wave of decentralization, activities that had traditionally continued to prefer the city center, such as producer services, also began to locate some or even all of their activities in outlying suburban centers (Coffey, Drolet, and Polèse 1996; Gong and Willer 2002; Halbert 2004; Han and Qin 2009; Shearmur and Alvergne 2002). This behavior stimulated new theoretical and empirical analyses of the spatial strategies at work. An important working hypothesis was that the decentralization of producer services heralded the decline of downtown areas as they lost their most dynamic, emblematic activities. In practice, this hypothesis has
not been corroborated: producer service firms have certainly relocated to suburban business centers but they have also continued to maintain a presence in city centers (Halbert 2004; Maoh and Kanaroglou 2009).

By shifting the conditions for intra-urban accessibility and the associated transportation costs, the rise of outlying suburban centers has modified intrametropolitan land price curves. Although central business districts have, in general, retained specific attributes, such as the high density and variety of service activities, the density of transit networks, and access to information and communication technologies (ICT), property rents no longer fall with increasing distance from the city center. Although downtown remains the most costly, it has been demonstrated empirically that there is an inverse relationship between costs/rents and distance from the city center or from a suburban center (Ahlfeldt 2011). This signifies demand for and the emergence of new intra-urban distribution of economic activities that includes producer services, their clients, their suppliers, and the workforce. In particular, when a company’s clients and employees are themselves located in suburban areas, it encourages relocation in order to reduce transaction and transportation costs (Aguiléra 2003). Lastly, polycentrism is accompanied by a multipolar demand for producer services accommodation that takes the form of a multipolar supply of office space. Developers, typically encouraged by the public authorities, have taken advantage of the demand from decentralizing economic activities by constructing new real estate, especially office buildings and retail centers (Cervero 2013). These are typically high quality facilities with rents considerably lower than downtown, combined with excellent intra- and inter-urban accessibility. Despite such advantages, in the majority of metropolitan areas these suburban business centers cannot replicate all of the functions and advantages of downtown (Gaschet 2002). They certainly do compete with downtown on the basis of lower cost office space, its quality and accessibility for clients and staff, and the relative absence of negative externalities such as traffic congestion. But the prestige and sheer diversity of national institutions and government functions as a whole range of related and often international services outweighs for many producer service firms the negative externalities.

Nonetheless, land rent theory has logically been completed by new models that reduce the influence of the monocentric assumption. They highlight two key parameters in order to explain the new intra-urban configurations of economic activities. Location decision-making by firms is now primarily dependent on economic sensitivities to workforce transportation costs, on the one hand, and to agglomeration economies, on the other. The latter are all of the benefits that each company derives from its proximity to other companies or to particular facilities (such as transportation infrastructure). Depending on the model, several types of agglomeration economies are considered, including access to transportation infrastructure, and sharing of a number of services like signage, cleaning, security, or informational externalities. These externalities include the formal, as well as informal or social, networks that are of benefit to companies, particularly with respect to innovation. Spatial proximity facilitates these interactions; information exchange is believed to increase in proportion to the number of nearby companies, and to decrease in proportion to the distance between them. Whether these circumstances result in economic activity organizing into monocentric or polycentric configurations depends upon the level of transportation costs and on the nature of the agglomeration economies. In particular, it has been shown that
the activities most sensitive to informational externalities tend to agglomerate downtown but that if they grow and can no longer fit their activities within that central area, they prefer to relocate in outlying suburban centers, rather than dispersing across the urban fabric. By selecting suburban centers, firms can continue to enjoy some informational externalities even if these are inferior to those offered downtown.

The decentralization of producer services, activities which are sensitive to informational externalities, will therefore only take place after the outer suburban centers have attained a certain level of maturity and are able to offer suitable positive externalities. Theoretical models also make it possible to consider the idea that the decentralization of producer services only involves, at least at the beginning, those activities that are the least sensitive to informational externalities. Producer services incorporate a wide range of activities, only some of which are considered “superior” (high value-added) services with a strong need for complex information and face-to-face contact with clients, like financial and insurance services or management consulting. Consequently, there are two possible scenarios with regard to the decentralization of producer services. The first is that only the more commonplace services, such as security services, cleaning services, rental services, or computer maintaining, will choose to operate from outlying suburban centers. The second scenario is that producer services firms will eventually be induced to subdivide their activities, in the interest of cutting their location costs, between the front office (design and decision-making) which is sensitive to transaction costs and client proximity, and the back office (fulfillment) which is essentially engaged in the exchange of codifiable (e.g., electronically transmissible) information with the front office or elsewhere. According to this scenario, the front office activities will remain downtown while back office functions will migrate outward to benefit from reduced real estate and other costs.

The empirical work conducted over the past 20 years in the United States and Europe, as well as in China and South Korea, generally upholds the outcomes predicted by the model. A sustained concentration of producer services in downtown areas or central business districts alongside them (such as La Défense (Paris) or Canary Wharf (London)) is shown to occur, but a broader distribution of such services in outlying suburban centers also occurs (such as Shanghai). However, producer services do not locate equally in all possible suburban centers but favor the most accessible location and also in general the ones that are not too far from the center. Moreover, geographic segmentation has been observed, based on the type of service provided: “superior” front office activities generally stay more central than their more commonplace, back office counterparts. For all this, downtown has maintained a strong appeal overall for producer services, particularly headquarters and high value-added services like finance and insurance services, advertising, or consulting. In other words, producer services are simultaneously characterized by strong centrality and tendencies to dispersal, but still remain more centralized than other economic sectors.

Amongst the other determining factors for the intrametropolitan location of producer services, empirical research has also confirmed the pre-eminent role of transportation costs, land and real estate costs, and agglomeration economies. It has been shown that some outlying suburban centers offer a viable alternative to the center, because they enjoy good access to highways while offering markedly lower property costs often of better quality (for example, in terms of energy consumption). A high value also tends
to be placed on parking facilities. Nevertheless, downtown maintains its high-value reputation for reasons of locational prestige, particularly in European cities. Surveys of the factors shaping locational decisions have also confirmed the influence of client locations: companies serving a primarily metropolitan catchment area often resolve the location choice between center and suburb on the basis of the (central or suburban) location of the clients with whom they need to interact most frequently (Aguiléra 2003). However, downtown is often preferred when clients are primarily located outside the urban area, because, in most cities, it remains far more accessible to national and international transport such as high speed rail or an airport than do outlying suburban centers. In the case of smaller producer service firms, the locational preferences (such as choice of residence) of the partners or senior employees of firms can also influence the choice of location. Finally, the most recent research has begun to explore the influence of the location of qualified manpower on the geography of producer services activities (Shearmur and Doloreux 2009). Processes such as gentrification, observed in many downtown areas, help to explain the enduring appeal of central business districts, because newly established producer service firms tend to be located near to central neighborhoods where their founders already live. In turn, the establishment of these services in central zones maintains the gentrification process which then reinforces the enduring attraction of downtown business locations.

SEE ALSO: Accessibility, in transportation planning; Localization/delocalization; Polycentricity; Producer services: definition and classification; Suburbanization; Urban geography

References


INTRAMETROPOLITAN LOCATION


Further reading

Iran: Anjoman-e Geopolitic-e Iran (Iranian Association of Geopolitics)

Founded: 2002
Location of headquarters: Tehran
Website: www.iag.ir
Ireland: Geographical Society of Ireland (GSI)

Founded: 1934
Location of headquarters: Dublin
Website: www.geographicalsocietyireland.ie
Membership: 156 (as of November 2013)
President: Frances Fahy
Contact: frances.fahy@nuigalway.ie

Description and purpose

The purpose of the GSI is to promote the status and study of geography in Ireland. Drawing on both professionals and members of the public for its membership, the society is open to all those who have an academic interest in geography and the geography of Ireland.

Journals or major publication series

Irish Geography is the primary publication of the society. This premier peer-reviewed journal devoted to the geography of Ireland has been published since 1944. It has an international distribution and is read on all five continents. Its reputation for quality is long established and standards are maintained by an internationally based editorial advisory board.

GeoNews is the newsletter of the society, providing news and information about geography in Ireland and about the activities of the society. The newsletter acts as a forum for the discussion of matters of general geographical interest. Currently GeoNews is published twice a year (usually in May and December) and includes submissions of interest to the wider geographical community (in the form of articles, commentary pieces, field trip reports, details on events, etc.).

Current activities or projects

The society achieves its goals through its journal, its website and newsletter, outreach and educational programs, and the participation of its members in numerous national and international conferences, especially its annual meeting of the Conference of Irish Geographers (CIG). The CIG is hosted annually in May on a rotating basis between all the universities on the island of Ireland. In 2013 the CIG was hosted at NUI Galway and the conference theme was Transformative Geographies: Critical Reflections on Environment, Sustainability and Governmentality. The keynote speakers were Professor David Harvey and Professor Doug Sherman and over 300 delegates attended.

The GSI presents annual awards including Lifetime Contribution to the GSI, Postgraduate Fieldwork/Travel Awards, Book of the Year Award, and Geographer’s Contribution to Community Award. In addition, regular society activities include the organizing of Geographical Awareness Week and hosting a series of public lectures and fieldtrips including the Farrington Lecture.

Brief history

The Geographical Society of Ireland was founded in 1934 to promote the status and study of geography, drawing on both professionals and members of the public for its membership. The affairs of the society are managed by an elected committee. One of its early aims was the publication of a journal that would have a strong focus on the geography of Ireland.

Submitted by Frances Fahy
Irredentism refers to a claim made by a state or by an ethno-national grouping to territory currently lying within the borders of a neighboring state. The term derives from the Italian *irredento*, meaning “unredeemed,” and originally applied to territory regarded as Italian but lying outside the boundaries of the unified Italian state in the late nineteenth century. Irredentist territorial claims rest on the idea that the portion of geographical territory in question is rightfully part of another nation. Unlike secessionist nationalist claims, it is not independence that is sought but rather a transfer of sovereignty. Such claims rest on historical and cultural factors. It may be argued by a particular ethnic or national grouping that having lived for a long time in a place it rightfully belongs to them rather than to the power that technically has sovereignty over it. The desire to incorporate people who are part of the nation but living within another state may be driven by concerns over the treatment of those minorities. An alternative basis for such claims may be the symbolic importance of the territory due to historic events, and the preservation of places of memory such as battle sites. For political, cultural, and historical geographers, irredentist claims are of interest because of their assumptions of links between place, culture, and national identity, the questions they raise about political boundaries, and the utilization of cartography to bolster territorial arguments.

While many irredentist claims remain aspirational, others become sources of conflict. German annexation of Austria and other Germanic speaking regions in the late 1930s was justified on the grounds that these were Germanic territories. In much of central Europe the periodic redrawing of boundaries has produced many minorities that find themselves within the territory of another state. Similarly, the colonial construction of borders in parts of Africa and elsewhere has divided ethnic groupings. Colonial legacies persist in such cases as the Falkland Islands, over which Argentina maintains an irredentist claim, while Spain maintains a similar claim over Gibraltar. Many irredentist claims hark back to a time in which the nation’s territorial extent was much greater than it currently is. This fuels a desire to recapture territory lost and to reinvent a territorially expanded version of the state. Irredentist claims are often depicted on maps indicating territory occupied by ethno-national groupings that transcend state borders.

The case of Northern Ireland is a classic example of an irredentist claim in which the region (technically part of the United Kingdom) is seen by Irish nationalists as rightfully part of a united Irish state. While cross-border relations have evolved over time, the partition of Ireland in the 1920s was seen to rupture the territorial integrity of the nation (Coakley and O’Dowd 2007). The Balkans conflict in the 1990s reflects complex irredentist claims. Ethnic cleansing in Bosnia and elsewhere in the region was predicated on making the territory ethnically pure. The autonomous Republika Srpska within Bosnia is viewed by many nationalists as part of a “Greater Serbia” (Toal and Dahlman 2011).
Kosovo, which declared itself independent from Serbia in 2008, is still seen as an intrinsic part of Serb territory by nationalists who mobilize history to emphasize the Serb-ness of the region. For their part, many Albanian nationalists view Kosovo (along with parts of neighboring Macedonia and Montenegro) as belonging to the Albanian nation in a region where many ethnic Albanians live outside the borders of the Albanian state founded in 1912 (Malcolm 2002).

SEE ALSO: Borders, boundaries, and borderlands; Nationalism and geography; Nation-state; Sovereignty; Territory and territoriality

References


Further reading

Irrigation

William E. Doolittle
The University of Texas at Austin, USA

Defined as the artificial application of water to land, irrigation is most commonly used in regard to agriculture, particularly the production of crops in arid and semiarid lands. Soils are pivotal to the process. All plants absorb water through their roots, and most cultivated plants need minimum amounts of soil moisture at specific times during their growth. This is not an issue in humid environments where rainfall is abundant but not excessive, and soils are of adequate fertility and temperatures moderate. In deserts and steppes, where rainfall is erratic in terms of amount, timing, and location, soil moisture can be problematic. Irrigation helps reduce this uncertainty.

There are several ways irrigation can be practiced. Some of these are simple; some are complex. Some are ancient; others are modern. Regardless of age and complexity, however, certain variable conditions underlie decisions about which form of irrigation is used. These are: the nature of the water source and distance to fields; available and appropriate technology; the size of the labor force; and the demand for irrigated products. The first of these conditions might best be described as geologic, the second cultural, the third sociological or demographic, and the fourth economic. In concert they are geographical.

All of these conditions must be considered to understand specific irrigation strategies. To illustrate, geologic and cultural conditions might be available for one type of irrigation in a certain locale, but demographic and economic conditions might be prohibitive. In another situation, it might be economically feasible to implement a given type of irrigation, but only if water is close at hand, technology is appropriate, and labor is available. Successful blending of these four factors underlies the currently fashionable concept of sustainability in terms of understanding the fluorescence and collapse of ancient civilizations, and in future development projects.

Water sources

If irrigation is deemed to be necessary in a given locale, the issue of where the water will come from is paramount. Sources of irrigation water are generally dichotomized as either surface water or groundwater (see entries elsewhere). This distinction, however, is not always clear.

Places at which the water table intersects the Earth’s surface are known as springs. This is where groundwater becomes surface water. The volume of water flowing from springs can vary considerably over time, and in times of drought when aquifer recharge is low, the water table can drop such that springs will dry up. An excellent example of a spring used for irrigation agriculture is near Balmorhea, Texas (30°56’40”N 103°47’18”W).

Also blurring the distinction between surface water and groundwater is subsurface flow. Many rivers in dry lands do not look like rivers at all, but appear as dry stream beds. In fact, however, water can be flowing just below the surface, through relatively thick permeable materials.
IRRIGATION

(e.g., gravel, sand) that form the stream channel itself, and above the impermeable underlying geology (e.g., shale, basalt). Subsurface flow typically provides only small amounts of water and hence is usually associated with small irrigation systems in near-subsistence situations.

Delivery methods

Getting irrigation water from a source to a field is at once simple and complicated. Perhaps the oldest and most basic of delivery devices is some type of handheld container, be it a gourd, animal skin, basket, jar, or, today, a plastic bucket. This is simple technology and is only feasible for watering very small plots, such as the “waffle gardens,” of the Zuni people of New Mexico ($35^\circ 4'2"N\ 108^\circ 51'5"W$). Other technology is clearly needed to irrigate larger tracts of land. Water naturally flows downhill under the force of gravity. If fields lie at an elevation lower than the source, then all that is needed is some type of conduit between the two. But again, nothing is as simple as it first seems, and it is here that technology becomes more complex.

Canals

A canal excavated into the earth from a water source to fields is a basic form of water conveyance. In a sense, irrigation canals are simply artificial stream channels. As straightforward as this may seem, however, it is not without complications. The canal has to be at such a gradient that it is neither too great (steep) nor too slight. If the gradient is too steep, the flow will be fast and may cause down-cutting and bank erosion. If it is too slight, the flow might be slow enough to result in the canal being filled with sediment (silt ing up), as irrigation water is rarely clean and clear. Given that gradient is a function of horizontal as well as vertical distance, canals with an ideal gradient usually appear somewhat sinuous or meandering when seen from above or on maps and aerial images. This is evident in the Lower Rio Grande Valley ($26^\circ 31’21”N\ 98^\circ 6’49”W$).

Unlined earthen canals are of great antiquity, and almost as soon as they came into use, problems arose, particularly with erosion. Lining the canal bottoms and sides with a hard plaster, especially at sharp turns, became a common solution. Today, of course, concrete is used to line canals, and some canals are actually made from prefabricated concrete segments that are linked together. A good example of some remnants of ancient canals, in this case those built by the Hohokam culture, paralleled by modern canals, can be found immediately east of the international airport in Phoenix, Arizona ($33^\circ 26’29”N\ 11^\circ 59’6”W$).

Excavating canals at proper gradients, and sometimes over long distances, can result in obstacles being encountered. Bedrock outcrops are not uncommon in dry lands where soils are typically thin and poorly developed. Chiseling through such materials is obviously labor intensive. Remarkably enough, in at least one case, China’s Red Flag Canal, workers bored 71 tunnels through the Taihang Mountains in order to convey water from the Zhang River on the west to farm land on the east around Linzhou ($36^\circ 21’47”N\ 133^\circ 13’4”E$).

Boring through bedrock is one problem, having to convey water from one side of a gorge to another or over low-lying ground is another. The proper gradient has to be maintained even above ground and through air. Narrow, shallow gorges that are obstacles for small irrigation systems pose a much smaller problem than huge gorges that affect the development of large systems. In the former situations it is not uncommon
for irrigators to use split and hollowed out tree trunks supported by a framework made of wood. Such devices have been used for centuries by Hispanic farmers in New Mexico where they are known in Spanish as *canos* (canoes) and in English as “flumes” (36°7′54″N 105°45′19″W). Spanning deep or broad valleys involves substantially greater structures, bridge aqueducts, typically referred to as simply aqueducts. Constructed of ashlar masonry blocks or concrete, the structures involve a series of arches that support a canal. Although aqueducts were built before and after them, the Romans are most commonly associated with their construction (41°50′40″N 12°33′47″E).

**Pipes and tubes**

Issues of gradients and obstacles can be mitigated in part and in certain situations by covering canals or using pipes. Some of the earliest recorded uses of pipe come from ancient Mesopotamia where slender ceramic funnels were fitted together to transport water. In a similar fashion, bamboo has long been used in China. Romans made pipes by rolling sheets of lead. Indeed, with the chemical symbol Pb from the Latin word *plumbum*, this is where the terms “plumbing” and “plumber” originated. In more recent times, pipes have been made of steel, and now polyvinyl chloride (PVC). Each of these materials has its advantages and disadvantages, but they all suffer from being relatively inflexible. Modern high carbon plastics are, in contrast, very flexible, with the latter being weather-resistant, malleable, and inexpensive. Whereas pipes can have inside diameters of several centimeters and carry reasonably large amounts of water, these new flexible tubes have small diameters and transport measured quantities of water.

**Extracting, lifting, and storage techniques**

Transporting irrigation water first involves tapping the source and then often lifting it from lower to higher elevations (e.g., from stream channels to adjacent field surfaces). In many cases, it also involves storage over periods of time. To compensate for rainfall deficiencies, irrigation water is often collected during periods of abundant precipitation and stored until it is needed.

**Dams and reservoirs**

The diversion of flow from springs and streams, including those with subsurface flow, is often aided by the use of dams (see entry elsewhere). Some dams are quite small, constructed diagonally across streams, and simply divert water out of channels and into canals. An excellent example exists near Villafranca del Bierza, Spain (42°36′34″N 6°48′40″W). Others are large, constructed in a distinctive arch-shape perpendicular to the river, and store huge amounts of water. A classic example is Hoover Dam on the Colorado River, which created Lake Mead – a reservoir – near Las Vegas, Nevada (36°0′57″N 114°44′15″W). Regardless of their size, the surface elevation of the water is higher behind dams (upstream) than it is below (downstream). As such dams result in canals much shorter than would have been required otherwise.

**Qanats, shadufs, and norias**

Innovative people in ancient Mesopotamia and Egypt developed some fascinating irrigation technologies that remain with us. Qanats are tunnels, much like near-horizontal mine shafts, extending from the water table to field surfaces...
sometimes several kilometers away. They were constructed by first digging a vertical shaft from the surface near the base of a mountain to the water table, and from there, digging a tunnel to the fields. Fearlessness and special skills were needed to work underground. Earth from the excavation of the tunnel was removed through a series of additional vertical shafts by means of a windlass (a hand-cranked horizontal shaft around which a rope is turned) and buckets, and deposited around the opening of each shaft. When viewed from above, the only evidence of qanats are rows of split bagel-like features extending from mountains to fields. One of the more conspicuous examples is at the Tafilalt Oasis in Morocco (31°20′54″N 4°5′14″W). Qanats terminate in canals that take water to the fields.

Two other ancient Middle Eastern water-lifting devices merit discussion because of their ingenuity, role in the evolution of technology, and applicability for local development projects. One is the shaduf, which consists of an upright frame across which lies a long pole with a bucket on one end and a rope on the other. It is operated by releasing the rope, allowing the bucket to pick up water, and then pulling the rope and swinging the pole such that water can be poured into a canal. Originating in Egypt 4000 years ago, shadufs remain in use in northern Africa and southwest Asia. The other ancient device is the noria, or Persian wheel. It consists in its most basic form as a vertical wheel with buckets attached to the rim. The bottom of the wheel is below water and as the wheel turns water is scooped up in the buckets and at the top of the wheel is poured out into a trough before flowing into a canal. In some cases norias are powered by stream flow as is the case of the most famous ones in Hama, Syria (35°8′6″N 36°45′10″E). In others, shafts are excavated from the surface to the water table, and rather than having the bottom of the wheel in the water, a chain of buckets is used. In these cases, a second wheel oriented horizontally and meshing with the vertical wheel like gears is powered by draft animals walking in circles.

Pumps

The vertical shafts used for norias, and those used for qanats, are in fact, wells. Indeed, the wheels used in norias are a development that evolved from the windlass. Today, we have perhaps two stereotypical images of wells, the proverbial “wishing well” with a roof over the windlass, and the windmill which is not actually a mill (grinding machine). Both devices lift groundwater to the surface. In the case of windmills, the vertical turbine powers a mechanical pump. Invented in ancient Greece, it is perhaps most identified with the Dutch who employed wooden wheels similar to those used with norias, but to drain low-lying areas (51°52′52″N 4°38′31″E). Settlements of the North American Great Plains, Argentina, Australia, and southern Africa in the nineteenth century were facilitated by the development of inexpensive steel windmills that used a gearbox and crankshaft to convert the rotary motion of the turbine into reciprocating strokes vertically through a rod to a piston-pump cylinder below. An excellent assortment of wind pumps can be found at the American Wind Power Center in Lubbock, Texas (33°34′45″N 101°49′26″W).

The mechanical wind pump was a precursor to the modern tube well. Each well of this type is constructed by boring a hole 10–20 cm in diameter from the Earth’s surface into an aquifer, often at a depth of more than 35 m. The hole is lined with a steel sleeve and water is pumped to the surface. At one time diesel was the preferred fuel, but electricity is becoming more common.
Field surface configuration

Regardless of its source, the type of lifting or storage device, and the method of delivery, water eventually must be applied to fields. Complicating matters, water has to be evenly distributed to every part of the field, in amounts that are neither excessive nor deficient. A number of methods have been developed to either reshape the land surface to accommodate watering or to account for topographical variations.

Wild flooding and border strip

The most basic and doubtless the oldest form of irrigation is wild flooding. Most commonly, this consists of opening sluice gates, or simply using a shovel to breach the banks of canals with parallel subtle surface contours, and allowing the water to wash down over cultivated surfaces. The gates are closed when the land is sufficiently wetted. Wild flooding is common on pastures of the Prairie Provinces of Canada. In locations where crops rather than pasture grasses are irrigated on slightly sloping land, parallel earthen ridges (10–30 cm wide and approximately 1 m wide and 5–10 m apart) are constructed perpendicular to the slope contours. These border strips help equalize distribution of water over the fields.

Contour ridging and terracing

Many areas are not slightly sloping, but rather consist of rolling hills or steep mountain slopes. Irrigating these lands has long required leveling, lest the water will flow off and not soak into the soil. On moderate slopes, this can be accomplished by building series of low parabolic earthen ridges (10–50 cm high and 2–15 m wide) parallel to the contours. Excellent examples can be found in central Texas (30°36′40″N 97°22′45″W). On steeper slopes a series of dry masonry rock walls, called risers, and varying in height from 1 to 4 m, are built along contours. The areas behind or upslope of the risers are filled with sediment that is either washed in through erosion or brought in and deposited by humans. The nearly level planting surfaces, called treads, can then be irrigated. Some of the best examples of past and present irrigated terraces are in Peru’s Colca Canyon (15°30′20″S 72°15′55″W).

Furrows and grids

Developing methods to keep water on slopes is one issue irrigators face; managing it with care is another. If farmers have access to plows and draft animals they most typically create a series of ridges and furrows. If they do not have this technology, they use hoes and shovels to build small rectangular ridges oriented parallel and perpendicular to each other that forms grids or micro water-retention basins. Fields to the east and south of the ancient earthen pyramid in Cholula, Mexico, for example, are in some years plowed into furrows and in other years shaped into grids, depending on the crop grown (19°3′18″N 98°18′07″W).

Sprinkle

Garden hoses and lawn sprinklers, be they oscillating, pulsating, or rotating, are common landscaping devices. Similarly, and increasing in popularity, are built-in lawn sprinkler systems that consist of buried tubes with a series of pop-up sprayers. Although involving small areas around residences or businesses and not farmland, these sprinkler systems are in fact irrigation. Using the same principles, agricultural sprinklers are much larger, each covering more than a hectare. And, because they involve
pressure, water is pumped from aquifers rather than flowing by gravity from surface sources.

Perforated pipe irrigation systems used to involve 4–5-m lengths of aluminum pipes, 10–20 cm in diameter, that one person could easily lift, clamp together, and be dismantled once an area was wetted and reassembled on a plot to be watered. As the name implies, the pipes have several holes drilled along their length. Increasingly, however, aluminum which corrodes is being replaced by PVC. Ranchers who irrigate pastures are the principal users of perforated pipes. Traveling sprinklers are an elaboration on perforated pipes. They consist of a horizontal pipe that doubles as a water conduit and an axle for multiple wheels that allow for the entire device to roll slowly over rectangular plots. The pipes are not perforated, but instead have sprinkler heads attached. The sprinklers are typically of the pulsating variety.

Without doubt center pivot systems are one of the most technologically advanced, and certainly the most conspicuous, form of irrigation. These systems consist of pipes supported by series of triangular-shaped wheeled devices. Several sprinkler heads are typically hung below the pipes. Rather than rolling over rectangular fields, one end of the pipe remains in place while the rest of the device rotates around it. The resulting landscape consists of a number of circular fields. These landscapes may look under-used as the corners between the circular fields are not cultivated. In many locations, however, such as the desert of central Saudi Arabia (23°46′35″N 49°6′6″E), lands would not otherwise be irrigated.

Trickle and subsurface drip

Farmers living in arid lands have always sought ways to use water more efficiently. The development of plastic tubing was a major step forward in this regard. Developed in Israel in the 1950s and 1960s, trickle irrigation involves hundreds of meters of small diameter tubing either laid directly on the ground or suspended on trellises, depending on the type of crop grown (e.g., on the ground for lettuce, above the ground for grapes). At various intervals along the tubing small spigots allow measured amounts of water to trickle onto the base of each plant. Excellent examples of trickle irrigation exist in the Champagne region of France (49°11′54″N 3°56′40″E).

A recently developed variation on trickle irrigation involves burying tubing a few centimeters below the ground surface. This allows for even smaller amounts of water to be applied directly to the roots of each plant.

Summary

As long as people have been farming dry lands – and the origins of agriculture can be traced to these very environments – they have been developing means to capture, transport, and apply water efficiently to their fields. Some ancient irrigation techniques remain in use today, and new techniques will doubtless be developed in the future.

SEE ALSO: Agriculture; Dams; Groundwater; Surface water

Further reading

Jamaica: Jamaican Geographical Society (JGS)

Founded: 1966
Location of headquarters: Kingston
Website: http://www.mona.uwi.edu/geoggeol/JGS/JGS.htm
Membership: under 100 (as of December 31, 2013)
President: Kevon Rhiney
Contact: kevon.rhiney@uwimona.edu.jm, jamaicangeographer@gmail.com

Description and purpose

The JGS was primarily established to advance the discipline of geography in Jamaica. The society is managed by an elected council and its members include scholars, educators, students, and employees in the public, private, and NGO sectors in Jamaica as well as members of the public with a general interest in geography. Since it was founded the society has been actively engaged in the dissemination of geographic research and knowledge through its newsletter and various outreach and sensitization activities including panel discussions, workshops, guest lectures, field trips, and occasional exhibitions.

Journals or major publication series


Current activities or projects

The JGS has embarked on a number of outreach activities in recent years ranging from panel discussions and guest lectures through to its various educational programs and workshops. Since 2010, the JGS has held four panel discussions on topical issues affecting the Jamaican society and economy. The society recently held a panel discussion looking at the state of the Jamaican housing market post the 2008 global economic recession. Previous panel discussions have focused more on the state of geography education in Jamaica, by (i) looking at the challenges Jamaican high school teachers face in preparing students for the Caribbean Secondary Education Certificate (CSEC) examinations and (ii) examining the reasons behind the decision of some high schools to replace geography with social studies. Over the past four years the JGS has been supporting the Dream Jamaica summer program by hosting whole-day workshops on GIS under the broad theme of “Why Geography Matters.” Dream Jamaica was formed in January 2008 to help high school students in Jamaica visualize and realize their dreams (for more details on the Dream Jamaica initiative go to www.dreamjamaica.org). The society has recently been awarded a Global Greengrants Fund grant to assist in raising awareness of climate change among Jamaican youth. The dissemination/exchange of geography research and knowledge is done through the society’s newsletter and Facebook page.

Submitted by Kevon Rhiney
Japan: Jimbun Chiri Gakukai (The Human Geographical Society of Japan (HGSJ))

Founded: 1948
Location of headquarters: Kyoto
Website: http://hgsj.org
Membership: 1323 (as of September 30, 2013)
President: Masahiko Yamano
Contact: abeno-yamano@juno.ocn.ne.jp

Description and purpose

The society promotes the studying of human geography and promoting its advancement and development. It publishes Jimbun Chiri (Japanese Journal of Human Geography) every other month (English issue is published once a year). It organizes geographical meetings (annual meeting in November every year, regular meetings three times a year, and other specialty group meetings more than fifteen times a year). The society also makes contact with geographical and other academic institutions in Japan.

Journals or major publication series

Jimbun Chiri (Japanese Journal of Human Geography)

Current activities or projects

The society published Jimbun Chirigaku Jiten, The Dictionary of Human Geography on September 30, 2013 (Maruzen, Tokyo). The society supported the IGU Kyoto Regional Congress in August 2013. The society completed the registration as a general incorporated association for nonprofit legal entities in Japan in October 2014.

Brief history

The Human Geographical Society of Japan was founded in March 1948 in Kyoto as the academic association organizing most Japanese human geographers. It was originally established as The Geographical Society of Western Japan in October 1946 with some eminent geographers who lived in the Kinki district (western Japan). At present, this society has 1323 members and nearly all Japanese human geography researchers are members. The society publishes its official academic journal, Jimbun Chiri, the Japanese Journal of Human Geography (JJHG) in six issues a year. One issue in every volume of JJHG publishes a survey of geographical studies in Japan, an annual review of human geography. This geographical society also hosts the following five special study groups: (i) History of Geographical Thought, (ii) Historical Geography, (iii) Metropolitan Area Studies, (iv) Education of Geography, and (v) Political Geography. During 2000–2005 an Asian area studies group was also active. Each study group holds several research meetings a year. This society has published a total of 12 volumes of Bibliographies in Japanese Geographical Research between 1953 and 2009.

Submitted by Masahiko Yamano
Japan: Nihon Chiri Gakkai (The Association of Japanese Geographers (AJG))

Founded: 1925
Location of headquarters: Bunkyo-ku, Tokyo
Website: www.ajg.or.jp
Membership: 3000 (as of January 1, 2014)
President: Noritaka Yagasaki
Contact: office@ajg.or.jp

Description and purpose

Since its inception in Tokyo in 1925, the Association of Japanese Geographers (AJG) has played a leading role in the advancement of geographic research and education in Japan. It now has some 3000 members who are geographers with diverse backgrounds, specialties, and interests. Given that the AJG attained the status of “Public Interest Incorporated Association” in Japan on April 1, 2012, it is expected to contribute further to the dissemination and application of geographical knowledge in society. The AJG publishes an online English journal, in addition to two Japanese journals. The AJG hosts two annual meetings in the spring and fall.

Journals or major publication series

Geographical Review of Japan Series A. www.jstage.jst.go.jp/browse/grj/-char/ja
Geographical Review of Japan Series B. www.jstage.jst.go.jp/browse/geogrevjapanb
E-journal GEO. www.jstage.jst.go.jp/browse/grj

Current activities or projects

The AJG hosts annual meetings in the spring and fall to promote scientific exchange and interaction. Thirty-one study groups are approved within the AJG, exploring various geographical subfields and topics. The AJG has started publishing a new English publication series entitled “International Perspectives in Geography: AJG Library.” The AJG provides two types of accreditation: the Certificate of GIS Skills/Advance GIS Skills and the Certificate of Regional Research/Advance Regional Research. In order to promote geography education, the AJG holds an open lecture series for the general public and supports the International Geography Olympiad.

Brief history

The AJG is the oldest and largest academic association of geographers in Japan. It was established by 49 geographers in Tokyo in 1925 as the Geographical Review of Japan. Membership has grown steadily over the years (272 in 1940; 612 in 1950; 1880 in 1960; 2810 in 1970; 3090 in 1980). Although AJG membership has been stable in the past three decades, geography in Japan has experienced substantial changes. The founding of the English language journal, the Geographical Review of Japan Series B in 1984 symbolized the effort of Japanese geographers to internationalize their academic activities. More recently, the English journal has become an open-access online journal and seeks to become a leading English geographical journal in Asia. The AJG, the leader among the geographical associations in Japan, obtained the status of “Public Interest Incorporated Association” as of April 1, 2012, marking the beginning of a new era of development.

Submitted by Noritaka Yagasaki
Just-in-time production system

Neil Reid
University of Toledo, USA

Just-in-time (JIT) is a system of production that allows manufacturers to simultaneously reduce production costs, improve product quality, and quicken the pace of both product and process innovation (Reid 1995). It is a system developed by the Japanese automaker Toyota, beginning in the 1940s. A key characteristic of JIT is its emphasis on having component parts delivered to the point of production “just-in-time” to be incorporated into the manufacturing process. This permits manufacturers to reduce costs by minimizing the need to store component parts at or near the point of production. Under JIT, parts (or modules of parts) are delivered frequently (several times a day) and in small batches. In the 1960s, the Japanese introduced a process known as gaichu kanban, by which parts do not leave suppliers until an order for them arrives from the customer. Gaichu means “order” while kanban refers to the written instructions that provide the supplier with delivery details (quantity and delivery deadlines, frequency, and destination within the receiving factory). Under a JIT regime parts can be delivered as frequently as every two hours to the point of assembly (Kaneko and Nojiri 2008). Today the efficiency of the gaichu kanban process is greatly aided by the use of information technology (e.g., barcodes and laser scanners) and automation (e.g., automated storage and retrieval systems) (Fields 2006; Bowen 2008).

While there are many different aspects of JIT, geographers are primarily interested in understanding the spatial implications of the system. Interest in the geography of JIT began to emerge with the appearance of Japanese automobile assembly and component parts plants in the United States in the mid-1980s. Honda was the first Japanese automaker to open an assembly plant in the United States when they started producing automobiles in Marysville, Ohio, in 1982. They were quickly followed by other Japanese automakers such as Nissan (Smyrna, Tennessee, 1983) and Mazda (Flat Rock, Michigan, 1987). The Japanese automakers were quickly joined by Japanese auto component part makers who established production facilities in the United States in order to supply the assembly plants with component parts. Today there are 14 Japanese-owned automobile assembly plants in the United States and literally dozens of Japanese-owned component parts makers (Japan Automobile Manufacturers Association 2012). While JIT is utilized across a wide range of industries it is in the automobile industry where its use is perhaps most pervasive. The automobile industry, therefore, provides an excellent case study to highlight JIT’s spatial characteristics.

In an early study on the spatial implications of JIT, Kenney and Florida (1992) found its use to be widespread among Japanese automotive companies in the United States. Eighty percent of first-tier suppliers (who sell directly to the automakers) and 43% of second-tier suppliers (who sell parts to first-tier suppliers) deliver
parts according to a JIT schedule. Furthermore, they found that 41% of first-tier suppliers are located within a two-hour drive of their major customers. Indeed, proximity to automobile assembly plants was the most important site selection criteria for 90% of Japanese automobile component parts makers (Kenney and Florida, 1992). Reid (1990) found that 79.1% of Japanese automobile component parts makers in the United States were located in a state that also had a Japanese automobile assembly plant. The findings of Kenney and Florida are supported by Estall (1985), who suggested that JIT requires suppliers to be within 62 miles (100 km) of their customers. This need for spatial proximity between customer and supplier results in a geographically clustered customer-supplier network.

Later studies by Klier (1995, 1999, 2000) have challenged the whole notion that JIT necessitates spatial proximity and geographic clustering. Klier (2000) argues that the level of spatial clustering exhibited in the Japanese automotive industry is nowhere near as high as Kenney and Florida (1992) suggest. The most geographically clustered supplier network identified by Klier (2000) was Honda’s. Yet, only 29% of Honda’s suppliers are located within 100 miles of the assembly plant. Klier (2000) argues that JIT does not result in local-scale geographic clustering. Rather, clusters of suppliers are more regional in nature, extending over hundreds of miles. Thus, the median distance between Auto Alliance (a Ford/Mazda joint venture) and its Japanese suppliers is 286 miles (460 km). Klier (1995, 1999), however, notes that spatial clustering tendencies do seem to vary from automaker to automaker and according to whether the suppliers are domestic- or Japanese-owned. For example, Honda’s supply network is more geographically clustered than Diamond Star’s (a Mitsubishi/Chrysler joint venture). Twenty-six percent of Honda’s Japanese suppliers are located within 100 miles, compared to only 8% of Diamond Star’s Japanese suppliers. In a similar fashion, Japanese suppliers to Japanese assembly plants tend to be more geographically clustered than domestic suppliers to the Japanese assembly plants. For example, 19% of Toyota’s Japanese suppliers are located within 100 miles of the assembly plant. In contrast, only 5% of Toyota’s American suppliers are located within 100 miles of the assembly plant.

In Japan the distant delivery of component parts on a JIT basis is facilitated by the existence of third-party logistics providers who organize the delivery of various parts to distribution depots (cross-docks) that are located in close geographic proximity to the assembly facilities. Upon arrival the parts are sorted and delivered to the assembly plant on a JIT basis. Parts that are best suited for such long-distance delivery, consolidation, and sorting are those that are standardized and whose production can be planned well ahead of time (Kaneko and Nojiri 2008).

It is generally recognized that access to a well-networked and modern transportation system greatly facilitates the delivery of component parts on a JIT basis (Reid 1990). In contrast, an informative study of the Indian automobile industry demonstrated the impact that a poor and inefficient transportation infrastructure can have on JIT production systems (Gulyani 2001). Maruti, India’s largest auto assembler, has struggled to reduce inventories at its assembly plant. The relatively primitive highway system that exists in India increases unpredictability with regard to the length of time that component parts will be in transit. For example, a parts delivery that would take a predictable 1.5 days in Europe (2000 km) will take a highly unpredictable 6–9 days in India (2500 km). Maruti has found that suppliers who are within an 80-km radius are able to deliver parts with acceptable
frequency. Beyond that, the degree of unpredictability becomes problematic. Ford, which has an assembly plant in India, has addressed the transportation challenge in a number of ways – by encouraging its suppliers to locate in close proximity, by utilizing a distribution depot (cross-dock) system for more distant suppliers similar to that used in Japan, and by ensuring that hard-to-ship components (e.g., seats and fuel tanks) are manufactured nearby (Gulyani 2001).

In any discussion on the spatial impacts of JIT it is worth noting that the definition of what constitutes a JIT delivery schedule can vary widely. If choreographed perfectly a JIT system would eliminate all inventories at the point of production. Zero inventories is very difficult to achieve, however, with the result that different producers tolerate different inventory levels and hence can be said to be operating JIT more or less efficiently. In comparing Japanese and American automakers, for example, Delbridge and Oliver (1991) found that the latter, on average, had parts delivered on a significantly less-frequent basis (an average of 1.59 deliveries per day versus an average of 7.9 deliveries per day) with the result that they carried larger inventories at the point of production (an average of 8.1 days of inventory versus 1.52 days of inventory).

In summary, JIT is a production system that has at its heart inventory control. The studies cited here suggest that while geographic proximity between customer and supplier may be desirable it is not necessary for the smooth and efficient functioning of a JIT production system. The utilization of other methods such as distribution depots and cross-docking allows component parts to be delivered on a JIT basis from suppliers who are at great geographic distance.

SEE ALSO: Economic geography; Industrial agglomeration; Industrial geography; Industrial linkage; Trade, FDI, and industrial development

References


JUST-IN-TIME PRODUCTION SYSTEM


Karst processes and landforms

Jo De Waele
University of Bologna, Italy

All landscapes of the Earth are formed by a combination of physical and chemical processes that have acted on the exposed rocks over long periods of time. The different lithologies are weathered and eroded away and leave erosional morphologies, while the resulting sediments are transported by various agents (rivers, ice, wind) and deposited elsewhere, sometimes thousands of kilometers away from their source, forming depositional landscapes on land or at the bottom of the ocean. These landscapes can take their name from the dominant processes that have shaped most of the surface features (e.g., glacial or volcanic landscape) or from the prevailing rocks (e.g., sandstone or granite landscape).

The karst landscape takes its name from a region between Italy and Slovenia dominated by outcrops of carbonate rocks. Karst is the German wording, adopted during the Austro-Hungarian period when the first studies on dissolution-related landforms were conducted. It derives from the old Indo-European term “karra,” meaning “stone” (Palmer 2007), known in Latin as “carusardius,” and still used in the form “kras” in Slovenia and Croatia. Karst landscapes are formed mainly by surface and subsurface rock dissolution and are best developed in tropical and mid-latitude regions where carbonates and evaporites are exposed. Karst dissolution is subdued or impossible in areas with no rain (e.g., deserts) or where temperatures are always below zero (polar regions, high mountain areas). When a karst area is buried by other rocks and dissolution stops, the voids are often filled with sediments or mineralizations. Some of these can become important mineral resources, such as the Mississippi Valley Type deposits. These ancient karsts are referred to as paleokarst (Bosak et al. 1989). By contrast with all other landscapes, in karst chemical dissolution largely overrules mechanical erosion. Most of the rock is carried away in solution by surface and underground flowing waters and physical sediment production is very low, restricted to insoluble material, mainly clays. Because of the dominant role of dissolution over erosion, karst areas are characterized by a very distinctive morphology and hydrology (Ford and Williams 2007).

Although karst typically refers to limestone terrains, mainly composed of the mineral calcite, dissolution can also be the dominant geomorphic process in other rocks. Extensive cave systems have also been discovered in dolostone, composed of the mineral dolomite which has slower dissolution kinetics than calcite. Good examples of such cave systems, mainly developed in tectonically complex areas, are found in the Dolomites in Italy (Sauro, Zampieri, and Filipponi 2013). Gypsum and anhydrite are around 100 times more soluble than calcite. Karst in these evaporite rocks has been described in all continents and in a wide range of different climatic settings (Klimchouk et al. 1996; Calaforra 1998). The most soluble rock on the Earth is rock salt (halite). In regions where rainfall is very low and/or rock salt is continuously brought to the surface by salt diapirism (e.g., Dead Sea area in Israel, Atacama desert in Chile, Zagros Mountains in Iran), surface dissolution can create the same
Karst processes and landforms

Karst morphologies encountered in limestones, but at a rate more than a thousand times faster (Frumkin 2013). Also, quartz-sandstone terrains, in tropical areas such as the tepuis in Venezuela and Brazil, can host typical karst morphologies such as solution rills, caves, and underground drainage (Wray 1997). Although referred to by some authors as “pseudokarst,” there is unmistakable evidence that dissolution is an essential morphogenetic process in these landscapes.

The chemical processes responsible for the dissolution of the rocks are various and depend on the mineralogy of the rock to be dissolved (Dreybrodt 1996). In gypsum and rock salt the dissolution is a simple two-phase (solid-liquid) dissociation. The solubility of gypsum in pure water is around 2500 mg/L at 20°C, roughly 140 times lower than that of rock salt (halite, NaCl, 360 000 mg/L) and three orders of magnitude greater than that of calcite (1.5 mg/L). Dissolution of any salt increases in the presence of other salts because of ion-pairing effects. Saturation is reached rapidly and dissolution rates are boosted in the presence of turbulent flow.

In quartz-sandstone terrains, quartz (SiO₂) is dissolved forming orthosilicic acid (H₄SiO₄), reaching saturation at 8.7 mg/L in pure water at nearly neutral pH and 20°C. Amorphous silica is slightly more soluble, reaching a solubility of around 100 mg/L. Both quartz solubility and rate of dissolution increase with temperature. The dissociation of orthosilicic acid, which also increases the solubility of quartz, starts at neutral to basic pH. In these conditions the orthosilicic acid undergoes up to four consecutive dissociations (Mecchia et al. 2014).

But the most intensively studied karst processes are those related to carbonate rocks such as limestones and dolostones. Solubility of both calcite and dolomite in pure water at 20°C is very low, similar to the solubility of quartz, but increases by two orders of magnitude in the presence of slightly acidic waters. In most natural systems on the Earth, acidity primarily derives from epigenic sources, mainly from carbon dioxide (CO₂) present in the atmosphere and in the soil. CO₂ slowly dissolves into meteoric waters forming carbonic acid (H₂CO₃), thus reducing pH and increasing the corrosion capability of the percolating waters. Besides carbon dioxide, acidity can also derive from organic acids (humic and fulvic acids) or oxidation processes occurring in aereate conditions. These aggressive waters penetrate into the bedrock through available permeable pathways (fractures, bedding planes) and flow down-gradient in the saturated (phreatic) zone towards the discharge points (i.e., springs). In carbonate karst areas, most of the dissolution occurs close to the surface, in the first tens of meters of rock, known as epikarst, and rapidly decreases downwards as the saturation degree increases.

Many factors influence the chemical reactions involved in the dissolution of carbonate rocks, the most important of which is mixing of different waters. In the mixing zone between two different waters, or at the interface between fresh and saline groundwater, typical in coastal karst areas, chemical dissolution is enhanced. Mixing of two solutions saturated with respect to calcite but at different partial carbon dioxide pressures results in an undersaturated solution capable of dissolving more limestone. This mixing effect, known in the literature as the “Bögli effect,” makes dissolution possible also deep in the aquifer. In young limestones close to the sea (e.g., small carbonate islands), enhanced dissolution at the interface between the freshwater body and the denser salt water gives rise to the so-called flank margin caves (Mylroie and Carew 1990).

The acidity of the solutions can also derive from deep (hypogenic) sources, in the form of CO₂- or H₂S-rich rising fluids. These sometimes highly aggressive rising waters lead to the formation of the so-called hypogenic cave...
systems (Klimchouk 2007), some of which are created by fluids enriched in carbonic acid (e.g., Black Hills, South Dakota), and others by solutions enriched in sulfuric acid (e.g., Guadalupe Mountains, New Mexico). Since the source of acidity derives from deep underground and is brought to the surface along major discontinuities (e.g., deep faults), these systems are often found in areas with thermal springs (e.g., Budapest). The dissolving power of the fluids diminishes moving upward, often leaving no geomorphic expression on the land surface. Hypogenic caves are also present in gypsum areas, where undersaturated waters enter the gypsum beds from below in an interstratal karst setting. These fluids dissolve the gypsum beds along every possible initial open fracture, leading to the formation of sometimes impressive regular maze caves (e.g., Ukraine) (Klimchouk 2007). Some of the largest cave systems in the world are actually hypogenic caves, such as Lechuguilla Cave in New Mexico (a 222-km-long sulfuric acid cave in limestone) or Optymistychna Cave in Ukraine (a 236-km-long gypsum maze).

In general the word “karst” refers to a landscape dominated by unique morphologies that form because of surface dissolution of the host rock. This process creates a wide variety of small-scale landforms typical of karst: these are called karren. They occur on all soluble rocks, especially limestone, gypsum, and rock salt, but also on quartz-arenites. They vary in form from small pits and grooves, to rills (rillen), to runnels (rinnenkarren), and flat-floored solution pans (kamenitze). These centimeter- to meter-sized rugged landforms are typical of high mountain karst (Figure 1) in the absence of soil and cover. Karren are also very well developed in coastal areas, where the corrosive, burrowing, and etching actions of organisms exacerbate their shape. In this environment bioerosion becomes dominant, and the landscape is often called biokarst or phytokarst.

Karst areas often have a drainage characterized by closed depressions and generally lack normal river networks. These closed depressions are probably the most distinctive landform of karst, and are known as dolines, from the Slavic word “dol,” meaning “valley.” In North America and in engineering language dolines are called sinkholes. The presence of dolines in an area is diagnostic of the dominance of dissolitional processes over erosional ones. Dolines can have many shapes, from shallow saucer-shaped depressions to large vertical shafts, depending on the dominant processes that formed them: dissolution, suffosion, collapse, or sagging. They can have diameters ranging from a couple of meters to over a kilometer and normally have a subcircular plan form. When numerous dolines coalesce they form complex depressions historically named uvalas, a term that is falling into disuse. Collapsed dolines that give access to the underlying aquifer are known as cenotes, typical in Central American coastal areas. Dolines occur both in carbonate and in evaporite rocks, with a greater frequency and dimension in the latter mainly because of their higher solubility (100 times more than the carbonate rocks), and their lower mechanical strength and higher ductility (Gutiérrez et al. 2014). Generally, dolines are classified as either solution dolines or subsidence dolines. The first develop because of focused dissolution in zones of higher permeability, resulting in the gradual and differential lowering of the ground surface. In general, this type of doline forms in areas with exposed or barely soil-covered soluble rocks. Subsidence dolines, on the other hand, form because of subsurface dissolution and gravitational downward movement of the overlying material. These are classified according to the type of material involved in the lowering (cover, bedrock, or caprock) and the main subsidence
process (collapse, suffosion, or sagging) (Gutiérrez et al. 2014). Cover refers to overlying material composed of unconsolidated deposits (e.g., soils, alluvium), bedrock is the soluble rock itself (e.g., limestone or gypsum), and caprock is a covering solid nonkarstic rock. The downward movement of the material can involve brittle deformation (collapse), slow washing away of loose deposits into underlying subsurface voids and settling of the overlying deposits (suffosion), or slow ductile bending of sediments (sagging). Cover collapse sinkholes (or dolines) are the most spectacular and dangerous ones. Sometimes more than one type of material and several processes can cause the formation of a complex doline, which will be classified using combinations of the above mentioned terms, putting the dominant material and/or process first, followed by the secondary one.

Dolines can occur as isolated individuals, in groups following the direction of fractured areas or the allogenic recharge boundary of the karst area, or can occupy all of the available surface area. This egg-box-shaped pattern of dolines is known as polygonal karst, and occurs in tropical and temperate areas characterized by high
values of rainfall (e.g., Papua New Guinea, New Zealand). The large density of dolines in these areas creates an extremely efficient autogenic drainage pattern, conveying surface waters very rapidly underground. The size of the dolines in these areas depends on a combination of lithological and structural factors, as well as climatic ones.

Another typical landform that derives mainly from dissolution of soluble rocks, in particular, limestones, is the karst plain, often referred to by the Slavic word “polje,” meaning “large plain.” These almost perfectly planed surfaces cut across the rock structure and their floor corresponds to the mean level of the water table. This means they flood during the rainy season, and remain dry for the rest of the year. Their floor is covered with a very thin cover of residual (insoluble) material with bedrock cropping out locally. Karst planation surfaces generally develop at the inflow and outflow boundaries and can be considered as the final stage of karst denudation in carbonate rocks. Often these large depressions are controlled by structural elements, such as grabens or large synclines. The soluble rock surface is lowered vertically until the water table is reached, after which dissolution enlarges these depressions laterally at the level of the water table. This produces a planation surface inclined parallel to the hydraulic gradient. Ford and Williams (2007) distinguish three main types of poljes: border poljes, structural poljes, and base-level poljes. The best examples of these poljes can be seen in the Dinaric karst.

Often the karst denudation is not complete, and some parts of the original bedrock are left standing out on the planation surfaces. Small hills like this are called hums in the Dinaric karst and are frequent in temperate areas. In a tropical climate, residual hills can instead be 200 m high and have steep to almost vertical sides. In China, isolated hills like this are called fenglin (peak forest), while clusters of residual hills sharing a common exposed bedrock base are known as fengcong (peak cluster). The first are formed close to the water table level (and thus erosional base level), while the latter are formed far above this level (and therefore lack the typical flat plain in between). The Spanish-derived term for the towers is “mogotes.” The intervening karst plains are formed in a similar way to the aforementioned poljes, as horizontal swamp notches at the base of the towers confirm. The presence of several of these notches along the tower walls reflects the changes in local water table level.

Surface rivers are absent in typical karst areas, and the drainage pattern is mainly characterized by dolines. Small allogenic rivers, for which the drainage basin is partially located outside of soluble rocks, normally disappear underground soon after they enter the karst area, forming blind valleys (valleys which end “blindly”). When allogenic streams are greater, water will flow on the karst surface forming a true valley, sinking along its path into more permeable areas (losing streams). When large amounts of water are conveyed underground through point sinks these are called swallow holes (also named swallets or ponors). The discharge carrying capacity of these swallow holes often does not allow all of the water to disappear, and the river continues downstream until all the water is absorbed by the karst rocks. Downstream of this last swallow the valley becomes dry for most of the year, and activates only during high discharge periods. Some rivers may maintain their flow throughout the entire karst area until flowing out of it. This may happen because the flow rate exceeds the capacity of the karst to absorb, or because the hydraulic gradient is very low, not allowing the water to penetrate underground rapidly enough. High hydraulic gradients in fact enhance the vertical percolation of surface waters. The river valleys in karst areas are generally characterized by steep walls, typically forming canyons (or gorges).
KARST PROCESSES AND LANDFORMS

Many of these canyons are formed gradually in a slowly uplifting region (antecedent valleys).

Since primary porosity in both limestones and evaporites is very low, surface waters in karst tend to penetrate underground through all available fractures. Passing through these variable pathways, dissolution continues and permeability gradually increases, resulting in an enhanced underground water flow and a consequent local groundwater lowering. Thus an aerate (or vadose) zone is created in the upper part of the karst aquifer, where water percolates downward or is held in place by capillary forces. This upper part of the vadose zone, extending immediately below the karst surface, is known as epikarst (or the subcutaneous zone), and can be regarded as the extension of the surface karren forms. This is where most dissolution of rock takes place in carbonates, since CO₂ concentrations are higher there. This network of fissures enlarged by dissolution often contains infillings of soil that can retain moisture over longer periods of time, gradually releasing their water downward by gravity. This subcutaneous habitat is also extremely interesting from a biological point of view. Epikarst usually ranges from a couple to some tens of meters deep. Below this epikarst, water flow is focused along a few larger voids that can extend very deep into the bedrock and give rise to cave systems, or are feeders of active caves below.

Dissolution processes, together with mechanical erosion along underground pathways, give rise to the formation of three-dimensional systems of conduits, and solutionally enlarged discontinuity planes (fractures, bedding planes), forming extremely complex three-component permeability aquifer systems. The direct connection between the surface and the karst aquifer causes underground waters to be extremely vulnerable to pollution. The high secondary permeability of karst enables waters and pollutants to be transported quickly and often without undergoing appreciable chemical and physical changes from their input points to the springs (Goldscheider and Drew 2007). Due to this rapid transfer, flow rates can also change by three orders of magnitude in a relatively short time span. During floods, pollutants stored underground during low flow conditions can be flushed out. Karst aquifers drain toward springs, which often have considerable flow rates. The largest springs in the world are most probably karstic ones, reaching flow rates of over 100 m³/s. The location of these springs is often controlled by the regional or local base level (e.g., sea level), or by a less permeable threshold (e.g., underlying or adjacent rocks, alluvium), sometimes caused by tectonic features (faults). Since sea level changed during the Quaternary, some springs can be located close to or even below sea level, giving access to underwater cave systems. These springs, typical of carbonate islands but also present in other geographical contexts, are known as blue holes. Headward retreat of the spring can produce a pocket valley with the spring at its upstream end, typically at the foot of a high limestone cliff. Base-level rise, due to valley aggradation, can cause vertical vadose passages to be flooded creating very deep springs, known as Vauclusian springs, named after the Fontaine de Vaucluse (France). Some springs become active only during floods, and are located above the regional water table. A special type of temporary springs is the so-called estavelles, which act as swallow holes during normal (low) flow, and become active as springs during floods.

Cave genesis (or speleogenesis) is a complex topic, involving different sciences, mainly geology, chemistry, and physics. Dissolution is an important process, but is generally not responsible for most of the created voids, since once mechanical erosion starts, this process largely overrules chemical dissolution. Dissolution,
however, is fundamental, especially in the initial phreatic phases of cave formation, when laminar flow is the only possible mechanism of water movement through the tiny little fractures and openings in the rock mass. This initiation phase in cave formation can last some thousands of years in limestone, and some decennia in gypsum. Once the width of the openings reaches a critical threshold, a little less than a centimeter, turbulent flow is established and flow rate increases. This allows the water present in these enlarged openings to remain chemically aggressive over longer pathways, enlarging this flowpath more than the neighboring ones. This causes, in the short term, the selection of this path as the main one, focusing most of the water flow through it, boosting its widening even more. This rapidly brings to the formation of a true cave passage, in a period similar to the initiation phase. These passages eventually will become air-filled and can accommodate sediments and speleothems, or continue growing through collapses. Complete infilling or surface dissection, with the creation of so-called unroofed caves, finally ends this cave’s cycle.

The overall cave pattern mainly depends on the origin of the dissolving waters, on the type of recharge, and on the nature of the guiding structures (Palmer 1991). In general, cave patterns depend on the local geological settings, including the distribution and type of soluble rocks, their structural setting, and the relative (and variable) position of recharge areas and points and the outlets (generally springs). Caves develop very rapidly (in a couple of years) in rock salt outcrops, in a matter of a hundred years in gypsum, and in some thousands of years in limestone and dolostone. The initial phases of speleogenesis will tend to favor the most permeable pathways, mainly consisting in fractures and bedding planes. The initial geological and structural conditions are often reflected in the morphology of the cave passage as a whole and on the cave wall sculpturing features. Differential corrosion-erosion creates typical cross-sections along more soluble beds, weaker bedding planes, and fractures.

As described above, the aggressiveness of the waters can have a surface origin (CO₂ in the soils and the atmosphere) or a deep origin. The first waters create so-called epigenic cave systems, the latter hypogenic caves. Recharge in epigenic caves can be allogenic (from outside the karst area) and autogenic (from inside). Allogenic input often is concentrated, while autogenic input can be either concentrated or diffuse. Recharge can also be indirect, through overlying nonkarstic aquifers. In general, epigenic caves form in unconfined conditions, with water flowing from a recharge area into the karst and then flowing by gravity following the hydraulic gradient to the outputs, in general important springs. Most of the upstream cave volume is produced in vadose settings, epiphreatic (flood-water) conditions prevail close to the water table, while in the downstream end phreatic conditions occur more frequently. Typical vadose passages are shafts and canyons, along which the waters tend to descend along the steepest available pathway, often represented by the dip of the beds. Once the water table is reached, the water no longer has a need to descend rapidly, and will follow the hydraulic gradient choosing the most efficient openings. Cave passages are often completely filled with water and tend to dissolve and erode the bedrock away in a homogeneous manner. This creates phreatic tubes that are roughly cylindrical in shape and, although they loop in both horizontal and vertical directions, they often more or less follow the strike.

The history of a cave can often be reconstructed based on the careful observation of the overall morphology of the system, but especially through the analysis of rock wall sculpturing
and its relation with the hosted sediments. Flow direction, for instance, can be inferred from the shape of scallops, erosional cusps left by turbulent flowing waters. Also the size of these scallops allows the flow velocity at the time of their formation to be estimated: the bigger they are, the slower the water was flowing. This ultimately allows flow rates in now abandoned conduits to be estimated based on scallop size and the passage's cross-section. Estimation of conduit cross-section size can be made difficult where sediments were present at the time of scallop formation. Underground rivers, in fact, can accumulate sediments on their floors, which shield the lower portion of the passage from corrosion-erosion. The water will thus be able to cut into the walls creating alluvial notches, slightly undulating subhorizontal lateral grooves on the walls. If alluviation fills the passage almost completely, phreatic conditions will be created and the water will corrode and erode the roof only. If sedimentation continues synchronously with this upward dissolution, ceiling channels or an anti-gravitative canyon can develop. Sediments can later be washed out from these passages leaving a canyon-like passage. This anti-gravitative erosional process is known as paragenesis, not to be confused with the minero-petrographical term. Para-
genesis in caves can be related to increased sediment input into the system, which can be induced by climate changes or can be caused by anthropogenic activities (e.g., deforestation).

The most common big epigenic caves are the typical underground rivers, and their genesis is strictly related to what happens at the surface. Developing over thousands of years, they reflect both climate and landscape changes, and their speleogenetic study can help unravel the paleo-geography and paleoclimatology of the area in which they have formed. Sinking points can shift upstream, creating new cave tributaries, while springs can move to lower elevations, following base-level lowering (e.g., valley entrenchment). This can create a set of cave “levels” that can be correlated with surface morphologies such as fluvial terraces.

In hypogenic caves, recharge comes from below and dissolves the rock upward. These cave systems are often located in regional ground-water systems where structural reasons cause fluids to rise (e.g., along major deep faults) and, when these waters come in contact with soluble rocks, create underground dissolution voids. In the typical intrastratal systems, the soluble bed is sandwiched between a lower-lying and an upper aquifer bed. Hypogenic caves in gypsum do not require acid waters, and caves can form as long as the rising fluids are undersaturated with respect to gypsum. This ascending speleogenesis acts along any available discontinuity and normally creates maze cave systems, some of which can reach kilometric development (e.g., Ukraine). Hypogenic caves in limestone, by contrast, need acidic waters able to dissolve the calcite (or dolomite). Since the waters often come from deep below the surface, the geothermal gradient causes them to be thermal. The decrease in pressure, due to rising, causes CO₂ to be released from the waters, while, conversely, part of this gas dissolves in the water again because of the decreasing temperatures, making these fluids become acid. Important thermal cave systems have been described in many parts of the world, such as in Budapest (Hungary) and the Black Hills (United States). In other cases the rising fluids can be enriched in H₂S, often deriving from the reduction of sulfates such as evaporitic rocks or hydrocarbon reservoirs. This gas, when arriving in an oxidizing environment, such as a cave or the shallow phreatic environment, turns into sulfuric acid that reacts with the limestone.
forming gypsum and releasing CO₂, which further enhances the dissolution of more limestone (Egemeier 1981).

Another peculiar type of cave is the flank margin cave, typically formed in relatively young carbonates close to the present or past sea level. In this zone freshwater-seawater mixing creates an area of enhanced dissolution forming a set of interconnected voids more or less parallel to the coastline and along a horizontal plane, corresponding to the sea level at the time of their formation (Mylroie and Carew 1990). Flank margin caves lack morphologies and sediments typical of underground streams. Cave passages tend to pinch out moving away from the mixing zone (and hence coastline). This type of cave is typical of carbonate islands, but has also been described in continental coastal limestone areas.

Once a cave has formed, it starts accumulating material, including clastic sediments, chemical precipitates, and organic debris. Cave entrances are often characterized by pitfalls, or sinks, and tend to collect surface material that falls in accidentally (animals) or is washed in by surface runoff. This creates preferential sedimentation sites of sometimes very valuable deposits that can be preserved over very long periods of time, while surface deposits are being stripped away by erosion. Caves can also be used by animals, including humans, as shelters or living places, or for ritual reasons. Caves have delivered some of the most important archeological sites for the reconstruction of the evolution of humans and their art (e.g., the Sterkfontein caves in South Africa with their Australopithecus finds, or the Lascaux cave in France for its wonderful Upper Paleolithic rock art).

Cave sediments can help to reveal at least part of the cave’s history, but because of extensive reworking and redeposition, they are often very difficult to study (Sasowsky and Mylroie 2004). Alternations of fine and coarse sediments in relatively undisturbed sedimentary sequences in caves reveal changes in flow in the underground stream. The mineralogical and petrographical composition of sediments can tell where they come from, and reveal possible changes in recharge areas. Sediments can sometimes be dated using paleomagnetism, in combination with U–Th dating of flowstone beds (Zupan Hajna et al. 2008). Allogenic quartz grains can be chronologically constrained using cosmogenic Al–Be dating techniques (Granger and Muzikar 2001). All these dating techniques deliver ages younger than the cave itself.

Inflowing and percolating waters introduce not only detrital particles but also dissolved species. Reactions between these fluids, the host rock, the sediments in the cave, and the atmosphere cause minerals to precipitate as single crystals or, more commonly, as aggregates known by the general name of speleothems. Over 330 different species of minerals have been described forming in caves (Onac and Forti 2011). Mechanisms of mineral formation include a variety of processes, such as CO₂ loss, evaporation, alteration of material already present in the cave, oxidation–reduction, cooling–heating, and changes in pH, just to mention the most important ones. The most frequent minerals are of course calcite and aragonite, followed by gypsum and other carbonates and sulfates. Where bat guano or bones are present, a variety of phosphates can be found. Some minerals are intimately related to the cave’s genesis, such as in the case of sulfuric acid speleogenesis, and their dating can give clues on cave formation (e.g., K–Ar and Ar/Ar dating of alunite) (Polyak et al. 1998).

But caves are especially appreciated for their hosted speleothems (Figure 2), such as the famous stalactites and stalagmites. These can be classified according to the nature of the water that deposited them (Hill and Forti 1997), or to the way single crystals and their aggregates
form the external shape of these cave formations (ontogeny, Self and Hill 2003). Most speleothems are formed of calcite or aragonite. Among the most common forms are: stalactites, stalagmites, columns, flowstones, soda straws, helictites, cave pearls, and rimstone dams. A review of most of these is given in Hill and Forti (1997).

Speleothems are important climate and environmental archives, characterized by annual layering of chemical origin, allowing the reconstruction of environmental and climatic changes at high temporal resolution (Fairchild and Baker 2012). The advantage of speleothems compared to other continental archives is the fact that they are datable with great precision with techniques based on the radioactive decay of uranium. Thanks to the introduction of MC-ICP-MS (Multicollector Inductively Coupled Plasma Mass Spectrometry) it is possible to date small (less than 0.2 g) samples of carbonate with high precision (2 sigma) of ±1% in the interval between 0 and 500,000 years BP. In recent years, the research on speleothems has focused on the isotopic composition (oxygen and carbon) of the carbonate of which they are composed and on their axial growth rates in order to reconstruct an annual and secular scale of climate variability, and to recognize the effect of internal and external climate forces on global climate changes. Recently, the study of the variability in speleothem trace element concentrations (Ba, Mg, Mn, P, Sr, U, etc.) through time has emerged as a useful hydrogeological marker that allows researchers to reconstruct rainfall, from seasonal scale to millennial scales, and to obtain information on the residence time of the water in the aquifer (often related to climate), with important consequences for the extension of phenomena of water-rock interaction. Speleothems have also been used for the reconstruction of Quaternary sea level changes, hydrological conditions in continental and marine environments, paleo-seismicity during early Quaternary, and relationships between climate change and archaeological events.

Many caves around the world have become tourist attractions because of their aesthetic value. A fee is charged to gain access to these caves, and normally visits are carried out following equipped pathways and under the surveillance of a cave guide. Although the first real show cave is Vilenica in Slovenia (where invited people have paid to visit the cave since 1633), most show caves developed in the second part of the twentieth century. The development of a show cave often requires modifications to be
made to the cave system, such as construction of pathways, opening of new or more easy access, installation of electric lighting systems, and a whole set of infrastructures both in the cave and above it or around its entrance. These operations often cause damage to the cave and the surrounding landscape, and environmental impact assessments should be carried out prior to show cave preparation to minimize these negative impacts (Cigna 2012). Special care should be taken in the selection of paths, lights and their position, and in the timing of the visits, in order not to exceed the so-called visitor carrying capacity, the maximum number of visitors that can enter the cave in a certain time frame without changing the environmental parameters irreversibly. In other words, modifications to the underground environment (e.g., temperature rise, CO2) should be recovered by the system at least over night, with return to natural conditions.

The protection of the cave and its environment is important to safeguard the mineral beauty of the cave (i.e., speleothems), but also to create as little disturbance as possible to the sometimes extremely fragile cave habitat. The underground voids, in fact, including not only caves but also the tiny fractures, are a habitat to an incredibly various and more or less specialized cave fauna (Culver and Pipan 2009). Some species are completely adapted to life underground: they have lost their pigments and eyes, have developed their other senses extremely well, and spend their entire life cycle in the dark and wet environment of a cave. These species, called troglobites, would not be able to survive outside and can be endemic to very small areas or even single caves. Troglophiles (cave-loving species) are species spending part of or their entire life cycle underground, being adapted to this environment, but can also be found outside (under rocks). Trogloxenes live close to the cave, and use caves as shelters (e.g., bats).

Even the slightest changes in cave environmental parameters can cause troglobites to migrate, and in the worst cases to succumb, resulting in loss of biodiversity. Modifications also include those that happen outside of the cave, at the surface or in the drainage area (water pollution, deforestation and soil loss, urbanization). While the extinction of a troglobite can pass unnoticed, the decrease in the population of large bat colonies in a cave can have a dramatic impact on the surrounding area, since bats are insectivores and therefore reduce the need for pesticides. Subterranean karst environments are also unique microbiological habitats. This biodiversity is worth protecting, and ongoing research on new cave-dwelling species and microbiological communities might allow the discovery of new substances useful for medical purposes (Barton and Northup 2007).

In the past decennia, public awareness of the unique character and the fragility of karst areas has increased enormously, especially in Europe and America, but still needs to be stressed in many countries. It is now understood that if we want to safeguard this fragile environment, and enjoy its multiple resources for centuries to come, we need to find a way of living on karst.

SEE ALSO: Applied geomorphology; Aquifers; Biodiversity; Coastal erosion processes and landforms; Earth system science; Exploration; Landforms and physiography; Natural hazards and disasters; Paleoclimatology; Quaternary geomorphology and landscapes

References

Barton, Hazel A., and Diana Eleonor Northup. 2007. “Geomicrobiology in Cave Environments: Past,


Further reading


Kazakhstan: RespublikalykoFamdybirlestigiKazaKstandyKUlttykGeografiyalykoFamyRepublican Public Association of Kazakhstan National Geographic Society (KNGS))

Founded: October 2013
Location of headquarters: Astana
Website: http://kazgeography.kz/
Membership: 122 (as of January 1, 2014)
Chairman: Karim Massimov
Contact: Assel Baibakisheva at kazgeography@nu.edu.kz

Description and purpose

KNGS is a public association founded in October 2013. KNGS unites the scientific and creative geographic community. Its goal is the development of geography and sciences of nature, society, and culture through the collaboration of the domestic scientific community and the development of cooperation with international associations of scientists.

KNGS concentrates on the following areas: geography and climate; environment and natural resources; biodiversity; tourism, travel and regional ethnography; and historical and cultural heritage, and ethnography.

Journals or major publication series

KNGS’s journals are under development.

Current activities or projects

As a young nonprofit organization, KNGS plans to expand its network nationwide and worldwide by opening eight branches in different regions of Kazakhstan and establishing ties with major international geographic societies. Scientific exchange and interaction occur through the KNGS’s annual conferences, joint research projects, and other collaborative activities. KNGS projects extend to a study of the Silk Road and different Kazakhstan regions, historical sites, nomad civilizations, national and international scientific expeditions, ecotourism and science tourism, restoration of fauna, study of migratory birds, and so on. KNGS also focuses on educational programs and popularization of geography.

Brief history

Before the establishment of KNGS there were a few geographic societies in Kazakhstan. KNGS was founded in 2013 by geographers of the Nazarbayev University, Kazakhstan, to unite the existing geographic societies under one umbrella.

KNGS’s mission is to promote research in the field of geography and related sciences, spread geographic knowledge among the population, and develop cooperation with international scientists, organizations, and geographic societies.
KAZAKHSTAN

KNGS is governed by a board of trustees and a managing council. The first conference of the KNGS, which gathered enthusiasts and experts in geography and related sciences, was held on October 8–9, 2013, at the Nazarbayev University. The conference was dedicated to the state of modern geography in Kazakhstan, its trends and perspectives, and the role of KNGS in solving problems at the national level with participation of academics, NGOs, and public organizations.

Among the first projects of KNGS are expeditions to Ulytau and Zhongar Alatau; restoration of Przewalski’s horse and the Turan tiger in Kazakhstan; a study of the snow leopard in Kazakhstan; the creation of a web museum of Kazakhstan that will list sites of historical, cultural, and natural heritage; the creation of short educational TV programs on historical sites of Kazakhstan; exploration works in south and west Kazakhstan; the development of infrastructure for local tourism; and many other projects.

Submitted by Vladimir Litvak
The term “knowledge-based economy” is used to characterize economies, national or regional, that are primarily driven by the “production, distribution and use of knowledge and information” (OECD 1996). The knowledge-based economy is thus different from economies that are mainly driven by other production factors such as capital, unskilled labor, or resource extraction.

The term began to appear in scientific literature in the 1990s, indicating that the development of the knowledge-based economy is a recent phenomenon (Foray and Lundvall 1996; Dunning 2000). Prior to the 1990s, capital and labor were usually regarded as the main drivers of economic development. With the shift toward stressing the importance of knowledge – or technology – as a production factor, a number of other related terms entered the academic discussion, for example “innovation.”

The knowledge-based economy has at least three important meanings. First, from the perspective of economic growth theory, it highlights the relative importance of different production factors. The standard production function in economics, developed in the 1950s and widely used in the field of regional studies/economic geography, highlighted the role of capital and labor. Economic growth was seen as a result of a balanced growth of both labor and capital, while fundamental changes in the way these two factors relate to each other, and changes in the effectiveness of their use, were attributed to technical progress, that is, new knowledge. The fact that knowledge was responsible for economic growth, and that it was more important than the growth of the production factors capital and labor, was empirically proven, and the focus of economic growth theory gradually shifted toward incorporating knowledge into the growth models (new growth theory; Romer 1990). It also resulted in a growing interest in basic innovation, technological paradigms, and their effects on economic growth and structural change.

With reference to growth processes in advanced economies, there is a broad consensus today that the creation and use of new knowledge is decisive, contrary to situations decades ago, during which economic growth was often fueled by an increase in labor (e.g., during times of rapid urban growth or mass immigration), the exploitation of natural resources, or a rapid growth in capital use (e.g., during the Industrial Revolution). Hence, the term “knowledge-based economy” denotes the attention paid to knowledge by economists.

Second, the knowledge-based economy acknowledges the fact that the industrialized countries, mainly the United States, Western European countries, Japan, and some other Asian countries, are investing heavily in the production of knowledge. Until the first half of the twentieth century, the production of new knowledge was usually not at the center of political and corporate decision-making. Technical progress was a result of trial-and-error processes in firms as corporate research and development activities were usually very limited. Public attention to knowledge creation meant providing schooling. Investment in basic science and tertiary education was limited to a few places, organizations, and individuals.
KNOWLEDGE-BASED ECONOMY

With the event of mass higher education in the 1960s and 1970s, and with a huge growth in the numbers and sizes of public research organizations, university research, and public research funding after World War II, the creation and dissemination of knowledge reached a new dimension. New knowledge was regarded as a tool for solving the great challenges of the modern world, ranging from providing cheap energy (or, today, clean energy), to fighting cancer, developing new means of communication, and achieving economic and military superiority. Hence, technical progress has accelerated and so has the dissemination of expertise with highly qualified engineers, scientists, and so on. The knowledge-based economy, and the growing importance of knowledge compared to other production factors, have thus been made possible by the expansion of science and education.

Third, the knowledge-based economy denotes changes at the corporate level that fundamentally affect the structures and dynamics of economies. The development of information-processing technologies (e.g., computers) and communication technologies (e.g., telephone, fax, the World Wide Web) has lowered the costs of interaction between firms. It has thus created opportunities for separating the activities necessary to produce a commodity (or a service), as independent firms that specialize in the different activities necessary for the production of one product can easily communicate with one another and bring together their contributions. Under these circumstances, companies have an incentive to concentrate on very few core activities that exactly match their core competence (Prahalad and Hamel 1990).

A concentration on core competencies and a specialization in related core activities require certain strategic actions. Companies then strive to be regarded as a producer of a distinct, special component or service. Technology leadership – or leadership regarding quality, customer orientation, and so on – is a necessary goal for a business strategy that is based on a narrow specialization. In order to be able to achieve such a leading role, a company’s core competence must be constantly increased. This requires recruiting specialists, carrying out research and development (R&D), and monitoring technical developments in related fields.

When the production of goods and services is not organized according to vertically integrated large firms (as it used to be until the late twentieth century), but is instead organized by a network of highly specialized producers, the whole economy is affected. For example, the need for a specialized workforce and stimulus for public research increases for each economic activity. Hence, the knowledge-based economy is an economy made of specialized firms with specialized personnel, and it calls for and draws on specializations regarding many economic activities that used to be covered by larger and unspecialized organizations.

From the point of view of the first and the second dimensions, the knowledge-based economy is necessarily related to hi-tech industries, innovation, and science-driven product development. The third dimension, however, is not necessarily linked to science and high-tech. There, the economy is based on knowledge, as it produces innovation through specialization in all parts of the economy, be they high- or low-tech. Companies can specialize in producing components that seem to be far removed from research and higher education, but the production processes or the logistics behind these components may still benefit from specialization and a core competence made of adequately skilled workers.

Hence, this understanding of knowledge-based development denotes strategic decisions at the firm level, most importantly the shift toward a
narrow focus on competencies and products, and, following on from this, certain structural changes in the organization of the economy, namely a higher degree of separation of production processes and of collaboration and networking. The growing attention to networks of all kinds, not only in geography but also in all social sciences, is inseparably connected to the emergence of the knowledge-base economy (OECD 1996).

From this brief discussion of the dimensions of the knowledge-based economy and the processes behind it, several important issues emerge that can only be touched on here. First, it is important to note that the knowledge-based economy is not only developing in the advanced economies. Although higher education, public science, and corporate R&D are concentrated in the advanced economies, many developing countries are catching up. The rise of China’s education and science sector is the most striking example of a change that fundamentally affects the composition of production factors in many developing countries, and provides the basis for knowledge-based developments. Moreover, companies in developing countries are part of global value chains and global production networks, and thus have the chance to specialize, to upgrade and learn, and to compete on the basis of knowledge. The latter point is illustrated by the rise of multinational firms from developing countries.

Hence, the knowledge-based economy has increasingly become a global description of the economy, and it is unlikely to be a short-term phenomenon. As long as companies are more profitable when they specialize and outsource, that is, as long as communication and transport are cheap, the knowledge-based economy will continue to be the key feature of today’s economic structures and dynamics.

This has profound implications for economic geography that are already visible in the major trends of the recent academic debate. Networks and networking, the nature of contacts and collaboration, the issue of university–industry linkages, the spatial consequences of public commitment to education and science, knowledge sharing and knowledge spillovers on various scales, as well as the notions of tacit knowledge and sticky knowledge, dominate a large part of the work of economic geographers.

Politics will have to respond to the need for education and science, and will have to ensure that conditions for corporate R&D are favorable. Moreover, communication and the openness of information flows, as well as trade flows, need to be on the agenda.

**SEE ALSO:** Cultural economy; Innovation and regional development; Information and communications technology; Technology

**References**


Knowledge-intensive business services and regional development: strategic importance of

Richard Shearmur
McGill University, Canada

Although it has become commonplace to refer to a certain category of services – high knowledge content intermediate services – as knowledge-intensive business services (KIBS), these services are also referred to as high-order business, or producer, services (Wood 2006). This semantic point is useful for understanding the connection between KIBS and regional development because much of the work by geographers on the topic, under the influence of Christaller (1966), uses the latter term (Coffey 2000).

This entry will briefly discuss the structural changes to the economy that explain the rise of KIBS, focusing more specifically on the way in which regional economists and economic geographers have perceived these services as actors in the revival of local economies. Whilst there is consensus on the role that these services play in assisting their clients in a wide variety of business functions, information gathering, and knowledge transmission, understanding of the geography of this assistance – that is, where KIBS firms locate relative to their clients – has evolved and continues to be the subject of debate (Wood 2006). The purpose of the paper is to flesh out these statements, explaining the way ideas have evolved.

The rise of KIBS

Since the 1950s, and especially since the rapid decline of manufacturing employment from the 1970s onwards, economies in the Western world have been shifting from secondary (manufacturing) to tertiary (service) employment. This has led to concerns about the decline in productive capacity, but also to detailed assessments of the causes of this shift and of the new role that service sectors – high knowledge content services in particular – play (Daniels 1985).

Broadly speaking, three processes are evoked to explain these structural changes (Bryson, Daniels, and Warf 2004). First, the “New International Division of Labor” has led manufacturing activities to seek cheap labor in developing countries, whilst Western countries specialize in high-knowledge content services. This has certainly occurred, though it probably is not the principal explanation of KIBS growth. Second, the vertical disintegration of firms, under the impetus of new technologies and new competitive pressures, led to the rise of an independent producer service sector as management, research, accounting, and other service functions were hived off. This same process – driven partly by the automation of production – also accounts for some of the decline in manufacturing employment. Third, there have been fundamental changes in the way value is created: in affluent societies in which all basic needs are met, consumption increasingly turns towards symbolic goods. (This does not mean that each person’s basic needs are met. It means that enough is produced to meet all basic needs: poverty and inequality cannot be attributed to lack of resources or productive capacity.) Firms
are required to innovate in order to distinguish themselves from competitors, and continuously renewed information and know-how are required in order to do so. In addition, as the economy globalizes, regulatory frameworks, logistics, management, and markets have become more complex. Specialized high-order business services are required to penetrate these information thickets, to identify and compile useful information and to pass it on to their clients, either by way of simple information, as subcontractors, or by way of collaboration (MacPherson 2008; Shearmur and Doloreux 2015).

There is still debate as to which of these processes carries most weight in explaining the rapid rise in the role of KIBS. Furthermore, some would argue today that creative industries are taking precedence, and that KIBS – almost systematically the fastest growing sector in Western economies from the 1970s to the early 2000s – may have run their course. The weakness of this argument lies in the difficulty in defining creative industries: there is no clearly defined “creative” function, creativity being a social process in which ideas are generated and confronted with an audience, which will ultimately decide what is creative or not. High-order business services, such as research and development (R&D), management consultancy, and design, are contributors to the creative process (which occurs across the whole economy) and in particular to cultural industries (which, like all other industries, require service inputs).

Thus KIBS continue to play a key role as knowledge intermediaries, information compilers, and transmitters of know-how across the economy (Strambach 2008). The tapering-off of their growth rates is partly attributable to the major structural changes having run their course, and partly to the simple arithmetic of growth rates (which necessarily slow down as the denominator gets larger, except in cases of exponential growth).

**KIBS and their strategic regional role**

Given the role played by KIBS in the economy, the connection between KIBS and regional economic development has been of concern to economic geographers. Three moments in the evolution of this concern will be described below.

**KIBS as footloose industries: an alternative to manufacturing**

In the 1980s, as widespread manufacturing decline was being experienced across most Western economies, certain regions were hit harder than others. In particular, resource-based regional economies (such as mining or smelting towns) and those relying on large factories (such as automobile or ship-building) lived through severe upheaval – some of these regions, such as the Nord-Pas-de-Calais in France, have never fully recovered. In this context, and bearing in mind the rapid growth in high-order services, this sector caught the attention of regional development specialists: the idea that producer services could stabilize, or indeed revive, lagging regional economies was explored (Marshall 1982; Bailly, Maillat, and Coffey 1987).

The argument is two-pronged. First, it rests upon the fact that KIBS were considered “footloose.” Unlike manufacturing, which requires machinery, large buildings, and other physical investment, KIBS were considered free to locate anywhere – their requirements being labor (available, almost by definition, in regions undergoing job losses), offices, and – at the time – a fax and phone line. Given the rapid KIBS growth, it was believed that they could be attracted
to lower-cost locations, or that, at the very least, local KIBS start-ups could be encouraged. However, as pointed out by analysts at the time, this idea is problematic since high-order services tend to locate in larger cities further up the urban hierarchy: this is partly as a consequence of Christallerian processes (high-order services require extensive markets, more readily accessible from central cities), and partly due to the specialized labor required (a preponderance of university educated personnel).

This approach, which may seem quaint in the light of work on innovation, cities, and knowledge exchange, is more relevant today than it was in the 1980s, since improvements in communication technologies are such that the advantages of city location may now be less decisive (Macpherson 2008). Already in 1996, Beyers and Lindhal (1996) documented “lone eagles and high-flyers” – that is, service professionals who, often in mid-career, decide to live, and run their consultancies, in remote communities (albeit communities located within an hour of a regional airport, allowing the consultants to travel easily). Currently, many regions in France are thriving because of consultants and other professionals who have chosen them as places of residence, working part of the time from the region, and traveling frequently towards the larger metropolitan areas where their jobs are officially located.

The second prong to the argument that KIBS can revive declining regional economies is that KIBS provide crucial inputs to manufacturers. Thus, notwithstanding the tendency of KIBS to locate towards the top of the urban hierarchy, Marshall (1982), for instance, argues that a concerted effort should be made to develop KIBS sectors in declining regions in order to enable local economies to revive. The idea is that start-ups and SMEs (small and medium-sized enterprises) – those best adapted to the new economic realities – can only emerge if they have access to KIBS, and that regional policymakers should ensure this access either by supporting the service sector directly or by providing the services through development agencies.

As we will see in the next section, this idea also retains some currency. The key difference between the ideas expressed in the 1980s about the strategic regional role of KIBS and those that are expressed today is their underlying vision of regional development and policy. In the 1980s regional development was understood to be, in large part, a consequence of the location decisions of manufacturers (and, marginally, of the service sectors): location theory, a series of ideas that explain the location of economic activity on the basis of cost minimization and access to markets, prevailed, and the role of KIBS was thought of within this framework. Furthermore, it was believed that central governments could and would implement regional development policies, such as incentives for particular sectors – KIBS, for example – to locate in particular regions.

KIBS as territorially embedded: milieu and innovation systems

The theorization of regional economies evolved in the 1980s and 1990s in tandem with the structural changes described above. At the same time there was a shift away from government intervention under the impetus of neoliberal ideas, epitomized by Mrs Thatcher (a UK prime minister) and Ronald Reagan (a US president). Two diametrically opposed, yet convergent (in terms of the way regional development is envisaged), lines of thought developed. On the one hand, a neoliberal reading of regional economies argued against policy intervention: regions and localities were understood to be in competition with one another, not part of an integrated system of interdependent territories but individuals in competitive struggle with
KIBS AND REGIONAL DEVELOPMENT

one another. On the other hand, regions and localities were seen by neo-Marxist geographers as the locus of resistance to globalizing mobile capital; and although the idea of resistance was watered down as the idea spread, the notion that regional economies are local systems which develop endogenously on the basis of local communities, institutions, and knowledge began to dominate the economic geography discourse. (From this perspective regions develop under the impetus of institutions, actors, networks, and resources (such as human capital) that exist within the region (i.e., that are endogenous to it). This contrasts with the view that prevailed in the 1960s and 1970s whereby regions developed under the impetus of economic actors choosing to locate in regions because of cost advantages and because of fluctuating world markets for their exported commodities (these factors are not within the region, they are exogenous to it).) Both the neoliberal and the critical perspectives emphasize that regional economies are identifiable entities which should (or need to) fend for themselves.

From these perspectives, KIBS are understood as strategic local assets, activities which enable innovation and economic growth (Cooke and Leydesdorff 2006). KIBS are seen as part of each region’s “technological infrastructure,” of its innovation system, and local economies are thought to be at a disadvantage if they do not house these services. Needless to say, not all KIBS are the same: of specific relevance to the local economy are technological KIBS, those concerned with engineering, design, and research.

KIBS’ role in local innovation systems is conceptualized in various ways. First, KIBS gather and enshrine specialized local knowledge, particularly tacit knowledge which is difficult to communicate in codified (i.e., written) form. Within a local innovation system KIBS firms act as knowledge and know-how relays between firms: as KIBS provide advice and interact with local companies they gain experience which makes them more valuable to the local system (provided, of course, that one accepts the idea that knowledge and know-how can be specific to a territorialized system). A second role played by KIBS is one of subcontracting: local firms may require specific service functions (such as accounting, R&D, product testing) for short time periods, and the local availability of these services facilitates their use. A final role is that of knowledge and information relay: Florida (1995) describes learning regions as those regions with the capacity to identify knowledge – whether local or nonlocal – that may be of relevance, and to translate it into terms useful for regional economic actors. KIBS, which constantly monitor new trends, new ways-of-doing, new regulations, and new markets – whether they emerge within the locality or beyond – are vectors of knowledge transfer within and towards the local economy.

If regional development is of concern, this view of KIBS in local economies can lead to two types of conclusion. The first is that only those regions with a well-developed KIBS sector are able to develop fully fledged endogenously driven local economies. Given that high-order services cluster in larger cities, this has led to the idea that large cities are the quintessence of well-functioning regional economies. Second is the idea that regions, whether urban or rural, should nurture local KIBS since they are key constituents of well-functioning local innovation systems.

KIBS, the Internet, and distance: back to basics

A recent approach to understanding the strategic role played by KIBS in regional economies
is, paradoxically, a return to some of the earliest work on the subject. Important work by Macpherson (2008) has empirically established that innovative firms do not rely on local service providers. He shows that firms in upstate New York identify potential KIBS providers by searching on the Internet for likely candidates (or by word of mouth), then travel in order to meet the providers, select them, and exchange hard-to-communicate information. They then rely on Internet and other modes of distant interaction for day-to-day communications. Shearmur and Doloreux (2015) show, furthermore, that the more strongly a KIBS subsector is connected with innovation the further the distance which separates the service provider from its client. Tether, Li, and Mina (2012), who examine two types of service provider (architects and engineering consultants), show that a Christallerian logic applies to architects whereas engineers seem to choose their location based on considerations largely unrelated to the urban hierarchy or to local markets (they are mainly global firms). In both cases local regional markets have little impact since these firms, especially the larger ones, tend to have a national or international client base. Finally, a pioneering study by Echeverri-Carroll and Brennan (1999) shows that innovative firms in southern US cities tend to search for strategic service providers up the urban hierarchy.

These studies have in common that the spatial economy is approached as a network of functionally interdependent regions. Endogenous regional development is not an applicable concept given these interdependencies; rather, it is understood that different functions will locate in different places according to costs, market extent, local availability of resources (including information and qualified workers; Polèse and Shearmur 2004), and communications technologies. It is not because particular functions are absent from a territory – KIBS, for instance – that the local economy will cease to perform. Quite the contrary: if a function is missing locally, economic actors will turn to the outside and identify providers elsewhere.

This does not preclude the fact that some KIBS will tend to function within and serve local markets: accounting and legal advice, for instance, will tend to be sourced locally – partly because they are required on a frequent basis. Other types of KIBS – particularly specialized innovation-related services required on an ad hoc or project-by-project basis – will tend to operate from accessible central places, with clients traveling occasionally to visit the consultant, or the consultant traveling occasionally to work with the client. The rapid rise and adoption of the Internet has made the identification of, and day-to-day communication with, KIBS far easier, thus undermining many ideas that focus too exclusively on local endogenous dynamics and territorialized knowledge systems.

Once it is recognized that KIBS need not be located within a region to play a strategic regional role, the way in which this role is conceptualized can evolve. Kautonen and Hyypiä (2010), for instance, view KIBS as knowledge mediators between different regional contexts. In their view, some KIBS can only play their role efficiently if they possess a geographically extensive client base. This enables them to transfer know-how between regions. KIBS too closely tied to a local system may in fact be detrimental to the local economy: KIBS which, from a metropolitan base, function both internationally (“up” the hierarchy or along “pipelines”) and regionally (“down” the urban hierarchy to a regional client base) transfer knowledge across borders and across scales.
Conclusion

The role of KIBS as a source of wealth creation, enabling firms to focus on their core activities whilst having access to specialized knowledge and know-how on demand, has grown rapidly since the economic upheavals of the 1970s. For a while it was felt that declining regions needed to attract KIBS either to replace lost manufacturing jobs, or to bolster failing local firms and fledgling start-ups. As the view of regional economies changed, as they were expected to become more entrepreneurial and competitive through endogenous processes, KIBS were increasingly seen as important elements of local knowledge and innovation systems: since knowledge and innovation are understood to be key drivers of economic development, the local presence of KIBS was seen as an enabling condition for regional growth.

Today, with almost ubiquitous availability of high-speed Internet and cellular coverage, there have been two important changes relative to the late 1990s. First, there is no reason to believe that geographically isolated firms cannot identify and interact with service providers from elsewhere. There remain geographical restrictions, since temporary face-to-face meetings are still important, but this places far looser constraints on the location of service users. Service providers, however, still tend to locate at the top of the urban hierarchy since accessibility – both to global markets and to national markets – is easier from these hubs. Furthermore, qualified labor is more plentiful in these places.

The second change relates, notwithstanding the previous sentences, to the fact that some service providers can also now locate away from major metropolitan areas. These relocations are unlikely to alter the basic hierarchical distribution of service providers, but can have sizeable impacts on receiving communities. Globalized service firms may not require location at the heart of a metropolitan area, and may consider location in places within easy reach – about an hour or so – of one: as Tether, Li, and Mina (2012) describe, these firms rely on internal economies of scale rather than on external ones. Accessibility remains important, but does not need to be maximized. Smaller consultants may likewise locate in amenity-rich areas, provided they are relatively accessible. Finally, KIBS workers increasingly perform part of their work from rural or isolated locations, traveling to their KIBS employer (or clients) on a regular but intermittent basis.

To conclude, it is important to emphasize that little space has been devoted to considering the way in which different types of KIBS subsector locate and/or interact with their clients (Tether, Li, and Mina 2012). Likewise, the increasing difficulty in distinguishing between KIBS and high-tech manufacturing has not been discussed (Daniels 2011); however, to the extent that these are beginning to resemble each other, the location and role of high-tech manufacturing (or of the knowledge-related functions within global value chains) may begin to resemble that described here for KIBS. Finally, given the focus upon regions, the entry does not deal with the distribution of KIBS globally.

SEE ALSO: Central place theory; Deindustrialization; Demand and supply for producer services; Economic development zones; Economic geography; Industrial districts; Industrial location theory; Information and communications technology; Innovation and regional development; Knowledge-based economy; Local development; Producer services and economic development; Regional development policies; Uneven regional development; Vertical integration
References


Throughout much of the twentieth century, geographers drew upon neoclassical economic theory to understand firms’ locational dynamics. At least four dominant approaches can be identified. First, a generation (if not more) of geographers used the work of Alfred Weber (1929/1909), who developed a “least cost” model which explained the locational decisions of independent, single-plant firms in terms of the relative weight of their transportation and labor costs, and what he called “agglomeration forces.” Second, an approach drawing upon Hotelling (1929), and exemplified by the famous example of ice-cream sellers clustering on a beach, focused upon the geographical behavior of businesses whose location is dependent upon that of their competitors. Third, the work of August Lösch (1954) was seen as a way to make the leap from understanding why individual firms behave the way they do to understanding what this behavior means for the creation of entire economic landscapes. Finally, the growth of behavioral geography in the 1960s led some to seek to understand firm dynamics in terms of the idiosyncrasies of various corporate decision-makers.

With the influx of Marxism (see Marxist geography) into geography in the 1970s, however, these approaches began to be critiqued. For one, they were seen as largely disconnected from consideration of the broader political economy of capitalist accumulation and the drive to extract surplus value from labor. For another, many were paradoxically aspatial in their approaches – they largely assumed that the economic landscape is simply an inert stage upon which economic actors interact, implying that the economic geography of capitalism is simply a reflection of the economic relations between firms. They also were derided for their positivist philosophy. Finally, by principally telling the story of the making of the economic geography of capitalism from the perspective of managers, their explanations largely ignored the role of workers. As it developed, this Marxist critique provided a way to link the behavior of firms into broader questions of the nature of capitalist accumulation, the extraction of surplus labor, the periodic crises that the capitalist mode of production experiences, and the production of economic landscapes. However, Marxist geographers initially conceptualized workers largely in terms of their being “variable capital” or a “pseudo-commodity,” rather than as embodied sociospatial actors. This is perhaps not surprising, as Marx himself was largely interested in understanding the nature of capital accumulation and how competition affects capitalists’ behavior. Nevertheless, since the 1990s this lacuna has led a growing number of geographers to focus upon labor as a geographical actor. Adapting Marx’s famous aphorism that “people make their own history but not under the conditions of their own choosing,” these geographers have sought to show how workers make their own geographies but not under the conditions of their own choosing. Framed as a shift from studying the “geography of labor” to studying
“labor geography”, the focus of such geographers’ writings has been to look at how workers’ spatial embeddedness affects their economic and political behavior, and how that behavior in turn affects how the geography of capitalism is made. Below is detailed some of the research conducted by labor geographers in three areas: how workers have shaped individual firms’ structures and investment patterns through their activities; how workers have shaped networks of firms; and how workers have convinced various elements of the state to implement policies which have affected the geographical structure of firms and corporations.

Workers as shapers of corporate investment strategies

Geographical research has explored how workers’ activities shape the spatial investment strategies of individual corporations in three principal ways. The first approach has shown how workers have influenced corporations’ organizational and geographical structures reactively, as workers make themselves more or less attractive to capital, which then responds to their actions. In such studies the economic landscape’s making is viewed through the eyes of corporate decision-makers who decide to invest (or not) in particular places because of the actions or inactions of workers. It is these corporate decision-makers, then, who are seen to be thinking in a geographically deliberative manner as they use the uneven spatial distribution of labor’s characteristics (workers’ wage rates, levels of unionization, gender, and so forth) to their own advantage. The focus is upon capital as actor choosing between workers, rather than directly upon workers’ own agency, even though such agency – workers’ militancy and/or quiescence – shapes what corporations choose to do. For example, Peet (1983) has shown how many firms relocated in the post-World War II era from regions with high wages, high unionization rates, and high regulatory burdens to those with little or no unionization, low wages, and more “business friendly” state and local governments as part of the so-called Snowbelt to Sunbelt migration of US capital. In related fashion, other analyses have looked at how workers have modified their behavior so as to make themselves more attractive to corporate decision-makers. Hudson and Sadler (1986), for instance, explored how British steel-workers toned down their rhetoric when the government-owned British Steel Corporation was deciding which steel mills to shut down as it rationalized the industry in the 1980s. Successful labor organizing, then, is one way in which workers’ activities shape corporate investment strategies – their unionism is something which mobile capital seeks to avoid. Others, though, have suggested that corporations have a more nuanced relationship to labor organizing as it affects their locational decision-making. Hence, Page (1998) has shown how meatpacking firms relocating from Chicago after World War II did not seek to avoid unions per se. Rather, they avoided Midwestern communities with militant unions and chose to relocate to communities with more passive ones. In many ways, such studies are similar to the “geography of labor” approach articulated by neoclassical and some Marxist scholars, although they see workers as embodied economic actors rather than simply as “factors of production” or “variable capital.”

The second way in which workers have shaped corporations’ organizational and geographical structures is proactively, through consciously engaging in behaviors that force particular corporations to reorganize their own structures in response. The difference with the situations described above is that here the making
of the economic geography of capitalism is seen through the eyes of workers rather than corporate decision-makers. These are what have come to be called “labor geography” approaches. For instance, with the introduction of containerization into the maritime cargo-handling industry in the 1950s, US dockworkers began to feel the effects of job losses on the waterfront. Dockworkers in places like New York, Boston, and Philadelphia also feared that the growth of the interstate highway system and the ease of transporting cargo due to the new container technology would encourage shipping companies to unload cargoes in the less unionized southern ports, where labor costs were lower, and then use the rail or road system to transport their loads to the cities of the Northeast. Herod (1997a) showed how dockers, in light of such fears, began a spatially conscious campaign to transform the industry’s system of bargaining from one in which contracts were signed on a port-by-port basis to one in which a national system of collective bargaining – one that would lead to national wage rates – was used. Dockworkers’ actions, then, forced shipping companies to restructure their operations and to create new national employer associations. The dockers also negotiated a series of work rules dictating that certain types of cargo-handling work had to be done in ports rather than at warehouses located further inland, rules that shaped which warehouses the shipping companies could utilize and thus the geography of warehouse work in the industry. Looking more globally, Wills (2002) has shown how unions, through signing international framework agreements, have forced firms in some economic sectors – such as the hotel industry – to change how they conduct business. Sadler (2004), meanwhile, has examined how unions and environmental groups collaborated to introduce resolutions at shareholders’ meetings of the global mining giant Rio Tinto, pushing the corporation to change its business practices with respect to the rights of indigenous peoples, the environmental consequences of mining operations, and respect for workers’ rights.

The third way in which workers and their organizations have been seen to shape corporations’ organizational structures is through practices which, though not directly aimed at such corporations, have nevertheless proactively opened up paths for investment that they have then sought to exploit – unions, in other words, have served as shock troops allowing corporations to come in behind them to invest in certain regions. For example, as part of its own efforts (separate from the interests of US corporations) to undermine communism in Latin America and the Caribbean, during the twentieth century the American Federation of Labor–Congress of Industrial Organizations (AFL-CIO) encouraged US corporations to expand their operations into the region (Herod 1997b). Its goal was to facilitate the spread of US capital into Latin America and the Caribbean, thereby more tightly binding the region’s countries to the United States and inculcating “democratic” (i.e., noncommunist), pro-American political organizations and economic practices. The AFL-CIO hoped that this would not only help the United States win the Cold War but also provide material advantages to US workers – dominating the economies of the region would, it was believed, translate into domestic jobs for US workers. Through working with pro-US unions, subverting communist ones, and participating in several coups over the years (as in Guatemala in 1954 and Brazil in 1964), the AFL-CIO paved the way for US corporations to establish many subsidiaries in the Western Hemisphere where they could work with compliant unions who frequently restrained worker militancy in their branch plants. The AFL-CIO engaged in similar policies during the
LABOR GEOGRAPHIES AND THE CORPORATION

1960s and 1970s in Southeast Asia, for similar reasons. Thus, its Asian-American Free Labor Institute worked closely with leaders of the Trade Union Congress of the Philippines to create a more favorable investment climate for American business during dictator Ferdinand Marcos’s rule. Similarly, the AFL-CIO, together with a number of other international labor organizations, such as the International Metalworkers’ Federation, supported the spread of Western corporations into Eastern Europe after the collapse of communism there, seeking to jump-start these regional economies and help establish democratic institutions. At a more local level, unions also have often been part of local growth coalitions seeking to encourage capital investment into particular communities (Humphrey, Erickson, and Ottensmeyer 1989).

Labor and global production networks

Whereas this first generation of labor geography research tended to focus upon how workers shape the actions of individual corporations, the rise of “globalization talk” in recent years has led many geographers to research the structure of global commodity/value chains (GCCs/GVCs)/global production networks (GPNs). Essentially, GPNs are networks of geographically embedded firms through which various commodities and parts thereof pass, as they are manufactured and brought to market. These networks’ social and spatial organization generally is viewed as responding ultimately to the demands of the corporation sitting at the GPN’s apex. However, labor and its institutions have often been seen to be rather passive entities in GPN analysis. Capital is pictured as shaping a GPN’s sociospatial structure through determining whence raw materials will be sourced and how and where various components will be brought together as a commodity is assembled and moves through the network, as well as by playing off workers in different locations against one another, given their militancy, wage rates, productivity, and so forth. Workers, especially those toiling for third- and fourth-tier subcontractors at the supposed bottom of the network, though, have frequently been seen simply as social beings who are subject to the locational whims of corporate decision-makers and have little impact on shaping GPNs (for more on this, see Global production networks; see also Global commodity/value chains).

There has been growing criticism, however, that treating labor as simply one factor amongst many underplays workers’ impacts upon GPNs. As Coe, Dicken, and Hess (2008) argue, this failure to see workers as active geographical agents within GPN analysis has resulted from the tendency to view the firm as a black box. However, given that labor process dynamics strongly influence work organization and conditions within any one firm in a GPN, together with the relationships between firms, workers have the capacity to exert great influence upon the geography of activities within a network and thus the network’s overall form. As a result, labor geographers have argued for placing labor at the center of GPN analysis, focusing upon how workers’ spatial embeddedness shapes their economic and political behavior and thereby the relationships between the firms making up a GPN. For example, whereas corporations like Walmart are often imagined to be unchallengeable, their reliance upon tightly coordinated logistics within their distribution systems is actually a significant weakness that workers can exploit to force improvements in wages and working conditions. Even workers in small firms, often perceived in the GPN literature to be quite weak, can exert tremendous pressure upon the corporations atop the networks. This potential power was clearly illustrated in a dispute at two
small components plants in 1998, which brought virtually all of the North American operations of General Motors (GM) to a halt because of GM’s use of just-in-time production systems (see Just-in-time production system), allowing the dispute to ripple quickly through its production network. Likewise, although the result of a natural disaster rather than a strike, the March 2011 Japanese tsunami closed several component manufacturing plants operated by suppliers who are part of Toyota’s and Honda’s keiretsu, which led both auto assemblers to halt production at several of their US and European assembly plants. The point, then, is that the closure of even fairly small second- or even third- or fourth-tier suppliers – whether due to natural causes or strike action – can force firms like General Motors, Toyota, or Honda, which are assumed to be presiding over the GPN, to have to close their facilities.

Whilst there has been much interest in how commodities are put together within GPNs, until recently there was little interest in what happens after these commodities have been used and disposed of. In the past few years, however, a growing number of scholars have advocated for examining the “after life” of commodities, as part of what has come to be known as “ongoingness research.” In this context, Herod et al. (2014) developed the concept of global destruction networks (GDNs) as, in some ways, “Other” to GPNs. Whereas GPNs link different communities and independent firms or parts of corporations together in the process of assembling commodities, GDNs link together places and firms involved in the process of commodity disassembly and the reclamation of their constituent elements. As with GPNs, labor’s power to shape GDNs’ structure has been clearly demonstrated. For example, the processing of electronic waste (“e-waste”) in different parts of the globe is fundamentally shaped by the labor practices involved in its disassembly. In the Global North, e-waste dismantling is typically capital-intensive, using large machines to break up old electronics. In the Global South, on the other hand, it is generally much more labor-intensive. As a result, the kinds of elements reclaimed from disassembled commodities are quite different. In the Global North there is a higher overall recovery rate for waste elements than in the Global South but the higher labor costs mean that it is less likely that parts thus recovered (e.g., computer chips, circuit boards, hard drives) will be repaired and directly reused in new products. Instead, these are usually melted down for the metals and plastics they contain. By contrast, in the Global South it is more common that components recovered from, say, brand-name computers will be resold as parts, either to repair other brand-name computers or from which to construct non-brand-name, locally assembled computers that are essentially built from scratch for sale nearby in the backstreets of a place like Bangalore.

Labor’s influence on the structure of a GDN is also evident in the global shipbreaking industry. Although the cutting apart of ships at the ends of their working lives largely takes place in the Global South, much of the industry itself is coordinated from global shipping capitals like London, Rotterdam, Singapore, and Dubai. The process works as follows. When a shipping company seeks to scrap a ship, brokers in these cities inform breaking yards in countries like India, Bangladesh, and China, to which the ships are sailed, to begin the process of retrieving metals and other components. In China, ship-breaking generally is undertaken in docks using large-scale machinery and relatively few workers. In countries like India and Bangladesh, on the other hand, ships are usually beached in shallow waters and then broken up by gangs of workers who often use little more than sledgehammers and axes. This difference in how the ships are
broken up primarily reflects variations in wages, which are much lower in Bangladesh and India, and in Chinese workers’ greater unwillingness to tolerate the poor working conditions faced by Indian and Bangladeshi workers. In short, workers’ willingness to do certain types of work fundamentally shapes where and how shipbreaking is done.

Labor and the state

A further way in which labor can shape corporations’ geographical organization is through convincing various governments to enact regulations that limit corporations’ abilities to structure themselves in particular ways and/or to invest in certain places. For example, Rutherford (2013) has examined the Canadian Health Employees’ Union’s legal victory when the Canadian Supreme Court recognized collective bargaining as a fundamental right in the country’s national Charter of Rights and Freedoms. As a result, corporations operating in Canada must now respect certain labor rights, which undoubtedly will shape how they operate. Likewise, Clark (1989) has shown how US unions’ successful exhortations to the National Labor Relations Board (the federal entity charged with adjudicating disputes between employers and unionized workers) have limited corporations’ freedom to do as they please, with significant implications for how the economic geography of the United States is made. For their part, in studying day labor markets Peck and Theodore (2012) have detailed how various local governments have passed ordinances regularizing worker centers. Such centers have arisen because of the growing use of day laborers and other contingent workers not only by small, local firms but also by large, transnational corporations (Peck and Theodore 2006). These ordinances restrict the abilities of firms and individuals to force day laborers into a “race to the bottom” through underbidding each other for work. As Peck and Theodore put it, worker centers represent an effort to counter neoliberal labor-market deregulation from the bottom up.

Other scholars have shown how workers and their organizations have managed to get various arms of the state to pressure firms directly concerning their operations and where they locate. Herod (1991) detailed how the International Ladies Garment Workers’ Union in New York City successfully lobbied the city government to create an area within its Manhattan garment district in which conversions of manufacturing lofts would be severely restricted through new zoning regulations. The union’s goal was to protect space for garment manufacturers at a time when market pressures were encouraging many building owners to rent their structures to corporations as office space. There are myriad other examples of unions agitating for laws that limit corporations’ abilities to structure their operations in the name of profitability. For example, the United Food and Commercial Workers and the Service Employees’ International Union have worked with local community groups to persuade municipal governments across the United States to ban Walmart stores. Numerous unions also have pushed the federal government to enact various national rules and regulations overriding state laws, undermining advantages that companies might otherwise gain by relocating from highly regulated to less-regulated US states. In the 1930s this included national legislation limiting child labor and creating a national minimum wage and various union protections; the 1960s and 1970s saw legislation preventing employment discrimination on the grounds of race, sex, religion, and national origin; and in the 1980s and 1990s it included federal laws providing for family leave and requiring advanced notice of mass layoffs.
Such legislative agitation can have unintended impacts on corporations’ structures, however. For example, after the United Auto Workers (UAW) helped pass protectionist legislation in the 1970s to keep Japanese imports out of the United States as a way to save American autoworkers’ jobs, Japanese corporations simply established assembly plants in the United States. Whereas the first such plant (Honda’s St Marys facility) was built in the fairly union-dense state of Ohio, most of the subsequent “transplants” have been built in southern states, where union density is much lower. This has undermined the UAW, both because these plants are hard to organize (they are mostly in right-to-work states, where state laws do not require all workers in a unionized workplace to belong to the union) and because the growth of a nonunion labor force in southern states has allowed the companies to press for wage and other concessions in their unionized auto plants in the American manufacturing belt.

Conclusion: beyond unionized labor

This brief entry has detailed some of the research demonstrating how workers have shaped corporations’ structure, behavior, and spatial organization. To date, much of this work within labor geography has focused upon unions. There are some good reasons for this. For instance, unions often have good documentary archives available to researchers and readily identifiable leaders whom they can interview. Given that labor geographers’ initial goal was to make a theoretical claim that scholars seeking to understand the making of the economic geography of capitalism cannot just focus upon the spatial actions of capital but also must look at those of workers, and given that workers tend to be more powerful when they behave collectively, it seemed reasonable to examine unions as collectivities of workers. However, with labor geography now an established field of study, what might be called labor geography 2.0 has begun to emerge. The geographers associated with this development – many also central in developing labor geography 1.0 – have suggested that (at least) two major developments are necessary. First, geographers need to explore the lives of more than just unionized, Global North workers if they are to gain a more complete picture of how workers shape corporations’ and firms’ structures and how they play active roles in making the economic landscapes of capitalism. Consequently, some have begun to research the activities of nonunion and Global South workers, as well as of those trapped in precarious (temporary and part-time) rather than full-time work (see Precarious work). Second, they have suggested that labor geographers need to develop a much more nuanced and situated understanding of agency, one which recognizes how sociospatial context differentially shapes workers’ abilities to exercise power in making the economic landscape (Coe and Jordhus-Lier 2011; Rutherford 2010). Through such efforts they hope to more fully develop Marx’s insight that people make their own histories (and geographies!) but not under the conditions of their own choosing.

SEE ALSO: Class; Economic geography; Industrial location theory; Radical geography; State, the; Trade unions; Unionism, community

References


Coe, Neil, Peter Dicken, and Martin Hess. 2008. “Global Production Networks: Realizing the


Labor geography

Peter Brogan
Steven Tufts
York University, Canada

Labor, human efforts to produce goods and services, is a key subject of geographic inquiry. The study of labor has, however, evolved theoretically from determinist treatments as a mere “factor of production” to a more contested agent in the formation of economic landscapes. Following a brief history of geography’s engagement with labor, we discuss four key themes in contemporary “labor geography”: agency, scale, class, and the state. We conclude with some recent theoretical developments and the future trajectory of labor geography as a field dedicated to analyzing the spatiality of work, workers, and workplaces.

Labor and geography

Modern geographic inquiry has studied what people do and how people organize themselves for social and economic reproduction over space. In the dark moments of environmental determinism, Ellsworth Huntingdon linked the industriousness of different groups to the specific climate conditions. The regional tradition within geographical research incorporated labor into descriptions of how patterns of production and reproduction develop across both time and space. Indeed, the different skills accumulated in local labor markets remain central to research on regional economic development and global economic competitiveness.

Labor, as a “factor of production” also played an important role in the quantitative revolution. Early location theory, inspired by Alfred Weber, conceptualized labor as a “factor of production” that, if its costs are lowered, “deviates” the location of a plant from its minimum transportation cost location. Industries that shift to lower labor cost locations are deemed to have a “labor orientation.” Economic geography’s quantitative revolution, similarly, was concerned with modeling the location of economic activity and included labor as a factor on par with transportation costs, resource availability, and market access.

It was the reduction of workers to just another factor of production, however, that was challenged by radical Marxist geographers. The emergence of Marxist geography in the 1970s and 1980s centered the conflict between labor and capital as key to the formation and transformation of economic landscapes. At the same time, feminist geographers also focused on questions of labor as fundamental to gender relations and economic relations. Here, not only were inequities within the gendered division of labor noted and investigated, but the importance of unpaid labor in the reproduction of capitalism was foregrounded (see McDowell 1993).

In the mid-1990s, critiques emerged against traditional Marxist geography’s theorization of economic landscapes that centered capital in analysis and overlooked the actions of workers within capitalist economies. Andrew Herod (1997) called for an economic geography that did not privilege capital and its agents over labor.
Labor Geography

in shaping economic landscapes. His call for geographers to assert labor as a more active agent in economic landscapes inspired a decade of research that shifted focus from the “geography of labor” toward a “labor geography” (see Lier 2007; Castree 2010).

Recently, Herod’s initial call has been critiqued for too easily dismissing economic geography as overly “capital-centric.” Storper and Walker (1989) in their now classic (yet difficult) schematic of employment and the construction of labor markets did center the employment relationship. Similarly, Peck (2013) has argued that the groundbreaking work of Doreen Massey (1995/1984) could easily be categorized as “labor geography”:

Intellectually and politically, Doreen Massey was a labour geographer long before the label existed. The path-altering analysis in Spatial Divisions of Labour not only marked but also helped realise a paradigm shift in economic geography – from positivism and location theory to critical realism and political economy – after which there have been many turns but very few returns. In a quite distinctive way, the not-so-little yellow book that all of us were carrying at the time put labour on the map, not just as an analytical category, but also as a social agent (Peck 2013, 113).

The counter criticisms against the “labor geography” project may overstate Herod’s original intentions in which he clearly indicates that his “capital is not all” premise should not be read as “labor is all” (Herod 1997, 3). In the 1980s, Marxist economic geographers in the United States and those directly involved in the locality debates in the United Kingdom, were undoubtedly challenged to celebrate the agency of labor in the midst of Thatcherist and Reaganist assaults on workers. In contrast, Herod’s minor corrective in the 1990s emerged at a moment where union decline had, momentarily at least, stabilized in the United States and the anti-globalization movement and other labor-friendly formations appeared to be gaining some momentum. Herod’s intervention also gave geographers a way to enter discussions of labor union renewal that emerged in Anglo-American countries in the 1990s.

At the same time, there were geographers in the 1990s that looked at issues of work and workplaces from feminist perspectives that were overlooked in the rush to establish a new “labor geography.” One example (among others) is Susan Hanson and Geraldine Pratt’s Gender, Work and Space (1995), which situated women’s power and place in the economy also without the “labor geography” label. Similarly, the work of Linda McDowell (1997) from the same period on the spatial dynamics of gender relations and work in capitalist landscapes must also be acknowledged as casting a feminist lens on the activities of labor and the geographies within which it is embedded.

Four themes

In the past decade there have been a number of rich assessments of labor geography as an emergent subdiscipline (Lier 2007; Castree 2010; Coe 2013). Four key central themes can be identified: agency, scale, class, and the role of the state. In some areas, advances building on Herod’s initial call have been significant, yet debate does persist in others.

Labor and agency

From its formal introduction in the 1990s, labor geography has argued that workers are active producers of the geographies of contemporary capitalism. Whether it is acknowledged or understood by workers or their organizations,
Labor geography is also a fundamental part of the structures workers confront as they change the terms and conditions of their employment and their lives more broadly. Social actors are geographically embedded, quite often in contradictory ways, and this embeddedness in turn shapes the possibilities for their social action. The focus on labor agency has led to a rich and wide-ranging register of case-study research that examines the question of what workers do besides work as they navigate the world. This scholarship has illustrated that neither workers nor employers are passive elements in the geographies of social relations, for both actively construct the world around them, although not always exactly in a way they would prefer. The control of space and place is thus crucial to the way that jobs are exported, created, lost, and fought over. This work demonstrates how thinking geographically enriches our understanding of labor’s agency and activism.

Unions and collective action have been the key mechanisms for workers to make changes in the economic landscape and both have been central objects of study in labor geography. Jane Wills and others have illuminated some of the ways in which various collective practices, cultures, and ideas are constructed in particular places and how these traditions are translated across space. Yet within early research on unions most labor geographers failed to differentiate between workers as individuals and workers as collective actors – working predominantly through trade unions. As such, while a major empirical focus in labor geography has been on the relevance of unions and how their changing strategies and tactics are being deployed in different places and across scales, the micropolitics of power and geographies of struggle within unions has not often been emphasized. While one of the identified limits of labor geography has been its institutional focus on trade unions, researchers have begun to expand the horizons of labor geography by conducting research into low-wage/no-wage geographies and alternative workers’ organizations as well as given greater focus to labor geographies outside of the advanced capitalist economies and the flows of migrant workers between and within the Global North and Global South.

The focus on labor’s agency has yielded important insights into how workers and their organizations are transforming political and economic geographies, so that we can see that how economic lives are shaped is not solely at the discretion of capital. While capital’s typically much greater geographic mobility can be a source of power, many political, social, and practical challenges often get in the way of allowing employers to move their operations anywhere they might choose. Yet as a number of studies in labor geography have demonstrated, the “rootedness” of labor may at times, and under particular circumstances, become a source of power when workers and their families, friends, and neighbors are able to create distinctive local communities, cultures, and organizations or use their location in global supply chains as a point of leverage against their employers. Capital and labor create different verities of place consciousness, both progressive and reactionary. For example, while capital often seeks temporary spaces for profitable production, for workers these spaces are also places in which to live, and thus places within which they have considerable individual and collective cultural investment. This has led some to conclude that whereas space is the domain of capital, places are the meaningful situations established by labor.

While this binary formulation lends perhaps too much power to capital as the dominant commander of space, it is a useful starting point for understanding that place is an important site and basis for workers to forge bonds of solidarity.
Indeed, workers will often utilize what David Featherstone calls “maps of grievances” to describe the place-based “dynamic practices through which political activity makes sense of and brings into contestation relations of power” (Featherstone 2012, 18). Yet, place-based identities or consciousness may also lead to an incredibly reactionary and dangerous politics, not simply through the forging of alliances with employers that obscure class conflict, but in producing, for example, violent xenophobic responses to outsiders. Workers are moving through systems of global migration in increasingly large numbers and their exit from and entry into different labor markets can be not only a form of resistance and adaptation, but also a significant source of conflict.

Critical commentary on the field has argued that the agency labor geographers lend to workers is often undertheorized and premature. Don Mitchell, for example, has compellingly argued that labor geography research needs to be more attuned to the materiality of the world in which labor exists, rather than overly focused on labor’s activities and power to change its geographies. That is, as we seek to see how workers create economic spaces and landscapes we must also closely examine those spaces and landscapes that they have not made, at least in any basic sense, but in which they find themselves and must live – those landscapes that are through struggles and the exercise of power, produced not for them but for others, those landscape that make “a new kind of community all but impossible” (Mitchell 2011, 567). Even harsher is Raju Das’s critique that “agency has often been used as a quasi-empirical category: a tool to describe how labor is making a difference to the spatial organization of capitalism, here and there. Agency in opposition to capital’s own existence, agency in collaboration with capital, and agency involved in gaining concessions, without challenging capitalist class relations, are all problematically put together” (Das 2012, 21). Put differently, simply focusing on the activities of workers within capitalism as a way to highlight the significance of labor’s agency does not get us very far, analytically or politically, if we operate with a conception of agency that captures any and all things that workers might do.

Coe and Lier (2011, 14) largely dismiss these concerns and continue to focus on “developing more precise concepts for describing the politics of work.” Their goal is to theorize agency more rigorously, and in so doing they turn to Cindi Katz’s typology of agency – resilience (adaptive, getting by), reworking (shifting distribution systems), and resistance (changing the forces of production, balance of power). They understand labor agency to be both embedded and constrained (Coe 2013) while others conceptualize any meaningful labor agency as a distinct force “autonomous” and independent from capital.

Despite these more nuanced conceptualizations of agency, a fundamental question remains of exactly “whose” agency is being spoken about? Recent scholarship conceives agency and consciousness (or subjectively more broadly) as being shaped by a whole range of contingent factors, structural locations, identities, and ideology. For instance, Bergene, Endresen, and Knutsen (2010) have suggested that all of these elements may come together and form a “chaotic consciousness” among workers, which can effectively be navigated and transcended by unions and other political organizations that workers construct to shape their economic landscape. Castree likewise notes that researchers need to pay extraordinarily close attention to the “geographies of employment and labor struggles not in and of themselves but as windows onto the wider question of how people live and seek to live” (Castree 2010, 468).
Production of scale

The second programmatic area within labor geography is the production of scale, the level at which resistance and negotiations with capital take place. For many traditional Marxist geographers action always begins locally, but it must quickly “scale up” to a level compatible with multinational capital in order to avoid the limitations of localized militant particularism. Geographers have focused on how workers organize at the microscale of the workplace, from local coalitions with community groups to tapping into national organizations and build transnational labor alliances (Waterman and Wills 2001). Some debate remains among scholars as to which scale (e.g., the workplace, local, national, global) is the most practical means of achieving greater power for workers and the ways in which different scales of organization both conflict with and complement others. Some emphasize that workers organize locally and globally by forming relationships with workers in different places and other classes (e.g., local developers) depending on their time and place specific needs. Others have emphasized the limits to organizing at international scales as local labor inevitably confronts a “geographical dilemma” as workers compete in a global economy for investment and jobs in their communities. Yet, there seems to be somewhat of a consensus emerging among labor geographers that organizing is often necessarily multiscalar as workers mobilize locally and globally simultaneously to influence their employers and public policy to improve their conditions.

Thus, one of labor geography’s main strengths is that it moved away from framing questions of scale within a rigid binary: either privileging organizing labor at a global scale to match the extent of capital or stressing the need for labor to dig into more local, community-based organizing. Labor geographers have contributed to a new labor internationalism/transnationalism with an understanding that organizing is inevitably multiscalar, both in the past and present. Further, labor geographers have recognized that targeting key nodes of the capitalist space-economy (e.g., global cities or logistics hubs in global production networks) is vital to labor power and that these key nodes shift with time and place. Even with more “local” organizing, geographers have called for an expansive research agenda that pushes our analytical focus beyond workplaces and trade unions arguing that unions need to build reciprocal alliances with communities as a key weapon in the struggle to reverse labor decline.

The centrality of class

The extent to which labor geography centers and advances class analysis is a question that continues to be raised. Coe and Lier (2011) contend that, aside from Wills (2008), labor geography has yet to contribute much to an analysis of class. As Wills (2008, 33) notes “when we try to map the working class, e.g. according to workers’ relation to the production of surplus value, we discern a landscape of highly uneven political possibilities.” Too narrow a focus on what constitutes “working-class” actors limits the possibilities and evidence of resistance. It is increasingly recognized that the working class is greatly fragmented and workers are unevenly embedded in economic systems, political regimes, and social networks. Numerous scholars have called for analysis of workers and their production of economic space and resistance in workplaces as constituted by gendered, racialized, and global imperialist power relations within and among classes. Labor obviously must be broadened as a category beyond unionized male industrial workers and unionized professional female public sector workers to include a diverse and expansive range of precarious workers.
LABOR GEOGRAPHY

It is not that the actions of declining male, blue-collar industrial workers are irrelevant, but rather that the labor geographies of fast food workers, unpaid interns, and the unemployed, for example, are just as worthy of attention by researchers. A labor geography that embraces the complex diversity of the working class while simultaneously respecting the unity among those who are required to sell their labor power to reproduce themselves has yet to be attained.

But there are more expansive understandings of work and workers inspired by feminist, antiracist, and anticolonial scholarship, such as Ruth Wilson Gilmore’s *The Golden Gulag* (2007), a rich antiracist feminist economic geography of the prison industrial complex and its relations to state restructuring and racialized labor in the United States. Here, we see how uneven labor geographies are produced by racialized bodies imprisoned in the prisons or trapped in marginalized urban labor markets. A narrow conception of labor geography that does not recognize these processes and fails to incorporate such theoretical and empirical insights in its research agenda will be limited.

There remain, however, concerns by those who fear that “de-centering” work and class in labor geography through a shift to intersectionality and extra-workplace activity may weaken analyses. Das (2012, 23), for example, argues that “reformist” labor geography conflates labor (i.e., paid workers and their institutions) with the working class (a much broader group) and an already anti-essentialist labor geography simply lacks a theory that encompasses the “unity that defines class.” Class is historically divided by sex, age, ethnicity, physical ability, and, of course, *geography*. Labor geographers will continue to study how workers can produce and remake economic geographies in ways that promote the necessary unity, or not, whether this is done with an increasing focus on class or other relations that constitute people’s identities will continue to be debated.

The role of the state

There is also the question of the role of the state within labor geography. Some scholars contend that labor geographers have demoted the state relative to other institutions and relationships. Again, this claim holds if a narrow definition of labor geography is used. Geographers have, however, paid significant attention to the role of the state and its links to work and economic development. Feminist studies of how access to affordable daycare and public transportation shape labor markets and urban economies had a huge influence on understanding the geographies of work.

Institutional relations have perhaps received less attention than the role of labor law in structuring worker action, but the two are interdependent. For example, Jordhus-Lier (2012) has argued that public sector workers have also been relatively neglected as an empirical focus, and one should add, theoretical subject of analysis. In many advanced capitalist economies experiencing significant declines in the private sector union density, a feminized public sector workforce has served as the reserve of union power and sometimes the only remaining institutional opposition to neoliberalism. And in the United States alone, there are over a million low-wage contractual workers employed by municipal and state governments, whose degraded and precarious employment conditions – and the struggles to transform the relations that underpin these relations and arise as a result of them beyond the workplace – are in desperate need of a critical geographical investigation that relates the experiences of these workers and their activities to the institutional transformation of both state and labor market institutions.
Although this lack of focused attention to the state may not be surprising given that the study of labor action within the context of neoliberal attacks has perhaps rendered the state as a given antagonist – and therefore no longer of any strategic value to workers, state power, and the changing contours of government employment relations in particular, will remain a pivotal research focus for labor geographers to explore as the field continues to evolve. At the same time, extra-state strategies of labor as its organizations, like unions, continue to attempt to out-maneuver neoliberal governments and private sector employers have and should continue to receive attention. And while labor geographers have begun to explore the variety of ways workers engage with local, regional, and national governments in order to shape economic development strategies, less has been done to develop a labor geography influenced theory of the state. In an era of intensified austerity, however, further attention to the state will become paramount.

**New worlds of possibility**

In many jurisdictions today, labor faces multiple challenges of declining institutional power, geographically fragmented workplaces, uneven capacities, and perhaps most significantly, the addition of millions of workers to the reserve army of labor each year. Presently, states continue to embark on austerity programs (although unevenly) across even the most regulated labor markets. How can a labor geography premised on the ability of workers to exercise agency survive eras of increasing austerity and deepening precariousness? What kinds of agency can labor exercise if researchers do not expand the scope of inquiry and analysis into questions of political strategy and new organizational forms for resistance? There is a continuing need to develop strategies and research that allow us to both understand and organize these new geographies of labor. A good deal of scholarship in labor geography has sought to explore the limits to traditional union organizing but the formal labor movement is not the only form of regulation and often fails workers in the informal economy where pay is low and conditions poor. If labor geography itself is to survive the current conjuncture of austerity, precarity, and environmental crisis, several theoretical and political issues will need to be addressed in the future, some of which have been discussed. Fortunately, there are already signs of labor geography rising to the challenge to confront these.

First, there is the lingering question on the centrality of class and the continued fragmentation of workers. For many, reading class as an essentialist concept – by which we mean a conceptualization that sees class as the principal sociospatial process shaping labor and its geographies – will remain central to advancing labor geography within an historical materialist framework. For others, the importance of addressing fragmentation of a working class that has failed to achieve any singularity in voice or action will involve exploration of how class is constituted by gender, race, sexuality, able bodies, and geography. Following the often neglected feminist and antiracist labor geography work, some of which has already been mentioned above, it is possible to address such “identities” as co-constitutive of class without de-centering a class analysis. Building class unity across space has been the objective of many social movements. Building such solidarity starts with a nuanced understanding of capitalism’s continued fragmentation of working people at multiple scales. These issues come together clearly in the experience of migrant workers in the global city, and fortunately recent work has focused on these workers (Wills et al. 2012).
LABOR GEOGRAPHY

Second, there is the question of shifting labor geography away from the study of sectors with strong institutional presence (e.g., unions in specific industries, the public sector). If labor geography is to understand how to overcome the fragmentation of workers, it must not “cherry-pick” sectors where unions have been established. Uncovering the “agency” of the marginal and unemployed must be part of the project. Peck and Theodore’s (2008) work on the role prison labor plays in shaping precarious day-labor markets and how churches have attempted to organize such workers is a relevant example. Labor unions are powerful, but are not the only example of workers exercising collective power. Mass social movements at urban, regional, and international scales in a variety of sectors must be studied in conjunction with capital’s attempts to restructure services and labor markets. Recent success in the US with implementing higher minimum wages at the municipal scale through community action is one example. Coalitions, alliances with migrant groups and the unemployed, are increasingly part of how labor builds power for political action. By rediscovering feminist scholarship on work that has been overlooked and integrating new antiracist and anticolonial approaches labor geography will be able to shift from a narrow investigation of workers in workplaces to workers in communities.

Moving beyond institutional actors, however, does not mean that labor geographers should completely abandon unions. There are those who undoubtedly see unions as creatures of capitalism and incapable of overthrowing its structures. Trade unions are uniquely situated in the economy and are vested with specific resources. It is premature to abandon, analytically or politically, institutions that continue to exercise significant power with employers and the state, despite the bureaucratic frustrations that plague such organizations. New social movements are often related to trade union successes and resources. Understanding how, when, where, and why trades union practices and structures can adapt to the challenges of precarious labor remains an open and timely empirical question.

Third, a labor geography that ignores the environmental crisis will be largely irrelevant in the midst of climate change and the very dependence of capitalism on the production of nature. Labor geographers have focused on the conflicts between workers and environmentalist and are now looking at how “blue” and “green” alliances are forming to deal with questions of more sustainable development within capitalist systems of exchange. Understanding that labor and capital produce nature in ways that benefit them materially will be increasingly important to labor geography. A labor geography that understands capitalism’s ability to produce nature for its own means will be able to counter neo-Malthusian interpretations that see the radical reduction of production (and perhaps the overall supply of labor) as the only immediate remedy rather than increased redistribution and economic security.

Lastly, there is the question of whether or not labor geography can ever truly break free of its theoretical roots in processes of capitalist reproduction. Das (2012) argues that it is necessary to discuss whether labor geography can encompass a more revolutionary project. For example, a “revolutionary” labor geography would map how workers build anti-capitalist economies. Das’s critique is aimed squarely at what he sees as a limited reformist, labor geography project. Ironically, however, he may have simply opened a new door of possibilities for labor geography as workers imagine new worlds of possibility.
**SEE ALSO**: Community; New Urbanism; Scale

**References**


**Further reading**


Labor market segmentation

Christian Berndt
University of Zurich, Switzerland

It has long been argued in the scholarly literature that the assumption that labor power is allocated according to the same universal laws that match the supply and demand of ordinary goods such as apples is fundamentally wrong. Yet this critique notwithstanding, this is exactly how labor markets are imagined and represented in public discourse and in our daily life. This is a good example of how orthodox economics works: statements like these are not about how real labor markets operate, rather they are about how labor markets should work.

It is against this truncated notion of labor and labor markets that scholars of different theoretical stripes put forward more realistic and complex representations. A prominent example is the heterogeneous body of work that has been assembled under the label “labor market segmentation theory” which emerged as a leading alternative to standard economic thinking in the 1970s. The various segmentation approaches share an understanding of the labor market as a multiple entity, fragmented into separate submarkets that subject market participants to different rules and work according to different logics. These submarkets provide employment opportunities that differ markedly in quality, resulting in a corresponding compartmentalization of the labor force into workers of higher and lower value (Peck 1996).

Scholars across the disciplines have engaged with labor market segmentation theories providing an overview of the evolution of this perspective and pointing to its strengths and also its weaknesses. It has been criticized that segmentation theorists explained labor stratification almost exclusively as a capital-driven process, attributing it either to the changing technical imperatives of the production process or to the controlling and disciplinary strategies of employers. Power asymmetries loom large in the latter account, capital implementing strategies of “divide and conquer” (Reich, Gordon and Edwards 1973, 361), exploiting existing ethnic and gender differences in the workforce, and in so doing actively reproducing and deepening these cleavages (Peck 1996, 53–54).

In geography it has mainly been Jamie Peck who introduced labor market segmentation into the discipline and addressed the blind spots of its various strands. This refers above all to the insight that labor segmentation processes are contingent both in time and space. Historically, interest in segmented labor markets emerged roughly at a time when the transition from competitive to monopoly capitalism made itself felt in industrial societies and in labor markets. Geographically, these arrangements were embedded in decidedly national territorial frames. This was the time of economies that combined cross-border mobility of finished goods with limited mobility of labor and capital, a national compromise that was to be challenged in the wake of the cascading rounds of outsourcing and offshoring that gave rise to the contemporary global division of labor. The emergence of what has variably been termed “network capitalism” (Boltanski and Chiapello 2005) or “supply chain capitalism” (Tsing 2009)
LABOR MARKET SEGMENTATION

posed new challenges for the “problem of labor,” challenges that are being met with a peculiar mix of old and new divide and rule strategies.

As commodity production morphed into a dizzying array of intertwined supply chains, production networks, and markets for semi-finished goods, newer generations of segmentation theorists put forward more complex explanations. There has been an increasing awareness that the primacy given to capital hides from view that employers are often simply reacting to changes in product markets, production processes, labor market constraints, and the regulatory environment rather than proactively formatting their environment according to their needs. With regard to labor supply, capital may put to profitable uses social differences in the labor force, but does not create these cleavages in the first place. All this led to a more detailed engagement with the “supply side” of the labor market, with labor being given a somewhat more active role. This crucially included a more systematic account of social reproduction (e.g., occupational socialization in the family, household division of labor), the regulatory role of the state at various spatial scales, and the role of organized labor (Peck 1996, 57).

These developments notwithstanding, however, labor segmentation theory appears to have lost traction in the scholarly community. More recent reviews within human geography and other social science disciplines rarely go beyond the comprehensive account given in Peck’s seminal contribution. This may at least partly be a reflection of the shortcomings of a still predominantly demand-side focused and capital-driven approach at a time of increasing geographical and social complexity. The remainder of the discussion therefore adopts a more flexible approach, putting emphasis on the question of difference and diversity in labor markets more generally. In doing so, attention is directed to three key developments associated with the rise of neoliberal capitalism.

Entrepreneurial subjects and labor as a diverse human resource

The first of these developments concerns the evolution of labor management strategies. In particular, this refers to recent developments within economic practice disciplines such as strategic human resource management and the emergence of diversity management strategies.

With the rise to prominence of the resource-based theory of the firm after World War II, Human Resource Management (HRM) emerged as a booming social science discipline. In its dominant mainstream variant, HRM is mainly about techniques to raise the commitment and performance of workers. It is generally assumed in the HRM mainstream that the interests of workers and employers are congruent and that both sides profit from good HRM practices.

The rise of HRM has a lot to do with the changing economic environment and the uncertainties arising from this. Insofar as individual knowledge and creativity and the social networks of the workforce replaced hard assets as key factors deciding about economic success, traditional management tools declined in importance. Because of the noncalculability of these resources there was an urgent need to develop different means of evaluation. The solution was provided by HRM: it is the individual worker, his/her skills, and his/her social connectivity that is becoming ever more important. In so doing, HRM developed into a powerful calculative tool to evaluate employees according to their individual capabilities.

Underlying the transformation is a radical shift in the way that issues of workforce difference are handled. During the heyday of the
Fordist corporation, the key to efficient production was mainly to organize labor in a way that social differences did not matter. This was helped by relatively homogenous workforces. More recently, however, the role of social difference in the private firm has been re-evaluated. In many countries of the Global North there have been marked shifts in the composition of workforces manifested in the feminization of employment, the aging of populations, and the increasing presence of migrant labor.

It is in the wake of these changes that diversity management emerged as an increasingly popular branch of HRM. A socially and culturally diverse workforce is valued as a competitive asset in its own right and HR departments are expected to actively manage, promote, and harness workforce diversity. Rather than just being an add-on to the “labor problem,” social differences are actively performed and utilized by HRM practices, making labor possible in today’s global production systems. In the newly emerging configuration the discovery of diversity and identity as a corporate asset and resource is underpinned by a liberal notion of equality. Workers are treated as separate individuals, capable of different, unequal performances. Accordingly, valuable resources controlled by employees due to their perceived gender or ethnic position are individualized. It is only those identity prescriptions that fit into the representation of workers as entrepreneurs that are acknowledged. Critical observers argue that by individualizing “diversity,” representations like these successfully hide from view that power asymmetries persist. In particular they point out that diversity management practices ignore collective sources of disadvantage, privilege those in society who happen to display the right mix of individual character traits, and render attempts to make claims based on collective, social aspects of identity positions extremely difficult (Gedalof 2013).

A concrete manifestation of this is the presence in many workplaces of a mosaic of employment statutes with extreme diversity of wages and other entitlements. The initiative is with capital and management in these processes, allowing the continuation of divide and rule strategies in different forms. However, just as in the case of segmentation during Fordism, division is never enough for effective rule. There is a corresponding necessity for (re)integration into a larger common frame and entity. The previous stratification of the workforce in primary and secondary labor markets was embedded in national systems of labor regulation. Today, the common ground is being provided by the corporate community, the mythical “firm as a family” transported in the mission statements and slogans that aim to rally workers behind a greater good: the globally competitive firm.

The new corporate world continues to distinguish core from peripheral workers. The latter include “unskilled” labor that has been virtually abolished in the Global North and moved to locations in the Global South, the temporary labor working “onshore” often next to more privileged colleagues, the rising labor market precariat unable to find anything else than spurious employment, and the unemployed. It is in these disadvantaged segments of the workforce that conditions accumulate which are not valued highly in today’s resource-driven organizations: workers who arrive at the labor market with the wrong set of qualifications and skills, who find it difficult to comply with the flexibility, communicative skills, mobility, and total commitment demanded by management (Boltanski and Chiapello 2005, 231–232).

It is frequently argued that it is the readiness on the part of workers to perform the subjectivities inscribed by the omnipresent rhetoric of workers as resources that explain part of the success of the new world of work. In her
account of the type of labor that is mobilized in what she terms “supply chain capitalism,” anthropologist Anna Tsing (2009) is adamant that tiring, repetitive, and physical work has certainly not disappeared. However, the problem of labor (i.e., the need to secure the consent and collaboration of wage-earners in order to realize capitalist profits) is now addressed in a different way. It is no longer done through the collective, political integration of workers into a society that equated economic and technological progress with social justice, but rather by means of developing entrepreneurial projects of self-realization that emphasize individual performance, responsibility, and mobility. In this logic, the individual worker’s autonomy is fostered as a means to her control, turning labor into active agents of self-regulation (Fraser 2003, 165).

Work tasks and job profiles are framed in particular ways in the corporate family. They are gendered, ethnicized, and coded according to nationality and age. And workers establish themselves by performing those very attributes. An investment banker must perform competitiveness, unscrupulousness, and readily embrace risk, a care worker empathy, affect, and availability. It is this double process of self-exploitation and external inscription that makes supply chain capitalism possible (Tsing 2009, 158).

Supply chains and the de-centering of the business firm

A second challenge concerns the ongoing blurring and reconfiguration of the organizational boundaries of the firm. The familiar argument is that the rigid corporate hierarchy has gradually given way to more spurious organizational designs, variably discussed with reference to concepts such as heterarchy, project-based forms of organization, or the network firm. As a result of this, it is no longer individual firms that compete with each other, but rather networks of firms bound together in ever more complex supply and value chains.

There is broad agreement in the literature that oligopolistic chain- and network-like governance structures are the key organizational units in today’s global economy. The rise in importance of practice disciplines such as logistics and supply chain management that deliberately extend governance regimes beyond the boundaries of the individual firm is testimony to this development. This development notwithstanding, however, it is often the case that a particular successful firm assumes the position of a role model for a specific configuration of capital–labor relations (e.g., Fordism, Toyotism). In the case of the network perspective with its imaginations of supply chains, logistics, and track and trace technologies, this role has most frequently been assigned to Walmart (Tsing 2009; Boltanski and Chiapello 2005, 223).

The apparently clear-cut boundary around the business firm is additionally dissolved with a view to what has traditionally been regarded as lying outside the realm of the formal economy, that is, the realm of social reproduction. This concerns both the emerging boundary zone of precarious, “informal” and “illegal” work that feeds the apparently formal production chains and networks as well as work that has been shifted from waged employment into a wide array of nonwage forms, be it voluntary, domestic, hobby, and forced work (Abbott 2006, 307–308). Social differences play a crucial role in maintaining both the hidden extension of ostensibly formal processes of the production of goods and services into less tightly regulated realms as well as successful attempts by economic and political decision-makers to reproduce an apparent clear-cut boundary between the economy proper and that which lies beyond. To a
LABOR MARKET SEGMENTATION

large extent it is the migrant woman that provides the labor in the latter realm, for instance, as cleaner, care worker, au pair, or maid. What those employment relations have in common is an expectation of almost unlimited flexibility and availability, connected with what André Gorz (2004, 27; author’s translation) termed “total mobilization of capacities and skills, including sentiments and emotions.”

One result of this is a blurring of the boundary between work and leisure time, between emotional engagement and the role of a professional, detached service worker. The task to police those boundaries is largely delegated to the worker herself. It is the individual care worker, cleaner, or maid who is responsible for making sure that she does not involve herself too much emotionally, for regulating her working time, or for making sure that the challenge to navigate five or six different workplaces per day remains manageable. Empirical research illustrates that these practices easily turn into self-exploitation, in so doing making the low wage and apparently unskillable character of this work possible in the first place.

Governing at a distance: the offshore logic of supply chain capitalism

A key reason for the changing role of social difference in the labor market has arguably been the growing social and spatial complexity of production, distribution, and consumption outlined above. As those organizational constellations are far more difficult to control than individual firms, the traditional solution of relying solely on stabilization via standardization can no longer be a viable option. This is another reason for the more indirect and subtle way in which to organize labor in today’s supply chains and production networks, a way of disciplining that governs workers “at a distance” (Fraser 2003, 167).

There is both a social and a spatial dimension to this. Starting with the former, this concerns the move toward more autonomy and responsibility for the worker, and a corresponding possibility for corporate decision-makers to detach themselves from the consequences of their actions. However, this “social distancing” is only one side of the coin, for the apparent organizational detachment is going hand in hand with tighter monitoring of other features of the supply chain, for instance, those involving prices and marketing arrangements (Tsing 2009, 156). And both moments come together if one looks at the increasingly prominent role of technological and quality standards in supply chains. At the same time as firms transfer responsibility for product quality and consumer satisfaction to workers and subcontractors, they increase the pressure on those actors with the implementation of ever more sophisticated track and trace technologies. These devices do not allow the slightest deviation from the script and do not forgive the tiniest mistake or lapse of concentration, making it possible to confront workers with their performance long after the product in question has left the shop floor or workplace. Supply chain technologies create an environment in which periods of idleness are abolished, creating a potentially unlimited repository of deeds and gestures that make workers subject to remote control and that are constitutive of “governance-at-a-distance” (Boltanski and Chiapello 2005, 247).

The geographical materialization of this is “offshoring.” Understood in an abstract form as a defining feature of contemporary capitalism, offshoring involves a double framing process. First, the articulation of people, things, and regions with global supply chains and markets, by way of transforming heterogeneity and contingency
LABOR MARKET SEGMENTATION

into productive labor, advanced production sites, special economic zones, and by disarticulating those elements from their social and spatial contexts. And second, a new moment of dis/entanglement that allows powerful interests to decouple the profitable exchange of labor and goods in global supply chains from the responsibility and liability for the social and geographical implications of these operations (Appel 2012). The attention turns to mobile standards, templates, and infrastructures as boundary-making objects. And it is with the help of various registers of social difference that this redrawing of boundaries is accomplished. This occurs by ritualized evocations of age-old stereotypes, such as the one manifested in the representation of southern female workers as “disposable” and “unskillable” (Wright 2006). Another example is the racialized legitimization of “old” disciplinary labor regimes in ostensibly advanced production sites, with hints at cultural traits that are apparently unsuitable for modern production and the corresponding paternalist “obligation” to be a strict northern father to southern factory daughters.

This is made visible with the help of ostensibly objective technologies that calculate defects per million parts, scrap rates, or productivity numbers. It is because these dis/entanglements are extremely complex and are constantly prone to fail and overflow that they require continuous logistical effort and financial investment. Methodologically, it is multisited ethnographic research, following things, people, and ideas and the plurality of relations they assemble, that is able to shed empirical light onto the realization of new regimes of labor control and stratification. Such an approach is also capable of illustrating how new organizational models travel to less privileged workplaces, above all in the Global South and how a key observation of early labor market segmentation, namely the dual labor market, has assumed a new global form. The last point takes up the observation that less desirable and overflow work continues to be outsourced to the Global South, either by moving jobs to locations in countries such as China or Mexico, or by mobilizing largely female migrant labor for reproductive tasks in northern households.

SEE ALSO: Corporate identity; Corporations and global trade; Corporations and the nation-state; Cultures of work; Emotional labor; Exploitation; Flexible labor markets; Fordism/post-Fordism; Gender, work, and employment; Global factory; Global production networks; Human capital; Inequality; Labor geographies and the corporation; Labor geography; Labor market; Trade unions

References

A labor market is most commonly defined as a platform that facilitates the buying and selling of labor power. Labor power refers to the mental and/or physical capabilities that can be used to generate use-values (e.g., the physical strength of a porter generates a capacity to carry luggage and thus replaces the physical strain that would otherwise be placed on the luggage’s owner) and/or exchange value (the same physical ability to carry luggage is priced monetarily and sold to any luggage owner who wishes to purchase this ability). In neoclassical economic terms, forces of supply and demand underpin the utility of any labor market, namely the generation of a price for which a transaction between a buyer and a seller of labor power becomes possible. Economists assume transaction prices are able to automatically “correct” and reach equilibria; when this occurs, the quantity of labor power supplied will match the price that buyers are willing to pay for that quantity of labor power.

To Karl Marx and Karl Polanyi, two major critics of neoclassical economic presuppositions, labor markets are not objective expressions of abstract forces of demand and supply – they are the precondition of a historically specific mode of production known as capitalism. This is because labor power is a fictitious commodity: people pre-exist labor markets as workers, builders, designers, and so on, and their capabilities are not inherently meant for sale. It is only when labor markets became recognized as legal institutions that these capabilities could be transposed into saleable commodities. And it is this saleable entity that constitutes the basis of value generation and capture: the labor power that goes into the production of a commodity could, in contemporary corporate parlance, be costed in monetary terms. Through this process, the producer is able to calculate how much value can potentially be captured from each unit of commodity – provided, of course, someone is willing to purchase this unit – through adjusting the quantity and quality of other variables that contribute to the production of this commodity (raw materials, machinery, output, etc.).

To understand labor markets as fictitious entities within the contemporary global system of capitalism is to understand these entities as unnatural in the first instance. This calls attention, in turn, to how this fiction is constructed and how it has come to be accepted as socially “normal.” After all, wage labor is taken as a primary basis of survival in many countries today. What is invariably more complicated is the appearance of labor markets: some labor markets are clearly defined for specific kinds of labor power (e.g., legal requirements for electricians or plumbers to pass province- or national-level tests before they are allowed to sell their labor within specified territories), while others have unclear definitions of what constitutes a labor market and where it exists (e.g., electricians or plumbers in many developing economies often do not possess formal qualifications; are not required to pass tests in order to sell their skills; and could move from place to place to “practice” as informal economic agents). It is the lack of a universally acceptable notion of a
labor market that reinforces its fictive nature: how well defined a labor market is in one place is contingent on how well regulators are able to weave together a coherent system of control, monitoring, and, perhaps most important, social compliance.

Noncompliance would lead to new spatial divisions of labor (Massey 1984). This refers to a situation in which the needs of buyers and/or sellers of labor power can no longer be met within pre-existing markets. For capitalists, the inability of local labor markets to satisfy growth objectives – often, if not always, expressed through the quest for increasing rates of profit – would trigger a search for labor power in alternative locations and/or the adoption of labor-substituting technology (e.g., cash deposit machines in banks taking the place of tellers). For laborers, the inability of local labor markets to satisfy their personal needs – such as the lack of sufficient job openings to match their skills or unsatisfactory remuneration – would cause them to relocate elsewhere to take up the same jobs or move into other local labor markets (e.g., a change in jobs).

Cognizant of the threat of noncompliance, state regulators operating at different scales often try to determine both supply- and demand-side practices in labor markets. On the supply side, they generate direct entry barriers (e.g., mandating a minimum working age and rigorous health checks for particular occupations, which in turn affect supply), leading in turn to labor market segmentation. A labor market becomes segmented when different subgroups are created on the basis of citizenship, residency, qualifications, gender, and numerous other variables. On the demand side, state regulators determine what count as “legitimate” and “fair” transactions (e.g., the imposition of a minimum wage, the control of foreign worker employment in specific sectors or locations, and the right to call an end to strikes, which in turn affect demand). The primary goal is not only to ensure that employers provide adequate social welfare; it is just as, if not more, important to minimize the number of interruptions to the capital accumulation process. In this regard, supply- and demand-side regulation of labor markets is the hallmark of a capitalist state.

It could be added, then, that a labor market is an institution – a political-geographic institution. This is because (i) its functions are governed by regulations that are underwritten by the state apparatus; (ii) these regulations apply to predetermined territorial parameters; and (iii) its reproduction is contingent on the extent to which existing regulations could serve the needs of buyers and sellers of labor power within these territories. Unfortunately, though in some ways unsurprisingly, social-scientific research has taken as unproblematic the state-centric presuppositions and quantifications of labor markets. After all, labor markets have become expressions and extensions of territories defined by the capitalist state. State agencies are also often in charge of collecting and publishing “official” labor market statistics, namely those on employment, unemployment, wages, and economic inactivity. These figures are often then further differentiated by industrial sectors, administrative geography, age, race/ethnicity, and gender.

Even when the state is taken as the primary unit of analysis, labor markets are inherently varied because regulations exemplify the ideology of the governing political party. Conservative or liberal political parties are likely to advocate more flexible labor markets through cross-border immigration and less regulation over minimum wages, while nationalist or left-wing parties tend to emphasize resident-first hiring practices and more secure employment tenure. Differing ideologies also impact on how the state
views labor movements. While some territorial jurisdictions constitutionalize the right of workers within particular jurisdictions to organize and represent themselves (including the right not to supply labor power) so as to secure improved employment terms, there are states that are fully committed to the demands of capital and refuse to legitimize autonomous collective bargaining in fear of strikes and social instability. The ideological orientations and/or vested interests of political actors in power at national, subnational, and supranational levels thus engender direct effects on market entry and exit for both buyers and sellers of labor power.

Economic geographers are aware of and have worked assiduously at transcending the limitations of state-centrism in labor market research. As Jamie Peck (1996, 84–85) argues in Work-Place, a highly influential account of local labor markets, unevenness in labor market phenomena and regulatory forms could just be as pronounced within as they would be between nation-states. While not conceding the importance of the national scale as a unit of analysis, Peck (1996, 105) calls for more incisive research on how local and regional forms are at once distinct and embedded within their national and international contexts (see Christopherson 2002). This emphasis on the interscalar socioeconomic linkages of labor markets is reflected subsequently in research on transnational migrant workers (Pratt 2004; McDowell 2005); on the relationship between policy transfers, the emergence of “workfare states” and the naturalization of precarious flexi-jobs (Peck 2001); and on the fast-growing field of labor geography which emphasizes the quest by working communities to carve out their own “spatial fixes” while remaining relevant to the increasingly transnational process of capital accumulation (Herod 2001; Lier 2007; Wills 2009). Where labor markets are taken as fixed within state space across much of the social sciences, economic geographers have demonstrated how the buying and selling of labor power is a gendered, multidimensional, and often transnational process that is always in the making – it is never about the self-correction of prices by the abstract forces of demand and supply.

SEE ALSO: Economic geography; International division of labor; Labor geography; Migration: skilled international labor; Political geography; Regulation/deregulation

References


LABOR MARKET

Further reading


Labor migration

Dennis Arnold
University of Amsterdam, Netherlands

Migrations have shaped modern history at least since the Atlantic slave trade and the dislocation, enclosure, and dispossession of the rural poor to populate the cities and fuel the booming labor needs of industry in England and other European countries (Casas-Cortes et al. 2015). In the century following 1820 about 60 million Europeans left for what they called the New World. These movements across the Atlantic were a critical aspect of industrialization and expansion into American frontiers. Less recognized are migration flows in Asia between 1846 and 1940. During that period, 48–52 million migrated from India and Southern China to Southeast Asia, the Indian Ocean Rim, and the South Pacific, and 46–51 million migrated from Northeast Asia and Russia to Manchuria, Siberia, Central Asia, and Japan. Despite the global nature of migration patterns over the past 200 years Western models and theoretical paradigms have, until recent years, been dominant.

The early moments of mass migration are important for at least three reasons (Casas-Cortes et al. 2015). First, the scale of Euro-Atlantic migration and capitalist development continues to inform the concepts used to investigate migration across the world. Second, migration studies emerged in the heyday of the process of mass industrialization and proletarianization in the early twentieth century. “Modernization” became the benchmark and model upon which much economic development theory rests, and the role of migrants in development processes has long been debated. Third, a concern for the social and economic integration of the migrant in nation-states has long dominated migration studies, marked by the bounded discourse of citizenship.

Contemporary labor migration challenges all of these points, and this entry addresses the following three aspects in turn. First, labor migration is globalized, moving well beyond the Euro-Atlantic world region. And, even when connected to industrial labor, contemporary labor patterns are qualitatively different, and characterized by precarization. The second section looks at the transnationalism turn in migration research, to understand both migrant subjectivities and the emergence of the migration–development nexus, which centers its attention on migrant workers’ remittances. The third section focuses on borders and the bordering of labor migration, to understand state attempts to manage both mobile labor and capital.

Globalization, precarization, and labor migration

A migrant laborer is generally defined as a person who goes from one place to another especially to find work. Migrants work across the entire occupational spectrum, from highly paid and skilled occupations in the knowledge economy, finance, and other sectors, to some of the world’s most exploitative jobs in agriculture, manufacturing, care work, and services. According to World Bank estimates, in 2011 more than 215 million people lived outside their countries of
LABOR MIGRATION

birth, an increase from 84.5 million in 1975. Of this figure, an estimated 105 million are international migrant laborers. Furthermore, over 700 million migrate within their countries. In China alone, some 145 million internal (rural-urban) migrants are currently working in coastal special economic zones and other urban and peri-urban areas. Migration figures are likely to be considerable underestimates, as “illegal” or irregular migrants are extremely difficult to collect reliable data on.

As early as the late nineteenth century, E.G. Ravenstein had observed that most migrants left their homes in order to search for work of a more remunerative or attractive kind. Ravenstein developed his 11 major laws of migration, which summarized the regularities that could be identified from data gathered from the 1871 and 1881 censuses. Some of Ravenstein’s findings remain valid, for example, the majority of migrants continue to move over short distances and large towns still grow more from migration than from demographic increase (Boyle 2009). In the 1960s, neoclassical economists built on Ravenstein’s laws and formulated a series of hypotheses concerning the volume of migration and its streams and counterstreams, with a focus on the push factors at the origin that stimulated moves and the pull factors at the destination that attracted movers. This approach continues to influence thinking around issues of labor migration.

Neoclassical economic theory, which assumes individual free choice and free flows of factors of production, predicts that workers will migrate from low-wage to high-wage areas. This theory attempts to explain both flows from the poorer South to wealthier countries in the North, and migration patterns between neighboring countries and within countries. The approach has been criticized for positioning migrant workers as being determined and overwhelmed by economic structural forces (Mezzadra and Neilson 2013). It does not reflect the complex set of motivations that lead migrants to leave their families and communities, nor does it consider migrants’ agency in their new jobs and communities.

People who migrate specifically for work purposes may involve at least four channels (Samers 2010). First, people may migrate because they expect better employment conditions, such as higher wages. Second, people migrate because friends, relatives, and acquaintances inform them of work opportunities. Such networks may help the migrants establish themselves, with help in housing, information on work permits and other state regulations, and other daily necessities. Third, people may migrate because an employer recruits them. Recruitment may involve friends, private (for profit) labor agents, and/or governments. Fourth, and related to the previous three points, migrants may be recruited illegally, often trafficked. Such migrants are most likely to do low-paid and highly exploitative work.

If one agrees that migrant workers are not simply pushed and pulled around the globe in an individualized, rational manner, then the question arises of how migration patterns can be conceptualized. Recent scholarship under the broad banner of globalization studies offers insight on political economic structures that frame migratory experiences. The following section turns to this literature, before considering labor migration as a social process shaping these structures.

Global precariousness

A prominent approach to understanding migrant labor is the dual labor market theory. This view claims that there are two sectors in modern, industrial labor markets: a primary sector with decent pay, job stability, and generally good working conditions, and a secondary sector
composed of work with low pay, lack of stability, and poor working conditions. Migrants have long been seen as most likely to take jobs in the secondary sector. This perspective was overly simplistic even at the height of the Fordist labor accord in the industrialized countries, when significant proportions of the working and middle class could expect decent wage and state welfare provision. The reality is migrants are no longer at the fringes of labor markets or viewed as an anomaly. In fact, migrants are now considered the paradigmatic worker, part and parcel of broader labor trends in global capitalism.

On the one hand, migrant labor is a response to competitive pressures from capital. On the other, capital's increased global mobility and need for flexibility is a reaction to the wave of struggles and strength of labor in the industrialized countries in the 1960s and 1970s. The saturation of markets, along with the high levels of competition that introduced the process of global outsourcing, obliged firms to develop techniques and technologies to enhance or create mobility and flexibility, which also created new barriers to labor organizing. These processes have diminished workers' bargaining power and rights across different countries and contexts, and at the same time generated new forms of struggle.

Over the past several decades, the growing power and reach of global capital has exceeded the ability of nations and labor movements to regulate it, exacerbating inequality, precarious work, and broader social precarity. Numerous labor trends have been associated with neoliberal globalization, including a decline in attachment to employers, an increase in long-term unemployment, growth in perceived and real job insecurity, increasing nonstandard and contingent work, risk shifting from employers to employees, a lack of workplace safety, and an increase in work-based stress and harassment. The lack of public and private investment in skills and development is accompanied by a lack of access to schooling, where women and ethnic and racial minorities disproportionately bear the brunt of these disadvantages. Yet precariousness is about not only the disappearance of stable jobs and other workplace-specific issues, but also the questions of housing, debt, welfare provision, and the availability of time for building effective personal relations.

These trends are not entirely new, and follow previous patterns. Capital is continually in search of spatial, technological, and product fixes, but with each phase of innovation the intensity of capital investment and productivity requirements increases. Not only have global production networks stretched across national boundaries to cover greater geographic scope and new “frontiers,” but also lead times have become shorter to respond to oscillations in consumer demand. As the geographies of production continue to expand, the process is often reproduced with tighter margins for lower tier producers and employers, which tend to provide lower remuneration for workers. Greater flexibility is the mantra of the day, from labor-intensive garment factories, to retail and service sectors, to knowledge work in the academe. In short, precarious labor regimes, and the central role states have played in configuring and reproducing them, have reconfigured the geography of the reserve army of labor. Migrants are at the center of the contemporary labor process (Wills et al. 2010).

In addition to their appeal to capital, they also have major political advantages as a flexible labor supply. Even those who cross borders legally find themselves politically disenfranchised. Lacking citizenship, they are more likely to encounter restrictions on their access to employment,
LABOR MIGRATION

welfare, and the political process. In contemporary globalization, migrants are not only less likely to organize trade unions, they are the sine qua non of flexibility that many employers seek. Migrant workers are attractive to employers precisely because they are migrants (Wills et al. 2010).

Studying migrant labor reveals a general shift of responsibility that follows the capitalist dream of an available labor force disconnected from the need for its reproduction. Not only has capital increasingly outsourced beyond national borders in the contemporary wave of globalization, so too has the social reproduction of labor been outsourced across borders, generally to migrants’ home communities and countries. The neoliberalization of social reproduction entails a reduction in state-provided social assistance for its citizens. Migrants are typically not afforded any support for the social reproduction of their own families, whether in their home or host country. This is a clear example of re-regulation that places burdens on families and individuals, away from capital and states.

This line of analysis leads to a simple equation: the less social support and the fewer rights migrants have, the more vulnerable and docile they are likely to be (Samers 2010). While not a universal truth, this contention helps to understand the rights claims that migrant labor may or may not make of employers and states. This is most salient when workers fear being sacked and/or deported. Clearly, a precarious migrant labor force has been central to processes reworking the balance of power between capital and labor. With that in mind, consideration should be given to how it is that migrant labor has come to be viewed by many analysts as a key contributor to economic development, and how migrant subjectivities are understood in relation to global-scale transformations.

Transnationalism and the migration–development nexus

It is only recently that the labor migration debate has begun to take account of the complexity of the development–migration interaction (Munck, Schierup, and Delgado Wise 2012). Munck and colleagues contend that in the 1950s and 1960s the dominant modernization perspective saw a synergy between the two – migrants were deemed “modernizers” who could contribute to economic growth in the receiving country through their labor, and in the sending country through remittances and their return with “modern” skills and values. In the 1970s and 1980s a structuralist Marxian–influenced model prevailed, which focused on negative consequences, such as “brain drain” and new patterns of dependency oriented around remittances. This “development of underdevelopment” model was replaced in the 1990s by middle range theories, most prominent the focus on transnationalism and transnational household livelihood strategies. The transnational turn in migration research increases the awareness of diverse strategies of mobility, while raising questions about migration–development precepts (Bailey 2010). Informed by post-left scholarship around the same time, it follows that gender, class, age, ethnicity, and racializations came to be viewed as key to mediating the migration–development nexus. Both transnationalism and migrant–development nexus approaches have received much attention, and have informed multiple actors’ views of migrant workers as agents for social and economic transformation.

Transnationalism

Transnational theorizing began in the early 1990s, when scholars found that the migrants
with whom they worked had developed transnational practices that conventional migration theories did not adequately capture. The approach critiqued traditional migration theory and its binary treatment of migrants as individuals who either departed (emigrants) or arrived (immigrants). Traditional approaches contributed to “methodological nationalism,” as the nation-state was deemed the primary unit of analysis in migration scholarship.

To overcome this dichotomy, and to better understand migrants’ multiple attachments, researchers proposed that migrants be understood as forming part of two or more dynamically intertwined worlds and transnational migration be seen as the processes by which immigrants initiate and sustain social relations that link together their societies of origin and settlement. Thus, sending and receiving societies became understood as constituting one single field of analysis, in which migrants often maintain a variety of social, economic, cultural, and political ties with their home. In short, migrants are more likely to develop complex affiliations, meaningful attachments, and dual or multiple allegiances to people, places, and traditions that lie beyond the boundaries of nation-states (Cohen 2006).

The transnationalism approach has also contributed to understanding labor markets, work processes, and industrial relations within localities and the ways in which they are also shaped by various flows across space. In contrast to a proliferation of research on the flows of investment capital, traded commodities, or organizational processes in transnational firms and the differential embedding of capital flows in space, transnationalism has contributed to highlighting the transnational flows and connections established by workers themselves. In many respects, studies of migrant transnationalism are the complementary inverse of scholarship on large corporations and intergovernmental organizations, representing instead a view of globalization from below.

Transnational migrants at various levels in the labor market are frequently separated from their immediate families. Transnational household arrangements often imply a significant restructuring of the division of unwaged domestic labor, including the reproductive work of running a household, maintaining a home, and raising children (Kelly 2009). This is pronounced in wealthier countries where migrants do domestic work, while impacting migrants’ families in poorer countries. This phenomenon is known as “care-chains.” It is characterized by the increase in dual-earner couples and single parents in wealthier countries, in part a result of the dissolution of the gender contract in which (typically) male breadwinners were paid a family wage and women did unpaid domestic labor. While care has always been labor, the social relation has changed with the increasing number of migrants who are paid to do the task of social reproduction. For the migrant families back home, social reproduction is often undertaken by an array of social networks, as these families typically cannot avail of state-derived assistance.

For many migrant workers in low-paid and insecure employment, it is precisely their transnationalism that undermines their capacity to assert rights in the work place. As citizens of one territory but workers of another, they may only be protected in theory by labor laws. In practice, their migrant status and their transnational obligations to family members back home mean that there would be a heavy price to pay for pushing back against abusive employers or the authorities of host states. Yet transnational migrant experiences are diverse. Many increase not only their incomes, but also their social status. Going further, Hardt and
Negri (2000) have argued that migrants are a special category of the poor imbued with an irrepressible desire for free movement. Mobility and migration are viewed as a powerful form of class struggle. Migrants play a role in constructing an alternative hegemony, in which they normalize movement and treat the globe as a common space. These and related claims to migrant “cosmopolitanism from below” have generated critiques that the migrant experience is overly romanticized. Against this backdrop, the migration–development debate opens another line of inquiry that considers both migrant subjectivities and global transitions.

Migration–development nexus

Analysis on the link between international labor migration and development has been contentious. On the one hand, many view the outflow of people not only as a symptom of underdevelopment but also as a cause of its perpetuation, and on the other hand are researchers who regard migration both as a short-term safety valve and as a potential long-term instrument for sustained growth (Portes 2007). Since the 1940s in the case of Mexico, with the implementation of the Bracero program, and the 1960s and 1970s for countries including the Philippines, Morocco, and Turkey, governments have encouraged under- and unemployed workers to migrate to the United States or Western Europe. The expectation was that the workers would contribute to economic and political improvements and stabilization in their home countries. However, the long-term results of labor migration schemes often proved disappointing (Castles and Wise 2007). This resulted in a generally negative view of migration, that it undermines the prospects for local economic development and yields a state of stagnation and new patterns of dependency (Massey et al. 1998). Fueling the negative view, “brain drain” became the popular view, arguing that it was generally “the best and the brightest” that left.

Beyond these debates, theories of national development have seldom paid much attention to international migration (Portes 2007). At best, migrant labor flows were treated as a marginal phenomenon – a reflection of the problems of underdevelopment. In recent years this has changed. The size of migrant worker communities and the volume of the remittances they send home have prompted theoretical models that emphasize financial flows. Over the past two to three decades migrants have been redefined as “heroes of development.” The links between migration and development are increasingly seen in a positive light, and “brain gain” replaces “brain drain” as a key metaphor.

In many countries, remittances now represent a significant contribution not just to their families and communities, but also to the national economies in which they are spent. According to the World Bank, remittance flows to developing countries were estimated to have totaled US$401 billion in 2012, an increase of 5.3% over the previous year. Global remittance flows, including those to high-income countries, were an estimated $529 billion in 2012. The top recipients of officially recorded remittances in 2013 were India ($70 billion), China ($60 billion), the Philippines ($25 billion), and Mexico ($22 billion). Remittances sent home by migrants to developing countries are equivalent to more than three times the amount of official development assistance.

For some, remittances play a key role in resolving past financial bottlenecks and providing the necessary resources for social and economic development. Furthermore, remitting involves economic and noneconomic resources and flows, and includes the exchange of money, knowledge, and universal ideas (Bailey 2010). Recognizing
LABOR MIGRATION

the economic importance of migrant workers and remittances, organizations including the World Bank have become champions of migrants as key contributors to development. Indeed, the World Bank (2014) argues that “Remittances generally reduce the level and severity of poverty and lead to: higher human capital accumulation; greater health and education expenditures; better access to information and communication technologies; improved access to formal financial sector services; enhanced small business investment; more entrepreneurship; better preparedness for adverse shocks such as droughts, earthquakes, and cyclones; and reduced child labor.” The Bank’s activities are centered on mobilizing migrant remittances to develop financing instruments for leveraging migration and remittances for national development purposes.

Bailey (2010) identifies other trends in the migration–development nexus literature. Drawing on transnationalism studies, it is recognized that promotion of remittances and the realization of any benefits occur through increasingly diverse and flexible forms of migration, notably temporary, return, circular, and transnational migration. Furthermore, the potential benefits of diverse remittances and diverse migration are most efficiently realized through coordination of the activities of an increasing number of stakeholders, including individual migrants, social networks, households and transnational families, communities, hometown associations, epistemic communities, diasporic knowledge networks, regional and national governments, and international organizations.

Looking at the matter from a different perspective, Castles and Wise (2007) contend that the World Bank and other organizations associated with the Washington Consensus emphasize the positive potential of the migration–development nexus as a result of the crisis of legitimation of neoliberal globalization. As discussed above, globalization has failed to deliver economic inclusion and greater equality for the majority of the world’s population. Migrants have become part of a solution to these and other failures of widespread liberalization, structural adjustment, and commoditization. Remittances have become part of a new development mantra, with the belief that they can be channeled into investments that will lessen or even overcome underdevelopment, thus making up for the failures of mainstream approaches to development. This approach has proven particularly attractive to states that lack a coherent national development strategy.

At present, the evidence for the links between migration and development remains weak. Portes (2007) argues that the rosy predictions are exaggerated. He contends that there is no precedent that any country has taken the road toward sustained development on the basis of the remittances sent by its expatriates. Importantly, migration can lead to vastly different consequences – economic stagnation, the emptying out of sending places, and the massive loss of talent with wide-ranging impacts on the energizing of local economies, new productive activities, and significant contributions for scientific and technological development (Portes 2007). Thus, many scholars have returned to the long-held view that international migration does not stem from a lack of economic development, but from capitalist development itself (Massey et al. 1998, 277).

Borders, the state, and migrant labor regimes

Transnationalism research has clearly been instrumental in the shift away from methodological
LABOR MIGRATION

nationalism. Yet there exists a polarity in migration studies between an economic consideration of migration under the headline of exploitation, explored in the first section, and the more positive views among many transnationalism scholars that highlight migrants’ hybridity and “cosmopolitanism from below.” Labor migration has always been a contested field in state regulation of populations in capitalism. Capital’s solicitation of labor mobility has always gone hand in hand with manifold attempts, particularly on the part of states, to filter, to curb, and even to block it (Mezzadra and Neilson 2013). Geographical literature has been attuned to heterogeneous migrant, state, and capital practices through the continuous reworking and reshaping of both social and geographical borders and scales, the focus of this section.

Cohen (2006) provides a summary of the shift from open immigration policies in much of Western Europe and the United States during the postwar era, to the more restrictive policies from the 1970s when sharp restrictions were first imposed (While in the United States there has not been an absolute decline in legal immigration, there have been significant qualitative changes in the occupational and legal categories admitted – from immigrants to refugees and from agricultural and mass production workers to the professional, technical, and independent proprietor categories. Japan has not permitted significant numbers of labor migrants throughout the postwar era.). Cohen attributes the shift to several factors, including:

- the oil crisis, which contributed to a wave of redundancies in energy-intensive industries;
- the rise of xenophobia, particularly among the working class fearing competition in the labor market;
- the organization of migrant workers who increasingly pushed for family reunification policies, contributing to many politicians and interest groups pressing for more restrictions;
- the rise in the cost of reproduction in terms of child care, language training, and education;
- economic restructuring, or the general shift from Fordist mass production to flexible specialization and offshoring of production and services.

Speaking to contemporary state policy, Cohen contends that the modern state has sought to differentiate the various people under its sway by including some in the body politic and according them full civic and social rights while seeking to exclude others from entering this “charmed circle” (Cohen 2006). Currently, formal citizenship and visa regulations are the most prominent form of regulation at national governments’ disposal. Typically, they are used to control access to and the duration of different kinds of work. This is known as “differential exclusion,” describing the incorporation of migrants into some areas of national society, especially the labor market, and their exclusion from others, such as social welfare and citizenship. While a good starting point, this binary concept is easily complicated when considering the different experiences of skilled and unskilled migrants, those “legal” or not, and variation across national contexts. It is more useful to view migrant labor management as creating different degrees of precarity and vulnerability, as well as opportunities and empowerment.

To understand how migrants traverse and reshape the multiple spaces of globalization, the study of borders and labor migration provides much insight. Borders are complex social institutions that can both connect and divide not only migrant labor, but also the times and spaces of global capitalism. This includes both political borders, whose relevance has altered yet certainly not gone away, and the multiplication
of social borders that very often shape migrant workers’ experiences. In migration studies, the border has typically been comprehended as a wall, both literally and figuratively, that excludes migrants from a national territory. The US-Mexico border is the prototype of the border-as-wall, an “economic dam” that seeks to prevent migrant flows from “contaminating” the US body politic, an approach and discourse that has been replicated elsewhere, including the Thailand-Myanmar border (Arnold and Pickles 2011). Border securitization has not only failed, it has contributed to riskier efforts on the part of migrants to evade authorities, while also leading to a proliferation of human traffickers and others capitalizing on the criminalization of certain migrant labor populations – very often police and military officials mandated to “protect” the border.

Yet national borders are no longer the only or necessarily the most relevant borders dividing and restricting labor mobility (Mezzadra and Neilson 2013). Migrant labor encounters complex sets of social relations that seek to reinforce labor power as a commodity by turning migrant flows into mobile governable subjects. In addition to the renewed emphasis on political borders, social borders have multiplied. Social borders play a central role in shaping labor markets by filtering and differentiating labor and bodies according to skills, citizenship, race, ethnicity, and gender. Indeed, the production and reproduction of differences is key to capital accumulation strategies. Impacts include the differential allocation of rights across populations. Very often, this entails being “in one’s place out of place” and “out of place in one’s place” (Casas-Cortes et al. 2015). All labor is subject to bordering, yet migrant workers’ experiences are most pronounced. In particular, capital has proven particularly adept at exploiting the continuities and the gaps – the borders – between different migrant populations.

Social borders are very often regulated and reproduced by states. Indeed, states remain a pivotal institutional apparatus that regulates the lives and politics of workers (Coe and Jordhus-Lier 2010), and are key to understanding power configurations and their articulation with capital-labor relations (Mezzadra and Neilson 2013). However, governments are far from being in control of labor migration, much less capital flows. One approach to understand the asymmetries among nation-states, migrant labor, and global capital flows is the continuous reshaping and intertwining of different geographical scales. The local, national, regional, and global have never been given, and a nested hierarchy of these scales has long been a running debate among geographers. Migrant laborers are a particularly interesting group, because they regularly traverse and intersect different kinds of scales and spaces, making the very concept of scale increasingly complicated in its constitution.

The variegated politics of scale and the centrality of migrant labor have received much attention. In their work on the Indonesia–Malaysia–Singapore Growth Triangle, Sparke and colleagues (2004) view cross-border economic zones as crystalized spaces of multiple ambitions. In this context the strong state withdraws to allow forms of economic autonomy that enhance profitability, or “exceptions” to state regulation of territory and (migrant) population in other spaces. Studies have suggested that the Growth Triangle is not so much a question of market versus state, or a transition from state to market, but that the creation of markets is occurring in settings where the state is sometimes very strong while in other areas it is nearly absent, creating a spatial variability in managing
LABOR MIGRATION

migrant and other populations, particularly in cross-border economic zones.

In Greater Mekong Subregion economic zones, a different state–labor arrangement has been observed (Arnold and Pickles 2011). At one Thailand-Burma border textile and garment manufacturing zone, the convergence between cross-border regionalization and globalization is “articulated in localized spaces conducive to mobile capital that, on the one hand, straddle and blur national boundaries, and on the other, redefine and reify borders, particularly in terms of flows of migrant labor” (1599). It is a softening of the border as far as its geopolitical function and its role in the control of the flows of capital and commodities are concerned. To this there corresponds a new rigor in the control of labor mobility. A multiplicity of borders surround the life and labor of migrant workers from Myanmar, ranging from widespread racism to differentiated legal statuses. In this context, it is certainly a zone of exception but not because the state’s role is compromised or because the state has withdrawn from that space. Instead, it is precisely the way in which state agencies, particularly the police, national security agencies, and border guards, work together to regulate migrant workers’ lives to create and sustain the conditions for a regime of precarious migrant labor that enables the garment industry to survive in the region. Neither the state nor employers are responsible for the social reproduction of labor power, and both avoid engagement in tripartite functions associated with industrial relations prominent in the Fordist era.

In these and other contexts in which migrants face multiscalar challenges, the question of labor organizing becomes critical. In US workplaces trade union efforts to defend the rights of their (immigrant) members elicit an “enclave of rights” (Parks 2014). Parks’s study shows that immigrants occupy their workplaces as both subjects of immigration policy and protected workers. The workplace is identified as a highly localized sphere of engagement that both generates its own practices and realizes the local manifestation of multiple scalar phenomena including global capital and migration flows, national state policy regimes, corporate management structures, and trade unions. For undocumented immigrant workers, the workplace produces both illegality and legal protection – the former under immigration law, the latter under labor and employment law. Unions negotiate the uncertain and often conflictual intersection of these two legal domains by inhering in place workers, rights (conferred by federal law) and rights attained through social membership (union registration). Thus, workers are accorded an enclave of rights at the local (workplace) level, contributing to the highly uneven space of (in)security of migrant workers in the United States.

Conclusion

Labor migration has profoundly impacted hundreds of millions of people worldwide since the mid-nineteenth century. Past labor migration scholarship tended to treat migration patterns as discrete events, auxiliary to labor markets and national development trends and economic growth. Contemporary migrations are typically viewed as heterogeneous, as is scholarship on the patterns, trends, and implications. There is an increasing recognition that migration is a complex and multifaceted process, rather than a discrete event, and labor migration is more fluid than many commentators have assumed in the past.

Over the past 30 years, global outsourcing and precarization of work has become central to the reconfigured geography of the reserve army of
labor. Both trends have contributed to delimiting the collective power of labor. Indeed, economic growth has occurred alongside a reduction in workers’ share of overall wealth. In the process, migrant workers have become the paradigmatic figure in the contemporary labor process. On the one hand, global capital drives migration and reshapes its patterns, directions, and forms. On the other hand, labor migration is an important factor in bringing about fundamental social transformation of both home and host countries, and is, in turn, itself a major force reshaping communities and societies.

The “transnationalism turn” in labor migration studies argues against both “methodological nationalism,” and depictions of migrants as passive victims to global forces. Focusing on transnationalism implies that local and national scales are not sufficient to understand contemporary migration processes. It centers migrant subjectivities and seeks to understand migrants’ multiple attachments that stretch across national borders, communities, and other relations previously viewed as bounded containers. Transnationalism research has contributed to a growing body of literature on the migration–development nexus, which claims that remittances can promote local and national development in their countries of origin. Migrant workers’ remittances are seen as a way of alleviating the ravages caused by globalization, which include increasing inequality and marginalization of large population segments. Both transnationalism and migration–development literature have, however, been critiqued for celebrating migrants’ “cosmopolitanism from below.”

The study of borders and states’ efforts to manage mobile capital and labor may privilege scalar analysis, hybridity, and the implications of racialization, gender, and citizenship in shaping labor markets. It is an approach that recognizes global capital trends without viewing them as paralyzing blows that determine subjective experiences. Instead, the spaces that migrants traverse are mutable and multiple, offering opportunities and constraints. The spatial reorganization of labor has multiplied and fragmented its forms. In these and other debates, geographical literature has been attuned to the continuous reworking and reshaping of migrant experiences. The theoretical and empirical diversity of research has made significant contributions to understanding important political, economic, and social changes.

SEE ALSO: Development; Domestic workers; Globalization; Precarious work; Remittances; Transnationalism; Unfree labor

References


Casas-Cortes, Maribel, Sebastian Cobarrubias, Nicholas De Genova, et al. 2015. “New Keywords: Migration and Borders.” Cultural Studies (special issue), 29(1).


LABOR MIGRATION


Lake climates

David A.R. Kristovich
University of Illinois, USA

Thermal and surface roughness characteristics of lakes can lead to significantly altered nearby climate conditions, referred to as “lake climates.” For example, locations close to large lakes tend to have moderated annual temperature variations, with cooler summers and warmer winters than outside the region of lake influences. Depending on the lake size, shape, depth, location, and surrounding topography, lakes can influence a wide range of climatic conditions, including atmospheric humidity, cloudiness, precipitation, and winds. Large lakes, such as those in the Laurentian Great Lakes in North America, Lake Victoria in East Africa, and the Great Salt Lake in Utah, affect climate conditions over areas of $10^4$–$10^6$ km$^2$ or more. Smaller lakes also can have significant local effects, but on correspondingly smaller areas.

Fundamental causes of lake climates

The fundamental reasons that lakes exert large influences on nearby climate conditions are twofold: (i) lake surface temperature variations tend to be delayed behind those of the air surrounding the lake, and (ii) lakes tend to be less rough (more flat) than the surrounding land surfaces (discussed later). The former reason is illustrated in Figure 1, which shows 10-year average annual temperature variations for Lake Michigan water temperatures (measured at Muskegon, MI) and air temperatures (Milwaukee, WI). The causes and effects of these two influences are described below.

Lake surface temperatures vary largely in response to changes in the overlying atmospheric conditions and in heat loss/gain through radiation processes. However, lake surface temperatures vary much more slowly than regional air temperatures, resulting in important differences between the lake and the overlying air. For example, at night when air temperatures cool below the lake surface temperature, heat is transported from the lake surface to the air through conduction, longwave radiative loss, and convective processes, resulting in a cooling of the lake surface. If the air has a low water vapor content, relative to the content slightly above the lake surface, upward moisture transport leads to evaporation, further decreasing the temperature of the lake surface. The following day, incoming solar (shortwave) radiation would tend to warm a thin layer of the nearby land areas rapidly, while slower warming of a deeper layer of the lake surface takes place. Eventually, the air near the lake can exceed the lake temperatures, resulting in downward heat transport. Finally, this downward heat transport slowly reheats the lake surface, which starts the diurnal cycle over again. Therefore, diurnal or annual temperature variations for lake surfaces are delayed relative to those of the air in the same region.

Lake surface temperatures vary more slowly than nearby air temperatures because water has a specific heat content that is approximately four times as great as air ($1.1 \text{ITcal g}^{-1} \text{K}^{-1}$ and $0.24 \text{ITcal g}^{-1} \text{K}^{-1}$, respectively). Therefore, it takes more than four times as much heat to increase the temperature of water by 1 K as it
LAKE CLIMATES

Figure 1 Ten-year average (1970–1979) air temperatures for Milwaukee, WI, and lake temperatures for Muskegon, MI. Modified from Kristovich (1988). Reproduced by permission of David Kristovich.

It takes to warm the same mass of dry air by 1 K. So, because air temperatures respond quickly to changes in nearshore land surfaces, but lake surface temperatures can only respond slowly to air blowing from the land over the lake, time periods of average “climatic” lake temperatures lag behind those of the air. For the Great Lakes (cf. Figure 1), these processes result in annual average lake surface temperature variations being delayed by about 1 month, with a decreased magnitude of variation, relative to the nearby air temperature variations.

The temporal lag in the annual cycle of lake water temperatures provide for time periods when the lake surfaces tend to be warmer than the overlying air (fall and winter) and times when the overlying air tends to be warmer (spring and summer). For example, Figure 1 shows that averaged over long time periods, air temperature tends to be 2–5°C warmer than the lake during the spring to early summer months, from about April to July. During the autumn and early winter time periods, September through February, average lake temperatures can be more than 10°C warmer than the nearby air temperature, which results in rising air parcels. Therefore, autumn and winter months are considered unstable lake climate seasons and spring and summer months are considered stable lake climate seasons.

Atmospheric and lake dynamics, such as winds and currents, also play a large role in the development of lake climate regions. During the cool season, air temperatures exhibit occasional “cold-air outbreaks,” with warmer temperatures between outbreaks. Those outbreaks are when the air is much colder than the lake surface, often producing the most intense lake-effect storms. Therefore, while the long-term climatic average conditions look smooth (as in Figure 1), the strongest modification of the local climate occurs episodically.

Water movements in lakes also have important influences on lake climate regions. Long-distance horizontal movements of lake waters are hampered by the lake shores, but vertical movement within the water column can greatly influence lake surface temperatures. During the stable season, warming of the lake results in a shallow warm layer near the top of the column of water. Transport of heat downward into deep layers is inhibited because the warm layer is less dense than the colder waters below. During the unstable season, cooling of the lake surface causes the water to become more dense (down to temperatures of about 4°C). Such dense water tends to sink and mix through the water column, slowing the rate at which the surface temperatures cool. For this reason, the average spatial variation in lake surface temperatures is highly dependent on the lake water depth during the cool unstable season, but not as dependent in warm months.

The magnitude of the influence of lakes on their local environments is largely determined by the rates of exchanges of heat, moisture, and momentum between the lake surface and overlying air. These rates are not only driven
by differences between the water surface and air temperature and moisture content, but are also strongly influenced by such factors as wind speed, lake water salinity, ice cover (Assel 2005), oils on the water surface, and so on. Wind speeds over lakes tend to be greater than over nearby land areas because of a lack of surface obstacles (i.e., trees, buildings, etc.), and this tends to increase heat and moisture transfer rates between the lake surface and overlying air.

It should be noted that lake climates can be quite similar to climates close to oceanic shorelines, due to many of the same physical causes. However, there are a few distinct differences: large-scale oceanic currents transport thermal characteristics of water over large latitudinal distances, which is not possible in lakes. In addition, because lakes are more confined to smaller areas, lake climate regions often experience the influences not only of the lake waters, but also of the land areas upwind and to the side of the lake. Therefore, lakes tend to influence climate conditions less than oceans and over a smaller area.

Cool, unstable season lake climates (autumn and winter)

Some of the most severe winter weather conditions observed in mid-latitudes occur through modification of the atmosphere by lakes. Such “lake-effect” storms are responsible for the Great Lakes snowbelts, and have large positive (tourist industry, building repair contractors and suppliers, etc.) and negative impacts (snow removal costs, health impacts, etc.) for affected communities. For these reasons, considerable effort has been made on characterizing the cool-season lake climate within the Great Lakes region and elsewhere.

Snowfall may be the most discussed aspect of cold-season lake climates, particularly after events dropping excessive snowfall (e.g., 1.75 m in northern Ohio in November 9–14, 1996, near 3 m on the Tug Hill Plateau, NY, in February 2007, 1.65 m near Buffalo, NY, in November 2014). However, lakes also modify many other climate conditions. In a comprehensive investigation for the Great Lakes region, for example, Scott and Huff (1997) compared climate conditions close to the lake shores (within 80 km) to those further away from the lakes, using climate data collected over a 30-year period. They found that during the cool season, lake climate regions were warmer (daily minimum temperatures by up to 8°C, maximum temperatures up to 2°C), skies tended to be cloudier (average of 1/10 sky cover increase), and the humidity (measured by vapor pressure) was greater. Average snowfall in the lake climate regions was more than double that outside of those regions in some locations.

Isolated and groups of smaller lakes also have been shown to have marked local influences on cool-season local and regional climate. However, many small lakes tend to become covered with ice, which can greatly decrease these influences during later periods in the cool season. Small lakes that are sufficiently deep (such as the New York Finger Lakes) or have high salinity (such as the Great Salt Lake in Utah) can remain ice free and influence local climatological conditions throughout the cool season.

Some of the most cited impacts of cool-season lake climates occur through changes in the atmosphere at the surface (ground level), but changes in deeper layers of the atmosphere can also be quite important to regional climate conditions. For example, many studies have indicated that the Great Lakes tend to lead to intensification or development of synoptic-scale low-pressure storm systems and weakening of high-pressure systems. Ultimately, these influences can lead
LAKE CLIMATES

to changes in weather conditions much further from the lakes than those typically considered to have lake climates (e.g., Petterssen and Calabrese 1959). Other, less obvious influences of lakes include moistening of lower layers of the atmosphere, which can ultimately lead to development of convective snow showers over land due to daytime surface heating or formation of dense advective radiation fogs when nighttime radiational processes result in cooling of the moistened layer (e.g., Eichenlaub 1979). In some cases, clouds that develop over the lakes can be blown far downwind, affecting temperatures of wide regions.

Typical patterns of movement of large-scale atmospheric features (synoptic-scale cold air masses, high pressure, wind patterns, etc.) largely determine the long-term lake climate conditions. However, since lakes tend to have the most influence during short, intense episodes of favorable conditions (on the order of hours to days, as discussed above), we must consider factors determining which locations will receive the biggest lake modifications in those short episodes. For example, wind patterns that develop over lakes through a combination of large-scale atmospheric features and local lake-induced wind circulations, determine whether a lake-effect snow storm will occur over a wide area of the downwind shore or in a single community. Such wind circulation patterns are called “mesoscale lake-effect” wind fields (tens to hundreds of kilometers) that organize the lake-effect snow into “mesoscale bands” (Figure 2).

Many factors influence the mesoscale lake-effect wind fields and the resulting lake-effect snow bands: lake–air temperature differences, wind direction relative to the orientation of the long axis of the lake, magnitude of the upward motions (convection) over the lake, atmospheric stability, wind changes with height (wind shear), and so on. A considerable amount of research has been conducted, and continues to be devoted to understanding what combinations of lake, land, and atmospheric characteristics result in different types of mesoscale wind fields. Some of the most fundamental controlling mechanisms of mesoscale wind patterns fall into two major categories: (i) surface frictional and topographic forcing, and (ii) thermal and isallobaric forcing. The former mechanism is important because wind direction and speed near the surface are influenced by the magnitude of “surface friction” (the amount of roughness of the surface due to such features as hills, vegetation, structures, etc.). Over land areas, obstacles make the surface friction large, which causes the winds to slow and move toward areas of atmospheric low pressure. Over lake areas, the wind is smoother, resulting in higher wind speeds with directions oriented between high- and low-pressure areas. Locations where the air moves between high and low surface friction can exhibit complex wind patterns, greatly influencing the local weather.

The second mechanism, thermal and isallobaric forcing, describes the atmospheric response

![Figure 2](image1.png) Terra MODIS visible wavelength satellite image of the Great Lakes at 16:25 UTC on February 13, 2016. Reproduced from NASA.
to the development of regions of warm air (due to heat fluxes from the lake to the air) and nearby cold air (which has not been modified by the lake). The areas of warm air exhibit decreasing atmospheric pressure, while the areas of cold, unmodified air maintain higher air pressure. These pressure differences, in turn, further change the mesoscale wind patterns. The balance of the two mechanisms discussed above, modified by several other factors (such as lake shape, wind direction and shear, lake–air temperature differences, ambient atmospheric stability and humidity, determine the mesoscale lake-effect wind patterns and the types of lake-effect snow bands that develop.

Many types of mesoscale lake-effect snow bands have been identified. Short lake-axis parallel (SLAP) bands, also called broad area coverage or widespread cloud streets, tend to occur when the mesoscale winds flow across the short axis of a lake (e.g., Kristovich 1993). Lake-effect snow bands that are oriented along the long axis of a lake, recently called long lake-axis parallel (LLAP) bands, have previously been called shoreline parallel bands and mid-lake bands. Lake-effect bands can also exhibit curving patterns, sometimes resulting in “vortices” of different sizes (e.g., Hjelmfelt 1990). Finally, in areas with multiple lakes (such as the Great Lakes region), lake-effect bands extending from one lake to another (lake-to-lake, L2L, bands) are also recognized as an important type of lake-effect mesoscale system.

Mesoscale lake-effect convective structures on smaller lakes have received much less attention, but have been identified for mid-latitude lakes as small as 100 km² in surface area or even less. Studies of lake-effect convection over the Great Salt Lake in Utah, Lake Champlain on the border of Vermont and New Hampshire, Lake Tahoe on the border of Nevada and California, and the New York Finger Lakes, indicate that these lakes can produce snow multiple times each year (cf. Laird, Sobash, and Hodas 2009; Steenburgh, Halvorson, and Onton 2000). Compared to lake-effect storms over the Great Lakes, small lake storms tend to occur less frequently, over a lesser range of environmental conditions (especially lake-air temperature differences and low-level wind direction), and have fewer mesoscale band types.

The combined influence of all the mesoscale band types determines the overall influence of lakes on the local climate. While long-term trends in most aspects of lake climates have not been investigated, there is a general consensus that lake-effect snowfall increased during much of the twentieth century. Recent studies have identified a slowing of the lake-effect snowfall increases with time or even a reversal to snowfall decreases in some of the snowbelt regions (i.e., Bard and Kristovich 2012). There is disagreement in the literature on the causes of these snowfall changes, and more research is needed.

Warm, stable season lake climates (spring and summer)

During the warm season, the influences of lakes on the local climate tend to be weaker than those during the winter, largely because the relatively cool lake waters only influence a relatively shallow layer of the atmosphere. In addition, since the lower atmosphere is cooled, the air resists rising vertically (“stable” conditions), further limiting the depth of the lake influences. Scott and Huff (1997) found that precipitation near the Great Lakes was, on average, 25–50 mm less than in areas far from the lakes, even though there was only a slight decrease in cloud cover. They found a weak tendency for decreased vapor pressures over and near the lakes. This might have been
LAKE CLIMATES

anticipated, since during warm, humid times of the year, water vapor in the atmosphere would tend to condense onto the lake surface much of the time.

The impacts of lakes on local climate conditions in the warm season can be more complex than during typical cool seasons. For many mid-latitude lakes, water temperatures usually only vary by a few degrees over the course of a day, while air temperatures can be more than 20°C greater during the afternoon than in the morning. The result is that, frequently, lake temperatures are cooler than the air temperature during most of the daytime hours, but can be warmer than the air for portions of the night. Indeed, Scott and Huff (1997) found that minimum temperatures tended to be slightly warmer over and near each of the Great Lakes by up to 3°C. However, considerably cooler maximum temperatures (up to 4–6°C over and near Lakes Superior and Michigan) were observed.

The change in the lake-air temperature differences between night (positive differences) and day (negative differences) gives rise to a unique oscillating pattern in the mesoscale wind patterns (e.g., Eichenlaub 1979). During the day, the warmer land surfaces cause the air to rapidly heat, resulting in a deeper planetary boundary layer. Because this expansion does not occur over the lake, winds between about 200 and 2000 m above the surface flow from land toward the lake. To adjust for this higher-level wind flow, winds near the surface move from lake toward land — the well-known “lake breeze.” The overall circulation provides an area of sinking, cloudless air over the lake and rising air with clouds over the land. Figure 3 shows an example profile of temperature, humidity, and winds at the shore in afternoon lake breeze conditions. During the overnight hours, if the lake becomes warmer than the air over land, the opposite circulation can develop — the “land breeze” circulation.

Lake breezes can play instrumental roles in relieving summertime heat waves in coastal urban areas such as Cleveland, OH, and Chicago, IL, but can also contribute to regional transport of pollutants, influencing air quality. Even though the afternoon lake breeze tends to decrease precipitation near the shoreline, the front edge of lake breezes can often be a location of thunderstorm events (Figure 4). In areas where lake breezes from multiple lakes tend to collide, frequent severe summer storms have been observed. Considerable research activities have focused on understanding how lake breezes move through urban areas, are affected by local topographic features, and interact with other types of mesoscale circulation patterns. For example, the Lake Okeechobee lake breeze in southern Florida frequently interacts with sea breeze fronts from the Gulf of Mexico and the Atlantic Ocean, thunderstorm outflow boundaries, and occasional fronts.
Relative to winter lake-effect snows, little work has been carried out to understand long-term trends in lake breeze occurrence. One of the reasons for this is the difficulty in identifying lake breezes in climatological surface datasets. Multiple ways of identifying when lake breezes were likely or have occurred have been developed to examine their climatological characteristics (e.g., Laird et al. 2001). Methods such as these will be necessary to understand the influences of long-term climate changes on warm-season lake climates.

Comments and knowledge gaps

Much is known about large-lake modification of the climate conditions in their vicinity throughout the year, particularly in locations typically downwind. Over time, communities and ecological conditions adapt to the usual conditions (such as average conditions and “usual” extreme conditions). The largest impacts are often experienced when climate conditions are extreme and outside those previously experienced in socially or ecologically relevant time frames.

Many fewer research papers have been published on the varying nature of lake climates, particularly during the warm, stable season. Additional field studies and numerical modeling studies are needed to understand interactions between severe thunderstorms and over-lake stable boundary layers, for example. Even in better-understood cool, unstable seasons, predictability of interannual variations is limited due to the complex interrelationships between large-scale oceanic temperature anomaly patterns, synoptic airflow configurations, lake surface (such as ice cover) and vertical profile conditions, and periodic air-mass conditions. For example, the question arises as to whether reported long-term variations in winter lake-effect snowfall are experienced as more extreme storms or as more frequent weaker storms.

Research continues on the ecological responses to climate change in many lakes, but little work has been reported on how biological processes, in turn, influence local climate. A critical area of continued research is how climate changes are manifested in changing extreme events and how, in turn, these events impact the lakes themselves and their surroundings.

Disclaimer: The views expressed in this entry are those of the author and may not reflect the opinions of the Illinois State Water Survey, Prairie Research Institute, University of Illinois, or the funding agencies.

SEE ALSO: Climatology; Lake ice; Lakes and limnology; Snow

References

LAKE CLIMATES


Lake ice

Étienne Boucher  
*Université du Québec à Montréal, Canada*

Mickael Lemay  
*Université Laval, Canada*

Yves Bégin  
*Institut National de la Recherche Scientifique, Canada*

**Definition**

In lakes, an ice flood occurs when ice and flood water overflows lakeshores and comes in contact with the riparian environment (Figure 1a). Such events can leave traces in riparian ecosystems when water levels, covered by drifting ice rafts, reach the upper part of the shore. The phenomenon does not occur each year. Ice floods generally occur at spring breakup, when the ice cover is lifted by rapidly rising water levels, due to a massive water input during snowmelt in the lake’s drainage basin. Ice rafts are pushed against the shore by wind and pile up on the shore (ice piling) and can induce severe mechanical erosion. Sedimentary levees (ridges) of coarse, ice-pushed sediments may form in the most exposed lakeshores. Such ridges are usually maintained by frequent and powerful ice push action. Disturbed trees growing on the shores where frequent ice push occurs may bear scars that are indications of past ice push activities. Such scars can be dated through standard tree ring analysis methods, and used to reconstruct the magnitude and frequency of past ice push events.

Drift ice season

The drift ice season can be divided into three phases. The first phase corresponds to the ice cover formation (ice buildup). Ice covers usually start forming during frigid weather conditions, most frequently at night and over calm waters protected from winds and waves. The thin, newly formed ice is relatively fragile and may disappear rapidly if temperatures rise above the freezing point, rainfall occurs, or even if the water surface becomes more agitated. If the newly formed thin ice cover persists over a few days, it begins to grow vertically towards deeper water layers, and horizontally towards more agitated and exposed water layers. Black ice forms when crystallization occurs in calm waters, and by opposition, white ice forms in more agitated, aerated conditions (Ashton 1986). Thermal accretion of the ice cover towards its base continues until water heat losses to the atmosphere become negligible. The distribution of heat in the water column is also an important aspect. Water’s maximal density is reached at precisely 3.98°C. Thus, when water cools down from the surface to the bottom at the beginning of winter, surface water become denser, less buoyant, and precipitate. This process therefore forces deeper, yet warmer water to become buoyant and mechanically climb to the surface while mixing with less dense waters. This mixing process occurs until water’s density falls below 3.98°C. Water’s density decreases and cold surface waters rapidly reach the freezing point, which accelerate the freeze-up. The phenomenon ends when water stops transgressing the freezing point, most often when the ice cover (but also the snow mantle) becomes thick enough to prevent water heat losses to the atmosphere.

*The International Encyclopedia of Geography.*
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0796
Figure 1  Ice flood processes, landforms, and dendrochronological markers. All pictures from Clearwater Lake, northern Quebec, Canada. Photos by Yves Bégin. (a) An example of an ice flood pushing against the upper shore of Clearwater Lake and damaging the vegetation. Note the presence of huge rocks carried by the ice during the process. (b) Ice cover degrades in early spring and breaks into large rafts that can be carried against the shore by the wind. (c) An ice push ridge (see person standing for scale). (d) Ice rafts pushed against riparian trees. (e) Multiple ice scars found on a single individual tree exposed to ice push disturbance.
When the ice covers the full water surface, a second phase begins: the ice pack period. The ice stops growing vertically and horizontally, and is immobile. The ice cover protects the low shore from the erosive actions of waves. In rare cases, cracks can form in the ice cover which allow for the penetration of meltwater (or rain) into the ice, which will rapidly freeze up. In such conditions, lateral movements of the ice cover become possible and, although they are limited in amplitude, the slight, marginal movements can damage the shore. However, such fractionation events can only occur if water levels fluctuate during the ice cover season following, for example, winter showers or winter mild spells. For this reason, the phenomenon is quite rare in dry and frigid climates. Nevertheless, since ice contraction and extension are a temperature-sensitive phenomenon, it is sometimes possible to observe some advance and encroachment of the ice pack on the lakeshore.

During the ice pack season, the ice behaves as an emerged surface capturing the falling snow. The amount of snow that the ice cover can support varies according to the climate (mean temperature and variability, amount and type of precipitation) and exposure to winds. Temperature controls the crystalline state of the snow. In climates where the temperature is maintained at below approximately \(-12\, ^\circ\mathrm{C}\), snow crystals are unaltered and are found at the particulate state, while under warmer temperatures, snow crystals fuse with one another and snow pack becomes cohesive (Michel and Ramseier 1971). Individual snow crystals may be mobilized by wind, which allows the redistribution of the snow cover over the lake. The size and shape of the lake, with respect to wind orientation, have important implications for snow mobility at the surface of the ice cover. In fact, the longer the fetch (the length to which winds can blow without encountering obstacles), the greater is the eolian force on the snow pack. A small lake bordered by a dense forest cover often represents a snow pit; a large lake exposed to wind gusts will capture snow which will ultimately be transported against island margins or against the shores. Thermal insulation of the ice by snow is proportional to the thickness of the snow accumulation. This will influence the rate of accretion of the ice, especially at the beginning of the ice season. Hence, small lakes that are rapidly covered by a thick snow mantle often exhibit a thinner ice cover.

In frigid climates with cold and dry winters, the characteristics of the ice cover breakup have an important influence on hydrological and geomorphological processes. An extended cold season is crucial in order to maintain a thick, dense, and coherent ice cover that will ultimately induce some severe mechanical effects on riparian ecosystems when it is mobilized. The lateral movements of the ice cover usually start at the time when the snow pack begins to melt. The thicker the snow pack is, the faster the melt will be, inducing a rapid rise in water levels. The hydrographic position of the lake in the watershed, the watershed’s shape and slope, and the efficiency of the lake’s outlet(s) are all important factors that control the amplitude of spring flood fluctuations.

The mechanism by which ice transports sediments and disturbs riparian ecosystems is also linked to wind action. Essentially, spring flood augments the open-water area at the margin of the ice cover (Figure 1b). The ice pack also gets thinner and reduces in size, starting near the shoreline. The combined effect of these two processes leads to the creation of an open-water fringe bordering the ice cover. Under the persistent force of wind, ice drifts and impinges on the islands and the peripheral shorelines. This results in various erosional and depositional phenomena that shape ice-affected shorelines.
LAKE ICE

Among other things, ridges (Figure 1c) represent alluvial levees recognizable from their linear, yet semichaotic, microtopography. The ice push therefore occurs during a very short period of time, at the moment of breakup. When ice breakup is advanced (mature), the ice becomes rotten, too friable to pile up, and thermally degraded, and the ice flood season comes to an end. It is inevitable that an ice flood can only occur when winds can mobilize the ice cover.

Latitudinal variations in ice flood properties

Winter and spring climatic conditions have a determining effect on ice flood magnitude and frequency. This activity will culminate at mid-latitudes, between temperate–cold climates and arctic climates. Under temperate–cold climates characteristic of the boreal zone, snow is abundant but stays for a very short period at the particulate state. It is generally sticky when it falls (temperatures often over 1–2°C), or becomes sticky in the following days when temperatures become milder. In spring, the snow pack melts in place, forming a thick water layer over the ice cover (Michel and Ramseier 1971). This ponding process is often mistaken for the “sinking” of the ice below the lake’s water surface. But this phenomenon would contradict the fundamental laws of physics, because ice has a lower density and cannot sink in fresh water. In arctic and subarctic environments, this phenomenon is rare, if not nonexistent, simply because less snow accumulates due to powerful winds and cold winter temperatures that maintain the snow at the particulate state. In such conditions, snow is blown away by wind and does not accumulate on the ice cover. Ice will therefore degrade either by thermal processes (under the control of rising temperatures, solar radiation, or heat transfers from the water to the ice) or by mechanical processes, after the agitation of water on the lake.

Lake ice floods’ magnitude and frequency vary with latitude, along with the corresponding impact on riparian ecology and geomorphology. Conditions that favor the most frequent and intense ice flooding activity are squeezed between the cold arctic climate and the more temperate boreal climate. Ice floods are therefore a typical subarctic phenomenon. A similar ice flooding activity could potentially be found in high-altitude environments; however, large water bodies are rarely found in mountainous relief.

By comparison to the geographical extent of river ice floods, lake ice floods are much more restricted spatially. River ice floods (see River ice and ice jams) result from a combination of factors that are found in a much broader range of conditions. Such events are caused by the blocking of the channel by the ice (ice jam), a situation that can be caused by an abnormally thick and resistant ice cover and/or by an early spring flood that occurs before the ice cover has thermally degraded. These conditions are commonly found in areas where air temperatures drop below 0°C for at least one month and can be aggravated by various channel properties that restrict the mobility of drifting ice rafts, for example, a sinuous or constricted channel. By contrast, lake ice floods are much more zonal (i.e., predominantly subarctic, but occasionally high-boreal) and require a precise combination of factors including (i) an important spring flood to lift and detach the ice cover from the lake’s bottom and shore, (ii) little snow accumulation to reduce the thermal insulation and spring ponding effects, (iii) a long fetch that will generate sufficient drag force to push the ice against the lakeshore, and (iv) an ice cover that is thick and mechanically strong enough to persist during the spring flood.
Lake ice flood reconstruction

Since most important lake ice floods occur during high water levels, it becomes possible to link spring flood magnitude to the abundance of ice-related damage left on the lakeshore (Bégin 2001). Most of the literature that aimed at reconstructing past lacustrine ice floods also aimed at investigating past water levels in lakes where local hydrological fluctuations have some regional significance (e.g., Bégin and Payette 1988). Reconstructing spring water levels allows the records to be extended beyond the observed (instrumental) variability in order to better appreciate interannual variability and long-term trends in regional hydrology. To do so, dendrochronology is often used to construct chronologies of past ice flood events and make attempts at dating damages left by ice during ice floods on riparian trees.

Most common ice flood dendrochronological markers are scars developed by trees on damaged portions of the trunk (Figure 1d–e). Such scars form when ice impacts against rigid and inflexible riparian stems and results in the partial destruction of the cambium (a thin envelope of dividing cells), a situation that hampers lateral growth. At the periphery of the injury, cambium cells are found to be intact, and lateral growth is uninterrupted. Wood cells form each year and are typically organized as tree rings that can be counted to date the injury. Using this technique, ice flood history can be reconstructed and the magnitude and frequency of past events are to be interpreted as a function of the abundance and height of the damages left on trees, according to statistically verified procedures.

In order to represent a useful marker of past ice floods, four groups of criteria need to be considered: the lake’s configuration, the shore’s exposure to winds, the forest structure (age, density, species), and the position of the trees on the shore. Headwater lakes are not related to large rivers and their water level fluctuations are closely related to precipitations, while lakes located downstream in a watershed collect water from higher-order basins and integrate the complex combination of variables that control the shape of the hydrograph. Circular lakes are also ideal for ice scouring phenomena to take place, because the fetch is long in all wind directions, therefore resulting in an equivalent probability of scouring (comparable exposure) for all shores. Steep slope shores are affected by the most intense ice push events. On such shores, it is not uncommon to observe ice pile-ups and ice push ridges. This situation is also true for highly exposed islands and capes jutting into the water. On these exposed shore segments, trees are not frequent. By contrast, lakeshores presenting a gentle subaquatic slope are less exposed, and ice impinges against the pre-littoral zone and rarely reaches forested ecosystems found on the upper shore. Only in rare extreme events will these ecosystems be disturbed by lake ice. In order to reconstruct the history of past lake ice floods, it is recommended to focus on the tree ring analysis of moderate severity, where the forest line comes into contact with water. Forest structure also influences the recording. The absence of trees does not imply the absence of a spring flood, but the probability of recording an event increases with the number of trees available for recording. It is therefore necessary to weight the number of ice scars from a given year against the number of trees available for recording that same year. It must finally be underlined that riparian tree populations growing along subarctic lakes can be disturbed by other factors, including forest fires and insect outbreaks. Finally, the position of trees on the lake shore is an important parameter to consider in the analysis, as trees growing close to the mean high water level are the most exposed to ice conditions.
scouring. The analysis of past ice scouring events and relationships with past lake water levels need to be analyzed while taking into account all these parameters.

SEE ALSO: Lakes and limnology; River ice and ice jams

References


Lakes and limnology

Sally P. Horn
University of Tennessee, USA

Lakes are inland bodies of relatively still water that occupy depressions formed by a variety of natural geomorphic and biological processes, or by human activity. They are the focus of the fields of limnology, which is the study of the physical, chemical, and biological characteristics of all inland waters, including fresh and saline lakes, streams, and wetlands, and of paleolimnology, which investigates lake sediments as archives of information on past climates and environments. Although lakes contain less than 0.5% of all fresh water on Earth, they are significant for human wellbeing as sources of water, food, and economic and recreational activities, and as avenues for water transportation (Wetzel 1975). Lakes in various parts of the world are considered sacred by modern human societies, and archaeological evidence indicates that they also held special significance for prehistoric peoples. They contribute to biological diversity, and may serve as sentinels of future climate and ecosystem change, just as they also provide valuable evidence, in their accumulated bottom sediment, of past changes in climates and environments resulting from natural processes and anthropogenic activities. Geographers have contributed to the study of modern lakes, lake sediments, and human interactions with lakes using both traditional, “muddy boots” fieldwork and new geospatial techniques.

Lakes and their study

Freshwater lakes range in surface area from less than a hectare to tens of thousands of square kilometers, in the case of the North American Great Lakes and the large lakes of East Africa, western Asia, and Canada. Lake Superior has the greatest area (83,300 km²) of any freshwater lake in the world, while Lake Baikal has the greatest depth and volume (1741 m and 23,000 km³) (Horne and Goldman 1994). The Caspian Sea, more than four times the size of Lake Superior, is the world’s largest inland water body, but this remnant of the ancient Paratethys Sea is saline (Cohen 2003). Small, shallow lakes are often termed ponds, but scientific definitions vary, and the use of “lake” or “pond” in naming water bodies is inconsistent.

Natural lakes and ponds of all sizes owe their origin primarily to geomorphological processes. Glaciation, volcanism, tectonics, fluvial dynamics, landslides, wind, changes in sea level and coastal shorelines, and the weathering of carbonate rocks create depressions by scouring, explosion, crustal extension, deflation, or collapse; by damming stream valleys with sediment or lava; and through faulting (Figures 1 and 2). Plant growth may be sufficient to dam some depressions (Wetzel 1975), and beavers create ponds by damming streams. Adding to the Earth’s natural lakes, humans have created large numbers of lakes by damming streams for hydropower and water management, and by creating farm ponds and other impoundments. Lake origin determines morphology, which
along with climate influences lake stratification and ecology. In his useful review of mechanisms of lake formation, Cohen (2003) made the important point that lake basins have histories that are partly independent of the lakes that fill them, and that basin evolution also affects both limnology and the nature of sediments available for paleolimnological study.

Using new statistical approaches, Downing et al. (2006) recently estimated that there are 304 million natural ponds and lakes over 0.1 ha (0.001 km²) in the world, covering an estimated 4,200,000 km². Small lakes dominate, with nearly 91% of these predicted lakes less than 1 ha in size, and a further 8% between 1 and 10 ha. Reservoirs and other artificial impoundments cover 260,000 km² of Earth’s surface, and farm ponds cover 77,000 km². Adding these water bodies to natural lakes leads to the estimate that 3% of the Earth’s surface is covered by lakes, ponds, and impoundments – an estimate more than twice as high as previous inventories, which underestimated small lakes.

In an earlier paper on lake distributions, Meybeck (1995) noted that the global distribution of lakes was addressed by geographers beginning in the eighteenth and nineteenth centuries, but that this information was not of interest beyond geography until much later. Only after the earth and biological sciences had evolved to a global perspective, a transformation Meybeck linked with the establishment of the International Geosphere-Biosphere Programme (IGBP) in 1987, did lakes and wetlands get due credit in consideration of regional and global water balances and biogeochemical cycles. Lakes are now seen as major regulators of carbon, nitrogen, and phosphorous cycles through the production of organic matter, sedimentation of organic detritus, and precipitation of carbonates and evaporites (Meybeck 1995) – and so their geography (distribution, sizes, volumes) now matters.

The latest estimates of global lake distribution by Downing et al. (2006) suggest that rates of material processing by aquatic ecosystems may be twice previous estimates, and that processes most active in small lakes and ponds may be of previously unrecognized global importance.

Lakes are significant sources of biological diversity, supporting terrestrial as well as aquatic biota, including endemic species of fish, sponges, bacteria, zooplankton, phytoplankton, and other algae that have evolved in particular lakes or lake systems. Differences in lake depth, water temperature, chemistry, and light create different environments within and between lakes, and over the seasons, affecting the geographic distribution of aquatic biota. Advances in remote sensing now enable remote monitoring of some physical, chemical, and biological parameters, at least for larger lakes. But other measurements, and sample collection, must be done on-site. Limnological field sampling typically involves the use of small boats or inflatable rafts to reach sampling sites, together with various devices to
measure lake transparency, water color, depths, temperature, oxygen, conductivity, and pH; samplers to recover water at different depths for chemical analysis and for quantification of phytoplankton in the water column; nets for zooplankton; and dredges for surface sediment (Figures 3 and 4). Advances since the early 2000s include data loggers that can be left in lakes to monitor temperature or dissolved oxygen until a return to collect them, or set up to transmit data to users. Self-propelled floating devices combining Global Positioning System (GPS) receivers with sonar and other sensors and probes are coming online that can speed limnological study; current versions are operated by radio control from shoreline to map lake bathymetry, measure temperature and selected chemical properties, and collect water samples. Limnological researchers of the future might also plan to use automated underwater vehicles to collect data and samples, or drones that collect data from above the lake surface.

The value of lakes in providing habitat and resources for aquatic and terrestrial wildlife, and various other benefits to the natural world and human society, are increasingly discussed in terms of *ecosystem services*, a concept that gained wide exposure in the United Nations
Figure 3 A Secchi disk used to measure lake transparency. The depth at which the disk disappears from sight when lowered below the water surface is the Secchi depth. Algal blooms and suspended sediment can change Secchi depths seasonally or over longer time periods, making this inexpensive device an important monitoring tool.

Millennium Assessment (Millennium Ecosystem Assessment 2005). Lakes and lake processes contribute ecosystem services across all four categories defined in this assessment, namely: provisioning services (water and food); regulating services (biogeochemical cycles, as described earlier, and in other ways); supporting services (nutrient cycling and aquatic photosynthesis); and cultural services (recreation, aesthetic, and spiritual benefits). The long-term viability of ecosystem services provided by lakes is a cause for concern, however, given the rate at which natural lakes are being altered or destroyed by human activities.

The increased recognition of the regional and global importance of lakes has spurred limnological research, which now exists for all continents and most countries of the world. Most of the Earth’s large lakes have been studied, but gaps

Figure 4 Temperature measurements made by lowering a probe through the water column reveal whether lakes are stratified, with a warmer, upper body of water separated from a cooler, lower body of water. This temperature profile from Lago Cote, Costa Rica, made on March 9, 1998, shows the lake to be stratified at 5 m. Repeat measurements reveal whether lakes maintain temperature stratification for long periods, or frequently overturn, mixing the water body. Lakes that mix frequently are known as polymictic lakes.
exist in knowledge of other lakes and in particular regions. The latest estimates of global lake distribution suggest that small lakes, in general, need additional attention, to quantify processes whose significance at regional and global scales has likely been overlooked. Limnologists have also called for additional monitoring of lakes as sentinels that can help to reveal effects of current climate change at the ecosystem scale in understudied areas, and to disentangle the effects of climate change and other influences on the environment (Williamson et al. 2009).

Growing interest in the reconstruction of past climates and environments has also driven research on modern lakes and on the records of past conditions preserved in lake sediments. Geographers have contributed to paleolimnological research at lakes throughout the world, recovering sediment cores (Figure 5) and studying plant and animal remains, stable isotope signatures, and other environmental indicators, or proxies, preserved within them. Studies of aquatic microfossils in sediment cores, such as diatoms and nonbiting midges or chironomids, have motivated regional limnological surveys to determine the relationships between lake parameters measured today and microfossils in modern sediments, information needed for interpreting past conditions. Much of what we know about the history of climate, vegetation, fire history, and human land use over the Late Glacial period and Holocene has come from the study of lake sediments. The pollen, diatom, chironomid, charcoal, stable isotope, and other proxy records that paleolimnologists have developed provide key information for understanding the development of aquatic and terrestrial ecosystems; interactions between climate, vegetation, and fire; and patterns of prehistoric land use. These records provide data for paleoclimate models and perspectives for modern ecosystem management, and can help society understand, anticipate, and perhaps mitigate the consequences of future global change.

History of research in limnology and paleolimnology

The field of limnology was established by the Swiss physician and scientist François Alphonse Forel (1841–1912) with the publication of his three-part monograph on Lake Geneva (in French Lac Lemán). The contents of his Le Lemán: Monographie Limnologique (Lake Geneva: Limnological Monograph) illustrate the interdisciplinary nature of the new science and the overlap with the subject matter of physical and human geography. Volume 1 (published in 1892) covered geography, hydrography, geology, climatology, and hydrology; volume 2 (1894) focused on hydraulic, chemical, thermic, optical, and acoustical processes and patterns; and volume 3 (1904) focused on biology, history, navigation, and fisheries. Forel coined the term
“limnology,” which at the time referred only to the study of lakes, and published the first textbook in 1901, *Handbuch der Seenkunde: Allgemeine Limnologie* (Handbook of Lake Study: General Limnology).

Though considered the father of limnology, Forel was not the first to systematically sample and describe lake characteristics or biota (Horne and Goldman 1994). Two centuries earlier, Antonie van Leeuwenhoek initiated studies in biological limnology with descriptions of the green algae *Spirogyra* from a shallow lake in the Netherlands. Important contributions to physical limnology were made in the 1700s with observations of periodic surface waves, measurements and interpretations of temperature stratification in lakes, and exploration of how light, heat, water temperature, and wind interact to structure water bodies. The discovery of zooplankton, and initial work on estuarine salt wedges and river pollution (all in western Europe) also predated Forel’s work, as did the earliest research on lakes in the United States. In 1850 Louis Agassiz, then three years into a professorship at Harvard, published *Lake Superior: Its Physical Character, Vegetation, and Animals*, which combined a travel narrative with observations on the lake (especially its fish) and its regional geography (Agassiz 1850). In 1887 Stephen Forbes presented a landmark paper before the Peoria, Illinois, Scientific Association. Entitled “The Lake as a Microcosm,” Forbes’s paper was the first to describe lakes as ecosystems, a key concept of modern limnology.

Early limnology shared techniques with oceanography, which was also developing in Europe during the latter half of the nineteenth century. In the 1860s the Italian astronomer and priest Pietro Angelo Secchi experimented with disks of various size, color, and material to measure the transparency of coastal waters of the Mediterranean. The metal or plastic disk that now bears his name also became a key tool for measuring the transparency of lake waters, used to this day by limnologists and in a global citizen science monitoring program developed by biologists and a geographer at Kent State University (www.secchidipin.org). Other traditional tools for lake study, including encased thermometers for measuring water temperatures at depth, silk nets for sampling plankton, and the Forel–Ule scale for gauging water transparency and color came into use in the latter half of the nineteenth century.

This period also saw exploration of the great lakes of Africa, supported in part by the Royal Geographic Society: Lake Malawi by David Livingstone in 1850; and Lakes Tanganyika, Victoria, Albert, Edward, and Turkana by others between 1858 and 1888. European and American explorers had reached the large lakes of North America earlier: Great Slave Lake in Canada in the late 1700s, and Great Salt Lake in 1824. The large lakes of Asia were known early by Chinese and Russian scientists, and the Caspian Sea was described by Greek historians (Meybeck 1995).

The first scientific society in limnology was the International Society of Limnology, established in Europe in 1922 as the Societas Internationalis Limnologiae (providing the enduring acronym SIL). The Limnological Society of America was founded in the United States in 1936, and reorganized to include oceanography and marine biology in 1948, when it was renamed the American Society of Limnology and Oceanography, or ASLO. In 2011 the society changed its name to the Association for the Science of Limnology and Oceanography to reflect its increasingly international membership (the acronym was conserved). SIL and ASLO each have several thousand members. SIL holds international congresses every three years and publishes two journals, *Fundamental and Applied*
Researchers who study paleolimnology attend limnology meetings but also have their own societies – the International Limnogeology Association (established 1993) and the International Paleolimnology Association (established 2003) – and journal, the highly ranked Journal of Paleolimnology, which launched in 1988 (2014 impact factor 2.1). Paleolimnologists are also active in organizations focused on the Quaternary period, such as the International Union for Quaternary Research, and national groups such as the American and Canadian Quaternary Associations; and on particular proxies, such as the International Society for Diatom Research, and AASP – The Palynological Society. Paleolimnologists in the United States also belong to the Association of American Geographers, which has a specialty group devoted to paleoenvironmental change; the Geological Society of America (Limnogeology Division); and the Ecological Society of America (Paleoecology Section), among other disciplinary societies.

A concise definition of paleolimnology, used by the Journal of Paleolimnology, is the reconstruction of lake histories. This definition suggests a focus on conditions and processes only within lakes, such as changes in algal productivity and community composition as indicated by diatoms in sediment cores. Early uses of the term were in this more restricted sense, for example by the ecologist E.S. Deevey in his 1953 book chapter on “Paleolimnology and Climate.” However, with the explosion of interest in using lake sediments to reconstruct past climates and environments, paleolimnology has come to include all studies that make use of these lacustrine archives – even those focused on proxies that derive from, and inform us about, processes outside of the lake itself, such as the pollen of terrestrial plants and charcoal fragments from watershed or regional fires.

Thus, although related to limnology, paleolimnology has a broader parentage, with stronger contributions from geology, physical geography, and terrestrial ecology and botany. As with limnology, the field of inquiry predates its name (which Deevey in 1953 may have been the first to use). Geologists were identifying and interpreting old lake deposits in the early to mid-nineteenth century, although relationships between Pleistocene lake sediments and climate did not become clear until the end of the century (Cohen 2003), and as late as 1953 Deevey called for better integration of geological and ecological approaches. Investigations of diatoms and pollen grains in sediments began in the early twentieth century, coincident with improvements in microscopes. Studies of pollen in Pleistocene and Holocene lake and bog sediments spread rapidly in northern Europe and the United States in the 1920s, 1930s, and 1940s. Most early pollen studies focused on the effects of climate change on vegetation, and early work on diatom communities focused on understanding the changing trophic states of lakes. In the second half of the twentieth century, paleolimnological studies began to address human impacts on lakes and
LAKES AND LIMNOLOGY

their watersheds more often, including prehistoric impacts. Diatom analysis expanded greatly in the 1980s as the technique came into use as a means to identify the impacts of acid deposition on lakes. The establishment of the Journal of Paleolimnology and professional societies and conferences in paleolimnology since the 1980s reflects the further growth and vibrancy of the field.

SEE ALSO: Biogeography; Climate change and biogeography; Environmental history; Geoarchaeology; Global climate change; Holocene; Paleoclimatology; Paleoecology

References


Land change science

B.L. Turner II
Arizona State University, USA

Land change science, also known as land system science, is an interdisciplinary subfield of study addressing the causes and consequences of changes in land use and land cover, especially as related to the themes and topics and explanatory lenses consistent with the global environmental change and sustainability sciences (Turner, Lambin and Reenberg 2007). Its historical and intellectual antecedents can be found in geography, perhaps more so than any other field of study, and professional geographers make up a substantial proportion of the practitioners in the subfield. The development of land change science as an international research endeavor, however, resides in the contemporary and interdisciplinary research programs addressing changes in the Earth system, landscapes, and ecosystems, and in sustainability and sustainable development. Land change science now spans a full array of environmental, social, spatial, and remote-sensing sciences, complete with new journals and journal sections devoted to various aspects of land change.

Formal programs

Land change science formed as an international and interdisciplinary subfield of study with the development in the early 1990s of the Land Use/Cover Change (LUCC) program of the International Geosphere–Biosphere Programme (IGBP of the International Council of Science) and the International Human Dimensions Programme (IHDP of the International Social Science Council), constituting the first joint research program of these two organizations. The IGBP, foremost its ecological dimensions, realized that understanding and forecasting the biophysical changes in the terrestrial surface of the Earth was inadequate, lacking a human or social dimension addressing land-use dynamics, and that the observed changes required long-term monitoring at a global scale. The IHDP was seeking ways to develop integrated human–environment research within the larger global environmental change agenda and welcomed a linkage to the IGBP. LUCC was envisioned as a truly integrated research effort joining the various disciplines interested in the observation, monitoring, understanding, modeling, and projecting of land change from the local to the global spatial scales. With a few important exceptions, however, the effort captured “tweeners” – those whose research is situated at the intersection of social and environmental sciences – interested in land systems, more so than disciplinary specialists interested in some facet of land change related to both its social and its environmental subsystems. Subsequently, a more expansive Global Land Project (GLP) emerged from the IGBP–IHDP, one that sought to be integrative in terms of the human and environmental dimensions and interactions of land systems, drawing on researchers from the social, environmental, spatial, and remote-sensing sciences. The level of intended integration was abbreviated, however, by the development of another international science program, DIVERSITES, focused on biodiversity and ecosystem
LAND CHANGE SCIENCE

change and the consequences for ecosystem services and human wellbeing. As result, the land systems and ecological landscape communities, represented by the GLP and DIVERSITAS, respectively, tended to work in a complementary, albeit informal, way. The newly minted Future Earth program of the International Council of Science blends both programs through a set of sustainability themes that require not only the kinds of knowledge gleaned from the two programs but also a union of them framed into “actionable” or “use-inspired” basic science.

Geographic legacy

The antecedents to the contemporary discipline of geography had a rich tradition of interests in human–environment relationships. Influenced by the works of Alexander von Humboldt, the modern geographic foci addressing land systems may be situated with the concept of *Landschaft* (landscape) in nineteenth-century German geography (Turner 2002). The various definitions of the concept notwithstanding, influential proponents championed a geography based in the systematic sciences in which the phenomenon of study was the landscape, a human–environmental or social-environmental system in contemporary parlance. This orientation, not necessarily linked to land systems, rose to prominence in the late nineteenth and early twentieth centuries under the label of “the geographic factor” – the environmental determinants of humankind. Championed by a number of prominent scholars, especially in elite, private institutions in the United States, this orientation morphed into environmental determinism, the negative repercussions of which scarred human–environment research in the discipline. As a result, a more reasoned human–environment approach – human responses to environmental conditions – championed by Harlan H. Barrows in the 1920s – was derailed until the latter half of the twentieth century, when it re-emerged as natural hazards or risk hazards research. Advocated by Gilbert F. White and his students, this human–environment subfield focused on socially defined problems with an “actionable science” orientation. An explicit landscape approach, however, commonly referred to as landscape morphology, had been established in the late 1920s by Carl O. Sauer. Directly tied to the *Landschaft* theme, it sought to give geography a phenomenon of study – the landscape or land system as an integrative, human–environment product – consistent with the foundational needs of the systematic sciences. Practitioners diverged in intellectual orientations, however, embracing natural history (as opposed to systematic science), cultural-historical, and humanistic approaches to landscape questions, themes, and understanding.

In the mid- to latter half of the twentieth century a science-based approach to land systems emerged under the rubric of cultural ecology (Turner and Robbins 2008), sharing many attributes with research under that label in anthropology and, within geography, sharing research space with risk hazard studies. This specialty began to splinter as critical and social theory entered the human–environment subfield, generating political ecology. Given the subsequent critiques of risk hazards and cultural ecology and the social-cultural theory turn of much of geography by the 1990s, practitioners favoring a systematic science orientation to human-environment issues opted to join – indeed, in many cases helped to create – new, integrated sciences and science-based subfields. Sustainability science is one example (Kates et al. 2001), a newly minted field of study that has embraced risk-hazard research, relabeled as “vulnerability” and linked to resilience, and land change/system science. Currently, many
of the topics of study undertaken in political ecology and land change science are similar, if not the same. What differentiates the two subfields is their framing: land change as post-positivist science and political ecology as a primarily structuralist and constructivist project.

Land change/system science

Land change science seeks to understand and explain land dynamics. These dynamics are viewed as a product of the interactions of the two subsystems comprising land use and cover, thus marking land systems as a type of social-environmental (or coupled human-environmental) system. Land change science seeks to identify and understand the causes and consequences of changes in land systems, which in turn require attention in the way of observation and monitoring, especially through the use of remote-sensing methods, as well as reconstructing land change throughout human history. Various modeling methods are used to test theories and themes of land change as well as to project these changes into the near-term (decadal) future. Significant portions of land change science overlap with the research interests of landscape sustainability, including attention to products that can be used to inform decision-making.

Causes and consequences

As a type of social-environmental system, land systems conform to the attributes of complex adaptive systems. They are associated with general drivers, causes, and processes, but the complexity of interactions between the components comprising the two subsystems and the variations in the base conditions in which they operate lead to place-specific outcomes (Kates et al. 2001). Land change science seeks to identify the general factors and principles underpinning land change and their mediation by a myriad of case-specific factors. For example, considerable research has demonstrated that over the long term and at the global scale the factors comprising the IPAT identity (environmental impact \( I = \text{population} \times \text{affluence} \times \text{technology} \) ) track with land change, pointing to these sources of demand and efficiency in production and consumption, or the causes of their change, in the reshaping of land uses and cover. At finer spatiotemporal resolutions, however, other factors, some more distal (e.g., markets, migration, environmental teleconnections) and others more proximate (e.g., local institutions, ecosystem services) reconfigure the outcomes of these demand factors or overpower them in regard to the explained variance of specific cases (types and locations) of land change (e.g., Lambin et al. 2001). For these reasons, the causes of land change are treated both quantitatively and qualitatively, although qualitative research tends to inform, not supplant, statistical approaches.

A large range of social, environmental, and interactive consequences of land changes has been addressed. Examples include such theories and themes as forest transition or the loss and gain for forest cover with changing economic conditions and household life cycles, the “hollow frontier” hypothesis addressing abandoned landscapes, “virtual” land change or the “out of place” consequences of “in place” land changes, and the agricultural intensification–land-sparing hypothesis. In these and other examples, such socioeconomic consequences as migration, remittances, and household income, and such environmental consequences as human-appropriated net primary productivity or biodiversity lost are addressed. Increasing attention is given to urban land change,
and urban–hinterland and urban–distant land interactions.

Observation and monitoring

Major progress has been made in land observation and monitoring, advancing the acquisition, processing, and analysis of data at multiple scales (e.g., NRC 2013). Workhorse imagery data, such as that from the Landsat or SPOT system of satellites, now provide a four-decade record of land change. As this archive of data grows, as well as the temporally dense MODIS dataset, land change studies have increasingly shifted toward more complex time-series assessment techniques. Previously constrained by data and computational power, current research is able to make use of growing cyberinfrastructure and supercomputers to analyze long time series of data, providing insights into changing phenology and system response times. The resolution and frequency of MODIS data have also created rich datasets that support and improve global land-cover models. Meanwhile, the availability of very fine resolution datasets, including multiband aerial photography, and new methods for image analysis (i.e., object-based image analysis) enhance opportunities to map microscale heterogeneous landscapes, facilitating land change observation and analysis in highly complex urban areas, treating such topics as land system composition and configuration on urban heat islands. Coupled with LiDAR data to detect elevation, fine resolution data increasingly offers opportunities to model three-dimensional changes in landscapes, and the implications for such fine-grain issues as the role of trees on air surface temperature and thus household energy expenditures and water use.

Past change

The historic human imprint on ambient environmental conditions and the function of land systems regionally and globally have been advanced with much improved mapping accuracy (e.g., Ellis and Ramankutty 2008). Paleo-environmental, archaeological, and historical evidence has been used to reconstruct past land system dynamics, providing strong clues about the use of prehistorical and historical land systems as analogs for current and future land systems. Past outcomes of social-environmental system dynamics are less useful than is information about the kinds, trends, trajectories, and rates of change of past interactions. Coupled with an understanding of complex land change dynamics, these improvements have led to new assessments of the role of land changes in proposed social-environmental collapses.

Modeling and projecting

The array of land change issues that can be addressed with multiple modeling approaches is staggering (NRC 2013). Land change models have been widely applied for descriptive assessment that simulate spatial changes in land use, as well as for prescriptive applications to optimize land use configurations for a given objective. These include, but are not limited to, land change models of agent (human actors) land use choices, including in urban areas, predictions of deforestation and agricultural change patterns, and time-series and economics models of land use change. In an effort to increase the relevance of land change models for social and ecological outcomes, recent research emphasizes models of land use and land functions, rather than simply land cover change. Perhaps of special interest to geography, various efforts are underway to specify not only what kind and degree of land changes will take place, but their location and
Land change science is a field that examines the spatial arrangement of land use and land cover changes over time. It is an interdisciplinary science that combines insights from the spatial sciences and land change science to inform one another. This synergy has begun but has much room for growth.

**Landscape and application**

Finally, advances in almost all facets of land change science have developed in tandem with land-based research in landscape ecology. The major distinctions between the two are landscape ecology’s emphasis on the ecological consequences of land change, especially on ecosystem function and ecosystem services, relative to that of land change science. Both subfields, however, share an interest in how changes in ecosystem services affect human wellbeing and sustainability in general (Wu 2013). In this regard, these two research communities have begun to address changes in land systems and landscapes as a means to mitigate and adapt to environmental change, global to local in kind, consistent with the international and integrated research efforts of the Future Earth program. Special attention is given to the environmental and socioeconomic consequences of policy interventions on land systems, such as REDD+ on the retention of tropical forest through direct payments, and to the design or reconfiguration of land systems, including those in urban-suburban areas, accounting for environmental tradeoffs to inform sustainability decisions.

**SEE ALSO:** Environmental determinism; Environmental hazards; Environmental science and society; Global environmental change: human dimensions; Habitat destruction and fragmentation; Land-use/cover change and climate; Landscape; Landscape ecology; Land systems science; Political ecology; Sustainability science

**References**


**Further reading**


Land degradation

Lindsay C. Stringer
University of Leeds, UK

The United Nations Convention to Combat Desertification (UNCCD) defines land as “the terrestrial bio-productive system that comprises soil, vegetation, other biota and the ecological and hydrological processes that operate within the system” (UNCCD 1994). Land degradation describes a broad range of processes, operating across scales, from the local to the global, that cause a reduction in the quality, complexity, and functions of land and its cultural, recreational, and heritage values. Some land degradation processes are relatively fast-acting and move quickly, for example, soil erosion following an extreme weather event. Other land degradation processes move slowly, operating over longer time frames, for instance, nutrient and organic matter losses. While some people view land degradation as an end point, it is more commonly seen as a process, with degradation occurring once a particular tipping point or threshold has been exceeded.

Broadly, land degradation takes three main forms which may occur separately or in combination. The first form is biological degradation, in which soil biota and biodiversity (including bacteria, fungi, algae, protozoa, nematodes, and insects) are lost due to changes in vegetation, temperature, and moisture that alter the amount of organic matter in the soil. The second form is the physical degradation of land. In this case, the land surface is removed by water or wind erosion, or is altered or damaged in situ by processes such as compaction (e.g., due to plowing) or sealing (covering of the soil with impervious materials). The third form is chemical degradation, linked to nutrient losses, salinization, or contamination. Each form of degradation can reduce the ecosystem services and economic benefits that society can gain from the land, both during the present time under the current land-use system and into the future. Many published definitions distinguish land degradation from environmental change by emphasizing that degradation is part of a long-term negative trend in the particular context of an area’s land-use system. Some forms of land degradation are effectively permanent. For example, degradation in the form of bush encroachment is permanent unless actions are taken to clear the bushes.

Land degradation is considered to be an important threat to sustainable development. It is of particular concern in light of an increasing global population and climate change which lead to growing competition for land and water, particularly for use in food and fuel production. The maintenance of land quality and its functions underpins food production and climate regulation as well as providing the basis of livelihoods for billions of people across the world. It is therefore seen as imperative that actions are taken to prevent and reverse land degradation, and rehabilitate degraded areas where it is economically viable.

Land degradation is caused by different interacting drivers in different places. It is widely accepted to result from a combination of human and environmental (including climatic) factors. Large- and meso-scale climate and hydrological processes and patterns are superimposed with sociopolitical and economic structures and processes such as markets, property rights,
LAND DEGRADATION

Technological changes, population and demographic changes, and human migration. These interact with current and historical land-use and land management practices at smaller spatial scales (e.g., watershed or farm scales). The effects of land degradation can be experienced directly at the landscape site at which the degradation occurs, as well as indirectly in other places, for example, through increased food prices, food insecurity, and water quality decline, as well as reduced ecosystem service provision at larger scales.

Policymakers often need to know what land degradation costs in order to determine how much to invest in addressing and preventing the problem. The diffuse nature of degradation impacts means that assessment methodologies are required that measure the costs of land degradation both on- and off-site. Some methodologies, such as total economic value (TEV) approaches (e.g., Turner et al. 2003), consider the costs and benefits of actions to mitigate degradation, to rehabilitate and restore degraded areas, as well as to instigate a shift toward more sustainable land management practices. TEV approaches can also encompass the costs and benefits of inaction against land degradation.

People who heavily depend on the natural resource base to support their subsistence are considered most vulnerable to the impacts of land degradation. While it is often presented as a developing world problem, affecting primarily agricultural economies, land degradation is an issue for several different sectors and occurs throughout the world. For example, transport and industry (including mining and ore smelting) can cause chemical land degradation through the emission and deposition of sulfur oxides and nitrous oxides. These chemicals acidify soils, damage vegetation, and negatively affect water quality in places far away from the original site of emission. Similarly, physical degradation in the form of soil sealing can occur as a result of infrastructure development and urbanization, as the land surface is covered with impervious materials, reducing, among others, its climate regulation and water purification functions. In Europe, nearly 10% of the total land area is estimated to be covered with impermeable material. Projected trends in population growth and migration suggest that soil sealing due to urbanization will continue into the future, in both developed and developing countries.

Whether land degradation is seen as a problem in need of remediation or not depends largely on the context and scale of analysis (Warren 2002). Land degradation for one group may be seen merely as an environmental change for another (Blakie and Brookfield 1987). This gives the management of land degradation a clear political dimension in many settings in which it occurs. Degradation takes place when the economic and ecosystem service benefits that people obtain from the land are negatively affected in the context of the local land-use system. For example, the encroachment of woody species in rangeland areas (unrelated to rainfall patterns and dynamics) may be seen as an environmental change by scientists. Yet, pastoralists who require a variety of grass species to feed their livestock may see that change as land degradation, in the context of their local land-use system. For them, bush encroachment represents a change in vegetation cover and community species composition and complexity, and has a negative economic and livelihood impact.

Various attempts have been made to estimate the extent of land degradation in recent decades. However, each assessment has used different methods, measured different variables over a range of temporal and spatial scales (using different indicators), and has focused on different systems. This makes comparison between the various assessments highly challenging. For
example, the Millennium Ecosystem Assessment (2005) estimated that 10–20% of the world’s drylands are degraded (with medium certainty; 65–85% probability), while Oldeman, Hakkeling, and Sombroek (1991) looked across all climatic zones in their global assessment of soil degradation (GLASOD) and suggested that 3.5 billion ha of land is suffering from soil degradation. The wide variation in these figures that focus on different systems and different components of land means it is difficult to identify the precise magnitude and speed of land degradation, as well as to ascertain the urgency with which it needs to be addressed.

In recent decades, emphasis has been placed on multispectral remote sensing analyses that allow calculation of the greenness of live canopy vegetation, using the normalized difference vegetation index (NDVI). Live, healthy, green plants reflect in the infrared part of the electromagnetic spectrum. The NDVI ratio is ascertained by dividing the difference in near-infrared and red color bands by the sum of the near-infrared and red colors bands for each pixel in an image. Although NDVI analyses provide useful insights that other methods cannot attain, the limitations of the approach are now widely accepted. These include, for example, acknowledgment that NDVI cannot determine vegetation height, making it difficult to tell the difference between grass-dominated rangelands, bush-encroached areas, and expanses of natural tree cover. Ground truthing is required within the local land-use system, alongside evaluation of meteorological data, particularly in dryland systems given their high levels of rainfall variability. Such additional layers of information are vital, particularly for time-series analyses, because without them, it becomes impossible to determine whether any change in green vegetation cover over time is due to land degradation, or whether it is due to changing rainfall patterns, changing land use, or changing land management practices.

Due to the challenges associated with measuring land degradation directly, using any single method, reliance tends to be placed on indicators, which assess a range of slow and fast variables. Indicators allow focus on a particular state or trend, and can be used to explain how land degradation is changing both temporally and spatially. Indicators also permit measurement of the antithesis of land degradation: sustainable land management (SLM). Many groups consider that focusing on SLM is more useful than measuring land degradation. This is because donors and investors, interested in improving the land quality, often want to know the extent to which their efforts are improving the land in relation to the baseline situation.

The international political community has agreed to strive toward land degradation neutrality by 2030 as part of commitments to the Sustainable Development Goals (2015, Goal 15.3). In order to be able to monitor progress toward this, agreement will be needed as to which variables and processes should be measured and at what scale(s) assessments should be carried out for the various types of land degradation. This, however, remains challenging. Land degradation is difficult to assess and compare across contexts, not only because of differing interpretations of what constitutes degradation, but also because the type of degradation experienced in one location may differ from that occurring elsewhere. For example, some parts of the world suffer from salinization due largely to inappropriate irrigation, while others may be experiencing soil erosion in the form of gullying due to a combination of overgrazing, fire, and extreme weather. Such diversity implies that different indicators will be appropriate to different contexts.
LAND DEGRADATION

Historically, indicators have been developed by scientists to allow assessment of the various land degradation forms. However, the UNCCD – the main international treaty that deals with land degradation – notes that the locally held knowledge of dryland communities can assist in recognizing, coping with, and adapting to degradation by anchoring indicators more securely in the context of local socioeconomic and land-use systems. Consequently, researchers have recently sought to combine scientific rigor and accuracy with greater relevance and sensitivity to local perspectives and context, in their use and development of indicators (Stringer and Reed 2007). Delving into locally held knowledge can also provide valuable insights into traditional SLM measures. These can include agronomic, vegetative, structural, and/or land management measures that help to maintain productivity and sustain the land’s physical, biological, and chemical properties. They include, for example, green manure and cover crops, intercropping, the use of stone bunds, and low or no tillage treatments. Such SLM measures are often locally appropriate, inexpensive, and can be tested by researchers with regard to their upscaling and applicability to larger areas. Computer models that consider the biophysical and economic characteristics of a particular area can be used to ascertain the feasibility of upscaling particular SLM measures. The findings of modeling exercises may then be usefully evaluated with stakeholders, allowing appraisal of their applicability within different local social and cultural contexts.

The international community has acknowledged that land uses and land management practices are shaped by interacting processes and decisions made at scales from the local to the international. This led to the negotiation of the UNCCD as a framework for international, regional, sub-regional, and national action to address land degradation, desertification, and drought. The UNCCD was one of the three Rio Conventions, agreed upon at the UN Conference on Environment and Development (UNCED) in 1992 (alongside the UN Convention on Biological Diversity (CBD) and the UN Framework Convention on Climate Change (UNFCCC)). Although the UNCCD is international in its scope, it places particular focus on desertification, land degradation, and drought in Africa and the arid, semiarid, and dry subhumid parts of the world, reflecting the understanding that the world’s drylands (i.e., arid, semiarid, and dry subhumid areas) are considered particularly vulnerable to land degradation. Countries that are party to the UNCCD and which consider their territories to be affected by land degradation are obliged to develop a national action program (NAP). NAPs identify the types and locations of land degradation in each country, alongside the actions necessary to prevent and reduce degradation, and to rehabilitate already degraded areas. The UNCCD further notes the need to draw on a range of different scientific and locally held knowledges about land degradation given its contextual nature. As such, it advocates an approach to NAP development that involves the active participation of affected land users.

Although the UNCCD entered into force following its fiftieth ratification in 1996, it has been widely criticized for its limited on-the-ground impact in terms of reversing, preventing, and reducing land degradation. Because it is grounded in soft law, there are no sanctions should parties to the agreement fail to meet their obligations. Consequently, the UNCCD has suffered from limited donor resources and, in many countries, a lack of institutional, financial, and human capacity for NAP development and implementation (Bauer and Stringer 2009). The challenge is further complicated because land degradation touches on so many sectors, ranging
from forestry, agriculture, water, energy, and mining to health, education, and development. This means that cross-sectoral capacity building is required. However, many capacity building investments by donors to date have been sector or issue specific (Akhtar-Schuster et al. 2011). In some countries, national focal points (those individuals responsible for interfacing with the UNCCD on behalf of their countries) are based in weak ministries. This provides a further limitation. In countries where tackling land degradation has limited political resourcing and capacity, the role of nongovernmental organizations (NGOs) becomes more important in raising funds for and implementing projects that tackle land degradation issues. Indeed, NGOs play a vital role in UNCCD implementation activities across the world.

At the international level, donor resources are perceived to be more easily available and accessible for initiatives under the CBD and UNFCCC. However, the problems dealt with by these conventions set at Rio are thought to be more clear-cut than that of land degradation. For example, the UNFCCC tackles climate change, a problem that can be clearly defined (i.e., rising average temperatures related to an increase in atmospheric CO₂ levels). It is relatively straightforward to set a target to reduce greenhouse gas emissions by a certain amount by a particular date and to monitor progress made toward that goal. Furthermore, the atmosphere is more obviously an open-access common property resource, changes to which by one group of actors will affect others. Land degradation is complicated because land is often seen as a national concern, which is subject to national laws and regulations (and their enforcement). Despite various different property rights that enable the sale or rent of land, many countries are reluctant to sign up to any agreements that could ultimately interfere with the ways in which their territory is governed. Indeed, even within the European Union, agreement recently failed to be reached on the development of a soil framework directive (which focused on only soils – just one element of land). Some member states felt that land matters should be managed at the lowest possible governance level in line with the subsidiarity principle. Consequently, discussions on a soil framework directive have currently been halted.

A further challenge to taking a common approach to tackling land degradation links to the ways in which the problem is measured, monitored, and communicated. The multisectoral nature of land degradation demands scientists take interdisciplinary and multidisciplinary approaches in defining both the extent of the problem and the urgency with which action needs to be taken (Stringer and Dougill 2013). Unlike the UNFCCC, which receives regular scientific updates from the Intergovernmental Panel on Climate Change (IPCC), and the CBD, which is supported by the Intergovernmental Panel on Biodiversity and Ecosystem Services (IPBES), the UNCCD’s institutional structure has historically hampered the ability of the scientific community to draw together a fragmented, multidisciplinary science base and channel research on land degradation into policy solutions. A decision by the UNCCD’s eleventh Conference of the Parties (COP-11), held in Windhoek, Namibia, in 2013 nevertheless marks an important step forward, with parties’ recognition of the need to develop an improved science-policy interface to bring together the current state of knowledge on land degradation and sustainable land management. It is hoped that this initiative will underpin efforts to deliver a reduction in land degradation and a widespread shift toward SLM.
LAND DEGRADATION

SEE ALSO: Agriculture; Desertification; Ecosystem services; Environment and development; Livelihoods; Soil erosion and conservation; Sustainable development

References


Recent land systems science (LSS) literature refers to interdisciplinary and multidisciplinary research that improves our understanding of land-cover and land-use dynamics as part of coupled human-environment systems. It includes theoretical and conceptual advances, improvements in data sources and innovations in numerical modeling, and applications relevant to environmental and societal problems at local to global scales. Rounsevell et al. (2012) identify four multidisciplinary aspects of LSS: behavior of people and society, the multilevel character of decision-makers and of land units, the connections between people and land globally, and past and future change.

The origins of contemporary LSS lie in research conducted in international science and social science programs that were organized by International Geosphere Biosphere Program (IGBP) and the International Dimensions of Global Change Program (IHDP), in particular the Land Use and Land Cover (LUCC) and the Global Change and Terrestrial Ecosystems (GCTE) programs. LSS emerged as the preferred term under the Global Land Project (GLP) and through the Danish Network for Land Systems Research (Reenberg 2009). While these programs comprise the terminological family tree, other key influences on contemporary LSS include the Millennium Ecosystem Assessment and the National Science Foundation’s Coupled Natural Human Systems (CNH) program. The GLP, which is now part of Future Earth, set itself two major challenges. First, to improve our understanding of the complex feedbacks between societal and environmental components of integrated land systems; and second, upscaling local and regional process understanding to enable global process understanding.

The critical role of land surface changes in the Earth system is underpinned by a substantial volume of research conducted in the last three decades (Turner, Lambin, and Reenberg 2007, and references therein) and has enabled land-use and land-cover change to evolve into land systems science. LSS is important because the Earth system supports over 7 billion people with higher per capita standards of living than at any time in history. This is increasingly understood in terms of provisioning, supporting, regulating, and cultural (ecosystem) services, which though only termed ecosystem services recently, have been changing the Earth system through land system transformations locally across the world for over 60 millennia, but in terms of major global impacts since the seventeenth century.

LSS has four objectives (Turner, Lambin, and Reenberg 2007): first, to increase our ability to observe and monitor contemporary land changes. Improvements in and better access to satellite image data, image archives, and innovations in information extraction have brought about improved surveillance. Though satellite image data is not the only data source for land systems science, it has a primacy that has survived the transition from land-use and land-cover change to LSS; second, to enable land system changes to be understood in an
integrated and holistic framework that couples human and natural systems; third, to provide spatially explicit modeling of land change, which has become a necessity because of the complexity of the parameters required to monitor and understand change, and the feedbacks between parameters. The requirement for spatially explicit modeling, as against nonspatial modeling, arises because of the availability of spatial datasets and the continued developments in GIS. Turner, Lambin, and Reenberg (2007) state that land change science is played out at the interfaces of social science, natural science, and geographic information-remote sensing science; and finally, to assess and evaluate land system outcomes in arenas that matter to society globally; arenas such as vulnerability, resilience, and sustainability.

The complexities in researching any of these objectives, but particularly the latter three, lie in the fact that land change trajectories, and therefore Earth system change pathways, are governed by local- and regional-scale human actions which involve interactions between decision-making, institutions and mechanisms of governance, patterns and trends in production and consumption of ecosystem services, technological development and diffusion, and the local and regional impacts of global environmental change. Moreover, once mapped and understood, these local- and regional-scale decisions have to be scaled up to understand their influences on LSS globally.

Rounsevell et al. (2012) identify three emerging questions to be addressed in the next phase of LSS. First, how can the analysis of empirical and historical land system datasets provide insight into human–environment interaction? Second, how can integrated modeling and the ecosystem services concept contribute to the testing of hypotheses about land system functioning and decision-making? Third, how can our current understanding of land systems inform the choices that society has about future landscapes? The first question is interesting as it creates a space for historical geographers within LSS and the ability to understand past land-use decisions which will help explain the current state of the land units, as well as providing better datasets for model testing and validation. The second question acknowledges the increasing integration of land system science with ecosystem services. The final question underlines the importance of this academic project to the future of Earth and its people, a point acknowledged by its inclusion in Future Earth.

LSS is situated in a strategic area of geographical enquiry – the nature–society divide. Its interdisciplinary and multidisciplinary nature suggests many disciplines have large stakes in LSS: yet geographers have been influential in the development of LSS globally, and continue to be at the forefront of the development of theory and practice in LSS.

SEE ALSO: Earth system science; Ecosystem services; Land change science; Sustainability science

References


Further reading

Landforms and physiography

David R. Butler
Texas State University, USA

Richard A. Marston
Kansas State University, USA

Landforms are recognizable topographic features on the surface of the Earth with a distinct shape and position in the landscape, as well as a characteristic local relief, composition of geologic materials, age, and processes of formation, maintenance, and change. Landforms are commonly found in assemblages that geomorphologists term a landscape. Landforms result from the interplay of the resisting framework, driving forces, and time. By resisting framework is meant the geological and biological features on which processes act, including lithology (rock type), stratigraphy (the layered sequence of rock types), and geologic structure (the geometry of rocks). Vegetation exerts resistance to landform change through effects such as the root-binding effects on soils, and the roughness against wind and flowing water. Driving forces refer to the endogenic processes of vulcanism and diastrophism (uplift, folding, faulting); the exogenic processes (weathering, erosion, transport, deposition) controlled by the agents of water, wind, ice, gravity, and biota; and the extraterrestrial processes (meteorites, asteroids, comets). Time refers to development of landforms over time. Because the resisting framework and driving forces both change over time, landforms can be active or relict, having formed under a former set of conditions. More often than not, landscapes are assemblages of landforms that are polygenetic in origin, that is, relict landforms are inherited from prior situations and are being modified by current processes. Almost all landforms observed on Earth today have been created during the Cenozoic Era, or last 65 million years. A relationship exists between the size of a landform and the length of time over which it typically develops. The difficulty in predicting (or reconstructing) landform change over time is nicely explained by Phillips (2007). On the one hand, all landscapes develop under the laws of chemistry and physics. On the other hand, each landscape develops with its own history of environmental perturbations that render each landscape as perfectly unique. The search for the “Theory of Everything” in geomorphology remains elusive as a result.

The earliest book on physiography can be attributed to the Scottish science writer Mary Somerville, who authored a book in 1848 titled Physical Geography, which featured detailed reporting on the topography of each continent. Thomas Henry Huxley published a book titled Physiography in 1877 in Britain, after which physiography became one of the most popular subjects in geography (Holt-Jensen 1999). In the United States, the earliest effort to describe regional geomorphology was authored by John Wesley Powell (1896) and included a map that was adopted by other geomorphologists in subsequent efforts. The term “physiography” was coined as a contraction of the words “physical geography” and was a steadfast part of academia in the early twentieth century.
Figure 1  The physiographic provinces of the contiguous United States, following the boundaries of Fenneman, with subdivisions. US Geological Survey (USGS) map colored by paleogeological areas and demarcating the sections of the US physiographic regions: Laurentian Upland (area 1), Atlantic Plain (2–3), Appalachian Highlands (4–10), Interior Plains (11–13), Interior Highlands (14–15), Rocky Mountain System (16–19), Intermontane Plateaus (20–22), and Pacific Mountain System (23–25). Reproduced from USGS.

The term physiography was clearly intended to serve as an overarching term for the universe of concepts typically taught in physical geography classes, including geodesy, meteorology, climatology, soils, plant distributions, and geomorphology. Widely used textbooks from the early twentieth century (Tarr and Martin 1917; Salisbury 1929) used the term physiography in their title and were topical compendiums of the entirety of the field of physical geography. However, after the publication of Isaiah Bowman’s *Forest Physiography* (1911), the field of physiography began to change and to mean regional physical geography. More specifically, the term began to mean regional geomorphology as associated with the boundaries of physiographic provinces as a result of the works of Nevin Fenneman, who was appointed by the Association of American Geographers in 1914 to chair a committee tasked with preparing a map of the physiographic divisions of the United States. The boundaries of the US physiographic provinces, which encompass regions with similar geomorphological histories and landforms, became standardized (Figure 1) and were subsequently the topics in Fenneman’s landmark volumes on the
physiography of the western (Fenneman 1931) and eastern (Fenneman 1938) United States. Beginning in the 1930s, Hungarian born Erwin Raisz produced thousands of maps and two atlases featuring landforms drawn with pen and ink for all regions of the world. Books by Atwood (1940), Thornbury (1965), and others continued the regional geomorphology (physiography) tradition in the United States well into the period when process geomorphology was becoming a dominant paradigm in the discipline. As part of the celebration of the centennial of the Geological Society of America, Graf (1987) edited a volume titled *Geomorphic Systems of North America* with 13 chapters that describe regional geomorphology and also synthesize process-based research for each region.

Vitek, Giardino, and Fitzgerald (1996) traced the changes in mapping landforms from paper mapping, through computer mapping, to virtual reality. The availability of digital elevation models (DEMs) has led to an area of research known as geomorphometry, where derivative properties of landforms can be measured. These properties, such as slope angle, slope aspect, slope curvature, and channel networks, can then be delineated for watersheds. Global DEMs are available from the shuttle radar topographic mission (SRTM), with 1 arc-second (approximately 30 m) horizontal resolution and vertical resolution in meters. Laser ranging surveys (LiDAR) are acquired with aircraft that can produce submeter spatial resolution. Accurate, high resolution (±1 m vertical) even exist for Mars. Starting with sampling land surface data from a DEM, a land surface model (landscape) can be generated and soil, hydrology, and land use/land cover can be added to the model. Remotely sensed data can be draped over the land surface model (Figure 2).

Bailey (1983) and Omernik (1987) introduced the concept of ecoregions, which are mappable units derived by integrating maps of physiography, soils, geology, potential natural vegetation (or land use/land cover), hydrology, climate, and wildlife. The schemes of both Bailey and Omernik are hierarchical in scale. Bailey tended to rely on one or two subject maps at each of the different levels in his hierarchy, whereas Omernik attempted to integrate most of the elements of the landscape. Ecoregions provide a template for designing field sampling schemes or formulating regional, continental, or global-scale geomorphic comparisons (Simon, Dickerson, and Heins 2004). Ecoregions more clearly represent regional physical geography to a greater extent than did the early classic books on the subject cited earlier, because ecoregions recognize that the landscape consists of landforms interacting with other elements of the landscape. Hence, ecoregions represent another step in the regional study of landforms and offer linkages between the study of geomorphology and the fields of biogeography and ecology (Figure 3).
Figure 3  Level III ecoregions of the conterminous United States. Oklahoma state boundaries are highlighted to show one of the most diverse states in the USA in terms of ecoregions. USEPA (2003). Reproduced from USGS.
SEE ALSO: Ecoregions; Geomorphology: history; Physical geography; Soil mapping and maps

References


Further reading

Landforms of other planets

René De Hon
Texas State University, USA

While resolutions and coverage of the surfaces vary considerably, the terrestrial planets and most large satellites of the outer planets have been imaged in sufficient detail to recognize their major landforms. Satellite-borne radar altimetry and imaging radar provide elevation data and images of surfaces hidden by an atmosphere. Physiographic maps at varying scales have been prepared for all of the planetary bodies that have been imaged by spacecraft. The first venture into describing and interpreting a landscape other than the Earth’s was by telescopic examination of our nearest neighbor, the Moon. Modern telescopic viewing limits resolution of lunar landforms to about 500 m. Landforms of other planets and satellites are identified by spacecraft imaging, which began in the 1960s and continues with increasing detail at this time. Table 1 provides a list of terms applied to various landforms in naming features on the planets.

Four chief groups of processes shape planetary surfaces: tectonism, volcanism, gradation, and meteor impact. Volcanism and tectonism are products of a planet’s internal temperature distribution and density variations which drive mantle convection. As the planetary body cools, that engine shuts down and internal movements cease. The lifetime of that engine is a function of planet size, as small bodies lose heat much more rapidly than large bodies. External processes are tied to surface gravity, atmospheric interaction with rocks, and infall of meteoric debris from space. Planetary bodies without significant atmospheres tend to preserve primitive landscapes only slightly modified by erosion. Table 2 summarizes the land-shaping processes observed on the various planetary bodies.

Outer surfaces of terrestrial planets are composed of rocky (silicate-rich) materials. Their early history records volcanically generated surfaces which were highly cratered by meteor impacts. Planets evolved along divergent paths dictated by differences in size, internal thermal cooling, atmospheres, and distance from the Sun. Large bodies such as Earth and Venus retain sufficient internal energy to reshape the surface within fairly recent geologic timescales. Smaller bodies such as Mercury and the Moon, having lost most of their internal energy, retain landforms generated during their earliest history. Mars, intermediate in size, retains some of its early cratered crust, but it exhibits an extended period of volcanism and hydrologic modification of the surface. The gas giants of the outer solar system retain primordial gases and have no clearly defined surface.

Satellites of the outer planets include a few extremely unique bodies such as Io’s active sulfur volcanism and Titan’s hydrocarbon lakes. Most large satellites have surfaces of ice or ice-rock mixtures. Landforms are impact craters and fractured ice terrains. Some icy surfaces are resurfaced by more recent eruptions of liquid water to form smooth, icy plains that have lost all evidence of early impact-dominated terrains.
# Landforms of Other Planets

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albedo feature</td>
<td>Geographic area distinguished by amount of reflected light; typically bright or dark</td>
</tr>
<tr>
<td>Arcus, arcus</td>
<td>Arc-shaped feature</td>
</tr>
<tr>
<td>Astrum, astra</td>
<td>Radial-patterned features on Venus</td>
</tr>
<tr>
<td>Catena, catenae</td>
<td>Chain of craters</td>
</tr>
<tr>
<td>Cavus, cavi</td>
<td>Hollows, irregular steep-sided depressions usually in arrays or clusters</td>
</tr>
<tr>
<td>Chaos, chaoses</td>
<td>Distinctive area of broken terrain</td>
</tr>
<tr>
<td>Chasma, chasmata</td>
<td>A deep, elongated, steep-sided depression</td>
</tr>
<tr>
<td>Collis, colles</td>
<td>Small hills or knobs</td>
</tr>
<tr>
<td>Corona, coronae</td>
<td>Ovoid-shaped feature; generally composed of fractures</td>
</tr>
<tr>
<td>Crater, craters</td>
<td>A circular depression; fresh craters are surrounded by a raised rim</td>
</tr>
<tr>
<td>Dorsum, dorsa</td>
<td>Ridge</td>
</tr>
<tr>
<td>Flexus, flexūs</td>
<td>A very low curvilinear ridge with a scalloped pattern</td>
</tr>
<tr>
<td>Fluctus, fluctūs</td>
<td>Flow terrain</td>
</tr>
<tr>
<td>Flumen, flumina</td>
<td>Channel on Titan that might carry liquid</td>
</tr>
<tr>
<td>Fossa, fossae</td>
<td>Long, narrow depression</td>
</tr>
<tr>
<td>Labes, labēs</td>
<td>Landslide</td>
</tr>
<tr>
<td>Labyrinthus, labyrinthi</td>
<td>Complex of intersecting valleys or ridges</td>
</tr>
<tr>
<td>Lacuna, lacunae</td>
<td>Irregularly shaped depression on Titan having the appearance of a dry lake bed</td>
</tr>
<tr>
<td>Lacus, lacūs</td>
<td>“Lake” or small plain</td>
</tr>
<tr>
<td>Linea, lineae</td>
<td>A dark or bright elongate marking; may be curved or straight</td>
</tr>
<tr>
<td>Mare, maria</td>
<td>“Sea”; large circular plain; usually volcanic except on bodies with stable liquid at surface</td>
</tr>
<tr>
<td>Mensa, mensae</td>
<td>A flat-topped prominence with cliff-like edges</td>
</tr>
<tr>
<td>Mons, montes</td>
<td>Mountain</td>
</tr>
<tr>
<td>Oceanus, oceani</td>
<td>A very large dark area on the moon; basalt plains</td>
</tr>
<tr>
<td>Palus, paludes</td>
<td>“Swamp”; small plain</td>
</tr>
<tr>
<td>Patera, paterae</td>
<td>An irregular crater, or a complex one with scalloped edges; usually volcanic</td>
</tr>
<tr>
<td>Planitia, planitiae</td>
<td>Low plain</td>
</tr>
<tr>
<td>Planum, plana</td>
<td>Plateau or high plain</td>
</tr>
<tr>
<td>Plume, plumes</td>
<td>Cryovolcanic features on icy planets and satellites</td>
</tr>
<tr>
<td>Promontorium, promontoria</td>
<td>“Cape”; headland</td>
</tr>
</tbody>
</table>
Table 1  Continued

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regio, regiones</td>
<td>A large area distinct from adjacent areas, or a broad geographic region</td>
</tr>
<tr>
<td>Reticulum, reticula</td>
<td>Reticular (net-like) pattern on Venus</td>
</tr>
<tr>
<td>Rille, rilles</td>
<td>Long narrow trough; narrow graben</td>
</tr>
<tr>
<td>Rima, rimae</td>
<td>Fissure</td>
</tr>
<tr>
<td>Rupes, rupès</td>
<td>Scarp</td>
</tr>
<tr>
<td>Scopulus, scopuli</td>
<td>Lobate or irregular scarp</td>
</tr>
<tr>
<td>Serpens, serpents</td>
<td>Sinuous feature with segments of positive and negative relief along its length</td>
</tr>
<tr>
<td>Sinus, sinūs</td>
<td>“Bay”; small plain</td>
</tr>
<tr>
<td>Sulcus, sulci</td>
<td>Sub-parallel furrows and ridges</td>
</tr>
<tr>
<td>Terra, terrae</td>
<td>Extensive land mass; high-standing terrain</td>
</tr>
<tr>
<td>Tessera, tesserae</td>
<td>Tile-like, polygonal terrain</td>
</tr>
<tr>
<td>Tholus, tholi</td>
<td>Small domical mountain or hill</td>
</tr>
<tr>
<td>Unda, undae</td>
<td>Dunes</td>
</tr>
<tr>
<td>Vallis, valles</td>
<td>Valley</td>
</tr>
<tr>
<td>Vastitas, vastitates</td>
<td>Extensive plain</td>
</tr>
</tbody>
</table>

Major physiographic provinces

Many planetary bodies can be divided into major physiographic provinces based on average elevations, origin of landforms, extent of cratering, or relative age of surface materials. Planetary bodies that retain surfaces from their earliest history usually have heavily cratered surfaces and later, volcanic plains materials. Planets with active tectonics and atmospheric-driven erosional processes may have lost evidence of extensive cratering.

Major provinces are seen as variations in the style of volcanism or tectonic deformation (Figure 1). Mercury is divisible into intercratered plains, heavily cratered terrain, and lightly cratered plains. Venus consists of mostly lowland, basalt plains with a small percentage of continental-like crustal plateaus surrounded in part by intermediate elevation shelf. Earth has distinct continents and oceans. The Moon consists of bright, cratered highland terrae and dark, low-lying maria. Mars is divided into a cratered upland and sparsely cratered lowland separated by a planet-encircling escarpment. Straddling the dichotomy boundary in Mars’s western hemisphere is a massive volcano-tectonic province known as the Tharsis bulge.

Landscape-forming processes

Meteoric and asteroidal impact

The early solar system was awash in asteroidal debris in planet-crossing orbits. Early planetary surfaces were subjected to intense bombardment that saturated surfaces with overlapping craters of
Table 2  Summary of processes and landforms on planets and moons.

<table>
<thead>
<tr>
<th>Planet/process</th>
<th>Impact</th>
<th>Volcanism</th>
<th>Tectonism</th>
<th>Surface water</th>
<th>Glacier</th>
<th>Wind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>Abundant</td>
<td>Lava plains</td>
<td>Scarp; rille; thrust faults; ridges</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Venus</td>
<td>Sparse</td>
<td>Lava plains; shields; coronae</td>
<td>Scarp; rille; ridges; riffs; tessor</td>
<td>None at present; possible in past</td>
<td>None</td>
<td>Sparse</td>
</tr>
<tr>
<td>Earth</td>
<td>Sparse</td>
<td>Ocean floor; lava plains; cones</td>
<td>Plate tectonics; faults; riffs; folds</td>
<td>Ocean; lakes; river channels</td>
<td>Polar cap; ice sheets; valley glaciers</td>
<td>Abundant dune fields, yardangs</td>
</tr>
<tr>
<td>Moon</td>
<td>Abundant</td>
<td>Lava plains; minor domes and lava channels</td>
<td>Scarp; rille; ridges</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Mars</td>
<td>Abundant</td>
<td>Lava plains; shields</td>
<td>Scarp; rille; ridges</td>
<td>Past ocean; lakes; channels; chaos</td>
<td>Polar caps; valley glaciers</td>
<td>Abundant albedo streaks, dune fields, yardangs</td>
</tr>
<tr>
<td>Io</td>
<td>None surviving</td>
<td>Pervasive</td>
<td>Scarp</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Europa</td>
<td>Rare</td>
<td>Cryovolcanic resurfacing</td>
<td>Abundant lineaments and grooves</td>
<td>Cryovolcanism</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Ganymede</td>
<td>Abundant</td>
<td>Cryovolcanism</td>
<td>Abundant fractures</td>
<td>Cryovolcanism</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Callisto</td>
<td>Heavily cratered</td>
<td>Cryovolcanism</td>
<td>Cryovolcanism</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Tethys</td>
<td>Heavily cratered</td>
<td>Minor resurfacing</td>
<td>Scarp; grooves</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Dione</td>
<td>Heavily cratered</td>
<td>Minor resurfacing</td>
<td>Scarp; grooves</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Rhea</td>
<td>Heavily cratered</td>
<td>Minor resurfacing</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Titan</td>
<td>Few</td>
<td>Cryovolcanism</td>
<td>Scarp; grooves</td>
<td>Grooves may be channels</td>
<td>None</td>
<td>Sparse</td>
</tr>
<tr>
<td>Iapetus</td>
<td>Heavily cratered</td>
<td>Equatorial ridge</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Triton</td>
<td>Abundant</td>
<td>Cryovolcanism</td>
<td>Scarp and chasma</td>
<td>None</td>
<td>Nitrogen polar cap</td>
<td>?</td>
</tr>
</tbody>
</table>
Figure 1  The topographic maps of Venus (NASA/JPL Magellan Laser Altimeter), Earth (NASA, USGS), and Mars (NASA/JPL Mars Orbiter Laser Altimeter). Venus and Earth have relatively high-standing continents surrounded by low-elevation plains. Neither planet preserves ancient terrains from the early period of meteorite and asteroid bombardment. Mars exhibits a sparsely cratered, low-elevation, northern plains and a higher-standing, highly cratered southern highlands. Reproduced from NASA and USGS.
LANDFORMS OF OTHER PLANETS

Figure 2  The far side of the Moon displays the heavily cratered surface associated with high impact rates from the early period of high-impact flux (NASA Apollo 16 photograph). Reproduced from NASA.

varying sizes (Figure 2). As planets cleared their immediate vicinities and orbits of debris, impact rates declined to the present very low frequency of impacts.

The relative velocities of meteors in planet-intersecting orbits are very high in respect to planetary surfaces. Impact at hypervelocity (velocity greater than the of speed of sound in the target) results in a near instantaneous conversion of kinetic energy of the meteorite’s mass and velocity to shock and heat energy in the target. The result is total destruction of the meteorite and excavation of an explosion crater surrounded by a rim of material ejected from the crater. Young craters typically have high albedos and sharp topographic features. They are also associated with crater chains of secondary impacts and bright ray patterns radiating away from the crater for great distances.

Crater morphology is a balance between impact energy, target material strength, atmospheric interaction, and planetary gravity (Melosh 1989). Small objects do not survive passage through an atmosphere. Large bodies that pass through the atmosphere retaining most of their kinetic energy leave indelible scars on the surface. Small craters, less than 5–20 km, are simple, bowl-shaped depressions with a well-developed raised rim. As impact energy increases, crater diameter increases and the crater develops terraced interior walls, a flat floor, and a central peak. At greater impact energies, the central peak breaks into multiple peaks and transitions into an inner ring of peaks. At still larger impacts, multi-ring basins greater than 100 km in diameter are created. Weird terrain is the vernacular term given to hilly and furrowed terrain antipodal to the Caloris Basin of Mercury. It is thought to have been formed by the convergence of shock waves from the Caloris impact.

Crater topography degrades over time. Crater rims are lowered, rays fade, secondary craters are lost, interiors fill with material eroded from the rims, and the number of superposed craters increases. Extremely old craters and basins may be reduced to faint circular features of little relief. On rocky planetary bodies degradation is accomplished by various erosion processes, including repeated impact of more recent craters. On icy bodies, basins are degraded by topographic relaxation in weak substrate materials (Schenk 1991).

Earth and Venus preserve a small number of craters because they have young, tectonically active surfaces. Portions of Mercury, the Moon, Mars, and many satellites of the giant planets retain heavily cratered surfaces from the early period of high impact flux. Heavily cratered moons of outer planets display surfaces of dark, ice-rock mixtures which are highly cratered. Small satellites and asteroids are little more than
irregular chunks of rock pitted by repeated impacts. Many are remnants of larger bodies that were broken apart by highly energetic impact with other large bodies.

**Volcanic landforms**

**Silicate volcanism**

Volcanism includes all processes by which material – gas, liquid, or solid – is expelled from the planet’s interior. Most planets were at least partially melted in their earliest history, allowing for differentiation and separation of materials into core, mantle, and crust components. Early volcanism spreads fluid lava across the planet’s surface to form vast, nearly featureless, volcanic plains. The timing and duration of plains formation is largely controlled by the size of the body, which determines how rapidly it cools. On small bodies, this episode ceases early, while on relatively large bodies, basaltic plains formation and volcanic structures persist to late stages of the planet’s history.

Vast low-relief, lightly cratered surfaces of planum and planitia on Venus and Mars and maria of the Moon (Figure 3) are formed of highly fluid flood basalts emplaced after the period of high impact rates. Slightly more viscous lavas build low, broad shields as seen on Venus, Earth, and Mars. Less fluid lava and lava extruded at low expulsion rates tend to build surface constructs as domes, cones, and caldera. The morphology of these features is a function of rates of extrusion, gas content, and silica content. Thus a range of features varying from very low, broad, basalt shield cones to steep-sided, andesite stratovolcanoes and rhyolitic explosion craters is possible. Most planetary volcanoes are simple shield cones similar to Hawaiian volcanoes. At small scales, individual lava flows and flow fronts are distinguishable, and lava channels incised into the surface of some planets can be observed.

Venus displays a wide range of volcanoes (Head et al. 1992), including eruptive shield volcanos (Figure 4), complexly fractured volcano-tectonic structures (coronae), and low, flat pancake domes. Radar-bright, rugged flows are found associated with volcanic constructs and rift zones, and as impact melt from large craters. Because the surface of Venus is so hot, lavas are able to cut very long lava channels across its surface.

Volcanism on Mars is represented by large shield volcanoes, wide, relatively low patera, and volcanic plains (Figure 5). Olympus Mons, 20 km high and 600 km in diameter, is the largest volcano in the solar system. Individual flows and flow fronts are identifiable in several regions.
LANDFORMS OF OTHER PLANETS

Figure 4 Radar image of the Sapa Mons on Venus. Radar-bright surfaces are rugged materials and radar-darker surfaces are smooth materials. Lava flows extend from the shield volcanoes on to the surrounding plains material. The scarcity of impact craters indicates that the surface is relatively young (NASA/JPL Venus Orbiting Imaging Radar). Reproduced from NASA.

Nonsilicate volcanism

Volcanism in the outer solar system is quite diverse. Io, the innermost satellite of Jupiter, has a diameter of 3660 km. Tidal flexing provides a continued input of energy that melts a large portion of Io. The surface is covered with very young flows and more than 400 active volcanoes (Figure 6). The surface consists mostly of plains of overlapping sulfur flows which color the surface in varying shades of yellow, red, white, black, and green due to the many polymorphs of sulfur and its compounds.

Cryovolcanism is common on the icy satellites of outer planets (Kargel 1995). Jupiter’s moon Europa, 3120 km in diameter, has the smoothest surface in the solar system. Its surface is a bright blue-white of water-ice crisscrossed by large fractures. Only a few impact craters survive. The youthful surface is maintained by tidal flexing, which allows a sub-ice water ocean to exist. Fractures in the icy surface allow for frequent resurfacing events. A few large areas of chaos occur in which the surface has been disrupted and large blocks of grooved ice float in a hummocky frozen ocean (Figure 7). Ganymede, at 5260 km in diameter, is the largest of the Galilean moons. The surface appears to be composed of an ice-rock mixture that is dominated by heavily cratered terrain and large patches of young, bright, grooved terrain (Figure 8). The grooved material is formed by floundering of old crust and flooding with liquid water from the interior. The bright satellites of Uranus, Titania, and Ariel are surfaced with fractured ice also.

Triton, the largest moon of Neptune, is 2706 km in diameter and has a very thin
atmosphere of nitrogen and methane. Its surface is poorly resolved by spacecraft imaging. An ice cap of frozen nitrogen covers much of the southern hemisphere. No heavily cratered terrain is preserved on Triton. Highly grooved surfaces were probably formed by repeated fracturing and subsequent filling with viscous volcanic ice or flooding by mixtures of molten ices of water, methane, or nitrogen. The Voyager 2 spacecraft imaged erupting plumes of dust entrained in nitrogen.

**Figure 6** The multicolored surface of Io shaped by sulfur volcanism. Active eruptions are marked by eruptive centers surrounded by dark haloes and by the plume of eruptive materials seen on the limb (NASA/JPL Cassini spacecraft). Reproduced from NASA.

Tectonic landforms

Tectonics includes all processes causing deformation of a planet's crust. On Earth, the movement of convective cells in the mantle leads to shifting of ridged plates on the surface. This system of *plate tectonics* began about 2.5 billion years ago when the crust obtained sufficient rigidity. Interaction along the plate edges is responsible for most of Earth’s large structural features. Plate interiors are only lightly deformed. The smaller terrestrial planets (Moon, Mercury, and Mars) have thick lithospheres that show little tectonic deformation over much of their history. Although plate tectonism is not observed on the other planets, convective movement in the interior and intrusions of igneous material into the crust are responsible for stresses in the crusts. Grooved and furrowed (Figure 8) terrains are formed on the surface of icy bodies by convection cells in underlying oceans (Hammond and Barr 2014).

The most common structural elements observed on the planets and their satellites are *scarps* produced by normal faults attributed to extension in the crust. A few tens of meters to several kilometers in relief, they occur singularly or paired as opposing scarps to form fault troughs (graben). Long, narrow fault troughs, called *rilles*, are common features on many planetary bodies (Figure 4).

Parallel low scarps, narrow graben, and large rift zones are indicators of tension in the Venusian crust. *Rift zones* consisting of complex graben formed of multiple inward-facing scarps are probably formed by extension above upwelling convection cells. Also on Venus, distinctive volcano-tectonic rings called *coronae* are formed above upwelling plumes in the mantle. Large coronae are circular features 200–600 km across consisting of gentle domes with superposed concentric fractures outlining a central depression.
Compressional stress is manifested by folded strata and thrust faults. On Earth, structural disturbance is revealed by deformed strata modified by erosion of materials of different resistance. Compressive stress, primarily at the edge of convergent plate margins, is recorded by crumpling of the crust into a series of tightly folded anti-clines and synclines. The largest mountain chains on Earth, such as the Alps, Appalachians, and Himalayas, are linear fold belts several tens of kilometers to 100 km wide and several hundred kilometers to 1000 km in length. Local folding and doming also is associated with vertical movements of igneous magmas or salt doming in the crust.

Low ridges (mare ridges or wrinkle ridges) a few hundred meters high, a few kilometers across, and 10–100 km long occur on volcanic plains material on the Moon, Mars, and Mercury (Figure 1). Wrinkle ridges typically occur in
LANDFORMS OF OTHER PLANETS

Figure 8  Ganymede is the largest satellite of Jupiter and the largest satellite in the solar system. The surface consists of an older, dark, heavily cratered rock-ice mixture and bright, lightly cratered, grooved terrain (NASA/JPL Cassini spacecraft).

parallel to the edges of circular maria at the junction of the shallow outer shelf and deeper inner basin. They also form rings that mark the position of buried craters. They are interpreted as anticlinal, fault-propagation folds accompanying local thrust faulting created as basalt flows cool and contract. Sinuous lobate scarps on Mercury, Venus, Mars, and the Moon are the surface expression of thrust faults generated by global contraction (Watters et al. 2004).

Wrinkle ridges and folded ridge belts are evidence of compression stresses on Venus (Solomon et al. 1992). Major fold and thrust belts are often located at the margins of kilometer-high plateaus. Intensely disturbed highland plateaus are thought to have been subjected to compression above downwelling plumes in the mantle. Tesserae are complexly folded regions cut by later tensional fractures.
LANDFORMS OF OTHER PLANETS

Gradation

Weathering

Weathering is the process of breaking down surface rocks to smaller sizes. Weathering prepares surface materials for erosion and redistribution by gravity, wind, water, or ice. Weathering may be chiefly physical or may involve chemical changes in minerals by reaction with an atmosphere. Chemical weathering produces new stable minerals in equilibrium with the atmosphere. On Earth and Mars, weathering has produced clay minerals. Weathering on Venus is very different from that on Earth or Mars. The style of weathering varies with altitude on Venus. At high altitudes and low temperatures, sulfur in the atmosphere reacts with lava flows to form the mineral pyrite ($\text{FeS}_2$). Pyrite is highly reflective of radar waves and may explain why mountain peaks and high plateaus of Venus are radar bright. At lower altitudes and higher temperatures, magnetite ($\text{Fe}_3\text{O}_4$) and anhydrite ($\text{CaSO}_4$) may be the stable minerals produced by weathering. In the absence of an atmosphere, physical weathering is primarily in response to changes in temperature and fragmentation by impacts. On airless bodies, space weathering occurs by the bombardment of surficial materials by fast-moving atomic particles and shortwave radiation from the Sun and deep space.

Mass wasting

Mass wasting is the downslope movement of rock and surface debris under the influence of gravity. This process is universal; it operates on all planetary bodies. Movement may be aided by the presence of ice or water, but it is not required. At largest scales mass wasting is observed as landslides and slump scars on crater walls and other steep slopes. Iapetus exhibits long runout landslides, or *sturzstroms*, possibly supported by ice sliding. Massive landslide scars in the canyon system of Mars are characterized by huge concave scarpss on the canyon walls and long runout deposits on the canyon floors. The basal scarp and the *hummocky aureole materials* that surround Olympus Mons are likely lahars or submarine landslides. On Earth and Mars, rock glaciers consisting of angular rock fragments and interstitial ice form lobate masses that creep slowly downslope. Rock glaciers often form in valleys previously occupied by true glaciers. The result of creep is observed on most planets as the accumulation of an *apron of loose materials* at the base of slopes. Individual tracks of boulders rolling downslope are observed in high-resolution images of the Moon and Mars.

Fluvial landforms

Surface fluids

Flowing water or other fluids are an active agent in erosion and deposition. Detrital materials are removed from their site of weathering and moved to regions where the flow ceases. On Earth, liquid water is found in oceans, lakes, and rivers. Driven by solar energy and gravity, flowing water shapes distinctive landscapes. The surface is dissected by a multitude of gullies, streams, and rivers. Intervening drainage divides and basins become the dominant elements of the landscape. The premier landform associated with streams and rivers is the channel eroded by water over time. Channel morphology and drainage patterns on Earth provide insights into both surface configuration of the land and underlying controlling factors of geologic structure. Sediments carried by channel flow are eventually deposited as channel fill, deltas,
LANDFORMS OF OTHER PLANETS

or basin filling materials. Young drainage systems generally exhibit inefficient, nonintegrated drainage; whereas, long-term drainage tends to develop well-integrated systems of channels which are effective in draining large regions.

Early Mars had a thicker, warmer atmosphere than at present. Rain and snowfall delivered water to the surface as did possible outbreaks from the subsurface. Large catastrophic flood channels (Figure 5), sinuous sapping channels, and smaller dendritic systems are found primarily in the highland regions (Mars Channel Working Group 1983). Deltas and alluvial fans are found at the terminus of some channels. Later, a significant portion of water escaped to space and, as temperatures fell, the remaining liquid water became locked up in the pore spaces of rock and soil as groundwater or ice. Since that time it has been released by magmatic heating or meteor impact.

Titan maintains a thick, nitrogen-rich atmosphere. At 5150 km in diameter it is larger than the planet Mercury. Surface relief is low. Presence of liquid lakes, rivers, and oceans is confirmed by images from the spacecraft Cassini. The surface of Titan, described as “complex, fluid-processed, and geologically young,” is marked by broad regions of bright and dark terrain. These include Xanadu, a large, reflective equatorial area about the size of Australia. The convoluted region is filled with hills and cut by valleys and chasms.

Groundwater

Surface water seeps into the ground and moves beneath the surface in pores and fractures. It may re-emerge at the surface in springs and seeps. Water reacts with and dissolves some materials which are moved as ions in solution. Dissolved ions recombine to form new minerals as water evaporates or local chemistry changes to favor precipitation. Newly deposited minerals may form new bodies of rock, such as limestone formed in the Earth’s oceans, or they may form the cementing material that helps consolidate loose sediment.

Solutions of limestone or other soluble rocks by groundwater produces karst topography characterized by caves, sinkholes, and solution valleys. On Mars, chaotic terrain and large outflow channels, such as Maja Vallis (Figure 5), are thought to be formed by the sudden release of large volumes of water from the subsurface to the surface causing collapse of surface materials. Theater-headed, sinuous channels with short, stubby tributaries are cut by groundwater sapping.

Eolian landforms

Wind-related landforms are restricted to bodies that have an atmosphere and loose aggregates at the surface – Earth, Mars, Venus, Titan, and Triton. At low resolutions, eolian processes are seen as albedo streaks aligned with planetary atmospheric circulation patterns. At higher resolution, albedo streaks become windswept, bare rock areas or depositional mounds. Dune fields and individual dunes can be seen. Dune shapes and distribution provide information regarding prevailing wind direction and sand supply sources. Wind erosional features include shallow pits and hollows – blowouts – and elongate, wind-sculpted hills – yardangs. Imaging of active dust storms and dust devils yields information about local, near-surface wind patterns.

In the dense atmosphere of Venus, wind speeds at the surface are very low, but high atmospheric pressure allows for effective transport of surficial materials. Dune fields are identified on a very small percentage of the planet – only 0.004%. Wind streaks and possible yardangs are also identified.
LANDFORMS OF OTHER PLANETS

Even at telescopic resolutions, planet-wide dust storms are visible on Mars. In spacecraft images, wind action is evidenced by wind streaks, sand dunes, sand ergs, and large areas of erosional yardangs. The north circumpolar sand sea (greater than 700,000 km²) is larger than the largest sand sea on Earth (Breed, Grolier, and McCauley 1979). Dune fields are found elsewhere on the planet, including in many craters. Dune types vary from barchanoid sand sheets, barchans, and linear dunes, to small dome dunes. Erosion by wind includes dark wind streaks presumably formed by scouring the surface of loose sediments and regions of exposed rock shaped into yardangs. Dust devils are observed to leave scoured paths on the surface over which they have traveled.

Lacustrine and marine landscapes

Lacustrine and marine-related landforms require standing bodies of water or other liquids. The most common landform, and the one that is most easily resolved, is the lake bed itself. Lake beds are perhaps the smoothest level surfaces to be found on Earth. Lakes may also be inferred by the presence of inlet or outlet channels associated with craters and other closed basins as well as by the presence of deltas. Ocean or large lakes with sufficient fetch have distinctive shoreline features such as bars, spits, and wave-cut cliffs or terraces.

Although channels on Mars exhibit many points of intermittent ponding, overall drainage eventually empties onto the northern plains. An ocean has presumably existed for a time in the northern plains (Baker et al. 1991). The basal scarp of Olympus Mons may be a wave-cut cliff and the aureole deposits around the volcano may be evidence of submarine landslides (Figure 5). Currently, Titan is the only body other than the Earth to have stable bodies of liquid (probably methane and ethane) at the surface.

Ice-dominated landscapes

Glacial and periglacial landforms are seen on planets with atmospheres that contain water and on which temperatures fall below freezing for significant periods of time. A glacier is a large body of ice deformed by the pull of gravity. Continental or polar glaciers cover very large tracks of land; alpine or mountain glaciers are confined to valleys. Glaciers erode the surface over which they move and deposit materials at their terminus as ice melts. Periglacial activity is produced by freezing and thawing of water-saturated loose surface materials. Currently, glacial landforms and polar caps are only seen on Earth and Mars. Morainal deposits, ice wedge polygons, and pingos indicate a period of glaciation in the equatorial belt of Mars.

In contrast to terrestrial planets, many satellites of the outer planets are surfaced with ice or ice-rock mixtures. Some of these satellites (such as Europa and Dione) consist of an ice cover several kilometers thick over a liquid ocean. Many satellites exhibit low-relief, grooved and furrowed terrain suggesting large-scale fracturing and rehealing of an ice mantle (Figure 8). Water escaping to the surface along fractures freezes to form large featureless plains. Those satellites with ice-rock mixtures (such as Ganymede and Callisto) retain craters, although their relief is severely reduced by ice’s tendency to flow (McKinnon 1985). Those satellites with mostly water-ice at the surface have very few craters.

SEE ALSO: Fluvial depositional processes and landforms; Fluvial erosional processes and landforms; Glacier lake outburst floods; Groundwater; Ice sheets; Lakes and limnology
References


Further reading


Landlocked states

Alexander C. Diener
University of Kansas, USA

Though a geopolitical condition known to all political geographers, landlocked states draw direct attention today largely in textbooks and from regional experts making specific case references. Rarely is a systematic analysis undertaken by border studies experts despite the fact that, at present, there are 48 landlocked states in the world including four partially recognized countries. Only Bolivia and Paraguay are outside of Afro-Eurasia. Europe contains 14 landlocked states including several that are rather prosperous by world standards. In contrast, the vast majority of landlocked states are considered “developing.” Thirteen African and four Asian landlocked countries rank among the United Nations’ 48 “least developed states” (UN-OHRLLS 2016). By this count, a third of the world’s poorest states have no coastline. Within the increasingly integrated system of global economics premised on minimal transport costs, lack of access to the sea is a major detriment (Collier 2008). This has not always been the case.

There have been times in history when isolation from the sea was valued as providing a measure of security from a wide range of dangers including hurricanes, tsunamis, pirates, and invaders. Changing transport technologies, however, catalyzed a reimagining of the value of sea access. For example, the relevance of overland routes between Europe and Asia known as the “Silk Road” was dramatically lessened by the advent of large seafaring vessels and the opening of ocean routes to the “Far East.”

Other advances in transport technologies were coupled with modifications to absolute sovereignty in favor of more liberal trade regimes in Europe. By the eleventh and twelfth centuries many European powers were giving access rights to landlocked countries and internationalizing rivers. By the nineteenth century, the heavy tolls and transit restrictions common to previous eras began to give way to recognition that development of industry and commerce required ease of movement. Railroads and other new forms of transport technology reduced the significance of waterways and offered new transit economies and bargaining power to landlocked states (Glassner and Fahrer 2004).

International conventions on transit in the early twentieth century sought to ensure access to the sea for Europe’s landlocked states. One of the more conspicuous efforts involved the creation of corridors through which landlocked countries would access short coastal frontages or navigable rivers. The most famous of these was the Polish Corridor, which proved more a geopolitical irritant than an economic salve. This period also saw important agreements forged concerning the transnationality of railways, pipelines, ports, power lines, and other transport facilities. What failed to manifest, however, was an international provision for freedom of transit to the sea and access to sea resources for noncoastal states.

Building upon the 1921 Barcelona Convention, Article 5 of GATT (General Agreement on Tariffs and Trade) and Article 33 of the Havana Charter incorporated transit provisions into an elaborate system of expanding world economies and bargaining power to landlocked states (Glassner and Fahrer 2004).
trade. The process of decolonization left many new states in Asia and Africa with limited or no coastal frontage. This condition inspired the 1964 Convention on Transit Trade of Landlocked States, which was the first international lawmaking conference to deal exclusively with the question of access to the sea. This convention heightened awareness of the plight of landlocked states and prompted its treatment within the United Nations Convention on the Law of the Sea (UNCLOS III 1973–1982). Though the focus of specific international law, the problems of landlocked states remain unsolved. Transit states retain the capacity to exploitatively or punitively enact sovereignty and thereby restrict resource, corporeal, and commercial movement across their respective territories.

Renewed efforts to improve the lot of the world’s landlocked developing states took form in the 2001 Millennium Development Goals and the 2003 Almaty Programme of Action (APoA). Integrating a wide range of actors (landlocked developing states, transit states, donors, private sector, international and regional organizations), the APoA places emphasis on five priority areas: (i) fundamental transit policy issues; (ii) infrastructure development and maintenance; (iii) international trade and trade facilitation; (iv) provision of development assistance and technical support from the international community; and (v) implementation and review. In the period since the APoA’s launch, development levels of landlocked states have improved (UN-OHRLLS 2013). Better roads, enhanced railways, larger ports, containerization, and cheaper fuel have opened vectors of transit previously unavailable, but establishing neoliberal regional institutions that would unencumber transit for landlocked states is far from a fait accompli. Moreover, to what degree such institutions can be achieved equitably and in an environmentally responsible manner remains a significant question.

While Switzerland stands as the foremost example of a successful landlocked state, the reality of most is more akin to the Kyrgyz Republic. Limited resources, politically assertive neighbors, black and gray market detractors from the national economy, and an unrelenting dependence on other states to facilitate exports and imports render the Kyrgyz Republic vulnerable. Like many other landlocked states, it possesses limited options in terms of geoeconomic policymaking and restricted prospects for development. At present no “right of innocent passage” exists on land as it does on the sea. Until it does, landlocked states will continue to be challenged by the territorial sovereignty of their neighbors.

SEE ALSO: Accessibility, in transportation planning; Borders, boundaries, and borderlands; Cross-border, transnational, and interregional cooperation; Maritime transport; Mobility gaps; Sovereignty; States and development; Sustainable transport; Transport and development; Transportation history

References


UN-OHRLLS (Office of the High Representative for the Least Developed Countries, Landlocked

Further reading


The term landscape ecology (in the original German *Landschaftsökologie*) was first used in 1939 by the German geographer Carl Troll as the name for a particular way of looking upon the landscape (*Anschauungsweise*) based on a holistic perspective integrating geographical and ecological insights into the nature of terrestrial environments (Troll 1939). Troll's previous research into the coupled relationship between vegetation, environment, and land use had brought him to conduct extensive fieldwork registrations of vegetation patterns in landscapes in Northern Europe. These experiences, in turn, had inspired him to take up aerial photography as an instrument to identify and describe spatial units of vegetation cover, the heterogeneous pattern of which he was then able to relate analytically to both social and environmental processes of change. On these foundations, Troll proposed a science of landscape processes that would be based on the new technology of aerial photograph interpretation, to classify the Earth's surface into discrete land units. On the basis of such inventory classifications, it was possible for Troll and his contemporaries to integrate analytical perspectives from a range of otherwise discrete sciences in their analysis of landscape processes. In this way, the development of landscape ecology, characterized by its specific focus on the ecological significance of spatial form and pattern, was closely associated both with the novel perspective provided by aerial photography and also with older geographical perspectives of analysis inherited from classical cultural and physical geography.

Troll had been trained in the traditions of geography defined by Alexander von Humboldt and Carl Ritter, the two great initiators of the field as a modern science in the first half of the nineteenth century. From them and their immediate successors he had picked up a certain breadth of scope and an interest in the interconnected nature of geographic phenomena, which he emphasized in his writings. He had studied the work of Humboldt and Ritter extensively and had been impressed especially by their ability to correlate natural and cultural phenomena. From Humboldt’s scholarship he had understood the need to abstain from disentangling research objects from their environment, but rather to delve into the multifaceted empirical relationships linking otherwise seemingly disparate elements with larger patterns of distribution and causality. For Troll and his contemporaries, it was never enough to understand the nature of a species or habitat type in its own right – the aim was to explain how species and habitats interact with each other and other facets of their environment, with reference to actual, empirical histories of change.

On this basis Troll was increasingly critical of the growing specialization of academic
geography in his day. Since the late nineteenth century the universalist approaches advocated by Ritter and Humboldt had been superseded in mainstream scholarship by an outspoken (and in Troll’s view unwarranted) reductionist stance towards the organization of scientific knowledge. Researchers within geography had sought to establish biological, physical, and demographic modes of analysis as distinct fields of research, and within these fields it had become the norm to seek for the smallest possible and most basic units of analysis in order to be able to study each environmental subprocess individually, as if in a laboratory. Troll saw this development as an understandable but tragic detour from the unified perspectives of earlier geography, and one that made it increasingly difficult to correctly observe and describe the cohesive, interdependent nature of geographic phenomena. Therefore, Troll sought to conserve and reformulate a unified perspective on terrestrial surface processes. Hence, landscape ecology from the outset was conceived as an approach to environmental research that was to be unrestrained by disciplinary and methodological dogma.

When, in 1959, Troll was asked to address the geographers of the British Association for the Advancement of Science on the centenary of the death of Humboldt and Ritter, he made his views on contemporary geography remarkably clear. In his speech he located the inspiration for a range of seminal landscape ecological concepts and instruments within the work of his two academic idols. These included: forms of spatial pattern analysis, which Troll became inspired to apply and further develop upon reading Humboldt’s pioneering work on ecological zones and associated patterns of plant distribution; the method of regionalization developed by Ritter, which Troll employed to classify heterogeneous areas (i.e., landscapes) into analytical units based on similar form, genesis or character; and types of early socioecological analysis developed by Ritter through his work on the relationship between land use, culture, and civilization (Troll 1960).

These concepts and instruments came to inspire the development of landscape ecology, as a counterbalance to the fragmentation or disintegration of geographic–ecological thought in the twentieth century that motivated landscape ecologists to establish their field as an explicitly integrative science. In the 1950s and 1960s, landscape ecology consisted of a synthesis of geography with soil science and ecology. From the late 1960s especially, island biogeography increased in importance (MacArthur and Wilson 1967). Parallel to this, a corresponding, spatially-oriented vegetation science developed, and among conservation-oriented zoologists a strong school of dispersal ecology and metapopulation theory developed (Gilpin and Hanski 1991). The methods and perspectives of these fields were combined by applying them to the same spatially defined land units at various scales – from single landscape elements or land units to larger complexes of functionally and historically associated sets of elements, up to whole landscapes defined by heterogeneous patterns of landscape elements.

Figure 1 illustrates the type of method for spatial analysis that is typically used. It allows researchers to organize the many multidisciplinary insights flowing into landscape ecology into a common empirical understanding of landscape dynamics, by referencing all relevant explanatory understandings to a common set of land units and the landscapes they constitute. In this way it becomes possible to examine the spatial distribution and overlap of explanatory understandings spatially based on map analysis. As the field developed, this
Figure 1  Landscape patterns analysis typically starts with a mosaic of vertical aerial imagery such as the sample shown here (left). Through imagery analysis supported by fieldwork observations, the remote sensed data are transformed into a classified map of internally homogeneous land cover units (center). Considered as a whole, the land cover units form a set of heterogeneous landscape patterns (right), each indicating a particular processual relationship between life forms on the Earth surface (including humans) and their geoeccological basis. The sample shown here illustrates how rural land use is adapted to the geoeccological conditions in an agricultural landscape near Skive in Northern Jutland, Denmark. In the north there is an intensively used and closely settled agricultural landscape dominated by arable fields situated on well-drained loamy soils (A). A narrow erosion valley covered by grassland habitats intersects the arable land from the north, draining into a broad glacial valley of low-lying waterlogged soils covered by paddocks and remnants of moorland at the center of the area (B). South of the central valley there is a relatively complex pattern of agricultural land use, which is adapted spatially and functionally to the heterogeneous soils of the area, characterized by a variety of gravel, sand, and clay deposits (C & D). Analytical approaches of this kind, which link spatial pattern with landscape functionality and landscape history, are used within landscape ecology to integrate perspectives from ecology, soil science, cultural geography, sociology, and the humanities into a common, spatially explicit framework of analysis. On this basis, otherwise potentially incongruent theories explaining landscape dynamics can be tested and compared. Source: Imagery recorded by the Royal Airforce in 1980, held by the Royal library of Denmark.
and postcolonial landscapes. These changes were pioneered in Eastern Europe, where they were associated with centrally-planned collectivization of agricultural properties into specialized, large-scale agroindustrial units. In Western Europe and North America, agricultural industrialization developed later, in most cases driven by a combination of government subsidies and competition among producers, so landscape ecology as an applied scientific practice developed more slowly and later here than in Eastern Europe (Ruzicka and Miklos 1990).

Cultural landscapes and holism

The growing influence of applied perspectives meant that landscape ecologists came into sustained conflict with traditional ideals of scientific practice, leading some researchers to define their field as an explicitly action-oriented, solution-driven practice integrating scientific discovery with direct intervention in landscape management practices. This has remained a widespread perspective within landscape ecology, which now includes a comprehensive array of planning and policy-related fields of research. In addition to these developments, landscape ecological research has also extended in the direction of historical and social disciplines. This has taken place gradually as researchers within the field have cultivated an increasingly integrated relationship with scholarship from the human and social sciences, which was considered a critical component to include within landscape ecology to fully appreciate the complex trajectories of change in human-dominated cultural landscapes.

Cultural landscapes have remained a dominant research object throughout the history of landscape ecology, which has progressed from descriptive accounts of human–landscape interaction to more intricate attempts at understanding, conceptualizing, and quantifying the nature of socioecological relationships in a landscape context. Such research demands a highly developed model for integrative research, because human landscape management is a process that mediates both between the ecological and social sides of human existence and between natural and anthropogenic facets of ecosystem functioning. As the German geographer Ernst Neef, one of the founders of socioecological research within the field, expressed it: “The transfer from societal changes to the natural systems is based on the spontaneous effect of the laws of nature, whereas the transfer from the natural area to the regulation of natural processes is achieved by cultural forces, human perception and decision making” (Neef 1984, 6). As such, cultural landscapes are a human construct and a biophysical system at the same instance, and any attempt to analyze, theorize or interfere with landscape dynamics must be able to account for both of these dimensions. Within landscape ecology the challenge of accounting for these two sides of landscapes in an integrated way has motivated the formulation of a landscape concept where the landscape is seen as a holon: an assemblage of interrelated phenomena that together forms a complex whole, which is “more than the sum of its elements” because “all parts are internally related to each other by the general state of the whole” (Naveh 2000, 11).

This perspective has become a characteristic feature of landscape ecological research. It explicitly challenges both the basic division between anthropogenic and natural phenomena, as well as the mono-disciplinary reductionism prevalent in mainstream science, by claiming that the socioecological processes constituting landscape holons and their constituent parts cannot be understood without accounting for
LANDSCAPE ECOLOGY

the multifaceted relationships that unite individual elements into a spatially and functionally integrated whole.

In terms of empirical work, such relationships have been investigated by focusing on the processes or “functions” linking ecosystems with each other and with human agency. In this way, empirical evidence of the organization of landscape holons is collected by characterizing and mapping the extent, magnitude and coupled interplay of ecosystems processes (ecosystem functionality), social processes (societal functionality), and intentional practice (transcending functionality) within landscapes (Brandt and Vejre 2004). By integrating intentional human practices related to human culture, cognition, belief, planning, and decision-making with respect to landscapes into the field on equal terms with other more tangible types of functionality, landscape ecologists were able to outline a broad-ranging holistic science devoted to examining the full interplay between socioecological processes and biophysical patterns in landscapes (Nassauer 1997).

Aerial vision: a new perspective in the study of landscapes

From the outset, landscape ecology was characterized by a visual approach to analyzing landscapes from a bird’s-eye perspective, making it possible to detect minute nuances in land cover patterns, while at the same time retaining a distanced landscape-scale overview. This perspective was made possible mainly due to the new technology of vertical mono- and stereographic aerial photography, which became gradually more widespread in the period after World War I. The technology of aerial photography had been pioneered with recordings from balloons over Paris in the 1850s and in Boston some years later, but due to navigational challenges the method proved unsuited for systematic recordings of large areas.

It was the advent of airplanes in the early twentieth century that kicked off what was to become an explosion of vertical aerial photography. At first, images were recorded in limited numbers and with rudimentary instruments, but technological development and funding for recording campaigns soon became fueled by public investment as the potentials for military use of the imagery became evident. During World War I recording efforts grew rapidly. Mobile laboratories for photographic processing became available in 1915, making it possible to analyze images swiftly at locations along the front lines. This meant that aerial photography became a critical technology to the war effort, and the increasing entrenchment of the western front raised a demand for intensive small-scale scrutiny of long stretches of frontline, to a point where at the height of the war the French side alone was processing more than 10,000 images each night. In the period after World War I, images and recording technology became available for other uses, forming a growing resource for landscape research, and a broad spectrum of researchers, planning, and policy professionals picked up the new technology.

The idea of an all-encompassing bird’s-eye view of the Earth’s surface goes back to antiquity and cartographic representations of the Earth as seen from above formed part of established practices within the scientific community long before airplanes made it possible to actually see the Earth’s surface from above. In spite of this, however, it had an unanticipated effect on landscape research when actual remote sensed imagery became available. Earlier map-makers based their delineations of areas, lines, and point features on fieldwork observations collected on the Earth’s surface, which were then transposed
onto a spherical or flat medium in order for patterns of landscape elements to be made available for visual inspection “as if” seen from above. In this sense, all earlier maps reflect a form of inquiry where research design, sampling, and data collection precedes the rendition of raw data spatially. In such mapping processes, the data becomes spatial in coverage only by way of analysis, not beforehand.

When remote sensed imagery eventually became available, it represented an altogether different type of resource for map-makers than previous types of spatial data. Mosaics of aerial imagery offered researchers a continuous field of spatial data to begin with, making the Earth’s surface an observable research object in its own right, rather than a canvas on which to drape existing observations. And when series of images of the same areas at different points in time were recorded, it also became possible to classify processes of landscape change by way of direct overlay analysis. This was the key condition for the development of early landscape ecological analysis methods, and for the particular concept of landscape processes that still characterizes the field.

In the period since World War II, the range, coverage, and resolution of remote sensed imagery available for scientific analysis has grown rapidly. The Cold War, which succeeded World War II and culminated with the formal dissolution of the USSR in December 1991, meant that American and Soviet intelligence gathering efforts were in constant preparation for armed conflict. New, more precise recording devices and new platforms for carrying them, such as the U2 high altitude jets and the Corona and Landsat satellite programs, are examples of the tremendous development of remote sensing technology that took place. When imagery from these programs was gradually declassified and older platforms began to be replaced by new in a succession of technological improvements, a broad range of new materials were added to the repertoire of data available for landscape ecological analysis. This spurred a rapid development of spatial analysis tools within the field, supported by the new technology of user-centered computing, which became available from the 1970s onwards.

As such, landscape ecology became one of several arenas where geographical information systems (GIS) and associated geostatistical tools were pioneered and developed. In conjunction with these efforts, an array of computational approaches and indexing-methods for measuring and comparing the configuration, complexity, and diversity of land cover patches was developed, including the important Fragstats software package that empowered individual researchers to conduct quantitative spatial analysis (McGarigal and Marks 1995). This initiated a shift in landscape ecology away from analogue pattern recognition towards the use of computational tools able to quantitatively assess landscape patterns and associated processes across vast expanses of land.

**Current methods**

Current landscape ecological methods involve a combination of fieldwork and analysis of imagery to identify the relationship between patterns and processes. Pattern in this sense refers to heterogeneity in the horizontal dimension of the landscape, i.e. differences between land units, while ecological processes typically transpire within a single land unit and refer to the vertical relationships between organisms and their abiotic environment. The various approaches within the field relate to either of these two parallel analytical perspectives: (i) the chorologic, dealing with the horizontal patterns and processes of
land units on the Earth surface, and (ii) the topologic, dealing with the vertical processes of energetic, informational, and material exchange between organisms within their habitats (Neef 1963; Zonneveld 1989).

In a chorologic perspective the primary dimensions of the landscape are its geometrical and temporal characteristics. Here landscape dynamics are approached analytically by measuring alterations in the shape, extent, and distribution or pattern of land units in the landscape, and by analyzing exchanges occurring between individual units. Trajectories of landscape change through time can be appraised by determining the rate of change, its frequency, and magnitude (Antrop 2000).

In a toplologic perspective the character of a land unit is investigated by examining its internal functioning as defined by the societies of species within it and its geoecological potentials in the form of substrates, climate, and hydrology, including the flows of information, matter, and energy taking place through vertical vectors within it.

Today, the patch-corridor-matrix model introduced by Richard T.T. Forman and Michael Godron has become the most widespread conceptual model with which to approach pattern-process relationships (Forman and Godron 1986). It consists of a spatial language designed to describe landscape patterns and related processes. The basic idea of the model is that landscapes are made up of a mosaic of patches (areas differing from their surroundings), connected by corridors (strips of land that infiltrate the landscape and support flows of information, matter, and energy) in a matrix (defined as the dominant, most extensive, and coherent landscape element type) (Forman 1995). In this perspective a landscape is defined by the pattern formed by patch, corridor, and matrix elements repeated throughout its extent. By measuring the size, shape, and distribution of these three types of elements, landscapes can be compared quantitatively in a number of ways. Key parameters include the connectivity, diversity, and composition of landscapes, which have been shown to be systematically associated both with types of human land-use practices affecting landscapes, with biodiversity, and with ecosystem functioning within patches.

Landscape ecology today and its challenges

Landscape ecology has grown to become a widely recognized approach to environmental research, distinguishable from other scientific traditions by the type of spatially explicit, interdisciplinary, and empirically-focused analytical perspectives detailed above. By 1980 researchers from across the world had become involved in the development of the field, with a world conference convened in the Netherlands in 1981. The following year the International Association for Landscape Ecology (IALE) was established as a focal point for dialog and mutual exchange of ideas by organizing regular conferences and through publication of newsletters and journals. These organizational activities take place at international, regional, and national levels through local chapters of the association. They have proven to be instrumental for the further development of landscape ecology, because the local working environment of many landscape ecologists is delimited by traditional mono-disciplinary or sectoral organizational structures and associated domains of expertise within the academic community. Therefore, most landscape ecological research relies on the application of a combination of skills and insights from researchers and practitioners who retain their commitment to one or more disciplines.
while contributing to landscape ecology through cross-, inter- and transdisciplinary cooperation. On this basis, a substantial literature on how to facilitate interdisciplinary exchange and organize action research has developed within landscape ecology, which has come to form a significant scientific contribution in its own right (Tress, Tress, and Fry 2005).

But, at the same time, landscape ecology is also haunted by its fragmented character, and one of the most persistent threats to its continued success is the challenge of translating concepts and methods that have been harvested eclectically from other disciplines into a common analytical framework. These efforts are made increasingly difficult by the fact that the field itself has become subdivided to some degree, because researchers from different parts of the world are basing their analytical efforts on differing theories of science and validity, and thus also on different criteria for scientific achievement within the field. Today, mainstream American and European approaches differ in research priorities and theoretical perspectives for example, which is reflected in books designed to provide an overview of the field from either of the two perspectives. The introductory text by Bastian and Steinhardt is a European example, while the reader compiled and commented by Wiens et al. is an American parallel (Bastian and Steinhardt 2002; Wiens et al. 2007).

In a global perspective, the main differences between perspectives within landscape ecology, including those between American and European scholars, have tended to reflect underlying disparities concerning: (i) the way in which human interference or engagement with landscape processes is handled analytically and theoretically; (ii) the degree to which researchers lean towards epistemologies derived from the human sciences and/or the natural sciences; and (iii) the degree of practical engagement with landscape management and policymaking.

When landscape ecological research is reviewed along these lines it is clear that a particular disparity within the field tends to inform and support the others, namely that between approaches advocating theoretical pluralism on the one hand, and approaches characterized by exclusive theories of truth and validity on the other. This appears to be a dividing line that cuts across other variations within the field, and it has become particularly clear with respect to research identified either with a strict positivist stance or with a constructivist conception of truth.

This may illustrate that while landscape ecologists have succeeded in reconciling their multifaceted perspectives on the landscape when dealing with tangible empirical problems, policy advice, and action research, it has proven more difficult to resolve differences pertaining to the variety of underlying, a priori theoretical foundations within the field. Landscape ecology was united not by a common theoretical debate but rather by a set of methods and research interests, and a fascination with a common empirical domain. At times this has been a great advantage to landscape ecology, which has been able to include and cultivate a host of different perspectives, but it also makes it a challenge to uphold a unified perspective able to make good on Carl Troll’s original ambition to understand landscapes holistically.

This would entail the further parallel development and amalgamation of three existing subject areas within landscape ecology.

1 Research focusing on the basic, persistent structural and processual character of landscape types and processes of landscape change, which provides insights into the long-term natural history of landscapes, enabling society to better understand and
adapt to general conditions for land-use management.

2. Studies of historic and actual, anthropogenic and natural landscape structures and their development through time, which illustrate past and present relationships and barriers within landscapes, supporting improved understandings of how to conserve and/or improve valuable landscape resources.

3. Studies of the cultural and ideological dimensions of human engagement with landscapes and associated types of management practices, making it possible to understand the sociocultural and political background for sustainable landscape management.

The main challenge for landscape ecology, therefore, is not to focus or broaden its scope, but rather to improve the way in which insights gained in one area of the field are communicated, compared, and combined with insights from other areas.

SEE ALSO: Areal differentiation (or chorology); Biogeography; Cultural geography; Geography and the study of human–environment relations; GIS: history; Habitat destruction and fragmentation; Interdisciplinarity and geography; Landforms and physiography; Landscape

References


Troll, Carl. 1939. “Luftbildplan Und ökologische Bodenforschung: Ihr Zweckmäßiger Einsatz Für


**Further reading**


Landscape iconography and perception

Yves Luginbühl
National Scientific Center of Research, France

For a long time it has been recognized that the evaluation of landscapes is based on the weight exerted by pictorial representations on the social imagination. It is undeniable indeed that the painting of landscapes, the pictorial genre of Europe that became established as an autonomous artistic movement with regard to religious painting, had a considerable role in our way of thinking about landscape. The painters, first of all, especially Italian, Flemish, French, English, and Spanish ones, built landscape models that were imperative in their representations of the European as a way of qualifying the landscapes, of appreciating them with regard to the reality of the landscaped forms offered to the eyes of European societies (Cosgrove 1998).

But the emergence of questions asked by the evolution of landscapes, in Europe as well as in the rest of the world, turned upside down the conceptions that dominated the thinking about landscapes: the transformation of landscapes was the object of conflict between diverse social groups and, little by little, from the 1970s, a new conception appeared that separated the seen landscape from the real-life landscape and gradually imposed the notion of the everyday landscape. This entered the legislative texts and, in particular, the European Landscape Convention adopted in October 2000 in Florence and, as of June 2013, ratified by 38 member countries of the Council of Europe (www.coe.int).

The everyday landscape means that the fundamental objective of landscape policies is one shared by the great majority of the population of the planet, whereas for several decades these policies had been centered on remarkable landscapes, in accordance with national measures peculiar to particular countries or in accordance with the 1972 World Heritage Convention of the United Nations Educational, Scientific and Cultural Organization (UNESCO). Nevertheless, these remarkable landscapes still take up an important place in public policies, because they catch the eye of the elected officials who try to highlight, with a kind of territorial marketing, their region or singular sites, sometimes considered exceptional, as well as landscapes experienced daily by the populations who live there.

These social representations of landscapes diversified, the American continent building its own model, after the first European models were strengthened; for a long time even Asia pursued its way of fengshui, today gone out of use and subjected to the new rules of the political systems of the emerging countries of the Far East. Social evolution, the new relationships of societies with nature, the development of mobility, and environmental problems have considerably modified social relationships with landscapes. New models appear, without the old completely disappearing, making the understanding of contemporary thought on the landscape complex. It is, however, necessary to understand what perceptions and social representations of landscapes mean before proceeding to an historic reading that will allow us to approach the overlapping and interweaving of the models.
LANDSCAPE ICONOGRAPHY AND PERCEPTION

Landscape perceptions and/or social representations

While they have different meanings in different geographical and scientific contexts, terms of landscape perception or social representation are used in common or administrative language without distinction. Perception is centered on the individual and his or her apprehension of the landscape through feelings and the senses, that is, through sight, hearing, touch, taste, and smell. For example, the ringing of cowbells evokes a landscape of the Alps, and the call of birds evokes a littoral landscape. An example of feelings might be the enjoyment of contemplating a beautiful landscape, or a landscape that one’s own work contributed to creating, or anger at seeing a degraded landscape, or sadness inspired by a landscape because it reminds the observer of a painful event in their life. In British and American scientific circles perception was used to explore the sensations that evoke landscapes of green countryside or dry country or an urban street. These methods gave interesting results, ranking the analyzed landscapes in order of satisfaction in a register of felt sensations: this landscape is “hotter” than that one, one evokes coldness, another agitation and noise, and so on (Luginbühl 2008).

Social representations of landscapes, used more by authors of southern Europe and Romance-speaking countries, express what landscapes represent for a specific social group: in other words, collective representations. A landscape can represent a place to which we are attached, or, on the contrary, an unknown, distant place, or a place where we had a difficult experience, such as that recalled by an inhabitant of a municipality in Central France, evoking a landscape of a forest of fir trees that reminded him of his captivity in a German camp during World War II. Social representations of landscapes are structured according to “scales of representation”: three scales relate to different images or meanings.

The first scale, which can be considered “global,” refers to “landscape models,” symbolic outlines that were developed in the course of history by the evolution of social relationships with nature. Ancient literature proposes at least one, well known to the readers of Virgil: the bucolic or pastoral model evokes a more or less wooded landscape where the shepherd guides his herd; it can be found in Psalm 23 of David’s hymn: “The Lord is my shepherd; I shall not want. He makes me lie down in green pastures. He leads me besides still waters. He restores my soul. He leads me in the paths of righteousness, for His name’s sake.” This landscape is also the one where man experiences a feeling of peace and justice, a sensation of freshness. Among the bucolic or pastoral landscapes, we can cite for example the landscapes of the English countryside with its enclosures, the western landscapes of Europe where the meadow dominates, or the landscapes of “dehesa,” the agro-forestry system made up of open forests of holm oaks or of cork oaks, as can be found in Andalusia or Estremadura.

The second scale of social representations refers to the landscapes lived locally by the population of a place, such as a city or a small area, and refers to the culture of the place (some specialists evoke the “spirit of the place”), that is, the empirical knowledge of the inhabitants of the place, which is not unanimous and which also relates to local and sometimes conflicting social relationships. It is also about the culture of the natural environment, that is, the knowledge, sometimes passed on orally by the inhabitants, of how nature has evolved locally. Planning professionals can resort to it by questioning elders to find out dangerous places, such as an avalanche corridor in a mountain, the hillside of a valley that is subject to landslides, or an easily
flooded river bed. It is social memory that is thus questioned, as empirical knowledge that can complement scientific knowledge.

The third scale is that of the individual: each individual builds his or her own experience of landscape from his or her culture of landscape, and it is different from one individual to another and not reproducible. It asserts at the same time the uniqueness of the human being and his or her own social identity. A father and a mother do not have the same landscaped culture as their children; however, this scale, although it has an undeniable interest in the theoretical field of knowledge of social representations of landscapes, is not useful for this inquiry, precisely because of the diversity of individuals and the uniqueness of the human being. Only the global and local scales can be interpreted by the inquiry, because they refer to symbols and collective meanings.

These three scales allow us to understand the complexity of social representations of landscapes and especially their sometimes fuzzy, ambiguous, or contradictory nature; the same individual can indeed build different representations of the same landscape, because of the interweaving of scales. A landscape can be considered beautiful because it is part of one of the landscape models previously mentioned, but it can also evoke a situation of injustice for an individual or for a certain social group: for example, some inhabitants of the Beauce, a French region of the Paris Basin, a vast cereal plateau, rather flat and not very wooded – only a few groves break the dullness of the landscape – consider that this landscape is too uniform, not very pleasing, but at the same time, they mention the effect produced by sunbeams at the end of the afternoon, giving a golden color to the wheat; either they compare this landscape to an ocean (as Emile Zola in La Terre 1980/1887) where the big isolated farms are similar to ships and the villages to archipelagoes, the landscape then becoming sublime, or the landscape is seen negatively because of the still recent memory of the times when farm workers were exploited by the local country squires of the region.

This aspect of social representations emphasizes the values attributed to landscapes: these indeed relate to positive or negative values (Figures 1 and 2). They appear in the language used by individuals to qualify landscapes: harmony or beauty and freedom, or conversely, ugliness or disharmony and constraint. But each of these values has diverse dimensions: esthetic, social, ecological. Harmony can mean esthetic harmony (that is, of form) but also social harmony (that is, harmony between the people and readable in a landscape). Both ecological harmony and harmony between people and nature can be read in the landscape of a clear, nonpolluted lake. Conversely, esthetic disharmony is an ugliness of form, but it can also involve social disharmony, such as that evoked by certain individuals in urban “cities” where crime or violence reigns; ecological disharmony is that which man can observe in polluted landscapes, filled with the smoke of exhaust fumes from cars or factories.

Furthermore, these social representations of landscapes are not unchanging: they evolve with change in society and with the change of relationship between societies and nature. For several centuries, in Europe landscape was indeed linked to the countryside produced by peasant farming – a normal situation for a continent in which the economy was essentially, until the eighteenth century, based on agricultural production. Since the 1990s, this conception has changed, agriculture becoming industrial and productivist. Since then the landscape has been more associated with nature, but not just any nature: it is about distant nature, that of Amazonia, the African savannas, or the taiga, landscapes where the individual supposes that
nature has not been degraded by the mindless practices of people. The American situation is doubtless different, for this continent does not have the same history: the American landscape is more evocative of the “wilderness.” However, in Latin America, the import of European models qualifies this situation, because certain regions were transformed by colonists from Europe,
such as the south of Chile, where Germans reproduced landscaped forms close to those of Bavaria, or Uruguay, where a Swiss colony created the “Nueva Helvecia.”

The situation is different still in Asia, in particular in China where landscape was for a long time created according to the theory of *fengshui*, which expressed the telluric strengths stemming from mountains and the ordering of valleys by the currents of waters arising from them: a house is positioned on a mountain hillside to overlook the valley where agricultural activity is concentrated. This situation was disrupted by the implementation of strong town and country planning policies, corresponding to the strengths of an economy that was both collectivist and capitalist.

The role of art in the construction of social representations of landscape

As previously described, the global scale of social representations of landscapes was powerfully influenced by pictorial and literary production, but also, and more recently, by photography and cinema. Some authors believe the term “landscape” to come from the artistic field, although it most probably originates in town and country planning in the fifteenth century. However, art played a considerable role in the elaboration of the landscape models that structure the global scale of social representations of landscapes. It is also true that terms equivalent to “landscape” appeared at the end of the fifteenth century when the painting of landscape released itself from religion and when it became an autonomous pictorial “genre.” It can be asserted that landscape began to dominate religious scenes and that it displaced representations of the deity, saints, and so on at the time of the Renaissance (Sereni 1955; Figure 3). The first pictorial representations of landscapes always involve religious scenes with inspirational landscapes, whether they are the supposed Paradise or symbolic places where religious events took place. If, for a long time, painters did not dare to represent landscapes of big nature in a fantastical way, it was the rural, bucolic, or pastoral landscapes that dominated, as in the painting of Le Lorrain, or landscapes where the abundance of fruits and produce reigned – a kind of land of plenty, as we find in Breughel’s painting for example.

The domination of these landscape models – bucolic, pastoral, or of plenty – can be understood in the economic environment of feudalism when the fundamental activity was the production of goods essential to the survival of the essentially peasant societies. But the medieval agricultural system gave priority to the production of cereal for bread-making while animal
breeding was a by-product of the cultivation of wheat, rye, or barley. The herds were for the most part left to themselves or led by a shepherd—frequently children—and rambled in search of grass, which was rare because people did not yet know how to cultivate it; pastoral spaces were mostly collective lands. The agronomists of the Renaissance strongly insisted on the need to reform this kind of development of the landscape by giving priority to breeding, animal protein being infrequent in the food of medieval populations and for a long time afterwards. We can then understand the domination, in the painting of this period, of bucolic or pastoral landscapes (Figure 4), which represented a kind of hope for improvement in human food.

It is in the eighteenth century with the reform of systems of agricultural production that pictorial representations of landscapes decline. England was a pioneer: English agronomists invented the cultivation of forage crops (ryegrass, fescue, lucerne, clover, etc.) which allowed the establishment of modern breeding with fenced artificial meadows where animals, oxen and sheep, were enclosed without the need for a shepherd (Hoskins 1955). It is the moment when England becomes the first economic power of the world, this revolution in fodder coming at the same time as the industrial revolution and the exploitation of mines, the production of steel, and the invention of the steam engine. It symbolizes the victory of humans over nature; it makes it possible to surpass nature’s strength and speed thanks to the laws of thermodynamics discovered by science. This is when the feeling of superiority of man over nature inspires the landscape model of the sublime, a new model that also evokes the practice of mountain climbing, as represented in Gustave Doré’s paintings or in the countless others paintings of mountain landscapes that appeared everywhere in Europe at that time. The nineteenth century is the time when pictorial representations of mountain landscapes, and then of swamps or coasts, dominated. They had rarely been painted before then because they symbolized the horror of nature, the danger and the risks incurred by travelers venturing into these places. At the same time, they represented the fulfillment of the Promethean dream of men: to be the master of nature.

The picturesque (i.e., that deserving to be painted) is elevated to the sublime, but it becomes attached to the most charming sites, prompting the first tourists to travel across Europe, in particular on the “Grand Tour.”
new picturesque landscape model is the tasteless counterpart to the sublime; it has considerable success with the bourgeoisie in Europe, who take advantage of the enrichment of their social category thanks to the development of the industrial and commercial economy. But if the economic dimension is present in the painting of landscape, another aspect asserts itself: its political dimension. It was already present, particularly in the Renaissance, when, in Italy, the big patrician families from Tuscany, Lombardy, and Venetia inspired essential changes in territorial governance in accordance with their economic interests. Ambrogio Lorenzetti’s famous fresco The Allegory of Good and Bad Government in the ducal Palace of Siena reveals perfectly this political dimension by actually evoking the territorial governance of the province over which political power reigned. The political meaning of landscape also appears in some English paintings representing the landscape of enclosures, such as an anonymous painting on four boards showing the individual appropriation of ground, the foundation of liberalism; or Gainsborough’s painting “Mr and Mrs Andrews,” showing a husband and wife from the landed gentry posing in front of a landscape of enclosures. In reality, this political dimension is often present, but it was rarely noticed by specialists except by some British or American authors (Mitchell 2002).

Social and pictorial representations of landscapes thus maintain powerful associations. The first cover the entire register of social sensibilities in landscapes, while the second constitute symbols that give a political, economic, ecological, and esthetic meaning to landscapes (Luginbühl 2012). Today, the political, social, and ecological dimensions acquire a new meaning, inspired by the desires of society to contribute to political decision-making; the United States ushered in these approaches with John Dewey. From hereon in, for societies it is a question of taking control of the fate of their living environment, the landscape of everyday life.

**SEE ALSO:** Cultural geography; Landscape; Scale

**References**


**Further reading**


Landscape

Mitch Rose
Aberystwyth University, UK

It has often been said that the field of landscape studies (LS) has an overinterest in, some would say an obsession with, its own history. It is a charge that is difficult to contest, given the number of monographs, edited volumes, and articles that have been dedicated to the term’s historical evolution and development. And, yet, it is ironic that, despite the amount of interest there is in landscape as a concept, there is relatively little work on landscape as an empirical phenomenon. Indeed, while geographers have written extensively on the history, development, and influence of LS, as a body of work, actual studies of landscape are relatively few relative to other geographic objects such as the city or the region. In this sense, it seems that LS punches above its weight; its “old man” position in the discipline allows it to operate as a touchstone or barometer for cultural geography, even as its actual subject matter remains underexplored. There are a number of positive and negative consequences for LS in this regard. No doubt it is positive for LS to be seen as one of the many crucibles of innovation in the discipline. But the lack of empirical work has engendered a somewhat conservative subfield. It is precisely the empirical exercise that creates innovation, albeit in a minor key. Thus LS’s tendency for major revolutions in theory, rather than minor transformations through empirical work, has kept the concept somewhat definable within particular schools – spheres of research that signal both a specific historical period and a distinctive modality of doing landscape research. While other geographic topics such as place and region have been associated with various schools, the surfeit of empirical work on these topics has forced many more theoretical iterations and thus suggest spheres of research that are far more varied and unruly.

It is for this reason that the following review of LS is organized around the three dominant modes of landscape research currently in operation. While not all landscape research comfortably fits within these schools, (i) they constitute three distinctive modalities for approaching what the landscape is and how it can be studied; and (ii) much contemporary work on landscape seems to work from, if not within, one of these traditions, although overlap and development are no doubt also present.

The Berkeley tradition

It was Carl Sauer who sought to convince American geographers that the landscape constitutes not simply a legitimate object of analysis but the discipline’s most significant and potent proxy for understanding the social forces that shape our world. Landscape interpretation, for Sauer, operated as an ideal type of geographic work, and his passionate pursuit of the topic effectively launched it as an independent field of geographic research for over half a century. Sauer’s main proposition for studying the landscape is the idea that the Earth’s surface is patterned and shaped by the material practices of past and present cultural groups. In this sense, the landscape is akin to a cultural artifact. Like shards of pottery...
or paintings on a cave wall, the landscape – in its organization and use – evidences traces of the cultural groups that settled it, its shape, formation, and adornment reflecting a community’s unique modalities of transforming the Earth. The purpose of geography, therefore, is to seek clues of inhabitation and reconfiguration. Sauer’s famous formulation that “culture is the agent, the natural area is the medium, the cultural landscape is the result” (1963, 343) presents the landscape as a canvas that cultural groups inscribe through the construction of material structures (e.g., barns, homes, and fences) and/or their distinctive patterns of tilling and organizing the land. The point, for Sauer, is that the landscape operated as a cultural hearth in which a community’s unique expression of land and life would be written and, crucially, read.

While there are a number of key texts that elaborate his thought and method in great detail, two broad points characterize Sauer’s conception of landscape and the distinctive approach to geography it establishes. First, Sauer resolutely understood landscape interpretation as a field-based practice. In order to study landscape, one must go outside and seek its material traces. The comparison to archaeology is apt, with the emphasis on exhuming signs of the past. But, rather than digging, geographers rely on their eye, relating clues in the landscape (an overgrown stone wall, a buried foundation, a system of field organization) to the imprint and layering of past cultural groups.

A second point concerns Sauer’s empiricism. Given that landscape is a record of culture, it is up to geographers’ visual acumen to interpret the signs they see. Landscape interpretation in the Sauerian tradition is an observational practice, steeped in inductive reasoning and an empiricist (but not positivist) epistemology. As Wagner (2009, 383) puts it, the geography that Mr. Sauer advocated was built on a foundation of direct immersion in field observations, interpreted in a broad, rather speculative fashion, informed by solid science, and directed above all toward environmental change over time in a close reciprocal relation with cultural development. Preformed deductive schemes could not compete with alert and insightful first-hand observation in concrete earthly context.

Where this tradition hit a conceptual wall, however, was in its theorization of culture. Because Sauer was primarily an empiricist, the trajectory, as well as legacy, of his work focuses on developing a coherent concept of what landscape is (a unique material expression of land and life imprinted on the material world) and how it could be studied (through visual observation). How and why culture imprinted the landscape in such a way was, for Sauer, a less interesting question. As Wylie (2007, 27) suggests, Sauer was “primarily interested in visible, material ‘evidence’ of culture in the landscape … and not with the ‘inner workings’ of culture in terms of shared cultural beliefs, rituals, ideologies, values, attitudes and so on.” This is not to say that Sauer ignored the question, but he dealt with it by borrowing from anthropology, and specifically Alfred Kroeber’s theory of superorganicism: the idea that culture’s rules, manners, and modes of living transcended individual experience and decision. Superorganicism was, in many ways, the “black box” that allowed Sauer’s concept of landscape to thrive, since it was culture itself (i.e., culture as an autonomous, causative force) that shaped the landscape. While there were many aspects of Sauer’s approach to landscape that were critiqued in the 1980s, the sharpest and most potent lines of attack pivoted on its weak conceptualization of culture. While many geographers who worked in the Sauerian tradition denied any abiding fidelity to superorganicism, this position was less about
rejecting a particular theoretical position and more about a commitment to an inductive project: that is, a belief that observation and description were the proper approach for a geographer and that any ambition to make broader claims must be born from evidence.

While this commitment to empirical description frustrated politically minded geographers emerging in the 1970s and 1980s, it is important to recognize that the empirical bias of Sauer and his students did not prevent their own theoretical revolutions. William Denevan’s work overturned long-held assumptions about the impact of native peoples on the precolonial landscape of the Americas. Drawing on his empirical work in Latin America, Denevan (1992) argued that the “pristine myth” – that is, the idea that native peoples’ impact on landscape change was minimal or benign – was not only racist but deeply Eurocentric, steeped in notions of the noble savage and ignorant of the industry and innovation of Indigenous peoples. Through a detailed discussion of native population figures, indigenous land forming, building technologies, and the extensive evidence of environment recomposition, Denevan precipitated something of a gestalt shift both in academia and in popular culture in terms of how presettlement America was imagined.

Today, work by Mathewson, Zimmer, and Sluyter continues to draw on fieldwork to analyze landscape change. While there is less emphasis on looking, regional expertise remains its defining attribute and maintains its distinctive relation to other forms of environmental history. Indeed, if there is one thing LS would clearly come to lose in its various revolutions, it is the emphasis on fieldwork and local knowledge. Few studies of LS in either the “new cultural school” or the “phenomenological school” from the Anglo-American world show any sustained interest in non-European case studies or in work in a language other than English. Today, work in the Berkeley School tradition overlaps with and intersects with cultural and political ecology. While such terms are no doubt broader and more encompassing than LS, the trajectory of LS interested in the cultural and political implications of landscape change certainly finds a comfortable home here.

In addition to mapping the way Sauer’s work has evolved, the innovative ways in which it has been retrieved are also worth noting. Ken Olwig, while not particularly interested in Sauer’s method (or in landscape morphology per se), follows Sauer into the Germanic pre-Renaissance tradition to recover what he terms the substantive nature of landscape: that is, the manner in which landscape operates as an expression of landed life – a life that is defined by, and expressed through, a relationship with the land. Explaining precisely what this relationship involves, and how it has changed, is a central feature of Olwig’s work, but principally it is founded on a conception of landscape as a unique legal administrative construct that binds a community to its geography and the forms of economy, community, and custom that geography bequeaths. In this sense, landscape is a comprehensive relationship between land and community – a legalistic administrative structure, to be sure, but one that positions land not simply as resource but as a means of defining communities. While much of Olwig’s work focuses on examining how this conception of land and life has been eroded since the Renaissance, there is a broad emphasis on recognizing the centrality of the land and life nexus that Sauer identifies as the engine of landscape development.

Ninety years after its initial publication, Sauer’s major work, “The Morphology of Landscape,” is cited by an increasingly smaller group of geographers. And yet there is little doubt that its legacy lives on. While landscape never took on the centrality and import in the discipline
that Sauer had hoped, this was perhaps more because of the venal critiques of Hartshorne and the scientism of the quantitative revolution than any shortcomings in the Sauerian method itself. Indeed, Wylie (2007) argues that there is an increasing appreciation of Sauer in the new biogeographical work of people like Sarah Whatmore, Steve Hinchliffe, Beth Greenhough, and Emma Roe, many of whom subscribe to Sauer’s insistence on the inseparability of nature and culture. Similarly, there is a more implicit recognition of Sauer’s influence in the phenomenological school of landscape studies, where landscape as an expression of land and life is being reconsidered. Most significantly, however, Sauer’s influence continues in the way North American cultural geography conceptualizes landscape as something material. While the British tradition has a tendency to approach landscape in visual and perspectival terms, American LS thinks about the landscape as a place—a hearth, or repository, for culture and cultural practices, however those may be conceived.

New cultural geography

If there was a moment when the critiques and new theoretical approaches that came to be called new cultural geography began, it was probably in 1980 when James Duncan published his critique of the superorganic theory of culture. Duncan’s central line of attack was that the theory was holistic, deterministic, and, crucially, long abandoned by anthropology and archaeology. The critique, and the response to it, are interesting events in the history (and sociology) of the discipline, but the event tells us less about the work of the Berkeley School and more about the issues and concerns of a new generation. As previously suggested, it is unclear to what extent traditional cultural geographers cared about culture. While their field ostensibly concerned the geography of cultural groups, their method focused almost exclusively on landscape change. Superorganicism, in this sense, was simply an easy way to sidestep the sticky question of culture. And, while some geographers (such as Wilbur Zelinsky) championed superorganicism as a ready-made solution, most of them ignored the question and hoped it would go away. Duncan, however, did care about culture. Steeped in the anthropology of Clifford Geertz and the symbolic interactionism of Irving Goffman, Duncan’s target was not the work itself but the atheoretical manner in which it proceeded, working with a deterministic theory of culture that was blind to question of power and meaning or indeed any issues that could be considered sociological.

It would be almost 10 years before these initial stirrings took shape in what would become known as the new cultural geography (NCG). In a flurry of publications that appeared in the late 1980s and early 1990s Denis Cosgrove, Stephen Daniels, Mona Domosh, James Duncan, Peter Jackson, Nuala Johnson, David Ley, and others would reorient and revive the study of culture in geography in both the United States and United Kingdom. For simplicity’s sake, their arguments can be broken down into ideas about what culture isn’t (the critique) and what culture is (the theory).

In terms of the critique, new cultural geographers responded to traditional cultural geography by arguing that culture is not deterministic, holistic, or innocent. It is not deterministic because it does not transcend individuals and individual choice; it is not holistic because it is made up of numerous subgroups and competing identity systems; and it is not innocent because the social frameworks it provides are imbued with relations of power. One of the keywords that came to
define NCG is *contested*. Culture is contested in the sense that it constitutes a realm of struggle where individuals strive to express the forms and signs that are meaningful to them onto the material world.

For new cultural geographers, landscape is not an artifact of identifiable cultural groups, but reflects social struggles to produce the material world in a manner that reflects one's ideas, values, and norms. Landscapes, in this framing, are *made to be read*. They are produced by powerful social groups to express a certain idea of how society is and/or should be organized. In addition, because landscapes are part of our everyday environment, they are a particularly insidious way of getting such ideas across, as Duncan and Duncan (1988, 123) point out:

> If landscapes are … read “inattentively” at a practical or nondiscursive level, then they may be inculcating their readers with a set of notions about how the society is organized: and their readers may be largely unaware of this. If, by being so tangible, so natural, so familiar, the landscape is unquestioned, then such concrete evidence about how society is organized can easily become seen as evidence of how it should or must be organized.

It is this combination of attending to the symbolic and meaningful dimensions of landscape, while being aware of the political implications of its meaningfulness, that is the hallmark of NCG. Stephen Daniels calls this the *duplicity* of landscape: that is, landscape’s capacity to be pleasurable and powerful at the same time. While the landscape provides a culturally specific manner of visual pleasure, it is also an expression of authority and ownership whose effects may be designed to exploit or control.

There are three further points to be made about NCG that explain both its impact and its enduring popularity. The first is that the rise of NCG coincided with a broader cultural turn in the discipline. While a systematic review of this work is beyond the scope of this entry, suffice it to say that questions of landscape became a minor player in a wider set of questions about the way power operates in and through space and place. Throughout the 1990s and early 2000s, a whole new set of geographies emerged, focusing on how space was produced by particular registers of power. Hence the emergence of an array of sub-subfields examining how taken-for-granted ideas about gender, race, sexuality, class, and youth insidiously insert themselves into our most everyday neighborhoods, parks, and playgrounds.

Second, it is often argued that the emergence of NCG was primarily a British phenomenon. This is true in the sense that many of its chief proponents were working in British geography departments, but also in the sense that it drew on the Birmingham School of Cultural Studies and British writers such as Raymond Williams, John Berger, and Stuart Hall. That said, the British character of NCG has been overstated. Certainly Meinig’s work on the symbolic aspects of “ordinary” landscapes is a significant contribution, as were Marxist geography (with its emphasis on the determinism of capitalist structures) and humanistic geography (with its emphasis on individual experience and sense of place), both of which were championed by geographers in North America. Indeed, Cosgrove (1978) argues that NCG was an attempt to bridge Marxist and humanistic geography, acknowledging the power-laden nature of landscape on the one hand and its role in generating meaningful subjective experience on the other.

This leads to the final point about NCG. If Cosgrove is correct that NCG represents a rapprochement between Marxist and humanist geography, it was by no means a comfortable one. For Don Mitchell, the representational emphasis of NCG fetishizes the look of landscape and
ignores how such imagery hides the exploitative labor relations that produce its aesthetic qualities. Similarly, Olwig illustrates how the development of landscape visualization (and the pastoral ideal) historically disrupted long-standing community relations with the land. For both Mitchell and Olwig, landscape aesthetics is a powerful social tool that is used to veil the implications of capital exploitation and/or land alienation. For others, however, the landscape is primarily a site of meaning and experience. Thus, in the work of David Matless and Karen Till, social and cultural identities that are vague or uncertain in the abstract become meaningful and concrete as they are expressed in and through the landscape. In Matless’s work on the Norfolk Broads, appropriate forms of leisure and conduct in the landscape are mechanisms by which Englishness is understood and enacted. Similarly, Till illustrates how German anxiety about its history is worked out in the landscape through the transformation of Berlin. In each case, the landscape is used to communicate abstract sensibilities rather than clear ideologies, and it is precisely the artful utilization of form and conduct that make ambiguous ideas meaningful and seemingly shared.

In terms of current work, a number of contemporary texts reflect the NCG tradition of providing regional and historical studies of landscape and landscape struggle. Dwyer and Alderman’s (2008) study of the memorialization of the Civil Rights Movement is an excellent example of how contests over material space bring to a head the subtle emotions that sustain segregation. In their study of the conflict to rename 9th Street after Martin Luther King in Chattanooga, Tennessee, Alderman and Dwyer make clear how latent and discordant feelings can be suddenly and eruptively made present by an act of geography. Similarly, Schein’s (1997) work on a suburban neighborhood in Lexington, Kentucky, draws attention to the diverse voices contributing to the development of a particular site over time. Schein’s argument is that landscape is discourse materialized, meaning that the landscape represents not only the voice of dominant authors but also the inconsistency and internal contentiousness present in all social discourse and ideology.

One of the distinctive features of NCG is its interest in theory. While NCG may be recognizable for its consistent attention to questions of representation and power, it draws eclectically from theory to make this point. Thus, Kirsch (2007) draws on sociology of science (e.g., Latour) to illustrate how landscapes operate not just as representations but also as scientific entities that carry within them their own forms of truth and necessity. Similarly, Della Dora (2013) draws together Cosgrove’s idea that “landscape is a way of seeing” and the concept of topia to examine how other cultures at other times have their own ways of seeing: that is, ways of making sense of the world through visual means other than perspective. Finally, there has been an array of sub-subfields that have emerged from the NCG tradition which highlight the various kinds of landscape struggle that can be potentially studied: for example, military landscapes, tourist landscapes, indigenous landscapes, imperial landscapes, and religious landscapes. In this sense, while NCG, as an approach to landscape, is over a quarter-century old, it continues to be practiced and developed in the discipline and has by no means been supplanted by emerging traditions.

Phenomenological school

The phenomenological school is often associated with the work of Paul Cloke, Owain Jones, Hayden Lorimer, Mitch Rose, David Seamon, John Wylie, and others. Its rise in the discipline is similar to that of NCG in three ways. First, its ideas were intimately connected
to developments in anthropology and archaeology, particularly the work of Tim Ingold. Second, the phenomenological school emerged from a broader critique of the way cultural questions are approached and pursued. And, third, the phenomenological school is primarily (and in many ways continues to be) a British phenomenon. Indeed, the ideas and concepts associated with the phenomenological school are closely aligned with nonrepresentational theory (NRT), a trajectory of thought developed by Nigel Thrift, Sarah Whatmore, Paul Cloke, and their graduate students at the University of Bristol in the late 1990s. While the connection between the phenomenological school of landscape and NRT are significant, that significance should not be understood as a seamless (or even comfortable) theoretical overlap. Indeed, as with many fields that were inspired by NRT, the points of connection flow from a broad approach to (and inspiration from) poststructural theory, rather than any consistency in theory or method. In this sense, it is useful to summarize some of the key ideas in NRT first, in order to properly understand its relation to the phenomenological school of landscape. While it is often suggested that NRT denotes an outlook and approach rather than a specific theory, NRT can nonetheless be characterized by three priorities: of ontology, events, and practice.

The prioritization of ontology can best be understood as a deprioritization of epistemology, the latter being the guiding conceptual frame for cultural geography since the mid-1990s. Prioritizing epistemology means understanding meaning, symbolism, and other forms of representation as coloring an otherwise objective world. Inherent in this epistemological framing is not simply a Cartesian dualism but, more significantly, a vision of the universe framed by the problem of knowing: that is, it is a framework born out of the question “How do we know the world?” Following Heidegger, NRT starts with a question not of knowing but of existence. Thus, it draws on ontologies oriented toward understanding how things exist rather than what they are. This leads to the second priority, which is NRT’s focus on events. The emphasis of the ontologies commonly used by NRT is that subjects are not self-standing beings but episodes, outcomes, or events that emerge from an ontological situation. While this situation is conceptualized differently depending on the theorist, there is a broad inclination in NRT to look beyond the nouns with which people engage on a daily basis (other people, pavements, pets, weather, etc.), in order to grasp the context from which these nouns emerge. In this framing, the world is not simply the Earth and its furniture but assemblages whose appearance is dependent on the relations that give them life. The final priority of NRT is practice. If NRT approaches the world not in terms of things but in terms of the background relations that allow things to appear, then concepts such as practice and performance take priority over questions of agency and will. In an NRT framing, space, place, and landscape are not produced because they cannot be wholly attributable to self-conscious agencies or a will. Rather, such events express the situation from which they emerge: that is, a context where certain things are (and are not) possible. To approach geography in such terms is to recognize how desires, affects, memories, energies, and other potentialities resident within a particular moment are concocted into forms and solidarities whose effects may be ephemeral or sustained, singular or repeating, limited or far-reaching, animate or barely audible. The point is that geography is not a noun but a practice; it is the cumulative effect of everyday enactments and events.
While this gloss of NRT leaves a number of significant topics undiscussed (e.g., affect), it is nonetheless indicative of ideas that have had an enduring impact on LS. Indeed, three characteristics of the phenomenological school of landscape broadly align with these themes: there is an interest in (i) the situation of landscape (ontology); (ii) dwelling (events); and (iii) practice.

Perhaps the most distinctive feature of the phenomenological school is its focus on the situation of landscape. Both traditional cultural geography and NCG focus on the landscape as an object, and (more specifically) an object that is produced. The phenomenological school, however, does not see the landscape first and foremost as something produced, but rather as a situation. When John Wylie ascends Glastonbury Tor, he attends to how the practice of climbing provides him and his companions with a means to see. Thus, the tor is not approached as an object to be viewed but as the means by which viewing takes place. It is itself a constitutive element in the emergence of a viewing subject: “the Tor is never an object for a gaze, nor a viewpoint on its own. It is a modulation of the visible world which lets there be a gaze to behold things open and hidden. Its elevation reveals and conducts the surrounding landscape” (Wylie 2007, 454; emphasis added). Thus there is an inclination to draw attention away from subjects and objects in order to focus attention on the conditions that allow a subject–object relationship to arise. Subjects come to be subjects through the practice of viewing: that is, through the process of being in an ontological situation that affords seer and seen to emerge through a set of practices and perceptions (e.g., walking and seeing). The emphasis, once again, is not on what the landscape is, but how it exists.

The second theme that characterizes the phenomenological school is an interest in dwelling. The dwelling perspective, as it is presented by Tim Ingold, presents landscape as a form of involvement rather than of representation. As he states, “the forms people build, whether in the imagination or on the ground, arise within the current of their involved activity” (Ingold 2000, 186). In his discussion of Koyukon hunters, he illuminates their capacity to use those aspects of their world that they have learned to see. Thus, it is only by being in the world in a particular manner that the world offers the subject particular possibilities. Similarly Cloke and Jones (2004) examine how trees in a cemetery in Bristol shape social memory in a manner that impinges on the expectations, planning, and management of the site. In this reading, the landscape does not exist as a representation of pre-established ideologies or ideas, but as expressions borne by inhabitants engaging with the world they are in. As Ingold (2000, 179) suggests, the world is not made before it is lived in: “the acts of dwelling are preceded by acts of worldmaking.”

The final and related characteristic of the phenomenological school is its emphasis on practice. If the landscape is an expression of involved activity, then it can be thought of as emerging from practices. This is why Ingold uses the term “taskscape”: that is, landscape as an expression of our everyday activities and tasks. In geography, Hayden Lorimer draws on journals and letters to explore how extraordinary landscapes can emerge from the most intimate tasks. In one example, Lorimer (2006) illustrates how the Scottish landscape emerges as a reindeer landscape through the activities of two people: Mikel Utsi (a nomadic Lapland reindeer herder) and Ethel Johns (a British anthropologist). What might be a story about the landscape being produced in a particular manner by two subjects becomes, instead, an exhumation of the relations and intimacies that bring this project to fruition: for example, Utsi’s familiarity with
reindeer, Johns’s attachment to Lapland, the ecological relations that make reindeer life possible, their love for each other, and their mutual sense of peace within the herd. The reindeer project is not the result of a particular vision but an expression of pre-established intimacies which are themselves informed by histories, attachments, desires, and relations.

While the phenomenological approach to landscape is fairly recent, it has been widely influential and has inspired an array of innovative work. This work can be broadly characterized by an attention to the manner in which insubstantial or affective modalities of sensation are critical components in the constitution of our engagement with, and experience of, landscape. This process can be seen in the work on absence and ghostly landscapes where the voices of the past, the remembered and the dead, generate (or undermine) gardens, housing developments, and heritage sites (see work by Franklin Ginn and Karen Till). It can also be seen in the increased interest in landscape experience and the various factors affecting our comportment to and involvement with a site (e.g., see Nina Morris’s and Tim Edensor’s work on luminosity, Chris Tilley’s experiential archaeologies, and Peter Merriman’s work on driving and mobility).

More broadly, the phenomenological school has had an impact on how landscape sensibilities are integrated into our expectations and even into policies relating to what people, communities, and governments desire from our environments and how those expectations are grounded in often precognitive expectations (e.g., see Catherine Brace’s work on climate change). In this sense, while the phenomenological school is relatively young, it is already influencing the trajectory of LS in geography, anthropology, and beyond, and it remains to be seen how this work will evolve and intersect with other long-standing traditions in the field.

Conclusions

This entry began with a justification for dividing the discussion into three sections, each representing a particular school of LS. Such a division leaves out and/or marginalizes a range of interesting landscape research. Indeed, inherent in such a division are the familiar dangers of categorization: overemphasizing difference, eliding similarity, ignoring the stuff in between. It is not that these schools supplant one another or even represent coherent spheres of research. They are traditions: modes of landscape practice that carry with them a particular perspective on, approach to, and interest in the landscape. In this sense, it is important to keep two things in mind: (i) scholars have always drawn on these traditions eclectically; and (ii) these traditions are very much alive and evolving. In the United States in the late 1990s, there were two schools of landscape research, traditional cultural geography and new cultural geography, each battling to define and defend their distinctive method and approach. Existential anxiety inhered in these debates: new cultural geographers fighting for acceptance and traditional cultural geographers fighting for their existence. Current debates do not show such anxiety. There is little sense of teleology in phenomenological critiques of new cultural geography, and there is no suggestion that issues of power and representation are no longer valid. This lack of tribalism and existential angst about the field has given LS greater range, and a number of distinctive approaches to landscape can be identified in current work as well as far greater eclecticism. In this sense, while the history of LS may have been written many times in the past, it has also been a fairly easy one to write. Future histories, if they are to be written at all, will be far more difficult.
LANDSCAPE

SEE ALSO: Berkeley School; Cultural geography; Culture; New cultural geography; Nonrepresentational theory

References


Further reading

Over the last several thousand years humans have increasingly modified existing land use and land cover to meet various socioeconomic needs. It is noted that by 1990, 45.7–51.3 million km² of the global land surface (35–39%) was being cultivated, with forest cover decreased by approximately 11 million km² since 1750 (Pielke et al. 2011). By 2000, only a few desert regions, the central Amazon and Congo Basins, and the Arctic and Antarctic had not been affected by land-use/cover changes (LULCC). Over the last several decades LULCC has been increasingly recognized as a major forcing of local, regional, and global climate on multidecadal timescales (e.g., Mahmood et al. 2014; Pielke et al. 2011). Observations and modeling studies demonstrate that LULCC impacts meso-, regional-, and potentially global-scale atmospheric circulations, temperature, precipitation, and fluxes. These effects can be observed over multidecadal to longer timescales. Moreover, global climate modeling research demonstrates that the impacts of LULCC on extreme temperatures, for example at continental or smaller scales, can be equal to or greater than the radiative effect of a doubling of atmospheric CO₂.

In this context, the primary aim of this entry is to highlight the role of LULCC in the climate system. The impacts of LULCC can be both biogeophysical and biogeochemical in nature. Both biogeophysical and biogeochemical type responses of the climate, therefore, are discussed in the following sections. Biogeophysical changes can result in the modification of leaf area index (LAI), surface roughness, and albedo, while biogeochemical changes are changes that occur in the carbon and nitrogen budget, and the budget of other trace atmospheric constituents. The following discussion addresses the impacts of long-term systematic LULCC (e.g., agricultural land-use change, deforestation) and short-term abrupt changes (e.g., rapid urbanization).

**Climate System and LULCC**

LULCC has both biogeophysical and biogeochemical impacts on climate. Biogeophysical changes modify the radiation balance, energy partitioning, and exchanges of energy, mass, and momentum between the land surface and the atmosphere as well as the terrestrial water budget (Figure 1). Biogeochemical changes impact, for example, the nitrogen and carbon fluxes.

The surface energy and moisture budgets for bare and vegetated soils can be of help in understanding the impacts of LULCC and can be written as:

\[
R_N = Q_G + Q_H + Q_E \quad (1)
\]

\[
P = E + T + RO + I \quad (2)
\]

where \(R_N\) represents the net radiative fluxes, \(P\) is precipitation, \(E\) is evaporation, \(T\) is transpiration, \(Q_G\) is the ground heat flux, \(Q_H\) is sensible heat flux, \(Q_E\) is latent heat flux, \(RO\) is runoff,
and $I$ is infiltration. Physical evaporation plus transpiration is also known as evapotranspiration ($ET$) and can be expressed in energy terms (latent energy flux). Transpiration is a plant physiological process which is linked to photosynthesis and a variety of other plant functions. These equations (1 and 2) also show that the energy and water budgets are linked and that LULCC affects these budgets. Surface budgets could also be completed for carbon, nitrogen, and other chemical species. Each of these budgets will be changed if any characteristic of the land surface is altered. This includes changes caused by land management (e.g., deforestation) and also ones that arise from phenology, drought, and so forth.

Similar budget equations can be written for carbon and nitrogen fluxes. The carbon budget, for example, involves the assimilation of carbon dioxide into carbohydrates within vegetation, respiration of CO$_2$ from plants and animals, decay of animals and plants, industrial and vehicular combustion processes, outgassing from oceans and other water bodies, and volcanic emissions. The nitrogen budget has also been segmented into its different components. Among these components, for example, nitrogen deposition clearly has major LULCC interactions as the amount of material deposited on land is very large. However, more studies are needed because vegetation will be affected differentially by this inadvertent fertilization of plants, with resultant impacts on albedo, as well as transpiring leaf area.

Changes in the heat and water budget necessarily alter the carbon and nitrogen budget (and that of other trace constituents), and alterations in carbon and nitrogen budgets (and in the other trace gases and aerosols) change the surface heat and water budgets. A major reason for all of these fluxes are closely coupled is that the transpiration of water vapor through the stoma of plants is not only associated with changes in the heat budget,
as shown earlier in this section, but also intimately involved with the assimilation of carbon into plant leaf, roots, and stems through their stoma.

If the amount of plant material changes, this alters the transpiration of water vapor into the air and thus the amount of carbon that is assimilated. The amount of nitrogen compounds and other trace nutrients affects plant growth and vitality. The intimate coupling of all of the surface budgets is a fundamental aspect of the climate system. In short, land management practices and LULCC that alter any one of these budgets necessarily alter all of them.

The following sections highlight the climatic impacts associated with some of the most notable types of LULCC. They include impacts of agriculture, deforestation and afforestation, and urbanization. Due to their relevancy and importance, also included are the impacts of CO₂ and aerosols.

**Agriculture**

The concept of human-induced vegetation cover changes having an effect on climate is not new. Pioneers of the Great Plains of North America believed that “rain follows the plow.” While there is evidence that human agricultural practices may have begun transforming the Earth’s atmosphere thousands of years ago, direct large-scale anthropogenic impacts on the Earth’s land cover are difficult to document before historical times. The possible first impacts of human activities were on vegetation through hunting and fires that may have led to extinction of many native species. Subsequently, agriculture and the domestication of grazing animals began to alter land cover. The following discussion includes examples illustrating impacts of LULCC due to agricultural activities.

It is found that the shift to agriculture often results in a cooling in daytime temperatures and presents the largest magnitudes of cooling. For instance, up to a 1.41 °C cooling of maximum temperatures was found at irrigated locations in the Ogallala aquifer region (Figure 2). In a subsequent assessment, a 1.56 °C increase was reported in average growing season dew point temperature over irrigated areas, and also up to a 2.17 °C increase was documented in dew point temperature for peak growing season months. These changes were primarily linked to modification of energy partitioning led by enhanced latent energy flux due to the availability of additional water through irrigation.

One study noted significant reductions in mid-June to mid-July maximum air temperatures, diurnal temperature range, and solar radiation of 1.7 °C decade⁻¹, 1.1 °C decade⁻¹, and 1.2 megajoules per meter squared per decade (MJ m⁻² decade⁻¹), respectively, over Canadian Prairies associated with agricultural activities. This research also found a precipitation increase of 10.3 mm decade⁻¹. Again, the increased latent heat fluxes from increased agricultural vegetation could lead to increases in dew point temperature without necessarily resulting in increased precipitation.

In northwestern India, up to a 0.34 °C cooling of growing season maximum temperatures was determined to be due to the introduction of irrigation. For individual growing season months, up to a 0.53 °C decrease of maximum temperature was reported and long-term temperature trends were largely negative and also statistically significant. These changes in temperature were associated with increased latent energy flux (Figure 3). Subsequent studies also suggest modification of dry season precipitation associated with irrigation and increased soil moisture and latent energy flux. Other modeling studies suggest changes in circulation over the
Asian continent due to the widespread adoption of irrigation.

In a modeling study for modern-day land cover, there was a statistically significant warming for the austral summer (December–February) for mean surface temperature of 0.1–0.6°C in eastern Australia and a warming of 0.1–0.4°C for the coastal southwest corner of Western Australia, which is still largely covered by tall native forests. However, there was a significant cooling of 0.2–0.4°C over the inland wheat belt for summer and austral winter (June–August). Further analysis of daily rainfall events indicates an increase in the number of dry and hot days, an increase in drought duration, and a decrease in daily rainfall intensity and wet day rainfall in southeast Australia.

From an observational study, (the 2005–2007 Bunny Fence Experiment (BuFex)) in southwest Australia, a favored area of cloud formation was found that was linked to land-use change (Nair et al. 2011). The cooling effect was most prominent during the summer when the land use in the agricultural area deviates most from the native land in terms of albedo, leaf area, vegetation cover, and surface roughness. Analysis of radiosonde observations showed, on average, the noontime planetary boundary layer (PBL) height was found to be higher by approximately 260 m over the native vegetation during summertime,
Figure 3 Latent energy flux differences (W m$^{-2}$) for March 23, 2000: (a–c) CTRL-DRY05, 10, and 15 and (d–f) CTRL-WET05, 10, and 15. Soil moisture was systematically decreased (DRY) and increased (WET) up to 15% with 5% interval (Sen Roy et al. 2007).
while over winter time the difference was approximately 189 m. Observations of energy fluxes from aircraft show that the enhanced boundary layer development was driven by sensible heat fluxes that were consistently higher over the native vegetation areas, with peak differences of approximately 200 W m$^{-2}$ and approximately 100 W m$^{-2}$ observed during the summer and winter seasons, respectively.

To address the effects of increasing cropped areas and afforestation on atmospheric processes in South America, Beltrán-Przekurat et al. (2012) performed sensitivity experiments using a fully coupled atmosphere-biospheric regional climate model. The general effect of agriculture was cooling when the shift was from grasses (with C3 metabolism) to crops, and warming when the shift was from grasses (with C4 metabolism), wooded grasslands, or trees. Near-surface temperatures decreased when crops (such as wheat and soybean) replaced C3 grasslands. Increases in temperature were found when crops replaced evergreen trees and wooded grasslands (e.g., in Brazil) and C4 grasslands (e.g., in the central region of the Pampas). Near-surface temperatures decreased when crops (such as wheat and soybean) replaced C3 grasslands. Increases in temperature were found when crops replaced evergreen trees and wooded grasslands (e.g., in Brazil) and C4 grasslands (e.g., in the central region of the Pampas). The coolest (up to $-0.8^\circ$C) and warmest (higher than 0.6 $^\circ$C) 2 m temperature differences (the difference at 2 m above the ground) appeared in the spring (October–November) averages. Maximum temperatures showed a larger decrease than minimum temperatures. Minimum temperatures slightly increased in summer in the southern Pampas. This also means a decrease in the diurnal amplitude of temperature. The results are consistent with the observed trends in temperature over Argentina.

Based on the observed responses of the climate system to agriculture-related LULCC, these regions where cooling occurs can be considered as “Agricultural Cool Islands” (ACI). This terminology could be applied to the landscapes dominated by both rain-fed and irrigated agriculture.

Deforestation, afforestation, and desertification

In many parts of the world, vast regions of original forest have been cut to permit agriculture, urbanization, and other human activities. Large areas of the Amazon are currently being clear-cut for agriculture or burnt by fire. This land-use change is referred to as deforestation. To reverse this deforestation, in some regions trees have been planted either to recreate an originally forested area or even to introduce forests where none exists. An example of the latter is Nebraska National Forest in the United States. Moreover, in response to various policy responses related to conservation efforts and to increases in carbon sink, some regions in the world have made efforts to plant vegetation and create new forests. This type of land-use change is called afforestation.

Observational studies have shown that deforestation results in changes in regional temperature, cloud cover, convection, and precipitation and is linked to changes in albedo, surface roughness, and land surface hydrology. Deforestation has been shown in a modeling study to cause changes in regional and mesoscale wind circulations and cloud formation in the Amazon. The dynamics involved in LULCC-driven climate change in the subtropics are slightly different from those in the tropics. Land surface and atmospheric radiation balance plays an important role in the establishment and maintenance of semi-arid regions. Positive feedbacks from a reduction in vegetation cover and its concomitant albedo increase can further exacerbate desertification. It is suggested by the scientific community that the interplay between vegetation, moisture, and
weather in the subtropics may create preferred modes of variability and preferred states of stability of climate in these areas.

Desertification occurs when a region that was covered in grassland or even forests is replaced by sparse desert plants or even bare soil. Studies have shown that changes in albedo are key to the climate response. Moisture is an important limiting factor in the subtropics, and modeling studies show that reductions in surface moisture tend to disrupt further precipitation by inhibiting moisture recycling. Modeling research shows that vegetation degradation in the Sahel produced rainfall anomalies consistent with the observed multidecadal drought in that area.

Model simulations were completed to assess modern era vegetation and that of earlier times. Restoring vegetation for the times of the Roman Empire, for example, caused summer rainfall in the model to increase in southern Europe and the Atlas Mountains, the lower Nile Valley, and the Levant (e.g., Reale and Dirmeyer 2000). Model responses to LULCC in the mid-latitudes tend to be more subtle and complex, because agriculture displaces both forest and prairie/steppes, with opposite effects on albedo and roughness, making the response highly sensitive to the details of the experiment’s implementation.

As expected, afforestation also inadvertently modifies biogeophysical and biogeochemical characteristics of a region and thus regional climate. Afforestation in temperate and tropical South America has led to cooler and wetter near-surface atmospheric conditions over most of the afforested areas, associated with higher and lower latent and sensible heat, respectively, than in current land cover conditions (Beltrán-Przekurat et al. 2012). Modeling research found temperature differences between current and afforested conditions were 0.6 °C and 0.8 °C in spring and summer, respectively. Higher evaporation and transpiration rates and lower temperatures in forest compared to grasslands were also found in an observational study of the area.

Precipitation is also affected by the vegetation changes in temperate and tropical South America. Domain-averaged precipitation showed an increase in natural versus afforested simulations: absolute mean values were higher, up to 30 mm month⁻¹ in summer, compared to current conditions. Relative changes were up to 30% in the afforested condition. Spatially, areas of increased rainfall were associated with changes in latent heat, roughness length, and areas of moisture convergence. Some dependency of afforestation pattern and moisture flux convergence was also observed, indicating that interannual variability in the fluxes may also affect the simulated patterns and magnitudes (Beltrán-Przekurat et al. 2012).

Urbanization

Humans tend to concentrate in cities. Although covering less than 0.5% of the Earth’s land surface, cities still exert a significant effect, out of proportion to their fractional coverage, on regional and global climate both from their LULCC, through their effects on the surface energy budgets, the water cycle, and surface biogeochemical cycles (e.g., carbon, nitrogen), and from their aerosol and gas emissions. This type of LULCC is referred to as urbanization.

The urban heat island (UHI) (Figure 4) is probably the most well-studied process with respect to weather and climate processes. A net surplus of surface energy over urban regions can be explained by enhanced surface sensible heat flux, ground heat storage, and anthropogenic heating as well as changes in physical evaporative and transpiration cooling. Because urban
regions typically have lower albedo values than rural areas (except, for instance, for urban areas embedded in coniferous forests), they absorb more shortwave radiation energy at the surface. Smaller values of the sky view factor (SVF) below roof level result in decreased radiative loss and turbulent heat transfer and add to the UHI anomaly.

UHI’s impacts can vary depending on existing climate regime and the vegetation type both within and surrounding the urban area, latitude, seasons, and even time of day. For example, over temperate regions, UHI can produce urban–rural sensible heat flux differences leading to a city center as a significant heat source. This is linked to the removal of forests, vegetation, and wetlands and the replacement by the built environment. On the other hand, over arid and semiarid regions urbanization may result in the introduction of irrigation and the resultant increase of latent heat fluxes. Moreover, it is found that over high latitudes, the urban–rural temperature differences are the highest during winter.

Urban land use can also modify, amplify, reduce, or initiate precipitation cloud systems since the surface heat, moisture, and momentum fluxes are altered from the surrounding non-urban landscape. Rainfall is often, therefore, different over the urban regions and downwind compared to nonurban areas, as verified by both in situ and satellite datasets. Thunderstorm rainfall can be increased over urban areas, for example, when the added heat from urban areas invigorates these storms. In semiarid irrigated urban areas, the cities also add water vapor to fuel the thunderstorms. UHI is a well-studied process; however, uncertainties still exist and a better understanding of the built-up environment and its effects on various aspects of the Earth’s weather and climate system is still emerging (e.g., Shepherd et al. 2013).

Carbon dioxide (CO$_2$) and LULCC

The association of LULCC and the carbon cycle is well known. Eastman, Coughenour, and Pielke (2001) investigated regional weather conditions in the central grasslands of the United States for three experimental scenarios using a model: (i) land cover was changed from current to potential vegetation; (ii) radiative forcing was changed from 1$\times$CO$_2$ to 2$\times$CO$_2$; and (iii) biological CO$_2$ partial pressures were doubled. Results indicate that the biological effect of enriched CO$_2$ and of land-use change exhibit dominant effects on regional meteorological and biological variables, which were observed for daily to seasonal timescales and grid to regional spatial scales. Simulated radiation impacts of 2$\times$CO$_2$ were minimal on these space- and timescales, with interactive effects between the three experimental scenarios as large as the radiational impact alone. Model results highlight the importance of including 2$\times$CO$_2$ biological effects when simulating possible future changes in regional weather.
Other modeling studies have shown that the inclusion of biospheric feedbacks due to increased CO$_2$ impacts the LULCC simulations. These studies have noted that these impacts are non-negligible and could be compared to changes associated with radiative forcing. Separate modeling research suggests that the impacts of LULCC initially remain over the areas of change and subsequently could be found over remote areas as well. In addition, it has also been noted that increased CO$_2$ does not negate the impact of LULCC but rather produces statistically significant regional effects. Regional modeling studies show that LULCC and increased CO$_2$ impact temperature and precipitation in China. It was found that increases in CO$_2$ resulted in increases in both maximum and minimum temperatures. Precipitation intensity has also changed due to LULCC and accompanying increases in CO$_2$. Moreover, modeling research demonstrated that LULCC and increased CO$_2$ modify return values of extreme maximum temperature and convective precipitation over Europe and China.

**Aerosols and LULCC**

It is well known that the impacts of aerosols on atmospheric processes, especially PBL layer evolution, are strongly coupled to surface processes and thus underlying land use. Depending on the albedo of a surface, aerosols can result in either warming or cooling of the PBL. Typically, aerosols over a high albedo surface result in warming, while those over a low albedo surface lead to cooling of the boundary layer.

Different atmospheric responses to aerosol radiative forcing over different land-cover types imply that aerosols modify horizontal temperature gradients and mesoscale circulations induced by land surface heterogeneity. Numerical simulations found that radiative forcing resulting from Saharan dust is rapidly reflected in the surface energy budget for land, causing a daytime reduction in downwelling radiation at the surface, and a corresponding reduction in sensible heat flux and boundary layer air temperature. However, over the ocean, since the radiation is absorbed over a deep layer compared to land, radiative forcing has little impact on the surface energy budget at short timescales. As a result, over the ocean, direct heating of air by the dust aerosols dominates and leads to an increase in air temperature. The net effect of aerosols, therefore, is to reduce the land–ocean thermal gradient in the PBL.

The coupled aerosol and land-cover change impacts on the modulation of mesoscale circulations may thus be especially significant in maritime environments. Sea breeze circulations, for example, are an important driver of convective activity tied to small islands. As a result, concurrent land-cover change and biomass burning in maritime settings and the combined impact of these two forcing mechanisms will influence deep convection, but this is not well understood.

A recent work by Junkermann et al. (2009) shows that land-cover change can affect the atmospheric distribution of aerosols and cloud condensation nuclei (CCN) which thus impact cloud microphysics. Aircraft observations over southwest Australia show that the differing biogeochemistry of salt lakes in the cleared areas and emission of aerosol precursors lead to approximately doubled CCN concentrations over this region compared to adjacent native vegetated regions. Corresponding differences in cloud particle size distribution are also observed in the clouds over the native vegetation region. Higher concentration of cloud droplets with diameters greater than 10μm were found over native
vegetation, while the cleared agricultural regions showed higher cloud droplet concentrations for particle diameters less than 10 μm.

Final remarks

This entry clearly demonstrates many impacts of LULCC on the climate system at all spatiotemporal scales through a variety of pathways and mechanisms. Moreover, after decades of scientific work around the globe, LULCC is now considered as a first-order climate forcing. However, additional research needs to be conducted to permit a further understanding of the pathways through which LULCC affects the climate system. This requires the establishment of in situ observational platforms in areas of rapid LULCC and also in and around where LULCC has already occurred, for example, in and around irrigated areas. It is also important that sustained funding be available for already established high quality in situ networks. Examples of such networks are the US Climate Reference Network (USCRN), and the Oklahoma, Kentucky, Nebraska, and Delaware mesonets. Similar funding stability should be provided to maintain the continuity of various satellite missions (e.g., Landsat). These observational capabilities should be complemented by simultaneous efforts for the improvement in modeling of the weather and climatic impacts of LULCC. We conclude that observational capabilities are essential for improvement and validation of modeling of LULCC effects.

Acknowledgements: The authors of this entry gratefully acknowledge the authors of Pielke et al. (2011) and Mahmood et al. (2010, 2014) which initiated many interesting and thoughtful discussions on climatic impacts of LULCC. For Roger Pielke Sr., the preparation of this entry was supported by NSF Grant AGS-1219833.

SEE ALSO: Agroclimatology; Climatology; Desertification; Land change science; Urban climatology

References


of Agriculture in the Climate System and in Climate Change.” *Agricultural and Forest Meteorology*, 142(2–4): 234–254.


Language and geography are intimately and intricately interconnected. The relationship between words and the world is pervasive, consequential, and often politically significant and divisive. This is because language, a distinctive marker of human cultural capacity and identity, can be used to name, categorize, and contest our world, and because language always bears the stamp of someone’s cultural and political authority. We use words to name things in the world (and beyond it), and to describe or explain feelings about our world or about other human or nonhuman agents. Variations within or between languages in how humans describe or explain their world, or understand local places, may be revealing of profound differences in meaning about place and space.

The beginnings of formal interest in the relationship between geography and language date to seventeenth-century Biblical scholarship and, from the eighteenth century in Europe and America, to discussions over the origins of language and the connections between language, national identity, and to the issue of social difference as expressed in terms of rank, status, or, to use the terminology of the nineteenth century, social class. In the nineteenth century especially, and in continental European scholarship in particular, the study of geography and language centered upon the mapping of language areas, by nation or by ethnic group.

In contemporary scholarship, the study of language and geography falls broadly within cultural geography. Three main themes may be identified. The first has to do with toponomy and taxonomy, with the words and categories which we use to describe and order the world, including the terminology used in geography (but it does not include the study of proper names, which is the preserve of onomastics). This first theme is most evident in the associations between place names, geographical features, cartography, and identity. The second theme incorporates the connections between language and geographical area. At different times and in different intellectual contexts, this is variously termed “language geography,” “linguistic geography,” “geolinguistics,” or “dialect geography.” For each of these descriptions, the focus is upon language’s areal differentiation over time and space. In the last two named, attention is paid to the geographical differentiation of language’s internal structures, the spoken differences in language, often in relation to social class, or to the rural–urban distinctions in a particular language’s dialects or lexicon. The third theme embraces the study of language within geography as an academic discourse, and of language’s social usage – “words-in-the-world” (McGeachan and Philo 2014) – as a means to intellectual authority and cultural and political power.

It is not appropriate to see these themes as either typologically or chronologically discrete, as methodologically distinct, or, even, as the concern of geography alone. Work on the politics of place names, for example, may have close affiliations with studies directed at the political authority behind the use of one language and not another, and explanation of the changing
Language geographies

National or regional geography of language or dialect areas similarly involves questions of power and cultural difference. Study of these issues is shared with linguistics, sociology, comparative philology, and ethnology, and, just as these themes are related, so, too, they have further subthemes.

Language, naming, and identity

The study of toponomy, the names of places and of geographical features, may reveal the linguistic history of a place or region since such names may contain elements of a previous language or variant meanings from that which is currently understood or dominant. There is, in Britain and in continental Europe in particular, an established intellectual tradition of place name studies as a distinct subset of enquiry in historical geography, history, ethnography, and in the study of folk practices, given that some such practices may survive, for example, only in a relict vocabulary. Capturing the nominative role of language and of given words either in their current form or in their former ethnological context may disguise, however, the processes and the authoritative acts behind naming. Understanding the names of places and the words used to label the world – on the ground, in writing, in speech, and on maps – requires that attention be paid to the relationship between the fact of description, the act of inscription, and the consequences of any linguistic proscription since given words, speech forms, or, even, entire languages, were often forbidden or deprived of their cultural currency.

Within modern historical geography and cultural geography, numerous studies have examined the ways in which identity, meaning, and notions of political and cultural order have been inscribed upon human landscapes through authoritative acts of naming, or, by contrast, have shown how meaning in place and in landscapes may be lost through linguistic change or, ultimately, language death. In eighteenth- and nineteenth-century North America, the contact between European trader-settlers with indigenes often resulted in the “writing out” of native names by their “writing over” in the mapped landscape and by parallel processes of “speaking over” as native languages were replaced, by English, or French, or Spanish, in the different spoken domains of daily life. In such encounters, the written word commonly dominated over the spoken word, and print culture over oral culture. The archival traces of such unequal linguistic encounters exist, where they exist at all, in the language of the colonizer, hardly at all in those of the colonized. In Australia, New Zealand, Hawaii, and throughout the South Pacific islands, the sustaining myth of *Terra Nullius* held by Europe’s settlers, geographers, and, importantly, map-makers – that these lands were unnamed and “empty” – led to them being inscribed with the names and languages of the colonial conquerors. The politics behind the descriptive, inscriptive, and proscriptive historical geographies of language and naming in colonialism and imperialism have modern parallels. In Israel’s administered territories, for example, Zionist ideologies are manifest in the names chosen for settlements in the West Bank and other disputed territories. In the Caribbean, tensions between users of the local creole patois or “speak” and speakers of the official language reveal linguistic differences to be matters of longstanding cultural and social stratification (for these and other examples, see Withers 2000).

Several points may be noted of this work and of its implications. Studies in geography on the authority and politics of naming in an historical context, and upon language acts in the contemporary world, share a concern with how the legacies of colonial pasts inform the
Language geography/linguistic geographies: areal differentiation and variations in language and speech

The mapping of languages by area, commonly in association with national identity or ethnicity, is an established feature of language geographies. From the early nineteenth century, such work was a distinctive element of German geographical scholarship. The earliest known thematic maps of language by area are the language maps of four continents, the *mappae geographicopolylgottae*, produced by the Silesian philosopher and theologian Gottfried Hensel as illustration to his *Synopsis Universae Philologiae* (1741). Hensel sought to explain all the world’s languages as having a common origin in Hebrew. In Germany, Julius Klaproth produced language maps of Asia in his *Sprachatlas* of 1823 in order to distinguish the different language families in that continent. Although dialect mapping exists for earlier periods, of variations in spoken German in an 1821 work by Johann Smeller, for example, most language mapping in the nineteenth century was concerned with language groups rather than with internal variations in language form. It was associated with the

colonialisms of the present. Both are motivated by the importance of that continued questioning of received historical and geographical “truths” which is a feature of postcolonial enquiry, in geography and in other disciplines. Although it is not always explicitly acknowledged, geographical work on language, naming, and identity has its parallels in studies concerned with writing in geography, with the act of writing as a matter of style (Billinge 1983), and, more than that, with different forms of writing (and of speaking) within geography as embodying unequal relations of power within the discipline. This issue is discussed in greater depth in the third section below.

Exploring questions of language, naming, and identity as questions of power requires that we pay attention to the nature of political authority and to those institutions and individuals in whom an inscriptive power to name is vested. This is particularly true of language’s central place in mapping and is a common feature of interpretation in the more recent history of cartography. Following the work of Brian Harley and his use of the work of Jacques Derrida and Michel Foucault, maps are seen as intrinsically powerful and power-full documents, capable both of inscribing space and of silencing claims to it by virtue of the omission of, or selection from, one language or word or worldview over those of other user groups. The map is an enduring form of geographical language. Maps may thus, as Mark Monmonier phrases it, “name, claim, and inflame” in relation to place names, topographical features, and the politics of identity (Monmonier 2006). In nineteenth-century Great Britain and in Ireland, Ordnance Survey was the state institution most involved with the naming of places, its acts of mapping often involving the transliteration of names from Irish, Welsh, or Scottish Gaelic into English substitutes, sometimes at the loss of earlier name forms or of an originating meaning (Withers 2000). In Britain, a Permanent Committee on Geographical Names for British Official Use was established in 1919. In the United States today, names are decided upon by the US Board on Geographic Names which reviews proposals for names and renaming and which adjudicates upon federal policy for dealing with toponyms within and beyond that nation’s borders. At the world scale, names are the concern of the United Nations’ Conference on the Standardization of Geographical Names and the UN Group of Experts on Geographical Names (Monmonier 2006, 98).
development of ethnological and anthropogeographical perspectives in geography, in Britain and, in particular, in continental Europe, and with developments in comparative philology. Language mapping in the modern world echoes this earlier tradition and, additionally, is used as a means to chart language shift within and between nations, the rise to global hegemony of English (Crystal 2003), and to plot the world’s linguistic diversity (Asher and Moseley 2007).

Drawing upon such earlier scholarship on language and geography, the distinction may properly be made between the *Sprachenkarte*, where the mapped subject is a language or language group, and the *Sprachkarte*, which is, by convention, a map of linguistic forms (phonetic, morphological, orthographical, synthetic, or lexical) with regard to a specified geographical area. This difference in focus, and in the representation of languages and linguistic variation over space and over time, speaks to the difference between, on the one hand, language geography or, sometimes, the geography of language, and, on the other hand, linguistic geography or geolinguistics. Language geography deals with the distribution over time and space of language as a cultural artifact and with the explanation of such changes in terms of social, cultural, and political processes. Linguistic geography is more the preserve of linguists than of geographers. Linguistic geography is geographical only in the sense that the linguistic phenomena in question are spatially located, although the term has been used to assess the political authority and cultural identity that is rooted in certain forms of, or variations in, a language in modern multicultural contexts such as the speaking of French and the position of so-called minority languages in Canada (Breton 1993). In linguistic geography, conceptualizations of geographic space have tended to focus upon the region as a static “container” of variations in linguistic form, rather than to consider the region or other forms of areal differentiation as contested arenas of social or political meaning as might more widely be understood in contemporary human geography (Britain 2010).

The term “dialect geography” is used within linguistic geography – and is, perhaps, even seen as a subdivision of it – to describe the study of local differentiation in words or expressions within speech (usually within a predefined administrative region rather than a cultural area), and the variable distribution there of given speech forms. In dialect geography as in traditional dialectology, the region is taken as the primary, and, often, the only independent geographical variable. In England, there is a tradition of enquiry in linguistic geography/dialect geography directed at rural, rather than urban, variations in words, expressions, and linguistic elements, and in recording such variations almost as a linguistic form of “rescue archaeology” before the folk practices and tacit knowledge they embody are lost. Notable publications in English linguistic geography include the work of the Oxford philologist Joseph Wright in overseeing the *English Dialect Dictionary*, published in six volumes between 1898 and 1905, and the *Linguistic Atlas of England* (1978), which was the outcome of a survey of English dialects undertaken between 1950 and 1961 by scholars based at the University of Leeds, as also was the *Atlas of English Dialects* (1996). In the United States, the *Dictionary of American Regional English* (DARE), published in six volumes between 1985 and 2013 and based on nearly 1800 recorded interviews undertaken between 1965 and 1970, mirrors the English dialect survey but extends from it in incorporating elements of historical usage as well as contemporary geographical variations in speech forms and lexicon. In Germany, geographers in the University of Marburg have continued and
developed that country’s traditions of linguistic mapping and language geography.

The study of dialectal and linguistic variations within urban settings is concerned more with questions of social class and of internal differences within language as presumed markers of educational status and social hierarchy. In that sense, urban geolinguistics is a form of sociolinguistics, and is a subject of enquiry for sociologists and social psychologists as well as for geographers. The emphasis is upon describing and explaining differences in speech forms in relation to social and cultural circumstances. In such work, the geographical scale of the linguistic enquiry is less commonly an administrative region and is more often a neighborhood, a locale, or, even, a particular community of language users with its own distinctive idiolect (Trudgill 1983). Explanation of specific linguistic registers or idiolects, particularly in an urban context, is widely seen as a matter of social and political conditions. In contrast, some recent work by sociolinguists has argued that varieties in certain speech sounds may be a direct consequence of the natural environment occupied by the speakers (height above mean sea level, and levels of relative humidity), not of cultural circumstance (Everett 2013). Such environmentally deterministic explanations of linguistic variation do not have currency in contemporary geography.

**Language, geography, writing, speaking**

Human geographers have long been attentive to the nuances of language and to the problems of geographical description (Darby 1962). Of course, good writing in geography means more than choosing the right words and knowing how to balance description with interpretation so as to offer clear explanation and promote others’ comprehension. Audiences matter – or, at least, they should – to how geography’s authors write authoritatively. Different styles in geography have been shown to transmit different meanings and, in some cases, to obscure any clear meaning at all (Billinge 1983). In general, such issues are not the concern of physical geographers. For them and other natural scientists, geography’s terminology is largely unchanged over time, is assumed to be unproblematic and to have universal currency, and can be easily manipulated within the overall language of science: *this* is the way the world is because the method, expressed not in words but in the language of numbers, makes it so.

The recognition, most evident within certain quarters of human geography, that there are pervasive and consequential connections between language, social power, and the practice of geography is part of a wider “linguistic turn” within twentieth-century philosophy and social theory. Language, it is claimed, is a medium through which intersubjective meaning is established and negotiated. As with other cultural practices, language does not exist outside social relations of power. There is always a politics of language. These issues are apparent in several ways.

The work of Gunnar Olsson has been influential in demonstrating in geography how the limits to our understanding of the world may, indeed, be the limits of our language. With others, he has argued for the importance of questioning the social and cultural relations we are talking in and through, rather than considering the object of what, as geographers, we are talking about (Olsson 1992). Similarly, issues to do with the politics of language and with the authority vested in language are central to certain strands of ethnomet hodological and qualitative procedures in geography as part of attempts to capture variant or subaltern meanings in the world and so to do
interpretative justice to others’ words and worldviews. The use of particular linguistic registers within geography is thus closely associated with the possibility of the more ethical distribution of resources within society, and, for some, of a more moral geography altogether.

The dominance of the English language as academic geography’s *lingua franca* (see Crystal 2003), and of the hegemony of Anglo-American geographical discourse, has been a subject of concern (for a recent review, see Desbiens and Ruddick 2006). English in its spoken and written forms dominates the subject’s major conferences and the majority of its specialist publications, although several journals incorporate abstracts in French, Spanish, and Mandarin Chinese. Paradoxically, some of the anxieties expressed over the issue of English’s global dominance in geography – the subordination in the discipline of other world languages (French, Spanish, Mandarin Chinese), the authority within the academy of certain styles of writing, even of speaking, and of gender biases in language which permeate geography’s dominant epistemologies – are not always equally a matter of concern for the speakers of other languages, for whom proficiency in English and publication in that language are markers of professional attainment. Even so, and recognizing that much ethnological fieldwork and qualitative methodology in geography requires competence in languages and idioms other than “standard” English, and that “foreign language” training is available to UK postgraduates, for example, most accounts of the subject’s practices and intellectual traditions too commonly assume Anglo-American narratives to be geography’s only voice.

Language is important to the making and remaking of local geographical meaning. This is, of course, particularly the case in contemporary human geography as the examples above indicate, and because nearly all studies in the field are, in one way or another, dependent upon both the spoken and the written word. Historical scholarship in geography has long been attentive to the power of written discourse and so has privileged archival analysis based on printed and manuscript materials. Nevertheless, it is possible to work with the spoken and the heard, as well as with the written and the recorded, in historical geography. Allan Pred’s examination of the conflicts of meaning articulated in different language usage in nineteenth-century Stockholm emphasized the potential in studying the locally spoken word as an element of geographical significance (Pred 1990). Words used in the workplace, or at home, words used to report upon others or to describe one’s self can be used to supplement the written record. Ogborn has shown how speech acts were, often, quite strongly regulated – in courts of law, for example – and how attention to different forms of speech, and to the speech acts themselves, can disclose the practices by which people made sense of their world: “Also, since speaking is simultaneously, and indivisibly, linguistic and embodied, a matter of ‘performance’ and ‘representation,’ understanding speech practices in the past or the present demands a continued rethinking of those distinctions so as to acknowledge representations as ‘actions themselves’ in the on-going making up of the world” (Ogborn 2011, 110).

In such claims, it is possible to see a shift in emphasis in the study of language in geography away from the mapped presence of different languages or of variations in the form and content of a language – whose language, dialect, or idiolect was where, and why, in given national or regional spaces? – toward an understanding of the power of words, within geography as a discourse, and in the world more widely. Rather than map *where* words were, geographers might more commonly now ask what work do words *do?* As McGeachan and Philo (2014) argue, it is
vital to recognize that different types or orders of words embody different ontological status relative to one another, and that different words embody and represent different ways of being in the world. As illustration, they explore the words “chatter,” “testimony,” “lore,” “report,” “opinion,” and “discourse,” and the different levels of authority and power that are inscribed within them and the work to which these words speak.

Language in its different forms and words, spoken and written, has an everyday and enduringly powerful presence. Mapping and naming the world, giving meaning to the world, and being in the world, are all embodied in the words and in the languages which geographers and others use in thinking about, writing about, and talking about the world.

SEE ALSO: Anthropogeography; Areal differentiation (or chorology); Cartography: history; Cultural diffusion; Discourse; Ethnography; Historical geography; New cultural geography; Place; Toponymy

References


Late Cenozoic polar glaciations

Rob McKay
Victoria University of Wellington, New Zealand

The commonly held view of the evolution and development of major glaciation in the Earth’s polar regions is that the East Antarctic Ice Sheet (EAIS) first formed near the Eocene/Oligocene boundary 34 million years ago (Ma) and was highly oscillatory until the middle Miocene (≈14 Ma). This was followed by development of the marine-based West Antarctic Ice Sheet (WAIS) in the Late Miocene/Early Pliocene (≈7–5 Ma) and major ice sheets covering large areas of North America and northern Eurasia by the middle Pliocene (≈2.7 Ma; Figure 1). Because of the difficulties in obtaining direct geological records from the polar regions, either due to extensive ice coverage or due to widespread glacial erosion, this view of Cenozoic glaciation has largely been derived from proxy records far removed from the polar regions (Figure 2). Many of these proxies, such as marine oxygen isotope records and sequence stratigraphy–based sea level reconstructions, provide excellent, indirect records of changes in the globally integrated volume of past ice sheets. However, these cannot provide information regarding the geographical location and extent of past ice sheets at either pole. More detailed information regarding the onset and subsequent variability of major glaciation in each hemisphere is critical in order to understand the cause of polar glaciations through the Cenozoic, as well as the potential sensitivity of ice sheets to environmental change not only in the geological past, but also in the future.

Early Cenozoic initiation of Antarctic glaciation

The Paleocene and Eocene epochs of the early Cenozoic (65 to 34 Ma) are commonly referred to as the “greenhouse” world, on account of highly elevated concentrations of CO2 (>750 ppm; Figure 2). Antarctica has been centered over the South Pole for the past approximately 80 million years, yet during the peak of the early Eocene warm period (≈55 Ma), pollen-based and lipid biomarker proxy reconstructions indicate a subtropical to temperate climate, with winters that were essentially frost free. Following this peak warmth of the Cenozoic, marine oxygen isotope records indicate a global cooling trend that appears to have culminated in a rapid climatic and ocean cooling at the Eocene/Oligocene boundary 34 Ma (Figure 2). However, the exact age of the onset of Cenozoic glaciation on Antarctica currently remains unconstrained. Numerical ice sheet models point to nucleation of highly dynamic ice caps centered on East Antarctic highlands, which could have been present throughout cooler orbital configurations during the relatively cooler time intervals of the Paleocene and Late Eocene. This is because geological records are inherently biased towards marine records, and therefore if these early ice sheets were entirely land-based, there would be no iceberg-rafted debris or glacial signal deposited in these marine...
sediments. However, far-field sea level records suggest that sea level variations on the order of tens of meters may have occurred during the Paleocene and Eocene on timescales <1 million years, which can best be attributed to glaciation. Despite a lack of age control, evidence for the earliest phase of Cenozoic glaciation is still preserved beneath the modern-day EAIS. Geophysical surveys of the Gamburtsev Mountains, which are currently covered by the EAIS, indicate the presence of a pre-glacial fluvial landscape as well as alpine-style glacial valleys formed during the earliest phases of Cenozoic glaciation in Antarctica. The clearest indication for widespread intensification of Antarctic glaciation is a two-stepped approximately 1.5‰ increase in marine oxygen isotope values over a time period of approximately 400 kyr at the Eocene/Oligocene boundary (33.4 Ma; Figure 2). This isotopic shift is interpreted as a 4°C cooling in deep ocean temperature occurring alongside an increase of 80 m of sea level equivalent in ice volume being emplaced on the Antarctic continent. This is a larger volume of ice than is currently on Antarctica, which is difficult to reconcile without invoking the presence of Northern Hemisphere Ice Sheets at this time. However, prior to extensive glacial erosion and subsidence due to tectonic rifting much of West Antarctica, which is now largely below sea level, was subaerial at this time. This allows for the development of high-elevation terrestrial ice sheets covering the entire Antarctic continent in the Oligocene, and ice sheet models...
run on this reconstructed topography can now reconcile this apparent offset without invoking major Northern Hemisphere Ice Sheets.

Glacimarine and subglacial diamictons recovered from the East Antarctic continental shelf, and from iceberg-rafted debris in the Southern Indian Ocean, confirm that a continental–scale EAIS had reached the continental shelf edge and was calving icebergs into the Southern Ocean by the earliest Oligocene (≈34 Ma). However, glacimarine sediments from drill cores in the western Ross Sea may be as old as the Late Eocene, indicating that localized glaciation may have been occurring in the Transantarctic Mountains prior to the Eocene/Oligocene transition.

Figure 2  Late Cenozoic climate history from: (top) direct geologic records of East Antarctic Ice Sheet (EAIS), West Antarctic Ice Sheet (WAIS), and Northern Hemisphere Ice Sheets (NHIS); (center) δ¹⁸O values of benthic foraminifera – lower values represent comparatively warmer oceanic temperature and reduced ice volume (modified from Zachos, Dickens, and Zeebe 2008); (bottom) reconstructed atmospheric CO₂ concentration derived from the geochemistry of marine plankton (modified from Zhang et al. 2013). Modeled CO₂ thresholds for ice sheet initiation are from DeConto et al. 2008.
The tectonic opening of the Southern Ocean, in particular the “gateway” between Australia and the Antarctic, has long been considered to have thermally isolated the Antarctic continent, preconditioning it for glaciation. However, general circulation and ice sheet models suggest tectonic changes may have played a less direct role than altering oceanic heat transport, and point to a potentially related change in the carbon cycle being more important. These models favor a mechanism whereby declining atmospheric CO₂ decreased to a certain threshold value, lowering the radiative forcing of incoming solar insolation and thereby setting the scene for glaciation (Figure 2). Once this threshold was met, it enabled the rapid development of large ice caps on Antarctica’s highland regions, which eventually coalesced to form a single continental-scale ice sheet. Once formed, such an ice sheet is difficult to melt, due to height-mass balance and albedo-related feedbacks. However, marine isotope records indicate that, following the initial inception of the EAIS, large, orbitally paced variations in ice volume did occur throughout the Oligocene and Miocene.

Neogene cooling and development of a “stable” EAIS

Sometime around the Oligocene/Miocene boundary, the vegetation in the Transantarctic Mountains shifted towards an increasingly tundra-style landscape (between 25 and 17 Ma), and stratigraphic drill core evidence indicates this cooling was associated with a large erosional event offshore of the Transantarctic Mountains. This coincides with a 200-kyr-long positive marine oxygen isotope excursion, and appears to represent a glacial expansion of the EAIS close to or exceeding its present-day volume. Terrestrial outcrops and geomorphology in the Transantarctic Mountains provide evidence of highly voluminous subglacial meltwater outburst floods from the EAIS margin prior to 14 Ma, consistent with a warmer, more variable ice sheet at this time.

The approximately 1‰ enrichment of marine oxygen isotope value at approximately 14 Ma is generally regarded as representing the transition from an oscillatory to relatively permanent EAIS. Geological outcrops in the Transantarctic Mountains indicate a permanent transition from wet-based to dry-based glacial deposition, the extinction of the Antarctic tundra and Nothofagus vegetation in the Transantarctic Mountains, and the development of a hyperarid polar environment at this time. However, while there is no evidence of large subglacial meltwater outburst floods at high elevation in the Transantarctic Mountains after 14 Ma, late Miocene sedimentary sequences in Ross Sea drill cores and outcrops in Prydz Bay contain glacimarine facies that indicate that significant discharge of sediment-laden subglacial meltwater from the low-elevation, marine termini of EAIS outlet glaciers persisted well into the Late Miocene (11.6–5.3 Ma). Tundra vegetation persisted in the Antarctic Peninsula region until at least 12.8 Ma.

The relative stability of the EAIS since the cooling event at 14 Ma has been disputed for the past three decades, with some studies suggesting widespread loss of the EAIS as recently as 3 million years ago. However, this is at odds with hyperarid polar climates in the Transantarctic Mountains through this time, and far-field marine isotopic variations. Far-field sea level estimates suggest sea levels were approximately 20 ± 10 m higher than present during the mid-Pliocene (≈3 Ma), inferring loss of both the WAIS and the Greenland Ice Sheet, and up to approximately 10 m sea level equivalent (s.l.e.) from the EAIS (currently 53 ms.l.e.). This sea level budget from EAIS can be reconciled by the
presence of uplifted marine deposits from the East Antarctic margin indicating there was some limited retreat of the EAIS marine margin during the Pliocene, and by geochemical fingerprinting of muds deposited offshore of Wilkes Land in East Antarctica that indicate erosion by an EAIS that had retreated inland during the Early Pliocene, potentially by hundreds of kilometers.

**West Antarctic Ice Sheet evolution**

An increase in the delivery of terrigenous turbidites in the late Miocene/early Pliocene to the Weddell Sea Abyssal Plain is commonly interpreted as representing the first major expansion of the marine-based WAIS. However, seismic studies and drill core studies in the Ross Sea indicate the presence of large ice caps covering Marie Byrd Land in West Antarctica, as well as localized highs in the Ross Sea since Oligocene times (34–23 Ma). As mentioned earlier, it is likely that much of West Antarctic was above sea level at this time and this is a critical boundary condition to identify as it allows for growth of terrestrial ice sheets in a much warmer climate than would be required to initiate marine-based ice sheet growth (i.e., a higher CO₂ threshold in Figure 2). Seismic reflection data to reconstruct the paleogeography exist around the entire West Antarctic margin, but their ages are only (partially) constrained by drilling in the Ross Sea and Antarctic Peninsula. Deep Sea Drilling Project site 270 in the central Ross Sea indicates a terrestrial to marine transgression in the central Ross Sea occurred approximately 27 Ma. Seismic profiles in the central Ross Sea further indicate that middle Miocene glaciation was characterized by local ice caps on basement highs, and that the continental shelf was shallow and seaward dipping. A major shelf-wide unconformity in the Ross Sea provides the first evidence for a greatly expanded marine-based WAIS, although the age of this event(s) is poorly constrained and could have occurred anytime between 14 and 4 Ma, with a potentially similarly timed event also occurring in the Antarctic Peninsula. Such an over-deepening of the continental shelf in the Late Miocene/Early Pliocene allows for greatly increased warm-water intrusion across the continental shelves of Antarctica which is the key process that acts to melt marine-based ice sheet. Thus, there is potential that the WAIS may have been greatly reduced in size during late Miocene to Early Pliocene times compared to early times, as the climate and oceans were too warm to support regular advances of marine-based ice sheets (Figure 2). Drilling beneath the modern-day Ross Ice Shelf as part of the ANDRILL program provided the first direct evidence that the marine-based WAIS advanced and retreated during the Pliocene, but was probably reduced to its terrestrial remnants during the warmest periods of the Pliocene. However, the WAIS was a more persistent feature, albeit with orbitally paced retreat events, after a Pliocene cooling event between 3.3 and 2.5 Ma.

**Paleogene glaciation in Greenland and the Arctic margin**

The evidence for the earliest onset of Cenozoic glaciation in the Northern Hemisphere is even more sparse than for the Antarctic, but it is widely believed that major ice sheets covering large areas of North America and northern Eurasia did not appear until the middle Pliocene at approximately 2.7 Ma. Diatoms and ice-rafted debris infer the presence of sea ice in the Arctic Ocean Basin during the middle Eocene (47.5 Ma). Dropstones with surface textures indicative of glacial erosion are present in late Eocene (38–30 Ma) marine sediments from drill
cores in the North Greenland Sea and Arctic Ocean, arguing for the presence of marine terminating East Greenland or Svalbard ice caps and mountain glaciers. Evidence for glaciation since this time is sparse until the mid-Miocene when ice-rafted debris increases at 16 and 14 Ma in Arctic Ocean cores. A core collected offshore of Southeast Greenland records sporadic pulses of ice-rafted debris from 7.3 Ma until the Late Pliocene (≈3.6 Ma) when ice-rafted debris becomes ubiquitous. Although potentially biased by poor drilling recovery, these data indicate that Greenland and the High Arctic have supported ephemeral glaciations since Late Eocene times.

The establishment of larger and more persistent glaciation did not likely occur in the Northern Hemisphere until the Late Pliocene. In the context of the general cooling and lowering atmospheric CO₂ trend of the Cenozoic, modeling studies support the notion that the atmospheric CO₂ threshold (i.e., radiative forcing) of Northern Hemisphere glaciation is significantly lower (≈280 ppm) than that for the Antarctic (≈750 ppm), and helps to explain a delayed onset of Northern Hemisphere glaciation relative to Antarctica (Figure 2).

Late Pliocene development of a bipolar glacial world

The Late Pliocene and Pleistocene is characterized by higher amplitude glacial–interglacial cycles and gradually increasing values in the marine oxygen isotopes, interpreted to be related to the onset and intensification of Northern Hemisphere glaciation. This increase is associated with a major increase in ice-rafted debris in deep-sea cores offshore of Greenland and North Atlantic. The increase in iceberg-rafted debris offshore of Northwest America at 2.7 Ma, and glacial outwash gravels dated at 2.64 Ma, suggests the Cordellian Ice Sheet may have been a nucleation center of the early North American ice sheets. The exact timing of the development of the Laurentide Ice Sheet over much of North America is more difficult to constrain due to the difficulty of distinguishing the source of ice-rafted debris in the North Atlantic. However, a recent provenance study has suggested that ice rafting in the North Atlantic was dominated by Greenland and Nordic Sea sources, with little input from North America until after 2.64 Ma. However, this indicates that the major centers of major North Hemisphere glaciation (Greenland, Fennoscandian, and Cordellian Ice Sheets) were all covered by large ice sheets at 2.64 Ma. An important caveat in the iceberg-rafted debris evidence is that it does not preclude the gradual onset of terrestrial ice sheets prior to this time. However, evidence for a lack of major glaciation prior to 2.7 Ma in the Northern Hemisphere is further supported by a recent drill core record from Lake El’gygytgyn in Northeast Arctic Russia that indicates that Arctic cooling was insufficient to support large-scale ice sheets until the latest Pliocene (2.64 Ma).

Early Pleistocene advances of the Laurentide Ice Sheet appear to have extended further south than Late Pleistocene advances, with expansions to approximately 39°N around 2.4 Ma. Because Early Pleistocene advances were volumetrically smaller than later ones, the elevations of these ice sheets must have been lower, implying that the glacial dynamics (i.e., flow) of these early ice sheets were different. This shift in glacial dynamics has primarily been attributed to the unconsolidated and deformable nature of the subglacial sediments that were eventually eroded back to bedrock by glacial erosion during subsequent glacial advances.

Hypotheses for the cause of Northern Hemisphere glaciation are numerous, but as yet this remains unresolved. Proposed mechanisms range
from shifts in oceanic and atmospheric circulation patterns arising from tectonic changes, such as the closure of the Isthmus of Panama or the Indonesian oceanic throughflow, through to declining atmospheric greenhouse gases that allowed the build-up of glacial ice in both hemispheres (Figure 2). However, these changes in Earth’s boundary conditions only pre-conditioned the Earth towards bipolar glaciation. Another key variable is the changes in the orbital configuration of the Earth, which regulate how the distribution of incoming solar insolation is distributed on the Earth’s surface over the course of a year. These orbital cycles operate at various timescales – from the well-known 20, 40, and 100 ka cycles to longer term cycles exceeding 1 Ma.

Building on the earlier ideas of Joseph Adhémar and James Croll, Milutin Milankovitch first developed his orbital theory of the ice ages in the 1910s, when he suggested that build-up of continental Northern Hemisphere ice sheets was the consequence of changes in the seasonality related to periodic shifts in the Earth’s orbit. His theory reasoned that reduced seasonality – that is, a warm winter but cool summer – would result in glaciation because the winter snowfall would be able to survive through the course of summer, and this would eventually build up to form an ice sheet. However, it was not until the 1970s that a relationship between the ice age cycles and changes in Earth’s orbital configuration was clearly recognized. From this, it became apparent that orbital changes initiated a series of oceanic, atmospheric, and ice sheet feedbacks that greatly amplified the influences of changing solar insolation over time. Despite the clear orbital pacing of the ice age cycles, fundamental questions remain regarding the response of the Antarctic and Northern Hemisphere Ice Sheets to orbital variations. Firstly, why did 40 kyr frequencies (relating to changes in the axial tilt of the Earth) dominate glacial cycles between 0.8 and 3 Ma, with precession (≈20 kyr) signal only playing a minor role despite its stronger influence on summer insolation. Perhaps more puzzling is that after 800 ka, the pacing of ice ages shifted to 100 kyr cycles, paced by eccentricity (i.e., changes in the circularity of Earth’s orbit) despite this being the weakest influence on solar insolation. This later shift indicates that there is a strong nonlinearity in how changes in incoming solar insolation at the top of Earth’s atmosphere are transferred to ice sheet growth and decay on the surface of the Earth. Numerous theories exist to explain the orbital pacing of the ice ages, but efforts to comprehensively resolve these puzzles are hampered by a lack of direct long-term continuous records of ice sheet variability in either hemisphere (see the review by Raymo and Huybers (2008) for more detail on these issues).

The Mid-Pleistocene transition and further glacial expansion

The transition from 40 kyr to 100 kyr glacial cycles was also associated with the final phase of glacial intensification, when an increase in the amplitude of the marine oxygen isotope cycles implies that sea level variations through a glacial-interglacial cycle increased from <80 m to >120 m. While most of the expansion is related to expansion North Hemisphere Ice Sheets, notable cooling is also apparent in the Southern Hemisphere, with an intensification of the sub-polar westerlies and expansion of the Southern Ocean sea ice belt. With the possible exception of the Greenland Ice Sheet, the Northern Hemisphere Ice Sheets likely experienced full glacial retreat during most of the Late Pleistocene interglacial periods. However, the response of WAIS during the Late Pleistocene is entirely unconstrained. The presence of Pleistocene
LATE CENOZOIC POLAR GLACIATIONS

diatoms recovered from beneath areas of WAIS now grounded indicates open marine conditions occurred in the interior of West Antarctica some time(s) after 1.3 Ma, but the exact timing of this event(s) is equivocal. A likely candidate is the last interglacial period, with far-field sea level reconstructions suggesting Marine Isotope Stage 5e (≈120 ka) likely had global sea levels >6 m above present. Recent ice core evidence suggests that only a modest reduction (≈2 ms.l.e.) in the size of the Greenland Ice Sheet occurred at that time, and thus sea levels require a loss of the marine-based sector of the WAIS. Sediments in the Ross Ice Shelf indicate the WAIS has oscillated between a grounded ice sheet and floating ice shelf since 800 ka, but there is no evidence of open marine conditions during the past approximately 1 Ma. To date, Marine Isotope Stage 31 (≈1 Ma) provides the only conclusive evidence of the last major WAIS “collapse” event.

The Last Glacial Maximum saw ice sheets in both hemispheres reaching their maximum positions between 26.5 and 19 ka (Figure 1). Following this glacial maxima, ice sheet retreat began for most of the Northern Hemisphere ice sheets at 19 ka, but was delayed in Antarctica until after 15 ka. However, the exact chronology of retreat of ice sheets in each hemisphere is poorly constrained and remains hotly debated. Accurately reconstructing these chronologies is essential to identify the mechanisms that initiate glacial retreat, in particular rapid “collapses” of the marine-based ice sheets. Rapid collapses of these ice sheets are evident in coral reef records of post-Last Glacial Maximum eustatic sea level rise – with one event in particular the focus of much attention. This event, termed Meltwater Pulse 1A, saw a sea level rise of up to 4 m/century for a period of 300 years, yet there is little consensus over the source region of this rapid input of glacial melt in the global ocean. However, identifying the source of this meltwater pulse is relevant for potential scenarios of future sea level rise. If the source was in the Northern Hemisphere, then much of the ice capable of rapid retreat (i.e., marine-based ice) no longer exists. However, if it came from Antarctica, there remains 22 m of sea level equivalent in ice sheets grounded below sea level in both West and East Antarctica. Some of this Antarctic ice is thought to be capable of more rapid retreat than the rates that we are currently experiencing (0.3 m/century), in particular that of the marine-based sectors of the WAIS (3.4 ms.l.e) which sits on an over-deepened, reverse-slope continental shelf – a configuration that is particularly vulnerable to rapid retreat via marine instability mechanisms.

SEE ALSO: Antarctica; Climate change and permafrost; Glaciations

References


Further reading


Description and purpose

The Geographic Society of Latvia is a non-governmental organization whose mission is to advance geography in Latvia. This mission is accomplished through a broad array of activities. The society coordinates scientific research in the field of geography carried out in different institutions, facilitates the advancement of methods of teaching geography in schools, coordinates the development of geographical terminology and toponymy research, encourages geographical local lore studies, national tourism, and travelling in Latvia, contributes to understanding geography and geology in Latvia and beyond, as well as popularizing geo-ecological and environmental knowledge and encouraging environmental awareness.

Journals or major publication series

Ģeogrāfiski Raksti, Folia Geographica. www.geo.lu.lv/petnieciba/geografiskieraksti

Current activities or projects

The principal activities of the society are the organizing of geographical workshops, conferences, regional symposiums and Latvian geographical congresses (the Latvian geography congress held in 2012 had 10 thematic sessions and symposia with 105 presentations); facilitating teaching and learning of geography, cooperating with the geography teachers association, and school geography olympiads; membership in the International Geographical Union (IGU) and the Association of European Geographical Societies (EUGEO) and participating in IGU congresses and conferences (the 32nd International Geographical Congress program had 12 presentations from Latvia); and publishing of Ģeogrāfiski Raksti, Folia Geographica.

Brief history

The Geographical Society of Latvia was founded in 1923. Its main purposes were to enhance geographic awareness and disseminate geographical knowledge across Latvia and beyond, to explore Latvia’s geography by organizing and carrying out research work, and to coordinate Latvian geography teachers for advancing the evolution of scientific and school geography. The society held regular meetings, geographical exhibitions, trips, as well as visiting scientific conferences. The activities of the LGB were suspended in 1940.

In 1952, a branch of the Geographical Society of USSR was established in Latvia. Its objectives were to develop and popularize geographic knowledge, to coordinate scientific research in the field of geography, and to promote the development of national economy. The society’s members explored Latvia’s natural resources.
and took part in physiographic and economic regionalization of Latvia’s territory.

In 1990, the general meeting of the Geographical Society of Latvian SSR made a decision to restore the Geographical Society of Latvia as an independent, voluntary, and creative organization. The society submitted the membership application to IGU in 1995. The Geographical Society of Latvia was officially accepted as a full-fledged member of IGU in 1996 at the 28th International Geographical Congress held in The Hague.

Submitted by Zaiga Krišjāne
Legal geography

Benjamin Forest
McGill University, Canada

Legal geography is a heterogeneous subfield that is defined partially by its topic (law) and partially by its approaches to the study of that topic. Some scholarship defines itself explicitly as “legal geography,” while other work addresses legal issues but may be typically categorized as political, social, or critical geography. Moreover, applied geographic work dealing with environmental regulation, natural hazards, or resource management typically address bylaws, codes, regulations, and/or laws but is not generally classified as “legal.” None of the major English-speaking professional associations (American Association of Geographers, Royal Geographical Society, International Geographical Union) have specialty groups or commissions devoted specifically to legal geography, although there are periodic attempts to form them. (The Institute of Australian Geographers does have a Legal Geography Study Group, however, and there are periodic attempts to form a specialty group in the AAG.) Moreover, law itself is a complex object of study that includes legislation (law-making), jurisprudence (court decisions and legal doctrine), police and policing, and legal actors (attorneys, magistrates, and judges). There are also various types of law, such as civil law (including property law and environmental regulation), constitutional law, criminal law, and international law. Customary law, which may be more or less formal, is also an important subject of study, particularly in developing areas and states. Legal geography is further fragmented because formal legal systems and practices are typically defined by sovereign state boundaries, making comparative or systematic work relatively rare. Finally, legal scholars may address “geographic” topics or use “geographic” concepts, with or without reference to work in the discipline of geography. For the most part, however, this entry focuses on work that is self-defined as legal geography, published by geographers or in geographic journals.

Defining legal geography

The subfield remains highly fragmented despite several attempts to institutionalize legal geography as an area of study. Gordon Clark initiated a special issue of Urban Geography in 1990 on legal geography, coedited with Nick Blomley, an effort that continued with the “Legal Geographies” series in Urban Geography, starting in 1993 under Blomley’s editorship. Their original impetus was to draw attention to the consequences of Brown v. Board of Education, the 1954 US Supreme Court decision that ended de jure segregation in American public schools. (Urban Geography subsequently published a special issue to mark the decision’s fortieth anniversary in 1994.) The Legal Geographies Series is marked by particular attention to major court decisions in Western liberal democracies, and hence tends to emphasize geographic analyses of formal jurisprudence. Authors in the series are primarily geographers or scholars well connected to the discipline, rather than legal scholars publishing on geographic issues. Nonetheless, the original and subsequent editors adopted an open perspective, so publications in the series have tended to reflect...
Legal geography has also been used as a starting point in efforts to critique and reshape the social sciences more generally. In the mid-1980s Gordon Clark (1985) drew on the practice of judicial reasoning in local autonomy cases to advocate an interpretive approach to social theory. More recently, David Delaney (2010) has developed the concept of the “nomosphere,” a way to characterize the interpenetration and mutual constitution of law and space, and to provide a unifying framework for legal geography. His ambitious, programmatic book emphasizes the way in which law and space are both performed together, and thus “worlded,” creating legal spaces (“situations,” “settings,” and “nomoscapes”) such as “home” and “public/private space.” In addition, he draws attention to the role of “nomospheric technicians” (especially lawyers and judges) who have privileged positions to manipulate, shape, and change the nomosphere. This framework has received attention in conferences and favorable reviews, but has not (as yet) achieved a significant presence in scholarly publications.

Several edited books have also sought to define – or at least describe – the contours of legal geography. Blomley, Delaney, and Ford’s *Legal Geographies Reader* (2001), edited by two geographers and a legal scholar, is probably the first comprehensive attempt, and is divided into sections on public space, racism, property, state formation, and environmental regulation. A second collection, edited by legal scholars and published as the fifth volume in the series *Current Legal Issues*, adopts a similar categorical approach, with sections on boundaries, land, property, nature, identity, culture and time, and knowledge (Holder and Harrison 2003). Scholarship in both works reaches well beyond disciplinary geography and includes numerous contributions from legal scholars and legal anthropologists; indeed, in the latter work, there are only three geographers contributing to its 27 chapters.

The inclusion of legal anthropology in such collections broadens the treatment of law beyond formal jurisprudence and beyond liberal states, by including studies of customary law in developing societies. For example, the phenomenon of legal pluralism (where formal legal systems coexist, and often conflict, with customary ones) is a common topic in legal anthropology. While legal pluralism occasionally appears in work by geographers or in geography journals, generally the scholars and scholarship on such topics are not well connected to disciplinary geography.

Similarly, it is rare for a legal or sociolegal scholar to publish in geography journals. Law review articles address a vast array of topics, including many “geographic” ones, but the practices of publication and scholarship in the two disciplines means that there is little incentive for law faculty to publish in social science journals, and it is very rare for social scientists to publish in law reviews. Likewise, although sociolegal scholarship often touches on geographic subjects, and even employs geographical terms such as “place,” it remains generally disconnected from disciplinary geography. While geographers do publish in sociolegal journals on occasion (and vice versa), this remains exceptional. As with the relationship between geography and many other disciplines, when legal scholars appeal to “geography” they tend to hold a highly conventional view of the term to mean basic spatial relationships, cartography, or even simply physical geography.

In short, while the term “legal geography” can potentially cover a broad range of formal and informal legal and regulatory systems, a large set of disciplines and disciplinary approaches,
and a variety of legal-political regimes, in practice the term refers to a more narrow set of scholarship by practicing geographers primarily (but not only) concerned with common law jurisprudence in developed Western democracies. Within that set, however, there are several areas of significant focus, including race, property, and policing.

Race

Arguably, the subdiscipline crystallized around questions of race and law in the United States (and, to a lesser degree, in Canada), and the close association between racial power and spatial segregation makes the topic a natural one for geographers. For early work in legal geography, racial segregation was a vehicle to explore the nature of local autonomy in the United States, and was arguably an extension of political geography’s conventional interest in the definition of territory and jurisdiction (Johnston 1984). While much of the scholarship in legal geography focuses on the latter half of the twentieth century, David Delaney’s work is notable because he analyzes legal struggles over racial segregation in both the pre- and post- Brown eras. Some work focuses specifically on residential segregation itself, while other research considers its consequences, particularly for education and political representation.

The American case presents several important legal conflicts and dilemmas. First, the federal structure of the United States means that there are often conflicts between federal and state law, and racial issues have arguably shaped these conflicts from the first drafts of the Constitution. As the ultimate arbiter of constitutional meanings, the US Supreme Court’s decisions often have significant geographic consequences, and rely on (assumed) geographic relationships.

Second, amendments to the Constitution in the post-Civil War era (especially the 14th Amendment, adopted in 1868) provided a formal guarantee of racial equality. Nearly all local and state jurisdictions, and (within a short time) federal institutions, were deeply hostile to this principle. Consequently, courts were receptive to various legal strategies to defend white supremacy. In the landmark Plessy v. Ferguson (1896) decision, for example, the court ruled that racial segregation was permissible if facilities were allegedly equal for whites and African Americans. Similarly, while the court ruled that cities could not establish apartheid-like racial exclusionary zones (Buchanan v. Warley 1917), it permitted systematic private discrimination in homeownership and sales until the mid-twentieth century.

Geographers have focused on two aspects of legal reasoning in such conflicts. First, the law’s use of abstract language and categories, such as formal equality, to mask the lived experience of racial discrimination. Here “the law” as a force of abstraction is opposed to “geography” as a carrier of specific experience. A second and often related principle is the separation of public and private actions. Particularly in the American tradition, constitutional measures can only affect governmental activities, so anything that is deemed “private” is beyond the reach of (constitutional) law. What matters, for example, are the formal rules used by a public body rather than the practical outcomes. Under this view, as long as a city, for example, does not regulate the ownership of property on the basis of race, there is no constitutional violation, even if it remains impossible for an African American to buy from a willing seller. Of course, the outcome of such cases often, crucially, turns on what courts define as public or private.

This highly formalized approach was challenged fundamentally in the Brown v. Board of Education (1954) decision, which essentially
overturned *Plessy*. The concurrent Civil Rights Movement also led to a raft of federal legislation over the next 15 years that expanded restrictions on “private” discrimination (in housing sales and rentals, public accommodations and services, mortgage provision, employment, and the like). Nonetheless, states and local jurisdictions used a variety of legal and quasi-legal strategies to maintain the de facto racial separation of whites and African Americans. Courts could order the desegregation of schools within a particular jurisdiction, for example, but refused to do so across municipal boundaries. By moving out of racially mixed cities to mostly white suburban jurisdictions, whites could isolate themselves from African Americans and other minorities. School segregation became driven by racialized housing patterns, whose causes were judged too complex to address through legal regulation (Johnston 1984). Whites (of sufficient economic means) also had the option of withdrawing from the public school system entirely by sending their children to private schools, which could maintain racial exclusivity through high fees, entrance requirements, and/or informal barriers.

Constitutional issues are also often at the center of disputes over racial identity and political representation, and in some ways they mirror the pattern of residential and school segregation. In the United States, the definition of electoral constituencies (districts) is ostensibly a political decision, typically under the control of state legislatures or local jurisdictions. In the mid-1960s federal courts became receptive to legal challenges based on unequal populations (malapportionment) and on racial discrimination (racial gerrymandering). Over the next 20 years or so, courts defined and refined constitutionally mandated standards for election districts. The jurisprudence of equal apportionment and racial gerrymandering both have complex histories, but the current requirement for (approximately) equal populations between districts is clear and settled law, while the use of race in redistricting remains complex.

Starting in the 1950s, but especially after the Voting Rights Act of 1964, federal courts began to rule against electoral redistricting plans that clearly weakened the voting strength of African Americans. Classically, such gerrymandering was a strategy of “cracking,” which divided minorities between two or more districts so that they did not form a majority or effective plurality in any constituency. These cases primarily concerned minority populations that were reasonably large and compact, so the principal empirical questions involved measures of voter polarization by race. In short, the “geography” of such districts was assumed, and the major legal question was whether a jurisdiction had discriminated against minorities by reducing their voting strength. By the early 1990s, however, various political and technical developments meant that the principal legal question involving racial gerrymandering shifted squarely to “geography.” In essence, it became possible (and often politically expedient) to create very irregular and/or elongated districts with African American or Latino majorities. Whether such districts were efforts to fairly represent minority interests, or unconstitutional classifications of voters by race, turned on legal judgments about the geographic nature of political communities. Generally, the Supreme Court hewed to a fairly conventional conception of geographical community, requiring some degree of propinquity but without articulating a formal standard.

The scholarship on race and law in geography strongly reflects an emphasis on jurisprudence, and on US constitutional jurisprudence in particular. This work typically analyzes the geographic concepts used in legal reasoning, including ideas that are literally geographic (compactness, scale,
boundaries, etc.) and figuratively so (public and private).

**Property**

Research on property draws on the analysis of a similar set of geographic concepts. Arguably, the protection of private property lies at the heart of the liberal legal tradition, so the distinction between public and private is central to scholarship on property. Critical legal geographers – Nicholas Blomley and Don Mitchell in particular – have pursued two major themes in the study of property. First, over the course of several works, Blomley has laid out a perspective on property that emphasizes its link to violence. Rather than being antithetical to violence, law and property are fundamentally rooted in violent (dis)possession. Legal protection of private property serves to freeze the original acts of violence in place. As with racial geographies, he argues that the abstraction of property as a category serves to obscure both its origins and its unequal effects on material relations. Coming from a similar critical perspective, Mitchell has focused on the use of property and property rights to limit the rights and activities of already marginalized groups and classes: the homeless, the poor, and labor.

At the same time, such analyses leave room for the complexities of the law. The 2005 *Kelo v. City of New London* decision over eminent domain provides a good example. Eminent domain is a well-established principle that governments can take private property for public use in return for fair compensation. In the *Kelo* case, the US Supreme Court ruled that the city of New London, Connecticut could use eminent domain to obtain property even if the direct recipient was a private party. The court accepted the city’s argument that the (promised) redevelopment would benefit the city through increased tax revenues, and was therefore a “public use.” The decision attracted intense criticism from both libertarians and civil rights advocates, and would seem to show the power of local government over private property rights. Yet careful analysis of the decision reveals that its logic privileges certain classes of property owners, and serves to reinforce a conventional private property regime.

**Police, policing, and social regulation**

In several ways, geographic scholarship on policing is distinct from studies of either race or property because it includes significant attention to the activity and strategies of policing, rather than to just criminal law or jurisprudence. The ethnographic work of Steve Herbert is especially notable in this regard. His earliest work focused on the territorial strategies used by police, in particular how the demarcation and enforcement of boundaries are a routine, essential, if often unnoticed, technique of contemporary urban policing. Such methods, however, can exacerbate divisions and conflicts between police and residents by creating informal “safe and controlled” and “unsafe and chaotic” zones of the city. Subsequent work examines the practice of “community policing” from the perspective of both police and residents, and suggests why this popular idea faces significant limitations in practice. More recently, Herbert and his collaborators have used ethnographic techniques to study the use of exclusionary zones as forms of social control, discipline, and (allegedly) therapy for drug offenders and sex workers. Such ethnographic work remains relatively rare in legal geography, as does research on legal actors such as attorneys and judges (but see Martin, Scherr, and City 2010).
Research on policing and social control also includes a substantial literature in the jurisprudential tradition, with much of the attention focused on immigration control, homelessness, and speech. At the same time, this work often embraces nonformal and nonconstitutional notions of “rights,” particularly Henri Lefebvre’s concept of a “right to the city.” In doing so, it goes beyond narrow concerns with legal reasoning.

The regulation of immigrants and the control of immigration gained renewed attention after 2001 as the United States sought to “fortify” its borders, but much of the work in legal geography has focused on legal conflicts within the country rather than at the boundaries. As with race, this scholarship concerns conflicts between federal and state (or local) jurisdictions. Although immigration is formally under federal jurisdiction, some jurisdictions have sought alternative means to regulate or exclude immigrants, or have embraced federal efforts to use local police for immigration control. In other cases, however, cities and states have actively resisted federal efforts to identify and deport immigrants.

The regulation of the homeless, especially efforts to exclude them from public space and public services, is another significant focus of legal geography, particularly by Don Mitchell. Work in this area typically examines ways that local jurisdictions make it difficult or impossible for people to live without access to (their own) private property. Restrictions on begging and access to basic hygiene facilities mean that the homeless find it difficult to earn a livelihood or simply care for themselves in public space, even though liberal norms hold that such space is defined by its openness and availability to all. Echoing the scholarship on race, formal rights to speech (in the form of panhandling), legal representation, or simple presence are shown to be ineffective protection for the homeless, in contrast to a more robust concept of a “right to the city.”

While much of the literature on homelessness concerns speech, other research focuses specifically on the right and regulation of speech in public space. Despite formal commitments to free speech, jurisdictions and courts have arguably become increasingly hostile to unpopular speech in public by increasingly recognizing a “right to be left alone” (i.e., to be free from hearing unwanted speech). The legal treatment of restrictions, however, often seems politically driven, with greater restrictions tolerated for anticorporate or antiglobalization protests, and relatively greater protection for religious or corporate speech.

SEE ALSO: Antiracist geography; Corporations and the law; Crime; Environment and law; Homelessness; Political geography; Property and environment; Public space; Race and racism; Residential segregation

References


Further reading


Leisure has and continues to be a term that is difficult to define, but it is most often associated with people’s free time behavior (Rojek 2010). Leisure and geography have long collided and Mercer’s (1970, 261) statement that “leisure remains a sadly neglected area of study in geography” remains relevant. Leisure geographies, opportunities presented by free time or what people believe and plan to do in time away from other obligations (most often assumed to be in juxtaposition with work), are produced, consumed, performed, practiced, embodied, and enacted across time and space. Leisure, as a concept, has become more permeable and now crosses subject and disciplinary boundaries. Geographers, and social scientists more generally, now prefer to define leisure by the activity itself. Leisure may involve cycling, running, traveling, skiing, gardening, shopping, watching movies, engaging with social media, online gaming, reading books/magazines/graphic novels, urban exploring, scrapbooking, sex, drugs, and many other hobbies, activities, and endeavors that are seen to be outside of those things that have to be done.

Yet participation in leisure remains contested. Cultural identities, issues of power, and access, gender, race, sexuality, meaning, and knowledge processes influence people’s understanding of what leisure is and the role it might play in their everyday lives. At the same time, concepts of lifestyle, wellbeing, mindfulness, and happiness are being increasingly extolled as the virtues of the freedom of leisure time. Nonetheless, researchers continue to explore the negotiations, constraints, and obstacles faced by individuals and communities whose ability to access leisure may be partial or nonexistent. As neoliberal societies and policies continue to dominate in many places, the commodification of leisure time increasingly calls into question whether this free time is actually “free.”

Leisure geographies can be seen to have outgrown the crises of leisure studies. They engage an interdisciplinary audience and recognize the plurality of what leisure is and what it means to individuals and society. They increasingly engage with poststructuralism, look beyond the mobile to consider (im)mobility, hybridity, stillness, and stickiness, and consider the more-than-representation. They deal with bodies, senses, performances, and emotions. Methodologically, leisure geographies are innovative in thinking through how to be a part of, or go with, when participating in leisure activities. Helmet cameras, video diaries, photography, and walk-along interviews complement more traditional methods. Yet despite theoretical and methodological growth, there remains a lack of focus on leisure geographies outside of the Western, predominantly Northern Hemisphere, world. Leisure, in all its forms and in its many spaces, needs to be more widely understood and leisure geographies are well-positioned to embrace the challenges to come.

SEE ALSO: Cultural geography; Nonrepresentational theory; Poststructuralism/poststructural geographies; Recreation; Tourism; Work–life balance
References


Further reading

American physicist Theodore Harold “Ted” Maiman invented the world’s first laser in 1960. In the following year, scientists put forward the idea of LiDAR (Light Detection And Ranging). Modern laser-based remote sensing technology began to be used in the 1970s by the US National Aeronautics and Space Administration (NASA), largely aimed at meteorology (John A. Dutton E-Education Institute 2014). Initial LiDAR systems were single-beam profiling devices and lacked accurate georeferencing on terrain mapping (Portland State University 2016). Between 1988 and 1993, the University of Stuttgart, Germany, invented the airborne laser scanner by combining laser scanning technology with a real-time positioning system. With the development of the Global Positioning System (GPS) and the Inertial Navigation System (INS), accurate real-time positioning and attitude determination could be achieved, which made it possible to deploy LiDAR on moving platforms, such as airborne LiDAR. Bathymetric LiDAR was actually one of the first uses of airborne LiDAR (NOAA 2012).

The LiDAR system

A LiDAR system includes a transmitter system and a receiver system. An airborne or satellite-based LiDAR also includes a position and navigation system. Figure 1 shows a basic LiDAR system (Chunlin 2008).

The transmitter system

The transmitter system is composed of beam expanders and one type of laser radiator, such as a carbon dioxide laser, diode-pumped Nd:YAG (neodymium-doped yttrium aluminum garnet), semiconductor laser, and so on. The laser can emit a laser beam with a fixed wavelength at a fixed pulse repetition, and the pulse duration is typically 10 ns. Different wavelengths have different resolutions, and a better target resolution is gained with a shorter wavelength. Each laser generates pulse energy at its wavelength. In general, an energy monitor is located within each laser canister to monitor if the level of the pulse energy is in the normal state. The role of beam expanders is to reduce the angular divergence between the transmitter system and the receiver system.

The receiver system

The receiver system is composed of a telescope, relay optics, electronics detectors, and signal processing devices. A signal returned from a target is received within a field of view (FOV) of 0.2–5 m and the stray light is blocked (Nayegandhi 2007).
LIDAR

The telescope has several mirrors and baffles to control light and to reduce the effect of a thermal gradient. There is also a visor mounted to reject solar illumination. The signal received by the telescope is collected by a hole mirror or a beam splitter (some polarization LiDAR are equipped with a polarization beam splitter) and sent to photodetectors (such as photomultipliers and silicon avalanche photodiodes), which convert the optical signal into an electrical signal. The electrical signal is amplified and digitized for computer processing.

The sensitivity of the receiver system is largely affected by the photodetector. A photomultiplier can provide low dark noise but has ordinary dynamic range, while an avalanche photodiode has good dynamic range but greater dark noise. The photodetector exhibits various quantum efficiencies at different wavelengths; for example, a photomultiplier has much better quantum efficiency at 532 nm than an avalanche photodiode. Figure 2 shows the basic LiDAR receiving system (Hongbo 2008.).

The position and navigation system

LiDAR sensors mounted on movable platforms (airborne, satellite) require the position and navigation system, as they must accurately confirm their absolute position and orientation. GPS and IMU provide good measurements in this situation.

Operation principles

The laser beam transmitted by LiDAR goes through the transmission media where it can be attenuated and scattered. The incident light is immediately reflected and scattered when it reaches the surface of the target. The LiDAR receiver telescope receives the reflected or backscattered signal falling into the FOV and transmitted to the photodetector. Through photoelectric conversion the optical signal is transformed to an electric signal (typically a current signal), which is computed and amplified and then finally displayed and recorded. Figure 3 shows the working principle of LiDAR.

For signals returned at different altitude, based on the increment of time between the laser emitting and receiving, the range from LiDAR to the target can be calculated by equation 1

\[ s = \frac{c \Delta t}{2} \]  

(1)

where \( c \) is the speed of light with a value approximately equal to 0.3 million km/s. Utilizing a combination of the laser height, scan angle, position of the laser determined by GPS, and emission direction determined by INS, the coordinates \( x, y \), and \( z \) of each spot on the ground can be accurately calculated. In general, the ground spacing of LiDAR spot elevations ranges from 2 to 4 m.

The general form of the signal to be calibrated can be written as equation 2, where \( P_R \) is the received laser power (unit: W), \( P_T \) is laser transmission power, \( G_T \) is the transmitting antenna gain, \( \sigma \) is the target scattering cross-section, \( D \) is the receiving aperture (unit:...

---

Figure 1  A basic LiDAR system.
Figure 2  Configuration of a receiving LiDAR system.

Figure 3  The working principle of LiDAR.

\[ m, R \text{ is the distance between LiDAR and the target (unit: m), } \eta_{Atm} \text{ is the one-way atmospheric transmission coefficient, and } \eta_{Sys} \text{ is the system parameter.} \]

\[ P_R = \frac{P_T G_T}{4\pi R^2} \times \frac{\sigma}{4\pi R^2} \times \frac{\pi D^2}{4} \times \eta_{Atm} \eta_{Sys} \quad (2) \]

\[ G_T = \frac{4\pi}{\theta_T^2} \]

where

\[ \theta_T = \frac{K \lambda}{D} \]

\( \theta_T \) is the laser beam width, \( \lambda \) is the laser emission.
LiDAR

wavelength, and $K_a$ is the aperture transmission constant (Dai 2002).

The LiDAR cross-section (LRCS) is the cross-section of a total reflection sphere on which the light intensity generated equals that received by the LiDAR receiver, and it is expressed as

$$\sigma_d = \rho_d \pi r^2$$

(3)

where $\rho_d$ is the spherical reflectivity, and $r$ is the radius of the sphere. Generally, the more closely an object resembles a sphere, the closer the relationship between the LRCS and the size of the object. The LRCS describes the target's ability to scatter the laser beam.

LiDAR characteristics

Small volume and light weight

Compared with the microwave radar, LiDAR’s volume and weight are much smaller. Radar’s volume is huge, with the size of antenna aperture a few meters or even dozens of meters and a weight of several tons. LiDAR’s structure is much more portable; its telescope’s diameter can be a few centimeters and it weighs only tens of kilograms. Assembling and disassembling are relatively simple, and it is easy to operate and maintain.

Good concealment and active interference resistance ability

Laser beams propagate in a straight line in the air and the beam width is very narrow. The information cannot be retrieved unless the receiver is on the propagation path. That is why LiDAR has good concealment. Microwave signals experience interference from electromagnetic waves widespread in nature, but a laser signal can resist electromagnetic interference. In a complex information environment, laser signals are not easily intercepted by enemies.

High resolution

LiDAR has a high angle, distance, and speed resolution. The range resolution can reach 0.1 m, and the velocity resolution can reach 10 m/s or less. Typically, the angle resolution is larger than 0.1 mrad (it is impossible for radar to achieve this), which enables two goals to be distinguished to within 0.3 m at a range of 3 km. In addition, LiDAR can simultaneously detect multiple targets. A high distance and range resolution means that the distance-Doppler imaging technology can be utilized to obtain the target image clearly. High resolution is the most significant advantage of LiDAR, and most applications are based on this characteristic.

Good performance at low-altitude detection

For LiDAR, reflection only occurs in the irradiation propagated area. So, no matter whether it is at high or low altitude, it is not influenced by ground object echo effects. Conversely, due to the influence of ground object echo effects, radar always has a certain blind area that it is unable to detect. LiDAR works far better than radar at low-altitude detection.

LiDAR classification based on LiDAR position height

Ground-based LiDAR

Ground-based LiDAR observation is fixed on the ground, with a high accuracy and signal-to-noise ratio (SNR) and convenient operation and maintenance, but its detection range is very limited.
and its detection height is restricted to no higher than the troposphere layer.

**Airborne LiDAR**

The airborne LiDAR measurement system is a kind of active remote sensing device. It is the leading international technology for acquiring high-precision three-dimensional (3-D) coordinates on the ground and image data quickly and synchronously and realizing the reproduction of the real-time, changing features of objects quickly and intelligently. Figure 4 shows airborne LiDAR in operation.

**Spaceborne LiDAR**

Spaceborne LiDAR can acquire global data and images with a wide range, overcoming the disadvantages of ground-based LiDAR regardless of weather conditions, but its low SNR ratio means that it is difficult to operate and requires a complicated algorithm. At present the most popular spaceborne LiDAR is the Cloud-Aerosol LiDAR and Infrared Pathfinder Observation (CALIPSO) which was launched in April 2006 to provide global vertically resolved measurements of clouds and aerosols. Figure 5 shows the CALIPSO satellite on-orbit and Figure 6 shows the receiver subsystem of the CALIPSO (Hunt et al. 2009).

**Applications**

**Military applications**

Lasers can collect 3-D data (including azimuth, dive angle, distance, speed, and strength information) and can display them in an image. Radiation intensity geometry images, range images, and velocity images can be generated at very high resolution. High-resolution LiDAR can produce real-time images of moving armored cars and the resolution is high enough to identify the vehicle model.

Since the late 1970s, LiDAR has been used in fire-control systems for tanks, artillery, ships, and aircrafts. Many specialists favor automatic tracking LiDAR on account of its precision range, accurate speed measurement, and perfect tracking advantages. Some developed countries have drawn up a plan to precisely track in-orbit satellites’ position and velocity, which provides support for research into gaining information on vehicle size, shape, and orientation from space.

LiDAR can also work for underwater detection. The blue-green laser with a wavelength of 0.46–0.53 μm can penetrate water from several hundred to several thousand meters. Using the high-power narrow blue-green pulse emitted by the laser and the reflected light, the military can obtain the parameters of a target’s velocity and position under the sea in high resolution. During the 1991 Gulf War, the American LiDAR “magic lamp” successfully detected torpedoes and mine chains.
Figure 5  CALIPSO working on-orbit.

Figure 6  CALIPSO receiver subsystem.
Agriculture

Lasers can help farmers make topographic maps which can be used to locate a region of high crop yield. With the maps fed into the computerized, variable-rate fertilizer applicators, farmers can divert more of their costly fertilizers to the highest-yielding zones and the least to the lowest-yielding zones (Comis 2010). This technology is important to farmers and has been tested in many areas.

Engineering survey

Much precision engineering requires a mathematical model for 3-D coordinates, for example, electrical power line selection, mining and tunnel surveys, hydrographic surveys, settlements, construction measurement, deformation measurement, and protection of cultural relics. LiDAR performs well in the detection of 3-D structures of buildings. A synthesis of high-accuracy laser point cloud data, textural information from wide-field cameras, and architectural models enables engineers to develop a building construction plan.

Atmospheric sounding

LiDAR has been widely applied in atmospheric sounding. The principle of atmospheric sounding is to use the backscatter signal of the emitted laser at different altitudes to analyze the vertical distribution of aerosol and cloud particles in the atmosphere. The latest achievement is to detect CO₂. The wavebands currently used include 532 nm visible green light and 1064 nm near-infrared light for aerosols and clouds, and 1.57 μm for CO₂.

There are several kinds of atmosphere detection LiDAR that are classified by measurement method. Mie scattering LiDAR is based on the Mie scattering of atmospheric molecules to detect cloud and aerosol particles in the lower atmosphere. As one characteristic of Mie scattering, the size of the scattering particles is similar to the wavelength of an incident laser and the scattering wavelength is equal to the incident wavelength. There is no energy exchange during this process.

Rayleigh scattering LiDAR can be used to detect atmospheric density, atmospheric fluctuations, and atmospheric temperature, based on a Rayleigh scattering mechanism in the intermediate layer at a height greater than 30 km. Rayleigh scattering happens when the scattering wavelength is much larger than the radius of the scattering particles.

Polarization LiDAR is used to detect and analyze the cloud moisture and the cloud micro-physical structure by measuring the nonspherical particles’ depolarization ratio. Figure 7 shows a polarization Mie LiDAR.

Resonance fluorescence LiDAR has been applied to study sodium and its related properties systematically. It can also be used to
LIDAR detect and study other similar kinds of atoms and ions. Because the resonance fluorescence cross-section is much larger than the Rayleigh scattering cross-section, atomic or molecular resonance fluorescence will be enhanced at some specific laser wavelength, so the atmospheric composition can be detected.

Doppler LiDAR, based on the Doppler effect, is used to measure atmospheric wind velocity. The temporal and spatial variation of the wind field can be obtained by measuring the scattering frequency relative to the Doppler frequency shift of the laser beam.

Raman scattering LiDAR can detect the vertical distribution of water vapor according to the Raman echo signal backscattered from water vapor and nitrogen molecules simultaneously. Raman scattering is a non-elastic interaction process between a laser beam and various molecules in the atmosphere. The major characteristic of Raman scattering is that the wavelength between the scattered light and the incident light is different. This LiDAR can be used to detect atmospheric humidity and temperature profiles at low altitude with high spatial and temporal resolution, and can also be used to measure the concentration of certain polluting gases in the surrounding environment.

Differential absorption LiDAR is used to detect atmospheric humidity and atmospheric pollution. It is widely used in environmental monitoring, such as detecting O₃, SO₂, and NO₂ in the atmosphere.

In recent years, the most popular atmospheric remote sensing LiDAR has been ground-based LiDAR and airborne LiDAR, as they have good SNR. In April 2006, CALIPSO, the world's most powerful spaceborne LiDAR, was launched together with the CloudSat satellite to observe the global distribution and properties of aerosols and clouds (Winker et al. 2009). The biggest advantages of satellite LiDAR are its wide measurement range and that it can provide observation on a global scale. However, it has relatively low SNR compared to ground-based and airborne LiDAR, on account of its very high altitude to geoid.

**Working for the power grid**

In the area of electrical communications network construction and maintenance, one advantage of LiDAR is that it can help find the terrain of the whole circuit design area and the situation of on-ground objects, with which information electrical engineers can design feasible schemes and estimate construction costs. When a disruption occurs in the circuit, LiDAR data can be used to quickly find the problem.

**SEE ALSO:** Distance; Optical remote sensing; Space; Spatiality; System dynamics; Tropical geography

**References**


St Petersburg, FL: ETI-US Geological Survey, Florida Integrated Science Center.

Life course approach

Irene Hardill
Northumbria University, UK

Janice Monk
University of Arizona, USA

Life course research is concerned not simply with exploring chronological age but also with addressing nondeterministic conceptualizations of age, and with individual and collective trajectories of experience in diverse spaces and places and through time as these are shaped by events, roles, memory, and retrospection. Transitions are seen as dynamic rather than determined. The studies attend to continuities, ruptures, and multiple pathways, and are taken up by researchers who focus on the Global South as well as the Global North. Currently the approach is more commonly used by geographers interested in the lives of children and youth, and of older people, than by those interested in midlife. Additionally, greater attention has been paid to examining women’s lives than men’s, though life course approaches have also been used to understand “coupledom” (primarily of heterosexual couples). Attention is paid to particular generations and also to intergenerational relations, to issues such as parenting, transitions from youth to adulthood, individual and household residential movements, grandparenthood, and retirement in older age. Given this breadth of interests, and the recognition that lives are linked, the range of research methods applied is diverse. They include quantitative analyses that often use census materials and other large databases, in-depth qualitative approaches, and, more recently, visual methods such as participatory video and experiments with the interpretation of sketches and diagrams created by research subjects (Fenster 2013). Less common, but emerging, is research that engages with geospatial technologies. The choice of methods is also acknowledged as having political and ethical implications and may adopt a single approach or increasingly be characterized by mixing methods.

Quantitative analysis

The application of quantitative methods in life course research is common in analysis of national or regional patterns of processes such as migration between communities and within countries, and intra-urban residential movements, as these vary across life stages, place, and time. These have considered both individual and household scales and draw on sources such as the long form collected by the US Census that includes data from two time periods, on specialized publicly available datasets such as the American Community Survey (ACS) and the Panel Study of Income Dynamics (PSID) in the United States which, since 1968, has measured the economic, social, and health aspects of a large representative sample of families and their descendants. Outside the United States, similar sources include the British Household Panel Survey, the German Socio–Economic Panel, and the Household Income and Labour Dynamics in Australia Survey. These include data that lend themselves to cartographic and graphic analysis and portrayal, and to statistical analysis with such
LIFE COURSE APPROACH

techniques as logistic regression modeling and multiple regression analysis. By disaggregating groups on multiple variables, they examine interactions, for example by gender, age, marital status, income level, education, and place of origin. In so doing, they demonstrate complexities of choices and behaviors in multiple stages of life beyond traditional notions of youth, midlife, and old age, and also reveal issues of intergenerational relations and changing patterns of marriage. In recent years authors have been able to link individuals within their household context, for example in the United Kingdom this has been possible since the 1991 Census of Population with the publication of household data in the Sample of Anonymised Records (SARs). The insights from these studies can inform and reflect policy changes, identify needs for social programs, and open questions for new research at other spatial scales. Among American geographers who have sustained applications of these methods are William A.V. Clark and David Plane (Plane and Jurevich 2009).

Qualitative approaches

An array of qualitative methods characterizes much life course research. Many studies employ in-depth qualitative interviews to investigate subjective interpretations and understandings of the life histories. The interview may take a number of forms, such as structured, semistructured, or open-ended, with structured and some semistructured interviews gathering interviewees' knowledge and experience of the outside world. In contrast, in more flexible interview formats all participants in interviews are seen as agents who construct meanings subjectively, not as objectively found, a view that is associated with feminist, life history and psycho-social studies. The in-depth interview is a collaborative process, constructed through the unique interaction between the interviewer and the interviewee, and as a technique it has been subject to evaluation and critique. Qualitative interviews, including field notes and digital recordings, need to be transcribed and analyzed. Researchers increasingly use qualitative software packages in combination with reading and rereading transcripts to facilitate data analysis. While such interviews yield rich data, the analysis is not always straightforward and geographers have identified situations in which interviewees express competing personal discourses on an issue.

Qualitative interviews are commonly undertaken with a purposive sample (or a judgmental sample), one that is selected based on the knowledge of a population and the purpose of the study. It can be very useful for situations in which a targeted sample exhibiting certain characteristics needs to be reached quickly and where sampling for proportionality is not the main concern. Purposive sampling is not confined to life course research but is used to select participants to explore issues such as personal and household biographies. It often forms part of a mixed methods approach. An example of this methodology is a study of the location and mobility decisions of heterosexual couples by Stockdale, McLeod, and Philip (2013) on midlife migration to a rural area in Ireland. They employed multiple stages and methods, beginning with a postal survey of a targeted set of households of recent arrivals including an item on willingness to participate in a follow-up interview. From 230 responses, they created a purposive sample of 13 households composed of couples in the 50–65-year age group who represented a range of mobility histories and backgrounds. In-depth interviews with these couples, in which they participated together, yielded diverse narratives of decision-making and assessments of experiences reflecting each
partner’s personal history in relation to the decision to move, their choice of location, and subsequent impressions of this community. Such research is not designed to seek a factual record of moves but to explore the interviewees’ interests, commitments, and “constructed subjectivity.” Interviews may be implemented in two stages in which both partners are first interviewed together and then separately, in the process revealing contrasting versions of their story.

As geographers have engaged with aspects of the subjectivity of individuals across the life stages, and in more depth than using the methods that include larger samples, some have engaged with phenomenology to explore deeper personal meanings. Though not widely practiced in geographic research, this approach is important to acknowledge for its capacities to reveal insights, to interpret and challenge preconceived notions of human experience over the life course. A key example bringing to light structures of consciousness and the meanings of places is the pioneering study of Graham Rowles (1978), who worked for a sustained period of time with five older men and women in a single neighborhood. The method seeks to reveal the first-person point of view, the significance of objects and events, of the self and others, and of the flow of time in and over life. As an approach this project required suspending the researcher’s potential judgments and stereotypes, following the subjects’ conversational leads, demonstrating sensitivity to their capabilities and preferences, considering daily and longer time frames, and engaging with their varied and multiple spaces.

Related, though not identified as phenomenological, are methods in which geographers reflect on their personal experiences of the life course and those in their own families or of people with whom they have close associations. An excellent example is Ann Varley’s (2008) blending of her developing responses and those of her father as he experienced dementia, in which she focused on meanings of home and of institutional spaces. Within that context she introduced ethical issues in research and writing about the personal. Other geographers, particularly those engaging with feminist work, have also used personal reflections to explore generational changes in mobility, work, and familial lives both within their own settings and cross-culturally.

**Participatory (action) research**

A growing number of studies engage with participatory (action) research, which refers to a variety of research practices that involve collaborative research, education, and action that is oriented toward social change. This method takes lived experience as a starting point, with approaches that aim to reduce the hierarchies between researchers, facilitators, and participants, and to generate accurate and reliable data using ethical and inclusive approaches. It has been used in both the Global North and South, and with diverse age groups, especially children and young people. It often employs contemporary technologies, such as community radio and TV, popular theater, public or community art, small format video, disposable cameras, Internet, information and communication technologies (ICTs), and auto-photography. Video is increasingly used to raise awareness of social difference or inequality for public education and advocacy. Although children may participate in collaborative and creative ways, they are not always involved in designing the research, choosing what methods to use, or in the analysis or dissemination that might result from it, such that there is participation in research rather than participatory research per se. While this work is valuable, it may not result...
LIFE COURSE APPROACH

in longer-term or deeper changes that may be needed if children’s power and influence are to be increased in decisions that ultimately affect their lives.

Mixed methods

For a number of years life course researchers have combined the collection and analysis of quantitative and qualitative data, bringing together census data or large-scale surveys with qualitative data derived from surveys conducted by the research team and sometimes by multidisciplinary teams. The central premise of combining methods in a single study or series of studies is that together they provide a better understanding of research problems than either approach alone. One such example is a study of the location and mobility decisions of heterosexual dual-career households in the United Kingdom which used a methodology with three main elements. Hardill, Green, and Dudleston (1997) first assembled and analyzed individual and household data and information from secondary sources, including the 1991 UK Census of Population and Labour Force Survey to provide a quantitative background for more detailed work in the Greater Nottingham area. They worked through contacts with five employers representing different economic sectors who helped the team identify approximately 140 dual-career households, who then responded to a semistructured self-completion questionnaire survey. The three-part questionnaire survey included sections on each partner’s career, including jobs held and places of residence over a specific time period, educational attainment, and “relationship career” (general household information). In most cases the female partner completed the section on “relationship career.” A subset (purposive sample) of 30 households was selected for in-depth interviews; both partners were interviewed together and then separately to gain insights into household decision-making relating to career prioritization. Another example using mixed methods is a Dutch study of the changing relationship between civic engagement and paid work undertaken by women in rural parts of the Netherlands between 1993 and 2007. Led by Joos Droogleever Fortuijn and Frans Thissen (2013), it combined survey data with semistructured interviews with women aged 18–50 years collected in two time periods, 1993 and 2007. The survey was conducted with a random sample of residents, while follow-up in-depth qualitative interviews were carried out with a representative sample of women, chosen in terms of life course stage and residential context. By comparing two time periods and two sources of data, they were able to discover how the labor market participation of women increased substantially while civic engagement decreased.

The burgeoning research with children and youth has perhaps been the most experimental in mixing methods. Groups of young people are engaged, for example, with visual methods such as creating videos, photographing their environments and/or making drawings, and in activities such as local walks or games. These may then be combined with participant observation by the researcher and semistructured interviews as illustrated by Nancy Worth’s (2011) study of transitions to adulthood that combined narrative interviews and audio diaries with life mapping as a participatory tool to capture “fateful moments” of transition to adulthood. Many examples of the use of these creative and experiential methods are reported in the journal Children’s Geographies.

In identifying this range of mixed methods, it is clear that geospatial technologies have not been
widely used in life course studies, though one such example by Oliver et al. (2014) compared the data from daily self-reported trip diaries kept by young people in an urban setting with data generated from GPS monitors that they wore for a week. This method revealed that only slightly over half the records were fully or partially matched, though the diaries tended to provide contextual information but not completely accurate sequencing.

**Ethics and politics of methodological approaches and the challenges they present**

Recurring themes in methodological discussion in the life course literature, especially that on children and those working with vulnerable older adults, are ethical and political aspects of research that have been elaborated in numerous articles in *Children's Geographies*. The complexities of these concerns center (though not exclusively) on ethical guidelines and clearance issues, including the ability to give informed consent; personal relations between researchers and those being studied; aspects of power relations when the research involves engagement with couples or groups, not only between a researcher and an individual subject. Protocols in the researcher's home institution, for example, may not mesh with those in a host community, especially where research occurs in a foreign cultural setting, and may involve multiple layers of review. Beyond such official requirements, however, are ethical and political concerns that arise when members of a group have differential power. These may involve what are parental rights in relation between adult researchers and young subjects, how an adult researcher establishes rapport in engaging with youth or children, or how gender and personal relations impact in group interviews or collaborative creative projects. The issues are further complicated if the research is being conducted in different cultural settings, for example in the Global South, by Northern researchers, where language differences and cultural expectations (for example, in relation to local hierarchies or questions of payment for participation) may arise.

Other examples that have been raised are whether and how “insiders” or “outsiders” should be engaged in interpretation or presentations. This is especially the case when the methods include participatory action in projects that aim not only to generate academic research but to bring about social change that crosses boundaries between subjects and practitioners in social or political community agencies. Joint interviews can reveal shared realities, and household dynamics, while separate interviews offer participants greater freedom to express individual views by allowing them privacy, but they can disrupt collective memory or understanding of events.

Finally there is a common assumption in the literature that “families” are composed of heterosexual couples, of a married couple, a husband and wife (possibly with children). But family arrangements today are also changing – they are diverse, fluid and unresolved, with a broad range of gender and kinship relations in the “postmodern family.” There is now a greater choice of lifestyle: to live alone, with a partner (of the same or other sex) or with other individuals; to stay single or marry; to remain in or terminate relationships and subsequently divorce/marry/cohabit; to forgo/postpone childbearing or to have children within/outside marriage or other consensual unions. In settings where diseases such as HIV/AIDS have disrupted families, researchers are finding particular challenges, while those working with older adults find that adults often live alone following bereavement.
LIFE COURSE APPROACH

Though greater choice may exist, living together remains a conjugal norm, and the heterosexual household, the most common form in many life courses.

SEE ALSO: Mixed-method approaches; World-systems theory

References


The concept of the life course was developed by researchers in the 1970s to counter the dominant tradition of social sciences in the 1950s and 1960s that had promoted the use of social surveys to generate grand theories. Against this abstraction of the social, the development of life course approaches sought to collate individual life histories to account for continuity and change in human lives and how these trajectories were shaped by interpersonal, structural, and historical forces. For example, Glen Elder’s (1974) major study, *Children of the Great Depression*, developed a historical, social, and psychological account of the relationship between childhood experiences and adult outcomes and behaviors. The development of life course theory has been characterized by interdisciplinary approaches, and the mainstreaming of life course in the social sciences has been codependent on the development of longitudinal data collection and analysis. A paradigmatic principle of the life course is that it is not prescriptive; it does not provide a blueprint against which individual lives are expected to develop, as in the case of other temporal configurations such as the life cycle. Instead, life course assumes a biographical approach to study connections between individual transitions and trajectories, to examine the sequencing and synchronicity of key events. Metaphorically, the life course can be compared to the unpredictable flow of a river rather than the more mechanical and predictable turning of a wheel. Yet this unpredictability cannot be reduced to individual circumstances and experiences. The foundation of life course analysis is not to study individual life events in isolation but to account for how individual lives are intertwined with social institutions through coordinated sequences of events, such as leaving home, completing education, and starting work. Life course analysis considers the collective significance of individual biographies and how these progress through institutional and social contexts.

**Geographical approaches to life course**

The founding principles of life course analysis were essentially geographical, through considering the significance of place in shaping formative experiences and transitions. Yet this does not imply that life course has been seamlessly adopted into geographical analysis. While more mainstream life course research has focused on analysis of individual data and appropriate quantitative methodologies, geographical analysis has foregrounded relationality and the contingency of individual biographies to more directly consider the spatial context of life course events. For example, the development of life course approaches in migration research has foregrounded the significance of the relationality of mobility, against more individualistic human capital frameworks. An important development in studies of “family migration” considers the tensions between individual trajectories within families and the potential for conflict between individual and collective mobilities, through exploring phenomena such as the tied mover/stayer (e.g., “trailing” spouses or “left-
behind” children) and how individual mobilities are moderated through relationships with others. Tom Cooke’s (2008) claim that all migration is family migration alerts us to the possibility that individual mobility is framed with reference to others, either through moving away, with or from or to be with others, or by corresponding immobility. Mobility practices can also anticipate life course events: for example, moving house prior to having children, or moving to be closer to relatives in older age in anticipation of frailty. Yet the significance of relationality cannot be reduced to interpersonal forces; mobility decisions are not just influenced by immediate family members and significant others, but are also given meaning in particular social-cultural contexts. For example, Ní Laoire’s (2000) biographical study of rural Irish youth migration considers the cultural construction of mobility biographies to reveal the contradictions of individual mobilities that are negotiated with reference to others and differential situated meanings.

Age

One genre of geographical research that has developed through life course analysis is the study of age and generation. While age is clearly a key dimension of life course, from a life course perspective it is not theorized as a linear development demarcated by concrete “stages.” Life course approaches resist the deterministic undertone of the life cycle; individuals do not simply move from one “age stage” to another. This resistance of a monolithic repetition of stage and time has been particularly apposite in geographical studies of childhood. The development of children’s geographies has been influenced by the social studies of childhood approach, which considers the social and cultural construction of childhood. In this approach children are not passive objects to whom things are done, but agents constructing their own identities and capable of resisting adultist interpretations. The revelation that the study of children should be carried out with, rather than on, children has stimulated a rich and varied empirical legacy, with appropriate participatory methodologies and analyses. This approach has also raised pertinent questions about the relevance of childhood itself and the extent to which it is necessary to dismantle rigid age demarcations. If life course is taken as an emphasis on flow, through which all experiences are relevant and the threads of different encounters can stretch over different spatial and temporal settings, then does it make sense to demarcate specific ages? In other words, do we need subdisciplines that are concerned with particular ages? Yet the demarcation of life course by age has a historical and cultural legacy: childhood is more than a category – it is also a social and political entity. A reasonable response to this tension between stage and flow is to consider the construction of childhood at particular times and places. Other researchers have argued for the importance of intergenerational relationships, and suggest that geographers are particularly well suited to analyze the complexity of intergenerationality, both familial and extra-familial, and how this complexity creates spatial networks of connectivity.

Transition

Life course is not a linear construct. While the notion of the trajectory is a key constituent, and this might suggest that the life course develops in a unified way over time, trajectories do not need to be straight lines. The development of the life course is not simply a matter of progression, but rather of interplay between continuity and discontinuity, pauses, fresh starts, and reversal
LIFE COURSE

as well as advancement. Thus, while life course does not assume linearity, it is demarcated by significant events that are associated with an acceleration of change. Much life course research is therefore concerned with demarcating the significance of these transition events. One main area of interest is the transition to adulthood. While denoted as a unified event, the use of the definitive article is misleading, as the study of the transition to adulthood is concerned with varied practices and developmental changes. This transition is more properly understood as a series of interconnected events, as young people move between different situations relating to education, employment, housing, and relationships. Thus the focus on transitions alerts us to changes in everyday practices and their deeper personal and social significance. Some of these changes are contemporaneous: for example, leaving home to go to university, take up a job, or move in with a partner. Yet many of these transitions are characterized by less celebrated events that are stretched out over time, and moreover key events, such as leaving home, may be reversed. For geographers, transitions also imply spatial practices; movement is not just in time but also spatial. However, space is not just the backdrop for these events, but is created and transformed through these transitions. For example, researchers have considered the meanings of home in leaving-home transitions. Fortier (2003) describes how, during this transition, the family home is always in construction, through the active emotional labor of all those involved and in individuals’ imaginations and memories.

Summary

There is considerable scope for life course approaches in geographical research, as its original conception incorporated the significance of events in time and place. Yet it could be argued that the potential of life course has not been realized, and, while many geographers might claim a life course approach, there have been fewer attempts to engage with what this signifies for geography. In particular, the spatial significance of life course can consider how spatial practices anticipate key transitions and the continuous development of spatial relations through life course events.

SEE ALSO: Aging; Children and youth; Migration: internal

References


Further reading

LIFE COURSE


Line simplification

Bin Jiang
University of Gävle, Sweden

Geographic features such as mountains, rivers, and city boundaries are fractal, and so are the cartographic curves that represent the geographic features. However, the fact that geographic features are fractal has not been well received in the geography literature, due to the limitation of fractal dimension. The definition of fractal dimension requires that change in scale ($r$) and change in details ($N$) meet a power relationship (Mandelbrot 1967, 1982). This definition is so strict that it would exclude many geographic features from being fractal. Recently, Jiang and Yin (2014) proposed a relaxed definition of fractal: a geographic feature is fractal if, and only if, the scaling pattern of far more small things than large ones occurs multiple times. The present entry attempts to clarify the recurring scaling pattern, or fractal geometry in general, and how it can be used to guide line simplification.

Line simplification removes trivial points or, equivalently, keeps vital points, without destroying the line’s essential shape or overall characteristic structure. Conventionally, the task of line simplification is conducted manually by trained cartographers while producing small-scale (or fine-scale) maps from large-scale (or coarse-scale) ones. Computer algorithms have been developed to automate the simplification process, but the most difficult and challenging issue is how to effectively measure the essential shape or the characteristic structure. In this regard, efforts in the past have been very much guided by Euclidean geometry, using measures such as angularity, distance, and ratio to characterize cartographic curves. However, these efforts have proved to be less effective. This is because geographic features are irregular and rough, while Euclidean geometry is used to describe regular and smooth shapes.

Euclidean versus fractal geometry

Euclidean geometry is widely known; it appears in high school mathematics and has a history of more than 2000 years. It can be used to measure many things, including cartographic curves in terms of angles, lengths, and areas, or sizes in general. Fractal geometry arrived much later with a history of three decades (Mandelbrot 1982). Despite its relatively short history, fractal geometry has found many applications in disciplines such as physics, biology, economics, and geography. Fractal geometry concerns or examines how things of different sizes form a scaling hierarchy; that is, whether there are far more small things than large ones. To put the two geometries in perspective, if one measures the height of a tree, he or she is using Euclidean geometry; if one measures not only the height of a tree, but also the length of all its branches, and recognizes far more short branches than long ones, this is fractal geometry.

Euclidean geometry is used for regular and smooth shapes, while fractal geometry is used for irregularity and roughness. The difference between Euclidean and fractal geometry can be further seen from Koch curves (Figure 1),
one of the first fractals, invented by the Swedish mathematician Helge von Koch (1870–1924). In the literature, Koch curve refers to the single curve when iteration goes to infinity, and therefore it has an infinite length. This infinite length is closely related to the so-called conundrum of length, that is, the length of a cartographic curve increases as the measuring scale decreases (Richardson 1961). Importantly, the relationship between the measuring scale \( r \) and the length \( L \) can be expressed by a power function, for example, \( L = r^{1.26} \) for the Koch curves — where the exponent 1.26 is termed the fractal dimension \( D \). This power function implies the recurring scaling pattern of far more small things than large ones or, more specifically, far more small triangles than large ones. The notion of “far more small things than large ones,” implying a nonlinear relationship, differs fundamentally from that of “more small things than large ones,” indicating a linear relationship. This nonlinear relationship does not exist in Euclidean geometry.

The plural “Koch curves” refers to curves with finite lengths at some limited iterations. Unlike the baseline of one unit, the structure of the Koch curves cannot be described simply using Euclidean geometry. A common feature of all the Koch curves is that there are far more small things than large ones. This is particularly true as iteration becomes high. For example, the Koch curve of the third iteration contains a total of 1, 4, and 16 triangles with respect to scales of 1/3, 1/9, and 1/27. The recurring scaling pattern of far more small things than large ones is captured by the ht-index (Jiang and Yin 2013) as an alternative index of fractal dimension for quantifying the complexity of geographic features. The baseline has a dimension of one, while the Koch curves all have a fractional dimension of 1.26. The Koch curves can hardly be characterized by their lengths. For example, the Koch curves have lengths of 4/3, 16/9, and 64/27, respectively, for the scales of 1/3, 1/9, and 1/27. The length information does not add anything new to the understanding of the Koch curves, and neither do the angles, for they appear to have the same 60°. Therefore, the Koch curves were referred to as pathological before fractal geometry was established.

![Figure 1](image.png)

**Figure 1** Koch curves and their basic statistic. (Note: The baseline of iteration 0 is also called the initiator, while the shape of iteration 1 is referred to as the generator.)
Figure 2  (a) Scaling hierarchy of the Koch curve; (b) the 21 triangles plotted in a rank-size distribution.

Ever since geography was established as a discipline in Ancient Greece, Euclidean geometry has been an important tool for measuring and describing the Earth and its surface. Although the Earth is not a perfect sphere, it can be approximated as a globe for the purpose of map projection. Euclidean geometry plays an important role in the transformation of the globe into two-dimensional maps. Beyond this, however, Euclidean geometry adds few insights into the understanding of geographic forms and processes. This is because Euclidean geometry can only be used to measure regular and smooth shapes, whereas geographic features appear irregular, wiggly, and rough. Unfortunately, Euclidean geometry has dominated geography for a long time and it is time to change our mindsets. The following section illustrates how line simplification, using the third iteration Koch curve as a working example, can be achieved with the guidance of fractal geometry.

Line simplification based on head/tail breaks

The third iteration Koch curve contains a total of 21 triangles of different sizes – specifically, 16, 4, and 1 with respect to sizes of $1/27$, $1/9$, and $1/3$. The 21 triangles, defined in a recursive manner, constitute a scaling hierarchy and are plotted in a rank-size distribution (Figure 2). The scaling hierarchy implies that there are far more small triangles than large ones, that is, 16 small and five large. The scaling recurs again for the five large triangles, that is, four small and one large. Therefore, the Koch curve has an $ht$-index of 3. Given the scaling hierarchy, line simplification can be conducted by removing small triangles or keeping large ones; in other words, removing the 16 smallest ones (blue) first, and then the four mid-sized ones (green). The line simplification sounds very simple and straightforward for the Koch curve. How can the idea be applied to a cartographic curve?

A cartographic curve is usually far more complex than the Koch curve, but both show the same scaling pattern of far more small things than large ones. What are the “things”? And how can

Figure 3  Illustration of geometric measures for a cartographic curve (note primary measures such as angle $\alpha$, distances $d$ and $x$, and secondary measures such as ratio $x/d$, and triangle area $d \times x/2$).
they be defined for a cartographic curve? The “things” refers to certain geometric measures, including both primary and secondary. The primary measures refer to angles and distances of various kinds, while the secondary measures are derived from the primary ones (Figure 3). One example of the measures is the distance \((x)\) of the furthest point from the line, with distance \(d\), linking two ends of a curve. Jiang et al. (2013) showed that both the primary measure \((x)\) and the secondary measure \((x/d)\) exhibit a heavy-tailed distribution, implying far more small values than large ones. This is the same for the measure area \((d \times x/2)\), that is, far more small triangles than large ones. Note that the values of the measures should be understood in a recursive manner and the recursive process ends up with a series of values for each measure. Once the things (or measures) have been determined, the next question is how to differentiate small and large things.

The small and large things are differentiated simply by the arithmetic mean. The key, however, is to differentiate small and large things recursively. This is what underlies head/tail breaks as a classification scheme (Jiang 2013) and visualization tool (Jiang 2015) for data with a heavy-tailed distribution. Given a series of values for measure \(x\) with a heavy-tailed distribution, head/tail breaks help to derive hierarchical levels of the measure. For the sake of simplicity, we employ the Koch curve to illustrate the head/tail breaking process. The arithmetic mean of the 21 triangles is:

\[
m_1 = \frac{1 \times \frac{1}{3} + 4 \times \frac{1}{9} + 16 \times \frac{1}{27}}{21} = 0.07
\]

The first mean of 0.07 split the 21 triangles into two unbalanced parts: five above the mean (about 24%, a minority called the “head”) and 16 below the mean (about 76%, a majority called the “tail”). The average of the five triangles above the first mean is:

\[
m_2 = \frac{1 \times \frac{1}{3} + 4 \times \frac{1}{9} + 5}{5} = 0.16
\]

The second mean of 0.16 further split the five triangles into two unbalanced parts: one above the second mean (about 20%, the “head”) and four below the mean (about 80%, the “tail”). The head/tail breaking process eventually leads to three hierarchical levels that we have seen already – that is, the ht-index of three.

Once the hierarchical levels have been derived, line simplification simply involves keeping those with large values or removing those with small values. Those with large values are characteristic points, which help retain the essential shape of a cartographic curve. The unique feature of the line simplification principle is that the levels of detail are automatically or naturally determined according to the scaling hierarchy of a cartographic curve. This immediately raises the question of how the levels of detail can be associated with different map scales. Another question is how to determine the levels of detail when multiple geographic features are involved in a single map or database. These two issues warrant further research in the future.

Further discussion

A very large body of literature on line simplification has been created over the past few decades (Li 2007). Previous studies on line simplification are strongly guided, and dominated, by Euclidean geometry in developing measures with which to characterize the essential shape of a cartographic curve. For example, over 30 measures have been developed for assessing the performance of line simplification algorithms (McMaster 1986). Fractal geometry
did stimulate many applications and research on line simplification, but without a great deal of progress (Buttenfield 1985; Muller 1986). The research was overly constrained by the notion of fractal dimension and failed to recognize that the essence of fractal geometry is the recurring scaling pattern of far more small things than large ones. In the spirit of fractal geometry, or fractal thinking in general, this section provides further discussion on two algorithms: the Douglas and Visvalingam algorithms (Douglas and Peucker 1973; Visvalingam and Whyatt 1993), which bear some fractal thinking.

The Douglas algorithm is probably the most widely used line simplification algorithm, since it can effectively detect critical points that reflect the essential shape of a cartographic curve. It starts with the straight line that links two ends of a curve and checks which point in between the two ends is farthest away from this line. If the farthest point is closer (measured by \( x \)) than a given threshold, it removes all in-between points; otherwise, the curve is split into two pieces around the farthest point. The above process continues iteratively for each piece until all of the farthest points are within the preset threshold. This is what is called the top-down approach, that is, starting from a curve as a whole and splitting and continuing splitting until all pieces are within a given threshold. The Visvalingam algorithm works the other way around and is bottom-up in nature. It starts with the individual point and checks which point has the smallest triangle with the point’s immediate neighboring points; that point is eliminated first. This process continues progressively until all points with the small triangles (smaller than a given threshold) have been eliminated. This idea of iterative elimination of the smallest triangles is very similar to the Bendsimplify function (Wang and Muller 1998) that is embedded with ArcGIS.

As elaborated, distance \( x \) and triangle area \((d \cdot x/2)\) in the two algorithms both exhibit a heavy-tailed distribution, which implies far more small \( x \) than large ones, or far more small effective areas than large ones. This scaling pattern provides new insights into line simplification. A cartographic curve is simplifiable because of its recurring scaling pattern. In this regard, the cartographic curves show no difference from the Koch curves. The only difference between the cartographic and the Koch curves is that the recurring scaling pattern is defined statistically for the former, but strictly and exactly for the latter. For either the cartographic curves or the Koch curves, head/tail breaks provide an effective means to derive an inherent scaling hierarchy for line simplification. Essentially, the Douglas algorithm is used to progressively select vital and critical points, while the Visvalingam algorithm is used to progressively eliminate trivial and redundant points. With the help of head/tail breaks, selection and elimination are applied respectively to the heads and the tails, again in the progressive and iterative manner. Both algorithms keep a cartographic curve as a whole in mind during the progressive processes of selection and elimination. It is in this sense that both algorithms share a certain spirit of fractal thinking, although both algorithms were developed without guidance of fractal geometry.

While keeping critical points, or equivalently removing redundant points, it is inevitable that convoluted curves will create crossings or self-intersections. A simple way to avoid crossings is to add a few more trivial or redundant points. However, some researchers tend to use the crossings to criticize a line simplification algorithm. The crossings are not a problem and should not be used as an indicator for line simplification performance. For an especially flourishing tree, crossings of branches are often...
LINE SIMPLIFICATION

seen in reality. For the Koch curve, if the scaling ratio is raised to a high value, it will also create self-intersections. Guided by the fractal thinking, the best way to assess the performance of a line simplification algorithm is to determine whether it retains the scaling pattern of far more small things than large ones.

The geographer Perkal (1966) studied the conundrum of length, that is, the issue of undefined curve length, for the purpose of measuring things on maps. However, it was the mathematician Mandelbrot (1967, 1982) who seized the opportunity and developed the new geometry of fractals. Despite many research efforts to apply fractal geometry into line simplification, our mindset still tends to be Euclidean rather than fractal. We see geometric details such as angles, lengths, and shapes, but miss the whole – the recurring scaling pattern of far more small things than large ones; in other words, we are used to seeing things linearly and individually rather than nonlinearly and holistically. Therefore, current line simplification, or map generalization in general, is very much (mis)guided by Euclidean geometry. It is time to change the mindset.

SEE ALSO: Cartographic design; Distance; Map generalization; Scale; Shape

References


Further reading


Until quite recently it was possible to recount the history of literary geography as a subfield of human geography in a relatively straightforward manner, with geographical work with literary texts following the shifts and turns of research trends in human geography quite closely. For example, early use of descriptive literary passages as data for regional geography was modified and complicated by a growing humanistic emphasis on the subjective experience of place. That shift was in turn challenged by more radical and critical approaches; and then, under the influence of the cultural turn, literary geography became more explicitly influenced by literary and cultural theory.

In the last few years, however, rapid development in three areas has complicated the definition of literary geography. First, work at the intersection of literary studies and geography has become more actively interdisciplinary, and as a result literary geography today appears to be moving into an era of greater cross-disciplinary collaboration. Second, and partly as a result of this first development, it has become more difficult to identify what should be included within the interdisciplinary configuration of “literary geography.” Third, bibliographic efforts to record the history and present range of work combining literary studies with geography have made a wider range of relevant material visible and accessible, and this has complicated the understanding of the history, spatial distribution, and multidisciplinary range of the field. While these developments make coherent overview and collaborative consensus more difficult to sustain they reflect the currently energetic state of work in literary geography and cognate areas.

The origins and early stages of geography’s literary geography have been well documented, with works from the mid-twentieth century on the role of the imagination in geography and on regional literary geographies such as Hardy’s Wessex and Faulkner’s Yoknapatawpha County most commonly cited as marking the establishment of the field. In these early stages of literary geography, emphasis was commonly placed on the depiction of setting in fiction and the relations between fictional and lived regions. Related work in literary studies that would in time become significant for literary geography from the same era includes M.M. Bakhtin’s argument for the connectedness of literary time and space in his theory of the chronotope and Joseph Frank’s introduction of the concept of spatial form. Where Bakhtin was interested in how distinctive literary space-time configurations shaped and characterized particular genres, such as nineteenth century French realism, Frank drew attention to the nonlinear structure of works in which the narrative is distributed spatially throughout the text.

Although the first comprehensive review articles would not appear until the 1990s, as a sub-field within human geography literary geography was already gaining coherence in the 1970s, partly as a result of the emergence of humanistic
LITERARY GEOGRAPHY

geography. Within human geography, literary geography in this era was performed in the context of an increased interest in human perception of the environment and a new approach to the idea of place as location invested with cultural and personal meaning. Humanistic literary geography tended to look to literary texts, usually fiction, for universal truths, and to literary authors for particularly acute and well-written descriptions of place and landscape. Not yet a combination of geographical and literary studies, literary geography turned to literary texts for geographical evidence while maintaining a disciplinary distance from the “complementary field” of literary criticism. Meanwhile in literary studies “landscape in literature” or “literary landscapes” remained an established theme.

In the same way that humanistic literary geography had emerged in reaction to positivist and quantitative geographies, over time it became subject to critique itself from various lines of work associated with more critical and radical geographies, as attempts were made to add a more socially critical and political dimension to the field. The 1970s and 1980s also saw the beginnings of work on the inclusion of maps in literary works and on the use of fiction in the geography classroom. In this period, even as debates continued over the extent to which literary texts could be understood to constitute geographical data, early work began to appear in various areas that would later become major themes for literary geography: literary tourism, for example, science fiction, children’s geographies, regionalism, empire, and postcolonialism.

Literary geography from 1994 to 2005

The appearance of Brosseau’s 1994 review essay in Progress in Human Geography marks an important moment in the development of literary geography for two reasons. First, Brosseau argued for a literary geography methodology built around close reading techniques characteristic of literary studies, criticizing the redirecting of attention toward the ways in which literary texts generate “particular modes of readability that produce a particular type of geography.” Second, because he was familiar with both French and English language literary geography and published in both languages Brosseau initiated early moves toward a greater internationalization of the field. If the period before 1994 could be characterized as the era in which literary geography was primarily focused on description and as a result, typically, realist description and recognizable settings, Brosseau’s work added a focus on the space-generating aspects of narrative as a second major stream in geographical literary geography.

Another new line of work in geographical literary geography beginning to appear in the 1990s emphasized the social, cultural, and political contexts of literary creation and reception. This new emphasis on reading and consumption developed from Sharp’s reminder (2000) that not all readers are trained in literary criticism and that in literary geography the reader-response of nonprofessional readers cannot be ignored. Meanwhile work on literary tourism continued to develop, as did work on the fictional representation of historical places, children’s literature, and the mapping of literary settings. Science fiction first emerged as a significant genre for literary geography in the late 1990s, with the collection Lost in Space (Kitchin and Kneale 2002) marking an important moment in literary geography’s expansion of range, as an interest in the textual production of imagined worlds was added to the conventional emphasis on realistic description and representation. Major themes in human geography in this era which influenced
work in literary geography included the post-colonial, feminist geographies, and geographies of identity and positionality.

Beyond the discipline of geography, the late 1990s and early 2000s saw an increase in work that combined literary studies with geography and spatial theory in established interdisciplinary fields such as American studies. While this work was not formally presented as “literary geography” its themes and methodologies were compatible with the geographical work with literary texts that was emerging in response to the new cultural geography. In contrast, Franco Moretti’s *Atlas of the European Novel* (1998) brought quite a different idea of a literary geography, disconnected from any existing tradition within human geography, to a wide audience. According to Moretti, a specialist in comparative literature, literary geography involved the identification of suitable textual features (such as settings), the collection of data, the cartographic reformulation of that data, and finally the use of the resulting maps to generate new insights into literary history: “this is what literary geography is all about,” he argued. With no reference to work by geographers, or to existing definitions of literary geography, Moretti launched a separate tradition with the *Atlas* that only later began to reconnect with academic geography, for example, in the literary cartography project at ETH Zurich, *A Literary Atlas of Europe*. With the development of GIS and the push to merge qualitative and quantitative methods in literary geography, this kind of work is now an important element in the field. In terms of interdisciplinary connections, this line of work in literary geography – which tends to focus on the mapping of settings – docks quite comfortably at present with work in narratology, which generally assumes a hierarchical container-space metageography and which, as a result, is difficult to integrate with lines of work in literary geography more interested in rethinking the nature of space and in relational and network geographies.

The so-called spatial turn in cultural studies and the cultural turn in geography of the late 1990s and early 2000s generated an increase in the interdisciplinary exchange of ideas and theories. Blair’s “Cultural Geography and the Place of the Literary” (1998), for example, which appeared in *American Literary History*, concentrated on the overlap between the new cultural geography and work in American studies on literary texts, history, and historiography. This line in American literary studies later produced work on the literary production of scale, on fiction and maps, and on literary metageographies. Despite referring to the work of Lefebvre, Harvey, Massey, Soja, Smith, and other spatial theorists, however, Blair’s article typically made no mention of geographical work in literary geography. Work on literary texts, geographies, and spatial theory – even as similar themes and methodologies were being taken up in various disciplines – in this way remained quite fragmented.

**Literary geography after 2005**

The 2004 conference of the Royal Geographical Society/Institute of British geographers included a significantly interdisciplinary session on “Textual Spaces, Spatial Texts” that brought together geographers and literary critics, with papers being published in *New Formations* the following year. While the conference session and subsequent publication represented a new move toward interdisciplinarity in literary geography, the contributions also highlighted the way in which the invention of new variants of literary geography also sustained disciplinary boundaries. The literary critic Andrew Thacker’s
article (2005), for example, calls for the establishment of a “critical literary geography,” based on the practice of “reading and interpreting literary texts by reference to geographical concepts.” Thacker’s question – “What would such a ‘critical literary geography’ look like?” – suggests the extent to which the literary-critical and the geographical strands in literary geography at this point remained disconnected.

To some extent, even today, work at the more literary end of the spectrum tends to focus on the ways in which geographical methods and spatial theory can illuminate critical readings, while work at the more geographical end tends to focus either on the analysis of map data or on close reading employed in the discussion of issues such as the literary production of space and geographies of writing and reading. Remaining differences in purpose and emphasis notwithstanding, however, there has been a notable increase in the interdisciplinarity of literary geography in the past decade. Collections of essays and journals are today much more likely to include work by both literary critics and geographers, and there are an increasing number of collaborative projects underway involving combinations of disciplines. The new open-access online journal Literary Geographies, for example, was jointly established in 2013 by geographers and literary critics collaborating in an attempt to reduce the disciplinary sorting and distancing effects associated with conventionally distinct literary and geographical journals.

Sorting work on literature and geography

One of the major questions in generating an overview of current work in literary geography is whether or not to limit the scope of the review strictly to work which self-identifies as literary geography, or to consider the whole range of work dealing with literary writing and reading and questions of space, place, and geography. While the distinction between critical literary geography and the more traditionally geographical literary geography is not so clear that the two cannot be understood within the same framework, and while work at the intersection of literary studies and geography within American studies (for example) is often clearly compatible with work produced within human geography, other recently emerging fields such as geopoetics, green cultural studies, ecopoetics, and ecocriticism are less susceptible to inclusion within the broad field of literary geography. This can be ascribed in part to the fact that approaches such as ecocriticism tend to embody specific sociopolitical goals in regard to issues such as globalization, ecology, and human–environment relations.

The relatively new field of geocriticism is commonly presented as an expanded form of the French-language concept of géocritique, a geocentered approach to literary criticism. In its English-language manifestation, geocriticism at present in practice bears many similarities with critical literary geography and in theory with cultural geography generally, but its relationship with literary geography is somewhat difficult to determine because of the way in which it has been presented as a new initiative in studies of literary spatiality and indeed spatiality more generally.

The field of literary cartography is more easily grasped in relation to literary geography, its history and current practices, and the terms are sometimes used interchangeably. While some studies dealing with literary texts and maps follow the Moretti “distant reading” path, relying on the accumulation of large data sets and (increasingly) GIS methods to generate cartographic representations of literary aspects such as fictional settings or author geographies, others take an approach more in line with the kind
of “close reading” typical of literary criticism, focusing in greater detail on single authors or texts. There is also a growing interest in the relationship of mapping to narrative: the use of maps within literary works, the mapping of fictional settings, and (particularly in the context of Web 2.0) the reading of maps as stories.

Bibliographies

One of the most significant developments in recent years in literary geography has been the establishment in 2012 of an open-access online bibliography providing a regularly updated database of work in the field. As of January 2014, the bibliography listed more than 1700 items organized into chronological and thematic lists with a separate bibliography for unpublished MA and PhD theses. Themed lists include literary geography and mapping, literary geography and tourism, poetry, science fiction, graphic fiction, and detective fiction, and there is also a page listing research project websites. Thacker’s (2005) and Hones’s (2008) review essays were able to specify clear limits to their reach, dealing, for example, almost exclusively with critical literary geography or with work on fiction published in geography journals. The pushing back of the horizons of literary geography and the increased visibility of similar work in other disciplines and on a wider variety of genres enabled by the literary geographies website means that this kind of strategic limitation is today much less likely.

The rapid emergence of online search engines and depositories in addition to the digitalization of journal archives has meant that the literary geographies bibliography website has been able to expand the horizons of literary geography by compiling citations for relevant but scattered and in many cases “lost” work, such as past unpublished doctoral dissertations or work published in collections and journals outside the mainstream of human geography. This bibliographic recuperation has rendered visible the very broad range of work in the field, bringing into sight work on such varied textual materials as English-language poetry from Singapore and the Niger Delta and from journals as diverse as Australian Literary Studies and the African Research Review. Taken together with the geographical range of the work included, the global spread of the page view metrics for the website indicate the extent to which even English-language literary geography today extends far beyond its traditional foundations in British cultural geography and literary studies and American studies and literary history.

New directions

In the period before 1994, literary geography was primarily concerned with questions of description and representation; the following decade saw the addition of a new focus on narrative and its role in the literary production of place and space. The shift in thinking that has moved geography from a view of space as a set of external coordinates to a view of space as something more unstable – the product of, not the container of, interrelations and networks – has led not only to a widening of the gap separating literary geography generally from narratological work on literary space but also to a closer relationship between literary geography and the work of textual theorists dealing with the concept of intertextuality, a combination which has introduced into literary geography the idea of literary space as a network of intertextual relations.

With continuing work on description, representation, and narrative in literary geography now including strong initiatives in literary cartography, the addition of a third major focus
means that literary geography today is also concerned with various forms of writing and reading practice. This new emphasis, in part a response to the emergence of nonrepresentational theory, has led to increased interest in studies of the physical and social geographies of inspiration, creativity, and production, of author–text–reader interaction and historical geographies of reader reception (Saunders 2010). The idea that the literary text itself can be understood as a spatial or geographical event happening in the interaction of multiple agents (including authors, editors, publishers, and readers) is an emerging line of research. Progress is also being made in work on the historical geography of the book, on author–editor interactions, and on textual materialities.

Overall, in recent years literary geography and its cognate fields have been characterized by a notable increase in productivity and diversity. The fact that the list for theses and dissertations at the literary geographies bibliography website has 53 citations for the period 1980–2008 but 57 citations for the subsequent four years, 2009–2012, indicates the growing interest in this area. Once focused almost exclusively on nineteenth century realist fiction, literary geography in the last decade has diversified extensively now including work on a wide range of other fictional periods and styles: the modernist and the postmodern novel, for example, short stories, poetry, and drama. The 2013 collection Poetry & Geography: Space and Place in Postwar Poetry (Alexander and Cooper 2013) provides a good example of the way in which the post-2000 line of critical literary geography, grounded in literary criticism, and the tradition of geographical literary geography, which can be traced back to the mid-twentieth century, are converging. With the launch of the new journal Literary Geographies in 2013, it seems likely that literary geography will continue to become more varied, collaborative, and interdisciplinary in the years ahead.

SEE ALSO: Cultural studies; Cultural turn; Imaginative geographies; New cultural geography; Representation; Text and intertextuality

References

Lithologic discontinuities in soils

Steven W. Ahr
*University of the Incarnate Word, USA*

Lee C. Nordt
*Baylor University, USA*

Randall J. Schaetzl
*Michigan State University, USA*

Accurate recording and interpretation of vertical changes within a soil profile are important for pedologists, geomorphologists, stratigraphers, hydrologists, and archaeologists. Describing soils is a fairly standardized process that follows a systematic cataloging of observable properties and their changes with depth. Many soil properties that vary with depth often do so abruptly, exhibiting stark contrasts with the upper or lower horizons. It is not always clear, however, even to those who routinely investigate soils, whether such vertical changes originated from sedimentologic/geologic layering (i.e., two or more parent materials), or from pedogenic processes. Where geologic layering has resulted in the vertical change in the soil profile, the point at which that change occurs is referred to as a lithologic discontinuity (LD); it separates the two layers (Figure 1). Each of the layers is assumed to differ substantially from the other layer in terms of particle size distribution (PSD) or mineralogy, reflecting differences in lithology. Because the sediment above an LD is always younger than the sediment below, LDs also denote age differences in the materials (Schaetzl 1998; Soil Survey Staff 2010). If no discontinuities occur within the soil material, pedogenesis is assumed to have proceeded within a single, uniform parent material.

LDs that separate different sedimentary layers, or strata, form when there is a shift from one depositional system to another (e.g., eolian to fluvial), or as changes take place in an otherwise similar sedimentary system. The former can be related to changes in parent material source, whereas the latter may reflect waning or advancing energies associated within a single depositional system. Loess overlying glacial till, alluvium overlying residuum, and colluvium overlying alluvium are examples of stratigraphic successions containing LDs that originate from different depositional agents and sediment sources. Changes in depositional energy within the same sedimentologic regime (e.g., fluvial) can also result in the formation of LDs, such as a significant shift in particle sizes due to changes in stream competence. For example, medium or fine sand overlying mostly coarse or very coarse sand can be assumed to be two different materials, due to differences in depositional energies, although they are of the same mineralogy (Soil Survey Staff 2010). Within soil-stratigraphic successions, LDs can also point to a chronological unconformity caused by a hiatus in sedimentation (beyond diastems). Such chronological unconformities within alluvial soil sequences are often marked by paleosols that form during episodes of landscape stability. The stratification of buried soils in an alluvial sequence can also be marked by LDs if the geologic materials comprising each alluvial unit are lithologically dissimilar, or if particle sizes between layers are strongly contrasting (Soil Survey Staff 2010).
LITHOLOGIC DISCONTINUITIES IN SOILS

Detection of lithologic discontinuities

Although easy to conceptualize, detection of LDs in the field is not always straightforward. Traditional approaches rely on methodologies designed to identify significant vertical changes in profile properties that have unquestionably resulted from geologic processes. Such approaches attempt to exclude those changes that were caused by additions, losses, transformations, and translocations that may have occurred during pedogenesis. For example, pedologists insist that LD indicators must not have resulted solely from translocation of clay during pedogenesis. The same could be said for any substance that is translocated in soils during pedogenesis. Thus, soil properties that are related to mobile soil constituents, including pH, clay content and mineralogy, organic matter and carbonates, as well as those that are affected by in situ weathering processes, should not be used for the detection of LDs. These vertical changes occur in soils due to pedogenesis, not via geologic processes. However, such properties can and often do abruptly shift at an LD. With this background in mind, we turn to the methods that have been developed to detect LDs, including morphological, textural, and mineralogical indicators.

Morphological and textural indicators

Classic morphological indicators of LDs include abrupt changes with depth that are unrelated to pedogenesis, especially those related to differences in lithology, shape, or distribution of coarse fragments. Significant horizon-by-horizon changes in the degree of weathering of soil rock fragments, rock fragment angularity, and the size and shape of resistant mineral grains observed in micromorphic thin section, can suggest different geologic origins for each layer. In areas where soils have relatively unweathered rock fragments, an LD can be indicated by an uneven distribution of rock fragments with depth. Stone lines, which commonly separate different parent materials, could suggest that a soil has developed in more than one kind of parent material, separated by an LD. Caution must be exercised, however, as stone lines are also commonly formed by pedoturbation in initially uniform materials (Johnson 2008).

In terms of textural indicators, abrupt changes in the depth distributions of sand and silt totals,
LITHOLOGIC DISCONTINUITIES IN SOILS

Mineralogical indicators

The analysis of depth trends in stable soil constituents such as Ti and Zr have long been used for detection of LDs. Both Ti and Zr are considered weathering-resistant when present in stable and insoluble mineral phases such as tourmaline and zircon, respectively. If these minerals can be isolated from the sand or silt fraction, which is normally not considered mobile in soils, the amount of Ti and Zr can be used to indicate the presence of lithologic changes. Nonuniform depth functions in Ti and Zr (and ratios between them) imply inherited layering in the soil profile and point to the presence of an LD. Alternatively, uniform values (and ratios) of these kinds of data point to parent material uniformity (Anda, Chittleborough, and Fitzpatrick 2009; Chapman and Horn 1968).

In some instances, Ti- and Zr-bearing minerals may become chemically reactive and/or mobilized, and as a result, their utility as a stable index element can become problematic. For example, Ti is normally present in ilmenite, rutile/anatase, mica, and biotite. But as these minerals weather, Ti may become incorporated into the fine clay (mobile) fraction, transported or leached, and reprecipitate lower in the soil profile (Taboada et al. 2006). In this case, Ti depth functions may reflect weathering and translocation processes, rather than an LD (Figure 2). Because Zr is found almost exclusively within the weathering-resistant mineral zircon, it tends to be ideal for identifying LDs, so long as it is present in the sand or silt fraction, and in detectable quantities. However, the use of Zr can also be problematic in some instances. For example, eolian additions of small amounts of zircon to the upper soil profile can significantly complicate LD detection. Zr has also been shown to be susceptible to redistribution as well as chemical weathering and leaching in extremely acidic or alkaline soils, and in volcanic soils.
LITHOLOGIC DISCONTINUITIES IN SOILS

Figure 2 A texture contrast soil (Grossarenic Paleustalf) from Burleson County, Texas. Clay-free textural and mineralogical data failed to reveal any LDs within this soil, that is, the textural differences have resulted from mobilization and translocation of fine clay by pedogenesis. Note that there is virtually no change in Zr with depth. In contrast, the slight shift in Ti at the top of the argillic horizon is likely attributable to weathering of primary minerals and incorporation of Ti into the mobile clay fraction, which is 81% correlated. Photo by S.W. Ahr.

Post-depositional factors affecting lithologic discontinuities

Some LDs are so subtle as to raise the question of whether they are even worth noting. In other situations, post-depositional processes such as pedoturbation have blurred the contact between the two materials, making detection (and perhaps importance) of the LD questionable (Schaetzl and Luehmann 2013) (Figure 3).

LDs within a profile are often obscured and blurred by weathering and mixing processes that disturb the original stratification between different layers of parent materials. Such horizionation and mechanical sorting processes include various types of pedoturbation, such as shrinking-and-swelling by smectitic clays, cryoturbation, and bioturbation. These processes are most pronounced in the near-surface zone, implying that LDs are more likely to be preserved at greater depths in the soil where relict bedding can persist through multiple phases of pedogenesis and pedoturbation – even within the highly decomposed and weathered soils of humid regions.

The same processes that obscure pre-existing LDs can, in some circumstances, also transform uniform, near-surface sediments into layered sediment, especially on older upland landscapes. For example, weathering in some soils transforms coarse particles in the upper profile into silt, clay, and fine clay, and moves these particles down the profile. This process often results in the formation of a texture contrast soil, in which coarse textured layers overlie finer-textured layers enriched in clay (see Figure 2). This type of contrast is commonly observed in well-drained Alfisols, as well as Ultisols (Ahr, Nordt, and Driese 2012; Koppi and Williams 1980). The contrasting textures, often separated by abrupt boundaries, can be mistakenly interpreted as geologic layering. Such vertical changes are more appropriately interpreted as pedologic discontinuities that resulted from the pedogenic segregation of mobile soil constituents. Because of these complicating factors, distinguishing between pedologic and lithologic discontinuities should be determined using multiple lines of evidence.
Figure 3  Examples of two different graphical methods that can be used to show LDs in soils. The data in (a) and (b) are from a profile composed of a 40-cm-thick, silt-rich, loess mantle over sandy glacial outwash. This soil has a mixed zone, rather than an abrupt and sharp LD. (a) Traditional depth plots of immobile fractions and ratios of immobile fractions, as well as modal particle size. (b) Continuous textural curves, taken by horizon. Note how these curves more readily show the presence of a distinct mixed zone in this soil, which is manifested mainly as loess mixed into the underlying sand, rather than sand mixed upward into the loess. (c) Similar data for a soil formed in 35 cm of loess over sandy glacial outwash, but with perhaps a more abrupt LD. Adapted from Schaetzl and Luehmann (2013).
LITHOLOGIC DISCONTINUITIES IN SOILS

**Figure 4** Historic-to-modern (1840–1960 CE) coal-rich alluvium, or “coal wash,” overlying a late Holocene alluvial surface. The contact between the two materials marks an LD of both a lithologic and a temporal nature. This soil is located along Nesquehoning Creek in eastern Pennsylvania. Photo courtesy of G. Stinchcomb and M. Stewart.

Impacts of lithologic discontinuities on pedogenesis

LDs are globally widespread, and probably more common than many soils researchers recognize. LDs have been reported in loess and sand covers over glacial tills in relic periglacial environments across Europe, in volcanic soils, in sandy Atlantic and Gulf Coastal Plain soils, in the loess-mantled landscapes in North America, in soil complexes in Australia, and in highly weathered soils of the tropics. Because they are so extensive, and because changes in lithology or mineralogy that are indicated by an LD can influence the trajectory of soil formation and impact pedogenesis, the ability to detect LDs in soils and to distinguish between different parent materials is important for pedogenic studies. For example, the presence of LDs in the subsurface, particularly in fine-over-coarse layering, can influence soil hydrology, and may also inhibit the eluviation/illuviation of mobile soil constituents. Also, identification (or misidentification) of an LD can greatly affect interpretations of soil laboratory data, particularly in terms of gains and losses of soil constituents. For example, mass balance calculations, which are used to quantify soil volume changes and open-system transport of constituents into or out of the soil during pedogenesis, depend on an assumption of parent material uniformity, or absence of LDs.

New research directions on lithologic discontinuities in soils

Although traditional approaches to investigating LDs are important, Lorz and Phillips (2006) suggested that current “top-down” genetic models fail to fully account for a dynamic regolith in soil formation. They point out the need to evaluate both pedological and geological (polygenetic) processes in soil genesis studies. This approach is warranted, particularly when dealing with dynamic soil environments such as those in alluvial or other aggradational settings, areas of dust deposition, and places where historic-era anthropogenic materials have been added to the soil column. Recently, Stinchcomb et al. (2013) integrated the concept of event stratigraphy with the pedologic investigation of lithologically distinct, nineteenth- and twentieth-century alluvial coal deposits in eastern North America (Figure 4). Such deposits are becoming increasingly identified and mapped around the world, which will have important implications for biogeochemical changes to the Earth’s surface, along with anthropogenic influences on deposition and erosion. This research underscores the need
LITHOLOGIC DISCONTINUITIES IN SOILS

investigations. Deciphering the complexity and influence of LDs is critical in soil genesis studies because the trajectory of soil formation is greatly influenced by parent material. Traditionally, identification of LDs has yielded important information about the genesis of “modern soils.” As the realm of pedology continues to expand, investigation of LDs will be instrumental in providing useful data on anthropogenic influences on local geomorphic settings, and in reconstructing ancient geologic depositional histories and soil-forming environments.

SEE ALSO: Soil mass balance; Soil taxonomy and soil classification; Soils and weathering; Soils in archaeological research; Soils in geomorphic research

References


LITHOLOGIC DISCONTINUITIES IN SOILS


Further reading


**Lithuania: Lietuvos geografin̆ draugija (LGD) (Lithuanian Geographical Society)**

Founded: 1934  
Location of headquarters: Vilnius  
Website: [www.lgd.lt](http://www.lgd.lt)  
Membership: 199 (as of January 1, 2014)  
President: Dovilė Krupickaitė  
Contact: dovile.krupickaite@gf.vu.lt, info@lgd.lt

**Description and purpose**

The Lietuvos geografin̆ draugija is a voluntary, independent, and public scientific organization that unites professional geographers and other individuals and organizations interested in the science and teaching of geography and promoting their progress.

The goal of the society is to enhance the significance of geography as a science through the coordination of the activities of its members.

One of the main objectives of the society is to publicize to the general public the geographical problems faced by the Republic of Lithuania so as to improve decision-making.

Membership has been fluctuating between 100 and 300 members over the years since inception. The members work in a variety of different fields, such as science, teaching, planning, and public administration.

**Journals or major publication series**


**Current activities or projects**

The LGS focuses on representing and promoting the interests of the geography community, promoting geography in society while defining and solving topical problems of geography, and international collaboration.

Principal current projects include the preparation of an academic dictionary of geography and the development of a national landscape management plan (in collaboration with other institutions). Since 1986 the society has focused on the education of young geographers with pupils in senior classes participating in a 3-year learning program. Annual national geographical contests attracting approximately 1000 participants from schools throughout the country have been organized by LGS since 2001.

The Lithuanian Geographical Society organizes an annual congress for the discussion and implementation of tasks and discussion of future goals.

**Brief history**

The beginning of geographical research in Lithuania can be traced to the sixteenth century when the first maps of the country appeared. The best known of these is “Moscovia”, mentioned in the *Cosmographia* of Sebastian Münster. In the seventeenth century meteorological research was started at Vilnius University and the *Compendium geographicae* was published in 1743 in Vilnius. By then many geographical subjects were taught and the first textbooks were published at Vilnius University with geography advancing significantly during the nineteenth century. Until 1917, some geographic research in Lithuanian was organized by the northwest division of the Imperial Russian Geographical Society. Geographical research...
LITHUANIA

intensified after 1922 following the restoration of Lithuania’s independence in 1918.

The LGS was established in 1934 with Professor Kazys Pakštas serving as the first and founding president. The LGS started their participation at the International Geographical Congress in 1938. Membership at the time was about 100. After the occupation of Lithuania in 1940, LGS actively functioned abroad (mainly in the USA).

The geographical society of Lithuanian (SSR) was founded in 1955. It was a subdivision of the All-Union Geographic Society of the USSR. Its scientific periodical *Geografijos metraštis* (Geographical yearbook) has been issued since 1958. The LGS was reestablished on April 26, 1989. In 2007 the LGS had about 300 members.

Submitted by Dovilė Krupickaitė
Livelihoods

Sarah Turner
McGill University, Canada

Livelihoods analysis has become a widely used approach in academic and applied development circles to better understand the numerous components of poverty and marginalization, as well as how people escape such circumstances. Since gaining prominence in the mid- to late 1980s, the term “livelihoods” has been the centerpiece of a variety of approaches, frameworks, and methodologies spanning diverse disciplines and scales. At their most fundamental level, these approaches aim to achieve a holistic, actor-focused understanding of how individuals and households work to create and sustain a means of gaining a living, often set within a context of poverty or vulnerability. At the same time, development discourse has increasingly prioritized environmental protection and the sustainable use of natural resources. Emerging scholarship on sustainable livelihoods thus lies at the interface of poverty studies and research on the environment. Since the terms “livelihoods” and “sustainable livelihoods” and the notions surrounding them are often used interchangeably, this entry discusses both concepts.

Two theoretical debates enhanced the initial appeal of the livelihoods concept within development theory and practice. One was the impasse that emerged in the mid-1980s regarding a relevant way to move forward with development theories. Scholars had become critical of the overly structuralist explanations of uneven development offered by the dependency and neo-Marxist theories of the 1970s and 1980s. While livelihoods perspectives kept attention focused on the unequal distribution of power and resources in society just as these former theories had, the livelihoods approach was argued to be more balanced, emphasizing the role of human agency in coping with vulnerability, poverty, and structural change. Second, the livelihoods approach presented an alternative to the inadequacies of modernization approaches to poverty that chiefly focused on income earnings and consumption levels. Advocates of a livelihoods approach thus argued for a more multidimensional and situated concept that perceived local development through the lens of lived experience, with a focus on households, networks, and communities.

Defining “livelihoods” and “sustainable livelihoods”

At its most basic, a livelihood is a means to a living. Livelihoods are dynamic, relying on emerging opportunities that are embedded within constantly changing social and political contexts. It is difficult – and, many would argue, simplistic – to try to measure complex local household realities by monetary income alone. Livelihoods analysis thus takes a holistic approach, looking at multiple components that contribute to the possible living of a household or individual. A frequently cited definition of a livelihood is “the assets (natural, physical, human, financial, and social capital), the activities, and the access to these (mediated by institutions and social relations) that together determine the living gained by the individual or household” (Ellis 2000, 10).
LIVELIHOODS

Because circumstances and situations change over time, an increasingly common feature of livelihoods studies has been a focus on sustainability. A livelihood is considered sustainable when “it can cope with and recover from stresses and shocks, maintain or enhance its capabilities and assets, while not undermining the natural resource base” (Scoones 2009, 175). Ideally, sustainable livelihoods result in maintainable opportunities for the next generation and contribute net benefits to other livelihoods at the local and global levels and over the short and long terms. Key authors involved in the initial creation of these definitions include Chambers and Conway (1992), Carney (1998), and Ellis (2000).

The sustainable livelihoods approach is particularly successful at conceptualizing livelihoods holistically. From this perspective, livelihoods comprise the capabilities, tangible and intangible resources (assets), and activities (or livelihood strategies) – plus access to these, mediated by social relations and institutions – which individuals or households draw on to gain a living. As Chambers and Conway (1992, 6) have suggested,

A livelihood comprises the capabilities, assets (stores, resources, claims and access) and activities required for a means of living; a livelihood is sustainable which can cope with and recover from stress and shocks, maintain or enhance its capabilities and assets, and provide sustainable livelihood opportunities for the next generation; and which contributes net benefits to other livelihoods at the local and global levels and in the short- and long-term.

Anthony Bebbington (1999) adds that sustainable livelihoods build on the assets people have at hand to provide them with outcomes such as material wellbeing, a meaningful existence, and the capability to shape the larger context in which they live.

The concept of sustainable livelihoods emphasizes the fact that people use a variety of assets and activities to build their livelihoods. By concentrating on the individual or household scale, the approach provides a crucial alternative to the abstract, macroscale discussions typical of sustainable development discourse. Sustainable livelihoods research also sees sustainability as a multifaceted and multiscalar notion, while adding a temporal component to earlier ways of conceptualizing livelihoods. Above all, advocates argue that sustainable livelihoods analysis allows for a comprehensive portrayal of the complex composition of livelihoods, including aspects of vulnerability and meaning, while tackling poverty in a transformative way. More specifically, sustainable livelihoods approaches have been considered to be a set of principles guiding evidence-based development interventions; an analytical framework to help understand what is and what could be done to make livelihoods sustainable as an overall developmental objective.

Livelihoods frameworks

Building on the livelihoods and sustainable livelihoods concepts, different frameworks have been designed as analytical structures to better understand the complexities and components of livelihoods and how interventions might best be made. The most familiar frameworks have been conceived of by the International Development Studies Institute, University of Sussex (Scoones 1998), the United Kingdom’s Department for International Development (DFID 1999), and Frank Ellis (2000). These frameworks focus on analyzing how impoverished people draw on their assets to formulate various strategies for making a living. The proponents of such frameworks suggest that the poor choose from
a range of available livelihood pursuits to meet their needs, within the constraints of their given locality and context. Livelihoods analysis considers how this context shapes people’s access to assets, the strategies they are able to pursue, and the resultant outcomes. In turn, the frameworks aim to depict how people subsist, and envision how contextual constraints to sustainable livelihood goals can be overcome. Although these frameworks all differ slightly, the general building blocks are fairly similar. One difference to note is that the DFID livelihood framework includes feedback loops, highlighting the complex and interlinked nature of livelihood creation and composition.

**Livelihood assets or capitals**

Assets or capitals – and how these may be used, combined, expanded, and reshaped – are at the core of determining the feasible set of livelihood strategies an individual or household can pursue to meet their material and experiential needs and motivations. Scholarship on sustainable livelihoods borrows from economics in seeing these assets as capitals which can be accumulated and invested in pursuit of livelihood outcomes. These capitals go beyond mere livelihood resources, though; they are assets that give people the capability to be and to act, permeating lives (cf. Sen’s (1985) capabilities approach). This broad view illustrates how low monetary incomes do not preclude one from having a range of other assets to provide livelihood options and wellbeing. By calling attention to what people have instead of emphasizing what they lack, the sustainable livelihoods approach, it is suggested, lends an outlook of optimism and individual agency to development discourse.

Five types of capital, together commonly known as the “asset pentagon,” are generally used in livelihoods and sustainable livelihoods frameworks:

1. **natural capital**: the local stock of environmental resources, including both renewable resources (such as water, forests, soil) and nonrenewable resources (such as extractive resources like metals and oil) on which livelihoods can be based;
2. **physical capital**: human-produced capital goods (productive resources), such as infrastructure (including roads, communication networks, irrigation schemes), tools, and machinery that facilitate production processes and market exchanges;
3. **financial capital**: wages, savings, credit, remittances, pensions, and precious metals (often in the form of jewelry) which are valued for their convertibility into other types of capitals or consumption goods;
4. **human capital**: a person’s knowledge base, education level, skills, and health status, and the resultant quantity and quality of labor individuals or households can draw on to fulfill their productive and reproductive duties;
5. **social capital**: heralded by the World Bank as perhaps the missing link of development, this includes the rules, norms, obligations, reciprocity, and trust that are found embedded in social relations, structures, and institutional arrangements and which enable people to achieve individual and community objectives, including gaining resources. Like other forms of capital, social capital is subject to accumulation and depletion.

Although these are the five most commonly recognized livelihood capitals, some authors have advocated for cultural capital to be treated as a separate category (rather than it being sometimes subsumed within social capital). Drawing on the work of Pierre Bourdieu (1984), cultural capital (simply put), relates to the skills, knowledge, education, and possible social advantages...
LIVELIHOODS

a person has that can give them benefits in society. Bourdieu divides cultural capital into embodied, objectified, and institutionalized forms. It is sometimes also argued that political capital should be considered as a separate asset, as discussed shortly.

These capitals are interrelated and affect each other; the acquisition or exploitation of one may hinge on the ability to mobilize or access another. For instance, social capital can mediate access to natural resources, while human capital is required to put natural capital to productive use. When necessary, physical capital (e.g., livestock) can be converted into financial capital (e.g., money). To a certain extent, capitals can be substituted for one another, with different assets demonstrating differing potentials for substitution. Fertilizer (physical capital), for example, can compensate for declines in soil fertility (natural capital) to some degree. Substitutions can also occur within asset categories, such as when labor is reallocated from domestic duties to cash cropping. Studies have demonstrated that the capacity to flexibly substitute and convert between assets is often critical for successfully coping with shocks and seizing new opportunities. Since livelihood strategies rarely draw on or enhance all the capitals/assets at once, their sustainability must be considered in light of the combinations of and substitutions between assets and their overall stock through time.

Access

Access determines how individuals acquire (or miss out on) the capitals needed for their livelihoods. While assets/capitals might be considered the building blocks of livelihoods, access to these building blocks is essential for creating a livelihood that meets current needs and can respond to changing circumstances. The three intricate mechanisms which mediate household and individual access to assets and activities include social relations, institutions, and organizations.

Social relations reveal how different social positionings mediate access, grounded in factors such as ethnicity, gender, kinship, religion, and class. Institutions can be defined as “the formal rules, conventions, and informal codes of behaviour, that comprise constraints on human interaction” (Ellis 2000, 38). Institutions are regularized practices or patterns of behavior, structured by the rules and norms of society, which have persistent and widespread use. Institutions are not absolute, but rather are created through social interactions and negotiations, structuring and regulating how society functions. Examples of institutions include land tenure arrangements, traffic rules, and the custom of marriage. Finally, organizations include government agencies, nongovernmental organizations (NGOs), cooperatives, and so on, organized around a common set of objectives.

Focusing on social relations, institutions, and organizations allows for a multiscalar analysis of access, for instance, revealing how a state policy or NGO can impact a household’s livelihood strategies. Concurrently, it is also important to note how social exclusion can occur through these mechanisms, with specific individuals and households (such as women-headed households) potentially locked out of access and opportunities.

Livelihood strategies and diversification

Livelihood strategies are the ways that individuals and households choose to use, transform, and reproduce their capital stock or resource base in pursuit of livelihood goals. Scoones (1998) suggests three main livelihood strategies that are often available in rural contexts: agricultural intensification, agricultural extensification, and
migration. A fourth strategy, livelihood diversification, may include a mix of these three.

Frank Ellis (1998, 4) defines rural livelihood diversification as “the process by which rural families construct a diverse portfolio of activities and social support capabilities in their struggle for survival and in order to improve their standards of living.” Diversification can occur at multiple scales, such as an individual pursuing a variety of activities such as farming, trade, and contractual labor, or members of a household specializing in particular activities but in aggregate bringing together a diverse portfolio of livelihood strategies. Migration is a common livelihood diversification strategy, making a household’s organization of livelihood activities multispatial. As with assets, the relative importance of different livelihood strategies changes through time, across regions, and even within a given household.

Individuals and households diversify their assets, incomes, and activities for various reasons. Push factors include the need to expand safety nets, mitigate risks, and resist and/or cope with shocks and stresses. Diversifying livelihood strategies in this way (also called “distress diversification”) can allow people to stabilize their income and regulate consumption in case of crises or seasonal shortfalls. Alternatively, pull factors, or “progressive diversification,” can involve the realization of new opportunities. In addition, as noted by Sarah Turner (2007), a strategy of “selective diversification” may result in flexibly engaging with certain livelihood strategies at opportune times and withdrawing from them during other periods, thus moving beyond a dualistic classification of diversification. Crucially, sustainable livelihoods scholarship recognizes that diversification plays a central role in livelihoods, and that the “rural” is not strictly synonymous with agricultural activities.

Examining diversification strategies highlights the importance of taking scale into account when researching livelihoods. Although livelihood approaches tend to be focused on the microscale, one still needs to consider whether the individual, household, or some other grouping should be the unit of analysis. Chambers and Conway (1992) situate livelihoods at the level of the household, which they define as a group of people that eat from the same hearth. Yet this analytical unit may not always be the most appropriate; households can be split by intrahousehold conflict or differentials in capitals and capabilities. As a result, different members will be impacted differently by changing circumstances: gender and age differences within the household, for instance, are likely to affect individual outcomes. Furthermore, the definition of a household is up for debate, since not all contributors to a household’s common livelihood necessarily live under the same roof, and remittances may constitute a significant portion of a household’s financial capital.

Vulnerability context and transforming structures and processes

The livelihoods literature calls attention to how the external environment affects people’s options and strategies for making a living. Some analysts distinguish between an exogenous “vulnerability context,” over which people have little or no control, and sociopolitical aspects endogenous to the norms and rules of society, deemed “transforming structures and processes.” The vulnerability context includes people’s exposure to shocks, trends, and seasonal changes. Shocks include human-induced or “natural” stresses such as natural disasters, droughts, conflicts, crop or animal disease, and death. Trends include fluctuating resource stocks and population growth,
LIVELIHOODS

as well as changes in technology and governance, while seasonality can affect production, prices, health, and employment opportunities.

In turn, transforming structures and processes are composed of structures which include public and private organizations such as the government, NGOs, and commercial enterprises, and of processes such as the laws, policies, cultural norms, and institutions that shape human interaction. Attention to these external conditioning variables, it is argued, links the livelihoods framework's micro and macro elements, providing a launching pad for policy-based interventions.

Livelihood outcomes

One of the key contributions of the sustainable livelihoods approach is its holistic conceptualization of what constitutes a sustainable living, challenging practitioners and researchers to question the desired ends of “development.” These ends, or sustainable livelihood outcomes, can include income, dignity, power, and sustainability; put differently, the approach implies an improved quality of life according to a household’s own criteria (Bebbington 1999). From this perspective, the assets that serve as inputs to livelihood strategies are also livelihood outcomes themselves. This bidirectional notion is a strong point of the sustainable livelihoods literature.

Methods

There is no set group of methods that is drawn on to undertake livelihoods analysis. Often a mixed methods approach is used, bringing together quantitative approaches such as village- or community-level surveys and qualitative approaches such as in-depth conversational and semistructured interviews, oral histories, focus groups, and observations. Participatory approaches are sometimes incorporated, including participatory rural appraisal (PRA) techniques such as participatory video and field-mapping. As with any participatory approach, such methods can raise concerns over representation and the “myth of community.” Particularly when used by development practitioners, a livelihoods analysis can also run into trouble if local interviewees downplay their assets because of taxation or theft concerns, or in order to benefit from possible interventions.

Critique of the sustainable livelihoods literature

Though it presents a dynamic portrait of the relationships between rural livelihoods, poverty, and the environment, sustainable livelihoods scholarship is limited in certain ways. One shortcoming is in its handling of poverty. While the concept of poverty is considered essential for discussing livelihoods, no explicit mention is made of poverty’s place within the notion of sustainable livelihoods, nor are the differences between definitions of poverty properly emphasized. Indeed, a focus on capitals privileges research and developmental interventions among poor people who already have some assets, at the expense of the completely destitute with minimal assets. By focusing on shaping environments that facilitate capital accumulation, the safety nets required to protect people from poverty can also be glossed over.

There is also concern that, since the 1990s, endless livelihood case studies have been amassed, but little research across differing scales has emerged. As such, no generalizable trends have become apparent from which comparative explanatory frameworks can be created or existing theories challenged. Academics such as Leo
de Haan (2012) thus argue for the combination of meta-analysis and comparative research to try to extend livelihoods research.

Critiquing capitals

While it may serve as a useful concept, the notion of capitals is contentious. “Capitalizing” key aspects of a livelihood using the language of neoclassical economics can be argued to be reductionist, as is the use of economic terminology to understand microscale resource use and household subsistence decisions. Often, those working through a human geography or anthropology lens aim to theorize capitals/assets within a relational context, while the language of neoclassical economics tends to remove assets from this crucial context.

Given the way sustainable livelihoods research couches both tangible and intangible resources in the language of economics, it is unsurprising that analyses can often end up ignoring ways in which capital can be drawn on to work toward nonmaterial ends. For instance, seeing human capital in terms of labor alone undervalues the way it enhances people’s capabilities to question, challenge, and transform the world. The same criticism applies to the concept of social capital, which is rarely considered for its intrinsic value, such as giving people voice, stimulating participation, and enriching lived experience. Sustainable livelihoods work also tends to ignore negative aspects of social capital, whereby certain individuals or groups are able to deny others access to resources and opportunities.

The asset pentagon is also subject to criticism since defining, measuring, and calibrating across assets is complex, as is calculating tradeoffs between often incomparable capitals/assets such as tangible resources and ongoing processes. Critics argue that this has weakened the sustainable livelihoods approach by not questioning or engaging the linkages and substitutions between different capitals/assets thoroughly enough. Yet proponents have suggested that these concerns can be overcome by working to understand what specific assets mean to particular groups in certain situations and contexts. For instance, an asset such as land is seldom just a tool to realize financial and material ends, but frequently gives meaning to a person’s world.

Power relations

A common critique of the livelihoods and sustainable livelihoods approaches is that they fail to adequately engage with the role of power relations and politics. According to Toner (2003, 772), the sustainable livelihoods approach provides an “individualized,” “depoliticized,” and “benign” portrayal of how people make a living and manage their resources. Its silence on power issues is puzzling given how power routinely (re)produces poverty. With a few exceptions (e.g., Ellis 2000), sustainable livelihoods frameworks have glossed over how everyday life and poverty are shaped by institutions like gender, property regimes, and power relations. This hides from view a crucial portion of the picture that should be analyzed, and underestimates people’s full potential for action. Some theorists therefore argue for the inclusion of political aspects in the definition of sustainable livelihoods, with political capital analyzed in its own right, as noted earlier. While there is some room in commonly used livelihood definitions for considering political power as part of social relations, institutions, and organizations, it can also be argued to be a capital asset which people accumulate or draw on in pursuing livelihood options. Others have instead turned to actor-oriented approaches to explore and explain how macro forces of change are experienced and negotiated locally, drawing on tools
such as the “social-interface” (Long and Long 1992).

Debating strategy

Another aspect of the sustainable livelihoods approach that scholars have questioned is the assumption that individuals act as strategic managers, rationally selecting from a range of livelihood options within given contextual constraints. Structuralist authors, on the other hand, argue that cultural, historical, and social forces govern human action to the point that individual agency cannot truly explain behavioral patterns. From this perspective, livelihood “strategies” are less the result of rational choices and actions than a reflection of power configurations and structural conditions. Structuralists ask whether selection between extremely limited options – for instance, starvation versus prostitution – truly represents a choice, contending that an emphasis on individual agency turns a blind eye to the historical, institutional, economic, and social conditions that people do not necessarily choose but that shape the ways they build their livelihoods. Analyzing livelihoods as outcomes of rational strategic decisions hence ignores how the factors that condition behavior are often cultural and ideological and not just a matter of individual personality or skill.

In response to such critiques, livelihood authors have underscored that, in order to comprehend livelihood choices and decision-making, it is vital to ascertain how these are framed by people’s own definitions and perceptions of wellbeing and poverty. Some, including Leo de Haan and Annelies Zoomers (2005), have advocated for the use of the term “livelihood pathways” in place of “strategies.” Rather than aiming to attain a preset goal, livelihood pathways highlight the iterative process through which a livelihood is fashioned, with goals, preferences, and assets being continuously re-evaluated.

Defining sustainability for livelihoods

The term “sustainable livelihoods” itself has been critiqued by authors who question whether, despite its title, this approach actually attempts to integrate sustainability with other concerns. The main question regarding the literature’s notion of sustainability is its opacity regarding the relative importance of assets like capital stocks and livelihood strategies. Though the sustainable livelihoods approach lends valuable insights by considering sustainability in terms of combinations and substitutions between the five capitals, some questions – such as the degree to which livelihoods require the conservation of certain environmental features in order to be considered sustainable – remain unaddressed. Some aspects of otherwise sustainable livelihoods may indeed be negative, such as when a stable, “sustainable” form of social organization supports the maintenance of a status quo that perpetuates poverty. Scholars working within the sustainable livelihoods framework also acknowledge that strategies can be sustainable at certain scales and not others, and recognize the tensions between short- and long-term livelihood goals; they do not, however, clearly address how to manage these issues. Likewise, the sustainable livelihoods literature has still not sufficiently theorized the temporal aspects of sustainability.

Concluding thoughts

In considering these critiques, it should be noted that no one model alone can fully capture the totality and complexity of lived experience, poverty, and the environment, yet such research does usefully advance understandings
of livelihoods and sustainability among social scientists and development practitioners. Perhaps using what is gleaned from this important work alongside other approaches (such as political ecology, actor-oriented approaches, and gender studies) is the best way to draw on its strengths while addressing its weaknesses.

SEE ALSO: Human capital; Networks, social capital, and development; Participatory development; Power and development; Social capital; Sustainable development; Vulnerability

References


Further reading

Staples, James, ed. 2007. Livelihoods at the Margins: Surviving the City. Walnut Creek, CA: Left Coast Press.
A living wage is the minimum hourly wage required to allow a worker to support her/himself and her/his dependents. Although the household circumstances of each worker will be different, the wage rate is calculated to reflect a locally determined minimum acceptable standard of living. Calculations are made to measure the cost of living, including everyday items (such as housing, food, transport, and the ability to engage in the local community), adjusted for household characteristics, such as the presence of a working partner and the number of children. Once agreed as an hourly wage rate, the living wage is designed to set an ethical standard for wages in a geographical labor market and this can be urban, regional, or even national.

The demand for a living wage first surfaced alongside the organization of workers into trade unions in the English coalfields more than 120 years ago. While the concept of a living wage was largely overshadowed by the implementation of minimum wages and the welfare state by national governments during the twentieth century, the demand has recently resurfaced in the context of growing rates of in-work poverty. In its more recent manifestation, the demand for a living wage arose in Baltimore in the early 1990s and subsequently spread to other cities across the United States. It has also been taken up in Canada, New Zealand, and the United Kingdom with related calls to include living wages as part of ongoing efforts to improve labor standards in international supply chains. In the United States, living wages have been agreed at the urban and county scales, whereas in the United Kingdom the campaign began in London and has subsequently determined an out-of-London rate to apply to the rest of the country. The geographical literature includes documentation of the early campaign in Baltimore (Walsh 2000) as well as the development and impact of the campaign in the United Kingdom (Wills 2009; Wills and Linneker 2014).

The growth of in-work poverty and the associated demand for a living wage have triggered a growing body of social science research that seeks to measure the impact of improved wages at the workplace, community, and national scales. Researchers have explored the impact of improved wages on productivity, worker morale, rates of labor turnover, and standards of service in a variety of contexts. In addition, there is a growing body of research looking at the wider impacts on the health and wellbeing of workers and their families, as well as savings made to the taxpayer in reduced welfare spending (Flint, Cummins, and Wills 2013).

Living wage campaigns have wider salience in relation to changes in social movements and the increased demand for economic justice and reduced inequality. The campaigns have also exposed the wider issue of in-work poverty as well as the related prevalence of subcontracted employment and heavy reliance on migrant and minority labor in the lowest paid jobs. As such, living wage campaigns have been associated with the development of new forms of political activism often led by marginalized migrant communities.
LIVING WAGE

SEE ALSO: Labor geography; Labor market; Social justice; Social movements; Trade unions; Unionism, community

References


Further reading

Local development

Costis Hadjimichalis
Harokopio University, Greece

Local development is an explicitly social and territorial approach to development including not only economic aspects but also employment creation, poverty reduction, quality of life, and environmental sustainability. In its ideal conceptualization, local development aims to empower local communities to shape, in a participatory and democratic way, the future of the places where they live. Successful local development depends on complex processes of consensus between local social actors – economic associations, firms, political leadership, unions, institutions and people – interacting in a particular territory, aiming at the welfare of everyone living in that territory or locality, without discrimination based on class, social status, age, gender, ethnicity, race, and sexual preference. Such processes may sound utopian under conditions of neoliberal capitalism but nevertheless may function as prerequisites for progressive local development planning. Application of a local development strategy will usually contain several steps: defining the territory and the problem or development issues under consideration; defining and analyzing the parameters of problems/issues and checking whether they generate from elsewhere; sensitizing local actors; promoting a local forum and organizing participation without exclusion; designing a local planning strategy and securing financial and political support; implementation; and evaluation and monitoring. In reality these steps often tend to feed into each other.

Theoretical background and novelties

Local is often conceived as the lowest scale (from “national” to “regional” to “local”) for framing research and policies. It arrived relatively late into the development repertoire with national development being discussed immediately after World War II, regional development from the 1960s onwards, and local development as an independent concept from the 1980s until today. In early studies, local was not adequately theorized – as was the case with regional – and was simply equated with the lowest available administrative boundaries, where politicians and policies are scrutinized and where statistics are available. The US federal system provides a nexus of formal political power to localities, and local means practically local government, which in turn manages local economic development (LED). A characteristic of the US system, in the absence of a clear regional development framework, is the rise of strong local interests, frequently forming local growth coalitions and the bitter competition among local governments to attract inward investments.

In other parts of the world, local development and its associated concept of endogenous growth came as a reaction to inadequacies of national and regional development frameworks. Since the 1950s and until the crisis of Fordism, dominant models of regional development were based on industrialization via capital incentives or large state projects in growth poles, provision of public infrastructures, and employment creation via inward investments from other regions. It was the period of, what is called, welfare regionalism largely based on Keynesianism and social
democratic political principles, predominant in Europe and beyond at that time. The economic crisis of the 1970s and the many inadequacies of the welfare state, including regional policies, paved the road to major changes and regional top-down welfare planning was hit from all sides.

Empirical investigations in the late 1970s discovered cases of “spontaneous” growth, away from old industrial cores and subsidized backward agricultural regions, without direct assistance from the central state or inward investment, where small firms with strong entrepreneurial spirit initiated a “bottom-up” local growth. Paradigmatic local areas included the Third Italy in northeastern-central Italy, Silicon Valley and Orange County in California, the “M4 corridor” in the south of the UK, and southern Bavaria in Germany, while subsequent research came to include many more examples in other parts of the world. Small firms in these localities successfully combined market opportunities mobilizing locally existing resources (particularly knowledge and learning) and new forms of production organization such as flexibility and networking. These new industrial spaces highlighted the role of competition and cooperation at the local level with the assistance of strong local cultural traditions, local institutions, and associations. Researchers and policymakers agree that a distinctive feature of these places and localities is the embeddedness of certain noneconomic factors such as social capital, trust, and reciprocity based on familiarity, face to face exchange, cooperation, embedded routines, habits and norms, and local conventions of communication and interaction, all of which, in turn, have a key role to play in successful local endogenous development.

Pioneering research on local development came from different directions. First, there was the Italian School of Third Italy’s Industrial Districts (IDs) which since the 1970s provides, mainly by economists, brilliant analyses of networked small and medium size industries in various parts of northeastern-central Italy. The Italian School made several groundbreaking contributions such as the concept of a local flexible productive system, the notion of diffused industrialization, the identification of the importance of small and medium enterprises and their networking, the role of local culture, local social relations and knowledge, and the reuse of Alfred Marshall’s notion of localized external economies. Key figures include G. Becattini, R. Camagni, A. Sforzi, G. Garofolli, G. Dei Ottati, and many more.

Second, there was the French School of Milieu Innovateur (innovative local milieux), developed by the economic Research Group GREMI (Groupement de Recherche Européen sur les Milieux Innovateurs) in the early 1980s. Their hypothesis was based in Philippe Aydalot’s idea that behavior initiating innovations are based on local or regional factors rather than national or global ones. Extending Schumpeter’s ideas on innovation, GREMI combined two French traditions: the older milieu concept from sociology and regional geography; and that of regulation regimes. Apart from Aydalot himself other contributors include M. Quévit, D. Maillat, O. Crevoiser, and, later, C. Courlet and B. Pecqueur.

Third, there was the British industrial restructuring perspective, applied in the mid-1980s in the research program “The Changing Urban and Regional System” (CURS), which studied industrial restructuring in particular areas via local labor market transformation. The program, known also as “locality studies,” focuses on seven localities which were chosen and studied by geographers, sociologists, planners, and few economists for their different experience of decline or growth. Among the protagonists was P. Cooke, later joined in this research perspective
by K. Morgan, A. Amin, and N. Thrift, while other CURS participants, such as R. Hudson and H. Beynon developed a much more critical view. Major findings of CURS include the negative effect of Thatcherism in regional policy, the need to link local to global processes for understanding restructuring, the importance of local proactivity, and the propensity for networking by local actors.

Fourth, in California, again in the 1980s, economic geographers developed innovative research on agglomeration economies and transaction cost analysis. Inspired by the Italian industrial districts and regulation theory, agglomeration and transaction are analyzed quantitatively and qualitatively. A. Scott, A. Saxenian, M. Storper, and D. Walker were among this group of researchers. Although their focus was on particular industrial localities, they underlined the importance of competitive regional economies in general at times of globalization, and they stressed technological learning, based on traded (input–output relations) and untraded interdependencies (labor markets, regional conventions, norms and values, and public or semi-public institutions).

These pioneering analyses, despite important differences among the localities studied, have been used as paradigmatic examples for the introduction of synthetic and policy oriented concepts such as “networked firms and regions,” “industrial clusters,” “learning firms and regions,” and “innovative firms and regions,” forming what has been called by others the “New Regionalism.” According to both theorists and policymakers, while in the past local development took place in a spontaneous manner, now it can be designed to implement a policy “from below,” and international organizations, such as the Organisation for Economic Co-operation and Development (OECD), the World Bank, the European Union (EU) and several national governments initiated programs to promote such policies, using the aforementioned success stories as “best practice.”

Misreadings, omissions, problems

In the 1990s local development became the new catch phrase, and the “New Regionalism” a new kind of analytical and development doctrine during a period of major political changes in the Global North. In the euphoria that followed the “discovery” of dynamic small firms in these paradigmatic industrial districts, several misreadings, omissions, and problems occurred. First, there was a simplistic binary opposition between mass production and flexible-specialization, whereas Fordism was never dominant in the regions in which growth was directed by flexible small industrial production ensembles. Second, a similar binary opposition occurred, with endogenous and exogenous factors guiding development. In reality development is always the outcome of the dialectical tension between the two, while the social and environmental effects of this tension are experienced mainly locally. Third, there was a major confusion of scale: in some cases local development was analyzed with reference to traditional agglomerations of small industrial firms with a few thousand inhabitants, to whole regions with multiple sectors, or to large urban conurbations with several hundreds of thousands of firms, while some cases were identified as “local” simply due to their coincidence with lower scale administrative boundaries. Fourth, and related to the problem of scale, is the study of localities (mainly by economists) as bounded territorial entities and as though they were firms. Fifth, in the relevant literature, a very selective appropriation of the complexity and richness of
local productive systems has taken place in which only certain general economic, organizational, and institutional issues have been taken on board. While others, such as power and inequalities among small firms and within industrial districts, the limitations of networking and learning, the real meaning of cooperation, reciprocity, and “social capital” for firms and labor, conditions of work, the informal economy, gender and ethnicity – to mention but a few – remain unexamined. Sixth, there is a lack of attention to the role of the state and other supranational entities in the introduction of various protectionist and assistance measures. Due to the important role of the devaluation of a national currency, labor legislation, and particular incentives for small firms and international agreements (such as the Multi-Fibre agreement), there was never such a thing as “without assistance from the central state.” Seventh, by looking only at success in the context of competitive capitalism, there was interest in few paradigmatic industrial sectors and few advanced service providers in a limited number of European and North American regions. Finally, eighth, and as result of the above, there has been a remarkable silence/neglect of other sectors such as tourism, trade, and agriculture in areas beyond North America and Europe and those millions of ordinary localities that form the majority everywhere and are in real need of development and prosperity.

Perhaps the most important omission in this research framework and in much of the “New Regionalism” literature, and what really remains unmentioned and unexamined, is the uneven relations among localities, successful and unsuccessful, in other words the hard question of uneven geographical development under capitalism. This omission was evident in the absence of attention to the capitalist crisis that was already visible from the mid-1990s in some IDs of Third Italy and in other emblematic localities and regions due to three interrelated processes: (i) mergers and acquisitions and formation of large, vertically integrated firms and groups of firms; (ii) de-localization of a part of production or of all firms to low labor cost regions and countries; and (iii) extensive replacement of local labor by medium- and high-skill immigrants to compensate for increasing labor cost and/or the lack of skilled native labor. These developments directly challenge grandiose claims about flexibly specialized small firms in IDs as models for the future and about the social and cultural continuity of the paradigm based on trust, reciprocity, and social capital. Just at the very moment that policy prescriptions based upon the assumed bases of success of these localities were being generalized through the design and implementation of regional and urban policies across the globe, the conditions on which success was based in these exemplar regions were being eroded.

Policies for local development

Policies promoting local development were not initiated in the 1980s, nor were they all inspired by the theoretical framework presented above. They have a long history and one can recall cases, among the many existing worldwide, in the northeast of England, in the Ruhr area in Germany, and in the polders of the Netherlands during the late nineteenth to early twentieth centuries, and, since World War II, in socialist Yugoslavia and Tanzania and in communist China. In these cases detailed local policies were implemented (without the intention of “development from below” as such) dealing with different problems in each case, from managing water, to building social housing, linking coal fields to ports, and promoting the decentralization and ruralization of industrial production. In the United States the Local Development
LOCAL DEVELOPMENT

Districts Policy in the Appalachia region has a 30 year long history dating from 1965, while community development policies in urban, working class neighborhoods were precursors of later LED policies in urban areas.

Since the 1990s many countries and major global institutions such as the OECD, the World Bank, and the EU started programs dealing explicitly with local development. In 1991 the EU LEADER initiative was announced for the period 1991–1994 as a Europe-wide experiment, and involved the first 217 local action groups (LAGs) which were representatives from small and medium-sized enterprises, local authorities, non-profit organizations, and other local players from various sectors who agree on a common approach to the interests involved in developing their local rural areas. Three other initiatives also embodied similar principles of local development, especially that of partnership: URBAN for urban areas; INTERREG for cross-border and interregional collaboration; and EQUAL for partnership in social integration, combating discrimination in whatever guise, inequality in the labor market and continued vocational training. In Europe another actor was LEDA, an independent association which brings together organizations, practitioners, research centers, and others who are active in the creation of local employment in EU countries.

The OECD’s involvement with local development dates from the 1970s with various programs and recommendations but perhaps the most advanced framework was initiated in early 1980 with LEED (Local Economic and Employment Development) with three main objectives: to improve the quality of public policy through continuous monitoring and assessment of current practices; to promote innovation in local economic and employment development across the globe; and to support the design, implementation, and evaluation of development strategies to help grow local economies. The OECD gives particular attention to tourism and culture as important pillars for local development, finances particular projects in many countries, and, since 2003, operates the LEED Trento Center for Local Development, in the Autonomous Province of Trento in northern Italy.

The World Bank also promotes local development projects worldwide, particularly in Latin America and Southeast Asia, without any particular spatial reference. For the Bank local development means creating a favorable environment for business success and job creation in urban areas. To do this optimally, LED is undertaken through partnerships between local government, business, and community interests. Finally, the UN in the 1992 World Earth Summit at Rio adopted the Local Agenda 21, which provides guidance to local governments in the preparation of local development plans integrating the principles of sustainable development. Local Agenda 21 introduces explicitly the environmental dimension in LED, such as the concern with the per capita ecological footprint, and, in this respect, is a major innovation and critique to previous models, unfortunately with limited applications.

Beyond recognized and formalized models of local development, however, there is a whole world of informal local development practices in both the Global North and South. These cases are classified by some development specialists as non-“modern” or non-“competitive,” while in reality they constitute the best way to increase wellbeing for particular localities. They include a large variety of productive and non-productive activities and local social actions such as – to mention but a few – the development of locally oriented agriculture and linking of producers directly with consumers, self-help housing, social health centers, associations and co-operatives, time banks, local social currencies, activities
LOCAL DEVELOPMENT

in the arts and culture, and larger scale initiatives such as participatory budgeting, operating through more ethical, gender, and environmentally sensitive processes. Often they are classified under the concept of social economy or solidarity economy, and they share a strong local vision of development and a much broader and socially sensitive approach to what constitutes the economic.

The way ahead

Since the 1980s policy options such as “fostering dynamic regional growth” came to replace the old “reducing regional inequality,” and, under neoliberalism, a transition occurred from integrated, comprehensive, and nationally organized regional welfare planning toward intensely competitive and geographically fragmented processes which may be called “entrepreneurial regionalism.” Local development in this context risks being used to introduce hard competition among localities or to substitute for other policies at times when direct welfare intervention on the part of the central state is discredited. Localities cannot be understood as competitive firms or as bounded administrative territories. Therefore, more nuanced interpretation of development trajectories is needed, using “de-localizing” and “de-scaling” discourses, while keeping the principles of progressive local development. Local development projects can learn from experiences in the Global North and South, both formal and informal, and from theoretical developments since the 1980s, keeping in mind the un-transferability of success and the need for cooperation and solidarity.

SEE ALSO: Regional development models

Further reading


Local embeddedness

John Pickles
University of North Carolina at Chapel Hill, USA

“Embeddedness” refers to the ways in which relational, institutional and cultural contexts shape economic and social life. In economic geography, it has two distinct albeit related referents and foundations. The first derives from Polanyi’s *The Great Transformation* (1944) and concerns the changing ways in which the economy is integrated into broader social systems. The second derives from Granovetter’s (1985) attempts to discern the role of social action in economic contexts, and to challenge the undersocialized conceptions of economic behavior found in transaction cost and agglomeration economies.

For Polanyi, economies were historically embedded in social and cultural institutions. With the deepening of capitalist market relations and market autonomy, the economy became progressively disembedded from these social and cultural institutions. Economic sociologists and geographers have responded to the disrupting effects of economic globalization by focusing on the ways in which economic actors are embedded in systems of interfirm relations, or chains and networks, and how these forms of organization and their corresponding forms of governance shape their behavior. For example, Gereffi and Mayer (2006) have drawn on Polanyi to suggest that the process of economic disembedding gave rise to global governance deficits as lead firms in global value chains were increasingly disembedded from national regulatory norms and laws. In response, social forces responded with efforts to create new forms of regulation and rule – norms about the length of the working day, health and safety regulations, struggles over wage rates and levels, consumer demands for product quality, and so on. This, they argued, was the Polanyian double movement.

Granovetter was much more skeptical of the Polanyian view of disembedding as a historical process, instead arguing that “embeddedness” refers to the fact that economies always take place “in concrete, ongoing systems of social relations” (Granovetter 1985, 487). It is these social relations that underlie economic transactions, enable relations of trust, maintain order in economic life, and explain why economic transactions can be normalized in conditions of great uncertainty. These relations are networks of interpersonal relations and attachments and, in order to understand economic practices, it is necessary to understand the networks of social relations that constitute them and have given meaning to them. Drawing on Albert Hirschman’s sociology of economic life, Granovetter (1985, 487) argued that actors do not behave or decide as atoms outside a social context, nor do they adhere slavishly to a script written for them by the particular intersection of social categories that they happen to occupy. Their attempts at purposive action are instead embedded in concrete, ongoing systems of social relations.

In this sense, it was important to avoid two traps. One trap was the undersocialization of the economy by focusing on atomized-actor explanations (such as utilitarian concepts of economic actors), with a corresponding overemphasis...
LOCAL EMBEDDEDNESS

on the utilitarian pursuit of self-interest. The other trap was the oversocialized explanation in which custom, culture, or values were seen to be full and sufficient explanations of why people behave in particular ways. Granovetter saw both traps as functionalist and deterministic. Instead, trust is a fundamentally social process that enriches a firm’s opportunities, access to resources, and flexibilities that are difficult to achieve in arm’s-length market arrangements (Bair 2012; Uzzi 1996).

While economic behavior is always embedded in and shaped by cognitive, social, cultural, and political practices, geographers have argued that such forms of embeddedness are also co-constituted by scalar networks that shape local, regional, and global economic geographies. The result is a spatializing of embeddedness that has informed much of the new regional economic geography over the past three decades. For Dicken and Thrift (1992, 283), the way we understand organizations is deeply linked to notions of “culture” as a set of conventions and intangible socially embedded processes that take particular geographical form, such as in research on Italian industrial districts and clusters. Parallel work on global production networks has similarly argued for a focus on the more complex relationalities of the networks of actors with which firms interact (Yeung 2005, 41). In this context, Yeung suggests that the concept of embeddedness represents “a major shift away from radical political economy’s focus on the social relations and social structures of production towards a focus on the socio-spatial economy organized around network paradigms, associational economies and relational geographies” (Yeung 2005, 41).

More recently, Hess has cautioned against the dangers of “over-territorialized” concepts of embeddedness, stressing the importance of not assuming that embeddedness is always a form of spatial rootedness in place, particularly when it is then used to refer to the “evolution and economic success of regions built by locally clustered networks of firms … industrial districts, creative milieu, learning regions or local knowledge communities” (Hess 2004, 166). For Hess, the strength of “new regionalism” is its focus on local and regional systems of economic relations and the social networks that ground them, but its weakness is that it overterritorializes the varying forms of institutional thickness and thinness and their corresponding levels of competitiveness in the global economy. These arguments bring the earlier sociology of embeddedness closer to actor-network theory, which sees in networks and assemblages chains of weaker and stronger association that produce the concrete places of the economy (see Callon 1998).

Cultural geographers have also developed notions of embeddedness through phenomenological, Marxist, and institutional economic traditions. Central to each have been the arguments of Antonio Gramsci, Jürgen Habermas, and the French school of regulation. For Habermas, for example, the economy is always embedded in lifeworlds of shared, often tacit, meanings and values. These taken-for-granted shared worldviews are not immutable but constantly shaped by economic and administrative logics that colonize the lifeworld in historical and geographically specific ways. Regulation theory similarly foregrounded the social relations in and through which economic life is organized. Here, social norms, codes, and beliefs are an essential component of any analysis of the specific regimes of accumulation and modes of regulation (Burawoy 2003).

SEE ALSO: Economic geography; Industrial districts; Industrial geography; Regionalism; Relational assets
References


Further reading


Local state

Mark Goodwin  
University of Exeter, UK

The local state refers to a set of state institutions and administrative agencies operating at a local scale. The concept of the local state first came to academic prominence in the late 1970s and early 1980s, following the publication of Cynthia Cockburn’s book *The Local State: Management of Cities and People* in 1977. The book was largely devoted to an empirical study of corporate management and community development in the London borough of Lambeth, in the United Kingdom, but it also set out a theoretical framework that viewed local government as a key part of the wider capitalist state. This was a crucial insight, which sparked a range of academic debates and empirical studies across the disciplines of political science, urban sociology, and human geography.

The usage and application of the concept of the local state have always been somewhat uncertain, and linguistically the term itself could equally well refer to “local institutions of the central state” or to an “autonomous local state.” In practice, where individual studies of any particular local state sit on this spectrum depends on both the theoretical framework employed and the empirical context of the research. The nature of the local state in a highly centralized and unitary political system such as the United Kingdom is bound to differ from that in a more federal and devolved system such as the United States. But in the aftermath of Cockburn’s book, work on the local state coalesced around three main avenues of inquiry, the first seeking to analyze the local state’s form and nature, the second investigating its functions, and the third exploring the relationship between the central and the local state. Sometimes researchers explored one or other of these avenues, but more usually the three were combined, as academics sought to understand what the local state was, what it did, and how autonomous it was.

For Cockburn herself, the local state was part of a whole, where the whole – the capitalist state – seeks to secure conditions favorable to capital accumulation by managing economic and social reproduction. The local state plays its part in this reproduction primarily through the highly gendered management of welfare and families locally; but it is always seen as part of the state in general, subject to central government and an integral aspect of the national state. Cockburn was writing from a Marxist perspective, which emphasized class relations and domination. Two years after her book, Peter Saunders published *Urban Politics: A Sociological Interpretation*, which also contained an empirical and theoretical account of the local state (Saunders 1979). Saunders again looked at a London borough, in his case Croydon, but he did so from a more managerial and corporatist perspective that stressed a degree of autonomy in local state activity. Saunders subsequently developed this work into his “dual state thesis,” which argued that distinct political processes operate at the national and local levels. According to Saunders, state intervention at the national level takes place mainly in relation to processes of production, through corporatist policy mediation (involving government, big business, and organized labor), and is addressed...
to the principle of private property and the need to maintain capitalist profitability. In contrast, local state activity is directed toward consumption processes, such as housing, health, and education, where policies are developed through a plurality of competitive political struggles, and are addressed to the principles of social rights and social needs. Saunders’s work highlighted the specificity of the local level and argued for a conceptualization of the local state based primarily on its uniqueness and on its separation from the center. While for Cockburn the local state was effectively the local arm of the central state, for Saunders the local state was local because it dealt with different issues through a different policy process.

Following the pioneering work of Cockburn and Saunders, the concept of the local state began to be widely used in human geography. Michael Dear, for example, used data from Canada to analyze historical trends in state expenditure on different functions at different levels, and concluded that the local state is primarily involved in providing services to reproduce the labor force. In later work with Gordon Clark he developed these ideas further, incorporating empirical data from the United States, to formulate the idea that the specificity of the local state arises from a crisis-absorbing function, whereby continued electoral legitimacy and local self-determination help to maintain political stability (Clark and Dear 1984). For Clark and Dear, surveillance and social integration of the population are facilitated by a decentralized state structure, and the local state helps to displace and manage structural crises at the local level.

In the United Kingdom, Simon Duncan and Mark Goodwin were developing perhaps the most geographically sensitive theoretical framework for analyzing the local state. For them, local state specificity is rooted in a set of factors that are inaccessible to the central state – namely, local social relations. In their 1988 book *The Local State and Uneven Development*, they trace the connections, both empirically and theoretically, between the uneven development of social and economic relations and the emergence of the local state. Put briefly, because such social relations are unevenly developed, there is, on the one hand, a need for different policies in different places, and, on the other hand, a need for local state institutions to formulate and implement these variable policies. Local state institutions are therefore rooted in the heterogeneity of local social relations, where central states have continued difficulty in addressing and responding to this differentiation. They go on to argue that this is a double-edged sword, for locally constituted groups can then use these local institutions to further their own interests, perhaps even those opposed to centrally dominant interests. This “representative” role of the local state is considerably sharpened by the development of representative democracy and universal franchise, for the extra leverage local state institutions give locally constituted groups is legitimized and even expected. For Duncan and Goodwin, then, local state institutions are at once an agent of, and an obstacle to, central control. Their stress on uneven sets of social relations highlights the fact that local states are more than just a group of locally based institutions and government structures. They are also viewed as a set of social relations themselves, and in this way local state activities contribute to the interpretation of how society works and why. Thus, local states have an interpretive as well as a representational role, and some will use this status to challenge the existing social order, while others reinforce it.

In contrast to the other theories considered, Duncan and Goodwin make no claims about the functions carried out at the local level. Indeed, they note that there seems no historical rationale
for any rigid division between the types of functions and types of politics characterizing local and national states, and point out that local state institutions are involved in activities in support of both production and consumption. They also pointed out that, because the local state is rooted in local social relations and performs both interpretive and representational roles, conflict with the central state is a constant possibility, although it is always an empirical question as to how this is realized locally and addressed centrally.

These theoretical frameworks were then populated by a range of empirical studies from the United Kingdom, North America, Western Europe, and Australia that looked at the operation of various local states in practice. Some of these concentrated on particular policy areas, such as housing or local economic development, while others studied the evolution and activities of the local state in particular places. A key message emerging from this work is that the nature and role of the local state are changing. The regulatory practices of Keynesianism, which relied on the local state to deliver collective public services as part of a nationally negotiated welfare settlement, have been replaced by neoliberal policies emphasizing lower welfare spending, competitiveness, and entrepreneurial governance. The abandonment of national redistributive strategies, and the emerging global mosaic of competitive regional economies, have led to the parallel mosaic of increasingly differentiated spaces of economic and social regulation. Within these new spaces of regulation, local state institutions are increasingly working in partnership with the private and voluntary sectors to promote globally competitive urban and regional spaces and to reorient their welfare provision away from a collective public services toward voluntarism, entrepreneurialism, and individual self-reliance.

SEE ALSO: Corporations and local states; Local development; Multilevel governance; Neoliberalism; Public policy; State, the; Urban politics; Urban regimes

References


Further reading

Local statistics and place-based analysis

Richard Harris  
*University of Bristol, UK*

Information such as census data, environmental records, and measures of neighborhood crime provide insight about the social and physical characteristics of different places, and about the people who live in them. Place-based analysis is about making sense of this information, through classifications of neighborhood types, by maps, charts, and other visualizations, and by the use of analytical techniques to highlight the prevalence of particular characteristics in some places more than others. Local statistics also look for geographical patterns and variations in a measurement at a subregional scale: for example clusters of diseases, hotspots of crime, or to consider whether the relationship between property price and the size and type of property varies according to where in a city that relationship is measured.

Some introductory examples

Consider a choropleth map showing a set of administrative areas located across a country where each area is shaded according to the local unemployment rate. Just looking at the map will reveal something that the national average unemployment rate does not: it shows variation at a subnational scale. Of course, that variation could be quantified in various ways: by determining the standard deviation of the (local) unemployment rates, for example, or by calculating the variance, the minimum and maximum values, or the interquartile range. However, the map offers something additional and fundamentally geographical: it tells us *where* areas of higher or lower unemployment are found. Knowing where something happens is an aid to understanding why it happens and is useful if we want to target appropriate policy initiatives to the areas of greatest need.

The map allows us to examine geographical patterns in the unemployment rate and to look for geographical differences. One way to make sense of the data is to divide the unemployment rates into quartiles, creating four map classes ranging from the lowest unemployment rates to the highest. The result would be a simple area-based classification, a rudimentary typology based on the prevailing level of unemployment in each place. More specifically, it would be a univariate classification, differentiating types of places on the basis of a single variable of measurement. We could then use the classification to look for associations with another variable. Imagine, for example, that the places with the highest unemployment rates contain 20% of the national population but 27% of a particular ethnic group. In this example, the ethnic group is disproportionately located in the high unemployment neighborhoods by a factor of 1.35 (27/20, sometimes multiplied by 100 to give an index value of 135).

We need to be careful how we interpret this. The co-location of the two occurrences (the ethnic group and unemployment) does not necessarily mean that members of the ethnic group are themselves unemployed. It is possible that they are working but living in areas of high
unemployment. To assume that a correlation between two attributes at a neighborhood scale necessarily applies to the individuals themselves is known as the ecological fallacy. Even so, the place-based analysis can be usefully suggestive; in many countries poverty rates do vary considerably by racial group and by neighborhood.

A second method of analyzing the data is to consider what the map shows for each area. It is, in fact, a local statistic – the unemployment rate at each location – where local means information for a specific part of the map as opposed to a statistic that applies to its whole. For simplicity, let us call these the local averages. We also have a global average, where global means not the whole world but the whole of the study region. Here it is the national unemployment rate. Subtracting the global average from each of the local averages tells us which places have high or low unemployment rates relative to the national benchmark.

Next we could map the places where the local average is highest (in the top quartile, for example) and see if they are regularly distributed across the country, randomly distributed, or, most likely in this case, clustered in some parts of the map more than others. In this way, we could study the map to look at the geographical outcome of social and economic forces that, in a market-based economy, sift and sort different people into different places. Note, however, that differentiating what we observe from what could be the outcome of an entirely random process is not always straightforward. This issue becomes especially salient in medical geography when, for example, an unusual cluster of mortality is statistically indistinguishable from a cluster arising due to chance. Early work in geocomputation (the application of computation to geographical analysis and modeling) considered a cluster of childhood leukemia cases around a nuclear processing facility. The fear was that the two were related. Yet, it could have been coincidence.

The analytical problem is that a random process can still generate what appear as clusters on a map. For example, Figure 1 shows a (simulated) cluster of cases to the bottom right and left of the map (shaded dark red), and an absence of cases elsewhere (shaded light yellow). Clearly the number of incidences is not the same across the map yet the numbers arise by chance: the 75 cases were placed on the map at random.

Third, we could use our choropleth map to look at how the unemployment rate in one area compares with those of other areas nearby. Doing so, we might discover a cluster of areas where above average unemployment is the norm. Alternatively, we might identify islands of high unemployment within regions of otherwise low unemployment. In fact, there are four possibilities: (i) areas of high employment surrounded by other areas of high employment (high–high, a “hotspot”); (ii) areas of low employment surrounded by other areas of low employment (low–low, a “cold spot”); (iii) areas of high employment surrounded by areas of low employment (high–low); and (iv) areas of low employment surrounded by areas of high employment (low–high).

Of the above, (i) and (ii) are examples of positive spatial autocorrelation, where similar values are found in closely located areas, and (iii) and (iv) are examples of negative autocorrelation, where dissimilar (or opposite) values are found together. Dependent upon your point of view, detecting spatial autocorrelation can be substantively interesting or a problem to be avoided. It is interesting where it is suggestive of a geographical process or outcome, or of a localized variation from the norm. Finding it invites the question, “What caused it?” However, it may also be regarded as unexplained variation, the presence of which violates statistical assumptions
of independence. If a map of regression residuals reveals patterns of spatial autocorrelation then those residuals are not independent of each other and that has consequences for how the results of the regression model should be interpreted. Specifically, it is likely that the standard errors for the estimated coefficients are underestimated, leading to an overconfidence in the statistical significance of the predictor variables (and suggesting methods of spatial regression should be used instead). The spatial autocorrelation may also point to the model missing an important explanatory variable, the model therefore being incomplete or incorrectly specified.

Figure 2 illustrates some of these ideas. Figure 2a shows the unemployment rate in and around London, UK, in 2011. The map makes clear that unemployment is highest in East London. Figure 2b reclassifies the map according to whether the local unemployment rate is in the top 25% nationally (highest unemployment), in the top 50% but not top 25% (above the median average), in the bottom 50% but not 25% (below average) or in the bottom 25% (lowest unemployment). The contrasting unemployment rates in London and the more rural areas to its south, west, and northeast are evident. Finally, Figure 2c shows how the local unemployment rates compare with those of surrounding areas. The concentration of higher unemployment rates in London is clear.

Geodemographic, neighborhood, and other classifications of places

Place-based classifications, especially at the finer scales of zip codes or census tracts, are usually referred to as neighborhood classifications, albeit with a rather administrative sense of neighborhood implied. Whether those neighborhoods also map to communities in any meaningful sense depends upon how and why they were designed. Some classifications are produced with a specific purpose in mind. For example, indices of deprivations have been developed in the United Kingdom and used to target regeneration funding to the poorest communities. Other classifications are designed for more general application and include many of the commercial classifications produced by multinational data analytical companies that are referred to as geodemographic classifications or as neighborhood segmentation tools. These have been used in applications as diverse as marketing, store location analysis, public service delivery, resource allocation, educational research, and medical research, amongst other fields of inquiry. Place-based classifications are not limited to social geography. Remote sensing can produce classifications of land use. Soil maps, ecological regions, biomes, and classifications of environmental risk all seek to characterize particular places in some way and to differentiate one type of place from another.
Figure 2  (a) The unemployment rate in and around London in 2011; the boundary of London is shown in green. (b) Showing how the local unemployment rates compare with the rates for all of England and Wales. (c) Identifying “hotspots” and “coldspots” of unemployment (see text for details).
As well as distinguishing between more specific and more generic classifications, we can also distinguish between univariate classifications based on a single variable (as in the introductory example) and those that are multivariate, drawing on a wide range of variables to produce the final classification. Some, like the United Kingdom’s aforementioned indices of deprivation, can be divided into a series of subdomains (income deprivation, employment deprivation, health deprivation and disability, education, skills, and training deprivation, crime deprivation, barriers to housing and services deprivation, and living environment deprivation). That is, there are actually seven different indices that can be chosen for specific tasks, as well as an overall index that combines them all. In this example, the classification is continuous – it assigns values to places but it is up to the analyst whether to split the data into distinct groups. The environmental performance index (EPI) uses twenty measures from national-level environmental data that can be combined into nine issue categories (including health impacts and air quality), into one of two overarching objectives (environmental health and ecosystem vitality) and into an overall index score. In 2014, Switzerland was ranked first of 178 countries, Somalia last (http://epi.yale.edu).

In the case of geodemographic products, they are multivariate classifications that assign neighborhoods into types. These are created by the vendor but the user usually has a choice of how many neighborhood types to analyze. Typically there will be a set of no more than ten different neighborhood types at the most aggregate level that then subdivides into a greater number of subclasses. For example, the UK’s Office of National Statistics (ONS) 2001 classification of census output areas had 8 clusters (neighborhood types) at its supergroup level, 13 at group level and 24 at a subgroup level.

It might be assumed that to use a greater number of classes is always better. However, the difficulty is that a large number soon becomes unwieldy in much the same way that it is hard to interpret a map with 20 map classes but relatively straightforward to interpret a map with four or five. The ONS classification mentioned here is of more than passing interest in that it was similar in construction to a commercial classification yet also an early example of “open geodemographics” – that is, a classification free to obtain and use, with the choice of variables and how the classification was created fully documented and accessible for scrutiny. The same cannot be said of commercial products, where that information is far more protected and the classification more of a black box. An open geodemographic classification has also been produced of the United States, and of London.

The origins of these neighborhood classifications are often attributed to Charles Booth and his 1889 Descriptive Map of London Poverty, the 1898–1899 revision of which can be viewed at http://booth.lse.ac.uk. From this, Booth developed a citywide classification of London’s streets where each was shaded on a base map to indicate the general socioeconomic conditions of the residents. For example, a black shading corresponded to Booth’s Class A, “the lowest class – occasional laborers, loafers, and semicriminals” (quoted in Harris, Sleight, and Webber 2005, 32). Dark blue was used to indicate Class B – “the very poor – casual labor, hand-to-mouth existence, chronic want” – and so forth. This classification was based on broadly qualitative information; specifically, school board records of families with children of school age. However, Booth also went on to devise a classification of London’s subdistricts using 1891 census data, creating a social index based on average percentage of poverty, percentage overcrowding,
birth rate, death rate, and rate of early marriage (Booth 1888, 1902–1903).

Booth’s method of classifying areas according to “type” is a precursor to modern geodemographics. His way of ranking neighborhoods according to domains of deprivation and then summing those to give an overall index foreshadows the UK government’s more recent indices of deprivation.

Viewed from the twenty-first century, the language (“occasional laborers, loafers and semicriminals” etc.) can seem paternalistic and patronizing or merely a product of its time. Unfortunately, the potential to stigmatize places and people remains in present day classifications and their use of labeling such as “claimant cultures,” “struggling estates,” or “shotguns & pickups.” The defense of such portrayals is that they create a vivid and instantly recognizable snapshot of a neighborhood. Critics say they are stereotyping. The classification produced by the UK ONS is again of interest here. No such portrayals were produced for it; to do so was judged inappropriate.

Charles Booth was pioneering in his mapping of area types and in his use of census data to create a multivariate social index. There were four further factors central to the development of neighborhood-based classifications. The first was the influence of the Chicago School of urban sociology and its interest in urban ecology, including the processes that led to the formation of what were regarded as natural neighborhoods – geographical areas within cities physically distinguishable from other surrounding areas and also socially distinct by the demographic and ethno-cultural characteristics of their populations. The second was the increased availability of small area data to measure neighborhoods and their populations, and the computational means to analyze them. In particular, the development of statistical techniques, including factor analysis and principal components analysis, which attempt to eliminate duplicated information in correlated variables and reduce them to their underlying factors or domains; also, methods of cluster analysis and agglomeration that are used to group small areas together on a like-with-like basis. Third, was the growth of geodemographics as a commercial industry from the late 1970s onwards. Fourth, is the periodic political interest in structural causes of poverty, deprivation, and inequality, the concern that these are entrenched and magnified by what people experience growing up in particular types of (low income) neighborhood, and the possibility of neighborhood-based policy solutions. It was political interest in the characteristics of neighborhoods, processes of neighborhood formation and decline, and the possibility of neighborhood-targeted regeneration funding that led to the indices of deprivation in the United Kingdom.

Today, a number of companies sell classifications intended to support place-based analysis. A retailer, for example, may notice that a disproportionate number of its customers reside in a given neighborhood type. Locating sites in similar neighborhoods would, therefore, make sense for future store openings. Alternatively, it may be found that instances of some health outcome, such as diabetes or obesity, are disproportionately associated with particular neighborhoods types. Those neighborhoods could be targeted for a health awareness campaign.

**Spatial analysis and local indicators of spatial association**

Despite the “geo” in its name, geodemographic and related types of neighborhood classification are not spatial methods of analysis. It is true that the input data are geographical – they are
information collected about people in places. It is true also that the results of the classification can be mapped because it is a classification of geographical areas. However, the central step, the methods that create the classification, is blind to geography. It takes the numbers and classifies them. It does not explicitly consider where the data were measured during the process of classification. Geographical data consist of two components, an attribute (the measured value) and a location (where it was measured). The processes creating the classification do not directly consider the second of these.

The same cannot be said of an analysis that compares a value at some location with the values for surrounding locations. Location is now important, specifically knowledge of the places that surround each observation. How we define surround is open to choice. For places represented by polygons on a map, it could be those that share a border. For places represented by points on a map, it could be those within a certain distance of each other. In either case, it could be the 10, 20, 50, 100, or any other number of nearest locations (provided that the number is an integer greater than zero and less than the total number of places measured in the study region).

Having settled on a definition, the places that surround each other — call them neighbors — can be recorded in a spatial weights matrix. If there are \( n \) places in total then this can take the form of an \( n \) by \( n \) matrix with a non-zero value at (row \( i \), column \( j \)) indicating that places \( i \) and \( j \) are neighbors. It is common for the matrix to be row-standardized, so that the sum of each row is equal to one. Row standardization prevents locations with large numbers of neighbors from dominating any subsequent analysis due to what would otherwise be the higher sum of their weights.

Recall the question of how the unemployment rate at one location compares with the rate at surrounding locations. One way of answering this is to consider whether closely located values are, in general, correlated with each other. For this a statistic known as Moran’s I can be used. Assuming we have a row-standardized spatial weights matrix, the formula for Moran’s I is:

\[
I = \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{1}
\]

where \( x_i \) is the unemployment rate at location (observation) \( i \), \( x_j \) is the unemployment rate at a neighboring location, and \( \bar{x} \) is the unemployment rate averaged across all locations.

There are three things to take from this formula. First, it provides a measure of spatial autocorrelation, usually assumed to range from \(-1\) (perfect negative spatial autocorrelation) to \(+1\) (perfect positive spatial autocorrelation) (this need not be correct: see de Jong, Sprenger, and van Veen 1984). Second, observe the presence of the spatial weights matrix, \( w_{ij} \). It is this that characterizes it as a spatial statistic. Third, note that the numerator (top line) of the formula can be rewritten as:

\[
\sum_j w_{1j} (x_1 - \bar{x})(x_j - \bar{x}) + \sum_j w_{2j} (x_2 - \bar{x})(x_j - \bar{x}) + \cdots + \sum_j w_{nj} (x_n - \bar{x})(x_j - \bar{x}) \tag{2}
\]

where the subscript 1 is for location 1, subscript 2 for location 2, and so on. The importance of this is that the final statistic is a function of what could be viewed as \( n \) local statistics summed together.

The idea that a global measure of spatial autocorrelation such as Moran’s I can be decomposed into a set of local indicators of spatial association (LISAs) was first developed by Anselin (1995). LISAs can be used in two ways: first, to look for local patterns of spatial correlation (e.g., hotspots) that are masked by or deviate from the
global average; second, as a diagnostic tool to investigate whether particular locations act as outliers distorting the global measure. LISAs are examples of localized statistics that measure local spatial clusters and identify spatial heterogeneity (meaning that what is present in one place is not the same as that present elsewhere). This broader class of statistics also includes Getis’ and Ord’s (1992) localized G statistics that are used to identify local clusters of low or high values and have been applied, for example, in research in ethnicity, looking at whether the enclaves occupied by minority ethnic groups have grown during the ten-year period between censuses in the United Kingdom. (They have, implying that the groups are not rooted to particular residences but are migrating outwards with a willingness to live amongst other ethnic groups and creating more mixed neighborhoods (Johnston, Poulsen, and Forrest 2015).)

Such statistics are important for highlighting geographical clusters and identifying spatial variations. Nevertheless, their use raises a number of issues. First, the centrality of the spatial weights matrix within their calculation means that if the matrix changes then so too will the result. Using a matrix where neighbors are defined as within 1 km of each other will not give the same answer as a matrix where neighbors are all locations less than 10 km apart. This suggests that, ideally, there ought to be a theoretical or research-related reason that guides the specification of the weights matrix. For example, if the interest is in whether there are spillover effects between regions that share a border, then this invites spatial weights based on contiguity. In other cases, it may be necessary to undertake a process of calibration, finding the distance threshold that gives the greatest average correlation between a location and its neighbors or, alternatively, the least average variation. Graphical methods including correlograms or variograms can help identify this threshold, plotting the amount of correlation or variation for various distances apart. Computational optimization procedures can also be used to minimize the variation or maximize the correlation.

Second, and as intimated earlier, establishing what is a statistically significant cluster of observed values is not entirely straightforward. There are two problems. First, a conventional statistical assumption that a test result is drawn from a Normal distribution where the observations generating the result are independent of each other may be either inappropriate or lead to the wrong values for determining significance in the presence of spatial autocorrelation. Second, if there are, say, 1000 observations located across the study region, and a local statistic is generated at each of them, then the same test has been repeated 1000 times. Imagine the null hypothesis states that there is no local spatial association and we seek to reject that hypothesis at a 95% confidence. The implication is that we will find about 50 false positives (5% of 1000), incorrectly rejecting the null hypothesis when we do.

A solution to the first of the above problems is to use a randomization procedure. This works by keeping the locations constant but randomly reallocating the data to them (such that the data are now in the wrong locations), before recalculating the local statistics. This is done over-and-over again to create a large set of values. If the local Moran’s value we are interested in is 0.30, then the idea is to determine for what proportion of the set a value of 0.30 or greater arises due only to a random process of allocation. If it is relatively often then the value is not particularly unusual given the geography of the study region and also the data.

In regard to the problem of multiple testing, a potential solution is to make a Bonferroni correction. If we have $\alpha = 0.05$ (a 95% confidence) but there are $n = 1000$ locations at which a local
statistic is generated then, arguably, $\alpha$ should be adjusted to $\alpha/n = 0.05/1000 = 0.00005$ (a 99.99995% confidence). However, as Anselin (1995) notes, this will reveal few if any significant clusters and is, in any case, unnecessarily conservative given that there are not really $n$ separate tests but some number less than $n$ because the neighbors at each location overlap.

Geographically weighted statistics

A final class of local statistics has grown in popularity in recent years. These are geographically weighted statistics. Tobler’s “first law of geography” (Tobler 1970) is often cited as the inspiration for the geographically weighted approach: everything is related to everything else but near things are more related than far things. It is not actually a law in any scientific sense but more a maxim about how patterns of positive spatial autocorrelation often are observed in geographical studies. If the maxim holds, then we can expect greater levels of spatial association between the closest observations in a data set and the spatial weights matrix should be formed accordingly.

With geographical weighting, those neighbors that are closest together receive a greater weighting than those that are further apart. This is described as inverse distance weighting – the greater the distance apart, the lower the weight given, reducing to zero beyond a threshold either chosen by the analyst or determined by a calibration/optimization procedure. As before, that threshold may be at a fixed distance (e.g., all observations within 10 km) or it could be at the distance to the $k$th nearest neighbor, a distance that will vary from one observation to another (unless it happens that the observations are all spaced out equally from each other). The latter is sometimes known as an adaptive approach in the sense of adapting to the local density of observations. It will take a larger distance to reach the $k$th nearest neighbor in parts of the study region where the observations are sparse and, conversely, a smaller distance where they are frequent.

Having defined the matrix, we can calculate a local and geographically weighted mean around observation $i$:

$$\bar{x} = \frac{\sum_{j} w_{ij} x_j}{\sum_{j} w_{ij}} \quad \text{(3)}$$

a local standard deviation:

$$s_i = \sqrt{\frac{\sum_{j} w_{ij} (x_j - \bar{x}_i)^2}{\sum_{j} w_{ij}}} \quad \text{(4)}$$

or even a geographically weighted LISA. These equations simplify if the weights are row-standardized (because in each case the denominator, the bottom line, becomes equal to one). The arguments in favor of row-standardization remain as previously. Note, however, that a consequence will be to disturb the relationship between distance apart and the weight given. For example, observations A and B could be the same distance apart as C and D but receive different weightings. That is because after row-standardization, the final weight of, say, B in relation to A, is not only a function of their distance apart but also the number and distance of other neighbors of A.

Figure 3 shows the geographically weighted (local) means for the unemployment data. They were calculated using a cut–off threshold of the 20th nearest neighbor. Comparing Figure 3 with Figure 2a it can be seen that calculating the geographically weighted means smooths the data – the more neighbors included in the calculation the greater the smoothing. In this regard it operates like a low-pass filter used in remote
sensing. Areas given a black border on the map are those where the local mean is significantly high (at a 95% confidence, determined using the sort of randomization procedure discussed earlier).

The idea of geographically weighted statistics has been taken forward within a regression framework. Geographically weighted regression works by comparing the results of a global regression model fitted to the whole study region to the set of results that are obtained if a local regression model is fitted at each individual location in turn, weighting that local regression to give most weight to the closest other observations. If there are $n$ observations, there are $n$ different regression models to compare with each other and with the global model. The interest is in whether the effects of the $X$ variables upon the $Y$ are broadly constant (the estimated beta values are approximately the same everywhere) or whether the relationships are nonstationary, meaning they vary in nature in different parts of the map. In matrix form, the formula for estimating the beta coefficients of ordinary least squared regression is:

$$\beta = [X^T X]^{-1} X^T y$$  \hspace{1cm} (5)

The equivalent for geographically weighted regression is:

$$\beta_i = [X^T W_i X]^{-1} X^T W_i y$$  \hspace{1cm} (6)

The detail of the mathematics need not concern us. Instead, observe the presence of the spatial weights matrices ($W$), which use inverse distance weighting, and note that what is obtained is a geographically localized set of estimates, as indicated by the subscript denoting location, $i$.

**Conclusion**

There is much to be learned from classifying and mapping data, by considering how the attributes of one place compare to those of its neighbors, by being able to establish where there are common characteristics amongst people, places, ecosystems, and environments, and by being
able to recognize and to understand where those characteristics diverge. Both local statistics and place-based analysis can be used to look for evidence of spatial clustering (shared characteristics of proximate observations) and also spatial heterogeneity (variations in those characteristics from one place to another). Local statistics, in particular, imply sensitivity to geographical difference and an expectation that what is measured in one location will not necessarily be measured in another. Identifying geographical difference is a first stage in establishing the causes of the difference.

SEE ALSO: Chicago School; Geocomputation; Geodemographic profiling; Geostatistics; Point pattern analysis; Quantitative methodologies; Spatial econometrics; Spatial weights

References


Further reading

Local/global production systems

Godfrey Yeung
National University of Singapore

Production systems (production chains) incorporate different production stages based on research and development (R&D), product design, the exploration of raw materials and minerals, and the manufacture of parts and components, which lead to the final assembly of a product before it is distributed and eventually sold to consumers. The networks of firms, the functions that they perform, and their relative power relationships in the various stages of production in different areas are some of the major focuses of research. The location of specific production processes can have a profound impact on the generation of economic value and local industrial development, as well as on interactions with the globalized world, for example the textile and clothing clusters in the Third Italy.

Local and global production systems

Local production systems focus on production stages conducted within a region while the global production system examines all stages of production across space. Manufacturing industries used to be integrated – either horizontally or vertically – to gain economy of scale and/or scope as well as the secure supply of parts and components. In horizontal integration, a firm incorporates pertinent industrial subsectors in-house, while a single firm (conglomerate) conducts all the processes, from R&D to in-house production, in the case of vertical integration.

The conglomerate mode of production is no longer financially viable in the globalized era when production costs in developed countries are rising and the location of manufacturing activities is becoming footloose, partly due to the international division of labor and the advancement of manufacturing and transportation technologies which allow different production processes to be located in different parts of the world. Under intense competition, conglomerates have to divest and focus on their core businesses through mergers and acquisitions (vertical disintegration). This leads to a massive industrial consolidation in manufacturing sectors, from iron and steel and shipbuilding to the pharmaceutical sector.

The divestment occurring in the chemical industry is a typical example. Traditional chemical conglomerates spanned all aspects of the business, from upstream oil exploration and production, through mid-stream ethylene and chlorine production, to downstream synthetic fibers, plastics, agrochemicals, and pharmaceuticals. The 1990s saw a revolution in these businesses, with most of the world’s leading firms espousing the breakup of the traditional chemical firm to focus instead on life sciences, with pharmaceuticals at their core. The establishment of Novartis from Ciba-Geigy and Sandoz is one such example. In the 1980s, Ciba-Geigy and Sandoz (both of Switzerland) were two of the world’s most powerful chemical companies, each particularly strong in speciality chemicals.
After a series of acquisitions and disinvestment in the mid-1990s, Ciba-Geigy was by then the world’s eighth largest chemical producer and ninth largest pharmaceutical manufacturer. To enhance the competitiveness and synergy of the chemical and pharmaceutical businesses, Ciba-Geigy and Sandoz radically changed the shape of the pharmaceutical industry through a simultaneous merger and demerger that arguably changed the corporate landscape of the world’s pharmaceutical and chemical industries in 1996. The main businesses of Ciba-Geigy and Sandoz combined in a US$27.5 billion merger to form the world’s second biggest pharmaceutical company, Novartis. As part of the transformation, Ciba-Geigy separated its slow-growing, speciality chemicals divisions into a new company, Ciba Speciality Chemicals.

Local-global linkages

World systems theory and dependency theory are grand paradigms for understanding the economic relationships between developed and developing countries; the former maps out such relationships and the latter highlights their power dynamics. These two theories do not fully recognize the importance of firms in the power dynamics. They also have no specific tool to unpack the relationships between actors and industries across space. Economic sociologist Gary Gereffi contributed to this research by developing a method to map out community chains in the 1990s.

Based on the asset specificity of transaction cost theory and the importance of trusts, social networks, power, and geographic proximity in the production network theory, Gereffi (1994, 2) introduced the concept of global commodity chains (GCCs), which incorporated “sets of inter-organizational networks clustered around one commodity or product, linking households, enterprises, and states to one another within the world-economy. These networks are situationally specific, socially constructed, and locally integrated, underscoring the social embeddedness of economic organization.”

He highlighted two specific types of commodity chains: producer-driven and buyer-driven commodity chains. Producer-driven commodity chains are characterized by capital- or technology-intensive manufacturers playing a central role in coordinating production networks, including their forward and backward linkages – computers, automobiles, and aircrafts – while large retailers, marketers, and branded manufacturers play pivotal roles in setting up decentralized production networks in labor-intensive consumer goods industries in buyer-driven commodity chains like footwear, garments, and toys. Gereffi, Humphrey, and Sturgeon (2005) further extended the concept to incorporate five specific governance types of global value chains (market-based, modular, relational, captive, and hierarchy value chains) to reconcile the asymmetrical relationships between transnational corporations (TNCs) in developed countries and their suppliers (local firms) in developing countries.

To highlight the nexus of the interconnected functions and operations of firms and non-firm institutions, where goods and services are produced and distributed globally, Coe et al. (2004, 471) proposed global production networks (GPNs) as a geographically attuned heuristic framework for the understanding of the global economy. This analytical approach highlights the spatial asymmetrical capture of value-added in various manufacturing activities across space as manufacturing activities, core knowledge, and technologies are still ultimately controlled by various TNCs and their first-tied
subcontractors. This reconciles the continuous divergence of growth between developed and developing countries despite the massive relocation of labor-intensive manufacturing industries to developing countries since the 1970s.

The GPN framework accounts for the possibility that selected local firms could reduce the manufacturing capability gaps between local producers and TNCs. Local original equipment manufacturers (OEMs) that sell their products to other companies (normally TNCs) to be marketed under the TNC brand could develop into dominant focal firms and spearhead the global (re-)organization of production networks through their corporate and marketing power as long as there is strategic coupling (linkages) with regional endowments and institutions. Focal firms are those with the capacity to operate major production functions, from R&D to manufacturing and assembly. Notwithstanding the analytical insights from relational economic geography in general, and the GPNs in particular, Sunley (2008, 8) highlighted the overemphasis placed on microscale processes and “ties and networks,” and the difficulty of validating the GPN framework empirically other than through qualitative case studies.

To catch up with TNC technological capacity in the knowledge-based economy, economic agents in developing countries have to absorb knowledge either through the internalization of codified knowledge, with the help of tacit knowledge acquired earlier, and/or through socialization that involves the exchange of tacit knowledge. This uncoded tacit knowledge is what Michael Storper (1995, 1997) termed geographically constrained “untraded interdependencies” where face-to-face exchanges, embedded routines, habits and norms, local conventions, and reciprocity and trust are crucial for its development. These interdependencies are not easily reproduced since they draw on the social properties of networked economic agents.

SEE ALSO: Global commodity/value chains; Global production networks; International division of labor; Knowledge-based economy; Vertical integration; World-systems theory

References


Further reading

LOCAL/GLOBAL PRODUCTION SYSTEMS


The International Encyclopedia of Geography
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0248

Localization/delocalization

Yu Zhou
Vassar College, USA

The term “localization” emerged as a response to globalization and thus delocalization. Since the 1960s, global production networks (GPNs) have become increasingly sophisticated and absorbed more and more regions into their networks of production, distribution, and consumption (Coe et al. 2004). GPNs are organized and led by top-tier transnational corporations (TNCs) in advanced economies to provide these companies with quick and low-cost accesses to resources, capabilities, and knowledge all over the world (Ernst and Kim 2002). The proliferation of such networks has not only transformed production processes but also profoundly shaped everyday lives and aspirations of peoples and places. For most parts of the world, localities can have little influence on or expect much accountability from TNCs headquartered far away. This transformation – referred to as “delocalization” – is represented by examples such as indigenous businesses being replaced by national or international chains, domestic factories becoming branch plants of TNCs, and local brands being eliminated by foreign brands. The fear is that, as TNCs standardize global practices, the world will eventually lose cultural diversity, and that power will be concentrated in the hands of the few (Sklair 2001).

It turns out, however, that this fear of homogeneity is overblown. Studies have found that TNCs do not just mold localities according to their desires, but also actively pursue localization by adapting capital investment, management personnel, product design, and marketing methods to localities in order to penetrate deeper into the local markets (Agarwal and Brem 2012; Coe and Lee 2013).

While management literature regards localization mainly as TNC strategies, geographers view localization as the vital means through which localities respond to the challenge of globalization. TNCs are only interested in selective local resources and opportunities. They are also footloose, and can quickly pull out of a region if the conditions become unfavorable. Localization from a regional perspective means that regions can develop local networks to broaden the spillover benefits of external resources and opportunities made available by GPNs. Regions can also deepen the embeddedness of TNCs in local economies to reduce their footloose tendencies. GPN studies, for example, advise regions to engage in “strategic coupling” with GPNs to articulate local resources into the global networks (Coe et al. 2004). Bathelt et al. (2004) use the terms “global pipeline” and “local buzz” to summarize the characteristics of innovative clusters that generate intense local exchanges and also build and maintain knowledge channels with innovative spots around the globe. In short, localization represents the regional strategies to create a codependency with GPNs for the purpose of increasing local leverage and autonomy in the global system. The coexistence of globalization and localization gave rise to the term “glocalization” (Valdivieso 2013).

SEE ALSO: Corporations and local states; Global production networks; Globalization;
LOCALIZATION/DELOCALIZATION

Local development; Local embeddedness; Local/global production systems

References


Location and multiplant firms

Isabel Mota
Università di Porto, Portugal
CEF.UP, Portugal

Economic agents – households and firms – have long tended to concentrate in a few places, cities or industrial clusters. But this raises the questions of what explains these locational outcomes, and whether the location choices of economics agents are influenced by their characteristics.

These questions are at the heart of location theory, which aims to explain the use of a finite resource – space – explicitly recognizing that economic activities consume space and are separated by costly distance. Location theory has a long historical tradition that goes back, for instance, to Adam Smith and David Ricardo, but interest in it has been cyclical. Following some remarkable developments, introduced by Von Thünen, Launhardt, Marshall, Weber, and Lösch, location theory has seen some less enthusiastic phases. Fortunately, in recent decades economists increasingly have come to recognize the relevance of space and location theory. It is possible to identify three main trends in economic geography modeling: the urban systems approach, which builds on externalities to explain the agglomeration of agents (Henderson); spatial competition models, which consider oligopolistic competition between firms (Hotelling); and the new economic geography approach, which assumes monopolistic competition among firms and increasing returns to scale (Krugman).

Notwithstanding this long history, it is surprising how little attention has been paid in location theory to studying the location choices of firms that set up multiplants (or multi-establishments or multistores). Typically, literature focuses on the location choices of single-plant firms. However, there are numerous examples of corporations that set up multiple stores in a city (e.g., Pizza Hut, McDonald’s, or Zara in a typical European city), multiple plants within a country (e.g., Cimpor has three plants (Cimpor 2015) in Portugal to produce cement; Talgo has two plants (Talgo 2015) in Spain to produce intercity, standard, and high-speed passenger trains), or even multiple plants across countries (e.g., the Volkswagen group operates 119 production plants (Volkswagen 2015) in 20 European countries and a further 11 countries in the Americas, Asia, and Africa; Bridgestone has 68 tire plants (Bridgestone 2015) in America, Europe, the Middle East, Africa, Russia, Asia and the Pacific, and Japan; Grohe has six plants (Grohe 2014) in Germany, Portugal, Thailand, and South Africa). Indeed, multiplant firms are not an exception, but rather a widespread feature of industrial economies.

Following Pal and Sarkar (2002), it may be argued that studying the location choices of a single-plant firm is merely a technical simplification, and that results for a single-plant firm can be extended to a multiplant firm. However, each plant’s decision influences the decisions of other plants in a corporate network, implying that the location choices of multiplant corporations are likely to differ from those of single-plant firms. The analysis of multiplant location choices may also be viewed as an extension to the theory of monopolistic price discrimination. Yet if the monopolist is not allowed to locate multiple
plants in different places, the ability to segregate the market is greatly restricted, and the price discrimination in question is in fact rather passive. As such, analysis based upon such a restricted setting has only limited implications.

The context in which single-plant firms and multiplant corporations make their location choices is dissimilar. While both search for the most profitable site, the relative importance of each location determinant differs for several reasons. First, new single-plant firms and new multiplant corporations are born at different phases of the life cycle of firms: new multiplants thus can take advantage of more mature entrepreneurship than new single-plant firms. Second, the birth of a multiplant corporation is frequently motivated by capacity needs or market expansion, implying that the optimization of multiplant locations justifies a distinct behavior. Third, branch plants of large firms are more sensitive to environmental regulation than plants in general. Fourth, given that location decisions involve incomplete information and uncertainty, we might expect multiplants to have access to more information about sites when they make their location choices. For example, multiplant corporations can benefit from economies of scale in conducting site searches. Multiplant corporations also can be less risk-averse than single-plant firms due to their presence in several markets, therefore idiosyncratic risk is null. Finally, multiplant corporations may be more flexible: the location of single-plant firms more likely reflects where the entrepreneur lives.

This entry discusses the location choices of multiplant corporations, examining the literature on this topic. In order to offer a broader overview of this research field, the location decisions of multistore and multi-establishment corporations are also included. The main purpose is to explore theoretical and empirical research focusing on location decisions. This is an important topic to review because, although there are many empirical examples, the scholarly literature on this topic is surprisingly sparse.

The second section discusses theoretical contributions to this topic, the third focuses on empirical studies, and the last concludes.

### Theoretical developments

There are three main trends in economic geography modeling: spatial competition models, new economic geography models, and externalities models.

Most research on the location of multiplant corporations builds on the spatial competition framework. This approach finds its roots in the seminal work of Hotelling (1929), who assumes that consumers are uniformly distributed along a bounded linear market with linear transportation costs. In this context, competition between two firms results in them clustering at the center of the market, a result known in the literature as the principle of minimum differentiation.

Hotelling’s paper catalyzed a large and rich literature on spatial competition. Usually, location models are classified into shopping or shipping models. A shopping (or mill pricing) model assumes that consumers bear the transportation costs, whereas a shipping (or delivered price) model assumes that firms deliver the product and pay for the transportation costs. Yet both models have the same centrifugal and centripetal forces, leading to similar location patterns under similar assumptions.

Spatial competition models may also encompass Bertrand (price) or Cournot (quantity) competition between firms. Under Bertrand competition, agglomeration only occurs if an additional differentiation dimension is allowed: for example, heterogeneity in consumers’ tastes
or brand specifications, or partial price collusion. Cournot competition has generated a relatively small body of location models, usually predicting that agglomeration equilibrium exists when population density is high. Recently, incomplete information has also been introduced as an explanation for spatial agglomeration under price competition by assuming, for example, asymmetric information about marginal production costs, product quality, or consumer preferences. Most work in spatial competition also utilizes the linear city model, although some analyzes location in Salop’s circular city and some introduces a triangular market.

Within this rich and diverse literature on spatial competition, only a few papers deal with multiplant (multistore or multi-establishment) corporate location choices. Teitz (1968) was the first author to study spatial competition among multistore firms. In the context of Hotelling’s linear city with linear transportation costs, he concludes that there is no location equilibrium if firms have multiple stores. Twenty years later, Martinez-Giralt and Neven (1988) considered both linear city and circular city models, assuming quadratic transportation costs (increasing with the square of distance) and mill pricing. They show that even if each firm is allowed to open two stores, its optimal choice in equilibrium is not to do so.

Most research on multiplant location emerged in the late 1990s. Some authors approach this issue by considering quantity competition. Chamorro-Rivas (2000) considers a circular market and a duopoly, with two plants per firm, concluding that in equilibrium all plants are equally spaced and paired around a circle. Li (2006) demonstrates that this result is also valid when one of the duopolists is a welfare-maximizing (public) firm. Pal and Sarkar (2002) study a multistore spatial duopoly in a linear city, finding that both agglomeration and dispersion location equilibria are possible. Extending their analysis to a circular city, Pal and Sarkar (2006) show that two firms with an equal number of plants result in a unique equilibrium location pattern, whereas multiple equilibria arise in most other scenarios. Yu and Lai (2003) study Cournot competition between firms producing goods that are either substitutes or complements. They show that a linear market always yields a central agglomeration, but that this result is not always valid for a circular market. For a circular market, they conclude that if the two goods are complements, there is a unique equilibrium where firms agglomerate at some market point; if the products are substitutes, there is a unique equilibrium where firms locate maximally far apart.

Other authors examine location choices on the basis of price competition. Assuming a two-firm two-plant setting, Pontes (2001) considers firms setting mill prices and making interdependent location decisions between two regions (or countries). He concludes that it is always a dominated strategy to locate a single plant in the smaller country. Dasci and Laporte (2004) make endogenous the number of stores opened by a monopolist through the introduction of a fixed cost, and conclude that the number of stores may decrease or increase with unit transportation cost. Janssen, Karamychev, and van Reeven (2005) incorporate consumer heterogeneity into competition between horizontally differentiated firms, and conclude that there is only one subgame perfect equilibrium where the location decisions of multistore firms are completely independent of each other. Karamychev and van Reeven (2009) adopt a random utility model to explore competition between multistore firms selling horizontally differentiated products and conclude that firms’ location strategies contribute to retail sprawl. Takaki and Matsubayashi (2013) focus on multistore sequential locations between two firms,
with fixed costs, which yield two opposing equilibrium location strategies for the leader: segmentation strategy (the leader monopolizes a market segment by partially deterring the follower’s entry) and equidistant strategy (in which stores are opened at equidistant locations throughout the market).

In the past two decades, new economic geography (NEG) models have become popular, emphasizing pecuniary externalities and merging these into a general equilibrium model of monopolistic competition. The seminal paper by Krugman (1991) combines the Dixit–Stiglitz model with an iceberg-type transport cost to examine the conditions under which a country can endogenously become differentiated into an industrialized “core” and an agricultural “periphery.” Extensions to the original model have introduced more realistic assumptions, including intersectoral rather than interregional labor mobility, positive transport costs for both the traditional and modern sectors, quasi-linear utility functions, and heterogeneous workers’ preferences. Although the details of the agglomeration process vary, there is general support for the emergence of a core–periphery structure for intermediate transportation costs with a tradeoff between transport costs and agglomeration. These conclusions agree with evidence of a secular fall in transportation costs coinciding with an intensified geographical concentration of economic activities.

There have been some attempts to model the spatial organization of multiplant firms within the NEG framework. The first was by Markusen and Venables (1998), although their results rely on numerical examples. The initial NEG model of Ekholm and Forslid (2001) also was rather limited. Neary (2002) focuses on the location choices of a multinational firm and identifies three distinct influences of internal trade liberalization: the tariff jumping that encourages plant consolidation; the export platform that favors foreign direct investment with only a single union plant relative to exporting; and reduced internal tariffs that increase competition from domestic firms. Fujita and Gokan (2005) developed a partial equilibrium model incorporating multiplant firms. Considering a two-region economy with declining communication costs, they conclude that firms producing low-trade-cost goods (such as electronics) tend to concentrate manufacturing plants in low-wage countries, while firms producing high-trade-cost products (such as automobiles) tend to open multiple plants serving segmented markets, even in the absence of wage differentials. Markusen (2002) offers an overview of partial and general equilibrium models with both horizontal and vertical foreign direct investment. He analyzes decisions about whether to build or acquire a foreign plant separately from decisions about where to raise the financing, assuming the interaction of scale economies, trade costs, factor endowments, and imperfect competition. Tsubota (2012) examines the outcomes of organizational choice between single-plant and multiplant strategies for two asymmetric regions. He concludes that a decrease in transport costs induces firms to agglomerate in one region, and also makes firms concentrate their production into a single plant. In addition, he observes that lower establishment costs favor the multiplant strategy, but when fixed export costs decrease, more firms choose single exporting.

Only a few studies consider multiplant corporations and the externalities used to explain urban systems (urbanization and localization economies). Using plant-level productivity data, Henderson (2003) finds that high-technology industries benefit more from localization externalities than do traditional manufacturing industries. He also finds that single-plant firms benefit more from localization economies than
do branch plants of multiplant corporations, which can rely more on corporate information networks. Moretti (1999) examines whether average educational attainment outside an industry affects plant level productivity, controlling for overall location fixed effects. He finds that a one-year increase in average education in the city increases plant productivity for single-plant firms by 7.7%, but has no effect for multiplant firms.

Overall, spatial competition models that assume price competition between multiplant firms focus on the optimal number of plants and mostly conclude that dispersion equilibria emerge. The models that assume quantity competition between firms and linear or circular markets frequently reach agglomeration or dispersion equilibria. Studies that use NEG to explain the location of multiplants typically conclude that a decrease in transportation costs favors agglomeration equilibrium and a single-plant structure. Finally, studies that focus on externalities to explain the location of multiplants conclude that both urbanization and localization economies are relevant for single plants and multiplants, but multiplants benefit more from urbanization economies.

Research on multiplant corporation location choices could particularly benefit from further study within the NEG and urban system approaches. It would be interesting to study differences between single-plant firms and multiplant corporations, in terms of how location profiles depend on different costs (transport, establishment, etc.), the types of good produced by each (final versus intermediate), the size of network externalities (i.e., how the value of a good or service changes with the number of consumers), or the qualifications of the workforce. The influence of multiplant corporation networks on urban systems also deserves further research.

**Empirical findings**

Empirical studies on the geography of multiplant firms are relatively scarce. Most research on the locational behavior of multiplant firms adopts a case study approach, employing survey methods to a specific sector. Case studies benefit from detailed information, but their results are difficult to generalize. Other researchers focus on small samples of firms in particular industries. Recently, however, there has been some research on the location choice of multiplant firms using large samples and econometric methods, allowing for more general conclusions. But econometric analysis faces several difficulties: location choices are not continuous and so standard regression specifications are not adequate; there is a whole range of possible locational determinants, requiring large datasets; and location choices are influenced by both local characteristics and neighborhood characteristics, requiring specific spatial econometric tools.

Early research on firms’ location choices, which assumed that firms were identical in terms of size, entrepreneur’s home base, and so on, has been challenged by authors claiming that the determinants of firms’ locations depend on, for example, firms’ size, industrial activity, or technological intensity (e.g., Arauzo–Carod and Manjón–Antolín 2004; Barrios, Görg, and Strobl 2006). There are few studies that focus on the influence of determinants of multiplant corporations’ location choices. Typically, they establish that the location profiles of multiplant and single-plant firms, and the influence of spatial variables on site selection, are different. It is observed that multiplant corporations are particularly sensitive to urbanization economies, land costs, and market size (Mota and Brandão 2013). Kim, Barkley, and Henry (2000), examining the influence of the number of plants
LOCATION AND MULTIPLANT FIRMS

on plant clustering, concludes that spatial clustering of establishments is positively related to industry average establishment size and labor intensity, but negatively related to multiplant structure and reliance on local product and input markets (Kim, Barkley, and Henry 2000). Other authors state that the clustering of industrial firms influences their birth and growth (Delgado, Porter, and Stern 2010), concluding that industries located in regions with strong clusters (i.e., a large presence of other related industries) experience higher growth in new business formation and start-up employment, while strong clusters are also associated with the formation of new establishments of existing firms, thus influencing the location decision of multi-establishment firms.

Some researchers study differences in the performance of multiplant and single-plant firms, frequently focusing on multinational firms. Usually, they establish that multiplant firms (e.g., multinational subsidiaries) have higher market shares and higher advertising and research and development ratios than single-plant firms (Anastassopoulos 2003).

There is a vast empirical literature on firm survival across different countries and time periods that identifies different variables (internal to firms, specific to the industry, specific to the business cycle, and geographical characteristics) influencing survival as well as different specifications for the survival function. New multiplant corporations are expected to have higher survival rates than new single-plant firms, since the former have better market information and can learn from other plants in the corporation. Multiplant corporations usually are larger firms, with fewer financial constraints; affiliates also can benefit from the potential technological and/or organizational superiority of the mother firm (Audretsch and Mahmood 1995; Mata and Portugal 2004).

A recurrent finding in the literature, however, is that plants owned by multinational multiplant corporations are more likely to close (Kneller et al. 2012). This may suggest that the main advantage of corporate networks of plants ultimately lies in the fact that they are bigger (Disney, Haskel, and Heden 2003). Researchers studying the location of functions within multiplant corporations conclude that proximity to corporate headquarters and to metropolitan areas influences a plant’s performance (Brunelle and Polese 2008; Toshihiro and Eiichi 2011; Cappariello, Federico, and Zizza 2012).

In short, much more empirical research into the location of multiplant firms is necessary, particularly into the influence of spatial variables – urban versus rural areas, agglomeration economies, fiscal policies – on the demography and dynamics of single-plant firms and multiplant corporations. It will also be important to develop large datasets and new tools for econometric spatial analysis that can deal with continuous space.

Conclusions

Notwithstanding the growth of both theoretical and empirical research on location economics, research in economic geography into the location of multiplant firms is sparse. With respect to theory, the majority examines spatial competition among firms, assuming either price or quantity competition, and either a linear or circular market. There are some studies examining multiplant corporations within general equilibrium models or the influence of multiplant corporate networks on urban systems. Empirical research is dominated by case studies or by recourse to small samples. There are attempts to develop large datasets and appropriate spatial analytical tools, but a great deal still needs to be done.
SEE ALSO: Firm foundation and growth; Industrial location theory; Localization/delocalization; Location-allocation analysis; New economic geography

References


Volkswagen. 2015. [http://www.volkswagenag.com/content/vwcorp/content/en/the_group.html](http://www.volkswagenag.com/content/vwcorp/content/en/the_group.html) (accessed April 26, 2016).

Location-allocation analysis is a process that develops mathematical location models, devises solution algorithms, and applies these methods to real-world decision-making situations that involve placement of facilities in a given space. Facility location decisions arise in a wide variety of public and private sector contexts. Examples of facilities whose locations have been analyzed in the location-allocation literature include schools, hospitals, libraries, post offices, ambulance stations, factories, retail outlets, warehouses, nuclear power plants, hydrogen refueling stations, and sensors for wireless networks.

Locations of these facilities greatly impact the level of services they provide, such as the accessibility of users to libraries and the distribution of market share of retail shops among competitors. Thus, constructing location models and solution algorithms that provide optimal locations of facilities is important. Location-allocation analysis provides a framework for the development of these location models and solution algorithms, and applying these methods to real-world situations is very important in providing successful services at facilities and understanding policy implications of locational decisions.

The modeling process is usually the first step of the location-allocation analysis, which involves identification of quantifiable objectives and constraints that depend on the locations of the facilities for the problem under study. An example of an objective is the minisum objective, which minimizes the demand-weighted total distance between users and facilities, where demands are given by the number of users at each location. Constraints define feasible solutions for location problems. For a factory, for example, these would include the number of facilities to be located, the budget available for locating facilities, and daily production capacity. The next step is to obtain optimal or desirable solutions to the facility location problem. Various algorithms have been proposed to solve location problems either optimally or heuristically. Finally, applications of location models involve collection of geographical data, analysis of solutions obtained, and analysis of policy implications for the problem under study.

Because of their rich set of applications, location models have been studied in such diverse fields as geography, regional science, urban economics, operations research, and applied mathematics. Each application has its own characteristics depending on the type of services provided at the facility. This has led location researchers to develop a large number of different models.

Types of location models

The basic elements of location problems are: (i) demand points, which are fixed in advance; (ii) facilities to be placed; and (iii) the space that
contains demand points and placed facilities. Demand points represent a set of locations where potential users of facilities exist, such as customers for retail shops and warehouses served by factories. Because facilities are small relative to the space in which they are sited, the problem can be modeled as finding a set of points in space.

There are a number of taxonomies to classify location problems. An important taxonomy focuses on the above three elements and classifies models into three types: continuous models, network models, and discrete models.

Continuous models typically assume that demand arises only at discrete points but that facilities can be located anywhere on a continuous plane. Objective functions are expressed as distances between users and facilities, where coordinates of facilities are decision variables. Since most distances are nonlinear functions of coordinates of facility locations (e.g., Euclidean distances), continuous location models are usually formulated as nonlinear optimization problems. Many of the continuous location problems are quite difficult to solve.

Networks are important modeling tools for describing transportation networks, such as road networks and railway networks. Network models assume a network composed of nodes and links. Both demand locations and facilities are restricted to the network, and travel between demands and facilities occurs only on a network. Network location models are further classified into those in which candidate locations are restricted among a set of nodes and those in which facilities can be located anywhere on the links as well as on the nodes. While demands are usually assumed to occur only at the nodes, some models assume continuously distributed demands along links of a network.

Discrete location problems assume that both demand locations and candidate facility sites are given by finite sets, with arbitrary distances between demands and facilities permitted. Discrete location problems are usually formulated as integer programming (IP) problems because location decisions can be represented by a set of binary variables that each indicate whether a facility is located at a candidate location or not. Often, variants and extensions of basic location models can be described as simple extensions of IP formulations of the basic models. Due to the recent development of various solution methods for IP problems and mathematical programming solvers, this approach becomes increasingly useful for a variety of problems.

To illustrate typical situations in which location-allocation models are applied, Figure 1 shows two simple solutions of a discrete location model. In the figure, circles represent both the sets of demand points and candidate facility locations. In general, these two sets may or may not be the same. The number of users residing at each demand point is the same for all locations. In both solutions, two facilities are located as shown by the black disks, and allocation of each demand to the nearest facility is illustrated by the solid lines together with user-facility distances. The sum of the distances for all users to the nearest facility in the solution on the left is 35 and that on the right is 26. Thus, in terms of the accessibility to the service provided at the facility, the right-hand solution is much more desirable than the left-hand one. As the example indicates, developing location models to provide optimal locational decisions is quite important.

Depending on the characteristics of each locational decision-making problem, goals and assumptions may differ greatly. For example, in some situations, instead of minimizing the total distance required for traveling to the nearest facility, minimizing the maximum distance
between all user-facility distances is more appropriate. In another case, the number of facilities is endogenous to the problem and is an output for the model rather than an input into the model considered in Figure 1. Another situation may require considering capacity limit for each facility, and each demand may not be allocated to the nearest facility. Just as real-world problems can be complex, so too can the problem formulation become complicated.

**Historical development of basic location-allocation models**

The historical development of mathematical facility location problems is described here and some of the most basic models are introduced. Location models have their roots in the pioneering work of Weber, who investigated an industrial location to minimize transport cost
LOCATION-ALLOCATION ANALYSIS

(Weber 1909). Weber considered the point that minimizes the sum of the transportation costs from two sources of materials to the facility location and the cost from there to a market. Today, the Weber problem is more generally formulated as follows: find a point in the plane that minimizes the weighted sum of the Euclidean distances from the point to a given number of destination points, where different destination points are associated with different amounts of demand, which determine weight. The objective function employed in this model is called the minisum objective because it minimizes the weighted sum of the distances. The Weber problem is a typical continuous location problem. This model can be solved numerically by an iterative method called the Weiszfeld algorithm (Weiszfeld 1937) or similar variants.

The classical Weber problem seeks to determine the location of only one facility. In many situations, however, locations of multiple facilities are simultaneously determined. There is a significant difference between these two cases. When only one facility is available, the facility provides service for all customers. However, when two or more facilities are to be introduced, decisions on which customer should be served by which facility should also be considered, as well as their locations. This process of assigning each demand to each facility is called allocation.

In the 1960s, more than half a century after the pioneering work of Weber, several important contributions were made that extended the classical Weber problem to deal with the location of multiple facilities. Among these, the multifacility Weber problem and the $p$-median problem are two of the most important location-allocation models, and these have had a significant impact on the development of later models. These two approaches employ the minisum objective (demand-weighted distance between demands and facilities), the same objective used in the Weber problem. This objective can be regarded as the maximization of accessibility when users access facilities directly and the minimization of total transport costs when facilities are source locations of products.

In the first of these models, Cooper (1963) considered the multifacility version of the classical Weber problem. As with the Weber problem, this problem assumes the Euclidean distance metric, known demand points in a plane, and a continuous planar region on which facilities can be located anywhere. The difference between this new formulation and the Weber problem is that the number of facilities to be located is generally two or more, which is why the problem is called the multifacility Weber problem. This problem seeks optimal locations of a given number of facilities and allocation of demands to specific facilities. Because of the strong nonlinearity of the objective, solving this problem is much harder than solving the original Weber one. Thus, heuristic methods are essential. The most well-known heuristic is the alternating heuristic presented by Cooper, which performs location improvement and allocation improvement phases alternately.

In the second of the multifacility extensions to the Weber model, Hakimi (1964) investigated the location of $p$ facilities so as to minimize the weighted distances between nodes on a network and the facilities. This can be seen as a network version of the multifacility Weber problem. This problem is perhaps the most well-known and important problem in location-allocation models; it is now usually called the $p$-median problem (PMP). Hakimi proved an interesting property: in any PMP, at least one optimal solution places the $p$ facility sites only on the network's nodes. Due to this property, which is often called Hakimi's theorem, the search
for optimal solutions can be reduced from an infinite set (i.e., every point on every link) to a finite set comprising only nodes of the network.

A mathematical programming formulation of the PMP was given by ReVelle and Swain (1970), which had a significant impact on the development of location-allocation models. They employed methods in mathematical programming such as integer programming formulation, linear programming (LP) relaxation, and branch and bound methods, whose approaches were later used by many researchers in the study of location models. Per Hakimi’s theorem, it is sufficient to search for the optimal solution on a network among the set of nodes. Thus, binary variables are introduced at each candidate node to site a facility or not.

The formulation of PMP is illustrated as follows. Let \( I \) be a set of demands points and \( J \) be a set of candidate locations on a network. The shortest path distance between demand point \( i \in I \) and candidate location \( j \in J \) is \( d_{ij} \). The weight \( w_i \) is associated with each demand location \( i \in I \). A binary decision variable \( x_j \) is defined at each candidate location \( j \in J \), taking a value of 1 if a facility is located at \( j \in J \) and 0 otherwise. Another binary decision variable \( y_{ij} \) is introduced that takes 1 when demand at \( i \in I \) is allocated to a candidate location at \( j \in J \). Using these notations, the PMP is formulated as follows:

| \( x_j \in \{0, 1\}, \forall j \in J \) (5) |
| \( y_{ij} \in \{0, 1\}, \forall i \in I, \forall j \in J \) (6) |

The objective function is the demand-weighted distance from each demand to facilities. Constraint (2) limits the number of facilities to be located to \( p \). Constraint (3) says each demand is allocated to exactly one facility. Constraints (4) states that a demand cannot be allocated to a given node unless a facility is located at the candidate site. Constraints (5) and (6) stipulate the integrality of variables. This formulation can also be applied to PMPs in which both demands and facilities are given by discrete points on a plane and arbitrary distances between these two sets are assumed.

The LP relaxation of the above problem, which results from replacing binary variables with continuous variables taking values between 0 and 1, often provides an all integer solution that is an optimal solution to the original problem. When one or more location variables take fractional values, branch and bound methods can be applied to obtain an exact optimal solution. The above formulation can be input into mathematical programming solvers directly together with input data for the model. Because of the recent development of such solvers, moderate sized problems can often be solved optimally. Daskin (2013) explains several basic methodological tools for solving discrete and network location models.

**Heuristics methods for \( p \)-median problems**

PMP is known to be \( \text{NP}-\text{hard} \), that is, there does not seem to exist any efficient solution method to solve all instances of the problem.
LOCATION-ALLOCATION ANALYSIS

This makes it difficult to obtain globally optimal solutions for large-scale problems. Thus, a number of heuristic algorithms have been proposed. The drawback of these approaches is that they find locally optimal solutions, which may or may not be globally optimal solutions. These approaches have been successfully applied to PMPs, especially for solving large problem instances. There are three fundamental classes of such heuristics for PMPs: greedy algorithms, neighborhood algorithms, and exchange heuristics. These approaches may also be used for obtaining solutions for variants of PMP and other types of location problems.

A greedy algorithm adds a facility at the best location at each step without considering the impact of the present decision on the final solution. This algorithm is a kind of a construction algorithm. At the $k$th step, with the $k-1$ locations obtained thus far taken as given, this method finds the best location for the $k$th facility that minimizes the total distance for $k$ facilities. After the $p$th step, $p$ facilities have been placed, and this is the final output of the method. The quality of solutions obtained by this method is not always good, but the algorithm is simple and easy to implement. Thus, the method is used to obtain initial solutions for neighborhood algorithms and exchange heuristics.

A neighborhood algorithm repeatedly attempts to improve a given solution. This algorithm is a kind of improvement algorithm. A neighborhood algorithm for PMP consists of two phases, a location improvement phase and an allocation improvement phase, and alternates between these steps until no further improvement is made. Given the locations of $p$ facilities, each demand node should be assigned to the nearest facility. This partitions the sets of nodes so that all nodes in each set are assigned to the same facility. These sets are called the neighborhood of the facility, and there are $p$ such neighborhoods. In each neighborhood, the facility should be located optimally. This can be done by solving the 1-median problem $p$ times by the method of exhaustion (i.e., checking the score for placement of the facility at each location within the neighborhood). This location improvement phase yields a new configuration of $p$ facilities (unless all locations were already optimal solutions to the 1-median problems). In the allocation improvement phase, each demand point is assigned to the nearest facility, and this creates new neighborhoods. These two steps are alternated until no further improvement is made.

Exchange heuristics are a third kind for solving PMPs. The neighborhood algorithm evaluates the impact of facility relocation only within its neighborhood; it does not consider potential improvement that might be achieved by relocating a facility to outside its neighborhood and reallocating demand. By lifting this limit, an alternative improvement algorithm, called an exchange algorithm, is created. Given the locations of $p$ facilities, the exchange algorithm evaluates the impact of relocating one of the $p$ facilities to any candidate node. In evaluating the impact of relocation that best reduces the total distance, the other $p-1$ facilities are fixed. There are several variants that vary according to how the relocation occurs. One common variant is to number each facility, from 1 to $p$, and relocate each facility to its best position in this order. When at least one of the $p$ facilities is relocated, the process is restarted. The algorithm continues until no further improvement can be obtained. Different initial solutions can lead to different locally optimal solutions. It is empirically accepted that the exchange heuristic is likely to produce a high-quality solution when a significant number of randomly chosen initial solutions are applied.
Variants and extensions

PMP has been successfully applied to a number of different planning contexts in both the private and the public sector. At the same time, the need to better describe various planning scenarios often requires modification of the basic model. Some typical extensions are introduced below.

For PMP, the number of facilities to be located is an input to the model. The fixed charge location problem is very similar to PMP, but a fixed cost of locating a facility at each candidate location is introduced. Instead of constraining the number of facilities to be located, the total cost of locating facilities is incorporated into the objective function. The objective function is given by the sum of the total transport costs from facilities to customers and the total facility construction cost. Thus, the optimal number of facilities is endogenously determined in the model. This type of modeling is often used in the private sector for location decisions when one organization both pays the costs and receives the benefits. For example, consider the situation in which a product maker decides the locations of warehouses from which goods will be delivered to retail shops. If the cost of opening a warehouse at each potential site and the cost of sending a unit amount of product from each potential site to retail shops is known, minimizing the construction cost of warehouses plus the total transport cost can be modeled as a fixed charge location problem. Depending on the transport costs, two typical locational configurations may be observed. When transport costs are high, many facilities will be selected, which will reduce the average transport distance. In contrast, when transport costs are low, a smaller number of facilities will be selected, which will reduce fixed charges.

Because PMP seeks to maximize global efficiency, accessibility of each user to the nearest facility often varies widely. Since weights at demand locations contribute proportionally to the minisum objective, areas of high population tend to receive significant benefits relative to areas of low population. As a result of this, less populous areas may be unacceptably far from the nearest facility. To reduce such variability, alternative constraints and objective functions have been suggested. For example, one simple idea is to introduce the requirement that the closest facility to each demand point must be within a specified distance. Another important approach that focuses on equity is the well-known \( p \)-center problem, which seeks to minimize the maximum of distances between each demand point and the facility nearest that point (Hakimi 1964). This objective function is called the minimax objective. Hakimi’s theorem does not apply to the network \( p \)-center problem, and therefore the following two cases should be clearly distinguished: in the first case, facilities can be located anywhere along the links of the network (the \( p \)-absolute center problem); in the second case, facilities are restricted to the nodes of the network (the vertex \( p \)-center problem). Solution methods for both types of problems have been developed.

In many location contexts, service to customers depends on the distance between the customers and the facility. Often, service is acceptable enough if the customer is within a given distance of the facility, and is not acceptable if the distance exceeds a given threshold. Covering location models address this situation. Although allocation is not as explicit in the basic covering problems as it is in the previously discussed problems, covering models are sometimes included among location-allocation models. Covering-based models assume that there is some critical coverage distance or time (e.g., 2 km or 10 min) within which demands must be met. Such models often apply to designing emergency services. For
example, it may be assumed that a given demand for fire-fighting services is covered when the nearest fire station is located within a 10-minute distance of households and businesses.

There are two main types of covering models. The first one is the set covering location problem that seeks to find the minimum number of facilities needed to satisfy all demands (Toregas et al. 1971). In practice, the solutions to the set covering location problem tend to be much larger than economically feasible due to the requirement that each potential demand point (e.g., customer location) must be covered. By using weights to reflect the number of customers at each demand point, the maximal covering location problem, the second type of covering problems, seeks to find the best locations for \( p \) facilities (Church and ReVelle 1974). This means maximizing the number of covered demands. These two basic covering models have been extended in a number of directions. For example, a definition of coverage may be changed so that a given demand is considered covered when at least two services are within a critical distance. This may often arise in emergency services when the nearest vehicle is busy attending other calls.

In PMP, it is assumed that each demand can be allocated to the nearest facility because of the minimization of the demand-weighted distance. In practice, facilities may have finite capacity to handle demands. This may represent, for example, a space limitation or production capacity at a factory. An important characteristic is that some demands may not be allocated to the nearest facility due to the capacity limit. Obtaining optimal solutions for capacitated PMPs is often much more difficult than for basic PMPs. Another well-known variation is the capacitated fixed charge location problem, in which each candidate for facility location is assigned a capacity along with a fixed charge.

A variety of solution algorithms have been proposed for this problem.

Many basic location models assume that all facilities being located are the same kind. Often, however, facilities are hierarchical in terms of the types of services they provide. An example is postal services. There are different levels of postal services: post offices at the lowest level, branch offices at the next level, and central post offices at the highest level. Health-care services and school systems are also examples of typical hierarchical systems. Often, higher-level facilities offer specialized services in addition to the services offered at lower-level facilities. In other cases, services are not replicated between levels. The former system is referred to as successively inclusive; the latter is called successively exclusive. A typical objective in this kind of problem is to find locations for each type of facility to best serve as many people as possible. Often, people can move between different levels of facilities. Thus, both the relative locations of facilities among levels and the locations of facilities within each level are important.

Many basic location models have a single objective, such as the total distance traveled to facilities or the maximum distance between facilities and demands. In reality, however, there may be two or more objectives, which should be considered simultaneously. Multi-objective approaches are commonly employed to deal with such situations. Often, the minisum solution and the minimax solution to a particular problem instance are quite different. Furthermore, the minisum solution is likely to rate poorly on the minimax objective and vice versa. However, there are very often many compromise solutions with good performance for both objectives, and so multi-objective approaches are quite useful in realizing desirable locations from various planning viewpoints. The Pareto optimal solution is an important notion in multi-objective
analysis, which characterizes good compromise solutions in multi-objective optimization problems. Solution S1 is dominated by solution S2 when solution S2 is as good as solution S1 in terms of each of the objectives and strictly better than solution S1 in terms of one or more objectives. When no solution exists that dominates solution S, solution S is a Pareto optimal solution. Obtaining some (or all) Pareto optimal solutions and displaying their performance in the objective space is quite an important approach in comparing different alternatives.

For many services provided at facilities, it is desirable to have them as near as possible to demand. There are, however, types of services in which remoteness should be considered. Waste disposal sites are such examples. Although these facilities are indispensable, most communities would like to have them as remotely located as possible. These types of facilities are called undesirable facilities. Other examples are nuclear power plants, hazardous waste dumps, and landfills. One of the common goals of undesirable facility location problems is maximizing a function of distances between facilities and demand locations. As examples, the maxisum objective seeks the maximization of the demand-weighted distance and the maximin seeks the maximization of the minimum distance between facilities and demand points. Usually, these types of problems have other conflicting objectives, and thus multi-objective approaches are commonly used to deal with the problems. For the example of waste disposal sites, the cost of transporting waste to the facility should also be considered. This cost is usually small when the facility is located near the population center. Thus, this type of explicit tradeoff should be properly modeled.

In many situations, some facilities are already placed, and the problem is to introduce one or more new facilities given the locations of existing facilities. An example of such models requires finding a given number of facilities such that the demand-weighted distance is minimized with the locations of existing facilities fixed. In some situations, relocations of existing facilities are also treated together with the introduction of new facilities. While a relocation policy increases the degrees of freedom, the cost of relocation cannot be ignored. Thus, maximum utilization of available resources and flexible relocation decisions should be balanced. This decision-making situation arises when demands for service change over time. An example is elementary school locations. When population changes occur, relocation and integration of existing facilities and construction of new schools should be considered. A class of competitive location problems is another topic requiring consideration of existing facilities. A basic problem formulation is to capture as much share as possible given the existing facilities of competitors.

**Modeling, data collections, applications**

To build adequate location-allocation models and to solve application instances, there are a number of things to consider. The following discusses some of them.

Usually, the first step in building location models is to identify quantifiable objectives that adequately describe the planning situation under investigation. For example, when planning ambulance station locations within a city, one possible objective function is the minisum objective, which seeks to minimize the total (or average) response time. Another objective may be to minimize variation in service among demands, thereby focusing on equity among people. When equity is more important than global efficiency, the minimax objective may be
an appropriate choice. The primary purpose of ambulance services is to save lives. This may lead to another objective of maximizing the number of people who can survive due to earlier access to medical services. This may, in turn, mean employing a nonlinear survival rate (as a function of access time to a hospital) for the objective function. Multi-objective models considering the above objectives are also very important.

Choosing an appropriate objective function (or functions) for the problem at hand is much more important than obtaining good quality solutions to a selected objective. Approximate solutions of well-formulated problems may have much more value than optimal solutions of problems that ignore important factors. It is important to check whether the model being formulated adequately reflects the real situation under study.

Allocation rules are also very important in location-allocation models. The simplest rule is to assign each demand to the nearest facility. However, when the facilities are capacitated, this allocation rule cannot be applied, because some facilities may have to handle more demands than they can cope with. Because of this, some demands may be allocated to facilities other than the closest facility. In some situations, each demand should be served by a single facility. For example, a retail store should be supplied by a single warehouse to operate effectively. Other situations require more complex allocation rules. In retail situations, good accessibility to a shop is only one of several important factors for customers choosing a particular shop. Often, attractiveness of the shop, measured by, for example, the size of the shop and the number of items at the shop, plays an important role. In medical emergency services, availability of ambulances should also be considered. The nearest ambulance depot may be busy attending to other calls. For such types of facilities, the location of not only the nearest depot but also those of the k-nearest depots should be considered. Other situations may require simultaneous deployment of two or more types of facilities to provide service for demands.

Collection of demand data is a very important factor in location-allocation analysis. Population data is typically used to represent demand. It should be noted that daytime population and nighttime population may differ, especially in urban areas. Thus, these measures should be explicitly distinguished and properly used according to the purpose of the analysis. When both nighttime population and daytime population are important, a bi-objective problem using two different inputs for a single objective function may be appropriate. Also, demand for services strongly depends on such demographic characteristics as age and income. For instance, demand for ambulance services depends on the age distribution, and this information should be considered when demands for ambulance services are estimated. When relocation and restructuring of elementary schools is considered, demands for services should consider not only the present number of elementary school students but also its future trend. As another example, when locations of stores are considered, income groups should be properly examined to estimate demand for services; spatial patterns of demands for luxury cars and groceries are generally different because the frequency of using the two stores differs by income group.

When constructing demand data, data are not always available for each potential demand point. In addition, it may not be computationally feasible to incorporate such detailed data for large-scale problems. Thus, depending on the purpose of the analysis, a level of detail adequate to represent the population of each community should be used. Often, aggregated population
data can be used to reduce the size or number of zones included in the model. Since aggregated data is only an approximation of real demand, there is an inherent tradeoff between accurate representation and amount of data; these should be adequately balanced. The proper level of aggregation may differ among types of facilities. One example is the locations of department stores and convenience stores. Each department store usually attracts customers from a wider area than any one convenience store. For locational decisions about convenience stores, detailed local population data may be quite important, but deciding the location of a department store may not require the same level of detail.

Aggregation of demand generates different kinds of errors that should be considered. For example, distances between facilities and demands are generally different after aggregation. For example, the nearest facility location for a given demand point in the original data may be the second or third nearest facility in the aggregated data. To apply location models to real-world decision-making, the extent and types of possible errors should be estimated, and proposed solutions should be carefully checked for acceptability.

A selection of candidate locations of facilities is an important issue in location-allocation analysis. An optimal solution of location-allocation models cannot be directly applied if the location selected in the solution cannot accommodate a facility. This situation is especially relevant for planar location problems, where a facility is assumed to be located anywhere on a continuous two-dimensional plane. The space required for a facility greatly affects the feasibility of sites. Facilities such as automated external defibrillators do not require much space; in contrast, a department store needs a lot. Regulations on land use and surrounding environments (such as zoning rules and environmental regulations) should also be considered in deciding feasible sites for facilities.

The method of measuring distances and travel times between two arbitrary points within a city is an important issue in location-allocation analysis. Sometimes, distances are measured using mathematical formulations, such as Euclidean distance (the $l^2$-norm) or Manhattan distance (the $l^1$-norm). Shortest path distances over a network are also commonly used. Often, travel times are more appropriate than travel distances. Using real-world data, travel times can be estimated by regression analysis for distances between various points over a network. Some more sophisticated models treat congestion of road networks and time-dependent link lengths (i.e., travel times by time of day). Uncertainty of travel times is also an important element in location-allocation models.

**Outlook**

Nearly half a century has passed since modern location models and methods were proposed by Cooper, Hakimi, ReVelle, and others. During this period, basic location models were modified and extended in a variety of directions, reflecting diverse planning situations. Such models include flow capturing (interception) problems, hub location problems, dispersion problems, and location-routing problems, to name only a few. In addition, various types of heuristic algorithms and exact solution methods have been proposed. These developments allowed researchers to introduce more complex factors into models, such as incorporation of the temporal dimension, stochastic elements, and facility failures and reliability issues, and so on.

Recent developments in the creation of mathematical optimization solvers and geographical
LOCATION-ALLOCATION ANALYSIS

information systems provide a strong framework for conducting applied location-allocation analysis. For example, by using an integer programming framework, students, researchers, and practitioners have easier access to location modeling and analysis of solutions. Integer programming solvers can be used to solve moderate-sized problems with location models formulated as integer programming problems. This approach does not usually require special knowledge about optimization techniques and solution algorithms. Another important factor is that various types of geographical data, such as population data and land-use data, are available to the public and can be used for various applications. Location-allocation analysis will continue to be an active research area for diverse disciplines and contribute to support decision-making in a variety of public and private sector problems.

SEE ALSO: Accessibility, in transportation planning; Central place theory; Demand and supply for producer services; Industrial location theory; Spatial optimization

References


Weiszfeld, Endre. 1937. “Sur le point pour lequel la somme des distances de \(n\) points donnés est minimum [On the point for which the sum of the distances to \(n\) given points is minimum].” Tohoku Mathematical Journal, 43: 355–386.

Further reading


Location-allocation models

Tom Storme
Frank Witlox
Ghent University, Belgium

Location-allocation (LA) models have been important and widely used instruments for location decisions for over half a century. A location decision typically involves substantial capital investment and the positioning of a facility has considerable consequences, so the decision should be based on firm ground (Badri 1999). A first major advantage of LA models is that they deal with multifacility location decisions, instead of evaluating sites individually (Ghosh and McLafferty 1987). This allows areas that are over- and underserved to be identified. Moreover, they handle two sets of questions: they not only address the question of optimizing location from the perspective of the facilities alone, but also aim to optimize the supply of a given set of demand points. They do so by allocating each demand point to a potential facility location and by continuously evaluating the resulting model in terms of a predetermined objective function. The objective functions maximize efficiency in a particular way. In doing so, the models keep the central place logic of more traditional location models but reject “the adoption of any particular spatial pattern of activities” (Rushton 1988, 99). As a result, and if applied well, LA models offer a valuable contribution to locational decision-making. In addition, the models are not only helpful for supporting location decisions, but they also can evaluate the quality and efficiency of current location patterns and/or diagnose viable alternatives for decision-makers, when conditions change (Rushton, 1988).

Common distinctions between LA models

LA models are mostly applied when planning the provision of public services or the location of development aid facilities (Rushton 1984; Møller-Jensen and Kofie 2001). Numerous examples can be found in the literature where LA models are used for site selection of emergency services, police departments, schools, post offices, museums, health care organizations, and the like within a given area (Murray 2010). However, the models can also be applied for determining the optimal location of private sector companies. For example, Goodchild (1984) applied the model for retail site selection, while Min and Melachrinoudis (2001) used the model in the banking sector. In fact, the often used distinction between public and private ownership is of minor importance, because what matters are the characteristics of the decision-making processes (Bach 1980). Another common distinction between LA models is based on a priori knowledge of potential selection sites. When there are few constraints in terms of potential locations and little a priori knowledge, the model determines the optimal locations across the complete study area, in a continuous plane across space (Wen and Iwamura 2008). Such analyses are more suitable for theoretical or macro-level analyses (Yeh and Chow 1996). In a particular (urban) context and built environment, there often are only a discrete number of candidate sites available, be it
LOCATION-ALLOCATION MODELS

already existing locations or new, potential sites. The model then searches for the best location among a given and discrete number of candidate sites, which is the type of LA model most often applied by spatial planners (Teixeira and Antunes 2008).

Predetermined objective functions

The optimal location pattern of facilities in LA models is subject to the prioritization of objectives or, in other words, the specific problem that has to be tackled. Several objective functions can be calculated. Three such objective functions are given as examples here.

One of the earlier and most popular LA models in operational research is the p-median model (Fotheringham, Densham, and Curtis 1995; Rahman and Smith, 2000), where p sites are chosen from a number of candidate sites to minimize the total weighted travel distance or time between solution facilities and demand points. This specific objective function starts from the assumption that supplier facilities offer a demand at fewer sites than there are points of demand and, as such, those facilities are located somewhere central to the demand points (Bach 1980). Given a set of demand points and transportation costs (per unit of distance, for example), this optimal central location can be determined. The p-median model ensures that the overall impedance for consumers/recipients is at its lowest. Such an objective function is particularly suitable to decide the optimal location of public services facilities, such as museums, schools, or libraries.

A second objective function aims at maximizing spatial coverage, or allocating as many demand points as possible to potential facilities. In the case of determining the optimal location of emergency services (e.g., emergency response service centers or fire departments), it is recommended that this objective function is applied, because it is often determined that these services need to be offered to all demand points within a specific response period. Other applications include, for example, the determination of the optimal location of bike-sharing stations (García-Palomares, Gutiérrez, and Latorre 2012).

Especially suitable for private sector companies are objective functions that maximize profits or market shares. For example, the Belgian case study of Dejonghe (2004) showed that such models can be used for constructing a market-oriented football league with professional clubs that can survive over a longer period of time. In addition, certain LA models take the location and service levels of competing facilities into account to obtain this objective.

Real-world location decisions, especially in an international setting, are nonetheless often complicated and require a well-informed trade-off between objectives. More information on structuring complex problems has been discussed by Badri (1999), who proposes to use hierarchical decision-making processes.

Input data and data challenges

Decision-making processes benefit from large amounts of data, which serve as an input to the algorithms. The models typically include information on elements such as:

1 the location, sets of demand and characteristics of demand points or “consumers,” for example, capacity and willingness to travel, specific consumer behavior;
2 the location, sets of objectives and characteristics of a company or “producer” of services and amenities. Characteristics
include, for example, a supply capacity. Information about current facilities or competing facilities can be included as well; operational and transportation costs, for example, accessibility by road/rail/air/water, travel and congestion times, parking facilities; spatial characteristics, which can include the availability of adequate sites and physical barriers.

To cope with these various and large sets of data, algorithms are often encompassed and run in a geographic information system (GIS) environment, which allows a relatively rapid evaluation of various scenarios.

LA models are also associated with a number of challenges, which can lead to suboptimal location solutions and can seriously cast doubt on the usefulness of the outcomes. It is both a strength and a weakness of the models that they rely on large amounts of data. With respect to the weaknesses, sufficient data are sometimes not easily available (e.g., data on competing facilities or facility capacities) and/or inaccurate or uncertain, which can lead to a considerable number of assumptions being made (e.g., all consumers have similar consumption behavior) or quite complex models (e.g., when applying the models in stochastic or fuzzy environments). Moreover, the modifiable areal unit problem (MAUP) problem, whereby demand points are aggregated in “central” and zonal demand points (e.g., when using census data), can seriously affect the outcomes as well (Goodchild 1979; Fotheringham, Densham, and Curtis 1995). Badri (1999) also stressed the difficulties of adding qualitative criteria to the model, for example, the political situation of a country when solving transnational, multifacility location decisions. It is, therefore, often necessary to perform sufficient research on the demand for facilities to ensure an adequate outcome.

**SEE ALSO:** Location-allocation analysis; Quantitative methodologies

**References**


LOCATION-ALLOCATION MODELS


Location-based games

Ola Ahlqvist
Ohio State University, USA

Games and play are part of human life, and in many games space plays a central role in determining the rules and interactions that are characteristic of each game. Location-based games, however, is a relatively new term that emerged from the increasing use of mobile and geospatial technologies in games that are situated in the real world.

Notions of play and games intersect, and many languages, for example German and French, use the same word for play and game (Salen and Zimmerman 2004). As such, games and play are complex concepts and have proven to be hard to define, but many efforts to do so seem to converge on at least some key elements. In an effort to give a comprehensive definition of games, Salen and Zimmerman (2004) propose the following definition: “A game is a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome.” Using this definition we can identify some key components of games. A game defined as a system means that there are parts (e.g., players, tokens, artifacts) and interactions between those parts. The system is then defined by the characteristics of those parts and the rules that govern how the parts can interact. The conflict or contest emerges from the system and a specification of how the outcome is quantified. As an example, the hide-and-seek game has a natural goal of finding all the participants who have tried to hide and the outcome then generates a conflict around how to best find and use the best locations for hiding. The quantified outcome can then be differently valued to determine, for example, if a player won or lost, or if no clear outcome is possible.

So far, this definition of games has avoided the idea of a game environment, that is, some type of framing or boundary of the game activity in both space and time. For a game like chess, the boundaries are defined by the game board of $8 \times 8$ squares and there are temporal limits, beginning with the start of the game and ending when one of the players wins or the game is tied. Traditionally, games take place either in the real world or in a represented world. Examples of the first category are tag, hide-and-seek, and football. Examples of represented games are chess, where the game board and pieces represent a generic battlefield and fighting troops, and Monopoly which in a similar fashion represents a generic city in which players, represented by tokens, move through and perform a variety of socioeconomic interactions.

Many popular games that are framed by a represented world include some form of a map. Historically, they go back several thousand years and often draw inspiration from war scenarios where players assume some role of an army leader with the goal of engaging in war-like interactions with the opponent in order to win the game. Only as late as the mid-nineteenth century did these map-based games take inspiration from and serve other purposes than warfare and conflict. Educational games and games drawing from current and noteworthy events started to permeate society and a variety of new games have continued to be produced to this day.
LOCATION-BASED GAMES

During the first half of the twentieth century, the industrial-military complex and the introduction of computers developed the foundations for what we know today as location-based games. With the advent of computer technology, many represented games could be translated into computer-based versions of existing real-world or represented games. Computers also allowed for entirely new games to be developed. Based on ideas from military simulations, other forms of simulated environments started to form the foundation for games like Space Invaders, and Pong.

Continued developments in computing and visualization technology allowed for increasingly sophisticated, realistic, and detailed renderings of various game worlds. Game developers continued to situate games in generic or artificial worlds, but as geographic information systems technologies have evolved and geographic data have become increasingly available, computer games have started to incorporate representations of real environments.

Still, games were traditionally either played in the real, physical world, or in a represented, sometimes digital and virtual reality. Only very recently have experimental computer games started to truly blend aspects of reality into a digital game system. However, the idea of combining the real and the represented and fictional is not new and some have argued that it dates back as far as *The Games at Abydos*, as described by Greek historian Herodotus (Ericsson 2004). While that example would probably be referred to as a live-action role-playing game today, contemporary media is increasingly blurring the boundary between fact and fiction, from Orson Welles’s *War of the Worlds* to current TV reality shows (Stenros, Montola, and Mäyrä 2007).

Recent developments and widespread access to mobile information, communication, and geospatial technologies have spurred a flurry of changes, including many variations of games that exploit these technical affordances. Because of their relatively recent nature, with first examples appearing around 2004–2005, there are many suggested terms for these games: pervasive games, ubiquitous games, augmented-, alternate-, and mixed-reality games, mobile games, geogames, and adatrionic games are just a few of the concepts proposed. Among them, location-based games were originally described as games that somehow involved a player’s relative or absolute location in the game play. Like many of the other terms listed above, the label draws from its enabling technology, location-based services for mobile devices.

Walther (2005) held location-based games to be a subgenre of the more overarching concept of pervasive games, which has been described as games that expand the spatial, temporal, and social boundaries of traditional games (Montola 2005; Stenros, Montola, and Mäyrä 2007), and these games are part of a gaming culture that “exist[s] in the intersection of phenomena such as city culture, mobile technology, network communication, reality fiction, and performing arts, combining bits and pieces from various contexts to produce new play experiences” (Montola, Stenros, and Waern 2009, 7).

Two early examples of location-based games are BotFighters and Pac-Manhattan. Players in Pac-Manhattan actually dress up as the yellow Pac-Man and the colored ghosts, run around the street grid of New York City, and while Pac-Man tries to collect virtual “dots” along the streets, the ghosts try to catch Pac-Man before all of the dots are collected. The game uses cell-phone communication, location-based services, and software to track players and broadcast their progress over the Internet for viewers from around the world (http://www.pacmanhattan.com/).
The BotFighters game also uses a portion of the real world as the game environment and mobile phones as a key communication device for game interaction and control. Players assume the role of robots that try to kill each other using text messages to find other robots, engage in battle using a variety of “weapons,” and receive messages about game progress. Integral to the game is the importance of location and relative distance between players to determine the outcome of battles.

Largely driven by continued developments in mobile, location-aware, information, and communication technologies, these and other games are increasingly blurring the boundaries between the game world and the real world. Because of this, it is often hard to draw a line between when a game is location-based and when it is not. Most observers would not consider a regular chess game as location-based, although spatial locations on the chess board are key to the game mechanics. Not even if we turn the board game version of chess into a large-scale, outdoor chess game would we think of it as location-based. Similarly, the game Pac-Man, in its most well-known form on a screen, would not be considered location-based until it is moved into the Pac-Manhattan context described above.

Despite a growing interest from academics to define the various pervasive game genres, including location-based games, they are elusive concepts that are constantly changing, and therefore hard to pin down with any one definition. So far the term has come to bear because of the need to label the types of games that are technically enabled by location aware devices, mobile technology, and information and communications technology in general. But this means that the definition would have to constantly change with changing technology. Better, then, is to characterize and identify location-based games using a set of key dimensions.

- **Location**: the degree to which the location of game entities, such as players, avatars, tokens, or other game objects determine the game system dynamics. Clearly, the location dimension is the first and foremost dimension to be satisfied for any game to be considered location-based. A location-based game, then, should be highly determined by spatial factors such as absolute location, distance, and connectivity. But this criterion alone would most likely include regular board games like chess in the category, and a few other dimensions are generally considered central to the broader category of pervasive games.

- **Spatial expansion**: the degree to which a game expands the game environment to larger spaces than a game board, a room, or a soccer field. Scale itself is not enough to be a determining factor but it helps to describe the spatiotemporal scope of a location-based game, and neighborhood-sized environments are common in existing examples. Another captivating aspect of spatial expansion is the ability to play a game anywhere and everywhere, even when that can introduce uncertainty about where the game space actually begins and ends.

- **Temporal expansion**: as with spatial expansion, location-based games often take on a less determined temporal scope and a common aspect of location-based games is their omnitemporal character; they continue even when you sleep or go to work.

- **Representation**: the way that the game represents (most often digitally) real space and time in an abstract information environment. This could take the form of a map of the neighborhood where the game is played such that
it allows for an overview of the game play. Represented game space and time can also be augmented, scaled, and warped to allow for the addition of virtual game items, and to show particular views, sequencing, jumps, and loops to support particular game rules and dynamics.

- **Pervasiveness**: the degree to which the game allows participants to move between real and represented environments. Wearable computers, head-mounted displays, sensor networks, and other pervasive computing technologies can allow the game to mix in with the real world such that the boundary between what is part of the game and what is not is blurred.

**SEE ALSO**: Augmented reality; Geographic information system; Geolocation services; Information and communications technology; Location-based services; Representation and presentation; Virtual geographic environments; Virtual reality

**References**


**Further reading**


Location-based services

Mohamed Bakillah
*Rupprecht-Karls-Universität, Germany*
*University of Calgary, Canada*

Steve H.L. Liang
*University of Calgary, Canada*

Alexander Zipf
*Rupprecht-Karls-Universität, Germany*

In 2010, a speaker at the Mobile World Congress emphasized that mobile smartphones were the “iconic” device of the moment, acting as an extension of yourself, aware of your location and displacements, and that will have an important impact on daily activities and life in general. Whether through smartphones or through the potentially more embedded devices of the future (which we may have difficulty visualizing today), location-based services (LBS), which are defined as wireless IP (Internet Protocol)-based services for mobile users that rely on geographic data, have definitely entered our reality. According to a recent survey conducted by the US-based Pew Research Internet Project, 74% of adult smartphone users said that they commonly use their phone to obtain information based on their location. Besides basic, well-known services such as discovering nearby restaurants or the whereabouts of a friend, LBS are quickly evolving toward more complex applications. The evolution of LBS is propelled by a unique characteristic: while LBS consume information to deliver it to their users, they also generate volumes of data. This self-contributing pattern today is mainly impeded by a lack of appropriate techniques and technologies to extract relevant information and knowledge from the data being generated, but it is only a matter of time before this issue will be addressed. As will be demonstrated in this entry, the research community from various disciplines, ranging from computer science, cognitive sciences, and law, to economics, geographic information systems (GIS), and sociology, is actively exploring the potential of LBS and the opportunities to improve it. In the meantime, other emerging paradigms such as so-called big data and the Internet of Things will converge with LBS to further stimulate their development. This entry reviews the historical development of LBS through an explanation of its initial enabling technologies, and provides an overview of the categories of existing LBS applications. Then, it analyzes the main impacts of LBS that have been discussed in the literature, whether at the economic or the social level. The current and future challenges related to the evolution of LBS are also discussed.

Historical development of LBS

A decade ago, official and commercial data producers and mapping agencies were virtually the only geospatial data producers. In addition, official and commercial data was available to experts, such as urban planning officials, but, in practice, not to “ordinary” citizens due to its high cost. In other words, geospatial data was data produced by experts for experts. Traditionally, the official or commercially produced data has been aggregated according to
LOCATION-BASED SERVICES

Figure 1  LBS-enabling technologies.

administrative units, which can be quite large (e.g., district-level). This constraint limits the possibility to conduct analysis and make decisions at a finer level of spatial precision. Moreover, traditional data update processes are lengthy, limiting access to up-to-date and real-time, or near-real-time, data.

Technological advances have changed the way in which data is being produced today, and have enabled the emergence, and ubiquity, of LBS. In the introduction to the 2013 review Progress in Location-Based Services, Gartner wrote that the first stage of LBS development was mainly driven by technology, as opposed to social, economic, or other considerations.

The technologies that have enabled the advent of LBS are illustrated in Figure 1. On the one hand, the core technologies underlying LBS include positioning techniques and technologies; broadband Internet access and web technologies that enable the sharing of location-based data; and mobile and web GIS. On the other hand, propelling technologies contribute to making LBS relevant and reliable: namely, the technologies, techniques, and paradigms that support the generation of publicly available location-based data; the increasing processing capacity that supports dealing with volumes of location-based data consumed and produced by LBS; the development of context awareness; and the standardization of LBS interfaces. Some of these enabling technologies are discussed in further detail below.

Positioning techniques and technologies

The evolution and availability of positioning technologies have played a major role in enabling
LBS. Various positioning techniques have been developed, with different costs and deployment scenarios:

**Global Positioning System (GPS).** GPS is a satellite-based localization system and probably the most well-known positioning technology. The majority of mobile devices are equipped with GPS chips that enable their position to be determined with relative accuracy. The uncertainty can be as little as 10 m or less (US Federal Communications Commission 2012). GPS can calculate the location of the device independently; the calculation is performed offline, which is an advantage when the connectivity of the device is limited or when privacy needs to be ensured for certain applications that are handling confidential information. However, a line of sight to at least four satellites is needed to calculate the device’s position. Therefore, it is unreliable in so-called “urban canyons,” as well as inside buildings. As a result, it is often used in conjunction with other positioning technologies.

**Network-based localization.** Wi-Fi technologies in mobile devices scan the surrounding area to find Wi-Fi networks, which include a wide variety of stations such as residential and business networks. The mobile device retrieves the coordinates of the Wi-Fi networks it found, and these coordinates are used to triangulate its current position. In contrast to the GPS system, network-based localization is generally conducted on a remote server operated by a third entity: for example, the Android localization system relies on Google’s Wi-Fi database, while iOS devices rely on Apple’s servers to compute coordinates based on incoming Wi-Fi signals. One main advantage of this technique is that it does not require building a costly infrastructure for the purpose of positioning itself, since it relies on existing networks.

**Cellular Sector/Base ID.** Cellular devices are constantly registering their presence with the nearest base station. The network operator has the exact location of each base station, so the cellular device’s location is computed through triangulation. The radius covered can vary by a significant amount, depending on the density of cells. For example, the coverage of a cell may vary from approximately 10 km to 30 km, depending on the urban density of the area. One of the drawbacks of this method is that since some networks for cellular devices have a quite wide coverage, the positioning may lack accuracy by a significant amount (e.g., more than the coverage of a cell).

**Other indirect positioning techniques.** Also worth mentioning are techniques to indirectly position a device through its user’s action. For example, check-in is an indirect way to determine one’s location and it is mostly used by applications that allow users to share their location with other users on the web. For a given place where the user has checked in (e.g., a restaurant), inverse geocoding can be used to retrieve the coordinates of the user at this time. Position can also be determined when a user scans a Near Field Communication (NFC) tag or a QR (Quick Response) code for which the location is known.

In addition, there are a number of hybrid techniques that combine two or more existing techniques, such as the Assisted Global Positioning System (A-GPS) technique, which refines GPS positioning with Cellular Sector/Base ID. While positioning techniques have reached a relatively satisfying level of spatial accuracy for most outdoor applications, today’s challenges with respect to positioning relate to indoor localization.
Concerning the mathematical models used for calculating a position, the most common ones are the following:

**Lateration:** the position of a point $A$ is calculated based on its distance to at least three other different points. The distance is usually calculated using signal runtime difference or received signal strength (RSS). The technique is also referred to as trilateration or multilateration, depending on the number of positions used.

**Triangulation:** the position of a point $A$ is calculated based on the angles $\alpha$ and $\beta$ between the segment that connects two points $B$ and $C$ and the segment that connects $A$ to $B$ and $A$ to $C$, respectively.

**Centroid localization:** the position of a point $A$ is calculated based on the location of origin of signals received by a mobile device. It is assumed that these locations are known. The position of $A$ corresponds to the centroid of a polygon formed with the origin of signals.

### Generation of publicly available location-based data

The other category of technologies that has enabled the advent of LBS is technologies that allow the collection of various types of data in a user’s surroundings, which can be called “location-based data.” Positioning technologies themselves are major tools in enabling the collection of such information. As today’s smartphones are equipped with positioning technologies, they allow so-called citizens-as-sensors, a term coined by Goodchild in 2007, to “sense” data about their environment and share it through collaborative web platforms or online social networks, generating huge volumes of data. Social network sites have been playing a significant role in the indirect generation of volumes of data, since they attract billions of users throughout the world. For instance, Facebook — probably the most well-known social network site — had more than 1.5 billion users in 2015. While social network sites are basically meant to support online social activities, due to their ability to localize users, the content generated through these sites is increasingly enriched with geolocated information. Also, location-based capacities of social network sites have been gradually improved to take advantage of this new trend, in addition to making some of this data available through Application Programming Interfaces (APIs). This has resulted in the labeling of these sites as geosocial or location-based social network (LBSN) sites (Roick and Heuser 2013). Examples of LBSN include Facebook Places, where users can visualize a list of nearby places and other points of interest (POIs) and can share their location with friends through check-ins, or Foursquare, where users are encouraged, through rewards, to check in frequently at the same place.

Emerging geosensor networks, which are increasingly accessible and cheaper to produce thanks to the miniaturization of sensing components and data storage, are also another major contributor of location-based data. In addition, data generated through geosensor networks are increasingly available online, especially due to the development of the concept of Sensor Web (Liang, Croitoru, and Tao 2005). The Sensor Web Enablement (SWE) initiative of the Open Geospatial Consortium (OGC) has contributed significantly to the online availability of sensor data through the development of interface and specification standards, such as the Observations and Measurements (O&M) data model and the Sensor Observation Service (SOS), to access sensor data through standardized protocols. Examples of Sensor Web platforms implementing SWE standards include the 52° North platform, mostly known in Europe, and the
GeoCENS platform, mostly known in North America. Such platforms can be leveraged to collect data on an LBS user’s environment, although concrete projects and experiments in this direction are still at an early stage due to limited ability to process and exploit real-time data streams, and to merge several semantically heterogeneous data streams.

Open data policies adopted by several governments, often put in place to foster economic opportunities, have also contributed to make official location-based data available to the public for location-aware services. While geospatial data have traditionally remained in the hands of experts (like governments and mapping agencies), paradigms such as open data, social media, and collaborative mapping projects make it possible for an increasing proportion of these data to be virtually available to anyone who has Internet access, with the potential to benefit businesses, civil society, and individuals in general (Bakillah et al. 2014a).

Context awareness

Context-aware applications take into account the physical environment or situation, and react proactively based on such knowledge. Research on context-aware applications started in the 1990s. Interestingly, several early context-aware systems were indoor services; in contrast, current LBS applications and associated context-aware services are mostly outdoor services, with a re-emerging trend to develop indoor services.

Context awareness technologies has played an important role in making LBS more intelligent and user-oriented (Anagnostopoulos, Tsounis, and Hadjiefthymiades 2004; Bakillah et al. 2014b). The notion of context as a component of ubiquitous and mobile computing has been studied for more than a decade, with some of the most prominent work being Dey’s (2001). Context is not an objective and comprehensive representation of reality, but rather an abstraction for which the degree of accuracy varies, and which may focus on different aspects of reality, depending on factors that include the users’ intentions and the perception of the real world. While Dey describes the context as “anything that can be used to characterize the situation of an entity,” there is no agreement on the definition of context, and a significant number of more specific definitions have been provided in relation to context awareness that depend on the application at hand. The elements that compose context can generally include the location of the user, his or her preferences, his or her neighbors in a social network, the most recent places visited and events that he or she attended, websites that he/she visited, and so on.

For the purpose of LBS, context can be acquired using two different approaches: explicitly, whereby the user manually specifies the context-related information, or implicitly, whereby technologies such as sensors are used to gather contextual information.

Standardization

Achieving the interoperability of LBS with various types of mobile devices and web services is critical for the potential of LBS to be fully exploited. Therefore, standardization has contributed and is likely to continue contributing to the development of LBS, even if existing standards are not necessarily widely adopted yet. More particularly, the OGC’s Open Location Services (OpenLS) initiative has been developing open specifications for LBS. The current official version 1.2 is dated 9 September 2008. The OpenLS core services are embedded in a service layer called the Geo Mobility Server (GMS), the role of which is to broker between the physical network and the service providers. LBS
LOCATION-BASED SERVICES

applications can be implemented either directly into the GMS or at the service providers. The OpenLS specification includes various interfaces that enable service providers to leverage existing services for their specific application. The interfaces, as described by the OGC specification, are the following.

- **Directory Service**: provides clients with access to a web directory that enables them to find a place, product, or service, often within a certain radius.
- **Gateway Service**: acts as an interface between the GMS and the Gateway Mobile Location Center (GMLC) or Mobile Positioning Center (MPC), which provides the position of the device.
- **Location Utility Service (Geocode/Reverse Geocode)**: when provided with a place name, street address, or postal code, this service returns geographic coordinates (geocoding). It also performs the inverse conversion (reverse geocoding).
- **Presentation Service**: this service enables the mobile terminal to display geographic information, such as the map of a selected area with selected data layers.
- **Route Service**: this determines a route between a starting point and a destination point. A navigation application is required to use this service.

The OpenLS standard specifications have been increasingly adopted in recent years, due in part to the efforts that were made to align them with other existing OGC standards. Still, more effort is required to increase their dissemination.

Categories of LBS applications

Today’s range of LBS applications is very diverse. The speed of innovation and market penetration has notably increased with the convergence of mobile communication, mobile Internet, Web 2.0 technologies, and new mobile devices such as smartphones, in addition to being driven by economic opportunities.

A categorization of LBS applications can help to visualize the spectrum of available services, but also services that are yet to come. Here, a categorization that is based on the applications’ purpose is offered (Figure 2).

![Categories of LBS applications](Image)
• *Monitoring and tracking applications* enable objects to be located in real time, often for business and management purposes, such as fleet management or parcel location. Applications also include systems that enable people, such as children and the elderly, to be tracked for safety purposes.

• *Entertainment applications* include location-based games such as real-life scavenger hunts.

• *Social networking applications*’ main purpose is to connect people online and enable them to share interests and location-related information, among others. Examples include applications where users can share places of interest through “checks-in,” such as Foursquare and Facebook Places. Social network sites that enable location-based information to be shared are also called “Location-Based Social Networks” (LBSN), and are reviewed in Roick and Heuser (2013). In their paper on the convergence of GIS and social media, Sui and Goodchild (2011) state that LBNS include check-in sites, social review sites, and social scheduling/events sites. LBNS are an important source of crowdsourced location-based data that could benefit a large variety of LBS.

• *Business applications* are meant to support and improve business processes and provide services such as targeted advertising based on a user’s location (e.g., PayPal Media Network, SkyHook, and go2 Media).

• *Information applications* are meant to deliver news and information that take into account the users’ location, and are often provided by traditional news services (e.g., The Weather Channel).

• *Traveling services* mainly include routing, navigation, and wayfinding services. Traditionally, the digitization of existing streets required to support routing and navigation services was performed by commercial services (such as Nokia’s HERE, formerly Navteq, and TomTom, formerly TeleAtlas). Later, with the emergence of Volunteered Geographic Information (VGI) platforms such as the OpenStreetMap (OSM) project, open and free routing services were developed. For example, the OpenRouteService.org uses OSM as a source of data. Also, initially available as car devices, an increasing number of applications are being developed for multiple transportation modes and pedestrians, as well as particular groups such as people with limited mobility (e.g., Rollstuhlrouting.de).

An analysis of the categories of LBS applications shows that at the moment, most of them offer services to individuals, but few support collaborative tasks that go beyond the activity of sharing data with no further commitment. In addition, most of them are based on a limited dataset and lack the capacity to use heterogeneous data that come from a variety of sources.

### Social and economic impacts of LBS

LBS have the potential to enhance people’s ability to make their environment more interactive and customized, their workplace more efficient, and their surroundings searchable, in addition to creating new economic opportunities. Still, LBS can also exacerbate social and environmental problems in an already crowded, polluted, and urbanized environment.

#### Economic impacts

Various sources emphasize that LBS have significant potential for economic growth. For example, Resch (2013) reported that in 2010,
the potential for economic opportunities generated by LBS was estimated to be US$20 billion. According to a 2012 report by the US Federal Communications Commission’s Wireless Telecommunications Bureau, revenues generated by LBS are expected to triple between 2012 and 2017. A review of location-aware applications from the iPhone, Blackberry, and Android Application Stores indicates that in July 2009, about 3300 LBS were offered, and that this number went up to about 7200 in February 2010 (Skyhook Wireless 2010). At the end of 2015 Foursquare’s website indicated that it had over 50 million users across the world and 8 billion check-ins, with several million new ones every day, and that more than 65 million companies were registered with them (https://fr.foursquare.com/about). With such numbers, it is likely that the LBS markets and opportunities will continue to grow in size and diversity.

Social impacts

As LBS are a relatively new technology that became available to the public only in the last decade, studies on their social impact, supported by experiments, for example, on analysis of social interactions through LBSN and analysis of human behavior in space and time as captured by LBS, are only starting to emerge. The main areas of concern relate to the impact of LBS on individual and collective behavior, privacy issues, and the threat or benefits associated with remote surveillance of people.

Regarding the impact of LBS on human behavior, in 2002, in *Smart Mobs: The Next Social Revolution*, Howard Rheingold predicted that mobile communication and instant access could result in disruptive social effects in various areas of life, ranging from political activism to intimate relationships and corporate management (Rheingold 2002). However, other authors, such as Miller, suggest that social media’s capacity to support many-to-many communication instead of the traditional one-to-one (through written correspondence, telephony) and one-to-many (through broadcasts and webpages) modes is also likely to foster collaborative behavior through the formation of groups (Miller 2013). This contrasts with the traditional social organization paradigm, where top-down, hierarchical organizations are created to support the achievement of a common, complex objective. As a result, scientists think that social media are likely to enable a new way of achieving complex objectives through technology-supported cooperation (Shirky 2008; Miller 2013). Still, it is not obvious that technologies such as social media and LBS will be sufficient to foster collaborative behavior. Several current problems are caused by behaviors that aim at optimizing an individual’s benefit but that, when adopted by a majority, lead to outcomes that are not beneficial for the society as a whole. Transportation is a common example of this type of problem: individuals use the shortest path from a starting point to a destination, as computed and recommended by some routing or navigation service. However, the large number of people relying on the same recommendation generates further congestion and pollution, which is the opposite of the expected result. LBS can contribute to this problem, unless LBS applications that take into account collective behaviors and outcomes are promoted and adopted by a majority. With regard to this concern, some LBS applications that aim to improve public transportation may be instrumental in achieving this goal – for example, the WaitLess bus tracking device, which allows users to locate buses. Other applications also reward users for environment-friendly behavior, like the TrafficPulse application (GeoSensor 2012). The TrafficPulse application provides subscribers with real-time updates on traffic conditions.
The application relies on GPS traces provided by volunteer smartphone users. Citizens are encouraged to provide updates through a reward system that emphasizes their contribution to reducing traffic. However, apart from such rare examples, as of today, most LBS applications are intended to benefit individuals first and foremost.

With regard to privacy and the remote surveillance of people, the tracking capacity of LBS obviously generates concerns regarding its potential use for monitoring individuals, as emphasized by Michael and Michael (2011). Some existing tracking services seem beneficial to their users, such as child safety location services that enable parents to track their children, or remote health monitoring services. LBS applications can also be used to make sure that people comply with certain rules; for example, employers can use them to supervise the displacement of their staff and verify if they have been engaged in work or in unpermitted activities. Even more controversial is the GPS Spouse Tracking service to check that partners are not being cheated on. Other emerging applications include vehicle monitoring devices in cars to assess how people are driving, which are used by insurance companies to evaluate risks and adjust prices accordingly. Authors argue that data gathered through LBS applications can be used to conduct profiling, since they can uncover a large spectrum of user patterns and behaviors. Concerns about the ability to track people through LBS are also exacerbated by the upcoming convergence between LBS and biometric recognition systems (remote iris scanners, facial recognition technologies, fingerprinting, etc.) (Michael and Michael 2011). Privacy concerns are therefore of utmost importance for the future development of LBS.

The disclosure of geographic information about individuals on the web can be a threat to personal privacy. A person’s location can be revealed in a variety of manners: someone can intentionally share their location with “friends” in the context of social networking activities, while an individual’s location can also be disclosed on the web by another user (Roick and Heuser 2013). The fact that users lose control of the information they share online, whether it relates to their location or their absence from a particular place, their identity, or their relationship with other users, is often itself considered a privacy threat.

Roick and Heuser (2013) report that users of location-based social networks experience difficulties in setting their privacy preservation rules. Researchers have attempted to address this issue, for example by developing systems that enable users to change their personal privacy settings when a location request is received (Xu et al. 2012). However, such a system is not yet common. Other strategies that were developed to protect privacy include cloaking algorithms, which generalize an individual’s current location (e.g., at a city level) or delay the publication of the individual’s location, and encryption protocols to prevent individuals and service providers from accessing geographic information on the user, as explained by Jonker, Mauw, and Pang (2012) in “Location-Based Services: Privacy, Security and Assurance.”

Current and future LBS-related research challenges

Developing LBS still poses several challenges to the research community in general and to the GIScience community in particular. The research challenges related to LBS include technological challenges for improving LBS themselves (core technologies), such as the need to develop better indoor positioning techniques, and research challenges related to the development of other
technologies that have an impact on LBS, such as big data and the Internet of Things.

Enabling indoor LBS

Indoor LBS is a relatively new trend, although it can be traced back to the early 1990s. Their relevance can be fully appreciated if we consider that, as reported by Li and Rizos (2014), human beings spend on average 90% of their time indoors, sometimes in complex environments like airports, large office buildings, huge shopping malls, university campuses, and so on. In addition, because they lose track of outdoor environmental markers, people lose orientation more quickly within buildings than in outdoor environments. As a result, at the moment, the most discussed applications of indoor LBS are routing, navigation, and wayfinding services, as Huang and Gartner (2010) reported in their survey of mobile indoor navigation systems. Indoor LBS are also needed for some groups of individuals, such as people with limited mobility, who must be able to find their way in indoor environments that are not always accessible (due to lack of access ramps, elevators, etc.). Providing indoor LBS requires very accurate positioning techniques, as the potential applications often have to detect in which room a user is located, for example.

While the accuracy of the GPS can reach approximately 4 m, which is suitable for the majority of applications, the GPS has major drawbacks for indoor positioning. The radio waves that are transmitted by satellites to mobile devices are mostly blocked by walls and roofs of buildings. As a result, other positioning techniques are required to enable indoor LBS. In their introduction to a 2014 special issue on indoor positioning and navigation published by the Journal of Location Based Services, Li and Rizos (2014) indicate that indoor positioning technologies can be classified into three categories:

- **Technologies based on pre-deployed signal transmission infrastructure**: the most common examples of such technologies include infrared signals, ultrasonic signals, and other radio frequency-based systems. Of note is that the first indoor positioning system was in this category: in 1992, Want and associates created the “active badge location system,” which used infrared emitters and receivers installed within a building to determine the position of people, who had to carry infrared emitting badges. A direct line of sight was required for the system to function, which is true for more recent infrared systems as well. Of note as well is that the latter system was able to achieve a 16-cm accuracy, but only within an area of 100 m². More recent technologies also include Bluetooth, which can be used to determine in which room a device is located and can achieve an accuracy of 98% for this task. The major drawback of this approach, however, is that the mobile device must not move for 3 minutes. This is hardly suitable for most of today’s contexts. Overall, while such a system can achieve good accuracy, systems that are based on a local infrastructure are costly to deploy over large areas.

- **Technologies based on signals of opportunity**: these systems exploit radio frequency signals and infrastructures that are not necessarily dedicated to positioning. Examples of such signals include Wi-Fi and mobile telephony signals. Wi-Fi in particular is seen as an interesting indoor positioning avenue as its wide use has turned it into a common standard today. For example, indoor positioning techniques based on WiFiLoc, which relies on the signal distribution of the widely used Institute of Electrical and Electronics Engineers (IEEE) 802.11 Wi-Fi networks, were developed. Such a system
has the capacity to position devices within rooms of a building. Still, for some indoor applications, such as navigation services for people with limited mobility, a higher level of accuracy is required. As another example, Laoudias, Piché, and Panayiotou (2013) state that Received Signal Strength (RSS) fingerprinting with wireless local area network (WLAN) is an interesting alternative for indoor positioning because WLAN access points are widely available and WLAN-enabled mobile devices can easily monitor RSS measurements. Fingerprinting is a technique where a reference device is employed to generate a radiomap within a given area; one drawback of this technique is that the positioning accuracy can be affected if the user has a different type of device than the one that was used for generating the radiomap (Laoudias, Piché, and Panayiotou 2013). Ongoing research aims at improving the accuracy of fingerprinting through calibration methods.

- **Technologies not based on signals**: these technologies are independent of any infrastructure and use embarked sensors to locate a device. Common approaches are based on inertial sensors, such as accelerometers and gyroscopes that measure changes in distance and speed, or camera systems. An active area of research is based on the improvement of the Inertial Navigation System (INS), which is continually adding changes detected by embarked sensors to the previously calculated position. The main drawbacks of this technique include many sources of errors, such as the fact that the accuracy of the computed position depends on that of the initial position and direction; the noisiness of accelerometer measurements; the angular deviation over the traveled distance (known as the Abbe error); and the uncertainty of the step length. Among the solutions proposed to improve accuracy, “map matching” is currently an active area of research. The idea of this technique is to compare the measured movement trajectory of the monitored device with the geometrical constraints indicated on an (indoor) map. An example of a recent approach within this area is the work of Kaiser, Khider, and Robertson (2013). Another technology in this category of positioning technique gave birth to so-called vision-based localization and navigation systems, where the comparison of images captured by an embarked camera with a pre-existing collection of images or maps enables the device carrying the camera to be located. A recent example of such an approach is given in Li and Wang (2014). As one can expect, this technique is computationally demanding.

In addition, a certain number of hybrid approaches exist that combine technologies from these different categories. Hybrid approaches are employed since, currently, there is no overall standalone solution to indoor positioning that would balance cost, coverage, and accuracy in a satisfying manner.

Another challenge related to the development of indoor LBS is the lack of solutions for indoor mapping. In particular, floor plans are usually not made available for open applications. The solutions being explored include the generation of indoor maps through the reconstitution of trajectories within buildings.

**Extracting implicit location-based information and knowledge**

As the huge volumes of data produced through mobile devices and disseminated through the web is an emerging, relatively recent phenomenon, research on how these data could be
exploited and used to improve LBS is still at an early stage. Nevertheless, LBS will become increasingly relevant to a large spectrum of activities when mature technologies to extract knowledge from data become available.

Nowadays, the research on extraction of implicit information focuses on the geolocation of features. Such techniques are needed since an explicit geographic reference is not always available for all features. For example, a message posted on a social media platform, such as a review about a specific venue or event, may not be associated with actual coordinates or a nameplace that can be automatically interpreted and processed. This limits the ability of location-aware services to use this information. Techniques that are currently being investigated to identify the geolocation of features include various text analysis and processing techniques, such as Named Entity Recognition, which enables the identification of a city, street names, or names of buildings within a text, such as in Twitter messages. It is worth noting that more research is needed in this area, as studies have generally resulted in low accuracy.

Besides text messages, keywords associated with photographs (for example, in Flickr) can also be used to extract location names, including colloquial names of geographical areas. Another related technique is to model the probability that a group of tags is associated with a location. Social media feeds can also be gathered and analyzed in order to locate social hotspots or to map social networks in a given geographical area.

The geolocation of users whose position is not made available is another active research field. For this purpose, researchers have employed similar techniques to those being used for geolocating features in general, such as using probabilistic methods applied on social media postings. Still, in this experiment, only 50% of users were located accurately. “Friends” in social networks like Facebook can also be relied upon to determine the location of users.

Besides geolocation, other efforts are being made to extract various types of location-based information from data, which we can refer to as “situational information.” Most of these efforts focus on geosocial networks. For example, situational information can be extracted from social media postings to coordinate responses to emergencies, while tourists’ preferences can be uncovered from pictures they have posted on social media, using data mining techniques. Experiments to extract the structure of geolocated events and patterns of real-world activity were also conducted with microblogging sites like Twitter. A well-known example of recent research is the use of social media to monitor natural disasters, where algorithms have been developed to calculate earthquakes’ epicenters and typhoons’ trajectories based on Twitter data. Twitter was also used to monitor the occurrence of particular diseases at a given location and detect the location of outbreaks, or predict the time and place of outbreaks. Landmarks and events can also be inferred from geotagged photo collections. Spatiotemporal analysis of check-in data can also reveal information on categories of POIs, based on a certain number of assumptions, such as the time of day when the number of visitors at a given type of venue is likely to be higher. Of note is that several of these approaches rely on domain-specific assumptions, and therefore are not readily applicable to other problems. In addition, they are mostly based on a single source of data, and do not tackle the problem of merging data from several sources.

Basically, due to these current limitations, most LBS applications use simple data of one type at a time – for example, finding POIs of a certain category within a given distance from the user – but the capacity to rapidly merge heterogeneous data from multiple sources in
order to generate value-added knowledge is still very limited, despite considerable research work, since current data conflation approaches include several steps that have to be performed and validated manually.

**LBS and “big data”**

Recent technological advances in data production and dissemination have enabled the generation of unprecedented volumes of geospatial data, giving rise to the concept of “big data.” In 2012, it was reported in *The Age of Big Data* that the global volume of data was growing at a rate of 50% per year (Lohr 2012), due to the increasing dissemination of digital sensors, smartphones, GPS-enabled devices, crowdsourcing applications, and social media, among others. VGI, that is, geographic information produced by volunteer contributors (Goodchild 2007), also contributes to big data, and is often produced through LBS.

“Big data” is a loosely used term that refers to either the huge volume of data itself, which no longer fits traditional database structures, or, perhaps less frequently, to the set of techniques that are being developed to deal with such volumes of data.

As for many applications, big data both raise challenges and open up new opportunities. Indeed, the possibility of exploiting massive datasets raises new opportunities to improve LBS with even more data on users’ environments. While traditional data storage and management approaches are not suitable for dealing with big data, the range of techniques and technologies dedicated to big data is already wide.

Meanwhile, storage technologies are reported to have significantly improved, and storage issues have largely been addressed through cloud-based platforms. Rather, major difficulties are related to the extraction of relevant information from massive datasets. Traditional data mining algorithms, such as spatial clustering techniques or classification algorithms, are inefficient when dealing with massive datasets. As a result, as reported by Russom (2011), ongoing research is being conducted to improve existing knowledge extraction techniques, such as Structured Query Language (SQL) queries, data mining, statistical analysis, clustering, natural language processing, text analytics, and artificial intelligence, so that they can deal with massive datasets. These analytic techniques, among others, could be deployed to improve social media analysis, change detection algorithms for high-resolution images, and improve complex object recognition, which would all help to enrich LBS. In particular, the ability to deal with massive datasets is supported by underlying technologies such as Google’s MapReduce, a programming model that supports the development of scalable parallel applications for big data. MapReduce is based on a distributed file system where data is represented according to key-value pairs. It is then partitioned in several machines before being processed. An increasingly popular open source implementation of MapReduce is Hadoop, which is now considered by some as the de facto standard in industry and academia. MapReduce can enable data mining algorithms such as clustering, frequent pattern mining, classifiers, and graph analysis to be parallelized to deal with massive datasets. Research is being conducted to further improve MapReduce, such as developing methods to improve join algorithms and indexing techniques. Research on big data could be highly beneficial for feeding LBS applications with more relevant knowledge on their users’ environments, such as data on road conditions to support efficient routing services.
Dealing with data quality

Data quality, while being a concern for most applications, is a special concern for LBS because an increasing part of the data used by LBS applications comes from volunteers who contribute through crowdsourcing applications and social media. Researchers have highlighted the value of crowdsourced data to enrich or update existing geospatial datasets (Goodchild 2007). This advantage is especially useful considering that traditional geospatial data producers may lack the capacity to generate datasets with comprehensive coverage and level of detail. In addition, the fact that crowdsourced data can be provided and disseminated in a timely, near real-time fashion suggests that it can help to provide a dynamic picture of the environment. Still, problems concerning the credibility, reliability, and quality of VGI and crowdsourced data in general have been reported by several researchers, as indicated by Elwood (2008) in a review of future directions for research on VGI. The lack of quality affects the ability to interpret data, to extract relevant knowledge from it, and to integrate it with other data, among other problems (Bakillah et al. 2013).

Crowdsourced data may lack credibility and reliability because it is produced in a context that differs from the “structured institution-initiated and expert-driven contexts” (Elwood 2008). For instance, in contrast with official or commercial data producers, who are expected to generate data with a certain level of accuracy, volunteer contributors are not obligated to do so. In addition, there are no standard formats and languages for collecting crowdsourced data and formalizing the context around it (Bakillah et al. 2013). Crowdsourced data is produced using natural language, rather than the agreed-upon terminology often employed by professional data producers. As a result, the terminology used by contributors is ambiguous. The lack of standardization, combined with the diversity of users’ experiences, suggests that the heterogeneity affecting crowdsourced data is more problematic than that of traditional geospatial data. In addition, the profile and motivation of contributors are often unknown. Still, while this remains to be thoroughly demonstrated, one can expect that the socioeconomic, sociological, and cultural background of contributors could have an impact on the quality of contributions. Being aware of the relevant characteristics of contributors could help to assess the quality of crowdsourced data.

A common method for assessing data quality is to compare it with reference data, whether commercial or administrative datasets. With respect to the quality of VGI, such an attempt has been made, for example, by Haklay (2010), who compared OSM data with ordnance survey datasets. However, accessibility to high-quality reference may be limited for crowdsourced data since, as mentioned, its coverage may be different from official data. Alternative solutions include methods for measuring intrinsic data quality indicators based on an analysis of the history of the data. Another approach consists of classifying contributors into categories such as “beginners” or “experts,” and using this categorization in the assessment of contributions. When the profile of a contributor is characterized, this contributor can be assigned a “level of trust” that can help predict the quality of his or her future contribution. Systems where contributors review each other’s contributions have also been developed and coupled with learning algorithms to improve their relevance. A contrasting paradigm was adopted by Foody (2013), where the focus is on extracting relevant information. Foody provides an alternative quality assessment method where crowdsourced contributions are considered as “imperfect indicators of the actual phenomenon under study”; a latent class analysis method is employed to extract underlying
information and to estimate the quality of the contributions.

“Preventive” approaches can also be implemented to try to prevent quality issues from arising by controlling the data production process. Some crowdsourcing applications use editors with templates to control data production. However, other authors report that a clear correlation between data accuracy and the use of editors has not been observed. For example, Yaari, Baruchson-Arbib, and Bar-Ilan (2011) have conducted a study that supports this claim, but it was conducted in Wikipedia. Further research is therefore needed to establish whether this statement holds, and if it holds specifically for VGI. Indeed, editors supporting the production of VGI must differ from those that could be used in Wikipedia. In addition, difficulties arise because of the need to reach a balance between the freedom and flexibility that is given to contributors and the necessity to improve the data production process. Therefore, further research is needed to determine how data production and control mechanisms can be successfully implemented within crowdsourcing applications. Some other VGI applications, such as OSM, recommend that contributors document and share with others the terms they use to describe geographic features, where terms are afterwards organized into a folksonomy. However, communities had difficulty in reaching a consensus regarding terminology; in addition, there are generally few mechanisms in crowdsourcing applications for checking adherence to the agreed-upon terminology.

The research on crowdsourced data quality is still at an early stage. According to Roick and Heuser (2013), the quality of crowdsourced data is one of the issues that has been the least studied so far. In addition, the suitability of quality management approaches for LBS has not been thoroughly explored yet.

LBS and the Internet of Things

Nowadays, users can produce geographic information via a variety of Internet applications. As a result, a “global digital commons of geographic knowledge” can be created without having to rely solely on “traditional” geospatial data production processes. In addition, it is expected that in the near future, smart devices that will be embedded in a variety of objects (vehicles, clothes, appliances, plants, buildings, etc.) will continue to blur the line between the virtual and real worlds (Liang, Croitoru, and Tao 2005). The upcoming availability of this data through the Internet of Things (IoT) infrastructure will create new opportunities to develop LBS applications that could better integrate information on the current state of the environment. In fact, we can consider that LBS are a building block of the IoT, since they both generate and consume location-based data that will be made available through IoT infrastructure.

Sui and Goodchild (2011) have put forward the idea of a generic data service that would retrieve, store, and provide access to geospatial data from LBSN, in addition to assisting users in the discovery of relevant information. Similar services are required to make the IoT a reality. However, before such a global data service can be implemented and before data from LBSN can be fully exploited and applied to different use case scenarios, several issues need to be addressed, such as enabling collaboration between applications and devices without human intervention, supporting the extraction and automatic formalization of context information so that applications can properly interpret the data, handling the IoT heterogeneities in an automatic manner, processing the IoT data streams with smart algorithms, and so on.

Today, some of the solutions that have been suggested for the development of the IoT rely
on existing standards and technologies for Sensor Web, such as OGC’s Sensor Web Enablement specifications, as well as on the Service-Oriented Architecture (SOA) principles.

Towards LBS of the future

While authors writing in “Location Matters: LBS 2030” explain that the first development phase of LBS was mainly driven by technologies such as positioning technologies and the Internet, they expect that coming development phases will be driven by data and users (Gartner 2013). Notably, the big data paradigm and related advances will stimulate the development of LBS applications that are based on the exploitation of important volumes of data, whether generated through sensors that will be widely disseminated in our environment, or crowdsourced applications. Then, while in the first decade of existence users of LBS were constrained by available technologies, once the main technological challenges have been addressed, the development of LBS will focus on users’ needs: for example, how to make LBS more useful to support collaborative work. Still, emerging technologies that have not been considered yet may challenge this vision of a “smooth” evolution where all technological challenges are behind us. For example, LBS may have to adapt to more embedded devices such as Google glasses in the future. One could also add that the legislation and regulations that will be adopted in relation to the use of LBS (privacy, data exchange, etc.) will have an impact on LBS evolution.

Whether the dissemination of information through LBS will benefit society and individuals from an economic, social, or environmental perspective will depend on how scientists, businesses, governments, civil society, and individuals address the challenges and opportunities created by LBS. The analysis presented in this entry reveals that the successful development of LBS will require integrated interdisciplinary approaches from a large variety of research fields that include computer science, GIScience, human–computer interaction, cognitive sciences, social sciences, and economics, to name only a few.

Still, the current advances suggest that, regardless of their impact on society, LBS will become even more widespread, until they eventually become part of the fabric of our daily life; so deeply integrated that it will become invisible.

SEE ALSO: Big data; Digital Earth; Ethics in GIScience; Geocoding; Geographic data mining; Geolocation services; Global navigation satellite systems; Open Geospatial Consortium standards; Routing and navigation; Volunteered geographic information

References


LOCATION-BASED SERVICES


Logistics

Markus Hesse
University of Luxembourg

Socioeconomic change and the modernization of logistics

Logistics includes the process of planning, implementing, and controlling the efficient, cost-effective, and synchronized flow and storage of raw materials, inventory, finished goods, and related information from point of origin to point of consumption for the purpose of fulfilling customer needs. Flows are also increasingly directed up-stream, from the point of sale or customer backwards to the supplier or producer, for example, in the case of vessels (e.g., empty containers) or reverse logistics, and also generally concerning data and information. While logistics initially emerged from the preparation and organization of military procedures, it is now used to optimize the functions of production, distribution, and consumption for serving businesses and households. Moreover, logistics was a major requirement for global trade to evolve to today’s volume, and it is a vital part of transportation operations of any kind.

Logistics once belonged to the three sciences of military operations, besides strategy and tactics. It emerged particularly during World War I, when large numbers of troops had to be moved and the supply of materials, vessels, and munitions had to be organized. It remained significant for any warfare until today, as recent experience for example, in the two Gulf wars in the early 1990s and 2000s, has revealed. The offspring of modern logistics goes back to the modernization of the capitalist economy, most notably the transformation of the modern corporation and the evolution of a networked economy. The development and constant improvement of industrial production and the emergence of a rather sophisticated division of labor across space and time would not have been possible without logistics. Particularly, the rationalization of production required an effective supply with materials and components. As a consequence, integrated supply-chain approaches were developed, including distribution and marketing. The practice of modern management, as described by Alfred Chandler (1977), started to integrate the flow of materials into overall managerial operations. However, logistics did not receive notable attention for long. In the early 1960s, management writer Peter F. Drucker called it “the economy’s dark continent,” still far from any particular consideration. Meanwhile, things changed significantly, due to both technological changes and innovations in organization. This change is also expressed by the move from the “visible hand,” as emphasized in Chandler’s treatise on management, towards the “vanishing hand,” as Langlois (2003) put it. The latter term refers to the discrete function logistics plays nowadays at various stages of the value-creating process, where a multitude of practices is conducted by an equally broad variety of actors and institutions (Table 1 indicates the major stages in the development of logistics).

The recent development of logistics and physical distribution was particularly triggered by a complex set of structural economic changes to which corporations had to respond. These
### Table 1: Phases in the development of logistics.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Developments in logistics</th>
<th>Aims</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-1970s</td>
<td>Traditional logistics, punctual improvements</td>
<td>Optimization of separate functions</td>
<td>Invention of the standard shipping container</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Logistics in military operations</td>
</tr>
<tr>
<td>1970s</td>
<td>Incremental improvements of traditional logistics</td>
<td>Optimization of separate functions</td>
<td>Increased outsourcing of specialized freight forwarding</td>
</tr>
<tr>
<td>1980s</td>
<td>Logistics as cross-section function</td>
<td>Optimization of processes</td>
<td>Emergence of third-party logistics service providers</td>
</tr>
<tr>
<td>1990s</td>
<td>Phase of functional integration</td>
<td>Establishment and optimization of process chains</td>
<td>Just-in-time production, lean management</td>
</tr>
<tr>
<td></td>
<td>Logistics integrates functions into process chains:</td>
<td></td>
<td>Retail logistics</td>
</tr>
<tr>
<td></td>
<td>Phase of comprehensive, interfirm integration</td>
<td>Establishment and optimization of value added chains</td>
<td>Emergence of fourth-party logistics service providers</td>
</tr>
<tr>
<td></td>
<td>Logistics integrates firms into value-added chains:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000s</td>
<td>Phase of worldwide integration of value-added chains</td>
<td>Establishment and optimization of global networks</td>
<td>Global production networks (GPN), e.g., in the apparel or electronics industries</td>
</tr>
<tr>
<td></td>
<td>Data integration in supply chain management</td>
<td></td>
<td>E-commerce in procurement and distribution systems</td>
</tr>
<tr>
<td>Recently</td>
<td>Green logistics, responsibility strategies</td>
<td>Enhancing sustainability, reliability, and control of logistics systems</td>
<td>Integrated reverse logistics</td>
</tr>
<tr>
<td></td>
<td>Enforcement of security concerns</td>
<td></td>
<td>City-logistics second generation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Certification systems</td>
</tr>
</tbody>
</table>

Source: Adapted from Baumgarten and Thoms 2002, 2.
changes comprise the rise of service economies, the increasing share of goods with high value and low weight, the upcoming high-tech and knowledge-based sectors, and certainly globalization, most importantly the rise of countries such as China or India to become the manufacturing center of the world. The rise of logistics became intensified by an accelerated competition among firms on increasingly open, deregulated markets. As a consequence, much of the storage, forwarding, and trucking activities of the firms that were previously organized in-house became outsourced to logistics subcontractors. The growth of these services helped in establishing specialized networks of freight forwarders (sea, land, air), of courier, express, and parcel services, and also of warehousing firms that were operating their own facilities on behalf of their customers.

Probably one of the most important catalysts of logistics development was the introduction of new information and communication technologies (ICT). They allowed for the truly integrated management and control of information, finance, and goods flows. Features like electronic data interchange, the automatization of product flow in dedicated warehouses, and distribution centers (DCs), computer-based tracking-and-tracing systems or radio-frequency based identification (RFID) of single items have changed corporate practice in logistics and physical distribution significantly. They are also the source of enormous productivity gains over the last two decades, as they permitted the operation of a completely new mode of inventory management, by adjusting materials management according to customers’ demand, thus reducing inventory costs substantially. The standardized container did the same; it made trade and transport more efficient, and thus much less expensive, from its very first introduction by a sea–land freight forwarding company in North America in 1956. The enormous growth of world trade, of which more than 90% of the volume is seaborne, would not have been possible without the economies of scale achieved by the efficient handling of containers. The throughput of containers via largely automatized terminal operations, their movement by ever-larger ocean carriers, and the challenge of maintaining access into and out of major seaport regions are among the most important challenges of the maritime economy due to logistics changes.

The Internet accelerated the application and distribution of new ICT in logistics firms, with research having some difficulties in properly assessing related impacts. Features such as the bar code and electronic data interchange (EDI) had already been practiced in logistics businesses since in the late 1970s/early 1980s. However, the advent of the web changed the situation dramatically. Now, data-handling technologies have become affordable on a general level, standardization of web appliances made the Internet a common feature, and corporations were able to establish interfaces with their customers. Now the web has become a major tool for rationalization, not only for individual firms and their subsidiaries, but for managing the entire producer–customer relationship and more complex networks. The accelerated pace of technological innovation contributed to further changes, for instance by introducing global positioning systems (GPS) that allow for identification and routing of vehicles (such as lorries, container ships) in a way that was not possible before.

As a consequence of these developments, corporate management has been switching to a flow-oriented mode of operation. Its core component is the supply chain, the time and space-related arrangement of the flow of goods, information, and capital between the stages of
supply, manufacturing, distribution, and consumption. These changes do result in an overall vertical fragmentation of the value added process, in modular manufacturing, worldwide sourcing, and big-box retail operations. The underlying generic idea of spatiotemporal fragmentation has to be compensated for by logistics integration. In order to describe the transformation or deconstruction of the firm in the context of fragmentation and networks, Suarez-Villa (2003) coined the term “techno-capitalism.” It is used as an interpretative scheme for analyzing structural changes that are primarily driven by new technologies and globalization. Network building, devolution of spatial hierarchies, and the reliability and speed of flows are its main characteristics. Thus, transportation costs and the supply of infrastructure lost their significance for explaining spatial differentiation, compared to corporate organization, institutions, and, particularly, the newly emerging logistics networks.

The geographical dimensions of logistics

When logistics is considered the art of managing and re-structuring material space–time relations, then it becomes clear that it has a substantial impact on geographies at various spatial scales (Hesse and Rodrigue 2004). Altogether with the expansion of infrastructure systems, logistics enabled regionally bound economies to further dissociate in terms of space, time, and organization. (In turn, the more complex and distanced the related corporate networks and trade systems were becoming, the higher was the demand for specialized logistics and freight distribution services.) On a global scale, logistics allowed for the structural rearrangement of value chains, production networks, and distribution channels. Astonishingly, logistics has received little attention by the pundits of global production networks (GPNs) so far – although the economic and social integration of regions into global economic relationships and interdependencies is hard to imagine without the properly designed flow of goods; the same applies for the increase in global trade rates that has been observed. As logistics and freight distribution are carrying the physical “load” of operating GPNs and trade flows, their significance is undoubted, however. Most notable is the emergence of major gateways alongside the trade routes, particularly large seaports and airports that offer the capacity to handle large portions of incoming or outgoing commodity flows. These gateways are bundling various flows oscillating between global, regional, and local scales, and provide connectivity with the larger hinterland regions (O’Connor 2010). As they concentrate major portions of national and international goods flows and related handling activities, they are subject to both positive and negative agglomeration externalities; ensuring the seamless flow of vessels through gateway regions still remains a costly and contested issue. Steering these regions is rather difficult due to an institutional fragmentation, consisting of corporations as diverse as shipping lines, carriers (of all modes), operators of warehouses, airfreight and container terminals, freight forwarding firms, receivers of goods (manufacturing, retail chains), and representatives from various public policy interests.

With respect to the landside movement and handling of consignments, the establishment of large distribution centers (DCs) is an important feature of current dynamics. This issue found significant coverage by geographical research recently. DCs have been developed in order to synchronize or buffer the materials flow between the scheduling of the supplier and the demand-side activities (sales, ordering) at wholesale or retail markets. They have replaced the old warehouses, as inventory has been diminished
until almost zero on both sides (production and wholesale/retail). DCs are often situated within regional logistics clusters, and big-box distribution centers have developed, as they offer advantages due to internal economies of scale. In locational terms, DCs and logistics parks have been subject to a significant sub- and exurban drift, with a certain preference for areas surrounding metropolitan cores that offer benefits for both long-distance haulage and regional distribution. Size, however, also tends to maximize the negative impact of related activities on neighborhoods and the environment. Besides traffic conditions in core city areas, land costs and policy pressure against a disturbing, locally often unwanted land use have pushed the movement of DCs out of the cities further.

DCs and logistics clusters also became the subject of regional development concepts and policies, particularly in the context of their localization as so-called transportation hubs. The innovation of the hub-and-spoke concept by airlines and freight shipping companies was based on an indirect routing of transport flows via major nodes. As a consequence, hubs were dislocated and got on the screen of regional development policies. Now, even places that lack any agglomeration potential could qualify for becoming a hub. Policy practices thus aimed at territorializing the dynamics of logistics, both locally and also in the context of state strategies (e.g., The Netherlands’ or Dubai’s attempt to create a global hub and “logistics city”). It was also argued that logistics corporations offer entry-level jobs, so this compensates for the loss of employment due to deindustrialization. However, research and practice are still challenged to make local sense of global flows in the context of emerging network dynamics, as hubs tend to be far more fluid, volatile, and fragile, compared to the traditional gateways. What they do have in common with gateways is that they were established on the ground of positive externalities of agglomeration (by routing and bundling flows), and they offer the same degree of negative externalities in terms of congestion, environmental, and neighborhood burden caused by the circulation of vehicles, aircraft, and the like.

In order to better understand the transformation of the logistics sector, research interest turned to the single corporation – particularly since large firms bundle a critical mass of flows and are proactive in using new technologies or organizational concepts. Such companies take advantage of a growing level of integration of services, finance, retail, manufacturing, and distribution. A good example of these powerful actors is the US retailer Wal-Mart, the largest corporation in the world, judging from relatively recent editions of the Fortune 500 list. Wal-Mart has already been coined the “template” for late-modern capitalism (Lichtenstein 2006). The company achieved this position by linking newly emerging places that provide the production and supply of cheap goods (e.g., in China) with serving major customer markets in North America. Wal-Mart thus successfully implemented a fine-tuned geography of retail outlets, distribution centers, and delivery concepts, with which it achieves an optimal coverage of markets and territories. This case also provides certain evidence that, besides the efficient management of financial resources, the core reason for its rapid success as a truly global company results from its ability to manage its supply chains and networks extremely efficiently (Brunn 2006). It combines minimum inventory with rapid turnover, fast and reliable supply chains, and a geostrategic know-how that helps in exploiting resources at the best possible level.

Whereas Wal-Mart has been considered a template for late-modern capitalism, the online retailer
Amazon.com might soon be regarded as the template for its digital version. At least it offers excellent empirical insight into the various ways that the material and the digital intersect. Again this feature coincides with the logistical capabilities of the firm and also its enormous rate of innovation. Within a time-span of about 15 years, Amazon became the world’s largest online store, investing massively in efficient logistics systems and optimal market coverage through a series of carefully placed distribution centers. Even more so, digital technologies are essential, allowing for not only the seamless flow of information, money and, last but not least, commodities, yet also the total control of assets, workforce, and customer behavior. The corporation is now seeking to further automate handling processes (e.g., by testing robots in DCs, planning delivery by drones) and also decentralizing distribution in major metropolitan regions. This may make same-day deliveries cost efficient on a regular basis. Hard to overlook is the impact that the further dominance of online retailers such as Amazon.com already have and will have in the future on stationary retail. It seems obvious that the rise of e-commerce severely affects one of the historical properties of towns and cities: to provide households and corporations with goods and services nearby. Will retail migrate into the back-end DCs of digital mail-order companies? The urban and societal consequences would be substantial.

**Future issues with respect to logistics**

There is a series of future issues that will raise further interest in logistics from a geography perspective, each including certain methodological challenges. First and foremost, an important issue will be energy and the related transportation cost. The circulatory economy is quite sensitive to the oil price, and the current rationale, for example, of locational concepts is very much based on the given cost of fossil fuel. If such limits are exceeded, the system will have to adjust, and it will be interesting to see how far and where this will go. Second, technological innovation will continue to develop. As projects such as the Amazon-drone or the Google-car already indicate, autonomous objects, robots, predators, and the like will take over an increasing array of functions within various supply chains. Spatial consequences are likely to occur, even though hard to break down at this point. Third, and this takes us back to the origins of logistics as warfare, there is an increasing “market” for security, surveillance, and intervention, secretly intersecting with geostrategic, economic interests. In the aftermath of 9/11, and with the NSA/GCHQ spy affairs not yet resolved, states and private agents will further use and thus revolutionize the management of flows. Logistics is becoming a core tool for regulating access and controlling mobilities of various kinds. In this respect, it seems most critical for the idea of an open society and has the potential to affect citizen rights severely (Cowen 2010). The astonishing degree and speed of change incorporated in recent developments has helped to create not only newly emerging material “logistics landscapes” (Waldheim and Berger 2008) but also a secret web of surveillance and control that deserves critical reflection.

**SEE ALSO:** Firms; Global commodity/value chains; Global production networks; Just-in-time production system; Trade and regional development; Transport geography

**References**


The Los Angeles School refers to a conceptual approach to understanding contemporary patterns of urbanization based on Los Angeles. The Los Angeles School emerged during the 1980s when a loosely formed group of predominantly Marxist urban scholars – geographers, planners, and urban historians – from the University of Southern California and the University of California, Los Angeles, turned to Los Angeles as, then, a comparatively understudied site for empirical and theoretical urban analysis (Davis 1990). While there has never been an official list of members, prominent commentators have been Michael Dear, Edward Soja, Mike Davis, Allen Scott, and Michael Storper. The work of these scholars is wide-ranging and presents numerous influences, objects of inquiry, and conceptual disagreements. Yet, the creation of the Los Angeles School was based on an understanding that Anglo cities were developing in new ways (post-1970), which rendered the classic modernist theories of the city increasingly obsolete (Dear 2000). The biggest critique was levied against the centralized, concentric-ring model associated with the widely known and influential Chicago School.

Based on dense empirical work on Chicago, the paradigmatic modernist city, the Chicago School was formed by sociologists at the University of Chicago during the 1920s and 1930s. Robert Park, Ernest Burgess, and Louis Wirth were its leaders. The Chicago School presented a theory of urban growth and change – often referred to as “urban ecology” – based on a central core surrounded by rings of distinctive zones. The concentric-ring diagram and model of 1920s Chicago is perhaps the best-known visual feature associated with this school of thought.

Los Angeles, conversely, was observed as a city void of a central core or of concentric rings even remotely reminiscent of the Chicago model. This observation led to the claim that the Chicago School was ill equipped to understand emerging trends and patterns of development that marked the post-Fordist, or what some Los Angeles School scholars called the “postmodern” metropolis. As most forcefully argued by Dear (2000), the polycentric, decentralized, and disorderly character of Los Angeles reflected what was unfolding across urban America and beyond, which formed the basis for claiming Los Angeles as the new prototypical urbanized landscape. Bluntly summed up by Joel Garreau in Edge City: Life on the New Frontier (1991, 3), “every American city that is growing, is growing in the fashion of Los Angeles.” As echoed particularly by Dear and Soja, Los Angeles epitomized what was happening across Sun Belt metropolitan regions and represented a new historical layer of development unfolding in many older, postindustrial cities, including Chicago.

The first explicit mention of a Los Angeles School is often attributed to Mike Davis in his landmark text, City of Quartz (1990). But it was not until Dear and Flusty’s (1998) article “Postmodern Urbanism,” published in the Annals of the Association of American Geographers, that the existence of the Los Angeles School gained widespread attention. On the basis of the dramatic departure of Los Angeles from the
LOS ANGELES SCHOOL

Chicago School model, Dear and Flusty argued that Los Angeles represented a “radical break” from modernist urban forms and that new epistemologies were needed to make sense of these changes. Modern industrial cities were supposedly being overwritten by a new hyperdifferentiated, “kaleidoscope” layer of development. And, to adequately understand this transformation, a new lexicon and taxonomy of postmodern urbanisms based on southern California was presented. For example, concepts such as “keno capitalism” were used to capture the perceivable lack of order or single driving force underpinning this emergent urban form. Dear subsequently became the primary spokesperson and most ardent promoter of the Los Angeles School.

This embrace of the postmodern movement, however, also marked a conceptual division within the Los Angeles School. Dear and Soja, notably influenced by Henri Lefebvre, Michel Foucault, Jacques Derrida, and Fredric Jameson, were among the earliest scholars to incorporate postmodern insights into human geography (Dear 2000; Soja 1989; 1996). But Dear and Soja were somewhat at odds in their approach to this incorporation. Soja’s postmodernism was based more on invigorating the role of space and spatiality in critical social theory, long subsumed beneath the dominance of time and temporality. Dear, in contrast, embraced a notably deeper, more epistemological basis for his postmodernism, criticizing Soja for still adopting a classic Marxist (and therefore) modernist metanarrative to structure his analyses and interpretations (Dear 2000).

But there were also more fundamental divisions within the group, with Davis rejecting postmodernism and Scott and Storper avoiding it altogether. Scott and Storper, for instance, rather than directly engaging with postmodernism, have tended to situate their work in industrial location patterns, flexible specialization, and the “regionalization” of contemporary metropolitan areas within the broad transition from Fordist to post-Fordist modes of production observed in advanced economies since the 1970s (e.g., see Storper 1997; Scott 1998).

Numerous criticisms have been levied against the Los Angeles School, notably the claims of this city as paradigmatic and that this new postmodern urbanism represented a radical break with the past (e.g., see Harvey 1989). This debate is now extensive, with the entire inaugural issue of City & Community (2002) covering the topic. A central criticism is that the Chicago School’s dominance was overstated, notably by Dear, to bolster the relevance of the Los Angeles School, as, by this time, the Chicago School was already considered by some as a relic of the past. Although Dear continues to advocate for the Los Angeles School (e.g., see Dear and Dahmann 2008), these debates have notably subsided. But, regardless of where one stands, the Los Angeles School stimulated significant and important critical debate in urban studies at the time, with many insights now generally accepted, such as the increasing polycentric patterning of metropolitan regions and obsolescence of the urban/suburb dichotomy.

SEE ALSO: Chicago School; Critical spatial thinking; Cultural turn; Edge city; Modernity; Postmodernity; Urban geography

References


Description and purpose

The institute’s objectives are the development of theoretical bases and methodology for physical, human, and regional geography studies on spatial processes and interrelationships; temporal and spatial surveys of the interaction between humans and environment; assessment of factors of the geographical environment with special reference to natural and socioeconomic resources and to the emerging socio-economic issues in the Carpatho-Pannonian area; international cooperation; and documentation and dissemination of research achievements through the publication of studies and periodicals in Hungarian and foreign languages.

Journals or major publication series


Current activities or projects

From 2012 scientific work has been organized into eight scientific project groups supervised by the director. These project groups are: National Atlas of Hungary; Geomorphology and Quaternary Studies; Loess and Quaternary; Recent Physical Geographical Processes and Landscape Studies; Lithosphere-Biosphere Interactions; Ethnic and Political Geography; Spatial Mobility; and Metropolitan Regions and Urbanization Processes. The project groups are supported by three functional groups: Cartographical Group, Laboratory for Sediment and Soil Analysis, and a Geographical Library.

Submitted by Károly Kocsis
Machine learning

Ningchuan Xiao
Ohio State University, USA

In his 1950 paper “Computing Machinery and Intelligence,” Alan Turing posed a provocative question: can machines think? To further understand the meaning of thinking and machine, Turing developed an imitation game, now more popularly known as the Turing test, to determine whether a computer can act as a human via a conversation with another human. Turing predicted that computers equipped with the ability to learn will pass the test. Tom Mitchell (1997) later defined the meaning of learning in a more specific context: “A computer program is said to learn from experience \( E \) with respect to some class of tasks \( T \) and performance measure \( P \), if its performance at tasks in \( T \), as measured by \( P \), improves with experience \( E \).” Examples of tasks in machine learning include recognizing spoken words, playing chess, identifying objects (e.g., houses, trees, parks, and roads) from an aerial photograph, and identifying places for ecological habitats.

For a computer program to learn, past experience must be presented to the program, which often takes the form of a set of data collected for the task. A machine learning method can be broadly categorized as being either supervised or unsupervised. A supervised method needs the desired result to be labeled or identified in the data, while an unsupervised method uses unlabeled data to uncover the structure in the data. A performance measure must be developed to indicate how well the machine or computer program has learned from the data. For example, if the machine is expected to identify housing styles from aerial photographs, then the performance measure will score how well the machine performs this task. To produce unbiased and robust learning results, the data are often partitioned into a training set that is used to calibrate the parameters of the learning program and a test set for evaluating the learning results.

Machine learning spans, and draws from, a number of different academic fields. Within computer science, machine learning is a special branch of artificial intelligence. Many methods used in machine learning are also derived from statistics. A closely related field to machine learning is data mining, which shares many common theories and methods with machine learning. However, data mining has a focus on the discovery of unknown patterns in the data while machine learning is more about prediction: using learned patterns to understand new data. For example, data mining approaches are useful in diagnosing the problems of a car; machine learning methods can be used to automatically drive an autonomous vehicle. Machine learning results must be generalizable, so that the result can perform under circumstances that are not necessarily reflected in the training data.

Clustering

Given a dataset, \( w \), each observation or data entry can be assigned to a cluster using a similarity or distance measure. In this way, observations assigned to the same group are similar to each other, while observations in different groups are dissimilar. A geographical example would
be classifying pixels in a remote sensing image as being one of several different land-use and land-cover classes (e.g., forest, agriculture, and water). The training data are denoted as $x = (x_1, x_2, \ldots, x_n)$, where $x_i (1 \leq i \leq n)$ is a $d$ dimensional vector. It is assumed there exists a set of $k$ mean (or center) data points $(\mu_1, \mu_2, \ldots, \mu_k)$ that is used to define a set of $k$ clusters $(s_1, s_2, \ldots, s_k)$ such that a data point $x_i$ belongs to cluster $s_j$ if, and only if, the distance between $x_i$ and $\mu_j$ is the nearest among all the $k$ center data points. The goal of solving the clustering problem is to find a set of center points such that the total distance between each data entry to its nearest center is minimized. Finding the exact solution to the problem is computationally intensive and the commonly used $k$-means method is an algorithm that can be used to return a good, though not necessarily optimal, solution.

The $k$-means algorithm has two steps. In the first step, $k$ data points in the $d$ dimensional space are randomly determined as the initial center points and each observation in $x$ is assigned to its nearest center based on the distance between the observation and the center. In the second step, observations assigned to the same center are used to calculate a new center as

$$\mu'_i = \frac{1}{|s_i|} \sum_{x_j \in s_i} x_j$$

If the new centers exhibit significant differences from the current centers, the new centers will be used to replace the current ones and the assignments of observations in $x$ will be updated. This step is repeated until no significant gain in the total distance can be achieved.

**Decision tree learning**

A decision tree represents a step-by-step classification process in which a question is asked in each step and, depending on the answer to each question, a decision can be made or more questions will be asked before a decision. For example, to decide on whether one should bring the umbrella today, one may first ask, “Is it raining now?” If it is raining now, then the decision will be yes (will bring umbrella). If it is not raining, one may further ask, “Will it rain later today?” If weather forecasting shows a high likelihood of rain, then the answer will be yes. Otherwise, the umbrella will not be brought. In this simple example, each question relates to an input variable or attribute (e.g., raining), which forms a node of the decision tree. When the answer to the question leads to the target (e.g., if raining is yes, then bring umbrella), the target answer is a leaf of the tree. The first question to ask is the root of the tree. Figure 1a shows the formation of a tree that captures the flow of decisions and questions in this simple example. In a more complicated situation, it is critical to determine which question will be asked first and the order to ask all the questions.

Decision tree learning is an example of the supervised machine approach that requires a training set where each data entry is labeled with a given classification. A learning algorithm is developed to associate the attributes of training-set members with the known target. The association will be represented in a series of decisions that can be illustrated as a tree. Different from the above example, where the sequence of questions seems to be straightforward, the training dataset can be more complicated and it is often difficult to know which question to ask first. The first question becomes the root of the tree.

To determine which question to ask first, it is necessary to identify which attribute best gives insight into choices. Assume there are $k$ target decisions to make (for example, in the above example, $k = 2$). Let $p_j$ be the probability for decision $j (1 \leq j \leq k)$ to occur in the data. The
The impurity of the data can be measured using an entropy measure

\[ H = -\sum_{j=1}^{k} p_j \log_k p_j \]

In information theory, entropy has been used to measure the disorder in the system. When only one class exists in the data, \( H = 0 \), indicating no further learning is necessary. When the \( k \) decisions are equally divided in the data, \( H = 1 \), indicating the highest level of disorder.

Suppose the question about an attribute takes \( m \) answers. By answering this question, the data can be divided into \( m \) subgroups and it is assumed that there are \( N_i \) data entries in subgroup \( i \) \((1 \leq i \leq m)\). The probability for the data to fall in class \( j \) in the subgroup is \( p_{ij} = N_j / N_i \), where \( N_j \) is the number of data labeled as decision \( j \). The total impurity of answering question \( v \) is

\[ H(v) = -\sum_{i=1}^{m} \sum_{j=1}^{k} p_{ij} \log_k p_{ij} \]

The information gain can then be calculated as \( H - H(v) \), which indicates how well attribute \( v \) can be used to classify the data. The attribute with the highest gain is then used as the root of the decision tree. For each of the subgroup, if the data in that subgroup contain more than one decision, a new node must be created and the data in that subgroup will undergo the same process described above to choose the attribute for the node. This process continues until the tree only has leaf nodes at the end of each branch, or all the attributes are used.

**Figure 1** Decision trees. (a) A simple decision tree to decide whether to bring an umbrella. (b) A decision using the sample data about likability of cities. The overall answer in the Like column has four yes answers and six no answers, which gives a total entropy of:

\[
-\frac{4}{10}\log_{2}\frac{4}{10} - \frac{6}{10}\log_{2}\frac{6}{10} = 0.53 + 0.44 = 0.97.
\]

At the root level, job opportunity has the highest information gain (0.97 − 0.32 = 0.65) and is chosen as the root of the decision tree.
Decision tree learning can be illustrated with an example dataset (Table 1) about the likability of 10 cities with three attributes: living costs, jobs, and air quality. The resulting decision tree (Figure 1b) is essentially a set of rules that can be used to determine the likability of a new city. For example, for any city with high (or low) job opportunities, according to this model it would be predicted that its likability will be yes (no). For a city with medium job opportunities and low air quality, it will not be likable. This example also shows some limitations of this method. For example, the model will not tell the likability of city with medium job options, high air quality, and low living cost, because this combination of attribute values is unknown in the training dataset.

Artificial neural networks

Artificial neural networks (ANNs) are inspired by the advances in animal central nervous system research. In essence, ANNs are meant to mimic how biological brains operate by representing how neurons are connected together in the brain. A typical ANN has three layers and each layer has a number of nodes that represent neurons in real biological brains. The input and output layers have the same number of nodes as the input and output variables of a problem. The third layer is called a hidden layer and each node in this layer is called a perceptron. Each perceptron takes the input from each input node and algorithmically transforms them into the values of the output variables. Figure 2a shows...
Table 1 A sample dataset. For each attribute, the values of H, M, and L refer to as high, medium, and low. Air quality only has two values: H and L. The Like column only has a binary value of Yes (Y) or No (N).

<table>
<thead>
<tr>
<th>City</th>
<th>Living cost</th>
<th>Jobs</th>
<th>Air quality</th>
<th>Like</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H</td>
<td>H</td>
<td>L</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>H</td>
<td>M</td>
<td>L</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>H</td>
<td>H</td>
<td>L</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
<td>M</td>
<td>H</td>
<td>Y</td>
</tr>
<tr>
<td>5</td>
<td>H</td>
<td>H</td>
<td>L</td>
<td>Y</td>
</tr>
<tr>
<td>6</td>
<td>H</td>
<td>M</td>
<td>H</td>
<td>N</td>
</tr>
<tr>
<td>7</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>N</td>
</tr>
<tr>
<td>8</td>
<td>M</td>
<td>M</td>
<td>L</td>
<td>N</td>
</tr>
<tr>
<td>9</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>N</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>L</td>
<td>L</td>
<td>N</td>
</tr>
</tbody>
</table>

The layout of an ANN with three input variables, two outputs, and a hidden layer of four nodes. Each node in the hidden layer is a perceptron that sums a set of weighted input information and converts it to a standardized value between 0 and 1 (Figure 2b). Formally, the output from a perceptron can be written as

\[ s_j = \sum_{i=1}^{n} w_{ij} x_i \]

where \( w_{ij} \) is the weight between the \( i \)-th input node and the \( j \)-th perceptron, and \( x_i \) is the input from the \( i \)-th input node. A sigmoid or “S”-shaped function can be used to convert \( s_j \) to a standardized value

\[ o_j = \sigma(s_j) = \frac{1}{1 + e^{-s_j}} \]

a step function can also be used for this purpose. A node in the output layer uses another set of weights to yield the estimated value:

\[ o_k = \sum_{j=1}^{m} w_{jk} s_j \]

where \( w_{jk} \) is the weight between the \( j \)-th hidden node (\( 1 < j < m \)) and the \( k \)-th output node.

A critical characteristic of a neural network is its learning ability, which it gains by virtue of adjusting the weights used in the above computation in order to minimize the overall difference between the estimated values \( (o_k) \) and the known target values (denoted as \( y_k \) for the \( k \)-th output) in the training set. For each entry in the training data, an error at an output node can be calculated as

\[ E = \frac{1}{2} (y_k - o_k)^2 \]

The partial derivative of \( E \) over the weight is

\[ \frac{\partial E}{\partial w_{ij}} = \frac{1}{2} \frac{\partial}{\partial w_{ij}} (y_k - o_k)^2 = (y_k - o_k) \left( -\frac{\partial o_k}{\partial w_{ij}} \right) = -(y_k - o_k) \frac{\partial o_k}{\partial s_k} \frac{\partial s_k}{\partial w_{ij}} \]

For the sigmoid function

\[ \frac{\partial o_k}{\partial s_k} = \frac{\partial \sigma(s_k)}{\partial s_k} = o_k(1 - o_k) \quad \text{and} \quad \frac{\partial s_k}{\partial w_{ij}} = x_i \]

Therefore for the output nodes, the error gradient is

\[ \frac{\partial E}{\partial w_{ij}} = -(y_k - o_k) o_k(1 - o_k) x_i \]

The weight can then be adjusted using a rate proportional to the error gradient. Based on this error correction principle, a back-propagation algorithm (so named because it operates from outputs back to the inputs) can be developed.
to systematically adjust the weights until the overall error is reduced to an acceptable level (Box 1).

### Box 1 Back-propagation algorithm.

Assign random values to all the weights

Repeat

Randomly pick an entry from the training data and compute the outputs

For each output node, compute

\[ \delta_k = (y_k - o_k) o_k (1 - o_k) \]

For each hidden node, compute

\[ \delta_j = o_k (1 - o_k) \sum_{k=1}^{m} w_{jk} \delta_k \]

For the entire network, update each weight as

\[ w_{ij} + \Delta w_{ij}, \] where \( \Delta w_{ij} = \eta \delta_j x_i \)

Until the termination conditions are met

In this algorithm, \( \eta \) is a small constant called the learning rate. The algorithm typically terminates when the error at the output layer is lower than a predefined level, or when the number of iterations reaches a predefined number.

### Genetic algorithms

Genetic algorithms (GAs) belong to a family of computational methods that is inspired by the biological theory of natural selection. To use a GA as a classifier that returns a class assignment to an input, it is critical to represent classification rules as an individual in the GA. Extending the example of the training data in Table 1, living costs and job opportunities can be encoded using three-digit binary strings, where values of H, M, and L are encoded as 100, 010, and 001, respectively. Value combinations can also be represented using this method. For example, a binary string of 011 indicates a value of M or L. For air quality, a two-digit binary string can be used and values of H and L are encoded as 10 and 01, respectively. A string of all 1’s is used to represent a condition when the corresponding input variable does not affect the classification result. Finally, a single binary digital can be used to encode likability of yes (1) or no (0). Table 2 lists a set of rules and their encoding in the GA.

A classifier often has multiple rules (see the result from the decision tree on the training data), which can be constructed by cascading across multiple binary strings. With such a representation strategy, a classifier genetic algorithm contains a population of \( P \) individual classifiers. Here, a single rule classifier is used as an example. An initial set of classifiers are randomly generated and then each classifier is evaluated by a fitness function that is often based on the number of data entries the classifier can correctly match. For example, the last rule listed above can match three instances in the data, while a rule of 001111111 does not match any instances.

The core mechanism in GA is that the individuals are selected to create new individuals on the basis of how well they solve the problem, with better ones being more likely to be

<table>
<thead>
<tr>
<th>Encoding</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>100010101</td>
<td>IF Living = H AND Jobs = M AND Air = H THEN Like = Y</td>
</tr>
<tr>
<td>010010100</td>
<td>IF Living = M AND Jobs = M AND Air = H THEN Like = N</td>
</tr>
<tr>
<td>111010010</td>
<td>IF Jobs = M AND Air = L THEN Like = N</td>
</tr>
<tr>
<td>111100111</td>
<td>IF Jobs = H THEN Like = Y</td>
</tr>
</tbody>
</table>
selected. After evaluation, each individual is assigned a selection probability as the ratio of its fitness value (a measure of how well it solves the problem) and the total fitness values of the population. Individuals with high probabilities will have a high chance to be selected to create the next generation of the population. Specifically, $\frac{P}{2}$ pairs of individual will be selected and, for each pair, a cross-over operation is used to produce two new individuals. For example, when two individuals of $H_1 = 010010100$ and $H_2 = 001111111$ are selected, the GA randomly decides a cross-over point, say, at after the third digit of each rule. Then the first part of $H_1$ is combined with the second part of $H_2$ to form a new individual of $H'_1 = 010111111$. Similarly, the first part of $H_2$ can be combined with the second part of $H_1$ to form $H'_2 = 001010100$. The cross-over operation will create a new population that is used to replace the previous one. Finally, a mutation operation can be carried out on each individual, on a rare and random basis, to flip a randomly selected digit in the individual. For example, flipping the seventh digit of $H'_2$, gives 001010000. The new population subsequently undergoes the selection, cross-over, and mutation operations to generate the next generation. The process continues until a predefined condition, such as the maximum fitness value or maximum number of generations, is satisfied.

GA performance can be significantly improved by controlling the pace of convergence. Because individuals are selected based on their fitness values, highly rated individuals may dominate the selection process and, therefore, take over the population at the early stage of the GA process. This may trap the entire population into a local optimum because individuals at the early stage tend to be inferior to the global optimal classifier. To prevent a GA from being premature in choosing its final solution, it is important to increase the diversity of the GA population so that individuals with low fitness values can be maintained in the population.

Though the fitness values of these individuals may be low, they may still contain useful binary blocks that can contribute to the overall search when they are selected and combined with other individuals.

**Support vector machine**

Support vector machine (SVM) is a relatively new development in machine learning. Here, a training data entry is denoted as $\{x_i, y_i\}$, where $x_i$ contains $d$ attributes and $y_i$ is the output that has two values of $-1$ or $+1$, representing two classes. Assume there is a $d$-dimensional hyperplane that can separate the data into the two sides of the hyperplane where all the $y$ values of $-1$ are on one side and $+1$ on the other. When $d = 2$, a straight line can be drawn to separate the data in a 2-D plane. The hyperplane can be denoted as $w \cdot x + b = 0$, where $w$ is the normal vector of the hyperplane and the operation ($\cdot$) means the dot product. The data points that are closest to the hyperplane are called support vectors and it is necessary to find a hyperplane such that the distance between the support vectors and the hyperplane is maximized. In Figure 3, the two points on the dashed lines are support vectors and the solid line between the two points is the optimal hyperplane. Given a hyperplane, there are a pair of planes that are in parallel with the hyperplane and cross the support vectors (as the two dashed lines in the 2-D case). The distance between the hyperplane and one of these parallel planes is called the margin of the support vector machine and can be computed as

$$\frac{1}{||w||}$$
To find the optimal hyperplane with a maximized margin, a specific optimization problem is designed that minimizes

$$\frac{1}{2}||w||^2$$

subject to $$\gamma_i(x_i \cdot w + b) \geq 0 \ \forall i$$. The constraints ensure any point $$i$$ in the training data is on one side of the dashed line. This problem can be converted to a convex quadratic problem that can be solved efficiently with well-known methods.

In many applications, the different classes in the data may not be separated by a linear function but instead by a more complicated one. In order to apply the linear separation approach, the data can be mapped into a higher dimension so that the linear separation can be applied. It is also possible to enforce a so-called soft margin on the training data, such that a few data points are allowed to be on the wrong side of the hyperplane. When multiple, instead

of two, classes are necessary, various approaches have been developed to convert the multiple classes into a number of binary classification problems, essentially taking a divide and conquer approach.

**Discussion**

A challenge that all machine learning methods face is overfitting, a phenomenon by which the results found from the training data turn out to be overly tailored to the training set at the cost of being generally applicable. These results, when used on test data or other new data, may yield errors higher than in the training dataset. The decision tree example described here, for instance, derived from the training data is perfect in the sense that it can correctly match every entry in the data. However, it would easily go wrong on a new data entry with the values of H, H, L, and N, respectively, because these values are not well represented in the training set but are still a reasonably likely combination. There are many ways to address the overfitting issue in learning. Penalty methods, for example, add a complexity penalty into the evaluation of each alternative, where the penalty is proportional to the complexity of the result. In a decision tree, a simple way to measure the complexity is the size (or number of branches) of the tree, where complex trees with many branches will be penalized by having a reduced evaluation score. Another method is called hold-out or cross-validation because it utilizes a smaller training dataset by holding out a subset of the training data to be used as experimental test data. While this method creates a smaller training set for learning, the subset that is held out can be used to test the performance of learning and, therefore, can prevent overfitting. The bootstrap method
handles overfitting by repeatedly drawing random subsets of data with replacement from the entire dataset. The statistical distribution of learning performance can be used to test the robustness of the result coming from the training dataset.

The machine learning methods described so far represent some important advances in the field. Other methods, such as Bayesian learning, instance-based learning, reinforcement learning, and analytical learning, are also important. Machine learning methods have been used in a wide range of application fields, such as speech recognition and computer vision. In geographic research, machine learning has been a focused area in remote sensing and geographic information science, for tasks such as land cover classification and spatial pattern analysis.

**SEE ALSO:** Artificial neural networks; Cluster detection; Decision analysis; Fuzzy classification and reasoning; Geographic data mining; Geographic information science; Spatial optimization; Spatiotemporal analysis; Supervised classification; Unsupervised classification

**Reference**


**Further reading**


Han, J., M. Kamber, and J. Pei. 2006. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.


Mainstream and shadow banking

Simon Xiaobin Zhao
Nam Hee Kim
University of Hong Kong, China

Mainstream banking

Mainstream banking can be defined as a system of commercial and retail banks that are closely regulated and supervised by government banks commonly referred to as central banks. Commercial and retail banks are privately owned institutions that serve as financial intermediaries for the safeguarding, transfer, exchange, and lending of money. Central banks are government entities tasked with supervising the safety and stability of the money supply and the banking system.

Commercial and retail banking

Commercial and retail banks accept deposits from businesses and households and pay an interest on the deposits. The banks then use the deposits to fund longer-term loans directly to borrowers. The activity of transforming the risk and timing of cash flows between savers and borrowers is called financial intermediation. An important aspect of financial intermediation is maturity transformation, which is the use of short-term liabilities, such as deposits, to fund long-term assets such as mortgages. This maturity transformation creates liquidity risk, which is the risk that it will not be able to raise cash to meet demand from depositors withdrawing their money or short-term creditors demanding repayment. Liquidity risk creates the possibility that an institution may be forced to sell its long-term assets at distressed prices in order to meet its short-term liabilities.

Another key function of banks is the provision of credit. In providing credit, banks accept and manage credit risk, or the potential that the borrower fails to repay the money that the bank lends. Through financial intermediation, banks transform savings into long-term sources of capital that allow individuals, corporations, and governments to engage in economic activity.

One of the most significant trends in mainstream banking since the 1990s has been consolidation. Some of the acquisitions were cross-border, resulting in a few large banks controlling more and more of the world’s money. The ultimate outcome of the wave of mergers in the 1990s and the globalization of banking has been a geographical concentration of banking decision centers and strategic functions. According to data compiled by Forbes in 2015 using the banks’ reported assets under their respective accounting regimes, China, United States, and Japan had the largest number of big banks, with, respectively, 13, 12, and 9 of the 100 largest banks by assets being headquartered there. The top 18 largest banks controlled assets of $40.2 trillion, which represent just over half of the total assets controlled by the top 100 banks. The top 30 banks were all located in North America, North Asia, the European Union, or the United Kingdom. Outside of those regions, Australia had the largest number of banks in the top 30 list, with five banks holding just over $3 trillion
of assets, followed by Brazil with three holding combined assets of $1.4 trillion. Only one bank from Africa, the Standard Bank of South Africa, made the top 100 list with assets of $182 billion under management.

Geography of mainstream banking

The role of geographical distance in banking activities is most often attributed to its effect on transportation and information costs (Elliehausen and Wolken 1990). Transportation costs include both the time and money spent by either the bank or its customer to facilitate in-person transactions. Several studies have examined how distance affects a consumer’s choice of depository institution and shown that consumers are more likely to choose banks based on geographic proximity (Chiappori, Perez-Castrillo, and Verdier 1995; Dell’Ariccia 2001; Villas-Boas and Schmidt-Mohr 1999). Information costs include the cost of advertising and maintaining relationships with existing and potential customers. Lenders who are geographically closer to customers may also gain a competitive advantage since they should be able to gather information about borrowers during the approval process as well as monitor the loans throughout the life of the loan at lower expense.

Automated teller machines (ATMs) have the potential to reduce the transportation costs for both customers and the banks that serve them. If consumers view ATMs as an effective substitute for bank branches, banks can compete for customers in geographic areas where they have few or no branches by providing services through an ATM. While this is unlikely to play any role in lending, ATMs may allow banks to expand their geographic reach at lower cost. However, using data on banks in Italy, Hester, Calcagnini, and de Bonis (2001) find that the number of ATMs in a province is positively related to the number of branches a bank has in that province, a result that appears contrary to the idea that ATMs are deployed by banks as substitutes for branches.

Not only does the distance between the bank and the customer matter in banking, but the distance between the branch (or location) and the headquarters (HQ) where the lending decision is made also impacts on banking relationships and decisions. Credit availability and innovation adoption have been found to be inversely related to functional distance (Alessandrini, Presbitero, and Zazzaro 2009). A large bank is, on average, subject to greater functional distance between its HQ and its branches and is less able to produce soft information; therefore it is less inclined to make relationship loans to small and medium enterprises (SMEs) compared to small banks which, in theory, are subject to a smaller functional distance (Berger, Rosen, and Udell 2007).

Banking activities vary greatly regionally and nationally because, perhaps more than any other industry, they are shaped by government-imposed rules and regulations distinct to each territory. Government regulation impacts the spatial distribution of mainstream banking across jurisdictions. In the United Kingdom the five “major” banks – Royal Bank of Scotland (RBS), HSBC, Barclays, Lloyds, and Standard Chartered – account for over 90% of bank deposits. These large banks maintain a network of branches primarily for the purpose of collecting deposits, but grant fewer loans to local SMEs. In Germany, by contrast, banking is dominated by almost 2000 small, locally headquartered banks distributed throughout the country whose activities are geographically restricted by law to the local area. These small, local banks in the regions account for 70% of bank deposits, while the larger, major banks account for only 13% of bank deposits.
In Islamic countries, banks and financial institutions have created financial contracts that conform to Islamic, or sharia, law, commonly termed “Islamic finance”; such banks do not pay or receive interest, or *riba*, on banking products. In lieu of *riba*, prohibited by the Quran, banks share in the profit generated from the use of funds that it loans to its customers. The depositors also share in the profits of the bank at a predetermined ratio. Most Islamic banks are headquartered in the Middle East including Iran, Saudi Arabia, United Arab Emirates, and Kuwait. Outside of the Middle East, Malaysia is home to several Islamic banks.

With increasing globalization and cross-border flows of capital, firms and households can choose not only between banks within their local area or jurisdiction but also make deposits and use the financial services of banks based in foreign jurisdictions, sometimes thousands of miles away. Many mainstream banks offer “offshore” banking to take advantage both of the absence of a global legal framework and of disparate tax structures. Although there is no precise definition of an offshore jurisdiction, significant offshore financial centers (OFCs) around the world include Bermuda, the Cayman Islands, the Channel Islands, Hong Kong, and Singapore (Roberts 2009). Onshore governments argue that regulatory efforts in banking are undercut by competition from offshore financial centers.

An increase in regionally concentrated imbalances may also be a by-product of globalization, creating a more uneven distribution of financial access. Access to mainstream banking services varies between both countries and communities within a country. A two-tier banking system has developed in the United States, with mainstream finance in most places but financial exclusion, predatory lending, and fringe banking (pawnshops, payday lenders, etc.) in certain communities (Caskey 1994; Dymski 1999; Immergluck 2009; Leyshon and Thrift 1997; Squires 2004). Predatory lending can increase the likeliness of loan default and housing foreclosure, which in turn can spiral into housing abandonment at the neighborhood level. Foreclosure rates vary widely between communities, especially during times of crisis. Smith (2008) finds that areas in Chicago that are over 80% minority exhibited more than five times the rate of foreclosures of areas with less than 10% minority during the Great Recession.

Central banks and international governance

Central banks are the government banks that manage, regulate, and protect the money supply and banks themselves. Central banks perform the role of lender of last resort (LOLR), providing loans to a financial institution when the market refuses to do so, possibly avoiding default caused by the lack of liquidity. In the nineteenth century, after the collapse of a major London discount bank, Overend and Gurney, the Bank of England became the first central bank to perform this role. As the role of central banks evolved through the twentieth century, macroeconomic stability, including a focus on inflation targeting, became one of their essential responsibilities. In some developed countries, the supervision of financial institutions was transferred from central banks to other agencies.

In the United States, the Federal Reserve System performs this central banking function and supervises banks, as well as acting as crisis manager with the authority to oblige banks to monitor their risks. In the European Union, the European Central Bank (ECB) has primary stability responsibilities when it comes to the payment system but not the explicit task of preserving the stability of the financial system in general. That responsibility falls on the national central banks, which also have the LOLR role.
The debate over the role of central banks persists, with some arguing that central bank interventions encourage banks to take more risks on the assumption that they will be bailed out. In times of crises or when an individual bank is in distress, bank runs can occur: a rapid loss of deposits that causes the bank to fail from lack of liquidity. Many countries address this problem by providing government guarantees for deposits or deposit insurance. According to the International Association of Deposit Insurers, nearly 100 countries offer government guarantees on deposits. Holding sufficient liquid assets also reduces the chance that investors would become concerned enough to cause a bank run. In most jurisdictions, regulatory requirements are in place for banks to hold sufficient reserves.

Established in 1930, Bank for International Settlements (BIS) is an international organization based in Basel, Switzerland, commonly referred to as the “central bankers’ bank.” Its aim includes fostering the cooperation of central banks and international monetary policymakers. Situated within the BIS, the Basel Committee on Banking Supervision (BCBS) was established in 1974 by the G10 central banks, Switzerland, and Luxembourg to supervise the activities of commercial banks. Committee membership now also includes Argentina, Australia, Brazil, China, Hong Kong SAR, India, Indonesia, South Korea, Mexico, Russia, Saudi Arabia, Singapore, South Africa, and Turkey. In 2004 the committee developed the capital adequacy framework, commonly referred to as Basel II. In response to the financial crisis of 2007–2008, an enhanced set of rules named Basel III was introduced in November 2010 at the G20 summit in Seoul. Excessive leverage and insufficient liquidity buffers in the banking sector are believed to have precipitated the financial crisis of 2007–2008, and Basel III standards are intended to safeguard against a similar crisis by requiring banks to maintain lower levels of leverage – the ratio of capital to assets – and higher amounts of cash than required under Basel II (BCBS, Basel Committee on Banking Supervision, 2015). The Basel process has gained focus among geographers as one of the first truly “glocal” institutional responses that coordinates the multilateral efforts of central banks, regulators, and market participants in an attempt to foster and maintain global and local financial stability (Bieri 2009; Christophers 2014).

While BCBS’s efforts may lead to an increased adoption of regulatory and supervisory best practice around the globe, this does not necessarily mean that local compliance with such standards is uniform. Regional variations in the quality of banking regulation and institutions persist, only some of which can be explained by differences in economic development (Bieri 2009).

Shadow banking

Broadly, shadow banking is defined as bank-like financial activities that are conducted outside the traditional commercial banking system, with little or no regulatory oversight. Some market participants prefer to use terms such as “market-based financing” rather than “shadow banking” to describe nonbank financial intermediation. Shadow banking institutions differ from mainstream banking institutions in that they cannot take deposits and do not benefit from deposit insurance. Notwithstanding widespread disagreement over which entities constitute the shadow banking system, it includes financial intermediaries that serve as proxies for traditional banks, such as investment banks, finance companies, money market funds, hedge funds, private equity groups, structured investment vehicles, payday lenders, and insurance firms.
A study of 23 jurisdictions by the Financial Stability Board (FSB, Financial Stability Board, 2014) estimated that shadow banking assets stood at $34.9 trillion at the end of 2013, up from $34.0 trillion in 2012. The United States had the largest shadow banking sector in terms of US dollars, followed by the United Kingdom and China. Brokerage or investment banks, accounting for $9.3 trillion of shadow banking system assets, are concentrated in the United Kingdom (39%), United States (28%), and Japan (21%). Structured finance or securitization vehicles, accounting for $5 trillion at the end of 2013, are concentrated in the United States (35%) and the United Kingdom (13%). Finance companies and money market funds made up $4.1 trillion and $3.8 trillion of total shadow banking assets, respectively. Money market funds are mainly concentrated in the United States and the euro area. The FSB study found that hedge funds accounted for an insignificant portion of shadow banking assets. However, the majority of hedge funds are domiciled in offshore jurisdictions such as the Cayman Islands, which were not included in the FSB study.

Brokerage and investment banking

Brokerage refers to bringing together parties interested in making a transaction such as buying and selling shares of stock, for a fee charged to brokers. Investment banks operate brokerages, assist in the issuance of securities such as bonds and stocks, and trade securities and derivatives for their own accounts. Investment banks have large holdings of securities that they generally plan to hold only for short periods, and fund a large portion of their balance sheets with repurchase agreements, in addition to commercial paper and long-term debt.

During the Great Depression in the United States, bank failures and investment losses nearly destroyed the banking system and the national economy. In order to safeguard against such bank failures, the Glass–Steagall Act of 1933 prohibited banks from owning or affiliating with brokerage firms. The intention of the law was to create a “firewall” between speculative businesses and banking institutions. Even as banks became deregulated in the United States in the 1980s, the prohibition of direct securities trading and insurance underwriting continued. In 1999, however, the Gramm–Leach–Biley Act (GLBA) allowed banks to pursue insurance and securities businesses directly, essentially repealing the Glass–Steagall Act and eliminating legal barriers that had separated the securities, insurance, and banking industries. By the 2000s, several commercial bank holding companies in the United States, such as Citigroup and JPMorgan Chase, had significant investment banking subsidiaries. Today, many of the largest investment banks globally are part of commercial bank holding companies. On Bloomberg’s 2013 list of top 20 investment banks globally by fees, only four (Goldman Sachs, Morgan Stanley, Jefferies Group, and Lazard) were not affiliated with a commercial bank.

Finance companies

Finance companies supply credit to businesses and consumers, but they do not take deposits. Instead, they raise funds by issuing commercial paper and bonds. While they securitize some of their loans, some finance companies also hold assets on their balance sheet, in a similar way to a traditional bank. Finance companies’ business activities may include providing business loans secured by assets of the firm, making consumer loans for purchase of appliances or furniture, and offering credit cards. In the years leading up to the financial crisis of 2007–2008, finance companies in the United States were also offering mortgages.
Securitization

Securitization is the process by which a pool of assets, mainly loans including credit card debt, mortgages, and student loans, are repackaged into securities, mainly bonds. The interest and principal payments from such a pool of assets are used to pay the debt service on the securities. Securitization began in the 1970s, when home mortgages were pooled by US government-backed agencies Freddie Mac and Fannie Mae. Other income-producing assets began to be securitized in the 1980s, and in recent years the market has grown dramatically. Shadow banking entities are critical players in markets for securitized products. The shadow banking system facilitates the securitization of a substantial fraction of consumer and business credit, including mortgages, auto loans, student loans, and credit cards. Finance companies fund loans prior to securitization, in many cases through short-term lending markets, while the investment banks structure and sell securitized products. Commercial banks or their investment banking affiliates often participate in the origination and trading of securitized products, supplementing their traditional banking business revenues. Higher-risk tranches of these securities are purchased mostly by hedge funds and special purpose entities.

Money market funds

Money market funds (MMF) are mutual funds that invest in short-term securities such as government securities, commercial paper, certificates of deposit, and repurchase agreements. MMF appear similar to deposits at commercial banks, and investors can even write checks using the funds in the MMF. MMFs attempt to maintain a stable $1 net asset value (NAV). Sometimes a parent company will provide support in order to avoid the NAV declining below $1, also known as “breaking the buck.”

MMFs began appearing in the United States in the early 1970s, following regulatory limits on the interest payable on bank deposits. In the 1980s the first MMF was created in Europe for similar reasons: French asset managers offered the product with similar returns and characteristics as bank deposits, avoiding restrictions on interest paid. The MMF sector is an important funding source through their investment in short-term debt issued primarily by banks. Many banks in the United States and Europe obtain a substantial proportion of their funding from MMFs and rely on them for rolling over short-term debt.

Hedge funds

Hedge funds are investment companies that manage funds for high net worth individuals or institutions. Global estimates for the size of the hedge fund industry vary. Hedge Fund Research estimates that $2.6 trillion of capital was managed by the global hedge fund industry at the end of 2013. However, regulators in the United States and the United Kingdom, where the greatest number of hedge funds are believed to be domiciled, have reported $470 billion and $5 trillion, respectively, of hedge funds assets under management.

Hedge funds originated in the United States. US-based managers manage approximately 70% of global hedge fund assets while European and Asian managers hold 21% and 5% respectively. Hedge funds are major trading partners with the large commercial banks and investment banks. These relationships form yet another link from the shadow banking system to the traditional banking system which could spread a crisis to the mainstream banking sector.
Peer-to-peer lending

A variety of peer-to-peer lending sites have emerged around the globe, using the Internet to facilitate the borrower and lender matchmaking. Although specifics vary, common characteristics of peer-to-peer lending include using social networks to facilitate connections, allowing lenders to earn interest by participating as part of a larger loan, and selecting borrowers to finance on the basis of their personal narrative in addition to their credit ratings.

Mainstream and shadow banking
interdependence and the 2007–2008 financial crisis

Systemic risk can arise from the relationship between shadow banking entities and the mainstream banking sector. This interconnectedness can take many forms. Direct linkages are created when shadow banking entities form part of the bank intermediation chain, are directly owned by banks, or benefit directly from bank support (explicit or implicit). Funding interdependence is another form of direct linkage, as is the holding of each other’s assets such as debt securities. Indirect linkages also exist, when the two sectors invest in similar assets or are exposed to a number of common counterparties. These connections create a contagion channel through which stress in one sector can be transmitted to the other, as occurred during the 2007–2008 subprime mortgage crisis and the ensuing financial crisis.

When the shadow banking entities came under pressure in 2007 and 2008, under agreements that were triggered by disruptions in the market, major US and European commercial banks were required to provide funds to these entities. As the commercial banks themselves came under funding pressure, even commercial banks that had not provided such backstop agreements reduced their lending in the interbank and other money markets. Banks began to hoard cash, which further froze the funding markets.

In late 2008 a US investment bank, Lehman Brothers, came under funding pressure which eventually pushed it into bankruptcy. The failure of Lehman Brothers had repercussions for both mainstream and shadow banking systems. Withdrawals increased from money market funds, and hedge funds and funding markets were in turmoil. On September 25, 2008, a commercial bank, Washington Mutual Bank (WaMu), filed for bankruptcy in the largest bank failure in US history, the result of a run on WaMu that began on the day of Lehman’s bankruptcy. The failure of Lehman Brothers had repercussions across Europe and throughout the world as a contagion effect started to spread through the entire banking system.

The Federal Reserve established a series of funding facilities in 2008 in an effort to restore investor confidence and liquidity in the shadow banking system, effectively extending its LOLR role (traditionally reserved for commercial banks and savings institutions) to the shadow banking system. The US government also offered temporary insurance for money market fund investments in 2008, which was necessary to restart financing in the many nonfinancial sectors that depended on the shadow banking system. In 2010 the Dodd–Frank Wall Street Reform and Consumer Protection Act was enacted, prohibiting banks from owning, investing, or sponsoring hedge funds, private equity funds, or proprietary trading operations. This provision, also known as the Volcker Rule, is aimed specifically at reducing commercial banks’ shadow banking activities in an attempt to weaken the interconnectedness of mainstream and shadow banking.

The evolution and impact of the Great Recession have been spatially uneven. A myriad of geographic research focus includes the nature of the governmental responses to bank failures (e.g.,
Grossman and Woll 2013) and geographical differences in experiences of the crisis (e.g., Aalbers 2009; Hadjimichalis 2011; Schafran 2013).

**Regulating shadow banking**

Regulators around the world have stepped up their oversight of nonbank institutions since the 2007–2008 financial crisis, recognizing that entities from money market funds to private equity firms can have a big impact on economic stability. The increased globalization of financial markets, the deepening of financial integration, and the rapid pace of financial innovation have created a global financial system that transcends national boundaries, resulting in increased capital flows, lower borrowing costs, and better price discovery. However, the system has also created and strengthened a network transmitting negative shocks simultaneously to different regions and markets. The worldwide Great Recession served as a wake-up call for the risks of a globalized financial market, highlighting the challenges for international regulation.

There is an ongoing international effort to improve the global financial architecture and to ensure more coordination and information sharing between national and international regulators. In most economic or financial crises, national governments have reacted in isolation from each other by introducing new national regulatory rules. Unless an integrated global framework is successfully implemented, market participants will continue to bypass national rules through cross-border regulatory arbitrage. A study by the International Monetary Fund (IMF) found that the 2007–2008 financial crisis has “revealed important flaws in the current global architecture” (IMF, International Monetary Fund, 2009, 1), identifying four areas in which reforms are needed: (i) surveillance of systematic risk; (ii) international coordination of macro-prudential response to systematic risk; (iii) cross-border arrangements for financial regulation; and (iv) funding for liquidity support or external adjustment.

The emergence of the G20 as a forum to deal with the Great Recession has been a major development toward such global coordination in regulating both mainstream and shadow banking sectors. The G20 communiqué from London in April 2009 agreed to replace the Financial Stability Forum (FSF) with a new supranational Financial Stability Board (FSB). In collaboration with other international institutions such as the IMF, this board is responsible for safeguarding global financial stability. Its mandate, as stated on its website, includes the following: “assess vulnerabilities affecting the financial system and identify and oversee action needed to address them; promote coordination and information exchange among authorities responsible for financial stability” (FSB, Financial Stability Board, 2015). The establishment of the FSB as a supranational board, combined with an enhanced role for the IMF, and the close collaboration of these two institutions with the BIS are among the efforts made to improve the international financial architecture and to contribute to the process of global financial stability.

Within Europe, the European Systemic Risk Board was established in December 2010 to provide an early warning to national regulators and to ensure that they take joint responsibility for the supervision of major banks. It is tasked with developing a common set of quantitative and qualitative indicators to identify, measure, and monitor systemic risk, and with recommending remedial action in coordination with international financial organizations, particularly the IMF and FSB.

The highly interdependent nature of the twenty-first-century global financial market, with assets that can instantaneously move from
MAASTREAM AND SHADOW BANKING

one country to another or from one country to
an OFC, means that rules set out by international
institutions such as the FSB or the IMF must
be fully supported by national legislation and
laws throughout the world to be effective. There
remain challenges for the FSB in collaborating
with all OFCs as a way of removing regulatory
arbitrage. The policy debate following the Great
Recession has highlighted the role of OFCs, and
the need to limit the shadow banking sector’s
ability to channel money through these lightly
taxed havens. However, the concern
over OFCs should not detract from how the
long-standing rivalry between New York and
London for dominance in both mainstream and
shadow banking was a major factor behind the
2007–2008 financial crisis (French, Leyshon, and
Thrift 2009). Despite the perceived competition
from OFCs, mainstream and shadow banking
institutions, including hedge funds, are still tied
to a location (Wójcik 2013; Sennholz-Weinhardt
2014), with disincentives preventing them from
relocating to another jurisdiction in order to
benefit from looser regulation. Reforming regu-
lation in the long-standing international financial
centers of London and New York should there-
fore be the highest priority, even if uniform
regulation can’t be achieved across regions.

Although the literature on mainstream and
shadow banking continues to expand following the
Great Recession, some areas for further
research remain. With the emergence of global
banks such as HSBC and Barclays, which have
multiple global headquarters (Hong Kong and
London in the case of HSBC; London and
New York in the case of Barclays) and branches
across continents, alternative and comprehen-
sive measures of banking activity, beyond assets
under management and market value, should
be explored in different countries and in inter-
ationally comparative ways. Although China
now hosts the four largest banks as measured
by profits (Chen 2015), research on banking is
still focused on New York and London. Further
investigation should be performed on the finan-
cial, spatial, and social implications of the growth
of Chinese banks and its regional and global
implications. It is also important to focus on all
dimensions of banking. The Great Recession
has focused the research spotlight on subprime
mortgage lending, but other sectors of credit
intermediation (including small business, stu-
dent, auto, and personal loans) have significant
societal implications. Geographers and other
social scientists also should not ignore the study
of spaces that seem peripheral, such as markets
that fall under the generalization of “emerging”
despite each having unique attributes. These
include alternative banking activities such as
microcredit and Islamic finance, which mainly
serve households and firms in the Middle East,
South Asia, and Africa. Qualitative and discursive
analyses, including case studies, could add to the
body of existing research given the dissonance
in scale and spatiality. Finally, the relationship
between mainstream and shadow banking and
the state is also an important avenue for future
research. Academic discourse on the regula-
tion of both mainstream and shadow banking
has increased since the Great Recession, but
empirical research on the interlinkages between
banking and government is relatively lacking.

SEE ALSO: Corporate financialization;
Corporations and the nation-state; Economic
geography; Financial geography; Globalization;
Regulation

References

Aalbers, M. 2009. “Geographies of the Financial Cri-
sis.” Area, 41: 34–42.


Manufacturing industry

Chun Yang
Hong Kong Baptist University, China

Does manufacturing still matter?

An essential component of the industrial sector of an economy, “manufacturing industry” refers to processes involving the transformation of raw materials into finished goods for sale, or intermediate processes involving the production or finishing of semifinished products. Unlike agriculture and mining, manufacturing takes place under human-made and controllable conditions. Unlike service activities, manufacturing provides vital material inputs into the continuation of human life.

For nearly three centuries, since the Industrial Revolution, manufacturing industries have helped drive economic growth and rising living standards; they continue to do so in developing economies. Global manufacturing value added grew from $5.7 trillion in 2000 to $7.5 trillion (in 2000 prices) in 2010, which represented 16% of global GDP. Both advanced and developing economies have experienced growth in manufacturing value added. Nevertheless, the role of manufacturing in global and national economies has evolved over time. Empirical evidence shows that as economies become wealthier and reach middle-income status, manufacturing’s share of GDP peaks (at 20–35% of GDP). Beyond that point, consumption shifts toward services, hiring in services outpaces job creation in manufacturing, and manufacturing’s share of GDP begins to fall along an inverted “U” curve. It is evident that manufacturing will continue to matter a great deal to both developing and advanced economies. However, the way manufacturing contributes to national economies and competitiveness changes over time as economies grow.

Moreover, the definition of manufacturing is changing. Manufacturing employs a rising number of workers in nonproduction jobs, and service inputs represent a rising share of manufacturing output. Manufacturing industries are increasingly large users of service inputs and employ many workers in service roles. Service-like activities – such as R&D, marketing and sales, and customer support – have become a larger share of what manufacturing companies do: 30–55% of manufacturing jobs in advanced economies are service-type functions, and service inputs make up 20–25% of manufacturing output (McKinsey 2012, 7). The distinctions between manufacturing and services have blurred.

Globalization of manufacturing industry

Over the past several decades, manufacturing industry has experienced significant changes as rapid globalization shifted a substantial proportion of manufacturing capacity from developed economies to the “newly industrializing economies”: that is, Hong Kong, Singapore, South Korea, and Taiwan, in the 1960s, and then to developing economies, for example, China and Malaysia in Asia and Mexico in Latin America, in the 1970s and 1980s. In consequence, global production has become increasingly fragmented and different stages of
production are now regularly carried out in different countries. This great unbundling of tasks, also known as fragmentation, offshoring, or vertical specialization, has implications for the organization and coordination of manufacturing activities around the world. Through the trade of intermediate goods and services, global production networks developed quickly in manufacturing industries such as textiles, automotive, and electronics. Traditionally, the global technologies/innovators group has been led by companies from advanced economies, such as Apple and Hewlett-Packard in the United States, Fujitsu, Hitachi, and Toshiba in Japan, and Ericsson, Nokia, Philips, and Siemens in Europe. Advanced economies contribute 70% of the global value added, and the United States retains the lead, with 27% of the group’s global value added in 2010. It is, however, worth noting that South Korea, with companies such as LG and Samsung, and Taiwan, with Acer and AsusTek, also have strong positions and have changed from followers to market leaders (Yeung 2007).

Globally, manufacturing output (measured by gross value added) continued to grow: between 2000 and 2007, by 2.7% annually in advanced economies and by 7.4% in large developing economies. Economies like China, India, and Indonesia have risen into the world’s top 15 largest manufacturing economies, according to their share of global nominal manufacturing gross value added (Table 1).

In order to reflect the increasing fragmentation of global manufacturing, the World Trade Organization (WTO) launched “Made in the World” in 2011. For a number of electronic products (iPods and laptops) manufactured in China, less than 3% of the export value actually goes to the manufacturers/assemblers in China, for example, Foxconn (Dedrick, Kraemer, and Linden 2010). The major part of the value is captured by firms in the United States, Japan, Korea, and Taiwan through the delivery of sophisticated intermediate inputs.

The future of manufacturing industry

Today’s global landscape of manufacturing is more complicated than that of a few decades ago. Although there are clear elements of continuity, dramatic changes have occurred. The overall trajectory of global manufacturing has become increasingly volatile. Within this uneven trajectory there has been major geographical reconfiguration. As claimed by Dicken (2011), a relatively simple global geographical division of labor has been replaced by a far more complex and multiscalar structure. The dominant factors that shaped the disaggregated supply chains today will not be the same as those through the next several decades, because the global environment is changing. On the one hand, many emerging economies used by transnational corporations (TNCs) as locations of low-cost labor have developed significant manufacturing and innovation capabilities, which allow them to produce increasingly advanced manufactured products. On the other hand, these economies have begun to experience a corresponding escalation in wages and costs, following in the footsteps of their developed nation counterparts decades ago. Greater prosperity and higher wages are helping to drive an increased ability and desire to consume by the growing middle classes, making them an exciting market of new consumers: for example, China and India.

Until the past decade, globalization of manufacturing was driven largely by cost arbitrage – essentially the search for low-cost production locations. Today, however, a more important driver for multinational manufacturing companies is to match production footprints to patterns of demand growth. At the same time,
Table 1  Top 15 manufacturers (% of global nominal manufacturing gross value added).

<table>
<thead>
<tr>
<th>Rank</th>
<th>1980</th>
<th>1990</th>
<th>2000</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>United States</td>
<td>United States</td>
<td>United States</td>
<td>United States</td>
</tr>
<tr>
<td>2</td>
<td>Germany</td>
<td>Japan</td>
<td>Japan</td>
<td>China</td>
</tr>
<tr>
<td>3</td>
<td>Japan</td>
<td>Germany</td>
<td>Germany</td>
<td>Japan</td>
</tr>
<tr>
<td>4</td>
<td>United Kingdom</td>
<td>Italy</td>
<td>China</td>
<td>Germany</td>
</tr>
<tr>
<td>5</td>
<td>France</td>
<td>United Kingdom</td>
<td>United Kingdom</td>
<td>Italy</td>
</tr>
<tr>
<td>6</td>
<td>Italy</td>
<td>France</td>
<td>Italy</td>
<td>Brazil</td>
</tr>
<tr>
<td>7</td>
<td>China</td>
<td>China</td>
<td>France</td>
<td>South Korea</td>
</tr>
<tr>
<td>8</td>
<td>Brazil</td>
<td>Brazil</td>
<td>South Korea</td>
<td>France</td>
</tr>
<tr>
<td>9</td>
<td>Spain</td>
<td>Spain</td>
<td>Canada</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>10</td>
<td>Canada</td>
<td>Canada</td>
<td>Mexico</td>
<td>India</td>
</tr>
<tr>
<td>11</td>
<td>Mexico</td>
<td>South Korea</td>
<td>Spain</td>
<td>Russia</td>
</tr>
<tr>
<td>12</td>
<td>Australia</td>
<td>Mexico</td>
<td>Brazil</td>
<td>Mexico</td>
</tr>
<tr>
<td>13</td>
<td>Netherlands</td>
<td>Turkey</td>
<td>Taiwan</td>
<td>Indonesia</td>
</tr>
<tr>
<td>14</td>
<td>Argentina</td>
<td>India</td>
<td>India</td>
<td>Spain</td>
</tr>
<tr>
<td>15</td>
<td>India</td>
<td>Taiwan</td>
<td>Turkey</td>
<td>Canada</td>
</tr>
</tbody>
</table>

Italics show the rising positions of India, Indonesia, and China (see text).

an increasingly diverse global customer base creates a demand for a broad set of localized, tailored products, also driving the need for proximity to customers. These shifting demand trends affect all manufacturing industries, although to differing degrees and on different timelines, and conventional policy is not adequate to deal with them. Demand is shifting to emerging markets at an accelerated rate and is becoming more fragmented. According to McKinsey’s (2012) research, consumption by developing economies could rise from $12 trillion annually in 2010 to $30 trillion in 2025. By 2025, it is estimated that developing economies could account for nearly 70% of global demand for manufactured goods. Demand shifts to emerging markets are being driven not just by large economies such as those of China and India, but also by economic growth in Indonesia, Kenya, Vietnam, and other smaller emerging markets.

Rising wage costs are most likely to affect industries in the labor-intensive tradable group and the assembly steps in global technologies/innovators’ businesses – where labor is a relatively large fraction of compressible costs. From 2005 to 2010, wage hikes averaged 10% per year. In 2010, the giant contract manufacturer Foxconn, which employs 920,000 people in China alone, doubled wages in its Shenzhen plant. Companies typically respond to rising wages by moving to lower-cost locations. This presents opportunities for “next frontier” developing economies to capture any labor-intensive work from countries with rapidly rising wages, such as China. Countries such as Bangladesh, Cambodia, Indonesia, Vietnam, and other developing economies are already experiencing
growth in labor-intensive industries because of their cost advantages. While the manufacturing of some goods will shift from China to nations with lower labor costs, these nations’ ability to absorb the higher-end manufacturing will be limited by inadequate infrastructure, a lack of skilled workers, scale, and domestic supply networks, as well as by political and intellectual property risks. Lower worker productivity, higher corruption, and greater risk to personal safety are also concerns for the relocation of manufacturing to these nations.

The future of manufacturing is unfolding in an environment of changing dynamics and greater risk and uncertainty than before. An exciting new era of global manufacturing is ahead – driven by shifts in demand and by innovations in materials, processes, information technology, and operations. The prospect is a more “global” manufacturing industry (Dicken 2011), in which developing economies are the source of new customers as well as the source of low-cost production. However, these opportunities arise amid a global shift in the cost and availability of factor inputs (e.g., labor and natural resources) and rising complexity, uncertainty, and risk. As manufacturing evolves, policymakers must adjust their expectations and look at manufacturing not only as a source of mass employment in traditional production work but also as a critical driver of innovation, productivity, and competitiveness. In the aftermath of the 2008 global financial crisis and world recession, the restructuring of global manufacturing is taking place. The new era of manufacturing will unfold in an environment in which old assumptions, strategies, and policies no longer suffice. Both manufacturing leaders and policymakers need to think and act in new ways, and develop new strategies.

SEE ALSO: Global factory; Global production networks; Industrial restructuring; Industrialization

References


Further reading


Map algebra

C. Dana Tomlin
University of Pennsylvania, USA

Map algebra is the language of cartographic modeling. Introduced in the early 1980s (Tomlin 1983), it is a computational pseudocode intended to represent the concepts, conventions, and capabilities of that methodology in terms that are concise, consistent, and comprehensive yet legible to a wide range of users. The result is a form of notation resembling that of assignment statements in imperative programming languages.

The capabilities represented by map algebra are common to fields ranging from remote sensing and geoprocessing to spatial statistics and digital photography. Map algebra expresses these capabilities by relating them to a coherent data-processing framework intended to make them more readily understood, more easily combined, more creatively utilized, and more likely to be extended. Among its extensions to date have been versions specifically oriented toward binary images, linear networks, areal features, vector fields, volumetric conditions, temporal phenomena, and relational data models.

The general nature, structure, and role of map algebra are introduced in this volume’s entry on cartographic modeling. By building on that foundation, the present entry offers more detailed descriptions of individual functions and their implications. To do so, it introduces the basic set of elementary functions and then illustrates how these can be used to generate a variety of compound functions and, thereby, perform more sophisticated and more specialized tasks.

The elementary functions of map algebra were initially implemented and are most fully developed for use in conjunction with geospatial data that are encoded in a raster format. Each operates on variables that are organized as shown in Figure 1.

Every function accepts one or more layers as its input variable(s) and generates a single new layer as its output. Since the output of any one of these functions can therefore be used as input to any other, functions can be easily combined, and

![Figure 1 Map algebraic variables. All of the map algebraic functions operate on variables in the form of “layers,” each of which depicts a geographical area as a bounded plane on which every “location” is defined by a particular pair of planar coordinates. Each location is also associated with exactly one of a set of a “zones” that are uniquely represented by numerical “values.” The zones of any given layer generally depict variations on a single geographical characteristic such as soil type, population density, or proximity to home.](image-url)
none of them therefore has to be complex. The elementary functions are, in fact, deliberately designed to minimize duplication and to avoid undue complexity.

The result is a set of some fifty functions that are associated with three major groups termed “local,” “zonal,” and “focal.” Among the focal functions, which operate on locations within a specified vicinity or “neighborhood” around any given location, four subgroups can further be distinguished according to whether those neighborhoods are “lateral,” “visible,” “accessible,” or “local” in nature. As indicated in Figure 2, this gives rise to three (or, more precisely, six) major types of map algebraic function.

![Map Algebra Functions](image)

**Figure 2** The major types of map algebraic function. Each of the map algebraic functions calculates new values for all locations (depicted as dots) such that the new value at any given location (red dot) is computed from the existing values of a particular set of locations (yellow dots). Local functions compute each location’s new value from its own existing value(s). Zonal functions use the values of locations sharing a common attribute. Focal functions consider the values of neighboring locations in one of four types of neighborhood. Lateral neighborhoods are defined by distance and/or direction to each neighborhood center or “focus” (red dot), while visible neighborhoods also require that neighbors lie within sight. Accessible neighborhoods are defined in terms of travel costs, and local neighborhoods only encompass immediately adjacent neighbors.
Local functions

Each of the local functions generates a new layer on which every location is set to a value computed from the value(s) of that same location on one or more designated layers. Local functions are specified as:

\[
\text{NEWLAYER} = \text{LocalFUNCTION of } \text{1STLAYER} \text{ and NEXTLAYER}
\]

where

- **NEWLAYER** specifies the new layer to be generated;
- **1STLAYER** specifies an existing layer of values to be processed;
- **NEXTLAYER** specifies another existing layer of values to be processed (in a phrase that may be used repeatedly);
- **LocalFUNCTION** specifies that new values are to be computed as follows:
  - **LocalSelection** sets each location to a value of 1 (for true) or 0 (for false) that indicates the verity of one or more specified statements asserting that the value of that location on one existing layer is greater than, less than, and/or equal to its value on another existing layer;
  - **LocalRating** sets each location to a value explicitly specified (either as a numeral representing that value or the name of an existing layer from which the location’s new value is to be drawn) according to the value(s) of that location on one or more existing layers;
  - **LocalMix** sets each location to a value that identifies the particular combination of its values on **1STLAYER** and one or more **NEXTLAYERs**;
  - **LocalVariety** sets each location to a value indicating the number of dissimilar values associated with that location on **1STLAYER** and one or more **NEXTLAYERs**;
  - **LocalMajority** sets each location to the most frequently occurring of its values on **1STLAYER** and one or more **NEXTLAYERs**;
  - **LocalMinority** sets each location to the least frequently occurring of its values on **1STLAYER** and one or more **NEXTLAYERs**;
  - **LocalAmount** sets each location to a value that indicates how often its **1STLAYER** value matches that of a **NEXTLAYER**;
  - **LocalRank** sets each location to a value that identifies the interval containing its **1STLAYER** value when its **NEXTLAYER** values are divided into a specified series of interval;
  - **LocalMean** sets each location to the average of its values on **1STLAYER** and one or more **NEXTLAYERs**;
  - **LocalMaximum** sets each location to the highest of its values on **1STLAYER** and one or more **NEXTLAYERs**;
**MAP ALGEBRA**

- **LocalMinimum** sets each location to the lowest of its values on 1STLAYER and one or more NEXTLAYERS;
- **LocalCalculation** sets each location to a value computed by subjecting its values on a specified set of existing layers to one or more standard mathematical functions;
- **LocalInteger** sets each location to a value computed by rounding or truncating its 1STLAYER value to integer form;
- **LocalBitSelection** sets each location to a value computed by subjecting its values on a specified set of existing layers to one or more standard bitwise operations.

An example of the LocalCalculation function is presented in Figure 3.

**Figure 3** The LocalCalculation function. DistanceInBetween is a layer on which each location’s value indicates the total distance from one destination to another (red dots) through that location. This layer was created by subjecting layers of planar distance from those two destinations (Distance and MoreDistance) to a function specified as follows:

\[
\text{DistanceInBetween} = \text{LocalCalculation of Distance} + \text{MoreDistance}
\]
Zonal functions

Each of the zonal functions generates a new layer on which every location is set to a value computed from the values of all locations that share its zone on a designated layer. Zonal functions are specified as:

\[
\text{NEWLAYER} = \text{ZonalFUNCTION of} \quad 1\text{STLAYER} \quad \text{within} \quad 2\text{NDLAYER}
\]

where

- **NEWLAYER** specifies the new layer to be generated;
- **1STLAYER** specifies an existing layer of values to be processed;
- **2NDLAYER** specifies another existing layer defining the zones within which **1STLAYER** values are to be considered;
- **ZonalFUNCTION** specifies that new values are to be computed as follows:
  - **ZonalRating** sets each location to a value explicitly specified (either as a numeral representing that value or the name of an existing layer from which the location’s new value is to be drawn) according to the set of **1STLAYER** values within its **2NDLAYER** zone;
  - **ZonalMix** sets each location to a value that identifies the particular combination of **1STLAYER** values within its **2NDLAYER** zone;
  - **ZonalVariety** sets each location to a value indicating the number of dissimilar **1STLAYER** values within its **2NDLAYER** zone;
  - **ZonalMajority** sets each location to the most frequently occurring of the **1STLAYER** values within its **2NDLAYER** zone;
  - **ZonalMinority** sets each location to the least frequently occurring of the **1STLAYER** values within its **2NDLAYER** zone;
  - **ZonalAmount** sets each location to a value indicating how often its **1STLAYER** value matches that of another location within its **2NDLAYER** zone;
  - **ZonalRank** sets each location to a value identifying the interval containing its **1STLAYER** value when all of the locations within its **2NDLAYER** zone are divided into a specified series of intervals by **1STLAYER** value;
  - **ZonalMean** sets each location to the location-wise average of **1STLAYER** values within its **2NDLAYER** zone;
  - **ZonalMaximum** sets each location to the highest of the **1STLAYER** values within its **2NDLAYER** zone;
  - **ZonalMinimum** sets each location to the lowest of the **1STLAYER** values within its **2NDLAYER** zone;
MAP ALGEBRA

**Figure 4** The $ZonalMinimum$ function. $DistanceToEachHill$ is a layer on which each location’s value indicates the distance from a particular destination (red dot) to the hill, if any, containing that location. This layer was created by subjecting a layer of distance from that destination ($Distance$) and a layer of individual hills ($EachHill$) to a function specified as follows:

$$DistanceToEachHill = ZonalMinimum\ of\ Distance\ within\ EachHill$$

$ZonalSum$ sets each location to the location-wise sum of $1STLAYER$ values within its $2NDLAYER$ zone.

An example of the $ZonalMinimum$ function is presented in Figure 4.

**Focal functions of lateral neighbors**

Each of the focal functions generates a new layer on which every location is set to a value computed from the values of its neighbors within a designated neighborhood. Lateral neighborhoods encompass locations that lie at specified planar distances and/or directions from each
neighborhood focus. Focal functions of lateral
neighborhoods are specified as

\[
\text{NEWLAYER} = \text{FocalFUNCTION of}
\]

\[
\text{1STLAYER at DISTANCE by DIRECTION}
\]

where

\text{NEWLAYER} \quad \text{specifies the new layer to be generated;}

\text{1STLAYER} \quad \text{specifies the existing layer of values to be processed;}

\text{DISTANCE} \quad \text{specifies the distance(s) bounding each neighborhood (which can vary from one location to another);}

\text{DIRECTION} \quad \text{specifies the direction(s) bounding each neighborhood (which can also vary from one location to another);}

\text{FocalFUNCTION} \quad \text{specifies that new values are to be computed as follows.}

\text{FocalRating} \quad \text{sets each location to a value explicitly specified (either as a numeral representing that value or the name of an existing layer from which the location’s new value is to be drawn) according to the set of 1STLAYER values within its neighborhood;}

\text{FocalMix} \quad \text{sets each location to a value identifying the particular combination of 1STLAYER values within its neighborhood;}

\text{FocalVariety} \quad \text{sets each location to a value indicating the number of dissimilar 1STLAYER values within its neighborhood;}

\text{FocalMajority} \quad \text{sets each location to the most frequently occurring of the 1STLAYER values within its neighborhood;}

\text{FocalMinority} \quad \text{sets each location to the least frequently occurring of the 1STLAYER values within its neighborhood;}

\text{FocalAmount} \quad \text{sets each location to a value indicating how often its 1STLAYER value matches that of another location within its neighborhood;}

\text{FocalRank} \quad \text{sets each location to a value identifying the interval containing its 1STLAYER value when all of the locations within its neighborhood are divided into a specified series of intervals by 1STLAYER value;}

\text{FocalMean} \quad \text{sets each location to the location-wise average of 1STLAYER values within its neighborhood;}

\text{FocalMaximum} \quad \text{sets each location to the highest of the 1STLAYER values within its neighborhood;}

\text{FocalMinimum} \quad \text{sets each location to the lowest of the 1STLAYER values within its neighborhood;}

\text{FocalSum} \quad \text{sets each location to the location-wise sum of}
MAP ALGEBRA

1STLAYER values within its neighborhood;

FocalProximity sets each location to its distance from the nearest of a set of neighbors identified by their 1STLAYER values;

FocalBearing sets each location to its direction from the nearest of a set of neighbors identified by their 1STLAYER values;

FocalNeighbor sets each location to the value of the nearest of a set of neighbors identified by their 1STLAYER values;

FocalConvolution sets each location to a position-weighted average of 1STLAYER values within its neighborhood;

FocalInterpolation sets each location to a distance-weighted average of the 1STLAYER values within its neighborhood;

FocalDistribution sets each location to a value computed by summing contributions from selected neighbors when each of those neighbors distributes its 1STLAYER value to surrounding locations in distance-weighted proportions;

FocalInsularity sets each location to a value that distinguishes it as part of an insular (or near-insular) group of locations that share the same 1STLAYER value.

Examples of the FocalBearing and FocalProximity functions are presented in Figure 5.

The Distance layer shown in Figures 3 and 4 was generated with a similar function specified as follows.

\[ \text{Distance} = \text{FocalProximity of Destination} \]

where

\text{NEWLAYER}, \text{1STLAYER}, \text{DISTANCE}, \text{DIRECTION}, \text{and FocalFUNCTION}

\text{SURFACE LAYER}

\text{specifies an existing layer on which each location's value indicates its elevation on an opaque (usually topographic) surface;}

\text{are all as described for focal functions of lateral neighborhoods (except for the absence of FocalConvolution)}

\text{Visible neighborhoods encompass locations that lie at designated planar distances and/or directions from each neighborhood focus and also within visual contact of that focus. Focal functions of visible neighborhoods are specified as:}

\text{NEWLAYER}

\text{= FocalFUNCTION of 1STLAYER at DISTANCE by DIRECTION radiating on SURFACE LAYER from TRANSMISSION LAYER through OBSTRUCTION LAYER to RECEPTION LAYER}
Figure 5  The *FocalBearing* and *FocalProximity* functions. *Direction* is a layer on which each location’s value indicates the compass direction from that location toward a particular destination (red dot). This layer was created by subjecting a layer depicting that destination (*Destination*) to a function specified as follows:

\[ \text{Direction} = \text{FocalBearing of Destination} \]

*TRANSMISSIONLAYER* specifies an existing layer on which each location’s value indicates the height above *SURFACE LAYER* at (or below) which that location is visually opaque; *RECEPTIONLAYER* specifies an existing layer on which each location’s value indicates the height above *SURFACE LAYER* at (or below) which that location is able to establish visual contact with other locations.

*OBSTRUCTIONLAYER* specifies an existing layer on which each location’s value indicates the height above *SURFACE LAYER*.
Visibility is a layer on which each location’s value indicates the distance to that location from a particular destination (red dot) but only if the location lies within visual contact of that destination. This layer was created by subjecting a layer depicting the destination (Destination) and a layer of topographic elevation (Elevation) to a function specified as follows:

\[
\text{Visibility} = \text{FocalProximity of Destination radiating on Elevation}
\]

If any of SURFACELAYER, TRANSMISSION-LAYER, OBSTRUCTIONLAYER, or RECEPTIONLAYER is unspecified, the layer assumed is one on which every location’s value is set to 0.

An example of one of the FocalProximity radiating function is presented in Figure 6.

Focal functions of accessible neighbors

Accessible neighborhoods encompass locations that lie at designated distances from each neighborhood focus, but where distances are measured in terms of travel cost. Focal functions of accessible neighborhoods are specified as:

\[
\text{NEWLAYER} = \text{FocalFUNCTION of 1STLAYER at DISTANCE by DIRECTION spreading in FRICTIONLAYER on SURFACELAYER through NETWORKLAYER}
\]
where

**NEWLAYER**, **1STLAYER**, **DISTANCE**, and, **FocalFUNCTION** are all as described for focal functions of lateral neighborhoods (except for the absence of **FocalConvolution**);

**FRICTIONLAYER** specifies an existing layer on which each location’s value indicates the amount of travel cost per unit of physical distance at that location;

**SURFACELAYER** specifies an existing layer on which each location’s value indicates its elevation on what is usually a topographic surface such that the physical distance between that location and any of its adjacent neighbors accounts for their vertical as well as their horizontal positions;

**NETWORKLAYER** specifies an existing layer on which each location’s value indicates which of its adjacent neighbors are directly accessible from that location.

If any of **FRICTIONLAYER**, **SURFACE-LAYER**, or **NETWORKLAYER** is unspecified, the layer assumed is one on which every location’s value is set to 0.

An example of the **FocalProximity spreading** function is presented in Figure 7.

**Focal functions of local neighbors**

Local neighborhoods encompass only those locations that lie on or adjacent to each neighborhood focus, and focal functions of these neighborhoods use them to characterize their foci as parts of larger geometric forms. Those forms may be 1-, 2- or 3-D in nature, and the value computed for each location within a given form generally indicates its unique contribution to the overall measurement of a geometric characteristic such as length, area, or volume. Focal functions of local neighborhoods are specified as:

\[
\text{NEWLAYER} \cdot \text{FocalFUNCTION of } 1\text{STLAYER} \quad \text{on } \text{SURFACELAYER}
\]

where

**NEWLAYER** specifies the new layer to be generated;

**1STLAYER** specifies an existing layer depicting 1- or 2-D forms;

**SURFACELAYER** specifies an existing layer depicting 3-D form (assumed to be flat if this layer is unspecified);

**FocalFUNCTION** specifies that new values are to be computed as follows:

- **FocalLinkage** sets each location to a value identifying the configuration of linear connections occurring at that location on **1STLAYER**;
- **FocalLength** sets each location to the total length of all linear connections occurring at that location on **1STLAYER** when projected onto a **SURFACELAYER** surface;
**Figure 7** The *FocalProximity spreading* function. *TravelCost* is a layer on which each location’s value indicates the cost of travel required to reach to that location from a particular destination (red dot). This layer was created by subjecting a layer depicting that destination (*Destination*) and a layer indicating the rates at which travel costs are incurred (*CostPerStep*) to a function specified as follows:

\[
TravelCost = \text{FocalProximity of Destination spreading in CostPerStep}
\]

- **FocalPartition** sets each location to a value identifying the configuration of areal boundaries occurring at that location on *1STLAYER*; that location’s portion of *1STLAYER* when projected onto a *SURFACELAYER* surface;
- **FocalFrontage** sets each location to the total length of all boundaries occurring at that location on *1STLAYER* when projected onto a *SURFACELAYER* surface; sets each location to a value indicating the surficial steepness at that location on a *SURFACELAYER* surface partitioned into *1STLAYER* zones;
- **FocalArea** sets each location to the surface area associated with that location’s portion of *1STLAYER* when projected onto a *SURFACELAYER* surface; sets each location to a value indicating the compass direction of surficial descent;
The *FocalDrainage* function. *TravelDirection* is a layer on which each location’s value indicates the direction of the minimum cost path to that location from a particular destination (red dot). This layer was created by subjecting a layer depicting the cost of travel from that destination (the *TravelCost* layer presented in Figure 7) to a function specified as follows:

\[
TravelDirection = FocalDrainage \text{ of } TravelCost
\]

An example of one of the *FocalDrainage* function is presented in Figure 8.

Also shown for comparison is a layer that was created as follows from the *Elevation* layer presented in Figure 6.

\[
DownhillDirection = FocalDrainage \text{ of } Elevation
\]
Convexity is a layer on which each location’s value indicates the degree to which a topographic surface is convex or concave in the vicinity of that location. This layer was created by subjecting a layer of topographic elevation (the *Elevation* layer presented in Figure 6) to a sequence of functions specified as follows:

\[
\text{MeanElevation} = \text{FocalMean of Elevation at } \ldots 100
\]

\[
\text{Convexity} = \text{LocalCalculation of Elevation} - \text{MeanElevation}
\]

The *EachHill* layer shown in Figure 4 was then created as follows:

\[
\text{AllHills} = \text{LocalSelection of Convexity} > 30
\]

\[
\text{EachHill} = \text{FocalInsularity of AllHills}
\]

**Compound Functions**

While it is easy to argue for additions to the elementary map algebraic functions, it is often more useful to note how such functions can be composed from those already available. Consider, for example, how elementary functions might be used to create general-purpose compound functions with names like *LocalAbsoluteValue*, *ZonalRange*, or *FocalStandardDeviation*. Of course, more specialized sequences of
elementary functions can also be composed, stored, and invoked as compound functions.

To demonstrate the power of these functions and the kind of thinking involved in their creation, consider a contrived example. It begins as described in Figure 9 by generating a layer of hills. As illustrated in Figure 10, it then goes on to indicate which of each hill’s internal locations is (or are) closest to a specified destination. Finally, as illustrated in Figure 11, each location is set to a value indicating whether or not any of those selected hill locations lie upstream on a 3-D surface of travel-cost proximity to that destination.

The technique used here to delineate minimum cost paths effectively drains those selected locations over the 3-D surface of travel costs just as rainfall might be drained over a 3-D surface of topographic elevation. Much of the power of map algebra relates to this sort of connection.

Figure 10  A ZonalLowPoint function. HillPoints is a layer that identifies the location (black dot) within each of a set of (tan) hills that is closest to a given destination in terms of travel cost. This layer was created by subjecting a layer of hills (the EachHill layer described in Figure 9) and a layer of travel cost (the TravelCost layer presented in Figure 7) to a sequence of functions specified as follows:

$$
\text{HillCost} = \text{ZonalMinimum of TravelCost within EachHill}
$$

$$
\text{HillPoints} = \text{LocalSelection of TravelCost == HillCost}
$$
Figure 11  A FocalRouting function. Paths is a layer delineating the minimum cost routes (red lines) from a set of origins (black dots) to a particular destination (red dot). This layer was created by subjecting a layer of those origins (the HillPoints layer described in Figure 10) and a layer of directions (the TravelDirection layer presented in Figure 8) to a sequence of functions specified as follows:

$$\text{Traffic} = \text{FocalSum of HillPoints spreading through TravelDirection}$$

$$\text{Paths} = \text{LocalSelection of Traffic} > 0$$

between seemingly unrelated phenomena. By transcending particulars and focusing instead on the general nature of geospatial variables and functions, it is often possible to grasp concepts and develop insights more effectively than might otherwise be likely.

Consider again, for example, the apparent similarity between those surfaces of elevation and travel cost. Applying the FocalGradient function to a layer of topographic elevation would yield a layer of topographic steepness. Applying it to a layer of travel costs would yield a layer very much like the FRICTIONLAYER layer from which travels costs were computed by using FocalProximity spreading. The reason for this reflects on the fundamental nature of the FocalProximity spreading.

Also consider again the apparent similarity between layers of travel cost and planar distance. In Figure 3 is a layer of ellipses between two
Figure 12 The cost of traveling from here to there at locations in between. \( \text{TravelCostInBetween} \) is a layer on which each location’s value indicates the total cost of travel from one destination to another (red dots) through that location. This layer was created by subjecting layers of travel cost from those two destinations (\( \text{TravelCost} \) and \( \text{MoreTravelCost} \)) to a function specified as follows:

\[
\text{TravelCostInBetween} = \text{FocalCalculation of TravelCost} + \text{MoreTravelCost}
\]

destinations that was generated by adding each location’s distances from those destinations. In Figure 12 is the travel-cost equivalent: a rather different pattern but one that also arises from the measurement of in-betweenness.

SEE ALSO: Cartographic modeling; Data structure, raster; Geographic information system; Models and simulation in biogeography; Quantitative methodologies; Spatial analysis

Reference


Further reading

Map generalization

William A. Mackaness
University of Edinburgh, UK

Dirk Burghardt
Dresden University of Technology, Germany

Cécile Duchêne
Laboratoire COGIT IGN, Paris

Reasons for generalizing

The ambition

Many tasks require the use of maps (whether in digital or paper form). They range in scale from the highly detailed (1:500), through the commonly used midscale topographic range (1:50,000), to the world maps found in atlases (1:50 million). There is increasing thematic diversity across those scales, fueled by an expectation that anyone can collect their own data, integrate it with others, quickly produce their own map (a “mashup”), and share it across the web. Now everyone is a cartographer (the so-called democratization of cartography) and can participate in solving the challenges of the world (cybercartography).

In reality it is difficult to source and integrate data at appropriate levels of detail (scale); users have poor cartographic knowledge and risk producing dreadful maps that mislead rather than inform. Whilst the default symbology of geographical information systems (GIS) might go some way to facilitating the creation of maps, GIS cannot deal with the notion that map design is a complex decision-making process – very much one of finding compromise among a competing set of constraints. The complexity of this design task was why large teams of highly trained and experienced human cartographers were required in order to produce series mapping at a wide range of scales; their manual solutions reflected a deep knowledge of the map generalization process and the ways and means by which geographic entities might best be illustrated at different scales. In the computer age, the human cartographer was seen as something of an anachronism and an impediment to high levels of automation. People sought ways to reduce the involvement of humans in the map generalization process by trying to capture and embed their “know how” within automated mapping systems. Thus the question became: Can we capture the map requirements of the user, automatically select appropriate contextualizing data, integrate those data and visualize them, at an appropriate scale, with clarity, taking into consideration aesthetic considerations, whilst concealing the complexity of this task from the user? This ambition very much defines map generalization research today. Can map generalization automatically produce maps at a range of scales with minimum human intervention (Figure 1)?

A definition

The opportunities and challenges of automated mapping first became apparent in the 1960s. The initial focus was on scanning and digitizing paper maps produced by national mapping agencies (NMAs). For example the Ordnance Survey, the NMA of Great Britain, completed this task in the 1970s. Typically, NMAs produced map series
at specific scales and each series had its own production environment independent of other map series. This “silo” approach to map production was echoed in the digital environment but it was quickly realized that there was huge redundancy in this approach. The question became: Could we not capture and represent the world at the finest scale, and from this fine scale derive more generalized representations? For example, could algorithms be developed whereby the streets, pavements, gardens, and individual houses could be grouped and represented as suburbs? Could these in turn be combined with representations of the commercial districts and green spaces in order to represent cities, which in turn might eventually be represented as labeled dots on a map showing an entire country? In such a context, changes made at the finest scale would be automatically reflected at the coarse scale. Free of the shackles of paper-based, map series production, that vision has been further expanded, reflecting a context of “scale-free” web-based mapping, in which the user can interact with the data in real time, pinch and swipe (zoom and “slide”) across the map using any device.

Early research into map generalization had topographic mapping as its focus in anticipation of an NMA’s paper-based requirements. Nowadays that context lies more with exploratory analysis and in anticipation of users who have limited appreciation of map production and design. It seeks to embrace the open source vision and accommodate the ambition of being able to mix together multiple sources of data (both formal and user-generated), captured at fundamentally different scales. The definition given by the International Cartographic Association in 1973 is that map generalization is concerned with “the selection and simplified representation of detail appropriate to the scale and/or the purpose of the map.” At its heart, map generalization is about the process of abstraction, conveying patterns and associations among a set of geographic phenomena. The context of use has changed dramatically but the ambition remains the same – namely, to efficiently communicate a particular geography of the world.

**Reasons for automating the map generalization process**

The first motivation for automating generalization comes from the need to produce maps quickly, and at low cost. For decades, manual generalization was used to produce topographic or more thematic maps, but this was time-consuming (expensive) and the typical 5–10-year update cycles were no longer compatible with current standards of data currency and productivity constraints (Duchêne *et al.* 2014). Automating the
generalization process seeks to ensure cartographic quality whilst reducing costs. Flexible automated generalization tools enable customizable integration of topographic data (backdrop data) with thematic map data (salient data). This is particularly relevant in the context of maps made by nonexperts in web-based environments – where the absence of generalization functionality can lead to maps of very poor cartographic quality (Gaffuri 2011; Figure 2).

Another motivation for automating generalization is the opportunity for data producers or integrators to improve geographic database management, by capturing the initial data at a fine level of detail and automatically deriving both coarser databases (referred to as digital landscape models) and maps at different scales (digital cartographic models). This improves consistency and data quality by making explicit the route by which different maps were derived.

Automated generalization helps both with data management and with derivation of customized databases and maps in the context of spatial data infrastructures (SDIs). SDIs are infrastructures that, via geoportals, provide access to institutional geographic datasets stemming from diverse sources and services. In the context of SDIs, an important process is the integration of datasets from diverse sources (e.g., different national agencies) into a seamless dataset that enables cross-border geographic studies. Whenever the levels of detail of the input datasets are different, automated generalization is used to homogenize the data as a first step to integration (Stanislawski and Savino 2011). Figure 3 shows pan-European data from EuroGeographics (via www.geoportail.gouv.fr) and reveals uneven hydrographic detail reflecting the varying definitions of what constitutes a “major river.” Generalization can be used to even out such differences.

Modeling the generalization process

Generalization operators

From observation of the human cartographer, it is possible to identify a set of discrete activities that collectively lead to the creation of a generalized map. For example, the line work representing rivers might be simplified in its form, the polygonal field data might be merged, or more generic text labeling might help reduce clutter. Much effort has been invested in creating a taxonomy of these operators (Figure 4 is an example), in the development of algorithms, and evaluative frameworks that enable assessment of the quality of any given solution. The difference between an operator and an algorithm is that an operator articulates how the cartographer conceptualizes the cartographic design decision while the algorithm articulates the execution of that decision (Roth et al. 2011). The taxonomy comprises a range of operators that can be loosely classified into one of two types of operator; those...
addressing the modeling of data (selection and reclassification) and the rest that seek to bring clarity to the map (model generalization and cartographic generalization). Some algorithms are specific to a particular feature type, whilst others can be applied more generically.

The subtle and complex epistemic nature of cartography is reflected in the many attempts to create a comprehensive taxonomy of generalization operators (Roth, Brewer, and Stryker 2011). Different taxonomies reflect various conceptualizations of the cartographic process and the difficulty in differentiating between outputs arising from various algorithms applied in varying degree. For example selection and elimination are conceptually similar while aggregation can be further differentiated into fusion of two connected objects or the merging of disjoint objects. Whilst the problem is often rationalized (and implemented) at a geometric level, the essence of the problem is a semantic one – namely, how to transform semantic information at one level to information at another. For this to happen, we need to take into account the underlying geography that we are seeking to visualize at the geometric level. For example, the simplification of lines representing streets is very different from the simplification of lines representing the outlines of buildings.

Generalization algorithms

The wide diversity and number of generalization algorithms reflect the breadth of feature classes, their different representational forms, and how they are stored within the database. Though algorithms have been developed to generalize data in raster form, the dominant focus is on vector-based algorithms. This entry first considers those algorithms that manipulate individual objects, and then those able to take a more
<table>
<thead>
<tr>
<th>Operator</th>
<th>Description</th>
<th>Before</th>
<th>After</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class selection</td>
<td>Select the classes of features that the model contains. Define the selection criteria for populating each class.</td>
<td><img src="image1" alt="Before" /> → <img src="image2" alt="After" /></td>
<td><img src="image3" alt="Before" /> → <img src="image4" alt="After" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reclassification</td>
<td>Common or dominant characteristics of the objects are preserved. Subtle distinctions are ignored.</td>
<td><img src="image5" alt="Before" /> → <img src="image6" alt="After" /></td>
<td><img src="image7" alt="Before" /> → <img src="image8" alt="After" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simplification</td>
<td>Important details are preserved, unimportant details are eliminated.</td>
<td><img src="image9" alt="Before" /> → <img src="image10" alt="After" /></td>
<td><img src="image11" alt="Before" /> → <img src="image12" alt="After" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collapse</td>
<td>Reduction in the geometric dimension</td>
<td><img src="image13" alt="Before" /> → <img src="image14" alt="After" /></td>
<td><img src="image15" alt="Before" /> → <img src="image16" alt="After" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enhancement</td>
<td>Make the shape of a feature more aesthetically pleasing.</td>
<td><img src="image17" alt="Before" /> → <img src="image18" alt="After" /></td>
<td><img src="image19" alt="Before" /> → <img src="image20" alt="After" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elimination</td>
<td>When congestion occurs, unimportant objects are eliminated.</td>
<td><img src="image21" alt="Before" /> → <img src="image22" alt="After" /></td>
<td><img src="image23" alt="Before" /> → <img src="image24" alt="After" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Displacement</td>
<td>Important objects remain on their locations. Unimportant objects are moved away.</td>
<td><img src="image25" alt="Before" /> → <img src="image26" alt="After" /></td>
<td><img src="image27" alt="Before" /> → <img src="image28" alt="After" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aggregation</td>
<td>Common boundaries, small gaps between neighbor objects are eliminated. General footprints are preserved.</td>
<td><img src="image29" alt="Before" /> → <img src="image30" alt="After" /></td>
<td><img src="image31" alt="Before" /> → <img src="image32" alt="After" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combine</td>
<td>Regrouping a set of feature into a more abstract features, often of higher dimension</td>
<td><img src="image33" alt="Before" /> → <img src="image34" alt="After" /></td>
<td><img src="image35" alt="Before" /> → <img src="image36" alt="After" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typification</td>
<td>Reduces the number of objects, preserves their distribution / pattern</td>
<td><img src="image37" alt="Before" /> → <img src="image38" alt="After" /></td>
<td><img src="image39" alt="Before" /> → <img src="image40" alt="After" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4** A taxonomy of generalization operators.
contextual approach by taking into account their spatial and topological relationships with respect to other map features. Space prohibits exhaustive review. Instead, a subset of algorithms is considered under the following three headings.

**Generalization algorithms for isolated objects**

Early forays into automated generalization focused on the simplification of individual lines using geometric transformations parameterized by tolerance values (e.g., the line simplification algorithm of Douglas and Peucker (1973)). The algorithm starts with a straight connection between the first and the last point of the line, followed by a recursive subdivision in which points are retained based on the largest perpendicular distance between the original line and its straight line connection (Figure 5). The storage within a hierarchical data structure provides fast and ordered access to the points describing the line, from the smallest critical set down to the finest level, depending on the tolerance length required ($\varepsilon$).

Another approach for line simplification was proposed by Visvalingam and Whyatt (1993) with elimination of points based on the smallest triangles created by successive triplets of points along the line. Both algorithms provide a sorting of points according to their importance. Whilst these algorithms reduce the number of points used to represent a line, a different class of algorithms was developed for the smoothing of isolated lines. A distinction can be made between the use of spatial filters (e.g., moving average filter, Gaussian line smoothing, Perkal’s $\varepsilon$-circle rolling), curve fitting (e.g., least squares fitting, spline interpolation), and application of frequency filters (e.g., Fourier or wavelet transformation). Algorithms can be applied to a variety of object classes, though problems arise.

![Figure 5](image.jpg)  
**Figure 5** The Douglas–Peucker algorithm: (a) pseudo code; (b) iterative point reduction; (c) a binary line generalization tree of the corresponding hierarchical data structure.
through the creation of self-intersections or changes in topology with respect to neighboring objects. These same algorithms can be applied to lines representing natural polygonal features such as woods and lake boundaries. However, a different class of algorithm is required for anthropogenic features such as buildings with their orthogonal shapes and sharp edges. The fitting and rotation of a minimum bounding box is one approach to simplifying such forms. Different algorithms are required for natural polygonal features (for example, a combination of dilating and eroding buffers can be used to merge and create simplified forms). Algorithms have also been developed for area elimination or for the collapsing of area objects to a linear or point representation (effectively reducing the dimensionality of the data). A center of gravity calculation is a simple way of representing an area as a point. More complex skeletonization algorithms are required to simplify areas to linear features (for example, reducing a detailed representation of a river down to a single linear form). Typical application cases are the derivation of large-scale topographic maps at a scale of 1:10,000 with streets and rivers modeled as lines, or the creation of network graphs from cadastral data at 1:1000 scale where streets and rivers are modeled as polygons (Figure 6).

**Generalization algorithms for connected objects**

Two types of algorithms exist for handling connected objects – those that generalize networks, and those that generalize polygon mosaics (or polygonal networks). Networks are modeled as graphs comprising segments connected by nodes. Polygon mosaics are modeled via the shared edges between the polygon boundaries. Algorithms for the transformation of network graphs (such as rivers or stream networks) utilize hierarchical ordering schemes based on branching complexity such as those proposed by Strahler, Horton, or Shreve. The Strahler stream ordering begins with numbering the unbranching streams with the value one. At the confluence of two branches with the same order the number will be increased, otherwise the higher number is taken (Figure 7). Selective omission can then be carried out through elimination of lower-order river objects followed by line simplification and smoothing operations.

Figure 6  Roads modeled as polygons within cadastral data (a) and skeleton algorithm applied on polygons (b) resulting in roads modeled as lines (c) for overview maps. Haunert and Sester 2004. Reproduced by permission of ICA Commission on Generalisation and Multiple Representation.
Algorithms for the simplification of transportation or road networks apply selective omission based on the road classes, topological constraints, and the good continuation grouping principle (Thomson and Brooks 2007). At each intersection the angles between connecting segments are calculated and pairs of segments are concatenated into “strokes” based on the analysis of that angle. A ranking of the strokes is achieved by considering geometric, thematic, and topological information and this provides an ordering by which the network can then be pruned.

Polygon mosaics which represent, for example, administrative boundaries or land cover data require the topological encoding of shared boundaries. Various criteria govern the merging of adjacent polygons including shape compactness and semantic similarity. One example of such topological data encoding is the variable scale representation of area partitions using topological generalized area partitioning (tGAP) developed by Van Oosterom (2005). The tGAP provides real-time access to a representation at a broad range of scales (Figure 8). Line simplification algorithms and skeletonization algorithms have been additionally utilized in mosaic generalization.

Generalization algorithms for contextualized objects

A third group of algorithms exist that are capable of analyzing the surrounding spatial context of an object. This is required where the generalization of an object is likely to impact on surrounding objects. For example, we might displace a group of objects in order to make them less ambiguous, but this approach is only possible where there is sufficient space to accommodate the displacement. Another example might be to retain the alignment of a group of features during displacement. Thus a range of cartometric techniques are required that measure changes in qualities such as the distribution, pattern, topology, and distances between a collection of features.

The process of typification seeks to reduce the number of objects whilst preserving their distribution pattern. A range of techniques have been developed that are based on clustering methods, neural networks, or mesh simplification techniques. Using hierarchical clustering methods such as single-linkage or nearest neighbor clustering enables individual objects to be sequentially combined in a hierarchical structure. Delaunay triangulation is an example of an auxiliary data structure that makes explicit information about neighboring objects; such information can be used in the process of typification, for example (Figure 9).

Selecting and combining generalization algorithms

Map generalization methodologies are generally reactive in nature. By this we mean that design
conflicts only become apparent once the data are symbolized. For example, consider how the symbolization of a road, river, and railway running through a gorge might result in overlapping and confusing symbology. Some remedies might be to remove a feature class; to simplify the shape of the features; to locally displace some of the feature classes; or to amalgamate some of the feature classes. Invariably, a combination of algorithms are used to resolve such conflicts, which raises the question, how do we select and combine these algorithms, and when is their application deemed to have resolved the conflict? Measures of content linked to changes in scale have been proposed as one approach. For example, a power law distribution between the ratio of source and target scale denominators \( M_1/M_2 \) and the number of objects before \( (n_1) \)

![Figure 8](image-url)  
**Figure 8** Two examples utilizing the tGAP model: from the most detailed (left), through a series of intermediate solutions, to the most generalized (right). Van Oosterom *et al.* 2014 © With permission of Springer.

![Figure 9](image-url)  
**Figure 9** Typification of points representing the position of buildings based on clustering.
and after \( n_2 \) generalization was empirically developed by Töpfer and Pillewizer (1966) and provides a guiding principle of selection:

\[
 n_2 = n_1 \sqrt{M_1 / M_2}
\]

However, knowing the proposed number of objects for a given scale does not help us to decide how best to combine and variously apply these many algorithms. For this task, four approaches are possible: human interaction models; condition-action modeling; constraint-based modeling; and flowline modeling (service chaining). In human interaction modeling the expertise of the user is drawn upon to select the data and the scale and theme of the map, to decide on the generalization strategy, and to assess whether the final map output is adequate. The focus is very much on an interface design that supports the intuitive execution of all these tasks. The anticipated environment of use is within NMAs in which there is a clear understanding of the map specification requirements. All the decisions are made by the human.

A second approach is condition-action modeling. The earliest attempts at automation used a rule-based approach, typically of the form: IF <this situation arises in the map> THEN <apply algorithm \( x \) to feature object \( y \>). For example, IF the building symbology overlaps the road THEN displace the building footprint away from the road. It was envisaged that such condition-action modeling could readily be distilled from detailed map specifications associated with topographic mapping, and that rules of thumb (heuristics) governing the ordering of algorithms could be derived from observation of the human cartographer at work. For example, it was noted that the process began with modeling (selection and classification of data), followed by graphical considerations focused on clarification of the data (aggregation, smoothing, and displacement). In reality, it proved hard to formalize these condition-action pairings and such systems proved inflexible in dealing with the various mixes of features comprising any given conflict. A third and more fruitful approach has been to use constraint-based modeling. In this approach various internal and external constraints are defined (e.g., minimum separation between features; minimum changes in area and distortion of shape; minimum loss of locational precision; and maintenance of distribution patterns). The idea is that various actions associated with each constraint are then triggered. An optimization engine is used to determine the maximum number of constraints that can be satisfied. Such an approach has been incorporated within agent-based methodologies in which the behaviors and interactions among map features can be modeled as a key component to constraint-based modeling (Ruas and Duchene 2007).

The fourth approach is referred to as flowline modeling or service chaining. It involves the sequencing (and possible iteration) of map generalization algorithms based on parameter values derived from research in constraint-based modeling. It has proved popular among NMAs because of its relative simplicity, transparency, and repeatability, thus ensuring consistency and quality control throughout the map production process. In effect it involves the hard wiring of the values derived from the optimization engine within a constraint-based approach. These four approaches are not mutually exclusive; one can readily envisage a hybrid approach and a sliding scale of dependency upon the decision-making capacity of the human. Different contexts and different levels of expertise will govern which mix of approaches is most appropriate in any given context.
Evaluation in generalization

The final generalized map reflects a compromise in terms of clarity and truthfulness: clarity in conveying the relationships among the various features (both metric and topological), and truthfulness in conveying the reality of what is actually there. The omission of something, or its displacement, makes the map less truthful, but improves its clarity. The question becomes: Are the distortions of truth acceptable given the context of use?

To measure any change in truth, and to assess the clarity of the map, we first need to apply a range of cartometric techniques that inform us of the properties of the features. Metric and topological measures can be applied at the micro scale (the individual feature or its parts), at the meso scale (examining local groups of features), and at the macro scale (all occurrences of a particular feature across the map). Thus we can measure changes internal to a feature (its shape, length, area), as well as changes relative to other features (their topology, density, and distribution). A range of techniques can be used to make explicit properties that are implicit in the arrangement of those features. For example, graph theory can be used to model changes among network features and Voronoi diagrams can be used to create surfaces that define neighborhoods.

But it is one thing to measure, and another to know the significance of any given measurement in terms of the generalized map. Beyond these many geometric measures, it is imperative to make explicit the semantic relationships inherent among a set of features (Touya et al. 2014). For example, if we no longer wish to show a river, then perhaps it is not meaningful to show the bridge by which the road crossed that river. If we model the semantics of this relationship, it provides the basis for this type of cartographic decision-making. Space precludes exhaustive review of these methods of geometric and semantic enrichment, suffice to say that they are critical to the detection of conflict and to the measurement of the degree of success in their remediation. In the absence of such data enrichment techniques (making the implicit explicit), our capacity to automate is curtailed and we must depend upon the eyes of the user to detect and resolve such conflicts.

Implementations and operational environments

One can argue that the veracity of these approaches is evident in their operationalization within NMAs and their incorporation within commercial off-the-shelf GIS systems and web-based services. Here some operational use cases are briefly reported on.

Automated generalization in national mapping agencies

In the 1980s, research into automated generalization was mostly driven by the needs of NMAs. Yet it was the early 2000s before we began to see automated cartographic generalization methods being comprehensively incorporated within production environments – an endeavor that depended on massive in-house developments based on research results. One example of this was work at IGN France with the production of the 1:100,000 map. The time allocated for manual generalization for one map sheet had typically been 1200 hours. As a result of automation, this reduced to 50 hours of automated processing and 100 hours of manual corrections per sheet (Lecordix et al. 2005; Figure 10).

More recently there has been a surge in adoption of automated techniques among a number of NMAs (Duchêne et al. 2014). For the first time, three NMAs have produced maps in fully automated environments: Kadaster Netherlands with their 1:50,000 map, and Ordnance Survey...
Figure 10 Results obtained in production with automated generalization at IGN France: production of the 1:100 000 map from the “BDCarto” database. Lecordix et al. 2005. Extracts of maps © IGN.

Great Britain and IGN France with their simplified versions of their standard 1:25 000 map (Figure 11). Two factors have contributed to their success, both reflecting an ability to industrialize results from map generalization research. First, the incorporation of customizable map generalization modules within commercial GIS software, and second, being able to draw upon in-house expertise to develop additional generalization functionalities. Fully automated environments produce meaningful and acceptable results. Where high-quality solutions are required, some manual intervention is still needed.

The traditional emphasis on paper maps is being replaced by a focus on delivery of maps via web-based geoportals. In these cases, a series of outputs are stored. Each output is bound by a minimum and maximum zoom level. Thus, as the user zooms in (or out), the initially displayed map is replaced with a map at larger (or smaller) scale. NMAs have begun harmonizing the symbols used on maps across scales for display via the web in order to smooth the transition from one zoom level to another. As a result of using vector digital cartographic models (DCMs) instead of raster maps, some NMAs also enable customization of the displayed maps, by interactive selection and deselection of layers, a case in point being the NMA of the United States, the US Geological Survey (USGS), with its multiscale “National Map” (produced semiautomatically). The associated web viewer enables the user to choose which layers to display. The USGS has placed great store by its focus on web-based delivery – providing more advanced kinds of customization, including the ability to integrate thematic data provided by the user. As NMAs become more confident, they are making increasing use of cartographic generalization techniques in the updating of their digital landscape models (DLMs), with the “capture once, use many times” model in mind.

Online generalization for web and mobile maps

Traditionally, the emphasis has been on the production of maps “offline” which are then
Fully automated cartographic generalization was first achieved in production in 2013, in three NMAs: Ordnance Survey Great Britain (left), IGN France (middle), and Kadaster Netherlands (right). Top left image reproduced by permission of Ordnance Survey. © Crown copyright and/or database right 2016 OS. Bottom left image contains OS data © Crown copyright and database right. Middle top and bottom reproduced by permission of IGN France © IGN. Right top and bottom reproduced by permission of Kadaster NL.

made available online as passive images and in vector format. The complexity of production (and processing time) necessitates this approach (i.e., it has not been possible to deliver generalized maps in real time). However, it binds the user to a predefined set of outputs. Therefore a number of research studies have tried to address this issue through a hybrid approach that creates a number of precomputed levels of detail (offline generalization) that can be subsequently modified in real time by the user based on (i) the zoom level, (ii) the addition of their own thematic data (perhaps also obtained in real time), and (iii) their location.

Figure 12 shows examples of such maps obtained via advanced prototypes stemming from three studies. The first example is from a European project named WebPark which focused on the real-time delivery of touristic and recreational information to national park visitors based on their location. In this context, Edwardes (2007) proposed on-the-fly selection and aggregation methods to generate real-time views of fauna locations at different abstraction levels (Figure 12a). Depending on the zoom level, a pie chart shows the proportions of species that can be observed from a given location (top), or a small image shows the majority species and their relative numbers (bottom). The second example is from the European project GiMoDig (geospatial info-mobility service by real-time data-integration and generalization), which sought to deliver data from different European NMAs to mobile users by means of real-time data.
integration and generalization. In this context, Harrie, Sarjakoski, and Lehto (2002) proposed a cartographic display where the area around the user’s position is emphasized by a circular “fisheye” lens (Figure 12b), while the rest of the area is generalized on-the-fly by simple selection and simplification algorithms. The third example is from work done as part of INSPIRE (infrastructure for spatial information in the European Community) – a European directive, which envisages the creation of a European spatial data infrastructure supporting the online publication of NMAs’ vector data and associated on-the-fly services. In this context, Gaffuri and Tóth (2014) have proposed a system that transfers the relevant portion of a vector dataset to the web client of the user based on pan and zoom “requests.” Simple cartographic methods are then applied on the client side to obtain a more synthetic representation (in Figure 12c, a density map is computed from a point cloud). Having the vector data on the client side enables quicker interactivity with the representation.

**Evolving contexts**

The value of a map lies in its abstraction of geography. A map is a caricature – a compromise between clarity and truthfulness. The overarching ambition of map generalization is well understood – requiring as it does the modeling of user requirements that results in a map specification, a choice of remedying algorithms, process modeling that sequences and variously applies those remedies, and map measurement and evaluation methodologies to assess the quality of the final design. As a complex decision-making process, it requires us first to model the underlying geography being visualized, and based on that to make explicit the knowledge used in that decision-making process.

Advances in technology have raised the expectations of the user; they now wish to interact with data in real time, through infinitely zoomable maps that reflect their own data gathering efforts, in both web and mobile environments. Thus there...
has been something of a shift towards utilizing map generalization methodologies in exploratory data analysis contexts. There is an expectation that, beyond the focus on topographic mapping, map generalization methodologies should be relevant in the context of thematic and even schematized maps and that they should be applicable over very large changes in levels of detail (or scale). In such contexts, we may need to find alternate paradigms, though ultimately the ambition will remain the same.

Ultimately, map generalization outputs seek to support human decision-making. From this perspective, map generalization methodologies should be built around a paradigm of efficiency in communication. In such contexts, it remains very challenging when it comes to measuring the information content of a map, or the effectiveness of our design in the communication of that information. As our view of the world becomes more generalized at smaller scales, it is not that we see less information, but that we see different information. How we model those changes in information content with respect to map generalization methodologies remains a key area of research.

SEE ALSO: Cartographic design; Cartographic modeling; Geographic information science; Geographic information system; Map algebra; Mapping mashups; Scale; Visualization

References


MAP GENERALIZATION


Websites

WebPark Project: cordis.europa.eu/project/rcn/60766_en.html
GiMoDig Project: http://www.fgi.fi/fgi/research-research-projects/geospatial-info-mobility-service-real-time-data-integration-and
One way to conceptualize the process of making a map is to view the flow from Earth observation to map as a sequence of transformations. These are the scale transformation that resizes the Earth to a geodetic model, the projection transformation that changes the spherical or ellipsoidal Earth into a flat surface, and, lastly, the symbolization transformation in which objects are selected, generalized, and converted into points, lines, areas, and text on the map. Of these, the projection transformation is critical. Map projection is the oldest form of cartography, having formed a major part of Claudius Ptolemy’s foundational work Geography. A map projection is an operation that has the geographic coordinates of features on the Earth or other mappable object as input, and as output creates a new coordinate space that is Euclidean. In converting a 3-D surface into one with only two dimensions, some of the geometric properties of the globe must be sacrificed. This takes the form of distortions of scale, shape, and direction at multiple locations on the map.

Controlling and minimizing the distortion has been an elusive goal of map projection science. This is especially the case when designing projections for standard coordinate systems. Coordinate systems are ordered systems that use one or more numbers to uniquely determine the position of a point within a Euclidean space. Standard coordinate systems have been developed for mapping and geographical application, such as the Military Grid Reference System (MGRS) and the State Plane Coordinate System. These systems can cover all of the Earth (a global grid) or part of it. Coordinate systems are suitable for positioning, the recording of a place’s location as a coordinate, and for navigation to a place. As the numbers underlying digital maps and geographic information systems (GIS), it is the coordinates that facilitate plotting and allow cartometry and analysis. Without coordinates, geography would be entirely descriptive and nonscientific. Deriving and using coordinates, therefore, and understanding their projective geometry, underlie all science within geography.

Map projections: a history

It is possible that projections date from about 200 BCE or earlier (Snyder 1987). However, in Claudius Ptolemy’s (90–168 CE) book Geography, he described four new projections for the first time. The book is centered around instructing the reader on how to make their own maps. It details methods in the first part, then assigns coordinates to all known geographical places and features on a global grid. This grid counted latitude from the equator, measuring off the length of the longest day rather than degrees. Parts 2–7 used a zero meridian at the Canary Islands. The remainder of the book includes instructions on drawing the four map projections, and topographic lists with map annotations for features and places throughout the known world of the time. While the maps did not survive, during the fifteenth century when Geography was translated into Latin from the Greek and Arabic, editions...
reproduced these four projections, sparking new interest in world mapping (Campbell 1987).

The ages of enlightenment and discovery led to a boom in the devising of new map projections. This was especially true after the first circumnavigations of the Earth and the discovery of the Americas, Australia, and Antarctica. Using as a source Snyder and Voxland’s (1989) Album of Map Projections, which catalogs the principal projections in use today, dates of creation were obtained for 93 map projections (Table 1). Eight projections date from before 1500 CE, then, other than the seventeenth century, the number increases rapidly. Important early additions were the Mercator projection (1569), the Sinusoidal (1570), and the Ruysch (1508) modifications to the equidistant conic. These projections were frequently used in maps showing the new discoveries in Asia and America.

During the eighteenth century, the work of Johann Heinrich Lambert linked the map projection transformation to mathematics and geometry. Lambert was the first to discuss the properties of conformality and equal area preservation and to point out that they were mutually exclusive (Snyder 1983). In 1772, Lambert published seven new map projections in Anmerkungen und Zusätze zur Entwerfung der Land- und Himmelscharen (Tobler 1972). Of these, three have been highly influential (the Conformal Conic, Transverse Mercator, and the Azimuthal Equal Area). During the nineteenth century important contributions came from Karl Brandan Mollweide, Heinrich Christian Albers, George Siddell Airy, and Ferdinand Rudolph Hassler’s Polyconic projection. Additional notable projections in the twentieth century were those by Robinson, Snyder’s Space Oblique projection and GS-50, Tobler’s Loximuthal Projection, and those by Osborn Maitland Miller. Snyder’s book on twentieth century projections remains a definitive survey (Snyder 1997).

### Table 1 Dates of creation for 93 map projections in Snyder and Voxland’s (1989) Album of Map Projections.

<table>
<thead>
<tr>
<th>Date range</th>
<th>Number of new map projections</th>
</tr>
</thead>
<tbody>
<tr>
<td>600 BCE–1 CE</td>
<td>3</td>
</tr>
<tr>
<td>1–500 CE</td>
<td>3</td>
</tr>
<tr>
<td>500–1000 CE</td>
<td>1</td>
</tr>
<tr>
<td>1000–1500 CE</td>
<td>1</td>
</tr>
<tr>
<td>1500s</td>
<td>7</td>
</tr>
<tr>
<td>1600s</td>
<td>1</td>
</tr>
<tr>
<td>1700s</td>
<td>9</td>
</tr>
<tr>
<td>1800s</td>
<td>18</td>
</tr>
<tr>
<td>1900s</td>
<td>50</td>
</tr>
</tbody>
</table>

### Classifications of map projections

Early classifications of map projections were based entirely on their mechanical means of construction. Prior to digital maps and computer programs, projected maps were generated by applying geometrical rules to the base graticule, and then filling in the Earth’s coastlines, political boundaries, and features by manual interpolation. In the early period this led to some significant geographical errors. Mechanical aids for projection drafting became quite sophisticated during the nineteenth and early twentieth centuries, including splines and projective ray geometry.

Largely inherited from this early era was the classification by the major “developable form,” that is, the surface onto which the maps were to be projected. Azimuthal projections, such as the gnomonic and stereographic, went straight from the Earth to a flat surface, either using projective geometry or the hemisphere at a time in the globular projections. Cylindrical projections projected onto a cylinder either tangent or secant to the ellipsoid or sphere. The cylinder was then unrolled, making straight parallels and usually straight meridians. Lastly, the projection could be
onto a tangent or secant cone, usually covering one hemisphere. The unwrapped cone made for a conic projection, in which the meridians were straight lines radiating from a common point, and the parallels formed concentric circles.

Variants on this simple schema included the “pseudo” projections, pseudo-azimuthal, pseudocylindrical, and pseudoconic. The pseudo designation is because one condition of the form can be preserved, while another is varied. For example, classic cylindrical projections such as the equirectangular, Mercator, and Miller projections have straight vertical meridians, while the Molleweide and Sinusoidal have only a straight central meridian, the remainder being curves that converge at the poles. Classifications that have extended the projection types to pseudo variants often add polyconical, “other,” and space map projections as categories (Yang, Snyder, and Tobler 2000).

There are many other ways to classify map projections. Bugayevskiy and Snyder (1995) simplified the types to those with straight parallels, those with parallels forming concentric circles, and those with parallels forming nonconcentric circles. These authors note that classifications have also been based on (i) distortion properties (conformal, equivalent, equidistant), (ii) the shape of the graticule, (iii) the aspect of the projection, (iv) the form of differential equations describing the projection, and (v) the method of construction. Snyder’s 1987 survey of US Geological Survey (USGS) projections used five categories: cylindrical, conic, azimuthal and related, space map projections and pseudocylindrical, and miscellaneous. Another classification is by the degree of interrelation of the projection coordinates: \( x = f(\lambda), y = f(\phi) \) (mostly cylindrical); \( x = f(\lambda, \phi), y = f(\phi) \); \( x = f(\lambda), y = f(\lambda, \phi) \); and \( x = f(\lambda, \phi), y = (\lambda, \phi) \). In general, the more traditional methods of classification of map projections based on methods of construction are yielding to classifications based on mathematical, distortion, and functional properties, largely influenced by the incorporation of projections into software packages and GIS.

### Map projections: methods

Map projections start with a vector of geographic coordinates on the sphere or ellipsoid \((\lambda, \phi)\) and end with Cartesian coordinates \((x, y)\). The transformations are embedded within computer mapping and GIS software, where the database is usually held in geographic coordinates. All transformations involve a scale factor \(s\), the radius of the Earth \(R\), and assume a central meridian for the projection, \(\lambda_0\). The simplest projections involve independent transformations from the spherical Earth model of \(\lambda\) to \(x\), and \(\phi\) to \(y\). The Plate Carrée or equirectangular projection (Figure 1), for example, sets \(x = R_s(\lambda)\) and \(y = R_s(\phi)\). More complex projections build algebraic geometric transformations; for example, the Mercator projection (Figure 2) sets \(x = R_s(\lambda - \lambda_0)\) and \(y = R_s \ln \tan (\pi/4 + \phi/2)\), but still with independent transformations on each axis. More complex projections use both \(\lambda\) and \(\phi\) to compute \(x\) and \(y\). For example, the Molleweide projection (Figure 3) first computes and solves for an intermediate variable \((\Theta)\) using \(2\Theta + \sin (2\Theta) = \pi \sin \phi\). This allows solving for \(x\) as: \(x = 2(2^{1/2}) R_s(\lambda - \lambda_0) \cos (\Theta) / \pi\) and for \(y\) as: \(y = 2^{1/2} R_s \sin \Theta\). So \(x\) is a function of both latitude and longitude, but \(y\) is a function of just latitude. Many projections make parts of the world unplottable, for example, the high latitudes on the Mercator. When lines cross the map edges or interruptions, special case corrections are necessary.

Other complexities with map projections involve using different scale factors for \(x\) and \(y\) (which change the proportions of the map),
rotating the axes by some angle, combining two or more projections by averaging or mixing, using different interruptions and central meridians, and doing any of these transformations on an ellipsoid rather than the sphere. As a result, some map projections are numerical approximations, so cannot be analytically inverted. Inverse transformations exist for most projections, that is projection formulas that start with \((x, y)\) and yield \((\lambda, \phi)\). Inverse transformations are necessary

**Figure 1** Plate Carrée projection showing seven days (February 26 through March 3, 2012) of Aquarius V1.2.3 SSS data collected during the ascending (evening) portion of each orbit. Reproduced from NASA.

**Figure 2** Mercator projection showing elevation from the shuttle radar topography mission (SRTM). Reproduced from NASA.
to capture data that has been mapped on a particular projection.

Variants for projection include aspect, interruptions, form, and tangency. Aspects can be equatorial, oblique, or transverse (Figure 4). The formulas simply apply an affine transformation to the computed $x$ and $y$ values. In equatorial aspects, the North Pole is customarily placed at the top and the equator runs horizontally across the middle of the map. Oblique aspects are rarely used, other than in the family of space projections, such as the Space Oblique Mercator, as they allow satellite ground tracks to be shown as continuous strips. Transverse aspects have the poles in the center, and a single meridian and its anti-meridian form a straight line across the map. Interruptions are where “seams” in the transformation are applied. Adding more breaks can reduce distortion, but makes the map more discontinuous. Interruptions can be chosen to fall in the oceans or on land, depending on the map’s purpose.

Forms of projection are the equivalent surfaces onto which the projection is made. The classic forms are the plane (azimuthal), cylinder (cylindrical), and cone (conic). Azimuthal projections can cover half or the whole Earth. Many sixteenth- and seventeenth-century projections chose two circles for the hemispheres, the globular projections. Cylindrical projections are common, have the advantage of being rectangular, and may or may not include the poles. Conic projections project first onto a cone, which is then interrupted and flattened. These projections have straight meridians that radiate out from the cone’s peak, and parallels that form arcs of circles. Tangency is the degree to which the 3-D globe and the form meet directly. For example, a cylinder can be scaled so as to wrap around the equator, with all other points being projected out onto the surface. This has the advantage of zero distortion, but only along the equator. Alternatively, the projection can be made secant. In this case the scale is set
so that the equivalent tangent line becomes two lines at parallels in each hemisphere. The lines of zero distortion are now at “standard” parallels, reducing the overall distortion. The parts of the map inside the secant cylinder are projected out, and those now outside are projected in. Many maps are made secant along a central meridian by changing scale to reduce distortion, such as the Universal Transverse Mercator (UTM).

Map projections: distortion properties

In the transformation from a 3-D surface to a 2-D one, properties of the surface must be distorted. Other than the obvious distortions at interruptions and map edges, in general one or more of four interrelated properties of the surface are distorted: scale, distance, shape, and direction. Projections that preserve scale across the map are called equivalent or equal area. For each point on the Earth’s surface, a small circle when transformed should have the same area at all points on the map. Differences in scale as a function of direction create different distances across the map. Generally, distances can only be preserved from a point on an equidistant projection. For shape, it is important that features on the globe, such as a county, should still be recognizable as the same features on the map. Shape can only be preserved locally, as only a globe can preserve shape completely. Such maps are called conformal. Conformal maps also preserve direction, but only over short distances. Conformality is desirable because compass directions are preserved locally, making them useful for navigation. Only a gnomonic projection can preserve directions, and only then for one point. Conformal projections have been used extensively since the twentieth century as the basis for standard coordinate systems for surveying and mapping. Projections that achieve neither conformality nor equivalence are termed compromise projections, but even maps that are equivalent or conformal do not preserve distortion at all points on the map.

It is possible to compute and display on a map the degree of each distortion (Mulcahy and Clarke 2001). First to do so was Nicolas Auguste Tissot in 1859 and 1871 in the form of his ellipse of distortion or indicatrix. Placed on the map at increments of longitude and latitude is a repeating small circle that results from projecting a circle of infinitesimal radius from the globe onto the map. Tissot proved that an ellipse whose axes indicate the two principal directions along which scale will be maximal and minimal.
at that point on the map. The ellipses show scale distortion by their size, shape distortion by their degree of lengthening, and direction by their angle (Figure 5). They are a highly effective visual aid in understanding projection distortion.

Lastly, many have attempted to optimize projections so that all distortions fall below a threshold, at least for a particular area. The Winkel tripel was an early example. The equal area Goode’s Homolosine joined two projections (Molleweide and the Sinusoidal) to improve conformality. Others have attempted to create the average of coordinates across multiple projections with each property. Snyder’s GS–50 was designed to be conformal, but to keep scale distortion below 2% over the land areas of the 50 United States (Figure 6).

In addition, recent scholarship has moved toward creating intelligent decision-support systems that assist in making choices about which projection to use for a given application. One such system is the “Decision Support System for Map Projections of Small Scale Data” from the USGS (USGS 2013).

Coordinate systems

Once a map has been projected into a Euclidean space, it is possible to use the mathematics of geometry and trigonometry to perform a large number of operations. Two technologies, however, favor systems that formalize the 2-D referencing into standardized coordinate systems. These are positioning, usually via a global navigation satellite system such as the global positioning system, and navigation, including web-based map use. Navigation needs coordinates both in the analog and digital sense, for display on aids to navigation and to create grids on maps. Tasks such as location-based services, search and rescue, point-to-point navigation by vehicle, and many others demand and use standard coordinates.

In some cases, geographic coordinates themselves can be used as coordinates. Many software packages store raw geographic coordinates in various forms and do map projection and coordinate conversion on the fly. However, the nature of the shape of the Earth makes the use of geographic coordinates difficult. An alternative is three axis (x,y,z) coordinates called Earth centered Earth fixed, as is common in geodesy, space geometry, and computer graphics. Mapping applications, however, favor flat map transformations so that the grids can be printed or drawn directly onto the maps. This implies two things: first a suitable
projection must be chosen and, second, the system must establish units, axes, and a point or origin.

Comparisons have been made of the properties of “good” global grid systems (Clarke 2002). Desirable properties are universality (fitting into existing alternative Earth models such as geodetic and astronomical models); authority (being supported by a permanent and credible agency, standard, or law); being succinct (using as few codes as possible to ensure location); definitiveness (providing reference to one and only one unique location); exhaustiveness (applying to the whole Earth); being hierarchical (allowing references to multiple scales depending on the desired level of detail); intuitive (easy to understand and use); and tractable (able to be used computationally). Not all grids in common use share all of these properties and, consequently, there exist multiple grid systems, in spite of national efforts at standardization around “National Grids” as in the United States and the United Kingdom.

Among the many global grid coordinate systems in common use are different versions of the UTM coordinate system (UTM, MGRS, the US National Grid), a referencing system for satellite imagery (Georef: The World Geographic Reference System), and various systems based on tesselated triangles and hexagons (e.g., QTM). There are also national grid systems for many countries, such as the Map Grid of Australia, the British Ordnance Survey National Grid, the Irish National
Grid, the QND95/Qatar National Grid, and in the United States the State Plane Coordinate system. Many national grid systems, though obviously existent, have little publicly available information about their details. At a third level, there are a host of special purpose grids such as the Maidenhead Grid Square system (amateur radio), the US Public Land Survey system, and the Canadian Dominion Land survey (cadastral), and the Universal Polar Stereographic, which completes the military and civilian versions of the UTM system for the two polar regions.

Covering all these systems is beyond the scope of this entry (see Cartesian coordinate systems). The reader is referred to Clarke (2015), Maling (1992), Goodchild and Kimerling (2002), and the other sources referenced in the Further reading section. Notable is that most online mapping systems and geobrowsers support only a subset of the systems: typically UTM and geographic coordinates, and some nonstandard grids. Many of the systems are based on the properties of the Transverse Mercator (or Gauss–Kruger) projection, usually made secant along the central meridian. Others mix projections, such as the United States State Plane Coordinate system, which uses the Transverse Mercator, the Lambert conformal conic, and the Oblique Mercator. The choice of projections and liberal use of zones and false origins is testimony to the difficulty of removing map distortion during projection. Therefore, global grids that extend across the whole planet or most of its surface by necessity introduce significant error due to scale and directional divergence. Projection distortion is always least on grid systems that are local, use conformal projections, and use local secancy or controlled extent to minimize distortion, strategically surrounding a particular mapping region of interest.

**SEE ALSO:** Cartesian coordinate systems; Geodesy

**References**


Further reading


Cyberspace from imagined to real

“Cyberspace is real” claimed President Barack Obama in his speech concerning American cyber infrastructure security in 2009 (White House 2009). In other words, cyberspace might be surreal, thus the president was obliged to make such an unmistakable statement about the existence of cyberspace. However, the uncanny thing is that cyberspace is a neologism coined by a science fiction writer through popular science fiction, which evokes massive public imagination about a future virtual digital space. The science fiction writer is William Gibson and the novel is *Neuromancer* published in 1984. So, what turned cyberspace from merely an imagination of the general public into an existence whose safety is fundamental to the security of a nation in less than two decades? More importantly, in order to map cyberspace in a meaningful way, there should be a clear understanding about the materiality of cyberspace.

Before striving to understand the materiality of cyberspace, it is useful to briefly review the concept of cyberspace, a virtual world, as it was envisaged by William Gibson. The original imaginative representation deserves a mention not only because it became a symbol in popular culture, but also because scholars and technicians were so fascinated by the idea to eventually bring forth the creation of the real cyberspace today (Bell 2007). In *Neuromancer*, cyberspace is “a consensual hallucination experienced daily by millions of legitimate operations. … A graphic representation of data abstracted from the banks of every computer in the human system.” (Gibson 1984, 67).

Nonetheless, although Gibson (1991) pointed out later that the neologism of cyberspace was suggestive and without concrete semantic meaning, the imagination of a virtual cyberspace had profound influence upon its subsequent materialization (Bell 2007). Gibson coined the word by borrowing the prefix “cyber-” from cybernetics and connected it to the suffix “-space,” whose combination pointed to an imaginary “datscape.” To a great degree, in the virtual world depicted by Gibson, the spatial dimensions of the cyberspace were implied. For the Greek stem of the word cybernetics, *kubernites*, means steersman (Wiener 1954). It was Benedikt, who orchestrated a spatial environment to spatially contextualize the constellation of immaterial data. According to Benedikt, the cyberspace is “multisensory, three-dimensional, involving, richly textured and nuanced virtual world converting oceans of abstract data and the intelligence of distant people into perceptually engaging, all-but-firsthand experience.” (Benedikt 1991a, 1991b) Thus, the cyberspace could be navigated in a way similar to geographical navigation in the real world, and a set of rules about discerning space, time, and objects could be fancied. That is to imply that, in Benedikt’s cyberspace, the virtual world could be mapped in the similar way to the way our physical world is mapped.

The goal of this entry is to introduce geographical knowledge regarding crafting maps of cyberspace. While cyberspace is beyond
humans’ direct perception, its skeleton, which is the modern day Internet, is as concrete as any other physical objects on the Earth’s surface. The rest of this entry is organized by first briefly overviewing the evolution of the Internet then in the second section focusing on the definition and categorization of many types of webs, such as the World Wide Web and other specific webs. After the third section, which gives a brief description of a cyberspace model, different mapping methodologies and maps of cyberspace are introduced. The entry concludes with a brief summary about the materiality of cyberspace.

Internet as the skeleton of cyberspace

In order to map cyberspace, the first step would be to investigate cyberspace about its constituent components. By carrying out a search in Google Scholar for articles containing the exact phrase “what is cyberspace,” the engine would give about 233 entries in less than 0.06s. Scanning through these entries you would realize that the definition of cyberspace is multifaceted by people’s perspectives. If you read them in detail, then you would discover that Benedikt (1991b)’s definition about cyberspace has been popularly cited. According to Benedikt, cyberspace is “a globally networked, computer-sustained, computer-accessed, and computer generated, multidimensional, artificial, or ‘virtual reality’.” In this article, Benedikt’s vision about cyberspace is adopted, and it was from this vision that we started to demonstrate the project of mapping cyberspace by first examining the evolution of cyberspace.

The progenitor of the present-day Internet, the ARPANET, was initiated 45 years ago in the shadow of global nuclear war. The motivation for the project of the Advanced Research Projects Agency (ARPA) within the US Department of Defense was to share the limited computing resources of powerful large research computers with users at geographically dispersed locations at that time (Leiner, Cerf, and Clark 2012). The first generation of ARPANET was launched in 1969 connecting hosts at four US universities: Stanford Research Institute, University of California at Los Angeles, University of California at Santa Barbara, and the University of Utah. ARPANET grew rapidly to have 23 hosts connecting universities and research centers in the United States by 1971. It became international in 1973 with two connections established to University College in London, England and the Royal Radar Establishment in Norway. The first commercial version of ARPANET was constructed by Bolt, Beranek and Newman (BBN) in 1974, called the Telenet. It is from the Telenet that the public started to vaguely savor the possibilities of networked computers. From the 1970s to early 1980s, more computer networks for specific usages were constructed, for example, BITNET for electronic mail and UUNET for Unix users. The dominance of ARPANET gradually faded out during this period, until the operation was permanently terminated in 1989 and the system was retired in 1990. Through its life cycle, ARPANET pioneered the research and development in global communication network construction, and left many legacies to its successors, such as the transmission control protocol/internet protocol (TCP/IP), which became the only endorsed information communication protocol for the future Internet.

While various networks were built in the 1980s, one gained extraordinary importance, since our present-day Internet was a spin-off of the original network, the NSFNET, which was funded by the US National Science Foundation (NSF). The initial NSFNET was constructed to support sharing precious supercomputing powers among
MAPPING CYBERSPACE

scientists at different locations of the United States. The first generation of the NSFNET initiated operation in mid-1986 and connected NSF-funded supercomputing sites. The sites comprised San Diego Supercomputer Center, University of Illinois at Urbana-Champaign, Carnegie Mellon University, Princeton University, which were all newly built out-of NSF funds, and the then-existing supercomputing facilities at the National Center for Atmospheric Research. It was also called the NSF backbone network, or simply the Backbone. Around this backbone network, many regional networks were deployed to connect countless local area networks (LANs). LAN was the first layer in the hierarchy of the layered NSFNET architecture. In the early days of the NSFNET, LANs were more likely to connect computers on university campuses, called campus networks. NSF called the NSFNET “inter-net,” because it was a net of networks. The structure of the Internet that the NSFNET adopted was developed by the Defense Advanced Research Projects Agency (DARPA), renamed from ARPA. Eventually, NSF decided to initiate the backbone service transition from a NSF-controlled Internet to a commercially operated Internet. The transition concluded in 1995, and the NSFNET backbone was decommissioned, which denoted the end of an era in the Internet history. As part of the solutions to decommission the NSFNET backbone, four network access points (NAPs) were established to enable information exchange among different networks. The four NAPs were deployed at four sites: New York, Washington DC, Chicago, and San Jose. At each NAP, peer Internet service providers (ISPs) connected with each other based on standard agreements. Business companies or individual persons can purchase connection services to Internet from ISPs. Soon NAP was replaced by Internet exchange points (IXPs), which gave the interconnected networks advantages of direct connectivity instead of via third-party networks.

With the Internet moving to the mode of ISP with IXP, the operational model of the modern Internet was formed. As ISPs dominate the accesses to the Internet, they may deny users’ connections to destinations. Thus, geographical proximity does not necessarily equal to cyberspace closeness. Recently, the development of wireless communication technologies has realized high-speed Internet connections on mobile platforms, which has fostered the nascent and fertile field of mobile networking. Grassroots organizations have been implementing wireless networking technologies to construct private communication networks at small areas, such as in Portland, OR. With many privately constructed networks mushroomed, there are efforts to connect them, and hopefully form an alternative Internet, independent of the current Internet and its supervision. In geography, there has been a persistent interest in mapping the Internet and the virtual space that was carved out of the bricks and mortar of the physical servers and routers. However, it is a daunting task. With the invention of the World Wide Web (WWW) and search engines as well as the growth and evolution of the Internet, the entire Internet has gradually developed into many types of networks, such as surface web, deep web, darknet, or dark Internet.

The virtual worlds of cyberspace

There are virtual worlds of cyberspace as different types of webs, either surface web or deep web. The web can be deemed an application that runs on the Internet. Although sometimes the Internet and the web are treated as the same thing, practically they are quite different. The Internet refers to the interconnected physical servers, and the web consists of networks of hypertext documents that can be accessed through the Internet.
Tim Berners-Lee invented the World Wide Web (or the web) when he worked at the European Organization for Nuclear Research (known as CERN). The original proposal was to improve the effectiveness of CERN’s communication system. But soon, Berners-Lee realized that the same design could be applied throughout the Internet, and thus the web was born. Inside the web, the hypertext document is identified by the uniform resource locator (URL) or the uniform resource identifier (URI). Hypertexts are formatted using the hypertext markup language (HTML) as web pages, which can be interpreted by web browsers. The exchanging and transferring of hypertext document follows the hypertext transfer protocol (HTTP). With hyperlinks embedded inside the body of HTML web pages, the conglomeration of hypertext documents forms another level of interconnectivity on top of the Internet. One fact that deserves to be pointed out is that although Berners-Lee proposed the universal authorship of the web, it actually took a very long time to realize fully the universal authorship for the majority of web users.

According to Morris and Ogan (1996), there are four types of relationships between information producers and receivers on the Internet. The first is one-to-one asynchronous information exchange, such as found with using email. The second is many-to-many asynchronous information broadcasting, and online public forum is the typical example. The third is synchronous communication, in modes from one-to-one, one-to-few, to one-to-many; for example, two users chatting via Skype. The fourth is synchronous information seeking, such as in the scenario when web users search and receive information from a website. For a very long time, the web was read-only, especially for multimedia contents, to the vast majority of web users, and the right to broadcast web content was held by a small group of producers that had the suitable skills for manipulating multimedia contents and the necessary computing resources for holding them. The producer's privilege was gradually transferred to the majority of web users after the emergence of search engine companies, such as Google, and social media companies, such as Flickr, Twitter, and Facebook. The web with universal authorship has been termed the web 2.0, in order to distinguish it from the first generation of web, or web 1.0, which was mostly read-only for the majority.

Indeed, the generational evolution of the web marked not merely the shift of power, but the web in a much deeper entanglement with the general public’s daily lives. In recent years, with the fast evolution of computing power of mobile platforms, such as smartphones and tablets, a new mode of web usage rises significantly, generally called the mobile web, which refers to the phenomenon of web surfing via mobile devices. It has been predicted that soon the use of the mobile web would surpass the use of the traditional web on fixed-line computers. The increasing fusion of the web 2.0 technologies and the mobile web has triggered dramatic socioeconomic and technological changes in both developing and developed countries. In developing countries, the use of mobile phones and social media has become a way to facilitate humanitarian operations by the United Nations. In developed countries, mobile social networking has gradually become the lifestyle of the younger generations. If the web 1.0 cyberspace was somehow like a detached virtual space from the geographical space where we live, then the emergence of the web 2.0 applications and the mobile web has indeed blurred the once distinctive boundary between the virtual and physical spaces.

The web is not the only application on the Internet, although it is one of the most popular applications and gives an immediate impression
about the Internet for general web users. There are many other Internet applications, some of them date back to the early days of the Internet, such as electronic mail (e-mail) application, Telnet, file transfer protocols (FTP), multiple user domain (MUD), and so on. E-mail has been the most popular application of the Internet. Telnet turns a user’s desktop into a terminal to a remote server, and enables the user to work remotely on files stored on the server. FTP lets users connect to remote servers without offering them capacities to work on it beyond browsing and transferring files. Also, there is multiple user domain (MUD, originally multi-user dungeon), which is a real-time virtual environment for users to explore. In a MUD world, a user imagines the world by reading the textual description. Meanwhile, the user can interact with peers through text messages. It is a very primitive role-play video game. But not all MUDs were games, some MUDs were contracted for educational and research purposes. Different from the MUD, virtual reality (VR) strives to visually create a mirror world of the physical world or an imagined world on the Internet. Google Earth is the most well known example of the mirror world, and World of Warcraft (WoW) is one of the most popular imagined online virtual worlds.

Just as the web is not the only application on the Internet, computers are not the only terminals that have the capacity for Internet communication. Things like the Internet phone, online fax, or Internet printer can all connect to the Internet. Actually, with the Internet communication functions built into more and more daily appliances, the web has gradually dissolved into humans’ daily life and the Internet communication technology becomes invisible. It has been envisioned that in future, with implementation of the Internet of Things (IoT) technology, which refers to the communication networks of material objects (e.g., clothes, food stocks), the webs of machines will greatly shift human societies towards more intelligent and autonomous.

Thus on the Internet, there are humans who interact with each other, machines which interact with each other, and humans who interact with machines. Different spaces were formed because of these activities. That is to say there exists not a single universal cyberspace, but many cyberspaces that can be defined at different levels of the Internet and with different types of terminals. In terms of mapping cyberspace, this complexity is significant and profoundly influences how we map cyberspace and why we map cyberspace.

The ways to spatially anchor cyberspace

The task to map cyberspace implies that the virtual space has spatiality and spatial geometry. Otherwise, there would be nothing to be measured and orchestrated to allow a space to be depicted on a map. Hence, an immediate question would be what defines the spatiality and spatial geometry of a cyberspace. Socially produced space is termed “spatiality,” which differs from the spatial geometry that is physical existence (Soja 1985). If the spatial geometry of cyberspace is somewhat obvious to us – the physical infrastructure of the Internet (for a comprehensive map about the global Internet between cities worldwide, see: http://global-internet-map-2012.telegeography.com/) then the spatiality of cyberspace is very obscure. In order to understand the spatiality of cyberspace, we must understand that the infrastructure of the Internet only provides the skeleton, on which information is transmitted. The operation of the Internet is dictated by software at various levels of the Internet. At the backbone level, the connection of information transmission conduit is regulated by the autonomous system (AS).
MAPPING CYBERSPACE

ISPs use AS policies to restrict the connectivity between different destinations of computers. The AS can be modified according to requirements. Hence, although the physical topology of the Internet is more stable, the information topology of the Internet is dynamic and ever changing. On top of the ISPs, companies or individual persons receive Internet connection services to satisfy their communication needs. These needs are fulfilled by leveraging all kinds of software. Indeed, it is through the utilization of software that humans interact with the information sphere of the Internet, and together they bring space into being. The socioeconomic studies of software have pointed out that software is not merely a technology product, it is also social constructions, which represent the idealisms of their creators (Kitchin and Dodge 2011). Thus, the being of cyberspace is socially constructed, which is the spatiality of the Internet.

Batty (1997) conceptualized the assembly of different types of hardware on the Internet with the different spaces they initialized in the following model. First, personal computers are the entrance to the Internet, and the interactions between humans and the Internet via the use of software on the computer create “cspace.” Therefore, individual cspaces exist locally in the geographical world, and the distributed networks of connected cspaces form cyberspace. In order for cyberspace to occur, it needs to be supplemented with the functioning physical hardware of the Internet. The interconnection of the physical hardware of the Internet forms the cyberplace. Batty further stated that the use of cyberspace created demands for extending the boundary of cyberplace to new geographical territories. The establishment of the Internet infrastructure modifies the socioeconomic relationships of geographical space. For companies, these modifications bring business advantages, and entice them to move towards increasing creation of cyberspace inside of the geographical space. Regarding mapping cyberspace, Batty’s model has provided a solid foundation for mapping cyberspace as the physical infrastructure, which is through measuring the topology of the hardware. However, it is much too complex to map cyberspace in terms of an information sphere, as hypertexts are mostly nonspatial, which indicates the fact that information in cyberspace lacks explicit spatial characteristics that belong to material things in the real world. Thus, spatialization is the first step for mapping the information sphere of cyberspace. Spatialization builds theoretical spatial forms of information in cyberspace.

Meanwhile, the evolvement of the information and communication technology (ICT) has mobilized Internet access. People can access the Internet from their mobile devices such as smartphones or tablet computers. Consequently, cyberspace at the physical level becomes dynamic as well. As the availability of global positioning system (GPS) technology, which is a satellite-based navigation system providing real-time location and time information, increases on mobile devices, it is much easier and accurate to pinpoint the geographic location of Internet accessing points by leveraging the GPS coordinates that have been inserted into the contents transmitted by the devices. Satellite technology not only supports versatile GPS applications that are widely demanded in situations where accurate geographical location matters, but also provides a way to observe the Earth’s surface by means of high resolution remote sensing imagery. These high resolution Earth surface images have been applied to create web-based virtual globes. One of the most popular virtual globes is Google Earth/Map. In the history of the Internet and cyberspace, a few companies stood at pivot points. Google is one of those companies and with its Google search engine, Google Earth/Map, and many other products changed users’ experiences
about the Internet and cyberspace. The Google search engine largely made obsolete the requirement for a comprehensive online catalog about webpages. Google Earth/Map really revived the public’s interests in geography. Using Google Earth/Map for virtual geographical discovery indeed becomes a fad, such that online forums and user communities were formed for sharing experiences and discoveries. The public’s enthusiasm about web-based mapping has been one form of a broader phenomenon of an increasing fusion of the Internet and people’s daily lives.

The web from 1.0 to 2.0 gives mapping cyberspace multiple meanings and methods. Firstly, mapping cyberspace can be done by approaching the ethereality of cyberspace via its material skeleton, the Internet. Secondly, mapping cyberspace can also refer to the activities such as building navigation map of the information sphere. Thirdly, mapping cyberspace can also broadly mean mapping in cyberspace and is a significant phenomenon. Web-based geographical information systems have become a major technological development trend for GIS, with companies like ESRI gradually moving its products into cyberspace. Google Earth/Map extended the means and applications of mapping in cyberspace. They have been one of the driving forces behind the volunteered geographical information (VGI) phenomenon (Goodchild 2007). VGI, or more broadly web 2.0, represents a new phase in the Internet and cyberspace evolvement.

Exemplar mapping cyberspace applications

There have been multiple ways to anchor digital information in the real world. Jiang and Ormeling (1997) among a few others started early attempts to adopt cartographic methods of creating maps for the geographical world, to plot maps for cyberspace. The maps were termed cybermaps. Cybermaps suggested solutions for questions such as how to map cyberspace and what to map. As cybermaps developed, they divided cyberspace into two parts: a physical part and an information part. More focus had been put on mapping the information part rather than on the physical part of cyberspace. Here, cybermaps provided several methodologies, with some still widely implemented in all kinds of tasks today. For example, using the last part of a domain name (e.g., www.google.com), the part after the last period (e.g., .com), is used to assign an organizational role to a website, indicating whether it is situated in the commercial domain (.com) or in the educational domain (.edu). Also, the last part of the domain name can indicate the geographical region, if the domain name ends with a country abbreviation, such as .uk or .us. Domain names, for example, google.com (no www), are purchased from registrars, such as Go Daddy Company. Once a customer owns a domain name, he/she can create as many subdomains as wanted. Each subdomain then can be allocated to a host machine. The allocation information is stored in domain name servers (DNS) in a form that connects the subdomain to the IP address of the host machine. Each host machine has a unique IP address on the Internet. Barrett Lyon initiated the Opte Project to map the Internet by using a tool called traceroute to record the route of data packets across an IP network. Maps produced by him have been collected by the Museum of Modern Art (MoMA) in New York City (for an example of IP maps, see: http://www.moma.org/collection/object.php?object_id=110263). The allocation of IP addresses is handled by regional organizations. Therefore, by examining the numerical numbers of an IP address, there is great chance to rightly estimate the geographical location of a host machine at the country level. For example, IP addresses between
The block from 3.0.0.0 to 3.255.255.255 are more likely to belong to host machines in the United States. There is software to convert domain names or IP addresses to the geographical locations (i.e., longitude and latitude) of the host machines. Thus, it has created a way to map the geographical dimension of cyberspace by using domain names or IP addresses. By estimate of the total numbers of hosts connected to the Internet, another type of cybermap was created to show the number of hosts per capita in different regions, such as in the United States. On the Internet, each access to a host machine is captured and recorded in a log file. By analyzing information kept in the log file, maps can be plotted to show the frequency of access as well as the geographical distribution of requests. By capturing the geographical existence of host machines, domain names, and IP addresses, cybermap managed to materialize the ethereal nature of cyberspace in geographical fashions.

The above methods portray cyberspace from a point that is close to Batty’s model about cspace and its assemblage. When getting inside cyberspace, that is, entering the information sphere of the cyberspace, the geographical location of the host machine does not provide much clue for navigating such an entirely virtual landscape. Cybermap used the networks of hyperlinks to indicate the position inside the information sphere. Thus, in order to reach a hypertext document, for example a home page at the end of a connection, a predefined passage of hyperlinks had to be followed. The assembly of passages and nodes created a network-like representation of cyberspace. The vast and ever expanding scope of networks of hyperlinks made the navigation a daunting task, even for someone familiar with the information sphere. Yahoo! has been one of the Internet companies that has achieved great success by cataloging these messy home pages according to their themes. A cybermap thence can be crafted by treating catalogue information captured by Yahoo! as a proxy of the structure of the cyberspace’s information sphere; and by representing such hierarchical structure in cartographical language, namely points and lines, the information sphere of cyberspace can be mapped. Such a cybermap is an indispensable compass for cyberspace navigation. Potentially, a cybermap can serve three types of functional purpose; navigation is the first functional category. The other two functional categories are cyberspatial analysis, which quantifies information flows on the Internet; and persuasion, which provides situational snapshots about the Internet.

For cyberspace of web 1.0, Dodge and Kitchin (2001) provided one of the earliest comprehensive surveys and illustrations about ways to map cyberspace. In the book, they presented methodologies about mapping cyberspace from different perspectives, from socioeconomic to art. Zook (2005) mapped the .com type of domain names about their geographical distributions in several US cities. The maps demonstrated the comparative effects of cyberspace on geographical distributions of socioeconomic developments. With Euclidean distance becoming ethereal and symbolic in cyberspace, the possibilities to map cyberspace is only confined by imagination. Tamara Munzner and Paul Burchard leveraged the multidimensions of hyperbolic space, which mathematically is a non-Euclidean geometry, to construct a navigable 3-D information sphere of cyberspace. Online forums and discussion boards normally adopt a hierarchical information organization strategy for displaying different topics. The method is that a new topic becomes the topmost entry of a conversation tree, and has all responses to the original topic listed as leaves. One drawback with this arrangement is that for a long conversation it becomes difficult to trace the threads between conversations. Xiong and Donath (1999) invented the PeopleGarden tool...
to visually portrait relationships between different conversations. Besides the various methodological depictions about cyberspace, there were artists who illustrated their imaginations about cyberspace in animated film and deconstruction views of webpage and its hidden codes.

Technically, hardware and software behind web 2.0 has mobilized Internet implementations, which resulted in greater fusion of cyberspace and geographical space, creating a type of hybrid space (Couclelis 1996). Inside hybrid space, there is in situ user experience of cyberspace. That is to say, users’ geographical location has been part of the criteria for an information search in cyberspace. Meanwhile, instant response from the cyberspace communities enables users to make real-time decisions that have concrete links to the real world. This technological trend is accelerating, with wearable technologies, such as Google Glass, and implantable technologies, such as neural implants, becoming realistic and affordable, which is a highly controversial trend (McGee and Maguire 2007). The emergence of hybrid space has generated new challenges for mapping cyberspace. One effect with the availability of hybrid space is that ideas or rumors can travel far more broadly as well as in greater speed through cyberspace, and impact people’s behaviors in the real world. For example, social media tools such as Twitter and Facebook have been commonly used for emergency responses or organizing protection. Scholars have started to study the new phenomenon and produce new methodologies, for example, for capturing the geographical footprints of real-world transitory events through pinpointing digital traces they left in cyberspace (Tsou et al. 2014).

The invention of search engines dramatically changed the way of information seeking in cyberspace. Automatic crawlers constantly navigate the accessible Internet and build an index of the information sphere. Keywords have to be properly designed for accurate search results, which are listed according to their relevancy to the keywords. From web 1.0 to 2.0 rendered it almost impossible to create catalogs for web contents that keep growing as web users continuously post new blogs or upload new photos. Search engines can better capture the evolving and changing information landscape, which can potentially be implemented for assessing public opinion about, e.g., presidential candidates (Tsou et al. 2013). The method starts with keyword searching that find the most relevant webpages on a subject. As each webpage has to be hosted by a computer that has an IP address, the location of the webpage can be approximated by the geographical location of the hosting server and generate a geographical point that can be put on a traditional map. The aggregation of webpages on the subject thus creates a density map that reflects the geographical distribution of the webpages. As webpages contain ideas that spread throughout the Internet, the method suggests a new way to study information diffusion in cyberspace.

Compared to webpages, social media are nowadays preferred for gauging public opinions. Part of the difficulty with mapping cyberspace using social media data is caused by the discrepancy between geographic locations of content creators and content hosts. There are explicit and implicit methods to inferring the geographical locations of the social media contents. If tweets or photos posted to Twitter or Flickr are from devices with geo-enabled technologies available, such as GPS or wireless triangulation, then explicit geographical locations can be easily gotten by reading the longitude and latitude coordinates. However, due to privacy considerations or the availability of geo-enabled technologies, explicit geographical locations are only contained by a small percentage of total public social media contents. Methods have been designed to indicate geographical locations from hints embedded
Mapping cyberspace reveals spatiality and geometry of cyberspace at different levels of the Internet. In his effort to explain the rearrangement of socioeconomic relationships in the digital information age, Castells (2011) defined two types of space, namely the space of places and the space of flows. The space of places refers to space that consists of physical places in the real world, where daily lives take place. The experience of space of places is confined by locality; therefore, the conduct of different socioeconomic activities has distinct geographical places. Through cyberspace, certain types of locality have been annihilated. Geographically distributed places can coexist inside of cyberspace, which forms the space of flows. New space, such as space of flows, expresses the spatially transformed society (Castells 2004).

Cyberspace maps convey the changes in a more understandable means, which provides indispensable support for decisions regarding future development of societies or cities. Finally, through the pursuance of mapping cyberspace at different levels from the Internet, the TCP/IP system, and to the virtual hypertext documents, the materiality of cyberspace has been revealed, which comprises human, code, software, and infrastructure (Kinsley 2014).

See also: virtual reality; Web-mapping services

References


Mapping mashups

Matthew Zook
Jessica Breen
University of Kentucky, USA

Mashing and mapping

A map mashup is a genre of mashup in which two or more pre-existing data sources gleaned from the Internet, via either data scraping or querying an application programming interface (API), are combined in a web map to create a novel, online cartographic representation. The original data sources can be either used directly (e.g., mapping a list of objects with latitude and longitude coordinates or addresses) or (dis)aggregated into new spatial categories. Originally termed a “Google Map mashup” (see Dalton 2013), the map mashup is a subset of a larger category of social practice associated with creativity and artistic endeavors including collage and musical remix, such as the methods used extensively in hip hop music (Butler 2006; McConchie 2008).

In addition to their connections to artistic practice, map mashups are also closely associated with the advent of user-generated data and social networking, often lumped under the category of Web 2.0 (O’Reilly 2005), which provided novel data sources to map (Zook and Graham 2007). While map mashups are not limited to this kind of volunteered geographic information (Goodchild 2007) – reuse or combination of non-crowdsourced datasets would also be included – the availability and variety of volunteered geographic information (VGI) data has featured prominently in these efforts.

An additional key element tied to the rise of map mashups was the unscrambling of civilian global positioning system (GPS) satellite data by President Bill Clinton in 1996, paving the way for an explosion in the use of GPS in consumer electronics, including the rise of location-aware mobile services and social networking. These applications leverage the new sources of geotagged data produced by Web 2.0 activities with handheld GPS-enabled devices to enable users to both create and consume map mashups, from relatively simplistic wayfinding (directions to a party) to spatial search (location of restaurants, apartments, etc.) to social and political commentary (historical tours, political protests, and geotagged Wikipedia entries). The flexibility and versatility of map mashups also means that they have become a widely adopted practice.

Ten years of map mashups

Arguably the most remarkable aspect of map mashups is how quickly they have faded into the background consciousness of daily life. They are used by hundreds of millions people every day in a matter-of-fact manner, and yet the origin of these remarkable and powerful cartographic artifacts only traces back to 2005.

Paul Rademacher is credited with creating the first mapping mashup, and it was in reference to his original mashup, Housingmaps.com, that the term “Google Map mashup” was first coined (Dalton 2013). In April 2005, Rademacher, who at the time worked as a software engineer at Dreamworks, reverse engineered the then...
unpublished Google Maps API so that he could map rental housing information gathered from Craigslist to aid him in his search for a new apartment. Google quickly discovered Rademacher’s unauthorized use of their map data, but rather than prosecute him for the data breach, they hired him. This choice by Google not to sue Rademacher – and moreover to publish the Google Maps API and actively encourage third-party applications – represents a key moment in the spread of mapping mashups.

In May 2005, Google Maps API went live for official use, with the proviso that users had to register their use with Google, a requirement which provided Google with greater control. This launch also created just enough changes in the API setup to effectively render pre-existing map mashups useless (Dalton 2013). This was followed by the release of map APIs from other online companies, including those of Yahoo!, Bing, and MapQuest. It also set the stage for OpenStreetMap, then in its infancy, to eventually adopt an open API as well. The availability of these online mapping tools and data sources attracted a number of individuals without formal cartographic training – the so-called neogeographers (Turner 2006). Many of these neogeographers come out of the field of computer science due to their familiarity with the tools used to create map mashups, namely APIs, data scraping, and coding.

As the number of APIs increased, so did the breadth of topics dealt with by map mashups. Expanding beyond the relatively simple first mapping of Craigslist apartment listings, map mashups have expanded to topics ranging from the mundane to the enchanting and the downright creepy. While a full accounting is well beyond the scope of this entry (for more examples see ProgrammableWeb.com, which tracks mashups by genre and API utilization, and googlemapsmania.blogspot.com, which highlights a range of online maps), examples to provide some sense of the range of mapping mashups include the following.

- Taxiwiz provides information on the cost of taxi fares for specific trips in about a dozen cities.
- Healthmap.org uses World Health Organization data to provide maps of contagious disease surveillance.
- Donor2Deed traces charitable donations to on-the-ground actions.
- Trulia lists properties for rent, for sale, and recently sold to create a house-hunting map.
- Goocam searches out and maps unprotected webcams connected to the Internet.

A key characteristic shared by all of these map mashups is the creation of cartographic products that make new patterns visible that were not as easily found in the data’s previous form or use. This often includes providing users with the ability to explore spatialized data in a highly individualized and granular manner (Elwood and Leszczynski 2013). Moreover, this listing illustrates the wide variety of data sources and uses (in particular the ability of Goocam to make unsecured webcams more visible) and reminds us that map mashups, like any technology, are deeply imbedded in the societal practices from which they emerged. As Castells (2001, 6) notes, “Neither utopia nor dystopia, the Internet is the expression of ourselves.” In the case of map mashups, this expression manifests across all aspects of society and begins with selecting data to be mapped.

**Getting data for map mashups**

In addition to using pre-existing and organized spatial datasets – often drawn from governmental sources such as the Census but also including directories and other online sources – mashups also build new datasets via online interfaces. For
example, one can obtain a listing of store locations from company websites that provide a store locator feature by conducting a series of queries by state or zip code (Zook and Graham 2006). Generally, there are two ways of building such datasets for use on a map mashup, data scraping and APIs, and both require a certain level of skill with computers and programming. Data scraping pre-dates APIs, which have emerged more recently to help the owners of data control access to their data; many data owners see usefulness, for either ideological or financial reasons, in sharing data more widely.

Data scraping (a.k.a. screen scraping) allows a skilled user to obtain data from a website by using computer code to simulate a user’s action and access websites or web-based databases to get data such as address listings, home sales, and so on and save it into a format that can be easily sorted and analyzed. It is an early Web 2.0 technology, but still widely used for pulling data from portable document format (pdf) files and non-API websites which do not offer an easier alternative. Because data scraping is a potentially illicit activity, data scrapers are often intentionally written to avoid raising red flags, for example, by limiting the number of requests sent to a server in a given amount of time.

In addition to being reliant on the technique of data scraping, map mashups rely on APIs for their base map content, and often for their secondary data sources. APIs are a set of commands or instructions through which users can request data from particular websites, for example, various social media services such as Twitter, Flickr, and Instagram. Depending upon the goals of a website, APIs can be documented, undocumented, partially documented, public, or not public. Public and well-documented APIs make it easy for mapping mashup to provide an up-to-the-minute reflection of a website’s content. Nonpublic, undocumented, and partially documented APIs can be accessed by users, but it requires the sort of reverse engineering of the site’s code that Rademacher initially employed with the Google Maps site.

A key issue facing map mashups is that the remixing of data – particularly in ways not originally envisioned by the creators of the data – can give rise to legal and ethical issues. As a result, data pulled via APIs are governed by the generating website’s terms of service or terms of use, and the more encompassing AUP (acceptable use policy), the code of rules set up by a website that a user is expected to abide by when using the site’s data. For example, Google (as do Bing, Yahoo!, and MapQuest) reserves the right to advertise on webpages where maps created via their map API are being used (Li and Yan 2010), while OpenStreetMap has a Creative Commons attribution license allowing for free reuse of their data. While somewhat of a grey zone – as aptly shown by Rademacher’s experience with Google – misuse of data could conceivably lead to prosecution, although a more common result is simply the cutting off of access – a death warrant for commercial applications. The use of user-supplied data is particularly complicated; as Li and Yan (2010) note, “In general, service providers do not hold the copyright but claim for free, non-exclusive rights to use, reproduce, modify, translate, publish, publicly perform and display, distribute, etc., the content.”

Challenges facing map mashups

While mashups have empowered individuals to create new and exciting maps, they have also led to the creation of maps that are problematic in various technical aspects, the result of little to no exposure to cartographic techniques and conventions. For example, a common failing of map mashups is related to projections – especially
the web standard of the Google Mercator projection. While this greatly distorts the size of continents at the global scale – the well-known “size of Greenland” problem – it has been adopted by Google since the company sees the primary use of Google Maps as a service for navigation, and the Mercator projection preserves angles and compass directions at the local scale. Of course, this distorts area, which can be extremely problematic for choropleths and other maps at the global scale. In the case of the “Gaza Everywhere Map,” a mashup which superimposes the outline of Gaza onto any zip code directed location in the world (see Nasri 2014), it is problematic given the area distortion inherent in the Mercator projection, which was completely unaccounted for in this map. To be sure, the goal of the map – to illustrate the size of this territory relative to an area familiar to a user – was likely achieved; nevertheless, it highlights how basic mapping knowledge, such as the distortion brought by projection, is often lacking within map mashups.

A second issue to arise with map mashups is the often naive approach to normalization of data, especially through the creation of heatmaps or interpolation of data that often is not appropriate given the type of phenomenon being mapped. As a result, non-normalized data about societal phenomenon are converted to heatmaps and presented as meaningful results despite the fact that these exercises primarily show population density rather than any real social phenomenon. This issue has become prevalent enough within amateur mapping that there’s even a famous Internet cartoon about it (Monroe 2014).

A third problem facing mapping mashups relates to the often underdeveloped sensibilities toward how mapping – particularly data drawn from Web 2.0 sources contributed by individuals – impacts privacy. The power of visualizing spatial data is that it allows a viewer to see patterns in the data that may not have been visible in the raw data themselves and to put the data to new uses (Elwood and Leszczynski 2011). The novelty of many map mashups is that they aggregate and map spatial data that have not previously been visualized; this is also what has led map mashups to occasionally run afoul of concerns surrounding privacy and surveillance. Examples of this include the website Rotten-neighbor.com, where individuals were invited to call out their neighbors for less than neighborly behavior, and the Girls Around Me app that culled public social media data of individuals checking in to a location via FourSquare to allow other users to surreptitiously view the social media profiles of women in their immediate geographic vicinity. While FourSquare check-ins were publicly available data, the aggregation of it and deployment via a map mashup upset many people, leading FourSquare to terminate their use of data and to the removal of the App from the Apple App Store (Bilton 2012).

These issues, however, are not simply limited to Web 2.0, as various map mashups using official government data have run afoul of privacy concerns – for example, the Connecticut Gun Ownership map published in a Connecticut newspaper following the Newtown School Shootings that showed the exact location of the home addresses of gun owners in Connecticut. While the information was already publicly available, it was available in table form and not as a map in which a user could zero in on a specific address (Maas and Levs 2012). The public outcry following the publication of the map surprised the newspaper publishers and ultimately led to the map’s withdrawal. A related example (FloatingSheep 2011) was the release of crime data in the United Kingdom that allowed for searches at the postcode level with returned results mapping these aggregated results to a specific point, perhaps the centroid of a road.
segment or postcode polygon. Concerns about this approach were that users of this search and mapping function would assume that crimes were taking place at specific points – rather than in an aggregated area – with repercussions for neighborhood perceptions and real estate values.

A fourth challenge relating specifically to map mashups drawing upon geotagged social media turns on questions of who is producing the data and how representative it is of the larger population. The claims to democracy on the Internet have time and again been discredited (Haklay 2013), with the majority of data being created by a small number of individuals (Stephens 2013). While these issues do not render geosocial media irrelevant (Crampton et al. 2013), they highlight the need for creators of map mashups to be cautious in how they are presented, a practice that is often downplayed. This mindfulness includes the recognition that “going beyond the geotag” – that is, considering nonspatial dimensions of social media data such as relationalities, temporal change, or qualitative measures – is important in understanding what map mashups might or should be visualizing (Crampton et al. 2013).

Conclusion

This entry highlights the history, data gathering techniques, and challenges associated with map mashups. Although a relatively recent phenomenon, the availability of new consumer mapping technology – including GPS and online tools – has resulted in the widespread creation and use of map mashups (Arnaud et al. 2010). In turn, location-aware mobile applications and social media have created new and more abundant sources of spatial data that can be gathered and funneled into new map mashups. This abundance, however, does come with a cost, particularly in new challenges in the way things are represented and the privacy of individuals and groups. As a result, map mashups are one of these most exciting, intriguing, and troubling spatial practices in the world today.

SEE ALSO: Volunteered geographic information; Web-mapping services

References


**MAPPING MASHUPS**


Maritime transport

Jean-Paul Rodrigue  
*Hofstra University, USA*

The geography of maritime transport

The geography over which maritime transportation operates is unique, with its combination of physical, strategic, and commercial imperatives. Physical issues are obviously stable across time, but strategic, and especially commercial, considerations continually shift with the ebb and flow of globalization. The physiography of maritime transport is composed of oceanic and river systems of circulation, which are well defined by criteria such as depth, currents, winds (historically significant), and the configuration of coastlines and passages. Although oceans account for 71% of the Earth’s surface, maritime transport mostly takes place only along specific routes that are regularly used in shipping itineraries. These routes are a function of obligatory points of passage, which are strategic locations, of physical constraints (coasts, winds, marine currents, depth, reefs, ice), and of political borders.

The configuration of the global maritime network is relatively simple and is organized along a circum-equatorial corridor linking North America, Europe, and Pacific Asia through the Suez Canal, the Strait of Malacca, and the Panama Canal (Figure 1). Limited use is made of the northernmost parts of the Atlantic, as well as the southernmost parts of the Atlantic, Indian, and Pacific oceans, because of hazardous navigation conditions (mainly ice) and their remoteness from the centers of economic activity. During the summer monsoon season (April to October), navigation may become more hazardous in the Indian Ocean and the South China Sea. Climate change may also provide maritime transport with shorter shipping routes through the Arctic, but such prospects remain limited at this point.

For obvious reasons, maritime routes try to follow the great circle route, a pattern that is readily observable on the configuration of transatlantic and transpacific routes. Core routes are those that are most used because they service major markets, while secondary routes are mostly connectors between secondary and main markets. Maritime transportation remains dominated by longitudinal (east–west) interactions, which are considerable. The advantage of maritime transport is not speed but continuity (regular services) and the capacity to handle large amounts of cargo. Rail and road transport systems are simply unable to support such a scale and intensity of traffic.

In part because of physiography, geopolitics, and trade, specific locations play a strategic role in the global maritime network. These locations are labeled as strategic passages and can be classified into two main categories. Primary passages are the most important since, without them, there would be limited cost-effective alternatives that would seriously impair global trade. Among these are the Panama Canal, the Suez Canal, the Strait of Hormuz, and the Strait of Malacca, which are key locations in the global trade of goods and commodities. The expansion of the Panama Canal, which is projected to come online in 2016, is an attempt to provide additional capacity, both in ship size and as a whole, to a strategic passage that has
been operating for 100 years. Secondary passages support maritime routes for which there are alternatives that would still involve a notable detour. These include the Magellan Passage, the Dover Strait, the Sunda Strait, and the Taiwan Strait. Historically, these passages have been subject to contention and the risk of closure, such as for the Suez Canal (which was closed between 1967 and 1975) and the Strait of Hormuz. More recently, maritime piracy has experienced a resurgence in light of the growth in the volume and value of trade, particularly in the vicinity of the Strait of Bab-el-Mandab, which links the Indian Ocean to the Red Sea.

However, several technical changes have transformed maritime transport in terms of its capacity and reliability. The first and most obvious is that most ship classes have become bigger, which has substantially improved the economies of scale that were always to the advantage of maritime shipping. The only remaining constraints on ship size are the capacity of ports, harbors, and canals to accommodate them. Second, the speed of ships has marginally improved, as containerships are faster than the conventional ships they have replaced. These slight speed improvements mean that, over transoceanic distances, a few days can be gained, which is important. Third, ships have become increasingly specialized, with many solely designed to carry one type of cargo such as containers, petroleum, vehicles, or liquid natural gas. Fourth, ship design has improved, enabling the construction of larger ships that are more energy-efficient. Fifth, automation has enabled ships to be manned by smaller crews while improving safety standards.

Like any transport activity, maritime transport is a derived demand supporting trade relations. These trade relations are also influenced by the existing maritime shipping capacity. There is thus a level of reciprocity between trade and maritime shipping as supply (maritime transport) and demand (trade) interact. Maritime transport
has been adapting to a number of commercial trends. A conventional trend is the growing demand for fossil fuels, raw materials, and grains, a process that is correlated with regional economic development. However, with the outsourcing and offshoring of manufacturing, it is the growth of the trade of parts and finished goods that has been the most common driver of change in maritime transport. The outcome has been far from uniform in the geography of maritime transport, with some locations better connected than others.

Not every location is directly connected to maritime transport, with landlocked countries facing the challenge of accessing global trade. The development of transnational infrastructure such as highway and rail corridors granting access to a port is of prime importance in facing this challenge. Being landlocked does not necessarily imply exclusion from international trade, but in many cases it does mean substantially higher transport costs, which may impinge on economic development. On average, landlocked countries have 50% higher transportation costs. These costs vary depending on the level of economic integration. For example, landlocked European countries have supportive trading agreements with their neighbors while agreements between landlocked African countries are much less effective.

Fluvial transportation is limited to the path of navigable rivers and canals and is slow and inflexible, but it still offers a high capacity and low costs. Ports are less relevant to fluvial transportation, but fluvial hubs are increasingly being integrated with maritime and land transportation, particularly with containerization. In regions that are well supplied by hydrographic networks, such as Western Europe, fluvial transportation can be a privileged mode of shipment between economic activities. In fact, several industrial regions emerged along major fluvial axes, as this mode was initially an important vector of industrialization.

The dominance of freight in maritime shipping

Maritime shipping has traditionally faced two disadvantages compared to transportation modes such as road and rail, even if such a comparison may be considered fallacious since they service different markets. The first is that maritime shipping has slow speeds, averaging 15 knots for bulk ships (26 km/hr) and over 20 knots (37 km/hr) for containerships. The second disadvantage is related to delays and time performance, particularly at ports where loading and unloading take place. This can involve several days of handling, particularly when break-bulk cargo is carried. Maritime transport is thus not attractive to shippers requiring rapid, reliable deliveries.

Maritime transportation is dominated by freight because there is no other means to move large amounts of cargo over long distances at a low cost. Maritime transport also used to dominate long-distance passenger transport, such as the North Atlantic liner ship trade, but this ended in the 1950s with the establishment of intercontinental air services. Passenger traffic is now a small niche market for maritime transport, mostly supplied by cruise shipping. There is, however, an active short-distance passenger market serviced by ferries, particularly in Western Europe (mostly in the English Channel and the Baltic Sea), Japan, Indonesia, and the Philippines. The dominance of freight in maritime transport has particularly been favored by:

- Energy and mineral cargoes: The rapid industrialization of developing economies has fueled additional movements of energy
(e.g., coal and petroleum) cargoes. These flows supplement the traditional north–south trade such as petroleum flows from the Middle East to the advanced economies of Europe, North America, and Japan.

- **Globalization:** The establishment of global supply chains has led to the maritime transport of a growing number of parts and finished goods.

- **Technical improvements:** Ships and maritime terminals have become more efficient and thus able to support long-distance sourcing. For instance, they are able to handle an impressive array of goods and commodities, particularly with containerization.

- **Economies of scale:** Ship and terminal facilities have become bigger, enabling a reduction of the unit cost of what is being transported. This was particularly the case for container shipping, a sector that has recently seen a rapid growth in the size of containerships.

There are two main types of maritime freight:

- **Bulk cargo:** freight that has not been packaged, such as minerals (oil, coal, iron ore, bauxite) and grains. These are dry or liquid bulk products that rely on the use of specialized ships such as oil tankers or ore ships and on specialized port and storage facilities. Bulk cargo usually has a single origin and a single destination, with services that are prone to seasonality, with the exception of energy transport. It supports the heavy manufacturing sector.

- **Break-bulk cargo:** general cargo that has been packaged in bags, boxes, drums, and particularly containers, which now represents the dominant break-bulk use of maritime transport. Break-bulk cargo usually has several origins and destinations. It supports manufacturing and retailing.

Technical improvements have blurred the distinction between bulk and break-bulk cargo, as both can now be unitized on pallets and also loaded into containers. Shipping grain and coal, which are bulk cargoes, in containers is becoming increasingly common.

About 100,000 registered commercial vessels are supporting global maritime transport. There are four major categories of these vessels. The first is the passenger vessels such as ferries and cruise ships. Passenger vessels used to be important conveyances for long-distance transport (liners), a role that has now been relegated to cruise ships. Still, the cruise industry has experienced a remarkable growth, particularly in the Caribbean and the Mediterranean, with about 20.3 million passengers taking a cruise in 2012.

A major category of cargo ships consists of bulk carriers, both liquid (particularly for petroleum) and dry (particularly for ores) bulk carriers (Figure 2). As a result of the propensity for economies of scale for this type of cargo, bulk carriers are the largest ships in existence. General cargo ships carry nonbulk cargoes and, prior to containerization, tended to be relatively small because the loading and unloading were done manually. These vessels have now been replaced by containerships that can be loaded and unloaded much more efficiently now that the cargo is in standard load units. Roll-on roll-off (RORO) vessels are designed to carry vehicles that can be directly driven on board. These specialized forms of transport were initially used as ferries but are now also extensively used as vehicle carriers. They transport vehicles (mostly cars and small trucks) from assembly plants to major consumption markets.

Maritime traffic is commonly measured in deadweight tons (dwt), which is the maximum amount of cargo that can be loaded on a ship without exceeding its operational design limits.
Between 2000 and 2012, the world deadweight tonnage almost doubled, propelled by a surge in global trade. Bulk ships tend to support regular services between two ports and may be used for specific trips depending on fluctuations in demand. The demand for maritime transport is often seasonal, as is the demand for grain transport and project cargo (e.g., construction material). General cargo vessels are commonly allocated to liner services whereby vessels will call at a series of ports according to a schedule.

Figure 2  World tonnage by cargo vessel type, 1970–2012 (millions dwt). Reproduced from United Nations Conference on Trade and Development (UNCTAD 2012).

Figure 3  Types of maritime routes.
Maritime transport networks

Maritime transport is an international industry, particularly in terms of ownership and the flagging of ships. The distinction between ownership and flagging is important: the corporations of developed countries remain the main owners of ship assets, but the flagging of the ships may involve different countries. While there are still many ships registered under national registries, the growth of the use of “flags of convenience” has been significant. A flag of convenience enables an owner to pay lower registration fees and lower operating costs, and be subject to fewer restrictions. Maritime transport is a highly deregulated industry, with open registry ships operating under fiscal shelter states such as Panama and Liberia. As of 2011, 68% of the global maritime transport tonnage was registered under a flag of convenience, up from 55% in 1995.

Maritime transport networks are designed to use the shortest route while being able to service the main markets. This leads to various compromises between the number of ports of call and the number of ships to allocate to specific trades. Containerization had a far-reaching impact on the configuration of maritime routes, particularly since, prior to containerization, loading or unloading a ship was a very expensive and time-consuming task; a cargo ship typically spent more time docked than at sea. The trend has now been reversed, and containerships spend more time at sea than in port, constantly on the move between ports of call. Therefore, containerships gave rise to a new network structure: what used to be mostly port-to-port has now evolved toward a more complex structure covering multiple markets. Maritime routes are organized according to the commercial services they support. These services can be divided into three main categories (Figure 3):

- **Port-to-port** represents the conventional service structure which concerns regular calls between two ports. Commonly ships move back and forth with a full load in one direction and an empty backhaul in the other. This low connectivity network structure is typical of raw material flows such as oil, minerals, and grains. These markets are usually serviced by chartered ships loading in one port (next to a major resource extraction area) and discharging their cargo in one to three ports.

- **Pendulum services** are typical of container shipping with regular itineraries covering a set of ports calls that are sequentially serviced. The sequence of these ports is obviously selected to maximize the load factor of the ships. The term “pendulum” refers to the shipping service moving back and forth between two maritime ranges (some pendulum services cover three maritime ranges). The most significant pendulum routes are between East Asia, North America, and Western Europe, the three main poles of the global economy.

- **Round-the-world services** also relate to container shipping and involve calling at a sequence of ports, often in both directions, such that the sequence involves a trip around the world. A limited number of ports per continent are serviced, but those ports are either major gateways or trans-shipment hubs. Thus, round-the-world services are an attempt to better connect longitudinal and latitudinal trade flows.

Historically, maritime transport was organized according to conferences, which are formal agreements between shipping companies operating over specific routes. Although they have the appearance of an oligopoly, they are designed to provide a level of rate stability in a context where rates have been subject to a
lot of variations, and to offer regular services. In view of the rising costs of ships due to their larger sizes, alliances have also been established. These alliances involve competing shipping companies pooling their vessels along major, and highly competitive, commercial routes with the expectation of higher utilization of their ship assets. This process has been taking place alongside a growing concentration of ownership and the emergence of a few very large shipping lines. As of 2013, the 20 largest shipping lines control 81% of the global container slot capacity, up from 42% in 1992.

Maritime transport is the linchpin of the global economy, acting as the physical support of its flows of freight. Although containerization was an important driver of change in maritime shipping, bulk cargo such as petroleum, minerals, and grains are fundamental and enduring trades, supporting large industrial sectors. Physical constraints have played an important part in the structure and organization of maritime transport. Such constraints have been partly mitigated by the construction of transoceanic canals, particularly the Panama (which is being expanded) and Suez canals. Climate change may also have an impact on maritime transport, with the opening up of Arctic routes, but such prospects remain speculative. It is the commercial factors related to the demand of energy, food, raw materials, parts, and finished goods that drive the dynamo of maritime transport. The fates of globalization and of maritime transport remain closely intertwined: maritime transport supports commercial flows, and more efficient maritime transport is also conducive to improved trade.

SEE ALSO: Containerization; Oceans and seas: human geography; Ports; Transport geography

Reference


Further reading

Marxist geography

Bernd Belina

Goethe University, Germany

Marxism – a global phenomenon with its own geography

Marxist geography, like Marxism, is named after one individual, the German thinker Karl Marx (1818–1883). This may sound like some kind of personality cult, but nothing could be more contrary to Marx’s own thinking, which focuses on how society as a whole functions and develops, and how capitalist societies are characterized by a mode of production based on exploitation. Academics and activists working within the Marxist tradition have become especially cautious when it comes to personality cults. Throughout the twentieth century, Marx’s name and ideas were exploited as a state ideology in the Soviet Union, Eastern Europe, and China, for example, in order to legitimize political regimes of highly problematic and sometimes outright dictatorial character that had very little in common with Marx’s own ideas.

During his lifetime, Marx’s work was adopted by thinkers and activists in many places all over the world to understand and to criticize both grievances caused by capitalism and capitalism itself. In the twentieth century, Marxism has become a global phenomenon, reflecting the spread of capitalist social relations and the struggle against their many different forms. Depending on historical epoch, national and language context, and strand within the discussion, there are numerous understandings of what Marxist thinking entails. While many of these understandings are closely related, there are cases where different strands of Marxism have little in common. Especially in the second half of the twentieth century, “Western” Marxists struggled to come to terms with Marxism–Leninism or Maoism, both of which drew on Marx but tended to portray his thinking as rather mechanical, as opposed to relational. Today, cleavages and disagreements within the Marxist tradition remain, but clearly take a back seat to the common goal: the comprehensive analysis of the functioning and steady development of capitalist society and the work toward alternatives. It was the latter that Marx famously called for when he wrote in the “Theses on Feuerbach” in 1845: “The philosophers have only interpreted the world, in various ways; the point is to change it.” On the most general level, the contribution of Marxist geographers to these endeavors lies in bringing to the fore the relevance of spatial differences in the analysis of capitalism and its alternatives, using concepts such as uneven development, geopolitics, urbanization, territory, place, region, landscape, and scale.

Marxist geographers are not particularly interested in defining a certain version of the academic discipline of geography. The division of academic labor into disciplines is mainly the result of struggles and exclusions within academia (with geography often having a hard time justifying its specificity vis-à-vis other disciplines) and, to a much lesser degree, a reasonable division of labor between academics. Just as Marx’s own work cannot be assigned to one academic discipline, the work of Marxist geographers is closely linked and sometimes indistinguishable from that of Marxists working in other disciplines.
Marxist Geography

Nature and metabolism

In volume 1 of his major work, *Das Kapital* (published in 1867 and revised in 1890), Marx makes reference to the “original sources of all wealth – the soil and the labourer” (Marx 1996, 508). Note that “the soil” is the translation for “die Erde,” which is more accurately translated as “the earth,” referring to nature in general. For Marx, as for geographers in the nineteenth century, nature forms the basis for all social processes. But, unlike the environmental determinism of nineteenth-century geography, he is very clear that the natural world – soil, air, water, minerals, plants, animals, and so on – does not determine culture but that human societies live in a complex relationship with nature which he terms *metabolism*, and which makes it difficult to even distinguish the border between the two spheres. It is through labor, understood as the “process in which both man and Nature participate” (Marx 1996, 187; note that Marx originally used the term “Mensch,” which is more accurately translated as “human being”), that this metabolism works. On the most general level, “arms and legs, head and hands” are themselves “natural forces” of the body, and through their use the human being “at the same time changes his [or her] own nature” (Marx 1996, 187). Marxist geographers have further developed this line of thinking by arguing that nature is actually produced in and through the metabolism with human societies, and that the production of nature under capitalism follows the imperatives of capital (Smith 1990). In another strand of the discussion in Marxist geography and in common with feminist literature, the human body, which was traditionally thought of as something merely natural, is discussed as produced in and through labor processes (Callard 1998).

Commodity production, value, and uneven development

The capitalist way to organize the metabolism between nature and society, and therefore the labor process, is marked by *commodity production*. In *Das Kapital*, commodities are the point of departure for Marx’s discussion of the capitalist mode of production. Commodities (including services) are useful things that are produced solely because they can be sold to others for a profit. The usefulness of things – the qualities that make up their *use value* – is reduced to a means to an end: it is only their quantitative *exchange value* that the capitalists who produce them are interested in. The exchange value is the amount of money that can be exchanged for the commodity because somebody else finds it useful and is willing and able to pay for it. It is only in the marketplace, in competition with other producers, that capitalists find out if they can realize an exchange value, and this is why it seems as if the value of a commodity is determined by what is going on in the marketplace: that is, supply and demand. But Marx argues that the exchange value is only the form in which the value of the commodity appears in exchange. *Value*, he argues, is inherent in the commodity prior to its entering the marketplace where it is realized. Following what is called the labor theory of value, he argues that the value of a commodity is determined by the average amount of socially necessary labor time that goes into its production. New value is produced in the labor process, not in exchange. Capitalists buy the commodity labor *power*: that is, the mere ability to work, in an exchange of equivalents for a certain price (the wage), and organize the labor process in a way that labor *time* is spent producing more value than is covered by the wage. This is how *surplus value* is produced and how capitalist
exploitation works: through the specific form in which the labor process is organized.

Marxist geographers have emphasized that capitalist production necessarily results in uneven geographic development (Harvey 1982; Smith 1990), since capitalists strategically use spatial differentiations to minimize costs (of means of production and/or labor power), to profit from the proximity to markets or to other capitalists such as suppliers, to create a spatial monopoly, or to use location as a competitive advantage (Harvey 1982). Spatial (re)location and organization thus play an integral role in the way in which production and exploitation develop, and result in a “spatial division of labor” (Massey 1984). In capitalism, uneven development is sometimes organized though spatial scales like the global scale of the world market, the national scale of the state, and the local scale, where production sites and labor power are concentrated (Smith 1990). Within labor geography, Marxist geographers look into the spatial strategies of laborers, such as how they mobilize their struggle for better working conditions and wages using spatial strategies, for example by organizing on a larger geographical scale (Herod 2001).

Classes, the state, and reproductive labor

The capitalist mode of production as described so far presupposes two classes: capitalists, as private proprietors of the means of production, and laborers, who have nothing to sell but their labor power. The existence of these two classes is the result of an (ongoing) history of “conquest, enslavement, robbery, murder, briefly, force” (Marx 1996, 705). Class struggle, enclosures, urbanization, colonialism, imperialism, geopolitics, coups d’état, police power, and wars were necessary, and are still used, to produce and reproduce these two classes. In Das Kapital, Marx emphasizes how wealth is produced and circulates in capitalism as a result of the distinction of these two classes with antagonistic interests. He further identified several more classes with particular interests, as well as other groups and institutions playing fundamental roles in the capitalist mode of production.

In “The Eighteenth Brumaire of Louis Napoleon,” first published in 1852, for example, Marx identifies the “small-holding peasantry” (Marx 1979, 187; emphasis original) as the class that would-be French ruler Louis Napoleon represented in his coup d’état in 1851. These small farmers were, as a result of the French Revolution, free from direct rule by their former overlords, but as they were not able to live off their land they ran up debts that made them unfree in a new respect. This economic position formed the basis of their particular interests that Louis Napoleon was able to represent. The proprietors of the banks to which they owned money, the “urban usurers” (Marx 1979, 190) as Marx calls them, are an example of yet another class: financial capitalists who lend money to other capitalists, laborers, and states. Their main concern is to secure the return of their credit with interest, for which reason they support the profit-making industrial capitalists and tax–collecting states while at the same time struggling with the former over the share of social wealth and the latter over the burden of taxation, banking regulations, and so on.

In the same work, Marx also mentions “the rivalry between capital and landed property” (Marx 1979, 128). Landowners who receive ground rent from capitalists or others for the permission to use a piece of land for a certain time constitute a class of their own, with interests that conflict with those of the renters. As landed property was discussed as a relevant source of wealth in the realm of agriculture in Marx’s time, he identifies the aforementioned rivalry with
“the old contrast between town and country” (Marx 1979, 128). Marxist geographers have contributed to an understanding of the role of private property over land, with a focus on the urbanization process that has shaped the world so fundamentally since Marx’s time. In order to understand the spatial structure of cities, the housing question, and gentrification, they emphasize “the co-ordinating functions that [ground rent] performs in allocating land to uses and shaping geographical organization in ways reflective of compensation and amendable to accumulation” (Harvey 1982, 333).

Also in the same work, Marx makes important contributions to an understanding of the state in the capitalist mode of production. On the one hand, he shows how the multitude of interests of various classes is represented within the state in an indirect manner, through parties, their campaigning, and propaganda, and through elections. On the other hand, he qualifies the state bureaucracy as “an artificial caste” (Marx 1979, 192) that is necessary to organize all other classes and that develops an interest of its own: the “maintenance of [the] regime” (Marx 1979, 192). Marxist geographers, often in close collaboration with Marxists from other disciplines, have made important contributions to our understanding of the state under capitalism and the way in which indirect rule, mainly in the interest of some classes over the vast majority of the population, is organized through coercion, but more importantly through hegemony, power relations, and the selectivities of state apparatuses. Geographers have tackled the following issues, among others: the way in which political coalitions of (parts of) different classes with competing interests form in places of different scale (cities, regions, nation-states), putting their common interest in the protection of value embedded in the land and the capitalist activities in these places on a similar level as their particular interests (Harvey 1982); how and why capitalist states are organized spatially in territories and scales, and how groups and classes try to strategically influence the scalar organization of states in their own interest (Smith 1990); and how states use their powers of policing, punishment, and incarceration to order spaces and populations in socially selective manners, such as along racial lines (Wilson Gilmore 2007).

Large groups and many activities seem to stand outside of the economic and political relations discussed so far, especially reproductive work and housework, which are mainly done by women. While Marx never explicitly engaged in a discussion of women’s role in capitalism, feminists in the Marxist tradition have highlighted the role of women in the biological reproduction of laborers (i.e., giving birth) and the social reproduction of capitalist social relations in general by guaranteeing the reproduction of the (often) male “breadwinner’s” labor power. Marxist geographers have contributed to these discussions by, for example, deciphering the way in which reproductive labor is hidden from sight in urban spaces and by showing how reproduction differs between places (Katz 2004).

Pathways of Marxist geographies

Marxist geographies, while based on the work of Marx, differ in time and space, reflecting different economic and political situations and emancipatory struggles in which Marxists in general, and Marxist geographers in particular, were engaged. In the anglophone world, Marxist geography was central to the theoretical development of geography in the 1970s and 1980s, before being challenged on a variety of levels by phenomenology, structuration theory, and especially poststructural and feminist theories that constituted a “critical geography” that
wanted to overcome the perceived shortcomings of Marxist radical geography. Building on the work of authors such as Michel Foucault, Judith Butler, Ernesto Laclau and Chantal Mouffe, many contributions of anglophone critical geography in the 1990s seemed to place a wedge between their endeavor and the Marxist tradition (sometimes ignoring the close relation of poststructuralist authors to the Marxist tradition). This development is rather specific to the anglophone context, though. It reflects, among other things, the political upheaval at universities in the late 1960s, the sometimes problematic relationship of Marxist political organizations to feminist movements and movements of people of color, and the specific “publish or perish” environment at British and American universities since the 1990s which gave an incentive to authors to focus on the novelty of their contribution. By contrast, in Latin America the work of Brazilian geographer Milton Santos (1926–2001), who included the category of race in his Marxist work, remains a central point of reference. Following his example, and as a result of the close relationship of Marxist geographers to social movements and their existential struggles, a far more open exchange between different kinds of critical theories has evolved. In France, important geographers such as Pierre George (1909–2006) were members of the Communist Party (a relevant political force between the 1950s and 1980s), but little Marxist theory is reflected in their work. In Japanese economic geography, Marxism was present from the 1920s and especially strong in the 1970s, following the country’s rapid industrialization after World War II and its negative impacts on Japan’s social and natural environments. In the Soviet Union and Eastern Europe, Marxism–Leninism functioned as a state ideology, but it also made possible an early and radical critique of traditional geography as well as interesting attempts to build an applied Marxist geography with a strong focus on territorial planning in the service of “real existing socialism.” In the current conjuncture, with English being the lingua franca of global academia and very few translations from other languages into English, and with excellent Marxist work being produced in English-language geography, anglophone Marxist geography has become hegemonic within Marxist geography (as reflected in the references of this entry).

The relational thinking of Marxism allows us to understand these different pathways as appropriations of Marx’s basic contributions to the understanding of capitalism and its alternatives in different economic and political contexts. In this sense, the development of Marxist geography illustrates one of Marx’s strongest phrases, which also captures nicely the inherently political nature of Marxist thinking: “Men [originally ‘Die Menschen,’ i.e., ‘human beings’] make their own history, but they do not make it just as they please; they do not make it under circumstances chosen by themselves; but under circumstances directly encountered, given and transmitted from the past” (Marx 1979, 103). Building on the insights of Marx and the tradition he founded allows us to understand the ways in which human beings make their own geographies in varying circumstances and to contribute to progressive attempts to change them.

SEE ALSO: Critical geography; Labor geography; Radical geography; Social geography; Social justice; Social movements; State, the

References

MARXIST GEOGRAPHY

Mass movement processes and landforms

Francesco Dramis
Roma Tre University, Italy

Mass movements may be defined as “downslope movements of rock (hard rock), earth (soft clayey rocks) or soil (loose debris) masses occurring when their resisting forces cannot resist gravitational stress.” They include a large variety of phenomena that mobilize with different modalities (type of displacement, type of movement, speed, run-out distance, etc.) masses of extremely various dimensions. Similarly to other geomorphological processes, these phenomena may heavily interact with human settlements and activities, being responsible for loss of human lives and damage to built-up structures and properties.

From the definition, mass movements are favored by conditions and processes causing either an increase in gravity stress and/or a decrease in shear strength in the involved material.

An increase in gravity stress may be induced by: (i) long lasting heavy rainfall causing higher groundwater levels in the slope materials and a related increase in their specific weight (the same effect is produced by water leakage from water pipes or sewers); (ii) earthquake shocks, which may increase (even if for a short time) the downslope stress component by the transitory addition of seismic acceleration (up to more than 0.5 g in high magnitude earthquakes) to the gravity acceleration; (iii) transient or permanent co-seismic oversteepening of slopes; (iv) oversteepening of volcano flanks during eruptions; (v) removal of lateral support because of valley incision, coastal erosion or glacial oversteepening of valley slopes followed by glacier retreat; (vi) removal of basal support due to underexcavation of the slope foot by streams, sea waves or human activity; (vii) slope overloading due to debris accumulation by other mass movements, running water, or pyroclastic fall. Comparable effects are produced by artificially constructed embankments and constructions, or by the growth of big tree forests.

The main parameters controlling the shear strength of a geologic material are cohesion, internal friction angle, and normal stress. Friction is the force that resists the sliding movement of a material over a surface. Its value depends on different parameters, the most important of which are the roughness of the potential sliding surface and the extent of the contact area. The internal friction angle is the threshold angle above which the material is unstable and moves down by failure. Cohesion is the tendency of the particles that compose the material to be linked; it depends on several factors, such as compaction, cementation, chemical bonds, and plant root networks. Normal stress corresponds to the weight of the material overlying the potential sliding surface.

Shear strength \(s\) is quantitatively defined by the Coulomb–Terzaghi equation:

\[
s = c + \sigma' \tan \varphi
\]

where \(c\) is the cohesion, \(\sigma'\) the normal stress, and \(\varphi\) the sliding surface angle. When water fills the pores, a part of normal stress is supported by it, resulting in a reduced shear strength. In this case, the Coulomb–Terzaghi equation is:
where $u$ is the neutral pressure, defined as the water pressure within the material interstices (pores). When air fills the pores, the neutral pressure is zero.

The strength of geological material is mainly dependent on its composition, permeability, and pore pressure. In the case of hard rocks, the following factors should be taken into account: rock composition and structure; weathering degree; joint spacing, orientation, width, continuity, and infill; and groundwater flow within joints.

A decrease in strength of the slope materials may be induced by different causes, including: (i) an increase in pore pressure because of infiltration of rain and snow melt or pipeline/wasteway leakage (this effect is significant where expandable clay minerals, such as montmorillonite, are present); (ii) seismically induced expansion/compression cycles of pore pressures in water-saturated granular sediments (in fine sands and silt, the material structure may break down (liquefaction)); (iii) seismically induced ground faulting and fracturing; (iv) progressive weathering; (v) gravity-induced enlargement of joints; (vi) joint enlargement by tree roots; (vii) climate-induced melting of permafrost ice and related high water pore pressures in ice-cemented debris; (viii) climate-induced warming of permafrost ice filling potential detachment/yielding surfaces (Davies, Hamza, and Harris 2001).

The stability conditions of a hillslope may be expressed by the safety factor ($F$), corresponding to the ratio between resisting forces and driving forces: with $F < 1$ if the slope is unstable; with $F > 1$ if the slope is stable. The safety factor of natural slopes commonly ranges between 1 and 1.3. The steepest slope angle that a cohesionless material can keep without losing its stability is called the angle of repose.

Gravity-driven masses may be displaced downslope through some basic mechanisms: (i) detachment from the slope by a newly generated fracture or shear surface; (ii) fall in free air; (iii) sliding over a newly generated shear surface whose shape is controlled by the mass structure; (iv) spreading of a subhorizontal layer over a softer material along a non-defined basal shear surface; (v) flow over the slope surface or in a preexisting slope incision; (vi) creep, an extremely slow gravity-driven deformation of slope material.

Mass movements may be distinguished in four main groups: (i) generalized near-surface slope wasting; (ii) landslides; (iii) deep-seated gravitational deformations; (iv) subsidence.

**Generalized near-surface slope wasting**

Generalized near-surface slope wasting includes two widespread gravity-driven processes: soil creep and solifluxion.

Soil creep is the slow (a few centimeters per year) movement of eluvial/colluvial material induced by several different actions (such as rain splash, shrinking/swelling, plant rooting, burrowing by worms or other animals) displacing the single grains in different directions but with the result of shifting them downslope. All the single grains make up a unique mass, involved in a continuous deformation of the creep type. Frost creep, a particular type of soil creep, commonly occurs on debris slopes affected by seasonal freeze/thaw cycles. Winter freezing pushes upward the debris fragments at a right angle to the slope; subsequently, the spring thaw of ground ice causes their vertical fall and downward displacement. The surface indicators of soil creep are hook-shaping of sedimentary layers, curved upslope tree trunks, bent walls and fences, inclined poles, offset roads, walls, and fences. Soil creep often forms long
narrow steps called terracettes, especially where animals walk across the slope.

Solifluction is the slow downhill flow of the soil layer saturated by rainfall or snowmelt. Gelifluction is a particular case of solifluction affecting the permafrost active layer: affected by seasonal freeze/thaw cycles, the meltwater cannot penetrate the ground because of the presence of permafrost ice, thus inducing high pore pressure in the soil particles and making them flow downward as a cohesive mass. Sometimes, the underlying permafrost table may act as a sliding surface for the saturated soil. Solifluction landforms include slope surface undulations with water ponds, lobes, sheets, and turf-banked terraces. Plowing boulders are large blocks that move slowly downslope pushing a soil rampart ahead of them, leaving a furrow behind.

**Landslides**

Landslides are the fastest geomorphic processes by which gravity tends to re-establish the disequilibria induced by tectonic uplifting on the Earth’s surface. The term includes a large variety of downslope movements that involve masses of various dimensions, ranging from small blocks to entire slopes and hills (large-scale landslides). Due to the large masses involved, large-scale landslides are occasionally characterized by scale effects, such as rock melting, carbonate dissociation, and fluid vaporization along the shearing surface (Goguel 1978). Landslides of different types and dimensions (up to several tens of cubic kilometers) also commonly occur in subaquatic environments as a result of earthquake shocks. They can affect the offshore areas near the coast and can extend backward to the coastal belt or the sediments accumulated on the edge of the continental scarp. Landslides are also common on the slopes of submarine volcanoes. The aquatic environment facilitates the movements and enables the displaced materials to travel considerable distances. Large-scale submarine or coastal landslides may generate catastrophic tsunamis.

Landslides are classified in different ways. The most generally adopted scheme, based on the types of movement and the nature of involved material, is that proposed by Varnes (1978) and more recently improved by Dikau et al. (1996). The landslide types shown here primarily refer to this scheme. As recognized for all classifications, the list is only a simplified representation of the real phenomena that, in many cases, may include more than one landslide type or may change downslope from one type to another. Different types of landslides are locally associated to form landslide systems, in which each component moves with a different mechanism and recurrence time. In some cases, landslides are superposed to deeper levels, the lowermost of which is reactivated only by major triggering events.

**Falls**

A fall is the detachment of a block of solid rock (rock fall), soft clayey–sandy rock (earth fall) or unconsolidated/poorly consolidated debris (soil fall) from a cliff, along newly generated or pre-existing rupture surfaces (joints, layering, fractures, etc.) progressively enlarged by gravity stress, recurrent earthquake shocks, frost shattering or plant-root wedging. The detached mass may move rapidly in free air or roll and jump on the slope surface. When impacting the ground, the fallen mass commonly breaks into smaller fragments that are deposited at the base of the cliff giving rise to a talus slope, an accumulation belt of debris whose slope gradient is controlled by the internal friction angle. The size of the falling mass may range from decametric blocks to single small fragments (commonly not considered as true landslides).
Toppling is the detachment of a prism of rock (rock toppling), earth (earth toppling), or soil (soil toppling) from a cliff, along a pre-existing subvertical surface. After detachment, the mass rotates clockwise on its base and tips over the slope foot or falls in free air, breaking into rolling or jumping fragments. The landslide heaps are similar to those produced by falls.

Rotational slides

Rotational slides are downslope movements of rock, earth or soil along newly generated, concave upward subcircular shear surfaces. Rotational earth slides (Figure 1) are commonly characterized by a step-like evolution, alternating short-lived reactivation phases with long periods of quiescence that may range from seasonal to multi-annual, decadal or centurial, up to more than 1000 years. There are different types of rotational earth slides: (i) single slides, which may enlarge upslope (retrogressive slide), downslope (advancing slide) or laterally (enlarging slide); (ii) multiple slides, made of more than one sliding units, each of which has a sliding surface intersecting a common, basal yielding surface; (iii) successive slides, consisting of a series of individual rotational slides one above the other. The main scarp is concave valley-ward and is commonly bordered upslope by tension cracks. Scarps and steps are often present also along both landslide flanks. The body of a rotational earth slide often consists of an almost undisturbed mass whose surface is less inclined than the original one or is
even inclined upslope, resulting in a topographic depression. This latter formation may host superficial water, causing concentrate infiltration into the ground and further slope destabilization. Transverse fissures, steps, trenches, and uplifted blocks on the landslide surface indicate the occurrence of minor shear planes. The landslide foot (the portion of the landslide beside the lower limit of the sliding surface) often shows transverse pressure ridges and compressional shear planes. The landslide toe (the lower margin of the landslide) often evolves to earth flow. Rotational soil slides, also known as soil slips or slumps, usually occur on steep slopes. They move quickly and may be extremely dangerous. Rotational rock slides are characterized by extremely large depth (up to several hundred meters) and involved volume (up to several thousand cubic meters). They commonly form on very high slopes and represent the final collapse of deep-seated gravitational deformations, which are described later.

Translational slides

Translational slides are downslope movements of rock, earth, or soil masses along a newly generated planar shear surface, emplaced over pre-existing (predisposed) “weak” structural elements, such as layers, faults, and fractures, that are inclined less than the slope. The detachment surface is subvertical in the upper part (main scarp) and less inclined (up to subhorizontal) in the lower part. Tension cracks are often present upslope of the main scarp. Low diverging downwards scarps commonly border the landslide body. On high steep slopes, translational rock slides move down with extremely high speed as rock avalanches impacting the ground and disintegrating into fragments that climb over the opposite valley-side (Figure 2). Occasionally, the landslide debris dams the valley bottom, causing the formation of a temporary lake. Translational earth slides frequently show a step-like evolution. The landslide bodies commonly consist of single (or a few) undisturbed units crossed by tension fractures. Also in this case, the landslide toe may evolve to earth flow. Translational debris slides on high steep slopes may change to extremely rapid, dangerous debris avalanches that can travel for great distances even along very low gradients.

Soil spreads

The term of soil spreading describes the gravity-driven translational movement of a soil mass over a softer material along a non-defined basal shear surface. This results from a loss of strength (liquefaction) in the upper layer (commonly quick clay soils) induced by prolonged periods of heavy rainfall/snowmelt or earthquake shaking. The expanding soil mass breaks off into tilting, subsiding, heaving, and overthrusting blocks that produce a typical hummocky topography. The movement starts in the lower sector of the slope and rapidly extends upslope and laterally, giving a pear shape to the failure area. Because of their rapid evolution, soil spreads may be extremely dangerous.

Flows

Flow is a fluid-like movement of loose sediments that have a velocity which ranges from extremely slow to extremely rapid in relation to the slope angle, grain size, and water content. Almost all flows move with a non-Newtonian behavior (their viscosity increases with an increasing applied stress), except for materials with high water contents. Debris flows are characterized by very high speed and frequently cause destruction and death. With increasing water content, debris flows gradually develop into torrential flows. Their typical geomorphic features include:
Deep-seated gravitational deformations

Deep-seated gravitational deformations are complex phenomena whose genesis and evolution are controlled by several factors, among which structure, relief, tectonic, and seismic activity play a primary role (Dramis et al. 1995). The most recurrent typologies are sackungs, rock spreads, and deep-seated block slides.

Sackungs

Sackungs, also known as gravity sags or rock flows, commonly develop in hard rocks over a high slope. They are characterized by the following morphological features (Dramis et al. 1995): (i) surface extension generally more than 1 km²; (ii) thickness of deformed mass ranging from tens to hundreds of meters; (iii) displacement reduced with respect to the involved mass volume; (iv) absence of a continuous shear surface; (v) extremely slow evolution with long periods of inactivity or reduced activity alternated with short-lived activations, often triggered by earthquakes; (vi) deformational mechanism of the creep type, with accelerations and ruptures; (vii) kinematics often influenced by active or residual tectonic stress; (viii) reduced influence of...
topography in respect to structure in controlling the deformation location and typology; and (ix) surficial landslides affecting the deformed slope. The upper part of the deforming slope shows high-angle extensional shear planes giving rise to graben-like depressions (trenches) and counter-slope scarplets. In contrast, the lowermost slope portion often presents compressional features such as bulging and low-angle shear planes. Smaller surficial landslides often affect the deforming slopes. Sackungs may be considered as preparatory stages for large-scale rock slides even if, usually, they may not complete their evolution (Dramis et al. 1995).

Rock spreads

Rock spreads are characterized by horizontal extension of the relief, balanced by shear or tensile fractures. They may occur in the following ways: (i) bilateral spreading without any shear surface or basal flow causing the formation of double ridges on top of high elongated reliefs in stiff homogeneous rocks (Jahn 1964); (ii) bilateral spreading associated with both trench-like summit depressions and shear planes at the base (Beck 1968); (iii) tectonic-gravitational spreading (Dramis and Sorriso-Valvo 1994) affecting morpho-structural reliefs produced by active thrusting (spreading starts because of tectonics...
but then evolves because of gravity); (iv) lateral rock spreading involving thick rocky masses horizontally overlying softer layers on top of a high relief (Cruden and Varnes 1996). The upper mass induces squeezing and expansion in the lower layer, which causes, in turn, the fragmentation of the hard rock into blocks, bordered by progressively enlarging fractures. These blocks may subside, translate, rotate, or tilt, giving origin to counter-slopes, steps, and trenches. The edges of the expanding mass are commonly affected by minor landslides.

Deep-seated block slides

Deep-seated block slides (Figure 4) are characterized by relatively small displacements of thick rocky blocks overlying gently sloping, less competent layers (Dramis et al. 1995). Their surface effects are large scarps and trenches in the upper and intermediate portions of the slope, and bulging associated with compressive shear planes in its lowermost portion.

Subsidence

Subsidence may be defined as a ground surface lowering or failure due to removal of underlying support. It may occur at different rates from extremely slow movements to sudden collapses that may result in subcircular cavities (sinkholes). In the latter case, catastrophic effects may be induced to buildings, infrastructures, and human lives. The possible causes of subsidence include both natural and anthropic actions. Among the first, the most common are: natural compaction of sediments; removal of fine sediments by circulating groundwater; collapse of near-surface roofs of karst caves in calcareous/gypseous rocks; dissolution of salt deposits; thermokarst in permafrost areas; upwelling artesian groundwater; seismic liquefaction; co-seismic ground lowering. The main anthropogenic causes are mining; extraction of gas, oil, and water; and overloading of the ground surface by buildings.

Figure 4  A deep-seated block slide near San Mango, Calabria, Italy. Note the large detachment scarp across the ridge (photograph by F. Dramis).
Gravity tectonics

Gravity-driven movements may also involve masses at the geological scale giving rise to imposing tectonic structures such as thrusts and nappes. These features may involve any rocks, including hard rocks, which move at an extremely low rate by creep deformation and sliding. During geological times, these phenomena have repeatedly affected orogenic structures, causing the removal of enormous masses of overlying rocks (tectonic denudation).

SEE ALSO: Applied geomorphology; Environmental hazards; Geomorphic hazards; Hillslopes; Mass movements in periglacial environments; Natural hazards and disasters

References


Further reading


At present, the level of confidence is high that the global average temperature of the past few decades was warmer than at any comparable period during the last 2000 years (IPCC 2013). Current evidence also suggests that temperatures during the last 25 years were higher than at any period of comparable length since BCE 900 for many, but not all, regions. The Intergovernmental Panel on Climate Change (IPCC) also reports that most land regions of the world show an increase of high temperatures for the past 50–100 years, usually expressed as the 90th or 95th percentile of the long-term record, such that the frequency of hot days has almost tripled in Europe since 1880. Based on simulation runs forced with different greenhouse gas emission scenarios, the IPCC (2013) also concludes that the rate of warming until the end of the twenty-first century is likely to be faster than ever recorded from historical or proxy records. In high-mountain regions, the evolution of mean and extreme temperatures will likely be comparable to what has been described above; however, studies specifically focusing on trends at high elevations have not been published so far.

The capacity of air to hold moisture is a function of temperature. As a consequence, global warming is likely to lead to an overall greater frequency and magnitude of heavy precipitation events. An increase in the frequency and intensity of extreme precipitation events has been identified in different sets of observational data from several regions of the world (IPCC 2013). For the future, projections likewise suggest decreasing return periods of extreme rainfall events.

Changes in temperature and precipitation are considered likely to have a range of secondary effects, including on the extent of glaciers, the distribution and duration of the snow cover, and on the temperature and 3-D distribution of permafrost. While there is a theoretical understanding for increased mass-movement activity as a result of predicted climate change in mountain environments, changes in activity can hardly be detected in observational records. Uncertainty also remains considerable as a result of error margins inherent in scenario-driven global predictions, and due to the lack of spatial resolution of downscaled projections.

This entry aims to document the role of climate variability and change on mass-movement processes in mountainous regions through the description and analysis of selected, recent mass movements where the effects of global warming and the occurrence of heavy precipitation are thought to have contributed to, or triggered, events. It addresses possible effects of future climatic changes on the occurrence of future mass-movement processes and speculates about possible consequences of climate and mass movements on hazards and risks.
Glacier downwasting and the formation of new ice-marginal lakes

One of the most obvious consequences of climate change at high-elevation sites is the widespread retreat and disintegration of glaciers (Zemp et al. 2007). The consequences for natural hazards are multiple and include the formation of ice-marginal lakes, ice avalanches, and mass movements originating from the recent debuttressing of previously glacierized landscapes.

A prominent phenomenon associated with glacier retreat and changes in glacier geometry is the formation and growth of ice-marginal lakes. Hazards related to glacier retreat and the formation of glacial lakes have been recognized for several decades. Indeed, severe disasters have occurred in the past as a result of outburst floods from glacial lakes in various high-mountain regions of the world (Worni et al. 2014). Rapid lake formation and growth has accelerated in recent years. Many of the current lakes have typically formed within the past decades and are often located at the terminus of glaciers, where subglacial topography has been overdeepened by the glacier. Lower Grindelwald glacier (Figure 1a), for instance, has been in strong retreat since its maximum glacier extension during the Little Ice Age around 1860. Downwasting of the glacier in its terminal part resulted in a loss of between 60 and more than 80 m of ice thickness between 1985 and 2000. In recent years a glacial lake started to form in the terminus area of the glacier (Figure 1a). In 2004 and 2005, the lake had a limited volume but has subsequently continuously grown in the spring and early summer seasons, resulting in lake volumes of 250 000 m$^3$ in 2006, 1.3 million m$^3$ in 2008, and 2.5 million m$^3$ in May 2009, and the occurrence of a glacier lake outburst flood in 2008 (Worni et al. 2014).

Glacier debuttressing and the occurrence of rock slope instability

The combined effects of glacier downwasting and debuttressing on rock and moraine slopes, permafrost degradation, rockfalls, and...
debris-flow activity, all interacting with the formation and growth of glacier lakes and further glacier decay, are often remarkable. At the Lower Grindelwald glacier, the rock slope failure above the glacier terminus (Figure 1b) is a textbook example of the effects of glacier retreat, downwasting, and associated debuttressing effects on rock slope stability, and could, in fact, serve as a model case for increasingly destabilized future high-mountain environments. The response of a rock slope to glacier downwasting has been reported to potentially result in (i) large rock avalanches, (ii) large-scale, progressive and slow rock mass deformation, and (iii) frequent rockfall events (Ballantyne 2002). The three modes of response are all consequences of stress redistribution and release and may act in a combined way. Rock slope failure thereby often represents the result of slope steepening by glacial erosion and unloading or debuttressing due to glacier retreat.

Examples that may be related to glacial oversteepening or debuttressing include the Brenva and Troiet rock avalanches in the Mont Blanc massif in the eighteenth and twentieth centuries, the Sherman glacier-rock avalanche in 1964 in Alaska (although earthquake-triggered), a significant number of rock avalanches in the Karakorum, and several rock avalanches in British Columbia, Canada, just to name a few. The nature, timing, and scaling of rock slope failures due to glacial debuttressing is strongly conditioned by geology, in particular by lithology and structure, that is, rock mass strength, orientation and inclination of discontinuities, and density and depth of joint networks. The timescale of failure and its delayed response in relation to glacier retreat has been much debated (Ballantyne 2002). Recent advances in geochronology have helped to better constrain the ages of many rock slope failures in alpine environments. The picture that evolves from the different studies and approaches is one of a varying response of rock slope failures to glacial retreat and downwasting. Failures can occur on timescales of $10^1$–$10^4$ years, depending on the glaciation history, topography, or geology.

**Temperature increase and permafrost degradation**

Important effects of climate change on mountain slope stability are, furthermore, related to warming and thawing of permafrost. Permafrost exists in many steep rock slopes in high-mountain environments and its degradation due to global warming can affect slope stability. Although this link might be intuitively clear, the mechanisms of permafrost degradation and related slope stability are rather complex, and the corresponding research field is relatively young. As a result, many aspects and links remain uncertain to date because of the complexity of interacting processes.

Evidence comes from a number of recent slope failures in permafrost areas, including mass movements at scales that range over several orders of magnitude from block and rockfall to rock avalanches (volumes of about $10^2$–$10^7$ m$^3$), observed predominantly in the Alps but also in other mountain regions. Several studies have demonstrated that the heatwave in the summer of 2003 and the related excessive thawing of the active layer of permafrost bodies have resulted in an unusually high number of rockfalls at high-elevation sites in the European Alps (Gruuber, Hoelzle, and Haeberli 2004). An increase in large rockslide failures has been observed in the Swiss Alps and neighboring areas for the past two decades as compared to the twentieth century, and more particularly in the Mont Blanc area, for which a detailed study exists.

Examples of large rock failures in the Alps include the 2004 $2.5 \times 10^6$ m$^3$ rock avalanche
from Punta Thurwieser, Italy; the 1997 $2–3 \times 10^6$ m$^3$ Brenva rock avalanche, Mont Blanc region, Italy; the 2006 about $10^6$ m$^3$ rockslides from Dents du Midi and Dent Blanche, Switzerland; the 2007 rock avalanche from Monte Rosa east face, Italy; and the December 2011 $2–3 \times 10^6$ m$^3$ rock avalanche at Piz Cengalo, Val Bregaglia, in the southern Swiss Alps. Many other regions have also experienced major rock failures. In the Chugach Mountains, Alaska, an approximately $50 \times 10^6$ m$^3$ large rock and ice avalanche was released from Mount Steller in 2005. Some $10–20 \times 10^6$ m$^3$ of rock and ice released from Dzhimaraikh-Khokh (Caucasus) entrained large parts of the Kolka glacier in 2002, resulting in a devastating $>100 \times 10^6$ m$^3$ ice-rock avalanche (Figure 2).

Notably, climate change affects permafrost in rock slopes on different spatial and temporal scales. Knowledge of the temperature distribution and dynamics at depth, and related 3-D effects, is, in fact, important in order to improve understanding of how climate change affects slope stability. Noetzli et al. (2007) modeled temperature and distribution of permafrost in idealized 3-D topography and demonstrated that

![Image of the Grabengufer rock glacier and Dorfbach debris-flow areas and the Grossgufer rockslide at Randa. Solid lines indicate the rock glacier and rockslide scar, and the dashed line the trajectory of the Dorfbach debris flows, fueled by the debris from the instabilities of the rapidly moving Grabengufer rock glacier (Base image from Google Earth, October 2009).](image)

**Figure 2** Image of the Grabengufer rock glacier and Dorfbach debris-flow areas and the Grossgufer rockslide at Randa. Solid lines indicate the rock glacier and rockslide scar, and the dashed line the trajectory of the Dorfbach debris flows, fueled by the debris from the instabilities of the rapidly moving Grabengufer rock glacier (Base image from Google Earth, October 2009).
contemporary permafrost temperatures at depth are significantly influenced by the climate of the past millennia, including the last Ice Age. In the perspective of such timescales of heat diffusion, the twentieth-century warming may only have reached a depth of some tens of meters on steep slopes at high elevations. Due to the large time lag of heat diffusion, permafrost at greater depth may be present where surface temperatures no longer favor its occurrence. As a consequence, potential effects on slope stability by recent warming may currently have penetrated to depths of tens of meters but will continue to reach increasingly greater depths with future warming.

Although 3-D thermal modeling of climate change effects on mountain permafrost is an important step forward, the processes through which air temperatures influence slope stability are not understood in sufficient detail. Davies, Hamza, and Harris (2001) demonstrated with laboratory tests that thawing permafrost is accompanied by a reduction of shear strength in ice-filled clefts. The resulting variations and the increase of hydrostatic pressure in previously ice-filled fractures might, therefore, result in reduced slope stability. Furthermore, slow growth of segregation ice over long periods of time during permafrost aggradation has been suggested to widen joints in their frozen state, and thus contribute to a reduction in stability when permafrost thaws under conditions of warming (Harris et al. 2009).

Permafrost degradation in steep bedrock is thought to be effective through heat conduction and by advection of heat through water percolating in fractures. Degradation by conduction is enhanced in convex topography (i.e., ridges, spurs, and peaks) due to warming from several sides. Several of the aforementioned large rock slope failures occurred from such ridge and spur situations. However, this understanding based on theoretical considerations and modeling assumes idealized homogeneous rock conditions without discontinuity systems. In nature, rock joints are likely to have a major influence on the rate and extension of permafrost degradation and slope stability. Recent measurements in rock slopes of the Matterhorn (Switzerland), for instance, revealed the importance of meltwater penetration into cleft systems as a driver of rock deformation.

### Changing sediment reservoirs and sediment supply

In addition to future changes in temperature and rainfall intensity, changes in sediment supply and land use are important determinants for mass-movement frequency and magnitude. Recent observations at several sites in the Swiss Alps indicate that sediment supply can, in fact, change significantly as a result of permafrost degradation of rock and scree slopes or mass movements related to other processes (Huggel, Clague, and Korup 2012).

While the average flow speed of rock glaciers in the Valais Alps (Switzerland) was usually below 1 m year$^{-1}$, ground and remote sensing based monitoring has revealed acceleration of rock glaciers’ surface flow-speed up to 4 m year$^{-1}$, and occasionally up to 15 m year$^{-1}$, in recent years (Figure 3) (Stoffel, Tiranti, and Huggel 2014). This phenomenon has been reported over wide regions of the Alps, with morphological features similar to those observed with landslides, such as transversal cracks, surface subsidence in the upper part, and rapid advance of the frontal part. Further field studies thereby indicate that rock glacier surface speed is increasing with increasing mean annual air temperature and, as such, warming exerts indirect control on debris-flow magnitude and frequency.
Reconstruction of the trajectory of the 2002 Kolka ice-rock avalanche in the Caucasus (Russia) overlain on a QuickBird false color infrared image acquired on September 25, 2002, five days after the avalanche. A massive slope failure in glacier ice and bedrock in the Northeast face of Dzhimarai-Khokh at about 4300 masl impacted Kolka glacier. A large portion of Kolka glacier was then destabilized to form a high-speed avalanche that traveled at maximum speeds of >300 km h⁻¹ down the Genaldon valley. The avalanche was dammed at the entrance of a gorge and formed a massive dam of about 130 × 10⁶ m³ ice and rock debris. Liquid parts of the avalanche traveled further downstream for about 15 km, devastated the valley, and caused about 120 fatalities.

(Fstoffel, Tiranti, and Huggel 2014). Permafrost thawing and increasing flow velocities of rock glaciers have been demonstrated to deliver more sediment into debris-flow channels under current conditions than in the past. The volume of the largest debris flows has, therefore, risen by one order of magnitude since the 1920s and is likely to increase further with ongoing climate change.

Whereas present-day debris flows typically deposit material on cones rather than eroding significant amounts of sediment, recent debris-flow events have been observed in the Swiss Alps that have developed sufficient erosive power to remobilize large amounts of sediment on Holocene fans. A model case is the August 2005 Rotlaui debris flow in Guttannen, which was the largest event in Switzerland for the past 20 years with a total debris-flow volume of 500 000 m³, about 300 000 m³ of which was entrained on the cone (Figure 4). The Rotlaui debris flow was triggered from extensive sediment uncovered by the recession of the local glacier in the past decades. As observed with previous debris flows, glacier retreat since the Little Ice Age and associated uncovering of large amounts of sediment was an important factor for particularly large debris flows in the Alps.

Future mass movements: implications for hazards and need for further research

The effects of changing mean and extreme temperature and precipitation are likely to be widespread and to influence both the occurrence (in terms of temporal frequency) and the magnitude of future mass movements in mountain regions around the globe. Despite uncertainties, slopes currently underlain by degrading permafrost will probably become less stable at progressively higher altitudes with ongoing climate change (Harris et al. 2009). It can also be speculated that the probability of rock instability and the incidence of large (>10⁶ m³) rockfalls will increase in a warming
Figure 4 View of the catchment of the 2005 Rotlaui debris flow at Guttannen, central Swiss Alps. A: zone of initiation consisting of an import body of partly frozen glacial sediment. B1: initial zone of debris-flow erosion in glacial sediment. B2: debris-flow transport zone with repeated channel erosion. B3: deep erosion on the Holocene fan, amounting to about 300,000 m³ additional sediment. B4: deposition of debris-flow material on the lower part of the fan, with run-ups on the opposite valley flank. C: destruction of the Grimsel highway and obstruction of the Aare River by debris-flow material (Base image from Google Earth, October 2009).

Glacier downwasting will result in the formation of further ice-marginal lakes and subsequent problems of glacier lake outburst floods (Worni et al. 2014). On steep slopes, warming firn and ice temperatures may result in new sites of ice falls and ice avalanches (Harris et al. 2009). Provided that sediment supply is not a limiting factor, debris flows have the potential to become larger in the future than they were in the past, but not necessarily more frequent and clearly conditioned by local site conditions (Stoffel, Tiranti, and Huggel 2014). The generation of cascading processes at high elevations might increase and result in chain reactions that are often difficult to predict (e.g., the impact of ice-rock avalanches into glacier lakes in the Cordillera Blanca, Peru, and other regions, with subsequent downstream flooding) (Worni et al. 2014).

Recent developments at high-elevation sites have shown clearly that the sensitivity of mountain and hillslope systems to climate change is likely to be acute, and that events beyond historical experience will continue to occur as climate change continues. Despite the lack of preserved analogs, and although paleoevidence for the link between mass movements and climate may be fraught with many questions of interpretation, the inclusion of paleorecords has the capacity to integrate signals over a wide area and to employ a broad range of supporting proxy information. In this perspective, observable evidence from the far and near past points to an increase in mass-movement activity and a major mobilization of sediment under warming and/or wetter conditions (Stoffel, Tiranti, and Huggel 2014).

Recent mass-movement processes originating from high-elevation sites have been observed at sites with no or little historical precedence (e.g., Rotlaui debris flow, Switzerland, Figure 4) and have threatened or destroyed critical transport and energy infrastructure. The risk of damaging events is often particularly high in developing countries, where both population and agriculture pressure on land resources lead to the exploitation of unstable slopes.

An improved understanding of the intrinsic complexity of mass-movement processes at
high-elevation sites, cascades of processes and their relationships with, and dependency on, climate variability and change is crucially important in planning appropriate and prospective measures to reduce the negative impacts of future events. This entry has attempted to shed light on recent developments and state-of-the-art knowledge on the possible effects of climate change on the occurrence and intensity of potential future mass movements. The contribution has also outlined the broad spectrum of recent research and the large suite of case study results available in the research field, but has also illustrated clearly that, despite all the progress made in climate–mass movement studies, many questions still need to be addressed in greater detail to overcome considerable gaps in knowledge that continue to exist in this challenging field of research.

SEE ALSO: Climate change and permafrost; Geomorphic hazards; Glacier lake outburst floods; Mass movement processes and landforms; Natural hazards and disasters; Periglacial processes and landforms; Permafrost: definition and extent; Quaternary geomorphology and landscapes

References


Measurement theory

Nicholas R. Chrisman
RMIT University, Australia

Many geographic questions expect a measurement as the answer. This result can be a simple figure, such as the height of some peak, the distance flown by an aircraft, or the area of given parcel. In other cases, the measurement is the result of more complex analysis, such as a city’s rank in a contest for most liveable city, estimated carbon sequestration over the forests of a region, or the total area of tea plantations in a given district. In each of these cases, the measurement element appears unambiguous, but there are many choices involved, and theoretical issues to address. Often what appear to be simple questions should have been clarified with some details about the process and result of measurement required.

Much of the trouble comes from tacit expectations tied up in a long history of making measurement seem obvious and unambiguous. The “classical” school of measurement has its origins in Aristotle’s metaphysics and continues through to the physical sciences at the end of the nineteenth century. In the classical view, measurement discovers the numerical ratio between a standard object and the one measured. Clifford, a mathematician in the nineteenth century, used a definition of measurement very little altered from that of Euclid: “every quantity is measured by the ratio which it bears to some fixed quantity, called the unit” (Clifford 1870). The property was seen as inherent in the object and the relationship took on all the mathematical properties of numbers without question (Michell 1993). Take the attribute “length.” Every entity in space can be measured by comparing its length to some other length. If we adopt a “standard” measuring rod (originally derived from the foot of the king or something equally available), the ratio between the length of the rod and the objects measured embodies a number implicit in the relationship. Reciprocally, the realm of numbers is seen as deriving from these basic relationships of ratios.

The classical approach can be discerned in many geographic applications, particularly when attributes are considered to be inherent to an object, and the measurement of attribute values is considered to be unproblematic. Measurement of agricultural productivity, volume of grain per unit area, ties back to centuries of practical experience. The current attribute reference systems are only slightly evolved from the city standard carved into some stone in the marketplace. Spatial reference systems, particularly local vertical datums, also depend on these kinds of measurement scales. Pick some starting point as zero, measure with a standard unit using traditional leveling. It is a technology in current use by municipalities around the world. While it looks simple, there is more to consider.

Through a series of theoretical developments in mathematics, this classical paradigm was overthrown by a “representational” school in the first half of the twentieth century (Michell 1993). The representationalists impose a firm separation between the empirical world and numbers, first expressed by Bertrand Russell (Russell 1897, 1903). In the place of ratios and rational numbers as the core of mathematics, set theory permitted a more generic range of relationships:
Measurement of magnitudes is, in its most general sense, any method by which a unique and reciprocal correspondence is established between all or some of the magnitudes of a kind and all or some of the numbers, integral, rational, or real, as the case may be… In this general sense, measurement demands some one-one relation between the numbers and magnitudes in question – a relation which may be direct or indirect, important or trivial, according to circumstances.

(Russell 1903, 176)

Representationalism is unified by the understanding that measurement assigns numbers to objects according to external rules. These rules are developed by humans, not inherent in nature. Thus, the measurement result depends on conventions and the observer, rather than being intrinsic to the object measured. Within this sweeping change in the theory of measurement, the original branch of representationalists, called “operationists” (and including a preponderance of physicists), focused on the influence of the measurement operation (Campbell 1920; Bridgeman 1927). They noted that even the standard meter rod in Sèvres varied in length due to changes in temperature and pressure. Thus, the number obtained depended on the whole circumstance of the operation, not just immutable and inherent properties. In this respect, they expanded the approach to measurement, but their opening did not include everyone. These operationists, having their origins in physics, espoused a restrictive theory centered on the role of addition in building the basic “extensive” building blocks and physical laws that defined “intensive” (indirect or derived) measurements mostly as ratios of the extensive measures. Extensive measurement applies to the seven base units of the Système Internationale d’Unités (SI); the operationists expected virtually everything else to be defined as ratios of these units. Exclusive focus on extensive measurement left almost no room for the social sciences to develop a measurement theory. The physicists could not consider phenomena like perceived loudness of sounds as a measurement, since it did not involve extensive properties like addition. They could allow the observer into their scheme, but not in such a perceptual framework.

Before tracing the subsequent theoretical developments, it is important to note that operationism, though not directly recognized by citations to the classics in physics, is alive and well in the geographic information sciences. Geodesy is acutely aware of the various influences on measurement and works to reduce error through more and more elaborate mathematics. Remote sensing adopts a more statistical viewpoint but still tries to limit the influences of the measurement operation.

Returning to the historical narrative, the issue of human perception (of sound or anything else) brought forth the next developments. Stanley S. Stevens, a psychologist at Harvard University, wrote a paper in Science in 1946 on the subject of measurement (Stevens 1946). This paper played a formative role in a larger movement towards a quantitative methodology in the social sciences. It responded to an impasse that characterized the work of a committee established by the British Association for the Advancement of Science. This committee had been chartered in 1932 to consider the possibility of “quantitative estimates of sensory events.” In eight years of deliberations, the unity of sciences was far from apparent, and the physicists could find no way to accommodate the psychophysical efforts into their scheme for measurement (Ferguson et al. 1940; Luce et al. 1990, 109–111). Stevens’ proposed solution recognized that not all measurements lived up to the expectations of the physical scientists, particularly that addition might not be the primary axiom at the base of all measurements. He based his scheme
on the group of transformations under which a measurement would retain its meaning (an automorphism group). Stevens described four “scales” of measurement (Table 1) a scheme now familiar to almost any social scientist. Stevens’ use of the word “scale” fails to make a distinction between a particular number axis (such as length measured in meters) and the generic grouping in which it falls. In geography and cartography the word “level” is more commonly applied for the groupings of scales (Chrisman 1998b).

Stevens remained inside the representationalist philosophy but articulated a “nominalist” approach (Michell 1993), defining measurement as the “assignment of numbers to objects according to rules” (Stevens 1946, 677). Stevens’ reliance on automorphism derived directly from the mathematical approach of Russell and Whitehead. Stevens prescribed a simple sieve of tests to decide in which group a measurement belongs. First he asserted: “We know that the values of all scales can be multiplied by a constant, which changes the size of the unit” (Stevens 1946, 680). This sweeping statement aligns his ratio level with the classical viewpoint, and creates some difficulties, which are considered below. His purpose is to introduce less rigorous levels by relaxing the restrictions one by one. If a constant can be added, while retaining the meaning of the measurement, it is “proof positive” (Stevens 1946, 680) that it is no longer a ratio scale. If it can be transformed in nonlinear manner, it is not even interval. If it can be transformed “at will” (meaning non-monotonically), “the nominal scale is the sole remaining possibility” (Stevens 1946, 680). Although these levels are now usually presented in order of increasing rigor, they were presented from the familiar extensive ratio towards the less restrictive in the original.

Stevens also tried to expunge the distinction that physicists had drawn between “fundamental” (extensive) and “derived” measurement. In Stevens’ automorphic viewpoint, the number system was more central, not the method by which it was generated.

A key element of Table 1 is the connection between the levels and “permissible statistics.” Many textbooks on statistics for social sciences (beginning with Siegel’s (1956) classic on

<table>
<thead>
<tr>
<th>Scale</th>
<th>Basic empirical operations</th>
<th>Mathematical group structure</th>
<th>Permissible statistics (invariantive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Determination of equality</td>
<td>Permutation group $x' = f(x)$</td>
<td>Number of cases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f(x)$ means any one-to-one substitution</td>
<td>Mode</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Contingency correlation</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Determination of greater or less</td>
<td>Isotonic group $x' = f(x)$</td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f(x)$ means any monotonic increasing function</td>
<td>Percentiles</td>
</tr>
<tr>
<td>Interval</td>
<td>Determination of equality of intervals or differences</td>
<td>General linear group $x' = ax + b$</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Standard deviation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rank-order correlation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Product-moment correlation</td>
</tr>
<tr>
<td>Ratio</td>
<td>Determination of equality of ratios</td>
<td>Similarity group $x' = ax$</td>
<td>Coefficient of variation</td>
</tr>
</tbody>
</table>
nonparametric methods) adopted this connection between a variable and appropriate techniques. The four levels of measurement spread rapidly from psychology to sociology (Blalock 1960) and other social sciences. The Stevens framework provided a simple means to organize the different kinds of data available and to structure textbooks and courses for decades to come.

In the earliest explorations of quantitative methods in geography, Siegel's book was the one that sparked the interest in levels of measurement (Morrist, personal communication). Siegel framed a set of nonparametric tests by level of measurement—a direct implementation of Stevens' concepts. The levels of measurement then become a cornerstone for the presentation of quantitative methods in geography: Haggett (1965, 212) references Siegel, not Stevens; Cole and King (1968, 54–55) presented the four levels without any reference whatsoever. In some curious way, the four levels became a kind of charter for the quantitative movement in geography, dispensing it from any further need to examine the literature on theories of measurement.

While in geography the diffusion was slow and noncontroversial, the levels of measurement generated some sharp controversy in psychology and sociology that continues to this day. This section cannot investigate all of the developments, but it will demonstrate that the four levels have been effectively challenged and that the linkage between level and procedures has also been questioned. These debates and extensions will be related to the particular problems of geographic information in the following section.

Redefining the levels

The four levels have been presented as a complete set but there are other possible levels not handled by this system. Stevens (1959) himself proposed another level at the same rank as interval for logarithmically scaled measures. Unlike the linear interval whose zero can be relocated at will, these measures can be rescaled by changing the exponent (the zero remains fixed). This “logarithmic-interval” scale is not cited in any of the geographic literature, though it is used for earthquake intensities (such as the Richter and Mercali scales) and similar measurements. Some measurement scales such as densities, which are dependent on more than one ratio-scaled measurement, could be rescaled by an exponent as easily as by the choice of an arbitrary unit; thus, they are log-interval not ratio (Krantz et al. 1971, 487). The use of persons per area (squared distance units) is just a convention; population density could just as easily be measured by average distance to the next person (in distance units), though these two scales are not related by a simple linear transformation.

Following Stevens' invariance logic to its conclusion, ratio is not the highest level of measurement. A ratio scale has one fixed point (zero) while the choice of the value of “one” (the unit of measurement) is essentially arbitrary. A higher level of measurement can be obtained if the value of one is fixed as well. Then the whole scale is predetermined or “absolute” (Ellis 1966; Narens 1985) and no transformations can be made that preserve the meaning of the measurement. One example of an absolute scale is probability, where the axioms fix the meaning of zero and one simultaneously. Bayes' law of conditional probability works because the scale is fixed between zero and one. Probability is just one example of a measurement incorrectly classified by a four-level scheme. The measurement literature provided workable definitions of log-interval scales and absolute scales thirty or forty years ago, but these developments did not penetrate into cartography or geography.
While Stevens’ levels deal with an unbounded number line, there are many measures that are bounded within a range and repeat. For instance, angles seem to fit into the ratio level, in the sense that there is a zero and an arbitrary unit (degrees, grads or radians). However, angles have a singularity at $360^\circ$. The direction $359^\circ$ is as far from $0^\circ$ as $1^\circ$ is. Any general measurement scheme needs to recognize the existence of repeating or cyclical elements.

Even at the bottom of the hierarchy, the nominal scale deserves re-examination. It is all too easy to criticize geographical analysis for using a model of sharp sets (Curry 1995). The standard rendition of the nominal level emphasizes pure “equivalence” of all members in a nominal category. However, in practice, categories can be much more complicated. Members of a category often have some degree of membership. Either they represent “prototypes” (pure types) or less clear cases, or there is some probability of membership, or some distance metric. All these forms of graded membership produce “categories” that cannot be treated as if all members were strictly equal.

To summarize, the list of four levels presented by Stevens is just not adequate even inside its own automorphic approach. Table 2 presents an expanded list of levels of measurement with the information required to specify a scale. The ten levels cited are by no means complete, in part because geographic information is not just a matter of simple scales of attributes taken without consideration of the spatial component.

### Multidimensional measurement

Measurement scales become much more difficult to define in a multidimensional situation. In the one-dimensional world of Stevens, an open-ended ratio scale seems to provide the most information content; the real number line contains the most promise for mathematical relationships. When representing a two-dimensional space, the classic Cartesian solution uses two orthogonal number lines. However, a radial coordinate system contains exactly the same information with one distance (open-ended) and one angle (closed range) (Figure 1). These two representations are equivalent, even though the units of measurement do not seem equivalent.

<table>
<thead>
<tr>
<th>Level of measurement</th>
<th>Information required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Definitions of categories</td>
</tr>
<tr>
<td>Graded membership</td>
<td>Definition of categories plus degree of membership or distance from prototype</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Definitions of categories plus ordering</td>
</tr>
<tr>
<td>Interval</td>
<td>Unit of measure plus zero point</td>
</tr>
<tr>
<td>Log-interval</td>
<td>Exponent to define intervals</td>
</tr>
<tr>
<td>Extensive ratio</td>
<td>Unit of measure (additive rule applies)</td>
</tr>
<tr>
<td>Cyclic ratio</td>
<td>Unit of measure plus length of cycle</td>
</tr>
<tr>
<td>Derived ratio</td>
<td>Unit of measures (formula of combination)</td>
</tr>
<tr>
<td>Counts</td>
<td>Definition of objects counted</td>
</tr>
<tr>
<td>Absolute</td>
<td>Type (probability, proportion, etc.)</td>
</tr>
</tbody>
</table>
The conversion from two ratio scales to one ratio scale plus an angle is not unique to geometric constructions. For example, radiant energy at various wavelengths coalesce into a particular color that can be represented in a cone-like object (Munsell's space or equivalent). The color cone is thus a “fact of nature, not a mathematical trick” (Suppes et al. 1989, 226). Multidimensional measurements create interactions not imagined in the simple linear world of Stevens.

Similarly, in GIS processing, the single phenomenon of “slope” is broken into two numbers for gradient and aspect (Burrough 1986). There are multiple methods in use (percent of rise/run versus angular measures) that are related through nonlinear trigonometric functions. These examples of “conjoint measurement” do not fit the linear scheme of Stevens. There is a substantial literature in psychology and sociology developing these kinds of measures (Krantz et al. 1971; Suppes et al. 1989; Luce et al. 1990). While this literature recognizes the contributions of Stevens, after decades of careful axiomatic development, everything does not reduce to fit into such simple levels. The system of automorphic groups may not be adequate to describe important characteristics of measurements.

**Linkage to procedures**

Almost as soon as Stevens’ original paper appeared, and before it became codified in statistical textbooks, there was a flurry of criticism in the sociology and psychology literature. The debate centered on the assertions about permissible procedures. For example, Lord (1953) wrote a mordant satire built upon a fictional analysis of the numbers on football jerseys. Savage, an established statistician, wrote a critical review of Siegel’s book, challenging the linkage between measurement scale and the choice of statistical test (Savage 1957). Such a reaction might be expected if the radical new paradigm was temporarily resisted by an “old guard,” but the controversy has continued without losing its heat. A self-styled “anti-measurement” group has continued the attack in a drawn-out series of articles (Gaito 1960, 1980, 1986; Borgatta and Bohrnstedt 1980), while others defend with equal intensity (Townsend and Ashby 1984; Maxwell and Delaney 1985). A more complete review (Khurshid and Sahai 1993) reveals more complexity than a simple dichotomy. The attackers correctly protest that some “permissible” procedures, such as Spearman’s rank correlation, involve addition and multiplication.
of ordinal scores, and use formulae that are equivalent to the nonpermissible procedures. Yet, the formalists (Luce et al. 1990) produce elaborate proofs of the central contention of Stevens that there are classes of invariance with consequences for validity of procedures. More recent work (Kampen and Swyngedouw 2000) has proposed five distinct forms of ordinal measures making the choice of method less direct.

Perhaps it does not particularly matter to geographers exactly how the sociologists and psychologists disagree. One lesson should be clear: the measurement level by itself does not guarantee a meaningful result following a particular procedure (Chrisman 1998b). For example, we can add the populations of counties within a state to derive the total state population, but we should not add city populations with county populations without understanding exactly how these objects nest within each other in that particular jurisdiction (in Virginia, the cities are county equivalents, in other states they are usually contained within the county figure, with a few exceptions). However, a figure such as median family income, though it falls in the ratio level, should not be added to another median family income. The result is not a median family income. Cartographers must also be aware of the date of the census material as well. Though the figures are in the same measurement scale, the addition of 1890 and 1990 populations does not produce a figure with an identifiable date. In other words, even for a simple operation like addition we need to know more than the scale of the population figures: we need to know about the spatial component and the temporal component.

Home-grown measurement theory for geography

To develop beyond the Stevens taxonomy, geographers need to re-examine some assumptions that have become a part of the discipline’s culture. The largest difficulties with Stevens’ scheme come not from the specific levels of measurement but with the overall model of the process. Stevens presumes a rather simple situation: the object to be measured is unambiguous. Most social sciences in the early quantitative period simply assigned numbers to “cases.” In some fields it might seem obvious what is a case, but there can be substantial questions even for those who study individual persons whether to measure households, neighborhoods, or other aggregates (Ragin and Becker 1992). The version proposed in geography was called the “Geographical Matrix” (Berry 1964), simply a matrix with “places” on one axis and attributes on the other. Places are problematic; geographic measurement requires many decisions involving time, space, and attribute.

One way to begin is to re-examine one kind of common cartographic data: counts aggregated over some region in space. Counts are discrete, since there is no half person to count, but a count captures more mathematical structure than the other discrete levels (nominal and ordinal). Since the zero is a fixed value, counts may seem ratios but, being tied to the discrete unit counted, it cannot be rescaled by some arbitrary factor. Counts have different properties from the absolute scale, as well. Counts do not have an upper bound in the same way as a probability or a proportion. Ellis (1966, 157) points out the difference between ratio scales and counting with the example that it is acceptable to present a unit of measurement by saying “Let this object be 1 minch long,” but it is not possible to say “Let this group contain one apple,” since it either has one apple or some other number when you start. The process of counting depends upon the recognition of objects, a procedure taken for granted in the Stevens tradition, but at the heart of the geographic data problem. A count only makes sense when you specify which spatial object was used to contain the counting process.
MEASUREMENT THEORY

The assumptions behind the assertion that a city “has” a population requires numerous assumptions. A population figure represents the estimate obtained by a particular agency applying a specific spatial definition of the city and a specific definition of people’s movements at a particular time. The agency selects its definitions for cities and for residents in response to a complex process of legal mandates, professional expertise and resource limitations. There are many possible population measurements that could be obtained, involving slightly different treatment of space, time, and attributes. The meaning of this number must remain contingent on understanding the whole set of circumstances that permit the measurement. A model of geographic measurement must handle much more than the number scales treated by Stevens.

A more fruitful model sees measurement not in terms of properties but in terms of relationships. Geographic information involves many distinct kinds of measurement and, as the example of census counts demonstrated, each kind invokes some combination of time, space and attribute rules. Wright (1955) attempted to enumerate all the possibilities – a huge task. Viewed from the perspective of measurement, these old issues take on a new clarity.

The simple taxonomy presented by Sinton (1978) provides a starting point to understand the relationships involved in measurement. In Sinton’s scheme, in order to measure one component, one of the others had to be fixed and one served as control. In this context, “control” does not refer to a geodetic reference system, or an experimental design. Control refers to a mechanism of restraint on the variation of a system to permit measurement of one component of a phenomenon while other components only vary within the limits of the control. In essence, the control is the choice of which components define the “objects” so that the last component is free to be measured. At the most basic, Sinton’s scheme distinguishes vector from raster because the first controls by attribute, while the later controls by space. This rough division provides a starting point but it does not explain the variations within these two approaches. A system of “measurement frameworks” articulates a home-grown theory of measurement for geographic information (Chrisman 1998a, 2001).

Attribute as control

When the attribute serves as the control, the spatial location can be measured. While this is the common basis for a vector representation, there are large differences between an isolated category and a connected system of categories. Table 3 summarizes some distinctions.

### Table 3  Attribute control frameworks.

<table>
<thead>
<tr>
<th>Isolated objects</th>
<th>Connected objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial object</td>
<td>Network</td>
</tr>
<tr>
<td>Single category distinguishes from void</td>
<td>Spatial objects connect to each other, form topology (one category possible)</td>
</tr>
<tr>
<td>Isoline</td>
<td>Categorical coverage</td>
</tr>
<tr>
<td>Regular slices of continuous variable</td>
<td>Network formed by exhaustive classification (multiple categories, forming an exhaustive set)</td>
</tr>
</tbody>
</table>
The simplest object control framework involves isolated objects, distinguished by a single category. Each distinct instance of the attribute identifies an “object,” an entity distinguished by its particular value of the attribute. Each object is described as a geometric whole, since it will forcibly occur in isolation. In the pure form of this framework, the only relationship is between the object and a position; there are no relationships between objects. These rules apply most directly to points and areas. Linear objects rarely remain isolated but connect with varying topologies, creating the need for the network framework (not isolated).

Isolines are formed by controlling for a specific value on a surface. Since isolines follow the contours and do not intersect, they have no topological relationships beyond the ordering of nested contours. Thus, the isoline is isolated from other isolines, requiring no structure for contiguity.

In the creation of GIS software, it was important to recognize that there were relationships between the objects in a database. When a coverage is formed with multiple categories, there will be topological relationships. Similar structure can be created by linear networks. The basic topology is required whether the categories form strict equivalence classes or some form of probabilistic or prototype categories. The distinction between the isolated objects and connected coverages is not a matter of database design but a recognition of the underlying measurement structure of the source material. Connected objects should not be treated as isolated objects. The decision to recognize a certain category as a part of an exhaustive coverage depends on the nature of the surroundings, not just on the category itself. For example, a given land use class is not simply defined by its own internal characteristics but also relative to its neighbors. This distinction creates some of the tension between an object-oriented technology that treats isolated entities and the topological approach that forces the world into exhaustive coverages. Though lumped into vector, these divisions are at least as deep as the divisions between raster and vector.

In terms of measurement, the attributes of all the attribute-controlled variants must be discrete. The spatial component can vary as needed to reflect the location of the edges of the discrete objects. One can dispute this approach on a variety of grounds, but the choices are bundled together. High resolution freedom in the spatial component comes from a choice to simplify the attributes into categories.

**Table 4** Spatial control frameworks.

<table>
<thead>
<tr>
<th>Point-based control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center point</td>
</tr>
<tr>
<td>Systematic sampling in regular grid</td>
</tr>
<tr>
<td>Systematic unaligned</td>
</tr>
<tr>
<td>Random point chosen within cell</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Area-based control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extreme value</td>
</tr>
<tr>
<td>Maximum (or minimum) of values in cell</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>Sum of quantities (e.g., reflected light) in cell</td>
</tr>
<tr>
<td>Predominant type</td>
</tr>
<tr>
<td>Most common category in cell</td>
</tr>
<tr>
<td>Presence/absence</td>
</tr>
<tr>
<td>Binary result for single category</td>
</tr>
<tr>
<td>% cover</td>
</tr>
<tr>
<td>Amount of cell covered by single category</td>
</tr>
<tr>
<td>Precedence of types</td>
</tr>
<tr>
<td>Highest ranking category present in cell</td>
</tr>
</tbody>
</table>
Spatial control

Rather than limiting the variation in the attributes to a certain set of objects, control can also be imposed by specifying a set of locations, often in the form of a regular spatial tessellation (Table 4).

Control by a set of points has different rules compared to control by areas. While both would be encoded in a raster representation, they must be understood differently. With a point-based control there are not too many rules. Center point provides a regular sampling of a landscape. Digital elevation matrices tend to expound a point-based sample in the documentation, though the actual operational procedures invoke models to arrive at this figure. Systematic unaligned is recognized in textbooks but rarely performed.

Control by area is more common for remote sensing and other applications of grid sampling. In each cell there is some rule that has been applied to all the possible values. Some sensors add up all the reflectance in a certain band width, others take the highest or lowest value (such as the minimum clearance elevation shown on aeronautical charts). A system that optimizes the choice within each cell by taking the most likely value may remove all traces of linear features and the minority elements. Unless these rules are known to the analyst, the information can be sorely misconstrued. GIS operations depend more on knowing how to combine these measurements than they do on the level of measurement.

Extending Sinton’s taxonomy

Control by object and control by space seem to be the only options, but they do not accommodate comfortably all the cases found in existing geographic information. The well-known choropleth map is an example of a composite framework, a sequential combination of the two forms of control. The base map is created using a categorical coverage for the set of collection units, then these objects serve as a secondary form of spatial control to tabulate the variable in question. Due to these two stages, the spatial measurements of the boundaries have little bearing on the precision of the eventual attribute measurement.

In the development of thematic mapping, the choropleth map dominated. The census viewpoint of tables of attributes for the same set of places was the origin of quantitative geography, expressed in Berry’s matrix. A whole set of techniques developed around standardizing these attributes to remove the unevenness of the varying sized spatial control. The “modifiable areal unit problem” (Openshaw 1984; Arbia 1989) and the early development of spatial autocorrelation (Cliff and Ord 1973) both addressed the peculiar situation of spatial control by uneven objects. Rather than generalizing from this special case to the rest of cartographic measurement, it makes more sense to see choropleth frameworks as composites of more primitive operations.

Triangular irregular networks (TIN) pose an additional problem. The TIN is composed of discrete triangles, but these do not arise from imposing a cookie cutter kind of control onto a surface. While the points may come from an isolated bunch of measurements (or one of the other cartographic choices for terrain representation), the TIN represents a set of relationships that cover space. The ideal TIN is constructed so that the triangles represent zones of uniform gradient and aspect, within the resolution available. Edges of triangles should conform to the topology of the surface, including ridges, courses and other breaks in slope (Peucker and Chrisman 1975; Mark 1979). These features of a surface are not based on an isolated value but on
the relationships of convergence and divergence, relationships of neighboring values. Thus, a TIN represents a novel class of measurement frameworks where relationships form the control, not the values of the attribute or the location alone.

A further example of a relationship control system involves attributes that attach to pairs of geographic entities. For example, international trade can be analyzed as flows of goods and money between pairs of countries. This kind of geographic information does not attach to any geometric object; it attaches to a pairing of objects – a relationship. This relationship creates the control for the measurement. This particular kind of geographic information is quite complex to portray using established thematic techniques. Edward Ullman (1954) built his economic geography around these flows and it became a long-standing data structure for regional analysis (input–output) and interregional trade (treated as its own specialization).

This initial collection of measurement frameworks provides a more durable foundation for processing geographic data. Compared to the levels of measurement that simply provided a hierarchy of linear relationships, this taxonomy mobilizes the interplay between geometry and attribute. It can also accommodate measurements in the temporal domain. As Langran (1992) has carefully documented, Sinton’s scheme of holding one element fixed, while it might be required for a two-dimensional display, is not a requirement in a database solution. Still, it seems a consistent constraint that only one component can be measured with full flexibility and resolution.

This system of measurement frameworks is grounded in the understanding that you cannot attain ultimate resolution in time, space, and attribute simultaneously. Measurement requires a choice of which component is more critical. From this scheme a taxonomy of geographic transformations (sketched in Chrisman 1998a) can be developed around the role of geographic neighborhoods and the rules of attribute combinations.

**SEE ALSO:** Cartographic modeling; Critical GIScience; Critical spatial thinking; Fiat and de facto objects; Fractal analysis; Fuzzy classification and reasoning; Geodesy; Geographic information science; Modifiable areal unit problem; Qualitative data; Qualitative GIS; Qualitative information: representation; Scale; Spatial concepts; Uncertainty

**References**


Measuring spatial dependence

Yongwan Chun
Daniel A. Griffith
University of Texas at Dallas, USA

Spatial dependence is one of the most prominent phenomena in the spatial sciences. It represents the spatial relationships of values in attributes that are measured at geographic locations. It refers to whether or not values in a single variable are correlated with other values in that same variable when they are inspected along with their spatial locations. This correlation is often positive in nature, so that observations at spatially close locations tend to have similar values. This correlation in georeferenced space is referred to as spatial autocorrelation because it involves only one variable, which distinguishes it from the conventional correlation between two variables. The presence of spatial dependence violates the independence assumption on which conventional statistical theories heavily rely. Hence, inferential statistics can become unreliable when spatial dependence is not appropriately handled. Measuring spatial dependence is an essential process to diagnose if conventional statistical approaches would produce proper results for a given dataset, and/or if spatial statistical methods that incorporate spatial dependence need to be employed.

Measuring spatial dependence requires several essential components. First, it needs an operational framework with which spatial locations are transformed to represent spatial proximity or spatial closeness among observations. This measurement involves spatial neighbors that covary with observations. A geographical topology structure, as well as geographical distances, frequently furnishes a basis for articulating this geographic neighborhood. Second, indices are used to numerically measure spatial dependence, and subsequently to perform an inferential statistical test. Numerical indices have been developed with extensive statistical concepts and theories, and can be considered as the core of measuring spatial dependence. Most of these indices are intended for quantitative attribute values (i.e., an interval or ratio measurement scale), but others are used for qualitative values. Also, indices are available to measure global spatial autocorrelation across an entire study area and local spatial autocorrelation for local subareas. Third, graphical tools are available to visually portray spatial dependence.

Spatial neighborhoods

Measuring spatial dependence requires identifying a set of values that has a spatial relationship among the values. The values in such a set are considered to be spatial neighbors. Spatial neighbors for a single observation within a set of \( n \) observations can be evaluated in terms of the other \((n - 1)\) observations. Identification utilizes spatial proximity or spatial closeness, with an assumption that spatially close observations are likely to be more strongly related to each other than to distant observations, an expectation often referred to as Tobler's first law of geography. Although distance is a natural way to represent spatial proximity, general geographical topology (i.e., spatial connectivity or adjacency) commonly is used, especially for areal features.
Figure 1 Representations of spatial neighbors based on the adjacency of 2010 census tracts in Dallas County, TX.

Figure 1a illustrates spatial neighbors for the 529 census tracts in Dallas County, TX, from the 2010 US Census, based on spatial connectivity. The dotted red lines connecting the centroids of these census tracts identify spatial neighbors that share a common boundary. The dotted red lines appear only for a small number (namely, 1380) of pairs among the \( \frac{n(n-1)}{2} \) (i.e., \( 529 \times 528/2 = 139656 \)) possible pairs. This neighborhood articulation can be represented numerically with \( n \) binary 0–1 indicator variables, where 1 denotes spatial neighboring pairs and 0 denotes spatial non-neighboring pairs. These numerical values are commonly organized in an \( n \times n \) matrix; here \( n = 529 \). Figure 1b shows the numerically coded spatial neighboring structure among the census tracts in a matrix, \( C \). Each \( c_{ij} \) element of matrix \( C \) denotes whether census tract \( j \) is \( (c_{ij} = 1) \) or is not \( (c_{ij} = 0) \) a spatial neighbor of census tract \( i \). Matrix \( C \) is called a spatial neighbor matrix and, for this specific case, a spatial connectivity matrix because it represents the physical contiguity among the areal units. Note that the diagonal elements of this connectivity matrix are zero. This means that areal units are not considered to have themselves as spatial neighbors.

Spatial neighborhoods can be configured in different ways. Spatial connectivity can be defined by two popular options: rook and queen. For rook connectivity, spatial neighbors are defined with areal units sharing nonzero length boundaries. That is, areal units connected only by a point (e.g., a zero length boundary) are not considered to be spatial neighbors. For queen connectivity, areal units connected by a point as well as a nonzero length boundary are considered to be spatial neighbors. These two options are similar to the movement options of rook and queen chess pieces. These two options produce a
substantially different spatial connectivity matrix for the regular square tessellations, such as a chess board or a remotely sensed image, but not necessarily for irregular tessellations, such as census tract boundaries where polygons usually do not meet only at a single point. Another set of approaches uses geographic distances to define spatial neighbors. When two areal units are within a preset distance threshold, they can be considered to be spatial neighbors; that is, \( c_{ij} = 1 \) when \( d_{ij} \leq d \), and \( c_{ij} = 0 \) otherwise, where \( d_{ij} \) denotes distance between areas \( i \) and \( j \), and \( d \) denotes a preset threshold distance. Types of transformation other than this binary transformation are possible. Inverse distance, especially, is widely used to represent spatial proximity. A general functional form of inverse distance is \( d_{ij}^{-\gamma} \), where \( \gamma \) is often set to 1 (i.e., \( c_{ij} = 1/d_{ij} \)). As \( \gamma \) increases, this specification approaches the structural form of the binary matrix. One characteristic of this distance-based matrix is that its off-diagonal elements have nonzero values and only diagonal elements have zeroes, while a spatial connectivity matrix contains a large number of zero elements.

Spatial neighborhoods are introduced as weights in indices to measure spatial autocorrelation. A matrix of these weights is called a spatial weights matrix. A binary spatial connectivity matrix, \( C \), can function as a spatial weights matrix. That is, each weight is 0 or 1 (Figure 1b). Furthermore, this binary spatial connectivity matrix can be transformed to a row-standardized matrix, \( W \). For example, if the \( i \)-th areal unit has two spatial neighbors, the two nonzero cell values (i.e., 1) in the \( i \)-th row of matrix \( C \) are replaced with \( 1/2 \) (i.e., 1 divided by the number of spatial neighbors). Hence, the sum of values in each row of matrix \( W \) is 1, and the total sum of its weights is \( n \). A distance-based matrix is also commonly used to construct matrix \( W \). Using matrix \( C \) results in comparing an attribute value at a given location with the sum of its neighboring attribute values. Using matrix \( W \) results in a comparison with the average of its neighboring attribute values.

### Indices measuring spatial dependence

Spatial autocorrelation measures can be categorized as global and local. A global measure furnishes an index of spatial autocorrelation across an entire study area, usually with a single value. It further involves a statistical test to confirm whether or not observed spatial autocorrelation is statistically significant. The most widely used spatial autocorrelation indices include Moran’s \( I \) and Geary’s \( c \). In contrast, a local measure investigates spatial autocorrelation associated with a specific areal unit. Measures and statistical tests are conducted for each areal unit. Their collective pattern of statistical tests can be used to examine local variations of spatial autocorrelation in a study area. The most well-known local measures include the local Moran’s \( I \), and the Getis–Ord \( G_i \) and \( G^*_i \); the local Geary’s \( c \) is also available.

These global and local indices are devised for quantitative values (i.e., interval and ratio measurement scale data). A join count statistic can be used to measure global spatial autocorrelation for qualitative values (i.e., specifically nominal measurement scale data).

### Global spatial autocorrelation measures

As a covariation measure, Moran’s \( I \) is directly related to the Pearson product moment correlation coefficient and has a similar functional form. A description of this functional form similarity can be found in spatial statistics textbooks (Chun and Griffith 2013, 10). Given
a variable \( Y \), the Moran’s \( I \) is calculated using equation 1:

\[
I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

(1)

where \( y_i \) denotes an observed value at areal unit \( i \), \( \bar{y} \) denotes the mean of variable \( Y \), and \( w_{ij} \) denotes a spatial weight for the pair of areal units \( i \) and \( j \). When binary connectivity is used as spatial weight, \( w_{ij} \) can be replaced with \( c_{ij} \) from matrix \( C \). Equation 1 returns a single number that represents the nature and degree of spatial autocorrelation.

As an analog to the Pearson product moment correlation coefficient, Moran’s \( I \) can be interpreted similarly while recognizing that it has different properties. First, the range of Moran’s \( I \) is no longer \([-1,1]\). For irregular tessellations, the maximum value of Moran’s \( I \) is slightly larger than 1, and the minimum also tends to be larger than \(-1\). Second, the Moran’s \( I \) value denoting no spatial autocorrelation is \(-1/(n-1)\), not zero. The theoretical range of Moran’s \( I \) values for the 2010 census tracts in Dallas County using matrix \( C \) is \([-0.7151, 1.1843]\); these extremes respectively are functions of eigenvalues associated with matrix \( C \) (Griffith 2003, 37). A Moran’s \( I \) value larger than \(-1/(n-1)\) represents positive spatial autocorrelation. That is, observations tend to have values similar to those of their spatial neighbors. In contrast, Moran’s \( I \) less than \(-1/(n-1)\) represents negative spatial autocorrelation, which indicates that observations tend to have dissimilar values with their spatial neighbors. A Moran’s \( I \) value closer to the maximum (e.g., 1) indicates strong positive spatial autocorrelation and, similarly, a Moran’s \( I \) value close to the minimum indicates strong negative spatial autocorrelation. However, a formal statistical test needs to be performed to confirm whether or not a calculated Moran’s \( I \) value is statistically different from zero spatial autocorrelation.

The statistical test for Moran’s \( I \) can be constructed under the assumption of a normal approximation. The expected value of Moran’s \( I \) is \( \text{E}[I] = -1/(n-1) \) and the variance can be calculated as \( \text{Var}[I] = \text{E}[I^2] - \text{E}[I]^2 \), where \( \text{E}[\cdot] \) denotes the calculus of expectation operator. Detailed derivations of the mean and variance can be found in Cliff and Ord (1981). Although the expression of its variance is complicated, Griffith (2010) shows that the variance is well approximated with the following simple equation for \( n \geq 25 \):

\[
\text{Var}[I] \approx \frac{2}{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}}
\]

(2)

The number 2 appears in the numerator because both \( c_{ij} \) and \( c_{ji} \) are involved in the calculation of Moran’s \( I \). Given the mean and variance, the z-score of Moran’s \( I \) value,

\[
Z[I] = \frac{O[I] - \text{E}[I]}{\sqrt{\text{Var}[I]}}
\]

(3)

where \( O[I] \) denotes the observed Moran’s \( I \), can be used as a test statistic. Here, the null hypothesis is that the observed Moran’s \( I \) is equal to its expected value, implying that observed values do not have spatial autocorrelation. A rejection of the null hypothesis yields the statistical inference that the observed values are spatially autocorrelated.

Geary’s \( c \) is also a widely used spatial autocorrelation index. It can be written as shown in equation 4.

\[
c = \frac{(n-1)}{2 \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - y_j)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]

(4)

The numerator of Geary’s \( c \) is a paired comparisons calculation that measures differences between the values of spatial neighbors. That
is, \((y_i - y_j)^2\) gets larger when spatial neighbors have dissimilar values. In the extreme case that observations have the same value as their spatial neighbors, Geary’s \(c\) becomes zero (in practice, where data are not uniform across a system, the index can only approach zero). As spatial neighbors have increasingly dissimilar values, \(c\) gets larger. The range of Geary’s \(c\) is roughly \([0, 2+\) with 0 representing extreme positive spatial autocorrelation, \(2+\) representing extreme negative spatial autocorrelation, and 1 representing zero spatial autocorrelation. Its theoretical range for the 2010 census tracts in Dallas County using matrix \(C\) is \([0, 3.1578]\); again, these extremes respectively are functions of eigenvalues associated with matrix \(C\). Geary’s \(c\) is inversely related to Moran’s \(I\). Chun and Griffith (2013, 12) show how these two indices are functionally related.

A statistical test for Geary’s \(c\) can be performed under the assumption of a normal approximation. Cliff and Ord (1981) show how the expected value and variance of Geary’s \(c\) are derived. The expected value of Geary’s \(c\) is \(E[c] = 1\), and its variance can be expressed with a relatively complicated expression. Meanwhile, Griffith (2003, 9) shows that the approximate variance of Geary’s \(c\) for \(n \geq 25\) is as given in equation 5.

\[
\text{Var}[c] \approx \frac{2}{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} + 2 \left( \frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{n} c_{ij}\right)^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^2} \right)^2 \tag{5}
\]

Using the expected value and variance, a Geary’s \(c\) z-score can be calculated in a manner similar to equation 3 and a statistical decision can be made about whether or not the null hypothesis, which states no spatial autocorrelation exists, should be rejected.

Figure 2 portrays 2010 population densities for census tracts in Dallas County. The map exhibits positive spatial autocorrelation with clusters of census tracts in dark red and other clusters of census tracts in light red. This can be statistically confirmed with Moran’s \(I\) and Geary’s \(c\). These population densities are transformed using a Box–Cox power transformation with \(\lambda = 0.33\) to better mimic a bell-shaped frequency distribution. The observed Moran’s \(I\) value is 0.4180 based on the binary matrix \(C\) and a rook definition of connectivity, which implies positive spatial autocorrelation. The corresponding expected value and standard deviation (i.e., the square root of the variance) are \(-0.0019\) and 0.0267, respectively. The z-score of Moran’s \(I\) (\(= 15.7128\)) and its \(p\)-value (\(= 0.0000\)) imply rejection of the null hypothesis, and accordingly indicates the presence of positive spatial autocorrelation. The observed Geary’s \(c\) statistic (0.6147) also implies positive spatial autocorrelation in the population densities. Its expected value and standard deviations are 1 and 0.0364, respectively. Its z-score is 10.5866, with an
accompanying $p$-value of nearly 0.0000, which implies rejection of the null hypothesis.

Join count statistics are intended to measure spatial autocorrelation for the nominal measurement scale. A binary definition with 1 denoting “black” (B) and 0 denoting “white” (W) provides a simple example of a nominal variable. A join is a classification of a $c_{ij} = c_{ji} = 1$ in matrix C as one of three types (i.e., BB, BW, or WW); join counts tally the number of joins for each type. These numbers of joins are compared with the expected numbers of joins based on a binomial probability model under the null hypothesis that no spatial autocorrelation is present. For example, a BB join count larger than its expected count under the null hypothesis may imply that B observations are co-located closely in space.

The functional form of the BB join count is

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} y_i y_j$$

(6)

where $y_i = 1$ if the $i$-th unit is black, and $y_i = 0$ if the $i$-th unit is white. The BW join count can be computed with

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} (y_i - y_j)^2$$

(7)

The number 2 in the denominator of these two expressions appears in order to avoid counting each join twice (i.e., $c_{ij}$ and $c_{ji}$). The WW join count can be calculated with the expression $J - (BB + BW)$, where $J$ denotes the total number of joins and

$$BB + BW + WW = J = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}$$

(8)

Cliff and Ord (1981, 38) furnish the expected values and variances for each join type:

$$E[BB] = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} p_B^2$$

(9)

where $p_B$ denotes the probability of a 1, which may be defined empirically as $p_B = n_B/n$ for $n_B$ black units;

$$E[BW] = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} p_B (1 - p_B)$$

(10)

and

$$E[WW] = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} (1 - p_B)^2$$

(11)

The variance of the BB can be calculated as $\text{Var}[BB] = E[BB^2] - E[BB]^2$, and the variances of the other two join types can be similarly derived. The equations for the variances can be found in Cliff and Ord (1981, 38). A statistical test can be performed under the assumption that the join count statistic is asymptotically normally distributed using these expected values and variances.

Chun and Griffith (2013, 13–14) discuss the relationship between the join count statistics and Moran’s $I$. The general formula relating equation 1 to BB, BW, and WW, using the empirical probability $p_B = n_B/n$, is given in equation 12.

$$I = \frac{p_B}{1 - p_B} + \frac{1 - 2p_B}{p_B(1 - p_B) \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} \times 2BB - \frac{1}{(1 - p_B) \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}} \times 2BW$$

(12)

$WW$ is accounted for implicitly by default, because $BB + BW + WW = J$. This equation highlights the role both similar and dissimilar joins play when measuring spatial autocorrelation.

Local spatial autocorrelation measures

Local indices of spatial association (LISA) are widely used measures for local spatial autocorrelation (Anselin 1995). Especially, the local Moran’s $I$ is frequently utilized to explore local
patterns of spatial dependence, and may be defined as shown in equation 13

\[ I_i = \frac{y_i - \bar{y}}{s_Y^2} \sum_{j=1}^{n} w_{ij} (y_j - \bar{y}) \]  

(13)

where

\[ s_Y^2 = \sum_{k=1}^{n} (y_k - \bar{y})^2 / n \]  

(14)

A local Moran’s \( I_i \) represents the level of spatial autocorrelation associated with areal unit \( i \). A positive local Moran’s \( I_i \) implies that the \( i \)-th areal unit has a similar value to its spatial neighbors, and a negative \( I_i \) implies that it has a dissimilar value to its spatial neighbors. Hence, a map portraying the whole set of \( n \) local Moran’s \( I_i \) values can reveal geographic clusters. Specifically, positive local Moran’s \( I_i \) values indicate clusters of similar values. However, this statistic itself does not distinguish a cluster of high values from a cluster of low values.

A test for significance can be conducted with its expected value and variance under the assumption of a normal approximation. Using a randomization hypothesis, the expected value of local Moran’s \( I_i \) is:

\[ E[I_i] = -\sum_{j=1}^{n} w_{ij} / (n - 1) \]  

(15)

Anselin (1995, 99) reports the corresponding variance. Equation 2 modified for local Moran’s \( I_i \) does not provide a good approximation of its variance because local statistics involve a small sample sizes (i.e., a small number of spatial neighbors). The z-score calculated with the mean and the variance, similar to equation 3, is a test statistic that can be referenced to the \( t \) distribution because of its small sample size context.

A local Moran’s \( I_i \) is directly associated with its parent global Moran’s \( I \). Specifically, the sum of local Moran’s \( I_i \)s is proportional to the global Moran’s \( I \). That is:

\[ I = \frac{1}{k} \sum_{i}^{n} I_i \]  

(16)

where

\[ k = \frac{1}{n} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \right) \sum_{i=1}^{n} (y_i - \bar{y})^2 \]  

(17)

In the special case that \( y_i \) is a z-transformed variable and \( w_{ij} \) are row-standardized, \( k \) becomes \( n \), and, consequently:

\[ I = \frac{\sum_{i}^{n} I_i}{n} \]  

(18)

This direct association allows us to investigate the contribution of each areal unit to the global spatial autocorrelation measure, and also to identify influential areal units.

Figure 3 shows an example of a local spatial autocorrelation map using z-scores of the local Moran’s \( I_i \)s. Significant positive local spatial autocorrelation is mapped in light red (at the

![Figure 3 Local Moran’s \( I_i \) for 2010 population density in Dallas County, TX.](image)
10% level) and dark red (at the 5% level). In conjunction with Figure 2, the positive local spatial autocorrelation around the boundaries of the county are clusters of census tracts with low population density. Positive local spatial autocorrelation in the north central part of the map shows clusters of high population density. Negative spatial autocorrelation is represented in light blue (at the 10% level) and dark blue (at the 5% level). The associated p values are calculated without considering multiple testing. Further discussion about multiple testing can be found in Chun and Griffith (2013, Chapter 6).

The Getis–Ord local statistics, $G_i$ and $G^*_i$, furnish another local spatial autocorrelation measure (Getis and Ord 1992; Ord and Getis 1995). These statistics were developed based on a distance rather than a spatial topology definition of spatial neighborhood structure. All areal units within a preset distance $d$ are deemed to be spatial neighbors. Distance $d$ can be considered a threshold for defining spatial neighbors. $G^*_i$ can be defined as shown in equation 19:

$$ G^*_i = \frac{\sum_{j=1}^{n} w_{ij}(d)y_j - \bar{y}\sum_{j=1}^{n} w_{ij}(d)}{s_y \sqrt{\left(n\sum_{j=1}^{n} w_{ij}(d) - \left[\sum_{j=1}^{n} w_{ij}(d)\right]^2\right) / (n-1)}} $$

(19)

where $s_y$ denotes the standard deviation of variable $Y$. The difference between $G_i$ and $G^*_i$ is whether or not a focal observation is included with its spatial neighbors. The $G_i$ statistic does not include this focal observation in its calculation. That is, $w_{ii} = 0$. In contrast, $w_{ii} = 1$ for the $G^*_i$ statistic. Because one less observation (i.e., the focal observation) is considered in $G_i$, its formula can be written by replacing $n$ and $n - 1$ in the denominator with $n - 1$ and $n - 2$, respectively. This distinction leads to a slightly different emphasis by these two statistics. The $G_i$ statistic focuses on spatial association between the $i$-th areal unit and its neighbors within a concentric zone of radius $d$. The $G^*_i$ statistic is intended to identify spatial clusters, because the $i$-th areal unit is considered conceptually to be a cluster member. Positive $G_i$ and $G^*_i$ values indicate clusters of high values that can be easily interpreted as hotspots. In contrast, negative $G_i$ and $G^*_i$ values indicate clusters of low values that can be interpreted as cold spots.

The $G_i$ and $G^*_i$ statistics follow the standard normal distribution when the underlying distribution of the variable $Y$ has a normal distribution. Hence, a calculated statistic itself furnishes a test statistic that can be referenced to the standard normal distribution. One weakness of these statistics is that the variable $Y$ must have non-negative values. Also, these statistics are sensitive to the distance value $d$. Hence $d$ should be carefully selected.

Figure 4 shows $G^*_i$ statistics for the 2010 population density in Dallas County. In this example, Euclidian distances between centroids of the census tracts are considered with $d = 5$. 

Figure 4  Local $G^*_i$ statistics for 2010 population density in Dallas County, TX ($d = 5$ miles).
miles (8 kilometers) defining spatial neighbors. The census tracts in red have significant positive $G_i^*$ values either at the 5% level (dark red) or at the 10% level (light red), indicating clusters of census tracts with high population density. In contrast, the census tracts in blue have significant negative $G_i^*$ values at the 5% level (dark blue) or the 10% level (light blue), indicating clusters of census tracts with low population density. Compared with the local Moran’s $I_i$ map in Figure 3, Figure 4 shows fewer significant values among census tracts around the edge of the county. A small number of spatial neighbors for these edge census tracts may have an impact on this outcome. Also, clusters of low values (i.e., cold spots) are identified.

Graphic portrayals of spatial dependence

Spatial dependence can be represented in forms other than spatial autocorrelation indices, which furnish summary numerical values. Especially, graphical methods provide a way to visually represent and qualitatively measure spatial autocorrelation. These methods are closely associated with exploratory spatial data analysis and usually do not involve a formal test of statistical significance. They include the Moran scatterplot, semivariogram, covariogram, and correlogram.

The Moran scatterplot (Anselin 1995) visually portrays the relationship between the value of the $i$-th areal unit and the summed spatially weighted values of its spatial neighbors. This graphic is a scatterplot between values at each location and their spatially lagged values. It plots the differences between each observation and the mean of the variable $Y$ (i.e., $z_i = (y_i - \bar{y})/s_y$) on the horizontal axis, and $\sum_{j=1}^{n} w_{ij} z_j$ on the vertical axis. As such, equation (1) is rewritten as

$$\frac{n}{\sum_{j=1}^{n} w_{ij}} \sum_{i=1}^{n} z_i \left( \sum_{j=1}^{n} w_{ij} z_j \right)$$

clarifying why the scatterplot is for $z_i$ versus $\sum_{j=1}^{n} w_{ij} z_j$. Figure 5 is a Moran scatterplot of 2010 population density in Dallas County using $z$-scores.

Each point in a Moran scatterplot is directly associated with a local Moran’s $I_i$ for the corresponding areal units. When four quadrants are considered based on the dotted lines (i.e., zero values on the horizontal and vertical axes), the points in the first quadrant represent areal units whose values are larger than the mean and whose spatial lagged values are also larger than the mean. That is, these areas have high values surrounded by high values, and are denoted by HH. In contrast, the points in the third quadrant denote areal units whose values are less than the mean and whose spatial lagged values also are less than the mean – that is, low values surrounded by low values – and are denoted by LL. Similarly, the second quadrant identifies low values surrounded by high values (LH), and the fourth...
quadrant identifies high values surrounded by low values (HL). Accordingly, while the first and third quadrants indicate positive local spatial autocorrelation, the second and fourth quadrants indicate negative local spatial autocorrelation. This direct association is revealed by the reduced form of equation 13:

\[ I_i = z_i \sum_{j=1}^{n} w_{ij}z_j \]  

(21)

where standard deviation of z-scores is 1. Hence, a Moran scatterplot can be used as a diagnostic tool to identify influential cases on the global spatial autocorrelation index (i.e., the global Moran's I). In Figure 5, more influential cases are differentiated with a different symbol, and generally are further away from the mean points, here (0,0). Also, the slope of the regression line without the intercept term is fundamentally equivalent to the global Moran's I that is calculated with the variable Y. With \( z_i \) for the horizontal axis and

\[ \sum_{j=1}^{n} w_{ij}z_j \]  

(22)

for the vertical axis, the formula for the slope

\[ \frac{\sum_{i=1}^{n} \left( z_i \cdot \sum_{j=1}^{n} w_{ij}z_j \right)}{\sum_{i=1}^{n} z_i^2} \]  

(23)

is the second term in equation 1. When row-standardized weights are used, the sum of weights is \( n \). Hence, the slope formula becomes exactly the same as the formula for Moran's I.

A semivariogram furnishes a graphical measure of spatial autocorrelation. It is often used to measure spatial autocorrelation among values of point features. It portrays the magnitude of the variance (the sill), the change of the variance with distance, and the distance interval in which spatial dependence appears (the range). It portrays how observations at a specific distance are dissimilar to each other, using a smooth nonlinear trend curve that is defined by a mathematical equation. Figure 6 presents a semivariogram using an exponential function (blue line) to describe the trend line; polygons are replaced with their geometric centroids to obtain point data. The blue line shows that semivariance is zero when distance, \( d \), is zero. It implies that observations zero distance apart (\( d = 0 \)) have zero dissimilarity; in other words, individual observations have zero semivariance and are perfectly correlated. As distance gets larger, the level of dissimilarity increases until a certain distance (the range). That is, as distance between observations increases, the level of correlation decreases (or the level of dissimilarity increases). Once \( d \) is larger than the range, observations become independent and have maximum dissimilarity described by the variance of the variable. An additional feature of a semivariogram is a nugget effect. Although the change in dissimilarity is continuous, it is not necessary zero at \( d = 0 \). Rather, a minimal change in distance may correspond with a sudden change in dissimilarity. This sudden increase in dissimilarity near \( d = 0 \) is captured with a nugget effect, which may indicate, for example, measurement error or specification error.

Estimation of a semivariogram requires exploring the dissimilarity of observed values together with the distance separating them. However, observation pairs do not provide a sufficient number of samples at every possible distance \( d \) (which is on a continuous scale). Therefore, a dissimilar pattern is discerned for bins that are defined with a range of distance intervals, such as is done when constructing a standard empirical frequency distribution. That is, a sample semivariogram, \( \hat{\gamma}(d) \), is estimated based upon:

\[ \hat{\gamma}(d) = \frac{1}{2 \cdot |N(d)|} \sum_{i \in N(d)} (y_i - y_j)^2 \]  

(24)
where \(|N(d)|\) denotes the number of pairs of observations contained in a certain distance interval. The points in Figure 6 that are plotted as coordinates for the center of each bin and its corresponding semivariogram value constitute an empirical variogram. Next, the best trend line is selected from among different semivariogram model specifications based upon parameter estimates and goodness-of-fit criteria. One common model is the exponential function defined by:

\[
\gamma(d) = \gamma_0 + \gamma_1 \left(1 - e^{-d/r}\right)
\]  

(25)

where \(\gamma_0\) is the nugget, \(\gamma_0 + \gamma_1\) is the sill, and \(r\) is the range parameter (i.e., spatial autocorrelation asymptotically converges on zero as \(d\) goes to infinity). The Bessel function furnishes another popular semivariogram model and is defined by:

\[
\gamma(d) = \gamma_0 + \gamma_1 \left(1 - \frac{d}{r} K_1 \left(\frac{d}{r}\right)\right)
\]  

(26)

where \(K_1\) is the modified Bessel function of the first order and second kind. A chosen semivariogram model with estimated parameters can be utilized further in spatial interpolation with kriging.

The summation term in the right-hand side of equation 24 can be expanded as equation 27:

\[
(y_i - y_j)^2 = [(y_i - \bar{y}) - (y_j - \bar{y})]^2 \\
\quad \quad = [(y_i - \bar{y})^2 + (y_j - \bar{y})^2] - 2(y_i - \bar{y})(y_j - \bar{y})
\]  

(27)

Accordingly, the semivariogram is a variance term minus an autocovariance term, which produces its concave downward increasing function shape. A plotting of the autocovariance versus distance creates a covariogram. In contrast with the semivariogram, this plot measures how observations covary with other observations that are separated from them by distance \(d\). This plot is also considered a portrayal of spatial autocorrelation. This covariation has a maximum value at \(d = 0\) because this case measures covariation of values with themselves. The maximum covariation at \(d = 0\) is equal to the variance at each location, \(\sigma^2\); that is, \(C(0) = \sigma^2\). As distance increases, covariation decreases. Similar to a semivariogram, a covariogram trend line can be described by a function of distance. A sample covariogram can be estimated as in equation 28:

\[
\sigma^2 = \int_0^d \gamma(d) \, dc \\
\sigma^2 = \int_0^d \left[\gamma_0 + \gamma_1 \left(1 - e^{-d/r}\right)\right] \, dc \\
\sigma^2 = \gamma_0 d + \gamma_1 \left(d - \frac{d^2}{2r} e^{-d/r}\right)
\]  

(28)

Figure 7  An illustration of (a) a semivariogram and (b) a covariogram.
Figure 8 Correlograms for 2010 population density in Dallas County, TX: (a) correlation based; (b) Geary’s $c$ based.

\[
\hat{C}(d) = \frac{1}{|N(d)|} \sum_{N(d)} (y_i - \bar{y})(y_j - \bar{y}) \quad (28)
\]

Figure 7 illustrates the inverse relationship between a semivariogram and a covariogram. Under the specific condition of intrinsic stationarity (Cressie 1993, 67), a covariogram is directly related to a semivariogram: $\gamma(d) = \sigma^2 - C(d)$. However, estimators from samples do not preserve this property.

A correlogram, similar to a covariogram, represents how observations covary using correlation instead of covariation. The correlation can be defined as $C(d)/C(0)$. Hence, correlation has its maximum value of 1 at $d = 0$ and decreases as distance separating observations increases. Covariograms and correlograms often can be utilized to visually represent spatial autocorrelation among areal units in terms of spatial lags (the horizontal axis). A spatial lag $k$ means that two areal units are $k$ steps apart when an immediate neighbor is considered to be one spatial lag. Figure 8a illustrates a correlation-based correlogram for 2010 population density in Dallas County. The observations at one spatial lag are highly correlated (>0.6), with the degree of correlation decreasing as the number of spatial lags increases. That is, observations separated by a large number of lags are weakly correlated. Figure 8b displays a correlogram based on Geary’s $c$. It has an increasing pattern similar to a semivariogram because Geary’s $c$ also measures dissimilarity between the value of an areal unit and those of its spatial neighbors.

Concluding remarks

Measuring spatial dependence requires underlying principles and theories, including a framework defining spatial proximity, numerical indices, statistical inference, and/or graphic representations. These principles and theories describe conditions for appropriately measuring spatial dependence. Such conditions include the geometric type of geographic features, measurement scales for attributes, and constraints of space. First, while measuring spatial
dependence is widely applied to either point or polygon features, line features have not been a common geometric type for dealing with spatial dependence. A lack of line features with attached attributes might contribute to the paucity of such analyses. Furthermore, a framework defining spatial proximity among linear features has not been extensively investigated. Papers related to this topic include Chun (2008) for logical linear features and Yamada and Thill (2007) for physical linear features.

Second, spatial autocorrelation indices often utilize a normal approximation that may not be appropriate for some variables. For example, counts hardly follow a normal distribution, and rates in the range of (0,1) are not likely to follow a normal distribution. Griffith (2010) provides insights about spatial autocorrelation measures for non-normal distributions. Sen and Soot (1977) furnish a spatial autocorrelation index for ordinal measurement scale variables. Third, spatial dependence is generally considered a 2-D continuous geographic space phenomenon, but other types of geographic space merit attention, too. Network space furnishes a popular example of a constrained space. For example, the spatial proximity of two car accidents on a street network can be better represented with network distance along the street network rather than Euclidian distance. Okabe and Sugihara (2012) provide details. Also, spatial dependence in 3-D space may need further development as the availability of 3-D data rapidly increases.

Measuring spatial dependence also requires tools for numerical processing. A number of implementations already are available. First, geographic information system (GIS) software packages include spatial dependence measuring functions. Second, specialized spatial statistical packages also include these functionalities. Finally, conventional statistical packages increasingly include more spatial statistical/econometrics capabilities.

SEE ALSO: Geostatistics; Quantitative methodologies; Spatial analysis; Spatial econometrics; Spatial weights; Tobler’s first law of geography

References

MEASURING SPATIAL DEPENDENCE


Further reading


14
Medical tourism

Valorie A. Crooks
Simon Fraser University, Canada

The term “medical tourism” is used to describe the intentional movement of patients from one country to another in order to seek medical care privately. As this care is intentional, emergency medical care accessed while abroad does not constitute medical tourism. Furthermore, as the care is privately sought (primarily through out-of-pocket payment, although sometimes through private insurance), formal cross-border care arrangements enabling patients to access in another country medical care that is paid for by their home health-care system are distinct from medical tourism. Medical tourism is a highly spatial practice that involves the movement between countries on a highly global scale of patients and the friends and family members who accompany them abroad, health workers and administrators, investors and their capital, and many other individuals.

Medical tourism is a global health services industry involving hospitals and clinics in destination countries that range in development and health indicators, and the procedures offered to international patients are highly varied (e.g., experimental stem cell procedures, fertility procedures, hip replacements, dental surgeries, cosmetic surgeries). Patients from countries around the world are motivated to go abroad as medical tourists for a number of reasons. They may be trying to avoid the high cost of care in their home health-care systems or to access procedures that are illegal in their home countries, among other factors.

It is thought that medical tourism can benefit destination countries by bringing in revenue and new investment opportunities, broadening the range of medical procedures available locally, and retaining health workers who may otherwise migrate elsewhere. At the same time, it is also thought that medical tourism can exacerbate health inequities in destination countries by draining resources from public health-care systems, lessening local citizens’ equitable access to care, and transforming aspects of health worker education (e.g., an increased focus on English language training). There are also concerns that patients can be exposed to a number of health and safety risks both at home and abroad through the practice of medical tourism and that the lack of neutral or third-party information about facilities and procedures threatens the process of informed consent. Meanwhile, advocates of the practice suggest that medical tourism enhances patients’ care options, thereby enabling greater choice and an opportunity for empowerment while lessening “system bottlenecks” in patients’ home countries through providing international alternatives.

SEE ALSO: Globalization; Health geography; Health systems; Tourism
Further reading


Mediterranean-type ecosystems

George P. Malanson
University of Iowa, USA

Mediterranean-type ecosystems exist in the five regions with what is called a Mediterranean-type climate (defined by mild, wet winters and warm, dry summers) (Figure 1). From an ecological perspective, net primary productivity, the base of ecosystem functioning and structure, depends on energy and water being available simultaneously. Thus, Mediterranean-type climates are special conditions for plant growth because the summer growing period, when energy is available, is one of increasing aridity, and Mediterranean-type ecosystems, biomes, have plants that have evolved a range of adaptations to this combination. The Mediterranean-type climate also creates conditions that make fire a regular phenomenon (enough fuel that becomes dry during warm summers) to which the ecosystem has adapted.

The primary adaptations of Mediterranean-type ecosystem plants to the Mediterranean-type climate are means to conserve water while maintaining photosynthesis through the dry summer, relying on soil water from the previous rainy season. The primary adaptation is evergreen sclerophyllous leaves: small leaves with waxy surfaces and stomata that can close when soil water is low. All of these characteristics increase water-use efficiency (the ratio of carbon dioxide uptake to water loss by evapotranspiration) while slowing net primary productivity. The leaves are evergreen because there is no need to expend energy on reproducing them if they can become dormant as necessary. In contrast, some vegetation on drier sites can be drought-deciduous, lacking the structures for water-use efficiency; species grow during the spring but then become dormant by early summer.

The limited annual growth leads to the dominance of woody shrub (many-branched from the base, around 2 m height) life-forms in most Mediterranean-type ecosystems. This physiognomy allows canopy development with low costs for respiration. Other forms include evergreen sclerophyllous trees, such as Eucalyptus species in Australia and a smaller shrub form in drier Mediterranean-type ecosystems.

The shrub form and the sclerophyllous leaves together mean that the Mediterranean-type ecosystem vegetation is extremely flammable in warm dry summers. The shrub forms carry fire because the branches are closely packed with a high surface to volume ratio, and because they maintain upright dead branches, created by age and drought, for years. The sclerophyllous leaves are high in volatile oil content. Other forms, such as eucalypt trees, and even non-sclerophyllous plants, such as Californian coastal sage scrub, also have this high volatile oil content in their leaves.

Mediterranean-type ecosystems have common and uncommon adaptations to fire. Shrubs can commonly resprout from their underground parts, and many Mediterranean-type ecosystem species are well adapted in this regard, storing carbohydrate in root crowns to promote post-fire resprouting. Some shrub species have become obligate-seeders, however, having lost the primitive trait of resprouting but producing more seeds that can remain dormant until a fire triggers germination. This contrast in reproductive strategies (and species exist on a
gradient between the extremes) is seen as an adaptation to differing fire intervals and intensities. Short intervals lead to less intense fire and favor resprouting, while long intervals allow the build-up of fuels and, thus, higher intensity fires that suppress resprouting and open space for reproduction from seed.

Mediterranean-type ecosystems vary among the Mediterranean-type climate regions where they are found. Near the Mediterranean Sea, Mediterranean-type ecosystems are dominated by 2–3 m evergreen sclerophyllous vegetation, such as maquis, with *Quercus ilex* the most common species, but in France a lower shrub form, garigue, also exists with *Quercus coccifera* the most common species. These shrublands can grade into pine (e.g., *Pinus halepensis*). Land use over millennia has affected the distribution of these vegetation types. Shrubs also dominate in the Western Hemisphere. In California and Chile, chaparral and matorral, respectively, are similar to maquis. Californian coastal sage scrub is a variant in the former, often found on lower or south-facing slopes. These forms give way to pines and oaks as precipitation increases. These Mediterranean-type ecosystems are an example of convergent evolution, where different plant lineages have developed similar physiognomies to adapt to the climate.

In South Africa and Australia, the differences in evolutionary history become apparent. The Mediterranean-type ecosystems of Australia, dominated by Eucalyptus, support a taller sclerophyllous woodland, mallee (whether the tall Eucalyptus forests of southwestern Australia are a part of this type of ecosystem may be questioned, but they are in a summer-dry climate), although shrublands do exist. The Mediterranean-type ecosystems of South Africa show the greatest floristic diversity, with fynbos being the heart of the Cape Floral Region. Based on a variety of Proteaceae and Ericaceae, the fynbos is known for its unusually high diversity of species in a small area.
Mediterranean-type ecosystems face several of the challenges of global change. They are adapted to a particular type of climate with relatively narrow geographic distributions in their five regions. The challenge of spatially tracking changing climates, if they continue to exist, is compounded by changing fire regimes. This secondary problem may be greater than the effects of changed temperature and precipitation. Changing land use is also an issue. Except in western Europe, where farm abandonment has recently increased the area of Mediterranean-type ecosystems, habitat destruction for agriculture and other uses has reduced Mediterranean-type ecosystem area to fragments. Whether climate change and fire will affect this land-use trend is, as yet, unknown. Lastly, invasive species are a challenge. The unique floras of Australia and South Africa are affecting each other.

**SEE ALSO:** Biodiversity; Biomes; Disturbance in biogeography; Ecoregions; Geography of evolution

**Further reading**


Megacity

Christopher J. Sutton
Western Illinois University, USA

The number of people living in cities has grown from nearly 750 million in 1950 to nearly 4 billion in 2015. The largest of all urban places are megacities — expansive, continuous urban population centers with at least 10 million people. Through their sheer size, megacities play a central role in the economies of their country and tend to be hubs of global urban connectivity.

Because megacities are defined by extent of urban buildup and not by administrative boundaries, the reported population of a megacity can vary widely. Since the mid-1950s, Tokyo generally has been considered the world’s largest megacity, with a 2014 population of 38 million. The World Bank considers the cities located on the Pearl River Delta in southern China — which includes the megacities of Guangzhou and Shenzhen — to be a single, expansive megacity. With a 2010 population in excess of 42 million, the Pearl River Delta surpasses Tokyo.

In 1950, only New York and Tokyo had populations larger than 10 million inhabitants. The number of megacities had risen to 10 by 1990, 23 by 2010, and the number is expected to increase to 41 by 2030. Of these 41, as many as a dozen are agglomerations with populations exceeding 20 million, and Tokyo, Japan, Delhi, India, and Shanghai, China, each have populations exceeding 30 million.

China and India are expected to account for much of the growth in the world’s population (see Population growth) during the first half of the twenty-first century as the number of persons residing in cities in the two countries is expected to grow by nearly 700 million. The number of megacities in the world’s most developed countries is projected to remain unchanged at seven until the middle of the twenty-first century. New York and Los Angeles will continue to be the only two megacities in the United States. The number of megacities in less developed regions is projected to increase from 22 in to 34.

By 2030, one-third of the world’s megacities will be in China and India. Much of the remaining growth of megacities will occur in Africa and, to a lesser extent, in Latin America, where Bogotá, Colombia, and Lima, Peru, are projected to achieve megacity status by 2030. The world’s largest slums — in excess of one million inhabitants — are found in the megacities of developing countries, particularly those in South Asia and sub-Saharan Africa. Managing growth in megacities, particularly within slum areas, will be a challenge for many countries.

SEE ALSO: Global cities; Slum; Urban geography; World cities

Further reading


Mental health geographies

Cheryl McGeachan
University of Glasgow, UK

Geographies of mental ill-health comprise a number of elements including the spatial variations in incidences of mental ill-health, the different environmental components employed in the creation of therapeutic regimes for those experiencing mental ill-health, and the range of locations occupied by the variety of institutions and facilities designed to diagnose and treat individuals experiencing mental health problems (Philo 2005). The concerns of “mental health geography” have intersected with the interests of a number of other subfields such as “health,” “medical,” and “social” geography. While some geographers remain content to deploy the models and terms of “mental illness,” others prefer to remain skeptical about conventional medical-psychiatric models and seek alternative ways of understanding the wide range of experiences felt by people with mental health problems. In doing so, a number of theoretical frames have been adopted, including insights from “anti-psychiatry” and “post-psychiatry,” utilizing a wide range of conceptual vocabularies from phenomenology, critical social theory, psychoanalysis, psychotherapy, and political economy. In part to connect with the ongoing attempts to reclaim “madness” as a politicized identity such as in the recent case of the “voice-hearing” movement (see Callard 2014), a number of geographers have elected to speak of “madness” rather than “mental illness.” Adoption of this term is also partly to acknowledge that “madness” has historically been the predominant mode of understanding different forms of mental difference (see Philo 2005).

In their substantial review, Wolch and Philo (2000) chart three “waves” of inquiry in mental health geography. The first “wave,” the authors suggest, was grounded in spatial science and often employed quantitative methods to investigate the distribution of “deviant” populations including those experiencing diagnosable mental illnesses. The second “wave,” in contrast and in line with the changing frameworks of the discipline, utilized more qualitative methods and remained rooted in social theory, centering more firmly on the social construction of difference. Trajectories towards a third “wave” are suggested by the authors as harnessing a range of theories and methods to uncover a more nuanced understanding of “place-specific happenings” alongside more structurally determined “space-compressing” processes (Wolch and Philo 2000, 149). This, the authors note, requires an ability to conduct research across scales and to pay acute attention to the uses and potential misuses of research findings, especially in relation to policy agendas. Following the pathways of the “waves” suggested above, this entry seeks to give a flavor of the subfield of mental health geographies. It begins with an introduction to so-called psychiatric geographies, continues through to the changing locational associations of mental health, specifically centering on the important shifts from asylum to post-asylum geographies, and finally returns to the experiences of mental ill-health and the new ways in which geographers are currently delving into the complex worlds these individuals inhabit.
MENTAL HEALTH GEOGRAPHIES

Psychiatric geographies

Throughout the early 1960s, and inspired by the pioneering work of sociologists R.E. Faris and H.W. Dunham (1939) on “mental disorders” in the city and their influential map of schizophrenia in Chicago, a number of geographers began to reconstruct what can be termed the “psychiatric geography” of the city (see Wolch and Philo 2000, 139). The resulting studies used a number of statistical methods to detect cross-correlations with spatially referenced data on other variables such as population densities and housing conditions to establish possible causal influences. From this work an interesting array of findings about inner-city concentrations of particular medicalized groupings (especially of schizophrenia) and their socioeconomic correlates was produced. They tended, in general, to confirm the earlier findings of Faris and Dunham with respect to schizophrenia that demonstrated a clear distance-decay in the prevalence of schizophrenia with increasing distances away from the city center. The quantitative rigor of such studies – which may be termed “spatial epidemiologies” (Wolch and Philo 2000, 139) – offers an elaborate way of describing spatial patterns and cross-correlations. For example, the work of Alun Joseph and Brent Hall (1985) into the concentration of social service facilities, including psychiatric group homes, applied a location quotient analysis to such services in Metropolitan Toronto. The authors argue that the localization data collected could be used in the formulation of placement policies to achieve a more equitable distribution of services; however, such studies do not necessarily succeed in explaining in any great detail the causal mechanisms involved in generating such patterns. Other geographers have attempted to develop their understanding of a city’s psychiatric geographies. Using methodologies that require researchers to enter more closely into the worlds of the range of individuals involved in the different processes of mental health practice allows a picture to emerge that accounts for more “contextual” factors.

Changing locational associations of mental health

An area of concern in the geographical literature focuses on the tensions and contradictions in attitudes and beliefs towards mental ill-health that can be revealed through an historical analysis of the spaces of care for mentally-ill individuals. Often referred to as “asylum geographies” (Wolch and Philo 2000, 138), which broadly refers to studies concerned with investigating the geographies of mental health services, attention has been given to the origins of mental health care facilities in the “lunatic asylums” that appeared across western countries. A number of geographers have used various archival sources to reconstruct the location adopted by asylums, “madhouses,” and hospitals run by a number of (state controlled) bodies and institutions. A key body of work in this respect is Chris Philo’s (see 2005) detailed investigation into “the spaces reserved for insanity.” Much of Philo’s work traces the changing geography of places and spaces associated with madness in Britain through the eighteenth and nineteenth centuries, paying particular attention to the discourses and practices that have created a succession of overlain and often disputed “landscapes of lunacy.” Questions raised in this type of work address the extent to which such locations – the specific spaces, places and environments – were shaped by medical, moral, or economic discourses, or by a deeper sense of wishing to remove certain “troubling” and “frightening” populations from “sight and
The different spatial dimensions of the regimes that operated in these institutions have also been examined. Studies have investigated what is known as “Jarvis's Law” (see Philo 2005, 587) of distance-decay in the numbers of asylum admissions decreasing with increasing distance from the asylum gates, while others have explored the tensions between surveillance and “moral architectures” present in the plans of many of the asylums.

Continuities have been traced between the asylum geographies of the eighteenth and nineteenth centuries through to the contemporary reinvention of carceral solutions to potentially dangerous mentally ill populations. From the eighteenth century onwards the asylum is presented in many western countries as a “carceral space,” with the wider literature presenting the asylum to be a “fortress” or “refuge,” protecting people with mental illness from aspects of harm and risk to which they are vulnerable in the wider community (see Curtis 2010, 185–214). These representations solidify psychiatric institutions’ characterization as “spaces of power,” and geographers have reflected, through figures such as Michel Foucault and his work on Bentham’s Panopticon, on the impacts this has for interpreting institutional settings as reflections on the power relations present in society more generally.

Asylum to post-asylum geographies

A key area of geographical concern has been the closure of older asylum spaces under the post-1960s drive to deinstitutionalization and community care in many western countries. This shift in mental health care ideology from institutional to deinstitutionalization solutions has had far-ranging spatial implications and has given rise to a new wave of geographical research on the different spaces and places of post-asylum care. Moving from large-scale “closed spaces” often located in set-apart areas and concentrating a substantial number of patients, carers, and other health professionals, to a maze of small-scale facilities such as drop-in centers, sheltered homes, general practitioner support, and others, has led to a changing pattern of dispersal across different communities. This significant transition, sometimes couched as one from asylum to post-asylum geographies (Curtis 2010; Wolch and Philo 2000) has, in effect, reintroduced into the spaces and places of everyday life people who had previously been spatially removed from community life.

Geographers have examined a wide assortment of care spaces in the community such as out-patient clinics, counseling services, and home spaces. While some have expressed concerns about the sufficiency of such spaces in terms of numbers and quality of service, others have questioned whether the new mechanisms, dispersed throughout diverse communities, amount to a tighter web of “psychiatric influence” (see Wolch and Philo 2000, 141–143). Two key texts in this respect, Michael Dear and Martin Taylor’s (1982) Not on Our Street and Michael Dear and Jennifer Wolch’s (1987) Landscapes of Despair, offer key statements on these issues. Dear and Taylor’s (1982) work sought to establish the nature, magnitude, and geography of community opposition in Toronto to proposed mental health facilities. In doing so they emphasized the challenges of many communities, notably middle-class suburbs, adopting a NIMBYist (“not-in-my-backyard”) attitude towards mental health facilities and their users. In contrast, Dear and Wolch’s (1987) study more broadly applies theoretical perspectives such as structural analysis and human agency to understand the downstream effects of asylum closure. They survey new
spaces of poverty, drug abuse, homelessness, and even reinstitutionalization through the penal and care systems that deinstitutionalized patients are inhabiting.

Further studies have illuminated the fractious politics involved in determining where facilities operate (often clustering in specific parts of the city, creating areas that are inundated with mental health clients), discussing the hierarchies of power and influence that are at play in these key locational decisions. Attention has also turned to the often abandoned asylum sites themselves. Research has shown the deeply symbolic value still placed on these sites by patients, workers, and communities long after the closure of the facility itself. Studies have demonstrated some of the aftermaths of the closure of certain long-stay facilities and, through detailed qualitative work, highlighted the significance of old hospital buildings and their surrounding sites (Moon, Kearns, and Joseph 2015). Connections can be drawn here with work conducted on the design and implementation of contemporary psychiatric care settings. For example, through their research into the perceptions of a newly built psychiatric unit in London, Curtis and colleagues (2009 in Curtis 2010) argue that important questions are raised in these newly restructured landscapes of care as “clinical space” is now extended into the community. In many ways, the authors argue, the psychiatric hospital has become a “permeable institution” which differs markedly from aspects of older asylum facilities. Attention has also turned to the afterlives of these asylum sites and their new economic, social, and cultural functions in contemporary society. Moon and colleagues (2015), for example, in surveying former public psychiatric hospitals in New Zealand, argue for the transcendent role of stigma in closure processes and subsequent building reuse.

Experiential worlds

A number of studies have also examined how post-asylum geographies have been experienced and encountered “from below.” This work has sought to consider the extent to which communities have proved to be “caring” places and has questioned the ongoing stigmatization of people with mental health problems as too different and therefore segregated in varying ways (see Wolch and Philo 2000). However, a range of studies have alternatively sought to uncover a selection of multifaceted caring practices surrounding people with mental health problems. Work on the “self-help,” “advocacy,” and “arts for mental health” networks materializing amongst people with mental health problems has focused on the friendships and relationships that are fostered in such informal spaces, such as cafes, parks, and street corners. This research focuses upon the “mad” geographies of everyday lives and has considered the role of the material spaces of the world and the different ways these are carefully negotiated by individuals experiencing mental health difficulties, as well as the virtual spaces and the rise of online communities (Parr 2008). In increasingly “insecure” times, research in mental health geography is crosscutting into the heart of broader geographical inquiry, tightly bound into issues such as environmental change and global health. Sustained engagements with the international population mental health agenda, such as the range of reports and statements from the World Health Organization, demonstrates a concern to investigate mental health at a global level and to view it as a crucial component of human health integrating once again with the work in “health geography.” For example, Sarah Curtis (2010, 215–238) discusses the complex question of how mental health might be improved at the population level, placing particular attention on strategies to provide and
promote good mental health in the population as a whole, as well as focusing upon treatment of individuals experiencing psychological distress.

Given the complex worlds that people experiencing mental illnesses present, it is increasingly evident that there is a need to consider the pain and suffering but also the hope that can be felt throughout both their “delusional” phases and everyday spaces of public interaction. Novel theoretical understandings of mental ill-health are being explored that are creating innovative cross-disciplinary studies. For example, recent scholarship is drawing on psychoanalytic and psychotherapeutic theory to inspire different routes of connection between space, psychoanalysis, and mental health geographies (McGeachan 2014). This continued desire to understand the created worlds and lived experiences of individuals experiencing mental ill-health at a range of scales, wrapped up within a continually restructuring landscape of care, demonstrates a continued and growing enthusiasm for exploring mental health geographies in creative ways.

SEE ALSO: Carceral geographies; Community; Health geography; Health inequalities; Psychoanalysis/psychoanalytic geography; Social geography; Therapeutic landscapes

References


Further reading

Mergers and acquisitions

Milford B. Green  
*Western University of Canada*

The core rationale for mergers and acquisitions is to increase shareholder wealth. Pursuit of this goal generates one of the most important activities that exist in the capital markets: mergers and acquisitions (M&As). They influence not only the firms involved but also the regions and cities of operation and the locales of the acquired firms. The terms “mergers” and “acquisitions” are used here in a generic sense. M&As describes nearly any business combination involving new partnerships. Legally, mergers require one company to sacrifice its corporate life. An acquisition, which is the purchase of another company, does not require the cancellation of a charter of incorporation. The term “merger” refers to both mergers and acquisitions. Spatial consequences of the two are indistinguishable and the legal differences need not be of concern here. Transfers of assets, corporate headquarters, control functions, and personnel can affect the losing region or city. Such consequences have spurred researchers to examine in detail the economic and social effects of mergers. The discussion here draws from Green (1990).

Some mergers are large by any standard and size bestows effects. The $202.8-billion takeover of Mannesmann AG by Vodafone AirTouch in 1999 is the largest merger on record. Such a huge sum is unusual, but mergers with values greater than $20 billion are now common. Many corporations are very active in the pursuit of mergers, with some having more than 150 acquisitions in the last 25 years.

There are two types of merger, cross-border and domestic, with domestic being the most common. Domestic mergers happen within a country’s boundaries. This does not preclude firms that have an international presence, but the firm buying and the firm being targeted are headquartered within the same country. The size of the United States’ merger market coupled with the periodic concern by the American public regarding merger consequences makes US mergers the most studied. United States merger studies provide the template for most studies of non-US M&As. In this tradition, the discussion here concentrates on the US experience.

Geographic merger research of the last decade exhibits a more quantitative nature than older studies. The results substantiate previous findings, of merger activity declining with increased distance, of increased likelihood as the size of the urban centers involved increases, and of lower success rates across longer distances. These geographical relations are tied to asymmetries of information. Information is situated on a continuum from explicit (information that is codified and available to a wide audience) to tacit (information associated with personal contacts). The quantity and quality of tacit information typically decreases with distance. More and better-quality information lowers risks associated with M&As. Generally, the more remote the target, and the smaller the city, the less is known, raising risk and possibly lowering returns (Chakrabarti and Mitchell 2013; Basu and Chevrier 2011).

Recently, some domestic megamergers have occurred because of the interventions of investment bankers or merger brokers.
promoting mergers for the fees. Such brokers search for potential merger targets for a firm on their own initiative, presenting the candidates to a potential acquirer. These brokers are spatially concentrated in New York, Chicago, and Boston (Davidson 1985), increasing the likelihood of acquisitions by firms headquartered in those cities. These third-party agents contribute to the geographic findings mentioned above.

Research interest in M&As grows and falls with levels of merger activity. The bulk of M&A research is either of a “how to” nature, or studies mergers as an aspatial economic phenomenon. Financial economists concentrate on efficiency gains that theoretically arise from mergers, while analysts in the industrial organization tradition argue that mergers do not generate significant gains. No consensus exists, which is not surprising given the complexity of the topic and the uniqueness of each merger. Neither body of scholarship shows much concern for the effects mergers may produce in social, political, spatial, or economic arenas. Until very recently, they have seldom considered how changes in merger patterns relate to space or how space influences target firm selection. The geography of mergers can affect the organization of the economy, however, altering regional and urban structures and impacting firm concentration, production consolidation and employment impacts.

Mergers and acquisitions in the United States and elsewhere:
A brief chronology

The following short history of mergers in the United States illustrates the roots of the most widely cited theories of merger origination and of how mergers affect industrial structure. The impact of government regulation and changes in funding alternatives on merger occurrences are also illustrated.

Interest in domestic merger behavior began in the late nineteenth century as the US public became concerned about the rise of monopoly trusts such as Standard Oil and US Steel. This was the first of at least six merger waves. Research generally concedes that mergers occur in waves, although the timing and amplitude of the waves vary by country. It is not well understood why such waves occur: Research into long-term activity is severely restricted by lack of time series data.

The first, the horizontal monopoly wave, was fueled by the drive to create large but short-lived trusts (1895–1904). This wave placed an indelible imprint on the structure of the US economy, leaving in its wake small numbers of very large corporations in basic industries. Antitrust legislation in 1890 was the first significant federal legislation to control mergers, targeting monopoly creation. Subsequent enforcement marked the end of this horizontal merger wave.

Second came the oligopoly wave (1919–1929). This resulted in a sharp rise in the share of manufacturing and mining held by the largest 200 corporations (FTC 1969). Again, this activity prompted US federal legislation to rein in such merger activity, prohibiting mergers that seriously lessened competition. Serious enforcement of this legislation did not occur until after the stock market crash of 1929, in response to public outrage.

The third wave – the conglomerate wave – peaked in the late 1960s. Conglomerates are characterized by weak, or no, linkages between the firms held in a conglomerate. Many conglomerate mergers during this wave were predicated on the false belief that management skills required by a new acquisition are transferable from the unrelated buying
firm. Subsequently, many conglomerates were dismantled due to poor financial performance.

Hostile takeovers, junk bonds, and larger leveraged buyouts characterized the fourth wave of mergers (1984–1989). The bust-up merger became common: selling off an acquired firm’s component pieces, resulting in a divestiture boom. The savings and loan crisis in the United States in the late 1980s helped bring this wave to an end.

The first megamergers (except for the purchase of the R.J. Reynolds Tobacco Company in 1989 for $31 billion) characterized wave five (1993–2000) – defined as mergers above $1 billion. Extremely large corporations were formed in pursuit of advantages that came from a bigger size, with negotiated deals the preferred form of acquisition. Visions of a “new economy,” in which economic boom–bust cycles were thought to be a relic of the past, triggered a speculative market in mergers. The economic environment at the end of the twentieth century was one of apparent prosperity, fueled by the creation of dot-com companies (the Millennium bubble). Lack of proper regulation, malfeasance, and overemphasis on short-term performance caused this bubble to burst.

The sixth wave (2002–2008) (Figure 1) was populated with mergers valued at more than $20 billion in value. Globalization and governments’ attempts to create national champions along with cheap financing were contributing factors. The global financial crisis of 2008 brought wave six to a halt.

Figure 1 shows domestic mergers for six major economies. These approximately mirror US trends, except for China where government policy has a greater impact on merger behavior. The dominance of domestic mergers in China reflects the fragmented nature of Chinese industry. Reorganization of and encouragement of growth by state enterprises played a strong role. In addition, high-technology firms were pushed to scale up, with a friendlier regulatory environment encouraging more mergers.

Explaining mergers and acquisitions

Although mergers vary by type, many theories try to explain the general nature of merger activity. Geography is generally not a primary consideration, although it likely influences the outcome. When undertaking a merger search, management will consider elements from both the internal and external environments of the firm. Internal factors include the general character of the firm, its size in assets or employment, its ability to generate expansion or operating capital internally, the current state of its production technology, allocation of resources among various products manufactured, and the adaptability and expandability of its organizational structure. Externally generated considerations include: competition in various product lines, present and potential size of market demand, production processes and technologies employed by competing firms, conditions in the labor and capital markets, government legislation related to the business environment, import tariffs and tax code changes.

Higher levels of merger activity are correlated with favorable economic conditions, such as high stock market prices, rapid new firm formation, and adequate funding sources. An amenable regulatory environment is often present, one of deregulation or lax enforcement. The process by which a firm decides to undertake a merger is complex. The firm lives in an uncertain environment that calls its continued survival into question. Its need for resources connects the firm to the outside world. To ensure access to resources a firm must itself adapt, or force others to do so. Merger is one
method of adaptation. The merger process is not a routine event in the life of the firm, and requires many nonroutine decisions.

The reasons for any particular merger reflect the reasons as described by theory but also idiosyncratic motives. The more regularly cited theories emphasize markets for corporate control, tax considerations, synergistic gains, and inefficient financial markets, market structure valuation gaps, or executive hubris. “Managerial motivations” are an additional consideration. These are outlined briefly here; a more detailed review can be found elsewhere (Cook 1988).

Explanations emphasizing the market for corporate control are premised on the argument that mergers replace inefficient management teams with more efficient ones. Takeovers are used by external groups that believe the firm could be made more efficient through an exchange of management. This is based on the supposition that the market’s valuation of the firm’s worth is an adequate measure of managerial efficiency: that is, that less efficient firms will be undervalued. The group successfully gaining control of the firm accrues the difference between the old value and the new value imputed by the market due to increased efficiency. This theory has some empirical support.

The free cash flow model applies when a firm has limited growth potential, and existing cash flow is more than necessary to undertake projects (Jensen 1987). Managers may undertake such acquisitions to reduce excess cash, or the acquirer may use the cash, along with dismantling some parts of the firm, to pay for the merger. Having no free cash on hand and being in debt makes a firm a less attractive target, as a potential acquirer cannot use cash obtained from the acquired firm to help pay for the merger.

For tax savings to be a motive, the merger must generate savings significant enough to offset the costs of the transaction. Warshawsky (1987) identified three principal sources of tax savings: a carryover of net losses, an increase in leverage, and a change in the tax basis in asset depreciation. The problem with identifying taxation as a primary motive for acquisition is that such savings may be obtainable through other means. Tax considerations are probably, at best, secondary considerations.

Synergy is among the most cited reasons for a merger. Synergy results when the profitability of a combined unit is greater than the profit that would have been realized had the merged units remained independent. Efficiency gains are realized at both plant and firm level. Empirical demonstrations of such gains are rare. Valuation gaps created by coordination gains are commonly called synergistic enhancements. Although the logic underlying the potential synergistic-type benefits accruing from mergers is secure, empirical evidence shows that expected performance gains are often not realized.
Mergers and Acquisitions

Mergers may also occur as a result of differences in management expectations or differing discount rates between the firms. Market structure valuation gaps revolve around the gains realized through reduced competition in an industry.

In managerial motives, managers enact a strategy of sales or growth maximization that induces merger activity. This type of behavior is fueled by reward systems related more to the size of the firm than its profitability. With management compensation schemes tied to absolute volume measures, a conflict of interest between inside managers and outside stockholders is definitely possible. Evidence shows that the empire-building or growth motive has played a leading role in the drive toward conglomerate mergers in the United States. Related to this is executive hubris and greed, whereby senior management overestimates its ability to manage an acquisition.

Eschewing any principal cause of merger behavior, Steiner (1975) derived a multiple-cause hypothesis. The model theorizes that certain merger motivating forces are always present. These motives are multiple, complex, and controversial. Once one of the motives is triggered, the merger process begins. History supports this view. To summarize, the motives and impacts of mergers are still not well understood. No general theory has been able to effectively explain all facets of merger activity.

Classes of merger activity

Four broad classes of mergers can be identified using the industrial relationship between the buyer and seller: horizontal, vertical, conglomerate, and reciprocal. These classes describe the fit of the acquired company to the current business carried on by the acquiring firm. This type of classification is particularly relevant to the geography of mergers because it helps define the spatial scope of the markets affected. Horizontal mergers involve expansion into similar or related product lines. Vertical mergers result in a logical and orderly arrangement of facilities in the successive stages of a continuous production process. Conglomerate mergers are more difficult to describe specifically. A merger of this type results in a heterogeneous mixture of parts. Reciprocal mergers require purchases by suppliers from a new acquisition. A brief exposition of each class is given here.

A horizontal merger leads to the elimination of a competitor, to an increase in the market share of the acquiring firm, and to an increase in concentration within the industry. There are two subclasses, which are sometimes considered conglomerate but are more properly horizontal. Market extension mergers involve the same products, but within a different geographical area. Product extension mergers involve buying a company that produces a similar but not identical product. Expanding control of a business activity in the same product line is an attempt to increase the firm’s dominance in its relationships with other firms, thereby reducing uncertainty. This strategy may be employed to expand geographically into unsaturated markets or to lessen local competition. Thus, it is horizontal mergers that are most commonly subjected to careful examination by regulatory bodies. The increased concentration generated by this process could, for instance, lead to inefficient pricing in the marketplace if firms coordinated their prices. Many studies of the impact of mergers on employment study mergers of this type.

The vertical merger expands a firm’s control into allied product lines. The acquiring company wants to increase its control over more sources of supply and distribution. A firm can expand vertically either forward or backward, integrating components into its existing business. By integrating forward, the firm attempts to ensure a
continuous demand for its product. For example, a paper producer might acquire stationery firms or publishing firms to guarantee a market for paper. Backward integration helps guarantee raw material, ensuring continuous supply and eliminating vagaries associated with dependence on independent suppliers. For example, a clothing manufacturer might acquire various textile mills to guarantee continuous input for its product.

Conglomerate mergers are diversification forays into unrelated areas, to reduce business risks. Often, conglomerate mergers occur when diversified firms are engaged in the production of goods in different industries with stable technologies. Other conglomerate mergers occur when diversified firms seek to make an efficient distribution network the linchpin of their strategy. The question arises, can conglomerate acquisitions foster a more efficient use of scarce resources? It is undoubtedly true that the potential economies available for horizontal or vertical acquisitions are unlikely to be produced by pure conglomerate acquisitions. However, scholars do identify sources of potential conglomeration economies.

A fourth class, the reciprocal merger, constitutes predatory behavior: an acquirer forces suppliers to purchase from its new acquisition. Such coercion might force competitors to the new acquisition out of business. These mergers are illegal under US antitrust laws due to their potential to reduce competition.

Mergers that entail a search for economies in producing services may also occur. These economies may arise from efficient use of certain inputs into the production of services, such as marketing and accounting systems. Productivity improvement may also result when assets are removed from the inefficient control of conservative managers by a company with dynamic, profit-oriented management. Capital economies may be derived from conglomerate mergers, as capital is transferred or allocated between many alternative components of the company with the objective of maximizing the rate of return while reducing uncertainty. The acquisition of management control makes mergers more attractive than portfolio management for risk reduction.

Cross-border mergers

The second type of geographically defined merger is cross-border mergers, a form of foreign direct investment (FDI). Cross-border mergers accounted for about one-half of all FDI in 2012 (UNCTAD 2013). Because these mergers cross international boundaries, the explanatory literature has identified additional possible rationales. The effects of distance, urban size, and information asymmetry, as discussed for domestic mergers, continue to be relevant.

Globalization is clearly a driving force of cross-border mergers. The pursuit of global firm scale and global markets compel transnational firms to strengthen their position. A cross-border acquisition is generally quicker and more cost-effective than pursuing a greenfield investment. International economic conditions may result in an undervaluation of target firms, making them attractive. The elimination of a local competitor also can be an additional benefit. Two major cross-border merger waves have occurred since 1990. The first wave peaked in 2000, with a value of about $900 billion; the second peaked in 2008 with a value of more than $1 trillion.

Geographical and societal implications

Each acquisition has consequences. Whether spatial or aspatial, they impact local and regional economies. The geographical component of mergers consists of two parts: spatial growth and locational adjustment. The former inevitably
means that new elements are added to a firm’s spatial structure. This spatial dimension complicates aspatial economic explanations, as it introduces an element of monopoly – control by a firm over the area surrounding it. This makes the impact of distance crucial, something that is inconsistent with the ideas of perfect competition and equilibrium integral to classical economic theory.

Most of the literature examining post-merger effects deals with the financial outcomes of the merger. Which shareholders benefitted and which didn’t? Was the merger a good decision? Much less common is a concern with how the change in ownership affects the level of economic concentration in a nation, or how a change to extra-regional control of a firm affects a region. The economic concentration concept deals most commonly with the possible negative effects.

Mergers can lead to a reduction in competition and to inefficient allocation of resources in the economy (FTC 1969). Although this has been a concern, evidence suggests that mergers have maintained current levels of corporate concentration rather than significantly increasing them. Measures of concentration in a particular industry may not be rising, but concentration may still be increasing because large firms now have representation in many industries. An allied concern is the political effect of economic concentration. The belief in the United States is that regions and municipalities should exercise control over their territory. Concentration of corporate power can threaten that autonomy. There is no question that large firms can affect local politics. Whether such involvement is good or bad is open to question. The geographic literature on headquarters relocations is relevant here: the geographic concentration of corporate headquarters is lessening even while corporate power is increasing.

Whether the overall outcome of any particular merger is negative or positive depends on its unique circumstances. Negative effects of absentee ownership or external control include declining employment; decreased power of local elites, creating a leadership void; a greater likelihood of plant closure; reduced expenditure on public institutions; increased infrastructure costs; lost tax revenue; and the risk that the acquired firm becomes a cash cow. Positive effects may include improved civic welfare (because of the participation of better-educated corporate managers); the provision of increased opportunities for local executive talent; and improvement in the tax base. In a rural industrialization setting, increased incomes, reduced income inequality, increased corporate philanthropy, and increased access to debt finance may occur.

Much of the concern with the geographic effects of mergers has been at the local community level and is merger-specific. There has not been any emphasis on a more general overview of mergers. Until recently, geographic merger studies have focused primarily on the United States. While following the templates of earlier studies, the newer literature now deals with Western Europe.

Little attention is given to macro-level-based analyses of the geography of mergers or of industrial class interrelations. Playing an important role in the growth of leading corporations, mergers provide a means by which small and medium-sized companies may create a stronger base for expansion and development. Mergers are formulated and completed in a complex economic system whose components are linked not only by transportation and communication systems but also by corporate structure and business relationships.

Most merger research is an examination of the relationship between mergers, market structure, and efficiency. Inadequate attention has been paid
MERGERS AND ACQUISITIONS

to the social implications of large corporations, and even less attention has been paid to that of mergers. Borrowing from Siegfried (1980), the social outcomes from market structure changes induced by mergers include the distribution of income and wealth and local community welfare. Mergers are a major component of changes in such outcomes because they may dramatically alter market structure.

SEE ALSO: Corporate financialization; Corporations and the nation-state; Cross-border, transnational, and interregional cooperation; Economic geography; Financial geography; Firm foundation and growth; Firms; Industrial geography; Mergers and acquisitions

References


Further reading

Metadata

Barbara S. Poore
US Geological Survey

Metadata, or data about data, describe the quality, content, and origins of a geospatial dataset. In the early 1990s, geographic information systems (GIS) technologies were becoming widely used. However, digital data about objects and their locations on the Earth were scarce and expensive to produce. Reusing geospatial data and exchanging data with other organizations became public priorities. Managing legacy data or using data produced by others can be problematic because geospatial data structures are complex and GIS technologies are inherently transformational. Each transformation results in a changed dataset and subsequent users must be made aware of the origins of the data and the processes that were used to produce them. Metadata that provided this information were seen as a way to increase data sharing. It is important to note that, even though metadata are created by the producer of geographic information, their main audience is meant to be the data user.

The metadata standard

The Internet and the World Wide Web, which also emerged in the 1990s, made widespread distributed sharing of geospatial information feasible and amplified the need for machine-readable metadata that could be exposed to users. In 1994, the Federal Geographic Data Committee (FGDC) (www.fgdc.gov), a committee of US federal agencies formed to promote data sharing among organizations, adopted the content standard for geospatial metadata (CSDGM) (FGDC 1998).

Metadata had roots in cartography. Although a visual medium, the map was always closely associated with written text organized into distinct linguistic codes of identification and explanation. Most maps contain a map collar that lists the map scale, projection, publisher, and other information. Metadata is the digital equivalent of the map collar.

The FGDC offered three rationales for widespread adoption of the CSDGM:

- metadata could help organizations manage their investments in data;
- metadata would promote data sharing among organizations by providing information to online data websites and clearinghouses;
- metadata could help users understand datasets produced by others (FGDC 1998).

Metadata were essential to the establishment of the national spatial data infrastructure (NSDI), part of the national information infrastructure, a Clinton-era technology initiative. The NSDI was described as the “technology, policies, standards, human resources, and related activities necessary to acquire, process, distribute, use, maintain, and preserve spatial data through all levels of government, the private and nonprofit sectors and academia (www.fgdc.org).” The NSDI had four components, an online clearinghouse or catalog of datasets available from different organizations, metadata describing these datasets, a series of standards for different types of geospatial data, and framework data – a cooperatively produced
and maintained national set of basic data such as elevation, transportation, and hydrography that could be used as a background for more specialized analyses. Federal agencies were required, by Executive Order to document their geospatial datasets using the metadata standard beginning in 1995 (FGDC 2000).

The NSDI clearinghouse was based on the model of a library catalog. Different organizations stored metadata records on their own servers. The main clearinghouse site had a gateway interface allowing users to search for datasets. The gateway server would query the distributed clearinghouse servers using the Z39.50 protocol – a search and retrieval standard promoted by the library community – and metadata records would be returned to the user (Figure 1).

The structure of metadata

The CSDGM is essentially a metalanguage – that is, a language about a language. It is structured as a nested hierarchy of object-elements, in conformance to the library model which depends on structured text. The main sections of the standard are shown in Table 1 and Figure 2.

In addition, there are several supporting sections whose contents are reused in various places throughout the standard. These are citation, time period, and contact (Figure 2). The categories that make up the standard were largely drawn from previous efforts in the geospatial community to standardize terms with which to describe spatial data quality for data transfer and interoperability.

Each section of the standard contains one or more compound elements that, in turn, contain one or more subsidiary data elements (Figure 3). In addition, the standard defines each element, the values that are permitted, and production rules for how the elements are to be organized. Some elements are mandatory, some are mandatory if applicable, and some are optional. The production rules are derived from the standard generalized markup language (SGML), an international standard sponsored by the library community. The syntax and semantics of the CSDGM are expressed formally in Backus-Naur form, a method for specifying context-free grammars in computer science.

The CSDGM does not specify any metadata presentation method. Because the metadata content and production rules are written using a formal syntax, it is possible to write computer programs that allow for many different types of presentations. The hierarchical structure of the metadata, derived from SGML, lent itself to the emerging use of the extensible markup language (XML), a subset of SGML that became the standard metalanguage for defining markup languages on the World Wide Web. XML is self-describing, thus a document encoded in XML could be read by all web browsers. In a typical CSDGM implementation, each data element of the standard is enclosed within markup tags that a parser can read, interpret, and display. The tags would allow clearinghouse users to search for particular fields in the metadata, for example, keyword or spatial domain.
**Table 1** Main sections of the CSDGM from the FGDC (1998).

<table>
<thead>
<tr>
<th>Sections of the metadata standard</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification information</td>
<td>Basic information about the dataset</td>
</tr>
<tr>
<td>Data quality information</td>
<td>A general assessment of the quality of the dataset</td>
</tr>
<tr>
<td>Spatial data organization information</td>
<td>The mechanism used to represent spatial information in a dataset</td>
</tr>
<tr>
<td>Spatial reference information</td>
<td>The description of the reference frame for, and the means to encode, coordinates in the dataset</td>
</tr>
<tr>
<td>Entity and attribute information</td>
<td>Details about the information content of the dataset, including the entity types, their attributes and the domains from which attribute values may be assigned</td>
</tr>
<tr>
<td>Distribution information</td>
<td>Information about the distributor of and options for obtaining the dataset</td>
</tr>
<tr>
<td>Metadata reference information</td>
<td>Information on when the metadata was written and by whom.</td>
</tr>
</tbody>
</table>

Adoption of metadata

Despite the mandate for federal agencies to document their datasets, adoption of the CSDGM was slow. The authors of the standard attempted to document every aspect of a generic geospatial dataset. As a consequence, the standard was long and complex, containing 334 different elements, of which 119 were compound elements that did not themselves contain values but expressed the relationships among elements. The length and complexity made it difficult to implement, especially for small organizations. There were attempts to adopt simpler forms of metadata. The Alexandria digital library (http://gcmd.nasa.gov/records/Alex-Dig_Gaz-00.html), for example, a distributed library of georeferenced collections established at the University of California at Santa Barbara in 1994 narrowed the CSDGM to 35 fields. Some organizations that serve geospatial data have adopted the Dublin Core metadata standard, an initiative of the library community to provide simple and generic resource descriptions (http://dublincore.org/metadata-basics/). The Dublin Core has only 15 elements, which may provide enough information to locate datasets in repositories, but not enough to fully evaluate their fitness use.

Paradoxically, many organizations also found they needed additional metadata fields for specialized data, for example, to fully document data from remote sensing instruments. There was confusion about the proper level at which to document data. Should one metadata record suffice...
to describe an entire collection of similar datasets such as a topographic data series? Should each individual dataset be documented? Or if objects in the dataset were frequently updated, should there be metadata at the level of the individual feature? The FGDC modified the CSDGM in 1998 to allow user defined profiles and extensions. This edition of the standard also somewhat simplified production rules (FGDC 1998).

The FGDC did not directly supply tools for implementing metadata. Some tools to automate a metadata workflow were supplied by interested volunteers. For example, Peter Schweitzer of the US Geological Survey wrote *mp*, a parser that allowed computers to read and display metadata in various formats, and also a document-type-definition (DTD) defining the markup elements for expressing the standard in XML (http://geology.usgs.gov/tools/metadata/tools/doc/mp.html). Due to a lack of mandate to document data beyond federal agencies, few commercial vendors initially supported FGDC-compliant metadata in their GIS software, making it difficult to collect metadata at the time of data creation.

**Metadata in the twenty-first century**

There are two trajectories for metadata in the twenty-first century. As the metadata standard was adopted it underwent changes, and technological developments questioned the complexity and user friendliness of the standard – but not necessarily the need for metadata. On the one hand, the CSDGM was internationalized and adopted by many large mapping organizations, becoming a baseline for other standards with a global reach, and encouraging the spread of spatial data infrastructures. Apart from the spatial data infrastructure movement, CSDGM was taken up by many communities of earth and environmental scientists, such as geologists, biologists, and oceanographers, to manage and provide access to scientific data that are not strictly geospatial but that may contain spatial footprints. These scientific communities frequently wrote profiles for the metadata standard that accommodated types of information specific to their disciplines. A profile for biological data is one example (http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/biometadata).

On the other hand, geographic information technologies and methods for sharing data online changed rapidly between the late 1990s and the present (Goodchild, Fu, and Rich 2007). These disruptive changes include the rise of open GIS standards and distributed GIS, the development of web mapping applications and services in the private and public sectors, the semantic web, and the rise of robust communities of volunteer mappers enabled by web technologies, mobile devices, and social media. Each of these changes depended on metadata, but also redefined metadata, posing challenges to the CSDGM.

**Internationalization of the standard**

The FGDC initially enlisted the ASTM (formerly the American Society for Testing and Materials), a formal US group that develops consensus standards for the engineering community, to publish the standard more widely. Meanwhile in the early 1990s, the European community, under the leadership of the French standards body the European Committee for Standardization (CEN) (www.cen.eu), developed a series of geospatial standards, including a standard for geospatial metadata. In 1994, this activity was absorbed into the newly formed Technical Committee 211 (ISO/TC-211) of the International Standards Organization (ISO) (http://www.isotc211.org). ISO is the world’s
largest developer and publisher of standards. It is comprised of networks of national standards organizations with one official voting member from each of 157 countries and participation from many other groups and individuals. ISO standards are developed by consensus through an open process, although unlike the CSDGM the standards are copyrighted and must be purchased. Since 1994, TC 211 has published more than 20 basic standards for digital geographic information. ISO 19115, Geographic information – Metadata (ISO 2014) was first published in 2003 and has been subsequently updated. Like the CDGSM, ISO 11915 is an abstract standard that does not deal with implementation. ISO 11915 adopted much from the CDGSM, but its current version has fewer mandatory and more optional elements, the ability to capture more specific information, for example about temporal accuracy, and support for the documentation of new geospatial technologies including geodatabases, web mapping applications, data portals, ontologies, and data models. The FGDC participates in the technical development process of ISO/TC 211 and worked with Canadian standards organizations to publish a North American Profile of ISO 19115. Agencies are currently in the process of implementing this profile.

Over the past twenty years, geospatial metadata standards have been critical to the development of a global spatial data infrastructure (GSDI) to facilitate data management and to post data to online clearinghouses and catalogs (http://gsdiassociation.org/). The first meeting of the GSDI Association was held in 1996 in Bonn, Germany. Its membership consists of organizations concerned with geographic information, including the European Umbrella Organization for Geographical Information (EUROGI), the FGDC, private corporations, and universities. The ISO metadata document is slowly being adopted by members of the GSDI; for example, the European Commission established a spatial data infrastructure for the European Community through its INSPIRE directive of 2007 (http://inspire.ec.europa.eu). The directive requires that member organizations document their geospatial data using the ISO metadata standard and post information to their data portal.

The Open Geospatial Consortium: portals and interoperability

In 1994, the Open Geospatial Consortium (OGC) (www.opengeospatial.org) formed a voluntary standards consortium with membership from commercial firms, governmental agencies, and academic organizations. This effort was an attempt to align the development of GIS software and standards with the concerns of mainstream information technology development, the World Wide Web, and the open source movement. OGC standards are developed through a consensus process and changes are subject to public peer review. OGC also participates in the development of international geospatial standards through ISO TC/211 and has endorsed ISO 19115, further spreading its influence throughout OGC’s constituent communities. The CSDGM and ISO 19115 are aimed at producers of geospatial data, providing them the means to share data through data catalogs based on the library model. As well as supporting both metadata standards, OGC also published standards for catalog services that specify how metadata should be organized in online catalogs (http://www.opengeospatial.org/standards/cat). Current online geospatial catalogs give access not only to geospatial data, as the NSDI clearinghouses did, but also to online services for displaying, editing, and analysis of geospatial data. These collections of data and services have come to be known as geoportals.
Geoportals are organized differently than the NSDI clearinghouse. In the clearinghouse, the metadata resided on distributed servers that could be searched simultaneously. In a geoportal, metadata about geospatial datasets and services are periodically harvested from distributed servers and placed in a searchable central catalog.

Following trends in the larger world of information technology, OGC emphasized interoperability or the ability of the end user to seamlessly integrate data of different types, produced by different distributed systems, independent of the technical details of how the data were produced or exchanged. There are many ways in which computers must communicate to achieve interoperability in the geospatial realm: network protocols, hardware and operating systems, spatial data files, database management systems, data models, and semantics. A failure of interoperability occurs if machines cannot talk to each other on any of these levels. Complete interoperability continues to be elusive; however, over the past 20 years OGC has produced consensus standards for many of these levels of interoperability, including standards that prescribe how metadata about geospatial data services can be passed among machines.

Disruptive ideas, disruptive technologies

The standards consensus process is time consuming and the culture of technology innovation does not always wait for standards. The year 2005 was a watershed for mapping on the Internet. In late 2004, Google acquired Keyhole, a company funded by the venture capital arm of the US Central Intelligence Agency. Keyhole technology allowed users to virtually “fly around” geographic areas over the Internet. In early 2005, Google launched Google Maps and, later that year, using the technology acquired from Keyhole, Google Earth. Also in 2005, Google provided developers with access to the application programming interface (API) of Google Maps. This meant that map and data services could be delivered over the Internet and used in webpages. That same year saw the crystallization of ideas about a second generation World Wide Web dubbed Web 2.0. Collective intelligence or the notion that the intelligence of “the crowd” was more powerful than any one person alone was advertised as the major feature of Web 2.0. Web 2.0 also signaled a switch from users’ passive consumption of information to active creation. The GeoWeb is a manifestation of Web 2.0. Civilian access to GPS signals, the availability of geospatial APIs, the increasing maturity of open-source geospatial software, and the rapid spread of geo-enabled mobile devices have made it easier for non-experts to access, use, and create geospatial data online. A culture of collaborative online mapping by users who are not GIS professionals has emerged. OpenStreetMap (OSM) (www.openstreetmap.org), an online map of the world collectively developed by millions of users, was also begun in 2005. These technical and conceptual shifts multiplied the uses for some forms of metadata while enabling great quantities of geospatial data to be placed on the Internet without formal content metadata.

Machine-based metadata

Standardized metadata describing protocols and services is essential to the operation of the Internet. Metadata is thus the primary means of negotiations between machines. In the current GeoWeb, programmers can request map images over the Internet in real time from APIs on a server. For these images to travel between machines, it is not required that they have content metadata conforming to the CSDGM, however, the requests themselves are a form of metadata — machine-based metadata. For
example, a program can be written to query the server to ascertain what capabilities the server has through a “GetCapabilities” request expressed in XML. The request specifies parameters describing the service type and name of the request among other things. The server responds to the requesting computer with metadata describing the service, permitting other actions.

The GeoWeb runs on machine-based metadata that has been standardized, often by OGC, but also through de facto standards, that is, standards that have emerged through accepted use rather than being negotiated from the top. Google’s keyhole markup language (KML), an XML notation for displaying map images for use in 2-D and 3-D browsers, was submitted to OGC and adopted in 2008, but it has limited support for content metadata of the ISO or CSDGM type. Google also presents sparse details of the source or quality of the map data it collects and displays to the users of its map applications.

User-centric critiques of metadata

Many current GIS software packages now store content metadata records along with the dataset, allowing users to read the metadata in ISO or FGDC formats. This may make metadata more accessible to end users of these professional systems but researchers have also argued that metadata needs to be even more user centric and include the effects of uncertainty on uses of the data, quality descriptions that can be understood by non-experts, and tools to allow users to determine the effects of data quality on results (Goodchild 2007). In the GeoWeb, databases are likely to contain elements mashed up from many different sources, arguing for more attention to metadata that tracks the changes to each feature. Since producer and user do not communicate directly, metadata must contain information about the semantics of a dataset, or what the objects mean (Comber, Fisher, and Wadsworth 2008). However, current metadata standards, even when they are used, do not accommodate data semantics. Schuurman and Leszczynski (2006) proposed additional metadata fields for the ISO and CSDGM that would be sensitive to semantics and contain qualitative ethnographic data descriptions to elicit the meanings of data objects.

Interactive metadata

Most of the user-generated geospatial data in the GeoWeb, often referred to as volunteered geographic information or VGI, do not contain standardized content metadata, limiting their usefulness in the eyes of many. OSM, as a prime example of VGI, has been studied from the standpoint of data quality. In the United Kingdom, OSM has been compared to Ordnance Survey datasets showing that, in more heavily edited areas, the quality is comparable to the reference dataset (Hakley 2010). Others have taken what little metadata is provided in the OSM database to infer data quality. A distinctive feature of the wiki software that underlies OSM is that all edits are maintained in the database, giving the user a detailed idea of the history of changes over time. Each node in the dataset carries its own metadata about when and by whom it was edited. These metadata also include location and various user-supplied tags that indicate the type of object, the source from which the node was derived, and the other nodes with which it is associated. Researchers have demonstrated that the quality of any object in the OSM database may be inferred from the contribution patterns, the reputation of the user-editor, and the number of nodes with many edits. Each of these methods for assessing data quality depends on the assumptions of collective intelligence. That is, that more eyes on each node will improve
its quality over time. This may be true, but it may also be an incorrect assumption due to the inherent biases of the user-editor base for VGI, which is heavily young and male, and also to some documented occurrences of sabotage.

There is also a vast realm of ancillary data describing OSM, explaining how to use it, and facilitating community discussion and debate. These data are quite diverse, consisting of computer programs, tiling schemes, Internet relay chat (IRC) chat rooms, YouTube video demonstrations, tweets, e-mail discussion groups, and the wiki pages. All of these are socially mediated, produced by the community, and accessible to any user. While this is not metadata in the accepted sense in the professional geospatial community, it is a body of data that could be analyzed by data mining techniques to produce a different kind of bottom-up metadata for VGI projects such as OSM. Potentially this is the next research frontier for metadata. An exploration of these ancillary metadata could inform research into how to better present data quality to the end user for more traditional forms of metadata.

**SEE ALSO:** Geoportals; Geospatial metadata; Geospatial Semantic Web; GIS: history; Ontology: domain applications; Spatial crowdsourcing; Spatial data infrastructures; Spatial database; Volunteered geographic information: quality assurance; Web-mapping services

**References**


Mexico: Sociedad Mexicana de Geografía y Estadística (SMGE) (Mexican Society of Geography and Statistics)

Founded: April 18, 1883
Location of headquarters: Mexico City
Website: http://smge-mexico.blogspot.mx
Membership: 55 academies and 33 members
President: Álvaro Sánchez Crispín
Contact: asc@igg.unam.mx, zamojj@igg.unam.mx

Description and purpose

The founding president was Justo Gómez de la Cortina who was interested in sciences, arts, letters, and socioeconomic studies. The initial principal objectives were to construct a map of the republic, published in 1850, to set up national statistics to consolidate the country as an independent nation, and to analyze the principal problems of the country in order to present alternative solutions to authorities. These actions permitted the establishment of laws to determine the geographic names of cities and settlements. They also promoted conservation of forests and archeological monuments, physiographic exploration of the territory, natural resources and their potential, and demographic, ethnic, and linguistic studies of the population.

Journals or major publication series

Boletín de la Sociedad Mexicana de Geografía y Estadística

La Carta General de la República Mexicana (The General Map of the Republic of Mexico)
The Jucutacato canvas
Portulano Mineado
All available at http://smge-mexico.blogspot.mx/p/joyas-de-la-smge.html

Current activities or projects

The principal current function of SMGE is the diffusion of culture across the separate academies and their affiliates in the states of the Mexican Republic. The Academy of Geography is one of the most active and is in charge of organizing national congresses of geography every two years and symposia on the teaching of geography in Mexico. These events are supported by fees of its members, registration of participants in the events, and assistance from local and state governments, and/or institutions of higher education where meetings are held.

Through the Academy of Geography SMGE has fostered academic excellence since 2000 with a yearly prize for the best bachelor’s thesis presented in oral examination as well as a yearly award for MA and PhD theses presented in oral examination. It also presents the Benito Juárez Geographic Medal of Merit to a distinguished member of the geographic community of Mexico.

Brief history

SMGE started its activities as the National Institute of Geography and Statistics between 1833 and 1839, and then as the Commission of Military Statistics between 1839 and 1849, until changing its name to SMGE at the end of 1850. It is recognized as the first institution to organize scientific investigations in Mexico.
and the first institutional body for the development and practice of geography in Mexico. Francisco Díaz Covarrubias, a distinguished member, was the founder of the first astronomical observatory in Mexico. SMGE represents the strategic value of science for the development of governmental projects. Its successful scientific history stems from its relation with powerful state agencies that were charged with creating bodies and institutions to professionalize their activities and generate scientific and intellectual results to support the earth and social sciences and the humanities. SMGE currently has close to 40 divisions. The SMGE Bulletin enables and supports diffusion and development not only of geography and statistics but also of astronomy, chemistry, natural sciences, medicine, archaeology, linguistics, literature, and history.

Submitted by Irma Escamilla
Microbial biogeography

China A. Hanson  
Queen Mary University of London, UK

Biogeography is the study of the distribution of biodiversity over space and time, including the presence, absence, and abundance of species. It offers insights into the ecological and evolutionary mechanisms that generate and maintain diversity, such as selection, dispersal, speciation, extinction, and species interactions. An understanding of species distributions and the mechanisms underlying them has implications for the management of the environment and its resources, for example, the ability to predict ecosystem responses if species ranges shift due to disturbance. Natural scientists have investigated the geographic distribution of plant and animal diversity for centuries. However, the biogeography of micro-organisms (here, defined as bacteria, archaea, and the smallest eukaryotes) has gained interest only in recent decades, driven in part by the development and application of molecular genetic methods that make it possible to sample and characterize microbial diversity independent of culturing. These methodologies also revolutionized our understanding of the sheer magnitude of microbial abundance and diversity on Earth, and an increasing number of microbiologists have turned to the field of biogeography as a useful theoretical context by which to interpret the factors controlling this immense diversity. Further, understanding the distribution of microbes is fundamentally of interest because microbial activity has many important ecosystem-level and societal consequences, including nutrient cycling, primary production, decomposition, parasitism, and disease.

The Baas Becking hypothesis

A central theme in microbial biogeography has been the so-called Baas Becking hypothesis, which states for micro-organisms, “everything is everywhere—the environment selects” (Baas Becking 1934). The claim “everything is everywhere” is based on assumptions that micro-organisms are so numerous, small, and resistant to changing conditions that they can effectively travel or disperse, rapidly and unhindered by passive means, to every point on Earth. In other words, their distributions are not limited by dispersal. The claim that “the environment selects” implies that different environmental conditions allow for selective growth of microbial taxa that are adapted to those conditions. Therefore the Baas Becking hypothesis predicts that, if distinctive microbial assemblages exist in different locations, this is due to the environment selecting adapted taxa from a globally available pool of diversity.

In recent decades, the Baas Becking hypothesis has generated much controversy, which likely helped spark a new generation of research expanding our understanding of microbial diversity. One reason the hypothesis became so debated is that the postulation that “everything is everywhere” suggests that microbes are not dispersal-limited, and therefore, unlike most plants and animals, there are no barriers to the mixing of microbial populations. Without such physical isolation, classic concepts of

The International Encyclopedia of Geography.  
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.  
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.  
DOI: 10.1002/9781118786352.wbieg0231
biogeography, ecological organization, and even speciation would be fundamentally different for microbes as compared to larger organisms.

Some of the first modern studies to describe the large-scale biogeography of any microbe, and thus attempting to test the Baas Becking hypothesis, found evidence in favor of “everything is everywhere.” These studies detected (i) virtually all globally known species of a particular group of eukaryotic microbe (defined by phenotypic identification techniques to group to the genus or near-species levels) within a single location, and (ii) these same species across a range of widespread and varied habitats (reviewed in Finlay 2002; Fenchel and Finlay 2004). This suggested that such microbes were indeed globally distributed, corroborating the first part of the Baas Becking dictum.

A wave of subsequent biogeographic surveys sought to challenge these results, further testing the ideas of the Baas Becking hypothesis for many different types of microbes in a variety of habitats, including soils, sediments, hot springs, marine, and freshwaters (as reviewed by Foissner 2006; Martiny et al. 2006; Lindström and Langenheder 2011; Hanson et al. 2012). Most of this work took advantage of newly available molecular genetic methodologies, allowing a more sensitive detection and characterization of microbes in environmental samples without the need for culturing or reliance on phenotypic techniques. Culture-independent methods also meant that more samples could be processed at once, so that the distributions of diverse microbes could be characterized at large scales in a statistically robust manner. Collectively, microbial biogeographic studies to date have generated much data, establishing that microbial communities are highly variable in space and time. However, whether this is due solely to environmental selection (as implied by Bass Becking), or whether dispersal limitation also plays a role, remains an ongoing debate, and much research in the field is currently focused on understanding the relative importance of these processes in controlling microbial biogeographic variation.

The processes shaping microbial biogeography

While the Baas Becking hypothesis provided a stimulating initial proposal for the processes potentially driving microbial biogeographic distributions, namely dispersal and environmental selection, a more comprehensive framework for understanding microbial biogeographic mechanisms has recently been adopted (Martiny et al. 2006). This framework borrows from a traditional biogeography perspective, which divides these processes into the effects of current environmental factors and the legacies of historical processes, both of which have an influence on the distribution of animals and plants (Ricklefs 2007). The contemporary environment includes abiotic variables (e.g., temperature, pH, nutrient availability) as well as biotic factors such as the interactive effects of other organisms. This is the same concept as that implied by the “environment selects” portion of the Baas Becking statement. Historical events include anything from the very recent to the geological past that might have influenced present-day distributions and caused differences in composition between locations, such as past environmental conditions, speciation and extinction, and evolutionary or ecological drift (stochastic changes in the composition of genes or communities). To observe evidence of historical processes, however, there must be dispersal limitation of some microbial taxa in the community; otherwise the effects would be long since homogenized, and microbial composition would only depend
on the current environment. This dispersal limitation allows for the effects of events in the recent or distant past to have a current signature on the microbial taxa present in a location. Thus, such a framework provides a more comprehensive way to evaluate the same notions proposed by Baas Becking, essentially the relative importance of environmental selection and dispersal.

**Evidence for the importance of contemporary environmental factors**

To assess the importance of environmental selection on microbial composition, many microbial biogeography studies examine the correlation between community composition (the identity and relative abundance of taxa) and measured environmental variables over space and/or time. For example, numerous studies have linked spatial variation in microbial communities to variation in pH or salinity (e.g., Fierer and Jackson 2006). Other studies have demonstrated that broadly defined habitat types, which are a proxy measure for a range of associated environmental variables, harbor significantly different microbial communities (e.g., Nemergut et al. 2011). Additionally, seasonal or temporal variation in microbial communities is increasingly being recognized, and such trends tend to be significantly correlated with temporal variation in environmental variables (Fuhrman et al. 2006; Gilbert et al. 2012). Nearly every microbial biogeography study to date observes some such relationship between similarity in community composition and similarity in environmental characteristics of those samples (Hanson et al. 2012), providing support for the importance of variation in environmental conditions in explaining why microbes are not uniformly detected in time and space.

**Evidence for the importance of historical factors**

In contrast, biogeographic variation in microbial community composition may arise because not all taxa can disperse to or establish in all locations equally, revealing differences based on the history of events at a location. Assessing the importance of historical factors in shaping biogeographic patterns is not straightforward, however. Because the movement and establishment of microbes are difficult to observe directly, the degree of dispersal limitation for microbes is often interpreted from the observed correlation between microbial community composition and geographic distance, after removing the influence of contemporary environmental factors. Since environmental variation is often spatially structured, as in an environmental cline or gradient (e.g., salinity or elevation), controlling for this environmental variation, either experimentally or statistically, is crucial to testing for the influence of historical factors.

When studies use this approach, some find a correlation between microbial composition and geographic distance after removing the effect of the contemporary environment (e.g., Cho and Tiedje 2000; Whitaker, Grogan, and Taylor 2003; Martiny et al. 2011). For instance, Whitaker, Grogan, and Taylor (2003) found clear genetic differentiation between hyperthermophilic archaea from geothermal hot springs on different continents. Significant genetic divergence between localities is not just a demonstration that the genetic diversity of microbes varies in space but also evidence that the sampled microbial communities are not completely mixed by dispersal of individuals between them – that is, that the communities are isolated from each other due to dispersal limitation.

A handful of studies have provided evidence that historical factors, manifested by some dispersal limitation, can indeed be responsible for
creaing biogeographic patterns in microbes. At the same time, other studies find no evidence for dispersal limitation in micro-organisms (e.g., Finlay 2002; reviewed in Hanson et al. 2012). Thus, a current question in the field is to determine when and where dispersal limitation in micro-organisms is more likely. In addition, if there is dispersal limitation, it remains to be investigated what historical processes are responsible (e.g., drift, speciation, extinction, or past environments).

Caveats to the interpretation of microbial biogeographic patterns

A balance between both contemporary environmental factors and historical processes likely shapes natural microbial communities. However, unraveling the importance of environmental selection, dispersal, and historical factors from observations of biogeographic patterns is not straightforward. First, these processes are likely highly intertwined over multiple scales of time and space (Ricklefs 2007), and pinpointing the singular influence of any particular process within natural communities will be difficult. Additionally, little is known about the ecology of the majority of microbial taxa in nature, so it is quite likely that many important selective factors go unmeasured. If any selective habitat features go unmeasured, they can cause a false relationship between microbial composition and geographic distance, and lead to a potentially incorrect conclusion that the biogeographic pattern is shaped by historical factors.

It is important to consider that the ability to detect biogeographic patterns and the importance of particular underlying processes may depend on various methodological aspects of the study, including the target microbial group and methods used to define taxa, as well as the environment type and spatial and temporal scale examined (Hanson et al. 2012). Microbial taxa likely vary in traits that affect their dispersal and colonization abilities (e.g., spore formation and dormancy stages) or their response to the environment (specific growth requirements versus more generalist strategies), thereby influencing their distributions. Similarly, studies using less sensitive techniques to define taxa may group together collections of microbes that differ widely in traits, making distributions appear more widespread and weakening correlations with environmental factors. Additionally, the importance of dispersal may depend on habitat type, as microbial communities associated with nonfluid (e.g., soils) or “patchy” habitats are expected to be more limited by dispersal than those in highly connected or well-mixed systems, such as the open ocean. Similarly, the likelihood of dispersal is expected to decrease as the distance between locations increases, and thus the importance of historical factors may emerge only at larger spatial scales. Last, because many microbial communities are temporally or seasonally variable, spatial biogeography studies should ensure that the timing of sampling is comparable across samples.

Finally, there are many unique aspects of micro-organisms and their communities that make it challenging to study their biogeography. Microbial communities are exceptionally diverse, consisting of a large proportion of taxa that are so rare (Sogin et al. 2006) that they will not be detected. Thus, characterizing the full complement of all taxa in a sample, or proving the absence of a microbial taxon within highly diverse natural samples, is nearly impossible with currently available techniques. Therefore, we may never be able to prove or disprove whether “everything is everywhere.” It is equally difficult to determine whether observed microbes are actually active in a given sample, and therefore
recognizing their surroundings as a suitable habitat. Certainly, many microbes are capable of dormancy, and evidence suggests that many taxa detected in environmental samples are dormant (Lennon and Jones 2010). Thus, without being able to determine activity from dormancy, the true habitat range of some microbes may not be distinguishable from a ubiquitous distribution.

Conclusions

Evidence suggests that many microbes exhibit biogeographic patterns, and that these patterns are likely shaped by a combination of both environment and history, similar to animals and plants. While at least some microbes appear to be dispersal-limited, it is often secondary to the influence of the contemporary environment. Thus far, these conclusions are broadly similar to larger organisms, although the relative strength of the underlying processes and relevant environmental factors may vary. But, given the immense diversity of microbes and their ecological roles, the distributions of all microbes are probably not uniformly shaped by the same mechanisms in the same ways. Thus the future of microbial biogeography lies in identifying the conditions under which the relative importance of these underlying processes may vary.

The interest in microbial biogeography extends beyond drawing and interpreting a map of microbial diversity. It aids in shedding light on countless microbiology questions and applications, for example, how much and on what scales to sample microbial diversity; the nature of microbial diversification or speciation; the function of microbes in ecosystems or in association with other organisms; and how to design and manage microbial communities with industrial applications such as wastewater treatment or biofuels. At the same time, given the unique qualities of micro-organisms, their varied roles, and their astounding abundances, they likely have much to teach us generally about the organization of life on Earth.

SEE ALSO: Biodiversity; Biogeography; Biogeography: history; Complexity in biogeography; Dispersal, diffusion, and migration; Ecogeography/macroecology (range and body size); Ecoregions; Ecosystem; Endemism; Geography of evolution; Island biogeography; Models and simulation in biogeography; Niche theory and models

References


**MICROBIAL BIOGEOGRAPHY**


Further reading

Microclimatology is a subfield of climatology that focuses near the ground where most living organisms exist and where most human activities occur. Microclimatologists are particularly interested in measuring and understanding exchanges of energy, mass, and momentum between the lower atmosphere and components of the Earth’s land surface, including the soil, vegetation, and surface water. As such, there are numerous applications for microclimatological information in fields such as ecology, agriculture, forestry, hydrology, zoology, air pollution studies, and urban planning.

The beginnings of microclimatology as a field are frequently attributed to pioneering work by Theodor Homen in the late nineteenth century. Homen investigated the differences in energy budgets among various types of soils. His careful observations demonstrated the wide variations in temperatures and humidities at different levels above the surface over the course of a day. Gregor Kraus, a German botanist and plant ecologist, is described by Rudolf Geiger as the “father of microclimatology” for his published text titled *Boden und Klima auf kleinstem Raum (Soil and climate at the smallest scale)* in 1911. This text provides a detailed account of the relationship between plants and microclimate conditions. In 1927, Rudolf Geiger published the first comprehensive book on microclimatology *Das Klima der Bodennahen Luftschicht* (1927) or the *Climate Near the Ground* (published in English in 1965).

The scale of microclimatology

Microclimate phenomena can be identified within certain temporal–spatial scales, although there is no exact consensus on the exact thresholds. Microclimates are often considered as ranging horizontally from $10^{-2}$ to $10^3$ m, so that they might encompass conditions under a rock, on a single plant, within an alley in an urban environment, on an agricultural field, or hillside. Scales slightly larger than this, ranging from $10^2$ to $10^4$ m in magnitude, are often referred to as “local scale.” The depth scale of microclimate phenomena is generally limited to phenomena that occur within the planetary boundary layer (PBL). The PBL is the layer of the troposphere that interacts closely with the Earth’s surface and where turbulence plays an important role in energy exchanges. On average it is about 1 km in thickness but its size is dynamic ranging from as low as 20 m at night up to 5 km during the day. Some limit the scope of microclimatology to phenomena in the surface layer, which is approximately the lowest one-tenth of the PBL ($<10^2$ m) and the zone where most biological activities occur or an even shallower layer called the roughness or canopy layer ($<10^1$ m). Similar to the latter depth scale, Miller (1965) defines the microclimate zone as extending to the top of the forest canopy to the rooting depth of plants. Finally, the timescales of relevance depend on the spatial scale examined, with processes $<24$ h...
within the PBL and as low as 1 s for processes at the lower end of the spatial scale (e.g., <10^-2 m), such as on a leaf.

Microclimatology is closely related to the field of micrometeorology, utilizing the same spatial scales and similar physical processes occurring near the Earth's surface. An important distinction, however, is that micrometeorologists are particularly interested on short-term fluctuations and averages of usually an hour or less while microclimatologists tend to focus more on longer-term statistical measures of meteorological variables, such as at diurnal and seasonal scales, as well as on long-term trends in those variables.

**Physical processes in microclimatology**

In considering physical processes that create different microclimates, the “active” surface is of fundamental consideration. The active surface is the boundary between two media such as air and land, and Oke (1987) notes that it is the principal zone of climatic activity for: energy absorption, exchanges, and transformations; precipitation interception; and where much of the drag on airflow occurs.

The energy balance across this mass-less plane can be defined as:

\[
Q_{net} = Q_h + Q_l + Q_g \quad (1)
\]

where \(Q_{net}\) is net radiation, \(Q_h\) is sensible heat, \(Q_l\) is latent heat, and \(Q_g\) is ground heat flux. Net radiation is the difference between incoming solar and longwave radiation and reflected solar and outgoing longwave radiation. Incoming shortwave radiation varies temporally due to Earth–sun relationships and sky conditions, and incoming longwave radiation varies with atmospheric humidity and cloud cover. Outgoing longwave radiation is controlled by surface temperature and emissivity (>90% for most terrestrial surfaces). Reflected solar is a function of changes in surface albedo, with values ranging from about 5% for a dark wet soil up to 95% for a surface with freshly fallen snow.

The partitioning of surplus or deficit \(Q_{net}\) at the surface plays an important role in microclimate conditions. It is dictated by surface conditions and the ability of the atmosphere to transfer heat. Surplus \(Q_{net}\) during daytime hours is typically removed via convection of \(Q_h\) and \(Q_l\) away from the surface and the transfer of \(Q_g\) into the ground. At night, deficits in \(Q_{net}\) lead to downward transfers of \(Q_h\) and \(Q_l\) to the surface and upwards fluxes of \(Q_g\). Convective exchanges will be increased with greater mechanical and convective turbulence. The exact partitioning between \(Q_h\) and \(Q_l\) depends on surface moisture availability with greater \(Q_l\) relative to \(Q_h\) over wetter surfaces. \(Q_g\) is transferred through conduction and its magnitude depends on the thermal conductivity, which is related to soil type, density, and water content. Of course, the relative importance of these different energy transfers depends on the scale of study. A microclimatologist studying small objects likes leaves or animals, for instance, would focus on radiative and conduction as the principal energy transfer mechanisms.

The water balance at the microscale for a volume can be considered as:

\[
P = E + \Delta r + \Delta s \quad (2)
\]

where \(P\) is water input (e.g., precipitation, irrigation, or dewfall), \(E\) (evapotranspiration), and \(\Delta r\) (net runoff) are outputs, and \(\Delta s\) (net change in soil moisture content) is water storage. Net runoff is often negative over steep slopes, indicating a loss of water from the volume. Variations in soil moisture play a particularly important role in the surface energy balance and partitioning of energy exchanges by altering surface albedo, soil
thermal properties, and the relative importance of latent versus sensible heat exchanges.

Aside from water, mass balances of carbon, nitrogen, and oxygen may also be considered at the microscale with important implications for ecosystem processes. The magnitude and rate of biogeochemical cycling may be altered on microscales by local environmental conditions. Factors such as soil temperature and moisture, for instance, may affect respiration rates from roots and microbial organisms, and therefore carbon cycling. Similarly, locally produced carbon dioxide (CO₂) emissions within urban microclimates have altered the carbon cycle, creating urban CO₂ domes.

Finally, momentum exchange within the microclimate zone affects the amount of turbulence and the transport of mass (e.g., water vapor, aerosols, air pollutants) and energy. The wind field near the ground is strongly influenced by the frictional drag from the underlying surface. Wind speeds decrease sharply close to the ground due to the increased friction, giving rise to a logarithmic profile with height. Variations in surface roughness and atmospheric stability may alter the wind speed profile. Under neutral atmospheric stability, rougher surfaces such as a city center tend to increase the depth of the PBL and produce wind speed profiles that have lower vertical gradients. Eddies may be created by mechanical turbulence near the surface due to surface roughness and vertical wind shear. At greater heights above the ground (>2 m), the relative role of convective turbulence in momentum transfers increase and is controlled by atmospheric stability.

Factors that create microclimates

A variety of natural and anthropogenic conditions can alter the water and energy budgets within a given area to create diverse microclimates within the broader macroclimate environment. Topography is a particularly important factor and can alter microclimates by influencing inputs of solar radiation and the creation or modification of airflow. Differences in the slope and azimuth angles can affect solar radiation intensity and therefore create differences in radiant heating across the landscape. In the Northern Hemisphere mid-latitudes, for instance, south facing slopes may receive considerably more solar radiation than north facing ones. These energy differences can be particularly important in influencing the distribution of plants and animals or dictating favorable locations for agricultural activities.

Topography can also alter many elements of the local water budget. Steep slopes, for instance, favor increased runoff while other topographic configurations may encourage local ponding and wetter conditions. The interception of precipitation at the microscale is controlled by interactions between local winds and topography. More exposed locations may receive less precipitation during windy periods with small rain droplets as they are deflected downwind to more protected locations. Also, under windy conditions, the interception of precipitation will vary based on the angle of incidence of the falling rain droplets. Slopes that are oriented towards and more normal to the wind-induced trajectory of the droplets will receive more accumulation. Finally, topography can influence snow mass balances via differential heating and wind redistribution of snow to more sheltered sites.

Local winds can be created by topographic variations. These winds are best developed under weak synoptic forcing, such as during periods with anticyclonic circulation. During the day, the air above the slopes and floor of the valley will be heated, creating a circulation with upslope (anabatic) flow along the valley.
sides. At night, cold air flows downslope under
gravity and may create cold pockets in low lying
portions of the landscape. Temperatures increase
upslope until the top of the pool of cold air is
reached in a zone called the thermal belt. Above
this point, adiabatic cooling with height occurs.
The thermal belt provides a favorable location
for crops such as vineyards that are sensitive to
low temperatures.

The presence of a vegetation canopy can alter
the energy budget as well as modify air flow. The
particular influence on mass, momentum, and
energy transfers, however, depends greatly on
the stand architecture and variations in foliage
density. Some plant communities like grasses and
cereals show little difference in foliage density
with height while others may have the foliage
concentrated near the top (e.g., many trees, sun-
flowers, maize), the base, or at many layers (e.g.,
rainforest). Most commonly, the presence of a
vegetation canopy alters wind speeds by displac-
ing the active surface above the ground, so that
wind speeds go to zero at some location within
the canopy. This height is referred to as the zero
plain displacement and is often located at about
two-thirds of the height of the canopy. Weak air
motion, then, generally occurs at levels below the
active surface. Of course, there are many scenar-
ios where conditions may vary from this gener-
ality. For instance, forests without a well-defined
understory may channel winds beneath the
canopy to foster greater wind speeds than those at
higher levels and gaps and openings in vegetation
can lead to complex local wind patterns.

Radiative exchanges are also considerably
affected by the presence of vegetation. Solar
radiation penetrating to the floor is reduced
as it is absorbed and reflected by the foliage.
The magnitude of reduction is related to the
characteristics of the foliage, such as density and
height, the particular species, and solar angle.
Approximately, 80–95% of the insolation may
be attenuated in a mature stand of trees. Further,
the amount of light penetration may change
seasonally with leaf development and loss in
deciduous forests. Longwave exchanges beneath
the canopy consist of downward emissions of
longwave radiation from the atmosphere and
vegetation, and outgoing transfers from the
canopy floor. As the canopy becomes denser, net
losses of longwave radiation from the ground are
reduced because of the smaller sky view factor.
Thus, reduced solar heating during the day and
radiative cooling at night, tend to moderate the
climate under plant canopies. Plant canopies can
alter the water budget by intercepting a portion
of precipitation. The amount of interception, for
instance, ranges from 10 to 25% for deciduous
forests and from 15 to 40% for coniferous forests.
As intercepted moisture is evaporated from the
canopy, it does not recharge soil moisture or
provide runoff into streams.

Finally, human activity has transformed the
landscape by replacing natural surfaces with
homes buildings, sidewalks and roads. These
new surfaces considerably alter surface energy
and water budgets, creating new “urban” micro-
climates. Urbanized areas tend to have low
permeability surfaces where precipitation is
quickly lost. Evaporation is reduced because
of less vegetation and moisture storage. Urban
locations are often warmer than neighboring rural
areas, especially at night, due to alterations of the
energy budget. Some of the warmer conditions
are attributed to anthropogenic sources of heat
released from manufacturing, transportation,
and the heating of homes and offices. Also, the
reduced sky view factors associated with buildings
minimizes losses of longwave radiation. The
reduced water storage means more surplus energy
is partitioned into sensible heat (i.e., raising
temperatures) than into latent heat (i.e., phase
changes). Lastly, urban environments tend to have
greater heat storage, which is released in the late
afternoon and evening, elevating temperatures. The presence of buildings and other structures may alter wind speeds within the urban canopy. Generally, the increased roughness will reduce wind speeds relative to those in a rural environment. However, strong winds can be generated in response to channeling of air flow or due to the urban heat island effect which can create pressure gradients drawing air towards the city center.

Microclimatological instrumentation

Microclimatologists frequently use in situ instrumentation to measure surface meteorological conditions and fluxes of energy, mass, and momentum. The focus here will be on the instruments used for the most commonly measured variables of air temperature, humidity, wind speed, and short and longwave radiation. For each meteorological variable, there are often multiple instruments with different response times and degrees of precision. The most suitable instrument depends on the intended application and use.

Air temperature can be measured with mechanical (liquid-in-glass) or electronic (thermocouples and thermistors) thermometers. Air temperature sensors should be shielded from direct solar radiation in a ventilated shelter. Ground surface temperatures or objects such as leaves can be measured with an infrared thermometer. Hygrometers use multiple approaches to obtain a value for atmospheric moisture (e.g., air temperature/psychrometric or variations in the electrical, chemical, or radiative properties as water vapor changes in concentration). Psychrometric and electrical hygrometers are commonly employed for routine measurements of humidity. Hygrometers such as the infrared gas analyzer, which consider the absorption spectra of water vapor, are particularly useful when very accurate observations must be made in short time increments. Similarly, multiple approaches (e.g., mechanical, pressure, thermoelectric, acoustical) are used to obtain wind speed measurements. The cup anemometer (mechanical) is the most commonly used for average horizontal wind speeds. For studies of phenomena like turbulence occurring on short timescales, pressure devices and acoustic anemometers are often employed. Radiative fluxes are measured with many different devices, including pyrradiometers (net radiation), pyranometers (solar radiation), pyrheliometers (direct beam solar radiation), and pyrgeometers (longwave radiation). Reflected shortwave radiation for computation of albedo can be obtained by placing the pyranometer so that it faces the ground. Additionally, a shade ring or “occuling ring” can be used to measure diffuse solar radiation by obscuring the sensing surface from direct beam radiation.

These instruments are often mounted on a tripod or tower to obtain a profile for computing vertical gradients of meteorological variables. Flux-gradient methods utilize these profiles to quantify energy and mass transfers. An alternative approach, called eddy fluctuation or eddy covariance, attempts to directly measure the properties of turbulent eddies as they pass the sensors to obtain energy and mass transfers. Eddy fluctuation approaches provide high quality estimates but require very sensitive instrumentation.

Ongoing developments in microclimatology

Since the latter half of the twentieth century, there have been major advances in the field of microclimatology, including theoretical developments, improvements in in situ micrometeorological instrumentation and remote sensing techniques, and the development of computer
modeling and geospatial techniques. Sophisticated computer models, for instance, are now used to simulate the surface energy balance and fluid flow within the boundary layer. Indeed, urban canopy models have been developed that can account for 3-D shapes of building and are capable of computing the energy balance of different components of the built environment such as roofs, walls, and roads. Similarly, fluid dynamic models are now being used to understand dispersion and transport of pollutants and to understand flow around complex shapes. Remotely sensed data are increasingly employed in microclimatic studies. Light detection and ranging (LiDAR), for instance, has been used to develop 3-D digital terrain models of urban environments and for identifying vegetation architecture within a plant canopy. Also, a combination of remotely sensed data from global positioning systems (GPS) and geospatial techniques (e.g., development of detailed digital elevation models) are used in precision agriculture to identify microclimate differences within a crop and to determine optimal inputs of water and nutrients for portions of the crop within these different microclimate zones. In sum, microclimatologists continue to make important theoretical and applied contributions to an array of fields using the latest ground-based and remotely sensed datasets, and geospatial analysis tools.

References


Further reading


SEE ALSO: Climatology
Microsimulation

Einar Holm
Umeå University, Sweden

Questions in microsimulation

Microsimulation is concerned with questions dealing with the demographic and socio-economic development of populations. Who will be born, get education, employment, and family over the years to come? What happens with income, wealth, and health? Who will die, when and where? Who moves from and to specific places? What about education, labor supply and income distribution, family formation, sick leave, geographic population distribution, settlement structure, and segregation? What is the impact on all these processes of assumptions regarding other parts of the system, such as changed taxes, transfers, immigration policy, medical services quality, fertility attitudes, and nutrition habits?

Such “what if” questions are studied extensively with the help of microsimulation. A recent article in the International Journal of Microsimulation (Jinjing and O’Donoghue 2013) defined a dynamic microsimulation model simply as “a model that simulates the behavior of micro-units over time.” With the help of such models, “what if” questions are often answered by comparing outcome of experimental simulations with and without the proposed policy, behavior or structural change implemented in the model. Or, the simulation produces an alternative, contra-factual development for a historic period to be compared with observed history for the same period.

How to perform a microsimulation

The basic structure of a time-driven dynamic cross-sectional microsimulation model, the most common variant, is quite simple and straightforward, as illustrated by the diagram and pseudocode shown in Figure 1.

The simulation moves an observed or synthetically created start population of individuals with attributes forward in time, year by year, person by person, and event by event while updating their existence and attributes, often via defined events such as death, giving birth, receiving an education, moving, partnering, and having a family, or getting a job and income. In the example, each “card” (representing a record in a computer database) carries information for one person for each year, $T$ and $T+1$, and so on. When an existence status event occurs (death, giving birth, immigrating, and emigrating), the card deck is reshuffled to mirror the new set of persons in the domestic population. Events not changing the basic features of a person’s existence might influence attributes of persons any year they exist in the model (e.g., education increases the attribute education level). The example person Per Andersson survives, becomes one year older, keeps his sex and birth country, does not improve his education but moves from Umeå to Lund to get a new job with substantially higher income while simultaneously divorcing from Lisa Andersson, and so his family status changes from “partner” into “single.” In addition, the link attribute to his specific partner Lisa Andersson is
For each year \( t \) and for each person \( p \) and for each event \( h \):

1. Calculate a probability for event \( h \) for this specific person in her specific context this year.
2. Fix an artificial lottery. If a drawn random number is less than the calculate probability, then the event will occur next year for this specific person.
3. Update corresponding attributes of the person, update families and context.

Calculate yearly aggregate results.

That’s it.

**Alternatives beyond microsimulation**

Microsimulation is about what could have happened or what might happen in the future to a large set of individuals, as is much of the other research and fictional accounts of individuals and the societies of which they are part. Comparing the life biography of a fiction writer’s main character with the series of events occurring to the agents in a microsimulation reveals distinguishing features. The talented fiction writer’s detailed story, told about the main character, is unbeaten in its internal logic of reason, action, and agency, and is often of a larger immediate interest to the reader compared to the story told by the relatively stripped-down sequence of life events produced by a simulation. Therefore, the bulk of biographies, fiction stories, and movies probably have had a much greater impact as decision support as have all microsimulation and other quantitative impact analysis studies together.

But the fiction writer’s substantive and communicative advantage comes with a price. The people surrounding the main character cannot be as thoroughly described as the protagonist. They pop in when convenient for the storyline of the main character and then disappear – without background history or a future life. They do not need to have a coherent life history of their own. Therefore, the invented life path of the main character is fiction in two ways. Not only is the person’s life trajectory invented by the author (no matter how internally consistent and plausible) but it also relies on interaction, at certain points in time and space, with other persons that eventually could not be there in the right moment with the right set of attributes because of constraints and events in their own life paths. These interactions are easy to manipulate in fiction, but harder in life outside the book covers.

In a microsimulation model, on the contrary, each person is a main character. Each and every
one develops their own life in parallel and interacts occasionally with the others. Their existence is not constrained to what is required for other person’s actions. But within that equal frame, their individual choices and actions are different because they are constrained by their own properties, preferences, and abilities – inherited or attained – and by resources and people in their vicinity and further away.

So, in a certain place, at a certain point in time, a person about to mate might not find a suitable and agreeable potential partner (and he or she does not have the fiction writer’s ability to invent one) and, therefore, has to postpone mating and family formation. The same goes for employment. If there is no vacancy in the person’s profession in the local labor market or if competition from other applicants is too fierce given the person’s experience and ability, then such a job is not available for the person at that time and place and a series of secondary adaptions might become necessary: commute to a more distant workplace, move to a living place closer to that workplace, abandon work for a while, go for a new education and profession instead, and so on. Similar constraints and differentiation apply to choice of education, living place, and housing.

On the scale between the fiction story and microsimulation, can be found (i) investigations based on life stories told to an interviewer, (ii) results from larger surveys based on answers to telephone or postal questions, and (iii) results of statistical analysis performed on the same kind of data (survey, census, and administrative registers) as used for starting a microsimulation. Contrary to fiction, interviews give observed information and they tell something about what actually happened to the selected individuals and, therefore, also what could have happened – but with less content than the fiction story. But it is hard to know how special or specific the stories are and whether they are somehow representative for the distribution of stories in the population at large. That problem is basically removed in a well implemented large survey, at the expense of even more detail and rationales compared to what is given in a few deep interviews. A description of the survey however, does not tell much about how different characteristics and actions in life and space link to earlier and later events, within and between individuals and places. That instead is the task for statistical analysis, which draws on a broad set of tools, including a substantial set of spatial regression methods. And, that is also the end of the line for the ambition in most contemporary quantitative social science research, including that performed in many kinds of human geography. So, why move further into microsimulation?

Why microsimulation?

One answer to the question of “why microsimulation” is exemplified by the initial “what if” questions posed earlier in this entry. A discrete choice regression equation might tell about the probability that a female gets a certain education. Another equation gives the probability that she moves if she gets that education, conditioned on a set of additional characteristics such as family status. A third equation (or a mating algorithm) might relate her education to the probability of family formation. When time goes on in the model (as in life), all those and other events occur, partly simultaneously and partly in a sequence, constraining some future choices and enabling others. The major advantage of microsimulation over a single regression equation applied on observables, then, is that the single equation does not answer the “what if” questions about a person some years ahead, as then many of the other drivers in that equation have also changed.
This ability to handle myriad “what ifs” is precisely the work performed by a dynamic microsimulation. Estimated equations and postulated rules for all intertwined individual events are put to work together and the outcome further on becomes contingent on the entire web of earlier life events and characteristics, for the person in focus and for those other persons and vicinities he/she was and became linked to. There are few workable substitutes for this kind of system-wide consistent computation of the development of each person’s existence and attributes in the pure statistical toolbox, even including those approaches that attempt to get at sequencing and relationships among entities and attributes, such as propensity score matching, panel data regression, and structural equation modeling.

Roots of microsimulation

Microsimulation originated with an article by Gay Orcutt (1957) followed by the first model in 1961. He realized the obvious shortcomings of analyzing economic policy based on aggregate statistics, especially those due to the averaging out of all differences between individuals. He proposed that micro-units (decision-making units) should be modeled individually in order to maintain heterogeneity in their responses to policy and development of their attributes. Macro-level results, such as the level and distribution of income, are thereafter obtained by simple summation over the individuals. A large number of microsimulation models has thereafter been produced, successively enhancing the original Orcutt model. Example model along these lines are: DYNASIM, CORSIM, APPSIM, SB3, MOSART, DYNCAN, LIAM, SESIM, LIFE MOD, MIDAS, DYNAMOD, SAGE, LifePaths, FAMSIM, SVERIGE, SMILE, and DYNAMITE. This list of acronyms can be extended into at least some 60 large, dynamic microsimulation models all over the globe but mainly developed in larger and richer countries (Jinjing and O’Donoghue 2013).

Alternatives within microsimulation

A static model changes some conditions and calculates the effect without considering time and side effects. Core examples are the large set of models used to evaluate income distribution consequences of major policy changes, such as a changed tax or benefit schemes. They often contain a huge set of tax rules and answers questions such as “How much more or less will this person/family get as disposable income if the rule is changed?” The implicit assumption is that the effect is instant, that nothing else changes in the meantime, and that no other behavioral adaption occurs (e.g., that the person, due to the rule change, reconsiders labor market participation or profession, or adapts by moving to a place with other rules).

A dynamic model, like the one exemplified above, instead explicitly considers time and duration, and by this focus on dynamics can accommodate side effects that stem from induced adaption or other behavioral responses such as those above. In a dynamic model, equilibrium might be approached but never reached (as outside the model in the real world).

A dynamic model might move just a single cohort (or one individual) forward in time or, instead, it moves each individual in an entire cross-section forward. Obviously, the cohort version cannot and the cross-section version can contain interaction between dynamically changing and adapting individuals.

The effects of events in a dynamic, cross-sectional model can be captured instantly or with a delay or slack corresponding to the time
period used. In an event-driven model, events immediately trigger other events (but they do not necessarily respond immediately) whereas a time-driven model gives each event just one chance to hit each time period (year). For low frequency events such as dying, moving, or getting education, the yearly time-driven event resolution is enough and not (much) influenced by who else dies and so on that year. However, for resource allocation or market-like events, such as mating in the partner market, or finding a job in the labor market, or finding a dwelling in the housing market, conditions change for other persons in the same market much more rapidly with each successful change. In these instances, an event-driven procedure is more appropriate than others (static or time-driven procedure).

Another source of alternatives within microsimulation is the nature of its starting data. The models’ start populations often consist of a random sample of the population obtained from a large survey or a census. In most cases, the number of sample individuals (or households) to simulate is less than 100,000. Several models are based on using a synthetic sample created by merging the information in a survey with detailed one, two or three-dimensional tables of the population as made available from the county’s census or administrative registers. A few models, such as the Swedish SVERIGE models and the Danish SMILE model, have access to individual, longitudinal databases covering each individual in the domestic population.

A demanding task is to accurately model the outcome for large choice sets, such as the choice of education/profession or the choice of destination when moving geographically. Such choices are often modeled with the help of multilogit-based equations or spatial interaction models pruned to produce large errors in the detailed outcome. Imitating the behavior of empirical “twins” with identical values on a core set of attributes (age, sex, place of origin, etc.) might sometimes stand out as a simple alternative. In the short term, such “alignment” to data can produce superior results at the expense of analytical/theoretical explanation, as too much of observed intrinsic, unexplained behavior will be preserved eternally (i.e., a systematic long-term shift in the impact of distance decay and population size on destination attraction will be lost).

Alternatives to microsimulation

Agent-based simulation, like microsimulation, maintains sets of individually interacting agents. It was developed some decades after the start of microsimulation. A dynamic microsimulation model furthers an observed population of individuals with all their changing attributes into the future, period by period. An agent-based model instead invents the starting context of agents with a few properties each (like age, race, location, ability, social network, etc.) but explores in detail the impact of goals, rules and agent interaction on emergence of structures not present in the initially created population. Both types of model applications might contain representations of a few (thousands) or many (millions) agent instances. There is a certain convergence between the two approaches. Agent-based models increasingly utilize microdata and some microsimulation models tend to contain more and more social interaction and ask questions related to emergence of order. Of note, long before the agent-based models appeared, substantial parts of their conceptual and theoretical underpinnings were formulated by the Swedish geographer Torsten Hägerstrand in his “time geography” (Hägerstrand 1970). The Swedish HÖMSKE and SVERIGE models were developed as much based on his time geography as on dynamic microsimulation (Clarke
and Holm 1987; Boman and Holm 1996; Holm and Sanders 2007).

**Spatial microsimulation**

Because of the relatively small samples available for most dynamic microsimulation model projects, very few micromodels produce and maintain fine geographic resolution. One reason for the small samples used is that it is enough for exploring nationwide income and wealth developments and distribution – a common focus of traditional dynamic microsimulation models. Increasingly, however, other objectives, including simultaneously demanding results with high spatial and demographic and socioeconomic resolution, have emerged on the wish list for impact studies. These demands argue for spatial microsimulation requiring very large samples, up to including each person in the target population. Related is the fast-emerging possibility to exploit relational, network-based information, be it the network of relatives from administrative registers or the new huge “big data” sources of personal contacts, purchases, and micromovements often derived from new technologies such as location-aware mobile phones. For example, the network of relatives based on mother and father pointers are useless in a sample because a 10% random sample does not contain the mothers of 90% of the children. Instead, the ideal situation is when an individual longitudinal database is available containing all individuals and all their attributes, including relations as required by the intended simulation. That situation is sometimes at hand in the Nordic countries based on annually updated official individual register data, at least for simpler models.

For most models, however, even in the data rich Nordic context, register data has to be supplemented by imputing survey data (e.g., registers do not contain attitude data) and by constructing synthetic individual data from aggregate cross-tables based on microdata that is not available. For the rest of the world this is the baseline situation. Whether or not annual longitudinal individual registers for the entire population exist, they are in any case usually not available for research and analysis on an individual record level.

The paucity of data is the main reason for creating synthetic populations. Such datasets can be constructed to mirror the population one by one and give a spatial resolution corresponding to the geographic division in used tables. A typical situation is that a large national survey is available and is then combined with, and constrained by, one, two, or three-dimensional cross-tables revealing aggregate information from the public registers. Several efforts to develop iterative algorithms to create useful synthetic populations exist using reweighting techniques, combinatorial optimization, iterative proportional fitting, record cloning, value imputation, and alignment (Rossiter et al. 2009).

The goal of creating synthetic populations is to find a set of individuals with attributes that, when again aggregated into the given tables, exactly reproduce all of them. When all tables derive from the same underlying population that goal should be possible to meet. The problem is that the tables together do not contain all the information in the missing target population. Therefore, several different distributions of attributes over the individuals, that is, different populations, might satisfy the constraints given by all tables and the question becomes “How different are those populations?” The result might be good enough for some analysis and simulation but the created synthetic population lacks some of the higher order interactions present in the unavailable target population and that might
therefore distort the simulation results. Neverthe-
less, since the current reluctance to give ana-
lysts access to existing large scale micro-level data
will likely persist, this development of “reverse
engineering” methods for recreating microdata
will remain necessary for furthering spatial
microsimulation into new realms of socially and
economically important impact analysis.

Future conditions for microsimulation

One obstacle for the development of microsim-
ulation is that their construction in most cases
requires extensive programming from the ground
up. Most current applied dynamic microsimula-
tion models so far, such as the MOSART model
in Norway, the Australian APPSIM model and
the Swedish SESIM model, are hard coded for a
specific national dataset and application range, as
were the original Dynasim and Corsim models
(Caldwell and Morrison 2000). Generic software
packages have not been available despite calls
from practitioners for greater cooperation in the
construction of such expensive models (Lymer,
Brown, and Harding 2007).

This lack of generic modeling approaches is
about to change. Promising efforts have emerged,
such as the ModGen toolset from Statistics Canada
and the LIAM2 development package from the
federal planning bureau in The Netherlands.
There is also heartening growth in the availability
of simulation tools based on the open-source
statistical language R contained in the UrbanSim
open source project (Waddell and Ulfarsson
2004). For modest numbers of agents, the growth
in agent-based frameworks, such as NetLogo or
RePast, helps considerably in the construction
of models with an explicit spatial representation.
Such tools are particularly useful for prototyp-
ing and for the construction of moderate size
models but probably packages will come which
also enable the construction of large, specialized
microsimulation models without requiring almost
any programming in the traditional sense.

This greater availability of more powerful yet
relatively easy-to-implement programming envi-
ronments would greatly enhance the possibilities
for new generations of scientists in academia
and in agencies to apply spatial, micro-based
dynamic modeling to urgent problems regarding
people and their environment instead of entirely
relying on regression analysis with its obvious
shortcomings in representing complex dynamic
interactions.

SEE ALSO: Agent-based modeling;
Behavioral geography; Big data; Data model,
object-oriented; Evolutionary economic
geography; Fertility; Geographic data mining;
Human geography; Microsimulation; Mortality;
Population geography; Representation: dynamic
complex systems; Representation: trajectories;
Residential mobility; Social geography; Spatial
social networks; Suburbanization

References

Boman, M., and E. Holm. 2004. “Multi-Agent Systems, Time Geography, and Microsimula-

Caldwell, S., and R. Morrison. 2000. “Validation of Longitudinal Microsimulation Models: Experience with CORSIM and DYNACAN.” In Microsimula-
tion in the New Millennium, edited by L. Mitton, H. Sutherland, and M. Weeks. Cambridge: Cam-
bridge University Press.


Further reading

The International Journal of Microsimulation (http://www.microsimulation.org/ijm/) is the main source for recent developments in microsimulation. The ModGen toolset from Statistics Canada, mainly aimed at longitudinal, cohort models and the LIAM2 development package from the federal planning Bureau in The Netherlands aimed at building cross-sectional models are readily downloadable toolsets for microsimulation development. The NetLogo agent based framework helps particularly in the construction of models with agent interaction and an explicit spatial representation.
Microwave remote sensing

Fulong Chen
Chinese Academy of Sciences

Electromagnetic radiation is a form of energy released by electromagnetic processes (visualized as a self-propagating transverse oscillating wave of electric and magnetic fields at the speed of light) that plays an essential role in remote sensing using radio waves, microwaves, infrared radiation, visible light, ultraviolet radiation, x-rays, and gamma rays; frequency increases from radio waves to gamma rays while the wavelength decreases. Microwave remote sensing, using radiation with a wavelength from approximately one millimeter to one meter, enables observations in all weather conditions with penetration capability (e.g., of vegetation and subsurface) in proportion to the wavelength. Since the rise in interest in its use in the 1960s, microwave remote sensing has had promising applications in studies concerned with the ocean, land, meteorology, ice, disaster, and even archaeology.

Microwave remote sensing can be separated into two technological approaches, active and passive, according to sensors used. Active microwave sensors have an onboard power supply used to transmit a periodic energy pulse toward observed features and measure the echo signatures; this type of sensor is represented by microwave scatterometers (e.g., QuikSCAT Seawinds scatterometer), altimeters (e.g., Jason), and imaging radars (e.g., Radarsat 1/2 and TerraSAR/TanDEM-X). Passive microwave sensors, on the other hand, do not send pulses of energy; instead, they receive and measure emissivity from the Earth and the atmosphere. This type of sensor is represented by microwave radiometers (e.g., NOAA AMSU and Envisat MWR). Microwave sensors currently in use can be summarized as follows.

- Radiometer: microwave radiometers are used to measure a part of the microwave radiated by thermal radiation from the ground surface and/or the atmosphere. Brightness temperature measured by a radiometer conforms to Rayleigh–Jean’s law. An object radiates an unique spectral radiant flux depending on the temperature and emissivity of the object, thus this radiation is called the thermal radiation.
- Scatterometer: microwave scatterometers are used to measure the backscattering reflected from the surface observed (e.g., 2-D velocity vectors of the sea wind). Microwave scatterometers can be either pulse type or continuous wave (CW) type. The pulse type uses a wide band and there are restrictions in obtaining a license to operate it and to avoid obstructions. The CW type has the advantage that the bandwidth can be reduced to 1/100 times that of the pulse type, reducing the costs of the operation considerably.
- Altimeter: microwave altimeters are used to measure the distance between the sensor platform and the surface observed. Based on the precise measurement of the satellite orbit and the geoid, microwave altimeters are widely used for applications involving ocean dynamics of the sea current, geoid surveys, and sea ice surveys.
- Imaging radar: according to the working function of sensors, imaging radars are classified into real aperture radar (RAR)
MICROWAVE REMOTE SENSING

and synthetic aperture radar (SAR). RAR transmits a narrow angle pulse wave beam in a direction at right angles to the flight direction (also called the azimuth direction) and receives the backscattering from the observed targets to form a radar image. To distinguish among different targets, measurements of angle and distance to a target are made by recording the arrival times of the received signals. Improving the azimuth resolution is the main technical limitation of RAR. Compared to RAR, SAR synthetically increases the antenna’s size to derive an effective large antenna to increase the azimuth resolution via the similar pulse compression technique adopted for the range direction.

Microwave remote sensing has developed rapidly and vigorously owing to the advantages it has over the more conventional optical remote sensing technologies.

- It can be operated in all-weather conditions and both day and night. Statistical analysis demonstrates that 40–60% of the Earth, particularly over the oceans and in tropical regions, suffers from intense cloud cover. Both optical and thermal infrared remote sensing in those areas suffer from data acquisition constraints. In contrast, microwave remote sensing is capable of penetrating clouds to observe the surface by using a longer wavelength; furthermore, illumination from sunlight is not needed, confirming its 24-hour observation capability.
- It has the capacity to penetrate vegetation and even soils and other ground surface materials. Based on its physical principal, the microwave signal is backscattered by a target in a geometrical manner comparable with the wavelength and, therefore, longer wavelengths exhibit greater penetration capabilities. This rule is basically applicable for vegetation categories; taking the forest as an example, the P-band can totally penetrate while the L-band can penetrate only partly. C- and X-band electromagnetic radiations, on the other hand, can only register canopy backscattering. The capability to penetrate into the soil is complicated by a number of factors. Among them the most relevant are moisture content and soil porosity. Generally, increasing moisture content and reduced porosity obstruct soil penetration. Thus, microwave electromagnetic radiation penetrates dry sand to a depth of tens of meters (particularly when a long wavelength signature is applied) but is only capable of surface backscattering and/or shallow penetration (e.g., a few centimeters) when the soil is wet.
- Variable sensing mechanisms can be used. Microwave remote sensing uses variable electromagnetic spectrums to acquire information linked to the geometric and dielectric property of targets. It can often be a complementary tool to observations using optical and thermal infrared remote sensing techniques

A number of characteristics of microwave remote sensing favor its application in a variety of fields.

- Superposition. When two or more wave sources are encountered in space, they obey the rule of independent propagation; that is, each wave maintains its original features (wavelength, frequency, propagation direction, etc.) for the propagation. Nevertheless, at the same time, the vector formation of electric and magnetic fields also follows the principle of vector superposition, resulting
in the vector combination of the actual displacement at the encountered point of waves. Accordingly, any complex waveform can be generated by combining multiple simple waves; that is, a complex periodic function can be resolved by the Fourier transform using a series of simple sine and cosine waves.

- Interferometry. Two or more waves with uniform frequencies and constant phase differences propagate themselves in the same direction; interferometry measures wave amplitude enhancement or decline when two waves are overlaid in space. For instance, the side-looking radar is an intrinsic coherence system. The occurrence of speckle noise is caused by the coherent superposition of randomly scattered echoes within a resolution cell, resulting in the reduction of effective spatial resolution as well as the difficulty of subsequent imaging processing and feature discrimination.

- Polarization. Electromagnetic waves are synchronized oscillations of electric and magnetic fields that propagate at the speed of light. The oscillations of the two fields are perpendicular to each other and perpendicular to the direction of energy and wave propagation, thus resulting in a transverse wave. The oscillation of these fields may be in a single direction (linear polarization) or the field may rotate at the optical frequency (circular or elliptical polarization). Linear polarization is commonly used in radar remote sensing and can be divided into horizontal (H, the electric lines of force lie in a vertical direction) and vertical (V, the electric lines of force lie in a horizontal direction) polarizations. Vertically and horizontally mounted transmitting and/or receiving antennas are designed to transmit and receive vertically and horizontally polarized waves, respectively. Copolarization (i.e., HH and VV) corresponds to the case of uniform polarization waves transmitted and received; alternatively, cross-polarization (i.e., HV and VH) corresponds to oppositely polarized waves being transmitted and received. Generally, polarization modes of electromagnetic waves demonstrate different scattering characteristics, providing a wealth of information for feature detection and recognition.

Thanks to the advantages described above, microwave remote sensing has proved to be of critical value in various applications in the ocean, on land, in the atmosphere, and in ice, as well as during disasters and even in archaeology.

Ocean

Nowadays, environmental and geophysical monitoring of oceans is of growing urgency. Observations from passive microwave radiometers provide global measurements on sea surface temperature, wind speed, water vapor, liquid water in clouds, precipitation rates, sea ice, and sea surface salinity and have led to significant advances in meteorological, oceanographic, and ecosystem research as well as improvements in forecasting climate, weather, and fisheries trends (Gentemann et al. 2010). For example, ocean salinity data are useful for shellfish-bed health monitoring and in detecting coastal water masses, which can be indicative of fish distribution. The accurate geophysical retrieval of microwave remote sensing data in oceanography depends on the selection of optimum electromagnetic wavelength, the development of calibrated brightness for temperature variations, and retrieval algorithms. The microwave radiometer is capable of providing a synoptic remote sensing measurement of salinity. Studies conducted imply the need for the choice of
MICROWAVE REMOTE SENSING

Optimizing wavelength. Calibrating brightness temperature gradients is essential to create highly robust quantitative ocean geophysical retrieval. Successful calibration procedures have integrated geolocation analysis, attitude adjustment, along-scan correction, absolute calibration, hot load correction, and antenna emissivity. Geophysical retrieval of radiometer measurements is determined by a radiative transfer model through regression algorithms to calculate the model-fitting coefficients. A large collection of ocean–atmosphere measurements is firstly assembled; then one half of the measurements is used to derive the regression coefficients while the other half is used to test and validate the algorithm.

Active microwave satellite observations of the ocean surface, particularly from scatterometers, radar altimeters, and SARs, provide information linked to the spatial wind variability over large areas. Wind resources based on microwave scatterometers, for example, ASCAT on board the Metop-A satellite of the European Organization for the Exploitation of Meteorological Satellites and the SeaWinds instrument on board the QuikSCAT satellite of the National Aeronautics and Space Administration (NASA), have been intensively used. Radar altimeters, for example, Cryosat-2, have the potential to monitor mesoscale structures and oceanic circulation. First generation (e.g., ERS-1/2, Radarsat-1, and Envisat ASAR) and second generation (e.g., TerraSAR/TanDEM-X, Cosmo-SkyMed, and Radarsat-2) SAR missions have been used for wind resource estimations. Estimation accuracy can be enhanced through the synergistic use of the integration of measurements via the various spaceborne microwave approaches mentioned above.

Owing to the high spatial resolution, SAR remote sensing has been frequently used in oceanography. One potential is the detection of ocean sand ridge signatures. An ocean sand ridge is a submarine feature of the continental shelf shallow sea formed by sand deposits and tidal currents. Under conditions of low wind and strong tidal current, sand ridges and sea mounts can be imaged by SAR using the sea surface imprints of shallow water bathymetric features. Detection and tracking of ocean oil slicks, resulting from spills from ships, oil platforms, and natural seeps from the ocean bottom, have drawn significant attention from communities dedicated to improving pollution control and mitigation and ecosystem sustainability. Oil slicks in SAR images appear as a dark tone, determined by the dominant mirror reflection. Based on this backscattering anomaly, SAR images play an important role for oil slick monitoring and removal in near real-time.

Land

As the primary living space for humans on Earth, the terrestrial land supports agriculture and other anthropogenic uses. Microwave remote sensing, a complementary synoptic tool when compared to solely optical techniques, plays an important role for the monitoring, evaluation, modeling, and prediction of physical dynamics of the terrestrial Earth.

- Agriculture. Soil moisture is a crucial factor influencing agricultural production. Poor soil moisture is associated with dry spells and drought, leading to crop losses; on the other hand, extremely wet soils reflect poor drainage, which hastens anaerobic microbial processes that reduce soil fertility (Champagne, McNairn, and Berg 2011). Current soil moisture assessment methods primarily rely on in situ meteorological records or borehole measurements. Overcoming the spatial coverage limitation of such point-based observations, microwave remote
sensing has great potential for quantitative estimation of land surface soil moisture, particularly the synergistic use of brightness of temperatures and backscattering coefficients measured by passive and active sensors. Moreover, thanks to the development of SAR, polarimetric SAR (PolSAR), and polarimetric interferometric SAR (PolInSAR), microwave remote sensing is also becoming useful for agricultural land-use classification, crop growth monitoring, and statistical yield estimation.

- **Built environments.** Urbanization is a prominent economic and social issue in the world, particularly for developing countries. Nowadays, more than half of the world’s population lives in cities and 1.4 billion more people are expected to move into cities by 2030. Definitely, urban is the core residential area for mankind. Although they occupy only a small percentage of terrestrial land cover, urban areas significantly alter climate, biogeochemistry, and hydrology at local and regional to global scales, because of their hotspot characteristics linked to production, consumption, and waste generation. Previous studies have shown that global urban mapping can be derived from spaceborne SAR data, such as the newly launched Sentinel-1 with global coverage, because urban regions are indicated as strong backscattering areas and can be easily discriminated from other surface features.

Indeed, urban locations and their surroundings belong to a whole environmental system. For sustainable development, essential elements in the entire environment, for example, land-use and land-cover change, forest, and other natural resources, need to be regularly monitored to derive objective assessments, forecasts, and management. Owing to the development of PolSAR and PolInSAR technologies, SAR remote sensing has found widespread application for land-use classification, particularly in the tropics. Investigations have demonstrated that deforestation and subsequent changes in biomass not only influence the hydrology at the local-regional scale, but also contribute to the long-term carbon balance driving global change. SAR remote sensing for biomass assessments has been widely used to demonstrate the potential for the use of long wavelengths, such as the P-band SAR that was selected as one of the future mission priorities of the Biomass satellite of the European Space Agency (ESA). Furthermore, owing to the high spatial resolution, SAR remote sensing is also preferred for interdisciplinary investigations of hydrology and mineral exploitation.

**Atmosphere**

Near-surface water vapor and precipitation are central hydrometeorological variables whose measurements pose significant challenges for conventional point-scale approaches, such as rain gauges and climate stations. For instance, traditional rain gauge measurements lack spatial representativeness and are susceptible to uncertainties introduced by wind- and exposure-induced errors. Although radar observations on weather overcome the spatial coverage limitations, the reliability of results needs further improvement. In contrast to the two technologies mentioned above, the terrestrial microwave network is becoming a new tool utilizing the interaction of microwave radiation with the atmosphere. The propagation of microwave radiation through the atmosphere is affected by many variables, such as pressure, absolute humidity, temperature, suspended water droplets, and precipitation. Studies have shown that it is feasible to extract liquid precipitation and humidity from microwave
attenuation and phase delay (the atmospheric disturbance estimation of spaceborne multitemporal SAR interferometry, MT-InSAR (Hooper et al. 2012). For the assessment and mitigation of extreme weather events, for example, hail and hurricane, microwave remote sensing also plays an important role. Studies found that there is a strong relationship between the occurrence of hail and microwave temperatures. Furthermore, SAR can yield high-resolution and low-level wind information below the cloud that is rarely detected by other sensors. These characteristics provide theoretical justifications for the use of microwave remote sensing for monitoring and assessment of extreme weather events.

Ice

Spatial distribution and temporal evolution of ice and snow in the cryosphere has tremendous influence on Earth system studies considering its close relationship with analysis of ocean currents, water balance, and atmospheric circulation as well as global climate change. Changes in ice and snowfall exert impacts on the environmental, agricultural, and economic sustainability of human societies. Generally, the ice and snow in the cryosphere of Earth can be categorized according to the landscape occurrence, including snow/ice cap and glacier, permafrost, and floating ice in oceans, lakes, and rivers.

• Snow/ice cap and glacier. It is clear that radar altimeters can easily detect mass changes of snow/ice by differentiation of the geodetic measurements of their heights. These point measurements tend to be of high accuracy (decimeter) but have relatively poor spatial coverage. Glacier flow velocity is a quantitative indicator for the study of glacier dynamics and mass balance in spatiotemporal domains, providing precursory warning of glacier-melting-induced flooding, debris flow, and global warming. SAR remote sensing could be a powerful tool for monitoring the evolution of glaciers using the methods of offset-tracking and differential SAR interferometry (DInSAR). For fast-flowing mountain glaciers, SAR-intensity-based offset-tracking may be the most appropriate technique to measure motion. DInSAR technologies could be applied for mild–medium flowing glaciers owing to their sensitivity to motions only if the interferometric coherence is maintained.

• Permafrost. Research on permafrost in polar, cold mountains and plateaus is still a relatively young but rapidly evolving field of science. The impacts of climate change as well as the response to human activities on permafrost are increasing. Changes in permafrost conditions can greatly influence land surface hydrology, ecosystems, carbon cycle, and landscape as well as engineering and construction. For example, permafrost degradation and expansion in the arid and cold Qinghai–Tibet Plateau would generate significant impacts on the Asian monsoon system, ecosystems, and engineering and constructions (Chen et al. 2013). Monitoring temporal changes in active layer thickness is crucial to understanding and interpreting permafrost degradation. Conventional approaches use available boreholes to measure the depth of zero annual oscillations and the corresponding temperature profiles of the active layer and permafrost beneath. Nonetheless, those point-based measurements have spatial coverage constraints and high-cost limitations. Considering the inversion capability for surface settlements, DInSAR technologies can be employed for estimating the long-term average active
layer by measuring the thaw-season surface subsidence. Previous studies have demonstrated DInSAR’s potential for permafrost monitoring over large areas at high spatial resolution (Short et al. 2011).

- Floating ice. Measuring the seasonal transitions of floating ice, particularly sea ice, is important in order to better understand climate variability for numerical weather prediction and designing global climate models. Monitoring of sea ice phenology is feasible using scatterometer image reconstruction, brightness temperature measurements obtained from advanced microwave radiometers, and spaceborne SAR images. Compared with microwave scatterometer and radiometer images, SAR images are more effective in retrieving distributions and detailing the temporal evolution of sea ice owing to their high spatial resolution. Salinity, temperature, and density of the ice affect the dielectric constant and the penetration depth of the radar wave into the ice, determining the backscattering of return radar echoes. Generally, the radar intensity of new and first-year sea ice is dominated by surface scattering (shown as dark tones in SAR imagery), whereas for older ice, volume scattering is the dominant response (reflected as medium tones in SAR imagery).

Disaster

The application of microwave remote sensing for disaster mitigation, forecasting, and evaluation is primarily based on its all-weather and interferometry advantages. Previous studies have shown that passive radiometers and active scatterometers together with imaging radar can monitor storms by retrieving key parameters of wind, that is, storm eye, force, and direction (Gentemann et al. 2010; Hasager et al. 2015).

SAR constitutes a valuable source of information for flood mapping because of its ability to capture the different scattering behaviors between flooded and nonflooded areas. When the spatial resolution is improved (1 m and even 0.5 m), such as that possible with the use of the second generation of TerraSAR/TanDEM–X, Cosmo-SkyMed, Radarsat-2, and ALOS-2, a fully automated processing chain for near real-time flood detection is feasible. Earthquakes (seismic events) and landslides are two typical geohazards with signatures of crust and surface movements. Quantification of co-seismic deformation is useful for seismic source characterization, evaluation of accumulated or released stress, and detecting indications of future earthquakes. DInSAR is capable of delineating surface deformation in the line-of-sight (LOS) direction of SAR data acquisition with accuracy at the subcentimeter level. Its application for seismic geohazards focuses on source characterization using parameters of position, orientation, and slip distribution along the seismogenic structure. For forecasting and mitigation of landslide hazards, the MT-InSAR and ground-based SAR interferometry (GB-InSAR) are preferred. Usually, spaceborne MT-InSAR methods only measure the superficial displacement of extremely slow (ES, velocity <16 mm year\(^{-1}\)) and very slow (VS, 16 mm year\(^{-1}\) ≤ velocity < 1.6 m year\(^{-1}\)) slopes due to constraints of wavelength as well as temporal baseline (Chen, Lin, and Hu 2014).

In contrast, GB-InSAR is a powerful terrestrial technique, widely used in engineering and in geological applications, to detect the fast displacement; thanks to the higher frequency of measurements as well as the optimal observing geometry, GB-InSAR allows the assessment of surface movement of faster landslides, even for cases that may not be visible from optical spaceborne platforms.
Archaeology

Taking into account the need for high spatial resolution, SAR is gaining interest and appeal among archaeologists, particularly with regard to the detection of archaeological remains and the preventative monitoring of heritage sites.

- Archaeological detection. Compared with optical imagery, penetration is one of the merits of SAR systems. This capability is useful for detecting vegetation-covered relics in rainforests and buried remains (settlements and ancient water systems) in deserts using backscattering. Penetration depends on the wavelength (the longer the wavelength the deeper the penetration), surface properties (roughness, soil porosity, and moisture content), and imaging geometry. The coupling mechanism of surface roughness, incidence angle, and the scattering and propagation mechanism of SAR signals interacting with archaeological targets beneath the surface could be studied in detail in order to establish SAR image interpretation rules for archaeology. Generally, the occurrence of archaeological targets is detected by backscattering anomalies in SAR images, providing clues for the identification of relics through changes in multitemporal SAR images. Furthermore, constructed by humans, archaeological sites tend to be regular in shape and scale, as well as having a distinct, vertical archaeological layer, which results in soil moisture or vegetation anomalies on the surface and even in the surrounding landscape. The strong response to polarimetric signals can be used for target identification by means of polarimetric image fusion, scattering element analysis, and polarimetric target decomposition in order to compensate for the weak information from topology, vertical distribution, and soil moisture content in archaeological sites.

- Preventative monitoring. Deformation is an indicator of vulnerability of ancient remains to natural hazards and anthropogenic activities. DInSAR approaches provide motion measurements for preventive diagnosis, particularly the MT-InSAR techniques, which can detect subtle deformation (with measurement accuracy up to several millimeters). Taking advantage of large spatial coverage, MT-InSAR using spaceborne SAR data could be employed for mass movement monitoring around heritage sites. With regard to the structural stability diagnosis, differential tomography SAR (D-TomoSAR) (Fornaro, Reale, and Serafino 2009) and GB-InSAR are recommended. D-TomoSAR integrates MT-InSAR and tomography into one framework to distinguish scattering at different heights and monitor their geometric displacement. It overcomes the unfavorable layover impact of MT-InSAR (no signature on layover regions) and is capable of detecting motions along the monument in the vertical plane, facilitating the monitoring of fractional motion anomalies on monuments. The GB-InSAR instrument carries a portable coherent radar that is able to acquire both the amplitude and the phase of the received signals echoing from observed targets. The derived wave phase reflects the motion of the targets during the observation period. Parameter specifications of signal (e.g., bandwidth, function distance, and incidence angle) make the GB-InSAR system suitable for fine deformation monitoring and preventive diagnosis of heritage sites. In summary, DInSAR techniques
have the potential to support the sustainable development of heritage sites through the detection of driving forces of instability that can lead to monument damage and collapse.

SEE ALSO: Digital Earth; Radar remote sensing; Spatial analysis; Technology and development

References


Further reading


Migrant division of labor

Jane Wills
Queen Mary, University of London, UK

The concept of the migrant division of labor (MDL) builds on existing understandings of the ways in which the labor market is segmented in relation to gender and ethnicity, by adding a focus on immigration status and/or citizenship. While it is well known that job opportunities are unevenly distributed in relation to gender and ethnicity, increased labor mobility has created additional segmentation processes on the basis of immigration status. The MDL refers to the way in which workers are segmented due to their status as migrants or foreign-born workers without full rights of citizenship in the country in which they are working. Such rights are variable but cover access to the labor market including rules about hours of work, access to welfare benefits for those in and out of employment, and the rules covering asylum applications. These rights play a key role in determining an individual’s incentives to work as well as their desirability in the eyes of employers.

In the context of oversupply, employers are able to differentiate potential workers on the basis of a combination of characteristics that might include skills, education, and immigration status alongside often stereotypical assumptions about nationality, ethnicity, and gender. As such, and in the context of increased penalties for employing those without the legal right to work in the country, those with unambiguous legal residence and rights to work are likely to be prioritized over those with less certain legal status.

The concept of the MDL was first developed through a number of geographical research projects conducted in London (UK) in the early years of the twenty-first century when strong labor supply from the extended European Union (most notably Poland) started to reshape labor markets that had previously been reliant on non-EU migrants with less secure status (including those arriving through the asylum system, international students, and irregular migrants). In this context, immigration status served to reinforce the re-racialization of the hiring queue, prioritizing white European workers over black African workers even if the latter group had been in the country for a much longer period of time (and similar trends have been identified in relation to the position of African American men in low-waged labor markets in the USA (Model 2002)). A growing body of research has highlighted the extent to which foreign-born workers are increasingly differentiated in relation to their immigration status in employers’ hiring queues, with different outcomes across time and space (May et al. 2007; McDowell, Batnitzky, and Dyer 2009; Pratt 2004; Wills et al. 2010).

SEE ALSO: Citizenship; Class; Labor market segmentation; Migrant labor; Migration: international

References


Migrant labor refers to work undertaken away from one’s usual place of residence, often requiring residence in another locale where the worker may not intend, or be permitted, to settle permanently. As a result, migrant workers often have a connection to employment or work in one place, and an ongoing set of connections to an economy, household, family, or community located in one or more locales elsewhere. This migration can be either inter- or intranational, but implicit within the term is the notion that individuals are crossing political, administrative, or institutional borders of some kind in order to engage in work. The incorporation of migrant labor into local labor markets is a trend that grew significantly during the twentieth century, and that continues to expand in the twenty-first.

On its own, however, this brief definition belies the diversity and complexity that characterize contemporary migrant labor. This is in part because the term “migrant” is often used to describe groups of individuals who may hold vastly different rights of residency and work, employment contracts, access to social protections, or settlement aspirations. In legal terms, migrant status can also be produced through diverse forms of territorial regulation – at times a patchwork of immigration, residency, and workplace rights laws. The ways in which these laws “touch down” on individual workers can vary immensely: a refugee claimant who finds employment while awaiting a decision on their claim, a young person working part-time while studying abroad on an international student visa, the unpaid domestic labors undertaken by the trailing spouse of another migrant, and an elite professional employed under a skilled foreign worker scheme may all be considered, in different ways, “migrant labor.” Discursively, the term “migrant” frequently carries distinct racial, ethnic, gender, and class connotations, and is often used to refer to lower-waged and lower-status workers instead of more elite labels such as “expatriates” which tend to be reserved for more affluent, professional migrants. Thus, whether defined through law, policy, daily practice, or public discourse, migrant labor is always a composite social construction produced through multiple – and at times conflicting – social practices and institutional arrangements. Moreover, the delineation of migrant labor is a process that is always spatially embedded in that the laws, institutions, and practices that regulate and construct it vary from place to place, and it is often shaped through complex linkages to “elsewhere.”

In contrast to earlier eras of labor migration which, particularly in Europe and North America, were marked more heavily by opportunities for permanent settlement, labor migration in the twenty-first century has in many cases become increasingly characterized by a variety of intersecting conditions of temporariness relating to employment security, residency rights, and social protections. This is in part the result of a number of broad political economic shifts over the past several decades that have facilitated the incorporation of migrant labor into contemporary economies. The first has been the establishment of a diverse array of new territorial arrangements governing flows of migrant labor,
MIGRANT LABOR

including temporary foreign worker programs, live-in caregiver programs, special economic zones, or political-economic-territorial institutions such as the European Union which have been intended to facilitate the movement of temporary migrant workers across borders. At the same time, associated shifts in the nature of contemporary work and employment have also fostered the use of migrant labor, such as the growing flexibilization and contractualization of work, in which jobs have become increasingly short-term, casualized, or part-time; subcontracted employment arrangements have become commonplace; and/or wages have either been eroded or remained stagnant. Growing wage erosion and employment insecurity have served in some cases to make certain jobs less appealing to “local” workers, while declining unionization in certain sectors has increasingly enabled employers to look to unorganized migrant workers as a source of lower-cost labor. The decline of employer-sponsored apprenticeship and training programs for “local” workers arising from labor market deregulation, meanwhile, has frequently provided the basis on which employers seek out migrant workers with the appropriate skill sets.

Despite the arguably global character of these shifts, the incorporation of migrant labor into local labor markets has been profoundly uneven. Sectorally, the demand for migrant labor tends to be particularly strong in immobile sectors of the economy where offshoring production to countries with lower labor costs has not been possible. These include services and goods which must be produced or consumed “in place”: hospitality and retail services, domestic work, care work, construction activities, financial services, and mining and agriculture are some of the most prominent sectors in which migrants – both high- and low-waged – tend to be disproportionately represented. Occupationally, migrant labor in most countries tends to be heavily concentrated in lower-waged, casualized, or contingent jobs. Particularly in locales where women have entered the waged workforce en masse, these shifts have had deeply gendered impacts on migrant labor flows as the rise of service sector economies and the commodification of previously unpaid and culturally devalued forms of labor in the space of the household, such as care and domestic work, have become key labor markets for migrant women. Geographers and others have thus pointed to the feminization of flows of migrant labor in the last several decades as an interlinked set of relationships accompanying these transformations in work and employment. Finally, at the national scale, workers from some labor-sending countries tend to be particularly concentrated within the ranks of migrant workers; as Lee and Pratt (2012) point out, currently over 8 million Philippine nationals are working abroad in over 190 countries, the majority (approximately 85%) of whom are on temporary labor contracts. The labor markets of some destination countries, such as the United Arab Emirates, meanwhile, have become so structurally dependent on temporary migrant labor in recent decades that these workers now constitute between 80% and 90% of the country’s total population. With respect to migration taking place within national borders, an unprecedented growth of new intranational flows of migrant labor has taken place in recent decades, particularly within countries across South Asia, Africa, and East Asia. Often involving the movement of individuals from rural to urban locales, processes of time-space compression have allowed for the greater movement of people across space within nation-states. In some cases, scholars point out, this has been accompanied by new processes of economic enclosure and dispossession in migrants’ sites of origin, which in turn have eroded local livelihoods and resulted in workers becoming more mobile in
search of jobs. Outside of immigration law and policy, thus, an array of other sovereign legal and institutional relationships can regulate the flows of migrant labor within nation-states, many of which are shaping the contours of urban citizenship for migrants working in cities today.

Perhaps no single national policy has had a greater impact on regulating internal migrants in the contemporary moment than the Chinese *hukou*, or housing registration system. The most recent incarnation of China’s *hukou* regime, established in the mid-twentieth century, designates every citizen as either a “rural” or an “urban” resident. While the constellation of policies known as *hukou* has had a great deal of power to limit the mobility of rural and urban populations, China’s ascendance as a central node in a vast array of global production chains has been underpinned by the incorporation of rural Chinese migrants into urban and peri-urban factories in recent decades. Despite this, rural migrants are not permitted to settle permanently in the city because of their *hukou* designation, thus producing what has been called a “floating population” of an estimated 150 million migrants who live and work away from their *hukou* locations. *Hukou* policies have also played a key role in the systematic marginalization of internal migrants because, without an urban designation, rural migrants in the city are denied access to an array of social welfare services such as unemployment insurance, and subsidized housing, food, and education (Chan 2010; Fan 2004). As Kam Wing Chan (2010) argues, this has effectively enabled local governments to treat internal migrants like foreign migrants. While enduring national regimes like the *hukou*, new international immigration policies, and transformational shifts taking place in the global economy are all extremely important in understanding the growing reliance of many labor markets on migrant labor, they do not explain why individuals migrate, who migrates and who doesn’t, how migrants end up in particular jobs in particular sectors, or how migrants’ identities both shape and are shaped by day-to-day work and employment. Geographical scholarship throughout the 1970s and into the early twenty-first century has sought to complicate macroscale analyses of migrant labor that viewed migrants’ movements primarily as the result of capitalism’s need for a reserve army of labor, and to question rational choice models which assumed that wage differentials between countries “pulled” migrants across borders. This work challenged economistic tropes of the individual migrant, detached from the agency emanating from kinship or community ties or from emotional or other noncapitalist desires motivating migration and mobility. Feminist geographers, in particular, demonstrated the limitations of these various frameworks by uncovering a wider and more complex spectrum of workers’ reasons for migrating; by highlighting how migrant men and women’s experiences of migration, work, and employment could be vastly different; and by foregrounding how unequal and often patriarchal household relationships can determine who migrates and who stays behind. These interventions mirrored the ways in which geographical research from the 1980s to the early 2000s was marked to some degree by a shift of focus from migrant labor as a structural concept to a growing concern for labor migrants – for workers’ own experiences, the microscaled politics and connections between their workplaces and their households, and their strategies of navigating conditions that might be deemed outside their control.

In this same vein, overly economistic accounts of the motivations, agency, and experiences of intranational migrants have received much needed correctives through research seeking to understand the cultural politics of migration.
Gidwani and Sivaramakrishnan (2003), for example, offer a nuanced account of how Indian intranational migrants’ counterhegemonic expressions of consumption—such as changing dress habits—have served in limited but meaningful ways to challenge entrenched social hierarchies of caste and class. Meanwhile, in her interviews with migrants returning to the South African settlement of Cancele for the Christmas holidays, Ngwane (2003) explores the household as a space that is performed, displayed, and talked into being through migrants’ aspirations of upward social mobility. Through these activities, categories such as “family,” “tradition,” and “home” become unstable sites under discursive construction and contestation through the movement of people back and forth between the city and the countryside. This frame serves to, in Ngwane’s words, “collapse some basic binaries that ... scholars had taken for granted between, for example, ‘family and factory,’ women and migrant labor, and town and countryside” (2003, 688). This research highlights the ways in which migrants’ imaginings and performances of alternative modernities—ones with the potential to perforate both dominant local power relations as well as the modernization scripts that can imbue Eurocentric migration research—demonstrate the inseparable and mutable character of social positions such as citizenship, race, gender, or ethnicity; how they can be produced and transformed through work and migrants’ movements across borders; and how their mobilization affects both migrants’ experiences of employment and their prospects for fair and decent work. Developed alongside the emergence of the body as a crucial scale of analysis in geography, this literature has also challenged the Marxian notions of “embodiment” that imbued geographical scholarship on migrant labor, which predominantly conceptualized it through the physical work of a universalized, waged (male) laborer. Gendered, racialized, and otherwise “embodied” forms of migrant work—for example, jobs requiring “service with a smile,” love and caring, tough manual labor, or sexual services—are geographical in very intimate ways, as they require the mobilization and valorization of particular bodily attributes, dispositions, biological characteristics, or other purportedly immanent qualities of migrant women and men.

Since the beginning of the twenty-first century, a wealth of new empirical and theoretical territory has been covered in geography on the
topic of migrant labor. Five strands of inquiry and debate are highlighted in what follows, not because they offer an exhaustive view of new work on the topic, but because together they show the diversity and depth of recent scholarship. The first has been an ever-deepening set of engagements with migrant households. In their case study of Singapore, Yeoh, Huang, and Willis (2000) demonstrate how the very production of the global city rests on the naturalization of ideas about the Asian family and women’s work within the household, pointing out that the city’s economy is sustained in part through the unpaid work of both the wives of Singaporean men working abroad and the trailing spouses of migrants in Singapore itself. This literature illustrates the silences in this scholarship on the theoretical place of the household in the global city, and challenges the broader ways that migrant labor has been overwhelmingly constructed as waged labor in both migration and urban theory. Research on migrant households has also been a source of innovative developments in ethnographic, collaborative, and multisited research methodologies. In this vein, Mattingly (2001) has shown how the commodification of certain households that employ migrant workers has involved the mobilization of unpaid household work by extended family members to sustain migrants’ own social reproductive needs. This work illustrates the multinodal webs of social reproductive labor required to sustain the commodification of a single household, and highlights the racialized and gendered power differentials between working women that operate within these networks. Ethnographic studies have also been integral in further challenging notions about the migrant household as a cohesive unit or a site with a cohering or equitable set of interests, and in demonstrating the important role that factors such as migrants’ age, gender, or emotional family attachments play in determining how much of their earnings migrants tend to remit back home. Collaborative and performative research, meanwhile, has starkly illuminated the profound and long-lasting social cleavages that can result from the separation of transnational care workers from their families back home (Pratt 2012).

The beginning of the twenty-first century has also been marked by a reinvigorated engagement with the role of the state in the sociospatial regulation of migrant labor. This is in part due to a number of major crisis events that galvanized efforts in geography to delve more deeply into questions about the production of citizenship and new border-making practices aimed at migrants. First, the profound new policy directives involving the securitization of space following the September 11 attacks have had major impacts on the regulation of migrant labor in many countries; in the United States and elsewhere, new disciplinary and surveillant activities by states, adopted under the aegis of homeland security have since become imbricated with agendas of identifying and rooting out “unauthorized” migrants. Second, the 2008 global financial crisis illuminated the multiple vectors of risk and vulnerability that transect migrant workers’ livelihoods; many migrants were not only overrepresented in the sectors hardest hit by the crisis – construction, hospitality, manufacturing, and finance – but they also tended to be employed in jobs in these sectors that were the most casualized and insecure, and were thus most vulnerable to retrenchment. The financial crisis spurred geographers to point to the ways that employing migrant labor has become a significant strategy of offsetting the novel risks imposed by globally interlinked and financialized economies in the twenty-first century. At the same time, research has also foregrounded innovative forms of labor organizing and cross-border solidarities that
Migrant workers and their allies are developing in the face of the growing destabilization and criminalization of migrant work and employment. The recent successes of upcaled worker solidarity movements such as the International Domestic Workers’ Network, the urban-scaled, multioccupational politics of the London Living Wage campaign or the large, yet entirely informal, strikes waged by unorganized construction migrants in Arab Gulf countries show that migrants, migrants’ rights groups, and others have adopted a variety of strategies to confront these challenges. These forms of organizing, moreover, mobilize a diversity of scaled relationships that cross a variety of political, occupational, and sectoral borders and boundaries.

There have also been a number of recent efforts to explore how the discourses, institutions, and day-to-day practices associated with the territorial regulation of migrant labor are integral to the contemporary construction of nationhood, state sovereignty, and citizenship. While poststructuralist feminist scholarship has illuminated how the very term “migrant” both refracts and reproduces the notion of a national state and a “citizen” subject (Yeoh and Huang 1999), there are more recent Agambenian engagements with new ways that migrant workers are deemed “exceptional” subjects who face both the de facto and de jure withholding of welfare, civil, and labor rights. In these readings, it is not simply that migrant workers exist outside the institutions of the state; rather, ascribing exceptionality to migrant labor is fundamental to the ways in which territorial sovereignty is maintained. The importance of producing migrant workers as exceptional subjects is particularly visible, though not unique, in Gulf Cooperation Council states such as Qatar, where all migrant workers are excluded from state welfare entitlements and must also pay a yearly fee to a national citizen sponsor, who in turn grants them the right to work temporarily in the country. This particular form of migrant labor rentierism is a key financial contract that the state has made with national citizens, and it is fundamental to the ways in which the legitimacy of the non-democratic state is reproduced. These financial and regulatory architectures of exceptionality in which migrants are embedded are thus integral to the entitlements and obligations encoding national citizenship. Other research on the role of the state in regulating workers has made key strides in illuminating the gendered social norms that encode particular scaled representations in migration geography. Work by Silvey (2004) and Fan (2004), for example, has been integral in gendering the nation-state by showing how labor-sending and destination country governments reproduce traditional patriarchal gender norms in excluding migrant women employed in the “private” space of the home from the application of host country labor laws.

Third, in a shift from migrant labor research that has been overwhelmingly focused on work in lower-skilled, low-status, or precarious forms of migrant work, there has been a growing focus in recent years on examining elite forms of power and privilege within migrant labor markets. Ley (2010), for example, has traced the multisited migrations among wealthy East Asian migrants to Europe, North America, New Zealand, and Australia in the 1990s, pointing to the ability of this group of migrants to transcend the interests of both host and destination states to retain them. Additionally, there has also been a growing focus on mapping how white privilege and constructions of whiteness operate within migrant labor markets. Notable among these efforts have also been postcolonial, anticolonial, and critical race analyses of migrant labor. Andrucki (2013), for example, traces the colonial institutional structures that continue to maintain racial privilege among
white South Africans who migrate. Other work has illuminated how the entangled hierarchies of whiteness and nationality encode the relative status of migrant sex workers in the United Arab Emirates (Mahdavi 2013), while scholarship on migrant nurses in the United Kingdom has complicated homogenizing narratives about “Third World” migrant women’s incorporation into “First World” households by showing migrant care work takes place in an array of sites and encompasses a diverse range of skill sets, working conditions, and occupations (Kofman and Raghuram 2006).

A fourth theme revolves around debates about the costs and benefits of circular migration, in which migrants oscillate between their home and host destination(s) periodically. This form of migration – which can include, for example, repeat migration through temporary foreign worker schemes which at once facilitate the movement of workers over borders while explicitly precluding permanent settlement – has been described as a “triple-win” scenario by some who argue that it allows host countries to address labor shortages, generates remittances for migrants’ home economy, and reduces “brain drain” in labor-sending countries as migrants are expected to ultimately return home to settle permanently, bringing with them new skills which will enable them to find better work back home. Critics, however, have voiced significant concerns with immigration policies seeking to encourage circular migration. They argue that such directives feed into xenophobic antiforeigner political discourses and absolve host states from providing social services to migrants who need them. Yet others suggest that policies seeking to foster circular migration help legitimate the myth of labor shortages in many host countries, and provide a justification for the creation of a two-tier labor market in which migrant workers are denied permanent settlement. Recent empirical case studies, moreover, have found that the promise of upward economic mobility through circular migration is not being realized for the vast majority of workers (e.g., Takenoshita 2011).

A fifth and final theme inflecting recent scholarship has been a growing engagement with migrant labor intermediaries. Recent research has illuminated how a diversity of actors, including multinational temporary employment agencies, private remittance-transfer companies, labor-sending country consular staff, and informal moneylenders can play crucial roles in shaping and sustaining migrant labor networks. This work has broadened depictions of the kinship and friend migration chains or “hometown ties” often viewed as underpinning migration networks. It has also illuminated the relationships in the production of economic precariousness among migrants; for example, migration debt can circumscribe poorer migrants’ lives more fully because they frequently pay more to migrate and are more vulnerable to high-interest, predatory forms of moneylending than more affluent migrants. Migrant labor intermediaries can also play a key role in the production of migrant identities in that recruitment agents, consular officials, and others tasked with the management of migrant labor markets often actively participate in constructing racialized and gendered norms around the construction of desirable workers. As a result, they often wield considerable power in employing constructions of what constitutes “good” workers to filter out “unsuitable” candidates. These perspectives provide an important counterpoint to the state-centric literature on the regulation of migrant labor, as well as research that has focused primarily on migrants’ own experiences. In particular, they raise important questions about the territorial, juridical, and practical limits of any single state to protect migrants when narrowly delineated
MIGRANT LABOR

protective frameworks focused on workplace rights fail to account for the grey transnational networks of power that extend well beyond the worksite. Together, these strands of geographical scholarship on migrant labor illustrate the immensely diverse experiences, conditions of work, regulatory frameworks, and migratory channels that can be associated with migrant labor, work, and employment. They also help to form a broader recognition of what may be considered key sites and spaces of labor migration – from home and host country households, to wire-transfer shops, to church basements, to corporate boardrooms – that shape and sustain migrant workers’ lives, livelihoods, and migration trajectories.

SEE ALSO: Gender, work, and employment; Labor market segmentation; Migrant division of labor; Migrant settlement; Migration: internal; Migration: international; Race, work, and employment; Remittances; Rights, labor; Transnationalism

References

Yeoh, Brenda, and Shirlena Huang. 1999. “Spaces at the Margin: Migrant Domestic Workers and the

Further reading

Migrant settlement

Wardlow Friesen
University of Auckland, New Zealand

There are many types of migration, ranging from relatively “permanent” to more short-term, circulatory, and transnational movements that appear to be becoming increasingly important. In discussions of settlement, the focus has almost always remained on longer-term international migrant settlement issues, although it is no longer possible to characterize these inevitably as outcomes of permanent migration, since transnational and other forces are at work. Much of the attention in migrant settlement research has been on “settlement countries” with proactive immigration policies, namely Australia, Canada, New Zealand, and the United States, but many of the same issues are relevant in other countries worldwide.

Arrival, settlement, and integration

The conditions of arrival for international migrants in a new country of settlement are dependent on the immigration and settlement policies of that country. A fundamental issue is the legal status of the movement, since this can have considerable influence on employment and education opportunities, family reunion, and sociocultural integration. In many cases long-term residence may be the outcome of undocumented arrival, and it is likely that these migrants will have little government support for settlement and will be cautious about seeking support from other agencies. Even in countries with proactive immigration policies, there may be minimal support for most migrants, often with the exception of refugees coming via the United Nations High Commissioner for Refugees (UNHCR) quota system.

Longer-term settlement policies vary considerably between countries, and these partly depend on policies on the integration of migrants as well as on the resources available for new migrants. Integration policies and their outcomes, in their many manifestations, ranging from assimilation to multiculturalism, have been the focus of more migration research and discourse than any other topic. Assimilation, implying the complete economic, social, and cultural absorption of a migrant minority into the mainstream host society, is considered an extreme and usually undesirable form of integration, although the term is still commonly used. This process has often been idealized in the metaphor of the American “melting pot,” which represents the objective of American migrant settlement policy since the nineteenth century, and elements of this approach remain part of explicit or implicit policy in many places.

Multiculturalism, in both its descriptive and its prescriptive senses, can be constructed as the antithesis of assimilation. Descriptively, multiculturalism is the presence of a diverse range of ethnic cultures, languages, religions, and so on in a city or country, usually as a result of relatively large-scale immigration from a range of countries. Probably the best-known prescriptive use of multiculturalism at a policy level was that of the Canadian government from the 1970s onward, which involved the active promotion and celebration of multiple
MIGRANT SETTLEMENT

ethnicities within Canadian society and polity. According to Mitchell (2004), multiculturalism serves not only the interests of pluralism and social liberalism, but also the neoliberal project of globalization because of the manner in which it valorizes and commodifies difference. More recently, conservative governments in Canada, as well as in Australia, have abandoned multiculturalism as a formal policy, arguing that a certain number of “shared values” are critical to nationhood and successful migrant settlement.

In geography, generalized models of integration have tended to reject assimilation as a model while at the same time advocating socioeconomic integration with only limited cultural integration so as to maintain the cultures of migrants. Many geographical studies have been undertaken to determine the degree to which migrants have become integrated into host societies, both socioeconomically and culturally. Attempts to determine levels of socioeconomic integration range from measuring the evolving characteristics of individuals or groups of migrants through to the spatiality of migrant settlement, while studies of cultural integration are more focused on cultural maintenance and expression.

Commonly used measures of socioeconomic integration include indicators of income, education, labor market, and health outcomes. Studies of these factors proliferated during geography’s “quantitative era,” and they continue to remain the focus of much qualitative and quantitative analysis. Cultural integration is perhaps more difficult to measure than socioeconomic integration and is often more the domain of sociologists and anthropologists, but it is also regularly studied by geographers. Migration- and population-focused journals such as Population, Space and Place, Journal of Ethnic and Migration Studies, and International Migration regularly publish articles on migrant integration worldwide, and geographers comprise a significant proportion of the authors of these articles.

Earlier studies of migrant integration, often based on the premises of the Chicago School in the United States, were focused on less wealthy migrants whose choices in relation to housing, education, and health care were limited by their socioeconomic position. In these situations settlement is often understood within the context of a relatively slow process of economic integration and social mobility. However, as a result of the increasingly selective immigration systems in settlement countries, a much higher proportion of immigrants are well educated and highly skilled, so there is a much greater degree of choice in terms of residential location and resultant housing and education decisions. For wealthier migrants, settlement is no longer concentrated in poorer central areas of cities, but is more likely to be suburban, as in the “ethnoburb” (Li 1998) model. To some extent, success in these areas has even created negative reactions from host societies, for example stereotypes about the “monster housing” of Asian migrants (Mitchell 2004, 112–113) and the perceived competitive study habits of Asian students.

A common debate about migrant integration revolves around the question as to whether “second-generation migrants” (who are not actually migrants but are often ethnic minorities) have achieved a higher level of integration than their first-generation parents. In the United States, evidence suggests that second-generation groups tend to have achieved higher educational levels and jobs than their parents, though this is more likely to be the case for those of European and Asian descent than for Hispanic populations. Portes (2003) concludes that, on average, the children of migrants have higher aspirations than the children of American-born parents, but there is great variation between groups, with Mexican Americans, the largest second-generation group,
falling well below the average in terms of second-generation aspirations and achievement.

Debates about integration have often been constructed from the perspective of acceptance into the host society, but at least as important is the discourse about migrant experience in relation to racism and discrimination within the host society. The study of race spans many disciplines. Within geography, there is a long tradition of the analysis of spatial segregation, but also of other aspects of the construction of race and racism. Cultural racism is perhaps the most significant form of racism in terms of host–migrant interaction in countries of migrant settlement. Rather than focusing on biological difference, there is a concern for cultural difference, and especially the impact that increasing cultural diversity might have on a dominant, sometimes “national,” culture and society. This may be expressed against individuals through specific acts of discrimination, through media and public stereotyping of cultural groups, or in some cases through institutional racism practiced by governments or other institutions discriminating in terms of employment or other services.

Academic attention to discrimination based on cultural difference and religious affiliation has increased significantly since the events of September 11, 2001, especially in relation to Muslim migrant populations. Within the United States, Muslims have often been targeted by conservative groups as representing a new or restigmatized other. In some countries, such as France and the Netherlands, Muslim migrant populations (and some other groups) have been perceived as acting outside a national ideal of secularism, especially within the education system, for example with the French state banning headscarves and other clothing representative of religious affiliation in schools.

### Migrant identities

As well as acknowledging that there has been a diversification of mobility types and durations, scholars also recognize that migrant identities, whether self-adopted or imposed, are fluid, multilayered, and contingent, and that significant perspectives on these ambiguities include cultural hybridity, transnationalism, and diaspora. Geographers have also been interested in identity politics and in the development of a “politics of difference.”

Broadly used, “hybridity” refers to the transgression and displacement of boundaries related to binaries such as other/same and them/us. When used in relation to race/ethnicity, the transgressed binary may be the “Western us” and stereotyped conceptions of non-Western people. Thus cultural hybridity is an important concept when considering the identities of migrant populations in which there may be a reconfiguration of characteristics and categories based on racial, cultural, linguistic, religious, historical, nationality, and other factors. In many cases this has involved the “return” of postcolonial subjects to the country of their former colonizers, with obvious examples including migration to France, Netherlands, Italy, the United Kingdom, and the United States. An example of this is the development of a hybrid “black British” identity among both migrants of African and African Caribbean heritage settled in the United Kingdom, and their descendants. “Black British” has been taken on as a term of pride, asserting difference but also the right to be British. In this and other cases the assertion of hybridity has been used to claim rights to citizenship in its broadest conception.

Since the 1990s, transnationalism has become one of the predominant paradigms within migration studies, and, although migrant linkages back to home countries and to third countries have
MIGRANT SETTLEMENT

a long history, it appears that the forces of globalization have accelerated transnational phenomena in recent decades. From one perspective, transnationalism may be seen as disrupting “migrant settlement” in that it suggests that many migrants are never settled but are rather in a (potential) state of constant mobility. However, an alternative perspective is that “transnational” is most helpfully used to focus on phenomena that involve continued cross-border mobility, including transnational spaces or channels of mobility, rather than to categorize individuals or groups (Collins 2012). These phenomena impact on (potential) migrants whether they participate in frequent and short-term movements or are largely “moored” in a single locality. In identity terms, these transnational phenomena influence the way in which migrants view their “place,” with the reality that many have dual or multiple citizenships in a legal and/or social sense.

Migrant identities are also manifest in terms of diaspora. The dominant usage of “diaspora” until the late twentieth century was as a term for victimized groups, especially Jewish, driven from their homeland. More recently, studies of diaspora have used the term more broadly, and, while a focus on dispersion has been maintained, a more positive perspective is common in relation to homeland orientation and boundary maintenance between members of a diaspora and others, especially the dominant society (King 2012). The larger diasporas such as Chinese, Indian, and Filipino are increasingly intergenerational and are spread worldwide, but a great diversity of smaller diasporas have developed under conditions of accelerated global mobility resulting from changes in transportation, communication, and global linkages generally. In some cases diasporic forces have resulted in resettlement and return migration, such as in the Indian case, where the Indian government has become proactive in seeking the return of skilled expatriate Indians, especially in the information technology industry.

“Identity politics” is a term often used by geographers and others to refer to the political rights and activities of an oppressed minority, whether that oppression is based on gender, ethnicity, sexual orientation, or other characteristic. In terms of migrant settlement, it relates to the political mobilization of ethnic groups who have been historically oppressed under colonial conditions or within hegemonic “host” societies. An alternative and much contested term is the “politics of difference.” Assimilationists may use the term to justify complete assimilation or, alternatively, total separation, whereas rights-based groups use the term to refer to political movements and forces which recognize the right of minority ethnic groups to social recognition and acceptance based on their shared histories and cultures. The latter perspective may be applied to a critique of governments and other institutions in relation to their policies toward racism and the support of diversity of various kinds. For example, Dunn, Hanna, and Thompson (2001) show how most local governments in Australia practice “assimilatory multiculturalism” rather than “radical multiculturalism” in that they support celebratory ethnic festivals and information provision to foster tolerance for ethnic minorities, but rarely confront racism directly.

Urban spatialities of migrant settlement

In most countries with large-scale immigration, migrants tend to settle disproportionately, at least initially, in larger cities, which are often described as gateway cities. The gateway metaphor suggests that in the longer term migrants will move on to other parts of the country and, while this is sometimes the case, often this does not happen
and migrant groups and their descendants remain concentrated in gateway cities. Ley (2010) has undertaken a detailed study of the way in which Vancouver serves as a gateway to Canada for wealthy Asian migrants, especially from Hong Kong and Taiwan. Other classic examples of gateway cities in settlement countries include Los Angeles, New York, Miami, Toronto, Sydney, and Auckland. Another term that has been applied to cities with large and diverse migrant populations is “cosmopolis.” This may be seen as an ideal for urban futures, in which cities globally embrace diversity, democracy, social justice, and livability, cities which Sandercock (2003) celebrates as “mongrel cities.”

One of the main areas of focus for geographers is the spatiality of migrant settlement within urban areas. The Chicago School proposed one of the original models of migrant settlement which highlighted processes of neighborhood “invasion and succession” as part of broader patterns of urban growth. According to this model, migrants initially settled in rundown central areas of large American cities, but as (and if) they became more established and well-off they tended to “filter” to suburban areas. Migrant groups that did not manage to progress socioeconomically often ended up in what became known as “ghettos,” a term adopted from the European context of Jewish enclaves.

The term “ghetto” is the most pejorative of those used to describe ethnic concentration and describes an area where disadvantaged migrant groups are located, as a result of socioeconomic disadvantage and/or discrimination, and this location is usually involuntary. The use of this term has been hotly debated by geographers (e.g., see Peach 2010), and they have often turned to quantitative measures of segregation to determine what might qualify as a ghetto.

Migrant and minority group segregation has been measured quantitatively by geographers in various ways. The most widely used has been the index of dissimilarity (ID), which measures the distribution of a minority population within an urban subarea in relation to the distribution throughout the city. The ID measures the percentage (sometimes proportion) of a minority group that would have to move to create an even distribution, so an ID of 0 denotes no segregation while 100 denotes total segregation. One weakness of this is that the ID is highly influenced by the size of the minority population within the broader urban population, with smaller groups receiving a higher ID because of random allocation as well as actual segregation. Many other measures have also been used, but according to Johnston, Poulsen, and Forrest (2007) there are just two broad dimensions to be measured: separateness (combining unevenness, isolation, and clustering) and location (combining centralization and concentration). They use these as a basis for a typology consisting of six types of area based on ethnic minority and host society shares of the local population in a census tract (or equivalent). This typology has been applied to cities within six countries (United States, England and Wales, Canada, Australia, and New Zealand) with conclusions including: (i) variable degrees of segregation can be attributed to common processes including discrimination, disadvantage, and self-segregation; (ii) there are significant across-country differences, including African Americans being more segregated than any other group in the United States, and South Asian groups in England and Wales being more segregated than Asian groups in other countries; and (iii) Northern Hemisphere countries are more segregated than Southern Hemisphere countries, which is partly explained by the recency of movements of relatively wealthy migrants to Australia and New Zealand.
MIGRANT SETTLEMENT

There are also a number of more qualitative descriptors applied to migrant clustering. The term “ghetto” has already been mentioned. The broader term “ethnic enclave” may describe areas of disadvantage but can also be applied to areas where an ethnic group (or groups) has had more locational choice based on the presence of coethnics, economic opportunities, and cultural or religious facilities catering to particular ethnicities, but there may also be an element of reduced choice in relation to the cost of living, especially housing. The term “ethnic precinct” is used in similar ways to “ethnic enclave,” although in academic writing it is often used in relation to the economic opportunities which migrant clustering offers for “ethnic entrepreneurialism.” Further, it is sometimes used to assess the political aspects of clustering, both in terms of the (potential) political power that ethnic groups may wield in areas in which they are concentrated, and in terms of the strategic imperatives of political parties.

Another group of descriptors of ethnic concentration is based on specific (or aggregated) ethnicities. The most common is “Chinatown,” which is used to describe usually inner city concentrations of Chinese population and/or business operations (Anderson 1991). This term is used worldwide, with well-known Chinatowns in San Francisco, London, Sydney, Singapore, Johannesburg, New York, and Vancouver, but they are also present in many smaller towns and cities in the Caribbean, Pacific Ocean, and elsewhere. In some cases (e.g., Vancouver), many Chinese have moved to other, more suburban locations but retail and restaurant functions have remained, which are often increasingly dependent on tourism. Similar terms have been used to describe the concentrations of other migrant groups, and these include Little India, Koreatown (or Little Seoul), Little Manila, Japantown, Little Saigon, Little Italy, Greektown, Little Havana, and Little Jamaica, as well as many others. Similar to Chinatowns, these named enclaves have generally originated as areas of significant migrant settlement but have increasingly become areas used in urban place branding. An important aspect of these ethnic concentrations and migrant settlement generally has been the development and evolution of “ethnic cuisines” serving new migrant populations but also reflecting the desire of host populations to “eat the other.”

While the Chicago School postulated that most migrant settlement was in central cities, and most Chinatowns as well, more recently immigration has often been predicated on migrants being admitted on the basis of high levels of education, skills, and investment assets, and one result of this is that migrant residential patterns have changed. Wei Li (1998) coined the term “ethnoburb” to describe the San Gabriel Valley near Los Angeles, where Chinese are the dominant minority population but live alongside other ethnic minorities as well as the white majority population. This illustrates the fact that many relatively wealthy new migrant Chinese have chosen to live in suburban areas, and, while this still represents significant clustering, it also represents choice and advantage rather than the disadvantage implied by earlier models. Ethnoburbs have also been identified in other cities, including London, Vancouver, Sydney, and Auckland, and can be seen as a phenomenon found widely in migrant settlement countries.

Although first proposed by the anthropologist Arjun Appadurai, the concept of “ethnoscape” has been adopted by some geographers to conceptualize the presence of migrants within urban areas. As one of a number of “scapes” resulting from global changes and linkages, an ethnoscape was envisaged as “the landscape of persons who constitute the shifting world in which we live:
tourists, immigrants, refugees, exiles, guest-workers and other moving groups and persons” (Appadurai 1990, 297). Although this definition seems to focus on tangible aspects of mobile persons and the landscapes that they create (neighborhoods, temples, shops, markets, festivals), ethnoscape has also been used to consider the less tangible or the intangible aspects of migrant presence (transnational “nations within states,” political projects, virtual linkages). It thus allows the conceptualization of outcomes of migrant settlement beyond the spatially contingent, to include possibilities of heterolocality, transnationality, and other types of linkages and processes.

Conclusion

The study of migrant settlement has engaged geographers for many years, and the models and paradigms used for this study continue to evolve. The analysis and naming of processes of integration have been a central focus of attention, and, while there has been overall movement from assimilative to pluralistic models, this has not been unidirectional. Constant themes of geographic research on migrant settlement have been migrant integration into institutions of the host society, and the barriers of racism. Studies of migrant identity have multiplied, just as the ways of conceptualizing these have diversified. Emergent paradigms of hybridity, transnationalism, diaspora, and identity politics have now entered the geographical mainstream. Geographers have always been prominent in the study of the spatialities of migrant settlement. Within urban areas spatial clustering of migrants and ethnic minorities have been quantitatively and qualitatively modeled through the use of concepts such as ethnic enclaves, ethnic precincts, ethnoburbs, ethnoscapes, and qualitative descriptors such as Chinatown. While these approaches are shared with other disciplines, in many cases, geographers have played a leading role in their development.

The recent “mobilities turn” and increased attention to transnationalism have reminded us that there are many types of migration besides “permanent” long-term settlement, including a variety of short-term, medium-term, and transnational movements, and these often have implications for longer-term settlement (King 2012). Movements for international education, which have escalated in recent decades, have often been facilitated by the location of diasporic populations and service infrastructures. In some countries without proactive settlement immigration policies (e.g., some countries in Europe), the movement of contract workers has sometimes resulted in longer-term settlement as a result of intermarriage, long-term overstaying and amnesties, and other factors. Settlement countries have increased the range of short- and medium-term migration opportunities related to employment and education, and in some cases these have been supplemented by “transition to residence” schemes. These changes and others generated by accelerating mobility and transnational forces will provide new challenges for geographers studying migrant settlement in the immediate and longer-term future.

SEE ALSO: Chicago School; Diaspora; Ghetto; Historical settlement; Migration: forced; Migration: internal; Migration: international; Multiculturalism; Refugees; Residential segregation; Transnationalism

References


Forced migration may be viewed as a particular type of population movement across or within nation-state borders marked by clandestine processes and irregular status, and tied to highly specific motivations to move where free will is said to be limited. Despite differences in various forms of forced migration – political refugees, environmental refugees, and trafficked persons in particular – they are united by the absence of choice in people’s decisions to leave and, oftentimes, where to move. Forced migration therefore raises several important issues for geographers about mobility, identity, violence, exclusion, and racism.

Important questions following from these issues concern the causes of forced migration, the multifarious ways in which forced migration relates to conflict, violence, racism, how territories receive (or not) forced migrants, and what exactly constitutes force in a person’s experience of mobility. Forced migration is often connected with renewed territorial vulnerabilities associated with conflict, natural disasters and climate change, and transnational crime over the past two decades. Four key groups of forced migrants have gained particular traction in discussions of this phenomenon in the contemporary era: political refugees, environmental refugees, trafficked persons, and those who move because of gendered discrimination. Insights developed in studies of transnationalism, postcolonial theoretical approaches, the meaning of home and methodological issues also present potentially novel insights into the experience and understanding of forced migrants as complex subjects of mobility.

Forced migration in retrospect

The most severe and large-scale episode of forced migration in history was arguably the African slave trade, which removed somewhere between 12 and 30 million Africans from their homelands to regions of North America, South America, and the Middle East. Geographers concerned with this movement have principally addressed two themes: the contemporary manifestation of racism against African Americans, and the reconfiguration of identity and community among slaves and their descendants. Other than this topic, geographers have been concerned more broadly to elucidate geography’s links to empire-building processes in the eighteenth and nineteenth centuries (Livingstone 1992).

The incidence of forced migration has increased starkly since the 1990s. In the pre- and post-World War II period until the early 1990s migrations associated with force were a peripheral concern in geographical discussions of population movements. The Armenian exile from Turkey (1915–1917) and the Jewish exile from Europe during the 1940s inextricably linked early episodes of forced migration with specific racial cleansing projects, most irrevocably genocide within war-torn Europe. The theme of racial cleansing and escape from political persecution were themes that united refugee movements in other regions, particularly in Asia where extensive refugee movements
accompanied Cold War political tensions in Southeast Asia, principally Vietnam. Geographers’ interest in forced migrations associated with the World Wars and the postcolonial conflicts in Asia has generally been limited to refugee settlement and integration in the countries to which they have fled, as well as the diasporic imaginings associated with their relationship to the homeland.

Both of these foci have interested geographers because of their potential to illuminate subjects of broader interest to the discipline, including racialization of and segregation in the city, and national identity and tourism. If geographers’ interest with urban assimilation and multicultural conviviality are not specifically reflective of a concern for refugee and asylum seeker mobilities per se, geographical scholarship on diaspora does take the experience of forced migration as a significant analytic frame in the study of these pre-1990s experiences of forced migration. Tim Coles and Dallen Timothy’s (2004) edited volume on tourism, diaspora, and space is provocative in its focus on diaspora tourism as a lens to explore refugee negotiations with “the homeland” as an imagined space marked by the violent history of their removal from it. Beyond this, little has been surmised about these movements through the geographer’s analytic gaze.

The Rwandan genocide of 1994 led to renewed attention to the subject of forced migration generally and in geography, and heralded the onset of the current era of forced population movements associated with civil war, ethnic cleansing, and racially motivated violence centered on disputed ethno-territorial claims. In the 1990s there were 34 territorial disputes within nation-states, of which most are ongoing at the time of writing, including protracted crises in Somalia and Ethiopia. Indeed, the break-up of the former Yugoslavia and the establishment of the new nation-state of East Timor are the only of these conflicts that have been resolved more or less successfully. They have nonetheless heralded further negotiations around memory and justice in the form of truth and reconciliation commissions (former Yugoslavia) and reconstruction (East Timor). These post-1990 forced migrations have attracted far more attention from geographers than the earlier movements briefly reviewed above.

Migrants who are forced today

Forced population movements tied to political violence, and often associated with ethnic or racial cleansing, have historically been seen as the most numerically significant and urgent expressions of forced migration. This emphasis is reflected in published work in geography which focuses on political refugee and asylum seeker (im)mobilities. Apart from refugee and asylum seeker populations two other key groups of forced migrants have nonetheless gained some traction in recent discussions of forced population movements as they are manifest in the contemporary period. These are environmental refugees and trafficked persons. The issue of gender in forced migration, especially as it relates to female forced migrants has also gained some purchase in recent discussions.

Political refugees and asylum seekers

The fact that most published work in geography on forced migration takes political refugees and asylum seekers as key subjects is no surprise since those writing on these issues are often based in major asylum seeker destination countries, including the United States, Canada, the United Kingdom, and Australia (Crawley 2010; Hyndman 2010; Mountz 2010). Migrants who
are forced by politically volatile situations in their countries of citizenship move because of threats to their life and liberty. These movements are both internal (in the case of internally displaced persons, or IDPs) and international (in the case of refugees and asylum seekers). Asylum seekers are a small, but politically contentious cohort of refugees. The sensitivities their movements generate are in large part explained by the fact that they make claims on destination states for protection (asylum) based on well-grounded fears of persecution should they return to their country of citizenship.

Geographical scholarship in this area has been particularly concerned to elucidate the ways these claims are productive of particular discourses and infrastructures of accommodation or refusal within destination states. Bordering processes concerned with the deflection of asylum claimants are spatialized in ways that aim to contain asylum seekers either onshore (principally through detention centers) or offshore (principally through offshore processing and detention centers). This manipulation of territory involves re-drawing the boundaries of a nation-state in jurisdictional terms (for example, by inscribing new functions to remote territories, such as Australia’s Christmas Island where a processing center for asylum claimants was established in 2006) or “renting” the territory of contiguous but independent nation-states to accommodate refugee claimants (for example, by building detention centers and offshore processing centers in neighboring countries in return for the conferral of development aid or other material benefits to these third states). Australia is exemplary in its execution of such political-territorial rezoning, and as recently as late 2013 added yet another element to these tactics by introducing the controversial “no boats” policy. In this policy the Australian government turns boats containing asylum seekers back to their Indonesian point of launch. This policy is achieved through active bilateral cooperation between the Indonesian and Australian governments, under which the Indonesian coast guard will intercept returning boats and detain asylum seekers in one of nine purpose built detention centers in Indonesia until their claims are heard by the UNHCR. What is most striking about these processes of containment, classification, and removal, however, is the way in which they themselves can present a further episode in the forced migration and human rights abuses of those seeking protection. Suicides, hunger strikes, and other dramatic embodied expressions of desperation and depression only highlight what human rights activists have been continuously documenting since these geographies of containment institutionalized through the infrastructure of detention became routinely invoked a decade ago.

Asylum seeker mobilities also raise important questions about the journey itself. Tracing asylum seeker trajectories and routes allows us to explore the geographies of transit, as well as arrival and reception in the destination. Michael Collyer (2007) has studied African asylum seekers in transit in North Africa, arguing that transit points themselves are inscribed with new meanings, communities, and landscapes as those who seek onward journeys become “stuck” at these staging points, often for years. The same trajectories of (im)mobility are evident in the Asian region, particularly as Indonesia and Malaysia navigate the intractable problem of large arrivals of refugees with the inability to secure onward passage. There is some evidence in both contexts of the widespread abuse and exploitation of these vulnerable migrants. Malaysian authorities, including police and immigration officers, were exposed transporting Rohingya refugees from Myanmar across the Malaysian–Thai border where they were sold into slavery on Thai fishing
vessels. In an obfuscation of vulnerabilities, these refugees also became trafficked persons. With the establishment of nine refugee detention centers in Indonesia, widespread detainment and abuses of unaccompanied refugee children have been documented recently. They invoke multiple geographies of violence, bare life, and spaces of exception.

Environmental refugees

Graeme Hugo’s (1996) seminal article on environmental refugees initiated what is now a substantial body of geographical scholarship concerning the vexed relationship between human mobility and environmental crises. In that original article Hugo argued for a distinction between migrants who are truly compelled to move (as in political refugees) and what he termed “environmental migrants” whose migration is associated with environmental problems, but lacks this same “overt force.” This definitional fixity is nonetheless unsatisfactory given the complexity of natural and human-induced disasters associated with climate change and the contextual variability of disasters. The degree of force that propels these environmental migrants to seek new homes can vary greatly and begs geographers to attend to the nuanced landscapes in which disasters and their relation to migration are to be best understood. Thus, geographers pursuing this line of inquiry have tended to focus on predicting the effects of climate change on migration by examining the impacts of environmental problems on migration (Reuveny 2007). Reuveny’s own efforts at predictability modeling have foregrounded the extent of environmental problems and migration capabilities as key factors determining the propensity of populations to move. Not surprisingly, in the developing world the propensity to move is the greatest because of poor capabilities to respond and the absence of systems that enable resilience in affected populations.

Myers, Slack, and Singelmann (2008) extended the possibilities for thinking spatially about environmentally induced population movements to wrest it from the “third worldism” that appears to slip rather unproblematically into much of the current discussion. They suggest that place-based social vulnerability to environmental crises as well as post-disaster mitigation are certainly regionally variable, but that this variability is directly related to the extent of social vulnerability of the population. Examining Hurricane Katrina in the United States as one example, they found that socioeconomically disadvantaged groups were more likely to suffer severe damage to their homes and property and therefore their propensity to outmigrate far outweighed that of other socioeconomic groups.

Trafficked persons

Trafficked persons are defined by the United Nations under the Trafficking Protocol of 2000 as those who are deployed in exploitative laboring situations (apart from those trafficked for organ removal) by the use of force, deception, or fraud in the recruitment process, and where there are often networks of connected actors (traffickers) working in concert to transport, harbor, and deploy a person into this exploitative laboring arrangement. In this sense some trafficked persons may in fact move willingly and voluntarily with their recruiter/trafficker, believing they are to be provided with legitimate job opportunities in the destination. This marks these subjects as significantly different from other forced migrants whose propulsion outwards is based on fear and discrimination in the home country, rather than a desire for economic betterment. Although some victims of trafficking are abducted or kidnapped (and so forced at the point of origin), by far the vast majority are willing migrants.
Unlike with political and environmental refugees, published work on human trafficking in geography is almost entirely absent from the discipline. Nina Laurie et al.’s (2011) work with Nepalese women and girls trafficked across the border into India for prostitution is an important exception to this general malaise. Their work foregrounds the way issues of citizenship, gender and moral geographies intersect in the figure of the sex trafficked Nepalese girl/woman. When these victims return to Nepal they are often met with societal and familial discrimination, institutionalized via exclusion from Nepalese citizenship, which is always granted to a woman by a male relative. These moral geographies of exclusion from the sociopolitical body are extreme responses to sexual transgressions, but other research has demonstrated the ways return from trafficking is productive of “moral geographies” of marginalization (i.e., where morality is spatialized in ways that produce exclusions from the social and political body), especially for women and girls sex trafficked abroad (Yea 2014).

Gender in forced migration

In these studies of trafficking, gender intersects with forced migration through moral lenses in the context of return and with gendered and class-based discrimination in producing vulnerabilities in the migration process. The Filipina entertainers in Yea’s ethnography in Korea, and the sex trafficked women and girls in Laurie et al.’s study are positioned not only through force in their migration trajectories, but they must negotiate their post-trafficking lives in ways that are yoked to their prior trafficking experiences. Nonetheless, the interstices between gender and forced migration remain a peripheral research topic in geography. This intersection offers many possibilities for feminist and migration geographers. Thus, for example, while the issue of female genital mutilation has been the subject of extensive discussion in the context of sexual health and in international law, migration driven by fear of female circumcision is a subject entirely absent in geography.

Reflecting on the groups of forced migrants, there are at least two areas of convergence in the experience of forced migration, particularly as it affects refugees/asylum seekers and trafficked persons. The first of these relates to the efforts of receiving/destination states to classify these populations into those with legitimate claims for protection and those whose experiences are deemed unworthy of redress under human rights norms. There is, in other words, an emerging hierarchy of legitimacy among forced migrants, which oscillates around the thorny question of who counts as a “victim” worthy of protection under state norms and international obligations. Geographers who have documented these sorting and labeling processes have found them anything but neutral, with gendered, ageist, and racial stereotypes predisposing authorities to particular interpretations of the worthiness of any claim being made. Crawley’s (2010) study of unaccompanied asylum seeker children in the United Kingdom, for example, exposes the ways these children’s applications for asylum are prejudiced by their exertion of “too much agency”; behavior seen as inappropriate for children, especially traumatized refugee children.

The second theme that unites the experiences of asylum seekers/refugees and trafficked persons is their relationship to third parties involved in the business of people moving. This is a fascinating area of inquiry for geographers given the intersecting issues of clandestine and illicit movement, the role of social networks, and the ways intermediaries – whether people smugglers or human traffickers – can act to compound the precariousness of vulnerable and forced migrants. What is often overlooked in the imposition of such rigid binaries between
MIGRATION: FORCED

traffickers/smugglers and those they move is that many of these individuals are driven not by criminal intent, but by concern for the welfare of the persons they move; assistance is often a muted subject when traffickers and smugglers are positioned as criminals profiting from the vulnerability of others. A fascinating book length account, *The People Smuggler* by Robin de Creispigny, of the first people smugglers convicted under Australia’s antismuggling law, draws out just such complexities. Iraqi national Ail Al Jenabi was variously a refugee, an asylum seeker, and finally an unwitting people smuggler before his arrest in Thailand and extradition to Australia to stand trial. The extraordinary account of his life emphasizes what is lost in much research in geography on both refugees and, more particularly, trafficked persons; namely, detailed accounts of their lives.

The question of force and the proclivity to intervene

One of the key undercurrents of much academic treatise on forced migration concerns the debate about what exactly constitutes force in a person’s migration trajectory. Are there, for example, degrees of force in people’s decisions to move, or should force be defined in a binary opposition to freely motivated movements? And if so, what should define “free movement” as opposed to compulsion? Geographers have been involved in debating these questions, with Russell King (2002, 89) arguing more than a decade ago in relation to Europe that previously dominant binaries in migration studies – internal versus international, forced versus voluntary, temporary versus permanent, legal versus illegal – blur as both “the motivations and modalities of migration become much more diverse.” In relation to the forced–free dichotomy both King and others have recognized the need for changing discursive conventions when speaking about the relationship between migration and force.

To date related disciplines – particularly anthropology – have embraced this task to a greater extent than have geographers, because of the methodological complements that ethnography yields to understanding the nuanced complexities in people’s experiences of mobility. Conceptual developments in cognate disciplines provide some clarity to discussions of the relationship between decisions made voluntarily or through force. Anthropologist Larissa Sandy (2009), for example, developed the particularly useful concept of “constrained choice” in her treatise on the decisions of Cambodian women to enter prostitution. According to this approach, women’s decisions result from a combination of deliberate and careful weighing up of the options available to them, including an appreciation of the fact that while most of these women are not forced to enter prostitution because of physical threats to their life or liberty, they consider themselves to have “no real alternative” open to them. This more complicated and nuanced understanding of agency and choice in forced migration is useful for geographers to further explore since it unbinds the idea of force from the violence of physical compulsion and enables the extension of discussions of force to compulsion felt by the subjects of these precarious movements in circumstances of heightened gendered and economic insecurity and vulnerability.

The other key issue that has dominated academic engagements with forced migration in the past decade concerns the question of interventions. Like development, forced migration is principally a field of humanitarian intervention characterized in large part by the application of high modernist rationality to technocratic solutions. The journals *Disasters* and *Forced Migration Review* both express this emphasis on
interventions and on-the-ground documentation of refugee situations, best practice, and the logistics of delivering effective relief and longer term development aid to forcibly displaced populations within their own borders and internationally.

The emphasis on interventions itself generates a further subdisciplinary focus on identification and mapping of refugees. Crawley’s (2010) research with unaccompanied asylum seeker children reveals how, despite the aura of transparency surrounding the bureaucratic procedures for sorting legitimate from illegitimate claimants, preconceptions of appropriate subjects and experiences of violence predetermine the outcome of claims. Jennifer Hyndman (2010) makes a similar point in relation to her examination of the differential treatment of two child soldiers, one of whom received international acclaim through the re-telling of his experience, and the other who was remanded in the notorious Guantanamo Bay Prison. In the reception of forced migrants, geopolitical imagination influences the way that law, protection, and rights are applied.

New directions for geographers studying forced migration

Many of the scholarly directions mapped by geographers to provide novel insights into the experiences of other migrants should arguably also inform discussions of forced and vulnerable migrants. There are four particularly significant directions, namely: the possibilities offered by transnationals; the value of postcolonialism as a theoretical lens; the meaning of home; and methodological impasses.

Geographers view transnationalism as a practice of connecting across borders that applies principally to the experiences of successful – and usually legal – migrants. Although geographers now clearly appreciate the importance of the dynamics of transnational practices as they differ – or perhaps overlap – between middling, elite, and lowly positioned migrants, illicit population movements have received little attention in these discussions. Al-Ali, Black, and Koser’s (2001) work is a notable exception, with a focus on transnational practices of refugee populations as the move to Europe. They identify “limits to transnationalism” within geography, which are visible through the absence of scholarly attention on refugees and asylum seekers. They focus on the empirical cases of Bosnian refugees in the United Kingdom and the Netherlands, and Eritrean refugees in the United Kingdom and Germany in identifying the role of history and social and political factors in explaining highly variable modes of transnational life. Several UK-based geographers have explored the relationship between political activism and transnationalism among Islamic state refugees. Whether political refugees engage in the politics of the homeland from the “safe space” of the United Kingdom is a question that opens up new insights into the modes and content of transnational relations. The focus on transnationalism within the experience of forced migration could be further developed to include emotional geographies of shame, stigma, failed migration, family reunification, and the anxieties generated by the simple technological impossibility of maintaining communication with family across borders.

Following the above point, the meaning of home has proved a significant concern in advancing the geographies of transnationalism. Geographies of home are, however, not just the remit of long settled migrants, but also figure in the mobility experiences of refugees who remain in limbo in refugee camps or as noncitizens in neighboring territories. These populations in limbo are sustained by the single hope of
return to the homeland. Home thus operates as a powerful territorial imaginary among both political and environmental refugees. For trafficked persons though, as both Laurie et al. (2011) and Yea’s (2014) work suggest, home may be a site filled with anxiety marked by further vulnerabilities. Events in the home may present new spatialities of shame, stigma, and marginalization, including return to the home country, without necessarily involving return to the home community. Nor should such concerns be restricted to trafficked women, with men trafficked for forced labor, for example, also subject to the scrutinizing gaze of home communities. The Bangladeshi men in Yea’s (2014) study of migrant workmen in Singapore reveals that labor trafficking and the anxieties of return home are equally felt for male victims as they are for women. The interest of geographers in return migration more generally could thus be unfettered from its location in the experiences of voluntary and “free” migrants and usefully expanded to include the complex and multiple homeward trajectories and imaginings of forced migrants.

The sociospatial practice of “othering” marks a representational stumbling block in enabling the emergence of more progressive popular and statist views of forced migration that cries out for further elaboration. For both the Australian and United Kingdom contexts geographers have documented the ways asylum seeker incorporation into society is hindered by the syndrome of NIMBY-ism (or Not In My Backyard) (Hubbard 2005). While these spatialities at the societal level reinforce institutionalized modes of exclusion operating within the state, their location within discourses of the racially and politically marked other are areas not examined in particular detail by geographers to date. Geographers are well positioned to advance such a project because of attention to the spatialities of marginalization that have been documented in relation to other excluded and stigmatized groups in society, such as Roma and prostitutes. To these existing popular and statist discourses asylum seeker mobilities present additional axes of exclusion that hinge on fear of political extremist and Islamic fundamentalism in particular, as well as economic concerns about refugees taking jobs and posing additional burdens on welfare budgets already stretched in the current era of austerity.

Forced migration presents many research challenges for geographers, especially those interested in the experiences and theoretical significance of groups such as trafficked persons or refugee camp occupants, for example, or subjects, such as the meaning of home, life in transit, and transnational relations. We may begin to understand these research predilections in the context of concerns about methodology and ethics. Methodologically, studying the “everyday” lives of forced migrants, including their transnational (dis)connections, their embodied experiences of living in transit, or their trajectories of return home, require novel and creative approaches. Yea’s (2014) study of trafficked Filipina entertainers in Korea, for example, required a sustained ethnographic engagement that involved living with many of these women in Korea over the course of one and a half years. This, she felt was necessary to disrupt the stereotypes associated with this form of forced migration that limited the insights gained in other studies on sex trafficking. Ethnography and participant observation are certainly valuable approaches to research that could produce novel insights into the experiences of forced migration; insights that may disrupt a violent, episodic framing in favor of a recognition that forced migrants also have every day, nuanced lives and aspirations. But such approaches are logistically hard and on occasion dangerous. Moreover, forced migrants – because of the clandestine nature of their movements – are often difficult to locate and/or unwilling to
share their experiences lest disclosure predispose them to further vulnerabilities in migration. Ethics problems also hamper research in this area, as it does with any vulnerable research participants. But ethical problems normally encountered by participants considered “at risk” by university ethics committees are further compounded in the case of forced migrants because of the ways their mobilities can challenge regimes of legality. These troubling methodological and ethical concerns need geographers who have studied the lives of forced migrants in-depth to reflect on these concerns in published academic forums. Pathways through this research impasse can only benefit this increasingly significant area of geographical inquiry and lived experience.

SEE ALSO: Borders, boundaries, and borderlands; Ethnicity; Migration: internal; Migration: international; Race and racism; Refugees; Transnationalism; Vulnerability

References


Internal migration refers to the movement of people within the boundaries of a nation-state. The term covers a multiplicity of different kinds of migration, including rural to urban migration, rural to rural migration, intraurban migration, displacement due to civil strife or environmental hazards, movement that is voluntary or forced, and movements that are permanent or circular. Studies of internal migration share many features of international migration research, including substantive themes, theoretical frameworks, and policy concerns, but have tended to fade into the background in migration work more generally. This is surprising, given the numerical importance of internal migration: while estimates suggest a worldwide figure of 214 million international migrants, this figure is far exceeded by numbers of internal migrants, currently estimated at 740 million, with 100 million internal migrants in China alone.

Internal migration has long been an important feature of human experience across various historical periods, but recent evidence suggests that it is on the rise, facilitated by better communication and transport links, and brought about as people move in response to the inequitable spatial distribution of resources, services, and opportunities, or to escape violence, natural disasters, or the increasing number of extreme weather events. Given that overall rates of fertility and mortality change quite slowly, internal migration may be the most volatile dimension of population change within a country, transforming settlement patterns and reinforcing or ameliorating geographically uneven development. As rapid increases of population in destination areas affect competition for food, housing, and other resources, and in many contexts exceed the provision of these, internal migration carries significant implications for human wellbeing, social justice, and equitable governance. At the same time, policies that facilitate internal migration are justified in terms of its development potential – as a strategy enabling low-income households to diversify incomes and mitigate livelihood risks.

The significance and extent of contemporary internal migration

The scale of contemporary internal migration underlines its importance as a fundamental process of social and demographic change. A recent estimate from the United Nations Development Program, based on national census data, states that around the world there were 740 million people living outside their region of birth in 2005, while further analysis using additional national datasets revised this figure upward to 763 million people (UNDP 2009). Ambiguities in the figures for internal migration reflect the fact that there is no single source of data capturing migration within countries, and variations in data collection methods make comparisons problematic. There are differences in the types of data collected (e.g., events or transitions), the intervals over which migration is measured (e.g., one year, five years, since birth, or latest move, regardless of the timing), and the distance over...
MIGRATION: INTERNAL

which internal migration, compared to other forms of local mobility, is defined.

In a recent study, demographers Bell and Charles-Edwards (2013) attempted to iron out some of these ambiguities in order to draw international comparisons of five-year and lifetime course internal migration. They concluded that the most mobile regions of the world are North America, Australia and New Zealand, and Europe, especially in terms of lifetime migration. This reflects an interplay between relatively inexpensive housing, flexible housing and financial markets, limited government regulation of housing markets, and long-term historical forces in so-called white settler societies. However, the researchers also note a decline in the intensity of internal migration in these regions in the 20 years between 1990 and 2010. This pattern runs counter to prevailing narratives of hypermobility and dislocation, but is seen as reflecting an aging population, increased rates of homeownership, and the increased number of dual-earner couples with a propensity to geographical fixity.

Compared to findings from North America, Australia, and New Zealand, Bell and Charles-Edwards (2013) report lower rates and intensities of internal migration in Asia, Africa, and Latin America, although China is noted as an important exception. Here, there has been a substantial increase in interprovincial mobility recorded over the 2000–2010 period. The emergence of China as an industrial giant has given rise to the largest migration the world has ever seen: more than 100 million people have been drawn to the expanding industrial zones of the east from the rural and less developed provinces of the interior. The visibility of internal migration within national statistics in countries such as Vietnam, which also reported an increase, reflects the existence of government policies restricting internal mobility (e.g., household registration systems). Elsewhere, it is likely that figures reported in quantitative studies of this kind severely underreport internal migration. Understandings of past and contemporary patterns of internal migration are best gleaned from case studies and smaller surveys, as nationally available census data is limited to interprovincial mobility. National data of this kind obscures the most common types of movement, including rural–urban migration and intraprovincial rural to rural migration, and movements of durations shorter than five years, none of which are easily captured by conventional population measures.

Historical contexts

While data of this kind suggest that internal migration is a relatively new phenomenon, this kind of mobility is a long-established facet of people’s lives throughout the world. However, histories of internal migration vary across different places, making generalizations awkward and unhelpful. Internal migration figured, for example, in response to events such as the enclosure of the commons and dispossession of the English peasantry, in movement from the countryside to burgeoning eighteenth- and nineteenth-century cities during the Industrial Revolution in Europe, and in the migration from the deindustrializing north of England to London and the southeast in the 1980s as neoliberalism took hold in the United Kingdom. In the United States, key internal migration streams have included the westward movement from the east coast of poor and working people in the mid-nineteenth century, the mass migration of African Americans from the rural South to Northern cities in the 1920s, and migration to California to escape drought and indebtedness during the Great Depression in the 1930s.
In regions such as Asia, Africa, and Latin America, internal migration has also followed particular historical phases, including, for example, migration for work on colonial plantations and trading posts, rural to urban migration following rapid industrialization and declining returns to agriculture from the 1980s onward in much of Asia and Africa, not to mention forms of displacement linked to infrastructure development, environmental protection, war, and conflict. For much of the Global South, internal migration is part of people’s lives, often forming part of a strategy in tandem with natural resource-based livelihoods, and in response to pronounced seasonality in monsoonal and semiarid regions. Examples include swidden cultivation systems, where soil fertility is managed through forest clearance, cultivation, and extended fallows coupled with the movement of settlements, and pastoralism in semiarid areas, where resources are managed by the movement of livestock. Migration in search of labor opportunities in other rural areas or in towns and cities is a feature of pronounced seasonality in sedentary farming in Asia and also in African farming systems, where household members depart to find temporary work to cover the slack season.

In many places, this is a pattern with a long history: in rural Vietnam during the colonial period as much as two-thirds of the peasantry were involved in circular movements between rural areas during the wet rice transplanting and harvesting seasons. The “rice floods” in the Mekong Delta are strongly linked to seasonal migration to cities and urban centers. Internal migration is also associated with various crop booms (e.g., oil palm in Sarawak, cocoa in Sulawesi, Indonesia, shrimp farming in Thailand, and coffee cultivation in Vietnam), which have encouraged rural to rural mobility from the colonial period to the present day. Furthermore, the intensification of internal migration since the 1980s reflects the simultaneous impacts of rapid industrialization and declining terms of trade for agricultural commodities, prompting rural to urban migration across a range of contexts (Rigg 2007). More recently, economic crises, structural adjustment, and the precarious nature of much urban employment have meant even more complex patterns of mobility, including rural–rural and peri-urban–urban connections, now predominate (Tacoli 2009).

Conceptualizing internal migration

Studies of internal migration have been influenced by the same theoretical frameworks that underscore migration theory generally. These include theories that focus on the drivers of migration and theories that focus on how migration is facilitated and reproduced. Aside from emphasizing the material dimensions of movement, theories have been developed to examine the role of power, subjectivity, and identity in internal migration, and the remaking of places through rural–urban interconnections. Theories focusing on the initiation of internal migration have fallen broadly into three camps. The first of these includes behavioral explanations that emphasize rational choice and voluntarism, in that the decision to migrate is seen as made by individuals in response to particular stimuli, such as wage differentials between place of origin and destination (Thadani and Todaro 1979). Critics have pointed to the ways this ignores the social dimensions of migrant decision-making, such as family and community context, and note that the wider political economic context of migration is elided.

By contrast, this latter dimension is given prominence in structural explanations, which emphasize wider forces shaping individual and aggregate migration patterns (e.g., the workings
MIGRATION: INTERNAL

of new international divisions of labor and labor market segmentation under capitalist and neoliberal systems; Sassen-Kooob, 1983). This approach enabled consideration of the gender, racial, and ethnic dimensions of labor migration to be included in analyses, emphasizing the ways in which migrant labor markets are segmented along racial and gendered lines. In Southeast Asia, since the early 1980s the growth of export-oriented industrialization has been closely linked to the migration of women from rural areas to work in urban and peri-urban factories (Mills 1999), while analyses of rural migrants in urban informal sector labor markets highlight the connections between particular jobs and ethnicity in Asian and African contexts.

The third approach has emerged from studies of the new economics of labor (Stark 1991) and was given renewed impetus in sustainable livelihoods approaches, which emphasize household livelihood strategies. Internal migration is part of a wider diversification of income sources across activities such as agriculture, petty enterprise and formal employment shaped by household labor availability, consumption needs, and production possibilities (alternative forms of income) insofar as these are conditioned by wider economic and sociopolitical shifts. Although this perspective has been useful for understanding the factors that prompt migration, it has the added benefit of directing attention to the role of migrant remittances in household livelihoods, and the links between agricultural and nonfarm work (Rigg 2007). As a result, this perspective on migration underpins much of the research and policy work on migration and development, and has been used as a framework by studies that demonstrate the positive ways in which internal migration mitigates environmental and economic risks, shocks, and stresses for low-income households (Tacoli 2009; UNDP 2009). Further nuance is added to this perspective by approaches that look closely at gender and intergenerational dynamics within households, in order to understand the ways in which internal migration is negotiated within the context of family power relations and thus has gender- and age-specific impacts on wellbeing (Resurreccion and Khanh 2006).

Theories that focus on the perpetuation and reproduction of migration explain how current flows can produce further internal migration, for example through the role of social networks that link the origin and destination of migrants. These approaches share much with those developed in relation to international migration, and include both formal and informal institutions. Of the latter, family and kinship are of critical importance in providing practical and material support for those who may follow migrants to other destination areas. Networks focusing on villages of origin are also of critical importance for similar reasons. Both of these kinds of network facilitate circular migration: the regular to and fro movement of people between city and countryside, which has been emblematic of much movement in African and Asian contexts.

Work on migration to forest frontier areas in upland Southeast Asia has noted the ways in which such networks link places of origin and destination, shaping the geographies of migrant settlement in forest areas as newly arriving migrants are able to stay with neighbors and work on their plots before establishing themselves. Informal networks of this kind can become increasingly formalized in efforts to channel and direct the ways in which migrant earnings are used. Elmhirst’s (2002) study of female rural migrants working in Jakarta factories showed how elite, city-based men from the same village of origin created a “village organisation” which took on the role of encouraging them to save, and send money home, with a view to improving life in the village. These networks work in a similar way to the kinds of
hometown associations described in the international migration literature and seen by the World Bank as potential agents for development.

Theories that examine the role of power, subjectivity, and identity in internal migration have drawn attention to the ways in which movement is motivated by matters beyond a simple economic calculus, and how individuals and spaces are transformed through processes of mobility. By focusing on the crafting of “modern” subjectivities, research has drawn attention to the aspirational dimensions, particularly of youth mobility from rural areas, where a desire to shake off a rural identity is manifest in migrants’ narratives. Migration for work or education is part of “becoming somebody” (Crivello 2011). Concepts such as modernity and cosmopolitanism are woven into analyses of how internal migration is integral to the ways people negotiate contrasting subject positions as “modern” or “traditional,” and how it contributes to the achievement of complex social goals linked to dominant cultural discourses about modernity and progress (Mills 1999). Other work that draws on similar bodies of theory explores the links between internal migration, identity, and place-making, as social and economic remittances connecting village and city produce rural villages as socially urban spaces and foster forms of rural cosmopolitanism (Gidwani and Sivaramakrishnan 2003).

Links with international migration

While many textbooks on migration would suggest that debates on internal and international migration have developed in parallel and with little cross-fertilization, migration scholars have suggested that this divide is artificial and unhelpful for three main reasons. First, shifting geopolitical boundaries blur the distinctions between internal and international migration. Europe’s Schengen Agreement has created a borderless zone for mobility within much of Europe, making international moves more akin to internal migration. The collapse of the Soviet Union and of Yugoslavia also turned internal migrants into international migrants or minorities in newly created states. The creation of nation-state boundaries in Africa following decolonization in the 1960s cut across areas where there was once free movement based on ethnic affiliations or nomadic circuits: in some cases these have continued, but in others they have been blocked (King and Skeldon 2010). Elsewhere, forms of graduated citizenship and differentiated entitlement accorded to ethnic minorities also have the effect of producing de facto borders within countries, as the examples of China and Vietnam attest.

Second, individual and aggregate migration trajectories reveal linkages between internal and international migration in important ways. King and Skeldon (2010) describe stepwise migration from internal to international, as rural migrants move first to the city in their country of origin before moving overseas: examples of this kind of flow include Turkey, Thailand, and Mexico. They argue that this kind of sequencing is shaped by the nature of global capital and geographies of labor demand. A second sequencing they identify is where international migration leads on to internal migration, as international migrants move on within destination countries. They cite studies of refugees in the United Kingdom, United States, and Sweden who move to larger cities in their host countries in search of better economic and educational opportunities. Third, King and Skeldon point to the gains to be had from linking theoretical approaches developed independently for internal and international migration. There is evidence that this kind of theoretical cross-pollination is already happening, as both kinds of migration research
are influenced by new developments in social theory, and as there is increasing recognition that the policy implications of migration (whether viewed as a negative or a positive phenomenon) are also shared.

Government policies on internal migration

While international migration is directly shaped by government policies on immigration and citizenship, there are only a few areas of internal migration that are directly controlled in this way. State-sponsored resettlement schemes of the kind found in Indonesia, Vietnam, and Laos involve the relocation of people to state-established resettlement sites from areas where they had been considered “out of place” (in remote or marginal places, in conservation areas, in the way of infrastructure developments). State-sponsored internal migration of this kind is driven by factors such as a desire to facilitate the delivery of services to remote and mobile upland populations and to remove upland swidden cultivators from fragile watershed environments by relocating them in lowland resettlement sites (Elmhirst 2012). In Indonesia, state-sponsored resettlement has also become entwined with corporate interests as transmigration settlements are being established around oil palm investments.

For the most part, however, internal migration is governed and shaped indirectly, through state policies relating to housing tenure and employment, and at a broader scale to planning, investment, and development strategies. Such policies have the effect of opening up and closing down opportunities for would-be internal migrants. In China, the household registration system (hukou) has been used to restrict migration and keep people in the countryside, as those seeking to change residence are required to obtain approval for hukou change from local authorities: possession of the appropriate hukou is a requirement for accessing services such as education and health care (Nguyen and Locke 2014). Critics have suggested that, while the system has not prevented rural–urban migration from taking place, it has created severe inequalities in the entitlements of migrants, leaving them precarious and vulnerable. Although reform of the system has recently been announced, there is skepticism about whether migrants will experience much change, given that reforms are limited to hukou in smaller cities and where there is a risk of corruption and rent-seeking behavior on the part of officials.

Rural–urban migration in Vietnam has been shaped by a similar system designed to restrict migration and, despite some relaxation of institutional controls, they continue to structure qualifications for residency and related social entitlements. While 80% of migrants in Vietnam have some form of temporary registration, fewer than 5% have permanent registration where they work, because they don’t meet the requirements (Nguyen and Locke 2014). Opportunities are also closed down by land-use zoning and restrictive tenure regulations that make it difficult for would-be rural migrants to settle in urban areas. Migrants without officially recognized tenure are easily displaced by processes of urban redevelopment, as has been the case in a number of cities across the world.

Indirect influences on internal migration have included development policies based around import substitution and industrialization, and urban-based investment in government services such as health and education, which created the conditions that encouraged migration to urban areas in search of economic and educational opportunities. In the 1980s and early 1990s structural adjustment and restrictions on government spending had the effect of creating uncertainties in urban incomes, prompting a
stabilization of rural–urban ties, and in some instances urban to rural migration. At an international level, the World Bank has set out a policy approach that is designed to encourage rural–urban migration, with urban agglomeration seen as crucial for fueling economic growth. Particular targets for this approach include registration systems such as those found in Vietnam and China, but other kinds of obstacles to urbanization (lack of transportation infrastructure, etc.) are also seen as important. This approach has drawn strong criticism for failing to consider the social and environmental impacts that follow from these kinds of urban mobility, where precariousness and poverty for some exist alongside privilege and prosperity for others.

Emerging areas of research on internal migration

Internal migration for development

There has been a resurgence of interest, led by scholars working from a sustainable livelihoods perspective, in the ways in which internal migration can foster positive development outcomes for low-income people (UNDP 2009). As part of an overall household livelihood strategy aimed at acquiring assets from a range of sources in order to insure against future shocks and stresses, internal migration is not necessarily a strategy of last resort (Tacoli 2009). This kind of approach highlights the potentially positive role of migration and migrant remittances in fostering development, through the transfer of economic and social remittances. Economic remittances may be used to pay for the education of children, the improvement of housing, and food and health care, thus enhancing wellbeing in areas of origin. Moreover, the migration experience may have fostered capacities for dealing with bureaucracies and figures of authority, thus enhancing household capabilities. There has been a mushrooming of empirical studies that seek to explore this dimension of internal migration and, taken together, emerging evidence suggests that optimism in relation to migration as a development panacea needs to be tempered by a recognition that the outcomes of migration depend heavily on the type of migration, the selectivity of migrants, and the destinations. Moreover, migrant remittances, as a form of bottom-up development assistance, are unlikely to solve deep structural development obstacles such as poor macro-economic policy, questions of social justice and uneven prospects for political participation, insecurity, corruption, and deficient infrastructure.

Internal migration and life course transitions

Scholars have been interested in exploring the link between internal migration and life course transitions of various kinds. Youth migration has provided a strong focus for recent research, reflecting the fact that a large proportion of internal migration is accounted for by the movement of young people, either for education or for employment. Geographies of student mobility have explored the internal migration of young people moving from home to university, focusing on the impact of student migration on the economic and social landscape of university towns in the United Kingdom and other parts of the world (Smith 2009). Some of this work has also explored the onward mobility of students after they graduate, and the link this has with the uneven development of labor markets and social mobility. In developing countries, the focus has been on youth transitions and migration for education or work. The World Bank has estimated that, across 29 developing
countries, young people were found to be 40% more likely than older people to move from rural to urban areas or to move across urban areas. The higher propensity of rural youth to resort to internal migration reflects a lack of employment opportunities in rural areas, but this is coupled with a generally negative perception among young people of agricultural labor, which is associated with low returns, drudgery, and low social status (Elmhirst 2002). Crivello’s (2011) study of internal migration for education in Peru shows how young people and their parents connect migration with the process of “becoming somebody in life” and view it as a strategy for achieving their educational aspirations. This is linked to dependencies between parents and children and, in particular, the roles that children play in mitigating family poverty (Crivello 2011). Rapid increases in internal migration and its potential impacts on wellbeing underpin recent work on marriage and social reproduction in areas with high levels of labor migration, particularly of women. Resurreccion and Khanh’s (2006) work in Hanoi is one of a number of studies that shows how parenting is affected by women’s rural–urban migration in Vietnam, particularly as gender norms limiting men’s domestic responsibilities prove intransient in the face of other kinds of social changes.

**Internal displacement, war, and conflict**

One aspect of internal migration that has received special attention in recent years concerns the displacement of people within a country’s borders, either as a direct or indirect consequence of war and conflict, or as a result of the structural violence of the state and other agents, who forcibly remove people to make way for development projects of various kinds. In relation to war and civil conflict, there are an estimated 33.3 million internally displaced persons (IDPs) worldwide, with the largest populations in Colombia, Iraq, and South Sudan. Conflicts in many other African countries, alongside Sri Lanka, Afghanistan, and Indonesia also have had large IDP populations, although recent increases reflect the escalation of conflicts in Syria and the Central African Republic. Unlike refugees, IDPs are the responsibility of their own government, since they have not crossed international borders, yet it is often this very government that has persecuted and displaced them. Despite the differences in legal status and of entitlement to aid from the international humanitarian community, the causes of displacement and the experience of being displaced are often similar for IDPs and refugees. Much like refugees, IDPs often feel like strangers in their place of refuge, where the local population may be from a different ethnic and/or religious group and/or may speak another language. Consequently, IDPs may not feel welcome, despite sharing the same citizenship as the host population. Identifying the processes that lead to conflict-driven internal displacement means recognizing multilayered causes, including the inequities of globalization and development, coupled with antipathies that have deep historical and cultural roots. Work has suggested that people only move when threats to their safety are perceived to exceed the risks of travel to a new and unfamiliar destination. Unless violence reaches certain levels, people are more likely to confine themselves to the safety of their homes, family networks, and surroundings they know and trust.

Other forms of displacement are associated with structural violence – that is, invisible forms of violence inherent in political and administrative actions that prevent poor people from making a living – and this too has formed a new avenue of research on internal migration. Development-induced displacement is one of the drivers of internal migration in
both urban and rural areas. The construction of spectacular high-rise apartments, shopping malls, and regional corporate headquarters has been a leitmotif of mega-urbanization in cities such as Jakarta, Bangkok, and Manila, and is accompanied by the displacement of the poor and politically marginalized. In addition, aggressive acts of “accumulation through dispossession” are evident in Vietnam – for example, in the establishment of the Thu Thiêm New Urban Zone, which led to the eviction of 14,000 low-income households, the razing of all built construction, and the filling in of marshes and streams to make way for new office space, luxury high-rise apartments, and new public spaces.

In a similar vein, recent research on migration and agrarian change in rural areas has drawn attention to large-scale land acquisitions by the state and the corporate sector for commercial or conservation purposes, which are associated with population displacement, the undermining of viable livelihoods, and subsequently the fostering of new patterns of internal migration. Processes of land acquisition for carbon sequestration, so-called green grabs, have become a feature of parts of Asia and Latin America, while the expansion of large-scale agrofuel plantations in parts of Africa and Southeast Asia are associated with the dispossession of local people, prompting new waves of internal migration as those displaced seek income alternatives elsewhere.

Internal migration, natural disasters, and climate change

Debates about the impact of climate change have focused attention on the linkages between the environment and migration, following alarmist reports that future environmental change will lead to mass displacement of populations from locations vulnerable to climate change effects. Although much of the research prompted by such pronouncements has focused on international migration, a growing body of research is emerging that draws attention to climate change and internal migration. Two key directions are evident in research. The first is work that looks at climate change as one of a series of drivers of migration, and that demonstrates the importance of internal migration as those displaced by climate-related environmental hazards follow already established migrant networks and relationships (Black et al. 2013).

A second thread explores the role of migration as an adaptive, resilience-building strategy, which enables individuals and households to accumulate assets and thus adapt to climate change more effectively. A key dimension of this work has been to encourage policies that support sustainable forms of local migration through better planning, economic diversification, and investment in smaller urban centers (Tacoli 2009). Critically, this work also recognizes the multiple linkages local migration has with various forms of environmental change (from slow-onset droughts to environmental shocks such as floods and typhoons), and thus points to the need for a nuanced, contextual approach that is not only restricted to examining internal migration but considers its connections with all other forms of mobility too.

SEE ALSO: Environment and migration; Informal sector; International division of labor; Labor migration; Livelihoods; Migration: international; Networks, social capital, and development; Remittances

References

Bell, Martin, and Elin Charles-Edwards. 2013. Cross-National Comparisons of Internal Migration: An
MIGRATION: INTERNAL


Migration: international

Valerie Preston
York University, Canada

Brian Ray
University of Ottawa, Canada

International migration, movements of people who leave one country to settle in another where they establish social ties (Migreurop 2013, 13), has stimulated a voluminous and growing geographical literature that examines the volume and diversity of migration flows, their meanings for migrants and the societies where they settle, and representations of migration. Geographical accounts have been particularly important in explaining the increasing spatial variegation and complexity of contemporary international migration and its historical foundations in ongoing processes of colonization and imperialism. Contemporary migration trends include the illusionary binary of permanent and temporary migration, the feminization of migration, the rise of migration induced by environmental change, and the challenges that ethnoracial diversity among migrants poses for immigrant-receiving societies across the world. In addition to efforts to explain international migration, geographical perspectives have been important in examining how representations of international migration and international migrants are linked to state efforts to exert and maintain control over national borders.

International migration is racialized and gendered. The geographies of contemporary migration are rooted in the European colonization of territories in the Americas, Africa, and Asia. The forced migration of colonized peoples contributed to the racialized hierarchies that still underpin many state efforts to control migration. Recent research has examined how gender affects migration, comparing the international movements of men and women; investigating how women’s roles as caregivers affect their decisions to migrate, their choice of destinations, and subsequent links to family and friends; and assessing how migrant women are represented by the public and policymakers (Pratt 2012). The interest in gender has also drawn attention to the heteronormativity of contemporary international migration and migration regimes, highlighting the barriers to international movement for many people who identify as LGBTQ (lesbian, gay, bisexual, trans, or queer) and the emancipatory potential of international migration.

Historical foundations of contemporary migration

Most introductory human geography textbooks deal with migration, often with a map showing population migrations over time. Such a map often gives the impression that the world’s population is on the move — arrows highlight flows to and from different continents and a dizzying number of arrows underline the intricate flows between states within the same continent. But what is new about such movements and why do the short- and long-distance moves of people across borders and boundaries continue to interest so many geographers?
Migration is not new. The movement of *Homo erectus*, and later *Homo sapiens*, out of Africa set in motion processes of social and biological change that shape the human experience to this day. The evolution of our species is tied to acts of migration, making it impossible to summarize human migration succinctly. Nonetheless, there are clearly important antecedents to contemporary migration because past migrations influence current geographical understanding of contemporary international migration, particularly thinking about its causes and impacts.

Empires, from the Holy Roman empire up to the Ottoman empire and the many European empires of the seventeenth to twentieth centuries, necessitated international migration. Maintaining control of far-flung territories required moving troops and administrators, and in many instances European colonial powers also facilitated the movements of their own citizens to exploit resources and deal with population pressures and indigence at home. Colonialism is tightly woven with processes of industrialization and migration. The industrialization of agricultural production in Europe created surplus labor that migrated to the emerging industrial cities and towns of England, Belgium, and the Netherlands to work in factories processing raw materials extracted from colonial territories. Some rural to urban migrants moved on to colonial territories such as the United States, Canada, and Australia. Their jobs were often taken by migrants from Ireland and southern and central Europe who were fleeing desperate economic conditions and political repression. The economic processes and migrations that colonialism encouraged also led to the decimation of aboriginal populations in the Americas, Australia, and New Zealand as European migrants built settler societies. At the same time, millions of Africans were abducted or sold into slavery and transported to work in the mines and plantations of the Caribbean and North and South America.

Slavery and indentured work are important antecedents to modern labor migration, as they illustrate how capitalism exploited and reinforced uneven geographies of resources and power, set in motion the emergence of ethnoracially diverse societies, and relied on involuntary movement. Highly sought-after goods such as sugar, tobacco, coffee, cotton, and gold were produced through slavery for European markets, and the labor of slaves underpinned the economic and political power of colonial states such as Britain, France, Belgium, the Netherlands, Spain, and Portugal from the seventeenth to the nineteenth centuries. It is estimated that, by 1850, 12 million Africans had been taken as slaves to the Americas under conditions of unprecedented brutality. Slavery involved women and men, and, while abysmal working conditions dominated every slave’s life, women’s experiences were also shaped by sexual exploitation. Systems of indentured work essentially continued forced labor practices in the nineteenth and twentieth centuries as migrants from India, China, Japan, and other regions of Asia were recruited to work in mines, plantations, and infrastructure projects across colonial territories. Although indentured laborers were paid, their wages and working conditions were usually dangerous, dirty, and undesirable.

Historically, migration has been highly uneven in terms of participants, motivations, and consequences. In the colonial period, migrants became essential parts of a capitalist economic system characterized by significant international variation in wages and working conditions that continue to this day. Ethnic, cultural, and racial diversity are inherent in the landscapes that migration, colonialism, and capitalism exploit and reproduce. Although migration had been a force in Europe long before the colonial period, and many cities and towns were marked
by cultural, linguistic, and religious diversity, colonialism drew the peoples of Africa, Asia, the Caribbean, and the Americas into a global system of trade, exploitation, and movement. The other became a potent force in shaping social relations and politics in settler societies and in the metropolitan societies that governed them.

**Trends in contemporary migration**

Contemporary migration builds on the uneven geographies laid down in the colonial period. The particulars of migration are structured by specific economic, social, and political histories and, as a result, contemporary migration flows vary geographically. The dynamics and implications of migration in Europe are different from those in North America, and different again from those in East Asia or Africa. Mixed migration flows predominate, with people migrating for many different reasons ranging from a search for better professional opportunities abroad to flight from conflict and environmental degradation. It is impossible to capture fully the diversity of contemporary migration, but four important trends stand out: permanent versus temporary migrations; the feminization of migration; migration caused by environmental degradation; and the challenges of ethnoracial diversity.

Because of the financial, social, and emotional costs associated with migration over long distances, many individuals leave their countries of origin knowing that the move will be relatively permanent. Historically, this was especially true when the financial costs of migration were high relative to income, long ocean voyages had significant risks, and weeks if not months were required to cross great distances. As a consequence, images of migrants leaving country, home, and family behind, never to return, are conveyed through novels, plays, and movies, and are strongly rooted in the popular imagination. Nevertheless, the distinction between temporary and permanent migration has never been straightforward. Historical research indicates that, even in the nineteenth and early-twentieth centuries, migrants maintained ties to their homelands and many viewed migration as a temporary relocation that would provide the resources needed to support a family and acquire land in the homeland. It is almost impossible to quantify the number of people around the world who move permanently and temporarily, in part because of the fluidity of migrants’ decisions. Immigration laws and regulations that permit some migrants to change their status from temporary to permanent also complicate attempts to count the two groups of migrants. Traditional settler societies such as the United States and Canada, as well as many European countries, are now admitting temporary migrants in response to domestic labor shortages. In many countries, there is a growing connection between temporary and permanent immigration systems, especially for skilled labor. For instance, in response to the boom in high-tech industries in the United States in the late 1990s, Congress increased the annual number of temporary worker visas (H-1B) from 65,000 to 195,000 each year from 2001 through 2003. In addition, many temporary admission categories in the United States now waive the requirement that individuals demonstrate intent not to remain permanently. As a consequence, some temporary visa holders can apply for permanent residency without leaving the United States. In this regard, the system in the United States is illustrative of the ever more complicated relationship between “permanent” and “temporary” migration.

The upswing in the number of “temporary” migrants has also been facilitated, if not encouraged, by several states in the global
periphery and semiperiphery which see remittances from individuals working abroad as an increasingly important component of their economies. The efforts of the Philippines government to arrange and certify employment opportunities for its citizens are perhaps the best known. The Philippine Overseas Employment Administration (POEA) manages a “temporary” labor flow of approximately 1 million Filipinos who work in jobs that range from domestic service to engineering and nursing (Tyner 2009). The majority of deployments are for work in the Gulf States and East and Southeast Asia, although many Filipinos also work in Europe and North America. As temporary foreign workers in domestic service, construction, and manufacturing, Filipinos are literally building countries but enjoy few of the rights afforded the national population. The direct involvement of governments in facilitating the temporary migration of their citizens, as well as the alignment of education and work experience with those demanded by employers, are new aspects of international migration. In many ways, the growth of temporary migration on a worldwide basis reflects an emerging agreement between origin and destination countries that are responding to distinct economic and political challenges. In many wealthy countries where permanent migration has become a politically charged issue, temporary migrants are an effective response to labor shortages. As long as “temporary” migrants remain temporary, the concerns about social, cultural, and economic integration and intergenerational social mobility that characterize debates about permanent migration are sidestepped (Khoo, Hugo, and McDonald 2008).

Debates about permanent and temporary migrants frequently overlook the women and men who make international moves. Migration is a gendered process that more women are undertaking on their own rather than accompanying husbands and other family members. Research increasingly highlights the feminization of migration, with women making up almost half of all migrants worldwide (King 2012). Although women’s motivations for migration are as diverse as those of men, a sizable number migrate in response to gendered aspects of everyday life. In many wealthy countries, growing numbers of women have entered the labor force since the 1970s. The feminization of the labor force has created a demand for child and elder care, and help with cooking, cleaning, and the other reproductive tasks needed for household survival (Romero, Preston, and Giles 2014). At the same time, many states have reduced spending on care, expecting family members to provide services that were formerly available publicly. The gap between care needs and the caregiving capacities of family members is often filled by migrant women. Ever more female migrants are also working in physically taxing and poorly paid occupations outside the domestic sphere, such as nursing assistance, food preparation, and clothing manufacturing. The numbers of female migrants leaving countries such as Sri Lanka and Indonesia have increased, with many working in private households as well as formal workplaces. Many of these women are part of global care chains in which they provide the financial resources that allow female relatives to care for their children and aging parents. Simultaneously, their physical and emotional labor ensures the wellbeing of families in the Gulf States, Southeast Asia, Europe, and North America. With strong ties to social reproduction in their home countries and places of settlement, many migrant women are part of two families, two cultures, and two economies, juggling conflicting obligations and loyalties.

Concentrated in Asia and Africa, forced international migration is also gendered. In 2013, the United Nations High Commissioner for
Refugees (UNHCR) estimates, there were 51.2 million forced migrants in the world, the highest number since 1989. Of this total, approximately 17 million were refugees, of which 49% are women (UNHCR 2014). Women experience as much political and social persecution as men, although sometimes in a different manner. For instance, although both genders suffer discrimination and persecution on the basis of sexual orientation or transgender identity, women are particularly vulnerable to sexual violence and gender persecution.

Among the new arenas of migration research, growing attention is being paid to environmental or climate change migrants who move as a consequence of environmental degradation (Black et al. 2011). Predictions about the number of people who may be displaced vary widely—from thousands to millions—but it is clear that climate change is associated with temporary, often local, displacement due to environmental catastrophes such as cyclones and floods, as well as with international migration from places where livelihoods are no longer sustainable. Migration is a long-established coping strategy for dealing with social and economic instability in ecologically vulnerable locations. Although there is general agreement that environmental and climate change migrants should enjoy the human rights guaranteed under international and national laws, existing legal instruments are most effective in dealing with people made homeless by sudden-onset disasters such as earthquakes, flooding, and landslides rather than with those who move in response to the slow-onset disasters and cascading environmental degradation associated with climate change.

One of the great challenges posed by contemporary international migration concerns responses in countries of settlement to ever more culturally, linguistically, socially, and economically diverse migrants. Research is already demonstrating that ethnoracial and linguistic diversity characterizes many cities across the world. Although many states, such as Australia, Canada, Argentina, and England, have long migration histories, the past provides few models for how to achieve integration, equity, and social mobility in contexts of high ethnoracial diversity. The roots of diversity in colonialism, capitalism, and the processes of climate change that do not respect international borders are seldom acknowledged. Contemporary migrants are being met by regulatory environments that increasingly attempt to “manage” who will cross borders as a response to domestic social, economic, and security concerns. Rather than discussing nation-building and the needs of migrants, contemporary discourses about immigration and diversity in countries of reception emphasize domestic economics and politics.

Explaining migration

Many geographers have tried to account for the volume, composition, and impacts of international migration. The “Laws of Migration,” published in 1885, which summarizes and explains migration flows prevailing in Western Europe at the time (Ravenstein 1885), is a seminal analysis that still influences geographical thinking about international migration. Ravenstein (1885) emphasized that migration flows diminish over distance according to a distance–decay relationship, in which the volume of migrants declines with distance from the origin. Geographers have also seized on Ravenstein’s observation that people migrated mainly for economic reasons. In the late nineteenth century, the minority of migrants who moved long distances in Europe were attracted to centers of industry and commerce where the Industrial Revolution created employment.
Geographers paid far less attention to Ravenstein’s observation that women were more likely to migrate than were men. In late-nineteenth-century Europe, young women were often viewed as economic liabilities and sent away to work, so their families no longer had to feed, clothe, and sustain them. As domestic and factory workers, many young women sent money home, enhancing their families’ economic wellbeing, and some even accumulated dowry monies that allowed them to return home for marriage. Social historians have documented migrant women’s vulnerability to economic and sexual exploitation as domestic and factory workers, but they also noted that migration could benefit young women. The independence and autonomy that migrant women sometimes enjoyed in large cities allowed them to forge new identities and lives that were not always sanctioned by their distant families.

Ravenstein was also prescient in recognizing that the national context influenced the volume and composition of migration. He distinguished countries of absorption that attracted more migrants than they produced from countries of dispersion where the opposite situation prevailed. After analyzing migration flows in the United States, he also acknowledged that the economic and political circumstances in each nation-state influenced international migration. For example, the promise of free land for settlers in the United States encouraged migration over longer distances than in Europe.

Neoclassical economics, gravity models, and international migration

One stream of geographical research took up Ravenstein’s observation that international migration was motivated by economic considerations, especially wage inequalities across space. Accepting the underlying economic logic of international migration, geographers drew on the ideas of economists who argued that people would migrate from locations offering low wages to locations with higher wages. Geographers suggested that the attraction of high wages could be offset, in part or in whole, by travel costs and the lure of intervening places that could be reached at lower cost than the original destination. They proposed a gravity model in which the attraction of high wages would be traded off against travel costs that increase with distance from the migrant’s origin. The gravity model draws on an analogy with Newtonian physics that the attraction between two bodies is related to their sizes and inversely related to the distance between them. According to the simplest adaptations of the gravity model, international migration flows will increase as the populations of two locations increase and decline as the distance between them increases. The gravity model was adapted in three respects. Rather than relying solely on population size to predict the probability of migration, locations were described in terms of their social and economic characteristics. Specifically, unemployment rates, wages, and number of working age adults were some of the variables used to describe the attraction of destinations and the propensity to migrate from various origins. The cost and effort involved in overcoming distance were also evaluated in terms of travel time and travel costs. Competing destinations and origins have also been incorporated in gravity models that take account of migration flows among a number of locations, rather than individual pairs of locations.

Despite these efforts to improve the empirical validity of the gravity model, serious criticisms remain. Gravity model approaches to migration are criticized for treating migrants as independent and atomistic social actors, detached from family and friends. Gravity models are also criticized as static descriptions of past migration
MIGRATION: INTERNATIONAL

trends. They offer no means of forecasting how changes in transportation and communications technology, the ever-changing social and economic characteristics of origins and destinations, and migrants’ evolving social roles and identities will influence migration flows.

Concerns that economic factors other than wages and migrants’ social relations influence migration encouraged new theoretical conceptualizations. Stark and Bloom (1985) suggest that migration decisions are made communally in the best interests of the household to reduce the risk of fluctuations in household income. Household members are sent abroad to diversify income sources and thereby reduce the risk of unexpected and sudden declines in total household income.

Taking these premises on board, gravity models have been modified to include a wide range of economic indicators to describe the economic attraction of competing destinations. Researchers have also devoted considerable effort to estimating travel costs for different migrants. Additional factors such as the presence of coethnics who could provide information about migration and settlement and state regulations have also been incorporated into gravity models.

An important legacy of Stark and Bloom (1985) has been the shift from investigating the motives and characteristics of individual migrants to examining the household context of migration. A growing body of research has investigated how familial structure and household concerns affect who migrates, and when and where people migrate (Kofman and Raghuram 2006). Studies of transnational migrants who try to maintain social identities and social links in multiple countries have revealed diverse migration trajectories. Highly skilled men and women, such as Ghanaian nurses, migrate alone for professional training and opportunities and often pay remittances to their extended families left behind. Likewise, Filipinas also migrate to take up domestic work that allows them to send remittances. “Astronaut families,” in which one parent, usually the father, returns to the origin country to take advantage of economic opportunities and the other parent and children settle at the destination, are highly dynamic. In astronaut families, decisions about who migrates, destination, and the timing of migration are influenced by family considerations at each stage in the life course (Kobayashi and Preston 2007). Recently, growing numbers of international students have travelled abroad to study, some of whom intend to establish residence at their destinations, while others simply seek the cachet of a foreign education. Like the children of astronaut families, they return to their countries of origin to benefit from their foreign education and credentials. Between Canada and the United States, there is a regular flow of seasonal agricultural workers from Central and South America whose families depend on their remittances. Similar flows of migrant labor occur between southern European countries and Africa, albeit on a more irregular basis. Much of this research illustrates how case studies that pay attention to the households from which migrants originate and which they struggle to remake at their destinations can enhance theoretical understanding of contemporary migration. Moreover, attention to the households of migrants recalls Ravenstein’s astute observations about the social contexts in which women migrate.

Systems theory and dual labor markets

One response to Ravenstein’s ideas and their legacy has been to move away from individual migrants as the focus of geographical theorizing and empirical analysis. In a seminal analysis of migration in Africa, Mabogunje (1970) proposed a systems approach that viewed migration
as a self-perpetuating and regular pattern of exchanges between places. The relationships that link migrants and their households across places are the central components of the migration system, which is in constant flux as a consequence of feedback between origins and destinations that alter migration flows. By specifying the components of the migration system, Mabogunje draws attention to explanatory factors at multiple scales ranging from the individual migrant to the nation-state. With its recognition that migration alters origins and destinations and the relationships between them, the migration systems approach moves beyond static gravity models.

Other systems approaches to international migration, world-systems theory and dual labor market theory, emphasize how the interactions between social and political structures affect international migration (Arango 2000). World-systems theory posits a capitalist world economy in which international migration is one of the processes that establishes and maintains a privileged core that exploits the periphery as a source of raw materials and labor. Dual labor market theory extends this analysis by arguing that migrant workers play essential roles in capitalist societies: as a flexible source of labor that can be laid off and returned to its origins when rates of profit fall, and as a reserve army of labor that disciplines the working class.

**Network theories of migration**

Recent explanations of migration focus on the social networks that link current and past migrants with nonmigrants at their origins and destinations. Social networks are seen as a potential source of social capital, resources that can be accessed through social relationships (Arango 2000). For international migrants, social networks are the crucial link between individual migrants and broad social structures. They offer crucial information about the timing and routes of migration and essential support such as temporary housing and referrals to employers that facilitate settlement. While acknowledging that migration decisions always involve individual people, network explanations of migration also highlight how, over time, individual decisions and the social networks that influence them may alter origins, destinations, and the relations between them. In this way, network explanations of migration are dynamic, helping us understand how migration can become self-perpetuating.

Geographers have used network explanations of migration to evaluate how and in what circumstances remittances and return migration are durable strategies for development. Feminist geographers have shown that many women learn about migration pathways, obtain referrals to employers, and care for children and family members at a distance through their social networks (Kofman and Raghuram 2006; Pratt 2012). Others migrate as sponsored family members whose legal status at the destination often depends on maintaining the conjugal relationship.

Research about social networks has concentrated on migration to the wealthy states of Europe, North America, Australia, and New Zealand. We know far less about the social networks associated with the growing migration flows within much of Asia, Africa, Central and South America, and the Caribbean. Researchers have also focused on personal networks, although there is growing evidence that other types of networks, such as those rooted in hometown associations, alumni organizations, and professional associations, influence migration decisions and settlement experiences.

Recent attention to the social networks that structure migration flows reflects a growing interest in mobility and mobilities within human geography (King 2012). As a form of movement,
MIGRATION: INTERNATIONAL

International migration draws attention to who moves and who stays, and to the social conditions that give rise to movement. Research in this vein challenges the normative assumptions that people live sedentary lives, which still frame much thinking about international migration. Recognizing the significance of social networks that connect systemic processes of globalization to individual migration decisions, geographers are investigating how gender, race, class, and sexuality influence migration, as well as how migration alters these dimensions of social identity. This research draws attention to the social, economic, cultural, and political contexts of international migration, highlighting its increasing spatial variegation.

Role of the state

Research is responding to the contradictions inherent in state efforts to control who enters national territory while facilitating the rapid movements of goods, money, and information across national borders. This tension has proved fertile ground for geographers who have found that the competing imperatives of globalization and security contribute to punitive policies that render many migrants temporary or without status at their destinations (Migreurop 2013). Temporary migration policies that are intended to ensure the recruitment of a migrant workforce that grows with labor demand, and shrinks quickly when demand falls, have proven to be highly racialized and gendered.

The attempts of receiving states to coordinate and enforce border policies have heightened the vulnerability of growing numbers of migrants who are desperate to escape poverty and conflict. On the United States–Mexico border, human smuggling has increased as the American government hardened the border, while the creation of Frontex, the agency that patrols the external borders of the European Union, is associated with a growing number of migrant deaths as states try to keep potential migrants at arm’s length by externalizing border control (Migreurop 2013). Packed into every form of transportation from boats to shipping containers, desperate people are being apprehended far from their destinations. Across the Mediterranean, the European Union has signed agreements with North African states such as Morocco and Libya, paying them to detain people who might otherwise migrate to European member states. Having risked their lives, many migrants are held for long periods in isolated locations, living in a legal limbo where they have few rights, and yet they are unwilling or unable to return to their countries of origin.

The competing imperatives of globalization and national security are evident in the growing numbers of migrants without permanent residency. The experiences of caregivers recruited as domestic workers around the world provide a particularly compelling example. Caregivers usually arrive in destination countries with temporary status and few legal rights with respect to employment conditions. As a consequence, they are vulnerable to the capricious and arbitrary demands of their employers and the states where they live and work. Even in the rare circumstances that domestic workers qualify for permanent residence, their experience as domestic workers is devalued in the labor market and may contribute to the tendency of many employers to discount their professional qualifications and experience. For example, Pratt (2012) found that women trained as nurses and teachers in the Philippines who entered Canada as caregivers experienced substantial downward mobility as employers simply ignored their Canadian work experience. Their status as temporary migrants required to work for specific employers in private residences, where there are few
workplace protections, exposes domestic workers to exploitation; however, the gendered nature of their work also contributes to their deskilling.

With no right to secure presence, migrants without status are even more vulnerable than those admitted on temporary visas. Geographical studies have revealed in detail how the border has been internalized as it has hardened. Everyday life is reconfigured as migrants without status try to avoid any interaction with state authorities. Informal taxi services replace driving, children are withdrawn from school, and social networks mobilize to share information about enforcement activities (Stuesse and Coleman 2014). Migrants without status often seek jobs in the informal economy, although their efforts to connect with employers are being stymied in many cities. Some local authorities use loitering bylaws to ban day laborers from congregating on street corners where they wait to meet potential employers. Lacking official status, migrants are being challenged on their rights to be present in the city. These studies underscore the ways that borderlands have expanded as border security has increased. The border has been internalized; it is present throughout the territories of many states.

Another research stream compares how migration policies are formulated and implemented by different jurisdictions within a single state, seeking explanations for policy similarities and differences. States emerge from these analyses as contested, multiscalar, and often contradictory. In the United States, the delegation of immigration enforcement to local police services and secure community policies revealed political divisions between big cities that are home to growing immigrant populations and the suburban municipalities surrounding them (Walker and Leitner 2011). Voting patterns also affected the propensity of municipal governments to institute sanctuary policies rather than enforce federal immigration law. Municipal jurisdictions that had voted Republican favored local enforcement activities, while those with Democratic-voting histories offered municipal services to all residents regardless of immigration status and promoted “don’t ask, don’t tell” policies to discourage municipal employees from asking about immigration status. Similar trends are apparent in Canada where the large metropolitan areas that are immigrant gateways are more likely to implement sanctuary policies than small towns and rural areas where anti-immigrant sentiments are sometimes expressed openly (Leitner and Preston 2011). In Canada, devolution of responsibility for the selection of migrants and the provision of settlement services from the national to the provincial and territorial levels has encouraged provinces and territories to pursue diverse goals. Some facilitated the rapid recruitment of workers to meet employers’ immediate needs while others sought migrants who had specific language and cultural competencies. The contested nature of migration policy is readily apparent in the European Union where states often pursue contradictory migration policies despite their commitment to shared enforcement of the European Union’s external borders. For example, Italy and Spain have a history of amnesties for migrants without status whereas Germany, the Netherlands, and other states in northern Europe have resisted calls for regularization on the grounds that they fuel irregular migration.

A parallel stream of research has compared the migration of highly skilled and business migrants with that of less skilled and less desired migrants. In some states, highly skilled and business migrants enjoy swift access to permanent residence, along with rights to family reunification and full access to national labor markets, while in others they are treated as privileged temporary migrants. Although their economic advantages and privileged migration
status protect highly skilled and business migrants from many of the difficulties experienced by their less skilled and less desired counterparts, the transnational lifestyle still presents challenges. Many transnational migrants struggle to maintain social ties and social identities at their origins and destinations at the same time as they try to make themselves at home at their destinations. The homing desire is highly spatialized. Numerous case studies describe the efforts of all migrants, even the most privileged, to achieve a sense of belonging by creating new homes at their destinations (King 2012).

Representations of international migrants and international migration flows underpin many state interventions. Highly skilled workers are viewed as attractive global talent entitled to permanent residence and citizenship because they will enhance the wealth and economic productivity of the nation-state (Yeoh 2006). Migrants who are considered less skilled are considered suitable for filling short-term and immediate labor needs rather than making long-term economic contributions. Although their remittances are now viewed as indispensable and effective development tools, the migrants that are the sources of these remittances are not seen as worthy of permanent residence. Although a few will benefit from the proliferation of two-step migration procedures by which temporary workers earn the right to apply for permanent residency, the majority will be expected to return to their origins or to live precariously without status. States often demonize refugee claimants and migrants without status as threats to the nation-state. Whether these migrants are labeled as terrorists or as queue jumpers, the representations are used to justify repressive state measures such as detention and deportation.

Geographical views of international migration highlight its social, political, and spatial complexity. Geographers have sought to describe and explain international movements, portray the experiences of increasingly diverse migrants, and investigate how representations of assorted migration flows and migrants influence state efforts to control international migration. There is growing concern that the spatial variegation of contemporary migration will exacerbate long-established geographies of inequality. As a result, geographers are asking how international migration can be a tool for achieving social justice.

**SEE ALSO:** Environment and migration; Migrant division of labor; Migrant labor; Migrant settlement; Migration: forced; Migration: internal; Migration: skilled international labor; Refugees

**References**


**Further reading**


Migration: skilled international labor

Jonathan V. Beaverstock
University of Bristol, UK

Expatriation, international assignments, and global talent mobility

The success of the multinational producer service firm rests on three major interlocking facets: the quality of its key asset, a highly skilled workforce, in terms of their knowledge, expertise, experience, skills and professionalism, and, in some sectors, the status and recognition of their formal qualifications (e.g., legal, accounting); the firm’s reputation, usually built up over a relatively long time period from a business model that has the firm–client relationship at its core; and the ability of the firm to manage its knowledge and practice, reputation and client relationship model on a global scale. Consequently, given that the firm’s knowledge, expertise, skills, and reputational assets are “embodied” within its highly skilled labor force, the strategic deployment of such labor within firms’ international office networks and/or to clients has become a crucial process for global knowledge management, transfer, and exchange. Skilled international migration within the international internal labor markets of producer service firms has become a normalized process of “transnational” work (Beaverstock 2007).

Producer service multinational firms have a cadre of highly skilled professional and managerial labor working outside of their home country on international assignments or other forms of mobilities. Such forms of mobility have been referred to in a historical context as expatriates. Back in 1977, Edström and Galbraith published a seminal text on the transfer of managers within multinational corporations (MNCs). They suggested that managers transfer across borders within firms to fulfill three main strategic functions: (i) to fill vacancies abroad that cannot be filled by home country nationals, across the entire range of the firm from key leadership positions to filling middle-ranking vacancies; (ii) as a part of the firm’s formal management development program to train and develop the global managers of the future, involving both junior and middle-ranking persons; and (iii) as a strategy to disseminate knowledge and practice within the entire firm to build an overall corporate strategy, ethos, and culture. Since 1977, thousands of texts have been published on “expatriation” and international assignments that have used Edström and Galbraith (1977) as the yardstick. For example, Sparrow, Brewster, and Harris (2004) suggest that firms send staff on international assignments for six main reasons: for career development; to establish a cadre of international staff; to fill local vacancies; to transfer knowledge and expertise; to manage local assets; and to provide leadership to roll out corporate strategy and best practice. Skilled international migration within producer service firms encapsulates all of these strategic roles for international assignments, with the addition of direct secondments to clients as a process to supply and deliver bespoke services in co-location.

From the early 2000s, Beaverstock (2007) reported a subtle but important change in the strategy of international assignments (transfers, exchanges, and secondments) within global
accounting firms and wholesale banks – the shift away from designated expatriation or formal assignment programs to much more flexible modes of mobility and movement. Millar and Salt’s (2008) analysis of global mobility in the aerospace and extraction industries corroborated these observations, and they advocated a “portfolio of mobilities” within organizations, including a combination of long-term assignments (1–4 years), short-term assignments (3–12 months), rotations (in months), extended business travel (30–90 days), and business travel (up to 30 days).

From the late 2000s, the discourse on expatriation and international assignments has switched to viewing such mobility of labor as “global talent.” For example, PricewaterhouseCoopers’ report on Talent Mobility 2020 concluded that there will be “more business travel, more virtual tools, and especially more quick, short-term, and commuter assignments,” because “along with short-term and long-term assignments we have frequent travellers, commuters, intra-regional, and virtual secondments to customers and supplier sites, as well as various assignee and talent types, such as executive, skill set and project-based, developmental, and employee initiated” (PricewaterhouseCoopers 2010, 4, 16).

Official statistics on skilled international migration

All OECD (Organisation for Economic Co-operation and Development) members collect official data on international migration (persons leaving or entering a country for more than a year). The collection of international migration statistics for the “highly skilled” are often categorized as “professional and managerial” (as in the United States). From a UK perspective, there are two main types of official data sources on highly skilled international migration: international migration and work permit/working visa statistics. Each will be discussed in turn.

The UK Office for National Statistics (ONS) collects data on long-term international migration (LTIM) for “professional and managerial” occupations, which can be categorized as the “highly skilled.” These include intercompany transfers, and highly educated, technical, scientific, professional, and/or other intrinsically talented persons, for example musicians, entertainers, and sportspeople. From 1980 to 2010, four main trends can be reported from an analysis of ONS LTIM statistics for British and non-British professional and managerial migration.

First, since 1980 and almost year on year for the United Kingdom, there has been an absolute increase in the inflow (immigration) and outflow (emigration) of both British and non-British professional and managerial migrants. In 2010 the number of all professional and managerial immigrants entering the United Kingdom was +245% (+108 000) greater than in 1980, and was +78% (+48 000) for emigration (Table 1). An analysis of the net migration balance (the difference between inflow and outflow) recorded more professional and managerial migrants of all nationalities entering the United Kingdom than leaving in 2010 (+40 000), 2000 (+34 000), and 1990 (+16 000), with more non-British professional and managerial migrants entering the United Kingdom than leaving (+60 000 in 2010; +57 000 in 2000; +25 000 in 1990; +8000 in 1980). For each year reported in Table 1, there have always been more British professional and managerial migrants leaving than entering the United Kingdom. Second, in each of the four time periods, there were more male than female
professional and managerial migrants for both inflow and outflow in all citizenship categories (Table 1). But what is greatly significant about the 40-year period is the growth in the proportion of female migrants entering and leaving the United Kingdom, for both British and non-British citizens. In 2010 the proportion of all female professional and managerial migrants entering the United Kingdom was 42% and of those leaving the United Kingdom was 38%, compared to 27% and 26%, respectively, in 1980.

Third, for all time periods, almost three-quarters of all professional and managerial migrants entering and leaving the United Kingdom were in the 25–44 age range (Table 1). This is not unexpected given that these migrants are by definition “highly skilled” and may have entered the labor market late because of higher or further education, and being in a position to migrate (to a new job vacancy or to seek work) only when they have gained a track record of work experience and/or an established career. Finally, the Commonwealth (new and old) accounted for the highest proportion of non-British citizen professional and managerial migrants entering the United Kingdom for each of the four time periods, representing 46% of total non-British citizen immigration in 2010 (52 000), 45% in 2000 (49 000), 44% in 1990 (24 000), and 46% in 1980 (12 000) (Table 1). Between 1990 and 2010, the number of EU citizens entering the United Kingdom who had been in professional and managerial occupations prior to migration increased from 13 000 to 30 000 (+131%), followed by citizens from other foreign countries, which increased by +82% from 17 000 to 31 000 during the same time period. With respect to the outflow of non-British citizen professional and managerial migrants, with the exception of 1980, EU citizens account for the highest proportions of this occupation group leaving the United Kingdom: 45% in 2010 (24 000); 36% in 2000 (19 000); and 43% in 1990 (13 000).

The UK Border Agency regulates the entry of non-European Economic Area (non-EEA) highly skilled immigrants through a points-based system of first entry and extension work visas (see Salt 2009). The highly skilled immigrant is included in one of two qualifying relevant “tiers”: Tier 1 encompasses high-value migrants like entrepreneurs, exceptional talent, and investors (who do not have a job offer); Tier 2, covering skilled workers who have a job offer and sponsor, includes skilled workers in general, intracompany transfers (ICTs), ministers of religion, and sportspersons. In 2012 a total of 45 740 (excluding dependents) Tier 1 and Tier 2 migrants were allowed to enter the United Kingdom for work purposes, with 82% (37 600) in Tier 2 (Table 2). Tier 2 migrants are non-EEA citizens who enter the United Kingdom to work in organizations, like firms, the public sector, and nongovernmental organizations (NGOs), and the highest proportion of these workers are given permission to work in the United Kingdom as intracompany transferees – that is, they come from outside of the EEA to work in a subsidiary, branch office, or global headquarters of their employing UK or foreign-owned MNC. Between 2009 and 2012, the number of working visas issued to Tier 2 ICTs increased from 8970 to 21 800 (+143%), and they accounted for almost a 50% share of all Tier 1 and 2 visas issued in 2012. These ICTs are in all sectors of the economy.

Prior to 2009, the UK government’s Highly Skilled Migrant Programme (HSMP) issued work permits across all economic sectors, including ICTs. Out of a total of 77 660 work permits issued in 2008, the highest numbers were in computer services (21 690, of which 87% were ICTs); administration, business, and management (10 249, 48% ICTs); financial services
Table 1  Long-term international professional and managerial migration trends by citizenship, age, and sex (thousands), 1980–2010.

<table>
<thead>
<tr>
<th>Year</th>
<th>All</th>
<th>British</th>
<th>Non-British</th>
<th>EU</th>
<th>EU15</th>
<th>Non-EU</th>
<th>All Commonwealth</th>
<th>Old Commonwealth</th>
<th>New Commonwealth</th>
<th>Other Foreign</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inflow</td>
<td>152</td>
<td>39</td>
<td>114</td>
<td>30</td>
<td>21</td>
<td>84</td>
<td>52</td>
<td>15</td>
<td>37</td>
<td>31</td>
</tr>
<tr>
<td>Male (%)</td>
<td>58</td>
<td>59</td>
<td>57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aged 25–44 (%)</td>
<td>76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outflow</td>
<td>113</td>
<td>60</td>
<td>53</td>
<td>24</td>
<td>20</td>
<td>29</td>
<td>21</td>
<td>13</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Male (%)</td>
<td>62</td>
<td>63</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aged 25–44 (%)</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Net</td>
<td>+40</td>
<td>−21</td>
<td>+60</td>
<td>+6</td>
<td>+1</td>
<td>+55</td>
<td>+32</td>
<td>+3</td>
<td>+29</td>
<td>+23</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inflow</td>
<td>162</td>
<td>53</td>
<td>109</td>
<td>25</td>
<td>25</td>
<td>84</td>
<td>49</td>
<td>32</td>
<td>17</td>
<td>36</td>
</tr>
<tr>
<td>Male (%)</td>
<td>60</td>
<td>58</td>
<td>61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aged 25–44 (%)</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outflow</td>
<td>128</td>
<td>75</td>
<td>53</td>
<td>19</td>
<td>19</td>
<td>34</td>
<td>16</td>
<td>13</td>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>Male (%)</td>
<td>63</td>
<td>63</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aged 25–44 (%)</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Net</td>
<td>+34</td>
<td>−23</td>
<td>+57</td>
<td>+6</td>
<td>+6</td>
<td>+50</td>
<td>+32</td>
<td>+18</td>
<td>+14</td>
<td>+18</td>
</tr>
</tbody>
</table>

(continued opposite)
<table>
<thead>
<tr>
<th>Year</th>
<th>All</th>
<th>British</th>
<th>Non-British</th>
<th>EU</th>
<th>EU15</th>
<th>Non-EU</th>
<th>All</th>
<th>Old Commonwealth</th>
<th>New Commonwealth</th>
<th>Other Foreign*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1980</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inflow</td>
<td>44</td>
<td>18</td>
<td>26</td>
<td>3</td>
<td>3</td>
<td>23</td>
<td>12</td>
<td>6</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Male (%)</td>
<td>73</td>
<td>72</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aged 25–44 (%)</td>
<td>75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outflow</td>
<td>65</td>
<td>46</td>
<td>19</td>
<td>5</td>
<td>5</td>
<td>14</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Male (%)</td>
<td>74</td>
<td>76</td>
<td>68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aged 25–44 (%)</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Net</td>
<td>+21</td>
<td>−28</td>
<td>+8</td>
<td>−2</td>
<td>−2</td>
<td>+9</td>
<td>+4</td>
<td>+1</td>
<td>+3</td>
<td>+5</td>
</tr>
</tbody>
</table>

Note: Calculations for net migration figures may be +1 or −1 from the difference between inflow and outflow due to roundings.
*From 2004, Other Foreign excludes the EUA8 countries that joined the European Union; from 2007, Other Foreign excludes Bulgaria and Romania, which joined the European Union in 2007.
Table 2 Points-based system (PBS) Tier 1 and Tier 2 highly skilled workers given permission to work in the United Kingdom, 2009–2012.

<table>
<thead>
<tr>
<th>Year</th>
<th>Total PBS</th>
<th>Tier 1: High-value migrants</th>
<th>Tier 2: Skilled workers</th>
<th>General</th>
<th>Intracompany transfers</th>
<th>Minister of religion</th>
<th>Sportsperson</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>8140</td>
<td>37 600</td>
<td>14 100</td>
<td>21 800</td>
<td>630</td>
<td>1080</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>11 700</td>
<td>33 700</td>
<td>10 100</td>
<td>21 900</td>
<td>460</td>
<td>1230</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>16 400</td>
<td>37 700</td>
<td>14 000</td>
<td>21 400</td>
<td>800</td>
<td>1570</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>18 600</td>
<td>24 600</td>
<td>8970</td>
<td>14 200</td>
<td>500</td>
<td>930</td>
<td></td>
</tr>
</tbody>
</table>

Source: Home Office n.d.

(7852, 51% ICTs); education and cultural activities (6274, 4% ICTs); and health and medical services (5883, 6% ICTs) (Salt 2009). Between 1995 and 2008, the number of work permits issued in the HSMP increased by +53 499, from 24 161 to 77 660 (+221%) (Table 3). In 2008, the highest share of ICTs as a proportion of work permits issued were present in these seven sectors: computer services (87% or 18 870 permits); telecommunications (82% or 2020 permits); extraction, manufacturing, and transport (each 59%, or 821, 1860, and 504 permits, respectively); utilities (53% or 500 permits); and financial services (51% or 4000 permits) (Table 3).

Corporate highly skilled international migration: the cases of global accounting and legal firms

Outside of official statistics on highly skilled international migration, case study research on organizations provides a rich source of data, often derived from a qualitative approach, to understand the organizational strategy for the global mobility of highly skilled staff within, and between, the internal labor markets of MNCs, and small and medium-sized enterprises (e.g., see extraction and aerospace: Millar and Salt 2008). The following case studies of highly skilled professional and managerial migration within large global accounting and legal firms illustrate the strategic importance of such global mobility in the firm’s armory to deliver knowledge across borders, to supply services to clients and cosuppliers, to develop the global ethos and human capital of the organization, and as a mechanism to switch labor resources to markets which require specific demand at certain times of the financial year.

Corporate international migration in global accounting firms

Since the late 1970s, the swift internationalization of the global accounting industry has led to it becoming one of the most prolific producer services sectors responsible for organizational highly skilled professional and managerial labor migration in the world economy (Beaverstock 2007). Despite the availability of smart information and communication technology in the workplace like videoconferencing and the availability of fast and frequent airline travel, global accounting firms continue to move high volumes of professional and managerial staff (partners, qualified accountants, technical staff, and trainees) on international assignments within their extensive international office networks, and outside of the firm in secondments to clients.
MIGRATION: SKILLED INTERNATIONAL LABOR

Table 3  Work permits (first and extension) approved for 1995 and 2008.

<table>
<thead>
<tr>
<th>Industry</th>
<th>1995</th>
<th>2008</th>
<th>Intercompany transfers (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administration, business/management</td>
<td>4041</td>
<td>10 249</td>
<td>48</td>
</tr>
<tr>
<td>Agricultural services</td>
<td>952</td>
<td>304</td>
<td>7</td>
</tr>
<tr>
<td>Computer services</td>
<td>1827</td>
<td>21 690</td>
<td>87</td>
</tr>
<tr>
<td>Construction and land services</td>
<td>182</td>
<td>2338</td>
<td>26</td>
</tr>
<tr>
<td>Education and cultural activities</td>
<td>1901</td>
<td>6274</td>
<td>4</td>
</tr>
<tr>
<td>Education and leisure services</td>
<td>2919</td>
<td>4770</td>
<td>4</td>
</tr>
<tr>
<td>Extraction industries</td>
<td>424</td>
<td>1392</td>
<td>59</td>
</tr>
<tr>
<td>Financial services</td>
<td>3194</td>
<td>7852</td>
<td>51</td>
</tr>
<tr>
<td>Government</td>
<td>46</td>
<td>396</td>
<td>3</td>
</tr>
<tr>
<td>Health and medical services</td>
<td>1774</td>
<td>5883</td>
<td>6</td>
</tr>
<tr>
<td>Hospitality, hotels, catering</td>
<td>320</td>
<td>3865</td>
<td>2</td>
</tr>
<tr>
<td>Law-related services</td>
<td>258</td>
<td>1004</td>
<td>27</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>1987</td>
<td>3153</td>
<td>59</td>
</tr>
<tr>
<td>Real estate and property services</td>
<td>5</td>
<td>140</td>
<td>16</td>
</tr>
<tr>
<td>Retail services</td>
<td>2826</td>
<td>1092</td>
<td>36</td>
</tr>
<tr>
<td>Security services</td>
<td>2</td>
<td>212</td>
<td>47</td>
</tr>
<tr>
<td>Sporting activities</td>
<td>544</td>
<td>2677</td>
<td>0</td>
</tr>
<tr>
<td>Telecommunications</td>
<td>458</td>
<td>2466</td>
<td>82</td>
</tr>
<tr>
<td>Transport</td>
<td>333</td>
<td>855</td>
<td>59</td>
</tr>
<tr>
<td>Utilities</td>
<td>168</td>
<td>945</td>
<td>53</td>
</tr>
<tr>
<td>Other</td>
<td>–</td>
<td>103</td>
<td>–</td>
</tr>
<tr>
<td>Total</td>
<td>24 161</td>
<td>77 660</td>
<td></td>
</tr>
</tbody>
</table>

Source: UK Border Agency, adapted from tables in Salt (2009, ch. 5).

and cosuppliers. In 2014 the Big Four accounting firms, Deloitte, PricewaterhouseCoopers (PwC), Ernst & Young, and KPMG, employed over a half a million fee-earning staff (partners, seniors/principals, and qualified staff) in an average of just over 150 countries around the globe (Table 4). Given the client-facing nature of accounting as the archetypal producer service, a firm’s professional staff will liaise and supply services to clients and cosuppliers, mainly through close physical relationships. Given the significance of embodied knowledge in the accounting profession, the competitive advantage of the firm rests on the quality, reputation, and client relationship qualities of its professional staff. In all of the major global accounting firms, the success of the organization is aligned to a wholly owned physical presence in the market, often with multiple offices, delivering services in co-location with clients. At the crux of service delivery is the ability of professional staff to maintain a highly mobile relationship with clients and to supply high-quality services that are accredited and require legal closure.
Table 4  Highly skilled international migration in the Big Four global accounting firms, 2011–2014.

<table>
<thead>
<tr>
<th>Firm</th>
<th>Total fee-earning staff*</th>
<th>Countries</th>
<th>Mobility program</th>
<th>New global assignments p.a.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2011</td>
</tr>
<tr>
<td>Deloitte</td>
<td>173 900</td>
<td>150+</td>
<td>International Assignments</td>
<td>3800</td>
</tr>
<tr>
<td>PwC</td>
<td>163 000</td>
<td>157</td>
<td>International Mobility</td>
<td>4330</td>
</tr>
<tr>
<td>Ernst &amp; Young</td>
<td>188 292†</td>
<td>150+</td>
<td>Global Exchange Program</td>
<td>NA</td>
</tr>
<tr>
<td>KPMG</td>
<td>133 890</td>
<td>153</td>
<td>Global Opportunities</td>
<td>5000+</td>
</tr>
</tbody>
</table>

*Includes partners, principals, and professional staff.
†Includes all staff.
NA = Not available.
Source: Firm websites.

For over 20 years or more all of the Big Four global accounting firms (and many below that ranking, like Grant Thornton, DBO, and Baker Tilly, for example) have formalized organizational highly skilled international migration within the firm, and between the firm and its clientele, in international transfers and in exchange and secondment programs. Such global mobilities have now become standardized within the firms’ international offices and organizational structures, involving many nationalities and international offices, superseding the primacy of firms’ headquarters as the major source of outwardly mobile staff (Beaverstock 2007). As of 2014, the Big Four firms have established global mobility programs: to provide structured pathways for career development, for example student trainee rotations and as part of formal management development systems; to fill organizational vacancies, often associated with partners required to go offshore to head divisions or departments, or for qualified staff to fill vacancies in tight labor market locations; to build the global culture and ethos of the organization by circulating staff of all nationalities between world regional offices; and as a mechanism to second professional staff to clients and cosuppliers (e.g., law firms). All of the Big Four have globally functioning mobility programs, which attract a few thousand new assignees per annum (Table 4). In 2014 Deloitte had almost 6000 professional staff on its International Assignments program, but for the firm (as for all the other firms), these are just the tip of their international assortment of staffing mobility. As Deloitte (2014, 17–18) states:

International experience at Deloitte means more than just a job in another part of the world. It includes everything from an international assignment in or transfer to another country, short- and long-term international business travel, and working on or managing international projects and teams. Last year [2013], ... nearly 6,000 mobility assignments helped Deloitte professionals develop cultural understanding, as well as language, networking, and managerial skills ... Professionals with global mobility experiences develop and enhance communications, coordination, and management skills, as well as global networks, to be able to better serve clients and effectively carry out global projects.
Ernst & Young (n.d.) offers an assortment of global mobility opportunities for professionals, which include:

- a Global Student Exchange Program, whereby part of an internship is spent abroad for several weeks in “an exciting international location … gaining new experiences … while obtaining global insight”;
- serving global clients abroad and working “alongside a team of EY people from all over the world, which provides you with exceptional cross-cultural experiences and learning opportunities”;
- short-term assignments, which “can offer you a taste of international work … when we have the need for additional resources in another country during busy season”;
- a Global Exchange Program, specifically for “high-performing seniors or managers” with partner potential who apply for a 18–24-month overseas posting in any function of the firm;
- a New Horizons Program, tailored for the “junior ranks to experience geographic mobility through short-term rotations”;
- a Corporate Responsibility Fellows Program, involving three-month assignments, directed at juniors who wish to travel to Central and South America to improve “the success of … high-impact … entrepreneurs in underserved communities”;
- strategic leadership and management opportunities for partners and senior managers, who are offered “a strategic opportunity to relocate to another practice or country to provide leadership or … management, … transfer your knowledge and share your experience”;
- international transfers: self-initiated mobility for professional staff who wish to apply for another position in the organization on a competitive basis, searching the firm’s worldwide vacancy list.

Corporate international migration in global legal firms

At the strategic hub of the global legal firm is the “embodied” expert knowledge of its partners and legally qualified lawyers, and their trusted and established relationship management of clients in home and international markets. As the leading global (mainly US and UK) law firms have internationalized their office footprint in the major world cities and international financial centers, a significant accompanying organizational strategy has been the international migration and mobility of partners, lawyers, and trainees between such office networks. As Beaverstock (2004, 166) notes, global mobility is, “a critical globalization strategy for knowledge management and transfer within professional service legal firms’ transnational office networks.” A recent research report by TheCityUK (2014) estimates that the largest UK law firms in London have about 45–65% of their lawyers working outside the country, compared to the leading US firms, which have about 25% of their lawyers working outside the United States.

At the nub of the organizational strategy for professional staff migration and mobility is the necessity for the global law firm to deal in multijurisdiction law in a single jurisdiction location when involved with global (or local) clients in cross-border and/or local cases or projects. Global law firms frequently use a combination of international staff, who provide expertise in an international jurisdiction (e.g., English) Common Law, and local qualified staff, who are registered to practice in the local jurisdiction, to manage their client work. For example, qualified English Common Law professionals are sent on an international assignment to New York,
Singapore, or Hong Kong to supply an English Common Law jurisdiction to clients in those locations that require such a service, working alongside locally qualified lawyers where appropriate. Global legal work is exceptionally multijurisdictional across the globe, irrespective of the location of the legal firm’s international office network. Also, it is important to recognize that partners especially are sent on international assignments to lead and/or manage a particular specialization of the firm in an international location, or to replace a partner returning to the United Kingdom who has been assigned to another international location.

Both the leading London- and New York-headquartered international firms, including members of the so-called magic circle, and beyond, continue to rotate early career staff (trainees and just-qualified lawyers) for training and to build their individual career paths and experiences, but also as an instrument for the individual to fully understand the global ethos and culture of the firm, outside of the main headquarters. A recent addition to trainee international secondments within the firm is the global mobility initiative of client secondments, where trainees are embedded in the client firm. Thus, many of the leading global firms have international secondment, international student exchange secondment, and client secondment programs, usually lasting 6–12 months for trainee lawyers. Often these are headlined in the recruitment literatures and websites for graduate trainees. For example, Slaughter and May (n.d.) have an international secondment program for trainees solicitors, providing “an excellent opportunity to experience working in a different environment and jurisdiction,” and in 2013 20% of Linklaters’ trainees went on international secondment within the firm, and 26% on client international secondment (Linklaters 2014).

The following quotation from CMS International (n.d.) summarizes very clearly the organizational rationale for global mobility, particularly client secondments during the training contract:

One of our key aims is to become a leading relationship firm. We have many ways of building relationships, but perhaps one of the most personally exciting … arises from our secondment programme. Client secondments really help us to forge the best relationships with our clients while giving you the chance to experience a new industry or profession … For us, international secondments help us work as one team, regardless of international borders – for you they can be hugely stimulating and rewarding life events … You may also find you have other opportunities to develop your career in other offices on a longer-term basis; again we will provide you with assistance, for example, with the logistics and expenses of relocation and with cultural and language training.

The leading global law firms rarely publish data tracing the international opportunities of their staff involved in assignment or secondment arrangements with international offices or other “best friend” firms or clients. Data from TheCityUK (2014), originally reported by the Law Society, estimates that in 2010 there were, “6000 solicitors on the Roll in England and Wales who are working abroad,” with almost a quarter based in Hong Kong (883, or 15%) and Singapore (524, or 9%). A word of caution, however: these solicitors on the Roll working abroad may be employed by a range of global and local law firms, but also in other allied sectors of the knowledge-based economy, for example in corporate finance (including wholesale banking), retail and private banking, shipping and maritime, accounting and consulting, and financial services.

Looking to the future, as producer service firms are at the vanguard of introducing innovation and
new corporate and knowledge management systems to establish, nurture, and fulfill client relationships, and are progressively being organized more across borders and in transnational project teams consisting of an array of cosuppliers (e.g., law firms, corporate financiers, real estate), the likelihood of more flexible modes (both physical and virtual) of highly intensive, “business travel” corporate mobilities are a greater possibility in lieu of traditional longer-term assignments.

SEE ALSO: Firms; Labor migration; Migration: international; Producer services: definition and classification; World cities

References


Further reading


Military geography

Rachel Woodward
Newcastle University, UK

Definitions of military geography

Military geography is the study of the ways in which militarism and military activities are geographically constituted and expressed. Central to this definition is the idea that military practices, both material and conceptual, not only have observable effects on space and on places, environments, and landscapes, but also that militarism and military activities themselves are brought into being through the inherent spatiality of social relations. Also significant to the definition is the idea that military geographies emerge both through the material practices of military institutions, organizations, and associated social groups, and through the conceptual and discursive processes of militarism. Militarism is the prioritization of military force in the resolution of conflict, and analysts talk also of militarization, a broad set of social, cultural, economic, and political processes by which military approaches to social problems and issues gain both elite and popular acceptance (Woodward 2004; 2005). Armed conflict itself, and all the indirect, non-conflict activities that armed forces undertake to prepare for the deployment of military power, are of equal interest to military geography.

Military geography, then, is a subfield of human geography, which takes as its central focus the spatialities of military practices. Because military geography engages with the outcome of primarily political practices, it is frequently understood as complementary to political geography, and shares with political geography many of that subdiscipline’s epistemological concerns with poststructuralist theoretical approaches and heterogeneous methodologies. It is notable, however, that many of the topics studied within military geography draw explicitly on broader economic, social, and landscape geography traditions and approaches to make sense of the spatialities of militarism and military activities. Also, and reflecting the wide-ranging impacts of military practices, materialities, and discourses, military geography looks across the social science disciplines (sociology, political science, international relations, feminist studies, and social psychology in particular), the humanities (particularly historical, cultural, archaeological, and landscape studies), and the environmental sciences, meaning that the subfield is defined less by disciplinary parameters and more by its multidisciplinary perspectives on a core set of issues relating to military forces and all this entails.

Military geography has existed as a term defining a field of study for as long as the discipline of geography itself has existed, with a formalization of its scope and approach emergent in the mid-nineteenth century as a consequence of the colonial and imperial ambitions of nation-states, particularly the United Kingdom and the United States. A definition established at that time and still in circulation within some academic groups defines military geography as the application of geographical tools and techniques to the solution of military problems (Galgano and Palka 2011). Although this – very applied – understanding of military geography still has some purchase, the majority of writers within contemporary
Anglophone geography would seek to distance themselves from the notion of military geographical scholarship as contributing to the military aims and objectives of states and armed forces and their efficiency, working instead with a more critical approach, which sees military issues as the outcome of social life and political contestation, rather than just accepting “military” as a functional and given category, and emphasizes militarism and militarization as simultaneously discursive, ideological, and material. With this critical turn, military geography has paralleled developments within military sociology and international relations, which are in their own ways working toward a more political, less functional, military scholarship.

Dimensions of military geography

The scope and dimensions of military geography have always been responsive to developments in military organization, strategy, and technology, and to shifts in the political relationships between military institutions and civil society. Military geography therefore has a dynamism reflecting this responsiveness, but fewer well-established traditions among its approaches. There are, however, four key dimensions to military geography through which most research has been focused: summarized as spatiality, place, environment, and landscape.

Spatiality

Spatiality, both as a capacity of social (and thus military) relations, and as an empirically observable feature or outcome of social activity, is the first key dimension in military geography. At its most basic and conceptually straightforward, and reflecting a long tradition of interest in battlefield strategy, military geography is interested in the operation of military activities across space – the “terrain and tactics” approach that seeks to understand how specific military campaigns or incidents have been shaped by environmental factors including local geomorphology, topography, climate, specific weather events, or biogeography, which in turn may or may not have proven significance for the outcome of that particular campaign or event. More nuanced and conceptually sophisticated approaches to the study of the practical deployment of military force have in recent years explored the battlespace in its entirety (so including the aerial and maritime as well as land-based dimensions), looking at how the geographical imaginations implicit in military planning and strategizing become played out across space and demand particular visualization and mapping practices (Williams 2011). The spatialities of interest to military geographers also include global, regional, national, and local power plays over territories, borders, and boundaries, and the ways in which military power and action shapes territorial blocks. Also significant here are changing modes of warfare, including developments in the technologies of armed conflict and deployment of lethal force, the ways in which these have transformed the nature of armed conflict in the twenty-first century, and the ways in which state and non-state actors can pursue armed violence. The pursuit of armed violence includes obvious kinetic activities, but also less visible practices around visualizing, mapping, and conceptualizing the battlespace in order to exert military control over space in ways other than direct physical presence or violent destruction of people and place. Increasingly, military geography’s interest in space is expressed less as empirically orientated descriptions of actual deployments of military power, and more as explorations of how space is conceptualized by military planners and strategies, and what
these theoretical understandings might in turn mean for the uses (and abuses) of military force in the present and future.

**Place**

Military geography has a range of interests in questions of place, a second key dimension to the topic. Reflecting a broader human geographical interest in the specificity of places as nodes within networks of social relations, work on place has tended to take specific types of military places as case studies for the exploration of wider social relations. So, for example, research on the economic geographies of the defense industry and defense economy has tied the growth of national and increasingly multinational defense industrial capacities to both the changing fortunes of (inter)national manufacturing industry and associated research and development functions, and to the use of the defense industrial base as a tool of both economic and military power projection. Defense economies are also evident in the economic relationships between military installations (i.e., bases and barracks) and their surrounding localities, and a key area of interest here has been the extent to which military bases do, or do not, support wider economic activities in the localities and regions in which they are based. Military bases also raise a host of issues concerning the social relations they shape and reproduce, and military geographers, along with sociologists and anthropologists, have sought to trace out the complexities and consequences of civil–military relationships in localities where a military base is present. The place-making practices of military power are also evident in the reproduction of urban and infrastructural forms and built environments around the globe as a consequence of the extension and consolidation of military power particularly (but not exclusively) by the US armed forces (Gillem 2007). Geographers have also focused on the ways places come to be imagined, and the specific ways in which military power, control, and objectives can be seen to provide particular readings of places at scales from the specific to the regional as part of mechanisms to legitimate a military presence (Farish 2010).

**Environment**

The effects of military activities on environments, and the ways in which discourses of environmentalism become articulated for specific purposes and ends, constitute a third key dimension of military geography. Military activities have profound environmental impacts, ranging from the modification of local ecologies and geomorphological features through to contamination and pollution. The environmental effects of military activities are associated primarily with instances of armed conflict; the physical destruction brought by the deployment of artillery, practices of aerial bombardment, or processes of infantry maneuver can be readily imagined, and accounting for environmental impacts in both the immediate aftermath of war, and over much longer-term time periods, remains a significant element of military geographical study. In addition, military geography has also been attentive to the more mundane and prosaic impacts of non-conflict military activities through, for example, the pollution effects of military basing or the environmental protections inadvertently brought to some environments by the blocking effects of military presence. The apparent paradox, that military sites can be simultaneously sites of military presence and of high environmental value, has prompted a wealth of studies that have looked to tease out both the validity of the claims made by military organizations about the benign effects of military practices, and the politics of such claims (Havlick
In parallel with the approaches of environmental historians, military geography seeks to explore also the role of representations of environments and their specificity in arguments about national identities and the pursuit of armed conflict in assertion or defense of those identities.

Landscape

The fourth key dimension of military geography is the focus on landscapes. Military landscapes can be conceptualized in a great variety of ways (Woodward 2014) but underpinning this range is a focus on the distinctive ways in which military and nonmilitary/civilian actors view, interpret, and represent landscapes, the political functions of these diverse interpretations, and the insights that a specifically landscape-focused approach can bring to understanding the operation of military power. Studies of military understandings of landscape include the specific ways in which ground is read for military purposes, at scales ranging from the efforts of small infantry patrols to assess territory for tactical purposes through to the mechanisms by which swathes of territory are interpreted for broader strategic purposes. As with assessments of military environmentalism, military landscape studies have been heavily influenced by concepts of representation, in turn reflecting dominant approaches in cultural geography since the 1990s. The focus on representation has generated a wealth of studies of, in particular, sites of memorialization, as geographers attempt to engage with the meanings and understandings of armed conflict conveyed through, for example, battlefield sites, sites of mourning and remembrance such as war memorials, and sites of wartime atrocity. Studies of military landscapes, again reflecting developments more widely in cultural geography, are increasingly attentive to ways of being in landscape, including awareness of and interest in the perceptual, haptic, sensory, embodied, and affectual effects of military landscapes. With the attention to the phenomenologies of encounter comes also an awareness of the politics of being in such place; military landscapes are not neutral spaces, and there are practical and methodological issues these sites raise for the geographer which shape the possibilities for their exploration.

Military geography in relation to human geography

In addition to its focus on the range of topics within these dimensions, military geography is also increasingly concerned with contributing to two significant developments within human geography. The first of these concerns the ways in which concepts of militarism and militarization are defined and used. Although definitions of these terms are given at the start of this entry, it would be misleading to suggest that they are fixed or that they have shared currency across and beyond geography. That these terminologies and associated conceptualizations are used in often markedly different ways by different writers is indicative, first, of the novelty within the social sciences of taking militarism and militarization seriously as objects of study. Reflecting human geography’s seeming disinterest in military issues in the postwar period, these terms and their conceptual uses have only relatively recently been subject to critical interrogation. Indeed, arguments have been made for their increasing obsolescence in a world where specifically military concerns and issues might be seen as part of, or on a continuum with, a broader range of issues pertaining to security and securitization. Conversely, the specificity of military organizations as agents of states, of military violence as planned and executed by state-legitimized actors who sit
in a relationship with civil society by virtue of that state sanction, and of military capabilities reflecting investments by the state, indicates that geography may need to engage with questions of military specificity – as well as related but distinct issues around security – for a while yet.

The second development in human geography has been the extension of specifically military-related studies across political (and to a lesser extent social) geography; the study of military geographies, in other words, is not the sole preserve of those self-identifying as military geographers. This extension of the study of military phenomena reflects the fact that in the first decade and more of the twenty-first century the US-led wars in Iraq and Afghanistan, the ongoing military violence in Israel and Palestine, and the long-running armed conflicts in central Africa, have made armed conflict inescapably real as a feature of daily life for geographers working in a variety of national contexts. Accordingly, political geography, and in particular critical geopolitics, has become much more alert to issues of militarism and militarization as part of the latter’s attempts to grapple with the spatialities of violence (Dalby 2010). This development in turn has prompted explorations of popular cultural engagements with the materialities and discourses through which armed violence is understood; the rich emergent literature on military-themed video gaming is a case in point.

In turn, geography has been significant for military studies more generally, particularly for critical military studies, which seek to question not only causes, consequences, and effects of armed conflict and military capabilities, but also the ways in which military processes and practices can be seen as the outcome of social life and political contestation, rather than just viewing “military” as a given functional category. In exploring the spatial constitution of military power, the multiple ways in which this works at a range of scales (including, significantly, the individual and embodied), and the intersections of different scalar effects, geographical approaches to military topics (whether branded as military geography, critical geopolitics, or something else entirely) have a significant contribution to make to a cross-disciplinary critique of the uses and abuses of military power. And while to date the engagements between those advocating serious study of military topics and those arguing for a focus on a geography centered on nonviolence and an ethics of peace have been limited, it is important to note that in its contemporary critical incarnation, military geography should not be assumed to be antipathetic to antimilitarism.

Future research directions for military geography

Of the many future directions and issues for military geographical study, three stand out for particular scrutiny. The first of these is methodological. Military phenomena continue to be studied, as they have long been, drawing the full range of methodologies and research techniques used across the social sciences (Williams et al. 2016). These include quantitative data collection and analysis, the usual variety of interactional methods including interviews of various types, focus groups, and ethnographic studies, textual analysis of documentation, visual methodologies, and field-based practices of walking and observation. Military research, however, raises questions about accessibility in ways that have quite pronounced effects on what can, and cannot, be studied, because of the nature of military institutions themselves. Although not always closed, secretive, or hostile to social scientific researchers, they frequently are. This hostility in turn has an effect on what
MILITARY GEOGRAPHY

can be studied and how. Methodological innovation as means of countering access problems is a hallmark of military geography – and much critical military research is not sponsored or endorsed by military gatekeepers. That said, the relative obscuring of military geographical studies within human geography is at least partially explained by the difficulties of doing military research, and this remains an issue for those concerned to explore the nature and effects of military power.

The second key area for future research lies with the political economies of military institutions and military power which arise from the movement toward the privatization of military functions and the consequences of the retraction of the state from the direct execution of military activities. The knock-on effects of neoliberal economic policies in many advanced capitalist nation-states since the 1980s have played out in distinct ways with regards to military issues. While geography is starting to become alert to the most visible of these effects, in, for example, the exponential increase in the number of private security contractors deployed by NATO governments and other agencies as part of force deployments, the less obvious, sometimes mundane, but equally profound effects, such as the transfer of personnel accommodation, facility management, specialist equipment training, and recruitment services to the private sector, are all in the process of prompting new spatialities of economic and social relations at scales from the individual to the global, and with effects from the battlespace to the home base.

A third area for future research is the contribution that military geography can make to understanding civil–military relations. There is considerable potential for a more spatially alert conceptualization of the relationships between military institutions and organizations and the civil societies in which they are located. Over the past decade, more critical approaches in military studies have looked to reconceptualize civil–military relations less as a managerial issue where the purpose of theoretical critique is the greater efficiency or management of armed forces as a tool of democratic states, and more as a discursive relationship where the categories of “military” and “civilian” are in constant states of imagination and rearticulation. The role of scale and of spatial relations in the process through which the categories consistently define themselves against each other is a pressing political as well as geographical question.

SEE ALSO: Critical geography; Geopolitics; Political geography; Postconflict geographies; Security; Violence; War

References


Mining and mineral resources

Gavin Bridge
Durham University, UK

Mining is one of two primary activities by which societies acquire raw materials, the other being cultivation. The extraction of mineral resources is part of a broader process of “resource metabolism” through which human beings create social, economic, and political order via the mobilization of materials. Understood narrowly, mining refers to a set of techniques associated with the removal of solid materials that, unlike fluids, will not move of their own accord: used in this way, mining describes the application of human labor and energy to separate and remove materials. However, mining is often used more broadly, describing the process by which all physical states of matter are taken from the Earth, whether as solids (coal, gravel, gold, peat), liquids (brine, oil, water), or gases (natural gas, helium). The association of mining with physical removal means it is also commonly used as a synonym for the depletion of exhaustible resources. In this context the application of the term to resources that are potentially renewable has a particular evaluative force: the mining of groundwater supplies or soil fertility, for example, describes a condition in which the rate of extraction exceeds the rate of recovery (see Natural resources and Resource extraction). Consequently the verb “to mine” has become a metaphor for the unsustainable use of resources.

To a physical scientist, a mineral is a naturally occurring substance with specific chemical and physical properties, composition, and atomic structure. To most geographers, however, mineral resources are concentrations of nonliving materials whose form, qualities, and quantity are valued positively by society. Modern societies are materially complex and the range of materials valued as mineral resources has grown over time. Material complexity has enabled improvements in the performance of products and the productivity of personal, corporate, and national economies: for example, a computer chip utilizes over 60 different elements (i.e., half the number of elements in the periodic table); and the nickel-rich “super-alloys” used in high performance aircraft turbine blades have over time incorporated a growing suite of additional elements – cobalt, molybdenum, niobium, hafnium, rhenium – in order to operate at progressively higher temperatures (Graedel et al. 2015).

The integral role of minerals within modern economic life means that mining and mineral resources occupy several core areas of policy concern. These include physical availability and the implications of resource depletion, the geopolitical consequences of resource distribution and patterns of resource consumption, the environmental effects of mining, and the challenges of mineral-based development (see Natural resources development). Geography has made significant contributions to understanding these issues. Over time, geographers have adopted a progressively more relational perspective on mineral resources: they have become focused less on “the mine” (as an industrial unit and point of extraction) and more on the interdependencies between mineral resource production and consumption and economic,
political, and environmental outcomes at a range of scales. Recent work, for example, highlights the accumulation of mineral resources in the built environment and entrained in consumer waste flows (see Waste and waste management), and the implications of international trade in fuels and metals for geographically uneven development.

Mineral resource classification

By convention, mineral resources are divided into 
*metals* (e.g., gold, iron, nickel), 
*fuel minerals* (e.g., coal, shale oil, uranium), 
*industrial minerals* (e.g., gypsum, phosphates, salt, sulfur), and 
*construction materials* (e.g., building stone, gravel, sand). However, because these categories reflect socially meaningful differences in how minerals are utilized, they may be differently combined or subdivided to distinguish, for example, *bulk minerals from specialty minerals*; *nonrenewable, recyclable* (metals), and *potentially renewable* mineral resources (such as beach sand, river gravels, sea salt); or to identify a subset related to particular policy concerns (e.g., *strategic minerals, critical minerals, conflict minerals*).

“Mineral resource” is therefore a social category applied to a range of materials, and, as a social category, it folds within it the unsettled nature of value, the possibilities of technological change, and the artfulness of industry. Whether or not the designation “resource” holds depends, among other things, on prevailing systems of valuation, the impacts of production technology on costs and prices, and the distribution of social power. Mineral resources, therefore, are always in a state of “becoming” (see Natural resources). This dynamic nature is reflected in formal classification systems – such as the McKelvey box (Figure 1) – that distinguish reserves from total resource base, and inferred reserves from those that have been demonstrated. These distinctions reflect the degree of geological certainty with which quantities and qualities of materials are known, and the economic feasibility of their extraction. As knowledge and/or economic conditions change, so do the size and location of mineral reserves. It is for this reason that the relationship between reserves and current rates of production – the so-called R/P ratio – is a poor indicator of future availability: the R/P ratio for most minerals has remained static or improved over time, despite prodigious rates of mineral resource consumption. A rise in the price of oil, from around $20 to over $100 per barrel, underpinned a 50% increase in the size of global oil reserves between 1990 and 2013 as formerly subeconomic reserves became commercially viable, and higher prices spurred new exploration. Similarly, the decision of the Chinese government to restrict the export of rare earth elements in 2010 drove up their price, increasing reserves by 25%. Although mineral resources are an inherently dynamic category, the everyday sense we have of materials such as oil, iron, and gold conceal the conditions of their possibility. However, unconventional resources – such as oil shale, tar sands, gas hydrates, or e-wastes – often reveal the work necessary to stabilize materials as mineral resources, as resource classification methods and the techno-economic assemblage through which value, costs, and markets are constructed are in the process of formation (Kama 2015).

For the most part, however, mineral resources take on the appearance of independent physical objects – stocks in nature’s warehouse – awaiting incorporation into economy. Mining moves mineral resources across a significant conceptual threshold dividing “nature” from “economy.” Once entrained within economy, mineral resources commence a trajectory from extraction, through material processing and product manufacturing, to final demand. This sense of mining as a physical and defining act of taking, in which
Figure 1  McKelvey box for mineral resource classification. http://www.rmi.org/RFGraph-McKelvey_diagram_for_coal_gas_resources. McKelvey 1972. Reproduced by permission of Rocky Mountain Institute.

materials innocent of human needs are bent to human will, is reflected in the popular association of mineral resources with “virgin” raw materials, a psychosexual gendering of Earth and extraction that has been integral (rather than incidental) to the expansion of extractive enterprise over the past few centuries (see Merchant 1980; Kolodny 1975). However, the relation between mining as a technique and the lithosphere as its target is not a necessary one: the technologies and practices of mineral prospecting, selection, and removal can also be applied to secondary material recovery and recycling in which mineral resources are extracted from the “ores” of accumulated urban infrastructures and wastes (Wallsten et al. 2013). “Urban mining” encapsulates the application of extractive techniques to concentrations of construction, municipal, and electronic wastes deposited in the environment (e.g., in landfills) and, potentially too, to the stockpiles of “hibernating” materials incorporated into the built environment during earlier periods of growth but no longer being used (such as old mobile phones that have yet to be recycled, or abandoned utility and transportation infrastructures). For some metals, such as copper, the volume and value of total “anthropospheric” mineral resources – consisting of wastes and hibernating materials, as well as materials currently in use – is broadly similar to known geological resources. For many elements, the concentration in industrial waste streams is richer than for geological ores (Kapur and Graedel 2006). Recent work to quantify the scale of above-ground stocks of mineral resources
MINING AND MINERAL RESOURCES

accumulated in the metropolitan built environment, and to map the complex trade networks around scrap metal and other wastes/resources (Gregson and Crang 2015), subverts conventional geographical imaginaries that locate minerals underground in the “resource periphery.” Material flow, life cycle, and material footprint analyses, for example, generate new understandings of the geographies of mineral resource stocks and the “anthropogenic mineral cycle,” and are indicative of a growing interest in how prevailing patterns of mineral resource production and consumption sustain socioeconomic and political-ecological relations (Müller et al. 2006).

Mineral resources in the Anthropocene

The scale of mineral resource extraction has grown at an unprecedented rate over the past century. For many mined materials, over half of all production has occurred in the last 50 years or less and, for some, the volume of material moved by humans now rivals background natural flows. Mineral resource acquisition has attained such scale that it has become possible to consider it equivalent in scope and significance to a geological force: mining and quarrying, for example, move around 60 billion tons per year, while the amount of material moved by water erosion is estimated to be around 53 billion tons (Douglas and Lawson 2002). This pattern is not limited to bulk minerals: the amount of sulfur mobilized through fossil fuel combustion and metal smelting, for example, is as large as that attributable to volcanic activity. Recognition of these cumulative effects re-imagines mining and minerals processing as a planetary-scale process of material mobilization in which humans act as geological agents. Such thinking has been central to the rise of the Anthropocene as a way of framing contemporary nature-society relationships (see Anthropocene and planetary boundaries).

The primary drivers of the enormous growth in mineral resource production and consumption have been Industrialization and the expansion of economic demand. These have been assisted by technological changes in production that reduce the costs of extraction, pulling down prices and expanding consumption, and by a growth in world population. At the broadest scale, four significant inflection points can be identified within the trajectory of mineral resource consumption, associated with (i) nineteenth-century industrialization in Europe and North America and the advent of bulk shipping for mineral resources such as coal and copper ores; (ii) the new materials of the so-called Second Industrial Revolution in the late nineteenth and early twentieth century; (iii) a postwar expansion in economic demand centered on “core” economies but also associated with peripheral industrialization; and (iv) the period from 1990s onwards, linked to the deepening of neoliberal economic reforms and a shift in the center of gravity of the world economy decisively towards Asia. As outlined below, these four periods map loosely onto distinctive techno-material assemblages over the past 150 years, in the manner described by Kondratiev’s longwave economic cycles. The material and technical shifts associated with the first three also correspond broadly with the characterization of eotechnic, paleotechnic, and neotechnic phases of civilization outlined by Lewis Mumford (1934).

Preindustrial uses of mineral resources were relatively constrained – stone and metals, for tools and armaments, building materials, arts, ornamentation, and exchange – and limited by prevailing techniques of physical and chemical separation. The so-called seven metals of antiquity associated with the material foundations of the ancient world – copper, gold, iron, lead, mercury, silver, tin – were either those that could be taken from the environment in
elemental form, or those that had melting points within the range of charcoal furnaces. A number of other metals were isolated during the Middle Ages, but it was industrialization from the eighteenth century onwards that brought about a fundamental shift from biomass to mineral-based metabolism (Barca and Bridge 2015). Industrialization drove an expansion in metal and fuel use, in construction materials associated with urbanization, and in the application of mineral resources to improve agricultural production (lime, guano). The revolution in organic chemistry in the mid-nineteenth century based on coal also gave rise to a number of novel materials, including aniline chemical dyes (the so-called coal tar colors, of which mauve is among the most striking). Recognition of the possibility for chemical recombination and synthesis laid the foundations for materials science, whose creative possibilities would later be given even greater rein in the context of petrochemicals.

The second major shift in mineral resource metabolism is associated with the materials revolution that occurred at the start of the twentieth century associated primarily with electricity and a growing understanding of atomic structure. This produced a suite of “neo-technic” materials: steel alloys (including stainless steel, via alloying with nickel and chrome), light metals such as magnesium and titanium, and the availability of materials in quantities enabled by electricity, such as aluminum (following development of the Hall-Héroult process in the late 1880s) and the widespread application of electroplating techniques. In comparison with the minerals most closely associated with industrialization – coal, iron, and steel – these materials were lighter and more easily worked: by understanding their molecular structure, they could be tailored for novel and precise applications, as in the development of specialized stainless steels. This period also saw a process of partial substitution, in which formerly biomass-based materials were augmented by mineral-based substitutes: in textiles, rayon substituted for silk, nylon for cotton, and polyester for linen and wool; while in agriculture “fossil” fertilizers – particularly natural gas and mineral supplies of phosphate and potash – largely replaced organic fertilizers, such as manure, guano, and bone meal. This second great shift in the scale and diversity of mineral consumption was centered initially on the United States: between 1918 and 1952 the United States consumed more minerals than had the rest of the world in the entire period before 1914.

A third shift – characterized by environmental historians as the “Great Acceleration” – occurred in the postwar period. It saw not only an intensification of mineral demand, but also a quite dazzling process of diversification in materials, particularly around nonmetallic mineral resources linked to the petrochemical sector. Growing consumer demand, and the need of a hitherto militarily orientated petrochemical sector to find domestic applications for a wide range of chemicals and plastics, drove a process of substitution. Plastics, for example, partially replaced glass and metal in packaging, domestic design, and industrial applications. A fourth and more recent shift in the scale of mineral resource production can be identified around the turn of the new millennium. This is associated, in part, with the proliferation of information and environmental technologies. Production of materials such as platinum and cobalt, for example, widely used in electronics, has increased fourfold in last 15 years, while lithium production has doubled in the same period. However, the significance of the past couple of decades is not limited to new materials: coal and uranium production have also doubled, and iron ore has followed a similar trajectory, with output tripling since 2000. A significant driver of these shifts has been geographical shifts in industrialization.
MINING AND MINERAL RESOURCES

associated with rapid economic growth in Asia, facilitated by neoliberal trade legislation including the accession of China to the World Trade Organization (WTO) in 2001. The current rate of global mineral resource Consumption is unprecedented, and has risen faster than the rate of population growth: global resource consumption per capita is estimated to be around 10 tonnes per year, implying an annual draw of 70 billion tonnes but also concealing wide variation in mineral resource consumption: industrial countries dominate material flows on a per capita basis, although a number of developing economies (and China in particular) have seen rapid and substantial growth in their material use in the past couple of decades.

Concern at the continuing expansion of mineral resource flows – and recognition of their complex geographical expression – has encouraged a number of efforts to calculate the material “footprints” associated with resource consumption. Direct material consumption is relatively straightforward to calculate on a national basis from domestic production and trade figures. Such statistics indicate, for example, that the average European will use 460 tons of sand and gravel, 166 tons of oil, 39 tons of steel, 1 ton of copper, and lesser amounts of other metals and rare earth elements in a lifetime. However, the weight of consumed resources is an imperfect measure of environmental impact for four reasons. First, it excludes the “hidden flows” of wastes generated during extraction, materials processing, and use which, for metals in particular, can be many times that of the mineral resource. Second, although sand, stone, and gravel account for the greatest weight of material moved, the complex geochemistry of metal ores and the use of reagents for mineral processing mean that the pollution effects of metal mining can be much more severe than those associated with the extraction of construction materials. Third, a measure of materials apparently consumed within an economy over a given period does not account for the ways in which some of them may be subsequently recovered and reused. And fourth, exposure to low-volume/weight materials, such as PCBs (polychlorinated biphenyls), mercury, cadmium, and benzene, may be far more significant for environmental and human health than exposure to those that dominate mineral resource flows by weight or volume, particularly where their effects may be amplified by bioaccumulation.

Nonetheless, mineral resource “footprint” calculations can be valuable because they point to the growing geographical interdependence of mineral resource trade. As the volume and diversity of materials consumed have grown, so supply chains linking production and consumption have also become increasingly global. Over the past three decades trade in raw materials has increased 2.5-fold (Weidmann et al. 2015) and many minerals have experienced large shifts in the geographies of demand and supply. In bauxite production, for example, the mid-twentieth century’s giant producers – Jamaica, Suriname, and Russia – have been eclipsed by Australia, Brazil, and China; similarly in iron ore, China, Brazil, and Australia now account for three-quarters of global production, replacing the dominance of Russia, the United States, and France several decades ago. The geography of mineral resource consumption is closely tied to industrial use and urbanization and therefore maps onto patterns of economic growth: China, India, and Brazil, for example, have rapidly become very significant mineral resource consumers. Their growth has also spurred significant domestic mineral production: for example, China was the world’s largest producer for 14 major minerals in 1995 but, by 2012, its lead role had expanded to 44 minerals (British Geological Survey 2014). The production of a few mineral resources has become more geographically concentrated over
time: the global supply of fluorspar, graphite, tungsten, and rare earth elements, for example, has grown much more dependent on production in China. However, other supply chains have become significantly more diverse: the long dominance of South Africa in gold and Canada in nickel production, for example, has sharply reduced, associated in the case of nickel with the development of new sources of supply in the Philippines, Indonesia, Russia, and Australia.

Perspectives on mineral resources and policy

The economic, political, and environmental issues associated with their production and consumption ensure mineral resources are frequently an object of social concern. The analytical perspectives of resource economics and political science have long dominated the policy space surrounding mineral resources. Geographers, however, have made a number of significant interventions, characterized by (i) a critical interest in the scales and spaces through which mining and “mineral problems” should be understood; (ii) adoption of a relational perspective, centered on the interdependencies between mining and other livelihoods, between extraction and consumption, and between Natural resources development and political institutions; and (iii) a thorough-going critique of Malthusian assumptions about physical scarcity and resource limits. Two broad areas of geographical inquiry into minerals are briefly presented here: mineral resource availability (including questions of strategic and critical vulnerability); and the environmental and political-ecological consequences of resource mobilization.

The availability and criticality of mineral resources

Availability is the original “environmental problem” surrounding mineral resources. It relates not to the impact of mining on the environment, but to the environment’s capacity to furnish society’s expanding demand for raw materials. Mineral depletion and the prospect of future Scarcity are hoary human concerns with roots in antiquity (Glacken 1967) which break into popular consciousness from time to time, often triggered by abrupt (although frequently short-term) changes in price. Rapidly expanding use of coal in the mid-nineteenth century led economist William Stanley Jevons (1865) to conclude Britain faced an imminent “coal crisis” that would inevitably lead to national decline. The country’s coal production did peak early in the twentieth century, although as a consequence of shifts in demand and fuel substitution rather than physical constraints on supply. The US Conservation Movement, most closely associated with fears of a national “timber famine” in the early twentieth century and the establishment of the US Forest Service, also expressed concern at the breakneck pace of coal, gas, and other mineral development. Gifford Pinchot, the first US Chief Forester, expressed this alarm, observing that “the five indispensably essential materials in our civilization are wood, water, coal, iron and agricultural products … We have timber for less than thirty years at the present rate of cutting … We have anthracite coal for but fifty years, and bituminous coal for less than two hundred. Our supplies of iron ore, mineral oil and natural gas are being rapidly depleted, and many of the great fields are already exhausted” (Pinchot 1910, 123–124, cited in Tilton 2002).

Pinchot’s prognostications on the availability of reserves proved misguided, as national reserves of oil, gas, and other mineral resources grew rapidly in subsequent decades. In the mid–1950s, right at the heart of the country’s postwar boom in
domestic oil production, geologist Marion King Hubbert projected US oil production would peak in the 1970s and then permanently decline. Crude oil production in the United States peaked in 1972 at 10 million barrels per day, and subsequently shrank to around 5 million barrels by 2010. However, the rapid development of unconventional “tight” oil resources, combined with offshore oil production from the Gulf of Mexico, has seen domestic crude oil production rebound to around 9 million barrels per day. The rate of rebound is such that the International Energy Agency have forecast (2012) the United States will outstrip Saudi Arabia as the world’s largest oil producer by 2020; others caution that the resurgence of domestic oil production cannot be sustained at such a level.

Growing environmental consciousness in the 1960s fueled a public debate over mineral resource constraints and other “limits to growth.” The Massachusetts Institute of Technology-led Meadows Report to the Club of Rome, for example, used computer simulation modeling techniques to conclude that mineral resource availability was a significant constraint on future growth. However, falling prices for many mineral resources during the 1980s and 1990s appeared to indicate scarcity was in retreat. Sharp increases in price in the first decade of the new millennium subsequently put resource availability and the prospect of peak supply back on the political agenda. Two mineral resources – phosphorus and oil – have been particularly significant in this regard. Phosphorus is an essential plant nutrient for which there is no substitute and its application as a fertilizer underpins crop yields worldwide (see Soil fertility and management). Uncertainty over the size of phosphate reserves has driven concern that global production of phosphorus will peak somewhere between 2030 and 2070. However, the most pressing limit (for import-dependent regions such as the European Union (EU) and Australia) is likely to be geopolitical: Morocco and the occupied territory of Western Sahara account for 85% of global reserves and the bulk of phosphate exports. In response, research is seeking ways to better manage on-farm phosphorus utilization, improve the efficiency of application, and instigate phosphate governance at a global scale (Cordell and White 2014).

Similarly, proponents of “peak oil” forecast an impending permanent decline in the availability of conventional crude oil as geophysical limits begin to bite. However, growing production from unconventional sources – such as Canada’s tar sands, the heavy crudes of the Orinoco Delta, or so-called tight oil formations such as the Bakken in North Dakota, combined with the erosive effect of high oil prices on demand growth – have allowed total oil output to keep pace with demand. The continued growth of output has been underpinned by a significant qualitative shift in global oil reserves that can be characterized as the end of “easy oil.” Although the abundant, low-cost, conventional oil reserves of the Middle East continue to dominate supply, reserves replacement increasingly relies on higher cost reserves that are more complex to develop, and more energy and greenhouse gas intensive. A similar process is underway in metals, where the transition to lower grade ores over time means rising energy, greenhouse gas, and process water requirements. As with oil, these sustainability constraints act as soft boundaries rather than geophysical limits: in the case of greenhouse gas and water requirements, for example, they can readily be exceeded, although at environmental and social cost. This suggests, therefore, that mineral resource availability is not simply a net effect of the forces of depletion and scientific knowledge, as resource classification systems such as that shown in Figure 1 would suggest. Mineral resource availability is, to a very significant degree, a social and political choice.
about what level of environmental and social disruption will be tolerated, and on whom their environmental and social costs will fall.

As the example of phosphorus illustrates, mineral reserves and mineral consumption are geographically uneven and frequently poorly aligned. As a consequence, mineral resource availability is very often a geopolitical concern. The result is a national focus on the availability of militarily strategic or economically critical minerals that parallels a concern for “global” resource availability and planetary boundaries (see Energy security). A strategic or critical mineral describes one for which there is little if any domestic production but domestic consumption is militarily or economically significant. The combination of supply risk (likelihood of disruption associated with import dependence) and impact of a disruption to supply (a function of a mineral’s importance in use) defines a “criticality matrix” (Figure 2). The EU’s Raw Material Initiative, for example, was launched in 2008 in recognition of Europe’s high dependence on raw material imports for sectors such as construction, chemicals, automotive manufacturing, aerospace, and renewable energy. As part of a wide-ranging initiative that includes resource diplomacy, innovation, and efficiency, the EU has identified 14 “economically important raw materials which are subject to a higher risk of supply interruption.” These critical “EU14” include antimony, fluorspar, graphite, germanium, indium, rare earth elements, cobalt, niobium, and gallium. A high degree of import dependence is not, in itself, a measure of criticality: also important is vulnerability to supply interruption, associated with reliance on a geographically limited range of sources and/or the possibility for politically motivated restrictions (including cartelization). In the case of the EU14, for example, criticality reflects the concentration of production for these minerals in China, Russia, Brazil, and the Congo, as well as limited options for substitution and low levels of recycling.

The concept of strategic minerals implies relevance to military or national emergency planning, and emerged to prominence in the United States in the wake of the Korean War. A surge in raw material prices during the conflict rekindled long-standing concerns about the availability of mineral supplies, and led to the formation of the President’s Material Policy Commission in 1952 (the Paley Commission). Written in an historic context marked by the defeat of Fascism and an escalating Cold War with the Soviet Union, the Commission’s report – Resources for Freedom – examined whether the United States had “the material means to sustain its civilization.” The Commission defined the country’s “Material Problem” as “consumption … expanding at

**Figure 2** Criticality matrix, in which the degree of criticality increases from lower left to upper right (i.e., mineral A is more critical than mineral B). [http://www.nap.edu/openbook.php?record_id=12034](http://www.nap.edu/openbook.php?record_id=12034). National Research Council 2008. Reproduced by permission of National Academies Press.
compound rates ... pressing harder and harder against resources which, whatever else they may be doing, are not similarly expanding.” While observing that the United States had outgrown its domestic resource base to become a major importer of many mineral resources, it nonetheless rejected self-sufficiency as a response to the country’s changed materials position. It proposed a mix of free-market and government-led response, encouraging US companies to invest overseas in search of low-cost mineral supplies, while also invoking the government’s capacity to extend price guarantees and requisition plant and equipment in a time of war, and making permanent stockpiling provisions for strategic materials.

During the Cold War the United States designated over 60 “strategic and critical” minerals and metals, holding reserves of many in the National Defense Stockpile. Strategic vulnerabilities associated with a high degree of import dependence have resurfaced as a concern of US mineral policy from time to time: in the 1980s, for example, dependence on South Africa for supplies of chromium, manganese, and platinum-group metals (platinum, palladium, rhodium, ruthenium, osmium, and iridium) was considered a source of strategic vulnerability, as the struggle against apartheid in South Africa had the potential to disrupt exports of these minerals. More recently, concern about strategic and critical supply has focused on the availability of minerals integral to the advanced digital economy and the so-called rare earth elements (REE) in particular (despite their name, REEs are relatively abundant in total quantity but occur in low concentrations so that their extraction and processing are difficult and costly). REEs such as neodymium, dysprosium, and cerium have key applications in information technology (smart phones, computers, flat screens, and light-emitting diodes (LEDs)), the clean-tech sector (in magnets for electric motors, for example), and in many defense applications (for guidance, control, and weapons systems). REE production has doubled in the past 20 years and, notwithstanding that mineral reserves are relatively widespread, around 90% of production has come from China since the 1990s. In 2010 China announced a policy of reducing exports in order to protect domestic consumption, igniting policy concern among importing countries and particularly in the EU, United States, and Japan.

From environmental impacts to the political ecologies of resource mobilization

Mining and mineral resource development take up a small proportion (less than 1%) of the Earth’s surface: estimates for the United States, a country with an extensive mining history, indicate that mineral extraction occupies only 0.25% of the land area, compared to 3% for urban areas and 70% for agriculture. In the nineteenth century, geographer and proto-conservationist George Perkins Marsh (1965/1864) largely dismissed the impacts of mining, noting they “must always be too inconsiderable in extent to deserve notice in a geographical point of view.” However, a growing body of research challenges this perspective by (i) expanding its analytical scale, away from the mine as a single point of extraction, to consider the ecosystems and biogeochemical cycles within which mines are embedded; and (ii) considering the broader political ecologies to which mining, and international trade in mineral resources, gives rise.

The extraction, processing, and consumption of mineral resources generate a wide range of environmental impacts, many of which have been long-standing sources of concern for individuals and communities exposed to them. Agricola’s (1998/1556) meticulous observations of mining in sixteenth-century Saxony record
not only mining’s effects on forests, rivers, and wildlife, but also how these impacts can undermine the legitimacy of extraction and rally those challenging the primacy of mineral development over other land uses and livelihoods:

[T]he strongest argument of the detractors is that the fields are devastated by mining operations … Also they argue that the woods and the groves are cut down, for there is need of an endless amount of wood for timbers, machines and smelting of metals. And when the woods and groves are felled, then are exterminated the beasts and birds … Further when the ores are worked, the water which has been used poisons the brooks and streams, and either destroys the fish or drives them away … Thus, it is said, it is clear to all that there is greater detriment from mining than the value of the metals which mining produces.

Because mining’s impacts can be severe, societies have wrestled for hundreds of years with the conditions under which rights should be granted to miners for access to land, or for the release of wastes to air and water. Agricola’s commentary neatly expresses how mining’s environmental impacts exceed the process of excavation, narrowly understood, to include mining’s energy demands (on timber and charcoal to fuel furnaces, or water to power lifting gear); the utilization and release of chemical reagents; and the effects of material removal on waterways and aquatic life. A large body of research now outlines mining’s role as a source of soil, water, and air pollution at local and regional scales, with significant work on acid mine drainage, mercury mobilization, the failure of tailings dams, and the impacts of smelting operations and fossil fuel combustion on land cover and water quality. By focusing on the transport, behavior, and fate of physical and chemical pollutants in the environment, this work opens up a broader perspective on mining’s environmental impacts including, for example, the way materials are modified (dilution, recombination, bioaccumulation) by biogeochemical cycles. Paralleling this concern with pollution, research on biodiversity and ecosystem health has examined the suite of ways in which mining modifies ecological processes operating over broad geographical scales, including mining’s role in driving regional land-use/cover change (see Land-use/cover change and climate), habitat fragmentation, and associated effects on biodiversity. One effect has been a growing recognition of off-site impacts, particularly around water, biodiversity, and the ways in which mineral resource production articulates with livelihoods. In the area of policy, this has included a growing recognition that some landscapes should remain off limits to mining, and has led to a greater concern with the cumulative effects of mineral resource development.

The field of Political ecology has made significant contributions to illustrating the wider social and environmental implications of mineral resource development, highlighting alternative scales and temporalities through which the significance of mining can be understood. While much of this work remains focused on spaces and communities affected by resource extraction (Bebbington and Bury 2013), political ecology has also shown an interest in understanding the wider political and ecological relations to which flows of mineral resources give rise. Its critical move is (i) to see site-specific conflicts over mineral resource extraction as part of a more generalized extension of the commodity frontier (see Frontiers), associated with bourgeoning material demand and the adoption of economic policies favorable to external investment in mining; and (ii) to focus on interdependencies between these “resource struggles” and the high rates of economic growth enabled in some parts of the world by the consumption of mineral resources. This deep relational perspective reframes mining and mineral resource development as a matter of global social justice: a
MINING AND MINERAL RESOURCES

distributional conflict that centers on the capacity of some groups to acquire and use mineral resources while displacing the impacts of these processes onto others (Muradian, Walter, and Martinez-Alier 2012).

Research on the ecological distribution conflicts associated with trade in mineral resources draws conceptual support from neo-Marxian arguments about capitalism as a process of spatial differentiation and uneven development by which country and city – core and periphery – become distinguished from one another and co-produced (see Cores and peripheries and Dependency theory). A useful heuristic in this work has been the notion of “metabolic rift” (Foster, York, and Clark 2011). In its original formulation, this term recognized how nineteenth-century urbanization significantly disrupted the historical pattern of nutrient cycling. The growth of urban food demand removed nutrients from the countryside (via the export of food crops) and effectively piled them up in the city (in the form of wastes), simultaneously setting in motion a geographical search for new sources of soil fertility, and efforts to manage urban wastes. Metabolic rift has been applied to planetary resource flows describing, for example, anthropogenic influences on biogeochemical cycles consequent to industrialization, such as disruption of the global carbon cycle and the flooding of atmospheric carbon sinks (Clark and York 2005). The concept’s critical value for contemporary work on mining and mineral resources is its capacity for understanding how international trade in mineral resources is simultaneously a process of political-ecological differentiation at the world scale based on the appropriation of energy and matter. It enables mineral resources trade to be reinterpreted as a form of ecologically unequal exchange, through which materials with enormous productive potential (metals, fuels) are siphoned from parts of the world with limited industrial capacity. Appropriation and consumption of these materials enables the creation of economic value, thereby differentiating core from periphery and ensuring the core continues to accumulate the technical and political-economic capacity to acquire future resources (Hornborg 2015).

Critical perspectives such as these have meant geographical research has tended either to be skeptical of efforts to promote development through the extraction and export of mineral resources, or to argue that its ability to do so is contingent on a specific set of institutional, technological, and environmental conditions. Such critical caution in the face of policy assertions is a hallmark of contemporary geography’s engagement with mineral resources. The continuing promotion of mineral resource exports as a pathway to economic development (notwithstanding its strikingly poor record; see Resource curse), and the emergence of “sustainable mining” as an active policy agenda, suggest such caution is warranted. For most mineral resources, the idea of sustainable harvesting or sustainable yield does not apply, because the resource is unable to regenerate over timescales that are meaningful to humans. The economic case for sustainable mining rests on three arguments: (i) that mineral extraction and processing convert stocks of natural capital into replaceable human capital; (ii) that wealth creation is compatible with the antipoverty objectives of sustainable development; and (iii) that a diversity of mineral resources are key to a “green growth” agenda, centered on high-technology, resource efficiency, and renewable energy (see Ecological modernization). Each of these arguments merits scrutiny: the first rests on strong claims for the substitutability of natural and human capital; the second (and third) on assumptions about the relationship between growth and the distribution of wealth; and the third that green growth
actively substitutes for other growth forms (via policies that squeeze out high-carbon material flows, for example, or which steer investment from primary extraction to secondary recovery) rather than serving as a supplemental “niche market” expanding overall metabolism.

SEE ALSO: Anthropocene and planetary boundaries; Ecological modernization; Energy security; Frontiers; Industrialization; Natural resources; Natural resources development; Resource extraction; Scarcity

References


MINING AND MINERAL RESOURCES


Mixed-method approaches

Timothy L. Hawthorne
University of Central Florida, USA

Many geographers increasingly recognize the value of mixed-method approaches in their research. Definitions of mixed-method research vary across and within fields, including in our own discipline of geography. Mixed-method research is defined here in a way similar to Creswell and Plano Clark (2011) as any form of research where the researcher intentionally combines multiple methods in order to harness the strengths of multiple approaches, while recognizing the limitations of using only one research method for data collection and analysis.

Geographers utilize mixed methods in their research for a variety of reasons. First, they recognize that added rigor may result from combining methods and analyzing datasets using both qualitative and quantitative methods. Such combinations often allow the researcher to explore and uncover deeper meanings and understandings in her/his research, which would not be possible using only one method or approach.

Qualitative methods on their own are often valued for the fact that they are subjective, inductive, and allow context, meaning, and individual lived experiences to emerge from a research study. In order for such themes to emerge, researchers often utilize methods that directly capture the experiences of individuals or groups of individuals. Such methods often include observation, ethnography, interviews, focus groups, phenomenology, case studies, and/or grounded theory. The data emerging from qualitative methods often allow the researcher to contextualize and understand geographic processes in a way that emphasizes the voices of an individual being researched, while demonstrating that all knowledge and individual experiences are socially constructed in a particular time and place.

Quantitative methods on their own are often valued for the fact that they are objective, deductive, and allow for theories and hypotheses to be proposed, tested, and analyzed. Statistical analyses are often at the core of such approaches where the goal is to understand relationships across variables in a way that can be quantified and analyzed with some statistical power. Quantitative methods are often lauded for their objectivity, their likelihood for replication across geographic settings, and their ability to generalize from a smaller sample to a larger population. Common quantitative methods may include descriptive surveys, statistical analyses, or case-control studies, to name a few. In many cases, geographic information systems (GIS) is often viewed as a predominantly quantitative method (though as discussed below, GIS can be viewed as a mixed method).

Mixed-method approaches are valued for their ability to combine the power of both qualitative and quantitative approaches to provide a more complex account of the issue, construct, process, or phenomena being researched. Often mixed-method researchers celebrate the intentionality of choosing multiple research methods to draw out the strengths of multiple approaches and to uncover the hidden silences inherent in any one method or approach. As many geographers move toward a more mixed-method
Mixed-method approaches are faced with a serious challenge in their research design: determining the appropriate way to integrate data gained from qualitative and quantitative methods. Creswell and Plano Clark (2011) demonstrate three viable integration options for mixed methods, including: merging data, connecting data, and embedding data. Merging integrates the qualitative data gained from interviews, group discussions, and images or videos with the numerical information captured from quantitative data. In many mixed-method studies, merging is achieved by reporting the different types of data separately to inform both sets of data. For example, a researcher studying health-care accessibility in an urban area might first report the quantitative data of travel distances computed in GIS between an individual’s home and her/his health-care provider. Then the researcher would introduce qualitative data in a series of themes or quotes to contextualize the travel distance and accessibility experience by demonstrating how difficult it was for an individual to access a health-care provider in an urban area.

A second integration strategy in mixed-method research, according to Creswell and Plano Clark (2011), involves connecting data. Connecting occurs when a researcher uses the analysis of one dataset or method to inform subsequent analytical methods. Extending the health-care accessibility research example discussed above, a researcher using a connecting strategy might choose to first examine the travel distances of participants in the study. The researcher might find that lower-income residents have to travel longer distances than higher-income residents to find an affordable health-care provider. Using these data and results through a connecting approach to integration, the researcher could then create a series of questions for use in an in-depth interview that asks participants to explain how they felt about their long commuting times and the experiences of their journey from home to the health-care provider. In this way the results of the quantitative GIS method to measure travel distance could be connected to the follow-on interview questions to understand the experiences of an individual’s journey to a health-care provider.

A third approach to data integration, according to Creswell and Plano (2011), involves embedding data. Embedding occurs when a primary research strategy and method are utilized and a secondary dataset is included to support the larger research design. For example, a health-care study might be designed to test whether or not a certain physical activity intervention can help with weight loss in obese individuals. An embedded qualitative method might be included in the intervention to quickly assess how an individual is doing with the physical activity intervention.

The recent, increased adoption of mixed-method research in the discipline of geography is a direct result of the field’s diversity and its greater internal debates about what counts as rigorous knowledge production. In an effort to avoid dualistic and divisive thinking, mixed-method researchers seek to move toward a more hybrid view of knowledge production where one method is not necessarily more valued than another method.

Though there are several examples of movements and debates about hybrid, mixed-method approaches in our field, one of the recent examples within geography is from the early 1990s GIS and society debates (see Sheppard 1995) and the field of critical GIS (see Schuurman 2000) which emerged
from these debates. Throughout the GIS and society debates of the early 1990s, some geographers called for a stronger recognition that knowledge was socially produced, and that technologies such as GIS were not value-free and objective, but rather simultaneously shaped by and shaping society. During the time of these debates, a pivotal disciplinary moment occurred in which many within the social theory and GIS camps proposed that multiple research methods were necessary to uncover the complexities of geographic processes and phenomena.

As Kwan (2004) argued in her centennial piece in *The Annals of the Association of American Geographers*, the GIS and society debates (and the subsequent cultural turn seen in emergent critical GIS studies) provided a unique moment for many in the discipline to discuss knowledge production. Using the GIS and society debates as one example, Kwan argued for a radical departure in thinking about knowledge production. She challenged Thomas Kuhn’s definition of scientific inquiry in which Kuhn argued that dominant paradigms must supplant other paradigms in scientific revolutions. Kwan contrasted Kuhn by arguing that a complete discarding of one paradigm for another would be unhelpful for geography as a whole and GIS specifically. She proposed the concept of hybridity for recognizing the strengths and weaknesses of two or more ways of thinking. At the time, Kwan stated that: “through these processes of polarization and boundary enforcement, alternatives are less likely to emerge, and geographers tend to be conveniently identified in terms of rigidified, incompatible, stereotypical, and binarized identities” (Kwan 2004, 757). Hybrid geographies as Kwan noted recognize the potential for unique findings and innovative ways of thinking at the intersections of seemingly incompatible knowledge bases and research methods.

Mixed-method approaches are a form of the hybrid geographies Kwan called for in 2004. A handful of recent hybrid, mixed-method studies that challenge dualistic thinking through the combination of qualitative and quantitative methods in order to capture the complexities of real-world processes and phenomena are introduced here. Knigge and Cope (2006) introduce grounded visualization as a mixed method and suggest that using only quantitative data fails to show important processes occurring in local neighborhoods as they research community gardens. In their work, qualitative interviews with participants, participant observation, and GIS analysis of land-use data are used together through a grounded theory approach in which “contextualized cartographic narratives in geographic discourse” are created (Knigge and Cope 2006, 2023) to reveal the everyday lived experiences and situated knowledge found in their research study site.

Another example of mixed-method research is evident in Pavlovskaya’s 2002 study of economic restructuring in Moscow. Pavlovskaya shows how GIS is used in a period of economic transition to understand the city’s economic structure. She notes that informal economies at the micro level (i.e., the household scale) are not considered in official statistics of the city. Through qualitative interviews with individual households, she uncovers the importance of these informal activities and complements official economic statistics of urban change. Such data suggest to policymakers that they may be missing an important component of the city’s survival.

Kwan (2007) uses 3-D geovisualization (focused on GIS, mapping, and individual interviews completed while riding through the city with research participants) to examine the emotions and feelings of Muslim women as they travel around a major US city after September 11, 2001.
11, 2001. In these 3-D visualizations, participant feelings and perceptions of safe and unsafe spaces in the city are revealed in a way that has not been done in conventional GIS research.

The mixed-method studies introduced above are just a small sample of a growing body of work within geography, but share a commonality that points to the value of mixed-method scholarship. Such studies openly recognize the situatedness of all knowledge and challenge so-called objective truths. By assuming that all knowledge is situated, mixed-method researchers move toward more context-specific analyses of people, places, and geographic processes.

In closing, it is important to recognize the intentionality of mixed-method scholars in designing studies with multiple methods and understand that all knowledge is socially constructed and situated within large debates about epistemology, ontology, and methodology. It is also prudent to stress that mixed-method scholars understand that qualitative and quantitative methods alone have value and provide different explanatory and descriptive powers in geographic research. The real power for mixed-method scholars, however, lies in combining the strengths of qualitative and quantitative methods to uncover hidden silences in the data and to provide a more complete understanding of the construct, process, or phenomena being studied in geographic research.

**SEE ALSO:** Quantitative methodologies

---

**References**


**Further reading**

“Mobility gaps” issues are concerned with whether every individual has adequate access to mobility to be able to satisfy the needs of their everyday life. Daily mobility depends on a wide range of determinants which can restrict or force it.

Some of these factors seem obvious. Depending on income, the economic budget that an individual can devote to transport either limits or opens up the size of the perimeter and the set of choices within which he or she can carry out everyday activities. The more one can spend, the further one can go and the more opportunities one can access. Likewise, residential location – which is also related to income – influences daily travel options. Individuals who live close to shops, schools, or jobs that suit their needs increase their chances of reaching many destinations within a realistic time or money budget, whereas living far from urban resources reduces it. The characteristics of a neighborhood also facilitate or impede travel because different neighborhoods provide uneven transport services as alternatives to car ownership. Among obvious determinants, there is no doubt that physical capacities (including the effects of age) also affect freedom of movement.

But less evident social characteristics also matter. The position of individuals in society, based on their social and cultural capital, can explain a certain relationship with space and a pattern of exposure to other individuals. For example, migrants with low proficiency in the host-country language may lack the know-how required to travel (find information, read directions or maps, or understand the symbols that serve to locate oneself in space), and people suffering social domination may feel unwelcome in some neighborhoods. The upper classes may feel a greater legitimacy to navigate through unfamiliar spaces.

Other social determinants of daily mobility involve the gender perspective. It has been observed that women have a smaller spatial range and average commute than men (Hanson and Pratt 1988) although the gap is narrowing. The major reason for this lies in the social role of women, who are much more active than men in domestic work and therefore have less time available for other activities, including work. As these obligations put pressure on their personal schedules, the need to reconcile their domestic and professional lives means that it is easier for them to work close to home. Thus, the allocation of paid and unpaid work within a household imposes variable limits to the geographical range of the household members. This has long been at the expense of the occupation of women, as spatial entrapment can restrict the proper matching of qualification and job.

These observations show that mobility determinants are not solely the outcome of individual characteristics: social, economic, and spatial environments play a considerable role. How easy it is to travel and access opportunities, or even...
whether this is possible at all, also depends on the environment at large: how the transport system is organized and benefits people; how populations, employment, and services are distributed in space, which itself depends on how land use and the land market are regulated; how working hours make the existing services available to people; how social norms influence behaviors; and so on. Finally, the conditions likely to impact individual mobility are those that define accessibility (i.e., the potential for reaching opportunities), namely the availability of nearby opportunities and/or the ease of traveling to more distant ones (Handy and Niemeier 1997).

The social effects of barriers to mobility

Barriers to mobility can impair life chances. A long tradition of surveys has studied the consequences of travel constraints on access to fundamental needs such as employment, education, health, social services, and so on. Initially in the United States, and then in other countries, the issue of accessibility has stimulated a vast amount of work that examines these effects. For example, all things being equal, it is acknowledged that low-skilled workers face poorer job accessibility than other workers, that is, the job opportunities available within a given travel time and budget from home are more limited. This is both because low-skilled workers often live at greater distances from the main employment centers and have fewer resources (access to car and travel budget) to overcome these distances. On this basis, poor accessibility is seen as part of the explanation for the longer periods of unemployment experienced by low-skilled job seekers (Kain 1992). More broadly, the accessibility issue is today related to the risk of social exclusion, whereby people are prevented from participating in many aspects of social life (Lucas 2006).

Nowadays, access to travel is judged to be a major issue because the nature of contemporary urban societies calls increasingly for mobility in the conduct of everyday life. Job insecurity demands the ability to be mobile in order to change job location. Working time flexibility increases the need for autonomous means of transport, as mass transit is often inadequate during off-peak hours. Women’s employment requires transport services that allow children to move on their own. Urban sprawl necessitates the ability to travel long distances and use a car. The increasing structural need for mobility and the adverse effects of restricted mobility make equity in transport a strategic issue for policy agendas, and its measurement a matter of concern. However, the meaning of mobility indicators is not obvious and the increasing need for mobility is contested.

Mobility as a problematic indicator of inequity

Unlike accessibility, which describes a range of choices, mobility refers to choices that are actually made: the decisions between options, in which personal preferences play a role. Indeed, daily mobility occurs in a world of constraints but is not exclusively the result of constraints. A trip is rarely the automatic consequence of either pure constraints or the strict need to perform an activity that does not contain any degree of choice. To some extent, journeys help an individual to fulfil personal goals. Residential location or lifestyle, in particular, are frequently primary choices that affect travel decisions and the nature of mobility, and which are often made possible by daily mobility. Since daily mobility combines constraints and choices whose relative importance is hardly discernible, interpreting mobility is complex. How would we analyze travel behavior? How would we assess inequity?

Daily mobility is measured by synthetic indicators which make it difficult to read. Traditionally,
the number of trips a day tells us about an individual’s level of participation in social activities and number of relations in the community. However, what about individuals who return home three times a day because of a fragmented working day or those who multiply contacts without travel, using distance communication? The distance traveled usually tells us how free a person is to travel – access to money, time, and a convenient mode of transport or freedom from cultural barriers. But what about individuals who have no choice but to travel far or those who have the choice to give priority to neighborhood-based living? Are long trips the sign of great freedom of action and short moves the sign of restricted options? Or, conversely, are short trips the sign of individuals who have the resources to control their environment and long trips a mark of forced actions?

In the classical economic vision, all things being equal, people are assumed to prefer short distances and low expenditures. In such a scenario, nearer is better. However, for a long time, wealthier groups were the ones moving the greater distances, and short trips were distinguishing low-income groups and were seen as a sign of restricted options. Given that all things are never equal, Zahavi and Talvitie (1980) have argued against the classical vision proponents that, with time and money limits, people may chose to satisfy more important personal aims – such as residential aspirations – than save on their travel distances and costs. In the second half of the twentieth century, there was a massive movement of population from the inner cities to the outer suburbs, as individuals were able to own their own homes in low-density environments, with increasing daily travel distances as the price to pay for achieving that goal. For these individuals, nearer was not better. However, nowadays, great distance seems less and less to be synonymous with free choice. Because of rising house prices, many households are forced to live far from employment centers and to endure travel over long distances.

Thus, it is difficult to say whether the level of mobility is a sign of restricted options or a sign of freedom. The significance of daily mobility indicators is uncertain because they say little about the range of options that prevailed before the actual choice. Since mobility behaviors include an element of choice, assessing equity or inequity would assume that it is possible to distinguish between behaviors that are chosen on the basis of individual preferences and those that are determined by exogenous constraints and determinants. Otherwise, mobility indicators would be unequivocally interpreted as measures of inequity according to a hierarchy of travel behaviors. The notion of mobility gaps thus seems questionable, because we cannot say that there is such a thing as a consistent norm according to which it could be categorically stated that an individual has a deficient or forced mobility.

In past decades, the hierarchy of travel distances actually used to conform more or less to the hierarchy of socioeconomic status. But preferences and social norms have evolved, and this may explain why previous traditional hierarchies of travel behaviors no longer apply and often present reverse trends. On average, men and women show more and more similar travel behaviors. The inhabitants of big cities now travel lower distances than in smaller cities. More and more, blue-collar workers are commuting further than executives; and cycling or the use of public transport becomes the privilege of the former while driving is the mark of the latter – while it used to be the opposite in the past. This blurs the assessment of mobility according to a single hierarchical scale. But it might not totally condemn the prospect of assessing behaviors in relation to new social norms that signal what is socially desirable.
MOBILITY GAPS

Contradictory goals

Equity in daily mobility is an important issue but it is difficult to assess and to address. Policy-makers are increasingly concerned about this crucial question, but to frame policy is challenging. Facilitating the satisfaction of everyday life’s needs by enhancing individuals’ facility to move has long been the prevailing option. But it becomes apparent that encouraging private autonomy is controversial in light of sustainable development goals, and financing alternatives such as public transport is challenging. In addition, it is clear that facilitating travel raises the norm of expected mobility and doesn’t reduce gaps (Orfeuil 2011). One group’s higher capability to travel creates a context that worsens the consequences of the other group’s lower travel capabilities (Fol and Gallez 2014). That is why more and more critics are opposed to the mobility credo. Facilitating the satisfaction of everyday life’s needs by promoting spatial proximity to opportunities (instead of mobility) is another option that is less often pursued because it demands complex and active land-use and land market regulation.

SEE ALSO: Accessibility, in transportation planning; Daily mobility

References


Modeling uncertainty in digital elevation models

Ashton M. Shortridge  
*Michigan State University, USA*

Jessica V. Fayne  
Matthew T. Rice  
*George Mason University, USA*

Digital elevation models (DEMs) provide representations of the Earth’s continuous surface elevation. While various data models and structures are possible, the raster model is the most common: a rectangular region of the landscape is depicted as a gridded array or square or rectangular cells, each of which stores a height value representing the elevation above sea level for its location. The Earth’s surface is of fundamental importance for a vast range of physical and human processes. Consequently, DEMs are used in an especially wide variety of applications, including suitability analysis, viewshed delineation, route optimization, image registration, environmental risk assessment, and as inputs to climate, hydrology, and soil models.

There are differences between any particular DEM and the real-world elevations it represents (Fisher and Tate 2006). These differences may be due to mismatches between the underlying ontology of the DEM and its reference terrain (ontological mismatch) or from errors introduced during measurement and processing (production error). Where these differences are unknown, either because they have not or cannot be measured precisely, uncertainty exists. This entry considers the sources of uncertainty in digital elevation models, their consequences for the application of DEMs, and approaches that have been taken or proposed to deal with uncertainty, particularly through statistical models.

### Error and uncertainty

This section begins with definitions for some critical terms – error, uncertainty, and fitness for use – before discussing sources of error in DEMs and exploring the nature of uncertainty in more detail. “Error” is the difference between the actual elevation and its representation in the DEM (Heuvelink 1998) gives a broader discussion of this definition). “Actual elevation” implies an independently collected ground truth measure to compare against a portion of the dataset, though this measure is itself subject to error. Typically, error is not known throughout the domain of the dataset but only at sampled locations. In contrast, “uncertainty” is the degree to which the fidelity of the DEM to the actual terrain is open to question. This is quite different from error. At sampled locations where the error is known, uncertainty is actually zero. It is at the other locations that uncertainty exists.

In general, DEM users are not interested in the dataset itself but rather in the real-world terrain that the dataset imperfectly represents. They need to know how imperfect this representation is, as it relates to their applications (Fisher and Tate 2006). A user deploying a DEM to identify what can be seen from an observation tower cares about how imperfection in that DEM may lead to imperfection in the calculated viewshed. Another user might wish to identify water runoff flow paths for the same area, and that user should
care about the consequences of imperfection in the DEM for the calculated flow paths. It may well be that uncertainty in the elevation values has different impacts on these different applications. “Fitness for use” is a paradigm in the quality standards community that poses the utilitarian question: is the fidelity of this dataset sufficient for my particular application? To satisfy fitness for use, uncertainty from DEM elevations must be propagated to the process output.

Earlier two quite distinct sources of DEM error were briefly described: that due to mismatch between the specifications of the DEM and of the reference data used as ground truth, and that due to production, either in measurement or processing. Both are considered in more detail here.

Ontological mismatch
As reflected in its very name, a DEM is fundamentally a model: a characterization of real-world terrain that necessarily involves simplification and abstraction from the conceptual model to its representation in a spatial database. A standard conceptual assumption is that each location has a single and theoretically measurable elevation. Real-world surface features with multiple elevations (e.g., cliffs and overhangs) are not supported by this model. Further, locations with ambiguous heights (e.g., unstable terrain like marshes or bogs, or with thick surface vegetation cover) may not be precisely measurable, or may have different heights at different points in time.

A key DEM computational modeling choice is the selection of an output raster spatial resolution. This resolution is connected directly to the spatial scale, or support, at which the terrain is represented. A coarse spatial resolution maintains a more general approximation to the actual landscape, with fewer fine-scale topographic features. Also important is the ontology of the reference data used to “ground truth” the DEM. These reference data have their own conceptual and computational model, the details of which may not match those of the DEM. As an example, consider the use of differential GPS gathered in the field as reference to calculate error in GDEM, a near-global product derived from ASTER stereopair imagery. GPS heights are measured at essentially point support, while GDEM is a 30-meter resolution raster, in which elevations within grids may be thought of as averages. Differences in measured elevations indicate mismatches in what is being characterized.

Production error
A range of production methods have been used for DEMs. In one general family of approaches, elevation measures for discrete locations are obtained from some primary source (e.g., manually surveyed locations, LiDAR point clouds) and then these measures are interpolated to regular grid points or cells to construct the DEM. In another general family, the primary sensor collects data in a gridded format initially (e.g., interferometric synthetic aperture radar (SAR), stereophotogrammetry) and through processing these data are converted to elevations, often, though not exclusively, at the spatial resolution of the primary sensor. Both the initial collection and the subsequent interpolation or assignment can, and do, contribute an error component to the elevation estimates. The nature of error in any particular DEM will have characteristics particular to the measurement technology and the production method used. The remainder of this section describes specific, characteristic errors for several common DEMs and links error properties to particular aspects of measurement or production.
Figure 1  Example DEM errors: (a) Contour artifacts in 1” NED near Bakersfield, CA. (b) Firth Effect in older 1” NED near Niles, MI. (c) Forest artifacts in 1” NED at Howard Prairie Lake, OR. (d) Aerial imagery for the Howard Prairie Lake site.

Paper contour map sheets (or the metal plates used to print them) were commonly used to develop twentieth-century DEM products. This approach required the digitization of contour lines and spot heights, followed by interpolation to grid. Artifacts from this process are evident in hillshaded DEMs, particularly in relatively level areas. Figure 1a shows contour artifacts in a gently sloping landscape. Positional accuracy of contours may be rather low in these level areas, and on surfaces covered by vegetation; elevation errors up to and beyond half the vertical contour interval should be expected in such areas. Different patterns of errors are evident in DEMs produced via stereophotogrammetric image processing, a process by which multiple
aerial or satellite images of the ground are taken from different perspectives, common points in both images are identified, and the difference in their relative positions due to the different angle of the camera is calculated to derive heights. Elevation errors in these DEMs may result from lack of precision in the identification of ground control points, changes in film media, reflectance from the sensor, instrument error, and operator errors. An example of this latter error is depicted in Figure 1b: on older semiautomated profiling stations, elevations were often underestimated by the operator when moving upslope and overestimated when moving downslope. The resulting systematic error is called the Firth Effect.

Interferometric SAR was deployed on the space shuttle in February, 2000 and processed to generate a near-global DEM. Surface heights obtained in this way were biased upwards in areas with substantial forest canopy cover, as reflectances came from the initial surface encountered. Figures 1c and 1d depict an example in the SRTM 1” product for an area around Howard Prairie Lake in south-central Oregon. The edge of the forest northeast of the lake is clearly visible in the DEM, with sharp-edged 10–15-meter high “cliffs.” This might be considered an ontological mismatch: the DEM has faithfully recorded the reflected surface height, while the DEM user could prefer bare surface elevations. Similar vegetation-related effects are observed with LiDAR elevation data, as returns of airborne laser pulses are likely to reflect from the top of vegetation, buildings, and bridges.

**Error measurement**

Error classifications have been useful for DEM producers. A common approach identifies three fundamental types of error: blunders, which are very large in magnitude; systematic errors, which have some regular pattern or association; and random errors, which are those remaining after accounting for the others (a good description is given in Fisher and Tate (2006)). These errors may be identified through internal or external validation. Internal validation uses properties of the dataset itself or production method to flag errors. External validation is the most common method for calculating error: it compares values in the DEM to a dataset of known higher quality. Defining error as the difference between one dataset and another requires a key modeling assumption: that one conceptual model (e.g., one based on a kinematic GPS survey) is superior to another (e.g., the one underlying the DEM). The preferred model may match our understanding of elevation more closely, or it may reflect better measurement or production methodologies. The reference dataset is assumed to be a

![Figure 2](image)

**Figure 2** A study area near Medora, IN: (a) SRTM 3” DEM. (b) MODIS vegetation cover percentage. (c) Error map of SRTM 3” using NED as reference.
proxy for ground truth, and the discrepancies between its elevations and collocated DEM elevations are defined as DEM error.

A standard summary error metric is root mean square error (RMSE), which is calculated from a set of $n$ reference elevations ($z_t$) and their corresponding DEM elevations ($z_i$):

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n} (z_{ij} - z_{tj})^2}{n}}$$  (1)

RMSE is a measure of central tendency of error. It is a useful but incomplete error metric (Hunter and Goodchild 1997). While the RMSE of an elevation model may be relatively low, local errors may be much larger. Further, the sample of reference points typically comprises a small and possibly nonrepresentative fraction of the DEM and may be a poor estimate of overall error. Other global measures such as average error and 95% error are subject to similar issues.

Further, spatial variation in DEM error is important and not well described by simple global measures such as RMSE or mean error. Consider Figure 2a, depicting SRTM 3” data for a rugged region near Medora, IN, and Figure 2c, SRTM error, using NED as reference data as it is a higher accuracy source. For this example it is assumed that NED is correct and any differences between it and SRTM indicate errors in the latter. Positive errors indicate cells in which SRTM is higher than NED, while negative errors indicate cells in which NED is higher. Average SRTM error is 6.8 meters. The inner-quartile range is from 1.5 to 11.4 meters, and the RMSE is 9.1 meters. These metrics paint a very incomplete picture of the error, however. A river flows through the northern part of the region; most low and negative error in that area is located in its level, largely treeless floodplain. Large, positive errors are associated with the ridges to its south, while valleys between the ridges have much lower errors. Some regions along the ridgetops, especially in the far south of the region, have low error as well. It is visually apparent that error is not distributed in a spatially random fashion, but rather is spatially autocorrelated and associated with other landscape features such as slope magnitude and forest cover (Figure 2b).

Error modeling

In the example illustrated in Figure 2, error is known everywhere, and the SRTM for that region could be fixed by simply subtracting the observed error from each SRTM elevation value. In real applications, reference data for a region of interest are either incomplete or missing entirely, giving rise to uncertainty about the magnitude of error at most or all locations. Although exact errors may not be known for some locations, particular characteristics may be estimated from samples or inferred from properties of the DEM product. In this way an error model can be developed that captures key statistical properties of these characteristics into a stochastic model.

A measurement of some quantity $z$ at a particular location $u$ may be characterized as a random variable $Z(u)$. The probability distribution of $Z(u)$ is a model of the possible values that $z$ can take at location $u$. A random field is a set of random variables defined at locations $u$ over the spatial domain of interest. Any number of cumulative distribution functions (cdfs) can be defined for the random field; they are characterized by the joint cdfs of any subset of its random variables. In practice, it is not possible to estimate the cdf for each observation. Therefore, additional stationarity assumptions about the properties of the random field are required. A very strong one might be that all variables in the field are Gaussian, with a constant mean and
independent, constant variance. With this strong stationarity assumption, DEM errors at measured locations could be pooled and estimates of the mean and variance of the distribution could be made, and the cdfs would be fully specified with just two estimated parameters. On the other hand, most spatial phenomena, including elevation and its error, cannot be well described by this simple model with its strong assumptions. Second-order stationarity is a more flexible approach with weaker assumptions but requiring more effort to model. Relaxing the constant but independent assumption of the first model and instead assuming that joint cdfs are not a function of their absolute locations but solely of their separation vector (distance and direction) \( \mathbf{h} \), enables the pooling of point pairs separated by \( \mathbf{h} \) to develop the model. Spatial covariance is then a modeling requirement. The spatial covariance is:

\[
C(\mathbf{h}) = E\{Z(\mathbf{u})Z(\mathbf{u} + \mathbf{h})\} - E\{Z(\mathbf{u})\}E\{Z(\mathbf{u} + \mathbf{h})\} \tag{2}
\]

and an alternative statistic, widely used in geostatistics, is the variogram:

\[
2\gamma(\mathbf{h}) = \text{Var} \{ Z(\mathbf{u}) - Z(\mathbf{u} + \mathbf{h}) \} = E\{[Z(\mathbf{u}) - Z(\mathbf{u} + \mathbf{h})]^2 \} \tag{3}
\]

Both statistics can be estimated from data. Either measure can be plotted against \( \mathbf{h} \). An uncorrelated random field will have constant semivariance equal to the variance for all \( \mathbf{h} \). For spatially autocorrelated random fields like elevation and elevation error, covariance and variograms have characteristic shapes: the covariances will be high for short lags (highly correlated over short distances) and lower for long lags (decay in correlation at greater distances), while the semivariances will be low for short lags (low variance over short distances) and higher for long lags (increase in pairwise variance at greater distances).

Three example error models of increasing complexity are presented in Figure 3. The first model (Figure 3a), characterizes a spatially random Gaussian field. Errors are drawn from a Gaussian distribution with mean-zero and variance 100. A realization from this model is shown in Figure 3c, and its empirical variogram (the blue circles) closely match the model variogram. Variances at all distances are constant, indicating a lack of spatial structure in either the model or the realization drawn from it. The second model characterizes a spatially autocorrelated process. While errors are distributed normally, nearby values are expected to be similar to one another (low semivariance), as depicted by the model semivariogram, which rises at longer distances (Figure 3b). A realization from this model is mapped in Figure 3d, and shows characteristic spatial patterns with some degree of random noise.

The third example model incorporates additional environmental variables. It attempts to model error in the SRTM 3” product for a study area in southern Indiana portrayed in Figure 2, and is based on an accuracy assessment in Shortridge and Messina (2011). That research conducted an extensive analysis of North American SRTM data, and determined that SRTM slope, aspect, and vegetation canopy cover were significant predictors of SRTM error. The continental model is shown in equation 4.

\[
\text{Error} = -0.90 + 0.007 \times \text{DF328} + 1.39 \\
\times \text{Sl} - 0.015 \times \text{DF328} \\
\times \text{Sl} + 0.078 \times \text{FCC} \tag{4}
\]

where error is predicted SRTM error, DF328 is SRTM aspect measured in degrees from 328 (northwest, and perpendicular to the shuttle’s orbital path over North America), Sl is degrees...
Figure 3 Example DEM error models. (a) Semivariogram model (line) characterizing an independent Gaussian process with mean equal to zero and variance equal to 100; measured semivariogram (circles) for a realization of that process. (b) Semivariogram model (line) characterizing a spatially autocorrelated mean-zero Gaussian process. (c) Random field realization generated from the model in (a). (d) Random field realization generated from the model in (b).

of slope, measured on SRTM itself, and FCC is the forest canopy cover MODIS product, measured in percent. On national samples the model R-squared was 0.583. Residuals from the model had modest spatial autocorrelation (Figure 4a). These residuals were themselves modeled, resulting in a compound approach called regression kriging in which the expected value of error for a location depends on the output of the regression model in equation 3 and deviations from the trend are characterized by an additive spatially autocorrelated random field. Figure 4b depicts a realization from this model for the study area; compare to the actual error for this area, shown in Figure 2.

Uncertainty propagation

Uncertainty modeling is typically motivated by the pragmatic goal of the fitness for use paradigm: what is the impact of data uncertainty
MODELING UNCERTAINTY IN DIGITAL ELEVATION MODELS

Figure 4  (a) Variogram of residuals from the SRTM error regression model. (b) Realization from a regression kriging model for a 3” SRTM near Medora, IN.

on my application (Fisher and Tate 2006)? This is the domain of uncertainty propagation. DEMs are integrated with auxiliary information like a high quality dataset at a subset of locations, or a vegetation dataset, as was used in the model illustrated in Figure 4. For a subset of problems, uncertainty can be propagated analytically from the error model to the application (Oksanen and Sarkakoski 2005), but the trend has been to more general simulation-based approaches (Heuvelink, Burrough, and Stein 2007). As depicted in Figure 5, the Monte Carlo simulation approach deploys the error model to generate a set of DEM realizations in a procedure known as stochastic simulation. Each realization conforms to all information known about the actual spatial phenomenon. The realizations are then independently processed through a geospatial operation just as the original data would have been. The resulting distribution of application results may then be analyzed to determine the impact of data uncertainty for the specific application (Heuvelink 1998).

As an example, consider an application to identify significant surface depressions in a DEM. If a set of deep pits are artificially drilled in the NED DEM for the southern Indiana study site, perhaps they represent human activity on the landscape that was subsequently filled in by the time the SRTM shuttle mission was flown. Can a change detection analysis be conducted to identify the locations of these former pits by comparing SRTM to NED? A straightforward approach would be to subtract the NED raster from the SRTM raster. Locations with a large positive difference would indicate spots with a vanished pit. This difference map is presented in
Figure 6 Identifying former pits. (a) A basic SRTM–NED approach; dark red values indicate positive anomalies. (b) Propagated error approach; cells are shaded to indicate the number of SRTM realizations (of 99 generated) with lower elevations than that of NED.

Figure 6a. Pit candidates are apparent as regular dark red rectangles.

However, this approach does not account for error in the DEM. This error may be in excess of the depth of a pit, resulting in many errors of omission and commission. Indeed, on this map, the bulk of the study region is shaded red, indicating large positive differences. From the previous discussion of SRTM, it might be expected that many of these areas are forested. To account for this and other errors, stochastic simulation via the regression kriging model presented earlier was used to generate 99 error realizations (one is presented in Figure 4). Each was subtracted from the SRTM DEM to produce a realization of what the actual terrain might look like. Finally, for each cell, the NED elevation was ranked with all 99 terrain realization elevations. The output value in each cell in Figure 6b is the number of terrain realizations with an elevation lower than the NED elevation. It is visually apparent that many more pit locations are visually identifiable in this result. Much of the red “error of commission” cells in Figure 6a have been filtered out; the error model was able to reduce the influence of vegetation on the outcome of the spatial operation.

A substantial advantage of this approach is that it is completely general and can be used to identify propagation to any application. If an analyst wanted to calculate viewsheds for this region, the same terrain realizations could be used to identify the probable viewshed (Fisher 1991), accounting for error. There are disadvantages as well. Most obviously the approach is computationally expensive, both for the stochastic simulation of realizations and the additional geospatial processing on every realizations. Second, the choice of model is not straightforward. Even if restricted to a basic univariate geostatistical approach, the number of valid model forms is endless. Variogram model type, number of structures, anisotropic form, and model parameters comprise an infinite domain. More research into the development and implementation of error modeling and propagation systems is clearly needed. Nevertheless, the general Monte Carlo propagation approach demonstrated above, coupled with data-specific DEM uncertainty models, appears to be the most effective way of satisfying the fitness for use requirement.
SEE ALSO: Data quality standards; Spatial resolution; Uncertainty; Validity and verification; Visualizing uncertainty

References


Further reading


Modeling uncertainty in categorical fields

Guofeng Cao
Texas Tech University, USA

Categorical spatial data, or equivalently, area-class data, represent distribution of categories or class labels in a geographic space. The class labels can be distinguished only by their attributes (nominal, e.g., land-use and land-cover classes), ranked orders indicating whether an observation has a higher or lower value than another (ordinal, e.g., health and sick), or truncation of continuous variables by specification of cutoff values (interval, e.g., 100–200). Spatially, the boundaries of class labels may be a type of bona fides or fiat depending on the nature and origins. The former exists mainly as the result of physical discontinuities (e.g., land-cover class data), and the latter exists due to human-related actions or decisions (e.g., a choropleth map of population density in California). In categorical spatial data, complex spatial patterns occur that are similar to those observed in continuous data and can be characterized with reference to many of the same general concepts, such as spatial autocorrelation and scale effects. For example, Figure 1 shows a land-use class map of the southern Santa Barbara, CA, area, a typical example of categorical spatial data, where the color green represents forested areas, brown represents woodlands, yellow represents grasslands and gray represents urban areas. In this map, the spatial patterns of different classes can clearly be seen; from the north to the south, the forested areas tend to be adjacent to class of woodlands, woodlands adjacent to grasslands, and grasslands adjacent to urban areas. A good understanding of such spatial patterns is critical for many analyses involving categorical spatial data, such as hypothesis testing of spatial data (e.g., land-use and land-cover change detection), spatial classification (e.g., classification of remote sensing imagery), and spatial uncertainty modeling.

As with other types of geographic information, categorical spatial data are approximate representations of complex real-world phenomena. Since it is impossible to create a perfect representation of the infinite complex real world, all types of spatial data are subject to uncertainty. Spatial uncertainty basically describes the disagreement between the spatial data and the corresponding true phenomena or processes they represent. The uncertainty of spatial data has been described as the Achilles heel of geographic information sciences (Goodchild 1998), and modeling and accounting for such uncertainty in spatial analysis has long been considered as one the most fundamental yet challenging problems in spatial science. Compared with other types of spatial data, the uncertainty issue is particularly prominent for categorical spatial data and rather challenging to model. Consider the derivation process of the example map in Figure 1. New sources of uncertainty can arise in every step of the derivation process and parts of the process are rather subjective. For example, the map in Figure 1 may be generated based on remote sensing images. In addition to the uncertainty in the remote sensing images, different map-makers may use different classification algorithms and label the same area different land-use classes due to the vague definition of land-use classes. From a representation perspective, categorical spatial
Figure 1  A land-use class map of the Santa Barbara, CA, area: a typical example of categorical spatial data.
data can be represented in discrete objects and categorical fields. In the former case, the basic properties of area class can be separated into feature attributes and geometry, and, accordingly, positional uncertainty can be modeled separately from the uncertainty of feature attributes. In the latter case, however, this separation of uncertainty is generally not possible, as geometry and feature attributes are more intimately intertwined. This discussion will primarily focus on the latter case of categorical fields.

Traditionally, one of the most commonly used approaches to assess the uncertainty of categorical spatial data is to compare the data with ground truth or reference data that are believed to have greater accuracy, and report the uncertainty, or agreement with reference data, in terms of a confusion matrix or its derivatives (e.g., kappa coefficient). By summarizing the overall false inclusion and false exclusion errors of the categorical spatial data compared to the reference, a confusion matrix provides a global accuracy measure that is assumed to be invariant across the geographic extent of the data. Consequently, locality (location-dependent) information and complex spatial effects (e.g., spatial correlation and scale effects) in area-class data uncertainty are ignored in confusion matrices.

Geostatistics provide a comprehensive framework for statistical modeling of spatial data while accounting for complex spatial effects and data heterogeneities (e.g., data support and reliability). The geostatistical approach has been widely used for characterizing spatial uncertainty and evaluating the impact of uncertain data in spatial analysis and spatially explicit modeling (Kyriakidis and Dungan 2001). The basic idea is to view maps as outcomes of spatial stochastic processes (or random fields); the characteristics of spatial processes are first inferred from the (sampled) maps and uncertainty distribution across the map may then be characterized through spatial variability measures of the spatial processes or an ensemble of stochastically simulated alternative maps. Figure 2 shows a flowchart of this typical process. The most fundamental component of this process is to stochastically model spatial data, while accounting for locality and spatial effects in a spatial context. Specifically, for uncertainty modeling in categorical fields, the essential task is equivalent to estimating the joint probability mass function of a set of georeferenced categorical variables or, more intuitively, the

Figure 2 Flowchart of a typical process for modeling uncertainty of categorical spatial data and assessing the impact of such uncertainty in physics-based models.
multipoint conditional (or posterior) probability of class occurrence at a target location (where the actual class is unknown) conditioned jointly on all known source data (class labels available at sample or source locations). Based on the specified joint probability or conditional probability, stochastic simulation (Figure 2), a broadly accepted tool for investigating a stochastic process, can be conducted to characterize the uncertainty distribution regarding the spatial occurrence of class labels, for example, a specific class tends to appear at certain locations on a map. The procedure of stochastic simulation generally involves generating an ensemble of alternative class labels that are consistent with the modeled conditional probability distribution from the (sampled) maps. By conducting spatial analysis or spatially explicitly models (e.g., climatic models or ecological models) on an ensemble of alternative realizations, the uncertainty propagation in the process can be monitored and evaluated.

To illustrate the statistical modeling of categorical spatial data, consider a spatial process $C(x)(x \in \mathbb{R}^d)$ on a categorical field, and $C(x_i)$ at an arbitrary location $x_i$ can take one out of $K$ mutually exclusive and collectively exhaustive class labels. Figure 3 displays an illustrative example for a spatially indexed variable with two categories, represented by blue and red. A target location $x^*$ (question-marked location in Figure 3) with unknown state $c(x^*)$ is surrounded by sample locations, $x_1, \ldots, x_N$ with states $c(x_1), \ldots, c(x_N)$, respectively. As discussed previously, for modeling the spatial process $C(x)$, it is necessary to estimate the conditional class occurrence probability for $x^*$ given sampled labels, that is, $P\{C(x^*)|c(x_1), \ldots, c(x_N)\}$. With the class occurrence probability, maximum a posteriori (MAP) could be applied to estimate the class label for the target location. It is a typical setting for geostatistical modeling, but the discrete nature of categorical spatial data (e.g., categorical attributes, sharp boundaries, and complex geometrical characteristics) limits the application of successful geostatistical methods that have been conventionally used for continuous variables including the celebrated kriging family of methods (Chilès and Delfiner 1999).

In geographic information system (GIS) literature, Goodchild, Sun, and Yang (1992) proposed one of the earliest algorithms for error modeling of categorical data via stochastic simulation. Further refinements and other developments regarding uncertainty mapping of categorical spatial data can be found in Zhang and Goodchild (2002). In the past decade, promising progress has been made and numerous methods have been proposed from different perspectives, particularly in the field of geostatistics and Bayesian statistics. According to the underpinning principles, these methods can be generally grouped into three categories: latent variables based

---

**Figure 3** An illustrative example of statistical modeling of categorical spatial data: a target location $x^*$ (question-marked location) with unknown state $c(x^*)$ is surrounded by a collection of sampled class labels.
approaches, multipoint geostatistics approaches, and probabilistic integration approaches. The multipoint geostatistics approaches were primarily developed and are mainly applied in the field of petroleum reservoir engineering, although recent efforts have been reported to apply multipoint geostatistics in addressing problems in geographical sciences, such as improving classification of remote sensing images. The remainder of this entry briefly introduces each group of methods with an emphasis on recent developments in latent variables based approaches and probabilistic integration approaches; it then discusses the future directions.

Latent variables based approaches

Categorical spatial data can be regarded as the result of discontinuities in underlying (latent/unobserved) continuous physical processes. This concept actually creates a convenient platform for statistical modeling of categorical spatial data by an explicit introduction of latent spatially correlated variables (latent random fields) to account for spatial effects. Figure 4 gives an illustration of the latent variables based approach. Latent variables based spatial models have been an active research area in the past decades; many methods have been proposed both analytically and computationally within this paradigm. The key to latent variables based approaches is to conduct inference on the introduced latent variables from the sampled categorical values, since the latent variables cannot be observed directly.

In geostatistics, truncated Gaussian simulation (TGS) has long been used for categorical spatial variables modeling. Assuming there is a single Gaussian-distributed latent variable and it can be truncated into different class labels by a specification of threshold values, categorical spatial variables can be simulated based on the specification of a Gaussian random field (GRF). TGS is attractive as it can take advantage of numerous available GRF simulation methods and is free of the order relation problem (see section “Indicator variants of kriging” for details). Characteristics of the latent GRFs (e.g., spatial covariance functions) could be inferred via stochastic simulation that is consistent with the categorical observations. TGS only generates simulations with ordered categories due to the simplicity of connections between categorical data and the single latent variable. Truncated plurigaussian simulation (TPS), an extension of TGS, was later proposed to address this issue by introducing more than one latent GRF.

The generalized linear mixed model (GLMM) is another commonly used venue for modeling spatially correlated non-Gaussian phenomena. As in TGS and TPS, latent GRFs are introduced to account for the spatial effects, and the connections between non-Gaussian variables and latent GRFs are usually modeled by a so-called link function (e.g., a logit function is often used for spatial binary data and a soft-max function for multinomial data in contrast to a specification of threshold values in TGS and TPS; Figure 5 shows an illustration of GLMMs).
Such spatial models, coined generalized geostatistical models or model-based geostatistics (Diggle, Tawn, and Moyeed 1998), have been successfully applied to model spatial binary and count data (Diggle, Tawn, and Moyeed 1998). In model-based geostatistics, Markov Chain Monte Carlo (MCMC) sampling is often employed to infer the distribution of the latent variables. This method, however, has been criticized for high computational demand and intrinsic convergence issues. To address these issues, approximation methods of Bayesian inference have been proposed. It is worthy noting here that ancillary information (e.g., observations of other related spatial variables) can be easily taken into account in the GLMM-based models. It has been shown that there is close connection between the latent variable based (Bayesian) approach and regularization theory. By exploiting this connection, a spatial multinomial logistic mixed model has been specifically proposed for categorical spatial data modeling (Cao 2011), and, accordingly, the proposed model could be represented as a multinomial logistic function of spatial covariances between target and source locations. This method was later extended to incorporate heterogeneous auxiliary information for spatial prediction of categorical variables (Cao, Yoo, and Wang 2014).

Multipoint geostatistics

Multipoint geostatistics (MPG) is a recently proposed approach to model the multipoint conditional probability by directly scanning carefully designed training images (Strebelle 2002). The training images are deemed representative of the expected spatial patterns in the study area and can be built by experts based on their prior knowledge regarding the phenomenon under study. The central idea of the MPG paradigm is to capture multipoint statistics or spatial patterns by scanning training images using a predefined template, and to generate pattern realizations using such multipoint statistics in stochastic simulation. Most recently, a pattern-based geostatistical approach was developed based on MPG by extending multipoint statistics to a pattern histogram. MPG-based methods avoid the assumption of
an underlying random field and alleviate the computational cost associated with large scanning templates (necessary to capture large-scale features) through a pattern search tree. On the one hand, the lack of specification of an underlying random field improves flexibility and efficiency; on the other hand, it impedes further inference in categorical fields, since the underlying model (training image) is hard to parameterize efficiently and may be difficult to acquire.

Spatial cumulants have been recently proposed as alternative multipoint statistics to measure multipoint (higher-order) spatial continuity in nonlinear and non-Gaussian random fields (Dimitrakopoulos, Mustapha, and Gloaguen 2010). Through higher-order cumulants of the underlying spatial distribution, multipoint statistics can be computed in a formal probabilistic way instead of via empirical feature matching in MPG. The computation flowchart of spatial cumulants is similar to MPG but cumulants can also be applied in continuous data. Studies have shown the great promise of spatial cumulants especially in non-Gaussian fields, but formal models of spatial cumulants and analytical methods based on cumulants have yet to be developed.

**Probabilistic integration approach**

In MPG, direct modeling of the multipoint conditional probability \( P\{ C(x^*)|c(x_1), \ldots, c(x_N) \} \) may become difficult as the number of samples increases. The basic idea of the probabilistic integration approach is to decompose the complex conditional multipoint class occurrence probability into a combination of two-point (or three-point) elementary spatial continuity measures. Specifically, two-point spatial continuity between \( x^* \) and each of its neighbors \( x_n \) can be evaluated through an spatial continuity measure \( \delta(c(x^*), c(x_n)) \). The objective of probabilistic integration methods is to best approximate the multipoint conditional probability \( P\{ C(x^*)|c(x_1), \ldots, c(x_N) \} \) with a combination of \( \delta(c(x^*), c(x_n)) \) while accounting for possible spatial dependence information. It could be represented as shown in equation 1.

\[
P\{ C(x^*) = k|c(x_1), \ldots, c(x_N) \} = f(\delta\{ C(x^*) = k, c(x_1) \}, \ldots, \delta\{ C(x^*) = k, c(x_N) \}) \tag{1}
\]

Figure 6 gives a graphical example based on Figure 3, where the class label for target location \( x^* \) depends only on its six nearest neighboring observed states. Each arrow indicates a pairwise spatial interaction, quantified by a spatial continuity measure \( \delta\{ C(x^*) = k|c(x_n) \}, n = 1, \ldots, 6 \).

A key concept in this probabilistic integration approach is the choice of the spatial...
CONTINUITY MEASURE (CHILÈS AND DELFINER 1999), \( \delta(x, x') \), WHICH USUALLY QUANTIFIES SIMILARITIES OF CATEGORIES AT TWO DIFFERENT LOCATIONS. INDICATOR (CROSS-)COVARIOMS ARE THE MOST COMMONLY USED SPATIAL CONTINUITY MEASURES IN INDICATOR VARIANTS OF KRIGING. MOST RECENTLY, SPATIAL TRANSITION PROBABILITIES (CARLE AND FOGG 1996; LI 2006) AS WELL AS BIVARIATE JOINT PROBABILITIES (ALLARD, D’OR, AND FROIDEVAUX 2011) HAVE BEEN PROPOSED AS ALTERNATIVE SPATIAL CONTINUITY MEASURES TO INDICATOR (CROSS-)COVARIOMS.

ONE OF THE SIMPLEST FUNCTIONS \( f \) IS THE WEIGHTED LINEAR COMBINATION OF THE SPATIAL CONTINUITY MEASURES (EQUATION 2):

\[
P\{ C(x^*) = k | c(x_1), \ldots, c(x_N) \} = \sum_{n=1}^{N} \lambda_n \delta\{ C(x^*) = k, c(x_n) \}
\]  

WHERE \( \lambda_n \) REPRESENT WEIGHS FOR \( \delta\{ C(x^*) = k, c(x_n) \} \) TO ACCOUNT FOR SPATIAL DEPENDENCY BETWEEN \( x_n \) AND ITS NEIGHBORS. INDICATOR VARIANTS OF KRIGING METHODS INCLUDING INDICATOR KRIGING (IK) AND INDICATOR COKRIGING (ICK) (CHILÈS AND DELFINER 1999), AND TRANSITION PROBABILITY BASED INDICATOR GEOSTATISTICS (CARLE AND FOGG 1996) CAN BE SEEN AS SPECIAL CASES OF THIS ADDITIVE FUSION MODEL.

ANOTHER SIMPLE FORM OF \( f \) IS THE MULTIPlicative FUSION, WHICH EXPRESSES THE MULTIPPOINT CONDITIONAL PROBABILITY IN A MULTIPlicative FORM OF SPATIAL CONTINUITY MEASURES IN CATEGORICAL FIELDS (EQUATION 3):

\[
P\{ C(x^*) = k | c(x_1), \ldots, c(x_N) \} = \prod_{n=1}^{N} \delta\{ C(x^*) = k, c(x_n) \}^{\varepsilon_n}
\]

WHERE EXPONENT \( \varepsilon_n \) REPRESENTS A WEIGHT FOR \( \delta\{ C(x^*) = k, c(x_n) \} \) TO ACCOUNT FOR SPATIAL DEPENDENCY BETWEEN \( x_n \) AND ITS NEIGHBORS.

THIS MULTIPlicative FORM IS USUALLY A RESULT OF BAYES’S FACTORIZATION RULE. THE RECENTLY PROPOSED MARKOV CHAIN RANDOM FIELD (MCRF) (LI 2007) AND BAYESIAN MAXIMUM ENTROPY (BME) BASED METHODS (ALLARD, D’OR, AND FROIDEVAUX 2011) CAN BE TAKEN AS SPECIAL CASES ALONG THIS MULTIPlicative FUSION MODEL.

### Indicator variants of kriging

AS EFFORTS ARE MADE TO ADAPT THE KRIGING FAMILY OF GEOSTATISTICAL METHODS FOR CATEGORICAL VARIABLES, IK IS PERHAPS THE MOST FREQUENTLY USED METHOD FOR ESTIMATING THE POSTERIOR (CONDITIONAL) PROBABILITY OF CLASS OCCURRENCE AT ANY TARGET LOCATION. AS FOR OTHER KRIGING VARIANTS, IK INVOLVES FITTING INDICATOR VARIOGRAMS AND THEN ESTIMATING THE PROBABILITY OF OCCURRENCE OF EACH CLASS INDEPENDENTLY VIA A KRIGING SYSTEM. IK CAN ALSO HAVE MANY VARIANTS DEPENDING ON ASSUMPTIONS OF THE LOCAL MEAN VALUES OR CLASS PROPORTIONS OF CLASS OCCURRENCES.

BASED ON IK, SEVERAL VARIANTS HAVE BEEN DEVELOPED TO IMPROVE THE PREDICTION ACCURACY OF PRIMARY CATEGORICAL VARIABLES. ICK, FOR EXAMPLE, IS A NATURAL EXTENSION OF IK FOR MULTIVARIATE CASES (CHILÈS AND DELFINER 1999), IN WHICH AUXILIARY VARIABLES ARE INCORPORATED INTO THE PREDICTIVE PROCESS VIA (CROSS-)COVARIANCE FUNCTIONS OF PRIMARY CATEGORICAL VARIABLES AND AUXILIARY VARIABLES. PRACTICAL APPLICATIONS OF ICK, HOWEVER, ARE CUMBERSOME OWING TO A NUMBER OF (CROSS-)COVARIANCE FUNCTIONS (OFTEN THROUGH THE LINEAR MODEL OF COREGIONALIZATION) HAVING TO BE JOINTLY FITTED. WHEN AUXILIARY VARIABLES ARE LINEARLY RELATED TO THE CLASS OCCURRENCE OF PRIMARY CATEGORICAL VARIABLES, THEY CAN BE INCORPORATED INTO AN IK SYSTEM AS DETERMINISTIC LINEAR FUNCTIONS (NONSTATIONARY MEAN). THIS IS REFERRED TO AS INDICATOR KRIGING WITH EXTERNAL DRIFT (IKED) (CHILÈS AND DELFINER 1999), WHOSE IMPLEMENTATION IN PRACTICE IS CHALLENGING, AS IT IS
often problematic to simultaneously estimate the parameters of external drift and the covariance function of the stochastic component. As a hybrid method of kriging and multiple regression models, regression-kriging (RK) has been developed to combine a regression of the dependent variables on auxiliary variables with kriging of the regression residuals. An indicator variation of RK, regression-kriging of indicators (RKI), has been proposed for categorical variable and this method has evolved into regression-kriging of memberships (RKfM) by substituting crisp indicator values with continuous membership values. Most of these IK-based methods, however, share the inherent problems that the original IK suffers from: the probabilities of occurrence are not guaranteed to be between 0 and 1 (e.g., IK, ICK, RKI), the sum of the predicted probabilities may not be equal to 1 (e.g., IK, ICK, RKI, and RKfM), and the outcome values of conditional cumulative distribution function may not be monotonic (order relation problem). A posterior correction of the resulting conditional probabilities is often necessary either through a Gaussian transformation or via a logistic regression model.

### Transition probability based geostatistics

The concept of transition probability is not new but, until recently, it has been regarded as a spatial continuity measure in categorical fields. Similar to the conventional definition of transition probability, spatial transitional probability $\pi_{k'k}(h)$ is defined as the transition probability from class label $k$ to $k'$ over a spatial lag $h$:

$$\pi_{k'k}(h) = P\{C(x + h) = c' \mid C(x) = c\}.$$

The transiogram is typically defined as a parametric diagram of transition probabilities as a function of a spatial lag $h$ (Li 2006). Similar to variogram or covariance models, sample transiograms can be collected by directly scanning sample data, and transiogram parameters can then be fitted from sample transiogram values. There is a relationship between the parameters of transiogram models and spatial distribution of class labels (class proportion, mean length, compactness, and class juxtaposition) (Carle and Fogg 1996; Li 2006; Cao 2011). This connection provides an interpretation of the behavior of transiogram models, and more importantly, offers a guideline for incorporating expert knowledge on the categories distribution in the construction of transiogram models.

Compared to indicator covariances and variogram models, transiograms are more interpretable in categorical fields and easier to integrate with ancillary information (Carle and Fogg 1996). Based on this concept, Carle and Fogg (1996) reformulated IK and ICK as systems of spatial transition probabilities according to their analytical connections with indicator covariograms. As with variograms in kriging systems, not every function of distances can serve as a valid transiogram for the reformulated system. It has been found that in most cases the exponential form and several of its variants are eligible for transiogram modeling (Cao 2011).

Based on the concept of spatial transition probabilities, the idea of a Markov chain has been applied for categorical data modeling in spatial settings. Most recently, MCRF has been proposed (Li 2007) by assuming that a single spatial Markov chain moves or jumps in a space and the state at any uninformed location is specified by a multipoint conditional probability conditioned on its its nearest neighbors and its last stay location. Under the conditional independence assumption, the multipoint conditional probability could be factorized into a multiplicative combination of (two-point) spatial transition probabilities.

From another statistical perspective, a BME approach has been used to address the statistical
modeling of categorical spatial data (Bogaert 2002). BME is based on a joint multidimensional multinomial assumption in the desired categorical random field, and estimates parameters (actually joint probability tables) by a nonsaturated log-linear model of main effects and interaction effects under certain marginal constraints (most often formulated in terms of bivariate joint probabilities). Most recently, a simplified variant of the BME approach (Allard, D’Or, and Froidevaux 2011), namely Markovian-type categorical prediction (MCP), was proposed to alleviate the heavy computational cost of categorical BME. For each target location, instead of modeling all possible interactions (in terms of joint probabilities), this simplified approach only considers pairwise spatial interactions between the target location and its neighboring known source data (in a perspective of multipoint conditional probabilities).

Despite the background differences between MCRF and MCP, both of these approaches were designed to decompose the difficult-to-get multipoint conditional probability into a combination of two-point probabilities, based on the assumption of conditional independence. The assumption of conditional independence has been widely used in the field of Bayesian statistics and, despite its simplicity, practice has shown that this assumption performs quite well. However, in the presence of actual data dependence when an independence assumption is difficult to corroborate, it tends to lead to nonrobust and inconsistent results (Journel 2002). A robust alternative is the assumption of permanence of ratios, which is based on the engineering paradigm that ratios (in this case, odds ratios of spatial transitional probabilities) are typically more stable than their components (Journel 2002). It has been applied to relax the assumption of conditional independence in categorical spatial data modeling (Cao 2011).

Future directions

In past decades, much has been learned about how to model spatial uncertainty in categorical fields and how to evaluate uncertainty propagation in spatial analysis and modeling. The majority of research tends to view categorical fields as static and focuses on the investigation of spatial effects in categorical fields. Geographic phenomena are essentially dynamic. With the advancement of spatial data acquisition and dissemination technology, temporal information of categorical spatial variables as well as other ancillary information (e.g., measurements of related spatial variables) is typically available. Incorporation of such information is expected to improve our understanding of the change dynamics in categorical fields. While much progress has been made in general spatiotemporal statistics, efficient and generally accepted theory for change detection and hypothesis testing in categorical fields that can effectively account for complex spatial effects (or geometry) are still largely lacking.

From a computational perspective, unlike simple models such as the Gaussian, models for categorical spatial data usually involve a large number of parameters, and inference of such models could become very computationally intimidating. In the past decade, computational science and technology has been dramatically advanced and provides a transformative paradigm for scientific research. Great progress has been made in the past few years to solve computationally intensive spatiotemporal problems by exploiting advanced high-performance computing capabilities, for example, the emerging CyberGIS framework (Wang 2010). It is still rare in the literature of uncertainty modeling to take advantage of this advancement.

Despite the progress that has been made for uncertainty modeling in categorical fields, a
large portion of the proposed methods is comparatively inaccessible, particular for domain communities outside of statistics and geographic information science (GIScience). This is probably due to the complexity of the involved mathematics and computing technologies, but it is, nevertheless, important to make the tools accessible to domain scientists. The rise of open source and the cyberinfrastructure paradigm in the scientific research community provide a promising opportunity to further increase the awareness of the uncertainty problem in categorical fields, and improve the adoption of existing methods and tools to address this problem.

**SEE ALSO:** Environmental uncertainty; Geostatistics; Modeling uncertainty in digital elevation models; Models and simulation in biogeography; Spatial analysis; Uncertainty; Visualizing uncertainty

**References**


Ecologists and biogeographers grapple with understanding the dynamics of systems that operate across an extended range of scales in time and space, from the almost instantaneous (e.g., physiological processes such as photosynthesis) to the millennial (e.g., range shifts in response to climate change). Despite the ever-expanding range of environmental proxies and data available to describe these dynamics, some scale-related issues are simply intractable for empirical approaches. In such settings modeling, and in particular simulation modeling, has become an important and widely used tool by biogeographers.

Models are simply representations of some phenomena (system or entity) of interest. This means that they are, necessarily, abstractions and are also, therefore, false. There have been many attempts to classify models on the basis of their implementation. A typical such classification would recognize mathematical, empirical, and simulation models as distinct approaches. Mathematical models represent systems using the formalisms of mathematics – the classical reaction–diffusion models of spread and invasion are archetypal examples. Such models are often abstract and their qualitative insights of more interest than their exact quantitative predictions. Mathematical models are typically top-down and seek to explain the behavior of broad classes of systems (e.g., free-living populations) rather than being site- or taxa-specific. Conversely, empirical models start with observations (data) and seek recurrent patterns in them – they are probably the most widely used of all models by biogeographers. Species distribution models, in which correlative relationships between environmental conditions and taxonomic distributions are established and then extrapolated across space and time, typify empirical models. The oft-repeated critique of such approaches is that they lack structure and causality, so have limited explanatory power beyond the specific conditions (scale, location, taxa, etc.) they were developed for. While this is a valid concern, modern model-building approaches, such as multiple model inference, are also explicitly concerned with hypothesis testing. The third broad class of models is simulation models. This class encompasses a broad range of approaches, all of which are concerned with the in silico representation of some system of interest; simulations typically contain elements of both mathematical and empirical models. A type of simulation that has received much attention among ecologists is individual-based models (IBMs) (Grimm and Railsback 2005), in which every entity of interest (e.g., trees in a forest, individual baboons in a troop) is represented and their behaviors and ultimate fate (e.g., growth, reproduction, mortality) tracked. In such bottom-up frameworks the broad-scale properties of the system arise from individual-level interactions – so, for example, shifts in composition across succession emerge as a result of interactions, such as facilitation and competition, between individual trees and their environment.

Another, perhaps more useful, way to think about classifying models is on the basis of their purpose. A simple, perhaps simplistic, dichotomy is
between modeling exercises focused on predicting the behaviors of the systems they represent and those focused on learning about them. Of course, these two goals are far from mutually exclusive. The purpose of a given modeling activity will determine where it lies in terms of fundamental tradeoffs, such as those between model simplicity and complexity, the level of mechanistic detail included, and the scale (grain versus extent) at which a model will operate. Simulation models span the full range from simple (but hopefully not simplistic) high-level generalized abstractions through to highly detailed mechanistically-grounded representations of specific locations, processes, and taxa. It is perhaps tempting to see the latter as more realistic and, hence, preferable – after all, detailed models contain more of the “real” world – but this realism comes at a cost in terms of interpretability and uncertainty in both model structure and parameterization. Modern approaches to model development, such as pattern-oriented modeling, aim to identify model structures that are sufficiently realistic to be useful but are not overburdened with unsupported detail. Unsupported detail acts both to increase model uncertainty (both in terms of parameterization and also in terms of representing poorly understood processes) and to inhibit the development of understanding of the system and processes of interest. In short, distinguishing between the processes that need to be included in a given model and those that can be omitted reduces to a tradeoff between realism and generality (as discussed in the classic paper by Levins (1966)). An important question given growing increases in computational power and data is (Held 2005 considers this tension): “Does it matter if we do not understand our simulation models?”

Such fundamental questions become even more difficult when we seek to explicitly represent humans and their behaviors in our models. As the role of humans in most ecological systems has grown in significance, there has been increasing recognition that models need to directly represent human activity and decision-making. It is not adequate to simply treat humans as boundary conditions or represent them as just another exogenous forcing factor, nor can the reciprocal nature of feedbacks between humans and the environment be ignored. Adequately representing human agency is particularly important, for example, when processes such as land-cover dynamics in peri-urban environments are being considered or, more generally, when socio-economic processes are stronger drivers of ecological change than biophysical ones. A framework that has proved particularly useful in such contexts is agent-based models (ABMs) (Heppenstall et al. 2012), which are closely related to the IBMs described earlier. Agents are goal-seeking entities whose behaviors are plastic and who interact in a reciprocal way with the environments they inhabit. The agents’ behavior is not necessarily rational, but may be designed to represent both the vagaries of the decisions that people make and the uncertainties involved in such decision-making.

There is growing interest in using coupled ecological and agent-based models to represent the dynamics of socioecological systems. Once any model is implemented it is important to evaluate its “usefulness.” While there is no doubt that the “quality” of a model’s outcomes need to be closely scrutinized, it is not necessarily obvious what quality means, with mimetic accuracy, heuristic usefulness, and/or the ability to provoke dialogue all among the qualities different evaluation frameworks emphasize. It is also important to distinguish between evaluations that focus on the implications of uncertainty in a model’s parameters (e.g., sensitivity analysis) while (usually) treating model structure as fixed versus those that seek to understand how a model’s structure (decisions about representation) affects
its outcomes (e.g., structural analysis). Arguably, structural uncertainty is at least as, if not more, important than parameter uncertainty but it rarely receives as much attention. Standard tools of model analysis, such as sensitivity evaluation and statistical measures of model predictive capacity, are useful but many other approaches have been developed for model evaluation and tools originally designed for statistical modeling are now being used for simulation models (Hartig et al. 2011). Ultimately, as with model design, what constitutes appropriate model evaluation depends on the context in which a model is being used—that is, its purpose.

In any context, understanding the role and usefulness of models and modeling hinges on carefully isolating the purpose of the exercise. There is no doubt that ever-growing computational power, coupled with the advent of “big data,” crowd-sourced information, and citizen science will enable models to explore issues that were previously technically limiting. On the other hand, to be truly useful models, of any sort, need to be more than highly detailed, but potentially inscrutable, representations of the system of interest. Finding the middle ground between the overly complex and the overly simple remains the fundamental challenge for all modelers.

SEE ALSO: Agent-based modeling; Biogeography; Complexity in biogeography; Ecosystem

References


Further reading


Geomorphological models are representations of the external reality of Earth surface processes and landforms. A primary distinction can be made between (i) conceptual models, (ii) scale models or analogs, and (iii) quantitative models (e.g., Huggett 1985).

Conceptual models are verbal or visual representations of a system or process, ranging from a sequence of relationship statements or a “story,” to flow diagrams and schematics, often in a graphical or cartoon form. Many historically (in)famous geomorphic models, and indeed paradigms, are of this kind, such as Davis’s (1899) “Geographical Cycle,” often represented as a sequence of landscape cartoons, or the general systems framework championed by Chorley (1962) and Strahler (1980), usually drawn as flow schemes resembling electronic circuitry diagrams. This type of modeling is referred to as an activity of conception in Haggett and Chorley’s (1967) taxonomy, representing a geomorphic system as an integrated whole synthetically (e.g., as a flow scheme), or as a black box model where the precise details of processes or constituent parts are not further defined.

Scale models and analogs aim to represent a geomorphic system on a (sometimes reduced) physical or temporal scale so as to replicate or mimic the processes and forms in an observable and reproducible form. This includes hardware models or physical analogs such as water flumes, wind tunnels, and wave tanks, where processes or properties are scaled in terms of their geometry (sizes), kinematics (velocities), or dynamics (forces), relative to the real-world system. An example of this is the investigation of subaqueous bedform development on moveable sand trays in small water tanks as an analog for understanding desert dune formation in the atmospheric boundary layer. This type of modeling relies heavily on the equivalence of dimensionless similarity parameters, such as the Reynolds number or the Froude number (e.g., White 1996). Haggett and Chorley (1967) refer to this type of scientific activity as translation, and it includes natural analogs as models: observable real-world systems that exhibit a behavior or characteristic that is similar in kind to (and thereby an explanation of) another geomorphic system, that is, using a known geomorphic phenomenon from a historical time sequence or from within a spatial hierarchy (e.g., a landform unit) as a model to understand another landscape or subset.

Quantitative models, lastly, quantify features, objects, and processes of a geomorphic system numerically as variables, parameters, and coefficients, representing the relationships between them as a collection of mathematical expressions. Such models range from single equations, such as the Universal Soil Loss Equation (USLE), to coded algorithms of varying levels of complexity running on computers. Quantitative models are usually developed out of initial conceptual models and often (partly) parameterized or tested against scale models and analogs. Malamud and Baas (2013) recognize three key distinctions in contemporary quantitative models of geomorphology: (i) traditional physically based computer models, (ii) cellular automata models, and (iii) statistical models of observations or simulated data.
MODELS IN GEOMORPHOLOGY

The first category of quantitative model, traditional physically based computer models, is grounded on an approach of what Haggett and Chorley (1967) refer to as dissection, to simulate geomorphic systems within a reductionist-deterministic context of quantifying small-scale physical processes, based on exact empirical or theoretical relationships (equations). The system is simulated within an explicit and discretized spatiotemporal domain (1-D, 2-D, or 3-D + time) and usually underpinned by the Exner equation, or the conservation of mass relating spatial gradients in sediment transport flux to topographic height changes, together with a diffusion- or advection-type transport forcing process. These are the most typical modern computer simulation models of geomorphic change, spanning a wide range of environments, including hillslope erosion by overland flow (e.g., WEPP, EUROSEM, MEDALUS), landscape evolution at a catchment or even continental scale (e.g., LAPSUS, SIBERIA), fluvial systems (meandering, braiding, river terrace deposits, etc.), and eolian dune development by wind. The detailed representation of small-scale physical processes usually involves a large number of parameters, complete definition of boundary conditions, and often a separate fluid-flow simulation component (e.g., for channel flow or surface wind), either as a complete Computational Fluid Dynamics solution or as a simplified flow theory, that require considerable computational resourcing.

The second category, cellular automata (CA), simulates geomorphic processes as a simplified set of rules of interactions between cells within a discretized spatiotemporal domain. The repeated application of interaction rules and corresponding changes in cell states over time between a multitude of elements produces larger-scale patterns, landforms, and collective behaviors that emerge as self-organizing phenomena. The number of parameters is usually small and the models can be run on ordinary computer systems. This type of modeling is also sometimes referred to as “reduced-complexity” modeling, and may also include Agent-Based Modeling (ABM) approaches (as yet rarely applied in geomorphology), where mobile agents such as animals move over a cellular domain interacting with each other and the domain environment. A key objective (and strength) of CA modeling is the identification and distillation of the most important, minimally necessary and sufficient, set of processes or rules (in their most fundamental form) required to reproduce certain geomorphic phenomena. They are often used to test or develop theories and hypotheses and produce generalized outcomes in terms of emergent patterns and behaviors, rather than customized site-specific projections. Typical examples of CA in geomorphology include sand dune models (Werner model, DECAL), models of river meandering (e.g., CAESAR), river braiding, coastal sand spit patterns, and landslide and forest fire models. Fonstad (2006) provides a comprehensive overview of CA modeling in geomorphology, as well as key issues and challenges.

The third category, statistical models, provides quantification and representation of relationships and structure found in data obtained from real-world geomorphic systems (or also sometimes in simulated data), without necessarily explaining the underlying processes or mechanisms. These models range from descriptive analysis, such as frequency–size distributions of events and processes such as landslides, to statistical tests that identify apparent relationships between variables (e.g., hypothesis testing and correlations), and the extraction of statistical trends in space and time (e.g., time-series models, fitted landscape surfaces). This category may also include sophisticated data models that use...
MODELS IN GEOMORPHOLOGY

statistical measures to optimize the representation of a geomorphic process or outcome, such as artificial neural networks (ANNs) and genetic algorithms (GA), where the internal realization of the working algorithm is essentially inscrutable (i.e., a black box situation). A GA has been used, for example, to optimize the parameter set of a traditional physically based simulation model of streamflow and sediment response in a river catchment (SWAT) calibrated to a specific real-world river basin.

The above gradation in quantitative geomorphic modeling approaches, in terms of the degree of specificity in representation as well as the differences in objectives, echoes a similar categorization by Dietrich et al. (2003). Within the realm of quantitative landscape modeling they propose four types of approaches or objectives: (i) detailed realism, which represents the – mostly unattainable – objective of exhaustively representing a geomorphic system with its unique mixture of site-specific details and general trends and processes operating at multiple scales of time and space – this type of modeling requires an unrealistic degree and detail of knowledge and data, but it is possible to reach detailed realism for some carefully circumscribed features or subsets of a landscape given sufficient information; (ii) apparent realism, which is the somewhat illusive quality of most large-scale spatially explicit landscape simulation models that generate detailed landforms or outcomes based on a distillation of key geomorphic processes on a coarse spatial resolution, but that should rightly only be considered as metaphors or qualitative representations of the real world; (iii) statistical realism, discussed by Dietrich et al. specifically within the context of self-organizing patterns and emergent behaviors, relating to the application of fundamental physical principles, such as minimized entropy production, and the general patterns and trends they produce that may be recognized across various different landscapes; and (iv) essential realism, which claims that geomorphic sediment transport laws can be identified and parameterized from empirical evidence and that model outcomes can be tested against field measurements – this type of modeling aims to understand differences between landscapes, as opposed to the search for commonalities across landscapes of statistical realism, and produces quantifiably testable outcomes, as opposed to the qualitative or metaphor representations of apparent realism.

A more extensive range of modeling approaches in geomorphology is distinguished by Odoni and Lane (2011), who place model types in a hierarchy of increasing complexity, but also recognize a differentiation between data-driven approaches and theory-driven approaches, as well as the interchange of data as inputs and outputs between these. On the data-driven branch, they recognize a sequence of modeling types starting at a basic level with data models, as capture and representation of empirical data (e.g., maps), moving on to hardware models, similar to the scale or analog models described previously, qualitative models, that is, conceptual representations, analytical models, mathematical analyses in continuous space and time, and finally, at the complex end, stochastic models, which provide probabilistic solutions. On the theory-driven branch, Odoni and Lane recognize a sequence starting at a basic level with theoretical models, based on logic and deduction, moving on to simulation models, as basic abstractions of some kind, quantitative models, relationships and properties as equations, numerical models, discretized time and space with computational techniques, and, at the complex end, deterministic models, which provide fully complete and repeatable outcomes. This theory-driven branch represents a progression of increasing degrees of specificity in inputs, parameters, boundary conditions, and outputs.
MODELS IN GEOMORPHOLOGY

There are many other model taxonomies (Odoni and Lane (2011) provide a comprehensive list of sources), and many different versions of classification have developed over the years, both in the past (e.g., those by Chorley (1967) and Huggett (1985), mentioned above) and more recently (Hooke 2003; Mulligan and Wainwright 2013). Another way of exploring attributes and underpinnings of different geomorphic models is to consider their characteristics along dimensions of dichotomy (e.g., Haggett and Chorley 1967; Slingerland and Kump 2011), such as:

- **static versus dynamic**: models that represent “equilibrium” features or structures as opposed to models that focus on processes and changes over time;
- **descriptive versus normative**: models that are concerned with a stylistic or simplified description of reality as opposed to models that attempt to predict outcomes under certain conditions;
- **stochastic versus deterministic**: models that represent aspects of a geomorphic system in terms of probabilities and statistics as opposed to models that quantify specific precise values;
- **process-based versus form-based**: models that simulate physical processes that are thought to operate in reality as opposed to models that represent shapes and forms found in the geomorphic landscape;
- **forward versus inverse**: models that attempt to predict a final or future state of a geomorphic system as opposed to models that aim to determine past initial or boundary conditions based on a current state;
- **black box versus white box**: models that use “invisible” or untraceable internal procedures (often involving advanced statistical methods) as opposed to models where all processes and relationships are transparent and precisely specified;
- **inductive versus deductive**: models that represent a system by generalizing and categorizing from empirical findings as opposed to models that assume a theory or framework to simulate consequent forms and processes.

Besides the characteristics and methods underlying different geomorphic models, another important question concerns what makes for a good model. Kirkby (1996) argues very precise criteria for a good model, stipulating four requisite attributes: (i) a physical basis, that is, that the selection of form/process equations, even if only in a black-box-style regression approach, is informed by some level of insight of the physical reality concerned, (ii) simplicity, that is, that the model includes only the necessary and sufficient number of processes and parameters for representing the geomorphic system, (iii) generality and richness, meaning that the model formulation is sufficiently general and versatile to allow application to a variety of areas and situations, and (iv) potential for scaling up and down, meaning that the model can be aggregated to represent geomorphology at a larger scale. There are some likely tensions between these criteria, however. The demand for generality and richness, for example, may require an extended number of input parameters and boundary conditions to enable representation of a wide range of situations, but this may conflict with the desire to keep the model as simple as possible. Pilkey and Pilkey-Jarvis (2007) argue on this score that many environmental models are bad, and indeed dangerous for end users, because they are too complex, with too poorly constrained parameters, yielding (unknowably) inaccurate predictions. Slingerland and Kump (2011) therefore maintain that a simple but properly validated model is much preferable over a more complex but less well-constrained model.
The meaning of the term “good” in a modeling context may involve various other considerations, ranging from practical utility, or the delivery of results that are useful and relevant to end users, through representation, or the synthesis and integration of concepts and understanding that help to consolidate knowledge, to inspiration, or the discovery of links and insights that lead to new research questions. A stochastic black box river flow discharge model, for example, may not offer much inspiration for novel research questions, nor a good representation of scientific understanding (and it would not meet Kirkby’s criteria either), but if it can predict a 50-year flood very accurately it is undoubtedly of good value to a water resource manager. Reflections on utility and representation are usually extensively considered in the context of model validity (Oreskes, Shraderfrechette, and Belitz 1994) and prediction (Wilcock and Iverson 2003), but the opportunities for inspiration are often only implicitly articulated. Some geomorphic models, such as reduced-complexity or CA models, however, can have a great capacity for inspiring new insights and research questions, even though they may not produce very accurate predictions for practical end users, nor represent particularly comprehensive integrations of detailed process-based understanding. Murray (2013), for example, argues that even though many researchers prefer models that resolve fine-scale physical processes and parameterizations, for simulation of large-scale landscape development and pattern formation emerging from sediment transport and biology interactions, for example, it may not be possible nor fruitful to pursue rigorous conservation of momentum principles at the small scale.

A final consideration for evaluating the merits of geomorphic models is their potential beauty or elegance. Some models may be considered beautiful because they combine, synthesize, and abstract a large and complex body of knowledge into a pinnacle of simplification, built on extensive foundations from various (often disconnected) areas of research. The opportunity for such beauty in geomorphic models is perhaps limited, but other ways of assessing elegance may be in terms of net change in information content (or Shannon entropy) between data inputs and model outputs, or in the formulation of the shortest, most efficient, algorithmic description (low Kolmogorov complexity) of a geomorphic system.

Models in geomorphology have a very long and rich history, with many reflections on their methods and merits, as the Further reading illustrates. Recent developments and future trends in geomorphic modeling include a widespread recognition and inclusion of biological components and controls (eco-geomorphic modeling), the pooling of expertise and resources in collective community modeling efforts, and the exploration and utilization of big data opportunities.

**SEE ALSO:** Agent-based modeling; Big data; Cellular automata; Geomorphic systems; Geomorphology: history; Geophilosophy; Hydrologic flow models; Parallel computing; Quantitative geomorphology

**References**


MODELS IN GEOMORPHOLOGY


Further reading


Modernity

Philip Howell
University of Cambridge, UK

“Modernity” periodically falls in and out of fashion. Its critics object to it as a portmanteau term that mixes up a range of phenomena, many of which are not especially new. Bruno Latour, sociologist of science and avowed contrarian, famously argues that *We Have Never Been Modern* (Latour 1993). In this view, modernity is no more than a matter of faith, a kind of “constitution” that draws fictitious divisions between object and subject, nature and society, premodern and modern. The whole idea should now be recalled, he has since gone on to say, much as one might recall a defective product. All the same, modernity continues to be a helpful, indeed essential, concept, and at least some objections can be met by keeping strictly separate modernity’s focus on technology, discourse, and experience. It is in this tripartite understanding of modernity’s multiple registers that its utility can be preserved, and the varied work accomplished by geographers under its sign highlighted.

First, for modernity as project, key technological innovations – steamships, railroads, and the electric telegraph are iconic – represent significant and non-negotiable advances in efficiency. They accompany a host of other putative features of modernizing societies – the growth of rationalism, individualism, secularism, scientific consciousness, and so on – that may be thought of as at least potentially liberating. Modernity as project does not have to be progressive in a political and ethical sense, however; technical advances were key “tools of empire,” for instance vital supports for the Western subordination of other peoples and cultures. Technologies and rationalities of modernity have also been linked to the greatest tragedies of the modern age, the Holocaust among them. Nevertheless, the sense of these developments as demonstrable, decisive, and unrecallable advances is crucial. The railroads, although they carried Jews and others to their deaths in the camps, also “annihilated space,” promoting mobility, communication, and trade. They were the age’s greatest symbol of modernity, a revolutionary rupture foretelling the wholesale industrialization of time and space. In the philosopher Charles Taylor’s (1995) words, they represent a “culture-neutral” definition of modernity. These technologies, and the associated phenomena of a rationalized, industrial, capitalist civilization, may have been foisted on the world by an increasingly powerful Western world, but they are not necessarily defined by that culture or tradition. It is impossible completely to divest ourselves of the idea that modernity, albeit in limited, carefully specified ways, represents calculable improvement, however incomplete and partial. David Harvey’s (1990) famous geographical thesis concerning time-space compression acknowledges these revolutionary technical achievements while decrying their role in the reorganization of capitalism.

These seemingly unimpeachable advances must be qualified by an understanding of the ways in which technology is embedded within particular cultures as well as technopolitical regimes. Charles Taylor noted that we cannot rely on “acultural” interpretations of modernity alone, for we would miss the ethnocentrism that,
for all this increased efficiency and economy, is installed at the heart of “modernity.” We may consider, then, in the second guise of modernity as discourse, the eminently contestable achievements of what have been often dubiously promoted as “modern.” So, while we might reasonably consider the railway an advance over the horse and carriage, we cannot make the same argument for, say, a Western-style suit in the colonial tropics as opposed to a loose-fitting robe like the Arabic *djellaba*. In terms of practicability, the latter, in its places of origin, is clearly preferable, but in the age of empire it still became the icon of a worn-out tradition as opposed to an up-to-the-minute modernity. In this discursive sense, then, modernity is no more than a claim to superiority, based on a particular provincial culture rather than on demonstrable rationality. In this geographic paradigm of modernity, Europe came to know itself, typically through the cultivation of superiority over other cultures, as the historical geography of “colonial modernity” shows (such as the role of mapping in the global, imperial gaze of modernity).

Modernity is both culture-neutral and culturally specific, then. But there is an additional element that needs to be recognized. This is the impact of modernity as project and modernity as discourse on ordinary as well as elite people’s consciousness and identity. For modernity as experience a whole series of lived phenomena in “modern times” have been put forward as part of the condition of modernity. In a justly famous analysis, Marshall Berman (1982) argued that the maelstrom of modern life opened up a contradictory and ultimately tragic tension between the possibilities and perils of the emerging modern world. Attempts to delineate this “structure of feeling” descend into cartoonish caricature when abbreviated, but the poet Baudelaire, in his well-known essay of 1863, “Le peintre de la vie moderne” (“The Painter of Modern Life”), helpfully puts the stress on the role of “the transitory, the fleeting, the fortuitous” in the life of the modern metropolis. This “lived modernity” has often attracted scorn from critics (by way of diagnoses of alienation, bureaucratization, commodity fetishism, narcissism, and so on). The sheer captivating excitement of those subjected to modernity’s thrilling disorientation should be registered, however, and it is captured in the aesthetic response to modernity’s pathos and delight. No less important is the mundane and complacent acceptance of modernity’s blandishments and material satisfactions.

The wider point is that myriad anthropologies of modernity are possible, often with the city as the crucial locus of concern for this mode of vital experience. The relationship between the making of the modern self and the modernization of urban environments and spaces has been an important theme for geographers (see, e.g., Dennis 2008). So, too, are the ways in which different groups of people (women, gay/queer men, immigrants, and people of color, say) responded to the challenges of modernity; here, the larger ethical dimension of modernity, its promise of individual as well as collective autonomy, is partially endorsed but also necessarily qualified, sometimes sharply.

What, then, is “modernity?” It is a concept that attempts to capture the acknowledged changes in technology (the growth of technical capabilities; scientific advances; the organization of economic, social, and political realms), the power of discourse (the normative, ideological, and aesthetic accompaniments of such “progress”), and, more simply, people’s experience of these destabilizing forces, operating from the sixteenth century but accelerating from the nineteenth. The tripartite schema suggested in this definition is far from exhaustive, and multiple modernities may be advanced. But it is worth emphasizing, in conclusion, that in the resultant dialectics or
trialectics of modernity, in the conjunction of modernity as project, discourse, and experience, a crucial spatial politics is always played out. Modernity is not about chronological newness; even less is it a well-defined period or state. It is best studied as a contextual and contingent historical geography, the product of a network of people, practices, knowledges, and objects that in their assemblage produce multiple “spaces of modernity” (Ogborn 1998). In these hybrid sites and landscapes, modernity is projected, planned, and programmed, hubristically celebrated as well as condescendingly denied, and at the same time lived and experienced, creatively represented, sometimes loved, sometimes loathed. To think of modernity in this multifaceted way is at least partly to address the critical concerns of Latour and other critics of the conceptualization of modernity, while retaining its unrivalled utility as description and provocation.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Cultural geography; Discourse; Empire; Historical geography; Imperialism; Modernization theory; Postmodernity; Space; Spatiality; Time–space convergence; Technology

References


Modernization theory

Sudarshana Bordoloi  
*St Cloud State University, USA*

Raju J. Das  
*York University, Canada*

Modernization theory (MT) seeks to elaborate the differences between what it calls traditional and modern societies and to explain what impedes the transition of traditional societies to modernization. MT is a significant part of the “development project” that began in the late 1940s (Parpart and Veltemeyer 2014). The concept of modernization arose from the dichotomous categories, such as “backward versus forward,” “traditional versus modern,” or “barbaric versus civilized.” The Western world (Europe and later North America) was seen as the center of modernity characterized by social mobility, equal opportunity, the rule of law, and individual freedom, whereas, the rest of the world comprised traditional societies based on “ascribed status, hierarchy and personalized social relations” (Kiely 1995, 37). Implicitly, the Western model of development was therefore staged as the “norm” and the West was viewed as the pinnacle of what modernization meant in terms of civilized development and human progress.

**Historical context of modernization theory**

All ideas originate and spread in determinate historical-geographical contexts and reflect certain social interests. MT as a set of ideas is no exception. MT developed in the context of the postcolonial Cold War (Baber 2001) and de-colonization in which there were distinct threats to the economic and geopolitical security of the United States. These threats came from processes such as the spread of Stalinism (often called communism), which needed to be countered through strategies of economic and political development in the Global South. Thus from the vantage point of the United States in the 1950s, MT was the response of the West to rising global socialist tendencies in the aftermath of World War II, particularly the rising influence of the USSR on third-world countries (Preston 1996, 178). In addition, the rise of the national bourgeoisie in the postcolonial countries and the rising legitimate aspirations of their masses for economic welfare meant that these countries wanted to catch up with the West. Attempted transitions to the status of a developed or modern society largely meant transformation of the economy through technological uptake and rapid industrialization as well as widespread social changes able to disrupt traditional forms of social and political order (So 1990, 17; Moore 1963, 69). Such changes would also create a market for goods and surplus capital from the United States and other advanced Western countries.

While there was a need to promote development in the Global South from the standpoint of the United States and from the standpoint of the Global South itself, the extant economic theory was unable to fulfill this need in part because of its neglect of the noneconomic conditions for development. Modernization studies, encouraged and funded by the US government, became a growth industry and emerged as a multidisciplinary effort
MODERNIZATION THEORY

to examine the prospects for development. As discussed below, MT has come in different disciplinary forms: sociological, economic, psychological, political, and geographical.

Contours of modernization theory

Sociological theory of modernization

Sociological MT has its roots in long-established ideas in sociology and, especially, structural functionalism, which stresses the significance of social institutions, structures, and social norms that discipline and socialize people in their struggle for survival and progress (Peet and Hartwick 2009). Renowned Harvard structural-functionalist sociologist Talcott Parsons (1902–1979) played an instrumental role in shaping and establishing MT. His most important contribution was his conceptualization of “pattern variables” that distinguish modern from traditional societies. These variables offered a cultural explanation for why less developed countries are less developed and why traditional values should be transcended such that modernization can occur.

The first pattern has to do with the role of affect in shaping relationships, what Parsons’s conceptualized as affective versus affective-neutral relationships. In judging a person, one can directly express one’s feelings in relation to the person (affective) or can set the feelings aside (affective neutrality). Traditional societies have affective (face-to-face, emotional, etc.) social relationships as compared to affective-neutral (impersonal, detached, indirect) relationships in modern societies. For example, in traditional societies, employees of a firm will not be fired even if the employer is losing money because the firm is considered a family. This means lack of economic efficiency and lower level of development.

Second, there is the distinction between particularistic and universalistic relationships. One may judge a person according to criteria that are peculiar to the person (particularism) or criteria that are applicable to many. Social relations in traditional societies are particularistic in nature where people interact with each other through interpersonal trust. In modern societies, however, social relations adhere to universal or more generalized norms, where interpersonal relationships are not necessary. In traditional society, people tend to associate with members of the same social circle (e.g., they work for a relative’s firm or they buy consumption items in a neighborhood store), showing attachment to a locality. People treat each other particularly. Customs determine the manner of economic activities. Particularism prevails in the distribution of economic roles. In modern societies, however, people tend to interact with strangers treating them in universal terms: for example, bank clerks or state officials ask all clients/citizens for proof of identity. Workers are mobile and unencumbered by artificial restrictions and seek employment where they could be more productive and earn the best wages. Universalism prevails in the distribution of economic roles and underlies efficient allocation of labor and resources (So 1990).

Third, traditional and modern societies can be distinguished on the basis of whether collective or self-orientations are fundamental to social order. A person can focus on work/relations that benefit a society as a whole or just himself/herself. Traditional societies are collective in their orientation, tied together by communal goals. Modern societies stress individualism and the focus on self-development. In traditional societies, people are asked to sacrifice their own interests for the sake of collective obligations. In modern societies, the opposite values lead to technological innovation and economic productivity.
Fourth, there is a distinction with respect to whether social status is derived from ascription or achievement. One can judge a person by what she does (performance or achievement) or by what she is by birth (ascription). In traditional societies, a person is evaluated by his/her ascribed status (based on social norms and practices), but in modern societies, by his/her achieved status (based on education, experience, self-improvement, etc.). In a traditional society, during a job interview, a worker is asked about her relatives’ or parents’ names, because there is a lack of reliance of individual achievement in distribution of economic goods (one gets things because of one’s ascription). In modern society, if employers hire incompetent people based on who they are by ascription, they will go out of business.

Fifth, and finally, traditional and modern societies differ with respect to the type of functional roles that individuals play, understood by Parsons as being either functionally diffuse or functionally specific. A person may be related to another through multiple ways (diffuseness) or in specific ways. In traditional societies, employers have functionally diffuse relationships where they have multiple responsibilities: employers hire, provide housing, and act as a guardian, and so forth for the workers. This diffuse relationship is considered highly inefficient in MT. In modern societies the emphasis is on functionally specific roles of actors (e.g., of employers in relation to employees); relations do not cross the boundary of the private enterprise. Since they can avoid other obligations to each other, the employer and employee can pay more attention to increasing efficiency and productivity. Additionally, the division of labor is less developed in traditional societies: one person does one thing, so she becomes more efficient (Parsons 1960).

Economic modernization

One of the most significant contributions to economic modernization theory was made by the economic historian W.W. Rostow (1960) in his seminal book, *Stages of Economic Growth: A Non-Communist Manifesto*. Drawing from earlier stage theories and consciously taking an approach that is opposite to that of Karl Marx who wrote *Communist Manifesto* and argued for a non–capitalist society, where resources are to be controlled democratically for collective use via class-struggle, Rostow (1916–2003) identified five historical stages of efficiency and consumption through which each country will go. The first stage that is characteristic of traditional societies has primitive methods of production, low labor productivity, low technological development, a hierarchical social order preventing social mobility, and persistence of religious mysticism. As traditional countries expand internationally and come into contact with more advanced societies, changes occur in the former as a result of emerging agricultural and industrial economies, the rise of new entrepreneurs, the expansion of global markets, and so forth; thus leading to the second stage – precondition for takeoff. The third stage – takeoff stage – occurs when economic development processes finally take off from all previously encountered constraints. Economic take off in this manner could happen only with external geographic discoveries, internal development of modern science, rise of entrepreneurship, and growth of urban industry, but most importantly, mobilization of capital and resources to raise the productive investment to 10% of national income (So 1990; Peet and Hartwick 2009). This stage is followed by the fourth stage – drive to maturity – where economic development has become an automatic process, with growth in one sector leading to development in other sectors of the economy.
MODERNIZATION THEORY

Societies that have achieved all these four stages then become high mass consumption societies – the fifth and final stage – with rising consumer demands, strong domestic markets, durable and consumer goods services industries, and the development of an urban skilled professional workforce. Technological innovations and development were fundamental in Rostow’s historical stages of development (Larrain 1989; So 1990). A direct impact of Rostow’s theory was on the “dual economy” model forwarded by W.A. Lewis (1950) which drew attention to how traditional sectors of the economy (agriculture) in third-world countries are advantageous for an advanced sector (industry) due to the scope of unlimited labor supply from the former and how the latter could rapidly grow through foreign capital investments (Potter et al. 1999).

Psychological modernization

Scholars have also analyzed modernization from a psychosocial/psychocultural standpoint. The political scientist and economist Everett Hagen (1906–1993) links personality patterns to technological innovation and change. Whereas an authoritarian personality and lack of creativity among individuals in traditional societies impeded technological progress, modern societies are endowed with innovative personalities who seek creativity to satisfy their own needs (Hagen 1962). On a similar note, McClelland (1917–1998), in agreement with Hoselitz (discussed below) and Schumpeter, stresses that human motivation for achievement, power, and affiliation is the driving force behind entrepreneurship; entrepreneurs are seen as leading the route to rapid economic development (Larrain 1989, 95). Resorting to the theory of social deviance, Hoselitz saw entrepreneurs as the “social deviant” – innovations are initiated by people who think and act differently and thus contribute to generating “creative adjustments in situations of change” (Larrain 1989, 90). Inkeless (1964) considered education and values (through informal curriculum) – particularly disseminated from the Western tradition – as a valuable contribution for the development of the “modern man” in non-industrialized societies (So 1990).

Political modernization

Modernization theory was also a political discourse. Drawing primarily from the structural-functionalist perspective, works of Huntington (1968) and Coleman (1971) were instrumental in shaping US foreign policy and promotion of political institutions as pillars of modernization. In his influential work, Political Order in Changing Societies, Huntington indicated that rapidly modernizing countries could easily come into conflict with pre-existing cultures and local identities. This would lead to political instability – conflicts among different civilizations or cultures often backed by religious principles or what he called “clash of civilizations” – within nation-states, although they may remain the most powerful actors at the global scale (Peet and Hartwick 2009, 115). MT has also stressed the association between economic development and liberal democracy (for a recent illustration of this linkage, see Jacobsen 2015). Since the United States is an advanced country and liberal-democratic, it is thought that democracy is a necessary aspect of economic modernization. Other scholars argue, however, that modernization need not generate democracy but democracies survive in countries that are modern (Przeworski and Limongi 1997).

Geographical modernization

From its earliest formulations, modernity as a concept and modernization as a theory have
had a clear geographical perspective founded on principles of spatial differences and spatial diffusion. The works of Swedish geographer Hägerstrand (1952) and Gould (1964) emphasized that spatial diffusion of innovations (Hägerstrand 1952) and communicative possibilities (Gould 1964) were constrained by distance – the closer a place is to the center of
MODERNIZATION THEORY

modernity, the faster is its pace of modernization. Hägerstrand argued that waves of change move across space and gradually lose power, owing to the friction of distance. Gould argued that people are persuaded to adopt innovations through communication with each other. Geographical theories therefore emphasized the concept of modernization through spatial diffusion.

Modernization takes place as people act in response to foci of change, that is, towns and cities. The modernization diffusion process originates in contact situations such as port cities or colonial administration centers with patterns of change moving down the hierarchy of urban areas to interior areas and along transport routes (Taaffe and Gauthier 1973). This process can be measured by the spread of education, medical facilities, and mapped as modernization surfaces. Modern transportation, geographic mobility, development of education, and transformation of parochial social organizations to embrace modern values and ideas were emphasized as key factors that drive the modernization process. As the geographer Soja (1968) mentions, “modernization is not simply increase in a set of indices. It involves profound changes in individual and group behavior” (1968, 4). Various geographical studies also employed statistical analyses to reveal “modernization surfaces” – case studies of Tanzania (Gould 1964) and Malaya (now Malaysia, Leinbach 1972) – at the global scale. These works tried to project and analyze the extent of spatial diffusion of innovations at multiple scales. Diffusion theories presented a “top-down” or “trickle-down” approach to modernization (Stohr and Taylor 1981) and were widely accepted by the public due to the paternalistic approach of the West toward other cultures (Hettne 1995, 64).

Modernization theory of secular self-expression values

Inglehart has sought to refine MT in the light of recent research on the link between economic development and cultural change. Economic development is associated with shifts away from absolute norms and values toward values that are increasingly rational, tolerant, trusting, and participatory. Inglehart’s neo-modernization theory talks about two kinds of cultural values on the basis of which countries could be categorized: traditional versus secular-rational values and survival versus self-expression values (Figure 1). Protestant Europe and English-speaking North America are at one pole and South Asia and sub-Saharan Africa at another. Traditional societies are religious and are characterized by: close parent–child ties; deference to authority; traditional family values, and rejection of divorce, abortion, and suicide; and a nationalistic outlook. Modern societies have the opposite preferences on all of these topics.

There is another set of opposite values, that is, survival versus self-expression values, which are linked with the transition from industrial society to postindustrial societies. In modern societies, self-expression values are prioritized, and survival is taken for granted; there is a shift from an overwhelming emphasis on economic and physical security toward an increasing emphasis on subjective wellbeing, self-expression, and quality of life. Here there is a high priority to rising demands for participation in decision-making in economic and political life, tolerance of outgroups, including foreigners, gays and lesbians, and gender equality. There is also shift in child-rearing values, from emphasis on hard work toward emphasis on imagination and tolerance as important values to teach a child. There is a rising sense of subjective wellbeing that is conducive to an atmosphere of tolerance, trust, and political moderation. People place
a relatively high value on individual freedom and self-expression, and have activist political orientations. This is conducive to democracy.

**Modernization theory: a multidimensional critique**

MT does bring to attention uneven socio-economic development across countries and within countries. It seeks to shed light on noneconomic causes of such unevenness, such as the lack of development in the Global South. In so doing, it undermines to some extent environmental deterministic explanation of under- and uneven development. However, MT suffers from several crucial problems. Both mainstream social scientists (Nisbet 1969; Eisenstadt 1974; Huntington 1976, cited in So 1990, 53) as well as more progressive scholars (Frank 1969) have mounted numerous criticisms of MT.

To begin with, MT claims that traditional values are an obstacle to technological and other forms of change. Such claims are idealistic: they explain social processes merely in terms of ideas and ignore the fact that there are material origins and material (i.e., developmental) effects of traditional values. Whether values affect development depends on, for example, state policies and the social relations (of production and reproduction) under which people transform resources to produce things for use and exchange.

Persistence of traditional values may have material preconditions. For example, peasant conservatism may be a reality but it exists because of definite material conditions. Vulnerable to crop disease, weather changes, fluctuations in the price of their produce, and to exploitation by social and political systems, peasants may bear conservatism as an attempt to establish some order in these precarious conditions, that is, some safety mechanisms. Traditional values may also have positive developmental effects in some cases. While the traditional MT, which, based on the Western experience, says that family ties can hinder development as they promote nepotism, weaken work discipline, thwart the free market selection of labor, and dilute incentives to invest, recent development scholarship, based on the non-Western experience argues otherwise. For example, Wong (1988) asserts that patron–client relations of dependency between industrial patriarchs and their employees retard class-consciousness among workers and help employers exploit and incorporate family labor, which is cheap, pliant, and hardworking. These relations, in a state of highly fluctuating global capitalist production regime, help the firms remain competitive (see So 1990). Wong argues for an economically dynamic ethos of entrepreneurial familism, which involves family as the basic unit of economic competition, providing the impetus toward innovation and risk taking.

Similarly, while MT has argued that individualist values of modern societies promote development and that collectivist orientation in traditional societies hinders it, research on social capital makes a different argument: social capital understood as norms of trust, reciprocity, and cooperation increases the value of existing resources (Das 2004). Indeed, promotion of social capital, including by the World Bank, is now seen as a low-cost way of promoting development, which is partly why social capital has become a new cottage industry in academic circles. That social capital and MT share something in common is a different matter: both stress norms/values as fundamental factors explaining development issues, in abstraction from class relations.

MT further says traditional society must emulate modern values/ideas, but it never asks whose goals/values these are. Clearly, these values are those of the West and are considered superior...
MODERNIZATION THEORY

to those of the rest of the world and universal; thus MT suffers from a form of ethnocentrism. Frank found the prescribed way of high mass consumption as the ultimate goal of modernization highly problematic as it does not take into account cultural specificities which make societies think about goals and outcomes of development differently (Frank 1969). Indeed many would wish to live well without the social and environmental problems associated with high mass consumption societies. The problem, for us, is not necessarily that certain ideas reflecting European history are being considered universal and therefore are ethnocentric. The problem for us is much rather this: the interests of capitalist development, which originally happened to take place in Europe, and ideas that are reproductive of capitalist development, are made to appear to be universal interests and universal ideas.

There is also a spatial fetishism that occurs in MT, one that undermines its context sensitivity. Diffusion of technology and modern ideas, which is necessary, is not merely a question of communication but an ability to use information and maintain economic control over resources, including via access to state power. Modern seeds may be available but peasants may not have the money to buy and use them (Yapa 1977). Mere diffusion of ideas and technology from the West or a few metropolises within the Global South linked to the West may not be beneficial to the less developed world and may indeed produce increased inequalities given the structural conditions that shape livelihood possibilities.

Related to this spatial fetishism is the manner in which MT ageographically homogenizes less developed and advanced countries. For example, Rostow’s stage theory relegates all “backward” societies with the same typology and uniform traditionalism. Such typology obfuscates the diversity of cultures in the third world along with utter disregard for the long and troubled colonial histories to which these countries were subjected (Frank 1969). Tribal, feudal, and bureaucratic empires are so different from each other that one cannot put them under the category of traditional society. To do so is to deny specific pre-capitalist histories of these countries/territories and to hide the under-developing effects of their contacts with Europe. In fact, Eisenstadt, an internal critic of MT, says that there are many differences between traditional societies and there are many traditional elements in modern societies.

It is not just that there are differences between traditional societies at the national level. Even within the so-called traditional societies, there are different types of values. There are no homogeneous traditional values. “Great tradition” – of the elites – like reading poetry, painting, hunting, philosophy, and so forth, and “little tradition” (working in the fields, thrift, etc.) coexist. For example, Indian executives may try to improve efficiency in their firms by scientific means, may worship Lakshmi to protect their competitiveness, and may consult with astrologers to avoid a heart attack or to determine the day of launching a new branch plant (Singer 1972).

Advanced countries are also homogenized in MT, and their modernity is exaggerated. In modern societies, particularistic values – those concerning gender, age, ethnicity, sexuality, and so forth – are present and are deployed in recruitment and promotion of personnel. In the United States, there are vast differences in income and wealth (including homeownership), which are based on racial differences. America still has not had a female president, and all but one of its presidents are whites, which means that not all groups experience equal opportunity for achievement, which is an important criterion of modernity. In modern societies such as Japan, promotion and post-recruitment salaries depend on age, and so forth. On the other hand, less
Developed countries may be just as universalistic, achievement-oriented, and functionally specific in many ways as developed countries are and yet still remain less developed.

We argue that whether in poor or rich countries, when inequality is rapidly rising, there is a structural pressure on the part of the economic and political elite (so-called modernizing elites of MT) to promote obscurantist (i.e., traditional) ideas among the masses to weaken their solidarity and ability to fight for equality (i.e., to demonstrate achievement orientation of MT): ordinary people may use modern cell phones and laptops and still be ready to kill each other on grounds of religion, race, or caste in the name of their respective gods, superior skin color, or ritual purity. We also assert this: since modern values coexist with traditional values partly in order to reproduce existing inequalities in the Global North and South, MT is mistaken to believe that the agency of development is the modern elite or that the West is a paragon of modern values.

MT also suffers from its teleological unidirectionality, assuming that the progress of human societies is predetermined in that each society is aware of the ultimate destination it has to finally reach in the course of its development process (So 1990, 19). Thus MT is developmentalist in its belief that the history of Euro-America is generalizable into a sequence of states of economic growth that all countries must follow (Pieterse 2010, 19). It also assumes that countries in the periphery want to and will achieve the same kinds of societies and economies that exist presently in the more developed West, and that “any contemporary differences in levels of development are ... merely different locations along the same path to development” that the West has traveled (Taylor 1992, 12; see Figure 2). Further, MT’s assumption that modernization followed a unidirectional path – the rest of the world following the West – ignores alternative paths to development that countries based on their unique history and cultural patterns may pursue. So in MT, advanced countries have history, but less developed countries do not have that luxury.

Related to this teleology is the underlying assumption of MT that there is one original developmental state common to all countries. In particular, MT assumes that the currently more developed countries may have remained underdeveloped at some point in the past. It mistakenly classifies today’s underdeveloped societies in the Global South with preindustrial underdeveloped societies in the West. Frank has argued that today’s developed countries may have been underdeveloped, not underdeveloped. A country may be undeveloped because of lack of resources and technology. That is different from a country being underdeveloped, which happens when its potential to become developed is underdeveloped (i.e., blocked) due to exploitation of it by other countries.

The insufficient recognition of such blockages and interdependencies means that MT blames the so-called third-world societies (their cultural values) for lack of development and ignores their oppressive/exploitative external connections. MT ignores colonialism and imperialism, including imperialism in its old form as well as in its new form (symbolized by interventions of international organizations) (Peet and Hartwick 2009). Relations between traditional and modern societies, to the extent these are considered in MT, are seen as conflict free and mutually beneficial. Frank asked how the Euro-American model of development could be repeated by the Global South primarily when the West’s development had already altered the context in which the Global South has to develop (Frank 1969, as cited in Peet and Hartwick 2009). As Kiely (1995) points out, “modernization theory ...
separates the modern international division of labor from ‘traditional’ nation states, and fails to clarify the mechanisms by which the latter were integrated into the former” (Kiely 1995, 43).

Finally, but no less significantly, there is MT’s elitism and ideological character. In the first instance, MT argues that only the elites can bring about development. While MT appears to be a scientific theory, its class commitment is to society’s rich elites (Peet and Hartwick 2009, 119). The Parsonian synthesis was aimed at protecting the interest of the dominant property owning classes of the society, thereby ignoring aspects of social inequality and associated exploitation, power imbalances, and social resistances of the masses (Gouldner 1970, 240).

Ideologically, MT says that the Global South has a lower level of achievement orientation, but when political mobilization aimed at achieving a higher standard of living happens, it is seen as a revolution of rising expectations or a crisis of governance, prompting the same MT thinkers to propose a lowering of newly acquired aspirations and level of political activity (Bernstein 1971). This is not surprising given that one of the main aims of MT is to help identify the threat of anti-capitalism in the Global South as a modernization problem. Clearly, what is a modernization goal for the masses is not a modernization goal for MT. If the Global South has to modernize, it has only one choice: move along the path that the United States has moved and away from anti-capitalism. To help achieve this, MT suggests profit-driven economic development, replacement of traditional values by commercial values, and institutionalization of liberal-democratic state forms. Not surprisingly, universities (e.g., Chicago), foundations, and nongovernmental organizations are lavishly funded to discredit anti-capitalist and anti-imperialist politics among third-world intellectuals (who encourage achievement orientation on the part of the masses).

All told, MT helps to legitimate the ameliorative foreign aid policy of the United States and
other advanced countries. Rostow and others say, “anticipating” the post-1970s neoliberalism, that underdeveloped countries should welcome foreign capital, be export-oriented, and promote free enterprise (Peet and Hartwick 2009). By arguing that developed countries can have only benign influence via the ideas and interventions of development agencies, the MT justifies the activities of these agencies. Rostow, the author of the Non-Communist Manifesto and a master modernization theorist, indeed became one of the closest advisors to a US president (Lyndon Johnson). If environmental theory developed in the historical context of colonialism and to justify it, MT developed in the context of Cold War and to justify Western domination (and Western-style capitalism). Slater says that the occidental deployment of MT signifies a will to power, that is, geopolitical power, providing an ideological legitimization for a whole series of practical interventions and penetrations seeking to subordinate and assimilate the West’s third-world other (Slater 1993). The purpose of MT was to create “an international order as a system dominated and manipulated by certain industrial countries [primarily the US] in their own interests” (Larrain 1989, 100). Indeed, the categories of modern and traditional or civilized and barbaric, used in the overall framework of MT, are ideological labels to justify Western superiority as well as Western interventions in the third world – in one word, imperialism – in the name of lifting it out of underdevelopment.

**Conclusion**

MT is a project of multiple disciplines. Each discipline has identified key issues concerning modernization. Sociologists: cultural obstacles to modernization. Economists: speeding up of productive investments including through Western aid and investment and ideas of free enterprise. Political scientists: enhancing the capacity of the political system to cope with the stress associated with modernization (e.g., public mobilization). Geographers: spatial diffusion of modern ideas.

MT basically seeks to explain lack of development in terms of the absence of modern values in poor countries. These values are not conducive to profit-driven economic development. MT also says that people in traditional societies are resistant to change (they do not accept new technology and new ideas easily) and therefore lack modern scientific knowledge about nature and society. And, of course, these societies lack investment, which advanced societies should provide in the form of aid, FDI, and so forth. Modern ideas, modern technology, and capital from the West are needed in poor countries, which should have a hierarchy of cities and settlements through which these modern ideas and things will diffuse. The actors in the drama of modernization will be modernizing sociopolitical elites, commercial profit-seeking entrepreneurs, and innovative individuals. The role of the masses – workers and peasants – is mainly to change their ways of thinking and to receive modern ideas from the West and from the domestic elites.

Following in the tradition of the dependency theories, which critiqued modernization theory for its justification of exploitation, dominance, and creation of dependencies by the core countries on the peripheries, alternative strategies of development that were fundamentally opposed to the idea of Western models started to emerge from the 1970s. While Emanuel Wallerstein’s world-systems theory (1974) positioned capitalist relations of domination and subordination at a world scale involving core, peripheries, and semiperipheries (Amin 1976), other theories, in the tradition of postmodernism and postdevelopment philosophies, have focused on indigenous knowledge, civil society based
participatory practices, postcolonialism against colonial discourses, and so forth as alternative strategies of development from a third-world perspective.

Despite the critiques and alternatives, modernization theory has found a way to live on through neoliberalism, the economic discourse of development since the 1990s. The fundamental principles that characterize MT have been strikingly similar to the main tenets of neoliberalism. With its emphasis on rationality, free markets, achievement of high mass consumption, withdrawal from state welfare interventions and state regulation of companies, integration of countries into the global capitalist order through trade and foreign investment, and development projects based on high-cost Western technology, MT could be seen as a prototype – a “harbinger” – of present day neoliberalism (Kiely 1995; Potter 2004; Peet and Hartwick 2009). In conclusion, as Peet and Hartwick (2009) argue, “modernization theory formed the more general theory of history within which neoliberal economics could be situated” (Peet and Hartwick 2009, 140).

A given set of ideas will not disappear as long as the material conditions which give rise to it persist, whether these interests are international or domestic. The problem of development, of modernity, persists in that millions of women, men, and children do not have access to nutritious food, health care, housing, safe neighborhood, transportation, culture, science and technology, and so on. The need to explain this problem and to act on it still exists. Some scholars will continue to produce explanations that deal with the surface causes, including people’s ways of thinking, without aiming to understand the material, political-economic, roots of these ideas. The traditional MT and those ideas which appear to be critical of MT (e.g., social capital theory) have done precisely this: trying to explain unevenness in material conditions in terms of the ways in which individuals or groups of individuals think and behave, and are therefore reproductive of the current system of production which is behind that unevenness. The ideas informed by the postmodernist turn, as well, share something important with MT: they do not fundamentally challenge capitalism or imperialist economic exploitation as a material reality which exists independently of the people who are thinking about it now. That is, given their stress on everything as dominantly a discursive construction and given their neglect of the systemic character of society and its materiality, they do not and cannot argue for overturning the system and for the creation of a socioecological order where there is democratic control of resources and production to be achieved through a multiscalar revolutionary action of the masses. They essentially argue for a system that is reformed through, for example, local-level (community) actions and discursive destabilization (i.e., academic critique of some relatively trivial or less important parts of the system). This is why we continue to need a theory that goes beyond modernization theory and similar ideas, postmodern or not. This would be a theory which not only recognizes the consequences of sociocultural ideas (as well as local initiatives people take) for economic development issues, but which also emphasizes that such ideas (and such local initiatives) are heavily conditioned by deeper structures of unequal relations defined in terms of access to productive resources and the ways in which people earn their livelihood.

SEE ALSO: Indigenous knowledge; Participatory geographies; Postcolonial geographies; Postdevelopment; Postmodernity; World-systems theory.
References


MODERNIZATION THEORY


Further reading

Modifiable areal unit problem

Paul A. Longley
University College London, UK

Geographic scale, or granularity, is fundamental to the fitness for purpose of a geographic representation. Yet the level of granularity with which we are able to represent the world may be restricted by constraints that lie beyond our control for reasons that are technical—such as the resolution of an airborne sensor or satellite—or societal and ethical constraints—as with the disclosure restrictions on individual level census or other survey data. The hypothetico-deductive approach to development of theory requires the specification of an appropriate scale range over which hypotheses can be tested, but it may happen that the researcher lacks either the technical equipment or the ethical mandate to record the phenomenon of interest in sufficient detail for the task in hand.

Other approaches are much more explicitly data driven, and inductive. Here, an insidious problem that characterizes much of geography is that, unlike many other scientific disciplines, there are few if any “natural” units of analysis into which elemental observations might be aggregated. What, for example, is the natural unit of analysis for measuring a soil “profile”? What is the extent of a “pocket” of unemployment, or a crime hot “spot”? In practice, the definition of objects of analysis in geography is often inherently ambiguous and, in practice, rarely precedes our attempts to measure or analyze their characteristics. In socioeconomic applications, the objects of study (or “geographic individuals” in the terminology of geographer Stan Openshaw) (Openshaw and Alvanides 2005) are usually aggregations, since the geographic locations that human subjects occupy are unique, and disclosure control usually requires that uniquely attributable information must be anonymized through aggregation. In natural environment applications, it may be tempting to think that all unique locations can be represented, but in practice the observable world is only viewed at a prescribed level of resolution. The accuracy of a land cover classification of a classified satellite image is only as good as the resolution at which reliable measurements were taken, for example. In each case, the fundamental problem is that the observable world is of seemingly infinite complexity and the units of analysis used in geographic analysis are often the outcome of choice, convention, and chance.

As a consequence, we cannot be entirely confident about ascribing even dominant characteristics of areas to true individuals or point locations in those areas. This source of uncertainty is known as the ecological fallacy, and has long bedeviled the analysis of spatial distributions. (The opposite of ecological fallacy is atomistic fallacy, in which the individual is considered in isolation from his or her environment.) The ecological fallacy problem is a consequence of aggregation into the basic units of analysis and is illustrated in Figure 1 (Longley et al. 2015). In this diagram, the function of a downtown central business district (CBD) has been usurped by a new neighboring retail center (not shown). This has resulted in high local levels of service sector employment. The operations of the factory located to the north of the CBD...
are unaffected, but its workers live in the same labor force zones as the service sector workers, generating the misleading impression that unemployment may instead be a blue collar (factory worker) phenomenon.

The likelihood of committing ecological fallacies such as this depends upon the administrative (or other) geography being used. If the units of analysis within each zone share the same characteristics or attributes, then the only geographic variation (heterogeneity) that occurs will be between zones. However, within-zone variation is usually the norm, and greater heterogeneity increases the likelihood and severity of the ecological fallacy problem. This said, it is also the case that geographers may have become over fixated about this potential problem over the years, and many general-purpose zoning systems are designed to maximize within-zone homogeneity in the most salient population characteristics – as with the UK Census geography, for example. Probably as a consequence, there are very few documented case studies that demonstrate the misleading consequences of ecological fallacy in practice.

The potential of aggregation to create problems in geographic analysis is compounded by problems arising from the different scales at which zonal systems may be defined. Gehlke and Biehl first identified this in 1934. It was subsequently demonstrated in a classic 1950 text

Table 1 1950 Yule and Kendall’s data for wheat and potato yields from 48 English counties.

<table>
<thead>
<tr>
<th>No. of geographic areas</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>0.219</td>
</tr>
<tr>
<td>24</td>
<td>0.296</td>
</tr>
<tr>
<td>12</td>
<td>0.578</td>
</tr>
<tr>
<td>6</td>
<td>0.765</td>
</tr>
<tr>
<td>3</td>
<td>0.990</td>
</tr>
</tbody>
</table>
by British statisticians Yule and Kendall, in an analysis that began by calculating the correlation between wheat and potato yields for the (then) 48 counties of England. Upon aggregating the data for the 48 counties to 24, then 12, then 6, and finally 3 zones, they demonstrated that correlation coefficients tend to increase with scale. Table 1 presents their results, ranging from near zero (no correlation) to over 0.99 (almost perfect positive correlation) — although subsequent research has suggested that this range of values is atypical.

Scale turns out to be important because of the property of spatial autocorrelation. A succession of subsequent research papers have reaffirmed the existence of similar effects in multivariate analysis. However, rather discouragingly, scale effects in multivariate cases do not follow any consistent or predictable trends. This has important implications for the analysis of big data and other secondary sources, where analysts rarely have control over the scale at which data are made available.

Further aggregation effects can arise because there is a multitude of combinations in which the mosaic of smallest available areal units can be assembled together into contiguous zones. This gives rise to the related aggregation or zonation problem, in which different combinations of a given number of geographic individuals into coarser-scale areal units can yield widely different results.

In a classic 1984 study, geographer Stan Openshaw applied correlation and regression analysis to the attributes of a succession of zoning schemes. He demonstrated that the amalgamation of elemental zones into larger areal aggregations could be used to manipulate the results of spatial analysis to a wide range of quite different outcomes. Moreover, using emerging geocomputational techniques, Openshaw was able to demonstrate that these outcomes could be prespecified by the analyst through careful, if usually contrived, zone design. These numerical experiments have some sinister counterparts in the real world, the most notorious example of which is the political gerrymander of 1812, in which the amalgamation of elemental voting districts was engineered to create an electoral constituency in Boston, MA, that would return Governor Gerry as its elected representative. The implication is that it is possible to assemble apparent spatial distributions that are unrepresentative of the scale and configuration of real-world geographic phenomena. The outcome of multivariate spatial analysis is also similarly sensitive to the choices and conventions that give rise to particular zonal schemes.

Taken together, the effects of scale and aggregation are generally known as the modifiable areal unit problem (MAUP). The ecological fallacy and the MAUP have long been recognized as problems in applied spatial analysis. Through the concept of spatial autocorrelation, they are also understood to be related problems. Geocomputational methods and big data analytics can be used to demonstrate the nature and potential extent of this problem, in particular empirical settings, and the results of experiments and simulations have focused interest on the effects of within-area spatial distributions upon analysis. This is an important contribution to balanced assessment of the quality of secondary datasets.

However, repeated scaling and aggregation experiments do not resolve the uncertainties inherent in spatial analysis. Zone-design experiments are merely playing with the MAUP, and do not provide any substitute for clear thinking about the validity of available units of analysis for particular investigations. There is no “solution” to the MAUP, but simulation of large numbers of alternative zoning schemes can gauge its likely effects. The imposition of
elemental zones that are to some extent arbitrary may be ad hoc, but it is rarely completely random. Measuring the range and variance of the MAUP in any particular case study does not really help us to accommodate its effects because there is no tenable analogy between “simple random zoning” and simple random sampling. Rather, the way forward is likely to lie in clearer application-centered thinking about the nature of within-zone heterogeneity that is concealed in aggregated data. In this view, MAUP will disappear if analysts understand the particular areal units that they wish to study.

Three other important considerations flow from this discussion. First, application-centered thinking implies a focus upon the uniqueness of places. Second, areal objects are not time-invariant but ever changing, and our perceptions of what is appropriate zonation will also change over time as a consequence. Third, at least within the socioeconomic realm, the act of defining zones can be self-validating if the allocation of individuals affects the interventions for which they are eligible (e.g., eligibility to apply for a place at a school, or eligibility for workfare incentives). Spatial selectivity affects spatial behavior, and so the principles of zone design are of much more than academic interest.

The abiding message of all of this is that the MAUP is not best thought of as an empirical problem but instead poses a theoretical requirement to hone statistics to the geographic context in which they are applied. The problem should not be considered in isolation, but rather as part of a grander design to develop explicitly geographical representations of the accumulated effects of historical and cultural processes upon unique places.

**SEE ALSO:** Geocomputation; Geographic information science; Spatial analysis; Spatial resolution

**References**


Moldova: Institutul de Ecologie și Geografie al Academia de Științe a Moldovei (Institute of Ecology and Geography, Academy of Sciences of Moldova)

Founded: 2005
Location of headquarters: Chisinau
Website:  http://www.ieg.asm.md/ro (Romanian), http://www.ieg.asm.md/en (English)
Mongolia: Mongol ulsyn
Shinjlekh ukhaany
Akademi, Gazarzüin
Khüreelen (Institute of
Geography, Mongolian
Academy of Science)

Founded: 1926
Location of headquarters: Ulaanbaatar
Website: www.mas.ac.mn
Monitoring and evaluation

Camille Gaskin-Reyes
George Washington University, USA

Monitoring and evaluation (M&E) methodologies are used in many fields to record and assess the performance of organizations, plans, and projects, and to involve the identification, measurement, and assessment of outputs, outcomes, and impacts. In the field of regional development M&E embraces activities in the economic, educational, health, and environmental sectors. In this entry the health sector will be used as an analogy to understand the basic concepts underlying monitoring and evaluation, and attention will focus on development projects, although the principles and methods have more general applicability.

In the health sector, physicians perform periodic and systematic checkups on their patients to assess their health status. Based on the interpretation of the symptoms that the patients report and the results of medical tests, the physicians make diagnoses and prescribe treatments. They usually monitor the situation for a while, and then ask patients to come in for follow-up visits to evaluate changes. From this example, one sees that monitoring involves an ongoing process of supervision and oversight during an event or intervention, whereas evaluation has to do with reviewing and assessing outcomes after a process of monitoring has been completed.

Monitoring can be defined more formally as a systematic procedure for checking for the effectiveness (outcomes) and the efficiency (outputs and impacts) of an intervention during implementation. Its main purpose is to identify progress and shortcomings and to recommend corrective measures to optimize desired results. In the health sector, the health professional monitors the patient and makes evaluative judgments. In development projects, the project manager or executing entity is responsible for monitoring the development intervention to gauge project progress.

Performance indicators are a key aspect of monitoring, whatever the sector or area. They are defined as metrics to track the progress of an intervention and the attainment of targets that a project is expected to meet. Their use also enables a comparative look at the situation, whether in a sector, project, or country over a specific time period. Using a baseline as a point of departure, project managers are expected to use the project information generated by the system to update and review project status on a periodic basis. The frequency of reviews varies from project to project, situation to situation; as a general practice, such reviews are carried out quarterly, yearly, and at the mid-term point.

In project design, the entire results framework and architecture are usually set up at design stage, using for example, a methodology such as the logical framework. The logical framework is an instrument that pulls the entire project results chain together: from inputs (resources) and processes to outputs, outcomes, and finally impacts. No logical framework is useful if it does not incorporate a sound monitoring and evaluation plan, including defined performance indicators.

While monitoring takes place during the life of the project, evaluation is the objective, largely third-party valuation of the impacts of an intervention after project completion and
final disbursements. Evaluation also uses pre-
determined criteria, focusing on factors such
as relevance, efficiency, effectiveness, impact
(understood as changes to beneficiaries’ health,
welfare, or living standard), sustainability, and
lessons learned for feedback into new design.
Hence, evaluation is about recording and
demonstrating change: changes over time for
the same group, changes in living conditions in
a geographical project area, changes in overall
poverty as a result of poverty reduction interven-
tions or strategies through the appropriate use of
project resources.

To sustain an evaluation system, however, and
for it to be valuable and valued, there must be
demand for the information. Other important
factors are an established, formal structure;
the timeliness, accessibility, and cost-efficiency
of the system itself; and the institutional and
technical/information technology capacity
to run and maintain it. A further important
hallmark of a sound evaluation system is the
credibility, transparency, and utility of the
information that it generates.

In development projects, monitoring and
evaluation go hand in hand like oil and vine-
gar. They have become indispensable tools for
project managers, officials, project stakeholders,
and civil society to assess the ongoing status of
interventions or activities under execution.

It cannot be overstated that project managers
can know if projects are really on the right
track only if the performance indicators and
data collection systems are set up and function-
ing smoothly. Much in the same way that the
physician monitors the vital signs of a patient,
the system helps the project manager to analyze
information and understand important changes,
in particular emerging issues, risks, or red flags
within the project context. Failure to address
problems that come to light through monitoring
could derail implementation and the attainment
of project results. A steady and reliable stream of
information also facilitates the ongoing review
and quality assurance of the project’s path toward
results.

Results consist of both outputs and outcomes.
Outputs are products and services delivered (i.e.,
health clinics, schools, training seminars, learning
materials) and outcomes represent the initial and intermediate effects of the use of those
outputs on the intended beneficiaries of projects
(improved health care, learning outcomes, etc.).
Projects also have impacts, defined as medium-
to longer-term welfare gains or improvements to
society. These impacts are the focus of the eval-
uation process, long after the completion of the
intervention or project.

Based on their in-depth knowledge of the
issues surrounding the project, project man-
agers should make timely decisions to resolve
problems. A project that exhibits continuous
bottlenecks must be closely monitored until
improvements are noted. Monitoring devel-
opment projects requires, in particular, the
participation of key stakeholders and beneficia-
ries. These groups are well positioned to provide
feedback on the quality and relevance of the
project’s products and services.

There are a number of steps to achieve success
in the monitoring, reporting, and ultimately the
evaluation of projects:

• plan for the monitoring and evaluation
  framework from the time the project is an
  idea, and set it up early;
• be rigorous in the implementation of a mon-
  itoring plan and cover any associated costs of
  the plan;
• conduct a readiness assessment with develop-
  ment partners and establish agreements on
  results and the factors for project evaluation;
• choose project performance indicators that are specific, measurable, achievable, relevant, and time-bound;
• anticipate and address risks throughout the monitoring process;
• pay attention to baseline data, data sources, data collection frequency, targets, and responsibilities for monitoring and results reporting;
• report results on achievements candidly, and supervise for risk management and course correction;
• use quantitative and qualitative information for timely decision-making by managers;
• make provisions for mid-term assessments and post-completion impact evaluations.

Once a project is completed, project managers usually prepare a project completion report. This report is a stock-taking exercise. It represents a preliminary assessment of initial or intermediate outcomes of the project, the use of budgetary resources, the outputs acquired, and preliminary outcomes.

At this point, the process of monitoring ends, and the evaluation phase begins. The evaluation of final outcomes usually takes place about four years after project completion, and involves persons not operationally linked to the project. The emphasis of this evaluation, sometimes called an ex post evaluation, is to generate final, objective conclusions with respect to the attainment of project goals, and recommendations for the improvement of future operations.

There are many methods and approaches for consideration. The specific choice of an instrument depends on a number of factors, such as the demand for the information, the costs involved, the technical capacity, and the time available for the evaluation exercise. Mostly, formal surveys are used for information gathering from a sample of beneficiaries or households to understand changes due to project interventions. For example, if a project manager wants to assess the project’s impact on how living standards of the population have improved, the type of survey selected would be a living standards measurement survey. The latter includes such aspects as spending, household composition, savings, and household income. Given the fact that evaluation is about trying to measure change in a target population, another tool used is a household survey, which assesses changes in a population’s access to social services such as health, education, clean water, and sanitation.

Client satisfaction surveys and focus groups are very useful to measure beneficiary satisfaction with goods and services delivered by the project. In focus groups of beneficiaries, people are asked about their experiences with the project, their perceptions about quality, timeliness, and usefulness of the services, as well as any bottlenecks experienced.

The information from these different surveys complements quantitative information provided by the monitoring system in place during the execution period. These findings help assess outcomes, the institutional performance of project managers, government entities, and funding agencies, and are also critical for accountability purposes. In recent years, accountability for results has become a critical component or even a condition for funding by international agencies. The latter rely on the information generated to exercise their own fiduciary oversight.

A well-established monitoring and evaluation system from start to finish and beyond project completion is, therefore, one of the main ways to track project progress and verify results. In fact, having a monitoring and evaluation system is now perceived as a key safeguard to manage fiduciary risks. In addition to providing project officials, managers, and the public with a tangible
MONITORING AND EVALUATION

way to corroborate project results, monitoring and evaluation delivers valuable lessons to improve future planning and budgeting, and demonstrate outcomes. Tangible evidence on outcomes reduces subjectivity and makes it easier for project managers to distinguish failure from success.

SEE ALSO: Infrastructure and regional development; Regional development policies; Regional planning: the resilience of an imperative; River basin management and development

Further reading

Monsoons

Shouraseni Sen Roy
University of Miami, USA

The word “monsoon” originates from the Arabic word “mawsim” for season. It was originally used by seafarers centuries ago to describe the alternating system of winds over the Arabian Sea. The most widely used definition of monsoons is the seasonal reversal of winds caused by differential heating of the land and sea, most frequently over southern and eastern Asia. This is primarily caused by the differing response of land and water vis-à-vis incoming solar radiation. The specific heat of water is much larger than that of dry soil. For instance, during the summer months heat conduction in soil takes place very slowly and only the few topmost centimeters of the land surface are heated. Thus, solar energy absorbed by land surfaces produces rapid warming, which, in turn, heats the air in contact with it. In contrast, the solar radiation absorbed by oceans is able to penetrate to a much greater depth, of at least 200 m, due to the vertical mixing of the uppermost water layers, which results in the turbulent transfer of heat. Therefore, land surface temperatures are considerably higher than the ocean along the same latitude, by about 5–10°C (Das 2002). Spatially, monsoons are mainly confined to tropical latitudes within 20° of the equator; temporally, monsoons are most pronounced in the summer months – June–August in the northern hemisphere and January–February in the southern hemisphere. Specifically, at the regional scale, the monsoonal circulation is most prominent over the Indian landmass, with similar seasonal wind reversals recorded above northern Australia (Hendon and Leibmann 1990), western and eastern Africa (Sultan, Janicot, and Diedhiou 2003), and the southwest USA (Adams and Comrie 1997). The focus here is on causal processes, patterns, and variability related to the wet Indian summer monsoon, locally known as the southwest monsoon season.

The majority of the precipitation across the Indian subcontinent occurs between June and September. During these summer months, conditions favorable to triggering the southwest monsoon result from the northward migration of the inter tropical convergence zone (ITCZ) to its northernmost position (30°N), the greatest annual contrast in land and sea temperatures, and the atmospheric steering effects of major physiographic barriers in the north, particularly the Himalayas. These forces set up a strong pressure gradient between the equatorial latitudes and the subtropical region to the region of maximum heating and lowest pressure, which is accompanied by a quasi-geostrophic westerly wind, deflected by surface friction towards the equatorial low pressure trough (Pant and Rupa Kumar 1997). The result is a weakening of the prevailing northeast trade winds, overpowered by southwesterly winds flowing from the opposite direction (Figure 1), the southwest monsoon.

In the upper atmosphere, the subtropical westerly jet stream moving across the Indian subcontinent shifts north to flow over the Tibetan plateau and is replaced by an easterly jet stream whose core overlies the equatorial Indian Ocean and extends westward into East Africa. The formation of this easterly jet stream is linked to the development of an upper level high-pressure ridge above Tibet. This anticyclone extends...
MONSOONS

Figure 1  Southwest summer monsoon circulation over the Indian subcontinent. Yellow arrows represent the Arabian Sea branch and green arrows represent the Bay of Bengal branch. The purple line represents the approximate location of the ITCZ during the summer months. The development of an intense center of low pressure in northwestern India is represented by the red circle with the letter “L” in the center.

vertically from the middle to the upper tropospheric levels of the atmosphere and is associated with mid-oceanic troughs over both the Pacific and Atlantic Oceans. Besides large scale, land–sea heating contrasts, the monsoon circulation also results from the marked sea-surface temperature contrasts and releases of latent heat that help to maintain strong horizontal and vertical wind shear (Pant and Rupa Kumar 1997). Given the relative significance of the summer monsoon in meeting the huge and growing water demands across the Indian subcontinent, it is of special interest in the explanation and prediction of the physical processes that shape monsoons. Ramage (1971) observed four principal features of the monsoon:
Figure 2  Normal dates for the onset of the southwest monsoon season across the Indian subcontinent. Source: IMD 2013.

1. A change in prevailing wind direction of at least 120° between January and July.
2. The average frequency of prevailing winds in January and July should exceed 40%.
3. Mean wind speeds in at least one month should exceed 3 m/s.
4. There should be less than one cyclone–anticyclone alteration every two years in either month within a 5° latitude–longitude grid.

Monsoon onset

The seasonal transition from premonsoon to monsoon is so abrupt that it is commonly referred to as the “burst” of the monsoon over the Indian subcontinent. The typical conditions that establish monsoonal circulation include surface winds blowing from the southwest, a sudden decline in near-surface air temperatures,
and a sharp increase in cloudiness and precipitation (Pant and Rupa Kumar 1997). The first monsoon rains occur over the southwestern state of Kerala, marked by a sustained increase in precipitation around the first week of June (Figure 2). More specifically, the main processes associated with the onset of the monsoon and factors affecting its performance are:

1. **Formation of a heat low and monsoon trough:** as the sun migrates northwards from the Equator, the land areas surrounding the Arabian Sea receive increasingly larger quantities of heat both from the sun and heat releases from the Earth’s surface. This produces a trough of low pressure that extends northward from Somalia across Saudi Arabia, Pakistan, and northwestern India (Das 2002). This trough is also a zone of intense horizontal wind shear.

2. **Mascarene anticyclonic system:** The Mascarene High located over the southwestern Indian Ocean around 30°S and 50°E near Madagascar, generates a significant outflow of air. This airstream moves across the Equator as a southwesterly flow known as the East African or Somali jet, reaching its peak between June and August (McGregor and Nieuwolt 1998). The intensity of this jet stream plays an important role in the intensity of the monsoonal rainfall over the Indian subcontinent.

3. **The Tibetan high:** this large cell of high pressure is centered above the Tibetan Plateau that lies poleward of the surface monsoon trough over northern India, and triggers upper-level divergence as well as lower-level convergence. The outflow of air from the Tibetan high is vital to the formation of the tropical easterly jet stream that lasts from June to September.

4. **Change in kinetic energy:** observations from several field experiments show the development of a sudden and rapid increase in kinetic energy over the Arabian Sea, which is associated with the formation of a vortex off the coast of southern India (Das 2002).

Accurate forecasting of the summer monsoon is critically important for this region. The Indian Meteorological Department (IMD) currently employs a two-stage approach for issuing southwest monsoon forecasts. The first seasonal rainfall forecast is issued in April, followed by an updated forecast in June for July rainfall across the entire country. Detailed information about long-range weather forecasts issued by the IMD can be found elsewhere (Rajeevan et al. 2007). The guidelines used to determine the onset and advance of monsoonal precipitation in India are (IMD 2013):

1. **Rainfall:** if after May 10, 60% of the 14 available weather stations report rainfall greater than 2.5 mm for two consecutive days, then the onset over Kerala is declared on the second day, provided the following criteria are also met.

2. **Wind field:** the depth of westerlies is maintained at 600 hPa in the area extending from the equator to 10°N and 55–80°E, and the zonal wind speed over the area bounded by 5–10°N and 70–80°E is maintained at or above 15–20 knots at 925 hPa.

3. **Outgoing long-wave radiation (OLR):** the OLR values derived from INSAT are below 200 W m\(^{-2}\) within the area bounded by 5–10°N and 70–75°E.

**Monsoon withdrawal**

The change of air circulation associated with the weakening of the monsoon is far more gradual than its onset. The withdrawal begins in northern
India around mid-September, and by the end of October it is complete across the entire subcontinent (Figure 3). This withdrawal is accompanied by the reassertion of the subtropical westerly jet stream and the dissipation of the easterly jet stream. The criteria currently used to determine the withdrawal of the southwest monsoon season by the IMD (IMD 2013) are:

1. Withdrawal from far northwestern India is not analyzed before September 1.
2. After September 1, the synoptic features analyzed for the withdrawal from the western part of northwest India are:
   i. Absence of rainfall activity over the area for five consecutive days.
   ii. Emergence of an anticyclone in the lower troposphere (850 hPa and below).
   iii. Substantial decrease in atmospheric moisture content as detected by satellite water vapor imageries and tephigrams.

Principal branches of the southwest monsoon

The progress of the southwest monsoon across the subcontinent is governed by the formation of two branches: the Arabian Sea branch and the Bay of Bengal branch (Figure 1). The Arabian Sea branch monsoon accounts for most of the intense heavy precipitation over the west coast, whereas the Bay of Bengal branch of the monsoon accounts for most of the precipitation occurring over the Gangetic plains. The two branches converge over northwestern India, resulting in renewed precipitation in the vicinity of Delhi. The Arabian Sea branch steadily advances northward to Mumbai in about 10 days by June 10. The Bay of Bengal branch moves northward into the center of the Bay of Bengal and expands into most of Assam state in northeastern India by the first week of June. After reaching the nearby Himalayas, the Bay of Bengal branch spirals upward and is deflected to the west, where it keeps advancing over the Gangetic basin while another branch hives off toward the northeast. The southwest monsoon months are also characterized by the formation of a monsoonal trough, which extends east–west over the northern plains of the subcontinent and is useful for determining the path of the rain-bearing winds. Several smaller low-pressure cells form during the monsoon season in the Bay of Bengal and they generally move along the axis of the trough paralleling the foothills of the Himalayas. These monsoonal depressions typically consist of 2–3 closed isobars at 2 hPa intervals (Pant and Rupa Kumar 1997), producing substantial precipitation in the northern plains, transporting heat and moisture upwards, and reinforcing the structure and activity of the monsoonal trough. Although these depressions usually weaken after crossing central India as their moisture supply wanes, some can be reinvigorated by the addition of moisture from the Arabian Sea branch, thereby restoring their intensity to produce heavy, sustained rainfall above western India (Pant and Rupa Kumar 1997).

Another important force shaping monsoonal rainfall is the impact of orography on spatial patterns of precipitation, such as in northeastern India where moisture-laden air is forced to rise against steep slopes, leading to some of the planet’s highest rainfall totals, as in Cherapunji (1087 cm of mean annual rainfall) and Mawsynram (1141 cm of mean annual rainfall) (Das 2002). A similar effect occurs along India’s west coast, where saturated onshore winds are forced up the Arabian Sea facing slopes of the Western Ghats, producing copious rainfall along
The amount of precipitation taking place over the Indian subcontinent varies over time and space, ranging from arid conditions in the northwest to a moist tropical regime in the far northeast and the western coastal strips, as is evident from the distribution of normal rainfall (Figure 4). Moreover, the amount and frequency of precipitation can be highly variable from year to year (Figure 5). The monsoons also undergo periodic pulses of active and weak cycles. For instance, Murakami (1977) discovered that the Indian monsoon has two significant peaks in its intra-seasonal periodicity, one at about five days (related to cyclonic systems embedded in

**Monsoon variability**

Figure 3  Normal withdrawal dates of southwest monsoon season over the Indian subcontinent. Source: IMD 2013.

the Malabar and Konkan coasts as far north as Mumbai.
Figure 4  Spatial patterns of normal seasonal rainfall June–September. Source: IMD 2013.
the monsoonal flow) and another at around 15 days (related to the strength of the monsoon itself). Krishnamurti and Ardanuy (1980) found a 10–20 day westward propagating modes and rainfall over India. Other studies that found significant periodicity in the monsoon rainfall, including 30–40 episodes of period in cloudiness and precipitation (Yasunari, 1979) as well as recurrence period of 40 days (Singh and Kripalani, 1986).

As well as the intra-seasonal variability observed in monsoon rainfall, there are substantial inter-annual variations in such rainfall across the subcontinent as shown in Figure 5. The analysis of 31-year moving averages of all India monsoon rainfall revealed two multidecadal periods of deficient rainfall (1901–1930, 1961–1990) and two unusually wet periods (1871–1900, 1931–1960) (Pant and Rupa Kumar 1997). Furthermore, significant spatial variations in monsoon rainfall have also been observed, such as the presence of a decreasing trend in monsoon rainfall across most parts of the country that ended in 1962 in northeastern India and West Bengal in 1972, while persisting in north-central India. According to the latest report of the Intergovernmental Panel on Climate Change (IPCC) there is medium confidence that the Indian monsoon precipitation is projected to increase (Stocker et al. 2013). The report also mentions the overall weakening of the monsoon circulation, which is compensated for by elevated levels of moisture leading to great precipitation. It is important to understand the possible causes for these temporal and spatial variations in rainfall over the years, because the amount and timing of rainfall is of crucial importance to socioeconomic conditions in this region. Some of the major driving forces

Figure 5  Average all India monsoon rainfall 1901–2010. Source: IMD 2013.
controlling the performance of monsoon rainfall across the region are as follows.

1 *Teleconnection and monsoonal precipitation*: one of the most widely researched teleconnections in relation to the southwest monsoon rainfall over the subcontinent is the El Niño–southern oscillation (ENSO). One of the first studies to link the southern oscillation and monsoon rainfall was undertaken by Walker (1924); it was further validated by later studies (Parthasarathy and Sontakke 1988). There is now a general consensus about the occurrence of droughts during El Niño events (the warm phase of the ENSO) in comparison to La Niña events (the cold phase of the ENSO). However, the relationship between the ENSO and Indian monsoons is not always the same, such as the occurrence of normal rainfall during the 1997 monsoon season that signaled a weakening association between the ENSO and monsoon rainfall during the final decade of the twentieth century (Kumar, Rajagopalan, and Cane 1999). Going further, Torrence and Webster (1999) noted substantial decadal-level changes between the ENSO and monsoon rainfall. More recently, the modifying roles of the Pacific decadal oscillation (PDO) (Krishnan and Sugi 2003; Sen Roy, Goodrich, and Balling 2003) and the Madden Julian oscillation (MJO) (Kane 2000) concerning the influence of the ENSO on Indian precipitation has been observed. Some other teleconnections that appear to play a role in the performance of the Indian monsoon include the quasi biennial oscillation (QBO), arising from the interaction between nonlinear intra-seasonal oscillations and annual cycles (Chattopadhyay and Bhatla 2002), the Indian Ocean dipole mode (IODM) (Gualdi et al. 2003), and the North Atlantic oscillation (Sen Roy 2011). For instance, during the developing phase the heat content and sea surface temperature variability associated with the IODM are modulated by a local response of the ocean to winds as well as a long-distance response to the agitation of Kelvin and Rossby waves (Gualdi et al. 2003).

2 *Eurasian snow cover*: the Eurasian snow cover is increasingly regarded as one of the driving forces of the monsoonal circulation across the subcontinent. The presence of extensive snow in northern Asia during the premonsoon months implies significant regional energy consumption in snow and ice melt plus the evaporation of water from melt-saturated ground. This leads to reduced premonsoon surface heating and results in weaker thermal lows (McGregor and Nieuwolt 1998). Analyses of empirical data underscore the delayed arrival of monsoon when Eurasian snow cover is greater (Barnett et al. 1989), and the negative correlation between them has increased markedly since the mid-1970s (Liu and Yanai 2002).

3 *Sea surface temperature (SST)*: variations in SSTs in the adjacent arms of the Indian Ocean play a major role in the overall monsoon rainfall across the Indian subcontinent. In general, the warm SSTs of the Arabian Sea result in above-average rainfall and vice versa (Shukla and Misra 1977). Additionally, the study of 10-day and monthly mean evaporation rates over the Arabian Sea and Bay of Bengal has revealed that evaporation was higher above those water bodies during a low-rainfall year (1987), indicating minimal influence of SSTs on subsequent monsoon activity on the subcontinent. In contrast, evaporation above the southern Indian Ocean was greater in July and September 1988 when compared to
the same months of 1987 (Ramesh Kumar and Schlüssel 1998). In yet another recent study Sen Roy, Goodrich, and Balling (2003) showed substantial spatial variations in the relationship between SSTs and monsoon rainfall, in the form of positive coefficients in peninsular India, while predominantly negative coefficients were observed across most of northern and eastern India.

Aerosols: recently, the Indo-Asian region has received substantial attention for its significantly higher concentrations of aerosols and their projected impacts as a result of several major field studies, which include the Indian Ocean Experiment (INDOEX) (Rasch, Collins, and Eaton 2001), the Arabian Sea Monsoon Experiment (ARMEX) (Moorthy, Suresh Babu, and Satheesh 2005), periodic observations carried out in the Bay of Bengal, and the recently concluded Gangetic Valley Experiment (GVAX). Some of the main findings from the INDOEX were the significant role of the Indian subcontinent in the circulation of anthropogenic aerosols from the landmass to adjacent oceanic areas (Reddy et al. 2004). The results from this experiment also revealed the growing impact of increased levels of aerosols on precipitation patterns across the subcontinent (Satheesh and Ramanathan 2000). The spatial patterns indicate a high concentration in the Gangetic basin, raising concerns about aerosol impacts on both social and physical processes (Kaufman, Tanré, and Boucher 2002), seasonal temperatures (Sen Roy, 2008), and the monsoon precipitation (Lau and Kim, 2006). Specifically, an increased loading of absorbing aerosols over the Indo-Gangetic Plain during the premonsoon season is directly related to the increased heating of the upper troposphere. This results in the formation of a warm-core upper level anticyclone over the Tibetan Plateau in April to May, which leads to increased rainfall over the Indian subcontinent (Lau and Kim, 2006). In another study, by Meehl, Arblaster, and Collins (2008), the role of black carbon aerosols in decreasing precipitation trends over parts of India was revealed.

In conclusion, it is evident that the monsoonal rainfall, which is extremely critical for this region, has significant spatial and temporal variations that are the focus of ongoing research studies. The role of processes at the global scale and processes related to global climate change need to be taken into consideration for the full understanding and accurate forecasting of monsoon in the subcontinent.

SEE ALSO: Climatology; Clouds; Intergovernmental Panel on Climate Change (IPCC)

References


Ocean as Simulated by a CGCM.” *Climate Dynamics*, 20: 567–582.


Moral geography

Philippa Williams
Queen Mary University of London, UK

Moral geography concerns the relationship between geography and moral philosophy. It is about the intersections between morality, space, and power within both the human and the nonhuman, and the material and the symbolic, dimensions of life. Some would say that it is the idea that people, things, and practices belong in certain spaces, places, and landscapes and not in others. To appreciate the meaning of moral geography, it is important to outline its key concepts. The idea of morality refers to the prevailing understanding by which things are judged to be right or wrong in an absolute sense, or better or worse in a relative sense. Such ideas often underpin the values by which individuals, institutions, or communities judge principles or standards in behavior, as well as what are understood to be important aspects of life. This capacity to reflect on and to reason about normative values, about how things should or ought to be, which things are good or bad, and which actions are right or wrong is what distinguishes humankind from other creatures. This practice of reflecting on such moral questions is also referred to as ethics, and forms a part of moral geography. Geographers, unlike moral philosophers, are less preoccupied with understanding whether something is judged to be right or wrong, moral or immoral, and more concerned with the ways in which moral meaning and value become produced and reproduced in particular places and articulated through space and time.

Moral geography therefore involves uncovering the social and spatial politics of morality, interpreting how moral meaning is infused with practices of power, and examining the implications of this reality for different people and different environments, both today and in the past. Moral geography connects to both the realities that geographers seek to know (ontology) and the way that geographers go about acquiring that knowledge (epistemology). And it concerns both the descriptive content of geography, as well as the actions of geographers, as researchers, practitioners, and teachers. Academics and students of geography often celebrate the discipline’s potential to change the world for the better, but, it is important to recognize that the meaning of “better” is highly subjective and depends on normative values produced through society. What is better for some people may not be better for others. Acknowledging the ways in which moral values or norms suffuse all research, practice, and teaching is therefore of paramount concern for all geographers.

Geography and the “moral turn”

The shift toward applying a “moral lens” (Philo 1991) within geography should be understood in the context of the “normative turn” in social theory, and geographers’ response to empirical realities during the 1970s and 1980s. During the “quantitative revolution” of the 1960s, the dominant way of seeing and understanding the world in social science was through a positivist lens that clearly demarcated the factual from the nonfactual, and rested on the capacity to determine the “truth” through a quantitative,
MORAL GEOGRAPHY

often technological, approach to social matters. The 1970s was a period shaped by the specter of poverty and uneven development, and not just for the “third world.” Notions of difference and identity, in particular race and gender, were becoming prominent political issues alongside growing concerns about an environmental crisis caused by rising pollution and the depletion of resources in the name of economic growth. These concerns underpinned a growing moral crisis that provoked a normative reaction in radical geography for greater social relevance within the discipline. Geographers began to place increasing emphasis on the ways in which social meanings and categories come to take on the appearance of truth or objectivity.

Although geographers already had deep attachments to topics charged with moral meaning, such as race, ethnicity, gender, and sexuality, the ethical or moral import of this work had not been the primary intellectual focus. This changed in the early to mid-1990s when the discipline underwent a moral turn that was reflected in the tabling of conference panels in the United Kingdom and the United States with explicit links between geography and ethics, such as the panel on “A/moral geographies” at the 1995 IBG–RGS conference, as well as the establishment of a new journal Ethics, Place & Environment. This intellectual shift was further augmented by the publications of geographers such as Sack (1997), Smith (1997), and Proctor (1998) whose work developed more explicit dialogues with moral philosophy and ethics in order to advance critical perspectives on the ways in which the spatial is implicated with moral meaning, such as justice, fairness, and difference.

These geographers engaged with different dimensions of moral philosophy such as metaethics, descriptive ethics, and normative ethics to inform their approaches to moral geography, and these continue to represent an important framework for understanding moral geography today. First, metaethics involves the big questions about what it means to think or do ethics. This is a hugely varied and contested area within moral philosophy but one that prompts important questions for geographers about the potential to accept the universality of certain grand moral values, and to recognize the spatial (and temporal) particularity of their application. Proctor (1998), for instance, distinguished between a “thin,” minimalist, or universal morality, captured by such grand values as justice and truth, and a “thick,” particular, or local morality that is culturally sensitive and situated. Drawing on this work, Smith argued that the task of the geographer is to “examine the contextual thickening of moral concepts in the particular (local) circumstances” of contrasting human contexts (1997, 587). Second, descriptive ethics identifies actual moral beliefs and practices and, as such, constitutes one of the most productive avenues for geographical research into spatial and social patterns and relations that invite a “moral reading” (Smith 1997, 587). In the context of geography’s rediscovery of the importance of context to geographical enquiry, Philo (1991, 19) proposes that an understanding of local culture alerts us to the geography of everyday moralities “which ‘glue’ together the assumptions and arguments of particular peoples in particular places.” And, third, normative ethics concerns the study of ethical action: how one ought to act. This prescriptive aspect of moral geography underpins concerns that geographers have a moral obligation to improve the world, and to effect positive change; some go so far as to say that geography should be a moralizing force within school and higher education. At a conference in 1990 on social and cultural geography Philo presented the case for geographers to consider the moral positions that they and other people take in and toward their lives without
becoming moralizers themselves. His views echo those of other scholars who caution that all geography is underpinned by certain norms and values that may not be universally constructive.

While moral issues were not entirely new to geographers, the moral turn did represent a shift toward a more coherent and visible relationship between geography and moral philosophy within the discipline. This agenda revolved around two key axes, first, geographers’ commitment to examining morality—the “geographies of morality”—and, second, the recognition that there are also ethical and moral dimensions to these geographies—the “moralties of geography.”

Places, spaces, and scales of moral geographies

The topic of moral geography has implications across the discipline of human geography and has developed significantly since the moral turn, reflecting changes within the discipline as well as responses to global experiences and transformations that have prompted new ethical questions concerning justice, equality, and power. It is possible to identify three key interlocking threads of inquiry that have attracted significant geographical attention.

The first of these concerns the application of a moral lens to questions of development, justice, and difference. Early precedents of work in this area include the place of justice and the spatial in urban and regional planning, as well as the relationship between people, nature, and matters of social justice in the city. As a key architect of the moral turn, David Smith (2000) built on these studies and directly applied questions of morality, ethics, and geography to understand how notions of difference and social justice were differentially experienced in contrasting places. Focusing on South Africa after apartheid, the changing moral landscape of an East European city, and conflict over land and settlement in Israel/Palestine he showed how such divisive histories and contested presents fundamentally shape human experience. This scholarship emphasized the significance of context for understanding morality and recognized that different places inform different moral geographies. It also highlighted the importance of applying a geographical lens to these questions.

Geographers’ engagement with notions of justice and the city have transformed over time. Scholars have shown how the focus has shifted beyond conceptions of wealth and poverty to more explicit concerns about difference. Second, geographers have become increasingly concerned with the value of identifying and constructing normative agendas for justice and with questioning the extent to which this is productive for challenging specific injustices. Third, there is growing debate around the ontological status of the spatial where some scholars see the spatial as a derivative of the social. More generally, it is interesting to note that geographers have typically documented instances of injustices rather than searching for examples of where justice is being realized.

Other scholars have considered ethics in development geography to debate whether, and in what ways, people in more privileged locations should intervene in the lives of less privileged people elsewhere. Shifting geopolitical relations underpinned by the rise of Brazil, Russia, India, and China have challenged US hegemony and prompted scholars to examine how different norms underpin different development partnerships. For instance geographers have examined the ways that North–South aid is often portrayed as unreciprocated charity to those less fortunate, whereas a positive moral tone of reciprocity is attached to development aid transacted between South–South partners. In the field of geopolitics scholars have also taken issue with the ways that geopolitical analysis has perhaps unwittingly
been infused with normative moral judgments concerning the “just” war, even while geopolitics itself eschews the adoption of normative political positions. The contention of some is that geopolitics has failed to imagine nonviolent alternatives and consequently needs to demonstrate greater reflexivity concerning the normative reasonings that underpin its analysis of war. As a result, the impact of geography on debates concerning military intervention outside of the discipline is arguably limited.

In other areas, the economic expediency of moral values such as justice, equity, and fairness has not been lost on scholars. A number of geographers have pointed to the political work done by certain value agendas and how these have become standardized in the neoliberal market economy. One example of these value agendas pertains to fairtrade, which was established on an ethic of care to alleviate poverty and economic injustice through a market-based form of solidarity exchange, but in reality these objectives are rarely realized on the ground. Scholars have revealed how the ideological power of fairtrade’s moral claims lies in the normalizing agenda of neoliberalism that obscures the unequal practices of power that sustain its existence and makes it both apolitical and irrefutable.

The second strand of geographical inquiry is linked to the notions of difference and justice, and concerns the complex relationship between distance, care, and responsibility. David Smith (2000) critiqued earlier calls by geographers to transcend local particularism in favor of universal values, and instead highlighted the productive tension in seeing moral universalisms and particular moralities together. This revised approach ultimately shifted the scale and scope of moral inquiry and action, and prompted geographers to re-evaluate perspectives with respect to ideas about “care at a distance” and “geographies of responsibility.” The question that animates geographical approaches concerns how attention to ethics can motivate a sense of responsibility toward different and distant others, as well as toward the people with whom we have more intimate and everyday encounters. Jeff Popke (2006) recognizes two responses to this position, first by linking care to more or less universal notions of justice against which acts of judgments can be made. In this context, responsibility toward others is ensured through collective rules, policies, and practices that embody and nurture an ethic of care. A second response emphasizes care as a fundamental activity of humanity and locates responsibility in the recognition of our intersubjective being. Scholars have turned their attention toward ethical consumption, which can be conceptualized as an act that links local practice with care at a distance, and embodies a form of ordinary or everyday ethics.

In her work on the geographies of responsibility Doreen Massey (2004) proposes that we need to go beyond the local and the politics of propinquity to think seriously about how our lives and actions are connected to and have implications for faraway places, and vice versa. She urges us to understand distant space in the same way that we think about distant time or “the past,” and therefore as something that is intrinsic in the present, and so too in place. As she argues, “we are responsible to areas beyond the bounds of place not because of what we have done, but because of what we are” (Massey 2004, 16). Other scholars have drawn on Kant’s notion of the perpetual peace and the imperative that we tolerate one another as neighbors to stress the spatially extensive sense of our responsibility toward others, our cosmopolitan responsibility. This political and ethical stance is particularly important with respect to an increasingly urban world that augments the thrown-togetherness of people, a world patterned by migration that demands certain forms of hospitality and the
legacy of postcolonial difference. Within these contexts the ethical imperative for dialogue, negotiation, and openness to and with the other are argued to be paramount.

The third strand of moral geography concerns “moral landscapes” and relates to the ways in which particular spaces and their social relations become imagined and organized in conjunction with moral meanings and values. This area of study owes much of its intellectual inheritance to an article by Driver (1988) on social science and the urban environment. Through an analysis of sanitary science, moral statistics, and medical geography, Driver pointed to the inherent moral concerns of human and social sciences, and proposed that, rather than simply reflecting patterns of urbanization, social science is in itself a moralizing force for the organization of space and society. Another important dimension concerns how landscape becomes linked to ideas about morality and how schemes of moral value are located in particular environments. Scholars had examined how space and society are organized to reveal the hierarchical constructions of “moral order” in the city that excludes morally marginal communities, the lower economic classes, and areas beyond the reach of economic planning. Largely focusing on urban and nineteenth-century concerns, historical geographers have explored a wide range of subjects from the moral locations of soldiers and sailors in nineteenth-century Portsmouth and mid-nineteenth-century New York fashions to attempts to regulate gender and consumption, alcohol, and prostitution in different sites. Other geographers have shown how rurality can be a repository of moral values and normative assumptions, and mark the people who live there as well as their dispositions toward life. David Matless (1997) has explored how visions of the English landscape promote various forms of order in contrast to disorderly characteristics associated with “untidiness” or “bungalows” that are simultaneously connected to appropriate forms of behavior and ultimately to national identity and citizenship. Meanwhile Duncan and Duncan (2003) show how aesthetics in an American suburb inform “landscapes of privilege” that constitute identities and promote patterns of distinction and exclusion.

Further developing this understanding of sociospatial inclusion and exclusion, scholars have shown how certain people are marked or stigmatized as outsiders deemed to be “out of place” because they do not conform to expectations about what the norm is. Geographers have examined these kinds of processes and their implications for different groups of outsiders, notably mad people, political protestors, non-white people, those who are homeless, deviant, or disabled, and homosexuals, among others. More recently geographers have focused on the moral dimensions of social encounters in our everyday lives to explore the moral evaluations we make about others concerning their social, economic, and cultural worth and how these give rise to different forms of class-based prejudice. This analysis shows the ways that poverty is deemed to be a personal moral failing rather than an intrinsic part of capitalism. Some have argued that, in the contemporary context of superdiversity and supermobility, this kind of prejudicial behavior is structured by a hegemonic ethic of individual self-interest that threatens the existence of socialized processes such as care, compassion, and responsibility.

Practicing moral geography

As well as applying a moral lens to geographical issues, a number of geographers have also highlighted the importance of both employing professional ethics during the research process and
teaching about ethical actions through geography education. Many argue that human geography, and its concern for the interactions between people and their relationship with the environment, is well placed to instruct students about their moral accountability as responsible citizens.

As Sack (1997) has argued, geography is now a discipline oriented toward improving the lives of others rather than in terms of self-interest or traditional practices. As such, he cautions that research should not be exploitative but should make a practical contribution to the lives of the people being studied. Others contend that geography should go further than simply “doing no harm” and instead actively contest hierarchies of power and enact transformative practices. Indeed, there is a rich history of social activism in geography motivated by an ethical commitment to make a difference or achieve social change; however, Cloke has expressed a degree of skepticism about the level of actual transformations and practical action that is realized by geographers, for, “whilst it is easy to think and write about ethical and moral issues, it is much more difficult to actually do something about acts of injustice in our everyday lives” (2002, 587). Cloke suggests that geographers are adept at recognizing difference and differentiating processes, but are less well equipped with the imaginative and practical guidelines required to address inequalities.

In his own research career Cloke has worked closely with homeless people and associated organizations, in both intellectual and practical capacities, and urges other scholars to forge stronger connections between their academic and nonacademic lives. It is becoming increasingly difficult, however, for scholars to live ethically and to act politically in the context of a neoliberal university system that increasingly values scholarly activity in terms of research output and funding grant incomes. Some scholars are finding ways to create opportunities within this environment, aligning their intellectual, political, and ethical commitments.

In order to act politically and effect change, geographers have been encouraged to identify appropriate ethical and political practices among the issues they study and to propose normative orientations. Some scholars have hesitated at the notion that geographers should be proposing their own moralizing frameworks, while others argue that geography has not done enough to inform political and moral philosophy and to shape notions of universal practice. The role of geography in documenting everyday moralities and proposing normative moralities is therefore contested.

Beyond the moral turn

Recent geographical debates about morality have turned inward, to reflect on how geographical scholarship deals with moral concerns and to consider the implications of particular normative assumptions for the production of geographical knowledge and understanding. Barnett and Land (2007) highlight the value of scholarship concerning care at a distance and geographies of responsibility. However, they take issue with the assumptions that underpin this work, in which placed-based local relationships are prioritized as the sites in and through which care happens and knowledge is produced, whereas caring at a distance is more problematic. Barnett and Land reorient these normative positions and suggest that people are already motivated to care by all kinds of factors, and that more knowledge does not necessarily drive responsible actions but can also create feelings of powerlessness. Instead, they propose generosity as a political concept that can help to challenge these assumptions about how we relate to others and expand a view around how responsible actions are both received and
responded to. In response to these ideas, scholars have suggested the need for a metaethical approach that brings together a theory of justice and virtue ethics.

In a series of reports for *Progress in Human Geography*, Barnett (2011, 2012) returns to contemporary moral and political philosophy to critically evaluate the relationship of geographical approaches to normative concepts, including justice, reason, and worth. He is concerned that a good deal of debate in geography has involved whether it is possible to develop normative foundations for critique, and yet contemporary moral and political philosophy has shifted toward more pragmatic, intuitive, and socialized understandings of ethics and practical reason. In line with this approach, Barnett proposes that theorizing justice, for instance, demands a more expansive approach that recognizes the situated and relative experiences of justice or injustice, rather than be constrained by the search for an absolute notion of the term. This approach might also encompass the importance of feelings about justice and injustice and turn attention to an emotional geography of justice. Meanwhile, other scholars have cautioned against seeing situated interpretations of justice as entirely independent of broader theoretical constructions of the concept, and emphasize the need for a dialogical understanding of justice.

More generally, Barnett has proposed that geographers would do well to rethink their disposition toward normativity. Moral geography’s concern has revolved around two axes – proposing normative moralities and charting everyday moralities; both approaches reflect a degree of skepticism about the degree to which normativity structures everyday life and the destructive role of power. And yet, Barnett contends, if we learn from shifting interests in moral philosophy, it may be possible to go beyond an excessively cynical view of the social as regulated by power, and instead expand our perspectives about why things matter to people, the nature of worth, and ultimately a broader view of the normative dimensions of social life in ordinary, routine, and practical ways.

**SEE ALSO:** Caregiving; Cosmopolitanism; Difference; Historical geography; Social justice

**References**


MORAL GEOGRAPHY


Mortality

James A. Tyner

Kent State University, USA

Mortality is commonly understood as the state or condition of being dead. So defined, the term acquires a particularly biological – and individual – quality. Mortality is what happens when a living organism, say a man or woman, dies. In actuality, though, mortality is a much more complex concept in that many political, economic, and cultural practices pivot on the conceptualization of death. Indeed, as Todd May (2009) explains, from an anthropocentric viewpoint, the fact that we die is the most important fact about us. Not only does death negate our mortal, material existence, the awareness of this inevitability is profound. For accompanying the inevitability of death – of our own mortality, as well as that of all other living creatures – is the uncertainty of death. We know that death will arrive; we just do not know when. Together, the inevitability and uncertainty of death has generated much scholarship toward the study and objectification of mortality.

Our understanding of death is far from simple. In many societies, throughout time, death has been understood as the termination of vital signs: the beating of the heart, the cessation of breathing. Ultimately, and most clearly, death was marked by the decomposition of the corporeal body. With advances in science and medical knowledge, however, the clear-cut separation between “life” and “death” has become blurred. This is more visible in the case of “brain death.” Beginning in the 1950s, but especially following a seminal report prepared by an ad hoc committee at the Harvard Medical School, the concept of brain death, and of a patient being brain dead, became established. This singular concept has arguably transformed our understanding of mortality more than any other development.

Prior to the twentieth century, when a person’s heart stopped or he/she stopped breathing, death was all but inevitable. With the advent of techniques for artificial respiration, coupled with advances in resuscitation techniques (e.g., cardiopulmonary resuscitation, CPR), death could be (temporarily) prevented. Later, with the development of “iron lungs” and the successful transplant of organs, death was once again delayed. The concept of “brain death,” however, and the attendant condition of being brain dead, changed forever how life and death were to be understood. Of fundamental importance is the fact that death has in some situations become relational. Whereas CPR is a practice conducted on one body, with the intent of saving that body’s life, the practice of organ transplantation is dependent upon two bodies: the organ “donor” and the organ “recipient.” This co-relational condition of life and death created a “demand” for organs – but also the necessity for a class of corporeal bodies that would occupy a threshold between “life and death.”

Biomedical advances made heart, lung, and kidney transplants (among others) possible. However, a major cause of failure in organ transplants was the use of cadaver organs that had deteriorated during or after the conventional death of the organ donor – waiting too long for a donor to die risked damaging viable, transplantable organs and thus an inability of saving the recipient-patient. For many medical
practitioners, therefore, interest was focused on those patients who were being kept alive via artificial means. The argument was that if a patient was irreversibly “dead” – meaning that he or she had suffered permanent brain damage; or was in a permanent “vegetative” state, it should be possible to retrieve those organs to save the life of another person. The concept of “brain death” subsequently entered the medico-legal lexicon and remains a contested ethical and legal minefield with a unique geography. In the United States, for example, there is no universally recognized medical consensus on the definition of brain death. Four different criteria are in use and individual states determine which criteria to follow. The result is that when someone is pronounced dead may be a function of where that person is pronounced dead.

Brain death as a concept calls attention to the fluidity to which life and death are understood and measured. There are many scientific terms and measures related to mortality. Life span, for example, refers to the oldest age to which any living organism can survive: in effect, a biological ceiling or limit to life. Biologists estimate that the life span for humans is approximately 120 years. By way of comparison, elephants are premised – under natural conditions – to live between 60 and 70 years. The qualifier “natural” is important in that it directs attention to the observation that most humans do not live to be 120; nor do most elephants live to be 70. For this reason, demographers and other social scientists refer to a concept known as “longevity,” which is the ability to remain alive from one year to the next. Longevity is expressed by a measurement known as “life expectancy” – the statistical average length of life. Currently in the United States of America, for example, the life expectancy of females is 81 years while that of males is 76; this compares to Somalia, where the life expectancy of females is 56 and that of males is only 53.

Differences in mortality, as indicated by life expectancy, are explained in part by biological factors (e.g., genetics); however, most inequalities in mortality result from societal conditions, such as a lack of access to health care and adequate nutrition. Geographic variability of observed and documented patterns of mortality, in other words, reflects unequal opportunities for life. Men and women, on average, live longer in the United States than their counterparts in Somalia not because of inherent or essential biological differences but rather because of deeper structural conditions that result in disproportionately higher rates of mortality in the latter.

The existence of inequalities in mortality has prompted geographers and other social scientists to confront “premature death” as an indicator of social injustice. Ruth Gilmore (2002), for example, argues that the cause of premature death is an application of violence. The salience of Gilmore’s argument is that it places the study of mortality squarely within the growing body of scholarship on violence (Tyner 2012). Here the purpose is to identify the direct and indirect causes and conditions – the myriad forms of violence – that result in death. It is common, for example, to distinguish between “direct” forms of violence and “structural” forms of violence. The former refers to interpersonal actions, such as hitting, stabbing, or shooting whereas the latter includes those actions (or inactions) where there is no identifiable actor. Poverty and malnutrition are considered forms of structural violence; mortality subsequently results from indirect conditions that may not be linked to any particular individual.

Recent scholarship has blurred the distinction between direct and structural violence; many alternative concepts and theories have been forwarded, including institutional violence, administrative violence, cultural violence, epistemic violence, and so on. Of note is that most conventional studies of mortality continue
to focus primarily on “direct” violence (e.g., homicide), secondarily on “structural” violence (e.g., famine); decidedly little on mortality has engaged in the broader debates on what exactly is captured by the term “violence.”

Beyond directing attention toward violence, an engagement with premature death is informative for a second reason, namely the politics of vulnerability. To begin, however, it is necessary to consider more closely the conceptual origins of premature death. Prior to the eighteenth century, and notably in Western societies, the knowledge of life and death were located within the realm of cultural and religious traditions. Death was a singular “moment” that marked the passage from one form of existence (life) to another (after-life). During the Enlightenment, however, a new demographic approach to the concept of mortality emerged, a transformation that coincided with the foundation of the modern state. What emerged was a biologics, a science that positioned the knowledge of life within a broader context of statistical patterns and regularities. Significantly, premature death was conceptualized as a “preventable risk”; this risk depended on the development of new scientific techniques for measuring longevity and mortality rates and for identifying the most common causes of death (Bayatrizi 2008). Moreover, premature death was inseparable from the growing political and economic interest in the management of life and death – exemplified best by the creation of life insurance. Simply stated, underlying the biological understanding of mortality – and the myriad demographic concepts of life expectancy, life span, and other vital measures (e.g., infant mortality rates, material mortality rates) – is a form of governance that continues to inform political practice.

The governance of mortality is well-illustrated in the voluminous work being conducted in on biopolitics, necropolitics, and even security studies (Rutherford and Rutherford 2013). These bodies of scholarship direct attention to the calculated management of life and death. Crucial here are considerations of how sovereignty, the law, and violence interact in generating those structures of inequality that result in empirically observed differences in mortality. “Biosecurity,” for example, is frequently understood as the prevention or elimination of infectious diseases in crops, livestock, and humans. More recently it has been applied to encompass broader governmental mandates to protect and secure human health (Hinchliffe and Bingham 2008). According to some geographers, for example, biosecurity is emerging as a new expression of sovereignty. To this end Eugene Thacker (2005) argues that sovereignty today derives its legitimacy from its ability to continually identify potential and actual “diseases” that threaten the body-politic. And while these diseases are commonly understood in a literal fashion (e.g., influenza, HIV/AIDS), it is important to note that biosecurity threats assumed – or, more properly, are constructed as – many different forms. Both obesity and aging, for example, have emerged as biopolitical threats that are presumed to jeopardize the national security and financial solvency of states. Here, both corporeal conditions – of excess weight and excess age – are premised by some pundits as contributing to a drain on the costs of health care with vastly different endpoints. Obesity, as a case in point, is framed as personal failure; but a failure that threatens to bankrupt the nation. From this position, premature death – which is held to be preventable through exercise and diet – is viewed not as an indicator of violence but rather as an indicator of a private failure to prevent one’s own death.

A biopolitics of mortality would appear to take us far afield from the realm of the biologic. In fact, recent scholarship indicates not a separation but rather a greater elision between the two.
MORTALITY

Consider the truism that – regardless of governmental intervention and prevention – all living organisms will die. Expanding on this idea, Judith Butler (2004) posits mortality as an equalizing possibility, one that hinges on the distinction between what she terms “precariousness” and “precarity.” Given that all humans, regardless of gender or sexual orientation, ethnicity or nationality, must at some point die, mortality holds the potential to create a shared morality. Recognition of mortality, Butler premises, constitutes an appreciation of “precariousness”; a form of corporeal vulnerability in which the fragility of life is shared by all mortals – including the privileged. As the empirical accounting of mortality, that is, differences in life expectancy, indicates, however, not all humans are equally vulnerable. In opposition to precariousness is the concept of precarity – the particular vulnerability imposed on the poor and disenfranchised: those living bodies endangered by war, natural disaster, and poverty – in short, the many structures of violence that unevenly affect life. Butler (2004) maintains that we are all subject to potential precarity through our common precariousness. In making this controversial argument, Butler effectively re-centers mortality into the political-economic realm and views our shared mortality as a path toward the promotion of a more socially just society.

While Butler’s optimism is perhaps debatable, the awareness of the finitude of death has contributed to a widespread engagement by geographers, anthropologists, sociologists, and other social scientists on the cultural meanings of mortality (May 2009). This research highlights that the fear of death and the awareness of mortality significantly inform the organization of all societies. Zygmunt Bauman (1992) explains that the fact of human mortality and the necessity to live with the constant awareness of that fact go a long way toward accounting for how societies interact. Geographers, for example, have long engaged in the study of “mortal geographies” – of how different cultures remember the dead. This is seen in the numerous studies that have documented the location, distribution, and cultural significance of cemeteries and other deathscapes (Maddrell and Sidaway 2010). This research has highlighted, among other aspects, the significance of “death” for the living and the politics of mourning. A related body of literature has focused on the remembrance and memorialization of death. Here the emphasis is on the contested meanings of death itself. More than simply a biological indicator of death, mortality is again revealed to be a peculiarly human condition informed by cultural, political, and economic factors.

SEE ALSO: Biopolitics; Biosecurity; Demographic and epidemiological transition; Population geography; Poverty; Security; Sovereignty; Violence

References


Mountain biogeography

Zehao Shen
Peking University, China

The discipline

The idea that Earth life originated from a mountain island has been in the tradition of Western culture tracing back to the Deluge. In accord with the legend from the Genesis, the greatest botanist and biogeographer before the nineteenth century, Carolus Linnaeus (1707–1778) hypothesized that Noah’s Ark landed at the Paradisical mountain – Mt. Ararat in Armenia. This mountain, with a significant elevational spectrum of climate types, harbored the animal and plant species adapted to different environments, and became the origin center of all creatures, which came down the mountain after the flooding and migrated around the Earth.

The voyages of exploration in the eighteenth century helped to establish a global perspective of modern biogeography. Covariation of climate and vegetation along elevational gradients, as well as the resemblance to the latitudinal pattern, were first elaborated by Alexander von Humboldt (1769–1859), the founder of phytogeography, after his field expedition to Mt. Chimborazo of tropical South America. In the same period, Augustine de Candolle (1778–1841) first pointed out that mountains act as one of the barriers to the migration of plants and animals, resulting in the boundaries of species ranges. Although the pioneers of this field invested their major efforts in the mountainous regions around the world, the most fascinating questions and insights in biogeography clearly came from the observations of insular biota, as did the Galapagos Archipelago for Darwin and the East Indies Archipelago for Wallace. Limited area, isolation from continents, and highly endemic biota combine to make an island (or a group of islands) a model system to all insular habitats for studying the patterns and causal factors of biodiversity, a fundamental question of not only biogeography, but ecology and evolution as well.

The equilibrium theory of island biogeography (MacArthur and Wilson 1967) provided the first theoretical model of biogeography. During the past half century, this dominant model has received hundreds of tests for the role of the spatial factors (i.e., area and distance) and processes (i.e., immigration and extinction) in the dynamics of island biodiversity. Mountains, among other locales, were taken as a typical insular context for testing MacArthur and Wilson’s theory. This is verified by the term of “sky islands” coined by Marshall (1957) to describe the mountains of southwestern North America, and its later applications to other mountains.

Although there is no such field specifically defined as “mountain biogeography” in the framework of modern biogeography, mountains are undoubtedly a special case of islands in the sense of spatial configuration, environmental pattern, and processes as well as biogeography itself. Given the equilibrium theory of island biogeography as the foremost theoretical paradigm of this discipline, the unprecedented environmental challenges in terms of climate warming, land use change, and biodiversity, conservation at global scale ask for an update of the
framework for biogeography, for which mountain biogeography seems to be an ideal candidate.

Environmental properties of mountains

Characterized by prominent topographic relief, mountains are only one type of geomorphology or terrain, however they are no doubt the hotspot of biodiversity of terrestrial ecosystems. Mountains cover about 25% global land area, while conserving 95% of vascular plant species, and 95% of bird, mammal, reptile, amphibian, and fish species (Kapos, Lysenko, and Lesslie 2000). Mountains possess distinct environmental properties that are critical in terms of biogeography.

Elevation gradient

Representing environmental variations independent of latitude, the primary biogeographical dimension, mountains are important for their elevation gradient, with the maximum of 8848 m in Mt. Everest of Himalaya, and the minimum of −392 m in the surface of the Dead Sea. Elevation gradient is accompanied by multiple environmental factors that are critical for the distribution of organisms:

Air pressure. Because of the effect of gravity, air density and also atmosphere pressure decreases as altitude increases, rapidly at first and then more slowly. Atmospheric pressure decreases roughly 50% (to around 500 mb) within the lowest 5500 m, above which, the pressure continues to decrease, at an increasingly slower rate, to about 10% (≈100 mb) at 12000 m.

Solar radiation. Reflected and absorbed by the atmosphere, solar radiation is weakened in total intensity and in the fraction of UV-B radiation, with the red/infra-red fraction increased, toward lower altitudes. But because clouds and fog often increase with altitude, there is not a consistent altitudinal pattern of solar radiation intensity or spectral composition, but a consistent global trend for maximum intensity of solar radiation and the UV-B fraction is found.

Climate gradient. The greenhouse effect of the atmosphere helps to keep a globally consistent lapse rate of temperature of about 0.6°C/100 m elevation, with regional and seasonal variability. In contrast, the altitudinal gradient of precipitation is more variable, although it generally increases (or at least keeps stable) from the lower elevations, but with the maximum precipitation height sometimes occurring at mid-elevation. Mountains as a whole are less windy on average, owing to the topographic shelter between or within mountain ranges, except for exposed crests, summits, and island mountains. Altitudinal climate gradient is often treated as an analog of the latitudinal gradient, although it is unique with regard to the latitude-related daily and seasonal cycles of climatic dynamics, and the 1000 times spatially compressed results compared with the latitudinal gradients.

Area distribution. The surface (or horizontally projected) area of altitudinal zones have a variable elevational gradient, depending on the relief and shape of the specific mountain.

Mass elevation effect

Mass elevation effect (or Massenerhebung effect) describes variation in the tree line (or other temperature-based biogeographical boundaries) based on mountain size and location. In general, mountains surrounded by large ranges will tend to have higher tree lines than more isolated mountains, due to heat retention and wind shadowing. This effect was first pointed out by A. de Quervain who found the treeline or snowline occurs at higher elevation at the central part of the Alps than at the outer margins.
Topographic variations

Dominated by geological structure and bed rock type but also reshaped by the climate, highly variable mountain physiography reveals heterogeneous environments at multiple spatiotemporal scales and dimensions, and provides habitat diversity for the organisms, in terms of solar radiation, soil moisture, wind velocity, snow accumulation, and so on. On the other hand, slope steepness acts as a primary driver of geomorphological and hydrological processes, regulates the pattern and dynamics of surface disturbance, as well as mountain landscape evolution.

Elevational geometric constraint

Measured by altitude, the range of mountain has geometric constraints on both ends of the elevation range, especially for the isolated mountains. Area decreases with elevation, and the upper end of mountains always means a hard boundary for distribution of the species.

Spatial isolation and connectivity

The spatial isolation within and between mountain ranges is ubiquitous. However, as the mountain bottoms are connected via lowland terrestrial ecosystems rather than water as in a real insular context, the media for terrestrial organisms should be much more permeable between the sky islands than water surrounding an archipelago.

Mountain uplift and evolution

Mountains (and mountain islands) have life cycles that are jointly determined by the agents of both the geosphere and atmosphere. On the one hand, inner-Earth energy uplifts mountains through the geological processes folding, faulting, and volcanisms. On the other hand, mountains are continuously worn by gravity-driven erosion. The height and configuration of a mountain, its area and elevation gradients, are the dynamic output of these interactive processes.

Patterns of mountain biogeography

Attitudinal zonation of vegetation and ecotone (treeline) dynamics

Zonation of altitudinal vegetation spectrum is a basic and prominent pattern of mountain biogeography, dominated by the compressed climatic gradient along the elevation. As first uncovered by von Humboldt, altitudinal vegetation zonation, including the type composition, altitudinal limit, environmental determinants, regional variation of mountains, as well as the global pattern of the altitudinal vegetation spectrums have been a fundamental part of mountain biogeography. The climatic determinants of altitudinal distribution of vegetation zones, as well as the response of vegetation types, particularly the vegetation boundaries (e.g., treeline) to climate change, have always been a contentious topic in mountain biogeography.

Elevational distribution of species and species richness

Although species distributions respond to climate gradients in an individualistic manner, at least two macro-ecological patterns reveal the holistic aspect of the response. First is the altitudinal vegetation zonation with significant boundaries (e.g., treeline) as indicators. Second, a consistent pattern, Rapoport’s rule, is speculated to regulate the altitudinal distributions of organisms, that is, species elevational ranges increase along with elevation itself (Stevens 1992).
The altitudinal pattern of species diversity, analogous to the latitudinal pattern, has been a core question of biogeography. After a half-century of study, four altitudinal patterns of species richness have been summarized, that is, decreasing, platform plus decreasing, hump-shaped, or insignificant (Rahbek 1995). The occurrence of a particular pattern and the contribution of the underlying mechanisms are still in discussion, based on mounting empirical and theoretical studies.

Species range disjunction, fragmentation, and endemism

Fragmentation within a mountain and disjunction within mountain ranges are normal patterns of intraspecific population distribution, providing higher probability of speciation. Endemism – restricted species distribution – is more common in mountainous regions than other areas, even though a general model for the elevational pattern of endemism is still elusive. The elevational pattern of endemism is generally different from that of species richness. However, in the insular environment, the rate of endemism tends to increase along the elevation gradient (Steinbauer et al. 2012).

Roles of mountains on species distribution and biodiversity

In the framework of island biogeography, species diversity in an insular environment is the sum of three processes: immigration, extinction, and speciation. These processes occur in ecological or evolutionary time scales and are affected by factors of these two categories, respectively. Contemporary mechanisms involve those factors hypothesized to play a role in the ecological time scale, while historical mechanisms are speculated to affect at evolutionary time scale and have cumulative impacts, through the interaction of three basic biogeographic processes.

Contemporary mechanisms

Climate limit on species distribution. The niche concept holds that each species occupies a preferred position and range width along the altitudinal gradient, regulated by the altitudinal climate gradient. Three basic hypotheses exist about the climatic constraint on species distribution: (i) productivity regulated by the combination of energy and water supply; (ii) ambient energy as a constraint on the rate of metabolic activity, generation duration, as well as genetic mutation probability; (iii) extreme low temperature limit and the distribution of different freezing tolerances. In the accumulated studies on the altitudinal pattern of species richness, climatic factors are able to account for more than 60% of variation in general.

Rapoport’s rule. Rapoport’s rule was speculated to apply to both latitude and altitude gradients, based on the hypothesis that the variability of climate (temperature or precipitation) increases along with elevation or latitude, and therefore species adapted to the habitat at higher elevation or latitude tend to have a larger niche width as well as a broader geographic range.

Topography, geomorphological processes, and feedback. Topographic heterogeneity in a mountain context typically reflects the content of the niche hypervolume concept, that is, a niche space with multiple dimensions and scales, including the temporal dimension as regulated by climate changes and the disturbance of geomorphological processes, as systematically documented in the alpine treeline. Mountain topography provides a diversity of habitats in a hierarchical framework to harbor the components of the mountain species pool. By niche differentiation in this context, the probability
of extinction driven by competitive exclusion is reduced, and the mountain acts as a museum of species diversity.

**Area.** Area is one of the two factors that affect species richness in the equilibrium theory of island biogeography; one interpretation is that species richness increases along with the accumulation of energy over larger areas. An unresolved issue is whether the larger surface area in mountains, relative to the planar projected area, contributes independently to the species richness.

**Mid-domain effect.** As a mountain has boundaries on both upper and lower ends of the altitudinal gradient, although not as a definitive at the lower end as for an oceanic island, the altitudinal range of species is geometrically constrained by the elevational boundaries according to the concept of mid-domain effect (Colwell and Lees 2000): random distributions of ranges will result in more overlap in the middle. The mid-domain effect produces a hump-shaped pattern of species richness along the altitudinal gradient. However, this hypothesis has been controversial for its assumptions, especially regarding the different range sizes of species.

**Historical mechanisms**

**Spatial isolation and disjunct distributions and speciation.** The debate between two biogeographical paradigms of disjunct species distribution and speciation – the vicariance model and the dispersal model – have lasted for more than a century, although both are based on spatial isolation, a basic property of mountain landscape and mountain ranges. Spatial isolation has potential impact on the three processes of migration, speciation, and extinction, depending on the interaction with other natural versus anthropogenic processes.

**Climate change and mountain refugia.** The Quaternary glacial reshaped the contemporary pattern of global vegetation and biodiversity, especially in the northern hemisphere. Refugia as a biogeographical concept refers to the locally favorable habitat condition for species to survive at a local scale in spite of hostile environmental change, especially climate change, at macro-scale. The role of glacial refugia is widely recognized in the survival of many high and mid-latitude species through the Quaternary glacial maximum in remnant populations in the northern mountains, instead of being completely eliminated or driven to the south, and migrating back after the glacial. The recently generalized refuge hypothesis includes micro-refugia of much smaller scale but more prevalent distribution. Although the glacial/interglacial cycling push the back-and-forth movement of species distribution, the survival in refugia or micro-refugia and recolonization is a rule rather than exception (Schmitt and Varga 2012). Glacial refugia in the northern mountains and interglacial refugia in the south mountains highlight the critical role of mountains in biodiversity conservation through the large climate changes of the Earth’s history.

**Mountains as the cradle of biodiversity.** Fossil evidence, genetic experiments, and phylogeographic studies repeatedly suggest that mountains act not only as a museum of biodiversity, by harboring species of different niches and reducing the extinction risk from competitive exclusion, but also play a critical role in the speciation process, in terms of both allopatric and sympatric speciation.

Through both spatial isolation and environmental differentiation, mountains act as a barrier to organism dispersal and gene flow between the populations of the same species. The following breeding isolation would eventually drive the separated populations into diverging evolutionary pathways. At large temporal scale, the fast uplift of mountain ranges was found to correspond with the increase of speciation rate in
time in the Andes and the Himalaya, supporting the idea of the mountain orogenesis process as a pump of regional biodiversity (Badgley 2010).

From island biogeography to mountain biogeography

As the first theoretical model for biogeography, the equilibrium theory of island biogeography suggested that species richness is the outcome of equilibrium between the immigration and extinction processes that are fundamentally determined by the area of the island and the distance between the islands or island to continent (MacArthur and Wilson 1967). The equilibrium theory has since been widely applied to the insular contexts in addition to islands, for example, mountains, caves, and even urban systems. Mountains, as “sky islands,” are an important locus for examining the theory.

The inadequacy of the equilibrium theory has been revealed by mounting tests, and the model was criticized for omitting the role of ecological factors, such as environmental heterogeneity, biotic interactions, as well as species traits in affecting insular biota. However, an early insight of island biogeography has also been reemphasized, namely the species specific island biogeography, which by integrating the evolutionary history of a particular lineage constructs a phylogeography (Lomolino, Brown, and Sax 2010). This approach aims at interpreting the biogeography of the lineage in terms of evolutionary adaptation to the critical environment variations, especially glaciations and mountain orogenesis.

Although mountains as a model are more representative to the isolated or fragmented domains generally applicable to all kinds of insular organisms, a theoretical framework for explicit mountain biogeography has yet to be develop. By absorbing the essence of island biogeography, and addressing the difference in isolation between mountains and real islands, as well as its impacts on the immigration, speciation, and extinction processes, mountain-based studies have a great potential to provide critical insights into the development of biogeography.

Climate change, human activity, and biodiversity conservation in mountains

The concentration of narrow-ranged species in the mountains entail mountains the priority of biodiversity conservation. This role is especially true in the face of the challenges brought about by global climate warming since the second half of the twentieth century, primarily as a result of increasing greenhouse gas emission primarily caused by fossil fuel consumption and land use change.

Observed species range shift responding to climate change is usually 1000 times greater in horizontal distance than in elevational distance (Chen et al. 2011), suggesting mountains as a safer environment for organisms to match the pace of climate change, especially for the taxa with weak dispersal capacity. However, as mountain species generally have narrower ranges owing to the steep environmental gradient, these species might be more vulnerable to the extinction drivers other than climate change, because their restricted and fragmented habitats are more susceptible to the risk from natural disturbance or disease. In addition, species distributed in the isolated mountains face an absolute boundary at the tops, and the upward shift driven by climate warming would eventually be stymied, leading to the risk of local extinction. Increasing records of upward shifts of mountain species imply the reality of the risk. On the other hand,
human activities at global scale have extensively altered the lowland landscape, turning mountains into sky islands separated by the land types difficult or even hostile to the migration of the organisms. By interacting with natural factors and processes at multiple scales, human activity as a pronounced component needs to be integrated into the framework of mountain biogeography, which poses a novel challenge for the development of the discipline.

SEE ALSO: Biodiversity; Biogeography: history; Climate change and biogeography; Geography of evolution; Island biogeography; Phylogeography and landscape genetics

References


Further reading


Mountain climatology

Bryan G. Mark
Ohio State University, USA

Mountain climatology is the study of weather processes and patterns resulting when a section of land is elevated with respect to the surrounding terrain. Mountains are renowned for being sites of unique and often dramatically variable weather, with characteristic winds, clouds, and colder temperatures (Figure 1). In more general terms, they are physiographic features that dynamically couple the atmosphere, biosphere, lithosphere, and hydrosphere. Moreover, mountains are resource-rich, high-energy ecosystems with cultural and biological diversity that are valuable to society (see Ecosystem services) and highly sensitive to changes in climate. Mountain climatology is a particularly apt context for integrating human and physical geography (see Interdisciplinarity and geography).

Because mountains are challenging locations in which to conduct observations, mountain climatology comprises a small and specialized branch of geography with a single comprehensive text, now in a third edition (Barry 2008). There is a long history to mountain climatology, as scientists have been drawn to explore these elevated environments and use them to gain access to observe the upper levels of the atmosphere. The insights of mountain climatology have direct bearing on many other areas of study ranging from water resources and hydrological management, forestry, ecology, and biogeography to human cultural studies.

Mountains are sensitive landscapes, and thus take on a particular interest for studying local responses to changes in global climate (see Global climate change). Going forward, certain dimensions of mountain climatology take on particular relevance in the context of human-induced climate change. These include archiving the nature of past climate variability, monitoring ongoing climate variation at high elevations, observing the rates and nature of ecological shifts in a warmer troposphere, and considering the nature of adaptation and resilience to climate change by diverse populations (see Climate adaptation/mitigation and Climate change adaptation and social transformation).

Defining mountain climates

Consideration of the climatology of mountains requires definition of mountain environments. What gets defined as a mountain is highly variable and even culturally specific, involving not only absolute elevation ranges above sea level, but also relative relief, steepness, and distinctive characteristics such as snowline, treeline (see Treeline ecotones), zone of permafrost (see Permafrost: definition and extent), and limits of glaciation. According to a typology to assess relative water yield, about 20–25% of Earth’s terrestrial surface is considered mountainous, and these regions contain a similar proportion of human population (26%), although they yield slightly more (32%) of the surface runoff (Meybeck, Green, and Vorosmarty 2001). Mountainous environments exist at different scales, ranging from isolated single volcanoes to extensive ranges of
summits and vast plateaus (e.g., Tibetan Plateau, Andean Altiplano). Their relief and altitude make mountains extreme environments, often evoking particular cultural reverence and awe. They also derive significance from containing high degrees of Biodiversity and endemism, and concentrations of natural resources. In particular, the climate and landscape characteristics of mountains make them effective “water towers” of humanity, given that enhanced Precipitation and favorable elevated landforms allow water, Snow, and ice to accumulate and flow from mountain highlands (Viviroli et al. 2007). Finally, mountains are regions of long fascination and opportunity for people, so that mountain climates have direct and cultural significance to society. Many studies have looked at the long-term significance of specific mountain climatologies, and form a genre of long-term observation and insight that mixes measurements with folklore.

Mountains are climatologically diverse settings since atmospheric density and temperature decline vertically through the troposphere. The range of climate variability is excessive, so that most general climate classification systems (e.g., Köppen) opt to simply group all highland zones as a separate category. Likewise, mountains transect many ecosystems and biotic zones vertically, and are often covered by snow and ice. Mountains also form boundaries for prevailing winds. This interrupts airflow, and forces parcels to be displaced vertically, cooling and precipitating. Hence, precipitation is often more intense on windward sides of mountains.
Conversely, the descending air warms adiabatically and dries, leading to “rainshadows,” zones of dry and often windy conditions downwind of mountains. Given this complexity of scale and processes, descriptions of mountain climate are delimited by twinned considerations of purpose and feasibility. Typical interest areas are on winds and precipitation formation, and often involve particular needs for monitoring conditions of hazards in the steep terrain of mountains, such as snow avalanches (see Snow and ice avalanches), runoff/flooding, and landslides (see Geomorphic hazards, Natural hazards and disasters, Glacier lake outburst floods, and Avalanche meteorology). Other implications in mountains for cold air pooling and boundary layer dynamics include atmospheric pollution.

For the geographer, mountain climatology holds a particular fascination and challenge given extreme heterogeneity of scale. Just as defining mountain areas is difficult given their myriad forms, their relative prominence (elevation above surrounding terrain) is responsible for an equally broad range of meteorological phenomena occurring over different scales of space and time. The fundamental regional controls of climate such as latitude and continentality (distance from oceans) that constrain radiation input, temperature, and humidity get further complicated in mountain regions by altitude and topography. These factors change the thermodynamics of the atmosphere by altering air density, pressure fields, winds, and exposure to radiation and moisture. Furthermore, atmospheric energy and momentum interact with the mountain landscape over scales ranging from the micro (e.g., individual rocks, plants, snow patches) through the local (e.g., slopes, valleys, ridges) to the regional (e.g., ranges, plateaus). All of these scales are themselves separate areas of climatic research with implications and relevance for many other realms of inquiry. How surface energy fluxes vary at the microscale impacts ecology, and leads to multiple niches along vertical expanses of mountain slopes. Diurnal heating patterns lead to thermal pressure gradients inducing valley winds, while cold air aloft sums sinks in katabatic flows. The orography of mountains can intensify precipitation along slopes, nourishing diverse tree species that form characteristic biotic ecotones with elevation.

Synoptically, mountains interrupt regional air flow, and cause gravity wave trains that induce subsequent storm generation in what is called lee-side cyclogenesis. Mountains thus play an important role in forcing or enhancing mid-latitude cyclones, and can induce orographic enhancement of precipitation. And on continental to global scales, seasonal variations in climate over mountain ranges and elevated plateaus induce large-scale monsoons. Interconnecting the dynamics of the atmosphere with landscape surfaces among all these scales challenges researchers to collaborate across disciplines and use multiple technologies to gather observations. Future advances will be made as this integration increases, and is able to be better simulated with models.

Measuring mountain climates

Progress in understanding mountain climate is hindered both theoretically and practically given the complexity of processes occurring on many scales. There is a general paucity of observations as many of the world’s mountain ranges lack long-term measurements. Logistical challenges to maintain and standardize data acquisition in different high-elevation locations make it difficult to conduct subsequent analyses or intercomparisons. Important initiatives by organizations such as the Mountain Research Initiative (MRI) are attempting to catalyze better collaborations internationally, motivated by a realization that
Mountains are key locations not only to preserve and herald the changes facing Earth, but also to test the hypothesis that temperatures at higher altitudes are changing faster than in lower places. Satellite telemetry and the use of low-cost embedded sensor networks are emerging in this context, and giving new data and insights, as are higher-resolution air- and spaceborne sensors.

The mountain range with most well-measured climate characteristics is the European Alps, where stations with reproducibly reliable instrumentation have been maintained for the longest duration; many summit stations have been in continuous operation for over a century. It was in these ranges that certain experiments were carried out as early as the seventeenth century to demonstrate fundamental principles of atmospheric pressure. Large science experiments have also been organized in the Alps, such as the meteorological research program ALPEX (Alpine Experiment) and the Global Atmospheric Research Programme (GARP) of the World Meteorological Organization. Many fewer continuous observations are available to the geographic community from the Andes, Himalayas, and more remote ranges. If observations have been obtained, data access has often been difficult, and there have been other issues of checking quality and standardization.

Specific climate variables have relevance for different questions, and researchers face different challenges in maintaining standardized measurement protocols for the highly varied conditions found in mountains. High-elevation locations are of particular interest for understanding the nature of climate changes from the past to the future. Temperature is arguably the most influential variable in mountain climate processes, and is also most straightforward to track for different research questions as it varies most persistently with terrain. Sufficiently dense arrays of measurements can allow components of regional and local physiography to be statistically decomposed to test their relative influence over time series of temperature. Precipitation, wind, radiation, and humidity are more prone to variability over complex topography and also present challenges for conducting standardized measurements to be integrated in models. Testing hypotheses about variability in these quanta on mountains is problematic because of difficulty not only in measurement technology, but also in selection of locations to place instruments; justifying any single place as “representative” is nearly impossible given the complex surface variations in mountains. Some advances have been made in embedding microsensors within the landscape, allowing for higher measurement density. Integration of satellite and aerial observations with ground measurements remains an effective way to address the landscape heterogeneity across scales, and major integrated field campaigns such as the Mesoscale Alpine Programme (MAP) in Europe and the Terrain-Induced Rotor Experiment (T-REX) in the Sierra Nevada have simultaneously deployed many ground and aircraft measurements to yield valuable new data and model simulations to understand dynamics. Mountains in all climatic zones have an impact on modifying precipitation, even if the effects vary. But recent theoretical work has postulated a linear theory of orographic precipitation that provides a simplified conceptual model of how stable atmospheric flow interacts with topography to yield precipitation. This predicts close correlation of topography with precipitation patterns, as manifest in observations from the Himalayas.

In the context of improving understanding of short-term phenomena in complex mountain environments, a host of observational techniques, data assimilation, and modeling all have been developed. Improving the understanding of multiscale meteorological processes in near real-time requires better models, data,
and analyses, motivating collaboration between academic and industry practitioners. Such has been the focus of a particular community effort: the Mountain Meteorology Committee of the American Meteorological Society sponsored a recent Mountain Weather Workshop to help “bridge the gap” between researchers and operational forecasters, which has motivated collaboration in the form of symposia (Chow, De Wekker, and Snyder 2013).

Mountains and climate change

Mountains are also highly sensitive to climate change. The need to sample climate variables directly and detect changes to the Earth system motivates a focus on mountains. This is particularly true as many observational and modeling studies have hypothesized that ongoing and future climate changes will impact high elevations preferentially. Glaciers and permafrost are found at the upper reaches; as temperature increases, these frozen bodies of water are prone to melt. Likewise the long-term and widespread loss of ice on mountains has been heralded as the “canary in the coal mine” of climate change. These changes have important implications for society. For instance, the loss of mountain Snow cover in the western United States has been a crucial issue for water supply, with associated social geographic issues related to the allocation and sustainability of resources for urban development and agriculture (see Snow cover changes).

Since topography and verticality are defining characteristic of mountains, the study of mountain climatology invokes a study of diverse ecology, variability, and gradients (see Hillslopes). Moreover, the existence of vertically defined ecologic zones with characteristic flora and fauna is constrained by climatic limits. Networks have been established (e.g., GLORIA – Global Observation Research Initiative in Alpine Environments) to record global mountain plant species assemblages using a consistent protocol, and revisited to document changes on mountains, and comparing these changes in different regions globally.

Mountain climatology has a broader heritage and legacy that comprises the study of past (paleo) climate (see Paleoclimatology) using proxies preserved in the mountain environment. Geoscientists have long been attracted to mountain environments for learning of Earth’s history; the environmental conditions preserved in the geologic record provide a means to infer past climatic conditions that appear to change radically over time. After the understanding of glacial ice ages (see Glaciations and Quaternary glaciations) was established in the early nineteenth century by reexamining geological evidence, many speculative ideas, often wildly divergent, emerged to explain the changing character of Earth’s climate over time. T.C. Chamberlain gave an address in Toronto to the British Association for the Advancement of Science in which he attempted to synthesize multiple working hypotheses on climate changes. He captures the trepidation of this early endeavor, which lacked data and theory, but makes an emphatic case for combining geology and climate studies that has subsequently borne fruit in the field of paleoclimatology: “While the atmosphere is the most active of all geological agencies, it has received the least careful study from geologists....The atmospheric element in geological history bids fair to long remain obscured by elusive factors and uncertain interpretations. None the less it is an element of supreme importance and should be persistently attacked until it yields up its truths. ... All our attempts at the solution of climatic problems proceed on some conscious or unconscious assumption concerning the extent and nature of the atmosphere at the stage involved” (Chamberlin 1897).
While the climate change theories reviewed were speculative at the time, research in many mountain ranges over the ensuing century has revealed the prescience of Chamberlin’s charge. Many important archives of past climates, from ice cores to lake sediments to ancient tree rings, and the paleoclimate records extracted, have given important context for understanding mountain climate variability over time. Assimilations of paleoclimate proxy databases have become important validation for global climate models. Furthermore, given the strong local topographic mediation of regional climate, more attention has been placed on achieving broader, more spatially distributed sample sets of multiple proxies. For example, using snowlines reconstructed from glacier geomorphologic data (i.e., moraines) to constrain regional to global climate change requires careful scrutiny of intraregional variability (Mark et al. 2005). Also, intercomparison of ice cores recovered from multiple mountain locations outside of the polar ice caps has broadened insights into interhemispheric responses to global climate dynamics.

While mountains are often considered static in relation to the dynamics of the atmosphere, growing interest in Earth systems research (see Earth system science) has shown that there is actually a full coupling that goes in the opposite direction. Geologic processes initiated in the atmosphere and characterized by climate actually impact the lithosphere and ultimately govern weathering and uplift of the mountains. The realm of landscape evolution is a growing focus area that brings together tectonics, geomorphology, and climatology (see Tectonic geomorphology and Quantitative geomorphology). Climate links surface uplift and erosion rates through several mechanisms, including orographic forcing of precipitation, onset of alpine glaciation, and other moderations of erosion efficiency (Willett et al. 2006). In this context, climate models, and even weather forecasting models, are being coupled with landscape evolution models to investigate the influence of mountain climate on erosional processes, landscape evolution, and the entire orogenic dynamics. Furthermore, weathering processes (see Weathering processes and landforms) in mountains and high-standing islands have implications that feed back to global climate by altering biogeochemical cycles such as that of carbon (Lyons et al. 2002).

**Research outlook**

With a summary look to the future, key questions continue to drive research in mountain climatology (see Figure 2): How do mountains impact the atmosphere in their vicinity? How
can mountain environmental changes record and reveal the nature of global climate changes? How do mountain societies adapt to climate change? In a general sense, mountains are obstructions to atmospheric flow, and thus play an important role in impacting synoptic weather regionally (even globally) given dynamic and thermodynamic processes they instigate. Mountains also present themselves as sampling platforms to sample the upper atmosphere over time, not only to test hypotheses about ongoing future warming, but also to better document past climate variability. And more recently, mountain climates have become relevant to understanding the interplay of all Earth systems, including global biogeochemistry and lithospheric crustal tectonic dynamics. Finally, understanding the nature and impact of human-induced climate change has motivated teams incorporating climatologists with other biophysical and social scientists to research mountain environments.

SEE ALSO: Atmospheric/general circulation; Climate change and biogeography; Climatology; Glacier hydrology and runoff; Landforms and physiography; Mountain biogeography; Mountain geomorphology; Mountain hydrology; Soils of mountainous landscapes

References


Further reading

Mountain geomorphology

Olav Slaymaker
University of British Columbia, Canada

Mountains occupy nearly one quarter of the land surface of the Earth and support 10% of its population directly. A further 40% of the Earth’s human population is dependent on mountain environments in terms of resource extraction and exposure to natural hazards originating in the mountains. Because most of the world’s major rivers rise in mountain regions, changes in mountain runoff affect a large proportion of the global population. Mountains are also sensitive indicators of climate change through the shrinking of glaciers and the upward migration or extinction of plants and animals. Mountains are core areas of global biodiversity and the origin of many of the world’s food crops. In many regions this global heritage is significantly endangered by changes in land use driven by a range of economic and political pressures. Because both natural and socioeconomic change is inevitable, fundamental understanding of mountain systems deserves high priority in sustainability strategies worldwide.

Distinctive features of mountain geomorphology

Mountain geomorphology is a “regional component within geomorphology” (Barsch and Caine 1984). The region in question is the world’s mountains defined by two conditions: (i) high absolute elevation above sea level and (ii) high gradient or local elevation range. Troll (1973) who was the creator of modern mountain geomorphology made a distinction between high mountains (“hochgebirge”), which rise above local timberline, and mountains (“gebirge”) which encompass more than one vegetation belt but remain below timberline. Troll’s distinction was important for at least two reasons: (i) his focus on biomes allowed global comparisons that were more difficult to make on the basis of absolute elevation above sea level and (ii) biomes integrate many of the effects of hydroclimate (Körner and Ohsawa 2005). The use of elevation versus biome zones in mountain geomorphology continues to be actively debated because there is no international standard definition.

Fairbridge (1968) provided a comprehensive classical description of mountains in terms of scale and continuity: (i) a mountain is a singular, isolated feature; (ii) a mountain range is a linear topographic feature of high relief usually in the form of a ridge; (iii) a mountain chain is also a linear topographic feature of high relief but usually reserved for features that persist for thousands of kilometers; (iv) a mountain mass, massif, block or group is reserved for irregular groupings of ranges with no simple linear trend; and (v) a mountain system or cordillera describes the largest continent spanning mountain features.

Fairbridge also provided a standard genetic classification of mountain types. Two broad categories of genesis were considered: (i) structural, tectonic, or constructional mountains and (ii) denudational or destructive mountains. Although these categories are useful for local scale description it is important to recognize that they were developed before the science of plate tectonics was fully understood.
Hewitt (1972) and Barsch and Caine (1984) provided creative discussions of the field of mountain geomorphology. Hewitt (1972) suggested that the high energy condition and the denudation history reflected in mountain forms express the distinctiveness of mountain geomorphology. Barsch and Caine (1984) built on Hewitt’s discussion by subdividing the high energy condition into (i) process dynamics and (ii) morphoclimatic models and subdividing the denudation history into (i) morphometry and structure and (ii) relief generation and history.

For the purpose of the following discussion, four broad categories of mountain geomorphology are proposed: historical, functional, regional, and applied.

### Historical mountain geomorphology

Historical mountain geomorphology concerns the evolution of mountains and mountain systems over both long and intermediate time scales. Since the plate tectonics revolution of the 1960s a strong relation between plate tectonics and mountains has been recognized. Most, but by no means all, large mountain systems or cordilleras are associated with plate boundaries, whether convergent, divergent, or transform. Large horizontal and vertical displacements are associated with plate interactions. Nearly all the literature on mountain building over the last 50 years has concentrated on active plate margins where collision and subduction can explain both mountains and the structures within them. There are four convergent plate settings in which some of the most rapidly evolving mountain systems of the world are located. They are (i) oceanic to oceanic plate convergence (e.g., Japanese Alps and the Aleutian Arc); (ii) oceanic to continental plate convergence (e.g., South Island, New Zealand, and Coast Ranges of the Pacific Northwest); (iii) continental to continental plate convergence (e.g., Himalayas); and (iv) displaced terranes along accreted margins (e.g., British Columbia). Divergent plate settings include (i) sites of ocean spreading (e.g., Iceland and the Galapagos Islands) and (ii) intra-continental rifts (e.g., the Gulf of Aqaba and the Scottish Highlands). Transform plate settings are: (i) ridge past ridge (Coast Ranges, California); (ii) trench past trench (Anatolia, Turkey); and (iii) ridge past trench (Pakistan-Afghanistan).

Mountains are also found in plate interior settings, such as (i) hot spots (Hawaii and Yellowstone National Park); (ii) continental flood basalts (Deccan, India, and Columbia Plateau, Pacific Northwest); (iii) shields (Hoggar Mountains, Algeria); (iv) intra-cratonic uplift sites (the San Rafael Swell, Utah); (v) post-tectonic magmatic intrusion sites (Air Mountains, Nigeria); and (vi) evaporite diapirs (Zagros Mountains, Iran).

Ollier and Pain (2000) disagree with the prevailing plate tectonic viewpoint by emphasizing that there are also mountains on passive plate margins and noting that mountains are made by uplift of continental areas. If the uplifted area is deeply dissected it will be a typical mountain chain with isolated peaks. Some mountains are at the edges of the plates delineated in plate tectonic theory, but many are not. The disagreement also arises in part from the different methodologies adopted by those who emphasize the close relation between mountains and plate tectonics, who are heavily invested in fission track analysis (e.g., Summerfield 2000), and the Ollier school, which emphasizes landforms, stratigraphic, and physical geologic data.

Relief development in mountains revolves around the interpretation of accordant summit levels and regionally consistent valley benches. Summit accordances and planation surfaces are present in most mountain systems. The European
Alps were early identified as having a “gipfelflur” (summit accordance) which was explained as a remnant of a pre-existing Pliocene erosion surface. Alpine crests and summit accordances in general have been explained as the product of equally spaced gully/river dissection which would lead to an accordance of summit heights between them. Uniform lithology over a broad area would certainly favor such a process, but this is not found, for example, in the European Alps. Timberline and alp slope accordances are explained as having been formed during warmer, interglacial periods under higher timberline. Rock benches along the sides of major valleys are variously explained as Tertiary, Pleistocene glacial, and interglacial effects.

Functional mountain geomorphology

Functional mountain geomorphology concerns itself with the various ways in which mountain geomorphic systems respond to disturbances, both natural and human. A valuable conceptual framework is provided by the sediment budget. The rate and nature of movement of sediments from sources to sinks and the associated rates of denudation of the mountain landscape can be estimated by determining all components of a sediment budget. It is common practice to examine the following four subsystems of the mountain environment and to aggregate the subsystem responses into a total basin response:

1 The mountain cryosphere includes snow, seasonally frozen ground, mountain permafrost, lake and river ice, and glaciers. Each of the elements of the cryosphere has a different response time to a disturbance. Snow responds on a daily time scale, lake and river ice on an annual time scale, permafrost and glaciers respond on annual to century time scales. Associated ecosystem responses are measured in decades to centuries and sediment systems may take decades to millennia to respond. With respect to sediment fluxes, in glacierized mountains, with their own active glaciers, the highest fluxes are found in the proglacial environment; in formerly glaciated mountains, where glaciers are no longer present, the highest sediment fluxes tend to occur at intermediate elevations due to patterns of sediment storage and availability; and in never glaciated mountains rates of denudation decline from high to low elevation.

2 The coarse grained debris system involves the transfer of debris between cliffs, through chutes to talus and associated colluvial deposits. Rockfalls, landslides, avalanches, and debris flows are the processes that drive the coarse grained debris system. This system is strongly influenced by the cryosphere and either supplies coarse sediment inputs into the fine grained sediment system or may be a closed system that remains uncoupled with its neighboring systems over long periods of time (Barsch and Caine 1984). Where the hydroclimate provides sufficient runoff, gravel-bed rivers are formed and they are commonly coupled with adjacent systems. Where the coarse grained debris system is uncoupled there may nevertheless be high local rates of denudation, as demonstrated in even the earliest sediment budget studies from the Swiss Alps.

3 The fine grained sediment system is well coupled with adjacent systems at progressively lower elevations of the sediment cascade. The upper size limit of fine grained sediments is commonly placed at 1 mm diameter. Weathering, surface erosion, and fluvial processes, which move sediment within basins, and aeolian processes,
importing sediments from external sources, drive the fine grained sediment system. Fluvial processes transport fine sediment in suspension, in solution, and in bed load into adjacent systems with variable lag times.

The geochemical system. The drivers of the geochemical system are solutional weathering, nivation, hydrogeological, and fluvial processes. The contact time between water and material sources and thus the residence time of water within the basin, temperature changes, vegetation cover, and the thickness of unconsolidated sediments dictate the importance of the geochemical system. As the sediment cascade enters more vegetation-dominated zones and soils are also more maturely developed at lower elevations, the breakdown of minerals into solutes and nutrients becomes more rapid and complex. Human disturbance of the land ensures the importance of pollutants in the geochemical system.

Regional mountain geomorphology

Relief type

Based on a consideration of form alone, Barsch and Caine (1984) suggested that there are four distinctive relief types in high mountain systems: (i) the Alp type, associated with an overriding influence of glaciation and glacial erosion; (ii) the Rocky Mountain type, with a less pervasive effect of glacial erosion, including low relief on flat summits and rounded interfluves; (iii) a polar type, which gives evidence of intense glacial erosion but frequently with local relief of less than 1000 m (e.g., Lofoten Islands, Norway); and (iv) a desert type, which is a high mountain according to elevational criteria but does not reach timberline and was only lightly glaciated during the Pleistocene (e.g., Tibesti Mountains, Chad).

Hydroclimate

Biomes integrate the effects of hydroclimate (Körner and Ohsawa 2005) and in many mountain regions, where hydrometric and meteorological data are scarce, biomes are the only reliable basis for global comparisons. Typically, ten mountain biomes are recognized as follows: polar nival, dry alpine, humid temperate alpine, humid tropical alpine, dry temperate montane, humid temperate montane, humid tropical montane, humid temperate hilly, dry tropical hilly, and humid tropical hilly. Of these, the humid temperate and humid tropical montane biomes occupy the greatest area (7 and 6 million km² respectively) followed closely by the polar nival biome (5.5 million km²). Although it is widely recognized that alpine biomes experience some of the most active geomorphic processes, they occupy collectively only 2.5 million km² (<8% of the terrestrial surface) and many of them are decoupled from the montane biomes at lower elevations (as previously noted by Barsch and Caine 1984).

Human activity

Slaymaker and Embleton-Hamann (2009) concluded that human influences on mountain systems have now become so pervasive that global comparisons of mountain geomorphology must incorporate a measure of human impact. They argued that human activity has two broad impacts on mountains: (i) inducing stability by sensitive environmental engineering and (ii) reduction of stability by serious soil and vegetation disturbance. Four distinct mountain geomorphologies are defined by them as follows: (i) polar mountains, such as Svalbard, which are relatively undisturbed by human activity, and which respond primarily to warming of the cryosphere; (ii) low population temperate mountains (such as Canada and
Tadjikistan) which are controlled by steep relief and harsh hydroclimate; (iii) high population temperate mountains (such as Austria and Japan) are strongly influenced by human impact (both positive and negative); and (iv) tropical mountains (such as Papua New Guinea and Ethiopia) are also heavily destabilized by human activity. Huddleston and Ataman (2003) noted that the humid tropical montane and humid temperate hilly biomes are by far the most populous, with >60% of all mountain peoples.

Applied mountain geomorphology

A geomorphic hazard results from any landform or landscape change that adversely affects the stability of a site or drainage basin and that intersects the human use system with adverse socioeconomic impacts (Slaymaker 2010). If there are no people affected there is no hazard and if the landform or landscape is unchanged there is no geomorphological hazard. Mountain geosystems are not exceptionally fragile but they do show a greater range of vulnerability to disturbance than many landscapes and their recovery rate after disturbance is often slow. The combination of extreme geophysical events, apparently increasing in magnitude and frequency during the present climate change scenario, with exceptional population growth and varied land use modifications underlines the urgency of greater understanding of applied geomorphology. Geomorphic hazards are intensifying and societal risks multiply accordingly.

During the past four decades, the world’s population has doubled but the population of mountain regions has more than tripled so that stresses on the physical and biological systems in mountains have intensified manifold. Human activity, in the form of population density and land use, is a direct driver of geomorphic processes in mountains. But geomorphic processes also influence population density and land use. This is what is known as a reflexive relation. It is not, however, the sheer numbers of people but aspects of population composition and distribution, especially the level of urbanization and household size which exercise the greatest demands on the land. High population densities in developing world conurbations, for example, may lead to better management, such as has been documented in Kenya (Tiffen, Mortimore, and Gichuki 1994).

Perhaps the deliberate change in vegetation cover by society is the most important driver of change. Nearly one third of Earth’s land surface is currently being used for growing crops or grazing cattle. Grazing and forestry are the dominant mountain land uses. Huddleston and Ataman (2003) noted that all mountain zones below 4500 masl have exceeded the population that is thought to be sustainable. Deforestation in particular provides timber for fuel and as construction material and, in order to clear ground for agriculture, sets in motion a series of geomorphic processes. Surface erosion of slopes, landslides, and debris flows lead inexorably to environmental degradation. Forest roads, highways, railways, electricity power, and communications networks alter water and sediment pathways. Excessive soil erosion, soil degradation, increase in the frequency of mass movement events and increased flood hazard are the results. But it should also be pointed out that many studies have shown how good management of land can also improve land stability. In either case (whether natural hazards have increased or decreased), the role of society is paramount.

Road and dam construction are also two of the most profound human influences on mountain geomorphic processes. Road building in Nepal has produced about 9000 m$^3$ km$^{-1}$ a$^{-1}$ of...
landslide material and it is estimated that every kilometer of road constructed will trigger circa 1000 tons of land loss from road failures. Many of the world’s large dams are located in mountain regions and there are massive changes in both discharge and sediment transport imposed by large dams.

There is also a need to understand, predict, mitigate, and adapt to the effects of mountain geomorphic processes. Unfortunately there has been inadequate representation of the interactive coupling between relief, land use, and hydroclimate. Based on magnitude, timing, persistence, reversibility, potential for adaptation, distributional aspects of impacts, and likelihood and importance of the impacts, some of these vulnerabilities have been identified as “key.” It is at the point of exceedance of thresholds, where non-linear processes cause a system to shift from one state to another that expresses the key vulnerability. The changing magnitude and/or frequency of geomorphic hazards are the geomorphologist’s favored measures of key vulnerabilities. Slaymaker (2010) has discussed mountain geomorphic hazards as a function of resilience, connectivity, and vulnerability and showed the importance of spatial scale.

SEE ALSO: Anthropogeomorphology; Applied geomorphology; Geomorphic hazards; Land degradation; Mountain biogeography; Mountain climatology; Mountain hydrology; Soil erosion and conservation; Tectonic geomorphology; Volcanic processes and landforms

References


Mountains cover approximately 24% of Earth’s land surface, but generate a much higher percentage of global runoff. Estimates of the proportion of mountain runoff to global total runoff vary from 32% (Maybeck, Green, and Vorosmarty 2001) to 40–60% (Bandyopadhyay et al. 1997). The term “water towers” has been used to describe the hydrologic role of mountains as wet islands within dry climate zones (Viviroli et al. 2007). In Wyoming, for example, only 15% of the state area experiences a water surplus – all of which occurs in mountains – and runoff from the mountainous highlands supplies water to the remaining 85% of the state that is situated in water-deficit lowlands (Figure 1) (Ostresh, Marston, and Hudson 1990). As another example, the European Alps cover only 23% of the Rhine River basin, but provide half of the total runoff over the year, ranging from 30% in winter to 70% in summer.

Wohl (2010) pointed out that mountains are situated in a wide range of climate regions, with profound differences in the flow regime as a result. Mountains in warm-arid climates (e.g., Basin and Range Province of the United States, Middle East, western South America, Africa, Australia) experience ephemeral flows with occasional flash floods. Cold-arid mountains (e.g., parts of Antarctica) experience ephemeral flows during the melt season. Semiarid mountains (e.g., Colorado Front Range of the United States, Pyrenees of Europe) can experience ephemeral flow from rainfall or perennial flows where snowmelt augments flows. Mountains in warm-humid climates (e.g., Appalachian Mountains of eastern United States, mountains of Japan) can generate perennial flows, occasionally with very high discharges caused by tropical storms. Mountains in cold-humid climates (e.g., western side of Canadian Rocky Mountains, European Alps, Chilean Andes) also support perennial flows, dominated by either rainfall or snowmelt. Mountains of the seasonal tropics (e.g., Northern Great Dividing Range of Australia, Caribbean region) experience strong seasonal shifts in discharge, in some cases highlighted by monsoon season flooding. Mountains in the humid tropics (e.g., New Guinea, headwaters of the Amazon in South America) tend to be perennial, with lower seasonal variation than in other types of mountain rivers. Finally, high-elevation mountain massifs are dominated by snowmelt and glacier-melt, although the Himalayas are also dominated by monsoon rains. Flow regimes in mountains could shift with climate change.
Figure 1  Water balance (net precipitation minus evapotranspiration) for the State of Wyoming, USA. Ostresh, Marston, and Hudson 1990, 25. Reproduced from Wyoming Water Development Commission.

Net precipitation (that which reaches the ground surface) is thus calculated as gross precipitation (measured above/outside of vegetation) minus the interception loss (absorption + evaporation). Interception storage is greater for low-intensity rainstorms with no wind. Thus, interception storage will be greater for trees in the forest than for isolated trees or trees along the forest margin. Interception storage is greater than snowfall, when measured in water equivalent, because rain generally adheres to vegetation surfaces better than snow. Total interception for deciduous forests may average 13%, contrasted with 22–28% for coniferous forests, and 10–20% for grasses. For crops, interception may be as low as 3% for corn and oats, increasing to 9% for soybeans, and 22% for alfalfa. The percentage interception loss varies inversely with storm size, and is greater for conifers than for deciduous trees. The interception loss is highest in regions where evaporation is highest.

Vegetation affects mountain hydrology in several key ways (Marston 2010).

1  It modifies soil moisture, through interception loss and transpiration. 
2 Organic matter in the soil increases water storage, infiltration, and percolation, thereby promoting vegetation growth and inhibiting erosion.

3 Roots bind the soil against piping, land surface erosion, and shallow mass movement.

4 Aboveground biomass creates microtopography on the land surface that affects overland flow. Roughness in the profile direction (upslope-downslope) slows overland flow; roughness along the contour concentrates sheetflow into rillflow, rills into gullies.

5 Aboveground biomass creates hydraulic roughness against overland flow.

In theory, one should be able to increase water yield by eliminating or decreasing losses by interception and evapotranspiration. Therefore, the conversion of coniferous forest to a broadleaf forest should increase water yield. Conversion from shrubs or trees to a grass cover (as undertaken with chaining of pinyon-juniper woodlands) should result in little or no change. Because the percentage interception loss decreases as storm size increases, it is important to note that decreasing interception loss is unlikely to have a significant effect on floods produced by major rainstorms in mountains.
When rainfall or snowmelt reaches the ground surface, some or all of it will enter the soil in response to gravity. Infiltration rates are commonly very high in forested mountain soils because of the coarse soil texture and high organic matter content. In contrast, soils are thin or absent in desert mountains, which keeps infiltration rates low. Water in the soil is held against the pull of gravity by capillary forces, which are especially strong in fine textured soils. When soil pores are filled to capacity, the soil is saturated. Further additions of water will cause the water to exert pore pressure and force soil grains apart from each other, destroying cohesion and creating deleterious effects on slope stability. The maximum amount of soil moisture after gravitational water has drained is known as the field capacity. Vertical percolation will carry water through the soil water zone, which extends through the zone of plant roots, to the intermediate (unsaturated) zone and eventually to groundwater (the zone of saturation return flow). Subsurface water also moves laterally by throughflow in the soil water zone, interflow in the unsaturated zone below plant roots, and by groundwater flow. Some subsurface water will reach streams as effluent seepage, contributing to the base flow of the mountain stream. Mountain streams also lose water to subsurface zones via influent seepage. Finally, water moves laterally beneath streams as intergravel flow. Available soil moisture is strongly controlled by topographic position in mountains (Figure 3).

Water is evaporated to the atmosphere from mountain lakes, streams, soils, and from intercepted water. Rates of evaporation are controlled by energy transfer, which increases with elevation in mountains. Vegetation pulls water from the soils and transpires it to the atmosphere through the pores in leaves. Transpiration rates also depend on energy transfer processes, but are controlled by many factors other than evaporation, including the phenology (stage of plant growth) and amount and type of pores.

Figure 3  A relative wetness index generated using GIS highlights the role of ridges, swales, and channels in controlling available moisture. Marcus, Aspinall, and Marston 2004, 359. © With permission of Springer.

Table 1  Human impacts on mountain hydrology.

<table>
<thead>
<tr>
<th>Indirect</th>
<th>Direct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Timber harvest and road building</td>
<td>Beaver trapping</td>
</tr>
<tr>
<td>Grazing and cropland agriculture</td>
<td>Dams and flow regulation</td>
</tr>
<tr>
<td>Urbanization</td>
<td>Off-channel diversions</td>
</tr>
<tr>
<td>Lode mining</td>
<td></td>
</tr>
<tr>
<td>Climate change and altered fire regime</td>
<td></td>
</tr>
</tbody>
</table>

Source: Modified from Wohl (2010, Table 6.1).
Runoff in mountains is generated by Hortonian overland flow, saturation return flow, permafrost melt, and effluent seepage. During Hortonian overland flow, rainfall or snowmelt exceeds the infiltration capacity of the ground surface for a period of time. It is not necessary for the soils to be saturated, but rather that the rate of water reaching the soils’ surface exceed the infiltration capacity. Next, depression storage must be exhausted – the surface water storage created by microtopography. Depression storage may be close to zero atop a paved parking lot, but can reach 25 mm in a tilled soil or even more in forested mountain soils. Finally, in the presence of a slope, Hortonian overland flow begins – this is very rare on forested mountain slopes. Saturation return flow occurs where subsurface flow (in the soil moisture zone, as vadose water, or groundwater) is conducted to the surface. Permafrost can be important in high mountain areas, especially in supplying local springs.

Human effects on mountain hydrology

Human activities can affect mountain hydrologic processes indirectly or directly (Table 1).

Indirect effects of humans on mountain hydrology

Timber harvest usually results in compaction along road and skid trails, which reduces infiltration and increases overland flow. Studies on the effects of clear-cut logging on mountain hydrology in rain-dominated flow regimes along the Pacific Rim generally agree that peak flows increase and low flows decrease. In snowmelt-dominated regimes, narrow clear-cuts enhance snow trapping and help shade snow so the melt is delayed. In such cases, low flows can be augmented without increasing peak flows. However, even in snowmelt-dominated flow regimes, clear-cut logging over a threshold value – 20% of the watershed in the Rocky Mountains in any one year – can increase peak flows to the point that irreversible stream channel changes occur. Studies in the Himalaya have shown that forest cover does not affect monsoon-period flooding, as the effects of land cover on local runoff are less significant compared to the combined effects of steep slopes, thin soils, and long-duration rain events (Marston 2008). Effects of vegetation change on mountain hydrology are most pronounced in mountain headwater regions, but gradually disappear further downstream from the area of impact.

Roads must be constructed as part of the timber harvest plans, with cut-and-fill required on mid-slope roads. Four primary hydrological effects of roads include the interception of rainfall, concentration of flows, rerouting of water pathways unnaturally onto hillslopes, and interception of local water tables at hillslope cuts. Forest roads increase the magnitude and frequency of peak flows by transporting runoff directly into streams at crossings, thereby reducing travel time and distance. Roads are nearly impervious compared to soil, which makes them efficient transporters of flow by limiting infiltration. Effects of roads on local hydrology will be more significant downstream from crossings or regions with high road densities. The hydrologic impact of mid-slope roads is to intercept subsurface flow while also capturing direct precipitation. The excess runoff must be directed via berms, roadside ditches, and culverts to mountain channels that may or may not have the capacity to conduct the increased flow.

Ranching and farming on mountain slopes can alter the hydrology. Heavy grazing on upland soils can compact the soil, reduce infiltration, reduce vegetation cover, and increase
Hortonian overland flow (Wohl 2010). One of the remarkable ways in which humans affect mountain hydrology is through agricultural terracing (Figure 4). Terraces are built to retain soils and water to support crop growth, while simultaneously conducting excess rainfall downslope. In ancient mountain civilizations, terraces have been in place for centuries if not millennia. For khet terraces constructed with a berm (or bund) on the outer edge to retain water, surplus runoff moves downslope from one ramped terrace to the next, and this occurs over hundreds and sometimes thousands of meters of local relief without triggering irreparable erosion. At higher elevations, rain-fed bari terraces built without a berm and sloping outward experience rates of erosion two orders of magnitude higher than khet terraces (Marston 2008).

The conversion of natural vegetation to impermeable urban land cover affects routing of water in mountains. Infiltration is reduced and Hortonian overland flow is increased, with enhanced local sheet flooding as a result. Hard rock (lode) mining can increase streamflow in mountain areas if discharge from active or abandoned mines is not controlled. The impacts of lode mining on water...
quality and sediment load are of greater concern than are the impacts on hydrology.

It is difficult to translate predicted climate change into mountain hydrology impacts, and those changes will vary between mountains located in contrasting eco-regions. Generally speaking, with warming temperatures, the elevation at which snow changes to rain is expected to increase. Glaciers with a lower elevation accumulation zone can be expected to shrink in extent and decrease in thickness. Glaciers with a higher elevation accumulation area can be expected to experience a net gain in mass. The timing of precipitation (e.g., monsoon rains) may also change. Permafrost on lower elevation mountain slopes can be expected to decrease in extent and melt at shallow soil depths.

Removal of vegetation by fire decreases interception loss and soil moisture by affecting infiltration rates and decreasing evapotranspiration. Intense fires, whether deliberately set after clear-cut logging or as wildfires, can create hygrophobic (water repellant) soils, which can enhance runoff. Intense drought is also known to create hygrophobic soils. Whether or not this results in significant alterations to streamflow depends on the path of surface runoff on the local hillslope scale, the proximity of burned hillslopes to water courses, and the cumulative area burned. Forest litter, downed trees, micro-depressions, and unburned segments can help retain runoff and encourage infiltration. Flash floods and associated mudflows are common in the steep San Gabriel Mountains in southern California; however, accelerated runoff was not a widespread problem on the broad Yellowstone Plateau after the extensive 1988 fires. Moreover, the effects of fires on accelerated runoff will diminish after a few years once vegetation recovers. One of the ways that climate change will manifest itself will be to alter fire regimes in mountains of the world, with accordant effects on hydrology.

Direct effects of humans on mountain hydrology

Beaver populations once flourished along streams of North America and Europe, but are presently experiencing reduced numbers (Wohl 2010). Beaver dams are constructed of wood and sediment to a height commonly between 1 and 2 m. The hydrologic effects are to promote more uniform streamflow and raise the water table, which often enhances riparian vegetation. Beaver populations were greatly reduced over past centuries by fur trappers, resulting in channel incision and an associated drop in adjacent water tables. Attempts have been made to reintroduce beaver for stream and riparian restoration, with mixed success. The beaver populations need to have access to sufficient wood to maintain their dams. If the dams breach during peak flows, channel incision occurs downstream.

Mountains are often preferred locations for dam construction because significant relief exists to generate a hydraulic head in combination with constricted valley sides comprised of competent bedrock to anchor the structure. The countries of the Alps have developed 76% of their hydropower potential, but Nepal has developed less than 1% of its potential (Romerio 2008). Dams alter mountain hydrology in the process of providing hydropower, flood protection, and reservoir recreation. Water is lost through evaporation and seepage. Storage capacity is lost with time since initial construction due to sedimentation. Downstream from a dam, peak discharges are reduced, the range of daily discharges is decreased, and the number of reversals in discharge is increased. Further alterations occur to the timing of low flows and peak flows, and to ramping rates. Additional effects to water
quality and riparian and aquatic ecology are widely reported.

Mountain streams are targeted for off-channel diversion of water, sometimes for local uses (e.g., mining, irrigation) and sometimes for transfer to large urban centers. A diversion structure begins with a headgate that allows water to be transferred from a stream channel into a ditch. Barriers are usually placed across stream channels directly downstream from the diversion structures to pool water and reduce stream power, making it easier for flow to enter into the ditch through the headgate structure. In some cases, barriers are built completely spanning the width of the channel to divert 100% of streamflow into the ditch. Water conveyance ditches then follow hillslope contours, transporting water from one watershed to another by transcending drainage divides. Water can be moved across multiple drainages, eventually flowing onto a private farm or ranchland for use as irrigation. Normally, some reduced amount of flow is able to continue down the natural channel past the diversion structure. Effects of diversion ditches on runoff patterns are fundamentally different compared to those associated with dams. A dam will reduce peak flow for flood control and increase low flow. Peak flows commonly pass by diversion ditches uncontrolled. Hydrological impacts of diversion ditches on streams downstream from the diversion include altered flow regimes, depleted stream flows, reduced channel capacity, and lowered local water tables. Total annual water yield and average annual discharge are shown to decrease downstream from diversion structures. Similarly, flow variability increases during low-flow years downstream from a diversion.

SEE ALSO: Dams; Glacier changes; Glacier hydrology and runoff; Glacier lake outburst floods; Glaciers; Hydrologic cycle; Mountain biogeography; Mountain climatology; Mountain geomorphology; Permafrost: definition and extent; Snow; Snow cover; Snow cover changes; Soil water; Soils of mountainous landscapes

References


Further reading


Movies and films help us to understand our place in the world and it is the extent to which they produce spatial meanings that interest geographers. Early studies focused on place-based meanings with an emphasis on how accurately places were depicted. The interrelations between movies and films on the one hand, and the politics of social and cultural representations on the other, became increasingly important to the discipline from the early 1990s onwards. At the time, a number of central works were emphasizing the importance of visual culture (Harvey 1989; Bruno 1993) and within the decade a critical visual geographic engagement with films and movies was codified methodologically (Rose 2001) and substantively (Aitken and Zonn 1994). Perhaps most critically, the last two decades has witnessed a flowering of spatial theories that elaborate new ways of knowing our cinematic world.

From a less critical stance, the discipline’s engagement with film may be traced back to when the Geographical magazine published a series of articles in the 1950s in collaboration with Dr Robert Manvell, then director of the British Film Academy. The series emphasized the national character and factual basis of filmmaking, focusing on the realism of the medium as its most significant contribution to geography. Most of the papers focused on documentary filmmaking, although some broached the issue of vérité or realism in narrative and fictional films (Manvell 1953). These articles outlined the ways film dealt effectively with the culture, customs, and behaviors of the everyday lives of people in the country portrayed. The pervasiveness of French cinéma-vérité and Russian Kino-Pravda had held sway in continental Europe since the Lumière brothers famously projected their first images in the late nineteenth century. British filmmaker John Grierson coined the term “documentary” in the late 1920s from the French adjective documentaire, used at the time to describe the accuracy of travel films. In the 1940s, a much older and perhaps less wiser (but perhaps presaging the influence of Hollywood and Bollywood) Louis Lumière took the idea of accuracy further by suggesting that the space of film is the space of reality, and film’s ambition is to “reproduce life.”

Aitken and Zonn (1994) coedited the first book on film geographies that attempted to bring together the work of film studies theorists like Christian Metz, Stephen Heath, Jean-Louis Baudry, and Thierry Kunstel with emerging critical theories of space and place. At around the same time, Wolfgang Natter and J.P. Jones (1993) published an acerbic critique of the “reality effect” of documentary filmmaking that highlighted performed narrative conventions as the outcome of social, cultural, and political mediations. Behind this work, and pushing a media savvy troop of young academics to aspire to new giddy heights, were the spatial and social theories of David Harvey, Doreen Massey, Michael Dear, and Ed Soja, who in turn drew on Walter Benjamin, Henri Lefebvre, Guy Debord, and other Marxist-leaning theorists. Caught in the rational
objectivity of the quantitative revolution where the most coherent spatial theories – like Waldo Tobler’s pronouncement that things close to other things have the most influence on those things – seemed banal and self-evident, these young academics were inspired by a set of social and spatial theories about the complex ways media and culture link insidiously with a capitalist-industrial-military complex.

Inspiration came from writers such as Frederic Jameson and Jean Baudrillard who spoke to how seeming interior psychic conditions and larger social structures intertwined with cultural products, and in particular the spatial arts of architecture, urbanism, and film. In addition, feminist writers such as Anne Friedberg and Guilliana Bruno (1993) countered through film and the spaces of movie theaters male dominated urban and architectural strictures, offering shrewd insights into what makes cities magical, fun, and empowering for women, despite their vulgarity, rabble, vice, and dangerously empty corporate plazas. Bruno (1993, 51) argues that cinema is a transgressive space because it allows women to experience the “erotics of darkness and (urban) wandering denied to the female subject.” Bruno and Friedberg in tandem relate society’s panoptic and objectifying gaze to movie-oriented experiences, arguing with Laura Mulvey that the audience looks with the male character and the female character is “looked at.” In her account of cinema’s role in culture, Friedberg positions the virtual gaze as a mode of looking that is based upon mobility rather than confinement and, by so doing, she opens an arena of speculation over the relations between visual spectacles and fluid subjectivities. As Mike Crang (1997) notes, the attenuation of other senses within the darkened interiors of theaters is an especial configuration and practice of viewing, which sets up the possibility of the illusory eye following the camera.

Geographers mobilized these ideas in new critical spatial theories. In his acclaimed The Condition of Postmodernity, Harvey (1989) uses Blade Runner (1982, directed by Ridley Scott), among other films, to argue for the ways movies elaborate what was referred to at the time as our postmodern condition and, in particular, the conflicts that contemporary capitalism wrought on time scales and spatial resolutions. Unwilling to cede too much power to film, Harvey sees Blade Runner as a parable in which postmodern conflicts are set in a context of flexible accumulation and time-space compression. Harvey is not willing to move beyond using film as a mimetic device that reflects changing societal structures, whereas other geographers wanted to understand also the ways that film changed society. Ed Soja and Michael Dear, for example, are sacrosanct about the spatial powers of film. Soja is particularly influenced by Fredric Jameson, who in The Geopolitical Aesthetic (1992) not only uses films as parables for conditions of postmodernity but argues convincingly about larger global/capitalist conspiracies that are mediated by the subtle and enduring influences of images and moving pictures. Jameson’s concern is not only to unpack cinematic representation in order to discern adequate allegories for our social and political existence, but to understand how it influences our political unconscious. The films that he analyzes are narrative figurations and examples of the way contemporary narrative “conflates ontology with geography and endlessly processes images of the unmappable system” (Jameson 1992, 4). Soja (1996) brings Jameson together with his mentor, Henri Lefebvre, to help elaborate the context of “real-and-imagined” places. Like Soja, Dear is also an advocate of postmodern theory and the postmodern city, but his focus on movies and films is much more direct. In The Postmodern Urban Condition (2000) he reworks urban theory through movies, architecture, and filmspace.
Also influenced by Jameson and Lefebvre, but trying to move beyond the work of Soja, Dear insinuates not only how spaces are produced through representations but also how it may be possible to create an urban and policy agenda for the twenty-first century that simultaneously incorporates film, local art, and global politics.

Through the 1990s, interest among geographers in understanding movies and film simultaneously as unstable constructions and quirky constructors of society rose significantly with a multitude of articles published in a wide array of journals. Unfortunately, with the rush to publish in mainstream journals many of these papers lacked geographic insight and came across as interesting but somewhat superficial place-based film reviews. To counter this trend, *Engaging Film*, edited by Cresswell and Dixon (2003), provided a set of thoughtful manuscripts that were theoretically grounded in issues of space, place, and spatiality. By so doing, the book mapped out a future for film studies in geography that reflected critical concerns of the discipline. Going beyond Friedberg’s understanding of mobility, for example, Creswell and Dixon argue that it can be thought of in a broader sense as a certain attitude, at times openly radical and at times quietly critical, toward fixed notions of people and cities as they may appear in film, art, and architecture. That is, an emphasis on mobility suggests a certain skepticism in regard to stability, rootedness, surety, and order, and yet it is not entirely about disorder either.

A large swathe of geographic endeavor these last two decades has focused on cinema and the order/disorder of the city. David Clarke’s (1997) focus on *The Cinematic City* as a “contained” spatial sphere of practice was a change away from earlier obsessions of some critical geographers, planners, architects, and cultural theorists with spatial metaphors that tended to hide and depoliticize the material conditions of lived experience. This wide-ranging book tackles issues of self, otherness, homelessness, community, race, and family through discussions of over a hundred movies. Elisabeth Mahoney’s essay, for example, begins boldly with a critique of Soja and Harvey’s postmodern geographies, and then follows with a fascinating discussion of Elizabeth Grosz’s (1995) conceptualization of the chora (a prosocial feminine and maternal space, p. 172). This theoretical “hook” is elaborated through empirical analysis of *Falling Down* (1992, directed by J. Schumacher), *Night on Earth* (1992, directed by J. Jarmusch), and *Just Another Girl* (1992, directed by D. Harris). Clarke’s continued interest in cinematic spaces narrows from cities to motel spaces in a 2009 volume coedited with Valerie Pfannhauser and Marcus Doel, and the important point about this emphasis is his enduring focus on movement and using the motel as an analytic device for understanding stopping (Clarke, Pfannhauser, and Doel 2009).

The idea of the cinematic city and narrower motel spaces joins a movement outward to consider cinema and national and international geopolitical issues (Power and Crampton 2007). It is reasonable to argue that this focus goes back to the beginnings of geography’s critical engagement with the elaborated on First, Second, and Third (World) Cinema, but today the focus is more on borders, boundaries, and territories in cinematic narratives, with concern for exploring interconnections between cinema and geopolitics (Power and Crampton 2007, 4). In one essay in this volume, the coming to film from a poststructural inspired critical geopolitics, for example, uses *Behind Enemy Lines* (2001, directed by John Moore) to elaborate gendered politics and what is called a pornographic gaze on weaponry in terms of larger structures of feeling around American militarism.

The turn to poststructuralism is important. Movies are about movement and, as such, they...
MOVIES AND FILMS, ANALYSIS OF

are fundamentally different from other forms of representation. Movement requires a different way of understanding your subject and a more flexible academic practice. Henri Bergson argued that it is only possible to understand movement intuitively because any kind of analysis that stops a frame loses the context of the frame. Standard, more structured, textual, and visual analyses do not apply to the mobilities, identities, and politics that are part of film and so many geographers are turning to poststructural methods and practices. Contemporary geographers writing about film note that the study of films and movies loses its transformative edge if it is reduced to language games and analysis of static visual symbolism. Bergson was one of the first to look carefully at what Grosz (2011, 1) calls “imperceptible movements, modes of becoming, forms of change, and evolutionary transformations.” Movement is about difference and transformation. It is the process of movement that makes and unmakes objects, including people and institutions. This is why the study of films and movies is crucial for understanding our geographic world and changing it for the better.

SEE ALSO: Art; Critical geography; Film; Imaginative geographies; Representation

References


Further reading

Multi-aperture telecentric lens

Bo Wu
Lei Ye
The Hong Kong Polytechnic University, China

A multi-aperture telecentric lens is a single telecentric lens with multiple aperture stops that enable it to capture multidirectional parallel light rays. Unlike conventional optical systems in photogrammetry or computer vision, which use a pair of stereo cameras for three-dimensional (3-D) reconstruction, a multi-aperture telecentric lens uses image(s) obtained from the light rays separated by different aperture stops for 3-D reconstruction. A multi-aperture telecentric lens is ideal for use in a variety of machine vision applications, such as noncontact measurement and inspection systems, 3-D reconstruction of close-range targets, and intelligent vision systems in robotic equipment.

The telecentric lens is discussed here firstly. Detailed configurations of the multi-aperture telecentric lens are then discussed. Key aspects of using a multi-aperture telecentric lens for 3-D reconstruction – including its geometric model and an evaluation of the potential accuracy – are then presented.

Telecentric lens

A telecentric lens is a compound lens used in an imaging system to make objects appear to be the same size independent of their location in space. Most imaging systems with conventional lenses exhibit varying magnification for objects at different distances from the lens. This causes several problems for machine vision and other applications. (i) Objects closer to the lens appear to be larger than those farther from the lens; for example, in an image of a cylindrical pipe the top and bottom crown edges appear to be concentric even though the two circles are perfectly identical. (ii) Object shapes vary with their distance from the center of the field of view (FOV); for example, circles near the center of the FOV appear to be egg-shaped when moved toward the periphery. (iii) Some features or objects may be hidden by objects closer to the lens. In contrast, a telecentric lens removes these perspective or parallax errors to produce an orthographic projection that provides the same magnification at all distances. Pioneered by Moore (1973), telecentric lenses have been pursued by researchers in various settings (Watanabe and Nayar 1996; Bai and Sadoulet 2007). In particular, telecentric lenses are commonly used in machine vision applications, where software analysis is simplified and more accurate due to the reduction of parallax. Imaging systems with telecentric lenses have made it possible to reach dimensional measurement accuracies that can be better than those generated by contact and laser-based methods (Djidel et al. 2006).

There are three types of telecentric lens. The first is an object-space telecentric lens, in which the aperture stop is placed at the front focal plane of the lens, resulting in an entrance pupil location at infinity. A shift in the object plane does not affect image magnification. Such lenses are used in machine vision systems because image magnification is independent of the objects'
distance or position in the field of view. The second is an image-space telecentric lens, in which the aperture stop is placed at the rear focal plane of the lens, resulting in an exit pupil location at infinity. A shift in the image plane does not affect image magnification. Such lenses are used in image sensors that do not tolerate a wide range of angles of incidence. The third is a double telecentric lens, in which the aperture stop is placed at the common focal plane, resulting in both the entrance and exit pupils being located at infinity. Shifting either the image or object planes does not affect magnification given that double-telecentric systems are afocal. Double telecentric lenses have magnification that is more precisely constant than those that are only object-space telecentric because the intersection position of the principal ray on the detector does not change, which allows for the precise measurement of objects regardless of their positions. Figure 1 illustrates an idealized double telecentric lens with two thin positive-powered elements, with the aperture stop placed at the common focal plane of the front and rear elements such that it only passes the light rays that are parallel to the optical axis.

In summary, telecentric lenses have the following advantages over conventional lenses: (i) constant magnification independent of shift in object and/or image planes; (ii) low distortion, normally in the range of 0.1% for high-quality telecentric lenses; (iii) reduction or elimination of perspective error; (iv) increased image resolution; and (v) uniform image plane illumination. However, they also have several disadvantages. First, more optical elements are used than in conventional lens systems due to the complex design. Second, large aperture optical elements in the region of telecentricity are required to provide a nonvignetted FOV. The large aperture and more optical elements lead to increases in the cost and weight of the imaging system. Third, traditional telecentric lenses use fixed focal lengths, which result in fixed FOVs. An investigation of varying FOVs requires the use of several fixed magnification lenses. Zinter
and Sanson (2001) presented an endeavor to develop a telecentric zoom lens for this type of task. Finally, it is not possible to obtain depth information from the telecentric images due to the orthographic projection of the telecentric lens. Despite the disadvantages inherent in the telecentric lens design due to its increased complexity, the numerous benefits make telecentric lenses a popular choice in a variety of applications. The last drawback evokes the development of a multi-aperture telecentric lens.

Multi-aperture telecentric lens

In applications such as machine vision, the 2-D information and the 3-D depth information from the scene must be extracted simultaneously. Stereo vision using multiple cameras is the conventional approach to infer depth information based on parallax from multiple perspectives, which has a long history in the fields of photogrammetry and computer vision. In recent years, a variety of techniques, such as motion parallax and depth-from-focus, have been implemented in 3-D imaging systems. However, these systems are relatively expensive and require complex camera calibration and geometric processing. As mentioned, the telecentric lens has the desirable property of orthographic projection, which makes it easier to measure or compare an object’s physical length independently from its depth in relation to the camera. However, it is not possible to obtain depth information from the image, as there is no foreshortening effect in telecentric images. The multiple aperture technique has been used in imaging systems to provide 3-D information. For example, Fife, Gamal, and Wong (2006) presented an image sensor comprising an array of apertures each with its own local integrated optics and pixel array. A lens focused the image above the sensor, creating overlapping fields of view between apertures. Image disparities from multiple perspectives in the focal plane facilitated the derivation of 3-D information. To take advantage of the orthographic projection of the telecentric lens and the image disparities from multiple, multi-aperture telecentric lenses have been investigated for use in 3-D reconstruction in recent years (Kim and Kanade 2011).

A multi-aperture telecentric lens has multiple aperture stops rather than one, as in a conventional telecentric lens. Figure 2 illustrates an idealized multi-aperture telecentric lens. There are two aperture stops $O$ and $O'$ on the focal plane, of which $O$ is located at the focal point of the lens and the other $O'$ is at a distance from the focal point. The aperture stop $O$ selectively passes light rays that are parallel to the optical axis because $O$ is located at the focal point of the lens and only passes the rays. The aperture stop $O'$ selectively passes light rays that are parallel to each other but not parallel to the optical axis. Assuming $O'$ is an infinitely small aperture stop, the light rays selected by $O'$ are parallel to the vector from the lens center to $O'$ (blue dashed line in Figure 2).

In a conventional stereo vision system, as illustrated in Figure 3, two cameras are used to form stereo images and the depth information of any object in the scene can be derived from the disparity of the same object imaged on the stereo images. From Figure 3, for an object point $P$, its disparity is:

$$d = x_1 + x_2 = \frac{f}{Z} \cdot B$$

where $f$ is the focal length of the camera, $B$ is the baseline between the stereo cameras, and $Z$ is the depth of point $P$. In equation 1, both $B$ and $f$ are fixed, and thus the disparity $d$ is linearly proportional to the inverse of the depth.

In a simplified multi-aperture telecentric lens system ($1 \times$ magnification), as illustrated in Figure 4, the distance between the two aperture stops is not constant but rather varies according to the object’s depth. This variation allows for the extraction of depth information from a single telecentric image.
MULTI-APERTURE TELECENTRIC LENS

stops is $B$, which can be considered the baseline between the two aperture stops. The focal length of the lens is $f$. The image plane is located at a distance $f$ from the focal plane. A spatial auxiliary coordinate system is used with its origin at the center of the image plane. Its $z$-axis aligns with the optic axis and points to the front of the lens while its $y$-axis points up. Its $x$-axis is perpendicular to the plane determined by its $z$- and $y$-axes. For an object point $P$ located at $(0, Y, Z)$, the light rays selected by the aperture stops $O$ and $O'$ pass through the points $p_1$ and $p_2$ at the image plane. The distance between $p_1$ and $p_2$ is the disparity of the same point $P$ imaged on the image plane through the two aperture stops. The coordinates of $p_1$ and $p_2$ on the image plane can be determined as:

\[(0, -Y) \text{ and } (0, -Y - \left(4 - \frac{Z}{f}\right)B)\]

respectively. Therefore, the disparity is:

\[d = \left(4 - \frac{Z}{f}\right)B \quad (2)\]

In equation 2, both the baseline length, $B$, and focal length, $f$, are fixed; thus the disparity, $d$, is linearly proportional to the depth, $Z$, of the object point. When $Z = 4f$, the disparity, $d$, becomes zero. It should be noted that $d$ can also be negative.

A multi-aperture telecentric lens system has the following major advantages over a conventional stereo vision system, as follows:

1. The disparity given by the multi-aperture telecentric system is linearly proportional to the depth, which is a critical difference from the conventional stereo vision system in which the disparity is proportional to the inverse depth, as already mentioned. This property results in a more accurate and stable determination of depth information.
The absolute location of an object point other than the depth can be easily determined based on the image captured by the aperture stop at the focal point, which maintains the advantages of a conventional telecentric lens.

When an object point goes farther, the angle between light rays in the conventional stereo vision system becomes narrower, as can be inferred from Figure 3. This results in less disparity in the image space, that is, less accuracy in depth determination. In the multi-aperture telecentric system, the angle between the rays remains constant because the disparity is linearly proportional to the depth, that is, the accuracy of depth determination remains at the same level.

Figure 3  A conventional stereo vision system.
Figure 4  A simplified multi-aperture telecentric lens system (1× magnification).

Unlike the conventional stereo vision system, the multi-aperture telecentric system does not require any photometric calibration or intensity adjustment between cameras because it is a single-lens imaging sensor.

3-D Reconstruction and accuracy potential using a multi-aperture telecentric system

The geometric model for 3-D reconstruction might differ among multi-aperture telecentric system types. Taking the simplified multi-aperture telecentric lens system illustrated in Figure 4, for example, the geometric model can be derived as:

\[
\begin{align*}
X &= -\frac{u_1 + u_2}{2} \\
Y &= v_1 \\
Z &= \left(4 - \frac{d}{B}\right)f
\end{align*}
\]  

(3)

where \((X, Y, Z)\) are the 3-D coordinates of the object point \(P\), \((u_1, v_1)\) and \((u_2, v_2)\) are the image coordinates on the images acquired from the aperture stops \(O\) and \(O'\), respectively, and \(d = v_1 - v_2\). From equation 3, it can be seen that the geometric model of the simplified multi-aperture telecentric lens system for 3-D
reconstruction is much simpler than other machine vision systems.

Camera calibration is an important step in 3-D reconstruction using conventional stereo vision systems, but with the multi-aperture telecentric lens system, because the telecentric lens offers orthographic projection and the projection center is at infinity, the lens distortions for telecentric lens systems are very small (normally in the range of 0.1%). Therefore, only the offset of the principal point needs to be considered in the calibration. This can be achieved by following the general methodology of calibrating the principal point in photogrammetry or computer vision by adding additional terms to the right-hand side of equation 3. The accurate principal point can be determined through several control points measured in object and image space. In a high-quality multi-aperture telecentric lens system, the calibration procedure might be ignored, as the influences of lens distortion and the offset of the principal point on the measurement accuracy are negligible.

Image matching is another important issue in 3-D reconstruction. In the multi-aperture telecentric system, image matching involves identifying the image coordinates on the images acquired from the two aperture stops, which belong to the same object point. The illustration of the multi-aperture telecentric lens in Figure 4 reveals that for an object point, the light ray should pass through two aperture stops and build a pencil of the planes containing the baseline. Because the image plane cuts the pencil of the planes, all of the correspondences should be on a line – the epipolar line. For an image plane parallel to the focal plane, all of the epipolar lines are parallel to the baseline and parallel to each other. Therefore, image matching is much easier than it is in conventional stereo vision systems, as it only needs to search for image correspondences along the epipolar lines, which are parallel to the baseline. Automatic and reliable image matching can be expected for the multi-aperture telecentric system.

To evaluate the potential measurement accuracy of the multi-aperture telecentric system, the simplified multi-aperture telecentric lens system illustrated in Figure 4 is used for analysis. Through an error-propagation derivation based on equation 3, the measurement errors can be calculated as:

\[
\begin{align*}
\sigma_X^2 &= \frac{1}{4} \sigma_{u_1}^2 + \frac{1}{4} \sigma_{u_2}^2 \\
\sigma_Y^2 &= \sigma_{v_1}^2 \\
\sigma_Z^2 &= \frac{(4f - Z)^2}{B^2} \sigma_B^2 + \frac{f^2}{B^2} \sigma_d^2
\end{align*}
\]

where \(\sigma_X\), \(\sigma_Y\), and \(\sigma_Z\) are the standard errors of measurement in the \(X\), \(Y\), and \(Z\) directions, respectively; \(\sigma_B\) is the standard error of the baseline length; \(\sigma_d\) is the disparity measurement error; \(\sigma_{u_1}\), \(\sigma_{u_2}\), and \(\sigma_{v_1}\) are the measurement errors of the image coordinates. \(\sigma_d\) can be determined by the accuracy of image matching. According to a theoretical analysis, pixel-level image matching (correlation) can reach an accuracy of about one-third of a pixel. Given the homogeneity of the imaging sensor:

\[
\sigma_{u_1} = \sigma_{u_2} = \sigma_{v_1} = \sqrt{1/2 \sigma_d}
\]

As equation 4 shows, the measurement errors in the \(X\) and \(Y\) directions are fixed because the CCD sensor is fixed, whereas the measurement error in the \(Z\) direction (depth) is related to the baseline length, depth, focal length, and disparity. To investigate the measurement error in the \(Z\) direction, a theoretical analysis is conducted based on the third equation in equation 4, referring to the parameters of a commercially available telecentric lens as listed in Table 1.

Because the best working distance is 500 mm, distances ranging from 300 mm to 1000 mm
are investigated. The baseline length error is assumed to be 10 μm. The measurement errors in the Z direction are calculated for each working distance under different baseline lengths from 10 mm to 60 mm. The results are shown in Figure 5.

Figure 5 shows that as the baseline length increases, the measurement error in the Z direction decreases, and the farther the working distance is, the larger the decline. When the baseline length is about 30 mm (about 14% of the focal length), a measurement accuracy in the Z direction of 0.1% (measurement error divided by distance) can be expected around the best working distance.

It should be noted that only a theoretical analysis has been performed here. The real performance of the multi-aperture telecentric system in 3-D reconstruction requires further investigation based on experiments using a real multi-aperture telecentric system.

There are several promising future research topics related to multi-aperture telecentric systems. For example, it remains difficult to obtain two separate images from the two aperture stops while maintaining relatively simple optical geometry. Previous studies have used complicated lens elements to separate images from multiple aperture stops (Yahav and Iddan 2000). A more productive approach would be
to design the apertures to be controllable and the two apertures to open successively to acquire two images separately. However, the actual implementation of this mechanism depends on future research. In addition, real applications in machine vision may require telecentric lens systems with better magnification capabilities (e.g., 10× or 20×). The geometric model for multi-aperture telecentric systems with magnification capabilities will be more complicated, as it will involve more optical elements, which requires future research.

SEE ALSO: Photogrammetry: 3-D from imagery

References


Further reading


Multicriteria decision-making

Piotr Jankowski  
San Diego State University, USA

Multicriteria decision-making (MCDM) denotes a structured approach to solving decision problems involving decision alternatives. The fundamental idea underlying MCDM can be described as follows: find feasible options (alternatives) for solving a decision problem, evaluate the overall merit of each feasible option by measuring its performance on multiple criteria (objectives), and compare the relative merits of evaluated options to choose the best performer. MCDM has been an academic field with practical applications in many domains, including finance, marketing, business processes, engineering, medicine, environmental management, and planning. In geographical sciences, MCDM methods have been developed to account for explicit spatial footprints of decision alternatives. Much of the development happened through integrating MCDM with geographic information systems (GIS) and putting MCDM methods at the core of spatial decision-support systems (SDSS). Applications of MCDM in decision problems involving explicit representation of geographical space include land-use allocation, site selection, environmental planning, urban and regional planning, ecological management, transportation, waste management, water management, agriculture and forest management, natural hazard management, and real estate and housing.

Roots of MCDM

Early papers on MCDM methods started appearing in the late 1960s and early 1970s, and the first MCDM conference was organized in 1975. However, the roots of MCDM can be traced to the nineteenth century and even earlier. According to Kõksalan, Wallenius, and Zionts (2011), Benjamin Franklin (1706–1790), the American statesman, scientist, and inventor, was a pioneer of a simple multicriteria approach to decision-making. In making important decisions, he used a two-column table to evaluate a decision by writing arguments for the decision in one column and arguments against in the other. He then employed an informal weighting system by identifying for and against arguments of equal importance and crossing them out. The column that had more uncrossed arguments remaining identified the decision to make.

Theoretical concepts that later contributed to the development of MCDM come from mathematics, economics, behavioral decision theory, and mathematical programming. Mathematical underpinnings of MCDM came from calculus methods of identifying function optima, known contemporarily as mathematical optimization, contributed by Pierre de Fermat (1601–1665), Isaac Newton (1642–1727), Joseph-Louis Lagrange (1736–1813), and Carl Friedrich Gauss (1777–1855). Other mathematical concepts, fundamental for MCDM, came from set theory created by Georg Cantor (1845–1918). In the field of neoclassical economics, Francis Edgeworth (1845–1926) introduced the concept of the indifference curve and set the foundations of utility theory, developed after World War II.
by Howard Raiffa (b. 1924), Ralph Keeney (b. 1944), and Peter Fishburn (b. 1936). An indifference curve can be depicted by a graph representing choices between two different categories (e.g., two different resources). The defining property of the indifference curve is that any point that belongs to it represents equal utility (satisfaction) that can be derived from the two corresponding resource levels found on $x$- and $y$-axes. Just as one can graph a utility (tradeoff) level between two resources by finding a proper form of indifference curve, one can also find a utility function for a given resource. The problem then becomes how to aggregate various utility functions into one synthetic, index-like measure. The question arises as to why this is a problem from a decision-making standpoint. Imagine selecting real estate based on a number of measurable criteria (price, square footage, number of bedrooms, distance to work, etc.). Each of the criteria could be represented by a separate utility function representing a specific level of utility for a specific value of a criterion in question. Providing a measure of the overall utility for each piece of real estate (e.g., a house), in order to facilitate their comparison, requires aggregating various individual utilities. The French-Italian economist Vilfredo Pareto (1848–1923) studied the problem of aggregating criteria into a single overall measure. The concept of efficient allocation of resources, also known as Pareto efficiency, is due to him. Accordingly, a Pareto-efficient choice is achieved if no other choice exists that performs better than the choice under consideration on most of the criteria without scoring worse on at least one criterion. The concept of Pareto efficiency has become influential in economics and decision theory. Along with the Pareto efficiency frontier representing a set of Pareto-efficient choices, it has also been one of the key concepts in MCDM.

Much of the theoretical work on decision-making that influenced MCDM came from behavioral economics. Working at the intersection of psychology and economics, Ward Edwards (1927–2005) published in 1961 an influential paper (Edwards 1961) probing into psychological mechanisms of decision-making and establishing the field of behavioral decision theory. Herbert Simon (1916–2001) postulated that people in their decision-making practices by and large did not adhere to a theoretical model of “rational man,” which was one of the cornerstones of neoclassical economics. Instead, Simon argued that people followed a satisficing behavior driving them to achieve subjective aspiration levels, which did not equal optimal levels. He termed such a decision-making behavior “bounded rationality.” The concept of aspiration levels was adopted by a number of MCDM methods developed during the 1970s, 1980s, and 1990s. Simon is also credited with popularizing a three-stage model of rational decision-making (Simon 1976), in which a decision-making task is subdivided into three stages: (i) intelligence – search for feasible alternatives, (ii) design – involving the determination of consequences/impacts of feasible alternatives, and (iii) choice of a particular alternative guided by a decision rule. Simon’s model of rational decision-making has been influential in decision-support systems (DSS) and SDSS including MCDM methods.

The search for feasible decision alternatives and determination of their characteristics can be facilitated by formulating and solving a constrained optimization problem or a mathematical program. Solving a mathematical program with many decision variables can be a difficult task, even if the program is composed of linear inequalities (i.e., a linear program). In 1947, George Dantzig (1914–2005) proposed the simplex algorithm providing an efficient procedure for solving linear
programs. This was a breakthrough leveraged by developments of mainframe computers and programming languages, which enabled solving large linear problems composed of hundreds, even thousands, of decision variables. Danzig’s discovery was preceded by a similar method of solving linear programs proposed in 1939 by Leonid Kantorovich (1912–1986) and applied to central planning in the Soviet Union. Unfortunately, Kantorovich’s work remained largely unknown to the Western world until the 1970s. The simplex method popularized linear programming and spurred many of its applications in the 1950s and 1960s in resource management, business and military logistics, and production management. Practical applications were accompanied by theoretical developments of linear programming extensions including nonlinear programming, goal programming, and multiple objective linear programming. In 1967, Arthur Geoffrion published an article in *Operation Research*, in which he showed how Pareto-efficient solutions can be generated for optimization problems with two objective functions. The period of the next 40 or more years witnessed the emergence of the MCDM field marked by developments in theory, methods, and applications, which proliferated into new domains including location analysis, land-use planning, and applications of GIS.

**MCDM concepts**

Fundamental concepts in MCDM relate to its structural components, including *objective, attribute, criterion, constraint, preference, decision variable, dominance, decision alternative, and decision rule*. Among the fundamental concepts are also two different ways of addressing MCDM, as either a problem of choice from a set of pre-existing decision alternatives, or a problem of a search for feasible decision alternatives. The former is referred to as the *multiple attribute decision-making* (MADM) approach, while the latter is called the *multiple objective decision-making* (MODM) approach.

*Objective* refers to a desired level of property or quantifiable result of some action that follows a decision. Examples include minimum cost, maximum profit, or some broader notion of desired outcome that can be quantified (e.g., increasing a suitable habitat). *Attribute* is a measurable characteristic of an entity and/or action comprising a decision alternative. The MCDM approach adopts the notion of more than one objective that is relevant for analyzing a decision problem. It follows that multiple attributes are employed to characterize each decision alternative and that the same set of attributes is used for all decision alternatives under consideration. A good example of such a set of attributes is a list of fields (attributes) in a GIS layer. Each spatial entity present in the layer, which could also be conceptualized as a spatial decision alternative, is characterized by a common set of fields in a layer attribute table.

A *criterion* is an attribute that has been qualified by a desired direction of improvement (e.g., maximization or minimization of attribute value). Hence, criteria in MCDM are the basis for measuring, either explicitly or implicitly, the performance of decision alternatives. The explicit measurement happens when decision alternatives and their outcomes are known. In such a case the performance of decision alternatives is represented by criteria values, which can be computed using various analytic methods. Feasible decision alternatives and their outcomes need to be identified in the first place. Under such a requirement, criteria and their predefined directions of improvement can be used to gauge progress in the search for feasible decision alternatives. To facilitate the search, criteria are used as quantifiable objectives, which can either be minimized or maximized.
Criteria can also be constrained by imposing a standard: a minimum value required, a maximum value allowed, or a range of admissible values. Constraints serve the purpose of ensuring decision alternative performance levels or representing the availability of resources required by decision alternatives. Thresholds and cutoffs are common forms of constraints requiring certain minimum criterion values to be achieved (thresholds) or certain maximum criterion values not to be exceeded (cutoffs).

Preference is an expression of criterion importance relative to other criteria under consideration. Criteria preferences are commonly represented by weights, with the property that the sum of weights equals 100% of preferences (typically the weight values range between 0 and 1 and the sum of weights equals 1). A criterion weight is in effect a scaling constant indicating a relative importance of a given criterion vis-à-vis other criteria. Another common way of representing criteria preferences is by means of tradeoffs. A tradeoff is a statement of indifference (equivalence) between \( x \) units of criterion \( c_1 \) and \( y \) units of criterion \( c_2 \).

The decision variable is a measurable quantity that has a value for every decision option. The simplest of all is a binary integer decision variable (1–0) often associated with a go–no-go (yes, no) type of decision. This type of decision variable can represent location decision alternatives, where not the site itself, but the question of whether or not to locate an activity at a given site, gives rise to location decision alternatives. Decision variables can also take on integer and real values.

The dominance principle, due to the already mentioned Vilfredo Pareto (hence called Pareto-dominance), states that a decision alternative is nondominated if there is no other feasible alternative that surpasses this alternative on any of the criteria without reducing performance on another criterion. Nondominated alternatives are naturally preferred over dominated alternatives and finding them has been the focus of many MCDM methods.

Decision alternatives (decision options) give purpose to MCDM regardless of whether there are only two alternatives representing a so-called go–no-go decision situation, or whether there are more than two feasible alternatives. In spatial decision problems, each decision alternative consists at the minimum of action specification (what to do) and location (where to do it).

A decision rule is an analytic procedure for processing performance data about decision alternatives in order to find a solution to a decision problem. Three generic approaches to solving a decision problem with MCDM are as follows.

- Selection: given a set \( A \) of alternatives, the selection task operation involves finding a subset \( A' \) of \( A \) composed of as small as possible a number of alternatives, judged by decision-makers as the most satisfying.
- Sorting: the sorting operation (also called classification) consists of assigning each alternative from \( A \) to one of the predefined categories. The assignment should be based on the intrinsic measure of a criterion for an alternative and not on its comparison with other alternatives from \( A \). However, in practice the assignment is often based on relative differences of alternatives along a criterion.
- Ranking: the ranking operation involves establishing a preference pre-order on the set of alternatives \( A \). The pre-order represents a priority list of the alternatives.

The MADM approach to MCDM requires that the selection, sorting, or ordering of decision alternatives be made on the basis of their criteria performance. MADM problems are
assumed to have an a priori known number of decision alternatives. Problems solved with the MADM approach are also called discrete because the number of evaluated decision alternatives is known and hence it can be enumerated. In the spatial decision domain, feasible decision alternatives are commonly identified with GIS suitability analysis leading to identification of suitable locations. Depending on the scale and spatial data representation model used, the number of suitable (feasible) decision alternatives can range from few to many with many corresponding to a situation, where each spatial unit (e.g., a raster layer cell) represents a feasible location decision alternative.

In the MODM approach, contrary to MADM, the decision alternatives are not known a priori. MODM uses a mathematical framework to search for a set of feasible decision alternatives. The search is carried out by an algorithm that evaluates different candidates for decision alternatives, composed of decision variable values, which are bound by constraints. The decision rule that guides the search is to optimize the objectives. Each alternative, once identified, is judged by how closely it satisfies an objective or multiple objectives. Since the decision variables can take on real number values, problems tackled with the MODM approach are also called continuous. The MODM approach can potentially result in a large number of feasible solutions (decision alternatives); hence, MODM methods focus on finding non-dominated solutions comprising a subset of feasible solutions. A comparison of both approaches contrasting their salient characteristics is given in Table 1.

**MCDM methods**

Many MCDM methods have been proposed over the last 40 years or more for solving either MADM or MODM problems. MADM methods, also referred to as scoring or multiple criteria evaluation methods, use simpler algorithmic procedures than MODM relying on optimization and heuristic algorithms. It is beyond the scope of this encyclopedic overview to provide a full account of various MADM and MODM methods. Hence, only the methods that have been widely used in spatial MCDM are briefly reviewed.

**MADM methods**

Two approaches to MADM, one based on multiple attribute utility theory (MAUT) and the other on outranking relation, led to the development of a number of MADM methods. The MAUT-based methods use the notion of utility/value function, which theoretically can be established for any decision criterion. If a criterion value attained for a given decision alternative is substituted for the function's argument, the function's value returned for the argument represents a standardized number ranging between 0 and 1 (or 0% and 100% representing the utility/value scale). The MAUT-based methods use various techniques of determining the shape/form of utility/value functions and aggregating the functions' values for multiple criteria in order to arrive at one number representing the overall utility/value of a given decision alternative. A simple and commonly used form of multiple attribute value function is the additive function given in equation 1, where:

\[ V_i = \sum_{j=1}^{m} w_j x_j \]

\[ 0 \leq w_j \leq 1 \text{ and } \sum_{j=1}^{m} w_j = 1 \]

\[ V_i \] represents the overall value of the \( i \)-th decision alternative

\( w_j \) is the weight representing the relative importance of the \( j \)-th criterion, such that \( 0 \leq w_j \leq 1 \) and \( \sum_{j=1}^{m} w_j = 1 \)
MULTICRITERIA DECISION-MAKING

Table 1 Comparison of MADM and MODM approaches to MCDM.

<table>
<thead>
<tr>
<th>Multiple attribute decision-making</th>
<th>Multiple objective decision-making</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicitly defined set of decision alternatives</td>
<td>Implicitly defined set of feasible decision alternatives</td>
</tr>
<tr>
<td>Decision problem is structured with a set of alternatives and set of criteria</td>
<td>Decision problem is structured with objective functions and constraints</td>
</tr>
<tr>
<td>Decision rule is based on a multiple attribute utility/value function or an outranking relation</td>
<td>Decision rule uses optimization algorithm to identify Pareto-nondominated solutions</td>
</tr>
<tr>
<td>Requires information on the decision-maker’s preferences</td>
<td>Information on the decision-maker’s preferences is helpful but not required</td>
</tr>
</tbody>
</table>

\[ v_j \] is the value function for the \( j \)-th criterion

\[ x_{ij} \] is the \( j \)-th criterion value for the \( i \)-th decision alternative.

The additive function form given in (1) can be simplified when the value function \( v_j(x_{ij}) \) is the same for all \( N \) criteria. This case is presented in the equations 2–4, where \( \min_j(x_{ij}) \) and \( \max_j(x_{ij}) \) are the minimum and maximum criterion values for the \( j \)-th criterion, respectively.

\[
V_i = \sum_{j=1}^n w_j v(x_{ij})
\] (1)

\[
V_i = \sum_{j=1}^n w_j v(x_{ij})
\] (2)

\[
v(x_{ij}) = \frac{x_{ij} - \min_j(x_{ij})}{\max_j - \min_j}, \text{ for the } j \text{-th criterion to be maximized}
\] (3)

\[
v(x_{ij}) = \frac{\max_j(x_{ij}) - x_{ij}}{\max_j - \min_j}, \text{ for the } j \text{-th criterion to be minimized}
\] (4)

MADM functions that combine criteria weights with criteria values, in a similar way to the function given in (1), are called aggregation functions. The additive value function in (1) results in ordering decision alternatives from the highest to the lowest performer. The resulting order is called a weak order since it follows a complete and transitive binary relation, which allows for an inference of the kind: if \( A > B \) and \( B > C \), then \( A > C \) (where \( A, B, C \) denote the decision alternatives). Ordering the decision alternatives requires that the criteria weights represent tradeoffs. In practical terms, this means that, for example, given two weights \( w_1 = 0.5 \) and \( w_2 = 0.25 \), the tradeoff between criterion 1 and criterion 2 equals \( 0.5/0.25 = 2 \), or in other words, the decision-maker is willing to trade off two units of criterion 2 for one unit of criterion 1. This type of information may not always be available, and if it is then its reliability is not always assured.

Equations 3 and 4 define a nonlinear value function that standardizes the criteria values to fall in the range from 0 to 1, with 0 representing the lowest and 1 representing the highest worth of a criterion value. Equations 2–4 provide the basis for weighted linear combination (WLC), which is one of the simplest MADM methods. In the spatial realm, WLC has been the most popular MADM method. According to an extensive
literature review by Malczewski (2006), of all articles describing research on integrating GIS with MCDM published between 1990 and 2004, 143 (39.3%) included WLC as the leading MCDM method.

Other prominent MADM methods that loosely follow the multiple attribute utility/value function approach include the analytical hierarchy process (AHP) – developed by Thomas Saaty (1980); ideal point methods, an example of which is the technique for order of preference by similarity to ideal solution (TOPSIS) – developed by Hwang and Yoon (1981); and ordered weighted averaging (OWA) operators – proposed by Yager (1988).

AHP has been arguably the most widely used MADM method, with applications in business, politics, engineering, and environmental management including spatial analysis with GIS. The method is based on structuring (decomposing) a decision problem into a multilevel hierarchy, and on pairwise comparisons of elements within the hierarchy. A typical hierarchy in a decision problem might include the decision goal, criteria, and alternatives. Criteria might be decomposed into two levels: higher-level criteria representing general categories (e.g., environmental, economic, and social) and lower-level criteria specifying each of the higher-level criteria. With the hierarchy established, the elements within each hierarchy level are compared to one another in a pairwise manner to measure the relative importance of their contribution to a higher-level element. For example, in a pairwise comparison of two lower-level environmental criteria, the importance of one criterion vis-à-vis a higher-level environmental criterion is judged against the importance of the other criterion. The judgments are expressed numerically using the 1–9 ratio scale, where 1 denotes equal importance and 9 represents the extreme importance of one element over the other element. The results of the pairwise comparisons obtained at each level of the hierarchy are then arranged in a pairwise, square comparison matrix and aggregated by computing its eigenvector. The elements of the eigenvector are weights \( w_j \), such that

\[
\sum_{j=1}^{m} w_j = 1
\]

expressing the importance of the \( j \)-th element relative to other elements at the same level of hierarchy. The procedure leads to rank-ordering of decision alternatives on the basis of a final score computed as a simple weighted sum for each decision alternative. In the weighted sum formula used by AHP, criteria weights are multiplied consecutively by a decision alternative weight and the products are summarized. The resulting sum is a ratio-scale score representing the worth of an alternative relative to other alternatives.

Various MADM methods have been developed around the concept of distance separating a decision alternative from an a priori reference point. A multidimensional reference point, or rather a vector of values corresponding to decision criteria under consideration, can be conceptualized as an aspiration level for each criterion, beyond which the utility/value of the criterion increases only marginally or remains unchanged. Another conceptualization of the reference point is used in the TOPSIS method where two opposing vectors of criterion values are established. One of them – called the ideal – is composed of the most favorable values for the criteria under consideration. This includes the maximum values for the criteria to be maximized and the minimum values for the criteria to be minimized. The other – called the nadir – is composed of the least favorable values, including the minimum values for the criteria to be maximized and the maximum values for the criteria to be minimized.
A distance between a given criterion value and the corresponding ideal and nadir values is computed for each criterion and the overall normalized separation metrics from ideal and nadir are computed for each decision alternative. The alternatives are ordered based on the proximity to ideal and the separation from nadir.

The OWA operators-based method computes for each decision alternative an aggregate of criteria weights and criteria values using two sets of weights: the criterion weights representing relative preferences attached to decision criteria and the order weights corresponding to a selected OWA operator. Common OWA operators include MEAN (order weight values are the same and they equal the ratio of 1/number of criteria), MIN (corresponding to Boolean AND), and MAX (corresponding to Boolean OR). In the case of the MEAN operator, the result of OWA-based calculation becomes identical to the result of WLC. The MIN and MAX operators represent two extreme cases, with MIN corresponding to a “pessimistic” decision choice behavior or the worst-case scenario, and MAX corresponding to an “optimistic” behavior or the best-case scenario. Other OWA operators residing between MIN and MAX can be selected to represent choice attitudes ranging between pessimistic and optimistic.

Methods based on outranking relation are associated with the European School of MCDM and with Bernard Roy (1996), his collaborators, and students. The outranking relation is a binary relation \( S \) on the set \( X \) of alternatives such that \( xSx \) (read: alternative \( x \) outranks alternative \( y \)) if there is sufficient information to reason that \( x \) is at least as good as \( y \), and \((x, y) \in X\). In most of the methods the outranking relation is established through a series of pairwise comparisons of the decision alternatives by employing the concordance – discordance principle. The principle states that \( xSy \) if (i) a majority of the criteria supports this assertion, thus fulfilling the concordance condition, and (ii) the remainder of the criteria representing the discordance condition do not outweigh in their aggregate the concordance majority. The crucial difference between the outranking relation approach and the value function approach is that the former uses ordinal comparisons and does not require quantitative tradeoffs between criteria. In practical terms this means that the outranking approach does not require a precise quantification of the decision-maker preferences in regard to the criteria. There are many situations where eliciting preferences from the decision-maker is difficult at best (e.g., when the decision-maker is not an individual but a diverse organization) and the preferences are only partially defined. In such situations, a less rigorous representation of preferences than the tradeoff weights, such as the one based on the concordance – discordance principle, has its advantages.

The oldest, simplest, and perhaps the best known outranking method is ELECTRE I (ELimination Et Choix Traduisant la REalité I (Elimination and Choice Expressing Reality I)). The method can be used effectively to sort the decision alternatives into those that are inferior and those that warrant further scrutiny (acceptable choices). The sorting is done by calculating measures of concordance and discordance for each pair of decision alternatives under consideration, and selecting a concordance cutoff and a discordance threshold. As long as the minimum level of support (concordance cutoff) is achieved while not exceeding the threshold for the opposite point of view, the concordance in favor of an alternative \( x \) outranking an alternative \( y \) outweighs the discordance. The successors to ELECTRE I include ELECTRE II, III, IV, and ELECTRE TRI.
PROMETHEE (Preference Ranking Organization METHod for Enrichment of Evaluations) methods comprise another well-known family of outranking MADM methods.

MODM methods

Multiple objective decision problems are typically structured by defining more than one objective function and constraining a solution space. A generic linear MODM problem has the following form:

\[
\begin{align*}
\text{Optimize} & \rightarrow f_1(X), \ldots, f_j(X) \\
A \times X & \leq B \\
X & \geq 0
\end{align*}
\]

(5)

where:

\[X = x_1, x_2, \ldots, x_j, \ldots, x_m\]
\[A = a_{ij}, i = 1, \ldots, n, j = 1, \ldots, m\]
\[B = b_1, b_2, \ldots, b_i, \ldots, b_n\]

\(X\) represents the vector of decision variables \(x_j\)
\(A\) represents the matrix of inputs \(a_{ij}\) required per unit of decision variable \(x_j\)
\(B\) represents the vector of available inputs \(b_i\).

MODM problems are typically solved with optimization techniques. A common solution strategy involves transforming a problem with multiple objective functions into a single objective function by weighting and summing each objective based on decision-maker preferences. For example, given two objective functions to be optimized separately, \(\text{Min} \rightarrow f_1(X)\) and \(\text{Max} \rightarrow f_2(X)\), and the decision-maker’s preferences for the two objectives expressed by the weights \(w_1\) and \(w_2\), where \(w_1 + w_2 = 1\), one can formulate the following single objective optimization problem: \(\text{Min} \rightarrow w_1 f_1(X) - w_2 f_2(X)\).

Optimization algorithms developed as extensions of linear programming are not very practical for solving large spatial MODM problems as these algorithms may be computationally expensive and do not guarantee finding optimal problem solutions. Various heuristic methods, including evolutionary multi-objective algorithms (EMO) and their subset, called genetic algorithms (GA), were proposed as alternatives to optimization algorithms for solving large MODM problems. Evolutionary algorithms are heuristics inspired by Darwinian ideas of evolution suggesting that fit members will emerge in a population through natural selection. A genetic algorithm begins with an initial random population of individuals, or potential solutions. In an optimization problem the fitness of individuals is evaluated based on their objective functions and constraints. Decision variables of the optimization problem are represented by chromosomes in the GA, which are often a string of bits or an array of values, where each value may be called a gene. For example, in a spatial decision problem a chromosome might be represented by a binary array of 1s representing locations that are initially selected and 0s representing nonselected locations. Typically, individual solutions with high fitness are chosen for reproduction, which consists of a series of operations. The most common genetic operators are selection, crossover, and mutation. Selection operations choose individual solutions, and then with a specified probability a crossover operation swaps a portion of chromosome genes between pairs of solutions to create new children solutions. A mutation operation may then be applied and considered as a random perturbation that will occur with some specified probability. Mutation operators often randomly alter one or more gene values, or initialize a new solution from scratch and add it to the population.
MULTICRITERIA DECISION-MAKING

MCDM in geography


The initial thrust of work on bringing MCDM into the realm of the spatial domain focused on the adaptation of methods proposed outside geography to spatially explicit representations of space (raster/field and vector/object). The crucial distinction between the two predominant models of planar space is that in the raster/field model, each fundamental element of model representation (i.e., a raster cell) is a potential location of a decision alternative, whereas in the vector/object model, each discrete object on the map (i.e., point, line, or polygon) is a potential location. Much of the initial work went into integrating MCDM methods with GIS software. Various integration forms were proposed, ranging from a simple file exchange between MCDM tools and GIS, through loose coupling, automating the exchange of data between MCDM and GIS modules, to full integration, embedding an MCDM module into a GIS. An example of the latter is the IDRISI GIS software containing a fully integrated MCDM module with multicriteria decision analysis functions.

Spatial implementations of MCDM included both MADM and MODM, with the majority of methods belonging to the MADM category. Owing to the early applications of MCDM in planning, relative simplicity of solution algorithms and mathematical representation, and obvious applicability to land suitability analysis, the ratio of publications reporting on the use of MADM methods in comparison to publications on MODM has been more than 2:1 (Malczewski 2006). The most popular MADM methods for solving spatial decision problems included WLC, ideal/reference point methods (TOPSIS and MOLA – an iterative procedure for multi-objective land allocation developed and implemented in IDRISI GIS software), AHP, and outranking methods ELECTRE and PROMETHEE. The MODM methods applied to spatial decision problems included multiple objective optimization, heuristic (evolutionary/genetic) algorithms, and reference point methods.

The range of MCDM applications in the spatial domain has been impressive. The major application domains include environmental management, transportation, urban and regional planning, water resources, agriculture, forestry, natural hazards, recreation and tourism, and real estate. The most common decision problem tackled with the MCDM approach has been land suitability analysis, followed by evaluation of decision alternatives, site selection, resource allocation, route selection, impact assessment, and location allocation.

Recent advances at the intersection of MCDM and geography have been centered on developing spatially explicit MCDM methods. In the area of MODM, spatial optimization methods have been proposed that take advantage of various representation models and address spatially explicit objectives such as contiguity, compactness, and area shape (Tong and Murray 2012). Using the principle of range sensitivity, according to which criteria weights may depend on a spatially variable range of criteria values, Malczewski (2011) proposed a local version of the WLC method including local and global
MULTICRITERIA DECISION-MAKING

criteria weights. Prior to this development, a conventional or global version of WLC assumed spatial homogeneity of criteria weights, meaning that the importance of a decision criterion is stationary throughout the study area. The local version of WLC adopts instead the idea of a heterogeneous criterion surface, in which the criterion weight (criterion preference) may vary depending on the range of criterion values within a local neighborhood.

Criteria weights are the obvious source of uncertainty in MADM models due to unstable human preferences and the difficulty of articulating preferences in regard to decision criteria. Accounting for this inherent bias in criteria weights has been traditionally a task of sensitivity analysis in MADM methods. The most common and also straightforward form of sensitivity analysis has been a so-called one-at-a-time (OAT) approach, in which a selected criterion weight is changed by a certain value/percentage up and down, while the other weights are kept unchanged, and the effect of change on the model solution (e.g., rank-order of decision alternatives) is observed. The OAT approach, however, does have its limitations, including a potentially arbitrary change of weight values and the inability to account for interaction effects among the weights (i.e., multiple weights changing simultaneously). To address these shortcomings in the sensitivity analysis of criteria weights in MADM models, Ligmann-Zielinska and Jankowski (2014) proposed an integrated and spatially explicit approach to uncertainty and sensitivity analysis. The approach and its methods allow generating an uncertainty map of an MADM model solution (e.g., an uncertainty map of land suitability ranking) and a number of sensitivity maps, which may reveal areas where the specific criteria weights influence the uncertainty of the model solution.

Future directions

It is reasonable to expect that further development of spatially explicit MCDM methods will continue by incorporating spatial representation and abstraction in decision problem modeling with multiple criteria. In addition, three research areas offer opportunities for future advances at the intersection of MCDM and geography: (i) flexible MCDM toolkits, (ii) behavioral considerations, and (iii) MCDM for big data.

Developing flexible toolkits has been a long-term objective of research on DSS and SDSS, which until now has eluded any satisfactory progress. The main motivation has been the high cost and considerable effort required to build realistic models of real-world decision situations, in light of which one often resorts to simplifications in problem representation and fitting a problem into the MCDM method, rather than the other way around. Advances in the Semantic Web, domain-specific model ontologies, and linked data might offer conceptual and methodological support for developing flexible MCDM toolkits, which might enable quick and inexpensive tailoring of MCDM models to fit a decision problem at hand.

Behavioral considerations of decision-making have been given attention since the 1950s in psychology, economics, and decision science. The Nobel Prize in economics awarded in 2002 to Daniel Kahneman amplified the importance of research on the behavioral foundations of decision-making. In the field of behavioral geography, some research has focused on the question of how cognitive processes, including environmental perception and cognition, wayfinding, and the development of attitudes about space and place, explain human decisions and behavior related to choice-making. Yet, despite the recognition of this topic, behavioral issues have received little attention from MCDM
MULTICRITERIA DECISION-MAKING

Researchers in geography and GIScience. Not much is known about how the attitudes about space and place influence possible biases in the selection of decision criteria and the elicitation of their preferences expressed by weights. Similarly, little is known about what role perception and cognition shaped by spatial organization play in the dynamics of group decision-making processes. These and other related questions offer ripe opportunities for infusing behavioral considerations in developing spatially explicit MCDM methods.

Big data present both a challenge and an opportunity for MCDM. In the geographic domain, large multi-attribute datasets are made larger by incorporating spatial and temporal dimensions. Optimizations—techniques developed in MODM often fail when applied to large spatial or spatiotemporal decision problems as exact optimal solutions cannot be found. Evolutionary multi-objective optimization methods have been shown to be effective at approximating exact solutions and there has been an increased interest in the use of EMO to solve large, computationally challenging spatial problems. As EMO methods typically generate a large number of discrete solutions approximating a continuous set of nondominated solutions, there is a need to aid the decision-maker in efficiently evaluating tradeoffs between different discrete nondominated solutions. This calls for development of new techniques facilitating visual exploration and assessment of tradeoffs between alternative solutions. Opportunities for research in this area can be found at the intersection of visual analytics, MCDM, and GIS (Andrienko et al. 2007).

SEE ALSO: Spatial decision-support system; Spatial optimization

References


Further reading


Multiculturalism

Katharyne Mitchell
University of Washington, USA

Multiculturalism is often used as an umbrella term to capture a wide range of ideas and policies associated with human difference – particularly cultural and religious diversity. The concept has a number of associated referents: to the lived realities associated with plural societies; as a set of perspectives and debates in political philosophy about the best management of group difference; and as a form of state-sanctioned politics of incorporation and accommodation. In all three of its formulations it has defenders and detractors spanning the political spectrum.

As a reality of everyday life, multiculturalism reflects the mixing and unmixing of populations throughout history. Few areas on the planet remain untouched by the movements of people, who bring diverse ways of thinking and being into new contexts. Plurality is thus a common feature of most societies and multiculturalization can be seen, in this light, as a de facto process that has occurred and will continue to occur regardless of the ideologies and policies of theorists, states, and citizens.

Nevertheless, multiculturalism is also a feature of everyday life that most individuals associate primarily with late modernity and the movements and flows of contemporary globalization. The intensive labor migration networks from North Africa and the Middle East to western Europe following World War II are often considered to be the primary galvanizing force of this period of accelerated mixing. Likewise, migration networks from less developed to more developed regions in North America and Asia are also part of the contemporary imaginary associated with the rise of a vigorously multicultural period. Multiculturalism is also connected to the rights claims of disadvantaged groups, who effectively expressed their demands in a series of identity-based social movements beginning primarily in the 1960s.

Thus, although in its broadest usage the term could be applicable to most urban centers, ports, and long-distance trade routes across the globe and throughout human history, it is most frequently defined in relation to the concentrated migration and associated cultural intermixing of the postwar era. It is also associated with the social movements, minority nations, and indigenous and other group-based claims of the civil rights era. Some variation of the term, consequently, remains closely identified with the state policies and politics of Canada, Australia, the United Kingdom, the United States, Singapore, Malaysia, and many countries of western Europe as they negotiated – and continue to negotiate – the social transformations of this period.

Multiculturalism as political philosophy

With respect to the effective management of group difference, multiculturalism has many proponents in political philosophy. These include a number of different positions and justifications, but all share the foundational proposition that disadvantaged groups who have been marginalized by the dominant culture have the right to be recognized and respected by the state; moreover, the patterns of representation and recognition
Multiculturalism

that have produced economic and political disadvantages for minority groups should be remedied.

Beyond this desire for recognition and remediation there are nearly as many different philosophical viewpoints and propositions as there are different interpretations of the meaning of multiculturalism. As a starting point the term is understood to refer to identity and to cultural difference and hence it is associated with forms of politics such as “the politics of recognition,” and “identity politics.” But culture itself is an infamously slippery term, variously encompassing religion, language, aesthetics, food, music, way of life, and/or the whole range of learned human behavior. In association with the word multiculturalism it is also often connected to race and nation. Thus when brought into the realm of political philosophy and the question of how best to negotiate diversity, there are many differences of judgment and outlook.

The varied positions in this body of thought can be best summarized in three major categories: liberal egalitarianism, communitarianism, and postcolonialism. In liberal egalitarianism, culture is perceived as important for the life chances of individuals; it brings feelings of positive self-worth and self-respect, but only if the culture identified with is given positive recognition by other entities, such as the state. Culture is also important for individuals because it can provide a kind of scaffolding that enables choices to be made and learning to occur within certain parameters of meaning; it thus brings a sense of autonomy and self-reliance necessary for negotiating one’s path through life.

Theorists such as Kymlicka (1995) argue that given the instrumental importance of culture for an individual’s life chances, all cultures must be protected. Without special protections afforded by majority cultures (and their representative bodies such as the state), individuals associated with minority group cultures will be irreparably harmed. Thus it is up to the state to provide the accommodations necessary for equal valorization of minority cultures, and citizens must help to bear the costs of this accommodation.

Kymlicka’s philosophy of liberal egalitarianism reflects the author’s professional experiences and particular interest in Canada, which has a significant First Nations (indigenous) population, two colonial powers, and extensive immigration. Along with other liberal multiculturalists, Kymlicka’s normative argument justifies a hierarchy of differential rights and expectations of accommodation to groups based on a perception of coercive or noncoercive incorporation into the state. From this perspective, Canadian immigrants, seen as voluntarily joining the majority culture, would expect less in the form of group-differentiated rights than First Nations groups, who were incorporated against their will.

Another justification for multiculturalism comes from the communitarian position. This political philosophy embraces the value of social goods and communal life over the liberal emphasis on individual rights and liberties. Because it recognizes all cultures as collective projects with equal social value, multiculturalism is upheld as a positive normative ideal for the nation-state.

Operating in a different register, a third set of arguments justifying multicultural politics and policy brings in the question of history and power. These claims can be broadly situated under the rubric of postcolonialism. Drawing on a much wider body of social theory, and critical of the “neutral” stance of liberal arguments, postcolonial claims in support of multiculturalism point to the long-term negative impacts of colonialism. In this view, dispossession and dislocation caused by the colonial encounter led to the fragmentation or loss of cultural
artifacts and practices for aboriginal peoples worldwide; additionally, colonialism spurred the forced migration of millions of often unskilled workers and peasants, who lost their land and livelihoods at home, and who had little recourse but to move and sell their labor in the colonial metropole.

According to the postcolonial position, this history of violence and subjugation both delegitimizes the state’s authority over the indigenous, and also upholds their demands for redress; likewise, it positions immigrants and immigrant cultures in Western nations as minority cultures deserving of validation and accommodation. As the well-known phrase “we are here because you were there” indicates, the relationship between colonialism and migration is profound, and demarcations of coerced versus uncoerced incorporation into the state are frequently problematic. Thus liberal egalitarian arguments about the hierarchies of multicultural rights claims, in which immigrants are accorded relatively low status because of the “voluntary” nature of their state incorporation, are eschewed in most postcolonial critiques. Rather, overarching protections for diverse groups that have experienced structural violence and seek redress are propounded, along with insistence on examining the limitations of liberal thought and practices more generally.

State-sponsored multiculturalism and backlash

Multiculturalism has a range of meanings in different historical and geographical contexts. Depending on the philosophies and the practices of different states, it can refer primarily to a particular mode of immigrant incorporation and/or to the rights of minority groups to state recognition and protection. Multiculturalism is one feature of a broader trend in liberal social thought emerging in the mid-1960s in which cultural difference was acknowledged and valued and the national imperative for “outsiders” to assimilate to a national ideal was greatly reduced. This “differentialist” turn, as some scholars have labeled it, resonated with broader epistemological critiques of foundationalism and universalism.

Perhaps the most important impact of multicultural ideas at this time took place vis-à-vis perceptions of effective immigrant and minority integration. From a strong assimilationist ideology of immigrant absorption that was prevalent in most nations until the late 1960s, multicultural ideals of the right to difference (and to an institutional recognition of that difference) led to perceptible shifts in both the rhetoric and policy of numerous state institutions associated with minority rights and immigrant acculturation.

Through struggles in political spaces and in institutions like schools, proponents of the right to a public recognition and affirmation of difference won numerous legislative and budgetary victories beginning in the 1970s. These included funding for the teaching of regional and immigrant languages in the United States and Europe, the recognition of minority rights in the charters or constitutions of several nations worldwide, the recognition and affirmation of indigenous cultural groups, and the introduction of educational curricula supportive of multicultural principles.

Canada, in particular, was one of the strongest early adopters of multicultural language, policy, and practices, advocating strongly for bilingualism, and making multiculturalism an official national policy beginning in 1971. Because of its leading position in multicultural debates, numerous geographers such as David Ley and Audrey Kobayashi have investigated the discourse and practices of multiculturalism in Canada through
MULTICULTURALISM
time. Ley (2007) suggested recently that one of the critical scales in which to investigate the successes and failures of multiculturalism is that of the city, where the encounter with difference is highest, and where political fears about segregation and political alienation remain strong. Despite recent pessimism and unease about the ongoing value of multiculturalism, he provides a defense based on the importance of immigrant gateway cities in Canada and elsewhere, arguing that a robust philosophy for engaging difference of this order is critical for defeating racism and supporting culturally plural urban environments such as these (see also Keith 2005).

Multicultural victories at the national scale, such as those in Canada and Australia, were often the result of strong and persistent demands in formal political venues, but they also reflected the micropolitics of the workplace, urban sites, public venues, social media and the arts, and the encounters of everyday life. Geographers have probed these formal and informal spaces with an eye toward elucidating the importance of sociospatial interactions in the encounter with and acceptance of difference. In recent work, Amin (2012) defends a politics of encounter that exceeds human engagement, arguing for the importance of material transactional exchanges that include the nonhuman: for example, objects, physical environments, affect, atmosphere, and other material entities. He believes it is these types of situated practices and exchanges, where many kinds of tangible and intangible things flow together to create the site and all-important framing of the encounter, that enable the achievement of diversity in the contemporary world.

In recent years the acceptance of state-sponsored multicultural programs and practices has declined, and both the discourse and the practices related to minority integration are currently undergoing a process of profound change. Although uneven, the immigration and integration policies of many European nations reflect a general movement away from an official recognition and facilitation of pluralism in the public sphere, alongside a renewed affirmation of the core values of laissez-faire liberalism.

In the United States, Australia, and several European countries, an active state-sponsored effort to achieve diversity has been largely abandoned, and the promotion of assimilation – in the sense of a renewed defense of a sharp separation of public and private realms – is increasingly apparent (Joppke and Morawska 2003). Although these transformations play out differently in different national settings, much of the state-based rhetoric now emphasizes the choice of individual immigrants and minorities to assimilate to the values of liberalism. If they choose not to do so, it is no longer considered the state’s responsibility to protect or accommodate them. Meanwhile, many of the state-provided social services that had formerly aided integration have been removed, subcontracted, or devolved to the community level.

Recent examples of these types of transformations include Denmark and the Netherlands, both of which have experienced stark shifts away from a strong official policy of multiculturalism over the past decade and a half. After many years of promoting the value of immigrant difference and of offering state recognition and providing numerous state safeguards for ethnic pluralism in the public sphere, the discourse and policy on minorities now favors a far more assimilationist tone and set of policies. Similarly, under the French presidential administration of Nicolas Sarkozy from 2007–2012, a sharp stance against publicly manifested difference was also intensified (Vertovec and Wessendorf 2010).

Perhaps the most dramatic and aggressive shifts in state policy have involved the early
twenty-first-century conservative governments of David Cameron in the United Kingdom, and Angela Merkel in Germany. Both of these political leaders gave speeches in the late 2000s indicating a crisis and failure of multiculturalism in their respective nations; Cameron went so far as to advocate the end of “state multiculturalism.” Their rhetoric implied that despite the benevolence of European nations toward postwar immigrants the multicultural experiment had failed and minorities remained a “problem” vis-à-vis national unity and shared values; part of this problem was seen to reside in the state’s effort to accommodate these outsiders by encouraging them to preserve their own cultural values and way of life.

Although a kind of de facto or noninterventionist multiculturalism remains in place in most liberal states, the pressure toward national languages and the acceptance of the core values of laissez-faire liberalism reflects a new discursive frontier that may well be the precursor to ever-harsher assimilative policies and practices. The discourse of multicultural failure and the renewed state policies of assimilation represent a return to a recuperative national project that seeks to relocate universalist notions of civil society firmly within the bounded contours of the nation-state. The economic and political component of this shift is rarely discussed yet critically important.

The current discourse of multiculturalism in crisis is intertwined with ongoing political and economic efforts to shift the responsibility for controlling, facilitating, and paying for immigrant integration from the institutions of state government to the local level. A generalized state rhetoric of multicultural failure and the necessity to rethink the advantages of national assimilation accompanies the devolution of responsibility for ethnic integration to the scale of the community and the individual. This is but one of many new technologies of knowledge/power under neoliberal regimes of governmentality, in which individuals are constituted as atomized, free-thinking, and entrepreneurial subjects who can “choose” to assimilate or not as they wish. Those who choose not to assimilate are represented as individuals unwilling to participate in public or civic life, who can, as a result, be excluded from society without incurring damage to the core ideals of a universalist liberal project.

A second political component of the contemporary backlash involves the jockeying of politicians for a position that allows them to address popular anxieties related to neoliberal globalization and immigration. Using multiculturalism as the scapegoat, Cameron, Merkel, and other political leaders can appear to confront the “problems” of social and economic dislocation and change, while at the same time absorbing and nullifying some of the even shriller claims of the far right. For example, following Cameron’s speech, Marine Le Pen, the leader of the French far right party Front National, congratulated him for his “endorsement” of her party’s anti-immigrant position, indicating their strong overlap on this issue.

Scholars critical of this sharp turn have noted how the state rhetoric of multiculturalism as failing provides an alibi for renewed forms of racism. Culture as “problem” inevitably references those groups and individuals outside of the dominant culture, which is assumed to be a neutral, universalist space. Consequently, the identification of an unassimilable culture – often represented as Muslim immigrants in the context of liberal Western nations – gives license to the practice of racial discrimination in the name of the nation (Lentin and Titley 2011).

The discourse of the unassimilable “other” is part of a larger right-wing critique of multiculturalism that rests on the idea of a “clash of civilizations.” In this view, the primary source of post-Cold War conflict is the inherent and
MULTICULTURALISM

insurmountable differences in people’s cultural and religious identifications. By this reckoning, multiculturalism will always be a failed project because certain kinds of integration are impossible, especially between those holding “liberal” and “illiberal” viewpoints and values.

Critiques from the left are also quite varied. In the current period, many scholars who have been critical of the concept for its theoretical and practical deficiencies have felt compelled to defend it in the face of contemporary conservative attacks and renewed racisms. Nevertheless, one of the primary poststructuralist critiques rests on the problematic notion of culture as hermeneutically sealed, as a contained entity that groups maintain over time despite all of the intervening influences of a contemporary world of constant movement, flow, and interaction. In this view there is no pure or autonomous culture, and to pretend that there is and try to protect it relies on a false view of both culture and individuals’ relationships to it. Moreover, group-based state protection of this kind risks harm to vulnerable members within minority groups. These may include women, sexual minorities, religious dissenters, and so on, who do not necessarily share the opinions or the embodied characteristics of the dominant members of the group, and who may experience internal discrimination from the group’s leaders and representatives.

Another critique challenges the prioritization of identity politics in general; this view holds that the politics of redistribution, with its stronger economic emphasis, is more critical than the politics of recognition and should be supported more vigorously. Relatedly, state-sponsored multiculturalism in its actual practices is often quite superficial – obsessing about aesthetic issues such as food and clothing rather than addressing the deeper issues of economic and political integration into society. Additionally, numerous scholars have observed that multiculturalism as a term and set of practices is often appropriated by capital to further its own interests (Mitchell 1993). Finally, many have argued that the real problem vis-à-vis the question of difference in most Western nations is racism, and that multiculturalism – as a political philosophy and as state-sponsored policy – does not sufficiently address the underlying issues of power and structural violence that allow racism to endure and flourish.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Globalization; Migration: international; Nation-state; Race and racism

References

Further reading


Multilevel governance

Andrew Beer
University of South Australia Business School, Australia

Multilevel governance is a wide-ranging theory about the ways in which governments act in contemporary society which emphasizes both the vertical exercise of authority and the horizontal sharing of power. It is an interpretation of government systems that foregrounds the social and political context within which power and influence are exercised, and which highlights the processes of dialogue, debate, and knowledge-sharing with other tiers of government and nongovernment actors. From its foundation with the initial work of Marks (1992) and the subsequent development of the concept (Marks 1993, 1996; Marks and Hooghe 2000, 2004) there has been increasing use by academics of the concept of multilevel governance to explain the decision-making processes of governments, especially in complex policy environments. Some researchers have suggested that the term has achieved almost universal acceptance within both policy-making and academic circles, and that the rate and pace of take-up of this term is virtually without precedent in the field of European studies (Stephenson 2013). Potluka and Liddle (2014) note that multilevel government remains a contested concept, but has a broad appeal and a shared concern with the increasing intricacy associated with government decision-making.

Multilevel governance recognizes that governments increasingly make binding decisions as a result of processes that include other actors, such as discussions with other nations or the influence of nongovernment agencies. As Piattoni (2010) has acknowledged, the concept emerged out of the attempt to uncover the processes shaping policy development in the European Union (EU), and while the EU represents a unique policy and governance environment, there has been sufficient resonance with other contexts to encourage the adoption of this concept elsewhere. Work in this area commenced with Marks (1992), who concluded that whereas the theory of the time suggested policies emerged from an orderly set of processes dictated by the nation-states of the EU, empirical observation suggested policy formation was more chaotic, and subject to the influence of a far broader range of actors. Researchers noted that policymaking in Brussels – the center of EU administration and deliberations – was not restricted to the staff of the European Commission, the administrative arm of the EU, and the representatives of the participating nations in the European Council and European Parliament. Instead, nongovernment organizations, provincial governments, and industries maintained a presence in Brussels and worked to shape the outcomes of policy deliberations. The idea of multilevel governance – with a constellation of government and nongovernment agents interacting to shape agendas and influence outcomes – was established to better explain how policies emerged and how governments took decisions. It acknowledged the pluralistic nature of decision-making in an environment marked by both complex issues and opaque institutional structures.

Multilevel governance as an observable phenomenon is premised on the idea of governance
MULTILEVEL GOVERNANCE

as a new model of government decision-making evident in many developed economies. Governance, therefore, is the precondition for multilevel governance, and lies at the heart of a substantial body of contemporary scholarship that addresses the role of governments in social and economic policy. It has been especially important in work on the development of urban and regional policy, one of the key policy domains of the EU (EU 2010). Despite its significance, governance can be a difficult concept to define categorically: Jordan, Wurzel, and Zito (2005, 478) observed that “there is no universally accepted definition of governance; there is not even a consensus on which set of phenomena can be properly grouped under the title of governance.” While there is merit to this argument, there is now a substantial body of work that acknowledges a number of common elements that have been accepted as typifying governance regimes and their outcomes. These include the movement away from restricting decision-making to the formal structures of government and instead incorporating a wider range of interests in decision-making (Whitehead 2003). In his working definition, Stephenson (2013) highlighted governance as system of power sharing both within governments and with nongovernment actors, including the private sector. Commonly, governance is associated with the rise of partnership arrangements and the reduced ability of governments to mandate or directly determine outcomes.

Governance takes different forms in different nations, with Blatter (2004) noting that there are significant differences between unitary systems of government (such as those in the United Kingdom, New Zealand, and Finland) and federal systems of government (such as those found in Australia, Canada, Malaysia, Germany, and the United States). Governance in federal systems, it has been argued, is marked both by horizontal links between agents and institutions, and by hierarchical, competitive, and cooperative modes of interaction. Governance, therefore, can lead to complex forms of interaction within federations, with both positive and negative relationships possible.

It has been argued that the adoption of governance is commonly associated with the rise of neoliberal policies and philosophies of government (Geddes 2005), which have effectively disempowered governments and forced them to find new ways of achieving societal aims, often in association with the private sector or community associations.

Explanations based on multilevel governance are differentiated from the broader literature on governance on a number of key dimensions. Perhaps most importantly, multilevel governance postulates that changes in the relationship between government, on the one hand, and society and the economy, on the other, occur as a result of shifts in policy domains, both internal and external to the nation-state. That is, political movements and strategic positioning by subnational governments is as likely to find expression in the international domain as domestically. Marks’s (1993) original conception highlighted the role of subnational governments in shaping decisions, and the fact that multilevel governance could be thought of as “a system of continuous negotiation amongst nested governments at several territorial tiers” (392). One of the arguments presented in support of multilevel governance is that it is more flexible, and therefore more adaptable, than the concentration of power in a single entity.

Hooghe and Marks (2003) suggested that multilevel governance could be divided into two forms. Type I was characterized as being constituted by general-purpose jurisdictions, with nonintersecting memberships, a limited number of levels, and a system-wide architecture
or remit. As they evocatively noted, under Type I governance “every citizen is located in a Russian Doll set of nested jurisdictions, where there is one and only one relevant jurisdiction at any particular scale” (Hooghe and Marks 2003, 236). Type II multilevel governance, by contrast, is characterized by task-specific jurisdictions, intersecting memberships, the absence of a limit on the number of jurisdictional levels, and flexible design. Examples of the latter could include a particular resource problem, such as the management of a catchment or the negotiation of trade treaties. Type I and Type II forms of multilevel governance are both seen to be departures from the centralizing state, but in different ways and with sharply differing outcomes. Inevitably, multilevel governance is seen to carry with it both costs and benefits. On the positive side, it can be flexible about the scale of action – with higher-level agencies responding to nationwide challenges or difficulties, while lower-level institutions respond to questions evident at the local scale. Such flexibility, however, comes at a cost, as there are considerable potential (and realized) coordination costs arising from the need to consult, deliberate, and potentially negotiate before acting.

Marks and Hooghe (2004) conceptualized multilevel governance as differentiated by processes as well as by form. They saw Type I and Type II forms of multilevel governance as being generated through separate, but related, processes: Type I processes were seen to be focused on the negotiation of powers and responsibilities at different levels of government, and Type II processes were an outcome of a concern with the interconnections between agencies and governments at the same level. These horizontal linkages were accepted as important in delivering coordination, building political consensus, and establishing a discourse for a particular course of action. Further work has broadened the original conception of multilevel governance to include a more expansive range of agencies and actors. These more embracing perspectives recognize that nongovernment organizations seeking to advance the rights of Indigenous Australians, for example, are as likely to argue their case in international forums – the United Nations (UN) in New York or at the Organisation for Economic Co-operation and Development (OECD) in Paris – as domestically. Campaigns for greater environmental or labor market protections are often argued in world forums. Multilevel governance is differentiated from comparable work on networks because, while the latter are intrinsically ephemeral, research on multilevel governance acknowledges the enduring nature of many institutions and their arrangements. Some researchers have suggested that economic policymaking – especially in federal systems – should be considered part of a single multilevel governance system (Stephenson 2013). To a degree, this idea is intuitively attractive, as economic policies inevitably involve multiple stakeholders, international negotiation and arrangements, and near-permanent processes of interaction and engagement. While the World Trade Organization (WTO) is an outcome of global engagement with the liberalization of trade, new initiatives in this field come about through the ongoing work of governments in bilateral and multilateral agreements. Industry pressure groups also play a pivotal role in advocating for change.

Multilevel governance as an explanatory framework has been applied to environmental questions, including the development of strategies for mitigating, and adapting to, climate change. Its application to this policy domain could almost be considered inevitable, given that climate change adaptation and mitigation strategies inevitably involve a multitude of participants, including national governments, provincial governments, industry bodies, local governments,
MULTILEVEL GOVERNANCE

community groups, and international forums. Bates et al. (2013) noted that the processes that typify multilevel governance – participation in forums, networking and lobbying across and between tiers of government, and so on – help mobilize businesses and institutions to environmental action. Such activities simultaneously add “trusted knowledge” and reduce the perception of risk. Multilevel governance also plays a more instrumental role in adaptation to climate change, as an effective response can only be achieved if all tiers of government – and many sectors of the economy – take action.

Bulkeley and Betsill (2005, 2013) commented on the importance of interrelations between global, national, and local actors in formulating responses to climate change. They argued these relationships have had substantial impacts at the urban scale, helping to reshape planning discourse and practice. Early academic work and policy development on cities and climate change was dominated by a “new localist” perspective that highlighted community-level action independent of context. To a degree, this outlook was typified by the Agenda 21 slogan, “think global, act local.” The researchers noted that this analytical framework did not adequately explain why some cities and communities took no effective action on climate change, while others led the way. Multilevel governance, by contrast, was seen to offer an effective explanatory framework. Bulkeley and Betsill (2005, 138) concluded from their analysis of urban planning in Newcastle upon Tyne and Cambridgeshire, in the United Kingdom, that in seeking to understand shifts to more sustainable development, “the most significant dynamics exceeded any purely local framing.” Reviewing their own work almost a decade later, Bulkeley and Betsill (2013) concluded that a multilevel governance perspective remained important in understanding the management of climate change in cities, but that increased attention needed to be paid to how these processes resulted in governance innovation across, and within, tiers of government, which in turn can usher in new policy outcomes.

Multilevel governance perspectives are deeply embedded in many accounts of local economic development initiatives and the agencies that support them (Eversole and Martin 2005). Cross-national work has consistently identified governance, and especially multilevel governance, arrangements as a defining feature of the structures established to promote local or regional economic development (Beer, Haughton, and Maude 2003; Halkier, Danson, and Damborg 1998). A number of authors have noted that many provincial, city, and local governments have embraced a multilevel perspective in attempting to promote their economic and other interests on the global stage. Multilevel governance in economic development sees national, local, and community scales of engagement in operation at the same time, and this simultaneous activity is inevitable. Importantly, economic development frameworks that ignore the other actors operating within this environment are likely to be less effective than those that encourage positive engagement vertically and horizontally.

Academic acceptance of multilevel governance has contributed to its adoption amongst policymakers, especially in the EU, where the distribution of significant monies in the form of structural funds to identified territories has encouraged cities and regions to establish a presence in Brussels. It has also lent a distinctive regional flavor to much of the discussion of multilevel governance. Stephenson (2013) noted that the EU itself explicitly acknowledged multilevel governance in its attempt to introduce the open method of coordination (OMC) – a system intended to simultaneously deliver both vertical
MULTILEVEL GOVERNANCE

and horizontal integration. And while OMC was not successful, its very failure has served to strengthen the conceptual and programmatic impact of multilevel governance as an efficient, democratic, and accountable set of political structures.

Other research has noted the usefulness of the concept of multilevel governance in dealing with economic shocks, especially the global financial crisis (GFC) of 2008 and 2009 (Allain-Dupré 2011). This research highlighted the fact that the policies of fiscal austerity that followed the immediate stimulus packages needed to be mindful of the ways in which reductions in national public sector outlays were transferred to state or provincial governments, and potentially undermined their capacity for future growth. This body of work acknowledged that there is a risk that budget cutbacks introduced at the national level have the potential to cascade through the more junior tiers of government. This in turn results in greater than expected reductions in public sector outlays, thereby undermining strategies to raise productivity over a range of time horizons. Subnational governments, it has been argued, play a leading role in the infrastructure investment needed to foster growth, and are best placed to target assistance to the areas of greatest impact. National economic policies that ignore multilevel governance and fail to acknowledge the significance of the more junior tiers of government may find their actions undermined in the short term, resulting in restricted economic growth in the longer term. Fortunately, these authors have identified a range of policies and practices mindful of multilevel governance that can help overcome these challenges. These include co-investment strategies between national and subnational governments, the use of existing institutions that embrace multilevel governance, and the creation of new entities that can overcome the challenges inherent in addressing economic crisis within a multitiered system of government. In 2008, the EU explicitly deployed its cohesion policy as part of its strategy to overcome economic crisis. Its actions included the modification of existing programs, the closure of some program areas, the introduction of a renewed commitment to the growth of regions, accelerating the cash flow associated with new projects, and the development of new methods for identifying the best possible investments in regional productivity.

More generally, the OECD (2009) has recognized seven important gaps in the multilevel governance of public policy, each with implications for the effective management of their economies (Allain-Dupré 2011, 21). These gaps are: an information gap associated with inevitable information asymmetries as subnational governments are not, and cannot be, party to all information available to national governments; an administrative gap resulting from a mismatch between the functional areas associated with an issue – for example, a functional labor market or a catchment – and administrative boundaries; a policy gap arising from the fragmentation of responsibilities across ministries; a capacity gap associated with insufficient technical or other expertise at the local level to meet needs; a funding gap that emerges from unstable or insufficient revenues at the subnational level; an objective gap as competing priorities direct national and subnational governments in contrary directions; and an accountability gap resulting from inadequate transparency. Often it is not clear which tier of government or government agency has brought about adverse outcomes or reduced the opportunities for growth.

Piattoni (2010) argued that multilevel governance has made important contributions to our understanding of government processes and the reshaping of cities and regions in three fundamental ways. First, she argues that it has
MULTILEVEL GOVERNANCE

encouraged researchers to think about the ways political mobilization takes place, both within the boundaries of individual nations and across boundaries. Second, she contends that “policy-making no longer neatly separates policy-makers from policy-takers, nor does it distinguish between public and private actors” (2010, 250). Third, she believes that multilevel governance results in political structures – polity – that look like nation-states but are not. The EU, she notes, is one example of a state-like entity, but it is not the only such institution. The actions and decisions of the UN, the World Bank, the OECD, the WTO (Knodt 2004), the Asia Development Bank, and even the International Olympic Organising Committee (IOOC), to name just a few examples, all contain elements of multilevel governance, including the influence of a host of interested third parties, some of whom are nation-states.

Some of the most important insights to come from the multilevel governance literature relate to the simultaneous interaction between provincial-level governments and nongovernment organizations, on the one hand, and supranational bodies, on the other. Piattoni (2010) suggests that the eagerness of state or provincial-level governments to become involved in matters above the level of the nation has the capacity to simultaneously unsettle the nation-state, while reinforcing the significance of supranational entities as a source of funding, legitimacy on the world stage, and/or political influence. Marks, Hooghe, and Blank (1996) label this process as one of “dilution,” with nation-states surrendering their capacity to influence the representation of domestic interests in international forums. Benz and Eberlein (1999) take a different perspective on the relationship between the nation-state and multilevel governance: they argue that national institutional structures both guide and limit the impact of multilevel governance on individual territories. They also discuss the significant costs associated with multilevel governance, to governments, nongovernment actors who participate in policy debate, industry, and the broader community. Perhaps most importantly, Piattoni (2010) argues that multilevel governance generates, and then maintains, fluid relations between a host of actors – nation-states, subnational governments, private sector representatives, and nongovernmental agencies – that remain in a state of flux.

Policy debates emerge, gather strength, and are then replaced by other discourses as new priorities emerge and new arguments for intervention garner support. Along the way, this process of policy formulation has real-world impacts, as programs of regulation (trade protocols and tariffs), pricing (emissions trading schemes to control the release of fossil carbon), and regional development (EU territorial development programs) are introduced, reshaped, or abolished.

Multilevel governance is not simply an unsettling influence on established policy frameworks and programs. Potluka and Liddle (2014), working in the Czech Republic, highlighted the benefits associated with the delivery of EU programs within a multilevel governance framework. They noted that the interactions that characterize multilevel governance resulted in better-tailored programs, able to achieve at the local scale policy objectives established by distant central authorities. They also generated an environment that was more open to participation in program delivery by a wide array of actors, including community-based organizations. Other researchers working in very different contexts have also identified benefits arising from multilevel governance. Homsy and Warner (2014), for example, used quantitative analysis to examine the take-up of sustainability actions by municipalities across the United States. They found that local governments that operated within a
Multilevel governance framework were much more likely to support sustainability initiatives. State governments across the United States played a clear role in promoting discussion of environmental protection, resulting in a “broad and co-ordinated discussion, which is important because most environmental issues cross local political boundaries” (Homsy and Warner 2014, 19). State governments also served as a reservoir of technical expertise, thereby adding to the stock of local social and political capital embedded within local governments. Active engagement with environmental sustainability was not simply a product of government actions; civil society institutions were also important for creating an environment supportive of environmental protection. Such engagement typifies systems of multilevel governance, and supports the extension of this concept beyond its European origins.

Multilevel governance has been criticized as an explanatory framework by a number of authors. Its popularity as an intellectual device and its widespread adoption by academics and the policy bodies of governments alike has, in some ways, undermined the utility of the concept. Some authors use the term to describe a wide range of circumstances that bear little resemblance to the original conceptualization. This includes labeling formal federal systems of government, such as the United States, Australia, or Malaysia, as systems of multilevel governance without acknowledging the constitutional basis for the division of power, or the fact that powers and responsibilities are formally prescribed, rather than an outcome of ongoing dialogue and debate. Some authors also appear to struggle to differentiate between multilevel governance and other types of associational activity. Potluka and Liddle (2014), for example, treat multilevel government, partnership formation, and network development as interchangeable entities. Such use of these terms would appear to dilute the impact of each.

Some authors have argued that, while multilevel governance provides a useful description of the interaction of agencies, it confuses involvement in the decision-making process with the real power to influence outcomes, and insufficiently analyzes differences in degrees of fiscal and decision-making autonomy, sometimes captured in the distinction between decentralization of administration and decentralization of choice. It does not predict which actors are likely to be important in shaping debates and the outcomes of consultations on an issue, nor does it provide insight into which types of agency are likely to be influential in the short, medium, or long term. Questions remain about the viability of multilevel governance as a concept outside the special circumstances of the EU. The EU allocates billions of euros through its multiyear structural funds and associated programs, and these substantial transfers – directed to regions, not nations – have generated their own political and economic dynamic. Sbragia (2010) notes that, in contrast to the EU, the North American Free Trade Agreement (NAFTA) contains strong measures to ensure a minimalist approach to implementation – effectively a guarantee against institution building. Most scholars acknowledge that multilevel governance as an explanatory framework has limitations, many of which have been exposed by the GFC and its aftermath. To a certain degree, multilevel governance is a concept that has focused inwards on the political dynamics of the EU and one of its redistributive functions. It is also a perspective intuitively attractive to scholars confronted by the reality of pluralist politics that are at odds with their structuralist intellectual traditions. Increasingly, it is accepted that there is a need for the discussion of multilevel governance to adopt a more global perspective, one which extends beyond the political structures and the territory of Europe. Others have suggested that multilevel governance remains a “statist”
concept, awarding considerable priority to the actions of governments – and especially central governments – while failing to pay attention to other drivers of change, including community action, small and medium-sized enterprises, global corporations, and political movements.

SEE ALSO: Governance and development; Neoliberalism; Regional political movements; Urban politics

References


Music

Michelle Duffy
Federation University of Australia

Geographical engagements with music and sound more broadly offer significant insights into the many and varied ways in which concepts of place, identity, and community are constituted. As many researchers have pointed out, music and sound should not be conceived as simply a background to place. Rather, they play a crucial role in the social relationships that help comprise a particular spatial context (Smith 2000). This is due to the ways in which certain acoustic properties are attributed particular and ideologically loaded meanings. For example, sound becomes music when placed within an appropriate context, be that the physical space of a concert hall or the marking out of a temporal frame as performance. Furthermore, as Jacques Attali argues in his seminal text, Noise: The Political Economy of Music (1992), the ways in which sound is defined as music or noise is an expression of power, one where the appropriation and control of sound reflects and affirms the sociopolitical structures of a community. Therefore music is not an aesthetic experience alone but a medium through which the social world is made and remade through various social relations.

Early twentieth-century scholars working in a variety of disciplines – including sociologists Georg Simmell and Alexander Ellis, and ethnomusicologists such as Alan Lomax – were interested in music for three main reasons: as an expression of meaning, as a basis for social differentiation through taste, and (perhaps the central reason) as a superorganic form of regional cultural traits or artifacts. Music has been understood to somehow reflect or represent a group of people, with the repertoire of different groups maintaining a highly patterned and stable sound profile that serves to identify the social group from which it originates. Geographers therefore sought to uncover the spatial diffusion patterns of these different social groups through differing musical styles and genres. Yet, a focus on music’s spatial distribution fails to recognize the significance of the social and cultural contexts in which music is produced and afforded meaning, nor does it consider the importance of interactions between local and global forms of music production for both creativity and industry. In addition, the framing of music as a cultural trait has led to a quest for authenticity and authentic practice. Yet authenticity is a highly problematic concept, and concern about authenticity masks the complex processes of music, as well as the ideological framings of tradition that serve to condemn change. Even so, as anthropologists acknowledge, there is some correlation between sound structures and social structures. As with language acquisition, we learn to recognize certain musical patterns or conventions as correct. Such conventions constitute a part of our social world and are the means by which we understand reality.

In their introduction to an edited collection of essays, The Place of Music (1998), geographers Andrew Leyshon, David Matless, and George Revill argue that space and place are not the sites in which music occurs or through which music diffuses but that place is formed and created through music and sound. This observation signals an important paradigm shift in human geographical research more broadly, a
MUSIC

poststructuralist turn that argues that there is no stable, fundamental link between music, identity, and place. Instead the interactions between composer/musician, listener, and the cultural context in which music is performed result in various complex sets of meanings that are constantly formulated and reinscribed. While recent research has moved away from attempting to categorize certain patterns of sounds as definitively representative of specific social groups, the use of music to create identity remains an important geographical interest. For example, some have examined how notions of identity and community are constructed through the deliberate associations attributed to music and lifestyle (Connell and Gibson 2003). Geographers have also examined the role of sound and music in constituting particular spatial entities such as the workplace, the city, the countryside, and everyday life (Bull 2000). Others have focused on the ways in which musical elements, such as song lyrics and language idioms, are strategically used to evoke images of particular places and to create a sense of connection to that place.

Music tourism, specifically festivals and events, demonstrate the intentional use of music to construct notions of identity and community through a focus on place-making and place-marketing practices. Much of the festival literature explores how these events sustain social benefits through the economic opportunities they generate, which arise out of deliberate representative practices that associate certain music with certain places. This area of research encompasses music in terms of an economic commodity, a cultural signifier, an expression of identity, a set of activities within the creative economy, and as a means to instigate political action. However, while a festival program tends to support notions of boundedness to particular places and people, the creation of community identity involves complex processes with potentially divisive outcomes. Some studies suggest that, rather than encouraging the formation of belonging and community, festivals often operate as spaces of exclusion.

Until recent geographic interest in performativity, embodiment, affect, and nonrepresentational theory, music geographies have most often focused on sound as an object, something seen rather than heard or felt. This emphasis on representation has meant an unintentional privileging of visual and textual discourses as research has examined such things as the score or lyrics, as well as associated paraphernalia such as instruments and promotional material, obscuring the ways in which the processes of music influence us and shape our actions in space and time. More recent research has started to shift. More recent research has started to shift this focus. One important way in which music operates is that it taps into our emotional and intuitive selves, opening up a means to examine how the emotions influence social interactions (Smith 2000). For example, Tia DeNora’s (2000) *Music in Everyday Life* examines how music in daily life demonstrates how the affordances of sound – that is, music’s melodic structure, rhythms, tone, and color – are an important means to organize our individual and collective lives both spatially and temporally. In addition, a focus on music as a process has meant closer examination of the characteristics of sound and how we process this information. It is not just the affordances of sound that make us who we are. Rather, where and how we hear and listen are integral to experiential practices of making sense of self in and through place.

Considerable empirical work has explored a phenomenology of sound, opening up new possibilities for what it means to listen. For example, some geographers have addressed the politics of music by drawing on a Foucauldian-type disciplining of listening. Yet, the type of listening
invoked in such studies relies on a simplified classification of the senses that limits auditory geographies to the ear, thus ignoring the significance of the whole body in sonic processes. Sound alters our perception of places and of others because it demonstrates the continuity and discontinuity of our subjectivity; we are immersed in its sounds and can either be drawn into feelings of connection or feel alienated. A multisensory approach that explores music in this way has drawn on the ideas of Deleuze and Guattari or that of nonrepresentational theory. Implicit in these approaches is the capacity of the body to sense sonic elements, such as rhythm, melody, and timbre, and how our affective and emotional responses to these then play a role in the emergence of bodily processes that can regulate the dynamics of social interaction (Duffy and Waitt 2011).

One of the challenges in investigating the experience of music is communication, as such experiences are ephemeral and difficult to express in words, a difficulty also recognized by musicologists. While interviews and participant observation can, to some extent, access how people may feel in response to music, geographers have begun to explore ways to capture nonverbal and noncognitive experiences. Some have begun to develop projects that bring together musical and geographical concerns in practice-based responses that facilitate what Derek McCormack (2008) calls a “co-intensive sensing” of emergent space. A deeper exploration of listening practices has therefore become important in recent geographical work. Such methodological and analytical approaches include rhythm analysis (based on the work of Lefebvre; see Edensor 2010), “participant sensing” (Wood, Duffy, and Smith 2007), sound diaries, phonography, and field recordings (Gallagher and Prior 2013), the extralinguistic elements of communication and the nonverbal components of performance (Kanngieser 2012), acoustic ecology and soundscape studies, and soundwalks.

SEE ALSO: Affect; Culture; Emotional geographies; Identity; Nonrepresentational theory; Phenomenology; Popular culture; Poststructuralism/poststructural geographies; Soundscape; Tourism

References

Edensor, Tim, ed. 2010. Geographies of Rhythm. Farnham, UK: Ashgate.


National and regional integration

John Agnew
University of California, Los Angeles, USA

National and regional integration involves the infrastructural and fiscal integration of polities at, respectively, national state and supranational scales. Progressively since the eighteenth century, but at a more rapid pace in the years since World War II, the world has been divided up into territories to which the term “nation-state” has usually been attached. Whether or not these states are true representatives of single nations or effective as governments remains an empirical question. Nevertheless, a central claim of governments of such states is that they have domestic sovereignty or the combined desire and capacity to authoritatively provide public goods and serve the material and symbolic interests of their populations, irrespective of where they live within the national territory. Spatial disparities in incomes and public good provision, therefore, can provide the basis for the politicizing of local cultural and political differences that are otherwise repressed. Of course, whether or not they are politicized depends on the emergence of political movements to mobilize local populations for either greater autonomy in fiscal policy or outright independence (Agnew 2013).

State formation and national integration

The worldwide rise of the nation-state is one of the most significant topics in social science. Such states have come about in numerous ways: nationalist secession (as in Mexico and Yugoslavia), revolution (as in the United States and Russia), transition from clan and feudal societies (as in Sweden and Thailand), unification movements (as in Germany and Italy), and top-down reform (as in Japan). Rather than global forces, world regional imitation and diffusion among neighboring nationalist elites drawing on the example of already established prototypes (France, the United States, etc.) as existing imperial centers and ancien régimes disintegrate seem to have been the dominant governing processes driving the spread of the nation-state since the early nineteenth century (Duara 2008; Wimmer and Feinstein 2010). The unmaking of empires is a fraught business, with some fault lines emerging relatively painlessly and other divisions, for example Russia’s recent wars in the North Caucasus and the conflicts in the former Yugoslavia, revealing historic fractures not readily healed. The stability and sustainability of specific states over time seems to have depended as much on their location relative to potential expansionist neighbors as on either their internal ethnic homogeneity or the existence of some putative ethnicity-spanning peoplehood (such as “British” or “Chinese”) and on the extent to which they have had external supporters. So-called buffer states, in particular, ones located between larger, pre-existing, and more stable states, are most vulnerable to annexation or partition by neighboring rivals (Fazal 2007).

Be this as it may, both the persistence of some nation-states and the revival of others after periods of demise is invariably the result of nationalist movements that thereafter use the typically contested claims to the precise territories
they occupy to constantly keep the “national question” on the political agenda. Within states, however, such nationalist projects frequently encounter resistance from localities and subnational regions that either favor some looser form of political organization than a centralized state or that see the new state as compromising the historic freedoms they have exercised as city-states, as provinces within decentralized empires, or as polities without much bureaucratic apparatus at all (see, respectively, e.g., Deudney 1995; Tilly and Blockmans 1994; Scott 2009). Consequently, a major challenge for states is often to ensure that state resources and infrastructure are distributed reasonably equitably across the national territory, for fear that the pooling up of advantages in some localities or subnational regions can spur efforts at secession or irredentism on the part of indigenous or neighboring nationalist movements.

“Sovereignty regimes” and the geography of state capacity

Once initially established, the legitimacy of centralized power is thus hardly unquestioned. In distinguishing despotic from infrastructural power, Michael Mann (1984) usefully identifies two different ways in which a state acquires and uses centralized power. These words refer to two different functions that states perform for populations and that jointly underpin their claim to authoritative rule: respectively, the struggle among nationalist elites within and between states, and the provision of public goods by states in the interest of placating various social groups and pursuing and legitimizing despotic power. In Mann’s (1984, 188) words:

Let us clearly distinguish these two types of state power. The first sense denotes power by the state elite itself over civil society. The second denotes the power of the state to penetrate and centrally co-ordinate the activities of civil society through its own infrastructure.

Before the eighteenth century the second type, infrastructural power, was relatively less important than it is today. This is because elites have been forced by political struggles to be more responsive to their populations through providing more public goods. But elite and popular pressures for economic development have also mandated the increased provision of roads, of standardized weights and measures, of elementary education, and so on, which will always be absent or underprovided with a simple reliance on private provision. This boosted the territorial sovereignty of states because economic demand was defined in terms of territorial populations and provision was oriented to satisfying that demand.

Since the mid-1970s, technological change, the increased intensity of all kinds of flows across borders, and the vulnerability of state populations to increased global economic competition threatening public good provision when foreign competitors provide fewer public goods have all conspired to make infrastructural power increasingly valuable for economic competition but also vulnerable to the fiscal crises of states as they struggle to attract global capital, demanding lower effective taxes (Frieden 2006). Increasingly, city-regions and supranational entities (such as the European Union) also challenge the state monopoly over public goods and international organizations, both public and private, and demonstrate the capacity to deliver regulatory, financial, and legal services hitherto usually associated with the bureaucracies of states.

At much the same time, despotic power has come to rely much more on popular legitimacy than it did historically. At one time, and this is at the root of Western territorialized conceptions of sovereignty, the ruler relied on the
claim to stand in apostolic succession to God, as in the divine right of kings. This eventually translated into the idea of the “body politic” as a territorialized people or nation. Rulers needed to establish at least a modicum of popular authority before they could pursue their goals. Such legitimacy as they have, however, is increasingly fragile. As interests and identities cease to conform to territorial norms, rulers must likewise adjust. This can involve pursuing increased influence elsewhere (as in an *imperium*) or ceding authority to other parties in order to manage dissent and resistance. There is no necessary correlation, therefore, between despotic power and central state authority. Elites are increasingly globalized both with respect to pursuing their goals through expanded political influence beyond home shores and in terms of alliances with multinational companies, banks, and other agents in a more networked world.

The degree of real state autonomy from outside influence or control, and the extent to which it is territorial in terms of delivering effective infrastructural and fiscal resources, varies from nation-state to nation-state. Four extreme or ideal type categories, or what may be termed “sovereignty regimes,” can be identified (Agnew 2009; see Table 1). These are relational in character. They refer to the differing combinations of effective central state authority and territorialized provision of public goods in different countries. They are not best thought of as entirely characterizing particular states in all their aspects; no particular state fits exactly into any of the boxes in question. But they do provide a heuristic basis for identifying the relative complexity of state fiscal and infrastructural authority. This is a patchwork of more and less sovereign spaces and flows, not a rigidly territorialized order, with some territorial states and space-spanning organizations more effective than others.

<table>
<thead>
<tr>
<th>Central state authority</th>
<th>State territoriality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stronger</td>
<td>Classic</td>
</tr>
<tr>
<td>Weaker</td>
<td>Integrative</td>
</tr>
</tbody>
</table>

Source: Agnew (2009).

Of the four ideal types, the first one, the *classic*, comes closest to the conventional story about state authority. Both despotic power and infrastructural power are largely territorialized and central state authority remains effective. Contemporary China perhaps best fits this case. The second case, the *imperialist*, represents best the case of hierarchy in world politics but with networked as well as territorialized reach. It is the complete opposite of the classic case. Central state authority is seriously in question, often exercised by outsiders if in collusion with local elites, and infrastructural power is weak or reliant on external support. Most of the states in the Middle East and sub-Saharan Africa fall under this regime.

The other two regimes are more complicated. The first, the *integrative*, is a regime where authority has migrated to both higher and lower tiers of government as a result of a sharing of sovereignty among states, and infrastructural power takes both territorialized and networked forms. Various sorts of unions or confederations of states take this form. The fullest contemporary example would be the European Union from the perspective of its member states. Other economic blocs, such as NAFTA, reproduce some features of this regime but remain much less integrated.

The second of the two more complex sovereignty regimes is the *globalist*. This regime is closely associated today with the globalization of the world brought about since the 1960s under
NATIONAL AND REGIONAL INTEGRATION

US auspices, but which has older roots. In this construction, the world city system, particularly the cities at the top of it, such as New York and London, provide the geographical nodes for the agents who are central to this regime. They exercise power wherever states have ceded authority to external agents (credit-rating agencies, accountancy standards boards, etc.) because of their debt dependence or need for independent regulatory oversight. This regime has a potentially worldwide reach but its effects are particularly strong in those parts of the world most integrated into the world economy and without the limits set on integration by, for example, managed exchange rates and capital controls, by states with greater effective central state authority (such as China). The historic basis of this regime in US global hegemony since World War II means that, at least until recently, US governments were able to use this regime as an alternative to the imperialist one in many parts of the world. But, as authority has slowly seeped out of US governmental hands worldwide, other regulatory and fiscal agencies, including many private as well as public organizations, have picked up the slack.

Contemporary challenges to national integration

Increasingly, to many of their residents as well as to multinational businesses and globalized elites, many nation-states have become either too big or too small with respect to their capacities to both represent emerging elites (despotic power) and effectively exercise infrastructural power – deliver public goods – in a globalizing world. Smallness in area and population were historically disadvantageous for states but are arguably much less so today (e.g., Alesina and Spolaore 2005). Small states were once easy prey for larger states or empires and could not enjoy the economies of scale in the provision of consumer goods and public goods that were potentially available to larger countries. In a world in which insertion into world markets has largely replaced import substitution based on large domestic markets, not only is a large national territory and population no longer necessary, but where serious spatial disparities in incomes and wealth emerge, allied to cultural and political fissures, a large territory itself can become a candidate for fragmentation. Smallness is less problematic in a world of flows than in one of territories. Of course, large states have hardly disappeared, but this owes less to any economic logic than to bureaucratic inertia and continuing nationalist loyalties. The five countries with the highest GDPs per capita in the world today are Qatar, Luxembourg, Singapore, Norway, and Brunei, none of which has a population greater than 6 million. The tax policies favorable to multinational businesses, resource bonanzas, and lax legal regimes that allow for anonymous firm ownership and transfer price invoicing (parking profits in low-tax jurisdictions) have in different respects allowed for this efflorescence of small jurisdictions. Moreover, large countries typically have more culturally diverse populations that increase internal political tensions, particularly if the economic benefits of larger size are not equitably distributed to minorities. Larger dependent populations also mandate that governments cannot adopt the low tax practices of small states even as they try to reduce their overall levels of spending.

Worldwide, and as a result of increased global economic competition involving the footloose corporate actors that has made smallness less problematic, the government-sponsored social rights that came to characterize large, developed industrial countries in the 40 years after World War II are in retreat. Under policies often labeled as constituting “neoliberalism,” governments
have been retreating from direct provision of what had been defined as public goods necessarily provided by them to a range of indirect and privatized approaches that can be seen as representing the decline of a territorialized “society” congruent with the territory of the nation-state. Regional policies geared to attracting investment and jobs in poorer regions have been replaced by “private–public partnerships” and reductions in infrastructure spending relative to richer regions. Yet, this decline in commitment to the nation as a collective enterprise opens up the possibility of a serious legitimation crisis for states both generally and more specifically in relation to the withdrawal of support from localities and subnational regions heavily dependent on state spending, not least because of the hollowing out of industries located in some of those places as a result of globalization (Brenner 2009). At the same time, there are limited possibilities for new employment in emerging or more competitive sectors because the new jobs are limited in number and require skills that the now redundant people do not have. The new jobs are also often located in places such as capital or other large cities with housing and other costs beyond the possible incomes of those displaced from elsewhere.

Even while it is under threat from the fiscal crisis outlined, it is still government spending that holds together the fabric of national integration in the face of increased interregional income inequality. Of course, the degree to which states have engaged in fiscal transfers (such as tax credits for lower earners and social transfers such as unemployment payments and food supplements, etc.) has always varied since such programs began early in the twentieth century, with Scandinavian countries making much higher payments/subsidies to individuals than countries such as the United States and Japan. Nevertheless, equalization transfers have been very much accepted as part of the governance process in both federalist and centralist states. In a federalist state such as the United States, for example, many payments over the years have been indirect, through defense spending and other stimulative measures to encourage employment and income growth in different places than through direct payments to individuals.

To take just one example, in the United Kingdom as of 2007–2008 dependence on public spending as a percentage of regional GDP varied enormously across regions: from 34.1% in southeast England and 37.0% in London to 57.4% in Wales and 62.75% in Northern Ireland. This is in a context in which major gains in incomes over the past 40 years have taken place overwhelmingly in London and the southeast compared to modest gains in Scotland and decline or stagnation in average incomes in most other regions. Such economic differences and the potential political tensions that follow from them are visible across a number of large countries from Italy to Spain and China, to name just three, where economic growth is increasingly concentrated in some localities at the same time as public spending to even out collective welfare across the national territory is in decline (Figure 1). Over the period 1990–2009, however, the United Kingdom and the United States showed the largest increases in regional income inequality compared to most other countries for which data are available; these countries have also restructured their public sectors the most, along with New Zealand and a couple of others, by privatizing the provision of many public goods and outsourcing government functions to private firms.

Regional redistribution itself is under attack in some countries, such as Italy and Germany, where there is a fear that such redistribution is merely a stopgap to maintain incomes in poorer regions while removing needed revenues for infrastructure spending and providing public
Figure 1 GDP per person of poorest and richest regions as a percentage of latest national average (OECD TL3 regions for all except United States (states) and China (provinces)) (CEIC Data; Eurostat; OECD; The Economist).

goods to keep the richer regions competitive in global markets (Figure 2). In the United States similar concerns over local economic competitiveness, allied to cultural differences, have generated the regional polarization between the main political parties, and growing regional divisions over federal government policies in relation to welfare and health-care spending. The long-standing consensus about the need for national policies to underpin national integration that lasted from the 1930s until the 1980s has been undermined.

This has provided a propitious historical context for the revival of autonomist, irredentist, and secessionist movements geared toward fragmenting existing states. This can be because of perceptions of lagging behind other regions within one’s country or redistributive programs that are seen as reducing the relative competitiveness of one’s region in world markets. If in the past new states were mainly the product of major wars, today it is internal challenges that take prime position. From the Uighurs and Tibetans of western China to the Scots and Catalans in Europe, to name just a few, centrifugal forces are challenging the authority of larger centralized states. Existing cultural fractures are undoubtedly important in this process. But recent political-economic triggers also count: “Growing spatial inequalities within states based on the success of some regions in attracting capital in the global market put pressure on national solidarity and give regions an incentive to mobilize in protection of their own interests” (Fitjar 2010, 1).

Regional integration

The possibility of such disintegration is enabled by supranational integration. Smaller states can have their putative independence and immediately join more spatially extensive regional organizations to enjoy the fruits of potential economies of scale. Supranational integration involves neighboring states entering into agreements with one another to lower barriers to trade, capital and human movement, and forming institutions to manage disputes and encourage both the enlargement of the enterprise to include new members and the deepening of its reach into policy areas such as subnational regional development which were hitherto entirely in the hands of the states. This is the integrative sovereignty regime described earlier.

Though regional integration offers a tentative resolution to the economic conundrum of state size, it more often than not also reflects immediately political goals, such as the desire to reduce frictions between states in the interest of common security, and to stimulate the spread of “values” and policies beyond current membership into the wider regions around which they are organized. Indeed, the most institutionalized example of regional integration, the European Union, had at its origins in the 1950s the twin goals of reducing the possibility of the interstate conflicts that had wracked Europe earlier
in the twentieth century and providing the economies of scale necessary to make European industries more competitive globally than they could ever be on a state-by-state basis. Whatever its successes over the years, it faces ongoing challenges from the other sovereignty regimes, not least the globalist regime and its hollowing out of whatever territorial forms, state-based or supranational, governance might take. Most of the other attempts at regional integration, such as NAFTA (involving the United States, Canada, and Mexico), Mercosur (in South America), and ASEAN in Southeast Asia are more like regional multilateral free trade agreements. Perhaps in time so-called subregional economic zones may provide the elements for a more bottom-up type of regional economic integration in regions such as East Asia. But nothing elsewhere has as yet produced anything like the range of policies or institutional depth of the European Union – covering executive, legislative, and judicial initiatives. For example, the European Union has its own parliament which oversees its budget (a relatively modest 1% or so of EU GDP) and a court devoted to human rights, and it has developed its own version of competition law to deal with antitrust and uncompetitive behavior by businesses. In other words, it has supranational institutions and policies at its heart rather than being simply a regional intergovernmental organization like the others.

In the 1960s many commentators saw the European Community, the earlier version of what became the European Union in the late 1980s, largely in terms of functionalist notions of integration. Over time, it was held, economic measures encouraging cooperation would lead to a politicization of actors’ purposes and new
loyalties beyond those of member states would emerge. More recently, however, the idea that one would replace the other has been thrown into question by the actual course of integration. Since the 1980s a multilevel system of governance has emerged, with different levels from supranational, through national state, to subnational regional, playing distinctive and shifting roles in the entire enterprise. This too is now challenged, particularly in light of the crisis that has enveloped one of the most ambitious of the European Union’s initiatives, the euro, since 2010 by a reassertion in some quarters, not least in the United Kingdom, of state sovereignty and a desire to selectively “opt out” of EU policies. Yet, as Keating (2009, 48) says:

The state can no longer be presented, even in an ideal-type theory, as the sole focus and framework for territorial ordering in Europe, but it is not being replaced by similar structures at the sub-state or supranational level. Rather we are returning to a diversified and complex order more typical of Europe’s long-term experience.

Quite what “integration” is to be, then, is open to question, notwithstanding the seemingly self-evident use of the term by proponents of whatever it is (Murray 2009).

Recent thinking about the course of regional integration in Europe tends to involve a number of approaches, not always alert to the historical difficulty, given the twists and turns in its geographical expansion and policy thrusts, of giving the European Union a singular purpose or form. One emphasizes questions of “Europeanization” or the emergence of identities and interests transcending those of the member states (e.g., Clark and Jones 2008). The roots of Europeanization are often seen as going back prior to the formation of the nation-states of Europe – in medieval and early modern epochs, when authority was not neatly territorialized and a common European culture was at work – and thereby providing a mythos for the European Union that does not simply involve starting from scratch. The rescaling of governance then can be seen as having old, if not necessarily primordial, origins. The European Union is taking Europe “back to the future.” In this view, the future of Europe and that of the European Union are not readily disentangled. Yet, as Kuus (2011, 284) shows, “the fetishization of history” is problematic. This is so not least because it tells a “progressive story of European integration” (Gilbert 2008), of the eroding of national sovereignty, and of the emergence, phoenix-like, of a new supranational authority.

Another approach focuses rather more on the character of the European Union as a “polity” in comparison to that provided by other governmental models such as the United States (e.g., Sbragia 2004). Here the mechanics of integrating disparate states, from a minute Malta to a gigantic but federal Germany, receive pride of place. The workings of the EU Council of Ministers can be compared to those of the US Congress. In this construction, the specialization of different US states in specific industries produces a different type of political representation nationally than a Europe in which most regions/countries are much less specialized than the US states and thus lacking in the “defenders” that industries have under the US federal system. Recall that a late senator from Washington State was called “the Senator from Boeing.” These differences have serious consequences for redistributive policies within the United States and the European Union. The United States is a much more territorialized system in that, as a result of the ways in which its economy has evolved, “territorially defined constituencies are central to the dynamics of Congressional policy making” (Sbragia 2004, 211). With the Council of Ministers, the European Union has a much
more flexible (if obviously not directly democratic) institutional capability. For one thing, it is engaged in interstate bargaining and can thus adopt policies that are not uniform and allow adaptations for different member states. Ministers and their civil servants without narrow territorial interests to defend are also much more open to negotiation and compromise than Congressional representatives subject to constituent and lobbyist influence. Finally, national representatives can make trade-offs much more easily, particularly when they come from the diversified economies more typical of the European Union, than US politicians who are beholden to the typically sectoral interests based in their districts. Despite these advantages, however, the growth of an EU menu à la carte, with some countries opting out of this or that policy and the problematic administration of the common monetary policy among those EU countries using the euro, has led to a major credibility problem for the European Union in European public opinion (Mamadouh and van der Wusten 2013).

A final approach tends to engage more with the problems facing the European Union: from the difficulties of expanding to incorporate new members, to the neoliberalization of its policy agenda since the late 1980s (particularly the switch of cohesion policies from more or less supporting declining regions to attracting private investment to them). There is also the challenge of managing a common currency for 17 of the 27 member states in the face of serious fiscal problems among a significant number of them following the recession of 2008–2013, and the lack of common fiscal policies across the Eurozone as a whole, and growing disillusionment with the European Union (and its central institutions) in the face of ballooning unemployment and antipathy to the free movement of other members’ citizens in many member countries. Several problems seem particularly serious from the specific perspective of “integration.” One of the most noted of these is the noncompliance of many member states with EU rules and regulations, including those on democratic governance. The failure to enforce the rules about government public debt and revenue collections in the making of European monetary union is perhaps the most devastatingly important of these. Opting out, when done explicitly, has become a serious problem for creating a common direction for the European Union as a whole, but noncompliance compounds the problem because it implies lip service to the project without bothering to put it into actual practice (Falkner 2013).

Even more urgent, the geographical patterning of the Eurozone crisis has brought to attention the serious fracturing of the European Union between core states, on the one hand, and various types of peripheral state, on the other. This signifies the conjuncture between economic problems and geographical placement (Agnew 2001). As a result, it has become commonplace to hear about the group of so-called PIIGS countries, Portugal, Ireland, Italy, Greece, and Spain (to which could be added Cyprus), which have had serious problems financing their public debts, and countries such as Germany, Finland, and the Netherlands, which are seen as “virtuous” in terms of their public finances. Beyond this, of course, lie the vastly different economic growth problems of states on the edge of the European Union as a whole (such as Romania and Bulgaria) and those awaiting the call to membership (such as Macedonia and Serbia, among others). Added to the growing skepticism about EU membership in some countries (particularly the United Kingdom, France, and the Netherlands), the fractures within the EU project have led some writers to refer to the prospect of “European disintegration” (Vollaard 2014). In a globalizing world in which national integration
NATIONAL AND REGIONAL INTEGRATION

is now inherently doubtful, the difficulties of regional integration in Europe highlight the limited possibilities of this alternative approach without significant change in the ways in which capital circulates worldwide.

SEE ALSO: Economic geography; Governance and development; Governmentality; Nationalism and geography; Nation-state; Political geography; Regional economic integration; Sovereignty

References


Nationalism and geography

Virginie Mamadouh
University of Amsterdam, Netherlands

Nationalism has proved to be one of the most powerful ideologies of the past two centuries and is certainly the most geographical. Political claims based on national identities generally involve strong territorial claims and strategies. “Space, place, territory are absolutely essential to the imagination of nationhood and to nationalist politics” (Brubaker, interviewed in Sturm and Bauch 2010, 186). This might explain the fascination of political geographers with its occurrences and their consequences. Nevertheless, the primary purpose of deploying nationalism is to mobilize people as a group by prioritizing national identity over all other social identities; it is a powerful form of groupism. Key to nationalism is the conceptual confusion of the state as political organization with the land (or territory), or with the people (or nation), under its jurisdiction. In English and many other European languages this confusion is expressed in the indistinct use of “people” and “nation,” “state” and “country,” and “state” and “nation” as if they were pairs of synonyms. Think of the adjective “national” used to qualify features related to a state, the term “international relations” for relations between states, or the name “United Nations” for a formal association between states. Modern territorial states are also routinely identified as nation-states (i.e., nation hyphen state) to stress the close association of the nation as a social body with the state as a political entity.

Nationalism comes in many guises, depending on the representation of the nation it advances, the political agendas it supports, and the societal effects it brings about. Agnew (2013) has warned against the easy dismissal of nationalism in contemporary social sciences, neglecting both its appeal and its emancipatory effects. Too often, it is seen as a dark force and an outdated one. The campaign for the 2014 Scottish referendum for independence showed how nationalism can mobilize large numbers as a progressive force, foregrounding cultural diversity, inclusion, and solidarity. Still, there are plenty of examples where/when nationalist ideologies are used to exclude, disempower, displace, and even murder the other (those who do not belong to the nation).

The importance of geography in nationalist ideologies

Geographical features are important in nationalist ideologies in several ways. A generic nationalist ideology asserts the hegemony of national identity over all other social identities, while specific nationalist ideologies assert the content and claims of a specific national identity. These specific nationalist ideologies, although they share the same generic nationalism, often clash with each other because they make competing and contradictory claims on specific groups and territories. Therefore, on top of opposing alternative ideologies such as socialism, liberalism, anarchism, communism, or Islamism, nationalists often oppose each other, and these claims can easily lead to conflicts and generate vast human suffering. But even when the territorial borders...
between national groups are pacified and largely uncontested, the generic ideology can put individuals in a difficult position when they do not conform, either because they have no national identity or it is not important to them, or because they combine several national identities.

Political geographers are particularly interested in territorial claims. The justification of such claims can be very diverse and may foreground different and largely contradictory arguments. For example, claims to a specific territory and claims about the specific positioning of the border between two territories can rest on historical arguments (“We have always lived here”; “Our ancestors lived there for centuries”), necessity and productivity (“We need more agricultural lands to produce enough food for our population”; “We can develop the land and utilize the available resources better than another group”), geopolitical logic (territorial integrity and the avoidance of enclaves and exclaves; practicalities regarding the defense of the territory against external attacks; buffers to protect a nearby city), a distinct physical geographical shape (islands), or teleological goals (i.e., the promise of a bright future). Historical claims are particularly likely to generate conflicts, especially in densely populated areas. Coexisting groups often claim historical events and past political organizations as heritage in the same way as they claim space, and they base conflicting territorial claims on such appropriation. Reactualizing the political organization from the past that best suits their present political goals, they can attempt to project power onto space, according to the most extended area this particular entity ever controlled.

Nationalist claims to territories are multidimensional. They entail claims to specific lands and the use of the resources in these lands: crops cultivated; labor force living there; ores, salt, or oil extracted from the soil; water running through the area. They also pertain to the territorialized bordering processes that consist of both the specific location of the borderline and the guarding of the border, especially the regulation of the entries into the national territory and the entitlements associated with membership of the national community. The most explicit and formal cases are nationalist ideologies promulgated by the state to regulate citizenship, but stateless nations also sustain strong views about who belongs and who does not, and a myriad of practices regulate these social boundaries.

Nation (and) state

The relation between nation and state is complicated. While nationalism first emerged in existing states to mobilize the population in its territory, that is state nationalism, it soon became a force to mobilize against the state, that is bottom-up nationalism against the existing state(s). Nationalism has notably inspired the articulation of struggles for autonomy and independence in multinational empires in Europe, such as the Habsburg empire, the Ottoman empire, and the Russian empire, and in colonial empires beyond Europe, in the Americas, Asia, and Africa. Bottom-up nationalism against the state is still a powerful force wherever groups see their cultural differences from the majority group of the state as a cause of discrimination against them; such differences are framed as sufficient conditions to warrant claims to autonomy and independence. Sometimes such differences are very consequential in everyday life and are expressed in different languages, religious beliefs, norms and values, and/or lifestyles, but at other times they may be very small and social boundaries between groups may be very fluid. Moreover, spatial identities can be sufficient by themselves: the distance from the capital city of the state, different climatic conditions and physical harshness, and different relations with the outside world may also foster a
feeling of alienation from the center of the state and its rulers that can be the very motivation for aspirations to autonomy. Finally, economic disparities can inspire such mobilization to resist exploitation from the political center, but sometimes it is in richer regions that national identities are mobilized against the state, to opt out of solidarity arrangements with poorer regions.

In general, it should be stressed that the very institutionalization of a territory as a separate territorial entity coevolves with the institutionalization of a group as a separate cultural entity. In other words, the nation and the territory co-constitute one another, even more strongly when state institutions are available to organize both the nation and the territory. Regulations promoted by the state, especially regarding language in public space, political institutions, media, and educational institutions, are strong vehicles for these identity-building processes.

Typologies of nationalisms reflect the key importance of the very nature of territorial claims to classify nationalist ideologies. Whether nationalism refers, in its generic form, to the isomorphy between the territory of the state and that of the nation, there are different ways to try to achieve this ideal state of affairs. State nationalism refers to nationalist ideologies enforced by the state to mobilize the population on its territory and to homogenize it to form a national community. In such discourse territory and nation are so intertwined in nationalist discourses that love for the nation as community and love for the land and its iconic landscapes can hardly be disentangled from one another, and are used next to each other, as in the British recruitment posters shown in Figure 1. Ethnoregionalism and substate nationalism cover a wide array of nationalist ideologies that demand a separate state for a community with a clear territorial shape through secession from an existing state or a multinational or colonial empire.

Many nationalist ideologies are more complicated, even if we consider only geographical aspects. Irredentism refers to groups that do not accept the delimitation of the border lines between two states and support claims to reunify a territory beyond the border to the state, hence the name – not yet recovered. Typically irredentism pertains to areas where people speak the same language in a contiguous area across an international border (originally the Italian speakers in Istria and Dalmatia), a situation, however, that is very common and not always conducive to such claims (think of French-speaking or German-speaking Swiss, Dutch-speaking or French-speaking Belgians, Swedish-speaking Finns, etc.). Irredentism can be either a local force (when the local population campaigns for a change of the borderline), a national one (when the state or a national movement claims a territory “back” from its neighbor), or both. The national variant might even completely negate local self-determination, when the adjacent territories are claimed on the grounds that they were once part of the national homeland and have been unduly annexed by others, while the local population has other linguistic and religious characteristics and publicizes other political affiliations – think of the claims of Serbian nationalists on Kosovo.

Another noteworthy form of nationalism, from a political-geographical point of view, is pan-nationalism: a nationalist ideology that unites populations of different states, generally on the basis of common cultural features, like language and religion: Italian and German nationalism led to the unification of Italy and Germany as modern states in the nineteenth century; pan-Europeanism was unsuccessful during the interbellum; pan-Slavism, pan-Turkism, pan-Arabism, and pan-Africanism were vigorous forces in the twentieth century; and
Figure 1  Two British recruitment posters encouraging men to enlist to fight against other nationalists during the Great War. The posters refer to (a) territory (i.e., the landscape) and (b) the nation (i.e., the women and children) to be protected. Both posters were produced and circulated by the Parliamentary Recruiting Committee in 1915. Image credit: Imperial War Museums (IWM).

pan-Americanism (aka Bolivarianism) still is in certain parts of South America.

Last but not least, diaspora nationalisms (or more neutrally labeled long-distance nationalisms) demonstrate the enduring strength of cultural bounds in exile and how feelings of belonging can be maintained while living far from the original homeland. They have been studied extensively, especially when the discursive, financial, and physical engagement of the diaspora has weighed in hot conflicts, such as the cases of the Irish diaspora in Northern Ireland and the Jewish diaspora in Israel.

It should be stressed that nationalist ideologies may have different relations to territory. Some ideologies may be more interested in the generic characteristics of territory (extent, natural resources, access to the sea) as resources for the state and the nation, and others more concerned with a specific place because it has a specific meaning for the nation and its history. The first type will accept land swaps – the
NATIONALISM AND GEOGRAPHY

The territorially equivalent of ethnic cleansing – to culturally homogenize the state population or to achieve a compact shape for the state territory more easily than the second, which will see a specific place as irreplaceable.

Nationalism and citizenship

The politically framed association between nation and territory matters for relations between nations or between states, but it also defines the individual’s membership of the nation and the (nation-)state. It is customary in this regard to distinguish civic and ethnic nationalism. In civic nationalism, or liberal nationalism, membership is rooted in a political definition of the community and the political will to apply the law, to participate in the civic community, and to advance the common good. Ethnic nationalism, by contrast, stresses blood ties and descent as the essence of membership. In this case it is difficult to become a member if one is born to parents who are not members of the national community. Sometimes an intermediate variant is distinguished, cultural nationalism, when blood ties do not matter but the willingness to endorse a specific cultural identity is expected. In civic and cultural nationalism, membership is generally based on birthplace and there are more generous procedures for naturalization (the conversion to citizenship) and more tolerance for dual citizenship (the combination of the citizenship of different states).

The role of (academic) geographers in nationalist projects.

It is well known that the institutionalization of geography as an academic discipline and as a school topic in the late nineteenth century in European countries has been strongly associated with their nationalist and imperialist projects (Hooson 1994). The making of modern (national) citizens was promoted through the spread of universal (state-sponsored and state-controlled) education; the formalization, homogenization, and imposition of a common national (i.e., state) language; and the circulation of a common body of geographical and historical knowledge that naturalizes the existence of the nation and state. Schoolbooks have proven illuminating sources to study how whole generations were socialized into a specific perception of the nation to which they belonged and its position in the world. Geographers have been instrumental in the production of knowledge and of teaching methods to achieve this task. Theirs could be labeled as the main “banal” contribution of geography to nationalism.

Another banal contribution common to all social sciences and humanities pertains to the nationalization of research objects, and is sometimes called methodological nationalism as it takes for granted nationally defined sets of social relations, that is, national societies. The states are accepted as given, as natural containers of social relations: we speak of Dutch politics, French media, German culture, Italian economy, as so on, as if they were material entities. Just as art historians tie painters to specific countries and therefore contribute to the idea that there is a national painting culture and a national identity essentially distinct from that in neighboring and competing states, geographers contribute to the idea that there are national landscapes and national geographic features essentially distinct from those in other states and contributing to the specific essence and vitality of the nation. Of course, geographers have also been involved in many other state policies that have contributed to the strengthening of the infrastructural power
NATIONALISM AND GEOGRAPHY

of the state, such as regional and urban planning, infrastructural decisions, and economic development policies.

Geographers have also played a role in “hotter” nationalist projects, with interventions in nationalist struggles, as participants in nationalist movements or as advocates of nationalist claims. Geographical knowledge and academic credentials have often been welcome to produce and circulate narratives and maps to support nationalist claims to territory or autonomy, or to defend the nationalist project of an existing state. Rudolf Kjellén (a Swedish law professor who coined the very successful neologism “geopolitics”) argued and campaigned against Norwegian independence from Sweden at the turn of the twentieth century. In the same period, Halford Mackinder, a key figure of British geography, promoted good relations with Russia to counter Germany as challenger of British hegemony. Paul Vidal de la Blanche, the influential founder of French regional geography, wrote a book to demonstrate that Alsace-Lorraine, the French region annexed by the unified Germany in 1871 after the Franco-Prussian war, ought to be returned to the French Republic. These scholars were certainly not exceptions.

Geographers have also been instrumental in the settlement of nationalist conflicts. They were especially influential at the end of World War I, when a new political map of central and southeastern Europe and of the Middle East was negotiated in a series of peace conferences to dismantle the multinational Habsburg and Ottoman empires. Guided by the 14 principles enunciated by the American president Woodrow Wilson, geographers were important advisers to the statesmen applying self-determination to the redrawing of the European political map. They were asked to provide scientific evidence for the determination of the nations (i.e., the nationalities, in the jargon of the time) that were entitled to a state of their own and the delimitation of the borders between the new states. Language and religion were the two main characteristics to determinate “objectively” the existence of separate nations, but there were many disputed cases. Indeed, when is a language a separate language, and when is it a variety of another language? A notorious case is the fate of Macedonians which depended on whether Macedonian was seen as a separate language, as a variety of Serbian, or as a variety of Bulgarian. The Serbian geographer Jovan Cvijić, who had published very different cultural maps of the region over previous decades, advocated, as the adviser of the Serbian delegation, the incorporation of Macedonians into the new Yugoslavian state federating the South Slavs (Wilkinson 1951). The negotiation of the borderlines was even more difficult, because the pattern of settlement in that part of Europe and the Middle East was such that different groups lived in mixed areas and it was virtually impossible to design (ideal-typically) contiguous territories inhabited by a culturally homogeneous population. Self-determination was applied very selectively, depending on the interests of the winners and their alliances.

Isaiah Bowman, the geographer advising the American delegation, published The New World (1922) afterward, in which he identified the contested areas and the territorial conflicts that would eventually emerge from the peace settlement and the new political map of the world. In Germany, the main loser of the war, resentment was widespread against the conditions and penalties imposed, including the loss of its colonies and of territories in Europe, the displacement of German-speaking populations, the demilitarization of the Rhineland, and huge financial reparations. Geographers involved in the establishment of a German school of geopolitics provided this revanchist political agenda with scientific arguments about the Lebensraum the German nation
needed, its precarious position surrounded by hostile neighbors, and the need to reunite German-speaking populations in one state.

To balance this account, it should be also remembered that others among the founding fathers of geography had completely different aspirations. Anarchist geographers such as Élisée Reclus and Peter Kropotkin clearly envisaged a different role for geography and the popularization of geographical knowledge. They saw geography as a tool to overcome cleavages between nationalisms, to promote peace universally, and to empower free world citizens.

The geographical study of nationalism

The geographical study of nationalism is characterized by case studies addressing nationalist projects of states as well as nationalist projects against states, and a wide array of topics, including the representation of the nation; the implementation of state- and nation-building policies in newly established states; the disintegration of multinational states, the emergence and diffusion of nationalist ideologies; the mobilization and organization of nationalist movements and their strategies and tactics, especially the use of violence; patterns of distribution and diffusion of nationalist protests and nationalist organizations; the electoral geography of nationalist parties; the representation of national identities in popular culture and the media; the reproduction of national identities in everyday life; and the dynamics of nationalist struggles. In these later types of study the case is not one nationalism but the clash between two nationalisms.

Much of the research pertains to European cases, the cradle of the generic nationalist ideology and for two centuries the battlefield of numerous hot and murderous conflicts between specific nationalist ideologies. It is typically a field in which the depth of a case study is the best way to achieve conceptual progress. Comparative studies are less common. They distinguish different types of nationalisms, and they try, foremost, to identify factors that fuel conflict and that enable pacification. For example, the Irish political scientist John Coakley (1993) examines the geographical determinants of success using two dimensions: concentration and dominance. The first measures the degree to which the group is concentrated in the claimed area, the second the degree to which the group dominated in the area or whether other groups are more important. Claims for a group with high concentration and high dominance are more likely to be successful. When a group is spread over a much bigger territory, autonomy in one region seems insufficient as it will create different situations for group members according to their place of residence that might lead to the (forced) displacement of those outside who will be resettled in the chosen area. When a group is not dominant in the area, autonomy might be resented and opposed by the rest of the population in the region or might set in motion new autonomist mobilizations at a subregional scale – think of the mobilization of the anglophone minority in Montreal after the status of French was enhanced in Quebec, the Canadian province with a majority of French speakers.

The scheme warrants at least two important remarks in this discussion of the relation between nationalism and geography. First, Coakley puts areas before groups. His examples are labeled after the area, not the group for which autonomy is claimed, and even less the movement or party making these claims, thus contributing to the blurring of the spatial container and the social entity it represents, and even more the political entrepreneurs that speak/act in its name. Second, his examples demonstrate that labeling is already constitutive of the nationalist discourse.
NATIONALISM AND GEOGRAPHY

and of the assessment of the situation, rather than a neutral observation. Whether a group is concentrated and/or dominant depends on how it is defined, named, and delimited with regard to the surrounding groups.

Coakley also provides a typology of state responses to territorial claims. These responses can be repression and the removal of difference, through compulsory assimilation or through ethnic cleansing (either displacement or murder). Europe’s history in the nineteenth and twentieth centuries has been rich with examples. The French state is the most famous example of a state that homogenized its population linguistically by actively suppressing the use of regional languages in public space. Exchanges of population were common after changes of the political map: Greeks and Turks in modern Greece and Turkey (much later, in 1974, between Cyprus and Northern Cyprus); German speakers expelled from Czechoslovakia and Poland; and Finns and Poles displaced from the Soviet Union. The list of extermination policies in the twentieth century is very long: Armenians in Turkey, Jews in territories controlled by Nazi Germany, Poles in Lithuania, Bosniaks/Muslims in the Serbian and Croatian territories of Bosnia–Herzegovina. Outside Europe nationalist ideologies have been deployed to homogenize the state population as well, the largest example being the displacement of population and the killing of Muslims and Hindus at the partition of the British Indian empire and the independence of India and Pakistan in 1947.

Alternatively, the state can respond positively to territorial claims and accommodate demands in a nonterritorial manner, for example by providing services and support for specific cultural institutions across the state territory. The response can take the form of a territorial arrangement: providing not only services and support but also autonomy to claimed areas. Such autonomy can be limited and conditional, with sovereignty remaining with the central state (as in the case of regionalist arrangements), or sovereignty can be shared (federalism) or even transferred to the smaller entities (confederalism or independence). Arrangements can be symmetrical when the tasks and competencies are delegated to all entities across the state territory in an identical manner (Belgium, France) or asymmetrical, giving more privileges to areas with strong mobilization and less to others (United Kingdom, Spain, Italy). New territorial arrangements set the stage for a new round of mobilization (Van der Wusten and Knippenberg 2001), and whether or not devolution or autonomy demobilize substate nationalists or encourage them to claim more autonomy is a matter of debate.

The study of nationalism from a geographical perspective has blossomed over the years, inspired both by the debates in the field of nationalism studies (primordialism vs modernization theories about the nature and continuity of nations and nationalisms) and the theoretical and methodological evolution of human geography. Geographical analyses of nationalisms demonstrate that place matters and that nationalism is not equally important everywhere but appears in different shapes at different times and places. It also underlines the importance of the context for the evolution of nationalist projects and of the interaction between nationalist movements and political institutions.

The role of nationalism in the disintegration of multinational federations, especially in the Soviet Union and in its successor states, including the position of Russian speakers in these new states, has drawn much attention. Nationalism was first seen as resurgent after decades of state communism but more careful examinations of Soviet policies demonstrated the complexities of Soviet policies regarding (titular) national
identities and showed how nationalism had been shaped by these state policies (Kaiser 1994). Nationalism was so important in 1991 not despite decades of state communism but thanks to Soviet policies toward nationalities.

The relation between state nationalism and substate nationalist movements has been explored by Mansvelt Beck (2005) in his study of Basque nationalism in the Spanish and the French part of the Basque Country. He shows how different state- and nation-building trajectories, the Spanish one and the French one, brought about different articulations of national (state) identities with Basque cultural identities and different national (substate) mobilizations of Basque nationalism, south and north of the state border.

Other geographers have stressed the intimate relation between national identities and geopolitical representations (Dijkink 1996; Paasi 1996), and how the representation of national identity and the nationalist (state) project are co-constitutive with the representation of the role of the nation in the world and of its relations with neighboring and other states. Following the cultural turn in geography, much attention has been given to cultural artifacts like national monuments and cultural events like national days and their role in fostering national identities, and to the circulation of nationalist discourses in popular culture and in the media. A particularly important development has been the study of nationalism from a multiscalar perspective. It was also inspired by feminist studies on the importance of the family and the everyday in the reproduction of national identities and of nationalism, and of gender roles in nationalist movements. Generally, the intersection of different identities and ideologies and the configuration of nationalist discourses with other ideologies such as socialism, and of national identities with language, religion, and race, need more engagement.

The local (re)production of nationalism is another theme of growing importance. Geographers have also followed the call of the British social psychologist Michael Billig (1995) to address banal nationalism instead of only looking at the more extraordinary expressions of nationalism, when people are prepared to kill others in the name of their nation and to be killed in the process. Sport appears to be a particularly powerful domain for the mobilization and the expression of such nationalist sentiments. Nationalism in the culturally diverse city is another. Inspired by the recent materialist turn of nonrepresentational geographies, attention is now firmly focused on the way nationalism is practiced and embodied.

The future of nationalism in the transnational world

Nationalism is a modern political ideology centered on national identity and territorial claims. It will eventually lose its appeal, but for now it remains one of the most pervasive ideologies in contemporary (local, national, and global) politics. Recent processes of globalization – including enhanced interactions stretching over large distances and the potential of new media – seem to have reactivated its salience rather than to have brought about its often announced dismissal. This may be because nationalism provides individuals with a strong sense of belonging that is easy to communicate and to ground in institutions and routines, and because the need to belong somewhere is felt all the more acutely when uncertainties and anxieties grow with globalization. At the end of the nineteenth century, nationalism competed with
two other main ideologies, liberalism and socialism; at the beginning of the twenty-first century nationalism and neonationalism compete mainly with neoliberalism and religious universalism.

In times of enduring transnational migration and transnationalism through the maintenance of cultural ties across state borders, migration and the management of the flows of people has become a particularly important issue in nationalist ideologies that often surpass the pre-eminence of drawing borderlines and the management of territorial claims proper. Nationalist and neonationalist discourses in Western countries – which in Europe are also directed against integration in the supranational framework of the European Union – often incorporate time in claims to authenticity and belonging to a certain place, scripting cultural elements associated with recent immigrants as less legitimate because of their recent arrival in the national territory. In turn, newcomers therefore feel excluded and this may encourage an orientation toward other cultural centers, demonstrating that place-making and home-making are at the core of the nationalist imagination of the world. Evidently the length of stay needed for individual residents and their families to qualify as legitimate inhabitants of the national territory along with established groups of residents, to stop being out of place, and to be seen as a host too instead of as a(n unwelcome) guest, is open to contention and a hotly debated issue.

While more traditional nationalist conflicts over territory and borders remain high on the political and research agendas (think of the Israeli–Palestinian case), new salience is given to the clash between inclusive nationalist discourses and national identities embracing multiculturalism and exclusive ones. Such inclusive national identities may also provide new horizons for institutional arrangements in traditional conflicts. In both types of conflicts, old and new nationalisms, geography can contribute to understanding these dynamics better, by providing studies that are sensitive to local differences and by conceptualizing the contingencies of nationalisms as they emerge and develop in time and place-specific sites.

SEE ALSO: Borders, boundaries, and borderlands; Diaspora; Geopolitics; Identity; Irredentism; Nation-state; State, the; Territory and territoriality

References


Sturm, Tristan, and Nicholas Bauch. 2010. “Nationalism and Geography: An Interview with Rogers
Further reading


Geographers have studied the state in a number of different ways, focusing especially on the state’s role as a regulator of the economy, provider of social welfare, and arena for political participation. The concept of nation-state draws attention to the state as a community, that is, the state as a focal point for identity, and the nation as a source of legitimacy for the state. These links between the state and nation define the conceptual core of the nation-state and provide a basis for distinctions with regard to state forms that are based on other conceptions of community, state territoriality, and state–community relations, for example, city-states, multination states, empires, and confederations. State–community relations are also the focal point for nationalist ideology, revolving around the principle of political and territorial congruence between a sovereign state and an autonomous national community.

Nations, nationalism, and nation-state formation

The focus on the community dimension of the state makes national identity an obvious analytical starting point for examining nationalism, nation-state formation, and nation-state models of citizenship. There is, however, no consensus on the meaning of the “nation.” Debates about its character, origins, and political roles have revolved around a polarization between primordialist and perennialist approaches, which see nations as essential, bounded, and stable, and modernist and constructivist perspectives, which view nations as constructed, changeable, and historically contingent (Özkirimli 2010).

Primordialism sees the nation as a natural category that is defined by objectively given organizing principles such as blood, race, ethnicity, language, religion, and territory. Such primordial bonds of human association, it is argued, have always and naturally divided populations. They hence precede and provide a given basis for the formation of nation-states. Perennialism does not share this view of nations as natural categories, but nevertheless sees them as long-standing and relatively stable social and historical phenomena. Contemporary nations stem from entrenched ethnosymbolic systems with deep historical roots. Both approaches hold that contemporary nations and their homelands are objectively defined and premodern in origin, and hence predate and shape modern state formations and nationalist politics.

Primordialism and perennialism held a hegemonic position in both popular and academic discourses on nations and nationalism for a long time, but have come under strong scholarly criticism because of their essentialist, static, and universalizing analyses of nations, nation-states, and nationalism. These perspectives are also viewed with skepticism as a result of their association with nationalist ideology. Primordialism provides a conceptual foundation and political legitimacy for the core argument in nationalist ideology that nations do exist and have unique characters, that the values and interests of nations take priority over those of other collectivities, and that the nation must be politically autonomous.
and protected through nation-states. Nationalism as a political movement seeks to achieve these core principles, and primordialism may provide a conceptual basis and political legitimacy for such actions (Herb and Kaplan 2008).

Modernist and constructivist perspectives share a view of nations as social constructions that are interwoven with nationalism. This means that nations are not given and fixed categories, but have come into being in the context of nationalist ideas about natural, bounded, and stable national communities. The principal difference in opinion between modernism and constructivism is with regard to the possibility of causal explanations behind the construction of national communities. Modernism holds that it is possible to identify economic, social, and political causes behind the invention and instrumental use of national identities and traditions. While some scholars emphasize capitalist economic transformations and the associated social division of labor as the drivers behind nationalism and the invention of modern nations, others emphasize the emergence of modern states that made nationalism a particularly appropriate form of politics, not least for the dominant classes in pursuit of hegemony in the face of radical politics by subordinate classes. Modernist scholars thus agree that the modern state and nationalist ideology preceded the invention of nations, rather than nations being an objective basis for nationalism and nation-state formation.

Constructivism rejects the modernist view of nations as invented traditions that can be causally explained with the needs of capitalism, the modern state, or dominant class interests. The focus is instead on understanding nations as discursive constructions – “imagined communities” – built around boundary demarcations of nations and national homelands, and the bonds between people and place (Jackson and Penrose 1993; Paasi 1998). Nations are constructed and institutionalized through narratives about who “we” are, where we belong, and how we are different from the “others,” who belong elsewhere. To examine the meaning of the nation is to examine who are rendered culturally and territorially included or excluded in representations of national communities, how imagined communities are institutionalized, and the manner in which discursively constructed identities constitute a basis for identification and everyday banal nationalism.

Nations and their “homelands” may be seen as given or constructed, but there is broad agreement that national identity and territorial self-determination are intrinsically linked to nation-states. The formation of nation-states has been a long-standing research tradition within and outside human geography. The existing literature revolves around two principal ways of pursuing congruence between national and state boundaries; either from nation building to nation-state, or through state building to nation-state (Penrose and Mole 2008). The state route to nation-state involves modern state formation as a precursor to building a nation of all people living within its territory. The nation is thus conceived as a civic collectivity built around universal citizenship, democracy, and liberalism. The process of creating cultural uniformity thus follows from rather than precedes the formation of the modern state. This focus on state building is in contrast to the nation route to nation-state formation, where self-determination is demanded for a culturally predefined nation. Both trajectories of nation-state formation are nationalist in the sense that they seek to achieve congruence between the boundaries of a modern state that is based on the idea of government for and by the people, and a nation that is construed as a cultural community of equals. Where they part
ways is in the sequencing of nation and state building in the process of nation-state formation.

Geographers have also given extensive attention to nation-state building, that is, the interrelated processes of constructing and politicizing the nation and to building the state and imbuing it with national meaning. Penrose and Mole (2008) observe that nation-state building includes at least some of five main processes: state building centered on creating state institutions for government and representation; establishing state monopoly over the legitimate use of force; building national systems of education and language; developing media institutions and mass communication; and fostering shared systems of meaning. All of these are always and unavoidably spatial and territorial processes, providing rich opportunities for geographic scholarship on the building and transformation of nation-states.

The nation-state and citizenship

While the analytical foci and theoretical perspectives may change, geographers continue to pay critical attention to state–community relations. One notable reorientation is that contemporary studies are increasingly framed by theories and debates about citizenship and nation-state transformation. Citizenship is conventionally taken to mean a formal legal status that is based on membership in a community and that provides a basis for citizen rights and responsibilities. The modern nation-state model of citizenship rests on membership within a nation, but the national community may be constructed in different ways. The previous section pointed to a general distinction between ethnocultural and juridical–political constructions, that is, nations that are built around cultural bonds and a historical homeland, or around people living within a territorial state under common law and legislature. Following from this distinction, citizenship has conventionally been granted on the basis of the parents being citizens (jus sanguinis) or on the basis of being born within the territory of a state (jus soli). Increased cultural diversity in the context of international migration has, however, challenged such simple distinctions and transformed the criteria of citizenship. Hybrid systems for legal citizenship have thus become more prominent as immigration countries have come to realize that the exclusion of large groups from citizenship has problematic implications in terms of social marginalization, political exclusion, conflict, and racism. In this situation, exclusion and stratified membership in communities of citizens have become key issues in citizenship studies within and outside human geography (Kofman 2006).

Citizenship studies have been marked by a cultural turn that has raised critical questions about cultural differences and cultural politics within the nation. Citizenship rests on assumptions about cultural homogeneity, and the principle of universality means that all members of a political community are granted citizenship on equal terms. It also means that commonality of identity among citizens is privileged at the expense of particularistic group belonging, and that all are treated equally in the sense that laws and rules apply to all in the same way. The liberal perspective on citizenship holds that the principle of universality constructs citizenship in a manner that transcends particularity and difference in favor of equality and justice. Critics argue, in contrast, that universality conceals how citizenship is defined in terms of dominant identities and thus puts other identity groups at a disadvantage, even if everyone is granted equal citizenship in legal terms. While universal citizenship is seen as an instrument of emancipation and justice, it simultaneously masks and
accentuates group oppression in terms of gender, sexuality, class, ethnicity, and so on.

Citizenship studies have drawn attention to contentions over the character of national communities and struggles for inclusive membership. Isin and Wood (1999) describe such politics of membership as “cultural politics” and make a distinction between identity politics revolving around claims on the basis of durable group identities, and politics of difference emphasizing recognition of differences. Identity politics and politics of difference present opposed views on the nature of identities, with divergent implications for political practices.

*Identity politics* refers broadly to political struggles around shared experiences of injustice based on involuntary membership in a particular social group. Conscientization about the mode of oppression that a group experiences is thus central in challenging stigmatizing representations and in demanding group-differentiated rights. Recent years have also seen the emergence of a *politics of difference*, focusing on the constructed character of identity and differences within any imagined community. This is in marked contrast to identity politics, which sees groups as defined by objective relations of oppression.

The question of membership in communities of citizens has thus produced two modes of cultural politics, separated by a divide between essentialist and constructivist views on identity. Whereas identity politics builds on essentialist notions of identity and seeks a transition from oppression to social and political inclusion, the politics of difference is based on a constructivist view of identity, critical attention to difference, and a broadening of what issues are considered relevant to citizenship.

Geographers have paid increased scholarly attention to the policies and politics of inclusion/exclusion in cultural membership and formal citizenship, including a critical focus on different forms of partial citizenship between noncitizens and full citizens. The prevalence of hybrid and stratified citizenship blurs the idealized image of binary distinction between citizens and noncitizens, and of equality among citizens. Differentiation at the level of *becoming* a citizen is also mirrored in diverse and stratified experiences of *being* a citizen. Renewed attention to citizenship has thus provided new opportunities for geographic scholarship on changing and contextual links between the state and communities.

Transformation of the nation-state model of citizenship

The nation-state model of citizenship has come under pressure as a result of increased diversity and cultural politics within presumably homogeneous nations, but also from the transformation of territorial state sovereignty in conjunction with increased globalization.

The modern model of liberal citizenship rests on assumptions about territorially bounded nations and states. The nation-state is seen as the principal site for citizenship, because of its sovereign authority to grant legal status as a citizen within its territory and to define the extent and content of rights and political participation. This territorial model of citizenship has come under pressure in the context of various forms of globalization. Increased economic globalization has reduced the sovereignty of the state and contributed to the emergence of multiscale forms of governance and citizenship. Global neoliberalization of governance also means that citizenship rights and participation are defined not only by the citizens’ relations to the state, but also by the market and civil society. Finally, increased international mobility has produced a growing number of people with dual citizenship or
transnational belonging, challenging the territorial nation-state model of citizenship. It can thus be argued that economic globalization, globalized neoliberalism, and international migration have introduced and reinforced tendencies toward cross-scale, multisited, and transterritorial citizenship. Citizenship has become increasingly complex in geographic terms, and the substance of membership, rights, and participation has come to be defined through multiple and relational scales and spaces (Desforges, Jones, and Woods 2005).

Citizenship has been decentered in regard to the territorial nation-states, although the state remains a prominent domain of citizenship, leading some scholars to the observation that citizenship is being transformed in postnational, denationalized, and transnational directions (Sassen 2002). A postnational trajectory means that citizenship comes to be located outside the framework of national communities in the sense that it rests on new forms of community and that it transcends the institutional framework centered on the nation-state. Denationalization, in contrast, refers to a transformation of citizenship away from the national scale that nevertheless remains within the general framework of nations and states. Such a denationalization trajectory is especially visible in the emergence of multilevel citizenship.

The most obvious example of multilevel citizenship is found in federal states, where individuals are citizens with rights and responsibilities both at the level of the component states and within the overarching federal state. Devolution of power to semiautonomous regions in a quasi-federal state or to municipalities and cities in a decentralized state structure provides additional examples of multilevel citizenship. Such subnational forms of citizenship can be characterized as scaled down versions of the nation-state model of citizenship. At the opposite scalar end, there are also examples of citizenship at the supranational level based on the nation-state model. The foremost example here is citizenship at the scale of the European Union, which also includes formal legal status, rights, and political participation.

The postnational trajectory is a more radical departure from the nation-state model of citizenship since it involves constructions of communities of citizens and rights outside the framework of the nation. In the past, citizenship was broadened when previously excluded groups (e.g., workers, women, children) were included in the national polity. In the contemporary period, some argue that rights that used to be reserved for nationals are increasingly granted to resident foreigners (denizens). This means that membership has come to be based on discourses of personhood and that rights are framed as human rights. It does not mean, however, that the state has become irrelevant. While universal personhood and human rights have become a legitimating basis for claims to rights, the realization of such rights is still tied to state institutions.

A third trajectory is found in the form of transnational citizenship. Transnational citizenship means that legal status and rights are situated in national institutions, while citizenship as membership (belonging) and participation (social, cultural, and political practices) is embedded in transnational fields. This creates a situation where belonging and active citizenship are with reference to multiple political and communal spaces in countries of origin and immigration as well as the transnational spaces of diasporas.

This brief review demonstrates the continued centrality of the nation-state as a research object in human geography. It also indicates that the questions that are asked and the approaches that are used have undergone important changes,
and especially that citizenship has become an important analytical tool for studying the concrete, context-specific, and changing meaning and politics of state–community relations within nation-states. Cultural and global changes have challenged the nation-state and its model of citizenship, and have been followed by increased attention to postnational, denationalized, and transnational forms of citizenship and to cross-scale, multisite, and transterritorial geographies of citizenship. Contemporary state–community relations in nation-states are thus characterized by an increasingly complex geography as institutions and political struggles for citizenship are constituted within and across multiple scales, spaces, and places.

SEE ALSO: Citizenship; Nationalism and geography; State, the; Transnationalism

References


Jackson, Peter, and Jan Penrose. 1993. Constructions of Race, Place, and Nation. Minneapolis: University of Minnesota Press.


Natural hazards and disasters

J.C. Gaillard  
*The University of Auckland, New Zealand*

The sharp increase in the occurrence of documented disasters worldwide over the past few decades has stirred growing interest within geography and other cognate social sciences. Geographers have given particular attention to disasters associated with natural hazards, the understanding of which requires blending knowledge of both the physical environment and the social fabric.

The definition and scope of both natural hazards and disasters have been debated. Some, especially within the North American geographical literature, have long considered natural hazards as the possible interaction between people and natural events (Tobin and Montz 1997, 5), while a disaster used to refer to the actual event significantly impacting society. The foregoing understanding of natural hazards, which was still prominent in the 1990s, has, however, progressively been replaced by a more consensual definition limiting natural hazards to natural processes and events which are possibly destructive to people and their resources, and disruptive of their activities (Wisner, Gaillard, and Kelman 2012, 21). In this sense, natural hazards only pose a threat to people who are vulnerable and lack capacities to face them. This interaction between hazard, vulnerability, and capacities mirrors the risk of a disaster or the risk that the occurrence of a natural hazard leads to casualties, infrastructure damage, or livelihood disruption – eventualities which are too important for the affected area and people to deal with on their own.

Such a definition of a disaster emphasizing the interaction between the natural environment and the social fabric has become central to geographical understandings of disasters as it refers to the essence of the discipline. In fact, most geographers’ efforts have been dedicated to better comprehending the causes of disasters, thus involving studies of both natural hazards and people’s vulnerability and capacities. In parallel, sociologists have focused on disruptions to the social fabric caused by disasters and how people actually react faced with these events, while psychologists have largely concentrated on disaster-induced trauma and anthropologists on long-term social and cultural change associated with disasters. All these other fields of scholarship have concurrently developed their own definitions of disasters, mirroring their disciplinary interests.

Apprehending the causes of disasters has stimulated much discussion among geographers who have debated the relative role of natural versus human factors in the increasing, but spatially unequal, occurrence of documented disasters. These debates have had a powerful influence on disaster risk reduction (DRR) as well as postdisaster reconstruction policies and practices.

**Trends in the occurrence of natural hazards and disasters**

All available databases recording disasters associated with natural hazards worldwide show a sharp increase in their occurrence from the
1980s onward. The most trusted database, EM-DAT, compiled by the Centre for Research on the Epidemiology of Disasters (www.emdat.be) records 11,632 events between 1900 and 2013, 9,504 of which have occurred since 1980, including some very high-profile events such as the 2004 Indian Ocean tsunami, Hurricane Katrina in the United States in 2005, the 2008 earthquake in China, the 2010 earthquake in Haiti, the 2011 earthquake and tsunami in Japan, and Typhoon Yolanda in the Philippines in 2013. The surge in the number of disasters is particularly pronounced in the less wealthy countries, especially those in Asia and Latin America, where 59% of the events recorded between 1900 and 2013 occurred. Meanwhile, Europe and Northern America seem to be less frequently impacted – disasters recorded in EM-DAT as 14% and 9% respectively. EM-DAT also shows that 87% of the disasters recorded worldwide between 1900 and 2013 are associated with hydrological and/or climatic events, with no significant disparities in the distribution of hazards across regions of the world. In addition, hydrological and/or climatic events are related to 88% of the casualties.

The severity of the impact of recorded disasters is increasing. The number of people affected has been surging over the past few decades, especially, again, in the less wealthy countries of Asia, Latin America, and Africa. Between 1980 and 2013, more than 6.2 billion people have been injured, forced to evacuate, or lost assets and resources in disasters, of whom 89% live in Asia, 6.5% in Africa, and 3.2% Latin America. In parallel, the number of people killed is also growing. EM-DAT shows that 2.3 million people died between 1980 and 2013, compared to 2.1 million individuals between 1900 and 1979. The economic impact of recorded disasters measured through the sole direct losses in monetary value is also surging. EM-DAT shows that cumulated losses since 1980 have reached US$2.5 trillion against US$87 billion between 1900 and 1979. The trend is exponential, as economic losses recorded since 2000 have reached US$1.6 trillion. Wealthy countries suffer the largest absolute losses, with Japan, the United States, and China topping the list of the most severely affected countries. However, less affluent countries, especially the smallest ones such as the Small Island Developing States (SIDS), record higher relative losses in comparison to their gross domestic product (GDP).

This brief overview of documented natural hazards and recorded disasters only mirrors the scope of existing datasets, which themselves merely reflect the interest of, on the one hand, international agencies involved in disaster response and, on the other hand, scholars engaged in disaster studies. In this regards, most databases recording disasters worldwide focus on large events which trigger international humanitarian attention and are known to a large enough number of people to stimulate research interest. For example, EM-DAT only records events which kill at least 10 people, affect at least 100 individuals, require the declaration of a state of emergency, and/or a call for international assistance.

In parallel with these large disasters, there are also a significant number of smaller events whose cumulative impact is supposed to be higher. These neglected disasters are partly captured in the alternative DesInventar database (www.desinventar.org), now managed by the United Nations International Strategy for Disaster Reduction. For example, DesInventar records more than 40,996 events, and 44,390 deaths, between 1914 and 2013 in Colombia alone. By contrast, EM-DAT records only 158 disasters, and 32,661 people killed for the same period. In Nepal, EM-DAT records 58 disasters between
1971 and 2007, while DesInventar lists 7263 events.

Small disasters are not only important for their cumulative impact and significance in the everyday life of those who are affected. They also challenge the common view of disasters as events which exceeds the ability of those affected to cope using their own resources. Because these events are overlooked or neglected by most institutional actors involved in disaster risk management (DRM), affected people are left on their own to cope and recover. Small disasters, in fact, constitute a category of events which fall between large disasters and the everyday crises associated with maldevelopment, that is, poverty and associated food insecurity, poor health, informal shelter. Figure 1 shows that small disasters, unlike everyday crises and large events, are overlooked by policymakers and people’s response is largely undocumented. They constitute what the United Nations’ 2009 Global Assessment Report on Disaster Risk Reduction labeled as “extensive risks” in comparison to “intensive risks,” which are associated with large disasters.

In addition, all available databases only provide a rather rough overview of the impact of disasters that tends to standardize data at regional and national scales. There is very little data indeed about the impact of disasters at the local level,
including across neighborhoods, genders, ages, and social statuses. This gap leads to an overall generalization of the impact of disasters to the detriment of a wide range of different local realities and experiences. It also affects the way disasters associated with natural hazards are explained and interpreted.

Dominant understanding and the hazard paradigm

Disasters have long been viewed solely through the lens of natural hazards and have thus been seen as prime concerns for physical sciences such as seismology, volcanology, climatology, geomorphology, hydrology, and meteorology. Up until the mid-twentieth century, the responsibility for the occurrence of disasters that include physical phenomena was therefore attributed to external natural forces or to the whims of angry or vindictive deities.

Rousseau was the first person known to attribute a disaster to human responsibility, in a 1756 letter to Voltaire about the previous year’s Lisbon earthquake and tsunami. However, it was not until the 1940s and White’s (1945) pioneering thesis on people’s adjustments to floods in the United States that the human dimension of disasters began to be widely accepted as part of a broader human ecology body of scholarship. However, although it addresses anthropogenic issues, the human ecology approach has continued to emphasize the contribution of natural hazards and has, in fact, further consolidated the so-called hazard paradigm. This paradigm stresses the importance of extreme (in magnitude) and rare (in time) natural hazards that exceed people’s ability to cope with them. Hazards are thus often considered extraneous to people and societies.

The extraneous and extreme dimension of natural hazards has led disasters being considered out of the regular social fabric. Scientists, institutions, governments, and media thus often mention extraordinary, uncontrollable, incredible, unpredictable, and uncertain phenomena, along with unexpected disasters and unscheduled and unanticipated damage. It is claimed that regions affected are unable to face such forces of nature, and such regions are often considered to be underdeveloped, overpopulated, uninformed, unprepared, and unplanned. Therefore, a clear border is delineated between regions of the world which are often struck by disasters and those that are supposedly safe. The construction of this divide is facilitated by the homogenizing nature of data available on the occurrence of disasters.

In this context, earth and climate scientists and engineers tend to focus on monitoring, predicting, and calculating probabilities and parameters for extreme natural hazards. In parallel, social scientists, led by geographers following White’s precedent, are interested in how people and societies perceive the potential danger and how they adjust to possible threats (Burton, Kates, and White 1978). Individuals and societies said to have a low perception of risk allegedly adjust poorly to possible threats. Conversely, people and societies considered to have a high-risk perception are assumed to adjust well to natural hazards. Factors that affect people’s perception of risk are hazard-related too (i.e., hazard magnitude, duration, frequency and temporal spacing, plus the recentness, frequency, and intensity of past personal experiences with hazards) and independent of the social environment.

Proponents of the hazard paradigm further stress the range of adjustment options known to those at risk. The larger the array of opportunities people are aware of, the broader their
portfolio of choices to adjust to natural hazards. Choices of adjustment therefore widen with technological progress and wealth (Kates 1971). Poorer people are therefore deemed to be less able to adjust to natural hazards and, in consequence, suffer the brunt of disasters. In many instances, the poor are said to aggravate the impact of disasters by deliberately choosing to live in hazardous areas and contributing to higher population growth, which leads to more casualties should hazards occur.

The study of people’s perception of risk relies on quantitative research methods, including the use of identical questionnaire surveys conducted in different regions of the world, following similar approaches, so that the outcomes are comparable (White 1974). These allow for the design of universal theoretical frameworks such as the choice tree of adjustment to natural hazard, which has been applied from individual to societal levels (Burton Kates, and White 1978). The choice tree distinguishes incidental adjustments to natural hazards from purposeful adjustments, which are further broken down into passive acceptance of losses to actions to reduce losses and initiatives toward choosing or changing land-use or location (Figure 2).

The hazard paradigm is further fueled by the dominant discourse on climate change. Uncertainties around the evolution of climate conditions constitute a powerful argument for considering “nature” as the major threat. Uncertainties are frequently associated with the probability of occurrence of rare and extreme natural hazards which should be addressed through scientific models and statistical probabilities. The contemporary focus on climate change thus reinforces a paradigm where nature is hazardous (even if exacerbated by human activity, as with climate change and many other hazards). In the face of such changing threats, people are expected to adapt to deal with them, which is very similar to the adjustment approach of the hazard paradigm. The vocabulary has shifted from adjustment to adaptation, but the meaning and scientific justification are largely the same.

The hazard paradigm and study of people’s perception and adjustment to natural hazards has provided a major breakthrough in the study of disasters. However, it has been widely criticized for being excessively deterministic, Malthusian, and technocratic to the detriment of the study of underlying social processes that make people vulnerable.

Alternative interpretation and the vulnerability paradigm

The role of natural hazards and people’s perception of associated risk in the occurrence of disasters has been progressively reconsidered since the 1970s through the work of geographers engaged in both political economy and political ecology research. For example, O’Keefe, Westgate, and Wisner (1976) have emphasized the root causes of people’s unequal vulnerability, or their susceptibility to suffer should potentially harmful natural hazards occur. The bottom-line evidence is that there is no disaster which causes all buildings in a particular place to collapse or all people to die. There are always buildings which withstand and people who survive.

Drawing on the precepts of an array of critical theories associated with the works of Marx, Weber and Sen, the proponents of the vulnerability paradigm argue that people affected by disasters are disproportionately drawn from the margins of society (Wisner et al. 2004). They are marginalized geographically and physically because they live in places exposed to natural hazards, which often coincide with locations
with inadequate services and contested tenure (e.g., informal settlers, shacks); socially and culturally because they often are members of minority groups (e.g., ethnic or caste minorities, people with disabilities, prisoners, and refugees); economically because they are excluded from the market (e.g., jobless); and, ultimately, politically because their voice is disregarded by those with political power (e.g., indigenous people, women, gender minorities, homeless, children, and the elderly). Disasters thus most frequently affect individuals with limited and fragile incomes, reducing their ability to deal with natural hazards, notably their ability to choose the location of their home and to invest in protective measures. Vulnerability and marginality also relate to inadequate social protection and limited social networks.

The unequal impact of disasters and people’s vulnerability to natural hazards reflect a lack of access to resources and means of protection. Lack of access does not mean that these resources and means of protection are not available locally. Most often they are available but access is limited.
to those with stronger economic, political, and social positions in society, thus reflecting an unequal distribution of power and opportunities. In cases of extreme marginalization, homeless people, migrants, traveling communities, squatters, gender minorities, lower untouchable castes, and other social groups whose citizenship and identity are legally and/or socially illegitimate, or at least not properly acknowledged through state identification systems, cannot even claim such access. This constitutes a further obstacle to protection and a driver of vulnerability. Therefore, the vulnerability–marginality nexus here goes far beyond poverty to also include issues of age, physical ability, gender, sexuality, race, ethnicity, caste, and religion.

People’s vulnerability to natural hazards varies in time and space and mirrors the nature, strength, and diversity of their livelihoods. The intimate relationship between livelihood and vulnerability explains why many people often have no other choice but to face natural hazards to sustain their daily needs. In many instances, threats to everyday needs, especially to food security, are almost always more pressing than threats from rare or seasonal natural hazards. The difficulty of meeting daily needs can further lead to environmental degradation, which often manifests itself in increasing natural hazards.

For the proponents of the vulnerability paradigm, people’s vulnerability to natural hazards therefore results from their limited ability to control their daily lives. In that context, disasters highlight or amplify people’s daily hardships and everyday emergencies associated with food insecurity, illnesses, fragile shelter, and poverty at large. Thus disasters cannot be considered as accidents beyond the usual functioning of the society (Hewitt 1983). Instead, disasters generally reflect development failures where the root causes of vulnerability have origins in other, usually contextual, development-related crises.

The root causes of disasters are grounded in an array of structural sociocultural heritages and political economy processes working at both local and international scales. These are further explained by deep-seated historical processes, colonial and postcolonial legacies, and powerful ideological frameworks (Watts and Bohle 1993). Ultimately, the root causes of disasters most often lie beyond the hands of those who are vulnerable and suffer.

Understanding people’s vulnerability requires fine-grained studies relying on qualitative, often ethnographic, research methods through undertaking what Chambers (2007) has called “anthropological particularism.” If the outcomes of the growing number of studies conducted under the vulnerability paradigm over the past three decades are not directly comparable, some powerful frameworks have been designed to confront them. The most influential of these is the pressure and release (PAR) framework popularized by Wisner et al. (2004). PAR tracks down the factors of vulnerability from the unsafe everyday lives of those at risk of dynamic pressures and, ultimately, structural root causes (Figure 3).

The foregoing tenets of the vulnerability paradigm constitute a radical departure from previous interpretations of natural hazards and disasters. It has been criticized for putting too much emphasis on the social to the detriment of the natural component of disasters. In addition, many scholars working in industrialized wealthy countries have found it more applicable to the less affluent regions of the world, where this approach emerged in the first place. In its early years, the vulnerability paradigm was also challenged for overly stressing people’s weaknesses to the detriment of their intrinsic capacities to face natural hazards and disasters.
From vulnerability and suffering to capacities and actions

Those who are vulnerable to natural hazards are neither “victims” nor helpless in time of hardship. They always display some form of ability to face threats and disastrous situations. Late in the 1980s, practitioners coined the term “capacities” (Anderson and Woodrow 1989). Since then it has been widely used through the practice of vulnerability and capacity analysis (VCA) – a core dimension of many DRM projects derived from Anderson and Woodrow’s framework for assessing vulnerability and capacities. However, the concept itself (not in association with other concepts such as in the expressions “adaptive capacity,” “capacity to resist/to face/to recover,” etc.) has been poorly defined in the academic world. Only recently have geographers pushed for its consideration in order to balance the sole negative dimension of vulnerability.

Capacities usually refer to the set of knowledge, skills, and resources people resort to in dealing with natural hazards and disasters. These capacities include, for example, kinship...
and solidarity networks, remittances, traditional medicines, vernacular architectures, experiences and knowledge of past hazardous events, traditional water management, hazard-resistant crops, knowledge of biodiversity resources, ability to hunt and fish, local leadership, and flexible decision-making processes. Figure 4 organizes these forms of knowledge, skills, and resources into arbitrary categories associated with particular strategies and outcomes when confronted with natural hazards and disasters. The concept of capacity further encompasses the ability to claim, access, and use these knowledge, skills, and resources.

Importantly, capacities are not the opposite end of vulnerability on a single spectrum, because highly vulnerable people may display a large array of capacities and vice versa. In fact, while the root causes of vulnerability are largely exogenous to those at risk, capacities are often rooted in resources which are endogenous to them. They are often the extension of everyday practices and roles within the society rather than extraordinary measures taken to face rare and extreme events.

The concept of capacities ultimately recognizes that those at risk take actions in facing natural hazards and disasters. In fact, they are often considered as the first line of responders. In times of disaster, outside assistance arrives, at best, hours or days after the event even though it is well known that the initial few hours are crucial in saving lives and livelihoods. Evidence collected by sociologists suggests that a very large majority of postdisaster survivors are rescued by their friends, kin, or neighbors who are on the spot at the time of an event and who rely on local knowledge and available resources and skills (Quarantelli and Dynes 1972).

Although the increasing emphasis given to people’s capacities emerged as a spin-off from the vulnerability approach and has not led to the emergence of a new paradigm per se, it still marks a significant evolution in the study and understanding of natural hazards and disasters. Because they are often embedded in people’s everyday life and culture, capacities are difficult to apprehend for outside researchers and practitioners. For this reason, the analysis of capacities has largely relied on participatory approaches led by those at risk along the line of what Chambers (2007) calls participatory pluralism, which includes tools, methods, attitudes, and behaviors geared toward fostering the contribution of a large array of diverse and usually excluded local actors in analyzing and finding solutions to problems which affect them.

Yet, much research has still to be conducted on the concept of capacity. Its contours have to be more clearly defined, notably in the context of other cognate concepts. Capacities indeed relate to the concept of subculture put forward by sociologists in the 1960s and the 1970s. Disaster subcultures include cultural patterns existing in particular places for overcoming social and nonsocial problems associated with natural hazards and disasters. Disaster subcultures guide the behavior of individuals and groups before, during, and after the occurrence of natural hazards. The ways in which capacities are mobilized in times of crisis also reflect coping strategies. Coping strategies refer to the manner in which people use available resources to overcome the unusual and adverse conditions of a disaster. Capacities also relate to many interpretations of the broader and controversial concept of resilience, which is increasingly used to similarly emphasize people’s ability to face natural hazards and disasters.

Reducing the risk of disaster

Since the early work of White in the 1930s and 1940s, geographers, more than any other
social scientists, have committed themselves, and significantly contributed, to enhancing policies and practices of DRM, in particular risk reduction, at all scales (Wisner, Gaillard, and Kelman 2012). Recognizing natural hazards as well as people’s vulnerability and capacities is essential for reducing the risk of disaster. In principle, risk reduction must therefore include...
actions to address the root causes of people’s vulnerability as well as initiatives to enhance their capacities. Because these are, respectively, largely exogenous and endogenous to those at risk, risk reduction requires actions from both the outside (or from the top down) and inside (or from the bottom up). Actions from the outside should be geared toward granting access to resources to those who are vulnerable as well as strengthening their everyday livelihoods. Actions from the inside should facilitate and improve the utilization of local knowledge, resources, and skills. In parallel, risk reduction should also consider actions to prevent hazards from occurring or, should they occur, ensure that they do not reach neighboring settlements nor affect people’s livelihoods. These could be both initiated from the inside and outside through bottom-up and top-down initiatives.

Risk reduction should therefore be an integrative process involving a large array of unequally powerful stakeholders active at different geographical scales (Gaillard and Mercer 2013). These stakeholders should include people at risk including the most marginalized, schools, faith groups, local governments, nongovernmental organizations (NGOs), scientists, the private sector, national governments, and international organizations. The role of both those at risk and governments is crucial. Disasters are local events which first and foremost affect local people. No one is, therefore, more interested in reducing disaster risk than those whose survival and wellbeing are at stake. Meanwhile, achieving large-scale outcomes requires appropriate institutional frameworks at the state level and the wide and thoughtful implementation of these by local governments. Internationally, these have been timidly recognized by policy guidelines such as the Hyogo Framework for Action 2005–2015 (titled “Building the Resilience of Nations and Communities to Disasters”) and the most recent Sendai Framework for Disaster Risk Reduction 2015–2030.

In practice, however, the policies and practices of many countries and actors reflect the influence of the hazard paradigm. These policies and practices are primarily geared toward the extreme and rare dimensions of natural hazards and often reflect strategies designed to organize battles in time of war. Indeed, in many countries, disaster policies are handled by the army or civil protection institutions, which rely on military chains of command and treat natural hazards as enemies to fight against. Most often these actions, designed at a global scale, include specific, technocratic, command and control measures such as engineering structures to control natural hazards, technology-based monitoring and warning systems to anticipate the occurrence of hazardous events, hazard-based land-use planning, and unilateral risk awareness campaigns. Such frameworks further rely on top-down transfers of knowledge, technology, and experience from the wealthiest and most powerful regions of the world, which are supposed to be safer because of their larger technological and economic resources, to poorer and less powerful regions, which are considered more vulnerable. These policies and practices draw on the abrupt and overgeneralizing geography of risk, opposing the safe North to the dangerous South (Hewitt 1983). Throughout history, wealthier countries have thus developed policies to assist poorer countries in addressing disaster-related issues because of the latter’s alleged inability to cope on their own (Bankoff 2001).

These dominant, hazard-focused, policies have largely failed to address the root causes of people’s vulnerability. The command and control nature of these frameworks does not yield easily to the amorphous and contextual dimension of people’s everyday life and vulnerability, which requires continuous attention and investments
rather than punctual consideration and sporadic allocation of resources. In consequence, the dominant strategies to reduce the risk of disasters have given paramount importance to large events to the detriment of smaller disasters, thus overlooking a large component of people’s hardship.

Alternative initiatives emphasizing local actions, following bottom-up frameworks, emerged in the 1970s and eventually spread widely in the 1980s and 1990s under the impetus of an increasing number of NGOs and other civil society organizations. These have been referred to as “community-based disaster risk reduction” (CBDRR) by policymakers and practitioners. CBDRR fosters the participation of those who are vulnerable according to both the evaluation of disaster risk and in ways to reduce it. It aims to empower those at risk with self-developed and culturally, socially, and economically acceptable ways of dealing with natural hazards (Maskrey 1989). It relies heavily on participatory learning and actions (PLA), which include all approaches, methods, and attitudes designed to empower those at risk to share, analyze, and enhance their knowledge of disaster risk and to plan, implement, monitor, assess, and reflect toward risk reduction. CBDRR thus gives prime importance to people’s capacities and preparedness in facing possible events through the design of early warning systems as well as evacuation and crisis management plans.

More often than not, CBDRR activities have had only a limited impact on reducing people’s vulnerability, which largely lies beyond the scope of such projects. In some extreme cases, the proponents of CBDRR reject all support from the outside/top down. In addition, vulnerable people’s participation in risk reduction is often skewed to serve the interests of outside actors who need to justify the involvement of local people in activities they have designed beforehand.

In such instances, participation is often conceived as an outcome that is accountable to Western funding agencies, rather than a process where accountability is downward to those at risk.

Ultimately, it is difficult to integrate the needed and positive outcomes of actions from the inside or the bottom up with initiatives from the outside or the top down. This gap is considered a major impediment to sustainable risk reduction. It reflects difficulties in appraising and integrating different forms of knowledge, resources, and skills as well as the continuing dominance of technocratic institutional frameworks, and a scarcity of appropriate tools to foster dialogue and establish trust between all stakeholders. Dialogue and trust are essential for all diverse stakeholders to recognize, value, and integrate the vulnerability and capacities of those at risk, and to make sure that actions to address and enhance these are integrated. Most of those who are vulnerable, including the most marginalized, know what their needs and resources are in facing hazards and disasters. The issue for these groups is usually to make their vulnerability and capacities tangible and recognized by others. In that sense, they should interact with those with power in the larger society, otherwise risk reduction initiatives remain clustered and fail to address the unequal power relationships which lead to disasters.

Postdisaster reconstruction and further opportunities to foster risk reduction

Should risk reduction policies fail and disasters occur, the subsequent process of reconstruction is often expected to provide further room for reducing the risk of future disasters. Opportunities for risk reduction following disasters are said to stem from increased awareness of disaster risk which opens a window for change. It is
therefore recommended to “build back better” (Lyons, Schilderman, and Boano 2010).

Christoplos (2006) identifies pathways which, in theory, favor positive change toward risk reduction in the aftermath of disasters: (i) disasters usually enhance individuals’ and the whole society’s perception of the risk associated with natural hazards, thus providing space for critical reflection on the causes of disasters; (ii) disasters thus often reveal the shortcomings of past development strategies and may lead to a better understanding of the underlying factors of disaster risk; (iii) disasters expose institutional weaknesses such as corruption, lack of human resources, and weak institutional structures, as well as weakening and discrediting those actors whose actions have contributed to creating disaster risk; (iv) in consequence, there usually is a stronger political will for change in the aftermath of disasters; (v) disasters clear away weak and vulnerable infrastructure and create the opportunity to construct stronger buildings; (vi) the frequent influx of financial and other resources which immediately follows disasters often prove important for initiating large-scale reconstruction.

Unfortunately, there is very little available evidence of postdisaster reconstructions which have actually significantly reduced the risk of future disaster. Most often, greater attention is given to short-term needs such as shelter, food, water, and sanitation over longer-term risk reduction. This gap largely results from postdisaster DRR planning allegedly requiring time and outside expertise to implement, whereas meeting survivors’ immediate needs supposedly entails quick, pragmatic, and straightforward actions. In fact, to meet large-scale short-term needs as quickly as possible, postdisaster reconstruction often relies on centralized and top-down policies which overlook day-to-day, collaborative, small-scale actions and local resources which are essential to risk reduction.

Overcoming this gap, again, requires integrating both actions from the outside/top down and initiatives from the inside/bottom up. Rebuilding following a disaster is often considered a long-term process, especially if it is to foster DRR and especially where it relies on external intervention, which requires often lengthy preliminary needs assessment and careful planning. However, a quicker alternative is to rely on local resources, knowledge, and skills, that is, those of affected people and institutions. Local actions need to be supported by appropriate initiatives from outside stakeholders for tasks which cannot be achieved by relying solely on local resources, for example clearing massive debris and reconstructing stronger large infrastructure, including scientific knowledge of hazards.

It is widely acknowledged that three key pathways facilitate the integration of actions from the inside/top down and from the outside/bottom up following disasters (Alexander et al. 2007). First, it is recommended that reconstruction begin immediately after the disaster, simultaneously with relief operations. Rebuilding should also be designed to balance change, which is critical to integrate DRR, with continuity, which is essential to ensure that the links between people and their territory be maintained, thus accelerating the recovery process. Too much change requires time and resources which are often unavailable locally. In this view, geographical relocation away from hazard-prone areas has often failed to foster the recovery of affected people. Integrating bottom-up and top-down actions toward DRR also requires decentralized, flexible, and straightforward decision-making processes. For this reason, the creation of centralizing superstructures dedicated to postdisaster reconstruction, which has frequently followed large disasters, has proved a hindrance for integrating local knowledge, resources, and skills.
NATURAL HAZARDS AND DISASTERS

Following these pathways toward risk reduction proves difficult when reconstruction is planned ad hoc following a disaster. Pre-disaster reconstruction planning is an alternative currently pursued by some local governments. It facilitates the mutual integration of reconstruction within risk and risk reduction within reconstruction, and a concurrent shift from a sequential approach to risk reduction and disaster reconstruction to a more holistic model of DRM which aims to transform development trajectories in line with a more sustainable future.

Future research directions for enhancing disaster risk management

The study of natural hazards and disasters has become a major component of geographical research. The field has grown so large and diverse that, since the mid-1990s, the initially clear divide between the hazard and vulnerability paradigms has become blurred. Some even suggest the emergence of a third, hybrid paradigm fostering the integration of natural hazard research, risk perception inquiries, and vulnerability studies.

Currently, people’s vulnerability, as well as their capacities, are often considered from a taxonomic and quantitative perspective in direct relation to natural hazards though the use of related concepts such as exposure (to natural hazards) and indicators like demographic data. Such an approach to natural hazards and disasters is much less grounded in geographical and broader social theories. It increasingly relies on the use of tools like geographic information systems (GIS) and remote sensing.

In addition, more recent scholarship regarding climate change and its impact on people and societies, which borrows many concepts from natural hazard and disaster studies, has emerged as a parallel stream of research. In this context, the long tradition of natural hazard and disaster research within geography and cognate social sciences has to be acknowledged so that climate change studies does not simply reinvent the wheel. This is of particular importance in the context of policies and practices of climate change adaptation, which are increasingly considered in tandem with DRM. The integration of both streams of research, policy, and practice is one of the major challenges of the coming decades in the aim to foster sustainable development.

SEE ALSO: Climate change adaptation and social transformation; Environment and everyday life; Environment and resources, political economy of; Environmental hazards; Human ecology; Livelihoods; Political ecology; Social resilience and environmental hazards; Social vulnerability and environmental hazards

References


The concept of resources occupies a major place in understandings of human environmental relations. Etymologically related to the Latin resurgere, meaning “to rise again,” the term “resource” conveys a sense of (re-)empowerment and opportunity, but also dependence and vulnerability. Without a resource, one may not be able to stand back up after falling. This fear is often reinforced by adjectives attached to resources, such as “critical,” “vital,” “crucial,” “national,” or “strategic.” Frequently deployed in mainstream geopolitical narratives of “resource wars,” these adjectives cast resources as the condition of (continued) “civilization” over which competing groups are bound to fight. Resources vary in their spatial location, relative abundance, physical characteristics, technologies of extraction and transformation, use, social and environmental impacts, and economic value. Resources and their relative availability or scarcity are, in other words, both (physically) material and (socially) relational. A simple example is diamonds, for which desirability, scarcity, and high prices are more functions of cultural construction and corporate control (by companies such as De Beers) than of intrinsic physical properties. Resources are also understood as objects influencing social relations. Such an idea does not mean ascribing a fixed and deterministic sense of agency to resources; oil itself does not declare war. Rather, it is about recognizing what Arjun Appadurai (1988) calls “the social life of things”: that resources both reflect and contribute to material cultures. Demand for resources is driven not simply by human needs but also by the social practices that resources, as objects, enable and influence. Similarly, human conflicts differ widely, and violent forms of conflicts include acts not simply of physical violence, but also of structural and symbolic forms of violence. As such, talking of “natural resources and human conflict” is problematic: first, because it reproduces a dualism between natural and human categories that fails to acknowledge the intrinsically social dimensions of resources, even if these are not obviously socially “produced” in the ways crops (“domesticated nature”) and manufactured goods (“transformed nature”) are; second, because it can deflect attention from human conflicts against nature — not simply the numerous environmental impacts of wars, but also the violence perpetrated against ecosystems by everyday consumption and through a prevalent utilitarian perspective of the nonhuman treating it solely as a “resource.”

**Historical perspectives**

Natural resources have long attracted attention in the study of human conflicts; accounts of resource plunder and destruction within war narratives date back to at least 3600 BCE. From a historical perspective, scholarly studies of resources and early warfare in preagricultural societies have largely focused on the role of material self-interest, the forms of conflict, the organization and the sedentarization of social groups, as well as the relative availability, density,
and predictability of resources. Ethnographic, archaeological, and evolutionary and comparative social ecology studies associate early warfare with the territorial control of abundant resources (mostly food) and uncertainty about resource access. Contrary to neo-Malthusian narratives, higher population densities and resource abundance would result in territorialization, with conflicts over resources occurring as a result of trespass or intrusion, while resource scarcity and low population densities would result in mutually beneficial cooperation rather than conflict. The transition to agriculture and exploitation of resources that reshape relations between human groups and the “natural world” are often understood as one of the main factors in the frequency of warfare. In common with the classical period, contemporary Western geopolitical perspectives on resources have been dominated by the equation of trade, war, and power. Following gold-focused conquest, colonial plantation economies of tropical slave-produced commodities became the core of Western imperial extension. Duties on sugar, tobacco, cocoa, cotton, coffee, and opium provided “modernizing” states with the finances to open up new markets through warfare. Since sea power itself rested on access to timber, naval timber supply became a critical preoccupation for major European powers from the seventeenth century onward, a situation comparable to the case of oil in the twentieth century. Given the strategic role of resources, concerns for resource scarcity and war received considerable attention from contemporary scholars such as Malthus, who saw “vices of mankind and able ministers of depopulation” usefully staving off (rather than resulting from) food scarcity.

The importance of resource flows for industrialization and militarization – most notably coal, iron, and later oil – reinforced an ideology of resource competition between European powers, found notably in the flurry of studies over access to raw material distribution during the first half of the twentieth century, and especially between the two world wars (Westing 1988). The growing assertiveness of “Third World” states during the 1960s and 1970s transformed the political landscape of sovereignty over natural resources and provided a (new) twist on ideologies of resource wars, with, for example, some Arab states leveraging their oil production in the form of an export embargo against some pro-Israeli states.

By the 1970s, broader geopolitical conceptualizations of security incorporated issues such as population growth, environmental degradation, and social inequalities in poor countries. The ensuing concept of “environmental security” came about to reflect ideas of global interdependence, illustrated through the debates on environmental “limits to growth,” political instability supposedly caused by environmental scarcity in the South, and more recently the consequences of climate change. The concept, however, represented for some a skewed and controversial “securitization” of environmental issues, calling for “military” and “international development” solutions, and constructing biased identities and narratives of endangerment in the Anthropocene that often blamed the poor (Dalby 2009).

With the end of the Cold War, a view emerged that violent scrambles for resources among local warlords, regional powers, and international actors were a major feature of contemporary conflicts, particularly given the supposed declining role of ideology in regional or local conflicts. Preoccupations with (militarized) “resource supply security” resurfaced again as commodity prices rose after 2002. China challenged US hegemony, and relations with resource-exporting countries became (selectively) rearticulated in the context of the “war on terror.” Most
contemporary accounts of resource wars are associated with a combination of rapidly increasing demand for raw materials, growing resource shortages, and contested resource ownership (Klare 2008). Many accounts of resource wars tend to pathologize conflicts and politically delegitimize protest and popular (armed) resistance, to criminalize small-scale mineral exploitation by local communities and regional migrants, undermining both livelihood coping mechanisms, and to prioritize certain types of economic growth (such as large-scale mining or logging) over local livelihoods and environmental practices.

Main explanations

Mainstream accounts of natural resource and human conflicts have thus relied on two often intertwined explanations. The first explanation views resources as a motivational factor for human conflicts. Raiding, looting, pillaging, grabbing, capturing, annexing, and conquering all carry a sense of violent dispossession and appropriation of resources. Such a motivational dimension can be mistaken for the consequences of social behavior during or after the conflict; state-led military annexation and house looting by individual soldiers are both violent dispossession, but they differ in scale, intentionality, means, and outcomes. The second explanation is that some resources are crucial factors for the conduct of warfare itself, which thereby achieve “strategic” status. Resources that are strategic for warfare have long preoccupied military planners, notably through anticipating wars to obtain control over those resources. In the extreme, a war would be “pre-emptively” conducted to access resources necessary for warfare. Accessing such strategic resources and denying access to potential enemies have constituted one of the many “great games” of military strategists. But the strategic and security dimension of a resource needs to be considered in light of the corporate interests that can motivate or at least benefit from such association. These two main explanations – resources as loot and resources as military means – continue to dominate much of the media, policy, and scholarly literature on resources and human conflicts, mostly through conventional geopolitical perspectives which define resource wars as armed conflicts revolving around the control of “critical” raw materials. To paraphrase the Prussian war thinker Carl von Clausewitz, resource wars are, from this perspective, the continuation of resource politics by military means. Mostly used in reference to interstate conflicts over the control and supply of “strategic resources,” the concept gives way to a narrow and militaristic notion of “resource security” (and in particular “energy security”). The term also became widely used in the application of the metaphor “trade wars,” such as in the examples of bananas or wheat. Finally, the term has been applied to describe popular struggles against large-scale resource exploitation projects and neoliberal reforms in resources and in public utility sectors.

Four main theoretical perspectives have conceptualized natural resources and human conflicts relations (Korf 2011). The first is Hobbesian, and conceptualizes resource conflicts as rational individualism in the absence of authority. From this perspective, life in the state of nature is not only “solitary, poor, nasty, brutish and short”, but also centered around competitive zero-sum game resource grabs. The second is neo-Malthusian, which combines the Hobbesian perspective with the anxieties of early urbanization, industrialization, and democratization to view “modern” progress and demographic growth as overtaking nature’s pace and resulting in scarcity-induced resource conflicts. The third
is positivist and suggests that resource opportunities affect conflict feasibility. Here *homo economicus* rationalizes the losses and gains of conflict over resources. Finally, the Schmittian perspective views historically complex identities reified in part through resource conflicts, whereby resources and conflicts over resources come to constitute identity narratives and to reproduce conflicts. The neo-Malthusian perspective has by far received the most attention within geography. David Harvey (1974, 256) warned of the “profound political implications” of supposedly ethically neutral scientific discussions of the population–resources relationship, especially a projection of neo-Malthusian views that invited “repression at home and neo-colonial policies abroad.” Contemporary debates about natural resources and human conflict mostly use three distinct yet overlapping concepts to capture different facets of resource-related conflicts (Le Billon 2012). First, the *resource curse* argues that resource wealth and dependence makes a society more vulnerable to conflict by weakening governing institutions, reinforcing authoritarianism, and undermining economic performance. The second, *resource struggles*, suggests that resource control and exploitation tend to increase the risk of conflict, including through competition, dispossession, and grievances over access and control of natural resources, and the distribution of costs and benefits from their exploitation. The third, *conflict resources*, recognizes resources as providing the financial means to sustain armed conflicts and potentially motivating belligerents to pursue hostilities. Overall, resource sectors may contribute to shaping greater vulnerability to, risk of, and opportunities for armed conflicts, most notably high-value resources exploited in resource-dependent countries already affected by ethnoreligious divides.

**Political ecology contributions**

In contrast to most studies focusing on the material characteristics of resources and their relative scarcity or abundance, anthropologist Arturo Escobar (2006, 9) has rightly pointed at the importance of accounting for cultural differences in explaining resource-related conflicts, for “many communities in the world signify their natural environment, and then use it, in ways that markedly contrast with the more commonly accepted way of seeing nature as a resource external to humans and which humans can appropriate in any way they see fit.” In this respect, political ecology approaches have emphasized contextualization and multiscalar relations, pointing also to the specific material and social dimensions of resources. Yielding nuanced and historically grounded analysis of uneven power relations and conflicts around natural resources, political ecology approaches seek to understand violence “as a site-specific phenomenon rooted in local histories and social relations yet connected to larger processes of material transformation and power relations” (Peluso and Watts 2001, 5).

As such, political ecology approaches make several contributions. First, it reconceptualizes scarcity, abundance, and dependence both historically and spatially, by considering how uneven power relations and resource entitlements reflect the antagonizing effects of conflicts on social identities, and by demonstrating how violence reshapes conditions of access and control over resources. Second, it questions not only at which location, but also at what scales, conflicts are occurring, and recognizes the chronic and multiscalar character of many resource-related conflicts. Third, it gives greater attention to both the material and the discursive dimensions of ecological processes and resource sectors, notably the “regimes of truth” that sustain and
seek to legitimate resource-based processes of capitalist accumulation (notably “accumulation through dispossession” in the form of enclosure of commons and other exclusive rights of access). Fourth, it accounts for a broader range of violence than geopolitical and political perspectives, thereby understanding different types of conflicts and multiple forms of violence. Fifth, it recognizes resources as a material object influencing sociomaterial practices, and as a social subject endowed with a set of character(istic)s through “spokespersons” such as campaigners, local communities, or resource companies granting them a voice; this combination of material enablement and social voice constitute a form of (indirect) agency, as recognized in the concept of the actant, from the perspective of actor-network theory. Sixth, it connects temporal and spatial realms, notably through site specificity and multiscalar analyses between resources and conflicts, helping to identify connections, actors and their motivations, and power relations – political ecology approaches allow for some degree of accountability beyond the immediate perpetrators of physical violence. Connections between resources and various forms of violence, in this regard, not only include commodification processes (i.e., how “things” become resources or commodities defined by their use and exchange values), but also fetishization processes (i.e., how imaginative aspects of resource production and consumption affect power relations and associated forms of violence). Many studies of natural resources and human conflicts are moving beyond simplistic conceptions of resource scarcity, abundance, and dependence. Rather, they consider how uneven resource entitlements and patterns of resource exploitation and consumption reflect and exacerbate the antagonizing effects of human conflicts, as well as informing the types of violence that are reshaping conditions of access and control over resources. This suggests mixed methods including historical and political economy analyses, as well as ethnographies of power relations and local perceptions of resources and violence.

SEE ALSO: Actor-network theory; Anthropocene and planetary boundaries; Energy security; Environment and resources, political economy of; Environmental (in)security; Mining and mineral resources; Natural resources; Political ecology; Resource curse; Resource extraction; Scarcity; Violence; War

References

Natural resource products are a significant component of the world’s economy. They account for one-third of world traded goods, and their share has been increasing over the last decade (WTO 2013). Natural resource products are particularly important for the least developed economies where they account for 78% of exports and 22% of gross domestic product (GDP) (WTO 2013). Natural resources provide significant nonmarket benefits in the form of ecological services, which are estimated to exceed the total value of the world’s GDP (Costanza et al. 2014). Natural resource development can also be a major contributor to environmental problems such as climate change.

This entry assesses issues and challenges of this important sector of the world economy. Dependency and comparative advantage paradigms and the efficacy of alternative development strategies are assessed. Environmental impacts of resource development are examined and a case study of tar sands development in Canada, which is among the world’s largest resource development projects, is analyzed. The entry concludes by identifying research questions that need to be addressed to better understand the role of natural resources in achieving sustainable development.

Natural resources and economic development

The starting point for understanding the role of natural resources in the development process is the “staple theory.” The staple theory originated with the work of Canadian economic historians Innis and Mackintosh and was developed into a more explicit model by Watkins (1963) and Hirschman (1981) to explain development of regions with a plentiful supply of natural resources. The staple theory bears similarities to North’s export-led growth theory (1955) and to Rostow’s (1960) stages of growth theory.

Staples are defined as natural resources and natural resource-related manufacturing. Traditional staple industries include agriculture, forestry and forestry-based processing such as lumber and pulp and paper, hunting and fishing, mining and mining-related processing, and energy.

Staples stimulate economic development by extraction of the natural resource and by a series of linkages that can be divided into four categories: forward linkages, defined as processing of the natural resources prior to export; backward linkages, defined as the production of inputs such as machinery and infrastructure used to extract the resource; final demand linkages, defined as production of goods and services required to support workers employed in the staple sector; and fiscal linkages, defined as the spending of rent generated by the natural resource.

The fiscal linkage deserves special mention because it is a linkage unique to natural resources.
Natural resources generate rent, defined as a return above normal returns to labor and capital. Capitalized rent is the market value of the in situ natural resource endowment and is a special dividend available to the owners of the resource above what would be generated in other economic development activities. Rent may be received as an ongoing payment tied to the rate of resource extraction or as a lump sum payment for the right to extract an in situ resource.

The rate of development in the staple region is determined by a number of factors that can be divided into endogenous factors under the staple region’s control and exogenous factors beyond the region’s control. An important exogenous factor is the quality and quantity of the natural resource. The quality of the resource determines the cost of extraction, which in turn determines the rate of development and magnitude of rent, while the quantity of the resource determines the longevity of resource extraction. A second exogenous factor is the production function of the staple, which determines the quantity of labor and capital required to produce the product, which in turn determines the linkages. Products such as fish and agricultural products may require little in the way of processing prior to consumption while other products such as minerals and some forestry products may require significant processing. Backward linkages also vary depending on the inputs required to extract and process staples. Another significant exogenous factor is external demand for the staple, which sets an upper limit on the size of the staple industry. World price for natural resources is also largely exogenous, but can be influenced by supply-side cartels such as OPEC (Organization of the Petroleum Exporting Countries).

Endogenous factors affecting the contribution of natural resources to development include resource management policies such as regulations governing the rate of extraction, ownership of the resource, development of linkages, collection and distribution of rent, provision of infrastructure, environmental regulations, and macroeconomic policies such as exchange rates.

The spatial distribution of extraction activities is determined by the location of the natural resource. A distinction is normally made between point resources such as mining and energy that are located in a specific place and diffuse resources such as agriculture and forestry that are spread across larger areas. The spatial pattern of point resources is normally small single industry towns in remote locations, or “fly in” camps if the life expectancy of the resource does not justify investment in long-term infrastructure. The spatial pattern of diffused resources such as agriculture is low-density settlement complemented by a hierarchy of more concentrated urban centers, a pattern that gave rise to central place theory. The spatial pattern of linkages is also determined by exogenous factors such as the production process that determines the optimal economic location of processing. Many forward linkages such as sawmilling, for example, are located close to resource extraction to minimize transportation costs because they are weight-losing activities. Other processing activities such as metal fabrication may be located in larger centers outside the staple region where costs may be lower due to economies of scale and access to other required inputs and markets.

A distinction can also be made between renewable resources that can sustain long run settlements and nonrenewable resources that may not justify permanent settlements. These spatial patterns can change over time with changes in the production function. Agriculture, for example, has experienced increased economies of scale that have led to larger farms and rural depopulation as capital is substituted for labor. Spatial patterns are also influenced by government policies impacting the location
of resource activities such as land-use controls, subsidies, provision of public infrastructure, and other regulations such as policies requiring local processing of natural resources.

Staple theory also helps explain the rise of political and institutional structures and public policy. The creation of countries such as Canada, for example, is often explained by the need to create a larger political unit to control and exploit the vast array of natural resources by building major transportation infrastructure and imposing order and controls conducive to resource extraction. The building of railways, creation of property rights, and the maintenance of social order are all nation-building endeavors necessary for resource exploitation. The location of natural resources is also a major factor stimulating conflict among jurisdictions vying for control of strategic resource deposits (Le Billon 2001).

Challenges of natural resource development

According to staple theory, an abundant supply of staples is an important asset providing regions with a significant advantage in the development process. More advanced regions seeking natural resources can provide demand, capital, technology, and entrepreneurial skills that stimulate more rapid development than would occur in a domestically oriented economy constrained by its internal rate of savings, market, and development capacity. Rent provides an extra return above normal returns to production. Stimulated by external demand and generation of rent, the staple region can prosper and diversify from its initial reliance on natural resources as it passes critical thresholds required to support more advanced industries with higher economies of scale.

Staple theory cautions that development of natural resources can also face inherent challenges that can offset the advantages (Watkins 1963; Auty 1995; Gunton 2003; Ploeg 2011). Development is constrained by the natural limits of the resource base. Nonrenewable resources extraction costs can increase as the most productive resources are exploited. Ultimately nonrenewable resources can run out, resulting in termination of production. Renewable resources can be exhausted by excessive rates of harvesting and, even if they are harvested on a sustained yield basis, extraction is constrained by the upper limit of the renewable resource’s biological productivity.

Resource development is often capital intensive: it requires high fixed costs to develop because the extraction of resources requires expensive geographically fixed investments at the site of the resource and infrastructure such as railways and pipelines required to transport the resource to markets. These “lumpy” and immobile investments create economic rigidity that impairs resiliency of the staple region to adjust to changing market conditions.

The rigidity in staple regions makes them particularly vulnerable to boom and bust cycles that characterize resource commodity markets. Resource booms stimulate a high rate of investment in staples fueled by excessive optimism that stimulates higher cost development, often financed by debt. The boom can create shortages of labor and capital that inflate costs and constrain development of other sectors of the economy. The rise in exchange rates resulting from a resource boom can produce the so called “Dutch disease,” characterized by contraction in nonstaple sectors suffering from reduced competitiveness due to currency appreciation and inflated costs.

The staple boom is followed by an inevitable downturn that can cause long-term economic
NATURAL RESOURCES DEVELOPMENT

stress. Export earnings fall, undermining the viability of high-cost resource development projects and the ability of the staple region to service external debts incurred to finance the investment boom. The recent write-off of $50 billion in uneconomic investments by the world’s major mining companies is a case in point (Koven 2013). The nonstaple sector is unable to compensate for the staple downturn because it has been weakened by the currency appreciation and inflated costs during the staple boom. The result is a legacy of uneconomic surplus capacity and debt that imperils the long-term economic health of the staple region.

Another challenge for staple regions relates to the corporate organization of the resource development. Some staple industries such as farming are smaller-scale locally owned enterprises, sometimes referred to as the entrepreneurial model. But many staple industries are large, capital-intensive, export-oriented industries that can only be undertaken by entities with substantial financial resources and management skills. Consequently, many staple industries, particularly in the energy and mining sectors, are owned and operated by large multinational corporations or state enterprises that control a major proportion of world production. Even the smaller locally owned enterprises may be indirectly dependent on large corporate enterprises that purchase their output. This can impose an additional impediment to development in staple regions because large multinational corporations may be less likely to locate higher-order functions such as processing, management, research and development, and equipment production in the staple region because these activities may already be established in the home location of the enterprise. Large multinational corporations can also extract rent from the staple region by repatriation of profits using intracompany transfer pricing to artificially reduce the export price, thus making it difficult for the staple region to collect rent (Gunton 2003).

The successful development of staple economies is very much dependent on government policies to collect rent and manage development in a sustainable manner. But staple industries are able to use their dominant position in the economy to pressure governments to implement policies to meet the needs of the staple sector, such as subsidizing infrastructure, reducing resource industry royalties, and weakening environmental and other regulations. These policies can retard development by expediting the leakage of rent, increasing the rate of extraction beyond sustainable limits, and impeding diversification by subsidizing the staple sector at the expense of other sectors. Governments may also have a propensity for dissipating rent by funding unsustainable or uneconomic public expenditures at the behest of stakeholders seeking a share of the rent.

Dependency and comparative advantage paradigms

While there is general agreement that natural resources provide advantages and disadvantages in the development process, there is disagreement over the relative balance. This has given rise to two schools: the dependency or resource curse school, and the comparative advantage or what could be termed the resource blessing school (Gunton 2003).

The dependency or resource curse school is decidedly pessimistic about staple-based development. Relying on staples is very risky strategy with a high probability of failure. The comparative advantage school is distinctly more optimistic. It acknowledges the challenges of staple-based growth but concludes that the advantages outweigh the disadvantages. The
comparative advantage school also argues that the problems of staple-led growth have to be compared to the problems of alternative development strategies such as indigenously owned manufacturing and import substitution, which have their own challenges. The conclusion is that regions with a plentiful supply of natural resources should focus their efforts on developing their staple sector. The key to success is not abandoning staples but managing their development properly by implementing policies such as collection of rent and development of economically viable linkages to maximize benefits and overcome challenges.

The debate over the resource curse versus resource blessing positions has spawned a plethora of empirical research evaluating whether natural resources are an advantage or disadvantage. Many studies show a negative correlation between resource dependency and growth (e.g., Auty 1995), while other studies show a positive correlation (e.g., Michaels 2011; Cavalcanti, Mohaddes, and Raissi 2011). The results vary with how development and resource dependency are measured, among other factors. Critics of the resource curse argue that even if there is a negative correlation between staple dependence and growth, this does not mean that slow growth is caused by resource dependency or that there would be more rapid growth if the region did not have natural resources.

While there is no consensus on the validity of the resource curse thesis there is agreement that many countries such as Norway and Canada seem to benefit from staple development while other countries such as Nigeria and Venezuela do not (Torvik 2009; Ploeg 2011). This raises perhaps a more productive focus of research: why are some countries more successful in achieving staple-led growth than others?

According to the research, the key variable determining success is the quality of governance (Torvik 2009; Ploeg 2011). Countries with high institutional quality – defined as clear and stable property rights and low incidence of corruption – exhibit a positive correlation between resource dependency and growth, while countries with a low institutional quality display a negative correlation.

Natural resource development and the environment

With the publication of the Brundtland Commission report in 1987 and the United Nations Conference on the Environment and Development held in 1992, sustainable development emerged as a primary consideration in public policy. The magnitude of the challenge is documented in many “state of the environment” reports such as the UN Millennium Ecosystem Assessment (MEA 2005), which concludes that nearly two-thirds of the world’s ecosystem services are in decline and 10–50% of species are threatened with extinction due to habitat loss and overexploitation.

The environmental damage generated by exploitation of fossil fuel resources is particularly well documented. The Intergovernmental Panel on Climate Change (IPCC) warns that unmitigated climate change will likely exceed the capacity of the natural, managed, and human systems to adapt (IPCC 2007, 73), while the International Energy Agency’s (IEA) 2011 World Energy Outlook concludes that the status quo policy regime will result in an increase in average world temperatures of 6°C, well above the 2°C target adopted by the world community in the Copenhagen Accord in 2009. The principal driver of these increases is increased production of fossil fuels, which is forecast to increase 18% for oil, 25% for coal, and 55% for natural gas by 2035 (IEA 2011, 71).
The environmental impacts of natural resource extraction and consumption, combined with the finite supply of natural resources, are dramatically illustrated by Meadows, Randers, and Meadows (2004) who popularized the limits to growth argument by showing through series of computer simulation scenarios that increased consumption would exhaust the world’s supply of natural resources sometime in the mid-twenty-first-century and put undue stress on the world’s environment. The limits to growth argument was criticized for failing to incorporate market adjustments such as improved productivity in resource extraction, development of substitutes, and changes in demand, all of which would occur in response to higher resource prices generated by increasing scarcity (Bardi 2011). Critics cite evidence of general declining resource prices over the last century and increased or constant reserve-to-consumption ratios as indicators that resources are not in short supply. But even the strongest critics of the limits to growth argument acknowledge that there is no assurance that these trends will continue and that at some point the supply of some natural resources may be exhausted. For example, more sophisticated modeling that incorporates prices concludes that oil and natural gas reserves could be exhausted within 35 and 37 years, respectively (Shafiee and Topal 2009). Further, even if world production is unaffected by supply constraints, production will shift spatially as some regional supplies are exhausted, resulting in significant economic dislocation, and pollution generated by unconstrained natural resource extraction will result in significant environmental damage.

Growing awareness of environmental impacts of resource development has resulted in most jurisdictions implementing environmental impact assessment processes to determine if large projects should be built and, if they are built, how the impacts should be mitigated. Governments have also implemented environmental regulations such as emissions controls and harvesting restrictions to reduce environmental impacts. Canada, for example, significantly reduced acid rain caused by sulphur dioxide emissions from its mining sector by the introduction of increasingly strict regulations implemented over the last several decades, and many jurisdictions are implementing regulations to reduce greenhouse gas emissions.

Environmental impacts of resource development have also spawned public opposition from nongovernmental organizations and aboriginal populations impacted by development throughout the world. In Canada, aboriginal populations and environmental organizations opposed timber harvesting by mounting blockades, legal challenges, and international boycott campaigns in the 1990s (Hayter 2003). The conflict, known as the “war in the woods” resulted in implementation of new timber harvesting practices, more than doubling the land set aside for protection and co-management of timber extraction (McGee, Cullen, and Gunton 2010). In North America there is opposition to the expansion of oil and gas production and the building of pipelines to transport oil to market. In Chile, opposition blocked a major power project and opposition in China over the last decade has blocked the massive hydro development along the Jinsha River.

Stakeholder opposition to resource development has given rise to concepts such as free and prior informed consent and social license, which are based on the principle that resource development must have the support of stakeholders as well as government before it can proceed. A specific tool to facilitate stakeholder support for resource development is the impact benefit agreement (IBA), which is a contract between resource development corporations
and impacted populations that specifies terms and conditions of development beyond those imposed by government. IBAs are common where resource development impacts aboriginal populations, and in some countries such as Canada, signing an IBA with aboriginal groups is a precondition for gaining support needed for project approval.

IBAs normally include provisions for financial payments such as royalties to local populations, profit sharing, co-ownership, provision of infrastructure and services, employment training, hiring, purchasing from local businesses, and protection of the environment. Evaluation of IBAs concludes that while there are challenges in developing and implementing them due to inequality in bargaining power, weak enforcement provisions, and uncertainty, they have the potential to increase benefits of resource development (Fidler 2009). IBAs will therefore become an increasingly common tool for managing resource development.

Case study: Canadian tar sands

Issues in resource development are well illustrated by the development of tar sands in Canada, which is one of the largest natural resource projects in the world. Oil production is forecast to increase almost fourfold to 5.4 million barrels per day by 2035, involving a capital investment of $220 billion (US 2011) plus an additional $30–40 billion in pipelines to transport oil to market (Honarvar et al. 2011). The development will entail dozens of individual development projects located across an area approximately the size of England.

Development of the tar sands is the subject of a major debate in Canada between proponents who argue that the development will strengthen the Canadian economy and critics who argue that development will undermine the long-run health of the economy by impeding diversification, degrading the environment, and relying on a risky nonrenewable resource subject to the vagaries of international commodity cycles.

Proponents cite economic impact studies that show that tar sands expansion generates significant economic activity (Honarvar et al. 2011) and increases per capita GDP in the three most resource dependent provinces (Alberta, Saskatchewan, and Newfoundland) by 40 and 60% above the Canadian average.

Critics caution that the tar sands industry is among the most capital intensive, generating only one job per $2 million of production, compared to manufacturing which generates 20 times more employment and the service sector which generates from 20 to 65 times more employment per dollar invested than the tar sands (Clarke et al. 2013). Although per capita GDP is much higher than the Canadian average for the resource-producing provinces, the differences in income are modest because returns to employees are such a small component of tar sands expenditures (Clarke et al. 2013).

Employment impact is further constrained by the failure to develop forward linkages related to processing of oil prior to export. Almost all future tar sands development will produce and export raw bitumen to foreign markets. Therefore, Canada will lose out on refinery processing employment and on many of the backward linkages such as supply of machinery and equipment, much of which will be imported from the US (Honarvar et al. 2011).

Critics also note that the oil sector in Canada is controlled by large multinational firms: the top 10 firms control three-quarters of Canada’s oil production and eight of these firms are foreign owned, including one owned by the Chinese government. Royalties will collect only 35–65% of the rent (Plourde 2010), leaving the rest to be
NATURAL RESOURCES DEVELOPMENT

retained by oil companies as excess profits, much of which will be leaked from the Canadian economy as payments to foreign-owned shareholders who receive almost one-half of Canada's oil and gas industry profits. The integrated structure of firms selling oil to their own refineries creates the opportunity for firms to reduce royalty payments to Canadian governments by discounting the export price using intracompany transfer pricing. Integrated firms transporting their oil to their own existing foreign-located processing assets are also less likely to build new processing facilities in Canada.

Critics also caution that development of the tar sands will distort the structure of the Canadian economy by inflating costs and the exchange rate and consuming capital and labor that could have been invested in other sectors of the Canadian economy. From 2001 to 2011 the Canadian economy gained 16,500 jobs in the oil and gas sector but lost 520,000 in manufacturing (Clarke et al. 2013). Critics acknowledge that there may be many reasons for the contraction of the manufacturing sector, but the evidence shows that a significant proportion of the contraction is caused by the Dutch disease impacts of natural resource development inflating the exchange rate (Clarke et al. 2013).

Critics also identify economic risks resulting from the cyclical nature of oil markets. Canadian tar sands development is among the most expensive sources of oil in the world and many of the current projects are suffering from significant cost overruns. During the current boom, tar sands expansion is growing so rapidly that it is inflating costs and reducing the export price by overwhelming transportation and refinery capacity, creating what has been termed the “bitumen bubble” (Clarke et al. 2013). The problem facing the tar sands is that because it is among the highest-cost producers, it has little capacity to absorb the inevitable fluctuations in oil markets.

Both proponents and critics agree that the expansion of the tar sands development will have significant environmental impacts. Despite regulatory controls that have helped reduce greenhouse gas emissions by almost one-third per barrel produced from 1990 to 2009, tar sands oil remains high-polluting relative to conventional oil (Grant et al. 2013). Greenhouse gas emissions resulting from tar sands development in Canada will increase threefold, from 45 Mt/year in 2010 to 155 Mt/year by 2030 (Millington et al. 2012) thus making it virtually impossible for Canada to achieve its greenhouse gas emissions targets, while other emissions will negatively impact air, land, and water (Grant et al. 2013).

Development of the Canadian tar sands represents the classic staple dependency versus staple advantage debate and there is evidence to support both sides. No doubt the tar sands will provide a fruitful area of research to test competing hypotheses about the role of natural resources in development.

Conclusion

Economically traded natural resources comprise a significant proportion of the world economy, especially for less developed economies, while noneconomic benefits provided by natural resources exceed the total value of the world’s economic production. The evidence suggests that management of natural resources is seriously deficient. Many countries with a plentiful supply of natural resources, that should provide a significant advantage in the development process, experience slower rates of growth than countries without natural resources. Environmental mismanagement of natural resources is particularly worrisome. The majority of the world’s ecosystems are being degraded, and increased consumption of fossil fuels is accelerating the rate
of climate change. The environmental impacts of some resource development projects may be so severe that society could be better off leaving the resources in the ground, thus turning the development paradigm on its head.

These challenges require a fundamental rethinking of management policies to achieve sustainable development. The priority in future research in the resource policy field should be to develop policies to achieve sustainable resource development and to identify strategies to ensure successful policy implementation. Focusing on policy research to collect rents, develop linkages, and mitigate environmental impacts will help ensure that natural resources are a benefit instead of a curse.

SEE ALSO: Monitoring and evaluation; Regional development models; Trade and regional development

References


NATURAL RESOURCES DEVELOPMENT


Natural resources

Daniel Banoub
University of Manchester, UK

For many people living in the Global North, capitalist society appears liberated from the material world. Economies are now “postindustrial,” with services, technology, and information replacing manufacturing and resource extraction. The natural world is no longer looked to for survival, but as a place of escape from everyday life for recreation. The combined effects of the 2007–2008 financial crisis, a period of rising commodity prices, and the effects of climate change upset this belief. The global economy is still a vast engine based on the appropriation and conversion of natural resources into a range of commodities and by-products.

The reemergence of the “problem” of natural resources, however, is more a question of perspective than reality. For most people outside the West, the “outdated” problems of resource availability and access have remained central to everyday life. The apparent liberation of the postindustrial economy is exactly that: an appearance. Society’s dependence on Earth as a source of material inputs and as a sink for by-products has only deepened over the past forty years. Ironically, the “liberation” of the economy from the material world is a result of the spatial expansion and intensification of resource extraction and use.

The strange combination of dematerialization and rematerialization, at first, appears to be a paradox. How can society be liberated from, yet more dependent upon, the material world at the same time? This seeming contradiction highlights the need for a geographical approach to the study of natural resources. Summed up concisely, the paradox is a product of uneven geographical development. The geography of natural resources is profoundly uneven as a result of a basic fact: natural resources are not evenly distributed across the globe. Even oxygen is scarce at certain altitudes. Therefore, natural resource availability does not map perfectly onto the geography of natural resource demand. This has two consequences. First, the modern economy is dependent upon long-distance flows of natural resources across space. Second, these flows produce an unequal distribution of positive and negative consequences, at a variety of scales. Spaces of dematerialized consumption cannot exist without (often far-flung) spaces of material extraction.

With this in mind, this entry will examine natural resources through the lens of critical human geography. The first section attempts to denaturalize the “natural resources” category. It concludes that natural resources must be considered relational, dynamic, and irreducibly social rather than simply “natural.” The following three sections are organized around the geographical contradiction between fixity and motion and the way it is experienced in three different registers: “stock–flow” and classification, “circulation-accumulation” and political economy, and “territory-networks” and geopolitics.

Denaturalizing natural resources

The term “natural resources,” at first sight, seems self-evident, almost banal. Natural resources are
any part of the nonhuman, biophysical world that satisfy a human want or need. They are most often understood as the raw material inputs that sustain the economy: minerals, fossil fuels, plant and animal life, soils, water, air, wind, and sunlight. Natural resources can also refer to a variety of ecosystem services, such as carbon sequestration, that help maintain the general conditions of life. Natural resources also provide certain noninstrumental uses, such as the aesthetic, spiritual, cultural, or recreational “consumption” of nature. The broad definition of “natural resources,” then, refers to a wide range of entities: a Douglas-fir tree, a plant’s ability to absorb CO$_2$, a resplendent sunset over a tranquil beach, to name three.

“Natural resources” is a blanket term intended to cover the diverse range of human uses of the nonhuman world. Yet, its conceptual clarity rests upon two subtle, yet questionable, implications that emerge from the use of “natural” as an adjective. Nature, as Raymond Williams famously asserted, “is perhaps the most complex word in the language” (1983, 219). It is so complex because it evokes a wide range of meanings. First, “natural” limits the object of inquiry to the nonhuman world, making a strict division between “society” and “nature.” Second, it naturalizes “human wants and needs,” reducing them to a question of supposedly innate requirements. “Resource-ness,” in other words, is assumed to be an intrinsic quality of the nonhuman world, existing unchanged throughout time, simply waiting to be discovered by resourceful humans.

The discipline of geography has played an historic role in developing applied, instrumental forms of knowledge meant to understand, appropriate, and manage the nonhuman world. For natural resource geographers working in the managerial tradition, the “naturalness” of resources is indisputable. Accordingly, they have devised highly technical, expert-based forms of knowledge predicated on treating natural resources as fixed and stable. These forms of knowledge and expertise are easily reconcilable with state administration and capitalist political economy and form the basis of the conventional understanding of nature.

The managerial approach views Earth as ruled by the law of the conservation of mass. Mass (matter or energy) can be neither created nor destroyed. Earth, therefore, is composed of a fixed stock of resources. The abundance or scarcity of natural resources is a geological or biological fact emerging from asocial processes. The managerial approach emerged in the nineteenth century to protect nature from the destructive tendencies of Western industrialization. From the late 1990s, the modern nature conservation movement extended administrative techniques to natural resources, seeking to improve the productivity of nature through its rational and ordered use. These centralized, hierarchical, and scientific forms of knowledge sought to radically reduce the complexity of nature to satisfy human needs. Nevertheless, this radical process of simplification was never fully successful, being resisted by both natural resource users and the natural resources themselves. In response to these failures, natural resource management has been reframed around questions of complexity, uncertainty, and adaptation. Rather than seeking to reduce complexity, so-called adaptive management embraces uncertainty, promoting active learning and continual modification.

In contrast to the instrumental forms of geographical knowledge, there is also a rich critical tradition in geography. This tradition refuses to treat nature as an ahistorical category and takes issue with the two subtle implications of the adjective “natural.” Critical geographers are located primarily in human geography, but also to some extent in physical geography.
They argue that natural resources are anything but natural. In other words, they contend that “nature” and “society” cannot be treated as discrete domains. Moreover, “resource-ness” is not an innate biophysical quality, but an emergent outcome of social and political work. Rather than pre-existing humanity, natural resources are produced socially. This does not deny the existence of biophysical processes and properties that humans refer to as nature. Rather, it argues that they cannot be divorced from the social means through which they are appraised, appropriated, distributed, used, and discarded. As Erich Zimmermann (1933) succinctly and presciently argued: “Resources are not; they become.”

Critical geographers reject the argument that natural resources are fixed, absolute, and asocial. They treat natural resources as dynamic, relational, and irreducibly social. In other words, what counts as a natural resource is seen as a social appraisal with two preconditions. First, there must be a sufficient level of knowledge to recognize, extract, and use them. Second, there must be an economic demand for services they provide. This helps explain why certain bits of the biophysical world slip in and out of the category of “natural resources” over time. For example, whale oil and guano were once highly valuable commodities. Now, they are no longer considered a profitable investment opportunity. “Resource-ness” does not exist in nature as an essential quality, but is a product of social work. Moreover, it is not immutable, but demonstrates remarkable dynamism over time.

Critical geographers, similarly, treat scarcity as contingent and relational. Scarcity is not predetermined by biology or geology alone. Rather, it is a social appraisal of nature that occurs in a specific historical context. A small shift in knowledge, technology, or demand can radically alter perceptions of scarcity or abundance. Oil, for example, is considered the perfect example of the relationship between scarcity and nature. Recent work in the political economy of oil, however, has shown that oil’s scarcity is not a simple function of geology. Rather, it is a relational outcome of social processes, such as the politics of physical access, control, and land ownership and the internal organization of the oil industry. The more pressing historical problem for oil firms has been to produce scarcity in the context of staggering liquid abundance. With this in mind, the recent narrative of “peak oil,” which posits the imminent exhaustion of finite oil supplies, must be taken with a grain of salt.

Natural resources purportedly satisfy human desires and requirements. There are certain biophysical requirements for the sustenance of human life, such as food, water, and shelter. On their own, however, they have very limited explanatory potential. Human “wants and needs” are fluid and ever changing. For example, sweetness is undeniably and innately pleasing. Nevertheless, the human predilection for sweetness cannot explain the role of sucrose in the modern diet. As Sidney Mintz (1986) has shown, the global adoption of sucrose as a source of calories was the result of social power, not humanity’s inherent sweet tooth. Human “wants and needs” are not reducible to biology. Rather, they are constituted socially and display remarkable flexibility across space and time. They are also riven with social differences, making them the site of intense political contestation.

Critical resource geography reframes the instrumental or managerial forms of geographical knowledge concerning natural resources. Rather than developing technical, expert-based management regimes, critical geographers seek to “denaturalize” natural resources. They demonstrate that the “naturalness” of natural resources is an effect of social processes rather than an innate quality. Critical resource geographers
excavate the deeply political character of natural resources. They broaden the scope of analysis beyond the narrowly circumscribed questions of management. They interrogate the more openly political questions of access to natural resources and the distribution of their benefits and waste. Critical geographers, in sum, seek to denaturalize, and thereby politicize, the study of natural resources by arguing that they are dynamic, relational, and irreducibly social.

**Stock–flow and the classification of natural resources**

As mentioned above, the term “natural resources” refers to a wide range of biophysical properties and qualities that humans find useful. Consequently, a number of classification systems have been developed to reduce the complexity of the biophysical world. Natural resources can be distinguished upon a number of axes. For example, a seemingly basic classification would be between biotic and abiotic resources. Biotic resources are living organisms capable of biological reproduction, such as plant and animal life. Abiotic resources are nonliving entities produced by nonorganic physical, chemical, and geological processes, such as minerals, sunlight, and air. This distinction, however, obscures the fact that biotic and abiotic resources are deeply codependent. Radical biologists argue that an organism and its abiotic and biotic environment are inseparable, existing in a dialectical, co-constitutive relationship. The most common classification system is based on the resource’s ability to regenerate. This system distinguishes between nonrenewable “stock” resources and renewable “flow” resources. Geographer Judith Rees (1991) refined this distinction by dividing stock and flow resources into different subcategories. She divided stock resources into three categories. First, those that are consumed by use, such as fossil fuels; second, those that can be theoretically recovered after initial use, which includes all elemental minerals; and third, stock resources that are recyclable with current technology, such as aluminum and gold. She divides flow resources into two categories. First, flow resources that are capable of exhaustion, if exploitation overwhelms regeneration, are defined as “critical zone resources.” Second, flow resources that human activity cannot affect, such as wind and sunlight, are defined as “noncritical zone resources.”

Yet, as Rees pointed out, the dividing line between stock and flow resources is “very fine.” “Conventional thought” justifies it more than biophysical realities (1991, 240). The stock–flow distinction is a function of the timescales over which they regenerate. Oil, for example, is a classic “consumed by use” stock resource. The decomposition of organic matter under conditions of heat and pressure over millions of years produces oil. In other words, the formation of

---

**Figure 1** Classifying natural resources by their spatial characteristics.
oil is a continual process, not a single event. Its exhaustibility is a relation between the rate of underground production and the rate of its extraction. The “exhaustibility” of oil is not an absolute, essential quality of the resource. It is a contingent outcome of the social relations of extraction and consumption. Similarly, many classic “flow” resources (fish, trees, water in aquifers, etc.) are capable of exhaustion. Like oil, this is a relational outcome of the social relations of production, not an essential biophysical quality. The stock–flow distinction is not an absolute division determined by biophysical characteristics. It is more accurately treated as a “resource continuum” (Rees 1991, 241).

Another way of classifying natural resources is around the contradiction between fixity and motion (Figure 1). This method is explicitly geographical. The system is organized around the spatial, temporal, and biophysical properties of natural resources. Some natural resources are mobile; that is, able to circulate freely through space – migratory animals and water are examples. Others are fixed, accumulating in particular locations – think of coal, land, and trees. The biophysical properties of resources – their phase, density, solubility, volatility, and so forth – can make them particularly amenable or resistant to efforts to fix them in place or make them circulate. This distinction is also dependent upon timescales, because what may appear fixed in space may actually be moving when examined over long periods. For example, while trees might be spatially embedded, forests are remarkably dynamic. Finally, the propensity to circulate or accumulate can be remade through technological or economic innovations. For example, the development of refrigeration, steam–powered transport, and canning increased the capacity of animal protein to circulate the globe by overcoming decomposition.

This distinction, moreover, highlights some of the inadequacies of the conventional stock–flow continuum. Oil and coal, for example, are both within the “consumed by use” stock category, yet they display different spatial characteristics. “Light, sweet” crude oil circulates, whereas coal accumulates in sinuous veins. Even within the category of oil – a “catch-all term that covers a diversity of liquid hydrocarbons” (Bridge and Le Billon 2013, 5) – there is a range of spatial characteristics. “Conventional” oil is far more mobile than the bituminous deposits of Alberta. Bitumen is mixed with sand and clay and requires refining before it can flow through a pipeline. In terms of the fixity–motion criteria, coal (stock) and trees (flow) may share more in common than trees (flow) and certain fish species (flow). The geographical lens of fixity–motion highlights similarities between stock and flow resources and dissimilarities within those categories. Like the classification system outlined by Rees (1991), this system is not a succession of fixed categories. Rather, the fixity–motion lens is best characterized as a continuum.

Accumulation–circulation and the geographical political economy of natural resources

The contradiction between fixity and motion permeates the entirety of capitalist political economy. The expanded reproduction of the capitalist mode of production is predicated on the capitalists’ desire to accumulate. This desire cannot be satiated and has no theoretical upper threshold. Their motto is: “accumulation for accumulation’s sake, production for production’s sake” (Marx 1976, 595). Yet, as Marx points out, the capitalist is not a miser. Rather than hoarding their money, capitalists set it free, throwing it into circulation, hoping that, like a true lover, it will return.
Marx distinguishes between “money as money” and “money as capital.” Money simply gathers dust in the miser’s hoard; capital is set in motion. However, to throw capital into circulation in the hopes that an equivalent amount will return would be as “purposeless as it is absurd” (Marx 1976, 251). Thankfully, as Marx notes, “circulation sweats money from every pore” (1976, 208). The circulation of capital is a process of self-valorization, a continual expansion of value. The initial sum of capital thrown into circulation, it is hoped, will return in an expanded form. The accumulation and circulation is characterized by a contradiction between fixity and motion. The desire to accumulate, to hoard wealth and fix it in place, can only be satiated by the circulation of capital, by throwing it into motion.

The dialectical tension between fixity and mobility shapes the dynamics of accumulation and circulation. As David Harvey argued in his book, The Limits to Capital (1982), it also shapes the geographical organization of capitalism. Capital is constantly striving to overcome spatial barriers, always seeking the “annihilation of space by time” (Marx 1973, 539). However, this can only be accomplished through the construction of immobile infrastructures. Circulation requires roads, communication systems, and cities. In other words, capital must be sunk into the built environment, literally fixed in space, to overcome spatial barriers. These built environments are not an everlasting solution to capital’s confrontation with space. Rather, they will at some point become the spatial barrier to be overcome. The contradiction between fixity and motion emerges as one of the driving forces in uneven geographical development. The mutating geography of capitalism is relentless, overcoming certain spatial barriers by creating new ones.

Natural resource geographers have made a significant contribution to geographical political economy. They argue that the materiality of natural resources shapes the organization of production and the geography of capitalism. They do not treat natural resources as a raw material input or a “free gift” of nature. Rather, they argue that biophysical properties constrain, enable, frustrate, and surprise natural resource industries. In other words, natural resources are not just an external limit to extraction. Every obstacle is an opportunity for capital. The exhaustion of wild capture fisheries creates the possibility of farmed fish. The materiality of natural resources actively constitutes the spatial organization of capitalism. Uneven geographical development does not simply produce effects on nature (depletion, pollution, etc.). It is also produced by natural resources.

Natural resources industries are essentially exercises in securing and controlling flow. Humans try to coax resources into the circuits of capital at the optimal rate for human needs, or at least the logic of capital. For fixed resources, the main difficulty is securing flow, bringing fixed deposits into circulation. For mobile resources, the difficulty is controlling flow, harnessing the mobility of nature for human needs. It is important to remember that these tensions and confrontations are not only encountered at the point of extraction. Rather, the tension between fixity and mobility permeates all phases of the circuit of capital. It affects the “upstream” phase of exploration and extraction, the “midstream” phase of storage and transportation, the “downstream” phase of processing, distribution, and consumption, and the “backstream” phase of waste production and assimilation (Bridge and Le Billon 2013, 37). The material properties of resources can either constrain or enable flow. This fact has profound effects upon the uneven geographical development of capitalism.

The propensity to accumulate or circulate provides the foundation for natural resource industries. Yet, ironically, it often emerges as the
central barrier to accumulation. For example, mining companies dream of finding a high concentration of valuable minerals in a limited area. However, the more concentrated and limited the area, the more difficult it is to find it. Factor in the distance from an economic center, and the efficiency of transportation, and the deposit may no longer be profitable. Likewise, water’s ability to flow is the basis of complex irrigation, hydropower, and drinking water systems. Yet, its innate tendency to under- or overflow can have disastrous consequences. Thus, the fixed/mobile character of resources appears as an external limit to accumulation. But, as Marx points out, in the capitalist mode of production, “every limit appears as a barrier to be overcome” (1973, 408). Accordingly, capitalists have devised many technical, scientific, and organizational innovations to overcome these barriers. Take deep-sea fisheries, for example: for a neophyte, finding large concentrations of fish would be like trying to find the proverbial needle in a haystack. Fish have complex life histories and migratory patterns and live in a dynamic environment. Over time, fishers develop refined understandings of the ocean environment, helping them track and extract fish. The development of scientific knowledge around fish biology and oceanography bolsters this traditional knowledge. Following World War II, the fishing industry adopted a range of new technologies. These include such things as sonar, radar, echo-sounders, synthetic fibers, and refrigeration. These technologies increased productivity and encouraged organizational changes in the industry. Large factory-freezer trawlers owned by large seafood corporations now dominate the industry in developed nations. These technical, scientific, and organizational innovations seek to overcome barriers presented by fish populations. They attempt to control fish mobility and regulate the flow of fish and fishery products around the globe.

The point of extraction seems like the most obvious point for examining the materiality of resources. But, as mentioned above, the different stock–flow propensities of resources shape all phases of capital circulation. For example, William Cronon (1991) showed how the steam-powered grain elevator restructured the grain industry. This invention reoriented grain’s distribution rather than overcoming the material obstacles to its production. Before the steam-powered grain elevator, grain shipments traveled to market in sacks by ship. This guaranteed their unmixed and uncontaminated arrival – a slow, treacherous, and labor-intensive journey. The grain elevator increased the scale and efficiency of Chicago’s grain handlers. It was able to move, sort, and aggregate vast quantities of grain. But it had one caveat: grain shipments had to arrive sackless. “Only then,” Cronon argued, “could corn or wheat cease to act like solid objects and begin to behave more like liquids: golden streams that flowed like water” (1991, 113). This transformation in distribution had profound effects on the entire organization of production. Before the grain elevator, legal claims were placed on a particular sack from a particular parcel of land. After, they were placed on a specific quantity of a particular grade of grain. This severed the connection between grain’s movement as a physical object and its movement as a commodity. It also opened the possibility of speculation and a futures market. Both results increased the pace of accumulation. The steam-powered grain elevator secured and increased the flow of grain by turning the fixed properties of the sack into the “golden streams” of liquid grain.

Natural resources, despite our efforts to secure and control flow, are unruly. They have a tendency to frustrate human desires. Consequently, they must be understood as a source of immense uncertainty and surprise. This problem
NATURAL RESOURCES

is not limited to the failure of human systems intended to secure and control flow, such as a flood caused by the malfunctioning of a dam. Surprises also emerge from the very successes of these systems. Industrial livestock and agriculture technologies have overcome some of the barriers to the regular flow of animal and plant products. Yet, they also produce a range of diseases and parasites that threaten the survival and profitability of the enterprises. Intensive open net-pen salmonid aquaculture produces infectious salmon anemia and sea lice. The large-scale pollination of mono-crops by bees leads to colony collapse disorder. Large-scale poultry farms breed avian influenza.

The transition to fossil fuels radically increased humanity’s ability to secure and control flows of natural resources. For millennia, biological sources of energy powered society: water, wind, and muscle. These sources were gradually replaced by fossil fuels – first coal, then oil. Yet the unearthing of vast lithospheric stores of hydrocarbons produced many unintended consequences. They unleashed many uncontrolled and destructive flows and accumulations that threaten the reproduction of capitalist society. The accumulation of CO₂ in the atmosphere is the most obvious, but these flows also occur on finer scales and micro-geographies. The extraction and processing of bitumen in northern Alberta is a good example. Bitumen is an “unconventional” source of oil comprised of heavy crude mixed with sand, clay, and water. Its extraction unleashes a flow of deleterious toxins in the immediate area and downstream along the Athabasca River. Toxins seep into the river through the storage of wastewater in “tailings ponds,” flowing northward to the Arctic Ocean. On their journey they bioaccumulate in plants and animals. More distressingly, they bioaccumulate in the bodies of indigenous people and settlers in northern Alberta. These uncontrolled flows have a predictable and devastating impact on northern, primarily indigenous, communities. For example, Fort Chipewyan has six recorded cases of cholangiocarcinoma, a rare form of bile duct cancer that typically affects 1 in 100,000 people. For a community of 1200 people this is an impossibly high rate. Despite our best attempts to secure and control the flows of natural resources, they will remain a source of surprise, shaping the organization of production and the uneven geography of capitalism.

Territory-networks and the geopolitics of natural resources

In 2008, the US Army Modernization Strategy listed resource demand and population growth as two of the key security threats in the era of “persistent conflict” (US Department of Defense 2008, 6). The US military’s interest in natural resources is based on a seemingly indisputable and pragmatic logic. Earth’s biophysical resources are limited. Thus, the exhaustion of Earth’s resources is inevitable in a rapidly urbanizing world. Fears over the imminent surpassing of planetary boundaries have intensified over the past two decades. On one hand, the problem of depletion, or “peak” everything, is causing fears of scarcity. On the other hand, the problem of pollution and climate change are causing fears of over-abundance. This logic has inspired terrifying visions of the near future. Poverty, starvation, population migrations, and mass mortalities are mere decades, or years, away. In this context, natural resources have reemerged as a central topic of geopolitical calculation. The negative effects of resource depletion and climate change are certainly lamentable. But the more pressing issue for most Western states is the possibility of violent conflict over resources, such as water, food, and energy. They believe that competition...
over natural resources will inevitably produce geopolitical instabilities. The return of natural resources as a geopolitical problem is far from expected. Indeed, the return of geopolitics as a problem at all in the twenty-first century would surprise many people. Swedish political scientist Rudolf Kjellen first coined the term “geopolitics” in 1899. It was the “science” of state security and interstate competition. The discipline viewed the state as an organic entity. Inspired by Darwin’s theory of evolution, states struggled for survival by imperial expansion. By the mid-twentieth century the openly racist and imperialist problem of state competition was reworked. The more abstract notion of “global” competition between the United States and the Soviet Union became the central preoccupation of geopolitics. With the end of the Cold War, the notion of interstate competition seemed outdated in the era of globalization. The “neoliberal” era emphasized uninhibited international trade and investment. In this context, few expected the return to violent nineteenth-century struggle over resources. It is surprising indeed, if not inconceivable.

Recently, conventional foreign policy experts have located an evolutionary transition from geopolitics to geo-economics. Despite this, geopolitical calculation has not been fully eclipsed by a focus on geo-economic flows. The military interest in natural resources demonstrates the continued relevance of geopolitics. Critical geographers have shown that the contradiction between fixity and motion is central to the social geography of power. The desire to fix and defend borders – to bound space and delimit movement – is one of the paradigmatic features of the modern state. Yet, the state is utterly dependent upon transnational flows of information, capital, people, and natural resources. The act of delimiting a boundary recognizes the inevitability of its transgression by movement. A border would be meaningless if no one or nothing tried to cross it. The territorial logic of geopolitics and the market logic of geo-economics are contradictory, but not mutually exclusive.

Many social scientists have shied away from questions of territory in favor of a focus on networked relations. Fixity and stasis appear to define the concept of territory. Moreover, it is rife with sociobiological connotations, positing animals and humans as biologically “territorial.” This is considered homogenizing, even reactionary. In a world characterized by transnational flows and diasporic communities, it is unacceptable. In response, many scholars turn to the concept of networks, focusing on relations, dynamism, and fluidity. Yet, as some scholars point out, territory and networks are not antithetical forms of spatial organization. Rather, they are dialectically related. The spatial organization of power is historically and geographically specific and inherently dynamic. Thus, the tension between the desire to enclose space and to open it up for circulation is one of political geography’s central contradictions.

Natural resources are one of the principal sites over which the tension between territorial fixity and economic mobility is laid bare. This is the result of a rather banal, geographical fact mentioned earlier. The geographies of natural resource availability and demand are not identical. This ensures that states and firms want to enclose the natural resources within their “own” territory. Yet, they also want to ensure that “foreign” resources are available for transnational mobility. In the “neoliberal” era, some argue that the state is losing relevance through privatization and deregulation. Nonetheless, it is far from irrelevant. There is an equally strong tendency toward the extension of state power. The recent saber rattling by the circumpolar states over claims to the Arctic demonstrates this. Politicians
NATURAL RESOURCES

are planting flags and fervently making speeches in an attempt to secure their nation’s access to the potentially valuable oil, gas, and mineral reserves.

The state has not abandoned its historical extra-economic role as an enabler of capital accumulation, but it has been seriously reworked over time. The focus on geo-economic flows and state withdrawal should not lead us to assume that geopolitics and territory are no longer relevant. Rather, it should highlight the centrality of territorial regimes in capital accumulation. Deregulation is better understood as a form of reregulation. The “free” market is never truly free, but operates within historically and geographically specific bounds. Natural resources are embedded in legal, institutional and cultural regimes of nation-states (Bridge 2008). Territorial regimes determine what counts as a resource. They control who is able to access them and what people can do with them. They shape the distribution of their positive and negative outcomes. The “territoriality” of natural resources still shapes the organization of geo-economic flows. The spectacular rise of the Chinese economy has produced a tectonic shift in the global distribution of geopolitical power. For many years, people considered China as a “developing” or “Third World” country. Over the past twenty years, China has emerged as an important player in the global political economy. For many, China’s emergence as global superpower has elicited considerable anxiety. Natural resources are a central concern. China’s unparalleled economic growth, at an average annual rate of 10% over the past decade, is dependent upon a colossal throughput of natural resources. Conventional foreign policy experts frame China as a nascent threat to American hegemony. They consider China a ruthless new competitor in the zero-sum game of geopolitics. China’s insatiable appetite for fossil fuels and raw materials will lead inexorably toward conflict with the West (Cáceres and Ear 2013). Environmentalists argue that China’s industrialization and urbanization is harmful. Powered primarily (70%) by coal, it is unsustainable and destructive. Others criticize China’s authoritarian domestic policies and “business only” foreign policy. China represses internal dissent and does not discriminate against nations with poor human rights records. Taken together, the negative social and ecological repercussions of China’s growth will be on a truly global scale.

China’s emergence as a global power is often framed in solely national terms. China’s internal, authoritarian conditions, however, cannot explain its economic growth. China’s current economy is export-oriented. It would be unthinkable without foreign markets to buy its cheap commodities. China’s quest for natural resources is a result of broader shifts in both Chinese geo-economic strategy and global political economy (Lim 2010). It is not simply an external “threat” to American hegemony, but inextricable from it. In 1978, the Chinese state “opened up” the economy, encouraging foreign investment within China. In the early 2000s, the state shifted to a “going out” strategy. It began to secure natural resources abroad to feed its export-oriented economy. Firms began investing abroad and the state pursued international trade agreements (Gonzalez-Vicente 2011). Chinese capital is increasingly assertive in securing an adequate supply of natural resources. Private Chinese firms are increasing foreign investments. The Chinese state, through its sovereign wealth fund and state-owned enterprises, is following suit. They are developing economic relationships with resource-rich nations in Africa, Latin America, and Asia. Chinese geo-economic strategy has recently shifted toward a spatially expansive quest for
natural resources. This is reshaping the geographies of natural resources and the distribution of geopolitical power.

Natural resources are geopolitical. The political economy of natural resources does not operate in some abstract Cartesian space. The processes of securing and controlling flow are deeply embedded in specific territorial regimes. Moreover, the geographical distribution of value across the circuit of capital is highly politicized. It is a central site of contestation, conflict, and compromise. For classical geopolitics, natural resources are only examined when they cause international competition. Critical “geopolitics,” in contrast, brings attention to the geographical character of natural resources politics.

Natural resources are conventionally defined as any part of the nonhuman world that satisfies a human want or need. Natural resources may be “nonhuman,” but they are not asocial. For critical geographers they are complex social phenomena. The geographical contradiction between fixity and motion helps parse the complexity of natural resources. First, natural resources are often classified in terms of their regenerative dynamics, but the propensity to circulate or accumulate presents another axis of comparison. Second, these propensities in turn help shape the contradiction between circulation and accumulation in capitalist political economy. Third, the geographical dynamics of capitalism are rooted in the geopolitical contradiction between territorial enclosure and transnational flows. Recent shifts in global politics ensure that natural resources will remain at the center of social debate in the twenty-first century.

SEE ALSO: Environment and resources, political economy of; Geopolitics of the environment; Marxist geography; Nature; Political ecology; Production of nature; Socio-nature

References


Further Reading


Nature and corporeality

Bethan Evans
University of Liverpool, UK

Ideas about (bodily) nature are central to understandings of corporeality and, as Castree (2013) demonstrates, questions about corporeality (race, genetics, DNA, food, bioengineering) are central to our understandings of nature. Since the mid-1990s, there has been a significant growth in work within geography on corporeality, stemming largely from feminist work on the body in the early 1990s. Within this work is an inherent tension relating to ideas of “nature” and the nature and materiality of bodies that has, until recently, been relatively undertheorized within geography. Before outlining some of this more recent work, it is necessary first to consider the entrance of the body into geographical scholarship and address the tension at the heart of geographical work on nature and corporeality.

Much Western thought is structured around a Cartesian dualism between mind and body which maps onto a distinction between culture and nature. These dualisms are central to understanding both the historical absence of work on corporeality within geography and the relative absence of work on nature within geographical literature on the body (and vice versa). As Rose explains, these dualisms have historically informed (and, arguably, continue to inform) hierarchies of knowledge production in which the clause considered closer to “nature” (female, body, private, nature) is subordinate to the masculinized clause associated with the mind and culture. Within geography – as a discipline “implicitly structured around the distinction between Nature and Culture” (Rose 1993, 73) – this dualistic thinking meant the exclusion of the body as an object of study and also the exclusion of those associated with nature and the body (along the lines of gender, age, class, sexuality, race, disability) from knowledge production. This is not particular to geography and, as Grosz (1994, 4) observes, is common across Western philosophical traditions where, “Given the coupling of mind with maleness and the body with femaleness and given philosophy’s own self-understanding as a conceptual enterprise, it follows that women and femininity are problematized as knowing philosophical subjects and as knowable epistemic objects.”

Early work on the body within geography therefore drew on and contributed to broader feminist work which sought to demonstrate how, as Haraway explains, the distinctions between pure and applied science and between nature and culture are versions of the philosophy of science that exploits the rupture between subject and object to justify the double ideology of firm scientific objectivity and mere personal subjectivity. This anti-liberation core of knowledge and practice in our sciences is an important buttress of social control. (1991, 30–31)

Both these elements raised by Haraway are important to understanding how geographers have challenged the distinctions between nature and culture and mind and body.

First, feminist work has sought to challenge the “double ideology of firm scientific objectivity and mere personal subjectivity” (Haraway 1991, 30). The exclusion of subordinate groups from
knowledge production was based on claims that those groups are unable to transcend their embodied natures to achieve objectivity (Rose 1993). Modes of knowledge production within the social sciences therefore valorized disembodied knowledges and ideas of objectivity that resemble a “god trick” – a universal view from nowhere and no body. Donna Haraway’s (1991) work on situated knowledges has been central to feminist geographical work which has challenged this. Instead of objectivity being about “the false vision promising transcendence of all limits and responsibility,” Haraway argued for a feminist account of objectivity which is about “situated and embodied knowledges” entailing “politics and epistemologies of location, positioning, and situating, where partiality and not universality is the condition of being heard to make rational knowledge claims … the view from a body” (1991, 340, 342, 349–350). This has involved recognition of the embodied nature of knowledge production through acknowledging the positionality and embodiment of the researcher. In practical terms, this has meant that, rather than writing as a disembodied voice, research has been written in the first person, directly addressing and setting out the position and situation of the researcher in order to write from a situated body. It has also allowed a pluralizing of bodies considered able to contribute to geographical knowledge production, with particular advances in recent years made in children’s geographies, participatory geographies, and geographical work by and with indigenous communities, partly through challenges to previous ideas that these groups, like women, were limited by their bodily natures.

Second, feminist work on the body has sought to highlight and challenge the ways in which the “anti-liberation core of knowledge and practice in our sciences is an important buttress of social control” (Haraway 1991, 30–31). Here, Haraway is referring to the ways in which essentialist and biologically deterministic knowledge claims within the social and behavioral sciences have been (and continue to be) used to justify social positions of domination and subordination for particular bodies and groups. As she explains, “The union of the political and physiological … has been a major source of ancient and modern justifications of domination, especially of domination based on differences seen as natural, given, inescapable, and therefore moral” (Haraway 1991, 30). Early work on the body in geography, as elsewhere in the social sciences, involved a distancing from “nature” and a focus on the body as socially and culturally constructed in order to challenge the subordination of those groups whose capacities were seen to be limited by their corporeal natures. Thus, early work bringing corporeality into geography was “in contrast to the commonsense understanding of the body as completely natural” (Rose 1993, 29). This work has been, and continues to be, important in challenging the subordination of particular groups via essentialist discourses which evoke the “nature” of particular bodies to justify their subordination (see, e.g. Nash 2003 on antiracist geographies, and Wright 2010 on feminist and queer geographies). Critiques of mind/body and nature/culture dualisms also extend to critical work on the normalization of particular bodily performances, forms, and practices through highlighting the cultural constructions and moral knowledges which position some as able to “correctly” control bodily functions, desires, and “natures” (see, e.g., Colls and Evans 2013 on fat).

While work that challenges essentialist accounts of bodies continues to be important, the dominance of social constructionist accounts of bodies in early geographical work on corporeality means there is a tension within geographical work on corporeality and nature. As Longhurst (1997, 489) explains,
Constructionist feminists argue that bodies are discursively produced and that essentialist discourses – that is, discourses which make reference to the physical, biological body – serve to naturalize what is in fact social difference. For constructionist feminists, references to the biological body are seen to reinforce patriarchal claims that women are naturally incapable of certain kinds of action.

Thus, in challenging the hierarchies of those scientific modes of knowledge production which legitimated the exclusion of women, and in challenging knowledge which legitimates the subordination of those groups regarded as inferior with reference to their bodily natures, early geographical work on the body involved a disavowal of the importance of “nature” to knowledge of corporeality. As Whatmore (2002) explains, there is a similar tension between geographical work which adopts constructivist and realist approaches to questions of Nature/nature. It is important to note here, that as Whatmore recognizes in relation to work on Nature/nature, only the “most vulgar” constructivist accounts focus exclusively on linguistic constructions, and early work on the body within geography was similarly not exclusively concerned with representations and discourses at the expense of lived embodiment. Informed by poststructuralist feminist politics, this early work “is interested not just in the social construction or representation of bodies, but also in ‘real,’ material bodies that occupy ‘real’ spaces and places” (Longhurst and Johnston 2014, 270). However, in disavowing claims to bodily “nature,” such work had left corporeal natures relatively undertheorized. In recent years, however, there have been calls to address critically, and in nonessentialist terms, the relationships between nature and corporeality. The resultant work spans subdisciplinary boundaries and is a literature too vast to cover in its entirety here. This entry therefore signposts two

key themes through which nature and corporeality have been considered within geographic scholarship: corporeal natures, and corporeality in/and nature. Before discussing this work, it should be recognized that, in line with the aforementioned work on situated knowledges, this entry is inevitably partial: the choice of and treatment of these themes is shaped by the author’s own position as a feminist geographer who works on questions of embodiment and health.

**Corporeal natures**

Since the mid-2000s, geographers have built on, and moved forward from, a focus on representations and social constructions to address corporeal natures through two related theoretical developments within the discipline more broadly: first, through work broadly referred to as nonrepresentational or more than representational geographies, and, second, through research concerned with bodily materialities. In the first, as Anderson and Harrison (2010, 7) explain, “the bodies which populate non-representational theory are, for the most part, relational bodies; ecological in form and ethological in apprehension.” Through a focus on the ways in which bodies develop habitual relations to others and environments, on precognitive and affective responses to stimuli, and on the role of our senses in shaping the ways in which we encounter the world, this research offers theoretical approaches through which geographers can question corporeal natures in nonessentialist terms. There have been some feminist critiques of this work, centered on the lack of a differentiated body-subject, but, as Colls (2012) suggests, there are productive possibilities at the intersection between feminist and nonrepresentational theories to develop geographical work here further. This range of theories has been influential in many of the areas outlined in this and the following sections.
Second, geographers have built on social constructionist accounts of bodies in ways that are more open to considerations of corporeal natures through attending to bodily materialities and bodily matter. This work draws on broader feminist theory (see, e.g., Grosz 1994) which challenges the dualisms between mind and body, science and the humanities, culture and nature which, while questioned by work on social constructionism, tend to leave the material body still in the realm of the physical sciences. As such, Longhurst (2001) suggested that, while there had, to that point, been significant growth in work within geography on the body, “the bodies articulated in geographers’ texts have tended to be theoretical, discursive, fleshless bodies,” meaning that “the leaky, messy, awkward zones of the inside/outside of bodies and their resulting spatial relationships remain largely unexamined” (2001, 1, 2). Since then, there has been growth in geographical work which does attend to bodily materialities. For example, Colls (2007, 357) explores fat bodily matter, offering a nonessentialist approach through arguing that “both the material and discursive are mutually implicated in the intra-activity of matter.” Other work, which similarly draws attention to corporeal natures through a focus on materiality, includes work on bodily organs and external and internal bodily surfaces (see, e.g., Colls and Fannin 2013 on placental surfaces); research on food which questions the ways in which ideas about “natural” food and “healthy” bodies are culturally connected through foregrounding the matter which crosses bodily boundaries through consumption (e.g., Guthman 2012); work on (the management of) bodily fluids and the porosity of bodies (Longhurst 2001); and research on environmental citizenship which foregrounds corporeal materiality and challenges distinctions between nature and culture through an approach that Gabrielson and Parady term “corporeal citizenship.” This approach “begins from the fact of humans’ inescapable embeddedness in both social and natural contexts and adopts an understanding of the human body as porous but resistant, plural and connected ... [It] highlights the connectivity of the material world across the nature/culture divide” (Gabrielson and Parady 2010, 382; emphasis in the original).

The corporeal in/and nature

Following from the final point in the previous section, the second (related) way in which geographical work addresses corporeality and nature is through considering the ways in which bodies and “natural” landscapes or environments encounter each other. This work is similarly multiple and there is space here only to signal some of the key areas of connection, many of which draw on the theoretical developments outlined earlier.

First, there is a long history within medical and health geography of considering the relationship between bodies and environments or landscapes. This includes work which adopts a disease ecology approach to consider the effects of environmental toxins on health, work which adopts a more socioecological approach to considering the impact of environments on health (e.g., work on obesogenic environments: see Colls and Evans 2013), and work on therapeutic environments (see, e.g., Conradson 2005). Much of this research explores the relationship between corporeality and nature through considering the proximity of particular “natures” to bodies and their effect on health.

However, as Colls and Evans (2013) argue in relation to work on obesogenic environments, bodies, nature, and the relationship between them have been undertheorized within health geography. Similarly, Conradson (2005, 338) argues that work on therapeutic landscapes is limited by “a tendency to frame such settings
as having intrinsically therapeutic properties … to equate physical presence within a landscape with the unproblematic receipt of its therapeutic influence.” There is, therefore, a need to explore more fully the relations through which “natural” landscapes and environments affect health. One possibility might come from engagement with the areas of work outlined earlier: nonrepresentational geographies and corporeal materialities. Here, for example, there is the potential to connect work on therapeutic landscapes with geographical literatures on sensory experiences of landscapes (see, e.g., Morris 2009 on naturalism). Moreover, Colls and Evans (2013) suggest that a combination of feminist and disability theory would be helpful to think in more nuanced ways about the interaction between bodies and socio–natural environments within work on obesogenic environments. Engagement between health geography and work in political ecology offers the further potential to interrogate nature–corporeal relations here.

Second, there has recently been a growth in work within political ecology which pays attention to bodies. For example, Bakker and Bridge (2006) suggest that engagement with literature on bodily materialities might provide a way for research on resource geographies to question the nature/culture distinction without falling prey to essentialism. Recently there have been a range of calls, in particular, for engagement between political ecology and geographies of health. Here, it is suggested that political ecology might provide a means to think critically about the nature–society question within geographies of health, moving away from an approach which breaks health down into its constitutive parts to think instead in more complex and connected terms (Jackson and Neely 2014). For example, Guthman (2012) suggests that political ecology might provide a means to “open up the black box” of the body in geographical work on obesity; and Mansfield (2008) uses the example of childbirth to illustrate the productive possibilities in health geography engaging with research on nature–society. Recent work on the biopolitics of disease (Braun 2007) also offers a more politicized understanding of corporeal–natural relations in health geography.

However, Guthman and Mansfield (2013) have suggested that, while political ecology has made important strides in problematizing ideas of nature, the same critical lens has not been turned to questions of corporeality (Guthman and Mansfield 2013). They suggest that new epigenetic approaches, which explore how environmental toxins and other stimuli affect phenotypical development, may provide a conceptual framework which does more to “open up the black box of the body” here. In particular, they argue that this approach offers a route for geographers to consider the permeability of the body and to think in more complex ways about the active nature of both bodies and environments, moving beyond simple deterministic models. However, as they acknowledge, there are risks here too and, as geographical work develops in this area, critical attention on the politics of such research needs to be maintained, as well as a wariness of any neoessentialist discourses surrounding race, gender, class, disability, and so on which may be legitimized here. Elsewhere, Greenhough (2010) suggests that adopting the concept of ontological politics may be one way to bridge the gap between political ecology, geographies of bioscience, and geographies of health to allow, for example, further exploration of the ways in which new biotechnologies are refiguring ideas of biological citizenship and the multiple ways in which bodies are connected to (nonhuman) others that are relevant for questions of health and health care. Further potential lies in engagements between feminist work on the body and work in
the subfield of feminist political ecology which is drawing on work on corporeality elsewhere in feminist scholarship, to demonstrate the importance of spatial and material ecologies to the lived and embodied realities of gender and race. More broadly, this work is illustrating that political ecology needs to attend to embodiment since “it is frequently at the level of the intimate that national and international power relations are produced and sustained” (Elmhirst 2011a; see also the other papers in Elmhirst 2011b).

Finally, geographical work on corporeality and nature also includes that which explores hybrid and cyborg bodies and our relations to nonhumans. Work here is interested in challenging the distinction between nature and culture in relation to bodies by demonstrating the multiple ways in which our corporeal being-in-the-world is enabled by connections to other humans and nonhumans. There has been a particular emergence recently of geographical work which considers human relations with nonhuman animals. As Buller (2014) explains, the “animal turn” which has occurred in the past 15 years across the social sciences poses particular challenges to human geography. This work not only destabilizes distinctions between human and animal but also between nature and culture, and mind and body, as geographers consider the agency of animals.

Other work which blurs the boundaries between human and nonhuman bodies concerns new geographical work on hybrid and cyborg bodies and on (bio)technologies (Kirsch 2014). Sarah Whatmore’s (2002) work on hybrid geographies is important here. Hybrid geographies, she observes, draw together thinking through space with thinking through the body “by attending simultaneously to the inter-corporeal conduct of human knowing and doing and to the affects of a multitude of other ‘message-bearers’ that make their presence felt in the fabric of social life” (Whatmore 2002, 3; emphasis in the original). Recent work in this vein is at the forefront of explorations of the production of natures and cultures and of what it means to be a (human) body. For example, Davies’s (2006) work on organ transplantation and stem cells addresses questions about the ways in which biotechnology may be framed as being “against nature.” Elsewhere, work that adopts actor-network theory is enabling geographic work which recognizes that humans are produced through their relations with others. Within urban political ecology the concept of the cyborg (defined by Haraway as an organism–machine hybrid) has been adopted in relation to cyborg urbanism, as a way of conceptualizing “the material interface between the body and the city” (Gandy 2005, 28). This work draws attention to the networks of pipes and wires that support and enable urban life. Much work on cyborg bodies refers to Haraway’s work (1991) which, as Callard (1998) reminds us, is about broad questions of human subjectivity in which Haraway (1991, 30) asserts that “Without question, the modern evolutionary concept of a population, as the fundamental natural group, owes much to classical ideas of the body politic, which in turn are inextricably interwoven with the social relationships of production and reproduction.” The concept of the cyborg thus provides one potential means to draw together geographical work in political ecology which questions ideas of “nature” with feminist work on corporeality in order to develop what Haraway (1991, 35) proposes is a socialist-feminist theory of the body politic which avoids biological determinism, recognizing that “neither our personal bodies nor our social bodies may be seen as natural, in the sense of existing outside the self-creating process called human labour. What we experience and theorize as nature and as culture are transformed by our work.”
Ways forward

There is no doubt that work within geography that questions corporeality and nature is some of the most theoretically and empirically exciting and dynamic within the discipline. New approaches in feminist geography, non-representational geography, political ecology, actor-network theory, material cultures, and hybrid geographies offer possibilities for exciting future work which engages with the multiplicity of human and nonhuman actors to question corporeal natures in nonessentialist ways. Geographies that allow a connection between work on nature and on corporeality are, as Whatmore (2002, 7) argues, important to “keep the promise of the geographical craft alive to the creative presence of creatures and devices amongst us and the corporeal sensibilities of our diverse human being.”

SEE ALSO: Actor-network theory; Affect; Biopolitics; Biotechnology; Bodies and embodiment; Children and youth; Cultural turn; Cultures of nature; Disease and illness, political ecology of; Environment and gender; Environmental health; Feminist geography; Feminist political ecology; Health geography; Human ecology; Hybridity; Nature; Nonrepresentational theory; Political ecology; Positionality; Production of nature; Socio-nature; Therapeutic landscapes

References

NATURE AND CORPOREALITY


Nature conservation

Paul Jepson
University of Oxford, UK

Nature conservation is a label for a cultural and policy movement that encourages societies to reflect upon, engage with, and govern their relationship with the nonhuman world. Approaches to nature conservation differ between countries due to the interplay of a variety of factors. These include cultural traditions of engaging with nature, what people consider natural and the landscapes, iconic species, and practices that shape this, the level of dependency on natural resources, the type of political order, and the time period during which nature conservation has gained high-level policy attention. There are distinct national and regional geographies of nature conservation. However, since the late 1970s these have interacted with an increasingly influential international conservation regime that formulates and promotes common standards for nature conservation. These have found greatest expression in less developed countries that are recipients of loans and aid from multilateral and bilateral government funds.

The origins of nature conservation can be traced to a group of social movements that emerged during the nineteenth and early twentieth centuries in the colonies and cities of Western Europe and East-Coast North America. They were a response to a major unsettling of beliefs on the human–nature relationship interacting with social agendas responding to industrialization, colonialism, and nation-building. Particularly influential were the debates provoked by Charles Darwin’s theory of evolution by natural selection (1851) and the extirpation of superabundant species, such as the passenger pigeon, at the hands of man. The generally accepted view that God created and destroyed the species on Earth was replaced with the realization that species could decline and cease to exist as a result of human actions. Thomas (1991) sees this as one of the great transformations in Western thought and avoiding extinction has remained at the heart of nature conservation ever since.

It is possible to recognize five distinct nature conservation movements that emerged or consolidated during the late nineteenth and early twentieth century. There was however considerable flow between each of them and across space, facilitated by the culture of letter writing during the Victorian era. These movements led to transnational networks of prominent citizens and organizations advocating distinct yet complementary values that manifested in a coherent body of law and actions to protect special sites, forestall extinctions, and govern the exploitation of wildlife and natural resources.

Briefly the social movements were as follows.

1 The wise use movement, which emerged among government scientists in the colonies of Western nations (eighteenth century onwards) tasked with finding economically valuable natural resources and developing new plantation crops. They identified the link between ecological and social systems and promulgated the value that natural resources should be managed for the greatest good for the greatest number in the long term.

2 The amenity or open spaces movement, which emerged in London and Philadelphia (mid-nineteenth century) among social movements.
NATURE CONSERVATION

reformers active on issues such as slavery, animal cruelty, and workers’ rights. They promulgated the values that access to nature and countryside is necessary for the health and wellbeing of urban dwellers, and that free enjoyment of nature is a human right.

3 The wildlife movement, which emerged in New York in the 1890s among an eclectic group of prominent big game hunters who were witnessing at first hand the impacts of human expansion on the American megafauna. They enrolled European big game hunters and together promoted the values that human conquest of nature carries with it a moral responsibility to ensure the survival of threatened life–forms, and that wanton and unnecessary slaughter of wildlife is cruel and barbaric.

4 The Naturdenkmalmovement, which emerged in West European cities (c. 1905–1925) among prominent citizens active in natural history who were concerned about damage to their favored excursion sites arising from clear-felling policies (Germany) and industrial expansion. They argued that aesthetic and intellectual contemplation of nature is integral to the biological and cultural inheritance of many peoples and that monuments of nature should be guarded from ruin in the same way as great buildings and works of art.

5 The wilderness movement, which was led by writers and philosophers in the United States (late nineteenth century onwards) who were active in forging an American cultural identity distinct from Europe. They argued that magnificent landscapes evoking wilderness should be preserved as a symbol of the vigor of the lands of the New World, as places for spiritual, aesthetic, and physical exploration and rejuvenation and as benchmarks from which to assess urban/industrial modernity (Nash 2014).

These values have found varying expression and emphasis over time and in different places. However, their contemporary versions remain evident in nature conservation policy and practice today. Perhaps the most significant value embraced subsequently relates to the rise of consciousness and activism concerning the rights of indigenous people during the 1980s. In 1989 the Coordinating Body of Indigenous Organizations of the Amazon Basin (COICA) called for Western environmental groups to include indigenous peoples in their vision of the Amazonian biosphere. This struck a chord with American conservation nongovernmental organizations (NGOs) at a time when public interest in the plight of indigenous people was growing as a result of the publicity campaigns of groups such as Survival International and films such as John Boorman’s The Emerald Forest. The social value associated with this development is less clearly articulated in histories of nature conservation but might be summarized by the statement that society has a moral duty to permit traditional peoples inhabiting natural landscapes to choose their own destiny in time frames appropriate to their history and culture (Jepson and Ladle 2012).

The concept of nature conservation may have diverse philosophical and political roots but there is a relatively small suite of causes that define nature conservation in the public mind. One is saving a rare species from looming extinction. A second is the fight to prevent unnecessary and wanton slaughter of wildlife. A third is the protection of places with special natural features or spectacles and/or where the human footprint appears light or absent. A forth is the sustainable use and management of natural resources. Household name organizations such as the World Wide Fund for Nature (WWF), The Nature Conservancy (United States), The Royal Society for the Protection of Birds (United Kingdom),
and Natuurmonumenten (Netherlands) trace their origins to one or more of these causes and the conservation movements outlined above.

During the twentieth century, nature conservation interacted with broader social policy and cultural trends that shaped or reasserted national identities; this was particularly so in North America, Europe, Southern Africa, the Indian subcontinent, Australia, and Israel. The timing and nature of these interactions are country specific. For example, UK national park policy was developed in the 1940s as a component of broader policy to create a national health service and assert a national identity more rooted in country than in empire. Similarly, India’s political leaders launched Project Tiger in 1969 to foreground India’s biological heritage as part of a strategy to construct a nationalist frame separate from British imperialism. The point is that nature conservation has many national variants although the geographies of these are yet to be mapped.

Here it is useful to position nature conservation in relation to other domains of conservation with which it overlaps. The two most significant are heritage conservation and biodiversity conservation. Heritage conservation focuses on the protection and care of tangible cultural artifacts such as artwork, architecture, archaeological sites, and museum collections, but extends to include cultural landscapes many of which have nature conservation value. The overlap between nature and heritage conservation is embodied in organizations such as the National Trust (United Kingdom) and international frameworks, notably the 1972 World Heritage Convention.

Biodiversity conservation is a 1990s scientific agenda within nature conservation. Simply put, biodiversity refers to the variety of life on Earth – ecosystems, species, and genes. Biodiversity conservation emphasizes the maintenance of species diversity for utilitarian purposes such as a genetic library for the development of new compounds, as a local livelihood resource, and as the basis for functioning ecosystems that provide services essential for life – oxygen, water, flood control, agriculture, medicine, recreation, and so forth. Nature conservation, in contrast, embodies a more relational value set that aspires to a better quality of life for humans and nonhumans alike (see Table 1).

Nature conservation policy

Nature conservation, in different guises, attained high-level policy attention at least four times during the twentieth century. Over time, nature conservation has interacted with wider policy goals relating to, for example, rational resource management, public health, urban regeneration, poverty alleviation, and sustainability. This has produced complex institutional legacies and served to embed nature conservation in public policy globally. An overview of these different phases of nature conservation is important for positioning contemporary scholarship, policy, and action.

Practices of natural resource management have ancient roots. They have been practiced by traditional peoples and by royal and colonial administrations for centuries. For instance, between 1764 and 1791 colonial powers established forest reserves in Mauritius and the Caribbean that can be regarded as the antecedents of forestry systems instituted in the Indian subcontinent in the nineteenth century. New cartographic techniques developed towards the end of that century produced powerful spatial visualizations of declining forest resources. These reinforced the strategic policy case for natural resource conservation, notably in the United States under the leadership of Gifford Pinchot.
who, from 1905 to 1910, was head of the US Forest Service. Pinchot understood the importance of popular discussion in creating policy legitimacy. He conducted massive publicity campaigns and debated rational resource use with those who opposed commercialization of nature (notably John Muir, a wilderness movement leader). This, together with the wider uptake of such ideas by colonial governments internationally, served to establish natural resource management as a key area of domestic and international policy. Pinchot’s strongly utilitarian framing of nature conservation has dominated in policy and bureaucratic circles since, not least because governmental competencies for nature conservation were initially assigned to forestry or agriculture ministries (e.g., Indonesia, Ghana, Brazil).

The cause of wildlife conservation attained high-level policy attention towards the end of the colonial era, peaking with the 1933 London Convention on African Wildlife. The wildlife movement, which included among its leadership presidents (e.g., Theodore Roosevelt, United States), foreign ministers (e.g., Earl Grey, United Kingdom), royalty (e.g., Prince Leopold, Belgium), and hunter celebrities (e.g., Frederick Courtney Selous), mobilized to avoid the African megafauna suffering the same fate as the megafauna of Europe and North America. They identified three major threats facing African wildlife: the disease rinderpest; market, sport, and subsistence hunting by white settlers; and the expansion of agriculture. Their policy response was the creation of parks and reserves in sparsely populated areas still rich in wildlife. Future human settlement would be limited in order to separate wildlife from disease-carrying domestic stock and for ease of governance (Hingston 1931). The London convention led to the creation of National Parks and wildlife sanctuaries in Africa, and also British India, the Dutch East Indies (modern Indonesia), and, to a limited extent, French Indochina. The guiding nature conservation ethos of this era can be understood as “nature should be protected to be useful, uplifting, indeed ennobling to mankind” (Winks 1997, 14). Nature conservation was an expression of humanity’s ability to rise above the chaos, opportunism, self-interest, and ignorance that had, for example, characterized the frontiers of the American West.

Nature conservation aligned with high-level policy agendas a third time during the 1960s in the context of rising post–World War II affluence and increased mobility and leisure time. Policymakers recognized that the protection of natural and historic heritage was part of building (or reasserting) national and regional identities and that growing urban populations needed space to play. Outdoor recreation linked these and other policy objectives: it catered for all social classes, provided health benefits, and generated urban–rural economic and cultural flows. This was the era of multi-use conservation or “nature development,” which President Lyndon Johnson identified in his 1964 State of the Union address as one of three pillars of his “Great Society” (Whitaker 1976). This policy frame is closely associated with the US National Park model but nature-development policies were enacted in many industrialized countries globally, for example, through the establishment of national, state, or country parks and urban “re-greening” projects. Given the complex interactions of nature, landscape, culture, and politics, the expression of this policy has many national variants. It is most evident in countries exhibiting a combination of progressive social policy, an active civil society, and traditions of nature-based sport and recreation.

The fourth and perhaps most significant era of conservation policy emerged from the interaction of these earlier nature conservation frames with the sudden forces of environmentalism.
<table>
<thead>
<tr>
<th>Conservation values and their social movement origins</th>
<th>Original protected area type</th>
<th>Relevant IUCN management category, type, and definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Places evoking wilderness should be preserved as a benchmarks to assess urban-industrial modernity and for spiritual, aesthetic, and physical exploration and rejuvenation (<em>Writers &amp; artists, US, nineteenth century</em>).</td>
<td>Wilderness area</td>
<td>Ib: Wilderness Area: managed mainly for wilderness protection</td>
</tr>
<tr>
<td>Humanity has a moral responsibility to ensure that its actions do not knowingly cause the extinction of species (<em>Elite hunters, New York &amp; London, turn of twentieth century</em>).</td>
<td>Wildlife sanctuary/refuge</td>
<td>IV: Habitat/Species Management Area: protected area managed mainly for conservation through management intervention</td>
</tr>
<tr>
<td>Aesthetic and intellectual contemplation of nature is integral to the cultural and scientific inheritance of many peoples and monuments of nature should protected (<em>Prominent citizens, W. European cities early twentieth century</em>).</td>
<td>Naturdenkmal/nature monument</td>
<td>III. Natural Monument: managed mainly for conservation of specific natural features</td>
</tr>
<tr>
<td>Benchmark/representative sites are required for the study of natural systems (<em>Ecologists, Europe, mid-twentieth century</em>).</td>
<td>Nature reserve</td>
<td>Ia: Strict Nature Reserve: managed mainly for science</td>
</tr>
<tr>
<td>Access to nature and countryside is necessary for the health and wellbeing of urban-dwellers (<em>Social reformers, London, nineteenth century</em>).</td>
<td>Urban/country/state park</td>
<td>Not recognized as PAs, but see National Park</td>
</tr>
<tr>
<td>Natural resources should be managed to support livelihoods of settlers/local people (<em>colonial resource managers</em>); natural resources should be managed for the greatest good for the greatest number in the long run (<em>US foresters, early twentieth century</em>).</td>
<td>Forest reserves and game reserves</td>
<td>VI: Managed Resource Protected Area: protected area managed mainly for the sustainable use of natural ecosystems</td>
</tr>
</tbody>
</table>
NATURE CONSERVATION

<table>
<thead>
<tr>
<th>Conservation values and their social movement origins</th>
<th>Original protected area type</th>
<th>Relevant IUCN management category, type, and definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy ecosystems are necessary to safeguard economic growth, quality livelihoods, and social stability (Colonial scientists, late eighteenth century. Several reiterations since).</td>
<td>Watershed protection forest</td>
<td>See National Park</td>
</tr>
<tr>
<td>Places that symbolize the above conservation values and their associated social practices can help create or reassert national identities (Political leaders/nation builders 1860–1980).</td>
<td>National Park</td>
<td>III: National Park: managed mainly for ecosystem protection and recreation</td>
</tr>
<tr>
<td>Sensitive management of cultural landscapes evoking beauty and heritage will bring cultural, economic, and conservation benefits (Broader cultural movements).</td>
<td>Landscape protection area (various national terms)</td>
<td>V: Protected Landscape/Seascape: managed mainly for conservation and recreation</td>
</tr>
<tr>
<td>Society has a moral duty to permit traditional peoples inhabiting natural landscape to continue their way of life and choose their own destinies in time frames appropriate to their history and culture (Indigenous leaders and urban activists, late twentieth century).</td>
<td>Indigenous and sustainable use reserves</td>
<td>VI: Protected area with sustainable use of natural resources.</td>
</tr>
</tbody>
</table>

The 1970 Earth Day event in the United States signified a profound shift in public consciousness on human–nature relations and galvanized high-level political responses. Until this time urban publics had largely viewed nature conservation as a moral and rational imperative that happened outside cities and did not directly impact on their lives. Popular books, notably Fairfield Osborn’s *Our Plundered Planet* (1948) and Rachel Carson’s celebrated *Silent Spring* (1962), challenged this view by making explicit the links between resource exhaustion, environmental pollution, and human wellbeing. They had impact because the physical manifestations of their arguments – water pollution, smog, uncontrolled waste disposal, and the collapse of predatory bird populations – were becoming visible to all, urban and rural alike. This was particularly so in the United States but also in many Western European countries.

The new environmental consciousness was emerging in many cities around the world, particularly in Western Europe. For instance,
between 1963 and 1970 the UK government convened a series of consultations on “the countryside in the 1970s” figureheaded by Prince Phillip, and less populist, but nonetheless influential, environmental books were appearing, such as Max Nicholson’s *The Environmental Revolution* (1970). However, histories of the time suggest that this new environmental consciousness gained a greater political and policy saliency in the United States because it was picked up as a key issue in political opinion polls, and because the cause was taken up by a young, well-educated, middle-class generation with an activist confidence instilled by the 1960s. Importantly, however, the Nixon presidency was able to assume sudden and progressive leadership in environmental policy domestically and internationally because nature conservationists had developed the policies, leaders, and networks to support a rapid political response (Winks 1997).

In 1948, Laurence S. Rockefeller (who had been a major philanthropic investor in the US National Park movement) together with Fairfield Osborne (above) had founded The Conservation Foundation (CF) as a high-level Washington-based policy “think tank” to work on formulating a “new conservation” incorporating what we now term environmental ideas. The foundation had links with the Swiss-based International Union for the Conservation of Nature and Natural Resources (IUCN), a body formed in 1947 through the influence of European–US conservation networks and with the status of a United Nations agency. This transatlantic conservation network included astute and eminent members of the political, scientific, and bureaucratic elite. They seized the policy opening created by Earth Day and mobilized their influence and networks to convene the landmark 1972 United Nations Conference on the Human Environment from which emerged a raft of new environmental legal frameworks. CF staff assumed influential positions in the new environmental institutions that resulted: for instance Russel Train became head of the new US Environment Agency and assumed the brief for international issues, whilst Raymond Dasmann became chief ecologist of the IUCN. In this position he assumed the influential role of preparing the first drafts of international conservation policy frameworks and guidelines and helped IUCN exploit the political opportunity to position itself at the center of a reinvigorated international conservation regime. To this end the IUCN leveraged three additional assets: (i) the US National Park ideal that offered an inspiring and populist goal, along with an international network of park managers (the Commission on National Parks and Protected Areas, later renamed the World Commission of Protected Areas); (ii) the emerging science of ecology (and later conservation biology) that provided a scientific justification for policy; and (iii) incipient technologies for metricizing attributes of nature (see below).

The events of this period are recorded and discussed in many excellent histories, biographies, and memoirs (see e.g., Whitaker 1976; Winks 1997). These histories point to a significant shift in the dominant worldviews underpinning nature conservation policy, activism, and public support. Earlier confidence in the ability of humans to protect, manage, and reshape nature was replaced by a sense of despair concerning humanity’s “wounding ways” and an awareness that the biosphere could no longer absorb human impacts. A logic appeared that areas needed to be left alone where nature could take its course, both as scientific benchmarks for restoring and managing Earth systems and because humans could no longer be trusted to treat nature responsibly. Nature conservation came to be seen as an end in itself rather than as
a means of improving the quality of life of both people and nonhuman species.

Such logics are clearly evident in protected area legislation developed subsequent to the environmental crisis in less developed countries such as Indonesia and Brazil during the 1980s and 1990s. These logics interplayed with the US wilderness ethic and the rise of ecological science to create a nature conservation policy that was more biocentric and exclusionary. Some social scientists argue that nature conservation embodies a troubling dualistic concept of the human–nature relationship with eco-colonialist overtones. Given the breadth of conservation such ideas are undoubtedly present; however, histories of conservation policy suggest that the so-called nature–society dualism may have come to the fore more recently.

The European Union (EU) exemplifies some of the political complexity subsumed in the term “nature conservation.” Several, though not all, member states have nature conservation legislation and institutions that pre-date their EU membership and reflect national ways of engaging with nature. During the 1960s, environmental governance started to emerge within the EU as an issue of common trading standards, leading to the creation of a new environment directorate within the European Commission (EC). A campaign by two informal bird-protectionists committees (one German and one Dutch/UK) against migratory bird hunting offered the new directorate a transnational issue where it could justifiably claim regulatory competence in nature conservation: individual member states could not protect migratory birds and the practice of killing tired birds on the way back to their breeding grounds seemed absurd to technocrats and cruel to influential urban publics.

Figure 1 Major milestones and the growth of the world’s protected area estate. Reprinted from Watson et al. 2014 by permission from Macmillan Publishers Ltd. © 2014.

Two directives emerged, both of which were co-drafted by tiny groups of EC and non-governmental organization (NGO) policy entrepreneurs. The 1978 Birds Directive and 1986 Habitats Directive are both examples of “top-down” prescriptive legislation that require member states to incorporate their provision in national law. With EC support, “insider” conservation NGOs built the capacity of national conservation groups across the EU to form partnerships that could assist and/or monitor member state compliance and report back to the
EC. Backed up by a powerful European Court, this simultaneously empowered the EC as a supranational governance entity and contributed to the emergence of a set of conservation NGOs as influential transnational policy actors (Jackson 2013).

The various institutionalizing processes of nature conservation that arose out of 1970s environmentalism assumed a maturity and stability with the rise of the term “biodiversity” in the 1990s. This represents the fifth, and present, era of nature conservation. The main focus of nature conservation policy over the past 20 years has been the protection of the world’s remarkable biological diversity (see Figure 1). The term biodiversity was deliberately introduced in the 1980s to make politicians and bureaucrats aware of the looming extinction crisis, particularly in the tropics. The term quickly rose to prominence and was cemented in international policy at the Earth Summit in Rio de Janeiro 1992 with the formulation of the Convention on Biological Diversity (CBD). This declared the preservation of biodiversity as a major element of “sustainable development.” Biodiversity was framed as a new type of natural resource, providing raw materials for activities ranging from plant breeding to development of pharmaceuticals, that could be systematically surveyed and prospected (Haila and Kouki 1994).

The term biodiversity gave rise to a powerful new policy and scientific discourse: it transformed nature into units that could be measured, modeled, and converted into targets. The Biodiversity Hotspot prioritization scheme developed and marketed by Conservation International galvanized strategic action and investment at a global scale. Less high profile, but perhaps more influential, are the requirements of the CBD for signatory governments to integrate provisions of the convention into national law and to prepare biodiversity action plans at each administrative level. In addition, biodiversity has been “mainstreamed” into other areas of policy, such as international development, and into influential standards, such as those of the International Finance Corporation. Biodiversity emphasizes a technocratic, resourcist approach to nature conservation that has marginalized earlier moral–aesthetic arguments for conservation and the benefits these generate. It is a domain of nature conservation that has assumed a distinct, and in many ways separate, identity.

Whilst the core idea of avoiding extinction through laws and setting aside land are similar, in the biodiversity frame, conservation strategy and policy have been guided by units of biodiversity that can be standardized and thereby quantified, compared, and modeled. Among the more influential units are biological species which gave rise to the notion of mega-diverse countries, Red Lists of threatened species and endemic species (technically termed “restricted-range species”), which have become central criteria for the global identification of Key Biodiversity Areas and ecoregions. The latter construct biota-scale units that aim to provide metrics for comparison and prioritization and for the identification of representative habitats and assemblages (see Ladle and Whittaker 2011 for overview). In the biodiversity frame the conservation of so-called cultural species – those that generate value for humans in the form of wonder, curiosity, identity, outdoor recreation, and so forth – is downplayed. This is because they are difficult to classify in a systematic way.

A fuller treatment of biodiversity is outside the scope of this entry but see for example Gaston and Spicer (2013). There are indications that “biodiversity” is losing traction as a policy frame: government funds for biodiversity in the international arena are declining. This is in part because conservationists have struggled to prove their arguments that losing species will
have significant negative economic and social consequences and in part because other causes have placed new demands on finite budgets. The term “ecosystem services” is gaining traction in policy as a more holistic and communicable alternative to the term biodiversity. Ecosystem services are defined as the benefits (goods and services) provided by ecosystems that contribute to making human life both possible and worth living. The ecosystem services frame aligns more readily with ideas of natural capital than does biodiversity and it enables clear articulation of the social, environmental, and economic returns generated by investments in conservation.

A key point from this section is that nature conservation must constantly adapt and align with changing agendas if it is to retain its policy saliency. This is achieved by a cadre of conservation policy advocates, entrepreneurs, and campaigners (lobbyists) who have secured access to high-level decision forums through the deployment of a suite of influence tactics (Box 1). Maintaining and supporting a strong advocacy force requires an authoritative and robust science evidence base. Further, success translates into laws being passed, guidelines being adopted and lands being designated. Conservation groups have responsibilities regarding the implementation and monitoring of the policies they advocate, for example as technical advisors, partners in management, watchdogs, and public communicators. All of this requires funds, and how to finance nature conservation is a real and pressing issue and the topic of the next section.

Financing nature conservation

Doing conservation requires large amounts of money. A 1999 study (James, Gaston, and Balmford 1999) estimated that a comprehensive global conservation program would cost approximately $300 billion annually, yet the total annual spend from governments and donor agencies was just $6 billion. Finance for nature conservation is generated from several different sources which can be grouped into three general types: (i) funds generated by conservation organizations which they can use more or less as they like; (ii) funds from governments and foundations that are for specified purposes; and (iii) funds “released” through the creation of economic devices such as environmental tax and compensation funds. Sources of income and the emphasis of nature conservation goals interplay: pragmatism dictates that in order to keep staff employed a conservation group may need to adapt its cause to align with the agenda of donors and/or act as a service provider for government and other agencies.

Fee-paying members and subscribers are a key source of income for many nongovernment conservation groups. They provide a relatively stable and “unrestricted” flow of income and a constituency who can be asked for additional donations (including legacies and planned giving). In addition, they act as a core customer base for the development of a retail arm and fundraising products (e.g., adopt an animal) aimed at the wider public. The United Kingdom’s Royal Society for the Protection of Birds generated £90 million in 2013 from a combination of membership fees, legacies, and commercial activities. In countries with strong traditions of philanthropy, such as the United States, donations from high-net-worth individuals are a significant source of income, and cultivating relationships with such people is a core conservation activity. Major philanthropists have created foundations and trusts to hold and manage assets that are to be used for charitable purposes. These vary enormously in size but the larger ones are
Although the goals of nature conservation have changed over time there is remarkable consistency in the influence tactics deployed by people and groups advocating nature conservation. The following approaches appear to characterize nature conservation.

1 **Relationship building** – Developing relationships with officials, bureaucrats, celebrities, and employees of corporations who are conservation minded and/or who see a benefit in siding with a conservation cause.

2 **Expert advice** – Generating and communicating expert and grounded advice and analysis about the state of the natural world, the threats to it, and possible solutions to deal with these threats.

3 **Creating standards** – Developing assessment, categorization, and measurement standards that turn agreements, laws, and commitments into tangible and meaningful practices that can be audited by others.

4 **Framing conservation issues** – Shaping how conservation issues are thought about (framed) and promoting public and media interest in the issues and in preferred solutions and actions.

5 **Direct action and mobilizing constituency** – Acquiring land, managing populations of endangered species, and engaging citizens in campaigns responding to threats to places, species, and ideals.

This is not an exhaustive list. For example, it does not include taking governments to court (used in the United States and EU but rarely elsewhere) if they fail to enforce conservation-related laws. Source: Jepson and Ladle 2012.

managed by professional staff and disperse grants to achieve particular goals. These often relate to broader human development and environmental goals and a conservation group will need to adopt and deliver these goals in order to be eligible for a grant. Grants from multinational agencies may be even more prescriptive and are invariably packaged into a time-limited project that is supposed to become self-sustaining in some way after the initial period.

Given the complexity of the situations where conservation operates, interventions can rarely be packaged into planned and costed three- to five-year work plans with predictable outcomes. The project-by-project mode of conservation delivery has generated inflexibility and inefficiencies and it incentivizes organizations to develop new projects rather than follow through on existing investments. In addition, the major conservation donors want to make fewer grants, but of a larger size, to reduce administrative costs. This favors large conservation NGOs with professional finance, human resource, and project capacity. Over time, smaller conservation NGOs have been “squeezed out” and the landscape of nature conservation action has shifted from a plethora of small, grounded organizations led by conservationist committed to a particular place towards a smaller number of centralized conservation delivery organizations.
Many governments have dedicated nature conservation agencies, and funding for conservation comes from a combination of direct budget support and co-financing from specific governmental programs. For example, the EU’s LIFE Programme has co-financed 3.4 billion euros worth of projects in EU member states since 1992. Since the 2008 Eurozone crisis government budgets for conservation have decreased: the US National Parks operating budget fell by nearly 13% between 2009 and 2013 and science positions in Parks Canada were cut by 13% during the same period (Watson et al. 2014).

As indicated above, underfunding for nature conservation is nothing new and since the late 1970s conservation economists have been innovating with the creation of financial instruments that either release new money or make it easier for people to support conservation. Conservation easements were one of the first such instruments. Developed in the United States, they involve a legal agreement between a landowner and a land trust or government agency that permanently circumscribes the use of the land in order to preserve its conservation value. The landowner benefits from reducing the taxable value of the property. Further, because “easements” remain binding if the property is sold, they empower landowners to protect their land from future development.

Debt-for-nature swaps are another clever financial mechanism. The underlying principle is simple: many less developed countries owe Western governments and banks billions and this “debt burden” is a major obstacle to development and sound environmental management. Such “debt” is traded on secondary markets and can be purchased at a price significantly lower than face value. This sovereign debt can be purchased and a deal negotiated with the debtor government to redeem the debt. This might involve restricting development around a protected area but more normally it involves the debtor country setting up a local currency trust fund to finance an agreed portfolio of conservation projects. The Nature Conservancy and Conservation International have become adept at facilitating and negotiating debt-for-nature swaps that secure redemption values higher than the purchase cost of the debt.

Brazil has pioneered environmental fiscal transfer schemes (EFT) to reward local government efforts toward sustainable development. Specifically, the Ecological Value-Added Tax (ICMS-Ecológico) introduced in 1991 compensates municipalities for land-use restrictions associated with protected areas. By creating the possibility of a new income stream, ICMS-E has put nature conservation on the agenda of public administrations. Unfortunately, its actual conservation impact is less clear. This is because Brazilian tax law provides that taxes are not bound to specific expenditures. This means that it is up to municipalities whether or not they wish to pass complementary legislation that earmarks ICMS-E income for activities to protect and manage the protected areas they have designated. In addition, this instrument is rival, meaning that tax revenues to a municipality are diluted as more municipalities join the scheme and as more area is protected.

The new horizons for conservation are creating mechanisms for payments for ecosystems services (PES) (e.g., forest carbon sequestration and storage) and the development of green bonds. The imperative to turn towards markets to meet the shortfall of public conservation funds is part of a trend towards a wider engagement between nature conservation, corporations, and investment markets.
Science of nature conservation

The science of nature conservation is, like the natural world it seeks to conserve, complex and multilayered. There are many excellent books on the science of conservation. The aim of this section is to briefly introduce the types of science that characterize nature conservation and to introduce key policy and scientific ideas that have influenced the work of conservation scientists.

Broadly speaking, conservation science can be understood as an interplay of: (i) policy-orientated science that constructs attributes of nature as in need of governance and governable. This interacts with high-level policy and provides expert advice on issues requiring policy attention and strategic policy responses; (ii) management-orientated science that informs the management of species populations, habitats, and, increasingly, landscapes. This can be located on two general axes – in situ versus ex situ and passive versus active management (see below); and (iii) a science of biological recording and monitoring, characterized by the involvement of citizens and that underpins (i) and (ii). In addition, conservation science informs, shapes, and is shaped by the scale at which conservation aspires to act. In the past 20 years an earlier focus on species and sites had been expanded to include habitats and landscapes.

The discipline of conservation biology established during the 1980s (Soulé 1985) developed an influential body of scientific theory and evidence that underpins and legitimizes contemporary institutions of nature conservation. However, the spatial and temporal resolution of the data informing conservation science is undergoing something of a step change with the emergence of new bio-informatics, sensing, and recording technologies. This, together with the critical engagement of social, political, computational, and economic scientists with nature conservation policy, is producing a new interdisciplinary conservation science that is simultaneously building reflexivity and prompting innovation in conservation theory and practice.

It is important to note that policy-orientated conservation science has its roots in the systematic heritage of conservation located in museums and herbaria. As a result the species is a key unit of analysis and action: over time conservation science has produced a portfolio of identities for species that combine ecological concepts and applied conservation practice; notable among these are flagship, keystone, umbrella, and threatened species. Flagship species are those with traits (such as charisma) that galvanize action; keystone species are those that play an important role in the structuring and functioning of ecosystems. The concept of umbrella species bridges the two and describes prominent species with large-range requirements whose conservation will provide protection for other species using the same habitat.

The threatened species identity formalized in the IUCN Red Lists is without doubt the most influential species construct in conservation. The first effort to systematically assess species at risk of extinction was published in 1958. Subsequent interactions developed a set of science-based criteria against which species are assessed and placed into one of six extinction risk categories (Mace et al. 2008). The Red List created a standard that enabled a more systematic approach to conservation prioritization and the development and enforcement of species protection laws at the international and national level. In so doing it reinforced the role of the IUCN as the expert intergovernmental authority on nature conservation.
An important concept for predicting the rate of extinctions is known as the species–area relationship (SAR). As a rule of thumb, theoretical models of the relationship between species numbers and area developed from studies of islands predict that a 95% reduction in habitat area will eventually result in a loss of half of the species that live within it. MacArthur and Wilson’s (1967) influential *Theory of Island Biogeography* provided a theoretical explanation of the causes of variation in the form of SARs that can generate predictions on the impact of habitat loss and fragmentation. The relationship has been used to predict future extinctions over a large scale. For example, in 2007 the NGO Conservation International created a constantly ticking “extinction clock” by combining extinction estimates based on habitat destruction with the latest models of the impacts of climate change on biodiversity. This science constructs the notion of a “sixth extinction” event that provides a powerful argument for the protection of natural habitats and at the same time serves to raise mass awareness of global conservation issues.

The species–area relationship led to an important applied debate concerning the merits of opting for a strategy of protecting a single large or several small reserves of the same total area (SLOSS) (Diamond 1975). The SLOSS debate was never fully resolved, but many commentators lent towards the large reserve end of the debate, and this legitimated the creation of vast reserves in many less developed countries during the 1980s, often exceeding 1 million hectares or more.

The SLOSS debate interacted with the so-called “representation principle,” a guiding goal of international conservation policy (first articulated in 1976, see Simberloff and Abele 1976) that aims to create global networks of reserves that together represent within their boundaries the variety of habitats and species found on Earth. It resulted in the design of protected network systems in less developed countries with a few very large reserves, often in excess of 1 million hectares, representing major habitat types (Sumatra in Indonesia is a good example). The representation principle initiated the science of systematic conservation planning (Margules and Pressey 2000) that seeks to design optimal, efficient, and resilient reserve networks. This science is increasingly interfacing with techniques from computational biology to widen the range of taxonomic and ecological units embraced in reserve network planning and to integrate social and economic spatial data into planning.

Two important and related concepts in conservation management are minimum viable population and the “metapopulation.” The first posits that once animal populations fall below a critical number they risk entering an “extinction vortex” caused by inbreeding, stochastic, and other processes. The metapopulation concept responds to the reality that as habitats become more fragmented populations get broken up into numerous small populations that are often loosely connected to each other through a greater or lesser degree of dispersal. Conservation scientists have developed a range of techniques to study how such populations behave and what the consequences are for species management and conservation. This translates to practical advice on reserve design and building reserve networks, reintroductions and translocations, and the creation of dispersal corridors.

The involvement of nonprofessional scientists is a feature in nature conservation science, particularly in the area of biological recording. The skills and enthusiasm of naturalist were first mobilized in 1927 when the Oxford Ornithological Society organized a national survey of heronries. Organizations such as the British Trust of Ornithology (United Kingdom) and eBird (United States) have developed an authoritative
body of conservation science based on their ability to mobilize citizens, and have developed sophisticated sampling, analytical, and visualization techniques to deal with the data they collect. Citizen recorders make vital contributions to the recording of many taxa and wider environmental issues such as the spread of invasive species, tree diseases, and incidents of road kill. Emerging technologies are widening participation and the contributions of citizen scientists. For example, web platforms such as iSpot and Zooniverse offer respectively “crowd identification” and data curation tasks, whilst smartphone technology enables the automated collection of time, locality, and other data and also the possibility to automatically detect and identify species (e.g., bats and orthoptera) (August et al. 2015).

Conservation citizen science is interacting with the scientific project of creating a biodiversity informatics infrastructure to realize a future whereby decisions and actions are coordinated, comprehensive, and sustain Earth observations and systems. Biodiversity informatics involves developing the computational architecture, data standards, and culture to link diverse biological data in museums, research institutes, and environment agencies in an interoperable data network accessible via platforms (see e.g., Global Environmental Information Facility) (Edwards et al. 2011). Increasingly, citizen-generated data input directly into informatics infrastructures via national gateways and are contributing to the transformation of conservation science from a data-poor to a data-rich science.

As already indicated, a feature of the last two decades has been the emergence of a more interdisciplinary science of nature conservation. One input into this was the alignment of biodiversity conservation and international development in the 1990s and the idea that if both could simultaneously be achieved the interests of both would be served. This led to the field of community-based conservation, which draws on scholarship in the areas of common property, traditional ecological knowledge, and political ecology (Berkes 2004). The engagement of social scientists with conservation institutions and local people, combined with the rise of social justice concerns, has given rise to a valuable critique of nature conservation worldviews, policy, and practice. This initially focused on the impacts and injustices associated with relocating peoples from protected areas: Brockington (2002) coined the phrase “fortress conservation” to describe such practices. Two particular complaints of sociologists, anthropologists, and geographers interested in these issues are (i) that conservation agencies have not adopted clear policy standards to deal with voluntary or involuntary displacement and the negative social, health, and economic costs incurred by the people involved, and (ii) that protected areas and the funding that comes with them impose or reinforce idealized versions of “wild” Africa (and other continents) – further, that portraying nature conservation (in its various guises) as a public good legitimizes the removal and/or marginalization of rural people. Perhaps at the core of such critiques is the reality that the agendas of indigenous peoples invariably begin from the desire to secure legal rights to, and authority to govern, what they see as their land or natural resources. In short, there is a political agenda seeking a degree of autonomy from the state. In contrast, the nature conservation agenda has typically strengthened the power of the state in less developed nations by extending rational and centralized land-use management into remote areas and by generating international funding flows into these areas.

The “fortress conservation” critique is now considered to be an oversimplification of the diverse biophysical, historical, political, and social contexts within which conservation acts and the contingencies that result. This and related
NATURE CONSERVATION

work are, however, giving rise to an insightful and sophisticated analysis of power (state, corporate, and private), of the institutional alliances that make up modern nature conservation and the underlying concepts and terminology that frame action. One strand, led by political scientists, examines the dynamics between “big” international conservation NGOs and other actors (such as the World Bank) involved in transnational governance. For instance, research by Duffy (2006) shows how in Madagascar private conservation NGOs have become part of powerful networks involving international and state government actors that act as conduits for global conservation interest at the national scale. In so doing, Duffey reveals the complex institutional arrangements and drivers of nature conservation acting internationally and raises questions concerning the representation and influence of national (Madagascar) ways of doing conservation in these networks.

A second strand, arising from environmental lawyers, philosophers, and political ecologists, examines the ideology and biases of the agendas with which nature conservation has aligned (or is aligning). An area of active scholarship is framing and critiquing the trend towards neoliberal conservation. One question being asked is whether conservationists are embracing the force that is responsible for the degradation of the natures they hold dear. By revealing and explaining the raison d’être of market-based policy and rhetoric, articles such as Penca (2013) promote reflexivity and accountability in nature conservation, particularly among biologically trained conservationists, who hold most positions of influence.

A third strand, popular among human geographers, is problematizing perceived dualisms in conservation, such as nature/society, which they argue have shaped the development of nature conservation policy and management. This academic agenda is concerned with promoting a more “hybrid” and “cosmopolitan” approach to nature conservation. Whilst it produces thoughtful insight, this strand of scholarship is sometimes constrained by its ahistorical approach and universalizing tendencies (e.g., belief that wilderness is a dominant natural archetype globally) and its insistence on dismantling binaries that may have only limited presence given the multifaceted character of conservation (see, e.g., Sundberg 2014).

The rise of a multidisciplinary conservation science is giving new vigor and depth to debates about the underlying ideology and purpose of nature conservation. For instance, the debate on whether the purpose of conservation is to protect nature for its own sake or for the benefits it provides to humans has gained renewed salience. Proponents of a so-called new conservationist science and new environmental pragmatism argue for an approach to nature conservation that seeks to protect or restore natural habitats in areas where people live, that is open to experimentation and the creation of novel ecosystems (those without examples in present or past nature), and that engages business and investors to generate forms of nature-based development (see, e.g., Kareiva and Mariver 2012). Traditionalists retort that claims that their approach has an undue focus on pristine and fragile natures, that they disregard human welfare, and that they have dwindling public support as being poorly evidenced. Furthermore, they argue that “refashioning conservation into a set of goals that primarily advance human interests means selling nature down the river, serving neither the long term interests of people nor the rest of the species with which we share this planet” (Doak et al. 2014, 82). Others such as Marris (2014) feel that new thinking on assisted migration and ecosystem restoration, which is blurring the native/nonnative species boundary
and producing novel species assemblies, is being unfairly aligned with the human-benefit side of the argument. She argues that these approaches are meant to complement and expand, rather than replace, traditional conservation and that they can equally serve anthropocentric and eco-centric conservation goals. In a powerful and comprehensive retort, Spash and Aslaksen (2015) accuse those associated with this new wave of conservation pragmatism of reducing “moral considerability [for nature] to instrumental usefulness” and of “being oriented towards the continued expansion of a market-based economic system of capital accumulation” (251). At the time of writing, conservation science appears to be entering a period of active discursive struggle.

**Nature conservation futures**

The interdependencies between international agencies, governments, academia, and civil society described above have created influential nature conservation institutions that have been relatively stable over a period of three decades or more. However, various developments suggest that we may be entering a period of “unsettling” with uncertain outcomes.

Continuing the themes of the last section, influential new research areas in ecology and Earth systems science are working to understand past and present ecosystem function to support resilience and adaption. Such research is enabled by step changes in data, computation, and associated analytics. It is generating a new “functionalist” agenda for nature conservation – the restoration of functional ecosystems – that is bringing into focus the “compositionalist” paradigm (conserve an example of each unit) that underpins contemporary nature conservation institutions. More radically still, new knowledge on the function of past ecosystems is posing fundamental questions concerning the baselines for nature conservation. For instance, is it still relevant and feasible for EU policy to focus on the maintenance of remnants of biodiverse habitats produced by preindustrial agricultural practices? Should the management goal of the Emas National Park in Brazil be to maintain an example of the impact of Holocene extinctions of mega-herbivores or to restore as far as is feasible the lost ecosystem functions (through introduction of functional analogues)? (Galetti 2004). The Oostvaardersplassen, a large-scale ecological experiment in the Netherlands, has given such questions practical expression by (in part) restoring a European mega-herbivore assemblage that has created a “Serengeti behind the dykes”: a radical alternative vision for nature conservation that has generated huge interest and controversy.

Such debates and the ideal of nature restoration are assuming the label “rewilding” (Lorimer et al. 2015). For many young conservationists rewilding offers an inspiring and positive conservation vision: one that engages with fresh and exciting questions and that encapsulates a positive environmentalism. For those who aspire to being a conservation hero, the age of charismatic conservationists working to document and save wild things in wild places has passed. The new conservation heroes are slated to be those who restore systems through recreating trophic cascades. To achieve this vision they will need to rally against constraints imposed by twentieth-century conservation institutions and those who defend these. In short, we may witness the emergence of a radical younger generation of nature conservationists.

The growing engagement of conservation groups with markets and corporations is also likely to shape the future of nature conservation. This can be understood as a natural progression of the “pragmatic turn” in conservation during
NATURE CONSERVATION

the 1980s that gave rise to a new generation of policy approaches designed to harness the power of consumers and markets to promote pro-conservation behaviors within supply chains of natural resources. Seminal in this regard was the Forest Stewardship Council certification scheme launched in 1994 by a partnership of conservation NGOs and retailers (Klooster 2005). Subsequently, the scope has broadened to include ecological footprinting tools that benchmark the ecological performance of firms, sectors, and cities, offset mechanisms that enable firms and projects to attain “not net loss” (e.g., of biodiversity), and new conservation finance and investment mechanisms. The combination of declining governmental support for nature conservation, globalization, awareness of the transformative power of markets, and professional conservation employment opportunities at this nexus is creating a new conservation governance, one that deploys “soft law” and gains influence through a blend of reputational risk, “best in class,” and business case logics. Such neoliberal conservation is increasingly operating in dynamic relation to the older state-centric regime and is creating points of friction and innovation.

Finally, new technological forces – the Internet, computing as a utility, and low-cost, miniaturized sensors – look set to transform the science and practice of nature conservation. The conditions of the late twentieth century required a centralized organizational form of nature conservation and led to an influential community of international and state agencies and “big” conservation NGOs. This organizational form served conservation well but it also restricted participation in conservation policy and action to those with the professional qualifications necessary to secure jobs with a conservation agency. Technology is democratizing access to knowledge, data, data analytics, networks, and fundraising and making once expensive research tools affordable to (nearly) all. New specialist conservation consultancies and services providers are forming at the conservation–industry nexus and technology seems likely to widen this trend. It may result in a more “situated conservation” involving groups pursuing agendas and visions tailored to their beliefs, interests, and socioecological setting. In short, the ordered institutional form of nature conservation may become unruly and messy, and challenge the late twentieth-century ideal of a rational, coordinated international management.

Nature conservation politics
and disconnects

The organizational, institutional, professional, and ideological landscape of nature conservation is diverse. Whilst there are many areas of common ground there are deep-rooted tensions which may be intensifying in response to social, scientific, environmental, and technological change outlined above. The following five axes offer a useful framework for understanding these tensions.

1. **In situ versus ex situ.** Many field conservationists are uneasy with wildlife being kept in zoos and worry that captive breeding consumes a disproportionate amount of conservation resources, has a poor track record of success, and lets politicians “off the hook” by providing an attractive alternative. Zoo conservationists recognize that zoos are part of the cultural makeup of many cities, play a vital role in reasserting connections between people and the natural world, and view captive populations as an important insurance policy for conservation.

2. **Animal rights/welfare versus wildlife management and hunting.** Many citizens supporting nature
NATURE CONSERVATION

conservation are animal lovers for whom the welfare (and sometimes rights) of individual animals is paramount. Conservation managers and hunters are primarily concerned with the health of specific animal populations and view culling of competing species and/or sustainable hunting as legitimate practices.

3 Passive versus active management of nature. Some conservationists subscribe to the ideal of a pure and pristine nature where natural processes (e.g., fire) should be allowed to take their course without human intervention. Others believe either that, given the restricted size of remaining natural areas and populations, such an approach is too risky, or that most natural landscapes are shaped by human intervention so it is appropriate to actively design and manage natural areas.

4 State versus market governance also expressed as centralized versus situated conservation. One group of conservationists holds to the ideal of strong centralized governments embedding the conservation rationale in laws that are effectively enforced. Another group sees the world as having moved on; they recognize the declining power of government and the rise of market logics and more networked modes of governance. In their view, conservation must embrace market-governance instruments and techniques such as certification, standards, and risk management to retain its relevance and influence.

5 Moral absolutism versus moral pragmatism. This axis conceivably underpins everything. It is the distinction between those who believe that certain actions are right or wrong regardless of the conservation context or purpose, and those who believe that whether a conservation intervention is right or wrong needs to be measured by the outcome it produces for nature conservation and human society.

The issue of international wildlife trade has long been an arena where such tensions and disconnects are fought out. The issue gained renewed media and policy profile in 2014 as a result of a speech by US secretary of state Hilary Clinton that framed illegal wildlife trade as a security issue. This was followed by a high-profile conference in London convened with royal patronage. Action to curtail wanton slaughter of wildlife and market hunting is an old and enduring cause within nature conservation. The story of the extinction of the passenger pigeon – the devastation of unimaginably large flocks by market and sport hunters, white farmers, and Native Americans between 1830 and 1880 – is deeply embedded in the Western conservation psyche. The species’ plight retains mobilizing power as a shocking illustration of the impact of market forces emerging from the expansion of commerce and infrastructure into remoter regions. The conditions that led to the demise of the passenger pigeon seem to be repeating in regions of Africa and East Asia. The hunting of elephants, rhinos, and pangolins appears to be out of control and there are real fears that, a century on, these species may go the way of the passenger pigeon.

A sense of urgency pervades the campaign against wildlife trade and this is producing a tense politics. One camp argues a prohibitionist stance: elevating wildlife crime to a serious crime, banning trade in wildlife products, and imposing strict enforcement. Another camp argues a situated, pragmatic approach that blends market and regulatory approaches according to context. It supports, or is relaxed about, the principle of selling ivory from managed herds, rhino ranching, and trophy hunting so long as they are effectively governed. The former position appeals to those who lean towards the in
situ, animal welfare, passive management, state, and/or moral abolitionist poles, and the later to the ex situ, wildlife management, active management, market, and/or moral pragmatism poles of the five axes above. The “enforcement” position is far more influential politically than the sustainable use position. This is because it appeals to urban publics who often have less understanding of the context but are more politically and financially empowered. During the 1990s, national organizations within the WWF federation moved their position from the sustainable-use to the more prohibitionist position.

These debates on conservation governance approaches are increasingly extending to debates on the morality of what actions are legitimate to avoid the extinction of a species. Political economists question the so-called militarization of conservation, pointing out that the framing of wildlife trade as a form of trafficking involving international crime syndicates and as a source of finance for terrorist organizations is based on flimsy evidence and legitimates lethal intervention against people labeled as poachers (Duffy 2016). As a consequence, nature conservation is becoming a core element of a global security project, especially in areas of US geostrategic interest in sub-Saharan Africa.

More generally, there is a growing disconnect between applied and academic conservationists. These two wings operate in different institutional cultures with different incentive structures. Applied conservationists sometimes assume that the role of academic conservation is to provide evidence in support of their causes and management challenges. In contrast, academic conservationists see their role as furthering the theory of conservation science and, increasingly, to study and critically appraise the practice of applied conservation. In some areas of nature conservation, particularly the management of conservation landscapes (see rewilding above) progressive academic conservation thinking increasingly poses a radical challenge to the institutionalized practices of nature conservation formalized a generation ago. In short, the nature conservation community is increasingly both divided within and exposed to external academic critique.

Wildlife trade apart, there are concerns that objectives of higher-level nature conservation policy – biodiversity and ecosystem service conservation – are becoming disconnected from those communicated to lay publics by the media, conservation groups, and nature enterprises such as zoos, and furthermore, that older conservation approaches that foreground flagship species and iconic landscapes are being downplayed or abandoned without critical evaluation of their cultural agency and impact. The historic, institutional, and scientific dimensions of conservation foregrounded in this entry are but one component of nature conservation and may be unknown or irrelevant to the many everyday practitioners working and volunteering for conservation groups.

In many countries around the world nature conservation remains primarily a practice conducted by civil society groups, nature-based enterprises (e.g., zoos), communities, and private individuals. Many frontline employees and volunteers are vocationally minded individuals whose conservation activities are steered by a mix of belief, experience, mentoring, and organizational polices. There is a level of disconnect between these “grounded” conservationists and their professional policy and science colleagues, and between the latter group and their colleagues in marketing, fundraising, communication, and campaigning who tend to portray versions of nature conservation that play to emotion, sentiment, cultural stereotypes, and fear. A small group of eight conservation NGOs have grown into major organizations in terms of turnover, employees, brand, and reach and are led by
executives recruited for their organizational management aptitudes rather than conservation knowledge. A corporate nature conservation is emerging that is able to assume high-level policy influence and act at larger scale, but this may weaken the social movement and democratic legitimacy of conservation.

**Conclusion**

Nature conservation can be understood as an interaction of complex social, scientific, and managerial visions of the human–nature relationship that manifest in different ways in different places and at different times. It is a dynamic concept that reformulates in different spaces as a result of both synergies and tensions between interest groups within and outside conservation, and forces of alignment and appropriation that manifest through the engagement of conservation ideals with other powerful political and policy agendas. As a result, nature conservation has multiple facets and identities such that it is unwise to discuss or study it as a unitary thing. Furthermore, the geographies of nature conservation are poorly mapped and understood, in part because of universalizing tendencies among both natural scientists working to develop a body of theory and evidence to legitimize and guide conservation policy and social scientists seeking to ask critical questions about the power and influence of conservation. This account has outlined key trends in nature conservation over time and the causes that define it as a domain of public life. It suggests that future directions are uncertain due to the combined influence of advances in ecology, new technological forces, the ageing of first-generation regulatory frameworks, deeper engagement with markets, and the increasingly polarized and fractious debates within the nature conservation community.

**SEE ALSO:** Biogeography; Biopolitics; Conservation and capitalism; Construction of nature; Ecoregions; Ecosystem; Environment and the media; Environmental policy; Social movements; Zoogeography

**References**


Nature, art, and aesthetics

Deborah P. Dixon

University of Glasgow, UK

Both art and aesthetics have been crucial to the making of the modern-day geographic discipline as a set of ideas and practices concerned with understanding, and engaging with nature. Art, understood as a creative practice that reworks materials as a means of affording them an expressiveness, continues to be an object of analysis for geographers interested in the role and value of the creative industries, the unfolding of creative practices and their products in the making of place, the everyday embodiment of creative energies, and so on. Historically, however, art has been a stalwart methodological practice underlying physical as well as human geographic research. Similarly aesthetics, understood as the making sense of sensory information, has played a key role across the discipline, particularly with regard to the import of geographic fieldwork. Though this historic role is difficult to discern in the contemporary practice of physical geographers, there is a renewed interdisciplinary interest in how art and aesthetics might be a productive means of bringing together human and physical geographers in their framing of nature as a highly complex site for critical investigation. Key here is a shared interest in multisensuous encounters with nature, and how these are perceived, experienced, and cognized, as well as in how the emotive dimensions of such encounters can help to reanimate the communication of geographic knowledge to a wide range of audiences.

Art and aesthetics have long and complex historical trajectories, manifest in other disciplines as well as geography. This wide-ranging presence stems in large part from the fluidity of Enlightenment thinking and practice, wherein the sharply observed disciplinary boundaries operative today were in the process of being posited, shaped, and reshaped. The work of Alexander von Humboldt (1769–1859), for example, often noted as one of the founding fathers of a modern-day geography, is representative of an expansive mode of inquiry into the form and workings of nature that consisted of systematic measurements of particular landscape features, hypotheses as to their construction, and also meditations on the role of an aesthetic sensibility. Here, art was by no means a rarified practice that took place under the aegis of expert cohorts such as the members of the Royal College of Art (established 1769), but it was a crucial component of the research process itself. Field sketches and paintings were an important means of getting to grips with the balance of forces (represented in Humboldt's design of the isoline, for example) that animated terrestrial phenomena, and that undergirded the universe. Humboldt's enrolment of artistic practice, and his appreciation for the sublime, as well as his desire for measurements of sensible phenomena, were both key to his philosophy of a holistic nature (Buttimer 2001). In similar vein, the work of G.K. Gilbert, often described as an inspiration for the mid-twentieth century's quantitative revolution in geography, illustrates an expansive mode of inquiry that draws on field sketches and photography as a means of making sense of a multisensuous engagement with nature (Dixon, Hawkins, and Straughan...
For both scholars, artistic practice was a critical component of scientific fieldwork and knowledge building. Notwithstanding this common grounding in artistic practice and aesthetics, the conceptual and methodological cleaving of the discipline in the twentieth century has largely placed both outside the remit of current physical geography research. Sensuous encounters with nature remain an important part of the fieldwork process, but are very much cast as the source from which systematic, methodological, and technically mediated procedures can produce data. Such data are of value insofar as they lend themselves to hypothesis testing. Furthermore, while this approach to aesthetics sees some value in artworks, it is their supposed accuracy in rendering the world they are understood to mimic that enables them to function as proxy datasets. As some physical geographers have observed, the emotive dimensions of wonder and curiosity that were so explicit in earlier eighteenth- and nineteenth-century geographic writing have largely been written out of the field (Baker and Twidale 1991).

It is unsurprising, then, to find that much of the debate on art and aesthetics has been carried out by human geographers, particularly those interested in cultural phenomena and practices, and that some have argued that such an interest is a useful, even necessary, antidote to an arid formalism that underlies physical geographic renderings of nature (Tuan 1993). For the most part, such geographers have sought to understand how art and aesthetics have been deployed by particular cohorts to frame and engage with nature. For some, this is primarily a matter of inquiring into how landscape has been creatively portrayed and shaped, insofar as this object of analysis has long been configured within the discipline as an interface between nature and society. Landscape paintings and literatures, as well as landscape design, are considered as testament to the diverse manner in which nature is perceived and interpreted. Techniques drawn from the formal arts and humanities, wherein emphasis is placed on the expert decoding of an artwork (according to style, genre, and the artist’s intention) resonate here with geographers’ accounts of reading the landscape with an “expert eye” and the role of human agency in affording meaning (Duncan and Ley 1993).

By the 1990s, however, as the deconstructive possibilities of Marxism, feminist theory, and poststructuralism teased apart nature as a contested term, landscapes were inserted into broader social relations of power and inequality (Daniels 1989). Moreover, other objects of analysis, such as sites and bodies, were to become a matter of interest to geographers working on art and aesthetics. In part, this was because sites were regarded as offering more complex and nuanced accounts of a nature that was not separate from society, but thoroughly entangled with it, such that a myriad of asymmetrical relations between and across species, as well as between and across organic and elemental domains, were seen to operate through, and to help comprise, such sites. In similar vein, bodies were no longer considered so readily distinguishable as human or nonhuman according to set taxonomies. Seemingly different bodies could have shared capacities that undercut the notion of agency, and even artistry, as a thoroughly human endeavor. But, also, the character of the relations between various bodies, such as symbiosis, was often considered as more important in a given context than shared biological antecedents. Such an approach belies the easy categorization of nature as composed of individual creatures rendered distinctive by their biology and occupying varying ecological niches. At the same time, in the humanities, the traditional, formal distinctions that maintained art and nature as separate
spheres were also giving way to more entangled understandings of the creative process, and the role of the nonhuman therein.

As what came to be termed a “postnatural” approach has become more prominent across academia, so the traditional figure of the “dis-attached” expert has largely been replaced with that of the immersed scholar undertaking research as a form of practice. Sensuous geographies, emotional geographies, and affective geographies, for example, have all foregrounded aesthetics as a matter of exploring how a “felt” experience of the environment is shaped, and with what import. Part and parcel of such accounts is an attentiveness to the doing of research that draws geographers close to debates in the arts and humanities on ecology, performance, creativity, and experimentation (Hawkins 2013). Indeed, an emerging trend within the discipline has been for geographers to collaborate with artists on specific projects using diverse mediums and with diverse venues and audiences in mind. What is more, there is a burgeoning series of “in the field” collaborations between physical geographers and artists focused around the mutual exploration of landscapes, and the harnessing of the emotive dimensions of such body–environment encounters in the production and communication of geomorphological knowledges.

At the moment, such efforts are heavily invested in a more traditional framing of human–environment relationships, wherein both individual human agency and meaning-making remain a central concern. Looking forward, however, there is substantial room for physical geographers to incorporate more critical accounts of the “nature” of human being and, in the process, reassess the practice of fieldwork as well as their understanding of how physical processes impact on, and are shaped by, people and groups. Likewise, there is room for human geographers to pay more attention to nature not as a bundle of forces, materials, and events awaiting human perception, experience, and analysis, but as existing in and for itself.

**SEE ALSO:** Cultures of nature; Fieldwork in human geography; Imaginative geographies; Nature and corporeality; Posthumanism

**References**


**Further reading**

NATURE, ART, AND AESTHETICS


The nature of geography

“Nature,” by definition, includes but also exceeds the meaning of “environment.” To consider how geographers have studied nature is, necessarily, to consider the discipline of geography in its entirety (rather than just some of its sub-branches). Why so? First, in various ways the first university geographers saw their task as the study of human–environment relations – with humans perceived as more or less responsive (in a physical and psychological sense) to natural stimuli issuing from their immediate surroundings. For decades, academic geography was in some measure about “Earth writing” in a literal sense: that is, the analysis of how the Earth materially inscribes itself on humans, and vice versa. Second, even though geography splintered into a largely nature-free “human geography” and an environment-focused “physical geography” after 1945, nature necessarily served as a “shadow concept” organizing the former. In other words, by presuming that something called “society” was different in kind from something called “nature” (human and nonhuman), human geographers were able to justify a largely exclusive focus on the spatial organization of people’s activities and creations. This focus was only possible by believing that the biophysical world could be “bracketed-out” and left to physical geographers to study – making it what some theorists call a “constitutive outside” or “absent presence” in a semantic sense. Third, in recent years many human geographers have taken a deep interest in the more-than-human dimensions of the world, while others have concurrently challenged conventional understandings of “the social” and “the human.” Though it has rarely involved collaborations with physical geographers, it has helped to dissolve the nature–society dualism that has for decades organized the academic division of labor in geography. Fourth, also in recent years so-called environmental geographers, who position themselves between human and physical geography, have increased in number. In large part this is because of the “global environmental change” research agenda now central to many funding agencies and intergovernmental organizations like the United Nations. But it also reflects the growing number of people who live in areas prone to perennial or periodic biophysical threats, like wildfires and floods. In short, whether our focus is on the discipline’s past or present, to write about the analysis of nature in geography is to contemplate the nature of geography. Despite this, there is no detailed history of how the two things have evolved over time (though see Castree 2005 for a sketch).

Since we cannot possibly cover such a large and complex terrain in a single entry, a parsimonious approach is required, with all the risks of oversimplification this entails. We begin in the obvious place by defining nature and offering some wider considerations about its place in Western understandings of the world. We then highlight ways in which nature was a central object of description and explanation in the early decades of academic geography, before summarizing developments after 1945, when it
became the almost exclusive preserve of physical geography’s various sub-branches. The fourth section focuses on the recent period in which, it is shown, “naturalist” and “denaturalizing” approaches developed side by side in geography, often without much contact.

Overall, we will discover that “nature” provides an illuminating conceptual lens through which to view both the evolution of geography and current fault lines (analytical and political) in the discipline. It will be suggested that geography has always had a “problem” with nature. Disagreements over what is and is not natural, who should study it, and how, have, it will be shown, been flashpoints for the successive reconstitution of geography as an academic enterprise since the discipline’s inception. Today, arguably, we are at a point where the ontological “obviousness” of nature’s existence is so widely doubted that the nature of geography may (again) be changing quite significantly. Far from “solving” the problem, this “nature skepticism” in its various current forms is symptomatic of ongoing struggles in geography to capture the hearts and minds of practitioners, students, and nonacademic stakeholders of various stripes. It only constitutes “progress” in how geographers think about nature seen from a certain viewpoint. Accordingly, we give previous approaches their due rather than rushing past them to focus mostly on recent developments, as if they eclipse all prior contributions. Most of these previous approaches share a philosophical commitment to nature’s naturalness (a commitment we will call “traditional”). As will be seen, more recent research aims to be post-traditional by questioning nature’s naturalness in various ways. Readers should note that there is a bias toward Anglophone research in what follows, and the analysis does not cover developments in continental Europe or elsewhere.

**What is nature?**

Nature is a very old word in the English language, and its meanings have varied in the detail through time. It refers to a wide range of phenomena in a plurality of different ways – this is why it is an unusually complicated word; some would argue the most complex of all. As a linguistic philosopher might say, it is a “signifier” (word, symbol, or sound) that has more than one “signified” (a specific meaning) and these signifieds are attached to an astonishing number and range of material things (“referents”), both human and nonhuman. Today, the term has four principal meanings which have endured through many decades. First, it designates the nonhuman world, especially those parts untouched or barely affected by humans (“the natural environment”); second, the entire physical world, including humans as biological entities and products of evolutionary history (or, as some would have it, a deity); third, the power or force governing some or all living things (such as gravity or the conservation of energy); finally, the essential quality or defining property of something (e.g., it is natural for birds to fly and fish to swim). As a shorthand, we can (respectively) call these meanings “external nature,” “universal nature,” “superordinate nature,” and “intrinsic nature” (see Figure 1).

Clearly, depending on the context of usage, the idea of nature can function as a noun, a verb, an adverb, or an adjective; it can also be characterized as object or subject, passive or active. As geographer Neil Smith once put it, “Nature is material and ... spiritual, it is given and made, pure and undefiled; nature is order and it is disorder, sublime and secular, dominated and victorious; it is a totality and series of parts, woman and object, organism and machine. Nature is the gift of God and ... a product of its own evolution; it is a universal outside [human] history and also the product of history, accidental
and designed, wilderness and garden” (Smith 1984, 11). In short, the word nature is – and has long been – promiscuous: not only polysemic and polyreferential, but utilized in a wide array of everyday and more specialized situations. We might therefore say it is a keyword rather than a buzzword.

Keywords, as cultural analyst Raymond Williams (1976) argued in his famous book of this name, have three characteristics. First, they are “ordinary,” which is to say used widely and frequently by all manner of people in all manner of contexts (private, commercial, and civic). Second, they are enduring rather than ephemeral – they do not come and go in a way that buzzwords like “globalization” or “post-modernism” do. Finally, keywords possess what cultural critic Tony Bennett and colleagues, in their update of Williams’s book, call “social force” (Bennett et al. 2005, xxii). In other words, because their various meanings become normalized in a given culture they are able to govern (i.e., steer or direct) not only our thinking but also a wide range of practices resulting therefrom. A simple measure of the importance of “nature” as a signifier is to imagine us dispensing with the term and its meanings altogether. The “hole” in our language would be enormous. We would be rendered both inarticulate and incapable in large areas of our thought and action. In sum, if we did not already have the term in our present-day vocabulary, we would probably have to invent it.

This becomes even more obvious if we consider two other things. First, Williams rightly referred to “particular formations of meaning” (1976, 13; emphasis added) when explaining to readers how they might interpret a signifier like “nature.” His point was that in any given society at one historical moment there are likely to be a family of keywords whose meanings bleed into, and borrow from, one another. We can say that the concept of nature is not exclusively associated with the word “nature.” Instead, the meanings are routinely signified by a range of other words that are (or have become) part of our collective vocabulary. In this sense, “nature” is something of a “ghost that is rarely visible under its own name” (Olwig 1996, 87). Its meanings often appear as collateral concepts (Earle, Mathewson, and Kenzer 1996, xvi). That is to say, they are signified by different keywords which refer us to similar or additional referents.

In the early twenty-first century nature’s collateral terms include the following, among others: “environment,” “wilderness,” “gene,” “genius,” “biology,” “race,” “sex,” “biodiversity,” “animal,” “life,” “intelligence,” “human,”

| Figure 1 | The principal meanings of the word “nature” in contemporary Anglophone societies. Source: Reproduced with permission from Castree (2014, 10). |
“instinct,” “blood,” “reality,” “climate change,” “mind,” and “ecosystem.” Some of these terms are relatively old; others relatively new. Some appear semantically simple and straightforward (though, in actuality, they are not); others more evidently complex. The precise ways in which they partake of the meanings of “nature” varies, according to both the word and the context of reference. Most of them feature in discussions of both human and nonhuman nature, but some are used more exclusively. Most of them are also freighted with meanings that go beyond those connoted by the term “nature.” These collateral words are thus only partly – rather than exclusively – synonyms for the last. It depends entirely on the circumstances of their invocation and usage (see Figure 2).

Second, circumstances aside, nature and its collateral concepts are members of a fairly select family of antinomies that can be said to structure Western thinking about the nature of reality. These antinomies include urban–rural, raw–cooked, wild–civilized, and authentic–artificial. Their meanings are best understood relationally rather than separately. Together they comprise a “semantic rule book” that at some level governs specific acts of verbal and visual representation. They are the epistemic means by which Westerners have assured themselves that there is a measure of inherent order to the world. The binaries create boundaries and differences that are presumed to be hardwired into reality, in part “by nature,” in part by design (see Box 1).

If all this seems to complicate matters too much, there are some signals in the noise. Arguably, whenever we use the word “nature” and its proxy terms, we typically think we are making (or anchoring) ontological statements of a cognitive, moral, or aesthetic kind. That is to say, we believe we are making (or vouchsafing) statements about a biophysical reality that exists independently of the words, concepts, and terms we use to make linguistic sense of it. In this respect, we are apt to assume that “nature” and its filial terms are “mimetic concepts”: that is,

---

**Figure 2** Nature and its collateral concepts. The principal meanings of the word “nature” (i.e., what it signifies) are routinely attached to all manner of material referents by way of other words (i.e., collateral terms). The words and meanings become conjoined in often complicated ways. However, these collateral terms may also signify meanings beyond the four signified by “nature.” Source: Reproduced with permission from Castree (2014, 18).
Box 1 Fundamental dualisms of Western thought since the European Enlightenment. Depending on the precise context of their use in acts of communication, each of the terms in one column can imply its opposite term in a “hard” or “soft” way. Some of the terms, again depending on the context of use, have ambivalent meanings, slipping and sliding between and across both columns. Reproduced with permission from Castree (2014, 24).

<table>
<thead>
<tr>
<th>“Hard” and “Soft” Antinomies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature ↔ Culture and nurture</td>
</tr>
<tr>
<td>Environment ↔ Society</td>
</tr>
<tr>
<td>Sex ↔ Gender</td>
</tr>
<tr>
<td>Wilderness ↔ Cultivated, cleared, and settled land</td>
</tr>
<tr>
<td>Race ↔ Social identity</td>
</tr>
<tr>
<td>Biology ↔ Social conventions and practices</td>
</tr>
<tr>
<td>Genes ↔ Cultural norms, habits, and rituals</td>
</tr>
<tr>
<td>Organism ↔ Social group</td>
</tr>
<tr>
<td>Real ↔ Fake</td>
</tr>
<tr>
<td>Instinct ↔ Free will</td>
</tr>
<tr>
<td>Rural ↔ Urban</td>
</tr>
<tr>
<td>Countryside ↔ City</td>
</tr>
<tr>
<td>Natural world ↔ Built environment</td>
</tr>
<tr>
<td>Emotion ↔ Reason/rationality</td>
</tr>
<tr>
<td>Woman ↔ Man</td>
</tr>
<tr>
<td>Mother ↔ Father</td>
</tr>
<tr>
<td>Savagery ↔ Civility and civilization</td>
</tr>
<tr>
<td>Object ↔ Subject</td>
</tr>
<tr>
<td>Reality ↔ Representation</td>
</tr>
<tr>
<td>Authentic ↔ Artificial</td>
</tr>
<tr>
<td>Raw ↔ Cooked</td>
</tr>
<tr>
<td>Wild ↔ Tame, domesticated, or disciplined</td>
</tr>
<tr>
<td>Traditional ↔ Modern</td>
</tr>
<tr>
<td>Physical ↔ Mental</td>
</tr>
<tr>
<td>Body ↔ Mind</td>
</tr>
<tr>
<td>Matter ↔ Spirit, ideology, and belief</td>
</tr>
<tr>
<td>Innate ↔ Manufactured</td>
</tr>
<tr>
<td>Animal ↔ Human</td>
</tr>
<tr>
<td>Fact ↔ Fiction, conjecture, and speculation</td>
</tr>
<tr>
<td>Truth ↔ Falsity</td>
</tr>
<tr>
<td>Ontology ↔ Epistemology</td>
</tr>
<tr>
<td>Observed ↔ Observer</td>
</tr>
<tr>
<td>Disorder ↔ Order</td>
</tr>
</tbody>
</table>

ones whose meanings capture in words actually existing phenomena that are “out there” (or “in here” if we are discussing our own physiology and neurology).

As we will see below, there is ample evidence of this being the case in academic geography across the decades. However, despite its apparently “obvious” existence, there is another
way of understanding what we call nature. Consider that some cultures have neither the word nor all (or even some) of the four meanings itemized above. For instance, in his study of several aboriginal societies, anthropologist Tim Ingold (1996) came to the following conclusion: “[H]unter-gatherers,” he suggested, “do not, as a rule, approach their environment as an external world of nature that has to be grasped conceptually and appropriated symbolically within the terms of an imposed cultural design …; indeed, the separation of mind and nature has no place in their thought and practice” (p. 120). Ingold’s conclusion chimes with that of Williams. As Williams (1980, 67) put it:

Some people, when they see a word, think that the first thing to do is to define it. Dictionaries are produced … and a proper meaning is attached. But while it may be possible to do this, more or less satisfactorily, with certain simple names of things, it is not only impossible but irrelevant in the case of more complicated ideas. What matters in them is not the proper meaning but the history and complexity of meanings.

Following Williams, we can suggest that all four meanings (signifieds) of the term “nature” are purely conventional, not once-and-for-all “correct.” Likewise, we can argue that there is nothing “natural” about the fact that the term refers to all the particular things it does, and not to others. If we want to know what nature is, and why we value or exploit it in the ways we do, we should, perhaps, look not to nature itself but to our ideas about what we call nature. When we talk about “nature,” Williams famously opined, we are talking about ourselves (whoever “we” happen to be) without necessarily knowing it or admitting it.

This is why I have devoted this section to discussing the concept of “nature,” rather than assuming that it is secondary to the material world it is intended to depict. Is this to suggest that “nature” does not exist, only the concept and its proxy terms? In one sense, yes. This is not the same as arguing that the material things to which the concept refers do not exist: they assuredly do. But it is purely a matter of habit to call them “natural” in the various ways and contexts that we choose to; indeed, it is arguably a matter of convention to divide the world into words and things, symbols and reality, mind and matter. For many (perhaps including some readers of this entry), this is no doubt a difficult argument to accept. It accords a lot of importance to language, its social origins and its practical effects, and it challenges the conventional idea that many or most concepts are semantic “mirrors” that faithfully represent the material world. As noted above, nature, by definition, seems to be that which lies outside of, and is irreducible to, linguistic frameworks. But appearances can be deceptive. In the end, some (like Ingold and Williams) would argue there is nothing “natural” about our habit of designating certain things as belonging to “nature,” nor about using those things as a reference point to secure our ethical or aesthetic beliefs. Conversely, it follows that any new habits, for example, ones that would dispense with concepts like “nature” and “society” altogether, cannot simply be justified with reference to the “realities” they supposedly respect. They too can be seen as contingent creations rather than necessities. This sort of “denaturalizing” approach to understanding nature has, as we will discover later, loomed large in human and environmental geography in recent years.

**Geography and nature: the early decades**

It is fair to say that the early university geographers took the ontological existence of nature as a given – it was considered “natural” in one or more of the four senses identified above. Their
concern was with phenomenal nature, rather than nature at the micro or cosmic scale. For academic geography’s founders, the interesting question was the various ways that this nature affected different societies across the globe. From the outset, geography was to be a “bridging subject” devoted to intellectual synthesis. This was abundantly clear in a key 1887 address by the young Halford Mackinder to the Royal Geographical Society in London. Entitled “On the Scope and Methods of Geography,” his lecture explained how and why geography should take its place alongside other disciplines within the academic division of labor. His strategy, at once simple and audacious, was to call that division of labor into question. Geography, Mackinder argued, can “bridge one of the greatest of all gaps,” namely, that separating “the natural sciences and the study of humanity” (1887, 145). He was not alone in defining geography as “the science whose main function is to trace the interaction of man [sic] in society and so much of his environment as varies locally” (Mackinder 1887, 145). At points east and west others were doing much the same, such as William Morris Davis in America, Paul Vidal de la Blache in France, and Friedrich Ratzel in Germany. The four men soon occupied important university positions and were followed by similarly vigorous proselytizers who quickly built on the foundations their forebears had laid.

So began geography’s career as a university subject and what historian of thought David Livingstone (1992, 177) called “the geographical experiment.” This was the attempt to keep nature and society under “one conceptual umbrella” in the face of “the incipient Balkanization of knowledge that accompanied the professionalization of scientific specialities” (Livingstone 1992, 177). There were two aspects to this. First, the likes of Mackinder, de la Blache, Ratzel, and Davis saw the need to study nature as a whole, not as a set of discrete parts. Where subjects like chemistry, physics, and botany specialized in investigating select elements of the natural world, geography would study all these elements in combination (as Alexander von Humboldt had famously sought to do early in the nineteenth century). This is what “physical geography” was, according to earliest proponents like Mary Somerville. In her 1849 book of the same name, she defined it as “a description of the Earth, the sea and the air, with their inhabitants the distribution of these beings, and the causes of their distribution” (p. 1). Her successors, influenced in part by Charles Darwin’s path-breaking theory of evolution by means of natural selection (1859), were sometimes given to discussions of “human nature” and how it varied geographically in response to the conditions of regional physical environment. Second, this commitment to studying nature as an integrated, multifaceted system was accompanied by a desire to explore its two-way (“vertical”) relationships with human societies in different places and regions (“horizontal” differentiation). For Mackinder and the other early geographers, it was important that nature be studied in context as something that forms the basis of (and is affected) by human practices of an economic, cultural, and political kind. In this sense, the study of human geography was to be deeply materialist, with its roots – quite literally – in the soil. The resulting investigations of “human–environment” relations could be conducted at a range of geographical scales from the local right up to the global.

Clearly, geography had high ambitions in its fledging years as a university subject: it was, in terms of subject matter and scope, very much a “world discipline.” But it became increasingly clear that its “bridging” aspirations were difficult to realize satisfactorily. The major problem was a practical one. Geography’s perspective
on the world was so comprehensive, and its subject matter so compendious, that it proved very difficult to demonstrate causal connections between the component parts of the nonhuman world, let alone all these parts and various societies worldwide. Some, like Mackinder and Ratzel, had hoped that a geographical equivalent (or version) of the theory of evolution might be developed. But this hope was in vain. It was time-consuming enough to provide mere descriptions of different societies and their physical environs, never mind plausible explanations. As a result, most early research publications by geographers were beset by what—with hindsight—were serious intellectual weaknesses. For instance, monographs that focused on specific places or regions were often impressionistic by present-day standards and filled with unverified speculations about how and why nature and human society were as they were in given situations. At worst, this shaded into what we would now regard as racism founded on supposed “natural differences” between humans rooted in biophysical conditions and their evolutionary imprint. For instance, in the United States, geographers like Ellen Semple and Ellsworth Huntington were apt to argue that certain physical environments produced human “races” less intellectually or physically capable than Europeans. These arguments were presented as descriptions of “historical accidents” that had blessed some races but not others. Indeed, both Semple and Huntington on occasion claimed that even “Europeans” would “decline” if they had evolved in a region like West Africa.

Such beliefs were consistent with certain versions of evolutionary theory—such as the work of Jean Baptiste Lamarck (1744–1829)—though certainly not with Darwin’s. With hindsight, they appear as academic attempts to justify the colonialism and clientalism—military, economic, politician and social—practiced by West European countries and the United States of America in Africa and elsewhere. Through all this, nature was understood as both universal and external, intransigent and enabling, restrictive and productive, in ways that were ultimately contradictory. Where an apparently “hard” environmental determinism was operative on some societies in some places, other societies elsewhere seemingly experienced only a “soft” determinism. Here environmental conditions were said to produce either a “superior” human nature or else “civilizations” that had, in effect, built a rich house of “culture” atop the inherent genetic and physiological universals of human biology. Intellectually, it was an approach that allowed Anglophone and continental European geographers to have their colonial cake and eat it too.

Post-1945 fragmentation: the emergence of a nature-free human geography, an environment-centered physical geography, and a small disciplinary middle ground

World War II was a turning point for geography and its approach to the study of nature–society relations. Many who would subsequently gain positions in university geography departments served in the military between 1939 and 1945. The experience was formative for most, instilling a belief that precision, measurement, and rationality were virtues to which one should aspire. Aside from the skills of mapping, surveying, and close field observation, pre-war geography struck many observers as offering precious little to the war effort. At the same time, three other developments were significant. First, the discipline had failed to produce major books or intellectual innovations comparable to, say, John Maynard Keynes’s 1936 General
Theory of Employment, Interest and Money. This became a cause for concern. Second, many outside geography had successfully argued that the physical sciences and the social sciences (with the humanities) had to be different by virtue of their subject matter. There could be no overarching analysis of people and nature, it was argued, because the former possessed ontological properties quite different from rocks, rivers, or ravines. For instance, humans are self-reflexive, linguistic, tool-making beings able to make their own history and geography – so argued philosopher Wilhelm Dilthey, among others. This argument drove a wedge between the two spheres – society and nature – that the original geographers had sought to bring together in a single intellectual frame. Finally, the embarrassment of “environmental determinism” – the pre-war argument made by Semple, Huntington, and others that some human “races” were mere reflexes of climate, soils, and resources – made some geographers determined to “raise their game” intellectually.

After 1945 academic geography progressively splintered into two major halves (human and physical), with each fragmenting into relatively discrete “systematic” subdisciplines. Consider some of the new studies of the natural environment published at the time. Bagnold’s (1941) The Physics of Blown Sand and Desert Dunes inquired into physical process-phenomenal form connections in arid environments (Bagnold’s British military service had been in dryland regions). R.E. Horton (1945) used his engineering background to argue that the action of water over and through different types of soil and rock had consistent physical consequences that could be measured empirically – and even predicted. Finally, Strahler’s (1952) “Dynamic Basis of Geomorphology” argued strongly that physical geographers should measure and explain how processes defined by universal laws create specific sorts of landforms, given a certain set of “initial conditions.” Together, works like Bagnold’s, Horton’s, and Strahler’s laid the groundwork for a physical geography where explanations were derived from the testing, by way of repeated observation and measurement, of refutable hypotheses. This was an altogether more specialized, more rigorous, less descriptive approach to the physical environment than almost anything found in pre-1939 geography. Within a decade key texts like Fluvial Processes in Geomorphology (Leopold, Wolman, and Miller 1964) were making this new kind of physical geography a serious proposition.

The turn to specialization was undertaken in the hope that human and physical geography could become “spatial sciences.” Scientists would discover – through careful measurement, experimental control (where possible), proposition testing, and use of statistics – laws explaining spatial patterns (such as the common tendency of rivers to meander, or migration flows to be inversely proportional to the size of destination cities). This implied a splitting of geography’s subject matter into two, with intellectual unity (many hoped) maintained at the level of the perspective taken on the subject matter (rather than subject matter per se). Accordingly, human geography increasingly abstracted the analysis of political, economic, social, and cultural practices from their biophysical integument. The pre-war fondness of some geographers for discussing “human nature” (in the biological sense of mind and body) was also quickly abandoned. On the other side, physical geographers produced increasingly “scientific” descriptions, explanations, and even forecasts of Earth surface phenomena. Specialization, new databases, new remote sensing capabilities, and new computer technologies made this possible. But lingering aspirations to synthesis paid the price: physical
geographers divided nature (in the sense of the planet's physical environment) into the five areas that comprise the field to this day (namely, geomorphology, biogeography, climatology, hydrology, and Quaternary environmental change). There was also a progressive move toward small-scale, short-time horizon studies because hypothesis testing was very difficult for macroscale analyses of (say) whole ecosystems or river basins. Where macroscale physical geography persisted, it was often hived off to other academic subjects – as happened with meteorology and climatology. Indeed, physical geography's major post-1945 branches bled into cognate subjects – such as ecology in the case of biogeography. This made them increasingly interdisciplinary and reflected their inability to police their own turf once the idea of a unified physical geography was progressively abandoned after 1945. Finally, all of the above meant that the study of human–environment relations became a minority pursuit, with “spatial analysis” and the search for general laws, models, and theories geography’s new modus operandi.

These changes together bolstered geographers’ intellectual self-esteem and improved their external image within the world of higher learning. Ironically, though, geography was effectively abandoning the study of human–environment relations at the very moment when the title of William Thomas’s 1956 book, *Man’s Role in Changing the Face of the Earth*, was becoming as obvious as it was profound. The Brit C.P. Snow’s famous complaint about the estrangement of “the two cultures” – one literary–humanistic, the other scientific–rational – was (ironically) applicable to postwar geography, the one subject that had made intellectual unity its *raison d’être*. The long shadow cast by pre-1945 environmental determinism made many younger geographers wary of inhabiting the “middle ground” between a newly scientific physical and human geography.

Notwithstanding all this, a number of geographers did try to occupy this terrain in ways that eluded the shadow. For instance, after World War II, many Western governments adopted a more hands-on approach to public welfare. This included a new desire to protect people from the effects of natural hazards, such as hurricanes, droughts, and landslides. In this context, the American geographer Gilbert White pioneered an approach in which people’s perceptions of hazards became the major focus. In *Human Adjustment to Floods* (1945) and subsequent works, White argued that many individuals and communities living in high-risk locations did not necessarily perceive themselves to be vulnerable and so failed to take adequate measures to mitigate the effects of hazard events. In this way, people’s cognition was granted a degree of independence and flexibility rather than being assumed to bear the “objective” imprint of their environs. Ian Burton and Robert Kates, among others, built on White’s approach and sought to identify different forms of “cognitive rationality” specific to certain groups in certain hazardous locations (Burton, Kates, and White 1978). The presumption was that people’s perceptions of hazards, however inaccurate or distorted, could be typified and rationally explained, leading to tailored policy solutions that might better protect them and their livelihoods. This kind of geographical research was important in US and Canadian geography, with natural hazards researchers being directly involved in public policy agendas for hazardous regions in both countries.

Meanwhile, cultural ecology came to prominence by the early 1970s. Situated on the marchlands between cultural anthropology and cultural geography, this approach to society–nature study was pioneered by Julian Steward, Andrew Vayda, Roy Rappaport, Marvin Harris,
and Clifford Geertz. It was a critique of anthropology’s mid-twentieth century “culturalism” – wherein cultural habits were thought to develop sui generis – and pre-war geography’s muscular naturalism. Typically focused on land and water-based communities in the Global South, cultural ecologists sought out those aspects of cultural belief and practice that seemed to be “functional” adaptations to local ecology. In this way they re-materialized culture to its biophysical base, without succumbing to environmental determinism or racism/cultural chauvinism.

The typical cultural ecologist would undertake detailed, long-term fieldwork into environmental usage and modification in one locality, as well as everyday and ritualistic cultural practices. The result was a set of holistic studies in which the metaphor of “homeostasis” – borrowed from systems theory – loomed large, and in which “culture” and “ecology” were regarded as codependent, mutually adjusted, relatively stable, and internally complex domains of process, relationship, and event.

Though cultural ecology was synthetic and hazards geography more specialized, they both prioritized immersion in local situations and eschewed the macro-focus of earlier human–environment geographers. Yet they also both, in their own way, held fast to the realism of their predecessors: “nature” was taken as an independently existing domain that could be studied carefully to ascertain the threats, constraints, and opportunities it afforded different people in different locations. This realism they shared with the newly scientific branches of physical geography. Though this conformed to the common sense of the time, the cultural ecologists, and hazards geographers’ efforts did not arrest the growing human–physical divide described above, and nor did they intersect with the Western environmental movement – which first gathered momentum in the late 1960s.

By then it was clear to many that population increase, economic growth, and mass consumption were having a profound effect on natural resource availability and the integrity of ecosystems. Mercury poisoning at Minamata Bay, Japan; the Torrey Canyon oil tanker spill; Rachel Carson’s (1962) bestselling account of how herbicides and pesticides got into the food chain – these and other events inspired the first Earth Day, the founding of Greenpeace and Friends of the Earth, and other seminal early 1970s environmental initiatives. Geography had a golden opportunity to make “human impact studies” its main business – a possibility foreseen in 1956 in the earlier-mentioned Man’s Role in Changing the Face of the Earth and foreshadowed a century earlier by geographer George Perkins Marsh (1965/1864).

Quite why this opportunity was missed is hard to say. Although it was grasped at the teaching level, this was not really the case at the research level. Richard Chorley’s (1969) Water, Earth and Man – which called for a new focus on human–environment interactions – was arguably among the exceptions that proved the rule. Geographers conspicuously failed to analyze the anthropogenic local and global “environmental problems” that became ever more apparent from the early 1960s. Morally, the discipline also virtually ignored the pro-nature (or eco-centric) arguments being made within the wider environmental movement. Instead, a relatively small number of geographers complemented the natural hazards work of White and associates with a rather anthropocentric focus on resource management. This resource analysis was usually empirical, quantitative, and conducted in what geographer Tim O’Riordan (1976) called a “technocentric” mode. In other words, this research looked at how best to conserve resources for present and future human
needs. It rarely took issue with the fundamental causes of resource depletion and was very human-centered. Relatedly, a number of physical geographers became interested in the impact of human activities on the parts of the environment that interested them (and vice versa) (e.g., Hollis 1975). Like resource management studies, their research had a policy dimension because environmental management needed to be based on a proper understanding of its objects (e.g., rivers, soil erosion, predator–prey relationships). Yet despite its apparent “ethical-neutrality,” this sort of research was arguably value-laden because it did little to challenge the institutions, human actions, and rule systems that generated environmental degradation in the first place. It was very much “status quo” research.

The 1980s, 1990s, and early 2000s: “re-naturalizing” human geography, denaturalizing an enlarged environmental geography, and finessing the nature of physical geography

By the mid-1970s, then, nature was a key object of analysis in geography but in a different and more circumscribed way than a century earlier. A decade later things began to change. A significant development was that “critical human geographers” – those politically on the left of the subject – began to take a serious interest in not only nature but its collateral terms and referents too. Concurrently, a new generation of environmental geographers (also critically disposed) began to focus hard on contemporary environmental problems, but in ways that challenged the status quo by de-emphasizing the purported contribution of “nature” (e.g., natural scarcity) to those problems. These developments, to be detailed momentarily, were at once unsurprising and novel. The radicalism reflected the formative influence of the worldwide anti-establishment protests that erupted in 1968 and lived on for some years; but the “rediscovery of nature” beyond physical geography was not, in fact, a delayed mirror image of the “green movement” whose heyday was the 1975–1985 period. As we will now see, many human and environmental geographers sought to de-objectify “nature” and were relatively unconcerned about things like “environmental limits to growth” or the “rights of nature.” In other words, and for reasons to be explained, they chose not to make geography the discipline of analyzing and trying to reduce “the human impact” on the nonhuman world (so continuing an existing trend). Their interest in nature took other forms and had other motivations.

Let us look first at developments in environmental geography, of which two are worth highlighting during this period: namely, the rise of (developing-world) “political ecology” and of “vulnerability” approaches to understanding natural hazards. Cultural ecology, for the most part, adopted a resolutely local focus and tended to treat cultural groups and their biophysical milieu in isolation from the national and global scales. But this began to change from the mid-1980s onwards. The reasons were presaged in cultural ecologist Barney Nietschmann’s study of the Miskito Indians of the Nicaraguan coast. A field trip in the early 1970s made Nietschmann aware that his chosen field area was beginning to be drawn into national and global commodity markets and was losing some of its former independence (Nietschmann 1973). A decade later, Piers Blaikie and Harold Brookfield formalized this insight in the germinal books The Political Economy of Soil Erosion (Blaikie 1985) and Land Degradation and Society (Blaikie and Brookfield 1987). For them, cultural ecology’s focus on homeostasis and the relative autonomy of different cultures were increasingly being rendered unrealistic by two things: first, the growing reach
of state power at the national level, and second, the internationalization of commodity production, distribution, and consumption. “Political ecology” was thus tasked with understanding how local resource use was being affected by wider social forces, and the accent was on asymmetries of power between ordinary people and the various actors (e.g., national states and multinational companies) affecting those people’s lives. A new generation of researchers, especially in North America, was inspired to uncover the complex chain of connections tying local land and water use decisions by (say) peasant farmers to global shifts in commodity prices, trade agreements, and so on. Political ecology was thus “political” in that it was critical of the actors and processes that were destabilizing local land-use practices. For instance, land users who degraded their local resources were typically seen in this approach as relatively blameless victims making difficult decisions in highly constrained circumstances explained by “external drivers” beyond their immediate control.

If political ecology radicalized cultural ecology and expanded its analytical horizons, so too did a new approach to natural hazards geography from the mid-1980s onwards. The studies of White and fellow travelers tended, like cultural ecology, to bracket off the local scale from its wider geographical context. They also abstracted people’s perceptions of hazards and the process of cognition from any broader consideration of social relations and social identity. This began to change when Kenneth Hewitt – a former adherent to the White approach – edited *Interpretations of Calamity* (1983). A decade later, Ben Wisner and colleagues published the important book *At Risk: Natural Hazards, People’s Vulnerability and Disasters* (Wisner et al. 2003). These texts had an international impact on Anglophone hazards geography. They suggested that “vulnerability” to hazards was not simply a question of living in a naturally hazardous location; it was also a question of wealth, power, identity, and social location. Drawing, like the political ecologists, on the radical ideas that flowered in the Western social sciences from the early 1970s onwards, Hewitt, Wisner, and others argued that vulnerability was (i) variable in degree and kind within populations, and (ii) explained by the positions people occupied in regional or national social structures and in world politics and economics more broadly. This kind of research directed policymakers’ attention away from people’s perceptions and conceptions of hazards (with associated public education programs), and from technocratic policies like building flood walls to defend against storm surges. Instead, attention was focused on how poverty and powerlessness are socially produced, and on how vulnerable communities can be better protected against the problems triggered when hazardous events occur, like a tsunami. Protective policies should, the argument went, be focused as much on social welfare as on physical engineering solutions. Vulnerability was seen as a *societal* problem whose downsides were starkly exposed when natural hazards (seen as “proximate,” not “root,” causes) occurred.

As these intellectual innovations were occurring, a disparate group of human geographers began to take an interest in nature. Their aim was not to understand its “real character” and how it materially affected people. Instead, they sought to *denaturalize* understandings of nature and its collateral terms. Publications by scholars influenced by “poststructuralism” and by Marxism respectively stand out. The former were part of a wider “cultural turn” in Anglophone human geography during this period, one that focused on the origins and effects of various forms of social “discourse.” The latter were part of Marxism’s consolidation as a major approach in
human geography. The denaturalizing perspective that became de rigueur had three broad aims: first, to show that claims made about nature contained culturally specific assumptions about how the world is and ought to be; second, to show how references to nature were a means of furthering specific social agendas; and third, to show how these references actively concealed the two things just mentioned. Together these aims defined a “nature-skeptical” attitude in the specific sense that, while the concept of nature was taken to be real, so too was the physical world it designated, but the concept was seen to be “unnatural,” as was its role in societal discourse (lay and expert). It is this attitude that was flagged in the latter half of the section “What is nature?” as an alternative to the presumption that “realism” is the only sensible attitude toward those things we, by convention, call “natural phenomena.”

Out of many key contributions to the new “cultural studies of nature” in the 1990s, human geography let us focus on just two authors. In a now classic essay and related book, Canadian geographer Bruce Braun (Willems-Braun 1997; Braun 2002) applied poststructuralist Jacques Derrida’s ideas to the topic of nature and wedded them to postcolonial theory. The latter proposes that colonialism lives on culturally even though it has largely disappeared economically and politically. Braun’s empirical focus was the early 1990s clash between environmentalists and a commercial logging company over whether to fell an area of “old-growth” temperate rainforest in British Columbia called Clayoquot Sound. The former saw Clayoquot as one of the last remaining spaces of “pristine nature,” while the latter regarded it as a valuable economic resource that should be logged in a responsible way for the good of the Canadian economy and those communities dependent on forestry jobs. In a detailed analysis of both sides’ representations of Clayoquot, Braun shows how the region’s “realities” were made to appear quite different, depending who was doing the looking. There was a clash between representations that made Clayoquot appear wild, intricate, threatened, and special on the one side, and those (on the other side) that made it appear as one more “resource zone” to be rationally harvested by high-tech logging firms. In both cases, the authors of the representations claimed to be depicting Clayoquot as it actually was. But Braun’s point was that these representations – which comprised books, pamphlets, press releases, and newsletters – reflected the specific agendas of those promoting them. In other words, one could not adjudicate between them by testing their veracity against the “nonrepresentational actualities” of Clayoquot’s old-growth trees.

Arresting though this insight was, Braun’s research contained a further surprise. Notwithstanding the differences in content and message between the environmentalists’ and forest company’s representations, Braun argued that they ultimately shared the same symbolic universe: a specifically Anglo-North American one that reflects the linguistic conventions and cultural assumptions of those colonists who spread through the United States and Canada from the seventeenth century onwards. He makes this point with reference to Clayoquot Sound’s small groups of remaining “native” or indigenous peoples. These groups had, historically, lived a peripatetic existence and had used forest, lake, river, and shore for generations to meet their material and symbolic needs. Yet both the environmentalists and the logging company fighting over Clayoquot’s future assumed that the region was largely empty. This, Braun argued, constituted a geographical expression of a specifically Western belief that nature and society are two separate things.
indigenous peoples, he concluded, were thus victims of symbolic violence, even in the supposedly postcolonial conditions of modern Canada. Their history and present-day claims to control of Clayoquot simply did not register in the unthinking assumptions made by the descendants of the original European colonizers. Even when seemingly “progressive,” Braun showed, references to nature can serve as politics by other means.

Around the same time as Braun was writing, Australian geographer Kay Anderson was reinterpreting one of nature’s oldest and most socially potent collateral concepts, “race.” In the 1980s, ideas about race as biological differences between groups of Homo sapiens made something of a reappearance in Western discourse, having been discredited in the 1940s and 1950s for reasons we cannot go into here. In part this reappearance was linked to progress in the various sciences of human biology, which led to an international attempt to “map” the human genome and an eventually abortive attempt to also map genomic diversity among humans. There was also, again courtesy of new scientific findings, a resurgent “nature versus nurture” discussion in the West. Within the disciplines of human biology, physical anthropology, neuropsychology, and the young field of sociobiology, a debate has raged over whether people’s mental and physical capacities are mainly a function of genes and the like, or a product of their sociocultural environment. As this debate has unfolded, new biotechnologies have been invented that promise to alter “human nature” so that behavioral “disorders” or congenial diseases can, potentially, be engineered out of existence. In this context, many people worry that we are witnessing a new biological determinism to rival the eugenic beliefs popular in some Western countries in the 1920s and 1930s. The concern is that beliefs about the supposed links between a person’s genes and their behavior or appearance will be used to target those with a supposedly “inferior” or “abnormal” genetic constitution.

Writing in this context, Anderson explored how discourses about the supposed “actualities” of a group’s “race” were used to exclude them from the rest of a society, both socially and spatially. In her 1991 book, Vancouver’s Chinatown, Anderson showed how, from the late nineteenth century, a majority white population of British descent in Vancouver stigmatized Chinese immigrants. She demonstrated how this majority fixated upon phenotypical and cultural differences between themselves and Chinese settlers, coding them in “racial” terms and hierarchically too. This coding then served to hold the settlers at a distance in key arenas like employment and housing, reproducing European privilege. A later book on the origins of ideas about significant differences between human beings, Race and the Crisis of Humanism (Anderson 2007), also focused on the way culturally specific ideas are often conflated with the physical world of which they seek to make sense. Anderson’s monograph explored the challenge Europeans faced in understanding aboriginal societies in Australia from the late seventeenth century. Though holding Enlightenment concepts of humanity as a singular species and of all humans as biologically similar, new arrivals to Australia struggled to fit aboriginal people into their schema. Over time, Anderson shows, it was believed that these people were somehow closer to nature, almost animalistic, and thus not “human” in the same way as Europeans were. Yet, far from being a “factual” corrective to the Enlightenment idea of a “universal human nature,” Anderson demonstrates that this idea was itself culturally infused with certain European perceptions of what it means to be “properly human.” The implication is that even contemporary ideas about human nature and
intrahuman group differences cannot escape the imprint of cultural preferences and dislikes.

Studies like Braun’s and Anderson’s paid careful attention to the origins, content, and effects of various forms of “nature-talk” and of representations of “natural phenomena” more generally. However, they paid little attention to the material referents invoked or to the political economic realm wherein people wrest goods and services from the Earth. The former oversight perhaps reflected a fear of reintroducing environmental (or biophysical) determinism by the back door. It underpinned publications focusing on how nature might be said to be a “social construction” (see, for instance, Demeritt 2002). By contrast, one of human geography’s leading Marxists, the late Neil Smith (1954–2012), focused on both things. His influential book Uneven Development (1984) argued that capitalism – the dominant system by which goods and services are today produced – generates geographically uneven development as a constitutive feature, not as a random occurrence or accident. This is because capitalism is inherently contradictory: its compulsion to increase wealth incessantly leads it to scour the Earth for new investment opportunities, new markets, new raw materials, and new labor forces. According to Smith, the generation of profits and their deployment on a vast scale not only explains the rise and fall of cities, regions, and whole countries; it also underpins what he called “the production of nature.” This, as Smith himself observed, is a counterintuitive idea: “We are used to conceiving of nature as external to society … or else as a grand universal in which human beings are but small and simple cogs. But … our concepts have not caught up with reality. … [C]apitalism … ardently defies the … separation between nature and society, and with pride rather than shame” (1984, xiv).

What did Smith mean by “production”? Capitalism, Karl Marx argued, is an historically specific mode of making and exchanging commodities that began in Europe during the early eighteenth century. What makes it distinctive, he argued, is that commodities are produced not for their use value (i.e., their practical utility) but their exchange value (i.e., how much money they can command upon sale). Since they do not sell in order to get the same monetary sum back as it cost to produce their commodities, capitalists are fixated on “accumulation for accumulation’s sake,” employing workers to do their bidding and facing competitive pressures from rival capitalists. These pressures, Marx argued, necessarily force all capitalists to innovate and to seek a competitive edge by making new products, hiring more or less workers, employing “smart machines,” relocating factories, and so on.

What has this got to do with “nature”? A great deal. According to Smith, capitalism has remade nature over the past two centuries, meaning that such things as genetically modified crops are only the latest in a very long line of material transformations of the nonhuman world. Nature has become a mere means to the end of profit realization, and in the process it has been physically reconstituted into an anthropogenic “second nature.” Though not formally referencing Smith’s book, a brilliant 1991 monograph by historical geographer William Cronon illustrated Smith’s arguments in bracing detail. Nature’s Metropolis tells the epic story of how the rise of Chicago during the nineteenth century was umbilically connected to the formation of entirely new agricultural landscapes throughout the mid-west of the United States. For instance, in a superb chapter on grain production Cronon (1991) shows how money invested in new railways, new storage facilities, and new farm equipment created a vast new geography of fields and fences that replaced the natural grasslands created by evolution and the actions of indigenous/native peoples. Capitalism, in both
metaphorical and literal terms, revolutionized Chicago’s hinterland in a few short decades.

Smith’s arguments may seem somewhat overstated, even in an era where the genetic composition of species is being altered by science and technology. After all, even the “produced” landscapes of modern, commercial farms are “natural” in the sense that the animals, seeds, and crops have not been created by people from scratch. So why did Smith not favor the use of “softer” verbs (like “modification” or “alteration”) when describing the capitalism–nature relationship? The answer is twofold. First, he wanted to challenge those who continued to reference a nonsocial, supposedly pure, external “nature” when advancing their own cognitive, moral, or aesthetic arguments. His PhD thesis supervisor, David Harvey, had already set a precedent back in 1974. Harvey strenuously resisted then-popular arguments claiming that the world was increasingly “overpopulated,” because there were more people alive than the Earth’s natural resources could sustain. He argued that advocates of policies to limit birthrates justified them by talking of “natural limits to growth” that were supposedly fixed. This, Harvey (1974) argued, was an “ideological” move because it neglected the fact that “optimum” population numbers can only ever be defined relative to culturally specific assessments of what a “suitable” standard of living is. Those neo-Malthusians arguing for population control, Harvey argued, were typically developed-world inhabitants using spurious claims about “natural resource scarcity” to imply that developing world poverty was caused by reckless procreation.

Second, Smith arguably favored the metaphor of “production” because he believed that in a capitalist world “nature” and “society” were no longer discrete domains of reality. In other words, they comprise an ontological unity, not two ontological spheres. Terms like “alteration” and “modification” suggest some residuum of “natural nature” untouched by capital. But Smith’s point was that those aspects of what we call “nature” that are germane to our everyday lives are increasingly defined in relation to the needs and actions of capitalists. It is not so much that all of nature is produced “all the way down” physically – the Earth’s molten core, for instance, is hardly “produced” by capital. It is more that so much of what we consider to be “modified” or even “pristine” nature is as it is because it serves the material and discursive demands of capitalist enterprises as they vary in time and space. And where it does not – as in the case of global warming, which is caused unintentionally by atmospheric pollution – this nature is the indirect result of the intentional production of other aspects of nature.

This said, Smith (ironically) said precious little about the material properties and causal efficacy of those things we by convention call “natural,” be they consciously “produced” or not. His was a one-sided materialism that accorded considerable power to capitalism. It was left to others to fill this intellectual vacuum. There were two aspects to this. On the one side, some showed how capitalism was able to physically mold some parts of nature in the interests of profitability. Examples include environmental sociologist Jack Kloppenburg’s First the Seed (1988) and Goodman, Sorj, and Wilkinson’s From Farming to Biotechnology (1987). Both books focused, like Cronon’s, on farming. They showed how science was used by agro-foods companies to break down the physical “barriers” posed by crops and farm animals to enhanced profits – barriers such as crops’ vulnerability to certain pests or a chicken’s inability to grow to adult size in a few weeks rather than months. In both books, the physical malleability of certain components of “nature” was made plain in
a way in which Smith only made theoretical mention.

On the other side, other Marxist researchers interrogated cases where capital (or, rather, certain capitalists) is unable to impose its will entirely and where the agency of “nature” becomes important. Two examples will suffice. Karen Bakker’s (2003) *An Uncooperative Commodity?* showed how the physical properties of water made a huge difference to the way that privatizing British water services unfolded post-1989. Its weight and bulkiness give it a certain insensitivity to which new water market had to adjust. Likewise, Scott Prudham’s book *Knock on Wood* (2005) showed how the Pacific coast forestry industry has, in key respects, had to adapt to the material challenges posed by softwood trees growing in mountainous environments. Bakker and Prudham were not reintroducing a concept of nonsocial “nature,” of which Smith would surely have disapproved. Their point is that what we call “nature” possesses a degree of agency and influence, but this is always defined relationally with respect to the changing needs and wants of capitalist firms, not *sui generis*. It is thus never absolute and is thoroughly contingent and conditional.

The Marxian work on capitalism and nature prioritized economic processes and motivations. As we have seen, the suggestion was that the conventional idea of “nature” has been rendered obsolete (in some or all of its four meanings) because biophysical entities can no longer be understood in abstraction from the influences of capitalist enterprises. As Smith and Harvey also argued, this idea can be used for ideological purposes even as it has outlived its usefulness. This implies a gap between “discourse” and “reality,” wherein the concept of “nature” is used as a smokescreen by those with power or influence. This focus on the concept’s “performativity” is one they shared with the likes of Braun and Anderson.

Compared to all of the above, physical geography altered in quite conservative ways through the 1980s and 1990s. It held firmly on to, while finessing, its commitment to the idea that (nonhuman) nature has objective properties that are amenable to impartial analysis. There were three noteworthy changes of emphasis. First, the balance between pure and applied research arguably tilted slightly in the latter’s favor. According to Gardner (1996), this reflected a growth in the “environment industry” after the first Earth Summit (in 1992) and, particularly, the field of environmental management. Issues such as desertification, water pollution, soil erosion, and deforestation increasingly made it onto physical geographers’ research and teaching agendas (for more on this, see Gregory 2000, ch. 7). These geographers often sought to aid environmental managers by pinpointing the physical changes caused by certain human actions (e.g., Burt, Heathwaite, and Trudgill 1993). Second, at a more philosophical level, physical geography moved away from the “steady-state” and “dynamic-equilibrium” assumptions that had underpinned much of 1970s and 1980s research. Instead, practitioners began to appreciate that the environment is complex, often disorderly, and even chaotic in its operations. As Barbara Kennedy (1979) presciently noted in the late 1970s, physical geographers are confronted with a “naughty world” (see also Kennedy 1994). This change in ontological assumptions was partly inspired by wider shifts in scientific thinking, notably the rise of “transcendental realism,” as well as complexity and chaos theory (see Phillips 1999). Third, the rise of Quaternary studies and a new emphasis on “global environmental change” meant that the study of environmental systems at large spatial and/or
temporal scales underwent a revival. In a sense, physical geography’s historical origins as what Simpson (1963) called a “historical science” were rediscovered, providing a counterbalance to the small-scale, process-form studies that had been so popular from the late 1950s. This meant that its credentials as an idiographic subject focused on novelty and difference were reasserted, not at the expense of a nomothetic (i.e., law or pattern-finding) approach but as a recognition that general laws and processes can have nongeneral (unique) outcomes (as some “critical realists” argue).

The nature of contemporary geography

We have just seen that human geographers had “rediscovered” nature by the turn of the new millennium by way of a seemingly paradoxical “denaturalizing” approach. This approach, we have also seen, animated an expanded domain of environmental geography, leaving physical geography’s branches focused on the nonhuman world and holding fast to a realist ontology. What of the most recent period? How do geographers today understand “nature” and what, therefore, is the current nature of geography?

In human geography the denaturalizing sensibility that existed by 2000 has to some extent given way to a new approach that challenges the idea either that “nature” is a socially constituted representation (or artifact) or that it is something existing regardless of human ideas and practices. This approach rejects the ontological distinction between society and nature that underpins most of the studies cited so far in this entry. It has roots in continental European philosophy and STS (science and technology studies): thinkers like Isabelle Stengers, Michel Callon, Donna Haraway, and Bruno Latour have strongly influenced several erstwhile “human” geographers since the mid-1990s. By the early 2000s two overlapping strands of research were already evident, both prevalent in British human geography: namely, one focused on “hybrid geographies” (closely associated with Sarah Whatmore 2002) and one on “nonrepresentational geographies” (closely associated with Nigel Thrift 2007). Since we do not have the space to explore their subtle differences, we will simply highlight their commonalities.

Hybrid and nonrepresentational perspectives maintain that dividing the world into “social” and “natural” entities simplifies reality to the point of misrepresenting it. The suggestion is that there is no neat Maginot Line that divides the world into two qualitatively different orders that somehow “come into contact” with each other in various ways in different contexts. Instead, Whatmore, Thrift, and like-minded geographers insist that we inhabit one reality, not two, and that supposedly “natural” things and “social” things cannot be separated if we are to understand their character and effects. This chimes with Neil Smith’s arguments, but unlike Smith these geographers do not believe that capitalism is calling all the shots in our syncretic, mixed-up world. They further suggest that human engagements with the world far exceed acts of “representation”: they also involve touch, smell, hearing, and physical interaction. We are multisensual actors, the argument goes, and what we are does not precede our relationship with all manner of other entities.

To flesh out these arguments, there have been numerous recent studies of the almost endless ways in which so-called social and natural things are so entangled as to variously co-constitute, stabilize, or alter one another (depending on the case). These studies are highly attentive to the fine details of different situations and amount to a cognitive, moral, and aesthetic call to attend to what the society–nature dualism has so long
hidden from view. For example, in some of her work on wildlife conservation, Sarah Whatmore has argued that the seemingly unproblematic category “elephant” needs to be called into question. Becoming an “elephant” is, Whatmore and Lorraine Thorne (2000) show, a process that is contingent on the specific network of actors, institutions, and physical environs in which individual pachyderms exist. There is thus, they argue, a notable set of differences between zoo elephants and “wild” elephants, even though they are conventionally regarded as belonging to a single species possessed of stable and singular characteristics common to any similarly named creature. Conservationists, they imply, could usefully pay more attention to the differentiated character of that which they are seeking to conserve, and to the varied networks through which “becoming an elephant” is achieved. To conserve “nature” one needs to understand that it is not a discrete object or site to be protected, but a whole set of differentiated entities bound into complex relationships that might be very extensive in space and time. This obliges us to rethink both what is being “conserved” and how to do it in all nature conservation policies (see also Hinchliffe 2008). More broadly, it means that any future politics must be attentive to the relationships between, and admixtures of, what can no longer be labeled as “social” or “natural” entities. This challenges most major extant political paradigms, be they environmentalist, socialist, liberal, or any other.

Research in this vein has grown in size and diversity since 2000. It is now known, variously, as inquiry that has a “more-than-human,” “postnatural,” or “posthuman” focus. Some of it examines the human body, yet without reintroducing fixed conceptions of “human nature.” It has also been called “neo-materialist” because it shares physical geography’s deep interest in the properties and effects of materials, while challenging the atomist and reductionist assumptions about matters evident in some of its branches. A fine recent example of this expansively relational approach is Lesley Head, Jenny Atchison, and Alison Gates’s (2012) Ingrained: A Human Biography of Wheat.

However, this is not to suggest that earlier strands of denaturalizing research by human geographers are passé. On the contrary, many studies continue to be published of who is speaking about “nature” and its collateral terms, in what ways, with what intentions, and with what effects. Many of these studies have used the analytical lens of “eco-governmentality” (or “environmentality”) to make sense of discourses about everything from global climate change to “green consumption.” This lens derives from French philosopher-historian Michel Foucault’s notion of “governmentality.” Foucault argued that the process of governing people (and non-humans) exceeds the institutions of the modern state and is dispersed among multiple organizations and actors. He further suggested that government not only acts upon people but, as it were, through them by altering their sense of self and world. Governmentality as an analytical concept combines two aspects of governing: namely, “rationalities of government” and “technologies of government” (Miller and Rose 2008, 15). The former specify the division of tasks and actions between authorities (e.g., political, spiritual, military, or familial) and express the ideals or principles that should direct government (e.g., freedom, justice, or entitlements). Such systems of thought also identify the nature of the objects to be governed (e.g., society, the nation, the population, animals, the economy) and characterize the persons over whom government is to be exercised (e.g., a flock to be led, legal subjects with rights, a resource to be exploited) (see Rose and Miller 1992, 178–179).
As Lövbrand, Stripple, and Wiman (2009, 8) express it: “Whereas rationalities of government … render reality into the domain of thought, technologies of government translate thought into the domain of reality.”

How does this connect with issues of “nature”? In recent years three such issues have become ever more prominent in academia, politics, business, and civil society. First, there are widespread concerns about the accelerating pace and heightened magnitude of anthropogenic climate change (“nature as threat”). Second, there are equally widespread worries about natural resource exhaustion and anthropogenic degradation of terrestrial and marine ecosystems (“nature under relentless pressure”). Finally, there is both excitement and concern about genetic modification, artificial life forms, and other science-led attempts to intentionally remake the building blocks of life (“nature denatured”). In this context, what studies of “eco-governmentality” do is trace the ways certain mainstream institutions are enrolling references to things like biodiversity, species extinction, resource conservation, genetic diseases, and carbon dioxide into new rationalities and technologies of government (see Rutherford 2011). For instance, something as mundane and seemingly positive as promoting “ethical shopping” can be seen as a contestable attempt by certain organizations (e.g., supermarkets) to govern the thoughts and actions of millions of people – in the process closing off other possible ways of responding to global environmental change.

This kind of research, like that done earlier by Braun, Anderson, and Smith, pays little attention to how the material world existing beyond rationalities and techniques of government might matter “in itself.” This may seem foolish in light of what many scientists tell us is the “dangerous” magnitude of anthropogenic environmental change coming at us in this century and beyond. Accordingly, some human geographers, and many environmental geographers, have been very focused on the following issues: how the science of environmental change is communicated to and interpreted by ordinary people (e.g., Boykoff 2011); how advanced industrial societies can make a rapid sociotechnical transition to a “low carbon” state (e.g., Bailey and Wilson 2009); how low-income countries can adapt to the biophysical impacts of a warmer, wetter world (e.g., O’Brien and O’Keefe 2013); how low-income communities and vulnerable people can be made more resilient in the face of anthropogenic environmental change (e.g., Wilson 2012); and how local alternatives to an environmentally destructive global capitalism can be fostered (Gibson-Graham and Roelvink 2010). This is an incomplete list, but it speaks to the partial “renaturalization” of both human and environmental geography in light of the end, as some earth scientists see it, of the Holocene (i.e., the 11,500-year period of relative environmental stability coincident with homo sapiens flourishing). The word “partial” is used here because the kind of research just itemized generally does not take a humanly altered “nature” as something whose character and impacts can be defined in absolute terms. In this sense, it carries forward the analytical sensibility first cultivated by the likes of Hewitt, Wisner, Blaikie, and Brookfield many years ago.

However, some environmental geographers now argue for a more profound renaturalization of geographical research and pedagogy. For instance, in his book Inhuman Nature: Sociable Life on a Dynamic Planet, Nigel Clark (2011) claims that the strands of “denaturalizing” summarized earlier in this entry suffer a key weakness. In his view they tend to ignore – or to not take seriously enough – the very largest, inhuman forces operative in and on Earth (like ocean currents and volcanoes). The twentieth century,
he argues, has fooled many Western geographers into believing that “high-magnitude” bio-physical events are low frequency and spatially dispersed. Yet our climate-changed future might assault global humanity with many such events in time and space. In this light, denaturalizing research – be it cultural studies of nature or “postnatural” research – risks being irresponsibly ignorant of the corporeal threats and radical opportunities for a new human future attendant upon the Holocene’s end.

Clark’s belief in the autonomous existence and raw power of “big nature” – albeit one imprinted by centuries of human activity – chimes with recent developments in parts of physical geography (and the wider environmental sciences). Many practitioners have regarded global environmental change as a further impetus to reclaim physical geography’s “big picture” integrative focus (see Clifford 2009). Many physical geographers have seen their research feature in the influential reports of the Intergovernmental Panel on Climate Change (IPCC) and a few are advocating the new idea of “planetary boundaries,” which humanity is said to compromise at its peril. Still others have, at times, lent their name to Earth System Science, a hoped-for super-science of the planetary environment that seeks to understand couplings, feedbacks, and thresholds among the various subsystems of land, air, and water (see Wainwright 2009). Through all this a rather traditional ontological realism prevails: global nature, one changing because of human impacts, is seen as amenable to analysis via remote sensing techniques, sophisticated computer packages, ground-truthing, and logical analysis based on known laws. So it is that the long-standing nature–society dualism is preserved, despite the critical interventions of Whatmore, Thrift, and others. Complementing this, a number of environmental geographers are keen to extend the scientific mindset to both understand and better manage the “human impact” on the world’s land, water, and air masses. Advocates of “land change science” and “sustainability science,” like Billie Lee Turner II (Turner, Geoghegan, and Foster 2004) and Robert Kates (2011) respectively, urge geographers to derive practical solutions to anthropogenic environmental problems by paying close attention to the scientific evidence. These geographers are, perhaps, less enamored of the denaturalizing impulses of many of their professional peers and closer in spirit to most physical geographers. Though in no way neo-environmental determinists, they apparently accept the is–ought dualism that parallels the nature–society, object–subject ones discussed in the section “What is nature?”

There are, however, exceptions. For instance, Mike Hulme – a physical geographer and former lead author of IPCC science reports – argues that the sciences of the Earth environment should be subject to sociocultural analysis. This is especially true now that they may be assuming unprecedented importance in shaping public and political decision-making. In Why We Disagree about Climate Change (Hulme 2010) and related publications he challenges the idea that the environmental sciences deal merely in “facts,” leaving others to trade in debates about values and goals. More than this, though a scientist himself, he worries that the global claims made by the IPCC can too easily underpin ideas of “global government” and “global policy” that are likely to alienate (even harm) many ordinary people whose lives are ineluctably local and situated. A new “climate determinism,” he suggests, may be afoot (Hulme 2011). If so, it comes on the heels of a neo-environmental determinism some detect in the writings of geography’s most famous environmental analyst, Jared Diamond. His popular books Guns, Germs and Steel (1997) and Collapse: How Societies Choose to Succeed or Fail (2005) are part of a genre of “highbrow”...
popular writing that recalls the grand analyses of geography’s founding figures such as Halford Mackinder. But they arguably suffer the same problems in their attempt to establish how the natural environment enables and constrains human flourishing. Gaps of data, questionable inferences, and contestable deductions are glossed over in the search for historic patterns and cross-cultural trends in humanity’s relation to land, water, and sky.

In light of all of the above, it should by now be clear that the various intellectual divisions internal to geography will not be overcome in the name of a holistic analysis of our new environmental era, what some are calling the Anthropocene. Though some geographers hope for a reunification of the discipline (e.g., Herbert and Matthews 2004), it is more realistic to expect greater interchange between certain physical, environmental, and human geographers interested in some aspect(s) of “nature.” This interchange has been interpreted as within-discipline experiments in, variously, “inter-,” “multi-,” “cross-,” and “trans-disciplinarity” (see, for instance, Oughton and Bracken 2009).

Though some of these experiments are intended to combine the expertise of specialists in order to reveal the “total picture,” many are designed to foster mutual learning between diverse perspectives. Hulme (2010), for one, sees the latter as vital for both geography and wider society. For him, current debates about “nature” – what it is, what it does, its value, and its very existence – are not to be resolved. Instead, they are incitements to explore diverse views on the sort of world we believe we inhabit and the sort of future we might want to create. For Hulme, the “problem” of how to understand what we call nature (and its collateral referents) is a productive opportunity, not a hurdle to overcome. It can and should spark democratic debate about the really big questions in our lives.

Conclusion

Despite its length, this entry has not represented the full range of work on nature by geographers past and present. Yet we have still covered an awful lot of ground. As we have seen, the centrality of “nature” to geographers’ inquiries has varied over the past century or more, but has never been marginal to their endeavors. In recent years it has captured the attention of physical, environmental, and human geographers, in contrast to the period 1945–1980, when it largely fell to physical geographers to study and teach about it. This means nature has again assumed the wider disciplinary importance it enjoyed during geography’s early years as a university subject – only now in very different ways than in the era of Mackinder, Ratzel, Semple, de la Blache, Davis, and fellow travelers. Today we can witness the influence of the accumulated history of geographers’ changing perspectives on nature (in any of its four main meanings and its myriad of referents). A traditional realist approach, albeit very different in the detail from a century ago, still persists and looms large in contemporary physical and parts of environmental geography. Meanwhile, two more recent denaturalizing approaches exert influence in parts of environmental and human geography – one focused on representations (images, discourses, signs, etc.) or material transformations (as in agriculture), the other on the dissolution of the supposed ontological divide between society and nature, human and nonhuman. Though we live in an era where many express alarm about the “realities” of humanity’s impact on both nonhuman nature and its own corporeal nature, this has not led to a synthesis of the three approaches or the triumph of one over the others. Instead, intellectual diversity prevails.

Is this a weakness, perhaps a sign of incoherence, or a failure to respond responsibly to...
the “grand challenges” of the early twenty-first century? Some would say so. However, there is another view. By juxtaposing very different perspectives within a single disciplinary matrix, geography today offers opportunities to combine and complicate forms of knowledge about the biophysical world that often exist in splendid isolation elsewhere in academic life. What can realists about “nature” learn from “denaturalizing” critics, and vice versa? What difference might it make to research, public policy, or scholar-activism? These questions challenge geographers and others to appreciate that, like all of the really crucial issues in life, there is no such thing as a “right answer.” Instead, there should always be struggles and disagreements over the way issues are framed and responses thereby delimited. Answers are thereby revisable, debate thereby stilled only for a period of years.

SEE ALSO: Actor-network theory; Animal geographies; Anthropocene and planetary boundaries; Climate change communication; Climate change, concept of; Cultures of nature; Ecological modernization; Environmental degradation; Environmental hazards; Environmental management; Environmental science; Environmentalism; Global environmental change: human dimensions; Hybridity; Natural resources; Nature and corporeality; Political ecology; Posthumanism

References


**Further reading**


Necrogeography

Hamzah Muzaini
Wageningen University, Netherlands

Necrogeography, or the study of “deathscapes,” is the inquiry into spaces associated with the mourning, management, and remembrance of death, dying, and the dead, of humans as well as animals, particularly how these are physically and symbolically constructed, negotiated, and sometimes contested.

Various aspects define its research parameters. The first pertains to the role of “place” in human mortality and end-of-life processes. This strand considers how rituals associated with such processes transform “places,” and the ways in which “places” in turn influence experiences of death and dying. The second relates to the critical mapping of the logic behind the planning and design of “deathscapes,” such as within cemeteries and crematoria. This applies to the processes and issues to do with how dead bodies (and animal remains) are spatially disposed of and managed.

Influenced by trends within the “new cultural geography,” a third strand focuses on the meanings invested in such spaces, and the co-option of these meanings for a variety of purposes (e.g., for identity-building, dark tourism, or bereavement). Underlying this strand is also how “deathscapes” mirror society, not only in its constitution but also its divisions, where contentions may arise over multiple uses of these spaces, and how they can embody exclusions (e.g., in terms of gender, ethnicity, and class).

A fourth strand is the analyses of death and “deathscapes” as these are depicted through art and other media. Of particular interest here is how such representations serve as a lens to reflect upon societal norms and beliefs.

Following more nonrepresentational lines of thinking, recent research strands include the study of the material and affective agency of corpses, emotional geographies of dealing with death, and everyday practices of engaging with the dead. From traditional sites like cemeteries and monuments, research has also moved to “new” spaces for handling remains and remembering the dead, such as through natural burial sites and virtual memoryscapes. In line with the emphasis on the notion of “practice,” there is also increasing interest in liminal spaces associated with the labor of taking care of dead bodies (e.g., within funeral homes). Cutting across many of these strands of research are the following common observations: that “deathscapes” reflect the sociocultural and political contexts in which they are found, that they are susceptible to cultural politics, and that they reveal as much about the living as they do about the dead.

SEE ALSO: Affect; Cultural politics; Emotional geographies; New cultural geography; Place; Population geography

Further reading

NECROGEOGRAPHY


Neighborhood, conceptual

Christian Freksa
Arne Kreutzmann
University of Bremen, Germany

The notion of conceptual neighborhood was introduced in the context of research in qualitative temporal and spatial reasoning to describe direct discrete transitions between temporal or spatial relations in dynamic worlds (Freksa 1991a, 1991b). As a simple example, consider the relative position of two (time or space) points on a directed line. The points P and Q can be positioned in one of three relations: P before Q (P < Q), P at Q (P = Q), or P after Q (P > Q). Under continuous motion of the points direct transitions are possible from P < Q to P = Q and from P = Q to P > Q; a direct transition from P < Q to P > Q, however, is not possible; a transition will go through the intermediate relation =. The relation before (>) therefore is called a “conceptual neighbor” of the relation at (=) and the relation at (=) is called a conceptual neighbor of the relation after (>), whereas the relations before (−<) and after (>−) are not conceptual neighbors, as no direct transition between them is possible (Figure 1). Spatial or temporal entities connected through relations form spatial or temporal configurations.

The role of relations

Qualitative relations are abstract concepts that correspond to discrete classes of continuous-valued relations. If one variable is fixed, qualitative relations either correspond to a single, precise value (e.g., P = Q), or they subsume a whole range of values that are not further distinguished (e.g., P < Q).

Qualitative relations play an important role in conceptualizing spatial, temporal, and other configurations. They are robust with respect to small distortions that do not affect critical aspects of the configuration. Depending on the relations chosen to represent configurations, critical aspects may refer to topological properties (e.g., disjoint, touching, overlapping, inside), to orientation properties (e.g., left, right, front, back, up, down), to angles (acute, right, obtuse, straight), to distance (near, far), or others.
Informal definition of conceptual neighborhood

The notion of conceptual neighborhood was first introduced for temporal relations (Freksa 1991a).

Two relations between pairs of events are (conceptual) neighbors if they can be directly transformed into one another by continuous deformation (i.e., shortening or lengthening) of the events.

This notion can be interpreted for spatial relations as follows.

Two relations between pairs of spatial entities are conceptual neighbors if they can be directly transformed into one another by continuous transformation.

This definition of conceptual neighbors addresses binary relations; it can be extended to ternary relations or to relations of higher arity in a straightforward manner.

The definition of conceptual neighborhood intentionally leaves room for interpretation. First, it does not specify exactly which continuous transformations are to be considered; second, it does not specify whether relations can be conceptual neighbors of themselves; and third, it does not specify what is considered to be a continuous transformation. In his book *Qualitative Spatial Change*, Galton (2000, Chapter 7) examines the implications of different types of spatial change, given various notions of continuity; this results in different answers as to whether or not a two-dimensional region can be continuously transformed into a line segment.

In early days, the terms “closest-topological-relationship-graph” (Egenhofer and Al-Taha 1992) and “continuity network” (Cohn and Hazarika 2001) were used synonymously with conceptual neighborhood graph.

Formal definition of conceptual neighborhood

In this section, the term “domain” refers to a mathematical structure. In geographic information systems the typical domain used is the domain of two-dimensional polygons extended to allow the inclusion of points and sometimes also of line segments. However, a circle or circle segment can only be approximated in such a domain.

In this section, the term “domain” refers to a mathematical structure. In geographic information systems the typical domain used is the domain of two-dimensional polygons extended to allow the inclusion of points and sometimes also of line segments. However, a circle or circle segment can only be approximated in such a domain.

Given a finite set of $n$-ary relations $R$ over a domain $D$, then two distinct relations $r_1$ and $r_2$ are called conceptual neighbors if, and only if, there exists a continuous function $f : C(D^n)$

\[
\begin{align*}
  f &: R \rightarrow D^n, \\
  f(0) &\in r_1, \\
  f(1) &\in r_2, \text{ and} \\
  f(t) &\in r_1 \cup r_2 \text{ for all } t \in [0, 1].
\end{align*}
\]

In the literature, concepts usually are not considered to be conceptual neighbors of themselves; therefore the above definition contains the requirement, that $r_1$ and $r_2$ are not identical. Further, conceptual neighborhood is often
regarded as symmetric, which also directly follows from the above definition with $f'(t) = f(1 - t)$. However, applications usually only allow a subset of all possible continuous transformations due to the semantics in the domain, geometric or physical laws, dynamic properties, and so on.

### Semantics

As conceptual neighborhoods are derived from transitions in the represented domain, they depend on structural and dynamic properties of the domain and of the entities in the domain. To illustrate this, this entry will look at a slightly more complex example: rather than considering (zero-dimensional) points on (one-dimensional) lines it will now look at extended one-dimensional spatial objects: line segments, which can move with respect to a directed line. Equivalently, time intervals (rather than one-dimensional spatial objects), which can be related to one another, can be considered. Thirteen qualitative relations can now be distinguished (Figure 2).

For each configuration of two intervals on the directed line there is exactly one qualitative relation; that is, the relations are jointly exhaustive and pairwise disjoint (the so-called “JEPD property”).

In the one-dimensional world of linear extended objects/events we can conceive of different kinds of dynamics depending on physical properties of the entities involved and on the forces that act on these entities. For example, objects may be assumed to be fixed in location, or mobile, they may be rigid (i.e., nondeformable), elastic (deformable), divisible, or mergeable, and these operations may be symmetric or asymmetric; some of these properties may be combined in various ways. For example,

<table>
<thead>
<tr>
<th>Relation</th>
<th>Symbol</th>
<th>Pictorial example</th>
</tr>
</thead>
<tbody>
<tr>
<td>before—after</td>
<td>&lt;</td>
<td><img src="image" alt="before-after" /></td>
</tr>
<tr>
<td>equal</td>
<td>=</td>
<td><img src="image" alt="equal" /></td>
</tr>
<tr>
<td>meets—met by</td>
<td>m</td>
<td><img src="image" alt="meets-met-by" /></td>
</tr>
<tr>
<td>overlaps—overlapped by</td>
<td>o</td>
<td><img src="image" alt="overlaps-overlapped-by" /></td>
</tr>
<tr>
<td>during—contains</td>
<td>d</td>
<td><img src="image" alt="during-contains" /></td>
</tr>
<tr>
<td>starts—started by</td>
<td>s</td>
<td><img src="image" alt="starts-started-by" /></td>
</tr>
<tr>
<td>finishes—finished by</td>
<td>f</td>
<td><img src="image" alt="finishes-finished-by" /></td>
</tr>
</tbody>
</table>

Figure 2 Thirteen qualitative relations between two linear extended objects on a directed line (or equivalently: two temporal intervals on the directed time line). Adapted from Freksa 1991a. Reproduced with permission of Elsevier.

(i) individual objects may expand or shrink at one end at a time; (ii) objects may be rigid (fixed size), but able to move relative to one another; or (iii) objects may be fixed in location but expand or shrink on both ends.

From a physical modeling, a cognitive, and a computational perspective, it may be desirable to consider exclusively those relations that can be physically realized; however, obtaining conceptual neighborhood in a formal system in the case of restrictions on the domain poses a difficult challenge and generally results in more complex structures. This is due to the fact that abstract formal systems are not structurally constrained in the same ways as physical systems are. Therefore, formal approaches often use a more general notion of conceptual neighborhood for practical applications that does not take into account the restrictions of specific domains; these restrictions then are taken into account at a later stage of modeling.

It frequently happens that, in domains where different perspectives such as empirical, theoretical, or practical may become relevant, these perspectives impose different desiderata on the concepts employed, as each perspective may employ a
different reference system. This is why some flexibility may be needed in order to relate these perspectives in a productive fashion. As applications generally require more information than those encoded in typical conceptual neighborhoods, various extensions have been proposed.

Adding further transition knowledge

Besides restricting the set of continuous transformations on the basis of structural properties of the domain, taking into account restrictions due to specific tasks or actions to be performed as well as implications from earlier processing can be of great value in applications. For example, in a control setting it is important to know which actions could result in a specific change on the qualitative level.

Action-augmented conceptual neighborhoods

Conceptual neighborhood allows the description of mandatory and possible changes but does not specify the cause or actions leading to such change. Assuming that all changes are caused by actions performed by the involved agents, the set of continuous functions can be divided into subsets such that each subset corresponds to a tuple describing the corresponding actions. For each tuple of actions the induced conceptual neighborhood can be determined. Dylla (2008) calls the resulting extension action-augmented conceptual neighborhood.

The asymmetrical nature of actions carries over to the action-augmented conceptual neighborhood. Further, due to abstraction, actions might not result in a change on the conceptual level (see the section “Perceptual uncertainty and coarse reasoning”) or could lead to different neighboring relations. Consequently, the resulting action-augmented conceptual neighborhood structure is complex and requires additional reasoning methods to be used for active agent control.

Dominance spaces and topological mode spaces

From a more theoretical point of view it is interesting to know what occurs at the transition from one relation to another and whether a relation can hold over an interval or only at a single instant of time. An example is the motion of a pendulum that reaches its extreme deflection only at a single instant where its speed is zero. This kind of knowledge is not only important when estimating what might be observable and what might not, but also when simultaneously employing conceptual neighborhoods from different relations.

A relation $u$ is said to dominate a relation $v$ if the relation $u$ has to hold at the transition point for relation $v$. Consequently, the relation describing a pendulum’s extreme deflection dominates the relation that describes a rising pendulum. A conceptual neighborhood structure extended by the notion of dominance is called dominance space. Topological mode spaces combine both dominance and possible temporal extent of relations. Galton (2000) proved a product theorem for topological mode spaces, which allows for directly computing the product of topological mode spaces. This is an important result for applications, as only basic topological mode spaces need to be determined, for example for single pairs of entities, to be able to calculate the topological mode space describing the simultaneous motions of arbitrarily large numbers of entities.
Conceptual neighborhood graphs

Depending on the ontology of the domain and the specific laws that rule the environment, different conceptual neighborhood structures will be obtained. Nonetheless, all of these structures form a graph, either directed or undirected. As an example, imagine two people looking into an aquarium seeing and describing the changing relations of two of the fishes. Essentially, they established a conceptual neighborhood structure, which is visualized as a graph in Figure 3. The observed relative size might change as the fish move in three-dimensional space while the viewer only perceives a projection. See also Figure 4.

Relation between conceptual neighborhood and spatial neighborhood

Tobler’s first law of geography states that “everything is related to everything else, but near things are more related than distant things.” This insight suggests that near things may deserve a different treatment than distant things. Qualitatively, we may distinguish near things from distant things through the attributes “directly connected” versus “not directly connected.” The special role of nearness and neighborhood is a particular property of space; it carries over to the domain of time, to which space is closely connected through the laws of physics.

Similar insights exist in pattern recognition and scene analysis: as a rule of thumb, it is a good heuristic to assume that a pixel (say on a satellite image) maps the same land-use category as the neighboring pixel; this rule is violated only at the (one-dimensional) boundaries of (two-dimensional) land-use regions. Dealing with these – comparatively few – exceptions is considerably cheaper than treating each pixel without prejudice.

Accordingly for conceptual neighborhoods: just as land use or other attributes are not randomly distributed over spatial regions, qualitative relations are not randomly distributed over conceptual spaces (Gärdenfors 2000). The rules for conceptual neighborhoods are even stronger than for attributes of spatial neighborhood: while attributes of spatial neighbors are merely a heuristic rule of thumb, the rules for conceptual neighborhoods correspond to intrinsic properties of spatial (or temporal) structures and processes. In these structures, there is no way that the neighborhood structure could be violated. This fact is exploited in conceptual neighborhood-based reasoning.

Conceptual neighborhood-based reasoning

Spatial reasoning in general can be reduced to the following inference pattern for three locations A, B, and C and two given relations r1 and r2: if A is in relation r1 to B and B is in relation r2 to C then A is in relation R to location C (Freksa 1991b). In this way we can compose each element of a set of relations with itself and with each other element of the set to obtain a complete set of composition rules that govern the respective domain; this has been done, for example, by Allen (1983) in his famous paper on Maintaining Knowledge about Temporal Intervals. By composing each of the 13 JEPD relations that may hold between two temporal intervals (Figure 2) with each element of the same set of relations, we obtain $13 \times 13 = 169$ relations that describe the knowledge we can infer by applying two of these relations in sequence. For example, when an event A takes place before an event B and the
event B takes place before an event C, we can infer that event A takes place before event C.

Not all of the 169 compositions are as straightforward as in this example: for certain compositions, the inferred relation differs from both constituent relations; for other compositions of constituent relations the inferred relation is ambiguous. For example, if event A overlaps event B and event B takes place during event C, then event A either overlaps or takes place during or starts event C; but other relations are not possible.

In the case of the 13 JEPD relations between temporal intervals, 72 compositions yield such ambiguous results. Three of these compositions do not permit any useful inference; for example, if event A takes place before event B and event B takes place after event C, any of the 13 base relations between event A and event C may hold.

Allen observed that compositions that involve the equal relation yield trivial inferences that are identical to the second constituent relation involved; he therefore reduced the $13 \times 13$ composition table to a $12 \times 12$ matrix (which he called the “transitivity table”).

But the structures of time and space exhibit much more regularity that can be exploited for reasoning. Specifically, if we arrange the composition table in such a way that neighboring entries correspond to conceptually neighboring constituent relations, we will observe that the inferences in the table behave in a rather regular and smooth way; this structure can be used to guide the inferences. Most notably, we will observe that only a small fraction of theoretically conceivable combinations of JEPD relations will actually occur as inferences in spatial and temporal reasoning. Upon further analysis, we will observe that the inference sets of relations (i.e., the sets of possible resulting JEPD relations that cannot be disambiguated through the inferences) consist of conceptually neighboring relations, that is, they correspond to situations that are obtained through continuous transformations in the domain.

This insight is not surprising if we analyze the reasons for the ambiguities in inferences: in qualitative temporal and spatial reasoning over temporal intervals or spatially extended objects the ambiguities in inferences are due to the fact...
that the premises yield coarse information with regards to the potential configurations of the entities as the boundaries of the objects are only partially fixed. As a consequence, only neighboring relations corresponding to the ambiguous sides of these boundaries are in question.

This consideration leads to a feature of qualitative spatial reasoning that is highly relevant for cognitive processing or, more generally, for processing incomplete information in benign environments. Here “benign environment” means an environment in which changes happen gradually rather than abruptly. In qualitative (categorical) views of the world, gradual changes correspond to smallest steps (transitions between conceptual neighbors) whereas abrupt changes would correspond to jumps between non-neighboring relations.

Distances between relations

The distinction between conceptually neighboring and other relations gives rise to the question of whether further distinctions between relations may be useful. In particular, a neighbor of a neighbor may be considered closer than a non-neighbor. This becomes computationally relevant if we exploit conceptual neighborhood for reasoning. Thus, we can define a conceptual distance between relations. Based on the conceptual neighborhood graph, a minimal distance between relations can be defined as the minimum number of neighborhood transitions between two relations (e.g., Goyal and Egenhofer 2001) for distances between cardinal direction relations.

Egenhofer and Al-Taha (1992) applied the notion of conceptual neighborhood to reasoning about gradual changes of topological relationships; for example, rising water levels of a lake that will cause a lake that originally is disjoint from the house first to meet and then to overlap the region of the house. The authors looked at conceptual neighborhoods within the framework of the 9-intersection model of topological relationships. The 9-intersection model describes topological relations between spatially extended objects in terms of the configurations of relations between their respective interiors, boundaries, and exteriors; this results in 9 subrelations, each of which can take the value “empty” or “nonempty.” They observed that certain topological relations are distinguished by only a single subrelation whereas others – including some conceptual neighbors – are distinguished by a larger number of subrelations. Consequently, they defined distance between topological relations by the number of distinguished subrelations in the 9-intersection model. Neighboring relations with the shortest distance were designated as the closest topological relationship.
NEIGHBORHOOD, CONCEPTUAL

Egenhofer and Mark (1995) used conceptual neighborhoods to describe topological relations between lines and regions; they confirmed that conceptual neighborhoods correspond largely to the way humans conceptualize similarity of spatial relations. Egenhofer also analyzes spatial relations on spheres in terms of conceptual neighborhoods.

Coarse relations

Conceptually neighboring relations, which are distinguished only through gradual transitions between semantically very close entities, lead directly to the concept of a coarse relation. A coarse relation can be defined by a set of fine (base) relations in which small distinctions (= transitions between neighboring relations) are ignored. The usefulness of coarse relations is twofold: (i) coarse relations can be used to represent incomplete knowledge (here, knowledge in which distinctions between certain details are missing); (ii) perceptual and conceptual knowledge frequently are structured in such a way that fine distinctions may be missing while the big picture is preserved. This seems to be particularly the case for knowledge relating to “benign” domains such as time and space, which are conceptualized as continuous domains in which changes happen gradually to a large extent. Such a type of “incomplete” knowledge frequently can be represented through coarse qualitative relations on the basis of conceptual neighborhoods.

Reasoning based on coarse relations

The next obvious step of what can be done with conceptual neighborhoods is to reason on the basis of coarse relations – particularly if they correspond in a natural way to the type of knowledge that is available through perception of space and time and through the conceptualizations we may have of them.

A classical approach to dealing with incomplete knowledge in artificial intelligence is to form disjunctions of hypothetical completions of this knowledge; in other words, the potential alternatives that may hold under the incomplete specification would be enumerated. Reasoning over disjunctions is a computationally expensive process, as each disjunction may result in new sets of disjunctions; this leads to an exponential growth of complexity. Furthermore, replacing incomplete knowledge by disjunctions of complete knowledge is intuitively not very plausible for the most part. It is a form of reasoning a detective may use to solve a puzzle with missing pieces; but for everyday tasks of dealing with coarse knowledge, what is looked for are computationally cheap approaches that can deal with the coarse level of knowledge directly and yield results on a comparable level of precision (or resolution) as the premises.

This is an area of application for which coarse relations based on conceptual neighborhoods work very nicely: reasoning can be carried out directly on the basis of the coarse relations rather than on the more precisely specified JEPD base relations. In principle, it may be possible to maintain the JEPD property on a coarser level; in practice, this may be neither natural nor necessary. In fact, by choosing coarse relations in such a way that their applicability overlaps for certain temporal or spatial configurations, advantages may be generated that have benefits similar to those of coarse coding techniques in biological and technical systems: by forming conjunctions between overlapping coarse relations the precision of inferences can actually be increased.
and the grain sizes reduced, provided that appropriate coarse knowledge is available. Freksa (1992) showed for the case of Allen’s interval calculus that a suitable choice of coarse relations (semi-interval representation) based on conceptual neighborhoods allows for calculations on the same level of precision as the high-resolution interval approach used in the Allen calculus. The conceptual neighborhood-based composition tables are much more compact than composition tables using base relations as they exploit inherent structures of time and space to a much larger extent than the finer relations.

Conceptual neighborhood-based spatial problem-solving preserves some of the special properties of time and space that make problem-solving in space as efficient as it is; this efficiency disappears when spatial structures are disintegrated by describing them on the level of atomic relations, as is preferably done in formal reasoning.

Perceptual uncertainty and coarse reasoning

The spatial resolution of visual perception in the eye or in a camera is limited by the granularity of the receptive fields in the eye or in the camera, or, as Zadeh (1973) succinctly puts it, “The closer one looks at a real-world problem, the fuzzier becomes its solution” (Zadeh’s Principle of incompatibility). As a consequence, certain conceptual distinctions between spatially similar situations may be difficult or impossible to perceive. For example, from some distance, we may not be able to visually distinguish whether two objects are almost touching one another but being spatially separated or whether they are actually touching, thus being spatially connected; similarly, we may not be able to distinguish perceptually whether two objects touch or barely overlap. All three conceptually distinct situations form a conceptual neighborhood of closely related spatial topological relations.

Interestingly, from a larger distance borders are perceived more clearly and look more orderly than from close by. For example, when we look down to Earth from an airplane or when we view satellite images of the Earth, certain spatial structures manifest themselves much more clearly even though the spatial resolution has been reduced. As transition zones between large regions become narrower they can be categorized more easily as lines at which the neighboring regions meet. This is a similar effect to that observed when looking at sketch maps or schematic maps that abstract from detail that is not informative for certain purposes.

Coarse spatial relations are more stable than fine spatial relations, as it typically requires many fine changes to change coarse spatial relations. When coarse spatial relations change, fine relations inevitably must have changed.

These observations are quite significant for information processing: anything that can validly be inferred on a coarse level of representation should preferably be inferred on a coarse level, as this (i) requires less information, and (ii) will be more stable.

Role of conceptual neighborhoods for perception and cognition

A crucial human ability to deal with perceptual and with conceptual information is to move between different object sizes (size constancy), different levels of granularity (resolution), and different levels of conceptualization of objects in a scene. For visual perception, this ability is
NEIGHBORHOOD, CONCEPTUAL

facilitated through physical/geometrical properties of space, motion in space, and optical projection: when visually perceiving cognitive agents move towards a scene, objects are projected at a larger size, more details can be differentiated, and entities that were perceived as single objects from a distance will appear as aggregates of smaller constituent objects from nearby. In other words, in our everyday spatial experience, “vertical neighborhoods” (Freksa and Barkowsky 1996) in a hierarchical or heterarchical representation that characterize relationships between visual objects and aggregates of objects that form new visual objects, are quite common and familiar to us from everyday experience.

The transitions between objects and aggregate objects have direct implications on spatial neighborhoods and on conceptual neighborhoods: obviously, only spatially neighboring objects will be optically separated/aggregated by moving towards/away from a scene; similarly, conceptually neighboring spatial relations between objects in a scene will be the first to be separated/aggregated as perception transforms between aggregates and their constituent objects. In other words, conceptual neighborhoods have a deep cognitive reality for our understanding of how parts and wholes relate to one another.

The relations between spatial entities and larger entities formed by them have significant implications on geography, map making, and map interpretation. An example is map generalization where individual houses are aggregated to make up residential areas, residential areas are aggregated to make up towns, and so on. On the level of paths and roads, minor routes need to be eliminated at coarser resolutions without violating certain connectivity constraints. Maintaining correct topological relationships is an issue that can be dealt with by means of conceptual neighborhood.

Similarity: conceptual neighborhood versus fuzziness

Conceptual neighborhoods have a distinct relation to fuzzy sets: conceptual neighborhoods can be considered discrete analogs to fuzzy sets. Whereas compatibilities of fuzzy relations typically change gradually as a relevant feature changes, the selection of a spatial or temporal relation changes in a transition to a conceptually neighboring relation in a discrete step. This corresponds to the way people typically talk about spatial and temporal relations: we select a label for a relation that appears to be appropriate, being well aware that a neighboring label also might fit. This awareness helps us to interpret semantically neighboring relation labels that may have been chosen by others due to subjective preference and/or due to not quite identical reference contexts.

Applications of conceptual neighborhood-based reasoning in artificial intelligence

There are a variety of applications of conceptual neighborhood; this entry will present four rather distinct examples.

Commonsense reasoning

When describing events, such as a car overtaking another car, humans abstract to a few key configurations assuming that it is clear how to fill the gaps. Commonsense reasoning focuses on methods to describe how “trivial” inferences
can be achieved, instead of saving reproducible inferences explicitly. Addition and subtraction are examples for such a kind of “inference,” as one could create a lookup table or describe the process of adding two numbers. Conceptual neighborhood-based reasoning fills a similar role in event recognition and planning (see the section “Conceptual neighborhood-based reasoning”), as often a key configuration cannot be observed directly but its occurrence can be inferred, for example, that a pendulum had reached its highest position.

Learning event models

Learning of event models is used in the case of complex domains or where no previous knowledge about the domain is present. In such cases, events and models are learned either unsupervised or semisupervised, that is, a teacher or ground truth is generally available. Often models can be simplified by exploiting the conceptual neighborhood. As discussed in the section “Perceptual uncertainty and coarse reasoning,” some relations are hard to distinguish using sensors, for example, if two objects are truly touching each other. Making the learning system aware of such uncertainty allows for more robust as well as more compact models.

Data integration and conflict resolution

When integrating data from various sources, such as database or sensors, usually conflicting information exists. Further, the resulting data should not violate consistency constraints, for example, two lakes cannot overlap. Using conceptual neighborhood combined with a valid distance measure allows to first determine the required changes on the conceptual level, called relaxation, before calculating the changes on the lower level.

Finding tractable subsets of calculi

Conceptual neighborhoods generally play an important role when determining tractable subsets of calculi. A generalized relation in a tractable subset only contains those relations that are pairwise conceptual neighbors. This has been shown by Nebel and Bürckert (1995) for the case of Allen’s interval algebra. As a result, conceptual neighborhood can be exploited to reduce computation from nontractable exponential to tractable polynomial complexity classes.

Concluding remarks

Conceptual neighborhood is a fundamental notion in cognitive systems. It reflects how cognitive agents including humans, animals, and autonomous robots perceive, describe, and perform changes in spatiotemporal environments. The notion relates to psychology of perception, cognitive geography, and artificial intelligence, as well as to cognitive modeling and process control. It has implications for visualization, autonomous robotics, and theoretical computer science. It is related to robustness, to switching between coarse and fine representations, and to efficiency. Consideration of conceptual neighborhood may play an important role in the design of future reliable technology.

SEE ALSO: Cognition and spatial behavior; Distance; Fuzzy classification and reasoning; Geographic information system; Qualitative spatial and temporal representation and reasoning; Tobler's first law of geography; Topological relations
References


Further reading


The concept of a neighborhood refers to the localized network of everyday social relations and senses of identity that form among residents and businesses that constitute that network. As an object of analysis, however, there has never been consensus among scholars on a specific definition of “neighborhood.” This uncertainty is reflected by the less than precise definition given by the Merriam-Webster Dictionary: “a section of a town or city” and “the people who live near each other.”

Neighborhoods are widely seen to come into existence through the formation of community organizations. These organizations consist of community members to represent interests of residents (based on housing, education, culture, identity, etc.) perceived as shared among most residents and businesses located within identified spatial boundaries. Here, more than just local networks of social relations, neighborhoods often become “politicalized” sites to the extent that members actively come together to mobilize resources in the name of preserving, enhancing, or changing particular neighborhood characteristics. In unison with other local institutions – churches, schools, community centers – these are the venues of social activity that work to organize everyday life and consolidate a sense of areal cohesion, control, and identity.

The boundaries and identification of neighborhoods are also frequently contested and debated. For example, the “community areas” officially recognized by the City of Chicago do not always overlap with the neighborhoods identified by the people who reside and/or work within these city-designated community areas. Chicago’s Bronzeville district signifies to local residents a loosely defined area of historical and cultural identity that actually encompasses four community areas (although some residents argue it is not this large), but is not included as one itself. The city, however, has designated a section of this loosely defined area as the “Black Metropolis Historic District,” an act that does recognize Bronzeville’s existence as a present and historic site in the city.

In this context, neighborhoods are not objective realities or passive creations, but social constructions which are often contested and negotiated by residents and vested interests. Rather than static entities, neighborhoods and their boundaries are porous, fluid, and evolving social formations, rendering any precise or broadly agreed upon definition elusive. There is also immense variation in the social and physical form of neighborhoods, both through time and across different cultural settings. It is in this sense that the neighborhood, as a concept and sociospatial phenomenon, has long been treated as an important object of analysis among urban geographers, planners, sociologists, and historians. The remainder of this entry briefly chronicles the evolution of neighborhoods through the modern period with emphasis placed on this entity as a spatial unit in urban geography and urban planning and its conceptualization in academia and beyond.

It has long been considered that humans have historically organized everyday social interaction and local identity within what can be
called geographically bounded areas, identified as neighborhoods or otherwise. In short, once human living patterns moved from nomadic to more stationary settlements (underpinned by increasing agricultural surpluses), the formation of distinct local-scale “communities” has likely existed in some shape and form (Mumford 1954). Through the premodern period, neighborhoods would have formed in compact, high-density spatial units featuring mixes of street-side businesses and apartment-style residences. The local market place is often considered as the adhesion that brought people together, and through which spatially bounded networks of social, political, and economic relations crystalized. While the physical appearance and layout has certainly varied, these foundational elements are considered to have been more or less standard.

These “traditional” neighborhoods were characterized by tight-knit social bonds, mixed populations in terms of class and ethnicity, and a vibrant street life. This spatial structure, however, was irrevocably ruptured by the industrial revolution, a process that has profoundly transformed pre-existing human settlements. In short, the preindustrial city erupted into the large-scale industrial metropolitan regions that have come to mark the modern period. This emergent scale of urbanization disrupted the material foundation that underpinned premodern or “traditional” modes of organizing everyday life (although this still persists in many rural villages across the world). The modern urban landscape, in contrast, is marked by a complex mosaic of districts organized by industrial, commercial, and residential land uses.

It is in this context that the concept, “neighborhood,” first emerged, to identify, almost exclusively, residential areas marked by increasingly homogenous population groups, such as “working-class,” lower-class, upper-class, and ethnic neighborhoods. In the modern industrial city of eighteenth- and nineteenth-century Europe, increased population densities, congestion, and innovations in transportation allowed the wealthy to relocate away from the emergent industrial districts and the community of workers that congregated near them. As a result, neighborhoods became increasingly segregated by economic class, employment type, and establishments that cater to particular population groups. While not all traditional neighborhood characteristics have vanished, they have increasingly been restructured to reflect this modern form in most metropolitan regions across the Global North and South.

This internal ordering of land uses and neighborhoods in the modern city, however, while rational to industrial capitalists, resulted in problems. Most notably, with few public services and little infrastructure (i.e., water, sewage) the living conditions in working-class neighborhoods declined horrifically as famously chronicled by Frederick Engels in The Condition of the Working Class in England, originally published in 1844. The rise of “slums” led to serious problems – starvation, crime, disease – and provided the basis for modern urban planning and consideration of the neighborhood as an important spatial unit of urban life.

**The neighborhood unit in urban geography and planning**

Urban geography and urban planning have been around for a long time. The spatial organization of medieval cities and villages around the town square, cathedral, and marketplace serves as an example of a precursor, with similar manifestations in Islamic, East Asian, and Ancient Greek, Roman, and Egyptian contexts as well. But the concept, “neighborhood,” was not explicitly identified or substantively incorporated into
planning theory and practice until the early twentieth century. Its origins can be traced to Ebenezer Howard's vision of the “garden city,” proposed as an ambitious vehicle aimed at resolving the increasingly prevalent problems associated with “slum formation” and inequality in the industrial city, particularly London (Hall 2002).

Howard envisioned the reorganization of city structure based on radiating concentric rings of avenues, housing, and businesses, which revolved around a civic center. There was no place for large-scale industrial factories, but rather small-scale, individual enterprises that would foster close interpersonal relationships among residents and constitute a kind of “collective” social identity. This was the “garden city,” limited to around 30,000 people, on 1000 acres of land, and surrounded by green belts of rural land uses. This would serve as the principle spatial unit of Howard’s broader vision of a polycentric metropolitan fabric constituted by many garden cities, functioning as individual collectives, and linked through rail transit.

This vision, with the goal of engineering more equitable social processes through physical planning, never came to fruition. But many of these ideas have influenced planned communities across Europe (England, France, and Germany) and, later, the United States. In the process, however, the driving force behind Howard’s vision – social justice – was lost, with planned communities (e.g., Hampstead in England and Riverside, IL, in the US) resembling more garden “suburbs” or “villages,” and safe havens from the worsening congestion, pollution, and poverty in the city.

The neighborhood unit emerged in this context, identified by the sociologist and planner, Clarence Perry. Inspired by the ideas of Jane Addams and his own experience living in Forest Hills Gardens, a garden suburb of New York, Perry advocated for the importance of design in fostering a strong neighborhood identity in planned communities. For Perry (1929), the neighborhood was what made these planned communities work in a socially volatile, rapidly changing, and increasingly auto-centric modern city. Perry emphasized community planning that safeguarded the neighborhood as a “political and moral unit” marked by strong local leadership and “intimate face-to-face association and cooperation” (Cooley, cited in Hall 2002).

In 1928, led by architects Clarence Stein and Henry Wright, Perry’s ideas were manifested in the design and development of Radburn, New Jersey, the first and perhaps best known garden city inspired planned community in the United States. In terms of building Perry’s neighborhood unit into the community, the key was the distinction between higher-volume arterial streets and intimate networks of local streets. The wider arterial streets were to provide the physical boundaries of the neighborhood by discouraging through (or unwanted) traffic. This hierarchical road system has since become the definitive component of the “Radburn layout,” with widespread implementation across urban America. As a result, neighborhoods in the United States have generally formed within the boundaries of arterial, commercially lined streets.

In this context, the planning for neighborhoods had exclusionary underpinnings. And after a few years, most Radburn residents were white and middle-upper class, a far cry from the socioeconomically mixed “garden city” envisioned by Howard. Moreover, the mixed land use component of Howard’s vision, inspired by the medieval town/village, was further complicated by the parallel advent of land-use zoning. While less an issue in Europe, where the pre-existing urban fabric has proved more amenable to preserving and crafting higher-density and mixed-use neighborhoods, the now standardized segregation of residential, commercial, and
NEIGHBORHOOD

industrial land uses via zoning has given way, in the United States, to the neighborhood understood primarily as a residential unit.

The commercial lining of arterial roads has provided walking access to businesses by interior neighborhood residents; but in the rapidly proliferating suburbs, the greater spatial separation between residential and commercial zones has prevented the cultivation of the kind of vibrant and necessary “street life” considered by Perry, Mumford, and others as definitive of healthy and cohesive neighborhoods. To these critics, the automobile and an emergent menace – modernist planning – was the culprit, as the mass suburbanization of American cities was producing a mode of residential life void of face-to-face interaction and, consequently, neighborhood spirit and identity. Here, zoning should cultivate, work through, and enhance, rather than obliterate, the neighborhood unit as the basis for social life.

By the 1950s, Howard’s “garden city” was a thing of the past, and the neighborhood unit lost its luster given the increased dominance of modernist planning principles. This loss was reflected not only in the “sprawling” American suburbs, but in the central city as well. The federal urban renewal program was demolishing large expanses of “blighted” urban spaces to pave the way for two things: limited-access expressways and public high-rise housing. These urban features not only marked the restructuring of American cities, they became ubiquitous features of modernist planning and architectural design around the world. The neighborhood was no longer the foundational unit of city making.

This dominance, however, did not go uncontested, with critiques coming from Lewis Mumford and urban activist and scholar, Jane Jacobs. For Mumford and Jacobs, one of the chief complaints of this proliferation of high-speed expressways and towering high-rises was the perceived obliteration of neighborhoods, and with it, the social and moral health of urban society. Mumford attributed many social urban problems to the loss of the neighborhood unit under the onslaught of suburban sprawl and urban renewal. Mumford emphasized the importance of neighborhood design in urban planning and advocated for a centrally planned “regionalism” of Radburn-type neighborhoods, mixed land uses, and mass transit systems that would link neighborhoods together in a vision inspired by Howard’s “garden cities.” Like Howard, the medieval city was also Mumford’s source of inspiration.

Jacobs levied a similar critique in the context of New York. Jacobs observed the disruption, if not obliteration, of the social networks characteristic of established neighborhoods by urban renewal demolition, new expansive expressways, and Corbusian-inspired high-rise towers that paid no attention to the importance and design of “traditional” neighborhoods. For Jacobs, the development of positive social relationships is based on close proximity between work and home, higher density living, and a vibrant and walkable “street life.” This new emergent city, however, was understood as the antithesis to this ideal vision of urban life. While Mumford and Jacobs are often discussed as oppositional figures, Jacobs taking a more bottom-up approach than the broader “regionalism” of Mumford, both have provided the inspiration for the increasingly popular “neotraditional” design principles of the more recent New Urbanism movement.

The modernist planning paradigm remained dominant through the 1960s, 1970s, and 1980s. But in the 1990s, amid the broader postmodern turn in architecture and planning, this direction began to shift as suburban sprawl and the design of public high-rise housing came under scrutiny. The public housing of the urban renewal era, which was now considered plagued by gang
violence, drugs, and crime was now treated as a horrific failure. Suburban sprawl was also increasingly demonized as wasteful, auto-dependent, and “unsustainable.” The monotony of large subdivisions, cookie-cutter homes, and lack of anything to walk to, prompted the accusation that this landscape had an antisocial effect and fostered an individualist mindset not conducive to the development of strong, cohesive neighborhoods.

The New Urbanism movement, led by architect and planner Andres Duany, arose as a response to the problems of both the “soulless” suburbs and poverty-stricken inner-city areas perceived as void of healthy, sustainable “communities.” Here, the neighborhood unit returned as an emphasis in urban planning and design. As an emergent and increasingly dominant planning paradigm, many new developments – across urban America and beyond – are now adopting “neotraditional” design elements, that is, increased housing density, mixed land uses, socioeconomic diversity, walkability and access to rail transit, access to civic centers and/or “town squares,” and close proximity between job and home. Inspired by Mumford and Jacobs, these design principles have become widespread, and are once again considered the basis for positive human relations.

Although modernist planning principles remain prevalent in many parts of the world, the New Urbanism now informs both suburban growth agendas and central city revitalization efforts in a wide variety of contexts.

The neighborhood unit in urban theory

The neighborhood also has been a fundamental object of analysis in urban studies through the twentieth and twenty-first centuries. As the social and spatial constitution of neighborhoods has continued to evolve, much research has sought to identify, describe, and make sense of the factors that drive change. During the early twentieth century, the Chicago School of Urbanism, led by Robert Park, Ernest Burgess, and Louis Wirth at the University of Chicago, presented a conceptual framework for describing and understanding urban life in the modern city, including the role and function of neighborhoods.

In this school of thought, neighborhoods are conceived as functional units within a broader, “organic” city. Not only is the constitution of neighborhoods of interest, but so are the flows and relations between them. During the early twentieth century, North American cities were experiencing an influx of immigration (i.e., from Europe and East Asia) as well as the Great Migration of African Americans from the American South to northern cities. Through this process, neighborhoods increasingly took an “ethnic” form, that is, black neighborhoods, Chinatowns, Italian, Polish, and Irish neighborhoods, and so on. This “ethnic” neighborhood formation has also marked many cities across the Global North and South.

Among Chicago School scholars, the emergence of “ethnic neighborhoods” informed what would become an influential model of neighborhood change: invasion-secession theory. Neighborhood change was propelled by a supposedly natural process: the “invasion” of one ethnic group moving into an already established neighborhood inhabited by another population. Invariably, the current population is forced to relocate elsewhere in the city to reestablish their neighborhood identity, constituting another invading force elsewhere. Typically, it is a lower-status (and lower class) ethnic group that invades and forces the secession of higher-status
NEIGHBORHOOD

groups, a seemingly inevitable process as immigrants continued to flood North American cities at the time.

The wealthy are first to vacate and reestablish new affluent (and typically white) communities, thereby creating a new ring of residential development around the central city. This process is reflected in the commonly referenced concentric-ring diagram of Chicago, where the outer ring of the city is characterized by higher-income residential areas while the “zones of transition” mark the older neighborhoods experiencing change. As the housing stock ages, older neighborhoods are seen to go through a cycle of successive population groups moving in and out, each one lower in status, which explains why, at the end of the cycle, lowest-status groups are observed to inhabit the least desired urban spaces, the so-called ghettos.

This model, however, was challenged in the 1970s by scholars working in a different paradigm: urban political economy. The Chicago School purportedly presented more of a descriptive than explanatory framework. Influenced by French Marxist scholars (notably Henri Lefebvre), those working in urban political economy emphasized a dimension of this process previously downplayed: segregation. For scholars such as David Harvey and Manuel Castells, the urban mosaic of ethnic neighborhoods formed via the power of structural barriers (legal, economic, political). The evolving differentiation of residential landscapes was not a “natural” process, but was shaped by discriminatory practices (i.e., redlining) designed to preserve existing social structures, power relations, and elite interests in the city.

Critical of the Chicago School, these scholars explained the patterning of residential differentiation as a necessary condition for reproducing the basic social relations of a capitalist society. In this view, the neighborhood is understood as the principal site for social reproduction by maintaining a social and spatial division of labor consistent with the dynamics of capitalist development. Life in a lower-income, working-class neighborhood is likely to lead to the reproduction of the same lower-income (and necessary) workforce in the future. Conversely, life in an affluent neighborhood, far removed from the realities of the inner-city, ethnic “ghetto,” immerses the next generation of elites into the upper echelons of the social structure.

At the same time, as metropolitan regions continued to suburbanize, creating sprawling landscapes of monotonous residential subdivisions, large swaths of central city areas were bulldozed to accommodate an evolving modern city. The erosion of the “traditional” neighborhood unit was now mourned by commentators who identified not only the modernist planning paradigm denigrated by Jacobs and Mumford, but also a new neighborhood-destroying force: globalization. During the 1970s and 1980s many geographers were linking this unraveling of local communities (which was being experienced across the advanced, industrial world) to a world-homogenizing globalization (Tuan 1977). The distinctiveness of neighborhoods was being eroded and replaced by a globalizing, monolithic culture, characterized by new massive shopping complexes, cookie-cutter subdivisions, and rampant consumerism.

A series of debates ensued, in geography and beyond, wherein local communities and identities were considered by some as less destroyed than transformed. While the social and spatial attributes of “traditional” neighborhoods were indeed disrupted, a new kind of neighborhood emerged, reconfiguring, rather than dissolving, the social networks and relations of communities. Here, supposedly, was the transformation of “traditional” communities to “modern” communities where the material conditions
that underpinned “traditional” communities no longer existed (Cox and Mair 1988). New conditions were underpinning the formation of different kinds of communities which now revolved around the workplace and sites of civic engagement rather than face-to-face interactions and tight-knit relations among neighbors.

For others, neighborhoods were conceived more as “concentrations of capital” than distinct cultural formations, whereby particular ensembles of local cultures and sociospatial networks form in relation to local employment conditions (Harvey 1989). But as employment conditions change, as they often destructively do during periods of economic restructuring, so does the social, cultural, and political constitution of the very local communities and neighborhoods that define places. In this setting, neighborhoods are understood as always evolving in response to the temporal rhythms of the broader, global economy.

Others have noted the reinforcement, rather than erosion, of local community cohesion and identity. In response to threats to their community or neighborhood identity, people are also seen to turn inward, seeking out new insular and increasingly balkanized forms of neighborhood and community development. This tendency has been observed in many contexts, such as the increasing presence of affluent “gated communities” in both urban and suburban contexts; Homeowners Associations, which function to organize civic engagement and neighborhood identity within an increasingly “placeless” landscape; and the new suburban frontier, which increasingly incorporates the principles of New Urbanism by modeling the design of new subdivisions on the neighborhood unit.

The neighborhood concept has also played an important role in urban redevelopment strategies, particularly in the context of gentrification. For proponents, “bad” neighborhoods are portrayed as being “revitalized,” a process often touted as reforming the social bonds and relations that make what are often simply called “good” neighborhoods. Gentrification, where a higher class displaces a lower class, is typically portrayed, among both proponents and critics, as unfolding at the scale of the neighborhood. Scholars engaged in gentrification research (Lees, Slater, and Wyly 2008) frequently debate the definition and role of the neighborhood within this process.

Among proponents, gentrification does improve neighborhoods by attracting more resources, jobs, and investment. For critics, this argument privileges the neighborhood as a spatial construct, as the process typically does little to benefit pre-existing residents, many of whom become displaced. For these scholars, the physical space of the neighborhood may have been improved, but it is typically a new population group that inhabits it, one that replaces the social fabric of what previously existed in favor of that which is deemed more desirable for generating revenue and attracting resources. The act of defining the neighborhood becomes an act of politics.

The neighborhood has historically been conceptualized and studied in myriad ways by urban geographers, policymakers, and planners. It has also taken many social and spatial forms. And while the neighborhood is primarily documented in the Anglo context, research is increasing across the Global North and South. Despite such diverse treatment, the common thread unifying conceptualizations of neighborhood is the scale of interpersonal interaction and social identity as dynamic and constantly evolving. The neighborhood will undoubtedly continue to represent an important analytical object among a wide range of social scientists.
NEIGHBORHOOD

SEE ALSO: Chicago School; Community; Identity; New Urbanism; Regionalism; Urban geography

References


Tuan, Yi-Fu. 1977. *Space and Place*. Minneapolis: University of Minnesota Press.
Neighborhoods and health

Karen Witten
Massey University, New Zealand

That geography matters to health is not a new idea. In “Life and Death over the Millennium,” Margaret Whitehead records fluctuating population trends in the United Kingdom, tracking the impacts of famine, diseases, and epidemics for over 900 years (Whitehead 1997). She draws on data published in the *Lancet* in 1843 that illustrate enormous differences in life expectancy between occupational groups, residents of rural and urban areas, and people living in different parts of the country. For example, laborers and artisans in Bath had a life expectancy of 25 years compared to 15 years in Liverpool. Gentry and professionals averaged 55 years at death in Bath and 35 years in Liverpool. These variations in mortality between places were attributed to varying exposure to hazardous living and working conditions associated with industrialization (Whitehead 1997).

In recent decades, in a number of industrialized countries, gains in life expectancy and living standards at a population level have not been equitably distributed across groups within populations, and health inequalities have widened. This situation has stimulated renewed policy and research interest in the social and spatial determinants of health inequalities. A notable example was the Commission of Social Determinants of Health established by the World Health Organization to collate evidence on how to achieve health equity (CSDH 2008).

Where a person lives has a profound impact on access to opportunities and the routines of daily life, many of which have implications for health status. For example, residential location determines access to educational and employment opportunities; the availability of public transport or cycling infrastructure may influence whether the trip to work or school is made by car or active transport; ready access to fresh and healthy food supplies can affect food consumption practices; and the social environment of a neighborhood can impact the breadth and depth of social networks and support available. All these factors have downstream consequences for people’s health status.

Investigating the relationship between place, or neighborhoods, and health poses a range of conceptual and methodological challenges: defining and operationalizing “neighborhood”; defining and measuring neighborhood effects; theorizing the pathways through which neighborhood environments affect health; identifying what aspect of health may be affected and for whom. Study in this field has benefited greatly from advances in geographic information system (GIS) technologies and statistical modeling. Virtual databases such as Google Street View can now be used to conduct street audits that have in the past required in situ observational methods.

Neighborhoods have spatial and social aspects. Most people have an intuitive understanding of where their neighborhood lies and if asked will likely define an area that encompasses the homes of immediate neighbors and the services and amenities they use on a regular basis. However, as a research construct, “neighborhood” can be difficult to define and operationalize. In quantitative research administrative boundaries or census units have been commonly used to define a neighborhood. They are convenient, can be
aggregated at various spatial scales, and enable the ready integration of other administrative data (e.g., a census). However they may not represent where residents spend time or what they consider to be their neighborhood. A more recent approach has been to define neighborhood as a buffer distance surrounding a residential address. This is termed an “ego-centered” neighborhood and is measured using the distance along a street network (rather than Euclidean or straight-line distance). A neighborhood defined in this way is more likely to represent the places a person spends time and buffer distances can be set according to the health outcome of interest and population group involved (e.g., a child’s neighborhood buffer may be smaller than an adult’s). The social aspects of a neighborhood are still unlikely to be captured by these methods (Chaix et al. 2009). A self-defined neighborhood is more feasible in qualitative studies.

A broad range of neighborhood factors that could affect health have been investigated, with a distinction often made between contextual or compositional effects. Among the contextual effects that have been examined are attributes relating to the neighborhood physical environment (e.g., air or water quality, urban design); institutions and services (e.g., transport, health, and education); amenities (e.g., parks and other recreational facilities); and the social environment (e.g., neighborhood social cohesion). The characteristics of the people living in a place have been termed “compositional effects.” In some areas of interest there is no clear distinction between contextual and compositional effects. For example, housing tenure can be considered a compositional variable when aggregated to a neighborhood level, yet it may have contextual effects if higher levels of homeownership result in increased social and financial investment in an area. Likewise, the aggregate level of educational qualifications held by residents in a neighborhood would be a compositional effect, but it is feasible that educational participation and outcomes may be influenced by the availability and quality of educational opportunities, a contextual effect.

How neighborhood factors influence health has been investigated for an array of health outcomes including heart disease, physical activity and obesity, low birth weight, smoking, and overall mortality. While many studies have found significant neighborhood effects on health, the effects are generally small compared to the effects of individual sociodemographic factors. In light of worldwide concern over rising obesity rates, a topic that has received considerable research attention in recent years has been the relationship between neighborhood design, residents’ physical activity for transport and recreation, and obesity. There is now strong evidence, in particular originating from the United States, Australia, and New Zealand, of associations between built environment characteristics such as dwelling density, mixed land use, and street connectivity and the physical activity and body size of residents. Individuals living in more sprawling, lower-density neighborhoods are more likely to be less physically active and have higher rates of overweight and obesity than those living in more compact urban areas.

Another topic of contemporary interest is the impact of housing and neighborhood interventions, such as urban renewal, on health and health inequalities. A systematic overview of intervention studies in these areas concluded that there is relatively strong evidence for interventions aimed at improving area characteristics and compelling evidence for warmth and energy efficiency for improving health outcomes (Gibson et al. 2011).

Quantitative research paradigms have dominated research on neighborhood effects on health but there have also been a number of
illuminating locality-based qualitative studies. Studies that use administrative boundaries to define neighborhoods in combination with cross-sectional or longitudinal survey databases containing health and sociodemographic variables remain common in this field. However, increasingly, specific neighborhood factors hypothesized to have an influence on a specific health outcome are being measured, aided in many cases by the use of GIS technologies. Associations between the neighborhood effects and health outcomes are modeled using multilevel statistical techniques, which enable an analysis of the extent to which observed variations in outcomes can be explained by differences at the individual and/or neighborhood levels. The direct measurement of specific exposure and outcome variables, particularly where common measurement tools are applied across studies, is increasing the precision and comparability of research in this field. Attention is also being drawn to differences in relationships between neighborhood effects and health for people of different ages, gender, and ethnicity.

Intervention and longitudinal studies are relatively rare in this field but are needed to substantiate understanding of the pathways of influence between place and people. A number of intervention studies have been conducted in the United States in which low-income families have been supported financially to move into more affluent areas, a shift which represents a change in neighborhood context and composition. Anderson et al. (2003) conducted a systematic review of 12 intervention studies including randomized controlled trials, and controlled and uncontrolled prospective studies. The review reported some support for improved household safety and fewer experiences of personal victimization in intervention households and mixed effects on health outcomes and health-related behaviors. Research in these areas is very challenging, as the individual and neighborhood factors that could potentially influence the health of residents are numerous.

Locality-based qualitative studies, involving in-depth interviews with small numbers of participants, are a useful complement to the larger quantitative studies. Residents’ experiences and understandings of the relationships between material and social aspects of neighborhoods and their health practices and wellbeing can be explored in these studies.

A concern for social justice in terms of more equitable health outcomes often underpins policy and research investigating neighborhoods and health. Place effects on health may be modest compared to the impact of individual-level determinants of health such as occupation, income, and education but they are potentially modifiable. Knowledge of how neighborhoods influence health can inform place-based interventions to improve access to positive health-promoting resources in neighborhoods where the chance of illness and premature death are high. To quote the Commission on Social Determinants of Health report: “Social injustice is killing people on a grand scale” (CSDH 2008: 4). It is probable that many of the known social determinants of health such as housing, education, and employment operate, at least in part, through neighborhood pathways.

SEE ALSO: Health geography; Health inequalities

References

NEIGHBORHOODS AND HEALTH

Neighbourhoods in Eco-epidemiologic Research: Delimiting Personal Exposure Areas. A Response to Riva, Gauvin, Apparicio and Brodeur.” Social Science & Medicine, 69(9): 1306–1310. DOI:10.1016/j.socscimed.2009.07.018.


Neoliberalism and the environment

Harold A. Perkins
Ohio University, USA

What is neoliberalism?

The prefix “neo” of course means new, or young. The root of the word is “liberal,” referring to an Enlightenment political philosophy. Put the prefix back together with its root and the word refers to a new version of a long-established political philosophy. But to understand better just what is meant by neoliberalism, it is necessary to momentarily separate the root from its prefix again and interrogate what each one means in relation to the other. Start first by interrogating “liberalism.” In the United States, political “liberals” are often people who hold political beliefs that lean left-of-center on the political spectrum. Liberals there are most strongly associated with the Democratic Party that historically tends to favor stronger social welfare, economic intervention, and environmental regulation. But, in fact, the application of the term “liberal” in the context of party politics in the United States is at best a conflation, and at worst, contradictory to what liberalism represents in its traditional and more widely practiced philosophical sense. It is important therefore to address the history of liberalism and delineate its deployment beyond the circumscribed context of current-day Democratic Party politics in the United States.

Liberalism as a philosophy was born during the Enlightenment period in the 1700s and diffused throughout much of the Anglophone and Continental European worlds as a dominant political practice in the centuries that followed. Locke, Mill, Montesquieu, Smith, and Kant wrote about and advocated liberal political philosophy. The geopolitical context in which liberal philosophy was conceived and promulgated in political action is an important consideration. Most of the states in Europe during the seventeenth and eighteenth centuries were ruled by monarchies or some equivalent that wielded power over their respective populations. Political and, to a considerable extent, economic freedoms were limited for most people living under feudal conditions prevalent in many monarchical states. The right to speak freely or own property was restricted in most instances to people born into dynastic systems based on heredity.

Liberalism, in contrast with despotism, eschewed hereditary privilege and instead espoused principles of societal equality, political freedom of expression, representative politics, and the individual right to accrue wealth (own property). Liberalism has a utilitarian component in that actions taken to govern society should be fair and maximize benefit for the largest number of people possible (Rawls 1999/1971). All of these positions were meant to seriously check the absolute power wielded by monarchs throughout Europe. Political liberalism expanded in many of the states that emerged from Europe’s colonial system as well. The United States is an example whereby liberal principles heavily informed the establishment of the state and its relation to its citizens and economy. In particular in the United States the central state was originally envisioned and established as a weak state with little power to interfere with the political and
NEOLIBERALISM AND THE ENVIRONMENT

economic concerns of its republic and citizens. This was the embodiment of the notion of a diffuse and individualist, Jeffersonian America. But of course under most liberal political systems, the only “citizens” who could originally participate in the affairs of governing the state were mostly white men who owned property. So from early on in liberalism’s history, it is exercised in haphazard and contradictory ways. Liberalism not only impacted the political realm, however, it was heavily involved in the development of capitalism.

Liberal political philosophy complemented the industrialization of Western Europe and in the United States liberalism’s emphasis on property rights and minimal state interference in economic affairs set the stage for the rapid expansion of industry. Concomitant with the growth of industrialization were its corollaries in free markets for commodity inputs and outputs requisite to the production process. Indeed liberal political philosophy justified laissez-faire economic regulatory policies by the state that facilitated capital accumulation into private hands. Quite important to this process of accumulation is that liberalism and its emphasis on private property rights enabled capitalists to appropriate common resources for their industrial processes. The privatization of common resources was pivotal to the development of a robust capitalism based strongly in the utilitarian notion that appropriation is justified so long as the action benefits the majority of individuals in society. There are, undoubtedly, many problems with how this has played out historically and geographically. Liberalism’s critiques will be discussed at some length later in this chapter. But before this, it is important to examine the prefix “neo.”

“Neoliberalism” implies that it is a new iteration of the old political and economic philosophy addressed above. But what has changed since the industrial revolution? A look back at history better contextualizes the rise of neoliberalism. Much of the blame for the Great Depression of the 1930s is put on the worst excesses of free-market capitalism justified by extreme interpretations of economic liberalism. Laissez-faire capitalism, while making some people extremely wealthy, was choking its own demand for commodities as wealth concentrated into fewer and fewer hands (Harvey 2007). Political economists of the time believed that pure liberalism needed to be tempered by increased state management of economic affairs to avoid another economic depression. The prominent economist Keynes, for example, believed that the state should pursue demand-side economic policies to stabilize the economy while regulating it more closely to prevent the development of future monopolies. This was the beginning of social democracy in Western Europe, the United States, and other states that expanded their powers to tax and redistribute income through social welfare programs. Keynesianism worked well so long as economic growth continued apace.

Crisis again struck the global economy in the early 1970s, however. Declining economic growth coupled with rising interest rates and rampant inflation in the 1970s spurred resurgence in the demand for liberalized market reforms. Firebrand conservatives like Ronald Reagan in the United States and Margaret Thatcher in the United Kingdom brought real power to bear in their shift away from Keynesianism toward what is now referred to as neoliberalism. The neoliberal turn, starting in the 1980s, was supposed to unleash the power of the market to manage economy, society, and the environment better. But what exactly makes this re-implementation of liberalism “neo”? Initially, liberalized market policies were embedded within, or grafted onto, Keynesian regulatory institutions including demand-side economic policy, social welfare programs, and nationalized
NEOLIBERALISM AND THE ENVIRONMENT

industries. “Neo” liberalization had to cannibalize its Keynesian cousin before it could reign supreme. Thus social welfare programs were rolled back while privatization of national industries, supply side policies, and tax breaks for the wealthy and corporations were implemented in the name of bolstering economic growth (Peck and Tickell 2003). Trade barriers like tariffs were greatly reduced between countries to facilitate global economic growth, too.

These shifts could not occur without a fundamental change in the way people view the relationship of the state to economy, society, and environment. With the return of liberalism come laissez-faire principles that once again have widespread appeal and form a major component of the predominant political economic discourse. Accordingly, the state should not unduly impede the workings of the economy. Bob Jessop refers to this as “the hollowing out” of the nation-state as it retreats from its former Keynesian-style obligations to drive the economy, society, and environment (2002, 235). Jessop and others note, however, that it is not so much about the shrinking of state power as it is a rearticulation and enhancement of state power toward creating rationalized metrics for sustained economic growth (Peck and Tickell 2003). In keeping, Moran (2003) suggests that the neoliberal turn in the United Kingdom, rather than diminishing powers of the state, has instead created a more formalized regulatory environment that constantly fosters novel and flexible policy innovations to steer its economy efficiently through globalization’s turbulence. Concomitantly, neoliberal states (especially in the developed world) continue to exercise their power over societal and environmental regulation; however, they do so indirectly through the coordination of civil sector institutions that provide social and environmental services in lieu of direct state provision.

Despite this rearticulation of state power, in a globalized world economy states still struggle to advantageously control flows of capital, information, and people across their borders. Even a neoliberal world economy thus requires rules for economic engagement at the international level that must be established and enforced through formalized supranational institutions and free-trade frameworks. These institutions include World Bank, International Monetary Fund (IMF), and the World Trade Organization (WTO) that coordinate global flows of capital at the supranational scale (Peet 2003). These institutions are orchestrated by wealthy and powerful states like the United States or those in Western Europe; perhaps more importantly, they represent the private interests in these states that have capital to lend to states and firms throughout the world. This is important because states – particularly in the developing world – participate in the global economy according to strict liberal edicts spelled out by these institutions. In fact, participation in neoliberal globalization is predicated on the adoption of comprehensive market and private property-based systems everywhere.

Many states in the developing world borrowed heavily from Western financial institutions, and are having a hard time servicing their debts. The World Bank and IMF lend states more money to repay their extensive obligations, but with certain conditions attached. Debtor states are frequently subjected to structural adjustment programs, or SAPs, in order to get funds from the World Bank/IMF. SAPs are imposed to “align” states’ economies with the global, free-market order. Terms include prioritizing export-oriented policies and the dismantling of social welfare programs and economic/environmental regulations believed to interfere with market expansion (such stipulations are even being imposed on debt-ridden,
developed states in Southern Europe). These terms are imposed to facilitate states’ participation in the global marketplace. It is important to remember that philosophically the idea behind liberalism is that marketization is supposed to eventually enhance the wealth and wellbeing of people through capitalist development rather than direct state intervention/provision. In fact, a global middle class (loosely defined) has grown by millions in the developing world with the rise of neoliberal capitalism. Critics point out, however, that liberalization of national economies (especially through SAPs) amounts to the transfer of common resources into lucrative forms of private property that disproportionately benefit investors from far-flung locations (Peet 2003; Harvey 2005a). People are sometimes harmed by these reforms that infringe on their resource base and their ability to make a livelihood. Thus it is important to note that neoliberal reforms and development land differently according to geographical context and therefore have different impacts on people throughout the world. Neoliberalism is therefore not monolithic in its outcomes as its reforms are grafted onto extant political economic landscapes, both material and discursive. One commonality to these reforms, however, is that they usually feature some pain and controversy. Strategies must therefore be deployed to deal with such exigencies.

A typology of neoliberal strategies concerning management of environmental resources

Noel Castree elaborated four interrelated ways in which supranational organizations, states, firms, and individuals seek to manage the environment within neoliberal capitalism (2008). As already noted, the expansion of neoliberal capitalism throughout the world increases prosperity but is also fraught with contradictions and tensions. Thus Castree notes there are ways that neoliberalization of the environment is legitimated in the face of such challenges. He refers to these strategies as neoliberalism’s “environmental fixes.” They have much overlap but are useful in helping articulate the primary forms by which the environment increasingly becomes a part of neoliberal capitalism’s material and discursive basis for existence. As he goes on to suggest, these fixes, too, are only partial and usually laden with contradictions. After briefly outlining the fixes here, examples will be provided to contextualize them in the following section.

Environmental fix 1: free-market environmentalism

In this neoliberal strategy the state steps away from direct environmental management to allow market mechanisms to take over conservation efforts. The notion here is that markets and their related price signals are just as capable of protecting “valuable” aspects of the environment as states with their more rigid practices of “fencing off” entire chunks of the environment. Note that market mechanisms for environmental conservation allow for the flexibility of economic transactions across spaces otherwise precluded under traditional modes of state environmental protection. For example, forests are assigned market value so that rational market participants are incentivized to protect the overall resource while sustainably harvesting some of the timber for sale. The argument here is that the market participants will not cut down all of the trees because they too desire to keep the market for timber products alive.
Environmental fix 2: accumulation by dispossession

Unlike the first fix’s emphasis on eco-capitalism, some iterations of neoliberalism appeal to the need for growth at the expense of the environment (and some people who depend on it directly). Here the plight of the economy is predicated on enhanced access by capital to more of the environment formerly managed by the state or by local communities. Often this requires reduced state protections and also can involve the direct appropriation/transfer of lands and resources by/to firms at the expense of communities and individuals who depended on them in the first instance. The idea is that cheaper inputs to production enhance a firm’s profitability—hence the accumulation of capital by the dispossession of resources from others.

Environmental fix 3: institutions and discourse

Parties involved in the neoliberalization of the environment recognize their efforts can be harmful to the environment and by extension damaging to the resource base on which other people depend. Thus strategies are needed to refute contestations and build consensus around environmental transformations that are in some contexts beneficial to capital alone. The creation of institutional frameworks is highly effective in this regard. The North American Free Trade Agreement is a good example, where the right of corporations within member states to consume and pollute free of state (or civil sector) interference is legally protected. By extension, these frameworks help create a discourse of inevitability that circumscribe alternative antiglobalization discourses including “limits to growth.” In many ways this fix is effectively self-referential. The rules of commerce are already well-established and are not easily challenged by those who feel wronged by them.

Environmental fix 4: state legitimacy

Castree notes this strategy perhaps involves the state more than the previous three. He notes the state must maintain its legitimacy and avoid/manage crises caused by conflicts regarding environmental management and resource procurement under neoliberalism. This is difficult because the state increasingly takes a neoliberal stance that promotes the fixes above, yet has to deal with the inevitable fallout these actions may spark, especially among citizens who feel they have lost access to a resource base they need for their survival. This puts the state into a dilemma: how much should it intervene in the management of its economy, society, and environment? Castree suggests this depends on the state itself. Developed states with a history of direct economic/environmental management increasingly depend on civil sector organizations to do the work of regulating the relationship between economy, society, and environment. Meanwhile, in lesser-developed states with little or no history of direct intervention, the state will often take a minimalist stance and rely mostly on the self-regulation of market mechanisms.

Implications and critiques of the neoliberalization of the environment

Free-market environmentalism?

It is important to recognize the atmosphere has long been imperative to the development of the globalized capitalist economy. It is absolutely essential to economic growth because it is treated as a giant pollution sink that absorbs all the
unwanted byproducts of industrial processes. Of particular concern these days is the carbon dioxide the atmosphere is made to absorb from the combustion of fossil fuels. Scientific consensus is that the planet, on average, is warming because of these kinds of greenhouse gasses (GHGs). For the most part in the market-driven world, however, states have rejected the idea of command and control regulatory actions that would force industries under their jurisdiction to reduce their emissions by a given amount. In a neoliberal political economy such a strategy is viewed as more costly, less profitable, and ultimately hindering economic growth. Polluters would likely push back hard against absolute reductions imposed by states. The neoliberal solution requires environmental protection that does not threaten the global economy. Hence a system is created whereby polluters are encouraged to innovate and reduce their pollution voluntarily through market mechanisms (Bumpus and Liverman 2008).

Treaties like the Kyoto Protocol and more recent Paris Agreement set targeted reductions in GHG emissions rates for participating states. Individual states and the European Union have set their own GHG emissions reductions goals in conjunction with their ratification of Kyoto and also independently of these agreements. Reductions targets are largely supposed to be met through the creation and implementation of a market mechanism known as “cap and trade.” Under cap and trade, states issue a certain number of GHG pollution “credits” that in total will limit the amount of GHG emitted into the atmosphere in a given period of time. Credits are sold on open market exchanges between firms that pollute the atmosphere with their emissions. The idea is that firms that innovate with their technologies will use less credits than they possess, and can then sell them to polluting firms that require more credits to be in compliance with the agreement. The hope is that firms will reduce emissions because they have incentive to profitably sell their unwanted credits.

A credit system under cap and trade treats the atmosphere as divisible into units with pollution absorbing capacity. In essence, polluters are buying and selling “pieces” of the atmosphere as a form of private property that they have the right to pollute. These credits are priced according to market signals (supply and demand). In dividing up the atmosphere, cap and trade places monetary value on its entirety and in doing so creates a common valuation scheme that is supposed to protect it as a whole. By extension the global economy is spared the damaging shock from large and immediate reductions in the combustion of fossil fuels. Indeed emissions under cap and trade, particularly in some developed states, have begun to level off. Cap and trade is not without contradiction and controversy, however. A fundamental concern arises as to how well the atmosphere (and everything that depends on it) is ultimately protected when polluting is still permitted.

It is possible that cap and trade does not limit GHG quickly enough. Some suggest command and control systems where states exercise more power to directly limit emissions from fossil fuels are imminently necessary. Other criticisms contend that cap and trade fails to spur much innovation and instead allows the most powerful and wealthy polluters (particularly in the Global North) to purchase their way out of reductions while making a tidy profit in the process. Another problem is that the United States, one of the largest total emitters of GHG, refused to even ratify Kyoto. A lack of participation by large polluters seriously undermines the effectiveness of such solutions and is a big concern for the future of GHG reductions recently negotiated in the Paris Agreement. Yet another concern is that it is difficult to
NEOLIBERALISM AND THE ENVIRONMENT

quantify, monitor, and enforce these schemes. What this all potentially means is that neoliberal, market-based solutions to emissions may not be enough to prevent catastrophic changes in the global climate that could potentially destabilize the global political economy, too. There is a growing sense this is contradictory in that capitalism may be destroying its productive basis in a previously stable atmosphere and climate regime.

Neoliberalization of the environment as “accumulation by dispossession”?

One of the major drivers of a shift away from Keynesianism toward a neoliberal global political economy was the stagnation of capital accumulation by the 1970s. Firms began to look out on the horizon for cheaper inputs to production that would lower costs and by extension increase profits. This drive to renew accumulation by the late 1970s and early 1980s fueled globalization in the sense that firms raced from the developed world with its high costs of production (materials, labor, regulation) to the developing world with its comparatively lower costs. An offshoot of this process was that the material base for enhanced accumulation needed to be secured amid considerable political and social uncertainty within an array of lesser-developed states. At the same time, many of the states with these untapped resources were highly indebted to creditors based in wealthy states. This proved convenient for capital as creditor states were able to bring power to bear over indebted states through global institutions like World Bank/IMF. These institutions devised and implemented the disciplinary frameworks mentioned above that further opened up the resources within developing states to multinational corporations. The transfer of resources has created an enormous amount of wealth globally and is one of neoliberalism’s most significant accomplishments.

Of course the imposition of market liberalization onto previous collective forms of resource management usually requires state support and intervention so that commons resources like water, forests, and minerals can be sold to private interests. Debtor states in the developing world, for the most part, have been eager to promote their resources on the global marketplace in order to enhance their own coffers and to establish their own primacy on the global economic stage. It is important to remember, however, that the transfer of public resources to the private sector represents a significant transfer of wealth. The recipients of these new forms of private wealth are often multinational corporations (MNCs) and their investors that conduct their business around the world and have enough capital to purchase large tracts of the Earth’s surface and the resources it contains. This process drives neoliberalized development as the logic of marketable, private property is extended into places it did not exist before. It sometimes accomplishes this goal at considerable expense to the people already subsisting on these resources prior to the entrance of globalized capital.

Opponents of neoliberal globalization believe accumulation by dispossession is a result of the global expansion of markets, commodification of the environment, and enclosure of the global commons. Accumulation by dispossession is said to occur when MNCs seek out more profitable investments by cheapening their inputs to production in novel private property arrangements and in turn the people who previously subsisted on those same resources lose their access to them (Harvey 2005a). Take the Congo Basin, for example. Expanding market logic into countries in the Congo Basin by enclosing the land and commodifying the trees potentially
NEOLIBERALISM AND THE ENVIRONMENT

dispossesses the people living off the forest from their resource. It is a profitable relation for firms looking for cheap and abundant sources of high-quality hardwood, but it threatens to sever the subsistence connection local communities have with their trees.

Dispossession, neoliberal detractors go on to claim, frequently requires some form of state violence on behalf of MNCs to subvert formerly collective modes of environmental governance by local and indigenous communities. This is in spite of the laissez-faire stance the state is supposed to take in the neoliberal global economy. State intervention is necessary to secure the market because dispossession (or the perception thereof) frequently sparks intense resentment by adversely affected parties. State power exercised in securing private property markets against local backlash is illustrative of the strong possibility that rather than a weakening of state strength under globalized neoliberalism, the state largely rearticulates its power toward protecting its growing resource markets and thereby enhancing its position and stake in the global economy. The Cochabamba water war in Bolivia is a good example of this possibility.

The World Bank encouraged the corporation Aguas del Tunari, in partnership with Bechtel, to acquire and manage the municipal water supply for Cochabamba and a few other Bolivian cities in the late 1990s. The corporations immediately raised water tariffs to the point where poor Bolivians were spending nearly 20% of their monthly income on their water bills, making it more difficult for them to access water. Increased rates were supposed to fund better delivery systems, but in Cochabamba, there was not more water in the taps after the rate increase. A small series of local protests in early 2000 quickly grew into a national strike when protesters shut down Bolivia’s highways and its economy. The Bolivian president declared a state of siege and prompted the military and police force to violently repress the thousands of protesters. Many protesters were injured and at least one teenage boy was shot and killed by the military during its crackdown. By April of 2000 the Bolivian government capitulated and abandoned its contract with Aguas del Tunari. Cochabamba is a rare instance where protest groups forced the state to back down from its neoliberal stance.

Critics sometimes liken accumulation by dispossession to an enclosure of the global commons (Harvey 2005b). Enclosure is a concept more often applied to seventeenth-century England when the growing capitalist political economy there pushed rural people off their land and concentrated them in cities where they were made into an industrial proletariat. Many other examples of enclosure have since occurred around the world including the removal of indigenous people by force from their lands in the Americas, Australia, Africa, and Asia. While the historical and geographical contexts of enclosure then and now are surely different, there are potential commonalities. Most important, perhaps, is that subsistence cultures dependent on a resource lose access to it when it is privatized and as a consequence must move, adapt, or fight back. Vandana Shiva writes much about how numerous indigenous communities in India had their forests declared as “wastelands” by the Indian state and transferred to multinational capital as private property (2013). Many subsistence cultures in South Asia have been destroyed in the process; some still hold out against the Indian state and capital. Shiva and others contend that neoliberal capitalism relies on enclosure much as colonizing states did centuries ago. Thus they argue neoliberalism is a form of neocolonialism.
Institutions, frameworks, and the discourse of inevitability

Almost every state, perhaps with the exception of North Korea and Cuba, is participating in the globalized free market. While some powerful states, like the United States, largely chart their own course for participation, many developing states have their participation in the global marketplace substantially orchestrated by the World Bank, IMF, and WTO. While these institutions outwardly seek to develop states, they are also trying to smooth the global economic surface for the maximization of capital flows and growth. This has profound implications for the people and environments subjected to SAPs. Again, it is important to remember that SAPs are neoliberal frameworks imposed on debtor states to help them avoid defaulting on their outstanding loans. The threat of default means that many states would otherwise be ineligible for future development aid and risk being further isolated from the global economy. This is a risk most states are unwilling to take, making the adoption of SAP frameworks by debtor states practically inevitable. SAP frameworks are fairly complicated agreements, but there are a few overarching commonalities of SAPs that impact people and their local environment.

SAPs frequently require trade liberalization (Peet 2003). Part of the idea behind trade liberalization is that developing states tax their growing export industries and use those funds to repay their loans. States subjected to SAPs are often expected to make their natural resources available to the global market. In other words they must be willing to open up their commons to private interests for exploitation (and in the process states should generate revenue to repay loans). SAPs encourage states to export tropical hardwoods like mahogany, for example, that command strong prices globally. States may also be encouraged to bolster their production and export of bananas, cacao, teak, palm oil, or coffee. Debtor states must, however, also open up their domestic markets to international competition. For example, small-scale coffee producers in Central American states lose protective tariffs that keep their product domestically competitive and must therefore compete with coffee producers of various sizes located all over the world. These actions bring many more low-cost products to the global marketplace and can thus be considered a success in neoliberal terms.

In addition to SAPs, international free-trade agreements also augment the shift to market liberalization (McCarthy 2004). These frameworks marketize the farthest reaches of member states’ hinterlands with consequences for local people now subject to their principles. Mexico’s participation in NAFTA, for example, was predicated on the Mexican state implementing reform of its land tenure policy long enshrined in its constitution. Essentially the state had to put on the market some of its lands guaranteed to farmers as part of its ejido system. Many small farmers relied on this lost land for their livelihood. This was the basis for the Zapatista Rebellion in the State of Chiapas where indigenous groups violently reacted to the sale of their land to MNCs at the behest of the Mexican state that had to follow NAFTA’s rules. The NAFTA framework also eliminated Mexican tariffs on the import of foreign sources of corn. Subsidized corn produced in the United States flooded the Mexican market, driving many small-scale producers out of business and off of their land. Large-scale protests occurred in Mexico City.
and elsewhere, but resulted in little change in the situation as the rules are encoded in the language of the NAFTA agreement. In these instances the state usually stands behind the inevitability of the legal rules for participation in the free-trade framework.

Liberalization is also finding its way into the developing world through the work of institutions other than the World Bank/IMF and the WTO. United Nations (UN) frameworks like its Millennium Development Goals and its efforts to provide potable water to the world’s poorest residents are a good example. Water security efforts coordinated by the UN frequently rely on liberalized, market initiatives to build and manage water infrastructure in resource-strapped contexts where there is little potential for state provision. Multinational water firms are also encouraged to privatize extant state and municipal water supplies if public utilities are deemed insufficient in their capacity to provide water. Part of the proceeds from the private management of water systems is supposed to fund the construction of new systems by the same corporations elsewhere. While there have been some successes, critics contend that the liberalization of water provision can be a continuing hardship for poor people who cannot afford the water tariffs (Shiva 2002). Violent reactions sometimes occur with water privatization, especially where public systems previously offered water at little cost and the new private managers cut off water supply to those not paying their bills. The Cochabamba water crisis mentioned above is one such example. Advocates of market-based strategies for water delivery insist, however, that other viable options for generating funding to build public systems are lacking. Again, this is another instance of the discourse of inevitability associated with neoliberal reforms.

A problem of state legitimacy?

The management of water provision can again be looked at to help answer this question. A growing human population, climate change, and a global economy predicated on unfettered growth are putting incredible strains on the world’s freshwater supply. A lack of potable water is the most pressing of public health crises and therefore is increasingly difficult for states to ignore. To do so runs the risk of losing legitimacy in the face of unrest/protest. Many states, supranational institutions, and municipalities increasingly believe the best way to regulate and protect the world’s freshwater supply is to turn it into a saleable commodity in an established water market (Bakker 2010). The liberalized idea is that when water is valued at its “true market price,” it will be conserved and delivered most efficiently. The growth of private water provision has thus expanded rapidly in the globalized, neoliberal political economy. A handful of MNCs and their subsidiaries are now responsible for much of the world’s potable water supply. The privatization process, however, plays out differently according to geographical context.

In many parts of the developed world, formerly public (often Keynesian) systems are transferred to the private sector and in essence are turned into private property. Public water systems are privatized in developed states for a variety of reasons, including the decline of infrastructure and the need to expand supplies. Mitigating these problems is expensive for the public sector. Some water systems in the United Kingdom, United States, and Greece, for example, were transferred to private firms. Transfer accomplishes two things. First, it bolsters the state’s reputation for fiscal responsibility; and second, it is a redistribution of wealth in the form of public sector assets transferred to the for-profit
private sector. These results provide states with legitimacy from a neoliberal perspective.

In lesser developing contexts, by contrast, states may be eager to expand water provision but political economic realities make it extremely difficult to afford the construction of systems. States, sometimes with UN cooperation, encourage corporations to build new water procurement and distribution systems in places where none previously existed, or perhaps where the extant public system is deemed to be insufficient (Cochabamba). The idea here is that firms use their investment capital to build new systems and thereby leverage more capital through the sale of water to newly connected consumers. The firm can then build new systems elsewhere. This is a particularly common strategy for solving water provision problems in sub-Saharan Africa. From a neoliberal, market-based standpoint, this is a win–win situation. Companies make money from water tariffs while more people gain access to safe drinking water. But there are problems with making water into a marketable commodity, particularly in the developing world, as the Cochabamba example makes clear.

Under neoliberalism people are treated as customers who have to pay to access it. This is probably acceptable for people who have enough money to pay their water bill in the United States or United Kingdom, but herein lies the potential contradiction. As critics suggest, the architects of these privatized water systems in the developing world sometimes underestimate the degree to which new customers have a difficult time paying their water bills. Evidence suggests that poor people often cannot afford the rates charged to them. Thus they end up walking to get water somewhere else, or they sabotage the new system and gain access to the water illegally (Loftus and Lumsden 2008). Overt acts of violence between residents and corporations/states do occur as a result of the contradictions of the privatization process. As was clearly the case in Bolivia, neoliberal fixes like water privatization can sometimes lead to further legitimacy crises for states.

**Concerns about environmental injustice**

Neoliberalism is controversial. In the developed world, the establishment of market liberalization requires the erosion of statist environmental protections won during the Keynesian era. Environmental justice advocates contend the rollback of environmental regulation and defunding of public environmental goods may benefit the overall economy and state budgets, but they say it also leads to a polarization of society where some people live in increasingly degraded environments. As already demonstrated, neoliberalism manifests differently in the developing world and as such it has created a different range of concerns about environmental injustices there.

It is true that millions of people living in the developing world are entering into some kind of middle-class lifestyle with the rise to prominence of market liberalism. Yet neoliberalism is fundamentally at odds with many communities’ (indigenous or otherwise) traditional modes of governance over resources and opponents suggest it will assimilate or dismantle anything that stands in its path (Shiva 2013). The lack of democratic capacity in relation to local communities’ control over the neoliberalization of their environment is well understood. A lack of local governance is important because there is little doubt that the world is a more polluted place than ever before. Intensified production, marketization, and associated consumerism in the name of global economic growth have led to severe environmental damage in the remotest corners of the world where people are subsisting.
From the fields of soya encroaching on the rainforest in the Amazon, to the petroleum polluted Niger Delta estuary, the world’s most productive biomes (and the human communities that live there) are often irreparably damaged by market expansion (Watts 2008). Such damage will be extremely difficult to remediate.

Critics contend there is little recourse for losers in these kinds of scenarios to make meaningful change in their respective states, let alone at the global scale. Neoliberal institutions like the World Bank/IMF and the UN as well as the MNCs that benefit from their frameworks all wield enormous amounts of power to coordinate global market relations. Negatively affected people will continue to protest the actions of these powerful interests and their market reforms. As is the case in Cochabamba, they will turn the tide against privatization of public resources in a few instances. But the cost of violent protest against neoliberalism and its supposed liberation of the individual is often a loss of human life in places like Bolivia and Nigeria. This is one of globalized neoliberalism’s most deep and intractable contradictions.

**SEE ALSO:** Commodification of nature; Corporate environmental responsibility; Ecological modernization; Environmental governance; Environmental (in)justice; Environmental policy

**References**


Neoliberalism

Jamie Peck
University of British Columbia, Canada

Neoliberalism is an umbrella term applied variously to a programmatic bundle of pro-market, pro-corporate, and pro-choice policy measures; to a rationality of small-government transformation, modeled on the principles of entrepreneurialism and competition; to the political-economic philosophy and belief system of “market fundamentalism”; and to a historically ascendant ideology (and pattern) of capitalist development, linked to globalization, financialization, and the hegemony of “market rule.” Signature policies associated with neoliberalism include privatization (the sale of state assets, the outsourcing of government services, and the reallocation of functions from the public to the private sector); deregulation (the diminution or recalibration of state control in favor of more competitive, market-based, or choice-oriented approaches); various forms of marketization (the adoption of market-like principles and subjectivities, based on competition and choice, across a wide range of governmental and nongovernmental spheres); structural adjustment (the enforcement, by way of loan requirements, of neoliberal restructuring measures by multilateral agencies, usually in the context of currency or debt crises, or “bailouts”); and austerity (the imposition of public expenditure cuts and strict fiscal discipline, especially since the 2008 financial crash).

Iconic figures of neoliberal thought include the philosopher-economist Friedrich Hayek, whose 1944 book, *The Road to Serfdom*, is considered by many to be a founding text in the movement for free markets, and the Nobel Prize-winning economist from the University of Chicago, Milton Friedman. Vanguard politicians especially associated with the neoliberal cause include Margaret Thatcher (UK prime minister, 1979–1990), Ronald Reagan (US president, 1981–1989), and General Pinochet (Chilean president, 1974–1990), although centrist, “third-way” politicians like Bill Clinton (US president, 1993–2001), Fernando Henrique Cardoso (Brazilian president, 1995–2003), and Tony Blair (UK prime minister, 1997–2007) also played important roles in the mainstreaming of neoliberal positions and policies, including those of free trade, privatized governance, and labor market/welfare reform. In institutional terms, the propagation and enforcement of neoliberal principles is closely linked, some would say organically, to the trio of Washington Consensus organizations – the World Bank, the International Monetary Fund (IMF), and the World Trade Organization (WTO) – together with an expansive system of free-market think tanks.

A slippery, contested, and polysemic word, “neoliberalism” is very much a critics’ term, one that is principally associated with the critical social sciences (where its explanatory reach and status are often questioned) and with activists and social justice movements working in opposition to corporate rule, financialized capitalism, and socioeconomic inequality. Neoliberalism has very little official currency per se. Politicians will only invoke the term, very occasionally, in order to distance themselves from it, as in the case of the declaration by Nicholas Sarkozy (French president, 2007–2012) that self-regulation, “laissez-faire,” and free-market deregulation...
were “finished” as a result of the Wall Street crash of 2008. (Only rarely will politicians even brand their opponents in these terms, as this would certainly beg the question of how far they, too, share in the neoliberal “consensus.”) In everyday usage, the word tends to be in much wider circulation in regions associated with wrenching experiences of structural adjustment, like Latin America, than it is in the supposed home of the Washington Consensus itself, the United States. As a result, neoliberalism represents, at the same time, a strong signifier in political and ideological terms, and an ambiguous and disputed term in both academic and popular discourse.

Positing a positive program for market freedoms – in the shape of an idealized, small-state, competitive order – neoliberalism is also very much defined by its antipathies, to socialism and statism, to public-sector solutions and government bureaucracies, to social collectivities like labor unions, and to programs of progressive redistribution. The prefix “neo” underscores the character of neoliberalism both as a successor to the influential strand of nineteenth-century political economy known as laissez-faire (against which it offers a sympathetic critique) and as an alternative to the various forms of state-centric government that rose to dominance after World War II, including Keynesian welfare states, state socialist systems, and developmental states (against which it offers a decidedly unsympathetic critique). In this respect, neoliberalism comes after, and refers back to, the laissez-faire vision of economic liberalism (based on freedom of exchange and trade, individualized rights, minimalist government, the spontaneous order of competitive markets, and an antipathy to state intervention beyond the enforcement of property rights), against which is advocated a more purposive but still constrained role for states as the active enablers, protectors, and propagators of markets, entrepreneurialism, and competitiveness. Neoliberalism also represents a historically “new” method for the regulation of capitalist economies, both as an ideological cum institutional alternative to Keynesian welfarist models and state-led approaches to development and, most fundamentally, as a Cold War antithesis to communism and socialism.

If the prefix “neo” serves a range of preformative (not to say performative) purposes, the stem “liberal” is not so easy to fix either. In the first instance, it should be taken to index an affinity to classical liberalism (freedom of thought and contract, equality before the law, the primacy of private property rights) and to liberal traditions in orthodox economics (deference to free markets, methodological individualism, and atomized competition; opposition to government interference), rather than liberal politics per se, at least in the received North American usage, where the label “liberal” usually applies to the politics of the center left and the traditions of New Deal and Great Society (welfare) liberalism. Even though some of the more conspicuous origins of neoliberal programs of government are to be found on the right of the political spectrum, with various conservative and “new right” traditions in North America and Western Europe, with authoritarian regimes in Latin America, and with the advocacy of pro-corporate positions more generally, the roots and reach of the project have long exceeded this. And, certainly in the contemporary context, neoliberal impulses, inclinations, and imperatives are yet more politically promiscuous, dominating the broad center (or mainstream) of the spectrum in many parts of the world, from the pragmatic left (notably, stemming from the Third Way movements of the 1990s) to various configurations of the corporate, establishment, populist, and radical right, while also fusing and interacting, in a usually more conditional and selective fashion, with an array of statist, authoritarian,
and neodevelopmentalist regimes, in China, Russia, India, Singapore, Brazil, and elsewhere.

In light of these complex geographies and genealogies, it should come as no surprise that “neoliberalism” can be a somewhat confounding and perplexing term – as a descriptor, as a signifier, and as an analytic. The influence of market-oriented politics may seem to be pervasive, and on some accounts even ubiquitous, yet no two manifestations of neoliberalism are the same. Since neoliberalism is a utopian vision, it cannot exist in “pure” form. Actually existing manifestations of neoliberalism must therefore always coexist with other social forms and political-economic tendencies – even where market rule is dominant or hegemonic. Its traces, furthermore, can only be found in complex hybrids, amid context-specific conditions, and in contradictory formations, where contradictions arise both from “internal” tendencies in neoliberal development (e.g., problems of market failure, or the experimental adaptation of strict privatization schemes into public–private partnerships and regulated privatizations) and from jarring relations with its coexistent others (e.g., conflicts with public sector unions and service users, differences with alternative conservative positions on issues like immigration).

Inevitably, then, idealized or shorthand representations of neoliberalism, as a generic condition of more market/less state governance, belie more complex and contradictory realities of hybrid coexistence, polymorphic state/market reorganization, and path-dependent restructuring. Nonetheless, the pervasive, “global” presence of neoliberalizing tendencies is commonly (mis)read as an indicator of incipient convergence or as a premonition of imminent conformity, when “total neoliberalism” remains a logical impossibility (and utopian fantasy) and as the reproduction of uneven geographical development is an endemic characteristic of “market rule” (Brenner, Peck, and Theodore 2010). Since all manifestations of neoliberalism are not only incomplete but also situated and context-specific (and not template reproductions, variants, or corruptions of, say, Thatcherism), it can only exist in the world in hybrid or “alloyed” form, while it also follows that neoliberalization functions as a tendency rather than an end state or equilibrium condition. Those that hold onto the term (and the concept) of neoliberalism, despite these challenges of conceptualization and operationalization, tend to do so not because it offers a blanket “explanation of everything,” but because it draws attention to complex connectivities, mutual relations, family resemblances, overlapping rationalities, repeating patterns, asymmetrical articulations, and dense relays between localized manifestations of variegated market rule.

A brief(er) historical geography of neoliberalism

The best-known introduction to the worlds of neoliberalism is David Harvey’s (2005) A Brief History of Neoliberalism. Underscoring the ideological integrity of the neoliberal project – for all its unevenly developed, actually existing forms – Harvey defines neoliberalism “in the first instance [as] a theory of political economic practices that proposes human well-being can be best advanced by liberating individual entrepreneurial freedoms and skills within an institutional framework characterized by strong private property rights, free markets, and free trade,” while at the same time recognizing that...
then, both Clinton and Blair could easily have reversed Nixon’s earlier statement and simply said, “We are all neoliberals now.” The uneven geographical development of neoliberalism, its frequently partial and lop-sided application from one state and social formation to another, testifies to the tentativeness of neoliberal solutions and the complex ways in which political forces, historical traditions, and existing institutional arrangements all shaped why and how the process of neoliberalization actually occurred. (Harvey 2005, 2, 13)

This is not to say that the idealized vision of neoliberalism had no part in these more haphazard, real-world experiences of neoliberalization, but rather that ideological (mis)representations of market freedom “primarily worked as a system of justification and legitimation for whatever needed to be done to achieve [the] goal” of restoring high(er) rates of profitability and returns to wealth, while securing the dominant positions of elites (Harvey 2005, 19).

The roots of neoliberalism as an explicit (and named) set of ideas go back to the middle of the twentieth century, and to two gatherings of prominent liberal intellectuals, first at the Colloque Walter Lippmann in Paris in 1938, and then at a Swiss mountaintop resort in 1947, the founding moment of the Mont Pèlerin Society. Hosted by the French philosopher Louis Rougier, the first meeting focused on the contributions of the American journalist and pioneering cold warrior Walter Lippmann, bringing together prominent representatives of various European liberal traditions, including Wilhelm Röpke and Alexander Rüstow of the German Ordoliberals (who maintained that a prerequisite for the effective functioning of free markets was a strong and well-organized state) and Austrian School adherents to a rather more classical, or paleoliberal, position, notably Ludwig von Mises and Friedrich Hayek. It was at this meeting that Rüstow made the case for a “new liberalism,” a purposeful reconstruction of the laissez-faire tradition, fit for contemporary conditions, and based on a principled repudiation of socialism, collectivism, and statism.

These plans to lay down the intellectual foundations for a new liberalism were disrupted by World War II. But the project was soon to be revived by Hayek in Mont Pèlerin, the Austrian serving as an interlocutor between an upstart cadre of Chicago-based economists, including a young Milton Friedman, who were beginning to define their “school” in opposition to the Keynesian mainstream (and the strongholds of New Deal economics on the East Coast of the United States), and the generally less buccaneering Europeans, including an Ordoliberal contingent with a hand in German reconstruction after the war. Echoing the line of argument sketched out in Hayek’s polemical defense of free-market reason, The Road to Serfdom (which had been a surprise bestseller at the end of the war and was promptly abridged by the Reader’s Digest magazine), these defenders of the liberal creed maintained that collective provisioning and state action, left unchecked, would lead inexorably to totalitarianism. As the founding statement of the Mont Pèlerin Society read:

The central values of civilization are in danger. Over large stretches of the Earth’s surface the essential conditions of human dignity and freedom have already disappeared. In others they are under constant menace from the development of current tendencies of policy … developments [that] have been fostered … by a decline of belief in private property and the competitive market; for without the diffused power and initiative associated with these institutions it is difficult to imagine a society in which freedom may be effectively preserved.

Believing that what is essentially an ideological movement [toward statism and collectivism] must be met by intellectual argument and the reassertion of valid ideals … The group does not aspire to conduct propaganda. It seeks to
As if to affirm their pessimistic reading of “tendencies of policy,” if not their alternative vision of a “free society,” the Mont Pèlerin group and its allies would spend most of the next quarter century at the intellectual margins and in an ideological wilderness. A cluster of Keynesian and developmentalist orthodoxies came to dominate not only the economics profession and the corridors of power, but also the “common sense” of mainstream political parties of all stripes across the Western capitalist countries, reflected in a bipartisan embrace of principles like full employment, social welfare entitlements, and the “mixed economy” of public and private provision, along with measures like price controls, countercyclical spending, and industrial policies. A line borrowed from Milton Friedman and frequently attributed to Republican president Richard Nixon, in the early 1970s, reluctantly acknowledged the status quo of the time: “We are all Keynesians now” (quoted in Harvey 2005, 13).

Far from idle during these years of isolation, however, those associated with what intellectual historians would later term the “neoliberal thought collective” (Mirowski and Plehwe 2009), even as they desisted from using the word “neoliberalism” itself sometime in the early 1950s, had been hard at work on the foundations of their “flexible credo.” This included a biting “monetarist” critique of inflationary tendencies in the Keynesian system, spearheaded by Milton Friedman and the Chicago School, which would later play a decisive role in the (very public) battles over the “stagflation” crises of the 1970s (the coincidence of high inflation and high unemployment, which exposed a mortal flaw in both Keynesian theory and practice). Alongside his achievements as a technical economist (recognized with the Nobel Prize in 1976, two years after Hayek, in the wake of which the field which has been dominated by Mont Pèlerinians and Chicago types), Friedman was by the late 1960s firmly established as one of the most prominent public economists in the United States. He was a formidable exponent of the argument–clinching sound bite, not just in the seminar room but also in the TV studio. A widely read magazine and newspaper columnist, Friedman also wrote some very successful mass market books, such as *Capitalism and Freedom* (1962) and *Free to Choose* (1980), the latter being serialized on public television in the United States in the year that ended with Ronald Reagan’s path-altering election.

Underlining the point that these were very much strategies of distribution and dissemination, rather than intellectual invention per se, the period between the mid-1950s and late 1970s was also a decisive one for what would grow into an expansive and tightly integrated ecosystem of free-market think tanks. The taproot of this system can be traced to a 1947 meeting between Hayek and Antony Fisher, a would-be conservative member of parliament turned businessman, who went on to devote both his life and personal fortune to the cause of economic liberty. On Hayek’s advice, Fisher put his financial and organizational skills to work in a pioneering effort to construct a quite new form of ideologically aligned think tank, at arm’s length from the universities but closely plugged into media circuits and a growing network of allies in the policymaking world. The first of the “Fisher think tanks” was the Institute of Economic Affairs (established in London in 1955), which fought a lonely fight for almost
two decades before the broader turn toward free-market ideas in the 1970s.

This decade saw the foundation of the most influential free-market think tanks, including the Heritage Foundation, established in Washington, DC in 1973 by conservative activists disillusioned with what they saw as Nixon’s excessively moderate policy positions. On the eve of President Reagan’s inauguration in 1981, Heritage would publish its Mandate for Leadership, anticipating both the general direction of (and specific measures within) the radical policy program of the new administration, into which several of its key staff members would promptly move. Rather more purist, the libertarian Cato Institute was founded by conservative billionaire Charles Koch in Wichita, Kansas, in 1974, later to relocate to San Francisco and then to Washington, DC. For his part, Antony Fisher was instrumental in founding the Fraser Institute (established in Vancouver, Canada, in 1974), the Centre for Independent Studies (established in Sydney, Australia, in 1976), the Manhattan Institute (established in New York City in 1978), the Pacific Research Institute (established in San Francisco, in 1979), the National Center for Policy Analysis (established in Dallas, Texas, in 1983), and the Atlas Economic Research Foundation (established in Washington, DC, in 1980), the latter with a mission to propagate free-market think tanks around the world, the extended network of which now consists of close to 500 affiliated organizations.

The underlying purpose, political orientation, and historical timing of the construction of this infrastructure of free-market think tanks might lend credence to a neat and tidy tale of linear, supply-driven ideational innovation. It is almost as if all that it took to win the battle of ideas was to reprocess the distilled spirit of Mont Pèlerin at the University of Chicago, and then to blend and dilute this in accordance with policymaking appetites and local tastes down the supply chain to the think-tank branch plants. The story, of course, more complicated, and it is significantly marked by fortuitous events and opportunistic openings. Constructing the think-tank delivery system may have been necessary, but on its own it was not sufficient. For his part, Hayek certainly believed that the new breed of free-market think tanks would have vital roles to play as “second-hand dealers in ideas”; since the basic principles of restrained government and liberalized markets were well understood (not to say, in these circles, beyond dispute), the downstream challenge was the continuous political and policy work of “permanent persuasion,” to borrow a Gramscian phrase (Peck 2010). As Antony Fisher liked to write in his fundraising letters, “the echo is more important than the message.”

Fisher was writing these letters, almost without exception, to wealthy business leaders and financiers. By the 1970s, some members of the business community in the United States were willing to write increasingly large checks to fund battles against burdensome taxes and bothersome regulations, against infringements of the corporate “right to manage,” and against creeping bureaucratization and social state expansionism, in order to defend the free enterprise system. But it had not always been so. After relocating to the United States in 1950, Hayek, Friedman, and their friends at the University of Chicago had struggled for years to raise adequate funds from the corporate sector and from private donors. The free-market cause had its supporters, but these were generally on the fringes of the business community. In the early 1970s, however, things began to change.

The turning of the tides was symbolized by the so-called Powell Memorandum of 1971, a confidential missive delivered to prominent members of the US Chamber of Commerce by Lewis F. Powell Jr, a corporate lawyer soon
to be appointed to the US Supreme Court. Entitled “Attack of American Free Enterprise System,” the memo traced the source of what was portrayed as a sustained and systemic assault on capitalist values and institutions to a “chorus of criticism … from the college campus, the pulpit, the media, and intellectual and literary journals, the arts and sciences, and from politicians,” in the face of which the business community had mistakenly adopted an implicit policy of appeasement. This had to change:

There should be no hesitation to attack the Naders, the Marcuses and others who openly seek destruction of the system. There should not be the slightest hesitation to press vigorously in all political arenas for support of the enterprise system. Nor should there be reluctance to penalize politically those who oppose it … The threat to the enterprise system is not merely a matter of economics. It also is a threat to individual freedom. (Powell 1971, 29–30, 32)

This is not to suggest that the Powell memo was the single smoking gun cause of the subsequent “strategic politicization” of the US business class, but it is widely credited as a germinal factor in (and portent of) the subsequent build-out of the free-market think-tank infrastructure; the ascendancy of organized assertion of shareholder values and interests; the massive expansion of business, financial, and “right channel” media; and the formation of conservative corporate policymaking operations like the American Legislative Exchange Council (ALEC), the influence of which did not fully come to light until well into the organization’s fourth decade of under-the-radar activities (see Block and Somers 2014; Peck 2014).

Few of these sparks would have led to any larger conflagrations, needless to say, absent the kindling of propitious cultural, social, and political-economic circumstances. In the pivotal decade of the 1970s, however, conditions began to favor, if not impel, the kind of concerted action advocated in the Powell memo and long anticipated by the free-market evangelists in Chicago. Sustained downward trends in corporate profitability, in labor productivity, and in the US share of world trade were contributing to an environment of macroeconomic insecurity, while a wide range of social movements (including organized labor; movements for civil, welfare, and women’s rights; and causes like environmental and consumer protection) had been achieving legislative traction as well as (re)shaping public opinion. The Powell memo had sounded the alarm that legislative and cultural incursions onto the corporate terrain had crossed the line to system-threatening status, arguing that the movement toward progressive taxation had gone too far and that increasing inflation would further erode the position of the holders of capital.

What some commentators would call the Great Inflation (the roots of which were traced to rising governmental and wage expenditures in the 1960s, not least as a result of the Vietnam War) was to prove catalytic for the subsequent turn toward neoliberal governance, the carrier of which was Friedman’s monetarist doctrine of tight control of the money supply (see Samuelson 2008). This turn echoed the way in which macroeconomic collapse and mass unemployment of the 1930s, the Great Depression, had provided the historical impetus (and the opportunity, if not necessity) for the rise of New Deal politics, for the establishment of the postwar social welfare settlement, as well as the foundations for the capital–labor accord of Fordism–Keynesianism (see Gordon, Edwards, and Reich 1982).

If these conditions described the path toward neoliberalism in the United States – culminating in the Reagan presidency and what would prove to be a long-run shift toward inflation control, tax restraint, and light-touch regulation – this is
not to say that they are replicated in each and every turn toward neoliberal politics and modes of governance. To be sure, there are recurring themes in neoliberal transformations around the world, such as preceding periods of aggravated macroeconomic stress (as reflected, for example, in currency, debt, or balance of payments crises), and the rise to power of “reformist” administrations (see Fourcade-Gourinchas and Babb 2002), but there is much more than analytical sophistry in the claim that neoliberal “transitions” are contextually and conjuncturally specific.

This point is underscored by what many regard as the “original” neoliberal counterrevolution, in Chile in the wake of the military coup of 1973. Under the dictatorial rule of General Augusto Pinochet, Chile was to become what Naomi Klein (2007, 53) has called, with requisite irony, a “laissez-faire laboratory” for Chicago School economics. True, there had been an experiment with neoliberal-style “shock treatment” in Germany, right after World War II, but Chile would become the most notorious case, and a forerunner to similarly severe applications of macroeconomic medicine in Central and Eastern Europe after the fall of the Berlin Wall in 1989 (see Yergin and Stanislaw 2002). In Chile, an elite cadre of US-trained economists — many of whom studied under Milton Friedman and his colleagues as part of an exchange scheme established in the late 1950s — were appointed by Pinochet to drive an unprecedented program of monetary restraint, privatization, deregulation, union suppression, trade liberalization, and tax reform aimed at taming runaway inflation and restoring economic competitiveness.

While Friedman and Hayek defended the economic record of the “Chicago Boys,” their enduring notoriety stems from the perverse political demonstration effect of Chile’s neoliberal turn: this program of economic liberalization was prosecuted through state violence, authoritarian governance, and repression of democratic institutions, undercutting the claim to an irreducible link between political and economic “freedom.” The embrace of neoliberalized modes of economic regulation would later be witnessed across a wide range of political regimes, including robust democracies and “reforming” state socialist systems. In some cases, neoliberal reformers can justly claim an electoral mandate, but there is increasing evidence that deepening neoliberalization is associated less with democratization and empowerment than with a long-run drift toward insider dealing, money politics, and technocratic management. In turn, this reflects established tendencies in neoliberal governance for upward expropriation over downward redistribution, for private gain over the common good, for shareholder value over social benefit, for speculative growth over sustainable development, and so forth.

Arguably, nowhere is the discrepancy between neoliberal economic reforms and democratic consent greater than in the case of structural adjustment programs (SAPs) managed by multilateral agencies like the IMF and World Bank, with the support of the US Treasury department. Beginning in the 1980s, these Washington Consensus organizations started to impose increasingly strict “conditionalities” on loans advanced to middle- and low-income countries in the context of macroeconomic (and especially debt) crises, conditionalities that essentially mandated the introduction of a range of neoliberal policy measures, including privatization, deregulation, financial liberalization, and public spending reductions (see Peet 2003). These policy prescriptions were later summarized by John Williamson (1990), an economist with close links to the IMF and the World Bank, in a manner that would pass into neoliberal lore:
1 fiscal policy discipline, focused on debt reduction;
2 shifting public spending from “subsidies” to pro-growth investments in skills and infrastructure;
3 tax restraint;
4 control of interest rates, which should be market-determined;
5 competitive exchange rates;
6 trade liberalization and the removal of tariff protections;
7 liberalization of inward foreign direct investment;
8 privatization of state enterprises;
9 deregulation;
10 legal security for property rights.

There has been a tendency, since the late 1990s, to apply structural adjustment measures in a less brazenly unilateral manner, with an increased emphasis on “good governance” and “pro-poor” policymaking, in the context of what has sometimes been dubbed the “post-Washington Consensus” – not least as a result of social protests triggered by earlier rounds of SAPs and by the (at best) mixed record of these programs themselves (see Craig and Porter 2006). There has been a resumption of stricter forms of structural adjustment, however, in the period since the global financial crisis of 2008. New rounds of “austerity” measures have been imposed not only on developing countries, but across the eurozone, especially in the Mediterranean states, and in bankrupt cities like Detroit – in some cases involving the formal or de facto suspension of established democratic procedures (Laskos and Tsakalotos 2013; Peck 2014).

Analytics of neoliberalization

The invocation of neoliberalism and its processual variant, neoliberalization, as analytical devices, mostly in the critical social sciences, exhibits a historical geography of its own. This overlaps with, but in other ways is quite distinct from, the evolution of its actually existing forms. Early explorations of the explanatory and conceptual (as well as political) status of neoliberalism can be found in Latin American development studies (see Colclough and Manor 1991), in the neo-Foucauldian governmentality approach (see Barry, Osborne, and Rose 1996), and in neo-Marxist debates around Thatcherism (see Jessop et al. 1988), although there were few connections between these islands of theoretical practice. In geography, the encounter with neoliberalism was principally associated with work in regulation theory (see Peck and Tickell 1994). The initial premise here was to question, in the British context at least, the status of the Thatcherite form of neoliberalism as a viable “successor” to Keynesian welfarism as a mode of social regulation in its own right, and as an effective enabler of an enduring pattern of post-Fordist macroeconomic development. Instead, the interpretation in these early regulationist treatments, which were largely concerned with the macro form(s) and consequences of neoliberalization, was that Thatcherism was more of a creature of the protracted British crisis than any meaningful guide to its resolution, being freighted with unmanageable contradictions and disabling externalities.

Regulation theory rejects economically functionalist conceptions of modes of social regulation (like Keynesianism, developmental statism, and would-be successors like neoliberalism), which are seen to possess relative autonomy from the needs of capital. Yet capitalism (and markets) cannot exist independently of social and institutional supports of different kinds. Markets and institutions are therefore found in (various states of) contradictory coexistence. It follows that the search for new institutional
NEOLIBERALISM

fixes (e.g., along the path of neoliberalization) is marked by a wide range of failed experiments, crisis-managing compromises, and “chance discoveries,” as well as more deliberative projects, putative models, and power plays. Over time, however, it became increasingly clear that various stripes of Anglo-American neoliberalization were beginning to evolve synergistically with deepening financialization, as they also became ever more profoundly entrenched with elite interests. These would soon be shaping the rules of the game more widely, even as this game was being played differently in different places.

Meanwhile, somewhat overlapping understandings of “neoliberalism,” as a synonym for corporate rule and free-market globalization, were emerging in various activist circles, beginning with the Zapatista uprising in Chiapas, Mexico, in 1994, resonating through the anti-WTO protests in Seattle in 1999, and provisionally uniting (albeit in opposition) the diverse array of social movement forces assembled at the World Social Forums, initially in Porto Alegre, Brazil, in 2001, as the dialectical other to the mobilizing slogans “Another World Is Possible” and “One No, Many Yeses” (see Fisher and Ponniah 2003).

This newfound political currency coincided with the recognition that, for many critical social scientists, the term “neoliberalism” had utility as a means to (re)politicize orthodox understandings of globalization, which had been ascendant during the 1990s, following the fall of the Berlin Wall and hubristic announcements of an “end of history” victory for US-style capitalism. (In mainstream usage, “globalization” was often portrayed as an uncontrollable process, like the weather, and somehow above politics itself.) Different conceptions of neoliberalism were soon being widely invoked, especially in geography and urban studies (see Brenner and Theodore 2002), in radical political economy and cultural studies, and rather more tentatively in sociology and anthropology, although still hardly at all in the more orthodox fields of political science or economics. Something of a step change occurred in the wake of the global economic crisis of 2008, however, which, even as it vividly and publicly discredited the model of “deregulated” finance, soberingly affirmed the grip of neoliberalism’s political hegemony in the uncomprising form of social state retrenchment and austerity programming (Brenner, Peck, and Theodore 2010; Hall, Massey, and Rustin 2013).

Notwithstanding the increasingly close attention that has been paid to the form and consequences of neoliberalization, it is notable that the basic definition, the explanatory status, and in some quarters even the very existence of neoliberalism all remain, to varying degrees, subjects of contention. In political rhetoric, the term retains polemical connotations. It is associated with a range of more precise meanings in the social science literature, but these often remain matters of dispute. Some of the theoretical currents include a Marxist depiction of neoliberalism as a (new form of) ruling-class hegemony tied to deepening financialization, as an ideological offensive or “revolution from above,” and in a more general sense as a synonym for the prevailing historical form of capitalism (see Harvey 2005); a (neo-)Foucauldian strand of governmentality studies, in which neoliberal impulses take a much more dispersed and socially complex form, as they are implicated in the enrolment and (re)constitution of enterprising and “responsibilized” subjects (see Barry, Osborne, and Rose 1996); and, working across and between these, various approaches to neoliberalism and neoliberalization as a variegated policy complex, as a restructuring paradigm or regulatory front, and as a prevailing rationality of government (see Brenner, Peck, and Theodore 2010).
In their different ways, all of these critical theorizations of neoliberalism must contend with the many discrepancies and slippages between, on the one hand, the reductionist and idealized discourse of neoliberalism (small state + free economy), which posits an essentially zero-sum relation between governmental withdrawal and the animation of (naturalized) markets, and, on the other hand, the much more prosaic, compromised, and contradictory forms (the plural being deliberate here) exhibited by actually existing programs of neoliberalization. Rarely, if ever, anything resembling the (constitutive) fairy tales of shrinking states and spontaneous markets, the wide array of actually existing neoliberalisms tends to be characterized by, in addition its idiosyncratic features, a tendency for complex reregulation rather than simple deregulation, shaped by recurring crises, policy failures, and opportunistic experiments; a significant degree of institutionalized path dependency, reflecting the specificities of those antithetical institutions targeted, and variably rolled back in neoliberal reforms, as well as the particularities of those rolled out in evolving programs of market-oriented reconstruction; and new forms of uneven geographical development, rather than convergence or standardization toward a generic or template form.

Disputes remain, however, around the utility and scope of neoliberalism as an explanatory device. Those skeptical of the term(inology) typically point to its normalizing effects, as an ideologically coherent “big story,” drawing attention instead to the diversity and vitality of lifeworlds outside the ostensibly all-encompassing ambit of neoliberalization, to local exceptions and alternatives, and to the irreducible social surplus beyond the instrumental reach of market rule. Those attentive to the variegated and sprawling form(s) of neoliberalization, on the other hand, tend to see recurring patterns, family resemblances, and constitutive connections, even as they recognize that neoliberalism’s totalizing tendencies will always be thwarted and exceeded. These different ways of “seeing” neoliberalism, or not, arguably reveal as much about theoretical predispositions as they do about social worlds in change. It is almost as if an analytical version of the Rorschach test is at work: some will focus on the menacing black shapes, while the eyes of others will be drawn to the open spaces beyond.

SEE ALSO: Globalization; Governance and development; Governmentality; Power and development; Public policy; Regulation/deregulation; Restructuring; Social movements; States and development

References

NEOLIBERALISM


Netherlands: Koninklijk Nederlands Aardrijkskundig Genootschap (KNAG) (Royal Dutch Geographical Society)

Founded: 1873
Location of headquarters: Utrecht
Website: www.knag.nl
Membership: 3000 (as of 2013)
President: Eelko Postma
Contact: e.postma@knag.nl

Description and purpose

The Royal Dutch Geographical Society is the professional association of academic geographers, geography teachers, geography students, and professional geographers in the Netherlands with about 3000 members. The main tasks of the society are the promotion of geography in general and the improvement of its position in primary, secondary, and higher education as well as in scientific and applied research. It also offers opportunity for discussion and activities among geographers in professions in commercial research, education, management, policy, and so forth. Members come from both the public and the private sector and they operate at national as well as international level.

Journals or major publication series

Geografie. www.geografie.nl

Current activities or projects

The society hosts the KNAG Onderwijsdag where about 800 geography teachers attend this annual national congress on geography education. The society gives pupils between the ages of 10 and 15 the opportunity to work with soil, water, and digital map technology while visiting businesses and government during GeoWeek. The society has educational programs that offer advice on curriculum and examination. The society also plans several excursions every year in the Netherlands and abroad, for example, northern Germany and Indonesia.

The society copublishes several monumental atlases on historical cartography, for example, the Grote Atlas van Nederland 1930–1950 (Comprehensive atlas of the Netherlands 1930–1950), Grote Atlas van Nederlands Oost-Indische compagnie (Comprehensive atlas of the Dutch East Indian Company) in seven volumes, and Grote Atlas van de West-Indische compagnie (Comprehensive atlas of the West Indian Company) in two volumes.

Brief history

The Royal Dutch Geographical Society (KNAG) was founded in 1873. Initially, the main aim of the society was to increase the scientific knowledge of the world and the Earth. Expeditions were organized into unknown territories. The members of the expeditions also had an eye on the commercial possibilities of the areas they explored. Artifacts from these expeditions – like authentic maps and original expedition equipment – are lent to various museums in...
the Netherlands where they are put on display. In total, around forty expeditions have been organized under the auspices of the society. All the map making, production of inventories, and the collection of samples has resulted in an impressive cultural heritage.

When the need to explore dwindled, the society turned its focus to geographers and geography in the Netherlands. Modern-day KNAG stresses the distinctive and useful characteristics of geography as an academic discipline, a school subject, and a profession. In addition, the society provides services for its members and functions as a physical and digital meeting place to share ideas, knowledge, and information.

Submitted by Eelko Postma
Network analysis

Kevin M. Curtin  
George Mason University, USA

Network analysis is one of the earliest and most persistent subdisciplines within the science of geography, and it has an extraordinarily broad reach across both theoretical and applied research areas. The networks analyzed in geography can be either natural (e.g., river networks) or human in origin (e.g., road networks), and they can consist of either physical structures or conceptual constructs. Networks represent a spatial domain over which an enormous variety of activities take place, from daily transportation to message communication to the flow of goods, services, and ideas. Activities that occur on geographic networks employ a diverse set of modes: cars, trucks, trains, airplanes, pedestrian traffic, and digital messages, among many others. Although the differences among the variety of geographic networks that can be modeled allow for the great diversity of research and applied study, it is the similarities in the network structure that provide the basis for analysis.

The history of network analysis in geography

Even the earliest practitioners of geography recognized the importance of networks in their descriptions of the world around them. Many of the earliest examples of cartography contained representations of rivers, canals, or pathways between and along important features. The ancient Greeks provided descriptions of network geographic features including shipping lanes, and the historical record contains maritime policies designed to protect them. It is well documented that Roman roads were considered vital to the building and maintenance of the empire.

The age of exploration brought with it a revolution in network analysis in geography. In the early part of this period long-distance transportation was dominated by sea travel, and the knowledge of transportation routes and the network of ports for supplies constituted a critical and highly valuable set of knowledge. Maps termed Portolan Charts showed the sea routes that could be exploited from and to a number of origins and destinations. In the later period, a tradition began of constructing series of maps that showed descriptions of routes travelled by explorers, frequently accompanied by directions that later travelers could follow, and pictorial or text descriptions of the sites they would find along the way. This tradition in travel cartography has persisted to this day.

With the advent of the modern university system, geography was recognized as a “mother discipline”; that is, one that is so intellectually broad that it tends to spawn other disciplines as they develop significant bodies of thought in their own right. This process held true for network analysis, as seen in the primary influence of transportation costs on land rent in Von Thunen’s economic location model (Von Thunen 1966). The distorting factors of transportation on the theoretical isotropic plane on which Christaller’s central place theory was posited resulted in a separate transportation principle for areas where the most efficient transport network was the primary locational consideration. Transportation
science itself became a separate discipline with research foci as broad as transportation network construction, to theories of traffic flow and queuing, to the science of urban travel demand modeling using transport networks as the domain for analysis.

The analysis of the movement of goods and services across networks took on new importance during World War II, when the field of military operations research demanded that transport networks be used to efficiently transport supplies and personnel based on operational objectives. This logistical approach to systems management entered the private sector after the war and became more broadly known as operations research. When an operations research problem contains location elements, including the location of objects on networks, or the locations of network elements themselves, this bridges operations research with geography in a subdiscipline termed location science. Location science remains a flourishing research subdiscipline of geography to this day.

During the period known as the quantitative revolution in geography (1950s and 1960s), network analysis became both a focus of research and a supporting concept in many of the most influential theoretical advances in geography of the time. Networks were considered the conduits of interaction in the area of interaction (or gravity) modeling. Diffusion models depended, in some cases, on hierarchical networks of places through which the goods, people, or ideas were moving. Investigations into the nature of the structures of networks and how networks can grow or evolve in different geographic contexts began to appear during this period (Haggett and Chorley 1970), and were tied particularly to economic analyses of regions. The related field of regional science embraced networks as the source of paths across which migrants traveled (and delineated the routes they chose to take on their migration), the conduits for commodity flows, and the connecting threads in interregional analysis (Isard 1960).

During this period transportation science itself developed into a discipline with strong ties to geography. This field encompasses elements of civil engineering (design of transportation network features), operations research (network flow modeling), and urban planning and design (transportation system integration, placement, and design), among others. The element of transportation most closely associated with network geography is that of urban travel demand modeling (Oppenheim 1995). The classic form of this model has four sequential steps, most with fundamentally geographic foci. The trip generation step determines the spatial origins of trip demands, usually through a combination of travel surveys and regression analysis using spatial demographic information as the independent variables. The trip distribution step employs variants of classic interaction modeling to determine likely trip destinations. The modal split step allows analysts to experiment with demand for the road network versus public transport networks based on costs and utility. Finally, the traffic assignment step associates trips with particular paths through the available transport networks. More broadly, transportation science is concerned with transport network design, routing across transport networks, queuing of objects traveling across transport networks, and other related transportation issues. However, since it is widely recognized that transportation applications are not the only research area that benefits from the flexibility and power of network structures, a broader science of network analysis has developed.
The science of network analysis

Network analysis in geography developed in concert with the development and growth of many other scientific disciplines in the modern university system. In particular, several subdisciplines of mathematics provided a quantitative theoretical basis for network analysis to complement the growing body of pure geographic theory. Those subdisciplines – graph theory and topology – examine the properties of connected sets of edges and vertices (in the case of graph theory), and connected spaces more generally (in the case of topology). Within these contexts networks or graphs (the terms are synonymous here) exist as a general class of spatial domain for geographic analysis. Thus, the properties of any network, regardless of the particular application or function, can be described, measured, and proven. Techniques for each of these levels of network analysis have developed over time.

Describing the properties of networks requires describing the elements that make up those networks (edges and vertices), describing the relationships among those elements, and describing the nature of the network as a whole. Among the most intuitive descriptive measures of networks are simple counts of the number of edges and vertices, and measures of the attributes of those network elements, such as the maximum, minimum, or total length of the edges, or the count of population or other demand that is presumed to exist at the vertices. These simple descriptive measures allow the most basic comparison of the size of the network under examination.

A very different means of describing networks consists of categorizing networks into idealized types based on the relative placement of features, the regular pattern of features in the network, or the topological properties of the network. Examples include the Manhattan network type, which consists of edges crossing at right angles creating rectangular “blocks.” As the name suggests, this network type (or an approximation of it) is frequently seen in more modern urban street networks. Distance measures across a network of this type are termed “Manhattan distances” and for some applications – such as pedestrian travel cost in a city – are perceived as much more realistic than other forms, such as Euclidean or “as the crow flies” measures. Hub and spoke networks, with the radial extensions that the name suggests, are a network type that has been widely used in distribution networks, including in some airline routing applications. When a network has branching edges, but no cycles, it falls in the category of tree networks. River networks and many types of hierarchical process networks are frequently modeled as tree networks. The “Bus” network type is a special case of a tree network consisting of a single main line with branches off to either side. In contrast to the name, this network type is frequently used in utility applications (electricity, water, and gas distribution, or telephone and cable television lines). Other types that are often used in more theoretical network analysis work include line and ring networks, or complete networks (where every vertex connects to every other vertex with a unique edge). Often these special cases of network models have distinctive properties that can be proven, and can be relied upon as the networks provide a platform on which network analysis takes place.

Another basic measure that considers the interaction of edges and vertices is the degree of vertices – simply a count of the number of edges incident to the vertex. The maximum, minimum, or average degree of the vertices in the network can then be recorded. These give some measures of the complexity of the interactions among the network elements. A series of more complex indices has been developed to specify
the level of connectedness of network elements (Kansky 1963). The beta index is simply the ratio of edges to vertices, where – intuitively – a larger number of edges associated with the same number of vertices represents a more highly connected network. The gamma and alpha indices represent a ratio of the number of links in a network to the largest number of possible links in that network (if it were a completely connected network), and the number of cycles compared to the maximum number of cycles, respectively. The maximum number of links can be determined based on a well-known proof, while the maximum number of cycles in a graph must be estimated.

These last network measures depend on graph theoretic proofs of the properties of the network in order to make comparisons across different networks. Further proofs of the properties of networks lead to a more robust set of network measures. Perhaps the most widely used proof of a network property is Dijkstra’s algorithm for shortest paths across a network (Dijkstra 1959). This efficient algorithm allows the guaranteed shortest path between any two points in a network to be computed, and variants of this algorithm allow all shortest paths between all vertices in a network to be computed. This, in turn, allows the determination of the diameter of a graph, defined as the longest shortest path between any two nodes in the network. This is a measure of the spatial spread of the network. As the Euclidean distance between any two nodes of the network is easily computed algebraically, the shortest path allows for a comparison of the network shortest path with the Euclidean shortest path between any two points. The ratio of the latter to the former is termed the detour index, and represents the efficiency of the network in overcoming distance. These are only a few of the most common network measures; new ways of measuring networks and their properties are entering the body of knowledge on a continual basis, from across a range of disciplines.

The measures discussed above allow analysis of a network to be performed but other research questions may require analysis on a network. There are many procedures that can allow routes across a network, paths among elements of the network, or optimal activities across a network to be determined. A limited sample of such procedures includes minimum spanning tree algorithms that find a least-cost tree network connecting all vertices in a network, maximal flow algorithms that determine the ways in which the largest possible amount of goods can be moved through a capacitated network (Ford and Fulkerson 1956), and methods to solve the transportation problem on networks where a set of supplies must satisfy a set of demands across the network with minimal system cost. The development of additional solution procedures for network analysis has moved forward rapidly as the technology of geographic information systems (GIS) has matured.

Network analysis in geographic information systems

As with much of the discipline of geography, network analysis has been revolutionized by the advent and growth of GIS. The extent to which network analysis has been integral in this growth stems from a combination of research interest in the community and fortuitous good luck. The latter element is an outcome of the needs for linear features and boundaries in the earliest GIS applications, and therefore their instantiation in the earliest GIS data models.

The earliest computer-based systems for automated cartography attempted to mimic the role of the cartographer in generating pictorial representations of linear features. Using a
series of consecutive symbols from a dot matrix printer, and later sets of individual lines drawn independently on a line plotter, the analyst could create maps showing a range of linear features. The structure of these independent yet adjacent geographic features came to be known as the “spaghetti” data model, as there were no formal connections between the features. In this data model each feature was stored individually, usually with an identifier and a list of coordinates that defined its shape. Because this data model does not preserve topological properties such as incidence or connectivity, it cannot readily be used for network analysis. The spaghetti data model persisted in computer-aided design software for several decades, where the shape of features was of primary importance for the generation of diagrams and drawings.

Although the idea of using topological information to store features developed independently in several places, it was the acceptance of a topological data model by the U.S. Census Bureau that indirectly led the way for significant advances in network analysis. Although generating maps for field work and for the location of addresses to which questionnaires could be mailed was of importance, the Census Bureau was primarily interested in a means of generating well-defined polygons that would represent the Census Tracts, Census Blocks, Counties, and other enumeration areas with which even casual users of GIS are now familiar. In order to meet this need researchers at the US Census Bureau developed a topological data model named the dual incidence matrix encoding (DIME) model (Cooke 1998). “Dual incidence” refers to the storage of topological information among nodes (which nodes are incident creating lines) and among lines (which lines are incident creating polygons). The DIME databases evolved into the topologically integrated geographic encoding and referencing (TIGER) files, and this data model for a time became the standard for vector representations in GIS.

While the goal of the Census Bureau was not to generate network data, the act of digitizing census block boundaries – many of which were roads, railroads, or streams – allowed for the rapid creation of network databases under the direction of (and with the resources of) a federal government agency. At a time when data for use in a GIS were not readily available and required time consuming and costly manual digitizing, the creation of these network databases was a great advantage for researchers pursuing network analysis. Shortly thereafter some of the analytical measures that had developed over many decades (described previously) began to appear as functions in GIS software.

However, because this data model was designed to support a set of hierarchical polygons it constrained network analysis techniques. More specifically, in order to maintain complete polygonal coverage, the data model had to enforce planarity. Planar networks are those that have no edges intersecting without a vertex at that location. In contrast, in real-world transportation networks it is common to have edges which cross in planar space without any connection between the edges. These include overpass–underpass crossings, bridge crossings, and even at-grade railroad and road crossings. No connectivity exists between the edges at the coordinates of the crossing. This limitation has necessitated the development of network-specific data models in GIS. These models variously allow for nonplanar intersections, for modeling flow across the network, and for identifying specialized features such as junctions or turn elements that restrict movement across the network. It has been found that networks can be more efficiently stored in structures that differ from standard relational database structures and, therefore, over time network data models have come to rely on
structures such as the forward star data structure (Evans and Minieka 1992).

Advances in network analysis in geography

With the full integration of network models into GIS the advance of network analysis in geography has seen a rapid move forward. These advances have taken the form of networks as the foundation for location referencing systems, networks as the basis for advanced location modeling, and networks as the spatial domain for new spatial statistical techniques.

A fundamental tenet of geographic analysis writ large is that a well-defined spatial domain must be chosen prior to analysis with some justification, and there must be a means of determining locations across that spatial domain. When the analysis consists only of specifying locations, the domain is frequently an ellipsoid representing the Earth, and the means of identifying locations is through latitude and longitude coordinates. When computations of spatial properties such as distance, area, or direction are part of the analysis, good practice requires that a surface be chosen onto which locations are projected, where it is known what properties are preserved or distorted through that projection. Further, a planar coordinate system is then used to determine locations on that flattened surface. However, it has been accepted in some application areas – particularly transportation science – that a flat planar space is not the spatial domain over which some phenomena occur; rather the network itself is the operational space. In these cases there must be a method of specifying locations across the network. The process of linear referencing provides these locations using the network as the spatial domain (Curtin, Nicoara, and Arifin 2007). A widely recognized application of linear referencing is the mile marker system along US highways. The mile markers represent a distance along a network feature from a well-defined starting point, just as a traditional coordinate represents a distance east and north of a well-defined origin in a traditional coordinate system. There are several benefits to using linear referencing for network analysis, including intuitive locations for work in the field, the ability to define features (sometimes termed routes) that are not limited by the storage parameters of the underlying spatial data model, and the ability to maintain many sets of network attributes using the network only as the underlying datum rather than as the storage set for attribute information.

The acceptance that network phenomena must be analyzed in the network spatial domain has led to significant advances in the realm of spatial statistics. It is now recognized that the spatial statistical determinations of density, clustering and dispersion, and interpolation (among others) of network phenomena cannot be treated as if they operate in the plane. Instead, a related but distinct subdiscipline with spatial statistics – network spatial statistics – has developed; this permits the justifiable analysis of network-based events (Okabe, Okunuki, and Shiode 2006; Okabe, Yomono, and Kitamura 1995). Examples of such phenomena include traffic accidents along roads, pipe breaks along water mains, or pollution samples along river networks.

Just as the prevalence of network data models in GIS has encouraged advances in location referencing and in network spatial statistics, it has encouraged the development of widely available tools for more advanced location science methods. In addition to the optimal routing procedures that have long been integrated, heuristic solutions for highly combinatorially complex decision problems are widely available. These include solution tools for the traveling salesman
problem, the vehicle routing problem (which is effectively a series of traveling salesman problems solved simultaneously), the P-Median problem (which seeks the optimal locations for P facilities on the network such that total cost between demand and the nearest facility is minimized), and the maximal covering location problem (which seeks to locate facilities in such a way as to cover as much demand as possible within a specified service distance across the network). When optimal solutions are needed, specialized software exists to generate such solutions, dependent on problem size. While only a small sample of the solution procedures that have been developed in network analysis have made their way into off-the-shelf GIS software, more are migrating to that user-friendly environment all the time.

The future of network analysis in geography

There is every reason to believe that the future advances in network analysis will continue to reinforce the importance of this subdiscipline within the science of geography. By examining current literature the increasing interest in network analysis across a broad range of applications areas can be seen. For example, a rapidly growing interest in social network analysis has led to contributions including additional network measures of closeness (average distance between nodes), betweenness (number of paths that go through nodes), and centrality (position within network neighborhoods) — all of which are fundamentally geographic concepts. Increased interest is being seen in the development of techniques for network-based spatiotemporal analysis, with the recognition that patterns that suggest an underlying process in both time and space are different from those that appear in only one dimension alone (Eckley and Curtin 2013).

It is likely that the past research on static measures of networks and deterministic analysis on networks will be complemented by more research into dynamic movement through and across networks. While network flow models have existed in the literature for decades, additional means of describing, visualizing, and measuring flow may come from work regarding the simulation of movement across networks. Advances in tracking objects as they move through space will likely support this area of research. There are also avenues for innovation in network analysis related to behavioral geography and the movement of teams across networks where they are seeking some group goal. As these problems become more computationally complex, the advances in computing power (e.g., massively parallel computing, cloud computing) may allow researchers to pursue problem instances that were not previously possible.

In summary, network analysis in geography is a major research area, with a strong theoretical history, a solid mathematical foundation, and wide range of methodological approaches. It is an area with broad interest across disciplines. Perhaps most importantly it is a vibrant and dynamic research area with both growth and refinement of existing methods, and the development of new methods over time. It is likely to remain a significant component of the larger area of geographic analysis for the foreseeable future.

SEE ALSO: Behavioral geography; Central place theory; Cluster detection; Data model, moving objects; Data structure, vector; Economic geography; Geocomputation; Geographic information science; Geographic information system; GIS: history; GIS for transportation; Graph theory; Location-allocation analysis; Location-allocation models; Logistics; Parallel computing; Point pattern analysis; Quantitative methodologies;
NETWORK ANALYSIS

Regional science; Road transport; Routing and navigation; Spatial analysis; Spatial database; Spatial interaction; Spatial optimization; Spatiotemporal analysis; Topological relations; Transport geography; Transport networks; Transportation planning

References


Networks, social capital, and development

Martin Hess
University of Manchester, UK

Development is a moving target. The 1980s saw a crisis in development studies, often referred to as the “development impasse,” when both development theories and real-world developments posed serious challenges to existing models of development at the time. Today, the latest indicators underpinning the 2015 United Nations Millennium Development Goals show that while progress has been made in tackling extreme poverty and a range of other areas of development, any improvements are highly uneven socially and geographically. Such uneven development outcomes occur against a backdrop of neoliberal economic policies, the roll-back of the state and the rise of increasingly fragmented global production networks in what Manuel Castells (2003) calls the network society. As conventional, top-down and state-led forms of development have failed on many accounts, albeit with some notable exceptions especially in East Asia, new forms of alternative development strategies and networked forms of development activism (nongovernmental organizations (NGOs), grass roots movements) have come to the fore. Since the 1990s, this has brought ideas of social capital into the spotlight both in academia and policy circles, often viewed as a development panacea and championed by global organizations such as the World Bank (Potter et al. 2012).

Critics, however, argued that the increasing focus on networks and social capital only helped to obscure some fundamental insights from political economy with regard to poverty and inequality and lead to a de-politicization of the development discourse and practice. Consequently, social capital as a concept has received a very mixed reception in human geography and cognate disciplines ever since its inception by Pierre Bourdieu and its subsequent formulations by James Coleman and Robert Putnam, with the latter becoming especially influential in framing the debate. The next sections will in turn address the role of networks as organizing principle underlying economic, social, and spatial development; Bourdieu’s ideas of social capital and the reproduction of power relations; Putnam’s concept of social capital, trust, and collective action; and the uptake of social capital in development discourse and policy, before offering some concluding thoughts.

Networks, geography, and development

The notion of networks has become almost ubiquitous, in academia as much as in public discourse and the media. Today, it seems an almost universal mode of social, political, and economic organization, from global sub-contracting networks and global production networks to terrorist networks like Al Qaida and Islamic State (IS), from transnational migrant networks and local ethnic business clusters to community unionism, activist networks, and NGOs. Networks are the primary structures and organizational forms through which the flows of goods, services, people, capital, and information
are organized and orchestrated (Grabher 2009). In human geography, since the 1990s, the rise of networks as analytical focus, ontological basis, and methodological foundation for research has found its expression in the so-called relational turn. This approach views networks neither purely as structures nor organizational forms situated between atomistic individual action and hierarchies, but as relational processes that produce observable patterns and outcomes in society and economy (Dicken et al. 2001). Network thinking and analysis has since then gained currency in a variety of subdisciplines, not least economic geography and, to a lesser extent, development geography.

Interest in networks was triggered in no small part by Mark Granovetter’s (1985) seminal article emphasizing the social embeddedness of economic action. Stressing the importance of concrete personal relations and networks of such relations for the generation of trust between actors, he laid down a foundation for the analysis of networks, their governance, and associated economic and developmental outcomes (Hess 2004). All networks share some fundamental characteristics. First, they are based on reciprocity. Any exchange processes (of materials, money, information, etc.) take place within the context of mutual relationships and network embeddedness is determined by the architecture, durability, and stability of these reciprocal relationships. Second, and in contrast to other forms of organization like markets or hierarchies, networks are held together by a multiplicity of interpersonal and organizational ties of different strength, from strong and durable connections to what Granovetter calls “weak ties” allowing for greater flexibility and openness thus avoiding the danger of lock-in and network inertia. Third, actors in networks are mutually interdependent, and such interdependency impacts on the stability and durability of collaboration between network agents. Fourth, networks are shot through with (asymmetrical) power relations that govern them and shape the economic and developmental outcomes of network activities for various actors and the places they inhabit.

Network formation takes place at a variety of scales, from the local to the global, connecting some places while bypassing others. Contemporary networks are made from a complex addition, crossing and entanglement of economic activities, of social and intellectual communities with considerable developmental consequences for the places in which global production networks “touch down” or local networks emerge in situ. There is now a plethora of literature that addresses regional development processes through the lens of networked forms of social and economic organization. Studies of economically successful regions, like the so-called Third Italy, Silicon Valley in California, or clusters of economic activity in the German state of Baden-Württemberg, identified formal and informal network building as main drivers of development. Dense networks at the local scale have been referred to as producing “institutional thickness,” creating the conditions for sustained economic growth, innovation, and development. Institutional thickness refers to a strong institutional presence, high levels of interaction in various political, social, and economic networks, defined governance structures of domination and patterns of coalition, and the mutual awareness of being involved in a common “regional” enterprise. Such conditions have been identified as key ingredients in creative milieus or innovative/learning regions, and, more recently, the thickness concept has been conceptually enhanced to theorize social innovation and community development (Moulaert and Nussbaumer 2006). Moreover, attempts to recreate such geographical clusters and emulate their success have become a policy
mantra in many regions across the Global North and the Global South. However, there also has been a growing recognition that such regions do not exist in isolation and developmental outcomes are also increasingly dependent on transregional networks and the nature of connections with global production networks (Coe and Hess 2011). In other words, development in any place is determined by a variety of local and translocal networks and the associated flows of people (through migration), materials, capital, and knowledge.

The distribution of power among the various agents constituting social and economic networks is clearly paramount for developmental outcomes and the distribution of any benefits from network activities. However, it has been argued that the network concept and metaphor provokes a one-dimensional understanding of power (Prey 2012). Referring to the work of Castells, network thinking has led to an emphasis on exclusion and inclusion as a fundamental form of exercising power. Indeed, much of the recent literature on economic and social upgrading for firms and regions has highlighted the importance of engaging with global networks in order to capture value as a means for development. But such a focus on inclusion and exclusion – following Prey’s argument – obscures the problem of exploitation, a still prominent feature of many global and local networks. What is needed, therefore, is a better understanding of the relations between processes of exclusion and exploitation, as well as inclusion and privilege. To this end, the following section turns to one conceptualization of social capital as developed by Pierre Bourdieu.

Social capital, class, and habitus

Societies and communities are not conflict-free or harmonious social structures, but made up of complex sets of asymmetrical power relations in various social fields such as class, gender, or race/ethnicity. Rooted in a structuralist tradition of social sciences and the sociology of conflict, Pierre Bourdieu was interested in understanding society and inequality, and more specifically the ways in which society and social relations are produced and reproduced. Understanding society as a plurality of social fields, he argues that the control of various forms of capital defines an actor’s position within social networks and society more broadly (Sissiäinen 2000). These forms are identified as economic, cultural, symbolic, and social capital and which need to be seen not in isolation but in relation to each other in order to understand how social relations are reproduced. The positions of actors in social space are conceived of as both the cause and the effect of capital accumulation in its different guises and particularly social capital. Bourdieu (1986, 249–250) defines social capital as:

the aggregate of the actual or potential resources which are linked to a possession of a durable network of more or less institutionalized relationships of mutual acquaintance and recognition – or in other words, to membership in a group – which provides each of its members with the backing of the collectively owned capital, a “credential” which entitles them to credit, in the various senses of the word.

This definition highlights two important components of social capital. First, it is a resource based on social networks and group membership and therefore is a collective phenomenon rather than the result of individual efforts even if actors are actively exploiting its potential. Within a social network or association, each actor occupies different positions and consequently varying possibilities to activate social capital as a resource as well as appropriating related benefits in the form of economic, cultural, and symbolic capital. Actors’ predispositions are fueled by inequality
which is rooted in the differential distribution of capital in its various forms. These predispositions in turn are mobilized to legitimate the differential structure of society and the inequalities in capital accumulation between members (Tzanakis 2013). Second, and following from this, social capital is based on mutual cognition and recognition, thus taking on a symbolic character in the form of symbolic capital. Social classes formed on the basis of uneven access to resources and capital only become meaningful when accepted and internalized by members of society, normalizing and legitimizing existing uneven distribution of capital within and between classes and individuals. In other words, social, cultural, and economic capital becomes socially effective only through the process of translation into symbolic capital and the exercise of symbolic power.

In Bourdieu’s account of social capital, social structures and institutionalized behavior are paramount (Bebbington 2009). Individual agency has to be accounted for in the context of structural constraints, which he theorizes through the concept of “habitus.” As an embodied materialization of an actor’s resources, or “internalized capital,” habitus puts an emphasis on the nonreflexive aspects of practice beyond the conscious, or the habitual element of human agency (Holt 2008). Similar to the role of symbolic capital, habitus acts as a key mechanism for reinforcing views and perceptions of social structure and context as legitimate and obvious. Acquired during the process of socialization from early childhood onwards, habitus describes an actor’s predispositions and forms the basis for the generation of practices within particular social fields. Consequently, the ways in which people engage in social networks and realize social capital can only be partly explained by rational choices or conscious decisions and is to a considerable extent influenced by individual dispositions and habitual practices of which the person may not be aware.

Social capital as developed by Bourdieu arguably provides an important contribution to the study of development by establishing a conceptual framework that helps to explain how social networks can contribute to the reproduction of socioeconomic inequality. Such a conceptualization has not been without its critics, however. Questions still remain about the explanatory power of social capital and the danger of succumbing to economic reductionism, underplaying the transformational potential of social capital. There are also unresolved theoretical and methodological issues of agency/habitus and aggregation in the form of class/class habitus. Despite these problems, social capital and habitus are useful analytical lenses through which to better understand the structural mechanisms as well as embodied practices of social network formation and the struggles over the appropriation of capital(s). Somewhat surprisingly, within human geography and development studies Bourdieu’s ideas about social capital, class, and habitus have been taken up quite reluctantly so far, albeit with some exceptions, for instance, in studies of inner city gentrification, rural communities, and transnational migration. Rather, it was another concept of social capital that from the 1990s onwards attracted the attention of geographers, development scholars and, importantly, policymakers and development practitioners.

Social capital, networks, and trust

In 1993, Robert Putnam published a study of institutional performance, civic society, and economic development in Italy titled *Making Democracy Work*. Like other observers, he was intrigued by the divergent economic development trajectories of different regions of Italy,
with a burgeoning economy in what became known as the Third Italy (northeast and central regions) and markedly lower growth rates or even decline in the old industrial heartlands (First Italy) and the Mezzogiorno in the south (Second Italy). Putnam’s conclusion based on his institutional analysis was that the north displayed much higher levels of collaboration than the south. Putnam (1993, 130) summarized the key empirical findings as follows:

In the North the crucial social, political and even religious allegiances and alignments were horizontal, while those in the South were vertical. Collaboration, mutual assistance, civic obligation, and even trust – not universal, of course, but extending further beyond the limits of kinship than anywhere else in Europe in this era – were the distinguishing features in the North. The chief virtue in the South, by contrast, was the imposition of hierarchy and order on latent anarchy.

Economic prosperity and development in Italy’s north were thus understood as an outcome of social capital emanating from networks based on three main components: trust, shared norms and social obligations, and civic engagement in the form of voluntary associations.

For Putnam, trust is a crucial element of and contributor to civil society. Essential to facilitate social cohesion, trust is also a prerequisite for political institutions and governments to be perceived as legitimate. Such an understanding of trust goes beyond immediate interpersonal relations and assumes a more generalized form, where citizens trust that their involvement in producing a common good for the wider community and society will be reciprocated through the development of positive communal relations (Siisiäinen 2000). In Putnam’s own view, such generalized trust creates reciprocity and various forms of civic engagement, which in turn produce and strengthen trust, in effect constituting a virtuous circle of social capital formation that ultimately fuels socioeconomic development. In line with other conceptualizations of trust (see Murphy 2006), it is also indispensable to reduce risk and lower transaction costs in a complex and contingent environment. In that sense, trust serves as the lubricant for mutually beneficial social relations in the form of bridging and bonding forms of social capital. Bridging social capital establishes connections between geographically and/or socially distant groups and expands social networks, while bonding social capital increases cohesion within geographically and/or socially proximate groups. Unlike Bourdieu, Putnam’s work provides a more contextual or “ecological” idea of social capital based on trust and mutual cooperation, instead of focusing on networks utilized for individual benefit and/or those that reproduce class structures and socioeconomic inequalities.

A second important element of social capital, and closely related to trust, is the role of norms and mutual obligations. Without a set of values that is shared by the members of a network or community, with the expectation of reciprocity, civic engagement erodes and becomes increasingly futile. Rather than resulting in a virtuous circle of social capital for the benefit of socioeconomic development, the absence of shared norms and values will result in a downward spiral of growing distrust and ultimately the disintegration of networks, community, and society. Voluntary associations constitute Putnam’s third central element of social capital, and to him they are a major source of trust generation, along with establishing interpersonal ties. Civic activism and voluntary associations are theorized as the cornerstones of civil society and democratic systems, in line with earlier work by social theorists like Niklas Luhmann and, in particular, Alexis de Tocqueville. The latter’s studies of American society in the nineteenth century – published in 1835 as
“Democracy in America” – were highly influential for Putnam’s own work on voluntary associations, democracy, and social capital in the United States for which he became famous after the publication of “Bowling Alone” (Putnam 2000).

What Putnam has developed, then, is a concept of social capital that represents a decidedly functionalist approach compared to Bourdieu’s sociology of conflict. Putnam’s emphasis on trust and shared norms portrays a notion of community based on common interest and he by and large ignores the unequal power relations inherent in any social relation. What is more, his approach to social capital tends to conflate the concepts of Gemeinschaft (community) and Gesellschaft (society) as developed by Ferdinand Tönnies, obscuring any social and economic tensions and struggles found even within successful regions. Instead, social capital is attributed to aggregate entities like regions, or even countries, without much consideration for social struggle between actors within these spaces. The strongest criticism leveled against theories of social capital, especially in Putnam’s formulation, is their propensity for circular reasoning and tautology without establishing clear causal relationships. Social capital is at the same time cause and effect of positive socioeconomic development. As Tzanakis (2013, 7) puts it, “[a]ll features of public life purported to reflect high social capital are accounted for by the high stock of social capital in that region.” And yet, Putnam’s work has had considerable influence on policy formulation in the realm of (international) development.

Development practice and social capital

The first major international organization to enthusiastically endorse the concept of social capital as a guide for policy formulation was the World Bank (Bebbington 2009). Against the backdrop of the highly problematic and counter-productive structural adjustment programs by international financial institutions during the 1980s and amidst increasing calls for more participatory forms of development as opposed to the prevalent top-down development strategies, social capital came to be seen as a possible route toward a more socially “grounded” approach to development. Such an approach also chimed with the emerging discourse of sustainable development, following the report by the Brundtland Commission in 1987, “Our Common Future,” as it offered a new metric in support of the so-called triple bottom line goals of sustainability: economic growth, social justice, and environmental protection. During the 1990s, social capital became one of the most popular development concepts, embraced by donors and development agencies in the Global North as much as by actors, activists, and NGOs in the Global South. Increasingly seen as “capital of the poor,” the possibilities of promoting social relations at the grassroots level were seen as both progressive and empowering.

One result of efforts by international agencies to mainstream social capital approaches for development has been the livelihood assets framework, aiming to address, in particular, the problems of rural poverty and vulnerability (Potter et al. 2012). New methods like participatory rural appraisal were deployed in order to understand development at the local level better, and to integrate the voices of the people which often had been considered as “objects” of development intervention instead of active subjects following their own strategies and objectives. Social capital was a vital conceptual ingredient in producing such participatory frameworks and the language of “capital” also resonated well with the overarching narrative of “inclusive” capitalism that could make it work for the poor. Critics therefore argue that mobilizing Robert
Putnam’s concept in development policy and intervention entails adopting a “Pollyanna”-style conception of society that ignores the fundamental political-economic problems of unequal access to assets, property, and resources. As Radcliffe (2004, 522, emphasis in original) notes: “Ironically, when social capital is placed at the heart of development projects they can in fact backfire, causing the destruction of local “social capital” or sociopolitical clout and knowledge.”

Applying social capital theory in a rather uncritical way to development policy and practice, then, has contributed to some serious shortcomings in mainstream livelihoods research (Sakdapolrak 2014). First, the incarnation of Putnam’s concept often aids in reproducing imbalanced consideration of the structure–agency relationship given that power and politics are not given the necessary attention. Second, social capital – like other forms of capital – is perceived as stock that can be utilized for economic and material benefits, without proper acknowledgment of the contested nature of assets. Third, while being useful in focusing on the local dimension of development there is a danger of neglecting translocal and transnational structures and dynamics and the ramifications of global processes for individual and collective livelihood strategies. A careful geographical, multiscalar approach to livelihood research and social capital formation is therefore essential.

Finally, the global financial crisis that has engulfed global society since 2007 has brought home the vulnerabilities of citizens not only in the Global South, but also an increasing polarization in the industrial and postindustrial countries of the Global North. Here, too, ideas of social capital have been mobilized by different groups with often incompatible goals. On the one hand, civil society and voluntary association are championed by some political elites as a means to compensate for the retreat of the state (for instance, the current British government-led “Big Society” discourse), while at the other end of the spectrum alternative forms of civil association are experimented with, as illustrated by the Occupy movements that have sprung up in numerous cities, from New York and London to Hong Kong. These examples illustrate the political and contested nature of social capital, its role in reproducing unequal social relations, but also its transformative potential.

Concluding thoughts

Networks and social capital have become important and influential concepts of understanding socioeconomic development in space. At the same time, both notions are highly contested with regard to their usefulness and explanatory power and in terms of their implications for the politics of development, economic, and social change. In geography, the surge of relational thinking has provided a strong basis for network analyses across various scales, from the global to the local, and the differential power relations that emerge from and are exercised within social and economic networks. While network ontologies and methodologies have made inroads in various subdisciplines of human geography – along with many other social sciences, the concepts of social capital as developed by Bourdieu and, in particular, Putnam have resonated most notably with development studies and geography, not least because of the uptake of Putnam’s ideas in policy circles. This has created an ongoing controversy between academics who see social capital as formulated by Putnam as yet another theoretical justification for neoliberal development and the retreat of the state, and others who perceive it as a way of thinking that takes civil society actors seriously and helps promote
more democratic societies and civic engagement (Bebbington 2009). Some observers have indeed reflected on the imminent “death” of social capital as a theoretical and heuristic tool (Radcliffe 2004). But rather than throwing out the baby with the bathwater and abandoning social capital theory altogether, there exists an opportunity to reframe the discourse of social capital and development in a way that sheds light on structural, political-economic conditions as much as social networks and civil society. A way forward therefore might be what Rankin (2002) called a “Marxian approach to social capital,” critically (re-)engaging with Bourdieu’s ideas of various forms of capital, habitus, and theory of practice. Sakdapolrak’s (2014) recent attempt to re-energize livelihoods research in development studies and development geography – drawing on Bourdieu’s theory of practice – can serve as an example of how networks, social capital, and development might be re-imagined in a more critical, constructive, and progressive way.

SEE ALSO: Civil society; Class; Community; Cultural capital; Development; Global production networks; Governance and development; Livelihoods; Relational assets; Social capital; Social movements

References


Capital is (Almost) Dead?” Progress in Human Geography, 28: 517–527.

Further reading

The new cultural geography’s emergence in the 1980s represented more than the revitalization of a sidelined subdiscipline, as “culture” became increasingly central to thinking about the world and, perhaps, to the world itself. Culture materialized in the most unexpected of places (e.g., airports, laboratories, the trading-floors of the world’s stock exchanges) and encompassed topics geographers had previously neglected (e.g., footwear, fascism, furniture, feng shui, film). Globalization, postcolonialism, and a plurality of socially constructed identities (tackling what it means to be a woman, gay, black, or disabled in an increasingly fluid world) underpinned renewed attempts to understand how cultural differences comprise the human world. The sense in which that world is shaped “discursively” – meaning that the human world is not distinct from (or at least not independent of) its representations – became a mainstream concern: culture was constitutive, performative, always-already in the making. Despite some outstanding debts to the Berkeley School of cultural geography – notably an abiding interest in landscape – the new cultural geography owed more to the growth of cultural studies than to any home-grown antecedents.

The new cultural geography shaped the entire discipline, spanning topics as diverse as nationalism and landscape painting, competitive masculinity and global finance, popular culture and geopolitics, subcultures and territoriality, and dance traditions and tourism. Yet it became an umbrella term for a heterogeneous mix of approaches rather than resolving any obdurate dualisms. Postmodernism’s influence on geography’s “cultural turn” saw accusations of structural factors being neglected in favor of free-floating identities. A lingering humanism, coupled with deconstructive tendencies, saw allegations that the new cultural geography privileged representations, eviscerating reality of its (material) substance – prompting calls for “non-representational theory.” Although such
charges typically attacked straw conceptions, they demonstrate the checkered history of the new cultural geography – from which the sobriquet “new” has long since been dropped.

SEE ALSO: Cultural geography; Cultural turn; Culture

Further reading

Cook, Ian, David Crouch, Simon Naylor, and James R. Ryan, eds. 2000. Cultural Turns/Geographical

The term “new economic geography” refers to a range of approaches that have emerged in economic geography since the early 1990s that share a central concern with the links between economic action and social and cultural practices in different places. These approaches derive theoretical inspiration from various sources, including postmodernism, poststructuralism, cultural studies, anthropology, economic sociology, and institutional economics. In general, they have served to direct attention towards questions of agency, difference, identity, language, and performance that were previously marginalized and neglected within economic geography. As Wills and Lee (1997, xvii) argue, “the point is to contextualize rather than undermine the economic, by locating it within the cultural, social and political relations through which it takes on meaning and direction.”

To a considerable extent, this “new” economic geography has been defined against the “old” economic geography of spatial analysis and Marxian political economy, which have been variously criticized for ignoring the wider social and cultural practices that shape economic life and for viewing individual and collective economic agents in essentialist and reductionist ways. Despite their initial counterposing, new approaches in economic geography have not always remained separate from work in geographical political economy as important areas of overlap and cross-fertilization have developed, reflecting the increasingly hybrid and fast-moving nature of the field. These include cultural political economy and efforts to develop a more institutionally – and relationally – sensitive geographical political economy. Crucially, the new economic geography that is the subject of this entry is not to be confused with the “new economic geography” developed by a group of geographically minded economists led by the Nobel Prize Winner Paul Krugman. While this program of research has successfully appropriated the term, it is a very different beast, based upon the application of mathematical modeling techniques to questions of industrial location and regional agglomeration. The remainder of this brief entry is structured around the discussion of three new economic geography approaches that are labeled cultural, institutional, and relational.

Cultural approaches have sought to reframe the economic in particular ways, demonstrating how economic relations are saturated with issues of identity, meaning, and representation. Four main strands of culturally inflected work can be identified. First, studies of consumption that have often focused on the creation and experience of particular landscapes of consumption such as shopping malls, supermarkets, and heritage parks. Second, research on gender, performance, and identity in the workplace and labor market, such as Linda McDowell’s account of work cultures in merchant banks in the City of London. Third, work on the importance of personal contact and social networks in financial and business services. Studies of financial centers such as the City of London, for example, have shown the importance of social networks and trust, encouraging geographical concentration...
NEW ECONOMIC GEOGRAPHY

through the need for regular face-to-face contact. A fourth strand of research is concerned with corporate cultures and identities, particularly in terms of how managers and workers create distinctive corporate cultures through particular discourses and day-to-day practices. Of particular interest here is Erica Schoenberger’s work on large American corporations such as Xerox, DEC, and Lockheed, showing the limitations of such cultures in dealing with a turbulent and unpredictable economic environment.

New economic geography has also drawn upon concepts from institutional economics and economic sociology, which also emphasize the social context of economic life and the role of institutions in shaping and “embedding” economic action. Here, institutions are broadly defined as the “rules of the game,” incorporating both formal and tangible rules and laws, and more informal and intangible conventions and norms. A related set of ideas from economic sociology, particularly Karl Polanyi’s conception of “the economy as an instituted process,” has been utilized to argue that the economy is always grounded in social relations. Characteristically, geographers have “spatialized” the sociological concept of “embeddedness,” emphasizing how particular forms of economic activity are not only grounded in social relations, but also rooted in particular places and regions. One particular area in which the notion of embeddedness has been applied is in studies of transnational corporate investments in particular localities and regions and the scope for local actors to embed these investments by, for example, fostering links with local suppliers and training providers. Institutional approaches are associated with the rise of a “new regionalism” in economic geography that examines the effects of social and cultural conditions within regions in shaping economic development. Successful regions such as Silicon Valley in California, North Eastern and Central Italy, and Baden-Württemberg in Southern German have attracted particular attention, as economic geographers have attempted to identify the social and cultural foundations of economic growth and prosperity.

Another strand of new economic geography research has been influenced by the rise of relational thinking. According to Doreen Massey, the relational approach to space is grounded in three basic propositions about space, viewing it as the product of interrelations, multiple, and always changing and becoming. Relational economic geography focuses on the relationships between economic actors and broader processes of change, informed by the institutional and cultural approaches discussed above. It is particularly associated with the study of networks, defined as “socio-economic structures that connect people, firms and places to one another and that enable knowledge, commodities and capital to flow within and between regions” (Aoyama, Murphy, and Hanson 2011, 181). In this respect, it overlaps significantly with work on the role of socioeconomic practices in economic geography, defined as “the stabilized, routinized, or improvised social actions that constitute and replicate economic space and through and within which diverse actors (e.g. entrepreneurs, workers, caregivers, consumers, firms) and communities (e.g. industries, places, markets, cultural groups) organize materials, produce, consume and/or derive meaning from the economic world” (Jones and Murphy 2011, 367). For example, a study of cultural embeddedness in Utah’s high-tech economy found that Mormon religious values were upheld by a range of day-to-day practices within firms. These include the reinforcement of group norms through routine associations with other employees, observation of fellow workers, the group ratification of culturally informed corporate decisions, and the recruitment of Mormon employees.
Relational conceptions of space have fostered a new concern with “the relational region” that re-imagines regions as open and discontinuous spaces, defined by the wider social relations in which they are situated. They are created for particular purposes by particular actors, meaning that they have no essential character or fixed identity. These arguments were illustrated by an analysis of South East England, the emblematic growth region of neoliberal Britain in the 1980s and 1990s. Within the South East, areas of economic decline and deprivation exist, complicating and confounding the overarching image of growth and prosperity, while the boundaries of the region can be seen as open and porous. This relational theory of the region has been accompanied by a critique of “new regionalist” thinking, rejecting its portrayal of regions as internally coherent and externally bounded.

As outlined above, “new economic geography” research is concerned with the relationships between economic action and social and cultural practices in different places. While the term is rather specific to the mid-to-late 1990s and not widely used today outside of its appropriation by geographically minded economists, it embodies a key moment in the reorientation of the field. This relates to the widening and diversification of economic geography beyond the narrower approaches of the past as it came into contact with new theories and literatures. The underlying concern with the relationships between the economic and the social and cultural spheres of action is central to the ongoing development of cultural, institutional, and relational approaches in economic geography. These approaches look set to continue to animate and inform research in the field over the next decade or so.

**SEE ALSO:** Cultural economy; Economic geography; Regional geography; Regionalism

**References**


**Further reading**


The concept of “new industrial space” was developed by Allen Scott in 1988 to describe the agglomeration of highly specialized firms that took place in various parts of the world in the late 1970s and the 1980s. Scott argues that these small firms cluster in one place because proximity to one another allows them to reduce transaction costs. The concept was often considered as a part of the “flexible specialization” thesis proposed by French regulation school scholars, who argued that capitalism has undergone a profound transformation since the 1970s, from “Fordism” to “post-Fordism.” Here, “Fordism” refers to the system of mass production in search of economy of scale. It is characterized by vertically integrated global firms, Keynesian welfare states, and unionized labor. In contrast, “post-Fordism” is a more flexible form of production that seeks to achieve economies of scope. It is characterized by vertically disintegrated firms engaging in batch and specialized production, neoliberal regimes of governance, and the emergence of informal, contract-based, or self-employed workers. While in the 1980s regulation school scholars had interrogated issues such as vertical disintegration (Piore and Sabel 1984) and the emergence of flexible labor (Brusco 1982), Scott points out that transformation to post-Fordism has a spatial dimension: that the big factories that used to dominate the industrial landscaping are now giving way to industrial districts.

One of the cases of new industrial spaces that Scott examined was the “Third Italy.” The term refers to the central and northeast regions of Italy (including Tuscany, Umbria, Marche, Emilia-Romagna, Veneto, Friuli, and Trentino-Alto Adige/Südtirol), which in the 1970s and 1980s witnessed the clustering of small artisan firms and workshops. Each region specialized in a range of loosely related products and each workshop usually had 5–50 workers (often less than 10). The wide range of products in each cluster suggested a shift from economies of scale to economies of scope. Additionally, these small firms and workshops were known for producing high quality products and employing highly skilled and well-paid workers. The workshops were also very design-oriented and multidisciplinary, involving collaboration between entrepreneurs, designers, engineers, and workers. This unique industrial structure was in sharp contrast to the large-scale mass production systems seen in Turin, Milan, and Genoa (the First Italy) and the underdeveloped South (the Second Italy). Scott also pointed out that, because these vertically disintegrated firms often do not need to have direct contact with established industries, and because they prefer places without the presence of big unions, new industrial spaces often appear either in depressed urban enclaves (e.g., the clustering of high-tech firms in Paris’s southern suburb) or in underdeveloped regions (e.g., the Third Italy and Silicon Valley).

Industrial agglomeration is certainly not a new issue in economic geography. In the late nineteenth century, Alfred Marshall, in Principles of Economics (1890), had examined the agglomeration of textile manufacturers in England. He argued that textile manufacturers
NEW INDUSTRIAL SPACE

tended to cluster in certain regions because these places offered common pools of resources that were important for textile production, such as skilled labor, land, capital, and transportation infrastructure. In other words, firms cluster because it is easier to find qualified workers and to exchange commercial and technical information through informal channels. Marshall’s study was followed by a series of research that examined agglomerations in many other parts of the world (e.g., Perroux 1950; Hirschman 1958; Chinitz 1961; Vernon 1960). Allan Scott’s approach to agglomeration, however, is quite different from that of Marshall and his followers. In emphasizing the external benefits provided by the “common pool,” Marshall assumed local economies as collections of atomistic competitors, being aware of one another solely through the intermediation of price signals. By contrast, Scott’s transaction cost thesis emphasizes the interdependence between firms, flexible firm boundaries, cooperative competition, and the importance of trust in the formation and reproduction of industrial districts (Harrison 1992). In other words, the theory of new industrial spaces posits a strong attempt to embed economic relations into a deeper social fabric, highlighting the importance of “trust” that is built in production networks. If trust can best be built through learning about the idiosyncrasies of the actors, and if this requires repeated human interaction, then such interaction is likely to be facilitated by personal contact, and that contact is in turn enhanced by geographical proximity. In this respect, Scott’s thesis of new industrial districts coincided with Granovetter’s (1985) attempt to “bring sociology back in” to the study of the economy: acknowledging the relevance of customs, interdependent preferences, or cognitive processes in the shaping of economic structures.

One of the major criticisms of the new industrial space theory is that transaction cost alone could not explain why firms agglomerate in certain places rather than in others. One important issue in subsequent discussions about the agglomeration of specialized firms is thus the territorial dynamics of agglomeration. In “The Resurgence of Regional Economies, Ten Years Later: The Region as a Nexus of Untraded Interdependencies” (1995), Michael Storper introduced the concept of “untraded dependencies.” He argued that interdependency between firms is based on collective tacit knowledge that cannot be removed from its human, cultural, and social contexts. In other words, since knowledge is place specific, agglomeration is unlikely to take place anywhere. In Regional Advantage: Culture and Competition in Silicon Valley and Route 128 (1994), Anna-Lee Saxenian proposed the concept of “competitive advantage,” which offered an alternative explanation. Saxenian argued that the organization of Silicon Valley’s industrial system is embedded in a local culture that accepted failure and encouraged risk-taking, and that this milieu of innovation is an important reason for the development of high-tech industry in Silicon Valley having been more successful than that of Route 128 in the Boston metropolitan region.

The proliferation of literature about industrial districts in economic geography in the 1990s drew concerns from some scholars, who questioned the extent to which the thesis of new industrial space helped grasp the dynamics of post-1970s capitalism. Some pointed out that some global forces, such as innovation in communications and transportation, the invention of information technology, and interregional migration, are not considered in new industrial space theory (e.g., Bathelt, Malmberg, and Maskell 2004). In addition, scholars such as Goodman (1989) criticized new industrial space for being overly silent on the role of multinational corporations. They point out that scholarship about new industrial space tends to
ignore the flexibility and innovation abilities of big corporations. Furthermore, new industrial space was criticized by scholars such as Markusen (1991) for being ignorant of the role of state and other institutions (such as nongovernmental organizations) in shaping post-1970s capitalism. Last but not least, scholars such as Peter Daniels (2004) pointed out that the research about new industrial space had paid little attention to the rise of the service sector in advanced economies.

Since the 1990s, large cities in Western Europe and America have been witness to the clustering of small business engaging in the production of cultural products. This particular kind of geography resembles many of the characteristics that Scott describes about new industrial space: specialized firms engaging in batch production, flexible labor, and interdependency between firms. However, Scott himself did not categorize these agglomeration economies as a form of new industrial space. Instead, he argues that a “cognitive-cultural economy” is now taking shape and it has reconfigured the contour of capitalism in Western economies (Scott 2000). Capitalism itself, he suggests, is moving into a phase in which the culture forms and meaning of its outputs have become critical elements of production strategy, and in which the realm of human culture as a whole is increasingly subjected to commodification. This transformation is represented by the emergence of industries such as high-tech, services, neo-artisan production, media and entertainment, and tourism. Scott argues that this transformation is made possible by the introduction of digital technology, which, on the one hand, allows many routine labor tasks to be replaced or reconfigured, and, on the other hand, enhances cognitive and affective work. Scott also points out that because creativity and innovation rely on mutual learning and personal interactions, cognitive-cultural systems of production and work come to ground pre-eminently in large urban regions. At the same time, the cognitive-cultural economy in contemporary cities is invariably complemented by large numbers of low-wage, low-skill jobs, and the individuals drawn into these jobs are often migrants from developing countries. In other words, the cognitive-cultural economy comes with a widening of social divides. In yet other words, even Scott himself has recognized the explanatory limitations of the concept of new industrial space.

**SEE ALSO:** Flexible specialization; Industrial agglomeration; Industrial districts; Local embeddedness; Local/global production systems; Networks, social capital, and development; Transaction cost; Vertical integration

**References**


NEW INDUSTRIAL SPACE


New Urbanism

Matthew B. Anderson
Eastern Washington University, USA

New Urbanism refers to an urban design movement that highlights the importance of traditional neighborhood principles in community planning. Its goal is to design communities that are less auto-dependent and walkable, that incorporate multiple modes of transit (i.e., rail, car, pedestrian), and feature a mix of commercial and residential land uses, socioeconomic diversity, increased housing density, access to civic centers, and close proximity between job and home. Such communities are thought to facilitate healthier, more “eco-friendly” and sustainable lifestyles, and tighter social bonds between local residents and businesses. The movement has gained considerable momentum and influence among urban planners and architects since the Congress of the New Urbanism, the central organizing community of New Urbanism supporters was founded by Andres Duany in 1993. At first an alternative vision of “sustainable” urban design and development, the movement has now proliferated across North America and beyond, and increasingly represents a dominant planning framework informing both suburban growth agendas and central city revitalization efforts.

The New Urbanism movement unfolded within the broader environmental movement during the 1990s in the United States as a planning initiative specifically designed to ameliorate the social and environmental consequences of “sprawl,” such as air pollution, loss of agricultural land, community-eviscerating insularity, teen alienation, and environmental degradation (Kunstler 1994; Duany, Plater-Zyberk, and Speck 2000). In this context, New Urbanism has represented a direct response and challenge to what had become a dominant regime of planning, land-use zoning, and development in the immediate postwar period, and the sprawling suburban world this regime had fostered (Duany, Plater-Zyberk, and Speck 2000). The term “sprawl” is typically invoked pejoratively in the context of denigrating this landscape – marked by low-density, mass-produced residential subdivisions, and expansive shopping complexes – as wasteful, auto-dependent, and “unsustainable.”

New Urbanism has its philosophical roots in the classic work of earlier critics, particularly Jane Jacobs and Lewis Mumford. Mumford, an urban historian and sociologist, attributed many social urban problems of the postwar city to the emergence of sprawl, and called for a more centrally planned “regionalism” to guide the horizontal development of metropolitan regions (Mumford 1962). To Mumford, important connections must be made between local-scale communities and the broader metropolitan fabric that engulfs them. The medieval city was a source of inspiration for Mumford, who argued for mixed-use, walkable communities linked by mass transit as the ideal vision for urban planning and design. Disregarded in this vision is the penchant of modernist planning to privilege the detached single-family house and segregate residential and commercial land uses.

Similarly, Jane Jacobs, urban activist and scholar, observed the parasitic relationship between suburbanization and urban decline, and sharply criticized “urban renewal” in her now classic text, Death and Life of Great American
Cities (1961). For Jacobs, writing about New York, the building of modernist high-rise public housing projects and expansive auto-centric avenues and expressways destroyed the integrity, vibrancy, and social fabric of neighborhoods. Here, urban life is understood to flourish thanks to the close proximity between work and home, higher-density living, and a vibrant and walkable “street life.” While Mumford and Jacobs are often discussed as oppositional figures – Mumford thinking more at the metropolitan scale than Jacobs – both provided poignant critiques of suburban America and advocated for the “neotraditional” design principles that New Urbanism has come to embody.

Despite these early critics, sprawling suburban growth remained the dominant vision among developers, builders, realtors, and mortgage-lending institutions through the 1960s, 1970s, and 1980s. This dominance, however, has since slowly eroded. In short, growing fears of impending resource scarcity and anthropogenic climate change now underpin myriad multiscalar policies geared toward “sustainable urban development” as an increasingly important global objective. It is in this context that “smart growth” emerged, in the 1980s, as a regulatory framework designed to curtail the negative effects of sprawl and foster more “sustainable” communities that promote and encourage efficiency in energy consumption. Neotraditional ideas of community planning and design subsequently informed many of the policy prescriptions promoted within this broader planning agenda.

One of the first, and most influential, communities planned around New Urbanism design principles was Seaside, Florida, which was featured in the movie The Truman Show (1998). The movie garnered national attention for the community and, with it, a significant boost to the New Urbanism movement. This was a pivotal moment, as knowledge of New Urbanism was now widespread among city governments, architects, and developers. Since the 1990s, many suburban developments across urban America have adopted New Urbanism principles as a means of not only (or necessarily) encouraging healthier lifestyles, social bonds, and energy conservation, but also to maintain demand within an increasingly demonized suburban landscape among potential consumers seeking the benefits of the suburbs (safety, good schools, etc.) without the drawbacks of traditional “sprawling” attributes (Anderson 2010).

Many urban revitalization agendas have also been influenced by New Urbanism, particularly in older, postindustrial central cities where abandoned warehouse districts have been targeted for redevelopment (i.e., inner-city Pittsburgh and Denver’s Lo Do district). In these places, New Urbanism is often melded with historical preservation efforts that seek to restore an older district’s “historical integrity.” This has been accompanied by many new downtown shopping and entertainment districts which often emulate an “old town square” and feature condominiums above storefronts (i.e., Atlanta’s Atlantic Station). The movement also influenced the federal HOPE-VI program to demolish and replace America’s most distressed public housing. Since the 1990s, demolished public housing projects have typically been replaced with “mixed-income” developments designed around New Urbanism principles as the presumed solution to concentrated poverty. The redevelopment of Chicago’s infamous Cabrini Green, Henry Horner, and Robert Taylor Homes are perhaps the most well-known examples.

New Urbanism, however, has been severely criticized. Many have noted its excessive reliance on consumer preference as a social force; others have identified the movement’s commitment to a nostalgic, mythologized past; and still others view it as reproductive of an exclusionary class-based
insularity rather than generative of increased diversity (see Harvey 2000). Many studies also now report mixed results at best at many of the HOPE-VI redevelopment sites, implicating the limitations of planning and design in fostering better social conditions, while others have charged HOPE-VI – and New Urbanism by implication – as a program mobilized by urban elites to facilitate gentrification. Here, New Urbanism is criticized as a component of the neoliberal city, a criticism only amplified by the favorable treatment of gentrification by some New Urbanism promoters, who view it as a benign vehicle for the movement’s realization.

Set against this, advocates insist that New Urbanism is the antidote to society’s myriad social ills: that is, poverty, inequality, alienation, and consumptive waste. In this context, New Urbanism has emerged as an increasingly dominant planning paradigm and certainly marks one of the latest chapters in the history of urban planning and theory in urban America and beyond. Its examples and applications are now increasingly widespread. To its advocates, it is an important step in the right direction, whether viewed as a “pragmatic” response to fifty years of woefully unsustainable urbanization or as a genuinely “progressive” movement capable of creating more healthy, socially vibrant, and equitable communities and ways of life.

SEE ALSO: Neighborhood; Neighborhoods and health; Sustainable cities; Sustainable development; Urban planning: human dynamics; Urban redevelopment; Urban renewal

References

New Zealand: New Zealand Geographical Society (NZGS)

Founded: August 1944  
Location of headquarters: Palmerston North  
Website: www.nzgs.co.nz  
Membership: 281 (as of October 10, 2014)  
President: Ann Pomeroy  
Contact: ann.pomeroy@xtra.co.nz

Description and purpose

The NZGS aims to promote geography by fostering the subject in secondary and tertiary educational institutions across New Zealand and by supporting geographic research. The society functions as a networking and communication node for New Zealand geographers by making original geographical research available and by liaising with the New Zealand Board of Geography Teachers, geography departments of New Zealand universities, and geographers involved in the private sector, Crown agencies, and community organizations.

Journals or major publication series


Current activities or projects

The NZGS has a core membership of 200–300, but is involved with a much larger group of geographers (more than 900) in New Zealand and internationally through emails and the NZGS website. Monthly newsletters and information specifically of interest to New Zealand geographers are sent through a geography community email list. The society’s activities are carried out by six local branches (Auckland, Waikato, Manawatu, Wellington, Canterbury, and Otago), a highly motivated executive committee and postgraduate network, and in partnership with the New Zealand Board of Geography Teachers. Experienced members that have made a significant contribution to geography in New Zealand can be elected as “Fellows of the NZGS.”

New Zealand Geographical Society conferences are held biannually and attract geographers internationally. Conferences are jointly held on a regular basis with the Institute of Australian Geographers (IAG). The society supports focused geographical research in special study groups. Outcomes of these are presented in seminars or symposia and are regularly published as special issues in the New Zealand Geographer journal. The NZGS is fortunate to have an active postgraduate network which has contributed to a variety of specialty sessions at conferences and quality publications in the New Zealand Geographer. The New Zealand Board of Geography Teachers is equally strong. One of their best contributions is the Network Magazine, an online newsletter that is published on the NZGS website and electronically distributed to all school teachers. Every year at their annual general meetings, the society awards outstanding geographers with distinguished awards and also acknowledges exceptional contributions by postgraduates.

The NZGS is registered as a charitable entity and is governed by an elected council, currently comprising 19 members, which elects the executive committee and specialized subcommittees. The NZGS is a member of the Royal Society
of New Zealand, the International Geographical Union (IGU), and the New Zealand Geographic Board.

**Brief history**

The first separate university geography department was established at Canterbury University College under geologist George Jobberns in 1937 (Auckland, Otago, and Wellington followed in 1946). In 1939, Kenneth B. Cumberland and a group of enthusiastic New Zealand geographers initiated an overseas branch of the English Geographical Association at Canterbury University College, although this lapsed in 1941. Efforts in 1943 led to formal incorporation of the New Zealand Geographical Society in August 1944, supported by the New Zealand Ministry of Internal Affairs, with Ben J. Garnier as secretary. Auckland and Otago branches followed in 1945, and Manawatu and Waikato in the 1960s.

The first issue of the *New Zealand Geographer* was published in August 1945 with three issues per year published electronically since 2005. NZGS conferences began in 1955 and have been held biennially since 1992, periodically together with the Institute of Australian Geographers (IAG). NZGS has co-organized regional conferences of the IGU (Palmerston North, 1974) and jointly with the IAG in Brisbane (2005).

New Zealand’s location in the Asia-Pacific region and its diverse physical features have continued to attract geographers that are energetic, highly motivated, and strongly committed contributors to the New Zealand geographical community. The NZGS remains a vibrant and invigorating society carrying out activities locally and nationally with annual general meetings, conferences, regular webinars, and quality publications in the society’s journal.

Submitted by Maria Borovnik
Newly industrializing economies (NIEs)

Shiuh-Shen Chien  
Liang-Chih Chen  
Dong-Li Hong  
National Taiwan University

The term “newly industrializing economies” (NIEs), also known as “newly industrializing countries” (NICs), refers to countries and regions undergoing rapid industrialization and economic growth, particularly following World War II. Generally speaking, NIEs have enjoyed largely continuous growth mainly because of export-oriented industrialization, which is characterized by (i) export policies focusing first on primary commodities, followed by manufactured goods, (ii) financial controls, including on currencies and interest rates, and (iii) a cheap and stable labor supply. The best-known NIEs in Asia are Hong Kong, Singapore, South Korea, and Taiwan, collectively referred to as the “Asian Tigers,” which averaged annual gross domestic product (GDP) growth rates in excess of 6% between 1970 and 1990.

There are at least three different, but interrelated, approaches to explain the development of NIEs. First, the liberal market approach considers the changing global economy as a main driving force, including the liberalization of international trade and investment. This approach focuses on the impact of free trade and market mechanisms on economic development. Second, the developmental state approach focuses mainly on the strong role played by the nation-states and their adoption of policies to promote economic development. Third, the technological upgrading approach is concerned with how NIEs learn and innovate to catch up with their industrialized counterparts. Although this distinction helps us to form a general picture of NIEs, it should be noted that, while these three dimensions are theoretically distinguishable, they are empirically intertwined in practice.

Explanations for NIE growth

The liberal market approach

This approach argues that the industrialization of NIEs is closely related to the relocation of industrial and manufacturing sectors from advanced economies in the context of global economic liberalization. In the 1930s, trade protectionism was widely seen as contributing to economic recession, which was a major factor in provoking World War II. In 1948, the General Agreement on Tariffs and Trade (GATT) was established as the first international governmental organization aimed at eliminating international trade barriers, followed by the International Monetary Fund (IMF) and the World Bank (WB).

Trade liberalization was followed in the 1960s by the lifting of restrictions on international investment. Labor-intensive and routine production tasks began to shift out of industrialized countries, initiating a further large-scale international campaign of deregulation and liberalization of cross-border investment. Moreover, capital investment also shifted from developed to developing counties for two main reasons: first, falling profits forced companies in labor-intensive manufacturing sectors (e.g., garments, shoes,
and plastics) to shift production to lower-wage countries; second, technological and organizational advances decreased the importance of geographical proximity and allowed Western manufacturers to divide complicated production processes into separable steps which were then moved offshore.

This trend brought about the emergence of a new international division of labor (NIDL). In the “traditional” international division of labor (starting in the nineteenth century), less-developed countries exported their raw materials for industrial production in the developed countries. In the NIDL, on the other hand, Western manufacturers were attracted to certain developing countries offering low-cost production conditions (including low-cost labor). Those countries adopted policies specifically to attract overseas trade and investment in a bid to accelerate industrialization and economic growth.

The development of the Asian Tigers is a case in point. In the 1960s, Taiwan and South Korea opened up their borders to attract overseas capital and production, mainly from Japan and the United States. By doing so, Taiwan and South Korea started to develop domestic labor-intensive light industries. Accelerated liberalization led to faster industrialization and more development. Similarly, a second wave of Asian NIEs, including countries such as Malaysia, Thailand, and Vietnam, began liberalizing their economies in the 1980s, attracting capital and production from Taiwan and South Korea. This pattern, with development flowing from Japan to the Asian Tigers and then to Southeast Asia, later became known as the “Flying-geese model.”

In sum, the liberal market approach posits that the new international division of labor plays a key role in the development of NIEs, and that states seeking to accelerate development should liberalize policies to facilitate international trade and cross-border investment. Although the liberal market perspective highlights the importance of economic liberalization as a key for NIEs, it does not explain why some countries benefit more than others from implementing such economic liberalization policies. To answer this question, we turn to the developmental state approach.

The developmental state approach

In the late twentieth and early twenty-first century, the Organisation for Economic Co-operation and Development (OECD) and the World Bank classified Greece, Portugal, Spain, Yugoslavia, South Africa, Mexico, Brazil, Indonesia, Malaysia, and Thailand as NIEs, but these countries generally did not experience the same level of economic growth as the Asian Tigers from the 1960s to the 1980s. The developmental state approach, mainly derived from the experience of the Asian Tigers, focuses on the positive role of the state in supporting industrialization through domestic policy. Scholars find that, instead of pure liberal policies, appropriate state inventions and extensive related regulations should play a part in NIE development. According to this approach, industrialization involves five dimensions.

The first dimension in this approach is the development of agricultural support, with the state initiating rural land reforms to offer farm-land ownership or long-term leases to private individuals, who are therefore motivated to increase agriculture productivity. Agricultural surpluses are tacitly extracted by the state to support industrial development. Farmers are encouraged or forced to use fertilizer to increase agricultural productivity, thus driving the development of other industries, which lay the foundations for early-stage industrialization. In addition, after raising productivity, fewer farmers are needed to generate the same agricultural yield, releasing labor resources to...
migrate to cities, thus triggering urbanization and industrialization.

The second dimension focuses on industrial policy. The state strategically selects certain industries for priority development, providing temporary subsidies, trade barriers, and related import substitution policies to support their protected development. State-owned enterprises and associated public research institutes are also established to facilitate the expansion of these strategic industries.

The third dimension concerns the spatial industrial policy. To attract investment, the state establishes special development zones with quality infrastructure and preferential policies such as cheap lands, low taxes, and loose enforcement of labor and environmental regulations. Such zones are also referred to as “export processing zones,” “science parks,” “industrial zones,” and so on.

The fourth dimension is the development of market-friendly labor policies. Labor union activity is restricted, keeping wages low and attracting incoming investors. The state also provides investors with an educated workforce.

Fifth, the state provides an appropriate capital environment. For example, the state can peg its exchange rate to a foreign currency (e.g., the US dollar) at a relatively low level, thus increasing the competitiveness of exports. In addition, the state strongly encourages households to save money in domestic banks, providing a source of funding for domestic investment. Finance policy also allows firms to retain a high proportion of their profits for reinvestment.

It must be noted that these market-friendly state-led development policies were implemented by far-sighted technocrats under authoritarian leaders. Those technocrats formulated policies that favored economic development, and the authoritarian leadership had an upper hand to implement these policies with no, or very limited, opposition. The state was also able to act autonomously to collaborate with private enterprises in the implementation of these policies, a practice which is referred to as the embedded autonomy of the developmental state.

In other words, unlike the liberal market approach (which argues that the state should refrain from intervening in the market), the developmental state approach emphasizes the need for a strong interventionist state to promote economic liberalization. And these two perspectives offer complementary understandings of the rise of NIEs.

The technological upgrading approach

The final approach is specifically concerned with technological capacity, asking how NIEs catch up with their advanced counterparts in terms of technological knowledge and expertise. The “latecomer advantage” thesis argues that the path of technological development in advanced economies serves as a useful reference for latecomers, allowing them to avoid redundant trial-and-error processes and risks. But not surprisingly, the learning process is found to differ widely between different NIEs and between industrial sectors within a single NIE, thus explanations must be found at the regional, and even firm, level.

At the firm level, at the outset of the globalization process, multinational corporations (MNCs) from advanced economies drive the emergence of NIEs. These MNCs diffuse their technologies to NIEs via direct investment or by subcontracting their production chains to local NIE firms. In the case of Singapore, for instance, its industrial development benefited greatly from technology transfer through foreign direct investment (FDI) from the West. In other NIEs, such as South Korea, policy is centered around
large-scale domestic firms in which the government directly subsidizes conglomerates (e.g., the Chaebol) to vertically integrate production in specific industries, thus fostering technological upgrading within complex production and supply chains.

In addition, in some NIEs, such as Taiwan, small- and medium-sized enterprises (SMEs) are key business agents for economic development. These firms are more flexible and dynamic, and excel at using informal channels (e.g., imitation, reverse engineering, casual counseling, etc.) to achieve technological advances. Finally, the interactions between MNCs, large domestic firms, and SMEs are highly dynamic and intertwined with other firms in so-called global production networks (GNPs) in various industries. In some high-tech industries, engineers who are born in NIEs but educated in the United States move frequently between the two economic spheres, serving as agents for the transfer of tacit knowledge, capital, skill, and know-how between Silicon Valley and their home NIEs. This means that technology upgrading is based not only on the NIEs, endogenous assets but also on their complex exogenous relationships and the strategic coupling of economic actors working in different regions.

At the regional scale, state-built development zones in NIEs provide diverse opportunities for innovations among spatially clustered firms and individuals. First, in NIEs the state establishes relevant universities and research institutes near such zones. Geographical proximity of these different organizations minimizes infrastructure costs and maximizes access to specialized information and talent pools, which collectively drives the development and diffusion of technological and business innovation. Second, such spatial concentration increases the likelihood of social interaction among firms, thus facilitating knowledge spillover and learning.

Challenges facing NIEs

The transformation of NIEs from “industrializing” economies to “industrialized” ones entails certain challenges, which can be also categorized as belonging to three different scales. First, the situation of the international political economy has changed. The development of the first wave of NIEs in Asia lasted from the 1960s to the 1980s, and occurred in the historical context of Cold War capitalism. NIEs were able to export their industrial products to the United States, the biggest consumer market at the time, because the United States strategically included NIEs as a part of its Cold War strategy of containment for China and the Soviet Union. After the end of the Cold War, this relatively stable international structure was replaced by a new, more dynamic global power structure characterized by the decline of US and European hegemony, further macroregional economic integration, and the rise of China and other developing countries (e.g., India) as emerging global consumption markets. These developments forced NIEs to periodically adjust their position in the changing international division of labor.

Second, the developmental state in NIEs also encountered challenges from within. Citizens who had enjoyed the fruits of economic development chafed under authoritarian rule, demanding democratic reforms including free and fair elections, media freedom, the independence of the judiciary, and the legalization of civil society organizations (CSOs). These challenges reduced the autonomy of state actors, and gave voice to concerns, such as environmental protections, social inclusion, human rights, or even anti-growth alternatives and so on, which potentially clashed with the state’s priority of economic development.
Third, NIEs face challenges in maintaining the pace of technology upgrading. NIEs developed from agrarian societies to industrial economies through the use of received technologies, but the next stage of development requires NIEs to innovate on their own. For example, in the current arrangement, many NIE firms are relegated to the role of original equipment manufacturers or designers, giving them little negotiating leverage when dealing with MNCs, and leaving them with limited resources for further technology upgrades. MNCs are reluctant to allow the development of potential competitors, and use international intellectual property law to prevent NIEs from making further technological advances through imitation.

**SEE ALSO:** Corporations and local states; Corporations and the nation-state; Export processing zones; Firm migration; Global production networks; Industrial agglomeration; Industrial districts; International division of labor; States and development; Trade, FDI, and industrial development

**Further reading**


Niche theory and models

Jennifer A. Miller
Paul Holloway
University of Texas at Austin, USA

The niche concept is one of the most fundamental ideas in ecology, and in spite of the fact that it has been a topic of books and journal articles for the past 100 years, several different niche theories and definitions have evolved. Within geography, recent attention to the importance of niche theory has increased in the context of developing inductive models to describe species–environment relationships that can subsequently be used to generate predictions for a number of biogeographic applications related to, for example, biodiversity inventory, risk of invasive species, and consequences of climate change.

The concept of a species’ ecological niche and its response to environmental gradients provides an important theoretical basis for quantifying species–environment relationships using these models. While these types of correlative models have been known by different names in the approximately 20 years that they have been used (e.g., “habitat suitability models” and “bioclimatic envelope models”), the terminology has more recently been split between “niche models” and “species distribution models.” Although these two terms have often been used interchangeably, there has been considerable debate about the fundamental and semantic differences between them, the types of data used to fit them, and the products they are used to generate. The main ways in which niche theory has been used to conceptualize these inductive models are discussed in more detail below.

In addition to the semantic differences associated with the development of niche models, the concept of an ecological niche has been heavily debated in terms of what it is, how it is defined, and how it can be used. Of the many different concepts and distinctions associated with niches, one of the main differences involves defining them in terms of the direction of the species–environment relationship: the effect of the species on the environment (“impact”) versus the effect of the environment on the species (“requirements”). This distinction is also implicitly tied to spatial scale: both the impact concept and the requirement concept are necessary at finer scales, while the requirement concept is more important at broader scales (species impact on its environment is also more difficult to measure at broader scales). These categories are also consistent with different definitions of niche that have historically been associated with the ecologists who first wrote about them in the past 100 years. Three of the main “author-ian” niche theories that are most important to the practice of niche modeling are described below. Many of the authors who originally wrote about niche theory incorporated similar ideas independently, and instead of summarizing them chronologically, they are summarized in terms of their contributions to niche modeling, with the most relevant contribution discussed last.
NICHE THEORY AND MODELS

Niche theory

Eltonian niche

The two main classes of niche theory have been described as “Eltonian” and “Grinnellian.” Elton’s (1927) original concept of a niche was related to a species’ functional role and the impact it had on its environment. It was characterized by ecological factors related to biotic interactions and resource consumption and was generally explored only with the aid of detailed field studies. Elton considered biotic interactions, specifically competition, to be a main driver of species composition and distribution. To test the importance of competition, Elton suggested comparing the observed species to genus (S/G) ratios for a location to the ratio that would be expected under randomness. Presumably intraindividual competition would prevent many similar species from occupying the same niche, therefore resulting in a lower S/G ratio. Subsequent studies that compared these ratios produced mixed results, some related to the confounding effect of diversity, and one of the main criticisms was the lack of an appropriate null model against which to compare the observed ratios.

Factors related to competition and resource dynamics are difficult to measure at the spatial scales at which most niche models are developed. This attenuation of importance of biotic factors as spatial scale decreases has been referred to as the “Eltonian noise hypothesis” (Peterson et al. 2011), although this hypothesis has been refuted in some macroscale studies. While original Eltonian niche theory is generally outside of the realm of the type of niche modeling discussed here, the importance of biotic interactions to species distributions has been incorporated in the concept of bionomic factors (see Hutchinsonian niche discussion below), which generally refer to dynamic resources that can be consumed, and for which competition may become an important aspect.

Grinnellian niche

Although he was not the first person to use the word “niche” in a biological context, Grinnell (1917) is generally credited with the original development and formal definition of the ecological niche of a species. Grinnell’s definition of a niche comprised all of the conditions necessary for a species’ existence, ranging from environmental factors associated with physiological tolerances to interactions with other species. Although he emphasized predator interactions over competition, the origins of the “competitive exclusion principle” – that two species must have different traits related to their fitness in order to coexist locally (also known as the “Volterra–Gause principle”) – can be traced back to experiments conducted by Grinnell.

Both Elton’s and Grinnell’s niche concepts were somewhat geographically abstract yet still considered to be an attribute of a place and fit in with a community ecology conceptualization of species composition (e.g., Grinnell’s exemplar California thrasher species was part of the chaparral community). A vacant or empty niche could exist where a particular habitat was unoccupied. While Elton’s concept was more related to a species’ functional role in its community or biotic environment, Grinnell’s concept was more analogous to what is considered habitat, and therefore involved factors that are important to species distributions at spatial scales that are more relevant to niche modeling, although some niche factors suggested by Grinnell such as “food size” are not easily conceptualized as occurring along gradients.
Hutchinsonian niche

Hutchinson (1957) advanced upon these niche concepts to provide a framework for quantifying the species–environment relationship inherent to niche theory and the basis for niche models. Similar to Grinnell’s niche concept, the Hutchinsonian niche is more related to a species’ environmental requirements rather than to its impact on the environment, but unlike both Grinnell and Elton, Hutchinson was the first to define a niche as an attribute of a species or a population instead of a place. The idea that a species had a niche was consistent with the individualistic theory in plant ecology as proposed by Gleason (1926) and extended by Whittaker, and led to the development of methods that quantified species responses to environmental gradients such as gradient analysis and ordination (Whittaker 1967).

Hutchinson’s most important and transformative contribution to the development of niche modeling was his characterization of a species’ niche in terms of environmental space. A Hutchinsonian niche was defined as a function of a species’ requirements (limiting factors) that were represented as gradients or axes along which a species’ tolerance could be plotted. In his seminal paper, Hutchinson (1957) considered limiting factors to be those which permitted a species to survive and reproduce, or more broadly, “to exist indefinitely” (416). The resultant $n$-dimensional hypervolume represented the sum of tolerable conditions across all factors in environmental space.

Hutchinson further distinguished a species’ niche as being fundamental or realized. A fundamental niche comprises a species’ full range of tolerable environmental conditions, the $n$-dimensional hypervolume, while the realized niche describes the subset of environmental conditions to which it is restricted as a function of interactions with other species.

Hutchinson formalized the link between environmental space – in which a species’ fundamental niche was defined – and the biotope, or the physical environment in terms of a species’ requirements. A species’ fundamental niche is defined in abstract environmental space and it is likely that only a subset of the requirements exist in geographic space; any point in niche space could correspond to many different (and geographically disparate) locations in biotope space, or none at all. It is this translation of a niche defined in abstract environmental space to geographic space – what has been referred to as “Hutchinson’s duality” (Colwell and Rangel 2009) – that provides the foundation for niche modeling.

Hutchinson also introduced the terms “scenopoetic” and “bionomic” to distinguish between environmental factors with which individuals did not interact and did, respectively. These factors are similar to the distinction between Grinnell’s niche, which would be defined by noninteracting scenopoetic factors, and Elton’s niche, which would be more closely related to bionomic factors. It is important to note that traditional niche theory generally accounts for only negative population interactions such as competition and predation, while positive interactions such as mutualism were part of what Hutchinson originally considered to be the fundamental niche.

Niche modeling combines aspects of all three niche concepts described above. Figure 1 illustrates the general process involved in formulating a niche model using presence/absence data for *Yucca brevifolia* and three environmental variables in a portion of the Mojave Desert in California. Data on species occurrences are combined with environmental factors (Figure 1a) to quantify the species’ hypervolume in environmental space (Figure 1b, in this case three dimensions: precipitation, temperature, and southwestness), which
Figure 1  Illustration of niche modeling process. Species and environmental data are shown in (a) (red is species absence; green indicates species presence); species data are plotted in environmental space; the relationship established in environmental space in (b) is then projected to geographic space to produce a map of predicted presence/absence (c).

is then applied to geographic space (Figure 1c) using the appropriate digital environmental layers in a geographic information system (GIS). This relationship between environmental space (species’ niche) and geographic space (species’ distribution) is described in more detail below.

Niches and distributions

Soberón (2007) introduced a framework useful for examining the three types of factors that determine the distribution of a species. The “BAM” diagram (Figure 2) can be used to visualize the individual and combined effects of biotic factors (B; biotic factors that affect resource use, e.g., interactions (similar to, but not synonymous with, Hutchinson’s bionomic factors)), abiotic factors (A; scenopoetic factors, e.g., temperature), and movement factors (M; e.g., accessibility based on dispersal capacities). Each circle represents the geographic space in which those conditions – biotic, abiotic, or movement – are suitable for a particular species. The entire circle represented by A can be thought of as the fundamental niche expressed in geographic space, or the potential distribution (Hutchinson’s biotope) of a species. A species’ actual distribution is represented by the intersection of all three factors (bam). The effects and limitations of each of these factors can shift in relative importance in a niche model application, depending on a particular study’s objective or goal.

In geographic space, the potential distribution and actual distribution are generally analogous to fundamental and realized niche, respectively. A distinction between the fundamental niche and potential distribution is that the fundamental niche exists in environmental space, and when it is translated to geographic space, not all of the environmental conditions exist at a particular time, and therefore the potential distribution is a subset of the fundamental niche.
A species’ fundamental niche is relatively static, but because the availability of suitable environmental conditions can change with time, a species’ potential distribution can expand or contract in future projections. These factors are conceptualized as occurring in geographic space, so the areas they represent can change as their respective suitabilities change. For example, a species would not be expected to be observed where only abiotic and movement factors are suitable (\(am\)) because of competition or a similar negative interaction factor. If negative competition was no longer a restriction, those areas could be reclassified as \(bam\) and become occupied. The intersection of suitable biotic and abiotic factors, but beyond suitable movement factors, (\(ba\)) represents areas where a species could survive, if they were accessible. Locations that were within this \(ba\) area could be targeted for translocation or reintroduction studies, or to investigate the risk of species invasions. Studies focusing on projecting future distributions as a function of climate change would generally aim to describe all of \(B\).

### Niche models

Considerable debate addresses what exactly these inductive models used to translate environmental niche space to geographic space should be called. Although terms ranging from “predicted habitat distribution models” to “bioclimatic envelope models” have been used in the past 20 years, the terminology has converged, but not coalesced, on the terms “species distribution model” (SDM) and “(environmental or ecological) niche models” (ENM). While both of these models are firmly rooted in niche theory, SDMs are used primarily to generate a prediction of a species’ distribution in geographic space rather than in environmental space. The use of the term “niche models” is more problematic because they are rarely formulated with data on species interactions, which suggests that they should be used only to predict the fundamental niche as opposed to the realized niche. In practice, however, the data on which they are calibrated were collected in the field and are therefore necessarily representative of the realized niche. The type of data (both species and environmental), as well as the statistical method used to quantify the species–environment relationship, also influences whether it is more appropriate to consider the technique an SDM or an ENM.

### Species data

The response variable of interest in species distribution and niche models can be represented by a variety of measures of species importance such as abundance or richness, but the most commonly used variable is presence-absence or...
NICHE THEORY AND MODELS

presence-only data. Species’ presence should be observed in the intersection of the biotic, abiotic, and movement circles (bam) in Figure 2. A species will most likely be absent outside of that intersection, although sink populations may be detected in other parts of the circle described as movement suitable (M). A species would be absent in areas outside of the B circle as a result of negative interactions and outside of the A circle as a result of unsuitable environmental conditions.

Keeping in mind the limitations inherent in using empirical data in niche models, a model calibrated with presence-only data would be more appropriate for predicting a result closer to the fundamental niche (A), while absence data would be necessary to distinguish between areas that were biotically suitable (e.g., ba, bam) from those that were unsuitable (A, am) in order to predict a result closer to the realized niche. Given that most niche models are developed using only abiotic factors, the information contained in observations of absence may be used indirectly to distinguish between abiotically suitable but biotically unsuitable and abiotically and biotically suitable locations.

Environmental data

Environmental factors vary in importance across species and predictor variables should be selected based on ecological relevance. The limits of several species are often coincident with particular combinations of climate, and abiotic climate variables such as temperature and precipitation, as well as more complex biophysical variables such as growing degree days and potential solar radiation, are often included in species distribution and niche models. These factors generally fall into Hutchinson’s scenopoetic variable category, although there are exceptions; for example at a broad spatial scale, solar radiation would be considered scenopoetic and noninteractive, while at the scale of a forest floor, competition for light could modify its designation to bionomic.

If modeling a species’ fundamental niche is the objective, abiotic factors such as climate should be sufficient. If the objective is to model a species’ realized niche, information on biotic interactions should be incorporated into models, although this has proved conceptually challenging. There is little empirical data on how one species directly influences another, and interactions can differ between populations of the same species. The most commonly used method of incorporating biotic interactions is to include the probability of finding another species alongside these abiotic conditions. In addition, the spatial scale will also affect the subsequent importance of interactions such that the signal of biotic interactions is often lost at coarse spatial scales (the Eltonian noise hypothesis).

Including information on mobility (M from Figure 2) has only recently been addressed, typically in the context of modifying predictions based on changing climate conditions. When it is incorporated, “movement” has been represented by using a measure of dispersal ability. The two most widely used involve the extreme scenarios of “full” dispersal ability, in which a species can access any future suitable locations, and “no” dispersal ability, in which a species is restricted to future suitable locations only within its current distribution. Dispersal ability can also be empirically derived. Data used to measure accessibility can also take the form of a separate movement model (i.e., dispersal kernel or cellular automata), which is then coupled with distribution and niche models.

Statistical methods

Many statistical methods have been used to fit ecological niche models and the specific method
used is often based on both the research objective and the type of data available for modeling. Environmental envelope and similarity are conceptually the most simplistic type of methods used in ENM, and they are also one of the few types of methods that are appropriate when only presence data are available. Envelope models define the hyper-rectangle that encompasses the species response data in multidimensional environmental space and represent a more direct implementation of Hutchinson’s concept of a species’ fundamental niche. In general, the statistical methods that use presence-only data tend to overpredict or result in high commission errors (“false positives”; locations that are predicted to be suitable or contain a species presence, but where the species was not actually observed). This output may be desirable if the objective is to identify potential locations for translocation or reintroduction, in which case areas where commission errors occurred could describe abiotically suitable locations that are currently unoccupied (e.g., as a result of inaccessibility) but could be occupied.

Another class of models that can be used with presence-only data compares the environmental factors associated with observations of presence to the entire range of environmental conditions within the study area. The “available” habitat or background data are also sometimes sampled as “pseudo-absences.” Maximum entropy (Max-Ent) is the most widely used of these methods, and in comparisons with envelope methods such as those described above, MaxEnt generally performs better and results in fewer commission errors.

When data on species presence and absence are available, regression-based models are still the most widely used technique in ENM and they extend the envelope method by modeling variation in the response data within the occupied environmental space and selecting predictor variables based on their observed importance in the final model. There has also been considerable progress in the application of machine learning methods such as decision trees and artificial neural networks in niche modeling applications. Newer ensemble versions of decision trees (bagging, boosting, random forests) have been shown to be of particular promise in quantifying more complex species–environment relationships that may be nonlinear, hierarchical, or involve interactions.

If the objective of a study involved selecting a reserve area, excessive commission errors would be problematic; therefore, a regression or machine learning method using presence/absence would be more appropriate. In general, the more complex statistical methods used with presence/absence data produce a result closer to the actual distribution/realized niche end of the spectrum while more simplistic envelope models used with presence-only data produce a result.

Model evaluation

Ecological niche models can be used for two main purposes: explanation or prediction. Understanding the environmental factors associated with a species’ niche is the main driver behind evaluating a model for its explanation. Niche models are intended to describe suitable conditions in environmental space, and therefore should be robust for predicting across space and time. This is particularly important in the context of climate change projections (see below).

These two distinctly different aims of an ecological niche model strongly influence the metrics used in model evaluation. When explanation is the aim, models are often assessed using model fit diagnostics, and an information criterion approach is becoming more widely implemented.
NICHE THEORY AND MODELS

When prediction is the aim, classification accuracy is assessed using an independent test dataset or a data resample. Several evaluation metrics are currently widely used (e.g., area under the curve (AUC), kappa, sensitivity, specificity) and as such it is considered good practice to report several metrics in evaluation. As with data and statistical method considerations, accuracy assessment metrics should be selected based on the study objective. When the selection of an area for potential translocation or reintroduction is the goal, many commission errors would be acceptable, but omission errors (“false negatives”; locations that are predicted to be unsuitable or contain absences but which really contain a species) should be minimized. Observations of “absence” may not be necessary in the assessment process, whereas in the reserve design example, absence data would be crucial for assessment.

Niche model applications

Several niche modeling applications are described in more detail below. It should be noted that while the niche theory and gradient analysis concepts on which niche models are based involve (plant or animal) species’ responses, the methods have been used to study the distributions of similar biogeographic phenomena such as biodiversity, fire, and vector-borne diseases.

Climate change impacts

Climate change studies are one of the most long-standing application areas for niche models. When used in niche models, abiotic variables are assumed to be somewhat static, but climate variables can change and, while this should not affect a species’ fundamental niche in short time periods (≈10–50 years), a species’ potential distribution in geographic space can change considerably. This can also induce changes in biotic interactions, and limitations associated with dispersal can result in substantial changes in a species’ actual distribution as well. By understanding which factors define a species’ current ecological niche, predictions can be made about the resilience of certain species in the future by extrapolating species responses in environmental space into future geographic space (under various climate change scenarios). Using niche models to assess extinction risk requires explicit modeling of changes in variables that define the actual distribution (bam). Ecological niche models can estimate habitat reduction using current environmental data, which can be used solely or coupled with population models, to estimate extinction risk. Studies focused on species reintroductions are more interested in the intersection between areas of favorable biotic and abiotic conditions (ba), as for some reason (e.g., local extinction, barrier to movement) the areas are not currently accessible to species. This modeling can then identify suitable locations for these species to be reintroduced or translocated.

Invasive species risk

The geographic spread of invasive species is also a phenomenon with global biological and economic consequences. Niche models have been used in this field of research to identify geographic areas that are currently unoccupied by the species but are susceptible to invasion if dispersal limitations are removed (i.e., the species in introduced). Niche models to predict spread of invasive species are fit typically using species and environmental data from one region (usually the native region) and then projected into new regions, where the species may be currently invading or where they could potentially invade. In terms of assessing the accuracy of
predictions from invasive species niche models, commission errors could be used to identify potentially invaded areas, while omission errors would be far more problematic and should be minimized.

Niche conservatism

One of the newest application areas of niche models so far examines the rate of change of niche characteristics both for closely related species and across time. In the niche model studies discussed here, relevant timescales could range from 10 years to several hundred years. Closely related species that share similar niche characteristics even if they are not close in geographic space provide evidence for “niche conservatism.” While niche conservatism would require that all three areas represented by BAM remain unchanged simultaneously, it is a prerequisite for several niche modeling applications such as projecting the impacts of climate change and determining invasive species risk. For example, determining niche conservatism with respect to climatic conditions can have important consequences for potential range expansion. It has also been noted that ignoring \( B \) can obfuscate studies that test for evidence of niche conservatism.

Conclusion

Niche models based on species–environment relationships have become a crucial tool in biogeographic research. While firmly grounded in niche theory and gradient analysis, the application areas in which niche models have been applied are continuing to grow. It is important to understand the assumptions and requirements associated with the niche concepts on which they are based in order to use these models appropriately.

SEE ALSO: Biodiversity; Biogeography: history; Climate change and biogeography; Community/continuum in biogeography; Ecogeography/macroecology (range and body size)

References


Further reading


NICHE THEORY AND MODELS


Nonrepresentational theory

Paul Simpson
Plymouth University, UK

Nonrepresentational theory refers to a diverse body of work that emerged in human geography in the mid-to-late 1990s and has since attempted to transpose a range of key concerns for human geography into diversified registers of thought and action. This movement of transposition originated primarily from a series of publications written by Nigel Thrift, but also evolved through the work of a number of his doctoral students from the University of Bristol and a range of interested others (see Anderson and Harrison 2010). As well as coming out of a variety of theoretical, technological, and social developments gaining momentum around this time, nonrepresentational theory also developed as a response to the perceived prominence of “representational” modes of thought in human geography. It was felt that the emphasis on seeing the world as “text” to be deconstructed provided a somewhat limited window onto a range of phenomena relevant to geographic inquiry. By way of contrast, nonrepresentational theory is concerned with the practical and improvised unfolding of the everyday that exceeds the representations that can be made of it but through which much of the world’s consistency, meaningfulness, and legibility come to be (re)produced and maintained.

Two immediate stumbling blocks that have impeded the reception of nonrepresentational theory are the words “nonrepresentational” and “theory.”

For many, the negativity of the “non” in “non-representational” meant (or at least implied) a move away from concerns with representations and texts altogether (see Representation). This impression is understandable given the tenor of some of Thrift’s early outlines and the way in which these positioned themselves very strongly in opposition to New cultural geography (Thrift 1996). This reaction in turn led to the suggestion of “more-than representational theory” as a more inclusive nomenclature that would soften some of this oppositional tone (Lorimer 2005). These debates aside though, nonrepresentational theory is in fact interested in representations. Within this work representations are considered for what they do, as being performative and so playing a part in the ongoing shaping of daily life (Dewsbury et al. 2002). Therefore, the critical target of the “non” is more a form of “representationalism” that reduces the world to, and fixes and frames it within, text or discourse alone, and not at representations in and of themselves.

Furthermore, nonrepresentational theory is not an actual “theory.” Rather, it constitutes a style of thinking that values practice and process. This means that it is better to think of nonrepresentational theory in the plural, as non-representational theories. Within this, a range of ideas from Poststructuralism/poststructural geographies, vitalists, Phenomenology, pragmatists, Feminist geography, and a collection of relational and constructivist social theorists mix in varying concentrations and combinations producing quite diverse and at times seemingly contradictory accounts of this happening of the world.

Nonrepresentational theories try to attend to the “onflow” of everyday life (Thrift 2007).
This attention lays emphasis on the processual registers of experience that exceed our ability to theorize or fully attend to them. Nonrepresentational theories thus openly acknowledge the partial and incomplete nature of our accounts of the world and so act against the reductive “vampirism” inherent in trying to fix the world within particular structures, models, orders, and frameworks posited by the researcher (Dewsbury et al. 2002). This acknowledgement has been manifest both in the required modesty of the researcher in the claims they can make about the world, and also in the way these accounts are produced and presented, as can be seen, for example, in Derek McCormack’s “A Paper with an Interest in Rhythm” (2002). Here McCormack deploys performative modes of writing in an attempt to animate the paper’s discussion of his encounter with a specific somatic movement practice and so avoid representing this practice in an overly fixed way. Inspired by developments in the performing arts, such work aligns itself with a sort of experimentalism that does not shy away from providing an open-ended account of the world.

Nonrepresentational theories have highlighted the importance of the pre-individual to such a process-based understanding of life. As such, this work has been concerned with the practices of subjectification that bodies undergo rather than with already formed subjects or demarcated identity positions. This subjectification arises out of the world being composed of a complex ecology of human and nonhuman things – objects, people, technologies, texts, ideas, discourses, norms, and so on – that are perpetually encountering and shaping one another. This means that nonrepresentational theories are often interested in the human body and its coevolution with the world (see Bodies and embodiment). As can be seen in John Wylie’s auto-ethnographic accounts of walking in the landscape, and particularly in the ways that embodied registers of experience and subjectification are foregrounded within these, the body is not counted as separate from the world but rather it is argued that the human body-subject is an emergent product of its interactions with the environments it moves through.

Being interested in the relations between various bodies/things leads to another key area of interest for nonrepresentational theories: Affect, sensation, and the precognitive aspects of embodied life. This has been a key area of development for nonrepresentational theories and a particularly contentious one given the relationship of this to other key areas of geographic inquiry, such as emotional geographies. While many interpretations exist, here affect does not refer to a personal feeling, but rather to a “pre-personal intensity” that corresponds to the evolving capacities of a body to act and be acted upon. Significant here is that capacities to affect and be affected emerge in the relations themselves, in the “in-between,” and so do not belong to the subject and/or object prior to that encounter. This emergent nature of affect means that affective registers of experience can become an explicit target in the design of various forms of social space. As James Ash has shown in the context of video games, for example, game designers are able to manipulate the affective experience of gamers by experimenting with the spatial and temporal characteristic of the game. Altering the responsiveness of controls, the speed of the “re-spawning” of characters, and so on, can then produce and/or suppress certain affective – and so relatively unreflected upon – interaction with the architecture of the game “world.”

Nonrepresentational theories also acknowledge the vibrancy of matter and so, rather than viewing the nonhuman as an inert supplement to the human world, give a more equal weighting to both the human and things.
(Anderson and Wylie 2009) (see Geophilosophy). Such theories then view the world as a multiplicity of heterogeneous networks and connections needing to be maintained, which in turn decenter the Cartesian notion of agency as belonging solely to the human. Anything from the microscopic infectious pathogens discussed by Steve Hinchliffe and others to the illuminated atmosphere present in urban spaces discussed by Tim Edensor have the capacity to impinge significantly upon the unfolding of human (and more than human) life. In addition to the strong influence of the work of Actor-network theory and “New Materialisms” here, work on Object Oriented Ontology and Speculative Realist philosophies, which seeks to pay attention to the existence and interaction of objects (or “things”) outside of the necessary presence of a human subject, has recently played an increasingly prominent part.

Finally, nonrepresentational theories have been concerned with the ways in which politics and ethics can be considered in light of the shifted and shifting points of reference outlined above. In particular, nonrepresentational theories have sought to consider how difference can be made and new ways of being in the world can be achieved when the parameters of the political and the ethical are expanded beyond solely representational realms (Anderson and Harrison 2010). Thus far, work has focused on notions of the event as a key site for change to emerge and considered how the agency of matter beyond the human is, in part, constitutive of human difference. Ultimately, the key theme here is uncertainty – over what counts and what might come to be – rather than the presentation of a legislative political program. For ethics, there has been a similar shift away from a normative understanding of what constitutes (im)proper conduct towards others towards a more “situated” ethics. Rather than stipulating how we should act towards others, emphasis is laid upon a consideration of how we negotiate the present with an ethos and a sensibility that is attentive to how we affect, and are affected by, the world and others.

The pursuit of these themes has touched a number of nerves within the discipline. One of the clearest indicators of the impact non-representational theories have had is the range, and vehemence, of the critique and criticism that have been made of it. For example, there have been suggestions that nonrepresentational theories are guilty of an ethnocentric universalism in the way they consider (or rather fail to consider) social difference through their movement away from collective identity-based forms of politics. Further, others have argued that nonrepresentational work embodies a form of masculinist rationality in its movement away from subject-based emotions and personal feelings towards a more reasonable language of “engineering.” Here it has been suggested that talking in terms of networks, junctions, pipes, cables, relays, and so on in the tracing of affective relations risks distancing research from important issues of being “human” – namely, our ongoing constitution as specific and often categorized “emotional subjects” in social spaces. Finally, it has also been said that nonrepresentational theory’s account of agency and subjectivity, and the inherent anti/posthumanism of such accounts, does not leave sufficient space for the acknowledgment of differing capacities to engage in political action or provide a ground from which such action can take place.

Arguably though, these critiques are products precisely of the challenge nonrepresentational theories present to the way in which geographic research is undertaken and what counts for knowledge. Nonrepresentational theories do not necessarily deny the importance of identity politics, collective action, and so on. In fact, work here has made substantial contributions to
understanding how racialized bodies are constituted, for example. Nonrepresentational theories also point, however, to the emergence of a range of new political techniques and technologies that act on pre-individual registers of thought and action. To do so requires the development of new vocabularies for expressing these emerging politics and new arts of political action for intervening into them. The extent to which such development is deemed sufficiently “political” though is an entirely different question.

SEE ALSO: Actor-network theory; Affect; Bodies and embodiment; Geophilosophy; New cultural geography; Poststructuralism/poststructural geographies; Representation and presentation

References


Norway: Norsk Geografisk Selskap (NGS) (Norwegian Geographical Society)

Founded: 1889
Location of headquarters: Oslo
Website: www.geografisk.no
Membership: 126 (as of December 31, 2013)
Leader: Tonje Blom (Oslo), Svein Frisvoll (Trondheim)
Contact: tonje.blom@elverbakken.vgs.no, styret@geografisk.no, svein.frisvoll@bygdeforskning.no

Description and purpose

The NGS is an association for geographers, students, and others interested in the discipline of geography. The society disseminates knowledge concerning geographical topics through meetings for members and others interested in geography. The objective of the society is to provide a collective forum for members and to contribute to the exchange of knowledge and experiences between various institutions where geographers work in teaching, research, administration, and other practical occupations. The society aims to further geographical knowledge and arouse interest for the discipline by arranging seminars, lectures, and field excursions for members and by publishing the Norsk Geografisk Tidsskrift – Norwegian Journal of Geography.

Journals or major publication series


Current activities or projects

Both the Oslo and Trondheim branches of the NGS arrange thematic meetings with lectures and discussions on current topics of interest, for example, climatic warming, permafrost, physical planning, polar explorations, and ongoing geographical research projects. The Oslo branch arranges an annual field excursion. The society organizes a national geographical conference every second year aimed at researchers, lecturers, teachers, planners, environmental administrators, cartographers, statisticians, and other professions. The society has recently published in Norwegian a history of its activities from its foundation in 1889 until 2000 (Jens Fredrik Nystad, 2012, Det Norske Geografiske Selskab 1889–2000, Oslo: Fram).

Brief history

The NGS was founded in 1889 in a wave of national enthusiasm greeting the return of Fridtjof Nansen (1861–1930) to Oslo after skiing across Greenland. It became an important arena of support for subsequent Norwegian exploration and research in polar regions and other parts of the world. Expedition reports and plans were presented at the society’s meetings, often with royalty and members of the government present. The society’s first chairman was Henrik Moen (1835–1916), founder of the Norwegian Meteorological Institute in 1866. Besides Nansen, prominent members included polar explorers C.A. Larsen (1860–1927), Carsten Borchgrevink (1864–1934), Roald Amundsen (1872–1928), Gunnar Isachsen

http://www.tandfonline.com/toc/sgeo20/current
NORWAY

(1868–1939), and Otto Sverdrup (1888–1957). Other explorers included Helge Ingstad (1899–2001) and Thor Heyerdahl (1914–2002). A committee of the society drafted Norway’s first Nature Protection Act in 1910. The society’s research promoted Norwegian sovereignty of Svalbard (annexed 1925), claims to East Greenland (rejected by the International Court of Justice in 1933), and annexation of islands and Queen Maud Land in Antarctica in the 1930s. By 1960, the era of exploration was over and geography became more professionalized and specialized. The Norwegian Geographers Association amalgamated with the society in 2000.

Submitted by Michael Jones
By the middle of the nineteenth century, marine biologists began to realize that the ocean was inhabited by species that appeared to be peculiar to certain geographic areas. This realization stimulated an interest in the regions that could be identified by the presence of endemic species, that is, species that were confined to a given area and found nowhere else. Our present knowledge of marine biogeography and evolution is based primarily on investigations of the biota of the continental shelf and the epipelagic zone, each extending from the surface to a depth of 200 m. This relatively shallow area is penetrated by sunlight, generates most of the primary productivity, and supports the highest level of biodiversity.

Our modern concept of marine biogeographical regions began with the work of James Dwight Dana, an American geologist, mineralogist, and naturalist. As a young man, Dana joined the United States Exploring Expedition to the South Seas. He served four years (1838–1842) as a geologist but also undertook much of the zoological work. As the result of his discoveries on the corals and crustaceans, he was convinced that sea surface temperature was the important factor that determined distributional patterns. Dana used isocrymes (lines of mean minimum temperature for the coldest month) to explain the geographical separation of species groups, and published (1853) an Isocrymal Chart (map) that illustrated the worldwide distribution of marine animals. He was the first to observe that the latitudinal distribution of marine animals was restricted by the cold of winter, not the average temperature. Dana’s chart laid the foundation for the biogeographic regions that are recognized today.

Edward Forbes, the English naturalist who first explored the depth distributions of marine organisms, also drew a map of horizontal distributions that was published (Forbes 1856) in Alexander K. Johnston’s The Physical Atlas of Natural Phenomena. In it, Forbes divided the world’s oceans into 25 provinces located within a series of nine “homoizoic belts.” In his posthumous book, The Natural History of European Seas, Forbes (1859) pointed out the evolutionary significance of zoogeographic provinces. He observed that: (i) each province was an area where there was a special manifestation of creative power, and that the animals originally formed there were apt to become mixed with emigrants from other provinces; (ii) each species was created only once, and individuals tended to migrate outward from their center of origin; and (iii) provinces, to be understood, must be traced back, like species, to their origins in past time.

Samuel P. Woodward published his classic work, Manual of the Mollusca, (1851–1856) but the part devoted to geographical distribution did not appear until 1856. Woodward described 18 marine provinces and located them on a world
map together with the names of typical genera. He was the first to characterize provinces by numbers of endemic species and suggested that endemics constituted more than 50% of the species in each province. Several of Woodward’s molluscan provinces are identical to those that are currently recognized. The next comprehensive work on marine distributions was Sven Ekman’s 1953 book *Zoogeography of the Sea*. This work was an updated English edition of a book originally published in German. It contained a review of all research accomplished until that time. Ekman recognized four climatic zones (warm, warm-temperate, cold-temperate, cold) under various names, and within the zones, he divided the marine world into regions and sub-regions. He discussed regional relationships in detail, paid particular attention to endemism, and speculated on the history of the various faunas. For the following 20 years and beyond, Ekman’s book was the primary reference for marine zoogeography.

The next book to analyze faunal distributions was *Marine Zoogeography* (Briggs 1974). The author’s intent was to continue the work of Ekman, but the enormous amount of new information needed to be organized in a more systematic manner. Within each of the four climatic zones, one or more biogeographic regions were described, and each region consisted of one or more provinces. Provinces were defined as geographic areas containing a minimum of 10% endemic species. So provinces, as Forbes pointed out, were geographic units with evolutionary significance. It was determined that the 20°C isocryme marked the latitudinal limits of the tropical faunas. The faunas of the adjoining warm-temperate provinces were viewed as primarily of tropical derivation. The major zoogeographic barriers and their influence on migration and colonization were identified. Although this volume was based primarily on fish distribution, the described patterns proved to be applicable to several invertebrate groups.

**New developments**

In recent years, numerous volumes on biological oceanography, marine biology, and general biogeography have been published but little attention was paid to the importance of marine provinces and the barriers that separate them. However, continued collecting and analysis (systematic and phylogeographic) produced information of considerable value. The continued migration of species from high-diversity centers to less diverse areas appeared to be a global phenomenon. But phylogeographic research revealed that a few reef fishes had been able to migrate toward higher diversities. It became practical to distinguish between hard and soft (semipermeable) biogeographic barriers; the latter proved to be important for the stimulation of parapatric or sympatric speciation. Accommodation was proposed as a theory to explain the acceptance of invasive species by native species that occupied the appropriate habitat. A re-evaluation of existing provinces, description of new provinces, and world maps illustrating all the marine regions and provinces were published (Briggs and Bowen 2012).

As the various provinces became better known, it became apparent that speciation, resulting in the development of high diversity, was not taking place evenly throughout each province, but was concentrated in smaller areas, called centers of origin, within some of the large provinces. In the case of the Coral Triangle (CT), a small part of the huge Indo-Polynesian Province, its diversity concentration had been known for many years. Diverse opinions had been expressed about the origin of the CT. Did it represent an
accumulation of species formed elsewhere, an overlap of Indian Ocean and Pacific biotas, or a center of evolutionary innovation and distribution? Recent research appears to have solved the problem. The CT has apparently passed through three developmental stages: (i) origination via migration from the Tethys Sea to the East Indies, (ii) speciation in place, and (iii) outward dispersal during the past 10 million years. There is also a “feedback” mechanism by which species in outlying provinces can occasionally penetrate and add to the diversity of the CT, as well as to the Caribbean center (Bowen et al. 2013).

In the Caribbean Province the highest diversity among the invertebrates occurs from Cuba along the Antilles arc to Colombia and Venezuela. This may be true for the fishes as well. In the Eastern Tropical Pacific, shore fish diversity peaks along the mainland coast of Costa Rica and Panama as well as at the tip of the Baja Peninsula. In the North Pacific, there are two centers of origin, one on each side of the Bering Sea.

The world’s greatest diversity of marine organisms, at least within the corals, molluscs, fishes, and mangroves, occurs within the CT. The location of world’s highest diversity of reef fishes, at least 1426 species, exists in the vicinity of the Bird’s Head region of Indonesia at the extreme western end of New Guinea. In their book, Reef Fishes of the East Indies, Allen and Erdman (2012) identified Cenderawasih Bay, West Papua, as a living laboratory of evolution due to its numerous short-range endemics. It has become clear that the CT does not comprise a single center of origin but contains multiple centers. A general characteristic of centers of origin is high species diversity, indicating an enhanced level of competition which produces dominant species that can spread into less diverse areas. As such species become accommodated in peripheral habitats they increase community biodiversity, stability, and productivity. At the same time, speciation has continued to take place in other parts of the world ocean resulting in a general buildup of biodiversity that has been apparent since the early Pleistocene. Although many marine species have small populations due to overfishing and habitat destruction, there have been very few extinctions. Therefore, despite speculations to the contrary, biodiversity loss in the sea has yet to occur and there is still opportunity to rescue threatened populations through prompt application of conservation measures.

The deep ocean

Below the continental shelf, the benthic deep sea consists of the continental slope (200–4500 m), abyssal plain (4500–6000 m), and the hadal zone of the trenches (6000–11000 m). The pelagic parts of the deep sea are often called bathyl or bathypelagic. The distribution patterns of organisms in these environments are poorly known, so that it has not often been possible to positively identify biogeographic relationships. About 84% of the ocean floor exceeds a depth of 2000 m. The enormous size of the deep sea, and the difficulty of adequately sampling its fauna, has prevented an overall characterization of biogeographic regions and provinces. Although generalizations remain elusive, many taxa appear to be broadly distributed across the deep-sea floor.

Broad distributions are apparent not only among soft-bottom taxa across the abyssal plain, but also in highly patchy environments such as hydrothermal vents and seamounts. Species in methane seeps and large food-fall (whale carcass) communities may have more restricted ranges but undersampling is a problem. At the generic level, evidence indicates high levels of
cosmopolitanism across all habitats. Many taxa display an amazing ability for larval dispersal, which may be augmented by extended larval development in cold deep waters. Dispersal apparently increases with depth but there are barriers consisting of currents with contrasting temperature and salinity, land masses, and other topographic structures. These barriers might define general biogeographic patterns, although many of them are semipermeable for numerous species.

Some distributional patterns bear the signature of historical events as evidenced by multiple points of evolutionary origin and the apparent role of deep-water oxygen concentrations. All current evidence points to a deep-sea fauna comprised of taxa, neither entirely old nor entirely young, filtered through many climatic fluctuations. Far less than 1% of the deep sea has been sampled, which is the primary impediment to the discernment of biogeographic patterns. Furthermore, the sampling of the deep-sea floor done so far has been concentrated in the Northern Hemisphere in proximity to American and European oceanographic institutions, preventing widespread biogeographic syntheses. Rare species, especially those that are narrowly distributed, are difficult to sample in the deep sea. This may lead to an overestimation of the contribution of broadly distributed species to biodiversity.

Future deep-sea research needs to concentrate on sampling over broader geographic areas, on phylogeography in order to clarify timing and points of origin for deep-sea invasions, and on species identifications via systematic and molecular studies than can characterize each deep-sea habitat. This kind of research, compared to that in shallow water locations, is relatively expensive and time-consuming but necessary in order to obtain an adequate knowledge about an environment that extends over 66% of the Earth’s surface.

**SEE ALSO:** Biodiversity; Dispersal, diffusion, and migration; Zoogeography

**References**


**Further reading**


That the ocean flows, and not just from the immediate effect of the local wind, has probably been recognized ever since humans have lived near coasts with tides. The Ancient Greeks were aware of strong currents in the Atlantic, beyond the Strait of Gibraltar, over 2500 years ago, while Polynesian seafarers implicitly understood the interactions between winds, waves, and currents, enabling them to navigate across open stretches of the Pacific for thousands of years. However, it wasn’t until the Age of Discovery that knowledge of surface ocean currents began to be compiled in a systematic way, and not until William Dampier’s 12-year circumnavigation of the globe (1679–1691) that the principal features of surface ocean circulation were first mapped. With Benjamin Franklin’s investigations of the Gulf Stream in the eighteenth century, Matthew Maury’s wind and current charts of the second half of the nineteenth century, and finally the global Challenger Expedition of 1872–1876, oceanography came of age. Developing understanding of the theory of ocean circulation has occurred ever since, and the World Ocean Circulation Experiment (1990–1998; Ganachaud and Wunsch 2000) produced a global perspective of ocean properties and currents to act as a baseline for interpreting the influence of climate change into the future. Nevertheless, discoveries continue being made, particularly in the polar seas and of the thermohaline circulation in the deep ocean, so this entry is a snapshot of our knowledge of current oceanic circulation.

Surface circulation

The surface oceanic circulation is largely forced by a mixture of surface winds and atmosphere–ocean exchange of heat and freshwater. The effects of tides are largely ignored here (but see the thermohaline section) as they lead to minor consequences for much open ocean circulation. The physical stress of the wind on the surface of the ocean – the wind stress – is the most obvious direct atmospheric forcing; however, it does not lead to direct acceleration of oceanic currents. Wind-driven currents actually come about due to a combination of forces in the ocean. The first of these is the direct effect of the wind on the surface waters. During Fridtjof Nansen’s expedition into the Arctic during 1892–1895 he noticed that the sea ice in which his ship, the Fram, was trapped moved 20–45° to the right of the wind direction. Ten years later, Vagn Walfrid Ekman, a Swedish oceanographer, developed what is now known as Ekman theory to explain this observation through a surface force balance between the wind, the Coriolis force, due to the Earth’s rotation, and friction. However, the frictional effect is transmitted layer by layer down through the upper 100 m or so of the ocean – the Ekman layer – leading to a spiralling velocity with depth. The mean flow in this layer is due to the overall balance between the wind stress and the Coriolis force accumulated with depth, leading to the Ekman transport, a net flow in this layer 90° to the right of the surface wind direction in the Northern
Hemisphere and to the left in the Southern Hemisphere. Ekman transport is a fundamental process affecting flow, and sea level, within the upper ocean (Figure 1). It underlies the reason for the existence of circulatory features of the ocean circulation – the oceanic gyres – shown in Figure 2, as well as the extensive zones of coastally upwelled water found along the eastern boundaries of the subtropical gyres.

However, in producing divergence and convergence of water and sea level slopes, pressure gradients are set up in the ocean that tend to add an additional force in the upper ocean. The final force balance from wind, Coriolis force, and wind-induced pressure gradients leads, on the large scale, to net upper ocean flow slightly to the right of the wind in the Northern Hemisphere. It also results in the main features of the global surface ocean circulation (Figure 2). Thus subtropical gyres are maintained in each ocean basin through this force combination by the overlying atmospheric circulation. Similarly, in the more poleward latitude band of 50–70°, oppositely rotating subpolar ocean gyres, always strongly geographically controlled by basin coastline shape, are driven by a constant succession of traveling low pressure systems. In the Southern Hemisphere, because all the ocean basins are linked by the Southern Ocean, with strong westerly winds above, a global-scale pressure gradient is set up between the high sea levels of the subtropics and the low sea level towards Antarctica. The latter is due to a combination of Ekman transport northwards and the dense, cold waters in the subpolar regions. This global-scale pressure gradient leads to a strong eastward flow called the Antarctic Circumpolar Current (ACC) encircling the Earth between approximately 50 and 60° S and extending 2000–2500 m below the surface. It is one of the strongest ocean currents on the planet, with a mean flux of approximately 130–140 Sv (1 Sverdrup = 1 Sv = 10⁶ m³s⁻¹).

The ACC meanders in places, due to interaction with seafloor topography and the narrowing of the flow as it passes through Drake Passage; it plays a significant role in the thermohaline circulation.

In the tropics, the south-north transition from westward-flowing South Equatorial Current through eastward-flowing Equatorial Countercurrent back to westward-flowing North Equatorial Current is also driven by the combination of meridional variation in wind stress and the sea level pressure gradient, modulated by seasonal shifts in the Intertropical Convergence Zone. Along the equator, the absence of the Coriolis force leads to a strong, equatorially
confined current approximately 100–200 m below the surface, along the thermocline, called the Equatorial Undercurrent, predominantly maintained by the balance of the westward wind stress and the eastward pressure gradient set up within the ocean by a resulting sea level slope. In the Indian Ocean, because of the dramatic seasonal wind reversal associated with the monsoon, this is only a boreal winter feature.

The subtropical oceanic gyres have an additional characteristic, namely that the poleward flow on their western side is largely confined to a narrow (≈100 km), deep (≈1000 m), and strong (≈2.5 ms$^{-1}$) boundary current. In the North Atlantic this is well known as the Gulf Stream, extending past the Florida Keys along the United States’ East Coast, separating from the shore near Cape Hatteras (≈40°N) before crossing the Atlantic towards northern Europe. However, each subtropical gyre possesses a similar feature. These western boundary currents occur because of the additional constraint that the water in a gyre needs to conserve angular momentum, in terms of a quantity better known in oceanography as potential vorticity, as both the Coriolis force and the rotation of the gyre change as water travels around the gyre. Where western boundary currents leave the coast and move into the ocean proper there are key locations for smaller scale (up to a few hundred kilometer) eddy formation and current meanders, supported by baroclinic instability along the strong horizontal temperature gradients typically found here. These features are major mechanisms by which water of different temperature and salinity properties is mixed horizontally, often accompanied by pronounced changes in marine biological productivity.

The pressure gradient forces set up by wind-driven sea level changes are further modified by the impact of heat, precipitation, and evaporation on the upper ocean density. Warming
the ocean or adding freshwater through rainfall or river inflows will lower ocean density, and allow the sea level to rise. Cooling the ocean or evaporation will raise ocean density, leading to a lowering of sea level. These changes will further modify the observed upper ocean circulation, and we will see their fundamental importance for driving the global thermohaline circulation in the next section.

Thermohaline circulation

The thermohaline circulation, often known as the Meridional Overturning Circulation, is the main oceanic circulation by which water is transported on a global scale. Components link all the ocean basins, allowing water to be mixed horizontally and vertically to such an extent that no part of the ocean contains water that was last exposed to the surface more than about 1500 years ago. The thermohaline circulation can also be thought of as the way by which the surface ocean circulation, which we have seen is predominantly driven by winds, connects to the deep ocean circulation, largely driven by density differences, through processes that transform the temperature and salinity properties of the water between extremes. The principal driving force of the thermohaline circulation is therefore the densification of surface water, allowing it to sink to a deeper surface of neutral buoyancy. This densification can occur in several ways. Heat can be extracted from seawater to the atmosphere by either latent or sensible heat transfer, cooling, and densifying, the water. Water can be evaporated from the ocean surface, leaving saltier, and hence denser, water behind. Finally, during the formation of sea ice at the ocean surface, a significant proportion of the salt from the freezing seawater is expelled into the water remaining, increasing salinity, and hence, density.

If each of these mechanisms is sustained, the upper ocean water will be made dense enough to convect to a deeper level of neutral buoyancy. When deep water is formed by this mechanism the convection does not occur spontaneously over the whole of a region where the particular process acts, but in preconditioned plumes, and concentrated in small (≈1 km) chimneys. These lead to mixing with the surrounding water and the effective transfer of properties from the surface to depth (Marshall and Schott 1999).

These regions of convection tend to be regional rather than basin-scale in extent, as much of the near-surface ocean is stratified all or most of the time. The deepest waters are formed where salinity is the driver of convection. At locations around the Antarctic continental shelf, but particularly in the Weddell and Ross Seas, sea ice formation leads to both cooling and salinification of water during the autumn and winter. This dense water sinks, and cascades off the shelf into the abyssal ocean. It then spreads north at depth, as cold, relatively salty, Antarctic Bottom Water (AABW). This is particularly noteworthy in the Atlantic, where deep water properties suggest some AABW penetrates into the mid-latitude North Atlantic from an original production of approximately 27 Sv. Cooling and salinification by sea ice formation also occurs on the Eurasian continental shelf of the Arctic, leading to the deepest waters of the Arctic Ocean, and a component exiting at depth through the Fram Strait into the Greenland-Norwegian Sea. Raised salinity is also crucial to deep water formation in the Mediterranean and Red Seas, but because of evaporation. The Mediterranean has a background salinity more than 10% greater than that of the Atlantic, so winter cooling by strong northerly winds leads to deep water formation in all its three main basins. About 1 Sv of this water leaves the Mediterranean through
the deeper parts of the Strait of Gibraltar, supplying a layer of relatively warm and salty water at mid-depths (≈1000–1500 m) in the central North Atlantic. Through winter mixing in the Atlantic this can help precondition the already salty upper ocean waters of the Atlantic for later convection in the basin’s northern seas. In contrast, the Red Sea water is made dense by intense summer evaporation, leading to warm salty water crossing the shallow sill into the North Indian Ocean, and spreading out, also at depths of around 1000–1500 m.

Cooling of surface waters is the dominant producer of deep and intermediate water in other regions of the globe. In the North Atlantic, principally in the Greenland, Irminger, and Labrador Seas, winter cooling of the relatively salty water of the Atlantic leads to deep water formation to a depth of up to several thousand meters, depending on the year and location. This, with an admixture of Arctic water, becomes North Atlantic Deep Water (NADW). About 20 Sv of this water is created, although the flux south at the only long-term monitoring transect of 26°N varies strongly seasonally and interannually (Srokosz et al. 2012). NADW spreads south in the Atlantic, with a stronger western boundary current at depth, eventually entering the Southern Ocean and mixing with water of other origins to form Circumpolar Deep Water, and hence entering the Indian and Pacific Oceans.

Other regions where large-scale intermediate depth water formation occurs are in the Southern Ocean, in the Antarctic Polar Front Zone between 50 and 60°N, where Antarctic Intermediate Water is formed, and in the North Pacific, north of the Kuroshio Current, where North Pacific Intermediate Water forms. Both of these water masses form due to winter cooling of relatively fresh seawater, beneath mid-latitude storm belts, and spread equatorward at depths of approximately 1000 m.

The deep waters of the global ocean described above eventually mix back to the surface ocean, to be drawn into the surface return flow of the thermohaline circulation. This mixing proceeds in a number of ways. There is a weak upward diffusion of water over the global oceans, but this is at an order of magnitude too small to explain the 1500-year residence time of water in the thermohaline circulation. However, much stronger mixing is concentrated in a number of geographically confined regions to compensate for the weak background mixing. A substantial amount of this stronger mixing is forced from the surface by the strong winds blowing over the weakly stratified Southern Ocean, but leading to mixing through eddies to considerable depths (Marshall and Speer 2012). Another substantial component of localized mixing occurs in the deep ocean, in regions of rough topography. A range of interactions between the ocean flow and this topography leads to enhanced mixing: tidal amplification through deep passages and shallower seas, eddy shedding, and the generation of internal waves. Note that while the oldest water in the global ocean is in the North Pacific, the processes leading to upwelling of the deep component of the thermohaline circulation occur throughout deep water flow, with especial concentration in the Southern Ocean.

The return flow of the thermohaline circulation acts to link the upper ocean circulation of the three main basins of the Pacific, Indian, and Atlantic Oceans. From the Pacific there is significant flux of water both westward into the Indian Ocean between the Indonesian islands, and eastward as part of the ACC, through Drake Passage into the South Atlantic. The former, called the Indonesian Throughflow, varies significantly from year to year (≈7–15 Sv), and supplies warm and fresh water to the Indian Ocean. The latter moves cold water between the Pacific and Atlantic. Most of this remains in the
OCEANIC CIRCULATION

ACC, but it is believed that approximately 6 Sv of this cold, fresh water from the Pacific enters the South Atlantic circulation as part of the East Falklands Current (Speich, Blanke, and Madec 2001).

The Indonesian Throughflow feeds into the Equatorial Currents of the Indian Ocean, helping fuel the western boundary current of the South Indian Ocean—the Agulhas Current—that flows southward along the southeastern coast of Africa. Most of the water in this current forms part of the South Indian Ocean subtropical gyre and so retroreflects towards the east, to the south of Africa. However, a component, perhaps 11 Sv, is entrained in the stream of warm core eddies that are shed from the tip of South Africa during this process (Speich, Blanke, and Madec 2001). These move northwestward, and are entrained in the equatorward arm of South Atlantic subtropical gyre circulation. This episodic process therefore acts as a warm water route by which surface waters re-enter the Atlantic following upwelling from the deep ocean further afield.

Once surface waters return to the Atlantic, a pathway exists whereby some of the water can be transported north, first in the South Atlantic subtropical gyre, then westward across the Atlantic in the Equatorial Current system, and finally northward into the Caribbean to be entrained into the beginning of the Gulf Stream off southern Florida. The Gulf Stream, and then the North Atlantic Drift crossing the Atlantic further north, returns water to the main North Atlantic convection regions, allowing the thermohaline cycle to be repeated. As a southerly component of the North Atlantic Drift, the Azores Current allows approximately 1 Sv of this water to enter the Mediterranean at the surface, also providing water to the Mediterranean arm of the thermohaline circulation.

The future of the study of the oceanic circulation

Much, although not all, of the description of the oceanic circulation given here has relied on the burgeoning of observational studies of the ocean over the last few decades. This is continuing apace. Over the last decade the Argo float program has led to an unprecedented knowledge of the seasonal and interannual variation of temperature, salinity, and currents in the top 2000 m of the ocean, and this looks likely to be maintained in the long term. Satellites have provided global measurements of sea surface temperature, sea level, and sea ice cover, amongst other variables, over the past few decades. Since 2011 it has also become possible to measure surface salinity from space. There are also a number of long-term programs to measure key fluxes in the global oceanic circulation. However, numerical, mathematical models of the ocean circulation have also become an important tool for studying the ocean, since their first development in the 1960s. Ocean models provide the opportunity to understand flows where observational data are not readily available, and to simulate interannual variability, which is still poorly observed in most locations. Ocean, and climate, models also permit study of how oceanic circulation is likely to respond to current and future changes in greenhouse warming. For example, for some years, and reaffirmed by the most recent model results from the 2014 Intergovernmental Panel for Climate Change Working Group 1 Scientific Report, climate models have been predicting a steady decline in the strength of the thermohaline circulation over the next century, which would have significant effects on the rate of climatic change, particularly over the North Atlantic and western Europe, but also further afield. It is known from study of paleoclimate and paleoceanographic records that the ocean circulation has changed abruptly on
occasion in the past; models allow the possibility of advance warning of such abrupt climate change. The World Ocean Circulation Experiment provided a snapshot of oceanic circulation in the 1990s, but environmental change means oceanic circulation will continue to evolve into the future.

**SEE ALSO:** Atmospheric/general circulation; Global climate models; Monsoons; Oceans and climate; Polar climates; Sea ice, ice drift, and oceanic circulation

**References**


**Further reading**


Oceans and climate

Weiqing Han
University of Colorado at Boulder, USA

Ocean basins

The World Ocean occupies about 71% of the Earth’s surface (Figure 1). Geographically, the World Ocean is divided into four individual ocean basins by continents. In descending order by area, they are the Pacific, Atlantic, Indian, and Arctic Oceans. The first three include their corresponding Southern Ocean region. While each basin has its own geographic distinction, all basins own some common features. They are the continental shelf (a shallow ledge next to the edge of the continent), the continental slope (a slope extending from the shelf toward the deep sea), the continental rise (a gentle slope at the foot of the continental slope), and the abyssal plain (ocean floor of the deep sea). Approximately 76.7% (in volume) of the World Ocean is occupied by the abyssal plain, 7.4% by continental shelf, and 15.9% by continental slope and rise. The abyssal plains are also filled with ridges and trenches. The deepest trench is the Mariana trench in the western Pacific, which is about 11 km deep.

The Pacific Ocean is the largest of all. It spans a zonal distance of approximately 20 000 km from the Strait of Malacca to Panama in the tropics, and a meridional extent of over 15 000 km from the Bering Strait to Antarctica. With all its adjacent seas (a sea is a division of an ocean or a large body of salt water partially enclosed by land) it covers \(178 \times 10^6\) km\(^2\) and represents 40% of the World Ocean surface, equivalent to the area of all continents. Without its southern ocean part the Pacific occupies \(147 \times 10^6\) km\(^2\), about twice the area of the Indian Ocean. This vast ocean facilitates strong air–sea interaction, and is the home for the ocean–atmosphere coupled climate phenomenon: the El Niño Southern Oscillation (ENSO).

The Atlantic Ocean differs from the Pacific in that it has a full meridional extent, from the Arctic Ocean to the Antarctic continent. In comparison, its zonal extent is small, with the largest distance approximately 8300 km between the Gulf of Mexico and the coast of northwest Africa. The full north-south extent exposes the ocean to the cold and dry air in the northern North Atlantic during winter, which cools the ocean and results in sea ice formation. Since the freezing point of saline seawater is \(-2^\circ\)C and that of fresh water is 0°C, fresh water freezes up first, leaving the ocean more saline after sea ice forms. The colder and more saline surface water increases surface density, which induces wintertime oceanic convection when heavier surface water sinks down and forms deep water. This comprises the sinking branch of the Atlantic and global thermohaline circulation (Figure 1). The Atlantic has the largest number of adjacent seas. The Mediterranean Sea produces Mediterranean water mass. Including all adjacent seas, the Atlantic Ocean covers \(106 \times 10^6\) km\(^2\). Variability of the thermohaline circulation can have a large influence on global climate (see “Ocean circulation,” below).

The Indian Ocean is small in comparison with the Pacific and the Atlantic. Different from the other two oceans, the Indian Ocean is bounded by the Asian landmass to the north, and its
northern boundary is located in the tropics. One consequence of this geographic location is that the Indian Ocean is subject to the strong influence of the monsoon, a seasonal reversing wind accompanied by changes in precipitation due to an uneven heating rate between land and ocean. Its meridional extent is 9600 km, from the Antarctic to the inner Bay of Bengal, and its zonal extent is 7800 km, from South Africa to West Australia. Including the southern ocean region, it covers $74 \times 10^6$ km$^2$.

The Arctic Ocean covers the Arctic north polar region. It is the smallest among the four and covers an area of about $14 \times 10^6$ km$^2$. Much of the ocean is covered by sea ice. The sea ice covered area varies with season, with maximum sea ice extent occurring during fall and winter. The Arctic Ocean is connected to the North Atlantic through the Denmark Strait between Greenland and Iceland, via the opening between Iceland and Europe, and to a lesser degree through the Canadian Archipelago. It is also connected to the Pacific through the narrow and shallow Bering Strait (45 m deep and 85 km wide). On the annual mean, seawater from the North Pacific flows into the Arctic, and subsequently exits the Arctic into the North Atlantic. The Arctic Ocean also exports sea ice into the Atlantic Ocean.

The three major oceans are also interconnected. The Pacific is linked to the Indian
Ocean via the Indonesian Throughflow near the Indonesian Archipelago, where fresher and warmer upper-ocean water from the Pacific enters the South Indian Ocean, flows westward, and part of it enters the Atlantic around the southern tip of Africa. The Indonesian Throughflow is an integral part and a choke point of the global ocean conveyor circulation (Figure 1).

Temperature

Temperature of the ocean varies at a wide range, from nearly freezing point at the poles to above 32°C in some regions of the tropical oceans, such as the Indo-Pacific warm pool of the tropical east Indian and west Pacific Oceans (Figure 2). Various instruments have been used to measure the temperature of the surface and subsurface ocean. They include bathythermograph (old method), expendable bathythermograph (XBT), CTD (conductivity, temperature, and depth), and protected/unprotected reversing mercury thermometer on board ships; thermistors on moored buoys, drifting buoys, and profiling floats (e.g., Argo floats). The sea surface temperature (SST) is the temperature close to the ocean surface. The SST values measured by in situ instruments usually represent bulk temperatures at a depth near the surface within the surface mixed layer, where seawater is fairly mixed by wind and turbulence and its density varies little with depth. Depending on the instrument type, the measurement depth of SST can vary from less than 1 m to approximately 20 m. Generally, temperature is high at the ocean surface and decreases with the increase of depth. In the subsurface ocean below the mixed layer there is a layer of water in which temperature decreases more rapidly with depth than it does in the layers above or below. This layer is known as the thermocline.

In recent decades, satellite-derived SST data have become available, including the SST dataset from the Advanced Very High Resolution Radiometer (AVHRR) on board the National Oceanic and Atmospheric Administration (NOAA) satellite since the 1980s, and the SSTs from the National Aeronautics and Space Administration (NASA) Tropical Rainfall Measuring Mission (TRMM)-Microwave Imager (TMI) since 1997. The AVHRR cannot look through clouds, while TMI can penetrate clouds but the data can be contaminated by strong precipitation. Generally, the satellite-observed and in situ measured SSTs agree; however, there are systematic discrepancies, because the satellites detect “ocean skin” temperatures within approximately the top 1 mm, whereas in situ measurements observe bulk temperature below the surface. Very few in situ measurements of the surface skin temperature are made on a regular basis, so the satellite SST data have been calibrated primarily by the bulk SST of in situ observations.

The SST is an important ocean variable that directly affects atmospheric convection (cloud, precipitation) and wind. It is therefore important for climate variability and change. For this reason, the dominant SST pattern and temporal variations are often used to identify natural climate variability modes on various timescales. For example, ENSO, which is an interannual climate mode of the tropical Pacific with a period of 2–7 years, Indian Ocean Dipole (IOD), and Atlantic Multidecadal Oscillation are all characterized by distinct spatial patterns of SST anomalies with temporal variations. They have large impacts on regional and global climate and can cause devastating floods and droughts in various regions around the globe.

Wide-ranging observational evidence shows that the World Ocean has been warming in the past few decades. Earth receives more energy
from the sun than that exits the top of the atmosphere and thus it has gained substantial energy, with most of the energy gain entering the ocean. The global upper ocean (0–700 m) has warmed and thus the upper ocean heat content has increased between 1971 and 2010. It is likely that the upper ocean has also warmed between the 1870s and 1971. The global total ocean (full depth) has stored about 93% of the warming, and the upper 0–700 m ocean has accounted for about 64% (IPCC 2013, section 3.2). The increased concentrations of human-induced greenhouse gases are shown to cause more than half of the observed increase in global average surface temperature from 1951 to 2010. The ocean warming can have various climatic, environmental, and social consequences, such as increased intensity of storm surges in certain coastal areas, sea level rise, coral bleaching, and increased drought or floods in some land areas owing to atmospheric circulation change induced by the ocean warming.

**Salinity**

A primary difference between seawater and pure water is the oceanic salinity. Salinity is defined as the number of grams of dissolved matter in one kilogram of seawater. Ocean temperature and salinity, together with pressure, determine seawater density. Due to the presence of salinity, the freezing point of seawater is lower (\(-2 \degree C\)) than that of fresh water. The oldest method for measuring salinity is to evaporate seawater with known weight and then weigh its residual. Later, the laboratory classical Knudsen method was used to determine the amount of chlorine, bromine, and iodine to give “chlorinity,” and then relate salinity to chlorinity by the equation: salinity = 1.80655 × chlorinity. This method was used until the International Geophysical Year in 1957. More recent methods determine salinity by measuring conductivity of seawater, such as salinity obtained by CTD. Salinity computed through conductivity is more closely related to
the actual dissolved constituents than chlorinity is, and is more independent of salt composition. Because the measurement is based on conductivity rather than mass, the Practical Salinity Scale (unitless) was introduced in 1978. In 2010, the Thermodynamic Equation of Seawater 2010 (TEOS-10) was introduced, which uses absolute salinity $S_A$ (mass fraction of salt in seawater, expressed in g/kg) as opposed to Practical Salinity. As SST, sea surface salinity measured by in situ instruments is the salinity close to the ocean surface. In June 2011, NASA’s Aquarius instrument aboard the Argentine SAC-D satellite was launched. It provided weekly global maps for ocean surface salinity, the skin salinity of the top approximately 1 cm depth, from August 2011 to June 2015.

Sea surface salinity is often referred to as nature’s rain gauge for detecting the global water cycle. In the open ocean, its value generally ranges from 32 to 37. In coastal oceans, salinity can be much lower adjacent to river mouths or melting land ice areas, and much higher near a concentration basin such as the Red Sea, the Persian Gulf, or the Mediterranean Sea. In tropical and subtropical open ocean basins, salinity is generally low in regions where precipitation is strong and exceeds evaporation, such as the Intertropical Convergence Zone (ITCZ), and high where precipitation is low and smaller than evaporation, such as the subtropical Atlantic, the subtropical Pacific, and the Arabian Sea of the Indian Ocean (Figure 3). As a result, changes in precipitation, evaporation, river discharge, and melting of land-based ice can all alter sea surface salinity. In addition, changes in ocean circulation can also affect salinity distribution via advection process.

Recent observational studies indicate that globally, sea surface salinity has increased over the past few decades in regions where evaporation exceeds precipitation and decreased in regions of excess precipitation. This spatial pattern of sea surface salinity changes is consistent with the notion that saltier regions get saltier and fresher regions get fresher under global warming. The surface freshening at mid-to-high latitudes extends downward and equatorward to intermediate depths, primarily through subduction of the fresher surface water into the thermocline, and subsequent advection of salinity anomalies by the mean current within the thermocline (e.g., Durack and Wijffels 2010). The observed salinity changes cannot be explained by natural variability, either internal to the climate system or from external forcing by variations in solar output and volcanic eruptions (e.g., Pierce et al. 2012); rather, they are consistent with the changes expected due to anthropogenic greenhouse gas-induced warming and the resultant amplification of the global hydrological cycle.

### Sea level

The local sea level is defined as the height of the sea with respect to a land benchmark. The local mean sea level (LMSL) is the sea level averaged over a period of time that is sufficiently long (such as a month or a year) to smooth out the fluctuations caused by waves and tides. Sea level variability and change can have direct impacts on coastal and island communities, economy, and natural environment such as marine ecosystems.

On a global mean, sea level variations can be induced by two causes: variation in the ocean basin shapes and change in the volume of ocean water. The variation of the ocean basins can result from vertical land movements, which occur because of isostatic adjustment of the mantle to the melting of continental ice sheets at the end of the last ice age and/or plate tectonics such as earthquakes. The isostatic adjustment happens because the weight of the continental
ice sheets depresses the underlying land, and when the ice sheets melt away the land slowly rebounds. The volume change of ocean water can result from a change in water density or a change in the mass of the ocean. Changes in water volume, and thus sea level, due to changes in density, which in turn result from thermal expansion (water volume expands as ocean water warms) and salinity variations, are referred to as steric change. For the global mean, changes in water density are primarily caused by thermal expansion, while salinity variations are also important in local and regional sea level variations. Changes in ocean mass are determined by water exchange with the continents. Melting of continental ice in a warming climate will increase the ocean mass and volume.

Regionally, changes in atmospheric pressure, ocean currents, and ocean temperature and salinity, which are closely linked to climate variability and change, can affect the LMSL as well. For instance, ocean circulation driven by
the anomalous winds associated with ENSO can cause large-amplitude sea level variability in the central and eastern equatorial Pacific and along the coasts of North and South America. ENSO-induced changes in the hydrological cycle can also cause global mean sea level variations (Cazenave et al. 2012).

Over the last century or so, tide gauges have been primarily used for observing sea level variability and change. Tide gauges, usually placed on piers, measure the local sea level relative to a nearby geodetic benchmark. Vertical land movements affect tide gauge measurements. During the past two decades, reliable measurements of global scale sea level from satellite altimetry have become available following the launch of TOPEX/Poseidon in 1992, which was followed by the Jason-1 satellite in 2001 and the Jason-2 satellite in 2008. The ocean surface elevation variations are mapped using measurements of sea surface height relative to Earth’s geoid by these satellites. The geoid is the shape that the ocean surface would take under the influence of Earth’s gravitation and rotation alone without other influences such as ocean circulation and tides. On the geoid surface, all points have the same gravitational potential energy. The geoid and the Earth’s pole may readjust with the changes in ground-based ice volume, which can affect local and regional sea level. Corrections are often made to tide gauge data to remove the vertical land movement effect, in order to obtain sea level signals that result from climate variability and change.

Tide gauge and satellite observations show that the global mean sea level is rising, and this rising sea level is likely caused by thermal expansion and land ice melt associated with global climate change. Over the twentieth century, the global average sea level rise is estimated to be \(1.7 \pm 0.2 \text{ mm year}^{-1}\), and this rate has increased to \(3.2 \pm 0.4 \text{ mm year}^{-1}\) since the early 1990s (IPCC 2013, section 3.7). Spatial patterns of global-scale sea level change have been detected since October 1992, when high-quality satellite altimetry measurements became available. In the western tropical Pacific, the rate of sea level rise since the early 1990s has been much faster than that of the global mean. This rapid sea level rise results from the combined effect of decadal climate mode – the Interdecadal Pacific Oscillation – and warming of the tropical Indian and Atlantic Oceans (e.g., Han et al. 2014; England et al. 2014).

Except for nonclimatic events like tsunamis, extreme sea level events (e.g., coastal flooding, high water events, storm surges, etc.) are often caused by storms, particularly when they occur with high tides. The IPCC (2013) report has projected that the global mean sea level will continue to rise throughout the twenty-first century. Sea level rise due to human-induced global warming has been suggested to be a major threat to low-lying coastal areas and island countries around the globe, because they might be facing increased risks of inundation, coastal erosions, and in some regions increased intensity of extreme sea level events (Church et al. 2010; IPCC 2013, Chapter 3). As such, understanding, monitoring, and predicting global and regional sea level variability and change are important tasks for adaptation. Sustained in situ and satellite observations are essential for carrying out these tasks.

Ocean circulation

The ocean is not static; rather, it is in constant motion. An ocean current is a continuous, directed movement of seawater generated by the forces acting upon the ocean. For instance, tidal currents are driven by gravitational pull of the moon. Horizontal current speeds have
a wide range, from approximately 200 cm/s in the swift western boundary currents (Gulf Stream, Kuroshio, Somali Current, etc.), through 10–100 cm/s in the equatorial currents, to a fraction of 1 cm/s in much of the surface layer and in the deep ocean. Vertical speeds are much less, in the order of $10^{-5}$ cm/s. Due to its slow speed, vertical current is difficult to measure within the levels of an instrument’s accuracy. Both direct and indirect methods are used to estimate horizontal currents. Surface drifters, subsurface floats, current meters, and acoustic Doppler current profilers are all used to directly measure ocean currents. “Geostrophic method” is also used to infer ocean current, based on the fact that large-scale ocean circulation generally obeys geostrophic balance. Due to the Earth’s rotation, currents are deflected to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. This effect is known as the “Coriolis force.” The deflection causes highs and lows of pressure directly proportional to the speed of the currents, referred to as “geostrophic currents” due to the balance between the Coriolis force and the pressure gradient force – referred to as “geostrophic balance.” Based on geostrophy, the global sea surface heights derived from satellite altimetry (see “Sea level,” above) are often used to infer global surface ocean circulation.

Ocean circulation is referred to as the large-scale movement of waters in the ocean basins. There are two types of basin- and global-scale ocean circulations: wind-driven ocean circulation and thermohaline circulation. For wind-driven ocean circulation, the most prominent features are the “subtropical gyres,” which prevail at the upper approximately 500 m of the subtropical oceans, and their deep parts can reach 1–2 km near the western boundary region, together with the equatorial current systems, subpolar gyres, and Antarctic circumpolar current. The wind-driven surface currents consist of two parts: a directly wind-forced component (Ekman drift) and a geostrophic current associated with the gradient of sea surface height. Surface wind also forces shallow meridional overturning circulations in the upper approximately 700 m tropical and subtropical oceans, known as the subtropical cells. They transport heat from the upper equatorial ocean to the subtropics, where the water gets cooled and subducted into the thermocline due to wind-driven, surface mass convergence. The colder, thermocline water subsequently flows back into the tropics, where it upwells to the ocean surface and completes the shallow overturning cells.

By contrast, thermohaline circulation is a large-scale, deep circulation depending on variations of temperature (thermo) and salinity (haline). It transforms low-density upper ocean waters from the tropics to higher-density intermediate and deep waters in high latitude, such as the northern North Atlantic away from the pole, where surface water gets cooled and may also get salty due to sea ice formation. Both lead to increased surface density and thus deep-water formation. While the conversion to dense waters of the thermohaline circulation occurs in restricted regions at high latitudes, the return to the surface involves slow upwelling, mixing, and diffusive processes over much larger geographic regions (Figure 1). The thermohaline circulation is frequently used synonymously with the deep Meridional Overturning Circulation (MOC), even though the MOC also includes a wind-driven component (IPCC 2013, AIII-19, 29).

The ocean can influence climate by storing heat and then gradually releasing it to the atmosphere to drive atmospheric convection and circulation. Meanwhile, variations of ocean circulation can affect climate by transporting large amounts of heat, fresh water, and carbon,
OCEANS AND CLIMATE

and exchanging these properties with the atmosphere. Changes in ocean circulation, therefore, can have direct impacts on regional and global climate. The cold, Younger Dryas event that occurred about 11,500 years ago is thought to be contributed from the shutdown of the Atlantic MOC (AMOC) due to the large quantity of freshwater influx from the melting land-based ice, which reduced sea surface salinity and density, and therefore stopped the deep water formation in the North Atlantic Ocean. The weakened AMOC reduced northward heat transport and thus caused North Atlantic cooling. It is possible that the cold SST of the North Atlantic induced changes in atmospheric circulation, which helped to spread the cooling to a global scale (Rahmstorf 2002). Climate model simulations suggest that the AMOC very likely will weaken over the twenty-first century relative to preindustrial values, due to the surface warming and melting of continental ice (IPCC 2013, section 3.6). Modeling studies suggest that a weakened AMOC can have important implications for regional acceleration of sea level rise along the northeast coast of the United States (Yin, Schlesinger, and Stouffer 2009). Further observational and modeling studies are needed to understand this effect, in order to provide informed decision-making for adaptation to sea level rise.

Ocean acidification

The ocean is a vast reservoir of carbon dioxide (CO₂). The storage of inorganic carbon in the ocean is roughly 50 times that of the atmosphere. As a result, small changes in the ocean CO₂ inventory can have a significant impact on the atmospheric CO₂ concentration. On the other hand, the ocean is also an important sink for CO₂ released by human activities, anthropogenic CO₂, which is the major human-induced greenhouse gas. Since the beginning of the industrial era, atmospheric CO₂ concentrations have increased from approximately 280 ppm to about 401 ppm in June 2014. Currently, roughly 25% of the total human emission of CO₂ to the atmosphere is accumulating in the ocean. While this oceanic absorption reduces atmospheric CO₂ levels and helps to abate global warming and its impacts, increased CO₂ in the ocean has a significant impact on the chemistry of seawater.

Ocean CO₂ uptake causes pH reductions and alterations in fundamental chemical balances for an extended period of time. Together such processes are commonly referred to as ocean acidification (Doney et al. 2009; IPCC 2013, section 3.8). Ocean pH is a measure of ocean acidity or basicity. A decrease in pH value means an increase in acidity and in hydrogen ion (H⁺) concentration. The pH of pure water is very close to 7. A solution with pH less than 7 is acidic and greater than 7 is basic or alkaline. The average pH (total scale) of surface waters ranges between 7.8 and 8.4 in the open ocean, so the ocean remains mildly alkaline at present. The mean pH of ocean surface waters, however, has fallen by about 0.1 units, from about 8.2 to 8.1, since the beginning of the Industrial Revolution. Uptake of anthropogenic CO₂ is shown to be the dominant cause (Doney et al. 2009; IPCC 2013), even though other natural and anthropogenic processes can also cause variations in oceanic total dissolved inorganic carbon and pH.

IPCC estimates of future atmospheric and oceanic CO₂ concentrations indicate that, by the end of this century, the average surface ocean pH could be lower than it has been for more than 50 million years. Ocean acidification can have large impacts on marine ecosystems, particularly on calcifying organisms such as corals, coral reef communities, and planktonic organisms. The increased CO₂ reduces pH and lowers calcium carbonate. The lower calcium carbonate concentration reduces calcification rate and therefore
affects the shell-forming organisms. The potential for marine organisms to adapt to more acidic water and broader impacts of acidification on marine ecosystems are not well understood. They are priority areas of further research.

SEE ALSO: Climatology; Ocean biogeography; Oceanic circulation; Oceans and seas: physical geography

References


Further reading

Oceans and seas: human geography

Philip E. Steinberg
Durham University, UK

Discovering the sea

Although historically few human geographers have identified the ocean as their area of specialty, the ocean has long had an important, if unrecognized, role in the development of human geographic thought. Since the Classical era, political geographers have highlighted the importance of maritime choke points as strategic sites of geopolitical interest. Following in that vein, the oft-cited works of late nineteenth- and early twentieth-century geopolitical thinkers Alfred Thayer Mahan and Sir Halford Mackinder focus on the balance between sea power and land power. Likewise, many early works on common property resources that inform today’s political ecologists focus on regional fisheries management systems and the ocean tenure regimes of coastal communities. Similarly, transport geographers have long noted the importance of maritime trade, an orientation that in turn impacted mid-twentieth-century urban economic geographers who researched the hinterland connections of port cities. And some of the iconic writings of mid-twentieth-century environmentalists, which went on to inspire a generation of environmental geographers, chronicle the causes and impacts of marine pollution.

Notwithstanding this rich history, however, prior to the 1990s the ocean was generally perceived as a secondary arena for geographic study: a space whose surface was used or impacted by society but not a fundamental space of society. That is, the ocean was typically understood and researched as a surface upon which land-based social processes were played out, rather than being studied as a space within which social processes, institutions, and power relations were constituted. Human geography until the last few decades was unapologetically terracentric.

There are likely a number of reasons for human geography’s terrestrial bias. First, despite the critique that geographers have long leveled at the idealization of the state as a naturally occurring, organic entity, geographers (and, more generally, social scientists) still have tended to conceive of the world as a universe of state territories. Because the ocean lies primarily outside state territory, it has tended to fall off the map of many state-centric human geographers. Second, these naive perspectives on territory have been complemented by naive perspectives on place. Humanistic and phenomenological geographers typically have understood place as a point in space at which people (individually or collectively) transform nature and embed social meaning through material and symbolic inscriptions. Because humans are unable to inscribe their presence at discrete points on the ocean’s surface, the ocean is often (if incorrectly) understood as a space within which communities fail to establish “roots.” The ocean thus has frequently been viewed as placeless and therefore uninteresting. A third reason for the historic lack of attention...
given to the ocean by human geographers is a naive separation between notions of “society” and “nature.” As a space without any obvious social boundaries or transformations, the ocean has appeared to many as a space of “nature.” Thus it is often deemed to be of little interest to human geographers, beyond the subset with a specific interest in bridging the nature–society, or physical–human, divide.

Today, all three of these concepts – the foundational nature of state territory, the association of place with rootedness, and the conceptual division between nature and society – are questioned by human geographers, and this has opened up the ocean as a new area for human geographic research. This explosion in ocean research began in the 1990s with a series of special issues of geography journals. These included a special issue of *The Professional Geographer* (Steinberg 1999) focusing on possibilities for integrating human and physical geographic ocean studies, a special issue of *Geographical Review* (Wigen and Harland-Jacobs 1999) that sought to reconceive the world as a series of ocean regions, and a special issue of the *Journal of Historical Geography* (Lambert, Martins, and Ogborn 2006) that sought to document the role of the ocean in historical geography. This has been accompanied by monographs, edited volumes, and review articles. As just one indicator of how the ocean has come of age as a topic of human geographic study, it is useful to compare the “Coastal and Marine Geography” entries in the two *Geography in America* volumes. In the 1989 edition, the discussion of human geographic themes (West 1989) is brief and devoted almost entirely to tourism issues. By contrast, the entry in the 2003 edition (Psuty, Steinberg, and Wright 2003) covers a broad range of human geographic topics including resource, cultural, political, and transport concerns.

Ocean geography as a mirror of human geography

Since its emergence in the 1990s, ocean geography has advanced in step with changing trends and insights in the broader field of human geography. Thus, while the ocean remains, in some senses, marginal to the dominant, land-centered current of human geography, it simultaneously is now deeply integrated with contemporary aspects of human geographic thought, and it is increasingly playing a role in enhancing geographic perspectives.

First-wave ocean geography: understanding globalization

During the 1990s, human geographers from a range of theoretical backgrounds made a sustained effort to interpret the rising tide of globalization, its historic antecedents, and its implications for nation-states, local cultures, and regional production and trade networks. Two of the first works in the blossoming of modern ocean geography, the “Oceans Connect” special issue of *Geographical Review* (Wigen and Harland-Jacobs 1999) and *The Social Construction of the Ocean* (Steinberg 2001), can be understood in this context. On the one hand, these two works take markedly different approaches. “Oceans Connect” attempts to challenge the terracentric division of the world into continents and states by conceiving a world defined by ocean–basin regions. *The Social Construction of the Ocean* adopts a more historicist and structuralist approach, seeking to understand how various phases of Western capitalism both reflected and were reproduced by specific ideologies and practices for incorporating the ocean into the space of society. Notwithstanding these differences, however, both works are rooted
in an attempt at understanding the ocean in an era of globalization and, conversely, both seek to explore how attentiveness to the ocean as an important space of society can aid our understanding of an increasingly “global” world.

In that respect, the first wave of ocean geography, as exemplified by these two works, can best be understood as fundamentally rooted in international political economy as they sought to uncover the unfolding political economy of the post-Cold War world. That said, first-wave works were not blind to complementary perspectives that were arising from further afield. In particular, works from this era engaged the focus that was emerging – in geography, but especially in the field of cultural studies – on the role of the ocean as a liminal space that both facilitates and symbolizes conditions of hybridity and postcolonialism. Key thinkers like Edward Said, Michel Foucault, and Paul Gilroy all, to a greater or lesser degree, highlight the ocean as a space where social practices challenge modernist ideals that link space, state, and society in coherent wholes, and several of the contributors to “Oceans Connect” as well as Steinberg in *The Social Construction of the Ocean* connect their analyses with these literatures. Nonetheless, even as both works engage theoretical insights from cultural studies, and even as they interpret cultural norms, aesthetic encounters with the sea, and (to a lesser extent) transformations of marine nature, the overall orientation of both is that of historical political economy, derived from the insights of development studies, Marxist studies of imperialism, and global history. Insights from cultural studies and cultural geography are secondarily employed, but generally to support arguments that are rooted in political economy.

Second-wave ocean geography: science and the sea

In the 1980s and, especially, the 1990s, structural explanations in human geography that sought to identify the singular political-economic “logics” of various eras were challenged. Feminist and poststructuralist geographers contended that structuralism was theoretically naïve, missing the intricate ways in which power relations were reproduced and challenged through social and discursive practices at a variety of scales, while empiricist historical geographers were critical of any explanatory system that sought to generalize across eras and geographic contexts. While these two critiques – the feminist/poststructuralist and the empiricist/historicist – were largely distinct from each other, they each, in different ways, pointed to the study of *science* as a key means for bringing the study of the ocean into human geographic research. For historical geographers turning to the sea, science was a crucial dimension of the human–ocean encounter because the ocean has been largely known through science (whether Western science or other, indigenous, techniques of assembling knowledge). Additionally, because ocean science, at least in the modern era, has required extensive state effort, its history is particularly intimately interwoven with histories of state formation and imperialist expansion. Historical geographers thus identified the study of ocean science as a fruitful means for gaining insights into the development of scientific knowledge and scientific institutions, and for understanding how these were connected with the rise of modern political institutions, economic relations, and cultural norms. Many of the articles in the special ocean issue of the *Journal of Historical Geography* (Lambert, Martins, and Ogborn 2006) thus focus on marine science or, more broadly, on how knowledge of the sea was supported, produced, and communicated during past eras.
A second force promoting a focus on science in geographic studies of the sea came from feminist and poststructuralist geographers who turned to science as a means for accounting for the agency of the nonhuman world. Drawing on insights from both feminist science studies (e.g., Donna Haraway) and science and technology studies (e.g., Bruno Latour), actor-network theorists and, somewhat later, posthuman and more-than-human theorists emphasized the role of interactions between the human and the nonhuman in ontogenetic assemblages. Some of the pre-eminent theorists in this area of study – including John Law and Gilles Deleuze and Félix Guattari – give important, if fleeting, attention to the ocean as an exemplary relational space where the human and nonhuman engage in mutual co-constitution. Poststructuralist geographers (and, in particular, actor-network theorists) thus turned to the ocean as a fruitful arena for understanding the role of the nonhuman in social relations. The ocean, after all, is a space where the primacy of the human as the dominant species is always precarious due to the marine environment’s hostility toward human habitation or vision. It is also a space in which it is starkly apparent that “nature” can never be relegated to the role of a stable background or a definable set of resources. Indeed, the ocean’s incessant mobility makes it difficult to consider as anything other than an actant, and this theme has been expanded upon by geographers employing actor-network and assemblage-oriented perspectives. For second-wave ocean geographers, the ocean is continually being co-constituted by human and nonhuman elements: ships, states, and sailors, but also waves, water molecules, and biota, and also regulations, navigation devices, and the unwritten as well as written rules of conduct that guide the lives of people as they travel the ocean’s surface, extract its resources, or gaze into its depths.

Third-wave ocean geography: affective encounters

If the first wave in post-1990 human geography of the ocean drew upon political economy and, secondarily, postcolonial perspectives to understand the ocean amid globalization, and the second wave drew upon science and technology studies to understand the ways in which humans simultaneously differentiate themselves from and integrate themselves with the nature around them, then the third wave has focused more specifically on the affective properties of the ocean encounter. From around 2008, geographers have stressed that while encounters between human and nonhuman natures are characterized by scientific rationality and cognitive understanding, they also are laden with affective meanings that emerge through the processes of connection and co-constitution. In such encounters – for example, between individuals or social institutions and the sea – affective atmospheres are created: not simply emotive reactions (fear, awe, etc.) but a joining of elements in unstable fusion in which, at least temporarily, the whole is greater than the sum of its parts. This perspective is evident in many of the chapters in the Water Worlds edited volume (Anderson and Peters 2014), in chapters ranging from the experiences of surfers and kayakers to scuba divers to offshore pirate radio broadcasters.

To be clear, these three waves of ocean geography have not so much contradicted each other as they have built upon the insights of previous inquiries. Thus some of the key “second-wave” works on science and the sea examine ocean science within the “first-wave” context of global trade and domination. Likewise, “third-wave” works on the affective nature of ocean encounters generally begin from a “second-wave” understanding of the ocean as a space where the human and nonhuman meet in a dynamic assemblage.
This, in turn, sets the stage for the fourth wave of ocean geography that is emerging today.

**Fourth-wave ocean geography: vibrancy, verticality, and volume**

If the first wave of ocean geography was associated with the rise of political-economic perspectives, the second with the rise of an interest in the more-than-human, and the third with a focus on affect and emotion, the fourth wave has emerged in the context of a “new materialism” that increasingly is guiding research across a range of human geographic subdisciplines. While avoiding the excesses of environmental determinism, new materialists stress that the material conditions (i.e., the physical environment) within which human relations occur is nonetheless significant. Drawing extensively on understandings of how the geophysical world is known, changed, and experienced, as well as how it limits and enables human actions, new materialists study, for instance, how economies and social institutions are inseparable from the transformation of matter across space and between physical states. Matter is understood as substance made “vibrant” through human encounters.

This turn in geographic theory has important implications for how geographers think of (and with) the ocean. The new materialism suggests a turn away from understanding the environment as static, flat, and surficial. Instead, the material environment is understood as one with depth, constituted by the incessant churn of molecules that produce a variety of social relations as they re-form through space and time. While this dynamic materiality is present everywhere (all landforms shift and re-form over the very long term due to climate variability and tectonic movement), the ocean presents a particularly poignant environment for revealing the Earth’s voluminous dynamism. Within geography, much of the early work in this genre focused on air-space, as a space that progressively has been incorporated into the social fabric through a variety of political and legal constructions: as a space for war, surveillance, and atmospheric geo-engineering. However, recent works have turned to the sea as well for appreciating the import of a planet whose surface – and whose depths – are understood as vibrant, vertical, and voluminous.

The effect of this perspective on ocean geography is to frame the ocean not as an exceptional space (beyond state territory and beyond human habitation) but as a paradigmatic space, integrated with the other spaces that constitute the world. Although all spaces are mobile, mobility in the ocean operates at a quicker speed and with more evident force. Similarly, although all spaces are voluminous, in the ocean there is particularly profound tension as states and shippers seek to render it as a flat surface. And while all spaces are in a constant state of change, this dynamism is particularly apparent in the ocean as it shifts between its paradigmatic liquid form and related forms of solid (ice) and gas (foam, spray). From this perspective, the melting of polar seas due to climate change or the inundation of cities due to sea level rise can be understood as an integral part of human geography, not an external, environmental stimulus but a constituent part of the dynamic Earth which social life is constantly resisting, adapting to, and enlisting as an ally.

**Destabilizing the sea**

A common theme throughout each of these four waves has been a recognition that the ocean’s difference provides opportunities for geographic inquiry. Prior to the 1990s, the ocean’s difference relegated it to the margins of critical human geographic analysis: while the ocean might have been an interesting space for observing the interaction of human (and physical) processes, it was considered too exceptional for providing
insights into how those processes might actually emerge or for facilitating research on their constitutive elements.

Since the 1990s, however, the ocean’s exceptionalism has been seized upon as presenting an opportunity because of the way in which it destabilizes accepted categories. In the first wave, the persistence of the ocean as a fundamentally global space was used to destabilize the seemingly foundational division of the world into ontologically distinct territorial state-societies. In the second wave, the relative dominance of the nonhuman over the human in the ocean was used to destabilize notions of exclusively human subjectivity. In the third wave, the complex dynamics of the ocean encounter were used to destabilize modernist distinctions between object and emotion. And in the fourth wave, the voluminous dynamism of the ocean is being used to destabilize notions of nature (and space) as background and surface. The ocean, long relegated to being a zone for applying geographic knowledge, has emerged as force for generating knowledge of the world.

SEE ALSO: Actor-network theory; Construction of nature; Containerization; Geopolitics; Imperialism; Maritime transport; Oceans and seas: physical geography; Place; Ports

References


Further reading


Oceans and seas are all around us. For starters, 70% of the Earth’s surface consists of bodies of water. Moreover, 96% of trade (on average) is carried by ship across the liquid voids that separate land masses. In effect, that means that the majority of items we own, clothes we wear, and goods we use on a day-to-day basis will have been touched by the seas and oceans that often seem so distanced from us. For those who live close to the sea, its impact is all the more obvious through storm surges, flooding, coastal erosion, and management schemes put in place to dampen the effects of the seas and oceans on the land and its populations. Nonetheless, in spite of the pervasiveness of the oceans and seas to our everyday lives, in obvious and not so obvious ways, geography has been a traditionally land-locked discipline. Scholars studying both human and physical processes have looked inwards to land rather than outwards to sea as a source of inspiration for their research. By and large the core concerns of geographers have focused on grounded and materially solid phenomena from landforms to sediments and on the sociocultural worlds of people living in cities, towns, and countryside locales. Yet the omission of seas and oceans from geographical enquiry has meant that a vast portion of our planet has escaped study. The word “geography,” translates as “Earth-writing.” If “Earth-writing” is taken to refer to writing about the planet as a whole, then two thirds of the narrative is missing if the seas and oceans are not included.

It might well be asked why geographers have failed to study the sea. Physical geographers have, for many years, focused on the processes borne from water – in rivers, lakes, and by the coast – and the changes to the global landscape that can be read through the processes of ice melt, longshore drift, erosion, and deposition. However, the seas and oceans as distinct areas of concern have been on the horizons of studies. For physical geographers, one of the prominent reasons has been the lack of accessibility of the seas and oceans as research sites and the dangers these locations pose. Although it is possible to explore the sea from the coast, such work tends to focus on the influence of the sea on the land (through erosion, deposition, and the creation and transformation of landforms) rather than the processes of the sea itself. Seas and oceans are vast and deep. This has traditionally made research beyond the shoreline difficult, hazardous, and expensive. That is not to say some land areas do not pose risks, but the scale and breadth of oceans make them a unique challenge to researchers.

Likewise, human geographers have had a long preoccupation with water, but less so with the vast liquid spaces of oceans and seas that separate (and also connect) continents. Sociocultural, economic, political, and environmental geographers have explored issues of water flow (or lack thereof) in developing nations, irrigation systems, damming projects, and flooding prevention, to name just a few engagements with what might be called “hydro worlds.” The sea, though, has been somewhere in the background. Again, the issue of access looms large. Human geographers have grappled with how to study...
a space that is beyond the solid physicality of land. But also the question of why comes to the fore. If, for the majority of us, we don’t live there or work there it may be questioned why seas and oceans are relevant as spaces for human geographers to explore. In human geography, this is embedded in contextual understandings of oceans and seas, which have marginalized them as spaces that matter in examining sociocultural, political, and economic life. As Philip Steinberg contends (2001), the industrial and postindustrial Western world has constructed the ocean as a void to be traversed for capital gain. The sea is an empty space that divides and connects spaces of importance on land. As such, the seas and oceans have been rendered inferior and outside of academic attention.

Yet it is no longer possible to say that geographers have not explored watery environments as part of their investigations. Over the past ten to fifteen years, geographers have increasingly paid attention to the large portion of the Earth that has been previously unreachable and omitted from study. Developments in geography have demonstrated the very relevance of seas and oceans both in their own right and in view of landed life. Indeed, broader advances in geographic thought have reconfigured the world not as constituted of discrete, bounded, and neatly defined parcels of space (land, air, sea, and so on), but as fluid, where spaces are connected and where elements meet, mix, and reform. With this vision of the world it becomes impossible to focus only on the land. Accordingly geographers have explored the ways in which the land is forged in relation to the sea (and also air); and likewise how the sea and air relate to the land. This “opening up” of physical and human geography alike to the interconnections between spaces has drawn increased attention to the seas and oceans as particularly salient spaces in the making of the world as we know it. Far from being irrelevant then, they are now appreciated to be wholly significant in better understanding a range of geographic processes, and re-visioning accepted knowledge of the world.

For physical geographers studies of the seas and oceans have centered on two areas – coastal areas and marine areas with the former focusing on the coast and the latter on the open sea. In view of coastal geography, scholars have investigated the morphologies of landscapes – beaches and coastlines – and how these have been changed by the mobilization of sediment and the velocity of waves. Coastal geomorphology has investigated, on various scales, these alterations to the coast from local to regional scales, driven by regular weather or storms, and the impacts wrought from human interactions with physical processes. Marine geography on the other hand has considered the geographies at work beyond the shore and is equated with marine science or oceanography. Here the seas and oceans have been investigated as a way of gaining a better understanding of the Earth system as a whole. Geographers have explored a range of interests under this branch, from the ecologies and ecosystems of oceans, the influence of weather patterns on ocean movement, to the impacts of pollution on water quality (Psuty, Steinberg, and Wright 2003).

Human geographers have, perhaps surprisingly, developed strands of interest that tie their work closely with the concerns of scientists. There has been a recent shift to explore the elemental mediums through which human life is lived: land, air, and sea. It is argued that the varying states of these mediums lead to particular co-compositions with human life. Indeed, if human geography is the study of the relations between people, space, and place, then it makes sense that such relations depend on the natures of those spaces and places. For human geographers, this interest in what is coined the “non” or “more-than-human” world has resulted in
some of the most diverse work concerned with geographies of seas and oceans. Geographers have become particularly interested in the ways that the physical processes of the sea co-join with sociocultural experience. Human geographers have moved away from understandings of the world as socially constructed (prioritizing the role of humans over the role of nature) to instead focus on the ways that human life is both shaping and shaped by the physical world. Such research draws consideration to the ways in which humans interact with oceans and seas, and moreover, the negotiations they face when trying to manage the force of water. Whereas humans have been traditionally understood as “mastering” nature (through the transformation of the landscape), attention to the seas and oceans demonstrates the complexity of such mastery, illustrating how those using the sea can seek to manipulate or harness its natural power but can also be victim to it. Such work that brings together human and physical dimensions in this context has also sought to consider the temporal force of oceans and seas at periodic moments; for example when tides advance and retreat (a result of broader extraterrestrial forces as the moon’s gravitational pull relates to oceanic movement). Here the oceans and seas are both a routinized and predictable feature of our planet; yet simultaneously irregular and volatile in how they combine with the human world.

Scholars have also, more experimentally, explored the shifting material form of our oceans and seas from the state of liquid to solid; salt water to ice. Research on ice confuses how we might typically think about seas and oceans. Such spaces cannot be understood simply as liquid zones between countries and continents; instead they can become part of countries and the continents they meet, as liquid transforms to ice. This has allowed geographers to challenge the uneasy categorizations that partition the world in identifiable “zones” of study. Rather, the oceans and seas permit an understanding of the constantly changing nature of the planet; its fluid shifting form as land becomes sea and sea becomes land. This work has gained attention not only by those geographers co-combining human and physical concerns (as described) but also by political geographers.

Borders, boundaries, and territory at sea are other core concerns of geographers working in this area. The sea is a space over which political control is contested (particularly in view of resources such as oil and activities such as fishing) and a space with a distinct historical narrative over its legal governance (Steinberg 2001). Traditional explorations of geopolitical control and governance over the oceans have been a concern of geographers who have sought to explore the zonation of ocean space and the politics over activities on and also under the ocean (from shipping violations between territorial and extraterritorial sea spaces to seabed rights where resources can often be gained). Yet the governance of the sea is different to the land. Unlike the creation of nation-states, the formation of the sea as an international space of shared stewardship results in differing conflicts in how it (and events on it, under it, and over it) should be policed. It requires a special form of governance, control, and partitioning. Land is territory to be acquired whereas the sea cannot be acquired in quite the same way. The Arctic Sea, for example, is a pivotal space for rethinking geopolitical processes. Where land turns to sea and sea turns to land through processes of ice formation and ice melt, new territory is perpetually formed and dissolved, creating new questions over ownership, political control, and rights over resources.

Moreover, while the sea might be fluid in its states from solid to liquid, in liquid form the sea’s fluid materiality means it is a constantly moving space that presents particular challenges to governance. Although possible to draw a line across a
OCEANS AND SEAS: PHYSICAL GEOGRAPHY

static representation of the sea on a map, drawing a line to demarcate space in reality is far more difficult. The sea moves – it does not stay still. Partitioning it therefore becomes a tricky task. Those studying fishing at sea have explored this tension. The reliability of sustainable fish certification is challenged when we remember that both seas and oceans move, as do fish (Bear and Eden 2008). Areas of ocean may be designated on a map as zones of sustainable fishing; but there is little to stop fish moving in and out of these areas. The seas and oceans therefore have allowed scholars to complicate and extend work in longstanding areas of geographical concern.

And just as the seas and oceans permit a vision of a more fluid world of connection and boundary dissolution, they also permit a vision of the world that moves beyond a horizontal perspective that looks out and across space to what is in front or behind us. The seas and oceans complicate ideas relating to surfaces. When liquid, seas and oceans open up what is called a “vertical ontology” whereby knowledge of the world can be extended through depth. The sea is not ever moved on. People or things (from rafts and canoes to boats and ships) do not traverse on top of the surface of the sea – they always move in it. Even in a buoyant condition, a floating subject or object is moving through the water with part above and part below. Studies of the sea have thus moved beyond traditional spaces of study to explore submerged, subterranean, and underwater worlds and their distinct natures and sociocultural engagement with them (Anderson and Peters 2014).

Engagement with oceans and seas has been a key intervention of geographers exploring planning and environmental management to prevent coast flooding. Yet human geographers have furthermore sought to unpack the embodied, nonrepresentational, and haptic experiences bound up in relations with the water world. It is argued that our sensory experience with the world is radically altered when we engage with seas and oceans as opposed to the land. Anderson’s work on surfing demonstrates the uniqueness of “being” in the water – a space moving, wet and dynamic (see Anderson and Peters 2014). Other work has investigated diving, swimming, sailing, and, in relation to development studies, the special relationship some societies have with oceans and seas as a means of sustaining life. Where a Western conceptualization of seas and oceans often renders them as distant spaces outside of everyday life, for many nonmodern worlds the seas are as familiar as the land and they are spaces of ritual, ceremony, and also of vital connection and resources central to sustaining life. Indeed, the oceans have also been explored as spaces across which mass migrations have occurred, reshaping societies and histories.

Such studies form part of a wider interest for human geographers in how people relate to seas and oceans, and have done so over time. Historical geographers have explored the relations between people and the sea through the lens of imperialism and colonialism, arguing that the seas were a key site of global expansion during this era. Lambert, Martins, and Ogborn (2006) use the sea as a way of reimagining global politics and change. They contend that the sea permits a way of thinking that moves beyond nation-state driven narratives. Such a focus on the oceans and seas as open and borderless allows us to rethink how the world has come into being. In short, many geographical areas of concern – with sediments, movements, fluctuations, change, management, politics, and society – can be decentered and rethought through attention to the oceans and seas, and this work continues to expand as interest in and access to the sea extend and deepen.

SEE ALSO: Coastal zones; Sea level rise
References


Ontology: domain applications

Nancy Wiegand
University of Wisconsin – Madison, USA

Geospatial data are characterized by containing either explicit spatial location information for geographic features or references to a place. Geospatial data may be used in a GIS (geographic information system) as part of viewing, creating, and analyzing maps and geographic features. Geospatial data are fundamental for many applications, including economic development, natural resources management, environmental protection, and emergency response. This entry explores ontology and applications in the geospatial domain and relates the geospatial domain to work in the area of semantics and ontologies.

Semantic technologies are being developed in the effort to create a Semantic Web and also to add a semantic layer to various kinds of applications. The Semantic Web vision includes the existence of formal ontologies as background semantic knowledge bases that can be referenced from webpages to provide context in a domain. Using such ontologies, automatic or semi-automatic methods can resolve semantic heterogeneity across webpages to relate disparate but related pieces of information. That is, intelligent agents would be able to interpret information on the web without human intervention and potentially return answers to complex queries posed over the web instead of returning links to webpages as done currently. Another aspect of the Semantic Web vision is to create linked open data in which similar data items are linked across multiple datasets through the use of common identifiers.

The World Wide Web Consortium (W3C, http://w3c.org) is developing standards to facilitate the creation of a Semantic Web and associated technologies, such as reasoners and rule and query languages. And, the Open Geospatial Consortium (OGC, http://www.opengeospatial.org/) is adding spatial aspects to semantic standards and technologies to create a Geospatial Semantic Web. The W3C and OGC have recently worked together on best practices for spatial data on the web (https://www.w3.org/TR/2016/WD-sdw-bp-20160119/).

Ontologies

Ontologies are integral in the vision for the Semantic Web and semantically enabled applications. An ontology contains concepts, terms, and relationships that describe a domain. An ontology may or may not be populated with instances. Formal ontologies are represented using an ontology language, such as OWL, a W3C standard for a web ontology language (http://www.w3.org/2001/sw/wiki/OWL).

An ontology is often organized into classes and subclasses but differs from a taxonomy in that attributes, relationships, and constraints can also be included. For example, a taxonomy could show that “river” is a type of (subclass of) “water body,” but an OWL ontology further allows specification of attributes (datatype properties) and relationships (object properties). For example, in OWL, it can be modeled that a river has a datatype property that specifies an
average rate of flow. And, an object property for
river called *flows into* could be modeled as being
transitive and having a range of type water body.

Uses for an ontology include:

- formulation of the concepts of a domain, that
  is, creation of a conceptual model;
- specification of terms to use to describe the
  concepts;
- organization of information;
- potential for ontology terms to be used as a
  type of standard for terms, resulting in future
  conformant terminology;
- enhancement of search;
- resolution of heterogeneous terms involved
  in queries and in data integration;
- enhancement of gazetteers;
- enablement of a Semantic Web and a
  Geospatial Semantic Web.

An ontology expressed in a formal ontology lan-
guage, such as OWL, is machine readable. This
allows automated processing by reasoners. A rea-
soner is an inference engine that infers logical
consequences from a set of asserted facts. A com-
mon type of inference is subsumption to find all
the sub or super classes of a class, for example,
all types of water bodies. A Semantic Web rea-
soner understands formalisms underlying Semi-
tic Web standards including the OWL language
and description logics.

Compared to schemas in database management
systems (DBMSs), which have a closed world
assumption, ontologies are part of an open world
assumption. That is, ontologies are meant to
define a domain more broadly than what would
be necessary for a particular database application.
Also, ontologies are meant to be re-usable and
extendible. Modular ontologies, in particular,
facilitate re-use of pieces of ontologies.

Although ontologies are valuable in enabling
the uses outlined above, they are not easily built.

The process of building an ontology is referred
to as ontology engineering. A typical approach
is to gather domain experts together in person
with ontologists who are then able to formally
express constraints on facts and relationships.
The Ordnance Survey in Great Britain did
seminal experimental work on methods to
create ontologies, such as developing tools for
domain experts to easily express facts in a certain
manner that enabled automatic formalization.
A current ontology editor, such as Protégé,
helps in entering and organizing information
and outputs an ontology in a formal language.
Because ontologies are difficult to build, a goal
is to automate building ontologies.

**Ontologies to enhance search**

for geospatial data

Being able to locate spatial data motivated the
development of national spatial data infrastruc-
tures (SDIs) as well as numerous state and local
geospatial portals. The prior US Geospatial
One-Stop (now subsumed under data.gov),
for example, made federal and other datasets
and resources accessible from one location.
Geospatial data are usually described in separate
metadata files using standard metadata formats
such as those from the Federal Geographic Data
Committee (FGDC), International Organiza-
tion for Standardization (ISO), or Dublin Core.
Metadata elements typically include theme key-
word, date, location, provider, and type of data.
Search in SDIs or portals is done over selected
metadata elements, now likely using a catalog
service for the web (CSW). The simplest user
search interface usually provides a method (text
or map based) to determine an area and also
provides a text box in which the user enters a free
form keyword. The search, however, may pro-
duce no results or too many results. Even when
a search system expands the user keyword with associated synonyms, there might not be any results due to a mismatch of terms and concepts between the data provider and the user. That is, user terms may not match stored metadata keywords or synonyms. Alternatively, a search may yield hundreds of results due to imprecision.

Because of the difficulties of searching for geospatial data, researchers have been working on improvements such as ontology-based search. An example is that theme keyword search can be expanded to include sub or super concepts through lookups to ontologies. For example, if an ontology has a pond as being a type of lake, that is, a subclass of lake, a search for lake could also return records referring to pond, as well as the other subclasses of lake. Also, reference to an ontology for a search term could restrict the domain of the term, enabling more precision, such as water in the domain of hydrography instead of water in a GIS city plumbing application. Another projected use of an ontology not fully exploited yet could be the ability to include information from relationships other than sub or super class relationships, such as flows into or part/whole relations. An example search expression in the near future could be “flows into the Mississippi River” to locate data on particular streams.

Work is ongoing in using ontologies and associated software in portals or SDIs to improve search. For example, Esri’s geoportal technology has been enhanced to expand or refine keyword search over metadata and data through access to ontologies. The International Coastal Atlas Network (ICAN, http://ican.science.oregonstate.edu/) provides ontology-enhanced search that includes an ontology browser and Semantic Web service. For example, a search for “coastline” will include a search for “shoreline” using its knowledge base (http://netmar.nersc.no/sites/netmar.nersc.no/files/ICANCookbook_ConnectingYourAtlas_v2.1_20120730.pdf).

The Museum Finland project (http://www.seco.tkk.fi/applications/museumfinland/) created a semantic portal and used ontologies for intelligent information retrieval across cultural collections. The project was part of larger work on semantic technology (http://www.seco.tkk.fi/projects/finnonto/). Ontologies are being used to improve precision and recall in searching for earth science data at the Oak Ridge National Laboratory using the Mercury metadata search engine. Ontology search is also being used in GeoSearch (Gui et al. 2013), in which multiple SDIs are being accessed through the GEOSS clearinghouse and search for data is enhanced using the SWEET ontology (http://sweet.jpl.nasa.gov/). GeoSearch also uses rules and semantic similarity measures to improve ontology-based search. Another example to improve ontology search mines the data itself using latent semantic relations to create an ontology for polar data (Li, Goodchild, and Raskin 2012). INSPIRE, the Infrastructure for Spatial Information in the European Community, also has initiatives to include semantic technology.

Resolution of heterogeneous terms: data integration and querying

Even after finding data, full use of geospatial data remains difficult because of the semantic heterogeneity in basic terms and concepts. Contrary to this, heterogeneity in file formats and coordinate systems for spatial data can be resolved using file and coordinate conversion routines. Also, some attribute names and values can be resolved using ETL (extract, transform, and load) techniques. But the use of different conceptual notions and vocabularies for nonspatial attributes, which comprises semantic heterogeneity, still poses
significant challenges. Semantic heterogeneity is due to each data provider using different descriptions and models. For example, one dataset may have subsets of terms or terms that are similar but not quite the same as terms in other datasets. Semantic heterogeneity prevents interoperability among spatial datasets.

The lack of semantic interoperability is not only a problem within one domain; it prevents needed collaboration across domains. Although spatial information has been referred to as an integrator for data across domains because different kinds of data can be linked through location, the data are limited in further integration due to different concepts, terms, or classifications of terms.

Data

Just as ontologies can be used to enhance search for geospatial data, they can be a solution for data integration. Data integration is the combining of data from different sources into a new unified view. An early study by Fonseca et al. (2002) presented the notion of ontology-driven geographic information systems to help integrate data and classify features across remotely sensed images. They described high-level ontologies as being an integrator over related applications that would each have more detailed ontologies. In their example, a top level ontology could cover multiple communities and have generic land use and land cover categories, such as forest, nonforest, and deforestation. A domain ontology at the next level of detail could then distinguish types of deforestation due to farms or cities. Finally, an application ontology for a specific task could further expand farms to soybeans, cattle, and so on. Combining these levels of ontology together allows relating specific and general land use or land cover codes to each other.

The approach of using upper level ontologies has also been used to integrate data from completely different domains. Upper ontologies, such as DOLCE, which contain abstractions relevant across knowledge domains, help in integrating data by connecting data items through high level concepts.

Aggregating data laterally, that is, across geographic areas or jurisdictions, also often results in needing to resolve semantic heterogeneity. For example, in local areas, such as neighboring counties, different classification systems in land use, zoning, wetlands, and land cover can occur. And data integration across local, regional, state, national, or world datasets will usually have semantic problems due to a lack of standards and also from language, culture, and landscape differences. For example, differences in landforms in various parts of the Earth and also how they are culturally perceived and described was studied by Mark, Turk, and Stea (2012).

Some solutions to semantic heterogeneity have involved developing standards to make all schemas and data compatible. For example, work in the Consortium of Universities for the Advancement of Hydrologic Science (http://www.cuahsi.org/) involved cooperation among data providers to use similar terms. But, in general, it is difficult to create standards and enforce them. The standards approach also does not cover legacy data unless effort is put forth to convert old data to the new standards. Further, setting standard terms may not be effective because such terms may end up at a high level of description, say, agriculture, and lose needed local detail (e.g., orchard) when comparing land use/cover classifications, for example.

Developing ontologies rather than standards helps prescribe terminology for others to use, enabling future data integration. And, ontologies used as a type of standard may be more
comprehensive than other standards and so help solve the issue of local detail being lost.

**Querying**

In addition to helping integrate data, ontologies can help in querying over heterogeneous data by providing a source of global terms and concepts needed to create semantic agreements (e.g., mappings) to local or legacy terms. Such mappings are then used to resolve semantic heterogeneities. For example, in prototype database or information systems that implement a semantic mapping approach, the query interface typically presents ontology terms for the user to choose to pose a query (rather than allowing the user to type in any term). The ontology terms are then looked up. This approach keeps data local and in its own terms yet allows local data to be part of a combined knowledge base. This is contrary to integrating all data by putting it into a unified form and copying it into a data warehouse, for example, in which local updates would keep needing to be propagated.

As mentioned, land use and land cover are example domains needing semantic resolution for querying. Although not the same, land use and land cover are sometimes combined. They are both represented by taxonomies created for local or national purposes without universal standards. These types of codes can be very heterogeneous even across local areas, and they may change through time. Climate change studies are an example use case that requires resolution of world-wide land cover codes as well as resolution of prior classification systems over time. Various methods have been experimented with to study and resolve land use or land cover classification systems including statistical methods, mappings, and merged ontologies (Ahlqvist, Macgill, and Guo 2004; Gahegan *et al.* 2011; Wiegand 2012).

Distributed query systems, such as federated database systems, that process queries over heterogeneous data also have to resolve differences in schemas and terminology. Global As View and Local As View approaches have been tried, for example. Using formal ontologies to resolve differences is a newer approach. Work has been done in the database community to add ontologies to relational DBMSs to support ontology-based querying, such as in Oracle (Das *et al.* 2004). Newer work is now being done on ontology-based data management to have DBMS technology accommodate formal knowledge representation and reasoning.

**Gazetteers**

Gazetteers provide another example application in the geospatial domain for which work is being done to enhance spatial information with ontologies and semantic relationships. Gazetteers contain a place name, type of feature, and location but do not include a place ontology, for example, in which additional information may be available, such as relationships between features. An example of information that is not now available in gazetteers is how administrative units are related especially when they are not completely nested. Another issue is whether spatial information, such as elevation or latitude and longitude, can be used to automatically determine semantic relationships, such as a change in elevation determining the direction of stream flow between lakes, for example.

**Linked data**

One of the visions of the Semantic Web is to have data within and across domains linked together and accessible over the web. The linked
open data (LOD) project (http://linkeddata.org/) builds and connects distributed data across the web. In the LOD diagram (http://lod-cloud.net/), each circle represents a knowledge base of linked data for a particular domain. Further, data are linked across domains. DBpedia is in the center. Initial data focused on the life sciences but more and more types of data are being added, including geospatial linked data. One value of the linked data initiative is to make data more open by putting it in a standard linked format and having it available over the web. An example is linked government data available through data.gov in the United States. A further value of the LOD cloud is the likelihood that new knowledge will be found through extensive linking of data. Although ontologies are not necessarily part of the LOD, formal domain ontologies and mappings would be helpful in data integration and querying for data not published using the same conceptual terms and identifiers.

The resource description framework (RDF) is a W3C standard data representation (http://www.w3.org/TR/rdf11-primer/) used to create linked data. RDF is a triple format consisting of resources organized as a subject, predicate, and object. Examples of triples are the “Mississippi River flows into the Gulf of Mexico” and “river is a type of water body.” If represented in graph form, the subject is a node connected by an arc that describes the relationship from the subject to either a literal value or a node representing another resource. Because the subject of one triple could be the object of other triples, as well as being the subject of more triples, an interconnected graph can be formed.

Each component of a triple (i.e., subject, predicate, or object other than a literal) is identified by an International Resource Identifier (IRI). IRIs are similar to URLs for web pages and uniquely identify a resource. Linkages are formed between triples when identical IRIs have been used for a resource or when IRIs are equated using owl:sameAs. In this manner, linking can be done between many domains and knowledge bases. As more and more data are linked, a larger graph is created enabling further relationships to be explored.

In the color version of the LOD diagram (http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.png), geospatial data are yellow and include Census data, GeoWordNet, Linked Geodata, GeoNames, and many more. GeoNames (www.geonames.org), for example, has information on over six million places and geographic features. Because place is a natural integrator, if data providers re-use the identifiers specified in GeoNames, data from various domains can be linked through place, enhancing the overall knowledge gained from the LOD project.

Various projects are creating RDF data related to the geospatial domain. For example, the United States Geological Survey (USGS) has put some of The National Map data into the RDF GeoSPARQL format (Varanka 2012). In the Consortium for Ocean Leadership, metadata was put into RDF to facilitate search over multiple facets having complex concepts, logic, and frameworks. Another project put weather sensor data into RDF and linked places to GeoNames.

### Storing and querying RDF data

RDF data are typically stored in triplestores, such as Allegrograph, Parliament, Jena, Virtuoso, Sesame, or others. RDF data can also be stored in tables in relational DBMSs, although Oracle, for example, also has native support for RDF graphs. Triplestores may provide a graph-based query language, such as SPARQL, the W3C standard RDF query language (http://www.w3.org/TR/rdf-sparql-query/). Further, an endpoint (a service that accepts SPARQL queries and
returns results) may be set up through which RDF data can be queried over the web.

An example SPARQL query to find all state parks is:

```
SELECT ?sp
WHERE
  ?sp a ex:StatePark .
```

Briefly, the variable sp is being stated to be of type state park. All features from the example database (ex) that are of type state park will be returned.

**Querying spatial RDF data:**

**GeoSPARQL**

Because SPARQL does not support spatial operators, a few spatial RDF query languages have been developed. For example, GeoSPARQL was developed as an extension of SPARQL (Battle and Kolas 2012) and is now a standard spatial RDF query language. It was jointly developed by the OGC and the W3C, with OGC providing the spatial expertise. GeoSPARQL’s design has a class SpatialObject that has two subclasses: feature and geometry. Separating the feature from its geometry allows a feature to have multiple geometries. The object property between the feature and geometry classes is called hasGeometry. For example, if an application has a feature type with a spatial representation, such as parks, parks would be modeled as a subclass of GeoSPARQL’s Feature class and would inherit the hasGeometry spatial property. The geometry class accommodates WKT (well-known text) and GML (geography markup language) representations. GeoSPARQL also accommodates topological relations expressed in any of three forms: OGC simple features, Egenhofer’s 9-intersection model, and region connection calculus-8 relations (RCC-8). These relations include spatial operators such as equals, disjoint, intersects, touches, within, contains, overlaps, crosses, and so on. GeoSPARQL also has additional spatial operators, such as distance.

An example GeoSPARQL query to find towers within state parks is:

```
SELECT ?t ?sp
WHERE {
  ?t a ex:Tower ;
  geo:hasGeometry ?tgeo .
  ?sp a ex:StatePark ;
  geo:hasGeometry ?spgeo .
}
```

This query specifies additional variables for the geometries, tgeo and spgeo. Using GeoSPARQL’s **within** operator, the result of the query will only consist of towers that are within state parks along with the state park containing them (i.e., not all state parks).

SPARQL and GeoSPARQL queries can become hard to write. Because of this and to allow more users to participate in a Geospatial Semantic Web, a graphical query interface, built specifically for spatial data and GeoSPARQL queries has been developed (Grove et al. 2014).

**Geospatial Semantic Web**

Initial work on technology for the Semantic Web did not include spatial considerations, although referring to spatial locations, such as to find the closest doctor’s office, was in the original Semantic Web vision. But, initial work on ontologies, linked data, and query languages, for
example, did not provide for spatial operations. Egenhofer (2002) introduced an early description of a Geospatial Semantic Web. He suggested that spatial functions be accommodated to be able to perform true spatial operations over the web.

In 2006, OGC had a Geospatial Semantic Web interoperability experiment (http://www.opengeospatial.org/projects/initiatives/gswie) to integrate Semantic Web technology into OGC web services. In the goal of creating an architecture that would allow automatic web processing of queries, a target query was to find airports in a specified area of the world that could handle a C-5 cargo plane. The solution included developing ontologies and rules to represent an understanding of the various domains that were involved and to resolve semantic heterogeneities. They also implemented interfaces and profiles to extend OGC web services. Their work helped contribute to the vision of a Geospatial Semantic Web.

Now, with spatial RDF query languages, such as GeoSPARQL or stSPARQL, accommodating the execution of spatial operators, a Geospatial Semantic Web is further enabled. Spatial queries can now be posed on the web over RDF data. As the next phase, to fully realize the vision, more data need to be put on the web in the RDF format.

Ontology design patterns (ODPs)

Creating a comprehensive ontology can be a large undertaking. Because of this, but also in their own right, ontology design patterns (ODPs) are being developed. ODPs are smaller ontologies intended to be re-usable that express overview, partial, or repeatable patterns. Patterns may be high level and generalizable across domains. An example use case is a generic route with beginning and end points that can be used to model various applications ranging from bus routes to bird migrations (Hu et al. 2013).

Ontology repositories

To promote semantic integration, ontologies should be re-used. Making existing ontologies available is the goal of the open ontology repository (OOR) initiative. The OOR initiative
ONTOLOGY: DOMAIN APPLICATIONS

aims to provide sharing of ontologies through modular open source software that uses best practices. As part of the OOR initiative, different repository code bases have been written to store and allow access to ontologies. Metadata for ontologies was developed to be used in searching for and browsing ontologies in these code bases. The repositories also provide for finding individual terms within ontologies and the ability to map terms between different ontologies. Further, the repositories provide web services to be able to access contents remotely.

In the OOR initiative, some ontology repositories are based on the BioPortal ontology repository code (http://bioportal.bioontology.org/), which was developed for biomedical ontologies. A few experimental instances of the BioPortal code were implemented by others including one by SOCoP (Spatial Ontology Community of Practice). The purpose of SOCoP's OOR was to collect ontologies related to the geospatial domain. SOCoP's ontologies have since been moved to Ontohub. Example ontology repository efforts that are part of the OOR initiative but that use different code bases are COLORE (http://stl.mie.utoronto.ca/colore/) and Ontohub (https://ontohub.org/).

Ontologies collected that are relevant to the geospatial domain include upper level ontologies such as DOLCE, SUMO, and BFO as well as domain-oriented ontologies such as SWEET (http://sweet.jpl.nasa.gov/), which encompasses the earth science domain. Other examples are GeoNames, Spatial Relations, Land Use and Land Cover Change (LUCC), Semantic Sensor Net, hydrography, ISO standards, and others.

Future

The use of ontologies and semantic technologies is in its beginning stages in general as well as in its beginning stages in the geospatial domain. New technologies are actively being developed, and a few prototype systems have been built. But semantic systems are not yet pervasive. More work is needed to build semantically enabled spatial systems and a Geospatial Semantic Web. And, more ontologies are needed to represent geospatial data and processes so they can be accessed from semantic systems. Also, more data need to be created in, or translated to, the RDF linked format so that more data are available to be queried over the web.

SEE ALSO: Data structure: spatial data on the web; Geospatial Semantic Web; Ontology: theoretical perspectives; Spatial feature classes

References


Further reading


Ontology: theoretical perspectives

Helen Couclelis
University of California, Santa Barbara, USA

Background

“Ontology” is a relatively recent addition to the geographer’s vocabulary. In its traditional meaning as the study of the essence and structure of what exists, it is a branch of metaphysics that goes back at least to Aristotle and continues as an important theme in philosophy to this day. In modern times the term has acquired certain specialized meanings beyond pure philosophy, as in parts of theoretical physics, where the ontology of space–time and that of quantum fields have been active research topics for several decades. More recently ontology was also introduced in computer science and more specifically in information science, where applications-oriented as well as theoretical versions can be found. The informational version of the concept is, however, much less ambitious than the original: according to the most widely accepted definition, an ontology is “a specification of a conceptualization,” to be used for clarifying the meaning of information items in the context of formal knowledge representation. In other words, the purpose of these ontologies is to enable computers to turn the massive volumes of available information into usable, sharable, actionable knowledge (Guarino 1998). To avoid confusion, these two quite different approaches are often referred to as “Big-O” ontology and “small-o” ontology, respectively. Both the traditional philosophical sense of ontology and the recent informational usage are represented in contemporary geography. The plural form “ontologies” is commonly used in connection with the small-o, applications-oriented kind, and more generally, in reference to the many alternative ontological frameworks that have been proposed over the centuries and in our time. In all cases an ontology is, at the very least, a classification of entities, concepts, or terms of interest, although relations (e.g., hierarchical relations) are typically included along with properties and other structural elements.

Ontology in the philosophical sense is represented in the work of several generations of geographers especially in humanistic and critical geography. These have been influenced by the writings of mostly European major philosophers and thinkers such as Kant, Hegel, Husserl, Heidegger, Habermas, Sartre, and, most notably, Marx, but also by contemporary scholars such as the British philosopher of science Bhaskar and more recently also by certain feminist thinkers. Small-o ontology on the other hand was introduced in quantitative geography in the mid- to late 1990s by researchers in geographic information science, following progress on the topic in the related areas of geoinformatics, and in information science and technology more generally. Some early evidence of ontological thinking is found in the Harvard Papers in Geographic Information Systems from the 1970s (Chrisman 1978), when researchers were exploring the nature of geographical space in the context of the nascent field of computer cartography. Clearly Big-O and small-o ontology have played different roles in the different branches of geography in which
they were adopted. The former has inspired new theoretical analyses and discourses, while the latter, originally meant primarily to facilitate the efficient and consistent retrieval and use of information across databases, is also increasingly raising philosophical questions of its own. In both cases geographers have made original contributions to ontology by exploring and foregrounding the role of space and spatiality in spatiotemporal phenomena and processes, whether natural or social. Despite major differences in interpretation and use that are reflective of the discipline’s diversity, the relevance of ontology for geography as a whole lies in its being one of the rare theoretical concepts that are common to the humanistic, critical, and quantitative streams of the discipline, with the potential to bridge disparate traditions.

**Philosophical foundations**

**Ontology and general human geography**

A number of the philosophical issues and debates surrounding ontology matter to geography. This is true of humanistic and critical geography and surprisingly, perhaps even more so of the geoinformatic (computational) sense of the notion as adopted in geographic information science and quantitative geography. In both cases the central issue is the question of representation: what exactly is being represented and how, and what the implications are for ontology development and use of the answer one gives.

Ontologies take on different forms and mean different things depending on their developers’ acknowledged or unacknowledged philosophical stance. For the purposes of this discussion the most relevant contrast is that between realist and nonrealist philosophies and the ontologies these tend to engender. Realism – which is a position, not a distinct philosophy – is based on two premises: (i) that the things and relations in the world really exist (more or less) such as they appear and with the properties that they appear to have; and (ii) that the reality of things and relations is mind-independent, that is, independent of how we observe, conceptualize, or talk about them. Nonrealist philosophies, of which there are many (for example, conceptualism, constructivism, cognitivism, conventionalism, idealism, instrumentalism, nominalism, poststructuralism, rationalism, perspectivism, phenomenology, and pragmatism), contest one or both of these premises, pointing to inconsistencies and other problems. Much of the critique of realism revolves around its inability to properly account for abstract notions not directly related to observable things, for morality and consciousness, for mathematics, for the role of context in assigning meaning to phenomena, and generally, for neglecting the relation between mental and social life on the one hand and conceptions of the world on the other. Strong versions of realism can be quite doctrinaire, a consequence of the belief in a single objective reality, and thus in only one true ontology. Realist philosophies also draw a sharp distinction between ontology and epistemology (what exists versus what we can know about what exists), whereas the distinction is much less clear in nonrealist philosophies to the extent that these emphasize the constitutive role of the mind, of context, and of situational interests in our conceptualizations of the world.

Ontological thinking in critical geography tends to be realist to the extent that it draws on earlier, strongly realist interpretations of Marx’s work, which was often seen as describing the “hard science” of society. The Frankfurt School’s cultural reformulation of Marxism has softened the edges of the original hard realism by refocusing Marxism on individuals and society and on the process of knowledge creation as part of social reproduction. Since the 1970s Bhaskar’s (1997/1975) critical realism has been very
influential as an ontology of the social sciences, and has been adopted in parts of post-Marxist geography and beyond. Feminist geographers have also involved ontology in their writings, largely in the context of their critique and rejection of Cartesian, “masculinist” ontologies of dualism and opposition (e.g., Stanley and Wise 2002). Both realist and nonrealist influences are found in the feminist literature, which is too diverse to pigeonhole. Humanistic conceptions of ontology on the other hand are more likely to be influenced by nonrealist philosophies such as Kant’s transcendental idealism or the phenomenology of Husserl, Heidegger, and Merleau-Ponty. For example, Pickles (1983) presents a strong critique of the physicalist ontology imported into human geography through positivist research and writings, and goes on to show how phenomenology can help uncover the distortions brought about by positivism while also illuminating spatiality as constitutive part of the human world.

This all is in stark contrast to poststructuralist and postmodern discourses, which tend to be explicitly antirealist and leave no room for visions of reality structured around stable ontological cores. More recently, however, the work of the French philosopher Deleuze (e.g., Deleuze and Guattari 2007) has led to a revival of interest in ontology among critical and cultural geographers and in the social sciences more generally, while also providing theoretical support for “nonrepresentational theory” and perhaps other intellectual movements in the discipline such as “actor-network theory,” or ANT. In any case, ontologies in general human geography, where they exist, are conceptual and often sketchy, unlike the elaborate, highly detailed, formal constructions of information science that are topics of rapidly growing interest in geographic information science and parts of quantitative geography.

Geographic information science and knowledge representation

Philosophical questions and debates made their appearance fairly early on in the development of computational versions of ontology (small-o ontology). Ontologies were originally introduced as tools for facilitating the integration of data from different databases (“interoperability”) so as to allow users to formulate queries about some particular topic that may be answered with data gleaned from more than one source. This can be difficult since data and data structures for the same (or same type of) phenomenon may differ significantly depending on when, where, why, how, and by whom the data are collected, organized, and documented. For example, water quality data collected by a water-supply agency over the summer months may be quite different from data collected from the same water source for research purposes at some other time. Likewise, mountains in the United Kingdom, for example, are known to have very different characteristics from mountains in Central Asia and rivers in northern Europe are very different from rivers in southern Europe so that creating a “world mountains” or a “European rivers” general-purpose knowledge structure is not that simple. Very soon it became clear that there were issues in ontology development reaching far beyond technical or definitional discrepancies and questions about interpretations and meanings, about contexts, about user needs and perspectives, about language and cognition, about place and culture, about the proper treatment of space, time, and information, and so on, became pressing. Increasing numbers of geographers are contributing to these debates through the angle of their own subfield, be it spatial modeling, spatial cognition and behavior, cultural linguistics, spatial decision-making and planning, physical geography, or some other substantive interest.
Perhaps most fundamental of all is the question of representation: what does a small-o ontology really represent? And what are the implications of the ontological commitments one makes in adopting this or the other ontology? While philosophy per se is not a common topic of discussion in the information science community, the contrast between realist and nonrealist positions is stark. On the one hand are those who subscribe to ontology in the traditional metaphysical sense – which is the representation of the things and relations that make up the real world – while making allowances for the distortions of context and cognition. The search is for truths that correspond to what “really” is out there, meaning nature, society, the economy, and so on. An interesting form of realism foregrounds cognition as the more tractable, measurable, and ultimately relevant reality on which to base ontology development. Thus several researchers are attracted to an approach from anthropology and cognitive linguistics known as experiential realism, according to which humans, qua embodied beings, can only know reality through “image schemata” deriving from their bodily experiences in the world. These schemata are expressed in a small number of elementary metaphors such as “container,” “path,” “up-down,” or “surface,” which seem to underlie much of our thinking (e.g., “up” is good, “down” is bad, life is a journey, a place is a container, and so on). Experiential realism offers a workable compromise to researchers with milder realist leanings because it implies that the intersubjective, cross-cultural, invariant reality of cognitive schemata can substitute for the reality that cannot directly be apprehended by humans. The obvious spatial nature of these schemata is an additional attraction (Mark and Frank 1996).

Related to experiential realism is cognitive semantics, another popular approach to ontology development that is also rooted in cognitive linguistics but lacks the former’s emphasis on bodily experience. Cognitive semantics relies instead on the notion of the “semantic triangle,” which is meant to represent the relationships between real-world things, their representations as concepts in a person’s mind, and the signs (words, icons, or any other symbols) used to represent these concepts. Being more remote than experiential realism from the physical reality of human bodies, cognitive semantics is compatible with both realist and nonrealist interpretations.

On the clearly nonrealist side, researchers less interested in cognition or philosophy may simply draw on their own training, practical experiences, and intuitive understandings to produce something that works in practice, resolving conceptual problems as they go. This is the classic engineering approach that views the design of artifacts (material or symbolic, the latter including ontologies) as solutions to constraint satisfaction problems. It is consistent with a view of informational ontology as the representation of the things and relations that make up not the real world itself, but the simplified, abstract artificial worlds that researchers and professional create with their quantitative and logical models. Under this interpretation engineering ontologies may be seen as ontologies of “artificial micro-worlds,” not claiming correspondence with the actual world except in limited, specific areas and for particular purposes. The best ontologies are simply the ones that help solve the most problems.

Both this pragmatist position and the realist one represent coherent conceptual or philosophical views and there are strengths and weaknesses to both. The realist position (more specifically, naïve realism – a technical term for the belief that we see reality as it really is) is the default position in the natural and social sciences: young people are attracted to the (empirical) sciences from a desire to “discover how nature (or society) works,” though this is by no means true
of all scientists. Indeed, a number of nonrealist philosophies of science were propounded by theoretical physicists, mathematicians, and others. But, by and large, the realist view of informational ontologies is in better agreement with most of empirical science. On the downside, small-o ontology development is arguably not an empirical science, but the discipline of designing tools to assist empirical scientists in their work with information. Realism in this case risks conflating two distinct levels: that of empirical science modeling and the meta-level of ontology modeling meant to assist empirical science modeling. This compounds the standard problems of realism mentioned earlier. In any case, since informational ontologies are about knowledge representation, they are closer to epistemologies, technically speaking, than to ontologies. As for the “artificial micro-worlds” interpretation, it has the advantage of justifying the “ontology” designation for something that is clearly an insufficient representation of the world as we understand it, while also encouraging more exploratory and creative approaches, not bound by the search for a single kind of truth. On the negative side, many geographers trained as empirical scientists may have trouble accepting a view of ontology that openly claims to not be modeling the real world. Yet it should be fully consistent for someone to be a pragmatist at the tools design level and still be a realist at the empirical level. An enlightening debate among small-o ontology developers of contrasting philosophical persuasions can be found in a special issue of the journal *Applied Ontology* (Smith and Ceusters 2010).

**Theoretical foundations**

Traditional computational ontologies are of two major kinds: foundational or top-level ontologies organize very general concepts that in principle cut across domains, while domain ontologies are adapted to the specialized concepts and requirements of particular applied domains, such as biology, medicine, law, or spatial recognition of ontology as a research theme in geographic information science may be traced to a 1998 workshop on “the Ontology of Fields,” organized by the National Center for Geographic Information and Analysis (NCGIA) in Maine (“fields” here refers to one of two principal ways of representation in GIS, the other one being “objects”). Geographic information science ontologies (geospatial ontologies for short) are strongly influenced by geoinformatics and by developments in computer science more generally. Ontologies addressing spatiotemporal phenomena are often adaptations of more general ones developed in the engineering disciplines, but original contributions are also beginning to appear out of geography. The growing literature on the subject includes several different forms, from conceptual sketches and frameworks and general theories to detailed domain ontologies addressing information needs in particular subfields. The “ontological turn” in quantitative geography has served to bring conceptual and philosophical issues back to the forefront of the discipline after many years of sometimes narrowly conceived technical advances in spatial modeling and GIS. It is yet to be determined whether the promise of increased clarity, relevance, and rigor that is attracting so many geographers and others to geospatial ontology research will be realized in terms of a better articulated, more powerful, more “interoperable” geographic information science.

**Computational ontologies and geographic information science**

Following several years of development in the United States and elsewhere, the official
decision-making (e.g., Spatial Decision Support Consortium 2008). Recently the idea of a third major kind of ontologies, lightweight or micro-ontologies, has been gaining ground (see next section). Well-known foundational ontologies include the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), the Basic Formal Ontology (BFO), the General Formal Ontology (GFO), the Unified Foundation Ontology (UFO), the popular SUMO (Suggested Upper Merged Ontology), and others. Some of these, especially the first two, have been used and in some cases extended by geographers. As may be gathered from its full name, DOLCE is built around concepts in their relation to language and cognition, whereas BFO is a realist ontology naming and organizing the categories of entities thought to exist in the world. Ontologies are axiomatized in some formal language that enables computational modeling and reasoning. Their conceptual structure is often represented graphically in the form of trees, lattices, concept maps, or other kinds of graphs. Most are hierarchically structured, placing the more general concepts or classes at the top, with subsequent levels becoming increasingly specialized (Figure 1).

Major themes

A number of major themes cut across most foundational ontologies. At the very core of the idea of (small-o) ontology is semantics, the notion that linguistic terms and any other kinds of symbols have meanings that associate them with the entities they refer to and with other symbols, and that these meanings may not be fixed. Semantics is of course fundamental to any ontology, but unlike traditional ontologies that may capitalize on the tolerance of natural language for vagueness and ambiguity, the informational kind must specify meanings rigorously enough to be conveyed to computers. Incompatible semantics is indeed one of the main reasons why interoperability among databases is a challenge, and why computational ontologies were developed in the first place. In the geospatial domain Mark (1993) was among the first to point out that meaningful (i.e., semantically aware) standards are necessary for successful data transfer. The purpose of the ontology is to specify these meanings clearly enough so that they may be formalized, thus allowing semantic compatibility to be verified and the meanings to become computationally tractable. Some people use the term “ontologies” for simple concept taxonomies or even for map legends, but these are not true ontologies at least until the “formalization gap” is crossed.

An important distinction that, like ontology itself, has preoccupied philosophers for the past couple of thousand years is that between universals and particulars. Universals are mind-independent entities that are part of reality and typically reflect the essential qualities of particulars (e.g., the “mountain-ness” of mountains and the triangularity of triangles), while particulars (or individuals) are the specific instances of universals. Universals thus characterize realist ontologies, while nonrealist ontologies deal with concepts and their instances, with linguistic terms, or with things as cognized by humans in particular domains or situations. Practically this means that strongly realist ontologies must first specify the immutable properties of universals before particulars can be handled, whereas ontologies that start with less absolute notions of concepts in agents’ minds, language, or individuals as members of classes have an easier time adapting specifications to the requirements of particular contexts of use. In fact, the widely accepted definition of (informational) ontology as “formal, explicit specification of a shared conceptualization,” further clarified as “a description … of the concepts and relationships that can exist
for an agent or a community of agents,” clearly leans on the side of nonrealist interpretations (Gruber 1993).

Of particular interest to geographers and others dealing with geospatial ontologies are of course the concepts of space and time. Few foundational
ontologies model space and time directly, but most include spatial and temporal concepts (e.g., notions of spatial and temporal qualities), or more general concepts such as regions and intervals that can be specified for spatial and temporal domains. This allows forms of treatment of space (and time) similar to those familiar from GIS and quantitative geography, with the already formalized notions of geometry, scale, resolution, and granularity, of fields and objects and their combinations, of discrete and continuous spaces, of time as fourth dimension, and so on, to be easily represented in most ontologies. Recent proposals in geographic information science for elementary building blocks that integrate spatial, temporal, and substantive characteristics can also be accommodated.

Relating to time, many foundational ontologies adopt a high-level conceptual distinction that is quite unlike anything familiar from geography (see Figure 1). Based on a proposal by the major philosopher of science, Whitehead, entities are commonly distinguished into continuants and occurrents (the terms endurants and perdurants are also used, though there are certain semantic differences between the two pairs). Continuants are the entities that remain stable through time (e.g., ordinary objects or most abstract entities), while occurrents are events, processes, and phenomena that unfold through time, albeit with the participation of continuants (e.g., a hurricane, a university education, a war, a life). However, the continuant–occurrent distinction may be relative to scale, context, and perspective. For example, a traffic jam may be described as an occurrent – something that happens in time and then dissolves – but it can also be described as a continuant – a relatively short-lived, slow-moving object composed of many cars. Users can of course decide which representation is more suitable for a given phenomenon of interest, but forcing a choice where both views might coexist is not fully satisfactory.

Generally there is room for improvement in how most foundational ontologies handle spatiotemporal concepts. Kavouras and Kokla (2008) discuss the reasons for that state of affairs, identifying three problem areas: the lack of a widely accepted geospatial knowledge corpus; the lack of clear objectives for geospatial semantic integration; and the fact that the traditional structures and tools of geographic information science were not developed for the requirements of geospatial knowledge systems. Despite major recent progress in the areas in question, problems remain.

Critiques and recent developments

Along with engineering ontologies in general, geospatial ontologies are still very much in flux. This is partly because of the relative novelty of the field, but also because of their close relation to computer science, which is itself undergoing continuous technical and conceptual transformations of momentous significance. The standard foundational ontologies are already considered outdated in some quarters and calls for “lightweight” or (micro-)ontologies are increasingly heard, sometimes from the very developers of the former (Davies 2011). This change of heart reflects both the past experience of trying to work with monolithic structures ill adaptable to changing needs and the future vision of versatile, easily developed tools that can be tailored to specific user purposes and are suitable for mining the disorganized riches of the rapidly expanding Semantic Web. Lightweight ontologies and formal languages are being developed specifically for the Semantic Web, and should greatly facilitate publishing, finding, sharing, using, and collaborating on information that is
“out there,” somewhere on the web, outside of properly structured and documented databases.

Critiques of the state of the art in geospatial ontologies address missing functionality and other shortcomings. To begin with, it may not even in principle be possible to develop an ontology that will satisfy most people most of the time. Basic conceptual, even philosophical, problems (such as that of realism versus nonrealism) linger. Furthermore, it is difficult to adapt any foundational ontology to the changing practical context of applications, even within a given domain; there is very little work as yet on the ontology of networks, a theme of growing importance in a connected world; adequately supporting the specification of complex dynamic phenomena within an ontology remains a challenge; and lightweight ontologies, despite their great promise, are still too young to be very powerful.

Partly in response to such internal critiques the research frontier in ontology development is represented by a number of efforts, theoretical as well as technical. Of great interest to geographers is work seeking to formalize the notion of place. The practical impetus is provided by the growth of location-based services and other aspects of mobile computing, but this search for clarification of a key geographical concept may also help build bridges between thus far separate domains of the discipline. How to account for the role of context in specifying semantics and in identifying the corresponding most appropriate data and how to make ontologies more responsive to specific user purposes and needs are two further research themes of general interest.

On the technical side, the fast-moving research frontier seeks to adapt ontology development to the even faster developments in computational power and data availability and to the expanding variety of types of data sources and range of applications. Two opposing trends are notable. On the one hand, a major national-scale effort is being deployed in the United States and elsewhere aimed at developing a geospatial cyberinfrastructure within which ontology is expected to play a major role (Brodaric, Fox, and McGuinness 2009). As part of this effort the US National Science Foundation also sponsors the “Earth Cube” initiative, which seeks to create an integrated data management infrastructure across the Geosciences. On the other hand, there is a trend toward downsizing – in addition to the lightweight ontologies mentioned earlier, a topic of growing interest is the development of ontology design patterns (Gangemi and Presutti 2010). These are formalizations of common concepts that are used in different domains and can be largely standardized with no loss of generality. Several such patterns have already been developed in geographic information science, for example, for the notions of “Point of Interest,” “Transport,” and “Trajectory” (Hu et al. 2013). Note that “transport” and “trajectory” are actually image schemata (see above), concepts that lend themselves particularly well to treatment as patterns (Kuhn 2007). Further, as part of making ontologies accessible to broader communities of users, user services such as visualization, editing, versioning, documentation, and search and collaborative utilities are also being constantly expanded and improved. Integrated platforms such as Stanford University’s “Protégé” aim to make ontology accessible to a wider professional public. These initiatives rejoin the broader efforts for the democratization of “big data,” also reflected in the development of the Semantic Web, which view ontologies as the premier tools for helping turn information into knowledge.

**Evaluation**

The place of ontology in geography may be evaluated from two different perspectives. On
the one hand, interest in geospatial ontologies is growing rapidly within geographic information science and related areas. On the other hand, in general human geography, the popularity of explicitly anti-ontological philosophies and influences, such as poststructuralism, deconstruction, and postmodernism, has to a large extent pushed aside the earlier humanistic and Marxist engagement with ontology. Post-Marxism continues to count some forms of realist ontology among its tenets, and certain feminist authors also present alternative ontologies, though the feminist emphasis tends to be on epistemologies. But for a while it appeared that traditional Big-O ontologies would no longer be an active theme in the discipline. Recent enthusiasm in parts of critical geography for Deleuze’s (Deleuze and Guattari 2007) innovative metaphysics of dynamic “assemblages” in space and time may presage a revival of general geography’s interest in ontology, with further such intellectual engagements to come.

The postmodern turn away from Big-O ontology that roughly coincided with the development of geospatial ontologies may be the reason why there has been very little critique of the latter coming from the side of critical geography, unlike what had been the case with GIS a couple of decades ago. In fact, Schuurman (2006) presents a cogent defense of ontology research in geographic information science from the perspective of critical GIS. She notes the great progress achieved in GIS in meeting these earlier critiques and considers ontology to be the latest stage in the development of a more conceptually mature and technically rigorous GIS. This is an important point because, one way or another, ontology is here to stay. It can greatly increase geography’s stakes in the information age, but only as long as societal issues are not lost in code. Here humanistic and critical geography can again play a major role, as they did before in the case of GIS, to the discipline’s major and lasting benefit.

SEE ALSO: Critical geography; CyberGIS; Data structure: spatial data on the web; Geographic information science; Humanistic geography; Ontology: domain applications

References


Further reading


The Open Geospatial Consortium (OGC) is an international voluntary consensus standards organization founded in 1994. In the OGC, hundreds of commercial, governmental, non-profit, and research organizations worldwide collaborate in an open consensus process to develop and promote the implementation of open standards for geospatial content and services, geographic information system (GIS) data processing, and data sharing.

History and purpose

In the early 1990s, sharing GIS data was very difficult and error prone. Each GIS vendor had proprietary data formats, application silos, and proprietary data conversion tools. Movement of data from one system to another was difficult and required the use of data translators and labor-intensive data conversion tools. Further, GIS procurement language often stated mandatory requirements that the provider must support specific proprietary geospatial data formats. This restricted open competition and increased the potential for users to become “locked in” to a specific proprietary product architecture. The net result was limited user choice, higher costs, and increased difficulty in sharing data in times of need. Perhaps worst of all, the GIS climate of the early 1990s severely restricted the market potential of the technology.

During the same time period, leaders of the open-source GRASS (Geographic Resources Analysis and Support System) community formed a nonprofit organization – the Open GRASS Foundation (OGF) – chartered to stimulate private sector support for GRASS and create a consensus-based membership process for the management of GRASS community affairs. Soon this group began looking at the issue of interoperable geoprocessing. Instead of focusing on open-source software, the group aimed to create an inclusive, open process that might (i) encourage the development of additional commercial as well as noncommercial geoprocessing choices in the marketplace, (ii) act as a sounding board for the user community to articulate its requirements to the developer community, and (iii) speed up procurement by aligning the needs of users with the product plans of vendors. The OpenGIS Project, which preceded the formal launch of the Open GIS Consortium, Inc. (OGC) in 1994, defined a vision of diverse geoprocessing systems communicating directly over networks by means of a set of open interfaces based on the “Open Geodata Interoperability Specification” (OGIS).

In response to the critical industry need for open, freely available standards to enable geospatial (GIS) data sharing, the Open GIS Consortium (OGC) was founded in September 1994 by eight charter members with this vision and mission:

- Vision: Realization of the full societal, economic, and scientific benefits of integrating electronic location resources into commercial and institutional processes worldwide.
OPEN GEOSPATIAL CONSORTIUM STANDARDS

- Mission: To serve as a global forum for the collaboration of developers and users of spatial data products and services, and to advance the development of international standards for geospatial interoperability.

In 2004, the OGC directors voted to change the name of the organization to the Open Geospatial Consortium to reflect a growing scope of activities that involved not only GIS but also imaging, sensors, location services, and more. From 1994 to 2015, OGC membership grew from 20 to more than 500 government, academic, and private sector organizations. GIS vendors are involved along with information technology integrators, data providers, universities, nongovernmental organizations, research organizations, and companies at the cutting edge of location and mobile services. As described below, liaisons with other standards, professional, and educational organizations have been an important part of the value that OGC provides for its members and the public at large.

While the OGC has its roots in North America, it now has more members in Europe than in North and South America together. The OGC was a founding member of, and continues to participate in, the Global Spatial Data Infrastructure Association, and OGC staff and members present on standards topics at conferences around the world.

Standards

The OGC has a set of formal policies and procedures that guide the consensus process for discussing, developing, processing, and maintaining OGC standards. The OGC also supports a rapid engineering and prototyping program called the OGC Interoperability Program (see below). Many new OGC standards, or enhancements to existing standards, are tested, validated, and demonstrated in interoperability initiatives as part of this program.

Most OGC implementation standards are grounded in a set of conceptual models for such topics as geometry (spatial schema), coordinate reference systems, and observations. These conceptual models collectively are known as the OGC Abstract Specification. These are conceptual models that cannot be implemented directly in operational code. OGC implementation standards based on these conceptual models provide the requirements and rules for implementing specific interface, encoding, and protocol instances. For example, the OGC Abstract Specification provides a set of mathematical definitions for geometry (e.g., points, lines, polygons). The Geography Markup Language (GML) provides the Extensible Markup Language (XML) grammar for these geometry types that can then be implemented by developers and used by end-user applications. Abstract models are technology and platform neutral. OGC implementation standards are technology or platform dependent, such as having a requirement to use the Hypertext Transfer Protocol (HTTP). More information on this topic is available at http://www.opengeospatial.org/standards.

Continual evolution of geospatial standards is necessary to address emerging technology as well as consumer trends and requirements. These include not only requirements to address such issues as the growth of mobile computing and RESTful (Representational State Transfer) programming, but also specific spatial communication and processing requirements in application domains such as hydrology, aviation, and emergency response. New requirements are often reviewed in OGC Technical Committee Domain Working Groups. The actual revision and development of OGC standards take
place in OGC Technical Committee Standards Working Groups. There is a growing need for member representatives in these groups to cross-participate in the working groups of other standards organizations whose standards may have a spatial component.

Figure 1 is based on the following documents that capture several years of Spatial Data Infrastructure implementation experience and policy:

- INSPIRE
- ISO 19119 – Services. The taxonomy of services is based on services defined in this ISO standard.
- The New Zealand SDI Cookbook
- The SDI Cookbook, Chapter 10
- W3C/OGC Web Services Model (Find/Bind/Publish)

The OGC standards baseline comprises more than 37 standards (not all of which are shown in the figure), including interface and encoding standards for:

Catalog services:

- CSW – Catalog Service for the Web: access to catalogs of information about services (including data services), symbols, devices, and other catalogs.

Encodings:

- Filter Encoding – XML and KVP (Key-Value-Pair) encoding for expressing projections, selection, and sorting clauses in a query.
- GeoXACML – Geospatial eXtensible Access Control Markup Language.
- GML – Geography Markup Language: XML grammar for modeling and encoding geographical information.
- Observations and Measurements (O&M) – General conceptual model and encoding for observations and measurements.
- Open GeoSMS – How to location-enable Short Messages (SMS).
- SensorML – Sensor Model Language: Conceptual model and XML encoding for describing a sensor.
- Styled Layer Descriptor (SLD) – Rule-based symbology encoding.
- WaterML – XML encoding and exchanging data describing the state and location of water resources, both above and below the ground surface.
- Web Map Context – Defines a structure for communicating information about the server(s) providing layer(s) in a Web Map Service (WMS) query.

Data services:

- GeoPackage – A universal file format for encoding and sharing geodata.
- Simple Features – SQL (SFS) – Standard Structured Query Language (SQL) schema for the storage, retrieval, query, and update of feature collections.
- Sensor Observation Service (SOS) – Requests an observation from a sensor or sensor archive.
- Web Coverage Service (WCS) – Provides access, subsetting, and processing for imagery and other gridded data types.
- Web Feature Service (WFS) – Provides an interface allowing requests for geographical features across the web using platform-independent calls.
Figure 1  A web services framework for OGC standards. Reproduced by permission of the Open Geospatial Consortium.
Portrayal services:

- Keyhole Markup Language (KML) – XML-based language for expressing geographic annotation and visualization on existing (or future) web-based, two-dimensional maps and three-dimensional Earth browsers.
- Symbology Encoding – XML language for styling information that can be applied to digital feature and coverage data.
- Web Map Service (WMS) – A standard protocol for serving georeferenced map images over the Internet that are generated by a map server using data from a GIS database.

Processing services:

- Sensor Planning Service (SPS) – Tasks a sensor.
- Web Coverage Processing Service (WCPS) – Provides a raster query language for ad hoc processing and filtering on imagery and other gridded data types.
- Web Map Tile Service (WMTS) – Provides map image tiles.
- Web Processing Service (WPS) – Facilitates publishing and ease of client access to calculations or models that operate on spatially referenced raster or vector data.

The design of most of these standards was originally built on the HTTP web services paradigm for message-based interactions in web-based systems. The OGC membership strives to keep OGC standards relevant for developers who employ new advances in distributed processing, such as cloud computing and the Internet of Things. At the same time, there is a strong commitment to and much deliberation over issues such as backwards compatibility, minimizing technology risk, maximizing the value of technology investments, and using the standards process to support technology planning.

Organization structure

The OGC has four operational units:

1 Standards Program: In the OGC Standards Program the Technical Committee and Planning Committee work in a formal consensus process to develop and maintain member-approved (or “adopted”) OGC standards. The OGC Technical Committee runs the standards development process with a published set of policies and procedures to carry out its work. Much of the work and discussion that goes into an OGC standard takes place in the Standards Program’s Committees Domain Working Groups and Standards Working Groups.

2 Interoperability Program: The OGC Interoperability Program organizes and executes hands-on engineering initiatives to accelerate the development and acceptance of OGC standards. These include testbeds, pilots, interoperability experiments, interoperability support services, and concept development studies. Some OGC standards originate in the Interoperability Program.

Since 1999, OGC’s Interoperability Program has conducted a multitude of initiatives that build and exercise public–private partnerships designed to accelerate the development of emerging concepts and drive global trends in interoperability through rapid prototyping of new capabilities. Such initiatives:

- reduce technology risk through accelerating development, testing, and acceptance of interoperability standards
OPEN GEOSPATIAL CONSORTIUM STANDARDS

with the refinement of standards and best practices;
- **expand the market and improve choice** by encouraging industry adoption of new standards and best practices, ensuring market availability of interoperable solutions;
- **mobilize new technologies** through providing participants with real-world experience and a platform to innovate while driving early adoption of standards; and
- **provide a cost-effective method** for sponsors and participants to share expertise and development while gaining early marketplace insight and advantage.

3 Compliance Program: The OGC Compliance Program provides the resources, procedures, and policies for improving software implementation compliance with OGC standards. The purpose of the OGC Compliance Program is to increase system interoperability while reducing technology risks. Vendors gain confidence that they are providing a product compliant with OGC standards, which will be easier to integrate and market. Buyers gain confidence that a compliant product will work with

![Diagram](image)

**Figure 2** Where the OGC fits in the standards world. Reproduced by permission of the Open Geospatial Consortium.
another compliant product based on the same OGC standard, regardless of which company developed the product. The OGC Compliance Program provides a free online testing facility, a process for certification of compliant products, and coordination of a community of developers.

4 Communications and Outreach Program: OGC staff, working with members, provide press coverage, media announcements, publications, workshops, seminars, and conferences to help technology developers and users take advantage of the OGC’s open standards. Technical documents, training materials, test suites, reference implementations, and other interoperability resources developed in OGC Interoperability Initiatives are available on the OGC Network.

Collaboration with other associations and standards development organizations

The standards development work of the OGC cannot happen in a geospatial technology standards silo separated from the larger world of information technology (IT). Therefore, the OGC makes significant efforts to collaborate and liaise with other standards development organizations (SDOs). The OGC’s earliest SDO collaboration was with the International Organization for Standardization (ISO). Since 1994, the OGC has maintained a close relationship with ISO/TC 211 (Geographic Information/Geomatics). Standards from the ISO 19100 series comprise significant topics of the OGC Abstract Specification. Further, the OGC standards WMS, GML, WFS, Filter Encoding (FE), O&M, and Simple Features Access (SF) have become ISO standards. OGC’s submission of its standards to ISO in a formal process for ISO adoption helps many organizations meet policy requirements or preference for de jure international standards (see Figure 2).

The OGC has developed a set of formal relationships with other standards organizations, with the details of cooperation often codified in memoranda of understanding. OGC works with other SDOs to help standardize the representation, processing, and exchange of geospatial/location information in a consistent manner throughout the standards stack. The OGC works with a range of national, regional, and international standards bodies including W3C (World Wide Web Consortium), OASIS (Organization for the Advancement of Structured Information Standards), TM Forum, buildingSMART, and the IETF (Internet Engineering Task Force), with more than 50 alliance partnerships listed on the OGC’s Alliance Partners webpage. Some of these are associations that do not develop standards but that provide the OGC with important knowledge about the geospatial needs of particular industry and professional domains. The OGC’s relationships with these standards groups and associations provide their members and those of the OGC with opportunities to resolve interoperability issues and to participate in joint outreach efforts to promote (i) participation in standards development, (ii) vendor implementation of standards, and (iii) market uptake of products that implement standards.

SEE ALSO: Augmented reality; Environmental management; Geospatial Semantic Web; Interoperability of representations; Location-based services; Map projections and coordinate systems; Sensor networks, the sensor web, and the Internet of Things; Sensor networks, wireless
Further reading


Open innovation

Faïz Gallouj
Faridah Djellal
Lille 1 University, France

Innovation is an old and fundamental problem that concerns all social sciences, including philosophy, sociology, economics, history, and, of course, geography. It is described as a categorical imperative, a condition for the survival of businesses, companies, territories, and nations. The popular concept of “open innovation” (Chesbrough 2003) is used to review a number of key aspects of the innovation issue. The concept of open innovation is a useful metaphor, a mobilizing semantic innovation, rather than a robust theoretical concept. This metaphor is used to report on the major contemporary openings made by experts in “innovation studies,” and in so doing, a number of important concepts in this field are reviewed. The openings in question concern not only the modalities of organization and implementation of innovation as implied by the concept of open innovation, but also the content and forms of innovation as well as the sectors that engage in these activities.

Open innovation: nature and forms

The first question to address is the definition of innovation and its typologies. Although this question is an old one, it is generally believed that Joseph Schumpeter (1934) is the author who provided the most complete answer. It will be shown how he truly helped open up the definition of innovation. Other pathways toward openness that have supplemented this Schumpeterian tradition of openness will then be examined.

The Schumpeterian tradition of openness

In the work of Schumpeter there are many concepts that not only have become “common knowledge,” but have also become part of our everyday language and culture beyond the academic sphere. This is particularly true of the distinction between “invention” and “innovation” and the concepts of “creative destruction” and “Schumpeterian entrepreneur.”

Thus, according to Schumpeter, innovation, also called new productive combination, is distinguished from invention. The former is a discontinuous socioeconomic process, whereas the latter is a continuous noneconomic process. Innovation is thus the result of the socialization process of the invention, that is to say, its meeting with the market. This relatively rare but more frequent meeting in contemporary economies between two disparate worlds, the scientific and technical world and the socioeconomic world, is brought about by an individual with particular psychological characteristics: the entrepreneur. The “Schumpeterian entrepreneur,” as opposed to the manager (the operator), is a risk-taker, motivated as much by the prospect of profit as by hedonistic reasons (pleasure, sense of adventure and challenge, etc.). He or she has a maieutic function and is characterized by the ability to enrol and mobilize networks (partners, investors, etc.). The introduction of innovation by the entrepreneur (and then by research and development (R&D) departments of large enterprises) results in “creative destruction.”
whereby innovative companies prosper while others are doomed to fail. Innovations that cause creative destruction occur in clusters in which a major innovation or breakthrough induces a myriad of other innovations.

These different concepts establish certain boundaries for innovation. Schumpeterian openness lies in the typology of forms of innovation that he suggested. Indeed, while according to the traditional economic view there are only two types of innovation, product innovation and process innovation (with a focus mainly on the latter), the Schumpeterian view identifies five types: new products, new production methods, new markets, new sources of supply, and new forms of organization. The openness obviously arises because more types of innovation are considered, but also reflects the importance given to nontechnological innovations. After all, some of these types of innovation are nontechnological in nature (organizational innovation, new markets), and product and process innovation can also take on both technological and nontechnological features.

Temporal or historical openness

Neo-Schumpeterian analyses (Nelson and Winter 1982; Dosi 1982) opened up the concept of innovation in time or history. Innovation is no longer seen as “manna from heaven,” an artifact set in stone, but as an endogenous problem-solving process for the specific problems of an organization. The journey is as important as the destination (the result) and is part of the essence of innovation that is therefore a cumulative process in which different forms of learning (learning by doing, by using, by trying, by consulting, etc.) play an essential role.

The combination of these two characteristics of specificity and cumulativeness has resulted in a number of important concepts that may have operational or managerial value. The concepts of irreversibility, technological lock-in, and path dependency reflect the reduced technological freedom over time and the difficulty of going backwards once technological choices have been made. Path dependency can be, in some cases, a strategic problem for an organization. An organization that has been successful in its field and is reluctant to experiment with alternatives gets locked into a competency trap and remains on its cognitive and technological trajectory.

The concept of technological trajectory is defined as a possible direction taken by a technological potential based on a technological paradigm, that is to say, a general model of solution of selected technological problems, based on selected scientific principles and on selected technologies (Dosi 1982). Technological trajectories can be characterized in different ways. They can be powerful or weak, localized or general, complementary or competing. There are so-called natural trajectories, that is to say, technological developments that are necessary and seem to be inevitable (mechanization, miniaturization, etc.). This concept of technological trajectory is a key element of Pavitt’s taxonomy (1984) in which the economy is divided into four groups of firms with similar technological behavior (trajectories): supplier-dominated, scale-intensive, specialized suppliers, and science-based.

The openness of institutional definitions of innovation

Building on the theoretical Schumpeterian openness, international statistical institutions, in particular the Organisation for Economic Co-operation and Development (OECD), have worked to develop and improve innovation indicators for measuring and comparing the characteristics of innovation dynamics using national and international surveys. Thus, while
in its initial version the Oslo Manual (of innovation indicators) distinguished two types of innovation, technological product innovation and technological process innovation, the latest version (OECD 2005) adds two other forms: organizational innovation (new or significantly improved knowledge management systems, a major change to the organization of work, new or significant changes in the relations with other firms or public institutions) and marketing innovation (significant changes to the design or packaging of a good or service, new or significantly changed sales or distribution methods).

This revision takes into account some nontechnological forms of innovation in industry and services and is undeniably a breakthrough, but other forms of innovations remain hidden, in particular services and social innovation (see below). The innovation gap remains, particularly in the following fields: nontechnological product innovations (a new insurance contract, a new financial product, or a new field of consultancy expertise, for example), nontechnological process innovations (methodologies and protocols), ad hoc and custom-made innovations, innovation in public services, innovation in complex packages (also called new concepts), and new formulas (e.g. in commerce or the hotel industry).

Beyond economic innovation: social innovation

Another important opening in the understanding of innovation involves what is known as social innovation (Franz, Hochgerner, and Howaldt 2012). In the context of growing concerns about the need for sustainable development, the concept of social innovation is increasingly important. Its success (both operational and theoretical) is explained in particular by the chronic nature of the socio-economic crisis in developed economies since the 1970s, demographic changes (especially population aging in rich countries), the failure of development policies, and growing environmental concerns. However, this does not mean that the definition of social innovation is clear. Despite many efforts to clarify the notion, it still remains particularly vague. It is heterogeneous and eclectic and covers a wide variety of concepts. Social innovation is often defined in terms that describe its shape or nature, its process and its actors, its target and its purposes: it is said to be immaterial, nontechnological, organizational, noneconomic, nonmarket, informal, local, or designed to solve social problems. However valuable these characteristics may be in defining social innovation, they do not provide indisputable technical criteria. It is easy to find exceptions that refute them. They should simply be regarded as a general indication of the social nature of an innovation.

The opening of innovation toward services

The recognition of innovation in services is probably the most important step forward (especially in terms of public policy) made during the past decades. Innovation studies were developed mainly with reference to industry and above all concerned the dynamics of technological innovation. Services long remained unconcerned by innovation issues, as these two fields ignored one another. The recognition of innovation in services is not independent of the various other openings mentioned above. After all, social innovation is often a service innovation and an innovation in services, and it is primarily the innovation gap in services that is reduced by the shifts to a wider definition of innovation, whether theoretical or institutional (incorporated in the revision of innovation indicators).
Attempts to interpret and understand innovation in services have evolved in four different analytical perspectives, reflecting the increasing importance of innovation in services and of services in innovation (Gallouj and Djellal 2010). The technologist or assimilationist perspective views innovation in services in the same terms as industrial innovation, with emphasis on its relationship to technical systems. Insofar as it is concerned primarily with innovations adopted by services from industrial sectors, the assimilationist perspective is a perspective of subordination. The differentiation (or demarcation) perspective focuses on the distinctive characteristics of services and seeks to identify innovation activities that are invisible to traditional (assimilationist) systems. The first in-depth studies favoring a demarcation perspective focused on knowledge-intensive business services (all forms of consulting), as well as financial services (banking and insurance). The inversion perspective breaks with the idea that services are subordinate to industrial innovation, giving knowledge-intensive business services (KIBS) an active role in the innovation of their clients (especially industrial clients), in a cognitive balance of power that can tip in favor of KIBS. In this case, services, to some extent, dominate industry and not the opposite. At the meso- and macroeconomic levels, KIBS are seen as essential components of local, regional, and national innovation systems. Finally, the integration perspective aims to develop an analytical framework that takes into account both goods and services, both technological innovation and nontechnological innovation. This perspective is justified by the converging trend in contemporary economies toward the servitization of goods and the industrialization of services, that is, a blurring of the boundaries between the two.

Open innovation in terms of organization and implementation

Strictly speaking, the concept of open innovation is not a new development; rather, as already pointed out, it is a particularly effective semantic innovation that covers often more robust economic concepts (but also more complex or esoteric ones). Open innovation reflects an “extroverted” and cooperative way of organizing innovation that has many different configurations. This mode of innovation is characterized by a multiplicity of actors and interactions.

The open innovation model contrasts with the traditional linear model. The latter implies a sequential (and specialized) organization of the innovation process that greatly limits interactions and feedback. In management science, this linear model is illustrated by a well-established theoretical tradition that views the production of new goods or new services according to the new product (or new service) development methodology that uses planned and systematic processes.

Open innovation is an “umbrella concept” that covers multiple types of collaborative innovation processes that are more or less sophisticated, more or less long, more or less formalized, and so on. Thus, the evolution (opening) of the model is illustrated not only by the multiplicity of actors and the diversity and density of their interactions, but also by the varying duration of the innovation projects or facilities (short- or long-term projects, transient or permanent facilities) and the varying degree of institutionalization (formalized projects, informal projects, ad hoc projects). Accounting for this diversity is achieved by distinguishing (simplistically and doubtless with a certain abuse of language) two levels of analysis of the opening of the organization mode: the microeconomic level and mesoeconomic (or even macroeconomic) level.
The microeconomic level

At the microeconomic level, the general perspective of open innovation covers various cooperative models that are more or less sophisticated and formalized. Note that some of them are actually intermediate forms or isolated components of reticular or systemic forms mentioned in another level of analysis: the mesoeconomic level described in the next section.

The perspective of open innovation covers a number of nonformalized and unplanned or emerging models that have been seen in services but whose application is broader, such as the rapid application model, the practice-based model, bricolage innovation, and ad hoc innovation. In the rapid application model, planning does not precede production, as it does in the traditional linear model. Once an idea emerges, it is immediately developed as the service in question is being provided. Thus, the process of providing the service and the process of innovation are one and the same. The practice-based model involves identifying changes in service practices and developing and institutionalizing them. The bricolage model describes change and innovation as the consequences of unplanned activities carried out in response to random events and characterized by trial and error and “learning on the job.” Finally, ad hoc innovation can be defined as the process of constructing a (novel) solution to a problem identified by the client company. This interactive process requires the participation of the customer and is described as ad hoc because it is “unplanned” or “emerging,” which means that it is consubstantial with the process of service provision, from which it can be separated only in retrospect. Ad hoc innovation is recognized as such only after the fact.

Open innovation also includes the chain-linked, or interactive, model introduced by Kline and Rosenberg (1986). This well-known model is doubtless a solid theoretical formulation of the idea of open innovation. It challenges the linear (sequential) vision of innovation and instead suggests a model characterized by multiple actors and places of innovation and the key nature of the interactions between them, at different stages of the innovation process. Thus, unlike the linear model, Kline and Rosenberg’s model has several related paths: (i) a central chain of innovation going from the invention (or design) to development, production, and marketing; (ii) short or long feedback loops that link these sequences together; (iii) multiple links between the central innovation chain and knowledge and/or research chains; (iv) a relatively rare direct link between science and invention that leads to radical innovations (e.g., the laser, nuclear energy).

A configuration variously known as the consultant-assisted, or interactional, innovation model is also part of open innovation. This model describes the role of consultants, and more generally KIBS, in their clients’ innovation processes. KIBS are information and knowledge-processing machines that produce information and knowledge in the form of technological solutions, but also organizational, strategic, legal, fiscal, and other solutions. They constitute an external source of knowledge and innovation for their customers. This model is also sometimes called the Schumpeter 3 model in that it supplements the Schumpeter 1 model (innovation generated by the individual entrepreneur) and the Schumpeter 2 model (in which the spirit of enterprise is associated with the R&D departments of large firms).

The opening of innovation to the customer (the user-driven innovation model) is not a new phenomenon. Statistical analyses all show that customers are the main sources of innovation in businesses. However, customers and users are not all on the same level, and the literature identifies “lead users” who are strategic resources in
innovation projects. In modern economies, the involvement of customers or users is increasing as two new situations arise: (i) the promotion of research and innovation by customers or users in some areas and (ii) the strategies of some companies to involve customers, in different ways and to varying degrees, in the innovation process. Examples of this include the crucial role of patient organizations in research and therapeutic innovation, crowdsourcing, or the use of social media in business innovation.

The meso- and macroeconomic level: networks and systems

Innovation networks and the associated concept of innovation systems are opening up innovation even more. These concepts have been undeniably successful for two decades and they will first be defined before exploring the reasons for their success as both theoretical and operational tools.

An innovation network can be considered both as a structure and as an intermediate mode of coordination (between market coordination and hierarchical coordination). As a structure, an innovation network is a mode of organization dedicated to innovation that comprises a certain number of actors and the learning relationships between those actors. This structure can be described in different ways.

1 According to the nature of the actors involved (and the nature of the relationships between them).
2 According to the nature of the innovation – traditional innovation networks are essentially devoted to technological innovation.
3 According to how the structure is formed – a distinction is made between planned networks, which are set up and orchestrated by one agent, and spontaneous networks, which emerge in a self-organized way because of the convergence of the activities of agents confronted with a problem.
4 According to how the structure operates – it can operate in either a top-down or a bottom-up mode. It can also operate in “caretaker” mode (there is a hub actor who acts as systems integrator) or “noncaretaker” mode (responsibilities are shared).
5 According to the life cycle – innovation networks are born, reach maturity, and may die.

An innovation network involves the establishment of relationships based on trust, reputation, and mutual dependence between selected partners. As a mode of coordination between economic agents, it is more effective than hierarchical coordination (integration in the company) that reduces transaction costs, but brings with it the risk of bureaucratization, which may endanger innovation (a risk already mentioned by Schumpeter). It is also more effective than the market since it is difficult to establish explicit contracts for complex and uncertain research and innovation outputs and there is a risk that strategic secrets might be divulged.

Open innovation also covers new specific innovation networks (public–private innovation networks in services, or ServPPINs) that develop in a dominant service economy. These ServPPINs describe the collaboration between public, private, and third-sector service organizations in the field of innovation (Gallouj, Rubalcaba, and Windrum 2013). They differ from traditional innovation networks in several ways. First, relationships between public, private, and third-sector service organizations are pivotal to the arrangement. Next, service providers are the key stakeholders in the process. Lastly (and the corollary of the previous characteristic), nontechnological innovation (service innovation), often neglected in the literature, is also taken into account. Gallouj, Rubalcaba,
and Windrum (2013) provide a long list of ServPPINs case studies in four service sectors: health, knowledge-intensive services, tourism, and transport. These include public–private networks to develop innovation in elderly care, new tourism products, new training programs focused on patient-centered diabetes education, innovative mobility support services to help public transport users in their daily mobility by providing them with information, and so on.

The innovation network concept has been experiencing growing theoretical success for several decades. It is characterized by its simplicity and its flexibility and it has considerable heuristic value. Its theoretical (but also operational) success lies within its ability, on the one hand, to assimilate many other more basic analytical tools and, on the other hand, to incorporate itself into other, broader concepts that have been undeniably successful as well. For example, among others, the following basic concepts play a major role in the constitution and dynamics of innovation networks: learning (in its various forms: learning by doing, by using, by trying, by consulting, by interacting, etc.), absorptive capacity, economies of scale, scope and agglomeration, transaction costs, positive externalities (spillovers), particularly network externalities, proximity (geographical, cognitive, organizational, social, institutional), and so on.

On the other hand, the notion of the innovation network is the basis for all the concepts in the systems tradition, whether they be national, regional, local, or even technological and sectoral innovation systems, innovative milieus, technological districts, clusters, and so on.

Empirical studies devoted to identifying innovation systems and networks and to the analysis of how they function and perform have steadily increased over the past decades. The pioneering theoretical studies on national innovation systems (NIS) always include concrete cases. For example, Nelson (1993) provides a description and analysis of the NIS of 15 countries. The OECD has been the initiator of many studies that map the NIS profiles of member states (and the underlying networks) and compare their effectiveness. There are also a great many empirical studies devoted to regional or local innovation networks and systems. The recent empirical literature emphasizes the notion of a cluster, which is the new emblematic form of local innovation network or system. It makes a distinction between traditional or low-tech clusters (i.e., automotive, textile, chemistry) and high-tech clusters (information and communication technologies, biotechnologies, nanotechnologies, environment, aerospace and defence, multimedia, etc.). The former follow established technological trajectories, which are sources of incremental innovations with a focus on exploitation rather than exploration. Conversely, the latter are characterized by intense R&D efforts and radical innovations and follow new technological trajectories.

Public policy is another domain of success for system- and network-based approaches to innovation. Theoretical approaches to innovation not only provide tools for mapping existing systems and networks and measuring and comparing their performance, but also provide concrete tools for action (e.g. from a benchmarking perspective). Inspired by the various versions of the concept, policymakers have implemented strategies, especially with regard to spatial development and planning, that seek (from a systems perspective, whether geographic or sectoral) to encourage agglomeration – in other words, to strengthen or create innovation networks of varying sizes and degrees of complexity. Thus, while NIS and the corresponding networks form the basis of science and technology policies at both national and European levels, regional innovation systems and networks and clusters
OPEN INNOVATION

(such as the innovative milieus and industrial districts of past decades) are notions that today lie at the heart of local and national policies in many countries, even though they may be known by different names: “skills clusters” in Germany, “knowledge clusters” and “industrial clusters” in Japan, and “competitiveness clusters” in France.

SEE ALSO: Information and communications technology; Knowledge-based economy; Network analysis; Outsourcing; Producer services and economic development; Public policy; Science and technology parks; Technology; Technology spillover

References


Optical remote sensing

Guofan Shao
Purdue University, USA

There are various definitions of remote sensing but none are universally recognized. According to the Cambridge English Dictionary, the word “remote” means far away in distance and “sensing” means being aware of something. Thus, “remote sensing” can be literally defined as being aware of something from a far distance. The word “optical” means connected with light. A simple definition of optical remote sensing (ORS) would be being aware of something from a far distance with light. This definition can be divided into four parts: “being aware of,” “something,” “a far distance,” and “light.” Here, “being aware of” is to identify, detect, measure, quantify, and so on; “something” includes objects, phenomena, and processes on Earth; “a far distance” ranges from short distances (as long as there is no physical contact) to hundreds or thousands of kilometers; “light” is solar radiation (i.e., sunlight) or human-made radiation. According to energy sources and radiation wavelengths, remote sensing methods can be classified into four remote-sensing fields: optical remote sensing, thermal remote sensing, LiDAR remote sensing, and radar remote sensing (Figure 1).

Sunlight

Sunlight is the energy foundation for ORS (Figure 2). Just as human eyes detect things that reflect sunlight, ORS sensors record light that is reflected from things on Earth.

Properties of sunlight

Sunlight is electromagnetic radiation (EMR), which is characterized by its wavelength and frequency. Each wavelength corresponds to a frequency and vice versa. Radio and TV signals are also EMR. Different channels have distinct frequencies. ORS signals are divided into channels or bands. Instead of frequencies, wavelengths are commonly used to distinguish ORS bands. The full spectrum of sunlight ranges from very short (<0.1 micrometers or μm) to very long (>1 m) wavelengths. Human eyes are sensitive to the band of wavelengths from about 0.4 to about 0.7 μm, which is called visible band in ORS (Figure 3). The visible band consists of three colors: red, green, and blue (RGB). Blue wavelengths are approximately 0.4–0.5 μm, green is 0.5–0.6 μm, and red is 0.6–0.7 μm. The three visible-light bands can be combined into a single band called panchromatic band or separately into blue, green, and red bands, respectively. EMR with wavelengths between 0.7 and 1.3 μm is the near-infrared (NIR) band; EMR with wavelengths between 1.3 and 3.0 μm is the shortwave infrared (SWIR) band (although there are various definitions for NIR and SWIR wavelength ranges). Because the NIR and SWIR bands are strongly reflected from the Earth’s surface, they both are together called the reflected infrared band. Human eyes cannot detect infrared light but ORS sensors can. Both the visible and reflected infrared bands are the most frequently used bands in ORS.

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0467
infrared image can provide information beyond the capability of a visible image (Figures 4 and 5).

Sunlight transmission in the atmosphere

Part of the sunlight incident on Earth is absorbed by the atmosphere and part is reflected. In other words, not all the incident sunlight passes through the atmosphere. The bands that are absorbed by the atmosphere are called absorbed bands because EMR in these bands barely reaches the Earth’s surface. The bands that pass through the atmosphere are called transmitted bands and mostly reach the Earth’s surface. Different atmospheric gases absorb different wavelengths of sunlight but the gas density determines the total amount of sunlight absorbed. This phenomenon provides the foundation for detecting gas (such as water vapor) content in the atmosphere: by measuring the magnitude of certain absorption bands on ground, the gas content of the atmosphere may be determined. However, only the transmitted bands are useful for detecting land resources with ORS. There are two distinct but interrelated remote sensing applications: atmospheric remote sensing and land resource remote sensing.

The reflection of sunlight in the atmosphere is multidirectional and so is called scattering (Figure 2). The proportion of scattered sunlight varies with wavelength – shorter wavelengths scatter more. For example, the proportion of blue-light scattering is about 15 times greater than red-light scattering. This explains why the sky (atmosphere) looks blue on a clear day. This also has implications for ORS data quality: a blue-light sensor is much more likely to capture scattered light than a red-light sensor. Scattered light makes images look foggy. Because the infrared band is at a longer wavelength than the visible band, an infrared sensor receives only a
small amount of scattered light, making infrared images look sharp.

**Sunlight reflected from the Earth’s surface**

Sunlight that reaches the Earth’s surface is reflected or absorbed by the surface (Figure 2). The reflected sunlight makes it possible to distinguish things or objects on Earth by eye or with ORS sensors. The amount of reflected sunlight varies for different objects and results in different shades of brightness. For example, fresh snow is bright white because it reflects a lot of sunlight; asphalt is dark because it reflects little...
sunlight. Within the visible band of sunlight, healthy vegetation in the growing season reflects more green light than blue or red light, making healthy vegetation look green. This suggests that the reflection of sunlight is a function of the sunlight that reaches the Earth’s surface, the type of object from which it is reflected, and the wavelength.

ORS data resolution

Radiometric resolution

ORS sensors convert object brightness into digital numbers to represent object brightness values, which range from 0 for extremely dark objects to a large number (e.g., 255) for extremely bright objects. If the range of brightness values is 256 (0–255), the data quantification level or radiometric resolution is eight bits (because $2^8 = 256$); if an ORS sensor provides 12-bit data, the range of the data recorded will be $2^{12} = 4096$ (0–4095). The earliest generation of digital-satellite ORS data was six bits (e.g., Landsat 1-3) whereas the latest satellite ORS data can be as high as 16 bits (e.g., Landsat 8). The higher the radiometric resolution, the greater the reparability power the data have. For example, two objects with brightness values of 25200 and 26000 when a 16-bit ORS sensor is used would have the same value of 25 when a six-bit ORS sensor is used.

Spatial resolution

The digital data collected by ORS sensors consist of frames or scenes. Each frame of data covers a rectangular area and may contain multiple bands.
OPTICAL REMOTE SENSING

Figure 4  Visual comparison between blue-band picture (left) and near-infrared-band picture (right). Broad-leaved trees and coniferous trees are hardly distinguishable on the visible-band picture but broad-leaved trees look much brighter than coniferous trees in the NIR. The two images received the same automatic correction with Microsoft Office Picture Manager. The area of each picture is about 4 ha. Data source: USDA National Agriculture Imagery Program. Image acquired from Tippecanoe County, Indiana, USA, on June 8, 2012. Spatial resolution is 1 m.

Each band of data within a frame is divided into rows and columns. Corresponding to any row and column, there is a picture element called a pixel, which is a square area with a single brightness value. The dimension of the pixels is the same within a frame of data and is called spatial resolution. The smaller the pixels, the higher the spatial resolution, and the more rows and columns for a given area of land covered by a frame. The position of each pixel is labeled with a row number and a column number. The spatial resolution of ORS data ranges from a few centimeters to hundreds of meters, and each frame can cover an area ranging from dozens of hectares to hundreds of thousands of square kilometers. Depending on the purpose of ORS applications, certain spatial resolutions are more preferable than others. An increase in spatial resolution is normally helpful for visual interpretation (Figure 6) but quantitative analysis of high-spatial-resolution data usually requires complicated methods that require high computational power.

Spectral resolution
ORS data vary with the number of bands, which is referred to as the spectral resolution. The typically available ORS data consist of three categories: panchromatic data (1 band), multispectral data (2–tens of bands), and hyperspectral data (hundreds of bands). Panchromatic data combine blue, green, and red wavelengths into one band, and commonly come with multispectral data. Multispectral data normally have at least three bands, including three visible wavelengths or green, red, and one near-infrared wavelength. Four-band ORS data normally contain three visible bands and one near-infrared band. Some satellite ORS data have more than four bands and their bands are continuous in wavelength. Hyperspectral data contain continuous bands with wavelengths ranging from the visible to up to 2.5 μm or greater. In general, the higher the spectral resolution, the more information an ORS dataset contains. However, in ORS data analysis, the number of parameters increases exponentially with the number of bands. This means that larger reference or training datasets are necessary for analyzing ORS data with a greater number of bands. A subset of bands sometimes is optimal for information extraction.

Temporal resolution
The unavailability of ORS data is often a limiting factor for ORS applications. Every corner of the Earth’s surface has been covered by ORS data but some regions have more repeated coverage than others. The frequency or revisit time of the repeatable coverage within a certain time frame is called the temporal resolution, which
Both natural grass (left) and synthetic grass (right) are used for football fields but their images have totally different colors when the near-infrared band is used instead of the red band.

Comparison between a 30-m-resolution Landsat image (left) and a 2–4-m-resolution image (right) of an illegally logged forest landscape in North Korea (Tang et al. 2010).
characterizes the potential availability of ORS data. For example, Landsat TM data repeat the same area every 16 days. Some other satellite ORS sensors have flexible viewing angles and can increase repeatability over the same area with oblique views. Multiple satellites of the same type can be used to increase temporal resolution. The USDA national agriculture imaging program (NAIP) has provided digital orthophotos for the entire United States once a year since 2003. The MODIS sensors have a temporal resolution of 1–2 days and are useful for tracking wildfires. In any case, repeated images over the same area imply permanent time-series records and provide the data foundation for detecting changes that happened in the past. ORS data from multiple seasons can distinguish phenological differences among different plants or vegetation types and are more powerful for vegetation classification than single-season ORS data.

ORS types based on distance

Airborne ORS

The first aerial photographs were taken from balloons in the 1860s. Aerial photographs were later taken from airplanes in World Wars I and II. Black and white, color, visible, and infrared films were used until the 1990s, when digital sensors gradually became more economical and effective. In many parts of Europe and North America, aerial photographs have been available since the late 1930s or early 1940s. Vertical aerial photographs helped the development of various map products, such as topographic maps. The current aerial photographs are almost exclusively in digital format. A typical example is the USDA NAIP. The relatively low altitudes (ranging from tens to thousands of meters) of airplanes cause locational errors for ground objects with different elevations; such a phenomenon is called displacement. Digital orthophotos are a type of data that are a mosaic of aerial photographs following displacement corrections so that they can be used as maps.

Airborne ORS has been rapidly advancing toward higher spatial and spectral resolutions. It is not surprising to see digital aerial photographs with a spatial resolution of a few centimeters. Hyperspectral digital aerial photographs are also available. Airborne ORS becomes even more flexible, safer, and less expensive when unmanned aerial vehicles (UAVs) are used to carry ORS sensors or cameras (Tang and Shao 2015). The USDA Forest Service has made systematic efforts with UAV ORS for wildfire control, forest health monitoring, and forest management (Ambrosia et al. 2011).

Spaceborne ORS

Spaceborne ORS or satellite ORS is commonly done with Earth-orbiting satellites, which have no limitations regarding national borders and makes ORS globally accessible. The orbits of remote-sensing satellites are between 600 and 900 km above the surface of the Earth, which is tens or hundreds of times higher than airplanes. Thus, vertical satellite images are less distorted than aerial photographs and have spatial properties similar to topographic maps. The first generation of ORS used films and was used for military purposes. Since the 1970s, civilian ORS has been developed for imaging Earth’s resources and meteorological purposes by using scanners. The earliest civilian spaceborne remote sensing was the United States’ NASA Landsat 1, which was launched in 1972. Landsat remote sensing has become the longest ORS over the world. Landsat 8 was launched in 2013 and its imagery products have the highest spatial resolution of all the free satellite data from the U.S. government (Table 1).

All the ORS products from the US government are free of charge. While the US
Table 1  Band wavelengths and spatial resolutions for Landsat 8.

<table>
<thead>
<tr>
<th>Bands</th>
<th>Wavelength (micrometers)</th>
<th>Spatial resolution (meters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band 1 – Coastal aerosol</td>
<td>0.43–0.45</td>
<td>30</td>
</tr>
<tr>
<td>Band 2 – Blue</td>
<td>0.45–0.51</td>
<td>30</td>
</tr>
<tr>
<td>Band 3 – Green</td>
<td>0.53–0.59</td>
<td>30</td>
</tr>
<tr>
<td>Band 4 – Red</td>
<td>0.64–0.67</td>
<td>30</td>
</tr>
<tr>
<td>Band 5 – Near infrared</td>
<td>0.85–0.88</td>
<td>30</td>
</tr>
<tr>
<td>Band 6 – SWIR 1</td>
<td>1.57–1.65</td>
<td>30</td>
</tr>
<tr>
<td>Band 7 – SWIR 2</td>
<td>2.11–2.29</td>
<td>30</td>
</tr>
<tr>
<td>Band 8 – Panchromatic</td>
<td>0.50–0.68</td>
<td>15</td>
</tr>
<tr>
<td>Band 9 – Cirrus</td>
<td>1.36–1.38</td>
<td>30</td>
</tr>
<tr>
<td>Band 10 – Thermal Infrared 1</td>
<td>10.60–11.19</td>
<td>100</td>
</tr>
<tr>
<td>Band 11 – Thermal Infrared 2</td>
<td>11.50–12.51</td>
<td>100</td>
</tr>
</tbody>
</table>


Table 2  Overview of major working spaceborne remote-sensing systems.

<table>
<thead>
<tr>
<th>Data name</th>
<th>Provider</th>
<th>Spatial resolution</th>
<th>Revisit time</th>
<th>Number of bands</th>
<th>Launch year</th>
<th>Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASTER</td>
<td>NASA/Japan</td>
<td>15–90 m</td>
<td>16 days</td>
<td>15</td>
<td>1999</td>
<td>Cost</td>
</tr>
<tr>
<td>AVHRR/3</td>
<td>NOAA/USGS</td>
<td>1090 m</td>
<td>1/2 days</td>
<td>6</td>
<td>1998</td>
<td>Free</td>
</tr>
<tr>
<td>GeoEye-1</td>
<td>DigitalGlobe</td>
<td>0.41–1.65 m</td>
<td>&lt;3 days</td>
<td>5</td>
<td>2008</td>
<td>Cost</td>
</tr>
<tr>
<td>GOES</td>
<td>NOAA</td>
<td>1–8 km</td>
<td>Minutes</td>
<td>6</td>
<td>1978</td>
<td>Free</td>
</tr>
<tr>
<td>Ikonos</td>
<td>DigitalGlobe</td>
<td>0.8–3.2 m</td>
<td>3 days</td>
<td>5</td>
<td>1999</td>
<td>Cost</td>
</tr>
<tr>
<td>Landsat 8</td>
<td>NASA/USGS</td>
<td>15–100 m</td>
<td>16 days</td>
<td>11</td>
<td>2013</td>
<td>Free</td>
</tr>
<tr>
<td>MODIS</td>
<td>NASA</td>
<td>250–1000 m</td>
<td>1 day</td>
<td>36</td>
<td>1999</td>
<td>Free</td>
</tr>
<tr>
<td>QuickBird</td>
<td>DigitalGlobe</td>
<td>0.65–2.44 m</td>
<td>2–6 days</td>
<td>5</td>
<td>2001</td>
<td>Cost</td>
</tr>
<tr>
<td>RapidEye1-5</td>
<td>BlackBridge</td>
<td>5 m</td>
<td>1–5.5 days</td>
<td>5</td>
<td>2008</td>
<td>Cost</td>
</tr>
<tr>
<td>SPOT</td>
<td>EADS</td>
<td>2.5–20 m</td>
<td>1–3 days</td>
<td>5</td>
<td>1986</td>
<td>Cost</td>
</tr>
<tr>
<td>WorldView-2</td>
<td>DigitalGlobe</td>
<td>0.46–1.84 m</td>
<td>1.1 days</td>
<td>9</td>
<td>2009</td>
<td>Cost</td>
</tr>
</tbody>
</table>

Sources: data providers’ websites.

government has continued developing medium- and coarse-spatial-resolution data for weather forecast and global change activities for more than three decades, private firms have made great contributions to the development of high-spatial-resolution ORS systems (Table 2). Each scene of high-spatial-resolution data is much smaller than that of the coarse-spatial-resolution data, but the high-spatial-resolution sensors have flexible lens pointing angles and, thus, can shorten their revisit time. The five satellites from RapidEye contain identical multispectral, equally
calibrated sensors, which makes them capable of covering relatively larger land areas with vertical images. The geostationary operational environmental satellites (GOES) stay continuously over one position on the Earth’s surface and constantly observe the weather.

Ground-based ORS

For the purposes of sensor calibration and in situ applications, ORS can be operated by hands, on vehicles, or on towers as platforms on the ground. Ground-based ORS is not mainstream ORS but is a flexible approach and can be used without temporal restrictions. The imagery of ground-based ORS can be obtained at spatial resolutions of as high as centimeters or millimeters, and its applications are commonly found in intensive agricultural experiments.

ORS data analysis

ORS data consist of numbers, and often a large amount of numbers. It is up to analysts how to present and interpret the numbers. The methods of ORS data interpretation and presentation vary with the purposes of ORS applications. It is essential to display the numbers as images that can be visually interpreted by humans (Figures 4 and 5). Various computer techniques or algorithms have been developed to assist analysts to carry out automated, quantitative analyses of ORS data. Useful information can be extracted through the quantitative analysis and often is presented in map formats, such as land-use maps and weather-forecast maps.

Data display

As computer technology has advanced, the capabilities of remote sensing have continuously been improved. ORS images were displayed only with single-color dots or symbols in the 1970s–1980s but are now available in “visual reality” formats. The current image-display techniques were unimaginable 3–4 decades ago when the first generation of civilian satellite imagery became available. If one band of ORS data is displayed, black-and-white images are formed; if two or three ORS bands are used, color images may be formed. ORS bands can be combined freely to obtain different looks for the images. For example, ORS data with six bands can have 216 ($= 6^3$) combinations of band displays. When blue, green, and red bands are displayed, it is possible to obtain images with natural colors that are familiar to human eyes; all other combinations, 215 (out of 216), result in fake or pseudo colors. In general, the more colors that are shown, the greater is the amount of information in an image. Researchers have developed various techniques to maximize information from a single display of multispectral and hyperspectral ORS data, including band selection and transformation (Landgrebe 2003; Jensen 2005). A typical band-transformation technique is principal component analysis (PCA), which compresses significant information on multispectral ORS data into a smaller number of new bands (Shao and Duncan 2007). One of the most astonishing band transformations is Tasseled Cap (TC) (Jensen 2005), which generates individual bands that have explicit land-surface implications, such as greenness, brightness, and wetness. Image enhancement is also a common practice with ORS data but is used primarily for visual interpretation rather than information extraction.

Information extraction

Information extraction from ORS data is the basis for various ORS applications. No two natural features are identical; neither are two
pixels if radiometric resolution is high enough. Furthermore, the reflected sunlight captured by an ORS sensor is controlled by many factors that vary with time and position (Figure 2). Therefore, there are large variations in pixel brightness and no two images are identical, even for the same type of ground features. That is why many methods or algorithms have been proposed to extract information from ORS data (Landgrebe 2003; Lu and Weng 2007). There are supervised and unsupervised algorithms of information extraction (Jensen 2005), although both actually require input from analysts. Blind ORS-data analysis without ground experience or knowledge is an unreliable and even irresponsible exercise. Following information extraction, the extracted information must be validated with independent sample data that should be 100% accurate (or close to 100% accurate) (Congalton and Green 1999; Foody 2002). ORS data with relatively low accuracy can have unexpected negative impacts on its downstream applications (Shao and Wu 2008). Regardless of pixel size, there are always pixels that cross boundaries of two or more ground features. Mixed pixels are another characteristic of ORS data and represent one of the error sources from information extraction. Fuzzy classification methods have been developed to extract proportional information from mixed pixels (Jensen 2005).

Remote-sensing software

Since the 1960s, many efforts have been made to develop computer programs for ORS data analysis; a computer program called LARSYS was one of the first systems. A number of systems in government laboratories, university research laboratories, and several commercially offered products were the descendants of LARSYS. MultiSpec is its direct descendant and contains advanced research results; it is available to download via the Internet (https://engineering.purdue.edu/~biehl/MultiSpec/).

There are a number of commercially available remote-sensing software systems. They include Erdas Imagine, ENVI, ER Mapper, INDRISI, PCI, TeraVue, and TNTmips. In general, commercial remote-sensing systems have a broad range of geospatial functions for users in different fields and at varied technical levels whereas free-ware, such as MultiSpec, may be specialized for research- and/or education-oriented capabilities.

ORS applications

ORS can be directly used for detecting things on Earth: their position, if they have changed, and so on. In fact, its indirect applications are even broader and deeper, and are playing an active role in sustainable development all over the world.

Day-to-day applications

Topographic maps on various scales were all derived from aerial photographs, which have become broadly available since the 1940s. Through digitalization, paper maps were converted into digital formats under the framework of geographic information systems (GIS). GIS maps are an important part of the navigation function of the global positioning system (GPS), and GPS navigation has now become an important part of daily life.

Google Earth (http://www.google.com/earth/) has a huge collection of ORS images that cover every corner of the Earth’s surface, and it integrates ORS images from other data sources. Thanks to state-of-the-art image computation techniques, ORS images from Google Earth can be freely expanded or contracted on computers that are connected to the Internet. One can also view ORS images in three
dimensions on Google Earth. Historic images on Google Earth are useful for detecting changes on the ground, be they due to natural causes or human activities. Because of their global coverage, ORS images on Google Earth are useful for assessing land use in places where ground access is impossible (Tang et al. 2010).

Specialized applications

Remote sensing is an effective and efficient technique for many land-use- and land-cover-related investigations at scales from landscape to global. The greater the spatial scales, the more advantages remote sensing can have. Any global or continental map products, such as the global tree-cover map (Defries et al. 2000), which is derived from remote sensing data, could not be obtained or would be too costly with ground survey methods. As a conventional remote sensing technique, ORS has been broadly used since the end of World War II. Airborne and spaceborne ORS has made unique contributions to understanding the dynamics of the atmosphere, oceans, and the vegetation cover of the land and has provided new perspectives in the study of the solid Earth (Goetz et al. 1985). Associated with continuous advances in remote sensing, ORS has been applied in many fields, including archaeology, aquatic ecosystems, biodiversity, climatology, forest measurements, global change, habitat detection and conservation, human health, invasive species, land use and land cover, urban morphology, vegetation stress, and wildfire management. In fact, there are too many applications to be listed here.

Considerations

Remote-sensing technology continues to advance on the producers’ side while its applications continue to grow on the users’ side. It is important to improve communications between the producers and users. The producers need to make data products as reliable as possible while the users need to understand the correct ways to use those products. Such considerations can be better explained with two examples: (i) two sets of global land-cover data were created with the same AVHRR imagery by two groups of reputable remote-sensing analysts, but their overall agreement for aggregated vegetation types was only 80.32% (Hansen and Reed 2000); (ii) two sets of national land cover data for the United States were produced with Landsat imagery acquired in 1992 and 2001 by the same group of professional analysts, but they were not directly comparable (Homer et al. 2007). In short, remote sensing is a useful and powerful tool only if used correctly.

SEE ALSO: Resampling, raster; Scale; Spatial resolution

References


Foody, Giles M. 2002. “Status of Land Cover Classification Accuracy Assessment.” Remote Sensing of
OPTICAL REMOTE SENSING


Further reading

In the most basic terms, history is a process by which people convey and try to make sense of incidents. This sense-making has characterized diverse cultures’ spoken traditions and narrative arts – speech, aphorism, recitation, song, ballad, folklore, symbols, architecture, and handicrafts among them. Through such means, traditional knowledge and culture have been passed on over generations, and have communicated the exceptional and the quotidian in our lives.

Whether spoken or written, history takes the form of chronicles, narratives, and modes of inquiry to judge how what has passed may affect what is present or expected; that is, history enables one to envision. As a formal discipline history is, among other things, a means to understand (i) the actions of individuals and groups; (ii) the influence of so-called natural and societal processes, structures, systems, and events; and (iii) the implications, meanings, and significances of them. Like other disciplines – geography not least among them – history is also concerned with its own internal workings: the values that inform it and the frameworks by which historians know and act in and on the world at different scales from households to international domains.

Much formal history has been “from above” rather than “from below,” descriptors coined by Lucien Febvre in the 1930s and popularized by E.P. Thompson in the 1960s. The initial focus of history from below was “the people” (and especially the working class) and, in the period since the 1970s, it has been further concerned with studies of subaltern groups, gender, sexuality, race, ethnicity, and so on. Eric Hobsbawm (2011/1997, 205) has noted that “grassroots” history’s source materials are elusive and require prospecting: there “is generally not material until our questions have revealed it.” Oral history is one of the mechanisms to elucidate historical knowledge from below and, while acknowledging that it is useful, Hobsbawm nevertheless questions its links to personal memory. This he views as a medium too unreliable for preserving facts, and one highly dependent on the corroboration of other verifiable and independent sources of information, or on experimentation – including collaboration with psychologists able to shed light on memory, association, and behavior.

Hobsbawm (and others) have expressed concern at the subjectivity of oral history. Assuredly, distinctions have been – and likely need to be – made between oral history as a form of research practice and as a mechanism for community publishing projects (Bornat 1989). Nevertheless, criticisms that oral history is overly subjective are contestable, and practitioners argue that this very characteristic may be among the method’s most important features – that is, its capacity to use conversation to elicit opinions, sensations, perspectives, and memories rather than (or as well as) “facts.”

To gain access to people’s memories of and views about certain experiences – including those related to singular and repetitive events, habits, practices, customs, and so on – the chief tool used in oral history is the recorded conversation. Initially written down, later audio-taped, and now often captured using
multimedia technologies, oral history conversations involve one or more extended engagements with participants to elicit recollections of the past (although contemporary settings and events are not discounted). At the same time, participants may be asked to share other forms of historical ephemera or information – letters, images, reports, artifacts, and so on. This style of conversation is often qualitatively different from interviews used in social science research. Interviews are more likely to be one-off encounters with participants (excluding interactions to set up discussions and later provide for member checking where participants affirm the accuracy of transcriptions). Typically, interviews will also be of limited duration and may, at least in part, be oriented to current and forthcoming or anticipated issues and events.

Oral history, then, is a method of approach to understanding in great depth the views and experiences held by those who possess direct information, knowledge, and insights about subjects that are also of interest to practitioners. Its enactment as a research method has been attributed to Allan Nevins and others at Columbia University’s Oral History Research Office, the purpose of which from 1948 was to record, analyze, and preserve the spoken recollections and reflections of national leaders. Nevins took the interview – a key tool in journalism, in which he first trained – and applied it in new ways to studies of the elite, framing his own analytical narratives around those stories in order to examine larger questions pertaining to political systems, leadership, and social processes.

Through the 1960s, and intensifying from the 1970s, the use of oral history started to influence reappraisals of the merits of learning about and from everyday life. Its application coincided with shifts in the philosophy of history and historiography – including debates on the relative merits of spoken narratives and written archives and – with broader transformations in epistemology, ontology, and axiology – studies of knowledge, being, and value. Part of the reason for this influence and attendant shifts is because practitioners of oral history articulated a defense of everyday narratives as valid evidence in knowledge production. These narratives were also used to prize open and critique the narrowness of history from above, which was understood to be dominated by public, political, white, wealthy, and masculine accounts. In this light, oral history has been construed as a form of consciousness raising that leads to action and transformational change (Thompson 1998).

Equally, oral history’s democratizing impulse and focus on direct access to living individuals have both challenged and augmented the use of statistical aggregation of the lives of ordinary people, humanizing history and other disciplines (Bornat 1989). According to Paul Thompson (1998, 28), a chief merit of oral history is that it exposes multiple realities and standpoints, facilitating nuanced historical understandings and “a much fairer trial,” which has radical potential (and some risks) for subaltern populations. Hence the need to pay attention to the ethical demands of oral history, among them (i) safeguarding participants’ rights to confidentiality where requested by them; (ii) affording them full respect and ensuring they are not reduced to “just” another source or use value; (iii) understanding the effects on participants of recollection and the ways in which remembering may harm as well as support them; and (iv) ensuring that, where analysis and interpretation of narratives are undertaken by researchers, participants are provided with opportunities to contribute to or comment on those processes.

In addition, particular care needs to be taken when working with indigenous and other cultures possessing traditional knowledge – these engagements must protect customary, spiritual,
and intellectual property rights (WIPO 2015). The same applies when conversing with children and young people, with the disabled or frail, or with those in compromised situations such as prisons or other institutional settings. In this regard, both international and national organizations exist to help oral historians and those using oral history methods demonstrate integrity and foster sound practice. Among these organizations are the International Oral History Association (www.iohanet.org), and like bodies in the United Kingdom (www.ohs.org.uk), Canada (www.canoha.ca), the United States (www.oralhistory.org), New Zealand (www.oralhistory.org.nz), and Australia (http://oralhistoryaustraliasant.org.au). Each also offers detailed instructional guidance about the practicalities of oral history.

Several other broad and institutional effects of oral history’s democratizing influence are worth noting. One of these effects is the method’s overlap with others in social research, for example the use of participatory action research by community groups seeking to engage researchers. In other words, it has been not uncommon for the impetus for oral history work to come from participants rather than practitioners. Another effect of oral history’s validation of the quotidian is that practitioners, participants, and audiences tend to reflect on the relationships between direct experience and its larger contexts – social, economic, political, cultural, and spatial. Such reflections have led to a recognition that oral history involves shared authorship, requires conscious decisions about the locus of interpretation of narratives, and warrants attention to the effects of the redistribution of intellectual authority.

In some oral history there may be cause to rethink the idea of authorship entirely, and to reframe the coproduction of knowledge as a form of curating. “Curating,” from the Latin “to take care,” involves keeping and interpreting heritage materials, and both participants and practitioners of oral history may be thought of as engaging in such activities; so, too, can audiences be invited to care for the everyday histories relayed to – and even coproduced by – them. Either way, practitioners must “respect, understand, invoke, and involve the very real authority [participants and] audiences bring” (Frisch 1990, xxii).

This idea of authority is important for at least two other reasons related to conceptions of “truth.” First, oral histories and the narratives they generate are sources of information and insights that inform and enable generalizations through analysis, comparison, conceptualization, modeling, and prediction. These generalizations also function as contingent truths and hold stable (if not durable) certain knowledge claims. Second, when oral histories and narratives are presented without apparent interpretation, they are seen to bypass the dangers implied by acts of analysis or construal – “apparent” insofar as interpretation might have preceded the production of written transcriptions or recordings and may be embedded in the selection of participants, questions, and conversational situations.

It is in this context that consideration is taken of the relationship of experience to memory and of memory to history and interpretation. Significant discussion arises in response to questions about the ways in which experiences are “translated” to memories later recalled and formalized as part of the historical record – hence Hobson’s call for the validation of oral recollections using other sorts of information. In this light, Frisch (1990, 10–11, 16) commends practitioners to remain alert to certain questions:

What happens to experience on the way to becoming memory? What happens to experience on the way to becoming history [or, one might ask, part of the geographical imaginary]? … what is the relationship of memory to historical generalization? … What sort of person is
ORAL HISTORY AND NARRATIVE

speaking? What sort of thing is he or she talking about? What sorts of statements about it are being made? ... At what distance, in what ways, for what reasons, and in what patterns do people generalize, explain, and interpret experience? ... what role [does] knowledge play in our lives?

Knowledge is, of course, created and, in relation to oral history, there are at least three expressions of this process. The first sort of knowledge produced is that derived from unadulterated transcriptions, where participants have opted not to check their narratives or have not been afforded such opportunity. There have been suggestions that these transcriptions are pure and direct, but they may contain errors of omission and commission that could distort understanding. The second sort of knowledge produced is that from transcriptions that are not edited by practitioners but are checked by participants in order to affirm, amend, augment, or remove what has been produced and thus to protect both themselves and third parties – a responsibility jointly held by practitioners and, in some cases, subject to the laws pertaining to slander. This second expression of knowledge production also allows participants to ensure that the knowledge engendered is coherent – that is, lucid and consistent. The third sort of knowledge produced is that derived from the work undertaken by practitioners who engage in substantial editing of transcriptions (which may or may not have been checked by participants for varied reasons such as lack of interest, absence, or death). To be masterful, such editing requires an intimate understanding of a raw transcription in its totality. It may be used to ensure flow; emphasize key insights; and level expression by removing speech disfluencies such as “um” or “er” and by avoiding the written representation of speech affects, sounds, or accents which risks caricature. In this sense, it should be immaterial to the quality of the outcome that oral histories derive from ordinary people’s recollections of everyday life. Rather, just as history from above is guided by various strategies to ensure its quality, mechanisms exist for ensuring the quality, depth, and salience of oral history transcriptions and their interpretations. These mechanisms – checking and editing not least among them – are deemed important if oral history is to remain estimable. Their use also justifies the considerable resources required to secure – in agreed and appropriate repositories, exhibits, and so on – the deposition of original written, sound, or multimedia files and additional interpretive artifacts (such as documentaries, monographs, or research papers).

Oral history has been taken up in human geography on the understanding that, as well as being a technique by which to understand the past and its importance to present and future knowledges, it is a powerful tool by which to understand the shifting terrain of diverse geographical imaginaries and materialities. According to George and Stratford (2010), geography’s chief concern is to understand people’s sociospatial relations in place, as well as the dynamics and scales of change that occur in locales and are mobilized across territories, regions, landscapes, and environments. Human geographers are particularly concerned to contribute to research-based outcomes that advance wellbeing by paying attention to the varied dynamics of sociospatial relations, spatial justice, embodiment, sense of place, and mobility. Their interest in oral history’s emancipatory potential is partly explained by this impulse. For some geographers, such orientations manifest both as philosophical and as political investments in understanding – rather than appropriating – important but hitherto subaltern perspectives. For others it gains expression as ongoing labors intended to understand the different loci and dynamics of power as it appears in hegemonic spaces. Either way, close attention is applied to considering, inter
alia, geographical imaginaries and memories (Lowenthal 1985): sense of place, experience of displacement, cognizance of connection or distance, capacity to move or to linger, or the ability to generate mental maps that anchor memories through the life course (Ostroff 1995).

Thus, in geography examples of the use of oral history and attention to narrative now abound, and here one example illustrates how this work performs multiple important labors. It relates to Hayden Lorimer’s (2003) consideration of the briefly merged lives of two individuals – a 14-year-old Glaswegian schoolgirl, Margaret Jack, and a 34-year-old geography instructor, Robin Murray – and their shared time and different reminiscences of a four-week residential field course at Glenmore Lodge in the National Forest Park below the Cairngorm Mountains in January 1951. Noting the intrinsic worth of their experiences at the lodge, Lorimer gathers written, visual, and material artifacts from Jack and from Murray’s widow, Catriona (who also knew the lodge), and converses with both – some 52 years after the original field course. He weaves through these reminiscences larger narratives about educational reform and the changing pedagogies of geography at mid-century; the democratization of land management in Scotland; the importance of “the region”; and ways to value “the minor figure” and acknowledge grassroots practices and geographical knowledge – including haptic, sonic, and kinaesthetic forms of experience. These juxtapositions Lorimer interprets as demonstrating the overlapping “fullness and diversity of still-lived lives … fragile connections and situated encounters” (2003, 204), rewarding and unsettling salvage operations that produce fluid knowledges. Just one among many, this case exemplifies the power of oral history in geography, and reveals the ways in which “small stories can tell of epistemic shifts on personal and intimate terms” (Lorimer 2003, 214).

SEE ALSO: Community; Environment and everyday life; Fieldwork in human geography; Interviews; Qualitative data; Subaltern

References


Further reading


Orientalism and Occidentalism are interrelated concepts. “The Orient” and “the Occident” have been historically produced as stable collections of texts, talks, and ideas that are juxtaposed, both representing fields of knowledge and sets of practices supported by often stereotyped ideas of other cultures and civilizations, and of their imaginative geographies. They are also a product of, and a response to, Eurocentric modernity and to its imperial and colonial projections.

Palestinian American cultural theorist Edward Said (1978), who famously proposed the first critical review of the term, defined “Orientalism” in three distinct ways. First, it is a body of scholarly work focused on “the Orient,” and largely implicated in the realization of the cultural and political conditions that have produced the broader colonial project, and also in the form of Oriental studies courses and programs in Western universities and intellectual circles. Second, it is a cultural and aesthetic interest in “the Orient,” manifested in the fine arts, and in painting in particular, starting from the eighteenth century onward, inspired by so-called “Oriental” themes. Third, it is a critical approach to the very idea of “the Orient” endorsed and implemented by European and American cultural authorities. Orientalist thought is here presented as a regime of truth about a specific region and culture imagined as a sort of discrete entity, as a theatrical setting available to Western scholarly inspection. This approach identifies “the Orient” – an object of study for the other two – as a European and American “invention,” which emerged as part of a constellation of power, knowledge, and geography. Drawing on the conceptual vocabulary of Michel Foucault, Said (1978, 6) maintained that Orientalism was not just a European fantasy but also a discourse made of a constellation of “theory and practice in which, for many generations, there has been considerable material investments.” The result of this long-standing and complex process was the production and circulation of imaginative geographies of “the Orient,” widely supported by popular and scientific understandings of “Oriental” culture and civilization.

It is hard to overestimate the impact of Said’s groundbreaking writings on the development of postcolonial studies in the humanities and the social sciences, including human geography. Said’s inquiry has inspired the proliferation of work in the past few decades on the othering of peoples and regions by imperial powers. Orientalist thought was (and remains) the result of ontological and epistemological categories endorsing a racial and unbreakable distinction between “the Orient” and “the rest,” and especially “the West,” a true practice of othering, based on writings, artworks, trade, military action, political statements, and even tourism. In this framework, the “Orientals” and their culture were imagined as objects that may be studied in isolation and whose presumed characteristics can be clearly and unproblematically defined. It is easy to imagine how such a standpoint could be the origin of a set of stereotyped understandings of the people labeled as such, and of their actions,
which often embraced an overall, albeit implicit, racist agenda. Indeed, what can be learned from Said’s 1978 key book, and from the decades of work that followed in the same vein, is that the imaginative geographies fueling the Orientalist discourse had – and in many cases continue to have – enormous practical implications, since they largely represent background knowledge that informs power relations with the presumed “Oriental world.”

Representations of “Oriental” individuals, typically marking them unreliable, irrational, barbaric, mysterious, and driven by a political economy of desire often matched by a latent sense of cultural inferiority, have traveled through the centuries and pervaded endless forms of encounter between “the West” and “the Orient,” largely based on military confrontation and moral judgment. These imaginative geographies of otherness have dramatized the distance and the difference between “the West” and some “Oriental lands” commonly and unproblematically perceived as culturally separate regions and geopolitical settings, as demonstrated on many occasions in past decades by the global media coverage of the wars in the Gulf region and the Middle East.

Departing from Said’s argument that deploying imaginative geographies is a predominantly European and American project, many scholars have responded with conceptions and analyses of self- and internal Orientalism. Instead of passively receiving the outcomes of the Oriental project and Orientalizing processes, “the Orient” has at times initiated and participated in constructing and circulating such imaginative geographies (Yan and Santos 2009). “Internal Orientalism” (Schein 1997) or “nested Orientalism” (Bakic-Hayden 1995) are identified as the attempt by dominant ethnic groups to exoticize (and consequently marginalize) other ethnic minorities within a specific political entity. Along these lines, other work has investigated the production of equally imaginative geographies of “the West” within non-Western cultural contexts, including the Arab and Muslim world. Such discrete visions of “the West” have been sometimes defined as Occidentalism: that is, a calculated and intended construction of “the West” or “the Occident” as a unified entity, described in scholarly debates as an inversion of the Western imaginary (Carrier 1995) or a counterdiscourse to Orientalism. In the first case, Occidentalism exists as a process in which non-Western cultures construct unifying and stereotypical framings and images of “the Occident,” in the the same way as European and American cultures produce and circulate Orientalist imaginative geographies. In the second, Occidentalism is a discursive and material strategy mobilized and deployed by non-Western cultures to offset the negative impacts of Orientalism and of “Western” cultural and political supremacy.

Many critics resist the idea that Orientalism or Occidentalism may be treated as similar processes, because of the significant historical and political differences that characterize their respective impact and reception. According to Said, the success of Orientalism as a political and cultural project was built on the historical depth, geographical scope, and durability of Western imperialism and colonialism. Unless similar projects of domination and dispossession are imposed on Western places by non-Western cultures, “no one is likely to imagine a field symmetrical to Orientalism called Occidentalism” (Said 1978, 50). Acknowledging that the use of Occidentalism as an inverse to Orientalism potentially distracts from the complicity of power-knowledge embedded in the expansive and extensive histories of imperialism that constituted Orientalism, Coronil (1996, 56) argues for Occidentalism to be framed as a condition of
possibility – as “a dark side of a mirror”; however, recognizing the existence of Occidentalist tropes at the same time allows us to engage in critical ways with the meaning of modernity for non-Western cultures and regions, especially those that have been subjected to colonial rule and the effects of imperialism, and to valorize alternative historical interpretations of the presumed East–West global divide.

Said’s approach to Orientalism was also widely resisted in academia. For example, conservative and influential Orientalist historian Bernard Lewis critiqued Said for his presumed historical inaccuracy and selective presentation of the field of Oriental studies. Another critique came from more sympathetic readers in relation to Said’s tendency to somehow essentialize “the Orientals” and their geographies – although in different ways compared to the Orientalist tradition – as if there were some deeper truth to be unveiled about those regions and cultures. However, goes the critique, the “objects” of past Orientalist investigations should not be seen as simply passive subjects unable to engage with the colonized gaze produced by Eurocentric universalist discourses. In fact, colonized subjectivities have engaged in substantial ways with the powerful discourses of Orientalism. Not acknowledging this engagement may indeed constitute new forms of objectification that deny the actual agency of these peoples in dealing with the Orientalist categories that have for so long qualified them with reference to the colonial and postcolonial contexts.

One final important aspect of Orientalism is the fact that its legacies still have multiple and sometimes subtle effects on the categories and the practices that, on an everyday basis, produce the “contact zones” and the related encounters between presumed “Western” and “Oriental” subjects. These effects have emerged on numerous recent occasions: for example, in the media representation of Saddam Hussein as a typical despotic “Oriental” leader, or in the horrific practices of the Abu Ghraib detention center. But they have also emerged as commonplace in discourses about Western culture and imperialism that feed into radical Islamic representations of Europe, or in local residents’ attitudes toward Western tourists traveling outside of Europe. Orientalism and Occidentalism, in the everyday practices. In their most popular accents they still play an important role in how the broader public, for example in Europe, understand the arrival of asylum-seekers and migrants from across the Mediterranean, and also in how politicians present their often populist views on the integration of people coming from different religious and cultural backgrounds.

The way forward for geography and geographers may indeed lie in persevering with the everyday of Orientalism. This approach can include an unrelenting engagement with how cultural, media, and travel imaginaries shape not only our travels abroad but also our everyday practices. The pervasive operation of Orientalism is in part founded on what has been argued as a (still) “colonial present” – a world that has persisted by dominating and dispossessing an “Oriental Other” based on politically constructed notions of “irrationality” and “unreason” in Afghanistan, Palestine, and Iraq (Gregory 2004). Such analysis can also entail an engagement of the “Oriental”/“Occidental” body as a site for the (un)intended deployment of imaginative geographies concerning “the Orient” and “the Occident” and the implications of such imperialist forms of othering. Such geographical inquiries can be located in the acquisition of the correct body for a beach vacation, in national population policies, racial profiling, biometrics and/in border security, and the “war on terror.” Despite more than three decades of critical work on the special effects of Orientalism,
ORIENTALISM/OCIDENTALISM

its practical implications are still very much among us, including the engagement with the mirrored responses of Occidentalist tropes and their powerful political and cultural geographies.

SEE ALSO: Discourse; Imaginative geographies; Postcolonial geographies; Tropical geography

References


Outsourcing

Ingo Liefner
Justus Liebig University Giessen, Germany

The term outsourcing is used to describe a company’s decision to stop carrying out a certain business activity and to buy it instead. Outsourcing decisions lead to a redistribution of the production of commodities (usually components) and services between firms. Widespread outsourcing has had a profound impact on economic structures by reducing the dominance of vertically integrated and large-scale producers and creating networks of interrelated specialized firms. Outsourcing has also promoted the emergence of many producer-oriented service companies (Mudambi and Venzin 2010).

Through the lens of manufacturing firms, suppliers enjoy certain advantages over final commodity producers, which justify the outsourcing of component manufacturing, certain manufacturing activities such as final assembly, or production-related services (Venkatesan 1992). First, suppliers find it easier to generate economies of scale in production when they can specialize in a certain component or service and sell it to many customers. Second, they can thus achieve a more favorable cost structure; third, they then have the incentive to improve their competence and recognition regarding the component or service in which they specialize. These advantages of suppliers reduce their production costs compared to vertically integrated firms and thus lower the overall production costs of the final product as long as the cost reduction achieved through outsourcing offsets rising costs of coordination and transaction. Outsourcing is thus promoted by the advancement of information, communication and information technologies, and the decrease in their costs.

Outsourcing is the decision to separate business activities that are necessary to produce one product in terms of business organization and production technology. However, the technical separation can easily lead to a spatial separation of production (Quinn and Hilmer 1994). When there is little need for close collaboration between supplier and customer, and labor costs and other costs that vary between business locations are of greater importance, outsourcing can go hand in hand with spatial relocations of production. Hence, outsourcing has led to a loss of millions of manufacturing jobs in the United States and Europe, while it has accelerated the growth of new companies and jobs that help to organize production processes that are spread over many countries.

However, outsourcing does not necessarily imply spatial relocations, and outsourcing decisions can also lead to the emergence of interrelated networks of specialized producers in close spatial proximity, namely clusters or industrial districts. Hence, the term offshoring should be used to characterize the relocation of business activities across national borders, even if they are initially motivated by outsourcing decisions (Manning, Massini, and Lewin 2008).

Outsourcing is sometimes differentiated into material outsourcing and service outsourcing, with service outsourcing receiving broad media attention. Business services, computing, information, and so forth were long regarded as immobile, and offshore outsourcing of these...
activities has raised fears about job losses, not only among poorly qualified service workers but also among specialists (Amiti and Wei 2004).

Hence, outsourcing is a process that underlies the formation of spatial economic structures that are typical of today’s globalized and knowledge-based economy. Outsourcing decisions are behind the “second great unbundling”; that is, the disintegration of production in spatial terms (Baldwin 2006).

SEE ALSO: Industrial linkage; Vertical integration

References


According to *Collins English Dictionary* (2000), “overlay” is defined as “to lay or place something over or upon (something else).” In human geography, overlay is a “process of draping layers of map information on top of each other. Such a process exposes how categories of data intersect and can reveal important patterns of convergence and divergence.” In Hoyt’s classic study *The Structure and Growth of Residential Neighborhoods in American Cities* (1939), he overlaid areas of low rent, high percentage of buildings requiring major repairs, older building stock, and high percentage of nonwhite population, suggesting a strong correlation between the variables (Castree, Kitchin, and Rogers 2013). As DeMers states (2005), the process of overlay is to “place the cartographic representation of thematic information of a selected theme over that of another. It is intuitive that the application long preceded the advent of modern electronic geographic information systems (GIS).” The traditional approach of overlay is to compare polygons over a coverage/area/sector/region, or a coverage/area/sector/region over polygons. A very early application of the analog overlay technique dates back to the days of the American Revolution. During the Battle of Yorktown, in 1781, George Washington engaged the services of the French cartographer Louis-Alexandre Berthier, who employed map overlay to examine the relationship between Washington’s troops and those of the British general, Charles Cornwallis (as cited in DeMers 2005). In modern geographic information systems, however, the term overlay has taken on a more complex significance. Overlay is a key and essential function in all GIS packages and is used to reveal spatial overlap and facilitate pattern analysis.

**Overlay operations**

In practice, map overlay consists of several features, including points, lines, and polygons.

**Vector overlay methods**

In point-in-polygon overlays, the input point features are assigned attributes of the surrounding polygons (Figure 1). One example is to investigate the association between wildlife habitats and classes of vegetation in ecological applications. For instance, Point 1 and Point 2 are two points of wildlife location. Type A and Type B refer to different vegetation cover types. The analyst will be interested in the extent of different types of wildlife that fall into different vegetation covers. Other point-like events, such as occurrence of disease, must be compared to the properties of the surrounding environment.

In line-in-polygon overlays, the identical line features appear in the output, but each of them is intersected by the polygon boundaries when the polygon and line overlay each other (Figure 2). Each segment on the output amalgamates attributes from the polygon and the line map (Chang 2004). One example is to identify pavement materials for a designed road. The input map contains a designed road,
Figure 1  Point-in-polygon overlay.

Figure 2  Line-in-polygon overlay.

Figure 3  Polygon-on-polygon overlay.

and the overlap map is a pavement material map. The output shows a designed road divided into segments, with each road segment having a different set of pavement materials from its adjacent segments.

Polygon-on-polygon overlays involve two polygon maps for which the output integrates input polygon boundaries and those from the overlay maps, in order to generate a new set of polygons (Figure 3). Each new polygon contains attributes from both maps, and these attributes are not the same as those of neighboring polygons (Chang 2004). One example is to examine the common area of land parcels and flood zones. Attributes of both feature classes carry through the overlay operation. A single land parcel can be split into two parcels if only one part is inside the floodplain. An attribute will be given to indicate whether land parcels are inside or outside the flood zone.

**Raster overlay methods**

In contrast to vector overlay methods, which include the geometric objects of points, lines, and polygons, raster overlays work on cells and grids. Raster-based map overlay operations can be implemented on fixed cell blocks, whereas the vector operation works on irregular polygon/line boundaries. In other words, the raster operation is computationally more efficient than the vector operation. However, an additional process of resampling is required to synchronize the resolution (cell size) of each map before graphical overlay. Raster data such as remotely sensed data, digital elevation models, digital orthophotos, digital raster graphs, scanned maps, and graphic files are capable of raster map overlay. Usually, numeric values are assigned to each cell (pixel), allowing mathematical manipulation (calculation) of the layers, such as addition, subtraction, multiplication, and division. A new value is assigned to each cell in the output layer.

Figure 4 shows an example of raster overlay by mathematical addition, that is, two input raster maps added together to generate an output raster map with the values of each cell summed.

**GIS software applications**

With the application of state-of-the-art GIS technology, overlay operations can be processed digitally and automatically. For instance, the analysis tools/overlay toolset of the updated commercial GIS software includes “identity,” “intersect,” “symmetrical difference,” and “union” functions.
First, “identity” calculates the geometric intersection of the inputs, and illustrates feature classes and feature layers (ESRI 2009) See Figure 5.

Second, “intersect” calculates the intersection of the input features geometrically. Features or parts of features which overlap in all layers and/or feature classes will be recorded to the output feature class (ESRI 2009) See Figure 6.

Third, “symmetrical difference” calculates a geometric intersection of the input and update features. Features or parts of features in the input and update features which do not overlap will be recorded to the output feature class (ESRI 2009) See Figure 7.

Fourth, “union” calculates a geometric intersection of the input features. All features will be inserted to the output feature class with the corresponding attributes (ESRI 2009) See Figure 8.

Map overlay is the “placing of multiple thematic maps in precise registration, with the same scale, projection, and extent, so as to allow a compound view” (Clarke 2011). Using modern GIS software, a three-dimensional (3-D) view of map overlay can be produced to enable analysis of shadow, sunview, viewshed, perspective view, wind ventilation, and also application in urban planning (Figure 9), for instance, analysis of urban morphology structure.

Another example is 3-D building temperature data overlaid on a thermal image (Nichol and Wong 2004) (Figure 10). The displayed color
OVERLAY, GRAPHICAL

of a pixel is defined according to the calibrated temperature of the thermal image. It shows the temperature difference between the national stadium of Hong Kong, having a cover of artificial turf, and the Hong Kong racecourse, with real grass cover. The building temperatures are also depicted on the facets of 3-D building models, which show the highest facet temperature on the national stadium of Hong Kong, at over 38°C at the time of image acquisition.

Case study

Urban environmental quality in Hong Kong

In a study by Nichol and Wong (2006), the GIS overlay technique was applied to integrate different datasets and derive urban environmental quality (UEQ) maps. The datasets included air temperature, vegetation density, air quality, noise, building density, and building height. The range of values for each parameter was ranked on a scale, such as 1–10 or quantiles. The resulting UEQ values then corresponded to the sum of data layers, at 4 m and 64 m spatial resolutions. The GIS overlay method could more readily accommodate data originating from different resolutions or units of measurement.

A region-based overlay approach at electoral district level was also evaluated in the study. The aggregate score for each district was obtained by summing the ranked scores for each variable. Higher values represented the most desirable environmental quality in the Kowloon peninsula.

Geospatial Information Hub in the Hong Kong Lands Department

Geospatial Information Hub (GIH) is a web-based information platform for searching, displaying, and sharing a large amount of geospatial data through the Hong Kong government intranet. Several geospatial datasets (e.g. digital maps, aerial photographs, land status information, public facilities, and building address) can be overlaid for display on a user interface. The GIH provides graphical overlay, scaling, panning, and query functions, which reduce resource duplication and search time. Up to December 2012, over 83000 officers in 32 government departments had gained access to geospatial data through this platform, thus greatly reducing the need for hardcopy printouts and paper usage (Tsoi 2007).

GIS applications in the Hong Kong Observatory

The Hong Kong Observatory enhances the tropical cyclone track information webpage by overlaying satellite images to depict the pattern and coverage of clouds associated with tropical
cyclones. The pathways of tropical cyclone movement can be illustrated and the accuracy of the forecast track is indicated (Hong Kong Observatory 2013). In addition, the “Regional Weather in Hong Kong” webpage designed within GIS and HTML5 (Hyper Text Markup Language 5) frameworks has been developed and launched. Several types of weather observations (e.g., rainfall, wind, radar image, lightning location) can be visualized on the webpage using graphical overlay techniques. The Hong Kong Observatory fully utilizes the geographic information systems overlay technique, and precise locational positioning techniques, for meteorological modeling, short-term weather forecasting, and long-term climate change studies.

Urban climatic analysis map from the Hong Kong Planning Department

An urban climatic analysis map (UC-AnMap) collates meteorological, planning, land-use, topographical, and vegetation data, and, based on their spatial relationships and their effects on air ventilation capacity and thermal comfort, these are analyzed and evaluated. In the UC-AnMap, layers of information are overlaid and displayed graphically in a geographic information system. These include building volume, topography, green space, ground coverage, natural landscape, proximity to openness, and wind information (Planning Department 2012). The UC-AnMap addresses the planning and design issues related to urban climate in order

Figure 10  3-D model using raster overlay method. Reproduced from Nichol and Wong 2004 © Elsevier.
to formulate urban development strategies in relation to urban morphology.

Conclusion

Overlay is a GIS spatial operation in which different layers of maps, features, and vector and raster data registered in a common coordinate system are superimposed. Overlay is also an operation for determining the spatial coincidence of features. The spatial relationships between features in the same geographical space can be analyzed so as to pinpoint the required information from a set of stacked layers.

The overlay function can work for diverse datasets and types of data (e.g., vector, raster). The vector overlay approach offers point-in-polygon, line-in-polygon, and polygon-on-polygon functions to determine the spatial coincidence of points, lines, and polygons. The raster overlay approach offers a more efficient operation to determine spatial coincidence from cells and pixels. In current GIS operational systems, functions of “identity,” “intersect,” “symmetrical difference,” and “union” are available for graphical overlay, and these are the key functions in spatial data analysis. In recent years, planning and environmental authorities, surveyors, and researchers have been using graphical overlay operations as objective topological procedures in their daily work.

SEE ALSO: Data structure, raster; Data structure, vector; Geographic information science; Overlay, topological

References


A topological overlay is used to combine and manipulate spatial data by the spatial relationship of their geometries. It relates multiple features and joins attribute tables using a quantitative algorithm that operates on their respective topologies. Attribute values from both input and overlay datasets are passed on to the output dataset. In a topological overlay, two or more layers of data are combined according to a specific set of rules. A new output dataset is created with both geometry and attribute data from the input dataset along with updated topological information. It is the most required and common GIS technique in spatial data processing.

A topological overlay is different from a graphical overlay (image overlay), which neither requires attribute updates nor creates a new dataset. Goodchild (2010) highlighted that many mash-ups simply overlay one layer on another in a visual spatial join rather than making a pure topological overlay.

Although overlay works in both vector and raster format, a topological overlay is mainly applied to vector formats because the overlay procedure involves the rebuilding of the topological relationships that make layers work. A topological overlay results in a new layer with attributes of the original input data layers. It also requires the rebuilding of topology for arc, node, and polygon. A new layer is generated to include the updated topological information. Thus, it is computationally expensive, CPU intensive, and often time-consuming. The processing time of a topological overlay is proportion to the number of vertices in the two input datasets.

The logic behind how topological overlay works is based on rules of Boolean logic. Boolean algebra is based on the basics of binary logical operations of the values one and zero. Boolean operations such as AND, OR, XOR, NOT are applied in GIS for the overlay of multiple data layers to check whether a particular condition is true (1) or false (0). The rules of Boolean logic are used to operate on the attributes and spatial properties of geographic features. The AND Boolean operator represents conjunction, where results are true for all areas that meet both the first and the second criterion. The OR Boolean operator is related to disjunction, where results are true for all areas that meet either the first or the second criterion, independent on the areas overlapping or not. The XOR Boolean operator means exclusive disjunction, where results are true for all areas that meet either the first or the second criterion but not both. The NOT Boolean operator corresponds to negation, where results are true for all areas that meet the first criterion but not the second.

A topological overlay allows users to combine two or more feature classes to intersect, erase, modify, or update datasets and results in a new feature layer with the updated attributes of the input and overlay dataset values. Using the topological overlay on feature data requires both an input dataset and an overlay dataset. The input feature class can include points, lines, or polygons, but the overlay feature class can only comprise polygons. Each in the overlay affects how the data from the input feature class and overlay feature class are combined, which affects
both the spatial data and attribute data of the output dataset. Exemplary methods of a topological overlay include erase, identity, intersect, symmetric difference, union, update, and so on.

Erase overlay creates a layer by overlaying the input layer with an erase polygon. It allows users to remove data from the input layer based on overlay polygons. This is based on the NOT Boolean operator. For example, the city generates a flood map for residents to inform which roads are not damaged by flood and are safe to drive on. The urban planner could input roads and use the erase with an overlay of the areas most affected by the flood to remove the segments of roads in areas affected by flooding.

Identity overlay creates a layer by combining the input layer and identity layer to contain features with attributes of both input and identity datasets where they spatially overlap. This method can be used in a situation where neighborhoods and school districts have independent and overlapping boundaries and an analyst would like to map the neighborhoods that constitute the same school district. With the identity overlay, the urban planner can also take into account proximity to libraries. The output dataset contains all of the data from the input dataset, with the addition of the overlay dataset where it overlaps with the input. If the input dataset consisted of polygons of neighborhood boundaries and school district boundaries, the overlay dataset could consist of a two-mile buffer around the library in that area, so the polygons of the output data will contain data about school district, neighborhood, and whether the area is within two miles of the library.

Intersect overlay creates a layer with only values that overlap from both the input layer and the intersect layer. It combines the input layer and overlay layer to produce an output layer that only contains values shared by both the input layer and overlay layer. This is based on the AND Boolean operator. This can be used in a flood risk assessment, where the input map of residential homes is intersected with a buffered-river intersect layer to determine the residential homes that are at risk of flooding. Imagine that after a flood, an urban planner wants to determine which families were most affected by the flood. This planner could intersect an input dataset of parcels associated with homes and families with a dataset of the areas most affected by flooding. The output dataset would only show the parcels highly affected by flooding and the parcels located within that area. Then the urban planner could use the attribute table to identify which families own those parcels that were most affected by the flood.

Symmetrical difference overlay combines two input layers and creates an output layer where any overlapping data is removed and both input datasets are merged to single layer. This method could be useful in calculating crime rates in the area managed by a police precinct compared with crime rates in the area managed by another precinct if the managed areas overlapped. On a very basic level, to remove the bias of the rate of crime in the area managed by both precincts, the symmetrical difference overlay could be used in analysis. It is often used to remove data. In looking at a residential school district, an urban planner is determining if access to only one library is enough to achieve a specific reading standardized test score. Students who have access to two libraries are excluded from this analysis so the urban planner can use the symmetrical difference overlay on a two-mile radius around libraries in the neighborhood. Both areas without any access and areas with access to two or more libraries will be removed from the dataset and only areas with access to one library will remain as a single dataset.

Union overlay combines two input datasets by including all features and attributes. This method
can be used in a situation where neighborhoods and school districts have independent and overlapping boundaries and an analyst would like to have a layer with polygons that include both school district and neighborhood values to obtain statistics from both datasets. This is based on the OR Boolean operator. With the union method, the neighborhood feature can be combined with the school district feature and the urban planner can select areas that contain attribute data indicating which neighborhood and which school district the polygon is associated with.

Update overlay computes a geometric intersection of the input layer and update layer. It replaces overlapping parts of the input layer with features from the update layer. It maintains features of an existing dataset and an updated feature in which the updated features and attributes replace input features. The attributes and geometry of the input features are updated by the update features in the output feature layer.

Some GIS software provides other types of topological overlay such as clip and paste. The resulting topology from clip overlay includes areas that appear in the input topology, except where they are outside the boundary of the clip layer. The input topology is clipped to the outer boundary of clip layer. The resulting topology from paste overlay includes the paste layer topology and any areas of the input topology that extend beyond the boundaries of the paste overlay topology.

Topological overlay among different feature types is possible. A point-in-polygon overlay is applied where input comprises points and the overlay dataset is a polygon. For example, it is illustrated by selecting houses (point) with children below the age of 18 within a school district boundary (polygon). A line-in-polygon overlay is used when the input comprises lines and the overlay dataset is a polygon. As is illustrated in selecting roads (line) from the flood-risk area, output is roads in a county boundary. A polygon-on-polygon overlay is often used when the input and the overlay dataset both comprise polygon, for example, determining which parcels are affected by a flood by combining parcel boundary with polygons of areas affected by the flood.

There are several challenging issues that could complicate topological overlay operations. Implementation of topological overlay may demand large amounts of memory and time to process the operation, depending on the number of vertices in the input datasets. Efficient algorithms for executing topological overlay in RDBMS are needed. Performing topological overlay functions on a road network dataset could be problematic, having topological errors such as connectivity, adjacency, missing segment, and so on. Although the road network can be rebuilt for integrity, it is not easy to fix those topological errors. Floating nodes, dangling arcs, and spurious or sliver polygons can be problematic during topological overlay although tolerance setting may resolve these issues. More details on indeterminate boundary issues by topological overlay, are given in Zhan and Lin (2003). While it is theoretically feasible to overlay multiple (more than 3) data layers at one time, most GIS software do not provide multiple data overlay operations where more than two data layers are involved.

See Also: Data structure, vector; Geographic information science; Overlay, graphical; Topological relations

References

OVERLAY, TOPOLOGICAL


Further reading


Paleoclimatology

Michael F.J. Pisaric
Brock University, Canada

Climatic change in response to anthropogenic activities during the twentieth and twenty-first centuries has stimulated considerable research into both anthropogenic-forced climate response and natural climate variability. While Earth’s climate is constantly evolving and changing as natural external and internal forcing mechanisms alter the receipt of solar radiation at timescales ranging from seasons to hundreds of thousands of years, the impact of anthropogenic activities on climate has led to much debate. To better understand the contribution of human activities on climate and gain a better understanding of natural climate variability inherent in the global climate system, studies of past climate are warranted. Paleoclimatology is the science that studies these changes throughout the history of Earth.

Paleoclimatology of the early-Earth, or “deep time,” has largely been the focus of geologists, geochemists, and geophysicists. Contributions by geographers are most often focused on the Quaternary period, representing the past 2.6 myr (myr = million years), or so, of Earth’s history. Thus, this entry on paleoclimatology will only briefly examine the paleoclimatology of the early-Earth and draw upon examples and methods to study paleoclimates of the Quaternary period. The Quaternary period is of importance from a paleoclimatological perspective, since this is the period of Earth’s history from which the greatest abundance of physical data remains today on the landscape to examine past changes in climate; it also represents the geologic time period during which the human species evolved and has now emerged as one of the leading drivers of change affecting global ecosystems.

The global climate system at the start of the twenty-first century appears to have entered a new state, as anthropogenic activities have significantly altered the concentrations of greenhouse gases such as carbon dioxide and methane. Understanding the potential impacts of these atmospheric changes requires an examination of climatic conditions during previous periods in the Earth’s history when climates were similar. Paleoclimatology provides a window through time to place current climatic conditions and the rates of recent climatic change in a long-term context. Paleoclimatic records are typically much longer than those afforded by instrumental data records. Thus, paleoclimatology not only provides a glimpse into past climatic conditions but also can assist in predicting how climate may change in the future. It is the ability to look backward in time to inform the future that makes paleoclimatology such a powerful and important tool.

Methods in paleoclimatology

Instrumental records of past climatic conditions (e.g., temperature and precipitation data collected using standard meteorological instruments, satellite-derived data) are relatively short and often overlap entirely with the period of anthropogenic impacts on global climates. Therefore, there is a need to develop longer records of past climatic conditions to place recent climate trends in a long-term perspective.
and to better understand the full range of natural variability inherent in global climate systems. To do this, paleoclimatologists rely on proxy climate data sources. There is a host of tools available to the paleoclimatologist to reconstruct and interpret past climatic conditions. These proxy indicators of past climate come from biotic or abiotic sources and cover a range of timescales, from decades to billions of years. Today, entire research communities are focused on the use of tree ring growth records (dendroclimatology), corals and other banded remains from oceans and caves (sclerochronology and speleothems), lake sediment analysis (paleolimnology), ice cores, and ocean sediment records (paleoceanography) to reconstruct past climatic conditions. While the breadth of paleoclimate methods is too vast to fully review here, several of the more widely utilized methods are discussed.

**Annually banded proxy records**

Instrumental data have the advantage of being collected at high temporal resolution (hourly, daily, monthly) but normally these records do not extend over long timescales. Some proxy data sources, such as tree rings and corals, provide annual proxy data information and, in some instances, subannual resolution, thus potentially allowing the extension of gathering climate data by hundreds to thousands of years.

**Dendrochronology**

Dendrochronology, the science of dating tree ring growth records, is a widely used paleoclimatic technique. Each year trees growing in regions that experience seasonal climate variability produce annual growth rings. While trees in many parts of the world form annual growth rings, dendrochronology is most useful for paleoclimatology in highly stressed environments, such as at the ecological limits of a particular species. For example, white spruce (*Picea glauca*) grows up to treeline across most of northern Canada. At the northern treeline, growth of trees is controlled primarily by temperature during the growing season (i.e., summer months), with warmer summer conditions leading to wider tree rings and cooler conditions leading to the development of narrower growth bands. This is an example of the principle of ecological amplitude, which represents the range of environments in which a tree species may grow and reproduce. Ecological amplitude is important in dendroclimatology because individual trees that are most useful to dendrochronology are often found near the margins of their natural range, whether that is latitudinally, longitudinally, or at higher or lower elevations.

While the principle of ecological amplitude is fundamental in dendroclimatology studies, the use of tree rings for studying past climatic conditions is also predicated on our understanding of several other guiding principles.

- Uniformitarianism, as applied to paleoclimatic studies in general, proposes that the physical and biological processes that link current environmental processes with current patterns of tree growth must have been in operation in the past. Championed by the Scottish geologist James Hutton, uniformitarianism is often linked with the idea that the “present is the key to the past.” This is critical in paleoclimatic studies, including dendroclimatology, where relationships between climate and proxy indicators are used to develop records of past climate. Within the field of dendroclimatology, there is debate about how stable these climate–growth relationships are. The divergence issue that has been identified at treeline sites in northwest North America and Eurasia suggests these relationships are not time-stable and may have been changing.
during recent decades. There is now ample evidence that northern treeline sites, which were previously temperature limited, are now showing evidence of increased drought stress related to warmer summer conditions. These findings raise uncertainty regarding uniformitarianism and dendroclimatology.

- The principle of limiting factors suggests tree growth (or the growth of any biological organism for that matter) can proceed only as fast as allowed by the primary ecological mechanism(s) that restricts growth. Sometimes, more than one mechanism operates to restrict growth but careful study of the current relations between the growth/presence of an organism and their climatic controls is critical in any paleoclimatic study. In the case of the northern treeline, the mechanism restricting tree growth is temperature during the growing season. The principle of limiting factors is important to dendroclimatology because annual growth rings can only be matched between trees (cross-dated) if one or more factors are in limited supply and persist for sufficiently long periods of time and across a wide geographic area to cause tree rings to vary the same way in many trees.

- Site selection is critical in paleoclimatic studies. By careful selection of sample sites, an attempt can be made to find trees sensitive to the environmental variable of most interest. The dendrochronologist must select sites that will maximize the environmental signal being investigated, and thus proper site selection involves an a priori assessment and knowledge for the selection of trees for sampling. Even so, as demonstrated by the divergence issue, there can still be surprises.

Sclerochronology and speleothems

The oceans and other water bodies cover approximately 70% of the Earth’s surface. Monitoring of these vast expanses of the Earth’s surface is temporally and spatially limited compared to terrestrial environments. Similar to dendrochronology, sclerochronology and speleothems make use of annually deposited layers as proxy indicators of past climatic conditions. Sclerochronology involves the use of accretionary bands formed in corals and mollusk shells and other structures such as otoliths (fish ear stones). These banded remains can represent various time steps from daily and subdaily accretions to annual deposits. Analogous to dendroclimatology, paleoclimatologists often make use of physical measurements of the size of the accretions and their chemical composition. Sclerochronology can provide proxy records of past sea surface temperatures, precipitation, and river discharge. The main limitation of coral reefs as proxy climate data sources is their distribution, as they form only in tropical and subtropical regions where mean annual water temperature is >18°C and water is clear. While continuous records of climate from coral sequences are relatively short (spanning the past several hundreds to thousands of years), discontinuous records can be obtained from corals extending back to the Permo-Triassic extinction event (251 myr) and possibly as far back as the Cambrian (540 myr) for some “reef-like” structures (Pandolfi 2011).

Mollusk records provide similar proxy data information as corals but have a greater spatial distribution. Recent efforts to cross-date multiple samples, similar to matching tree ring width chronologies, are leading to the development of robust mollusk chronologies covering the past millennium and more.

On land, cave systems can provide ideal environments to preserve past changes in climate. Speleothems are accumulations of the carbonate mineral calcite (CaCO₃), which is released from rainwater when it percolates through the
PALEOClimatology

Ground and into a cave. Speleothems collectively include stalagmites (accretions on the floor of caves), stalactites (accretions extending from the ceiling), and flowstones (sheet-like deposits that form on cave walls and floors). Paleoclimatologists analyze chemical changes preserved in these deposits, including stable isotopes (e.g., $^{18}$O), or make measurements of the thickness of accretions. Paleoclimatic studies using speleothems can extend back several 100 kyr (kyr = thousand years) (Vaks et al. 2013) but periods of no deposition (hiatus) during dry periods are common. It is important to note that, while some speleothem records are annually resolved, most have decadal to centennial scale resolution.

Ice cores

Our understanding of the paleoclimate of the Earth has been advanced through the analysis of ice cores recovered from ice sheets in Greenland and Antarctica, in particular. Ice core studies have provided detailed information about Quaternary period climate and atmospheric change, with recent ice core drilling projects providing records of atmospheric greenhouse gas composition and climatic change during the past eight glacial/interglacial cycles or 800 kyr (EPICA 2004). Ice cores from Greenland and Antarctica have been recovered through several kilometers of ice by international teams of researchers. Air bubbles trapped in the ice provide a snapshot of atmospheric greenhouse gas composition during the Quaternary period and analysis of isotopes of oxygen provide key insights into changes in climate.

Paleolimnology and paleoceanography

Paleolimnology is the study of lake sediment records to reconstruct past climatic and environmental change from inland water bodies. Similarly, paleoceanography makes use of many of the same techniques but focuses on sediment records collected from the world’s ocean basins. Unlike the proxy sources discussed thus far, lake and marine sediment records are not normally banded and annual in nature. Instead, sediment records from lakes and marine basins are massive and exhibit limited structural components that are discernible to the naked eye (Figure 1a). Thus, chronology development is a critical step in paleolimnological and paleoceanographic studies. Paleolimnology normally makes use of radiometric dating techniques, such as $^{14}$C dating (for sediment deposited >150 years ago and as much as 40–50 kyr ago) and $^{210}$Pb dating for sediment deposited within the last 150 years. Occasionally, the depositional environment may be such that annual laminations (known as varves) occur in lake and marine sediments (Figure 1b). The deposition of varves normally occurs under anoxic conditions and in deep basins where seasonal overturn and mixing does not occur.

The analysis of lake sediment records in paleolimnological studies is based on the study of a variety of biological subfossils preserved in the sediment in addition to the physical, chemical, and mineralogical properties of the sediment. There are many different types of biological remains and chemical signatures preserved in lake sediment records that can provide insights into past climatic conditions, including siliceous algae (diatoms), charcoal, chironomids (nonbiting midges), Cladocera, pollen, pigments, and stomata (Figure 2). New techniques such as biomarkers and environmental DNA have great promise for unlocking new insights into past climatic conditions.

Lake sediment records are collected using a variety of coring equipment that is enabled either by gravity or by physically pushing, driving, or vibrating the coring apparatus into
PALEOCLIMATOLOGY

Figure 1 Example of massive sediment with little stratigraphy and a thin section scan of varved sediments. (a) ITRAX core scanner image of massive lake sediment from Fort McPherson, Northwest Territories, Canada. Sediment profile illustrated is approximately 50 cm in length (photo taken by the author). (b) Thin section scan (rectangle is 5 mm long) of varved sediments from West Lake, Cape Bounty, NU (Courtesy of Dr Jaclyn Cockburn). The couplets represent seasonal deposition with dark layers representing clay caps deposited under ice cover and lighter gray bands being coarser material deposited under higher energy regimes during the short melt season.

Paleolimnology and paleoceanography can provide insights into past climates across a range of timescales, from very recent changes occurring over the past few decades to millions of years in some of the larger inland lake systems and ocean basins. Ocean sediment coring and the climatic information it can provide are discussed in greater detail in the later section on Quaternary paleoclimates.

Paleoclimatology of the early-Earth

With the coalescence of materials remaining after the formation of the sun, the Earth and other planets were formed. The early sun produced far less solar energy than it does at the current time. These conditions had the potential to thrust the early-Earth into a glacial state but instead hyper or near-hyperthermal conditions characterized this period in Earth’s early history. In spite of the lower solar output, temperatures were believed to be much warmer than today during the Archean eon (2.5–4.0 gyrs) because the composition of atmospheric greenhouse gases was much different. Today, our atmosphere is comprised primarily of nitrogen (78%), oxygen (21%), argon (0.93%), and carbon dioxide (0.04%). While greenhouse gases such as carbon dioxide and methane comprise only a small proportion of the atmosphere today, during Earth’s early history volcanic activity likely produced higher concentrations of carbon dioxide and, later, with the emergence of the first primitive forms of life during the Archean eon approximately 3.5 gyr ago, atmospheric methane concentrations likely increased as well via anaerobic decay processes (Kasting 1993). Some estimates suggest carbon dioxide levels may have been 100–1000 times above current levels (Kasting 1987) and methane may have exceeded 1000 parts per million (ppm) (Kasting 2005). Today, methane concentrations in the atmosphere are around 1.8 ppm. With less solar output, an enhanced greenhouse effect due to higher concentrations of carbon dioxide was
the most likely mechanism that prevented the early-Earth from plunging into glacial conditions. This is known as the faint sun paradox.

One constant of climate today and throughout Earth’s history is change. While climate throughout the majority of Earth’s history can be typified as hyperthermal, long periods of glacial conditions have also occurred. In fact, paleoclimatic studies indicate there have been at least three episodes of glacial conditions that have occurred on global to near-global scales and on timescales of millions to tens of millions of years. This includes the earliest evidence of glaciation during the Neoproterozoic and widespread glaciation during the late Carboniferous and early Permian period. The Quaternary represents the most recent period of widespread glaciation, especially in the Northern Hemisphere. These extensive glacial episodes in the Earth’s early-history are thought to have enveloped the globe in ice and,
thus, are termed “snowball Earth.” The drivers of these extensive glacial episodes are difficult to ascertain but hypotheses include increased volcanic activity or a supervolcano eruption, changes in the composition of greenhouse gases in the atmosphere or changes in solar output. Any of these could lead to a decline in global temperatures and subsequent growth of ice sheets. However, the idea of a snowball Earth is still debated.

Quaternary paleoclimates

The Quaternary period or the most recent “ice age” encompasses the past 2.6 myr of Earth’s history. The Quaternary is characterized by numerous alternating glacial and interglacial episodes, commonly divided into the Pleistocene and Holocene epochs (Figure 3), with the Pleistocene epoch representing the glacial periods and the Holocene epoch (about the past 10 000 years) the most recent interglacial period. Using isotopes of oxygen \(^{18}\text{O}/^{16}\text{O}\) preserved in the shells of marine organisms, long sediment records from deep-sea basins record over 100 episodes of cold/warm climate fluctuations during the Quaternary period (Lisiecki and Raymo 2005). These are commonly referred to as marine isotope stages (MIS).

On land, long terrestrial records of climate change during the Quaternary period have come from the continental ice sheets on Greenland and Antarctica. The longest of these ice core records that have been recovered are from the Antarctic ice sheet, where a drilling program led by the European Project for Ice Core Drilling in Antarctica (EPICA) (1996–2006) recovered a 3270 m ice core from Concordia Station at Dome C that captured the past eight glacial/interglacial cycles (EPICA 2004). On Greenland, the Greenland Ice Core Project (GRIP) (1989–1992), Greenland Ice Sheet Project (GISP) (1971–1981), and GISP2 (1988–1993) recovered long records of paleoclimatic conditions from the Greenland ice sheet. While GRIP and GISP drilled through similar depths of ice as EPICA (GRIP penetrated 3028 m in depth; GISP2 reached a depth of 3053 m), the Greenland records span only the past 100–120 kyr (GRIP 1993; Johnsen et al. 1992), which is much shorter than those from the Antarctic continent.

Ice core studies utilize the changing ratios of isotopes of oxygen \(^{18}\text{O}/^{16}\text{O}\) preserved in air bubbles within the ice to study paleoclimate. Traces of these gases can be extracted from the ice cores, the composition of greenhouse gases can be determined, and the ratio of oxygen isotopes can be measured. By analyzing the composition of gases in these air bubbles, scientists have reconstructed not only the composition of the atmosphere in the past but also how climate changed and oscillated between glacial and interglacial conditions.

The Quaternary glacial/interglacial cycles were primarily driven by changes in the orbital geometry of the Earth over long timescales (Figure 4). These variations in the Earth’s orbital geometry are referred to as Milankovitch cycles or Milankovitch theory of climate (Hays, Imbrie, and Shackleton 1976), after the geophysicist and astronomer, Milutin Milanković. Milankovitch cycles define changes in the Earth’s orbital geometry on timescales of about 100, 41, and 21.7 kyr. These orbital changes led to variability in the receipt of solar radiation (Berger and Loutre 1991), which, in turn, led to the waxing and waning of the glacial ice sheets during the Quaternary period.

Orbital eccentricity, or the shape of the Earth’s orbit around the sun, varies from nearly circular (low eccentricity = 0) to somewhat elliptical (high eccentricity = 0.05) with periodicities of around 405 and 100 kyr (Figure 4a). Orbital
Figure 3  CO₂ concentrations (Lüthi et al. 2008) (red) and reconstructed temperature (Jouzel et al. 2007) (black) from the EPICA ice core, Antarctica during the past 800 kyr. Note recent CO₂ concentrations of around 400 ppm far exceed estimated concentrations during the past 800 kyr. Also apparent in the data is the change in magnitude and cycle length from a 41 kyr cycle to a 100 kyr cycle during the past 500 kyr. Numbers above and below time series indicate select MIS. The time frames of the Pleistocene and Holocene are also indicated.

eccentricity has been of interest regarding Quaternary paleoclimates due to the similarity in the timing of changes in eccentricity and glacial/interglacial cycles during the last 500 kyr, when a pronounced 100 kyr glacial cycle has dominated. However, it remains unclear how eccentricity drives the 100 kyr Quaternary glacial cycle since changes in eccentricity lead to only small changes in solar radiation receipt compared to those generated by changes in precession and obliquity (Berger and Loutre 1991). This is known as the 100-ka enigma or 100-kyr problem.

Obliquity, or the tilt of the Earth’s rotational axis, varies from 22.05° to 24.5° with a periodicity approaching 41 kyr (Figure 4b). When the tilt of the Earth’s rotational axis is closer to 24.5°, the difference between winter and summer solar radiation increases. This leads to warmer summer conditions and cooler winters. As the tilt of the Earth’s rotational axis decreases towards 22.05°, the opposite occurs and summer season conditions tend to cool.

The final component of the Milankovitch theory of climate is precession or the change in the orientation of the rotational axis of the Earth (Figure 4c(i)). With a periodicity of about 21.7 kyr, the Earth’s rotational axis wobbles like a spinning top and the direction it points varies.
Figure 4 Changes in the orbital geometry of the Earth based on the Milankovitch theory of climate. It is believed that these orbital variations are the drivers of glacial/interglacial cycles during the Quaternary period. (a) Eccentricity or the change in the Earth’s orbital geometry around the sun that varies with periodicities of around 400 and 105 kyr. (b) Variations in the tilt of the Earth on its axis from about 22° to 24.5° with a periodicity of about 41 kyr. (c) Changes in precession that alter the timing of perihelion and aphelion (i); (ii) represents precession at the current time, with the Earth being nearest the sun (perihelion) during the austral summer; (iii) precession during the late Pleistocene–Holocene transition when perihelion occurred during the Northern Hemisphere summer.

This causes perihelion (when the Earth is nearest the sun) to occur during the austral summer or the Northern Hemisphere winter (Figure 4c(ii)). During the Pleistocene–Holocene transition (around 10 kyr), precession was such that the Northern Hemisphere summer occurred during perihelion (Figure 4c(iii)).

Precessional change of the rotational axis increased insolation during the Pleistocene–Holocene transition, leading to treelines across the circumpolar boreal region migrating poleward. Paleoclimatic evidence from pollen records and radiocarbon-dated macrofossil stumps from North America and Eurasia, suggest that summer temperatures were several degrees warmer than at present and remained so until summer insolation began to decline again during the past 4 kyr or so.

While early to mid-Holocene epoch climate was warmer than present, there is also
paleoclimatic evidence that storm tracks and precipitation regimes changed. At the Pleistocene–Holocene transition, the Sahara desert was wetter than at present and was covered with savanna-like vegetation. As solar insolation declined after 4 kyr, the Sahara gradually became drier. In the American southwest, large lake systems (pluvial lakes) developed in regions that today are semiarid.

The Anthropocene and the Great Acceleration

The Holocene epoch is often described as a period of relatively stable climatic conditions given the lack of wide climatic swings similar to the glacial/interglacial cycles during the Pleistocene epoch. However, the past 10 kyr of Earth’s climate history have been witness to important periods of climate variability, including the Medieval Climate Anomaly and the Little Ice Age. We are now on the doorstep, and have possibly crossed the threshold, of the next period of climatic variability in the Earth’s history. Many believe we now find ourselves in the Anthropocene epoch, the geologic period of Earth’s history where the impact of humans has become one of the primary drivers of climatic and environmental change. While still not formally recognized as a geologic unit, usage of the term has become widely accepted.

Many argue the Anthropocene epoch commenced with the start of the Industrial Revolution in Europe during the latter part of the eighteenth century (Crutzen and Stoermer 2000). Based on evidence obtained from ice cores, it is at this time that human activities began to noticeably alter the chemical composition of the atmosphere. Carbon dioxide increased from about 280 ppm in 1750 to current levels of around 700 ppm and methane from about 700 ppb in 1750 to current levels of around 1700 ppb.

Regardless of how human impacts are defined from a geologic perspective, anthropogenic impacts on recent climate have been significant. Since the mid-to-late twentieth century, climate has warmed at an accelerated pace, driven by changes in the concentration of greenhouse gases in the atmosphere. Since 1950, synchronous changes in socioeconomic and Earth systems trends occur and begin to accelerate in tandem for the first time in Earth’s history. These changes are certain to continue impacting global climate in the future.

Paleoclimate studies have taught us much about how the Earth’s climate system has responded to changes in climate drivers in the past, providing a window to look back in time and better inform us for the future. However, these studies also confirm that global climate systems are now changing at rates that exceed the range of natural variability noted in many paleoclimatic studies. The divergence issue in dendroclimatology is perhaps an emerging example of this. In a rapidly changing Anthropocene world, interpretations of the relationships between climate change and proxy data sources will continue to yield important findings about past climatic conditions, and likely many new surprises as well.

SEE ALSO: Anthropocene and planetary boundaries; Climate change, concept of; Dendroclimatology; Global climate change; Holocene; Quaternary glaciations

References


Further reading


Paleoecology and methods

The study of paleoecology involves reconstructing past interactions between organisms and their environment. Such biogeographical data about the past provides valuable information relevant to understanding modern and future responses of plants and animals to environmental changes, including climate shifts and human impacts on terrestrial landscapes and aquatic ecosystems. Organisms that offer such paleoecological data are those preserved as fossils in suitable geologic settings, such as those buried in the anoxic sediments of lake bottoms and peatlands, and those remains sheltered in caves, particularly in arid environments. These fossils become “paleoecological proxies” of past environments when they are identified to species or genus by comparison to the remains of modern representatives of the same taxa, that is, the “modern analog concept.” For instance, fossil cones and pollen of white spruce can be directly compared to the components of a living white spruce tree to confirm species identification. And then by applying the modern analog concept a paleoecologist can use data about the modern habitat and range of this species today to infer the paleoecology of this taxon in the past. Using the example of white spruce, this species today occupies the boreal forest biome of northern North America and its southernmost limit is dictated by the 18 °C July isotherm, as summer temperature warmer than this curtail seedling growth of this boreal tree (Yansa 2006). Therefore, a paleoecologist who finds white spruce fossils buried at the bottom of lakes in eastern and central North America, south of the boreal forest today, can infer a boreal climate and habitat for that area in the past. Such a subarctic boreal reconstruction for the Late Pleistocene (≈18,000 to 11,600 years ago) vegetation of this area of North America has been made by several scholars, even though the present-day vegetation is deciduous forest or grassland correlative to a warmer climate.

Pollen and plant macrofossils (cones, seeds, leaves, fruits, and other macroscopic plant remains) are common proxies studied in paleoecology in that they reconstruct the species composition of vegetation communities over past millennia that changed in response to shifts in temperature and moisture regimes. Certain plant fossils can also indicate past human impacts on landscapes, particularly the clearing and burning of land for agriculture, by a decrease in forest pollen abundance, increase in the pollen of grass and open vegetation types, and the appearance of fossils of domesticated plants including maize (corn), wheat, rye, beans, and wild rice. Also, such pollen of crop plants is usually associated with high concentrations of charcoal fragments, indicative of fire, and such information is correlative with archaeological records of prehistoric human–environment interactions. Analyses of charcoal records with an eye toward identifying peaks of greater charcoal accumulation can reconstruct local fire histories of the past and intervals of greater fire occurrence, but cannot discern individual fires.
PALEOECOLOGY

These and other fossils can be collected from peatlands and lake bottoms by excavating trenches to provide vertical exposures (if the water table is lower today than in the past) to sample sediments containing fossils. Or, more commonly, existing lakes are sampled using coring equipment (such as a Livingston piston corer) from a boat in open water or from the frozen surface of a lake in winter. The coring device is typically 1 m in length, so it needs to be put down the same hole to repeatedly collect several meters of lake sediment before reaching an impenetrable layer, such as sand or glacial till. Plant macrofossil and/or charcoal from certain sampled levels of cores and trenches can be sent to radiocarbon (\(^{14}\text{C}\)) laboratories for dating, thereby providing chronologies for sites. The limit of \(^{14}\text{C}\) dating is approximately 50,000 years, but the study of pollen and plant macrofossils can go back further in time, however, the chronologies of these are based on other, indirect dating techniques. The majority of plant fossil studies focus on the last 20,000 years for which fossil preservation is best and the results can be directly applied to understanding modern plant distribution patterns and prehistoric human impacts on landscapes. The last 150 years, up to the present day, cannot be dated using \(^{14}\text{C}\) dating, due to half-life limitations of this isotope, however, \(^{210}\text{Pb}\) and \(^{137}\text{Cs}\) are isotopes that cover this period for fossil records.

Plant fossils are isolated from the sediments of lakes and peatlands by two means: (i) sieving 50 to 250 cm\(^3\) of mud for the larger plant macrofossils and charcoal fragments under tap water, and (ii) for pollen and microscopic charcoal, treating 1 cm\(^3\) of sediment in test tubes to a series of chemicals to remove carbonates, silicates, and cellulose. The macroscopic remains are viewed in petri dishes in distilled water under \(40\times\) magnification, whereas the microscopic pollen and charcoal are dispersed in silicon oil, a tiny amount of which spread on a microscopic slide and viewed under \(400\times\) magnification. For both, the fossils are identified and counted to the lowest taxonomic level, species for most plant macrofossils and usually to genus for pollen. The low numbers of seeds and other macroscopic fossils recovered in a lake or peat core permits only displaying the results as counts.

In contrast, the greater abundance of pollen allows for counts of 400 or more pollen grains per sample and hence statistical analysis and modeling can be performed to elucidate on pollen–climate relationships. Specifically, canonical correspondence analysis (CCA), detrended correspondence analysis (DCA), and principal components analysis (PCA) can be done to derive paleoecological transfer functions. These transfer functions are equations based upon variations in the present-day abundance and geographic distribution of pollen taxa relative to modern climate gradients. These types of inverse modeling assume that climate is the only environmental factor limiting the distribution of plants, which is true at the biome level, but not at the local scale where soils, topography, and disturbance also play important roles. Hence, modeling of pollen data is mainly done to understand biome shifts over past millennia. Other proxies, described below, are more climate-dependent and hence provide transfer functions through the analyses of calibration sets, which yield better paleoclimate reconstructions.

Other biotic proxies

Conveniently, the same lake and peatland sediments that contain pollen and plants macrofossils also hold several other plant and animal proxies, with each providing unique data about past environments, including paleoclimates. Diatoms, for instance, are a group of microscopic eukaryotic algae (phytoplankton) found in freshwater and
marine environments, with the focus here being on those from lakes. Most diatoms are unicellular and comprised of valves (theca) that can be identified to species. Given the great taxonomic diversity of these algae ($\approx 100,000$ species worldwide), each species provides very detailed data of the chemistry and temperature range of the water body it once occupied, and thereby offers valuable paleoecological information. Diatoms are comprised of silica (SiO$_2$) and hence are best preserved in cold, soft water lakes of high latitudes and alpine locations where dissolved silica is abundant and are correspondingly poorly preserved in warm alkaline or saline lakes of the low latitudes. The majority of diatoms are either planktonic, being mainly suspended in the water column, or benthic, those attached to plants, stones, and sediment around the margins of lakes. The ratio of planktonic versus benthic species provides meaningful information about water depth changes and inferred precipitation regimes over time.

For each lake where fossil diatoms will be investigated by deep cores, researchers sample all modern diatom habitats using sediment traps to capture and identify living diatoms at the sediment–water interface as well as collect water samples for chemical analysis. These data are then compared to modern climate data. CCA is used to provide an ordination of each living species against a modern environmental variable, such as temperature, water pH, or phosphorus, and through this method diatomists produce a “calibration training set” for each lake (Battarbee et al. 2001). Then the training set is applied to the fossil species data from that lake to reconstruct the paleoecology of the area, again invoking the modern analog concept. Fossils diatoms are collected using the lake sediment coring methods, described above, and a series of chemicals clean the valves so details of cell walls are visible for species-level identification under 750× magnification. Typically, a diatom study involves the investigation of a series of lakes with a focus on a gradient of interest (e.g., salinity, pH, or total phosphorous) within a region. Fossil diatom analyses have provided invaluable information about (i) past and recent climate change from diatom-inferred salinity changes of lakes situated in semiarid environments, and species changes in lakes attributable to (ii) acid rain pollution by fossil fuel burning, and (iii) cultural eutrophication caused by elevated lake phosphorus levels from fertilizer runoff and other anthropogenic sources.

There are several types of animal fossils preserved in and recovered from lake sediments that are studied in paleoecology. One such animal proxy is ostracods (Ostracoda). These are microcrustaceans in the same family as lobster, crab, and shrimp species, but are tiny, about the size of plant seeds, and have clam-like shells formed by two valves of calcite (CaCO$_3$). Each ostracod species is either planktonic (near the water surface), swimming, or benthic (crawling on and into the mud at lake bottoms). Many ostracods eat decaying organic matter and some fish species prey on them. Interestingly, the eggs and adults of some ostracod species are able to tolerate months of desiccation and freezing and they will emerge from dormancy once suitable water conditions occur. Like diatoms, each species of ostracods is correlative to a certain temperature range and to specific gradients for dissolved oxygen, salinity, and other chemicals in lake waters, and hence calibration training sets (based on modern data to interpret fossil data) have to be constructed as well. Being arthropods, ostracods grow by molting and most individuals will molt eight times prior to adulthood, so the adult shells are the ones identified to species based on differences in shell morphology (size, shape, muscle attachment scars, etc.). The growing shell absorbs
different weights (various isotopes) of a chemical depending on the temperature and salinity of the surrounding water. Geochemical analyses of stable oxygen isotopes (the ratio of heavy $\delta^{18}O$ vs. light $\delta^{16}O$) from ostracod shells are commonly performed to determine the relative amount of fresh water in lakes over time, which is used as a proxy for determining moisture sources in the past. For example, such research has reconstructed the relative input of moisture from the Gulf of Mexico air mass versus moisture from the Pacific air mass for the last 12,500 years for central North America and hence can infer the dominant patterns of atmospheric circulation over these millennia (Yu 2000).

Another small animal proxy derived from lakes is chironomids (Chironomidae), commonly referred to as midges or “lake flies,” which are mosquito-like, aquatic insects found in lakes and slow-moving streams. Like ostracods, they are an important food source for fish. The larval head capsules of midges are made of a hard, semi-transparent material (chitin) that become embedded in sediment and morphological variations of these permit species-level identification at 50× magnification. A calibration set, modern data regarding living midge species in association with current water temperature and chemistry data, are also required to interpret the fossil chironomid data, and hence the paleoecology of a site. Such work has been used to reconstruct summer surface water temperature for lakes over past millennia, based on the optimum ranges of cold-water versus warm-water taxa (e.g., Walker 2001).

Another animal made up of chitin is beetles (Coleoptera), which are found in freshwater and terrestrial habitats throughout the world. These insects are important food for fish, reptiles, birds, and small mammals and beetles play critical roles as pollinators and nutrient recyclers. Those Coleoptera buried and preserved in lake, pond, and river sediments are recovered as separate parts, specifically heads, pronota (thoraces), and electra (wing cases), which are identified to species based on morphological variations. One limitation to studying this proxy is that large volumes (several liters) of sediment are required to find enough remains (using a kerosene flotation and sieving) for meaningful paleoecological reconstructions, and hence only trenched excavations can provide enough material.

Where fossil beetles are abundant, they provide excellent paleoecological data as these insects are dependent on temperature for all phases of their life cycle, and each species has its own temperature range and is associated with a particular biome. Calibration sets are also created to derive more precise paleoclimate information from the fossil beetle data, following a methodology similar to that employed for midges and ostracods. For example, paleo-beetle research in central North America reveals that the area south of the Laurentide ice sheet during the last Glacial maximum, 25,000 to 18,000 years ago, had a mean July temperature that was 10–12°C lower than present, so it was essentially an arctic environment (Ashworth 2001). Subsequently, the climate became warmer so that boreal forest-adapted beetles had arrived by 14,700 years ago, which agrees well with the pollen and plant macrofossil data for this area (Yansa 2006). Sometimes conifer bark beetles are also identified which further corroborate paleovegetation reconstructions.

Another excellent paleoclimate proxy is testate amoebae (Rhizopoda), which are unicellular shelled animals (protozoa) that live on the surface of most Sphagnum acidic peat bogs. These protozoa have tests made of smooth secreted material, preformed plates, and/or cemented particles gathered from the surrounding environment and can be identified to species. Certain taxa are known to live above and others below the water
So the coring of peat bogs and analysis of testate amoebae can reconstruct times when bogs were drier and other intervals when moister conditions prevailed, thereby providing records of past water-level fluctuations. For instance, the 3500-year water table depth reconstruction of Minden Bog, Michigan, based on testate amoebae data, is remarkably similar to the water-level curve of nearby Lake Michigan inferred from dated shoreline elevations (Booth 2008). Of all of the paleoecological proxies, testate amoebae is thus the best to reconstruct past hydrological conditions, but unfortunately it is limited to peat bogs, so is largely restricted to the boreal forest and arctic biomes.

The majority of paleoecological proxies and sites that preserve fossils are found in humid environments, such as eastern North America, Europe, and parts of Asia, however, a few proxies are able to survive oxidation inherent to semiarid and arid climates and thereby provide valuable paleoecological information for these data-sparse areas. One such proxy is phytoliths, which are microfossils comprised of silica found in soils. Plants, particularly grasses, uptake silica from the soil and deposit the silica as an amorphous (opal) solution within and between cells in leaves and stems, where the silica hardens and takes on various sizes and shapes dependent on the cellular nature of different plant types. Anyone who has cut his/her hand by pulling grasses has encountered the abrasive phytoliths (“plant stones,” from Greek), which may be an adaptation by plants to reduce grazing by herbivores. After the plants decay, the phytoliths are liberated into the soil in great numbers and most remain in place. Fossil phytoliths are acquired by the trenching and sampling of soils, and heavy liquid separation of the lighter phytoliths from quartz sand in the laboratory. The study of this proxy has been mainly used to reconstruct the composition of grasslands of the past and to delineate past shifts of grassland-forest ecotones (boundaries) in high-latitude, temperate, and tropical environments.

Phytoliths are identified to morphotype and so can distinguish between tribes of grasses, offering greater precision than do pollen analysis that lumps all native grass types together. Phytolith analysis can differentiate between the Festucoids, which are cool-season C₃ grasses (where the product of photosynthesis is a sugar containing three carbon atoms), from the warm-season C₄ grasses (have a sugar containing four carbon atoms). Phytoliths can further distinguish the C₄ grasses into two types, the short-grass Chloridoids found today in the arid western Great Plains, from the Panicoids, a tall-grass prairie type native to the moister eastern Great Plains and in what is now the Corn Belt. Therefore, the study of phytoliths of buried soils can reconstruct approximate temperature conditions for the past by comparing the abundance of C₃ to the C₄ phytolith types and provide estimates for paleoprecipitation by the relative quantity of Chloridoid versus Panicoid morphotypes.

In the arid and semiarid mountainous areas of western North America, packrats (woodrats), species of the genus Neotoma, are known to collect plant material from the surrounding area to form middens (nests) within rock shelters. Over time these middens become cemented by dried urine (amberat), abandoned by the rodents, and the arid climate prevents dissolution, so that these deposits become uncontaminated “time capsules,” which can be analyzed for plant remains and ¹⁴C dating. Such research has elucidated that there was no deserts in the southwest United States during the last Glacial maximum, the peak of the last glaciation at approximately 20,000 years ago (rather an open conifer woodlands existed), and tracks the subsequent development of the desert biome in this region (Betancourt, Van Devender, and Martin 1990).
The packrat midden data are displaced as abundance classes (1–5), as concentrations, or weight percentages per plant species. The key shortcoming of the midden data is that each species of packrats prefers different kinds of plants; most collect twigs and shoots, but only some species collect seeds, fruits, acorns, and/or cacti. So the absence of certain plant species in midden records not necessarily means that these plants were not present in the landscape at the time the nests were constructed. However, the presence of certain plant species in midden records does provide important paleoecological information for a habitat where lake and peatland records are absent due to aridity. For instance, by using the modern analog concept, the macrofossil presence of saguaro cactus means “hot and dry,” pinyon pine (piñon) equates with “cool,” and mesquite indicates “wet.” Such work can also be applied to archaeology. For instance, a packrat study in Chaco Canyon, New Mexico, indicated that the area was pinyon-juniper woodland until an Anasazi occupation deforested the area between 1000 and 800 years ago, and this vegetation failed to regenerate afterwards (Betancourt, Van Devender, and Martin 1990).

The nature of packrats to collect and store food in their middens is a perfect analog for the archiving of paleoecological data, hence Neotoma is the name of a recent amalgamation of separate proxy-specific databases into one under an umbrella website (www.neotomadb.org). This website provides valuable information and some visualization tools that allow the public to become familiar with the various proxies and to learn of the valuable insights these proxies provide about past environments.

The proxy that receives the greatest attention from the public is the remains of ice age megafauna: the behemoth mammoths, mastodons, and giant sloths; the exotic American lion, saber-tooth cat, and giant beaver; and other mammalian oddities that went extinct between approximately 15,000 and 13,000 years ago, at the end of the Pleistocene. Fossil bones and teeth of over 35 genera of extinct animals are found in a range of environments south of where the ice sheets covered the northern and mid-latitudes of North America and Eurasia. Fossils of these large fauna have been recovered from sinkhole lakes, river deposits, and tar pits, the most famous site being the La Brea tar pits of Los Angeles.

No scientific consensus exists as to the cause(s) of this extinction event at the end of the Pleistocene, a time when the majority of large mammal species disappeared in North America, Eurasia, and in several other locations throughout the world. Several hypotheses have been proposed that solely or in combination can explain this event, including: climate and vegetation change; overhunting by humans (Paleoindians); diseases from humans and their dogs; and the controversial idea of an asteroid impact (Holliday and Meltzer 2010). There are several well-dated archaeologist sites that definitely document the first Native Americans having occupied the west coast and Midwest of North America south of the ice sheets by 14,700 years ago, so humans may have played a role in this extinction.

Late Pleistocene refugia and megafauna

The remainder of this section will weave together paleoecological information from the abovementioned proxies into a short summary of the landscape changes that occurred from the peak of the last glaciation to present. The focus will be on North America, but these patterns are similar to those observed in other mid- and high-latitude locations in the Northern Hemisphere. During the last glaciation, the Laurentide ice sheet expanded south from Hudson Bay and
merged with the Cordillerean ice sheet that covered the mountainous areas of western Canada and the northwestern United States. These ice sheets extended south to approximately 38° N latitude, which resulted in plants and animals being confined to a “refugia” south of these glaciers. The environment closest to the ice sheets during the last Glacial maximum (25,000 to 18,000 years ago) was periglacial, with the mean annual temperate being 10–12°C lower than today. A tundra vegetation was able to tolerate the cold winds coming off the ice sheets and low temperatures caused by the high albedo (sunlight reflectivity) of the glaciers. Occupying this frigid, open habitat were wooly mammoth (extinct), two species of muskox (one of which went extinct), and caribou (extant, living today in the arctic).

South of this tundra resided open boreal woodlands that were inhabited by the now extinct giant beaver, stag-moose, mastodon, and Jefferson mammoth and more familiar (extant) deer, elk, and moose. Immediately to the south, a mosaic of plant communities existed, including mixed coniferous-deciduous forest, open deciduous forest, and pine woodlands, which covered the southern half of the eastern United States. Meanwhile, the vegetation of the western United States south of the Cordilleran ice sheet, was drier, but still diverse, comprising a mix of grassland, temperate scrub and woodlands, and open conifer woodlands. Therefore, the vegetation of the refugia (south of the glaciers) at the end of the Pleistocene was more open and had greater habitat diversity than later during the Holocene. The mosaic of different plant communities at the end of the ice age can explain the great diversity and abundance of herbivores, especially in the central and western parts of the United States, many of which are now extinct. These herbivores included species of horses, llamas, camels, Columbian mammoths, and a larger form of bison and thus supported a greater diversity of carnivores, such as the short-faced bears, American lions, American cheetahs, saber-toothed cats, and scimitar cats.

This great mammalian diversity collapsed at the end of the Pleistocene and only a limited number of herbivores and carnivores survived. Vegetation changes can at least partly explain this event. The recession of the Cordilleran and Laurentide ice sheet permitted plants to colonize barren, water-logged glacial sediment and the first vegetation to appear was tundra. Later an extensive spruce parkland/sedge wetland covered the central and eastern portion of this formerly glaciated area which reduced the food variety for herbivores and probably contributed to population decline and the extinction of some of the larger herbivores, with many of the carnivores following suit. The transition from this boreal-like spruce vegetation to deciduous forests and farther west grasslands and deserts resulted in fewer biomes and correspondingly reduced habitat diversity that could support fewer species as compared to the Late Pleistocene mosaic. Such paleoecological information is relevant to us today given the current habitat destruction and species extinctions caused by climate change and human impacts.

In summary, a variety of plant and animal fossils can be used as proxies for reconstructing past climates and habitats and how they changed over time. Certain proxies are better than others to study particular environments and time intervals and the use of multiple proxies in an investigation provides more robust paleoecological reconstructions. Some of this research has focused on interpreting past habitat changes of Pleistocene megafauna to explain, in part, their extinction. And such information about past responses of plants and animals to environmental changes, including range shifts, can provide insights into
how organisms today and into the future will adapt to climate changes this century and beyond.

**SEE ALSO:** Biogeography; Boreal forest ecosystems; Climate change and biogeography; Ecoregions; Environmental history; Holocene; Lakes and limnology; Paleoclimatology; Quaternary geomorphology and landscapes; Quaternary glaciations; Radiometric dating/techniques; Wetland biogeography; Zoogeography

**References**


**Further reading**


Paleofloods

Victor R. Baker
University of Arizona, USA

Paleofloods are past or ancient floods that are indicated through natural recording processes. The recordings of paleofloods occur because of direct causal connections of past floods to deposition, erosion, or other markings that can subsequently be documented by knowledgeable investigators. More specifically, paleofloods are defined (Baker 1987) as floods that occurred prior to either (i) systematic measurement of flood flows using modern hydrological procedures (e.g., at stream gages or by surveys conducted after directly observed flood events), or (ii) other observation and recording by human agents (in which case the floods are termed “historical floods”). The natural archives of paleofloods extend information on the phenomenon of flooding well beyond the artificial limitations imposed by the opportunities afforded to humans for the direct observation of floods. This is especially important in regard to the largest and most intense flood phenomena, because (i) these are rare and thus less likely to be observed on human timescales; (ii) they are dangerous for direct measurement and also difficult to approach because of associated damage to infrastructure; and (iii) such floods can be so energetic as to commonly destroy the recording instrumentation at gaging stations.

The scientific study of paleofloods is associated with a long tradition in geology that was classically concerned with field evidence for ancient floods in regard to their shaping of landscapes. This concern remains important as one of the applications for paleoflood data. A scientifically interesting development in this regard is the recognition that especially immense and intense flood phenomena occurred during the terminal phases of the last ice age, that is, the late Pleistocene epoch of the geological timescale. These paleofloods are studied under the general topic of “megaflooding” (Burr et al. 2009). Although the term is somewhat loosely applied to a variety of high-discharge flood flow events, a “megaflood” can be more precisely defined as freshwater flow of 1 million cubic meters per second or greater. There are no well-documented humanly observed flood flows of this magnitude, but some Late Pleistocene outburst paleofloods from glacially dammed lakes achieved peak discharges of 10 million cubic meters per second or greater (Burr et al. 2009).

The geological study of intense flooding was highly stimulated by studies in the 1920s and 1930s of the Channeled Scabland region of the northwestern United States (Baker 2009). Professor J Harlen Bretz of the University of Chicago proposed that the distinctive landscape of this region had been eroded by spectacular megaflooding during the terminal phase of the Pleistocene. It was subsequently recognized that this flooding emanated from an immense ice-dammed lake in western Montana that held approximately 2500 cubic kilometers of water. About 16 000 or 17 000 years ago the failure of this lake, Glacial Lake Missoula, released water that was 600 meters deep at the ice dam location in northern Idaho. As the high-velocity water flows spread over the mostly basalt plateaus of eastern Washington state, they carved huge, irregular channels into the bedrock, deposited

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0104
immense bars of gravel and boulders, and created other distinctive features indicative of the past flooding magnitude. These features provide extreme examples of the natural recording process that are so important to the recognition of paleofloods.

In contrast to the megaflooding of the last ice age, the types of flooding associated with present-day hydrometeorological processes are represented in paleofloods that occurred during the last 10,000 years, the Holocene geological epoch. It is these paleofloods that are of particular interest for characterizing the hazards posed by very large, rare floods for which modern hydrological records, or even historical accounts, are inadequate.

A formal scientific approach to the study of ancient floods began to emerge in the mid-twentieth century. The term “paleoflood” first appeared in the scientific literature during the 1970s, and paleoflood hydrology was formally defined as a scientific discipline in 1982 (see Baker 2008; 2014). Paleoflood hydrology is actually a synthesis of two somewhat different scientific approaches to the study of floods. As noted, the geological approach emphasizes the recognition of various signs or traces that can be interpreted by the experienced geological investigator as evidence for past flood processes. Thus, the geological investigator of floods works out a history of past flood events in much the same way that a history of past biological organisms is worked out through geological studies of their fossil forms. This history is used to make discoveries about the nature of the flooding, including its patterns in time and space.

The hydrological approach to flood studies developed out of origins in which hydrology acted as a scientific arm for practical hydraulic engineering. The latter is concerned with the design of hydraulic structures, such as large dams. Since these structures can be at risk from extremely large floods, it is necessary to generate numeric measures of risk that can be used to achieve designs within some level of tolerance. Risk in this context is defined as the peak flood discharge multiplied by the probability for its exceedance or nonexceedance. The necessary probability estimates for risk assessment are achieved through a flood-frequency analysis of the available systematic stream-gage data. Given the lack of information on extremely large, rare floods, the prediction of flood probabilities for risk analysis must entail certain assumptions, including (i) that floods are random in time and space; (ii) that the floods at a specific place, such as at a stream-gaging station, arise from a single probability distribution; and (iii) that the mean and variance for this probability distribution are time-invariant, that is, these statistical moments are stationary in time. Modern hydrological research has shown that all these assumptions are wrong to varying degrees (Baker et al. 2002), though they continue to be applied because of the necessity to make engineering design decisions in the light of what is presumed to be a lack of data on extremely large, rare floods. Of course, this is exactly the kind of data that can be provided through the study of paleofloods.

There are many types of indicators associated with the recognition of paleofloods. For alluvial rivers, in which the beds and banks of the river are composed of the same types of sediment that the river transports, the size of the channel and its stage at overbank levels correspond to relatively low-magnitude paleoflooding that commonly recurs every year or two. This is the regime-based approach to identifying paleofloods that makes use of various empirically derived relationships to estimate the magnitudes of bankfull and other relatively high-frequency paleoflood magnitudes (for a review of this topic see Baker 2014).
More interesting evidence of high-magnitude floods can be found in their effects on trees that grow on floodplains or along valley bottoms. The stages reached by the floods produce scars on the trees that can serve as paleostage indicators (PSIs) for past flooding. Moreover, the sequence of tree rings related to the scarring events can rather precisely determine the year in which flooding occurred. Paleostages can also be inferred from the highest levels eroded by energetic flooding along the margins of channels in canyon and valley bottoms.

The boulders transported by very intense floods can be related to measures of stream competence, usually through expressions of flow velocity, bed shear stress, or stream energy expenditure (power). These can all be related through various hydraulic formulae to measures of flow depth, with the latter indicated by the preserved PSI evidence provided by the paleoflood. Additional parameters also need to be estimated, including the channel geometry, the energy gradient of the stream, and the flow resistance of the channel bed. Though estimating these parameters is subject to many uncertainties, paleocompetence calculations provide general information for approximating the magnitudes of the paleofloods that emplaced various boulder deposits.

A relatively accurate means of characterizing both the magnitudes and frequencies of paleofloods was developed in the latter part of the twentieth century based on the recognition that extreme flooding along bedrock streams conveying sandy sediments will deposit well-preserved sequences of those sediments in appropriate, protected riverine settings. In these environments, during especially intense flooding, sedimentary particles with sufficiently high settling velocities accumulate relatively rapidly from suspension in areas of flow separation and slack water, such as at the mouths of back-flooded tributaries or in bedrock alcoves set into canyon walls. The resulting slackwater deposits (SWDs) commonly comprise thick accumulations (Figure 1), in which individual sandy layers correspond to the causative flood events and layers of nonflood materials or boundaries between the SWD layers correspond to intervals between the flood events. The nature of the preservation sites for the SWDs is such that they are protected from the erosive action of later floods. The well-preserved evidence provided from sequences of slackwater deposits (SWDs) is combined with other paleostage indicators (PSIs) to comprise the mode of investigation that has come to be known as SWD–PSI paleoflood hydrology (Baker 1987).

Several technological developments contribute to the greatly improved accuracy of SWD–PSI studies for estimating both the magnitude and frequency of paleofloods. Frequency estimation requires understanding of the timescales in which flooding occurs, and there are now many new analytical tools for assigning timescales to sequences of paleofloods. The tool that was first applied to SWD–PSI paleoflood hydrology in its initial phases of development was the radiocarbon-dating of organic material intercalated with the flood deposits. This method is applicable to paleofloods that occur throughout the Holocene epoch. Recent developments in tandem accelerator mass spectrometry now allow organic grains with a mass of only a few milligrams to be dated within an error range of several decades.

Of course, many flood deposits do not contain organic material. Recent developments in the geochronological tool of luminescence dating are filling this void by making possible the direct dating of mineral grains of quartz and feldspar that were transported by paleofloods. Another new geochronological tool, terrestrial cosmogenic nuclide dating, makes possible the direct dating of flood-eroded rock surfaces and flood-transported boulders.
PALEOFLOODS

Figure 1 A sequence of sandy paleoflood slackwater deposits (SWDs) exposed in a bedrock alcove developed in cliffs along the margins of the Escalante River, a tributary of the Colorado River in south central Utah. Note the multiple layers indicating deposition by separate flooding events. (Photograph by the author.)

The calculation of paleoflood discharges was originally accomplished using simple hydraulic formulae, such as the Manning equation, which relate the discharge of the paleoflood to its cross-sectional area, flow depth, energy slope, and a measure of flow resistance. Starting in the 1980s, computer flow models began to be employed to perform the hydraulic calculations. The first models performed step-backwater calculations in one dimension, along the thread of the main channel. Multiple cross-sections are used along a channel reach that is long enough to achieve an energy-balanced calculation of multiple water-surface elevations for various potential paleoflood discharges. By matching the flood paleostage evidence to the calculated water-surface profiles, paleodischarges are estimated. More recently, further increases in computational capability have enabled the use of two-dimensional flow models. These are appropriate for the more complex channel geometries that can be accurately represented in the one-dimensional models.

A variant of PSI paleoflood hydrology was developed in the 1990s by the US Bureau of Reclamation for the study of potential risks posed to large dams by very rare, extreme flooding. This methodology identifies the various threshold indicators of paleoflood nonexceedence. It then calculates the paleodischarge values for these thresholds, uses geochronology to determine the timescales over which the nonexceedence thresholds apply, and finally puts the discharge values and their timescales into a flood-frequency analytical scheme that determines the likelihoods of nonexceedance.

In contrast to these applications by the US Bureau of Reclamation, much of the US engineering community has been relatively slow to embrace paleoflood hydrology in its operational approaches to the evaluation of extreme flooding hazards. One reason for this seems to be reliance on engineering rather than scientific approaches to the phenomenon of flooding. Some of the engineering literature continues to repeat criticisms of paleoflood hydrology that were expressed during the early development of that science in the 1970s. As reviewed by Baker, Webb, and House (2002), those 1970s concerns included (i) the impossibility of determining the ages for very ancient floods; (ii) the impossibility of accurately determining the discharges for
ancient floods; (iii) the lack of appropriate statistical procedures for incorporating paleoflood data into risk analyses; and (iv) the presumption that future changes in climate and/or land-use practice will invalidate the incorporation of information on past floods into the statistical analyses that determine the risk of future floods. As noted, concerns (i) and (ii) have been met by new technology for geochronology and for hydraulic modeling. Concern (iii) has been met by developing the appropriate statistical procedures, as exemplified by the US Bureau of Reclamation approach. Finally, concern (iv) is merely a kind of skepticism that applies to all historical information. Its logic would hold that all empirical information about floods, including that derived from stream gages, is invalidated by the reality of the future not being exactly like the past.

Though the modern science of paleoflood hydrology was developed in the United States, some of its more extensive applications have been taken up in foreign countries. SWD–PSI paleoflood hydrology spread to Australia in the 1980s, and to India and Israel in the 1990s. These countries all have many areas that are highly appropriate for the SWD–PSI methodology. Extensive paleoflood investigations in the 1990s and the new millennium have also been conducted in southern Europe, mainly in Spain, France, and Greece; in southern Africa, both in Namibia and in the Republic of South Africa; in South America (Chile and Peru); and in parts of Asia, including China, Japan, and Thailand. A recent survey (Baker 2013) documents 60 paleoflood study areas from 21 US states and 25 other countries. High potential for very accurate applications of SWD–PSI paleoflood hydrology occurs in other parts of Africa (e.g., Algeria, Libya, Somalia), Asia (e.g., Turkey, Iran, Mongolia, Sri Lanka), Europe (e.g., Russia, Romania, Portugal), and South America (e.g., Argentina, Brazil, Venezuela).

Very interesting paleoflood hydrological research is currently being done in China. As a result of war and turmoil in the mid-twentieth century, China did not have the infrastructure of stream gages to provide the data necessary to hydrological design decisions for the large dam projects of recent decades. During the period immediately after World War II China made extensive use of historical accounts of river flooding that were documented over the past 2000 years by the bureaucracies of the various Chinese emperors. These included actual marks and inscriptions on canyon walls showing the stages reached by ancient floods. Moreover, largely independent of the work on paleofloods in the United States and other Western countries, some Chinese investigators in the 1980s were supplementing the historical accounts by field studies of slackwater sediments along major rivers, such as the Huang He (Yellow River) and Chang Jiang (Yangtze River). In more recent years Chinese investigators have fully embraced SWD–PSI methodologies (Baker 1987), using them to study the risks posed by flooding to major dam projects, water-transfer schemes, and the hazards posed to very dense human populations along river bottomlands.

Future global climate change is widely recognized to be a likely consequence of well-documented anthropogenic increases in radiatively active atmospheric gases, particularly carbon dioxide. There are sound physical reasons to expect that the increases in global mean temperature associated with these atmospheric changes will result in more intense flood phenomena. In essence, the increased radiative forcing associated with increased tropospheric carbon dioxide, methane, and other gases should lead to greater surface heating, which will increase air temperatures and evaporation. Much of this effect will occur in the tropics, probably enhancing tropical flood-generating phenomena and
also promoting more atmospheric water transport from the tropics to higher latitudes. This will, in turn, lead to more atmospheric water vapor at those latitudes, which will contribute to areas of more intense precipitation and increased storm intensities. Recent examples of extreme flooding phenomena in parts of the United States have led to concern that this phenomenon may be a consequence of global warming.

Because conventional engineering approaches to flood-risk analysis rely on flood-frequency procedures, the expected changes in flood magnitudes from global climate change pose a major concern for the evaluation of future flood risk. Clearly such changes violate the stationarity assumption noted earlier as underpinning engineering design decisions that are based on mathematical extrapolations from flood-frequency distributions. Thus, it may not be inappropriate to ask whether this entire engineering approach to estimating extreme flooding phenomena is flawed scientifically. As an alternative, it is clear that the largest paleo-floods are very real manifestations of the most extreme kinds of flooding that can be produced by storm systems. Global warming will produce climate change, that is, it will result in shifts of storm patterns. Global warming is not going to create entirely new kinds of meteorological phenomena that have never before occurred on the planet. Evidence for the kinds of extreme floods likely to be associated with the new storm patterns is available from the natural archives of paleo-floods. To ignore these archives of SWD–PSI and other paleo-flood evidence would be the anti-intellectual equivalent of ignoring the lessons of human history when facing the expected uncertainties of future human affairs.

As a final observation, paleo-floods have the potential to play a unique role in regard to the communication of flood risk to the general public. Such use sharply contrasts with prevailing authoritative government pronouncements regarding the so-called hundred-year flood. The latter is a by-product of the engineering flood-frequency approach to flood-risk assessment described earlier in the entry. Unfortunately, the “hundred-year flood” designation is nearly universally misunderstood. It does not refer to an actual flood, and it has nothing to do with real years. Instead, it is the inverse of annual exceedance probabilities derived from mathematical extrapolations that are based on questionable assumptions. In contrast, paleo-flood information derives its authority from the natural recording of ancient (but very real) cataclysmic processes with the obviously documented potential to cause harm. The commonsense recognition that what has actually happened will very likely happen again has much more potential to provoke engaged and wise public response than the invocation of abstract terminology that befuddles rather than informs.

**SEE ALSO:** Dendrogeomorphology; Environmental risk analysis; Fluvial depositional processes and landforms; Fluvial erosional processes and landforms; Geomorphic hazards; Rivers and streams

**References**


Paleosols

B. Brandon Curry
University of Illinois at Urbana–Champaign, USA

A Paleosol is the evidence of an ancient soil, either at ground surface (a “relict soil”), or buried at depth by sediment (a “buried soil”). Relict soils are evident at places where the surface soil characteristics are incongruous with the prevailing environment, for example, a deeply weathered soil (implying abundant moisture) in an arid climate. An “exhumed soil” is a special case in which sheet erosion brings a buried soil horizon to the surface. Exhumed soils are especially common in semiarid regions. Here, mesas may be capped by thick, resistant carbonate horizons (hardpans, caliches, or calcretes) that originally formed at depth in the past. In areas with episodic sedimentation, such as on floodplains and in glaciated regions, Paleosols constitute marker beds that allow differentiation and correlation of past soil-forming versus sedimentation/burial events.

Most studies of Paleosols require (or prefer) that the age of the soil be known. If it is not known, determining the age of the soil is often a key part of its study. Soil “age” in this context can be defined in a number of different ways. The period of time during which soil actually formed, from its period of inception (commonly called timezero) to its burial, is called a soil-forming interval. Most soil-forming intervals represent times of geomorphic stability when the surface was neither eroding nor getting buried, hence, a soil could form. If the age of the soil-forming interval is known, then the characteristics of the Paleosol can be used to interpret what the environment was like during that period of time. Another key aspect to soils, with respect to age or time, is establishing the date of timezero — exactly when the surface stabilized and the soil started to form. And lastly, one often wants to determine when a buried soil was actually buried. This date (the end of the soil-forming interval) helps define when in the geologic past conditions became unstable, locally or regionally, causing sediment to be deposited on the soil, effectively burying it.

The upper surface of a buried Paleosol is a non-conformity, representing a time when little or no sediment was deposited and the surface was geomorphically stable. However, in many pedogenic environments surface soils are said to be cumulic, that is, the geomorphic surface slowly incorporated new material, gradually grew upward, and thickened as a result. Thus, an important attribute of Paleosols is the time span represented by the nonconformity, that is, the soil-forming interval, or at the very least, the slowdown in sediment accumulation rates. These temporal or chronometric issues may be addressed by various dating techniques, particularly by methods such as radiocarbon or luminescence dating. The stratigraphic and time elements of Paleosols are encapsulated in the fields of pedostratigraphy and allostratigraphy (American Commission on Stratigraphic Nomenclature 2005).

Because Paleosols possess attributes that reflect past climate and environmental changes associated with past climates, they serve as important paleoenvironmental proxies. Indeed, the first field geologists who studied stream cuts and other
outcrops in the upper Midwest of North America used Paleosols to infer that Earth’s recent climate had shifted over time between “glacial” and “interglacial” periods (Leverett 1899).

Paleopedology is the study of Paleosols. It is closely allied with pedology, the study of modern soils. Many paleopedologists have couched their interpretations of Paleosols in the context of modern soil surveys (i.e., Retallack 2001). Soil surveys provide soil maps and descriptions of key soil horizons and other features, using an established classification system. The United States Department of Agriculture’s latest effort, *Keys to Soil Taxonomy* (Soil Survey Staff 2010) classifies soils into 12 orders with diagnostic surficial layers (such as the mollic epipedon, a diagnostic horizon for Mollisols, which are characteristic of the prairies) and subsurface soil horizons (such as the argillic horizon, a diagnostic horizon for Alfisols, which are characteristic of temperate forested ecosystems), as well as other compositional and climatic criteria. The International Union of Soil Sciences system includes additional classes of soils not found in the United States, such as anthropic soils with very high phosphorus content due to centuries of human occupation (IUSS 2006). Key soil horizons, also observed in Paleosols, are rigorously defined in these systems. In general, the morphological features of all soils become coarser with depth, which is a characteristic for identifying a soil profile. This type of depth trend is normally absent from geologic strata.

**Soil-forming factors**

At many scales, Paleosols exhibit predictable variability. The vertical and spatial variability observed in Paleosols is perhaps best interpreted through Jenny’s (1941) state factor concept, in which soil (and Paleosol) characteristics are viewed as a function of five soil-forming factors: climate (cl), organisms (o), relief (r), parent material (p), and time (t). For buried and exhumed Paleosols, one could add the d factor – diagenesis, which includes postburial changes in the physical and chemical characteristics of Paleosols, such as rubification (reddening) of fossil B horizons due to oxidation of iron (typically a bacterially mediated process) (Retallack 2001). Diagenesis includes compaction, which can dramatically affect soil structure (Follmer *et al.* 1979; Follmer 1998). In leaching regimes, diagenesis also commonly includes the recarbonation of buried, acidic Paleosols, as percolating water moving through overlying, calcareous sediment adds carbonates to the buried soil. A theme that quickly emerges is that the state factors are intertwined; one cannot fully assess the impact of one factor on Paleosol development without examining its effects on the others.

**Organisms**

The role of organisms in soil development is complex. Most soils form in equilibrium with the current vegetation of the site, allowing interpretations of soil paleoenvironments from soil morphology and chemistry. For example, Paleosols with thick, dark A horizons likely formed under grass vegetation. Paleofauna impacted Paleosol development in ways that are readily observable (such as burrows of ants, worms, wasps, cicadas, crayfish, and gophers) (Johnson 1989; Follmer 1983; Retallack 2001). In addition, micro-organisms play vital roles in pedogenesis, and they remain active after burial (Driese and Nordt 2013). Most diagenetic reactions are mediated by micro-organisms, primarily bacteria. The once-living components of the soil, such as plant roots or burrowing ants, were affected by the prevailing climate and other factors (such as parent material porosity) which impacted critical conditions such as the depth...
and degree of saturation during the growing season, and depth of frost.

Parent material

Parent material is the sediment in which the soil or Paleosol has developed. Careful studies have revealed that, rather than a static entity, soil parent materials are dynamic through additions of new material of eolian or alluvial origin (or both) that can become mixed into the upper parts of the soil (solum). Consider a soil on a low-lying river terrace that receives increments of fine sediment from floods. In this scenario, thin layers of new sediment may be rapidly incorporated in the soil through mixing agents such as worms or ants. A less obvious example would be a soil which had been undergoing additions of eolian dust to its surface, which eventually become incorporated into the profile. Significant dust sources include outwash surfaces in glaciated areas, dry lake beds, and volcanic ash. The nature of such eolian additions may be revealed through careful textural and chemical analyses, which may show enrichment of certain constituents in the upper solum that may not be present in relatively unaltered material at depth. For example, some Paleosols in the deserts of North America have meters-thick calcrete (Bkkm) horizons, and yet the unaltered parent material may not contain any carbonate. Dust traps installed in such areas have revealed that the primary carbonate source was carbonate-rich eolian dust. In many areas, groundwater is also a source of carbonate. Recent studies have shown that dust literally travels across the globe; North African dust carried on easterly trade winds reaches Florida, Barbados, and Amazonia (Muhs et al. 2007). Incorporation of dust into soils also affects their physical properties, especially if the original parent material is coarse grained. Soil chemistry and fertility are also altered and diversified, because eolian dust contains minerals rich in carbonate, iron, and phosphorus.

Relief

Not only do Paleosols exhibit changes in the vertical (“z”) direction, but they show much variation along former slopes in the x-y (lateral) direction, producing a characteristic pattern of soils across the paleolandscape. This pattern of slope-dependent, physical and chemical variation in a sequence of soils is referred to as a catena, or toposequence. Along catenas, certain types of soil features or profiles tend to develop in certain slope positions (summits, shoulder-, back-, foot-, and toeslopes). For example, cumulic (overthickened A horizons) occur in areas of deposition at the bases of some slopes, for example, footslopes or toeslopes, and erosion is common on steeply sloping shoulder slopes and backslopes. An example of variable Paleosol characteristics along a catena might include those on summit positions that have reddish hues and well-defined soil horizonation, but which transition to poorly drained, gray (gleyed) weathering profiles in the footslope or lower positions (Follmer et al. 1979). The redness versus grayness of the soil in the previous example is related to soil aeration, which is a function of water content. Soil water content, in turn, impacts the soil flora (organisms) that mediate chemical conditions and valence of iron-bearing minerals. Slope-dependent soil moisture characteristics also impact terrestrial vegetation and hence the chemistry of the soil horizons (Follmer 1983; Retallack 2001). Of the five soil-forming factors, paleoslope and paleorelief require the least interpretation. That is, determining past changes in effective moisture, temperature, soil biota, terrestrial vegetation, and its variability over time, for Paleosols, may often be inferential. If one is fortunate, paleorelief may even be fully revealed in outcrop.
Climate

Seasonally effective moisture and temperature are parameters that control rates of leaching and chemical transformations in soils, and also dramatically affect the composition of flora and fauna living in and on the soil. Due to natural climatic variability, interpretation of past climates is more inferential and complex the longer the time span associated with development of the Paleosol. Even under ideal circumstances, specific climatic conditions inferred from Paleosols (such as mean annual temperature or mean annual precipitation) are, at best, generalized estimates (Retallack 2001).

Regionally significant and well-developed Quaternary Paleosols typically have formed over a relatively long period of time, for example, over 5000–75,000-year time spans or more (Follmer 1983). During this time span of soil formation, climatic characteristics may have changed significantly, especially for the longer time spans. A traditional school of thought suggests that most pedogenic chemical work is done under optimal soil-forming intervals (i.e., climatic “optima”), that is, a synergy of moisture regime and temperature resulted in high physical biotic activity (burrowing, homogenizing soil material), biochemical production (organic acids), and physiochemical transformations (mineral weathering) to produce a well-developed soil. Evidence for accelerated soil formation in some areas, however, has been attributed to rapid rates of dust incorporation, rather than to increased weathering rates under an “optimal” climate. Nonoptimal conditions for pedogenesis in the geologic past may have been “too dry,” “too cold,” or “too hot,” resulting in slower rates of soil formation. Conditions that are “too wet” lead to waterlogged conditions or erosion, and little profile development. Another school of thought suggests that overall rates of pedogenesis are more or less steady over time. In this scenario, under changing environmental conditions a waning pedogenic process is replaced by another (Birkeland 1999).

Time

Time (soil age) is of considerable interest in paleopedology, in both applied research and stratigraphic studies (including geoarcheology). For applied studies, the age of tectonically displaced Paleosol horizons may be critical in determining the capability or periodicity of seismic activity of a particular fault or fault zone (McCalpin 2009). The chronologic age of the Paleosol’s parent materials, however, can raise several complex issues, and the paleopedologists must specify the context of the dated fraction. Minimum ages for \( t_{\text{zero}} \) may be provided by radiocarbon ages of in situ tree stumps (rooted in the soil, and subsequently buried by a deposit not related to the soil). Alternatively, optically stimulated luminescence (OSL) ages of eolian material above the soil may help to isolate the date of burial and the end of soil formation. Relative to the stump age, older and less precise ages may be obtained from radiocarbon dating of disseminated organic material (humus) in A and O horizons (organic-rich surficial soil horizons) which reflect the mass balance of additions of “new” organic matter with the oxidation or consumption of “old” organic matter. These ages will conceptually lie between \( t_{\text{zero}} \) of soil formation, which often is equivalent to the age of the parent material, and the soil’s minimum-limiting age (provided, for example, by the stump). Some workers provide ages of the whole soil or organic fractions that are separated either chemically (into organic fulvic and humic acid groups) or thermally into volatile and more stable fractions (Wang et al. 2003). The longer the time span during which the soil formed, the larger the potential error associated with the estimated soil age.
Polygenesis

Polygenesis refers to soils that have developed under more than one set of soil-forming conditions, usually as associated with the five soil-forming factors. In contrast, monogenesis implies soil formation under nearly unwavering soil-forming conditions. Although polygenesis can imply changes in parent material (e.g., through additions of dust) or relief (through erosion), it typically refers to changes in vegetation and/or climate. Normally, polygenesis is complex and involves multiple factors, such as a local erosion event that results in changes in slope, parent material, and possibly organisms. Paleosols are more likely to have become polygenetic if they have formed over long time spans, that is, long enough to involve periods of significant climate or environmental change. For example, the Sangamon Geosol in midwestern North America formed in most places over a period of about 50,000–75,000 years during the last interglaciation (Curry and Follmer 1992). This soil shows morphological evidence of polygenesis, but overall its features indicate development under a temperate, humid climate, similar to today’s climate (Follmer et al. 1979; Follmer 1983). However, coeval pollen and ostracode records from lake deposits in the “type” region of the Sangamon Geosol show that climate was considerably more variable than what is implied by soil morphology (Curry and Baker 2000). The upshot is that the most significant and readily observable features of polygenetic Paleosols are ones that reflect periods when the soil environment was conducive to weathering and biotic activity, namely, warmer, moister intervals. Paleosols with shorter soil-forming intervals typically are best preserved in areas where there has been episodic sedimentation, as revealed by proxy data that have more detailed resolution, such as rhythmic beds of thin A-O horizons separated by thin increments of loess.

As noted above, diagenesis refers to changes to the original soil’s physical and chemical makeup that occur during, but mainly after, burial. The most important diagenetic changes in Paleosols include increased density (through loss of pores via compaction, and decay of roots and other organic masses such as fungi), changes in wetness due most commonly to perched water above the buried soil, and chemical interactions with groundwater. The net effect of diagenesis is that the material in the buried soil slowly loses its original pedogenic character. Commonly, sediment cores will more readily reveal subtle paleostructures and fabrics that have been affected by diagenesis after the material has been allowed to be exposed to air, but not completely dry. Virtually all interpretations of features observed in Paleosols are impacted by equifinality, that is, similar features may have been formed by different processes. Equifinality can imply that two Paleosols that were quite different when they were forming became increasingly alike after burial and diagenesis affected their morphology and chemistry. The effects of diagenesis and equifinality require that paleopedologists constantly entertain alternative hypotheses when interpreting a Paleosol. For example, many features thought to be purely abiotic, such as soil reddening, may be initiated or stimulated by biotic processes, particularly those of bacteria.

In the study of buried soils, the description of “klumpen” has been added to the traditional United States Department of Agriculture methods (Schoeneberger et al. 2012) to focus on hand sample-size features ranging in size from 0.02 to 2.0 mm in diameter (Follmer 1998). The most valuable klumpen features for identifying buried A and E horizons are pores, aggregates, and organic materials, especially charcoal. These features aid in the identification of the upper horizons of buried soils; here, coarser features and organic matter are often lost. In buried soils,
due to diagenesis, the original granularity of surface horizons is usually compressed into a massive aggregate nature that breaks into irregular clumps and granules upon drying (“klumpen”).

Geosols

A buried Paleosol with regionally consistent stratigraphic relationships among its parent materials and its superjacent, younger geologic material may classify as a Geosol. Geosols are the basis of pedostratigraphy, one of several independent stratigraphic systems used to describe and interpret layers of Earth materials. Other related but independent classifications include chronostratigraphy (time stratigraphy) and lithostratigraphy (material stratigraphy) (American Commission on Stratigraphic Nomenclature 2005). Pedostratigraphy is related to allostratigraphy, which is a stratigraphic system based on unconformities (paleosurfaces) and their cross-cutting relationships.

In the context of the Sangamon Geosol, we note the Indian Point and Farmdale Geosols which formed in advance of, or adjacent to, the Lake Michigan Lobe over periods of time ranging from hundreds to about 5000 years.

Figure 1  Poorly drained pedofacies in the Athens Quarry, Athens, Illinois. Moss and wood fragments impart the brown color to the organic-rich Peoria Silt (loess). The Farmdale Geosol is developed in colluviated loess of the Robein Member (Roxana Silt), and the gleyed Sangamon Geosol is developed in the Vandalia Member of the Glasford Formation. Curry and Follmer 1992. © 1992 University of Illinois Board of Trustees. All rights reserved. Used courtesy of the Illinois State Geological Survey. Photography by B. Brandon Curry.
Figure 2  A, E, B, and C horizons of the polygenic Sangamon Geosol capped by Wisconsin Episode loess at the Sisters Section near Peoria, Illinois. McKay et al. 2005. © 2005 University of Illinois Board of Trustees. All rights reserved. Used courtesy of the Illinois State Geological Survey. Photography by Richard C. Berg.

(Curry and Follmer 1992; Figures 1 and 2). The parent material for these Geosols is colluviated and locally cryoturbated loess and alluvium. As the colluvial parent material thins away from the source areas (e.g., river valleys), so does the evidence of separate pedostratigraphic entities.

SEE ALSO: Paleoclimatology; Paleoecology; Soils in geomorphic research; Soils as relative-age dating tools

References


PALEOSOLS


Parallel computing

Shashi Shekhar  
*University of Minnesota, USA*

Daniel Cintra Cugler  
*University of Campinas, Brazil*

Parallel computing is a computational technique in which multiple operations are executed concurrently rather than sequentially. It is a broad area of computer science that encompasses processor technologies and computer communication infrastructures in order to provide high performance and redundancy in high-availability systems. Parallel computing has become essential for managing and processing the massive volumes of data generated or acquired by different technologies, such as smartphones (e.g., location-aware applications), satellites (e.g., remote sensing), sensors (e.g., temperature, humidity, and seismic sensors), and social media platforms (e.g., Twitter). Vast amounts of these data have spatial features that make it inherently challenging to manage. Unlike sequential computing, which relies on one central processing unit (CPU), parallel computing solves complex problems quickly by harnessing the power of multiple CPUs simultaneously. Computations that in the past took days, months, or even years can be done in minutes, seconds, or even milliseconds using parallel computing.

Parallel computing plays an important role in geography (e.g., processing spatial information for weather forecasting and climate change prediction), as well as in many other fields, including astrophysics (e.g., computing galaxy formation), electrical engineering (e.g., electromagnetic field computation), biology (e.g., genome mapping), and physics (e.g., hydrodynamics). Parallel computing that specializes in spatial information is known as parallel GIS (parallel geographic information systems). In order to exploit the full potential of parallel GIS, technologies need to be developed that focus on solving challenges surrounding spatial data, such as methods for parallelizing GIS (Shekhar et al. 1996; 1998). The evolution of parallel computing can be classified into two eras: pre-cloud computing and cloud computing.

**Parallel computing in the pre-cloud era**

Single, multicore, multiprocessor, and special purpose processors

One of the most important parts of a computer is the central processing unit. The CPU, also known as the core, is responsible for performing arithmetic and logical calculations. The main technologies adopted for processors are the single processor, the multicore processor, the multiprocessor, and the special purpose processor. In the past, most computers were equipped with a single processor (with one core), which could perform only one operation at a time (sequential computing). The way to increase processing speed in such processors was to increase their clock frequency. However, the fastest speeds consumed too much power. This problem hampered the creation of computer processors faster than 3 GHz (Hey, Tansley, and Tolle 2009). The cheapest solution adopted by the main computer processor manufacturers was to execute tasks in parallel, by replacing the single-core processor with a multiprocessor. Modern processors have from two to four cores. The clock speed of these processors has reached 4 GHz, and the number of transistors in them has reached hundreds of millions, making them highly versatile and powerful.
PARALLEL COMPUTING

with a collection of processors on the same chip, known as a multicore processor (Asanovic et al. 2009). A multicore processor has two or more cores that enable the parallel execution of computer operations. The performance of such processors is relative to the number of their cores and the clock frequency. Multicore processors are responsible for making parallel computing more popular in the past few years. Figures 1 and 2 illustrate the core of a single processor and a multicore processor.

The computational power provided by multicore processors is limited by the number of cores that could be inserted in one processor. One solution to this problem is to provide computers with two or more processors (multiprocessor). In a multiprocessor computer, computations are performed not only in different processors but also within the different cores of each processor. The evolution of such kinds of processors settled a number of architectural issues in computer science, such as the definition of memory access, for example, shared and distributed (Asanovic et al. 2009). Figure 3 illustrates the processors and their cores in a multiprocessor computer.

Processors can be designed to execute general or specific tasks. General-purpose processors (GPP) are widely used in desktops, laptops, and server computers. They treat a large class of problems but can lack high performance when used to execute specific applications. For these cases there are special-purpose processors (SPP), also known as application-specific processors (ASP). SPPs combine
low cost, low power consumption, and high performance for devices such as smartphones, LCD monitors, digital cameras, and wireless routers. A well-known example of an SPP is the graphics processing unit (GPU), a kind of processor that specializes in efficiently generating graphic displays. Nowadays, some GPUs also enable software developers to add support for GPU acceleration in their own applications. Such kinds of GPUs are called *general-purpose graphics processing units* (GPGPU). GPGPUs provide parallel computing platforms and programming languages (e.g., NVIDIA CUDA) in which software developers develop general purpose applications which are traditionally handled by CPUs.

**Clusters and grids**

Clusters and grids are another kind of technology that use parallelism to provide high-performance computing and redundancy in high-availability systems. A computer cluster is a set of linked computers working together that behave as a large single computer. The machines have similar hardware and are usually connected to each other through a local area network (Kumar et al. 1994). The Titan supercomputer is a well-known example.

A grid is also a set of linked computers working together and behaving as one large computer. Unlike clusters, however, the computer nodes in a grid tend to be heterogeneous and geographically dispersed (e.g., SETI@home project). This means that the grid’s nodes are connected through some kind of wide area network (WAN), such as the Internet. Grids may be made up of machines with very different hardware configurations. Normally, users have little or no knowledge of where the computer nodes are located or which kind of hardware and operational systems have been used (Jacob et al. 2005). A cluster is tightly coupled, whereas a grid is loosely coupled.

Besides providing high-performance computing, clusters and grids also support high-availability computation by providing redundant nodes that can assume computational duties when other computer nodes fail. In the past, grids and clusters were available only to governments and big companies that could afford the high financial cost of their design and upkeep.

**Software aspects**

Parallel computing hardware alone does not guarantee significant gains in computing power. Harnessing the power of different kinds of processors, grids, and clusters requires computer programs that are developed specifically for each parallel environment. Algorithms are needed to divide complex problems into simpler and smaller parts so that they can be solved concurrently. Existing serial computer programs need to be reimplemented in the parallel format. All these processes can be time-consuming, and different tools have been created to support them. Examples of programming languages that support parallel programming are C*, SequenceL, CUDA, OpenCL, Unified Parallel C (UPC), Parallel Fortran, OpenMP, and MPI.

Parallel programming languages have more complex data and control models than serial programming languages. The data model defines how processors communicate with each other. The control model determines how processors are coordinated. Usually programmers do not need to know such features, since high-level programming languages encapsulate this information.

In the pre-cloud computing era, high-performance environments were usually developed to provide a small number of computational
services for a specific domain. Their capacity to provide a broad range of services was limited. With the advent of cloud computing, a broad range of general-purpose computational services has become available to the wider public, for example, Microsoft Office web apps and Google Calendar.

Cloud computing era

Cloud computing, which is intimately linked to parallel computing, enables ubiquitous and on-demand network access to a shared pool of configurable computing resources, such as networks, servers, storage, applications, and services (Mell and Grance 2011). Cloud computing enables groups of users to host, process, and analyze large volumes of multidisciplinary data (Hey, Tansley, and Tolle 2009). Although the cloud approach does not necessarily speed up the computational power provided by parallel computing, it has revolutionized parallel computing, providing access to inexpensive high-performance systems and many new remote services.

According to the American National Institute of Standards and Technology (NIST), these services fall into three main categories. (i) Software as a service (SaaS) is the most widely used kind of cloud computing service. It provides sophisticated applications that are accessed and executed through the Internet, usually using web browsers (Yang et al. 2011). Google apps, Salesforce CRM, GoToMeeting, and Workday are typical examples. (ii) Platform as a service (PaaS) is a cloud-based platform service for the development of new applications. PaaS provides hardware and software infrastructure, thus freeing users from the need to maintain server hardware and operating systems. Software or services developed to run over cloud PaaS can be executed entirely on the Internet. Examples include Google App Engine and Microsoft Azure (Yang et al. 2011). (iii) Infrastructure as a service (IaaS) provides computer infrastructure, such as processing, storage, networks, and other fundamental computing resources as virtualized services (Yang et al. 2011). Users cannot control the cloud infrastructure but they can control the operating systems, storage, and deployed applications (Mell and Grance 2011). A widely known example of IaaS is the Amazon Elastic Compute Cloud (EC2).

According to NIST, all categories of cloud computing services share five essential features: (i) on-demand self-service, where consumers can reserve more computing capability (e.g., server time) without the need for human interaction with the service provider; (ii) broad network access, where services are provided for different kinds of client platforms (e.g., mobile phones and laptops); (iii) pooling of resources, where a variety of resources are provided as a single pooled resource; (iv) rapid elasticity, where capabilities can be rapidly and elastically supplied, making them appear to be unlimited and available for purchase in any quantity, at any time; and (v) measured service, where resource usage can be monitored, controlled, and reported, providing transparency for both the provider and the consumer. These five features distinguish cloud computing from other kinds of parallel paradigms, such as grid computing (Yang et al. 2011).

As cloud computing has become cheaper and more popular over time, access to high-performance environments is no longer limited to large organizations with deep pockets. Pay-as-you-go, in which users pay only for the services they use, is a good example of a pricing model that is used for selling cloud computing services. Moreover, many cloud-based services, such as Google Docs and Dropbox, are provided at no cost to end users. All of these benefits have been made possible thanks to the parallel computing infrastructure provided by multiprocessors.
(with multicores) working in large clusters, as well as by new highly efficient programming models and associated implementations that are able to orchestrate large numbers of parallel tasks within clusters.

MapReduce

MapReduce is a popular programming model used for processing and generating large datasets in cloud environments. Created at Google in 2004, it was inspired by the map and reduce primitives present in Lisp and other functional programming languages. Software for MapReduce platforms is designed to divide a problem into many small parts, each of which may be executed on any node in a cluster. The core of MapReduce is the map and reduce functions, which parallelize computation across large-scale clusters of machines, handle machine failures, and schedule intermachine communication to make efficient use of the network and disks. These features allow programmers with little or no experience with parallel and distributed systems to utilize the resources of a large distributed system and process enormous amounts of data (Dean and Ghemawat 2008).

A popular free and open source implementation of MapReduce is the Apache Hadoop framework. It provides an environment in which programmers do not have to deal with issues of parallelization, remote execution, data distribution, load balancing, or fault tolerance. Hadoop libraries have been created for programming languages such as Java, Perl, Python, and PHP. Apache Hadoop also provides the Hadoop Distributed File System (HDFS), which enables high aggregate bandwidth across the cluster as well as reliable storage of very large files across machines in a large cluster. Both Hadoop MapReduce and HDFS are designed so that node failures are automatically handled by the framework.

Spatial problems such as spatial autoregression (SAR) or spatial graph algorithms (e.g., breadth-first search and shortest path) challenge MapReduce because these algorithms are iterative. Although processing one iteration is parallelizable, the synchronization overhead across iterations for cloud environments may be too large to maintain speedups. Recently, approaches such as Spark (Zaharia et al. 2010) have been proposed to address iterative computations that are popular in spatial problems. Since Spark is memory-based, it may have a cheaper “reduce” step, leading to speedups for iterative spatial problems.

Spatial cloud computing and parallel GIS

Parallel computing specialized to spatial applications is known as parallel GIS. Spatial data have long been recognized as presenting unique challenges (Shekhar et al. 1996; 1998). Today, spatial data volumes are exceeding the capacity of commonly used spatial computing and spatial database technologies to manage and process data with reasonable effort. Such spatial data are known as spatial big data. Current tools supporting the common activities in data-intensive processing are often limited in the face of challenges posed by emerging big spatial datasets, due to the volume (e.g., petabytes), acquisition/search demands (e.g., terabytes per second) and variety (e.g., maps, GPS tracks, sensor measurements). Spatial big data require new advances in GIS.

Spatial cloud computing (SC2) adds parallel GIS services and technologies to cloud computing. Similar to cloud computing, SC2 encompasses the entire cyberinfrastructure, including hardware, software, and networking. Spatial cloud computing can also be optimized with spatiotemporal principles to make efficient use of available distributed computing resources.
There are an increasing number of software tools to manage spatial big data in SC2 platforms. For example, SpatialHadoop is a MapReduce framework, developed at the University of Minnesota, designed especially to work with spatial data. It can be used to analyze large spatial datasets on a cluster of machines. Another example is Google Earth, a virtual globe that uses a large cloud computing infrastructure and the BigTable data structure. BigTable is a data storage system developed by Google which provides compression and high performance for large datasets.

Large spatial data repositories can be accessed using data as a service (DaaS). Spatial DaaS enables spatial data sharing and reuse, and it plays an essential role in spatial cloud computing. Although the term DaaS concept is not defined by NIST, it is commonly used in the literature. Spatial DaaS centralizes spatial data in a repository that can only be accessed through service interfaces, while providing fast and secure access to Earth observations. Data services can perform preprocessing steps (e.g., using validation filters) before effectively saving them in the cloud, making DaaS suitable for improving the quality of spatial data. DaaS enables data sharing, discoverability, accessibility, and usability on the fly to support “science on demand” (Yang et al. 2011).

Spatial cloud services can provide several kinds of functions to compute, for example, Normalized Difference Vegetation Index (NDVI), Earth surface temperature, ocean temperature, thermal anomalies and fire, burned areas, digital land elevation, and other functions that often involve a complex series of geospatial processes. Today there are many different kinds of spatial cloud computing software, such as the ArcGIS implementation in the cloud, GISCloud, CartoDB, Google Timelapse, GeoCommons, MangoMap, OpenGeo Cloud Edition, SpatialStream, and MapBox.

Software development in cloud environments

Developing software to run in cloud or non-cloud parallel environments requires different strategies. In a noncloud environment, software development can use implicit or explicit parallelism. In implicit parallelism, programming language compilers or interpreters translate computer program codes into parallel codes automatically. However, if the codes are created in a sequential (serial) way, parallelism can be lost. In explicit parallelism, programming languages (e.g., C*) provide specific functions to deal with parallelism.

Cloud environments provide their own application program interfaces (API). Therefore, programmers developing software for cloud environments must make use of specific APIs for each environment. For example, Google provides APIs for their cloud computing infrastructure (Google Cloud Platform) in different programming languages, such as Java, Python, and PHP.

SEE ALSO: Technology; Technology and development

References


Austin, TX: IBM, International Technical Support Organization.


Participatory action research

Jeffrey R. Masuda
Queen’s University, Canada

Participatory action research (PAR) refers to a wide range of collaborative principles and practices that aim to increase the relevance and impact of research by positioning communities at the center of inquiry into problems that affect them. PAR has been embraced by researchers who have recognized how traditional expert-driven approaches have often prioritized government and corporate interests to the exclusion of individuals, communities, or populations who are the subjects of research inquiry and are often exploited by it. PAR theorists accuse the dominant research industry of being complicit in a range of injustices that affect these people because of its inability to take on problems linked to patriarchy, racism, sexism, ableism, ageism, and colonialism that both structure various forms of oppression in the world and are deeply embedded in research institutions and practices themselves. In the latter case, feminist PAR has been particularly influential in exposing the unchecked power of research disciplines and institutions to historically privilege some, often white, masculine, heteronormative, ableist, and Eurocentric perspectives over others. PAR directly confronts this epistemological crisis by shifting notions of expertise to include nonacademic, experiential, and indigenous forms of knowledge, and by ensuring that all aspects of research, from design to dissemination, are initiated, vetted, and controlled by those who are affected by its conduct and outcomes. This approach departs from conventional research where academic or professional researchers place themselves in the position of inquirer, and often position individuals, communities, or whole populations as the inquired (or research subject). In this sense, research is seen to be complicit in human oppression insofar as its practice strips people of their agency, whether by quantitatively reducing them to variables of statistical manipulation or, in qualitative approaches, condensing, cleansing, and transforming their insights into “data” and granting them the status of knowledge only after they have been subjected to the interpretation of the researcher.

In contrast, subscribers to PAR prioritize the pursuit of justice in the world by communities as the primary aim of research. They achieve this by offering several specific principles which equip researchers and communities to work in concert to expose the political and economic underpinnings of the oppressive social and environmental conditions in which they live and often suffer. Commonly cited principles include: (i) leveraging the power of collaboration between researchers and communities to provoke a transformation of societal power structures; (ii) treating knowledge production as a democratic process; (iii) widening the scope of research practices in ways that accommodate a more pluralistic notion of expertise; and (iv) promoting collective emancipatory action, either within or outside of existing structures of power, leading to social justice. The idea of community empowerment is often cited as both a means and an end of PAR. However, while the placing of
PARTICIPATORY ACTION RESEARCH

communities in greater decision-making roles over research has proven effective in achieving many local aims, some have sounded a note of caution over how the rhetoric of empowerment is easily co-opted into neoliberal agendas of governmental downloading, a rhetoric of individual responsibility, and “third way” politics more generally (see Jupp 2007 for a discussion of, as well as an empirically substantiated response to, this critique within geography).

In PAR, those affected by a particular health, social, or political problem learn to become investigators into the conditions that affect their own lived experiences. But PAR is not a methodology per se, although methodological preferences often gravitate toward the qualitative and hermeneutic, rather than the quantitative and positivist. PAR may include research that does not involve people at all, as for instance in research problems focused on environmental problems such as pollution monitoring or investigations of housing conditions. There are many examples, including in geography, of quantitative studies that have been undertaken by and with communities to confirm suspected environmental problems as a way to mobilize action. For example, numerous US-based PAR studies of environmental determinants of asthma have deployed epidemiological methods and geographical information science (GIScience) to successfully support community-based policy advocacy and outcomes (see Minkler (2010) for one of many examples). Whatever methods are undertaken, the overarching goal of PAR is to prompt political action that not only addresses the immediate problems but provokes more systemic societal change that corrects the social, environmental, or health injustices that give rise to them.

Many variations of PAR exist (see Kindon, Pain, and Kesby 2007 for a concise review) owing to numerous historical origins around the world and a concomitant plurality of philosophical foundations, terminological traditions, and methodological preferences that coincide with the disciplinary and political climates of their founding practitioners. For brevity, only main lineages will be reviewed here. The so-called Southern School of popular education originated in the seminal work of Brazilian academic activist Paulo Freire who, from the 1940s through the 1990s, elaborated a critical pedagogical approach that aimed to increase the political literacy of peasants toward the conditions of their oppression. Influenced by his own lived experience of poverty and oppression as a child, Freire sought to join together political and functional literacy in the delivery of education to (or rather with) the illiterate poor as a means to restore their humanity as political citizens. Thoroughly Marxist and anticolonialist in its orientation, Pedagogy of the Oppressed (1968; translated into English in 1970) provides a philosophy of inquiry that places human oppression in a relational space between the oppressed and their oppressors; both are required to participate in education if there is to be true liberation, otherwise one form of oppression is merely replaced by another: “the oppressed, instead of striving for liberation, tend themselves to become oppressors” (Freire 2000/1968, 45). At the heart of Freirian pedagogy is a rejection of the “banking approach” to education, which has a dehumanizing effect on recipients who are treated as passive receptacles, to be filled with the very systems of knowledge that oppress them. This approach, which applies equally to research as it does to education, is seen to have a colonizing influence on learners, subjugating them into a relation of domination that reinforces the power of societal elites. Rather, Freire’s popular education involves a radical shift in the concept of education, in which all inquirers, whether oppressor or
PARTICIPATORY ACTION RESEARCH

oppressed, engage in a process of collective political consciousness raising (conscientization) which facilitates a more radical scrutiny of the underlying conditions of oppression that are responsible for an alienation from justice that is shared by all. Often considered revolutionary in nature, popular education has been influential throughout the world in prompting political action among the poor, women, and other politically oppressed groups, particularly in the Global South. Those in the North who draw inspiration from Freire have tended to tone down this revolutionary impetus into more of a philosophical barometer that gauges the level of community engagement and solidarity required by researchers who wish to work within marginalized communities.

The so-called Northern School of PAR, often known as “action research,” is a more conservative version of PAR that originated in the 1940s with philosopher and psychologist Kurt Lewin. Born into a Prussian Jewish family and educated in Germany, Lewin immigrated to the United States before World War II, where he elaborated an approach to research that could be specifically applied to solving social problems. Influenced by Lewin’s theoretical forays into field theory in which group participation is central, both to the accomplishment of tasks and to the functioning of democracy, action research emphasizes a deliberative process of collective discussion as fundamental to personal and organizational change. The role of research in this approach is as an agent of change, responsible for facilitating purposeful collaborative efforts toward group goals. In action research, the artificial distance that is held up in conventional, individually oriented research between knowledge production (of the researcher) and actions to be taken (by the user) is seen as detrimental to the effectiveness of organizations. This distance can be reduced by building a more collaborative approach to inquiry that places research projects within an iterative cycle of planning, action, and evaluation, leading to direct social action as the desired result.

Community-based participatory research (CBPR) is a term that has gained traction since the mid-1990s, in many ways as a compromise response to the radical and conservative orthodoxies of the Northern and Southern Schools respectively. CBPR has come to refer to research arrangements that are based on an allegiance to postpositivist forms of research undertaken in partnership with nonacademic organizations. Studies that have employed CBPR range from modest efforts to evoke the priorities of communities in local affairs that concern them to those that deploy a more radical politics that mobilize communities to expose manifest injustices and to scale up action into national and even international arenas. A term often used by researchers working in the area of health, CBPR-inspired projects tend to harness local activisms surrounding chronic social or environmental deprivations that have affected community health and well-being. Many communities, particularly indigenous ones, are attracted to CBPR not just as an alternative to conventional approaches, but as a way to directly confront the colonizing agenda of research itself, including countless injustices and atrocities that have been invoked on racial minorities and indigenous peoples in the name of research (see Meyer 1981 for the classic example of the Tuskegee syphilis experiment, a study that followed but did not treat over 400 African American men with syphilis; see Mosby 2013 for a recent and explosive example of intentional starvation by Canadian government scientists of over 1000 First Nations children in northern Manitoba and in residential schools during the 1940s and 1950s).
Indigenous PAR has become increasingly influential among anticolonial scholars working in both the Global South and North who are interested in correcting long-standing social and environmental injustices inflicted on indigenous communities. Inspired by Linda Tuhiwai Smith’s influential book *Decolonizing Methodologies* (1999), indigenous PAR implicates research within current and former colonial powers as both part of the problem and the solution in the struggle for indigenous rights. Smith is unrelenting in her evisceration of the colonial agenda underpinning Western forms of inquiry, including its long history of racial oppression under the guise of Enlightenment philosophy and its valorization of the “civilized” society. An indigenous research agenda of self-determination, decolonization, and social justice can be achieved for indigenous people only by research that positions indigenous worldviews centrally. This agenda goes to the core of the research enterprise, indicting the ontological status of knowledge itself as being inseparable from the people whose bodies, histories, and experiences have produced it. It is with this understanding that indigenous research principles such as OCAP (ownership, control, access, possession) have begun to counter the dispossessing legacies of nonindigenous research (Schnarch 2004), in part by ensuring that indigenous knowledge stays in the hands of indigenous peoples.

**Motivations in geography**

Contemporary applications of PAR in geography can be found in all of the aforementioned traditions. Generally speaking, geographers who embrace PAR have been motivated by a search for the root causes of localized spatial inequalities, whether measured in human outcomes, such as socioeconomic status and health and wellbeing, or in the environmental conditions that precede them, such as urban air pollution, housing conditions, or water quality. The PAR field has been receptive to geographers’ conceptual and methodological repertoire, as communities that become engaged in research want to know more about the local context and conditions that directly affect their lives. Investigations of community problems have been enhanced by geographers’ ability to provide and interpret thick descriptions of every life, as well as by their proficiency in mapping, in photography, and in other forms of conceptual and empirical sociospatial visualization. Among these, photography-mediated interviewing has been favored by geographers since the late 1990s (Castleden and Garvin 2008). Most PAR studies in geography point to place-based processes of racialization, class-based economic marginalization, colonialism, or other forms of discrimination as being responsible for observed and experienced inequalities. In tying local conditions to these systemic social structures, PAR provides geographers and their community partners with a mechanism to scale up research insights and impacts into broader political arenas.

A strong uptake of PAR by critical geographers, starting in the 1990s, coincided with a demand for more relevant “public geography” (see Martin 2001), following the discipline’s reorientation away from positivist spatial science in the 1970s as well as the meandering, albeit often critically oriented, postmodern turn of the 1980s. With the ushering in of more culturally substantive modes of inquiry into social and health inequities, and the corresponding elaboration of social theories pertaining to communities, cities, and “place” more broadly, geographers have become acutely aware of
the epistemological tensions implicit in acts of inquiry that take place at a distance from the world, as well as the claims geographers make about the world, including their complicit role in global imperialism since the nineteenth century. Since the 1990s, PAR has given geographers of a more critical orientation a more praxis-focused way to critique the structural forces of capitalism and neoliberalization, following long-standing complaints about the failure of the “armchair” theory of radical geography to offer viable alternatives to the status quo (Pain 2004).

One epistemological objective of PAR-inspired geography is the reassertion of human values connected to place within often abstract and dehumanized governmental, corporate policies and practices. For example, issues of environmental contamination affecting particular communities involve more than the scientific assessment of levels of exposure and resulting health impacts, important as these are. Rather, environmental risk affects wellbeing, disrupting not only people’s physical health but also the ontological security of whole cultures and societies that are tied to universal human values of personal dignity, home, family, and community (Edelstein 1988). Furthermore PAR provides a practical solution to the problem of distrust in environmental science and policy, including the professionalized forms of knowledge inquiry that produce them, which are widely perceived to have been given over to governmental and corporate control and to be serving agendas that are incommensurate with grassroots preferences and priorities connected to place. More recent PAR applications in geography have attempted to reconcile the need for evidence and human values, adopting mixed-methods approaches that engage citizens not just in introspective studies of their own perceptions and experiences but in scientific agenda setting, research processes, and policy-influencing activities, literally working alongside scientists and policymakers in their own highly specialized and professionalized domains. In this vein, public participatory geographical information systems (PPGIS) deserve mention as a burgeoning method of problem and solution mapping made possible by technological advances and increasingly user-friendly GIS software and its uptake in web-based platforms. Finally, in widening the scope of participation in research to marginalized groups, geographers have been particularly adept in engaging with youth. In recent years, youth-led PAR (YPAR) has been developed as a special form of PAR that prioritizes both youth-accessible and -preferred processes (often replacing more rigid, linear, and formalized methods with more flexible and creative, fun, and age-appropriate activities) as well as youth-specific outcomes, particularly in terms of confronting urban spaces (and the policies that produce them) that attempt to “put youth in their place” (see Skinner and Masuda 2013 for a more in-depth account of these priorities).

In summary, the combined effect of the burgeoning literature in PAR has given geographers new inroads in their abiding insistence that place matters in the production of space and spatial inequality. As PAR approaches evolve, they are becoming ever more effective in garnering community involvement in knowledge production, displacing conventional modes of inquiry long favored by academic practitioners with more creative and often arts-based forms of inquiry that are seen to be accessible to a more diverse spectrum of community researchers and more effective in reaching public audiences.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Community; Environmental (in)justice; Indigenous people and regional
PARTICIPATORY ACTION RESEARCH

development; Inequality; Marxist geography; Neoliberalism; Patriarchy; Power; Race and racism; Radical geography; Social justice

References


Participatory development

Tony Binns
University of Otago, New Zealand

“Top-down” and “bottom-up” development

Participatory development (PD) aims to engage local people in planning for development interventions which are normally intended to benefit them. PD strategies have become progressively popular since the late 1980s and 1990s, and now represent development orthodoxy, being popular with many nongovernmental organizations (NGOs), as well as with bodies such as the World Bank.

More democratic “bottom-up” PD methodologies are seen as a significant step forward from the somewhat dictatorial “top-down” development planning that was undertaken in the early decades after World War II. A classic example of this was the ill-fated East African (or Tanganyikan) Groundnut Scheme, designed to meet a serious shortfall of vegetable oils in postwar Britain. Launched in 1947, after a superficial nine-week survey, the plan was to clear and plant 1.3 m ha of land with groundnuts, mainly in Tanzania, then the British Protectorate of Tanganyika. Using converted ex-army equipment, 1000 Europeans and 60 000 Africans were engaged in land clearing and planting. Despite expenditure of £49m, crop yields were lamentable and the project was canceled in 1951. This top-down, highly centralized project suffered from inadequate soil and rainfall data in a drought-prone region, and an unrealistic faith in Western technology. Sadly, no one thought to talk with the local Wagogo agropastoralists about environmental conditions (Binns, Dixon, and Nel 2012).

Many postwar development projects followed European-based modernization strategies and showed little awareness of local knowledge among the communities that were supposedly “being developed.” Rural communities in developing countries were often portrayed as being remarkably uniform. Anthropologists such as Polly Hill, working in West Africa in the late 1960s, challenged the idea of an “amorphous peasantry,” and observed that among such communities there were “rural capitalists” who were engaged in adopting new crops, forward planning, and investment in infrastructure development. In criticizing current data collection methods, Hill strongly advocated giving greater attention to “producers” rather than to “production” (Hill 1970).

Early PD methodologies were inspired by the work in the 1960s and 1970s of Brazilian educator and philosopher Paulo Freire. He encouraged teachers to mingle with communities, asking questions and understanding the people in their broader social context. During the 1970s, frustration with brief and superficial “rural development tourism,” and lengthy and costly questionnaire surveys, led to the emergence of Rapid Rural Appraisal (RRA) and a range of new participatory survey methods such as transects, sketch mapping, and semistructured interviews, which would, it was hoped, deliver information that was relevant, timely, accurate, and usable. Agroecosystem analysis in the 1980s contributed further data collection methods.
PARTICIPATORY DEVELOPMENT

The growing popularity of participatory development

Robert Chambers has played a key role in championing PD methodologies. His 1983 book *Rural Development: Putting the Last First* (Chambers 1983) was a landmark in raising awareness about the need for a better understanding of rural communities in poor countries, and paved the way for the wider use of PD methodologies. Chambers argued that, “Outsiders (whether in universities, government departments, research or training institutes, or NGOs) have believed their professional knowledge to be superior, and so have behaved in ways which have almost universally inhibited the expression of local people’s capabilities” (Chambers 1993, 98). Chambers was concerned about issues such as rural people’s knowledge, power and the poor, and the need for change and “reversals” to “put the last first.” He argued, “Reversals require professionals who are explorers and multi-disciplinarians, those who ask, again and again, who will benefit and who will lose from their choices and actions. New professionals who put the last first already exist; the hard question is how they can multiply” (Chambers 1983, 168).

In the late 1980s, lessons from RRA and agroecosystem analysis led to the development of Participatory Rural Appraisal (PRA), and a range of participatory methodologies designed to empower local and often marginalized people, allowing them to express their views, share their knowledge, and engage in action that has meaningful benefits for themselves and their communities. Working within and in close collaboration with communities, PD methodologies involve a range of data collection techniques including: focus group discussions, participatory surveys, transect walks, participatory mapping and modeling, matrix ranking, scoring, and diagramming. These methods

should lead to developing empathy, building up trust with communities, understanding “what makes them tick,” and “handing over the stick” to the participating communities in leading the data collection processes, identifying needs and aspirations, and ultimately playing a greater role in development interventions.

The challenges of participatory development

Many academics and development practitioners working among poor communities would probably agree that PD represents a significant step forward from the “blind surveys” and “top-down” development strategies of the 1950s and 1960s. There now seems to be a genuine concern to work “with,” rather than “on,” communities in shaping development interventions which will hopefully empower and uplift livelihoods (Kyamusugulwa 2013). But often the question is asked as to whether PD is a means to an end, or actually an end in itself. In simple terms, it is probably both, as the process should assist in both identifying and implementing more appropriate development interventions, while also helping to articulate and empower participating communities. But the key test is whether outcomes from PD include tangible benefits for all members of the community.

As PD strategies have become more widely used, the nature of participation has been closely interrogated, since complex issues of power and control are involved. The key parameters seem to be: Who participates? How do they participate? When do they participate? As Cornwall observes, “‘participatory’ processes can serve to deepen the exclusion of particular groups unless explicit efforts are made to include them” (Cornwall 2008, 277). A key challenge for those who are conducting PD methodologies
among communities is to be sure that everyone is involved, that certain groups are not excluded, and that different viewpoints are equally articulated and incorporated into decision-making processes. Inclusion, exclusion, and degrees of involvement in PD methodologies are important considerations, making the implementing of such methodologies a skilful undertaking. Some individuals or groups may exclude themselves, for example, by intentionally not attending public meetings. Incorporating the views of marginalized groups such as the poor, the elderly, the young, and women is a particular challenge. Much concern has been expressed about women’s views being overlooked or marginalized (Gujit and Kaul Shah 1998). For example, in a 1996 survey of urban farming activities in Kano, northern Nigeria, it was apparent that relatively few women were visible in the fields due to strict Hausa Islamic seclusion protocols, so it was necessary to engage a team of women enumerators to go into family compounds to talk with women vegetable growers.

Local power hierarchies need to be fully appreciated in relation to levels of participation and possibilities for delivering tangible outcomes. But fully appreciating such power structures can take time. One of the criticisms of PD is that communities and groups may be perceived as being homogeneous and isolated, while in reality there are complex social networks both within and beyond the community, and development interventions can easily be shaped by pre-existing relationships. While power relations commonly exist among community members, they also exist between development facilitators and possible beneficiaries. A particular concern is the possibility of “elite capture,” or “elite control,” where participatory methodologies are dominated by elite members of the community who have particular objectives which can shape development interventions (Kyamusugulwa 2013). But, in reality, any interest group within a community could control the PD process.

More recently, and drawing on PD methodologies, the “Reality Check Approach” (RCA) was initiated in 2007 by the Swedish Embassy in Bangladesh – supported by the Swedish International Development Cooperation Agency (SIDA). Through a process of “immersion” and living with households, RCA aims to “engage with, listen to, observe and document the voices, opinions and experiences of ‘people living in poverty’ in relation to the policies and interventions that are carried out in their name.” The main purpose of RCA is to understand how specific policies lead to tangible benefits at grassroots level by developing two-way relaxed and informal conversations that can be held without interruption to household duties. Using RCA, strong and trusting relationships have developed through revisiting households each year (SIDA 2012). There is some concern that engaging with communities through PD methodologies can be time-consuming and possibly even stressful, and that “participation fatigue” may lead to self-exclusion by certain individuals and groups.

Despite its attendant difficulties, PD can yield some real benefits for communities and households, and interaction, dialogue, and decision-making among community members and with development teams might be regarded as a valuable expression of citizenship.

SEE ALSO: Community-based natural resource management; Development; Networks, social capital, and development; Participatory geographies; Participatory modeling
PARTICIPATORY DEVELOPMENT

References


Further reading


Participatory geographies

Kye Askins

University of Glasgow, UK

Participatory geographies do not (only) describe geographic research in which people participate; they are not (only) methodological/a set of methods; neither are they (only) community engagement, or “impact,” or limited to and by “knowledge exchange.” Participatory geographies include yet exceed all of these. And it is necessary to start our definition in such a way, since “participatory geographies” are increasingly being claimed within and beyond the discipline, loosely and out with the epistemological and political framings that we seek to define here.

Participatory geographies are political and politicized; they are activist and aim to resist oppressive structures and hegemonic discourses that serve to socially and spatially marginalize individuals and communities; they explicitly set out to enact progressive change; they are about working together with communities and individuals, ever cognizant and making visible issues of power, inequality, voice, agency, and difference. Participatory approaches destabilize academic-as-expert: they suggest new ways of collecting data, coproducing knowledges and determining outcomes, with more horizontal focus on who steers and benefits from research.

Participatory geographies are closely aligned to participatory action research (PAR): both center around collaborative processes of research, education and action as/toward social change; both are produced through and productive of a range of actors, including academics and nonacademic co-researchers and participants; both involve investigation that leads to action, which leads to reflection and learning, which then feeds back into investigation, in ongoing processes of “verification” of research data and outcomes.

Such approaches have been in resurgence since the early 2000s, with a “participatory turn” across contemporary social and environmental sciences, representing radical, creative, and exciting provocations for/in human geography (Kindon, Pain, and Kesby 2007). Participatory geographies and PAR are often referred to interchangeably; yet recently participatory geographers/geographers orientated to PAR are foregrounding more explicitly the role of geography and emphasizing the more-than-action-research elements: we return to this possible distinction later.

First, we need to acknowledge issues regarding voice here. This is a multi-authored piece, in that its ideas and examples come from a decade of learning, debating, and dialoguing about participatory geographies with a diverse range of academics and publics. Collaborative writing, unfortunately, remains anathema to normalized academic publishing structures, and meaningful collaboration was unfeasible for this entry, due to deadlines and word length. We thus adopt Cook’s use of “et al.” (Cook et al. 2014) to indicate the co-authorship inherent in writing this entry, through which many voices resonate.

Mindful of our privileged position in the (Western) academy, we also highlight the gaps and omissions throughout.
PARTICIPATORY GEOGRAPHIES

Theory and praxis

Participatory geographies are epistemological. Paulo Freire, a South American educator, is often cited as a key inspiration for involving participants directly in research, knowledge production, and its outcomes. In the 1960s, Freire developed an epistemology foregrounding research as social transformation, explicitly connecting with local struggles and knowledges, to enable a process of conscientization (conscientizacao) through which impoverished people increase awareness and understanding of their marginalized positions, and the structural forces involved, to (better) resist such positioning through political and social action. These theoretical underpinnings and political perspectives reverberated across the Global South/majority world throughout the 1960s/1970s, challenging ongoing legacies of colonization and resisting capitalist oppressions and early neoliberal “development” paradigms. Versions of anticolonial approaches to research echoed across India, Africa, and South America and among indigenous people in Global North countries.

Participatory geographies resonate, then, with postcolonial and subaltern geographies: listening carefully and dialoguing, reframing issues, and nudging learning and scholarship in new directions, while always foregrounding power relations and paying serious attention to issues of positionality and subjectivity. Critical race theory also informs such work, unpicking and addressing identity politics and exclusions (material and discursive) within communities as well as across cities, states, and the world. Participatory geographies are described as “enabling” or “facilitating” participation, critically aware of the need to reconceptualize who is enabling whom, in line with postcolonial critiques that argue against such academic orientalism. And also mindful of issues of representation: who is speaking for whom.

Feminist theory is central too: challenging extractive research as normalized within academia, through alternative methods and approaches that work to destabilize any public-academic dualism or “ivory tower” notion (see here commonalities with “public geographies”). Key here is addressing the power inequalities that enable and are reiterated through academic-as-expert. Informed by feminist ethics of reciprocal care in its most profound sense, as a deep respect for relationships and humanity, participatory geographies conceive public–private and global–local as interconnected rather than binary, engaging an analysis of how the intimate and global intertwine, and connecting to critical geopolitical work on social and spatial in/justices as interscaling/interscaled.

Poststructural theory further offers us alternative perspectives on the circulations of power in/through discourse. Paying attention to language, representation, and narrative is vital to foreground relational, fluid, plural subjectivities that emerge and are (re)made through intersubjective engagements in research. Such epistemological approaches in geography and other disciplines, as well as across (some) governmental/nongovernmental policy, have increasingly precipitated a renewed uptake of participatory action research.

Participatory action research

Participatory geographies have been developing from PAR since the early 2000s in response to renewed debates around the “relevance” of the discipline, and also in parallel with and connected to re-emergent focus at that time on more overtly activist ways of doing research, learning, and teaching. It is important
to consider here participatory geographies’ heritage and continuing overlap. PAR itself is embedded in a long trajectory of emancipatory approaches to research and learning across a range of disciplines, including legal studies, education, liberation theology, and social/critical psychology. Kurt Lewin is credited with the phrase “action research” in 1946, to describe the idea of iterative, cyclical rounds of action and reflection in “spiral science.” This approach evolved to become science that is not hierarchical, extractive, imperial; research that critically questions its purpose and relevance, taking seriously the idea that academia embodies, and, often and problematically, reproduces classed, raced, gendered, and other hierarchies.

Both participatory geographies and PAR, then, are engaged with/in everyday realms of struggle, working with communities, identifying issues to challenge and change, and raising consciousness (after Freire) – wary of privileged/ing academic positions and aware of the capacities of people to raise their own conscious understandings of sociospatial exclusions. In PAR, action is central, and emphasis is on research outcomes/actions directed by and of utility to co-researchers/participants. This can be booking university IT labs for community group access to computers for learning; training people in research methods and data analysis; producing policy-relevant publications and/or education/awareness-raising materials, crucially, writing with co-researchers; staging direct action campaigns; and so forth. (For vibrant examples see http://publicscienceproject.org/.)

More narrow approaches to PAR, when limited to methods, have been and should be critiqued for re-paternalizing development approaches in the Global South, and simplifying complex lived realities for managerialist needs: another neoliberal tool disguised as “participation,” tokenism at best. This critique highlights knowledge as a powerful normative construct, produced through social norms and cultural practices, embedded within logics of power relations. “Local knowledge” is thus not some fixed, uncontested commodity held/owned by local communities, which may be appropriated by authority. As such, PAR and participatory geographies must remain alert to these issues to avoid epistemic violence. Certainly, methods are key to deconstructing dominant discourses and social hierarchies (see later), but the wider deployment of ethics, politics, and relationships is crucial to ensure that research progresses through co-ownership, mapping out relationships between social structures and injustice in everyday life experiences (Cahill 2010). Such deployment emphasizes working as/toward equality and justice, rather than simplistic “solutions” that may be completed, precisely because identities and social relations are always becoming and shifting through participatory geographies as process.

Participatory geographies as process

Participatory geographies are “messy” encounters with and across intersectional differences, involving coproduction of knowledges (including but far in excess of academic knowledge) and foregrounding the process of collaborative research (Askins and Pain 2011). Inherent is a commitment to understanding diverse and marginalized experiences in affirmative ways, through an ethics of encounter that remains open, responsive, and supportive. Such iterative and becoming relationships, between academic researchers and co-researchers and between/among co-researchers, are always fragile and at risk and full of transformative potential.

Taking process seriously demands opening out to emergent, innovative, and creative methods
PARTICIPATORY GEOGRAPHIES

that may draw from and surpass one or more of the following:

- art;
- drama/theatre/role playing;
- diagramming;
- mapping;
- storytelling;
- media;
- dialoguing;
- photography;
- video and filmmaking;
- educational camps;
- exchange programs;
- focus groups/interviews/standard qualitative techniques;
- surveying/questionnaires/standard quantitative techniques;
- shared analysis;
- collaborative reporting, presentation, and writing;
- direct political action and advocacy.

More than research methods, we highlight here participatory geographies’ educational potential, both in learning and teaching praxis in formal classrooms, lecture and seminar spaces, and beyond – the list here pertains to participatory pedagogies as well as part of action research. Regarding the latter, participatory video (PV) is well known and developed and shares a politics of bringing forth issues for communities in context to effect change. Co-researchers are trained in all aspects of filmmaking, from choice of topic, through storyboarding, filming, and editing: key is putting people behind as well as in front of the camera. Participatory mapping is also particularly effective, and often emergent through research with indigenous communities, increasingly utilizing GIS and digital media. Indeed, the potential of new social media and technologies is being explored through participatory geographies, with exciting and creative “methods” being developed by co-researchers all the time, emergent in process and the material realities of place.

Emphasizing process raises issues regarding the temporal dimensions to participatory geographies. It is critical to engage in slower ways of working that build trust and reciprocal relationships, to engage reflectively and iteratively with co-researchers through all stages of a project, including analysis and “outputs.” At the same time, there is often an urgency as co-participants suggest outcomes and actions that need to be fast paced to respond to local public policymaking. The long delay of peer review or high level policy report writing, to which co-researchers including academics may aspire, is useless to many projects in the reality of everyday struggles. Thus participatory geographies simultaneously involve long term engagement and commitment and more immediate responses and activities (see activist geographies).

Given the importance of process, and emergent nature of method/activity/outputs, we need to consider more carefully the role of place and space.

The role of place and space

Geography matters. Participatory geographies are embedded in and across the everyday neighborhoods and locales of people’s lives, opening up new spaces for knowledge coproduction and action – spaces that are as much a part of the process, mutually co-constructing identities, relations, and encounters. While PAR certainly pays attention to the place of research, it is on this point that geographers have specific spatial theorizing to bring to bear. We raise three key, interrelated issues here that speak to the theories discussed earlier.
First, a feminist understanding of scale as interconnecting allows us to work closely with the ways in which global hegemonies and dominant structures are not (only) top down but also hierarchically influencing everyday experiences in different locations. Participatory geographies are centered on the ways in which local, embodied, and intimate spaces speak and act back, such that the body, nation, and global are conceived as caught up in similar processes rather than operating as/at different scales. This further destabilizes any separation between public and private spaces, deconstructs the “state” versus the “domestic” scale binary, and demands attention to the complexities and contradictions within individual/community lives caught up in webs of regional, national, and global processes. Carefully conceptualizing across place, space, and scale brings a specifically geographical lens to participatory approaches.

Second, in line with the above and drawing on activist work, participatory geographies may be enacted as “convergence spaces”: engagements that are place-based but not place-bound, centrally concerned with developing a politics of solidarity capable of transferring across space without flattening out contested positions, struggles, and beliefs. For example, the Participatory Geographies Research Group (PyGyRG) is UK-based, but endeavors to research, learn, and write collaboratively and act in solidarity with communities and academics across the world (see PyGyRG 2012). Here too, reflective participatory pedagogies can connect students with struggles outside the university; for example, see Mrs. Kinpaisby’s (2008) notion of “communiversity,” which remakes/redoess university-community as interrelational, connecting research and pedagogy, and engaging with communities as positioned in cross-scale structures of governance. Indeed, participatory geographies offer genuine pedagogic opportunities and respond to problematic, dominant discourses regarding “service learning” provision and “community engagement” practices.

Third, in taking issues of voice, agency, and power seriously, Torre et al. (2008) draw on “contact theory” to understand participatory approaches as “contact zones” within which all involved encounter each other through close negotiation, emergent identities, and contested diversity. In this vein, participatory geographies may work toward creating autonomous spaces as action in itself. Autonomy is explicitly about a prefigurative politics in which a “politics of affinity” examines ways in which multi-issue politics and diverse, contested positions come together to address overlapping oppressions (see Autonomous Geographies Collective 2010). It is precisely the messy negotiation of different bodies within communities and local spaces that often remain masculinized, raced, homophobic, and so forth, to which participatory geographies explicitly attend.

Foregrounding what we might call an ethic of space, participatory geographies incorporate a more-than-action-research agenda. What we mean here is that the material realities and discursive productions of/in local neighborhoods are central to co-making participatory geographies whatever they need to be in that environment: research, action, education, reflection, and other aspects ebb and flow at different stages of any project, depending on context. Critical PAR attends to these issues too, opening out to process without a research component. Our argument here is that participatory geographies involve greater attention to the centrality and understanding of place and space. And such focus is increasingly concerned with participatory approaches as embodied and emotional.
PARTICIPATORY GEOGRAPHIES

Participatory approaches as embodied and emotional

There is a burgeoning literature regarding emotional and embodied geographies, exploring the relationality of bodies, emotions, and affect, and their role in coproducing spaces and knowledges. Being/becoming emotionally involved in/through contact zones and processes, it is not surprising that participatory geographers are paying attention to the ways in which social and spatial relations are produced through and productive of emotions and affect. Furthermore, the approaches and epistemologies outlined above suggest the creation of space for emotion in participatory dialogue and process as central to challenging inequalities within and across communities, paying attention to emotion in all aspects of research, from bid-writing/proposal stage, inclusive of participants thoughts and feelings, through to dissemination and after, cognizant of how “leaving” any project affects researchers and participants and incorporating these issues into the approach/work (e.g., Askins 2009). Such embodiment connects to debates regarding nonrepresentational and actor-network theories, and more-than-human geographies, considering everyday identities and lives as felt, facilitated, and performatively practiced through ontologies of becoming.

Participatory geographies are starting to explore how the materiality of objects, landscapes, environments, and urban fabrics – and how they are mobilized – is co-constructive of identities, emotions, affects, and in/exclusions, crucially maintaining a humanism/responsivity to human agency through engagements that practice ontological politics (Gibson-Graham 2008). Wider calls for methods that attend to more than spoken or textual elements of the world, across a range of senses inherent to embodied experiences, are being taken up through the creative and open approaches to developing more-than-methods outlined above. Indeed, since participatory geographies are a diverse set of praxes, with ongoing, emotional self-reflection, among all co-researchers, working in ways that closely attend to emotions and affect is one of the key future challenges.

Future challenges

Participatory geographies share the complex challenges of the twenty-first century: widening gaps between rich and impoverished in late capitalist “austerity”; multiple climate crises; and ever-rising urban populations. If we ask ourselves who we are as academics, what is our purpose; it seems clear that participatory geographies are increasingly relevant, required, and demanded. Some may identify their particular academic position as an educator, others as scholar, most of us with elements of both, recognizing the broad continuum along which we may sit. Wherever we are, our argument here is that working with students and research participants is central to co-developing democratic, grounded knowledges that may better actively address issues of social, spatial, and environmental injustices.

One concern among radical geographers and other academics is regarding the ways in which neoliberalism suture capital as inevitable, and how capitalist structural compulsions attempt to recoup resistance, at the moment of resistance, by re-appropriating the “marginal” in a post-political or postdemocratic frame. Autonomy, creativity, and fluidity have become buzzwords of contemporary capitalist business organizing. We worry at the start of this entry that “participatory geographies” are increasingly being claimed outside the epistemological and political framings that we seek to define, because “impact” and
engagement activities are being brought into academic audit frameworks across the world, not necessarily in critical ways. Pain, Kesby, and Askins (2012, 120) discuss the ambivalence of participatory geographers here, arguing “for the potential of more participatory praxis to undermine neoliberal structures within the academy precisely through subverting notions of “impact” toward a social justice agenda.”

For participatory geographies to retain an emancipatory politics, they must continue to challenge increasing commodification and competition in academic knowledge production and practices, destabilizing entrenched disadvantage within and beyond the academy. We end, then, with a focus on the opportunities that participatory geographies afford: staying alive to hope, possibility, and the potential of for people to resist oppression and the surprise of collaboration and solidarity such that neoliberal and exclusionary structures and discourses continue to be exceeded.

SEE ALSO: Antiracist geography; Feminist geography; Participatory action research; Participatory video; Postcolonial geographies; Poststructuralism/poststructural geographies

References


Cahill, Caitlin. 2010. “‘Why Do They Hate Us?’ Reframing Immigration through Participatory Action Research.” Area, 42(2): 152–161.


Participatory modeling

Klaus Hubacek
Christina Prell
University of Maryland, USA
Federica Ravera
ICAAM, University of Evora, Portugal
CREAF, Spain
David Tarrasón
CREAF, Spain

The case for participatory modeling in environmental decision-making

The decision-making needed for addressing increasingly complex environmental issues requires input from multiple disciplines. But that input, often presented via complex scientific models, can be difficult for many stakeholders to comprehend, leaving many of them excluded from the decision-making process and/or unable to make good use of the knowledge offered to them. Participatory modeling (PM) is one way of addressing these concerns. PM links science and modeling to stakeholder knowledge and experiences, thus supporting and enabling participatory decision-making. PM brings stakeholders’ interpretations, experiences, and know-how together with researchers’ scientific knowledge, and integrates this information via models ranging from conceptual frameworks to computer programs. Scientists are thus seen as only one stakeholder group who learn from and interact with other stakeholders in order to address issues with high complexity and uncertainty. Whereas traditional science emphasizes objectivity, rationality, and expert-driven collection of information and modeling of problems, PM emphasizes collective deliberation, where learning outcomes emerge from ongoing discussions. As such, learning is often linked to practical action, and thus, PM is seen as a means to bridge science with policymaking.

International support for participatory processes came in 1992, when the UN Rio Declaration on environment and development and its local Agenda 21 called for community planning as a vehicle to re-engage people with their community and society. More recently, participatory actions have gained support in Europe, where policies such as the Aarhus Convention on Access to Information, Public Participation in Decision-Making and Access to Justice in Environmental Matters require environmental managers and policymakers to increase the involvement of stakeholders in their decision-making and management work. In addition, the term “government” has frequently been replaced with “governance” in both academic and policy circles to reflect the need to involve citizens, nongovernmental organizations, and social movements directly in the development and implementation of environmental policies (Lane 2005). In addition to policy drivers, public participation in science research has been applied in a number of academic fields ranging from development studies and public health to natural resource management. Citizen-science, which refers to the engagement of nonprofessionals in scientific investigations, has a long tradition within academia. The data, technology, and skills emerging from citizen-science provide multiple examples of how
PARTICIPATORY MODELING

engagement of nonscientists in scientific research can yield fruitful, often innovative outcomes.

There is broad agreement that participation can help with complex social-ecological problems or, as Arnstein (1969, 216) noted: “The idea of citizen participation is a little like eating spinach: no one is against it in principle because it is good for you.” Indeed, the use of participatory procedures has been an increasingly common feature of environmental management, although the approaches employed are wide-ranging and reflect different rationales. First, there is an ethical or normative rationale. Stakeholders have a legitimate right to participate in model-building processes because the outputs and results of the model directly impact them. Participation is an important part of the trend towards democratization of society and is increasingly a statutory requirement. Second, there is a substantive rationale. Local people may be the best source of knowledge about their surroundings, providing comprehensive understanding of historical and current processes of change. Such local knowledge can thus supplement, or be triangulated with, scientific knowledge, potentially resulting in better models that are likely to be more “fit for purpose” and create outcomes that are more likely to be in tune with what is needed and wanted. Third, there is an instrumental rationale, by which engagement is a potential means to build greater trust between scientists and nonscientists. Forth, there is a learning rationale, the hope being that through PM, stakeholders can engage in meaningful dialogue, potentially resulting in innovative solutions and knowledge creation by changing values and mental models. Fifth, there is an empowerment rationale, where PM builds local people’s confidence, skills, and ability to cooperate, thus leading to a sense of community and stewardship as people feel more attached to the environment they interact with.

People often want to be involved in shaping their environment and seem to enjoy the process.

The underlying rationales for stakeholder participation in modeling processes can be regarded as mutually reinforcing, cutting across different levels of involvement. A first level of involvement is simply providing information, that is, “knowledge transfer,” in order to raise awareness of the model’s functionality, capabilities, and practical applications. A second level of involvement is consultation: here the concern is to elicit precise information from stakeholders to inform the model’s development. This form of engagement is a more interactive but structured process where stakeholders fulfil data needs elicited through interviews or workshops. The deepest level of involvement is “collaborative model building.” Here, stakeholders engage in more fundamental aspects of model development. They engage in collective problem framing, priority setting, option formulation, and model evaluation. Models are coproduced in an iterative learning process between modelers and stakeholders.

Evolution of PM approaches

Any form of stakeholder participation in modeling processes, or conversely, the use of modeling to support decision-making processes that involve stakeholders, can be labeled as PM. We review here some of the more influential approaches that have been developed from diverse intellectual sources.

“Group model building” emerged in the 1960s, based on system dynamics applied to industrial systems, and later to urban systems studies (Forrester 1969). In the 1980s, group modeling was pioneered by Royal Dutch Shell for developing business strategies. The approach mainly consisted of involving diverse stakeholder groups to capture and model
their worldviews and help experimentation, cooperation, and learning (Akkermans 1995). In group model building interventions stakeholders contribute with partial views of the system, but they are affected by the system as a whole.

“Adaptive management” is a decision-making process rooted in ecological studies, which emerged in the 1970s to address the twin challenges of learning and management of complex ecological systems and human–environment interactions (Holling 1978). Grounded in systems thinking, “adaptive modeling” confronts the complexity and uncertainty of natural resource management by integrating modeling within a cyclical learning process. Recently, the adaptive management literature has moved more in the direction of “collaborative learning,” which focuses on actively engaging stakeholders throughout a given learning cycle. Indeed, many PM approaches have begun adopting a collaborative learning approach grounded in soft-system methodology (Checkland and Scholes 1990).

Participatory “integrated assessment modeling” (IAM) emerged in the mid-1990s (Van Asselt Marjolein and Rijkens-Klomp 2002). Integrated assessment was a modeling approach applied to human–environment phenomena such as climate change or land change that sought to transcend disciplinary and policy boundaries and single policy arenas. Expressly participatory integrated models advocated reflections on the management of uncertainty and engagement of multiple perspectives, including those of lay people, often under the aegis of ecological economics as “mediated modeling” or “collaborative modeling,” which emerged in the 2000s as a variation of group model building. Mediated modeling broadens the participation of stakeholders to increase the shared understanding of a system’s dynamics through consensus building over complex environmental challenges (Van den Belt 2004). Other methods to support participatory decision-making processes, such as scenario development and multicriteria evaluation, have been used in combination with integrated modeling, especially in integrated water resources management.

“Companion modeling” (ComMod) (Etienne 2014), which is rooted in complex system thinking and participatory action research, was developed by researchers from the Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD; French Agricultural Research Centre for International Development) in the mid-1990s. Despite the diversity of applications of ComMod, their common basis resides in an underlying constructivist approach and the combination of diverse modeling tools inspired by social psychology (e.g., conceptual modeling, agent-based models, role-playing games, and spatially explicit models) to help stakeholders confront some of the implications of their own perceptions, beliefs, and behaviors. Stakeholders interact within a virtual, computerized world, to resolve conflicts or manage natural resources.

“Participatory mapping and geographical information system (GIS),” broadly defined as using participation-based methods for eliciting and modeling spatial data, was developed in geographical information science, in the late 1990s. For instance, the pioneering Varenius initiative investigated how to extend conceptual frameworks for spatial decision support systems to help groups of decision-makers generate tractable solutions to ill-defined spatial problems and to enable them to work together by providing a set of tools that handle tasks such as data exchange, group evaluation, consensus building, and voting. The initiative also guided research on how GIS can be democratized through placing it in the hands of a wider spectrum of society and by considering broader
PARTICIPATORY MODELING

ethical and legal implications. Since then, participatory GIS (and other variants, such as Critical GIS, GISocial, and participatory geomatics) has gained followers from urban planning, development research, community development, and landscape ecology, as well as from environmental and natural resources research.

Differences and commonalities of PM approaches

Current approaches in participatory modeling share a number of commonalities but differ with regards to goals, elements of the process, the level of stakeholder engagement, and models used or developed during the process. The key commonality among PM approaches is the emphasis on bringing nonscientists and modelers together to inform model development. Here we also highlight some additional commonalities and differences.

Aims and principles

PM approaches can differ with regards to the following aims:

- to gain insights on specific decision problems and complex interactions between socioeconomic and ecological dynamics of systems,
- to gain clarity on the impacts of decisions,
- to increase the legitimacy of the process,
- to facilitate social learning and consensus building.

Though their focus and application may vary with regards to aims, PM approaches tend to be guided by common principles. These include the following.

- Emphasis on iterative learning. The models emerging from PM approaches are less about prediction, and more about helping stakeholders learn and think through uncertainties in complex environments and potential implications of their values, preferences, and choices.

- The traditional role of scientists is questioned. PM approaches value the insights of nonscientists, seeing their inputs as complementary to their own.

- Emphasis on the process. Given the complexity of environmental issues, PM focuses less on certain one-time changes but emphasizes what kinds of knowledge gets created, how behaviors, perceptions, and institutions change and adapt, and in what ways innovative thinking and learning can be improved.

Types of stakeholders’ engagement

Classifications for PM may differ with regards to: (i) the timing of stakeholder engagement in different modeling stages; (ii) roles stakeholders take on during the process; (iii) the level and mode of participation; and (iv) direction of information flow between researchers and stakeholders. Depending on the modeling approach adopted, stakeholders may be involved individually or in groups. Their level of involvement can range from being asked to be informed and being informants at discrete points in time, to iteratively interacting and steering from the beginning and throughout the modeling stages.

Moreover, scientists assume new roles, such as scientific knowledge providers, process facilitators, and mediators. They differentially interact with stakeholders through techniques taken from participatory research and communication. At earlier stages of the modeling process, scientists may adopt a broad range of methods from social sciences and participatory action research to collect primary and secondary
data, share knowledge, or identify actors and resources. Other approaches are necessary during model conceptualization and design, including developing the modeled system’s ontology, the model boundary, and the user requirements. Stakeholders may be asked to choose analytical tools and, in some case, to help create prototypes. During model construction, scientists may build in participatory workshops, jointly with stakeholders, the operative simulation model. Then, at the stage of model calibration, verification, and validation, stakeholders may be involved through workshops or questionnaires to elicit feedback on content, the model’s validity and usefulness, and their evaluation of the learning process. Finally, the use of models by stakeholders may be direct or mediated by scientists.

Differences in models

There are two main categories of models being built during a participatory modeling process: (i) qualitative or conceptual models, which explain the system on the basis of multiple knowledge sources and mental maps; and (ii) quantitative, computational models of the system. Conceptual models have been used to point the way toward future research and can help stakeholders to articulate hypotheses to be tested. Many proponents of this methodology, especially those from sociological, anthropological, and development studies traditions, stop at the stage of a diagrammatic conceptual model. Others argue that although all models are simplifications, using mathematics to create different simulations of the future is necessary for exploring the consequences of what is assumed at the conceptual modeling stage.

Quantitative models applied to natural resource management have been developed by different research groups and they span a range of methods, including dynamic system models, multi-agent models, Bayesian networks, and mixed models adopted in integrated assessment. Dynamic system models usually represent social and ecological systems in a causal loop model, where stocks are connected by flows. Multi-agent systems (MAS) are intended as simplified, virtual metaphors of reality (social, biological, and physical) via creation of artificial worlds composed by agents with some degree of autonomy that interact with other agents and with the environment. Frequently microworlds and games are adopted, including role-playing games, which generate new knowledge by creating common understanding of shared situations (and at the same time allowing to reflect on divergences among agents), via learning-by-experimenting within the model, and by establishing the rules and goals of negotiation processes. Bayesian networks provide a probabilistic representation of the relationships between input parameters and their dependent variables. They integrate different knowledge sources and can include expressions of stakeholder perceptions of uncertainty. Overall, natural resource management has mixed and matched quantitative and qualitative modeling tools, such as coupled socioeconomic and ecological models in combination with storylines and all integrated within a PM approach.

Outlook, critical issues, and barriers to implementation

In traditional science-intensive environmental decision-making processes, the science is more or less separated from policy processes. Given the increasing frustration with the outcomes and shortcomings of traditional approaches to model building, alternative processes have been developed to facilitate the interactions between scientists and other stakeholders. But new forms of knowledge production may lead to an institutional split of science into an academic and a
Table 1 Criteria on a continuum from traditional to alternative approaches to model-based decision support.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>“Traditional” approaches</th>
<th>“Alternative” approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree of stakeholder participation</td>
<td>Individuals: stakeholders are individually consulted for their inputs</td>
<td>Groups: interactive team learning situation where the model is almost a by-product of the modeling process</td>
</tr>
<tr>
<td>Timing of participation</td>
<td>At the beginning and/or the end of the process</td>
<td>Throughout the process</td>
</tr>
<tr>
<td>Role of stakeholders and scientists</td>
<td>Stakeholder is seen as data provider and as layperson whereas the modeler is seen as expert</td>
<td>Stakeholders are seen as co-creator of knowledge; the modeler is seen as one of the stakeholders; multiple legitimate knowledge sources exist</td>
</tr>
<tr>
<td>Framing the research</td>
<td>Stakeholders work with given questions/problems and within a given frame</td>
<td>Stakeholders guide problem formulation and help design the conceptual frame of the model</td>
</tr>
<tr>
<td>Purpose</td>
<td>To develop better models</td>
<td>Multiple purposes: (i) Improve model; (ii) group learning/capacity building – improves acceptance of outcome; leads to consensus – increases stakeholder buy-in and improves implementation; (iii) transparency; (iv) legitimacy</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Predictable outcomes</td>
<td>Less predictable outcomes as stakeholders shape the process</td>
</tr>
<tr>
<td>Power sharing</td>
<td>Little</td>
<td>Potentially high</td>
</tr>
<tr>
<td>Model integration</td>
<td>Models of natural and human world are separate</td>
<td>Coupled models of natural–human dynamic systems</td>
</tr>
<tr>
<td>Validity of models</td>
<td>Scientific standards as metrics of success of the model</td>
<td>Untraditional metrics of success of the process, e.g. measuring new knowledge, evaluating adaptive behavioral, social learning, and institutional change</td>
</tr>
</tbody>
</table>

Public policy branch (Haag and Kaupenjohann 2001) with potentially quite different characteristics, processes, and outcomes (see Table 1).

Not all decisions and decision contexts are equally suited for PM, thus the analyst has to make choices and consider tradeoffs with regards to purpose of the modeling exercise, model type and method, and the participation mode. There is a need to better match the choice of participation level and processes to the problem it attempts to solve.

These challenges hint at the fact that in spite of advances, the combined application of participatory processes and computer models to support environmental decision-making raises a number of theoretical and practical issues. The difficulties in filling the philosophical gap underlying qualitative and quantitative research and operationalizing their integration remain widely unexplored. Research is needed to close the gap between natural science models that simulate change and stakeholders’ experiences...
and perceptions of change. Moreover, the use of computer models may lead to a risk of developing fixed paradigms of the future through unified conceptual or computational models with generalized knowledge, and this counteracts one of the original aims of participatory processes of including a variety of knowledge types and backgrounds (Crane 2010). The integration of knowledge in PM exercises carries the danger of discouraging less accepted and more tacit forms of knowledge, which points towards the problem of managing and facilitating power within such processes. More generally, having the real power needed to affect the outcome of the process and translate findings into actions is often listed as a key impediment to PM. But this criticism might arise mainly from a focus on formal and relative power rather than an evaluation of outcomes from deliberative processes.

Moreover, environmental problems are the result of aggregate actions of numerous actors rather than a single decision-maker, a fact that also adds credence to the importance of the deliberative and social learning aspects of PM (Lane 2005). But these processes take time and can be fairly complex and thus are often not well managed because facilitators and participants are not aware that knowledge and practices have developed over time. This problem is often amplified by unrealistic expectations about what environmental stakeholder processes ought to be, and what they can achieve, resulting in ineffective use of resources and time. Stakeholders frequently struggle to balance requirements of their day-to-day job with their participation in deliberative processes. Similarly, scientists and modelers are also pulled in different directions when confronted with producing scientific outputs versus the needs of the process. The exploration and inclusion of local assumptions and knowledge as main inputs of the recursive PM process can frustrate researchers and modelers who might feel that this kind of process is not rigorous and robust enough. In addition, the modeler is cast into an arena that can be very different from the usual modeling environment, where he or she must contend with stakeholders’ agendas and the need to translate model outputs in ways that all stakeholders can understand. Models also tend to ignore the emotive side of decision-making, even though emotions play a very important role, especially when resource conflicts are involved. PM can provide a framework in which these conflicts can play out by simulating implications of choices in a safe environment, by organizing knowledge and contributing to the integration of scientific and nonscientific knowledge, and by catalyzing mutual learning processes.

**SEE ALSO:** Critical spatial thinking; Ontology: theoretical perspectives; Participatory action research; Public-participation GIS; Qualitative data, acquisition; Qualitative information: representation; Representation: dynamic complex systems

**References**


ecologyandsociety.org/vol15/iss4/art19/(accessed October 27, 2015).

Participatory video

Sara Kindon
Victoria University of Wellington, New Zealand

Participatory video, in both community development and academic research, has been growing exponentially since the late 1960s. It can be thought of as a field “coming of age” (Mitchell, Milne, and de Lange 2012, 10). Its diverse forms involve groups or communities in the production of videos that tell their stories, explore their concerns, and/or enable them to mobilize and take collective action. It is also sometimes known as process video, community video, or video-as-dialog and is used within participatory processes to raise awareness, share meanings, and facilitate discussion, often with distant stakeholders.

Participatory video processes may result in the production of documentary films, but participatory video differs from most forms of documentary filmmaking in three main ways. First, it emphasizes the ethics of access by placing the means of production into the hands of subject-communities, rather than filmmakers/experts. Second, it attends to the ethics of representation enabling people to self-represent according to their own aesthetic and cultural conventions, and, third, it foregrounds the ethical responsibilities of audiences by seeking to affect them, as a vehicle to effect change. Its origins and uses are entwined with participatory (action) research, participatory development, and indigenous media and research practices, although the latter remains generally unacknowledged in mainstream accounts.

The Fogo Method is frequently attributed as the formal origin of participatory video processes. Carried out by Colin Low and Fred Earle within the isolated fishing communities of Newfoundland in 1967, the process was innovative in many ways. Funded by the Canadian Film Board through a program called Challenge for Change, Low produced a number of short single-person “vertical” films as opposed to a longer, more wide-ranging “horizontal” documentary involving the intercutting of opposing perspectives. He enabled community members’ input into the focus, shooting, editing, and final content of the films produced, and emphasized the closing of the subject-feedback loop, screening the films produced to different island communities, as well as to distant academic and policy-oriented stakeholders, to generate discussion and dialogue. Such an orientation reflected earlier innovations in documentary filmmaking by American Robert Flaherty (Nanook of the North, 1922), British filmmakers involved in the 1935 documentary Housing Problems, the work of film-maker Raymond Garceau in New Brunswick and Gaspésie, Canada, in 1965, and the influence of British documentary filmmaker John Grierson within the Canadian Film Board at the time. However, the Fogo Method differed in its outcomes. The screenings of the films produced resulted in new filmic responses as interrelated stakeholders chose to use the medium to “talk to each other” across geographic distances, and to mobilize for local change.

The Fogo Method is credited with giving its participants a “voice” facilitating their empowerment. Over the next two decades, it influenced filmmaking processes across the world. Its spread was aided by the reduction in size, weight, and...
PARTICIPATORY VIDEO

costs of the recording and editing technology needed, and as video began to replace 16 mm film as the main medium of production, it became easier for nonexperts to transport and use. It was also supported by wider human and environmental rights movements across the world advocating for greater democracy and participation. In New York, Martha Stuart established the Village Video Network in 1984. She traveled to India to work with the Self-Employed Women’s Association (SEWA). Members of SEWA have gone on to produce more than 400 videos and have used them to advocate for women’s and labor rights.

Elsewhere, what has become loosely defined as participatory video emerged from diverse sources. In Tanzania, one influence was the approach of Soviet propagandist filmmaker Aleksandr Medvedkin, who in 1932 created and ran a film-train consisting of three railway cars carrying a film crew, production equipment, projection room, laboratory, and film-printing machinery. It was a self-contained film studio that could stay on location for months at a time reaching remote communities to produce critical films on their conditions. These films were shown locally to stimulate discussion and development initiatives. Also in East Africa in the mid-1930s, the Bantu Kinema Educational Experiment attempted to create a cinema produced by and for the local people. Medvedkin’s work inspired others, including French avant-garde filmmaker Jean-Luc Godard, who lent his 16 mm film camera to workers in a factory in Lyon in the 1960s. His practices were then taken up by video activists in the community television movement in the United States and Canada in the 1970s.

Further south, in Latin America, uses of video were informed by the critical pedagogy of Paulo Freire to enable people to document and take action against the oppressive aspects of their lives during civil war or dictatorship.

Throughout this period, equipment was still cumbersome despite the emergence of more mobile synchronous sound 16 mm film cameras like those used by Godard. It was not until the 1960s when portable video cameras were developed with an immediate playback function that participatory aspects of production became a more frequent possibility within project-based applications and academic research.

During the late 1960s and early 1970s, academics in various disciplines, particularly anthropology, began experimenting with participant-led forms of visual ethnography and data generation. Such initiatives embraced a more inclusive ethics of access and sought to understand the worlds of others through the generation of their own representations. At first they sought to capture “the native’s view” as somehow more authentic and “real” than that of the outside filmmaker – as a transparent window into a different reality. However, over time, scholars such as Jean Rouch, David and Judith MacDougall, and Eric Michaels embraced a more reflexive or intersubjective orientation to their work.

Rouch used the camera as an agent of investigation and to initiate ethnographic encounters. He advocated the practice of “feedback” whereby the first rushes of a film were shown to the people who were represented in it to invite their comments and ideas. This “audiovisual countergift,” as he termed it, acted as a stimulant for mutual understanding and participant dignity. He also envisioned a “shared anthropology” where the camera was placed in the hands of those usually in front of it. The MacDougalls developed a practice of “participatory cinema” in which they sought to make the filmic encounter – the process of production – visible to audiences by including discussions with participants about the making of their films and shots of the equipment being used. In Australia, Michaels used video with Aboriginal Walpiri peoples to
enable them to tell their own audiovisual stories via television. These practices called attention to the role of the researcher or filmmaker and to the politics of cultural production associated with the technology itself in shaping the products and representations produced (Pink 2007).

In the 1980s and 1990s, feminist, postcolonial, and indigenous critiques, coupled with the uptake of participatory and social learning methodologies, foregrounded the “participatory” dimensions of participatory video practice more prominently in both development and research applications. As mentioned above, advances in the technology also enabled greater participation in filming by nonexperts. As examples of successful practice around the world were shared, claims about participatory video’s ability to empower others and give them a voice led to the construction of a dominant discourse that was celebratory, romanticizing, and largely uncritical (White 2003).

Alongside these developments, the indigenization of visual media by people dominated by white eurocolonial settler populations such as in the United States, Canada, Australia, and Aotearoa, New Zealand, represents a different but related form of participatory practice. As a particular form of cultural activism, it has involved a political dynamic or intercultural negotiation implicated in a politics of identity and broader movements for cultural autonomy and political self-determination. One of the earliest and most striking examples of indigenous media use was by the Kayapo Indians in Brazil, who used video to document historical encounters with Brazilian state power and internal political events. In Mexico, the Chiapas Media Project/Promedios enabled indigenous communities to use video to draw attention to their plight and to challenge dominant media representations of themselves and their struggle. The international dissemination of their videos produced worldwide solidarity and political pressure on the Mexican government.

Through these uses of video, the ethical responsibilities of the audience emerged as a key, if undertheorized, element of various approaches to participatory and advocacy video (Corneil 2012). They also highlighted a key tension in participatory video practice: that of the promotion of a more objectified or fixed notion of social reality in order to communicate across or between cultures, while giving people more control over that process of objectification.

Since the late 1990s, visual and participatory research methodologies have become more mainstreamed within the social sciences and geography. The rise of digital technologies and online connectivity have sustained this trend and brought new logistical and ethical challenges. Two main strands of current scholarship are evident: first, the use of participatory video in discrete citizens’ media projects drawing on participatory development and/or critical pedagogy to facilitate the expression of participant agency and self-representation through video; second, participatory video as a research method, which can epistemologically challenge extractive social science approaches, and generate rich qualitative data for collaborative analysis. In some cases, academic engagements with participatory video bridge both project and research orientations.

Parallel to, and feeding into, both of these trends has been the growth of UK-based organization InsightShare as the most prominent provider of participatory video training around the world. Not only have they secured major contracts with international organizations like the World Bank and national governments such as Myanmar to integrate participatory video into their ongoing development programs, they have also partnered with British academics through the establishment of PV-Net since 2007. Funded by the Economic and Social Research
Council to bring researchers together to develop their capacity to integrate participatory video into academic research, InsightShare’s form of participatory video is now proliferating via PV-Net activities and within academic research in similar ways to the adoption of participatory rural appraisal techniques in the 1990s. For example, the majority of British-based geographers deploying participatory video since 2008 have been trained through InsightShare.

Throughout these developments, critical questions about the relationships between participation, video, and audience remain relatively undertheorized. Scholars rarely make explicit what they mean by participation, agency, voice, or representation, despite their contested meanings. Studies frequently imply that participation is evidence of agency and that expressions of agency demonstrate personal or social empowerment or transformation, “as if telling one’s story necessarily means taking charge of, or making changes to, one’s like and community or acting in the wider public sphere” (Low et al. 2012, 55). Yet the histories and politics of representation complicate understandings of the relationships between participation and the use of video.

In addition, there is increasing recognition that processes of filming, screening, and discussing video products simultaneously express and produce subjectivities, and that having subject-community or indigenous members operating the technology does not necessarily make the representations produced more authentic, respectful, or politically powerful. That said, many authors argue for the value in continuing to have so-called marginalized peoples carve out a discursive space for their own understandings and uses of audiovisual media, and that within this space, indigenous media is a specific kind of representational project.

These more recent developments in the field suggest encouraging efforts to better understand the relationships between affect and effect, aesthetic, artistic, and political expression, and how they are informed by the means of cultural production being facilitated. They also recognize that handing over the camera to participants and an ethics of access is only one part of a complex process. Researchers are starting to explore how the media logic of the technologies in use shape, limit, and fix what can be represented. The recent critical reappraisal of participation as tyranny has fueled a re-examination of the processes of facilitation and group work at the heart of participatory video (Kindon 2012). There is also much to be learned from indigenous media discourse where more explicit attention and acknowledgement has been paid to the highly charged intercultural dynamics and colonial continuities associated with the politics of seeing, researching, and representing (Deger 2006).

Questions are beginning to be asked about how we can document and theorize the intangible and/or affective aspects of people’s engagements with video, such as pleasure, fun, and frustration. In addition, there is growing awareness that not everything can be rendered visible, nor is there a clear and consistent relationship between the realization of a particular vision and the performance of agency or voice. In this respect, the work of long-time participatory video practitioner and academic Jackie Shaw (2012) is significant in its attempt to theorize the gaps between participatory video’s ideals and realities with respect to empowerment. And as the dissemination of participatory video products into distant spaces and online environments increases, there is much work needed to inform more appropriate ethical practices associated with consent and ownership. There is also a need to better understand the diverse audiences being both targeted and reached by participatory video products.
Out of this greater focus on praxis across different stages and spaces of interventions there emerges no singular “participatory video” nor a solution to its enduring tensions, only a call to remain vigilant to the contextual interplay of people, power, and audiovisual technologies. Making sense of how participatory video is continually remade in different places provides opportunities for further geographic inquiry.

SEE ALSO: Indigeneity; Mixed-method approaches; Participatory action research; Participatory development; Participatory modeling; Public-participation GIS; Qualitative information: representation; Social justice; Visuality

References


Further reading

Human dominance has been established for thousands of years and human activities have been changing the world prodigiously. Forests are cut down to make fuels and build houses; artificial rivers are dug to facilitate irrigation and transportation; interstate highways are constructed, breaking up vast fields, to make our lives more convenient. However, various environmental crises have arisen due to such human-induced changes. The forest may never grow back; tons of topsoil are lost; imported plants take up large area and original local species start to die out, and so on. By drastically disturbing natural habitats, human beings have greatly disturbed the ecosystem. The decrease in biodiversity and the emergence of different environmental problems have pushed humans, especially resource managers and scientists, to search for ways to better manage the environment.

The relationship between spatial patterns and ecological processes forms the overarching question in the science of landscape ecology. Under the premise that the change of landscape patterns interferes with critical ecological processes, scholars have put much effort into quantifying landscape patterns, which is essential before the pattern–process relationship can be understood. The landscape is composed of various elements, such as mountains, rivers, vegetation, buildings, and so on. According to the phenomenon being considered, these elements form homogeneous areas, that is, patches, which collectively make up a distinct pattern or structure of the landscape. Over the past three decades, landscape metrics have been developed and have become one of the most popular methods to quantify landscape patterns at different scales. In the beginning, indices from traditional community ecology were adapted to quantify the composition (what and how many patches). Since then, a number of new metrics have been proposed to add a new dimension of information: spatial configuration (position and arrangement of the patches) of the landscape. Specifically, the metrics computed for each patch are defined as patch metrics. Additionally, a set of metrics can be derived for all patches belonging to the same class, or for the entire landscape. Correspondingly, they are referred to as class metrics and landscape metrics, respectively, and thus are not included here.

McGarigal, Cushman, and Ene (2012) summarized most of the prevalent patch metrics according to which aspect of landscape pattern is measured. As such, patch metrics are put into five different categories.

1. **Area and edge metrics**: for example, patch area, patch parameter, radius of gyration (extent of the patch). Metrics in this category provide important measures of the edge effect.
2. **Shape metrics**: for example, perimeter–area ratio, shape index, contiguity index. The effects of the patch shape are also related to edge effects. For example, patch shape has been found to affect inter-patch processes, for example, animal migration.
3. **Core area metrics**: for example, core area. The core area is the area of a patch excluding a
certain buffer from the edge. Therefore, core area is a metric that takes patch size, patch shape, and edge effect distance into consideration simultaneously.

4 **Contrast metrics:** for example, edge contrast index. This describes how different the patch is from the adjacent patches with regard to one or more attributes.

5 **Aggregation metrics:** for example, Euclidean nearest neighbor distance, proximity index, similarity index. These metrics describe how dispersed (dispersion), fragmented (subdivision), or isolated (isolation) the patches are. Furthermore, by comparing a certain patch measure (e.g., area, shape) to the mean value within the class and the landscape, the deviation and percentile of one patch from the class and the landscape also form a set of important patch metrics to identify patches with extreme values.

Often, patch metrics are defined on the basis of thematic maps. Depending on the definition of thematic classes, different landscape patterns and structures may be derived. For example, if deforestation is the underlying topic, all of the tree areas will be defined as the same patch type, “forest,” whereas if the topic switches to bird habitats, it may be desirable to separate yellow birch and white ash as two different patch types. Nevertheless, when studying the problem of fish habitats, it is not necessary to define the ocean as patches. In summary, it is important to correctly define classes according to the question being addressed before patch metrics can be calculated.

Habitat or biodiversity conservation has been the top concern that patch metrics strive to address (Wu 2013). In this regard, studies have been carried out on all kinds of species, including mammals, birds, plants, fishes, insects, and so on. Specifically, patch metrics are usually factored in the ecological models as predicting variables to assess the effects of environmental change on biodiversity. Two examples that illustrate the implications of patch metrics are: (i) the increasing fragmentation of habitats, which has been found to be a major threat to biodiversity all over the world; (ii) the different responses of native and invasive plant species to the same landscape pattern, for which patch metrics are helpful in predicting the distribution of invasive species.

Land-use/land-cover (LULC) changes benefit from patch metrics study. Urbanization, deforestation, and cultivation are three main processes of human interference that are taking place in the environment. With comparison of the patch metrics for the initial and the changed LULC landscape, the change of landscape patterns is quantified – for example, how much land has been taken over by new buildings or constructions, how much forest has disappeared, and how much land has become new farmland. Besides the questions of “how much,” patch metrics facilitate the explanation of more important issues. For example, they can address the questions of whether urban areas are expanding by merging with neighbors in all directions or by growing in one direction; which scenario, cutting trees from the outer boundary of woods or fragmenting the forest by cutting trees, suffers worse deforestation; or what LULC type is more likely to be replaced by farmlands. To sum up, patch metrics provide more in-depth information than pure comparison of two LULC maps. Alternatively, prediction models are also engaged to predict future trends of LULC change. In the model-based approach, patch metrics are utilized as one set of predicting variables for models such as cellular automata and Markov chain.

Landscape regulation with the help of patch metrics received considerable attention recently.
Studies have revealed that different environmental components are influenced by various aspects of the landscape, which can be characterized by patch metrics. For instance, water quality is more related to landscape composition than to spatial configuration; the spreading of fire is well explained by fragmentation metrics and the severity of fire depends more on the composition of the forest; after a forest fire, a more fragmented landscape pattern is usually created; soil erosion is affected by watershed landscape patterns, and so on. Therefore, regulation issues regarding water quality control, fire risk reduction, and soil preservation benefit much from the utilization of patch metrics. Additionally, researchers started to understand microclimate (e.g., urban heat islands) using patch metrics as predictors. The findings have led scientists to believe that patch metrics are promising in helping to analyze the relationship between larger-scale (e.g., national, global) climate change and ecosystem functioning (Uuemaa and Mander 2013).

The potential to facilitate so many environmental research topics attracted researchers to create a variety of patch metrics. By the end of the twentieth century, a number of patch metrics had already been developed and convenient software packages produced. Despite the development of these metrics, three limiting factors still exist. As a result, a major question now concerns how to effectively relate the values of patch metrics to corresponding ecological processes without developing new measures.

First, scale of view is a widely discussed issue on landscape patterns and metrics. With the rapid development of remote sensing, the availability of global data at different spatial resolutions reveals the problem of scale. “Scale” here has two different meanings: grain and extent. Grain, or resolution, refers to the smallest spatial size of the data to be studied. Different patterns will appear at different grain sizes and the metrics will also change. Taking the edge effect as an example, when the grain size decreases, more details become visible and the straight edges may become sharpened. Therefore, comparing edge effect at different scales will not be very informative. Extent, as the name implies, refers to the area of study. Compared to changing the grain size, the effect of changing extent on patch metrics is rather more random. It’s often infeasible to deal with the scale problem by interpolation or extrapolation. To use the patch metrics wisely, for instance, to link the pattern to certain processes, or to compare the patterns at different times, it’s important to keep the scale at the same level.

Second, although the number of available patch metrics is huge, they are not mutually independent. Great redundancy exists between different metrics. Some metrics are inherently correlated by design. For example, the perimeter–area ratio is the ratio of patch perimeter and patch area. Hence, these three metrics are completely correlated. However, all of these correlated metrics are maintained because each of them can be more convenient to represent the landscape, depending on what aspect is to be emphasized. On the other hand, some metrics may be empirically dependent under certain landscape conditions, such as scale, dominant patch size, number of classes, and so on. When selecting metrics, it’s better to avoid using the highly correlated metrics simultaneously. Techniques such as principal component analysis and factor analysis are often applied to reduce the redundancy in the set of metrics used.

Last, patch metrics themselves stop at being a quantitative description of the landscape pattern at a certain time point. To be more useful, the metrics should be applied to compare landscape pattern at different times or in different scenarios. The application in LULC change is one example of utilizing the metrics. However, usually only a few discrete time periods are
chosen to study the change and the majority of temporal dynamics are still omitted. Patch metrics can also serve to predict variables in modeling. However, the relationship between the metric value and a certain pattern may not be one-to-one. The pattern represented by patch metrics, the pattern that causes certain processes, and the pattern that is seen by humans can all be different from one another. Therefore, further exploration is still needed of how to better relate patch metrics to the function of patterns.

**SEE ALSO:** Biodiversity; Land-use/cover change and climate; Landscape ecology; Microclimatology; Scale

References


Path dependence

Jinn-Yuh Hsu
Yu-Kai Liao
National Taiwan University

The concept of “path dependence” emphasizes that the set of decisions one can make is influenced by the decisions one has made in the past. More specifically, it argues that certain historic events would disproportionately cause later conditions. One of the most well-known examples, offered by economist Paul David, is the adoption of QWERTY keyboard. Although the layout of QWERTY keyboard is not the most efficient one in comparison to many other keyboard models that have existed in history, it has become the industrial standard (David 1985). Paul David explained that this is because in the initial stage of the development of typewriters, most typists were trained to learn and memorize the arrangement of the keys in the QWERTY layout. And, as it is difficult to change typists’ habit, more and more typewriter manufacturers adopted the QWERTY keyboard. As a result, the less efficient keyboard model became the mainstream. In short, path dependence emphasizes that economic configurations cannot be explained solely by some timeless logics such as cost or benefit; history matters in the shaping of economic structures.

Since its introduction in the late 1980s, the concept of path dependence has undergone several rounds of elaboration and modification. One important contribution was made by economist William Brian Arthur, who, in Increasing Returns and Path Dependence in the Economy (1994), proposed four concepts – large initial fixed set-up costs, dynamic learning effects, coordination effects, and self-reinforcing expectations – to further elaborate on the relationship between history and the economy. “Large initial fixed set-up costs” refers to the fixed capital that a company has to expend prior to the production of a particular product. To cover this cost, most companies, in the initial stage of the production at least, will focus on increasing production outputs, thus bringing inertia to changes that may make the production process more efficient. “Dynamic learning effects” emphasizes the importance of learning by doing and learning by interacting, which would increase human capital. “Coordination effects” argues that cooperation between firms confer advantages to going along with other firms taking similar actions. Lastly, “self-reinforcing expectations” refers to the situation where increased prevalence of a particular product, technology, or production practice enhances beliefs of further prevalence. As a result, more suppliers and consumers would invest in the common standard, and products will fulfill the expectation and reduce the risk of investing.

David’s work on the economics of technology (the adoption of the QWERTY keyboard being his most famous example) and Arthur’s thesis about increasing returns together inspired a generation of economists to investigate how economic structures are shaped by historical accidents and contingencies. This strand of scholarship is often referred to in literature as evolutionary economics (EE). It is also worth noting that in economics insights from path dependence theory encouraged some scholars to bring geography into the study of the

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0574
PATH DEPENDENCE

economy. In *Geography and Trade*, for example, Paul Krugman (1991) points out that increasing returns is one of the determining factors behind trade. This approach to trade is in sharp contrast

**Box 1**

The case that Krugman examined was the formation of the Rust Belt in the United States. He developed a simple model that shows that the interaction of increasing returns, transportation costs, and demand can give rise to the spatial concentration of the manufacturing sector. In the early history of the United States, where manufacturing was marked by few economies of scale and where transportation was costly, no strong geographical concentration could occur. As the country began its industrial transition, manufacturing grew in areas that contained most of the agricultural population outside the South. As this economy of scale increased, transportation costs fell, and the share of the population in nonagricultural occupations rose. As a result, the initial advantage of the manufacturing belt was locked in.

to Ricardo’s notion of comparative advantage, upon which mainstream economics has been based on. In economics, literature works by Paul Krugman and his followers are often referred to as *new economic geography*.

The influence of path dependency theory on other social sciences disciplines – human geography, sociology, and political sciences in particular – has been no less significant than it has been on economics. In human geography, the introduction of the path dependence theory is closely associated with the emergence of the literature of *evolutionary economic geography* (EEG), which is our primary interest here.

**Evolutionary economic geography**

Evolutionary economic geography concerns the evolution of economic landscape, or the transformation of the spatial organization of production, distribution, and consumption. Two lines of inquiry characterize this body of research. The first is the study of the evolution of economic landscape in a given location. For example, in his study about the coal and steel industrial complex in Rhur, Grabher (1993) identified three types of “lock-in” – functional (based on hierarchical relations between firm), cognitive (consisting of a common world-view), and political (a thick and dense institutional structure that hampers restructuring) – to explain why the region become inflexible, unable to absorb new ideas, and ultimately unable to respond to competition.

The second is the study of the novelty of economic landscape across locations. Here, the scholar’s primary interest is in the spatial logic of the emergence of new industries. Boschma (1997), for example, argues that the requirements of newly emerging industries are often discontinuous with the environment of pre-existing industries. This is because previous local context does not necessarily provide an advantage to stimulate for the development of new industries. The implication of this proposition is that regions that are lagging behind are often more attractive to new industries. The case that Boschma used to develop this argument is the transformation of industries in Belgium since the nineteenth century. He carefully examined the locational patterns of five industries: coal-based iron making, cotton, electronic engineering, automobile, and microelectronics. He found that the development of electrical engineering and automobile industries (often referred to as the “second industrial revolution”) require specific technological knowledge, skill, and capital.
These requirements are very different from those of coal-based iron making (which required proximity to the coal and ore mines) and the cotton industry (which required a pool of experienced labor and skilled entrepreneurs). As a result, the electrical engineering and automobile industries preferred to locate in particularly old industrial regions or lagging regions that could provide the specific technology knowledge, skill, and capital.

Without doubt, prior to Boschma’s research, economic geographers had investigated the issue of industrial locations. In *The Capitalist Imperative*, Michael Storper and Richard Walker (1989), for example, proposed the concept of “windows of locational opportunity” to bring political economy into industrial location theory. They pointed out that the geography of capitalism is uneven and inconstant. Through technical innovations, organizational changes, and labor intensification, capital is able to escape from the past and creates a new industrial landscape. This means that new industries would not necessarily develop in established industrial regions. What is new about Boschma’s research, however, is that it points out that every place has its history, and the history leads to different regional capacity and possibility. An analysis of the spatial logic of the emergence of new industries thus has to consider historical contingency.

Between EE and EEG, there is thus a clear difference in terms of research agenda. Evolutionary economic geographers investigate how place dependence (as a form of path dependence) shapes the trajectory of economic development, an issue that is largely absent in the scholarship of evolutionary economics. Such a unique research focus reflects an important theoretical stance that has been widely accepted among human geographers: that space and place are not just stages upon which activities and events occur, but also (and more of) driving forces that bring about social, economic, political, and cultural transformations.

Within geography, EEG is often seen as a critic to *institutional economic geography* (IEG), which also studies the spatial logic of industry. Researchers regarded institutions as carriers of geographical–historical context and sociocultural product. Ash Amin and Nigel Thrift (1995), for example, use the concept of “institutional thickness” to emphasize the importance of institutional decisions, such as strong institutional arrangements, high level of interaction among institutions, well-defined structure on institution, collective mobilization, and the spatial organization of industries.

For them, the transformation of industrial landscape cannot be understood without attention to instructional transformations. EEG criticizes the concept of institutional thickness for lacking a clear definition and being difficult to measure. In general, there are two main
differences between the two approaches. First, while EEG is concerned about the dynamics of the changing economic landscape during a particular period of time, IEG is less concerned with the issue of change. Second, while EEG emphasizes the importance of the routines of firms (which are shaped by past experiences and learning by doing), IEG argues that communities and territories produce a macro-context, which influences on the routine of firms and guides the firms to work.

Without doubt, attention to history, or inquiries about spatial transformation through time, existed in human geography long before the introduction of path dependence theory. The Marxist turn in human geography during the 1970s had drawn many scholars to the approach of geographical–historical materialism, which views space as the product of historical processes. In the broadest sense, Marxist geographers see an existing spatial structure (e.g., uneven development between the developed world and the developing world) not as a state of equilibrium brought about by some abstract economic laws (e.g., the invisible hands of supply and demand) but as the product of the dynamics of capitalism. David Harvey (1982) and Neil Smith (1984), for example, emphasize uneven development as an inevitable condition under capitalism. To solve the inherent problem of overaccumulation, capitalists have to constantly create profitable terrains for surplus production. This can be achieved through, for example, induced unemployment, technological innovation, immigration, or exporting capital. If all these methods failed, capital accumulation is blocked and capital can be devalued or physically destroyed. The spatial implication of this dynamic is that development in one place is often interconnected with underdevelopment in another place. Yesterday’s boomtown may thus become a ghost town tomorrow (as in the case of Detroit), and previously deprived regions may become new centers of production (as in the case of post-Mao China) or speculation. Along a similar line, Doreen Massey, whose research focuses on the restructuring of industries in postwar Britain, drew attention to the spatial division of labor (Massey 1984). Her main argument was that the economic landscape inherits the legacies of its past development and that these legacies exert an influence on its present and future development.

There exists, however, one profound difference between the ways in which Marxist geographers and evolutionary economic geographers treat history. While Marxist geographers’ concern is the meta-narrative of political economy, many evolutionary economic geographers, as Martin and Sunley (2010) pointed out, build their research upon universal Darwinism, complexity theory, or panarchy theory (Boschma and Frenken 2006; Martin and Sunley 2006). In other words, not all the evolutionary economic geographers are concerned with the political economy.

**Geography’s contribution to path dependence theory**

The relationship between human geography and path dependence theory is not unilateral. While path dependence theory inspires geographers to ask new research questions, geographers have also, in turn, helped shape the contour of the theory. One good example is Martin and Sunley’s (2006, 2010) critique and modification of the basic path dependence model.

The initial path dependence model developed by Paul David and William Arthur was composed of four difference phases: pre-formation, path creation, path lock-in, and path dissolution (Figure 1). The pre-formation phase, also often called multiple equilibria, refers to the situation in which a lot of choices and possibilities are available
Box 3

Within institutionalism, there are also discussions about incremental changes. In *Explaining Institutional Change: Ambiguity, Agency, and Power*, James Mahoney and Kathleen Thelen (2009) argued that significance of gradual evolution of institutions should be considered and the endogenous forces are also the driving force. Mahoney and Thelen viewed institutions as distributional instruments that raise tensions between different groups. Different actors will compete or ally with others to create, maintain, or revise institutions according to their interests. Actors change institutions through four modes of institutional change – displacement, layering, drift, and conversion. The occurrence of these four modes of institutional change is dependent on political context and institution form not only directly induces institutions to change but also shapes the emergence of the type of dominant change-agent and the strategies which the agent adopted.

(or could be discovered). The path creation phase refers to the situation in which a certain path is formed as the result of random events or historical accidents, and along the path new technologies, organizations and institutions are developed. The path lock-in phase refers to a rather stable condition (as the result of network externalities and increasing returns) in which a certain economic system reproduces (or being locked in) the conditions of the chosen path. Lastly, the path dissolution phase refers to the destabilization of the lock-in and the decline of the path as the result of some “external shocks,” such as innovation.

![Figure 1](image-url) The basic path dependence model developed by Paul David and William Arthur. Source: Martin and Sunley (2010).
of technology, competition, or institutional change.

Martin and Sunley criticized the basic path dependence model for being, first of all, not evolutionary enough. They pointed out that, in the basic model, history matters only in the initial stage. The model implies that there would be a state of equilibrium once an economy is locked in certain practices. Martin and Sunley insisted that the evolution of an economy should be seen as a continuous process rather than a series of states of equilibrium. Second, they questioned the presumption that there exists a phase of development that is free of pre-existing conditions (as how the phase of pre-formation was described). Martin and Sunley emphasized that economic structures are always shaped by pre-existing technological structures, knowledge, and competence. Lastly, they criticized the emphasis on the influence of “external shocks,” which, according to the basic model, would result in the decline of a certain path. Martin and Sunley pointed out that the concept of “external shock,” on the one hand, denies the possibility that endogenous forces can also lead to significant changes while, on the other hand, it fails to take into account incremental changes and modifications (which can be endogenous or exogenous).

The alternative model that Martin proposed emphasizes that the environment for the emergence of new technologies and industries can

Figure 2  Regional path dependence model developed by Martin. Source: Martin (2010).
be either constraining or enabling (Figure 2). A constraining environment would often lead to a “static state,” where selected technologies and existing structures and knowledge of the firms tend to last and reproduce themselves until there comes an internal or external shock. An enabling environment, in contrast, would make a selected path more dynamic. Here, Martin elaborated how layering, conversion, and recombination effects would lead to incremental, path dependent evolution and the renewal of industries or technology. The layering effect refers to the addition of new rules to an existing institution, which allows the institution to change incrementally. The “conversion effect” refers to the addition of new rules (or the abolition of the old) that would convert an institution’s original logic of operation. Lastly, the “recombination effect” argues that “any particular existing social-political-economic structure is a system of resources and properties that actors can recombine and redefine to produce a new structure” (Martin 2010, 15).

Martin used the case of the development of the high-tech cluster in Cambridge, UK, to further elaborate the proposition that economic evolution is a process instead of a series of states of equilibrium. In the early 1960s, the Cambridge consultant proposed to recruit academic staff from the University of Cambridge to industry. The result of this action was the development of a high-tech cluster in the region, which includes computing and science instrument, software, and bio-tech companies. The formation of this cluster greatly contributed to the region’s economic growth. Based on this case, Martin argues that path creation frequently emerges from existing industrial regions, which allows innovation to happen. Innovation thus is place-specific and locally contingent rather than accidental or random.

SEE ALSO: Institutions and development; New cultural geography

References


PATH DEPENDENCE


Broadly defined, patriarchy refers to social structures governed by hierarchical power relations that privilege masculinity over femininity, and permit the domination, oppression, and exploitation of women by men (Walby 1990). Derived from the Greek words patria (family) and archy (rule), patriarchy initially referred to the autocratic rule of male heads of household and older men in government. Western, radical, second-wave feminists deployed the concept of patriarchy in critically and politically strategic ways to reveal the universality of a distinct social system that disadvantages all women irrespective of other forms of social stratification. Second-wave Marxist and socialist feminists theorized patriarchy as a product of the overarching social structures of capitalism that govern class and gender relations; while liberal feminists drew attention to the “small-scale” deprivations women experience (e.g., in employment and domestic division of labor) (Walby 1990). Such theorizations of patriarchy, however, have been criticized, not least by women of color, for producing totalizing and tautological effects that presuppose what patriarchy defines, as well as for producing ahistorical and acultural generalizations that overlook nuances and variations in power relations (Kandiyoti 1988). As a consequence, overt use of the term has waned and some of its meaning is now conveyed with terms (such as “male domination,” “sexual inequality,” and “gender asymmetry”) that carry less essentialist conceptual baggage.

In the 1990s anti-essentialist poststructural and postmodern approaches to feminism complicated singular concepts of gender and the dichotomization of gendered individuals into women and men. Critiquing the neglect of differential power relations between women and men, and patriarchy’s inherent heterosexism, these studies also foregrounded the forms of agency women employ to resist systems of male power and to “bargain with patriarchy” (Kandiyoti 1988). The local specificity of such bargaining strategies reveals the way in which patriarchy is entangled with the historical, transterritorial, and transnational processes and particularities of capitalism and colonialism.

Extending out of debates in Antipode initiated by Foord and Gregson (1986), feminist geographical scholarship examines the ways in which patriarchy operates across a range of spatial scales from the body to the nation-state: from the construction of an abstract modern liberal subjectivity to gender inequality in law, as well as constructions of ethnic and national identities. A state can be interpreted as patriarchal in that it is neither hegemonic nor monolithic, but rather a mediator of many powers that shape women’s lives. Within the context of Western liberal nations, secularism emerges as a crucial component of patriarchal state sovereignty by serving as a pretext for the marginalization of communities based on religious background, ethnicity, and race. In Turkey, for example, Arik (2015) reveals how secularist national security discourses work to control the public representation of Muslim women through regulations about “contentious” religious embodiment (e.g.,
Patriarchy

headscarf) and practice (e.g., gender segregation). For feminist geographers, the body remains a significant spatial scale of intersectional analysis that permits an intricate understanding of the insidious ways in which patriarchy continues to operate through the political, social, cultural, and economic institutions of society.

SEE ALSO: Feminist geography; Gender; Heteronormativity; Reproduction: social; State, the

References


Patterned ground

Carol F. Sawyer
University of South Alabama, USA

Patterned ground (PG) refers to a collective set of surficial geometric landforms associated with periglacial environments located in polar, subpolar, and alpine tundra locations. Permafrost is not a requirement for PG; however, it may be present at the site. The landforms are classified into five shapes: circles, polygons, nets, steps, and stripes. Typical classification of PG is based on visual inspection rather than determination of formation processes, allowing both relict and active landforms to be categorized and compared. The surface of the landform may be flat or raised above the surrounding surface, either along the edges or in the center—or middle in the case of stripes—of the landform; however, a landform with an elevated center is not automatically PG, as other landforms have raised centers, including frost boils, mima mounds, and hummocks or thufur (see Goudie 2004). PG is categorized into sorted or nonsorted features, with sorted landforms exhibiting a surficial appearance of a progression from the smallest to largest material sorted by size from the center outward, with the inside containing sediment or a combination of small pieces of gravel and sediment, and the outer edges of the landform (sometimes referred to as the “gutter”) containing the largest clasts. Nonsorted features do not exhibit any size organization, with different sizes of clasts and sediment mixed together. The geometric appearance of PG is influenced by the slope of the surface, with a general positive relationship existing between landform shape and its slope: as slope increases, PG features become more elongated in appearance. Circles are located on the flattest slopes, whereas polygons, nets, steps, and stripes will be found, in the order listed, on slopes of increasing inclines. Washburn (1980) noted that nonsorted features may not follow this progression as predictably as sorted features. One of the most well-known examples of PG is in Spitsbergen, an island in the Svalbard archipelago, Norway (78°45′ N 16°00′ E), where sorted circles with raised edges measure 3–5 m in diameter (Hallet and Prestrud 1986). These distinctive large sorted circles were featured on the cover of Science in conjunction with an article on computer modeling PG formation by Kessler and Werner (2003).

Washburn’s (1956, 1980) series of works produced the seminal classification scheme for periglacial PG based on visual inspection of landforms. Washburn’s (1956) goal in creating the original classification scheme was to eliminate duplicate terms and to consolidate the terminology of PG features. He later added further categorization by noting that the features can be classified as “small patterned ground” when smaller than 1 m (Washburn 1980). In the years following Washburn (1980), “miniature patterned ground” was added to further distinguish geometric features measuring smaller than 20 cm. The evolution of this category began appearing in the literature in the 1980s, with Wilson and Clark (1991, 369) finally outlining quantitative parameters defining miniature patterned ground, specifying that landforms have a “mesh diameter or stripe width not exceeding ca. 20 cm.” These parameters are still in use.
today by both alpine and polar researchers to classify the smallest forms of PG, resulting in a three-prong classification scheme to describe PG: (i) shape: circles, polygons, nets, steps, or stripes; (ii) organization: sorted or unsorted; and (iii) size: large, small, miniature.

In a periglacial environment, types of PG encompass a majority of the landforms but are not the only landforms present, as pingos, palsas, and blockfields or felsenmeer are also found in alpine and polar and subpolar regions (Washburn 1956). Geometric landforms found outside of periglacial environments may or may not be considered PG, depending on the researcher. Goudie’s (2004) definition of PG included gilgai (a landform whose process is attributed to the shrink–swell properties of clay), polygons on playas (formed from salt deposits or shrinkage from water evaporation), and brousse tigrée or tiger brush (alternating light and dark bands of vegetation that give the impression of stripes on the surface) all within the suite of PG landforms.

Processes responsible for and that maintain PG are associated with frost action, a general phrase referring to processes experiencing water freezing in the subsurface in response to cold temperatures, including freeze–thaw cycles, needle ice formation, frost weathering, and frost heaving. Freeze–thaw cycles and other processes encompassing the freezing of water are the result of a phase transition of liquid (water) to a solid (ice) with the force of that process driven by the expansion of the water when frozen (to approximately 8–9% greater volume than its liquid state). Frost action is consistently cited in studies involving currently active PG in polar and alpine regions. However, no single process is known to start the initial formation of PG, nor does a single process control all continued appearance of PG. In addition, processes responsible for the formation of PG may not be the same process actively modifying active PG (Washburn 1980).

Consequently, determining why PG initially formed at a location is difficult, as is the length of time PG was present prior to discovery. Washburn (1956) identified 19 possible mechanisms or processes responsible for PG formation; he did not associate a particular type of PG (e.g., sorted circles) with a mechanism. The list included processes associated with the freezing of water (e.g., seasonal frost, frost heaving, primary frost sorting), salt (salt heaving), and water (rillwork). Washburn (1980) elucidated on his original list by categorizing the mechanisms into two basic processes: essential surface cracking and nonessential surface cracking, both of which are linked to sorted and nonsorted PG landforms (Table 1). Surface cracking is caused by the contraction and expansion of the ground in relation to active processes, which may not necessarily be frost-related (e.g., the expansion and contraction of clay). Desiccation cracking is the result of soil drying and contracting, and is associated with small and miniature PG, both sorted and nonsorted features, containing silt and clay (Washburn 1980); however, the amount of either silt or clay needed for activity has not been specified. Seasonal frost cracking has also been associated with sorted and nonsorted landforms (Washburn 1980). Unlike desiccation cracking, silt and clay do not need to be present; however, moisture is required. Frost cracking begins each fall, and on the ground thawing in the spring the cracks may disappear (Washburn 1980), making confirmation of the process difficult.

Freeze–thaw cycles, cryoturbation, and frost heaving have typically been attributed to active nonsorted PG, with only frost heaving specifically mentioned by Washburn (1980) as part of his 19 mechanisms. These three processes are similar to each other as they all involve freezing temperatures, with cryoturbation attributing the churning of sediment or soil, as is seen in cryosols, to nonsorted PG movement. Washburn
Table 1  Washburn’s mechanisms for PG formation.

<table>
<thead>
<tr>
<th>Essential surface cracking</th>
<th>Nonessential surface cracking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desiccation</td>
<td>Frost heaving (differential)</td>
</tr>
<tr>
<td>Dilation</td>
<td>Mass displacement (nondifferential)</td>
</tr>
<tr>
<td>Frost action along bedrock joints</td>
<td>Mass wasting (differential)</td>
</tr>
<tr>
<td>Permafrost</td>
<td>Primary frost sorting (nondifferential)</td>
</tr>
<tr>
<td>Salt</td>
<td>Rillwork (nondifferential)</td>
</tr>
<tr>
<td>Seasonal frost</td>
<td>Salt heaving (nondifferential)</td>
</tr>
<tr>
<td></td>
<td>Thawing and eluviation (nondifferential)</td>
</tr>
</tbody>
</table>


(1980) listed eight processes where surficial cracking was not essential, including frost heaving (Table 1). Within that list, he created subcategories of differential and nondifferential, thereby noting that some processes may occur at unequal rates. Differential frost heaving, where the ground heaves in unequal amounts across a surface on freezing, can explain the uneven heaving of measurement tools such as nail and wooden dowels placed in PG. Additionally, differential frost heaving may be responsible for surficial clasts moving, turning, and becoming buried as the surface expands at different rates, causing sediment on the edges of the expansion to fall to the side or be buried by other sediment.

Processes responsible for PG formation studied by other researchers since 1980 include convective cell movement, self-organized criticality (SOC), and chaos theory. For example, Kessler and Werner (2003) developed a computer model based on SOC to study the hypothesis that sorted circle formation can occur through convective cell movement. Kessler’s and Werner’s (2003) study did not provide the seminal answer to PG formation, nor did it eliminate any of Washburn’s proposed mechanisms; however, the study illuminated a continued central quandary in PG research: the lack of a single formation mechanism.

The study of PG continues today and should be considered a topic with great research potential. For example, the daily, seasonal, or annual rate of movement within PG features is not well documented in periglacial environments, but it should be, considering that the change in the amount of activity of PG in response to climate change cannot be made without knowing current movement rates and condition. In addition, as resolutions of remote sensing platforms increase, the ability to study PG in remote polar and alpine regions over greater temporal scales continues to increase, opening up several avenues of research of PG. The use of remote sensing, specifically multiseasonal imagery, would be a great help to many researchers, as field access to many sites, particularly high alpine locations, is limited during the winter months. The study of PG tends to focus on processes for movement or the amount of activity in PG; however, one key research topic of the future will be the interaction between vegetation and active PG. The interrelationship between active PG and vegetation growth, and establishment is currently understudied, as a majority of PG research occurs in polar and subpolar regions whereas interaction between vegetation and PG is most notable in the alpine tundra, where treeline may advance with climate. Past studies have shown that the movement of
material in sorted and nonsorted PG prevents seedling establishment at the alpine treeline in the western United States. As climate changes, PG in the alpine regions should be observed for the effect its movement has on the treeline, and vice versa (Resler, Butler, and Malanson 2005).

SEE ALSO: Periglacial processes and landforms; Treeline ecotones

References


Peace

Sara Koopman
York University, Canada

Peace is all too often understood as being simply not war. It is also,ironically,widely used to justify violence and oppression. Peace for some means pacification through military victory. For others it means free reign for corporations. Yet even if it is defined only as a lack of direct political violence between states, peace is never clearly distinct from war. It is widely argued that, in effect, war is inside peace, shaping everyday political life, institutions, and sociospatial order. But the norms of armed conflict have also changed, and war no longer has a clear beginning or end in time or space, and the lines between civilian and combatant are also increasingly blurred.

Peace, then, is not a separate end point to achieve in time or space. Peace also happens inside war, not only in peace zone enclaves, but in everyday peace building by all sorts of actors. But, whether or not it is made in the midst of armed conflict, peace is always precarious and must be constantly remade.

The hegemonic understanding of peace among elites is of a liberal peace, although the liberal descriptor is rarely used by those who believe that (neo)liberal democracies will not go to war against one another. This belief is closely linked to the idea that (capitalist) development will bring peace. Paradoxically, though, rushed liberalization can be deeply destabilizing.

The growing belief in the “West” after World War II in a liberal peace led to the establishment of peace studies as an academic field. There are now encyclopedias of the field that cover various sorts of peace. The distinction most widely used in peace studies is between positive and negative peace. In a letter from a Birmingham jail in 1963, Martin Luther King Jr defined negative peace as “the absence of tension” and positive peace as “the presence of justice” (King 1991). In peace studies these terms are widely associated with Johan Galtung, who in 1958 founded one of the first and the most well-known peace institutes, the Peace Research Institute of Oslo (PRIO). In 1964 Galtung argued that negative peace was the absence of direct violence, that is, bodily harm, and that positive peace was the absence of structural violence, that is, social structures with life-shortening consequences. Thirty years later Galtung (1996) added a third type of violence, cultural violence, which he defined as the ideas used to legitimize both direct and structural violence. In Galtung’s triangle, structural violence leads to cultural violence, which in turn leads to direct violence. Negative peace is the absence of direct violence (although not the absence of all conflict). Positive peace, Galtung now argues, is the absence of both structural and cultural violence. Ironically, though, even in Galtung’s definition of positive peace, it is defined by what it is not rather than by, as King defined it, the presence of justice or of life-affirming values and structures.

Ideas of a positive peace can even be used to promote a liberal peace and hide an economic liberalization agenda. In 1997 Dietrich and Sützl, working at the Burg Peace Centre in Austria, put forward “A Call for Many Peaces,” in which they argued that the vision of the “one peace” has long been tied to an understanding of development born of the civilizing mission. They write...
that, without being based in “concrete places,” peace is simply an abstraction, and to argue for one peace is an act of intellectual violence that can lead to other violences (Dietrich and Sützl 1997; emphasis added). They imply that peace can be conceived and achieved only on a small scale.

Place and Scale are the classic remit of geographers, and it would seem that such a call would inspire a turn by peace studies to geography. That has been slow to come, but perhaps this is because, until recently, there was not much work in geography that explicitly used the term. Yet geographers are increasingly turning to the study of peace, and making similar calls to understand peace as varied across place, space, and scale, and as shaped in and through the spaces and times through which they are made, as they too shape those spaces.

Even in one place, for one group of people, peace is not a static end point but, rather, is always changing and must be made anew daily. Geographers now write widely of peace as a process rather than a steady state, and of peace as having different meanings across time and place, as Loyd (2012) argues in her useful overview of work in the field. (For an earlier and also essential genealogy of work by geographers on and for peace see Mamadouh 2005.)

Peace is always situated. It is made in some way but also somewhere for some people. Haraway (1992) argues that it is because knowledge is always situated that we need to enter into noninnocent conversations that open us to understanding and alliances with others. Likewise, it is because peace(s) are situated that it is important to draw connections between them to build more just peace(s). Geographers are well placed to connect, as Williams and McConnell put it, situated knowledges of peace. They call for geographers to research more “sites and scales to show how peace is differentially constructed, materialized and interpreted through space and time” (Williams and McConnell 2011, 930).

Does this work actually have to use the word “peace”? Williams and McConnell “propose a more expansive and critical focus around “peace-full” concepts such as tolerance, friendship, hope, reconciliation, justice, humanitarianism, cosmopolitanism, resistance, solidarity, hospitality, care and empathy” (2011, 930). The notion of peace is so contested that other terms are often used instead. Peace is understood not only in an entangled relation to war, but also in relation to the (also contested) concepts of development, security, and human rights.

If peace is broad enough to encompass all of this, it might be questioned whether the term is useful. The danger is that, if the term is left undefined, it is too often assumed as a universal across time and place, and sentimentally idealized as either simply not-war or as all that is good. But the benefit, as Williams and McConnell have argued, is that using a “richer theorisation of peace” as both messier than and more than not-war can, as they put it, “intersect and engage” with research agendas across geography. Loyd (2012), for example, makes a strong argument for geographers of peace to learn from the ways these broad connections are made by the antiviolence movement, and in particular the work being done on and against racial state violence within the United States and Canada.

Framing research in terms of peace allows for a different engagement with scale. Security is traditionally understood to be kept at the national scale, while human security is measured at the individual scale but generally understood to be kept by nations, intergovernmental organizations (IGOs), and nongovernmental organizations (NGOs). Peace, meanwhile, can be experienced as both intimate and global. Peace can be created
at an individual, family, neighborhood, community, and other scales, and use of the term can foster seeing these scales as intertwined.

It may seem odd to turn to geography for peace, for geographers have long offered advice to the prince and justification to the great powers for their colonial exploits. Ó Tuathail (2011) argues that war made geography, and gives a history of how war has spurred the academic institutionalization of geography since the Franco-Prussian War of 1870–1871.

Even so, Ó Tuathail argues that the discipline of geography is not “beholden to battlefields,” even though it has been shaped by them (2011, 478). Many geographers today are resistant to and critical of the discipline’s military ties, but those ties continue to be strong. These have come under greater scrutiny after the Bowman expeditions were accused by research subjects in Oaxaca, Mexico, of not divulging their US military financing, yet the board of the American Association of Geographers (AAG) has voted not to investigate that issue or to form a commission to discuss the relationship between geographers and the military.

Given that the discipline of geography has long been a tool used for violence and domination, it may seem odd to turn to it as a way to study and build peace(s). Yet geography is well placed to study and build peace precisely because war and peace are entangled, and both are inherently spatial. Geography itself can also be enriched by a focus on peace, for if war and peace are intertwined we cannot fully understand the making of space if we focus only on how war shapes space.

After a series of edited collections on war and peace which focused much more on war than on peace, and tended to leave peace undefined, the most recent collection is entitled simply Geographies of Peace (Megoran, McConnell, and Williams 2014), and takes great care to problematize the term.

Geographies of peace can take many forms, but a traditional global map of peace is problematic. Dan Smith (2003) has attempted to map peace over the years in several iterations of his Atlas of War and Peace, which has various versions of maps of war and conflict, overlaid with things like poverty, rights, and political systems – but only two maps of “peace,” limited to negotiated peace processes and militarized peacekeeping. In 2007 the organization Vision of Humanity developed an annual Global Peace Index. It includes an interactive online map that now has 23 indicators, which have changed over the years. Yet both what is tracked, and how, are contentious. Colombia, which ranked 146 out of 162 ranked countries in 2015, exemplifies many of these issues. The number of dead and disappeared in Colombia overall, as well as which to attribute to the conflict, are heavily disputed. It is often not safe to report deaths. The index does not track violence against women, unionists, journalists, or any other targeted violence, such as by sexuality or race. But, then again, even if it did track these, numbers would vary wildly depending on systems for reporting and the safety of doing so. These and other issues with the data are also a concern with arguments that violence has declined globally.

The Global Peace Index has been focused on a negative peace, but did begin to include positive peace in its report in 2015, although this is not yet included on its map. For this purpose, it uses eight indices, including a sound business environment and levels of corruption. It does not track access to food, health, or education. Yet, even if it did, it would be mapping one peace, imposing one definition of peace equally across time and space. These concerns were raised at the 2015 AAG meeting, during two panel sessions discussing the possibility of a future atlas of peace. Creative proposals were presented, including crowdsourcing
online maps that would show various peaces as ongoing processes.

Maps of peace may work best at a small scale, but peace is not necessarily only to be found in the local, nor must it be global. Peace is built both everyday by all of us in all of our interactions, and in high-level international peace negotiations. Peace means different things at different scales, and different groups, and at different times and places. Peace is not the same everywhere any more than war is. When peace is portrayed as a mythical singular, it can become so abstract as to be seen as unobtainable. Or it can become so unspecified that it can be attached to violent pacification. Instead, geographers see peace as a located and spatial process – and, as such, as necessarily plural.

SEE ALSO: Human rights; Peacekeeping; Security; War

References


Peacekeeping is a conflict management strategy whereby nonaligned soldiers and civilian personnel are deployed to a postconflict country in order to provide security and, in some cases, humanitarian assistance. On a global scale, peacekeeping is largely, although not exclusively, carried out by forces operating under the auspices of the United Nations, wearing the anti-camouflage blue helmets or berets that signify neutrality. In some ways, peacekeeping is a departure from the original purpose of the United Nations, which was founded in October 1945 under the mandate to prevent wars between states but not necessarily to intercede in civil wars or to engage in nation-building exercises. Peacekeeping operations have been embraced by international powers (especially in Europe, North America, and East Asia) to end hostilities, maintain ceasefires, establish stability, mitigate fallout from regional violence, and distribute humanitarian aid. In several countries in Africa, as well as in Haiti, UN peacekeeping forces have become a semipermanent feature of the military policing landscape. While UN peacekeepers have been instrumental in establishing and maintaining order in several countries, peacekeeping forces nonetheless have come under scrutiny due to accusations of corruption, arbitrary violence, and, in some cases, sex crimes against civilians.

Historically, peacekeeping can be usefully divided into two periods: the Cold War period (1956 to 1991), and the post-Cold War period (1991 to the present).

The Cold War period

The origins of peacekeeping cannot be separated from the great revolutionary movements of decolonization that reconfigured world politics in the 1950s and 1960s. Most peacekeeping missions in the Cold War era were designed to monitor the often violent fallout following colonial withdrawal from newly independent countries. In 1956 the first large-scale UN peacekeeping operation occurred in the aftermath of the Tripartite invasion of Egypt (known in the West as the “Suez Crisis”), where peacekeepers oversaw the withdrawal of Israeli, French, and British forces from the Egyptian Sinai. After the Tripartite Aggression mission UN peacekeepers were immediately deployed in 1960 to the Congo at the request of the first Congolese prime minister, Patrice Lumumba. This large-scale mission monitored the withdrawal of the Belgian colonial forces, and worked to prevent the outbreak of a civil war, in the wake of Lumumba’s murder by Katangan authorities in January 1961. Throughout the 1960s, 1970s, and 1980s, UN peacekeeping forces operated in multiple countries, predominantly in Africa, but also with sizable missions in the Middle East, Central America, and the Caribbean. Compared to the post-Cold War period, United Nations-led peacekeeping operations during the Cold War were occasional and largely focused on preventing interstate conflict. Between 1945 and 1989, 15 peacekeeping operations occurred globally. After 1991 the...
frequency, focus, and conduct of peacekeeping activities changed dramatically; as of December 2014, there were 16 active UN peacekeeping operations worldwide.

**The post-Cold War period**

In the post-Cold War era, peacekeeping operations underwent two distinct phases – conveniently marked by the September 11 terrorist attacks in the United States – in terms of force configuration, the types of institutions involved, and the frequency of operations. As Western capitals celebrated the end of the Cold War and hyped the coming era of economic liberalization and globalization, new interconnected global threats and human insecurities were identified: the rapid movement of rural populations to urban slums; a rise in ethnic cleansing and civil wars; religious extremism; structural poverty; a booming black-market arms and drug trade; organized crime; human trafficking; state instability in the Global South; and the impact of climate change on agriculture and social destabilization (the list is not exhaustive).

In the 1990s novel political concepts were introduced to general publics, like “failed states,” “new wars,” “humanitarian intervention,” and “nation building,” and new industries emerged to tackle global threats, such as nongovernmental organizations (NGOs) and intergovernmental organizations (IGOs) to private security contractors. The geographical imaginaries often employed to conceptualize the post–Cold War “conflict environment” relied on a neo-Orientalist binary, whereby the Global North was framed as an orderly hemisphere of tech-savvy, humanitarian states that could cruise the frictionless processes of globalization to address problems plaguing “failed states” at the frontiers of modern civilization. Despite the neo-Orientalist baggage, peacekeeping was a predominant instrument used in the 1990s by the “international community” to address global instabilities in the Global South. Moreover, peacekeeping became multilateral and “multifunctional” insofar as the remit of peacekeeping moved beyond just monitoring interstate hostilities to integrating “diplomatic, military, and humanitarian activities in an overall political strategy” for intrastate conflicts (Chopra 1996, 355).

Since September 11, the global geopolitical landscape has undergone even more profound transformations, especially in the Middle East, Central Asia, and western, northern, and central Africa – from the acute instability caused by the US–led wars in Afghanistan and Iraq, to the lingering fallout of the 2008 global financial crisis. As Perkins and Neumayer (2008) have shown, not only have the number of the peacekeeping operations increased, but the composition of peacekeeping forces has changed, with the inclusion of intergovernmental organizations, “third-party” missions, and ad hoc “coalitions of the willing” conducting such duties.

The approach to conflict management has also changed with the introduction of new technologies, such as remote sensing and geospatial analysis (Herscher 2014). By allowing operators to analyze conflict conditions at a distance inside humanitarian hubs, advocates claim that such technologies enable more “efficient” peacekeeping and aid operations. Moreover, in recent years, the goals of peacekeeping seem to have moved away from “nation building” toward what the US military calls “stability operations,” which forgo grand schemes buttress central states in the Global South in favor of basic police operations that provide basic security and population control (Hills 2009). In the immediate future, peacekeeping operations will likely continue to rely heavily on
so-called technological solutions, as governments in the Global North continue pursuing austerity measures which affect their ability to put large numbers of peacekeepers on the ground.

SEE ALSO: Postconflict geographies; War

References


Further reading


The term “pedoturbation” refers to processes by which soil is physically mixed or disturbed. Although its chief cause would seem to be biological, in the form of digging animals and falling trees, a diverse range of processes can actually lead to soil mixing, many of which are abiotic. Indeed, soils can be mixed by a wide variety of vectors, such as freeze–thaw and shrink–swell activity, seismic shaking, slope failure, and even exploding bombs. And although, historically, pedoturbation has been associated with profile simplification, it is now viewed as a process that does not always destroy but may sometimes form and maintain genetic soil horizons. Although long underappreciated and only minimally studied, pedoturbation is a measurable process in almost all soils, and has important consequences for soil genesis, properties, and behavior.

Expressions of pedoturbation

In its various forms, pedoturbation is studied either by observing the process (such as termites digging tunnels and, in so doing, moving soil particles) or by examining and interpreting the end products of pedoturbation within the soil itself. Signatures of pedoturbation are primarily expressed as within-profile, morphological imprints and as surface topographic features. Within-soil expressions include slickensides, fecal pellets, microfabric alterations, stone lines, broken and disrupted soil horizons, and open and infilled burrows (krotovinas). Because many forms of pedoturbation cannot effectively move larger fragments upward in soils (with tree uprooting and freeze–thaw activity being the major exceptions), they tend to settle and become concentrated at the lowermost depth of the process, as a stone line or zone. This type of depth distribution of coarse fragments is commonly used to infer long-term pedoturbation in the upper parts of a soil. Surface expressions of pedoturbation occur as microrelief such as gilgai, tree-throw mounds and pits, ant and termite mounds, patterned ground, and depressions associated with caved-in krotovinas.

Agents of pedoturbation

Types of pedoturbation are identified by the vectors that cause them. The list of such vectors, originally compiled by Hole (1961) and expanded on by Johnson et al. (1987), includes faunal- (animals), floral- (plants), gravi- (soil movement under gravity), congeli- or cryo- (freeze–thaw cycles), argilli- (shrinking and swelling of clay minerals), seismi- (earthquakes), aero- (passage of air, wind), aqua- (passage of water through soil), crystal- (rupture by growth of salt and other types of crystals) and impact- (comets and meteorites) turbation. To these, Hupy and Schaetzl (2006) added bombturba- tion. Normally, the term bioturbation is given to mixing by biota, including plants and/or fauna.
Bioturbation has received more attention than perhaps all other forms of pedoturbation combined (Wilkinson, Richards, and Humphreys 2009). Each of these agents/vectors may be prominent in some environments and negligible in others.

In most cases, pedoturbation is associated with increases in soil volume and porosity, and with concomitant reductions in bulk density. Positive feedback can readily be imagined, especially with respect to bioturbation — as habitation opens up the soil, rendering it even more habitable for bioturbators. Synergy of other kinds, between various pedoturbation vectors, may also occur, for example, the combination of high precipitation, steep slopes, deep weathering, and giant trees in the humid tropics results inevitably in isotropy or haploidization — minimal differentiation of the soil profile into recognizable horizons (Hole 1961). The following sections briefly cover some abiotic processes of pedoturbation, eventually focusing on the role of bioturbation, given the prime function of soil as terrestrial habitat for biota. Biotic processes also include past and future human activities, which disturb soil for better or for worse (anthropoturbation).

Abiotic processes

Although bioturbation is the most commonly viewed and studied form of pedoturbation, many other forms also exist, and most of these are abiotic in nature. For example, in graviturbation, soil moves from elevated upland positions to bottomlands by piecemeal erosion or mass movement processes. This movement can be imperceptibly slow (as surface wash and soil creep) or catastrophic (in the form of debris flows and landslides). Any degree of mixing can occur during these processes. The end results of such mixing are evident in the colluvial or alluvial end products near the bottom of slopes. Textural sorting may be evident here, although not easily distinguishable from that of animals. Mass movement can also be triggered by seismic activity, in which case it is considered a type of seismiturbation.

Climate is an important mediator in all forms of pedoturbation, even the abiotic ones. For example, the flow of water on and within soils can mobilize the solid constituents within, causing mixing. Water is an agent of volume change in soils, as enacted by the swelling of wetted clays or the expansion of ice crystals. It is also a generator and concentrator of salts through weathering, leaching, evaporation, and crystallization. Soil mixing caused by the growth of such crystals is called crystalturbation. The pedogenic development of Vertisols is a celebrated example of abiotic pedoturbation by the expansion caused as smectite clays wet up (Jackson 1965; Southard, Driese, and Nordt 2011). In wet dry climates, such soils undergo many cycles of wetting and drying, causing volume changes and churning of the soil profile; this form of pedoturbation is called argilliturbation.

Abiotic influences on pedoturbation are also, in turn, mediated by biota. Rates of infiltration, for example, are affected by faunal burrowing, and by vegetation cover which reduces raindrop impact. Less surface runoff means reduced sheet erosion, but it also makes mass movement more likely because of the increased incidence of saturation.

Biotic factors

Floralturbation (soil mixing by plants) is largely accounted for by the mixing that occurs when trees are uprooted (Hole 1961; Schaetzl et al. 1990; Šamonil, Kral, and Hort 2010). In this case the real agents are wind and gravity, with the tree itself playing a passive role. Trees tear up
soil as they uproot, often forming a pit at the former location of the roots, and an adjacent mound, located where the soil slumps off the roots. Soil materials within the mound can be extremely mixed, although in some cases, they are simply overturned in a more or less intact manner (Schaetzl 1986).

A spectacular example of tree uprooting can be observed where rainforest occurs on bauxite terrain in the interior of Guyana. Here, to a depth of about 2 m, the soil consists of a friable brown earth containing abundant, angular, disoriented fragments of saprolitic bauxite (Figure 1a), below which the saprolite shows an inherited jointing pattern, indicating preservation in situ. When wet, the tree canopy becomes top-heavy and the anchoring substrate loses much of its coherence, needing only a light breeze to topple the trees and tear up deep bauxite fragments. The disorganized morphology of the upper bauxite profile is therefore often attributed to uprooting. Loose surface gravels above hard laterite in the ancient jarrah forest of the Darling Range in Western Australia are possibly due to a similar mechanism, in which the laterite cap is progressively broken up by roots of falling trees (Figure 1b).

In southern Africa, a noteworthy accessory agent of floralturbation by uprooting is the elephant. In pushing trees over while foraging, elephants indirectly create a surface armoring of uprooted gravel in terrain with pedogenic calcrete (Figure 1c). The surface armor forms as the uprooted soil slowly erodes, leaving the coarse fragments at the surface as a lag concentrate. This example shows how flora and fauna may combine to mix soils, as well as illustrating that most soil processes are both biotic and abiotic in nature. The way in which animals interact with the plants on which they forage has many examples, especially among rodents and pigs which burrow for roots, bulbs, and truffles. Some rodents, such as North American gophers and Southern African mole rats, live below-ground in large family groups and have extensive tunnel networks. The constant bioturbation they cause is easy for the casual observer to miss.

**Figure 1** Examples of floralturbation by tree uprooting in contrasting environments. (a) The top 1 m of lateritic bauxite profiles in the Kopinang district of Guyana consists of randomly broken bauxite fragments mixed with a friable, earthy matrix. (b) Roots of fallen giant jarrah trees (*Eucalyptus marginata*) in the Darling Ranges, Western Australia, exhume large fragments of the laterite hardcap. (c) The combined effect of animals and plants near Lake Etosha, Namibia. Thorn trees (*Vachellia* sp.) pushed down by foraging elephants break up the pedogenic calcrete layer, eventually leading to armoring of the soil surface with limestone fragments. (Photos by Martin Fey.) *Continued opposite.*
Figure 1  Continued
Floralturbation is not always passive. The emergence of seedlings and the growth and death of plant roots are examples of active floralturbation processes, even though their scale is small compared to that of tree uprooting.

Faunalturbation (soil mixing by animals) is perhaps the most conspicuous category of soil mixing, with the greatest diversity of expression. Every continent has its own evolved assemblage of small and large animals that either live underground or dig, burrow, and scratch in the soil for various purposes, for example worms, insects, spiders, snakes, frogs, birds, and mammals (Johnson et al., 1987; Fey, Milewski and Mills 2010; Fleming et al. 2014). Quite likely, the largest overall amount of faunalturbation is produced by smaller animals, especially earthworms, ants and termites (Figure 2) – an observation that even Charles Darwin (1881) noted. Associated with the activity of these smaller fauna are the burrowing and nutrient-cycling effects of larger animals that prey on them. Their digging is additional to that of animals that forage for vegetable matter and of still others that simply make their home in the soil.

Figure 2  Termite mounds on an open grassy dambo (marshland) in northwestern Zambia (photo reproduced by permission of Tamara Knudsen).
Mima mounds are a famously controversial surface expression of bioturbation. Named for the Mima Prairie in Washington State, United States, Mima mounds are dome-shaped earth mounds, often >2m high and 10–50m in diameter (Figure 3). They are widespread on many grassland landscapes and can cover vast acreages at densities exceeding 100 ha$^{-1}$. Mima mounds occur only on soils shallow to bedrock or a subsurface pan, such as a duripan, or in soils that have high water tables (Horwath Burnham and Johnson 2012). The origin of these features has long been controversial, and many have attributed their formation to abiotic processes. Modes of origin include wind and water erosion of intermound lows, sediment accumulation at the sites of individual plants (shrubs or clumps of grass), trapping eolian or fluvial sediment, and paleoperiglacial origins, like Arctic stone circles, in large part because areas between the mounds were often bare of vegetation and strewn with large rocks. The fossorial rodent hypothesis of Mima mound origin, widely accepted today, was in fact once ridiculed. In this hypothesis, Mima mounds are formed as pocket gophers or similar burrowing rodents, for example moles or tuco-tucos, tunnel outward (not so much downward), pushing soil material behind them and building up a mound. The mounds serve as nesting chambers in the thin or wet soils, providing the increased soil thickness necessary to protect them from predation, winter cold, or high water
tables. Because these animals are so territorial, their mounds come to be located almost in perfect, regular arrangements on the landscape. Where soils are thick, gophers are not restricted in siting their nests; thus, they move from place to place, burrow more deeply, and mounds per se are not formed. Between the mounds one commonly finds a zone where soil is thin and rocks are numerous; the rodents have removed the soil from these areas. The mounds contain small stones only, equal to the size that the rodents can carry upward. Stone lines containing stones too large for gophers to move (>≈6 cm diameter) commonly underlie and ring the mounds. Thus, the mounds are essentially overthickened biomantles formed by point-centered burrowing (Horwath and Johnson 2006).

Mounds similar to Mima mounds are conspicuous over large areas of the African continent. In South Africa they are locally known as _heuweltjies_ (Afrikaans) or _isiduli_ (Zulu), and termite activity seems to be central to their formation (Fey, Milewski and Mills 2010), although, as is the case with Mima mounds, more than a single factor may be involved (McAuliffe et al. 2014).

Bioturbationally mixed parts or layers in soils are referred to as biomantles. Usually, a biomantle represents the upper part of the soil that is or has been thoroughly mixed and disturbed by biota (Johnson, Domier, and Johnson 2005). The idea of the biomantle was perhaps first introduced (but not named as such) by Charles Darwin (Johnson 2002). Johnson, Domier, and Johnson (2005) have argued that the role of animals in soil formation has until quite recently been underestimated because of the agricultural orientation of soil science. Especially useful is their honing of the concepts of proisotropic and proanisotropic pedoturbations, giving rise to regressive and progressive expressions of soil development, and resulting in soil profiles becoming either haploid (simple) or differentiated (horizonated), respectively.

It is appropriate to include all types of human activities in the category of faunal turburbation. Human actions include the ploughing and draining of farmland, diverse urban excavations and cut and fill operations, mining and reclaiming of mined lands, and various land-use practices that accelerate soil erosion. Some of these activities can fall under the rubric of pedoturbation. Anthrosols and Technosols are two groups of soils defined in the World Reference Base for Soil Resources (FAO 1998), that accommodate soils with properties markedly affected by human activity, some of which would involve mixing or anthropoturbation. The human factor may also have operated negatively with respect to bioturbation, because much human activity has led to a decline in the number and species of animals, that is bioturbators, in most environments.

### Consequences of pedoturbation

Of all the effects which soil mixing and disturbance produce, perhaps the most important involves soil structure and, more specifically, its corollary, soil porosity. Soil without pores would be lifeless. Porosity, pore size distribution, and the connectivity of pores all contribute to life-sustaining functions of soils such as the infiltration, storage, and discharge of air and water, including both nutrients and gases that may be either necessary or harmful. Furthermore, the penetrability of soil by roots and burrowing animals is mediated by porosity, which imparts a softer, more friable consistence to soils, especially finer-textured soils. Some clay soils such as Oxisols are highly porous. Others, such as Vertisols, are dense, compact, and deficient in functionally useful pores, except for either extremely fine ones that inhibit water
uptake by plant roots, or wide transmission cracks that allow rapid water infiltration into dry soil but close up soon after wetting. In Oxisols, porosity generated by the deep burrowing and construction activity of small fauna, especially termites and ants (Reatto et al. 2009), is preserved through a combination of soil climate and clay mineralogy which results in minimal shrinkage and swelling, and negligible clay dispersion. In Vertisols and related clayey soils, mixing by swelling and clay dispersion (argilliturbation) likely combine to destroy larger soil pores not long after they have been generated via bioturbation. Shrink–swell cycles also create ephemeral porosity in the form of spaces between peds, which are most evident when the soils are dry.

A number of other, secondary consequences of soil mixing also stem from the development of soil porosity. Many of these take the form of feedback loops, through which pathways of development, having crossed a pedogenic threshold, become increasingly divergent. The clay minerals that characterize the Oxisol–Vertisol catena derived from mafic rocks in tropical landscapes are promoted by the flushing of silica and bases from the Oxisol and the accumulation of these solutes in the Vertisol (Jackson 1965). In considering the chemistry of such pedogenic divergence, it might be interesting to assess the contribution made by pore-initiating soil biota.

Pedoturbation has other interesting consequences. In agriculture, there is increasing awareness of the fuel-saving and soil structure-preserving benefits that may accrue from minimal tillage or no-till farming. In reality, soil tillage does not cease with no-till farming; it continues, but in the form of biotillage, driven by bioturbators and fueled by crop residues. Worms and other soil fauna do the work of tractor and plow. In repairing landscapes disturbed by mining, the long-term stability of restored vegetation hinges on the extent to which a new ecosystem can be established, including soil organisms, especially burrowers that can counter the effects of human-induced soil compaction. Water yields from catchments are affected by faunal pedoturbation, including preferential subsurface flow through krotovinas. The ways in which bioturbation affects land use and ecosystem function are reviewed by Wilkinson, Richards, and Humphreys (2009). The development of subsurface stone lines below a biomantle, as is so common in many landscapes (Nye 1954; Johnson et al. 1987; Figure 4), is of special interest to archaeology because human artifacts are often included in these stone lines.

Other consequences of pedoturbation in its broader sense are those of an engineering nature. Landslides produce characteristic soil morphology that can serve as a warning to builders. Slickensides, cracks, self-mulching, and gilgai microrelief are all indications that the site may be unsuitable for construction due to argilliturbation. And the presence of cryoturbation (frost heave) may sometimes also be revealed by characteristic spatial patterning (Johnson et al. 1987).

Theoretical considerations regarding pedoturbation

For decades, soil science has focused on how pedologic order (anisotropy, soil horizonation) can evolve from sediments that were initially disordered, isotropic parent materials such as loess or dune sand, or ordered, anisotropic ones such as stratified alluvium. Parallel to this, the notion that pedoturbation is a regressive soil process – one that blurs soil horizons or prevents them from forming – has also been historically
Figure 4  Stone lines marking the boundary between the fine earth-textured, humus-rich biomantle above, and underlying B horizon containing weathered saprolite. The soils shown here are both derived from gabbro on the Nolangeni Mountain near Kokstad, South Africa. Buried artifacts, originally left on the surface by paleo-peoples, are lowered by bioturbation and thus are commonly encountered in such stone lines. (Photos by Martin Fey.)
PEDOTURBATION

dominant. Few studies prior to 1960 concluded that pedoturbation can actually create, or even preserve, the anisotropy that often directly results from pedogenesis. Thus, pedoturbation has traditionally been viewed as a process that destroys horizons or that acts to slow or reverse horizon-forming processes.

Current research has found, however, that pedoturbation can also form horizons, or at least be neutral with regard to horizonation processes. Thus, Hole (1961) and Johnson et al. (1987) classified pedoturbation into one of two categories: proisotropic or proanisotropic. The former term implies a condition tending toward (pro) isotropy or disorder, while proanisotropic pedoturbation means a tendency toward layering and order. Proisotropic pedoturbation disrupts, blends, or destroys soil horizons and geologic layers, or impedes their formation. When this type of pedoturbation dominates a soil, a morphologically simplified profile evolves from a more ordered one. Proanisotropic pedoturbation forms or maintains soil horizons and geologic layers, usually causing an overall increase in profile order.

Seldom are pedoturbation processes entirely proisotropic or entirely proanisotropic. Instead, they usually have elements of both, with one form of pedoturbation often being more strongly expressed than the other. Most soils have components of each of these two sets of interacting processes. The balance between the two ultimately determines the morphological makeup of the soil. For example, earthworms may mix organic litter into the A horizon, and in so doing blur the two horizons – a proisotropic phenomenon. But, by doing this, the worms thicken the A horizon at the expense of the litter layer above, thereby promoting horizonation – a form of anisotropic mixing. The expression of this suite of processes is a soil with a slightly thicker A horizon.

Conclusions

Soils are constantly in a state of flux. The flux can be turbulent and catastrophic, but for most of the time and in most places it is barely perceptible, permitting stable anchorage for the growth of plants, shelter of animals, evolution of ecosystems, and the emergence of prosperous agricultural communities. And yet, in such settings, pedoturbation of various sorts may be ongoing almost continually, although usually unseen, below the surface, and in the background. In other landscapes, such as those grasslands where Vertisols dominate, or the tundra where evidence of frost churning is everywhere, the importance of pedoturbation is clear, omnipresent, and less likely to be underestimated.

Normally, pedoturbation rejuvenates and refreshes the soil. Only rarely does it constitute a hazard. Future research could profitably explore the extent to which the soils we study owe some of their attributes to pedoturbation. These attributes may disappear if the flux is not maintained or may have done so already because pedoturbation has ceased. On the other hand, traditional soil tillage may be seen as a substitute for natural pedoturbation but it may in the long run be insufficient to maintain the soil porosity needed for profitable agriculture.

SEE ALSO: Biogeomorphology; Soil biology and organisms; Soils in archaeological research; Soils in geomorphic research

References


Periglacial processes and landforms

Forrest D. Wilkerson and Ginger L. Schmid
Minnesota State University, USA

The term “periglacial” was first used to describe a variety of landforms that existed in cold climates within the Carpathian Mountains of central Europe (Lozinski 1909). The term refers to landforms that develop near or surrounding glacial environments and originally concerned landforms thought to be influenced and created by glacial climates. It was quickly realized that periglacial landforms can and do develop in numerous localities that are not necessarily located anywhere near glaciers or permanent ice fields. The formation mechanisms that create periglacial landforms are complex, multi-faceted, and to a large degree still poorly understood. The one unifying factor is the requirement of cold temperatures combined with a warmer summer season that allows for numerous freeze-thaw cycles. The majority of periglacial landforms are located in arctic, subarctic, and cold alpine areas where conditions are wet enough for the formation of ground ice, but dry enough to preclude the formation of glacial ice. There are of course exceptions where patterned ground has been documented as forming beneath glacial ice and exposed only as the glacier retreated (Washburn 1956).

As research concerning these landforms began to grow, it was realized that there are numerous landforms that develop in cold temperatures and that the processes that lead to the development of these landforms is complex and often poorly understood to this day. In addition, a variety of researchers began to name these landforms in terms that made sense to their particular language and research background. The nomenclature became so complex and convoluted that it became a serious burden to geomorphologists and other Earth science researchers. The first attempt to streamline the confusing nomenclature occurred in 1933 and this effort was followed by Troll’s (1944) excellent synthesis of the geographic variations, distribution, and variety of formation mechanisms for periglacial landforms. However, it was not until Washburn’s seminal paper was published (1956) that the confusing nomenclature was formalized into a classification system that is still widely used today. Washburn’s paper reduced the confusion and standardized the terminology used for each of the numerous landforms he described (56 landforms). In addition, the processes that lead to the development of these landforms were also described and discussed (19 processes). Only the most significant landforms and processes described by Washburn are discussed in this entry.

Modern investigations understand the complex nature of periglacial landform development, including the concept that the landforms are polygenetic. This term refers to the creation of similar looking landforms being developed by different processes as well as different looking landforms developing through similar processes. The type, number, and activity level of each individual landform is highly dependent upon the presence or absence of water, number of freeze-thaw cycles and depth of freeze penetration, the presence or absence of permafrost, and the depth of the active layer. In the Northern

The International Encyclopedia of Geography.
DOI: 10.1002/9781118786352.wbieg0480
Hemisphere, periglacial landforms are more likely to be found on slopes with a north to northeast facing aspect, within soils containing enough fines (primarily silts and clays) to hold moisture for longer periods, sufficient precipitation, and enough freeze–thaw cycles to allow for the sorting and development of landforms.

The presence of permafrost, ground that is frozen for at least two summers and the intervening winter, is also very common in periglacial landscapes that are presently active. Patterns are also more likely to develop in depressions and shallow ponds that tend to pool water and are wet when the winter freezing front penetrates to depth. It is generally understood that the 9% expansion of water as it changes phase from liquid to ice is the driving force for the inception, development, and maintenance of periglacial landforms.

The majority of periglacial landform formation processes that lead to the inception and development of periglacial landforms have been reviewed (e.g., Washburn 1956; Kessler and Werner 2003; Hallet, Sletten, and Whilden 2011). A useful example can be found in the development of sorted patterned circles. Water in liquid form can migrate through soils by capillary action toward a freezing plane. The freezing plane is usually the base of a stone or boulder that lies within a nonfrozen soil matrix. The stone is denser and will freeze before the soil due to insulating pockets of air contained within the soil. The capillary movement is driven by vapor pressure gradients and can actually occur in soils whose temperature is as much as \(-2^\circ\)C. The liquid water freezes instantly upon contact with the already frozen boulder and begins building an ice lens. As the lens expands and grows through time, the boulder is heaved from the ground and the sorting process begins. The freezing and 9% expansion of water stands out as the primary mechanism for the implementation and development of landforms, but many other factors influence the final landform type and many of these processes appear to be present but as of yet have not been proven. For example, sorted circles appear to be created by convective currents that develop within the soil profile. However, true convection has yet to be proven and others claim that, rather than convection, freezing segregates the larger clasts and forces them to the sides of the pattern. Once all of the larger clasts are separated, the freeze–thaw process no longer sorts any material and the landform is considered to be fully developed.

Even in extremely cold environments, periglacial landforms do not develop everywhere. Stable surfaces are necessary to accommodate the slowly developing landforms. Snowfall is also a significant factor, where too much snow, particularly at the beginning of winter, will insulate the ground and prevent deep freezing, and too little results in inadequate moisture for sorting to occur. Late season snowfall also can help to insulate the frozen ground and preserve the landforms rather than allowing the summer thaw to penetrate deeply enough to melt the permafrost. Coarse soils also have difficulty maintaining enough water to allow sufficient ice expansion when winter freezing occurs. Pockets of fine materials often lead to sporadic landform development and remarkably different rates of activity over very short distances (Wilkerson 1995). In most landscapes containing periglacial landforms, subtle differences in microclimate lead to the development of classic landforms in one area and the complete lack of landforms in adjoining locations (Washburn 1980).

The complexity of periglacial landforms is demonstrated by the diversity of landform distribution and location. Most periglacial landforms are located in the cold dry climates of the Arctic, particularly in northern North America, Europe, and Asia. However, numerous mountain ranges...
in the mid latitudes also have the same types of climate as the High Arctic and contain a number of similar landforms. Landforms can also be found at sea level in high latitude environments, but are also found on the equator in high alpine environments. Troll (1944) provides an excellent and thorough description of the landscapes where patterned ground develops and the types of landforms that develop in each area. The concept of macro-patterns, those greater than 2 m in diameter, and micro-patterns, with diameters less than 1 m in diameter, were defined and described. Activity rates within these macro- and micro-scale landforms vary within landforms in certain mountain environments, with macro landforms being relict and micro landforms nested within the larger landforms showing high rates of activity (Wilkerson 1995).

Washburn’s (1956) classification included 56 distinct periglacial landforms. Only a few of these can be specifically addressed here. In general, these periglacial landforms can be separated into two broad categories of sorted and nonsorted patterns. Nonsorted patterned ground includes symmetrical circles, polygons, nets, and stripes, as well as, ice, sand, and soil wedges that are fairly common in arctic environments. The majority of nonsorted patterns are bordered by slightly raised or depressed soil ridges and not by sorted stones. Different types or amounts of vegetation also often define the borders of nonsorted patterns, which give the pattern its characteristic shape. Solifluction lobes and nonsorted circles, polygons, nets, and stripes are often created by freeze-thaw segregation which allows either a raised or depressed border that can contain denser vegetation, or often different species of vegetation, that form the pattern border. Interestingly, nonsorted patterns usually develop within a soil matrix that contains a variety of different sized material, but boundaries are created by contraction, expansion, and soil mounding rather than by sorting (Washburn 1956). As an example of the complexity with periglacial landform development, ice, soil, and sand wedges are known to rarely exhibit signs of sorting; however, sorted stones have been noted within the borders of some of these landforms.

Sorted patterned ground requires the presence of a soil matrix that contains at least two different sizes of material. Sorting involves the separation of larger and smaller particles and the segregation of larger particles usually to the outsides of the pattern. On flat ground, this sorting produces sorted circles that can be remarkably beautiful and uniform (Hallet and Prestrud-Anderson 1986). As slope angles increase, the circles are stretched into polygons and nets, and if slopes are steep enough, sorted stripes are the result (Figure 1).

As with many periglacial landforms, there are exceptions. For example, stone pits form as opposite or mirror image landforms where the larger stones are forced into a center area or pit and the finer materials surround the central pit. The sorting process and the resulting formation of sorted patterns has been excellently modeled by Kessler and Werner (2003), who eloquently describe the number and property of freeze-thaw cycles that are needed to produce identifiable landforms.

Pattern development, or self-organization, is common in many areas that are considered periglacial environments. In the formation of periglacial patterned ground landforms, self-organization is produced by the repeated freezing and thawing of the soil matrix. This physical process sorts the larger stones by heaving them from the ground and forcing them to the sides of the landform. The heaving is driven by the expansion of ice lenses as water freezes and is spatially differential in nature due to the ability of finer sediments to hold more water within their soil matrix. This heaving process produces
Periglacial sorted circles and polygons formed on a low angle surface (foreground) and stretched into nets and stripes on steeper angles (background) in the White Mountains of California (photo by author).

Sorting and results in a pattern where larger stones define the borders and the interior of the pattern consists of finer material, primarily small gravels, silts, and sand.

Ice, sand, and soil wedge polygons develop in very cold climates where winter temperatures often reach lows approaching or exceeding $-40\,^\circ\text{C}$. As the ground freezes, it contracts to the point where a split or crack develops along a failure plane. Thawing at the surface allows liquid water to flow into the crack and freeze against the still frozen deeper layers, which forces the wedge to widen and grow by accretion. In drier climates, sand or soil will often drift into the crack and fill the void. The crack continues to widen through subsequent winters and continuing deep freezing cycles. In uniform soils, cracks will often develop at $120^\circ$ angles and will through time develop into distinct polygonal patterns. Eventually, the cracks can grow to be a meter wide or larger at the surface and taper down until they pinch out right at or just above the permafrost layer. These often intricate and beautiful landforms are common along arctic coastal plains in the Northern Hemisphere and are usually found in poorly drained marshy environments.

Solifluction is the slow downhill movement of saturated soils due to gravity. Solifluction occurs in soils underlain by seasonally frozen ground,
while gelifluction is a specific type of solifluction that forms over permafrost layers. These landforms develop on shallow slopes and move at rates measured in millimeters to centimeters per year. They are usually nonsorted, but can occasionally develop stone banked lobes along the front of the landform. Solifluction lobes are relatively common in both arctic and high alpine environments and relict landforms often leave behind a series of distinct terraces. In certain environments, sorted patterns develop on the treads of the lobes and the risers contain stripes due to their steeper topography.

The pingo is an unusual ice-cored landform that is found primarily in arctic environments. Liquid water migrates through capillary action to the freezing core of the pingo. The core grows through the accretion of ice and forms a mound or small hill as it develops. The soil on top of the ice core acts as an insulating layer and prevents the ice core from melting. If the pingo grows to the point where it ruptures the overlying layer of insulating soil, warmer conditions often penetrate into the ice and melt the core, which leads to the demise of the landform. Pingos require the presence of permafrost to keep liquid water near the surface and provide a source for the continued growth of the core.

Pingos can be open or closed. Open-system pingos develop where water drains from an elevated source and feeds the growth of the pingo through gravity. Closed-system pingos primarily develop in drained marsh or pond areas where water lying on top of the permafrost layer is added to the pingo through hydrostatic pressure. The palsa is a landform that is somewhat similar to a pingo, though usually much smaller and more often associated with peat bogs rather than drained lakes or gravity-fed soil water. Ice lens development forces the peat upward, forming a distinct mound. The pingo is a distinct landform and similar looking landforms have been noted on the surface of Mars. This suggests that similar mechanisms, whether open or closed, are occurring there.

Climate change and periglacial processes

Climate change, particularly warming, has already had a noticeable impact on permafrost and periglacial landforms in low elevation environments of the Arctic. The permafrost layer of the Arctic is melting noticeably, which has a direct impact on future climate change and periglacial activity. As the permafrost melts, large amounts of methane are released, which produces a negative feedback for continued warming. Melting permafrost also removes the ability of the ground ice to keep liquid water near the surface during the early winter freeze-up. Freezing and ice expansion are the driving mechanisms for the majority of periglacial landforms. Well-drained soils tend to be moisture poor and therefore do not allow the added expansion of ice during re-freeze. If enough warming occurs, the number and depth of freeze-thaw cycles will be compromised and landforms will cease to be active. In certain alpine environments water seems to be more significant than temperatures, as long as freeze-thaw cycles are maintained. If climate change reduces precipitation or significantly impacts the water balance of a periglacial environment, the lack of sufficient water during re-freeze will have a direct impact on the rates of activity and in some cases will lead to periglacial landforms becoming inactive or relict. If climate warms enough to significantly reduce the number of freeze-thaw cycles and the depth of freeze penetration, then all periglacial activity will cease in that particular environment.
PERIGLACIAL PROCESSES AND LANDFORMS

The next frontier: extraterrestrial periglacial environments

With the advent of advanced satellite-based remote sensing, periglacial landscapes can now be monitored in one location and compared with other cryosphere environments. Comparisons can now be made between Earth and the surface of Mars, where it is now known that there are numerous landforms that resemble many of the patterned landforms found on Earth (e.g., Soare, Conway, and Dohm 2014). A number of researchers have documented landforms that resemble ice wedge polygons, pingo fields, collapsed pingos, and possibly even patterned ground. The presence of periglacial landforms on Mars suggests that similar processes are occurring on the surface of that planet as well. These processes are either physically similar to earth-based processes or mimic those found on Earth. It also suggests that freeze-thaw processes, most likely occurring with water but potentially occurring in other substances such as carbon dioxide, are also continuing to impact the surface of Mars.

SEE ALSO: Climate change and permafrost; Cryosphere: remote sensing; Cryosphere studies: history; Mass movements in periglacial environments; Patterned ground; Permafrost: definition and extent; Soils of cold and permafrost-affected landscapes

References


Further reading

History

The first mention of permanently frozen ground comes from Siberia, Russia. The first written text dates back to 1598 and the first report documenting its occurrence comes from Yakutia by Tatishchev in 1725. M. Lomonosov in 1757 gave its first scientific description. In 1837 a 116.5 m deep shaft was excavated in a permafrost layer in Yakutsk. The temperature measured there showed the thermal regime of the frozen ground for the first time. The first mammoth body, preserved in the frozen sediments of the Lena River, was discovered in 1799 and caused a sensation on being brought to St Petersburg. Work of M.I. Sumgin, published in 1927, is considered to be the first synthetic description of permafrost. Leonard Jaczewski was the first to establish the southern boundary of permafrost in Siberia and differentiate it into mountain and Arctic permafrost in 1889.

Definition

Permafrost is a thermophysical condition of the lithosphere (rock or soil), the temperature of which is equal to or lower than 0°C for a period of at least two subsequent years, along with the summer season dividing them. In terms of the encompassed area it is the most important component of the Earth cryosphere.

Permafrost is defined on the basis of the substratum temperature and its duration. This means that it is not an object or an Earth surface form, but physical state (Figure 1). As such, it is a direct subject of geophysical research, since the only direct measurement to be applied to permafrost is the measurement of temperature. It is expressed in Celsius due to the important role of water, which usually accompanies the occurrence of permafrost; the solid/liquid phase changes in the water generate a series of geological and geomorphological processes. These processes and the accompanying relief forms are the subject of research both for Earth and other celestial bodies. The geosciences refer to permafrost with the term “phenomenon,” which reflects its essence very well as a subject of research that does not have a material form and is only a geophysical feature. This is why it is probably the only phenomenon in geosciences that remains invisible for the observer.

Characteristics

Permafrost is a geological phenomenon. It is formed as a result of climatic impact, in conditions where cooling causes the seasonal accumulation of “cold” to be greater than the seasonal accumulation of “heat”: the depth of ground that freezes in winter is greater than the depth that thaws in summer. The notions of “cold” and “heat” are colloquial references for <0°C and >0°C, respectively; they facilitate the description of the process of formation and
PERMAFROST: DEFINITION AND EXTENT

Figure 1 Thermophysical model of permafrost of a periglacial environment (bold lines) and associated terms used to describe the ground temperature relative to 0°C.

Freezing takes place in the temperature range of ≤0°C.

The most conducive climates for permafrost are freezing and dry climates, specifically continental ones, whereas cold and humid climates with a large amount of precipitation in solid form are appropriate for glaciation.
Permafrost layer

Permafrost occurs in the form of a geophysical layer in the lithosphere inside which the temperature is constantly equal to or lower than 0°C. Its upper limit is usually the surface of 0°C-permafrost table, that is, the bottom of the seasonally thawing active layer. When the permafrost is of relict nature, not connected with the current climate, both from the top and from the bottom it is limited by a 0°C surface, beyond which the temperature is always positive. Since permafrost is not a sediment in geological terms, there is no sedimentological variability in it. It may be referred to as a “climatic deposit;” however, the stratigraphy of this “deposit” comes down to variability in temperature in its entire thickness. This is its primary feature. Secondary features may be the properties of the material encompassed by permafrost: rock type and the presence of various forms of water and/or ice, which influence changes in the solidity of the ground, for example, by water freezing-cementation.

Cryotic state, presence of water/ice

Permafrost that is developing in unconsolidated sediments may contain ice depending on the type of material. Fine-grained materials can contain water and, thus, are referred to as frost-susceptible. Drainage is better in coarse-grained materials and not all pore spaces are filled with water. Freezing causes ground ice to be the major component of permafrost in both fine and coarse grained materials. Its most important types are: pore ice, segregation ice, vein ice, and intrusive ice. As a decrease in temperature to below 0°C does not have to cause, and usually does not cause, full freezing of all water present in the ground, water occurs both in liquid and frozen form within permafrost. Mineralization, water pressure, capillarity, and surface adsorption in the material encompassed by permafrost cause the freezing point to drop below 0°C; this is the so-called freezing-point depression. In soils and loose sedimentary rocks the freezing point can even be several dozen degrees below 0°C. A rock or soil in which the temperature is ≤0°C without freezing occurring for at least two consecutive years is referred to as remaining in cryotic state. Cryotic ground is a synonym of permafrost. Gradual freezing causes cryosuction, causing water to migrate towards the freezing front, where it freezes, forming lenses of segregational ice. Obviously, the presence of water in the form of ice is much more common in permafrost. The ice in permafrost can be:

- a component of the frozen ground, with homogeneous or heterogeneous nature;
- the medium where permafrost occurs.

In permafrost of the periglacial environment, the accompanying ice can be of different genesis and in various volumes. Permafrost contains mainly pore ice, that is, ice formed in the water freezing process. The terms accompanying permafrost cover over 20 different types of ice. The descriptions include types of ice connected with the particular relief forms, such as pingo ice or ice wedges, as well as types formed as a result of the influence of various processes inside the permafrost layer. Its formation process is usually described in geological categories. Formation of pingo ice is similar to magma intrusion into the sedimentary material. The pressurized water may be derived from groundwater flow beneath the permafrost layer or from the expulsion of water from pores ahead of a penetrating freezing front in saturated unconsolidated sediments. A crystalline laccolith of the intrusive ice is formed in this way: pingo ice is referred to as hydrolaccolith. The ice formation connected with the permafrost environment, as specified
above, is homogeneous. Ice formed in a different environment and then transported to the permafrost environment is heterogeneous. It is usually a fragment of glacier: dead, buried ice. It can also be buried river or sea ice. Melting of ice in nonconsolidated sediments during permafrost degradation can be accompanied by ground subsidence. The irregular relief formed this way is referred to as thermokarst.

**Subsea permafrost**

Subsea or offshore permafrost refers to geological materials that have remained ≤0°C for two or more years and that occur on the seashore at or below sea level. It is formed as a result of sea water flooding of a shore encompassed by permafrost or as an effect of the influence of water on the sea bottom. Most commonly it is of Pleistocene origin and was formed when the lowering of water level during glaciation uncovered a shallow continental shelf and allowed it to freeze. Similarly as on the land, the subsea permafrost also does not have to be frozen due to the lowering of the water freezing point below 0°C resulting from the presence of salt or capillary effects in the fine-grained material. Narrow zones of coastal permafrost are present along most arctic coasts, particularly in the East Siberian Sea, Laptev Sea, and Beaufort Sea. Subsea permafrost can extend up to 100 km from the shore. Sea-bottom temperatures on the Arctic shelves reach a temperature of −1.8°C and salinity up to 34 psu.

**Caves permafrost**

The karst forms, caves, can also be encompassed by the occurrence of permafrost. Its presence and development is related to the specific type of air circulation, called chimney effect or Balch circulation, which is based on the inflow of colder, heavier air that sinks to the lowest point and allows the cold winter air to remain during summer. The diversified relief of mountain areas is particularly conducive to this phenomenon. This allows not only for ground freezing to develop for many years but also for the accumulation of underground ice, the genesis of which is sometimes very similar to accumulation of the glacial ice. These are cases that can be referred to as underground glaciation. Such caves can occur in areas where the average yearly temperature reaches even +7°C, but where it is frosty in winter. This circulation type is also noted in sufficiently high mountainous scree slopes.

**Glacial permafrost**

Until the mid-1970s, glaciers and ice sheets were included into the scope of permafrost. Later, the occurrence of permafrost was excluded a priori from the glacial environment. It is obvious that glaciers are encompassed by permafrost. Firstly, their temperature is never higher than 0°C and, secondly, ice is a mineral – and in greater amount, rock – therefore, it belongs to the lithosphere (Dobinski 2011).

Glacial permafrost occurs in the form of the thermal structure of glaciers where “cold ice” and “temperate ice” can be found (Figure 2). Treating “cold glacial ice” as permafrost is obvious, since it has been frozen for many years. “Temperate ice” also meets the requirements of the definition, as its temperature is always ≤0°C. However, as it is at the pressure melting point, it also contains water in the cryotic state. Since the cryotic state is a synonym for permafrost, each ice sheet, ice cap, and glacier is always and fully encompassed by permafrost. This implies that permafrost occurs also under the glaciers.

The volume of ice in permafrost can be very diversified: from almost 0% in freezing magmatic rocks to 100% in buried glacier ice.
or glacial permafrost. In this context the term “dry permafrost” is also applied; this refers to the perennially frozen rocks that do not contain H₂O. Since the ice, by definition, is dry, the term dry permafrost should cover such a state where all the water within the permafrost is...
PERMAFROST: DEFINITION AND EXTENT

frozen. The notion “wet permafrost” could be justified in relation to the cryotic ground both in glacial and periglacial permafrost.

Active layer

Above the permafrost layer there is a surface layer in direct contact with the atmosphere; it is called the suprapermafrost layer. The part where the temperature increases seasonally above 0°C is called the active layer. This can cover the suprapermafrost layer fully when the permafrost is in direct connect with the current climate. When it is relict and located below the layer that freezes seasonally, there is a nonfrozen ground between the active layer and the permafrost – the talik. The active layer is not a part of the permafrost but is a place where the highest number and complexity of geomorphological processes that accompany the occurrence of permafrost take place, especially when the process of transition through 0°C is accompanied with the presence of water. The active layer has an important role in the cold regions of the world, as the geomorphological processes are most often connected with the seasonal freezing and thawing of the active layer and water present in it. In the case of active permafrost or partially-active permafrost, the active layer is simultaneously the top of the permafrost – the permafrost table. In the case of relict, inactive permafrost which is usually located deeper, the active layer is not in contact with the permafrost table. The temperature in the active layer can change within a great range of positive and negative temperatures, since it reacts to seasonal weather changes. Therefore, its thickness can vary every year. It depends on multiple factors, including the ground surface temperature, surface thermal properties and its cover, soil moisture, snow coverage period, altitude above sea level, slope aspect, topographic conditions, chimney effect, terrain depression, and others.

The thickness of the active layer in peat is very small owing to the very good insulation properties of peat. In the subarctic region the thickness of the active layer in such a material reaches 60–80 cm, whereas in the clastic material in the same region it is 1.5–2 m. Where the thick clastic material is abundant, the thickness of the active layer may reach even several meters. Nicholson and Thom (1973) state that at the Timmins 4 permafrost experimental site in Labrador (55°N, 700 m above sea level (a.s.l.)) the active layer reaches up to 12.3 m deep. The thickness of the active layer changes along with geographical latitude and altitude above sea level. In the northernmost part of Arctic it reaches only about 15 cm and becomes thicker in the subarctic regions. The characteristics of Earth lighting as well as lack of seasonal climate change at the very low latitudes causes the daily nature of the variability of freezing and thawing of the active layer. In equatorial areas at an altitude exceeding 5000 m a.s.l. this process covers only several centimeters of the ground or, in shady places, it does not occur at all, and the perennially frozen ground begins from the surface, thus sometimes the active layer is not present at all. A similar situation can be found only in the central part of the Greenland and Antarctic ice sheet.

Permafrost processes, active permafrost

Since permafrost is a thermal state of the ground that depends on the interaction between the atmosphere and lithosphere, a change in the permafrost temperature is a derivative of the change of atmospheric temperature at its ground layer. There are only two processes acting here: aggradation and degradation. Aggradation is a process reflecting the cooling of the climate,
PERMAFROST: DEFINITION AND EXTENT

consisting of a lowering of the permafrost temperature, usually accompanied by an increase of permafrost thickness, whereas degradation is the reverse process, connected with climatic warming. In both cases the phase transition of water, freezing, can take place if permafrost contains it. The seasonal variability of permafrost temperature can cover its upper part, up to the depth of zero annual amplitude (ZAA) if it is of great thickness, or its entirety when it is appropriately thin. ZAA is then below the permafrost layer. It may be referred to as, correspondingly, partially active or active. This is in respect of the permafrost influenced by the contemporary climate. The relict permafrost, located at a greater depth, is not affected by the current climate, thus it can be referred to as inactive. Its temperature remains constant. The process of aggradation and degradation has influence not only on the thickness of permafrost, but also on its temperature.

**Thickness**

The thickness of permafrost depends on the degree of cooling influencing the ground surface, its length, and the thermal properties of the ground, including water content. A two-year-old permafrost can be an unperceivable centimeter-thick frozen or nonfrozen layer in the ground. In areas of volcanic activity, such as Kamchatka or Iceland, the permafrost thickness is not substantial despite the conducive climate. In the latter, it reaches to 5–30 m. In Canada, in its continental climate, at the southern border of occurrence where the mean annual air temperature (MAAT) is close to 0°C, the permafrost thickness reaches 15 m; in the north of the continent it reaches 60–360 m and on the islands of the northern archipelago it reaches 400–530 m. The estimated thickness reaches up to 1000 m in the Paleozoic rocks quartz sandstones. In Spitsbergen, the thickness of permafrost reaches several dozen meters in valleys, while in mountainous regions it reaches 400–500 m. In the mountains of mean latitudes permafrost thickness reaches from several to several hundred meters. On Monte Rosa in the Alps (4500 m a.s.l.) the thickness can even exceed 1000 m. The thickest permafrost in the world has been measured in western Yakutia, near the Arctic Circle (111°E, 66°N), where the depth reaches 1500 m. The maximum possible depth on Earth is estimated to be about 2600–3000 m in Antarctica.

**Age**

Determination of permafrost age by absolute dating, for example, with isotope methods, is very difficult, as it is a state not a material object. It is an attempt to determine the “age of temperature.” Indirectly, it is possible to date permafrost, bearing in mind that its minimal age is the time needed for its formation with the given thickness. When the temperature at the surface is too low, the penetration of “cold” in the ground is slow and formation of a thick layer of permafrost takes a long time. The age of a layer of permafrost several hundred meters thick is usually estimated to be between several dozen and several hundred thousand years. Most of the Earth’s permafrost dates back to the Pleistocene age. The Siberian permafrost is estimated to be over 1 Ma. The oldest permafrost, located in the Dry Valleys region in Antarctica, is estimated to be over 2 Ma. If the glacial permafrost of Antarctica is taken into consideration, the age of the oldest permafrost on Earth can be estimated at 20 Ma.
PERMAFROST: DEFINITION AND EXTENT

Occurrence

As geosciences are applicable outside the Earth as well, permafrost also commonly occurs in space due to its low temperature: about 3K. This affects planets, moons, and other celestial bodies of the solar system. The surface of Mars, besides the presence of two circumpolar ice sheets, is most probably fully encompassed by permafrost. Many other celestial bodies of the Solar System have an icy lithosphere. These include the planets Uranus and Neptune, as well as the so-called icy satellites: Europa, Ganymede, Callisto, and Io. Their icy lithosphere is encompassed entirely by glacial permafrost.

The main problem regarding the scope of permafrost occurrence on Earth is its presence in glaciers and ice sheets. During the last almost 70 years from the development of its commonly assumed definition, glaciers are taken into account or not in the global area of permafrost occurrence. The solution of this problem is significant, as it concerns approximately 15 × 10^6 km^2 of the Earth’s surface, that is, mainly the ice sheets of the Antarctica and Greenland. Another important factor influencing the determination of permafrost occurrence in the world is the issue of relict permafrost. Its presence has been found in Western Siberia and, recently, also in the Polish lowlands. Research regarding this matter has only just begun in America.

The dominant permafrost is the zonal permafrost, also referred to as the Arctic or latitudinal permafrost, as it is related to the climatic zones of Earth. It occurs from 84° N in Greenland, covers vast areas in Russia (Siberia), the United States (Alaska), China (Tibet), Canada, and Mongolia. Due to the uneven location of lands on Earth, it covers especially the continents on the Northern Hemisphere, Arctic, and polar regions, especially in the continental climates where large annual temperature amplitudes are conducive to deep frost penetration in the lithosphere. Usually, the temperature limit allowing for formation of permafrost is specified as −1°C, mean annual air temperature.

With increasing altitude or latitude, first the isolated patches zone occurs (0–10% ground underlain by permafrost), which then becomes the sporadic permafrost zone (10–50%), then discontinuous permafrost (50–90%) and, finally, the continuous presence of permafrost (90–100%) (Zhang et al. 1999).

When climatic cooling is connected with temperature drop along with the increase of altitude above sea level, altitudinal or Alpine permafrost is formed. These conditions allow for the occurrence of permafrost at all latitudes on Earth; however, only in the Northern Hemisphere can latitudinal permafrost overlap altitudinal permafrost: for example, the Ural Mountains, Scandes (Scandinavian Mountains), and Rocky Mountains. The border between them has been assumed as the altitude of 500 m a.s.l. The highest point is 8848 m (Mount Everest, the Himalayas). In the southern Himalayas it reaches 26° N. In the Southern Hemisphere it reaches a maximum of 6962 m a.s.l. (Aconcagua), in Africa near the equator (03°04′) it reaches 5895 m a.s.l (Kilimanjaro), and in New Guinea (04°04′) it reaches 4884 m a.s.l. (Carstensz).

According to Brown and Haggerty (1998), the land surface of the Northern Hemisphere, together with the ice sheet, is 100.407 × 10^6 km^2. Of this, 25.5 × 10^6 km^2 is encompassed by permafrost, which accounts for 25.4% of the Northern Hemisphere’s surface (Table 1). In the Southern Hemisphere, permafrost occurs at all latitudes. It covers the entire Antarctic: (i) as glacial permafrost, (ii) as subglacial permafrost partly in the cryotic state, and (iii) as the traditionally understood permafrost encompassing the nonglaciated area of this continent. Mountain permafrost encompasses mainly the great
### Table 1  Permafrost areas of the Northern Hemisphere based on extent (continuity).

<table>
<thead>
<tr>
<th>Permafrost extent</th>
<th>Total, $\times 10^6$ km$^2$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous (90–100%)</td>
<td>10.95 (43.1)</td>
</tr>
<tr>
<td>Discontinuous (50–90%)</td>
<td>4.42 (17.3)</td>
</tr>
<tr>
<td>Sporadic (10–50%)</td>
<td>3.99 (15.5)</td>
</tr>
<tr>
<td>Isolated (0–10%)</td>
<td>3.89 (15.3)</td>
</tr>
<tr>
<td>Relict</td>
<td>0.12 (0.5)</td>
</tr>
<tr>
<td>Glaciers</td>
<td>2.12 (8.3)</td>
</tr>
<tr>
<td>Total area</td>
<td>25.49 (100)</td>
</tr>
</tbody>
</table>


mountain range of the Andes and the mountains in New Zealand.

French (2007) states the permafrost area for the Northern Hemisphere to be $22.7 \times 10^6$ km$^2$, whereas for Southern Hemisphere it is $13.5 \times 10^6$ km$^2$, which encompasses the Antarctica. The total of these values, that is, $36.2$ million km$^2$ (out of $149 \times 10^6$ km$^2$ entire land surface of the Earth), is 24–26% of all land area.

According to Gruber (2012), the global surface covered with permafrost, along with the Antarctic and subsea permafrost, is estimated to be $16–21 \times 10^6$ km$^2$. This accounts for 9–14% of the land surface on Earth. The author claims that the maximal expectation for the occurrence of permafrost is the area of $25 \times 10^6$ km$^2$ (ca. 17%).

Such a great discrepancy in the estimation of the extent of permafrost results from the doubts regarding the ice sheet or the bedrock of the Antarctic.

**SEE ALSO:** Antarctica; Cryosphere studies: history; Glaciers; Ice sheets

### References


### Further reading

Petrocapitalism

Matthew T. Huber
Syracuse University, USA

It has become a truism that oil is the foundation of modern capitalist economies. It is the basis of entire states’ budget revenues. It drives global transport networks from North American suburban sprawl to air-based logistics. It is the number one traded global commodity. Its importance is not only because of its strategic importance as an energy source, but also because of its material qualities – its liquid propensity to flow makes it very cheap to transport over land and sea. In one sense, oil is the only energy commodity that circulates so widely (there are serious material and economic constraints to transporting coal, gas, and, for obvious reasons, “renewable” energies like hydro, solar, and wind power). Oil is also said to be a central object of “geopolitical” conflict – oil wars and imperial power rest on who has control over its flow.

Given oil’s centrality to modern industrial life, it is no surprise that geographers – particularly those with interests in political economy – have attempted to theorize the precise relations between petroleum and capitalism itself. As concerns around peak oil, the war in Iraq, climate change, and the resource curse began to proliferate in the early 2000s, geographers began to theorize a specific form of “petrocapitalism.” At first, this concept was used to understand a very specific constellation of forces wrapped up in petroleum extraction or production (particularly in the developing world) – violence, state power, money, local environmental degradation, and, of course, resistance (Sawyer 2001; Watts 2004). As concern with oil’s embeddedness in the reproduction of industrial societies intensified some geographers, and others, began to think about the wider relations between capitalism and oil consumption. While geography of extractive petrocapitalism is focused on specific points in space – what Gavin Bridge (2010, 1) calls the geographies of “the hole” – consumptive petrocapitalism is dispersed and virtually includes the entirety of capitalism itself. Such a wide view of petrocapitalism had led scholars to link oil with some of the foundational values of capitalist liberal economies – from democracy to freedom.

Extractive petrocapitalism: violence, money, and power

In its initial usage, petrocapitalism referred to specific forms of accumulation tied up with the extraction of oil. In his study of Nigeria, Watts (2004, 199) examined how, “petro-capitalism operates through a particular sort of ‘oil complex’ (a configuration of firm, state and community) that is territorially constituted through oil concessions.” This “oil complex” represents both the set of institutions meant to facilitate the extraction of petroleum wealth (multinational private companies, state companies, and state officials), and the local communities who often suffer the environmental and human rights abuses that follow. While the geography of oil extraction might mirror many other forms of extractive mineral economies, what distinguishes petrocapitalism is the tremendous wealth generated from this critical commodity. Often encompassing over 90% of state revenues in oil
exporting countries, oil becomes a “magical” commodity that promises “instant wealth” and dreams of national development (Coronil 1997). As Polish journalist Rizard Kapucinski remarked of oil-rich Iran, “Oil creates the illusion of a completely changed life, life without work, life for free” (quoted in Watts 2004, 213).

Perhaps the most important “institution” within the oil complex is the “petro-state.” In most areas of petroleum extraction (except the United States where private subsurface rights exist), all subsurface minerals are the property of the state. As Coronil (1997) discussed in relation to Venezuela, this means the state acts as the landowner – or the “landlord state” – that grants “concessions” to private oil companies – increasingly in conjunction with a state owned oil company – to explore for and extract oil wealth. The companies then pay the landlord state “rent” in the form of royalties, taxes, and other payments. Even in the most spectacular instances of geopolitical conflict and “oil imperialism” – like the invasion of Iraq in 2003 – the institutional power of the state as “landlord” retains considerable power to act as a “barrier” to forms of capital investment (see Labban 2008).

As Watts suggests, these “concessions” are territorially constituted in the specific locations in which the extraction happens. On the other hand, the wealth generated through this extraction is often funneled elsewhere. Watts (2004, 212) describes the accumulation of money in the petro-state “fiscal centralization.” As many in political science suggest (e.g., Ross 2012), this “oil curse” creates the phenomenon of the “rentier state” where massive windfalls of oil money create state corruption and lack of accountability to domestic populations. For example, while most Nigerians suffer incredible levels of poverty (per capita income is about US$290 per year) the Nigerian government spends lavishly and sizable funds have simply gone missing (Watts 2004, 201). Despite the tendency to pathologize the governments in this literature, it is also worth pointing out that in the context of increasing neoliberalization of oil policy the vast majority of the wealth generated is captured by private transnational corporations. For example, Sawyer (2001, 163) estimates that Texaco generated an astounding $23 billion between 1972 and 1991 – extraction that became even more profitable because the company did not invest in basic environmental safeguards (e.g., the lining of waste pits).

Given that oil extraction can be extremely damaging to both public health and the natural environment, local communities are often forced to bear the brunt of the costs of extraction while the wealth flows elsewhere. Watts (2004, 208) describes the Niger Delta as an “ecological disaster” where “76 percent of natural gas in the oil producing areas is flared.” This lead a visiting journalist to remark that, “some children have never known a dark night even though they have no electricity.” The water – so critical for fishing livelihoods in the Delta – has been decimated by an estimated “300 spills per year” leading to hydrocarbon levels an estimated “360 and 680 times the European Community permissible levels” (Watts 2004, 208).

Bridge (2004, 226) identifies a “scalar mismatch” between the local environmental and human consequences of all forms of mineral extraction and the national–global flows of benefits. This mismatch often leads to substantial forms of political unrest, resistance, and violence. In Ecuador, local indigenous tribes are still attempting to force Chevron (formerly Texaco) to clean up the toxic legacy of their nearly 20 years of operations. In Nigeria, there has been both violent and nonviolent forms of political mobilization – both of which have been brutally repressed by the Nigerian government with thousands killed, including famous nonviolent
environmental activist Ken Saro-Wiwa. Violence almost always accompanies “clearing” the site for extractive activities with displacement of local populations. In Sudan, oil concessions were advertised to investors as unoccupied land – “land, that is, which paramilitaries have first systematically rendered uninhabited by driving off the residents” (Ferguson 2005, 378–379). The sites of extraction themselves become heavily securitized “enclave economies” where companies often make use of state security forces and hired private mercenaries to effectively militarize extractive operations (Ferguson 2005, 378–379).

In its initial usage, petrocapitalism was a particular geography of accumulation centered upon the extraction of oil in the relatively few spots on the Earth where it is profitable to do so. Ultimately the wealth flowing to all the actors within the “oil complex” – corporations, petro-states, private security firms – hinges upon that particular moment and place of extraction. Yet, the travels of oil itself do not end in these sites of extraction. In fact, crude oil by itself is a relatively useless substance (a few oil-rich countries might burn it for electricity but 99% of crude oil is not used in its crude state). It must travel by pipeline and/or ocean tanker to a refinery where it is transformed into a multiplicity of “use-values” that then flow into countless channels of consumer markets – most notably transportation fuel like gasoline and diesel. If, as Mitchell (2011) suggests, we “follow” this dispersed geography of oil consumption, how does our vision of “petrocapitalism” change?

Consumptive petrocapitalism: flowing through daily life

In order to understand oil consumption as a form of “petrocapitalism” it is first necessary to avoid the many platitudes that exist around ideas of “oil addiction.” First, it is increasingly common to recognize that petroleum products have become absolutely central in the reproduction of industrial lifestyles. Most notably this includes gasoline fired automobility, but also the countless petroleum products that saturate daily life from chemicals, pharmaceutical products, plastics, and even food. Often taken for granted, oil appears to be the commodity that holds life together. Second, especially among critical political economists, this “addiction” to oil is often explained as the result of narrow corporate conspiracy among oil and automobile companies. The most notable example of this is the case of “National City Lines” (a conglomerate of auto, oil, and tire companies) who conspired to destroy much of the United States electric streetcar public transit system and replaced them with oil-fired buses.

An understanding of oil-based petrocapitalism requires a wider analysis. It is not enough to recognize oil’s centrality to social reproduction; the real question is how oil emerged historically as such a central commodity in the reproduction of labor power in capitalist societies. Again, oil is not consumed in its raw state, but as particular petroleum products that only become use values in relation to other commodities like the internal combustion engine. The story is also broader than an elite conspiracy. To understand mass oil consumption in the United States you first need to understand the historical process of suburbanization (especially in the post-World War II era) and its relation to wider regimes of capital accumulation. Specifically, the crisis of capitalism during the 1930s and state-led restructuring of capitalism led to a set of institutions – labor regulations, housing reforms, infrastructure development – that created the material conditions for Fordist mass consumption.

Essentially, the state empowered certain privileged sectors of the American working class to
PETROCAPITALISM

gain access to a new widely generalized form of space and energy-intensive life. While it may be true that this form of life is “wasteful” and environmentally disastrous, it served to save capitalism both economically and ideologically at a moment when its historical survival was very much in question. Only in this wider social context where masses of the population actually had the wages and government supports to purchase and use cars, single-family homes, and countless other petroleum-based commodities, does the centrality of oil to social reproduction take hold. Moreover, this form of life also became linked to wider ideologies of “The American Dream” – ideas of leisure, family, freedom, and space – that created ideological legitimacy for oil-based life in the postwar era (legitimacy that did not require a corporate plot).

Of course, this story is not confined to the United States. Essentially the world economy in post–World War II was constructed on the basis of increasing oil consumption. The United States not only ensured the development of Middle Eastern oil fields (e.g., most notably in Saudi Arabia and Iran) by Western capital – sometimes by force, as in the 1953 CIA-organized coup of the popular socialist leader Mohammad Mosaddegh in Iran. The United States also helped “rebuild” the European and Japanese economies to become more oil dependent. For example, the Marshall Plan included massive amounts of refinery construction in Europe to process new oil flowing from the Middle East (Painter 1984). While European states tax gasoline at much higher levels than the United States – reinforcing public commitments to public transit and “compact” cities – most nations remain almost entirely dependent upon “foreign” sources of oil mainly from Russia and the Middle East.

It is also critical to not singularize oil’s importance to postwar capitalism. The postwar economy was created upon the basis of a multiplicity of petroleum products. The “petrochemical” revolution of the post–World War II era was itself a product of the massive mobilization of all industry toward the war effort in the 1940s. Specifically, the government investment in catalytic cracking refinery capacity – to aid the mass production of high octane aviation fuel – created a massive expansion in the production of “olefins” like propylene and ethylene that became the basis of the postwar petrochemical industry based on plastics and other consumer chemical products. Other advances include the use of natural gas as the preferred feedstock in ammonia synthesis (as opposed to coal) which laid the basis for the massive expansion of synthetic nitrogen fertilizer production in the postwar era and the so-called green revolution in crop yields. By 1978 our food system had become so reliant on hydrocarbons – from diesel fired tractors to petrochemical pesticides – a prominent scientist claimed, “Modern agriculture is the use of land to convert petroleum into food” (Bartlett 1978, 880).

The embeddedness of oil in everyday social reproduction has increasingly led geographers and social theorists to attempt to grapple with the ways in which oil-based cultural practices prefigure wider political and ideological discourses. Urry (2013) calls for a “sociology of energy” that confronts the centrality of oil (and other fossil fuel energy) to the relational ties that constitute what we think of as “society” itself. Mitchell (2011) argues that carbon based fuels – and the various forms of movements and contestations over their flow – underlie the political forces we often equate with “democracy” (even as the sites of “extractive” petrocapitalism discussed above are often pathologized as “undemocratic”). Huber (2013) argues oil helped power and provision ideas of freedom
and “entrepreneurial life” that animated sub-
urban right-wing politics as America became
“neoliberal.”

All these insights combine to present a much
more dispersed understanding of petrocapital-
ism. While a focus on extraction has led to
insights into the politics of oil – as a specific object
extracted in a specific point in space – we are
now thinking about oil and energy as ways to
think about the ecology of politics more broadly.

Conclusion: beyond petrocapitalism?

Moving beyond petrocapitalism is relatively
straightforward in technical terms – the goal is to
find an alternative energy source in the transporta-
tion sector (which, for example, consumes 70% of
every barrel in the United States). While there
are a number of alternatives with varying degrees
of feasibility (e.g., hydrogen fuel cells, agrofuels),
the most promising recent developments have
been centered on the electrification of the transport
sector – that is, running cars (and maybe some
day trucks and airplanes) with rechargeable bat-
teries. Of course, as it stands today, the electric
power sector is fueled by a combination of fossil
fuels, nuclear, and some marginal “renewable”
sources (most notably hydropower, but increas-
ingly the green beacons of hope, solar, and wind
power). Thus, in order for electrification to
become a truly “green” alternative, the electric
power sector itself must be weaned off fossil fuels
and dangerous nuclear power.

Yet, it is critical to point out that the real
barriers to transcending petrocapitalism are polit-
ical. First and foremost, there are concentrated
and powerful institutions – from multinational
oil corporations (i.e., “Big Oil”) to petro-states
and their associated government-owned oil
companies (e.g., Saudi Aramco) – who have
much to lose from any decrease in our economic
dependence on oil. Their powerful control
over many countries’ “energy policy” (or lack
thereof) is substantial and entrenched. Yet, from
a more Gramscian political perspective, we must
also recognize that oil “hegemony” is rooted
in much broader political forces than simply
the control of “Big Oil” and the petro-states.
Again, the embeddedness of oil-based practice
with “consent” bound up in everyday notions
of freedom, home, and individuality are much
more formidable political challenges to over-
come. This is clearly the case in the United States
where even the prospect of a marginal increase in
the gasoline tax is met with widespread populist
refusal and outrage. Ultimately, if we insist on
coupling oil with the mode of production “cap-
talism,” it must be self-evident that efforts to
move beyond oil might require moving beyond
capitalism itself – specifically its requirements
for growth and accumulation over questions
of social and environmental benefit. Thus, a
move beyond petrocapitalism needs to be less
understood as a struggle over “technology” and
more a struggle over the political and economic
organization of society as a whole.

SEE ALSO: Energy resources and use; Energy
security; Environment and everyday life;
Environment and resources, political economy
of; Geopolitics; Livelihoods; Resource curse;
Resource extraction; States and development

References

Energy Crisis.” American Journal of Physics, 46(9):
876–888.
Bridge, G. 2004. “Contested Terrain: Mining and the
Environment.” Annual Review of Environment and
PETROCAPITALISM


Phenology

Mark D. Schwartz
University of Wisconsin–Milwaukee, USA

Phenology “refers to recurring plant and animal life cycle stages” and is also the scientific “study of these recurring plant and animal life cycle stages, especially their timing and relationships with weather and climate” (USA-NPN 2015a). Examples would include the opening of leaf buds on a shrub, the beginning of autumn leaf color on a tree, or the first appearance of a migratory bird species in spring. The timings of many of these phenomena can be quite sensitive to small changes in environmental conditions, such as air or soil temperature. Thus, phenological data provide an important independent measure of the impacts of climate change on the biosphere, which can be assessed at scales ranging from the individual plant to entire biomes.

History

Phenology before the dawn of history

Observing changes in the timing of plant and animal life cycle stages in the natural world has likely always been a part of the human experience. Early nomadic humans would have looked for signs to help pinpoint the migration timing and presence of animals, as well as any information to help them locate and take advantage of fruits, nuts, and other vegetative sources of food. The relevance of phenology must have increased considerably once humans started establishing permanent settlements and practicing agriculture, horticulture, and animal husbandry. Being at the same location over multiple years, and eventually over multiple generations, early farmers would realize that phenological changes in the plants and animals around them could be used as a generalized natural calendar to help direct their planting, care, harvest, and other management activities, especially in temperate regions with pronounced seasonal transitions. Ancient Chinese references, stories with phenological themes in both the Old and New Testaments of the Christian Bible, and other sources offer abundant evidence of the long understanding and utilization of phenology as an integral part of human agricultural practices over many millennia (Schwartz 2003, 3). However, such knowledge would have been highly localized, and much more useful in average than extreme years.

Relatively recent scientific study and terminology

While the practical utilization of general phenological principles in association with agriculture is ancient, the emergence of phenology as a distinct area of scientific inquiry is quite recent. The father of modern scientific phenology was the Swedish botanist Carolus Linnaeus, who outlined methods for developing annual plant calendars, and suggested that these botanical data when combined with climatological observations could be used to show how areas were different (Lieth 1974, 25). However, it was another hundred years before the term “phenology” was coined by Belgian botanist Charles Morren in 1853 (Lieth 1974, 25). Even so, the term and its definition are somewhat
PHENOLOGY

problematic. Phenology is derived from the Greek word *phaino* (which means to show or to appear). While practically speaking the study of things that “show or appear” could mean just about anything, originally the focus was almost exclusively on plants (especially woody plants and wild flowers), with just a few exceptions such as birds, until around the 1950s when more biological organisms began to be considered (Lieth 1974). In recent decades the definition has been generalized even further to include changes in remotely sensed reflectance values (which change in relation to vegetation growth stages) as well as nonbiological events driven by seasonal changes in climate, for example spring ice thaw on fresh water lakes.

Published comprehensive volumes

Schnelle’s book (1955) *Pflanzen-Phänologie* (Plant Phenology) appears to be the first comprehensive volume devoted to phenology as a distinct area of scientific inquiry. He noted the multidisciplinary nature of phenological science and its traditional place as a supporting science within numerous fields (botany, phytopathology, agronomy, agricultural meteorology, climatology, and geography) that explore environmental influences on plant development. However, Schnelle asserted that phenological studies are a holistic way to use plant development as integrative tools that provide a critical focus for evaluating the simultaneous interactions of multiple environmental measurements. This book stimulated increased interest in phenological observations and studies in Europe (including development of an International Phenological Garden Network); however, its worldwide impact has been reduced due to being written in German, and never translated into English.

Lieth’s book (1974, 3) *Phenology and Seasonality Modeling* was a direct result of the International Biological Program (IBP) and work of the United States’ IBP Phenology Committee. This seminal volume clearly established a broader scope for phenological studies as encompassing not only plants but also animals, whole ecosystems, and newly emerging remotely sensed measurements available from sensors on board satellites. Lieth and his colleagues introduced numerous perspectives through which phenological analyses could enhance ecological studies, as well as new modeling and graphical presentation techniques.

Without question, rising recognition in the late 1990s that phenological observations could serve as an additional independent measure of the impacts of global climate change on the biosphere, led to exponential growth of the scientific community’s interest in phenology. The need for an updated volume to summarize the current state of research was realized with the publication of *Phenology: An Integrative Environmental Science* (Schwartz 2003). This book documented the continued broadening of phenological studies to continents beyond Europe and North America, bioclimatic zones outside of temperate moist forests, and additionally to life forms other than plants. Updated synopses of analyses techniques included plant and animal life cycle modeling, new developments using satellite remote sensing-derived measures, and emerging applications in global change studies and ecosystem assessment through carbon/energy flux measurements. The recently published second edition of *Phenology: An Integrative Environmental Science* (Schwartz 2013) provides an overall update on progress in phenological research over the past decade, while highlighting newly emerging approaches such as the use of near-surface cameras as a source for remotely sensed observations of seasonal vegetation changes, and examples of developments in field-based college- and university-level
Phenological education which take advantage of the latest technology.

**Techniques**

**Types of information**

Phenological observations of *higher plants* have traditionally included the timing of visibly distinguishable stages of growth and development, as well as senescence, which occur during seasonal progressions over the course of a year. So for woody perennial plants this would start with the swelling and opening of leaf and flower buds in spring, progress to full leaf and flower development in early summer, in some cases continue through fruiting in mid- to late summer, and end with leaf coloring and fall in autumn. In annual *grasses*, *herbs*, and for *fungi*, observations are generally more restricted to initial appearance, flowering, and in some cases fruiting and senescence over a shorter time period.

*Animal* phenological observations for organisms are generally much simpler than those recorded for plants. Often for warm-blooded animals this can be arrival dates for migratory species such as birds, or appearance dates for those that undergo hibernation, for example various rodents. Other typical life cycle events recorded for warm-blooded animals would including behaviors related to reproduction (nesting, mating, egg-laying, birth and development of young). Cold-blooded species observations are more diverse, depending on the life form types. Observations of reptiles tend to be similar to those for warm-blooded creatures, while those for insects would typically include growth and development stages (egg, larva, pupae, and adult) that would parallel approaches used for plant observations.

Traditionally, (as in the definitions above) phenological observations have been reserved for use solely with organisms. However, recently, no doubt in part due to the exceptionally broad nature of “to show or to appear,” coupled with the renewed visibility of phenological science, publications have appeared applying the term phenology to studies of *nonbiological natural processes* that are also driven by seasonally modulating environmental factors. Some of the best examples of these come from hydrological science, such as timings of winter ice cover and spring thawing in freshwater lakes and rivers, as well as timings of peak and low water flow in rivers. So while it appears likely that the “biological only” definition of phenology will remain dominant, it is important to recognize that a wide range of nonbiological seasonal phenomena are now considered by some to be included in phenological studies as well.

**Types of observations**

For *ground-based visual phenological observations* to be most scientifically valuable they need to be recorded according to a clearly defined and reproducible procedure, typically referred to as an “observation protocol.” First, a couple of terms need to be defined which are crucial to the structural development and subsequent understanding of phenological protocols: “phenophases” and “phenological events.” A *phenophase* is “an observable stage or phase in the annual life cycle of a plant or animal that can be defined by a start and end point.” Phenophases usually last from a few days to several weeks (USA-NPN 2015a). Further, a *phenological event* is “a precisely defined point in the life cycle of a plant or animal, typically expressing the start or end point of a phenophase.” The occurrence of a phenological event can at least theoretically be tied to a specific date and time (USA-NPN 2015a).

Thus, protocols are generally constructed so as to allow multiple observers at a single site, as well...
PHENOLOGY

as different observers at multiple sites, to consistently and precisely identify a specific point in the life cycle of a plant or animal using features that are clearly visible. The most successful protocols are those that define phenological events that are easily recognized visually and then comprehensively described using both a written description and an image. The complexity of a protocol (in terms of the number of events and their detail) is generally related to requirements of the proposed objectives of the studies for which they are designed. Due to the time and effort involved in taking visual phenological observations, most protocols are designed with a small number of events, so as to maximize the likelihood that the majority will be recorded in all years and sites where it is utilized.

Phenological observations are traditionally recorded as the calendar date on which they occur. Thus, one day is the smallest difference that can be measured between two individual phenological events. Further, in order to make comparisons between observations taken in different years and at different locations, calendar dates are often translated into a day-of-the-year format, most commonly with January 1 equal to day number 1. Several limitations of this approach to recording phenological observations have been noted. First, during periods when no events are occurring, no information is recorded. This imposes a certain temporal resolution on the sequence of phenological information being recorded; meaning that unless the protocol includes many events and details, large periods of development may not be well represented. Lastly, there is essentially no information available to assess the error of the observation itself. While some approaches to overcome this problem have been suggested, such as requiring the observer to simultaneously record an image of what they observe (largely impractical in the past, but becoming more feasible as data storage technology improves), the supervising authority responsible for data collection can do little more than hope that the observer followed the protocol faithfully.

For plant phenology, a number of growth processes proceed in continuous fashion, such as leaf development in spring and leaf senescence in autumn. Thus, when monitoring these processes, it is possible and desirable for protocols to be established such that whenever an observation is taken, a level of development is able to be recorded. The advantage in this approach is that it allows phenological information to be recorded at up to the highest temporal resolution possible (every day).

Status monitoring is a different approach to recording phenological observations which addresses some of the limitations inherent in the traditional event date-recording technique. Observers employing status monitoring still regularly check the phenological status of specific individuals from plant and animal species at selected locations. However, phenological information is reported as answers to “a series of yes/no questions.” “This approach ensures the capture of negative data (when the phenophase is not occurring), repeat events …” (such as secondary blooming later in the same year), “… and allows for an estimation of the uncertainty around the beginning or end date of a life stage” (USA-NPN 2015b). Specifically, when status monitoring is used, negative data recorded immediately before positive event data would indicate low error/high reliability, and the periods when neither positive nor negative data are recorded can both contribute to overall assessment of measurement uncertainty.

While phenological observations can be (and often are) taken in isolation at individual sites, their value increases considerably when observations using the same protocol are taken at multiple sites. Such phenological networks allow
assessment of the impacts of varying environmental conditions on phenological events at different locations, and in turn that of the collective effects of changing environmental conditions across all sites. Another consideration is scale. Most phenological observation networks are set up to assess differences across regions or nations. Networks designed to span large countries, continents, or climatic zones face difficulties in species selection, as few or no species are present across all portions of such large areas. In some cases phenological networks can be set up at high spatial density in small areas, in order to explore and separate the impacts of environment, genetics, and other factors on phenological responses within a single species.

Satellite-based sensor-derived observations
The advent of sensor systems on board satellites able to capture reflected and emitted radiation in specific portions of the electromagnetic spectrum systematically has been transformative to many areas of science, including plant phenology. Information about the growth stage of vegetation can in theory be assessed using the same methodology over regional to continental scales at regular time intervals. However, there are a number of important issues which must be considered when utilizing these data in specific applications in practice. Among the most universal are: (i) orbital mechanics and cost generally impose an inverse relationship between the spatial resolution of the smallest individual image element (pixel) size and the frequency at which imagery can be obtained; (ii) cloud cover and other atmospheric interference will reduce the amount of usable data obtained; and (iii) comparisons of conventional ground-based visual phenological observations to satellite-derived reflectance measures are often problematic, given the inherent spatial mismatch of these data as well as the different observational approaches. Despite these difficulties, phenological applications of remote sensing data obtained from satellites, especially when merged with surface-based data, remain a fruitful area for future research.

Near-surface sensor-derived observations
A recent and exciting development in remote sensing phenology is the use of sensors deployed from fixed structures, such as towers or buildings, rather than moving (airborne or orbital) platforms. Such “near-surface” remote sensing data collection offers a number of distinct advantages to phenological studies, especially when used to bridge the spatial and technical gap between satellite-derived and conventional ground-based visual phenological observations. Among the most important of these advantages are: (i) providing essentially continuous data with minimal atmospheric interference; and (ii) allowing detailed information about well-defined groups of plants across the entire yearly cycle of vegetative growth and senescence (Richardson, Klosterman, and Toomey 2013, 414–415).

Model output (plants and animals)
Since a fundamental aspect of phenological information is its intimate link with environmental drivers, a large portion of phenological research has been directed toward understanding the nature of these connections. When successful, such analyses can serve as the basis for constructing models of phenological behavior. Phenological models can subsequently provide predictions of phenological responses in places and times where the environmental driver data are available but actual phenological data are lacking. Such models can serve many purposes at multiple scales, such as predicting the flowering
and fruiting of a specific horticultural species within one orchard based on daily air temperatures (from a combination of past data collected and future forecasts) or reconstructing the timing of the start of the growing seasons from daily air temperatures over the period of instrumental records across North America.

Selected applications

Agriculture and horticulture

Given their ancient connections, it is not at all surprising that agriculture and horticulture remain a healthy area for phenological applications (Chmielewski 2013). Traditional areas that continue to be quite useful include defining the growing seasons for various crops at selected locations, assessing an area’s climatological variables for compatibility with new crops, and evaluating the risk of frost damage to crops based on their state of phenological development. In addition to these uses other common areas include the timing of management practices where the growth stage of the plant is important, such as the timing of pesticide or fertilizer application, as well as irrigation schedules. Agricultural modeling of plant development and yield is also highly dependent on phenological information in order to interpret the interactions of microclimate and other site-dependent factors (Chmielewski 2013).

Global change

The Intergovernmental Panel on Climate Change (IPCC) noted that phenology “is perhaps the simplest process in which to track changes in the ecology of species in response to climate change” (IPCC 2007). Further, phenological data (together with climatic and remote sensing data) provide a third independent measure of these impacts which can be assessed at scales ranging from the individual plant to entire biomes. Phenological data collection and research will continue to inform global change science as researchers grapple with understanding the processes and impacts, as well as search for adaptation strategies for both managed and natural ecosystems.

Education

The integrative natures of phenological data make them well suited to education applications in both formal and informal settings. For example, when individuals participate in data collection networks as observers, they not only gain an immediate and practical understanding of the life history of the organisms they observe, but often also become sensitized to the variations in environmental drivers. In more formal settings, phenological data can facilitate course assignments designed for almost any age level that span the entire range of scientific endeavors, including field methods, hypotheses formulation, numerical analyses, and graphical display of results.

SEE ALSO: Agroclimatology; Climate change and biogeography; Earth system science; Ecosystem services; Environmental management; Environmental science; Global climate change

References


Phenomenology

Alex Jeffrey
University of Cambridge, UK

Phenomenology, or the study of human consciousness and experience, has a long historical lineage within the study of philosophy. Immanuel Kant used the term *phenomena* in the eighteenth century to refer to events and categories that could be apprehended by the physical senses. This approach was later challenged by Hegel who sought to explore how human consciousness of the external world was formed. Consequently, Hegel’s work focused on the challenge of comprehending an external reality, in doing so encouraging reflection on the ways in which objects external to human consciousness are constructed in terms available to the human mind. Such questions of consciousness and human experience were later taken up by Edmund Husserl, Martin Heidegger, and Maurice Merleau-Ponty (among many others), where phenomenological reflection was directed at foundational philosophical questions relating to the objectivism of science, the ability to establish stable truths about the world, and the separation between humans and other animals and objects.

Phenomenology has a shorter history within the study of human geography. As critiques of positivism emerged in the late 1960s and early 1970s questions of existence and being, previously felt to be secondary to concerns of distribution and location, came to the fore. Humanistic geography sought to encourage a more “anthropocentric” (human centered) understanding of the world, and drew on phenomenology — alongside other humanist philosophies such as existentialism — as part of its philosophical toolkit. We must be clear that this was not always a case of wholesale rejection of science: many scholars felt that spatial science was adequate for *describing* geographical processes but they simultaneously felt less confident that they could use scientific approaches to *understand* the meanings behind human choices, distributions, and inequalities. Early humanistic geography was concerned that the models and quantification of spatial science had “de-humanized” geography inquiry, where the complexity of human decision-making was reduced to rational economic choices, pointing in particular to the reliance of modelers on price mechanisms and other market forces to predict human behavior. We must be careful not to characterize the sophistication of quantitative work in summarizing these critiques, but on a philosophical level the production of models was seen by humanistic geographers as focusing on the general rather than the particular, thinking of humans as populations as opposed to individuals. But this was not simply a philosophical reflection: by emphasizing variation and, therefore, inequality, humanistic perspectives could also focus on the ways in which social processes of marginalization or exclusion operate in spatial terms.

A number of traits can be identified that characterize phenomenology’s influence on geographical scholarship. The first is an interest in human subjectivity: phenomenologists are concerned with how humans experience their environment in unique ways. The second is a preoccupation with the construction of knowledge: phenomenologists focus on how individu-
PHENOMENOLOGY

give meaning to their objects of consciousness. The third and perhaps most contentious is the search for shared categories: the goal of such study is the identification of elements in individual consciousness that control the allocation of meaning, what sometimes are referred to as “essences.” But while this three-part division gives a sense of some of the thematic concerns of geographical phenomenologists, it is perhaps more appropriate to think of phenomenology as a methodological approach than a body of theory. For example, uses of phenomenology in geography often draw on the work of Edmund Husserl, who was interested in developing a philosophical method that would allow the scholar to go “back to things themselves” (Entrikin 1976, 617). At heart, then, phenomenology has been concerned with challenging abstraction and thinking about the lived experience of human existence. In terms of scholarly practice, this approach advocates qualitative empirical inquiry and the importance of the humans at the center of this process. But in getting “back to things themselves” Husserl was also advocating setting aside all preconceptions of the nature of human experience and, instead, examining the world afresh.

We can interpret this move to phenomenological reflection in two ways: it seems to reflect a return to a more tangible sense of human experience while also posing a radical empiricist challenge to the production of scientific categories. In Husserl’s view, phenomenologists rejected preconceptions and traced instead a science of beginnings, where the very origins of particular understandings of the world are examined. In many ways, and this is a point that John Pickles (1985) makes across his analysis of phenomenology in geography, the phenomenological challenge to objective science pre-empts many of the more recent genealogical investigations of the production of scientific institutions and expertise (after Michel Foucault, among others) or the wider field of the sociology of scientific knowledge. It is the case that phenomenological studies have explored how subjective elements of human consciousness shape science, in doing so exploring how the influences of history, life, and bodily cognition influence scientific knowledge. But this was not in order to argue for untrammeled subjectivity, an unmooring of knowledge from any exterior reality, but rather to explore how particular themes, categories, and abstractions are produced in specific social and scientific conditions.

Perhaps the most tangible expression of this tension between objective and subjective knowledge production in geography comes in the distinction between space and place. While space has often been presented as an objective category, a plane or volume that expands over a particular physical extent, place is often understood as its antithesis, the subjective experience of space, constructed from human physical sense, emotions, and memories. Of course, this distinction is challenged in thought and practice, not least since the idea of a stable, detached, and abstract image of space shared between all human subjects is itself open to intellectual and political challenge. But place has been a key theme of phenomenological work in geography, a means of capturing and conveying the subjective experience of particular sites. A key proponent of this school was Edward Relph (1976, 43) who emphasized the phenomenological method of centering on human experience:

The basic meaning of place, its essence, does not therefore come from locations, nor from the trivial functions that places serve, nor from the community that occupies it, nor from superficial or mundane experiences … The essence of place lies in the largely unconscious intentionality that defines places as profound centers of human experience.
By drawing on Husserl’s philosophical approaches, phenomenologists have attempted to study human consciousness and in so doing discern shared cognitive categories that underpin experiences both across space and time. An example of the ambition of this style of scholarship can be found in Yi-Fu Tuan’s (1974) *Topophilia* a text that sought to connect sentiment with place. What strikes the reader most forcibly in Tuan’s account is the wide diversity of examples marshaled to explore how motivations, values, and emotions shape human interactions with the environment. Tuan argues for the supposed *universality* of certain forms of cognition (e.g., five senses and stereoscopic vision) in contrast to the *variability* of other factors (e.g., culture and biography). The copious material, from an eclectic array of sites and epochs, aims to discern the individual from the transcendental aspects of place interpretation. This may seem rather subjective and almost playful, but Tuan directs such analysis toward political and environmental ends:

The themes taken up here – perception, attitude and value – prepare us, first of all, to understand ourselves. Without self-understanding we cannot hope for enduring solutions to environmental problems, which are fundamentally human problems. And human problems, whether they be economic, political or social, hinge on the psychological pole of motivation, on the values and attitudes that direct energies to goals.

In this excerpt Tuan sets out two principles of a phenomenological approach that have been discussed: an interest in subjectivity (“understanding ourselves”) and the construction of knowledge (“values and attitudes”). These seem consistent with a question for the variability in human values of place. But Tuan (1974, 2) also alludes to universality, and in particular the search for *essences* of human values concerning place:

What can be the common ground between a detailed analysis of the shopping behavior of housewives in Ames, Iowa and a grand survey of the Christian doctrine of nature? Or between the study of color symbolism as a universal trait and the history of landscape painting? A possible reply is that somehow they all bear on the way human beings respond to their physical setting – their perception of it and the value they put on it.

There are numerous criticisms of phenomenological perspectives: on the basis of its potential blindness to questions of gender, sexuality, and race; for its lack of attention to political and economic structures, and for its underlying humanist philosophy. Reflecting wider critiques of geography in the 1980s and early 1990s, feminist scholars criticized the rather uniform imagination of human subjectivity promoted in some phenomenological work, a subject unmarked by gender difference. This critique exposes a troubling essentialism that underpinned some phenomenology: that it was seeking to uncover general transcendental essences that were shared among all humans. But in doing so, forms of inequality and oppression that mark human experience may be underplayed or ignored. By emphasizing issues of positionality and embodiment feminist geographers have been at the forefront of inserting cultural, bodily, and historical contexts into debates concerning the marginalization and oppression of women. Reflecting a wider feminist critique of the masculinist nature of human geography, “gender blindness” was anything but blind; it led to studies focusing on male-dominated spaces and lifestyles but imagined as universal human experience. In this sense, a gender critique is not necessarily a philosophical challenge to the tenets of phenomenology, but rather a challenge to a sense of an undifferentiated human consciousness.

Reflecting this challenge to the unmarked human subject, critics of phenomenological and humanistic approaches also questioned the
PHENOMENOLOGY

extent to which such studies engaged with the different structural and material factors that shape life choices. As Gregory (1981, 16) states “the materiality of social life is weakly developed in modern humanism, and as a result it inevitably encounters severe difficulties in comprehending objective societal forces.” This lack of attention to structural factors, in particular questions of class, was only amplified by the parallel development of humanistic and Marxist perspectives in human geography over the 1970s. This dual intellectual flowering can be crudely seen as two sides of the structure–agency debate, an ongoing discussion between those who see human actions as determined by external forces (structure) and those who sought to foreground the significance of human decision-making and free will (agency). In neither Marxist nor humanistic cases would proponents see their work as fitting an image of human action as either determined by overbearing economic structures or the product of free-floating and voluntarist human consciousness. But, nevertheless, the focus on human subjectivity is often interpreted as phenomenology’s overemphasis on agency, where – in the words of Entrikin and Tepple (2006, 31) – “experiences, attitudes, and beliefs, as well as moral and aesthetic judgment” inform perceptions of the environment.

The question of human agency is also central to a third line of critique emerging from posthuman (or antihuman) philosophical approaches. This field is diverse and evolving, but at its root is a desire to challenge the centrality of human reason and experience as the center of philosophical and social inquiry. As Braun (2004) notes, the flourishing of deconstruction as a form of geographical research method has challenged the centrality of the human subject, as critical scrutiny turned to the form and effects of language, the relational nature of identity, and the instability of essential truths. Rather than focusing on the operation of human consciousness and experience, this turn to deconstruction has involved exploring instead how the “figure of the human” is formed as an identity separate from other objects or beings. This point flows into a more general philosophical perspective that seeks to examine the world as more-than-human, widening the scope of geographical and philosophical inquiry to include objects, animals, and atmospheres. Such approaches use theoretical frameworks derived from Latour’s actor-network theory or Deleuze’s concept of assemblages (among others) to grasp the spatial complexity of the social world. Of course, scholars have challenged the extent to which this move constitutes a genuine shift from the pre-occupations of phenomenologists, interested as they were in the relational and the material aspects of the world. In particular, Simonsen (2013) encourages a re-reading of phenomenological literature, particularly that of Merleau-Ponty, to examine how this perspective can encourage an understanding of experience that fosters an openness to (human and nonhuman) others: one that does not reduce them to being the same, but does not preclude mutual understandings.

It is in this final point that we can see the emergence of what has been termed a “critical phenomenology,” challenging a sense of the rational human subject but retaining an interest in (diversely experienced) human subjectivity. This perspective retains what is perhaps at the core of phenomenology’s contribution to human geography: the phenomenological method. Rather than an overarching philosophical message, this method involves understanding engagements with place within the terms and categories of everyday life and individual human subjects. It is interested in the generative nature of consciousness, that terms and objects do not preexist the dialogic (that is, two-way
and relational) processes through which they are comprehended and understood. And they unfold in distinct embodied situations, as sensitive to questions of gender, class, race, and sexuality (among other lines of difference) as any other critical account of human geography.

SEE ALSO: Existentialism; Humanistic geography; Marxist geography

References


Photogrammetry is defined by the American Society for Photogrammetry and Remote Sensing as “the art, science, and technology of obtaining reliable information about physical objects and the environment through the processes of recording, measuring, and interpreting photographic images and patterns of recorded radiant electromagnetic energy and other phenomena.” Simply, photogrammetry allows 3-D measurements (e.g., position, orientation, shape, and size) of objects from photographs.

Photogrammetry is as old as modern photography and can be dated to the mid-nineteenth century (Konecny 1985). Over the past 80 years, the principal application of photogrammetry has been in the compilation of maps from aerial photographs. In recent decades, the development of high-resolution satellite imaging and close-range techniques have facilitated the application of photogrammetry to many other fields, such as Earth observation, environmental monitoring, smart cities, architecture, industrial inspection, robotics, and so on.

This entry provides a brief review of the historical development of photogrammetry and presents the fundamental techniques for deriving 3-D information from imagery via photogrammetry. Some observations and considerations about the future development of photogrammetry are also presented.

Historical development of photogrammetry

Photogrammetry began soon after the invention of photography in 1839. In 1849, Aimé Laussedat was the first person to use terrestrial photographs for the compilation of a topographic map and is now referred to as the “father of photogrammetry” (Birdseye 1940). In 1893, Albrecht Meydenbauer was the first person to use the term photogrammetry. He also designed the first wide-angle lens for topographical mapping and architectural surveying (Meyer 1987). The 1900s were pioneering years in the development of photogrammetry, with achievements including the generation of topographic maps based on techniques such as a “photographic plane table” using terrestrial photography or aerial photography supported by kites or balloons (Konecny 1985).

With the Wright brothers’ invention of the airplane in 1903, the development of photogrammetry entered a prosperous era due to the better camera platform. Since then, the development of photogrammetry has followed three development stages: analog, analytical, and digital photogrammetry.

Analog photogrammetry

The theory of stereoscopic vision widely used in the 1900s provided the foundations for analog photogrammetry. When looking at an object at a particular distance, our eyes simultaneously focus on and converge upon the object. The angle of convergence is called the parallactic angle. Objects at different distances from the
viewer are perceived through different parallactic angles. Due to the parallactic angle, the image of an object falls on different locations on the retinas of the left and right eyes. Points at different distances from the eyes appear at different relative locations. The difference in the position of the points on images is called parallax, thus changes in the parallactic angle result in parallax. The measurement of parallax offers an accurate method of measuring height or depth from stereo images. Figure 1 illustrates the principle of height determination from stereoscopic vision.

In Figure 1, assuming that two cameras (E_l and E_r) take images of an object at point P, the object’s image will appear as two image points (p_l and p_r) on the left and right images. Their locations in the images along the x direction are x_l and x_r. The parallax is determined as \( p = x_l - x_r \). The height can then be computed from the parallax measurements:

\[
H_p = H - \frac{Bf}{p}
\]

where \( B \) is the baseline length between the two cameras and \( f \) is the focal length of the camera. It can be seen that parallax of any point is inversely proportional to its distance from the camera and that parallax due to height occurs only on the x axis (the direction of flight for aerial images).

If looking at a stereo pair of images, with one eye looking only at the first image and the other looking only at the second image, the scene would appear as a 3-D image. Viewing a stereopair is actually quite difficult without the aid of a mechanical device, the simplest of which is called a stereoscope. A stereoscope set up for viewing a stereopair is shown in Figure 2.

From the beginning of the twentieth century, analog photogrammetry was developed based on the above theory of stereoscopic vision. In 1908, Eduard von Orel invented the first stereoautograph. The development of this plotter was significant because its construction principles made terrestrial photogrammetry practical in mountainous areas by allowing the operator to trace elevation contours directly (Collier 2002). After 50 years of development, this type of instrument reached maturity in the 1960s. As these instruments used an optical or a mechanical projection device, or a combination of them, to simulate the imaging process and to intersect the 3-D positions of the objects in the images, they were called analog photogrammetric instruments. During the evolution of analog photogrammetry, the main focus of development was on expensive stereoscope instruments.
Analytical photogrammetry

Analytical photogrammetry began with the invention of the computer in 1950 (Konecny 1985), which allowed digital projections to replace the physical projections (e.g., the optical or mechanical projections) used in analog photogrammetry. Digital projection uses a computer to calculate the 3-D positions of objects in images in real time, based on the colinearity equation. Fundamental to the colinearity equation is a perspective projection, where a point in the real world, its image point, and the perspective center of the image lie on one straight line. The colinearity equation is described as:

\[
\begin{align*}
    x - x_0 &= -f \frac{m_{11}(X - X_S) + m_{12}(Y - Y_S) + m_{13}(Z - Z_S)}{m_{31}(X - X_S) + m_{32}(Y - Y_S) + m_{33}(Z - Z_S)} \\
    y - y_0 &= -f \frac{m_{21}(X - X_S) + m_{22}(Y - Y_S) + m_{23}(Z - Z_S)}{m_{31}(X - X_S) + m_{32}(Y - Y_S) + m_{33}(Z - Z_S)}
\end{align*}
\]  

\(2\)
This equation represents a direct link between an image point \((x, y)\) and its 3-D position \((X, Y, Z)\) in the object space. \((x_0, y_0)\) is the principal point (the foot of the perpendicular on the image of the perspective center) and \(f\) is the focal length of the camera. \((X_S, Y_S, Z_S)\) are the coordinates of the camera center in the object space and \(m_{ij}\) are the elements of a rotation matrix that is determined by three rotation angles \((\phi, \omega, k)\) of the camera frame with respect to the object space. However, as each equation represents a straight line, the conjugate image points from a stereo pair of images need to be measured to compute the object point’s 3-D position. This process is referred to as space intersection. The colinearity equation provided the theoretical foundation for analytical photogrammetry.

In the early 1950s, Everett Merritt published works on analytical photogrammetry. He developed a series of analytical solutions for camera calibration, space resection, interior and exterior orientation (EO), relative and absolute orientation of stereo pairs, and analytical control extension (Doyle 1964). In 1955, Duane Brown developed new approaches for camera calibration and the mathematical formulation of the bundle adjustment. This was a significant development because it involved the simultaneous solution of the EO parameters of the camera and the coordinates of the survey points, along with the interior orientation (IO) and systematic radial lens distortion. In 1957, Uuno Helava developed the first analytical stereoplotter (Konecny 1985). A computer was used to drive the instrument around the stereomodel and to digitally transform the coordinates between the image and the map. Various types of analytical stereoplotters were developed, reaching a climax in the 1980s due to the technological improvements in large-scale integrated chips, personal computers, and interface technology. Figure 3 shows an example of an analytical photogrammetric system.

Significant technological developments in analytical photogrammetry were also made at this time. For example, the bundle adjustment for large photogrammetric blocks with self-calibration, developed by Duane Brown in

![Figure 3](image-url)  
**Figure 3**  An example of an analytical photogrammetric system.
the 1960s, improved the accuracy and reliability of photogrammetric adjustments. The direct linear transformation method developed by Sam Karara in 1971 enabled photogrammetric applications with non-metric cameras (Wolf 2001). These developments in technologies and instruments led to immense growth in the application of analytical photogrammetry in various fields during the 1980s.

Digital photogrammetry

The previously described analytical methods required an operator to view the photographs and place floating marks in the correct positions to derive the 3-D information. Digital photogrammetry originated from the search for ways of automating the manual work involved in analytical photogrammetry. For example, the task of placing the floating marks on photographs was replaced by image matching, which involves identifying the conjugate points (points representing the same image texture) on a pair of digital images, from which the 3-D coordinates of the point can be computed. Such processes automate and speed up the extraction of 3-D information from a stereo pair of images. Digital photogrammetry also automated other processes, such as image orientations.

Although digital photogrammetry originated in the 1950s, major research activities did not begin until the 1980s, sparked by significant advances in electronics and computing, such as digital cameras, parallel processing, and increased storage capacity (Schenk 1999). In 1957, Gilbert Louis Hobrough first demonstrated the concept of image correlation on a Kelsh plotter. Due to the technology at the time, the correlation process was analog and hardware was used to compare the gray levels of the images (Schenk 1999). In 1967, Hobrough developed the gestalt photo mapper, an automated orthophotographic system that used the correlations between stereo images. The system consisted of a scanner, correlator, computer system, operator console, and input/output device. Uuno Helava also played a central role in the development of digital photogrammetry, helping to develop digital photogrammetric workstations for the Defense Mapping Agency in 1986. Zhizuo Wang of China presented his ideas and solutions for a fully automatic digital photogrammetric system in 1978 and developed the WUDAMS, a fully automatic digital photogrammetric system (Wang 1998). In the 1990s, the WUDAMS was upgraded to a digital photogrammetric workstation, Vitruvzo, and a new generation, DPGrid, was recently developed based on network computing and cluster parallel processing (Zhang et al. 2011). In the early 2000s, Airbus Defence and Space in France released a new generation, fully automatic photogrammetric system named PIXEL FACTORY. Using mass parallel computing technology and native open architecture, PIXEL FACTORY is capable of automatically processing vast numbers of images to produce a wide range of 3-D cartographic end products such as digital surface models (DSMs), digital terrain models (DTMs), and TrueOrtho images.

3-D from imagery via photogrammetry

Deriving 3-D information from imagery via photogrammetry actually involves the reversal of the photographic process. If the position and orientation information of the photography ray when taking the images can be recovered, the reverse process can be achieved so that the 3-D information can be derived from the image. The process of recovering the photographic ray’s position and orientation information is called image orientation.
Image orientation

Image orientation includes interior orientation (IO) and exterior orientation (EO). The former derives the relationship between the image measurement and the image-space coordinate systems, while the latter derives the relationship between the image-space and object-space coordinates. The parameters derived from the image orientation enable the 3-D positions of objects to be calculated from their corresponding image points through the aforementioned colinearity equation.

The interior orientation relates the coordinates measured on the image to those of the object to be measured. To do so, it is necessary to establish the location of the principal point \((x_0, y_0)\) in the image-space coordinate system, as illustrated in Figure 4. The principal point \((x_0, y_0)\) and the camera focal length \(f\) are referred to as the IO parameters in the colinearity equation, as they are the camera’s intrinsic parameters and, thus, do not change when the location and orientation of the camera change.

Taking aerial images as an example, normally there are at least four fiducial marks distributed in the four corners of the image. These marks have known coordinates in the image-coordinate system; these are measured and used as observations. From these observations, a transformation between the observation and image coordinate systems is computed, which can then be used to determine the principal point \((x_0, y_0)\) in the image-space coordinate system and transform other coordinates measured on the image to the image-space coordinate system.

The images taken by cameras may have distortions, as illustrated in Figure 4, that need to be estimated and calibrated for accurate measurement. In the case of film-based images, distortions may occur because of lens distortion and other factors, such as stretching or shrinkage of the film due to handling, processing, or storage. Distortions in digitally recorded images may occur because of lens distortion, dissimilar pixel spacing, or differences in the pixel dimensions of the imaging device. Ultimately, the shape of an object measured on the image must be the same as that of the image when it was recorded. This can be achieved by adding distortion correction parameters to the left side of the colinearity equation in equation 2. The camera distortion parameters also belong to the IO parameters.

The image IO parameters are normally calibrated separately through a special control field with precisely measured targets as ground truth, or through self-calibration approaches.

Figure 4  (a) Image IO parameters \(f\) and \((x_0, y_0)\); (b) image distortions.
that incorporate the IO parameters into a photogrammetric bundle adjustment process so that they can all be solved together and simultaneously with other unknowns.

There are six EO parameters describing the relationship between the image’s and the object’s coordinate systems: the aforementioned three rotation angles ($\phi$, $\omega$, $k$), which describe the angular relationships, and $(X_S, Y_S, Z_S)$, which describe the location of the point of exposure of the image in the object coordinate system. These exterior orientation parameters can be derived in one of three ways:

1. Direct space resection,
2. Relative orientation followed by absolute orientation,
3. Simultaneous orientation by bundle adjustment.

1. **Direct space resection.** Based on the colinearity equation (equation 2), if one point in the object space and its corresponding point in the image space are known (called the control point), they contribute two observations. If three control points are available, the six EO parameters can be solved. In practice, four or more control points are normally used to calculate the EO parameters for improved accuracy, using a least-squares adjustment. This direct space resection method is normally used to determine the EO parameters of single images. For a stereopair or an image block, the exterior orientation parameters are derived by one of the other two methods because they require fewer control points.

2. **Relative orientation and absolute orientation.** Relative orientation is used to establish the relationship between two images without knowing about the object. Relative orientation is based on the coplanarity condition that two image points on a stereopair, the perspective centers of the two images, and the object point lie on the same plane (Figure 5). It assumes that the orientation and position of the left image are fixed, and the relative relationship between the left and right images can be determined by the five relative orientation parameters $b_y$, $b_z$, $\omega$, $\phi$, and $\kappa$ (assuming $b_x = 1$). $\omega$, $\phi$, and $\kappa$ are the rotation angles that are needed to make the right image coordinate system parallel to the left image coordinate system. $b_y$ and $b_z$ are the translations required to correctly position the right image’s perspective center.

   With the coplanarity condition, each pair of conjugate points identified on the stereo images generates one observation. At least five conjugate points are required for a unique solution of the five relative orientation parameters. Normally, six points are used for this process, including the two principal points and four other points distributed on either side of the principal point along the $y$ axis of the images. These are often referred to as von Gruber points. Once the five relative orientation parameters have been obtained, a 3-D model of the imaged scene is established. At this time, however, the 3-D model is not accurately scaled and its coordinates are based on an arbitrary coordinate system.

   Before the 3-D model derived from the relative orientation can be used for measurement, it must be scaled and oriented to the object coordinate system. This procedure is called absolute orientation. Absolute orientation is a 3-D conformal transformation that converts the model coordinates obtained during the relative orientation into correctly oriented mapping coordinates. There are seven parameters in this 3-D conformal transformation, including three rotation angles $R_\omega$, $R_\phi$, and $R_\kappa$, three translation components $T_X$, $T_Y$, and $T_Z$, and a scale factor $s$. At least three control points with known horizontal and vertical positions are necessary to achieve a result. The accuracy of absolute
orientation depends on the quality of the relative orientation and the accuracy of the control points.

The relative and absolute orientations can be performed on an individual stereopair or on complete image blocks covering a large area. In the latter case, the term aerial triangulation is often used to describe the procedure.

3. Simultaneous orientation by bundle adjustment. An alternative to the relative/absolute orientation methods is bundle adjustment. This method is based on the principle that, given a tie point in an image, it is possible to produce an observation based on the aforementioned colinearity equation. This observation represents an optical ray originating from the measured image point that goes through the center of the camera to the ground point. From the conjugate tie points identified on a stereo pair of images or multiple images in a block, a bundle of optical rays defined by the tie points can connect to the images themselves to form an image network and, thus, connect the images and object space. Ideally, the optical rays from the same conjugate tie points on different images should intersect at exactly the same ground point in the object space; however, in reality this may not be the case (Figure 6a) due to various uncertainties and errors. Therefore, bundle adjustment is used to adjust and solve the accurate image orientation parameters so that the corresponding rays intersect correctly (Figure 6b). This process is basically a least-squares adjustment based on the colinearity equation.

Given a few 3-D control points and tie points identified on the images, the image orientation parameters and the object coordinates of tie points can be computed simultaneously through bundle adjustment. This involves a
simultaneous resection and intersection process. The process may be applied either on a stereopair-by-stereopair basis, or on the entire image block. In the latter case, the term block adjustment is often used to describe the procedure. In the stereopair-by-stereopair case, there needs to be at least three 3-D control points in the overlap area of each stereopair. Observation of the image-space coordinates of each control point in the overlap area allows the image orientation parameters to be computed for both images. For an image block, a few 3-D control points and a number of conjugate tie points identified from the images are necessary, and all of the orientation parameters for the images are computed at once. The entire image block is processed in a single homogenous coordinate system.

In the bundle adjustment process, different weights can be assigned to different observations depending on their a priori precisions. The residuals of all parameters can also be computed and used to evaluate the performance of the bundle adjustment, which is an important advantage of this method.

**Figure 6** (a) The optical rays from the conjugate tie points do not intersect at the same ground point due to various uncertainties and errors, and (b) the improved results after bundle adjustment.

**Automatic solutions in digital photogrammetry**

Along with the development of digital photogrammetry since the 1980s, significant advances have been made in digital image processing; these have been applied to almost every aspect of photogrammetry, from image acquisition to image processing. These developments have also facilitated the automation or semi-automation of photogrammetric processing in modern digital photogrammetric workstations. The automation of digital photogrammetry strongly depends on the automation of image matching and the aforementioned IO and EO processes.

**Automatic image matching**

Image matching is the process of finding conjugate image correspondences (points or patterns) in the overlapping regions of two or more digital images. The process is based on either examining and matching the grey levels of small portions (image patches) of both images in a stereopair, or matching an image patch with an
PHOTOGRAMMETRY: 3-D FROM IMAGERY

image template. The matching may be done on a pixel-by-pixel basis (area-based matching) or by examining and matching the individual features of the image patches (feature-based matching). The most important applications of image matching are the automatic IO and EO of images, and the automatic creation of DTMs from multiple images. For the former, only a small number of reliable image correspondences with a favourable distribution are required, whereas for the latter, dense and reliable image correspondences are necessary.

Tremendous work has been done on image matching in the fields of photogrammetry and computer vision. The most straightforward method is normalized cross correlation (NCC), which directly assesses the degree of agreement between two local image windows by cross-correlation of their grey levels (Lhuillie and Quan 2002). An important development in image matching is the scale invariant feature transform (SIFT) method (Lowe 2004). SIFT detects points of interest based on local 3-D extrema in the scaled-space pyramid invariant over a wide set of transformations and matches the points according to descriptors defined by their gradient distributions in the detected local regions. SIFT provides automatic robust matching results even in the presence of scale changes and distortions. However, SIFT only provides sparse matching results. A representative method of area-based matching is semi-global matching (SGM) (Hirschmuller 2008), which combines the concepts of global and local stereo methods for pixel-wise matching. SGM approximates the cost of the global aggregation of matching from a number of 1-D cost paths, which provides accurate dense matching results. A representative method of feature-based matching is self-adaptive triangulation-constrained matching (SATM) (Wu, Zhang, and Zhu 2011), which uses triangulations to constrain the matching of feature points and edges. An important characteristic of SATM is that the triangulations are dynamically updated along with the matching process by inserting the newly matched points and edges into the triangulations. The most distinctive features are always successfully matched first, so that the densification of triangulations automatically self-adapts to the changes in image texture, and provides robust constraints to generate dense and reliable matching results. Figure 7 shows an example of SATM’s results for the automatic matching of Mars ground images acquired by NASA’s Mars Exploration Rover, Opportunity, and the generated DTM. SATM was also used to generate high-precision lunar DTMs for selecting the landing site for the Chinese Chang’E-3 lunar exploration mission (Wu, Hu, and Guo 2014).

Automatic image orientation

IO and EO are the fundamental orientation procedures in analog and analytical photogrammetry. In digital photogrammetry, for example in an interactive setting using digital photogrammetric workstations, the orientation tasks are essentially performed in the same way as on an analytical plotter. This section focuses on automatic orientation procedures, highlights their important differences, and discusses some solutions for digital photogrammetry.

Automatic IO is the starting point of the automation chain. For aerial images from scanning film, automatic measurement of the fiducial marks is the key step in automatic IO. Images from a metric camera contain at least four fiducial marks in the corners. Template matching can be used for this task. This process is driven by the structure description of the template that can be constructed for a fiducial mark, which automatically matches the predefined template with the
Figure 7  (a) A Mars satellite image showing the mapping area; (b) a stereopair of ground images of the area taken by the Mars Rover, with the matching results marked in red; (c) another stereopair showing large perspective changes and the matching results; and (d) the generated DTM from the matching results.

fiducial marks. The autonomous process requires a general and robust solution to accommodate the different types of fiducial marks. Subpixel localization of the fiducial marks should also be emphasized, as the pixel size is likely to be larger than the expected precision of the fiducial centers.

The development of automatic relative orientation has primarily focused on image matching to identify conjugate points. The aforementioned image matching methods can be used to find conjugate points automatically. The existing point-based algorithms developed to determine the orientation parameters in analytical photogrammetry can be used for this purpose. At least five conjugate points are necessary to determine the parameters. In practice, dozens or hundreds of matched conjugate points with a favorable distribution are normally used to identify the relative orientation parameters using least-squares adjustment. To determine the EO parameters of single images through automatic space resection and of stereopairs through automatic absolute orientation, it is critical to automatically establish the relationship between the image and object spaces. Schenk (1999) discussed the theory of using linear (e.g., road boundaries) or surface features (available from existing DTMs or from laser altimetry) as entities in the adjustment procedure, to compute the EO parameters automatically. More recently, Tommaselli and Berveglieri (2014) presented a more practical solution for automatic EO. They used a special camera attached to a GPS receiver to collect panoramic images in nadir view, while simultaneously collecting the 3-D coordinates of control points using the GPS. The panoramic images were then automatically matched with the aerial or satellite images to identify the locations of the control points in the latter to compute the EO parameters.

Future trends in photogrammetry

There is a tremendous worldwide demand for 3-D data, yet the methods for generating 3-D
data from imagery are still not fully automated and remain relatively slow. The development of a more automated photogrammetric process for deriving 3-D data from various types of images is a challenge. High redundancy (e.g., with every ground point visible in about 10 images) may play a major role in the design of fully automatic photogrammetric methods; for example, the UltraCam digital aerial mapping system has eight lenses, providing high imaging redundancy. Multi-image matching with high redundancy ensures reliable matching and enables high-quality 3-D reconstructions to be produced entirely automatically.

Laser scanning has been popular since the 1990s, as it enables much faster delivery of 3-D information than traditional photogrammetric techniques. It has even been argued that laser scanning may replace traditional photogrammetry in the future. However, in the past few years, reliable and automatic image matching has become one of the most active research areas in both the photogrammetric and the computer vision communities. The generation of accurate and dense 3-D information from multi-angled images from multiple sensors with a high level of automation is now a reality. A typical example is the “Building Rome in a Day” project by Microsoft Research. The development of oblique photogrammetry in recent years also offers encouraging solutions for 3-D city data generation and modeling. Photogrammetry and laser scanning have distinct characteristics that render them preferable for certain applications. The respective advantages and disadvantages of the two techniques suggest that their integration would provide better performance than can be achieved by either method alone. Wu, Hu, and Guo (2014) provided an example of integrating lunar imagery and laser altimeter data for consistent and precise lunar topographic mapping.

The timeliness of photogrammetry is currently attracting plenty of attention. In applications such as industrial measurement and real-time monitoring, the timeliness of photogrammetry is the key to success. The development of real-time (or quasi-real-time) photogrammetry has become an urgent task. Compared with traditional photogrammetric techniques, the research and development of real-time photogrammetry needs to overcome certain challenges, such as the full automation of photogrammetric image interpretation, the development of embedded algorithms, chips, and hardware systems for real-time image processing, and the development of new real-time photogrammetric sensors and multisensor collaborative technologies.

In addition, there is a trend of integrating methods developed for computer vision, such as shape-from-shading, shape-from-shadow, and structure-from-motion, into photogrammetry to allow better derivation of 3-D information from images. Regardless of the challenges ahead, photogrammetry remains the most complete, economical, portable, flexible, and widely used approach for deriving 3-D information. With the further developments mentioned here, photogrammetry has a bright future.

**SEE ALSO:** Geodesy; Geographic information system; Optical remote sensing

**References**


Phylogeography and landscape genetics integrate molecular genetics with geography, ecology, and evolutionary biology to provide an understanding of patterns in the spatial distributions of biological diversity. The historical and contemporary processes that underlie the distributions of organisms are inferred from signatures left in spatial patterns of genetic variation. Phylogeography and landscape genetics bring advanced molecular and geospatial tools to bear on issues that are vital to the field of biogeography, including how and where biodiversity arises, how species respond to changing climates, and how and where conservation efforts should be focused. The two disciplines contribute to and integrate with conservation biology, evolutionary biology, paleoecology, geomorphology, and emerging climate and sustainability sciences.

Substantial conceptual and methodological overlap exists between phylogeography and landscape genetics, but the two fields differ in their spatial and temporal scales of study. Phylogeographers investigate impacts of historical geographical conditions on genetic variation within or between species over broad spatial extents, while landscape geneticists focus on contemporary genetic variation within species at the scale of individual landscapes. The spatial and temporal scales investigated using the two approaches can be viewed as a continuum, but individual studies are necessarily limited to discrete time periods and spatial extents. Thus, the appropriate molecular and analytical tools to test a given biogeographical hypothesis may differ with the specific scale(s) under study. Ultimately, however, phylogeography and landscape genetics are highly complementary when used in combination to tell a more complete biogeographical story than either one alone could have done.

Phylogeography arose in 1987 as a synthesis between advancing molecular genetic technology and traditional biogeography. Phylogeographers use both phylogenetic and population genetic methods to investigate evolutionary responses such as speciation, adaptation, and demographic changes (e.g., population expansion or contraction) to broad geographical features (e.g., glaciation, changing climates). The timescales of study range from thousands to millions of years and spatial scales tend to range from continental to global. Phylogeographers sample genetic variation in one or a few closely related species. For example, genetic relationships between modern and ancient (fossil) European red deer indicate that this species persisted in northern refugia during the last glacial period, but that most modern red deer in Europe descend from ancestors that survived in unglaciated areas on the Mediterranean coast (Meiri et al. 2013). Comparative phylogeographers investigate biogeographical histories across distantly related species. Typically, comparative phylogeographic studies compare the responses of different organisms to the same geographical conditions. Rissler and Smith (2010) used genetic information from 37 amphibian species to show a high density of suture zones (areas of overlap between distinct genetic lineages) concentrated at the southern...
end of the Appalachian Mountains, indicating that this region is a hotspot of North American amphibian diversity and an area of special conservation concern.

The field of landscape genetics, which combines population genetics and landscape ecology, was introduced in 2003. Landscape geneticists focus on ecological time periods (the present to hundreds of years) and on finer spatial scales (hundreds of square kilometers). Population genetic methods are used to investigate how ecological processes such as invasive species dynamics and biological responses to human-induced land-cover change are impacted by geographical variation in the matrix (the intervening land cover between habitats) and arrangement of habitats within the matrix. For example, a study in central Germany indicated that both the river Rhine and a nearby highway restrict contemporary genetic connectivity between populations of endangered wildcats (Hartmann et al. 2013).

The evolutionary and ecological processes underlying patterns of genetic variation depend intrinsically on spatial context. As geospatial environmental data have increased in availability, the incorporation of geographical features in phylogeography has shifted from qualitative (e.g., visualization of genetic variation on maps) to quantitative. The burgeoning discipline of landscape genetics has included quantitative geographical analyses since its inception. For example, Mantel tests are commonly used to detect correlations between pairwise genetic and geographical distances. Other geographically quantitative methods in current use include principal components analysis and canonical correlation analysis of genetic covariation with geographical variables.

The use of geographical information science (GIS) to assess spatial patterns of genetic variation is an area of active synthesis in both phylogeography and landscape genetics. One relevant GIS-based method is species distribution modeling (SDM), in which climate maps are used to model historical and future species distributions based on assumptions about ecological niche parameters. For instance, SDMs of European birds, mammals, and plants predicted that diversity will shift broadly northward over the next 70 years under scenarios of climate change (Thuiller et al. 2011). Phylogeographers and landscape geneticists also use GIS methods such as wombling and least-cost path (LCP) analysis. Wombling combines interpolated raster surfaces for genetic and environmental variables to identify spatial locations of boundaries in genetic variation. LCP analysis uses resistance surfaces (hypothetical or, ideally, empirically based spatial quantification of differences in resistance of land cover types to organismal movement) to model likely paths of migration or dispersal. In a comparison of LCP-modeled dispersal routes and estimates of gene flow between bighorn sheep, anthropogenic barriers (e.g., urban areas) explained inconsistencies between observed and predicted dispersal patterns (Epps et al. 2007).

Phylogeography bridges the fields of phylogenetics and population genetics, while landscape genetic analyses are primarily population genetic. Phylogenetic analyses take the form of trees or networks representing relationships within or between species. Population genetic analyses in both fields frequently rely on coalescent theory, which models the path to the most recent common ancestor backwards in time from modern sampling of genetic variation. The coalescent helps to resolve signatures of recent and historical events like range expansion and loss, migration, and population bottlenecks caused by colonization events.

Phylogeographers and landscape geneticists must carefully consider which molecular genetic tools are most appropriate for a given
research question. Phylogeographers use genetic variation, in part, as a powerful complement to the highly incomplete fossil record of spatial distributions of biological diversity. Likewise for landscape geneticists, genetic variation is a surrogate for direct dispersal data (e.g., global positioning system (GPS) tracking or mark recapture data) that may be difficult to obtain. Genetic variation in organisms arises via mutations in deoxyribonucleic acid (DNA) sequences. The variants caused by mutation then change in frequency as a result of processes such as random genetic drift, demographic changes (e.g., in population size), gene flow, and natural selection. Researchers in the two fields focus primarily on sampling neutral genetic variation (i.e., variation not subject to natural selection). Because mutation rates differ for different portions of an organism’s genome, various types of genetic markers capture snapshots of biogeographical processes on different timescales.

The primary genetic markers used in phylogeographical studies have traditionally been uniparentally (often maternally) inherited organelar DNA sequences, including mitochondrial DNA (mtDNA) in animals and chloroplast DNA (cpDNA) in plants. Useful mtDNA markers contain variable regions (informative over thousands to tens of thousands of years), along with regions that are far less variable (informative over millions of years). Because inheritance is exclusively from one parent, analyses using cpDNA or mtDNA sequences may tell a sex-biased story. For instance, comparison of genetic variation in humans between maternally inherited mtDNA and paternally inherited Y chromosome sequences showed evidence of historically higher migration rates in females than in males (Seielstad, Minch, and Cavalli-Sforza 1998). The genetic markers typically used in landscape genetic studies reflect a focus on recent and ongoing processes. The most commonly used markers for landscape genetic studies are microsatellites. These markers change rapidly enough to be informative over tens to hundreds of years. Microsatellites are biparentally inherited nuclear markers that consist of short, repeated sequence fragments of variable length.

Increasingly, both landscape geneticists and phylogeographers combine data from as many molecular markers as possible to provide robust comparisons of biogeographic hypotheses because individual markers tell only a small part of a complex story. A substantial challenge in both fields is the lack of readily available genetic markers for all but a few “model” (extensively studied) organisms. Markers such as amplified fragment length polymorphisms (AFLPs) provide an inexpensive way to sample genetic variation without pre-existing sequence data, but may be less informative. DNA sequences for entire genomes are becoming progressively more attainable for a wide range of organisms. Genomic datasets simplify identification of large numbers of markers, including microsatellites and single nucleotide polymorphisms (SNPs). The use of large SNP datasets allows for easier detection of adaptive genetic variation (i.e., subject to natural selection). Adaptive variation is increasingly being incorporated in phylogeographic and landscape genetic studies.

SEE ALSO: Biodiversity; Biogeography: history; Ecosystem; Geography of evolution

References


Further reading

Physical geography

Antony R. Orme
University of California, Los Angeles, USA

Definition and terminology

In literal terms, physical geography seeks to describe Earth’s natural surface features (Greek geo, Earth; graphos, writing). In practice, it has come to be identified with explaining the nature and distribution of Earth’s landforms and formative processes (geomorphology), weather and climate (climatology, meteorology), water in various states (hydrology, glaciology, oceanography), soil and sediment (pedology, sedimentology), and plants and animals (biogeography, ecology) (Figure 1). Some scholars focus on the more physical themes and treat biogeography (plants, animals, organic soils) separately. In reality, however, no study of Earth’s physical or biological realms can now ignore their interdependence or their interaction with human activity. Physical geography is thus better defined as the geography of the natural world, explained both for its intrinsic interest across space and time, and in its human context.

Scientific progress should always be viewed against its cultural milieu because social mores, religious beliefs, economic wellbeing, political arrangements, and technological assets may promote or restrict new ideas, investigation, and debate. This entry focuses sequentially on the changing nature of physical geography in a human context over time and assesses how present practices predict the field’s future. Not all practitioners would have been called geographers at the time but their work has become the essence of physical geography today (Orme 2013a–f).

Physical geography of the early human world

From earliest human times, people bent on survival learned from experience about weather, potable water, edible plants, and natural hazards. Archaeology reveals much about prehistoric life and livelihood, cave dwellings affording shade and shelter, campsites above flood waters, trackways along rivers and through wetlands, and the use of fire. Early peoples understood the impact of changing seasons on water resources, plant life, and animal and fish migrations. They also fashioned useful tools, weapons, fishing tackle, containers, and boats to serve hunting and gathering needs. They acquired folklore, incidences of which have emerged from archaeological diligence, providing insights to contemporary environments, and gathering momentum among the advancing Paleolithic and Mesolithic cultures of late Pleistocene and early Holocene time.

These early peoples lacked literacy in a modern sense but acquired measures of numeracy and graphicity. Petroglyphs carved on rock reveal an environmental awareness. More dramatic are the cave paintings that survive at Lascaux in France and Altamira in Spain, which with associated materials reveal much about animals hunted and plants eaten. They were also fascinated by the Sun, Moon, and stars beyond. An awareness of the march of the seasons led communities to build structures, such as Stonehenge, which showed a sense of astronomy.
Figure 1  Earth’s natural systems: the basis for physical geography. Modified from Orme 2002. Courtesy A.R. Orme.
Physical geography among early literate peoples

When folklore was written down, the documentary record began, albeit beguiled by ambiguity, superstition, and mythology. But literacy likely implied some cultural stability, as sedentary peoples fostered by Neolithic farming and animal domestication allowed those not needed for food production time to reflect on and record the world around them. Among these were priests who encouraged literacy in ways appropriate to their beliefs and provided environmental indicators for later cultures (Orme 2013a).

Early scribings contain incidental observations on environments and astronomy, as depicted on Minoan frescoes in Bronze Age Crete, cuneiform scripts on Sumerian and Babylonian clay tablets, and hieroglyphs and pictographs on Egyptian papyrus scrolls. These show an awareness of Sun and sky, rivers and coasts, storms and high seas, plant and animal life, and measurement units. Around 3000 BCE (Before the Common Era), the Sumerians had creative gods for heaven, Earth, air, and water, and lesser gods for mountains and rivers. By 2500 BCE, astronomers in the Nile valley had devised a solar calendar based on a year of 365 days.

For agriculture and settlement to prosper amid the seasonal floods and droughts of the Nile, Tigris, Euphrates, and Indus valleys, these early cultures intuitively grasped the links between climate, hydrology, and food production. Irrigation ditches were being built in Mesopotamia before the date later proposed for Earth’s creation by Judeo-Christian biblical scholars. The success of Egyptian irrigation works, Phoenician aqueducts, Persian qanats, Indian canals, and Chinese flood controls shows an empirical appreciation of water, or conversely drought. Seafarers understood the links between storms and high sea, the perils of shoaling water and lee shores, the tranquility of safe anchorages, and the nature of tides.

Physical geography in the classical Mediterranean Basin

Greek scholars laid the foundations of physical geography in the Mediterranean Basin about 3000 years ago, whence their ideas diffused widely. Although imperfect, their science yielded much that was fundamental to understanding the physical world. Then, as more rational enquiries began probing their habitable world (oikoumene), an ordered semblance of physical geography emerged from philosopher-scientists. Their observations advanced knowledge because they sought explanation, which involved speculation and mathematics. Although most of this work has been lost, partial reconstructions by later scholars reveal their geographies.

Thales of Miletus (ca. 624–546 BCE) was among the first philosophers to explain nature without invoking mythology. Anaximander (ca. 610–547 BCE) pondered water’s origins. Pythagoras (ca. 582–500 BCE) saw a spherical Earth rotating on its own axis, and Aristarchus (ca. 310–250 BCE) favored a heliocentric view of the solar system. Herodotus (484–425 BCE) saw seasonal floods deliver sediment to the coast and equated Nile distributaries with the Greek letter delta (Δ). Aristotle (384–322 BCE) measured Earth’s sphericity, speculated on weather and water, and implied cyclic uplift and subsidence. He followed Hippocrates (ca. 460–370 BCE) in linking climate to human health and behavior! Theophrastus (371–287 BCE) defined the hydrologic cycle and its effect on rivers, and described winds, plants, and animals. Archimedes (288–212 BCE) defined principles for mechanics and hydrostatics. In Geographika, Eratosthenese (ca. 276–194 BCE) discussed Earth’s
PHYSICAL GEOGRAPHY

circumference and its frigid, temperate, and torrid climate zones. Polybius (ca. 200–118 BCE) opined that streams eroded valleys, and navigators, from Pytheas (ca. 325 BCE) and Posidonius (135–51 BCE) in the Atlantic to Seleucus (ca. 150 BCE) in the Indian Ocean, pondered the Earth-Moon-Sun-forcing of tides. Crude maps of the known world improved.

Ancient Rome’s original contributions to science were more practical, although Greek traditions of geography continued. Roman engineers who built aqueducts, highways, and harbors understood many surface processes, exemplified by a surviving text on aqueducts by Sextus Julius Frontinus (40–103 CE (Common Era, equivalent to AD)). Thus, in his 17-volume Geographica, Strabo (63 BCE–24 CE), a Greek scholar flourishing under Roman rule, attributed landforms and ocean changes to earthquakes and volcanic activity, discussed sediment transport, related delta size to drainage-basin erodibility, and bemoaned the effects of forest clearance on denudation. But uncertainties persisted. Astronomer and mapmaker Ptolemy (ca. 100–170 CE) of Alexandria reverted to a geocentric planetary system, with a stationary Earth at its center. Around 540 CE, the Byzantine traveler Cosmas Indicopleustes described a flat Earth on which sailors ventured beyond the horizon at their peril, and explained night and day as a response to the Sun circling a heaven-piercing mountain. While such thoughts may have confused ordinary folk, the torch of scientific progress was relayed to more distant cultures.

Physical geography in early Islamic and Asian cultures

With the decline of classical Mediterranean civilizations, the expansion of Islam from its seventh-century roots absorbed Greek and Roman legacies, while environmental cognition, physics, and mathematics were advanced by Arab, Persian, Indian, and Chinese scholars. Geography flourished under the Abbasid Dynasty, after the caliphate moved in 766 CE from Damascus to Bagdad where Greek science survived in Persian culture. Abu Ma’shar’s ninth-century work defined the astronomical forcing of tides. Astronomers in tenth-century Basra conceived a cycle of erosion and sedimentation that reversed the roles of land and sea over a Great Year of 36,000 years. In Quadrans Habilitatis, Abu Rihan Birunensis (973–1038 CE) described south Asia and east Africa. Al-Idrisi (ca. 1099–1165) and Ibn Batuta (ca. 1304–1369) described features across Eurasia. Ibn Khaldun (1332–1406) identified seven climatic zones, which he dubiously correlated with human traits and behavior. Meanwhile, Persian scholar Avicenna (Ibn Sina, 980–1037) pondered time, rocks, mountains, and denudation in central Asia:

Either [mountains] are the effects of upheavals of the crust of the Earth, such as might occur during a violent earthquake, or they are the effect of water, which, cutting itself a new route, has denuded the valleys, the strata being of different kinds, some soft, some hard … It would require a long period of time for all such changes to be accomplished, during which the mountains themselves might be somewhat diminished in size. (Translation of Avicenna, 1027)

Chinese scholars addressed similar issues. Shen Kuo (1031–1095), noting coastal progradation and sediment-laden rivers, suggested that land is reshaped over time by uplift, erosion, and deposition. He inferred climate change from fossil bamboo found far beyond its present range. The seven voyages of Zheng He (1371–1435) from China to the Indian Ocean (1405–1433) recorded water depths, winds, and currents, while Chinese usage of printing and the magnetic compass were introduced to Europe.
Physical geography in medieval and Renaissance Europe

As the western Roman empire collapsed and its eastern counterpart in Byzantium atrophied before barbarian onslaughts, much of Europe was plunged into intellectual gloom. Even so, amid the uncertainties of these times, glimmers of light persisted as Philoponus (c. 490–570) of Alexandria challenged religious concepts of Earth’s origin. When barbarian leaders became better educated, they encouraged observation, as noted by the Venerable Bede (672–735) in Britain. But many people still thought of a geocentric universe and flat Earth.

Classical learning saw some renaissance under Charlemagne, king of the Franks and Germans from 768 to 814, but this later declined amid dynastic strife. During the Middle Ages, with social contacts mostly local, literacy rare, and religion persuasive, science struggled for debate. Space and time were poorly understood. Maps were few and, following biblical texts, mappae mundi centered on Jerusalem, as shown on the disk map from 1300 now in Hereford Cathedral.

Natural hazards were noted but rarely analyzed: Gerald of Wales (1147–1220) described erosion and Heligolanders measured their island coastline shrinking from 192 km in 800 to 72 km in 1300 (to 5 km by 1900!). Viking documents suggested reduced sea ice around Iceland and expanded cereal cultivation in northern Europe between 900 and 1200, when the climate was warmer than today. Later, as the Little Ice Age approached, glacier advances were noted but scientific studies were limited.

Even the emerging academies of medieval Europe were constrained by faith and rarely excited by nature. Nevertheless, some Europeans were attracted to the geographies of Asia reported by travelers, notably William of Rubruck (Rubruquis, 1220–1293) and Marco Polo (1254–1324) and his kin who traded with China between 1260 and 1295. Stimulated by Arab and Persian studies, enquiring minds began questioning aspects of nature, although the Franciscan friar Roger Bacon (1214–1294) was admonished for advocating experimentation over faith.

Between 1300 and 1600, the bonds of feudalism, religious dogma, and folk tradition on Europeans were loosened by the Renaissance, notably after the socioeconomic catastrophe of the Black Death in the 1340s. Apart from impacts on religion, politics, and art, the Renaissance revived classical mathematics and astronomy, and favored explorations overseas and scientific experiments at home. In France, studies of gravity, acceleration, and momentum by Jean Buridan (ca. 1300–1358) anticipated Newton by more than three centuries. Physical geography revived, based in part on translations of Arabic transcripts of earlier Greek scholarship.

The European age of exploration and discovery

From the late fifteenth century, European maritime explorations yielded much new information for late Renaissance scholars (Table 1). Bartolomeu Dias (1451–1500) and Vasco da Gama (1469–1524) sailed around Africa into the Indian Ocean, aided by local navigators familiar with monsoons. John Cabot (1450–1499), Christopher Columbus (1451–1506), and Amerigo Vespucci (1454–1512) reached Atlantic America, and Ferdinand Magellan (1480–1521) and Francis Drake (1540–1596) passed Cape Horn into the Pacific. Coastal charts or “wagoners” were valuable assets, notably the Spieghel der Zeevaerdt (Mariner’s Mirror, 1584) of Lucas Waghenaer (1533–1606). Military forays by Hernán Cortés (1485–1547) in Mexico
PHYSICAL GEOGRAPHY

(1519–1520) and Francisco Pizarro (1476–1541) in Peru (1531–1533) opened lands for Spanish colonization, even as others were entering the Americas. Not only did these discoveries bring Europeans and their trappings to new lands, but their ships returned with strange plants and animals, accelerating the relentless transfers between continents of alien species that were to transform landscapes far and wide.

Meanwhile, at home, European scholars reaffirmed Greek calculations of Earth’s shape, and new maps began rolling off printing presses from the late fifteenth century. Scale globes became popular and Ptolemy’s mostly forgotten maps reappeared, but, as Greek mapmakers had found, projecting Earth’s curvature onto a flat surface might work locally but was no easy task over larger areas. New facts reached a wider audience through Gerhard Kremer (Mercator, 1512–1594), whose map projection became the basis for sea charts, and Abraham Ortelius (1527–1598) in an atlas of 70 charts of the then known world, Theatrum Orbis Terrarum, in 1570.

The Scientific Revolution in seventeenth-century Europe

By 1600, the intellectual excitement of the Renaissance was waning before the wars of religion and the realignments of emerging nation states. Many cultures were changing—from country folk dependent on nearby field and forest to townspeople for whom outward-looking ports and foreign trade, and thus exposure to new ideas, were becoming influential. Traditional faiths also came into conflict with new ideas. The Scientific Revolution and later Enlightenment responded to the spirit of curiosity and openness spawned by these events (Table 1).

The Scientific Revolution in Europe was favored by Renaissance thought, growing awareness of Islamic and Asian science, overseas discoveries, and printed works that disseminated ideas new and old. Between the birth of Leonardo da Vinci (1452–1519) and the death of Galileo Galilei (1564–1642), major changes occurred in science as metaphysical paradox gave way to measurement and experiment. In 1530, Nicolaus Copernicus (1473–1543) reaffirmed a heliocentric system for the solar system involving Earth’s rotation and precession that, though suppressed by the Roman Catholic Church, gained support from Johannes Kepler (1571–1630) and Galileo’s 1632 work on celestial mechanics. However, as Galileo found at his trial for heresy in 1633, inquisitive scientists could not always rely for support on those more closely linked to the religious and political establishments of their age (Orme 2013a).

When the revolution peaked later in the century, mathematics, astronomy, physics, chemistry, and biology emerged in idioms which, if not modern, are recognizable today. Concepts relevant to physical geography began to appear—momentum and energy conservation from René Descartes (1596–1650) in 1644; air pressure linked to volume by Robert Boyle (1627–1691) in 1662; centrifugal force from Christiaan Huygens (1629–1695) in 1673; kinetic energy from Gottfried Leibniz (1646–1716) in 1686; and laws of motion and calculus from Leibniz and, in 1687, Isaac Newton (1642–1727). Scientific progress also saw better instruments: the optical microscope (ca. 1570), the thermometer (1612) from Santorio Santorre (1561–1636), and the barometer (1643) from Evangelista Torricelli (1608–1647). Even so, the hydrologic cycle was long misconstrued: a popular view offered in 1664 by Athanasius Kircher (1602–1680) showed ocean water suctioned upslope by lunar forces in
Table 1  Time frame for physical geography, 1500–2000. Courtesy A.R. Orme.

<table>
<thead>
<tr>
<th>Date</th>
<th>Cultural milieu</th>
<th>Science in general</th>
<th>Biogeography biology, soils</th>
<th>Climatology glaciology</th>
<th>Geomorphology geology</th>
<th>Hydrology oceanography</th>
</tr>
</thead>
<tbody>
<tr>
<td>1600</td>
<td>Reformation</td>
<td>Copernicus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1700</td>
<td>Scientific revolution</td>
<td>Galileo Descartes Pascal Boyle, Huygens Leibniz Newton</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1800</td>
<td>Agrarian revolution</td>
<td>Bernoulli</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1900</td>
<td>Enlightenment</td>
<td>Lavoisier</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td>Industrial revolution</td>
<td>Malthus Carnot Clausius C. Darwin Wallace Thomson [Kelvin] Le Châtelier Marie Curie Aerial photos</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Depression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>World War I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>World War II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Electronic revolution</td>
<td>Radar Nuclear weapons Plate tectonics Moon landing Environmental laws, conservation Planetary landings Climate concerns</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Globalism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
subterranean conduits to spill from upland caves into rivers that returned water to the oceans (Figure 2).

Mining heralded interest in Earth’s physical systems while the need to supply water to growing towns advanced hydrology and fluid dynamics. Building on the work of Bernard Palissy (1510–89) and Simon Stevin (1548–1620), Benedetto Castelli (1577–1644) and Torricelli explained streamflows in terms of water and channel properties. A more realistic hydrological cycle appeared when Pierre Perrault (1611–1680), Edmé Mariotte (1620–1684), Domenico Guglielmini (1655–1710), and Edmund Halley (1656–1742) began integrating precipitation, evaporation, and infiltration into models linking atmospheric and soil processes with streamflows (Figure 3).

Climatology began with weather data collected from rain gauges, barometers, and thermometers, notably in London and Paris in the 1660s. Inferences regarding average temperatures and precipitation would later come from these data, adjusted for observer errors, inconsistent instruments, and missing records. William Dampier (1651–1715) described typhoons and George Hadley (1685–1768) explained trade winds in terms of Earth’s rotation.

Biogeography slowly emerged from a curiosity for newly discovered plants and animals. This increased interest in botany and zoology, which in the sixteenth century began diverging from, but never severing, their practical roots.

**Seventeenth-century perspectives on physical geography (left to right)**

**Figure 2** Kircher’s hydrologic cycle (subterranean conduits in black; rivers in white). Kircher 1665. Reproduced from *Mundus subterraneus*.

**Figure 3** Perrault’s humorous hydrological cycle, drawn to counter Kircher’s model. Perrault 1674. Reproduced from *De l’origine des fontaines*.

**Figure 4** Burnet’s *Theory of the Earth* (1689) showing the imagined past, present, and future. Clockwise from top right: initial chaotic liquid, 6000 years ago; water-filled globe; Noah’s Flood with Ark; flood abates, crustal wreckage forms Earth’s relief; conflagration precedes second coming; Earth restored to water-filled globe; and becomes a star. Thomas Burnet (1681, 1689) *Telluris Theoria Sacra*, W. Kittilby, London [1816, *The Sacred Theory of the Earth*. T. Kinnersley, London].
PHYSICAL GEOGRAPHY


Volumes on Earth’s emerging (mostly physical) geography began to appear. Philip Cluverius (1580–1622) wrote an *Introduction to Universal Geography* published posthumously in 1624, and Nathanael Carpenter (1589–1628) a *Geography Delineated Forth in Two Booke* in 1625. *Geographia Generalis* by Bernard Varenius (1622–1650) came in 1650 and was translated into English as *The Compleat System of General Geography*. Impressed by the works of Copernicus, Kepler, and Galileo, Varenius opined: “Geography is that part of mixed mathematics which explains the state of the Earth and of its parts depending on its quality, namely its figure, place and magnitude.”

Scientific debate was also stimulated by universities expanding beyond their medieval foundations and by new academies, notably the Royal Society of London (1662) and the Académie des Sciences de Paris (1666). In its *Directions for Sea-Men, bound for far Voyages* (1665–1666), the Royal Society urged mariners to record ocean processes and seabed conditions, and to improve navigational aids, sounding gear, and water sampling.

Explanation in physical geography was long confused by questions of Earth time: short time constrains what is possible; longer time offers more flexibility. Certainly, early Greeks had been puzzled by fossils in rocks, Avicenna in 1027 had invoked long time for mountains to be raised and eroded, and Leonardo da Vinci had Po valley sediment accumulating over 200,000 years. However, James Ussher (1581–1656) and others reverted to biblical sources to claim that the six days of creation (Psalm 90, 1 day = 1000 years) implied that Earth was only 6000 years old and that Noah’s Flood began in 2349 BCE. This required Earth’s surface to be shaped and clothed in very short time by divine providence and then changed by frequent catastrophes. A prolonged tussle between persons of faith and observers of the real world ensued, pitting divine revelation against scientific deduction. Carpenter (1625) opined that “Mountaines, Valleyes, and Plaines were created in the Earth from the beginning, and few made by the violence of the Deluge.”

In 1681, Thomas Burnet (1635–1715) ascribed Noah’s Flood to the bursting of a water-filled globe that “at one stroke dissolved the frame of the old world and made us a new one out of its ruins which we now inhabit since the Deluge” (Figure 4). Catastrophism gained support from earthquakes near Port Royal, Jamaica, in 1692, and Lisbon, Portugal, in 1755, and eruptions of Vesuvius. Hooke saw gorges as earthquake rifts. Such time constraints confounded scholarship and were not countered objectively until the advent of radiometric dating in the twentieth century.

The Enlightenment of eighteenth-century Europe

The Enlightenment was a cultural process in which reasoning came from a discerning interest in nature and reality rather than from uncritical acceptance of received texts such as the bible. Studies of the natural world led to conflict with religion and to debates between students challenged to think anew and theologians intent on preserving tradition. The Enlightenment
was led by philosophes such as Montesquieu, Diderot, Voltaire, and Rousseau in France, Kant in Germany, Hume in England, and Franklin and Jefferson in America; some even conducted experiments and offered mathematical solutions for natural relationships. A broader public became engaged in coffee houses and academies, exposure to artists and poets imbued with burgeoning Romanticism, and Grand Tours to civilization’s assumed roots in the Mediterranean and Middle East. The Enlightenment closed with the French Revolution and Napoleonic Wars, from which emerged a very different scientific world, more pragmatic, more aware (Orme 2013a).

Many scientists visited little-known lands in Pacific and polar waters, with navigation aided by improved sextants, marine chronometers, and sea charts. The invention by John Harrison (1693–1776) of a timepiece sufficiently robust for ship travel was a welcome replacement for lunar tables in defining longitude. Notable voyages included those by Vitus Bering (1680–1741), James Cook (1728–1779), George Vancouver (1757–1798), Fabian Bellinghausen (1778–1852), James Weddell (1787–1834), Otto von Kotzebue (1787–1846), and Dumont d’Urville (1790–1842).

English and Dutch trading companies charted Indian Ocean coasts, and Hudson’s Bay Company Canadian Arctic shores. Such charts attracted mapmakers, such as Bourguignon d’Anville (1697–1782) of Paris and the Blunt family of New York. Charts were also made by naval hydrographic agencies such as those founded in France (1720), Spain (1753), and Britain (1794), and the civilian Coast Survey in the United States (1807). While mapmakers still struggled to depict onshore relief, chartmakers were already using contours (isobaths) to link soundings of equal depth to a common datum, notably Luigi Ferdinando di Marsigli (1658–1730) for the Golfe du Lion (1725) and Philippe Buache (1700–1773) for the English Channel (1737). Marsigli also wrote an ambitious *Histoire Physique de la Mer* (1725) (Orme 2013f).

Onshore surveys, hitherto mostly conducted by landowners and military engineers, became systematized under national agencies, such as the Trigonometrical Survey of Great Britain (now the Ordnance Survey) in 1791. Physical geography was further stimulated by geologic mapping and by topographic maps on which hachures were superseded by contours in 1728.

Wisdom spawned by the Scientific Revolution was refined during the Enlightenment. Attempts to explain Earth’s surface features saw biblical Catastrophism and its kin Diluvialism (based on Noah’s Flood) challenged, first by Neptunism and then by Uniformitarianism. Neptunism, proposed by mining geologist Abraham Gottlob Werner (1749–1817), suggested that Earth’s rocks had been precipitated from a universal ocean that had shaped surface features as it receded. In an attempt at reconciliation, naturalist Georges Cuvier (1769–1832) separated the six days of biblical creation with the six catastrophes (révolutions) he found in the fossil record (Figure 5). In various guises, Catastrophism persisted well into the nineteenth century, supported by Benjamin Silliman (1774–1864) in whose journal, the *American Journal of Science*, J.W. Wilson (1821) wrote “is it not the best theory of the Earth, that the Creator, in the beginning, at least at the general deluge, formed it with all its present grand characteristic features?”

In contrast, Uniformitarianism held that the present is the key to the past, that Earth’s surface is mostly shaped by slow but observable processes over a very long time. The concept evolved from Mikhail Lomonosov (1711–1765) in Russia, Giovanni Targioni-Tozzetti (1712–1783) in
Italy, and Jean-Etienne Guettard (1715–1786) and Nicolas Desmarest (1725–1815) in France, who each asserted the then novel idea that valleys are shaped gradually by streams, rather than suddenly by biblical floods and earthquakes.

James Hutton (1726–1797; Figure 6), Scottish physician and farmer, is often hailed as the founder of the Earth sciences in a relatively modern idiom. His writings between 1785 and 1795 include some incisive statements:

But if the succession of worlds is established in the system of nature, it is in vain to look for anything higher in the origin of the Earth. The result, therefore, of our present enquiry, is that we find no vestige of a beginning, – no prospect of an end. (Hutton, *Theory of the Earth, or an investigation of the laws observable in the composition, dissolution, and restoration of land upon the globe* 1788, 304)

The natural operations of this globe, by which the size and shape of our land are changed, are so slow as to be altogether imperceptible to men who are employed in pursuing the various occupations of life and literature. (Hutton, *Theory of the Earth, with Proofs and Illustrations* 1795, II, 563)

Hutton’s ideas were supported in *Illustrations of the Huttonian Theory of the Earth* (1802) by mathematician John Playfair (1747–1819; Figure 7), and in *Hydrogéologie* (1802) by biologist Jean-Baptiste Lamarck (1744–1829). Playfair understood the temporal implications of Hutton’s work in writing “time performs the office of integrating the infinitesimal parts of which this progression is made up.” He defined denudation thus:

“water … in every state from transparent vapour to solid ice, from the smallest rill to the greatest river … attacks whatever has emerged from above the level of the sea and labours incessantly to restore it to the deep. The parts loosened and disengaged by the chemical agents, are carried down by the rains, and, in their descent, rub and grind the superficies of other bodies … [T]he
consequence … is a system of universal decay and degradation, which may be traced over the whole surface of the land, from the mountain top to the sea shore.” (Playfair 1802, 99–100)

In 1830–1833, Charles Lyell (1797–1875; Figure 8) published The Principles of Geology – Being an Attempt to Explain the Former Changes of the Earth’s Surface, by Reference to Causes now in Operation. This extended title had broad implications for questions of time. Hutton’s “succession of worlds” with neither a beginning nor an end, and the slowness of Lyell’s “causes now in operation” denied Catastrophism and encouraged concepts of cyclicity and evolution.

Meanwhile, practical studies of Earth’s surface processes were being made by scientists and engineers across Europe, notably the influential Corps des Ponts et Chaussées (1715) in France. These fostered understanding of processes in ways fundamental to physical geography. Major works included explanations of air and water currents by Daniel Bernoulli (1700–1782), open-channel flow by Antoine de Chézy (1718–1798; Figure 9) and Pierre du Buat (1734–1809), slope stability by Charles Augustin Coulomb (1736–1806; Figure 10), and waves and tidal dynamics by Joseph-Louis Lagrange (1736–1813) and Pierre-Simon Laplace (1749–1827; Figure 11).

Studies of Earth’s atmosphere now acquired a sounder scientific framework. Temperature scales were rationalized in 1724 by Daniel Fahrenheit (1686–1736) and in 1742 by Anders Celsius (1701–1744). Weather data were being collected more consistently in many European locations and from 1730 in Philadelphia. In England, John Dalton (1766–1803) recorded pressure, wind, temperature, and humidity for

Eighteenth-century understanding of physical processes (left to right)
Figure 9  Antoine de Chézy (1718–1798), hydraulic engineer. Reproduced from Ecole des Ponts et Chaussées, Paris.
Figure 10  Charles Augustin Coulomb (1736–1806) clarified slope stability and slope failure. Reproduced from Ecole des Ponts et Chaussées, Paris.
Figure 11  Pierre Simon de Laplace (1749–1827) explained dynamic tides and waves. © The Royal Society. Reproduced with permission.
PHYSICAL GEOGRAPHY

forecasting purposes. The Mannheim Academy of Meteorology was founded in 1780.

After the Little Ice Age (1500–1850) peaked around 1700, receding glaciers led naturalists to question glacier motion. Johann Scheuchzer (1672–1733) proposed downslope displacement by meltwater freezing in crevasses; Horace-Bénédict de Saussure (1740–1799), viewing striated rock emerge from beneath retreating glaciers, saw basal gravity sliding as the driving mechanism; and Svein Pálsson (1762–1840) in Iceland compared glacier motion to a resin-like viscous fluid. In 1830, Franz Hugi (1791–1855) measured the speed of the Unteraar Glacier.

Biogeography also traces roots to the eighteenth century, notably to the taxonomies of Carl Linnaeus (1707–1778; Figure 12) in his Systema Naturae (1735 on), and to Georges-Louis Leclerc, Comte de Buffon (1707–1788), in his multivolume Histoire naturelle, mostly on animals (1749–1788). Both Linnaeus and Buffon struggled with prevailing Catastrophism. In Des époques de la nature (1778), Buffon suggested from foundry experiments that it would take up to 75,000 years for Earth to cool from a molten state and 3 million years to assemble its sedimentary cover. Links between climate and vegetation were reflected in Carl Wildenow’s (1765–1812) montane zonation in Grundriss der Kräuterkunde (1792). In Zoönomia (1794–1796), Erasmus Darwin (1731–1802) addressed environmental influences on animals, and anticipated evolution, later developed by his grandson Charles. Interest in landscaping, herbaria, and menageries (zoos) generated further impetus.

As the Enlightenment progressed, European scholars saw physical geography as a basis for the broader discipline. These included Anton Büsching (1724–1793) in Neue Erdbe- schreibung (1754); Johann Rheinhold Forster (1729–1798) and son Georg (1754–1794) who sailed on Cook’s second voyage to the Pacific (1772–1775); philosopher Emmanuel Kant (1724–1804); and the versatile Alexander von Humboldt (1769–1859), who measured elevations, climate, and vegetation, and collected specimens across Eurasia and the Americas (Figure 13). These he categorized in terms of natural features (Physiographia), temporal change (Naturgeschichte), and spatial relations (Weltbeschreibung). His five-volume Cosmos (1845–1862) was remarkable for its time.

Understanding biogeography (left to right)

Figure 12 Carl Linnaeus (1707–1778), Swedish taxonomist, author of Systema Naturae (1735 on). 1858 painting after Per Krafft the Elder, 1858; Royal Botanic Gardens, Kew. © RBG Kew. Reproduced with permission.

Figure 13 Alexander von Humboldt (1769–1859), German geographer. 1806 painting by Friedrich Weitsch. Reproduced from Alte Nationalgalerie, Berlin.

The agrarian and industrial revolutions of the eighteenth and nineteenth centuries

The seventeenth-century scientific revolution spawned the agrarian and industrial revolutions across much of Europe and eastern North
America over the next two centuries, with a concomitant change from dispersed, mostly self-sufficient, rural farming communities to proliferating urban societies dependent on industry, trade, and services. These changes shaped new landscapes, which provided fresh challenges for physical geography. They also raised questions, notably from political economist Thomas Malthus (1766–1834), who in 1798 pondered how human population, expanding exponentially, could in future be supported by food resources, expanding arithmetically.

The agrarian revolution raised awareness of soil properties, water chemistry, land drainage, and reclamation potential, and encouraged crop rotation and stock improvement, which began changing landscapes far more profoundly than in times past. Recognizing the agricultural potential of wetlands, Dutch engineers initiated bold reclamation schemes from the English Fens to the Baltic. Gerard Boate classified Irish bogs in 1652 and scientists from Scotland to Russia began fertilizing farm crops with peat moss and exploring the industrial uses of peat. Scottish agriculturalist John Morton (1781–1864) linked soil types to parent material, American Amos Eaton (1776–1842) addressed soil fertility, and German chemist Justus von Liebig (1803–1873) enhanced soil productivity with nitrate-rich crops and fertilizers. Water was brought to drylands by British irrigation engineers in India, which from the 1850s inspired American politicians to envisage grand schemes for irrigated agriculture across the American West.

But there were also negative impacts from these changes – loss of open common land and wetland resources, and displacement by enclosure and better tools of country folk, who migrated increasingly to industrial towns in search of work. The diversion of water from rural sources to cities began to pose questions about the use and misuse of this valuable resource.

The furnaces of the industrial revolution, which relied on mineral ores, coal, and fluxes, fed the growth of geology, while needs for better transport and water supplies generated advances in science and engineering. Mining pits and canal cuttings exposed rock strata and fossils to enquiring scientists, such as William Smith (1769–1839) who compiled the first useful geological map of England, Wales, and southern Scotland in 1815. Embankments and cuttings needed for new roads and railways took Coulomb’s principles of slope stability into new arenas. Better transport conveyed more people farther, exposing them to nature as never before. In North America, completion of the first transcontinental railroad in 1869 began opening vast western lands, hitherto lightly explored. Supplying water to growing industrial towns improved understanding of climate, groundwater, and channel flows, exemplified in *Recherches hydrauliques* (1865) by Henry Darcy (1803–1858) and Henri Bazin (1829–1917). Further, while steam engines powered this revolution, their workings also favored thermodynamics through Sadi Carnot (1796–1832), Rudolph Clausius (1822–1888), and Henri-Louis Le Châtelier (1850–1936), whose studies of energy, entropy, and feedback were later invoked to explain natural systems.

The application of industrial technology to steamships encouraged further studies of oceans and distant lands. In his *Physical Geography of the Sea* (1855), Matthew Fontaine Maury (1806–1873; Figure 14) captured much knowledge about the oceans, although his bathymetry of the North Atlantic, with its 1000-fathom contour interval, barely hinted at the Mid-Atlantic Ridge. Oceanography was energized in 1872–1876, when the steam-assisted corvette HMS *Challenger* (Figure 15) carried the first modern oceanographic expedition on a world cruise, directed by Charles Wyville Thomson (1830–1882) and later
John Murray (1841–1914). The ship returned with a vast array of sediment samples, biological specimens, and data on ocean bathymetry, circulation, and chemistry, yielding 50 volumes that set new standards for marine geography. The first *Carte générale bathymétrique des océans* appeared in 1904 (Orme 2013f).

The broader impacts of these revolutions encouraged awareness of how people were changing their environment. *Man and Nature; or, Physical Geography as Modified by Human Action* (1864), by George Perkins Marsh (1801–1882; Figure 16), became a recurring theme. Stirred by Henry David Thoreau (1817–1862), American concern for nature began to differ between preservation (John Muir, 1838–1914) and managed conservation (Gifford Pinchot, 1865–1946).

**Physical geographers of the nineteenth century**

The plethora of new data from expeditions on land and sea was grist to the mill of physical geography during the nineteenth century. Most geographers focused on description, interesting to a civilizing world intrigued by then little-known regions and life forms, but the more discerning sought to explain these new facts, involving such concepts as evolution and cyclicity.

Evolution may be defined as change through time and, in a biological sense, as change in life forms through the survival of the fittest. Charles Darwin (1809–1882) commonly receives the accolade for proposing the concept *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (1858). In fact he was building on a growing awareness of fossil evidence for changing life forms, extinctions, and mutations addressed earlier by Buffon, Erasmus Darwin, Lamarck, and Cuvier, and by contemporaries such as Lyell and Alfred Russel Wallace (1823–1913). Young Darwin was so impressed with Lyell’s *Principles* that he took the first volume on board his global voyage in HMS *Beagle* (1831–1836). Regardless...
of its precise lineage, evolution profoundly impacted science in the years ahead.

Cyclic thinking was not new but had been revived by Hutton’s “succession of worlds” based on limitless time punctuated by episodic changes in an evolving Earth system. Cyclicity was also espoused over short time by catastrophists, notably in Cuvier’s six “révolutions” in the fossil record. In time the model begat proposals for diverse cycles by which to organize and explain the plethora of information now emerging on the natural world. The cycle mania peaked during the nineteenth century: climate cycles, orogenic cycles, hydrologic cycles, erosion cycles, life cycles, and so on became fodder for physical geography, as the following three examples show.

First, when Swiss paleontologist and impeccable catastrophist Louis Agassiz (1807–1873; Figure 17) lent support to glacial theory Glacial Theory in 1840, scholars began seeking to explain what Karl Schimper (1803–1867) had termed the *Eiszeit* (Ice Age; 1837). Some, such as James Croll (1821–1890; Figure 18), revisited classical sources to ponder the relevance of cyclic changes in Earth–Sun relations to climate oscillations, a theme later expanded by Milutin Milankovitch (1879–1958; Figure 19). Rhythmic bedding in older rocks was also attributed to orbital perturbations, especially the precession of the equinoxes. By the close of the century, the multiple Pleistocene glaciations, found in the American Midwest (Chamberlin, 1883) and the European Alps (*Die Alpen im Eiszeitalter*, Penck and Brückner, 1901–1909), made explanation much more challenging.

Second, in 1847, based on stratigraphic evidence for sedimentary rhythms, James Dwight Dana (1813–1895; Figure 20) proposed that

---

**Concepts of cyclicity if the nineteenth century (left to right)**

**Figure 17** Louis Agassiz (1807–1873) supported glacial theory from 1840. Sonrel 1850. Reproduced from Antoine Sonrel, PD-1923.

**Figure 18** James Croll (1821–1890) advocated astronomical forcing of climate. Irons 1896. Reproduced from J.C. Irons.

**Figure 19** Milutin Milankovitch (1879–1958) invoked astronomical forcing of climate change. Milankovitch 1943. Reproduced from Papa Jovanovic and Vasko Milankovitch.

**Figure 20** James Dwight Dana (1813–1895) recognized cyclicity in Earth history. Merrill, G.P. 1924. *The First One Hundred Years of American Geology*. Reproduced from Yale University Press.

**Figure 21** Joseph Le Conte (1823–1902) recognized cyclicity in Earth history. Merrill, G.P. 1924. *The First One Hundred Years of American Geology*. Reproduced from Yale University Press.

**Figure 22** William Morris Davis (1850–1934) advocated the cycle of erosion. Reproduced from Association of American Geographers.
“epochs” in the geologic record reflected periods of prolonged quiet alternating with pulses of rapid change and mountain building, the latter then explained by Earth’s contraction. Joseph Le Conte (1823–1902; Figure 21) stated cyclicity thus:

Geological history, like all other history, has its periods of comparative quiet, during which the forces of change are gathering strength, and periods of revolution, during which the accumulated forces manifest themselves in conspicuous changes in physical geography and climate, and therefore in rapid movements in the march of evolution of organic forms. (Le Conte, On Critical Periods in the History of the Earth … 1877, 100)

Third, the erosion cycle proposed in the 1880s by Harvard geologist William Morris Davis (1850–1934; Figure 22) reflected the evolutionary dogma of the age, expressed as a “cycle of changes” by Thomas Huxley (1825–1895) in his popular Physiography (1877). Davis’s erosion cycle was initiated by mountain building, followed by prolonged crustal stability during which landscapes were denuded, through stages of youth and maturity to old age, to plains just above sea level. This cycle could be interrupted and reset by renewed orogeny or climate change. Seemingly logical, this model relied heavily on evolutionary thinking and on the bold denudation of the American West described by John Wesley Powell (1834–1902), Grove Karl Gilbert (1845–1918; Figure 23), and Clarence Dutton (1843–1912; Figure 24). Inspired by artist renderings, notably of the Grand Canyon by William Henry Holmes (1846–1933; Figure 25), many adherents to the Davisian cycle were oblivious of its limitations. But, in reality, Davis selected just what he needed to support his model, ignoring the more profound views of Gilbert and Dutton on landform dynamism, crustal instability, and isostasy (Orme 2007; 2013c).

As the nineteenth century progressed, Hutton’s limitless “succession of worlds” was reined in by
those seeking Earth’s age in cooling rates from an initial molten state, sedimentation rates, and ocean salinity. Further, Lyell’s “present as the key to the past” was confounded by evidence for past climate changes, multiple glaciations, fossil ecologies, and sudden events such as glacier outburst floods and earthquakes. Such events, neither catastrophic in a biblical sense nor rigidly uniformitarian in Lyell’s sense, needed to be engaged for Earth’s changing landscapes.

Meanwhile, meteorology and climatology advanced, especially where time series of data were collected from fixed weather stations and ships at sea, initially using kites and balloons to probe the atmosphere. National weather services emerged after the Napoleonic wars and the International Meteorological Organization came in 1873. In 1806, British naval hydrographer Francis Beaufort (1774–1857) codified a scale relating wind force to the amount of sail a frigate could carry. This Beaufort Scale was later calibrated to anemometer data (1850) and sea state (1906). French mathematician Gaspard de Coriolis (1792–1843) analyzed the kinetic energy and rotation of moving bodies relevant to the deflection of air and water currents. In the United States, Elias Loomis (1811–1889) described tornadoes and thunderstorms, William Ferrel (1817–1891) studied cyclonic and anticyclonic circulations, and Lorin Blodget (1823–1901) produced a climate map and base data that became useful for later studies of climate change. Alexander Buchan (1829–1907) recognized weather spells for Scotland. Vasily Dokuchaev (1846–1903; Figure 28) defined the zonal influence of climate and vegetation on pedogenesis in Russian Chernozem (1883). Nikolai Sibirtzev (1860–1899) identified intrazonal and azonal exceptions to the climate forcing of zonal soils, and Konstantin Glinka (1863–1935) studied soils in Asiatic Russia. Russian emphasis on soil profiles influenced pedology in North America where Eugene Hilgard (1833–1916) examined links between Mississippi valley sediment and farm soils.

As the nineteenth century closed, physical geography was being promoted by many eminent scholars, notably in Germany by Ferdinand von Richthofen (1833–1905; Figure 29) in loess studies, Albrecht Penck (1858–1945; Figure 30) in defining multiple Pleistocene glaciations, and Otto Krümmel (1854–1912) in marine geography. Also prominent were Elysée Réclus (1830–1905) and Emmanuel de Martonne
Prominent physical geographers of the late nineteenth and early twentieth centuries (left to right)

**Figure 26**  Wladimir Köppen (1846–1940), proponent of climate classification. Reproduced from Federal Maritime and Hydrographic Agency, Hamburg, Germany.

**Figure 27**  Eugen Warming (1841–1924), Danish ecologist and biogeographer. Budtz Muller 1880. Reproduced from Royal Library, Copenhagen, Denmark.

**Figure 28**  Vasily Dokuchaev (1846–1903), Russian soil geographer. Reproduced from Russian Academy of Sciences, Moscow.

**Figure 29**  Ferdinand von Richthofen (1823–1905), German contributor to loess research. Chorley, R.J., A.J. Dunn, and R.P. Beckinsale. 1964. *The History of the Study of Landforms*, vol. I. Reproduced from John Wiley & Sons, Ltd.

**Figure 30**  Albrecht Penck (1858–1945), German contributor to Pleistocene glaciations. Reproduced from United States Library of Congress.


(1873–1955) in France, and Gilbert and Davis in the United States.

**Physical geography in the Electronic Age: the twentieth century and later**

For physical geography, the new century continued, even reinforced, many evolutionary and cyclic paradigms of late nineteenth-century science. Davis's cycle of erosion prospered, except in central Europe where structural explanations of landforms persisted. The concept of plant succession, identified by Henry Cowles (1869–1939) for dune vegetation in 1899, was refined in 1916 by ecologist Frederic Clements (1874–1945; Figure 31). Like Davis's model, Clements's scheme advocated an ideal sequence through time, from pioneer plants to climax vegetation, which could be interrupted and reset by disturbance.

Physical geography in the United States was given impetus in 1904 by the founding, on Davis's initiative, of the Association of American Geographers (AAG). The AAG was dominated early on by geologists (e.g., Gilbert, Fenneman), ecologists (e.g., Cowles, Adams), and soil scientists (e.g., Marbut), who identified with physical geography. Of the 21 papers presented at the inaugural meeting in Philadelphia, all but one were in physical geography. This new forum stressed Davis's model and discouraged views involving crustal instability and surface processes, a selectivity that caused later problems. Eventually, the Davisian model, withering by 1930, crumbled away in the mid-twentieth century under the onslaught of the plate tectonic revolution, resurgent isostasy, radiometric...
PHYSICAL GEOGRAPHY

dating, process studies, and quantitative testing. Clementsian plant succession suffered similarly, notably from analyses by Robert Whittaker (1920–1980).

Notwithstanding this seeming continuum of physical geography across the century divide, change was coming, some revolutionary, most linked to new technology. The agrarian and industrial revolutions had been driven primarily by mechanical processes applied to farming, mining, factories, and transport. In contrast, the twentieth century became an Electronic Age during which electricity and electrons in various guises were increasingly applied to agrarian, industrial, mechanical, and information systems. Powered flight appeared, offering remote views of Earth earlier seen only from the surface (and some balloons). Remote-sensing from airborne, shipborne, and space platforms became a major boost to physical geography. Twelve salient changes that have influenced the field since 1900 are now outlined.

1 Communication, information technology, and digital revolution: The Electronic Age began with experiments on electricity and electromagnetism during the early nineteenth century. The electric telegraph came around 1840, the telephone in the 1870s, wireless radio transmission around 1900, practical television around 1930, the Internet in the 1960s, and electronic mail in 1971. Meanwhile, echo sounding from 1904, and sonar (sound navigation and ranging) from 1906, improved underwater research. Radar (radio detection and ranging) emerged in the 1930s, Doppler radar in the 1940s, synthetic aperture radar in the 1950s, and LiDAR (Light Detection And Ranging) in the 1960s. Since then, satellite altimetry and geophysical surveys have probed changes in air, land, and water bodies with increasing precision. Manned, tethered, and autonomous submersibles have become essential to deep-ocean research, as have aircraft, radiosondes (from 1929), rockets, and drones to atmospheric research. Photography began recording information using sensitive photochemical film from the 1840s and computer-based electronic digital cameras from the 1960s. Over this time frame, computers evolved from mechanical calculators, through the large slow electronic mainframes of the 1960s, to the high-speed miniaturized computers of today, which enable accelerated transmission and processing of large datasets and geographic information systems. These events reflect a digital revolution that, via virtually unconstrained electronic media, has made vast amounts of data available to human and robotic receptors. Electronic publishing, often with few constraints, began augmenting or replacing printed books and journals in the 1990s.

2 Cartography: Cartography progressed rapidly with the aid of the gyrocompass (1906), aerial photography, photogrammetry, radar, lasers, computers, and, from the 1960s, satellite remote sensing and global positioning systems. At sea, closely spaced soundings and bottom sampling preceded laying transoceanic telegraph cables in the 1840s and, as technology improved, deep-sea data proliferated and sonar revolutionized the speed and accuracy of underwater mapping. As the twentieth century progressed, mapping filled many lacunae, notably for polar regions and oceans, while remote sensors began providing detailed information on subtle changes in continental elevation, ocean surface, and bathymetry. Since the mid-twentieth century, spacecraft and satellite landers have
compiled large data banks for mapping planets in the solar system.

3 Radiometric dating: The principles of radioactivity, discovered in 1896, were soon applied to mineral decay and resolutions of Earth time. In 1890, Earth was thought to be 20–400 million years old, depending on the source. By 1911, applying uranium–lead decay methods, Earth was found to be at least 1300 million years old (and approximately 4600 million years old today). Thus, if one accepted the premises of radiometric dating, Hutton’s “no vestige of a beginning – no prospect of an end” became untenable. An extended but not limitless time provides ample opportunity for the many changes in Earth’s geography recorded in rocks and readily accommodates changes over the past few million years most relevant to present geographies. In recent decades, radiometric techniques have resolved the ages of many events over Earth time, although their resolution usually diminishes with increasing time while issues of sample provenance, contamination, and age calibration remain.

4 Climate and environmental change: During the twentieth century, climate change emerged as an important issue for physical geography and society in general. Multiple Pleistocene and older glaciations and changes shown by lengthening weather records demanded explanation. In the 1920s, mathematician Milutin Milankovitch (1879–1958; Figure 19) resurrected James Croll’s orbital forcing of Earth–Sun relations as the principal driver of cyclic climate change, but doubts persisted. Then fossil-rich deep-sea cores began revealing cold-warm cycles reaching far back in Cenozoic time. Advances in coring technology, isotope geochemistry, micropaleontology, paleomagnetism, and radiometric dating now allow comparison with cycles revealed by Greenland and Antarctic ice cores (Figure 32), and from lakes and terrigenous archives, notably loess. These data support degrees of orbital forcing for global climate change, while remote sensing and field studies of glacier retreat confirm inferences on recent warming trends (Figure 33). Even so, variable solar radiation, asteroid impacts, tectonism, volcanism, surface changes,
and human activity are interwoven into climate-change scenarios (Orme 2013b; 2013d; Figure 1).

**Relative sea level change:** It was long unclear whether observed sea level changes were caused by absolute changes in ocean volume (e.g., growth or decay of continental ice sheets) or by changes in ocean-basin capacity and continental freeboard (e.g., tectonism, isostasy, sedimentation) (Orme 2013f). In 1888, Eduard Suess (1831–1914; Figure 34) invoked “eustatische Bewegungen” (eustatic movements) for sea level change, which he attributed to ocean subsidence and sedimentation. Confusingly, the term “eustasy” was then applied to climate forcing of global sea level, a usage that long prevailed until, later in the twentieth century, the issue was revisited in the context of plate tectonics and isostasy. It is now known that many forces influence sea level, depending on the magnitude of change, the spatial context, and the timescale. For example, melting of Antarctic and Greenland ice has contributed 50–80% to recent global sea level rise, which has averaged approximately 1.0–2.4 mm yr\(^{-1}\) since 1850, depending on location. Sea level studies relate to physical geography, not only for the evidence from geomorphology and sedimentology, but for their implications for climatology, hydrology, biogeography, and people.

**Continental mobility:** The growing awareness of Earth’s crustal mobility, culminating in the plate tectonic revolution of 1950–1970, dramatically altered the playing field for physical geography. Mobilist notions were not new, having been invoked earlier to explain fossil biota, but made no headway against prevailing stabilist paradigms wherein deformation was mostly confined to continental margins, mountain building to crustal contraction, and biotic migrations to failed land bridges between fixed continents. When Alfred Wegener (1880–1930; Figure 35) offered a continental drift hypothesis in 1912, expanded in *Die Entstehung der Kontinente und Ozeane in 1915*, his ideas were mostly rejected, even ridiculed. Nevertheless, in *Our Wandering Continents* (1937) Alexander du Toit (1878–1948) felt that “a great and fundamental truth is embodied in [Wegener’s] revolutionary Hypothesis,” and, in 1944, Arthur Holmes (1890–1965) suggested subcrustal convection currents as a possible mechanism for shifting continents. The plate tectonic revolution reset Earth’s stage in ways relevant to explanations of past and present landforms, climates, and biota. Ocean ridges became crustal spreading centers, ocean trenches became subduction zones, and most orogenic belts attributed to collisional tectonics. As early as 1924, Wegener and
his father-in-law, Wladimir Köppen, wrote *Die Klimate der Geologischen Vorzeit*, a study of past climates that became a recurrent theme in later research in paleogeography, notably reinvigorated by the plate tectonic revolution. In recent decades, palinspastic restorations of shifting continents and ocean basins to earlier locations have enhanced understanding of past geographies. Measurements of recent and continuing crustal mobility allow, with appropriate caveats, some prediction of future geographies (Orme 2013b).

7 *Atmospheric dynamics*: Three-dimensional views of the atmosphere began emerging from Scandinavia after 1917, notably from the Bergen School of Meteorology in Norway, a country with a large merchant fleet but whose neutrality in World War I deprived it of weather data from warring nations. This School set an array of weather stations and ships across Scandinavia and nearby seas to study atmospheric dynamics. Vilhelm Bjerknes (1862–1951; Figure 36) studied frontal waves between air masses; his son Jacob Bjerknes (1897–1975; Figure 37) and Halvor Solberg (1895–1974) examined polar-front dynamics and life cycles of depressions, and Tor Bergeron (1891–1977) formulated air-mass analysis. Sverre Petterssen (1898–1974), Erik Palmén (1898–1985), and Carl-Gustav Rossby (1898–1957) added refinements. The School’s work spread widely: Jacob Bjerknes, at UCLA (University of California, Los Angeles) from 1940, studied upper-atmosphere and jet-stream dynamics, and refined studies by Gilbert Walker (1868–1958) on conditions favorable to ENSO (El Niño Southern Oscillation). Rossby, in Chicago from 1940, studied jet streams and planetary waves. Since then,
improved instruments, remote sensing, and modeling have expanded ways for evaluating the processes affecting weather and climate (Figure 38).

8 Advances in soil science: Building on Russian concepts, Curtis Marbut (1863–1935) translated Glinka’s work into The Great Soil Groups of the World and their Development (1927) and refined podsolization, laterization, and calcification in Soils of the United States (1935). In 1938, Charles Kellogg (1902–1980) of the Soil Conservation Service (SCS, US Department of Agriculture) stressed the soil’s organic evolution in stating: “there can be no life without soil and no soil without life; they have evolved together.” Hans Jenny (1899–1992) defined climate, vegetation, relief, parent material, and time as the five Factors of Soil Formation (1941); bioturbation was added later. After 1950, a more quantitative approach began under Guy Smith (1907–1981) at the SCS (now Natural Resources Conservation Service), which led to Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys (1975). The resultant hierarchy of soil types and terminology has since been widely applied, limited only by soil complexity.

9 Process studies: Processes of environmental change, whatever their timescale, were long poorly understood. As Leonardo da Vinci opined in 1500: “although nature begins with the cause and ends with the experience, we must follow the opposite course, namely … begin with the experience and by means of it investigate the cause.” This was long a problem for scientists who inferred processes from visible legacies, such as landforms and vegetation. In critiquing the Davisian cycle of erosion, geographer John Leighly (1895–1986) stated: “Davis’s great mistake was the assumption that we knew the processes involved in the development of landforms. We don’t; and until we do we shall be ignorant of the general course of their development. In his eagerness to set up a general system, Davis jumped over the preliminary, necessarily painfully slow study of processes, and so left his system with an inadequate foundation” (Leighly, AAG Symposium, 1940, 225).

This criticism could be applied across physical geography well into the twentieth century. Certainly processes had been studied earlier, mostly by hydraulic engineers and meteorologists, but this work entered mainstream geography but slowly. By the mid-twentieth century, however, scholars were instrumenting field and laboratory studies designed to explain process-response systems across physical geography, notably the hydrogeomorphology of Luna Leopold (1915–2006; Figure 39) and colleagues...
after 1950. Today, we focus on the nature and rate of such processes, such as glacier retreat, sea level change, warming climate, and habitat degradation (Orme 2013d).

10 Quantitative revolution and modeling: The mid-twentieth-century quantitative revolution affected all geography, moving it toward measurement and statistical testing to augment, even replace, what had been a mostly qualitative and descriptive discipline. Even so, reliance on numerical data had long existed in climatology and hydrology, and geometric progressions of arterial and stream systems long recognized. Statistical methods had been advanced before 1900: probability theory by Adolphe Quetelet (1796–1874), regression and correlation by Francis Galton (1822–1911), and multivariate methods and principal components analysis by Karl Pearson (1857–1936) (Amalie Orme 2013). But most geomorphologists and biogeographers had resisted quantification while qualitative methods, deductive reasoning, and conceptual models sufficed. For geomorphology, the revolution began when hydraulic engineer Robert Horton (1875–1945) began quantifying drainage systems in the 1920s; his seminal essay in 1945 introduced a wider audience to numerical measures and statistical tests of physical properties. In 1952, Arthur Strahler (1918–2002; Figure 40) advised geomorphology to “turn to the physical and engineering sciences and mathematics for the vitality which it now lacks.” There followed, from geomorphologists, climatologists, hydrologists, and biogeographers, a suite of quantitative devices by which to improve explanation, including derivation of some spurious “laws.”

Meanwhile, conceptual and scale models spawned into probabilistic and mathematical models to better serve the needs of scientific prediction; new models were devised and old conceptual models, such as Davis’s erosion cycle, were abandoned. Weather forecasting, climate change, river behavior, flood potential, and erosion hazards became attractive avenues for predictive modeling, reliant for success on secure numerical data and time series, informed analysis, and experience, but prone to mathematical sophistry.

11 Nature and People: Building on nature as viewed by Enlightenment philosophies and changed by agrarian and industrial advances, the tenets of Marsh’s Man and Nature (1864) became a rite of passage for geographers. Early disciples of this theme included Russian climatologist Alexander Voeikov (1842–1914), French anarchist Réclus in L’homme et la terre (1905), English geologist Robert Sherlock (1875–1948) in Man as a Geological Agent (1922), and contributors to Man’s Role in Changing the Face of the Earth (1956, edited by W.L. Thomas). Such studies usually bemoaned human impacts on nature but offered few palliatives. The latter often passed unheeded, even by nations keen to advance to the assumed security of a “developed” world suffering from prolonged degradation to their land, water, air, and biota.

Human attitudes to nature have since changed, owing to life in an ever-crowding world, concern for degrading resources and threatened ecosystems, and reassessments of human interactions with nature. Stimulated by the industrial and urban revolutions, Earth’s population grew from 1 billion ($10^9$) in 1800, to 2 billion in 1930, and 7 billion in 2011; absent disaster, it may reach
PHYSICAL GEOGRAPHY

9 billion by 2050. But this growth has been unbalanced: over 1000 people km$^{-2}$ inhabited farmlands of the Ganges-Brahmaputra delta in 2011, and at least 10 megacities, each exceeding 20 million people, are anticipated on Earth by 2025. Overcrowding and the potential for starvation, disease, and conflict focus the mind, as Malthus saw more than 200 years ago. Accordingly, since about 1960, many nations have renewed attempts to conserve the best that nature has bequeathed and to plan for the future, reflected in environmental protection laws, air and water quality controls, and prudent use of resources in managed ecosystems. In the United States, the Environmental Policy Act in 1969 led to creation of the Environmental Protection Agency in 1970. Agencies and scientists from many nations have since exported their experiences, although such advice rarely sits easily with developing societies for whom economic development may take precedence over environmental protection. Modern physical geography, in its literature, curricula, and practices, reflects these challenges.

Environmental planning and management: For physical geographers, a major objective of lessons learned should be their application to environmental planning and management. Planning has a mixed heritage. Many modest plans have been implemented but designing integrated plans is more challenging for large populous regions suffering land degradation, air pollution, foul water, and influxes of alien biota (Orme 2013e). Planning for climate change also poses problems, especially where industrial reorganization and population relocation are resisted. Such challenges are exacerbated by increasing climate variability, including more frequent fires, floods, and droughts that threaten water resources. Ample scope thus exists for physical geographers to design sensitive management plans in a human context (Figure 41).

Ideally environmental planning should lead seamlessly to effective management of plans adopted. The reality is rarely seamless, especially where environmental desiderata are diluted by stakeholders, politicians, and conflict. Despite earlier ventures, effective environmental management only began to crystallize during the twentieth century and is still being refined. This process often pits local interests against regional and national priorities, conservation against development, and new money against old traditions. Most countries now have laws for the protection or conservation of “natural” landscapes and the remediation of degraded areas. But such are the problems confronting the enforcement of environmental laws that the objectives of sound management based on scientific principles may be obscured, sometimes lost. Physical geography is well placed to engage in applying its scientific heritage to the common good.

The future for physical geography

Physical geography is alive and well. Despite unevenness across nations, it is widely taught in primary and secondary schools, and in universities. It is often an integral part of curricula in geography and environmental science, serving to provide students with an introduction to nature in a human context. Portions of the field are also shared by other disciplines, such as geology, biology, engineering, and planning. But physical geography is not merely an academic enterprise; it is a method and a pathway to useful employment in such fields as hazard prediction, weathering forecasting, land
Figure 41 Lake Tahoe: physical geography as the basis for environmental planning and management. A.R. Orme 1975. The Shore-Zone Plan for Lake Tahoe, Tahoe Regional Planning Agency, California and Nevada. Reproduced by permission of A.R. Orme.

degradation, resource evaluation, urban planning, biotic sustainability, and environmental management. These opportunities may be found across the public sector and in the private sector in agriculture, industry, transport, and finance.

Physical geographers bring to specific tasks a practical awareness of the natural world and, in a human context, an understanding of how environments and people have interacted in the past, how they function today, and their prospects for the future. The time frame for understanding nature varies: it may be very long (e.g., geologic time, tectonic change), variably long (e.g., biological time, climate change),
Physical geography is a discrete but comprehensive study probably peaked in the 100 years between 1870 and 1970. Before 1870, the field was being assembled from knowledge accumulating across the natural sciences. Between 1870 and 1970, geography as a whole became a distinct discipline in secondary and tertiary education, dominated initially by its physical roots. Geography became a part of national high-school curricula and geography departments were formed in many older and most newer universities. After 1970, physical geography began declining within university geography, even as its methods and interdisciplinary strengths were being rediscovered by cognate disciplines. Flagship journals in geography began publishing fewer papers in physical geography, in large measure because the growing body of research by physical geographers was being welcomed in kindred publications. Meanwhile, geography as a whole continued to figure prominently in high-school and university entrance curricula in Europe, Asia, and the British Commonwealth, but not in the United States where, for historic reasons, the discipline had never prospered at these levels (Orme 2013e).

There are several reasons for this change in the context for physical geography at university level. First, a reawakened environmental consciousness led many universities to entertain new curricula that often bypassed existing geography departments. Some of the latter closed; others revised their emphasis. Among the latter, physical geography was sometimes relegated to introductory courses designed to attract students to geography as a whole, although more advanced components, such as geomorphology and climatology, often flourished in the absence of competitive offerings elsewhere on campus. Second, many of the traditions of physical geography were recaptured by the Earth and space sciences (reawakened by plate tectonics and planetary ventures), by life sciences (where ecology and biogeography gained fresh impetus), by engineering (where human impacts were now considered and interdisciplinary practices encouraged), and by planning (where environmental impacts and sustainability became de rigueur). Third, compared with this revived interest in the natural sciences, geography departments pigeon-holed in the social sciences were poorly placed to provide the laboratories, materiel, and initial funds needed for effective research in physical geography. Indeed, programs in the latter were often founded in, or moved to, the natural and engineering sciences. Fourth, from the 1970s onward, flagship journals in geography had to compete with new journals, some devoted to physical geography, others catering to growing cross-disciplinary subfields (geomorphology, climatology, hydrology, biogeography, and environmental science). Meanwhile, established journals and publishers in other fields, such as geology and ecology, welcomed contributions from physical geographers.

In short, physical geography did not die after 1970; it broadened its base by returning to former roots in the natural sciences, which by then were flourishing amid new challenges and aspirations, and by finding fresh applications in the engineering sciences and in urban and regional planning. Physical geography is indeed alive and well, but its practice, content, and emphatic human significance now extend far beyond its base established earlier within the discipline of geography. Herein lies the future of physical geography.
SEE ALSO: Biogeography; Biogeography: history; Cartography: history; Climatology; Climatology: history; Cryosphere studies: history; Geomorphology: history; Glaciers; Landforms and physiography; Oceans and seas: physical geography; Rivers and streams

References


Phytosociology is a branch of vegetation science that deals with current plant assemblages (communities) at a spatial grain size of vegetation stands. Its principal goals are the delimitation and characterization of vegetation types based on the complete floristic (species) composition. Floristically similar concrete stands (phytocoenoses), represented by plot records (so-called relevés), are combined into abstract vegetation types (syntaxa; singular: syntaxon). In analogy to plant taxonomy, phytosociological classification (syntaxonomy) places vegetation units into a hierarchical system based on varying degrees of floristic similarity.

History

The fundamental concepts, methods, and terminology of the discipline were developed by the Swiss-French botanist Josias Braun-Blanquet from the 1920s onwards (Braun-Blanquet 1964). He combined standardized sampling of plant communities on plots, sorting of species-by-plot matrices, delimitation and naming of vegetation types, and their placement into a hierarchical classification. Phytosociology is often called the “Braun-Blanquet approach” (Westhoff and van der Maarel 1973) or, in accordance with the founder’s main places of work, the Zurich-Montpellier school of vegetation science. Methodological developments and the new possibilities of computers have changed phytosociology significantly since the days of Braun-Blanquet, but in essence modern phytosociologists still apply the same ideas that have proven to be so productive.

When plot-based vegetation classification emerged as a scientific task in the first half of the twentieth century, there were several regional “schools” with methodologies different from the Braun-Blanquet approach, the latter prevailing in central and southern Europe (see De Cáceres et al. 2015). Particularly in the Anglo-American scientific realm, the value of a typology of plant communities was fundamentally questioned at that time, based on an individualistic community concept. In retrospect it is evident that only phytosociology managed to create comprehensive and widely applied classifications of vegetation types for larger spatial areas, while all other approaches remained regional or ceased to exist.

Nowadays, phytosociology is the mainstream method throughout Europe and is applied also in northern Asia and various regions of Africa and Latin America. While its application in North America remained limited, relatively recently the US National Vegetation Classification (USNVC; see Peet and Roberts 2013) was launched, which acknowledges the importance of consistent hierarchical classification systems and adopts ideas of the Braun-Blanquet approach in a modified terminology.
PHYTOSOCIOLOGY

Placement and size of plots

While traditional phytosociologists thought that one could represent a whole stand (patch) of vegetation with a single “typical” relevé, current phytosociology is seen as a statistical approach that aims at characterizing vegetation types by the combined information from many different plots. Therefore, it is generally better to sample many smaller rather than fewer larger plots. Sampling strategies normally aim at covering the full variability of vegetation within the defined geographical and ecological extent of a study, while minimizing the within-plot heterogeneity. These two targets can hardly be achieved by a spatially random sampling because this would miss geographically restricted types, which, however, may hold a distinct place in ecological space. As statistical inference across geographic space is usually not the intention of phytosociological sampling, a stratified-random or even a subjective placement of plots is preferable.

Formerly, phytosociologists thought that one could determine a “minimal area” for each vegetation type above which the number of species does not further increase. Sampling vegetation on plots corresponding to or larger than the specific minimal area would yield comparable results, irrespective of different plot sizes. According to modern empirical and theoretical knowledge, however, mean species richness monotonously increases with area and a minimal area is a deception caused by the nonlinear nature of species–area relationships. As most approaches for classification of vegetation are sensitive to different plot sizes, the recommendation is to apply uniform plot sizes within a few structurally delimited formations. In Europe, typically, areas between 4 and 25 m² are used for herbaceous types, and those between 50 and 400 m² for scrublands and forests. In regions without extensive legacy data, a single standard area of 100 m² might be advantageous.

Phytosociological relevés

The data of a single plot are called a relevé (French for “record”); they consist of “header” and species data. The header data comprise plot identification, information on methodology (e.g., date, researcher, plot size), geographic localization (name of the locality, geographic coordinates), environmental data measured/determined on the plot (e.g., altitude, slope and aspect, microrelief, soil parameters, land use), and structural data (i.e., cover and height of the distinguished vegetation layers). The selection of parameters and the precise approach for their measurement will depend on aims and resources of the project in whose context they are sampled, but there are attempts to establish common minimum standards to increase comparability.

The core of a relevé is a list of plant taxa (species and infraspecific taxa, further called species) found on the plot and their attributes. A peculiarity and, at the same time, a major reason for its success is that phytosociology aims at compiling a complete census of species instead of focusing on dominant or woody species only. All visible living vascular plants are included, often also the terricolous bryophytes, lichens, and macro-algae, and sometimes even species growing on substrates other than soil (e.g., epi-phytes). Every species is assigned to a layer (e.g., tree layer, shrub layer, herb layer, cryptogam layer), and species occurring in more than one layer (e.g., young individuals of tree species in the herb layer) are recorded several times. Each observation of a species in a layer is assigned
Table 1  Extended Braun-Blanquet cover-abundance scale as it is widely used (in the original Braun-Blanquet scale 2m, 2a and 2b were joined as “2”). For numerical interpretation, either the mid-points of the cover range or the ordinal values are used.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Abundance</th>
<th>Cover (%)</th>
<th>Ordinal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>1</td>
<td>0–5</td>
<td>1</td>
</tr>
<tr>
<td>+</td>
<td>2–5</td>
<td>0–5</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>6–50</td>
<td>0–5</td>
<td>3</td>
</tr>
<tr>
<td>2m</td>
<td>&gt;50</td>
<td>0–5</td>
<td>4</td>
</tr>
<tr>
<td>2a</td>
<td>Any</td>
<td>5–12.5</td>
<td>5</td>
</tr>
<tr>
<td>2b</td>
<td>Any</td>
<td>12.5–25</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Any</td>
<td>25–50</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>Any</td>
<td>50–75</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>Any</td>
<td>75–100</td>
<td>9</td>
</tr>
</tbody>
</table>

Phytosociological classification

Classifying vegetation has three fundamental goals (Dengler, Chytrý, and Ewald 2008):

- delimiting and naming parts of the vegetation continuum to enable communication about them
- predicting a multitude of ecosystem attributes (e.g., species composition, site conditions, ecological processes, conservation relevance, appropriate management)
- making multispecies co-occurrence patterns representable by verbal descriptions, tables, diagrams, and maps.

Since vegetation classifications serve many different purposes in fundamental and applied ecology, it is not surprising that no single classification can serve all purposes optimally (Dengler, Chytrý, and Ewald 2008; Peet and Roberts 2013). There is a fundamental tradeoff between having ecologically meaningful and easily recognizable units. Defining temperate European forests according to their dominant tree species, for example, would be easily conceivable even for laymen, but a category “beech forests” would include ecologically dissimilar stands while ecologically similar stands dominated by other trees would be excluded.

Phytosociological classifications stand out from other types of classification in four ways:

- They classify vegetation based on the assemblage of species, making the outcome independent of a priori knowledge of environmental factors and letting the species “tell” what is relevant for them (Peet and Roberts 2013).
- They aim at being suitable for many purposes.
- They aim at being integrated into a single consistent classification system.
- They are hierarchical, allowing for the coding of information effectively at different degrees of floristic and ecological similarity and for the use of different levels depending on specific needs.

All these aspects made phytosociological classifications an efficient system for collecting
and storing information on the vegetation continuum. While in syntaxonomy, as in taxonomy, no two classification systems are identical, the approach is so robust that in regions with phytosociological tradition, syntaxa became the most widely used and understood system for habitat classification, even for non-phytosociologists.

Historically, phytosociological classifications were achieved by manual sorting of species-by-plot matrices, so-called relevés tables. The aim of the sorting was to find groups of floristically similar plots and join them into vegetation units. This was achieved by arranging the discriminating species within the study in “blocks” along the matrix diagonal. Later, algorithms like the still widely used TWINSPAN emulated this sorting approach in a software tool. Nowadays, a wide array of different numerical tools for unsupervised classification are available to phytosociologists, each with specific pros and cons (see Peet and Roberts 2013; De Cáceres et al. 2015; IAVS 2015). However, unsupervised classifications can only use the information contained in the specific dataset. As, for practical and computational reasons, numerical classifications are always carried out for geographically and ecologically restricted datasets, the challenging task is to combine these separate classifications into a single consistent classification system. This latter step still involves much manual work and interpretation, but increasingly tools of supervised or semisupervised classification become available that help with this task (see De Cáceres et al. 2015).

Once a classification is achieved, its units should be validated, both internally (e.g., how well they are separated floristically) and externally (e.g., how well they predict site conditions), and, if necessary, be refined (see De Cáceres et al. 2015). The next step then is the characterization of the distinguished vegetation types (De Cáceres et al. 2015). While the original classification procedure was based on the full species composition, the floristic characterization of a syntaxon aims at highlighting those species that are most important for its recognition. These species are the dominant (those with the highest cover), constant (those with the highest frequency), and diagnostic (those with the highest “fidelity” to that unit) species. The latter can be determined by numerical criteria, such as constancy ratio or total cover ratio, between the target unit and other units or with statistical measures of concentration in the overall dataset, such as the widely applied phi values. Among diagnostic species, one can distinguish character species, discriminating the target unit against all other units of the same rank, and differential species, being diagnostic only within the next-higher syntaxon. Characterization of a vegetation type also includes the determination of its altitudinal and distributional range, frequency, site conditions (through direct measurements in the included plots or inference from the species composition via mean indicator values; see Dengler, Chytrý, and Ewald 2008), disturbance and management regimes, and conservation status.

A specific aspect of phytosociology compared to other schools in vegetation science is that it provides a system for consistent naming of the derived vegetation units that is applicable globally. The naming of formalized units of phytosociological classifications, so-called syntaxa, is regulated by the International Code of Phytosociological Nomenclature (ICPN; Weber, Moravec, and Theurillat 2000). The rules are similar to those that have successfully governed the naming of taxa since the times of Linné. The two basic principles are the priority principle (the oldest valid name must be applied) and the type principle (that attaches each name to a type element). Together they ensure that the system can be built inductively,
Table 2  Syntaxonomic ranks of the ICPN. Not included are the nonobligatory ranks subclass, suborder, and suballiance.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Termination</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>-etea</td>
<td>Festuco-Brometea</td>
</tr>
<tr>
<td>Order</td>
<td>-etalia</td>
<td>Brachypodietalia pinnati</td>
</tr>
<tr>
<td>Alliance</td>
<td>-ion</td>
<td>Bromion erecti</td>
</tr>
<tr>
<td>Association</td>
<td>-etum</td>
<td>Gentiano-Koelerietum pyramidatae</td>
</tr>
<tr>
<td>Subassociation</td>
<td>-etosum or “typicum”</td>
<td>Gentiano-Koelerietum pyramidatae agrostietosum tenuis</td>
</tr>
</tbody>
</table>

that is, bottom-up, before all vegetation types of a region are recorded, and that an inflation of names is avoided when units are joined or split. Phytosociological nomenclature knows four principal ranks, with association being the basic one (Table 2). The names of syntaxa are derived from the scientific names of one or two (or up to three in the case of subassociations) species typical for the type, which are joined with connecting vowels and specific endings. The results are short and easily memorable names, which, despite their conciseness, carry an idea of the character of the unit.

Phytosociological databases

Over 100 years of phytosociological surveys have resulted in several million published and unpublished relevés. Nowadays, they become more and more available in regional, national, and supranational vegetation-plot databases. Presently, the databases registered in the Global Index of Vegetation-Plot Databases (GIVD; www.givd.info; Dengler et al. 2011) alone comprise data from nearly 3 million plots, most of them from phytosociological studies. Combining and harmonizing data from different databases for joint analyses is still a big challenge, given the differences in structure and software of individual databases, different header data fields used, and varying species concepts followed. However, both in Europe (European Vegetation Archive, EVA) and in the Americas (Botanical Information and Ecology Network, BIEN3) initiatives are ongoing to establish continent-wide vegetation-plot archives. The potential of national and supranational databases for fundamental and applied ecology is enormous. Unlike coarse-grain distributional data of individual species as so far used in macroecology and species distribution modeling, vegetation plots contain information on actual species co-occurrence, true absences, species’ relative performance and, in most cases, environmental variables determined at the plot and not just derived from geographic information system maps.

SEE ALSO: Biogeography: history; Biomes; Community/continuum in biogeography; Ecoregions; Ecosystem; Supervised classification; Unsupervised classification

References

PHYTOSOCIOLOGY


Further reading

Place

Paul C. Adams
University of Texas at Austin, USA

Place is arguably one of the two or three most important concepts in the discipline of geography. The ways in which geographers address questions of place are inevitably tailored to their particular subdisciplinary interests, but there is a prevalent understanding of place as the coming together of heterogeneous processes. This interest in dynamism and heterogeneity indicates a refusal by geographers to reduce place to location, where things merely happen to come together, or to a container that tidily isolates certain phenomena from the rest of the world. Understanding place as essentially a process of synthesis allows room for the real world’s complexity, with its interplays of space, time, and matter, rather than resorting to an idealized “view from nowhere” in which phenomena are theoretically separated from place. What happens in a place is part and parcel of the place, both materially and conceptually, so the motivation for studying geography can be seen as a full appreciation of “the inter-relationships among the physical environment, the built environment, and the people” (Johnston 1991, 101).

In regard to the various causal chains that operate in the world, place is more than just the end of the line where we can observe the playing-out of effects. Place is also a constant source of causal chains that radiate out into the world. Many places are known for their roles in affecting the world, and some place names serve as coded references to ripples of effects that are still radiating outwards: Seneca Falls (women’s rights in the United States), Maastricht (the European Union), Hiroshima (the era of nuclear weaponry), and so on. Place meaning is a collective production and as such it gives coherence to collective actions such as protecting women’s rights, respecting EU policies, or avoiding a nuclear holocaust. Place identity also contributes to the constitution of personal identity, even as it serves to reflect identity, as in the meaning of being a woman or man, a European or non-European, a supporter or opponent of nuclear armament.

Despite shared, intersubjective strands, place means something different to every person. Pulling the strands together in a coherent fashion poses a major theoretical challenge. Geographical understandings of place are distinct from everyday, vernacular understandings of place, but they engage critically with these vernacular meanings, as well as with the narrowly focused place meanings from other areas of scholarship and creativity. “Place” in the geographer’s lexicon cannot lose touch with place as it is understood in daily life or other branches of scholarship. Daily life is intensely bound up in place so place itself escapes conscious awareness much of the time. Place references pervade everyday communication but the meaning and role of place remain paradoxically indefinite. Academic thought, in contrast, constructs very precise ways of defining place, but only by grossly simplifying place, either to location (in the natural sciences), to social categories and hierarchies (in the social sciences), or to a type of experience (in the humanities). The first of these, scientific knowledge, is often produced in laboratories but it is limited by its dependence
on such places since they lack the messiness of the real-world places where most things happen. In short, ordinary people live place more than they can articulate it, while scientists achieve ontological coherence by walling themselves off in very peculiar places. Either approach obscures place, creating a need for an approach that analyzes the phenomenon of place in general, along with the processes, patterns, and phenomena that are evident in particular places. Geography works in this direction, most explicitly through social science and humanities approaches, but also implicitly, through a shared commitment to maintaining the cohesion of geography itself, since the discipline addresses place in epistemologically diverse ways.

Geographical interest in place stems from an assumption that processes of all sorts (political, economic, cultural, biological, geological, and so forth) happen with places rather than merely in places. This distinction is subtle, but important: for geographers, processes are shaped, affected, reworked, modulated, organized, coordinated, and inflected by places. Place thereby plays an integral role in the world, making things possible and allowing them to emerge and develop in particular ways. Places of geographical interest come in all sizes, from a room to the cosmos, but it is now generally accepted that scale itself is a construct, so that, rather than simply being larger or smaller, places are actively scaled larger or smaller. A room is a place, but a country can also be understood as a place, and even the Earth as a whole is a place, at least when we capture it in images like “spaceship earth,” a blue marble floating against a field of stars (Cosgrove 2003).

The betweenness of place

A formative moment in the evolution of geography’s understanding of place was the Anglo-American geographical debate of the mid-twentieth century between advocates of chorographic versus nomothetic approaches. The chorographic (or chorological) approach built on foundations established by the Greek philosopher and geographer, Strabo, and the seventeenth-century writings of Varenius, but it found its clearest theoretical justification in Immanuel Kant’s eighteenth-century concept of geography as a synthetic discipline. According to him, the geographer’s task is to explore how phenomena come together in space just as historians explore how phenomena come together in time; both strive for disciplined and rigorous descriptions, whether of a particular time or place. Richard Hartshorne developed this idea in the mid-twentieth century, drawing on the German geographers Alexander von Humboldt, Ferdinand von Richthofen, and Carl Ritter to characterize geography as the study of areal differentiation. According to him, places are “element complexes,” combinations of phenomena revealed by the systematic sciences ranging from geology and biology to anthropology and economics. What makes a place significant is the unique coming together of phenomena, so the geographer’s task is ideographic: writing about the particular.

This approach was critiqued as “exceptionalism” by proponents of a nomothetic approach. Under the banner of spatial analysis it was argued that geographers, like other scientists, must search for generalizable relationships and ultimately discover laws governing interactions between and within places. Distance and population would serve as explanatory variables, much like distance and mass do in physics. While the idiographic versus nomothetic debate focused mainly on whether geographers should synthesize the findings of other sciences or develop their own body of theories and laws, it also involved a fundamental disagreement about the
meaning of place. The nomothetic geographers treated place as a location in space while the idiographic approach, identifiable as chorology or chorography, insisted on treating place as a much messier thing, albeit under the rubrics of “region” and “area.”

While spatial analysts continued formulating increasingly complex location-based models into the 1960s and 1970s, a newly emerging approach took a different angle and blended place’s particular and general sides. Humanistic and phenomenological geographers described place experiences in unprecedented depth and detail, borrowing from the social sciences, philosophy, and the humanities, including history. If the ultimate goal of spatial analysis was to predict interactions between multiple locations, and the ultimate goal of chorography was to capture the uniqueness of place, the countervailing goal of humanistic and phenomenological geography was to articulate the universal phenomenon of place experience as revealed by particular people living in particular places.

In contrast to these approaches, several distinct approaches emerged in the 1980s and 1990s that were committed to exposing how places are caught up in, and contribute to, social power relations. These included Marxist/Marxian, feminist, poststructural, postmodernist, and postcolonial methods. While divergent in regard to the power relations exposed by their critiques, these approaches shared the assumption that places are best understood in terms of social power relations. Groups that are at odds with each other almost inevitably divide around the issue of place as they appropriate different places, appropriate the same places in different ways, appropriate different parts of places, scale places differently, or do all of the above. The goal of studying place through power relations is to find out what unites them. These various approaches all seek to intervene in the dynamics that make place a contested terrain.

These three meta-approaches could be parsed as place-as-location, place-as-experience, and place-as-contested-terrain. They pervade the discipline, adding richness but also mutual misunderstanding. Tensions arise between them when divergent goals channel teaching and research efforts. Many geographers attempt to navigate between these extremes, however, and articulate an ideal, albeit difficult to achieve, in which the essence of place is its ability to synthesize multiple characteristics. For Entrikin (1991) such efforts boil down to an exploration of the “betweenness of place,” with place being a lens that combines objective and subjective realities. He argues that the study of place calls for ontological reconciliation between a decentered view from nowhere and multiple views from somewhere, between objective and subjective perspectives.

Agnew and Duncan (1989) conceive of place in terms of a triad: location, locale, and sense of place. This approach incorporates geometrical/physical, social/cultural, and experiential/emotional elements. While location equates place with position in space, locale consists of “the settings for everyday routine social interaction provided in a place,” and sense of place is a form of “identification with a place engendered by living in it” (Agnew and Duncan 1989, 2). Rather than mutually incompatible alternatives, the elements of this triad are offered as complementary dimensions of place and they have been taken up in this spirit by many authors engaging with the topic.

Lynn Staeheli (2008) adds detail, identifying five distinct aspects of place that have been treated in the geographical literature: place as physical location or site, place as a cultural and/or social location, place as context, place as something constructed over time, and place as
process. What produces a place, for Staeheli, is the systematic way in which each facet of that place structures the others. Whether we adopt Entrikin’s notion of betweenness, Agnew and Duncan’s tripartite system, or the five aspects indicated in Staeheli’s review, careful study reveals place as a complex synthesis and interplay of heterogeneous elements. These attempts emerge out of a shared assumption that place is essentially integrative or synthetic. The synthetic quality of place is astonishingly pervasive. Even the planet Mars, out of reach in the nineteenth century, was transformed into a place by the debate among “experts” on its imagined geography (Lane 2011). The social relations between these astronomers imposed competing senses of place time and civilization onto the distant body, although it was barely more than a dot in their telescopes.

A close look at any given place reveals its synthetic quality. This quality does not produce singular outcomes, although it typifies all places. Rather, its multidimensionality permits the creation and evolution of a multitude of distinct places, making place as difficult to analyze as it is important.

Location

Location is primary in some geographical approaches to place. Cartography, remote sensing, and GIS subsume sense of place and locale to locational attributes. Studies of geographical patterns emphasize location, as do efforts in economic geography to quantify and map the global flows of goods, capital, and information, or in political geography to trace flows of bodies, power, and authority. Perhaps the most locationally deterministic approach developed thus far is central place theory which assumes that the relative locations of places (their distances from one another) and their relative masses (populations or economic activities) together determine how places will interact and develop. In the early twentieth century, Walter Christaller theorized that large, medium, and small urban centers would ideally be distributed in hexagonal patterns. The emergence and growth of population centers would be governed by the imperatives of markets, transportation, and administration. In each case settlements would ideally form hexagonal patterns, but the relative frequency of small and large settlements would differ depending on the driving forces in a particular area. It was hoped that once local contingencies were taken into account such patterns could be identified in the distribution of actual urban settlements; however, the patterns predicted by central place theory bore scant resemblance to the actual layout of urban centers. Distance alone did not determine geography, owing to contingencies of site and situation such as terrain, soil, climate, vegetation, water bodies, resource distributions, historical factors (such as the layout of transportation infrastructure), and the choices made by individuals, groups, and governments.

Geographers in the spatial analytic tradition rejected the notion of place as singular, assuming that all relevant characteristics of place could be quantified and their interactions would yield a coherent set of laws. Systematic analysis of the spatial patterns of such thinned-down places appeared to offer a way of rescuing geography from widespread accusations that it lacked scientific rigor, but it became increasingly difficult to defend the relevance of this approach because it lacked predictive power, and in addition it turned people into little more than dots on a map, while avoiding moral and ethical questions. Ultimately, the pared-down notion of place inherent to spatial analysis proved untenable. The reduction of geographical place to location is logically flawed because it separates space from
time and matter, and laws of geometry cannot, on their own, explain the patterns observed in geographic space (Sack 1973). In addition, the key figures in the quantitative revolution were deeply embedded in particular places, such as academic departments and research laboratories, despite their aspirations to transcend the peculiarities of place.

Despite its limitations and contradictions, spatial analysis contributed useful concepts to the study of place. The most important is time–distance. Torsten Hägerstrand initiated time geography, and he and his followers explored the ways in which people’s movements through space are structured by time, authority, and access to various forms of transportation. The options for face-to-face interaction are constrained in multiple ways such that people inhabit invisible prisms demarcating the farthest points they can reach in a given span of time, using a particular mode of transportation and assuming a certain period to be spent at the distant location. Activity occurs not within an unlimited space but within the spatiotemporal constraints imposed by the authority of territorial systems, the physical limitations of one’s own body and vehicles, and the spatiotemporal constraints of one’s social contacts, colleagues, clients, and customers.

Insofar as distance is more often encountered in terms of travel time or communication time than in terms of geodesic distance (as a crow flies), technological changes can be understood as bringing places closer together functionally and experientially, producing time–space convergence. A complementary process of divergence also occurs through various place-based cyclical processes. Donald Janelle argued that time–space convergence leads to centralization (e.g., fewer and larger stores) and specialization (e.g., proliferation of different types of stores), both of which create a demand for improved accessibility. This demand motivates transportation innovation, which may result in further time–space convergence but could also have the unintended consequence of creating divergence; for example, traffic congestion resulting from increased automobile dependency. Related ideas have been advanced, although often with different research objectives, under the name of “time–space compression.” David Harvey argued that time–space compression arises from the acceleration of geographical flows of people, goods, and capital in support of capital’s imperative to conquer space. Time–space compression leads inevitably to tensions between the acceleration of spatial flows and the construction of costly and immobile infrastructure to support those flows. From these various perspectives, innovations in transportation and communication can be seen as radically changing the functional separation between locations, transforming the practical impacts of distance, and thereby creating, transforming, sustaining, and destroying places.

Locale

Throughout the social sciences place is often equated with community but this equation is misleading, since all but the smallest places bring together multiple communities, and there are communities that bind lives across multiple locations. Confusion arises when elements of place are thought to be derived from social relations rather than co-constituted with social relations. A geographical reply points out that a town cannot be reduced to a list of its inhabitants and their social ties any more than it can be reduced to the physical infrastructure at a particular location. A place is the three-way interaction between social relations bound up in one or more mutually interacting communities, varieties of place
Reducing place to community is, in fact, ethically problematic since the idea of community often masks insidious assumptions about who belongs and who does not belong, or in Cresswell’s (1996) terms, who is “in place” and who is “out of place.” Prejudice can lead to an exclusive definition of the community as a group of people that belong in a place, while people not fitting a certain definition are subject to treatments such as expulsion, marginalization, exclusion, and oppression. Shopping malls may expel youth who are just “hanging out” rather than consuming, enforcing an idea of community as consisting of consumers even while evoking nostalgic images of a less materialistic kind of community. Small towns and suburbs may be refuges imbued with purified place images that foster prejudiced ideas of outsiders and stereotypical ways of understanding insiders. In similar fashion a much larger place, the nation, mobilizes racially and ethnically exclusive concepts of identity at borders, ports, airports, and frontiers, where civil servants seek to control the bodies and actions of people moving through. In any place, at any scale, a dynamic tension can be identified between multiple communities in the place, each of which competes to define the place through its communications and actions. Through these processes we see how sense of place interacts with the social dynamics of locale as groups are “denied a place in society through a particular construction of place” (Sibley 1995, 107).

Place’s importance to the formation of collective identity also makes it a lever with which to bring about societal change. A gay pride march, a protest occupation, a flash mob, a work of performance art, or a riot each in its own way uses place to challenge the social order. Their potency derives from the dependency of the social order on place, and specifically on sense of place and locale. Such extraordinary events are essentially performances that “take place” in the dual sense of claiming place and happening. The seizure of public space by groups may be permitted, for example, in the United States under the freedom of assembly clause of the Constitution, but in most cases the actual use of place to challenge collective notions of order and community is tightly controlled and regulated in both space and time. New digital technologies and social media such as Facebook and Twitter have facilitated the strategic occupation of places, allowing protestors to mobilize quickly and to relocate or change tactics in response to police and military intervention. These media have also facilitated the task of gaining the support of distant audiences and using them as a means to put pressure on authorities. Technology’s influence on struggles to seize territories and redefine places can be seen in “developing” and “developed” countries alike. The power of place is therefore reworked by coordinated manipulation of performance in place and technologies that permit people to communicate through space: that is, among scattered protest participants, and also to audiences who live in distant locations and have access to different political resources.

For most people, the place with the most resonance is the home. Adopting a place as one’s home is “the most highly charged commitment one can make in that realm wherein the personal is political, [and] it is also an affirmation to be present in the moment” (YoungBear-Tibbetts 2007, 158). Part of the appeal of home is that, notwithstanding power differentials based on age, sexuality, and gender, one is mostly “in place” there. Home offers a refuge from globalizing processes, but this sense of place can be deceptive insofar as home offers at best a permeable membrane shielding the private world from the economic, political, and physical risks of the public world. The dynamics are complicated and
people may understand their homes as places of refuge or escape even when these homes are colonized by the control mechanisms of a global space of flows. Some people bear the deeply compromised status of “homeless.” The lack of a home is socially (and structurally) imposed on them, since they may “think of themselves as having places to call home even when they were sleeping rough but these places simply do not count in the eyes of those who encounter them” (Cresswell 2004, 117). Locale therefore involves warmth and conviviality but also vicious cycles of exclusion and marginalization.

An interest in locale also directs us to consider how places are colonized by globalizing political and economic forces. While capital is mobile, place’s fixity permits the expansion of capital flows. Globalizing forces relate to places in terms of exploitable resources, incorporating extraction and production facilities as well as infrastructure for shipping and distribution (Harvey 2010). Economic changes can disrupt sense of place through, for example, the leveling of forests, the exhaustion of fisheries, the demolition of inner-city neighborhoods, and the closing of factories. Memories of an earlier time, along with a subsequent sense of loss, can linger and haunt a place. Economic development therefore reworks places, modifying the elements of locale to fit social structures based in larger scales, including national, regional, and global economies. In an ironic twist, not only the resources in a place, but even a sense of place, can be sold to tourists and other outsiders as symbolic carriers of uniqueness, packaged to cater to standardized tastes. Cancun, Mexico, for example, has been reworked to suit the tourist’s gaze and desire for entertainment, and has become what local observers refer to as gringolandia (Torres and Momsen 2005).

In response, place attracts its defenders, those who seek to protect a “space of places” in opposition to capitalism’s “space of flows” (Castells 1989). The symbolism mobilized in such struggles can be progressive. The discovery of common causes and interests among multiple communities can help make a place into a shared home. Such homes can resist the forces of globalization; however, any social order erases certain understandings of a place while promoting others, and when a group defends “our place” they are quite often solidifying one group’s social power and privilege over that of other groups with equally valid claims to that place. Place is therefore deeply ambivalent, providing a scale and focus to movements that resist the expansionary, globalizing forces of capitalism while at the same time fostering conservative and even reactionary politics.

In contrast, there is what Massey (1993) would call a global or progressive sense of place – one that recognizes the many place-making activities that are nonlocal. This view resonates with highly mobile lifestyles incorporating a multitude of communication technologies into place-making. Flows of ideas, information, and emotion enter and leave a place through communications, not only dissolving the boundaries between places but also accentuating the distinctiveness of places as switching points or nodes. Human bodies increasingly occupy transnational spaces, routinely circulating between scattered, rather than clustered, places. The challenge is not just to think carefully about the sending and receiving places of migration, with their push and pull factors, but also to reflect more generally about how “migrant and other mobilities are embodied, embedded and grounded” (Blunt 2007, 691).

While place has traditionally been assumed to require spatial fixity, many people are obliged to develop a sense of place that is not limited to a particular locale; their sense of place incorporates multiple locations or movement itself. Not only refugees, migrant workers, and expat retirees but
also tourists and business travelers “dwell through travel, and vice versa” (Ralph and Staeheli 2011, 519–520). Mobility is nothing new. In many cities of the world there are neighborhoods that have changed hands and identities multiple times as older immigrant populations have given way to newer immigrants. However, over time the “balance between ‘relations of presence’ (local ties) and ‘relations of absence’ (extra-local ties) has moved in the direction of the latter” (Agnew 1989, 24; emphasis original). The proliferation of communication technologies, their permeation of the places of daily life, the acceleration of message transmission, digitization, and media convergence all play roles in this evolution of locale. Claims have been made that new communication technologies will allow people to stay at home or otherwise reduce their mobility, strengthening ties to place. Rather than substituting for travel, however, information and communication technology complements, accompanies, directs, and stimulates travel, migration, and mobility. Even relatively poor immigrants remain in touch with family through interpersonal and social media, as well as by sending remittances, and their sense of place may be split between multiple homes separated by international borders and thousands of miles. At the same time, poverty, political oppression, and lack of education lock some people into comparative immobility, turning place into a kind of trap or prison.

Sense of place

Sense of place builds on the locational and social elements discussed in the previous sections, creating an “organized world of meaning” (Tuan 1977, 179). In a familiar place, every object has a meaning so the places one knows best become “centers of felt value” (Tuan 1977, 4). People attach meanings and emotions to locations and a sense of place develops over time, building up layer on layer. Variations in sense of place indicate the inherently subjective nature of this experiential phenomenon and the challenge of understanding or empathizing with others’ place experiences. A school may be an exciting and fun place to one student and a dull and oppressive place to another; likewise, a home may be experienced as a sanctuary by a man and a prison by his wife, or vice versa. Such variations in sense of place are often deeply politicized as people defend a sense of place rooted in one narrative and dismiss countervailing narratives as distortions and delusions. Nonetheless, a sense of place cannot be mechanistically derived from social structures or dynamics. Locale affects sense of place and vice versa.

If we follow Martin Heidegger in framing the experience of human life as being-in-the-world, then place arises phenomenologically from human intentions and actions, rather than abstractly from categories and measurements. Sense of self and sense of place are formed through a constant interplay. The implication for place scholarship is that deepening one’s understanding of place encourages a critical understanding of oneself and, more generally, of the construction of self-identity and subjectivity. The recent return toward issues of practice and performance under the rubric of “nonrepresentational theory” revitalizes key questions of the phenomenological approach, although the terminology and many of the assumptions have evolved. While Thrift calls for a nonrepresentational “style” of research that would be “a means of valuing and working with everyday practical activities as they occur” (2008, 112), this is remarkably similar to Heidegger’s interest in Dasein (being-in-the-world), which grounds reality in particular purposes and actions, or Husserl’s interest in Akterlebnis (act experience).
The main differences between phenomenology and nonrepresentational theory may be terminological, although the latter advances somewhat more modest claims.

Sense of place includes a range of emotions from profound attachment to dread. The former has been called topophilia (Tuan 1974). The latter may arise from incidences of violence or tragedy, which lead to community responses such as the designation of the site as important to collective memory, the obliteration of lingering signs of the tragic event, rectification of the site by returning it to the status of an ordinary place, or in some cases sanctification of the place as a kind of shrine (Foote 1997). These countervailing notions of topophilia and places of tragedy are ineluctably subjective and interlocked. People migrate to places in pursuit of their own topophilic ideals, but an influx of new residents may cause long-term residents to perceive a deterioration in their place's identity. Here we can see one person's topophilia leading to another's topophobia.

Mediated communication involving books, radio, television, the Internet, and many other media is closely tied to sense of place. Place experiences circulate as generic and particular images in the media: for example, war zones on the nightly television news or beach resorts in advertisements for packaged vacation tours. Such place images interact with collective memories and with each individual's personal experiences, then feed back into first-hand experiences of particular places. Media are encountered in particular places and media represent places, but in addition media also create places, in that they bring together spatially distributed audiences or participants in locales of shared experience. Particular media, texts, and discourses can function much like gathering places when people interact around these media products and their conversations touch base on certain themes, stories, characters, and worldviews. In this regard, mediated place images and media-altered places connect with other trends toward mass production: for example, the proliferation of chain restaurants and stores, international architectural styles, and standardized spaces of travel such as freeways and airports.

In connection with these processes, academic disagreement has revolved around questions of authenticity, and particularly whether there has been a cumulative weakening of place attachment and a progressive loss of local distinctiveness. For some observers, there is a growing homogenization of place. These discussions bring together questions related to sense of place in a way that elicits strong emotions, captured quite often with Relph's term "placelessness" (1976). The term indicates that mobility, tourism, commerce, and other elements of modernity have eroded place-based authenticity. When a place attracts substantial levels of attention from outsiders, for example by catering to tourists, processes are set in motion that rework the landscape to reflect the outsiders' expectations. Relph outlines a range of relationships to place, from existential outsiders at one extreme, in which places are seen in the homogenizing gaze of a visitor who is just passing through, to existential insideness at the other extreme, which entails that "a place is experienced without deliberate and selfconscious reflection yet is full with significances" (1976, 55). Recent work challenges the notion of "authenticity" along with "tradition" and "heritage," since these ideas are often linked to oppressive, exclusionary, and reactionary politics. The sense that place is under siege by globalizing forces that are undermining local distinctiveness can be attributed, in large part, to a more specific concern with Americanization. What passes for "placeless" may in fact be particular place associations that are perceived as alien and invasive because they are redolent of a single country's
iconography, that of the United States. The debate about authenticity has much to do with notions of community; the close ties between locale and sense of place carry normative weight and place images provide neat packages for how things “should be.”

Doreen Massey challenges the assumption that increased mobility and communication erode the foundations of sense of place. Defining place in a way that emphasizes particularity but not separation, her “global sense of place” is one in which places are defined by connections rather than contents, links rather than boundaries, and the people in a place are situated differently with regard to flows of social power. She challenges the assumption that place depends on spatial isolation and stasis, as well as the Marxist concept of place as resistant. Massey argues that “what gives a place its specificity is not some long internalized history but the fact that it is constructed out of a particular constellation of social relations, meeting and weaving together at a particular locus” (1993, 66). By concentrating on places as “moments in networks of social relations and understandings” (Massey 1993, 66), she is able to tell a story about “power-geometry” in which some people control mobility and communication, some move and communicate quite a bit but do not control the processes favoring mobility and communication, and some merely suffer the consequences of living in an increasingly interconnected world.

New directions

Attention across academia to the works of place-sensitive nongeographers such as Michel Foucault, David Morley, Manuel Castells, Saskia Sassen, John Urry, Mimi Sheller, and Judith Butler, among others, indicates a broad “turn” toward place. This turn is also evident in philosophical works like Edward Casey’s Getting Back into Place (1993) and The Fate of Place (2013). There have been particularly earnest turns toward place-based research in anthropology, sociology, and communication theory, as well as in environmental psychology. Scholars in these fields, as well as in architecture, design, and planning, have drawn inspiration from geographical writings of the 1970s and 1980s, including works by Tuan and Relph, as well as the concept of a “place ballet” (Seamon and Nordin 1980). Such place-based social science forays are proliferating, whether signaled by the local as opposed to the global, the symbolic power of home, the economic power of world cities, the consumption of places, or the Deleuzian interest in “topography” as something one lives through rather than merely on.

In geography, the emergent nonrepresentational approach or style of geography revitalizes many long-standing interests of humanistic and phenomenological geography, seeking to understand aspects of place experience. Body language, cooking, dance, everyday life, haunting, play, sports, tool use, and touch are among the topics of interest (Thrift 2008). Embodied passions, emotions, practices, and encounters are now linked to “affective realms,” “encounters,” “events,” “performances,” and “waves of influence” so the details of place experience are in the foreground more than ever. This interest in embodiment, instead of in static patterns and hierarchies, directs attention back to place, suggesting yet another round of conceptual revitalization for this term.

Technological change has also changed how people (including scholars) think of place. The diffusion of the Internet has led to considerable fascination in popular culture with novel ideas of place such as cyberspace and virtuality. In
the mid-1980s, the science fiction novel *Neuromancer* described cyberspace as a “consensual hallucination” consisting of graphic representations of digital information: “Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding” (Gibson 1984, 69). While the novelty of computer networking has faded considerably and the reality of the Internet has turned out to be somewhat more prosaic, Gibson’s basic insight regarding place is still valid: digital media have altered the relationships between location, locale, and sense of place. The emergence in the early 1990s of graphic user interfaces and online service providers inspired many variations on a theme: for example, online “chat rooms” and “dungeons,” “going online,” “posting” things on one’s “wall,” “pinning” images on one’s “board,” and so on. Such place-based terminology is employed casually but it reveals collective recognition that online experience is related to place: digital code supports interactions in an array of locales, each with its own characteristic sense of place, based on hardware and software rooted in particular locations. All three elements of place are present in and around online interaction, even when bricks and mortar have been left behind.

Geographers are beginning to engage with the idea of virtual place. New technologies call for familiar themes to be revisited, including accessibility, centrality, marginality, privacy, and surveillance. This new research challenge does not prevent critiques of the notion of virtual place or efforts to identify problematic social dynamics in virtual places, but it can potentially reinvigorate earlier discussions of authenticity and placelessness, or “senses of home that are spatially extensible” (Ralph and Staeheli 2011, 520). Insights about how people make a place or even home in complex media like the Internet will help shine light on older transformations in and of place, including those arising from earlier broadcast media, and prior to that the emergence of writing, printing, and literacy. As computing capacity and speed constantly increase, places have become sites for the extraction of “big data.” It is impossible to enter or leave many places without being detected and identified, for example through RFID (radio frequency identification) tags, facial recognition software, and biometric sensors. Where places once offered seclusion and privacy, they now increasingly serve as mechanisms for monitoring and surveillance, which in turn produces a demand for alternative means of carving out privacy, seclusion, and containment of place within the transparent space of digital flows.

**SEE ALSO**: Areal differentiation (or chorology); Big data; Home; Homelessness; Human geography; Orientalism/Occidentalism; Privacy, personal privacy; Regional geography; Representation: indoor spaces; Scale; Space; Spatiality; Territory and territoriality; Time–space convergence; Toponymy

**References**


PLACE


Further reading

Vegetation influences hydrological fluxes between the atmosphere and the ground surface. The canopy of individual plants acts as a barrier, preventing a proportion of precipitation from reaching the surface and contributing to runoff or groundwater. Precipitation is intercepted by components of the canopy (stems, branches, flowering parts, and foliage), and much of the intercepted water evaporates directly back into the atmosphere (Levia and Frost 2006). Interception ($I$) is calculated as $I = P_g - (T + S)$, where $P_g$ is gross precipitation, $T$ is throughfall, and $S$ is stemflow. All of the variables in this equation are expressed in units of depth equivalent, such as centimeters. Gross precipitation is the quantity of precipitation that falls on the top of the vegetation before it is in contact with the canopy or passes through the canopy. Throughfall is the quantity of precipitation that reaches the ground minus the stemflow. Two pathways of throughfall exist. The first pathway is precipitation falling directly through the canopy during a rainfall event without being intercepted. This quantity of precipitation does not contact vegetative surfaces and falls through canopy gaps. The second pathway is precipitation that drips from the canopy after precipitation splashes on vegetative surfaces. In this pathway, water on vegetative surfaces can coalesce to form large water droplets. Stemflow is the precipitation that reaches the ground after channeling down the branches to the stem of a plant. Because this hydrologic pathway occurs along vegetative surfaces, it is distinguished from the throughfall pathway in which drops of precipitation free-fall to the surface from the canopy.

The most common field method to measure interception is to use precipitation gauges, throughfall collectors, and stemflow collars. For accurate measures of interception, the instruments need to be in close proximity to each other. For this reason, precipitation is best measured from the top of a canopy tower projected above the height of the canopy. Because the construction of a canopy tower is cost prohibitive, the next best solution is to position precipitation gauges in a clearing or large canopy gap in close proximity to the throughfall collectors and stemflow collars. The two most common types of throughfall collectors are precipitation gauges and throughfall troughs positioned under the vegetation canopy. Vegetation canopies create a large variation in size and quantity of gaps. Throughfall collectors positioned under a large canopy gap will potentially measure more throughfall than a throughfall collector positioned under a large branch. As such, the cumulative orifice of throughfall collectors needs to create a large area to represent the heterogeneity of canopy gaps created by the vegetation. To create a large cumulative orifice for throughfall collectors, the number of precipitation gauges collecting throughfall is increased. Throughfall troughs may be better than small throughfall collectors because they may provide a larger area covering a longer swath of canopy gaps and obstructions. Regardless of the type of throughfall collector used, it is best to reposition the throughfall collectors underneath
the canopy on a regular basis to increase the representative area of the canopy. Stemflow collars are constructed of plastic tubing that is cut longitudinally and fastened to the trunk with sealant and small nails to collect the stemflow drainage. Stemflow drainage is channeled into an uncut section of the tubing and into a covered precipitation gauge. Stemflow is calculated as a depth equivalent per unit projected area of canopy. Published research commonly reports that stemflow is only about 1–5% of annual precipitation and is zero in small storms.

Another term related to interception is “canopy storage capacity.” Canopy storage capacity is the amount of water held on the canopy at the end of a rainfall event, after drip has ceased and in the absence of wind. The quantity of canopy storage capacity represents the minimum amount of water that can be stored in the canopy and not the total amount of water that was intercepted by the canopy during a precipitation event. Once the canopy has become saturated, the amount of interception exceeds the canopy storage capacity by whatever is evaporated during the rainfall event. Because the rainfall intensity normally exceeds the evaporation rate by at least an order of magnitude, the canopy will stay wet and evaporate during precipitation events. Direct measurements of canopy storage capacity are labor-intensive and often rely on destructively sampling vegetation. Aston (1979) suspended small trees under a rainfall simulator to demonstrate a mass-balance (gravimetric) approach to the measurement of canopy storage capacity.

Approximately 10–25% of annual precipitation is lost by interception, depending on meteorological and vegetation factors. The meteorological factors that influence interception include rainfall characteristics (e.g., rainfall intensity, precipitation duration, precipitation frequency) and the evaporative potential of the atmosphere. The length of time to achieve canopy storage capacity is reduced with higher rainfall intensity events. Additionally, canopy storage capacity is larger in many broad-leaved species during higher rainfall intensity storms than in lower rainfall intensity storms because of greater amounts of rainsplash on the underside of leaves. The duration of the rainfall event also influences interception. During short duration, low-intensity rainfall events, 100% of precipitation can be intercepted. The evaporative potential of water intercepted by the canopy is influenced by available energy, vapor pressure gradient, and the aerodynamic resistance of leaf surfaces. Interception loss is greater in environments with high temperature and low humidity than in environments with low temperature and high humidity.

Vegetation factors that influence interception, throughfall, and stemflow include leaf area index, stand density, vegetation age, and species composition. Leaf area index is the ratio of total leaf surface area of vegetation to ground area and is expressed in units of m²/m². A positive relationship occurs between leaf area index and interception. Vegetation with a larger leaf area index has a greater potential to intercept more water during a rainfall event because there are more surfaces for the water to adhere. Because leaf area index strongly influences interception, leaf phenology is also an important control on interception. Deciduous species during the growing season intercept a greater proportion of precipitation than deciduous species during the dormant season because of differences in leaf area index between the seasons. Stemflow may be enhanced during the dormant season in deciduous species. Although needles persist year-round for evergreen species in temperate and high latitude regions of the world, the leaf area index is less during the winter months than during the growth season. Small changes in leaf area index can produce significant changes in interception.
The presence of epiphytes in the canopy of cloud forests and temperate rainforests increases leaf area index and canopy storage capacity.

Vegetation age influences interception insofar as the leaf area index of younger forests is lower than the leaf area index of older forests. Stand density can influence the spatial distribution of throughfall beneath a canopy. For example, forests with low stand densities create larger canopy gaps between individuals. This allows more throughfall and reduces interception. Well-stocked stands have greater interception losses than understocked stands. The variety of spacing between individual trees produces large variation in quantity of throughfall to the forest floor.

The influence of species composition on interception is perhaps the most understudied aspect of plant–water interactions. Of the thousands of different plant species that occur on Earth, only a small percentage have been used to study interception, throughfall, stemflow, and canopy storage capacity. Besides leaf area index, the canopy of each plant species has unique properties that influence hydrological fluxes. These properties include the canopy architecture, leaf characteristics, and bark characteristics. Branch angle is a component of canopy architecture. In general, branches that are upright tend to channel intercepted water down to the stem and generate more stemflow than branches that are horizontal with the surface. Abrupt changes in the angles of branches in the canopy create drip points for stemflow that increase throughfall. The anatomical structure of the bark also varies between species. For species with flaky or rough bark, the surface area of the canopy is increased, which causes water to be easily retained by the canopy, increasing interception and canopy storage capacity. Species with smooth bark do not retain as much water during a rainfall event.

Variation in leaf characteristics (i.e., area, angle, shape, surface properties and chemistry) also influences interception and canopy storage capacity. Leaf angle or leaf orientation relative to the horizontal of the petiole and lamina of a leaf influences the ease with which water drains from the leaf surface. Several plant species have developed drip tips to channel water off the leaf surface. Leaf angle, leaf surface structure, and leaf surface chemistry affect leaf hydrophobicity and water droplet retention (Rosado and Holder 2013). Leaf hydrophobicity is a measure of how repellent a leaf surface is to a water droplet, and is commonly measured by calculating the contact angle between a sessile water droplet and the leaf surface. Water droplet retention is a measure of how easily a water droplet drains off a leaf surface, and is commonly measured as the angle at which a sessile water droplet will begin to drain off a leaf surface as the leaf surface is incrementally tilted. The significance of leaf characteristics such as leaf hydrophobicity and water droplet retention as a mechanism that influences hydrological processes is an underexplored research topic in ecohydrology. Several studies have found that the contact angles between a leaf surface and a water droplet range from 22° to 180°, depending on the species and the chemical properties of the leaves. Many leaf surfaces producing contact angles exceeding 150° may not be true surfaces, with the droplet instead held away from the surface by large numbers of projections of wax or trichomes. A contact angle exceeding 110° indicates leaf surfaces that are nonwettable, or repellent, and leaves with contact angles <110° are considered wettable, or nonrepellent. Although the contact angle between the leaf surface and the water droplet may be large, the ability of the water droplet to drain off the surface of the leaf (water droplet retention) may be small because of contact angle hysteresis. A
leaf surface has a low water droplet retention if a water droplet sheds from a leaf surface that is tilted <20° from horizontal. A leaf surface has a high water droplet retention if a water droplet remains on a leaf surface tilted >60°.

Leaf hydrophobicity and water droplet retention may be common adaptations among plant species in habitats exposed to daily precipitation during the summer growth period. The selective pressure for reducing the wettability of leaf surfaces is usually thought to be physiologically driven. Water droplets on leaf surfaces restrict photosynthetic carbon exchange because carbon dioxide diffuses approximately 10,000 times slower in water than in air, and the beading of water on leaf surfaces enables the maintenance of high photosynthetic rates. Older leaves are less hydrophobic and wet more evenly than newer leaves. Leaf hydrophobicity and water droplet retention may be important variables that influence canopy storage capacity during a rainfall event (Holder 2013). Several studies have addressed the influence of canopy architecture and precipitation characteristics on canopy storage capacity, but few studies have directly addressed how leaf hydrophobicity and water droplet retention influence canopy storage capacity. Species with highly repellent leaf surfaces may reduce canopy storage capacity, and therefore, result in a decrease in interception.

The hydrology of cloud forests and temperate rainforests differs from other environments because of the presence of fog precipitation. Although the term “fog precipitation” is not universally used, early writers in various locations suggest people have had knowledge of the process for many years. Fog precipitation is called horizontal precipitation, occult precipitation, fog drip, and fog stripping by different authors. Fog precipitation occurs when fog droplets pass through the canopy of the forest and are filtered by vegetative surfaces. The intercepted fog droplets coalesce on the vegetative surfaces and drip to the forest floor. The environments that are especially likely to have fog precipitation are high-elevation regions where cool temperatures result in the condensation of water vapor and coastal regions on the western side of continents where cool air off the oceans condenses and moves inland. Local factors influencing the quantity of fog precipitation include canopy height, canopy architecture, wind velocity, leaf surface chemistry and structure, hillslope orientation, and orientation of foliage and branches.

Fog can contribute 25% or more of total precipitation in localized cloud forests. Fog precipitation represented a greater proportion of hydrologic input in the cloud forest of Guatemala during the dry season (19%) than it did in the rainy season (<1%) (Holder 2004). Because the dry season has a significant quantity of fog precipitation, management of the cloud forest is particularly important in maintaining water resources for arid lowland communities that ration water during the dry season. Without the canopy, fog precipitation would not exist. Deforestation of the cloud forest will impact this hydrological input and may reduce water yield because of the decrease in fog precipitation. With changes in species composition and stand characteristics, the filtering of water droplets to produce fog precipitation may be directly affected and fog precipitation inputs will likely change. Very few details are known about plant–water interactions and the canopy storage capacities of the cloud forest species that are being cleared by humans.

In temperate, high-latitude, and high-elevation areas, snow can be the primary form of precipitation. Snow interception totals vary between species. Dense conifer forests intercept the largest amounts of snow and consequently cause the greatest loss from the canopy by sublimation.
Thin and leafless hardwoods or open lands intercept the least amount of snow. Snow accumulation and water equivalent are greater in the open than in the forest. The forest canopy, however, modifies microclimates at the forest floor by producing lower temperatures, less shortwave radiation, higher relative humidity, and slower wind speed. These microclimatic conditions reduce energy inputs to snowpacks under the canopy and make the rate of snowmelt slower beneath the canopy. Snow intercepted by plant surfaces that is exposed to direct shortwave radiation and greater wind speed sublimates at a faster rate than snow at the shaded forest floor. During the low-sun season, canopy aspect can influence the rate of sublimation by creating different canopy microclimates based on exposure to direct shortwave radiation.

Numerous studies have examined the influence of vegetation on rainfall interception and stream discharge from undeveloped experimental watersheds. Although the results presented in these studies expanded knowledge of plant–water interactions, experimental watersheds that examine the role of vegetation in hydrological fluxes do not often include urban landscapes within the catchment area. Because the majority of humans live in urban areas, urban ecology has emerged as a rapidly growing discipline. Urban ecohydrology is an important subdiscipline within urban ecology. Urban watersheds represent a complex mosaic of land uses, including buildings (residential, commercial, and industrial), infrastructure (roads, sidewalks, etc.), landscaped vegetation, agriculture, and undeveloped open space (forests, grasslands, etc.). Leaf and canopy characteristics between common species in urban watersheds vary significantly (Holder 2011). Information on how changes in vegetation (e.g., via landscaping, conversion of open spaces to housing developments, etc.) influence hydrological processes such as rainfall interception and discharge within watersheds is important for city and environmental planners as the population of citizens continues to increase and land cover continues to change. Additional studies on the interactions between the physical properties of canopy traits and water droplets on vegetative surfaces would be useful to develop enhanced models of ecohydrological processes.

SEE ALSO: Biomes; Ecosystem; Precipitation; Soil water; Tropical rainforest ecosystems; Water budget

References


PLANT–WATER INTERACTIONS

Further reading

Plate tectonics deals with the motions of the Earth’s rocky shell, or lithosphere, which cause the gradual shift of continents and are also responsible for mountain building, volcanic islands, and thermal vents. These geological processes have influenced the distribution of biodiversity. Prior to the identification of a geophysical mechanism that could drive movement at the Earth’s surface, its effects had been predicted. The concept of “continental drift” is usually attributed to Alfred Wegener, who presented his ideas on the origin of continents and oceans in 1915. However, in the 1500s the Belgian mapmaker Abraham Ortelius had noted that the coincidence of the coastal outlines of the Americas, Africa, and Europe implied former juxtaposition. In the 1800s, the “father of biogeography,” Alfred Russell Wallace, developed his ideas with the benefit of geographer Charles Lyell’s Principles of Geography. Lyell had noted that “Continents … shift their positions entirely in the course of ages.” Nevertheless, in the 1900s continental drift was strongly opposed by many, and the rejection of this hypothesis, which was built on extensive evidence, represents a notable failure of twentieth-century scientific method. The claim that consideration of the theory required demonstration of a geophysical mechanism was unjustified, but was nevertheless assuaged in the 1950s and 1960s when a tectonic model was presented. Today, paleomagnetism, radiometric dating, stratigraphy, and spatial information allow detailed interpretation and paleogeographic reconstruction.

One type of evidence presented in support of continental drift by Wegener and others was (paleo)biogeographic: the widespread distribution of certain fossil animals and plants across landscapes that are now disconnected (Figure 1). However, although widely reiterated even today, the occurrence of fossils did not in truth inform on the process of continental isolation. An alternative idea (land bridging) resolves the spatial discontinuity of biota problem equally well and makes the same biological assumptions. As with continental drift, land bridging assumes that the organisms of interest were unlikely to disperse between patches of suitable habitat. Thus both hypotheses require habitat continuity if species (or their lineages) are to be shared among areas. In biogeography, distribution patterns are often taken as evidence of past processes but they provide only the basis for formulating alternative hypotheses about the processes that might explain them. This is a subtle but vital distinction that underpins scientific method; a proposition cannot be simultaneously proposed and tested using the same observations.

In later editions of his book, Wegener recognized that his continental drift hypothesis was not to be tested with fossils, or for that matter a coincidence of continental crust outline. Rather, the best approach to testing the hypothesis was geodetics (Earth measurement), because the hypothesis that continents had moved in the past generated the prediction that continents would still be moving. Wegener died in his tent during a trip to Greenland to confirm longitudinal drift,
PLATE TECTONICS IN BIOGEOGRAPHY

Figure 1  Distribution of four Permian and Triassic fossil groups used as biogeographic evidence for continental drift, and land bridging. *Mesosaurus* included a small number of meter-long marine reptile species recorded in early Permian time (299–280 Ma). It is said that *Mesosaurus* could not have crossed the Atlantic but this is speculative, and at least one popular web resource incorrectly states that these were freshwater reptiles. *Lystrosaurus* included a debated number of species of pig-like land reptiles less than 1 m long. The genus is unusual in being present on either side of the Permian–Triassic extinction boundary (255–241 Ma). *Cynognathus* (*C. crateronotus*) was a meter-long predatory mammal-like reptile usually treated as one species that existed in early to mid-Triassic time (247–237 Ma). *Glossopteris* was a group of woody gymnosperm plants with wide but mostly southern (i.e., Gondwanan) distribution in the Permian (290–252 Ma). The identification of *Glossopteris* from northern areas has been questioned, and estimates of diversity vary.

Fossil sites were mapped using the database and tools of the Fossilworks Paleobiology Database, Macquarie University (John Alroy). Paleogeographic reconstruction is for Permian–Triassic time (250 mya). The inset on left is a version of a popular iconographic but misleading depiction of the fossil distributions (The Snider-Pellegrini Wegener fossil map; WikiCommons). Permian Map by C.R. Scotese, PALEOMAP Project. Reproduced by permission of C.R. Scotese.

but geodetic data have since demonstrated the fact that continents move. Plate tectonics explains how. Geological (rather than biogeographic) data informs paleogeographic reconstruction, and in doing so provides the context for biogeography. Plate tectonics is not, however, only about drifting continents; it generates short- and long-term environmental changes that influence the distribution and evolution of biological diversity, in many different ways.

Continental drift vicariance

Recognition of the potential for continental-scale effects of plate tectonics on biology developed
PLATE TECTONICS IN BIOGEOGRAPHY

side by side with plate tectonic theory. The founding concept of continental drift was that most of the Earth’s land had formerly been connected in a single supercontinent (Pangea). Fossils of Permian age indicate that Pangea existed more than 250 mya, and modern evidence supports this (see Figure 1). In fact the key fossils occur primarily across the southern part of this continent, a region usually referred to as Gondwana (or Gondwanaland), and it is noteworthy that this term was also used by land-bridgers who also envisaged a former extensive continent, parts of which subsequently disappeared. Although the geophysical evidence for a supercontinent is very clear, we now know that it did not include all continental areas at any one time. The idea that past breakup of a supercontinent (up to 200 million years later) could result in the establishment of biotas that are visible today makes many assumptions, among them that a supercontinent would have a homogeneous environment with continuous distributions of plants and animals, that other global events did not cause local changes in biotic assemblages, and that dispersal between continents had little influence on biological assemblages. In fact, paleoecological evidence indicates diverse and sometimes extreme environmental conditions existed across Pangea. Continental drift itself resulted in changes in climatic conditions so that natural selection on the biota would change, an extreme case being the wholesale extinction of Antarctic biota due to the drift of the continent over the South Pole. Mass extinctions across the globe radically altered biotas and dispersal between continents influenced regional biotic composition through time.

Intriguingly, although vicariance biogeography associated with continental drift is a highly attractive and popular idea, contrary evidence of the power of biota to disperse and colonize across habitat discontinuities is abundant and well understood (e.g., colonization and speciation on oceanic islands such as the Galapagos). In studies of living biota, evidence for plate tectonic vicariance can come from dated phylogenetic analysis (see Phylogeography and landscape genetics). Appropriate sampling and fossil calibration (see Figure 2) allow the timing of evolutionary events and continental drift to be compared. A match between the estimated time of origins of regional biota and the time of continental separation is consistent with, but not proof of, vicariance.

Land bridges and sea barriers

Plate tectonics can result in the connection of formerly disjunct land areas and simultaneously the sundering of marine environments, through volcanics, accretion, orogenics, and deformation. Although land bridging has been largely rejected from consideration of ancient biological history of Earth, younger examples show its influence. The Central American (or Panama) isthmus between North America and South America is a narrow strip of land (60–177 m wide) that finally shut off the equatorial link between Atlantic and Pacific oceans (the Central American seaway) in the late Pliocene (about 3 million years ago). Several tectonic plates intersect in the region (Figure 3) and their interaction appears to have led to the formation of numerous volcanic islands starting at about 5.5 mya. Sediment accumulation between these islands is one mechanism proposed for the land bridge.

Closure forced the Gulf Stream to carry warm equatorial water into the northern Atlantic Ocean, where it influences the composition and ecology of regional marine biota. Separation was accompanied by other environmental change; upwelling on the Pacific side see ms to have increased while salinity has increased in the Caribbean. Around Central America,
Figure 2  Dated phylogenetic analysis of four Madagascan mammal lineages using multiple nuclear DNA gene sequences and fossil calibrations. Origins and diversification of the Madagascan mammal fauna substantially postdate the pre-Cretaceous plate tectonic separation of Madagascar (red) from Africa (yellow) (≈83 mya). Very different dates for each of the mammal groups suggest no single mechanism for their arrival in Madagascar is supported. Tectonic crustal movements may have resulted in some parts of the Mozambique Channel being raised above sea level in the upper Eocene. Although there is no evidence for a continuous land bridge, islands might have facilitated mammals colonizing Madagascar. However, this speculation is not supported by the current phylogenetic analysis. Red branches lead to Madagascan species, yellow branches to nearest African relatives, grey branches represent outgroup taxa. Green and blue bars at nodes indicate 95% credibility intervals (bounds of uncertainty) for Madagascan stem and crown group, respectively. Tertiary time periods are Paleocene (Pa), Eocene (Eo), Oligocene (Ol), Miocene (Mi), and Pliocene (Pl). Adapted from Poux et al. 2005, by permission of Oxford University Press. Animal photographs from WikiCommons.
populations of marine creatures (and even seabirds that are reluctant to fly over land) were split between western Atlantic and eastern Pacific communities. On land, the isthmus led to changes in rainfall and climate patterns and facilitated the exchange of land creatures (a.k.a. the Great American Biotic Interchange). Many animals are inferred to have moved between continents and contributed to biotic mixing and thus elevated biodiversity in Central America.

The late Pliocene formation of the Panama Isthmus demonstrates a way that plate tectonics influences biogeography, which is counter to the effects of continental drift vicariance. This geologically, ecologically, and phylogenetically well-studied phenomenon is precisely the type of land bridge proposed by Charles Schuchert and others to explain exchange of biota between continents in past times. Their ideas replaced earlier, simplistic land bridge models that required special, unknown, processes to destroy large parts of continents. Systems similar to the Panama Isthmus might be geologically short-lived and thus hard to detect after the passage of time. Clearly, however, small and short-lived geophysical phenomena can have a profound influence on biogeography.

Intriguingly, while many mammals appear to have walked between continents once the Panama Isthmus formed, molecular and fossil data indicate that habitat connectivity is not essential for exchange. For example, the best explanation for the initial presence in South America of several mammal lineages involves oversea dispersal from Africa. This includes some that were subsequent North American colonists (e.g., porcupine). The marsupials, for instance, are represented by just one species in North America, about 100 in South America, and about 230 in Australia. However, the oldest fossils (65 mya) are from North America. Expansion southwards through South America and Antarctica probably took marsupials to Australia and eventually Borneo and Sulawesi, while the group disappeared from North America. Genetic analysis suggests that all living marsupials have their ancestors in South America. This makes sense because the only living marsupial in North America (opossum) arrived over the Panama Isthmus. Little is known about how these types of animal interacted when they met, and consequently ad hoc interpretations are common. For example, the extinction of placental mammals in Australia is often attributed to the arrival of marsupials, yet the reverse (replacement of marsupials by placentals) is suggested for North America. Very little is known about the marsupial fauna that must have existed in Antarctica, and none of course are there today. Modern species distributions are therefore revealed to be poor indicators, in many cases, of biogeographic history.

Other land bridges have resulted not from tectonic activity but from sea level change. Most recently, during the Pleistocene (<2.6 mya),...
Figure 3
lowered sea level resulting from water being locked into polar ice caps created land connection in many parts of the world. Habitable continents were linked as recently as 13,000 years ago through features such as the Bering land bridge between Siberia and Alaska that enabled the migration of many plants and animals, including humans.

**Mountains**

Linear mountain ranges form when continental plates collide, and it is routinely suggested that the formation of mountain ranges is a likely driver of allopatric speciation in terrestrial ecosystems. Populations sundered by geophysical barriers such as mountain ranges could evolve independently of one another if gene flow between them was sufficiently impeded. Mountain formation provides an appealing scenario for terrestrial vicariance because it generates abrupt environmental discontinuity (similar to land bridges dividing marine environments). Abrupt changes in environmental conditions are expected to influence population density, range sizes, and natural selection, but proven examples of the emergence of mountains causing speciation through vicariance are few. Cases where related species occupy habitat either side of mountains provide the expected distribution patterns but might also have resulted from occupation of habitat patches after mountains formed. They might also be the remnants of species distributions that predate mountain formation.

Mountain building contributes to biogeography in another way. The steep environmental gradients and habitat mosaics that mountain ranges produce appear to have stimulated adaptive radiations in many taxa. Examples include alpine buttercups in New Zealand, paper daisies in South Africa, and Andean hummingbirds in South America. Species that evolve to use alpine conditions in one location may also be successful on other ranges where similar ecological opportunities exist even though these are island-like, and this can result in disjunct distributions that are suggestive of vicariance.

Numerous studies in Europe and elsewhere show that plants and animals responded to Pleistocene climate cycling by range shifting, and this is contrary to the idea that mountains (and other geophysical features) routinely make biological barriers. The distribution across Europe of genetic variation within species and species groups shows that only in some cases do some mountain ranges correspond with species limits. In these cases, of course, the mountains are already present but the climate zones and species move. Distinguishing between mountains (and similar features) as drivers of lineage splitting and their longer-term influence on environmental gradients and ecosystem processes is challenging and exposes a general problem for biogeography. Mountains may come to be correlated with the limits of species and assemblage ranges or they might provide convenient landscape markers for biogeographers, even if they were not the mechanism of population sundering in the first place. Instead of looking for absolute interactions between geology and biology that focus on allopatric processes of lineage splitting, modern biogeography now recognizes the ways gene flow, environmental gradients, environmental fluctuations (e.g., climate change), extinction, sympatry, and uneven selection pressures across species ranges influence biotic composition.

**Deep sea hydrothermal vents**

Mineral-rich hot water emerges at temperatures of 400°C or more in submarine conditions where tectonic plates meet. The heat and chemical resources fuel specialized faunas
that constitute the only major biological systems that do not draw their energy from the sun. Secondary production from Archaea and chemosynthetic bacteria supports faunas that include giant worms, shrimps, clams, and limpets, at depths of 1–4 km below the sea’s surface. The habitat of hydrothermal vent faunas is dictated almost entirely by tectonic processes that result in a patchy distribution around the globe. These islands have histories tens of millions of years long but their biological isolation appears to be influenced by their chemistry and the ecology of their fauna. Gene flow between invertebrate populations living around widely spaced geothermal vents is mediated by planktonic larvae. Plate tectonic activity may be implicated in the origins of life in the ancient seas of Earth as it has been proposed that the thermal, chemical (e.g., acidity), and physical (e.g., high pressure) conditions around hydrothermal vents could have supplied the environment suitable for initial emergence of replicating organic molecules and protocells. There is evidence for hydrothermal vents in the Earth’s oceans in the appropriate time frame more than 3 billion years ago.

Volcanoes

Oceanic volcanoes create virgin land that is physically separate from existing, inhabited land areas. Although separated from other pieces of land, islands are not isolated. Island biogeographers recognize that the fauna and flora of islands develops through the interaction of colonization, speciation, and extinction and these components are in some cases quantifiable. There is some evidence that the rate of colonization is influenced by the size and distance of islands from potential biological sources, but increasingly it is apparent that establishment of migrant species is strongly influenced by the composition of the biota that has already arrived.

Classically, Charles Darwin’s observations of life on the Galapagos archipelago were instrumental in the founding of evolutionary theory (Figure 3). Volcanic island systems of this sort provide natural laboratories where the effects of colonization, natural selection, and population isolation can be observed. Because volcanic rocks are readily aged using radiometric analysis, maximum ages for the local biota can be determined.

Volcanoes both destroy life and create opportunities, as has been demonstrated on a local, recent scale in systems such as Krakatau in Indonesia. On a much more profound scale, volcanic activity is implicated in one or more ancient global mass extinctions. Extreme volcanic activity associated with the Central Atlantic Magmatic Province is strongly indicated as driving end-Triassic (201 mya) extinctions. Flood basalt eruptions at the end of the Permian (252 mya) are likely to have contributed to extinction via increased atmospheric carbon dioxide, climate warming, and associated ice melt methane production. The end of the Permian was marked by the extinction of about 70% of land animal species and 90% of marine animals. Among the casualties were the formerly widespread Glossopertid plants, the fossils of which were so influential in paleogeography. The Triassic saw the emergence of new ecological strategies in mammal and dinosaur diversification. Ammonites emerged in the oceans, and crocodiles, pterosaurs, frogs, and sphenodonts diversified on land.

Conclusion

On large and small scales, short and long time frames, plate tectonics influences the formation and distribution of biological diversity; however,
many of the details of how this happens remain to be thoroughly tested. A traditional focus on spatial effects (vicariance) is gradually being replaced by more in-depth analysis of the way geophysical attributes of the planet influence evolutionary ecology.

**SEE ALSO:** Biodiversity; Biogeography: history; Mountain biogeography; Ocean biogeography; Zoogeography

**Reference**


**Further reading**


Point pattern analysis

Seth E. Spielman
University of Colorado at Boulder, USA

Point pattern analysis is a set of methods for statistical inference about the pattern of points on a map. A point pattern is a set of observed coordinates, typically the digital analog of a paper map with locations indicated using push-pins. Point pattern analysis is typically used to determine if points are clustered relative to some reference distribution or spatial process. For example, Austin et al. (2005) examined the spatial distribution of fast food restaurants relative to the spatial distribution of schools and found that fast food restaurants “cluster” near schools. Intuitively, a cluster is simply a group of geographic events/objects; however, this naive definition belies some of the complexities and subtleties of point pattern analysis. The identification of clusters, or any patterns of points on a map, depends on some a priori assumptions about the expected number of points at any given area of a map; a cluster of points is a statistically significant deviation from that expectation. The following entry defines key terms, describes the analytical methods used to test for patterns in maps of points, and discusses the future prospects for point pattern analysis.

Point pattern analysis consists of two sets of methods: those that examine the absolute number of points within regions of the map (e.g., quadrant analysis) and those that examine the distributions of points relative to each other by counting the number points within some distance \( (h) \) of each point on the map (e.g., Ripley’s K-function). The example in the previous paragraph is an instance of the latter; it examined the number of fast food restaurants around schools, not the absolute number of fast food restaurants in a given area of the map. Methods which characterize the distribution of points relative to each other, work through the estimation of a function describing the expected number of points as a function of distance. This distance function is used to perform statistical inference on the observed spatial pattern of points on a map.

In order to make any statistical statement about the pattern of events on a map, the location of the events has to be compared to some reference distribution. Consider Figure 1, a map of all deaths due to car accidents in the State of Colorado during 2012. These deaths clearly appear clustered along a line in the center of the state. This line represents the eastern edge of the Rocky Mountains and the Front Range metropolitan corridor extending from the city of Colorado Springs in the south to Fort Collins in the north, inclusive of the Denver metropolitan area. The pattern of vehicular deaths may merely reflect the pattern of traffic: there are many more cars in the Front Range than there are in the mountainous interior of the state, and thus there are more deaths.

Statistical inference about point patterns depends upon some assumptions about the expected pattern of points on a map; these assumed patterns are called null models or point process models and play a central role in point pattern analysis. These reference distributions...
form the null hypothesis against which a map of points is compared. In spite of the relative simplicity of points as a mode of representation, point pattern analysis is a complex endeavor; key terms in point pattern analysis, such as cluster and pattern, can only be defined relative to these reference distributions.

The simplest null model is called complete spatial randomness (CSR) or a Poisson point process. Under this null model it is assumed that points are equally likely in all parts of a study area. In this model, the number of expected events in a region of the map is a function of the size of the region and an intensity parameter (denoted with the Greek letter \( \lambda \)). This intensity parameter is the mean/variance of a Poisson distribution; under CSR the same Poisson distribution applies to all locations on the map. CSR is not really an appropriate assumption for Figure 1 because car accidents are constrained to roads, and their frequency is related to the amount and speed of traffic, thus not all locations within the State of Colorado have an equal probability of a car crash occurring. Within a fixed study area, a higher-intensity point process will result in more points than a lower-intensity point process. With this definition of a null model, a “cluster” is an area of the map that contains more points than would be expected under the Poisson distribution describing the null model. For example, if we had a 1 km\(^2\) region of the map, and expected five vehicular deaths per square kilometer (\( \lambda = 5 \)), we would expect to find five points within the square kilometer. However, we live in a random world and will never get exactly what is expected; sometimes there will be slightly less than five events in the square, sometimes more. The Poisson distribution could be used to determine the probability of observing a given number of points; given \( \lambda = 5 \) it would be exceptionally rare to observe 13 or more points in the square kilometer – by random chance this would only occur 1 in 1000 times. The excess number of points relative to our expectation can be understood as a “cluster.” However, if we expected 10 points per square kilometer, 13 or more points would happen 50% of the time due to random chance. The key point is that the null model determines the definition of a cluster.

Point pattern analysis is rapidly gaining relevance as mobile devices facilitate the large-scale collection of point datasets. However, not all geolocated information can be appropriately analyzed with point pattern analysis techniques. Point pattern analysis operates in a specific conceptual framework that views points as events for which an expected probability of occurrence at any location within the study area can be estimated. For example, the locations of restaurants, as in Austin et al. (2005), is constrained by municipal regulations and macro-economic trends, making it difficult to view their locations as events governed by some probability distribution.
Defining a point

From a geographic perspective, a point is a set of coordinates designating a location. At a minimum, a point is described by two numbers, an x and a y coordinate. These two coordinates allow a point on a plane to be located; in the discipline of geography, this plane is usually some section of the surface of the Earth.

The ubiquity of global positioning systems makes it relatively easy to collect point data. However, not all point data are well suited to point pattern analysis. Typically, in point pattern analysis, a point is defined as an event. Events are things that occur with some nonzero probability at all locations in a study area. For example, the location of a lightning strike could be conceptualized as an event, but the location of a lake would be difficult to see as an event. Viewing points as events has certain implications; most importantly, it implies that at any location within the study area, the probability of an event occurring is known. In practice, this probability is seldom known and has to be estimated.

Events occur at a location in space but also at a specific time. The concept of a point can be extended into the temporal dimension, such that events can occur at a discrete location in space and during a specific instant of time. This extends the notation for points to include x, y, and t. Historically, points were often collected on a static map; events occurred at a location but little was known about the instant of their occurrence. In a modern computing environment where data are collected by mobile devices, it is common for events to be tagged with a time stamp. Furthermore, these same mobile devices can collect elevation data. While not commonly used in applied analysis, points often contain a vertical dimension and are not necessarily constrained to the surface of the Earth. For example, a map of property crime in a large city might include the story (floor) of the building on which the crime took place. Thus, points can contain a vertical dimension (z), extending the notation for points to include four coordinates: x, y, z, and t.

In point pattern analysis, it is assumed that all events within the study region have been captured. A study region is defined spatially or as a spatiotemporally bounded region. Point pattern analysis generally does not deal with samples of events; instead, the assumption is that events have been completely measured and mapped within the study area. This complicates analysis of some types of events— for example, some small-scale property crimes may not be reported to the police, or a disease may be undiagnosed or misdiagnosed; unless these omissions were purely random in space and time, they would make it difficult to analyze the locations of such events using point pattern analysis methods.

Defining a cluster

Often the goal of point pattern analysis is to identify the presence (or absence) of clusters of events in space and/or time. Point pattern analysis is sometimes called “cluster analysis,” which is an understandable but unfortunate label because of potential confusion with a large literature of the same name in pattern recognition and statistics. The object of point pattern analysis is the identification of “clusters” of events in space and/or time; it differs from “cluster analysis” in the use of a null model for statistical inference. Spatial point pattern analysis explicitly tests null hypotheses about the arrangement of events whereas nonspatial cluster analysis aims to identify latent structure in a dataset; the two sets of methods should not be confused. On
the other hand, ontological problems arise in both literatures because the meaning of the term “cluster” is sometimes vague.

In point pattern analysis a map is said to exhibit clustering if an arbitrarily drawn region of the map contains more points than would be expected under some assumed data-generating process (i.e., a null model). This is an odd idea – how many points should be expected in some arbitrarily drawn region of a map? A toy example may make this concept easier to understand. Imagine a landscape filled with magic coins that flip themselves when one says “flip” and magically land in the exact same spot. Assume that the coins are evenly spaced 1 m apart and fill the landscape. When one says “flip,” a map of heads emerges (also magically) in your computer; the map shows you the locations of the heads but not the tails. Each time you say “flip” there will probably be a different outcome and a different map. All the magic coins are fair and equal, so heads/tails are equally likely in all places. If you selected a 4 × 4 m square region of the landscape, you would expect to find 8 heads because there are 16 coins, half of which would land heads-up. Of course, we live in a random world and might not get exactly 8 heads; sometimes you might get 9 or 10 heads but this would not be too surprising. However, if you got 14 heads it would be unusual indeed. You might say that the 4 × 4 m square study area with 14 heads and 2 tails constituted a “cluster” of events (points), where an event is the act of a coin landing heads-up when tossed. Conversely, it would be equally surprising if you found 2 heads; this would be an instance of dispersion or anticlustering. The opposite of a cluster is either the absence of events or the uniform, nonrandom patterning of events; it is a substantively interesting finding that can be empirically identified via point pattern analysis methods.

In the above example the null model was a random flip of a fair coin. What if the coins were unfair and landed heads 90% of the time? Fourteen heads out of 16 coin flips would not constitute a cluster in this case because it would be in line with expectations. A cluster can only be defined relative to some hypothesis about the probability of events (points) occurring within some arbitrarily defined area.

Scale in point pattern analysis

All point pattern analyses are constrained to a bounded region of space (or space–time) within which data are collected. This bounded region, or window, describes the full extent of the area for which events were monitored. Events can, and probably do, occur outside of the observed area, and this must be accounted for in statistical procedures. For example, if the State of Colorado was the analysis window, the motor vehicle related fatalities in Figure 1 are clearly clustered in the major metropolitan areas of the state. However, if the window shrunk to a section of Denver, and only events within the red box in Figure 1 were observed, the accidents might not appear to be clustered. The window within which events are collected has an important effect on statistical inference. Patterns and the interpretation of patterns are scale–dependent.

Scale dependence in point pattern analysis raises important questions about the design of studies. Since statistical inference is scale–dependent, the analyst must choose an analysis window that has some substantive experimental/policy relevance. If possible, given the sensitivity of inference to the analysis window it is sometimes advisable to ensure that results are not affected by a random shift of the analysis window.
Point process models

Point pattern analysis conceptualizes points as events. Statistical analysis of a map of events requires estimating the probability of an event occurring at a given location. In the simplest and most widely used form of point pattern analysis this probability is assumed to be uniform across the study area; that is, the probability of an event occurring is equal at all locations within the study area. This assumption of constant intensity yields a relatively simple point process model CSR.

A full understanding of this idea, of a constant intensity, requires some discussion. The idea that the probability of an event occurring is equal at all locations in a study area implies that the locations of events are independent from each other. That is, the location of a lightning strike, or a car crash, does not affect the location of other events. The chances of lightning striking the same location twice are the same as the chance of lightning striking any two separate locations; the same may not be true of car crashes. CSR means that the location of an event does not affect the location of other events. This assumption is entirely reasonable for certain types of phenomena, but it is not reasonable for others.

Thus there are two general ways to specify a null hypothesis for point pattern analysis: the process perspective and the probability perspective. In the process perspective the null hypothesis takes the form of a point-generating process. For example, the patterns of seedling locations for plants with low seed mobility are constrained by the locations of their parent plant. If seeds cannot travel long distances, seedlings have to be located near their parent. This type of process is called a Neyman–Scott point process. A Neyman–Scott point process is different from a Poisson point process because the probability of an event occurring at a given location is related to how close the location is to a “parent” and the amount of dispersal from a parent. By contrast, in the probability perspective spatially varying covariates might be used to estimate a surface describing the probability of events occurring at all locations within a study area. This type of approach is also called an inhomogeneous model because it contrasts with CSR, which assumes that the probability of events occurring does not vary within the study window (i.e., lambda is homogeneous within the analysis window). Viewing intensity as a function of covariates allows the creation of a point-generating process based on the relationship between events and covariates. For example, in Figure 1, if the volume of cars is associated with vehicular mortality, traffic volume could be used in the construction of a null model.

Test statistics

In order to evaluate the pattern of points observed on a map it is necessary to describe the pattern on the map quantitatively. This is done through the construction of statistics that counts points as a function of distance \( h \). Counting points as a function of distance allows the pattern of points on a map using a graph to be statistically described. For example, consider a simple test statistic, called the “nearest neighbor” function (or the G-function; equation 1).

\[
G(h) = \frac{1}{n} \sum_{i=1}^{n} (I(d_{ij} < h))
\]

\( n \) is the number of points on the map, \( d_{ij} \) is the distance between the \( i \)-th point and its nearest neighbor, and \( I(d_{ij} < h) \) is an indicator function that takes the value 1 if the expression \( d_{ij} < h \) is true, and zero otherwise. \( G \) is simply the cumulative distribution function of the
nearest neighbor distances. The test statistic is bounded by \([0,1]\); however, the rate at which the curve approaches 1 will be determined by the spatial distribution of points. If points tend to be separated by a fixed distance, say \(h = 3\, \text{km}\), \(G\) will rapidly approach 1 around that distance. If, however, interpoint distances are random, \(G\) will monotonically approach 1.

The obvious problem with \(G\) is that it neglects to fully exploit the information contained within the map of points. Only the distances between nearest neighbors are measured. A fuller picture of the distribution of points could be obtained if a test statistic was constructed that considered all interpoint distances. Ripley’s \(K\)-function measures the number of points within \(h\) units of each point. Rather than simply counting the number of nearest neighbors, pairs were taken into account. The \(K\)-function, because of its treatment of interpoint distances and the ability to modify null hypotheses, is one of the (if not the) most widely used test statistics for point pattern analysis.

Cressie (1993) writes the \(K\)-function as in equation 2.

\[
\hat{K}(h) = \frac{1}{\text{intensity}} \mathbb{E}(\text{number of points within } h \text{ of an arbitrary point})
\]  

Expressing the \(K\)-function in this way makes it fairly intuitive. Think of a circle of radius \(h\) centered on each point. As the area of the circle increases the number of points contained by the circle will grow. The \(K\)-function measures the average number of points around each point on the map as a function of distance. In this formulation, “intensity” (or \(\lambda\)) is the number of points per unit area of the analysis window. This can be written as equation 3.

\[
\hat{K}(h) = \frac{A}{n^2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} I(d_{ij} < h)
\]  

\(n\) is the number of points on the map, \(A\) is the area of the analysis window (in square units), and \(d_{ij}\) is the distance between the \(i\)-th and \(j\)-th points. However, when points are near the edge of the analysis window, or when \(h\) is large, the circle used in the calculation of \(K(h)\) spills over beyond the boundary of the analysis window. Thus, an edge correction factor \((w)\) is often applied to \(K(h)\), yielding equation 4.

\[
\hat{K}(h) = \frac{A}{n^2} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} I(d_{ij} < h) w
\]

This correction factor is generally \(w_i = 1\) if the \(i\)-th point is more than \(h\) units away from the edge of the window. If the \(i\)-th point is close to the edge of the study region then \(w < 1\). Various software packages implement different (or several) edge correction strategies, with the main practical difference between them being computational efficiency.

One nice feature of the \(K\)-function is that if points are randomly distributed within the analysis window, \(K\) rises proportionally to the area of the circle described by \(h\). This makes

\[
L(h) = \sqrt{\frac{\hat{K}(h)}{\pi} - h}
\]

a useful transformation because the expected value for a random pattern will be nil and positive or negative values suggest spatial patterns. More details on statistical inference regarding the presence (or absence) of patterns is given in the next section.

There are a number of extensions of the \(K\)-function that allow for testing of dependencies among “marks” in point patterns. A map of points is said to be “marked” if points contain categorical information, for example if points represent different species of trees or different types of criminal events (property versus violent crime). These marked points modify the logic of a \(K\)-function; instead of considering all interpoint distances the distribution of one type of
event is studied relative to other types of events (equation 6).

\[
\hat{K}(h) = \frac{1}{\text{intensity}} E(\text{number of type } i \text{ points } h \text{ of an within } h \text{ of an type } j \text{ point})
\]  

For example, the interest might be in understanding if murders are co-located with other types of violent crimes. In this case, one type of mark would be “violent crime, not murder” and another type would be “murder.” Case-control data, as collected in experimental situations, are a special case of marked point pattern. Test statistics for marked point processes are similar to those for unmarked patterns (equation 7).

\[
\hat{K}(h) = \frac{A}{n^2} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{I(d_{ij} < h)}{w}
\]

Inference

Test statistics, like the K-, G-, or L-function, simply provide a numerical summary of the pattern of points on a map. Further work is necessary to understand if the pattern represents a statistically significant deviation from expectations. The methods used to make statistical inference about point patterns have evolved substantially since the mid-2000s even though the test statistics have remained basically unchanged during that period.

Inference about point patterns involves simulation of the null hypothesis. These simulations are used to establish a range of plausible point patterns conditional upon the null hypothesis being true. These simulated point patterns are summarized using a test statistic. The range of the test statistics computed from the simulated data is used to establish a range of expected values. For example, Figure 2 shows two plots, each containing a grey region; this region represents the range between the highest observed K-function under the null hypothesis and the lowest observed K-function under the null hypothesis. The red dashed line shows the mean of the simulations. The black line measures the actual map of points/events. It can be noticed that in the graph on the left the black line is entirely contained within the grey simulation envelope; this is considered evidence that the observed pattern is consistent with the null hypothesis. On the right, however, the black line is outside of the simulation envelope; this might be considered evidence that the observed pattern is not consistent with the null hypothesis.

However, recent work by Loosmore and Ford (2006) shows that using graphs like Figure 2 to evaluate the null hypothesis can be misleading; the use of simulation envelopes to visually evaluate point pattern statistics leads to an elevated rate of Type I error. That is, the null hypothesis is rejected more frequently than it should be. In some situations Loosmore and Ford (2006) found that the null hypothesis led to patterns that exceeded the simulation envelope 74% of the time; that is, random chance alone can, and often will, lead to an observed point pattern falling outside of the simulation envelope. If one used this chance occurrence as evidence of a statistically significant pattern one would be committing a Type I error. Thus, while it is common practice to evaluate point patterns visually using simulation envelopes, there is some reason to believe that this approach is not effective. An alternative set of tools for inference about point patterns has emerged in recent years. These newer modes of inference also exploit simulations of the null hypothesis, but they use simulation to calculate a single statistic by examining the variance within the simulated patterns.
Many of the simulations used to assess point pattern analysis are unconstrained in space. Thus, for networks (constrained patterns, like the car accidents in Figure 1) special tools and/or simulation models may be necessary. Software for the analysis of network and 3-D point patterns exists (see Okabe and Yamada 2001, and the SANET software (http://sanet.csis.u-tokyo.ac.jp)). In addition, it is possible to simulate “inhomogeneous” point patterns (see SpatStat R Library (http://spatstat.github.io)). These are useful when the intensity of events is a function of spatially varying covariates (like population density). These inhomogeneous intensity surfaces can be estimated at all locations within an analysis window using regression-like methods.

**Software**

Many geographical information system (GIS) packages have basic facilities for point pattern analysis. The National Institutes of Justice publishes CrimeStat, a free, stand-alone software package for point pattern analysis. Special purpose software, such as SANET for network-based point pattern analysis, exists. The most advanced and flexible tool for point pattern analysis is the free and open source Spatstat R Library, which is well documented and structured such that specialized null hypotheses are relatively easy to implement and simulate.

**Prospects for point pattern analysis**

Point pattern analysis is unique among spatial analysis methods in that it requires an analyst to explicitly develop a null hypothesis. This can be a challenge because much of the geographers’ toolkit is descriptive. Point pattern analysis methods can be used to comparatively describe patterns but are generally used to make statistical inferences. These inferences, because they require the use of simulation, force an engagement with geographical processes, that is, there has to be a meaningful null hypothesis for the results of an analysis to be interesting. In the absence of such a hypothesis, all comparisons are
made with CSR, which is naive for most social and geographic phenomena.

Points may be the geographic data type of the future. Area-level statistical data, such as those collected by national statistical agencies, are in decline as agency budgets shrink and response rates dwindle. Spatial “big data” is often point-level data. For example, the growing interest in Twitter among geographers is an example of using point-level data in a purely descriptive, non-inferential mode. These novel point-based data sources suggest point pattern analysis is an area ripe for innovation.

References


Further reading


Humans ascribe names to features on the Earth’s surface, and use those names in everyday discourse to obtain directions, to associate information with features, and for many other routine purposes. Recently a number of new applications of such names have emerged, in support of Internet-based services. An index of named features (place names) is needed by mapping services (“Where is San Diego?”), wayfinding services (“Find me a route from my home to the nearest Starbucks”), and information-retrieval services (“Where can I get a digital map of recent earthquake activity around California?”). When these needs were first recognized in the early 1990s two existing sources of such indices existed. Gazetteers are lists of officially recognized place names and their locations, compiled by mapping agencies such as the US Board on Geographic Names in the interests of standardizing spelling and avoiding duplication, and often providing an index to the named features shown in atlases or on topographic maps. Point-of-interest databases (POI databases) were originally compilations of businesses and their locations, assembled by companies such as Dunn and Bradstreet and sold to businesses seeking to evaluate potential sites. Both of these traditional databases included names, locations in geographic coordinates such as latitude and longitude, and limited additional information such as the type of feature that is named.

Traditionally, gazetteers have been limited to officially recognized place names, and to the types of features normally shown on comparatively coarse-scale maps: cities, lakes, mountains, rivers, and so on. Because of the economics of map-making such features tend to be permanent, or at least semipermanent. POI databases have been limited to businesses, prominent features that provide landmarks for wayfinding (churches, parks, etc.), and locations providing services (libraries, schools, etc.). Such features are more likely to change through time than gazetteer entries, and may also be too small to be shown on topographic maps. Today, however, these distinctions are rapidly becoming irrelevant as both types of database have the potential to support the same Internet-based services, and as the costs of acquiring and maintaining such data continue to decline. The discussion that follows avoids making any distinction, using the term “place-name database” (PND) throughout.

### Place and space

Geographers have long distinguished between concepts of place and space. Places are human constructions, associating characteristics such as esthetic qualities, safety, and wealth with named places, without necessarily giving those places precise geometric locations or limits. By contrast, geographic information technologies have traditionally taken a spatial perspective, using coordinates such as latitude and longitude to define, measure, and represent position on the Earth’s surface, computing properties such as distance
and direction, and using geometry to define topological relationships between features: adjacency or containment, for example. Measured locations have been used to link together disparate kinds of information, a primary function of geographic information systems (GIS). Place names traditionally played a minor role in these technologies, for example being omitted from the list of seven datasets identified in the 1990s as the framework layers of the National Spatial Data Infrastructure. Geographic information systems were seen as tools for performing spatial analysis based on coordinates, and little attention was paid to place names, their associations, or their study.

More recently, however, place names and the human dimensions of geographic information technologies have assumed much greater importance, and there is growing interest in a platial as distinct from spatial perspective. The average citizen is now engaged as a user of geographic information, and expects to be able to interact with it through an intuitive interface. Citizens should not have to take advanced courses in GIS to undertake such simple tasks as wayfinding; instead, systems should emulate the everyday language of human discourse, which means using place names rather than coordinates. Suppose, for example, that a citizen is searching for real estate in and near Niagara Falls, NY, and compare two equivalent responses: “Three-bedroom house at 43.08097N, 78.95198W” and “Three-bedroom house in Love Canal neighborhood.” The second is clearly much more useful, since it immediately conjures important (and largely negative) associations with the notoriously polluted Love Canal. In short, to people place names have associations, while coordinates largely do not. Place names can be used to link information in ways that may be more powerful than coordinates, and may be less subject to errors of measurement.

Place has also become much more important as citizens have become engaged not only in the consumption of geographic information, but also in its production. The phenomenon of volunteered geographic information (VGI), and crowdsourcing in general, has created a new source of readily available data on named features and for their updating, and companies providing Internet services rely heavily on such processes to keep their PNDs current. The term “neogeography” has been suggested as a catch-all term for this engagement of the average citizen in both the consumption and the production of geographic information, and in new kinds of mapping that address the day-to-day needs of citizens.

This is not to claim that place is now as important as space in GIS. Many of the less tangible associations of place, ably discussed by generations of human geographers, lie beyond the simplifications and formalizations of geographic information technologies. Instead, PNDs offer only a crude caricature of the concept of place, but one that is rapidly growing in importance and suitable for use in computational systems. A suite of interesting questions emerges. If machines reason about nearness by evaluating the spatial property of distance, how do humans traditionally reason about nearness, when they have no such ability to compute? Could it be, for example, that humans reason that San Francisco must be nearer to Seattle than is Mexico City because the first two lie within the higher-level place United States while the third lies in a different higher-level place, Mexico? In other words, is the place name hierarchy, which lacks any measured geometry, nevertheless the basis for such reasoning processes? Or similarly, do humans reason that San Diego must be west of Reno because the West Coast runs roughly north-south and San Diego is on the coast while Reno is not? In other words, is reasoning based on place enabled by caricatures of real feature geometry, such as a rectangular continental
United States, that are simple enough to be represented in human memory?

Space can be examined and represented at any level of detail, from global to very local scales. The principles of GIS can be applied equally readily to the analysis of patterns of global climate change and to the organization of furniture within a room. Similarly, place names exist at every scale, from the continents, oceans, and countries of the global scale to the street addresses, building names, and internal building layouts of the very local. Places form hierarchies: my house lies in the (vaguely defined) Bryant neighborhood of Seattle, which lies in the (formally defined) State of Washington, which is one of the United States, which is a country in the (again vaguely defined) continent of North America. Consistent with the possible vagueness of their geometry, places are not necessarily subject to precise rules of hierarchical containment: France is in Europe, but the nation of France includes islands in the South Pacific; and most people think of Disneyland Paris as part of Paris, though it lies outside most formal definitions of Paris, such as the boundary of the municipality or the Périphérique. Unlike the polygons of GIS, named places are likely not enclosed within well-defined boundaries. This inherent vagueness associated with places (and with many aspects of human communication) has often been used as an argument against attempting to deal with them in computational systems, and as an argument for the greater formality and potential precision of spatial perspectives.

Unlike spaces, places are social constructions that reflect cultural, linguistic, and other human agendas and contexts. The decision to name a feature is driven by a host of social priorities, as is the very identification of a feature as something worthy of naming. Different cultural and linguistic groups will attach different names to the same feature – the English Channel to the anglophone, and La Manche to the francophone. Strong emotions are sometimes associated with these differences, as in recent arguments over the use of the name Persian Gulf. Thus it can be important to identify both the name of a feature and also who calls it that.

With this background, the next section addresses issues in the construction of PNDs.

**Construction of PNDs**

**Names**

Most applications of PNDs begin by searching for a name through a list. Many web services assist the user with autocompletion, that is, by suggesting names that the user might be in the process of typing. Often these lists are compiled from previous searches by the user, and from knowledge of the user’s current position, obtained from his or her IP address, an Internet Protocol code that is automatically sent to a server by a user attempting to access a web service located on that server. By typing “3” into Google Maps, for example, I am immediately given five suggestions of street addresses beginning with 3, all of them in the Seattle area, since my IP address identifies me as likely located in Seattle. Typing the full house number, 3334, results in the suggestion “3334 NE Blakeley St, Seattle,” my full home address, based on my having frequently used that address in requests to Google Maps in the past. Typing the same sequence into Google Maps using the same computer, but located now in the Netherlands with a different IP address, results in an entirely different sequence of suggestions.

Large features often have names that are globally unique – there is only one North America, for example. But the names of smaller features
may be repeated many times, and there are hundreds of millions of features named “The kitchen” in the United States alone. Humans resolve such ambiguities in a number of ways. Context provides a frequently successful strategy: spoken in a house, “The kitchen” will be assumed to refer to the kitchen in the same house, while “London” spoken in London, Ontario, Canada, will be assumed to refer to that London, and not the much larger and better-known London in England. Autocompletion achieves a similar result: typing the highly ambiguous “Downtown” into Google Maps on my computer in Seattle results in several suggestions, the first being downtown Seattle.

Consider the name “Springfield.” In the United States, 32 states have recognized places named Springfield, according to Wikipedia. Wisconsin has a Springfield in no fewer than five counties. Context suggests that the Springfield most likely being sought by the user is the one closest to the user’s location; but some Springfields are more important than others. Springfield, IL, is the only state capital, while Springfield, MA, is the most populated; neither is the closest to Seattle. Google ranks Springfield, IL, as most likely when I query the name without attempting to resolve ambiguity by specifying the state.

Human behavior differs significantly from the actions of such web services in one important respect, however. While web services attempt to resolve ambiguity through context and by offering suggestions, humans often resolve ambiguity through extended discourse. Thus “London” may be followed by a response: “Which London do you mean?”

Many features have multiple names, as we have already seen. Transliteration of names is a frequent source of problems between language groups that use different alphabets – Beijing has been rendered as Peking in the past, and still is in the case of Peking University, though Peking’s two Chinese characters have not changed for centuries. Names change through time, and successive names may bear no linguistic or phonetic relationship to each other. Hastings (2008) has documented six names that have been used at various times and by various groups for the feature commonly known today as Lake Tahoe. All of these issues contribute to the practical difficulties of matching a user’s request to a list of names.

Vast amounts of digital text are being created in the form of emails, blogs, tweets, and other user-generated content, and the conversion of audio records such as phone conversations to text is now a well-developed technology. Optical character recognition allows text to be extracted from images, such as photographs of maps. Recognition of place names in such data is of interest in many fields, including intelligence. Thus there has been significant research in recent years in what is sometimes termed “geoparsing,” the automated recognition of place names in text. Metacarta, for example, is a powerful and widely used set of geoparsing tools. The task is made complex and difficult because of the vagaries of spoken and written syntax, the complications of different languages and alphabets, and the importance of context. For example, a reference to “Shanghai” in text could be interpreted as a reference to the Chinese city or to kidnapping. A reference to “Boston, New York” could be a reference to two large cities or to a single small town in upstate New York near Buffalo. Another challenging problem occurs with place names that may exist in text but not in any PND, that is, how one can determine that a sequence of letters is in fact a place name, and, if it is, its likely location. Syntax can provide clues, as in the use of certain location-related prepositions such as “near,” as well as direction-related phrases such as “north of.”
Location

The power of PNDs stems from their ability to link names to locations, and to marshal the functions of geographic information technologies to address user needs, whether they be in query, wayfinding, information retrieval, or a range of more sophisticated decision-making. The historical focus of PNDs has been on providing a single point location, the result of least effort, irrespective of the physical extent of the feature. Because they dealt with small features, this was a reasonable choice for POI databases. For gazetteers that were built as indices to atlases it was sufficient to locate a feature to a page in the atlas, and perhaps to a part of the page, but again for all but the largest features a point was sufficient to direct the reader to the appropriate point in the atlas. In other cases, conventions were adopted: the point location for a large feature was sometimes given as the location of the center of the feature’s label; the location of a mountain is conventionally its summit; and in the case of rivers, location was sometimes given as a point at the mouth. Today’s PNDs still often reflect these conventions and simplifications. For example, the US Geological Survey offers a digital gazetteer known as the Geographic Names Information System. A query for “Mississippi River” will return a point in Plaquemines Parish, LA (at the river’s mouth), despite the 3734 km length of the feature. A query to Google Maps for driving directions from “Colorado” to “Wyoming” will return a complex itinerary, using roads that are sometimes seasonally closed, from an uninhabited point in the geographic center of Colorado (CO) to a similarly uninhabited point in the geographic center of Wyoming (WY), despite the fact that the two states share a lengthy common border. Clearly this service could be greatly improved if the states were represented as areas rather than as points, and if the user were required to specify the point locations of the origin and destination.

Some well-defined features have complex geometries that are readily available in digital form, and it would be easy to provide a more intelligent response to the CO-WY problem. In other cases, however, it would be costly to capture the feature’s complex geometry, and in still other cases the feature lacks a formally defined and agreed geometry. There is no accepted geometry, for example, for the feature “downtown Seattle”; instead, services such as Google Maps use a single, representative point that may or may not be at the focus of the user’s interest.

Montello et al. (2003) interviewed a sample of subjects, asking them in various ways to identify “downtown Santa Barbara” on a map, and exploring alternative ways of representing the results in a spatial database. Recently, attempts have been made to use novel sources of information from which to impute feature geometries. Many collections of user-generated content include references both to place names and to geographic coordinates: this is true, for example, of many of the photographs stored in the Flickr database, and of some tweets. Figure 1 shows the result of searching the Flickr database for photographs that have been both georeferenced and tagged using the words “Eiffel Tower” in any of several languages. As with many such sources, the data include various kinds of noise: photographs taken from distant vantage points such as the Sacré Coeur basilica or Montparnasse; photographs that are georeferenced or tagged as a group, and may have been taken over a period of time at several locations; or cases where the georeferencing malfunctioned. It is possible also that some errors are malicious.

The result of such efforts is a continuous surface that can be interpreted in one of two
ways, depending on how the surface is constructed and normalized. First, the value $z$ at some location $\mathbf{x}$ is the probability that $\mathbf{x}$ lies within the feature; and second, the probability of the feature lying in some area $A$ is the integral of $z$ over the area $A$. The first interpretation makes $z$ a probability surface, constrained to the range $\{0,1\}$, while the second makes $z$ a probability density function whose integral over the entire plane is 1. Surfaces can be represented in spatial databases using any of the standard methods for representing continuous fields, including rasters and digitized contours. It is also possible to extract a suitable isoline from the surface as one possible representation of the extent of the named feature, though there is no guarantee that the extent will be convex or singly bounded.

Recent years have seen a rapid growth in the variety and volume of such data sources, consistent with what is often termed “big data.” It seems that in future, if not already today,
there will always be more than one source of the geometry of a named feature, so the integration or fusion of such sources to obtain or synthesize a single, best estimate is becoming a significant challenge. The general problem of conflating PNDs is discussed at greater length below.

Another challenging problem arises in dealing with museum and other collections of artifacts, if, as is often the case, the location from which an object was collected is given in text rather than coordinates. Today, of course, it is trivial and routine to record GPS coordinates whenever an observation or collection is made in the field. But verbal descriptions are often the only information available about earlier locations, especially when such locations were visited before an area was mapped in detail. Verbal descriptions often refer to prominent features, with estimated directions and offsets (“five miles northwest of the river junction”). GEOlocate is a widely used tool for converting text descriptions to coordinates, though its ability to deal with offsets is limited. Liu et al. (2009) have developed methods for capturing both the most likely location and also uncertainty bounds, when offsets are used. It is important in all such cases to capture both the most likely location and also uncertainty bounds, when offsets are used. It is important in all such cases to capture both the most likely location and also the methods used to impute it, along with associated uncertainties, since these can be of great value later in trying to conflate data from several sources.

Type

The third element in the traditional gazetteer is the type of feature. Many national and subnational mapping agencies have developed classification schemes, and these have been widely used to categorize features in gazetteers. For POI databases, however, the typing of features is traditionally more nuanced. The type “grocery store” is important for assessing competition, but it also encompasses a range of distinct subtypes, from the full-featured supermarket to the corner convenience store. Today’s web services support multiple types: for example, coffee shops often offer a range of baked goods and thus may call themselves bakeries, while many supermarkets also include both bakeries and coffee shops.

The principle of spatial heterogeneity holds that the geographic world is fundamentally heterogeneous, exhibiting uncontrolled variation. Left to its own devices, therefore, it follows that any administration whose responsibility covers only a fraction of the Earth’s surface will adopt a classification scheme that best fits its own purposes, based on the variation found within its jurisdiction. A mapping agency for Saudi Arabia will develop different types from one responsible for Alaska, for example, and will include terms for ephemeral streams or oases that have little relevance in Alaska, where periglacial features such as pingos may be more important. Similarly, and at a more detailed scale, a predominantly urban county in the United States will likely adopt a classification scheme for land use that is different from that adopted by a rural, and perhaps neighboring, county. Such differences are likely exacerbated by cultural variation: as Mark and Turk (2003) have shown in their work on what they term “ethnophysiography,” the landscape features that are important to one culture may be of no importance whatever to another, co-located culture.

Many well-meaning efforts have been made over the years to find universal typing schemes. In 1913 an international agreement was signed to create an International Map of the World at a scale of 1:1 000 000 – the “Millionth Map.” Some 2500 sheets were planned, using a uniform standard of content and cartographic rendering,
but only a fraction were produced before the project was finally abandoned. More recently, a Japanese-led International Steering Committee for Global Mapping was established in 1996 to achieve a similar objective in the Internet era. In 2011 the schema.org project was established by a consortium of major web service providers, as a universal markup that could be used to type objects, including geographic features. Although it has universal aspirations, the types recognized clearly reveal the scheme’s origins in the dominant consumer-oriented culture of the United States and its manifestation in Silicon Valley.

Several research projects have been dedicated to the development of methods to match such universal feature-type taxonomies to locally derived typing schemes. But matches are rarely simple and 1:1, and it is clear that any universal scheme must be much coarser than local schemes. Debates continue over whether the benefits of a universal scheme outweigh the costs of its inevitable local suboptimality.

Typing presents additional issues in the volunteer world of crowdsourced PND content. In principle, it is desirable that such typing schemes emerge from the community, rather than being imposed from above as in the traditional gazetteer. The term “folksonomy” has been coined to describe such bottom-up approaches to taxonomy. But many types are inherently vague and may even carry emotive and political associations. Lake Tahoe, for example, can legitimately be classified as a reservoir rather than a lake, since its level is controlled by a dam. In the case of the VGI project OpenStreetMap, differences between American and British English have led to separate types “harbor” and “harbour,” and “tag wars” have erupted between contributors who are unwilling to accept each other’s assigned tag or type (Mooney and Corcoran 2012).

In summary, it seems that any attempt to build a comprehensive PND in today’s environment must accept that a feature can have multiple types, and that each type should be identified with the community or culture for which that type is appropriate. This would allow a user to see different “views” of a PND depending on his or her preferred perspective. This perspective might be permanent, as in the case of cultural differences, or might be transitory, as in the case of a user searching for a bakery.

Hierarchy
Reference has already been made to the importance of hierarchy in human reasoning about places. Links between places and the higher-level places of which humans consider them to be part is important in resolving ambiguity: for example, identifying the containing US state immediately resolves ambiguity over which Springfield is meant, while “my kitchen” immediately identifies the specific kitchen referenced in a conversation by implicitly linking it to the higher-level place “my house.” We have also seen that hierarchy can be a useful basis for reasoning about relative proximity, based on the principle that two places that are linked to the same higher-level object are likely nearer to each other than to a third place that is linked to a different higher-level object. Other heuristics for relative proximity are discussed below.

Hierarchy is not a rigid or well-organized structure, despite the clarity of the hierarchy of administrative units that exists in most countries. Places can span administrative boundaries (Kansas City extends to two states), and vernacular places such as “the tri-counties” or “the Midwest” are frequently used in human discourse. A lower-level place is frequently part of more than one higher-level place, and while administrative hierarchies may formalize the
concept of a level (county, state, nation), in other
cases the concept of a level in the hierarchy can
be vague. It seems that two types of relationships
can exist within a place hierarchy: those between
features at adjacent levels (“is in,” “is part of”),
and those between features at the same level (“is
at the same level”).

Applications of PNDs

Reference has already been made to various
applications of PNDs, even in their simplified
form as today’s gazetteers or POI databases.
Place names and street addresses are used to
drive a host of online mapping and wayfinding
services, typically by converting user-supplied
names to coordinates and then using those coor-
dinates as an interface to simple GIS functions.
Mention has also been made of geoparsing, the
identification of place name references in text.

Other more sophisticated applications con-
tinue to emerge. The term “mashup” refers to
the linkage of information, in this case through
the occurrence of common place names, to
create new information. Many new kinds of
maps have been created by mashing independent
sources, and such operations are now enabled
by highly sophisticated tools. For example, it is
now possible to take the information in an Excel
spreadsheet and to use a place name in each
record to link the data to a digital map, creating
a thematic map of any of the variables in the
spreadsheet.

In another example, the textual information
in travel blogs has been mined for references to
places, and for terms that can be interpreted as
associations of those places. Adams and McKen-
ze (2013) have analyzed such data, finding
groupings of terms that co-occur frequently,
and the places associated with each grouping,
and producing a series of compelling maps of
the concepts associated with geographic places
and regions (Figure 2). It is clear that techniques
such as these will permit the acquisition and
synthesis of much richer data about places and
their associations.

As PNDs proliferate, and as new sources of
data on named places become available, the
problem of conflation, or the merging of PND
records about the same place, is becoming more
and more important. If a place can have multiple
names and types, if its location measured in
a geographic coordinate system is subject to
uncertainty, and if its geometric representation
can range from a point to a line to an area, the
question arises of how it is possible to know if
two records actually refer to the same feature.
Conflation can have several useful purposes:
combining the geometric representations of
a feature can produce a better representation;
and one dataset may have better geometry,
but another may have richer associations, so
their combination will be more valuable and
useful. Conflation is one form of synthesis or
mashup, and is likely to become more and more
important as data sources proliferate.

Hastings (2008) has compared the three
attributes of a traditional gazetteer – name, loca-
tion, and type – as keys to successful conflation.
Of the three, he finds that location is the most
important guide to whether two records refer to
the same feature, because locations will always
have some degree of similarity however they
are represented. Type is also useful, because the
alternative types assigned to a feature are likely
to have some degree of similarity. But the names
assigned to a feature can bear no relationship
whatever to each other. In his example of Lake
Tahoe, a point representation will almost cer-
tainly lie within a polygon representation, and
both “lake” and “reservoir” are types associated
with water bodies. But there is no similarity in
the alternative names given to the feature in the
past, some of them by different language groups.
Figure 2  Place associations of the word cluster “wine, bottle, glass, …” based on analysis of travel blogs. Courtesy of Ben Adams.

SEE ALSO: Ethnophysiography; Geographic information system; Mapping mashups; Place; Spatial data infrastructures; Volunteered geographic information

References


Further reading


Points in polygons

Yu Liu
Peking University, China

Definition

There are three possibilities for the topological relationship between a point and a polygon: the point can be inside, outside, or on the polygon’s boundary. A point-in-polygon (PIP) algorithm returns one of the three relations, or a binary value indicating whether the point is located inside the polygon.

Algorithms

Nordbeck and Rystedt (1967) summarized a number of PIP algorithms, in which the ray-casting algorithm is widely used. Later Huang and Shih (1997) and Žalik and Kolingerova (2001) also made a summary of PIP algorithms.

Ray-casting algorithm

The ray-casting algorithm indicates whether a point is located inside a polygon according to the number of intersections between a ray starting from the test point and the polygon. In general, an odd number of intersections indicates that the point is located inside the polygon. Hence, this algorithm is also called the even/odd rule algorithm or crossing-number algorithm. The implementation of the algorithm is rather straightforward. It can be simplified by comparing the \( y \)-coordinate of the ray and edges of the test polygon. It is suitable for both convex and concave polygons. The time complexity of the algorithm is \( O(N) \), where \( N \) denotes the number of polygon edges. If the intersection point coincides with a vertex of the polygon, then the intersection counts only when the second vertex of the edge is below the ray. As shown in Figure 1, edges \( a, b, \) and \( c \) are considered but the intersections with edges \( d \) and \( e \) do not count.

Winding number algorithm

The winding number algorithm is designed to compute the polygon’s winding number for a test point. The winding number is the number of times that the polygon’s boundary travels counterclockwise around the point (Figure 2).

If the winding number is non-zero, then the point is located inside the polygon. The axis-crossing algorithm can be applied to determine the winding number (Alciatore and Miranda 1995). Let the test point be at the origin. Determine whether edges of the polygon cross the positive \( x \)-axis and in which direction. Then count the winding number according to the principles shown in Figure 3. This algorithm is applicable to both convex and concave polygons. For simple polygons (Figure 4a), the ray-casting algorithm and the winding number algorithm provide consistent results. For complex polygons, however, the two algorithms may yield different results. As shown in Figure 4b, point \( a \) is outside the polygon according to the ray-casting algorithm. But its winding number is not zero. Hormann and Agathos (2001) pointed out that
Points in Polygons

Figure 1 Ray-casting algorithm.

Figure 2 Winding number algorithm.

the criterion can be changed from a non-zero winding number to an odd winding number to get the same result from the winding number algorithm as the ray-casting algorithm.

Sum of angles algorithm

For this algorithm the polygon is divided into \(N\) triangles by linking the test point with node-pairs of the polygon edge in sequence. If the sum of angles is \(360^\circ\), the point is located inside the polygon. This algorithm is applicable to both vertex and concave polygons. It is very slow to calculate angles using inverse trigonometric functions, the angles have rounding errors, and there is accumulation of these errors.

Sum of areas algorithm

In this algorithm the triangles are formed in the same way as in the sum of angles algorithm. The test point is located inside the polygon if the sum of areas of triangles is equal to the area of the polygon. Using the determinant formula to calculate the area of triangles is very easy and the result does not contain round-off errors. This algorithm is not applicable to concave polygons. The determinant formula is shown in equation 1.

\[
2T_n = \{P_1, P_2, P_3, \ldots, P_{n-1}, P_n\}
\]
\[
= \begin{vmatrix} x_1 & x_2 & x_3 & \cdots & x_{n-1} & x_n \\ y_1 & y_2 & y_3 & \cdots & y_{n-1} & y_n \end{vmatrix}
\]
\[
= (x_1y_2 + x_2y_3 + x_3y_4 + \cdots + x_{n-1}y_n)
- (y_1x_2 + y_2x_3 + y_3x_4 + \cdots + y_{n-1}x_n)
\]  

where \(T_n\) is the area of the polygon. Note that the area should be the absolute value.

Swath algorithm (Salomon 1978)

For this algorithm the polygon is first divided into a set of horizontal swaths. Then the swath containing the test point is selected. Finally, the ray-casting algorithm is applied to the edges inside the swath. This algorithm is efficient for multi-point queries because only the edges inside the swath are selected to count the number of intersections, instead of all the edges. Even though the preprocessing procedures take a little time, the overall efficiency is high, especially for multi-point queries.
**Points in Polygons**

**Figure 3** Winding number update criterion. Source: Modified from Alciatore and Miranda (1995). $V_i$, vertices of the polygon; $W$, winding number.

**Figure 4** (a) Simple polygon; (b) complex polygon.

**Wedge algorithm (Preparata and Shamos 1988)**

Here the polygon is first divided into a set of wedges using a random point inside the polygon. Then, similar to the swath algorithm, the wedge containing the test point is selected. Finally the point is tested to see whether it is located inside the polygon, based on whether it is on the same side of each triangle edge. Three methods can be applied to determine the position of the test point relative to the triangle edges:
POINTS IN POLYGONS

1. the sign of $ax + by + c$ can reflect the side of the point relative to the edge when the edge equation is $ax + by + c = 0$; 
2. let $p_1, p_2, \ldots, p_n$ be the vertices of a polygon and $q$ be the test point. The sign is the cross product between vector $a = (p_{i+1} - p_i)$ and $b = (q - p_i)$; 
3. the third-order determinants to calculate the area of a triangle can be applied to determine the sign. If the sign is the same for all edges, the test point is located inside the triangle.

This algorithm is only applicable to convex polygons. Similar to the swath algorithm, the wedge algorithm is suitable for multi-point queries.

All these algorithms are vector-based. Raster-based PIP is rather straightforward (Huang and Shih 1997). Huang and Shih (1997) compared the complexities of eight algorithms and found that for practical applications the ray intersection algorithm and the swath algorithm are suitable.

Applications

Alciatore and Miranda (1995) summarized the application fields of PIP, such as determining whether a point is within the reachable workspace of a robot. In geographical information systems (GIS), PIP can be used in a wide range of applications. Point queries are a typical example. Point queries can be divided into two parts: single-point and multi-point queries. Single-point queries determine whether a single point is located inside a polygon. For example, when selecting one from a set of polygons, we click a point inside the polygon. In this procedure, the PIP algorithm is applied to determine which polygon the point is located inside. For multi-point queries, one needs to determine whether many points are inside a polygon or which polygon they are inside. When points with associated attribute data are available, comprehensive and summarized information can be obtained for administrative or arbitrary regions (Nordbeck and Rystedt 1967), such as the average real-estate price of neighboring blocks, or the number of schools in a county.

Points-in-polygon overlays are another application in GIS. The points-in-polygon overlay assigns the attributes of the polygon to coincident points, creating a new point layer. The new point attributes come from the polygon in which the point falls. The geographic data on polygons are not incorporated into the point attributes. In practice, we might have questions such as which category of land use a point of interest falls into (Figure 5). Each point of interest would be added to the land-use attribute of the polygon it falls into using the PIP algorithm.

Implementation in spatial databases

Determining the spatial relationship between two geometries in structured query language (SQL) is implemented based on the nine-intersection model (9IM; Egenhofer and Franzosa 1991) and the dimensionally extended 9IM (DE-9IM). The two possible relations between points and polygons, Within and Contains, are defined in equation 2 (Open GIS Consortium 1999).

\[
\begin{align*}
\text{a.} & \text{Within}(b) \iff (a \cap b = a) \land (I(a) \cap I(b) \neq \emptyset) \\
\text{a.} & \text{Contains}(b) \iff b.\text{Within}(a) \\
\text{a.} & \text{Contains}(b) \iff (a \cap b = b) \land (I(a) \cap I(b) \neq \emptyset)
\end{align*}
\]

where $I(x)$ denotes the interior of a geometry $x$. In terms of the PIP problem, both the Within and
the Contains relations can determine whether the point is located inside the polygon.

The 9-IM and DE-9IM are used in many spatial databases like PostGIS, Oracle Spatial, and ArcSDE. In PostGIS, the Within and Contain relation functions are ST_Within (PostGIS 2012c) and ST_Contains (PostGIS 2012a). The function ST_Relate (PostGIS 2012b) outputs the standard Open GIS Consortium’s DE-9IM string code. In Oracle Spatial, the Within and Contain relation functions are SDO_INSIDE (Oracle 2009a) and SDO_CONTAINS (Oracle 2009b). The spatial predicates can be used to build SQL statements for point query and spatial join operations.

**Extension of conventional PIP**

In practice, the position of a point or a polygon may contain uncertainties because of the random errors produced in the procedure of data capture and processing. Thus it seems irrational to determine definitely whether a point is located inside a polygon. Cheung, Shi, and Zhou (2004) proposed a probability-based uncertainty model, which calculates a probability index to describe the closeness between an uncertain point and an uncertain polygon. The probability density function (PDF) of a point $P$ is applied to denote the probability that any point is the true location of $P$. The uncertainty of a polygon $A$ is determined by the PDF of the vertices of $A$.

When $A$ is certain, the probability $Pr$ that $P$ is located inside $A$ is given by:

$$Pr(P \in A) = \int \int_{(x,y) \in A} f_P(x,y) dx dy$$

(3)

where $f_P(x,y)$ is the PDF of $P$. Whether the point is located inside the polygon is determined by the probability index $Pr$, equal to the volume of the bell-shaped figure bounded by $A$.

When $A$ is uncertain, the equation becomes the conditional probability $Pr(P \in A|A)$. The probability $Pr$ that $P$ is located inside $A$ is given by equation 4.

$$Pr(P \in A) = E(Pr(P \in A|A))$$

$$= \int \ldots \int h_A(x_1, y_1, \ldots, x_N, y_N)$$

$$\times Pr(P \in A) dx_1 dy_1, \ldots, dx_N dy_N$$

(4)

where $h_A(x_1, y_1, \ldots, x_N, y_N)$ is the PDF of all the vertices of $A$. To reduce the number of variables in equation 3, the problem is divided
POINTS IN POLYGONS

into different cases according to the intersection between the polygon and the error ellipse of the point. This model is feasible to solve the PIP problem when positions of points or polygons are uncertain.

SEE ALSO: Data structure, vector; Qualitative spatial and temporal representation and reasoning; Spatial database; Topological relations; Uncertainty

References


Poland: Instytut Geografii i Przestrzennego Zagospodarowania (IGiPZ PAN) (Institute of Geography and Spatial Organization (IGSO PAS))

Founded: 1953
Location of headquarters: Warsaw
Website: www.igipz.pan.pl/
Membership: 124 (as of 2014)
Director: Marek Degórski
Contact: igipzpan@twarda.pan.pl

Journals or major publication series

Geographia Polonica. www.geographiapolonica.pl
Prace Geograficzne (Geographical studies). http://www.igipz.pan.pl/geographical-studies.html

Current activities or projects

The institute carries out studies for public and private activities involving, among others:

• spatial development at the national, regional and local scales;
• evaluation of investment projects and developmental programs (including EU projects);
• transformation of urban agglomerations;
• transformation of Polish transport systems, analysis of traffic and demand for transport, and transport accessibility (http://www.igipz.pan.pl/accessibility/pl/);
• the geography of agriculture (e.g., land use, spatial differentiation of crop and animal production, rural population);
• assessment of environmental quality for the purposes of spatial planning on the basis of potential vegetation, sedimentation, and sediment transfer in lowland and mountainous fluvial systems;
• assessment of climatic and bioclimatic conditions in spas and holiday resorts; and
• specialist spatial analysis utilizing cartographic computer techniques as well as a wide spectrum of GIS methods.

Brief history

The Institute of Geography and Spatial Organization PAS, established in 1953, is an important Polish research center in the field of socioeconomic geography, physical geography, and spatial organization. The Central Library of Geography and Environmental Protection is part of the Institute and is one of the largest thematic library centers of this kind in the world. The IGSO PAS carries out scientific research studies and participates in numerous applied projects. Its active involvement in international scientific collaboration that, in recent years, has resulted in its participation in more than fifty scientific and practical programs and undertakings, including EU framework projects, as well as the ESPON scientific platform – the Institute’s most significant project. Furthermore, in the
last decade there have been approximately two hundred national research and practical projects, expert evaluations, and so on, carried out by the IGSO PAS. Our research staff’s annual publishing output amounts to a total of approximately four hundred published items (e.g., peer-reviewed articles, monographs, notes, etc.). The institute is also actively engaged in collaboration with Polish ministries such as the Ministry of Infrastructure and Development (contract agreement), the Ministry of Agriculture and Rural Development, as well as with local authorities, especially at the voivodeship (provincial/regional) level. In recent years it participated in, among others, preparation of the National Spatial Development Concept 2030 and was involved in activities associated with the Polish presidency of the EU council in 2011.

Submitted by Jan Peliwo
Poland: Polskie Towarzystwo Geograficzne (PTG) (Polish Geographical Society)

 Founded: 1918  
 Location of headquarters: Warsaw  
 Website: www.ptg.pan.pl  
 Membership: 1441 (as of April 2013)  
 President: Antoni Jackowski  
 Contact: ptg@uw.edu.pl

Description and purpose

The Polish Geographical Society (PTG) is one of the oldest geographical institutions in Poland. Among its members are teachers, academics, practitioners from a wide spectrum of professions, and students.

The purpose of the society focuses on the popularization of geographical knowledge, education, and gathering and representing geographers. The main activities of the PTG are research, organization of conferences and seminars, excursions, and publications. The society is the largest geographic conference organizer in Poland and publishes the largest number of books and magazines related to the discipline.

Journals or major publication series


Current activities or projects

The Polish Geographical Society represents the discipline as well as geographers on both national and international levels. The society supports periodic conferences and seminars organized by the different committees (e.g., on cultural landscape, hydrology, rural areas, etc.) and numerous publications (20–30 per year). It organizes an annual competition for the best master thesis of the year. An important activity of the society is the organization of the Geography Olympiad and participation (with a lot of success) in the International Geography Olympiad. In 2014, the Polish Geographical Society, as a member of the consortium of eight Polish geographical institutions, will organize the IGU Regional Conference which will be held in Krakow.

Brief history

The Polish Geographical Society was established in 1918. In addition to geographers, the founding members came from different professions and scientific disciplines, including economists, historians, sociologists, meteorologists, geologists, botanists, zoologists, and anthropologists. In 1934, the Society co-organized the Congress of the IGU in Warsaw. After a forced break caused by World War II, the society resumed its activity in 1945. In 1953, the existing structure of the PTG underpinned the establishment of the Institute of Geography and Spatial Organization of the Polish Academy of Sciences. The institute took over a large part of the society’s research activity.
After 1953, and throughout the period until 1989, the society was mainly involved in the promotion of geography and its teaching methodology. It published a scientific monthly aimed at the general public, *Poznaj Świat*, and also organized research expeditions.

The 1980s and the early 1990s saw a dramatic collapse, both in membership and in the financial base of the society. The PTG had to face the new economic reality in which state financing of all scientific societies and associations was terminated. In the late 1990s, the situation of the society considerably improved. In 2001, certain provisions of the statute were amended to adapt to the new socioeconomic reality. As of 2006, the PTG has operated as a public benefit organization.

Submitted by Jerzy Bansi
Polar climates

Roger G. Barry
University of Colorado Boulder, USA

Basic characteristics of the polar regions

The two polar regions exhibit distinct geographical differences. The Arctic comprises a sea ice-covered ocean that is surrounded by the landmasses of North America, Greenland, and Eurasia, but is well connected to the North Atlantic Ocean. The Antarctic is occupied by the massive Antarctic continent, which is 98% covered by a 3–4-km-thick ice sheet, surrounded by a circumpolar ocean with seasonal sea ice. The Greenland ice sheet, which rises to 3200 m, is much smaller and has only regional effects on the Northern Hemisphere’s climate. In climatic characteristics there are similarities. In both regions there is a six-month winter night and six-month summer day. Surface temperatures are well below freezing in the winter in the Arctic and year-round in most of Antarctica. Surface-based temperature inversions predominate in both polar regions, especially in the cold months. Both regions are overlain by a deep polar vortex in the troposphere around which circle the global westerly winds, extratropical cyclones, and mesocyclones, but their surface climates are in strong contrast due mainly to their geographical characteristics.

Following a long interval of mild climatic conditions, the climate cooled and the Antarctic ice sheet developed rapidly about 33 million years ago (Ma). It has essentially been present ever since, although its extent has fluctuated significantly, particularly in West Antarctica. The Greenland ice sheet probably formed about 5 Ma, but perennial sea ice may have covered the Arctic Ocean since 14 Ma. Its recent sharp decline in extent and thickness in late summer is unprecedented in the palaeoclimatic record.

Observational networks

The observational database in both polar regions is limited because the regions are remote and weather conditions are extreme, making the setting up of weather stations costly and difficult. The climate of the Arctic Ocean has been monitored through the North Pole (NP) Drifting Station Program operated by the Soviet Union with NP1 in 1937–1938, NP2 in 1950, and then continuously with one or two stations from 1954 to NP31 in 1991, installed on multiyear sea ice or ice islands. The program resumed through the Russian Federation in 2003 with NP32, and continues to operate with NP39 established in October 2011. The United States has operated several ice island stations, notably T-3 from 1952 to 1978 and ice station SHEBA (Surface Heat Budget of the Arctic) from October 1997 to October 1998. For coastal and island locations there were 65 Russian and 24 Western stations for a period that includes the early 1950s through 1990. Many arctic coastal stations in Eurasia were closed in the 1990s, whereas most of those in the North American Arctic continue to operate. A valuable compendium of arctic climatological data and charts spanning 1893–1995 is contained in Arctic Climatology Project (2000). Corresponding sea ice data are collected in
Polar Climates

an atlas of the Environmental Working Group (2000) and updated for 1933–2006 in Arctic and Antarctic Research Institute (2007). There are coastal weather stations around Greenland, Summit Station that was installed in 1980, and the Greenland Climate Network of automatic weather stations (AWS) around the ice sheet, that began in 1991 and expanded to 19 stations.

In Antarctica, 51 coastal and island stations were operating during the 1957–1958 International Geophysical Year (IGY). Currently there are 11 manned stations and over 60 AWS sites that were mainly installed in the 1980s and 1990s. A majority of these are in the Ross Sea sector and West Antarctica. Beginning about 2004, ARGO floats were deployed in the Southern Ocean and these now provide extensive real-time oceanographic data.

Satellite visible and infrared coverage began around 1970, primarily from National Oceanic and Atmospheric Administration (NOAA) polar orbiting satellites. Currently, there is a wide range of passive microwave, radar, and altimetric sensors that are operated mainly by the National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA). These provide a wealth of global data on numerous environmental variables, including sea ice and Greenland ice sheet melt.

Radiation regime

In winter in each polar region there is an absence of sunlight and the surface radiation balance is negative due to the outgoing infrared radiation from the snow and ice surfaces. The net radiation budget over the Arctic Ocean is $-80 \text{ MJ m}^{-2}$ in January. In July over Antarctica it is approximately $-50 \text{ MJ m}^{-2}$, as a result of the low surface temperatures that reduce the longwave emission. In each case there is a small downward transfer of sensible heat to the colder surface. Over the Arctic Ocean in summer there is initially strong surface reflection of incoming solar radiation by the snow-covered sea ice. The snow cover has an average depth of about 40 cm. Gradually, this snow melts and melt ponds form on the surface, increasing the absorption of solar radiation. The surface albedo (reflectivity) decreases from 0.8–0.9 in May to 0.4–0.5 in July and August. The net radiation budget at the surface of the Arctic Ocean in July is 50–60 MJ m$^{-2}$.

By comparison, over the tundra, net radiation amounts in mid-summer reach 1000–1300 MJ m$^{-2}$ as a result of the long days and large amounts of solar radiation absorbed at the surface, where the albedo is in the range 0.12–0.18.

The temperature structure in the lower troposphere in winter features a pronounced surface-based temperature inversion in both polar regions. At the South Pole the temperature at 500–600 m above the surface can be up to 23°C higher than those at the surface, and the inversion can extend up to 1000 m altitude. In the Arctic inversion depths average 350–800 m and intensities range from 5 to 14°C. Values are greatest in northeastern Siberia. Inversions are stronger and deeper in winter and autumn than in spring and summer. The inversion is caused by a combination of warm air advection in the atmosphere and strong radiative cooling at the surface, which is usually snow-covered.

Polar amplification

A notable feature of the high latitudes is the regional enhancement of global warming during the twentieth century. The global mean temperature increased by about 0.7°C over the twentieth century, but this increase was amplified several times in the Arctic and around the Antarctic Peninsula. Analysis for land stations north of
59°N shows a warming rate of 1.36°C per century for the period 1875–2008, almost twice as strong as the Northern Hemisphere trend (0.79°C per century).

The polar amplification has been more pronounced during the past three decades, particularly over the Arctic Ocean, and is especially apparent in autumn and winter. A major factor is the decrease in autumn sea ice extent that acts to strengthen the ice-albedo feedback process. The increased area of open water leads to greater absorption of incoming solar radiation and hence more heat transfer to the atmosphere, raising the air temperature. Increased low cloud as a result of the open water provides a further positive feedback and warm advection is another major factor. Amplification is observed in the Arctic throughout the lower and middle troposphere and is evident in all seasons, implying a significant role for heat flux convergence in the atmosphere.

Recent work by Francis and Vavrus (2012) suggests that arctic amplification is having an effect on middle latitude atmospheric circulation. They identify weaker westerly winds and increased Rossby wave amplitude, both of which contribute to a slower eastward progression of the upper tropospheric Rossby waves and may lead to more persistent weather patterns. The effects are most apparent in autumn and winter, linked to the reduction in sea ice area.

In the Antarctic, warming has been most rapid in the Antarctic Peninsula, where temperatures have risen by about 3°C since the 1950s. Sea ice in the Bellingshausen-Amundsen seas has declined over this interval and air masses from those seas affects the peninsula. Elsewhere in Antarctica, there have been only minor temperature changes.

The effects of polar warming are evident in the accelerating mass loss from the two ice sheets. Based on changes in ice sheet elevation measured by ICESat-2 for 2011–2012, Helm, Humbert, and Miller (2014) determine a loss of 96 ± 93 km³ a⁻¹ for the Antarctic and 353 ± 29 km³ a⁻¹ for Greenland.

**Atmospheric circulation**

In winter there is a ridge of high pressure at sea level over the Beaufort-Chukchi seas and low pressure over the eastern Arctic Ocean (Figure 1). The ridge links the persistent, eastern Siberian high pressure and highs over the Yukon. North Atlantic cyclones track northeastward from the east coast of the United States, across Iceland, into the Barents Sea, where they decay. Only in spring is there an arctic anticyclone situated over the western Arctic Ocean, north of the Canadian Arctic Archipelago. In summer, there is a weak mean low located over the central Arctic. This pattern is associated with an influx of lows generated over the Eurasian continent and cyclogenesis that is attributable to baroclinicity over the Arctic Ocean itself.

The arctic frontal zone in summer is located near the arctic coasts, especially over Eurasia, and its location is largely determined by the large thermal gradient between the bare land and the cold ocean. It involves maritime arctic air and continental polar air, although the latter is relatively moist at low levels due to the high proportion of small lakes and the often-waterlogged peat vegetation cover that supplies moisture to the air through evaporation.

On monthly timescales there is an Arctic Oscillation (AO) or Northern Annular Mode (NAM) between low pressure over the pole and high pressure in mid-latitudes (37–45°N). This pattern extends through the troposphere and is a major control of Northern Hemisphere climate, especially in Europe and North America. A positive index, with strong westerlies, tends to keep arctic air to the north, while a negative index
can lead to cold air intrusions into mid-latitudes. Its sign fluctuates irregularly and its time series shows no periodicities. The index was mostly positive from 1900 to 1930, then mainly negative to 1990, and subsequently mainly positive until 2010. Components of the AO at sea level are the North Atlantic Oscillation (NAO) between low pressure over Iceland and the Azores anticyclone and the North Pacific Oscillation (NPO) between the Aleutian low pressure
and the subtropical anticyclone in the eastern North Pacific. These are sometimes linked and sometimes independent of one another.

In the Northern Hemisphere troposphere there is a deep polar vortex accompanied by strong westerly winds. In winter, there are two deep wave troughs – over eastern Asia and eastern North America – with a weaker one over eastern Europe. The former two are determined by two factors. One is upwind topography – the Tibetan Plateau and the Rocky Mountains, respectively. These act through the conservation of potential vorticity to set up a ridge in the airflow over the high terrain and a downstream trough over the eastern part of the two continents. The second factor is the thermal gradient in winter between cold land and warm ocean, which leads to a trough developing. In summer, there is a single weak trough over eastern North America as the land–sea thermal contrast is now reversed. The wave pattern irregularly breaks down and blocking anticyclones may develop that disrupt the westerlies. There is either an omega pattern or split flow in the westerlies. There are about 25 such events annually around the Northern Hemisphere. Each event lasts between about one and four weeks. The preferred regions for the blocks to develop are over the North Pacific between 60° and 70°N and over Scandinavia between 50° and 60°N.

The Antarctic is overlain by a deep polar vortex with extratropical cyclones and mesoscale polar low systems. The former spiral in towards the continent over the Southern Ocean (Figure 2), whereas the polar lows move equatorward. The polar lows are comma clouds and spiraliform systems, with an approximate ratio of 10:1. In many, but not all, winters there is a correlation between more extensive sea ice and an increased frequency of polar lows. Unlike in the Northern Hemisphere, the southern polar vortex is nearly circular, making for stronger, more persistent westerlies. In winter it expands equatorward but, in contrast to the Northern Hemisphere, the southern westerlies show little seasonal variation in strength. There is a basic wave number 1 pattern as the vortex is displaced towards the Indian Ocean. This is related to the fact that the most remote sector of the ice sheet (the “pole of inaccessibility” at 83°S, 55°E, and the highest part of the ice sheet) is displaced away from the pole towards the Indian Ocean and southwest South Pacific. Hence, the continent is asymmetric with respect to the South Pole. Occasionally, there is a wave number 3 pattern with the troughs over the major ocean basins. Blocking is less common in the Southern Hemisphere and is concentrated primarily around New Zealand. Close to the Antarctic continent the mean flow is easterly as a result of average high pressure over the ice sheet. There is a circum-Antarctic trough around the continent that has centers near the major embayments of the Ross Sea and Weddell Sea, and off Wilkes Land in East Antarctica. These centers reflect the locations towards which mid-latitude systems move, undergoing stagnation and cycloysis. Cold air outbreaks on their western sides often give rise to the formation of mesocyclones.

The Antarctic trough shifts equatorward at the solstices and poleward at the equinoxes in response to the semi-annual (half-yearly) pressure oscillation. This is a mid-tropospheric feature that involves the different annual cycles of air temperature between the Antarctic and the surrounding mid-latitude oceans. The temperature difference is strong at the equinoxes. At latitude 50°S cooling in autumn is rapid compared with the warming in spring (September), the opposite being the case at latitude 65°S. The semi-annual maximum tropospheric temperature gradients shift the circumpolar trough poleward during the transition seasons. In turn, this produces a semi-annual oscillation in the
pressures (≈5–6 mb) and in the winds over the area affected by the trough.

Corresponding to the NAM there is a Southern Annular Mode (SAM) in the atmospheric circulation between the Antarctic and southern mid-latitudes. This is the dominant circulation mode in the Southern Hemisphere. The SAM index was mostly negative from 1980 to 1993, positive to 2000, then negative to 2008. There has been an increasing trend in the SAM in the austral summer since the mid-1960s, and in austral autumn since 1958, that has been linked to the intensification of the ozone “hole” in the lower stratosphere. This develops each spring as a result of chemical interactions between the solar ultraviolet (UV) radiation and chlorofluorocarbons (CFCs). The UV radiation interacts with the CFCs, producing free radicals that break down the ozone molecules catalytically. This hole is projected to be repaired by
around 2050 as the CFC concentrations are reduced by natural decay following the implementation of the Montreal Protocol in 1989, when nations agreed to phase out the production of harmful chlorine chemicals.

The Pacific–South America (PSA) pattern links pressure anomalies in a standing wave train across the South Pacific and South Atlantic oceans. There is an out-of-phase relationship between pressure anomalies in the Bellingshausen-Amundsen seas and the Weddell Sea sector. This relationship east and west of the Antarctic Peninsula comprises an Antarctic Dipole. It appears to link anomalies of temperature and pressure in the seas around Antarctica with tropical Pacific sea surface temperatures and surface air temperatures; the relationship is best developed during cold La Niña events in the eastern tropical Pacific.

An Antarctic Circumpolar Wave (ACW) in the atmosphere that consisted of a wave number 2 pattern, that propagated eastward around the continent over an eight-year interval, was observed from 1985 to 1994, but was largely absent before and after that time. It appears to have originated in the southern extratropics. The troughs and ridges gave rise to a suite of ocean anomalies that propagated eastward, maintaining their signatures over seasons and the different regions of the Southern Ocean.

The phases of El Niño Southern Oscillation (ENSO) in the tropical East Pacific have substantial effects in the Southern Ocean and over Antarctica. During El Niño (warm events in the tropical Eastern Pacific) most of Antarctica is colder than usual, with enhanced katabatic winds. East of the Antarctic Peninsula there are more southerly winds giving cold conditions with more extensive sea ice. In La Niña, with opposite sea surface temperature anomalies, there is a strengthened Amundsen Sea low and a weaker Weddell Sea low. Cold conditions west of the Antarctic Peninsula then lead to more extensive sea ice in contrast to the conditions to the east.

Climatic elements

Air temperature

Mean annual air temperatures over Antarctica average −60°C or below over the 4000-m-high East Antarctic Ice Sheet (EAIS) and around −10° to −30°C at the coast. Along the Antarctic Peninsula in winter the temperature increases about 1°C per degree of latitude. The mean annual range is about 30°C over the plateau with a range between extremes of 70°C. At the coasts, temperatures in summer rise to near freezing; they are affected by onshore winds and in the Dry Valleys they generally increase with distance inland. In winter at the coast, katabatic winds generally prevent very low temperatures as a result of turbulent mixing in the boundary layer.

Over the central Arctic, mean temperatures range from −30°C in January to near 0°C in July. Away from the coasts, temperatures drop to between −40 and −50°C in January and rise to between 10 and 20°C in July.

Cloudiness

There is high cloudiness (80–90%) over the Southern Ocean throughout the year, while over the central plateau amounts average 40–50% annually, and less than 10% in summer, due to the very dry air and the absence of cyclones.

In the Arctic the areas with open ocean have high cloudiness all year round. Over the pack ice, cloud amounts in winter average around 50%, while in summer there is a high percentage of low stratiform cloud cover due to a combination of evaporation from melt ponds and leads, and moisture advection from the south.
Polar Climates

Precipitation

The Antarctic Plateau is a polar desert with very low snowfall amounts; 50 mm or less is typical. “Diamond dust” (suspended ice crystals) and rime deposition are common. There is considerable drifting and blowing snow. Along the coasts, annual totals of 200–300 mm are typical.

Over the Arctic Ocean, precipitation totals are around 250 mm, with at least half of it falling as rain in the summer months. Totals are higher over the Kara Sea (up to 400 mm) and over the Bering Sea (400–800 mm) as a result of more cyclone activity and higher winter temperatures and specific humidity.

Winds

The surface winds in Antarctica are a persistent, downslope, katabatic flow driven by gravity and thermal gradients. They turn to the left of the maximum downslope gradient as a result of Coriolis acceleration in the Southern Hemisphere. They are shallow systems (100–200 m deep). Winds are strongest at the coast where the ice sheet slope is greatest and topographic channeling is common; mean annual speeds of 20 m s\(^{-1}\) are recorded at Cape Denison, for example. These winds play a major role in opening leads (linear cracks) and coastal polynyas (irregular-shaped openings) in the sea ice.

Climatic types

There are three major climatic types in polar latitudes on land: tundra, polar deserts, and ice plateaus. Over the open polar oceans there is a polar maritime type and over sea ice there is a seasonally dependent regime.

Tundra

Tundra means “treeless” in Lapp, and this zone occurs north of the treeline in arctic Canada, arctic Siberia, the Siberian islands, coastal Greenland around the ice sheet, and in Svalbard. Winters are long and cold while summers are short, with temperatures around 5–10°C. In winter, temperatures are around −25 to −30°C, except when low-pressure systems bring warmer air masses from the North Pacific or North Atlantic. Figure 3 shows mean monthly temperatures for Alert on the north coast of Ellesmere Island, in stark contrast with South Pole Station in Antarctica. Most of the limited precipitation falls as rain in summer. Annual precipitation totals are between 150 and 250 mm in high latitudes. Winter snowfall amounts are low and snow depths range from about 60 to 160 cm. The region is windswept and blowing snow is common in winter. Snowmelt is rapid in June, as a result of the long days, and the ground thaws to depths of 30–100 cm above the permafrost table. In consequence, the ground is frequently waterlogged. Permafrost (permanently frozen ground) is continuous beneath most of the arctic tundra.

Figure 3  Mean monthly temperatures (°C) at Alert (81°N) and South Pole Station. Source: Barry and Hall-McKim (2014). Reproduced with kind permission from American Meteorological Society.
and has typical depths ranging from 100–600 m in arctic Canada and Siberia. At its southern margins it is rapidly degrading, and ground temperatures in much of the tundra are rising.

Cold polar deserts

Cold polar deserts occur on the poleward margins of the tundra in the northern portion of the Canadian Arctic Archipelago, northern Greenland, and the Siberian Arctic islands, and in the Dry Valleys of Antarctica. They are attributable to low atmospheric humidity and infrequent weather systems that result in low precipitation (100–300 mm). Temperatures average between −30 and −40 °C in winter and between 1 and 5 °C in the summer. In polar deserts the ground is usually frozen below a shallow active layer in summer, and the permafrost may be hundreds of meters thick.

Ice plateaus

The two major ice plateaus are the Greenland and Antarctic ice sheets. Greenland is basically an ice dome rising to 3200 m at latitude 72 N. It extends 2400 km north to south and is over 1000 km wide at its widest point, covering 1.7 million km². Numerous small ice caps and glaciers surround it. Antarctica is 98% ice and it covers 14 million km². It is made up of the larger EAIS that rises to 4500 m and the smaller West Antarctic Ice Sheet (WAIS), flanked by the floating Ross ice shelf in the Ross Sea and the Filchner-Ronne ice shelf in the Weddell Sea. There are many other smaller ice shelves around the coastline and along the Antarctic Peninsula. The ice shelves are composed of floating land ice that flows out into the ocean and they are typically several hundred meters thick.

Greenland is affected by storm systems that enter Baffin Bay from the south, or move across it from the west. They also redevelop off the southeast coast, bringing significant precipitation in southeasterly onshore flow. The northern part of the ice sheet, however, is very dry. Mean annual accumulation on the ice sheet is approximately 340 mm water equivalent (w.e.), with over 2000 mm in the southeast. The higher part of the ice sheet is dry snow but near the margins there is substantial melt in summer. Ponds and small lakes form on the ice surface, some of which drain rapidly through moulins (shafts in the ice). At Summit Station (73°N, 39°W, 3207 m) the winter mean temperature is −41 °C and the summer mean is −15 °C. Temperatures at coastal stations average 4–6 °C in summer. The wind regime is dominated by katabatic drainage winds that can reach gale force, especially near the coast where they are channeled by valleys and fiords.

The climate of Antarctica is the most severe on Earth. The mean monthly temperature at South Pole Station, located at 2835 m altitude, varies between −29 °C in December–January and −59 °C in July–August (Figure 3), with extremes of −13 °C and −83 °C. Notice that, unlike Alert, the winter temperatures remain low for six months in a pattern referred to as a “coreless” winter – there is no mid-winter core of cold. This pattern is caused by the consistently negative radiation balance, in the absence of solar radiation. Vostok Station (78°S, 107°E, 3488 m) on the East Antarctic Plateau has recorded a world record low temperature of −89 °C. Satellite data have recently been used to calculate record minima of approximately −93 °C near the East Antarctic ice divide. As noted earlier, there are strong surface temperature inversions. The surface winds over the continent are almost always downslope due to gravity and thermal gradients, deflected by the Coriolis effect. They are typically decoupled from the synoptic gradient winds above; they are shallow and best
developed on the slopes of the ice sheet. In winter (April–October), Antarctica is encircled by extensive sea ice, covering approximately 20 million km² in September, which augments the low temperatures of the continent. This sea ice shrinks and largely disappears in March.

**Polar maritime climate**

The polar maritime (or marine) climate is delimited by the mean temperature of the coldest month being below −6.7°C, where the poleward margins approximately coincide.
with the maximum winter extent of pack ice. Summers are cool (5–10°C) and cloudy. This climatic type occurs in three regions: (a) a southwest-to-northeast zone in the North Atlantic, from south of Greenland through the Denmark Strait to the Barents Sea north of
POLAR CLIMATES

Norway and the Kola Peninsula; (b) the southern Bering Sea and Aleutian Islands; and (c) a circumpolar zone over the Southern Ocean and sub-Antarctic islands between about 49 and 60°S (Figure 4). There is a cold season precipitation maximum and a large proportion of the annual precipitation falls as rain, with many days of precipitation. It also has a high storm frequency, extensive cloud cover, and strong winds.

Ice-covered seas

The extent of sea ice in each polar region changes dramatically with time of year. In the Arctic it ranged from 13–15 million km² in March to 4–5 million km² in September over the last decade. In September 2012 there was a record minimum of 3.4 million km². In the Antarctic it ranges from 15–16 million km² in September to 2 million km² in March. The seasonal sea ice zone (SSIZ) in the Antarctic extends to 55–60°S in September and its overall area has increased slightly over the past few decades, perhaps in relation to a reduction in sea surface temperatures. Earlier, from the 1930s to the 1950s, the sea ice edge was about 2–4° latitude north of its average location in the 1980s. The climatic conditions both determine and in turn are affected by the surface properties – snow-covered sea ice, melt ponds, leads, and polynyas. In the Arctic, in January, temperatures average just below −30°C, rising to 0°C in July. Precipitation peaks in late summer/autumn in response to the cyclonic activity. Even in July about half of the precipitation falls in solid form.

Little is known about the Antarctic ice zone except in the Weddell Sea sector. Snowfall depresses the sea ice and flooding of the floes leads to the formation of snow ice and slush. Ice thickness is typically 0.6–0.8 m at maximum, but in the western Weddell Sea it ranges up to 1.5–2.5 m along the Antarctic Peninsula.

The sea ice zone is a transition region between Antarctica and the sub-Antarctic. Near the coast, the sea ice moves westward, steered by the persistent easterly winds. Farther north, in the westerly wind belt, cyclones form in the lee of the Antarctic Peninsula, where winds are forced over the topographic barrier. Generally, lows from mid-latitudes move southward, slow down, and decline near the continent, but give rise to little precipitation.

SEE ALSO: Atmospheric/general circulation; Climatology; Sea ice, ice drift, and oceanic circulation

References


Further reading


Polar region ecosystems

David M. Cairns
Texas A&M University, USA

Polar ecosystems represent some of the extremes of life on Earth. The continental landmass of Antarctica contrasts strongly with the sea ice-dominated oceans of the north polar regions. In addition, the lack of significant ice-free land at south polar latitudes results in limited tundra environments there. In general, the predominant polar ecosystems can be divided into terrestrial and marine systems.

Terrestrial polar ice-free environments are dominated by tundra. The arctic tundra and polar desert biomes constitute about 20% of North America. Traditional views of arctic vegetation have relied upon vegetation maps with bioclimatic zones as the defining top-level hierarchical unit. These conceptualizations of the Arctic have divided the region into the High and Low Arctic based on a multivariate set of environmental characteristics including temperature, soils, and vegetation cover and functional type. Recently, the Arctic has been mapped using bioclimate subzones and vegetation physiognomy (growth form). The circumpolar arctic vegetation map (CAVM) retains the High and Low Arctic differentiation, but further divides each zone into finer units (Figure 1). Bioclimatic subzones A, B, and C represent the traditional High Arctic conditions and are associated with low-stature vegetation on mainly mineral soils. Subzones D and E are the Low Arctic with typically more vegetation cover on organic soils. Overall, the vegetation of the circumpolar Arctic is classified as barrens, graminoid tundras, prostrate-shrub tundras, erect-shrub tundras, or wetlands. Barrens and graminoid tundras dominate the High Arctic (bioclimatic zones A and B). Shrub tundras are present in the warmest part of the High Arctic (bioclimatic zone C) and common in the Low Arctic (bioclimatic zones D and E). Wetlands are topographically determined and present in all bioclimatic zones.

The amount and nature of woody vegetation within the Arctic is indicative of summer temperature regimes. Woody vegetation gradients thus tend to follow the temperature gradient. As such, the zonal vegetation of the coldest bioclimatic zone lacks woody shrubs and is nearly barren of vegetation or characterized by a thin cover of herbaceous plants, mosses, and lichens. Woody vegetation shifts from prostrate dwarf shrubs (e.g., Dryas and Salix spp.), to hemiprostrate dwarf shrubs (semi-erect or trailing stems and less than 0.15 m tall: e.g., Cassiope tetragona) to erect dwarf shrubs (<0.5 m tall: Betula, Vaccinium, Ledum, and Empetrum spp.), and finally to low shrubs (0.4 to 2 m tall: many of the same species as erect dwarf shrub tundra with the addition of Alnus spp. and some taller Salix spp.) with increasing temperatures.

Surface ice dominates the vast majority of land at southern polar latitudes. In ice-free locations some plants do survive. Antarctic flora is unique because it is the only terrestrial continental flora dominated by mosses, lichens, and a few liverworts. There are only two species of flowering plants, and therefore classifications of antarctic vegetation rely on two broad classes of vegetation: lichens and bryophytes. Bryophyte distribution is associated with ground moisture. The gradient in lichens, however, is
more complex and is associated with substrate conditions, the availability of nitrogen, and physiological responses to moisture.

Phytoplankton form the base of the trophic chains in polar seas and warm-blooded air-breathing predators are at the top of the food chain. Historically, polar marine ecosystems have been thought to be characterized by short, low-diversity food chains. More recent research, however, has indicated that the low-diversity conceptualization of the marine polar ecosystems is overly simplistic. The diversity of plankton forming the base of the food chain is no different than in other nonpolar systems. There are differences between the poles in the key prey organisms for the higher trophic level vertebrates. In the Antarctic, crustaceans are the primary prey for vertebrates, whereas in the Arctic it is fish and
benthos. High-level predators such as whales and seals have been subject to human exploitation for centuries and it is difficult to disentangle the impacts of human populations from decade-scale fluctuations in climate and hydrography.

Quaternary history of polar vegetation and animal communities

Prior to the Pleistocene glaciations, a circum-polar arctic flora that consisted of around 1500 species developed. By contrast, the modern circumarctic flora is around 1000–1100 species. The modern flora biodiversity of the Antarctic is far less and is comprised mostly of lichens and mosses.

The most recent Pleistocene glaciation resulted in ice sheets that covered much of the north polar region. Large areas of Beringia (from Alaska to eastern Siberia) were left ice-free and were a dry arctic desert. Much of northern Europe, extending from Scandinavia east to the Urals, was also covered in ice. Arctic-adapted plants survived either in refugia associated with the ice-free areas or migrated southward beyond the glacial margins.

The reduction in size of the modern flora for the Arctic relative to that of the pre-Pleistocene glaciations resulted from multiple advances and contractions of the polar ice caps. During the last glacial maximum (LGM, about 18 ka) in North America, polar vegetation species survived in ice-free refugia in central Alaska and to the south of the ice sheets. Modern genetic analysis indicates that Beringia served as a refuge and a source of recolonization for arctic flora during the LGM; whereas western Siberia served primarily as a refuge, but not as much as a source of gene flow for recolonization of newly ice-free areas.

The late Pleistocene and early Holocene can be divided into three vegetation periods for the Northwestern part of Canada. The oldest, the full- and late-glacial, extended from 25 ka to 11 ka and was dominated by herbaceous arctic and alpine taxa. During the early Holocene (11–7 ka) there was a period of rapid vegetation change. Dwarf shrub tundras and forests became common at the southern margin of the arctic landscape. The late Holocene (7 ka to present) began with an increase in the amount of alder pollen and has led to a period of relative stability in species composition.

During the LGM, populations of large herbivores, such as the wooly mammoth (*Mammuthus primigenius*), decreased and moved southward. After the LGM the populations migrated northward along with the vegetation characteristic of the mammoth steppe, but eventually died out, probably due to the speed of the changes in climate, the presence of humans in North America, and the increased prevalence of conifers, peatlands, and birch (*Betula* spp.) in the North American arctic landscape. Birch spp. can be toxic to cecal digesters such as mammoths.

The importance of ice

In addition to ice having been a dominant determinant of the distribution of life within the polar regions over long time periods, it remains an important driver of patterns of life and diversity today. Direct effects of sea ice are felt most strongly in the marine portions of the polar environments. Sea ice serves as the substrate for an array of micro-organisms and also acts as habitat for a set of vertebrates such as walrus, polar bears, and seals. Sea ice algae and sub-ice phytoplankton account for more than half of the Arctic Ocean primary productivity. Ice-algae and phytoplankton blooms are tightly
POLAR REGION ECOSYSTEMS

linked to the extent, duration, and timing of melt of the sea ice. Ice-algae and phytoplankton are the base of the tightly coupled polar marine food web. Zooplankton populations rely upon the phytoplankton blooms which are linked to fish such as cod and their avian and marine mammalian predators. In addition to providing substrate for ice-algae and modifying the light environment for phytoplankton, the sea ice also facilitates foraging, reproduction, and resting of vertebrate species in the polar regions. Polar bears, seals, and walruses are the most prominent users of the sea ice in this fashion.

Most of the land in the Arctic is ice-free; in contrast, only about 2% of the antarctic continent is ice-free. Permafrost, however, is a common occurrence in both the Arctic and the Antarctic. Permafrost is defined as any subsurface material that remains at or below 0°C for longer than two years. Polar regions may experience permafrost up to 1500 m deep, and although much more prevalent in continents, offshore permafrost conditions do exist below unfrozen marine sediments in some cases. Deep permafrost has relatively little influence on the ecology of the surface. Permafrost in areas that are rich in ground ice has significant impacts on topography, hydrology, and ecology. An active layer (substrate that freezes and thaws seasonally due to air temperature oscillations) is characteristic of terrestrial permafrost zones. Complex topographic surfaces that are created by the differential thaw of ice-rich permafrost (thermokarst) have been developing across the Arctic as temperatures have risen over the past century.

Melting of ice-rich permafrost has the potential to alter the hydrology of an environment producing significant shifts in vegetation on the surface. For example, large areas of birch forest in central Alaska have been transformed into bogs and fens. There is also the potential for the release of carbon to the atmosphere from its storage in frozen ground as the climate warms.

The presence of permafrost associated patterned ground influences the distribution and patterning of vegetation. For example, in the Russian Low Arctic (and other locations around the Arctic) alder savannas are common. Individual alder shrubs and clumps of alders are regularly spaced across a landscape. There is a spatial association between patterned ground microsites and the alders. At these locations “circles” or “frost boils” develop in which a circular patch of mineral soil develops surrounded by larger clasts. These circles are regularly spaced and zones of these types of features may be very widespread. Recently, using a combination of satellite imagery with field data collection, Frost et al. (2013) have shown that there is a strong correlation between the patches of mineral soil and the presence of alders. Alders preferentially establish in the mineral soils as opposed to the more rocky surrounding environment. This preference results in a regular pattern of alder on the landscape. The underlying geomorphology is controlling the pattern of the vegetation in this environment which in turn has implications for the distribution of nutrients, patterns of permafrost, and future colonization of other vegetation.

Interaction of plants and animals

Plants and animals are linked in the Arctic through their phenology (the timing of events in their life cycles) and by the disturbances or perturbations that their presence can have on the ecosystem. Phenological linkages can occur through the timing of migrations, and in the synchronization of life stages with available forage. For example, caribou migrations in Alaska, Canada, and West Greenland have historically shown tight coupling of the calving
season and the emergence of plants from winter dormancy. Similar phenophase synchronization occurs in nonmigratory animals. For example, in Fennoscandia, the autumnal moth (*Epirrita autumnata*) is an important component of the dominant subarctic forest systems. The timing of the hatching of the moth eggs and their early development is linked to bud flush of their primary forage – the leaves of the mountain birch (*Betula pubescens ssp. czerepanovii*). The chemical composition of the leaves changes through time so that as the moth larvae grow, the leaves become less palatable to the moths. Disruptions of the linkage between the phenophases of the two organisms can cause significant perturbations to the ecosystem.

Animals on a landscape influence the vegetation communities in both positive and negative ways. One conceptual model of the impact of animal activity on vegetation outlines three possible responses of the vegetation to increasing amounts of animal activity. Increased animal activity may have a positive, a negative, or an intensity-limited effect on vegetation communities. In cases where increases in animal activity lead to increased productivity or abundance of certain species, the driving mechanisms include animal-facilitated dispersal of plants and the fertilization of the environment.

Animals may disperse plant seeds by carrying them either inside or outside of their bodies. The distance that species are carried in any single dispersal event varies with the travel distances of the animal vector. Mammals tend to travel shorter distances than birds, for instance. In the Arctic, reindeer and caribou have been shown to carry seeds in their digestive system, and the presence of the animals on the landscape facilitates the dispersal of those seeds. Most of the seeds are for herbaceous plants. Birds are also important dispersal vectors and can carry seeds very long distances.

An example of how increased animal activity may influence the productivity and diversity of a site through fertilization comes from Antarctica. At some antarctic locations, penguins serve as a conduit for transferring nutrients from the marine to the terrestrial environment. In the vicinity of penguin colonies, diversity of vegetation tends to increase over short periods after abandonment (5–6 years). Areas fertilized by penguin guano tend to have higher amounts of vascular flora compared to unfertilized areas that are dominated by bryophytes and lichens. This example illustrates that animal activity may not be good for all species in an area. The presence of animals therefore has the potential to alter vegetation composition. In the antarctic example, vascular plants were facilitated by the fertilization of the environment by penguins, but at the expense of the bryophytes and lichens.

Defoliating outbreaks are an example of the second (negative) mode of interaction between plants and animals. For example, increases in autumnal moth larvae populations reduce the productivity of the birch trees in arctic Scandinavia that are their primary forage. Clear evidence exists from tree rings illustrating the reduction in growth of mountain birch for between two and three years after an outbreak. Outbreaks of the autumnal moth not only have the potential to reduce the growth and productivity of the large trees that are their primary forage, but in some cases they also reduce the productivity and biomass of the ground vegetation. The regular rodent cycles in these systems are also tied to the variability of the vegetation biomass, and consequently the impacts of one species in a system may cascade through the larger ecosystem. Similarly, the presence of reindeer significantly reduces the ground vegetation in terms of both the size and the abundance of keystone species such as *Cladina stellaris* (one of the main species of lichen
**Polar Region Ecosystems**

in the region). In Alaska, ptarmigan (*Lagopus* spp.) browse heavily on shrubs above the snow surface at sites north of the Brooks Range. Browsing by ptarmigan has an influence on their height and growth form and stimulates stump shooting.

Importance of disturbance and human activity to polar ecosystems

No ecosystem is static. Changes to an ecosystem occur naturally through autogenic processes of succession over time or can be imposed by exogenous factors. Polar ecosystems are no different in this regard. To understand the impacts of exogenous disturbances resulting from development at the poles (mostly in the Arctic) it is important to first understand the natural successional pathways that occur there.

In contrast to successional processes at temperate and tropical locations, polar ecosystems tend to undergo successional changes very slowly. For example, in the High Arctic where vegetation cover is often less than 1%, the interspecies interactions are quite limited. The physiological and genetic attributes of colonizing individuals determine successional pathways rather than interactions between the plants. In these environments, physical disturbances or climatic conditions can continually reset the successional clock thereby extending the time for passage along the successional trajectory. Although the vegetation cover is typically higher at Low Arctic sites, the rate of change remains slow due to incomplete plant cover, physical disturbance, and a low resource base. Limited resources for plant growth in the Low Arctic are associated with low temperatures, a short growing season, and limited nutrient turnover due to slow decomposition rates.

Normal successional trajectories in the Arctic can be modified by nutrient additions. Additional nutrients result in the rapid establishment of graminoid vegetation and can ultimately change the end point of the successional process. For example, at fertilized sites, the growth of the woody shrubs *Betula* and *Salix* occurs at the expense of the tussock cottongrass (*Eriophorum vaginatum*).

The depauperate flora of Antarctica create more simplified successional trajectories than in the Arctic. On continental Antarctica, primary succession is understood better than secondary succession. Studies of successional trends on recently deglaciated foregrounds have shown that the lichen and bryophyte communities that dominate these environments are divided into three stages of development. Pioneer communities may develop within a few decades of glacial recession, intermediate-aged communities develop on surfaces 300–400 years old, and climax communities occur on the oldest terrains. Age is the most dominant factor that influences the vegetation composition of these sites, but duration of snow cover and substrate characteristics may modify successional trajectories. Although newly deglaciated terrains can be invaded relatively quickly, their successional pathways can be impeded or halted at intermediate stages due to environmental conditions.

There is a long history of human impacts on polar ecosystems. The type of human impact differs between the Arctic and Antarctic. Resource exploitation and human settlements are common in the Arctic, but are virtually absent in the Antarctic. The Madrid Protocol to the Antarctic Treaty was adopted in 1991 and has designated Antarctica as a “natural reserve, devoted to peace and science.” It has banned mining and oil exploration on the antarctic continent. Recent increases in research and tourism activity there have begun to produce human impacts that
are having an influence on soil biota and may also alter ecosystem processes such as carbon cycling. Trampling due to limited foot traffic in the Dry Valleys near McMurdo can also reduce populations of various soil biota. In contrast, the subantarctic islands are and have been less well regulated and human impacts are more common.

The natural ecosystems of subantarctic islands were significantly impacted by the arrival of people involved in whaling and sealing prior to the middle of the twentieth century. There was a transfer of nutrients to terrestrial environments by scavengers and an introduction of a variety of exotic species (both plants and animals). Because there were no baseline ecosystem studies performed before the human activities began, there is limited understanding about what the “natural” conditions of these islands were like. In some cases, the extirpation of local seal populations allowed for increased vegetation productivity.

Exploration for and extraction of natural resources have become more prevalent in the Arctic during the last 50 years. The development of oil and gas fields by both North America and Russia has produced changes in the local ecosystems. On the Alaskan North Slope, winter seismic exploration has had an effect on vegetation on permafrost soils. Vehicle traffic due to oil and gas exploration in the mid-1980s resulted in plant community composition change. Over the long term, vegetation communities lost cover of evergreen shrubs and mosses and increased cover in gramminoid plants. The changes persisted over at least two decades, and the intensity of the disturbance influenced recovery rate. Vascular plant cover declines along the exploration lines in all types of tundra relative to undisturbed reference plots. Diversity decreases, as does abundance.

At High Arctic sites, vascular and cryptogamic species richness has been shown to be greatly reduced after low-intensity disturbance by tracked vehicles in both mesic hummock vegetation and wetter meadows. Seedling establishment is virtually nonexistent at sites subjected to multiple passes by tracked vehicles. There is some evidence that the persistent vegetation changes that result from older exploration activities (10–30 years) have reduced bird abundances.

In addition to the direct effects of resource exploitation in the Arctic, on the Yamal Peninsula in northwest Siberia, the discovery of large natural gas fields in the 1960s has resulted in changes in land use. Large portions of the land have been taken out of traditional usage (primarily for reindeer herding) and have been set aside for oil and gas exploration. The result has been to concentrate the traditional activities of the native Yamal Nenets into much smaller areas. The increased density of reindeer in these areas has resulted in the trampling of and grazing of the tundra. The changes in vegetation have resulted in the development of thermokarst, and initiated Aeolian erosion. There has been an extensive transformation from shrub- to grass- and sedge-dominated tundra. Land-use changes in northern Finland have also significantly affected the reindeer pastures and the herding system.

Disturbed arctic tundra sites recover very slowly due to the arrested successional processes. Comparisons between archeological sites and more recent disturbances due to pedestrian traffic show similar responses in the Canadian Arctic and other Low Arctic locations. The relatively low number of species present at High Arctic locations provide only a limited number of avenues to recovery from disturbance. Assisted revegetation of arctic sites disturbed by human activity has occurred across the Arctic. The success of these attempts has varied, and has occasionally relied on the introduction of non-native species. Rates of revegetation differ between wetland soils and those characterized more by gravel, sandy, and saline soils.
Climate change implications for the polar ecosystems

As the temperature in the Arctic has risen over the past 100 years, the land surface has also begun to change. The permafrost underlying much of the terrestrial arctic land masses has begun to thaw, changing the terrain, the hydrology, and the availability of nutrients. One impact of these changes has been the “greening of the Arctic.” In the late 1980s, satellite data became available to illustrate the distribution and magnitude of this change. Advanced Very High Resolution Radiometer (AVHRR) data indicated that the Arctic was becoming significantly greener. This change was concentrated particularly in much of Alaska, eastern Siberia, and the northern part of Fennoscandia. Subsequent monitoring efforts using similar data have shown that the pan-arctic greening trend has continued into the twenty-first century.

Although the satellite data have been able to illustrate that some change is occurring, they are at relatively coarse scales and are not capable of clearly illustrating what is happening on the ground. Other methods of monitoring change have been necessary to illustrate the details of what is happening. One example of this is the increased shrub cover across the Arctic. Using aerial photographs taken from low-flying airplanes in the mid-twentieth and early twenty-first centuries, Sturm, Racine, and Tape (2001) have shown that for the Coleville River basin on Alaska’s North Slope tall shrubs have been growing larger, filling in the non-shrub interstices, and expanding into new areas. Research from other locations in the Arctic have confirmed that this pattern is occurring world-wide and that it is a component of the greening that had been previously identified using satellite data. Recently, some evidence has become available that addresses the spatial heterogeneity of the response of these shrubs to the changing arctic climate.

The tundra has also been changing over the past 50 years. The international tundra experiment (ITEX) is a coordinated set of experiments using similar methodologies at tundra locations around the globe. The results of this experiment have illustrated that the impacts of climate warming differ between sites: shrubs increase with warming at sites that were warmer at the start of the experiment, and graminoids increase more at sites that were colder at the start of the experiment. Fertilization studies at tussock tundra sites have indicated that increased nutrient availability, coupled with warmer temperatures, leads to increased biomass and losses of carbon and nitrogen from deep soil layers. These changes in soil nutrients may form a positive feedback to climate warming if they are sustained.

Polar marine systems are also likely to experience changes under a warming climate. The most dramatic change documented to date has been the decrease in sea ice. Sea ice change occurs as a reduction in extent, thickness, and duration. The lowest arctic sea ice extent based on a satellite record occurred in 2005. Since 1979 there has been about a 3% per decade reduction in sea ice extent measured in February in the Arctic. Such reductions produce fewer days for foraging of polar bears, and a reduced area for walrus to pull out of the ocean for reproduction and as a base for foraging. Changes in sea ice may also have an impact on gene flow, and the transmission of pathogens.

In the Antarctic, sea ice is linked to population dynamics of the emperor penguin (*Aptenodytes forsteri*), and declines in sea ice are predicted to result in severe declines of their populations. Changes in synoptic climatology related to changes in sea ice may influence atmospheric and oceanic circulation patterns. The impact of these changes may outweigh that of temperature
change for terrestrial subantarctic ecosystems characteristic of the subantarctic islands.

The poles of the Earth represent integrated systems that are more than the sum of their parts. The processes that determine the functioning of polar ecosystems are linked through direct connections and also by feedbacks that can either lead to stability or contribute to change. The most readily observable of these feedbacks is the surface albedo-temperature feedback. As the surface albedo changes (due to the melting of sea ice, changes in the distribution of plant communities on land, and the melting of glaciers), temperatures are expected to increase at faster rates than have been yet observed. The changes in temperatures will induce continued change. Polar ecosystems will adapt to these changes and may eventually reach an environmental state that is very different from the current system state.

SEE ALSO: Biodiversity; Biogeography; Climate change and biogeography; Permafrost: definition and extent; Polar climates; Sea ice, ice drift, and oceanic circulation

References


Further reading


Political ecology and scale

Matthew D. Turner
University of Wisconsin–Madison, USA

The relationships between humans and their biophysical environment operate at multiple spatial and temporal scales. For example, the clearing of a patch of temperate forest for agriculture leads to a rapid transformation of its biochemical cycles and ecological populations. The farming of the cleared piece of land has a slower but ultimately more transformative effect on soils and the prospects for forest regeneration. Forest regeneration after field abandonment occurs over a longer time period with multiple possible trajectories shaped not only by the characteristics of how the forest was originally cleared and by subsequent farming, but also by climate, the surrounding landscape matrix, human uses of the forest, and chance during the decades of forest succession. Clearly, what one observes in terms of ecological change depends on the length of time that one observes and when one observes. Moreover, one’s ecological assessment may vary depending on whether one is making observations at the edge or in the middle of the cleared patch, which would both differ from observations over a broader spatial extent that encompassed many such patches in a shifting mosaic of transition. The spatial and temporal scales of society’s material exchanges with the biophysical world are affected not only by intentional resource management and policy but also by the scaling effects of the broader political economy.

Given the importance of spatial and temporal scales in both producing environmental change and affecting our observations of it, the way we talk and think about human–environment relations often involves simplified, sometimes underexamined, categorizations of spatial and/or temporal scale. Terms such as “landscape scale,” “watershed management,” “community-based conservation,” “connectivity,” “sustainable use,” “buffer zones,” “disturbance,” and “resilience” very much animate the worlds of environmental research and practice. The use of these and other terms in environmental science and policy discourse invoke particular spatial or temporal scales. For instance, watershed management, a popular approach in water quality and soils management, is seen to improve previous management approaches in that the “natural” unit of the watershed is adopted as the scale of management. In this way, the downstream effects of the (mis)management of a farmer’s field could be incorporated into policy and management decisions. While it may better match the scales at which water moves, watershed management may not match the scales of rural land-use dynamics or existing management institutions. Effective management, therefore, necessarily results from a recognition of both biophysical and social scales, not in isolation from one another, but as coproduced.

Political ecology, as an approach, seeks to understand the material landscape transformations and contestations around them within the context of broader social change. This commitment has compelled political ecologists to explicitly or implicitly address scaling issues in their work. As a result, political ecological work has made important contributions to
our understanding of scaling questions in the fields of environmental change and governance. The topic of political ecology and scale will be addressed here by providing: (i) a basic understanding of what is meant by questions of scale in the social and biophysical sciences; (ii) a description of the treatment of scale in political ecology’s conceptual framework; and (iii) a discussion of the treatment of scale in political ecology’s study of environmental governance.

The concept of scale

Scale is an important concept for all subfields of geography, with a corresponding diversity of meanings attached to it and the processes that reflect and produce it. Questions of scale have also elicited significant attention within ecology and the biophysical sciences more broadly. Much political ecological work can be characterized as analyses of society–environment relations, contextualized by history and place, with a particular emphasis on the environmental and social justice implications of broader political economic change (Zimmerer and Bassett 2003; Robbins 2004). The field of political ecology includes work that privileges ecological concerns over political concerns (a minority), and that which is less concerned with ecology per se than with the politics surrounding the meanings attached to, the use of, and struggles over access to the nonhuman world. Given this diversity, political ecologists’ engagements with scale span the epistemological divisions between and within human and physical geography. As such, it potentially offers much to contemporary understandings of scale, particularly with respect to the interplay of the social and biophysical processes that produce scale. Prior writings on this subject by Sayre (2005, 2009), Cohen and Bakker (2014), and Neumann (2009) are particularly insightful and have informed this entry.

Sayre’s work (2005), comparing treatments of scale in ecology (sharing commonalities with physical geography) and human geography, is particularly helpful – not only for introducing greater clarity to the scale debates in human geography at the time, but also for its discussion of scale in ecological and social processes (operational scale). As Sayre (2009) observes, the multiple meanings of scale include size, level, and relation. Scale as size is the meaning attached to scale as a quantitative measure that dominates in cartography and/or geographical information science, as well as in the biophysical sciences, with its use both referring to the extent and resolution of observation and display. This meaning of scale sits squarely within what Sayre describes as the epistemological moment – namely, what one observes (and therefore understands) about socioecological dynamics is shaped by the temporal and spatial scales that one adopts as a researcher, policymaker, or land manager. Thus, as described, one’s assessment of ecological response to human activities can be quite different depending on the temporal and spatial scales that one adopts to make observations.

Scale as level is a more qualitative measure, referring to the scale categories that we impose on the world (epistemological) and the discontinuities that exist in human societies and ecologies (ontological). Social and biophysical worlds are structured and as such show discontinuities in the nature of social and biophysical processes. The watershed represents one such discontinuity in hydrological systems (ontological). Still, the branching drainage and river patterns reveal that watersheds of different orders are nested within each other. Therefore, uses of the term “watershed scale” reflect not only ontological discontinuities but also, owing to the choice of the level within the nested hydrological network
to call the “watershed,” observational or epistemological discontinuities. We as observers have simplified the complexity of hydrologic networks. Similarly, social scientists may, in referring to the “household” or “village” implicitly as a closed social group, ignore or downplay the social networks that transcend these social group categories. In short, the categories that we impose on complex realities may impose harder boundaries or discontinuities on social or biophysical processes and properties than actually exist. In both these cases, the levels or categories in question reflect both discontinuities in social and biophysical processes and discontinuities introduced by the observer. Therefore, scale as level often has what Sayre describes as both ontological and epistemological moments. These moments are highly intertwined in the sense that our conceptions of the world can help produce material discontinuities in the world. Sayre goes on to argue that a major source of confusion is the conflation of these two moments – focusing on one of these moments of scale as level to the exclusion of the other – not only among social scientists but also among ecologists.

Scale as relation focuses on the biophysical and social processes that move through scale to produce, through their interaction, a scaled, leveled, or structured world (ontological). Most political ecologists or ecologists would argue that they are not inherently interested in scale alone but instead are interested in the social and biophysical processes that reflect and produce the ontological moment of scale as level. Using the earlier example, the patterns of conversion of forest to cropland (distribution and sizes of cleared parcels) is shaped not only by the physical environment (soil types, topographic position) but also by the infrastructure (roads and access to markets), technology, and the distribution of the control over productive resources (labor, capital, land). The resulting mosaic of land cover will influence scale-dependent variables in landscape ecology such as connectivity and edges, which will shape not only forest succession (at different levels) but also wildlife movements (at different levels) and human–wildlife conflict. In these ways scaled landscapes are both produced by and shape social and ecological dynamics over time (at annual, decadal, centennial, millennial levels).

“Chains of explanation” and spatial scaling

Scale also lies at the heart of the persistent attractiveness and challenges facing political ecology as an approach. As admirably reviewed, for example by Rangan and Kull (2009), Neumann (2009), Brown and Purcell (2005), and Robbins (2004), a major feature of early political ecology was its embrace of the interaction of social and ecological processes operating at different scales. It promised to link broader political economic processes to on-the-ground management and environmental change through the concept of “chains of explanation” (Blaikie 1985). In referring to the power of the approach, Blaikie (1985, 88) states that “soil erosion problems can be analyzed in a framework of Chinese boxes, each fitting inside the other,” with the boxes being social organization levels such as the individual, household, village, district, and so on. Very early on, the concept was critiqued by those aligned with political ecology as an approach that was theoretically underdeveloped, leading to seemingly ad hoc, contingent, and empirically derived explanatory sequences tying broader political economic changes to changes in local society–nature relations (Peet and Watts 1996). Since then, political ecology as a field has grown and diversified from its agrarian
POLITICAL ECOLOGY AND SCALE

political economic roots (Blaikie 1985). Political ecologists have become more comfortable with muddling through without an overarching theory tying political economic change at broad scales to local communities while still raising concerns about the implicit hierarchical or fixed container model that underlies the chain of explanation concept (e.g., Rangan and Kull 2009; Brown and Purcell 2005; Zimmerer and Bassett 2003; Robbins 2004). As a result, it has been argued that Blaikie’s analogy of nested Chinese boxes has had a persistent effect, with subsequent political ecologists relying on a hierarchical framework to interpret the processes that animate “scales of relation.” A number of scholars have argued for a greater reliance on networks (horizontal and vertical), conceptualized variously, to analyze relations that work across and produce scale (Rocheleau and Roth 2007; Robbins 2004; Birkenholtz 2011).

Brown and Purcell argue along these lines to raise the point that political ecologists’ rigid adherence to this scaling hierarchy leads them to fall into a “local trap” where the key to desirable environmental and social justice outcomes is seen to be the “devolution of power to local-scale actors and organizations” (2005, 608). This represents an important caution. Still there is reason to believe that these issues are not as inherent or ubiquitous to political ecology as some may argue. First, there is much political ecological research, some of it very early in its history, that critically engages with “the local” and calls for checks on local-scale authority. More importantly, Neumann (2009) argues that political ecologists, while they do not often cite the scale literature coming out of political economic geography (exceptions include Zimmerer 2000; Rangan and Kull 2009; Cohen and Bakker 2014; Swyngedouw and Heeynen 2003; McCarthy 2005), treat scale not as rigid but as fluid, historically contingent, and socially malleable. Indeed they are not theoretically constrained to a particular scale, and in fact there are many examples of political ecological work (e.g., Turner 1999) that address Swyngedouw’s call for scale research that “never resides in a particular geographical scale, but rather in the [social] processes through which particular scales become (re)constituted” (1997, 169, quoted in Brown and Purcell 2005). Moreover, one can point to the literatures on territorialization, decentralization, counter-mapping, commodity chains, and boundary construction, to which political ecologists contribute, as very much treating “scale” as politically contested and constructed. In this work, social processes working across scale or political contestations having scalar implications are the concern, with little analytical leverage lost by not explicitly invoking scale per se.

In short, treatments of scalar politics by political ecologists demonstrate that many have not been captured by a hierarchical vision of scale or fallen into a “local trap.” Still, the local trap debate provided a caution to the methodological commitments shared in political ecology. Political ecology is dominated by place-based approaches through which the researcher makes causal claims connecting observed society–nature relations in place to broader political economy and ecology through a combination of secondary research, tracing networks, or progressive contextualization as interpreted by a combination of mediating middle-level theories. Place-based research does not necessarily result in the researcher being captured by the local, but it can result in local places being more richly characterized and understood. This is not a problem faced solely by political ecology but by all place-based approaches seeking to make causal connections beyond that experienced and observed in research.
Scale and environmental governance

In this section, the various ways in which political ecology has treated the malleability of scale (size, level, relation) in environmental governance and politics are described — in other words, how scale is produced, invoked, and circulated. Environmental governance is a major area of inquiry within political ecology. It not only encompasses government and non-governmental programs that specifically focus on environmental conservation or resource management but also the environmental consequences of other forms of governance or use of social power. There is a politics to all of these instances — power is contested and therefore environmental governance research is necessarily concerned with resource-related conflict. The politics that surround environmental governance includes not only overt contestations over authority in electoral, bureaucratic, legal, and less formal political venues, but also the “knowledge politics” over acceptance of different understandings of environmental response to human activities. The politics of knowledge takes place not only in scientific labs, environmental NGO offices, ministry buildings, debates at conferences, or the pages of journals but more generally in the broader discourses surrounding environmental management and governance that shape what counts as valid environmental knowledge and what serves as widely accepted common knowledge within epistemic communities.

Observational scale and knowledge politics

How is scale implicated in such politics? Environmental analysis, policy, and politics are very much shaped by the temporal and spatial scales adopted by different observers (e.g., Sayre’s observational scale). At a fundamental level, environmental changes are seen as problems by human society (and social groups) only if they are observable, valued by humans as important, and elevated as problems by individuals and institutions with power. In short, the scaling of environmental change in relation to the observational scales used by human observers has a significant effect on what, among the myriad of biophysical variations that are occurring, is treated as directional change and a problem to be addressed (Cohen and Bakker 2014). Different temporal frames placed over environmental variation will often result in quite different assessments whether change is directional, seemingly random, or cyclic. Gradual environmental changes over timescales that exceed the human life span are often ignored or are not measurable owing to the limitations of data or measurement technology. Even over shorter timescales (decades), directional change in an environmental variable that fluctuates widely at shorter timescales (years) may also not be identified as change. As landscape ecology has clearly demonstrated, high variation at more spatially resolved levels can mask changes at broader levels.

Environmental monitoring and assessment are very much bounded in that they necessarily focus on the change of particular variables within circumscribed areas. This bounding defines the spatial scale of observations, which in turn shapes the processes that can be observed and our normative assessments of changes. The statement that someone’s environmental improvement is another’s degradation reflects not only different values held about changes in the biophysical environment but also different people’s placement within the spatially specific array of changes resulting from a human activity. Soil erosion is a common example of this, whereby topsoil loss from a farmer’s field is associated with material flows which, as one
POLITICAL ECOLOGY AND SCALE

expands out spatially, may result in a gain of topsoil to other farmers’ fields or stream silation downstream in the watershed. Almost all environmental changes have this characteristic, with the effective observational scale affecting normative assessments and political responses to environmental change.

In these ways, scale as size plays an important role in environmental monitoring and assessment and their influence on environmental governance. Terms such as scarcity, biodiversity, competition, or connectivity, which very much frame contemporary ecological management, are highly scaled concepts – their value, no matter how they are defined, is strongly shaped by the scale of observation (temporal and spatial). Thus we can see differences in how scientists of different disciplines and agencies interpret the same landscape, and how, in turn, their perceptions differ from those living and working on these landscapes. On a deeper level, differences in knowledge systems – spanning the problematic dichotomy between Western scientific and indigenous knowledges – are often associated with different temporal and spatial framings, either explicitly or as matters of belief or tradition. In this way, strategic or nonstrategic spatial and temporal framings underlie the knowledge politics that surrounds environmental questions of interest to political ecologists.

Scale as level in the politics of scale

Our social and biophysical worlds, while open-ended and dynamic, do display levels of structure. As described earlier, social and biophysical scientists make sense of their worlds by dividing the world into organizational categories that variously relate to spatial scale. Social scientists speak of the scales of the individual body, household, community, district, nation-state, and globe. Ecologists similarly speak of the population, community, ecosystem, and biome or, in a more spatially explicit fashion, watershed, landscape, and ecosystem. It should be clear that these organizational typologies are variously related to scale as size. In addition, they incorporate organizational framings (social, biological, jurisdictional, population ecological, hydrological, etc.). The confusion and ambiguities that exist in how such categories relate to scale have produced significant debate in ecology, physical geography, and human geography. The confusion stems in part from the fact that these levels do reflect the scaled nature of different processes (scale as relation) that produce structure in the world and, therefore, are not simply epistemological but have ontological moments (Sayre 2009). Still, their epistemological content, often unexamined, has created barriers to the integration of social and ecological analyses in political ecology and other fields.

Such barriers are often described as mismatches – a mismatch between the scales of sociopolitical research, governance, or jurisdiction with the appropriate scales of ecological functioning. This has led to calls for greater attention to ecologically informed spatial units such as the watershed or eco-region to address the problems associated with community-based or national approaches to development and conservation. For effective programs, it is argued that social and ecological scales should match (e.g., the soil erosion example provided in the previous section). However it is difficult to assign a single scale for management to a web of interacting processes. Ecological communities are composed of organisms that interact with each other and the physical environment at quite different scales. The assignment of an appropriate scale at which human society should manage diverse ecologies has therefore proven illusionary.
When ecologists or land change scientists seek to bridge ecological and social scales, they are often seeking to identify levels of management or governance that would encompass the ecological processes threatened by a human activity. This view of social scaling is different from that offered by the political ecologist who may focus on the levels of social organization at which social processes and relationships work to shape a human-organized activity that produces the particular threat. Policy or management responses to these two approaches may be quite different, with political ecologists more likely to identify multiple areas of intervention at different scales.

It is not surprising that ecological and social scales have entered into environmental management not as spatial scales in their abstract form but as bounded spatial units. The “watershed scale” has become the bounded watershed, with the underlying assumption that, as a hydrologic unit, it represents, like an island, some degree of ecological separation from other such units. The watershed, long a unit of international hydrologic management, has also become a unit for local environmental management, particularly in highland areas. Given the networked, hierarchical nature of river systems and associated watersheds, the watershed represents a convenient biophysical unit for management jurisdictions of different sizes. For low flat or dryland areas of the world, visible hydrological units such as the watershed are not as useful in this regard. As a result, we are more likely to see a switch in conservation discourse where levels of social organization (e.g., community) are used to define the scale of conservation and development programs in these parts of the world. This does not necessarily reflect their ecologies (or social geographies, for that matter) but their hydrologies and the seeming lack, to the human eye, of clear ecological boundaries.

It is important to restate that these categories — whether ecological or social — may reflect ecological and social organization, but their use will often sharpen boundaries and distinctions. Moreover, the invocation of a particular scale for governance — whether defined through physical (e.g., watershed, eco-region) or social (e.g., community) geographies — affects the loci of authority over natural resources and people. An obvious example would be whether management authority is accorded to a local community or to a national ministry (community vs a national forest reserve). Not only do these entities rely on different modes of decision-making, but shifts in authority across these different levels will fundamentally affect the allocation of authority. In this way, there is a politics of scale surrounding the Sayre’s epistemological moment of scale as level (Swyngedouw and Heynen 2003). The politics of scale of concern to political ecologists encompasses but transcends the scalar mismatch of concern to ecological managers. Despite the rhetoric, much of the scale shifting in environmental governance is less about proper ecological management and more about control over natural resources. By engaging with questions of how scales of governance are defined and manipulated by science and politics in combination, political ecologists can consider both the distributional and ecological implications of such moves.

**Scale as relation and the production of scale**

Shifts in the scales of governance and the associated reterritorializations of resourced landscapes have material effects on peoples’ livelihoods and on the ecologies on which they depend.
Political ecology offers much to the understanding of how social and biophysical scales are coproduced, an area of inquiry that has been identified as very important by scale scholars and others (Swyngedouw and Heynen 2003; Cohen and Bakker 2014; McCarthy 2005). Social and ecological processes interact across scales (scale as relation), with these relationships potentially reshaping social and ecological organization. For example, as the body of work on forestry in India has shown, shifts in the authority over forested land between community and national ministries not only have an effect on how forests are managed, but also exert a shaping influence at a broader landscape scale, affecting the size and distribution of forest cover, with significant implications for biodiversity and ecosystem functioning.

Political ecology’s early focus on understanding the livelihood practices of people in place (Blaikie 1985) lends itself well to exploring scales of governance in relation to people’s livelihood practices rather than relying on preconceived categories of social or ecological levels. Political ecologists, through their engagement with not only ecology but people’s livelihood practices, have identified inappropriate features of prescribed resource management practices which have been transported from one socioecological context to the other. An example of such work by political ecologists is work that critically engages with the equilibrium assumptions that underlie many forms of conventional resource management (range management, fisheries, forestry, conservation biology). In the West African Sahel, such assumptions, coupled with sporadic, crisis-led environmental attention, led to circular logics that equated poor range condition with livestock presence. Such assessments contributed to the malign neglect by governments and donors of mobile forms of livestock husbandry. These views have changed but nongovernmental and government organizations continue to pay little heed to the mobility requirements of livestock husbandry in these arid lands. More recent development emphases of granting greater authority to local “communities” have worked in favor of certain social groups, with the consequent enclosure of key pastoral resources. As a result, there is some evidence for a decline in the mobility of livestock owned by pastoralists and farmers alike, which in turn has increased the potential for localized overgrazing and social conflict.

**Conclusion**

Political ecologists, who are interested in how social and ecological processes interact across time and space, necessarily grapple with serious questions of scale in their work. The way they address scale has been largely through careful identification of the interactions of interest and tracing those interacting relations or processes over space and time. Therefore, despite early invocations and the persistent appeal of concepts such as chains of explanation, political ecology’s understanding of how socioecological scales are produced in the world has developed through the tracing of networks, relations, and processes within and between interlinked places (Rocheleau and Roth 2007; Birkenholtz 2011). Scalar ontologies are generally seen as emerging from these interactions.

In environmental governance, political ecology has also exposed the malleability of both observational scale and scalar categories or levels. The spatial and temporal scales that one adopts very much affect what one observes in terms of anthropogenic environmental change. The use of different spatial or temporal scales by different actors reflects not only assumptions
embedded in standard methodologies but, in some cases, choices made by scientists or policymakers to portray society–environment relations in a certain way (scale is an important part of environmental narrative construction). Moreover, environmental governance is very much shaped by a politics of scale where the appropriate scalar levels at which authority is invoked over resource access and monitoring are highly contested. From public lands policy in the western United States to wildlife management in East Africa to forestry policy in Southeast Asia, scale and its malleability figure prominently.

SEE ALSO: Environment and development; Political ecology; Scale

References


Political ecology is a cluster of theoretical and methodological approaches to the study of nature–society relationships, developed largely through and with respect to the study of agrarian dynamics in the Global South, but increasingly applied to a wide range of human–environment relationships throughout the contemporary world. Within Anglophone geography, it has a specific genealogy, recounted below. Yet it has increasingly become a capacious umbrella term used to denote research into any aspect of nature–society relations that shares certain analytical and methodological assumptions and commitments. Moreover, there are other vibrant strands of inquiry and practice, related but distinct from the lineage in Anglophone geography, developing under the label “political ecology” in other linguistic, regional, and political traditions, and there is increasing traffic among these multiple political ecologies (see Bridge, McCarthy, and Perreault 2015, 6; McCarthy, Perreault, and Bridge 2015, 620). Such breadth makes any concise or tightly bounded definition of the field challenging and necessarily incomplete. This entry focuses on political ecology in and from the perspective of Anglophone geography. Attempts at comprehensive introductions to and overviews of the field can be found in Neumann (2005) and Robbins (2011/2004), while a sense of the contemporary composition and range of the field can be found in the volumes edited by Perreault, Bridge, and McCarthy (2015) and Bryant (2015).

Origins

Political ecology grew out of multiple research traditions, in geography and related disciplines, organized around understanding the interactions between human societies and their biophysical environments. In particular, it grew directly out of analyses of how human societies adapted both to the everyday parameters of their environments (e.g., human ecology, cultural ecology, and ecological anthropology) and to their extremes (e.g., the natural hazards tradition). These, in turn, should be understood relative to the understanding of geography as the analysis of human responses to naturally given environments, the dominant conception in the early twentieth century. Political ecology emerged directly out of a critique of such traditions, particularly their conception of adaptation as a functional, unidirectional social response to an environment conceived of as ontologically prior to and separate from society, and their failure to theorize social structures or processes in the modern world that directly impacted the human–environment relationships at the very hearts of those fields (see Watts 2015).

Both the term and the field “political ecology” had multiple moments and sites of origin in the 1970s, which is unsurprising inasmuch as they reflected larger social and intellectual currents (see Bridge, McCarthy, and Perreault 2015). Two of the most directly relevant antecedents for political ecology in geography were uses
by anthropologists: Eric Wolf used “political ecology” in a 1972 article examining property relations in land and resource management, but did not really elaborate on the term beyond the title (Wolf 1972), while in 1977 medical anthropologist Meredith Turshen published an article on “the political ecology of disease” that directly linked human disease patterns in Tanzania to the political and economic relations of the colonial and postcolonial eras. In geography, Richard Walker published a pair of articles in 1973 and 1974 arguing that wetlands management was inseparable from its political-economic context and from power relations, while in 1974 David Harvey published a critique of Limits to Growth and the neo-Malthusian environmentalism it represented, which has since become one of the canonical analyses of environmental questions from a Marxist geographic perspective (Walker 1973, 1974; Harvey 1974). While the latter three pieces critiqued events and actors in the Global North, analytically they contain much that became central to political ecology. Likewise, while Michael Watts’s classic Silent Violence (1983a) does not use the term “political ecology,” all the major elements of what became the field are present in it. Given all of these antecedents, it is surprising both that so many accounts of the development of political ecology begin with Blaikie’s admittedly classic 1985 study, The Political Economy of Soil Erosion in Developing Countries (Blaikie 1985), and that the phrase “political ecology” does not appear in a research article in a geography journal until 1988 (Bassett 1988). The quest for first usages and single origin points is not especially productive, however. What matters is the clearly shared and growing sense in the 1970s and through the 1980s that human–environment relations are always, necessarily, social and political.

Cultural ecology was the most direct antecedent of political ecology, along with the intellectually and disciplinarily closely related fields of human ecology and ecological anthropology. This discussion emphasizes cultural ecology in geography, but the major analytical points apply to all three fields. As noted above, cultural ecology attempted to analyze how human societies adapted to their biophysical environments, with culture understood as a key regulatory mechanism of such adaptation (i.e., specific activities central to how a given society used and reproduced itself in its environment, such as planting or harvesting certain crops, engaging in or abstaining from hunting of a given species or in a certain area, holding feasts, going to war, and so on, were understood as mediated through culture in such a way as to maintain a dynamic equilibrium in the local ecosystem) (see, e.g., Rappaport 1984/1968).

Several important points are critical to an understanding of cultural ecology in the geographic tradition. First, there were strong, direct continuities between the “Berkeley School” of Carl Sauer and its work on cultural landscapes and what became “cultural ecology” by the mid-twentieth century: the former arguably bequeathed to the latter a central empirical focus on premodern agricultural systems, a methodological commitment to fieldwork as essential to understanding such systems, and an analytical focus on purportedly discrete “cultures” or communities as the most significant actors in shaping landscapes, rather than either individuals or extra-local networks. Other key sites and figures for work in cultural ecology in geography included the University of Chicago, the University of Wisconsin, Madison, and the Australian National University in Canberra. Closely related was much work in mid-twentieth-century anthropology, including the work of Julian Steward, who coined the term cultural ecology, and cognate work under the banner of ecological anthropology (e.g., Rappaport
Second, cultural ecology became both more positivist and more intellectually ambitious than Sauer’s more descriptive approach, making serious efforts to trace, quantify, and analyze flows of energy and materials through local human–environment systems, and to identify systemic dynamics such as positive and negative feedback loops. Such calculations and analyses, while inevitably incomplete, made real and substantial contributions to knowledge and understandings of human–environment systems, and especially of the complex causal pathways within them (Robbins 2011/2004; Watts 2015; Huber 2015). Third, the shortcomings of cultural ecology provide nearly a negative image of what became political ecology. Specifically, by the late 1970s and early 1980s, cultural ecology was increasingly criticized for treating small-scale human–environment systems as if they were closed systems in which all material and energy flows were internal and traceable, rather than as connected to a wider world, and for underestimating or leaving unanalyzed increasingly important extra-local sources of change in those systems; for assuming that the societies it examined normally maintained a homeostatic relationship with their environments; and for viewing culture through a functionalist lens, in which everything that a “culture” did somehow functioned to maintain that equilibrium with its environment (Rappaport 1984/1968; see Neumann 2005 for a discussion). Grossman’s (1984) work in Papua New Guinea, which explored precisely how relatively isolated subsistence communities were reorienting their production toward global markets, identified these problems and emblematized a shift within the field toward reversing each of those assumptions (see Watts 2015, 33).

Similar critiques were being raised within and about the natural hazards tradition, a line of work that also informed political ecology. The natural hazards tradition shared with cultural ecology a focus on human adjustments to biophysical environments, but it emphasized, as the name suggests, specifically hazardous aspects of, or events in, those “natural” environments: floods, droughts, and the like. Those hazards were generally treated as “natural” in origin, meaning arising from outside of society, and the hazards tradition focused on analyzing the wide range of human responses and adjustments to them, which could range from decisions about land use and the evolution of building styles to variations in market prices and the development of insurance. In geography, the hazards tradition was especially associated with the geography departments at the University of Chicago, Clark University, the University of Toronto, and the University of Colorado, Boulder (Watts 2015). Analytically, the hazards tradition typically assumed that society was made up of rational individual actors, and it theorized social responses to hazards as the collective outcomes of countless individual decisions. As with cultural ecology, the hazards tradition was subject to substantial internal and external critique in the 1970s that questioned its core assumptions and lack of attention to critical aspects of the phenomena it claimed to study. First and foremost, critics noted that exposure and vulnerability to hazards were not uniform within any society, but were in fact highly unequal based on entirely “social” structural inequalities within any social formation (of gender, of class, etc.) and that such differential exposure was often intimately connected to contemporary economic and geopolitical dynamics (O’Keefe, Westgate, and Wisner 1976; Watts 1983b; Watts and Bohle 1983). In the common formulation, a drought might be natural, but the resulting famine is inescapably social, at least as much about the production and distribution of food and risk as about physical conditions. For example, in
one of the field’s foundational studies, Watts (1983a) argued that the famine in the Sahel in the 1970s was due less to drought than to the fact that colonial and postcolonial shifts in land use, including changes in which crops were planted in the context of a strong national focus on maximizing export earnings, and a shift toward an increasingly monetized economy in which wage labor, debt, and class differentiation were all increasingly prominent, had eroded earlier agrarian production systems and associated moral economies that had successfully reduced risk, redistributed some surplus, and seen people in the region through many previous droughts without famine. The critical point is not merely that exposure and vulnerability to hazards is differential within a given society, but that those differentials are direct results of social structures and processes, such that truly theorizing the social experience of hazards would require a theory of power and inequality within a given society. Second, critics argued that the natural hazards tradition failed to account for the fact that many environmental hazards, far from having origins exogenous to society, are themselves the result of social processes and structures (e.g., climate change). Watts (1983a, 1983b, 2015) argued in particular that dominant versions of the hazards tradition and cultural ecology erred in treating adaptation as a functionalist and unidirectional process in which rational social actors make strategic choices in response to the pressures of a nature external to society.

The foregoing critiques of cultural ecology and the natural hazards tradition were partially informed by, and very much animated by, a combination of peasant studies and specifically third world Marxism that became important across the social sciences in the 1960s and 1970s for both analytical and political reasons (see Watts 2000, 2015; Bridge, McCarthy, and Perreault 2015). In the aftermath of widespread decolonization from the 1950s through the 1970s and in the context of Cold War geopolitical struggles over political allegiances and development trajectories in the resulting “Third World,” political and scholarly interest in so-called area studies (and, especially, in the character and dynamics of peasant-based production systems and political movements) surged, including an analytical focus on the household as the production unit in that context (Watts 2000; Bridge, McCarthy, and Perreault 2015). Such regional production systems and their articulations with the global economy were often analyzed through the related lenses of Marxist dependency theory and world-systems theory. Not surprisingly, “local” communities and their relations with their environment looked quite different through this lens than through a lens that assumed discrete local cultures were working to maintain equilibria with local environments. As Watts (2015, 28) puts it: “In the face of transformations wrought by global capital flows, the second great wave of marketization, and the dynamics of postcolonial development, talk of self-regulating systems, of third-world communities adapted to the ecological niches they occupied, and cognized models in the service of evolutionary flexibility all appeared increasingly problematic.”

The final critical element of the origins of political ecology within geography is the broader turn toward Marxist analysis in economic geography (Bridge, McCarthy, and Perreault 2015; McCarthy, Perreault, and Bridge 2015; see also Bakker 2012; McCarthy 2012; Robertson 2012). Led by David Harvey (e.g., Harvey 1974), many geographers in the late 1960s and 1970s interested in environmental questions began turning to Marxism for ways of thinking about the origins and dynamics of environmental problems that provided an alternative and counter-narrative to the
neo-Malthusian analyses that dominated popular discourse in the Global North in the decade and that were typically punitive and restrictive in their implications for the Global South (see Bridge, McCarthy, and Perreault 2015, 4; Watts 2015, 32). Alternative analyses rooted in Marxist geography that emphasized the unity of society and nature and the economic origins of environment-development dynamics provided critical intellectual resources and frameworks as political ecology developed over the subsequent decade. Moreover, the tension between dominant neo-Malthusian analyses and predominantly Marxist ones espoused by political ecologists became a continuing theme in the field, as seen in the repeated predictions and analyses of so-called resource wars or resource conflicts produced by mainstream sources and the alternative analyses provided by scholars such as Le Billon (see, e.g., Le Billon 2005).

The intertwining of all of these strands – internal and external critiques of cultural ecology and the hazards tradition; a focus on peasant-based production systems in the Global South and how they were changing in the context of postcolonial integrations into global circuits of capital; and anti-Malthusian, Marxist-inspired analyses of the “global environmental problems” newly discovered and articulated by mainstream environmentalists and policy apparatuses – gave birth to what we now refer to as “political ecology” in geography. Again, this is very much a genealogy of political ecology within and from the perspective of Anglophone geography. There are other “political ecologies” as well, particularly in the present (discussed later).

**Places and topics**

Given the origins above, it is no surprise that early work in political ecology focused almost entirely on the dynamics and consequences of changing patterns of access to and control over rural lands and resources in the Global South (see Bridge, McCarthy, and Perreault 2015, 5; Watts 2015, 32). Indeed, Bryant and Bailey (1997) later defined the field as specific to the “third world.” Several themes were especially prominent in this early work, each of which can be illustrated clearly through one or two major early works.

First, one of the greatest differences between earlier cultural ecology and emergent political ecology was the degree of attention paid to the effects of increasing integration with global markets. For rural regions in the Global South, this most often took the form of shifts toward the production of commodities for export (often at the behest of colonial or postcolonial national governments), most commonly agricultural crops or of natural resources such as timber or oil. A focus on increased primary commodity production and export figures prominently in, for example, Watts (1983a), Grossman (1984), and Bassett (1988). The point in part was that increased production of commodities for export usually came at the expense of production of a more varied and resilient range of subsistence goods for local use, but the capitalist relations and dynamics that tended to accompany this shift – toward wage labor, private property, more debt, class differentiation, and infinite accumulation – were also inherently antithetical to and corrosive of the sorts of human–environment equilibria posited and ostensibly studied in cultural ecology. As Watts put it (1983a, xxii), “famines were and are organically linked to the rupture of the balance between peasant subsistence and consumption precipitated by the development and intensification of commodity production.”

Second, political ecology recognized from its inception that states were active architects
POLITICAL ECOLOGY

of the market transformations above, and had additional agendas of their own as well. The emphasis on primary commodity production and export was often driven primarily by the desire of colonial or postcolonial administrations to increase revenues and export earnings, as above. States had other, and broader, agendas as well: to pursue particular development strategies and trajectories; to consolidate control over territory and, in many cases, over the state apparatus itself; and to demonstrate compliance with evolving international norms regarding conservation – even when such projects came at the expense of pre-existing human–environment relations. New postcolonial national governments often pursued such projects aggressively. Analyses of the reasons behind and consequences of state projects figured centrally in, for example, the work of Hecht (1985), Hecht and Cockburn (1989), and Peluso (1992) – all classics of the “first” generation of political ecology.

Third, these questions of use of, access to, and control over the rural lands and resources above were connected to strongly policy-relevant debates regarding the political and political-economic causes of environmental degradation. In two of the field’s foundational works, Blaikie (1985) and Blaikie Brookfield (1987) made powerful arguments that the roots of soil erosion (and, by extension, many other forms of environmental degradation) in such locations lay not in the ignorance, poor judgment, or reproductive decisions of those who directly worked the land (assumptions that still seemed to shape many development interventions and environmentalist discourse), but in the larger political and political-economic contexts in which those “land managers” lived and worked. Specifically, they argued that social and economic marginality each contributed to environmental degradation, as producers, burdened by debt, struggling to feed their households, and often forced on to less productive lands, knowingly “mined” the soil’s fertility in an effort to make ends meet. In the context of debates at the time, the critical point was that economic and social marginalization were causes of environmental degradation, rather than inevitable results of exceeding absolute natural carrying capacities.

A fourth prominent topic that emerged early in political ecology was the criminalization of customary patterns of resource use, often in the context of the establishment or expansion of protected natural areas in the name of conservation. This line of work drew strongly from the work of British Marxist historian E.P. Thompson and his emphasis on “history from below,” particularly his volume *Whigs and Hunters* (1975), as well as from Scott’s (1976) subsequent work on “moral economies.” Its emergence in political ecology was marked by Peluso’s *Rich Forests, Poor People* (1992). This theme connected with the field’s roots in cultural ecology inasmuch as it emphasized how landscapes and livelihoods were co-produced over long periods in ways that, while often informal, were deeply structured and effective, in ways that the expansion of centralized state control often disrupted. It also forged a direct link between political ecology and the study of common property regimes, which sought to delineate and analyze such systems, and environmental history, which was also engaged in a critical reexamination of the concept of “wilderness” and its effects. The point that “wilderness” and “protected nature” were simplifying constructs imposed on complex landscapes of which humans were important co-producers is captured perfectly in the title of another major work in this domain, Neumann’s *Imposing Wilderness* (1998), which historicizes conservation in East Africa. The fact that such impositions are backed by powerful
discourses in which all human activity is assumed to be negative for the environment (sometimes erroneously) was demonstrated by Fairhead and Leach (1996) in a critical analysis of standard accounts of deforestation in West Africa.

Methods

Political ecology is defined and distinguished almost as much by its methodological commitments and approaches as by its theoretical or geographical ones. From its inception, the field has been committed to, and is in large part defined by, intensive, place-based research that seeks to understand the complexities of specific nature–society interactions in one or a handful of cases in historical and geographical context. Ethnographic and other qualitative methods typically figure prominently, as political ecologists, following a model inspired by anthropology and social history, endeavor to understand what the dynamics in question mean for participants from the “bottom up.” Fieldwork is essential, inasmuch as political ecology tends to focus its inquiries on precisely those elements of nature–society relations that cannot be seen or explained from a social or spatial distance – through, for example, national survey data, or remotely sensed images (McCarthy, Perreault, and Bridge 2015, 623). While political ecologists might certainly use the latter sort of data, they would also seek to contextualize them and to see how the local subjects of such surveillance viewed them. Such an approach also means that the methods employed will vary from case to case, and so political ecologists typically employ mixed methods in highly varied ways. In all of these respects, there are strong continuities with cultural ecology, human ecology, and environmental anthropology.

Such a methodological approach requires multiscale analysis (see Neumann 2015; Sayre 2015). Two highly influential formulations in the field are those of Vayda’s (1983) “progressive contextualization,” and Blaikie and Brookfield’s (1987) “chain of explanation.” While there are subtle differences between them, in each, the researcher begins from an observed phenomenon (e.g., soil degradation, deforestation, shifting patterns of resource access or use), and then constructs a causal chain of explanation reaching outward, upward, and backward as necessary. So, to return to our previous examples, soil degradation might result from intensified production of export crops, which might in turn be the result of farmers trying to pay higher taxes, which were in turn put in place by the national government raising taxes precisely to spur greater production and export of cash crops in order to increase revenues to the national treasury, which were needed to pay back international loans dating to efforts to promote national development in the years immediately following decolonization and independence, with the focus on cash crops occurring in a context of national participation in international trade agreements on unfavorable terms due to unequal geopolitical and geoeconomic relationships traceable directly to the colonial era. The larger-scale variables could then themselves be interpreted in light of, for example, changing food regimes and the global historical geographies of capitalism and imperialism.

This approach, which directly or indirectly informs most work in political ecology, brings with it significant strengths and significant challenges. At its best, it refuses many of the dramatic simplifications and reductions so common in academic analysis, including the ontological division between nature and culture. It is historically materialist both in its insistence on beginning from and working through
what is actually happening in people’s lives and environments, and in its acute consciousness of the ways in which history and geography matter. It takes seriously the knowledges and priorities of the people researchers engage, and it maintains a constant attention to the multiple ways in which power works, as well as at least a presumptive sympathy with the less powerful.

Yet, such an approach also carries significant challenges, methodological and analytical. First, tracing relevant causal chains may sound simple in theory or in the hypothetical example above, but it is much less so in reality: causal chains proliferate in many directions, not all of which can be followed. In fact, choices must be made at every moment: in a complex, dynamic world, what temporary stability or claimed observation do we begin from as the alleged “phenomenon” at the start of the chain of explanation? As we move upward and outward, which of many potentially relevant causal pathways do we follow? Also, causality is often inferred more than conclusively demonstrated: what if others do not agree with our assessments that a given branching pathway is clearly the most relevant causal one? Each such choice and its justification is thus necessarily shaped by theory and politics: a purely empiricist or inductive approach is neither possible nor desirable. Peet and Watts criticized Blaikie’s chain of explanation approach for its lack of a strong theoretical framework (Peet and Watts 1996, 8), which would provide a way to narrow, guide, and justify such choices along the way. Conversely, other scholars in this domain have criticized political ecology for the opposite reason, contending that commitments to particular theoretical frameworks lead researchers to seek only evidence that confirms conclusions preordained by the theory (e.g., Vayda and Walters 1999).

Political ecology’s greatest methodological challenge may be the inclusion of ecological data and methods. While the field’s name and alleged objects of study would imply their presence, and they certainly figured prominently in the cultural ecology that preceded it, in-depth ecological methods, data, and analysis have become less prominent in most political ecology (with important exceptions, to be sure: see, e.g., Zimmerer and Bassett 2003 and Goldman, Nadasdy, and Turner 2011). For this reason, the field has been subject to a steady critical refrain that it emphasizes politics over ecology (Vayda and Walters 1999), reinscribes an ontological division in privileging social over natural processes and scales, and more generally struggles to integrate its critical perspectives on science and politics (see Forsyth 2015; Zimmerer 2015). Most political ecologists still advocate methodological pluralism and indeed see openness to the same as one of political ecology’s strengths (see, e.g., Doolittle 2015), and many do practice it, but the constraints of graduate training and other disciplinary pressures toward specialization make it an ongoing challenge. Debates about the epistemological, ontological, and political significance of the divergent trajectories of political ecology and other contemporary approaches to nature–society research in geography, such as land-change science, and whether and how those trajectories can fruitfully converge again, are ongoing (see, e.g., Brannstrom and Vadjunec 2013).

Another important methodological challenge is the question of historical depth: one of the choices above is how far back in history to go in constructing explanations. Political ecology has arguably been very focused on what we might call a meso–temporal scale of somewhere between several decades and roughly two centuries: less would almost certainly be considered too shallow, but more is rare indeed (see the
commentary by Wilson 2014). This focal length presumably reflects the field’s original central focus on the consequences of recent or current incorporation into a modern global capitalist economy. Yet many have argued that it is too shallow, and that the fuller understanding of many of the nature–society dynamics at the heart of political ecology demands a deeper and perhaps more rigorous engagement with historical evidence and analysis (see, e.g., Biehler 2009; Davis 2015; Zimmerer and Bell 2015). The treatment of historical depth and analysis remains a plane of friction between political ecology and environmental history (again, see Wilson 2014).

Questions of generalizability and comparability present an additional set of challenges, both methodological and analytical, for political ecology. Given the field’s insistent focus on the specific and the local (albeit contextualized), it is difficult to treat its substantial corpus of fine-grained case studies as strictly comparable, or to draw generalizable results from them. Yet policymakers, other academic fields, and students frequently ask for such generalizations and comparisons, and the cumulative knowledge production in a field that they often represent (see Castree 2008a, 2008b on “neoliberal natures”). To be clear, many political ecologists do not aspire to produce such results as inputs into the circuits that demand them. Yet in their absence, it can be difficult to convey or translate the field’s knowledge beyond its boundaries, and political ecologists frequently find themselves in the position of criticizing work – of policymakers, development specialists, or those promoting climate resilience, for example – without being able to offer concrete alternatives.

Development of the field

Political ecology grew out of the political and theoretical developments of the 1970s, including the turn toward Marxism in geography and other social sciences. But, while the pivotal works of David Harvey and others were important intellectual resources in the field’s development (see Bridge, McCarthy, and Perreault 2015; see also Bakker 2012; McCarthy 2012; Robertson 2012), the earliest works of political ecology in geography were all tightly focused on how agrarian production systems and social relations, especially those of peasants, in the Global South were being transformed by capitalist transitions and associated state strategies and interventions (e.g., Watts 1983; Hecht 1985; Blaikie 1985; Bassett 1988; Blaikie and Brookfield 1987; see also Peet and Watts 1996, 5). That first wave of political ecology was overtly Marxist, focusing on how individuals, households, and even particular classes were responding to and being transformed by the extension and deepening of capitalist social relations in particular places (although different authors used different variants of Marxist theory, in varying depths). So, questions of capital accumulation, changing relations of production and class structures, new commodities and commodity chains, and the unequal distribution of economic surplus were central (see Watts 2015; McCarthy, Perreault, and Bridge 2015). These early works of political ecology also paid attention to the materialities of agricultural ecology, including soil fertility and forest cover, with a consistency that later became much more uneven in the field.

Three points about political ecology’s Marxist foundations are critical. First, they are arguably what most decisively distinguished the field from its antecedents in cultural ecology, both in advancing an explicit theory of contemporary society, including its macroscopic structures and processes, and in then making those macroscopic structures and processes central to the analysis of “local” nature–society relations in rural areas around the globe. Second, they built into
the field a set of political as well as analytical commitments: most political ecologists neither are, nor aspire to be, neutral with respect to their objects of study (McCarthy, Perreault, and Bridge 2015, 622). As Robbins has put it, what makes the field “political ecology” is precisely that it is an alternative to supposedly “apolitical ecology” (Robbins 2011/2004, 5). This remains a fundamental point of difference with many otherwise overlapping approaches to the same objects of study. Third, these theoretical and political commitments have made the issue of the field’s engagements with policy a difficult one: on the one hand, most political ecologists strongly believe the situations they study can, and should, be different in ways that would make them more just and sustainable: that is, they are anything but politically neutral or clinically objective with respect to intentional change, and they often claim to have insights into the key processes and relations that matter the most in producing objectionable outcomes. And so, many have worked with communities, national governments, and international agencies, including both funders and deliverers of development interventions, in order to try to put their knowledge to work. On the other hand, many believe that the changes needed to truly alter the processes and relations at the heart of the field’s critique are extensive, structural, and beyond the parameters of conventional policy discussions and options. They believe that working with the major players in international environment and development is a form of co-optation that dulls critique, and they can contribute more to change by keeping their analyses as sharp and unvarnished as possible. This tension has been present in the field since its inception, in ways explored in depth in a 2008 special issue of Geoforum focused on the career of Piers Blaikie (Muldavin 2008); see also Carr and Simon (2014), Bebbington (2015), and Wisner (2015).
others detailed the ways in which access to and control over resources and surplus, distribution of the costs and benefits of socioecological transformations, environmental knowledge, modes of political engagement, and more were deeply structured by gender (see Feminist political ecology). Analytical attention to categories of social difference other than class threw into question analytical units such as “the household,” showing that, far from being unitary economic units, they were themselves sites of exploitation and contestation, and that they and other units sometimes taken as given, such as “communities,” were shot through with multiple, intersecting axes and categories of social difference that affected the dynamics political ecologists studied (see, e.g., Carney 1993; Schroeder 1999). From an initial focus on gender, this line of work broadened to analyses of other categories and dimensions of identity, such as indigeneity, ethnicity, and race, and their imbrications with environmental politics (see, e.g., Moore, Kosek, and Pandian 2003; see Yeh and Bryan 2015 for a review). Deeply informed by poststructuralism (see below) and anthropology, thinking around these questions in political ecology has come to focus on intersectionality, the effort to understand the multiple dimensions and categories of identity as always existing and operating in fluid (yet not voluntarist) constellations (see, e.g., Sundberg 2004). These insights from feminist theory have become as much part of the fabric of political ecology as Marxism: the field now takes it as established that categories such as gender and race are dynamic social constructs shaped by unequal power relations, and that they affect the starting points, dynamics, and outcomes of all socioecological processes.

Deeper engagement with feminist theory and questions of gender coincided and overlapped with a turn toward poststructuralist theory in political ecology in the 1990s. Drawing on Foucault, Gramsci, and others, political ecologists broadened their understanding of the sites, means, and substance of environmentally relevant politics. Authors such as Moore (1993), Fairhead and Leach (1996), and Agrawal (2005) began to investigate power and politics as pervasive, and as operating through micropolitics, discourses, scientific knowledge, and the production of new subjects, subjectivities, and consent as well as through formal state policies, capital investments, outright impositions, and coercion. Already evident in the publication of a 1993 special issue of Economic Geography which formed the basis for the first issue of Liberation Ecologies (Peet and Watts 1996), this poststructuralist turn led political ecology to focus increasingly on the investigation of new terrains and dynamics of politics created by environmental interventions, including new processes of subject formation in which “environmental” categories and goals figure centrally (see, e.g., Agrawal 2005; Li 2007). While not strictly poststructuralist, Gramsci’s work on hegemony and the production of consent became increasingly central to work in this vein, providing critical resources for thinking about how environmental claims were used to contrast and contest political legitimacy and create new political identities and coalitions.

As political ecology broadened theoretically, it broadened empirically as well, with the two trends amplifying each other. To the early focus on changing peasant production systems in the immediately postcolonial third world were added studies of the new commodity chains that resulted, producing work that examined not only some of the classic sites and settings of political ecology but also the cultural economies of the global cities where those commodities were consumed (e.g., Freidberg 2004), and the evolution and uneven consequences of myriad attempts to make such commodities and commodity chains more transparent and ethical (see Otto and
Conservation and its impacts on local, and especially indigenous, peoples became a major topic as well, with studies examining both conflicts around protected natural areas (e.g., Neumann 1998) and indigenous social movements that made strategic use of claims regarding identities rooted in, and knowledge of, their environments (see Yeh and Bryan 2015 for a review). Drawing largely from work in anthropology (e.g., Agrawal and Gibson 1999), political ecologists were among the few to question an otherwise largely celebratory turn toward “community-based” resource management and conservation that emerged in the late twentieth century as an alternative to fortress conservation, based on the recognition above that “communities” are typically not unitary actors but rather internally divided terrains of struggle between unequal actors.

As the example of community-based resource management indicates, approaches to resource management travel widely and quickly through global networks in the modern world: consider other examples such as scientific forestry, payments for ecosystem services, or the many iterations of sustainable development. Commodities also flow round the world. As political ecologists began to pay more attention to such networks and flows, undertaking ethnographies not only of small producers but of the international organizations and networks that so affected the latter’s livelihoods (e.g., Goldman 2001), they began to be more interested in the points of origin of some such ideas, and also in the counterflows that complicated simplistic models of the top-down or center-out diffusion of ideas (e.g., foundation program officers who brought ideas with them back to their points of origin). It is thus not surprising that political ecologists began to see many of the dynamics studied in the field thus far at work in other settings and situations: in marine biosphere reserves, where traditional fishers were being displaced by conservation efforts; in North America, where rural people and others engaged in direct primary production also complained that the environmental priorities of environmental nongovernmental organizations (NGOs) and bureaucracies of the central government were displacing and undermining their livelihoods (e.g., McCarthy 2002; St. Martin 2001; Schroeder, St. Martin, and Albert 2006), and where informal and subsistence activities in forests persisted despite dominant discourses that relegated them to the distant past (e.g., Emery and Pierce 2005). This could also be seen in the rapidly growing hegemony of neoliberal approaches to conservation, in which the assignment of individual property rights and payments for ecosystem services were touted as solutions to nearly any problem of environmental management (see, e.g., Dempsey and Robertson 2012); and for that matter in cities around the world, where access to and benefits from environmental goods and amenities such as potable water or green space, as well as exposure to environmental hazards, were markedly unequal, based on race and class and demonstrably shaped by processes of capital accumulation and state formation (see, e.g., Heynen, Kaika, and Swyngedouw 2006; see also Urban political ecology). Finally, political ecology’s increased attention to not just primary production, but the full array of metabolic processes involved in the contemporary global economy, led also to research in the field on waste, using roughly the same mixture of theory and methods used to study production assemblages and their consequences (see Moore 2012).

Current and future directions

Given the growing ambit of political ecology, no summary discussion can do justice to the field’s
POLITICAL ECOLOGY

current and likely future directions. Yet some major trends, reviewed here briefly, are clearly evident.

First and foremost, perhaps, is a surge of work on the interrelated topics of climate change and energy production. With respect to climate change, the topic seems in many respects ideally suited to political-ecological analysis: geographers can contribute careful analyses of how multiple actors, especially the most marginalized and vulnerable, are affected by and respond to changes in their local environments wrought by climate change, understood as a set of physical processes driven by greenhouse gases released mainly in the service of capital accumulation and left unchecked, due in part to the political limitations and imperatives of an interstate system that is itself largely an artifact of the colonial and Cold War eras. Moreover, much of the discipline of geography and some related fields, including many policy networks, claim to have internalized the insights and lessons of political ecology in their thinking about climate change, dutifully acknowledging highly differential vulnerability within social formations based on gender, poverty (although rarely class), and other axes of difference, and at least superficially questioning a single-minded focus on economic growth as the sole good. Indeed, much of the currently tremendously popular and influential work on “resilience” claims to incorporate precisely these insights. Yet, as Bassett and Fogelman (2013), Watts (2015), and others have scathingly demonstrated, many of these recent efforts to “mainstream” the insights of political ecology arguably miss the heart of its critique. First, they rely upon notions of adaptation and adaptive capacity quite similar to those that dominated cultural ecology, in which local actors react to current or imminent changes in their proximate environments in ways that allow them to survive – or not. But, as Watts argues, “it was precisely the limits of adaptation as a form of thought which constituted the very ground on which political ecology emerged during the 1970s and 1980s” (2015, 21). It is a long way from an acknowledgment of differential exposure and vulnerability to a coherent theory of the global capitalist economy and its imbrications with formal and informal political structures and power relations that produce those differential exposures and vulnerabilities, and it is precisely those latter things that are absent from, and indeed incompatible with, currently popular theories of resilience and adaptation. Also mostly absent is a sense of human beings as active producers of environments as well as inhabitants of them. So, while political ecology has contributed to analyses of climate change and should certainly continue to do so (see Liverman 2015 for a review and commentary), it is vital that such work retains the field’s defining critical insights and perspectives.

Closely related to the interest in climate is a boom in work on the political ecology of energy and extractive industries. While the two are not identical, a major impetus for much of this work is undeniably a keen attention to the continued extraction of fossil fuels, especially so-called unconventional ones, in the context of accelerating climate change, as well as the new geographies of energy production, transmission, and consumption occurring as a result of both the ongoing extraction of fossil fuels, and the increasing use of renewable energy sources (see, e.g., Zalik 2009; Bridge et al. 2013). Political ecology’s attention to the energy flows that sustain the global capitalist economy echoes, at a global scale, cultural ecology’s earlier efforts to quantify energy flows within local ecosystems as essential to analysis of those systems (see Huber 2015).

Political-ecological research on fossil fuels (focused mainly on the sites of their extraction)
can be seen as part of a recent growing wave of research on what Bebbington (2012) has called “underground political ecologies.” Bebbington argues that until quite recently, political-ecology research focused overwhelmingly on biotic resources and systems on the surface of the land. While there is a certainly a long tradition of research on the political economy of extractive industries and regions, a large body of research on these topics explicitly identifying as, and drawing on the cumulative insights and frameworks of, political ecology has developed only in the past decade (see Bebbington 2012 for citations and commentary). The intensified search for fossil fuels above is one reason (Zalik 2009); the turn by many putatively postneoliberal Latin American governments to extraction as a source of revenue is surely another (see Bebbington and Bury 2013).

Another important development in political ecology has been the increasing attention to how neoliberalism has reworked environmental governance in many places and sectors, with profound consequences for human societies, nonhuman species, and ecosystems alike (see, e.g., many of the cases collected in Heynen et al. 2007; see Castree 2008a, 2008b for a review and commentary emphasizing the challenges of definition, comparability, and research design). Two things stand out about this body of work with respect to the development of political ecology. First, most of this work draws from, and shares, the key characteristics of political ecology: the research is nearly all intensive and place-specific, often focused on a single natural resource or resource-based industry, overtly Marxist or neo-Marxist, attentive to social context and complexity and the other theoretical resources and developments detailed above, and normatively anti-capitalist. Yet overall, this body of work draws as much from, and identifies with, economic geography as much as political ecology. In this sense, it represents a strong return to one of the early theoretical wellsprings of political ecology – the Marxist turn in economic geography that began in the late 1960s and gathered strength in the 1970s and early 1980s – but also draws directly and strongly on more recent economic geography, including the proliferation of work theorizing the neoliberal turn. It thus draws tighter and more explicit connections between political ecology and economic geography (see Bakker 2012; McCarthy 2012; Robertson 2012 for discussions of this relation). Second, this body of work is defined more by theme than region: its common feature is examination of the consequences of the neoliberal turn for the production of particular socio-natures; inasmuch as that turn has been extremely widespread (if variegated), this body of work examines cases from both the Global North and Global South and emphasizes certain similarities between them, and so perhaps represents a somewhat different geographic imaginary within political ecology (see, e.g., Fletcher, Dressler, and Büscher 2015).

If one new direction for political-ecological research has been downward, looking underground, another has been “inward,” looking at human bodies, their health, and their immediate domestic environments as new objects of research within political ecology. While “traditional” political ecology looked primarily at how individuals and households used the surrounding landscape to support themselves within the constraints of highly structured and rapidly changing contexts, emphasizing accumulation, marginalization, and degradation, it took the actual bodies, health, and homes of those actors largely for granted, beyond placing them with respect to categories and dynamics related to class, gender, race, and ethnicity. By contrast, a new wave of work building on political ecology is looking,
on the one hand, at how human health and disease are deeply interwoven with the production and reproduction of specific livelihoods and socio-natures (see, e.g., King and Crews 2013), and on the other at human bodies as themselves terrains through which capital circulates and accumulates, producing environmental transformations and often degradation in the process (see, e.g., Guthman and Mansfield 2015), including in specific domestic environments not usually examined in political-ecological research focused on the “nature” of productive landscapes and protected conservation areas (Biehler 2009; Biehler and Simon 2011).

Still other new research in political ecology decenters humans entirely, breaking with the field’s undeniably anthropocentric past and focusing instead on how many of the dynamics above – of capital circulating through bodies and landscapes, of new commodity chains, of states exerting control over their territories, of conservation projects and displacements – affect nonhuman animals and ecosystems, including those at the heart of agrarian production systems, domesticated animals and pests, “wild” ones, and ones that blur such categories, regarding them all as significant actors, legitimate subjects of ethical concern, or both (see, e.g., Biehler 2009; Emel and Neo 2015; Gillespie and Collard 2015). Collectively, this direction in the field can be read as breaking new ground for what Sundberg (2011) has called a “posthumanist political ecology.”

One final development in political ecology merits discussion here: as should already be evident from the preceding review, political ecology is an increasingly polycentric field in disciplinary, linguistic, and geographical terms. Intellectually, it draws from and overlaps with anthropology, economic geography, environmental history, science studies, and many other fields. Linguistically and regionally, it is a term increasingly used and recognized in other linguistic and national research traditions beyond Anglophone geography and beyond the Global North, with a plethora of journals, conferences, research networks, and other undertakings developing under the label of “political ecology” in Europe, Latin America, and elsewhere, and connecting political-ecological work in English with work in French, Spanish, Portuguese, and other languages (see multiple chapters and the editors’ introductions and conclusions in Perreault, Bridge, and McCarthy 2015 and Bryant 2015; see also ENTITLE 2015). Recognizing and promoting such connections is a major theme of the collections edited by Perreault, Bridge, and McCarthy (2015) and Bryant (2015), as well as the EU-funded ENTITLE network and its training activities (see www.politicalecology.eu). Geographically, the field is responding to ongoing changes in the structure of the global economy and geopolitics that call into question the North/South framing central to its early decades (McCarthy, Perreault, and Bridge 2015, 628). While political ecology may once have been a field specific to, or about, the third world (Bryant and Bailey 1997), recent and contemporary political-ecological scholarship belies such framings (see, e.g., McCarthy 2005; Schroeder, St. Martin, and Albert 2006; Joshi 2015; Zimmer 2015).

Conclusion

Such a large and growing diversity of topics under the heading of “political ecology” raises the question of what, if anything, this range of work has in common. But careful readings of the above work reveals a continuing commitment to several of the core projects and themes discussed above. Political ecology remains the ambitious effort to understand
specific human–environment interactions both in all of their dynamism and social complexity, and in their larger historical-geographical contexts. The field retains a conviction that this goal requires place-based, intensive research on the one hand, and multiscalar analysis of relevant actors and processes operating at larger, often international or global, scales on the other, inasmuch as the latter directly affect the former. Capitalist processes, state projects, and the priorities of other powerful international actors, such as NGOs and scientific networks, still figure prominently in the latter set. Marxist, feminist, poststructuralist, and increasingly postcolonial theory animates the great majority of this work, which assumes that actors always make decisions in contexts highly structured at multiple scales under conditions of radical asymmetries of power. The goal of such intensive, place-specific research is both an understanding of the specific case as an end in itself – often as part of a long-term engagement between the researcher and a particular group of people – and the generation of insight into other cases, which, while not directly comparable, may have important similarities inasmuch as they are also affected by the same, or parallel, multiscalar actors and processes. The field also remains deeply normative and political, typically focusing on developments researchers find objectionable, often overtly taking sides between opposing claimants, and advocating alternative, often non-capitalist, ways of interacting with environments. What has changed the most, perhaps, is that political ecologists have increasingly applied this framework to the interrogation and analysis of nearly any aspect of contemporary society, whether urban or rural, production or consumption, in the Global South or North, and whether focused on natural resource industries or other sectors. In sum, political ecology is increasingly used to examine the metabolism of global society, often in ways that deliberately span or reject such binary categories.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Disease and illness, political ecology of; Environment and development; Environment and resources, political economy of; Environmentality and green governmentality; Marxist geography; Natural hazards and disasters; Nature conservation

References


POLITICAL ECOLOGY


POLITICAL ECOLOGY


Further reading


Political economy and regional development

Danny MacKinnon
Newcastle University, UK

Political economy approaches became highly influential in geography and regional development during the 1970s and 1980s, and have remained important. In geography, they emerged out of a critique of the so-called quantitative revolution in the late 1950s and 1960s for its perceived lack of social relevance and concern at a time of intensive social and political ferment and profound economic restructuring in the late 1960s and early 1970s. In response, these critics turned to political economy, and to the writings of Marx in particular, in their search for new intellectual foundations. By the 1980s, political economy had became “the dominant discourse of human geography influencing debate, research and the very sociology of the discipline” (Barnes 1995, 423). It was subsequently displaced from the cutting edge of the discipline by a host of competing, though sometimes overlapping, approaches which have variously emphasized culture, institutions, and processes of historical evolution. Yet political economy remains an important source of insight in economic geography and regional development studies, not least in terms of providing a framework for understanding the uneven development of cities and regions in the wake of the financial crisis of 2007–2008 and the subsequent recession.

Political economy should be seen as a broad framework of analysis, encompassing a range of strands and traditions, rather than as a single theory or concept. Its origins lie in the classical political economy of Adam Smith and David Ricardo in the late eighteenth and nineteenth centuries, becoming most associated with Marx’s radical critique and reformulation of their work. Political economy analyzes the economy within its social and political context, rather than seeing it as a separate entity driven by its own set of rules based on individual self-interest. It is concerned not only with the exchange of commodities through the market but also with production and the distribution of wealth between the various sections of the population, while it differs from mainstream economics in its conceptualization of economic life and of the principal drivers of growth and development.

While the classical political economy of the nineteenth century was not concerned with geography, economic geographers have sought to apply a political economy framework to geographical topics such as regional development and urban restructuring. This has given rise to what Sheppard (2011, 320–321) terms “geographical political economy,” which is defined by three key characteristics. First, capitalism is just one way of organizing the economic activities of a society rather than being historically inevitable or natural. This reflects the historically specific nature of political economy as an approach concerned with analyzing the transformation of social and economic relations over time. Second, geography is not passive or external to the economy but is actively produced alongside economic activities. This is
evident in terms of new forms of economic activity giving rise to new growth regions, often alongside the decline of older industrial regions. Third, economic processes must be considered in relation to parallel biophysical, social, and cultural processes such as the utilization of natural resources, social class, gender relations, and identity formation. It is in this sense that geographical political economy approaches should be seen as open and adaptable, being receptive to insights from other perspectives, and evolving in line with capitalism as their main object of analysis.

The remainder of this entry is organized into three main sections. First, it examines the rise of political economy approaches in the 1970s and 1980s and discusses a number of landmark contributions to the study of regional development. The next section assesses the repositioning of political economy approaches to regional development in the 1990s and early 2000s in response to some high-profile criticisms and the emergence of a number of other approaches. This is followed by a section which highlights the continuing relevance of political economy to contemporary issues in regional development, and indicates how it continues to inform research in several key areas.

The rise of political economy approaches, 1970s and 1980s

In response to the pressing social issues of the late 1960s and early 1970s, such as racial divisions in US cities, the Vietnam War (symbolizing the imperialism of US foreign policy), gender inequalities, and the rediscovery of poverty in inner-city ghettos, a group of US-based geographers sought to fashion a new radical geography. This radical geography turned to political economy and the writings of Marx, in particular, to provide a framework for the critical analysis of advanced capitalism. In general, Marxism emphasizes processes and relationships rather than fixed things, adopting a dialectical perspective which sees change as driven by the tensions between opposing forces. Capitalist society in particular is characterized by continual change and flux, driven by the search for profits in the face of competition which requires, as Marx and Engels wrote in *The Communist Manifesto* of 1848, the continual revolutionizing of production in search of competitive advantage and thereby the wider relations of production in society.

While the writings of Marx provided only a few scattered comments and insights into the geography of capitalism, Marxist geographers such as David Harvey sought to build on these by developing a distinctively Marxist analysis of geographical change. From this perspective, the economic landscape is shaped by the conflict-laden relationship between capital and labor, mediated by the state, which provides a stark contrast to the harmonious equilibrium state of regional balance posited by mainstream neoclassical economic theory.

The first phase of Marxist geography concentrated on establishing how capitalism produces specific geographical landscapes. In *The Limits to Capital* – the classic landmark theoretical text of Marxist geography – Harvey (1982) identified a central contradiction between the geographical fixity and motion of capital. There is a need, on the one hand, for fixity of capital in one place for a sustained period, creating a built environment of factories, offices, houses, transport infrastructures, and communication networks to enable production to take place. Such fixity is countered by the need for capital to remain mobile, on the other hand, enabling firms to respond to changing economic conditions by seeking out more profitable locations. This may require
them to withdraw from existing centers of production in which they have invested heavily. Capital is never completely mobile but must put down roots in particular places to be effectively deployed. Nonetheless, its relative mobility lends capital an important spatial advantage over labor, which is more place-bound.

Harvey argues that capital overcomes the friction of space or distance through the production of space in the form of a built environment that enables production and consumption to occur. Indeed, such investment in the built environment can act as a “spatial fix” to capitalism’s inherent tendency toward overproduction by absorbing excess capital, hence performing an important displacement function. As economic conditions change, however, these infrastructures can themselves become a barrier to further expansion, appearing increasingly obsolete and redundant in the face of more attractive investment opportunities elsewhere. In these circumstances, capital is likely to abandon existing centers of production and to establish a new spatial fix involving investment in different regions. The deindustrialization of many established centers of production and to establish a new spatial fix involving investment in different regions. The deindustrialization of many established centers of production in the Rust Belts of North America and Western Europe since the late 1970s and the growth of new industry in Sun Belt regions and the newly industrializing countries of East Asia can be understood in this light.

One of the most notable contributions to Marxist political-economic thinking about regional development is Neil Smith’s theory of uneven development. This is based on the idea that capital moves to the areas that offer the highest profits for investors, leading to differentiation as these areas experience rapid development while other regions are left behind. Over time, the process of economic development in a particular region tends to undermine its own foundations, leading to higher wages, rising land prices, lower unemployment, and the development of trade unions, reducing profit rates. In other regions, underdevelopment leads to low wages, high unemployment, and the absence of trade unions, creating a basis for profit that attracts capital investment. Over time, capital will seesaw from developed to underdeveloped areas, “jumping” between locations in its efforts to maintain profit levels. It is this movement of capital that creates patterns of uneven development. In this sense, “capital is like a plague of locusts. It settles on one place, devours it, then moves on to plague another place” (Smith 1984, 152).

In addition to the abstract analyses developed by Harvey and Smith, another strand of Marxist work was developing more concrete analyses of specific situations and circumstances. One key research question was how particular places were affected by wider processes of economic restructuring, focusing attention on questions of uneven regional development. In a seminal text, Doreen Massey investigated the changing location of industry in the United Kingdom, developing the concept of “spatial divisions of labor.” Massey demonstrated how production was becoming “stretched” across different regions as large corporations sought to capitalize on pre-existing regional differences, particularly those related to the availability of labor. Increasingly, higher-order functions such as research and development were being located in metropolitan regions like the southeast of England, where large reserves of qualified, professional labor could be found, while skilled manufacturing was being concentrated in industrial regions such as the Midlands of England, which contained skilled engineering labor. At the same time, less skilled work related to the assembly of components was being relocated in peripheral regions of the United Kingdom and in less developed countries, where labor costs were much lower.
As such, different types of region were specializing in different parts of the production process. This approach successfully brought together a concern with general processes of capitalist development, relating the locational decisions of corporations to the dynamics of accumulation or investment, and an emphasis on local uniqueness in the form of individual regions.

Another strand of political-economic theory research which was influential in regional development during the 1980s and 1990s is the regulation approach. Derived from the work of a group of French economists in the 1970s, the regulation approach stresses the important role that wider processes of social regulation play in stabilizing capitalist development. These wider processes of regulation find expression in specific institutional arrangements which mediate and manage the underlying contradictions of the capitalist system, enabling renewed growth to occur. This stabilization occurs through the coming together and consolidation of specific modes of regulation: that is, the institutions and conventions which shape the process of capitalist development. When these act in concert, a period of stable growth, known as a regime of accumulation, ensues. In the 1980s the regulation approach provided theoretical support for the idea that postwar Fordism – based on mass production and mass consumption, linked through rising wages for workers and increased productivity in the workplace – had given way to a new regime of flexible accumulation, defined by the growth of more individualized consumption patterns and new information and communication technologies.

The shift to post-Fordism and flexible accumulation was seen to be underpinning new geographies of industrial production in the 1980s, based on the rise of new industrial spaces and regions in hitherto nonindustrialized areas such as Silicon Valley in California and other areas of the US Sun Belt, the M4 corridor in southern England, and the “Third Italy” of Emilia-Romagna, Tuscany, and the Veneto. In addition, the regulation approach was also deployed to study the politics of regulation at the regional scale. For instance, Alain Lipietz, one of the original French regulationists, developed the concept of regional armatures to highlight the role of regional elites in regulating conflictual social relations through the construction of a distinctive regional interest and identity. The activities of this group are instrumental, Lipietz (1994) argues, in the conversion of a “space-in-itself,” which is defined by the prevailing social relations of production in a region, to a “space-for-itself,” which is endowed with the institutional capacity to intervene in processes of economic and social change. At the same time, Jamie Peck and Adam Tickell (1995) drew on earlier research on industrial restructuring and uneven development to develop the concept of local modes of social regulation. These are distinguished by their “unique position within wider (national and international) structures of accumulation and regulation” (Peck and Tickell 1995, 27). They argued that the “regulatory deficit” of southeast England, reflecting the contradictions of uneven development under Thatcherism, meant that it was unable to reproduce the conditions for sustained growth into the recession of the early 1990s.

Beyond political economy? Rethinking regional development in the 1990s and 2000s

By the late 1980s, Marxist political economy was becoming subject to increasing criticism, informed by the emergence of postmodern thought. Marxism had become rather out of
POLITICAL ECONOMY AND REGIONAL DEVELOPMENT

touch with the new times of the 1980s, marked by the dominance of neoliberal ideas, particularly in the United Kingdom and the United States. In the realm of left-wing politics too, the focus was shifting from the traditional politics of distribution, concerned with work, wages, and welfare, to a politics of identity, concerned with asserting the rights to recognition and justice of groups such as women, ethnic minorities, and gay people. Such claims were channeled through broader social movements rather than the traditional labor movement.

Three main criticisms of Marxist political economy in geography can be identified.

1. Its apparent neglect of human agency in terms of an impoverished view of individuals and a failure to recognize human autonomy and creativity. Instead, Marxists tend to privilege wider social forces such as class, and to see people as bearers of class powers and identities, reading their behavior in terms of such attributes instead of seeing them as unique individuals.

2. Its emphasis on economic forces and relations. While the political-economic concept of production is, as has been shown, much broader than conventional notions of the economy, political economists have been criticized for stressing the determining role of economic forces.

3. Its overwhelming emphasis on class, and neglect of other social categories like gender and race. Leading Marxists such as Harvey were attacked by feminist geographers in the early 1990s for their neglect of gender issues and accused of subsuming these within a class-based Marxist analysis.

These are important criticisms, although it is questionable whether Marxist political economy really was as economically deterministic as its critics allege. The criticisms were, nonetheless, influential in encouraging the reorientation of the field of regional development toward a range of new approaches from the early 1990s, emphasizing the institutional and cultural foundations of economic processes. The key context here is the cultural turn which took place in human geography and the social sciences in the late 1980s and 1990s, fostering a new cultural geography focused on questions of identity, meaning, and discourse.

This led to talk of a “new” economic geography which examines the links between economic action and social and cultural practices in different places. As Wills and Lee (1997, xvii) put it, “the point is to contextualise rather than undermine the economic, by locating it within the cultural, social and political relations through which it takes on meaning and direction.” The links between economy and culture are of central importance here, with many observers agreeing that the economy has become increasingly cultural in terms of the growing importance of sectors such as entertainment, retail, and tourism, while culture has become increasingly economic as it is viewed as a set of marketable commodities. Three main strands of culturally orientated research on regional development can be identified.

1. Consumption, with studies focusing, for instance, on the creation and experience of particular landscapes of consumption such as shopping malls and heritage parks. The marketing of localities and regions for tourists and potential new residents has been a particular theme of research in this area.

2. The development of the so-called cultural industries, encompassing art, performance, music, film, broadcasting, museums, and theaters, among others. These cultural industries were identified by governments...
as key growth sectors in the 1990s and 2000s. They tend to be highly concentrated in large metropolitan regions, and are often clustered together close to central business districts, which gives them access to suppliers and provides large markets for their products. Notable examples of such clustering include the motion picture complex of Hollywood, the media complex of Manhattan, and the publishing industry of London. More recently, the focus on cultural industries has given way to an emphasis on the importance of creativity in regional development and the need for cities to attract talented “creatives,” based on the influential writings of Richard Florida.

3 Culture has come to play a key role in local economic strategies in recent years. In particular, industrial cities in Western Europe and North America that faced catastrophic declines in their manufacturing sectors in the 1970s and 1980s have sought to attract and develop cultural facilities and attractions in response, generating new investment and employment opportunities. Bilbao provides a good example of culture-led regeneration, alongside UK cities such as Glasgow, Liverpool, and Newcastle–Gateshead.

A second set of ideas are drawn from institutional economics and economic sociology, emphasizing the social context of economic life. Institutions, defined broadly as rules, conventions, and norms, are important because they link the economic and the social through a set of habits, practices, and routines. The notion that economic processes are grounded, or “embedded,” in social relations is derived from economic sociology. This has been spatialized in regional development studies through the idea of territorial embeddedness, emphasizing how particular forms of economic activity are rooted in particular places. One area in which this idea has been particularly influential is in studies of foreign direct investment (FDI) in response to claims that multinational corporations’ (MNCs) affiliates were becoming increasingly embedded in host regions in the United States and United Kingdom during the 1990s, partly as a result of the increasingly sophisticated set of supports offered by regional institutions, but research has found limited evidence of this.

Institutionalist ideas have encouraged the rise of a new regionalism in economic geography which examines the effects of social and cultural conditions within regions in helping to promote or hinder economic growth. In particular, inherited institutional frameworks and routines are held to be of considerable importance in influencing how particular regions respond to the challenges of globalization. Individual places have, as such, attracted renewed attention, as economic geographers have attempted to identify the social and cultural foundations of economic growth and prosperity in successful regions such as Silicon Valley in California, and Cambridge and “Motorsport Valley” in southeast England. In contrast to the political economy approaches of the 1980s, institutionalist perspectives emphasize the importance of internal conditions within regions in shaping their experience of economic development as opposed to external processes, and treat localities and regions as active participants in economic development rather than as passive arenas that are exploited by capital (e.g., large MNCs).

Third, research in regional development has also been reshaped by an engagement with concepts from evolutionary economics, particularly the notion of path dependence. This means that the ways in which economic actors respond to wider processes of economic change are shaped and informed by past decisions and experiences. As such, the past is expressed and reproduced
through technology, machinery and equipment, and organizations as well as a broader set of attitudes and habits. Regional culture is a product of the past in this sense. The process of path dependence can be illustrated with reference to “old” industrial regions such as the Ruhr Valley in Germany, and northeast England, where cultural factors have been closely associated with decline. The industrial cultures of these areas seemed to have become rigid and fossilized, meaning that firms and institutions were tied to obsolete production systems and methods. This militated against more positive responses to economic change in terms of generating new products and methods. These declining areas can be seen as what Gertler (2003, 134) terms “hard-luck cases”: once successful places where local cultures have become overly rigid and fixed, blocking the emergence of new ideas and new ways of doing things. In this sense, declining regions tend to become locked in to outmoded practices and habits, which prevents them from acquiring new knowledge in an economically effective manner.

The influence of these diverse approaches has not, however, completely supplanted that of Marxist political economy, which is still important. Economic geography and regional development remain haunted by Marx, and Harvey continues to be an important figure, with critics still feeling the need to define their position in relation to his (Sheppard 2011, 320). Some of the key arguments advanced by political economists in the 1970s and 1980s around the unstable and crisis-prone character of capitalism as a mode of production, the inseparability of the economy from the social, political, and ecological spheres, and the need to relate regional problems to wider processes of uneven development continue to be widely accepted by many researchers in the field. The continuing influence of Marxist political economy into the 1990s and 2000s was sometimes explicit, for instance in work on labor and geographical scale, and sometimes more implicit and understated than in the 1970s and 1980s, such as in research on the commodification of places through heritage and tourist consumption. There is also evidence of cross-fertilization between political economy and cultural, institutional, and evolutionary approaches: for example, through the development of a cultural political economy (CPE).

Marxist political economy continued to inspire research in labor geography and regional development through the 1990s and 2000s. While a previous generation of Marxist work had treated labor as a relatively passive factor, considering its distribution and organization across space, the “new labor geography” emphasized the active role that labor plays in shaping the economic landscape (Herod 2001). In particular, labor geographers have been pursuing research that highlights the proactive role of organized labor in creating their own spatial fixes, focusing attention on the need for trade unions to rethink their own geographical strategies to respond to the more global production networks being developed by MNCs. At the same time, political-economic approaches have also dominated geographical theorizing on scale, with particular scales such as the regional or national being seen as products of wider social, political, economic, and cultural processes rather than as predefined arenas within which such processes unfold. This insight has been exemplified by numerous studies of rescaling processes since the 1980s which have analyzed the upscaling of powers from the national to the global and supranational levels, and a parallel downscaling to the regional and local, which has underpinned the new regionalism.

The emergence of CPE in recent years shows how aspects of political economy have been integrated with elements of a cultural approach. More specifically, CPE combines a critical semiotic analysis of language, narrative, and
The continued relevance of political economy to regional development

The new regionalism attracted increasing criticism in the early 2000s, with some of these criticisms invoking long-standing themes of geographical political economy. The first concerned a tendency to take regions for granted as objects of analysis, failing to consider how they have been historically constructed and institutionalized as spaces of policy intervention. Second, the prevailing focus on endogenous capacities and relations entailed a neglect of exogenous networks and institutions, particularly those associated with MNCs and national states. Third, an orientation toward providing accounts of successful regions at particular points in time meant that longer-term questions of regional adaptation and renewal were overlooked. In addition, such accounts were often divorced from wider issues of uneven development and socioeconomic inequality. The continued, indeed increased, pertinence of these issues in the wake of the economic crisis of recent years underlines the ongoing relevance of political-economic categories for analyses of regional economies subject to processes of crisis-driven restructuring aimed at restoring the conditions for profitability. In particular, the adoption of a political-economic perspective helps to connect regionally based research to the big questions of uneven development, social justice, and environmental degradation on a global scale. In the remainder of this short section, the continuing influence of political-economic
thinking is exemplified, with reference to three main areas of contemporary research.

The first of these is work on neoliberalism and the institutional variety of capitalism. In essence, neoliberalism involves the promotion of market forces, individual choice, and a limited state as key principles of economic and social organization. Two of the three principal perspectives on neoliberalism in geographical research are political-economic in nature and origin: Harvey’s conception of it as a class project designed to restore profitability following the economic crisis of the 1970s, and Peck and Tickell’s regulationist-inspired definition of it as a set of constantly evolving policies and practices which have spread globally, but unevenly, through elite policy networks (in addition to the poststructuralist understanding of it as a form of governmental). From a regional development perspective, the adoption of neoliberal doctrines of free trade and capital mobility at the national and international scales have effectively set the underlying rules of the game, establishing a framework of interregional competition which has led to the proliferation of regionally specific competitiveness strategies. At the same time, geographical political economy approaches have focused attention on variegated forms of capitalism at different scales of analysis, going beyond the national orientation of the varieties of capitalism school of comparative political economy to emphasize how economic development is structured by locally and regionally specific institutional arrangements.

Research on global production networks (GPNs) is the second area of interest. This is informed by political-economic research on global commodity chains and writing on actor-network theory (ANT and value chains). A GPN is defined as “the globally organized nexus of interconnected functions and operations by firms and non-firm institutions through which goods and services are produced and distributed” (Coe et al. 2004, 471). As a broad relational framework for the study of economic globalization, the GPN approach signals a renewed interest in extra-regional relations. This is redolent of earlier political economy approaches, although extra-regional processes are conceptualized in relational rather than structural terms, emphasizing multiple actors and relations rather than external domination and control. Regional assets provide an important resource for regional development, but must be harnessed by regional institutions to complement the strategic needs of lead firms in global production networks. From this perspective, regional development is a product of the strategic coupling of global production networks and such regional assets. MacKinnon (2012) argues that strategic coupling should be seen in dynamic terms, focusing attention on processes of decoupling and recoupling between regions and GPNs. Decoupling involves rupturing the relationship between a region and a GPN through plant closure and disinvestment. Recoupling, by contrast, is related to the attraction of repeat investment, recombining existing plants and regional assets with a new round of investment. In many respects, the GPN approach represents a political economy informed framework that is better able to capture the multidimensional and fast-moving character of FDI in contemporary regional economies than the clunkier formulations of the 1970s and 1980s.

The third area of contemporary political economy informed research highlighted here concerns the impact of the financial and economic crisis and subsequent austerity policies. Political economy analyses interpret this recent episode within the broader context of uneven development under capitalism and the evolution of neoliberal governance, focusing attention on the distributional politics of restructuring.
in terms of the unequal effects on different social groups and places. Contrary to initial expectations that the greater impact of the crisis would be on financial services and the relatively prosperous cities and regions where much financial services employment is located, research has demonstrated that workers in manufacturing and lower-income groups, often based in poorer regions, have been worst affected. In the United States, the pattern is complicated by the magnitude of the preceding housing boom in dynamic Sun Belt regions, such that the impact of the initial wave of foreclosure in 2007 was felt most severely in states such as California, Nevada, Arizona, and Florida, in addition to the traditional Rust Belt of Ohio and Michigan. In the United Kingdom, the effect has been to widen existing spatial disparities, with research showing that “the places and people most severely impacted by the last five years of recession and hesitant recovery are those that were already the worst-off” (Townsend and Champion 2014, 3). Thus, the post-2008 decline in employment was steeper and more persistent in the provincial city-regions of the north and the Midlands compared to the United Kingdom as a whole, while London diverged strongly by experiencing a marked recovery from 2010. In both the United States and the United Kingdom, fiscal austerity has been disproportionately concentrated on local government, involving the displacement and downloading of financial responsibility from the national and federal to the local scales, resulting in severe cuts to services and the erosion of collective entitlements. This has been most severe in the cases of municipal bankruptcies in the United States, most notably Detroit, where unelected administrators and emergency managers have assumed draconian powers to slash budgets, cancel contracts, and privatize services.

Conclusions

This entry has assessed the influence of the political-economic approach on regional development research since the 1970s. Political economy has been an important source of theoretical inspiration and direction over this period, exhibiting considerable flexibility and an openness to engage with other perspectives, although its precise influence has fluctuated over time. It became the dominant theoretical approach to regional development by the mid-to late 1980s. Subsequently, however, political economy became the target of criticism for its perceived determinism and neglect of agency, and its dominance was supplanted by the development of new cultural, institutional, and evolutionary approaches. In particular, the new regionalism became highly prominent in the 1990s and early 2000s, based largely on institutionalist thinking. But political economy remained significant and continued to structure and inform research on issues such as labor and scale. It became increasingly complex and diffuse in nature, particularly as a result of its engagement and cross-fertilization with insights from other perspectives, as exemplified by the development of CPE. If anything, political economy has become more prominent in regional development research in the wake of the 2008 financial and economic crisis, as its insights have been deployed to assess the uneven geographical effects of the crisis and its attendant policies of fiscal austerity. It is likely to remain influential, continuing to provide a theoretical framework for linking regional development issues to the larger problems of uneven geographical development, social inequality, and environmental degradation that may themselves become more acute and politically charged over the next few years.
SEE ALSO: Economic geography; Industrial restructuring; Marxist geography; Neoliberalism; Radical geography; Regional inequalities; Regionalism

References


Further reading


Political geography

John Agnew
University of California, Los Angeles, USA

Political geography is about the geographical distribution of power, how it concentrates in some hands and some places, the human and environmental consequences of such concentration, and of how it shifts between places over time. Political geography at its origins as an organized academic field in the late nineteenth century saw the geography in its name as determining the political. More recently, the political is seen as conditioning what is understood as geography. Geography was long understood as the distribution of such physical features of the Earth as mountains, rivers, and oceans but it has since broadened to include politically mediated human reactions to and impacts on the physical environment from settlement and economic-development patterns to human-induced climate change. Compared with political sociology and political science, political geography is less concerned, respectively, with the politics of social groups and the political preferences of individual persons and more with how social groups and people organize and orient themselves in terrestrial space for political purposes. Obvious examples include the field’s long-term focus on such phenomena as the dynamics of interstate borders, the history of modern statehood, electoral geography (including the geographical organization of elections), and the strategic ranking of world regions in foreign policies. Historically, the field has focused on the links between space, regions, and the natural environment, on the one hand, and politics (and polity), on the other. Such linkages were important to ancient and early modern political theory but have weakened in recent political theory, as its practitioners have tended to become largely state focused. Political geography, therefore, recalls a historically more integrative and perhaps more wide-ranging approach to understanding political phenomena.

General intellectual trajectory
1890–present

As a modern field of study, political geography dates from the 1890s as initially an “aid to statecraft” in organizing their empires and confronting their adversaries on the part of the great powers of the day. Largely analogous to what was also known after 1899 as geopolitics, this political geography took contemporary national political identities and reason-of-state as givens. The “needs” of territorial states (given their own personhood in this perspective) and the role of relative location on the Earth’s surface and the resources available in driving and determining the outcome of competition between them were the main concerns of the field. Lurking in the early history of political geography is the history of thinking about how nature relates to nation-state as inherited from the eighteenth-century European enlightenment and the early nineteenth-century Romantic reaction against it. If the idea of levels of development associated with different national territories comes from eighteenth-century enlightenment thought, then that of hierarchies of national territories on a racial–natural basis competing...
with one another for domination comes out of nineteenth-century German idealism.

This intellectual trajectory long dominated the field. Indeed, until very recently, many political geographers were either cameralists (advocates of state-based economies that maximize their self-sufficiency and minimize their transactions with others) or imperialists (favorable to empire building and controlling distant territories). Some of both are still around, even if many advocates of cameralism now think of themselves as on the political left rather than on the nationalist right. Changing times can produce strange bedfellows. But over the past twenty years, liberal perspectives, pitting states against markets, and social perspectives, looking to a plurality of forms of governance, have tended to become more influential than the older statist vision. Of course, disputes among cameralists (both nationalist and state socialist), imperialists, liberals, anarchists, and romantic localists have deep roots in many genres of modern political thought.

But it is not only the range of implicit political projects informing the field that has changed. Intellectual attitude has moved 180 degrees. Since the 1960s, a more independent and critical approach has begun to develop, acknowledging the need to question critically rather than to serve actively the particular interests of the “home state” of the political geographer. At the same time, the empirical scope of the field has widened to consider questions about the origins, spread, and support for political movements and parties, the links between places and political identities, and geographies of nationalism and ethnic conflict. Along with many other parts of political study, political geography has gone from presuming that states (particularly “my” state) are everything to seeing “the political” as everywhere. This intellectual expansion of the field has simply transformed political geography from a particularly state-centered field at its origins to one interested in the range of ways in which geography intersects with “the political” broadly construed: from the material and discursive construction of states and their interrelations to the connections between places and political identities.

Over the entire history of modern political geography a fundamental issue has been the combination of making claims to universal or scientific objectivity while serving national or other particular interests or purposes. This project revolves around the implicit claim to offer a “view from nowhere,” or the idea that the whole world can be known totally from beyond the particular geographical and intellectual location of the person making the claim. But as the philosopher Thomas Nagel (1986, 25–28) has pointed out, there can never be complete objectivity: there is always a necessary “incompleteness” to objective reality. The solution for Nagel is to “enrich the notion of objectivity” by adding the views of oneself and of other selves so as to incorporate the insights of subjectivity. As Nagel concludes: “to insist in every case that the most objective and detached account of a phenomenon is the correct one is likely to lead to reductive conclusions. I have argued that the seductive appeal of objective reality depends on a mistake. It is not a given. Reality is not just objective reality.” So, even though acknowledging the prospect of knowing that is not completely subjective or the result of a singular experience, Nagel is suggesting that the quest for complete objectivity is a fruitless one. There can never be a view from nowhere that is not also, and profoundly, a view from somewhere. In recent years this sort of philosophy has provided an important justification for developing theoretical perspectives sensitive to the discourses, representations, and social practices of those studied around the world rather than simply focusing on the “objective”
causal relationships presumed by the external but invariably privileged observer.

Notwithstanding this critical turn, the early history of the field has had persisting influence even if often in terms of negative appraisal. Canonical writers have cast a long shadow. Friedrich Ratzel and Halford Mackinder, in particular, were a looming presence across the first fifty years of modern political geography and have continued to exert fascination long after their ideas might seem to have faded into the intellectual sunset. They and their major critics constitute the historic canon upon which more recent political geography has been built and against which it has often revolted. There are obvious discontinuities and differences across the writings of the various figures important to the early history of the field. There are, however, a number of continuities that also provide an important backdrop to what was to happen in political geography later on (Claval 2010; Agnew and Muscarà 2012).

The first is the rather unrelenting focus on territorial states as the singular geographical units par excellence of political geography. Although not without its critics (such as the anarchist geographer Elisee Reclus), the enterprise was largely state-oriented almost to the exclusion of political processes operating at other geographical scales and in other ways. The conception of the political was almost entirely statist and territorial, with weaker liberal currents eddying around the edges. The “hard-nosed,” masculinist, and realist conception of the world is here rooted in the geographical facts of an Earth that rewards only those who take what they can. The second is the merging of a naturalized claim to knowledge, based on various mixes of the “view from nowhere” and biological metaphors, with the idealist goal of serving one’s own nation-state. Perhaps only Reclus and much later Karl Wittfogel stand out as major examples to the contrary, at least in so far as serving a particular state is concerned. Although, by the 1930s figures such as Owen Lattimore (1941) in the United States and Jacques Ancel in France can be seen as offering much more critical perspectives on the conventional wisdom: Lattimore because he was largely self-taught and brought his intimate familiarity with Central Asia to bear in all that he wrote; Ancel because he hoped to counter Nazi geopolitics with a more “open” version of geopolitics of his own. The third is a “problem solving” orientation that animated all of the major figures. They aspired to influence policy: to whisper in the ear of the Prince, to paraphrase the Renaissance-era Florentine political philosopher and diplomat Machiavelli. Though geared towards establishing a presence within the new universities of the time, this goal could be served only by appearing “useful” to raison d’état. The prospects for political geography were thus tied to the national flag.

The fourth continuity, and one that distances many of us today so much from the thinking of the time, was the ready acceptance of the language of racial difference and the possible environmental causes of racial divisions. There was no sense of the social construction of racial differences. The American authors, often so critical of European colonialism for its subjugation of other peoples, nevertheless had blind eyes for the reality of American racism at home and in its colonies (such as the Philippines). The fifth, and final, continuity is that Europe and, at least since 1918, the United States are seen as being at the center of the world. The rest of the world is ancillary: bit players or pawns in a world politics driven almost entirely by Great Power competition and inter-imperial rivalries. Of course, the world wars of the period 1875 – 1945 did begin in Europe. But the rest of the world had long figured and increasingly did so in the machinations
of the powerful. And it did so not just as a passive object of desire for the powerful but as an active participant in both its victimization and its own incipient liberation.

These five continuities from the late nineteenth century have not been readily transcended. Indeed, elements of all of them live on today in political geography and beyond: an urge to naturalize knowledge claims to grasp the mantle of science, a Euro-American centered view of the world in which all others are seen as backward when compared to a Euro-American based modernity, and the aspiration to influence national policies to underwrite disciplinary success and gain the ear of political leaders. But beyond these continuities much has also changed in both purposes and subjects of study. Changing political and economic circumstances have had important effects on the empirical substance and theoretical orientations of the field.

Geopolitical contexts and political geography

A good case can be made that the geopolitical context of the time has been crucial to the making of political geography over the past one hundred years. The field has not evolved simply as the result of an internal dynamic, as one paradigm simply replaced another because of intellectual fancy or academic competition. It is not that such considerations have been absent. But they have been relatively less salient to the making of the field than the nature of the world that political geography has claimed to directly report on and interpret. Obviously, the logic of the historical periods used and the geopolitical themes (including the main geopolitical protagonists) identified as crucial to those periods are subject to dispute. Contexts (be they social, cultural, historical, or geopolitical), as Peter Burke (2002, 172) reminds us: “are not found but selected or even constructed, sometimes, consciously, by a process of abstracting from situations and isolating certain phenomena in order to understand them better. What counts as context depends on what one wishes to explain.” The dilemma is that in trying to avoid both “the assumption of eternal wisdom to be found in American political science and great books programs” (Burke 2002, 170) and an internalist account of political geography as a succession of paradigms or great men theorists (the typical alternatives) we can “imprison” all ideas in their contexts not just those that had the closest “fit” with the times. To try to avoid doing so is why identifying ideas that either remained largely marginalized in their time or that arise in times of geopolitical transition out of the merging of previously distinctive perspectives is necessary. But a focus on historical–geopolitical contexts is vital in providing us with summaries of the very materials that political geography has always tried to understand.

The age of inter-imperial rivalry

The time of modern political geography’s founding in the 1890s was one of burgeoning inter-imperial rivalry and arms races between a set of great powers – Germany, Britain, France, the United States, Russia, and Japan – that reached its twin peaks in the two world wars. This period gave rise to the political geography that privileged the role of physical geography and relative global location in determining or conditioning state prospects and limits. Making arguments for control over oceanic sea-lanes and articulating geopolitical pretexts based on the relative propensity of different states to expand were major features of political geography in this phase.

The late nineteenth century was both the zenith of European empire building and the
time when new extra-European great powers, the United States and Japan, emerged into global prominence. So, not only were the European great powers, particularly Britain and France, renewing their colonial activities under stimulus from the colonialism of newly unified Germany and Italy, Europe was no longer the sole center of global imperialism. At the same time, and as noted by the British political geographer Halford Mackinder in 1904, the relatively easy expansion of worldwide empires into “open spaces” that began with Columbus’ voyage in 1492 had come to an end. With the exception of the polar regions, the expansion of any one empire now had to be completely at the expense of the others. More great powers and shrinking space for their expansion spelled possible doom for established and up-and-coming alike unless strategies could be fashioned to protect and enhance what they had from the threats posed by the others.

Of course, the “open spaces” were almost invariably occupied, but by peoples who had succumbed to European deceit, military prowess, and disease. As a result of this history, such peoples had long been judged the civilizational inferiors of Europeans but now they were also categorized as natural racial inferiors too. Into a political atmosphere of intense inter-imperial rivalry, therefore, came a new way of explaining the global political hierarchy based on the increasing acceptance of ideas of environmental and racial determinism. European “success” was, henceforward, to be explained in terms of the climatic and/or racial characteristics of Europeans relative to those whom they had conquered or dominated. If in the past providence had shone down or God had offered a helping hand, now natural characteristics associated with different world regions and the peoples who lived in them were to become popular ways of explaining global political arrangements and offering insights into how best to plan for this or that empire’s future success or brilliance.

The enmity between the great powers was not based solely on sheer willfulness or the will to power of this or that domestic group, such as the military or armaments manufacturers or other capitalists. The period from 1875 to 1945 was one of fundamental instability in the world economy because of the arrival of new great powers wanting to muscle their way into markets controlled by others and the declining capacity of Britain, the main commercial as well as colonial power, to provide its lending services to the world economy without damaging its colonial position. It is worth saying a little more about this because it provides the historical context in which modern political geography came into being.

The period 1815–1875 was one in which Britain held the balance of power in Europe, enjoyed a significant edge in sea power that allowed it a coercive role in imposing its trade and monetary policies around the world, and sponsored a set of doctrines – comparative advantage, free trade, and the gold standard – that, though appearing universal, benefited influential interests in Britain. This combination of European concert and British hegemony elsewhere began to collapse after 1870 once other states with powerful economic and military assets began to challenge Britain. Germany was by far the most important of these. Its capabilities could not be translated into an enhanced global political role without upsetting both the concert and the global flows of trade and capital centered in Britain. Concurrently, the increased industrial production of the United States and the European states undermined British industrial pre-eminence and led British business and governments into the use of nontariff barriers and colonial trade to restrict global free trade and price competition. The net result was an erosion
of the system of trade and finance centered around Britain and the emergence of a set of competitive imperial states dividing the world into zones based on territorial monopoly.

The Cold War era

The Cold War of 1945–1991, with its emphasis on global ideological competition between two models of “modernity” – the democratic capitalism of the United States and the state socialism of the Soviet Union – initially produced a diminished interest in the study of political geography. The field as it had existed before World War II did not seem to offer much food for thought in the new circumstances. Of course, the period did encourage a freezing of political borders and a seemingly permanent standoff between the two sides. Ideology not geography was what mattered. Nevertheless, even during this period geopolitical claims to spheres of influence and the definition of “buffer states” between the two “sides” were important parts of the overall conflict.

The political tenor of the times played an important part in the relative eclipse of political geography. Between 1945 and 1949 a whole new geopolitical order was constructed based on the values, myths, catchwords, and political-economic orientations of the two dominant states: the United States and the Soviet Union. The intellectual genealogy of political geography was ill-suited to this new world. For one thing, in their rhetoric both sides offered different conceptions of the political, on the American side liberal and on the Soviet side Marxist, from the statist one that had long dominated political geography. For another, the entire world was drawn into the bipolar conflict in a way that the world had never been divided previously. But this struggle was seen by all sides as ideological more than territorial, even if it had obvious geographical correlates (the “Iron Curtain” running through Europe, the threefold division of the world into US and allies, the Soviet Union and allies, and a “Third World” of mainly formerly-colonized countries in which the US and Soviet Union competed for influence, etc.) Of course, a more adaptive field in touch with intellectual currents in adjacent fields such as political science and diplomatic history might have had something of a more positive response. The personal danger of questioning the conventional wisdom needs emphasizing. In the United States, figures who did offer alternative perspectives were likely to receive subpoenas to appear before congressional committees investigating “un-American activities” or find difficulty in acquiring academic posts. As it was, in the 1950s nothing very new emerged in political geography in Western Europe and North America that was not based on a conventional wisdom that in eschewing the imperial hubris of pre-World War II geopolitics substituted descriptive gazetteers of border disputes and possible resource conflicts.

As the Cold War slowly eroded, however, political geography underwent something of a revival in the United States and elsewhere, as what had seemed frozen melted in unpredictable ways. The disastrous impact of the Vietnam War on the American economy and politics, the civil rights struggles in the United States, the major economic challenges facing the United States and the world economy following the collapse of the postwar economy, and the rediscovery of “the market” as an abstract guiding political concept all conspired beginning in the late 1960s and early 1970s to create what Daniel Rodgers (2011) called an “Age of Fracture” in which old assumptions about political and economic predictability and certainty were openly called into question. Few fields have been untouched by the wrenching changes of the period beginning at
this time. For political geography in the United States in particular, not only did the distinction between the “domestic” and the “foreign” begin to break down as a meaningful device for organizing thinking in the face of a globalizing world economy but all manner of new topics emerged for study; from the regional-ethnic movements empowered by the end of the Cold War and the invocation of an oft ill-defined “neoliberalism” as the ruling ideology of the age in explaining just about anything, to the challenges facing definitions of citizenship and understandings of nationality in a world of vastly increased international migration.

The revitalization of geography as a whole in the United States in the 1960s had a stimulating effect on political geography. A new generation of geographers discovered an interest in the spacing of social forms, such as urban settlements, land uses, and migrant flows. They increasingly addressed these empirically using the quantitative research methods popular in fields such as economics, demography, and urban sociology. The net effect was to improve the social status of geography as a university subject, at least among adjacent fields in the social sciences. Though the initial revival was to become the subject of controversy, particularly over theoretical and methodological issues, it set the scene for a revival of political geography.

Increasing opposition to Cold War sensibilities was also very important to the revival. While the nuclear arms race gathered pace and Cold War hysteria entered one of its darkest moments, a new generation started to react democratically by beginning to ask how the Cold War had arisen and how it could be made less dangerous. The paradoxical atmosphere of the time is well caught in Stanley Kubrik’s film Dr. Strangelove that in exposing the continuity between World War II and the Cold War also underlined the tragic risk of an accidental nuclear holocaust by presenting it as a comedy.

The events of the late 1960s in North America and Western Europe brought practical politics to the fore in the everyday lives of both ordinary people and academics. The race riots, civil rights marches, Vietnam War demonstrations, and student rebellions of the time had a deep impact on social science. In all fields they brought political questions to the center of concern in ways that they had never been before. In geography this tendency was expressed in three ways. One was by bringing issues of the distribution of power into the analysis of economic and social phenomena, such as residential segregation by race and class in American cities and the global distribution of economic development. These were no longer seen as purely market driven or a matter of free choice but the outcome of systematic bias in political institutions, such as local governments and school districts, managing the distribution of public goods. From this point of view, all of geography became political geography, at least of a kind. From the mantra of “states are everything” that had long animated political geography, geography as a whole confronted the claim that “politics is everywhere.”

Another expression of the political quality of the period was the politicization of geography through public analysis of the field itself: who runs it, for whom, with what ends in mind? In the 1960s higher education had expanded massively in the United States and elsewhere. An important consequence was the recruitment into student bodies and the ranks of university teachers of people from social backgrounds hitherto largely excluded from this world. They did not always accept the norms of personal conduct and political outlook that had evolved in academia during the postwar years. They also challenged the benign acceptance of statehood and international geopolitical hierarchy as “facts
of nature.” A significant example of this comes from France where Yves Lacoste (1976) crafted a new “geopolitics” based on a critique of the old version as essentially geography’s contribution to war-making by powerful states and insisted on the irreducible spatiality of politics within as well as between states. Lacoste’s “geopolitique” is one of the most obvious fruits of the explicit politicization of political geography to emerge in the early 1970s.

In addition, there was a dramatic increase in academic mobility, particularly within the English-speaking world, that brought people with very different social and national backgrounds to those countries, such as the United States, Canada, and Australia, that were most active in recruiting new graduate students and staff. The politicization was not simply disinterested or idealistic, however, because power within the field over staff appointments, publication outlets, professional reputations, and external influence was also very much at stake. The expansion of universities had the effect of dramatically increasing the number of students and faculties, increasing the volume of research and publication, and encouraging new intellectual trends and “niches” so as to find employment for graduate students and gain access to funds for research and graduate financial support.

Finally, political geography was discovered by a new generation without much background or interest in the old studies. They tended to view the subfield naively as something they were inventing afresh rather than an old enterprise into which they should be inducted. Without the old obsession with defining disciplinary boundaries, they searched around in other fields for inspiration and found it in economics, radical sociology, anthropology, economic history, and, even, in political science. At the same time, they embarked on explicitly redefining the field as the geography of politics more than the geography of states. Whatever the geographical scale or context – urban, regional, national, world-regional or global – as long as power pooled up in some places, political organization privileged some in some places over others elsewhere, and territorial boundaries were used to exclude and include, political geography had research questions of interest.

After the Cold War

The final collapse of the Cold War system first mapped out at international conferences at Bretton Woods and Yalta in 1944–1945 had actually begun two decades earlier than 1990. These two agreements imposed geopolitical order, respectively, on world monetary and territorial affairs. Bretton Woods created a fairly stable system of fixed currency exchange rates and Yalta froze the boundaries of postwar Europe and established the spheres of influence of the Cold War. The Bretton Woods Agreement ended in 1971 with the US abrogation of the use of the dollar as the world’s reserve currency and of the US as the lender of last resort. In 1989 Yalta came to a close. The end of the bipolar order started at the urban scale with the downing of the Berlin Wall (1989). It soon extended to the European scale with the reunification of Germany (1990), the collapse of the Iron Curtain and the disintegration of Yugoslavia, and reached its peak at the global scale with the implosion of Soviet Union (1991), officially ending the Cold War in 1992. After an interregnum of a decade in which a unipolar world dominated by the United States seemed in evidence, the geopolitical context began to take on a more complicated cast after 2001. The attacks by agents of the militantly Islamist al-Qaeda Network on the New York Trade Center and the Pentagon building of the US Department of Defense in Washington DC on September 11, 2001 were followed by the
US government’s declaration of war on “international” terrorism. Many Americans reacted by flying the national flag both as a symbol of their threatened identity and as a sign of their mobilization and support of a military response. These events seemed to portend the emergence of a geopolitical context in which states, however mighty, would confront shadowy networks of discontented and fanatic groups following this or that objective, often of a religious or ethnic nature. As this feature of a new world order, and other dimensions of it, such as increased flows of money, goods, and people between localities and regions in it, have taken geographical shape, political geography has also changed its shape in order to deal with what has changed. More recent events signal a very different global political environment from that of 40 years ago. Such events include the financial crisis of 2007–2009 and its pointing towards a restructuring of the world economy around an ascendant Asia and a declining Europe and United States, if also with an apparently enhanced role for national governments; the emergence of global environmental challenges such as climate change, a polluted food chain, and regional water shortages; the rebirth of militant nationalism around the world; and the collapse of many erstwhile states, such as Somalia.

One major difference with the past, therefore, is the complexity of the geopolitical context in which political geography is today being remade as opposed to the relative simplicity of the previous periods. Two background conditions are important in understanding much of what follows. First, a conversation rather than a dialogue of the deaf has begun to emerge between advocates of the three major theoretical “waves” that have swept over political geography since the 1970s. Some examples are provided in the section on Theoretical Perspectives in Contemporary Political Geography. This perhaps reflects the de-radicalization and routinization of what were once seen as totally opposing perspectives as their proponents achieve professional recognition and higher intellectual status. But this trend is also possibly the result of changing times in which geopolitical instability and uncertainty makes the theoretical certainty (even if about the certainty of uncertainty in some varieties of postmodernism) of the different positions increasingly untenable. While an increasing theoretical eclecticism would seem both more necessary and likely, then, as students educated more broadly than past generations emerge into prominence to deal with an increasingly volatile and geographically complex world, in some cases, post-Cold War geopolitical instability may produce an opposing trend: a hardening of perspective boundaries to defend against heresy.

Second, this world to which political geography directly addresses itself seems divided by identity-based (religious, national, local or economic) divisions that take very different geographical forms from the state-based colonial and worldwide-ideological systems that characterized the periods of inter-imperial rivalry (1875–1945) and the Cold War (1945–1991) during which political geography was previously made and remade. On the one hand, the world economy and many political movements (including religious ones) are organized in terms of spatial networks more than territorial blocs, as noted above; on the other hand, claims are made by leaders of dominant states and political movements on behalf of the welfare of huge geographical areas and their populations (such as the West, the Islamic World, etc.) as if these were homogeneous entities. These claims are often informed by a tremendous nostalgia for a period in the past (from medieval Islam to a golden age of state-based capitalism or the European welfare state) and for the “purity” of territories from “foreign” contamination but which now must confront a world in which people of diverse
cultural origins increasingly live intermingled with and interdependent on one another on a practical everyday basis.

In this context, three trends in political geography appear particularly significant. The first is a return to the old physical–human nexus of political geography, now in relation to how human threats to the natural “environment” are mediated by political institutions and movements. As climate change becomes increasingly connected to geopolitics, the risk of a reappearance of environmental determinism seems still far from being removed. So, one task of political geography is to offer approaches that steer well away from this reinvention of what previously proved to be a dead end for the field. The second represents the continuing drift away from state-centrism towards according a central role for “geographical scale” in understanding the geography of power. If the global war on terror has regressively renewed the importance of nation-states, the global financial crisis and the increased importance of the financial markets seem to subtract from conventional understandings of state territorial sovereignty. The third is the question of the nature of politics and its connection to the diversity of human experience. Political geography is finally addressing normative political issues concerning citizenship, democracy, group rights, and the role of intellectuals. These issues are increasingly fraught in a world of increased migration, economic interdependence, jurisdictional disputes over tax evasion, and the revitalization of ethnic and other identities. They are briefly addressed in the section on Emerging Themes in Political Geography.

From naturalized to critical knowledge

Political geography had a history before the term itself came into more widespread use in the 1890s. For example, the seventeenth-century Englishman William Petty’s idea of “political arithmetic” and his Political Anatomy of Ireland (1672) can be seen as historical precursors of late nineteenth-century political geography. In mid-eighteenth century France, Anne-Robert-Jacques Turgot used the term political geography to refer to the relationships between the facts of geography, seen as all physical and human features of spatial distribution, and the organization of politics. It is also apparent that many of the great figures in the history of political thought, from the ancient Greeks Aristotle and Thucydides to the early modern Florentine Machiavelli and later writers such as Hobbes, Locke, Montesquieu, Turgot, Madison, Rousseau, Hegel, and Marx, had ideas about political territoriality and the effects of geographical location and access to resources on conflict and war that can be regarded as basic elements of political geography. They picked up on the practical realities facing political elites and offered their solutions in the context of the historical periods in which they lived. Thucydides’ great work, The Peloponnesian War, concerns the two decades of war between Athens and Sparta (431–411 BCE) and forms the first example of use of the opposition between sea and land powers that later political geographers such as Halford Mackinder (1861–1947) in his famous work on control of the Eurasian Heartland used as a basic organizing principle of world politics.

The founders of modern political geography as such, therefore, could draw upon many centuries of relevant thought to inform their research and writing. But they were also creatures of their time. An important continuity across the early twentieth century is the naturalized understanding of knowledge that tended to dominate geography in general and political geography in particular. The university field of geography as a whole was invented in the late nineteenth
century in part as an offshoot of the growth of national geographical societies devoted to exploration, collecting information about exotic peoples, and the opening up of foreign lands to commerce and/or conquest. The other part of its origins lay in detailed mapping and portrayal of the regions and landscapes of national territories to communicate the material basis of national identity in the burgeoning elementary schools of the era. In this respect, of course, geography was one of a panoply of subjects with ancient roots that were reinvented under their old names to service the needs of statehood and empire-building: from anthropology’s measuring of physical differences between human groups and literature’s capture of national literary genius to history’s telling of distinctive and noble national histories. New fields such as sociology, economics, and political science acquired their own niches in the national service.

An increasingly prestigious and dominant thrust in all of the new “disciplines,” however, was toward a “naturalization” of knowledge claims. This is the tendency to want to explain human and social phenomena largely if not entirely in terms of natural processes, either physical or biological. In other words, they wanted to explain using processes assuredly not of mental construction that lay outside the questionable “human” realm in which values, interests, and identities were all subject to divergent interpretations and hence less amenable to “expert opinion.” The German thinkers who initially dominated the field, such as Friedrich Ratzel (1844–1904), were particularly given to this sort of approach.

Naturalization of knowledge claims had two vital intellectual preconditions. One was the separation of the scientific claim from the subject position of the particular writer (as in the “view from nowhere” mentioned previously). Claims were made to universal knowledge that transcended any particular national, class, gender, or ethnic standpoint. So even as a particular “national interest” was addressed, that focus was theoretically framed by a perspective that put it into the realm of nature rather than that of politics or society. This “view from nowhere” was by no means new, but it was very important to the new university fields in supporting their assertion of expertise and relevance to addressing the problems of the age. The second precondition was preference for the use of arguments drawn from the natural sciences to explain social and political phenomena. Thus, the principle of natural selection as proposed by Charles Darwin filtered down into popular culture and into fields such as geography largely in terms of the idea of survival of the fittest. This not only encouraged organic conceptions of nation-statehood (the state as a type of organism) but also stimulated ideas about racial competition, degradation, and dominion. Much of what passed for Social Darwinism, however, was inspired by the older evolutionary ideas of Jean-Baptiste Lamarck (1744–1829). These were both more open than Darwin’s reliance on variation over extended time periods to the immediate effects of the physical environment on social processes and, crucially, to the impact of “will” or intervention in creating more successful organisms. This allowed for the packaging of seemingly contradictory elements into a single study, such as races as biological categories arrayed according to their superior “consciousness” for which there was no natural basis whatsoever. Such ideas were widely shared among elites, not least the new academic ones, across all of the great powers.

After World War II, such ideas began to fall into disfavor, particularly in the United States. There had always been those who thought such logic limited or fallacious. But particularly in light of the expropriation of the term geopolitics
by supporters of German and Japanese expansionism in the interwar period, undoubtedly influenced in their reasoning by precursors such as Mackinder, political geography was tarred with the brush of a now defeated geographical determinism. Political geography went into something of a slump. It started to recover under two influences in the 1960s and early 1970s. One was the attempt to bring to bear classical concerns with territory back into political geography but without the biometaphysical bias. Crucial to this innovation were such figures as the French/American geographer Jean Gottmann (1915–1994) (e.g., *The Significance of Territory*, 1973) and the Norwegian political sociologist Stein Rokkan (1921–1979) (e.g., “Territories, Centers, and Peripheries,” 1980). The emphasis on state territories as the outcome of historically institutionalized and geographically variegated processes involving sovereignty claims and nationalism was a marked break with previous biological conceptions of territory. Later work, such as Robert Sack’s *Human Territoriality* (1986), both deepened appreciation for this contribution and extended it theoretically in new directions. Another influence came from a revival of electoral geography, first explored by Andre Siegfried in France in the early 1900s, with its focus on geographical patterns of election results and what they said about how people came to vote in the ways they did. Kevin Cox, Ron Johnston, and Peter Taylor were particularly influential in this regard (e.g., *Geography of Elections*, Taylor and Johnston 1979). This direction later developed into a more fully articulated linking of popular political views to the place settings or contexts in which people live (e.g., *Place and Politics*, Agnew 1987). If the approach of Sack and others tended to intersect heavily with positive political theory, electoral geography related more to research in fields such as political sociology and political science. In both cases, however, knowledge was seen as the outcome of human agency rather than of the direct effect of natural process.

Perhaps the most important theoretical departures, however, date from the 1980s when, respectively, Marxist ideas about statehood and capitalism and feminist and postmodern ideas about the discursive construction of geopolitics and political identities acquired increasing influence. In identifying the central role of the state in global capitalism and the social construction of political identities as a process beyond the realm of what was conventionally considered “politics,” these self-consciously critical approaches to knowledge called into question both the restriction of the “political” to a separate sphere and the independence of the state from broader economic and cultural considerations. As a result, political geographers referred to the limits of thinking about states in purely territorial terms (e.g., “The Territorial Trap,” Agnew 1994), argued for thinking about geopolitics in “critical” discursive terms (e.g., *Critical Geopolitics*, Ó Tuathail 1996), and refused to separate the economic from the political, although often with a tendency to privilege the former (e.g., *The Condition of Postmodernity*, Harvey 1989). Explicitly normative critiques of the modern state system and the competitive war machines that drive it, proposals for transnational democracy, and increased attention to looming global environmental disasters for which the geography of existing political arrangements seems ill prepared to respond, lay behind much of the renewal of political geography in the early twenty-first century.

**Theoretical approaches in contemporary political geography**

Three general approaches tend to dominate the field today, even though the lines between
them are not hard and fast. Indeed, more recent innovative research tends to work across them. The first, spatial analysis, involves the correlation of spatial patterns showing how a dependent variable, such as a vote for a specific political party, covaries geographically with various presumed independent or predictor variables such as class, ethnicity, religion, and age. In its stress on empirical data collection and analysis, this is akin to much of what goes for quantitative political science except, crucially, the emphasis is on how geographical variance fundamentally affects the nature of the correlations simply because closer locations are more highly related to one another than to more distant ones. Debates about important theoretical-methodological questions, such as the dependence of individual political preferences and behavior on spatial context (the local economy, whom you talk to, etc.), the “levels” or geographical scales at which variance is more or less concentrated (are differences local or regional?), and the clustering of different political phenomena (e.g., votes, riots, strikes, civil wars, territorial disputes, diffusion of different political institutions) in different places, tend to animate the approach (O’Loughlin 1986, 2003). The popular idea of a United States divided up into so-called red and blue states reflects a primitive version of the notion that where you live can have an independent effect on how you think and act politically separate from simple demographic indicators added up irrespective of how their effects intersect in people’s everyday lives in particular places.

The spatial-analytic approach is exemplified by the theoretical logic provided by Kevin R. Cox and David R. Reynolds in a book published in 1974. After a brief review of the neglect of geographical considerations in studies of power and conflict, the authors identify two factors that they see as leading to an increased concern for space in political studies: the increasing effect of externalities (effects on others who are not parties to a transaction) on people in industrial societies and the adoption of “systems” perspectives in political science that tend to increase the attention given to outcomes of the political process (who wins where) rather than just the political process itself. So, even if politics in the past could be thought of as “spaceless” this is no longer the case. The focus on urban settings and the spatial patterns produced by externalities (think of pollution plumes from factory chimneys and the benefits from living inside the catchment boundary of a high-quality high school) firmly distinguished the approach from previous ones. The logic, however, is not specific to either the urban scale or to such local external effects. It can be extended to interpret national and international relations in similar terms, as can be seen in Cox (1979).

During the 1970s geography experienced something of a turning away from the dominance of this spatial-analytic perspective. In the context of an extended period of political and economic crisis in many Western countries, many political geographers turned towards theoretical perspectives that could encompass the current situation as well as offer fresh understandings of old topics. A revived political economy proved especially attractive. Drawing in particular from Marxist and neo-Marxist writing in political economy, scholars framed political-geographical phenomena in terms of global patterns of uneven development and the processes that they claimed produce them. One variety of this approach that of the world-systems theory associated with Immanuel Wallerstein, proved particularly influential in political geography. This particular theoretical perspective was first popularized in Peter J. Taylor’s textbook (1989) and has led to numerous publications by Taylor and others, such as a paper showing how elections in a country such as Ghana have much to do

POLITICAL GEOGRAPHY
with not just that country’s but also different regions within that country’s commodity trade links to the world economy (Osei-Kwame and Taylor 1984). Theoretically eclectic, drawing its main tenets from such different thinkers as Fernand Braudel, Karl Marx, and Karl Polanyi, this perspective tends to explain most other phenomena with respect to where they are located within a global division of labor (core, periphery, semiperiphery) produced by the historical workings of the capitalist world economy. Under the political-economic rubric, however, can be found a range of perspectives, some adhering closely to an orthodox Marxism privileging the process of capital accumulation pure and simple and others exploring the autonomous powers of states. What joins them together is their view of space as a surface upon which political-economic processes (whatever the specifics) are inscribed and embedded but which is nevertheless essential to the outcome of the processes (e.g., providing the “spatial fix” to the declining rate of profit by moving investment from one place to another, defining the spatial limits to state autonomy, etc.).

This second approach is clearly committed to a political-economic analysis that foregrounds the geography of uneven development at a variety of geographical scales, from the global to the metropolitan. From this viewpoint, over time, the cycles of global economic expansion and contraction are seen as particularly influential in structurally shifting geographical patterns as new places are incorporated and old ones shed their historic roles. By way of example, political parties can be viewed as rising to represent distinctive political-economic interests that tend to cluster in different places because of the history of uneven development.

Finally, a third perspective tends to reject the overt rationalism of the first two modernist ones. In a register that emphasizes the role of the observer, the world is seen as written about rather than discovered or explored. Although writers under this rubric differ in the degree to which they see an “external world” as having an independent reality, the commonality is the rejection of the simple correlation and cause-effect relationships that are the basis for the other two. In a postmodern vein, language and discursive strategies become the focus. So-called critical geopolitics, for example, involves deconstructing the representational and communicative strategies employed by politicians in constructing foreign-policy crises, situations, and wars. From this viewpoint, these things never just “happen.” Some of the narratives help to anchor national identities, whereas others relate to the global agendas of elites in pursuit of this or that interest or policy (Campbell 1992; Sharp 2000).

Over the years many thinkers have questioned the pretensions of “grand theories” and “master narratives” in the social sciences, pointing to how they “over reached” the empirical evidence used to support them. Others have suggested that knowledge was more a product of language, conventions of study, and the relative power of different scholars than based on independent “facts” about the world. As a result, one critique, associated particularly with feminism and postcolonialism, came to emphasize the partiality or “situatedness” of knowing; knowledge is a function, at least in part, of the standpoint or subject position at which a scholar is located, particularly the historical experience of power relative to others and, thus, the capability to tell your story and those of others like you. Another critique came to play up the role of language and writing in offering meaning to readers. From this point of view, the world is entirely as it is written about. In other words, in this poststructuralist or deconstructionist view, writers recycle metaphors and tropes rather than discover new knowledge. Finally, some identified the tenuousness of all claims to tell “stories” about other
people and their places. Even “emancipatory” narratives, stories told to benefit the interests and identities of others, involve a quest for transcendence that disciplines and limits the aspirations of presumed beneficiaries. In this postmodernist view, respect for irony, ambiguity, and the paradoxes of existence remain as the only guarantees against imposing order on others. To the extent that it is possible, one looks for the stories that groups share to understand their self-constructions. By way of example, Sankaran Krishna (1994) uses the metaphor of “cartographic anxiety” to convey how discourses about an Indian “nation” are used to define the borders of contemporary India as a putative nation-state. The “body politic” of India is defined in terms of a physical map that tries to conjure up a “historical original,” a “homeland” that never existed prior to British colonial rule. So, not only are those at the borders of India caught up in an exercise in spatial self-definition that is the essence of nation-statehood: abstracting from history a set of stable, legitimate boundaries that fix the history of the state in place and guarantee it a place in future. The map, therefore, is an attempt at answering definitively the anxiety that comes from being “suspended forever in the space between the ‘former colony’ and the ‘not-yet-nation’” (Krishna 1994, 508).

There are signs of theoretical rapprochement, particularly in some research and writing on geopolitics and the rise of deterritorialized forms of power, such as those associated with so-called global cities and world-city networks. For example, Agnew and Corbridge (1995) attempted to bridge the divide between political-economic and postmodern perspectives by showing how “geopolitical order” – based on trends in the practical political economy of world politics – and “geopolitical discourse” – the ways of seeing and thinking about world politics – interrelate to produce the everyday practices of world politics. They identify different historical periods in which order and discourse relate to one another in distinctive ways to resolve the difficulty that arises in ahistorical accounts of giving priority to one over the other. Their historical geopolitics represents an attempt at engaging with both political-economic and postmodern perspectives rather than favoring only one of them.

In a different vein, an interesting combination of political-economic and spatial-analytic perspectives is apparent in the Beaverstock, Smith, and Taylor (2000) project on world-city networks. Criticizing the state-oriented understanding of the world that dominates both world politics and the social sciences, they propose instead a focus on the world of flows, linkages, and connections among the world’s cities. They bring together the explosive growth of service industries, the increased importance of information technology, and the tremendous development of worldwide direct investment to propose a theoretical framework for mapping what they call “the intercity global network.” Taking a roster of 55 world cities, they show how firms from some set up shop in the others. Empirical analysis of these linkages shows that there is a definite hierarchy to them, with some cities, such as London and New York, currently sitting on top. This pattern suggests that “networks” are beginning to challenge “territories,” if not for the first time, as an organizing principle of global geopolitics. Importantly, however, the authors note that “World cities are not eliminating the power of states; they are a part of a global restructuring which is ‘rescaling’ power relations, in which states will change and adapt as they have done many times in previous restructurings” (Beaverstock, Smith, and Taylor 2000, 132).
POLITICAL GEOGRAPHY

In the study of nationalism and ethnic conflict there has been a different rapprochement between spatial analysis, on the one hand, emphasizing the spatial patterning of conflicts and, on the other hand, postmodern approaches to interpreting the narratives or stories told by the combatants. Good examples come from the joint research of John O’Loughlin and Gearóid Ó Tuathail on ethnic conflicts in Bosnia and the Caucasus (2009) and Gerard Toal and Carl Dahlman (2011) on the return of displaced people to Bosnia in the aftermath of conflict. The latter study is an exemplary attempt at weaving together what can be termed “facts on the ground” about the geography of ethnic “cleansing” and the political debates and different stories of various parties to the resettlement of displaced people including foreign participants. Such research is not classifiable in terms of a single “approach” and is suggestive of future possibilities beyond the threefold categorization that tended to characterize political geography at least into the 1990s.

In the future more “violation” of the theoretical boundaries that have subdivided political geography in the recent past can be expected, if only because the world to which political geography is applied is changing rapidly and in ways that make the previous “intellectual division of labor” increasingly irrelevant. In particular, the emerging “geopolitical order” is one that challenges the fixed spatial claims of spatial-analytic perspectives, the unchanging political-economic imperatives of many political-economic perspectives, and the focus on dominant discursive representations of postmodern approaches.

Emerging themes in political geography

Trying to identify the emerging themes of political geography requires divining from recent trends those likely to emerge into prominence because of their fit with the developing geopolitical context. The purpose here is certainly not to exclude other topics from any future political geography or to offer an encyclopedic inventory of all contemporary trends in subject matter. Rather, drawing attention to the current “making” of political geography allows us to show how active the redefinition and reworking of the field actually is. This is not to simply endorse what we consider the avant-garde and disregard what has been emphasized in the past. The purpose is to illustrate the flow of influence between the contemporary geopolitical context, on the one hand, and the making of political geography, on the other.

In brief, three trends stand out. While the politics–environment nexus may have been disregarded within the geopolitical context of Cold War political geography, there are now significant examples of research that show how political geography can fruitfully engage with different aspects of the environment: from the analysis of how the uneven geographical distribution of resources could contribute to conflicts, to the politics of the environment, to the environmental connection to development issues. Moreover, the increasingly pervasive topic of climate change – perhaps as pervasive today in the public arena as security discourse – has opened multiple connections to questions of development, security, governmentality, scales of governance, and many other issues. Perhaps the possibility of resource wars, environmental security, and some of the many implications of climate security, in their relation to state sovereignty, violent conflicts, and human security stand out.

A second theme concerns the continuing debate about geographical scale and political analysis. The political-economic approach, for example, emphasizes scale as a social construct
incorporating and expressing power in terms of the dominance of one scale over the other, from the global to the local. The poststructuralist approach emphasizes the relational character of scale and how scales come to be represented and enter into different discourses. The two perspectives are not necessarily mutually exclusive in their understandings of scale (and space). As Anssi Paasi (2004) observed, relational networks and bounded territorial spaces may coexist. Therefore, the two approaches can actually complement each other, even if the former privileges direct and the latter intellectually mediated accounts of causation. Four examples can serve to illustrate the importance of considerations of geographical scale in contemporary political geography. The first is the rising role of cities and their transnational networks in climate governance. The second is the role of territorially embedded networks of power in world politics. One particular case is that of the Al-Qaeda network in its association with the September 11, 2001 attacks in the United States. The third is the changing “balance” of geographical scales in electoral politics, showing how “nationalization” is not the singular fate of electoral outcomes but can also involve localization (or fragmentation). The fourth example is of the emerging global system of finance linking major global cities and offshore banking centers in loosening the bonds of national financial regulation and currency markets and, consequently, undermining the close match between financial flows and national territories. In each case, however, it is the intersection of processes across scales rather the singular dominance of one scale (i.e., global networks versus states or the local) that is at work in the analysis.

Third, unlike mainstream political theory’s rigorous focus on concepts such as interests, power, and sovereignty, little or no attention has been paid until recently in political geography to critical analysis of the concepts that the field has been based on: territory and territoriality, space and place, nation-states and statehood, nationalism, sovereignty, and national identity, power and hegemony, geographical scale and networks, violence and boundaries, and so on. The theoretical aphasia finally began to dissipate in the 1970s with critiques of spatial analysis from broadly analytic-philosophical and Marxist perspectives. But only in the 1990s did a critical attitude towards normative assumptions begin to permeate the field as a whole, extending from analysis of conventional views of state territoriality to the language of foreign policy, understandings of power, the violence of national boundaries, and place and political identity in democratic practice (Elden 2013; Barnett and Low 2004). Particularly important in political geography have been deployments of the idea of “governmentality,” drawing from the writings of Michel Foucault. In this locution, power is seen as inherent in all relationships and institutions rather than being a singular property of states or other formal political entities (Allen 2003; Legg 2007). Currently, four areas seem particularly important for the further development of a normative political geography, engaging with debates about actual changes in the world of politics and with important normative aspects of political geography as a field as yet without the attention they deserve: transnational democracy, weapons and warfare, caring for distant strangers, and democratic states and intellectual freedom (Agnew and Muscarà 2012, Chapter 5). The future of political geography remains to be made.

SEE ALSO: Borders, boundaries, and borderlands; Corporations and the nation-state; Cross-border, transnational, and interregional cooperation; Nationalism and geography; Nation-state; Transnationalism
POLITICAL GEOGRAPHY

References


Polycentricity

Christian Vandermotten
*Free University Brussels (ULB), Belgium*
*Belgian Royal Academy, Belgium*

The term “polycentricity” designates an urban system showing a more or less equilibrated functional, and possibly population, pattern. It is opposed to “monocentricity,” where one single city concentrates all the main functions, or at least to very pyramidal and hierarchical urban systems.

Monocentric urban systems are seen to engender such strong territorial disparities that they can even handicap overall national economic development, as was suggested in France by Jean-François Gravier in his book *Paris et le désert français* (Paris and the French desert) in 1947. In France, this was the basis of the policy of “métropoles d’équilibre” (equilibrium metropoles), the promotion of investment in big provincial cities to counteract the weight of Paris, from the 1960s onwards. In a sense, this policy anticipated the promotion of polycentrism, the term being then used not only as a description but as an aim. At another scale, in the Paris metropolitan area, the same concept has been used to found the policy of creating new towns around the French capital. The birth of the concept of polycentricity is thus implicitly embedded in the implementation of regional economic policies, as well as in suburbanization processes and the increased integration of cities and their surrounding hinterlands.

However, the concept remains polysemic, unsharp, multiscalar, and simultaneously descriptive (polycentricity) and normative (polycentrism).

As a descriptive tool, the concept is often used to describe, as more or less polycentric, the pattern of cities and their functional urban areas (FUAs) at the highest levels of the urban hierarchy, or even the business networks linking cities at a continental or worldwide scale. But it can also be used at larger scales for describing more local urban networks, metropolitan regions, or even intra-urban patterns, focusing on the location of business activities (one or more central business districts; possibly “edge cities”) or on retail patterns.

A lack of explicit definition of the scales (and the variables examined, morphological or functional) can lead to opposite conclusions: the German urban system is the archetype of a polycentric one, but considering for instance only northeastern Germany, it is strongly monocentric, centered on Berlin (which at the intra-urban level is a polycentric city); the French urban system is always presented as the archetype of a monocentric one, but it is very polycentric, considering, for instance, southeastern France. In a sense, Christaller’s and Lösch’s classical central places theories describe both monocentric and hierarchical structures as polycentric, depending on the scale at which the patterns are considered.

It also depends on the variables considered: if the United Kingdom is less monocentric than France when considering the population pattern, due to the importance of the conurbations linked to its coal mining and manufacturing past, concentration is at least as high as in France in terms of the location of company headquarters in the capital metropolitan region. Looking at
POLYCENTRICITY

population distribution patterns, Belgium, especially Flanders, presents a very polycentric urban network, but in terms of the functional hierarchy and the international links of the cities, and even the daily commuting pattern, it is much more monocentric, centered on the Brussels-Capital Region, which produces 20% of the Belgian gross domestic product (GDP), with the whole Brussels metropolitan area accounting for nearly one-third.

Different measures of polycentricity have been proposed: the share of the main city in the (population or economic) system; the ratio between the main city and the next cities in the rank order of cities, the slope of the line plotting the ranked distribution of cities, and so on. But, mainly in international comparisons, these measures can be strongly influenced by possible heterogeneities of the territorial definitions on which the data are based. The question arises as to whether the population considered is for the administrative units, for the morphological areas, or for the functional areas; and whether contiguous urban areas should be combined into one or not. Also the sizes of the countries have to be considered: a similar percentage of the population or of economic production or command in the main city of a small country doesn’t have the same meaning as in a large country.

At the metropolitan level, polycentricity can be described from a morphological point of view, characterized by different historical cores and major peaks of rent prices or population densities, or at a functional level, with several main employment cores, important multidirectional commuting flows, and possibly specialization and cooperation between these cores.

As a normative concept, the success of polycentrism grew dramatically during the past decades, in parallel with the development of network theories, in the context of the globalized post-Fordist economy.Implicitly, a polycentric urban system, with each city well related to the others by strong and complex material and immaterial links, is supposed to lead to better chances in the contemporary tertiarized economy. Through better networking in polycentric systems, smaller cities hope to get better opportunities. At the metropolitan scale, a multicore structure is supposed to avoid high congestion costs.

However, cooperation and specialization between cities in polycentric systems is far from being systematic. In many Marshallian industrial polycentric districts or areas, such as Silicon Valley, cities are not specialized but are inserted into specific proximity networks, enabling cross-individual relationships between enterprises. In the reverse case, cities located in polycentric urban areas, such as the early manufacturing or coal-mining basins, suffer generally from weak functional linkages, due to their similar obsolete economic structures: polycentricity is then only an “image on the map.” Cooperation links between cities are often not related to vicinity.

Even considering more or less homogenized scales, there is no clear empirical evidence that national polycentric patterns as such perform better or engender more interregional equity than more monocentric ones, either economically or even taking into consideration sustainable development criteria. It remains very unclear why regional morphological polycentric urban systems should imply more cooperation between cities inside the system, and consequently more economic growth. Linkages between polycentricity and the spatial scope of agglomeration externalities remain badly theorized.

At the metropolitan and intra-urban scale, the environmental situation is no worse in more monocentric systems, and public transport even appears to be better within the latter.
The question then arises: What explains the political success of the normative concept of polycentrism, largely promoted in European strategic planning documents, such as the European Strategic Development Plan (1999), the Territorial Agenda of the European Union (2011), and in the same way by many national planning authorities?

The promotion of polycentrism at European and national scales, on a par with the insistence on globalized networks, appears to be politically in line with the transition from priorities associated with Keynesian cohesion policies towards more neoliberal competitive ones. Such a transition is supposed to improve global economic growth, in terms of GDP, but instead concentrates it spatially. Through more polycentrism, implying a reinforcement of strong nodes within a national (or a European) system, each country or each part of a large country should thus be able to attract some profits from this more competitive (and concentrated) economy.

So, this vague concept allows a European consensus, with each region or big city, and even second-tier cities, hoping to benefit from strengthened competition, even if regional funding is in fact reduced. They then have good reasons for launching benchmarking studies, trying to improve their exogenous attractiveness in the framework of international business networks.

Despite this strategic promotion of polycentrism by the European (and national) authorities, the spatial concentration of economic command functions is not fading globally, nor at the international or national scales. More monocentric or more polycentric urban networks and patterns are historical products, reflecting long-term national and economic histories and associated spatial developments. It is doubtful that European or national public policies could strongly change these patterns, especially in a period of reduced funding for regional policies.

Finally, at the local scale, related to spatial planning, the idea of polycentrism can draw near to the one of “deconcentrated concentration,” referring to a better organization of suburban sprawl areas, for instance by concentrating residential areas around main railway stations. It can also refer to new towns or edge district developments, or even to the promotion of several new business or shopping districts and flagship projects inside the more or less central parts of the morphological area.

SEE ALSO: Central place theory; Network analysis; Regional inequalities; Regional planning: the resilience of an imperative; Spatial organization and structure; Urban geography

Further reading


**POLYCENTRICITY**


Popular culture

Thomas Stieve
Chris Lukinbeal
University of Arizona, USA

Marcel Danesi’s Dictionary of Media and Communications (2009) defines popular (pop) culture as cultural products such as television shows, pop music, fashion, shopping, comic books, and so on, that are geared toward mass consumption. Such different forms of culture and representation have offered fertile ground for new areas of investigation and theorization in geography. The history of geographic research in popular culture shows how paying attention to these cultural products has created different units of analysis and implications for the discipline. Starting in the 1960s with humanism and landscape studies, geographical inquiry started to include place experience in examining popular cultural forms, such as novels and music. Subsequently, in the 1970s and particularly the 1980s, philosophical and methodological debates also arose between those favoring behaviorist and those espousing Marxian interpretations of popular culture. By the 1990s, geographers were expanding their analysis to focus on still more diverse forms of media, and adopting different theoretical lenses in their analysis. Film was now used to analyze the crisis of representation, and television in the understanding of mediated places. The impacts of globalization and commodification have been examined in the new century. The effects of the international trade of cultural items and the influence of commodified music on space and time, for example, have drawn research attention.

J.K. Wright was one of the first geographers to focus attention on popular culture as a valid area of study. In his presidential address to the Association of American Geographers in 1947, he suggested that geographers should not only be concerned with the traditional formal or objective areas of studies, but also “peripheral areas,” such as travel books, magazines, newspapers, fiction, poetry, and paintings. Wright proposed the term “geosophy” to encompass this kind of study of geographic knowledge (Wright 1947). Although this concept was groundbreaking for its time, it took many decades before others would explore Wright’s terrae incognitae, or unknown lands.

Interest in popular culture emerged again in the 1960s–1980s, as humanist and landscape studies sought to move away from regional geographies and the quantitative revolution to the individual, the everyday, and popular culture. Relph (1976) suggested that popular culture affects how a person experiences place, and that popular culture via mass media encourages placelessness as it destroys authentic place experiences. Mass media suggest one way to experience place but are often equated with the dominant culture’s ideology. This theme was later addressed by researchers interested in globalization and the commodification of culture.

In the 1970s and 1980s popular culture was only selectively studied. Even with its rejection of positivistic investigation, the humanist movement’s study of popular culture concentrated only on literature and art. For example, Pocock (1981) pointed out that human geography is an art and social science, but the latter uses objective techniques. Authors in this collection of essays eschewed the positivistic framework.
and suggested approaches that engage the experience of region, place, and the environment through literary criticism. While analyzing the Sherlock Holmes series, Yi-Fu Tuan (1985) created a bridge between literature and popular culture by suggesting that research should focus on “nowness” and “permanence”: “nowness” means that a work of art reflects the social milieu of an era, while “permanence” questions a text’s impact based on its popularity within the era in which it was created, as well as its staying power. Continuing popularity occurs when a work of art speaks to what we call the human condition.

Burgess and Gold (1985) argued that geography does not value research on popular culture and the media because they are considered mere entertainment, and went on to suggest that existing research fell into two categories: those that adopted the American approach and those that privileged the European approach. Where the Americans use a behaviorist approach, the European approach favors a Marxist one. Geographers using a Marxist approach seek to uncover hidden ideologies written into mediated text, which naturalize class inequalities as well as other power relations that reinforce the status quo through popular culture. The authors’ type-casting of the two camps reflects a positivistic versus critical theory divide which, in practice, does not exist. This simplistic distinction ignores humanistic geography and its relationship with environmental perception and behavioral geography. While behavioral geography and environmental perception both use cognitive and social psychology theories, the former relies on the scientific method but the latter does not. This is a key distinction, as environmental perception, especially through transactional theory, was widely used in the 1990s to examine media. Transactionalism focuses on the person in the environment as a dialectic contingent on specific spatial and temporal contexts. In Place Images in the Media, Zonn (1990) used a transactional approach to situate the filmmaker, the medium, and the audience.

In 1994 Stuart Aitken and Leo Zonn coedited the first book about film by geographers. Place, Power, Situation, and Spectacle was published around the time that geography was going through the cultural turn and the crisis of representation. “The cultural turn” was an intellectual shift in which culture was not only brought to the forefront but also, through the crisis of representation, questioned the mimetic role of positivistic epistemologies. Aitken and Zonn (1994) pointed out that film, as a form of popular culture, has been neglected because geography emphasized material over nonmaterial culture studies. In their view, as a result of the postmodern condition, there is little difference between real and reel life, and film must therefore be central to geographic investigation. Cresswell and Dixon (2002) claimed that film is one of the most popular areas for geographic research and teaching.

While film has received lots of attention by geographers, television has not. Paul Adams (1992) argued that television can be considered a complex organization of people, technology, and symbols, where social practices overlap with television depictions. Like a monument that offers a focus to a community, it provides a gathering place for social life. Television also structures participation along a social division based on who can and who cannot appear within the monument. Those who appear on television have power conferred on them. In Place, Television, and the Real Orange County, Fletchall, Lukinbeal, and McHugh (2012) explored television place-making through a case study of three popular US television shows: The OC, Laguna Beach: The Real Orange County, and The Real Housewives of Orange County. More than just representations of reality, mediated places are an inextricable part of our daily lives and directly.
engage the processes of place-making by affecting our perception, senses, and subjectivity. This book showed that mediated places, especially as seen on television, matter.

More recently, geography has used popular culture to understand the commodification of culture. Rosati (2007) researched how the MTV network uses statisticians and psychologists to analyze audience reactions to songs. Like empires in the past, media enforce a monopoly on cultural life by destroying time with space: that is, the cultural product becomes so popular that it controls the time period of its airing and thus eliminates all other activities in space. Whereas Rosati’s focus was on music in the mass media, Miller (2014) analyzed how shopping malls in Argentina are engineered to create a certain affect for consumption by the playing of pop music. Music acts to encourage people to spend time in the mall and be a part of pop culture, regardless of purchases made.

The relationship of globalization to pop culture has become an area of interest to geographers in the past decade. Jakob Nobuoka (2010) studied how the Akihabara area of Tokyo, Japan, has had a worldwide influence. This small place, an area surrounding a railway station in the Japanese capital, is where many high-tech department and outlet stores offer Japanese cultural products: manga, anime, anime collectibles, video games, and so on. It also provides a space where fans and creators can come together to exchange ideas and knowledge about the meaning of these cultural items, which, in turn, influences their production. With this newly elaborated knowledge, new cultural products can be manufactured and sold globally to an ever-receptive market.

In addition to cultural products such as anime, globalization can be seen in music. Connell and Gibson’s Sound Tracks: Popular Music Identity and Place (2003) is often cited as the first book to deploy critical theory to engage music geography. For the authors, music is embedded in local practices that can transform social relations, as well as define and shape settings for social action and identity; however, music can also be thought of as a global commodity, a cultural product that is exchanged and used. Music allows popular culture to resonate through mobility, different scales, and transnational landscapes, for example Andean music played in Times Square in New York City. With new technological, cultural, and economic shifts, music is embedded in social networks as well as local–global exchanges of cultural products and practices. In another study of the globalization of popular culture, Stanley Waterman (2006) argued that, whereas European classical music has been considered the national music of Western-identifying Israel, globalization has also led to the importation of Western commercial music, causing many to worry about an Americanization of the country’s culture. Therefore, discourses of high and low culture (classical and commercial music, respectively) have been mobilized in globalization debates.

SEE ALSO: Film; Globalization; Literary geography; Music

References


Popular geographics

Jerome E. Dobson  
*American Geographical Society, USA*

Geography and public education go hand in hand today as they have done for at least 2400 years. Even the earliest known attempt to write for the public lay firmly in the realm of geographic education. Ancient and modern scholars often credit Isocrates (436–338 BC) with being the first professional writer, founder of the modern approach to higher education, and the most successful teacher of classical rhetoric. What truly distinguishes his writing is that it was the first ever directed to the public … and his topic was urban geography (Marsh 2013).

Today, geography is taught from the kindergarten, through doctoral programs, and to the public at large, but its academic and public fortunes have waxed and waned wildly along the way. At times geography blossomed among scholars and the public as it had done in Greece and Rome during the classical age and in Europe and the Americas from the Renaissance to the mid-1900s. At times it lost favor and slipped into fantasy as in Europe during the Dark Ages and the United States today. There were always distinctions between what was taught in school and what was broadcast to the public, and there were always geographical patterns of where it was taught and where it wasn’t.

Among all disciplines, geography stands out for its vast bodies of phenomenological and methodological information. The amount of geographical information and data required to describe and explain a gargantuan Earth in all its complexity, covering all physical and human entities and processes, exceeds that of any other discipline. To convey even a generalized portion of that knowledge from one generation to the next requires the collective efforts of teachers at all levels of education, but that is not enough. Thankfully, society fills in with popular magazines, atlases, documentaries, and digital sources; some parents teach their children basic map and navigational skills; and other disciplines infuse geography into their own subject matter.

Likewise, the body of knowledge required to analyze all physical and human processes in their geographical contexts worldwide exceeds the phenomenological breadth of any other discipline and presents unique challenges due to geography’s insistence on spatial analysis, scientific integration, place-based research, and fieldwork. Throughout history, geographic methods have been taught almost exclusively in institutions of higher education, but certain exceptions among the public are conspicuous. We are currently witnessing a societal movement of massive proportions. Private individuals are using geographic technology, mapping geographic features of interest to themselves, volunteering to map information of great usefulness to other people in all sorts of crises, and building software for free and open use by vast segments of the public.

Many names – neogeography, GIS 2.0, FOSS4G – have been suggested for the nebulous collection of geographic technologies – popular computer mapping and imaging websites, volunteered geographic information (VGI), public domain software, spatially enabled social networking, crowdsourcing, and so on – that so excite the public today. It is, in fact, the

---

*The International Encyclopedia of Geography.*  
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.  
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.  
DOI: 10.1002/9781118786352.wbieg1188
democratization of geography, automated geography, and geographic information science that has been advocated since 1983.

“Popular geographics” serves well as the covering term for all geographic information technologies that can be understood and employed at little or no cost and without much formal training, in the same sense as “popular science” and “popular mechanics.” Subsumed terms and definitions, gleaned from pertinent publications and websites, include the following.

• Neogeography: Literally, “new geography” is the use of geographic techniques and tools for personal and community activities or by a nonexpert group of users (Turner 2006).
• GIS 2.0, open GIS, open source geo: “a new generation of digital geographic tools … a converging set of technology trends including open source software, decreasing computation costs, ubiquitous data networks, mobile phones, location-based services, spatial database, and cloud computing … dramatically expanded access to geographic data and spatial analysis by lowering the barrier to entry into geographic computing. This expansion is leading to a new set of business models and organizations built around geographic data and analysis” (Disruptive Geo 2013).
• FOSS4G (www.foss4g.org): An organization called Free and Open Source Software for Geospatial (FOSS4G) holds a conference every year somewhere in the world. About 1000 individuals collaborate voluntarily to advance the cause of free and open software specifically for geographic applications. FOSS4G brings together a mix of developers, users, decision-makers, and observers from a broad spectrum of organizations and fields of operation for five days of workshops, presentations, and discussions under the theme of “Geo for All.” The 2015 meeting in Seoul, Korea, drew 562 attendees from 48 different countries.
• Volunteered geographic information (VGI): a georeferenced type of citizen science, a growing area of information gathering. The term was coined by geographer Michael F. Goodchild who, in exploring the world of user-generated content on the web, noted that “a remarkable phenomenon … has become evident in recent months: the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies. They are largely untrained and their actions are almost always voluntary, and the results may or may not be accurate. But, collectively, they represent a dramatic innovation that will certainly have profound impacts on geographic information systems (GIS) and more generally on the discipline of geography and its relationship to the general public. I term this volunteered geographic information (VGI), a special case of the more general Web phenomenon of user-generated content” (Goodchild 2007, 212).
• Public domain software: software in the public domain with absolutely no ownership such as copyright, trademark, or patent, often because it was generated through government-funded activities.
• Spatially enabled social networking: automatically adding geo-coordinates to messages generated by Facebook, Twitter, Instagram, Foursquare, and other social networking software.
• Crowdsourcing: According to Wired magazine’s Jeff Howe, who coined the word, “Simply defined, crowdsourcing represents the act of a company or institution taking
a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers” (Howe 2006).

- Open publication: Open access (OA) means unrestricted online access to peer-reviewed scholarly research. Open access is primarily intended for scholarly journal articles, but is also provided for a growing number of theses, book chapters, and scholarly monographs.

- Participatory GIS: making geographic information technology available to disadvantaged groups in society to enhance their capacity for generating, managing, analyzing, and communicating their own geographic information (http://www.ppgis.net/). The American Geographical Society’s Bowman Expeditions, for example, provide equipment, training, and advice to indigenous communities so they can map their own lands and resources and thereby defend their ownership rights.

- Google Earth (www.google.com/earth): Google Earth is a virtual globe, map, and geographical information program, originally called EarthViewer 3D and created by Keyhole, Inc., a Central Intelligence Agency-funded company acquired by Google in 2004. It maps the Earth by the superimposition of images obtained from satellite imagery, aerial photography, and geographic information systems (GIS). The Google Earth Community, specifically those who register and add a location, consisted of approximately 160,000 individuals in April 2015, when the forum switched to even greater management by volunteers.

- Popular computer mapping and imaging websites and services: Many popular websites serve individuals wishing to locate or map Earth features of personal interest to them. Google Earth is by far the largest. Other sources include vast collections organized by the Federal Geographic Data Committee (www.fgdc.gov) and offered by various federal departments, plus many nongovernmental sites with data derived from them. Trails.com (www.trails.com), for example, offers digital topographic maps developed by the US Geological Survey. Many geographic databases are available for free, as well, from states and local governments throughout the United States. Worldwide, however, many governments insist on cost reimbursement for geographic data.

Today’s popular geographics movement is reminiscent of the eighteenth and nineteenth centuries when gentlemen scholars abounded and ladies of means hosted salons for debating the latest mysteries and discoveries of science, often including geography and Earth exploration. Now everyone can have a little or a lot of the globe to study and enjoy at leisure, earn a living, or help others. Significantly, both movements have included methods and techniques as well as information and data. Men and women often volunteer their time and talents to produce new geographic information, sometimes assembling this face to face, sometimes via Internet, and always digitally connected. The American Geographical Society (AGS), known since 1851 for advocating geography in government and industry, has adopted popular geographics as a major cause. Executives of organizations engaged in this burgeoning new industry now constitute nearly half of
its governing council. Not-for-profit volunteer mapping organizations, with snippets of description from their own websites, include the following.

- OpenStreetMap (www.openstreetmap.org): OpenStreetMap is a map of the world, created by people like you and free to use under an open license. Its 2014 workforce amounted to about 1.5 million volunteers.
- Crisis Mappers (http://crisismappers.net): The International Network of Crisis Mappers (Crisis Mappers Net) is the largest and most active international community of experts, practitioners, policymakers, technologists, researchers, journalists, scholars, hackers, and skilled volunteers engaged at the intersection of humanitarian crises, new technology, crowdsourcing, and crisis mapping, with 8000 members.
- MapGive (mapgive.state.gov): Map data are key to humanitarian and development missions. MapGive helps new volunteers learn to map and get involved in online tasks.
- MapStory (http://mapstory.org): MapStory empowers a global community to organize knowledge about the world spatially and temporally. With MapStory, people of all kinds turn into storytellers who can create, share, and collaborate on MapStories, and ultimately improve public understanding of global dynamics, worldwide, over the course of history.

Thus, large segments of the public are educating themselves earnestly and effectively in pursuit of multiple objectives, some altruistic and some commercial. For-profit companies dedicated to facilitating crowdsourcing and VGI include the ones listed below.

- Spatial Networks (http://spatialnetworks.com): According to its website, Spatial Networks build “geospatial technology products that solve real problems right now. Field data collection including crowdsourcing.”

Academic geographers have long been interested in promoting geographic literacy among the general public, but clearly a different force motivates geographic learning outside academia today. Indeed, this public movement seems to be driven by seething demand among consumers, not by supply offered from traditional producers of geographic knowledge. A nationwide survey administered by the AGS in 2011–2012 uncovered startling new evidence of a strong American constituency for geography. Compared to reading, writing, and mathematics, 66.8% of respondents felt that geography was just as important, an astounding affirmation of geography, considering the usual high regard the public has for “the 3 Rs.” (reading, writing, and arithmetic).

The AGS Geographic Knowledge and Values Survey found overwhelming public support for geography at all levels of education and widespread recognition of the discipline’s value in government and business (Kozak, Dobson, and Wood 2015). The results indicate that Americans predominantly want more geography education to be offered in schools, colleges, and universities throughout the United States. Approximately 9 out of 10 respondents wish they themselves had more geographic education and nearly all want more for their children. To meet this need, respondents favor offering geography courses at every education level, from elementary school to PhD programs in the most elite universities. And they insist that instructors should have formal training in geography. In
summary, geography’s problem is not with the American public but rather with the educational establishment itself, which has been killing academic geography programs relentlessly for seven decades.

Respondents recognize the importance of geography in today’s society. A large portion of respondents say they use geographic knowledge and skills in their everyday lives. They report using geographic concepts and spatial thinking when engaging with the world around them on local to global matters. Respondents correctly identify professions and government agencies in which geography and its skills would be highly useful.

Respondents know that geography is about far more than just knowing their states and capitals. They show sophisticated understanding of the purview of geography, which is defined not by content but rather by its emphasis on spatial reasoning, though many did not know to call it geography. They understand that the discipline covers a wide range of topics, methods, techniques, and technologies. However, their specific knowledge about real-world geography (places and processes) is no better than the dismal levels found routinely in other public surveys of geographic knowledge.

Surely normal human curiosity is a motivating factor for the kind of self-education we are seeing today. Undoubtedly, however, self-interest plays a key role, and nothing draws a crowd like jobs, jobs, jobs. The US Department of Labor, the Guardian newspaper, MSN, Money magazine, and PayScale all report higher salaries, lower unemployment, and faster industry growth for geography graduates compared to other disciplines (Dobson 2014).

Many academic geographers are responding to these issues by interacting with the public more intensely than ever through geographically enabled citizen science, variously called participatory GIS, participatory research mapping (PRM), or public participation in GIS. This movement, now about a quarter of a century old, follows the classic model of didactic geography, with geographers teaching and students eagerly learning. The classroom, however, is the field and the students are dedicated citizens seeking to map their own lands, understand their own resources, and defend their rights from encroachers. Most often, the goal is to help disadvantaged minorities or indigenous communities, but the model can work with any group that is eager to learn the knowledge and skills that geographers have to offer.

Geographers also continue to reach general audiences through their writing. Informative and instructional geographic writing for the public has a long history from Isocrates to today. In the West, for example, Herodotus (ca. 484–ca. 413/425 BCE), Plato (ca. 428–347 BCE), Strabo (ca. 64 BCE–24 CE), and Ptolemy (ca. 90–ca. 168 CE) wrote vividly about real-world phenomena, theoretical geography, cartography, and geographic investigations, and their works profoundly impacted subsequent generations both within and outside academia.

Marco Polo (1254–1324) wrote for the public, sometimes flavoring fact with fantasy. Indeed, fiction can be an effective medium for educating and informing the public. We tend to think of novels as pure fiction, but they can be a means of communicating geographic truths that otherwise may not reach audiences beyond academia. Charles Frazier’s historical novel Cold Mountain (1997) connects readers with the Appalachian Mountains, in a lilting fashion that matches the best writing of landscape geographers. James Michener’s Hawaii (1959) is a pageant of geomorphologic evolution and sequent occupancy, enchanting and informing at the same time. Jules Verne is known for using fiction to publicize his
predictions about the future. He wrote about airplanes, submarines, television, guided missiles, rocket ships, and satellites almost a century before their actual invention. But his writings were also thoroughly grounded in geography, and his first novel, *Five Weeks in a Balloon*, might be seen as the 1860s version of remote sensing, though he tried at first to publish his idea as a serious scientific essay.

Certain popular writers have a talent for writing about geography with a flavor of its own that doesn’t need fiction or fantasy to make it appealing. Harm de Blij, Jared Diamond, Barbara Tuchman, Timothy Egan, Erik Larson, Toby Lester, Charles Mann, V.S. Naipaul, Jan Morris, and John Noble Wilford are masters of this genre. Wilford’s *The Mapmakers* (1981) and Lester’s *Fourth Part of the World* (2009) were bestsellers as well as textbooks for the classroom.

Academic geographers who want to reach the public often favor short articles on timely topics. Not long ago, that meant op-eds in newspapers, and the American Geographical Society facilitated access by editing and distributing manuscripts to more than 1200 newspapers and magazines. Many individual op-eds were accepted and published by a dozen newspapers or more, including major market outlets such as the *Chicago Tribune*, the *San Francisco Chronicle*, and the *Guardian*. Now, it means blogs and tweets too, and a lot more competition in the world of ideas.

Last but not least is the medium of popular magazines. The National Geographic Society’s (NGS) flagship publication immediately comes to mind, but others are prominent around the world. Published in English and 40 local-language editions, *National Geographic* magazine, now majority-owned by 21st Century Fox, has a global circulation of about 6.8 million. Clearly it falls within the long-established public tradition of real-world geography, especially through exquisite photography and engaging writing, even though it doesn’t focus deeply on current geographical topics such as the explosive growth of geographic information systems, through which geography is dramatically transforming society. The NGS has movies, documentaries, books, and websites, and also uses social media and offers opportunities to teach geography to popular audiences on a deeper level. The Royal Geographical Society’s *Geographical* magazine and the Australian Geographic Society’s *Australian Geographic* fall between the popular emphasis of *National Geographic* and the academic tone of the American Geographical Society’s *Geographical Review*.

Looking ahead, popular geographics faces dramatic changes. New digital, network, social media, and geospatial technologies offer expansive opportunities to engage the public even as traditional media decline in impact. These opportunities hold all the traditional challenges of their predecessors but with a new twist. How can geographers keep geography alive and recognized in a web of communication, interaction, and movement they do not control? How can geographers simplify methods, techniques, and technologies to allow a broader public to use their innovations? How can participatory or crowdsourced geography involving hundreds or thousands of participants be designed and implemented with little centralized control? How can this work be pursued with the assurance that the privacy and personal rights of everyone involved will be respected? But perhaps the greatest challenge is how geographers can inspire the public with the same sense of wonder and curiosity, exploration and discovery, that lies at the heart of the discipline, and continue to make geography relevant to all audiences in today’s world.
SEE ALSO: Geography education, informal and for public engagement; Geography education: primary and secondary; Geography in higher education

References


The debate about population and natural resources refers to whether human population growth will increase consumption levels and lead to the ultimate exhaustion of natural resources. The discussion is usually divided between neo-Malthusians who believe that population increases will lead to a collapse in resource availability and market optimists who argue that human ingenuity and price mechanisms will allow societies to devise means to avoid collapse. A third group of analysts, however, is neither Malthusian nor market optimists, and instead look at how social factors control the ability to adapt to resource scarcity, or how scarcity is presented in order to justify different approaches to economic development. In turn, the debate about population and natural resources has also influenced parallel discussions about scarcity, adaptation to scarcity, and access to resources.

**Limits to growth?**

Most discussion about population and natural resources dates from essays written by the British cleric, Thomas Malthus (1766–1834), between 1798 and 1826. These essays outlined Malthus’s projections that population growth would lead to poverty through greater competition for wages and a decline in the availability of resources. At the heart of Malthus’s writings was the principle that agricultural production increases arithmetically (i.e., 1, 2, 3, 4) whereas population grows geometrically (i.e., 2, 4, 8, 16). Consequently, unchecked population growth could be predicted to lead to a sharp decline in population and human welfare as resources run out – often referred to in terms of the classical constraints of famine, pestilence, war, and death.

Malthus’s writings referred explicitly to population. When applied to natural resources, these ideas have been called neo-Malthusian. During the 1960s and 1970s, various influential books adopted this theme. *The Population Bomb* by Paul Ehrlich (1968), for example, predicted famines between 1970 and 1985 because of population growth outstripping food supply. The *Limits to Growth* report (Meadows et al. 1972) added to this concern. It presented detailed analysis of how current natural resources might remain available based on current known reserves, and estimates of resource use based on current growth in their use. Using this formula, the report predicted that natural resources would run out within specific estimates of time, such as, in the case of crude oil, by 1992 (later editions of the book adjusted this date). It concluded, “Short of a world effort, today’s already explosive gaps and inequalities will continue to grow larger. The outcome can only be disaster” (p. 195).

The implication of neo-Malthusian analysis is that human societies need to be aware that resources are ultimately finite, and that social and political unrest will become more common as resources become scarcer. Much research has sought to identify where and how...
local or global limits might be reached. For example, some analysts have argued that societies with scarce resources and poorly developed conflict-resolution mechanisms might be locations where population growth and competition for resources lead to violent conflict (Homer-Dixon 1994). Sub-Saharan Africa, for example, has been cited as a region where long-term environmental changes such as desertification or the emergence of refugees might lead to violence when different groups seek to access a dwindling supply of water or grazing land. Jared Diamond’s book, *Collapse* (Diamond 2005), identified various cases where human societies have, or have not, managed to avoid social breakdown as the result of controlling population growth – or the impacts of population on land cover such as forests in order to control erosion. Diamond, for example, reviewed evidence about the genocide in Rwanda (where ethnic Tutsis killed ethnic Hutus in 1994), and concluded that population growth, with its concurrent demand for land, was a key driving force.

In global terms, the concept of planetary boundaries (Rockström et al. 2009) has been used to identify ultimate limits to human productivity. This concept seeks to define safe operating spaces for humanity by identifying key tipping points of resource exhaustion after which environmental change will occur abruptly and will impact severely on human society. This type of analysis is different to earlier discussions of limits to growth because it reflects the spaces within which resources can be used, rather than focusing on resources alone. The proposed nine zones of planetary boundaries include climate change, biodiversity loss, biogeochemical changes (such as removal of nitrogen and phosphorous from ecosystems), ocean acidification, conversion of land to cropland, freshwater consumption, atmospheric ozone depletion, atmospheric aerosols (or particulates), and chemical pollution (including toxic substances).

**Challenges to limits to growth**

Various analysts, however, have questioned the appropriateness of limits to growth as a way of understanding the relationship of population and natural resources. The first, and most commonly heard, criticism is that the neo-Malthusian concerns about scarcity simplify the operations of markets and price mechanisms when competition for resources occurs. Consequently, so-called market optimists argue that scarcity will fix itself because goods that are in demand will become more expensive over time, and therefore competition for these goods will decrease. Perhaps the best example of this way of thinking was the famous wager undertaken between the Malthusian Paul Ehrlich and the market-optimist, Julian Simon, in 1980 when Ehrlich predicted that the price of chromium, copper, nickel, tin, and tungsten would increase by 1990, and Simon predicted the price would decrease. Ehrlich lost this bet – his presumed scarcity did not occur – in fact, there was a relative oversupply of these commodities during this period. Since 1990, commodity prices have risen, however.

In addition, market optimists have pointed out that demand for resources do not simply reflect demand and supply for known resources. Economic trends also determine which resources are valuable, and human ingenuity can create new responses to scarcity, or new trends in which resources are needed. Accordingly, a famous statement is that “resources are not; they become” – or that society should not focus explicitly on the commodity that offers an economic value, but instead seek which properties are offered by commodities, and
search for alternative sources for these properties. For example, copper wire used to be used for telecommunications. In time, this was replaced by fiber optics, and then again by wireless technology. Consequently, society should not look at the scarcity of current sources of resources, but at multifarious ways that social needs can be addressed. Another famous statement is that “the Stone Age did not come to an end because of a shortage of stone.”

Indeed, the work of Ester Boserup (1910–1999) provided a direct challenge to Malthusian notions of scarcity and collapse based on the transformation of agricultural production in developing countries (Boserup 1965). Boserup used the term adaptation to refer to how different societies regulate scarcity of resources by two key stages. First, societies identify and respond to the threat of scarcity proactively through practices, such as regulating resource consumption, or sharing resources such as agricultural fields. Second, societies can increase agricultural production in more than arithmetic ways through processes of innovation, such as converting sloping agricultural land to irrigated terraces. These forms of social organization and agricultural intensification are effectively different forms of technologies – both social and physical – that can allow societies to live with scarcity.

In response to these criticisms, Malthusian analysts have argued that debate about resource scarcity should focus mainly upon those commodities that are not substitutable. Another response is that the short-term adaptations to scarcity observed by people such as Boserup do not change the underlying reality that resources are finite, and that human societies will ultimately have to live within limits.

But the critics of the concept of limits to growth are not always analysts who believe in human ingenuity or the operation of market price mechanisms. Various analysts have instead looked at the political economy of which societies or individuals experience scarcity, or who can avoid scarcity through adaptive mechanisms and technologies. David Harvey (1974), for example, argued that the relationship of population and natural resources needs to be assessed through the lens of social class and the operation of capitalism. Rather than seeing scarcity in absolute terms (as proposed by Malthus), Harvey argued that questions of subsistence, or consumption, are also determined by the particular circumstances of the mode of production faced by working classes vis-à-vis higher classes. These distinctions can also be applied globally. For example, during the 1974 United Nations Bucharest Conference on Population, the United States proposed a population plan that sought to slow down population growth in rapidly growing developing countries. Developing countries, however, resisted these plans on the grounds that the focus on population numbers alone overlooked the social and economic context in which these numbers are considered problematic, and by whom.

Despite these international controversies, domestic policies in some countries sometimes introduced coercive population controls. In the early 1970s, the government of India introduced a program of forced sterilization that targeted men with two or more children. But the scheme also targeted many poorer, uneducated, men, and political opponents of the government. From 1976, the Indian government adopted a more persuasive approach based on propaganda, plus offers of money, housing, and land, aimed at both men and women. In 1979 China introduced a policy restricting urban populations to one child only, a policy that has been claimed to increase forced abortions and female infanticide where parents prefer sons. During the 1980s under the South African Apartheid regime, official family planning practices frequently injected...
the contraceptive Depo-Provera into women immediately after childbirth as the so-called fourth stage of labor – these injections generally targeted non-white African women, while other races were offered a choice of contraceptives. Concerns about the links between scarcity and population have therefore often been linked to human rights abuses, or opportunistic policies by specific governments (Mosher 2008).

Moreover, it is also clear that scarcity (and the ability to avoid scarcity) is also not uniform. Various analysts within international development have therefore proposed theories of access, or the means by which different societies or individuals benefit from resources or adaptive responses. Amartya Sen (1981) in particular provided a radically different perspective to Malthusianism by demonstrating how large famines during the twentieth century did not occur in democracies, and sometimes happened during times of national food surpluses. Sen used the term endowment to indicate direct access to resources such as food, and entitlements to refer to indirect access such as through family connections or employment. This “entitlements approach” was fundamental in changing thinking about international development away from questions of macro-economic food production alone, and toward an understanding of social vulnerability. For example, Sen pointed out that in India, the people most at risk from food insecurity were landless rural migrants, often of low caste, who lacked the social connections and assets to survive a period of food shortage. According to this approach, limits to growth are not simply driven by population growth or market forces, but by exclusionary social structures that make some sections of society more vulnerable than others. This point reiterates Harvey’s (1974) point above about the role of capitalism, but also acknowledges the role of other social influences such as caste, gender-relations, or other forms of social inequality that can influence vulnerability in association with market forces.

Representations of scarcity

The work of Harvey, Sen, and others showed that scarcity is neither universal, nor separate from social contexts. Consequently, these analysts have argued that there is a need to consider who experiences scarcity and why – plus there is a need to rethink what is meant by scarcity to begin with. Critics have therefore argued that social science needs to understand how definitions of scarcity are made, and how these can be used for political agendas. One paper (Taylor and Buttel 1992), for example, argued that the Limits to Growth report made the mistake of assuming that all individuals in different countries would respond to scarcity in the same way, or that scarcity would lead to social collapse. Instead, this paper argued that there was evidence from the work of Boserup and others that human societies respond to scarcity in different ways, including anticipating and responding to scarcity in adaptive ways. Consequently, this paper criticized the limits to growth model for projecting an image of immanent collapse when it was making unreliable assumptions about social behavior across the world. Scholars have therefore realized that models such as the limits to growth have the potential to represent society in a uniform or overly simplified manner, and that policies based on these models might therefore repeat these clumsy universal assumptions.

Another model of human behavior that illustrates this pattern is the so-called tragedy of the commons. This is another neo-Malthusian framework, which assumes that unregulated access to resources will lead to the degradation of the resource, as each individual will seek to
maximize their own benefits. The “tragedy” is often represented as an inevitable outcome, which should therefore justify the regulation of individualists. But analysts have also argued that this predefined vision of a tragedy is a development narrative – or more of an expectation than a reality – and there is a need to question why people believe in the tragedy rather than acknowledge diverse existing and informal regulations of resource use worldwide (Mehta 2010).

It is also important to ask whether causes of research degradation and scarcity under some circumstances are universally applicable. Diamond’s book *Collapse*, for example, identifies deforestation and soil erosion as key driving forces of societal decline, and that population growth is a key driving force behind both of these changes. Yet, this form of analysis simplifies the contexts in which erosion and deforestation might be problematic. Moreover, Diamond argues that population control is essential for international political order. He wrote: “People in the Third World can now, intentionally or unintentionally, send us their own bad things: their diseases like AIDS, SARS, cholera, and West Nile fever … unstoppable numbers of legal and illegal immigrants … terrorists; and other consequences of their Third World problems” (Diamond 2005, 518). He also suggests, “As a result of those problems [they] are also creating problems for us rich First World counties, which may end up having to provide foreign aid for them … or may decide to provide them with military assistance to deal with rebellions and terrorists, or even have to send in our own troops” (p. 516). Statements like these clearly reflect a worldview that minimizes development intervention or the creation of adaptive capacity in developing countries. They stand in contrast to discussions of scarcity by analysts such as Sen, who focus instead on how social deprivations drive scarcity, and which interventions might prevent scarcity. Consequently, there is a need to consider which political norms drive visions of resource scarcity, and vice versa.

**Governing population and scarcity**

Despite the popular debate about population growth and the possible exhaustion of natural resources, much international development policy in the twenty-first century tries to achieve equitable access to resources, rather than overall controls on population. This process is, however, far from easy. Both population planning and access to resources has to be undertaken sensitively and with regard to the vulnerabilities of different social sectors. For example, the United Nations International Conference on Population and Development in Cairo, 1994, emphasized reproductive rights – especially for women and girls – as a means to maximize individual choice in reproduction. This emphasis promotes individual empowerment and barriers to gender equality as part of the means of regulating population, rather than population control in Malthusian terms.

Similarly, international development programs seek to build institutions – or the shared rules and norms governing resource use – which can be considered part of adaptation to scarcity. These forms of adaptation, however, can be influenced by national priorities for development and controls on local participation. For example, in some countries of the Sahel, governments wish to control local livestock numbers and agriculture because of the feared impacts on desertification or forest loss, but development agencies wish to diversify access to resources and gain more empowerment for local actors or communities to take responsibility for governing resource use. Another area of uncertainty is in the potential role of migration as a way to regulate resource
use, or to adapt to local scarcity. Migration, sometimes across international borders, has been used as a means of responding to scarcity for centuries. But under international law, and under common uses of climate change policy, migration is often resisted and communities need to seek responses to scarcity where they stand.

Achieving successful governance of population and natural resources, therefore, requires flexibility in defining the problem of scarcity and in identifying who is vulnerable to it. This situation has become a lot more sophisticated since the generalized prediction of collapse by Malthus and his followers. But neo-Malthusian ideas still remain popular and often shape public debate about how population growth should be addressed by controlling population growth and movement, rather than the actual means by which vulnerable people might govern access to natural resources, achieve livelihoods, and avoid violent conflict.

SEE ALSO: Climate change adaptation and social transformation; Ecological footprint; Environment and consumption; Environment and development; Environment and migration; Environment and resources, political economy of; Environmental degradation; Natural resources and human conflict; Political ecology; Population growth; Resources and development; Social vulnerability and environmental hazards

References


Population geography

K. Bruce Newbold  
*McMaster University, Canada*

Population geography is a division of human geography that focuses on the study of people, their spatial distributions, their characteristics, and their density. It is the study of the ways in which spatial variations in births, deaths, and the distribution, composition, migration, and growth of populations are related to place. Population geographers seek to understand the society around them, the structure of a population, and how populations change through population movements and processes.

Despite its emphasis on location, place, and space, population geography is interdisciplinary and regularly draws upon related fields in economics, sociology, planning, and anthropology in order to understand population processes. It is also closely related to other disciplines, including demography, which is the study of population statistics and trends, spatial demography, which refers to the formal empirical tools to connect people and places and reflects the role and importance of geographic information systems (GIS) and spatial methods, and social demography, which seeks to explain population patterns based on different disciplinary perspectives.

While the disciplinary boundaries between demography, spatial demography, and population geography have become increasingly blurred, there remain important distinctions. Demography has been traditionally more interested in births, deaths, and marriage, and has typically had less interest in space and migration. Population geography, on the other hand, has taken greater interest in population movements, including residential mobility (which captures local moves), migration (crossing some administrative boundary such as a county or state), and international migration. Moreover, given geography’s emphasis on space, location is particularly important in population geography. But population geography is not just about population movement, and the subdiscipline has grown to encompass topics typically seen within demography, as well as other areas reflective of the diversity of geography. Indeed, the patterns of population growth and change through space and time are fundamental to understanding other processes that geographers frequently tackle, including urbanization, segregation, crime, and health. The study of population geography has influenced other geographic studies, including health, ethnicity, and economic geographies, as well as other disciplines.

Growth and diversity of population geography

Although early examples of population geography and thinking spatially about population processes could arguably include Ravenstein’s theory of migration from the 1880s and John Snow’s mapping of cholera deaths in London, population geography as a formal subdiscipline of geography dates from the early 1950s. Prior to this time, geography as a discipline was typically rooted in descriptive analyses of cultural and physical landscapes. The emergence and recognition of population geography have largely been attributed to Glen Trewartha’s presidential
speech at the 1953 meeting of the Association of American Geographers (AAG). Trewartha called for the increased study of population, arguing that geographical studies should be organized around the cultural landscape, the physical Earth, and population. As Trewartha (1953, 83) commented: “Population is the point of reference from which all other elements are observed and from which they all, singly and collectively, derive significance and meaning. It is population which furnishes the focus.” Following Trewartha’s call for population geography to assume greater prominence within the discipline, the study of population geography grew through the 1960s and 1970s. This growth corresponded to the growing interest in population issues more generally, along with the establishment of the Population Council and the first World Population Conference in 1954, held in Rome, Italy. Like geography more generally, population geography was initially largely descriptive in nature, content to describe a population through its characteristics, including density, distribution, and age structure. Following Trewartha, a number of geographers, including Torsten Hägerstrand and Wilbur Zelinsky, grew the field of population geography to include explanations for the spatial configuration of populations and the geographical analysis of population phenomena. Work by Hägerstrand in the 1970s, for example, situated people in time and space, a concept that population geographers continue to explore. Zelinsky (1966) further separated population geography from demography by highlighting the importance of space and time, and the influence of space on population processes.

Since Trewartha’s speech, geography has typically been divided into physical and human geography, with population geography a sub-discipline of human geography. By the 1980s, population geography had fully developed as an important subdiscipline, reflected in the participation of population geographers in a number of organizations such as the AAG, and particularly in the Population Specialty Group (see: https://popgeog.wordpress.com), the Canadian Association of Geographers (CAG), and the Royal Geographical Society–Institute of British Geographers in the United Kingdom (RGS–IBG). Further work continued both to refine and to expand the focus of population geography. Methodologically, it drew upon spatial analysis, logical positivism, and quantitative methods, reflecting its empirical origins and scope. In addition, population geography was increasingly linked to economic, cultural, and ethnic geographies, as well as regional science. Population geography and spatial demography were also increasingly cross-fertilized with the development of GIS and spatial econometrics, reflective of work by Luc Anselin and Michael Goodchild.

While population geography has not ascended to the importance envisioned in 1953 by Trewartha, and instead remains a subdiscipline of human geography, it has been an important influence upon other areas of geographic study as well as other disciplines. Not surprisingly, population geographers have actively published their work in journals including the AAG publications, the Annals of the Association of American Geographers and The Professional Geographer, as well as the CAG and the RGS–IBG publications, The Canadian Geographer and Transactions, respectively. Beyond geography journals, population geographers also regularly publish in journals including Demography, Papers in Regional Science, Annals of Regional Science, International Migration Review, and Environment and Planning A, reflecting the interdisciplinary nature of their work. As the subdiscipline grew and recognized the need for a more discipline-centered publishing outlet, the International Journal of Population Geography was established in 1994, the first journal to
POPULATION GEOGRAPHY

publish almost exclusively research by population geographers. While population geographers continue to publish in the journal, it changed its name in 2004 to *Population, Space and Place*, to reflect the diversity of disciplinary backgrounds and theoretical perspectives that were being published within the journal.

**Importance of geographic scale**

Scale is an important issue for population geographers. Population processes are influenced by individual preferences, local contexts such as the cost, age, and type of housing, neighborhood composition (race, income, education, age, etc.), employment opportunities, greenspace, and so on, as well as broader-scale processes and events such as the overall economic climate. The scale of the analysis influences the findings, such that conclusions reached at one scale of analysis may not be applicable at another scale. Identified as the modifiable areal unit problem (MAUP, Openshaw 1984), spatial units such as a neighborhood, census tract, county, or state that are used in many studies are often arbitrary and modifiable. More formally, MAUP represents a source of bias resulting from the aggregation of data from one scale to another, with results influenced by the choice of boundaries. Therefore, given that results can be influenced by the scale chosen for the analysis, it is typically more important to understand what scale is the most appropriate for the analysis than to assume one “correct” scale. The problem, however, is that geographic scale can run from the individual and household scale to the global, with all scales being potentially relevant. For example, population movement can occur across a number of scales, with an individual moving across the street, within the neighborhood, within the city, within a region (i.e., state, province), across regional (i.e., between states) lines, or internationally. In each case, the determinants of migration are different. In studying mobility and migration, the number of migrants moving across a comparatively small spatial scale, such as within a city, is always greater than the number moving across larger spatial scales, given that we are more likely to move short rather than longer distances. These differences reflect changes in housing preferences at small spatial scales, and the smaller number of individuals or families that are willing to invest in longer distance migrations that are often associated with changes in employment and labor markets.

But MAUP does not only affect the counting of migrants and the reasons for moving. Population geographers often explain differences in the behavior of populations in relation to local contexts (Entwisle 2007). Local spaces, such as neighborhoods, will have different contextual factors, such as local bylaws and different sociodemographic or socioeconomic profiles, with differences ultimately influencing behavior and outcomes. Fertility, for example, reflects choices made by the individual, but such choices are mediated by partner/family desires and the local community, as well as by broader economic conditions or state policies and programs that may discourage or encourage fertility. The role and importance of studying local contexts are found throughout the geographic literature, including health, crime, and segregation studies.

**Topics in population geography**

Societies are shaped by their populations, along with the processes and characteristics that define those populations. Population movement, for example, links people, countries, regions, and space, and it is not surprising that population
geographers have made substantial contributions to the mobility and migration literatures given the inherent spatial structure of population movements. Likewise, fertility and mortality processes, such as the generally higher levels of fertility and infant mortality in the developing world as compared to the developed world, influence security, economic growth, and politics.

While population geography has traditionally focused on fertility, mortality, and migration, the scope of research has expanded tremendously, with population issues underlying many of the challenges facing the world today, including resource and environmental issues, ethnic relations, politics, and power relations. Population geographers study a diverse range of subjects, including demographic phenomena (fertility, mortality, growth rates, etc.), population change, population mobility (including migration and immigration), occupation structure, and how places react to population change. In the 2006 publication Geography in America at the Dawn of the 21st Century, Patricia Gober and James Tyner authored an entry on “Population Geography,” and identified six research subdiscipline themes: (i) internal migration and residential mobility; (ii) international migration and transnationalism; (iii) immigrant assimilation and adjustment and the emergence of ethnic enclaves; (iv) regional demographic variation; (v) social theory and population processes; and (vi) public policy. Within these six broad themes, population geographers have contributed to migration studies by advancing understanding of direction and distance in migration, who moves (and who stays), net and gross flows at different scales, and modeling economic and social causes and consequences of migration, as well as the characteristics of domestic, international, and seasonal/cyclical movements. In addition, geographers have considered the wider implications of legal and illegal immigration, including the assimilation and adjustment of new arrivals to the destination country; economic, social, and political responses to population movement; and population ageing.

Population geographers continue to explore these themes, albeit with new methods or the addition of greater theoretical depth. Beyond these themes, Gober and Tyner noted additional opportunities for population geography to reach other subdisciplines in geography by engaging environmental geography and GIS. Indeed, since Gober and Tyner’s chapter was published, population geographers have extended their interests to include these realms, as well as other subdisciplines such as cultural or ethnic geography. The emergence or inclusion of new theories, data sources, methods, and tools, including GIS, remote sensing, and spatial analytical techniques, has increased the prominence of space and place in population and other studies, and these methodological and analytical tools are visible in the development of spatial demography, which is essentially a hybrid of population geography and demography. Concurrently, the work of some cultural, ethnic, rural, or economic geographers also constitutes population geography. Indeed, for geographers to consider issues such as the evolution of cities requires consideration of the population trends that accompany these shifts and the characteristics of the population within the city.

Research interests and methods of population geographers have shifted over the years, with such shifts chronicled in the chapter by Gober and Tyner, along with the review by Suzanne Withers (2010). Withers offers thoughts on research needs within population geography, including the need to consider the interdependencies between places and scales, human security issues, the population–health–environment nexus, economic development, and poverty issues. While mobility and migration studies
have remained a common topic among population geographers, they have been increasingly explored through life course and longitudinal analyses, gendered studies, and intergenerational studies. In particular, many population geographers have situated their work in a life course perspective, reflecting and linking events occurring over time related to someone’s life, and enabling the examination of the “careers” or trajectories of people’s lives (Bailey 2009). While life course analysis can be applied throughout population studies, it has been most readily applied to migration research, with work by Plane, Henrie, and Perry (2005) illustrating the diversity of motivations for migration and destination choice from this theoretical perspective, with the authors noting that many of the migrations across the metropolitan hierarchy in the United States are related to migrants in particular age groups, including the preference among 25- to 29-year-olds to migrate into the country’s largest urban areas, while retirees tend to move into nonmetropolitan areas, expressing preferences for lower cost of living or rural amenities (Plane, Henrie, and Perry 2005).

Data and methods in population geography

Reflecting the quantitative revolution that was sweeping geography in the 1960s, along with close ties to demography and the use of large datasets and an emphasis on spatial and quantitative analysis, population geography has been closely linked to logical positivism. Additionally, demography and social demography have also remained closely aligned with positivism and empirical analysis. Given the empirical scope of the subdiscipline, population geographers regularly use tools and measures such as population density, median age, sex ratios, migration rates, birth and death rates, population pyramids, labor force participation rates, and other measures to characterize populations. While many of these are descriptive measures, multivariate techniques including regression analysis, spatial econometrics, and GIS techniques are also frequently used by population geographers.

Population geography is traditionally data-rich. Most nations conduct national censuses containing information on housing, socioeconomic status, gender, age, education, and other attributes. Data sources such as the census or labor force surveys are typically representative of the total population, allowing generalized conclusions to be drawn. Beyond the census, population data are also found in other sources, including other government-run surveys such as the American Community Survey (ACS) or the Current Population Survey (CPS), as well as birth and death certificates or so-called vital statistics that are available through government agencies. In much of Europe, in particular, there is a strong reliance on these vital registry systems which record events such as births, deaths, cause of death, marriages, divorces, and moves/migrations.

Spatial demography and population geography more generally have benefited from the development of GIS and spatial analytical techniques that have offered new ways to view spatial population processes along with the opportunity to incorporate spatial relationships and interactions not available through aspatial analytical methods that had dominated the field for so long. Techniques, including network analysis, location optimization techniques, spatial interaction modeling, agent-based modeling, geodemographics, and spatial econometrics, have placed additional tools at the disposal of population geographers.
While quantitative analysis has not disappeared, it has faded along with the quantitative revolution elsewhere in geography and the social sciences. Instead, population geographers have increasingly turned to qualitative methods to provide new insights into and alternative understandings of population processes. This shift reflects a questioning of the reliance on traditional methods of inquiry, the reliance on positivism and empirical analysis, and the relevance and meaning of large, representative datasets. In particular, it has been argued that empirical techniques could not provide adequate explanations and representative datasets could not provide local or other details. Large datasets and statistical methods could not, for example, adequately capture the complexity of population processes, nor could they reveal the subtleties of individual (as opposed to collective) behaviors. Others, including White and Jackson (1995) argued that, given their dependence on data and processes, population geographers tended to overemphasize events without concern for longer biographical histories and the economic, social, or cultural context within which population processes occur. Moreover, the reliance on data such as the census tended to restrict understandings of the role of gender or age by confining these to specific categories such as male and female, or specific age groups. Other geographers have emphasized the need for ethnographic approaches to population processes in order to reflect on the broader set of features, including life histories and circumstances that could alter choices related to migration or fertility behavior.

Shifting analysis from quantitative to qualitative methods has revealed the details of social and cultural influences on population processes, while removing the emphasis on specific economic events and outcomes. Population geographers have increasingly embraced qualitative methods and regularly use a range of methods including document review, interviews, focus groups, or surveys to collect information. Population geographers have also embraced a range of theoretical perspectives to explore population processes, offering the ability to reveal the complexities of the locale and situate local knowledge. The greater complexity in population research has been revealed through an emphasis on people and their connections to multiple places, where it is argued that people on the move live in the present through reflections on their pasts and the potential of uncertain futures. Such work emphasizes the role of ties to place and the issues, concepts, and contexts that root an individual to a particular place or allow them to move.

Theory in population geography

Population geography’s link to demography has meant that it has long been associated with empirical, positivistic research and theoretical underpinnings, and it has historically drawn heavily on spatial choice theory, along with microeconomic and macroeconomic models. Premised on an assumption of individuals’ desire to better themselves, Ravenstein in the 1880s described the spatial, population, and economic determinants of migration. His work formed the basis for migration theories and analyses for decades. Among the more important generalizations, Ravenstein concluded that migration occurs in a “stepwise” manner; each migration stream tending to have a compensating counter stream; the majority of migrations are short-distance; and the major cause of migration is economic. These often quoted generalizations have stood the test of time and formed the basis of scientific discussion and theoretical development over the years. Lee (1966) advanced Ravenstein’s ideas, creating a framework for
migration analysis that involves the “pull” effects of the destination, the “push” effects of the origin, intervening opportunities, and personal characteristics. Several other theoretical perspectives have emerged within population geography to address questions of migration and mobility studies. These include Zelinsky’s mobility transition theory (1971), which links the types of migration with the development of a region or country; Sjaastad’s (1962) human capital theory, which sets migration as an investment in human capital whereby the costs of migration are balanced against future expected returns measured by lifetime earnings; and Rossi’s 1955 book, Why Families Move, which introduces concepts of the life cycle (now life course) and “residential stress” as primary motivators for migration.

All of these theories drew heavily from economic theories and were based on assumptions including economic rationality and utility theory. Starting in the 1990s, the emphasis on theories that reflected the empirical nature of the discipline drew criticism from population geographers, particularly as human geography more broadly reinvented itself by adopting various critical theoretical standpoints, including social theory, feminist theory, and cultural geography. Given theoretical advances in other fields, numerous calls emerged in the 1990s for population geographers to more directly engage in theory by placing it within population studies. Such calls included White and Jackson’s 1995 argument that the actions of individuals should be embedded within broader social, cultural, and economic contexts, so that population geography could benefit from new theoretical foundations that better situate population processes. Such calls reflected the introduction of critical perspectives across geography, including Marxist or feminist geographies in the 1970s. These interventions marked an important turning point for geography by enabling geographers to explain rather than just describe phenomena, in part by understanding how individuals make and perceive space and make locational decisions. For example, through these perspectives, the residential mobility literature turned its focus towards the impact of neighborhoods on peoples’ daily experiences and their opportunities.

Population geographers have been somewhat slow to incorporate critical perspectives. Population geography has retained its empirical and positivistic roots much longer than other subdisciplines. Despite calls for theory to inform the work of population geographers and some progress in incorporating theory within population geography, its empirical roots run deep and the inclusion of new theoretical insights has remained elusive. Elspeth Graham (2000) offered a number of reasons as to why population geography has been slower to adopt theory than have other areas in geography. For instance, the existence of large datasets continues to encourage empirical analyses with less attention to theory, particularly when the end result is policy engagement, where theoretical frameworks complicate the picture. That is, it is easier to base policy on empirical and generalizable results than to address underlying structural or ideological settings. Likewise, Graham suggests that the complexity and layers of theory complicate their application to population issues. Furthermore, spatial demography has also been largely atheoretical, limiting it to a largely analytical role within population studies.

At the same time, new and alternate theoretical viewpoints are increasingly being incorporated within the subdiscipline, deepening the understanding of human (population) behavior. In addition to life course theory, a wide range of critical theoretical frameworks have been incorporated within population geography studies, including feminist, gendered,
and transnational theories, both drawing from and contributing to other related fields. The use of theory is perhaps particularly evident in migration studies (i.e., Silvey 2004), although this may simply reflect the continued emphasis on migration-related research within population geography more generally. Moreover, other subdisciplines within geography may be better placed to theoretically inform the work of population geographers, resulting in greater collaboration across subdisciplines.

**Conclusion**

Although population geography has never attained the prominence and role suggested by Trewartha, its contributions to both geography and other disciplines have meant that it remains an important subdiscipline, valued for its emphasis on how place and space affect population processes. Population processes have continued significance for society, in terms of their connections to economic and livelihood systems, social structures, consumption, and cultural processes, or even geopolitics. Despite its relevance, there has also been critical discussion within the literature that a continued reliance on empirical and positivistic analyses, along with the failure by population geographers to engage more fully in theoretical developments, has meant that the discipline has become less engaged in geography’s mainstream (Graham 2004). Since these concerns were raised, in the 1990s and 2000s, there has been greater—but not complete—adoption of theory, particularly within migration research, although there is room for further developments as the discipline continues to evolve. As Paul Robbins argued in his Progress in Human Geography/CPESG Netting Lecture at the 2015 AAG meetings in Chicago, current population issues such as the implications of declining global fertility levels and population decline remind us of the continued relevance of geographical perspectives on population. They also provide an opportunity for the revitalization of population geography through the infusion of political economic theory to explain and solve such issues.

Echoing Trewartha’s call for increased attention to population issues, population geographers have also called for greater research in particular areas, leading to the evolution of the discipline, or more specifically the research areas that it touches upon, over time. The future topics of interest and research in population geography are just as critical as its theoretical expansion. This raises the question of what will be the next major area(s) of research for population geographers and whether this/these will reflect critical questions that are relevant to current events and issues. Emergent research has focused on links to environmental issues, including topics related to so-called environmental refugees. Other subdisciplines within geography, including health geography and economic geography, have also incorporated population geography within their research agendas.

While highlighting developing areas of research interest is always tricky, the coming years will offer new research opportunities for population geographers to expand the scope of their work beyond the traditional pillars of fertility, mortality, and migration. Issues such as environmental degradation, climate change, conflict, race/ethnic relations, and health care are potentially important areas for consideration by population geographers, given their inherently spatial implications. Moreover, there is the opportunity for both theoretical and policy development associated with these topics.

The journal Progress in Human Geography offers insight into topical issues tackled within the realm of population geography and the broader
discipline. Over the past decade, authors have identified a number of areas that could be further explored by population geographers. Writing in 2009, Adrian Bailey called for greater research into the role of the life course in population decisions and how population geographers could explore the interrelationships between population and climate change, including notions of vulnerabilities and vulnerable populations, risk, and policy discourse. Graeme Hugo (2007) argued for the need for population geographers to explore topics within the Southern Hemisphere, a geographic focus that has been less represented within the literature despite the size of its population, along with various fertility, mortality, and migration issues (i.e., international, rural–urban, population development). Hugo argued that there is substantial room for population geographers to add additional insights and policy discussion based upon population processes and phenomena in the Global South. More recently, James Tyner (2016) identified death, mortality, and vulnerable populations as areas that need additional attention from population geographers. Separately, Coulter, Van Ham, and Findlay (2015) identify the need to rethink short-distance mobility from a life course perspective, with an explicit focus on individual relations through time and space, while also considering the context and structure that define mobility.

It is interesting to note that the majority of these authors call for further work associated with mobility, migration, and immigration – the areas that population geography has typically focused upon – while there seemingly remains less interest in fertility and mortality. Given that population processes have implications for finance, property values, marketing, and other day-to-day events, population geography underlies many broad-based discussions, meaning that it remains a fertile area for study, research, and discussion.

**SEE ALSO:** Aging; Demographic and epidemiological transition; Fertility; Life course; Migrant settlement; Migration: forced; Migration: internal; Migration: international; Mortality; Population growth; Residential mobility

### References


POPULATION GEOGRAPHY


Further reading


Throughout much of humanity’s history, world population was small and it grew slowly. Although fertility rates (which express the average number of children born to a woman over her reproductive years) were high, unpredictable food supplies, short life expectancies, famine, war, and disease meant that the population grew slowly and the global population remained small. Even by the early 1700s, the estimated global population was just 500 million people. Yet, by mid-2015, it had exploded to nearly 7.3 billion, with ongoing growth assured (see Figure 1). These rapid changes raise a number of critical questions: How did the world’s population grow so quickly and in such a comparatively short period of time? How is the global population distributed? What are the implications of a large and growing population?

Beginning in the mid-1600s, the era of slow and stable population growth ended. Roughly corresponding to the start of the Industrial Revolution in Europe, death rates started to fall and life expectancy slowly increased with improvements in commerce, food security and production, and nutrition. At the same time, birth rates remained high, reflecting long-standing needs for larger families. This difference between birth and death rates allowed the population to grow, such that by 1800 the world’s population was approximately 1 billion, growing to 2 billion by 1930. The pace of growth continued to increase through the twentieth century, with global population passing 3 billion by 1960, 4 billion by the mid-1970s, and 5 billion in 1987. Just 12 years later, in 1999, global population reached 6 billion, and it surpassed 7 billion in 2011. Despite falling fertility rates and slowing population growth rates since the 1960s, current projections made by the Population Reference Bureau and other groups estimate the global population will reach 9.8 billion by 2050, with some projections estimating a total world population as large as 10.7 billion in the not too distant future beyond 2050.

Population growth patterns and rates are not the same across the globe, meaning that there is a clear geography to the world’s population growth both historically and in modern times. Many Western nations, including the United States, Canada, and several European countries, experienced their most rapid growth during the Industrial Revolution, with the population of Europe doubling between 1800 and 1900. Immigration to North America in the 1700s and 1800s resulted in rapid population growth during this same time. In the developing world, population growth was typically slower, although countries in the developing world already represented the largest proportion of the world’s population.

Much of the current population growth is occurring in the developing world where fertility rates and growth rates remain high, with its population projected to grow from 5.9 billion in 2015 to 8.2 billion in 2050. At the same time, the population of the developed world will be largely unchanged. Regions with rapid population growth include portions of Africa and Asia and countries with large populations or high fertility, including India, Indonesia, Pakistan, and the Philippines. Fertility rates in
India, the world’s second most populous country, remain high (2.3) and the population is growing at a rate of 1.2%, meaning India’s population continues to expand, with projections that India will overtake China as the world’s most populous country by 2028. Conversely, although China is currently the world’s most populous country with a 2015 population of 1.372 billion, it is growing at just 0.5%, reflecting government policies restricting the number of children per family. In sub-Saharan Africa, fertility rates remain high and average 5.0, resulting in a population growth rate of 2.7%. This means that its population can double in just 26 years.

Population growth rates are much lower in Central America, South America, and the Caribbean, although populations continue to grow because fertility rates remain in excess of 2.1, the “replacement rate” that is required to exactly replace the current generation. In comparison, the fertility rate in North America is just 1.8, combined with net migration gains this corresponds to a growth rate of just 1.0%, meaning the population will take 70 years to double. Europe has some of the lowest fertility rates and lowest growth rates, with some countries, including Germany and Portugal, experiencing population decline, while many other countries are experiencing negligible population growth.

The shift from slow and steady population growth to rapid population expansion can be largely explained by the demographic transition theory or the shift from high birth and death rates (pre-transition) to a low fertility and mortality regime. Initially, both birth and death rates are high and largely cancel each other, resulting in slow population growth. With development, modernization, and improved health care, death rates decline but fertility remains high, resulting in a period of rapid population growth given the difference between birth and death rates. Ultimately, birth rates decline as well, and population growth stabilizes. In the demographic transition theory, the two most important determinants of population growth are (i) level of fertility before the transition occurs, and (ii) the amount of time between the initial decline in mortality and the delayed decline in fertility. In other words, both higher initial fertility and/or increased time between the decline in mortality and the decline in fertility are associated with greater population growth.

Although the demographic transition theory can be applied globally, it is important to note that the timing, pace, and triggers of changes in fertility and mortality vary from country to country and region to region. The demographic transition theory has also been widely criticized for its Western focus. In the developed world, transitions to lower mortality and fertility occurred simultaneously with public health improvements that led to declines in infant mortality rates and increased life expectancy. The demographic transition has also not yet
completed in many parts of the developing world: mortality rates have generally fallen given improved nutrition and health care, yet fertility rates remain above replacement level in many locations in the developing world, averaging approximately 2.6 children per woman in the less developed world and 4.4 in the least developed world.

Consequently, population growth in the developing world has been aided by higher pre-transition fertility rates in the developing world and typically slower reductions in fertility relative to the developed world. This is because, in contrast to mortality, fertility choices are very personal decisions shaped by societal structures. Therefore, the pace and degree of declines in fertility will vary from country to country due to effects including different social, cultural, and religious expectations (for example around the number of children and the preference for male heirs), lower literacy rates, the likelihood that children will survive into adulthood, lower female participation rates in the workforce, and the availability, acceptability, and cost of family planning programs.

Although world population growth rates have started to decline, population growth remains a concern, in large part due to the continued population growth observed in Africa and other populous nations. Recognizing the potential for economic, social, and political problems of rapid population growth, some countries have moved to actively reduce fertility rates, with China being the most well-known. Beginning in the late 1970s, China’s leadership moved to drastically reduce population growth with its “one-child policy.” Although restrictions on family size were relaxed in 2015, families that participated in the program typically received cash bonuses, preferential school admission, and improved housing, while there were corresponding disincentives for families that had more than one child. While successful in reducing fertility rates that exceeded 7 at the start of the policy to just 1.7 today, the policy has been controversial given the state’s expectations that the policy would be followed and its use of coercion in ensuring that it was. China’s one-child policy has also led to an imbalance in the sex ratio at birth. Typically 105 males are born for every 100 females but the sex ratio in China has increased to 118 (Gilles and Feldman-Jacobs 2012), with even higher ratios in some parts of the country. The excess number of male births as compared to female births has been associated with the one-child policy and the preference amongst Chinese families to have a son (Hsu 2008), while the broad availability of ultrasound imaging to identify the sex of the child during pregnancy has facilitated selective abortion.

India has also pursued population programs aimed at reducing the country’s high fertility levels with a “two-child norm” policy. It has promoted lower fertility rates through family planning programs, economic incentives, increased accessibility to contraception, and sterilization. India’s program, like that of China, has been criticized for being heavy-handed and has resulted in skewed sex ratios (110), with more males than females, given similar preferences for male children (Gilles and Feldman-Jacobs 2012). In both India and China, the “missing women” problem is critical (Gilles and Feldman-Jacobs 2012), with potential implications for civil order if men are unable to find partners.

Many developed countries are faced with the opposite situation: low fertility rates, aging populations, and declining populations. Europe’s fertility rate is just 1.4, and its growth rate is stagnant (0.0%). Lower birth rates have meant an aging population: Germany’s share of people over 65 is 21%, and the proportion over 65 in Japan is 26%. Even in the United States, where the fertility rate is comparably high, the US
Population Growth

Census Bureau reports the population is aging. By 2030, over 20% of Americans will be over 65, compared with just 15% in 2015.

Concerns with housing, health, and the provision of other services for an aging population have prompted countries to promote fertility through pronatalist policies to increase fertility rates and ultimately population growth. Such policies often include financial incentives, such as paying women to have children, tax incentives, paid maternity or paternity leaves, free or reduced-cost childcare and/or restriction to contraception and abortion services in attempts to promote fertility. While several countries, including France, Australia, and Russia, have implemented pronatalist policies, there is typically only a small increase in fertility levels (Haub 2008). In Russia’s case, fertility rates had dipped as low as 1.2 in the late 1990s. Following the introduction of pronatalist policies, fertility rates had increased to 1.8 by 2015.

While population decline is cause for concern, the ongoing concern with high fertility (and consequently population growth) stems from discussions linking population growth with resource consumption and particularly food or the ability of the world to feed itself. Written in 1798, Thomas Malthus’s “Essay on the Principle of Population” has long shaped the discussion around population growth and food security. In his essay, Malthus argued that food supply would increase in a linear fashion (1, 2, 3 …) and that population would increase geometrically (2, 4, 8 …). In his scenario, population would ultimately exceed agricultural output unless population growth was “checked” by famine, plague, or war, which would ultimately decrease the population. Alternatively, population growth could be controlled through “preventative checks” with individuals imposing their own limits on reproduction. Given that Malthus saw little hope for preventative checks, he forecast a future of population decline and widespread poverty.

Malthus’s ideas have continued to generate discussion within the broader literature, and alternate viewpoints have emerged. Following Malthus, neo-Malthusians such as Paul Ehrlich have argued that limited resources place limits on the growth of the human population and consumption, and point to food shortages as evidence of the finite means to produce food. Murton’s 1999 analysis of population growth, poverty, and agriculture in Kenya demonstrated increasing disparities and a loss of self-sufficiency, particularly amongst those dependent on agricultural labor markets. Alternatively, “distributionists” such as Karl Marx argued that people were poor due to the organization of societies, and that an equitable distribution of resources aided by technology would allow unlimited population growth, meaning that societies needed to be reorganized (Harvey 1974). “Economic optimists,” including Ester Boserup, see few limits to population growth and prosperity, provided the economic system and market mechanisms work correctly. Boserup (1965), for example, argued agricultural methods depend on the size of the population, whereby a scarcity of cropland can stimulate labor specialization, increased productivity, and changes in agricultural practices. That is, larger and denser populations are associated with greater food supplies able to support a larger population.

Regardless of the standpoint, the basic question is whether the world has sufficient resources – including water and food – for a growing population. There is clearly anecdotal evidence that resources, including clean water and agricultural land, in areas with rapid population growth are stressed and limited (De Souza, Williams, and Frederick 2003). Likewise, there is mounting evidence that the global food
supply is precarious, with disruptions to supply chains, a poor growing season, or increasing fuel costs effectively reducing the food supply (De Souza, Williams and Frederick 2003). Climate change could further jeopardize global food security as rising temperatures and shifting precipitation patterns alter agricultural production. Moreover, there is emerging consensus that rapid population growth reduces economic growth, increases poverty, weakens investment in education, decreases household savings, and degrades resources. Population growth and limited economic growth also appear to be self-reinforcing (Homer-Dixon 1999), resulting in greater disparities across populations.

The broader concern associated with population growth is the potential for conflict over scarce resources, including water supplies. The threat is potentially the greatest in the developing world, where population growth is the greatest and resources are the scarcest. Homer-Dixon (1999), for example, argued that future conflicts are more likely to be related to environmental degradation, population displacement, or disruption to economic systems.

Given that population growth rates have been declining since the 1960s, some authors have argued that population growth is no longer a concern. It is argued instead that declining populations and population aging are the new population problems (Longman 2004). Yet, the young age profile of much of the less and least developed world means that their populations still have to enter their reproductive years. Even if fertility rates decline, population momentum will ensure the world’s population continues to grow, and failure to recognize the potential for future growth is problematic. As noted earlier, the global growth rate is 1.2%, meaning that the world’s population could double in just 58 years. While this scenario is unlikely, some growth is assured. Even as fertility rates have dropped in some regions, they remain high in the least developed countries, including much of sub-Saharan Africa where the 2015 fertility rate was 5.0. Moreover, young populations translate to a population that is entering or already in its prime reproductive years. Consequently, global population growth will continue, and it remains unclear whether the world can accommodate this growth.

SEE ALSO: Aging; Demographic and epidemiological transition; Fertility; Mortality; Natural resources and human conflict; Population geography; Population and natural resources

References


POPULATION GROWTH


Further reading


Website

Population mobility and regional development

Tony Fielding
University of Sussex, UK

This entry is about the many relationships that exist between regional economic development and population migration. At one level, these relationships are fairly straightforward—a region with high wages experiencing rapid growth is likely to attract migrants from regions that have low wages and are in decline. But, dig deeper, and one finds many unexpected outcomes—these relationships are often paradoxical and are not as simple as they seem.

Population mobility comes in many forms, only a minority of which will be covered here. Figure 1 shows some of these mobilities plotted on a graph with axes for distance moved and duration of stay. Broadly speaking, this entry will focus on those who move further than distances (often judged to be around one hour’s traveling time—say about 50 miles) over which people are likely to commute rather than move house, and on those who stay in the new location for longer than about six months, that is, longer than one would stay if the move were a regular seasonal migration such as for agricultural work or spending the winter months in a warm, sunny environment.

Figure 1 is, however, very useful in pointing out the fuzzy boundaries of migration between city-regions. On the lower boundary of the shaded area are the forms of mobility that lie on the border between definitive migration and temporary stay. Sometimes the intention of the person or household is to move permanently to a new city or region, but, in the event, things do not work out as expected, and the stay is cut short. More commonly, however, the shift is in the other direction. Without the intention to do so, a person’s temporary stay in a new place, perhaps to visit a relative or to do a temporary job, turns into a permanent move as friends are made, job opportunities realized, and an almost imperceptible embedding of the person in the new location occurs.

On the right-hand boundary of the shaded area in Figure 1 are the forms of mobility that lie at the border of internal (intranational) and international migration. Some countries are so large in population and area, and contain so many socially and culturally different city-regions, that an internal migration can take on the characteristics more typical of an international migration; this is especially true of China where the household registration (hukou) system means that a person’s social rights (to health, education, housing, etc.) are associated with their official place of residence, and changing that (e.g., by moving from a rural area to the provincial capital or to one of the larger cities in the country) has been difficult if not impossible for most people, especially in the past. Conversely, an international border quite frequently cuts through the territory of a functioning city-region; a striking example of this is Singapore, where the island-state of Singapore increasingly constitutes only the core section of the city-region, with the state of Johor in Malaysia and the province of Riau Islands in Indonesia forming the outer sections of that city-region.
Finally, on the left-hand boundary of the shaded area are forms of mobility that involve long-term but relatively short-distance moves. Many such moves combine housing and neighborhood considerations with those of employment location. For example, the solution to the problem of a long commute is, for some people, to move nearer to work; for others, especially those in a two- or multiple-job household, one person will work away from home, sometimes far away, coming home only at weekends or public holidays. Certain types of work tend to produce these kinds of arrangements (e.g., serving in the armed forces, working on construction sites or...
offshore, or working in domestic/care occupations), but dual-career professional households in high-income countries are also increasingly taking this form, resulting in the acronym LAT (living apart together).

The concept of regional development is taken to mean differential urban and regional economic change, measured by indicators such as gross national product (GNP) and GNP per capita, unemployment rates, and standardized mortality rates, at local, city-region, and subnational scales.

Absolutely central to an analysis of the relationships between population migration and regional development is the appreciation that the direction of causality can take place, and typically does, in both directions; that is, the rapid growth of a region’s economy will usually result in net in-migration, and high net migration gains will usually lead to rapid growth in a region’s economy. This creates the strong possibility that there will be positive feedback loops linking, for example, migration and employment, and that processes of circular and cumulative causation will be at work (Myrdal 1957). A more recent microeconomic version of the path dependency arguments developed by Myrdal is represented by evolutionary economic geography. A representation of this approach, when applied to migration, is provided by Figure 2.

Starting in the lower left-hand corner, regional economic growth will be largely determined by national forces such as technological change and changes in consumer behavior, mediated by both regional productivity change (investment and innovation) and the inherited regional employment structure. This employment structure, together with the volume and direction of employment change (new jobs created/jobs lost), will largely determine the region’s in- and out-migration flows (Boyle, Halfacree, and Robinson 1998). Interregional migration will affect population growth not only directly through the net migration balance, but also indirectly through the age composition (and other characteristics) of the migrants. As regional population growth (or decline) occurs, the region gains (or loses) critical mass for its selection as a suitable location for investments related to market size or labor pool size. Finally, interregional migration greatly influences regional economic growth through the selectivity of migration. Since migration is generally positively selective of people who have ambition and courage and qualifications and skills, possess social and cultural capital, tend to be flexible and adaptive, and are in the young adult age group when their economic potential is at its height, a net migration gain can bestow an enormous advantage to a region’s economy. Conversely, a net migration loss can have all the opposite effects—a brain drain, an aging population, and a loss of commitment to adjust or change. Interregional migration can also affect regional employment change; migrants may well only take up those jobs that are attractive—that are secure and that offer good pay and prospects. The rejection of the other jobs will lead to either the stemming of employment growth or the jobs being filled by foreign workers. This interdependence between domestic interregional migration and international migration has become of very great significance in high-income countries around the world (see, e.g., Fielding 2015, ch. 10, on the new immigration model applied to East Asia).

To emphasize in this way the bidirectional flow of causality between population migration and regional development, and its potential to result in patterns of cumulative growth and decline, is not to deny the strong logic of neoclassical economic approaches to these relationships. Regional income convergence in the United States, for example, has been explained as resulting from the combination of the migration
of labor from low-income, high-unemployment regions to high-income, low-unemployment ones (thereby raising the scarcity of labor and therefore wages in the former and reducing labor scarcity and wages in the latter) and of the flow of capital in the reverse direction (to take advantage, of course, of the lower costs of labor in the low-income region) (Borts and Stein 1964; for a useful survey of migration models in this tradition see Plane and Rogerson 1994). The problem is that many of the expectations about interregional migration based on the neoclassical economic approach are simply not borne out by the facts: paradoxically, the unemployed are, generally, rather immobile, whereas homeowners on the other hand are surprisingly mobile; on balance, high-income cities often lose interregional migrants rather than gain them, while low-income rural areas often gain migrants rather than lose them. More generally, those who have the greatest need to migrate (because they have the lowest incomes and live in places with the worst economic prospects) tend to migrate the least, while those who have the lowest need to migrate (because they are in secure, well-paid employment and live in places with the best prospects) tend to migrate the most.

Much of the remainder of this entry is concerned with providing explanations for these anomalies while focusing on just a small part of Figure 2 – the one-way causal link between regional economic growth and regional employment change on the one hand, and interregional migration on the other. But, before this, a note of caution is appropriate. In the past, say 50 years ago, the link between migration and
employment was exceedingly close – very few could afford to live away from where jobs were located. Today, there have been two changes, especially in high-income capitalist economies, which have weakened this link. The first is the decline in the returns to labor relative to the returns to capital. This means that a significant proportion of people – those living off their investments – can choose to live in places that are geographically distant from the places where the economic activities on which their wealth depends are located. Within countries, this can result in environmentally favored regions having wealthy in-migrants and net migration gains even though these regions are renowned for their low wages and high unemployment. In the United Kingdom, the classic example of this is the southwest region and, above all, the county of Cornwall (Fielding 2012). Internationally, the benefits of this separation favor Caribbean island countries, and cities like Dubai and London. The migration of wealthy foreigners to London has knock-on effects on internal migration within the United Kingdom. By raising housing and other prices in London, these migrants make it more difficult for ordinary families (such as those of teachers and nurses) to migrate to the capital region, and at the same time encourage others to cash in their London properties to enjoy a high standard of living elsewhere.

The second change that has weakened the link between migration and employment is communications technology. It has become possible to work physically apart from your work colleagues while remaining as much in contact with them as if they were in the office next door. As part of this tendency, web conferencing and online meetings have become a part of many people’s working lives. This facilitation of the physical separation of home and workplace is of great importance to professional men and women in dual-career households; it has the effect of injecting flexibility into situations that might otherwise threaten career development and relationship commitments.

We return to the economic drivers of interregional migration. The first step is to categorize the many economic processes that affect interregional migration on the basis of the time spans over which they operate. This step is essential; far too often one finds economic analyses of migration floundering due to the unproblematic conflation (e.g., in econometric models) of very different kinds of migration processes; specifically, business cycle-related processes are mixed up with regional economic restructuring processes, which in turn are mixed up with underlying and only very slowly changing patterns of regional inequality.

Figure 3 shows all the different economic processes affecting interregional migration. It has three layers or levels: the top layer consists of economic changes that occur frequently and quickly; these are typically related to the business cycle. The label adopted to describe these processes is “conjuncture”: this word was chosen because there is no term in English that matches exactly the economic use of the term conjuncture in French. Conjuncture means the economic circumstances of the moment, which, in turn, usually means the position of that moment within the business cycle.

We know that there are close connections between interregional migration flows and the business cycle; these links operate through both the labor market and the housing market. Four stages can be identified. In stage 1, the economy is in its “rise” phase: confidence is growing, firms are investing, people are moving to new jobs and are thinking about moving house. For a high-income region within a national space economy, the typical migration responses to this stage in the business cycle are: in-migration low but rising; out-migration high but falling; with
house prices rising, people put their houses on the market and plan to move to new places; with unemployment falling, people contemplate taking the risk of changing jobs, often with the implication that this will provoke moves to new places; the effect of these changes, which can happen quite suddenly, is a net migration gain for the high-income region.

In stage 2, the economy reaches the peak of the business cycle; investment levels and job opportunities remain high; there is a lot of construction of dwellings and of commercial property; and unemployment is low. Turnover in both the labor and housing markets is high. In-migration peaks high, out-migration bottoms out low, so a high net migration gain is the result.

In stage 3, the economy is in its “fall” phase: confidence is declining, firms are cutting back, people are reluctant to move to new jobs or to move house. House prices begin to fall and unemployment begins to rise. The migration responses in a high-income region are: in-migration is high but falling; out-migration is low but rising; and net migration shifts from net gain to net loss.

Finally, in stage 4, we reach the trough in the cycle – recession. Unemployment goes up sharply, house building grinds almost to a halt, and confidence in the economy plummets. People and firms tend in these circumstances to hunker down, that is, to reduce their activities, expenditures, overheads, and commitments so that they can weather the economic storm and be in a suitable position to benefit from the upturn when it comes. The migration responses to recession tend to reflect the low turnovers in both the labor and the housing markets, so rates of migration flow are at a very low level: in-migration bottoms out low; out-migration peaks high; and net migration loss is to be expected.

However, there is a contradiction in this account. During a recession mobility rates tend to be low, and yet high net out-migration is expected. In times past, when business cycles tended to be more local, perhaps even differing from one region of a country to another, it would

Figure 3  Population migration and regional development: the Fielding three-layer model (modified from Fielding 2012).
have been possible for the out-migrants from a region in recession to have gone to a region that was booming. But today, in general, that is not what happens; globalization tends to ensure that there is a synchronization of business cycles, so recession tends to affect a whole country, indeed many countries, at the same time. There is, therefore, nowhere that is booming where those who are adversely affected by recession can go.

Furthermore, in the real world, business cycle migration effects tend to be modified or even covered over by migrations that are obeying a different economic logic. This point can be illustrated by looking at the migrations to and from the London city-region (southeast England) since the mid-1970s. Migrations to and from London tend to reflect the business cycle in that they go up during boom times and down during a recession. But as a high-income region with low unemployment rates, London should logically be a region of net migration gain from the rest of the United Kingdom, and it should be gaining most by migration during the boom years. This is most assuredly not the case. Most of the time London is a net loser of population through its migration exchanges with the rest of the United Kingdom, and it tends to lose most during boom times and least during a recession (indeed, it has sometimes gained by migration during a recession). This paradox can be explained by focusing on the difference in the dominant stage in the life course of migrants toward the London region and those moving away from the region. Those coming to London from the rest of the United Kingdom are, of course, very diverse in qualifications and social backgrounds, but they tend to be young adults from middle-class backgrounds straight out of university; they are single and consume very little housing space. In boom years, they may go directly into well-paid employment; in a recession, they enter the same or similar jobs but at a much lower level (perhaps even as unpaid interns). In other words they come to London whatever the state of the economy – whatever stage in the business cycle the economy is in. Those leaving the London region tend to be older, often much older, and are, typically, partnered with children (so there are many of them), and they are also predominantly middle-class homeowners. During a recession, not only are house prices low, but it is also difficult to sell property. As the economy experiences an upturn, confidence in the housing market suddenly increases, houses become easier to sell, and the prices obtained are higher. Those migrating away from the metropolis can now “surf” the high house price wave as it spreads outward to the nearby regions and often, ultimately, to the whole country. They can sell high and buy low, thereby maximizing the value of the assets that they have stored up in their metropolitan properties. The implications are striking: many of those who leave the metropolitan region are responding very intimately to business cycle processes; in contrast, most of those who enter the region are indifferent to these processes.

The second layer in Figure 3 refers to economic processes that take place over a much longer time span, typically over a period long enough for three or four business cycles. These processes are called “economic restructuring” because they imply new ways of organizing what, how, and where goods and services are produced. In this author’s judgment, the best way to envisage them is through the concepts of changing spatial divisions of labor.

In the period 1950–1970 the dominant spatial division of labor was regional sectoral specialization. What is meant by this is, first, that the nation-state largely defined the limits of the economy – most economic transactions were internal to the national space economy; second, that regions specialized in the branches of
production for which their special endowments of physical and social resources (local climates, mineral resources, labor skills, work cultures, etc.) were best suited; third, that the incomes earned from the sale of these goods and services to people living in other regions of the country would be used to purchase the goods and services produced in those other regions; and, finally, that the agency for the exchange of these goods and services would be the market – this implies that the spatial or geographical division of labor would coincide with the social division of labor, that is, the distribution that would result from market exchange. The implications of this for migration were profound. If the specialization on which a region’s economy depended suffered decline, for example, through changing consumer preferences, technological change, or a low income elasticity of demand, then businesses would fail, jobs would disappear, and out-migration would follow. In the 1950s and 1960s most high-income countries experienced rapid rural depopulation. This came about because technological change was revolutionizing food production, leading to much lower labor inputs; food and agricultural products had a low income elasticity of demand (demand increased very little as incomes went up), and rural businesses in all three macrosectors (agriculture, etc.; manufacturing; and services) tended to be small, technically backward, labor-intensive, and uncompetitive. In contrast, large factory industrialization and services-based urbanization were making the cities rich in employment opportunities. The result was a mass movement from the countryside to the city. The classic study of this migration was Guy Pourcher’s (1964) account of the migration of men and women to the Paris region. However, rural depopulation was not the only migration pattern for this period. Old industrial areas built on coal and steel, shipbuilding and textiles, were also losing out economically to the centers of Fordist production. By Fordist production is meant the mass production of standardized goods for mass markets – products such as automobiles, white goods such as cookers and refrigerators, and electrical goods such as radios and televisions – goods that were very much in demand during this period. Fordist forms of production called for very large workforces and very large markets; these were to be found only in the largest and wealthiest of cities – the metropolitan conurbations and national capitals.

Connecting and, for a time, coexisting with regional sectoral specialization, the new spatial division of labor emerged in the 1970s, coinciding with a rapid growth in the size of both private sector firms and of public sector organizations. Instead of a region being defined by the sector of the economy that dominates its labor force, it was defined by the role that its labor force played in the production process – as headquarters staff managing the affairs of the whole company (located in the capital city-region and/or main financial center); as research and development staff bringing new products to market or evaluating new production technologies (located in urbanized countryside close to the metropolis or in a major university city); as skilled workers needed to produce those goods and services calling for particular knowledge or experience (located in provincial cities and in new industrial areas); and as raw labor engaged in routine unskilled production (located in regions with large labor surpluses such as agricultural regions and old industrial areas). The spatial division of labor now coincided with the technical division of labor – that is, a planned separation of tasks designed to maximize the efficiency and/or profit performance of the whole organization. Once again, the migration implications were profound. Since the new spatial division of labor brought the work to the workers (at the same time, incidentally, as
welfare payments helped unemployed people to survive in situ), working-class migration became much less important, and indeed in most high-income countries total interregional migration rates began to fall during this period. But, below the surface, two important new trends were developing. The first was the growth of intraorganizational transfers – as organizations managed their large and widely distributed workforces to achieve efficiency, they deployed their managerial, professional, and technical staffs to their various operations located in different regions of the country. So middle-class mobility was being boosted just at the time when working-class mobility was being suppressed. The occupational class compositions of interregional migration flows were being transformed (Stillwell, Rees and Boden 1992). Second, the labor and other costs incurred by companies in the metropolitan cities were too high to sustain the more routine parts of their production processes. As a result, there was a major decentralization of industrial and routine service sector employment away from the metropolitan city-regions toward small towns and rural areas where labor was cheap and plentiful. With this change in the geography of investment there was a migration “turnaround” – a shift from urbanization toward counterurbanization. This counterurban migration trend of the 1970s and early 1980s was found in all high-income countries (Champion 1989); in some, notably the United States, it was largely interpreted as a response to the changing place preferences of populations (favoring the rural over the urban), but in most cases it was seen as the result of shifts in the political economy of production, often combined with demographic factors (Ishikawa 2001).

From the late 1970s, and building up over the next 20 to 30 years, a further change in the spatial division of labor occurred. This time the driving force was not organizational change affecting firms operating within the national space economy, but the strong development of links external to the national territory – in short, globalization. The effect of this at the level of regions within a country was to produce regional functional disconnection (Dunford and Fielding 1997); this means that regions are not linked to each other through market mechanisms (exchanges of goods and services) or through the planned separation of tasks (as in the new spatial division of labor), but are instead disconnected from each other – each one linked independently through global connections of ownership, management, and the flows of goods, money, information, and people to places outside the national territory, that is, to other parts of the world. The London city-region represents, perhaps, an extreme version of this process; so seemingly cut off from the rest of the United Kingdom has it become as its global city character has displaced its national and regional economic roles that it is sometimes spoken of as if it were a “different country.”

Globalization also weakens labor’s relations with capital, allowing forms of flexible specialization to emerge along with new industrial districts and industrial clusters. As a result of these changes, regional functional disconnection has two major implications for migration. The first is that it further reduces the many economic connections between regions that have previously acted as conduits for migration (e.g., intraorganizational transfers). The second is that, by forging relationships at the global scale, it promotes international migration and thereby helps to bring about the new “superdiversity” of metropolitan cities in high-income countries. The social significance of these international migrations lies in the fact that they add populations at both ends of the social spectrum – at the top in the form of highly paid professionals and managers, and at the bottom in the form of gap-filling immigrants.
who do the jobs that the locals, including interregional migrants, cannot or will not do.

Since about the mid-1990s there has been another change in the political economy of advanced capitalist countries, a change that has again tended to focus net migration gains in the major metropolitan city-regions. This change does not involve the end of globalization but rather its transformation from being primarily about the integration of markets for the production and distribution of material goods to the establishment of global networks for the trading of immaterial goods and services – a shift to what is sometimes referred to as the “knowledge-based economy” or “cognitive capitalism.” This trend toward the financialization and commoditization of key social relationships, and the partial replacement of wealth creation by wealth capture, favors the regions that are the linchpins of the global economy, places where rent-seeking behaviors are most likely to be rewarded – New York, London, and Tokyo, of course, but also globally significant cities like Hong Kong, Singapore, Shanghai, and Mumbai.

The third layer in Figure 3 refers to economic processes that work themselves out extremely slowly, say over a person’s life span or over three or four economic restructurings, and are therefore called “deep structural” processes; this change is so slow that it is almost imperceptible. The underlying macrogeography of wealth and poverty, both within and between countries, is the kind of thing to which this label refers. At the global level, the deep structural processes include those that help to maintain the inequalities between developing and developed countries, and the slow but steady shift of wealth and power since the 1950s toward East Asia. At the regional (i.e., subnational) scale, certain metropolitan city regions seem to dominate the space economy in wealth and power from one generation to the next – Paris does so for France, and Beijing and Shanghai for China. One of the principal reasons for this stability of wealth concentration is interregional migration. These metropolitan city regions attract the brightest and the best from all over the country; they promote them socially at rates that are much higher than in other regions; and they lose them to other regions at later stages in their careers or at, or close to, retirement from the labor market. The metaphor of the escalator has been used to describe this process – hence these regions become “escalator regions” (Fielding 1992).

Conclusion: population mobility and regional development in Japan

It is sensible to draw some of these ideas together through the use of an empirical example, Japan. The focus will be on the net migration gains and losses of the three major metropolitan regions: Tokyo, composed of the four prefectures of the southern Kanto region; Osaka, composed of the four core prefectures of the Kansai region; and Nagoya, composed of the three core prefectures of the Chubu region (see Figure 4).

The first thing to notice is the variability of the net migration curves, and the tendency for the peaks to coincide with periods of business confidence and the troughs with periods of economic difficulty or recession. Second, the early 1970s, that is, before the oil crisis of 1974–1975, marks a turning point in the migration flows to Japanese cities. Before this, all three metropolitan regions were the beneficiaries of net migration from the rest of Japan, and the order was simple – the largest gained the most. The 1950s and 1960s were the peak years for rural depopulation in Japan, and hundreds of thousands of people, mostly young men and women from farming and small business backgrounds, were migrating
each year to the major industrial cities as regional sectoral specialization favored these centers of Fordist production over traditional provincial Japan. After the turning point, however, something very interesting happened – Tokyo took off on its own as a major destination for internal migration. This is largely explained by the transition to the new spatial division of labor; Tokyo is the corporate headquarters location for Japanese domestic and international business. The 1980s was a period when these major corporations were achieving astounding success both in Japan and elsewhere, and Tokyo, not Osaka or Nagoya, was the major beneficiary of that success. Tokyo’s global city success, however, suffered a severe blow during the Japanese financial crisis of the early 1990s; the net migration gain momentarily became a net migration loss as asset values tumbled and business confidence plummeted. Since that time, significant net migration gains have returned and held up until the current crisis which started in 2008.

But Figure 4 also shows us something else – the enduring attractiveness throughout business cycles and restructuring periods, of the Tokyo metropolitan region to internal migrants. Tokyo is the “one-point concentration” of the Japanese space economy. Where would an able, active,
and ambitious young person go? To the center of bureaucratic power and corporate wealth, of luxury consumption, artistic creativity, and celebrity success – they would go to Tokyo. Tokyo is itself a major hub for university education (Tokyo University being the most prestigious in the country), but it is also the destination for many students graduating from other university cities around the country. These other university cities, especially Kyoto, but also major provincial cities such as Osaka, Nagoya, Fukuoka, Sendai, and Hiroshima, act as “siphon cities,” recruiting bright students from nearby regions and from further afield and channeling them to Tokyo after graduation. And, when they get there, they benefit from the massive opportunities for upward social mobility that Tokyo offers relative to other places in Japan (especially, incidentally, for women). In the past, once a person had made it to Tokyo, they stayed there, but recent census results show that Tokyo is now joining many metropolitan cities around the world that lose older migrants to provincial, coastal, and rural areas as they enter retirement age groups.

In all these respects – stepping on, being taken up, and then stepping off – Tokyo displays the characteristics of an “escalator region,” a region that acts as a kind of permanent engine for upward social mobility within contemporary society.

Every country is unique, and Japan is arguably more unusual than most, and yet, using this example, most of the key theoretical relationships between population mobility and regional development outlined earlier in this entry have been exemplified.

SEE ALSO: Cumulative causation, endogenous growth, and regional development; Political economy and regional development; Regional development models; Regional inequalities; Regional unemployment and regional labor markets

References


Ports

Peter V. Hall
Simon Fraser University, Canada

Ports provide safe harbor for ships and boats so that goods and people can be transferred between water- and land-based transportation modes. This functional definition is complete, yet it obscures a fundamental global–local binary that makes ports a central and fascinating object of geographical inquiry. On one hand, ports are spaces situated in global transportation networks that connect widely dispersed points of production and consumption, and hence they are forged by the actions of steamship lines, railroads, cargo owners (shippers), workers, and governments. On the other hand, ports are places with distinctive local and site-specific characteristics, combining particular coastal morphologies and development histories, imbued with political and cultural meaning. Traditional geographical approaches focused on the conception of ports as places, while post-World War II analysts focused on ports as spaces; today, a more balanced appreciation of ports as multiscalar and multidimensional activity systems is emerging.

The simplest ways of classifying ports reference the physical characteristics that influence which vessels may call there: water depth and the width and other dimensions of the approach or shipping channel and berth limit the draft, beam, and length of ships. While natural characteristics of the waterways are important, often these are determined through dredging, the construction of coastal infrastructure, canals, locks, and berths. Hence the body of water on which a port is located provides one axis of classification: sea (or ocean or maritime), river, or lake ports. The ships which may call at these ports, and the cargos they will handle, are constrained in different ways: where seaports may face high and low tides, river ports may face seasonal flooding or drought, while lake ports may be constrained by ice during winter. Ports face other physical constraints – one of the most important is the availability and configuration of land for cargo-handling activity and expansion.

Ports are made up of one or more terminals, the term which describes the distinct combination of facilities for loading ships: wharf, storage yard, cranes, and other equipment. Different cargos require different terminal attributes. Hence categories of cargo also describe cargo-handling statistics and terminal types: container, ro-ro, and break-, liquid-, and dry-bulk. Containers, metal boxes which carry a huge variety of goods from manufactures to refrigerated foods, are lifted on and off ships with specialized cranes. Container terminals are prized because of the value and diversity of cargo moved in containers (see Figure 1). As container ship sizes have increased, container terminals have become larger and more mechanized (and increasingly, robotized) in order to achieve the desired economies of scale. Ro-ro (or roll-on, roll-off) cargos refer to automobiles and other self-propelled vehicles loaded via ramps. Break-bulk cargo is typically undifferentiated cargo loaded on pallets, in crates, or in bags; it is also increasingly common to distinguish neo-bulks, which refers to differentiated cargos such as steel, paper bales, and granite. Liquid-bulks such as oil and liquefied natural gas (LNG), and dry-bulks such as coal and grain are loaded unpackaged. Where
Figure 1  Millions of container units (twenty foot equivalent units, TEUs) per port, 2012. Most of the world’s largest container ports are located in East Asia. Adapted from Containerization International, port authorities and port associations, created by Jean-Paul Rodrigue.
liquid- and dry-bulk cargo may be poured or pumped onto ships, break- and neo-bulk cargo is typically lifted by crane. Most of the world’s largest ports have multiple terminals, and hence are multi-functional. Rotterdam, for example, is known as a container gateway to northern Europe, but it also handles huge volumes of bulk cargos. There are also non-cargo terminals and ports: cruise, ferry, fishing, and recreational.

Ports vary according to the inland and offshore location(s) they serve. Local ports serve the city and regional territories adjacent to the port area, while gateway ports serve more widely dispersed areas that could be continental in scale. The concept of the hinterland is fundamental to this distinction; local ports serve a captive hinterland in which they enjoy an effective monopoly, while gateway ports are likely competing with other ports to serve a contested hinterland. Of course, terminals within ports may also compete with each other to serve local and remote hinterlands; indeed, in a process called terminalization, the operators of port terminals today are asserting a more active role in value chains (see also Global commodity/value chains).

Ports may also be characterized according to their insertion into water-based transportation networks. The concept of the foreland describes this waterside or overseas trading territory of a port (Rodrigue and Notteboom 2010). Hub ports collect and redistribute cargo from multiple smaller feeder or peripheral ports located within the foreland of the dominant port. For example, containers from ports in the Pearl River Delta might be barged to Hong Kong and there loaded onto larger container ships. While transshipment occurs by definition at all hubs, ports such as Singapore which are specialized in this activity are known as transshipment ports. Also known as intermediate hubs, several of the world’s newest container terminals and ports have been built at locations which connect the major oceanic routes: for example, Algeciras and Tangier Med ports are located near the Strait of Gibraltar to connect Atlantic and Mediterranean–Indian ocean traffic. Hub and transshipment ports are most closely associated with containerized cargo.

Containerization is associated with other profound changes in the geography of ports. Traditional port locations were physically determined, sited at convenient river crossings (e.g., Paris), the head of a delta (Cairo) or estuary (Antwerp), on either side of an isthmus (Manila), or in a protected bay (Durban). These sites were important points in the “break of bulk,” hence there was a close historical relationship between ports and human settlements. Not only was lots of physical work and hence employment involved, but also warehousing and financial and legal administration to support trade. Merchants and traders left a distinctive impact on historic port cities in both metropolitan and colonial locations.

Industrialization heralded a change in the scale of shipping that was later greatly amplified in the container era. Fordist industrial ports were associated with processing imported raw materials for nearby assembly lines, and the export of manufactured products. Contemporary versions of such port-centric industrialization inform the widespread creation of export processing and Free trade zone around ports, notwithstanding concerns about whether the economic benefits will accrue to the port region (see Hall et al. 2011).

Changes in the relationship between ports and cities intensified with containerization. Larger ships require deeper water and more space for cargo operations. In his site-centric “anyport” model, Bird argued that ports would migrate downstream toward deeper water and away from the land-constrained port city, while Hayuth argued in his situation-centric “load center” model that a small number of ports would win out in the competition for the larger ships
PORTS (see Olivier and Slack 2006). In large measure, both predictions have been proven correct, but have also been surpassed by subsequent events. In many parts of the developed world ports, especially terminals and industrial sites in core urban areas, were abandoned. Some waterfronts have been redeveloped as sites of consumption, residence, and commerce (see Desfor et al. 2010; see also Urban redevelopment).

Yet the goods kept moving in ever more complex global supply chains. In order to secure the massive investments required to build and operate container terminals, port authorities, the government organizations which manage port lands, as well as terminal operating firms, have become increasingly active and strategic about their investments in both hinterlands and forelands. Notteboom and Rodrigue (2005) have termed this process “port regionalization.” The resulting changes in the relationship between ports and urban economies have been extensively documented (see Wang et al. 2007). In general the literature indicates that the relationship between ports and economic development today is positive at larger (continental, national, and mega-regional) spatial scales, but more uneven and even negative at smaller (neighborhood, port district, and city) scales. There are also differences in the port city economic relationship in different world regions, which relate to differences in colonial history and recent industrialization.

Such economic geographies are, of course, made by actors. Olivier and Slack (2006) draw attention to the increasingly important role played by transnational corporations (TNCs) in shaping patterns of port development: these TNCs include specialized global terminal operating firms, the major container steamship lines and their terminal operating divisions, as well as increasingly influential shippers such as the major North American and European retail chains. The entry of TNCs has been enabled by changes in the governance of ports, including privatization of some terminals, the corporatization of port authorities in formerly publicly owned and operated ports, and the financialization of the port industry.

All of these processes open up the possibility for new conversations between port and other human geographers around economic and urban development, institutions, and even culture and the health impacts of port activity. Terminalization, for instance, which involves the simultaneous fragmentation of ports as coherent places but also the opening up of new global and local connections to hinterlands and forelands, shares much with other Glocalization processes. Port geographers are also engaging more with physical geographers with ongoing attention to marine environments as well as to climate change and sea level rise and options for reducing transportation-based greenhouse gas emissions.

SEE ALSO: Airports; Coastal zones; Containerization; Corporations and global trade; Infrastructure; Logistics; Maritime transport; Rail transport: freight; Road transport; Transport geography; Transport technology

References


Portugal: Associação Portuguesa de Geógrafos (APG) (Portuguese Geographical Association)

Founded: 1988
Location of headquarters: Lisbon
Website: www.apgeo.pt
Membership: 800 (as of 2013)
President: Rui Pedro Julião
Contact: rpj@fcsh.unl.pt

Description and purpose

Associação Portuguesa de Geógrafos (APG) is the professional association for geographers in Portugal and it has just celebrated its 25th anniversary. The APG’s mission is to promote geography, supporting its professional development, and fostering the recognition of the geographers’ role in society.

Journals or major publication series


Current activities or projects

The APG organizes a congress every two years (odd years) where the broad field of geography and its major academic, research, and professional achievements are presented and discussed. This conference alternates every year with the Iberian conference (even years) as there is a strong liaison between the two countries. The APG also publishes Inforgeo, a magazine that is an open platform for paper publication.

The APG sends out electronic newsletter three times a week to all members and registered users with relevant news and the agenda for the upcoming period.

Brief history

APG was founded in 1988 by a group of academic and professional geographers and currently has about 800 members, most of them active as public servants at different levels of the public administration (municipalities, regional, and central administration) and working in a wide range of subjects.

The history of Portuguese geography is not confined to APG activities, but, undoubtedly, APG is now an important part of recent developments.

In 2012 the APG received a Public Interest Organization award from the Portuguese government.

Submitted by Rui Pedro Julião
Portugal: Sociedade de Geografia de Lisboa (SGL) (Geographical Society of Lisbon)

Founded: 1875
Location of headquarters: Lisbon
Website: www.socgeografialisboa.pt
Membership: 1280 (as of December 31, 2013)
President: Luís Aires-Barros
Contact: geral@socgeografialisboa.pt

Description and purpose

The Sociedade de Geografia de Lisboa (SGL) was created in 1875. Its origins and objectives were connected to the Portuguese and European contexts of that time and with problems in relation to Portugal’s position regarding its overseas colonies, especially in Africa. Currently its fundamental concerns are the maintenance and reinforcement of scientific and cultural connections with the countries in which Portuguese is the official language through conferences, seminars, books, and the systematic publication of the *Boletim* (Bulletin).

The SGL library is recognized nationally and internationally as a mandatory resource for all investigators dedicated not only to the study of the Portuguese expansion and history of discoveries but also to geography, history, and ethnography of the Portuguese speaking countries. It is used annually by students and researchers from all over the world.

The SGL has a museum rich in artifacts coming from Africa, Asia, and Australia.

Journals or major publication series

*Boletim da Sociedade de Geografia de Lisboa*
*Memórias da Sociedade de Geografia de Lisboa*

Current activities or projects

General cultural activities covering a great variety of subjects are carried out by about thirteen general commissions (such as the Africa Commission, Asian Commission, European Commission, Commission for International Relations) and by specific sections (such as Section of History, Section of Anthropology, Section of Ethnography, Section of Mathematical Geography and Cartography, Section of Ocean Geographic, Section of Literature and Fine Arts, Section for the Patrimony Studies, etc.). Many of the presentations made at the public sessions are organized by those commissions and are published in the *Boletim da Sociedade de Geografia de Lisboa*. Some workshops or colloquia publish their papers in special issues of *Memórias da Sociedade de Geografia de Lisboa*. Some special themes regarding our rich collections (at the library and/or museum) are developed as research projects supported by the National Foundation for Science (Ministry of Science) and normally involve university researchers.

Brief history

After the Berlin Conference (1840) organized by Bismarck, a rush of geographic exploration took place in Africa (the so called *Terra Incognita*), crossing this continent from the Atlantic Ocean to the Indian Ocean (e.g. Livingston, Stanley,
PORTUGAL

Cameron, Brazza). These geographical voyages were promoted by some recently founded geographical societies (such as those of Paris, London, and Berlin). It was thus considered mandatory for Portugal to have representation in that movement, leading to the birth of SGL in 1875 proposed by people interested in overseas matters. It received official support from the outset. Since then, the SGL has been sustained as has no other private institution specially connected to the national cultural life. It is alive and dynamic, in some cases taking initiative, in others coordinating activities with other national institutions.

A large number of congresses and colloquia, some promoted by SGL, same in cooperation with other institutions, have been held at SGL premises. It has created an ethnographic and historic museum in its headquarters. This museum has received not only documents and remainders of documents by Portuguese navigators from the fifteenth century in the African coast, but also donations to form an outstanding collection of artifacts from Africa, Asia, and Australia. No less important is SGL’s cartographic archive which offers the visitor a rich of variety of maps, atlas, plants, both Portuguese and foreign, and from all times in history.

Submitted by Luís Aires-Barros
The concept of positionality was introduced into human geography by way of feminist and poststructural interpretations in the early 1990s. Positionality highlights how people, including researchers, come to know and interpret the world from different social locations; positionality shapes research, and may inhibit or enable certain research insights. A person’s embodied subjectivity mediates the research process, including the relations with the people, places, and materials of the research project. Initially conceived in reference to face-to-face fieldwork, it was soon apparent that positionality had purchase in other sorts of research methods. As a concept, positionality raises important questions about the politics and ethics of research, and since the mid-1990s it has become highly influential in the design, conduct, and representation of research across human geography.

The roots of interrogating the impact of positionality on the research process date back to broader feminist challenges to the supposedly neutral, impartial claims and theories in the social sciences and the humanities. Also significant are contributions of feminist scholars of science studies who illuminated the ways in which scientific thought and practices are gendered. In brief, the critique is that seeking universal truths and all-encompassing knowledge (whether in the humanities or the social or natural sciences) constitutes what Donna Haraway (1991, 189) famously describes as an all-seeing “god trick … seeing everything from nowhere.” In short, all research is a view from somewhere and research is never value-free (even “hard science” research). Intellectual activities are shaped by the historical circumstances, normative expectations, and cultural beliefs of the time and place in which they are created. Knowledge is partial, situated, and power-laden.

The implication of situated knowledge for positionality is that both researchers and participants (and everyone else) are differently situated by their cultural, intellectual, and spatial locations, as well as their intellectual history and lived experience, all of which mediate our understandings of the world and the knowledge we produce. Thus research is created in a world already steeped in meanings and interpreted by people, including the researcher, who live their lives in it. When conducting research, our values and location in webs of power frame what we take to be “facts,” how we develop a particular research strategy, and what research questions we ask and what we “see” when exploring those questions. Researchers are a visible and integral part of the research encounter, and, as Liz Stanley and Sue Wise (1993, 157) put it, “researchers remain human beings complete with all the usual assembly of feelings, failings, and moods.”

As researchers are an embodied and integral part of the research process (rather than external, detached observers), positionality has been further extended to include considering others’ reactions to the researcher. Both our embodied presence as researchers and the participants’ response to them mediate the information collected in the research encounter. Moreover, subjectivities are contextual and shifting: different personas may be performed in particular
Positionality

circumstances and locations. In addition, positionality is a relational concept: the researched
are not passive; they are knowledgeable agents and, as such, the research encounter is structured
by both the researcher and the researched, both of whom construct “the data.”

Like most useful and robust concepts, positionality has been refined and adapted over
time. Two sets of theoretical, conceptual, and empirical interventions have been especially
significant: intersectionality and embodiment. Positionality became popular in the context of
the rise of identity politics and the early 1990s poststructural crisis of representation. Part of
that was also a challenge to white, Western feminists’ epistemic reliance on the universality
of a monolithic category of women as oppressed. Since then there has been growing literature
on the discourses, practices, and processes of systems of social difference and inequality (such
as gender, class, “race,” ethnicity, (dis)ability, and other social identities). Intersectionality involves
considerations of how systems of difference intersect differently in different places, times,
and situations. Positionality has and continues to be enriched by researchers paying attention
to how they are positioned (by the self, by others, by particular discourses) in relation to
multiple relational and intersectional processes of difference. Critical awareness involves addressing
how researchers are differently positioned in hierarchies of power and privilege from each
other and from those they engage in their research.

These more thorny issues have arisen particularly in the context of the positionality of those
wishing to address multiple and cross-cutting positions of privilege and oppression, and the
politics of white Westerners researching people from/in the Global South. Rather than avoid
such research encounters, Richa Nagar and Amanda Lock Swarr (2010) suggest building
cross-cultural collaborations and critically aware transnational praxis based on “alliances that are
created and sustained through deeply dialogic and critically self-reflexive processes of knowl-
edge production and dissemination” (2010, 3). One such example is Christine Schenk’s (2013)
discussion of her experience of doing fieldwork in the “antagonist context” of a non-Muslim
researcher in the province of Aceh in Indonesia. She argues for a psychoanalytically flavored
relational positionality that embraces a division of power through “cross-cultural collabora-
tion and a gender balanced team composition of researchers (that could) balance differences
arising from different backgrounds” (2013, 352).

Understandings of positionality are also being nuanced via provocative discussions of embodied
methodologies and fieldwork. The body was evident, but not marked as such in early discus-
sions of the feminist politics of both research and positionality. For example, in their influ-
cential mediation on the research process Liz Stanley and Sue Wise (1993, 157) pointed out
that our “consciousness is always the medium through which the research occurs; there is no
method or technique of doing research other than through the medium of the researcher.”
Since then theorizing of bodies, corporeality, and embodiment has propelled discussions of not
only the discursive constructions of bodies, but also the fleshy materiality of actual bodies. For
example, Longhurst, Ho, and Johnston (2008) reflect on their research on the role of making,
eating, and sharing food in the construction of home for migrants in Hamilton, Aotearoa/New
Zealand; and make the case for “our own bodies and our participants’ bodies as ‘instruments of
research’” (2008, 214).

It is now broadly accepted and indeed expected that human geographers will acknowledge their
positionality when constructing, analyzing, and writing up their research. It has become
commonplace for researchers to include a statement of self-representation explaining that they are writing as someone positioned within a particular set of subject positions. This is, of course, an important corrective to the trope of the omniscient, unmarked researcher. But a serious engagement with positionality requires more than this. Positionality is about being embodied and embedded in the shifting processes and practices of identity formation, affect, and power. Remaining at the category identification version of positionality may devolve into self-stereotyping by fixing the researcher’s subjectivity into generic categories, and further runs the risk of solipsistically placing the researcher at the center of the analysis (Robertson 2002; Kobayashi 2009). Positionality’s critical edge is that it is both a reflective and a political resource. The latter demands an accountable, politically aware progressive use of research to capture people’s everyday lives in all their shifting, messy complexities.

SEE ALSO: Ethics in geography fieldwork; Feminist geography; Feminist methodologies; Fieldwork in human geography; Reflexivity

References


Further reading

Postcolonial geographies

James D. Sidaway  
National University of Singapore

The department of geography at the University of British Columbia (UBC) in Vancouver has long been a key site for the discipline, with a community of scholars and generations of graduate students who have made significant inputs to its evolution over the last few decades. Directly opposite the four-story building that houses the department is another lower, longer building, UBC’s First Nations Longhouse (Figure 1). A relatively recent addition to the campus, the longhouse cites a presence on the land that long predates that of the modern university and its departments. Yet that history has only become visible to most visitors in recent years, through developments like the Long House and First Nations museum on campus and a series of art projects across the campus that signify indigenous presence (Figure 2). For decades, however, those histories were largely erased and overwritten as the structures that house UBC and their systems of knowledge and naming were established.

Today Vancouver sits at the intersections of imperial projects. Foremost for Canada and British Columbia has been the British empire, whose linguistic, legal, cultural, and political legacies deeply shape the landscape, nation, and city; however, that Canada accommodates two European settler projects (anglophone and francophone) is frequently visible, since Canada’s constitution requires that both European languages appear on official documents and signage. Moreover Canada is part of the military alliance structure in which its neighbor to the immediate south is the leading power (and, according to many observers, now acts as an empire). But Vancouver also faces west, across the Pacific where British and American power have clashed and melded with other empires in the twentieth century (chiefly that of Japan, but also those of France, Britain, Germany, the Netherlands, Portugal, and, in the century before, Spain) and where American power now faces the successor state to the Qing empire: the People’s Republic of China. China itself is both a successor state to a continental empire that became an object of foreign imperial colonization (chiefly by Britain, Germany, Portugal, and Japan). Elsewhere, Canada played a supporting role in the American-led invasions and occupations of Afghanistan and Iraq, whose course (together with Israeli policies in the West Bank territories it occupied in the 1967 war) led a UBC-based geographer to designate the early twenty-first century as “the colonial present” (Gregory 2004). For Gregory:

The colonial present is not produced through geopolitics and geoconomics alone, through foreign and economic policy set in motion by presidents, prime ministers and chief executives, the state, the military apparatus and transnational corporations. It is also set in motion through mundane cultural forms and cultural practices that mark other people as irredeemably “Other” and that license the unleashing of exemplary violence against them. This does not exempt the actions of presidents, prime minister, and chief executives from scrutiny (and, I hope, censure); but these imaginative geographies lodge many more of us in the same architectures of enmity. It is important not to allow the spectacular violence of September 11, or the wars in Afghanistan, Palestine, and Iraq, to blind
us to the banality of the colonial present and to our complicity in its horrors. (2004, 16)

More specifically however, geographical knowledge and geography as a discipline have long been deeply implicated in imperial projects. Mapping for nation and empire, imperial education, and the codification of strategic geographical knowledge offered much of the rationale for the nineteenth- and early twentieth-century establishment of geography as a formal discipline in many European states (Clayton 2011). Arguably, therefore, much geography was an expression of and was shaped by wider colonial discourses, frequently caught up with formations of “race” and the operation of racism. Awareness and critical scrutiny of these colonial discourses have developed since the 1970s in the wake of and accompanying formal decolonization in much of the Caribbean, Africa, and Asia and ongoing struggles for national liberation.

Orientalism and representation

The intellectual project of postcolonial geography was bolstered by the publication of Orientalism by the Palestinian-American literary scholar Edward Said (1978). For Said:

Orientalism is a style of thought based upon an ontological and epistemological distinction made between “the Orient” and (most of the time) “the Occident.” Thus a very large mass of writers, among whom are poets, novelists, philosophers, political theorists, economists, and imperial administrators, have accepted the basic distinction between East and West as the starting point for elaborate theories, epics, novels, social descriptions, and political accounts concerning the Orient, its people, customs, “mind,” destiny, and so on. (Said 1978, 2–3)

Not only that, for, according to Said, geography – though in a broader sense than the discipline alone, and therefore in the form of wider geographical imaginations—is at the heart of Orientalism and hence imperialism. So the “Orient” is not simply objectively there; it is represented as a historical and geographical object in Western discourses. In turn, those discourses are projected onto the Orient which is then cast and remade in an imperial light. The West is thereby defined vis-à-vis the Orient and vice versa. This process requires one side of the world (the West) to be represented as the source of civilization,
progress, and order, while the other side (the Eastern, non-Western, Oriental world) is represented as lacking these features and hence in need of Western stimulus to progress. For Said,

much as the West itself, the Orient is an idea that has a history and a tradition of thought, imagery, and vocabulary that have given it reality and presence in and for the West. The two geographical entities thus support and to an extent reflect each other. (Said 1978, 3)

While he focused in particular on European (and later American) representations of the Middle East, Said’s account of the power of imperial era definitions has been applied to other parts of the West/non-West binary, such as representations of East Asia, Africa, and Oceania and of labels such as the Balkans and Eastern Europe and “marginal,” “less developed,” and “backward” areas within continents and countries, such as the US South. In turn, the ways that such representations are internalized in the Orient, as well as the possibility of counternarratives and resistances, have meant that the critical study of Orientalism has thrived. This includes the study of a so-called “Orientalism in reverse” or “Occidentalism,” whereby the West is represented by non-Western thinkers as having characteristics (such as lack of religiosity) that invert the prior Orientalist notions about which side of the (imagined and preformed) East/West divide is superior.

Yet, despite this wealth of work, Orientalism was relatively slow to be cited by geographers and it was not until the early 1990s that it became a frequent reference point. It was first encountered through radical geography’s interrogation of the discipline’s intimate relationship to (especially European) imperialism. The history of academic geography, especially in the nineteenth and early twentieth centuries was thereby repositioned as part of a wider Orientalist narrative, complicit with racism and expressed through the discipline’s encounter with environmental determinism, as well as a corpus of colonial and allied geographies of the non-West. By the 1980s, radical geography sought to distance itself from these legacies and they were critically reassessed in subdisciplines such as historical geography (with its sensitivity to the representations and focus on empires), and development geography (in which the core object of study is the non-West). Subsequently, the advent of postmodernism and poststructuralism in human geography foregrounded the politics of representation in ways that Said and other literary theorists had forged. Postcolonial influences thereby became established within critical geography. While this happened, there was also what seemed to be a revival in Western imperial interventions in the form of the American-led wars in Yugoslavia, Iraq, and elsewhere, which led Derek Gregory to denote a colonial present and many others to reconsider the meaning and endurance of empire.

**Postcolonial geographies**

Today such debates continue against the backdrop of an array of theoretically sophisticated postcolonial readings of the spaces of past empires, including critical works on the intersections of sexuality, race, space, and empire (Phillips 2006; Legg 2014). In the meantime, moreover, literary studies have registered a wide broadening of the canon of texts worthy of study, to include postcolonial literature from former colonies or where colonial themes or power is foregrounded. More widely, literary studies were enlivened by the development of postcolonial approaches, inspired by Said and allied work that situated the canon vis-à-vis imperial cultures. These developing postcolonial approaches increasingly also impacted on academic geography through
the 1990s. Reflecting the richness and density of some of the key literary postcolonial theorists who emerged after Said (Homi K. Bhabha and Gayatri Chakrabarty Spivak in particular) some of this geographical work (for a recent example, see Jackson 2014) has been quite densely theoretical.

The publication of an edited collection of *Postcolonial Geographies* (Blunt and McEwan 2002) and a textbook on *Geographies of Postcolonialism* (Sharp 2009) signified the consolidation of postcolonial approaches in human geography. A few years earlier, Daniel Clayton (2001, 749–750) had suggested a series of departures for “postcolonial-geographical lines of enquiry that (variously and among other things)” seek to document and theorize multiple, contradictory, and fragmented identity spaces. He called for grounded work with postcolonial theory that was attuned to the discordant postcolonial politics of the places being studied. In turn, Clayton argued that all critiques of colonialism must recognize themselves as “situated knowledges,” bearing limits derived from their means and sites of study. For Clayton, colonialism needed to be examined at a range of scales, power intersections, and vantage points. He argued for serious study of the past and judicious use of archives, but that these needed also to transform understandings of the present and should have “an anti-colonial effect.” Clayton was wary, however, of a straightforward “postcolonial template for geography” (2001, 750), stressing that “the rudiments of a postcolonial-geographical way of thinking” required much further work, elaboration, and debate.

Analytically it can be useful to distinguish between postcolonialism as a condition, as in a postcolonial state or society that has been through a process of decolonization, and postcolonialism as an orientation, theory, or vantage point, although these postcolonialisms are frequently intertwined (Sidaway 2000). Moreover some geographers have criticized what they discern as a tendency for the discipline to focus on colonial discourse analysis of past colonial geographies at the expense of a focus on contemporary expressions and forms of colonial power. The historiography of the discipline has been greatly enriched by the impact of postcolonial approaches. Yet, according to Gilmartin and Berg (2007, 122), that focus has been “on faraway pasts and geographically distant spaces as the spatio-temporal containers of ‘postcolonial geographies’” comes at a cost:

As a consequence, British postcolonial geography is dominated by critiques of the discursive construction of historical colonialism, rather than focusing on the aftermath or continuation of colonialism. This effective hijacking of the postcolonial within geography means that it loses theoretical force, becomes one of a long list of “posts” that often alienate geographers who might like to, or already do, engage with the challenges posed by postcolonialism to the ways in which we construct knowledge. (Gilmartin and Berg 2007, 123)

While the social geography of migration and diaspora frequently registers postcolonial perspectives, more widely (reflecting postcolonial theory’s formation in literature and allied humanities), social, cultural, historical, and urban geography have registered the impacts of postcolonial perspectives to a greater degree than, for example, political or economic geography. The partial exception has been development geography, which as noted earlier, was an early domain for postcolonial perspectives. But the wider agenda for and possibility for other postcolonial geographies has been asserted by Pollard *et al*. (2009, 139), who argue

that postcolonialising economic geography is not confined to an engagement with the more obviously postcolonial worlds of diasporic communities and the global South. Nor are we suggesting
that it is only those who work in area studies who can practise a postcolonial economic geography. Far from it. The postcolonial economic geography we advocate is not a geography of “the South,” but an economic geography more conscious of its own perspectives and more open to embracing different perspectives through which to view economic practices. It is as important to turn these perspectives on the North to disruptive effect as it is to break the silences from the margins.

Future directions

Several decades on from postcolonialism’s arrival in human geography, what prospects loom now for further postcolonial geographies? Although the issue has become widely debated and contested, it is still the case that the parameters of disciplinary progress are largely defined from Britain and North America, whose relative hegemony (along with that of the English language) in setting agendas endures. That a discipline that was so implicated in colonial projects and therefore caught up with racism remains comparatively white in terms of faculty members at its core British and North American nodes has been flagged by critics. In respect of the opening themes of this entry, however, challenging questions about “indigenous geographies” have been raised by writers such as Panelli (2008, 802) who also draw attention to the relative lack of indigenous faculty in Aotearoa/New Zealand but also stress the ways that indigenous geographies “speak to many core interests in social geography: place, social power relations, and identity politics to name a few.” Subsequently Coombes, Johnson, and Howitt (2014) set out the ethical, epistemological, methodological, and political challenges arising from sustained engagement with and from indigenous geographies. They conclude that

The challenge is contextual, engaged and performative – it requires not just methods of research, but approaches to being-in-the-world. As feminist methodologies shifted the discipline’s view of its purpose a generation ago, debates about approaches, collaborations and ethics in Indigenous geographies unsettle our journey as researchers. They reset the compass by which we guide our endeavours and judge our achievements. (Coombes, Johnson, and Howitt 2014, 851)

In related terms, “decolonial” manifestos that have been elaborated by Latin American writers operate between recognition of the continued power of Eurocentrism and a refusal of its logics. Kiran Asher (2013) describes the premises of this literature, arguing that engaging further with them would be especially productive for the field of development geography. This begins with further study of how colonial conquest and racialized colonial practices constituted the modern world-system and how Western thought claims to be “universal and total,” placing “alternatives” as local and/or marginal. For Asher (2013, 832–833), “such decolonial alternatives or ‘non-Eurocentric’ forms of knowing and being in the world can emerge from the different wisdom and experiences of those who have been on the borders of colonial modernity,” and “Latin America and the past and present experiences of Latin Americans are a key, though not the only, loci of enunciation for decolonial thinking.” And, in a recent article, Jazeel (2014, 101) advocates “a postcolonial geographical research imagination to more effectively, and far less imperially, engage radical alterity.” His aim is to consider the analytical utility of “subalternity,” a category and style of analysis that developed first in critical historical writing on/from India. Indian subaltern studies sought to look beyond the power structures and discourses established there through empire and their categories of analysis that were arguably reproduced by
the successor states to the Raj. Subaltern has become a frequently used and contested term in postcolonial geography and wider postcolonial historical literary, cultural, and historical studies. In the same special issue of the *Singapore Journal of Tropical Geography* that contains Jazeel’s paper and another cited here, Sidaway, Woon, and Jacobs (2014) have suggested (rather like Asher, but without direct reference to her article) “five pathways for a postcolonial geography.”

The first of these (which draws on writing by postcolonial literary theorist Gayatri Chakravorty Spivak) they termed “narrating the planetary,” arguing that planetary thinking brought forth different conceptions (of human relations to nature, for example) to the more frequently used term “global,” which risked seeing the world as a grid, defined by universal systems of commerce and connection that relate to colonial visions and tend to overwrite diversity and difference. The second pathway required greater recognition of a diversity of postcolonial writings in languages other than English (they cited recent Portuguese work, for example) and in fields – such as postcolonial theology – that have hitherto not been widely drawn on by geographers but that are full of insights about categorizations that may inform postcolonial geography. A more central place for indigenous perspectives in postcolonial geography was their third pathway, one already being trodden by others but that required further thoughtful engagement with categories, movements, and epistemologies associated with indigenous peoples. Their fourth path stressed a broadening of one of the enduring strands of postcolonial geography, the critical study of “seeing like an empire,” requiring geographers to examine the presence and reproduction of imperial ways of thought and action. But they argued that this must not be confined to the West but should also scrutinize the legacy of empire among other powers, such as in China’s narratives about its place in the world-system. China’s empire was just one (though arguably the most prominent) among other non-Western empires, such as the Mughals, Japanese, Ottomans, and Safavids, suggesting that the scope for more diverse understanding of what falls under the remit of imperial discourses is in order, along with deeper historical geographies that study how empires, subalterns, and spaces intersected. Their final pathway is related. It stresses the complexity and importance of translation, signaling how the circulation and translation of geographical terms and categorizations have been mediated through colonialism, and may have their origins beyond the West, as well as arguing for a more nuanced recognition of how putative universals, like capital and the urban, internalize and disguise difference and need to be approached through diverse vernaculars and a historical-geographical sensitivity to their origins and uneven circulation.

As this entry has sketched, postcolonial perspectives have become part of a wider critical geography since the 1990s. This has followed an interrogation of geography’s past, specifically its relationship to racism and imperialism. It has also accompanied a developing debate about the geography of geography and theory, what places and cases set the parameters of theoretical debate in the discipline and how these need to be broadened beyond the metropolitan centers. Moreover, since neither racism nor imperialism is dead, the potential for postcolonial geography to elicit shifts in the geographical curriculum and to question who writes and delivers it remains. The decolonization of the discipline is incomplete. And the capacity for human geographers to contribute much more to wider developments in postcolonial theory remains. This contribution might include more careful tracing of the contemporary and historical geographies of postcolonial theory itself and its
relationship to allied fields, notably poststructuralism (Ahluwalia 2010), as well as connections with and contemporary reconfigurations of “nature” (Chakrabarty 2014), sovereignty, and security. There is much work to be done.

SEE ALSO: Antiracist geography; Critical geography; Development; Poststructuralism/poststructural geographies; Race and racism; Radical geography; Whiteness

References

Postconflict geographies

Jon D. Unruh
McGill University, Canada

The powerful spatial aspects of armed conflict and its repercussions create the particular features that constitute postconflict geography. Wars are almost always about some form of spatial contest and control. While the focus of armed encounter on military objectives does influence the postconflict period, it is the profound disruption of economic, livelihood, ethnic, and identity-based attachments to, and dependencies on, particular landscapes that constitute the greater domain of postconflict geographies. The sociospatial repercussions of violence, dislocation, lack of access to and destruction of lands and properties, together with the breakdown of formal and customary institutions that manage land, property, and territory, have a much more pervasive and lasting impact on postconflict geography than the aggregate repercussions of battlefield encounter (Holbrooke 1998). Chronologically, these repercussions begin with the control and negotiation over the spatial spoils of war (landscape-based resources and locations that are or will become valuable), and continue with large-scale population return to areas of origin from which dislocation occurred, and the subsequent focus of war-affected populations on the pursuit of postwar livelihoods, including interaction with areas containing landmines.

How wars end can influence certain aspects of postconflict geography, but generally wars come to a close through either a peace accord or victory. As a civil war moves toward a ceasefire when negotiations for a peace accord are to begin, there is often increased battlefield activity as the different sides attempt to gain control over as much land, or over valuable areas, as possible before the ceasefire comes into effect, so as to have an advantageous position during negotiations or for future profit or administrative or population control. At the onset of a ceasefire, combatants then maintain their positions in the field and hence their control over specific areas, resources, and people. The result is often a spatial patchwork of control, along with a parallel patchwork of alliance and enmity with local populations. This then has a significant influence on postconflict geography, as positions and areas of control, alliance, and enmity harden, are negotiated away, or become contested again, sometimes even with a return to armed conflict. Control of specific areas, resources, and local populations can often be treated as spoils of war to be used in postconflict patronage politics. This control also influences the return of dislocatees, and can become volatile when the pre-war occupants of such areas return and attempt to reclaim lands. With the signing of a peace accord, international peacekeeping forces usually provide security, keep belligerent parties separated, and operate disarmament, demobilization, and reintegration programs, but can have little influence over the boundaries of areas of control by the opposing sides (Stedman, Rothchild, and Cousens 2002).

A war that ends with one side prevailing over another can take a postwar geography in a couple of possible directions. There can be a takeover of territory, lands, and land-based resources that were previously occupied by a population allied with the losing side, which usually then results in their resettlement elsewhere. This resettlement...
often involves impoverished livelihoods and may be where a subsequent insurgent movement emerges. A less common direction involves spatially based attempts at winning hearts and minds with the intention of reducing the prospects of future instability. Such measures can include the return of lands, assistance with reintegration, and support for livelihoods. What distinguishes a postconflict geography resulting from victory from a peace accord is that, with the former, there is no patchwork of control of territory, land, resources, and people, because all areas come under the control of the winning side. In addition, there is usually no allowance given to the losing side to form a political party as there often is where a peace accord has been reached. And there is much less of an international community presence because there is less desire by the victor to engage in disarmament, demobilization, and reintegation (Holbrooke 1998).

The forced dislocation of populations from pre-war livelihoods is one of the primary features of what becomes a postconflict landscape. Both refugees and internally dislocated persons (IDPs) often number in the hundreds of thousands and occasionally in the millions. The spatial repercussions involving the landscapes from which they depart, in which they arrive and stay for the duration of the war, and to which they then return are profound and long-lasting. Such forced dislocation, whether as a result of populations fleeing the effects of war or as part of purposeful military-on-civilian objectives of ethnic cleansing, secondary occupation, expropriation, emptying productive lands of their agricultural occupants so that opposing forces cannot resupply themselves, control over high-value resources, or the use of landmines, is a pervasive and acute aspect of today’s conflicts, their conclusion, and recovery (Fay and James 2009; Leckie 2009). International community participation in the return of IDPs and refugees to areas of origin after a war is a high-priority, costly, time-consuming, and complicated endeavor that puts great pressure on governments, donors, nongovernmental organizations (NGOs), and local communities (Leckie 2009).

An ability to return to predislocation livelihoods in a home area will depend on the length of the war, the degree of intactness of the return community, relationships between those who left and those who stayed, the scale of physical changes at the field, village, and landscape scales, and the degree to which individual and community changes during dislocation are still compatible with previous livelihood patterns. Also relevant to going back are the presence and activities of other actors, including squatters and commercial interests, who may seek access to lands thought to have been abandoned during the war (Holbrooke 1998).

Returning to areas of origin or elsewhere subsequent to the end of a war does not mean that populations will quickly re-engage in livelihoods that existed prior to dislocation. The productive landscape assets that would allow for resettlement in both rural and urban settings will have been destroyed, neglected, degraded, or taken over by others. The result for returnees is often the pursuit of a postconflict-specific set of livelihood activities that interact with the landscape in very short-term approaches to fulfilling the immediate needs of food, water, shelter, and security. While such approaches are usually not sustainable, even in the near term, they can nonetheless provide for survival. This form of landscape interaction includes the rapid overextraction and depletion of certain spatially based resources, such as overhunting, overfishing, freshwater resource depletion, destructive farming (i.e., no fallow period, quick farming with inappropriate crops, farming on unsuitable terrain), and the consumption or sale of reproductive livestock, along with the overextraction and sale
of timber and other plants, alluvial minerals, and wildlife parts. The pursuit of such postconflict livelihoods can be facilitated by incursions into national parks, commercial farms, and peri-urban areas. At the same time, however, a postconflict livelihood strategy also includes an attempt to reacquire or rebuild the landscape assets that allow for the (re)construction of a more durable livelihood strategy, such as (re)acquiring farmlands, grazing lands, farming implements, livestock, and so on. Essentially, then, postconflict livelihood strategies involve an attempt at the reacquisition of productive assets, together with the pursuit of a constantly changing menu of extractive, low-investment, opportunistic, short-term, and often dead end (resource-exhausted) activities which, taken as a broad set of pursuits, significantly alter both the human and physical geography in postconflict scenarios. Such strategies then influence the landscape constraints and opportunities of recovery and development. This point is significant because, subsequent to a war, the lack of food, jobs, markets, and infrastructure result in a much greater proportion of a national population living very close to the land in extractive and disrupted agricultural formats (Peluso and Watts 2001).

A further factor influencing postconflict geographies is the presence of landmines. The strategies for laying land mines highlights the highly spatial nature of violent conflict, particularly civil conflicts in developing countries. The legacy of landmines and other explosive remnants of war (ERWs), used offensively in area denial, social disruption, ethnic cleansing, and contests for high-value land resources such as diamonds, or defensively as a deterrent to enemy incursion and for the protection of specific locations, exerts considerable spatial influence on the recovering of livelihoods. The enormity of the ERW problem in war-affected countries is a primary, enduring obstacle to recovery and development. Apart from the highly visible impacts on mortality and morbidity, the presence of ERW alienates a population from important agricultural, forest, water, and pastoral resources. They also delay the return of displaced populations, thwart the reconstruction of infrastructure, and disrupt important fundamentals of peace building such as land and property restitution, investment, and personal security. The presence of ERWs not only leaves a profound social, economic, and physical footprint on specific areas, but has significant repercussions on larger adjacent areas where affected populations must reside in order to pursue farming and pastoralist livelihoods, in many cases coming into conflict with the original occupants of such areas (Unruh and Corriveau-Bourque 2011).

**SEE ALSO:** Migration: forced; Military geography; Peacekeeping; Refugees; War

**References**


Further reading


Postdevelopment

David Simon
Royal Holloway, University of London, UK

The shortcomings and failed promises of conventional or mainstream development have long spawned lively theoretical debates, attempted reforms, and alternative visions. Beyond the confines of particular groups or institutions, consensus has never existed in respect of development theory, policy, or practice and probably never will. Critical engagement should be regarded as healthy and debates have differed considerably in terms of the extent of problems identified and changes proposed by the protagonists. Postdevelopment is emblematic of approaches that reject not just the outcomes but also the fundamental assumptions and approaches of conventional development.

In terms of intellectual history, the watershed moment was the so-called impasse in development first identified in the mid-1980s and which has been characterized as a profound crisis of existence for the discipline. Its significance lay in the coincidence of several major trends or processes, including the widespread disillusionment at the failed promises or deleterious effects of developmentalism and many development projects, increasing recognition of the environmental costs of conventional development, and the need for greater sustainability. The impact of the debt crisis and reversal of major previous gains by the enforced curtailing of development expenditure, and the profound challenge to modernist ideologies and discourses (which encompassed the major contending development theories and approaches) posed by postmodernism were also important.

After some years of foment and ferment, a series of poststructural approaches emerged that reject the monolithic, universalizing, and totalizing approach of modernist discourses like modernization theory, dependency, and neoclassical and structural Marxism, all of which contended to be “the” single truth. Underpinning poststructuralism, by contrast, is the embracing of difference, diversity, polyvocality, hybridity, and some degree of cultural relativism. The best-known such approaches include postmodernism, postcolonialism, posttraditionalism, and anti- and postdevelopment. Conventional modernist approaches privilege Western ideas, norms, values, and prescriptions as embodied in the bulk of post–World War II development theory, policy, professional expertise and the interventions based on them. In contrast, poststructural perspectives reject such singularity and acknowledge multiple truths and knowledges, including indigenous traditions and systems, all of which have potential value. Accordingly, processes and projects need to be formulated inclusively, drawing appropriately from each according to the specific context, and perhaps with hybridized outcomes.

Understanding anti- and postdevelopment

Since postdevelopment is itself something of an umbrella for diverse interpretations, it is helpful to distinguish between those that constitute essentially rejectionist critiques of conventional development(alism), often termed antidevelopment(alism), and those seeking to envision or formulate
alternative approaches. Probably the most widely cited “foundational” antidevelopment texts are *The Development Dictionary* (Sachs 1992), especially Gustavo Esteva’s essay therein on “development,” and Arturo Escobar’s (1995) *Encountering Development*. Several other volumes from the early to mid-1990s contained more diverse perspectives that preclude simple categorization, for example *The Post-Development Reader* (Rahnema with Bawtree 1997). By the late 1990s, critiques of these perspectives began to appear, and debate continued through the 2000s, not so much among proponents of anti- or postdevelopment themselves as with progressive critics of these approaches.

That said, it would be inaccurate to suggest that anti- and postdevelopment were entirely original interpretations, born of the development impasse and poststructural turn. Similar concerns, objections and alternative formulations have abounded through most of the history of development as a worldview, discourse, discipline, and set of practices. They were expressed in terms of opposing theories (e.g., the variants of Marxism and political economy) and paradigms or initially radical propositions and approaches to tackle perceived problems, violations, and omissions. Such propositions have included meeting basic needs, bottom-up or grassroots development, feminism and gender mainstreaming, environmental sustainability, and participatory development.

The stimulus for all such critical engagements has been strong objections on philosophical, ethical, and/or practical grounds to the flawed nature of postwar developmentalism and its impacts. Had such interventions been overwhelmingly positive, the critical concerns would have been far smaller in scope and number. However, conventional development as espoused and practised by the principal international financial institutions, donor agencies, national elites, and middle classes in low- and middle-income countries has often distorted conditions on the ground, had unintended negative consequences, or actually failed. Lessons of experience are often not learned and mistakes are then repeated as a result of “institutional failure” of various kinds (including corruption) or deliberate ignoring of evidence and recommendations on account of inconvenience, additional cost, pressure of time, vested interests, and/or misplaced faith that the implementing “experts” know best. Examples range from large dams to overambitious integrated rural development schemes, industrial “white elephant” projects, inappropriate urban housing projects, and diverse “soft” programs ill-suited to local conditions and perceived needs.

From an anti- or postdevelopment perspective, the common feature of all these problems and examples is their underpinning by an absolutely or relatively uncritical Western-centric developmentalist worldview. This perceives the objective of development as being to bridge material and other gaps and through large-scale interventions and modern technology to enable poor countries and communities to “catch up” with rich countries. Such “modernization-as-development” is generally driven by economic and technological considerations in an ideological lineage stretching back through the Cold War to the Marshall Plan, in a complex mixture of altruism and material and geopolitical self-interest (Simon 1998). However, this frequently neglects or devalues local institutions, traditions, cultures, and practices, and transforms or undermines relative or absolute self-reliance, environmental sustainability, and collective senses of wellbeing and quality of life. As a result, supposed beneficiaries as well as those who are excluded and deprived of previous resources are often alienated. Moreover, from this perspective, not only are resource-intensive, consumption-oriented,
and individualistic Western lifestyles almost universally desired, but they would be unsustainable at present for all 7 billion people on Earth.

**Critiques of anti- and postdevelopment**

One of the most widely articulated criticisms, especially of antidevelopment, is that of oversimplification and exaggeration. In particular, the considerable diversity of visions and versions of development among official agencies and institutions, as well as of practice and outcomes, are elided into a monolithic and demonic stereotype of deliberate and even cynical exploitation. To argue that the World Bank and the United States Agency for International Development (USAID), for instance, are homogeneous bodies, let alone indistinguishable from other multilateral or bilateral agencies, is not credible. If nongovernmental organizations (NGOs) and other actors are included, the diversity on every variable is greater still.

A variant of this critique is that of unjustified extrapolation, in which particular case studies are used as the basis for global generalizations. A good example of this is the way in which Escobar's (1995) very detailed anthropological dissection of his native Colombia's development experience and the interventionist role of USAID, in particular, then forms the basis for globalized antidevelopment claims. This is doubly ironic because these claims constitute precisely the same homogenizing and universalizing discursive fallacy for which anti- and postdevelopment advocates castigate conventional developmentalists, and because it is probably mainly on the basis of this broader argument that the book has achieved its classic status. The principal reason for regarding this book as a canon of anti- rather than postdevelopment in terms of the distinction being made in this entry, despite Escobar (1995, 215) himself positioning it as post-, is that it constitutes overwhelmingly a critique of development; there is only a very preliminary attempt to consider alternative possibilities at the end. It is only fair to point out, however, that Escobar's subsequent writings reflect a moderation of this strident position as he engages more actively with the search and struggle for indigenous/local alternatives to development “from below.” He thus remains a leading postdevelopment voice.

The second major critique is that part of the oversimplification for which antidevelopmentalists are often taken to task is their overlooking or denial of the real gains achieved worldwide in terms of improved educational standards, health and welfare, and material standards of living that are reflected in widespread increases in life expectancy, improved public health, and reduced poverty. To be sure, this progress has been geographically and socially uneven but the Human Development Index and numerous other indicators bear this out. Moreover, the increased numbers of people worldwide living in absolute and relative poverty reflect growing populations rather more than people falling into poverty through development failures.

Third, as mentioned above, criticisms of development flaws and failings from within and beyond the ranks of conventional development organizations date back to the earliest overtly economic development initiatives. Some of these critiques have been far more than cosmetic and have generated entirely new understandings and approaches, the value of which has come to be widely understood to the point that they have been adopted or adapted by such agencies – the process often dubbed as “mainstreaming.” These provide evidence that institutions are not impervious to criticism or cynical and can “learn by doing,” to paraphrase an old World Bank motto.
Other radical and emancipator visions of development do not renounce development in its entirety but seek to harness appropriate elements in promoting positive, endogenously driven change. To this, postdevelopment proponents counter that reforming development is futile on account of its inappropriate (some would say objectionable) philosophical and ideological underpinnings and hence needs to be rejected or at the very least challenged fundamentally at that level. Essentially, it is the intellectual, philosophical, and/or political enterprise or “project” of development(alism), that is, the explicit intent to develop others and its underlying motives, that are fundamentally flawed rather than people’s or communities’ legitimate desire for and need of absolute and relative improvement in their material and psychological wellbeing. These latter constitute the basis of postdevelopment, starting with the experiences and aspirations of those “other(ed)” people and communities.

In addition and somewhat ironically in this sense, postdevelopment has been accused of reactionary or neopopulism in terms of a rather romantic and uncritical view of (particularly precolonial) local/indigenous traditions and cultures. Such positions avoid or ignore gender, caste, and other inequalities, exploitation, violence, and other ills in the implicit construct of indigenous cultural authenticity. This constitutes the fourth element of critique but, as Ziai’s nuanced assessments (2004, 1050; 2007) point out, this – like most other criticisms – is unfair in relation to many postdevelopment authors and highlights the dangers of generalization in relation to a very diverse and heterogeneous field. Indeed, even particular authors and individual works by an author reveal inconsistencies in terms of their vulnerability to the range of criticisms outlined here – for instance accepting the dynamic potential of cultural or biomedical hybridization while insisting on total rejection of development and all it stands for.

Fifth, some two decades since the emergence of identifiable anti- and postdevelopment critiques one might pose the question of what difference these dissident voices have made on the ground. In other words, how have they affected actual policy and practice in support of the aspirations and agendas of those at the bottom of the global pile? Undoubtedly some specific grassroots initiatives and NGO operations have been informed by such analysis and are operating more appropriately as partners and facilitators/advocates to people and communities. Despite many exhortations, however, very few published case studies exist to validate this assertion. Moreover, as with most of the poststructural perspectives and agendas, there is little evidence that anti- and postdevelopment have gained significant purchase within the larger development actors and agencies, where it appears to be business more or less as usual. Given the foundational principles of anti- and postdevelopment and the great momentum and vested interests behind such institutions, this is hardly surprising – indeed, anything else would be surprising. This in turn might be seen as at least partial validation of the common critique of anti- and postdevelopment as being more of an intellectual fashion than a practical guide and hence principally of value in academic debates and career promotion rather than being useful in the toolkit or field. Such criticisms emerged rapidly in the late 1990s and have also been leveled at most poststructural approaches (Simon 1998, 2006; Matthews 2008). It is also noteworthy that the leading postdevelopment voices have, with few exceptions, been academics or academic activists from or now based in wealthy countries. Whether the lack of purchase of these ideas in “southern” academia reflects perceptions that they have little relevance or
whether – as postdevelopment adherents would surely argue – such academics are themselves too intimately bound up with modernity and development, to abandon them is a moot point.

Transgressing boundaries in a globalized world

Such dilemmas are inevitable in the highly contested intellectual, policy, and project terrains constituted by development and no definitive resolution appears imminent or indeed feasible precisely because of the widely divergent perspectives, interests, and experiences. Beyond academia and some radical NGOs, it is likely that the principal recruits to agendas identifiable as substantively postdevelopmental will be individuals and groups (however defined) who feel irretrievably marginalized from the benefits of some form of development or who have been severely adversely affected by development schemes. Those who perceive some gains or hope of benefits are more likely to avoid the probable risks of total rejection, if indeed they even have any option. For instance, inhabitants of a river valley where a large hydropower dam is proposed and who will be displaced by the rising waters will surely oppose such a destructive scheme unless offered generous and appropriate compensation. In their struggle, they will probably seek support from civil society, NGOs, and internationally networked pressure groups. That does not mean, however, that they would reject or not appreciate access to reliable electricity (which such a scheme would be very unlikely to deliver them, even if they were relocated to safe ground nearby) and its potential benefits. Hence they would probably be far more positive about a meso- or micro-hydro scheme that would deliver them safe and sustainable water and electricity supplies without the massive capital cost and social, livelihood, and environmental destruction of a large scheme. Indeed, they might even lobby proactively for such a project to be commissioned. If it were undertaken as a genuinely participatory partnership between local communities and other partners (NGOs, private firms, official agencies) that ensured an appropriate degree of local “ownership” and output that was principally for local use in the first instance, all stakeholders would benefit. This would be a good example of appropriate and sustainable development and would address the objections raised by postdevelopment critics as outlined here, but whether it would qualify as being consistent with postdevelopment is a moot point.

Perhaps one lasting legacy of postdevelopment thinking will be a certain shift and sensitization of more critical thinking rather like that achieved by other poststructuralist approaches. In particular, as now acknowledged from different theoretical positions, ongoing globalization and progressively greater differentiation among countries between as well as within conventional “worlds” or other global categories have rendered such categories increasingly redundant except perhaps in world-system historico-evolutionary terms. That being so, persistent questions about why “development” is undertaken only in former colonial regions and not in some of the poorer countries of southern, central, or eastern Europe, for example, or in impoverished and deprived inner cities and peripheral regions of wealthy countries, become even more apposite. Historically, western Europe was the proving ground for postwar development efforts through the Marshall Plan; if semantic sensibilities nowadays preclude redeploying it in such contexts, perhaps it really is time to abandon this terminology and all it represents and to apply other, potentially less paternalistic or problematic vocabularies globally. Paradoxically
too, postdevelopment sensibilities suffuse the growing concern to develop “southern theory” as a counter to the traditional “northern” dominance of social theorization, although this self-ascribed label unhelpfully perpetuates the very dichotomous thinking that poststructural approaches supposedly eschew.

SEE ALSO: Dependency theory; Development; Developmentalism; Modernization theory; Postcolonial geographies; Power and development; States and development; Sustainable development

References


Further reading

Posthumanism

Franklin Ginn
University of Bristol, UK

Posthumanism is a collection of diverse ways of thinking that are highly skeptical about the centrality of the “human” in Western philosophy and politics. Where humanism holds that unique properties such as language, tool use, culture, and so on enable humans to transcend nature, posthumanism emphasizes the different ways humans are continually produced through material forces, discursive regimes, and through nonhuman agencies. One of posthumanism’s key aims is to dissolve binary distinctions that characterize humanism, most notably culture/nature and self/world. Posthumanism is associated with the rise of poststructuralism through the latter half of the twentieth century, as well as struggles against hegemonic definitions of the human in feminist, postcolonial, anti-racist, and queer politics. It is also closely allied with philosophies that emphasize the embodied, material, and vital nature of human life. Within geography there has been some skepticism about whether posthumanism is any more than another episode in a procession of -isms. More receptive geographers have emphasized that posthumanism is broader in scope and in political and philosophical ambition, and captures many other -isms within its own orbit as ancillary concepts.

Posthumanism is in some sense easier to understand and identify due to recent changes in biological sciences that blur the boundaries of the human, such as genetic modification or xenotransplantation, as well as new technologies which can subvert or extend human agency in various ways, from electronic surveillance to unmanned military drones. Posthumanism is commonly identified therefore as both an epoch and a theoretical perspective. While the “post” in posthumanism may be taken to signal a transition to a new stage after humanism, many commentators argue any quantifiable historical shift is less important than the radical ethos of posthumanism’s attempt to deconstruct and move beyond a unitary, Western-specific idea of the human. Some scholars have argued that posthumanism is symbiotic with humanism, in that it relies on the figure of the human for a constituting other; from this perspective posthumanism risks reifying humanism as something grander than a historically and geographically located mix of practices, ideas, and technologies. Furthermore, some strains of posthumanism can often reproduce certain humanist commitments to ideals like freedom and democracy. In this sense posthumanism may present an evolution of humanist thinking, rather than a radical break. Others argue that posthumanism offers productive ways to move decisively beyond the human, but only if we reorient our horizons toward an ontology of life that is more radically open-ended, relational, and emergent.

Decentering the human

It is possible to identify a series of four logics through which posthumanism works to decenter the human. Each of these logics implies that the human should be taken neither as a common sense dogma, nor as an unproblematic grounding for ethics or politics, and works to replace the
human as the sole concern of analysis. Each has been influential in geography, the social sciences, and humanities.

The first is biological. Western Enlightenment thought enshrined a firm divide between humanity and the rest of nature. Advances in biological science have rendered such a view increasingly untenable, however. Beginning with Darwin, who placed humanity in a genealogical continuum with other species, research in diverse fields, including ethology, comparative genomics, and neurophysiology, has concluded that many other species share supposedly exclusively human characteristics, including language, tool use, and consciousness. Social theorists such as Donna Haraway have also emphasized the way the human body relies on alliances with other organisms, including slime, bacteria, or domesticated animals, for its continued operation. For these theorists, posthumanism involves tracing the ethical, political, and ontological implications of this biological decentering. Animal geographies have charted the constitutive inclusions and exclusions of nonhuman creatures for human culture and identity in varied settings. Eco-Marxism has, at a larger scale, shown how nature is not separate from society, but is produced by capitalism, demonstrated most spectacularly by the power of biotechnology to make new life. Others have pointed to shifts in governmentality wherein the object of intervention is no longer simply the traditional human subject, but aspects of human biology, including genes. Posthumanism therefore emphasizes distributed biological interdependencies and shared life, rather than a divide between nature and culture.

The second decentering concerns subjectivity. Where humanism posited an autonomous, rational being in charge of its own actions and destiny – the knowing subject – posthumanism emphasizes how the subject is given over to and produced by other forces beyond its own control. Around the turn of the twentieth century, the founder of psychoanalysis, Sigmund Freud, showed how much human behavior was driven by unconscious desires and fantasies. While psychoanalysis has had a modest impact on geography, recent theories have been more influential. Emerging from poststructuralism, Judith Butler’s notion of performativity was crucial in eroding the idea that a subject had any kind of essence at its core. Butler argued that gender difference was not a matter of nature, but of social performance. Performativity stresses the malleable, motile, and historically contingent character of subjectivity, in contrast to an essential human nature that takes the same basic form regardless of time and space. In geography, since the 1990s great attention has been paid to difference in the creation of subjectivity, both through discourse and representation, and also more recently through embodiment, materiality, and emotion. Far from being an autonomous, knowing subject, the human is seen by posthumanism as a process made through many more-than-human flows.

Third, technology has decentered the rational human. Karl Marx was one of the first commentators to write systematically about how technology and wider systems shape the conditions for human nature. Actor–network theory has also been deployed to show how technology is not a passive object operating at the behest of a potent human agent. Instead, actor–network theory exposes how agency circulates between heterogeneous actors, including technologies of all kinds, in complex, hybrid circuits. Posthumanism therefore stresses that technology is not something added to an already existing human, but is something that makes up humans. This is what Jacques Derrida called “originary technicity” and Graham Harman, building on Heidegger, calls “tool-being.” From a posthuman understanding, the human is always

---

POSTHUMANISM
POSTHUMANISM

a human-with-technology (and always has been). Indeed, posthumanism emphasizes that the human itself is a product of what Agamben (2004) called the “anthropological machine,” a system which produces self-recognition and partial differentiation of humans from other entities. Aspects of quantitative geography have also worked to decenter the dominant notion of the human. In the 1950s and 1960s, cybernetics conceptualized humans not as subjects of their own lifeworlds, but as nodes within a system that merely reacted to external forces. While cybernetics and systems theory were only models they were influential in producing ways of thought, for example in managing environmental and economic crises. More recently, advances in computational power and complexity theory have contributed to the rise of agent-based modeling in geography. Modeling human decision-making and behavior as part of emergent systems further decenters the idea of an independent, self-aware agent in favor of one reacting in complex ways to stimuli and feedback in sociotechnical systems. In addition, under the auspices of government and corporate surveillance a growing ecosystem of digital information, including self-replicating code, adaptive algorithms, and “big data,” increasingly shapes the way that humans think and react.

The fourth way in which the human is decentered is more recent, and can be termed “the planetary.” While the deleterious impact of human activities on Earth systems and other creatures has long been of concern to geographers and others, in recent years awareness of the unprecedented temporal and spatial scale of these impacts has led some to christen a new geological epoch, the Anthropocene. The Anthropocene remains contested. For some it represents the apoctheosis of humanism, with progress reaching a new level of intensity and extensiveness. For others, this new epoch firmly embeds humanity in the Earth’s material and energetic flows. This latter version of the Anthropocene shows the extent to which modern humans have internalized ecological and geological Earth forces, and also humanity’s differentiated vulnerability to such forces. This planetary decentering means the human is no longer an agent in charge of its own destiny: instead, the human is beholden to planetary forces beyond its control.

Each of these four logics of decentering has been influential in posthumanism. We can also discern three broad approaches to posthumanism: as epoch, as epistemological style, and as ontological condition.

Posthumanism as epoch

This populist strand of posthumanism takes the proliferation of new biotechnologies and technoscientific objects as a hallmark of a new historical epoch and a new stage of human evolution. Broadly speaking, much of what was previously limited to the realm of science fiction has come into being and changed the way many humans interact with the world. Such a list would include, but by no means be limited to: from William Gibson’s cyberspace to the pervasive chatter of social media; from cyborg military machines (the films RoboCop, Blade Runner) to unmanned drone warfare; from the paranoia of the cult 1990s television show The X-Files to the burgeoning discipline of exobiology; from speculative tales of colonizing other planets (Kim Stanley Robinson’s Mars Trilogy) to proposals by private corporations for extra-earth mining and space exploration ventures; from dreams of electric sheep (Phillip K. Dick’s Do Androids Dream of Electric Sheep?) to cloning Dolly the sheep and the creation of synthetic bacteria.

The list would encompass the breakdown of boundaries between flesh and information,
POSTHUMANISM

with genomics and bar-coding life for medical, scientific, and commercial purposes now part of biocapital enterprise across the world. These technological novelties are part of both spectacular and mundane processes changing the nature of everyday life for many. They each involve many “things” – DNA (deoxyribonucleic acid), images, bodies, instruments – that defy easy classification as natural/cultural or artefactual/given. Posthumanism as epoch suggests that all these processes have passed some threshold, putting the human in new terrain never before realized.

Optimists see this new horizon of humanity as an event to be celebrated. This position is popular among libertarian pro-technology and innovation groups, in particular certain science fiction writers or authors, in particular certain science fiction writers or authors in, for example, Wired magazine. They see playing with new ways of being human as exhilarating and technological creativity as something to be encouraged. They welcome boundary crossings between machine, human, and animal. They also inflate the potential for science and technology to make good on current speculative proposals, such as downloading human consciousness as software. This strain of posthumanism is also, however, ambivalent about inequalities, skeptical of state power, and welcoming of the creative destruction that capitalism brings. Instead, ingenuity is crucial, as amid an ever-faster pace of change the slow are left behind and the most adaptable posthumans thrive. This populist strain therefore in fact follows quite closely the established tenets of humanism: the potency and sovereignty of the individual; the march to ever-greater progress and prosperity; the promise of tomorrow to make good the problems of the present. Critics see this strain of posthumanism as an opportunistic attempt by advanced capitalist interests to graft a promissory posthumanism on to existing neoliberal subjectivities.

Pessimists agree that we have crossed a threshold into a posthuman era, but dispute that this is something to be celebrated. Rather, they see the proliferating evidence of posthumanity as deeply worrying. Pessimists include religious commentators, deep green environmentalists, and those who wish to defend the legacy of the European Enlightenment and its humanist commitments. They oppose, each for different reasons, genetic modification, nanotechnology, geoengineering, and other developments that they see as polluting the sacrosanct boundaries between humans and other creatures. Conservative thinkers like Jürgen Habermas and Francis Fukuyama, for example, argue that genetic manipulation and xenotransplantation are morally wrong on the grounds that they alter human nature. Deep greens oppose genetic modification because they believe that biotech corporations do not have the right to make new forms of life, and also because they worry about the unpredictable mixings that occur when new creatures are added to the Earth’s biodiversity. More broadly, mainstream environmentalists remain wedded to the idea of saving a transcendent nature, and erasing signs of human pollution of that separate realm. By contrast to the optimists’ desire for ever-faster change and ever-greater motility of human form, the pessimists want to halt current changes and where feasible turn back the clock to a time when categories were clear and the human was pure.

Posthumanism as epoch tends to describe hyperbolically new technological developments for rhetorical advantage to support existing ideological commitments. Both optimistic and pessimistic interpretations of posthumanism as epoch share a reliance on the figure of the human. For the optimists, the potency of posthumanism is to be measured by its distance from the historical figure of the human. For the pessimists, the dangers of posthumanism can be gauged from
the degree of deviation from the established norm of the rational human. Posthumanism as epoch is generally seen as the conceptually weakest of the three approaches to posthumanism.

Posthumanism as epistemological style

Feminist, postcolonial, disability, anti-racist, and queer studies have critiqued the normative assumptions at the heart of humanism. The central insight of this deconstructive stance is that “the human” acts with regulatory force as a standard of sameness to produce insider and outsider groups. This strand of posthumanism therefore interrogates different ways of knowing the human and their political effects.

Historically, a powerful standard of the human emerged from the Western Enlightenment and colonial encounters with various other peoples deemed by Europeans to be inferior in one way or another: due to their environment, physiology, innate intelligence, or distinct cultural evolution, for example. The figure of the white, male, full-bodied, and heterosexual human was deployed to justify a series of epistemic and physical violences against subaltern groups. Edward Said was one of the first thinkers to articulate a reasoned critique of the colonial experience and its injustices. Said’s Orientalism exposed how ideas of European exceptionality relied on a constitutive other, the Orient. Said argued that reason/enlightenment and violence/barbarism were not mutually exclusive, and that both could be found within humanism. Said turned Western, rational thought back against itself, critiquing humanism in the name of a more cosmopolitan humanism.

Michel Foucault’s ideas were another major force in anti-humanist critique from the 1960s on. Foucault argued against the autonomy and self-determination of the human subject, and against seeing the human at the center of history. Foucault’s work on madness, sexuality, and biopower showed how the production of difference was an internal dynamic within Enlightenment humanism. Instead of a dominant ideal human, Foucault emphasized how the human was socially constructed and power-laden. Posthumanism as epistemological style therefore replaces fixed identities with lines of difference that cut through gender, ethnicity, corporeality, race, sexuality, and so on.

One of posthumanism’s political aims is to enable people to decode, resist, and transgress their subjectification as certain kinds of subject. Feminist geography has examined the construction of gender and the inequalities that flow from the naturalization of rational man as the measure of all things. Feminist geography has been a particularly influential subdiscipline, with an important landmark being Gillian Rose’s Feminism and Geography (Rose 1993), which exposed the patriarchal assumptions inherent in the discipline. Postcolonial geography has brought self-critical awareness of how the discipline helped produce discourses of European exceptionalism, as well as to the political implications of how history is written. Postcolonialism critically engages with the continued legacies of colonialism, including how humanist ideals and their exclusionary universalism are written into social, cultural, architectural, and economic forms.

Posthumanism as epistemological style also encompasses the tenets of knowledge production dear to humanism. While the scientific method, steeped in empiricism and critical rationalism, was more a model of and less a rule book for scientific inquiry, science and technology studies (STS) has exposed the supposedly objective, disembodied, and universal idea of knowledge as a myth. STS emphasizes the role of culture, political economy, and the mobilization and alliance of myriad objects with
contextually specific procedures in producing and mobilizing knowledge. Scientific knowledge remains thoroughly marked by its geographic and historical context, rather than floating freely. Feminist epistemologies have taken the situatedness of knowledge as a methodological and ethical imperative. For instance, standpoint theory and situated knowledges both stress that knowledge production should acknowledge its partiality and locality, and that done well this produces more acceptable knowledge than false claims to universality. Posthumanism is closely allied with these ways of knowing, in which the subject is neither all-knowing nor transcendent of their particular circumstances.

Geographers have also deconstructed the divide between human and nonhuman, although not always under the banner of posthumanism. In the 1990s, animal geographies were reinvigorated by geographers drawing on animal philosophy and Foucauldian ideas of discipline and spatialized difference. This subdiscipline explored the subjugation and suffering of nonhuman creatures, which is often justified by the “great divide” (Latour 1993) between humans and all other animals. Animal geographies also exposed the ways in which representations of animals and other parts of nature provided ideological justification to naturalize certain forms of human behavior, such as heteronormative sexuality. Conceptually, animal geographies sought to deconstruct the nature/culture dualism not only in exterior space but also as it ran through the human, examining what Agamben called a great “caesura” at the heart of the human. Posthumanist animal studies therefore seek to account for the animal within as well as animals in diverse settings such as the city, the farm, and the laboratory. Feminist scholars have been among the foremost to point to the ethical implications of denying animality in favor of a constructed rational human. This denial has led to a “crisis of reason” and fuels environmentally destructive behavior (Plumwood 2002), so deconstructing ideologically charged human rationality became not just a matter of epistemology but also an ethical necessity.

Posthumanism as epistemological style maintains certain commitments to the importance of representation, and often works by deconstructing the self/other binaries at the heart of identity-making practices. This strand of posthumanism is therefore often concerned with denaturalizing nature, offering counter-readings to the messages encoded in nature and in representations of nature. The impact of posthumanism as epistemological style has been great. Replacing a pre-given, unitary standard of the human with the ongoing production of different subjects has enabled geographers to study the power relations of different processes of subjectification and has opened up important political spaces for anti-racist, anti-colonial, and anti-heteronormative politics. These perspectives emphasize the need to go beyond extending existing modes of representation to subaltern groups; they call for a deeper reassessment of the model of rational man as the foundation of modernity. The commitment to working at an epistemological level, however, means that this branch of posthumanism eschews the bodily, material dimensions of life. In other words, it is more concerned with the figure of a certain kind of human and posthuman than with the fleshy beings themselves. Therefore some scholars have pushed posthumanism to think of a more relational subject with embodiment, sexuality, affectivity, empathy, and desire as core qualities. Moreover, some critics have suggested that, ironically, posthumanism as epistemological style remains too reliant on humanism, not just as a necessary straw figure for criticism, but also because it usually ends up invoking humanist
POSTHUMANISM

ideas such as freedom, democracy, and progress to articulate its political goals.

Posthumanism as ontological condition

This final approach to posthumanism is rooted in anti-essentialist, process-oriented, and vitalist philosophies. It is associated with the ideas of Bruno Latour, Gilles Deleuze, and Isabelle Stengers, among many others. These thinkers emphasize dynamic processes of becoming instead of a world composed of discrete entities. In this ontology, human characteristics are merely temporary intensifications of matter and energy, indivisible from other agglomerations of matter. Accordingly, posthumanism as ontological condition is closely allied to geographic studies of embodiment, materiality, emotion, and the nonhuman.

Within geography the ideas of Bruno Latour and others associated with actor–network theory have been particularly influential. Latour argued that the modern settlement between nature and culture was merely an illusion and that the world was constituted through flows and connections between heterogeneous actants. Latour argued that the human was a result of purifying practices that worked constantly to separate a hybrid world into two realms, nature and culture. Actor–network theory therefore conceptualizes the human as an ongoing performance spun out of many different agents and agencies. Latour’s ideas have been recently supplemented by several other bodies of thought. First, concepts of multiplicity have augmented the flat ontology suggested by actor–network theory. Instead of one network, there is a multiplicity of overlapping networks, which create a denser, more complex sense of space. Second, geographers and others have drawn on the geophilosophy of Gilles Deleuze to produce assemblage theory, which accounts for conditions of multiplicity and describes open, emergent systems.

Critical of the text-bound excesses of postmodernity, scholars such as Nigel Thrift have highlighted the importance of forces that work on a pre-, ex-, or postrepresentational register in human becoming, formulating what has become known as nonrepresentational theory. Of particular importance has been the concept of affect, which is broadly defined as the capacity of a body to affect and be affected by other bodies through extralinguistic forces. Through ideas like affective atmosphere and affective politics, geographers are beginning to understand how life is shot through by transhuman energies. In related vein, cultural and historical geographers have reconsidered phenomenology, and in particular the philosophy of Merleau–Ponty, to explore how the world is brought into being through sense and experience, always in particular circumstances.

Feminist and postcolonial scholars have also cautiously brought the body back into their research agenda. While deconstructive posthumanism argued powerfully that race, gender, and sexuality were socially constructed rather than naturally given categories, scholars such as Rosi Braidotti have argued for re-ontologizing the body and examining the ways that difference is lived, felt, sensed, and performed. For this feminist perspective, bodily – and in particular sexual – difference becomes a political position from which to build alliances, rather than a problem to be overcome. Similarly, geographers have shown how race is materially produced through affective and bodily assemblages, as well as constructed through discourse. Both feminist and race scholars are mindful of the dangers in returning to the body, given the political effectiveness of constructivism and the dangers of naturalism. But the political message of such work is that a shift in focus from discursively
POSTHUMANISM

codified subject to processes of subjectification that encompasses body, technology, and the stuff of life generates a politics open to reappropriating the play of differentiation. Posthumanism emphasizes the mutability of being such that politics becomes a matter of interfering in flows of energy and matter to produce new possibilities for becoming otherwise.

This branch of posthumanism also brooks no ontological hygiene between humans and other species. Of particular resonance has been Donna Haraway’s material-semiotic cyborg and, more recently, companion species. Haraway argues that beings are a “knot of relationality” that draws in many species and many histories. Haraway’s work has also consistently reproached environmentalists for their anti-technological stance. Posthumanism as ontological condition draws on the concepts of originary technicity to argue against any possibility of human/technology purity, and instead argues for a critical appropriation of the technologies of capitalism.

There are several key differences between posthumanism as ontological condition and other forms of posthumanism. First, it denies that there has been any historical transition to a new, posthuman era. Instead, the argument is that we have never been human (Haraway 2008). In other words, the doctrine of humanism never fully encapsulated the condition of becoming human. This branch of posthumanism does recognize, however, that processes of mixing and decentering have accelerated in intensity in recent decades. Second, its vitalism – the notion that matter is agentic, self-organizing, and mutable – distinguishes this approach from the priority accorded to meaning, representation, and identity politics by the other forms of posthumanism. Third, its ethics and politics are less grounded in humanist commitments to universal justice, rights, freedom, and so on. Instead, the ethical-political commitments of ontological posthumanism are situated, context-specific, and open-ended. They exhibit a strong experimental flavor, and are organized through lines of connection and alliance which cut across national, species, and other traditional forms of difference. Often, research practice uses participatory methods and is aimed at encouraged resubjectification of both researcher and researched. Therefore much turns on an enlarged sense of embeddedness and awareness of relation to others, which creates new geographies of ethical response and responsibility. This posthumanist ethical-political project is generally underspecified, with speculative appeals to care and flourishing common but still not fully substantiated or worked through.

These ethical and political commitments mean that posthumanism has been met with some skepticism by geographers committed to more established forms of politics. Some critics point to the similarities between an open-ended, experimental, and context-bound ethical project and forms of uncertainty and precarity produced by neoliberalism. Others note that the posthuman possibilities presented by complexity and the mutability of matter have quickly lost their political potential as biocapital incorporates such resources into new regimes of accumulation. Eco-Marxists and political ecologists are debating the utility of posthuman immanent ontologies which do not use structure or meta-explanatory devices such as capital. Still others have identified a blind spot in posthumanism, in that it critiques humanism from the inside of Western thought. These critics have argued that there are plenty of human cultures that never shared Euro-American ideas of the human and its separation from nature and technology. They see the lack of attention to Indigenous knowledge and cosmography as extending the dangerous Euro-American exceptionalism of humanism. From such a perspective, while posthumanism as
ontological condition cleaves to posthumanism as a transhistorical theory rather than as an epoch, this maneuver still denies the way in which posthumanism has emerged from particular historical conditions and particular sites of privilege to become a key concern of our times.

**SEE ALSO:** Actor-network theory; Animal geographies; Bodies and embodiment; Difference; Nature and corporeality; Nonrepresentational theory; Postcolonial geographies; Poststructuralism/poststructural geographies; Subaltern; Whiteness

**References**


**Further reading**


Postmodernity

David B. Clarke
Swansea University, UK

On the face of it, postmodernity refers to a new historical epoch, beyond modernity. When it first rose to prominence in the 1980s, the idea was enthusiastically greeted by many geographers: the suggestion was that, whereas modernity privileged history and time, postmodernity privileged geography and space (Foucault 1986; Soja 1989). “The great obsession of the nineteenth century was … history,” wrote Foucault (1986, 22), “with its themes of development and of suspension, of crisis and cycle, themes of the ever-accumulating past” – whereas, Foucault declared, we now seem to have entered the “epoch of space … the epoch of juxtaposition, the epoch of the near and far, of the side-by-side, of the dispersed.”

Even without beginning to evaluate such a bold contention, the sense in which “postmodernity” defies easy comprehension may already be apparent. Since postmodern sensibilities regard “history” as itself a modern notion (not what actually happens out there in the world, but a particular human view of it), the end of modernity would also spell the end of history – which would seriously undermine the very idea that there can be a “next stage” of history such as “postmodernity.” If we are to avoid plunging into this kind of paradoxical tailspin, the implication must be that postmodernity, if it means anything, cannot refer to a new historical epoch. Rather than a periodizing concept, the alternative would be that postmodernity marks the realization that the modern sense of history as a chronological linear progression – onward and upward; destination Utopia – was always the hopeful vision of a particular kind of society. Modernity’s vain hope, the postmodern thesis contends, has revealed itself as supremely illusory. The specters of dystopia, ruination, the return of the repressed with a vengeance (fear, terror, risk, unreason) are once more on the prowl. “I define postmodern as incredulity towards metanarratives” (“big stories”), pronounced Lyotard (1984, xxiv), attesting to a loss of faith in the grand march of progress. Viewed in retrospect, the strikingly modern idea that reason, securely entrusted to the hands of the powerful – bureaucrats, scientists, politicians, industrialists, experts of all kinds – could deliver a utopian future for all has come to be recognized as a self-legitimating modern myth.

Although modernity’s penchant for imposing a rational order on an unruly world sounds superficially appealing, those on the receiving end – from nineteenth-century craft laborers, forced to adapt to the harsh regime of factory production, to those deemed genetically inferior and condemned to the Nazi gas chambers of the twentieth century – might offer a very different perspective. Tellingly, the most iconic of modern spaces is Bentham’s eighteenth-century design for an ideal prison, the Panopticon, which imposed order by actively soliciting the compliance of its inmates, placing them in a position where they were, at least potentially, perpetually subject to the gaze of the prison guards. Several centuries’ experience of modernity, the balance sheet of which is hardly limited to entries on the plus side alone, means that the promise of a final solution to the world’s ambivalence no longer
POSTMODERNITY

appears as either viable or, indeed, desirable: the twentieth century went down as the most violent in history. Released from the delusion that ambivalence and uncertainty are temporary irritants, set to be eradicated once and for all, we are finally in a position to realize not only that such qualities represent the world’s ineluctable condition (its “default setting,” so to speak), but also that they are not necessarily irritants. The idea that there might be something to celebrate in postmodernity resides precisely in this sense of release from the stranglehold of power-assisted “rationality.” To paraphrase Nietzsche, “the way does not exist,” and only by shattering the idols of modernity can new possibilities emerge. So what does “postmodernity” look like? All we may say with confidence is that it “means many different things to many different people:"

It may mean a building which arrogantly flaunts the “orders” prescribing what fits and what should be kept strictly out to preserve the functional logic of steel, glass and concrete. It means a work of imagination that defies the difference between painting and sculpture, styles and genres, gallery and street, art and everything else. It means a life that looks suspiciously like a TV serial, and a docudrama that ignores your worry about setting apart fantasy from what “really happened.” It means licence to do whatever one may fancy and advice not to take anything you or others do too seriously. It means the speed with which things change and the pace with which moods succeed each other so that they have no time to ossify into things. It means attention drawn in all directions at once so that it cannot stop on anything for long and nothing gets a really close look. It means a shopping mall overflowing with goods whose major use is the joy of purchasing them. It means the exhilarating freedom to pursue anything and the mind-boggling uncertainty as to what is worth pursuing and in the name of what one should pursue it. (Bauman 1992, vii)

None of this entails that postmodernity is some sort of panacea, allowing for the proliferation of spaces capable of accommodating life in all its infinite variety. If postmodernity is sometimes held up as the chance for difference to flourish, on other occasions it looks to be closer to the glossy, hyperreal, commodified world of the consumer society.

In this light, it is unsurprising that one of the most influential geographical contributions to the debate dismissed postmodernity as nothing more than the “cultural clothing” of late capitalism. For the Marxist geographer, David Harvey (1989), the fact that Lyotard (1984) sought to disqualify the insights of Marxism was enough to reveal postmodernism’s complicity with neoliberal ideology. A fundamental principle of historical materialism is that social consciousness (part of society’s “superstructure”) reflects society’s underlying mode of production (its “economic base”), and not the other way round. Adhering to this principle, Harvey sought to demonstrate that his alleged “metanarrative” was perfectly legitimate – and eminently capable of exposing the ideological basis of fashionable discourses pronouncing the “end of ideology.” To caricature, this was something of an “irresistible-force-meets-immovable-object” confrontation: for Lyotard, the distinction between “base” and “superstructure” no longer held; for Harvey, the economic base of late capitalism was responsible for generating precisely the illusion that the distinction was no longer operative. In toeing the Marxist line, however, Harvey created a backlash from feminist geographers, who saw in postmodernity the chance of an analysis of society and space that embraced more than class struggle. The idea that postmodernity represented a “merely cultural” phenomenon – signaling business as usual for Marxist theory – was more generally seen as inadequate. For Jameson (1984), another
Marxist, postmodernism represented the “cultural logic” of late capitalism: not simply its “cultural clothing” but a fully integral logic, as if base and superstructure had been fused by an unforeseen short-circuit. For Jameson (1984, 83), the vertiginous experience this engendered issued from the “latest mutation in space – postmodern hyperspace,” of which the disorienting interior of downtown Los Angeles’ Bonaventure Hotel provided an apt example: an appropriately commercial, pleasure-orientated postmodern counterpart to the iconic modern space of Bentham’s Panopticon.

Despite tendencies to associate postmodernity with particular spaces and places, of greater significance was the sense in which the changed view of knowledge ushered in by postmodernity pointed to a spatial analog. For Dear (2000, 140), a distinctively postmodern urbanism “is based on a simple premise: that just as the central tenets of modernist thought have been undermined, its core evacuated and replaced by a rush of competing epistemologies, so too have the traditional logics of earlier urbanisms evaporated, and in the absence of a single new imperative, multiple urban (ir)ratationalities are competing to fill the void.” But haven’t cities always been like this? As a number of commentators have shrewdly observed, “postmodernity” has been modernity’s indispensable alter ego from the start; the dissenting voice deflating its grand claims and exposing its illusions that drove it on. What has changed is the balance between the two. Postmodernity is the continuation of modernity by other means or, more accurately, the continuation of the means in the absence of the ends that provided its initial legitimation. Progress and history rumble on, leaving perpetual upheaval and transformation in their wake – but where they are heading has long since been forgotten. If the terminology of “postmodernity” has, after an initial wave of enthusiasm and contestation, by now been roundly abandoned, its disappearance may, paradoxically, signal its triumph rather than its demise: the situation it was meant to grasp having become generalized to the point where the need for specialist terminology has become redundant.

**SEE ALSO:** Marxist geography; Modernity

**References**


It is difficult to distinguish a specific “postsocialist” geography. The scale, the structure, and the level of geographical research and education in former socialist countries depended on their historical legacy and traditional schools created by leading geographers, and on a relative openness to the outside world and other factors. For instance, Polish geography was historically closely related with German anthropogeography, had a strong “natural” wing and developed landscape studies. Polish scholars had more opportunities to keep working contacts with Western partners than did their colleagues from the Soviet Union or Romania. Today, national geographical communities of former socialist countries are facing mostly the same problems as world geography as a whole; however, some common features inherited from the socialist past still can be observed.

From the institutional perspective, a common problem for postsocialist national geographical communities is, first, the decreasing importance of geography at school: it is often withdrawn from the curriculum of certain classes or replaced by a “general” discipline like “natural studies” embracing the elements of different sciences. This tendency undermines the positions of geography in pedagogic institutes/universities and partly the role of academic geography in society. Second, in the universities where geography was represented at the first-tier level (faculty), it is now often incorporated into new larger units with a wider focus (geosciences or natural sciences) or problem-oriented “institutes” (architecture, regional and urban planning, municipal economy, etc.).

On the positive side, in some countries, like Poland, geography remains a valued subject at high school and plays a decisive role when being enrolled in a university to study economics, finance and accountancy, management, tourism, and political science (Piróg 2012). The survival of specialized geographical institutes as parts of national academies of sciences should be particularly noticed. Such institutes play the leading role in geographical research in Hungary, Moldova, Poland, Romania, Russia, Slovakia, Ukraine, and other postsocialist countries where the academies were created or reshaped following the Soviet model, that is, directly run research institutes in different fields. After the collapse of communist regimes there were strong voices calling to relocate most research activities to universities or other institutions but the academies were kept. Even in Germany, where the academy of sciences of the former GDR was abolished after reunification, its geographical institute in Leipzig was transformed into the well-known Leibniz Institut für Länderkunde; however, in early 2014 the Russian Academy of Sciences was deprived of control of its research organizations, and the destinies of its geographical institutes in Moscow, Irkutsk, Vladivostok, Barnaul, and so on are still unknown.

As in other parts of the world, in the postsocialist countries the institutional position of
geography, combining the natural and human sides, is not easy. Usually, like in post-Soviet countries, Poland, Hungary, or Romania, geography is considered as part of the earth sciences but its position is weak compared with geology, geophysics, and so on. Human geographers have a particularly complicated status, as their field does not belong to either earth sciences or the humanities. But this model offers more opportunities for integration of natural and social approaches in studying the interaction between society and the environment, often not really used yet.

A common trend in the development of geography in postsocialist countries is its internationalization and Europeanization – a much wider involvement in the activity of international associations, research networks, and joint projects. The accession of Central European countries to the European Union (EU) allowed geographers to become fully-fledged partners of such programs as Erasmus or ESPON, the European Observation Network for Territorial Development and Cohesion. Together with specialists in regional economy, geographers play a central role in compiling comparable data bases on territorial development, mapping, and analyzing spatial structures of the enlarged EU, carrying out applied research which contributes to territorial cohesion and a harmonious development of the European territory – turning the boundaries between old and new EU members from dividing lines into seams between territorial structures of neighboring countries. Partners from post-Soviet countries can enter into consortiums applying for grants of European Framework Programs. There is also a great number of bilateral agreements between national foundations.

The creation of geographical journals in English is another manifestation of internationalization. For instance, in Russia there are now two new journals in English. The objective of Geography, Environment, Sustainability is first of all to publish papers written jointly by Russian and foreign authors that result from joint projects. Regional Research in Russia is focused primarily on the translation of the best papers that have recently appeared in Russian journals but also accepts original works.

Postsocialist geographical studies focus on a number of common main themes. First, geographers led the work on national atlases, which appeared or are being published in most postsocialist countries – Hungary, Slovakia, Slovenia, Romania, Russia, Ukraine, and so on. A good national atlas presenting a portrait of the country, an inventory of its potential and resources to its citizens, and “introducing” it to a foreign user is a question of prestige and an element of identity and state-building. The work on national atlases justified the geographers’ raison d’être and contributed to a better visibility of the discipline.

Second, the transition to market economy, the change of the political regime, and the accession to the EU have given rise to a great number of new themes and provoked important transformations in the structure of human geography. Economic reforms and European integration are related to growing territorial disparities, polarization of space – further concentration of economic activity around the largest cities usually in capital regions, and marginalization of the periphery, especially eastern regions – of Central-European countries. So, human geographers focused on new social mechanisms producing space, new methods and forms of spatial planning and regional policy, path dependence in regional development, assessment and mobilization of the regions’ endogenous potential, the impact of globalization and integration on different types of regions, and relocation of industry. Much attention is paid to border regions and cross-border cooperation. An important stream of projects and publications concerns...
urban geography: networking of urban systems at different levels and competitiveness of cities within the enlarged EU, metropolitanization, migration, social segregation and social exclusion; recovery of industrial cities and urban edges, and so on.

Both human and physical geography have faced the process of further differentiation and the rise of relatively new fields, such as the study of cultural landscapes, ethnic mapping, political geography, geographical images and representations, critical geopolitics, and so on.

Third, the traditional landscape analysis has been further developed and complemented by the study of the local impacts of global climate change. A strategic direction of physical geographical studies is also sustainable development and landscape management aimed at the improvement of the quality of life. Physical geographers continue to be involved in the selection, delimitation, mapping and management of protected areas, assessment of natural risks and hazards, particularly floods.

Fourth, a highly important new theme is geoinformatics – a tool of geographical analysis integrating natural processes and human activities.

In former socialist countries Russian geography has a place apart because of the size of the country and a relatively large number of geographical institutions, the duration of the socialist period, and specific academic legacy that was related to the original concept of territorial-productive complexes (TPC), or the economic regions as the basis of spatial planning and the administrative division. They were applied in national planning not only in the former Soviet Union but also in other socialist and third world countries.

In the early 1990s, the Soviet heritage in human geography was criticized. Yet in the postcrisis years geographers and economists, particularly from Siberia, found that the TPC theory had a lot to do with the cluster model developed by M. Porter and M. Enright, and some elements of this theory have maintained their relevance in the contemporary conditions. Some other studies pretend to modernize the Christaller–Lösch heritage, or generated some original ideas, such as the geo-evolutional concept of transportation networks, developed by S. Tarkhov.

Soviet/Russian physical geography was focused on the concepts of natural-territorial complexes (geosystems, landscapes) and studies of the relationships between their elements, their functioning, specific time, stability, and dynamics. Pioneering studies on the theory of rational use of natural resources, very close to the contemporary concept of sustainable development, were published as early as in the 1960s but unfortunately remained unknown in the West. Soviet/Russian physical geographers were among the founders of evolutionary geography and the authors of unique paleo-climatic reconstructions, which help to assess the impact of global climate change.

Physical geography traditionally dominates in terms of the number of laboratories, scholars, and papers. In 2008–2010, articles on physical geography made up 74.9% of all papers published in the main four national geographical journals – even more than in the 1980s–1990s (about 70%); however, as the increasing number of papers is now devoted to environment and sustainable development, it is sometimes difficult to separate publications in physical and human geography. The main journals are: Izvestia Rossiiskoi akademii nauk, seria geograficheskaya (News of the Russian Academy of Sciences, geographical series), Izvestia Russkogo geograficheskogo obschestva (News of the Russian Geographical Society), Vestnik Moskovskogo universiteta, seria geografia (Bulletin of Moscow
University, geographical series), and Geografia i prirodnye resursy (Geography and Natural Resources).

In human geography, an obvious shift from economic themes (geography of industry, agriculture, etc.) to social, cultural, ethnic, and political problems can be observed. During the period 2008–2010 these new themes accounted for 26.2% of papers that appeared in the same four journals (during 1980–1982: only 8.0% and during 1990–1992: 15.0%) (Kolosov and Treivish 2009). In the late 2000s the share of articles on economic geography increased again and in 2008–2010 reached 24.0%. This rise was related to the ongoing economic growth. The share of papers on population geography and urban geography, as well as on general, theoretical, and regional topics remains relatively stable—respectively, 13.8 and 22.2%. The interest in the geography of natural resources and environmental problems interpreted from the perspective of properly human geography seems to be declining (13.8% of papers). All in all, human geography has thus become more pluralistic, both theoretically and empirically.

SEE ALSO: Economic geography; Internationalization; Political geography

References


Further reading


Poststructuralism is a response to structuralism, a mid-twentieth century French intellectual movement dedicated to developing “scientific” approaches within the social sciences and humanities. The structuralists were organized against the abstractions of Western humanism. Poststructuralism addresses both movements with a range of critiques. It claims that the West’s traditional universal narratives about “mankind” invariably pushed many gendered, queer, raced, and non-Western “Others” to the margins. At the same time, it challenges structuralism’s binary abstractions – such as nature−culture, emotion−reason, space−time, nonhuman−human – that effectively reproduce past hierarchies of difference. In the place of binaries, poststructuralists offer a broad variety of perspectives and approaches for exploring difference both critically and productively. These include deconstruction and decentering, interrogations of discourse and power, and differential ontologies.

The history of geography’s encounter with poststructuralism is nonlinear. The discipline discovered these thinkers relatively late compared to other critically minded departments within the English-speaking academy, partly due to the nature of the publishing industry, which published translations (usually from French) out of historical order and in clusters that focused on one thinker while ignoring many of her or his contemporaries, creating a scene that was ripe for confusion and misinterpretation in early encounters with poststructuralism.

Geography’s engagement with poststructuralist thought began in the mid-1980s. Today, its influence can be seen in every area of “human geography.” Initially, poststructuralist geographies were fixated on critiquing representation, deconstructing disciplinary and ideological boundaries, and reading space as a complex plurality. Geographic interest turned to the porousness of identity categories and their performance and the agency of nonhuman objects. Later, geographers broached problems of nonrepresentation and affect. Today, poststructuralism and poststructuralist geographies are the subjects of numerous debates that have exposed certain theoretical limitations. It remains to be seen if these interventions will constitute a break from or an extension of poststructuralism.

Intellectual precursors

If poststructuralism is one of the most influential intellectual movements of the twentieth century, it is also one of the most complex. This complexity is lessened when poststructuralists are considered in the broader context of their precursors. As the name suggests, poststructuralism extends and responds to structuralism. Indeed, many celebrated poststructuralists initially contributed to the French structuralist movement. At the same time, structuralists and poststructuralists were influenced by earlier intellectual tendencies in French thought – most notably humanism, phenomenology, existentialism, and
Marxism. The following section offers a brief survey of these contexts before moving on to the turn to poststructuralism.

The three H’s

Starting in the 1930s, French intellectuals experienced renewed interest in German thought, which had been out of vogue since the Franco-Prussian War (1870–1871). The publication of Martin Heidegger’s *Being and Time* in 1927 attracted French students to Germany to study phenomenology – the study of human experience and consciousness – under Edmund Husserl and Heidegger. Neither had yet appeared in French translation. Instead, their thought filtered into France through the work of their French students. Most notable among these were Emmanuel Lévinas (who merged phenomenology with an ethical obligation to the Other), Jean-Paul Sartre (who explored it in the context of existentialism and free will), and Maurice Merleau-Ponty (whose phenomenology of bodily perception challenged the traditional separation of mind and body).

During roughly the same period, French intellectuals developed an interest in Hegel. Between 1933 and 1939, Marxist émigré Alexandre Kojéve delivered famous lectures on Hegel that were attended by Lévinas, Sartre, and Merleau-Ponty as well as Jacques Lacan and other future structuralist thinkers. Kojéve interpreted Hegel’s treatment of history anthropologically, correlating the “end of history” with the advent of the struggle between the United States and the Soviet Union. Kojéve’s thought influenced the official ideology of the French Communist Party – a mixture of Marxism and nationalism – which eagerly embraced a Hegelian interpretation of Marx. French Hegelianism hit its apex just after World War II. During this period, Jean Hyppolite (under whom poststructuralists Michel Foucault, Gilles Deleuze, and Jacques Derrida studied) published the first French translation of Hegel’s *Phenomenology of Spirit* (1939) along with key commentaries such as *Genesis and Structure in Hegel’s Phenomenology* (1946) that anticipated the structuralist movement. Hyppolite’s contributions cemented Hegel’s importance in France. Thus, by mid-century, French philosophy’s universalistic and human-centric obsessions with the “three H’s” – Hegel, Husserl, and Heidegger – had set the scene against which the poststructuralists would develop many of their best-known interventions.

Structuralism

The French structuralist movement broke with phenomenology and Hegelian humanism in favor of “scientific” approaches. During the early 1950s, anthropologist Claude Lévi-Strauss found an ideal figurehead in the early twentieth century semiotician Ferdinand de Saussure, whose “structural linguistics” employed “a scientific method rather than a philosophy or speculative thinking” to map the structure of signs and their creation of meaning (Dosse 1997, 115). Saussure’s structuralism distinguished abstract rules of language *système* (la langue) from individual speech *actes* (la parole). Studying the infinite variability of speech, he argued, reveals a finite, representable system – linguistic structure – capable of illuminating similar elements across multiple languages. Saussure divided the smallest unit of semiotic meaning – the “sign” – into two different elements: a “signifier” (for example, the word “cat”) and its “signified” (the idea that corresponds to that word). Saussure insisted that meaning is not a matter of a necessary relation between signifier and signified: the idea “cat” might just as easily have attached to the signifiers “schlemihl” or...
“schlemazl.” Rather, meaning emerges through the interaction of a series of signifiers. Structural linguistics studies meaning created through the play of signifiers. This dual concern with signs and structural abstraction became a definitive element in French structuralism.

If modernity had long obsessed over transcendent human “truths,” Saussure’s system offered many antihumanist escape routes for the humanities and social sciences. At the vanguard, Lévi-Strauss employed Saussure’s semiotic break to analyze different cultural practices and beliefs in terms of broader overarching systems. Like Saussure, he argued that broader transcultural structures rested behind apparent cultural differences. His *The Raw and the Cooked*, for example, describes disparate communities in terms of similar binary structural relations (nature—culture, feminine—masculine, and so on).

In psychoanalysis, Jacques Lacan repurposed the signifier (detached from its signified) as the germinal element of the “symbolic order” in the subconscious. He replaced Freud’s system of social taboos and repressed desires (for example, the Oedipal system of incestuous desire and the fear of castration) with a structure of signs (or exchanges of signs between subjects: for example, the Father’s “No”) that proceed one after another. The resulting “signifying chain,” in which signifiers are attached to each other and detached from their signifieds, is constitutive of the subconscious.

Finally, in the realm of Marxist political economy, Louis Althusser argued against the statist Hegelian argument that history is the manifestation of human reason. Taking a cue from Friedrich Engels, Althusser instead argued that in “the famous last instance” social relations are fundamentally determined by the economy. The state, he countered, was merely a structural effect of repressive and ideological “apparatuses.” These circulate through the practices of state agents such as police, bureaucrats, and professors. Althusser felt that autonomous subjectivity was a humanist specter. Rather, individuals are interpellated or “hailed” by ideological apparatuses (via police, school teachers) that *subjectivize* them: enlisting the individual to identify with a structural relation (e.g., to the state). Thus, too, structural class relations prevail over individual identities.

**The birth of poststructuralism: Johns Hopkins University, 1966**

Poststructuralism was born in Baltimore, Maryland, during autumn 1966. There, the Humanities Center at Johns Hopkins University hosted a conference on “The Languages of Criticism and the Sciences of Man: The Structural Controversy” that included several structuralist luminaries, including Roland Barthes and Lacan. Among them, a young philosopher named Jacques Derrida presented a talk entitled “Structure, Sign and Play in the Discourse of the Human Sciences” that took aim at Lévi-Strauss and structuralism’s sacred object: the sign. Derrida questioned structuralism’s tendency to maintain a “center” to which its structures were anchored (culture, ideology) and its conservative view on the play of signifiers within those structures. Both, he argued, are symptomatic of Western thought’s dependence upon a “metaphysics of presence”; that is, its insistence upon centering concepts (Being, existence, God, the sign) that stabilize a system but are invisible to it. Derrida appealed to a structuralism to come that would fully embrace the *play* of signs and discourses and their proper mode of analysis.

American academics heard Derrida’s call as a continuation of the structuralist project and the announcement of something entirely new.
For this, they invented the term poststructuralism. “What Americans call post-structuralism,” as Dosse explains, “existed even before the structuralist paradigm waned” (Dosse 1997, 17). Indeed, Althusser’s structuralist Marxism, for example, was not widely popular until after the worker-student revolts of May 1968. Thus, even the birth of poststructuralism (an appellation that its supposed members widely rejected) was marked by confusion.

Deconstruction and différence

Jacques Derrida (1930–2004) was born and raised in French colonial Algeria. The child of Jewish parents, he was expelled from lycée (French high school) due to the Vichy government’s anti-Semitic quotas. This seemingly paradoxical position—simultaneously one of privilege and one of marginality—anticipated a main concern of his work. Derrida’s philosophical project rejects the privileging of speech over writing (or logo centrism). Western thought has long preferred presence and existence to absence and inexistence. Following this tendency, structuralists favored speech (the present, the spoken, the witnessed) over writing (the removed, the disembodied, the recorded). Indeed, Saussure viewed writing as merely derivative or imitative of speech. Derrida’s famous invention “deconstruction” enabled him to disassemble and reverse (de-construct) uneven binary constructions.

For Derrida, writing does not follow from or represent speech. Indeed, he insists that, because it amplifies the polysemy of signs, writing is the more limber and playful of the modes of signification. This has to do with the way that we encounter writing (and it is here that Derrida merges his training in phenomenology and structuralism). Consider Derrida’s famous neologism différence, which plays on the French term différence, meaning both “to differ” and “to defer.” Différence introduces a “graphic intervention” (Derrida 1982, 3)—replacing the “e” with an “a”—that irreversibly entangles difference’s temporalization (to defer) with its spatialization (to differ). This space-time describes the way readers experience the emergence of meaning in writing. Words accrue meaning through their relation to other surrounding words (what Derrida calls “spacing”). But this means that, as we read, there are always “traces” of past words and more words to come that inform the meaning of any given word. The future and the past impinge on the present, multiplying rather than narrowing meaning. Thus, Derrida uses writing to deconstruct the binary of passive space and active time, rendering both dynamic in the process.

By the time Derridian poststructuralism arrived in British and American geography in the mid-1980s, it had already exerted a deep influence across the humanities. Geography was still in the thrall of debates over Marxist structuralism and humanist agency, and deconstruction was met with widespread confusion, suspicion, and accusations of relativism. Urban geographers such as Michael Dear cautiously introduced Derrida’s thought into journals, such as Society and Space, sometimes crowding many disparate thinkers under the umbrella of deconstruction. By the late 1980s, heated debates were flying between established Marxist geographers determined to organize the discipline around a common struggle and the poststructurally curious who are eager to experiment with critique. In many ways, these positions echoed the debates in 1970s France between the Maoists and the poststructuralists.

Geography’s first landmark poststructuralist intervention was cartographer Brian Harley’s “Deconstructing the Map” (1989). Blending
Derrida with Michel Foucault (see below), Harley interrogated cartography’s history of symbolic representation and its complicity in reproducing Western meaning systems. A decade later, deconstruction was sufficiently entrenched for Deborah Dixon and John Paul Jones III to broaden its focus to the intra-divisions that had arisen around differences in geography’s methodological ideologies. Written as a conversation between a Derridian and a spatial scientist, Dixon and Jones’s “My dinner with Derrida, or Spatial Analysis and Derrida Do Lunch” (1989) is poststructuralist geography par excellence. Through the use of literary dialogue, they offer glimpses of the interpretive gulfs between dominant discourses – a reminder that reading and writing are always confronted with the prospect of reconstructing or deconstructing structures of domination.

Drawing upon an old critique of Freud’s masculine bias (or, “phallocentrism”), Derrida invented the term “phallogocentrism” in the early 1970s to challenge the predominance of masculine perspectives in writing. Feminist thinkers had also deployed nuanced critiques against phallocentrism in Western representations and social institutions. During the 1970s, several poststructuralist feminists – such as Hélène Cixous, Julia Kristeva, and Luce Irigaray – merged these currents into the écriture féminine (“women’s writing”) movement, which sought to overturn phallogocentrism by writing bodily gendered difference into their texts. Within two decades, American feminist theorists such as Judith Butler and Eve Kosofsky Sedgwick would view gender and sexuality as something that was performed (rather than innate), leading them to consider the body itself to be a kind of text.

Such maneuvers were crucially important for cracking open the latent sexism in much of geography – even in the radical currents – in the 1970s, 1980s, and 1990s. An early chapter of David Harvey’s The Condition of Postmodernity (1989), for example, illustrates the cultural shift between modernity and postmodernity through a series of representative portraits and advertisements, each of which featured nude women. In an article entitled “Boy’s Town,” Rosalyn Deutsche critiqued Harvey’s use of the feminine body in a book that seemed otherwise oblivious to questions of gender or feminist politics. Harvey responded by adding a brief note to subsequent editions of Condition of Postmodernity that largely dismissed Deutsche’s critique as a product of postmodern feminism. His seeming blindness to his own phallogocentrism throughout the exchange was characteristic of the debates of the period.

Poststructuralist feminism insisted that, like texts, spaces are multiple and hybrid. Spatial experience is gendered. Thus, what might be a scene of abandon or enjoyment for men, can be a site of danger and terror for women. Gendered power is defined in part by the ability of the powerful to remain oblivious to spatialized differences. In geography, this myopia also surfaces when people of privilege (the discipline is disproportionately white and male) insist that their own spatial experience is universal. Antisexist geographers have responded by deconstructing geographic masculinism and, in the spirit of écriture féminine, creating new approaches to representing and narrating gendered spaces. Similarly, critical race theorists such as bell hooks, Gloria Anzaldua, and Katherine McKittrick speak back to white normativity by detailing marginalized raced spatialities. Finally, geographers J.K. Gibson-Graham (a portmanteau name for co-authors Kathy Gibson and Julie Graham) brought the attack on “centers” to bear upon political economy. Their critique of “capitalocentrism” blended feminist critique with an emerging interest in alternative economies.
Why, they asked, should political economy fixate on one mode of exchange – capitalism – to the exclusion of all others (such as cooperatives, gift economies)?

In the late 1990s, geographers unearthed the deconstructive kernel in postcolonial studies to critique Western constructions of nature. Bruce Braun’s “Buried Epistemologies” modeled this approach by examining wilderness preservation strategies by white Canadians that erased the indigenous communities living in and using the contested “natural” spaces. A decade later, Jake Kosek’s _Understories_ (2006) called for “new natures” after deconstructing the cultural politics surrounding dynamic forest spaces near Los Alamos National Laboratory in New Mexico. Kosek’s “understories” (a forestry term for leaf litter resting on the forest floor) are the many material and symbolic traces (names, mysteries, uses, and even radioactive isotopes) that expose the porous boundaries between nature, culture, and politics. We remain blind to these as long as we construct insurmountable differences between nature and culture. Other cultural geographers, such as Rich Schein and John Wylie, have recently deconstructed the “material semiotics” of vernacular landscapes. Schein, for example, critically dismantles the construction of a park built to celebrate horseracing in Lexington, Kentucky, that also erases from view a neglected African American enclave sitting just north of downtown. Meanwhile, Wylie reminds us that memorials make absence present in ways that ultimately exceed our capacity for representation.

**Discourse, power, and the body**

Michel Foucault (1926–1984) studied philosophy at the École Normale Supérieure under Hyppolite the Hegelian and Althusser the structuralist Marxist. In his early works, which many consider to be structuralist, he examined the underlying assumptions (or, _épistémé_) that enabled historic shifts in scientific knowledge during modernity. He argues that disciplinary and institutional knowledge (medicine, psychiatry) is a mode of power capable of inventing human subjectivities (the infirm, the abnormal) by introducing new categories and differences into social life.

Foucault broke with structuralism after 1968. His attention shifted to the complex relationship between power and “discourse,” a set of “utterances” – statements – that resonate across intellectual and institutional fields. To this end, he invented the “archeology of knowledge,” an archival research practice that seeks out the conditions of possibility (rather than the origins) of discourses. These tend to emerge across tangles of documents, appearing “sometimes as the general domain of all statements, sometimes as an individualizable group of statements, and sometimes as a regulated practice that accounts for a certain number of statements” (Foucault, quoted in Dosse 1997, 241). Rather than inscribing order upon this jumble, Foucault employed a “genealogical” reading strategy of “coupling together … scholarly erudition and local memories” (Foucault 2003, 8). This approach rendered a view of disciplinary power that, rather than being concentrated in a single point, was dispersed across various banal sites and statements.

Foucault and Derrida appeared on the Anglo American geographic scene during the same period, leading these very different thinkers to be sometimes confused with one another. Today, Foucault’s influence can be found in much of the discipline’s critical language. Like Derrida, his uptake by geographers was gradual. The earliest engagements in the mid-1980s focused on the “power/knowledge” relationship and its
implications for governmentality (the management of populations). This analysis also rounded out Harley’s deconstruction of mapping, which interpreted cartography’s semiotic complicity in state power. Other geographers, such as Chris Philo, took a page from Foucault’s histories of modern institutions by analyzing the institutional spaces of madness. Later, the genealogical strategy informed explorations of power, “expertise,” and the complex realities they manipulate. For example, Paul Robbins merged poststructuralism with Earth processes to track the paper trail of experts recommending invasive tree planting in deforested areas of Rajasthan, India, during the 1990s. His turn to the discourse of expertise helped initiate a new critical focus in political ecology on governmental structures’ imaginaries of ecosystems. Most recently, Stuart Elden’s *The Birth of Territory* (2013) employs genealogy on a grand scale to trace the shifts in the territory concept from antiquity to early modernity.

Poststructuralist geographers also celebrate Foucault for his nuanced engagements with metaphoric and material spaces, something he famously discussed with the editors of the French political geography journal *Hérodote* in 1976. Spatial metaphors figure prominently in his archeology and genealogy. Both reject linear historic perspectives in favor of spatial relations, such as the “proximity” of certain discourses within a tangled field of statements and documents. Foucault’s fascination with material spaces is evident from the opening lines of his *History of Madness* (1961), which invokes the “large, barren uninhabitable areas” left by recently vanquished leprosy “at the edges of the community, at town gates … where the disease no longer reigned but its ghost still hovered” (Foucault 2006, 3). By the 1970s, Foucault understood space as a dynamic power mechanism. In *Discipline and Punish* (1975), he famously analyzed British philosopher Jeremy Bentham’s “panopticon,” a circular prison surrounding a central surveillance tower. Believing themselves to be constantly under observation, prisoners would internalize self-discipline. Foucault described such spatial structures as power “diagrams,” arguing that the panopticon signals a particular disciplinary shift from central sovereignty to modern forms of dispersed self-government. His genealogy discovered the same “dispositif” (institutional power mechanism) in documents concerning army training, factory work, and schoolhouse discipline. All frame the body as a site for exercising not the power to oppress subjects, but the power to produce them: brave soldiers, fast workers, strong students. These “biopolitical” relations enroll components of human biology as elements of disciplinary “power strategies.”

Foucault’s conceptualizations of biopower and governmentality have influenced several approaches in critical geography. During the 1990s, some geographers explored the possibility for liberation from power relations via what he called “heterotopias,” complex sites that move out of phase with repressive structures. Meanwhile, studies of the government of bodies – particularly women’s bodies and reproductive labor – have gravitated toward the critical discourses of biopolitics. In the years since the start of the United States’ interminable “Global War on Terror,” Foucault figured prominently in critical geographies of surveillance, governmentality, and “necro-politics” (the power to determine who lives or dies). Important to these discussions are David Murakami Wood’s analyses of the globalization of surveillance and Ben Anderson’s work on emergency preparedness, which shifts biopolitical governmentality to the regime of anticipatory reflex. Derek Gregory’s study of Western geographical imaginaries of the Middle East, which help constitute the “colonial present,” shares much in common with Foucault’s lectures on governmentality, colonialism,
and race. Gregory is currently expanding this analysis through to the necropolitics of the US drone program in Iraq and Afghanistan.

**Difference, becoming, and ontology**

The French philosopher Gilles Deleuze (1925–1995) is best known for his collaborations with Félix Guattari (1930–1992), a trained philosopher and experimental psychotherapist. Together, they developed several important interventions in contemporary social theory that challenged Lacanian-Althusserian structuralism. Overturning psychoanalysis’ phallocentric “Oedipal” organization of sexual desire, their *Anti-Oedipus* (1972) invents a political psychology based on flows of “machinic desire” (bodily affections). According to Freudian psychoanalysis, personal desires are structured by innate drives that, when constrained by social taboos, fill the unconscious with anxiety. Deleuze and Guattari countered that unconscious desire is not pre-directed but rather operates upon a body organized by diverse social structures. A line of resistance seems to manifest in the “body without organs,” a body that shrugs off social dictates regarding the normal uses of its organs (e.g., the equation of sexuality exclusively with the genitals). Deleuze and Guattari saw a parallel between psychoanalysis and political ideologies that are prescriptive about what “counts” as legitimate objects of political struggle (a characteristic of the statist Marxism of the 1960s). Against this depiction, they introduced the notion of “micro-politics,” diverse, localized politics that emerge in minor ways from a wide range of social relations. These contributions have since been deeply influential for theorizing sexual liberation and other marginalized political movements.

Deleuze and Guattari later extended their interest in minor structures to systems as diverse as material processes, linguistics, social history, and perception. *A Thousand Plateaus* (1987) addresses these many fields and intertwines them with several others. Each system involves a dynamic relationship between “major” (dominant, normative, central) and “minor” (rebellious, abnormal, marginal) elements. Interacting major and minor elements transform one another in a process Deleuze and Guattari call “becoming”: the major enters a process of *becoming-minor* while the minor is enrolled in *becoming-major*. They become, in other words, decentered, emergent structures. In geography, Cindi Katz’s “Toward Minor Theory” (1996) uses minor theory to “scratch at major theory” (1996, 487) such as Marxism that continues to dominate and structure disciplinary debates on politics and relevance. Katz also interrogates the potential traps to discourses about becoming-minor, most notably the notion of “becoming-woman.” In this, she echoes many debates about the dangers of fetishizing the “minor” in social theory and social movements.

Deleuze and Guattari’s account of the minor is sometimes misread as a moral imperative, despite their warnings against this. For example, during the mid- to late 1990s, Nigel Thrift’s early nonrepresentational and affect theories tended to frame becoming-minor as a fundamentally liberatory activity that is easily accessed through activities such as dancing. Others objected that dance is a thoroughly representational activity. More to the point, any dance involves a variety of interacting major and minor elements. More recently, many affect theorists have been drawn to Deleuze’s use of seventeenth-century Dutch philosopher Baruch Spinoza to explore the dynamics of bodily capacities to affect and be affected by other bodies. These studies have inspired reconsiderations of working life, care labor, and human–nonhuman relations.
Because the elements of becoming are mutually transformed through interaction, Deleuze refers to such relations as “differential.” This is the fundamental component of Deleuze’s ontology (the study of what exists), which he defines as “pure difference.” Ontology speculates about the fundamental building blocks (or metaphysics) of the world. In the West, these have often been defined as things (or identities) such as “Being.” By defining ontology in terms of a relation rather than a thing, difference rather than identity, Deleuze created a picture of the world that emphasizes contingency and change rather than normativity and order. In the early 2000s, this emphasis would influence what is sometimes called the “ontological turn” in geographic theory. In their critique of the scale concept in geography Sallie Marston, John Paul Jones III, and Keith Woodward proposed that structural scalar concepts be replaced with a suppler notion of “site ontology.” They defined sites differentially, as the convergence of any number of disparate elements. Geographers do not prefigure the composition or character of such elements but rather describe them “immanently” from within the given spatial relations. Other geographers have been drawn to Deleuze and Guattari’s notion of “assemblages” (which, like Marston et al.’s sites, consist of disparate elements relating immanently) to explore the dynamics of urban and socio-natural spaces. Most recently, Deleuzian ontology has inspired an exploration of “new materialisms” that ask critical questions about the nature of material reality and seek to redefine its metaphysics and ethics.

The geography to come

Translations of French critiques of poststructuralism (some dating back to the 1970s) have appeared with greater regularity during the past decade. One longstanding critic is Alain Badiou (1937–), a former Maoist and student of Althusser, whose philosophy of events argues that mathematics is ontology. Although very critical of Derrida and Deleuze throughout their lives, his recent work has incorporated elements of their thought. Some key Marxist geographers involved in the 1990s debates on poststructuralist geography have begun to embrace fragments of Badiou’s earlier thought, along with that of fellow Althusser student, Jacques Rancière (1940–). Some geographers formerly associated with nonrepresentational theory have also embraced these thinkers. Currently, Badiou occupies a position in geography similar to that of Derrida and Foucault during the mid-1980s: the discipline does not yet seem to know where he fits.

François Laruelle’s (1937–) “non-philosophy” claims that several poststructuralists remain firmly situated within the “prior decisions” (presuppositions) of Western philosophy despite their decentering critiques. Nonphilosophy is a “science of philosophy” that Laruelle claims can interrogate such prior decisions from a nonphilosophical viewpoint outside the discipline. While geographers have yet to engage him, nonphilosophy presents a potentially interesting tool for interrogating the “prior decisions” of geographic social theory. Badiou and Laruelle lean heavily on science (mathematics and nonphilosophy), but it is not clear if this reflex heralds a return to structuralism of one form or another.

Quentin Meillassoux (1967–), a student of Alain Badiou, has introduced speculative questions about what it means to think what cannot be thought. This is an important question for nonrepresentational theorists; however, these thinkers have characteristically anchored their analysis to problems of the body, such as affect. Meillassoux’s project is much larger. He critiques the “correlationism” (the tendency to
define “truth” as a correlation between objects and thoughts) that has dominated Western thought since Kant. What are the implications of correlationism, Meillassoux asks, for something that preceded human existence and thus could not be correlated with thought? He calls this “something” the *arche-fossil* and uses it to develop an anticorrelationist philosophy grounded in contingency. Given geography’s reflexive empiricism and its tendency to demand that theory be met with practice, it seems unlikely that the critique of correlation can find much fertile ground in the discipline. As we move into the so-called Anthropocene, however, the Earth’s future state – which may or may not sustain human life – does suggest future states of the world that are essentially unthinkable. In this sense, Meillassoux might offer a fertile, antihumanist perspective to help frame this discussion.

**SEE ALSO:** Affect; Biopolitics; Biopower; Borders, boundaries, and borderlands; Carceral geographies; Care work; Critical geography; Cultural turn; Difference; Discourse; Feminist geography; Gender; Geophilosophy; Identity; Intersectionality; Nonrepresentational theory; Ontology: theoretical perspectives; Orientalism/Occidentalism; Positionality; Postcolonial geographies; Posthumanism; Power; Queer geographies; Race and racism; Representation; Sexualities; Social constructionism; Subjectivity

**References**


**Further reading**


Poverty

Lakshman Yapa
Pennsylvania State University, USA

Although poverty in the urban slums of South Asia, Africa, and Latin America is materially far more intense than poverty in any US inner city, there is no need for two different theories of poverty – one for poor countries and another for rich countries – as a single theory suffices. The United States is the greatest wealth-producing engine ever conceived in human history, but, nevertheless, the persistence of poverty amid great wealth remains a continuing paradox (Glasmeier 2006, 1). A discussion of poverty in the United States should therefore prove instructive for understanding poverty elsewhere in the world.

This entry contrasts an exchange-value approach to poverty with a use-value approach. Economics has always been central to the study of poverty, prioritizing exchange-value to increase income, employment, and economic growth. In contrast, the use-value approach looks at details of basic needs in complex substantive networks of food, health, energy, and so on, instead of subsuming everything under a single “income” rubric. The two views lead to very different outcomes and to the conclusion that poverty is a normal manifestation of an economy that has privileged exchange-value over basic use-values.

Persistence of poverty

Poverty has persisted, despite 60 years of initiatives from the World Bank and the United Nations and 50 years of the “War on Poverty” in the United States. Poverty status in the United States is determined through income thresholds that vary by family size. If a family’s total income is less than the family’s threshold, then that family and every individual in it is considered poor (US Census Bureau 2013). When official poverty calculations began in 1959, the poverty rate was 22%. The rate steadily declined through the 1960s, but since then has remained steady between 12% and 15%, despite trillions of dollars being spent on social welfare.

The World Bank defines extreme poverty as earnings of less than US$1.25 a day at 2005 prices. In 2008, 1.29 billion people lived on less than $1.25 a day, compared to 1.94 billion in 1981. These numbers represent a steady decline in poverty rates from 52% in 1981 to 22% in 2008. However, almost all of this poverty reduction came from China: between 1981 and 2008, the proportion of Chinese people living on less than $1.25 a day is said to have fallen from 85% to 15.9%; that is, by roughly 600 million people. But China achieved this poverty reduction with an unsustainable rate of economic growth of about 10% per year, which has come at an enormous cost to the environment. If the poverty line were increased to $2 a day, nearly 42% of the world’s population would be
POVERTY

deemed poor, and the World Bank itself admits that rates of poverty reduction at this threshold are not impressive (Chen and Ravillion 2010). The World Bank’s statistics on poverty reduction mask the stark reality that nearly 1 billion people live in slums, the fastest-growing human settlements in the world, in the megacities of poor countries (UN–Habitat 2003).

Why poverty persists

Conventional explanations of poverty such as lack of resources, lack of capital, and lack of economic growth are not helpful. Defining poverty as the lack of income privileges exchange-value over use-value. Modern economies systematically fail to produce the basic use-values needed by households. In fact, this can be used as an alternative definition of poverty: people are poor when the larger economy fails to produce basic use-values. There are four major reasons why poverty cannot be eradicated within an exchange-value framework: (i) failure to provide full employment; (ii) nonpayment of living wages; (iii) social construction of scarcity; and (iv) undertheorizing of use-value. Lack of full employment and low wages are well known as causes of poverty, but less recognized is the idea that these conditions are functional to the expansion of exchange-value. Another mechanism for expanding exchange-value is creating demand for high value-added goods through socially constructed scarcity. Finally, the exchange-value circuit is so hegemonic that it has created a condition that can be called “econo-normativity,” where it is almost impossible to visualize the production of goods outside that framework.

Nonpayment of living wages

Apart from unemployment, poverty also results from low wages, hence the presence of the working poor in the United States. According to the living wage calculator developed at MIT by Glasmeier (http://livingwage.mit.edu/), in every US state the poverty wage per hour is far less than what was calculated to be a living wage. The tertiary sector of the economy, which includes retail and food services, has been the principal job provider for the US economy since the 1960s, but this is also a sector where a disproportionate number in the workforce make very little money. Apart from the classic downward pressures on wages that Marx spoke of, neoliberal globalization has opened up whole new areas for the expansion of exchange-value. While globalization has broken down all national barriers to the movement of
capital, there is a veritable apartheid restricting the movement of labor from poor countries to rich countries. Whenever wage levels rise in a poor country global multinationals shift to other low-wage countries. The desperation of economic migrants from Central America to the United States, from North Africa to Europe, and from South and Southeast Asia to the Middle East is stark evidence that the model of development is not working for the poor.

Socially constructed scarcity

The claim that socially constructed scarcity is the leading cause of poverty needs a lengthy explanation as it is not a commonly held view (Yapa 1996, 2015). Resources such as arable land, oil, and food are indeed finite but, contrary to what economists claim, this is not the origin of scarcity, which is in fact socially constructed. Here is how it happens: the exchange-value of a commodity is usually a single definite amount of money, but the same commodity can have many use-values. For example, a meal at a given price can satisfy hunger, provide nutrients, express hospitality, or even confer social status. Most commodities have a range of use-values, with one or two being primary for the consumer. The primary uses can be described as “end use-values” – what we use the commodity for in the end. Once the primary end use-value of a commodity, say A, is specified, it can be seen that there are other ways of meeting that end use-value without consuming the commodity A. This is important to understand in order to see that scarcity is socially constructed. One way to increase the demand for commodity A is to remove all the other ways in which the end use of A can be met. The history of exchange-value is actually a history of how alternatives ways of meeting end uses have been eliminated, de-developed, or made unavailable. Many students of capitalism have observed that many needs appearing in a capitalist society are intentionally created to realize profits; that is, they are the needs of capital, not of people (Illich 1978).

The examples of socially constructed scarcity are legion in every sector of the economy throughout the history of the world. A most egregious example of this pattern is the increased demand for automobiles in the United States accompanying the elimination or de-development of alternative modes of transport such as public transit, carpooling, bicycling, and walking. Other examples include the increased demand created for infant formula caused by discouraging breast-feeding, the substitution of expensive meat protein for less expensive vegetable protein, the medicalization of health by discouraging preventive practices, and narrowing housing options by making the single family suburban home the most desirable; examples of socially constructed scarcity are everywhere.

The underlying mechanism for the social construction of scarcity can be represented through a simple demand and supply diagram. In economics scarcity is said to exist when the demand for a commodity exceeds its supply. In Figure 1a, when the demand curve moves to the right from D1 to D2, more of a commodity is demanded at each level of price, which increases the scarcity of that commodity. Scarcity is also increased when the supply curve of a commodity moves to the left from S1 to S2 as shown in Figure 1b. This is the case with the chronic underproduction of basic goods which, although essential to life, offer limited opportunity for extending commodity chains, adding value, and increasing profit, a situation clearly demonstrated by the nonavailability of fresh produce in contrast to the abundance of processed food in what are called “food deserts” in US inner cities.
Modern economies systematically *underproduce* basic use-values such as food, health, housing, energy, transport, and education. There are simply not enough of these goods available at affordable prices, yet there is simultaneously an overproduction of high-profit, high value-added goods. We can say, then, that the creation of the scarcity of basic use-values at affordable prices is a fundamental cause of poverty.

**Undertheorizing of use-value**

The exchange-value view of poverty is the dominant conventional view of poverty, and it enjoys support across a wide spectrum of political philosophy. At one end is the conservative view that poverty is a result of individual failure to adapt to unfettered markets. Liberals view household poverty as reflecting structural inequity, caused by class, race, and gender, that limits opportunity for social mobility, a matter that could be corrected through government regulation of the market (Krugman 2009). Inspired by the writings of Marx in *Capital* (1990/1867), geographers such as Harvey (2010) believe that poverty is an inevitable manifestation of capitalism’s inherent tendencies toward uneven development. Through this entire range of poverty philosophies, from individual failings to inherent tendencies of capitalism, one concept remains constant — people are poor because they have no money. This represents a classic instance of ideological hegemony: the power of ruling classes to convince not only the general public but also their formidable critics that poor people lack access to such things as food, health, and housing because of their marginal status in the formal economy. The net effect of the dominance of exchange-value in poverty knowledge is the complete absence of a discourse on use-values.

The historical movement of the household from the circuit of use-value to that of exchange-value, beginning with colonialism and continuing today with increased globalization, has had enormous consequences for the poor. First, the poor have been corralled into the formal economy, knowing that the

Figure 1  Shifts in demand and supply, end use-value, and constructed scarcity.
prospects of their full participation are dismal. Second, privileging gross national product (GNP) growth rates has prevented the critical examination of economic performance with respect to the production of affordable basic goods. Third, metanarratives of income poverty have disempowered the poor: a vast discourse has been generated that has invoked a raft of explanatory variables of poverty – lack of free markets, exploitation of labor, racial and gender discrimination, overpopulation, lack of natural resources, lack of investment capital, lack of education, the persistence of a culture of poverty, welfare dependence, and capitalist exploitation. By invoking these factors as causes of poverty, academics have created metanarratives at a discursive scale that are not correlated to the power of poor people, and that thereby rob them of agency. The poor have been acted upon through our agency – of corporate executives, development planners, state bureaucrats, and academics – because our theories don’t allow us to see the poor as potential agents of change.

A poverty discourse on use-values will be different from that of exchange-value because it will focus directly on subject-specific topics related to food, good nutrition, health of the body, and so on. The key claim is this: A large body of literature already exists that shows that resources such as land, labor, and capital are not limiting factors in the attainment of basic needs for the poor. A turn to a theory of the production of basic values is a good example of what Foucault (1990, 92–114) meant by “non-sovereign power”: power is the ability to solve our problems, and it does not always come from a single sovereign source, the state, or from above; power is resident in the multiplicity of force relations immanent in our daily lives. Knowledge of use-values exists to help ordinary people exercise power and agency at specific sites in specific fields at scales that are correlated to their power. The overwhelming presence of exchange-value in poverty knowledge has simultaneously led to the absence of its other: namely, use-value.

Use-value and the basic needs economy

The concept of use-value offers a very useful framework for thinking about solutions to poverty; however, as Marx himself noted, the concept of use-value does not lend itself to general theory. While the contradiction between use-value and exchange-value embodied in commodities is the source of all major contradictions of capitalist production, Marx believed that the proper object of study for political economy is the laws governing exchange-value (Marx 1990/1867, 138–154). It is not possible to make general claims about use-values as they are by necessity tied to specific physical, chemical, and locational properties of resources. A discussion of use-values involves subject-specific substantive literatures, so the claim that land, labor, and capital are not limiting factors for the production of basic use-values has to move forward through description of specific cases. The selected example here is food and agriculture.

An average person needs about 2000 calories of food energy, 50 g of protein, and some vitamins and trace elements in his or her diet each day. A perfectly healthy diet can be had from unprocessed whole grains, beans, vegetables, fruits, nuts, and modest amounts of fish and meat. The world as a whole produces enough calories for all human beings, but many do not have enough money to bid for foods that are also used as inputs for animals and automobiles. Since large numbers of people are, and will continue to be, marginalized by the formal economy, it
POVERTY

makes sense to develop an economy that focuses directly on basic use-values such as food.

Nearly 870 million of the 7.1 billion people in the world were suffering from undernourishment in 2010–2012 (FAO 2012). Almost all the hungry people live in developing countries, in areas that are tropical or subtropical. Plants that fix solar energy by means of photosynthesis are the sole source of food for humans and animals, and it is ironic that most of the hungry people in the world live in tropical and subtropical biomes that have the highest photosynthesis, and therefore the highest potential for producing food. The food and agriculture sectors of the United States economy generated about $1.3 trillion in 2010 (USDA 2014) and it is also one of the largest exporters of agricultural commodities. Nevertheless, the number of food-insecure individuals in the United States in 2010 was 48.8 million, 16.1% of the total population (Coleman-Jensen, Nord, and Singh 2013).

Mechanisms for the social construction of scarcity are at work in creating scarcity of nutritious food at affordable prices. For example, the average US diet is mostly made up of processed food, meat, and dairy with high levels of salt, sugar, and fat (Pollan 2006). Many observers believe the US food system promotes exchange-value by implicating food in long commodity chains between farm and household, a situation that can be described through the model of socially constructed scarcity presented earlier. Broadly speaking, nutrition can be obtained from two classes of food. Let us call whole foods and fresh produce class A foods, and processed foods with long commodity chains and value added at each step class B foods. For example, the profit margin for apple juice or a packaged apple tart (class B) is much greater than that of a fresh apple (class A). When supermarkets leave poor neighborhoods, they leave behind convenience stores selling class B goods. Thus the demand curve for processed food moves to the right (class B) and the supply curve of fresh produce (class A) moves to the left, aggravating food scarcity for poor people in the city (Figure 1).

Or consider sources of protein for human nutrition. For decades the idea that animal protein was superior to vegetable protein was official nutrition doctrine in the United States, but nutrition science now states that vegetables and beans can fulfill all our protein requirements, and, contrary to public perception, are not inferior sources of protein. The idea that vegetables are an incomplete or inferior protein removed vegetables as a source of protein supply, and thus increased the demand for meat. It is one thing to consume meat as a matter of taste and food preference, but to represent it in official government documents as an essential food group is an example of socially constructed scarcity (Pollan 2006). The mass production of animal protein creates additional scarcity, from the inputs needed to produce meat (Pimentel and Pimentel 2008, 67–75), and from high meat consumption, which is linked to poor health (Campbell and Campbell 2005).

Despite the “food deserts,” nutritious food can be produced in US cities at affordable prices because land, labor, and capital are not limiting factors. Methods of biointensive farming show that large quantities of food can be grown in very small spaces with few external inputs and little capital investment. The science is both new and old. Jeavons (2012) and many others have demonstrated that, compared to conventional agriculture, biointensive mini-farming can produce far more nutritious calories per unit area; use less water, energy, and fertilizer; increase soil fertility; and produce more income per unit area. Jeavons (2012, 215) also claims that it is possible to grow all the food for a person’s diet in as little
as 2000 ft², and that includes the area required for growing organic matter for producing compost to maintain soil fertility.

During the deindustrialization of Northern cities in the 1970s there was a mass movement of whites to the suburbs, leaving large areas of vacant land in the city. A survey of 70 cities in the United States (Pagano and Bowman 2000) revealed that on average 15% of urban land was vacant. Even these figures are gross underestimates as some cities do not have good records of vacant land parcels. Most urban farms in the United States grow crops on raised beds filled with compost made from readily available raw materials such as food waste, wood chips, and leaves. Often the compost is enriched with worm castings, a very low-cost method of using worms to break down organic waste into nutrient-rich soil amendments. Almost all poverty areas in US cities have very high unemployment rates and large numbers of previously incarcerated youth with limited job opportunities. Labor is not a limiting factor of production in urban agriculture. Biointensive urban agriculture requires little capital, uses little or no industrial inputs, and depends on the reproductive power of nature for growth and decomposition, rather than on the productive power of industry for inputs. Many urban farms, such as the ones run by Growing Power in Milwaukee and Chicago (Allen 2012), are not yet self-financing; their sales revenue is supplemented by modest grants from foundations and city governments. But it is too early to apply profit/loss calculations to urban mini-farms as they are an evolving technology in the stage of research and development. Moreover, urban farming has not received a fraction of the support that the federal government gives to conventional farming through crop subsidies, crop insurance, cooperative extension, and research from land grant universities. Urban farming exemplifies the use-value approach to the problem of poverty, and it also shows how each basic use-value draws on a specialized field of substantive knowledge.

Poverty and sites of agency

The most common definition of poverty – a lack of income – leads to the most conventional of solutions: increase employment in the corporate economy. This is the exchange-value view of poverty. But the economy does not produce full employment and does not pay living wages to even those who do find work. The economy systematically underproduces affordable basic use-values such as food, health, housing, energy, transport, and education. This is because the intentional construction of scarcity facilitates the expansion of exchange-value. Existing theories of poverty – conservative, liberal, and radical – fail to answer the question “What is to be done?” in a manner that confers agency on poor people. Conservatives point to welfare dependency and the need for unfettered markets. Liberals focus on race and gender as inhibiting social mobility. Radicals point to inherent tendencies in capitalism toward uneven development. These theories are not wrong; each emphasizes a particular root cause in what is a mutually constituted overdetermined system. Free markets, race, gender, and class are conceptual categories discursively aggregated at scales that are not correlated to the power of poor people. Existing theories of poverty have robbed generations of poor people of their agency. One solution to this problem is to explicate a use-value framework for poverty; it is a discourse that requires subject-specific information in the fields of food, health, housing, energy, transport, and education. Beyond the example of food production detailed earlier,
POVERTY

the use-value approach can be used to construct similar arguments in health (prevention of disease), housing (high-density living), energy (reduce entropy by matching energy sources to energy end uses), transport (decentering automobiles), and education (develop curricula to serve the sustainable production of use-values). The use-value framework turns away from the universalizing logic of economic growth and development to knowledge that is place-specific. In answering “What is to be done?” it produces knowledge that confers agency on place and community. Geography is central to the task of demonstrating how subject-and location-specific knowledge can be deployed to show that land, labor, and capital are not limiting factors for the production of basic use-values.

SEE ALSO: Dependency theory; Development; Discourse; Food security; Globalization; Inequality; Neoliberalism; Postdevelopment; Postmodernity; Poststructuralism/poststructural geographies; Scarcity; Social constructionism; Subculture of poverty

References


Power and development

Mary Lawhon
Joseph Pierce
Florida State University, USA

Development is, broadly speaking, the improvement of people’s wellbeing. Orthodox development thinking is often largely apolitical, despite the inclusion of vocabularies of power. The development community uses terms such as “empowerment,” “increasing capacities,” or the “ability to obtain resources” as key motivating vocabularies for the project. These terms are often used, however, without explicitly asking questions about how this increased power arises and where it comes from; what new relations are formed; which power relations have produced the existing patterns of capacity and resources; or how a redistribution of power might be necessary to achieve any form of improved wellbeing. Geographers have contributed to this depoliticized yet highly political development agenda both directly (by providing spatialized economic data to public and private actors) and analytically (by arguing for the benefits of well-executed development in the scholarly literature), and some continue to write in this vein.

Power is certainly a slippery concept, with diverse epistemological influences that shape how we understand its influence. The precise meaning of power in geographical writing on development is often implicit rather than explicit: most often, power has been transparently operationalized as the ability of some (usually Western/Northern/“developed”) actors to exert their will on other (typically Southern/“developing”) regions. This notion of power as the ability to force outcomes is often short-handed as “power-over,” as in “power over those with fewer resources.” However, over time geographers are becoming both increasingly explicit and more varied in their conceptualizations of power: some scholars argue that it can also be a means of enablement (“power-to” act and organize), and the distinction between power-over and power-to is increasingly important in understanding how geographers have used the concepts of development and power next to one another. Some of the specific treatments of power found in the literature on geography and development include power as an inscribed capacity which is located somewhere (a classical version); as relational and mobilized by groups towards specific outcomes; or as diffuse and mobilized through various processes.

This entry seeks to explicate the importance of power in development as well as how the concept of power is used in geographies of development. The development literature makes a distinction between immanent development (broad processes of structural or systematic change) and intentional development (specific interventions) (Corbridge 1998). This review discusses the role of power in both of these interrelated processes, with attention to different uses of the concept of power as well as key theoretical influences and epistemologies. It surveys significant trends in the discipline, while acknowledging that power/development is an important conjunction in the field and that there is wider geographical literature beyond this review.

One of the ways in which power has been exerted in the context of development is through language, and thus, before continuing, this entry...
briefly introduces some of the historically shifting regional terminology used, and its impact. The terms “First World,” “Second World,” and “Third World” emerged out of the Cold War context, in which the United States, the United Kingdom, and their allies were the First World, the Soviet Union and its allies were the Second World, and nonaligned countries were the Third World. Despite new geopolitical conditions, these terms remain in use today. While some consider the term “Third World” to be condescending, it is also used by progressive organizations such as the Third World Network. “Developed” and “developing” were subsequently popularized, although they have also been critiqued in turn for containing implicit normative assumptions and hierarchical power relations. More recently, the terms “Global North” and “Global South” have been increasingly used in an attempt to avoid the negative connotations typically associated with the Global South, although they are routinely critiqued for their geographical inadequacy/inaccuracy. Contemporary geographers have also sometimes opposed the underlying project of grouping of nations in its entirety, which inevitably essentializes diverse regions of the world and overlooks diversity at smaller scales. In the text below, the terms Global North and Global South are predominantly used, although alternative terms have been chosen where appropriate.

The powers behind development

Geographers (and the discipline of geography) were intimately involved in the expansion of Western/Northern power through the colonial project, as well as later expansionist American foreign policy. Throughout much of the nineteenth and twentieth centuries, geographers were complicit in the colonization of the Global South, contributing geographical data such as maps and regional studies. This era of scholarship began with the survey and publication of spatial data, but in later decades moved toward what are now seen as problematic efforts to explain spatial differentiation of privilege through race and climate. These now-discredited efforts helped to provide both an empirical basis and a theoretical justification for the exertion of what is typically described as power-over through colonialism, imperialism, and expansionism, as well as for later intentional development projects. Modernization became a critical aim, and development became central to global politics, particularly as the Cold War began to make international strategic allegiances increasingly important.

In this context, the World Bank (1944, originally known as the International Bank for Reconstruction and Development (IBRD)) and the International Monetary Fund (IMF, 1945) were founded in order to promote the internationalization of finance and support ongoing intentional development initiatives. Associated interventions sought to increase economic growth by bolstering the diffusion of technologies, markets, institutions, and practices from industrialized to developing contexts. However, optimistic projections of economic growth associated with development policies in the South often failed, leaving many countries (particularly in Africa and Latin America) unable to pay their debts. Geographers have demonstrated how this has shifted the site and scale of governance, giving international institutions power-over the governments in many nation states. Many countries in the Global South lost national autonomy as these international institutions pushed client nations into policies of “structural adjustment” which empower markets and private sector actors as key deliverers of development. Such programs generally reduce welfare support for lower-income citizens through market-centric programs, privatize state-owned enterprises
and public utilities, strengthen private property regimes, and reduce international trade barriers and tariffs. Collectively, these policies expose nations to the transfer of ownership of (and thus control over) key national industries to Global North actors, further reinforcing Northern decision-making prerogatives.

In addition to the IMF and World Bank, there has been a notable growth of nongovernmental organizations (NGOs) promoting intentional development, often through actions at a smaller scale aimed at improving wellbeing through specific projects. The power of NGOs is sometimes understood as “power-to” help Global South actors to achieve development. In this vein, a geographical lens can help us understand why NGOs in different contexts are comparatively more or less successful in achieving development outcomes; why NGOs are more present in some places than in others; and why their resources (human, financial, and other kinds of capital) flow to some places and not to others. Geographers have also attended to the shift in power relations between governments and NGOs: the growth of development NGOs means that these nonstate actors have moved to rival governments as the primary font of resource distribution (Bebbington, Hickey, and Mitlin 2008). This shift has rescaled the power relations of the aid community from that of various nation-states or NGOs toward internationalized systems and networks, largely understood as located in the Global North. More recent work has sought to move beyond individual case studies of organizations and towards understanding networks of social movements and the relationship between organizations, knowledge, and ideas. Taking this one step further, geographers have used the concept of assemblage, which promotes the understanding of social movements as performative and dynamic composites, rather than having distinct nodes between established entities (McFarlane 2009). McFarlane uses a case of Slum/Shack Dwellers International in Mumbai as an example of an NGO with international linkages, but not located in the North. The language of assemblage enables an understanding of power that is not about having power-over or power-to, but instead is a distributed, diffuse assemblage capable of mobilizing resources.

There has also been a significant reconfiguration of power relations that shape immanent development, or the broad processes of capitalist development. While trade has long been globalized, and economic relationships were a fundamental consideration of colonial geographers, increasing integration and new global production networks have shifted power. This includes a change from explicit and often violent power-over during the colonial era towards the greater importance of economic power-over, including transnational corporations which have in many cases become of greater significance in the development process than nation-states. New global production networks (GPNs) have become a focus of geographical examination, as geographers stress the need to move away from the typical focus of development studies on the state and towards the importance and cross-border activities of global firms. Geographical scholarship also seeks to move beyond the conceptualization of value chains towards understanding production and consumption as networked, nonlinear, and spread across space (Coe, Dicken, and Hess 2008). Economic and geopolitical power are also being simultaneously redrawn empirically and reworked conceptually as countries such as China, India, and Brazil, previously deemed part of the Third World, become increasingly central to production and consumption. Geographers have contributed overarching commentaries and theoretical developments as well as detailed localized examinations of the
POWER AND DEVELOPMENT

impact of such globalizations in the context of a capitalist economy, such as Hart’s work connecting South Africa with East Asia (Hart 2002).

Power as a lens of critique

While geographers have made important (and sometimes historically problematic) intellectual contributions to both the practice and our understanding of the shifting power relations associated with colonialism, globalization, and development, they have also been importantly involved in critiquing both their own intellectual history and international power relations. Critical scholars have suggested that the shift away from colonialism did not signal an historical shift away from the application of power in international politics, nor a move in the direction of international altruism. Rather, the rhetoric of power was transformed from one of political dominion to one of social and economic uplift in the orbit of successful examples in the Global North, while the practice of power shifted from an overt power-over to (neoliberal) economic and discursive expressions of power. Here are identified two key threads of critique, not entirely distinct from each other. The first emphasizes the need to understand and contest power in order to actually achieve development’s elusive ideal outcomes. The second questions the end goal, and suggests that the desire to achieve development itself is rooted in the imposition of power.

Analyzing power to better achieve development

Geographers have drawn on a number of different theoretical approaches to critique the dominant economic system, and most recently its neoliberal expressions, as well as that system’s impact on imminent development. In development studies generally and geography in particular, a number of these critiques begin with the articulation and rise of Marxist geographies from the 1960s onward. Questions of power slowly became central to understandings of development and underdevelopment; while the end goal of improved wellbeing remained relatively clear and uncontested, political-economic scholars began to question who gains and who loses in the development process, and whether development could ever be achieved through the actions being taken or processes being reinforced in the name of development. For example, drawing on dependency theory and world-systems theory, scholars have argued that historical power relations are merely reinforced through the subsequent efforts to develop. Increasing interconnectedness can, as Hart (2002) suggests, be disabling, reducing the autonomy of individual actors as well as nations in the context of established political economy. In short, macro-scale power relations established through colonialism and capitalism continue today, and must be recognized and contested in order to achieve development.

Economic geographers have also looked at the role of micro-scale power relations and the ways in which they constrain immanent development in the context of broad patterns of the contemporary capitalist economy. They have argued for the need to broaden scholars’ understanding of what constitutes the “economic” in order to see the networked, relational characteristics of development, and how power is enrolled in these relations. By emphasizing the roles that trust, associations, learning regions, institutional “thickness,” or placed networks may play in economic activity, these scholars offer ways of thinking about development that focus on its sociality and embeddedness, and how both power-over and power-to can be expressed through these
relations. In focusing on participants’ positionality, these scholars collectively provide a critique of development by suggesting that unless and until development attends to relationality and differences in power, it will fail to engender the improved wellbeing toward which it is nominally directed. As with Marxist scholarship noted above, such work typically implicitly retains a belief in the possibility of improving economic conditions and wellbeing through development.

Geographers have also critiqued the role of power in intentional development as it is conceptualized and performed by governments, donors, and NGOs. Some scholarship has been critical of the agendas of international government aid as well as that of NGOs, highlighting their tendency to reinforce the power relationship between donors and recipients and more generally their role in reinforcing broader international political economic relations. This includes the ways that government and non-government donors maintain their power to define agendas (a point geographers have particularly emphasized in research on the funding of conservation and sustainable development), have promoted particular technologies that increase dependency on donor economies, and promote export-oriented production to provide resources for more capital-intensive, profitable industries in the donor countries (Bebbington, Hickey, and Mitlin 2008). This critique has recently extended to include consideration of ethical consumption as a new funding mechanism through which intentional development occurs, and geographers have interrogated the use of development language as a marketing strategy that reinforces perceptions of difference and unequal power relations.

Another aspect of power relations that has been seen to shape international development is gender. One important insight that emerged from development theory, and continues to be significant in feminist development geography, is that women were typically overlooked in development plans and projects. Yet the decisions and activities of women are essential to the provision of everyday household livelihoods, and women often place greater emphasis on the wellbeing of the household (and particularly children) than men do. This insight and the associated critique have had practical outcomes, resulting in the allocation of funding and development of projects that sought to empower women by giving them power-to, both within the household and collectively. This included reconfiguring gendered power relations by increasing the decision-making and income-earning capacity of women in households and communities. This has had important consequences, not just in positive cases where changing the gendered power relations leads to improved wellbeing; it has also been noted to have threatened the empowered positionality of male household members, at times resulting in resistance, confrontation, and even violence within the household (Radcliffe 2006).

Development as the imposition of power/knowledge

Development has often been presented in both economic and specifically geographical scholarship as an unambiguously positive thing: surely no one would oppose improving wellbeing? In sites of development practice, however – largely in the Global South – the concept and its concomitant practices are much more contested. Resistance and opposition are as old as the concept, but critical geography was somewhat late in its engagement with postcolonial theory. Postcolonial theory suggests that the question of power is not just the contestation over resources or material, economic power; development is also about the power of ideas and the power to define norms and aspirations (Corbridge
1998; Crush 1995). Influenced by the work of Said, Spivak, and Escobar, but also drawing on Northern scholars such as Foucault, geographers have shown the discursive power of the development discourse. The aim here has been to problematize the self-evident; while it may surely be the case that improving wellbeing is desirable, geographers argue that there is much more to the practice and language of development, and a multiplicity of ways of understanding wellbeing (Crush 1995). The critique of the language and concepts is more than theoretical; it is rooted in a materialist desire, as so much is done in the name of development. This discursive critique seeks to show the connection between Western/Northern knowledge and power; power in this case is neither repressive power-over nor a liberatory power-to, but power as a universalizing discourse that shapes norms and desires.

An important locus of geographical scholarship has been the reconsideration of the aspiration of a single, unitary “modernity” that often underpins development theory and practice. Critical scholars suggest that modernization takes for granted that Northern economic patterns are “modern,” desirable, and beneficial for the residents of developing nations. Local case studies have been combined with broader theory to show how an understanding of geography helps to draw attention to the myriad networked influences on place, the diverse experiences of modernity, and the different possibilities and futures that may emerge. Geographers have also attended to the ways in which modernity is opposed, reappropriated, and reworked; Pred and Watts’s (1992) notable work examines wide-ranging examples, from technological changes in rice production in West Africa to labor conflicts at a steel mill in the United States, in the context of globalized ownership.

Geographical research drawing on postcolonialism and postdevelopment has also critiqued conventional development narratives (and previous scholarship) that treat recipients as objects of development. One of the consequences of this has, arguably, been the internalization of the development discourse by the people to be developed, thereby creating passive “objects” of development (Crush 1995). Instead, geographers and others argue for the need to recognize and respond to people as pluralistic, complex subjects with agency to resist, reappropriate, and subvert development projects and agendas (Crush 1995). Examining the power that these subjects have highlights another rescaling of analysis away from global/geopolitical outcomes of development (e.g., the degree to which development as a whole produces outcomes), as well as from the success or failure of individual plans and projects, and toward the networks of resistance and responses to development at the sites of investment or extraction.

**SEE ALSO:** Dependency theory; Development; Globalization; Networks, social capital, and development; Postcolonial geographies; Power

**References**


Further reading


Power

John Allen
The Open University, UK

Power, you would have thought, needs no introduction. A working definition captures its distinctive feature; namely, the ability to get people to do things that they would otherwise not have done. A long line of thinkers, from the political philosopher Thomas Hobbes, through Max Weber and Karl Marx, right up to present-day geographers David Harvey and Richard Peet, would have little difficulty in recognizing this definition of power. It is a definition of instrumental power; instrumental in the sense that it involves the exercise of power over others, whether through an ability to dominate, coerce, or subject someone to the rules of authority. Power, on this view, is a zero-sum game; it is always exercised at the expense of others.

Such a view chimes with the experience of imposition, of not being entirely in control when on the receiving end of an organized or institutional power. It is a view of power from which it is hard to digress. Yet for an equally long time, another less strident definition of power has run alongside it, in tandem, so to speak. Less instrumental in its force, it simply involves the power to make things happen, to make a difference, or to secure an outcome. Power, on this view, is a means to get things done, an enabling tool or facility, and has its historical roots in the statecraft of Niccolo Machiavelli. In more recent times, the legacy of this view is to be found in the work of such disparate figures as Hannah Arendt, Michel Foucault, and, Michael Mann.

Geographers have made their own use of these two different readings of power, often adapting them to their concerns and only showing fidelity to one or the other conception when it aligns with their particular interests. Such slippage arguably owes less to any profound eclecticism and rather more to the fact that when the likes of space, territory, and distance enter into the analysis they begin to trouble primary distinctions of power and authority.

“Power over” and “Power to”

It is possible to conceive of the “power to” act as a general precondition for all kinds of action, of which the power that is wielded at somebody else’s expense is a particular type of constraining action. On this view, the ability of the weighty institutions of the world to hold “power over” others reduces itself to one among other ways of acting in the world. Of late, however, the subsumption of instrumental power under a more general, catch-all, facilitative definition of power has given way to a more direct contrast between the two approaches. Whereas the ingrained idea of power as a force to be reckoned with, something which is held over others, speaks to domination and constraint; the looser notion that power is something that makes things happen, speaks to an enabling sense of action. Where the former accentuates the repressive side to power, the latter stresses its potential for empowerment. Where one is negative, the other is positive.
about the consequences of power and its actions. The distinction has hardened between the two senses of power and with it a gulf has opened up over what power is deemed to be for and how it works.

Instrumental power, or power over others

With only a slight hint of exaggeration, it is probably fair to say that an instrumental conception of power represents the default position within geography, especially political and economic geography. The ability to bend the will of others, to exercise power to get people to do things they otherwise would not have done, lends itself to questions of state rivalry and economic domination, as much as it does to issues of leaders and led, hierarchies, and the legitimation of authority. All of these characteristics imply that there is no “half-way” house about the distribution of power; either you have it or you don’t have it, and those that have it use it to further their objectives despite possible resistance from others.

What is also evident about this default position is that power is something that is “held” – it is held over others to gain advantage or to bring them into line. On this reasoning, power has something of a dispositional quality; that is, a capacity or disposition toward acting in certain ways, even if that potential is never actually realized. The capacity to do something, in this instance, is separate from the actual exercise of power. Such capabilities represent latent rather than observable qualities of power, so that the capacity for domination held by state agencies or economic corporations is conceived as present and capable of being exercised should the need arise or circumstances dictate. In practice, however, the possession of power and its exercise tend to be conflated, as if they were one and the same thing.

It is not uncommon, for example, to “read off” the power of an economic institution or governing body by the size of the resources at its disposal, regardless of how they may be used, or indeed misused. Size matters in this instance because it can be an indication of a concentration or “bloc” of power. The greater the resources concentrated at one location, the greater, it is often assumed, is the capacity for domination and control. Barry Hindess once described this as a quantitative conception of power, where more power prevails over less and it is possible to determine in advance who wins and who loses, regardless of context. In the “real world,” as he points out, one would expect a somewhat messier, qualified outcome, yet the obviousness of this zero-sum imagery underpins much of what is taken for granted when power is exercised over others in a hierarchical, top-down fashion.

The clue, perhaps, is in the instrumental language itself; when something is exercised “over” others it implies that power is located above the mainstream, at the apex of political organizations or economic enterprises. We owe it perhaps to Max Weber for this view of organizational domination, whereby those in authority attempt to order people’s lives through the bureaucratic character of rules and regulations which are more or less impartial. The removal of choice and the imposition of constraint lie at the heart of this process and institutional power is extended downwards through delegation with the attendant risk of losing authority or destabilizing the decision-making process. On this asymmetrical view of power then, effectiveness cannot be wholly guaranteed in advance and the pay-offs do not automatically sum to zero. Nonetheless, the zero-sum imagery holds, as does the spatial hierarchy of power. Weber’s legacy within geography is less clear than that
of Marx’s, but the more complex calculations of organizational power that he first outlined have left their mark, not least on those who have had to engage with the state’s territorial significance and the distribution of its authority.

Bob Jessop (2008), a contemporary political theorist, is one such figure whose work demonstrates a grasp of such complexity, although his account of the state’s instrumental use of power draws primarily from the Marxist tradition, in particular the writings of Nicos Poulantzas. For Jessop, the dispositional character of the state’s power is best spoken about in the plural, as a series of state capacities, not as a unitary subject in possession of preformed powers. On this view, the state is an ensemble of institutions made up of different interests with different capacities who influence and mediate the balance of class forces at large. The effectiveness of state power in this context is contingent upon the changing balance of political forces located both within and beyond the apparatus of the state itself. In place of a simple disposition or capacity to dominate, the political landscape is one of potential structural powers that may or may not be actualized depending upon the precise balance of forces in place.

If that sounds a little bit convoluted, that is precisely the message that Jessop wishes to convey. A crude instrumentalist approach is rejected in favor of a complex institutional system, one that nonetheless serves to reproduce capitalist social relations through the state’s ongoing political domination of society. If state power is institutionally mediated in such a complex, circuitous manner, however, it remains a hierarchical structure of potential powers, albeit embedded within a variety of “power centers.”

The importance of knowing where power is located is a constant in any instrumental treatment of power, Marxist-inspired or otherwise. One of the most influential treatments in this vein is David Harvey’s (2003) account of the workings of contemporary US power in The New Imperialism, in which he identifies two key axes, one political, the other economic. Taking his cue from Giovanni Arrighi, he outlines two logics of power, one territorially based, which involves the state as the main political actor orchestrating its powers to control the often highly charged landscapes in which capital accumulation takes place and a second capitalist logic predicated upon flows of economic power across territorial boundaries in the endless pursuit of capital accumulation. The two logics combine, yet each has a life of its own, one awkwardly fixed in territorial space, the other highly mobile, a fact which builds into the system the likelihood of antagonism and their volatility and unpredictability.

Political power, on this view, thus acts as the stabilizing force in the system, in much the same way as in Jessop’s analysis, as it works to balance the haphazard movements of capital and its aligned forces. For all its built-in tensions, however, power operates along familiar hierarchical lines with a Wall Street–Treasury–IMF complex situated at its apex. The concentration of capabilities at this one location tells you all that you need to know about the disposition of US power. Where the analysis departs from that mold, however, is through the recognition that the rest of the world is not so much under the command of this power center as “networked and successfully hooked into (and effectively ‘hooked on’ usually by way of credit arrangements) a structured framework of interlocking financial and governmental (including supranational) institutions” (Harvey 2003, 134).

Such slippage in the language of instrumental power is itself indicative of the troubling effect that geography can have upon more well-worn treatments of power. We return to such matters after first considering a different understanding of
power; one that focuses on power as a means of enablement rather than a capacity for domination and control.

Enabling power, or power to make things happen

If instrumental power represents the default position within geography, something that is adopted for want of an obvious alternative, enabling power has made its way into the discipline by a more roundabout route. The looser sense of the “power to” act can be spun out in a range of different directions, designed to secure mutually beneficial ends as much as far-flung goals of a more normative persuasion. Here, the clue is in the word itself – enabling – in that power can be made to “work” for a variety of ends in what can amount to a positive, rather than a zero-sum, game. As a means to an end, power as such does not have to serve any particular interest or be tied to practices of domination and control. It can be designed to further common purposes or to foster transformation, in so far as it can “make a difference” to eventual outcomes. That does not mean to say that such outcomes are always benign, however, only that they may be so.

With this more nebulous conception of power, it is harder not only to pin down, but also to locate. In contrast to an instrumental view of power, when power is conceived as a medium there are no readily identifiable concentrations of power, only a sense of how it is generated through the mobilization of collective or individual resources. On this view, it makes little sense to talk about “blocs” of power when the means itself can expand in line with the resources available or contract once collaborative, short-term goals have been achieved. Rather, it is sustained through patterns of association that bind people together in the pursuit of particular ends. As such, its geography is less centered and more dispersed over tracts of space and time, in line with the forms of social interaction generated.

It is perhaps not that surprising then that one distinctive twist to this account comes through the recognition that the mobilization and control of resources takes place through networks of social interaction. The work of Manuel Castells is probably among the best known for such an orientation. In Communication Power (2011), drawing upon his earlier account of networks as means to overcome the barriers of distance to secure outcomes, he focuses explicitly upon what he calls “network-making power”; that is, the ability to program the content of networks in line with specific goals as well as ensure a unity of purpose through the control of the points where different networks connect. Programmers and switchers are said to shape the communication content of networks that connect the local with the global and give them extensive reach. Different types of networks – political, financial, military – interact, although each is programmed to achieve its specific goals across a range of interfaces. This understanding of power networks, however, only really comes to fruition in the detailed treatment it receives at the hands of Michael Mann.

In his magisterial four-volume work on the sources of social power, he sets out a notion of power resources as something which, historically, have been organized and extended over space in loose or tightly orchestrated ways to achieve identifiable goals. The expansion of power and its consolidation are said to take their shape from a series of networks organized over space which overlap one another, of which the most significant in terms of power resources are economic, ideological, political, and military. Differences in the makeup and dynamism between the networks ensure that they reach out across
space in different ways and to varying extents, but the most effective networked practices, for Mann, are those which blend different forms of organizational reach. In his account of the makings of power in the twenty-first century (Mann 2011), political relations are presumed to channel and institutionalize each of the other power resources, although in practice he details how powers are often poorly organized, uneven and incoherent, especially when it comes to the manner in which US power has been extended across the globe in the contemporary period. Nonetheless, power for Mann is the product of pooled resources that are used to fuse and modify patterns of interaction to achieve far-flung goals.

As a fluid medium, power in this pooled sense has also been interpreted in at least two different ways to give a more associational spin to networked power. The first draws inspiration from actor-network theory, principally the contributions of Bruno Latour (2005) and Michel Callon (1998), and the second, from a different angle, is rooted in the mutuality of Hannah Arendt’s thinking (1958).

In the former set of hands, the “work” of the network is stressed; that is, the “power to” bridge, broker, and connect people and things together to bring about some form of alignment. The ability to forge associations by enrolling, translating, and channeling others into networks of meaning in such a way that they extend and reproduce themselves through space and time represents a mediated exercise of power; one where distances are overcome by the successive enrolment of others to form what amounts to a single will across the network. Leverage, on this account, is achieved through association rather than simply domination. The same associational tone is evident in Arendt’s understanding of power, but for her the formation of a common will is less something forged and more the product of mutual action, of people acting together to pursue common, agreed goals. The powers of association, on this interpretation, are thus something that is experienced through other people. As such, it is more about exercising power with others where the power to act is empowering in and of itself.

Power, on this view, has a positive side, one that can open up possibilities rather than simply close them down. This positive side to power was also pointed out by Michel Foucault (1984), where the emphasis upon people being free to govern themselves works through a kind of normalizing force. On this view, there are no direct constraints on behavior, no overt sanctions, only indirect techniques of regulation through which people freely fashion their sense of self. Subjects are constituted by the spacing and timing of their own activities as much as they are by those seeking to influence their behavior. As such, their conduct is said to be shaped as much by what they absorb and imagine the “truth” of their circumstances to be as it is by the spatial layout, distribution, and organization of their surroundings.

Foucault’s account of dispersed government and his later concerns with biopolitics and liberalism have had a significant influence on geographers, most notably in the ongoing body of work which falls under the heading of critical geopolitics and, more generally, through those more concerned with the technologies and rationalities of government. The former, although less concerned with the relationship between politics and space, has moved beyond a concern with the power/knowledge of statecraft to embrace a wider cultural and political agenda, while the latter has found a renewed significance with the publication of Foucault’s lecture series. On both counts, though, the power that is brought to bear upon people’s actions is often more about closing down rather than opening up possibilities. It is at this point that the “power to” provoke can fold over into the “power over” others. Such
ambiguity is plainly evident in the writings of Foucault, but when geographers introduce space and territory more centrally into the analysis the notion of how power actually works out in practice seems to become more troubled.

**Powerful geographies**

It would be false to say that instrumental and enabling notions of power have developed without taking geography into consideration, but for the most part the actors involved have been sketched against a given spatial backdrop. Even Harvey’s account of US power has an implicit geometry that rarely troubles the operation of power itself and makes for easy distributions. Since the publication of John Agnew’s seminal work on “the territorial trap” (1994), however, which highlighted the strain between how state power is practiced and the inability of bounded territories to contain the relationships involved, it has been harder to ignore the fact that spatiality itself can make a difference. Indeed, there is now a greater willingness among geographers to problematize power and spatiality, and also to question the place of instrumental power as the default option.

**Territory in question**

Neil Brenner (2004), following Bob Jessop, is perhaps the most consistent exponent of the view that state power has been liberated from the idea of containment and contiguity, in terms of units of spatial authority seemingly piled one on top of the other. States increasingly face pressures from both “above” and “below,” which have resulted in the displacement of their authority. Administrative and regulatory functions have, as it were, moved “upwards, downwards, and sideways” (Jessop 2008, 196) as national states find themselves destabilized from above by the growth of supranational institutions and undermined from below by the devolution and decentralization of decision-making powers to subnational institutions. On this view, the larger number of institutional interests on the political landscape, in particular the multiple sites of authority, from numerous quangos and private agencies to local administrative units, have undermined the notion that the central state apparatus is the prime locus of political power. In place of the conventional assumption that the central state is the only actor of any real import, the institutional playing field is now shared with nongovernmental organizations, multinational enterprises, and other supranational as well as interstate organizations. At any one point in time, the relative significance of each scale may change in that the power can be scaled up or down through the different units of spatial authority both transnational and subnational. This is an altogether different arrangement of powers from one that considers regulatory or devolved authority as something which is slotted into pregiven spaces, be it local, urban, regional, or national. In recognizing that space is not simply “filled out” by power relationships or acts merely as a container within which things happen, the geography itself becomes pliable, open to manipulation and modification. The lines of authority, as much as where the edges are drawn, and the scales of regulation, as much where their extension halts, are themselves the subject of political construction, relative rather than absolute in their geographical form. Multicentric, multiform, and multiscale are among the descriptors used to capture this new found spatial complexity as the coordinates of power lose their territorial fixity.

An equally nuanced use of the notion of territory is to be found in Stuart Elden’s
Terror and Territory: The Spatial Extent of Sovereignty, which attempts to work through why territory matters in an age when boundaries have not only become more porous, but state territorial sovereignty itself is increasingly under threat. Territory, or more specifically the connection between the state and its territory, is for him one of the key sites of struggle in the contemporary geopolitical moment. In line with Brenner, territory is not seen as a static backdrop with fixed coordinates, but as an arena through which particular geographies of fear, threat, and division are played out. Elden’s prime concern is with territorial integrity: the spatial extent of a state’s sovereignty, that is, its effective political control over a given territory. Such integrity, he argues, has been fractured, as more or less failed states have had their sovereignty undermined by the international community in the name of security and the “war on terror.” Yet, at the same time, that very same community has worked hard to defend territorial integrity in the interest of preserving stability and the right of states to exercise power legitimately within their borders. On the one hand, the sovereignty of states has been drawn into question when they are deemed a threat to the vital interest of others, yet, on the other hand, the spatial extent of their powers has been defended and their borders deemed inviolable.

The image is one of territorial political actors, the more dominant of which extend their powers over the territory of others when “global” danger threatens. There is no pre-giveness to territorial power and, as with Brenner, territory is conceived as an achievement, something that is actively produced. Unlike Brenner, however, there is less of a preoccupation with the redrawing of bounded spaces, of the need to continuously recalibrate spatial units of authority or, in Elden’s case, territorial boundaries. Rather, the focus is more about the ability of certain powerful states to curb the powers of other states elsewhere and, in so doing, throw into sharp relief the making and remaking of territories.

A similar concern with the making of territory is at the heart of Saskia Sassen’s (2006) Territory, Authority, Rights: From Medieval to Global Assemblages. For her, the mix of spaces and times that inhabit the national setting defines a new, overlapping geography of power within which different actors – state authorities and jurisdictional agencies, corporate firms and supranational institution, civil society movements, and transnational activists – jostle, co-exist, and interrupt one another to gain advantage. As she sees it, the disembedding of state functions and the growing authority of nonstate actors in the public realm that Brenner also describes opens up spaces within the formally exclusive territory of nation-states where global firms are increasingly subject to extraterritorial forms of authority. Parts of global cities like New York and London, for instance, predominantly their corporate finance sectors, are seen to be partially detached from the geographically circumscribed authority of the state. This, however, is not a geography where some parts of national territory are said to hover or float above it; rather, such economic spaces remain firmly embedded in national territories, yet are subject to wider geographical authorities when it comes to regulation and control.

The detachment of authority from territory also brings into question the idea that borders are always at the edges of any given territory. For Sassen, such is clearly not the case, as borders are not so much redrawn as re-embedded within national territory, in line with the form of authority under consideration. The disruption to a fixed geometry that such insights presuppose, however, does not lead her to problematize the notion of extraterritoriality, nor question
instrumental power, in particular authority, as the only relation that really matters.

Far-reaching power

Extraterritorial power can be translated literally as an addition to a state’s territorial powers, but to do so offers little insight as to how such spatial reach is obtained. For geographers more concerned with such questions, the ability of the state or any institution for that matter, to exercise power at-a-distance is more often than not a work of associational or networked power, although an alternative immanent form of reach is evident in the work of Hardt and Negri, which had an impact within geography. Their debt to Foucault (and Deleuze) is evident from the manner in which power is said to take its shape from the simple act of people living out their lives in circumstances that mold their behavior and actions. Unlike associational or networked accounts of power, however, the mediated nature of reach is largely absent when immanence comes to the fore.

For that, we have to turn to the work of Joe Painter (2006) who has shown how everyday life is suffused by state practices in a rather mundane, prosaic manner. Rather than try to identify the actions of the state from above, he argues that, following Timothy Mitchell, there is an element of reification to the power of the state as something separate from civil society. Both the state’s power and its territorial base, in his view, are best understood as an effect of networked practices within and outside of state institutions. In line with Latour and Callon, the “work” of the bureaucratic apparatus relies, in the main, upon routine certification, registration, and licensing practices that reach directly into people’s daily lives and serve to both shape as well as enable their behavior. The enabling side to such practices is clearly in direct contrast to structuralist accounts of state power, exemplified in the work of Jessop.

The focus upon state practices as ordinary, often tending to pass unnoticed, draws attention to the agency involved and leads quite naturally to accounts of state power which point up its hegemonic qualities. Among the most prominent within geography is Agnew’s (2005) Hegemony: The New Shape of Global Power, which is an attempt to show just how much state power relies upon both consent and cooperation. On this view, the ability to persuade and induce is as much in evidence as coercion and constraint. Such work chimes with Joseph Nye’s influential (2004) Soft Power, where the stress upon cultural as well as political values as a source of attraction and desire for others is said to form part of an exercise in molding people’s expectations about what it is that they want and aspire to. The eventual outcome of such practices may be instrumental in shape, favoring one set of interests at the expense of another, but the routine nature of their exercise relies upon people buying into those interests rather than having them imposed upon them. Indeed, the employment of possibly more artful practices by state agencies which may mask their actual intent draws attention to the slippage that can take place between instrumental and enabling conceptions of power, where the gains of a cultural positive-sum game, for instance, may turn out to be illusory.

Such hegemonic accounts tend to rest upon the assumption that power is capable of extensive spatial reach; that is, an ability to permeate daily life in such ways that people bring themselves to order. Painter, however, is hinting that power is more than simply an extensive force. For him, the ability of the state to become present in the here and now of daily life suggests a more intensive reach where the barriers of physical distance are not at issue. When the “reach of government” is understood as intensive, such
actions are said to be topological, in that the reach of government bodies enables them to make their presence felt in more or less pervasive ways by cutting across proximity and distance (Allen 2003). It is not just the reach of the state that can be viewed in this manner, however, so too can the actions of far-flung corporations in making their presence felt at close quarters, or the ability with which social movements are able to fold distant harms into local campaigns. Such power-topologies are said to transcend a landscape of fixed distances and well-defined proximities and pose a challenge to conventional territorial accounts of geography and power.

Conclusion

The two main readings of power, instrumental and enabling, have had an uneven yet influential presence across the social sciences. Among geographers, they have been adopted and troubled in roughly equal measure. The subjection of power to a spatial agenda, whether that be through the production of territory, the mobilization of networks, or through a more topological register have, in their own ways, brought conventional accounts of power into question. But the challenge of geography today is not only spatial, the natural or environmental also has a part to play in the figure of the more-than-human. The inspiration for thinking more broadly about the powers of life, rather than simply power over life, largely comes from Foucault’s biopolitical critique, but it is extended further when the question of power relates to the powers in and of things. How geography engages with the provocations of vital materialism, the powers of the nonhuman, especially the powerful physical forces that work themselves out on a planetary scale, has the potential to trouble mainstream accounts of power even further.

SEE ALSO: Biopower; Geopolitics; Governmentality; Power and development; State, the

References


Precarious work

Kendra Strauss
Simon Fraser University, Canada

Precarious work (also called precarious employment) is a relatively recent addition to the conceptual toolkit of economic and labor geographers. Developed by labor scholars to describe what was understood, conceptually and empirically, as a significant shift in the nature of employment in the advanced industrial economies after the 1970s, the concept encompasses several processes and dimensions. These include the rise of temporary and intermediated forms of employment, the changing nature and allocation of social and economic risks, the feminization of paid work, and the experiences of migrant workers at the “bottom end” of the labor market. In this sense research on precarious work in other disciplines overlaps with evolving preoccupations of labor geographers (Coe 2013) and economic geographers interested in subcontracted and temporary work and migrant labor (McDowell, Batnitzky, and Dyer 2009; Wills et al. 2009). Not solely confined to low-paid work, however, geographers have also identified precarious employment relations among workers in the creative and knowledge economies (see e.g., Hracs and Leslie 2014; Sanchez-Moral, Mendez, and Arellano 2014).

At the same time, work by geographers such as Louise Waite, Gill Valentine, and Hannah Lewis (2014) has explored the precarity experienced by asylum seekers and refugees in the United Kingdom across multiple dimensions. Here, precarious work – in some cases forced labor – interacts with precarious legal status in a more generalized way. This approach aligns with broader currents in political and cultural geography, which build on theorizations of the condition of precarity and which understand vulnerability as socially and politically constructed and combine different scales of analysis – including the microscale of the body.

The concept of precarious work

Precarious work is a concept that has developed to describe changes in labor markets, especially but not exclusively in the Global North, from the late 1970s. The development of the notion of precarious work aligns with a broader set of debates about political economic and social shifts described by neoliberalism, post-Fordism, and the rise of the service economy. A diverse range of geographers have written about these trends, with work by inter alia David Harvey, Jamie Peck, Ash Amin, Allen J. Scott, and Doreen Massey examining the changing organization of production, the globalization of labor markets, deindustrialization, and the forms of technological change associated with the knowledge economy. In relation to the last mentioned, economic geographers have examined the expansion of high tech industries and associated forms of geographical concentration (agglomeration, clusters) with ambivalent outcomes for labor; the extensification of production has been explored via the frameworks of global value or commodity chains, and global production networks.

The picture that has emerged in many economies is one of increasing polarization.
between highly remunerated workers in professional services at the top end of the labor market, and low-paid insecure conditions for workers in food services, hospitality, agriculture, care, and domestic work at the “bottom end.” Associated with the creation of “flexible” labor markets, the political economic shifts that have underpinned this polarization include legal and regulatory changes to systems of labor law, employment standards, and occupational welfare (pensions, unemployment insurance). These shifts have in many cases been designed to reduce what orthodox economists refer to as rigidities: the protections (e.g., associated with hiring and firing workers) that, in their school of thought, interfere with the efficient functioning of labor markets. The abandonment after the 1970s, with Keynesian macro-economic policy, of the goal of full employment also allowed governments to deal in very different discursive and policy terms with the unemployment created by industrial restructuring and deindustrialization.

These trends have produced a hollowing out of the occupational structure of labor markets, with formerly middle-income jobs – from unionized blue collar to clerical and lower level white-collar work – becoming either professionalized or more precarious. Privatization and subcontracting in public service professions like health and education have reduced pay, benefits, and security for those in previously solid middle-class jobs. While these trends have been uneven and contested, like other dimensions of neoliberal de- and re-regulation, labor market insecurity and polarization have increased in almost all countries, including the formerly social democratic Scandinavian states – albeit from a much lower level. There is also variability within what Gosta Esping-Anderson described as the three worlds of welfare capitalism, so countries with some similar characteristics (like Sweden and Norway) have taken different paths to labor market reform. Declining unionization in many countries has compounded, and been exacerbated by, these changes, and in the Global South structural adjustment policies have also been significant. Labor geographers like Andrew Herod, Andrew Cumbers, and Danny Mackinnon have explored these trends in relation to traditional unionized sectors, emphasizing that workers have not been passive victims of the processes that have increased precariousness for many.

Yet, despite many common themes in these accounts of political economic change, the idea of precarious work is not a straightforward one. As Linda McDowell and Susan Christopherson (2009) have noted, precariousness and precarity have different meanings, grounded in different intellectual and political genealogies. If the former is associated with the political economy of work and employment, the latter is grounded more directly in Marxist approaches and political mobilizations in France (around the concept of precarité) and Italy (the San Pracario movement). In France, the idea of precarity was first associated with Pierre Bourdieu’s work on contingent and nonstandard forms of employment. In France and Italy, however, and arguably in Europe more broadly, an initial focus on wage work and employment shifted to a wider political concern with, and mobilization around, generalized conditions of insecurity across – and connecting – the domains of production and social reproduction. Some commentators have noted that direct political action under the rubric of precarity seems to have peaked in countries like Italy around 2004–2005, while academic work on precarity and precariousness (still relatively rare at that time) has since proliferated.

In the Anglo-American nations, however, explorations of precariousness have been more directly associated with academic research on the changing nature of employment in the
post-Fordist period. The concept of precarious work emerged out of research by theorists of labor market segmentation and “nonstandard” work in disciplines such as sociology, political science, and feminist legal studies. Influential studies by Leah Vosko, Judy Fudge, and Arne Kallenberg built on earlier research by Gerry Rodgers that identified four key dimensions that could indicate insecurity in an employment relation, and hence the condition of precarious work (also called by Vosko the “precarious employment relationship”). These are uncertainty about the continuation of employment, lack of control over the labor process (related to the presence or absence of employee voice), the absence of regulatory protection, and low pay. Rodgers emphasized, however, that the concept of precarious work encompasses more than just the form of employment and recognizes wider conditions of insecurity including a lack of legal and union protection and forms of social and economic vulnerability. This approach is related to theories and studies of labor market segmentation and forms of social stratification which, like Jamie Peck’s work in the mid-1990s, also built on earlier models, for example, of the dual labor market with primary and secondary sectors.

A key focus of theorizations of precarious work, however, has been the decline of the standard employment relationship (SER) and related standard employment contract in the advanced industrial economies in the Global North. The SER was understood as representing, in a sense, the mirror image of the conditions of insecurity identified by Rodgers. It encompassed full-time, relatively secure, continuous, direct employment, often with a single employer over the life course, and was associated with both industrial and white-collar jobs in the “golden age” of industrial citizenship after World War II. It was also associated with both the male breadwinner model and family wage, and with the social wage (into which occupational welfare benefits like health, disability and unemployment insurance, and pensions were enfolded). As feminists and critical race scholars have long highlighted, the SER was always only a reality for some: women, workers of color, and immigrants were fully or partially excluded from its protections, which in any case varied spatially and temporally. Others, like Neilson and Rossiter (2008), argue that Fordism’s “golden age” was not only partial but also exceptional and that capitalism has always been characterized by precarious work. What labor market restructuring, flexibilization, and re-regulation in the post-Fordist (or neoliberal) period thus represent, according to them, is a return to longer term historical trends.

In terms of reproduction, too, it has been argued that the male breadwinner model associated with the SER was not only normative and aspirational for many (rather than descriptive), but also exceptional. Certainly, changes in relations and institutions of societal reproduction have also occurred in the period that has seen the rise (or resurgence) of precarious employment. On one hand, women’s increasing educational attainment and entry into the labor market in large numbers has been of huge significance in relation to the changing composition of the workforce over the last half-century. The composition of the household, from falling birthrates to increased divorce, unmarried co-habitation, and single parenthood, has changed enormously. Thus wage stagnation in the middle and lower income brackets in, for example, the United States has been offset by the rise of the dual-earner household. On the other hand, however, state decommodification of services like childcare, education, and health has been impacted by welfare state retrenchment in relation to what feminist political economists have characterized as the re-privatization of
social reproduction. As women often depend upon these services and sectors for support and paid employment, such changes have gendered implications for both paid and unpaid work, making life more precarious across these dimensions. In Asia and Latin America, capital has sought the “nimble fingers” of young women for its new manufacturing workforce, while states like the Philippines and Indonesia depend on remittances from women who migrate to perform domestic and care work in the homes of middle-class and wealthy women.

Thus not all workers have experienced the trend toward increased precarious work in the same way, or to the same extent, since the late 1970s. Vosko, in her theorization of precarious work, highlighted the importance of social location and social context. She defined precarious employment as “work for remuneration characterized by uncertainty, low income, limited social benefits, and statutory entitlements,” but went on to argue that precarious employment is shaped by the relationship between employment status, form of employment, dimensions of labor market insecurity, and social context and social location (Vosko 2010, 2). While social context signifies characteristics like occupation, industry, and geographical location, social location relates to the interaction of social relations and categories, like gender, with legal and political categories, like citizenship. From a geographical perspective, social location has distinctly spatial and scalar dimensions. The role of place in the social construction of local labor markets, the interrelationship of spaces and places of work (and home) through varying and evolving types of linkages, and the increasingly “glocal” nature of economic activity all relate to the highly uneven nature of the restructuring of employment relations. At the same time these processes shape, and are shaped by, the social construction of categories of difference like gender, race, and class.

Much work on precarious employment thus overlaps with and incorporates concepts like feminization: the idea that “mainstream” employment relations increasingly resemble conditions in the secondary sector in general, and women’s employment in particular. Part-time, temporary, contract, and agency work do not figure in the SER. Historically these forms of employment were associated with women and racialized workers who, because of caring responsibilities or their subordinate labor market position, were unable to secure full-time permanent wage work. Geographical research, like the work of Susan Hanson and Geraldine Pratt, has shown both how the spatial distribution of home and work sites affect women’s choices and opportunities for paid employment, and how (in Pratt’s work on live-in caregivers in Canada) middle-class women’s labor market participation often depends on the availability of low-paid racialized domestic workers to “replace” their labor in the home. Subordinate labor market position and caring responsibilities are thus often connected, but subordination is not uniform among women. Many women of color held down full-time jobs doing paid reproductive labor in homes and public institutions, as well as having caring responsibilities in their own households — which often went unrecognized in accounts of women’s work.

Vosko’s argument was that the kinds of insecure employment conditions experienced by women temporary agency workers in clerical occupations, previously the exception, are coming to look more like the norm. Like geographers including Jamie Peck, Nik Theodore, Neil Coe, and Kevin Ward, Vosko is particularly interested in labor market intermediaries and how they function within labor markets. Intermediaries include temporary employment agencies, labor brokers, labor dispatch operators and labor contractors. While temporary agency
work was previously limited in its scope, and confined to the secondary sector (pink collar clerical work in Canada was the example used by Vosko, but other sectors like agriculture are also characterized by the presence of labor brokers), it has now permeated all sectors. Heightened precariousness in employment is directly related to increased intermediation in the employment relation. In other words, although historically particular to marginalized groups, many workers are no longer hired directly by the firm or individual on whose worksite they labor. Moreover, such intermediated employment relations often involve short-term and/or temporary job placements, fluctuating and irregular hours of work, lower levels of pay than those earned by permanent employees in the same job, and low levels of job security (Strauss and Fudge 2014). Again, however, not all workers experience precariousness associated with intermediation in the same ways. Higher skilled workers in sectors like IT, while precarious across several domains, need to be distinguished from workers in low-wage occupations with little mobility, power, or discretion in the employment relation. However, Peck and Theodore (2007) have demonstrated that labor intermediaries now play a key role in providing numerical flexibility not only at the firm level, but in the macro-context of overall labor market adjustment during periods of crisis and recession. Temporary workers are increasingly the buffer that firms and sectors use to smooth business cycle adjustments.

Two additional trends related to the rise of precarious work are political-legal and institutional in nature: changing regimes of labor law and employment regulation, and changing immigration regimes. Kallenberg has argued that the decline of the SER indicates the decline of “good jobs” relative to “bad jobs.” The prevalence of “bad jobs” in the US labor market is associated with changing labor market regulation, which in turn relates to the changing balance of employer and employee power. These trends are uneven; research has highlighted polarization in incomes and access of occupational benefits, but less of a direct shift to insecure job tenure. These outcomes, however, need to be put in the context of an already highly flexible labor market. In parts of Europe, where employment protections (so-called labor market rigidities) are stronger, the proportion of temporary workers is higher but so are the regulations protecting core workers.

In a more generalized sense, the norms underpinning both employment policies and protections, and related welfare models, are shifting to what Dyer, McDowell, and Batnitzky (2011) call the “adult worker” model. This means that the insecurity associated with precarious work overlaps with changing norms that cast both men and women, regardless of their caring responsibilities, as labor market participants. Changes to legal and regulatory frameworks governing employment thus emphasize procedural equality (equality of opportunity), and greater flexibility, while simultaneously rolling back substantive protections in the areas of job security, health and safety, and right to collective action. Research in Canada, for example, has shown how employment standard regimes have shifted toward complaint driven models with little proactive enforcement (Gellatly et al. 2011). Workfare regimes increasingly require people to engage in the labor market in return for welfare benefits, at the same time that the spread of precarious work means that paid employment provides less security and social mobility than before. For some workers, the discourse of work–life balance has produced advances in the ability to (at least request) flexible working arrangements, but such benefits are often available only to higher paid and higher skilled workers, while low-paid workers may
precarious work

struggle with too much flexibility, as in the case of zero hours contracts in the United Kingdom.

Finally, precarious work is increasingly associated with changing migration dynamics in many labor markets. Bridget Anderson, a migration scholar, has noted the ways in which governments use immigration controls like a tap to regulate the flow of labor and control the type of worker that can enter the local labor “pool” – and under what conditions. Points-based systems pioneered in countries like Canada are increasingly being adopted elsewhere to establish channels of entry for skilled workers. At the same time, temporary migration (guest worker) programs often attach highly restrictive conditions to “low-skilled” workers who are granted visas to work in areas such as domestic service, agriculture, hospitality, and food services. As mentioned above, migration makes “work–life balance” possible for many in the Global North, but migrants and racialized workers are employed in the most flexible and low-paid jobs of the postindustrial economy. Precarious employment conditions thus create a fragile work–life balance for many migrants, especially those who leave families behind to migrate with no route to settlement and family reunion in the host country.

Scholars have thus coined the term “precarious migrant status” to describe the condition of migrants whose precariousness is constructed in and through the legal and political regimes that structure their entry, and continued presence, in “host” countries. They argue that, in relation to migration status, dichotomous approaches (legal–illegal) do not capture variable forms of illegality and irregularity, therefore precarious status helps draw attention to blurred boundaries between citizenship and illegality where the policy context (alliances between the state and capital) create pathways to precariousness (Goldring, Berinstein, and Bernhard 2009).

Precarious migrant status again signals the ways in which precarious work is mediated by institutional contexts beyond the labor market and economy proper.

Much of the research on precarious work, tied to the concept of the SER, has focused on the Global North. It remains an open question the extent to which the concept can be profitably deployed in contexts where informalit is the norm. In countries like India, for example, the vast majority of the population has never had access to a bundle of employment rights and conditions resembling the standard employment relationship. Yet at the same time, processes of casualization, outsourcing, and subcontracting in those parts of the public and private sectors where formal employment had been institutionalized do suggest parallels with forms of labor market restructuring in the advanced industrial economies. Both local and transnational firms and factions of capital are identified with these processes, as in the Global North. In China, the destruction of the “iron rice bowl” – a set of rights to employment and occupational welfare benefits associated with urban, mostly industrial employment in state enterprises – has created new groups of precarious workers. Many workers in formerly state-owned enterprises are now subcontracted by labor dispatch companies. Internal migrants from rural areas are, because of the hukou system of household registration, subject to precarious employment conditions that are compounded by a lack of access (and rights) to services like housing and healthcare.

Similarities and differences between labor market realities in different places have left theorists searching for ways of linking up processes that share common characteristics but play out in a variety of ways in different places and at different scales. European policymakers have sought to extend Scandinavian and Dutch models of “flexi-curity” to create social safety
nets (encompassing unemployment benefits, but also robust training and skills upgrading programs) to facilitate labor market flexibility and mobility in ways that do not create social and economic insecurity. Applied outside the relatively generous welfare states in which they were first designed, however, flexi-curity policies have been critiqued for emphasizing flexibility over security. At the other end of the spectrum Guy Standing, formerly of the International Labour Organisation (ILO), has coined the term precariat to describe a “new class” of workers defined not by traditional class relations but by their shared experience of precarious employment. Standing’s concept of the precariat is broad enough to encompass highly skilled and well-paid workers on temporary contracts in the creative industries, migrant workers doing the “dirty, dangerous, and degrading” labor at the bottom end of the labor market in advanced industrial economies, and workers in the informal sector in the Global South (from rag pickers in India to petty traders in Lagos to migrant construction workers in Dubai). This opens Standing up to the critique that describing the precariat as a class ignores the very limited points of intersection of the positions and experiences of such diverse workers – or, at the very least, downplays the differences in labor market power, income, and social and cultural context among them. Academics remain divided on its utility and the concept has been little discussed in geography despite garnering widespread attention in other disciplines and beyond academia.

Precarity and precarious life

It is in relation to geographies of migration, and political geography, that the concepts of precarious work come into conversation with concepts of precarity and precarious life. Drawing heavily on the work of the feminist philosopher and gender theorist Judith Butler, geographers have utilized her articulation of a social ontology of precarity to explore how bodies and communities experience precariousness across a range of dimensions. Butler’s notion of precarity draws attention to the forms of mutual dependency that are the condition of human life. In Butler’s conceptualization we all depend on each other in various ways, but precarity is also unequally distributed: certain populations are rendered precarious through the allocation of social, economic, and political risks. Waite, for example, examined political, social, and spatial dimensions of precarity in the lives of asylum seekers and refugees in the United Kingdom for whom precarious work (asylum seekers are prevented from seeking formal paid work) interlocks with generalized precarity (Waite 2009). Asylum seekers in particular endure “hyper-precarious” conditions because of their precarious legal status and vulnerability to precarious, and sometimes unfree, conditions of work in their desperation to secure the conditions of everyday life.

Another theoretical influence on geographers theorizing precarity is Michel Foucault. Nancy Ettlinger (2007) argued for an expansive conception across time and space, using Foucault’s concepts of biopower and governmentality to examine how the US government’s attempts to create illusions of security and certainty after the attacks of September 11, 2001, instead recreated and concentrated precarity in certain populations. This approach has resonances with other work in political geography on vulnerability and insecurity, including at the scale of the body. Feminist geopolitics have long argued for understandings of the body as a gendered, racialized, and classed site through which norms of nationhood, citizenship, and exclusion are rendered visible. The body is likewise the field through
which imperatives of capitalist accumulation in the organization of the labor process are materialized and labor control exercised.

The concepts of precarious work and precarity thus resonate, but are also informed by quite different intellectual and political traditions, especially in the Anglo-American world. In geography, in recent work on migrants, asylum seekers, and refugees, political economic, and political philosophical traditions are being explored in ways that amplify those resonances, exploring precariousness across economic, sociocultural, and political dimensions. These approaches are in keeping with geographical traditions of social constructivism, for example, in understandings of scale and of labor markets, but also with poststructuralist-influenced approaches to ontologies of vulnerability and insecurity. At the same time, political economic understandings of precarious work remain important in relation to epistemologies of labor market change, new geographies of paid and unpaid work, and economic and social polarization in increasingly unequal societies.

SEE ALSO: Flexible labor markets; Fordism/post-Fordism; Gender, work, and employment; Intermediaries; Labor geography; Migrant labor; Race, work, and employment

References


Further readings

Precipitation describes hydrometeors that fall from the Earth’s atmosphere to its surface (American Meteorological Society 2012). Rain, snow, sleet, glaze, and hail are all common forms of precipitation. Often the general term is used because it can be difficult differentiating between the liquid and solid phases of water when making forecasts or remote sensing observations. The field of “precipitation research” can be divided into four main themes: precipitation physics, quantitative precipitation estimation (QPE), quantitative precipitation forecasting (QPF), and the climatology of precipitation.

Precipitation physics

Not all clouds form precipitation, which is not surprising considering millions of cloud droplets must somehow coalesce to form a raindrop. For clouds that are entirely above freezing (>0°C), known as warm clouds, precipitation formation occurs through the collision-coalescence process. Large cloud droplets, having a greater terminal fall speed, collect small cloud droplets much like a bulldozer. The efficiency of this process varies. Clouds with a diversity of droplet radii and strong vertical winds will be more likely to form precipitation. However, collision is also limited by small droplets being repelled by the airstream around large droplets. Another process aids collision-coalescence in clouds that contain ice crystals (cold clouds). The “ice crystal theory” or “Bergeron theory” was proposed in a series of papers by Wegener (1911), Bergeron (1935), and Findeisen (1938). It states that when ice crystals and supercooled (liquid below freezing) cloud droplets coexist, the deposition phase change will overwhelm condensation and the ice crystals will grow at the expense of the liquid droplets. However, this vapor deposition mechanism depends on the humidity of the atmosphere, and in many cases both ice crystals and supercooled water droplets will grow or shrink simultaneously (Korolev 2007).

Further growth is achieved through accretion, where supercooled droplets freeze onto ice crystals to form graupel (small snow pellets), or aggregation, where ice crystals clump together to form large snowflakes. These larger hydrometeors precipitate but will often melt into raindrops, where the collision-coalescence process takes over. The melting at the 0°C isotherm in the atmosphere is clearly evidenced by large reflectances within vertical profiling radar images, the so-called bright band.

In a cold cloud, the type of precipitation that eventually reaches the ground depends largely on the vertical temperature profile. Snow falls if the atmosphere is below freezing at all levels. All other precipitation types go through the melting process. If the atmosphere is sufficiently warm rain falls. However, sometimes graupel will circulate within a thunderstorm cloud sustained by strong updrafts. The graupel will grow when it is below freezing, as supercooled droplets freeze on the exterior, or when it is above freezing, when the pores of the ice stone soak up liquid water like a sponge. When the hydrometeor is sufficiently heavy or enters a region of downdrafts, it will...
fall as hail. Detailed physics on hailstone growth can be found elsewhere (List 2014a, 2014b).

Near the surface, raindrops will freeze and become ice (sleet) if there is a deep enough layer of below-freezing temperatures. Glaze or freezing rain requires a strong temperature change close to the surface, with temperatures dropping quickly to well below freezing. In this case rain will become supercooled and freeze upon impact with the ground. Sleet and glaze can lead to treacherous driving conditions and the riming of ice can down trees and power lines. These societal hazards motivate the forecasting of winter precipitation types. However, predicting the temperature profile of the lower troposphere remains a significant challenge.

**Quantitative precipitation estimation**

Precipitation is easy to measure; the earliest records can be traced back to India in the fourth century BCE (Strangeways 2010). All that is necessary is a container to collect precipitation. Rain gauges have become more sophisticated over time, including the popular automated tipping bucket gauge. Rain or melted snow enters a collection funnel and is directed into a seesaw apparatus. A cup at one end of the seesaw is filled (usually 0.01 inch or 0.25 mm) and then tips under the weight of the water, pivoting the second cup upward to collect more precipitation. Each “tip” is automatically recorded. Other gauges measure the size distribution and fall speed of precipitation (called distrometers); they are useful for understanding the microphysical processes involved and can distinguish among different types of precipitation.

Paradoxically, even though precipitation is easy to measure, it is very difficult to adequately quantify the total amount falling from any particular cloud or storm. This is due to precipitation being highly variable in time and space. Precipitation collected in one gauge is not necessarily representative of neighboring areas and depends on storm motion and size, and moisture flux (Hendrick and Comer 1970). This issue has implications in hydrology and has led to concerted efforts in deploying new rain gauges (e.g., The Community Collaborative Rain, Hail, and Snow Network), compiling existing rain gauge information (e.g., The Global Precipitation Climatology Centre), and data rescue operations. In addition, meteorologists have turned to the development of remote sensing estimates of precipitation and research in quantitative precipitation estimation (QPE).

Radar is one means for estimating precipitation. Microwave pulses are emitted from a weather radar. When a radiation pulse intercepts a hydrometeor, it reflects back to the radar dish. After correcting for range and calibrating the equipment, the strength of the reflectance value (Z) can be converted to the intensity of rainfall (R). Very large reflectances are associated with hail, but otherwise scientists have developed empirical “Z–R” curves to estimate rain rates. The Z–R curves continue to be refined by relating echoes to physical measurements of rainfall from gauges.

Satellites can also infer precipitation at the ground by sensing radiation coming from the atmosphere. Infrared radiation or “thermal energy” is detected from cloud tops. The lower the infrared temperature, the higher the cloud top must be in the atmosphere. Studies have shown that thick clouds produce more precipitation than low, thin clouds, and that the infrared technique is most appropriate in the tropics. Microwave energy is more directly related to precipitation as low-frequency microwave channels are emitted from liquid hydrometeors and high-frequency channels are scattered by solid hydrometeors within the cloud. The first
NASA mission dedicated to the estimation of precipitation from space was the tropical rainfall measuring mission (TRMM). The core satellite was launched in 1997 and included microwave sounders and the first space-based radar platform. The active microwave instrument greatly advanced QPE and paved the way for the follow-on global precipitation measurement (GPM) mission. The core satellite was launched on February 27, 2014, orbiting at a more inclined angle relative to TRMM, and thus covering the mid to high latitudes. Advances in instrumentation will allow for the retrieval of light rain and snow. The most exciting advances in QPE research are those that take advantage of all available estimates of rainfall – gauges, radar, and satellites – to produce accurate and timely global precipitation datasets.

Quantitative precipitation forecasting

One of the biggest forecasting challenges is correctly predicting precipitation amount. Therefore, in concert with QPE, quantitative precipitation forecasting (QPF) has evolved as an important research agenda item. Currently, precipitation forecasts are given as a percent chance of rain, or the probability that 0.01 inches (0.25 mm) of rain will fall somewhere in the forecast area during the period of forecast. Numerical models produce grid-scale precipitation as a function of hydrometeor size distributions over the cell. Subgrid-scale precipitation is parameterized through a number of convective schemes. In the United States, QPF is performed operationally by the Weather Prediction Center (WPC), formerly the Hydrometeorological Prediction Center. In general, QPF starts with an ensemble of model output and then determines which models are producing reasonable quantities of precipitation in time and space. Estimates are then extended into the future, after being adjusted manually by forecaster experience. QPF skill decreases with lead time, and is generally not performed past five days, but WPC verification shows greater skill compared to individual model runs.

Precipitation climatology

The distribution of plants, animals, and people on this planet (its habitability) is strongly dependent on amounts and extremes of precipitation. Precipitation is one component of the overall hydrologic cycle – comprised of water evaporating from oceans, rivers, and lakes, condensing into clouds, and then returning to the surface as precipitation. The exchange of precipitation in this cycle has recently been estimated at $486,000 \text{ km}^3 \text{ year}^{-1}$, with 23% falling over land and 77% over the oceans (Trenberth et al. 2007). Of course, the distribution of precipitation on the planet varies greatly (Figure 1). The wettest place on Earth according to the Guinness Book of World Records is Mawsynram, India, which receives an average of nearly 40 feet (12 m) of rain in a year, while some areas of Antarctica have not seen rain (or snow) in two million years.

The history of geography includes the classification of Earth’s climates based partly on precipitation. In fact, in most of the tropics the annual cycle is marked by variations in precipitation, and not temperature. To understand Earth’s precipitation climatology it is important to recognize that rainfall is primarily dependent upon two things: (i) a humidity source and (ii) uplift. The presence of moisture is insufficient if the atmosphere is stable or the air is subsiding. Some of the driest regions on Earth are over the oceans (Figure 1). Air rises because of orography, surface convergence, frontal boundaries, and convection. Some of the wettest (driest) areas in
the world are on the windward (leeward) sides of mountains. The intertropical convergence zone (ITCZ) is where the trade winds converge and coincide with copious clouds and precipitation (Figure 1). The South Atlantic Convergence Zone and the South Pacific Convergence Zone act in a similar way and are easily recognizable in global maps of rainfall (Figure 1). The east coasts of Asia and North America receive consistent amounts of rainfall throughout the year due to the baroclinic zone that naturally occurs between land and ocean. Extratropical storms generate and intensify in these coastal environments, and the tracks of storms across the North Pacific and Atlantic and encircling Antarctica can be identified in satellite-based precipitation climatologies (Figure 1). Finally, convection is a type of energy transfer in the atmosphere, where air rises from a warm surface. The Earth’s dominant convection centers are over the tropical landmasses of South America, Africa, and the Indonesian archipelago (Figure 1). Each of these areas receives abundant precipitation, thus supporting the biologically diverse rainforest ecosystem. However, convection can also lead to severe thunderstorms, hurricanes, and extreme rainfall. It was found in a recent study that the most intense thunderstorms are not located within the climatological convection centers or oceanic convergence zones but in semi-arid regions of the world such as northern Argentina and the south central United States (Zipser et al. 2006). The combination of insolation and a nearby water source lead to strong convective updrafts.

Earth’s precipitation climatology cannot be discussed without mentioning the monsoon. While “monsoon” is technically defined as a reversal of planetary-scale winds, monsoons are associated with copious amounts of rainfall. In particular, the southwest monsoon, when the
wind blows from the Indian Ocean into the Indian subcontinent, brings rains from May to October and is the main reason for the rainfall record in Mawsynram, mentioned earlier. The opposite northeast monsoon brings rains from November to April to portions of the southern hemisphere, such as northern Australia. Other areas of the world also experience monsoonal rains but to a lesser extent, such as West Africa and southwestern North America.

It should be noted that Earth’s precipitation climatology is not static, but can undergo variations on many timescales. Changes in expected rainfall can lead to societal disruption, with an excess of rain resulting in floods and too little rain leading to drought. So it is not surprising then that the rise and fall of civilizations have been linked to these extremes. People have attempted to mitigate the drought disaster by trying to control the weather, and make it rain. While history is teeming with stories of medicine men performing rain dances, modern techniques have primarily focused on cloud seeding, where aerosol particles, like silver iodide, are released at the top of clouds to initiate the growth of ice crystals and thus kick start the Bergeron precipitation process. However, cloud seeding has met with limited success and, thus, is not routinely practiced.

One of the biggest contributors in disrupting global precipitation patterns is the El Niño Southern Oscillation (ENSO). The periodic warming of the eastern equatorial Pacific disrupts the convection cell (also known as the Walker circulation) that would normally bring about abundant precipitation in the western Pacific (tied to the Indonesian convection center) and a deficit of precipitation in the eastern Pacific. During El Niño the most intense precipitation shifts eastward across the basin causing the western Pacific to dry out. The new distribution of rainfall affects the amount of latent heating in the atmosphere, which subsequently has a ripple effect on the rest of the world by altering the major pressure centers and jet streams. Impacts of the ENSO on global precipitation are reviewed in Curtis (2008). Other climate phenomena, such as the North Atlantic Oscillation, can cause changes in climatological precipitation as well, but to a lesser extent than the ENSO and usually only over limited regions and seasons.

Long-term trends in precipitation are not as readily apparent as interannual variations or the anthropogenic global warming signal. The Clausius–Clapeyron relationship directly relates temperature with atmospheric moisture. Warming of 1°C equates to an approximately 7% increase in total water vapor (Allen and Ingram 2002). However, global climate models also predict that the convection cell in the equatorial zone (i.e., Walker circulation) will also weaken. Since, as stated earlier, rainfall is dependent on humidity and uplift, the net result is little to no change in global precipitation. Most observations from the satellite record have corroborated this theory (Adler et al. 2008), although some would suggest that global precipitation is increasing and the hydrologic cycle is accelerating (Wentz et al. 2007). The contradictory findings, lack of long-term reliable datasets, and the small signal-to-noise ratio, all lead the scientific community to having a low-to-medium confidence in global trends in precipitation.

Observations have shown, and climate models project, an increase in precipitation extremes, due to the availability of moisture. If, in general, the frequency of storms decreases due to a background environment that inhibits vertical motion, storms that do spawn will be laden with water vapor and produce heavy precipitation – often coined as the “more bang for your buck” theory. Regarding global patterns of precipitation, climate models also suggest that wet regions (e.g., the ITCZ) will become wetter

**PRECIPI TATION**
and dry regions (e.g., subtropical desert belts) will become drier with the warming (“the rich get richer” theory). This is likely due to an expansion and intensification of the Hadley circulation cell (Held and Soden 2006). Finally, it should be noted that other changes in the atmosphere caused by humans, such as the addition of aerosols, can have equally strong, and sometimes contradictory, impacts on precipitation when compared to global warming.

SEE ALSO: Atmospheric/general circulation; Climatology; Clouds; Droughts and water shortages; Global climate change; Global climate models; Hydroclimatology and hydrometeorology; Hydrologic cycle; Water and climate change

References


Further reading

Pre-Quaternary landforms and landscapes

C.R. Twidale
University of Adelaide, Australia

Convention

Common sense suggests that land surfaces exposed to the elements surely must be worn away. The concept finds support in the sediments eroded and transported by rivers, glaciers, and waves and on the wind, and in measurements of contemporary erosion. With the exception of exhumed forms, most landscapes are perceived as of Quaternary age. True, some may date from the Pliocene but continuous chemical weathering arguably has ensured that few older features could have survived. And enormous tracts of the Earth’s land surfaces, including the mid-latitude deserts and continental areas of the Northern Hemisphere that were occupied by ice sheets or glaciers during the Pleistocene, manifestly are youthful. Similarly, the many fluvial plains, some local features but others of regional extent, like the erosional and depositional plains of the Lake Eyre catchment of central Australia, which occupies about 22% of the continent, are of Quaternary age.

Yet the basic argument propounded by Hutton, Lyell, and Davis, that weathering and erosion continually and rapidly destroy and renew land surfaces, is flawed. This is demonstrated by the existence of pre-Quaternary landscape elements not only in uplands but also in the midst of youthful plains. In northwest Queensland, Australia, for example, laterite-capped plateaus of Miocene age survive on the divides between major river valleys eroded in Cretaceous strata.

Moreover, exhumed landforms – some of them of great antiquity – are widely distributed in both space and time. The age of the resurrected surface is bracketed by the age of the youngest rock exposed in the buried surface and the oldest of the cover material, but it most likely immediately predates the cover.

Paleoforms of epigene-etch origin

In addition to such relics of past land surfaces, very old forms that have never been protected by burial also persist in the contemporary landscape. They are overwhelmingly of etch type. They are two-stage forms that have been shaped at the weathering front or lower limit of discernible alteration, and then exposed as landforms by the stripping of the regolith or weathered mantle. They are not untouched by weathering and erosion, but their essential characteristics survive.

Thus, some 70 million years ago, what is now the steep-sided inselberg known as Uluru was a low soil-covered rise located in the Amadeus Basin of central Australia. Differential weathering by moisture retained in the regolith caused the weathering front to become dimpled and grooved. As the bed of Lake Eyre subsided, the rivers flowing to it, including those draining the Amadeus Basin, incised their valleys. The regolith on the rise was stripped away...
and the rocks around it were further weathered and eroded by the waters flowing from higher ground. A siliceous duricrust was widely developed on the lowered plains during the Eocene (Figure 1).

Once standing in relief, the hill that was to become Uluru continued to shed water and was little changed. By contrast, the surrounding plains were lowered and the bounding slopes of the residual were steepened to produce the present well-known monolith rising abruptly from the desert plains. The dimpled and grooved bedrock crest of the inselberg came to stand high in the local landscape and was, as such, a dry site. It has stood basically unchanged throughout the Cenozoic.

Figure 1  Dissected plateau capped by Eocene silcrete, southwest Queensland.

Examples of very old Australian paleoforms

No direct method of determining the age of pre-Quaternary land surfaces has yet been discovered. Physical dating procedures have been applied in some areas but with confusing and in some instances widely different results that are moreover at variance with dates produced by a consideration of stratigraphic/topographic criteria.

Plateau remnants are typical of central and northern Australia. Those capped by laterite stratigraphically determined to be of Miocene or Oligocene–Miocene age occur throughout the monsoonal tropics of Australia, from north
Queensland in the east to the Carnarvon Basin of Western Australia. In the lower Lake Eyre topographic basin plateaus associated with Eocene silcrete are widely developed and conserved (Figure 1), but in southwest Queensland the silcrete has been folded, giving rise to cuestas that serve as a reminder that old surfaces survive despite virtually continuous and continuing minor tectonic disruption.

Using stratigraphic data – later confirmed by numerical dating – Hills (1955a, 1955b) showed that summit bevels developed in Paleozoic strata and preserved in the ranges of eastern Victoria are separated by valleys floored by lava flows, some of them of Eocene age. The lavas postdate the valleys, which are younger than the summit surface into which they are incised. Thus, the lavas provide a minimum age for the summit bevels. Similar features preserved in southeastern New South Wales were initially ascribed to the Pliocene but dated valley lavas later showed them to be older and comparable to their Victorian counterparts. Later regional mapping suggests that throughout southeastern Australia the higher bevels may be as old as Triassic, with the Permian glaciation a useful back-marker for landscape development here and in other parts of the former Gondwana.

The method of dating articulated by Hills (1955a, 1955b) is based in topographic relationships: in an undisturbed landscape the older surfaces persist high in the relief, whereas younger forms are found lower in the topography and the most recent are in valley floors. This relationship is illustrated by the stepped morphology characteristic of the many inselbergs that have been shaped during periods of subsurface weathering and intervening phases of erosion and lowering of the surrounding plains. Such episodic exposure caused the residuals to become more and more prominent features of the local topography. The crests of some of the inselbergs are of considerable antiquity.

In southeastern Australia, lavas provide numerical age control, but elsewhere, dated sediments and duricrusts serve the same purpose. The southern Flinders Ranges in South Australia is a region of ridge and valley. During the Middle Eocene the pre-existing westerly drainage of the intermontane northern Willochra Basin was blocked, forming a lake. Silicified lacustrine sediments extend up adjacent valleys. Thus, the ridge-and-valley assemblage was in existence some 60 million years ago. The beveled crests of the quartzite/sandstone ridges probably date from the Cretaceous and may be relics of a planate surface shaped by rivers graded to an Early Cretaceous marine shoreline. The stratigraphy of Triassic intermontane basin deposits hints at regional planation during that Period, but no landforms demonstrably associated with that event have been located.

To the west of the Flinders Ranges the prominent summit surface of the Arcoona Plateau transects various gently dipping Proterozoic quartzites. Early Cretaceous marine beds (and Eocene silcrete) are preserved in valleys cut into the surface, which therefore predates the marine transgression.

In the arid interior of Australia, the beveled bornhardts of the Everard Ranges apparently stood as islands in the Jurassic seas, for in eastern sectors of the upland tongues of Late Jurassic sediments extend up the intervening valleys. Similarly, the quartzitic Kakadu massif of Arnhem Land is bordered by marine beds and shorelines of Early Cretaceous age, showing that 120–130 million years ago it stood as an island in the then seas.

Dissection of the east–west trending Macdonnell Ranges developed in Proterozoic strata and folded in the later Paleozoic has produced a ridge-and-valley assemblage that existed prior
to the Eocene, as indicated by silcrete-capped mesas in valley floors and the southern piedmont. Remnants of a much higher and hence older bevel are evident in the Ranges and have tentatively been construed as having been shaped by streams graded to a Cretaceous shoreline.

Though some of the laterites of the southwest of Western Australia are of Late Cenozoic age, others appear to be older. Jutson noted that the so-called New Plateau or high plain of the Yilgarn Craton of Western Australia is an etch surface resulting from the dissection and partial stripping of the lateritic duricrust that forms and underlies the Old Plateau (Figure 2). The topographically lower New Plateau is an etch surface eroded by rivers and streams initiated during the Eocene and still extending inland. This evidence indicates a pre-Eocene or Mesozoic age for the lateritized Old Plateau. The regional setting also suggests a pre-Eocene age for the laterite but later Cenozoic paleomagnetic ages complicate interpretation.

In any event, stepped residuals protrude above the lateritic surface. Their beveled crests are even older than the duricrusted high plain. The Humps and other stepped forms display a dome-on-dome morphology, some with flared sidewalls associated with former piedmonts and, possibly, planate surfaces that predate the Old Plateau (Figure 3).

The laterite that forms plateaus in southern Eyre Peninsula, Kangaroo Island, and southern Mt Lofty Ranges – the Gulfs region of South Australia – appears to predate the Middle Eocene, as indicated by lateritic debris recorded from the basal Middle Eocene beds in one of the fault-angle depressions. Stratigraphic evidence from Kangaroo Island points to an even earlier
Triassic age. This interpretation is based on potassium–argon (K–Ar) dating of basalt that appears to overlie the laterite, but other dating methods (particularly paleomagnetism) indicate a Miocene or even Pliocene age for the duricrust, the former consistent with the high Miocene sea level noted in the area. However, whether the laterite developed on a pre-existing surface or planation and pedogenesis were concurrent remains to be determined. Suffice it to state that regional stratigraphic and topographic evidence favors a Mesozoic age for the duricrust. For example, the early Tertiary sediments deposited in southern Eyre Peninsula are not lateritized as they ought to be, had the intense humid tropical alteration occurred in the later Tertiary as suggested by paleomagnetic analysis.

In some instances, fossiliferous deposits provide evidence of the age of landscape of the source area from which the sediments were derived. Thus, the summit bevel preserved on the bornhardt massif of the Gawler Ranges is an etch surface. During the earlier Mesozoic the area was reduced to a planate surface underlain by a regolith of variable thickness, thick over fracture-controlled bedrock valleys, but thin over bedrock highs. Uplift of the southern margin of what is now the upland caused rejuvenation of streams flowing north to the Cretaceous seas that occupied what is now the Eromanga Basin. The fluvial deposits, including cobbles of Gawler Range Volcanics – corestones derived from the previously developed regolith that covered the whole of the Ranges – were laid down as the Mt Anna Sandstone. Thickness variations and cross-bedding also indicate a southerly provenance for the formation. It was interbedded with fossiliferous marine beds of the Cadna-Owie Formation that show the Sandstone to be of Early Cretaceous (Neocomian–Aptian) age. As
deposition implies erosion, this age determination applies also to the exposed bornhardt landscape. Silcrete of probable Eocene age preserved in the piedmont and within the upland suggests that, apart from marginal valley floor dissection, no later comparable stripping of the upland is in evidence.

**Paleosurfaces in other parts of Pangea**

Very old landscape elements are not confined either to Australia or to the oldlands of cratonic regions. Thus, and like the Kakadu massif, the various uplands of the Roraima Plateau of northern South America may have stood as islands in the Cretaceous ocean. The beveled upper surface must be considerably older. The uplands defined by the Great Escarpment of southern Africa also are bordered by coastal and offshore plains underlain by Cretaceous and younger sediments. In the interior, very old Gondwanan plateau remnants that transect Jurassic basalt flows and Late Cretaceous kimberlite pipes survive on uplands such as the Drakensberg (Figure 4).

But, by contrast with the subsiding crust and interior drainage of central Australia, southern Africa has experienced recurrent uplifts that induced repeated phases of stream rejuvenation and landscape revival. Older and older surfaces separated by more-or-less prominent scarps form concentric zones, decreasing in age-range from the interior toward the coast.

In Zimbabwe, for instance, extensive plains of Pliocene age occupy the lower reaches of major catchments, such as those of the Limpopo and the Zambezi rivers. They are succeeded inland by the regionally dominant post-African Miocene plains, which in turn give way to the landscapes of African, that is, mid-Cretaceous to Oligocene, derivation. This stepped landscape is protected by its interior location but each surface is being reduced as the younger scarps and surfaces migrate inland. Stepped landscapes with bauxitic and ferruginous cappings of Jurassic, Cretaceous, and various Tertiary ages are prominent in several parts of West Africa (Niger, Cameroon), and to the east in Kenya, Uganda, and Tanzania.

In Britain, and citing stratigraphic indicators, Jones speculated that the summit surface preserved in the uplands of central Wales is of Triassic age. The upland bevel preserved on monzonite in the western Sierra Nevada of California has been linked to Cretaceous age.
deposition in the adjacent Great Valley. But correlations are not everywhere simple and obvious: four prominent planation surfaces have long been recognized in the Appalachian Upland of eastern North America but only three phases of rapid weathering and planation (earlier Jurassic, mid-early Cretaceous, and Middle Miocene) are identified in the record preserved in the strata beneath the adjacent continental shelf. Double planation associated with etching may account for this discrepancy. Alternatively, it may further confuse the issue.

**Reality**

Common sense and field observations point to the ever-changing youthfulness of the Earth’s land surfaces. However, several very old surfaces have been recognized based on a range of evidence and argument based in topography and stratigraphy. Ages derived from physical procedures and at variance with these results are cause for concern and call for explanation. Yet stratigraphy cannot be denied. As to survival, as they exist, the very old paleosurfaces cited here must be possible. The structural factor is vital to concatenation: structural contrasts have led to the development of an unequal distribution of weathering and erosion that has in turn induced positive feedback, episodic exposure, and stepped forms and terrains. This complex of factors may, in part at least, account for the persistence of landscape elements that are relics of the former Pangea.

**SEE ALSO:** Landforms of other planets; Landforms and physiography; Quaternary geomorphology and landscapes; Tectonic geomorphology; Volcanic processes and landforms

**Further reading**


Definitions and conceptual traditions: geographies of personal privacy

Personal privacy is multifaceted, not amenable to simple definition, and subject to diverse and wide-ranging categorization across the social sciences. Dictionary definitions focus on being alone or free from public attention. Recent papers have identified three persuasive taxonomies, outlined in Table 1. The majority of these definitions cluster around a person’s ability to be alone, to control information about oneself, and to freely behave in a manner consistent with religious, sexual, and so forth, practices and beliefs. Privacy in this sense has strong links to individual freedoms, and links between personal privacy and rights of the person and individual are hence emphasized in some disciplines.

Concern with personal data, personal communications, and autonomy of association, beliefs, and practices is writ large in geographical thought. This is especially the case methodologically. The methodological implications of personal privacy, especially as they relate to research ethics, have been one focus of geographical research on personal privacy (Dowling 2005). The collection of information by researchers (or other agencies) can involve “invading” someone’s privacy. Geographical research may entail recording information considered private, or observing interactions in people’s homes that are customarily considered private. Geographical research, and university
Table 1  Categorizing forms of personal privacy.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Left alone</td>
<td>Privacy of the person (bodily privacy)</td>
<td>Privacy of the person</td>
</tr>
<tr>
<td>Secrecy (conceal facts about oneself)</td>
<td>Privacy of personal behavior/practices</td>
<td>Privacy of behavior and action</td>
</tr>
<tr>
<td>Intimacy</td>
<td>Privacy of personal communication</td>
<td>Privacy of personal communication</td>
</tr>
<tr>
<td>Control information about oneself</td>
<td>Privacy of personal data</td>
<td>Privacy of data and image</td>
</tr>
<tr>
<td>Restricted access</td>
<td></td>
<td>Privacy of thoughts and feelings</td>
</tr>
<tr>
<td>Mixture of above</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Human ethics boards, thus develop and follow guidelines about not releasing such private details into the public domain. This includes protocols about the storage of data and ways to ensure that research participants are not identified. The rise of the Internet has transformed geographical research, opening up a plethora of new sources of data such as social media, new ways of collecting qualitative data like online surveys, and new ways of recruiting research participants. These methods and sources pose myriad challenges to research ethics, especially to informed consent, privacy, and confidentiality. Using online questionnaires brings confidentiality to the fore. Online surveys are often conducted using commercial software and web-hosting services, which means that the researcher doesn’t have complete control over protecting participants’ identities. The web service, for example, may track IP (Internet Protocol) addresses and hence enable the identification of participants. Using the various elements of Web 2.0 as sources of qualitative data is even more contentious in terms of privacy and informed consent.

Geographical scholarship on personal privacy, however, tends not to produce or work with taxonomies such as those contained in Table 1. Geographers accept the existence of multiple dimensions of personal privacy and acknowledge that they are complexly interrelated. As such, the multiple definitions of personal privacy in Table 1 infuse rather than direct geographical engagement with the concept. Though it recognizes that privacy is a phenomenon experienced by an individual, geographical analysis does not focus solely on the individual or on individual experiences. Rather, personal privacy is understood as a social construct: the ability, characteristics, and places of personal privacy are determined by the social and political context. Geographical attention is hence more regularly turned to the production of personal privacy: the conditions under which it is enabled and experienced. More specifically, understanding personal privacy geographically has focused on whether, how, and where personal privacy is experienced by an individual, and is influenced by power relations, social positioning, and discursive framings.

Taking power relations first: personal privacy is strongly influenced by a powerfully reproduced dichotomy of public and private. In this dichotomy (see Blunt and Dowling 2006) the private is defined in opposition to the public: personal privacy is created and experienced in retreat from, and separation of, the public sphere – of work, politics, and the street, to name just three examples. Privacy here takes on a particular meaning because of its political-economic context. And it is used to define and discipline
populations. Feminists have long argued, for instance, that privacy, and in particular the perceived sanctity of home, have been one means through which the oppression of women has been reproduced. Privacy as a social construct also entails attention to sociopolitical framings. The meaning and shape of personal privacy is produced through discourses of the state, of economy. In legal geography, for example, the ways in which laws and regulations enable, constrain, and spatialize personal privacy are highlighted.

A second point, which also draws from feminist analyses, is that personal privacy is socially differentiated. For example, the ability to keep information about oneself secret is dependent on the degree of autonomy that an individual has. Likewise, being able to claim a space for privacy is highly dependent on access to physical, psychological, and financial resources. Even within private spheres like the home, women’s ability to claim independence can be curtailed by patriarchal relations that involve surveillance by men. The question for geographers is typically to examine the ways in which social and geographical contexts circumscribe and shape experiences of privacy.

A final point is a discursive and cultural analysis of privacy as a social construct. This means that definitions of what privacy means are historically and geographically specific. In many Western societies, for example, privacy is thought of as a virtue, something to aspire to and defend. Indeed, some legal scholars speak of a “right” to privacy. Yet, in other cultural or historical contexts, the propensity to be alone and the ability to control information about oneself are not highly valued in comparison with shared knowledge and understanding. For example, in communities in the West developed with the intention that they be shared – with communal housing and activities like cooking and child care collectively rather than individually undertaken – different expectations and characteristics of personal privacy are expected and enacted.

Human geographers translate this understanding of personal privacy as a dynamic and context-specific experience into a concern with the relations of personal privacy and place. As Somerville (2009, 532) observes, privacy requires “the possession of a certain territory with the power to exclude persons from the territory and prohibit surveillance by others.” Securing personal privacy is a geographical process. In the process of seeking to be, and being, undisturbed and free from public attention, geographical barriers are assumed, made, and remade. Three geographical implications of personal privacy are commonly excavated. The first is that there is a rightful place for privacy, that certain places are constructed as obvious places of privacy. Here, geographers examine how such places come to be sites of personal privacy, how they are experienced by diverse social groups, and how they are contested and remade. A second strand focuses on the experiences of personal privacy on the move, especially, but not exclusively, in public. Here, the focus is on the ways such places are created as private, and the forms of personal privacy they support. This strand directly challenges the public/privatedichotomy by considering the experiences of privacy in public spaces like town squares, streets, and public buildings. A third strand considers privacy at a distance. Here, the focus is on the ways in which privacy is being transformed in the contemporary world. This entry outlines each of these in turn.

**Personal privacy in the private sphere: home**

Privacy in geographical terms is conceived of as a spatial location, secured by the policing of
boundaries around a territory (Squires 1994). Given the strong associations between privacy and private property, home is considered the quintessential private space. Home is not the state or government, home is not work, home is not the church, home is not the realm of politics, and home does not encompass commercial activities. Rather, home is separate from and outside of the collective institutions of public life and, as such, is the primary site through which personal privacy is enacted. It is a space for intimate, confidential relationships and a space beyond the gaze of the state and other institutions. In Blomley’s (2005, 619) terms, “in our home we are, supposedly, able to shield ourselves from the views of others, whether our neighbors or the state.” Home, and the relations within, foster personal privacy. Home offers a space of freedom and control for the individual, and for intimate, caring relationships. Home as haven or sanctuary is a central idea in the privacy of home. While public space outside the home is seen as imposing and dangerous, home space is inside, enclosed, and safe. It is a sanctuary, a place to retreat into, to be oneself, away from the gaze of others, and to freely express one’s beliefs. Home also facilitates the notion of privacy as familial intimacy. The activities of caring for family members, caring for the self (washing, dressing, grooming), engaging in intimacy with others and in related “private” emotions like love and desire, occur within, and are associated with, home spaces. Indeed, the emergence and consolidation of the nuclear family as the idealized family formation coincided with the delineation of home as a private space. Home is assumed to be, and is constructed as, the natural place for personal privacy.

In geographical work on privacy and home, multiple definitions and contestations of privacy are highlighted. One important dimension is the social and geographical specificity of the personal privacy afforded through home. Feminists ask whether home as sanctuary is pertinent to the lives of many women. In the context of domestic violence and emotional abuse, for many women home is certainly not a haven or a foundation for autonomy, but a site and source of alienation and upheaval. Gendered experiences of privacy are culturally specific. In Qatar, for example, privacy at home is especially significant for women in that it “reduces or eliminates the need for covering and modest self-presentation and conduct that is required of them in public” (Sobh and Belk 2011, 130). The class specificity of home as a sanctuary is equally important. The privacy of home is similarly never completely available. It is, for example, not an opportunity for many who rely on state-provided housing. In jurisdictions like Australia and the United Kingdom, for example, state housing authorities have established standards of cleanliness and occupancy which they often enforce by entering an individual’s home without warning. Home as sanctuary is hence an ideal and a practice connected to both class and social differentiation. A final instance of complicating home as private is what privacy means for those who do not have a home. Again returning to the discursive and legal associations of personal privacy with private property, it has been shown that homeless people not only lack the privacy afforded by home, but that this lack also functions to reproduce the political marginalization associated with homelessness (Sparks 2010). Personal privacy at home is politically produced and socially differentiated.

A second theme is the porosity of public and private boundaries, both material and metaphorical. Privacy at home is at once enabled because of the symbolic and physical boundaries around the house that deter and prevent the entry of uninvited others. Yet these are also fictitious boundaries. For example, the concerns of the state and economy are writ large in home-based
experiences of personal privacy. Building regulations, home energy and appliance standards, sanitation policies, to name a few, shape, and indeed make possible, experiences of solitude. This porosity similarly applies to the privacy experienced by families within the home. Technology, and in particular television, has long been responsible for bringing the “outside” worlds of politics, war, and popular culture into the home (Silverstone 1994). Personal computers and other devices connected to the Internet challenge the solitude and independence of individual activities at home. The World Wide Web enables constant connection between home and beyond in historically unprecedented ways. Privacy at home is being transformed.

A final geographical concept is that there are multiple experiences and forms of privacy in the private sphere. The notion of home as private implies that all members of a family do not desire privacy from each other; indeed, the ideal of familial togetherness requires that family members embrace the sharing of space within the home. A geographical, dynamic conception of personal privacy leads to a questioning of the totality of home as private. Family members seek sanctuary from each other, a respite within the home. Munro and Madigan (1993) aptly term this “privacy in the private sphere.” In this construction of privacy, the ability of individual family members to secure and maintain a sense of physical or social separation from other members is equally important. Such privacy is also secured through social negotiation between inhabitants, whether through formally agreed scheduling of activities within a shared space or through informal negotiations as occasions demand. These social negotiations are arguably gendered, with men having greater freedom to achieve privacy and independence within the home than women (especially those who are mothers), who more often fit their own needs and work requirements around the activity, space, and time demands of children. Thus we see separate areas for children (rumpus rooms, play rooms) and parents (parents’ retreats), multiple bathrooms to accommodate the individual desire for privacy, and multiple electronic devices (televisions, computers) scattered throughout the home to facilitate the fulfillment of each family member’s individual aspirations. Through these, privacy in the private sphere is both sought and achieved.

Thus far this section has presented a quite conventional geographical understanding of privacy and home. There is a further complication to this description that draws on recent concerns with sensory geographies. Privacy is also an olfactory experience, as illustrated in Sobh and Belk’s (2011) analysis of homes in Qatar. There, before guests arrive, homes are scented to create a different, nonprivate smell, and after guests leave homes are treated to return them to their original, private smell. Privacy at home can also be auditory – keeping unwanted sounds out.

Creating privacy in public

While personal privacy at home is the predominant focus of geographical scholarship, it by no means exhausts the spatialities of privacy. Indeed this possibility is implicit in the conception of privacy operationalized in the previous section, namely privacy as an accomplishment shaped by social, political, and cultural framings, and imbued with both porous and fixed relations to the public sphere. The seeking of personal privacy transforms public spaces into private spaces, illustrated through the following two examples.

Cars dominate cities around the world. While they are often associated with traffic, movement, and road infrastructure, geographers also point out that cars are spaces of personal intimacy.
They are physical cocoons. Within them drivers are strapped into seats and afforded some protection from surrounding objects and people by metal and locks. This material cocooning is correlated with a sense that cars are also individual, providing a space to relax, be alone, and be oneself. According to Kent (2014), “the car is explicitly perceived and felt as a very personal and private space. It is a place where the body is shielded from others, and from the biophysical environment.” Kent’s work also highlights car drivers’ use of the car as much more than a means of traveling from A to B. Cars facilitate time to zone out and “me time.” Cars are spaces that are personal and personalized, in which individuals can choose to interact. The car is also a site for familial privacy and intimacy. As with the home, it is in the car (sometimes through information and communications technology (ICT), sometimes through the thinking space it provides) that family arrangements are made. For women who are mothers, the car becomes a site valued for its facilitation of private domestic conversations. In this respect, cars are mobile sites of personal privacy that are private in the way they are experienced, and public in that occupants are visible to outsiders and traveling through public space.

A second example is also of privacy in public: mobile privacy. Hatuka and Toch (2014, 3) have articulated the concept of “portable private-personal territory” which they define as “a personal space that individuals develop and that is characterized by a multi-dimensional set of social and spatial relationships.” This form of personal privacy on the go is facilitated by the proliferation of communication technologies, in particular mobile phones. These allow people to be private in public, bringing ostensibly private activities like chatting and watching television into new spaces. Smartphones, in other words, provide personal visual and audio spaces regardless of a person’s location, and allow users to ignore the presence and actions of people around them. Personal privacy is, thus, becoming disconnected from the physicality of space.

Personal privacy in the twenty-first century

The twenty-first century has heralded significant shifts in the meaning of, and the influences on, personal privacy, especially in relation to information about the person that may be deemed private. These shifts have occurred partly because of new roles and relationships among privacy actors. First, there has been an explosion of social media platforms – Twitter and Facebook being the most well-known examples – designed solely for sharing personal information. Second, individuals are more likely to reveal highly personal details online, such as in blogs. The first two points converge to create a third – an explosion of personal information outside the control of the individual and often in the hands of corporations (Elwood and Leszczynski 2011). In essence, personal privacy in terms of information about the person has become more public and less private.

Smartphones and other locative (GPS-enabled) devices bring to the fore questions of locational privacy – the ability of a person to move about without being tracked or monitored (see de Souza and Frith 2012). For example, when it was discovered that users of some smartphones were being tracked in 2011, ostensibly without their permission, and that this data was available to others, public consternation and discussion ensued (see Kumpu 2012). Likewise, the geotagging of online content, or technologies like Google Street View, are provoking public debate about what personal privacy means. Moreover, as Elwood and Leszczynski (2011, 7) argue, these
debates are conceptually important for the ways they “reconstitute the objects and practices of privacy concern, and in so doing alter the roles and relationships of state, civil and corporate actors in the constructions of privacy.” In sum, geographers are interested in these privacy violations not just for their detail, but for what they reveal about transforming the spatialities of privacy.

SEE ALSO: Home; Public space

References


Processing LiDAR data

Maria Antonia Brovelli
Carolina Arias Muñoz
Politecnico di Milano, Italy

Processing of Light Detection and Ranging (LiDAR) data can vary depending on the different technologies used for data collection and on the desired output mapping products. One thing that is common for all technologies and mapping products is the generation of digital terrain models (DTMs) – bare earth – and digital surface models (DSMs), which are the most used and fastest growing LiDAR observations products. In almost all LiDAR applications, point cloud filtering, or ground filtering, is a necessary step in LiDAR data processing to determine which LiDAR returns are ground surface features, and which are nonground surface features. The former are used to generate DTMs, while the latter are used to extract DSMs. DSMs are surfaces containing features that reflect the first return of each laser pulse (e.g., trees, buildings, or towers), while DTMs are surfaces representing the bare ground without any objects like vegetation or buildings (see Figure 1).

The general procedure to generate a DTM from a point cloud, as described in Meng, Currit, and Zhao (2010), can be summarized in four steps: (i) Error filtering; (ii) Interpolation, resampling, or reorganization; (iii) Ground filtering; and (iv) DTM generation. Error filtering consists of the elimination of outliers which normally have excessively high or low elevation values. To identify these outliers an analysis of the elevation value’s frequency distribution can be done, as well as a comparison of each point to a local elevation reference, a residuals analysis after interpolation, or simply a manual examination of the dataset. After error filtering, some interpolation, resampling, or reorganization is needed to create raster datasets from the point cloud. This procedure could be necessary because certain ground filtering algorithms depend on raster-based neighborhood search logic. Methods for point cloud rasterization assign to the pixel value the lowest or nearest point within an area, or interpolate the pixel value using kernel functions. Ground filtering, as previously mentioned, is the process of separating ground and nonground points or pixels. Finally, to generate a DTM, the ground points identified after ground filtering have to be interpolated to produce a continuous surface. The interpolation methods include inverse distance weighting, splines functions with regularization, absolutely minimizing Lipschitz extension (AMLE), kriging, and other hybrid methods that combine linear and non-linear interpolation. Steps i, ii, and iv are common to all field measurements; ground filtering, meanwhile, is specific to LiDAR observations.

In order to select a ground filtering method, several aspects have to be considered: the terrain characteristics, the number and type of LiDAR returns, and the definition of neighborhoods, among others. Concerning the number of LiDAR returns, first and last returns are used for ground filtering; intermediate returns are mostly used for separating vegetation from solid objects. The definition of neighborhoods is essential to describe the height context of ground features. The height context is defined by a local neighborhood function which assigns weights to each point within that neighborhood. The function...
could also be performed as a series of linear, circular, or rectangular windows. Furthermore, most filtering algorithms iteratively modify the neighborhood size to improve filtering accuracy.

Though filtering algorithms can be differentiated on the basis of different operative characteristics, the more interesting distinction can be made upon the assumptions on which the different filters work.

Tóvári and Pfeifer (2005) point out four distinct algorithm types: morphological or slope-based, progressive densification, surface-based, and clustering/segmentation algorithms. Other less commonly used methods include block-minimum and multiscale (Lee et al. 2013).

*Morphological or slope-based algorithms* use height differences to recognize points belonging to objects. The classification is performed using a threshold that determines the admissible height difference between two neighboring points such that one can be considered as bare earth. This approach can also use fixed slope or gradient threshold values instead of height differences. *Progressive densification algorithms* are filters that at first identify some points belonging to the ground and then, depending on those, classify more points as ground. Usually the points used as seed are the ones with lower height. Additional ground points are determined by investigating their neighbor on the reference surface. *Surface-based algorithms* are filters that use a parametric surface that iteratively approaches hypothetical bare earth. The surface is modified depending on the influence of the individual input points. Common interpolation methods include: hierarchical robust interpolation; bilinear/bicubic spline functions with Tychonov regularization in least squares approach; thin-plate spline (TPS); the facet model; and statistics methods such as minimum values. *Clustering/segmentation algorithms* are filters that are based on the idea that a cluster of
points belong to an object if it has height values greater than its neighbors. In these cases the classification is performed in two steps: first, a segmentation is carried out, and then the segments are divided into different classes depending on the differences in height between segments (Sithole and Vosselman 2003). Block-minimum filters search the minimum value in a given spatial extent within neighboring points. Multi-scale filters are based on the fact that surface properties are scale dependent: feature variations exist at a smaller-scale range than terrain variations. The point cloud is interpolated into multiresolution grids, and points are then classified based on a threshold that varies according to the procedure used. Decision algorithms include: multiscale Hermite transform (MHT) filter; multiscale curvature classification (MCC); and multidirectional ground filtering (MGF).

Automated filter algorithms are built into LiDAR software tools. These tools have several data-processing and information-extraction functions, with ground filtering as their core function. Among the most widely adopted proprietary software is TerraScan, which is the most complete, advanced, and powerful software available for LiDAR data processing and analysis. It has the capability of handling big datasets by creating projects and breaking the big datasets into smaller tiles that can be processed individually. LAStools Software Suite is a collection of highly efficient, batch-scriptable, multithreaded command-line tools for LiDAR data processing, which can be run via a graphical user interface and are available as a LiDAR processing toolbox for ArcGIS versions 9.3, 10.0, and 10.1 of the Environmental Systems Research Institute’s (ESRI) mapping software. General tools such as those for data format conversion, header information extraction, or data precision are free, but more sophisticated tools for filtering, classification, and interpolation require licensing.

Free and Open Source Software (FOSS) offers alternatives for LiDAR tools, and some of these suites have the same processing capacity as proprietary ones. Among the FOSS alternatives, GRASS GIS includes a suite of tools related to processing LiDAR data (Brovelli and Lucca 2011) as described on the GRASS LiDAR wiki (https://grasswiki.osgeo.org/wiki/LIDAR). It has a series of commands to perform DEM/DSM separation: v.lidar.edgedetection detects edges from last return data; v.lidar.growing generates buildings from detected edges; v.lidar.correction postprocesses data; and finally v.surf.bspline computes DTMs and DSMs. BCAL Lidar Tools is another FOSS tool, developed by the Idaho State University Boise Center Aerospace Lab as a plug-in for the ENVI (ENvironment for Visualizing Images) software package. It is written in IDL (Interactive Data Language) programming language and includes a bare-earth digital elevation model (DEM) creation and a height filtering tool optimized for open rangeland (sagebrush) vegetation. SAGA GIS is an open source geographic information system (GIS) package that includes tools to work with point clouds, like reclassifier, attribute calculator, subset extractor or gridding, and interpolation. FUSION is a LiDAR/IFSAR (interferometric synthetic aperture radar) data conversion, processing, and display FOSS suite. It processes raw LiDAR data into a number of vegetation metrics; and it has a processing tool called Canopy Fuel Estimator (CFE) that uses LiDAR data to characterize canopy fuels over a landscape. MCC-LiDAR is also an open source command-line tool for processing discrete-return LiDAR data; it is used especially for forested environments, with the MCC algorithm for ground filtering.

In the future, LiDAR studies for data processing need to focus on large LiDAR datasets processing optimization, full waveform LiDAR data analysis, and integrated analysis of LiDAR data.
and other remote sensing products, like multispectral image data. Until now, ground filtering algorithms have had performance problems with surfaces with irregular terrain or discontinuous slope, low vegetation regions – often ignored by many ground filters – and dense forest canopies that reduce the penetration of laser beams. This means that ground filtering algorithms research is far from being over.

**SEE ALSO:** Ground-based LiDAR; LiDAR

**References**


Producer services and trade liberalization

Holger Breinlich
University of Nottingham, UK

Trade liberalization and recent advances in communication and information technologies have increasingly permitted the trade of services, including producer services, across borders. This entry discusses recent research on the nature and importance of such trade, on existing trade barriers, and on the progress and consequences of services trade liberalization. The concept of trade liberalization is broadly interpreted here to refer to the elimination or reduction of any policy measure which explicitly or implicitly discriminates against foreign providers of a service relative to the domestic providers of the same or related services. The analysis will focus on the economic consequences of services trade liberalization and look at the impact of liberalization on variables such as firm- or sector-level output and productivity. The entry presents the historical context of this research; describes the most common definitions of trade in producer services and the forms it can take, and briefly looks at issues of measurement and data quality; discusses trends in aggregate services trade flows and summarizes recent findings on transactions at the firm level underlying these aggregate flows; and provides an overview of existing restrictions on services trade, attempts to liberalize trade, and the economic consequences of such liberalizations.

Historical context

Research on services trade has a relatively short intellectual history. This is in stark contrast to the much older research efforts related to goods trade which go back to at least the works of Adam Smith in the late eighteenth century. While there were some important early contributions in the 1960s and 1970s, a broader research program on services trade was established only in the 1980s, largely in reaction to the trade negotiations during the Uruguay Round that resulted in the General Agreement on Trade in Services (GATS) in 1994 (see Francois and Hoekman 2010). Since then, the research agenda has evolved to span a wide range of issues such as the nature, definition, and measurement of services trade, the collection of comprehensive international statistics on trade flows and barriers to services trade, and the evaluation of the consequences of trade liberalization. While initial research has concentrated mainly on insights into broad trade flow patterns and the aggregate consequences of trade liberalization, more recent research has focused on the firm-level decisions underlying these aggregate trends.

Trade in producer services: definitions and measurement problems

The provision of producer services across borders is in many ways fundamentally different from the trading of physical goods. For example, Hill (1977) notes that an important characteristic of most services is a “joint production requirement.” As services often
PRODUCER SERVICES AND TRADE LIBERALIZATION

cannot be stored, both producer and consumer need to be present at the same time, and possibly in the same location, for a service transaction to take place. This means that services trade often requires a combination of direct cross-border transactions as well as a local presence through the movement of persons and the establishment of foreign affiliates.

The two most common definitions of services trade reflect these peculiarities. First, the residential or balance-of-payment definition of services trade defines it as service transactions between the residents and nonresidents of an economy. For example, the provision of call center services from India to customers in the United States is such a transaction. A presentation given by a London-based consultant to clients in Paris, or the attendance of the same clients at a training course in London, also represent services trade. Note that, while these last two examples do not involve cross-border transactions, the producers and consumers remain residents of their respective countries; thus, their interaction represents transactions between residents and nonresidents.

Second, an often used alternative definition focuses on the modes through which services transactions take place. This GATS definition classifies services trade into four modes. Mode 1 (“cross-border supply”) comprises transactions in which the producer and consumer of a service remain in their respective countries. The provision of call center services mentioned earlier would be a mode 1 transaction. For transactions in mode 2 (“consumption abroad”), the consumer moves to the country of the producer (the Paris-based clients attending a training course in London), while in mode 4 transactions (“presence of natural persons”) the producer moves abroad to provide the service (the London-based consultant giving a presentation in Paris). While these three modes overlap closely with the residential definition, the GATS definition also includes mode 3 (“commercial presence”), where companies supply services internationally through the establishment of foreign affiliates that are legally resident in the host country, and their transactions with local consumers do not count as services trade according to the residential definition.

The discussion of these two definitions illustrates that the measurement of services trade is substantially more complicated than that of goods trade. The shipment of physical goods across borders can be monitored with relative ease, and a long tradition of levying customs duties has led to the establishment of sophisticated statistical systems in most countries. In contrast, services are intangible and much harder to record (World Trade Organization 2013, 202). Many countries rely on surveys in which the coverage of service-producing establishments is often incomplete. If central bank records are used instead, transactions that do not use financial intermediaries go unobserved. Services that are transmitted electronically or between the establishments of individual multinational enterprises are frequently not registered at all. Also, while substantial progress has been made in setting up internationally comparable methodologies for data collection, there remain important gaps in terms of coverage, in particular for developing countries. As a consequence, data on services trade is much less complete than data on goods trade and is generally assumed to be of inferior quality (Lipsey 2006).

Overview of recent trends in services trade

Figure 1 shows the evolution of aggregate world exports for the principal categories of international trade. Trade in commercial services, of which producer services are the
dominant category, stood at US$4.35 trillion in 2012. Since 1980, nominal yearly growth rates have averaged 11.9%. Despite this strong growth, the share of commercial services in total world exports has increased only slowly over the last 30 years, from 16% in 1980 to 19% in 1990 and 20% in 2012. As seen in Figure 1, this is because other categories of international trade (manufactures, in particular) have been growing quickly. Nevertheless, trade in commercial services has been the fastest-growing component of international trade since 1980.

The subcategories of commercial services do not neatly subdivide into producer and consumer services (i.e., services for intermediate use and for final consumption), but Table 1 shows that categories that mostly incorporate producer services (transportation, computer and information services, royalties, construction, and business services) accounted for 62% of the total value of commercial services trade in 2012. Those comprising a mixture of consumer and producer services (insurance, finance, and communications) accounted for 12%, and categories dominated by consumer services (travel, and personal, cultural, and recreational) for 26%.

The most important type of producer services in 2012 were business services which consist mainly of professional and technical services such as legal, accounting, management consulting,
Table 1  Structure of international trade in producer and consumer services, 2000–2012 (%).

<table>
<thead>
<tr>
<th>Type of service</th>
<th>2000</th>
<th>2005</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal, cultural, and recreational services</td>
<td>1.0</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>Insurance services</td>
<td>1.9</td>
<td>2.0</td>
<td>2.3</td>
</tr>
<tr>
<td>Communications services</td>
<td>2.3</td>
<td>2.4</td>
<td>2.6</td>
</tr>
<tr>
<td>Construction services</td>
<td>2.0</td>
<td>2.2</td>
<td>2.6</td>
</tr>
<tr>
<td>Computer and information services (IT)</td>
<td>3.1</td>
<td>4.2</td>
<td>6.0</td>
</tr>
<tr>
<td>Royalties and license fees</td>
<td>6.2</td>
<td>6.4</td>
<td>6.7</td>
</tr>
<tr>
<td>Financial services</td>
<td>6.5</td>
<td>7.2</td>
<td>7.0</td>
</tr>
<tr>
<td>Transportation</td>
<td>23.2</td>
<td>22.7</td>
<td>20.5</td>
</tr>
<tr>
<td>Travel</td>
<td>31.9</td>
<td>27.9</td>
<td>25.5</td>
</tr>
<tr>
<td>Business services</td>
<td>21.9</td>
<td>24.2</td>
<td>26.0</td>
</tr>
<tr>
<td>Total</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Source: Author’s calculations based on Time Series on International Trade, World Trade Organization.

The information in Figure 1 and Table 1 excludes an important channel through which providers of producer services can access foreign markets: the establishment of a commercial presence through foreign direct investment (mode 3 in the GATS definition of services trade). Data on the activities of such local affiliates is much less comprehensive and mainly available only for OECD (Organisation for Economic Co-operation and Development) countries. However, it shows that local affiliate sales exceed the combined value of trade through GATS modes 1, 2, and 4 in many services sectors. For example, Oldenski (2012) estimates that mode 3 sales of US multinationals were on average twice as large as their sales through all the other modes put together in 2005. In contrast, US multinationals sold as much manufactured goods through direct exports as through local affiliate sales. Similar findings hold true across a range of OECD countries (WTO 2013) and further highlight the fact that, despite substantial progress in communications technology, proximity to customers remains of paramount importance for services providers (Oldenski 2012).

Research on trade in producer services also increasingly makes use of the firm-level data used to construct balance of payments aggregates. Such data have recently become available in a number of countries. Using British data, Breinlich and Criscuolo (2011) show that only around 6% of firms export producer services, but that these exporters show a number of positive performance characteristics, such as higher sales and productivity than nonexporters. Exports of producer services are also concentrated among architectural, or engineering services. The overall share of producer services has also increased markedly since 2000, the first year for which disaggregated data are available. Particularly striking is the rise of computer and information (IT) and business services, which by 2012 accounted for a third of all services trade. Explanations often advanced for this trend include technological progress that has made the cross-border transmission of information easier, as well as changes in organizational practices such as outsourcing and offshoring.

The information in Figure 1 and Table 1 excludes an important channel through which providers of producer services can access foreign markets: the establishment of a commercial presence through foreign direct investment (mode 3 in the GATS definition of services trade). Data on the activities of such local affiliates is much less comprehensive and mainly available only for OECD (Organisation for Economic Co-operation and Development) countries. However, it shows that local affiliate sales exceed the combined value of trade through GATS modes 1, 2, and 4 in many services sectors. For example, Oldenski (2012) estimates that mode 3 sales of US multinationals were on average twice as large as their sales through all the other modes put together in 2005. In contrast, US multinationals sold as much manufactured goods through direct exports as through local affiliate sales. Similar findings hold true across a range of OECD countries (WTO 2013) and further highlight the fact that, despite substantial progress in communications technology, proximity to customers remains of paramount importance for services providers (Oldenski 2012).

Research on trade in producer services also increasingly makes use of the firm-level data used to construct balance of payments aggregates. Such data have recently become available in a number of countries. Using British data, Breinlich and Criscuolo (2011) show that only around 6% of firms export producer services, but that these exporters show a number of positive performance characteristics, such as higher sales and productivity than nonexporters. Exports of producer services are also concentrated among architectural, or engineering services. The overall share of producer services has also increased markedly since 2000, the first year for which disaggregated data are available. Particularly striking is the rise of computer and information (IT) and business services, which by 2012 accounted for a third of all services trade. Explanations often advanced for this trend include technological progress that has made the cross-border transmission of information easier, as well as changes in organizational practices such as outsourcing and offshoring.
a small number of firms; 50% of the total is accounted for by fewer than 15% of exporters. Similar results have been found for a number of other countries such as Italy, Germany, France, and Turkey.

Trade barriers and trade liberalization

Some characteristics, such as the need for proximity to customers, make trade in producer services inherently difficult, but there are also a large number of policy measures that prevent firms from accessing foreign markets. Prominent barriers include limits on foreign ownership of local affiliates, licensing requirements for foreign firms and professionals, aspects of domestic regulatory regimes, or visa restrictions (relevant for mode 4 transactions).

While the variety and complexity of these barriers make measurement difficult, most forms of trade restrictions fall into one of three categories (Borchert, Gootiiz, and Mattoo 2014). The first consists of measures that explicitly discriminate against foreign service providers, such as restrictions on foreign ownership of domestic firms. The second consists of policies that do not explicitly discriminate but may still disproportionately affect foreign providers; for example, domestic qualification requirements to practice professional services such as law, auditing, and accounting impose an additional burden on foreign professionals who have already met requirements in their home countries but whose qualifications are not recognized abroad. Finally, an absence of regulation can also represent de facto barriers to foreign services producers. This is the case, for example, if regulation fails to ensure equal access for all providers to essential infrastructure such as telecommunication networks.

Even when the complexities of measuring services trade barriers are taken into account, the availability of data on such barriers can only be described as poor. This is the case even though the share of producer services in world trade is rising and services trade barriers are increasingly important in bilateral and multilateral trade negotiations. Initially, research in this area focused on attempts to quantify the services trade commitments made by members of the World Trade Organization (WTO) as part of the GATS negotiations in the 1990s (Hoekman 1996). However, subsequent work has established that actual services trade policy is much more liberal than countries’ WTO commitments would suggest (Borchert, Gootiiz, and Mattoo 2014). Put differently, it seems that services trade commitments only guarantee a minimum level of openness which is usually surpassed in actual policy.

More recent research has tried to provide better measures of the de facto openness of countries to services trade. This is usually done by surveying private sector representatives, local law firms, and government officials to obtain information on the most relevant barriers for each sector and GATS mode across different countries. Indicators of the presence of individual policies are then aggregated to the sectoral or country level, using subjective weights of the importance of individual measures.

One challenge any such aggregation faces is to take account of possible substitution between GATS modes. For example, if a given producer service can be delivered either cross-border (mode 1) or through the movement of professionals to the other country (mode 4), substantial trade barriers in only one mode need not imply high effective barriers – producers can simply switch to service delivery via another mode. On the other hand, if service delivery requires a combination of modes, such as a
local commercial presence (mode 3) combined with cross-border trade (mode 1), the higher of the two trade barriers will be relevant for a producer trying to access the market in question. To some extent, the use of subjective weights in the construction of restrictiveness indices allows such issues to be addressed, although extensive validation exercises have not been carried out so far. It is also unclear whether the use of fixed weights can really adequately capture the complex substitution patterns between modes.

A number of studies have tried to quantify the barriers affecting individual sectors or countries, but there are currently only two databases that aim at providing a more comprehensive picture. The Product Market Regulation Index of the OECD provides data on all the important types of producer services for OECD countries for 1998–2013. The World Bank’s Services Trade Restrictions Database (STRD) offers more limited time coverage but also includes information for 79 developing countries. Both databases use the combination of survey data and aggregation via subjective weights discussed earlier.

The picture that emerges from these data is one of substantial heterogeneity across countries, types of services, and GATS modes. Using the STRD, Borchert, Gootiiz, and Mattoo (2014) show that there is a negative correlation between the restrictiveness of a country’s policy regime and its per capita income. By and large, developing countries impose higher services trade barriers than high-income countries (Figure 2). However, there are important exceptions to this general pattern. For example, the Gulf Cooperation Council (GCC) countries – Bahrain, Kuwait, Oman, Qatar, and Saudi Arabia – are high-income but exhibit some of the most restrictive policy regimes. There are also developing countries such as Rwanda, Madagascar, Senegal, and Mongolia that are quite open to services trade, while some of the most rapidly growing Asian economies (China, India, Indonesia, Malaysia, the Philippines, and Thailand) impose above average trade barriers.

In terms of types of services, professional services (here accounting, auditing, and legal services) encounter the highest barriers to trade (Figure 2). This is also true for OECD economies, which maintain substantial immigration restrictions and licensing and qualification requirements, making the delivery of such services by foreign professionals exceedingly difficult. Substantial barriers remain in transportation services as well, whereas access in banking, insurance, and telecommunication services is relatively unrestricted, at least in high-income OECD countries. However, Borchert, Gootiiz, and Mattoo (2014) caution that a high degree of market access uncertainty remains in most countries, even if formal services trade barriers are low. For example, the allocation of licenses is often opaque and highly discretionary; regulators in some countries do not have to provide justifications for rejecting license applications; and foreign service providers do not always have a right of appeal against unfavorable regulatory decisions.

To date, attempts to liberalize trade in producer services have taken a number of forms. Best known are the multilateral negotiations leading to the implementation of the GATS at the end of the Uruguay Round in 1994. More recently, services trade liberalization has also played an important role in a number of regional trade agreements. Since the 1980s, in parallel with these multilateral and bilateral initiatives, countries have undertaken unilateral liberalizations in a range of sectors such as telecommunication and transportation.

Services trade liberalization in the GATS takes the form of so-called commitments, whereby countries guarantee market access and national treatment (i.e., the nondiscriminatory application of domestic regulation and taxation to both
Figure 2  Trade barriers for producer services, 2008–2011. Horizontal axis denotes country groups categorized by income level; STRS measures the overall restrictiveness of a country’s policies on services trade (0 = open without restrictions; 100 = completely closed) (author’s calculations derived from Services Trade Restrictions Database; Borchert, Gootiiz, and Mattoo 2014).

foreign and domestic providers) for specified sectors and modes of supply. The degree of commitment is codified in a country’s schedule of commitments and can vary across modes for the same service activity. For example, a country may guarantee market access and national treatment for cross-border service provision (mode 1) while imposing restrictions on the establishment of a commercial presence (mode 3).

Countries can also choose not to commit themselves by writing “unbound” in their schedule for a given service type. This does not mean that market access will be necessarily restricted, only that the country does not guarantee access and national treatment. As a consequence, the GATS schedules only provide an incomplete listing of policies that are actually applied. This is known as a “positive list” approach because only unrestricted service types and modes are explicitly enumerated. More recent regional trade agreements have generally followed the commitment approach of the GATS and have also relied on positive-list schedules. An exception is the North American Free Trade Agreement (NAFTA) where the
presumption is that market access and national treatment are guaranteed unless restrictions for a sector or service type are explicitly mentioned (a “negative list” approach).

How successful have multilateral and bilateral trade liberalization initiatives been in reducing services trade barriers? The general view in the literature is that, with some exceptions, they have not been effective, and that most meaningful liberalizations have taken place on a unilateral basis (Francois and Hoekman 2010). As mentioned earlier, analysis of GATS commitment schedules has revealed that actual policies are often substantially more liberal, and that this was the case even before the implementation of the GATS (Borchert, Gootiiz, and Mattoo 2014). Even the European Union, the deepest regional integration initiative at present, has encountered serious resistance to attempts at services trade liberalization such as the Services Directive.

Why bilateral and multilateral liberalization initiatives have been mostly unsuccessful is still not fully understood. Possible explanations include conflicting regulatory objectives of national regulators, weak exporter lobbies, and a lack of suitable reciprocal offers (Francois and Hoekman 2010). Many service sectors, such as professional services, are subject to substantial regulation, and it is often impossible to agree common minimum standards across countries. Because services sectors tend to be less export-oriented than manufacturing, exporter lobbies seeking market access abroad may also not be sufficiently strong to initiate liberalization initiatives. Finally, while service providers in developing countries would stand to benefit most from the liberalization of immigration regimes that would allow easier market access through mode 4, such access is strongly opposed in most developed economies, reducing the scope for meaningful exchanges of concessions between countries.

Given the lack of success of bilateral and multilateral negotiations, most of the evidence regarding the effects of services trade liberalization comes from unilateral liberalizations. One strand of this literature relies on cross-country panel data and tries to link sector-specific liberalization initiatives to performance measures such as sector-level output and productivity. The findings of these studies are generally supportive of a positive economic impact of services trade liberalization. For example, Fink, Mattoo, and Rathindran (2003) analyze the impact of policy changes relating to ownership and competition in the telecommunications sector in 86 developing countries in the period 1985–1999. They find that these reforms led to better sectoral performance, measured as increases in fixed line communication connections and in labor productivity (i.e., production per employee). However, a common problem encountered by Fink, Mattoo, and Rathindran and similar cross-country studies is the difficulty of disentangling the effects of specific policy changes from the myriad of confounding factors that also impact sectoral economic performance, an issue reminiscent of the cross-country literature on (goods) trade and growth.

More recently, a number of papers have used firm-level data and focused on liberalization episodes in individual countries. For example, Arnold, Javorcik, and Mattoo (2011), in a study of services liberalization in the Czech Republic, find that foreign entry into producer services contributed to improved productivity of downstream domestic manufacturing firms that used the inputs produced by the foreign-owned service providers. Similar results have been obtained for India (Arnold et al. 2010) and Chile (Fernandes and Paunov 2012). Given the greater ability of firm-level studies to disentangle channels of influence and to pinpoint specific mechanisms, this latest research provides more convincing
evidence for the beneficial effects of services liberalization on economic activity. It remains to be seen, however, to what extent these findings can be generalized to other countries and to the liberalization of GATS modes other than mode 3.

Conclusions

Trade liberalization and recent advances in communication and information technologies increasingly permit the trade of producer services across borders. Trade in producer services has been growing rapidly since the 1980s and now accounts for close to 70% of services trade and 14% of total world trade. Services trade liberalization has featured prominently in multilateral negotiations in the World Trade Organization and in many regional trade agreements. However, progress in liberalizing services trade has been slow and has been achieved mainly via unilateral initiatives. As a consequence, important barriers remain between many countries and across types of services, despite the fact that a small but growing body of research has shown that trade liberalization in producer services can have substantial positive economic effects.

SEE ALSO: Corporate spatial organization and producer services; Outsourcing; Producer services: definition and classification; Producer services and economic development; Regional economic integration; Trade, FDI, and industrial development

References

Over the past century, the Western world has experienced significant economic restructuring, from manufacturing to services and then to producer services, which is depicted as a “quiet revolution.” In contrast with consumer services, which are primarily consumed by individuals and households, producer services are intermediate-demand inputs supplied to enterprises or public institutions for the production of goods or other services, including economic activities as diverse as finance, insurance, real estate, business services, and transportation and communication services. Intensively involved in the production process and providing state-of-the-art knowledge and expertise, producer services not only generate direct economic benefits as a substantial source of output and job creation, but also significantly contribute to regional economic performance as a catalyst of technological and organizational change.

The move from Fordism to post-Fordism and the “global shift” of industrial production since the 1970s has given rise to a world economy characterized by “flexible specialization” in production and “vertical disintegration” in organization. As an outcome of the deepened division of labor, producer services have undergone a dramatic expansion in response to recent changes in the economy, and as a consequence, replaced the manufacturing sector as the principal engine of urban and regional development. Producer services exhibit “a geographical logic” of location different from that of the manufacturing sector. The location of manufacturing activities is influenced primarily by production and delivery costs, such as transport, labor, and land. Producer services, however, reveal a strong tendency to concentrate in certain places that provide access to the benefits arising from spatial agglomeration, such as a reduction in transaction costs and a localized process of “collective learning” (Coe and Townsend 1998). From the urban system perspective, producer services are more likely to locate in large metropolitan areas. Within such cities, these service activities display a strong clustering propensity in the form of agglomeration in central business districts (CBDs) (Illeris 1996).

Thus, the rise of producer services has dramatically reorganized urban and regional economies, the structure of the urban hierarchy, interactions between cities, and urban form and landscape. In the “transnational urban system,” “world cities” or “global cities” have emerged to dominate the system as the strategic sites for producer services. In subsystems at the national and regional levels, cities that were not able to successfully transform from manufacturing-based to service-based economies have been downgraded. In recent years, an outward movement of producer services from large to medium and small cities at the urban system level and from cities proper to suburban areas at the city level has been observed (Coffey and Shearmur 2002). Advances in telecommunications and transportation enable producer service suppliers to enjoy lower wages, lower rents, and a better living environment in small cities and suburban areas.
areas, while still retaining their economic and social contacts with clients in large cities and elsewhere. The decentralization of producer services does not reduce the importance of large cities and downtown areas for attracting producer services; rather, it gives rise to inter- and intra-city functional specialization, leading to the emergence of specialized service centers in the urban system and new suburban centers within cities.

The phenomenon of producer service growth and its impacts on urban and regional development are not confined to developed economies. Over the past two decades, even though a degree of “catch-up” with the West is still required, producer services in developing countries have undergone rapid expansion and become an increasingly important force driving economic, urban, and social change (Daniels 1998). For example, the development of the Asia-Pacific region into a “workshop to the world” has enabled it to be one of the fastest-growing global regions. Although specializing in manufacturing, the region is increasingly dependent on recent producer service growth and its “strategic” impacts in order to retain its competitive advantage. The growth of producer services in the Asia-Pacific demonstrates an uneven pattern similar to that of its Western counterpart but, because it is embedded in a different economic and political context, it also shows attributes distinctive from Western experiences. As Daniels, Ho, and Hutton (2005) argue, service development in the Asia-Pacific is characterized by rapid tertiarization, parallel development with manufacturing, and strong state intervention.

China, the second-largest economy in the world, is one example. It has undergone a profound economic transition based on the implementation of reforms and an open-door policy towards investment and trade since the late 1970s. The rapid economic growth and the acceleration of industrialization that has followed, especially since the mid-1990s, has driven up demand for producer services, leading to their rapid expansion. In 2011, producer services generated US$1618.2 billion (10 032.8 billion RMB) in gross domestic product (GDP) and created 22.15 million jobs, accounting for 21.2% of total GDP and 15.4% of total employment, respectively. Some of the general attributes of service development in the Asia-Pacific, as identified in the existing literature, are certainly replicated in China. However, a history of central planning and the recent transition of the country to a socialist market economy have complicated the use of existing theories derived from the experiences of Western and other Asia-Pacific economies.

It is widely known that China adopted the planned economy framework during the Maoist era. The central government was the sole agent in allocating resources and arranging production. Although a market economy has been actively developed since the economic reforms in 1978, the Chinese state has adopted a gradual mode of market transition, which has resulted in distinctive state–market relations that affect the growth of producer services and, consequently, their role in economic and urban development. Indeed, the Chinese party-state still plays a crucial role in shaping the growth trajectory of producer services (Yeh and Yang 2013). Prior to the economic reforms, when production and distribution were arranged within the “plan,” demand for producer services was rare. While economic reform started to remove obstacles and provide opportunities for the growth of producer services, their dramatic expansion did not happen until the Chinese government recognized their importance in the economy and adopted a series of policies and approaches to facilitate their growth from the mid-2000s onwards, such as the emphasis of producer services in the two
recent Five-Year Plans (eleventh and twelfth). Producer services in China are still relatively less developed than those in Western economies, even though they show a higher growth rate. Nonetheless, they have contributed significantly to changes in the Chinese urban system and within Chinese cities in the past two decades.

First, in contradistinction to most developed and developing countries, industrialization in China did not only occur in urban areas, but also occurred in the countryside, giving rise to a large number of small and medium-sized cities. But producer services show a strong inclination to locate in large metropolitan areas to enjoy the advantages of agglomeration economies. It is also the case that the recent decentralization of producer services in Western countries has yet to become evident in China (Yang and Yeh 2013). The net effect is a significant reinforcement of the role and significance of extra-large cities in the urban hierarchy, leading to significant reorganization of the Chinese urban system. The growth of producer services is thus an important underlying process that has ushered in a new wave of urbanization in China since the 2000s. Second, the boom in producer services has underpinned the emergence of “new urban spaces” such as CBDs that have substantially reconfigured the internal structure of Chinese cities. As the “natural habitats” for producer services, CBDs are a common feature of Western cities, but before the rapid growth of producer services in the 1990s they were absent from Chinese cities. More significantly, rather than being a spontaneous form of development, the growth of CBDs in China reflects active state action in urban planning and land leasing. The distinctive state–market relationship and its interplay with the forces of globalization mean that the theoretical basis for understanding the relationship between producer services and economic development in China requires some reconstruction.

Because producer services deliver strategic benefits to the economy, they are increasingly regarded as an important determinant in the formulation of urban and regional policies. However, as noted earlier, producer services demonstrate an uneven pattern of growth in favor of certain cities and locations. The locational attributes of producer services should be seriously considered in order to formulate appropriate policies for promoting urban and regional economic and spatial transformation.

SEE ALSO: Corporate spatial organization and producer services; Deindustrialization; Demand and supply for producer services; Flexible specialization; Global cities; New economic geography; Producer services: definition and classification; States and development; Uneven regional development

References


PRODUCER SERVICES AND ECONOMIC DEVELOPMENT


Further reading


Based on which economic activities use them, services are often classified into those used mainly by other businesses and those primarily used by households for final consumption. The former are classified as producer services and the latter as consumer services. Producer services are “part of the supply capacity of an economy; they influence its adjustment to changing economic circumstances [by enabling users] to adapt skills, attitudes, products, and processes to changes, or to reduce the structural, organizational, managerial, and informational barriers to adjustment” (Marshall et al. 1988, 13). The OECD (2000, 83) defines producer services as “intermediate inputs to further production activities that are sold to other firms, although households are also important consumers in some cases. They typically have a high information content and often reflect a ‘contracting out’ of support services that could be provided in-house.” Producer services serve markets that are internal as well as external to individual firms and those that are involved with external markets may only have transactions with intermediate users while others may also supply services for final consumption. These are sometimes described as “mixed services.”

Producer services are therefore intermediate inputs to the production of goods as well as other services, including public institutions that also purchase producer services as inputs. This reflects the way in which most producer services activities provide specialist knowledge and expertise to their clients that might be required infrequently or for only limited periods each day (such as office cleaning services); this does not justify the cost of retaining their own staff to fulfill their requirements. Indeed, a rapid expansion of the range and specialization of producer services over the last 30 years, with information and communications technology playing an important role, has been encouraged by a tendency for many firms to “contract out” or to outsource to other providers all but their core business functions. Such externalization is not only about potential cost savings for firms but also about gaining access to the most up-to-date information and expertise at the time and place when they really need it in order to ensure, for example, that their products retain a competitive edge, incorporate the latest innovations in their product lines, or utilize the most advanced marketing techniques to reach their customers.

There are variations in the statistical classification of producer services that depend on the resources available to national statistical institutions and the length of time that they have been collecting appropriate information. Most, such as the North American Industrial Classification System (NAICS), will attempt to conform to the International Standard Industrial Classification of All Economic Activities (ISIC) (United Nations 1989). Using Rev. 3 of the ISIC, producer services comprise the following:

Tabulation Category: J – Financial intermediation
Division: 65 – Financial intermediation, except insurance and pension funding
Group: 651 – Monetary intermediation
The classification is organized as a hierarchy, starting with the highest order Tabulation Categories (J and K) and gradually introducing increasing levels of detail down to the four-digit classes. It will be apparent that the range of activities identified as producer services is considerable and it is necessary to drill down to the lowest levels in the classification hierarchy to achieve a satisfactory level of detail. Note that the most populated groups and classes are “other business activities”; there the ongoing challenges associated with placing producer service activities into specific categories, not least the fact that new types of these services are continually emerging and, until they are represented in sufficient critical mass or have a clear identity, they will invariably be placed in the catch-all “other business services n.e.c.”

It should also be acknowledged that researchers rarely adopt a consistent classification of producer services. For some the above classification is too narrow. There are other types of services with intermediate markets, including real estate, finance, insurance, or parts of transportation services such as air and sea freight. Some are not unambiguously producer or consumer services, rather they are “mixed” services: accountancy, tax, or banking services are examples of mixed services.

SEE ALSO: Deindustrialization; Restructuring

References

Production of nature

Alex Loftus
King’s College London, UK

The “production of nature” refers to the historically and geographically specific practices through which humans “make” their environments. This making should be viewed as a coevolutionary or metabolic process in which “nature” and “society” – categories that lose their meaning within the production of nature thesis – are mutually transformed. As a thesis, it represents a profound, perhaps scandalous, challenge to the belief that nature is a pristine realm, free from human influence. Most studies have focused on the production of nature within capitalist societies. As such, the production of nature thesis, first advanced by Neil Smith (1984) in his landmark book *Uneven Development*, has come to be closely associated with a critique of nature–society relationships under capitalism and a simultaneous critique of the commodification of nature.

Smith’s original formulation is framed as a reconstruction of Marx’s concept of nature and is juxtaposed with “a bourgeois” or ideological conception of nature (Smith 1984, 15). Smith revels in the perverse claim that nature is produced, a claim that appears no less quixotic when one realizes that Smith constructs the argument as a means of better understanding processes of uneven development and gentrification. Although building on a thorough reading of Marx, Smith’s thesis goes against the grain of almost all Marxist conceptualizations of the environment: he therefore produces one of the most brilliantly original readings of a range of geographical processes.

The success of Smith’s approach lies, in part, in the relative failings of earlier theorizations of nature. *Uneven Development* was published at a moment when Marxism was characterized by a strongly urban focus and an apparent inability to speak to a growing environmental consciousness. The production of nature thesis intervened in this moment and preceded several key debates on the relevance of Marxism to environmental questions (Benton 1989; Grundmann 1991) while exploring fundamentally different ground from previous writings on Marx and nature (Engels 1976; Lukács 1971; Schmidt 1971). The thesis simultaneously chimed with the growing emphasis on anti-essentialist readings of nature that emerged in the late 1980s and early 1990s: it thus became one part of a new constellation of ideas seeking to challenge dominant readings of nature–society relations. The production of nature thesis, perhaps inevitably, came to be read alongside William Cronon’s (1991) *Nature’s Metropolis* as the theoretical counterpart to Cronon’s more empirically grounded and liberally oriented effort to read nature in the city.

The approach

The concept of nature, Smith argues, is predominantly ideological. Taking his cue from Lefebvre, Smith describes this ideology as “an inverted, truncated, distorted reflection of reality” (Lefebvre, quoted in Smith 1984, 15). This distorted worldview has changed depending on the context in which ideas about nature are developed. Although it is not an example that Smith draws
PRODUCTION OF NATURE

on, we might think of the framing of nature within apartheid South Africa. The main environmental debates under apartheid concerned wilderness conservation, often through excluding those who had a justified claim to using that land for agriculture and settlement. With the end of apartheid, environmental justice activists sought to reframe nature as a lived, practiced, and embodied realm. Challenging the ideology of nature that had existed previously, the key environmental question became how best to provide basic services for those most in need. Nature is thus recast.

For Smith an essential dualism has come to dominate the social concept of nature. Understood simultaneously as that which remains outside of human practice (an external conception) and that which is essential to the human subject (a universal conception), interpretations of nature throughout Western philosophical and scientific traditions have been plagued with the fundamental contradiction between these two understandings. Marx, Smith claims, pioneered a fundamentally different approach that liberates his writings from such contradictions; nevertheless, how one gets to grips with Marx's approach is not an easy task, given that the latter never laid out any separate treatises on “nature.” Having dispensed with Alfred Schmidt's The Concept of Nature in Marx, a book that had come to be regarded as the “definitive” account of Marx's concept of nature (Smith 1984, 18), Smith then seeks to reconstruct, largely from the Grundrisse, The German Ideology, and Capital, a fundamentally different approach.

Smith’s strategy is to reconstruct the methodological coordinates of Marx’s approach:

Marx nowhere talked explicitly about the production of nature. But in his work there is implied an understanding of nature which leads firmly in this direction … The first major task has been to detect these clues; the second is to lay them out and complete the jigsaw puzzle. Marx has given us the four corners and most of the straight edges; he has also given us most of the common pieces necessary to complete the picture, but these pieces are presented in the context of wholly different analyses. What must be done in order to recognize their significance is to turn these pieces over and, as it were, to reveal their nature face. (1984, 32–33)

The first step is to historicize nature–society relations. From the beginnings of human history, acts of production have served to break down the perceived divide between nature and society which is held to be sacrosanct within the philosophy that runs alongside it. Building on such an approach, Harvey (1996) challenges his readers to try and separate out their own lives into that which is natural and that which is social. The task quickly becomes impossible, as we consume, breathe, walk, type into a computer, and so on. Human relations with nature are, as Marx writes, metabolic. Against both environmental determinism and socialized conceptions of nature, Marx and Smith (and Harvey) see “nature” and “society” as code-determining. Capitalism establishes important sets of social relations in this process. Numerous examples of such a produced nature abound. William Cronon’s Nature’s Metropolis (1991) is one example where an environmental historian has sought to demonstrate how the emergence of a city like Chicago is dependent on a symbiotic relationship with its hinterland, and how nature is progressively transformed through this symbiosis. To take a somewhat more banal example, when an individual wakes up in the morning (in an “environment” owned by a landlord to whom he or she pays a monthly rent), the organic oatmeal and fairtrade coffee that he or she first consumes are all transformed aspects of “nature” mediated through a set of colonial, gendered, and capitalist relations of production and exchange. Although
the individual’s metropolitan life seems utterly divorced from nature, it is, in reality, entirely dependent on “nature’s” bounty. Swyngedouw and Kaika (2000) similarly recount a tale of standing on Piccadilly Circus while tracing the socio-natural connections that produce that specific place: music floats from the Rainforest Cafe, high roast coffee fills the air with its pungent odor, and the vast TV screens selling new commodities are fueled by the transformation of oil resources drawn from disparate locations.

If Smith’s work (and Kaika and Swyngedouw’s) contrasts with that of Cronon’s, it is in the greater influence accorded to capitalist relations of production and exchange. (Cronon was roundly critiqued by geographers in an entire issue of Antipode (26(2)) dedicated to Nature’s Metropolis for failing to emphasize the importance of capitalism.) Nevertheless, far from a narrowly deterministic reading of capitalism as the sole influence on nature in the present moment (a critique that Cronon seems to level at some of his own critics), Smith sees human activity (not capital, money, or capitalists) as having forged this coevolutionary relationship throughout history. Smith quotes the Grundrisse, in which Marx writes “some determinations belong to all epochs, others only to a few,” before re-emphasizing the point that the need to produce means of existence is a transhistorical act: without production, humans would, quite simply, starve, freeze, or expire. However, in tarrying with this notion of transhistoricity, Smith inadvertently replicates the dualistic conception of nature that he is seeking to refute. Thus, “Human beings are born with certain natural [universal conception] needs – food, sex, warmth, social interaction – and they are born into a world where nature [external conception] provides, either directly or indirectly, the means for fulfilling needs” (Smith 2010, 54). This brief dalliance with a dualistic conception is then quickly dispensed with, and, instead, Smith considers, albeit at an abstract level, the practices through which metabolic relations have been forged. Practical activity and the natures produced out of such practical activity shift according to a variety of determinations (the new needs that are created out of fulfilling original needs; the conceptions of the world that circulate within particular societies; and emerging divisions of labor).

Although attentive to the changes in nature’s production, Smith’s thesis can be contrasted with epochal arguments that detect a blurring of the nature–society relation within the contemporary moment. (In many respects, discussions of the Anthropocene can be read as one more recent manifestation of such an epochal argument, against which the production of nature thesis would, quite rightly, present a more historically and geographically specific reading of what has always been a metabolic process.) Challenging shallow, and often apocalyptic, claims about the death of nature, Smith is quite explicit in stating that “production in general is the production of nature” (1984, 53). Nevertheless, some confusion has emerged around the relationship between the transhistorical (the production of nature) and the historically specific (the capitalist production of nature), leading to Castree’s (2000a, 643) claim that Smith “used Marxian economics to argue that capitalism has replaced a non-human ‘first nature’ with a socially produced ‘second nature.’” Nowhere does Smith make such an argument, even if he appears to come close to doing so. Much of the confusion arises from Smith’s idiosyncratic reading of “first” and “second” nature in Uneven Development. Although they have some historical purchase in their reference to the human and nonhuman, first and second nature acquire a qualitatively different meaning in the present moment. Thus, against the
PRODUCTION OF NATURE

commonsense reading of first (nonproduced) and second (produced) nature (as interpreted by Castree), Smith (1984, 55) claims:

This distinction ceases to have real meaning … the distinction is now between a first nature that is concrete and material, the nature of use values in general, and a second nature which is abstract, and derivative of the abstraction from use value that is inherent in exchange value.

Reception

In the first readings of Uneven Development, the production of nature thesis was often overlooked. Even now there seems to be a reluctance to see Smith’s theorization of the production of nature as fundamental to the production of space. Nevertheless, by the late 1980s and early 1990s the production of nature thesis had become a fundamental influence on the fast-moving tide that sought to retheorize nature in nonessentialist ways (Braun and Castree 1998; Castree and Braun 2001). Elsewhere, Castree has noted the distinctiveness of the production of nature thesis within Marxist readings of socio-natures. He argues that Smith’s work, although much less known, avoids the pitfalls of other theorizations such as O’Connor’s and Benton’s. Writing in 2000, Castree offered

a long overdue presentation of an alternative tradition of Marxist work on the capitalist production of nature … It thus promises to take us beyond the Nature–Society dualism organising both previous Marxian work on nature and versions of bourgeois technocentrism and radical ecocentrism. (2000b, 10)

If Castree gave a long overdue presentation of the production of nature thesis in 2000 (and was in many respects fundamental to the revival of the fortunes of the production of nature thesis within and outside geographical scholarship), Smith’s work has been engaged with more readily since then; however, it has still perhaps not quite achieved the status of a “classic” thesis within Marxist debates.

Perhaps the main subfield to explore the production of nature thesis has been urban political ecology; thus, in the writings of Swyngedouw, Heynen, Kaika, and others, the production of nature has been seen as the starting point for an analysis of the power-laden ways in which nature comes to be urbanized (Heynen, Kaika, and Swyngedouw 2006). Loftus (2012) places the production of nature thesis in conversation with an immanent critique of everyday life and thereby seeks to position the thesis more centrally within the Marxist tradition. However, elsewhere the sources turned to for an analysis of Marx’s approach to nature remain works produced by sociologists, economists, and political scientists.

Several engagements with the production of nature thesis have, unsurprisingly, taken offense at the stark, and seemingly paradoxical, claim that nature is produced. For both Marxists (Benton, personal communication) and non-Marxists (Lorimer 2009), this assertion is one step too far; the thesis appears to place too much agency on humans and quickly lapses into dualisms. The production of nature is portrayed as hubristic, insensitive to nonhuman agency, and even dangerous. The normative implications are then read as the demands of a demagogue whose wish to fully “humanize” nature will be frustrated by nature’s own “liveliness.” Whatmore (1999) has taken issue with the dialectical framework. Thus, against Castree’s claim that the production of nature thesis is nondualistic, and against the clear overall intent of the production of nature thesis, Whatmore mistakenly writes: “far from challenging this a priori categorization of the things of the world, dialectics can be seen to raise its binary logic to the level of a contradiction and
engine of history” (1999, 25). These criticisms are perhaps to be expected from a thesis that is so deliberately provocative. To place labor squarely within an environmental politics was always likely to raise the ire of those who would seek to defend nature’s beauty, its vitality, and the agency of nonhuman over human. However, beneath the provocations and the aggressive prose, the production of nature thesis is acutely sensitive to nonhuman difference, to the interrelationships between human and nonhuman, while it also remains profoundly open to nuanced readings of coproduction and nonlinear science.

Future directions

Given the mass of scholarship now accumulating on nonhuman agency, it seems likely that such themes will be explored further within work on the production of nature. Donna Haraway, whose writings on the figure of the cyborg and on companion species chime rather beautifully with Smith’s ideas, would seem one of the most brilliant of fellow travelers on such a journey. Indeed, the relational approach to socio-natures developed by Haraway (2008) has many similarities to that of Smith: one can sense how indebted Smith, Hartsock, and Haraway might have been to one another in their initial shared readings of Marx and also in their collective efforts to elucidate a nondeterministic understanding of the socionatural. Hartsock’s writings on feminist standpoint theory may well provide another valuable source of inspiration for future work, enabling a more thorough exploration not only of the gendered division of labor out of which “produced natures” emerge, but also of the conditions of possibility for nonbourgeois conceptions of nature (see Loftus 2007). These conditions of possibility are enabled by the practical acts of those working within capitalist patriarchy.

Such an emphasis on the gendered division of labor suggests another crucial area for development. Abstract labor and the role of the exchange abstraction are crucial to the kinds of nature that Smith sees being produced and reproduced within capitalist societies. Nevertheless, acts of laboring simultaneously exist as concrete practices: labor under capitalism is a differentiated unity of both the concrete and the abstract. By neglecting concrete practices and the multiple relationships out of which workers are not only classed, but raced, gendered, and sexed, the production of nature thesis risks assuming a white working-class laborer as the subject of history. Better historicizing the concrete acts out of which historically and geographically specific natures are produced would not only enable the thesis to speak more directly to a range of concerns outside of the Marxist canon but also open up conditions of possibility for thinking beyond abstract natures to other more democratic environments. Ekers and Loftus (2013) have suggested that Antonio Gramsci’s absolute historicism provides valuable conceptual resources for rethinking such a position.

Opening up the labor process to a broader range of determinants also necessitates a deeper consideration of mental, affective, and immaterial laboring practices, in addition to, while not excluding, the manual labor which is often prioritized within secondary writings on the production of nature. Smith discusses the importance of these emerging divisions of labor and their influence on the production of nature and associated conceptions of nature (1984, 52–53). As elsewhere, one of the primary inspirations for his discussion is Sohn–Rethel’s (1982) discussion of intellectual and manual labor and the real and conceptual abstractions that emerge from such concrete practices. However, in subsequent discussions little sense has been given of the kinds of nature produced through the laboring
practices of the call center worker, the caregiver, or the artist. As these forms of laboring have gained greater prominence within late capitalist societies, the production of nature thesis can seem a somewhat redundant basis for considering produced environments and conceptions of the world. However, the greatest advantage of the production of nature thesis in being able to respond to such questions is the way it is framed not as a doctrine but rather as an approach through which historically and geographically specific practices might be better understood and critiqued. The production of nature thesis thus encompasses a flexibility to accommodate other modes of working and an openness to a wide range of historical and geographical practices.

SEE ALSO: Marxist geography; Nature

References


Further reading

Productivity

Andrés Maroto-Sánchez
Universidad Autónoma de Madrid and Instituto Universitario de Análisis Económico y Social (IAES), Spain

Productivity is commonly defined as a ratio of a volume measure of output to a volume measure of input use. While there is no disagreement on this general notion, a look at the specialized literature and its various applications reveals very quickly that there is neither a unique purpose for, nor a single indicator of, productivity.

Despite these controversies over its particular definitions and different uses, “productivity” is probably one of the most widely used terms by economists and noneconomists alike. Political leaders use it in their electoral programs and proposals. The media refer to it when they discuss economic performance. Trade unions and business leaders use it in wage negotiations and demands. But productivity tends to get the most attention when it is poor. That is, when one of the indicators that are normally used to analyze productivity shows very low or negative growth. Productivity is a concern because, as Krugman (1990) said, “productivity isn’t everything, but in the long run it is almost everything.” Productivity affects a country’s capacity for economic growth and its global competitiveness, as well as the standard of living of its citizens. In the long term, whether or not the welfare of an economy improves depends essentially on improving productivity.

Within the ongoing debate around productivity, the case of the services sector is particularly important, both in terms of the theory and of its political and economic applications. The developed economies are clearly service economies. Most of the production structure, in terms of both employment and added value, is directly attributable to activity in the services sector. As a result, in the long term the aggregate productivity of these economies should converge toward the rates registered by the services sector. However, not only is the productivity of the services sector important in itself, but it is also important indirectly because of the strategic effect of intermediate services – mostly producer services such as transport, financial services, communications, and certain business services – on the improvement in productivity in other economic sectors.

One problem that always arises when writing about productivity is that the concept itself is often confused. The notion of productivity has been approached from various angles and has been used for a range of purposes. This has given rise to significant ambiguity. First, the term “productivity” is not univocal. It can refer to the economy as a whole but also to a particular sector or activity, or to a specific company. Second, it can be measured based on one factor of production or on a combination of them all. It is also important to bear in mind that other terms, such as “efficiency,” “efficacy,” “competitiveness,” and “effectiveness” are sometimes used incorrectly to denote productivity.

Rethinking the concept of productivity

It is therefore necessary to clarify the different definitions that have been formulated for the
concept of productivity throughout the history of economic thought. The first to approach the subject of productivity were eighteenth-century physiocrats who referred to it as the capacity to produce. Over the years, its meaning has gradually been refined and by the twentieth century economists were referring to productivity as the relationship between the end product and the factors required for its production. This definition has been accepted for several decades: first, because it evokes what productivity is considered to be in the context of a company, an economic sector, and the economy as a whole; and, second, because it does not vary depending on the kind of production or socioeconomic system being analyzed. According to this definition, productivity is simply an arithmetical quotient that can be understood as the efficiency with which factors are used in the production process.

Based on this definition of the term, “productivity” is understood as the ability of an economic player to convert inputs into finished goods. It is therefore a relative concept, that depends on whether comparisons are made between time periods or between different units of production, or whether it is used to refer to a unit within a company, a company, an area of activity, an economic sector, or the aggregate economy.

However, the socioeconomic reality – increasingly deregulated, globalized markets; constant changes in consumer preferences and expectations; a reduction in the life cycle of services; new structures and work organization systems – is leading us to reconsider the notion of productivity. Traditionally, the term was related to productive efficiency – insofar as it analyzes to what extent the use of resources to create a particular end product is optimal – but the concept of effectiveness or efficacy is gradually being incorporated into this definition, understood as the way in which companies dynamically adapt to the needs and demands of consumers. Productivity, therefore, depends intrinsically on the value of products and services – which is why terms like usefulness, quality, convenience, availability, and originality arise – and on the efficiency with which they are produced and supplied. Consequently, this far broader definition of productivity calls for a broader, clearer set of indicators that embrace these new elements, as well as those related to processes and methods that improve productivity, to the environment and sustainable development (with the emergence of concepts like green productivity), to the management of value and supply chains, or directly to the human factor as a key element.

**Measuring productivity**

However, the description of the concept of productivity thus far does not allow for the difficulties that arise when this concept is applied to the economic reality. For this reason, the problems and biases that arise when measuring productivity have received a great deal of attention in recent years. Most current research seeks to improve the quality and quantity of existing estimates, or to present and develop new methods that overcome the biases and errors. The main elements when measuring and analyzing productivity, especially in the services sector, are those related to output estimates: adjustments deriving from improvements in the quality of the services offered and their effect on the price of the product; estimates of activities that operate outside the market, such as certain public services, or in underground (or informal) economies; international comparisons based on common terms; and problems stemming from the aggregation of factors and prices, both in terms of time and space, among others.

Productivity can be measured using various approximations, indicators, and methodological
PRODUCTIVITY

Deciding which of these to use falls to the analyst, who usually makes a decision based on the objective of the study and the availability of data with which to undertake the research. If we take productivity as an economy’s capacity to convert inputs into end products and services, it is a relative concept. Different productivity indicators will arise depending on the different kinds of productive factors analyzed.

Measurements such as **apparent labor productivity** or **apparent capital productivity**, where only one input is analyzed, are known as **partial productivity indicators**. Despite the potential problems with these – such as the fact that a change in the proportions of the composition of factors affects productivity – these measurements are useful for analyzing potential growth, so whether they are used will depend on the time frame of the study in question. In the case of services, the indicator traditionally used to measure productivity has been the relationship between production – measured through value added – and the work factor – whether that be total workers or total hours worked (OECD 2001). Nonetheless, when we try to apply this indicator to some services, its significance is questionable since the added value of many services, especially those that are not destined for sale, is practically equivalent to the work factor cost. This means that there is a direct link between the evolution of production and productivity in these activities.

Due to the limitations of these partial productivity measurements, the concepts of **total factor productivity** (TFP) and **multifactorial productivity** (MFP) have been introduced for cases where all factors of production (TFP) or more than one factor of production (MFP), respectively, are included in the indicator. These multifactorial measurements are not, however, independent of the partial measurements; they form part of the sources of growth of the apparent labor productivity. When all factors of production are included, the growth in TFP can be conceived residually as the growth in output that cannot be explained by the growth in inputs. It is in the estimation of these multifactorial indicators that most innovations have taken place in recent decades. Multifactorial productivity can be estimated through frontier analyses using both parametric techniques, such as stochastic frontiers and Bayesian estimations, and nonparametric techniques, such as data envelopment analysis. Likewise, other kinds of analysis have been introduced, such as the average response function and the use of index number theories.

Once we have overcome the problem of choosing the right indicator with which to measure productivity, further problems and issues arise. These become more significant still when we attempt to analyze the services sector generally and some producer services specifically (Maroto 2011). This has led to an intense debate in recent decades and to a case-by-case analysis of the different types of services, in an attempt to identify potential biases or errors when it comes to measurement of their productivity.

The traditional estimation technique used for other economic activities, based on double deflation, is not the most appropriate for service activities. The use of double deflation incorporates various problems that would be inconsequential or nonexistent in the production of manufactured goods. But, in the case of services, one of the main problems is how to account for intensity or quality. To account for these within the productivity indicator, we would have to add two more indices to the index of services production: one related to the complexity of the set of services and another related to their intensity. Another aspect to bear in mind for many services, especially those related to the transfer of knowledge such as education, health care, research and development, and some business services, is that their provision gives rise
to a series of results in the medium and long term. For example, a lawyer could charge for his or her services based on direct output (legal work undertaken as reflected by the services provided and the time taken), but also on the basis of the results of his or her output (a proportion of the payout obtained by the client through the lawsuit). It is often difficult, however, to distinguish between direct output and medium- and long-term effects.

Traditional view of services and low productivity

Regardless of the indicator used to measure productivity and the aforementioned issues with these indicators, the literature traditionally cites the relatively low productivity of the services sector as one of the key factors behind its growth in modern economies. This idea is fundamentally based on analyses developed from the seminal work of Baumol (1967). These attribute the differences in productivity in services to the role played by labor. Baumol’s cost disease, as this phenomenon is widely known, if it were to take place, would result in a slowdown in economic activity and the aggregate productivity of the advanced economies, along with an increase in service prices.

The empirical evidence at an international level for the developed economies for different time periods appears to indicate a negative relationship between the size of the tertiary sector – both in terms of occupation and of production – and the growth of aggregate productivity. Therefore, for the services sector as a whole, this would appear to confirm conventional thinking about its relatively low productivity. However, the different ways in which services are defined and measured and how they are included in the productive process could have an impact on the results. Some of the effects many services have on aggregate productivity are undervalued, especially for producer services such as communications, transport, and financial and business services.

More positive perspectives on services and productivity

Since the mid-1990s there has been progress toward a less negative view of services and productivity. This been driven fundamentally by two elements. First, traditional theses have been adapted to take account of the kinds of service being analyzed. Many authors state that the aggregate nature of some studies leads to the real productivity of tertiary activities being undervalued. A microeconomic analysis would therefore be the most appropriate option. Second, new ideas have been introduced that contradict, or at least limit, the traditional theories. Some authors even go so far as to state that Baumol’s cost disease has been cured (Triplett and Bosworth 2004).

One of these new ideas is to incorporate the role of innovation in services productivity, especially in some high-growth activities. In addition, related innovation, the role of knowledge, and human capital have been emphasized, as they facilitate significant increases in productivity in many tertiary activities, especially knowledge-intensive business services (KIBS). Another important factor helping to revise the image of many services is the positive indirect effects they have on other productive sectors. This is the case for activities that are aimed at intermediate demand and are highly weighted within the intermediate consumption of other highly productive sectors, such as logistics services and transport, telecommunications, information technology (IT), many advanced business services, research and development (R&D), and design services. All these services, which are also
closely linked to information and communications technologies (ICTs) and the new economy, have productivity levels and rates that are not only comparable with, but even higher than, those observed in some of the most dynamic manufacturing industries (Maroto and Rubalcaba 2008).

Finally, another of the strongest lines of revision in recent years is the idea that the low productivity of many services is merely a measurement issue; the data for many service productivity analyses is simply undervalued. The effect of this bias will depend on the weight that poorly measured tertiary activities have in the productive landscape as a whole. Among the best examples are the many producer services whose real effect on the aggregate productivity of the advanced economies is far greater than that expressed by traditional statistics, owing to not only progress in innovation and knowledge in those services, but also the indirect stimulation they provide for other productive sectors.

Conclusion

In summary, the problems of measuring the productivity of services stem from two sources. First, there is an issue around how to estimate inputs. For example, it is important to measure the work factor through the number of hours worked, especially in those services where there has been a reduction in the number of hours per worker or where part-time or seasonal work accounts for a substantial proportion. Likewise, it is important to bear in mind the relationship between the work factor and other intermediate consumption, especially in services such as commercial retail services or financial services, where outsourcing and offshoring have become more widespread in recent years. Second, output estimates could also be improved. In the case of services, for example those related to ICTs, these improvements would stem from a differentiation in the calculation of constant prices, between price increases derived from purely inflationary processes, and between price increases that are attributable to improvements in the quality of services. However, not all of these are easy to analyze. There are therefore very few studies that comprehensively examine all of these measurement biases. For this reason, work is currently under way on alternative measurement procedures or improvements to the existing statistics on the productivity of services.

However, the evidence to date can only give an initial indication of the extent of the measurement biases and their effect on the calculation of productivity variations. It does not resolve many of the doubts about the reliability of measures of productivity in services. Possible channels for future improvements could arise from the wider application of the best practices developed by some countries (such as Canada, the Netherlands, and some Scandinavian countries) for the measurement of services productivity, and from a stricter and more careful application of existing production models to some service activities such as health care, transport and communications, and business. Some countries have begun to take steps to improve these measurements and international organizations are working – along with their member states – on various specific sectors, such as finance, insurance, and IT services. Nonetheless, it is clear that much more effort is required if we are to achieve accurate measurements of productivity for the services sector in general and for producer services in particular. Only then will we reach a better understanding of the factors that drive services productivity and the differences that exist at an international level.

SEE ALSO: Competitiveness; Factors of production; Industrial geography; Innovation
PRODUCTIVITY

and regional development; Labor geography;
New economic geography

References


Property and environment

Nicole G. Graham
University of Technology Sydney, Australia

Definitions

The property regime of a human society is the regulatory mechanism that controls the distribution of real and abstract things both within and beyond that society. Consequently, concepts of property profoundly affect, and are affected by, the culture and the environment of a society. Property regimes are arguably the most important of all regulatory mechanisms with regard to the environment because they articulate and actualize culturally acknowledged principles of, and processes for, the human use and disposal of various environmental elements and processes. The specific character of a property regime, its method of distribution, is determined in part by political models and in part by environmental conditions. Property, from the broadest view, also necessarily implies and requires corresponding corrective processes (redistributive and/or punitive) for instances of improper distribution, access, use, and disposal of the environment. The numerous histories, philosophies, and geographies of property are so fundamental to human economies and cultures that they encompass the entire history of the relationship between human societies and the environment.

Ideas of property create, protect, and defend culturally specific environmental epistemologies. For example, cultures in which the environment is regarded principally as a set of economic resources have property regimes that position humans as separate and superior to the environment. These property regimes articulate property primarily (but not exclusively) in terms of entitlement (rights). In contrast, cultures in which the environment is regarded partly as a set of economic resources and partly as objects and processes invested with noneconomic cultural, ecological, and/or metaphysical significance have property regimes that position humans both within and connected to the environment. These property regimes articulate property primarily (but not exclusively) in terms of obligations (responsibilities).

In comparative legal terms, property is understood as a set of principles concerning access to and disposal of things. Access to environmental things, as objects of property, is described variously in terms of inclusion, exclusion, and possession (contingent on the status, permissions, transactions, and privileges of social membership), rights (actual, future, potential, mediated, presumptive, subsidiary, individual, familial, tribal, communal, sovereign, sui generis), responsibilities (customs, duties, and obligations), and security of tenure (relation to other property interests). Disposal of property is described variously in terms of alienability (transferability, assignment, sale, and gift) and waste (permissive, voluntary, and externalities).

Categories of property

There are many different kinds and categories of property across time and place, more than can be described here. The following familiar categories are neither exhaustive nor mutually
exclusive. However, they do provide a helpful basis on which to consider some key features of property regimes generally and in relation to the environment specifically.

Open access

An open-access regime is not properly a property regime because it is characterized by the absence of regulation. Open access is in fact an unregulated regime in which a thing is regarded as being simultaneously the property of everyone and no one, that is, a thing to which all may have access but for which none bears responsibility. The reason that a thing is so regarded is that it is an abundant and inevitable geographical feature of living in that particular place (for example, snow in Alaska), or that human society has yet to successfully remove it from this category and place it a different category with restrictions of access and use. In legal terms, there are no rules (of exclusion and alienability, for example) regarding the access, distribution, use, and disposal of environmental resources within an open-access regime. In political terms, open-access regimes can be the source of intense and ongoing political conflict due to competition for the resource in circumstances where the resource becomes scarce (seasonally or permanently) and no responsibility is taken for the health of the resource base and/or for social stability.

Historically, human societies have determined open access to (and use of) parts of their own (or another community’s) local habitat for food and fuel which were believed to be, but were not in fact, infinitely abundant and inevitable geographical features of their environment. It is thus often the case that things that were once regarded as objects of an open-access regime were so regarded on the basis of ecological misunderstandings and attendant economic miscalculations, or political immaturity (through which ongoing social conflicts over the objects remained unresolved). For this reason, open access is increasingly regarded as a regulatory failure rather than as an evidence-based economically sustainable property regime. Another, longer-term view is that because human knowledge of certain environmental resources is acquired through economic experience with those resources and because that knowledge necessarily develops over time, open-access regimes may be more generously regarded as indicative of immature or developing knowledge bases rather than regulatory failures per se.

Once an environmental resource is recognized as no longer being (or never having been) abundant, renewable, or easily reproduced, it is common for societies to remove that resource from this category and thereafter restrict and regulate access and use of the object. For example, marine resources (fisheries) beyond the jurisdiction of individual states are no longer regarded as an open-access environmental resource. Modern laws of the sea with regard to the distribution and use of marine resources have replaced open-access regimes whereby societies depleted and/or exhausted fish (including whale) stocks.

The property regimes of nations and communities colonized by Western European imperial forces since the sixteenth century were often mistakenly regarded as open-access regimes. Indeed, it was often precisely on the basis of this erroneous classification and the corresponding international law doctrine of *terra nullius* that those nations and communities were colonized.

Common property

Common property regimes, also known as “the commons” and “community commons” are those in which entitlement to and responsibility for certain environmental resources are held in common by members of a certain society.
Common property is determined in part by the agreement of a society (with a defined membership) to both enable (entitle) and restrict (prohibit) its members to access, use/consumption, and disposal of certain environmental resources at certain times and under certain conditions. Common property is determined in part also by the geographical conditions and limits of those resources. Common property is thus a regime that can control the allocation and distribution of environmental resources effectively over long periods of time (for example, centuries) because it combines (evidence-based or experience-based) human knowledge of environmental resources with locally adapted and highly particular prescriptions of access, use, and disposal of those resources. For this reason, common property is often regarded as suitable for environmental resources that are less abundant than open-access resources, or less resilient to unregulated access and use. For example, in parts of pre-Enclosure England, the timber branches, sticks, and twigs of certain species of tree already fallen to the ground or easily pulled down by a simple hook were regarded as common property. The corresponding corrective rule applicable to the same context prohibited the taking of branches of certain trees if that involved using a sharp utensil such as an axe (Bellamy and Williamson 1987).

However, while common property regimes often share the quality of providing environmentally sustainable regulatory practices, there is a wide diversity of sociopolitical structures at their foundation. For example, some common property regimes are characterized by egalitarian structures while others are hierarchically structured. Successful common property regimes (those that have endured over the longest periods of time) have been associated with small human populations and/or strict administration. For example, the violation of common property laws regarding the killing or taking of a certain animal, or the use of freshwater outside the fixed (and sometimes biologically determined) rules of membership or outside the appropriate (seasonal or cultural) timing, have in some regimes been severely punished. Rules of exclusion and alienability are central to common property regimes. The access to and use of common property is restricted to the members of the society owning the resource and excludes all others (noting certain exceptions for cultural purposes including hospitality and economic exchange with other societies). Common property is by definition not individualistic and so the alienation (transfer through gift, trade, or sale) of common property by individual members of that society is therefore generally prohibited by its operation.

The property regimes of nations and communities that were colonized by Western European imperial forces were often held in low esteem, maligned, or not recognized at all for predominantly colonial purposes (for example, the doctrine of *terra nullius* was the basis for the assertion of British sovereignty in the colony of New South Wales in the eighteenth century). However, since then, many of those property regimes have been belatedly recognized, and some have been classified as common property regimes. Owing to the association of successful common property regimes with environmentally sustainable economic practices, indigenous property regimes in former colonies are occasionally conflated with environmental knowledge that is regarded as somehow inherent to an indigenous society on the basis of race, rather than on long-standing experience and evidence-based economic expertise with certain environmental resources in particular places. The tendency to essentialize and conflate indigenous property regimes and environmental knowledge (and to identify indigenous peoples with nature in
the Enlightenment’s nature/culture paradigm) arguably perpetuates the misunderstanding and misdescription of indigenous laws particular to the “other” of colonial discourse (Fitzpatrick 1992; Krech 1999).

Common property regimes can be used to resolve persistent tensions and dilemmas about the access to and use of critical environmental resources, and have been used in hybrid forms within capitalist economies. For example, in the Murray–Darling Basin of Australia, irrigation cooperatives and corporations were established by the New South Wales government in 1994 to transfer the ownership and management of irrigation schemes in the Basin from the state to the irrigators themselves. Where once farmers were water entitlements holders under a state-owned irrigation scheme, now they are shareholders owning and operating cooperatives and corporations. Water-sharing plans regulate the distribution (including the trade) of the water in the catchment to and between the members of the cooperative or corporation.

Crown, state, and public property

Crown, state, and public property regimes include a broad category of different interests in land and environmental resources that are defined by the fact that they are controlled by the sovereign of a given jurisdiction (whether monarchy, presidential systems, parliamentary republics, theocracies, or one-party states). Indeed, these property regimes do not exist in the absence of a legally constituted and observed sovereign power. Most property interests within this category exist within hierarchical political structures and are abstract rather than material (environmental) in origin, logic, and extent. Public property, unlike open-access and common property regimes, has been said to be nonrivalous because the resource base is not diminished by the use of others. For example, whereas the taking of 20 fish (from an open-access or a common property regime) by one person will decrease the fish stock available for others, the walk through a national park by one person does not decrease the quantity or quality of walks available for others. However, even public property (beaches, parks, rivers, and state-owned mines) is subject to diminishment of the quality and quantity of resources through human use over time.

In England and in Commonwealth countries, the feudal doctrines of tenures and estates determined that the ownership of all land was vested ultimately in the Crown through whom smaller private tenures (landholdings) and estates (interests defined by time) were created and distributed to the Crown’s subjects. In addition to being the ultimate “owner” and “radical title holder” of all land within the jurisdiction, the Crown also reserves land to itself called Crown Land. Crown Lands are now administered by legislation which determines access, use, and disposal of these lands and the resources therein. For example, in (the Commonwealth of) Australia, extensive leasehold interests in Crown Land include mining leases and pastoral leases to facilitate and control specific land uses. Crown Land includes but is not limited to public property. Public property in monarchical societies owes its origins in part to common property as after the enclosure of common property (the Enclosure of the Commons), land and resources that remained in common usage, such as parks, became regarded and eventually regulated (through the local municipal government, for example) as public property. Crown land is inalienable (it cannot be transferred by gift or sale other than by the Crown itself). Access to and use of the various kinds of Crown Land depend on the relevant governing legislation and so the
Crown ultimately determines whether, when, and who to exclude from that access and use.

Unlike the tenurial system of Commonwealth jurisdictions, the property regimes of presidential systems and parliamentary republics are allodial (land without a lord). Allodial land interests are the interests of a private citizen (rather than a subject) and are not derived from the Crown. As such, they are regarded as a form of absolute ownership. However, such private property interests, while enjoying legal and cultural protection through constitutional law and powerful political lobby groups (property rights movements), are ultimately creatures of political hierarchies and are accordingly subject to acquisition by the state from time to time. In recent years the extent to which the state does compulsorily acquire or “take” the private property of its citizens for broader environmental benefits has become the subject of significant debate. State property includes property that was never successfully claimed by private interest holders and property that has been acquired compulsorily (or was never available to the private sphere) by the state. This is called compulsory acquisition or takings. State property includes but is not limited to public property. Public property in presidential systems and parliamentary republics is similar to public property in monarchies in that it is state lands and environmental resources that are reserved for public access and direct use (such as national parks and cultural heritage sites). State property that excludes, prohibits, or strictly limits public access and uses, such as military bases and nuclear power plants, is invariably rationalized by the state in terms of the ultimate benefit (economic or security, for example) to and on behalf of the public, its citizens.

Private property

Private property is a regime in which the access, use, distribution, and disposal of land and environmental resources is controlled not by the Crown/state, a community, or a collective, but by individual entities (which can be natural persons or legal persons such as a company) through the notion of an entitlement or “right.” The twin rights that characterize private property are the right to exclude all others from the land or resource, known as “exclusive possession,” and the right to alienate (to transfer by gift or sale) the land or resource.

Because private property consists in multiple rights (the right to possess, the right to exclude all others, the right to alienate), it is common for private property rights to be shared between different individuals. For example, the ownership of a parcel of land may be shared between co-owners in various proportions and under various conditions; or the right to exclusive possession may be sold temporarily to other individuals through lease arrangements or permanently through easements of different kinds. There are many other ways in which private property in a single parcel of land or resource is fragmented through legal mechanisms so that the “ownership” of something becomes complex and complicated for those wishing to interact with, experience, or impose conditions on its access and use. In theoretical terms (to be explored in more detail below), this complexity is referred to as a “bundle of rights” in which each “stick” within the bundle indicates a different right or interest in the object of property. Some interests pertain to the abstract dimension of time (such as future, current, reversionary, and redemption interests) while others pertain to the more material dimensions of space or place (such as easements of right of way, or covenants to not build something, or licenses to use or take certain things from the land).
The legal apparatuses of private property regimes regulate the normative and discursive conditions through which objects are commodified, exchanged, and accumulated as wealth. Private property is therefore a category of property that is indispensable to the purpose and operation of capitalist economies. Accordingly, private property rights have been afforded strong legal protection and legal priority within the overarching hierarchy of legal rights and obligations in a given legal order, thus stabilizing and then entrenching the normative dimensions of private property. For a long time, private property rights were regarded as absolute rights free from external (usually government) interference and unshackled by obligations to the object of property as well as to all other non-owners of that object of property. However, in practice, private property rights are not absolute as they have always been subject to the demands of social and technological change. The acquisition of private property by the state or Crown is enshrined in the constitutions and conventions of several jurisdictions and has been long practiced to achieve, for example, improvements in infrastructure (through the construction of railways, freeways, airports, etc.) and the realization of mineral wealth beneath the land’s surface (through the legal fragmentation of land ownership into state and private strata).

Since the late twentieth century, private property has been used as an instrument of public environmental protection policy through the extension of the concept of private (and therefore tradable) property rights to environmental phenomena such as biodiversity (bio-banking and offsetting), pollution (carbon credits), water, and genetic material, for example. The extensive use and strong policy preference for the abstraction of environmental phenomena into environmental assets, environmental services, and environmental markets is consistent with the logic of commodification within capitalist economic theory. However, private property regimes have also been identified as a major legal and cultural agent of adverse anthropogenic environmental change (Sax 2008; Graham 2011). It is in this context that the use of private property rights in environmental policy is the subject of strong debate, and the extent to which environmental policy can further develop without evaluating the fundamental role of private property in environmental change (for both sustainable and unsustainable outcomes) is unclear.

Private property regimes have been closely related to the forceful termination of the property regimes of indigenous nations and communities through acts of dispossession and declarations of imperial sovereignty. The expansion of the British Empire throughout the seventeenth and eighteenth centuries depended heavily on private property concepts, including the notion of exclusive possession, which were applied in non-British jurisdictions against the different property regimes in operation. The notion of exclusive possession, that is, that land cannot be shared, is central to the operation of private property, and it was this way of thinking about property that made it possible for imperial forces to justify to themselves the dispossession of First Peoples from their lands physically and in abstract legal terms. In the twentieth century, private property regimes were radically challenged by native title claims. However, invariably native title has been regarded as incompatible with private property rights due to the right of exclusive possession, and private property is therefore legally immune from native title claims, which can only be made over nonprivate property.

Private property in the environment can logically include any and all of the components and processes of all of the Earth’s systems even prior to any necessary technological innovations designed specifically to capture or harness them.
This is because, owing to the abstractness of the notion of a private property “right,” it is possible to create property interests in abstract things, including things that (may) exist in the future, ideas, and innovations. Similarly, property in the environment is not limited to property in the elements and processes of the Earth, but may extend (subject to technological possibilities) to the Earth’s solar system and beyond. It is important to acknowledge also that broader definitions of environment, including contemporary ecocentric definitions that situate humans within the environment, and anachronistic colonialist definitions that conflate colonized peoples with “nature,” have facilitated private property in human beings (alive, deceased; fluids, cells, tissue, and body parts). Private property in the environment can mean therefore property in almost anything.

Stewardship and custodianship

Beyond the categories of property described above, there are also property regimes known as land or country stewardship and custodianship. Western property regimes articulate property in terms of “rights” (entitlements) and “negative duties” (to avoid certain behaviors, for example, restrictive covenants in private property regimes; prohibitions of access to and use of certain species of plants and animals in common property regimes). However, the property regimes of custodians and stewards articulate property in terms of “caring for country” through locally particular land-use practices such as fire-stick farming (Altman and Kerins 2012) and responsibilities (Lucy and Mitchell 1996). Western legal discourse would translate “caring for country” as “positive duties.” Stewardship and custodianship property regimes are highly culturally and geographically particular and their legal apparatuses vary significantly across time and place. However, both share long histories and similar environmental philosophies of a nonhierarchically structured connection between humans and the environment. Indeed, the characteristic intellectual premise of these regimes is that humans and the environment are epistemologically and ecologically integrated, rather than separate. This philosophy is sometimes apparent in the religious and spiritual discourses of societies with stewardship and custodianship property regimes which anthropomorphize environmental elements and process, and which adopt totemic systems and embed totems with enduring economic, social, and spiritual significance.

Theories and critiques of property

Overview

Owing to the broad reach of property as a key concept in cultural and legal discourses and as a powerful social institution and regulatory mechanism, it has been the subject of the research of scholars from multiple disciplinary backgrounds. Conventionally, theories and critiques of property have been anthropocentric in focus, emphasizing the role and means through which property organizes the structure, quality, meaning, and experience of human life and human society (Davies 2007). Such work can be found largely within and across economic, sociological, anthropological, political, geographical, and legal scholarship. Given the predominance of capitalist economies in the anglophone world, a great deal of property theory written in the English language deals with questions of distributive justice, economic efficiency, and personal liberty arising from the ubiquity of private property regimes.

However, property theory of the late twentieth and early twenty-first century has increasingly broadened its scope beyond the human world, to include a focus on the relationship between
property and the environment. Such scholarship considers the environmental potential of extending the concept of rights at law to things in nature (Stone 1972; Burdon 2014) and of redeveloping the current model of private property as a “bundle of rights” into a more environmentally sustainable “web of interests” (Arnold 2002). It also critiques the adverse environmental legacy of the dominant paradigm of private property as a legal institution designed to incentivize limitless growth (Sax 2008; Alexander 2011). Contemporary property scholarship has also pointed to the ways in which private property regimes have been based more on abstract principles than on material conditions, which adversely affects the viability and durability of those regimes (Sherman 2008; Graham 2011).

### Labor, improvement, and growth

Theories of private property broadly classified as liberal have presented rationales for and defenses of private property regimes. However, despite the predominance of the anthropocentric lens throughout this theoretical tradition, the environment, referred to as “nature,” constituted the intellectual and material basis of these ideas. The gradual transition from feudalism to capitalism in England depended heavily on the rise and acceptance of a new concept of property, which in turn was based on a new concept of nature. Consequently, liberal property theories date back to these profound theoretical changes. The work of John Locke in the seventeenth century captures the broader socioeconomic and political debates of his time relating to the enclosure of the English commons and the colonization of North American lands and peoples. These debates were articulated through the discourses of improvement and progress, both of which carried aspirational sentiments regarding the place of humans in the world and particularly of (certain kinds of) people transcending nature, becoming cultivated and cultured in contradistinction to their primitive counterparts (Fitzpatrick 1992; Brace 2001).

John Locke’s theory of property was premised on the idea that the environment, as “nature,” was without inherent value, chaotic and disordered, from which state human intervention would improve and order it through cultivation. From this premise, Locke argued that improving land (through certain kinds of agricultural land use) provided the moral basis for a claim to private entitlement to that land. His “labor theory” of property was used to defend the enclosure of the commons and the dispossession of First Peoples and the colonization of their lands. The nexus between humans and the environment in the “labor theory” of property is the greatly increased quantum of agricultural land use. It is precisely this point that links the labor theory of property with the idea of a growth economy. Locke’s property theory explicitly advances the idea that land should be used to exceed economic need rather than to meet it, so that profits become available. There is a large literature defending and critiquing Lockean property theory, most of which relates to questions of distributive justice and political and economic philosophies. However, more recently scholars have revisited Locke’s provisos on waste, in particular to examine what restrictions, if any, there are in his original work regarding the environmental limits of property and environmental protections. Economic theorists critiquing the current growth economy are concerned to question its intellectual origins and for this reason it is important to review the work of Locke and his contemporaries.

Locke has cast a long shadow over property theory. It has been 400 years since he anonymously published his justification for the dispossession and enclosure of the then common and community property regimes in place.
in England and North America. Despite this, the idea that the labor and cultivation of land improves both the environment and the society in which those land-use practices are adopted prevails. Even up to the early twenty-first century the case for economic growth and development has been so long advocated as to have become normalized. Theories of property that question this orthodoxy are either part of, or at least perceived to be, the counterargument of antiliberalism known better as Marxism, socialism, and materialism. However, it is important to note that environmental critiques of liberal theories of property are not always, or even predominantly, concerned with this classical framing of the property debate in exclusively political terms. Environmental critiques may or may not be also concerned with distributive justice and other political issues at the heart of critiques of liberal property theory. What is radical about some environmental critiques of liberal property theory is that their various degrees of nonanthropocentrism (from ecocentrism to network and relational theories) indicate a shift away from debating property in the exclusive terms of human experience towards debating property in broader ecological terms that include humans but do not place them at the center of the epistemological model on which the debate is founded.

Materiality, sustainability, and connectivity

Since Locke’s advocacy and justification of private property regimes, and since their proliferation within and for capital-growth economies, property theory has become increasingly and remarkably abstract and detached from material and, especially, environmental concerns. Some property theorists’ work deals directly with the abstractness of the dominant private property regime. In legal scholarship, this abstractness is referred to sometimes as “dephysicalized” property. The origins of dephysicalization are traced back to Locke, although more commonly it is the work of English lawyer Jeremy Bentham in the eighteenth century and American lawyer Wesley Hohfeld in the early twentieth century that is cited in scholarly enquiries into the nature of property’s dephysicalization. The perspective of proponents and analysts of dephysicalized property alike is that private property relations are not about real things in the environment, but about abstract legal rights between people. The emphasis on rights and on the abstractness of property relations has been the source of much scholarly analysis, with some theorists arguing that dephysicalized property is politically unstable and others arguing that the emphasis itself is misplaced. Nineteenth-century German economist Karl Marx and twentieth-century American philosopher Hannah Arendt insist that private property regimes are certainly framed within the discourse of abstract rights, but that their conditions and effects are inevitably material and specifically environmental (in which they include humans). Both Marx and Arendt caution against forgetting the materiality of real property as things in the world in the interests of sustainable economies.

However, beyond the original writings of Marx and Arendt, and like liberal theories of property, Marxist critiques of liberal property theories are largely anthropocentric in focus, and deal mostly with the distributive justice questions of political theory (liberty and individual personhood) and economic theory (the efficiency of natural resource use). Some of this theoretical work examines directly the material conditions and effects of private property in human terms such as social inequality, poverty, homelessness, and diaspora linked to the privatization of lands and environmental resources. More recent attention to questions of distributive justice has also included consideration of the impact of climate
change on developing countries and island states, referred to as environmental justice. A distinctive recent trend in this literature has begun to deal directly with material conditions and effects of private property with specific regard to environmental considerations. Interestingly, much of this emerging literature is produced not by legal theorists but by human geographers, anthropologists, and environmental economists whose work is influenced directly and indirectly by Marxism and the philosophy of materialism generally.

Since the 1970s, environmental law and environmental governance scholarship have also contributed substantially to property theory directly through sustained and rigorous attention to the potential and effects of adopting market mechanisms in environmental regulation methodology (Freyfogle 2003). Most of this literature frames the discussion in terms of property rights and environmental sustainability. However, while some of this literature promotes the use of private property rights to enhance greater natural resource-use efficiencies (for example water trading for irrigated agriculture) and to resolve environmental problems (internalizing the previously regarded “externality” of industrial pollution through pricing units of atmospheric carbon), other literature concerned with the use of property rights for environmental sustainability outcomes critiques the logic and/or the outcomes of this increasingly common policy approach.

Both the supporting and the critical literature on property rights and sustainability from an environmental perspective engage explicitly or implicitly with the issues raised nearly 50 years ago in Garret Hardin’s 1968 essay “Tragedy of the Commons.” That essay incorrectly refers to an open-access regime as a common property regime (an error which several legal, economic, and geography scholars have discussed at length). Putting aside this error of classification, the issue Hardin identified remains that anthropogenic environmental problems are indeed created by regulatory failure and almost always by inappropriate property frameworks. While Hardin’s essay contended that the use of private property regimes offered the most environmentally sustainable outcomes, almost any property theorist and environmental researcher would agree that any property regime providing natural resource access and use restrictions and prohibitions is superior to an open-access regime. The conventional political analysis that compares private, common, and public regimes has tended to overshadow some of the more complex environmental management questions raised in Hardin’s essay, which remain as urgent today as they were then.

What Hardin’s essay has shown, in addition to the significance of property to anthropogenic environmental crises generally, is, inadvertently, the importance of transdisciplinary education and integrative thinking. Indeed, theoretical and critical property scholarship that is transdisciplinary has revealed the environmental potential of redirecting property research from its predominantly anthropocentric focus towards an alternative approach that is significantly decentered. Network theory and the ideas of relationality and connectivity appear increasingly relevant to this most recent body of research concerned with the intersection of property and environment. Some of this work derives from in-depth analyses of the property regimes of indigenous peoples and links the notion of custodianship with connectivity between peoples and places (Weir 2009). Other work derives from discursive analyses of regulatory frameworks constructed on the premises of individualism and the precept of autonomy, advocating greater attention to relationality and responsibility (Nedelsky 2011). Much of this work
emphasizes the intellectual and practical possibilities of rethinking the human experience (with each other and in the world) beyond the terms of Enlightenment philosophy (progress, improvement, growth, autonomy, possessive individualism).

This growing trend in recent property theory suggests that the incorporation of geographical knowledge, and specifically of environmental conditions and limits, into property laws, and the alignment of property and environmental laws rather than their taxonomic separation, could address some of the current challenges to private property regimes – for example, the inadequacy of the abstractness of private property rights to helpfully contribute to the pressing environmental policy issue of the physical erosion of coastal lands. Given the strong influence of private property regimes (and the attendant ideas of progress and growth) on natural resource management and policy, it is likely that environmental problems will continue to be examined by reference to the relationship between property and the environment. Indeed, it is likely that new categories of property will emerge to replace those described above through unilateral and multilateral efforts to adapt to changing climatic circumstances in particular. The role of integrative transdisciplinary innovations of both the cultural discourses and regulatory institutions of property cannot be underestimated to this effect.

SEE ALSO: Climate change adaptation and social transformation; Commodification of nature; Community-based natural resource management; Ecological imperialism; Ecosystem services; Environmental governance; Environmental regulation; Indigenous knowledge

References


PROPERTY AND ENVIRONMENT


Psychoanalysis is a form of inquiry about society, a therapeutic practice, and a theory of personality emanating from the work of Sigmund Freud (1856–1939). The legacy of Freud’s ideas, which developed throughout his lifetime, has been hugely influential on modern thought and the ways in which we understand contemporary society, our relationships with each other, and human mental functioning. Since the publishing of his work, which spans nearly 50 years, Freud’s theories have been enthusiastically embraced, refuted, rejected outright, molded, and reshaped by many including Carl Jung, Melanie Klein, and Jacques Lacan. The results can be seen in different regional, national, and cultural settings across the world where distinct (re)interpretations and forms of psychoanalysis have emerged – from the British Kleinian and object relations school of thought and the (French) Lacanian linguistic formulation, to the American ego–psychological paradigm, the Latin American Kleinian and Bionian perspective, the Franco–German philosophical and hermeneutical contingent, and so on.

While each school of thought may have its own theoretical and professional foci, one thing that unites these different traditions is a convergence on the existence of the unconscious. This dynamic part of the mind strongly influences our conscious lives and our sense of identity and how we become aware of ourselves and, as a consequence, of who we are. For Freud, this part of the mind lay outside the boundaries of consciousness and was constructed in part by the repression of ideas and thoughts that were too painful or dangerous to be allowed to enter the conscious mind. Some of his earliest works examined how the unconscious, memories, and “phantasies” were indistinguishable. He soon realized that not everything in the unconscious was repressed, but that these things were just not expressed at the time.

In 1923 Freud published a book on a new dynamic model of the mind, known as the structural theory of the mind. Although it did not contradict or exclude his earlier topographical descriptions, this work was an attempt to describe the way in which the whole mind system determines personality and its main motives, drives, and desires. According to Freud, the mind is made up of three parts: the id, ego, and superego. The ego is the mediating part of the psyche that reacts to external reality and that which a person sees as “self”; the id is the unconscious part of the psyche that is concerned with inherited, instinctive (primitive) impulses; and the superego is the part of the mind that acts like an “inner parent,” acting as a social conscience that reflects society’s cultural norms and values. The ego and superego are essential for keeping the id in check.

For those curious about psychoanalysis an obvious question follows: How do we know that the unconscious exists? While there are some differing understandings, evidence of the unconscious can be experienced and felt through dreams, slips of the tongue (parapraxis), Einfall (the German word for mental content that simply drops into conversation but is not necessarily invited), and our “gut feelings” or intuition.
about situations, people, and places. Sometimes we may do something that is out of character, that is not entirely expected of us or “rational.” We may, in conversation, intuit how somebody else is feeling (maybe even finish their sentence), subtly building our unconscious communication with one another. Sometimes our feelings of envy, guilt, aggression, rivalry, and so on may seem incomprehensible or appear from out of the blue. Phenomena may occur that cannot be explained by rational knowledge, highlighting theoretical contradictions that, according to Thomas Kuhn and Karl Popper, indicate the extent to which our existing knowledge is always probabilistic and provisional. The way we make decisions and navigate our lives, therefore, intrinsically connects to the kaleidoscope of our inner worlds. This process does not happen by simple cause and effect, from inside to outside, but through a number of complex processes such as projection, denial, resistance, splitting, and so on (Kingsbury 2009). These psychoanalytic terms attempt to give a name to what is really going on for us, and to explain how our behavior is rooted in deeper unconscious communications.

In a clinical setting, psychoanalysis has been fundamental to the development of a number of psychological therapies. Psychotherapy, coined an “old art and a new science,” is one treatment and method (among many) that is directed toward the psyche. While analysis does not purport to be a quick fix (it can take years), the framework for therapy is rooted in the relationship between patient and therapist. It uses the most pragmatic form of material available – communication – both spoken and somatic. In other words, it is not just what is said by the patient but how it is conveyed (including body language, silences, hesitations, and repetitions). The therapist and patient attempt to unravel and untangle the unconscious workings in the consulting room as a microcosm of the patient’s world. In this therapeutic setting the patient may play out his or her concerns, feelings, thoughts, and wishes, which the therapist in turn holds, listens to, and re-presents in a manageable form. While there are considerable divergences in theory and approach, it is through this complex interaction that analysis purports to bring unconscious material to light and to help relieve the patient’s problems.

One might argue that, by understanding and making sense of the patient’s world, therapists become well versed in geography (Kingsbury and Pile 2014). The experiences, coincidences, and everyday narratives retold in the therapeutic encounter are (slowly) unfolded and mapped by therapist and patient together. Creating a space in which both therapist and patient are allowed to think is of central importance to the psychoanalytic endeavor. Geographers, in turn, have found psychoanalysis to be a rich resource, despite the different aims of each profession (Kingsbury and Pile 2014). While much controversy exists about Freud’s topology of the mind, the idea that the unconscious destabilizes human agency as rational, knowing, and predictable can be very useful to human geographers. Geographers have engaged with psychoanalytic theories since the 1930s, although it was during the late 1990s that psychoanalytic ideas in geography became more prominent, leading to the inaugural entry of psychoanalytic geography in the *Dictionary of Human Geography* in 2000.

Geographic engagement with psychoanalysis has not always been a comfortable journey. As Chris Philo and Hester Parr (2003) once remarked, geographers are “squeamish” about concepts such as sexuality, threat, desire, repression, and abuse, as well as cautious about the unconscious as “raw material” for geographical inquiry. The task of assimilating psychoanalytical concepts and ideas into geographical inquiry
PSYCHOANALYSIS/PSYCHOANALYTIC GEOGRAPHY

presents a challenge when developing understandings of spatialities and subjectivities that do not necessarily fit together neatly. In addition, Liz Bondi highlights how there is much confusion about different approaches: psychology, psychotherapy, and psychoanalysis have often been equated unproblematically, without recognizing their respective diversifications (as well as similarities) between and within theory and practice.

Concerns and critiques have been valuable, however, in developing psychoanalytic geographies. Geographers have continued to ask what the unconscious might mean for geographical inquiry, coupled with how psychoanalytical “material” may be used in geography. Conversely, geography is currently recasting psychoanalytic concerns by working with active therapists to think through spaces of trauma and distress, and notions of therapeutic space. Such collaborations indicate that geographers are also having an impact on psychoanalysis, including in the domains of neuropsychoanalysis and psychiatry (Callard 2014).

Building on earlier feminist work in human geography, a rethinking of the politics of positionality in geographical research has raised important (ontological and epistemological) questions about the production of knowledge. Through a psychoanalytic lens, the ontology of the subject is somewhat unknowable and requires new thinking (Thomas 2010). If we acknowledge the unconscious, then our knowledge about the world and of thoughts is not transparent and straightforward: the unconscious processes of the research encounter (of subjects and ourselves) create a number of epistemological issues, raising questions about truth claims, relationships between data and theory, and how to appropriate subjects’ narratives. There are no easy answers. Nonetheless, psychoanalytic geographers continue to formulate practical resources in response, advocating a research alliance that has “intimate distance” between participant and researcher and that draws on earlier formulations in psychoanalysis (Pile 2010); and providing a framework to conceptualize narratives rather than inferring personalized unconscious workings (Thomas 2010).

Liz Bondi has emphasized how empathy and identification can enhance geographical understandings of fieldwork. Drawing on Winnicott, Bondi explores the differences between empathy and identification and how reflecting on these can enrich existing methodologies. For example, empathy involves a capacity to think about and be engaged in responses while not being incapacitated by such responses, something which she acknowledges is not easy. Empathy is different from identification (which draws on projection and introjections of bad and good objects) because empathy for a researcher involves offering, rather than expelling, unconscious material in a way that can be handled by the participant. Recently, Bondi has further discussed the ways in which we can better understand the emotional dynamics of interviewees’ accounts by developing our own “receptive unconscious.” Research is replete with unconscious processes and “transmissions” that require us to think about positionality, trust, and rapport in fieldwork (see Bondi 2014).

Moreover, there are a number of applications of psychoanalytic thinking that have provided critical analyses of contemporary social problems or issues. Steve Pile, one of the earliest proponents of the subdiscipline, suggested that some geographers (erroneously) assume that psychoanalysis is preoccupied with the person(al) and that the unconscious is somewhat fixed, meaning it is thus unable to inform critical analyses of social power relations. Classic work in the subdiscipline by David Sibley and Robert Wilton drew on the works of Freud, Klein, and Kristeva, however, to address critical questions of exclusion, difference,
and marginalization. Paul Kingsbury notes how, by producing discursive and stimulating accounts, psychoanalytic geographic research can illustrate the political psychospatialities of issues such as sex(uality), racism, exclusion, embodiment, and community. More recently, conversations between psychoanalysis and other areas in human geography have drawn on affect and emotions as well as intersubjectivities and the nonhuman.

The breadth and depth of psychoanalytic geographies continues to gain critical mass. Research (and the list here is not exhaustive) has engaged with mental health and health geographies as well as psychiatry and neuroscience (Callard 2014); the politics of global warming and the Anthropocene (Healy 2014); industrial capitalism, masculinity, and the geographies of the machine (Nast 2014). Further, Mary Thomas (2010) has delved into the spaces of girls’ identities and daily practices in schools in the United States, marked by racial segregation and violence. Using a psychoanalytic lens, she examines the psychic investments and identification practices that actually serve to maintain (rather than resist) racist, sexist, and heteronormative relations. While feminist geographers have been critical of alliances between psychoanalysis and feminism, particularly Freud’s (early) formulations of women, psychoanalysis has been used as a resource in challenging normative gender roles. For example, the use of psychoanalytic concepts such as splitting and conflict are central to gendered discourses on young women’s alcohol consumption (Stepney 2014). Here, the contradictions and ambivalences in young women’s own narratives about their drinking behavior relate dialectically to neoliberal negotiation and the ordering of social space.

As Liz Bondi notes, applications of psychoanalytic theory offer powerful ways of understanding how unconscious aspects of human life are manifest in the social world. More widely, the transmigration of people, the social tra7uma of conflict, the rise of far right movements in several parts of the world, the fear of terrorism, and the re- enactment of imperialism, along with all the joys and ecstasies of life, are all in some way underscored, internalized, and remade by our inner dynamic worlds. In other words, the outer world becomes internalized, while our inner workings are a microcosmic mediation of the whole world, or at least our world. It is often at points of crisis or in times of conflict that our inner worlds reveal themselves. Connecting worlds, geographers have found psychoanalytic theory to be wholly social, relational, and spatial.

SEE ALSO: Affect; Emotional geographies; Feminist geography; Fieldwork in human geography; Gender; Health geography; Health and wellbeing; Human geography; Identity; Mental health geographies; Positionality; Sexualities; Subjectivity; Therapeutic landscapes

References


Public health activities focus on any source of ill-health that an individual cannot fully control by their own actions, but which can be more effectively controlled by a larger entity, often, but not always, a government entity. Through time, the common sources of ill-health in a population change, primarily as a result of new knowledge, as well as by acceptance by the public of the need for actions to prevent, control, or eradicate the effects of a disease. As this knowledge changes, so does the scope for public health activities. Omran (1971) influenced the thinking of many people when he introduced the concept of “the epidemiologic transition.” This transition described how the typical causes of death changed as societies developed and the proportion of deaths from communicable diseases declined while the proportion of deaths attributable to diseases of the aging process increased. During this time, the average years of life expected at birth increased. He noted that the role of public health changed as the prevention, control, and treatment of the diseases that caused death changed. Along with this transition came the increased realization by the public that different societal behaviors were necessary to reduce the burden of these “new” diseases. Thus the human dynamics of public health also needed to change.

Human dynamics enter this discussion at several levels. First, knowledge of the mechanisms by which a disease develops in any population follows a path of knowledge acquisition and diffusion. Second, pathways through which diseases spread among a population change as typical behaviors of populations change. Finally, the types of diseases that cause death change through time from diseases that spread from human contacts to diseases that develop in populations from the ageing process. Concerns about the health of the public change and more of the activities of public health focus on regulating the exposure of people to broader environmental conditions that, individually, they cannot control.

The focus of this entry on human dynamics concerns the different behaviors in societies involved in these processes, and the different outcomes that occur as a result. The entry distinguishes between two types of human dynamics that affect public health: first, the everyday behaviors of individuals that affect their health, and second, the behaviors of societies who act in the broader general interest. These behaviors are often different between individuals and between societies. The entry begins with a consideration of two notable diseases that have ravaged societies through time, cholera and smallpox. Then, it turns to a consideration of principles surrounding these two types of behaviors, particularly in the context of how they operate differently through time, and throughout the world. The results are complex differences in the geography of public health throughout the world.

Cholera is a disease which illustrates the different kinds of human dynamics that have affected...
the burden it has brought to human populations over long periods of time. The disease “is a severe bacterial infection of the intestine that results in watery diarrhea, vomiting, and when not treated, may lead to death within less than a day. The illness in humans is by ingestion of various species from the genus *Vibrio*,” (Collins 2003). One of the most important public health discoveries of the nineteenth century was that the disease of cholera was transmitted among people by their exposure to organisms in fecal matter in the environment. These vibrios flourish in warm alkaline mediums, such as are found in the brackish waters in estuaries in tropical and subtropical environments (May 1958). After 1854, consensus rapidly developed in societies in North America and Western Europe and most other temperate areas of the world that the spread of these contaminants could be arrested by constructing sanitation systems in which the sources and distribution of fresh water were separated from the collection and dispersal of public waste. To control the importation of cholera from the areas where the disease was endemic, these temperate areas focused on the quarantine system in which the health status of people who reached them, usually by ship, was tracked and, if there was any suspicion that they may have the disease, any contact between them and the local population was prevented for a sufficient length of time that the chain of transmission of the disease from person to person was broken. In many tropical and semitropical areas, however, the cholera bacillus was endemic and this disease was virtually constantly found in local populations. Later, the focus changed to eradicating the disease in areas in which it was always found.

Three different methods of control were available (see May 1958, 56): attacking the causative agent in nature (Collins 2003); changing the susceptibility of the potential hosts, for example, by inoculating travelers; and lessening the opportunity for contact between agent and host. The disease is most often fatal in young children and, for much of the twentieth century, it accounted for a large number of deaths in developing countries. An important medical discovery of the 1960s was that many children could be saved from death by receiving a package of salts that replenished the loss of body fluid from dehydration, thus enabling many young bodies to escape death through rehydration. Thus, the story of the control of cholera in the nineteenth and twentieth centuries is a complex story of advances in biological, medical, and civil engineering knowledge and changes in individual and societal behavior. The story is typical of many other public health problems that rely for their solution on regulating and controlling human dynamics in a population. To be specific, in the case of cholera, regulations related to human behavior began in the 1850s and were so successful that Jacques May, often called the father of medical geography, could conclude, prematurely as it turned out to be, that the last worldwide epidemic of cholera was in the past. The disease was now, on a worldwide scale, “effectively contained.” The reason May’s conclusion proved to be premature is that he did not envision that the regulation of cholera would relapse as societies increasingly relied for their health on vaccinations of the population that traveled in these areas of the world where cholera existed. Experiences with cholera epidemics were largely forgotten and controls, such as requiring vaccinations among those traveling in areas where cholera was endemic, ceased. It was now possible for an immigrant from a cholera-endemic area to travel to other areas of the world without requiring proof of a currently active vaccination, as would have been required 30 years before. Collins (2003) reported that during the late 1990s more than 74 countries
reported cases of cholera. Recently, Richardson et al. (2013) have referred to “the spatial turn in health research.” They note the increased use of geographical information systems to study and monitor the status of such diseases as cholera, where many geographic layers of data interact and where “overlay analysis” increasingly sheds light on the effectiveness of public health efforts to control the spread of the disease.

Not all efforts to control the spread of diseases were without success. Notable was the decision by the World Health Organization (WHO) and by a Cold War agreement between the Soviet Union and the United States in 1958 to cooperate in the goal of stopping the transmission of smallpox throughout the world. Under the auspices of the WHO, many countries participated in a “surveillance and containment strategy” that eventually stopped the person-to-person transmission of smallpox. Surveillance meant prompt and accurate reporting of cases of the disease; containment meant the isolation and vaccination of all persons who had contact with persons known to have the disease. This geographically inspired strategy replaced the older strategy of attempting to vaccinate a very large proportion of all people at risk of the disease. Since May 1970, when the WHO declared that smallpox was eliminated as a disease everywhere in the world, other diseases that depend on the chain of human-to–human transmissions have been targeted for elimination.

The disease that currently appears to be tantalizingly close to elimination is polio, and for more than a decade the WHO has declared its imminent demise. Since polio is spread by interpersonal contacts, and since an effective vaccine is now available to prevent its transmission to protected individuals, polio has, for many years, been a candidate for worldwide eradication. In 2014, the WHO (see WHO 2015) declared a “Polio Public Health Emergency” with temporary recommendations aimed at reducing the international spread of the polio virus. By May 2015, only two countries remained on their endemic list for the polio virus: Nigeria and Pakistan. Several other countries remain under watch. The reason that the worldwide eradication of polio has not yet been achieved is a breakdown in accepted protocols for vaccination in the affected regions and disruptions in implementing these protocols caused by civil wars and the flow of refugees. Each year new cases are found and new steps are taken to contain its spread beyond the geographical area in which it has been found. The question remains as to why it is not possible today to eliminate a disease with largely the same person-to-person transmission characteristics as smallpox. The answer has two elements: first, there have been changes in cultural views about the efficacy of vaccines or the motivations of health workers in areas in which they are sometimes regarded as not working in the interests of local populations; second, there have been increases in the spatial mobility of persons enabling them to move large distances without being required by authorities to show health vaccinations in the areas to which they have moved. In short, the human environment of prevention and control of communicable diseases has changed at the same time as the numbers of people and distances over which they move have increased by many orders of magnitude. The environment of poverty, warfare, and malnutrition, and the absence of a public health infrastructure, also contributes to an environment in which many communicable diseases flourish. More recently, the outbreaks of Ebola in several West African countries have reminded the world of the jeopardy that all may experience as new diseases arise in remote corners of the world.

With cholera and smallpox as illustrations of two diseases in which public health has played
PUBLIC HEALTH: HUMAN DYNAMICS

a notable role in their eradication or control, the roles played by individual behaviors and public behaviors can be discussed. In the age of communicable diseases, individual health behaviors conducive to controlling ill-health differed according to the specific process by which particular diseases spread. Tuberculosis accounted for a very large proportion of deaths during the age of rapid industrialization in parts of Europe and America. The disease is spread through close interpersonal contact, especially within populations who are already weakened by poverty and ill-health. The high density of populations in work areas and residential areas provided fertile ground for human interaction that led to the spread of the disease. Before the discovery and use of drug therapies in the 1950s, however, the proportion of the population with tuberculosis plummeted in many areas of the world as society regulated these conditions by reducing the opportunities for close contact between members of poor populations. This experience is noted because it is emblematic of many other diseases that thrived in certain environmental and medical conditions. These examples led many to discuss the role of society itself in controlling the conditions in which diseases flourish. Many countries decided that there was a distinct role for public health to play to reduce the burden of ill-health on defined populations. Thus was born a concern over “health disparities” and “health inequalities.” This concern, found in many societies, is to find the causes of these disparities. Evidence shows that the health of poorer populations, and other disadvantaged groups, is distinctly poorer than of that of more advantaged groups. Consequently, many countries are trying to identify the causes of these health inequalities and to provide these populations with services that will improve their health. Some of the inequalities are shown to be found in specific areas and some conclusions spread to spatial interventions to reduce inequalities.

Human behaviors, which are the core of public health, operate at both individual and group levels. Noteworthy in the nineteenth century were behaviors related to personal and public hygiene. These activities followed new understandings of how many diseases involved person-to-person disease transmissions and that personal and public hygiene protected the individual from acquiring many diseases. Civil engineering was the key intervention as the wastes of urban living were collected, conveyed, and disposed of separately from the collection and distribution of potable water. The degree to which a person could protect themselves differed markedly from one disease to the other. With the development of vaccines, many of these transmissions could be avoided, and protective methods became available to the public. Not all vaccinations were 100% effective, and not all protections were permanent. A large literature on this subject exists (Rosen 1958).

A second area of focus in human dynamics and public health is the idea that the diseases from which many current populations die can be attributed to behaviors in which individuals choose to engage. Many diseases now can be traced to patterns of eating and drinking that often lead to obesity. Fewer people engage in day-to-day activities of movement and exercise. Instead, far more people are sedentary. Many people smoke tobacco or consume other drugs with harmful health effects. People sometimes drive after drinking alcohol, leading to deaths from vehicular homicide. By the 1960s, major causes of death in the developed world were cardiovascular disease and stroke, cancer, pulmonary disease, and injuries. Vallin and Meslé (2004) show how some countries in Europe were successful in reducing rates of death from cardiovascular disease while others were not.
They show that a group of countries in Western Europe were able to focus on changing behaviors that increased the risk of individuals developing cardiovascular disease and worked to accelerate the development of medical treatments which, to a large degree, reduced these risks and reduced the rate of death due to these causes.

Determining the behaviors in question that needed to be regulated and controlled involved research, education, and changes in public policies that many people disputed were issues for the public to decide. One behavior was smoking tobacco products. The story of how smoking came to be recognized as responsible for higher rates of death from many diseases is a fascinating one, and the slow adoption of measures to control the exposure of nonsmokers to smoking by others was long and arduous in all countries. Currently, individual countries are found at many different places along the continuum of control over the availability and use of tobacco products by the population. The success of many legal challenges to the rights of tobacco companies to promote their products, especially to the young, and the penalties many of these companies have been required to pay as a result of their actions, has led to a marked reduction in the rate of smoking of tobacco products in the United States, Western Europe, and many other countries across the world (Eriksen, Mackay, and Ross 2013). Nevertheless, international sales of tobacco products continue to grow in many countries as marketing science continues to play a strong role in promoting the use of tobacco products throughout the world.

Many well-known causes of ill-health are the subject of controversies over the role of public decisions to control them. Common questions that are still highly debated in developed countries include whether the number of retail outlets that sell “fast food,” liquor and tobacco products, and guns and ammunition should be regulated; whether fines should be increased for driving under the influence of alcohol, or for not wearing a seat belt or motorcycle helmet; whether infants should be required to be restrained in rear seats; and questions concerning countless other interventions known to reduce death and illness. On these and many other similar questions, people disagree on the role of the state in regulating personal decision-making. Libertarians are generally viewed as a group of persons who want to protect the rights of individuals to make their own decisions about their patterns of consumption and their rights to offer products to others.

Other circumstances that affect the health of populations relate to philosophies of life held by different population groups. An example would be whether the decisions of women concerning childbirth should be regulated by legislatures or whether they should be left as individual decisions made by the persons affected by them. In another example, research has been conducted on the local circumstances in which people become addicted to drugs. This research led to many policy recommendations that would likely reduce rates of drug abuse in specific populations if adopted (Thomas, Richardson, and Cheung 2008).

Societies differ on whether the geographical accessibility of essential medical services should be regulated by governments. Where many medical personnel are employed directly by governments, control of access is commonly found, but some other countries, such as the United States, generally rely on incentives that will encourage individual medical personnel to choose to practice in areas for which there is a known need for expanded services. Models linked to geographic information systems have been developed to facilitate this purpose.

New areas for public health intervention are environmental health and cancer control.
and prevention. In many countries, the public has reached the conclusion that one source of ill-health that could be controlled are chemicals in the environment that have adverse health effects. Materials that affect the respiratory functions of children and others are common. These chemicals, it is asserted by many of the public, could be reduced or eliminated from the environment. Notable examples are smog and fine particles that originate in the transportation system through the extensive use of oil products. Other industries also produce large quantities of dust and chemicals during the manufacturing process. Many children are exposed to environmental substances that are now known to cause ill-health and even death. Some mining products also contribute to ill-health. Asbestos and lead proved to be very dangerous products used extensively by the public in their everyday lives until, in the interests of public health, their use and people’s exposure to them were drastically limited.

An example of a successful program to eliminate an environmental health threat is the program started in many US cities to reduce exposure to lead. Research showed that lead in the blood of infants and children originated in lead particles in the environment from the time when lead was a common ingredient in paint. A strong correlation was found between the age of a residence and the levels of lead in the blood of infants and children. Children living in pre-1970s residences had higher blood-lead levels than those living in residences built after the mid-1970s, which were not painted with paint containing lead. In the United States, individual cities and towns needed to make the decision to clean up lead droppings from around older residences and enforce the ban on using lead-based paints. Many incidences of developmental disabilities were found to be associated with the levels of lead in the blood of children.

As the number of cases of cancer increases sharply in countries across the world, public health increasingly is involved in activities designed to prevent their occurrence and control the numbers of deaths caused by them. Human dynamics enter this fight in many ways. First is the decision to research the characteristics of cancer. Increasingly, cancers are identified using their molecular components rather than the traditional identification by the organ in the human body in which it is found. Second, an important recognition is the development stage at which the cancer is identified in the body. The earlier the stage, the more likely the cancer can be successfully treated and death avoided. Accordingly, the role of public health often is to map the proportion of the people at risk of having the cancer who are in the late stages at the time of their first diagnosis – the larger the percentage, the higher the death rate from the disease. The goal of public health is to increase the “surveillance” of the cancer in question in the population at risk in order to increase the number of people at risk whose cancer is recognized at an early stage when it can be treated and, often, death avoided.

After reviewing the dynamics of public health activities, the number of disciplines involved in them, and the essential role of the public and the responsibilities of individuals and of society at large, one essential question remaining concerns the development and control of the infrastructure of public health. This infrastructure is none other than the collective actions of individuals, groups, institutions, and countries. The dynamics of this collective action are very large, as well as the effectiveness in the grand task of ensuring the health of the people. The complexity is enormous, the grand challenge, breathtaking in scope, the actors, many, the final healthy state of the world, essential.
SEE ALSO: Disease diffusion; Environment and health; Geographic information system; Health and development; Health geography; Health inequalities; Health-related behaviors; Health systems; Health and wellbeing; Mortality

References


Further reading


Public policy

Kevin Ward
University of Manchester, UK

Public policy is policy named, designed, delivered, monitored, and evaluated for the public good. There are many areas of public policy, including crime, education, environment, housing, and transport. In these, government has decided to establish a course of action, setting aside a budget to finance its intervention based on a particular set of objectives. Rarely are the decisions to label something as “policy” and to allocate a particular budget to its delivery and evaluation neutral, however. Rather, they embody a particular strand of thinking among those in government whose remit it is to make policy. Decisions over public policy involve decisions over prioritizing or targeting some areas over others (e.g., crime over housing), which means prioritizing some groups of people over others and some geographies or places over others. There are always “winners” and “losers” in public policy even if who they are is not immediately apparent. Public policy is the product, then, of a complex set of decisions involving those inside and outside of central, regional, and local government.

It is important to place public policy within its wider geographical context. Each country has its own constructions of “public policy.” In part these reflect the role that government has historically played in the country. In some cases governments have adopted a relatively interventionist strategy, expanding their remit in term of public policy. A good example would be the Japanese government’s approach to industrial policy. While there have been changes over the years, for the most part the government has sought to invest in a number of industries that were identified as those in which Japan might be able to develop a comparative advantage. This approach to industrial policy in Japan was probably at its peak in the 1980s. In other cases, governments have been circumspect. Their remit has been more laissez-faire. An example of this approach is that of the British government to banking policy during the early 1980s. It liberalized the supply of money and reduced the reserves required to be kept by banks. The differences between types of public policy making reflect the particular ways in which cultural, economic, political, and social factors have combined in each country, even among those countries that appear to share some common characteristics. This point applies equally to countries in the North and in the South. There is a variety of public policy types, and thus considerable differences between countries in terms of how they name, design, deliver, monitor, and evaluate public policy, and the overall thinking that underpins these decisions.

It is also possible to identify a number of patterns across countries in constructions of “public policy.” First has been the emergence of the notion of “partnership” at the center of public policy. In a number of countries governments have historically taken the lead on the naming, design, delivery, monitoring, and evaluation of policy. They worked with representatives of business and labor unions nationally in corporate arrangements. While government continues to be the dominant partner, the last three decades have seen the emergence of “government +”
PUBLIC POLICY

arrangements in most areas of policy; that is, the uneven and partial emergence of governance as the dominant mode of public policy making. Whether an area-based policy – such as that focused on downtown or a whole city – or a thematic policy – such as that focusing on the banking or environmental sector – a range of others have been brought more centrally into the public-policy-making framework. While the role of business has been expanded, in many cases labor unions have found themselves less involved. Others, such as community or resident groups and neighborhood associations, have been brought into partnerships with government, not just nationally but also regionally and locally. So, not only has partnership become the new orthodoxy in public policy, but the makeup of the partnerships has changed as their number has grown.

Second has been the growth in the notion of “best practice” in public policy. This term refers to those locations – cities or countries – that are generally understood to have been successful by governments and others involved in public policy. For example, in urban regeneration public policy, Barcelona in Spain is often referenced (Garcia-Ramon and Albet 2000). The city is constructed as an example from which others around the world can learn, in terms of how the regeneration was financed, what partners were involved, and the nature of the outcomes. The establishing of “best practice” takes place in a number of ways. One is through the circulation of policy documents, allowing others involved in urban regeneration policy to read about Barcelona. A second is through exchanges and visits, as those from elsewhere visit Barcelona or those from Barcelona visit other cities. Of course, the making of public policy has always involved an element of comparison between examples; however, in recent years there has been an expansion of the geographical reference points.

The third has been the growing role of consultancies in the various aspects of public policy. In some cases this growth has been at the expense of government. So, in such examples there has been a withdrawal or downsizing of government. In its place consultancies have seen their role expand. In other cases the increased involvement of consultancies has constituted an expansion in the work done around the naming, design, delivery, monitoring, and evaluation of policy. Private sector consultancies, both the large transnationals such as KPMG and the smaller independents, have seen their involvement expand in the public policy process. This growing market for consultancies has occurred as local, regional, and national governments in a number of countries around the world have purchased their expertise, experience, and reputation.

Fourth has been the proliferation of ways of monitoring and evaluating public policy. While its different aspects have always been subject to scrutiny and review, the nature and means of this monitoring and evaluating have changed in recent decades. There has been a growth in the data generated on the different aspects of policy. So, every stage is now subject to attention in a way that it has not been in the past. This enumeration of the aspects of public policy has translated into a growth in the way that data are represented. Benchmarks, key performance indicators, and league tables are just three of the most high-profile ways in which the data are brought together and used to compare and assess the effectiveness of policymaking. In the United Kingdom, the emergence of New Public Management (NPM) during the late 1980s involved the restructuring of government and the making of public policy. A central feature of NPM was the production and use of indicators as a means of assessing and evaluating the performance of policymaking.
For some, taken together, these general patterns in the making of public policy are evidence of how neoliberalism has emerged as the dominant institutional context for policy-making, at least in some countries (Peck 2003). Here, neoliberalism is being used to refer to the qualitative restructuring of the state, involving new forms of statecraft, such as the widespread emergence of partnerships across many areas of public policy. This attention to its broader institutional field reflects a renewed interest in human geography in relation to the study of public policy, and includes some reflection on the role of academic geographers in the different aspects of public policy (Ward 2005). For, as the making of public policy has been restructured to involve a wider range of partners, so alongside community groups, consultants, and think tanks, universities and academics have featured more prominently. This has led academic geographers to reflect on their roles in the making and evaluating of public policy, as, despite claims to more open and transparent government, participation by those who are most affected by policy remains rather limited.

SEE ALSO: Human geography; Neoliberalism

References


Further reading


Public space

Erin DeMuynck
University of Wisconsin–Fox Valley, USA

Public space is understood as diverse, vibrant, and inclusive of all members of society. It offers opportunities for sociality outside of work and home, which can add pleasure to the everyday. Beyond its capacity to enhance enjoyment of everyday life, public space is identified as an essential element of urban life and a vital site of democracy. It serves as a site for making struggle and dissent visible and audible, where different groups have the opportunity to declare and debate diverse opinions and viewpoints and to express their needs and wants. More casual planned and unplanned meetings with others also take place here, allowing for exposure to ideas and ways of life that might not otherwise be encountered or considered, and which can enhance understanding, tolerance, the free flow of ideas, and the potential for greater inclusion and participation in the making of urban spaces and urban life.

This notion of public space, in which it exists as a truly open, democratic, and peaceful place, is an idealization – a goal to strive toward rather than something that presently exists or has ever existed in a pure form. Even the ancient Athenian agora, for example, often considered a quintessential public space, was accessible only to men. True publicness has remained elusive in contemporary cities. One reason for this is that a degree of conflict is inherent in public space. The rights or comfort levels of certain individuals or groups may be encroached on when individuals or groups with conflicting values, or with differing ideas about the appropriate use of a certain space, come into contact. Exclusion from an ostensibly public space may be explicit, enacted through rules and regulations or even violence, or it may be very subtle, achieved through a feeling of uneasiness or through real or perceived incompatibility with the dominant values associated with a space or its dominant uses.

Whether exclusion from public space is implicit or explicit, it is constantly being challenged. Through these challenges, whether subtle or overt, public space is continually being made and remade. Moreover, it is through these challenges and continuous remaking that public space does its democratic work. Many examples throughout history can be used to illustrate how public space plays important roles in the redefinition of who is included in “the public” and in the advancement of human rights for a number of groups. Through claiming and acting in public space, groups that were at the margins or excluded from many aspects of society have challenged their exclusion. The rights of women, people of color, people with disabilities, and the LGBTQ community, among others, have been fought for in and through public space.

The women’s suffrage movement in the United States in the early 1900s made extensive use of public space. Women occupied streets, plazas, and urban parks to raise the profile of the issue of women’s suffrage rights to national consideration. Entering spaces that were considered public, but in practice were male-dominated, helped women redefine what “public” meant, broadening the dominant conceptualization of “the public” to include women. African
PUBLIC SPACE

American civil rights activists made use of public space during the 1950s and 1960s as a way of claiming rights as members of the public. Activists carried out marches and sit-ins, claiming rights to white-only spaces from which they were explicitly excluded. Later in the 1960s, the gay rights movement brought gay identities out of the realm of the private or underground and into public space in order to legitimate gays and lesbians as members of the mainstream public. Much progress has been made toward women’s rights, African Americans’ rights, gay rights, and the rights of many other groups, but work remains to be done. Rights continue to be fought for today, in and through public space, throughout the world.

For people experiencing homelessness who may not have access to private spaces, public space is crucial. Without the right of entry to private space, the alternative for carrying out daily activities and simply living must therefore be public space. Without public space, people experiencing homelessness may have no place to legally exist. Local regulations against activities like panhandling, as well as lying down, sitting, and even body odor have been increasingly implemented as a means of targeting people who are homeless and other groups. These types of regulations, along with the advent of benches designed to restrict lying or sitting comfortably, strategically placed sprinkler systems, and increased surveillance that are identified as unwanted in these locations, make spaces uncomfortable and often difficult or impossible to use.

It should be mentioned here that, even for the most open and accessible of public spaces, as the values and beliefs, uses, or social needs of one group are being declared or claimed, the ability of another group to use the same space in ways that meet their own interests and needs is likely to be diminished to some degree. This limitation disproportionately affects marginalized groups, but may affect anyone. For example, people experiencing homelessness find it increasingly difficult in today’s cities to find spaces where they are not excluded by law and oversecuritization. On the other hand, a public park that is a congregating space for men may be viewed by women as a space that makes them feel uncomfortable or unsafe and one that they avoid. Publicness is always subjective to some degree and the extent of a space’s publicness may vary from moment to moment.

Nevertheless, a general reduction in the inclusiveness of public space has been identified in contemporary society. The possibility of encountering a variety of social groups or of raising alternative and dissenting viewpoints has become rare in cities today. New regulations, design elements, and increased policing have been enacted that restrict the use of public space. This is a consequence of the increasingly instrumental stance taken by cities toward public space. Public space has been reconceptualized in contemporary cities as a component of sanitized city livability primarily for middle- and upper-class consumers, hampering the capacity for public space to fulfill its democratic role. Efforts to commercialize, sanitize, and securitize public space, while they may appeal to some and be beneficial to capital, curb the rights of others who are excluded from these spaces, sometimes punitively.

Geographic research places the changes in the way public space is perceived, used, and experienced within the context of broader urban restructuring processes. To understand current processes impacting public space, we can start by looking at some of the consequences of suburbanization. Suburbanization and white flight, along with deindustrialization and decentralization of employment, led to dramatic urban...
changes starting in the 1950s and 1960s. Population numbers in city centers declined along with the median income of the residents who remained. The relatively well-paying industrial jobs that had allowed for middle-class lifestyles had become scarce. An increased percentage of the urban population was living in poverty at the same time that federal support for urban social programs was greatly reduced. The local state could no longer afford to provide for citizens’ needs or to maintain urban infrastructure. In this context, public spaces deteriorated. Crumbling streets and sidewalks, litter-strewn vacant lots, and unkempt parks with dilapidated playground equipment and broken benches came to symbolize urban crisis. By the 1980s, many urban centers were viewed as no-go zones for the middle and upper classes.

Urban entrepreneurialism was developed as a strategy to solve this urban crisis. The goal was to attract an economically active population back to the city at a time when funds for services and infrastructure had been drastically reduced, unemployment was rising, and social need was high. As a mode of adaptation, any remaining energy and resources were redirected away from social programs and toward pro-growth and pro-business strategies such as corporate tax breaks, privatization of public services, the creation of public–private ventures, and cultural development and image-building projects. The goal was for postindustrial cities to be perceived as cleaned up, safe, and welcoming to business owners, developers, and wealthy tourists and residents. As part of this strategy, formerly neglected public spaces have been selectively redeveloped in certain parts of the city. This process continues today.

Through the lens of urban entrepreneurialism, which continues to dominate urban policy today, public space is seen as a redevelopment and city imaging tool that can be put into the service of attracting investment to the city. Public space is often given a central role in urban planning and redevelopment as a way to beautify the city, demonstrate quality of life, and gesture toward sustainability, a value that is thought to be held in esteem by the population regarded as being the key to cities’ success; however, because of the drive to create a sanitized, orderly urban image, democratic public space is increasingly under threat. Messiness, disorder, and tension, which are inherent in spaces that are open and accessible to all, are at odds with the goals of contemporary urban governance seeking to attract business investment and middle- and upper-class tourists and residents. As its latent potential for disorder and conflict is increasingly constrained through design, regulations, and increased security and policing, public space loses the capacity to fulfill its democratic potential. People and actions deemed undesirable or incompatible with business interests and land uses are increasingly being formally excluded from public space. Its democratic capacity cannot be carried out when entry is allowed to some while those who neither consume nor add to the aesthetic desired for the city are kept out.

Although exclusion from what is conceptualized as public space, ostensibly free and open to all, is not a new phenomenon, contemporary forms of exclusion place added constraints on public space. Presently, from the perspective of entrepreneurial governance, public space is considered an urban amenity that enhances quality of life and is promoted as a way to encourage private investment and attract tourists and middle- and upper-class consumers. Although public space has become an important discursive and urban imaging strategy, the projects and developments that are actually produced are for middle- and upper-class consumers only. They
PUBLIC SPACE

are public in name alone. The element of autonomy in public spaces is small in today’s cities and continues to decrease. Few spaces today offer freedom from an obligation to conform to externally imposed sets of rules, from the pressure to spend money, or from surveillance and police presence.

The results of catering to a particular demographic slice in urban redevelopment and public space planning can be visually observed in cities. Upscale outdoor shopping districts and nightlife zones have exploded in cities. Urban parks, riverwalks, downtown plazas, and urban greenspace are increasingly incorporated into urban redevelopment plans as well. While these plans discursively express a renewed commitment to public life, the actual publicness of the spaces produced is questionable. The development of business improvement districts and other forms of public–private partnerships adds to the concern as management of public space is controlled by business interests, which displaces public space even more from democratic processes. Resources are directed toward attracting middle- and upper-class consumers. Increased surveillance, policing measures, and regulations aimed at unwanted populations are introduced to keep revitalized parts of the city free from those deemed undesirable. The result is an intensely uneven landscape in which public space has been reinvented in showcase urban districts as a lifestyle amenity for the middle and upper classes while those who are poor continue to be shunted into increasingly neglected and deprived parts of cities.

Securitization has further eroded public space as a response to the real and perceived threat of terrorism, which escalated after the September 11 attacks on the World Trade Center and the Pentagon. Streets and open space around government buildings, for example, are physically cordoned off and people have fewer places where they can protest or congregate. Urban scholars suggest that limitations on public space that are legitimated in the name of security and the prevention of terror attacks are in reality being secured from its users rather than for them. Public protest, social activism, and political mobilization are being repressed as access to public space is taken from all but those with security clearance to enter. In effect, formerly public spaces are further privatized. The implementation of free speech zones is an overt reduction of free speech rights associated with public space in the name of security. When an event surrounding a controversial issue is planned, protesting may be outlawed in nearby public spaces and confined to a zone that is removed from the event itself. While business owners, employees, and shoppers continue to be granted access, protesters and activists are kept away. This control is exercised in the name of safety and security, but the result is that the targets of the protest have no way to see or hear the message of the protesters unless it is reported in the media or on the Internet.

The Internet, for its part, has been suggested as a contemporary form of, or alternative to, public space. Some public space functions can be fulfilled through participation in the virtual space of the Internet. People can connect with others, and disseminate and discuss ideas. Used together, virtual and public space may have allowed the Occupy Wall Street protests and, similarly, the Arab Spring revolutions, to expand their impact beyond what might have been possible without the use of the Internet. The virtual public space of social media was used to coordinate and mobilize protests in physical public space, warn activists of danger, as well as communicate what was happening on the ground in specific public spaces like Tahrir Square in Cairo or Zuccotti
Park in Manhattan with people both nearby and in other parts of the world.

While the role of virtual space was important to these movements, it should be noted that they would not have had the same impact without physical public space. Coming together in physical space remains important in the digital age. Moreover, the ability of the Internet to function as a democratic public space currently faces similar challenges to those faced by physical public space. Both are threatened by uses that privilege profitability. Net neutrality, the notion that all Internet traffic should be treated equally, is regularly challenged. While the US Federal Communications Commission has ruled that Internet service providers are not allowed to prioritize content on the Internet for those who pay more for this service, they continue to work to undermine the open Internet. Without net neutrality, Internet content created by those without wealth would be buried beneath content created and paid for by corporations, making it difficult for alternative messages to be disseminated. Beyond the issue of net neutrality, the digital divide continues to be an important consideration. Access to the Internet is not universal but is contingent on where one lives and on socioeconomic status.

The emphasis on the economic function of public space over its use-value precludes certain conducts, activities, and political practices. Patterns of ownership, design, and management schemes are at times subtly restrictive and/or overtly exclusionary. If making claims, achieving visibility, and influencing public opinion are to occur, public space cannot favor one type of use by one segment of the population and deny other uses and groups. However, although public space has been increasingly privatized, commercialized, and regulated, a great potential remains. In spite of regulation and exclusion, the parks, plazas, streets, sidewalks, and urban squares that fall under the umbrella of what's traditionally considered public space may still offer the best opportunity for social life and social change. Research on public space has focused primarily on the impacts of economic restructuring and urban entrepreneurialism on the decline of public space. Future paths for public space research include interrogating the openings that exist for public space to be claimed in this political-economic context and for demands to be made that challenge the use of public space as a tool of an exclusionary form of urban redevelopment.

**SEE ALSO:** Neoliberalism; Space; Spatiality; Urban redevelopment

### Further reading

PUBLIC SPACE


Public-participation GIS

Rina Ghose
University of Wisconsin–Milwaukee, USA

Spatial data and geographic information systems (GIS) have been essential to resource management, conservation, planning, and policymaking activities since the 1960s, as most decision-making tasks require spatial information. Yet, access to GIS has been uneven due to its cost and complexity. Based on the notion that increased use of spatial knowledge leads to more informed participation in policymaking, public participation GIS (PPGIS) aims to broaden the access of GIS technologies to socially marginalized groups, leading to more informed and empowered communities.

The evolution of PPGIS as a research agenda and a practice is rooted in the tumultuous debates surrounding GIS in the early 1990s. GIS was critiqued for its perceived positivist epistemology, which reduced complex social processes to points, lines, areas and attributes (Pickles 1995). Further, GIS was seen as an elitist technology with uneven access that led to unequal power relations. Critiques considered GIS to be an instrument for production of rational, expert knowledge, implemented in policymaking through a top-down approach. The presence of a GIS digital divide along class and race lines was seen as inherently disempowering to marginalized citizens, for it discouraged their knowledge and their participation in resource allocation and governance.

Led by the National Center for Geographic Information Association (NCGIA), the Friday Harbor conference in 1993 united the critics and the proponents of the GIS to formulate the GIS and society research agenda, in order to address these criticisms. PPGIS arose out of this research agenda as an effort to advance informed participation of marginalized communities through inclusive access to GIS and spatial data. Therefore, the unique nature of grassroots groups and their challenges need to be understood.

Grassroots organizations of marginalized communities tend to be more decentralized and more fragile in financial and staffing support when compared to larger nonprofit organizations and public–private sector agencies. Their resource-poor conditions make it difficult for them to afford the cost of implementing GIS in their organizations. Purchase of data, software, and hardware or provision of ongoing GIS training to their staff members continue to be a challenge for these organizations. Moreover, these organizations are also hampered by frequent staff turnovers that result in higher staff training costs. Finally, grassroots organizations have faced the additional challenge of not gaining easy access to public databases that are often the repository of valuable community-based spatial data at multiple scales.

These difficulties have created a technological divide between the planning agencies that are frequent users of GIS and grassroots groups that face challenges in using GIS. Class inequalities are then further heightened by inequalities concerning the knowledge of and access to such information technologies, for it is clear that such access and understanding create greater opportunities for conferring political, economic, and social power upon the citizens of distressed neighborhoods. The establishment
of “empowerment, marginalization and public participation GIS,” as a critical research agenda by NCGIA and University Consortium for Geographic Information Science (UCGIS) in the late 1990s significantly advanced PPGIS work. A wide range of initiatives emerged across the world, aiming to enhance accessibility of data/GIS to disadvantaged groups and to incorporate multiple voices and local knowledge within a variety of contexts.

An examination of global case studies of PPGIS research and practice reveal certain key goals. PPGIS aims for: (i) equitable access to both spatial data and GIS technologies among socially marginalized citizens; (ii) incorporation of grounded, indigenous, experiential, and local knowledge with public datasets for marginalized citizens to contest or reshape policies; (iii) discouragement of top-down, rational planning approaches in policymaking; (iv) inclusion of both qualitative and quantitative data in GIS; (v) alternate forms of mapping, representation to capture complex social processes and cultures; and (vi) alternate forms of geospatial technology designed to suit the needs of indigenous or socially marginalized groups.

In response to the establishment of PPGIS agenda, various international conferences and workshops have been held that have significantly shaped PPGIS work. Notable among them are the Varenius workshop on “Empowerment, Marginalization and Public Participation GIS” (1998) held in USA, the workshop on Access and Participatory Approaches in Using GIS (2001) held in Italy, PPGIS international conferences held in the USA (2002–2005), and the PPGIS e-seminar (2007). Such workshops and conferences have been sponsored by significant scholarly bodies such as the US National Science Foundation, Urban and Regional Information Systems Association (URISA), European Science Foundation, Institute of British Geographers/Royal Geographic Society, which demonstrate the critical attention that has been given to this topic. A rich body of literature containing both theorization and case study analysis has been published in various journals and books. Notable among these are the special issues in Cartography and GIS (1998), URISA (2003), Cartographica (2004), as well as a seminal book on global PPGIS titled Community Participation and GIS (2002). Literature reviews of PPGIS practices also provide significant assistance in understanding this complex process (Sieber 2006). Other notable volumes, such as the Handbook of GIS and Society Research (Nyerges, Couclelis, and McMaster 2011), have also devoted significant sections on PPGIS. Individuals engaged in PPGIS have thus gained awareness as a distinct community with the creation of new spaces of discourse such as the PPGIS conferences and listserv.

PPGIS initiatives have been carried out by a multitude of social groups, such as activists, universities, government and nongovernmental agencies, community-based organizations, and grassroots communities. University–community partnerships and collaborative planning programs have provided particular opportunities for GIS practitioners and scholars to collaborate with disempowered and marginalized community groups, leading to grounded, bottom-up spatial knowledge production that could be strategically employed in policy activities.

Over the past two decades, the areas of PPGIS applications and theoretical conceptualizations have evolved and mutually informed each other. PPGIS practices have been carried out globally to address a multitude of policymaking, ranging from rural land use, to conservation and natural resource management, to urban planning, to social and environmental activism, and to advocacy (Craig, Harris, and Weiner 2002). PPGIS research has been shaped by an integrated
theoretical framework derived from social and political theory, science and technology studies, human computer interaction, organizational theory, and feminist theory. PPGIS as a process has been critically scrutinized to explicate the notion of public and the nature of participation. Such critical introspection has led to creative tension, with suggestions to rename it to participatory GIS (PGIS). Both terminologies are currently in use; however, PPGIS is also considered by some to be subsumed within the “critical GIS” moniker.

Key issues in critical GIS research span a diverse array of topics – differing national level procedures for spatial data access, effective PPGIS practices from initiatives around the world, and the use and impacts of digital geographic data in spatial decision-making. The nature of participation and the process of empowerment through PPGIS have also been interrogated. Though their areas of inquiry are quite diverse, research indicates that geographic data access and PPGIS projects are place specific, highly contingent, and strongly shaped by the local context in which they are situated. Place thus plays an important role in shaping participatory approaches to spatial decision-making. Further, the nature of these participatory processes is crucial to understanding the differential impacts of PPGIS initiatives for the individuals and communities affected by them.

Examining access to data/technology in PPGIS

PPGIS stresses the importance of enhancing access to existing spatial data/GIS and to participation opportunities in policymaking processes. Within this context, appropriateness and accuracy of data are important considerations, as data and analysis that are found to be appropriate and accurate will lead to greater action. To properly assess this issue, a model has been proposed which asks several questions – Are the data and material produced appropriate to the organization issues? Can the organization use the information in action-oriented ways to support decisions, enhance communication, and inform actions? Is information available to the organization in a timely manner? Is the information pertinent to organization issues? Do the results have a temporal and cross-comparison component, that is, a time perspective? Are the available data sufficiently accurate?

In the case of data/GIS, access is shaped by factors such as existing spatial infrastructure developments, the presence of a supportive network of actors, and supportive local policies. Access to data and technology also lies beyond the simple availability of data, hardware, and software. Issues of appropriateness of data, data accuracy and updates, forms of data representation, and costs of training are important and integral parts of effective PPGIS practices.

Access to public datasets is affected by different legal structures for copyright and licensing, and existing traditions such as the freedom of information access to public data. Data sharing between public agencies and citizen groups also depends upon the degree of government agencies’ openness to accepting grassroots citizen groups as authoritative participants in the planning process. Conversely, budgetary cutbacks and fiscal constraints may compel government agencies to sell their data to the public and/or limit data sharing with selected stakeholders.

Within PPGIS, access to technology leads to questions of GIS implementation in grassroots organizations. The adoption and use of GIS by grassroots organizations has been studied under the moniker of GIS implementation. This body of work is based on the research on GIS
PUBLIC-PARTICIPATION GIS

implementation in local governments, which, in turn, has roots in organizational theory. Here implementation refers to the decisions made by an organization to acquire, install, implement, and use GIS in accordance with organizational needs and tasks.

The presence of supportive actors that assist grassroots groups to gain access to data is another key factor. Because of the resource-poor nature of grassroots organizations, they often rely on external expertise for access to GIS data and technology. External GIS providers in PPGIS tend to contain certain common characteristics. In particular, they tend to be housed in larger institutions, which provide the budget, stability, and credibility to those using its products. However, the institutional homes of the supportive actors are varied, including academia, larger nonprofit organizations, public libraries, and city government offices. In the case of urban PPGIS studies, Leitner et al. (2002) found six major models of PPGIS provision to neighborhood organizations: community-based GIS; university–community partnerships; GIS facilities in universities and public libraries; map rooms through government GIS; Internet map servers; and neighborhood GIS centers. The six GIS provision models are assessed through the following aspects: costs of maintaining GIS (hardware, software and GIS training); ease of data provision and responsiveness to community organizations’ needs; stability and longevity; and ability to support collaborations among stakeholders and among grassroots organizations.

Studies also indicate that access to and participation in the use of geographic information depend on social context such as local culture and institutions. Factors such as a culture’s ability to absorb uncertainty, its level of masculinity, and its ability to accommodate human inequality can significantly shape PPGIS. Other factors include a community’s tolerance of expert solutions, its sense of collective control, and its level of individualism. Prevalent cultural and political norms can limit the type of participants to a specific gender, class, race, or caste. Politically repressive cultures can restrict public access to critical data through legal controls and may control citizen participation to such an extent as to render PPGIS efforts ineffective and irrelevant.

Finally, internal characteristics of a grassroots organization also significantly shape its PPGIS activities, including its data access. In particular, Elwood and Ghose (2001) provide an organizational framework to examine neighborhood organizations’ usage of GIS. The framework includes four major dimensions: organizational knowledge and experience; networks of collaborative relationships; organizational stability; and organizational priorities, strategies, and status. In terms of data access, the organizational ability to form and sustain relationships of support with local government agencies and GIS provider actors is vital. Further, a grassroots group’s awareness of the value of spatial information in policymaking will shape its efforts to access data. Lastly, availability of data in terms of appropriateness, accuracy, access, ownership, and representation are equally pertinent issues.

The context of equitable access to GIS is dependent not only upon mere availability of software and hardware but also upon the GIS skills that reside among the grassroots group. Within the PPGIS agenda, GIS represents a socially constructed technology that can be redesigned to suit the goals of the users, provided they have a fairly high level of technical skills that can be adapted to suit their goals. PPGIS practitioners in the environmental conservation arena often display a high level of technical skill and are adept at using the technology to suit their needs. Therefore, in-house GIS use by such grassroots user groups is common. In the developmental
context, the integrated approaches to participatory development (IAPAD) group developed a process of blending GIS, physical terrain models, and community participation into “participatory 3-D modeling.” This approach exemplifies an interface that is nontechnical. On the other hand, many community mapping projects exist in both rural and urban groups where community members do not engage directly with the GIS but work in collaboration with technical teams to provide their input (such as indigenous, experiential or local knowledge) and evaluate output. Here, it is necessary for community groups to acquire map reading skills, as outputs are in the form of maps. Some projects increase an applicant’s ease of use and lessen the need for GIS skills by enhancing the human–computer interaction (HCI). Other research has produced a bird’s-eye viewer for GIS participants who experienced difficulty in comprehending their community from a 2-D planimetric map.

Human–computer interactions can be characterized from no direct use, to passive use, active use, and proactive use. These are not hierarchical; proactive is not necessarily the optimal level of usage. Nor will stakeholders benefit equally. An effective PPGIS application depends on understanding how much and when technology should be brought into a process. The corollary is how much GIS must be learned by individual stakeholders and what technologies can be supported by available resources.

Advances in Internet technology have significantly enhanced the ease of data access and mapping, and have facilitated data input and communication between different groups. Government agencies have facilitated public access to spatial data and analysis through their Internet GIS sites. However, access to high-speed Internet connection is vital for grassroots groups to effectively access data and perform analysis. Internet GIS is also commonly utilized to elicit the public’s input and opinions in resource management and planning.

Most recent studies have emphasized the development of geospatial technologies in combination with Web 2.0 technologies (Goodchild 2007). Google Maps, Google Streetview, and Google Earth have emerged as popular, user-friendly mapping interfaces, which can also be linked with GIS technologies to create analytical functions. Citizens can locate places and view them at a high resolution through street maps, detailed aerial photographs, and satellite images. Such efforts have significantly increased spatial awareness and mapping among average citizens and have resulted in individuals providing user-generated content through “volunteered geographic information” (VGI) (Goodchild 2007). VGI has provided any citizen with the opportunity to add information to place, upload videos or images, share information, and enrich the understanding of a place. For managing emergencies and disasters, or even for citizen science, such user-generated content and mapping through VGI activities have been very helpful. Both VGI and PPGIS thus provide the opportunity to any Internet–using layperson to contribute geographic information for a broad spectrum of purposes, allowing users to act as both information consumer and information provider. The distinction though is that while VGI is more concerned with individualized information sharing and mapping, PPGIS is oriented towards group decision-making that includes the voice of marginalized citizens and explicitly seeks social change.

Development of free and open source GIS or Open GIS is a significant achievement in eliminating the factor of software cost in PPGIS (Sui 2014). The principles of Open GIS include: the creation of free, open source software; the harnessing of collective intelligence through a bottom-up information flow and user-generated
content; the idea of the web as a platform and the development of web services, cloud computing, and lightweight programming models allowing for application development, such as mash-ups; software that extends beyond a single device to laptops, mobile smartphones, and tablets; and the rich usability of an easy-to-use graphic interface. Moreover, the increasing employment of smartphones with special apps provides great ease in collecting geocoded spatial data and enables new ways of data collection through participatory photomapping. The diverse, mobile, and flexible nature of open GIS provides greater opportunities to collect local knowledge and build spatial narratives in cost-effective ways.

Overall, web-based GIS moves away from one-way information dissemination to two-way interactive communication to three-way public–public communication. Grassroots groups have a better chance to participate in public policy formulation if spatial knowledge and GIS is incorporated into the organization and utilized as management tool, a form of community interaction, and for resource allocation. The ongoing development in web-based geospatial technologies provides community-based organizations a space to represent their knowledge of local places through the maps they create. Web-based GIS mapping programs are a popular way for organizations to incorporate a GIS as part of their internal management and planning strategies and have also made it easier for community-based organizations to share information with one another as well as with the different social actors with whom they are collaborating on participatory projects.

Web-based GIS mapping applications have been cited as effective tools for public participation in collaborative research projects, urban governance, and spatial planning initiatives. Community organizations have used Web 2.0 mapping technologies for a variety of reasons, including, first and foremost, to collect information that can be shared in a way that effectively communicates a message to the organizations members, associated partners, and with whom an organization is attempting to form a relationship. Increasingly, the boundaries between the spatial data user and spatial data creators have blurred.

**Representation and integration of local knowledge**

The nature of spatial knowledge production in PPGIS has been explored through numerous studies. PPGIS research has emphasized the value of indigenous or local knowledge and ways of incorporating local knowledge and indigenous knowledge into GIS databases. Significant efforts have been made to incorporate local knowledge and value-based data into GIS (Cope and Elwood 2009), including community organizations’ active reworking of the meaning of mapping through traditional GIS tools, creative engagement of visualization and multimedia representations, and efforts to rewrite GIS software to embody multiple forms of spatial data.

A number of studies have sought to investigate the existing everyday practices of conventional GIS that integrate local knowledge. Integration of local, experiential knowledge with public datasets creates nuanced spatial knowledge that has proved to be significant for marginalized communities. In the context of urban revitalization processes, studies have shown that community organizations use such spatial knowledge to enhance inter- and intra-organizational communication, to legitimize existing experiential knowledge for obtaining action, to monitor neighborhood conditions for strategic planning, to prepare for organizational tasks and funding recruitment efforts, to enhance service delivery
tasks, and to explore spatial relations to challenge or reshape urban policies. PGIS projects from the Global South also show that there are three major uses facilitated through community mapping: communicating information within communities, communicating information between neighboring communities, and communicating information to outside groups.

Visualization has been acknowledged as an important approach used by community groups in PPGIS to state their position, communicate with stakeholders, and contest official discourse. PGIS projects increasingly incorporate multiple ways to enhance such visualization. Maps are a popular form of visualization used by community organizations. In inner city revitalization, maps help organizations to create spatial strategies to both combat problems and identify opportunities of growth. While mapping of neighborhood problems are important to seek action, community asset maps that showcase the neighborhood’s positive aspects are also frequently used to attract investment opportunities. Such maps can range from simple thematic maps to more complex maps employing kernel density functions. Mapping of neighborhood indicators can be used to compare a neighborhood’s condition with that of other neighborhoods, represented at multiple scales (inner-city scale, municipal scale, metropolitan scale). Such scale jumping can be used as a spatial strategy to highlight the low quality of life in a target neighborhood in order to demand resources and policy changes.

The intersection of critical (or interpretive) epistemologies and qualitative methodologies has led to innovative modes of visualization in PPGIS, as it continues to grapple with underlying questions concerning human identity, lived experience, situated and contingent knowledge, power and positionality, as well as concerns of social oppression and spatial exclusion. Representation is particularly critical where local knowledge is to be integrated. In order to incorporate local knowledge into the building of GIS databases, efforts have been made to include value-based, traditionally intangible information through efforts such as geoethnography. Feminist geographers employing qualitative ethnographic research have created feminist visualization techniques to more effectively understand subjectivities and represent the different experiences and emotions of individuals (Kwan 2002). Such approaches provide multiple ways of integrating indigenous and experiential knowledge in PPGIS. PPGIS studies have thus incorporated pictures, photographs, narratives, videos, sketches, architectural footprints, and so on to construct spatial knowledge. Through the use of a multimedia community integrated GIS project in South Africa, Harris and Weiner (1998) demonstrated how different forms of community knowledge could be integrated to create powerful spatial narratives. Through innovative approaches, PPGIS projects show that multimedia files can be attached as attributes to point layers to represent oral histories, nontraditional weighting schemes for site suitability can be evaluated, and language-specific user interfaces can be created.

PPGIS projects have also demonstrated the use of alternate mapping techniques, such as mental mapping, sketch mapping, and participatory photo mapping, in order to include qualitative data into GIS. Through the act of drawing, a mental map reveals significant dimensions of memory and human perception. Features become emphasized or de-emphasized, revealing individual perceptions and experiential knowledge about places. However, it is important to avoid any normative use of mental maps that assume humans possess perfect locational knowledge.
The aforementioned critical and qualitative turn should not be interpreted as a comprehensive rejection of quantitative research methods. Sketch-maps and participatory photomaps, on the other hand, are spatially referenced and can be considered as examples of methodological pluralism. Narratives and videos can also be georeferenced and used in conjunction with traditional GIS activities. Through these mixed, multi, or hybrid-methodological approaches to human spatiality, qualitative GIS offers innovative opportunities to bridge the intractable qualitative–quantitative divide in the direction of synergism and holism. Applications of such qualitative GIS techniques in PPGIS projects have provided greater opportunities for marginalized individuals and communities of varying age groups, race, and gender to counteract hegemonic narratives of exclusion, to incorporate multiple perceptions of reality, to emphasize local issues of importance and concern, and to draw attention to persistent silences in data.

There have also been notable efforts to rewrite existing GIS software to accommodate the needs of community groups and to challenge hegemonic forms of data representation. In an effort to bridge the qualitative–quantitative methodological divide, qualitative GIS projects have been created to blend GIS software with computer-aided qualitative data analysis software. In order to facilitate the ease of GIS use among marginalized communities, efforts have been also made to create interfaces customized to the unique needs of communities. Nonetheless, while rewriting the code of GIS opens greater opportunities of multiple data representations, given the level of technological expertise required this approach is largely initiated by academia or public agency rather than by the community groups.

Further, while access to spatial data and GIS can enable disadvantaged groups to more effectively participate in planning and policymaking activities, there is also evidence of contradictory outcomes, as such access may empower certain groups within the organization while disempowering others. The ability to use computers and spatial technology varies widely among marginalized groups, leading to exclusion of some members of the community. The contradictory outcomes of PPGIS can be reflected in web-based mapping projects as well. Access to high-speed Internet differs among different social groups and in different regions. The ability to use the Internet for PPGIS is also quite variable, based on race, class, age, and gender divides. It is, therefore, important to inquire into whose knowledge is being represented through PPGIS.

Evaluating outcomes in PPGIS

Scholars have called for critical examinations of the role of GIS usage in a variety of contexts as to whether GIS has empowered marginalized communities or consolidated existing power relations. PPGIS outcomes can differ based on the goals of the project, the local contexts of participation and access, and the internal organizational characteristics.

It is recognized that there are multiple levels of public involved in PPGIS practices within particular contexts. Representing the various perspectives of multiple publics within spatial data structures remain key research challenges. Similarly, participation is a multifaceted process. Many studies have drawn upon Arnstein’s (1969) concept of a participation ladder to illustrate the levels of participation, ranging from tokenism at the bottom to collaboration in the middle and to citizen control at the top.

Another concept that has been widely associated with PPGIS is the notion of empowerment.
PPGIS research shows that it leads to empowerment of marginalized groups. The concept of empowerment has a wide array of definitions. In particular, Elwood (2002) provides an insightful multidimensional analysis of empowerment in the context of GIS usage in community planning. In this framework, empowerment contains three dimensions: distributive change, procedural change, and capacity building. Among these three dimensions, distributive change is the least sustainable and capacity building is the most sustainable. A PPGIS project may result in all of the three dimensions of empowerment and be reflected in multiple forms. Sustainability of PPGIS projects is also a factor in such discussions. Ultimately, empowerment must be seen as a gradual process and not as the end product of PPGIS.

Notions of expanded participation also may presuppose some degree of homogeneity of benefits among those involved in PPGIS. However, various studies have noted that the notion of participation in PPGIS is complex and needs explanation. The notion of individual participation varies according to cultural norms and the notion of participation in indigenous or non-Western cultures is significantly different from that of Western culture. Social structures based on class, race, and gender divides also shape participation. Further, even in the same group, certain individuals may be better able to participate than others. Similarly, all factors being equal, certain organizations are better able to participate than others. In other cases, widespread participation may not be desirable. Similarly, the use of the word participatory may also be problematic, as it implies a need for an intermediary actor or institution.

Evaluating the effectiveness of increased participation through PPGIS has not been easy and is shaped by its contexts. PPGIS activities among indigenous communities show a more activist and confrontational stance. Their goals include gaining recognition of land rights, protecting traditional land, gathering and guarding traditional knowledge, and achieving social justice. In contrast, PPGIS activities in urban governance are undertaken by community organizations in a framework of accommodation and collaboration, which may not lead to any alterations in power.

Political and economic policies that shape governance and policymaking are also deeply influential in shaping the nature of PPGIS. The impacts of neoliberal policies have significantly affected the process of governance and citizen participation across the world, and such impacts have been felt on PPGIS practices as well (Ghose 2007). In particular, the shift to collaborative governance emphasizing public–private partnership and advocating a leaner government has reconfigured the role of the state and led to the rise of new intermediary stakeholders. Simultaneously, citizen participation in governance has been more formalized, in order to shift many of the responsibilities of the state upon the citizen groups. Collaborative governance has led to considerable data sharing between the government agencies and citizen groups, and, consequently, the use of spatial information and mapping among citizen groups in such collaborative governance has become more common. However, cutbacks in state budgetary and funding support for community organizations have reduced the capacity of organizations to participate effectively in collaborative governance. Consequently, community groups still lack sufficient resources to implement in-house GIS, resulting in variable PPGIS productions. The effectiveness of participation and spatial knowledge production is now increasingly dependent upon internal organizational factors (Elwood and Ghose 2001). In particular, the organizational ability to navigate the local politics of turf and building relations among stakeholders plays an
important role in shaping PPGIS outcomes. Relationships among stakeholders in PPGIS range from cooperation, compliance, and collaboration to control. Studies show that grassroots organizations must build dense networks of relationships to build and sustain their PPGIS efforts. Further, dynamic networks of support must be built at different scales (local, regional, national) among community groups and other stakeholders (Ghose 2007). Community organizations in urban governance thus increasingly undertake a cooperative stance rather than a confrontational stance. Within the context of urban governance, marginalized community organizations therefore tend to utilize PPGIS projects for effective navigation into the mechanisms of governance. Their projects have primarily four functions: administrative, organization, tactical, and strategic. These functional categories mirror public and private sector goals, and are similar to increasing efficiency in tasks such as map production, reducing redundancy in databases, and improving effectiveness in decision-making. Certain scholars have identified PPGIS as a mode of “collaborative decision support.” Here, PPGIS has been perceived to add value at several stages of the decision-making process, improving the articulation of stakeholders’ views, increasing individuals’ or groups’ understanding of technology, making complex decisions more transparent and objective, augmenting deliberation and consensus, furthering communication and linkages among internal participants and between internal and external parties, disseminating or sharing information, resolving conflicts, and enabling greater exploration of ideas.

Evaluation of the effectiveness of PPGIS projects has been overall a difficult process. Certain scholars argue that PPGIS projects should be assessed based on (i) appropriateness and match with an organization’s existing activities, (ii) adaptability to local conditions such as local culture, political climate, (iii) fitness to current organizational goals and capacity, and (iv) ability to be integrated into broader societal goals. Others argue that the contingent and place specific nature of PPGIS make it difficult to craft quantifiable assessment measures, and qualitative assessments of satisfaction with PPGIS projects should instead be practiced.

In conclusion, PPGIS practices have been increasingly employed in diverse areas across the world. Major barriers of access to spatial data and technology remain a significant challenge for many disadvantaged groups. At the same time, there have been numerous creative engagements with GIS to incorporate local and multiple forms of spatial knowledge through traditional GIS software, rewriting GIS as well as the recent combination of Web 2.0 technologies with geospatial technologies. Research has shown that PPGIS practices have been highly context-dependable and there are variable outcomes of empowering community organizations. Existing studies have shown the depth and breadth of PPGIS practices. However, there is very little research on the longitudinal investigation of PPGIS practices. This may partly reflect the challenging issue of sustainability of PPGIS projects. Yet, a long-term investigation of PPGIS practices within a particular place will provide important insights into the above three theoretical streams reviewed.

**SEE ALSO:** Critical GIScience; Empowerment; Environmental (in)justice; Participatory action research; Social justice

**References**


Qualitative data, acquisition

Steve Herbert
University of Washington, USA

Places are made through meaning-laden processes.

This statement seems simple enough, but reckoning with it fully is not. Although a wide range of research questions animate academic conversations in geography, place-making remains a key touchstone to the discipline. Social groups importantly construct themselves through the various means by which they build places. Through the construction of the built environment, through various symbolic acts that convey meaning, through the maintenance of boundaries, societies create an identity and a presumed durability.

Notably, these acts of simultaneous place- and group-construction are ongoing processes. Certainly many enduring structures, such as the capitalist space-economy (Harvey 2007), guide these processes. Yet all such structures are only reproduced through daily action. Their endurance is, thereby, a product of thousands of conscious and unconscious acts. Further, these acts are laden with meaning; they bear a symbolic significance to the group members who perpetuate them (Delyser 1999; Herbert 2000).

Because of the importance of these meaning-laden processes to the construction of places, many questions central to the discipline of geography require recourse to qualitative data. Without in-depth understanding of how and why people build places, and how they render such action meaningful, how social processes and spatial forms are mutually reinforced cannot be fully comprehended. Indeed, there are few research agendas within the discipline of geography that would not profit from qualitative data.

There are, of course, multiple specific goals that might be pursued through gathering qualitative data; for example, the wish to understand the relationships between talk and action. It is one thing for social groups to create any number of discourses through which they articulate a worldview and a set of desired practices. It is another thing entirely for them to actually enact those practices. Indeed, disjunctures between talk and action frequently provide insights into such questions as the importance of ideology in maintaining power (Hall et al. 1978; Willis and Aronowitz 1981). Alternatively, the wish might be to understand how certain cultural schemas are unique to a given group, and how they generate a specific identity that is re-created daily through spatial practices (Herbert 1997; Sewell 1992). If such schemas are contested, either within or across groups, then conflict can emerge. These conflicts may take spatialized form, as when residents of one area resist the incursion of differently-situated others (Coleman 2005; Nevins 2001; Varsanyi 2011). These conflicts can best be dissected through sensitively gathered qualitative data.

These examples hardly exhaust the productive analytic uses to which qualitative data can be put. But they do illustrate how vibrant geographic research profits immensely from attention to the meaning-laden processes through which socio-spatial life is built.
QUALITATIVE DATA, ACQUISITION

Some of the central challenges that a qualitative researcher faces in acquiring useful data are reviewed in this entry. Also reviewed are the two key modes of qualitative data gathering – ethnography and interviews. The key strengths and weaknesses of each of these methods is outlined and when each might be employed most productively considered. As will be clear, gathering high-quality qualitative data is neither quick nor easy, yet such data enable geographic analyses of great depth and insight.

Key challenges in gathering qualitative data

Regardless of which type of qualitative data a researcher seeks to acquire, s/he will face numerous challenges. Though these challenges are rarely overcome easily, they can be approached with sufficient sensitivity to minimize their disruptive potential. Four key challenges are considered here: choice of case, access, trust, and validity.

Choice of case

The intrinsic value of qualitative data lies in its depth. Through extended exposure to a particular group, the qualitative researcher can discern many of the less visible wellsprings of social action. Core group values emerge through exposure to both discourse and behavior. Fissures between group members become more apparent; their significance can be analytically revelatory. Rituals can reveal how meaning is created, and perhaps contested. The meanings of such rituals, or any other form of patterned social action, are often opaque when fieldwork begins. As the qualitative researcher gains greater exposure to the group, what was once mysterious becomes more susceptible to thoughtful analysis.

Such fieldwork takes time. Multiple observations of the same phenomenon are commonly necessary before its significance is clear. Prolonged familiarity with the group’s members is often required to enable the creation of trust, and the more open conversations such trust allows. Because of these extensive investments of time, the qualitative researcher must restrict the number of cases s/he examines. For this reason, qualitative research is often critiqued for its failure to create data that can be generalized across multiple cases.

This is no small concern but it need not cripple the qualitative research enterprise. It does mean that any case needs to be selected carefully. Three common strategies can achieve the goal of finding appropriate cases to study.

One strategy is to find a case that is representative of others. It is often true that a particular social group possesses dynamics that occur in other, similar groups. Because of this, any generalizations that emerge from an in-depth analysis of one group can be extended to others (Stake 1995). A second strategy is to find a case that is somehow unique, one that does not conform to expected patterns. The effort to understand just how and why that group’s uniqueness occurs will likely require the researcher to extend existing theory in productive fashion (Burowayi 1998). A third strategy is to adopt an explicitly comparative approach. Here, the researcher seeks to find two or more groups that are similar in some respects but different in others. The effort to explain these similarities and differences through analysis of in-depth qualitative data should also work to enable the refinement of existing theory (Miller 2000; Ragin 1987).

The investment of time in a single case is the hallmark of good qualitative research. Because this reality places limits on how many cases to open for examination, the selection strategy
must be careful. Yet judicious choices here can yield generalizability.

Access

Careful case selection is the necessary starting point of qualitative inquiry. Yet that does not mean a researcher can gain access to the group most ideal for study. Indeed, the granting of such access is often no small matter for the members of the group in question. Those members must tolerate a newcomer for an extended period. They will likely be asked any number of probing questions; they may have their routines disrupted. They typically will engage the researcher with the full knowledge that they will not control the ultimate results of the work. For groups wary of the attention the researcher may bring to them, it may prove difficult, if not impossible, to grant access to their daily behavior.

Given these realities, the researcher must be thoughtful about how to approach a group to which s/he desires access. A key consideration here is how the researcher wishes to be positioned vis-à-vis the group in question. Classic fieldwork placed the researcher in the role of the seemingly dispassionate scientist. Here, the goal was to remain as “objective” as possible in delineating the kinship structures or other key cultural determinants of social action. From this perspective, the group under study is primarily fodder for academic analysis, and arguably profits little from interactions with members of the research community. While it remains the case that academic researchers are required to publish rigorous and theoretically oriented analyses to advance their careers, it is increasingly likely that any group that grants access will want something more from the encounter. Because of this, the researcher needs to consider just how the group in question might itself benefit from the research. It may be the case that the group welcomes insights from an outside observer. A public agency, for example, might wish to better understand its internal processes. A more politically motivated group might wish to see the researcher conduct work that will assist it in pursuit of particular goals (Routledge 2010). It might even wish to negotiate the contours of the research itself, and remain in active collaboration with the researcher. Such “participatory action research” holds great appeal to researchers who seek to combine scholarship and advocacy (Kindon, Pain and Kesby 2008; Reason and Bradbury 2001).

Yet such collaborations must be approached carefully. However well-meaning a researcher may be, the reality remains that the career advancement of an academic requires a particular type of analysis that is likely distinct from that desired by the group. There will thus always be some degree of self-interested expropriation from the group by the researcher, and it is disingenuous to pretend otherwise. Ethical practice demands that the researcher never lose sight of this reality, and communicate that to the group under study. In addition, the researcher must be careful just which point of contact to approach initially. It may be that certain individuals who can grant access will wish to steer the researcher in particular directions, without revealing their motives for doing so. It may also be that entering the group under the aegis of one individual may lead others to cast suspicion on the researcher. In this author’s experience, the fact of being granted access to the Los Angeles Police Department by a new and reform-minded police chief meant that many feared a “spy” intent on uncovering officer misdeeds. It required significant exposure to the group for that suspicion to erode (Herbert 2001).

The challenge of gaining access to a group is a significant one and needs to be approached thoughtfully. It is necessary to think carefully
about just how a plausible claim to assist the group can be made without losing sight of the fundamental reality that the researcher is doing rigorous social science. And it is necessary to try to discern a route into the group that will allow the development of trusting relations as quickly as possible.

Trust

Of course, trust is not always easily won. Even if a researcher finds a good entry point, much work must still occur to ensure that group members will be forthcoming. Research subjects can commonly be expected to possess concern about how the author will use the data s/he gathers. How will the group be represented? Will the analysis be critical? Will it be accurate? Will individual group members be identifiable? Who is the intended audience? Even if such questions are not explicitly articulated, they often influence a research subject’s initial interactions with the qualitative researcher. It is, therefore, necessary to address these concerns. Several strategies can assist here.

It is imperative that the researcher approach the group in a nonjudgmental fashion. Not only is this stance respectful of the group, but it conforms with sound social scientific practice. The researcher’s task is to understand the world as it is experienced by the group members. This is impossible without bracketing pre-existing conceptions to the greatest degree possible. If the task is to discern the underlying logic of various sociospatial practices, then open-mindedness to group practices is essential. Research subjects will quickly discern how genuinely open a researcher is to their experiences, and this will strongly influence their ability to trust.

A second key component of trust building is the process of informed consent. Human subjects review is now an inescapable component of qualitative data gathering. Such review will invariably require that researchers secure written consent from their subjects to participate in the research. The consent form will include a description of the research and its intended purposes, and a discussion of how participation might provide unwanted consequences for the subject. The researcher is expected to review the form with the subject and to address candidly any questions that emerge. Indeed, this discussion will often be the first substantive conversation between researcher and subject. Trust will develop if the researcher approaches this discussion in an open-minded fashion, and allows the subject ample time to make a considered decision about participation.

A third means to increase trust is to provide explicit protections against publicity to the group. This will commonly include grants of anonymity to individual participants. It may also include disguising the group as much as possible, by providing both a fictional name and locale. Such fictions might be considered an unwelcome component of social science practice, which encourages researchers to make their data gathering replicable. Yet without these protections, group members may well refuse to participate, or they may choose to provide only superficial information. The necessity of trust in the researcher–researched relationship may demand that one disguises enough about the group to prevent a reader from discerning its identity.

The ability to gather substantive qualitative data requires the researcher to assiduously work to earn the trust of the research subjects under study. Such work should be motivated not solely by the researcher’s instrumental goals of advancing his/her career, but by a genuine interest in understanding the world from the subject’s perspective. If such interest exists, and can be communicated both directly and indirectly, trusting relationships between researcher and subject can develop.
Validity

Without trust, the qualitative researcher cannot gather the data necessary for insightful analysis of group processes and meaning structures. The significance of such data cannot be overstated. Without it, the researcher cannot analyze a particular phenomenon with sufficient depth and breadth. The researcher must be able to persuade the reader that s/he understands the group well enough to make generalizations about its patterned practices. Further, the researcher must convince a reader that the data under analysis are reasonably comprehensive, and that they were gathered via ethical and effective practices. In other words, the data must be considered valid if the researcher is to generate high-quality analyses.

The acquisition of valid data requires trust, but it also requires more. The researcher must be ever attentive to the reality that any single group member’s perspective is partial. This is true even if the subject is being completely forthcoming, but is especially the case if a subject is choosing to structure a narrative about the group for some strategic purpose. An informant may, for example, be looking to use the researcher to tell stories that would enhance his/her own status. Because of this, the researcher must endeavor to engage as many subjects as possible as deeply as possible. Further, the researcher will want to observe the group in multiple settings, to be able to discern just how specific contexts shape group processes. Otherwise, the analysis could rest on data that cannot bear the weight of generalization.

The data must thus capture multiple perspectives from within the group. They must also capture those perspectives in sufficient depth. In almost all instances, this will require extensive involvement with the group in question. For that reason, qualitative research is typically time-intensive. Ideally, the researcher can stay involved with the group for long enough to reach a point of saturation. In other words, the researcher gathers data until new information becomes scarce. In many instances, the lack of time or resources may prevent the researcher from reaching this ideal situation. In those cases, the researcher must rely on the data s/he possesses, and explain to the reader the degree to which the data might be limited.

As the foregoing discussion illustrates, the qualitative researcher has some important decisions to take before entering the field, and some important work to do once there. Without considering these decisions carefully, the researcher may end up with insufficient or invalid data. Yet through judicious decisions about which case to study, and assiduous attention to establishing trusting relations with research subjects, the qualitative researcher can capture data that enable insightful analyses.

Another set of important decisions revolve around just which qualitative method to pursue. This is considered in the following section.

Choice of method

It is one thing to thoughtfully choose a case and to establish ethical and productive relations with informants. It is another thing entirely to actually collect data. Just what type of data does the researcher desire and what is the best means by which to acquire the data? Which of the available data collection methods is most sensible in any given situation? These questions do not commonly yield easy answers, but require thoughtful consideration before entering the field, and sometimes even after data collection starts. Each of the principal qualitative methods – ethnography and interviews – possesses certain advantages and disadvantages. Without complete cognizance of what any given method can and cannot provide,
QUALITATIVE DATA, ACQUISITION

the researcher cannot ensure the acquisition of data most essential for the project at hand.

Ethnography

An ethnography consists of a prolonged exposure to a particular social group. In the process, the researcher becomes attuned to the daily rhythms and routines of the group. Indeed, the researcher will be required to become active in the group dynamics, at least to some degree. For that reason, ethnography is sometimes described as “participant observation” (Watson and Till 2010). This term suggests one of the key advantages of ethnography: the researcher, as newcomer, must acquire the necessary cultural competence not only to understand the group but to participate as a member in good standing. All social groups require some specific knowledge to enable competency as a member. The ethnographer acquires this knowledge through daily practice, but does so with sufficient reflexivity to document the process.

The key advantages of ethnography, then, derive from sustained exposure to the group and some degree of participation in its processes. A lengthy fieldwork experience should also enable the steady accumulation of trust and, with it, more forthcoming revelations from group members. Ongoing observations of the group also position the researcher to make assiduous distinctions between talk and action. Group members may say one thing, but their actions may suggest another. For that reason, a qualitative researcher who relies solely on interviews may gain an incomplete understanding of how the group actually functions.

The main disadvantage of ethnography is simultaneously its main advantage: it takes time. This is a resource that few possess in abundance. In addition, ethnography requires sufficient interpersonal skills to develop trusting relations with subjects. Further, the investment of time in a single case places significant pressure on the case selection decision. The researcher must make sure that an analysis of the case under examination can yield broader generalizations of interest to other social scientists (Herbert 2010).

That said, ethnography uniquely enables the acquisition of data of remarkable depth. Provided that the researcher has the necessary time and interpersonal skills, and chooses the right case for examination, s/he can generate social scientific work of considerable significance, as the rich ethnographic literature more than amply illustrates. Of course, not all researchers have the time required to carry out intensive ethnography. Fortunately, high-quality data can result from interviews.

Interviews

Insight into group processes and meaningful structures can emerge from fruitful conversations with group members. Interviews built on thoughtful questions and empathic listening can yield valid data. If done well, interviews can be a very efficient means to achieve a researcher’s goals (Rubin and Rubin 2011; Spradley 1979).

The collection of useful interview data requires several considerations. The first is to isolate a population of interviewees that can provide illuminating data. Success here rests upon the same considerations as for ethnography: identifying well-positioned informants, approaching them respectfully, establishing trusting relations. The second is to structure the interviews properly. Some researchers may wish to ensure a high degree of standardization across the interviews, and hence will employ a well-structured schedule of questions that is repeated identically with all informants. Most interviewers, however, will want to create enough elasticity in the interview to enable the emergence
of unanticipated ideas and themes. For that reason, semi-structured interviews are more common. These ensure that similar terrain is covered in all interviews, but allow the conversation – and potentially the research itself – to expand to include additional and relevant data. Semi-structured interviews also allow the interaction to develop into something approaching a conversation, and thereby help the subject relax. Success here rests upon a third key ingredient in interviewing – establishing trust. Such trust is more likely to result when the researcher is fair, ethical, and honest in all interactions, and when the interviewee feels as if he or she is being listened to empathically.

Unlike ethnography, interviews typically disallow much comparison between talk and action. If unable to observe the research subjects engaging in the social relations under study, the interviewer is left largely to take subjects’ at their word. For that reason, any research project that relies heavily on interviews can benefit from even a minimal amount of participant observation. Even if this is not possible, a well-crafted interview-based project can yield useful data. Targeted conversations with knowledgeable individuals can be enormously efficient means to gain insight into a particular social group.

Most interview-based studies rely on a series of conversations with single individuals. Some, however, bring together multiple interviewees together for a group conversation. Such focus groups can be advantageous (Bosco and Herman 2010). One advantage is that members of the group can challenge accounts from each other. This limits the ability of any one interviewee to expand on an idiosyncratic – and perhaps even erroneous – narrative about key group processes. A second advantage is that group interaction processes can be rendered at least partially visible. In the process, the researcher can learn of intergroup tensions or other dynamics that may be quite relevant for efforts to answer key research questions.

Given the researcher’s desire to cultivate a wide-ranging conversation in the focus group, the schedule of questions is commonly short. This reality places pressure on the researcher to construct the best possible set of questions. It also means that the conversation will necessarily be limited, and hence reduce the researcher’s capacity to uncover unanticipated components of group processes. That said, if focus groups are carefully constructed and skillfully handled, they can unearth insightful data.

Of course, not all qualitative research projects rely exclusively on either ethnography, interviews or focus groups. Indeed, given the inherent strengths and weaknesses of any of these used in isolation, a robust research strategy might employ these in some combination. A project based largely on ethnography may profit from more extensive conversations with particularly well-placed informants to enable a targeted consideration of one particular dynamic. A project based largely on interviews with individuals may profit from focus group discussions to enable some insight into group processes. Not all researchers possess the time or other resources to accomplish a mixed methods approach, but these may provide the best means to ensure the collection of comprehensive and valid data.

Conclusion

The reality that places are made through meaning-laden processes guarantees that the skillful collection and analysis of qualitative data will remain essential to the practice of human geography. While such data are never gathered easily or quickly, they possess the capacity to illuminate and explain a wide range of sociospatial behavior. The challenges of gathering
Qualitative Data, Acquisition

High-quality qualitative data are numerous and substantial, but they can be addressed effectively, and thereby provide unparalleled opportunities to understand and document the complex undercurrents of sociospatial life.

See Also: Ethnography; Focus groups; Interviews; Qualitative data

References


Qualitative data

Iain Hay
Flinders University, Australia

Qualitative data comprises non-numeric information recording and describing the ways in which people regard, experience, understand, and construct the places they inhabit. It is typically drawn from naturalistic settings and reveals what participants understand about how and why specific events or relationships have come to pass. It can also capture the worlds of researchers’ and participants’ senses, conveying ideas like how a place smells, how something tastes or feels, what it looks like, or what memories it evokes. As data, qualitative information comprises records or descriptions, but does not include explicit judgments about content – notwithstanding the subjectivity or positionality of any observer gathering the information. Qualitative data is customarily and contentiously distinguished from quantitative data, which comprises numerical information, such as “how many,” “how much,” “how often,” all capable of various forms of statistical manipulation and interpretation.

Qualitative data: variety and verstehen

Qualitative data emerges from research seeking to understand the qualities of things, rather than their quantity. It is something that can be observed, but not measured. Although qualitative data might include accounts of taste and smell (e.g., describing the reek of an abattoir visited on a fieldtrip), it more typically encompasses text, images or sounds, including the words and actions of participants as read, heard, or observed by a researcher (e.g., a researcher’s observations of teenagers’ behavior in a shopping mall and those teenagers’ own explanations for their behavior). It may also describe or express emotions and sentiments (e.g., how a refugee feels about their former homeland).

Qualitative data also emerges from research that seeks to reveal subjective understandings or the meanings of action from the perspective(s) of participants (e.g., the value Australian suburban homeowners attach to recreational gardening; how people regard their own daily commuting contributions to global warming). Rather than treating research participants as objects of the study, they are held to be subjects, with their own capacities to comprehend and craft their worlds. In this regard, qualitative data materializes from the pursuit of that which German antipositivist sociologist Max Weber termed verstehen (i.e., understanding the meaning of action from the actor’s point of view).

Qualitative data can take many different tangible and intangible forms. In some modes of research, qualitative data has been taken to mean records that have a nominal or categorical scale of measurement drawing, perhaps, from a single-word answer to a question (e.g., “From what country have you most recently traveled?”) or describing, gender, occupations of residents in a neighborhood, or car brands (e.g., “what type”). These categories can then be assigned numerical values, describing, for instance, the number of: men or women; carpenters, medical practitioners, and teachers; or Fords, Toyotas, and Renauls in the neighborhood. While some

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0421
QUALITATIVE DATA

Statistical tools may be applied to such qualitative data, they are largely limited to descriptive statistics (e.g., mean, mode, standard deviation).

Numerical representations of data are used in some qualitative contexts, perhaps most notably in content analyses. For instance, in their study of The Australian newspaper’s reporting of Seattle antiglobalization protests, McFarlane and Hay (2003) included simple numerical analyses of the extent and type of reporting associated with different groups involved in the disputes (e.g., protestors, politicians). However, understanding of qualitative data as categorical information that can be enumerated offers a much narrower appreciation than is now most commonly understood.

It is much more usual for qualitative data to be regarded as including photographs, audio recordings, weblogs, policy documents, diary entries, web content, documentaries and movies, personal correspondence, observational and interview notes, or even advertising brochures. And, as these forms imply, qualitative data may be gathered or produced through a broad array of techniques, including interviews, focus groups, observation, and participatory methods (Hay 2016), or is retrieved from public and private archives.

Patton (1990) aligns qualitative data with the three dominant forms of qualitative methods. First, the data from interviews or focus groups comprises direct quotations from informants, outlining matters such as their experiences, feelings, beliefs, and interpretations. Second, data from observations includes the researchers’ thorough descriptions of observable activity. Third, data from textual sources, including quotations and other selections, comes from a diverse array of documents, including letters, diaries, organizational records, official reports, newspaper articles, or government memos and randa. These textual sources might also be expanded to include those other forms described previously, such as photographs, movies, or brochures.

Qualitative data need not be gathered first-hand. It is increasingly being made available through national and other archives, such as ADA (Australian Data Archive) Qualitative (https://www.ada.edu.au/qualitative/home), the Irish Qualitative Data Archive (http://www.iqda.ie/), the United States’ National Opinion Research Center at the University of Chicago (http://www.norc.org/Pages/default.aspx), and the UK Data Archive (http://www.data-archive.ac.uk/). These agencies typically acquire, curate, and provide a central access point for qualitative social science data produced in or about a specific jurisdiction or territory. Their collections offer access to high quality contemporary and historical information that might otherwise be unavailable to researchers for reasons such as cost, time, or physical inaccessibility. Moreover, the collections facilitate a range of work, such as comparative research and re-analysis, as well as offering useful real-world materials for research training. Some researchers are now also making available qualitative data, such as interview recordings or transcripts, through scholarly journal websites (e.g., Geographical Research) as part of those journals’ emerging dataset and code availability policies.

Trustworthiness, quality, and quantity

Whereas qualitative data is associated with a range of epistemological positions (e.g., postmodernism, poststructuralism, interpretivism), most quantitative data gathering is underpinned by positivist philosophical foundations. The hegemony of positivism and that philosophy’s long-term methodological associate, the so-called scientific method, with their conjoint
Table 1  Provisions that may be made by a qualitative researcher wishing to address Guba’s four criteria of trustworthiness.

<table>
<thead>
<tr>
<th>Quality criterion</th>
<th>Possible provision made by researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credibility</td>
<td>Adoption of appropriate, well-recognized research methods. Development of early familiarity with culture of participating organizations. Random or purposeful sampling of individuals serving as informants where appropriate. Triangulation via use of different methods, different types of informants, and different sites. Tactics to help ensure honesty in informants. Iterative questioning in data collection dialogues. Negative case analysis. Debriefing sessions between researcher and superiors. Peer scrutiny of project. Use of “reflective commentary” (i.e., researcher’s evaluation of the project as it develops). Description of background, qualifications and experience of the researcher. Member checks of data collected and interpretations/theories formed. Thick description (behavior and its context) of phenomenon under scrutiny. Examination of previous research to frame findings.</td>
</tr>
<tr>
<td>Transferability</td>
<td>Provision of background data to establish context of study and detailed description of phenomenon in question to allow comparison to be made.</td>
</tr>
<tr>
<td>Dependability</td>
<td>Employment of “overlapping methods.” In-depth methodological description to allow study to be repeated, though not necessarily to achieve identical outcomes.</td>
</tr>
<tr>
<td>Confirmability</td>
<td>Triangulation to reduce effect of investigator bias. Admission of researcher’s positionality, including their beliefs and assumptions. Recognition of shortcomings in study’s methods and their potential effects. In-depth methodological description to allow integrity of research results to be scrutinized. Use of diagrams to demonstrate “audit trail.”</td>
</tr>
</tbody>
</table>

Source: Adapted from Shenton 2004, 73.

Demands that data be measurable and that data gathering be replicable, have led to common beliefs that qualitative data lies outside the scope of legitimate, “scientific evidence.” Indeed, the very trustworthiness of qualitative data is sometimes questioned by sceptics whose key positivist principles of validity and reliability cannot be tested in true-to-life work in the same ways as may be possible under closely controlled conditions. In response to the ideology of scientism, educational researcher Egon Guba (1981) proposed four guiding principles to help ensure the trustworthiness of qualitative data.
QUALITATIVE DATA

1 Credibility (compared with internal validity) – would the data/findings be recognizable to study participants and comprehensible to those outside the study?
2 Transferability (compared with external validity/generalizability) – do the data fit in other relevant contexts?
3 Dependability (compared with reliability) – have proper research practices been followed?
4 Confirmability (compared with objectivity) – are the data a reflection of the participants’ ideas and experiences rather than those of the researcher?

As the points indicate, Guba’s influential ideas parallel the four quality criteria employed by positivist researchers, but they are not intended to be obligatory mechanisms for evaluating the worth of data.

Building on other scholars’ subsequent work elaborating these four principles, Shenton (2004, 73) very helpfully sets out in summary form a range of steps that might be taken by a qualitative researcher seeking to enhance the trustworthiness of their data (Table 1).

In other related and important work making a point similar to that of Guba, Andrew Sayer (2010) proposes that the qualitative counterpart of replication is corroboration. For example, to test qualitative data a researcher might seek to verify or substantiate the opinions or experiences of one interviewee by observing the context within which they are placed, or by interviewing other people involved, or those with similar experiences. It is useful to conceive of corroboration as resembling a legal trial in which various forms of evidence and testimony are called upon to test the veracity of police, witness, or defendant claims, although the crime itself is not repeated! But while this metaphor has some use, it cannot be stretched too far as it implies a quest for a singular truth, whereas qualitative data may in fact disclose multiple, competing, and valid perspectives on reality – which may indeed be its aim.

Yet another useful and related way of assessing the quality of specific qualitative data is to examine the extent to which it helps to produce an account that would make or does make sense to the actors themselves. In some types of research, such as participatory action research, where data and other outcomes are generated with participants, this form of corroboration should be more-or-less inevitable.

As part of the process of demonstrating characteristics of credibility, transferability, dependability, and confirmability, and therefore the trustworthiness of its data, a qualitative study needs to detail its workings, outlining for audiences explanations of how and why particular settings were selected (e.g., Why were the three specific factories chosen for a study?); how and why specific research activities were undertaken (e.g., Why were focus groups employed as the primary data gathering approach?); what actions were taken to ensure the ethical conduct of the research; and how and why different analytic, synthesizing, or causal themes were selected. Moreover, contemporary qualitative research also demands reporting on the researcher’s positionality (i.e., their social or political position relative to the people with whom they are working) and the ways in which researcher(s) and participant(s) regard one another. Rather than presenting themselves as dispassionate observers, qualitative researchers increasingly acknowledge and make explicit the various subjectivities they bring to their work, describing, for example, their politics, their socioeconomic status, and their ethnicity, acknowledging the significance these subjectivities might have for their analyses and for the ways in which their narratives may be written.
This leads to another key difference in the production of qualitative and quantitative data. Most commonly, the relationship between quantitative researcher and participant is mediated by an instrument (e.g., questionnaire survey, experiment), whose fitness-for-purpose can be readily assessed. The instrument itself may also be reused to assess the reliability/validity of the data gathered. By contrast, in qualitative research, the researcher is the data-gathering instrument, conducting interviews, leading focus groups, observing behavior... The value and potency of qualitative data that is produced depends, therefore, on the researcher’s integrity, their methodological skill and critical capabilities, the nature of their relationship(s) with participants, and mutual perceptions of one another. Generating good data from an oral history, for example, calls for much more than having a friendly chat with the informant about some period of their life. Instead, it requires: a strong background in interviewing, knowing when, for example, to prod and when to stay silent; a detailed background knowledge of the informant; and a highly developed appreciation of the context about which the participant is talking.

While most quantitative studies have statistical measures that can be used for determining how many data are required to make general statements about them and the degree of certainty that can be applied to those statements, there are no such numerical indicators of, for example, sample size, that can be applied readily to qualitative data. Instead, the concept of saturation (less commonly termed “redundancy”) is employed. At its core this concept challenges the researcher to whether sufficient informant/participant experiences have been gathered for development of the conceptual framework, theory, or explanation. Are informants’ accounts and interpretations being repeated and are incidents, features, and facts being verified by different sources? If no new or relevant information is emerging and the theory or account being produced appears to be robust, saturation is understood to have been achieved.

Analysis of qualitative data

The texts, sounds, and images that constitute qualitative data may be examined either as objects of analysis themselves (e.g., interpretations of a television travel documentary or regional tourism brochures) or as proxies for experience (e.g., diary notes of pioneer women in the US West). Given that qualitative data is not as rigidly defined in quantity or form as quantitative data, it is unsurprising that its analysis is not characterized by the formal rigor and consistent structures customarily associated with the more standardized approaches of quantitative research (and which arguably yield “rigor mortis,” characterized by banal empiricism, reductionism, and the ecological fallacy). This variation is not to imply, however, that high-quality analysis of qualitative data requires any lesser degree of exactitude or expertise than is required for quantitative analyses. Qualitative data analysis runs the gamut from artistic endeavor, performed by a scholarly virtuoso, to a more structured process, suitable for establishing credibility and providing clear conceptual trails for subsequent evaluation. While the procedures of the former cannot be elaborated, making sense of qualitative data in the second sense is most commonly termed coding – a simple idea often overcomplicated. Coding refers to a range of different programmatic ways of configuring the sometimes messy raw data from interviews, observations, and texts into materials that are more coherent, comprehensible and which might better inform research.
QUALITATIVE DATA

The form(s) of coding that might be chosen depend on data, methods, and research intent. Richards and Morse (2013) usefully identify four successive forms of coding. Descriptive coding typically involves simple categorization of data by some of its key characteristics, so the data can be more easily retrieved in future. For example, a researcher might code interviews by type of informant (e.g., sex, age, occupation), interview location (e.g., workplace, home), or time/date. Topic coding involves reviewing and categorizing data by subject, taking up questions like: “What were they talking about?” “What was this document about?” This form of coding can describe accurately the content of material as well as offering new ways of making sense of the data. Metaphorically speaking, analytic coding takes the researcher above the data, searching to develop ideas, categories, and concepts that seem apparent within it. It is about evolving or germinating categories. And theme coding, which may occur at any stage of data interpretation, concerns identifying pervasive themes, compelling “stories,” or undercurrents that flow through and from the data. Coding is a foundation to analysis: it moves the researcher towards analysis through the process of creating and developing abstractions from the data (Richards and Morse 2013, 151). Coding can be completed manually or using some of the range of computer assisted/aided qualitative data analysis software (CAQDAS) now available (e.g., ATLAS.ti, MAXQDA, NVivo).

Whether performed by a wild-thinking scholarly prodigy or a more pedantic researcher sweating over a keyboard, qualitative data analysis is open-ended and largely inductive (i.e., inferring likely explanations or accounts from particular instances) or abductive (i.e., seeking the likeliest explanation for an incomplete set of observations) and, therefore, its detail cannot be described in the form of a predetermined sequence of analytical stages.

On the powers and liabilities of qualitative data: clearing up misunderstandings

In a very widely cited paper, Danish scholar Bent Flyvberg (2006) points to key misunderstandings associated with case studies, regarded here as being emblematic of much qualitative data. These five troublesome confusions and Flyberg’s correctives are summarized in Table 2. This table is intended to offer a summary indicator of both the ways in which qualitative data is frequently devalued (through the ideology of scientism) and some corrective points about how and why such data can, in fact, contribute to the development of enhanced understanding of social (geographic) worlds.

As several of Flyvberg’s correctives imply, qualitative data offers depth of understanding of sociospatial worlds, providing comprehensive and informative descriptions and explanations of processes occurring in specific contexts. They can allow events to be traced chronologically and spatially and to reveal detailed causal explanations of specific events (e.g., how and why recent government policies of austerity have moved back and forth between particular countries of the global “North” and “South”). More than this, qualitative data can spark new and revised conceptual frameworks or yield unexpected revelations. For example, in his substantial study of medical liability insurance and health-care cost containment in the United States, Hay (1992) found hitherto unnoticed causal connections between malpractice insurance and the global asbestos crises confronting insurers and their offshore reinsurers.
Table 2  Five misunderstandings of case studies and correctives.

<table>
<thead>
<tr>
<th>Misunderstanding</th>
<th>Corrective</th>
</tr>
</thead>
<tbody>
<tr>
<td>General, theoretical (context-independent) knowledge is more valuable than concrete,   Predictive theories and universals cannot be found in the study of human affairs. Concrete, context-dependent</td>
<td></td>
</tr>
<tr>
<td>practical (context-dependent) knowledge.                                         knowledge is, therefore, more valuable than the vain search for predictive theories and universals.</td>
<td></td>
</tr>
<tr>
<td>It is not possible to generalize on the basis of an individual case; therefore, the case study cannot contribute to scientific development.</td>
<td>It is often possible to generalize on the basis of a single case, and the case study may be central to scientific development via generalization as a supplement or alternative to other methods. But formal generalization is overvalued as a source of scientific development, whereas “the force of example” is underestimated.</td>
</tr>
<tr>
<td>The case study is most useful for generating hypotheses; that is, in the first stage of a total research process, while other methods are more suitable for testing hypotheses and theory building.</td>
<td>The case study is useful for both generating and testing of hypotheses but is not limited to these research activities alone.</td>
</tr>
<tr>
<td>The case study contains a bias toward verification, that is, a tendency to confirm the researcher’s preconceived notions.</td>
<td>The case study contains no greater bias toward verification of the researcher’s preconceived notions than other methods of inquiry. On the contrary, experience indicates that the case study contains a greater bias toward falsification of preconceived notions than toward verification.</td>
</tr>
<tr>
<td>It is often difficult to summarize and develop general propositions and theories on the basis of specific case studies.</td>
<td>It is correct that summarizing case studies is often difficult, especially as concerns case process. It is less correct as regards case outcomes. The problems in summarizing case studies, however, are due more often to the properties of the reality studied than to the case study as a research method. Often it is not desirable to summarize and generalize case studies. Good studies should be read as narratives in their entirety.</td>
</tr>
</tbody>
</table>

Bazeley (2013, 5) makes clear the conspicuous revelatory and contextual power of qualitative data when discussing the significance of cases, rather than variables, as qualitative data. As clarification he points to the difference between the statements: “With their lack of education, Bill and Stephen struggled to find meaningful work;” and “Low education predicts poor employment options.” Whereas the latter statement simplifies, disembodies, and dislocates relationships between employment and education, the former simultaneously reveals actors and invites further examination of Bill and Stephen’s struggles.

Qualitative data can be used in a variety of ways to make a scholarly argument (Mason 2002). It may be employed evidentially, to help create a very carefully arranged case. It can be used interpretively or narratively, to create a convincing story or to suggest that a specific argument is reasonable. It might also be used to evocatively, illustratively, or aesthetically inform
or add color and substance (e.g., an excerpt from a focus group discussion that clarifies the reactions of a group of high-rise residents to local greenspace development). Clearly, data such as photos, quotations, or video are especially useful for this. It might also be used multivocally, so as to allow a range of perspectives to be presented or voices heard. And, of course, qualitative data can be used in several or indeed all of these ways in a single document such as a scholarly paper, report, or thesis.

Quantitative and qualitative data have, unarguably and problematically (Oakley 1998), been gendered, with the former being represented as masculine, hard, statistical, and objective while the latter has been rendered feminine, soft, nonscientific, and subjective. While this has led some feminist researchers, for example, to discount quantitative methods and their data in favor of those which admit subjective knowledge, the dichotomy and divisions it spawns are generally unhelpful and unproductive. Indeed, the power and versatility of qualitative data have been acknowledged across much of (human) geography. Despite the continuing – yet waning – dominance of positivism within the discipline, the rise and widespread influence of critical theory, poststructuralism, and the cultural turn in geography and other social sciences have all spurred use of qualitative data, which is used either exclusively or in conjunction with quantitative data. Because so much geographical research spans different scales and deals with complex human–physical relationships, multiple data sources have long been required. Indeed, Elwood (2010, 109) acknowledges that so-called “third geographers” (bridging qualitative and quantitative approaches) have been undertaking mixed method and multimethod research for decades, in acknowledgement perhaps that different data sources and data types can enable exploration of different causal relationships and illuminate different explanations of complex processes.

Just as there are significant shortcomings with quantitative data derived from positivist techniques, there are some significant issues associated with qualitative data. Notable amongst these is what has come to be termed the “crisis of representation.” This can be illuminated by a famously perplexing remark from British–Austrian philosopher Ludwig Wittgenstein (2009, 327) who wrote, “If a lion could talk, we could not understand him.” Although the meanings(s) of this inimitable quotation are contested, it can be argued that it questions the capacity (and authority) of a researcher to represent the lives of others. Researchers can never speak completely of others’ lived experiences (or even their own for that matter), nor can they portray social reality in any conclusive or ultimate ways. There cannot be any final, definitive, or conclusive account of lived experiences; instead, we must satisfy ourselves with interpretations, debates, and approximations. Full appreciation of the lives, sentiments, and experiences of others will always be beyond researchers’ grasp. Writing can only ever offer up partial accounts – even when that writing is about the researcher’s own life, as in the case of autoethnography – for every story or representation necessarily has its inclusions and exclusions. And representations of data will always be “political,” reflecting the positionality of its producer. These challenges are certainly not cause to cease the pursuit and production of data revealing experiences, understandings, and sentiments. Instead, the way forward lies in acknowledging and working with these certainties, aspiring to produce insightful accounts of sociospatial worlds that ring true to participants and shed light on hitherto obscured or concealed aspects of social lives.

To conclude, what is currently understood to be the shape and character of qualitative
data are changing and enlarging as the scope of geographical work is transformed. Challenging new work on areas such as affect, performativity, and nonrepresentational theory is emerging. As Elwood (2010, 110) points out, some of this work resists familiar and conventional acts of data collection, analysis, and communication, disrupting for example any clear separation between representation and practice. A good deal of effort will be required in coming years to determine how best to engage with these new developments and give meaning and function to the transforming expressions of qualitative data in geographical research.

SEE ALSO: Fieldwork in human geography; Mixed-method approaches; Qualitative data, acquisition; Qualitative GIS; Qualitative information: representation; Quantitative methodologies

References


Richards, Lynn, and Janice M. Morse. 2013. Qualitative Methods, 3rd edn. Los Angeles: SAGE.


Only about a decade old, the field of qualitative GIS (QGIS, or qualitative geographic information science) has rapidly grown and expanded into many disciplines outside geography. Today, it straddles the intersection of geographic information science (GISci), social science, humanities, computer science, and the geoweb. In the past two decades, methods became less hard wired to epistemologies, and the seemingly profound epistemological disconnect between GISci, on the one hand, and qualitative research, on the other, no longer appears clear and immutable. That qualitative GIS has successfully established itself as a major research direction was reflected, among other ways, in the 2009 publication of *Qualitative GIS: A Mixed Methods Approach* (Cope and Elwood 2009). Qualitative GIS has been vividly discussed in scholarly journals, books, and blogs across geography and geographic information science. Spatial turn in social sciences and humanities made qualitative GIS an actively sought methodology that combines representation of place, qualitative inquiry, and information technology. It has been used in research in sociology, history, and digital humanities, and interdisciplinary research on information technologies and society in the health sciences, environmental perception, and even environmental engineering.

As social sciences and humanities turn their attention to space, place, and mapping, the epistemological debates in geography about GIS can fruitfully inform these disciplines’ engagements with this technology. Unfortunately, these debates to a large degree remain confined to geography. Social scientists and humanities scholars, consequently, appear to be only moderately familiar with the geographic research on qualitative GIS, critical cartography, and critical GIS. They also tend to overlook rich theorizations of space and place developed by human geographers. When pursuing qualitative GIS, they often do not use the term or trace connections to these important literatures. Yet, the exciting spread of qualitative GIS creates opportunities for geographers to engage in mutually enriching conversations with other disciplines.

This entry, therefore, will not only examine the role of qualitative GIS in geography but touch upon research in other fields that the author, as a human geographer and critical GIS scholar, would identify as directly related to qualitative GIS. The entry begins by briefly reviewing the aspects of geospatial technologies that do align closely with qualitative information and qualitative research. It then discusses how geospatial technologies have been used in qualitative research and the ongoing integration of GIS with qualitative research methods. The recent but wide-ranging forays of GIS into the field of digital humanities, spatial social media, and neogeography are examined next. In conclusion, the entry reflects on the future of qualitative GIS.

### Epistemological interventions of feminist and qualitative GIS

Useful sketches of history of qualitative GIS can be found elsewhere (Cope and Elwood 2009),
QUALITATIVE GIS

so highlighted here will be those epistemological interventions that, in the author’s opinion, played a particularly important role in development of qualitative GIS in human geography and related disciplines.

Qualitative GIS stems directly from earlier interventions by feminist geographers and feminist GIS scholars. When the first feminist geographers argued that mainstream geography ignores experiences of women, they also advocated for qualitative research as necessary for addressing the nonquantifiable aspects of gendered experiences and informal social and economic relations. Once the limitations of quantitative methods were exposed, feminist geographers began to reassess their role and concluded that, if practiced reflectively within the feminist epistemology, quantitative analysis and statistics could be indispensable in the geographic research on gender, together with qualitative methods.

A decade later, feminist scholars have played an equally pivotal role in reclaiming GIS for feminist and critical geographic research. The fierce debates in the 1990s between GIS scientists and critical human geographers acknowledged the growing role of this technology in academic practice. But they also highlighted its ties to the interests of the empire, corporations, state surveillance, and militarism as well as to quantitative geography and positivism. Feminist geographers have specifically focused on the masculinist nature of GIS technology and the gender-blind research that it informed (Kwan 2002). They have not, however, rejected GIS altogether. Instead, feminist scholars have simultaneously led the efforts to reconstruct GIS as a mixed method that uses both quantitative and qualitative techniques. Neither the technology itself nor its applications are fixed, they argued; they evolve and depend upon who is using GIS and to what end (Kwan 2002; Cope and Elwood 2009).

To become a tool for feminist and critical geographies, however, GIS had to be reimagined as suitable for qualitative research. The prevailing assumption that GIS is only good for quantitative computation and modeling, therefore, had to be challenged at the epistemological level (Pavlovskaya 2009; Cope and Elwood 2009). Feminist GIS scholars pointed to the fact that we falsely associate computation with quantitative methods because computer technologies in general handle well qualitative information and tasks such as writing, graphic design, and visualization. Visualization of information in map form constitutes arguably the most important function of GIS. Maps, and digital maps in particular, have a powerful effect; they fascinate and convince the viewer regardless of whether the presented information is quantitative or not.

Moreover, the affective impact of maps provides them with ontological power to produce landscapes. Once visible, the mapped places and phenomena become real; they exist and require explanation. Being on the map, therefore, incorporates actors and phenomena into social imaginaries (this applies to both quantitative and qualitative data but social sciences have been suffering from a dearth of qualitative information because this information was not treated as scientific). Omission from the map, either intentionally or by ignorance, in contrast, leads to theoretical as well as socioeconomic and cultural marginalization (Pavlovskaya 2009). Since mainstream GIS mainly works with quantitative data sets, it fails to represent phenomena, processes, and experiences of paramount importance that are not quantified. Informal social practices, alternative economies, unpaid domestic work, barter exchanges, gift economies, emotions such as trust or fear, local and indigenous knowledge, class oppression, gender exploitation, racism, and other social relations remain invisible in our landscapes and conceptually marginalized.
Qualitative research methods offer insights into these processes. Therefore, integrating them with GIS begins to address their conceptual (and, by extension, political) marginalization.

Geospatial technologies and qualitative information

Assuming that GIS works only with quantitative computer-coded information, we have long ignored the fact that qualitative data such as text, images, and sounds are also coded for computation and, therefore, can be part of GIS analysis. Today, the production of digital spatial information is rapidly expanding (the so-called big data phenomenon) and so is its qualitative component. Some of this qualitative content is a result of conversion from nondigital sources. It includes historical and library archives, diaries and literary narratives, photographs, newspapers, and many other sources. The rest is born out of the digital age itself and incorporates webpages and blogs, geotagged social media such as tweets and Facebook posts, global positioning system-tracked movements, and digitally recorded consumer information. Numerous grassroots mapping projects also generate new digital and often qualitative data about their communities. GIS technology is increasingly capable of handling these diverse newly available qualitative data. In addition, GIS keeps converging with other digital media and this process transforms it into a multimedia visualization and analytical tool popular with digital humanities, social computing, geoinformatics, neogeography, volunteered geographic information (VGI), public participation GIS, augmented reality research, humanitarian aid, and many other projects.

In contrast to the 1990s and 2000s, desktop GIS packages no longer dominate the geospatial field. Various open source and free Internet-based mapping applications are widely available today. Many of them provide new methods and algorithms for handling large volumes of information while the producers of digital spatial information adopt formats compatible with these new analytical tools. Initially a source of information for people, the Internet is becoming a geoweb that act at once as an analytical tool and as a source of data for the new tools. GIS scholars and computer programmers focus on the algorithmic handling of these data (DeLyser and Sui 2013) but constructing meaning from these data in conjunction with social theory requires the insights and methods of qualitative GIS.

Doing qualitative research with geospatial technologies

Qualitative GIS researchers first attempted to integrate GIS into their work by mapping data from in-depth interviews. Initially, they analyzed interview data using grounded theory methods and related coding techniques and then mapped the results in a desktop GIS using the conventional cartographic symbols of different color, size, and thickness (Knigge and Cope 2006). It was vital for them to visualize the marginalized social phenomena that lacked quantitative representation (e.g., household economies or community gardens).

Local knowledge mapping projects also play exactly this role. They reveal understandings of the landscape that differ from the dominant representations created by states, experts, or corporations. Often called counter-mapping projects, they translate local spatial knowledge into claims to resources and land. In some cases, this is achieved by contrasting local ecological knowledge of resource users with professional expertise and satellite image classification. Other projects attempt to secure community access to resources...
by mapping informal land-use rights or how different groups of fishermen collectively use fishing grounds (Weiner et al. 1995; St Martin 2001).

The spreading use of GIS and other digital technologies by indigenous communities to represent indigenous knowledge and organize together gave birth to a new term, “indigital” (Palmer 2009). Digital indigenous mapping practices contribute to renegotiations of colonial legacies in many parts of the world, advancing and also posing challenges to postcolonial struggles. Here qualitative GIS can also make a lasting impact. In addition, public participation GIS (PPGIS) scholars worked to map assets of economically marginalized and minority communities in urban areas as a way to involve these communities in planning and economic decision-making. When visualized and shared as maps, their collective local knowledge has a potential to transform planning practices, together with the ways the communities perceive and value their neighborhoods (Elwood 2006; Knigge and Cope 2006). Qualitative GIS can inform such projects in terms of understanding exclusions created in the collective production of knowledge.

Feminist scholars pushed the “bounds of GIS” to map the experiences of women in a number of ways. Mei-Po Kwan’s pioneering work combined the computational power of GIS, feminist economic geography, and a new take on Hägerstrand’s time–space geography. She analyzed gender differences in access to urban opportunities by mapping daily movements of women based on information from their diaries. The resulting three-dimensional (3-D) visualizations showed that women’s daily life paths differed considerably from those of men, while differences between women by class and race also stood out (Kwan 2002). These visualizations revealed how hierarchies of gender, class, and race continue to shape women’s use of urban space.

In another project, Kwan (2008) turned to mapping emotions and feelings, another under-represented form of experience. She created a geography of fear experienced by a Muslim woman during her daily travels around the city before and after 9/11. Segments of her daily path reflect the strength of her emotion with different colors. This powerful visualization shows that the woman felt unsafe for weeks in her car and even at home after the tragic events.

This author’s work on postsocialist Moscow involved qualitative interviews with households with children in Moscow about the economic practices that helped them survive the economic devastation during the transition to capitalism in the 1990s. Each household used a combination of economic practices in which the share of informal, unpaid, and in-kind exchanges and production of goods and services has increased following privatization of urban services (Pavlovskaya 2009). The theory of transition ignored household economies because it focused on macroeconomic scale and structural transformation of the formal economy. Mapping multiple economies of households made it clear that blindness of the national policy to the scales of households and the form in which the transition was experienced by ordinary people made households bear the everyday brunt of the economic transformation.

Matthews et al. (2005) illustrate the effectiveness of using GIS in combination with ethnographic research that seeks to understand the challenges of the urban poor who struggle to maintain both family and employment responsibilities. Their large-scale ethnographic research project produced geo-ethnographies of five US cities that encouraged new policy approaches by changing the ways in which social scientists and social workers understand the experiences of poverty.

Joining ethnographic methods with GIS has proven fruitful in research on artistic
communities and creative industries. Artists are increasingly engaged in the production, and understanding the meaning, of urban space, but artistic creativity is hard to measure using formal indicators. In their research on distribution of artistic activity in the Australian city of Darwin, Brennan-Horley and Gibson (2009) interviewed artists and mapped their creative spaces (spaces in which they produce art) as well as their travel between those spaces. They found that most creative activity takes place outside the formal cultural centers (such as theaters or art galleries) and concentrates outside the formal cultural hubs. When designing policies that support creative production, local governments need to consider these geographies.

**Integrating GIS with qualitative research**

In addition to mapping the results of qualitative analysis, scholars attempt to integrate qualitative analysis with spatial visualization more directly. Knigge and Cope (2006) developed an approach called “grounded visualization” that involved close interaction between methods of grounded theory (e.g., coding) and GIS analysis. Their spatial database of community gardens in Buffalo, New York, included information derived by continued rounds of qualitative grounded theory analysis. Visualizing this information on the map, in turn, provided new ideas for such an analysis.

Further methodological innovations combined the rigor of qualitative and geospatial analysis at the level of software that allowed for conducting both types of analysis jointly. One such innovation is the so-called computer-aided qualitative GIS (CAQ-GIS) that made it possible to query and retrieve interview information and spatial locations related to a particular qualitative code or concept with greater ease (Jung and Elwood 2010). Another innovative integration involves adding select qualitative analysis functionality to a GIS. By programming qualitative analysis functions in ArcGIS software, Kwan and Ding (2008), for example, constructed the so-called geo-narratives that dynamically linked the interviewed respondents’ stories to the places about which they were told. Visualizing these connections demonstrated a high degree of spatialization of the individual narratives. Kwan and Ding’s (2008) research has also inspired others to visualize human mobilities. These developments are important because conventional GIS still offers only a limited functionality to represent movement.

These innovations nevertheless push the development of market-based and open source software to integrate qualitative research with space. Thus, qualitative analysis software now incorporates basic mapping capabilities while computer scientists within and outside the GIS field pay increasing attention to qualitative visualization. Qualitative GIS, however, can uniquely contribute to the task of explaining these visualizations using social theory.

**Humanities GIS, neogeography, and beyond**

The expansion of spatial digital qualitative information has occurred when both proprietary and open source tools for visualization and spatial analysis are becoming widely available. Some of them replicate nondigital forms of analysis while also making it faster and more expansive (for example, applying grounded theory methods to qualitative data using ATLAS.ti or a similar software package). Other digital tools, however, enable distinctly new kinds of analysis, especially for algorithmic processing of big data. They include, for example, visualization of large
qualitative data sets, knowledge mining, rhythm-analysis, dynamic and semantic analysis, and others (see DeLeyser and Sui 2013; Manovich 2016).

The emergence of digital humanities in the past two decades at the crossroads between humanistic inquiry and digital technologies, according to some authors, might represent a radical break with existing scholarly traditions because of the drastic impact of the technological and informational revolutions on knowledge production. The new field of geohumanities, or spatial humanities, focuses on the meaning of place constructed through digital literary geographies, histories, and memories, and local community participation. (The Association of the American Geographers has recently launched a new peer reviewed journal called GeoHumanities.) In contrast, social computing and geoinformatics are primarily concerned with direct visualization of the information from social media and other high-volume and dynamic data generated by contemporary publics (Manovich 2016). Finally, the ongoing decentralization and increased collective production of place-based geographic information known today as neogeography might represent another pivotal shift (Warf and Sui 2010).

The impact of geoinformatics, social computing, digital humanities, and neogeography on how we represent and understand place would be hard to grasp without the theoretical and methodological contributions of qualitative GIS and critical human geography. These remain vital to understanding the new forms of geographical knowledge because of their insights into the role of spatial (quantitative and qualitative) information in the analysis of place and space; the ontological power of mapping that transforms conceptually marginalized processes into objects of theory and practice; and their commitment to rigorous social theoretical analysis when explaining the emerging spatial patterns.

**The use of GIS in literary geography**

Understanding place using literature as a form of geographic data has been a long-standing tradition in humanistic and cultural geography. Invigorated by critical scholarship, literary geographers have brought into sharper focus the relations between social power and literary landscapes. While it has been common to analyze literary texts through narration, digital humanities and qualitative GIS began to map the spatial settings of literary narratives in order to reveal new aspects of their spatial organization and gain additional insights into their meaning.

It is important to note that the efforts to map literary landscapes originate in different theoretical perspectives. Literary geographers are likely to use advanced visualization capabilities of GIS together with social theoretical insights from humanities, geography, and critical GIS. Travis (2014), for example, combines critical literary theory with feminist and qualitative GIS in order to create sophisticated 3-D visualizations of Irish literary texts that incorporate the nonlinear and even nested literary spaces (e.g., a novel within a novel).

In contrast, humanities scholars who discovered mapping as part of their relatively new focus on space and place construct literary landscapes without using geographic theories of space, place, and time or critical cartography and GIS. They also bypass desktop GIS in favor of online mapping tools (e.g., Google or Open Streets Maps) that allow them to quickly add the needed placemarks to the base maps. The literary cartography project “Mapping St Petersburg,” for example, visualizes in this way the geography of Dostoevsky’s novel *Crime and Punishment* (Young and Levin 2013). These fascinating experimental efforts, however, could benefit from the insights of cultural geography and critical cartography and GIS.
Digital historical atlases

Digital historical atlases have become a hallmark of digital humanities projects. Highly interactive, they put online previously inaccessible information from historical archives. The website Digital Harlem (http://digitalharlem.org/), for example, aims to portray the everyday life in this famous neighborhood for the period between 1915 and 1930. The website includes information “drawn from legal records, newspapers and other archival and published sources” and allows users to search it by different events, persons, and other categories. The results are presented as maps overlaid on contemporary and historical maps of Harlem. The amount of information available through the website is astonishing; however, police records predominate because they are the most systematic source. This puts a particular slant on how everyday life in Harlem at that time is depicted. Critical cartography and feminist and qualitative GIS scholarship on the epistemological dangers of silences created by mapped data would be highly relevant here. This would stimulate more balanced data sourcing and the greater inclusion of, for example, memories, diaries, and newspaper descriptions that are harder to incorporate but would prevent the (unintentional) marginalization of the cultural heritage of this iconic neighborhood.

Neogeography and crowdsourcing geographic data

Neogeography, including VGI, public participation GIS, community mapping, and citizen science, provides new sources of highly decentralized, collectively constructed, and often qualitative geographic knowledge. Warf and Sui (2010, 201) characterize neogeography as part of the citizen science that uses the geoweb to produce and analyze geocoded data. Despite being developed only recently, neogeography, many authors agree, has profoundly changed how geographic information and knowledge are created. Neogeography projects flourished not least because the newly available geospatial tools are easy to learn and work well for basic geovisualization. This has made mapping considerably less dependent on expensive GIS software and technical expertise.

The range of neogeography projects varies from crowdsourcing data on car emissions to humanitarian emergency mapping (most famously in the aftermath of the Haiti earthquake) to deep mapping of cities (see next section), mapping noncapitalist economic practices such as the solidarity economy (www.solidarityeconomy.us/), collective urban resources (such as #MapJam by Sharing Cities Network in 60 cities worldwide http://mapjam2014-shareable.nationbuilder.com/), or activist mapping of election violations (www.kartanarusheniy.org) and wells drilled for fracking (Inman 2015), among many others.

Qualitative GIS has much to contribute to neogeography in terms of understanding the collective practices of geographic knowledge production. This research on community participatory mapping and public participation GIS commonly involves academics who mediate the interaction between communities, GIS technologies, and policymakers. Neogeography, in contrast, is often self-organized, and academics play a less prominent role. In this situation, control over spatial data and resulting knowledge is moving away from a relatively tight circle of GIS academic, government, or industry experts into the hands of nonacademic communities. The degree of decentralization is clear when comparing 1 million GIS users with the 400 million neogeographers worldwide (Warf and Sui 2010, 201). The latter form self-selected groups
with access to the Internet that generate spatial information that serves their interests. This stands in contrast to both the conventional GIS research that seeks to discover the objective truth and PPGIS that aims to generate a representative community view. Qualitative GIS scholars could offer methods for analysis of data produced by neogeography as well as insights into the politics of representation, partiality of knowledge, and political economy of neogeography.

**Deep mapping and digital urbanism**

Humanities scholars have been inspired by GIS because it allows for digital “deep mapping” (Bodenhamer, Corrigan, and Harris 2013). Deep mapping creates rich, textured, and multilayered descriptions of places using environmental, social, historical, political, economic, and cultural information from official sources such as natural history, census statistics, newspapers, and historical archives as well as less formal sources such as personal histories and memories, photographs, observation, interviews, and conversations. The goal is to articulate a collective sense of place. GIS provides a way to generate meaning of place by juxtaposing spatial layers with diverse information, and such digital deep maps can, compared to a narrative, represent place more directly. New digital media allow for creating multimedia deep maps that exist in cyberspace and also interact with the inhabitants of the place. Historical maps are layered over contemporary street networks, and census statistics are combined with memoirs of past residents while local communities articulate contemporary sense of place by posting photographs, comments, and other materials. Digital deep mapping projects could thus become collaborative modes of spatial knowledge production that develop at the intersection of artistic, scholarly, and participatory (neogeographic) representation of places.

One such “digital urbanism” project is edmontonpipelines.org at the University of Alberta. It depicts urban space in Edmonton, Canada, and seeks to build inclusive citizenship. The metaphor of pipelines supports the idea of channeling into a single website the past and present life of this oil city. Among other things, the website includes the indigenous people’s memories, although somewhat superficially, which raises questions similar to those that qualitative and PPGIS scholars have asked in their own research. In particular, why these and what other citizens of Edmonton have been underrepresented in this digital urbanism project? Qualitative GIS research has addressed problems of exclusion through mapping in many contexts, making it directly useful in digital urbanism projects.

**Spatial social media**

Social media (such as FaceBook, Twitter, and other platforms) constitute a new source of large volumes of dynamic qualitative data that social computing scholars seek to analyze algorithmically (Manovich 2016). The increasing number of posts are geotagged, which transforms them into geospatial data that can be semantically analyzed and mapped. Such maps are of great interest to the public and can potentially change how people think about the world (see, for example, www.floatingsheep.org). One popular method to analyze tweets is to generate content clouds in which the most frequently used words are plotted, for example, with larger fonts, and mapped across space. This method is attractive because it maps unmediated textual information. However, it pulls words out of context and shows the frequency of the use of a word instead of the meaning of the tweet. Thus, mapping the frequency of tweets that mention a political
candidate during the election campaign, for example, would not indicate whether people tweet in support of or against this candidate (Jung 2015). To address this problem, Jung (2015) uses a qualitative GIS method that combines grounded theory coding techniques with the semantic analysis of the spatial media content. Codes are words or expressions that show the meaning of blocks of text or tweets. Mapping the resulting “code clouds” would not conflate statements in support of or against politicians, preventing claims to popularity based simply upon how often a name is mentioned in tweets. Mapping the codes reflects spatial differentiation in political support as opposed to a mere presence in mass media.

While the processing of spatial social media information is high on the agenda in big data analytics, very often the data are parsed and visualized mechanistically without using rigorous methods for construction of meaning. Qualitative GIS methodologies, therefore, can usefully inform and work in conjunction with textual analysis of the geoweb. In addition, it is sometimes forgotten that spatial social media suffer from the old and new forms of digital divide. Critical and qualitative GIS scholarship, however, points to the fact that cyberspace is highly fragmented along the lines of class, race, and gender as well as Global North and Global South. This research adds an important dimension to analysis of social media.

**Conclusion**

The formerly dominant view of GIS as a primarily quantitative tool was owed to its origins in military applications, quantitative geographic tradition, and the computer industry. Yet, following critical GIS debates, qualitative GIS has emerged as a field that pioneered ways to map new types of data derived from qualitative interviews, historical archives, literary texts, and, more recently, social media, neogeography, and artistic visions of place. It also advanced integration of qualitative research methods with geospatial analysis in order to account for nonquantifiable, uncounted, and conceptually marginalized but important experiences and socioeconomic practices. Qualitative GIS, therefore, constructs new imaginaries of place and space that can contribute to inclusive citizenship. Advancing the fusion of critical human geography with GIS mapping would help to realize better this potential of qualitative GIS.

In less than a decade, qualitative GIS became widely used not only in geography but across social sciences and humanities. Involving GIS in mixed methods and qualitative research would not have been possible without epistemological interventions by feminist and qualitative GIS scholars. These scholars asserted the importance of qualitative data, argued for its compatibility with GIS, countered conceptual marginalization of unmapped subjects and processes, and theorized the ontological role of mapping and the necessity of participatory generation of geographic knowledge. These interventions both legitimized qualitative GIS as a field and reconfigured GIS technology in the light of feminist and critical epistemologies.

The recent but intense diversification of Internet-based and open source geospatial tools has coincided with the rapid growth of digital qualitative information. Historic archives and library collections are being converted into digital form on a massive scale, making it possible to digitally analyze them for the first time. Tweets and consumer data are increasingly geotagged. In addition, public science, neogeography, crowdsourcing, VGI, and similar projects provide self-selected groups with the means to generate new forms of geographic information. New tools for processing and visualizing the volumes of often qualitative big data appear overnight. New fields are flourishing, including...
QUALITATIVE GIS

social computing, geoinformatics, geo-/spatial humanities, digital urbanism, and indigital networks, among others, that utilize these new data and tools to present and use information in spatial form.

While the context in which qualitative information and mapping intersect today has changed, the debates within qualitative GIS about the nature of geographical knowledge and representations of space remain highly relevant and current to these new fields. The latter, however, are not always aware of qualitative GIS interventions and often use the qualitative GIS approach without naming it as such. Yet, learning about contributions of qualitative GIS would prepare scholars in social sciences and humanities for the challenges of big data and the digital age. Thus, they would be equipped to deal with space and geographic knowledge in a profound way and see their work in the context of ongoing urban, economic, cultural, and environmental struggles as well as the politics of geographic knowledge and barriers to participation in its production.

As these lines of research and social practice continue to evolve and expand, the epistemological significance of qualitative GIS stands to grow. It is well positioned to equip scholars practicing qualitative GIS with theories, critiques, and ideas that enable progressive struggles while also allowing them to generate new insights by learning from the new digital spatial contexts.

SEE ALSO: Critical GIScience; Feminist methodologies; Geographic information science; Public-participation GIS; Qualitative data

References


Qualitative information: representation

Stephen Burgess
Cardiff University, UK

What is qualitative information?

In order to understand qualitative information representation, it is necessary to first understand what is meant by information. While it has a common lay meaning, information is a technical term not commonly used in the qualitative geographical methods literature. However, there are some areas of geography where the term information is regularly used, and that is in the related fields of GIS (geographic information systems) and geovisualization. GIS and geovisualization have long borrowed from visualization techniques and worked alongside visualization practitioners in the sciences and computer sciences (Dykes, MacEachren, and Kraak 2005, 4–5). This relationship has been reciprocal with many of those working in visualization also interested in representing variation in parameters that have been central to much geographical analysis such as location and time. Therefore, to understand common use of the term information in geography, it is necessary to turn to the visualization literature.

Visualization is a set of practices which emerged at the intersection of the natural sciences, computer science, and graphic design and is concerned with producing and conveying meaning by visually representing complex data. Although it initially struggled to gain currency within the scientific community, several shifts from the 1960s onwards, including the increasing efficiency and proliferation of computing technology and its visual interfaces for scientific analysis, saw visualization become more embedded within scientific practice (see Unwin 2008). As data sets continue to become bigger and more complex, the uses of visualization also continue to expand.

In the overlapping literatures and practices of scientific visualization and information visualization, there is an important distinction made between data and information, where information is created by a process or processes which give meaning to data. This specific contrast between data and information made in the information visualization literature often draws on Shroff’s Continuum of Understanding (1999) which differentiates between data, information, knowledge, and wisdom. Shroff suggests that data are the product of research and the raw material of analysis which, by and of itself, lacks meaning and is not suitable for communication to an audience. Information is produced by “organizing [data] into a meaningful form, presenting them in appropriate ways, and communicating the context around them” (Shroff 1999, 272–273). It is only once researchers have begun to produce meaning, or information, that it is appropriate to communicate understandings to audiences through representational practices. Knowledge is created when information is represented to and interpreted by people (individuals, groups, institutions, etc.) based on individual or shared experiences (i.e., that which contributes to creating understandings, preferences, beliefs, etc.). Shroff argues that the design of visual representational practices should focus on...
knowledge production through the representation of information. Beyond knowledge is what Shedroff calls wisdom, where individuals move the information forward and create their own understandings.

While it should be noted that there are alternative models to that of Shedroff, his model is useful in illustrating that the representation of qualitative information is concerned with representing that which constitutes and/or emerges from the various processes which seek to bring meaning to data (i.e., the analysis and communication stages of qualitative research, and not the data acquisition stage). Because of the relationships between visualization, geovisualization, and GIS, the term information representation is more common in these areas of geographical practice. In both geovisualization and GIS there has recently been some interesting development in the representation of qualitative information as these fields have begun to consider the challenges this presents.

The “crisis of representation” and beyond

While information representation may not be a term that occurs commonly in qualitative geography, issues of representation are themselves important and need consideration in a wider context. Representation, or re-presentation, takes many forms; for example, representation is an activity that is claimed by art, literature, cinema, photography, theatrical performance, and music. To these can be added representational activity familiar to the academic, such as the writing up of ethnography or the graphing of a regression model. These are practices which all concern depiction, including the portrayal of thoughts, ideas, and observations. The proliferation of representational practice across society and culture is reflected in debates around issues of representation which occur across many disciplines including art, literature, the social sciences, humanities, and geography, and within popular culture beyond the academy. These debates, in many instances, continue to cross over with one another and representation is consequently a complex and dynamic notion.

Within qualitative geographical methods, various practices and methods are adopted to capture qualitative data, and produce and represent meaning at different stages of research including data acquisition, analysis, and communication of the findings of analysis. Methods and techniques for acquiring data have been, and still are, dominated by a suite of approaches including interviews, observation, ethnography, and analysis of documents such as diaries and policies. The data acquired by these methods are most often represented as text for the process of meaning making: for example, interview recordings are transcribed and notes written up, producing texts which are the representations of conversations that are used to create meaning. The organization of textual data to make meaning often involves the labeling, or coding, of these textual representations to identify patterns and relationships using a variety of analytical approaches including discourse analysis, conversation analysis, thematic analysis, and semiotic analysis. This may be done using one of several specialized computer software packages, sometimes known collectively as computer-assisted qualitative data analysis software or CAQDAS. The meaning that is made through analysis (which the visualization literature would refer to as information) is then also often represented as written text for communication to different audiences. Different textual formats may be used for different audiences such as academic papers for an academic audience, reports for a policy audience, and newspaper articles for a lay audience. Notably, even where non-textual data
are collected, such as photographs, video, audio, and sketch maps, they will often eventually be represented textually for analysis and/or communication of information. The dominance of textual representation of qualitative information within geography is certain, in terms of both making and communicating information.

Importantly, qualitative approaches have struggled with issues surrounding representation. In 1986, anthropologists Marcus and Fischer coined the term “crisis of representation” to describe a change in thinking that was occurring across the social sciences, including within some areas of geography. This broad change in thinking questioned whether it was possible to ever adequately describe, or represent, social reality. This aligned with epistemological positions (e.g., what might now be called interpretivism) embedded within ontological positions (e.g., what might now be called social constructionism) which conceive of a social world that is continually made, interpreted, and remade by social actors and which can only be understood through the actions and interpretations of these social actors, including academic researchers.

The universal claims and general “laws” made by positivist methods had begun to be rejected by some, who saw knowledge as specific to particular contexts such as different individuals, societies, cultures, times, and places. Interest in the complexity and diversity of understanding eventually turned attention to consideration of the positionality of the researcher in the research process and of the power relations embedded within research: researchers were no longer seen as being able to impartially represent an external reality which they observed but instead were understood as contributing to the creation of knowledge both through their co-participation in the research process and through their analysis, interpretation, and communication practices (i.e., their information representation). For example, postcolonial writing has addressed the ongoing interrelationships between issues of positionality, power relations, and representation between colonized and colonizing groups, including consideration of research practices. Often located within the emergence of postmodernism, changes in thinking also began to challenge the way Western academics thought about representational practice itself: rather than being seen as reflecting an external reality, or “truth,” that was present in the social world, representations were understood as interpretations that themselves were a part of the social world and therefore helped create that world. Often drawing on linguistics, these ideas have developed particularly within poststructural discourses. The presence of varying approaches reinforces the complexity of the term representation and how it has been operationalized by human geographers.

In geography, one response to this thinking has been greater consideration by geographers of how they undertake a key area of academic work: that of writing. The act of writing is inescapable as an academic and is geography’s dominant mode of representation for analysis, meaning making, and communication (see DeLyser 2010 for a useful overview). This entry is itself a textual representation of a collection of ideas and is part of an ongoing process of meaning making, including the author’s engagement with previous ideas and various readers’ engagement with this text. Some academics have argued for the need to write clearly and simply in order that the interpretations and understandings of research participants are not obscured by the researcher. At the same time, particularly emerging within feminist geographies, there has been an acknowledgment of the role of the researcher in making meaning by influencing the production of knowledge in many ways, including their research agendas, how they frame
research questions, the methods they implement, their presence in the research setting, their analysis of data to produce information, and the communication of the information produced. Consequently, attention has been paid to the need to write reflexively, identifying how the self might impact on the production and representation of knowledge and meaning. There has also been an increase in representing multiple viewpoints and interpretations in the production and communication of meaning through writing. Examples include the adoption of multiple authorship, collaborative writing with those who take part in research as well as with other academics, and also the exploration of multiple perspectives around an issue. Geographers are not alone in these challenges and responses, which are reflected across the social sciences and the arts, including within literary criticism and practice.

Another response across geography and the social sciences has been to try and create research methods which attempt to break down the power relations embedded within research. Participatory research, for example, has sought to try and redress the imbalances in research toward voices other than the researcher by focusing on the idea of collaboration. Trying to move away from a top-down model of research, researchers work alongside stakeholder groups addressing topics and research questions which those groups have identified as important. In facilitating the influence of local knowledges and perceptions on the research process and the production of knowledge, multiple voices, including those from beyond academia, become represented within the knowledge that is produced and communicated. Where this research is designed in order to bring about appropriate change through this collaborative effort, it is known as participatory action research. However, it should be observed that the implementation of these processes is variable, and there are questions as to exactly how democratic these methods can be and to what extent they represent other voices in order to address power imbalances within the production of knowledge.

The crisis of representation in geography has also led to the development of thinking around nonrepresentational theory, as originally proposed by Nigel Thrift (1996). A diverse body of work has developed from Thrift’s ideas which have focused on what can be thought of as the “doing” of culture and social life. Meaning, society and culture (indeed all of life) are continually made and remade through the “things” that happen every day. The word “things” is deliberately used here, being a vague word, as the happenings that create the world are myriad, including practices, processes, events, moments, movements, encounters, and interactions, including the mundane and the exceptional, including the human and the non-human (leading to more than human, or posthuman geographies), including the articulated and the unarticulated, including the cognitive and, of particular interest, the noncognitive. In this shift, representation was replaced by performance and a body of work around performative geographies emerged. Performance as both metaphor and substantive research area has developed across the social sciences. In focusing on performance, geographies have become embodied, focusing on how what individuals do, experience, and understand help create meaning, society, and culture. This has allowed a recognition of the importance of individuals and included a focus on affect defined as “properties, competencies, modalities, energies, attunements, arrangements and intensities of differing texture, temporality, velocity and spatiality, that act on bodies, are produced through bodies and transmitted by bodies” (Lorimer 2008, 552). This has led to the recognition of the importance of multisensual understandings of the world as well as the role
of emotion and the development of emotional geographies. It is important to note that within nonrepresentational geographies there is space for representational practices which have affect and can be seen as part of the doing of social and cultural life.

Nonrepresentational thinking has led to an interest in the development of methods which address how complex, multifaceted, and embodied meaning relates to the happenings of life. This includes being sensitive to affect and to multisensorial and emotional experiences of these happenings. Development of research practice includes attention to data acquisition and the production and communication of meaning. Some argue that existing, and dominant, research methods cannot adequately produce these understandings therefore new qualitative methods are necessary. Others, notably Davies and Dwyer (2007), suggest that, actually, it is hard to envisage a wholly new set of qualitative practices and that most change has been to the ways in which existing methods are conceived, conducted, and also to the claims made for the understandings and representations they produce. This latter position is the one most evidenced by the methods that can be found adopted in the empirical nonrepresentational literature. Davies and Dwyer argue that this reconceptualization embraces that there is ambiguity in researchers’ understandings of the world and that it is not possible to claim a precise representation of a world that is so textured (e.g., contingent, subjective, embodied, and often unarticulated) that there is no capacity to fully understand it. Empirical research is still worthwhile, but, intertwining with earlier and broader developments outlined above, it necessitates a reflexive approach. This includes awareness that the action of undertaking (or performing) research, from fieldwork to data representation for analysis and communication, itself contributes to creating the social and geographical contexts from which meaning is made. Reflective practices need to include a reconceptualization of the knowledges that are produced from the research process: that there are various alternative ways of knowing than have been engaged with previously (e.g., embodied, multisensory, emotional) and that knowledges are contingent. Examples of research practices include: the adoption of narrative analysis to study how meaning is created through language rather than just what people say; haptic geographies which are attuned to multisensory knowledges such as sonic geographies; and an awareness of embodied practices such as those found in the mobilities literature. Conveying these practices and the meaning that is made through them still necessarily engages with representational practice, most often in the form of written text, but this engagement aims to be reflexive.

Geovisualization and mapping: visual representation of qualitative information

Despite the broader issues of representation within qualitative geography, the notion of information representation is more often encountered within GIS and geovisualization literature than in qualitative geographical research. Geovisualization is concerned with the visual representation of geographical data to facilitate the exploration and analysis of data in order to produce and communicate geographical knowledge or information. Geovisualization is a key element of GIS in which geographic databases (which can also be analyzed statistically) are visually represented to facilitate the exploration of data and the creation of meaning or information. Primarily, the geographical emphasis within these has been on mapping and modeling spatial relationships. For this reason, fundamental to the geographic databases underlying GIS is that
features (e.g., a road) and their attribute data (e.g., speed limit of road, road name) are georeferenced or linked to a location using a co-ordinate system. For the most part, and reinforced by the adoption of practices and literatures shared with visualization as it has emerged in the natural sciences and computer sciences, information representation within geovisualization has been identified as a quantitative practice embedded in positivism. The juxtaposition of information representation with the term qualitative is suggestive of mixed-methods projects that have emerged around GIS and geovisualization since the 1990s.

The developing importance of information visualization within academia has been reflected in popular culture, with a growing interest in the use of visual representations to deliberately convey complex information. Recent years, for example, have witnessed the increasing use of infographics within the news media. A contraction of information and graphic, infographic has become a popularized English-language term for information visualizations with infographics clearly part of the visualization genealogy, drawing on practices familiar to visualization and geovisualization. A more interactive web environment, sometimes referred to as Web 2.0, has helped bring the tools of visualization to a wider audience as an increasing number of websites appear which help create various information visualizations. Tag clouds (or word clouds) are a common text visualization tool, where the most commonly used words in a piece of text are visually represented in juxtaposition to one another, with the frequency of each word being indicated by text size or font color, for example. Figure 1 shows a word cloud for this encyclopedia entry as produced using www.worldclouds.com, a free web-based application.

Visual representation for meaning making is not itself new and has arguably always been important within geography. Mapping is one of the longest-standing examples of visual representation in geography, its own origins often being tied to the origins of geography itself, with those often highlighted in Western histories of the discipline as being early geographers (e.g., Ptolemy, around the first century CE) being involved in map making. Cartography is an important example here as geovisualization is located at the intersection of cartography and scientific visualization (Unwin 2008, xii). Changes in thinking across geography have been paralleled in cartography and also in geovisualization and GIS. Postmodernist approaches and the crisis of representation helped shift the understanding of maps; no longer were they viewed as objective and accurate reflections of an external reality, but as authored interpretations and representations which are also themselves a part of the social world and which help to create the social world. For example, postcolonial discourses helped maps to be interpreted as tools of power, where the drawing and labeling of maps by colonizing powers both reflected and sustained colonial rule. These perceptions of mapping which politicize maps and draw attention to the power saturated

Figure 1 Word cloud showing frequency of words in the current entry. Source: created using www.worldclouds.com.
within cartography emerged during the 1990s and are often referred to as critical cartography. Over the first decade of the twenty-first century, further changes have happened to mapping as visual representative practice. It is argued by some that mapping has become more democratized. As the World Wide Web has become more interactive, then interactive mapping tools have also become increasingly available on the web. Anyone with an internet connection can now be a cartographer. Open-StreetMap is an example of a global mapping project where the mapping is undertaken by a community of users using various techniques. Famously, in the wake of the large earthquake which devastated areas of Haiti in 2010, there was a lack of detailed mapping of the area. Working mostly remotely, the OpenStreetMap community addressed this by creating detailed local mapping which became the main map used by organizations working on the ground in the wake of the disaster. Such collaboratively created maps bring different perspectives to bear on the cartographic process, the creation of maps as visual representations, and the geographical information they produce. Nonrepresentational theory has also influenced cartography with a rise in practice that is interested in mapping embodied, multisensory, and emotional experience. Often sitting on the juncture between art and academia, and sometimes referred to as counter cartography, these practices are further questioning how, why, and what is mapped. Examples of this include Christian Nold’s biomapping/emotion mapping project (Nold 2014), which seeks to map subjective responses to what Nold calls the “external world,” and a book of geographies of the imagination collated by Katharine Harmon (2004). Through these changes, mapping has seen a shift from being heavily embedded within positivist discourses to being seen (i) as a subjective geographical practice of visual representation which is highly influential in the interpretation and creation of meaning within the social world and (ii) as a practice that has engaged with the challenges of representing information on maps that might be called qualitative.

With geovisualization sitting on the intersection of cartography and scientific visualization, the debates in geovisualization and GIS have followed a similar path to those in cartography. In the early 1990s, a series of debates developed around the limitations and usefulness of GIS and geovisualization methods. These debates can be located within broader debates (some of which are outlined above) that were occurring across geography at the time as new approaches and schools of thought challenged the positivism of the quantitative revolution and the focus on spatial analysis that had come to prominence within geography throughout the 1960s and 1970s while those involved in spatial analysis sought to defend it. These debates within GIS and geovisualization are sometimes characterized by those involved either as resulting in an ongoing standoff between those on different sides of the debate or as resulting in collaborative work. Arguably, both of these positions still exist, but it is the collaborative projects that are pertinent to qualitative information representation. Schuurman (2000) argues that throughout the 1990s these debates moved from hostile to conciliatory with the endpoint being that “[c]ritics were drawn into the upper echelons of GIS research, and the present era of increased co-operation between critics and the GIS community was initiated” (2000, 570). One outcome of this has been the emergence of a self-reflective set of approaches that go under the name critical GIS. This has been characterized as a small, but important subset of GIS practice which seeks to transcend ontological and epistemological divisions.
One outcome of critical GIS has been the emergence, particularly in North America, of participatory GIS (PGIS) which has applied participatory methods, as discussed above, to GIS and geovisualization as representational practices. Another response to critical GIS has been termed qualitative GIS and has emerged in the literature since the mid-2000s, a threshold moment being a special edition of *Environment and Planning A* which was dedicated to qualitative GIS. Qualitative GIS is concerned with mapping the “nonmeasurable properties of place, human experience, social hierarchies, power relations, and theoretical relationships that are of concern to critical geography” (Pavlovskaya 2006, 2015). As well as conducting spatial analysis including these properties, they are also represented, through geovisualization, for the purposes of data exploration and analysis in order to make meaning from the data and produce information. This information, which includes qualitative information, is also represented through geovisualization, often in the form of mapping, for the purposes of information communication. In this way, qualitative GIS has been argued to engage with qualitative information representation with many of its practitioners describing it as a mixed-methods project. A good outline of the emergence of the qualitative GIS project is provided in the collection edited by Cope and Elwood in 2009. Since around 2010, the term qualitative GIS has appeared to decline in use. This decrease in use may be linked to an emerging argument that, actually, qualitative GIS may not be very dissimilar to the GIS and geovisualization practices that have preceded it. In digitizing data to make them suitable for a GIS databases, GIS researchers have often drawn on data that may be termed qualitative. However, an active geographical community remains engaged with this work and this is reflected in very similar practices, although sometimes under different nomenclature, in other disciplines such as the digital humanities and archaeology. Finally, it should also be noted that, given the nonrepresentational turn in geography, consideration must be given to the ways in which these practices emphasize the visual over alternative ways of knowing.

### Representing for different audiences

In making and communicating qualitative meaning, researchers often find themselves engaged with representational practice of one kind or another. Across all these practices (e.g., geovisualization, written text, audio reports, films) the format and character of the representation might be influenced by the intended audience. For the most part, academics have communicated and still do communicate their understandings to other academics via journals, seminars, and conference papers. Often, these follow certain academic conventions depending on the form of representation adopted. However, geography, like other disciplines, also seeks to engage other audiences requiring different conventions and practices. Since the mid-1970s, and influenced by the publication of books such as David Harvey’s *Social Justice and the City* (1973), there has been a self-examination within geography as to how it is relevant to and, indeed, how it can contribute to wider society. A debate around public geography continues to be concerned with these questions. These sit alongside a common observation of the absence of geographical discourse from public debate, reflected by a lack of high profile public geographers in popular culture, unlike academics from other disciplines. Understanding how knowledge and those who make geographical meaning from inside the academy can engage with knowledge and those who make (sometimes geographical) meaning in
other contexts (e.g., policymakers, news editors, and publics) has to pay attention to representational practices. The conventions of geographical academic writing are not familiar to public audiences, for example, and alternative writing styles need to be adopted in order to make geographical information clear and available to engage with. As another example, policymakers tend to have their own conventions for writing (e.g., concise points with the inclusion of executive summaries) and engagement with these will further help the chances of geographers to engage policy and policymakers with the broad spectrum of geographical information.

SEE ALSO: Critical GIScience; Geographic information science; Geographic information system; Nonrepresentational theory; Postmodernity; Poststructuralism/poststructural geographies; Qualitative GIS; Reflexivity; Representation and presentation; Visualization

References


Qualitative spatial and temporal representation and reasoning

Alexander Klippel
Jan Oliver Wallgrün
The Pennsylvania State University, USA

Qualitative spatial and temporal representation and reasoning (QSTR) is the research area concerned with theoretical and experimental approaches to represent spatial, temporal, and spatiotemporal information, and with performing symbolic reasoning through qualitative abstractions of spatial and temporal information. Representation and reasoning apply to both spatial and temporal properties as well as relations between entities. The main distinguishing characteristic is that spatial and temporal information is abstracted into a finite and typically small number of categories, also referred to as equivalence classes. The field has roots going back to work on naive physics and commonsense reasoning in artificial intelligence (AI), on naive geography, and on topological data modeling. This entry provides an overview of the field covering the main motivations, ideas, and concepts as well as discussing current trends and developments.

There are many different approaches that address how data and information about space and time are represented in formal systems such as computers, and how humans, as the users of spatial and temporal data, engage in making sense of this data. Research on the commonsense understanding of spatial and temporal information has spurred the development of symbolic approaches, particularly qualitative formalisms, in which only a small number of distinctions is made with regard to spatial and temporal properties or relations. Qualitative formalisms are ubiquitous in areas where it is important to design interfaces between human and AI such as natural language understanding, geographic information retrieval, or human–robot interaction.

One of the first things that can be observed about commonsense representation and reasoning about space and time is that humans neither think nor represent the different types of spatial and temporal knowledge with high precision. That is, humans do not demonstrate a natural and intuitive capacity for dealing with infinitely precise information about spatial and temporal relations. In most situations, qualitative information about space and time – in the sense of a fairly small number of equivalence classes (categories) – seems to be sufficient for natural cognitive agents. The way these equivalence classes capture continuous information about space and time in a qualitative way has been referred to as qualitative metrics (Montello and Frank 1996). For example, various studies show that although humans may perceive angular information quite precisely, they conceptualize, think about, and remember it with less precision. To illustrate, imagine the directional relation between two point-like entities in geographic space. Whether humans use an absolute reference frame (cardinal direction) or a relative one to describe this relation, they always will use only a small number of equivalence classes to organize and describe this spatial information. The absolute reference frame may be reduced to north, south, east, or west, and the
relative directions may become simply left, right, in-front-of, and behind.

Inspired by this idea of qualitative metrics, AI and the spatial sciences have developed numerous approaches to formalize fundamental aspects of space and time that are considered equally important to cognitive and artificial systems. Examples are (mero-) topology, direction, and distance. The notion of a fundamental concept that reaches across disciplines is, however, generally not well defined, but it can be approached with the concept of invariance. Researchers in many fields from the cognitive to the spatial sciences have addressed the topic of invariance in the context of cognitive information processing, and used invariance to formally distinguish fundamental concepts of space and time. Worboys and Duckham (2004), for example, employed the concept of invariance as a framework for their characterization of fundamental spatial concepts. First proposed by Felix Klein for the Erlangen program (Klein 1872), geometries can be distinguished based on invariant properties under certain transformations. This approach allows for the differentiation of Euclidean geometry from set-based geometry and from topology, which can be seen as “rubber-sheet geometry” concerned with properties that remain invariant under topological transformations such as bending and stretching. Paralleling these formal approaches are perceptual and cognitive invariants, which are also found to be associated with conceptual primitives, and which have long been of interest to the cognitive science community as well as behavioral geography. The human mind, in adapting to its environment, identifies invariant characteristics of information that form the basis for cognitive processes. This idea is prominently featured in the classic work by Gibson (1979), who refers to temporally constant characteristics of environments as structural invariants. Many recent approaches feature the concept of invariance to explain, for example, perception, categorization, and cognition of spatial, temporal, and spatiotemporal information. The importance of identifying invariants of environmental information is prominently noted by Galton (2000), who speaks of our ability to intersubjectively identify invariants of space and time, which then allows a shared understanding of our physical (and social) environments to be constructed. Without the agreement that certain characteristics of spatial environments ground our meaningful understanding of spatial environments, the concept of a shared reality and our ability to communicate about this reality would not be possible. This is true for human–human communication as well as the communication of humans and artificial systems.

Many researchers have pointed to topology as the cognitively most important qualitative formal theory that also allows for a more precise interpretation of invariance. Although there are other theories that identify potentially relevant invariants for humans’ fundamental understanding of space and time, topology is unquestionably important for understanding cognition. In this brief overview a mereotopological approach is used, namely the 9-intersection model (Egenhofer and Franzosa 1991), arguably the most prominent example of a spatial calculus, to illustrate the ideas of QSTR. Other (more extensive) overviews on this area of research are listed in the “Further reading” section.

**Qualitative calculi: the example of the 9-intersection model**

Prominent in QSTR is the notion of a qualitative calculus. A qualitative calculus defines a finite set of basic relations for a particular spatial (or
temporal) aspect, such as topology (e.g., contains, disjoint, overlaps) or direction (e.g., north, south, east, west). The set of basic relations is typically defined such that for any pair of objects exactly one relation holds, meaning the basic relations are jointly exhaustive and pairwise disjoint (JEPD). Here, as an example, the 9-intersection model (9-IM) for topological relations is described (Egenhofer and Franzosa 1991), discriminating eight basic relations for two spatially extended entities: disjoint, meets, overlaps, contains, inside, covers, coveredBy, and equal (Figure 1). The 9-IM approach is largely equivalent to the region connection calculus RCC-8 by Randell, Cui, and Cohn (1992). Reasoning in a qualitative constraint calculus is performed on the power set of the basic relations. Unions of basic relations are typically written as sets. For example, the relation $A \{\text{disjoint, meets} \} B$ describes that region $A$ is either disjoint from or meets region $B$ (meets is defined as only the boundaries of the two entities overlapping). The universal relation (U) consisting of all base relations denotes that nothing is known about the relation of two objects.

Given a qualitative spatial calculus, the spatial arrangement of a set of objects (Figure 2a) can be described in terms of relations from the calculus by specifying the basic relations holding between each pair of objects. Such a qualitative description is often visualized as a directed graph called a qualitative constraint network (QCN).

**Figure 1** Eight basic relations of the 9-IM arranged in a conceptual neighborhood graph indicating how relations can change over time. Source: author.
The nodes represent the involved objects and the edges are labeled with the relations holding between the objects (Figure 2b). In such a case, the constraining relations in the network are labeled with a single base relation from the calculus selected to represent the scene formally (here 9-IM). Such a qualitative representation, that is, a scene for which all relations are known, is referred to as a scenario.

However, in QSTR qualitative descriptions that are incomplete or imprecise are often dealt with. For instance, the information extracted from observations may provide information about which basic relations hold between certain objects but not between others. Or it may only be known that the relation must be one (out of several possible) from the set of basic relations corresponding to a disjunction as introduced above (e.g., A {disjoint, meets} B). Figure 2c shows a constraint network that corresponds to such a situation: the edge between the spatially extended entities D and C is labeled with the disjunction of the four basic relations {disjoint, meets, overlaps, contains} because it is assumed that nothing more precise is known about the relation between these two objects. Similarly, entities A and C do not have a connecting edge, which should be interpreted such that the constraining relation is the universal relation U, meaning nothing is known about the entities’ spatial relationship.

Given a QCN that is not a scenario, that is, not all qualitative relations between entities are known precisely, one may be interested in inferring as much information as possible from the given relations. The composition operation, an important relation-algebraic operation that needs to be defined for a qualitative calculus (typically in form of a table), specifies the relation between entities X and Z as implied by the given relations between X and Y and between Y and Z. For instance, in the example shown in Figure 2c, given the relations {A disjoint B} and {B contains C}, it can be inferred that the relation between A and C must be disjoint. One challenge is that not all compositions yield a single basic relation that can hold between X and Z as the result. More often, the result is a disjunction of several basic relations or even of all basic relations, so the universal relation U. In the example here (Figure 2c), the composition of relations {D overlaps B} and {B
contains C\} implies that D can be in any of the following relations with C: disjoint, meets, overlaps, contains, or covers. This excludes only the equal relation from the disjunction in Figure 2c.

As this simple example shows, applying the composition operation to a triple of nodes in a QCN may allow the inference of new information not part of the originally observed information. The resulting network is a refinement of the original network in which some basic relations have been removed from the disjunctions. The result of applying the composition operation to infer as much as possible from the given information in the QCN from Figure 2c is shown in Figure 2d as a refined network.

Another important problem, the one most closely studied in theoretical QSTR research, is the consistency checking problem, also termed the satisfiability problem. A QCN is consistent/satisfiable if it has at least one solution. A solution is an assignment of concrete objects from the involved spatial domain to each object variable (node) in the QCN, such that all the relations in the QCN are satisfied. One method to check for consistency is the algebraic-closure algorithm based on the previously introduced composition operation. For most calculi, the algebraic-closure algorithm has to be integrated into a backtracking search procedure that considers different refinements of the original QCN. The algebraic-closure operation performs the composition operation as indicated until no further basic relations can be removed from the disjunctions. The result can be (i) an empty constraint, which means that the network is inconsistent and no solution exists; (ii) a scenario, meaning it was possible to derive complete information from the originally incomplete network; or (iii) a network that still contains disjunctions as in Figure 2d. The algebraic-closure algorithm has a time complexity of $O(n^3)$, where $n$ is the number of involved objects.

### Conceptual neighborhood

Qualitative spatial and temporal reasoning is also concerned with how spatial configurations described on a qualitative relational level may change over time. The notion of conceptual neighborhood has been introduced by Freksa (1991) to capture how the relation between two objects can continuously change over time in terms of relations from a given qualitative calculus. Two relations from a qualitative calculus are said to be conceptual neighbors if they can be continuously transformed into each other without resulting in a third relation in between. For instance, for two polygons A and B in the plane which can continuously move through 2-D space, the relation between them can change from \{A disjoint B\} to \{A meets B\} without any other 9-IM relations holding in between. Therefore, disjoint and meets are conceptual neighbors. In contrast, disjoint and overlaps are not conceptual neighbors because the transition has to go through meets to get from \{A disjoint B\} to \{A overlaps B\}. Figure 1 visualizes the conceptual neighbor relation between 9-IM relations by arranging them into a so-called conceptual neighborhood graph: relations connected by an edge are conceptual neighbors, while the others are not. Conceptual neighborhood depends on which transformations (move, change size, or change shape) the involved objects can perform. Conceptual neighborhood and the conceptual neighborhood graph have been generalized to configurations of more than two objects. Reasoning based on conceptual neighborhood allows for simulation, planning, and explanatory analysis with qualitative relations for both space and time.
Cognitive evaluations of qualitative calculi

QSTR formalisms are inspired by an interest in how humans intuitively make sense of their spatial and temporal environments, that is, their commonsense understanding of spatial and temporal relations between and properties of entities in their environments. Two goals are achieved by relating natural cognitive and formal artificial approaches to space and time. First, cognitive processes can be understood in great depth and detail if it is possible to formally ground hypotheses on, for example, how many spatial relations humans naturally distinguish between spatially extended entities. Second, the ubiquity of computational/formal systems has fostered the desire to make information systems intuitive for seamless integration into humans’ personal and professional lives. While experimental work on evaluating qualitative calculi in behavioral studies has started to catch up with the developments in calculus design, there remain many open questions and challenges in understanding human conceptualizations of spatial relations in relation to formal QSTR approaches. The goal of behavioral evaluation is to confirm, reject, or refine qualitative calculi, and improve their performance from a cognitive and computational perspective in different application areas such as interface design for geographic information retrieval. Important questions in this regard are: What are intuitive levels of granularity for different fundamental concepts such as topology, direction, or distance? How do humans assess the similarity of individual spatial and temporal relations or spatial arrangements? How do natural language expressions relate to QSTR formalisms? How are other forms of human communication and reasoning such as sketch maps captured using qualitative formalisms? Some of the main results of cognitively evaluating QSTR can be summarized as follows: the majority of QSTR calculi makes finer distinctions than natural cognitive agents, for example, humans almost always aggregate relations of containment (see the relations contains and covers shown in Figure 1); and while topological relations are fundamental to how humans understand space, the semantic context (the identity of the entities in question), will change which relations are aggregated. In other words, which relations are considered more similar to each other than others is dependent on what a particular relation means in a particular context. For further illustration see Figure 1: the relation meets can have neutral (two buildings are side by side), negative (an oil pest reaches the shore), or positive (a castaway reaches an island) meanings, which will change its salience for a natural cognitive system. Depending on the context, the order of importance of relations and properties can change. While a hurricane making landfall is an important spatial relation, it may be even more important depending on whether the hurricane is severe (e.g., category 5) or not (e.g., category 2).

Recent trends and developments

Although the field of QSTR is more than three decades old, new calculi are still being proposed, and their theoretical properties investigated. A recent trend has been the development of multigranular formalisms whose resolution can be adapted to the task at hand. Examples of such formalisms are the STAR calculus, a calculus for absolute directions, and the OPRA model for relative directions. Another ongoing research effort is the combination of several spatial aspects (e.g., topology and direction, or direction and distance) in a single reasoning approach.
Important novel theoretical results have also been achieved relatively recently, for instance, the finding that reasoning with relative direction information that distinguishes left and right is computationally intractable. Results like this have spurred new research on reasoning procedures that are tractable but only approximate the sound and complete reasoning methods as well as possible.

Researchers have also started to look at reasoning problems beyond the deduction of new information and consistency checking in different domains. Applications in different domains have resulted in an increased interest in reasoning approaches based on conceptual neighborhood in the areas of spatial planning, explanatory analysis, and data integration. Moreover, the problems of deriving the most compact qualitative spatial description equivalent to a given QCN, and of deriving a geometric representation that corresponds to or exemplifies a given QCN, have recently attracted the attention of the QSTR community.

These developments have been accompanied by the emergence of freely available QSTR software tools. Specialized reasoners, such as GQR, and general-purpose toolboxes, such as QAT and SparQ, relieve application developers from having to implement models and reasoning techniques described in the literature, and in doing so promote the dissemination of QSTR methods into future spatial information and assistance systems.

SEE ALSO: Behavioral geography; Climate adaptation/mitigation; Cognition and spatial behavior; Neighborhood, conceptual; Overlay, topological; Qualitative information: representation; Representation; Representation: time; Spatial thinking, cognition, and learning; Topological relations

References


Further reading


Geography, at least from the perspective of quantitative geographers, can be defined as the study of spatial processes – the processes that generate both the human and the physical environments we experience on the surface of the Earth. Geographers observe these environments and try to understand how they have arisen – that is, they try to answer the question: “What are the processes that produce the things we observe in the world?” For instance, we try to find answers to such questions as the following.

1. How do land values decline as distance from the city center increases?
2. What causes sea temperatures to vary spatially and temporally?
3. How and why do systems of cities evolve?
4. Why do the locations of people with multiple sclerosis often exhibit a very strong geographical pattern?
5. What are the factors that lead to various movement patterns such as commuting, migration, or shopping?
6. Why do social inequalities exhibit strong geographical patterns?
7. What are the factors producing the particular spatial distribution of an animal or plant species?

The list of such questions is endless and points to both the importance and the prevalence of quantitative geography. However, all these questions have the same underlying basis for being answered – we want to understand the processes that led to the geographical pattern we observe on the earth’s surface. Quantitative methodologies in geography supply this understanding by attempting to provide a linkage between form and process through either empirical analysis or mathematical theoretical reasoning. This type of approach has been a major component of geographical studies since the 1960s and it encompasses an extremely broad variety of applications covering every substantive part of the discipline. Quantitative methodologies include statistical methods, mathematical models, visualization, and computation and, although they have been a part of mainstream geography for over 50 years, they also form the major component of relatively modern terms such as “geographic information science,” “geocomputation,” and “geoinformatics.”

The different ways form and process can be linked have led to the evolution of different types of quantitative geography, so there is no “one” quantitative methodology within geography, and the variety of approaches taken is touched upon below. However, before this discussion, it is useful to consider the common features of spatial data that link the different forms of quantitative methodologies practiced.

What makes quantitative geography special

Quantitative geography proceeds from a description of the human and physical environments. This description typically generates spatial data which can be analyzed in various ways, with the
QUANTITATIVE METHODOLOGIES

objective being to improve our understanding of how these data arose. Spatial data contain attribute information, such as the presence or absence of a disease, the level of a pollutant, or the rate of household burglary in a neighborhood, and locational information about where this measurement is recorded. The locational information, usually in the form of coordinates, is key to analyzing the data through a distinct set of quantitative geographical methodologies because spatial data contain three properties that make them unique: spatial dependency, spatial heterogeneity, and spatial scale.

The data we observe about the real world are rarely, if ever, randomly distributed over space, and spatial dependency refers to the general observation that data recorded for nearby locations are often more similar than are data recorded for locations farther apart (Tobler’s First Law of Geography). One impact of this is that we need to be careful about any inferential tests we employ that rely on an assumption of randomly distributed data, such as parameter estimates obtained from a regression of spatial data. A wealth of literature has been generated on this topic (inter alia, Anselin 1998; Cliff and Ord 1981).

Another impact of spatial dependency is that we can use the property that nearby data values are likely to be more similar than those from locations farther apart, ceteris paribus, to estimate values of a continuous variable at locations for which no measurements are available. For instance, suppose we have a set of air quality monitors at fixed locations throughout a city and we want to estimate a continuous surface of values from these finite measurements. Because of spatial dependency in the data, we can estimate the values of air quality at any location from the values observed at nearby locations. If we have some means of measuring the degree of spatial dependency in the observations, we can weight the data from nearby locations according to their distance from the location at which an estimated value is required. A large literature of techniques for what is known as “spatial interpolation” exists and although there are many different types of interpolation measures, such as inverse distance weighting and kriging, they all rely on the property of spatial dependence for their applicability (Lam 1983).

Spatial heterogeneity exists when the processes generating our observations of the real world vary over space. Traditionally, it was assumed that these processes were constant over space so that a “one model fits all” mentality prevailed. For instance, suppose we were interested in the determination of house prices across a major city and related the variable “house price” to a set of attributes describing the property and its surroundings in a classic multiple regression framework. This would yield one parameter estimate for each relationship in the model so that it would be implicitly assumed that each of these relationships is constant throughout the urban area. However, it is quite reasonable to think that what determines house prices in one part of the city might not be the same as what determines house prices in other parts of the city. For instance, the presence of a swimming pool or a garage might have a differential effect on the price of a house, ceteris paribus, depending on whether the house were in the city center or the suburbs. Similarly, suppose we were investigating the sales of surfing equipment in different cities across the USA – we would not really expect that increasing the advertising budget by a fixed amount, ceteris paribus, would have the same effect on sales in Honolulu as in Iowa City. These are examples of where the processes generating the data we observe in the real world might vary over space. A substantial literature has arisen in the past decade on what are referred to as “local models” – that is, models which recognize that processes might be spatially
heterogeneous (Fotheringham, Charlton, and Brunsdon 1996; Fotheringham, Brunsdon, and Charlton 2002; Lloyd 2007). Such models produce a whole new geographical world – that of relationships – as a complement to our usual focus on geographical variations in data.

“Spatial scale” is a term used to capture two properties of spatial data. One is that the processes that produce what we perceive as the human and physical environments might operate at different spatial scales. For instance, in the study of soil moisture content, we know that there are some processes, such as rainfall and temperature, that operate at a regional scale (albeit with minor variations) and others, such as shade, slope, and aspect, that operate at a much more local scale. Similarly, there are macro-level processes, such as regional economic performance, that affect house prices as well as micro-level processes, such as how well the property is maintained, whether it is near a major highway, and so on.

The other effect of scale relates to the units for which spatial data are collected and analyzed. One of the challenging issues in quantitative human geography is that, for confidential reasons, data relating to individuals are typically only released in aggregate form for some set of arbitrary administrative units such as census divisions or political units. However, it has been shown repeatedly and in different contexts that the inferences we draw from the analysis of spatial data can be dependent on the spatial scale at which the data are analysed – a problem often referred to as the modifiable areal unit problem and well documented for bivariate relationships (Openshaw and Taylor 1979), multivariate relationships (Fotheringham and Wong 1991), and mathematical models (Fotheringham, Densham, and Curtis 1995). Consequently, we may draw different conclusions about the nature of the relationships being investigated, depending on the set of aggregate spatial units for which the data are reported. An unseen, but equally worrying, problem is that the results of our analysis might vary over space if the spatial data units we are using are not generated in the same way. A classic case of this is the use of county-level data in the United States – the process by which counties are defined varies by state and the physical sizes and the populations of counties vary enormously across the country.

A brief history

Initially, quantitative methodologies were introduced into geography with little if any recognition of the subtleties inherent in spatial data described above. Prior to the 1960s, geographical studies were almost exclusively descriptive, but mathematical and statistical techniques became increasingly employed for the study of spatial patterns in what is often referred to as “the quantitative revolution.” Such techniques were often imported into geography from other less spatial disciplines, such as statistics, physics, and economics, and although this was a necessary and important initial step in the development of quantitative methods within geography, with hindsight the techniques used were sometimes naïve and lacked a real understanding of explicitly spatial processes. Mathematical models were often geometric and normative, and empirical analyses were often devoid of real spatial awareness. Unfortunately, critiques of quantitative geography often still focus on this era and are repeated by individuals who fail to grasp that the discipline has moved on considerably to address specifically spatial properties and processes. An enormous amount of progress has been made and, today, quantitative geography has matured to the stage where it is a net exporter of ideas and techniques to other disciplines in which researchers recognize that a
large number of processes and datasets in their areas of study are spatial.

Although there are literally hundreds and perhaps thousands of techniques that are incorporated under the umbrella of quantitative methodologies in geography, they can be categorized into two broad groups: mathematical models and statistical techniques. The former emphasize statements about how the world works and include spatial interaction models, optimization models, and differential equation-based models of climate change. The latter emphasize either the description of spatial patterns, such as spatial autocorrelation statistics, or the identification of relationships between variables exemplified by regression models and various other types of multivariate models, such as factor analysis, discriminant analysis, and so on. The distinction becomes blurred, though, in situations such as spatial interaction modeling, where the models are developed from mathematical principles but are then calibrated by statistical processes such as regression and maximum-likelihood estimation.

It is probably reasonable to state that quantitative human geographers have been more concerned with statistical approaches than their physical geography colleagues because of the inherent uncertainty in human spatial behavior. It is both a weakness and a strength of human geography that it is extremely difficult, and perhaps impossible, to develop any statements akin to scientific “laws” because of the great difficulty in understanding and predicting human behavior. However, there are certainly regularities in human spatial behavior that are useful to identify and quantify, an obvious one being that our propensity to interact with a location decreases as the distance to that location increases, *ceteris paribus*. This concept, known as “distance-decay,” is extremely powerful in explaining patterns of movement such as migration, retailing, and commuting, and even patterns of communications where the friction of distance is in theory virtually nonexistent, such as telephone calls, texts, and tweets. However, there are many other factors which also can play a role in determining patterns of movements and communications, so that spatial interaction models are never without error and therefore always contain a stochastic component representing the myriad of unknown or unmeasurable processes that affect our decision to interact with a specific location or person.

The presence of uncertainty does not mean, however, that modeling exercises in geography are doomed to failure and should be abandoned, as some would have it. Modeling human spatial behavior generates very convincing and powerful evidence on the major processes affecting the outcomes we observe whilst recognizing that some uncertainty exists. The power of statistical modeling is that we can quantify this uncertainty and thereby have standards to assess the probability that our inferences about the determinants of human spatial behavior are wrong. In this way, we accumulate evidence to make a compelling case for a relationship. This ability to quantitatively assess the quality of our results and inferences is an extremely powerful reason for the use of quantitative approaches in geography. The element of uncertainty has also acted as a catalyst to the development of new techniques in order both to incorporate it and to minimize it, so that the subject matter of human geography is now one of the most fruitful areas of research in academe, as evidenced for example by the number of nongeographers “discovering” geography (inter alia, Noulas *et al.* 2012; Simini *et al.* 2012).

One final comment on the use of quantitative methodologies by both human and physical geographers is that they act as a means of communication between two potentially diverse groups. In an era when geography has become
QUANTITATIVE METHODOLOGIES

increasingly fragmented and broad, encompassing both humanistic and scientific approaches, it is increasingly important that the constituent components of the discipline have a means of communicating with each other – quantitative methodologies provide this bridge between human and physical geography.

Recent trends

From the early days of geography’s “quantitative turn” in the 1960s, the application of mathematical and statistical modeling to the study of spatial processes has seen a blossoming of approaches as our understanding becomes ever more sophisticated. Five trends in particular are worth commenting upon. These are: (i) the role of geographical information systems or science (GIS); (ii) advances in computing power; (iii) the focus on the local; (iv) the arrival of the “big data” era; and (v) the adoption of visualization as a means of conveying information.

Although it began a decade or so after the so-called quantitative revolution in geography, the development of GIS has roughly paralleled that of quantitative geography. The positive effect GIS has had on the external perception of geography cannot be overstated. Although GIS (particularly geographical information science rather than systems) is for some people synonymous with quantitative geography, the two areas were initially quite distinct, and significant differences remain. GIS arose from the cartographic tradition, and initial developments largely involved computer-based mapping software that allowed different types of spatial data to be linked, queried, and displayed. Statistical and mathematical analysis was almost nonexistent. Even today, when most GIS contain a sophisticated array of quantitative techniques, including many commonly used statistical techniques plus specialized routines for interpolation, optimization, and network and surface analysis, there are many GIS professionals who would not claim to be quantitative and many quantitative geographers who would not claim to be GIS experts. However, GIS has had a profound effect on quantitative geography in that the demand for quantitative geographical techniques has soared as GIS has become virtually ubiquitous. GIS exposed people to the power of spatial data through visualization and simple types of analysis, such as buffering, overlay, and point-in-polygon operations, and as users have become more sophisticated, their needs for spatial data analysis have grown. This has created new markets for quantitative geographical techniques that have been bundled into newer versions of GIS software (an example being that of geographical weighted regression (GWR), which first appeared in ArcGIS 10) and also for stand-alone spatial analysis software such as GeoDa and PySAL (Python Spatial Analysis Library) (Anselin and Rey 2014).

A second trend affecting quantitative methods in geography has been the rise in computing power over the past 30 years. This has had two obvious consequences. One is that statistical inference no longer relies on theoretical distributions being available because brute force simulations can be used to generate empirical distributions for any statistic. This has been particularly beneficial to the application of new types of spatial statistics where theoretical distributions are not always available or are extremely cumbersome to work with. The other is that quantitative geomorphers have been able to develop and calibrate increasingly complex models (see below). Forty years ago it was relatively easy to produce spatial models which were essentially inoperable because of a lack of computer power and/or data. Today neither of these constraints is an issue: the only limitation
QUANTITATIVE METHODOLOGIES

to the development of improved spatial models and spatial statistics is our ability to devise them.

Aided by the advances in computational power, a powerful trend in quantitative geography over the past 20 years has been the development of local statistical techniques and models. Previously, an unquestioned, and probably unrealized, assumption of most quantitative research was that the processes being examined were stationary over space. That is, a “one size fits all” mentality predominated in which a single parameter estimate or a single descriptive statistic such as a measure of spatial autocorrelation was assumed to be constant over space and hence representative of every location. Initially, through spatial interaction modeling, with the development of origin- and destination-specific models, it came to be realized that processes might be spatially heterogeneous and hence new model forms were needed which allowed this variation to be captured. Ultimately, this has led to the development of local spatial modeling techniques such as GWR and local descriptive statistics such as local indicators of spatial association (LISA) (Fotheringham, Brunsdon, and Charlton 2002; Anselin 1995). These techniques, pioneered by geographers, are now commonplace in many other disciplines. They provide potentially enormous amounts of information on spatial process and a whole new geography to investigate – that of relationships rather than data.

Allied to developments in local modeling, which can be termed “big models” because of their complexity, have been advances in data collection technologies and data dissemination resulting in what are known as “big data” (Goodchild 2013). Many big datasets contain geocoded data from new technologies such as LiDAR (Light Detection and Ranging), global navigation satellite systems, and satellite imagery or from new forms of social media such as Twitter and from a recent phenomenon whereby individuals want to make their locations known to others, such as FourSquare check-ins and geocoded photograph repositories. Consequently, we are now in an era when data availability is no longer the bottleneck it once was; rather, the new bottleneck is our ability to provide useful information on spatial processes from these big data. Currently, big datasets in geography are of the order of terabytes and petabytes but, with GPS chips becoming virtually ubiquitous, datasets of these magnitudes will seem relatively small within the next 10 years. The need to distinguish the signal from the noise in these large, multivariate datasets will be even greater and the process even more difficult than it is today.

Big data and big models both provide similar challenges in trying to make sense of often very noisy data. In the former, the data are observations of the real world; in the latter they are model outputs, usually in the form of thousands of parameter estimates. The need to provide useful insights into such large arrays of data has led to an increased use of visualization as an aid to understanding. The challenge to present large, multivariate datasets in a visual form that can provide insights in combination with the human brain has spawned the new field of visual analytics (Dill et al. 2012). A large set of techniques has been developed, and continues to be developed, in order to convert dry tables of data into appealing visual displays which the brain can process much more easily. Often these techniques rely on quantitative methods linked to some form of mapping and this has proven to be a fruitful avenue of research for geographers who can think both visually and quantitatively.

The above trends – the rise of GIS, the development of massive computing power, the stressing of the local over the global, the advent of big data, and the increased role of visualization – have all combined to make quantitative geography more visible but also more data-driven.
and less theoretical. The relative ease by which empirical analyses can now be conducted has pushed theory into the background; we need a steady flow of innovative ideas to maintain our leadership in spatial data analysis and modeling.

Challenges

To conclude this brief overview of quantitative methodologies, it is useful to consider what challenges lie ahead. Three generic issues, all interrelated, can be identified: (i) model choice (or variable selection); (ii) inference; and (iii) robustness.

The problem of model choice or variable selection is common to many disciplines, ever more so with the increasing availability of highly multivariate datasets. Faced with large numbers of potential explanatory variables and very little theory to guide us, the question arises as to how we select the appropriate combination of variables to help us understand the spatial distribution of a particular attribute. If we include too many overlapping variables in our model, we incur the problems of multicollinearity – confusion of parameter estimates and inflated standard errors. If we leave out relevant variables, we incur the omitted variable problem – biased parameter estimates. Of course, there are some standard statistical procedures that provide guidance on model selection, such as various stepwise regression procedures, but they are not always helpful, and the question remains how such techniques are to be applied to local models. In the presence of spatial heterogeneity there could be different optimal models for each location, and global diagnostics would therefore not be very helpful.

One answer is to develop better theory, but in many instances, theory is of little use. Suppose, for example, we are interested in understanding why migrants chose certain destinations over others and we have a model that relates destination choice to attributes of the destinations and the individuals making the choice. Suppose we think one variable that might influence destination choice is a measure of house prices at the destination. We could argue theoretically that the expected direction of this relationship is negative: higher house prices make a destination less attractive, ceteris paribus, due to fewer houses being available in one’s price range or one’s disposable income being reduced due to higher housing costs. On the other hand, we could argue theoretically that areas with high house prices tend to have higher rates of return on housing over time and therefore represent a wise financial investment making a destination more attractive, ceteris paribus. It is extremely difficult to separate these effects, and the empirical result globally might be that the variable appears to be unimportant to the decision. Local models might shed some light on this issue – if the pattern of local parameter estimates is positive in some areas and negative in others, this may give us a better indication of the effects of house prices on the migration decision.

Linked to the above is the issue of statistical inference. In an era when complex local models can be calibrated with huge datasets containing hundreds of potential explanatory variables, we now have the ability to estimate unprecedented numbers of parameters. In theory, this ability should lead to ever-greater insights into spatial processes. However, it also brings greater problems of separating “real” positives from “false” ones. Issues such as the multiple hypothesis testing problem (Williams, Jones, and Tukey 1999) become more important as we add more variables to our models and test the same hypothesis with repeated samples. There is a growing need, therefore, to examine statistical inference more carefully to reduce the chances
of being misled by noise in the data. As datasets become larger and as our models become more complex, it becomes more difficult to identify what is really meaningful.

The issue of robustness has long plagued quantitative geography. The empirical results of our models cannot generally be replicated when applied to a different area. There are countless examples of the same, or similar, models being calibrated with different datasets and yielding different results. In some applications a parameter might be significant, while in others it is not; it might be significantly positive in some cases, yet significantly negative in others. We now have a potential explanation for this in terms of spatial heterogeneity. If we accept that processes might vary over space, it follows that parameter estimates, which measure these processes, should also exhibit variation in excess of that we would expect from random sampling. However, we are still left with the problem that few of our models seem to be robust enough to be applied to situations other than that in which they are calibrated, and therefore lack accuracy in terms of prediction.

Final thought

The above issues should be viewed as challenges and not just as problems. They are what make the area an exciting one in which to work. There is no doubt that quantitative methodologies have much to offer because they can provide strong evidence for how various spatial processes shape both the human and the physical environments. Mathematical and statistical models of spatial processes are useful – nobody claims they are perfect. To criticize them for not being perfect is to misunderstand the application of quantitative methodologies. We are at an unprecedented era in terms of the demand for individuals with skills to analyze spatial data. Geographers with quantitative skills are leading the way in this area whether they are labeled quantitative geographers, data scientists, geoinformaticists, geocomputationalists, or GI scientists. Investment in spatial data capture and analysis has never been higher; society is increasingly not just data-driven but spatial-data-driven; countries around the world are developing centers for spatial data analysis; and demand for such skills in the private sector is booming. The future for quantitative methods in geographical enquiry is very bright.

SEE ALSO: Distance decay; Geocomputation; Geographic data mining; Geographic information science; Geostatistics; Measuring spatial dependence; Models in geomorphology; Modifiable areal unit problem; Network analysis; Point pattern analysis; Scale; Spatial analysis; Spatial interaction; Tobler’s first law of geography; Visualization

References


Further reading


Today, “quantitative geomorphology” could be considered redundant, as geomorphology is fundamentally based on quantifications. In fact, quantitative geomorphology refers to a specific way of exploring geomorphological questions and takes its roots in the early debates between qualitative and quantitative approaches. When reviewing the use of the term “quantitative geomorphology” in the scientific literature, it predominantly appears during two periods: the 1950–1960s and the 2000s.

The first period includes pioneer researchers such as Horton, Strahler, Morisawa, or Chorley, who first published contributions entitled “quantitative geomorphology…” or “quantitative analysis…,” and argued the need to think about processes and explain how forms are related to processes.

The pioneer quantitativists defended the need to move from “qualitative” to “quantitative” geomorphology to study the spatial organization of forms (stream ordering, stream network, and basin characters) as related to driving factors (mainly hydraulics), by developing process-based and systemic approaches to understanding landscape complexity; these are the theoretical framework for the discipline today (Church 2013). This period is seen as the “golden age,” wherein emerged the concepts of “dynamic geomorphology” and “functional geomorphology.” These concepts developed fairly independently of historical geomorphology, which, in turn, gave birth to paleoenvironmental studies and geoarcheology.

A second quantitative revolution occurred in the 2000s with the emergence of new geographic data and expanding computer capacities accessible to scientific teams around the world. Nearly five decades after the pioneers of geomorphology, these new contributions consistently refer to “quantitative geomorphology,” such as in a special issue published in Géomorphologie in 2011, entitled “quantitative hydrogeomorphology” that described the emergence of a new scientific era, similarly argued by Marcus and Fonstad (2010) in “Remote Sensing of Rivers, The Emergence of a Subdiscipline in the River Sciences.”

Qualitative information (mainly visual observation) is biased by the perception and acuity of the observer, his/her experiences, memories, beliefs, and/or attachment to a given theory, as discussed in the case of the Davisian theory of cycles of erosion (Rhoads and Thorn 1996). Qualitative information that supports shared truth is limited to observable facts, even though a large set of them is not visible or is difficult to observe. The qualitative approach is also limited to those processes that are notably changing, many of which are not observable without measurements over time. Quantification is necessary to simulate processes in order to understand and generalize cause and effects based on order of magnitude and changes in volume or mass (e.g., sediment budgeting, appropriate timescale estimates). A qualitative approach is often linked to an inductive approach, so that the observer lets the field speak without formulating any preliminary hypothesis (Rhoads and Thorn
1996). Even if a qualitative approach guided by a critical attitude, multiple-hypotheses, and deductive reasoning reached a truth shared by a large set of people, it would still be limited by a lack of precision and set of facts. Convergence of evidence is also a critical issue to give truth to interpretation. In this domain, quantification is useful as it can provide a wider set of solutions and make interpretations more robust.

Quantification is based on data collected by instruments. Quantitative geomorphology supports the idea that objectivity should be reinforced by quantification, reliable and precise data, and shared procedures to avoid any bias. It is a way to gather more precise information that can be shared by observers. Such factual information can then be interpreted to reformulate questions so as to design a new research strategy to gather additional facts. It allows one to distinguish and interpret facts and introduces a strong experimental framework: the so-called deductive approach of hypothesis validation.

If quantification can be done in different ways, two complementary strategies are often observed, with sometimes disciplinary debates to consider which one is the most valuable, one focusing on physical laws and the other exploring the Earth complexity using mainly statistics. Physics, mainly mechanics, aims to understand the elementary laws controlling the evolution of processes and forms, is mainly grain-based, and is supported by in situ and ex situ experiments, numerical models, and simulations. It simplifies the real world to understand step by step how it works. Earth complexity analysis is usually more holistic, approaching Earth organization and functioning by ordering, classifying, and understanding factors by discriminating temporal states and spatial units. For this type of analysis, statistics has become important, similarly to other integrative sciences such as ecology, to explore complex systems.

Geomorphology now fundamentally relies on quantification to answer different questions, to characterize the magnitude and frequency of events and critical geomorphic phenomena (i.e., sediment transport, landslide), and to assess the evolution and organization of forms and their connections. Quantification is founded in physics. It is made possible by technological advances in optics, mechanics, electronics, or geoinformatics, including geocomputing, statistics, mathematical morphology, geometry, modeling, remote sensing techniques, and geographic information systems (GIS). It deals with dating, geopositioning, and imaging the Earth surface and the water column using a set of innovative techniques such as optical stimulated luminescence (OSL), isotopes (U/Th isotopic ratios, $^{14}$C, $^{137}$Cs, $^{210}$Pb, etc.), airborne and terrestrial LiDAR, various satellite and airborne imagery (hyperspectral, hyperspatial, radar, etc.), ground penetrating radar (GPR), and ground sensors such as radio frequency identification (RFID) and cameras (Thorndycraft, Benito, and Gregory 2008; Carbonneau and Piégay 2012). In this field, temporal resolution is also becoming significantly improved, notably with ground sensors that record phenomena almost continuously (e.g., a few images per second for ground video cameras to a few images per year with the Pléiade 1A and 1B satellites with a spatial resolution equivalent to the conventional ortho-rectified photos provided by national agencies (0.7 m versus 0.2–0.5 m)) or UAV technologies. With the development of the Internet, more and more state services offer archived data online, increasing the data availability for research purposes. These sites include online metadata/data catalogs and online geotreatment tools or tool packages to explore geomorphic systems within GIS. The computer revolution not only allows for more complex modeling in terms of hydraulic simulations but also provides new techniques to
explore complex sets of data needing large computer capacities (e.g., multiple-agent systems, segmentations and classifications, analyses of uncertainties based on probabilistic approaches and randomization tests, etc.).

New techniques and new data provide new insights. There is now the ability to more intensively observe fine sedimentary structures, topography under forest cover, and river bathymetry, in order to study processes which were not yet fully examined (e.g., flux of wood, bank erosion during a flood event, landslides or debris flows, etc.). This has allowed us to widen our range of observations and our vision, which was extremely limited in space and time by field-based approaches. Traditionally field based, geomorphology is now reaching the point in its evolution where remotely collected images are providing more data than field observations, opening exciting prospects to explore processes and forms not only at a local scale but at much wider scales by combining observations and simulations. In a certain way, geomorphologists not only base their knowledge production on field observation but also on remote sensing information, a complementary perspective of observing Earth. In widening our space and time framework, these new data should improve our understanding on how local observations can be generalized.

Methods promoted by reductionists developed in the 1950s are now being progressively adapted for applications at larger spatial scales. Reductionist quantitativists must now collaborate with holistic quantitativists to explore phenomena in this big data era recognizing geomorphology as a “system science.” This new big data era more closely connects physicists with statisticians, who explore Earth complexity from two joined perspectives. It is influencing not only functional/dynamic geomorphology but also historical/evolutionary geomorphology as shown by Thorndycraft, Benito, and Gregory (2008) or illustrated by works such as Notebaert et al. (2009) opening new bridges between these two main branches of the discipline.

**SEE ALSO:** Big data; Cloud computing; Digital elevation model and digital surface model; Geocomputation; Geographic information system; Geomorphological mapping and geospatial technology; Modeling uncertainty in digital elevation models; Models in geomorphology; Photogrammetry: 3-D from imagery; Quantitative methodologies; Representation: 3-D; Representation: dynamic complex systems; Synthetic aperture radar; Technology; Unmanned aerial vehicle (UAV)

**References**


Since the recognition in the mid-nineteenth century that glaciers had been considerably more extensive than at present, the Quaternary has been synonymous with glaciation of the mid-latitudes. Today evidence from both the land and ocean floor sediment sequences demonstrates that the major continental glaciations occurred repeatedly over what are now temperate regions of the Earth’s surface. Knowledge about the number of glaciations has increased, and also knowledge about the extent of the Pleistocene ice sheets. Progress has been rapid since the 1970s and can be seen in the INQUA Work Group 5 project “Extent and Chronology of Quaternary Glaciations” (Ehlers and Gibbard 2004a, 2004b, 2004c) and the “Extent and Chronology of Quaternary Glaciations, a Closer Look” (Ehlers, Gibbard, and Hughes 2011) (Figure 1). While the maximum extent has undergone comparatively few changes, the Late Weichselian ice sheet of northern Europe and Siberia has changed fundamentally. Today, a single ice sheet is envisioned instead of the previous three major glaciation centers. Major differences include the (i) glaciation of the Bering Sea, (ii) non-glaciation of the North Sea, (iii) non-glaciation of the northern Urals, and (iv) limitation of glaciation in Siberia to the Putorana Massif and the Taimyr coast. Extensive ice-rafting, an indication that glaciers had reached sea level, is found from the earliest cold stage (2.6–2.4 Ma) in both the North Atlantic and North Pacific oceans. The initiation of conditions that resulted in glaciation resulted from the long-term cooling trend in world climates that began early in the previous Tertiary period. Apart from some limited activity in the Eocene, significant glaciation began in the late Oligocene (ca. 35 Ma) in eastern Antarctica. It was followed by mountain glaciation through the Miocene (23–5.3 Ma) in Alaska, Greenland, Iceland, and Patagonia and later in the Pliocene (5.3–2.6 Ma) in the Alps, the Bolivian Andes, and possibly in Tasmania. From the Neogene, glacially derived ice-rafted debris is found in ocean-sediment cores from the North Atlantic region, including the Barents Sea, and areas adjacent to Norway, north and southeast Greenland, Iceland and northern North America, and in the Southern Ocean off Antarctica (De Schepper et al. 2014).

Glaciation during the Quaternary

The climatic variation of that characterizes the Earth’s climate during the late Cenozoic and indeed before, is controlled by variations of the planetary orbit around the sun that therefore controlled the receipt of solar energy at the Earth’s surface. These Milankovitch variations, named after their discoverer, are responsible for the cyclic climate changes that characterize the Quaternary and indeed much of Earth’s history. One of the most critical ways they are expressed is through the development of “ice
ages” or periods when glaciation extended across large areas of the Earth. The Early Pleistocene (2.6–0.8 Ma) was characterized by climatic fluctuations dominated by the 41 Ka precession cycle, during which relatively few cold periods were sufficiently cold and long to allow the development of substantial ice sheets. Only 14 of the 41 cold stages of that period currently show evidence of major glaciation. They include the Plio-Pleistocene boundary events Marine Isotope Stage (MIS) 104, 100, and 98, together with Early Pleistocene MIS 82, 78, 68, 60, 58, 54, 52, 36, 34, ?30, and 26 which reach $\delta^{18}O_{\text{ocean}}$‰ of c. 4.6–5. It is not until the transition in dominant orbital cyclicity to the 100 Ka cycles, that began circa 1.2 Ma and was fully established by about 800 Ka (“Middle Pleistocene transition”), that the cold periods (glacials) were regularly cold and long enough to allow ice-sheet development on a continental scale, outside the polar regions. However, it is during MIS 22 (ca. 870–880 Ka) that the first of the “major” cold events occurred that reached critical values of circa 5.5 or above $\delta^{18}O_{\text{ocean}}$‰ equivalent to substantial ice volumes that typify glaciations of the later Pleistocene (i.e., MIS 16, 12, 10, 6, 4–2). Potentially therefore, it is likely that there were a minimum of 20 periods during which extensive glaciation could have developed during the last 2.6 Ma, with the most extensive (ca. 5–6 periods) being limited to the last 900 Ka (Table 1, Table 2, and Table 3).

Precisely where these glaciations occurred and how far they extended is very difficult to determine, given that the remnants of less extensive early glaciation tends to be obliterated and mostly removed by later, more extensive advances. Although this is so in all terrestrial areas, it is especially difficult in mountain regions where the preservation potential of older sequences rapidly
Table 1  Occurrence of glaciation in Europe through the Cenozoic based on numbers of observations presented in contributions to the INQUA project “Extent and Chronology of Quaternary Glaciations” (Ehlers and Gibbard 2004a). ?, questionable glacial deposits; arrows, glaciomarine deposits; MIS, Marine Isotope Stage. Geomagnetic polarity epochs: B, Brunhes (normal); M, Matuyama (reversed); G, Gauss (normal).

<table>
<thead>
<tr>
<th>Stage/Age</th>
<th>MIS (approx.)</th>
<th>Austria</th>
<th>Belgium</th>
<th>Bosnia-Herzegovina</th>
<th>Croatia</th>
<th>Czechia</th>
<th>Denmark</th>
<th>Estonia</th>
<th>Finland/Scandinavia</th>
<th>Masowia</th>
<th>Norway</th>
<th>Poland</th>
<th>Eastern European Russia</th>
<th>Sierra Nevada</th>
<th>Sweden</th>
<th>Bavaria</th>
<th>Slovakia</th>
<th>Spain</th>
<th>Switzerland</th>
<th>Ukraine</th>
<th>North Sea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weichselian</td>
<td>2</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td>Saalian</td>
<td>6</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td>Elsterian</td>
<td>12</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>218 or 206</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td>Cretaceous</td>
<td>18</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>22+</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>34–38</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>64–72</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>176–182</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td>Phocene</td>
<td>ca. 98–104</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>Upper</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>Lower</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td>Miocene</td>
<td>Upper</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
<tr>
<td></td>
<td>Lower</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
<td>△</td>
</tr>
</tbody>
</table>
Table 2  Occurrence of glaciation in North America through the Cenozoic based on numbers of observations presented in contributions to the INQUA project “Extent and Chronology of Quaternary Glaciations” (Ehlers and Gibbard 2004b). ?, questionable glacial deposits; arrows, glaciomarine deposits; MIS, Marine Isotope Stage. Geomagnetic polarity epochs: B, Brunhes (normal); M, Matuyama (reversed); G, Gauss (normal).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wisconsin</td>
<td>2</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>△△</td>
<td>?△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Illinois</td>
<td>6</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Pre-Illinoian A</td>
<td>8</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Pre-Illinoian B</td>
<td></td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Pre-Illinoian C</td>
<td></td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Pre-Illinoian D</td>
<td></td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Pre-Illinoian E</td>
<td></td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Pre-Illinoian F</td>
<td>M</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Pre-Illinoian G</td>
<td></td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Pre-Illinoian H</td>
<td></td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Pre-Illinoian I, J</td>
<td></td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Pliocene</td>
<td>G</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td></td>
<td>Upper</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td></td>
<td>Lower</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
<tr>
<td>Miocene</td>
<td></td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
<td>△△</td>
</tr>
</tbody>
</table>
### Table 3  Occurrence of glaciation in the rest of the world through the Cenozoic based on numbers of observations presented in contributions to the INQUA project “Extent and Chronology of Quaternary Glaciations” (Ehlers and Gibbard 2004c). ?, questionable glacial deposits; arrows, glaciomarine deposits; MIS, Marine Isotope Stage. Geomagnetic polarity epochs: B, Brunhes (normal); M, Matuyama (reversed); G, Gauss (normal).

<table>
<thead>
<tr>
<th>Stage/Age</th>
<th>MIS (approx.)</th>
<th>Antarctica &amp; Southern Ocean</th>
<th>Argentina</th>
<th>Patagonia</th>
<th>Bolivia</th>
<th>Chile</th>
<th>China – Tibet</th>
<th>China – Taiwan</th>
<th>Colombia</th>
<th>E. African mountains</th>
<th>Ecuador</th>
<th>Mexico</th>
<th>New Zealand</th>
<th>Tasmania</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weichselian/ Wisconsinan</td>
<td>2</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td></td>
<td>3, 4</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Saalian/Illinoian</td>
<td>6, 8</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Elsterian/pre-IllinoianB</td>
<td>?12</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Early Middle Pleistocene</td>
<td>?14, 16</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td></td>
<td>?18 or 20</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Early Pleistocene</td>
<td>M</td>
<td>22+</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td></td>
<td>34–38</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td></td>
<td>64–72</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td></td>
<td>78–82</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Plio/Pleistocene</td>
<td>c. 100–104</td>
<td>▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Pliocene</td>
<td>G</td>
<td>Upper ▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lower ▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Miocene</td>
<td></td>
<td>Upper ▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lower ▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>Paleogene</td>
<td></td>
<td>Paleogene ▲ ▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
<td>▲</td>
</tr>
</tbody>
</table>

diminishes with time and subsequent glaciation. However, examination of the frequency of glaciation through the Cenozoic indicates that glaciation in the Southern Hemisphere having been established first, principally in Antarctica and southern South America, occurred continually from the Early Neogene to the present day. By contrast, Northern Hemisphere glaciation, although initially somewhat restricted, increased markedly at the beginning of the Quaternary, increasing again in frequency in the latest Early Pleistocene and reaching very high levels in the Middle–Late Pleistocene. While this pattern is not unexpected, the striking increase in ice sheets through the Quaternary clearly emphasizes that worldwide glaciation is in effect a northern-hemispheric phenomenon.

Examination of the evidence accumulated in the INQUA project (Ehlers and Gibbard 2004a, 2004b, 2004c; Ehlers, Gibbard and Hughes
QUATERNARY GLACIATIONS

2011), supported by other published sources, demonstrates the current state of knowledge. When examining the resulting tables (Table 1, Table 2, Table 3), it is important to bear in mind that the stratigraphical control between regions, beyond the range of radiocarbon-dating, is weaker than might be desired. This is particularly a problem outside Europe where biostratigraphy is less developed and the sheer extent of the unexplored regions makes future discoveries likely. Thus the presentation below can only be a first step toward a comprehensive overview. The correlations applied here are based on the Global Correlation Chart (Gibbard and Cohen 2013).

Plio-Pleistocene glaciation

Evidence of glaciation is widespread from throughout the Quaternary and indeed the Neogene in the Northern Hemisphere. The longest sequences are restricted to Alaska and the adjacent northwest territories of Canada which, together with Greenland and the Rockies, preserve evidence of glaciation from the Neogene to the present. In northern Canada and Alaska, the oldest till and accompanying ice-rafted detritus in marine settings, dates from the early Miocene, with regionally widespread glaciation occurring in the Pliocene and regularly throughout the Pleistocene. In adjacent British Columbia a comparable sequence is found, particularly in the north. Similarly, in Greenland and Iceland glaciation began in the Miocene, occurring regularly through the Pliocene and onwards to the present day in the mountains. Likewise, in Norway, its adjacent offshore and the neighboring Barents Sea, glaciation is recorded from the early Miocene, early Pliocene and Plio-Pleistocene. By the late Miocene, inland ice shields were periodically present in Greenland, especially in the mountainous east, with ice reaching the sea in southeast Greenland, although contiguous ice sheets have occurred since the earliest Pleistocene (ca. 2.3 Ma). In the eastern Rockies of the United States a much shorter glacial sequence occurs. Here Plio–Pleistocene-aged till is known from Montana, North Dakota, and California. On Mount Kilimanjaro in East Africa the first glaciation is recognized at circa 2.0 Ma (ca. MIS 68). In Europe glaciation before the Middle Pleistocene glaciation is represented only by ice-rafted material, outside the mountain regions (e.g., in the Netherlands, lowland Germany, European Russia, and Britain). Substantial glaciation of the Baltic region late in the Early Pleistocene is indicated by erratic materials in the Netherlands (1–1.2 Ma; MIS 34–36). Glaciation is also established in the Alps from the Plio–Pleistocene.

In the Southern Hemisphere, glaciation is much longer established, as noted above. Here the ice already formed in the late Eocene–early Oligocene in East Antarctica and built up in a step-like pattern through the Neogene (De Schepper et al. 2014). The present polar conditions were already established by the Early Pleistocene after 2.5 Ma. A similar history is known from the Piedmont areas of Argentina and Chile where substantial ice caps were established by 14 Ma. Widespread lowland glaciation between 2.05–1.86 Ma (ca. MIS 68–78) followed by the “Great Patagonian Glaciation” took place at 1.15–1.00 Ma (ca. MIS 30–34). Further north, the earliest glaciation recorded in the Bolivian Andes dates from 3.27 Ma with extensive events at 2.2 Ma (ca. MIS 82). In Columbia the record also begins at 2.5 Ma. The earliest records in Australia are found in Tasmania and New Zealand from the Plio–Pleistocene (2.6 Ma: MIS 98–104). Only slightly younger is New Zealand’s oldest known glacial event (the Porika Glaciation).
The “glacial” pleistocene

The “glacial” Pleistocene effectively begins with ice sheets spreading over vast lowland areas, particularly around the North Atlantic region, and the intensification of global cold period (glacial) climates in general. It coincides with the “Middle Pleistocene transition” (1.2–0.8 Ma) when the transition from the dominant 41 to 100 Ka Milankovitch orbital cyclicity resulted in periods sufficiently cold and long to allow the development of continental-scale ice sheets.

The till sheets of the major glaciations of the “glacial Pleistocene” are found throughout large parts of Europe and North America and especially in the lowlands and under the sea. Widespread lowland glaciation began in the early Middle Pleistocene shortly after the Brunhes/Matuyama palaeomagnetic reversal (780 Ka). In Europe, the phases represented include the Weichselian (Valdaian, MIS 4–2), Saalian (Dniepr and Moscovian, MIS 6, 8, and 10), Elsterian (Okan, MIS 12), and the Donian (Narevian, MIS 16). More limited glaciation may also have occurred in the circum-Baltic region during the latest Early Pleistocene (MIS 20 and 22). Curiously, evidence for early Middle Pleistocene glaciation is absent from the North Atlantic and Norway, while it is certainly present in Denmark, the Baltic region, and European Russia. In the Italian Dolomites, glaciation becomes established in MIS 22. Comparable evidence is also found from north of the Alps in Switzerland and southern Germany. Further to the west, in the Pyrenees, the oldest glaciation identified is of late Cromerian age (MIS 16 or 14). Widespread lowland glaciation again is first seen throughout North America in MIS 22. From this point onwards, major ice sheets covered large regions of the continent during the Middle Pleistocene pre-Illinoian events MIS 16, 12, 8, and 6 (Illinoian sensu stricto) and the Late Pleistocene MIS 2–4 (Wisconsinan). In Mexico, the oldest moraines on volcanoes have been dated to 195 Ka and probably relate to a pre-MIS 6 glaciation. Evidence from east Greenland suggests that its quasi-permanent ice sheet may have almost disappeared during the Eemian Stage interglacial (ca. MIS 5e). Glaciation of Tibet and Tianshu is not recorded before the Middle Pleistocene, of which that during MIS 12 was the most extensive. In Tianshu older glaciation (?MIS 16) may have also occurred. This apparently delayed glaciation of the Himalayan chain might reflect a late uplift of high Asia. Subsequent events took place during MIS 8, 6, and 4–2, and continue today in the highest peaks.

As in Europe and North America, glaciation increased in intensity throughout the Andean chain from 800 Ka to the present day, but in the south it was less extensive than during the Early Pleistocene events. In Australasia, following a 1 Ma break, the glacial record continues in MIS 12, followed by MIS 6, 4, and 3. In Tasmania, an early Middle Pleistocene event, possibly during MIS 16, is followed by glaciations during MIS 6 and 3. The glacial record during this time in Africa is restricted to the East African mountains, Mount Kilimanjaro, Mount Kenya, and the High Atlas, where glaciations appear to be broadly equivalent to those elsewhere, that is, during MIS 12, 6, and 2.

Last glaciation

The term Last Glacial Maximum or LGM is widely accepted as referring to the maximum global ice volume during the last glacial cycle corresponding with the trough in the marine isotope record centered on c. 18–14 C Ka BP (Martinson et al. 1987) and the associated global eustatic sea level low also dated to 18–14 C Ka BP (Yokoyama et al. 2000). It has also been assigned
Figure 2  Distribution of glaciers and sea ice in the (a) Southern and (b) Northern Hemispheres during the Last Glacial Maximum.

Chronozone status (23–19 or 24–18 Ka cal BP dependent on the dating applied), the event being centered on the calibrated date at 21 Ka cal BP (i.e., LGM *sensu stricto*). However, since the last maximum glaciation after MIS 5 occurred much earlier in some areas than in others, the term LGM should be used with caution. For the purposes of this entry it is defined as the interval 27.5–23.3 Ka (= Greenland Stadial 3) (see Hughes, Gibbard, and Ehlers 2013; Hughes and Gibbard 2015).

During the LGM, the extent of the glaciation of the Southern Hemisphere differed very little from that of the Pleistocene glacial maximum. Glaciers in Antarctica still reached to the shelf edge and on New Zealand, Tasmania, and in South America the glacier tongues were only slightly smaller than during earlier events. On mainland Australia, local mountain glaciation occurred. It seems that the LGM in the Southern Hemisphere began earlier than in the Northern Hemisphere, probably around 27 Ka. The high mountains of East Africa were glaciated. There is no unequivocal evidence of glaciation in South Africa, although minor glaciers have been postulated by various authors. However, it must be borne in mind that South Africa is located relatively closely to the equator to the north. Were it in the Northern Hemisphere, Cape Town would be situated at the same latitude as Atlanta in Georgia, USA or, if placed relative to the European ice sheet, it would be south of Tunis. Consequently, sea ice cover did not reach the southern end of Africa (Figure 2a).

Likewise, during the LGM, ice in many parts of the Northern Hemisphere reached an extent very similar to the Quaternary glacial maximum. In North America, the differences are very small. Again, most parts of Canada were ice covered, including the shelf areas. It is the same in Greenland. The map of global ice distribution (Figure 2) shows that while sea-ice cover went very far south, most parts of the land areas of Beringia remained ice free.
In Europe, however, the situation was different. New evidence suggests that the North Sea was not fully glaciated during the Weichselian glacial maximum but slightly earlier. An ice sheet covered the Barents Sea and extended well into the Kara Sea but hardly touched the Russian mainland (Svendsen et al. 2004). The ice sheet was markedly smaller than during the Quaternary glacial maximum. It seems that the northwards drainage of the Ob and Yenisei rivers was not impeded. Not only has the glaciation of the Alps been mapped in great detail (Figure 3), but also knowledge of other mountain glaciations has increased considerably. In glaciated mountain areas outside the major ice sheets, such as in Italy and Greece, the maximum Middle Pleistocene glaciations were much bigger than the local last glacial maxima. This is attributed to a change in equilibrium-line altitude (ELA) because it has a much bigger impact on glacier size in areas characterized only by mountain glaciation than in areas where ice covered the lowlands during multiple glaciations.

There were major ice sheets or mountain glacier systems in the Siberian mountains further to the east. Because of the lack of detailed investigations to date, in most cases it is not possible to differentiate between the extent of the LGM and earlier glaciations. Likewise, the age and extent of the glaciers in Iran and the mountain ranges and high plateaus further to the east, and especially the extent of glaciation on
the Tibetan plateau, are still matters of debate. In contrast, the small mountain glaciers of Japan are well mapped and dated and, unlike those of the adjoining Asian mainland, the maximum ice advance of the last glaciation occurred in MIS 4. The same date for the maximum extent is found in New Zealand.

**SEE ALSO:** Climate change and land ice; Climate change and permafrost; Climate change and sea ice; Glacial depositional processes and landforms; Glaciations; Glacier changes; Ice sheets; Oceans and climate; Quaternary science

References


Quaternary science

Xiaoping Yang
Zhejiang University, China

Quaternary science is the multidisciplinary studies focusing on the Quaternary period, which encompasses the last 2.6 million years according to the chronology defined in 2009 by the International Commission on Stratigraphy (ICS). The Quaternary period is further divided into Pleistocene and Holocene (Figure 1). While the Quaternary science focused on describing and interpreting geomorphic features and glacial landforms on the Earth surface earlier, it now largely examines all kinds of proxy evidence to reconstruct the past environments during this period in order to decipher knowledge about the climatic and environmental changes and to assess triggering factors of such changes.

From a geological perspective, Quaternary refers to the youngest and shortest system of the Earth’s history. The term was adopted for the first time by the Italian mining engineer Giovanni Arduino (1714–1795) when he distinguished four orders of strata – Primary, Secondary, Tertiary, and Quaternary. The first two terms were discarded long ago, the last two are still used although Tertiary has been out of the 2004 updated geological timescale by ICS. Quaternary was mistakenly seen as an outmoded relic also while the ICS deleted the Quaternary and the Tertiary out of the ICS timescale altogether in 2004. Then the Quaternary communities around the globe went into open revolt. After open discussions and formal proposals from the Quaternary communities, two rounds of voting took place at the ICS in 2009 and the proposal passed. The Quaternary is redefined and its beginning is placed at 2.6 million years ago, it was defined in 1985 to cover the last 1.8 million years. Different from other periods of geological histories, the Quaternary is the time period in which humans evolved and live today. During this period the bipedal, toolmaking, fire-use hominids emerged from Africa and gradually moved out to occupy Eurasia, Australia, and the Americas, as well as the oceanic islands throughout the planet. Thus, the Quaternary is not only part of the Earth’s history, but also the time during which humans developed and the most active part of the Earth system. Humans now live in the recent epoch (Holocene) of the Quaternary period of the Cenozoic era (meaning “new life”). The famous US geologist Grove Karl Gilbert commented already in 1890 on the importance of the youngest period: “When the work of the geologist is finished and the final comprehensive report is written, the longest and most important chapter will be upon the latest and shortest of the geological periods.”

Quaternary science studies not only all aspects of this geological period but also the connections and interactions between Quaternary and modern as well as future times. It is the field where geography interacts most extensively and intensively with geology. In the United States, Quaternary science is often heavily affiliated in geological departments and geological institutions, but it is based strongly in geographical schools in Western Europe where this subdiscipline was originally initiated. At a time of widespread concern over the negative human impact on our environment, Quaternary science can offer key options and concepts for reconstructing the
history of past environmental changes and for drawing some relevant lessons from research on former human interactions with their natural environments (Eitel et al. 2005). The attempt to predict possible future climate fluctuations, possibly triggered by anthropogenic increase in the atmospheric concentration of carbon dioxide, methane, and other greenhouse-enhancing gases, has considerably invigorated the Quaternary science in the last several decades, since the scientific communities assume that past climatic events could be used as analogs for possible future climatic changes (Williams et al. 1998). Quaternary science, earth system science, and sustainability science are suggested to be the three major interacting and mutually reinforcing subfields of geosciences in the early twenty-first century. Among the three subfields, Quaternary science is considered to be the most firmly established (Matthews et al. 2012).

Part of Quaternary science is about studies of clearly visible landscape changes and their mechanisms. In such areas it has a strong interaction with geomorphology and sedimentology, since the landforms and their constitutions are the main research materials. Glacial theory has been an important development of Quaternary science. The glacial system is controlled by two major climatic parameters: the input of snowfall (accumulation) and the loss of snow and ice (ablation) resulting from evaporation or melting. The boundary line along which summer melting is equal to all of the preceding winter’s snowfall is called snowline. Glacial ice covers about 10% of Earth’s land area at the present time and alpine glaciers can be found in mountainous regions on most continents. The Pleistocene Epoch was the time of recent geologic history that experienced several major advances and retreats of continental ice sheets over large portions of the land (Figure 2). With the improvement of the physical dating methods of glacial sediments, the precise estimate of the events of glacial advance and retreat, as well as changes of snowlines is now potentially possible. The so-called surface exposure dating has been successfully used to date glacial moraines and has helped to establish precisely the timing of glacial advances and retreats. The surface exposure dating is based on the measurement of abundance of accumulated nuclides in certain minerals at the surface through cosmic ray activity. The more recent dating results, partly via the surface exposure method, shows that fluctuations in mountain glaciers are not always synchronous in different regions owing to differing regional patterns of rainfall and temperature response to global-scale climate variability. At present, most of the alpine glaciers on Earth are receding, possibly owing to global warming.

Ice moved into the northern United States after 26 000 $^{14}$C years BP (radiocarbon dating, before present), yet the extent in the sense of the last glacial maximum was also reached at different times in different places in the United States (Mickelson et al. 1983). Lobes in the Great

---

**Figure 1** The updated definition of the Quaternary, as approved by the International Commission on Stratigraphy in 2009.
Lakes and New England regions reached their maximum extension before 21000 $^{14}$C years BP, but lobes to the west of the Great Lakes reached their maximum extent at about 14000 $^{14}$C years BP, out of phase with the rest of the ice sheet margin. Rapid retreat of the ice sheet began after 14500 $^{14}$C years BP, but the re-advance of lobes are found all along the southern Laurentide ice sheet margin at 13000 $^{14}$C years BP, 11800 $^{14}$C years BP, and 9800 $^{14}$C years BP. After 9800 $^{14}$C years BP ice retreated out of the northern United States (Mickelson et al. 1983).

Another important field of Quaternary science is to investigate the Quaternary environmental changes in the world’s arid and semiarid regions, because arid and semiarid regions occupy about one-third of the global land surface, occurring in all continents. While the formation of ice and glaciers is considerably controlled by temperature, the occurrence of deserts is much more dependent on the aridity or the ratio between precipitation and evaporation. Many imprints of climate change in arid zones are large scale and easily visible, in particular the desiccation of lakes and rivers, shifting of sand seas’ boundaries, as well as the stabilization or reworking of sand dunes. Sedimentary sequences occurring in desert dunes reflect changes in desert systems and as such may contain signals useful also for recognizing spatial and temporal changes of deserts and their response to regional or even global climate fluctuations. One of the earliest
recognitions of Quaternary climatic instability was from the geomorphological field studies in arid regions of western China (*China: Ergebnisse eigener Reisen und darauf gegründeter Studien* (China: The Results of My Travels and the Studies Based Thereon) by Ferdinand von Richthofen in 1877). This work even pre-dates the landmark publications on Quaternary glaciations in the Alps of Europe (*Die Alpen im Eiszeitalter* (The Alps in the Ice Age) by Albrecht Penck and Eduard Brückner in 1909). In more recent years, a large amount of new information on the nature, timing, and extent of past climatic changes in arid and semiarid regions has emerged. Optically stimulated luminescence (OSL) dating enables age control for aeolian deposits that were previously hard or impossible to be dated directly. Since the temporal and spatial density of chronological datasets on past environments in deserts has increased, the spatial variations of the patterns of dune activities has become increasingly evident, requiring revisions and reinterpretation of the regional and global correlations of aeolian processes (Lancaster, Yang, and Thomas 2013).

Multiple Quaternary periods of both increase and decrease in the level of aridity have been found in all arid and semiarid regions on Earth. During the Pleistocene glacial period linear dunes of the western Sahara extended onto the continental shelf, accompanied by dramatic shrinkage of lakes in the region, southward shifting of sand dunes, and large-scale contraction of Guinean forest in West Africa. At the same time, the dryness level increased in South Africa and Australia as well as in parts of the Americas due to a decrease in precipitation. The glacial increase in aridity, however, was much less clear and inconsistent in the drylands of Asia. Lacustrine records in western China suggest that an increase in effective moisture at around 20,000 to 18,000 years BP was likely due to an intensification of mid-latitude westerly winds bringing moisture from the Atlantic and an expansion of glaciers in the surrounding mountains. Between 35,000 and 25,000 years BP, the world’s largest sand sea, Rub Al Khali, experienced moister grassland to semi-desert conditions.

Even during the Holocene, at least one distinctly wetter period has occurred in every part of the arid and semiarid regions of the world, although the precise timing of this epoch did not occur synchronously globally nor regionally. In contrast to the earlier assumptions of the extreme longevity of the deserts in northern China, more recent studies of the physical and biochemical indicators in the sediments and their OSL chronology show that most of the current dunes in the mid-latitudes of eastern Asia are of Late Pleistocene or even Holocene age and the dune system has responded sensitively to climate change. During the middle Holocene a large-scale stabilization of the dunes occurred in the eastern part of the desert belt in northern China. In more arid parts of the region, however, the increased rate of annual precipitation was not sufficient to fully stabilize dunes, thus some active dunes persisted even during this long-lasting wetter epoch (Figure 3). In general, Holocene dune stabilization due to paleosol formation varied along the climate gradients across the various sandy lands of northern China and in general it began earlier and lasted longer in the east than in the west (Yang et al. 2013).

Although paleoclimatic records can provide an analog for assessing the impacts of global warming on environment, climatic parameters in the past and potentially for the future need to be carefully compared to make a realistic forecast. Yang *et al.* (2013) estimated that the annual precipitation of the Hunshandake Sandy Land in northern China (Figure 3) during 9600–3000 years BP could be 30–140 mm more
Figure 3  Comparison of the dune fields in northern China during the mid-Holocene (a) and at the present (b), showing the obvious changes of dune landscapes during the Holocene. Individual deserts (sand seas) and sandy lands (dunes partly stabilized by vegetation): Key: 1 Taklamakan; 2 Gurbantunggut; 3 Kumtag; 4 Chaidamu; 5 Badain Jaran; 6 Tengger; 7 Wulanbuhe; 8 Kubuqi; 9 Maowusu; 10 Hunshandake; 11 Horqin; 12 Hulunbeier. Modified from Yang et al. (2013).
than at present. Driven by the higher insolation in the summer, this period was characterized by higher temperatures globally. The warming in this century in the eastern part of the desert belt differs from the globally warmer climate during the early and middle Holocene with respect to magnitude and seasonality of temperature changes, both of which would not be favorable to the vegetation in the entire region in the future and not the same as during the middle Holocene warmer period (Li and Yang 2014).

Quaternary scientists around the world have been reconstructing changes in flora and fauna in terrestrial areas for more than a century. Stratigraphic analysis of macrofossils and pollen of plants in terrestrial, lacustrine, and marine sediments and the sedimentary features of various deposits has enabled the reconstruction of vegetation history and the interpretation of climatic fluctuations and human impacts on environment throughout the world. The studies of ice cores, speleothems, and tree rings have greatly improved the temporal resolutions for Quaternary paleoenvironmental reconstructions. Results of the studies of ice cores from the Greenland and Antarctica ice sheets produced continuous paleoenvironmental records for the last several glacial–interglacial cycles.

Studies of marine and coastal environmental changes are other important areas of Quaternary science. Various international efforts, including the Deep Sea Drilling Project (DSDP), Ocean Drilling Project (ODP), and Integrated Ocean Drilling Project (IODP) have studied the deep sea and continental margin sediment cores that have shed light on ocean basin history, changes of Quaternary climate, and ocean circulation. The interpretation of marine and speleothem records and the numerical simulations of Quaternary climate changes have also helped the Quaternary communities to reconsider the astronomical theory by the Serbian mathematician Milutin Milankovitch who used the potential changes of orbital parameters of Earth (precession, obliquity, and eccentricity) to explain the periodic alteration of glacial and interglacial climates in 1941. There have been also significant glacial–interglacial changes in sea level during the Quaternary, for example, the global sea level has risen by circa 150 m from the last glacial maximum to present. During the last glacial maximum, roughly 5.5% of the water on Earth was locked up in the form of ice (the corresponding number today is 1.7%). When current sea level rise is a great concern for human wellbeing in coastal areas, a complete understanding of natural cycles of sea level change as well as anthropogenic effects is imperative for future development in the coastal areas and beyond (Murray-Wallace and Woodroffe 2014).

While Quaternary science progresses are presented in various geological, geographical, geomorphological, climatological, and anthropological meetings and events, the Quaternary Geology and Geomorphology Division (QG & Q) of the Geological Society of America (GSA) organizes sessions and field trips every year at the GSA Annual Meeting that has become a large annual gathering for Quaternarists all over the globe. In more recent years, the organization PAGES (Past Global Changes, based in Bern in Switzerland) has also become an international coordinator for studying past global changes, one of the key areas of Quaternary science nowadays. The large number of Quaternary publications appearing each year and the continuing high impact factors of specific Quaternary journals (e.g., *Quaternary Science Reviews*, *Quaternary Research*, *Quaternary International*) also confirm the vital status of Quaternary science in a global scientific context.
SEE ALSO: Climate change and health; Geoarchaeology; Geography and the study of human–environment relations; Geomorphology: history; Historical geography; Quaternary geomorphology and landscapes

References


Further reading


The Quaternary is the most recent period in the Cenozoic era comprising the Pleistocene and Holocene epochs. The Pleistocene epoch extends from the beginning of the Quaternary 2.6 million years ago to the beginning of the Holocene epoch 12000 years ago. The most recent 12000 years make up the Holocene, which includes the present-day interglacial period. One of the most distinctive features of the Quaternary period is the expansion of ice sheets and mountain ice caps in many parts of the world.

During the Quaternary, Earth’s climate has shifted back and forth between glacial and interglacial states. The cyclical nature of the shift between these states is often attributed to the astronomical Milankovitch cycles of eccentricity, obliquity, and precession. Eccentricity refers to the change in Earth’s orbital path around the sun, which has a periodicity of about 100 000 years. Obliquity, or axial tilt, refers to the change in Earth’s axial incline. Over a time period of about 41 000 years, Earth tilts on its axis from 21.5° to 24.5° relative to the equator. Today, Earth’s axial tilt is about 23.5°. Precession, or Earth’s “wobble” on its axis, has a periodicity of about 23 000 years.

The variation in the Milankovitch cycles affects the seasonality and impact of solar radiation on the planet’s climate system. These shifts largely account for the swing between glacial and interglacial periods. While the astronomical cycles play a large role in the behavior of Earth’s climate, they do not exist in a vacuum. Other factors, such as plate tectonics, have an effect on the movement of glacial ice from the poles and also impact Earth’s overall albedo, or solar reflectivity. The chain of relationships continues in the alteration of ocean currents and global atmospheric composition.

While there is strong evidence that the Milankovitch cycles play a role in regulating factors that drive glacial and interglacial cycles, it is unlikely that these cycles alone drive the beginning or end of an ice age, or series of glacial periods resulting from a long-term decrease in Earth’s overall surface temperature. In fact, the ice ages observed through the geological record happen on an irregular basis, which contradicts Milankovitch patterns. This means that additional factors are necessary in exploring the causes of ice ages as well as shifts between glacial and interglacial cycles.

**Landforms**

During glacial periods, the mid to high-latitudes of the Earth are covered by large ice sheets extending from the poles. Interglacial periods, such as the one we have been experiencing since the beginning of the Holocene, witness glacial recession and an overall increase in Earth’s temperature. The Holocene is characterized by the disappearance of the ice sheets that expanded during the Pleistocene. The most recent glaciation is known as the Wisconsin glaciation, which was at its maximum about 20 000 years ago and drove the formation of many landforms observable in the current day. Residual glaciers still exist in Antarctica, Greenland, the high
Canadian Arctic, Iceland, and some high alpine environments.

As glaciers grow and recede from the poles, sea levels pulse, currents shift, and winds change direction. The combination of these responses affects global precipitation patterns and acidity levels, factors that play a large role in weathering and erosion patterns. These factors drove the evolution of topography on a global scale. Landforms, or features that are the result of the surficial processes impacting topography, arose during the Quaternary, which allowed for the expansion of flora and fauna species.

Geomorphology is the study of the processes that shape Earth’s landforms and the combined landforms in an area, or landscapes. Geomorphologists, or scientists who study these processes, investigate changes in Earth’s landscape over time. Thus, Quaternary geomorphology is the study of landscape changes during the recent Pleistocene and Holocene epochs. While geomorphologists typically study landforms with a historical perspective in mind, Quaternary geologists focus on the surficial sedimentary cover as well as deeper layers of sedimentary deposits, which provide a window into the past of Earth’s changing climate.

Paleoclimatology is a field concerned with the study of past climate system activity. Paleoecologists develop a timeline of Earth’s geological past through proxies, which help to tell the story of Earth’s history, including the Quaternary period. Both land and ocean features provide a geological record of the Quaternary period.

The only direct record of Earth’s atmospheric history exists in the form of ice cores. The most recent of ice core investigations covers a history of nearly 1 million years, and serves as a calibration tool for other geological records (see Figure 1).

Quaternary sediments make up the surficial layer of land, marine, and shoreline environments and are particularly useful in reconstructing paleoclimates. The coring of lacustrine (lake), fluvial (river), and marine sediments allows for paleoclimate reconstruction using stable isotope analysis. Other paleoclimate proxies include the study of dendrochronology (tree rings), foraminifera (planktonic organisms), fossils (plant and animal), pollen, and speleothems.

The Quaternary period left behind distinctive landforms in the form of glacial erosion during glacial advance and glacial deposition during glacial retreat. The pulse of ice sheets from the poles during Quaternary glacial periods left evidence of climate shifts through a variety of landforms while also setting the stage for mountain glaciers to build in high alpine environments. Glacial features are the primary characteristic of high alpine environments such as the Rocky Mountains, the Alps, the Cascade Mountains, the Himalayas, the Sierra Nevada Mountains, the Andes, and the Tibetan Plateau.

As a glacier flows, it scours the environment with collected debris through erosion and deposition. This process of abrasion repeatedly shaped the land of the Quaternary. Landforms that
exist in the current day are those carved by the most recent glacial activity. Evidence for the timeline of Earth’s most recent glaciation exists in features such as moraines, drumlins, kettles, kames, outwash plains, V and U-shaped valleys, and arêtes.

The creation of glacial features allows for the study of glacial history. The features that arise from glacial advance and retreat provide a window into the past movement of glacial ice, which plays a large role in paleoclimate reconstruction. The aggregated information from paleoclimate proxies and physical features allows scientists to model past, present, and future glacial activity.

Glacial activity from the Quaternary also left behind evidence of past geological activity through the disruption of the surficial drainage system and the resulting creation of glacial lakes. As glaciers moved across Earth’s surface, they eroded the bedrock below, creating new features that would remain long after glacial retreat. As glaciers began to recede, meltwater discharge filled depressions in the landscape, which produced a notable number of lakes in previously glaciated landscapes (see Figure 2).

Periglacial activity

The oscillation between glacial and interglacial periods during the Quaternary directly conditioned periglacial environments through melting and refreezing processes, often on the margins of glacial areas. These environments are characterized by frost action, which modifies the ground surface with the movement of water through the area and the redistribution of glacial sediments. Today, periglacial processes dominate more than one third of Earth’s land surface.

While the melting and refreezing processes are the signature mark of periglacial environments, these areas are also associated with perennially frozen ground. When the ground is frozen for two or more years in a row, the frozen soil, rock, and sediment is called permafrost. (Other definitions of permafrost include ice that is frozen throughout the year or on an annual basis.) While many periglacial areas contain permafrost, others are dominated by the freezing and melting processes, absent of permafrost.

Location is one of the most important factors affecting the extent of periglacial environments, which are concentrated in high latitudes. The distribution of Earth’s land masses is such that there is more land in the high latitudes of the northern hemisphere than in the high latitudes of the southern hemisphere – this means that a larger portion of the Earth’s periglacial regions exists in the north than in the south.

While periglacial regions are more common in high latitudes, lower latitudes can see the existence of periglacial regions in areas of high elevation, such as in mountainous regions. Finally, location relative to the coast plays a large role in the freeze–thaw processes that dominate periglacial regions. Areas subject to polar surface currents near a coast can support periglacial activity. Additionally, inland areas without a nearby water body to moderate the temperature can experience the seasonal freezing and thawing characteristic of periglacial regions.

The result of climate variation during the Quaternary period was an assortment of ground ice features, some of which were preserved in the fossil record. During the cycle of freeze–thaw, the ground contracts as temperatures drop, causing cracks to form. As temperatures rise, the ground expands and moisture builds in the cracks. When this water freezes, it fills the cracks, creating an ice wedge. When temperatures rise again, soil, rock, and sediment collect and allow for the fossilization of these formations (see Figure 3).

The extent of periglacial environments varies with the Earth’s climate. Changes in the climate
Figure 2 Kettle Lakes of the Turtle Mountains, 2015. About 25 kya, ice sheets expanded southward and covered the Turtle Mountains. As the climate warmed between 15 and 13 kya, the glaciers stopped flowing. Over time, a debris layer formed on top of the stagnated glaciers, which insulated the ice. This process resulted in the present-day landscape of the Turtle Mountains. Photo courtesy of NASA/GSFC/METI/ERSDAC/JAROS and US/JAPAN ASTER Science Team, August 16, 2015.

during the Quaternary period affected periglacial regions by impacting the hydrological cycle, sedimentation, atmospheric composition, and vegetation.

Vegetation

The oscillation between glacial and interglacial periods during the Quaternary affected vegetation distribution worldwide. The paleovegetational record combined with present-day species distribution provides an indirect history of global climate change. While the spread and abundance of plant species track climate shifts, the present-day variety of plant species tells the story of Quaternary vegetation movement.

The shifts between glacial and interglacial cycles during the Quaternary impacted atmospheric and oceanic circulation patterns, which shaped global plant distributions across the planet. The changes in vegetation that occurred during this time played a part in a feedback loop between climate variation and plant life. While
climate influenced vegetation, this feedback loop meant that changes in vegetation also impacted the climate. Plant species across the world influenced the global energy cycle by affecting the planet’s *albedo*.

The changes in Earth’s albedo amplified temperature swings worldwide, creating additional feedback loops that shaped the Quaternary period. As ice spread from the poles during this time, the overall abundance of vegetation at high latitudes decreased, leaving these areas with a higher albedo, which amplified cooling. Evidence of this influence exists in the latitudinal organization of plant species in the modern day. Furthermore, changes in evapotranspiration, hydrological, and carbon cycles drove feedback loops that defined the biosphere of the Quaternary.

**Climate change**

While evidence for glacial activity comes in a variety of forms, subsequent glacializations erase the history of prior ice extent, complicating the interpretation of the geological record. Despite the difficulty of teasing apart the signal and noise within Earth’s climate record, analyses of
the Quaternary time frame show that the shift between glacial and interglacial periods is related to a variety of factors.

Today, atmospheric temperatures are warmer than they have been throughout 90% of the Holocene (Marcott et al. 2013). Although high temperatures are unsettling, the rate of warming that characterizes the twentieth century is particularly concerning – temperatures have risen faster in the most recent century of our history than they have in the last 2000 years. Scientists have learned that the most recent changes in Earth’s climate system, in the time from the 1850s to the current day, are largely attributed to human influences on the global system. In fact, anthropogenic, or human-caused, influences on the climate system are the driving factor of global warming in the present day (see Figure 4).
SEE ALSO: Glacial depositional processes and landforms; Glacial erosional processes and landforms; Glaciations; Glacier changes; Glaciers; Mountain geomorphology

Reference


Further reading


Queer geographies

Lynda Johnston
University of Waikato, New Zealand

Queer theory and queer geographies are difficult to define. The concept of queer is often associated with identities such as lesbian, gay, bisexual, transgender, and intersex (LGBTI). Yet, queer is also a perspective, or a way of thinking and knowing. It calls attention to taken-for-granted assumptions regarding bodies, genders, sex, and sexualities. One of the main contributions queer theory has made to studies of sexualities is a critique of sexual identity politics. Queer theory sits precariously with constructions of queer identities, places, and spaces. The word “queer” is not fixed, and its meaning is socially produced through language. What it means to be queer, and what can be queered, have multiple meanings and are subject to change over place, space, and time.

Studies of sexualities are heavily influenced by postmodernism, poststructuralism, and anti-essentialism. Social theorists, including geographers, employ different understandings of bodies, places, and spaces in order to question grand claims and single truths of the modernist era. A postmodern approach recognizes that all knowledge is partial, fluid, often queer, and contradictory. There is recognition of difference and multiple truths, realities, voices, and identities. Queer theory – viewed through a poststructural theoretical lens – challenges ideas of preconstituted sexual subjects and allows for the examination of power as productive rather than as merely oppressive. This approach moves beyond humanist understandings of essentialized, gendered, and sexual subjectivities. It has been argued that queer theory is anti-normative – therefore anti-essentialist – and can be employed to critique the normalization of bodies, spaces, and places. Any attempt to reduce, mainstream, or even define and summarize queer theories is potentially anti-queer.

How have critical geographers used queer theory? There has been a tendency to claim queer spaces as existing in opposition to—and in transgression of—heteronormative spaces and places. Heteronormativity, as discussed later in the entry, is the normalization of heterosexual so that it appears natural. Since the mid-1990s geographers have been adopting the concept of queer space to illuminate and create spaces of sexual dissidents (Bell and Valentine 1995). Some geographers have gone beyond a sexual politics of recognition and produced queer geographies that engage with other theoretical frameworks, such as postcolonial, feminist, and critical race theories. Such accounts bring to the fore questions of race, colonialism, migration, globalization, geopolitics, and nationalism. Queer geographies, therefore, have expanded beyond the confines of sexual identity politics.

Queer activism to queer theory

The term “queer” was, in the 1970s and 1980s, slang for “homosexual” and was often associated with homophobic abuse. In the early 1990s the term began to perform different identity work and took on other meanings. Notoriously indefinable, queer became in the 1990s a key word for radical theorists and activists whose
sexual politics began with reclaiming the very word “queer” from its use as homophobic slang to be a positive label used by a variety of sexual and gender minorities. At this time, being queer meant being countercultural. It was associated with the rejection of mainstream, capitalist profit margins, patriotism, patriarchy, and assimilation. Being queer was anti-elitist and about being on the margins.

Queer politics gained traction at a particular historical juncture, when several Western governments failed to respond adequately to the HIV/AIDS crisis. It was a time when numerous pieces of antigay legislation were put forward by, for example, the homophobic Thatcher government in the United Kingdom. Queer activists in North America formed groups such as Queer Nation and ACT UP. They were visible in pride parades like the 1991 Montreal pride parade. Here queer activists rejected the “official” route that ran only through the city’s gay district. A march through Montreal’s gay neighborhoods would have, they argued, been affirming and empowering of Montreal’s sexual dissidents without challenging or confronting the city’s heteronormative culture. A similar queer struggle occurred over the Auckland, New Zealand gay pride parade when it was moved from the main street, downtown, to a “gay” suburb.

Current understanding of queer theory has evolved from its US origins in literary theory and humanities. Adrienne Rich’s (1980) early intervention argued that homosexuality should not be understood as misplaced (as sexologists were advocating at the time) but, rather, the focus should be on heterosexuality. Rich moved attention away from the marginalized “other” to heterosexuality, which she argued is a “compulsory fiction” designed to perpetrate patriarchal ideas that women are biologically and innately sexually oriented only toward men. This fiction, she claimed, was maintained via state policies and practices that valorized heterosexuals while leaving lesbians and gays invisible. These ideas were further refined by other queer theorists who also recognized the inadequacy of existing theoretical frameworks.

Queer politics are closely linked to queer theory, a diverse literature that is often seen as having emerged out of gay and lesbian literary criticism, particularly in the work of Eve Kosofsky Sedgwick – for example, her Epistemology of the Closet (1990) – who decentered normative understandings of sexualities. Sedgwick claimed that “the closet” had shaped gay life, beginning in the twentieth century and continuing. The “epistemology of the closet” is another way of saying that homosexuality exists through regimes of an open secret: that is, something that is known through not knowing. This scholarship is most commonly situated within postmodernist and/or poststructuralist frameworks that decenter normative notions of sexuality. Humanity scholars, therefore, forged new ways of conducting cultural studies with a focus on fragmented subjectivities.

“Queer theory” is a term originally coined by Teresa de Lauretis that first appeared in her introduction to the special issue of differences called “Queer Theory: Lesbian and Gay Sexualities” (de Lauretis 1991). She traces the way the conceptual framework for analyzing sexualities shifted in the twentieth century from an early emphasis on “homosexuality” to a gender-undifferentiated “gay,” and more recently to “lesbian and gay.” Concerned with the maintenance of differences, she remarks that specificity is elided “in the contexts in which the phrase [‘gay and lesbian’] is used, that is to say, differences are implied [in the phrase] but then simply taken for granted and even covered over by the word ‘and’” (de Lauretis 1991, v–vi). The turn to the word “queer”
is meant to mark “difference between and within lesbians, and between and within gay men, in relation to race and its attendant differences of class or ethnic culture, generational, geographical, and socio-political location” (de Lauretis 1991, viii). De Lauretis questioned the normative stance of heterosexuality against which homosexuality is defined as marginal. There has been a productive – and sometimes fraught – relationship between feminist theorizing and queer theorizing.

Queer theorists understand heterosexuality and homosexuality not simply as subjectivities but as categories of knowledge (Jagose 1996). Queer theories challenge the supposed naturalness of sexual and gendered hierarchies and disentangle any established understandings of gender, sex, and sexuality. To reiterate, it is language that frames constructions of bodies, desires, sexualities, and identities. Queer theorists highlight that normative language shapes moral boundaries and political hierarchies. Queer theoretical frameworks are employed to study homosexuality – a sexual minority – but also to study those knowledges and social practices that organize society as a whole by sexualizing – heterosexually or homosexually – bodies, desires, acts, subjectivities, relationships, knowledges, cultures, and institutions. Queer theory provides a framework from which to draw on critical social theories that challenge heteronormative discourses. Put another way, the task of queer theorists is to critique notions of identity to their fullest extent rather than only searching for radical queer subjects. This task focuses attention on advancing a critical approach to the workings of gendered and sexual normativities and non-normativities.

Judith Butler (1990) has also contested the naturalization of sexual and gender categories, to undo and destabilize binaries such as heterosexual/homosexual, gender/sex, male/female, and mind/body. Queer theorists are committed to the recognition of fragmented power across and between these binaries, pushing their scholarship beyond binary categorizations of social construction versus essentialism that conceptualizes bodies, subjectivities, and desires as fixed. Instead, Butler argues that heterosexuality is an ideal that can only be approximated and performed. It is within the heterosexual matrix that the binary opposition of man/woman finds its cultural and social currency. The gendered norm of male or female is solidified in the heterosexual matrix. Butler argues that this matrix is a masculine sexual economy in which bodies become gendered in order to support a gender hierarchy deemed necessary for compulsory heterosexuality. Her theorizing of repeated stylizations of bodies shows how, over time, bodies produce the appearance of that which is deemed natural and normal. The performativity of gender is described as “the stylization of the body, a set of repeated acts within a highly rigid regulatory frame that congeals over time to produce the appearance of substance” (Butler 1990, 33). By repeatedly conforming to gendered norms and established ideas about sex and sexuality – in dress, language, and everyday behaviors – embodied performances solidify hegemonic heterosexual identities so that they appear normal and seemingly natural.

While some theorists contend that heterosexuality came into being after the delimiting and othering of homosexuality, Butler argues homosexuality is indeed the starting point for heterosexuality. Heterosexuality was, and continues to be, constructed through its otherness to homosexuality. Such queering of heterosexuality also challenges the fixity of categories such as male and female. Butler illustrates the fiction of gendered and sexed embodiment by discussing drag queens and drag kings. Drag queens, for example, are doing femininity in, with, and from male bodies, and this mismatch exposes
the sexed body as the base for gender as a fiction (Butler 1990). Butler’s theory of performativity, however, has been critiqued, and she acknowledges that there is a difference between performance – as an occasional drag act – and performativity. Bodies, spaces, and identities are re-created through performativity, which is the process of repeated behavior and acts. Bodies are disciplined and (self) policed through dominant norms and understood codes of behavior.

The political imperative of a queer critique is to understand how norms and categories are deployed. Butler (2004, 15) notes: “If I am always constituted by norms that are not of my making, then I have to understand the ways that constitution takes place.” Butler’s work has been crucial in the development of queer geographies.

Queer geographies

Queer geographies question the relationships between gender, sex, and sexuality and assert that multiple genders and sexualities exist. At the core of this scholarship is the understanding of sexuality as a social process that allows for expressions of desire (in all directions and beyond dichotomizing discourses). Queer geographers have become dissatisfied with essentializing, limiting, neoliberal, and often Westernized sexual categories. They have turned to fluid, unstable, and ambiguous conceptualizations of sexualities. Spaces and places, like sexualities, are also considered fluid and ambiguous.

Scholarship on sexuality, place, and space started to appear in the 1970s. It was not until the 1980s – when humanistic approaches gained prominence over positivist approaches – that more sexuality and space studies emerged. This period of scholarship is perhaps best known for its emphasis on structuralism, such as theories of Marxism. The intersections of class and gay identity dominated studies of, for example, gay men and gentrification in US cities. There is a move away from reified notions, such as “gay gentrification” and “gay neighborhoods” (which also tend to be celebrated) to understanding bodies, places, and spaces as fluid, ambiguous, and often contradictory.

The first lesbian geographical scholarship appeared in the early 1990s and, in retrospect, this research may be considered “queer geography.” While not explicitly labeled as such, the research attempted to understand why lesbian communities had a quasi-underground character, thus recognizing difference, not sameness. A groundbreaking article entitled “All Hyped Up and No Place to Go” (Bell et al. 1994) also queered geography by considering the ways in which sexual identities – lipstick lesbians and gay skinheads – were performative parodies of hyperfemininity and hypermasculinity. In the mid-1990s the influence of poststructural thinking provoked an array of new research in this field, much of which can be found in a collection that mapped desires and sexualized landscapes (Bell and Valentine 1995). Notably, these studies show the desire to draw on feminist, poststructural, and queer theories in order to examine difference at a variety of spatial scales. This work challenges the notion of fixed identities and spaces by focusing on the proliferation of identities, rather than structural inequalities. Bell and Valentine (1995, 4) introduce their edited collection by noting:

The women’s, gay and civil rights movements emerged in North America and Europe in the 1960s and 1970s on a wave of social and political upheaval. But despite a growing awareness amongst geographers in the following decade of the need to study the role of class, gender and ethnicity in shaping social, cultural and economic geographies, sexualities were largely left off the geographical map.

While this collection is known more for its empirically rich case studies and less for
its advancement of sexuality theories, it did, however, open up new avenues for scholars to focus on queer spatial theories of desire.

While geographers have been slow to question taken-for-granted sexualities, some have been able to show the ways in which people unconsciously use heterosexuality to structure their lived everyday spaces, from bodies to homes to nations and beyond. This normalizing of heterosexual geographies is also incorporated into the canon of thought, which means that heteronormativity permeates the ways in which geographers research place and space. While some geographers have charted the ways in which heterosexism pervades the institutional practices of geography, others have focused on examining the creation of heterosexist scholarship. The turn to queer geographies is, in part, a reaction against heterosexist geographies but also an opportunity to form the subfield of sexuality and space alongside feminist geography. Queer geographies, in particular, have benefited from feminist geographers' critical engagement with gender and sexuality. Feminist geographers have been quick to question the authority assigned to science and “objective” claims of universal truth. For example, feminist geographers have used queer theories to upset the naturalness and legitimacy of global capitalism. Capitalism is often scripted as violently irrepressible, and as a discourse it gathers strength from heteronormative language and practice. Some geographers employ queer theories in order to deconstruct, and “de-phallic” globalization, by highlighting various points of excess, where inscriptions can be seen as uncontrollable or indeterminate, or as potentially inscribing noncapitalist identities (Gibson-Graham 1996).

The term “heteronormative” is a concept now widely used by geographers. The usage can be traced back to Judith Butler (1990, 2004). Geographers such as Hubbard (2012) have researched cities – their acts of planning, health reforms, and environmental modifications – to understand the ways in which regulations are heteronormative and, in turn, discipline and control diverse sexualities. Heteronormativity can be found within an array of geographies from the body to the globe. Language and images circulate globally and often represent maternal bodies and nuclear families. Geographers, taking up the notion of heteronormativity, argue that repetitive embodied performances of heterosexuality – that is, conforming to dominant gendered ideas of how to dress, talk, behave, and perform their embodied identity – make most heterosexual identities appear normal, even natural. This powerful, yet often invisible, performance means that bodies and spaces are mistaken as inherently heterosexual.

This is not to say that heterosexuality easily conflates with heteronormativity, and queer with homosexuality. Natalie Oswin (2008) offers a critical reading of the deployment of the notion of “queer space” in geography and highlights an alternative queer approach that is inseparable from feminist, materialist, postcolonial, and critical race theories. She notes that there is an important distinction between heterosexuality and heteronomativity, which frequently goes unnoticed and unacknowledged in queer geographies. When “queer” is understood to stand solely for LGBTI it forecloses possibilities of understanding the hierarchies of heterosexual privilege, and other identities that intersect with sexuality.

There has also been a great deal of discussion about homonormativity. Geographers have adopted this concept to show how certain performances of homosexuality incorporate the logic of heteronormativity. In other words, becoming a normative homosexual usually means gaining the privileges associated with neoliberal politics and practices that are aligned
Queer Geographies

with heterosexuality and class privileges of consumption (Ossein 2008). Such practices produce a new “homonormative” identity that gains popularity at the expense of radical performances of homosexuality. For example, the emergence of so-called acceptable and trendy gay public spheres is typically aligned with processes of neoliberal gentrification.

Globally, a new kind of normativity is emerging as countries around the world debate and reform marriage laws so that all people, regardless of gender or sexual orientation, have the right to marry and the right to adopt as a couple, if they so choose. These public debates focus on the ways in which state and religious institutions regulate normative sexualities (both hetero and homo). Geographers are adopting queer theories to argue against the privileging of married coupledom as the defining relationship by which access to social justice and forms of belonging are measured. Some debate that LGBTI rights are part of an assimilationist agenda in which the materiality of homo sex becomes rendered as less perverse, and less “scary.”

Hetero- and homonormativities are continually reproduced, and hence there is always the possibility that the norm will be reproduced in different ways. Many queer geographers examine the multiple ways in which genders and sexualities “fail” to be performed in the “correct” way. The potential of this “failure” is that new ways of queering the norm are brought into being. These “failures” also make obvious the normative frameworks that discipline gender, sex, and sexual acts.

There are many articles on queer geographies sprinkled throughout a variety of social, cultural, and feminist geography literature. This has led to studies of “the closet” (Brown 2000), which is used both as a metaphor and also as a spatial practice of power-knowledge that takes up the notion of spatial scale. The body, city, nation, and globe are spaces in which “the closet” is not always about disempowerment but can also be understood as a setting for creative and transformative sexual, social, cultural, and political resistances to heteronormativity.

Queer theories are applied to a number of different empirical geographical projects, including the pink economy; gay tourism; pride parades; bisexuality; sexuality in rural areas; sexuality and the cities; sexuality and citizenship; the globalization of sexuality; HIV and AIDS; and sexuality and ethnicity. The notion of deviancy and/or sexual marginalization is often the topic of queer geographies and queer methodologies, as well as gender transgressions (Browne, Lim, and Brown 2007).

This blossoming of queer geographies since the mid-2000s means that it is now an area taught in a number of undergraduate courses in countries such as the United Kingdom, United States, Canada, Australia, and New Zealand, and has succeeded in striking up conversations with a range of other subdisciplinary areas such as cultural and social geography, gender studies, migration studies, and tourism. Hence queer geographies is not a discrete intellectual field but one that has emerged through its relationships with other disciplinary areas.

Geographers are also considering a number of queer intersections. Sexuality and space studies have (and have not) intersected with other identities, such as gender, race, class, age, (dis)ability, and religion. There is a building consensus that if studies of sexualities and space are to be truly queer, then they cannot focus solely on sexuality, but that connections with other social identities must be considered. This kind of critique has been leveled at queer theories and queer activism.

So how has sexuality intersected with other identities to produce differently queered spaces? Most scholarship looks at the importance of place in bringing different identities together,
and less research has been done on the convergence of identities across spaces. One of the most researched intersections is sex, gender, and place (Johnston and Longhurst 2010). A number of geographers group together lesbians and gays as single populations, thus highlighting the importance of intersecting gender and sexual identities. Increasingly this scholarship speaks to the significance of lesbians and gay men at a variety of spatial scales, from bodies and homes to the nation.

After many cries for critical scholarship on heterosexuality and space, there are now publications that attend to the intersection of genders and heterosexual identities. Some have focused on the ways in which particular women may be disciplined and controlled in heteronormative and patriarchal ways, and attention to heterosexual men is a recent trend. Much of this work seeks to highlight the privileges and oppressions at the intersections of genders and sexualities.

Trans identities have added further complexity to the scholarship on gender and sexualities (Halberstam 2005). Recent studies by geographers have shown how transphobia and sexism are experienced by trans people in homonormative gay and lesbian spaces, and some have suggested alternative places where support and security can be found. This scholarship challenges any notions that gender is stable and, at the same time, illustrates the intersection of gender and sexuality, including the persistent “tyranny” of gendered spaces (Doan 2010). What is clear from this scholarship is the political urgency to continue questioning, challenging, and subverting the binary system of gendered and sexed spaces. Trans geographies is one way to do so.

Queer geographical scholarship may highlight the ways homonormative spaces can privilege whiteness and, as a consequence, are deeply racist. Attention too has turned to the prevailing whiteness of queer scholarship. There have been compelling accounts on American homonormativities after 9/11. Other researchers have shown how homophobic discourses are put into operation in order to seek ethnic and national purity by, for example, reproducing narratives of Irishness (and attempting to exclude LGBTI performances) at St Patrick’s Day parades in New York City and in forming national closets.

Class and sexuality is another fruitful intersection that is receiving more attention in queer geographies. This approach can be summarized into three themes: the classing of epistemology and methodology within sexuality studies; the relationship between culture and the economy in framing research on class and sexuality; and the significance of space, place, and intersectionality. Privilege and oppression can be illuminated when considering the intersections of class positions, gender, and sexed identities. Some scholars delve deeply into class divisions of lesbian identity and spaces. The consideration of classed identities fractures the notion of collective lesbian identities, both emotionally and materially. Others have highlighted the class privileges that align with homonormative gay men and lesbian spaces and identities. The confluence of poor, working-class, and heterosexual has also been considered in Vancouver (Boyd 2010); Butte, Montana (Dando 2009); and Nepal (Richardson, Poudel, and Laurie 2009).

Work on sexuality, space, and age is another intersection of identities that is beginning to gain prominence. There are accounts of the way in which heterosexual patriarchal discourses in Vietnam work to control young women (Rydstrom 2006). Bringing together race, sexuality, age, and education, one study, for example, considers the space of a high school campus for Latina, Armenian, Filipina, African American, and Anglo girls. This research shows that, while the girls use the “everyone just wants to get along” discourse, they also perform racism and
QUEER GEOGRAPHIES

uphold spatial segregation (Thomas 2011). Gay, lesbian, and bisexual youth are the focus of an Australian family home study (Gorman-Murray 2008).

There are important connections between disabilities and queer sexualities. Geographers have studied the way in which non-normative sexualities have been categorized as illness or disease, and some studies highlight the challenges that disabled bodies pose to hetero- and homonormativities. Consequently researchers have started to investigate the association of physical ableism with hegemonic notions of beauty and sexuality. Underpinning this research is a critique of embodied normativities, yet much more work is needed in this area.

Queer trouble

While queer geography has troubled normative understandings of genders and sexualities, it must be remembered that queer theory has its origins in US university humanities programs. This point is important because the set of (Americanized) ideas comes out of a place with a particular history within a so-called developed Western nation. Hence, queer theory not only causes trouble, but needs to be troubled itself. Geographers are well placed to consider the spatiality of queer epistemology. In other words, adopting the word “queer” may be appropriate in some spaces and places but inappropriate in others. Similarly, the use of the term “queer” as synonymous for gay, lesbian, or other identities is troubling because it collapses cultural nuances and eliminates the diversity of sexualities beyond “developed” Western countries.

Another trouble with queer is the way in which identity categories are brought into question. Political organizing and activism often rests on identity categories – such as LGBTI – in order to gain rights and be “equal” with individuals and groups that have societal privileges (e.g., men and heterosexuals). Queer theory calls into questions all categories and normative identities, which in turn undermines the basis of identity politics. The deconstruction of identities such as “lesbian,” “gay,” “woman,” and so on means an erosion of the platform from which to argue for social justice. Queer theorists have been accused of throwing the baby out with the bathwater; when identity categories cannot form the basis of political action, the power of collective action is lost. Queer, for some, may be too fragmenting when basic human rights are on the agenda.

A number of feminists have argued that queer erases gendered differences when considering sex and sexuality. Furthermore, the term “queer” – for lesbian feminists – may mean that women become invisible as it collapses differences between gay men, lesbians, bisexual women, bisexual men, and so on. Others assert that using the broad term “queer,” rather than the gender-specific term “lesbian,” is similar to the use of “man” to refer to humans. In other words, women once again disappear under a universalizing signifier. When “queer” encompasses too much, it is rendered of little use.

There is also concern that the deconstruction of categories such as “woman” means that any empowerment associated with that category is undermined. Feminist scholars argue that categories of women are necessary for collective action and collaboration due to the persistent nature of patriarchy and gendered inequalities. Similarly, queer approaches to lesbian and gay identities may weaken the political power of already disenfranchised groups. Many of these scholars recognize that identity categories are not homogeneous, yet need to be enacted to affect legislative, social, or cultural change. This type of strategic essentialism sits awkwardly beside queer theory, yet many scholars work back and forth between the two theoretical positions.
Some geographers ask if queer theory – with its beginnings in the humanities – has any relevance to “real” places and people. For feminist, lesbian, gay, and trans geographers concerned with power relations and challenging inequalities, queer theory may seem too removed from people’s day-to-day struggles. When politicized identities are deconstructed, fragmented, and fractured, it becomes difficult to critique gendered and sexualized harassment, abuse, and violence. Yet, living struggles accelerate changes in language, and language evolves and reforms into new expressions.

The final trouble with queer theory is that it is often considered elitist and exclusionary because of its focus on language, concepts, and ideas. Many of the concepts are not translated into accessible language. Further, queer theory tends to be employed to critique embodied normalizations, with very little focus on empowerment of those who are marginalized by normalizations. The advancement, however, of concepts, ideas, and constructions of identities, and the dismantling of normalizations are powerful ways in which to critique geography’s hegemonies.

Conclusion

Contestation surrounds queer theory and queer geographies. “Queer” is a highly contestable term which, used too frequently, normalizes identities. “Queer” can be used as a term which collapses a range of sexual and gendered identities; it also contests the formation of identity categories. Feminists and sexuality scholars have used queer theory to contest the naturalization of binaries such as heterosexual/homosexual, male/female. At its core, queer theory is anti-normative and thereby contests all assumptions that gendered and sexual identities are fixed, coherent, or stable. Queer theory is a powerful tool used to question hegemonies and hierarchies within and between sexual subjects. The strength of queer theory is its reach into all normalizing practices, including within LGBTI communities.

Geographers have adopted queer theory in order to challenge the aspatiality of identity categories as well as to question the globalization of rights from the West and their application across different nations. Geographers are well placed to consider the performativity of both place and bodies. Queer theory allows geographers to understand not only how place is normatively gendered and sexualized, but also how place sexualizes and genders bodies. Queering geography further has, for some, meant understanding bodies and places at the intersection of multiple identities.

For geographers who are committed to engagement with disenfranchised and marginalized groups, adopting queer theories may be a controversial choice. When identity categories are called into question, so too are civil rights and the ability to organize collectively to fight against patriarchy, homophobia, and transphobia. This is contested, as many queer geographers work from feminist, postcolonial, and antiracist positions. Queer theory has had a major impact on geographers as they work through analyses of social, cultural, and institutional practices and discourses that are productive of gendered and sexual knowledges, subjects, spaces, and places.

SEE ALSO: Antiracist geography; Bodies and embodiment; Cultural geography; Feminist geography; Gender; Heteronormativity; Identity; Postcolonial geographies; Sexualities

References

QUEER GEOGRAPHIES


Race and racism

Audrey Kobayashi
Queen’s University, Canada

The concept of race as a means of sorting and segregating human beings is ancient, but its modern manifestation, as a means of sorting bodies according to their putative abilities based on phenotypical or genotypical differences, dates to the time of European ocean exploration, starting in the fifteenth century, when the quest to classify bodies based on skin color was folded into the colonial/imperial project, with ramifications for global economic power, the transfer of resources to Europe, and the contest for global geopolitical power. Notwithstanding interaction between Europe and Africa since at least the twelfth century, the contest to control the globe from the so-called age of exploration until the so-called Enlightenment of the eighteenth century involved not only mapping and documenting the globe, but also the active formulation of racial difference in both scientific and popular ideas. Immanuel Kant, who began teaching “physical geography” at the University of Königsberg during the 1750s, lectured on the classification of the human species, a topic that was of immense controversy among scientists at that time. His novel but fundamentally racist theories were based on an attempt to understand the differences between so-called races. He postulated a relationship between skin color and distance from the equator, and believed that those born in Africa are originally white, but become black as the sun interacts with capillaries under the skin, affecting not only their color but also their capacity for intelligence (Kant 1997/1775).

Race is a product of racism, and it is deployed through a process of racialization. Over the next 200 years, both popular and scientific understanding of racial difference advanced to create an overwhelming Eurocentric assumption of the supremacy of the white “race.” The global division of labor was heavily racialized as well as classed, structured to include plantations dependent on slavery from Africa and indentured labor from Asia, as well as labor in situ, for example to work the mines, clear the forests, and build the railways in colonized areas of the world. The forced movement of millions of people around the world resulted in new population geographies, while scientists continued their work of classification, with climate playing a major role in explaining racial differences. By the time Darwin presented his work on human evolution during the mid-nineteenth century, many scholars were prepared to adopt a historical explanation of differential rates of human evolution, while others clung to a polygenetic view of races as independent from one another. Both monogenetic and polygenetic views, however, saw race as a fundamental differentiator of human qualities, including intelligence and capacity for civilization, in what David Livingstone (1991) has termed a “moral discourse” that tangled geography, anthropology, and esteemed groups such as the London Geographical Society.

By the early twentieth century, environmentalism was the dominant paradigm in geography. Geographers produced global maps of racial categories, often associated with putative climatic influences. Such ideas were widely diffused in school textbooks throughout the world.
Ellsworth Huntington (1876–1947), perhaps the most adamant adherent of environmental determinism, wrote many articles and books linking climate and civilization (e.g., Huntington 1924), in which people of the tropics, especially, are depicted as lazy, dishonest, or immoral. Ellen Churchill Semple (1863–1932), influenced by the ideas of Friedrich Ratzel (1844–1904) while she was a student in Leipzig, Germany, is generally depicted in the literature as an environmentalist. Her efforts were directed to understanding how human beings have adapted to and modified the physical environment, however, and not to race itself as an explanatory factor. To the extent that she used the term, it was in reference to groups of people identified with a particular region, and she did not address questions of variations within regions. Indeed, Semple acted as an adviser to President Woodrow Wilson, himself no defender of racial equality, during the post-World War I peace talks. She advocated that the ethnic qualities of national populations be taken into consideration when redrawing European boundaries (Silver 1984).

As debates over environmentalism developed, various countertheories were forwarded, including the belief that, while climate is not the determinant of racial variation, human beings fall into naturally different categories, with naturally different characteristics, in different regions of the Earth. The challenges to environmentalism were many, however, and increased throughout the twentieth century, although in the discipline of geography such challenges were not directly linked to understanding, or opposing, the concept of race. The American geographer Carl Sauer (1889–1975) was perhaps the most vocal – or at least the most cited and listened to – advocate of rejecting environmentalism. Sauer rejected Huntington’s work, as well as that of German scholars such as Ratzel who had informed Huntington. Sauer was deeply influenced, however, by other German scholars such as Franz Boas, whose understanding of cultural processes informed a generation of anthropologists. Sauer did not, however, challenge the assumption of racial differences; rather, he advocated that it was culture rather than race itself that explained variations in human practices in different places. Urged by Sauer and an increasing number of other voices, geographers by and large abolished environmentalism from their analytical lexicon. It has been speculated that so strong was the push to distance the discipline from environmentalist influences that post-World War II geographers seem to have had little taste for any speculation about race. Geographical scholarship was therefore not a part of scholarly discussions that emerged in other disciplines, especially anthropology, challenging the idea of race. Geographers did not take the opportunity, at least in the pages of their scholarly journals, to make clear links between environmentalism and racism.

In Germany, there is no doubt that the determinist ideas of geographers such as Friedrich Ratzel, linking putative racial qualities and the power of the state, had an influence on Nazi ideology. If the same connections had been made with the work of avowedly racist determinists such as Huntington in America, or even of Halford Mackinder in the United Kingdom, the Nazi ideology would have found even more support. On the other hand, Walter Christaller (1893–1969), another German geographer who had published a treatise on the distribution of urban centers according to what he called “central place theory,” in which race did not play even the smallest role, was hired by the Nazi administration during World War II to develop his rational ideas as a basis for planning the repopulation of areas purged of Jews, Roma, and others of a non-Aryan background. His work was later taken up by geographers everywhere as one of the most
important contributions to modern urban geography. In hindsight, the widely different ways in which Christaller’s work was taken up illustrates that scholarly ideas are invoked within particular contexts. It is perilous to make simplistic connections between racist ideas and geopolitical acts. A much more complex understanding is needed of the ways in which ideas are understood, lived, used, challenged, modified, and deployed.

Ashley Montagu (1905–1999), an American anthropologist, published just such a complex analysis in *Man’s Most Dangerous Myth: The Fallacy of Race* (1997/1945). He then became part of a United Nations commission that in 1950 brought together a group of scientists to study the “race question,” who arrived at the definitive opinion that there is no scientific basis for differentiating human beings on the basis of race, a finding that was published by UNESCO in a series of papers during the 1950s. Montagu’s influential text was republished in several editions, the last in 1997. But he was a person before his time; he was denied tenure at Rutgers in 1955 at a time when his ideas were still controversial, notwithstanding the findings of the United Nations in the aftermath of World War II.

A history of the concept of race is more than simply a description of how ideas emerged. It is also an accounting of how the past and the present are interconnected (how past ideas condition and are challenged by present ideas) and also how past silences and omissions are imbricated in present understanding (how silence emerges as the contrapuntal interval of the spoken). Scholarship is in many ways a dialectic of the spoken and the unspoken. By the 1940s, the attention of the entire world was focused on questions of racism. It would take some time before the shock of what had been done in the name of anti-Semitism would resonate in the Civil Rights Movement, but a decade after the war, in 1955, the Montgomery bus boycott in Alabama provided an opening event that would lead over the next decade to a surge in public support for the concept of racial equality. One of the main organizers of the boycott was Thelma Glass (1916–2012), a geographer at the University of Alabama. Notwithstanding her efforts, and doubtless the efforts of many more geographers who may have been involved in the events of the time, the geographical literature remained curiously silent about race and racism.

Also notwithstanding these developments, the uncritical and biologically determinist assumption of racial difference continued to inform the discipline, especially in urban geography. This subdiscipline owed much to work by members of the Chicago School of the 1920s and 1930s, and especially to sociologist Robert Park (1864–1944) and his colleagues, who mapped the spatial patterns of the city according to race and ethnicity. Based on concepts of both urban ecology and an urban moral order, Park believed that, much like plants, groups of people complete for space across the city and engage in activities akin to ecological invasion and succession, thus traveling through the city’s residential areas from the center to the periphery as their circumstances improve and their housing options increase. Park believed that urban residential location depended on a racialized science of space from which African Americans were excepted because their moral order prevented them from engaging in this process (Kobayashi 2014). When urban geographers took up Park’s ideas decades later, they ignored his writings on moral order and concentrated only on his theories of space. But they nonetheless mapped and analyzed a racial spatial order, with no indication that they challenged the fundamental assumption that races exist as given.

It is indicative of the academic politics of the time that geographers and other social scientists were drawn to Robert Park and the prestigious
Chicago School rather than to the African American scholars of the time. Park had, in fact, worked with Booker T. Washington, of the Tuskegee Institute, who was well known for believing that the circumstances of African Americans could be understood only as a result of their own cultural capacity, a view that Park shared. At the same time, Park clashed publicly with the African American scholar W.E.B. Du Bois (1868–1963), a graduate of Harvard, who had a very different view. Among his many books and articles addressing what he called the “double consciousness of blackness,” which he depicted as the struggle to remove the color line that separated black and white in America, Du Bois (1967/1899) created maps of ghettoized poverty in Philadelphia based on an 1899 study that reached beyond its times in every respect. Du Bois was very clear that race is a concept invented in order to create the color line, and he was uncannily prescient about the ideas that would develop among geographers more than a century later.

During the 1960s, however, many urban geographers were bent on elevating the science of space as the mainstay of the discipline. They paid little attention to Park’s ideas of race or moral difference, but they leaned heavily on his idea of spatial science to create a new version of urban ecology based on spatial models (e.g., Berry and Kasarda). With the power of new methods of computing, extensive datasets, and sophisticated survey methods, geographers mapped residential segregation to show that, in particular, blacks lived in residentially segregated areas of the city, and that those areas by and large coincided with segregated poverty. Such quantitative analyses paved the way for a small number of geographers who by the 1970s had made a link between spatial patterns and racism (Morrill 1965; Rose 1970). While using the same techniques of mapping, these geographers focused on racism as a form of human behavior, and as such were among the first to question deterministic racial understanding.

Also during the 1960s, a small number of radical geographers Radical geography pushed the discipline to become more socially relevant. Their work was informed by the Civil Rights Movement, as well as by the antiwar movement and (later) second-wave feminism, but their major focus was on class and on the link between capitalism and space. The journal Antipode featured a spirited discussion of race and residential segregation during the late 1960s – probably the most attention the topic had ever received in the discipline – but such discussion was soon eclipsed by other issues. One exception was James Blaut (1927–2000), for whom racism was a product of imperialism. He wrote: “Imperialism … is white exploitation of the non-white world, a plague that began some 500 years ago on the West African coast and spread across the globe” (Blaut 1970, 65). Another exception was William Bunge, who undertook controversial “expeditions” into North American inner cities, working directly with African Americans and others to map poverty (Bunge 1971). But these are the rare examples of white male geographers who have made a difference in bringing the issue of racism to the attention of the discipline.

Racialized geographers at the time were few in number, but they were actively involved in the Civil Rights Movement, albeit under the radar of most of the discipline. Thelma Glass, mentioned earlier, took her work off campus as a member of the Women’s Political Council, the group that organized the Montgomery bus boycott in 1955 and supported Rosa Parks a year later in the Supreme Court case that led to the abolition of legal segregation (George and Monk 2004), but she received no recognition in the discipline until decades later. Others noticed the structural racism of geography itself, and advocated
for social justice within the discipline. At the behest of Saul Cohen, then executive director of the AAG, Donald Deskins became the first organizer of the Commission on Geography and Afro-America (COMGA), an effort to increase the number of “minority” students in graduate programs in US Geography departments (AAG Diversity Task Force 2006). Deskins also compiled a bibliography of geographical scholarship on race, comprising a list of overwhelmingly white, male scholars who wrote primarily about either migration or residential spatial patterns (Deskins 1969).

Notwithstanding these attempts to address and overcome racism, it was not until the 1990s that geographers began a theoretical challenge to the concept of race itself. That challenge was inspired by and had its roots in the Pan-African movement that began after World War I, and by poststructuralist ideas that emerged from it, particularly in France, after World War II. After World War I, it was none other than W.E.B. Du Bois who emerged as a leader of the Pan-African movement, after he attended the Paris Peace Conference of 1919. The movement then consisted of a loose-knit group of black activist intellectuals who met at various places. Post-World War II Paris was a place where a new generation of Pan-Africanists congregated: individuals such as Léopold Sédar Senghor, future Senegalese president; the poet Aimé Césaire; the Congolese independence leader Patrice Lumumba; the Ghanian Léon Damas; and Frantz Fanon, a medical student born in Martinique – as well as a small number of European scholars, most notably Jean-Paul Sartre, in whose apartment the group convened. This movement provided much of the intellectual (and in some cases actual, on-the-ground) support especially for French decolonialism over the next several decades. Their work also profoundly influenced the development of postcolonial studies (see Gilroy 1992; Kobayashi and Boyle 2014).

In geography, the genealogy of antiracist scholarship has two clear lineages. First, James Blaut’s (1993) work on the destabilizing role of colonialist racism in developing countries drew explicitly on the Pan-African movement, and profoundly influenced a generation of geographers to address the international context of racialization. Second, geographers trained in the tradition of ethnic residential segregation, particularly those who studied under Oxford geographer Ceri Peach – such as David Ley, Susan Smith, and Peter Jackson – shifted their focus from studying race to studying racism. Peter Jackson’s (1987) edited collection on Race and Racism was the first expression, especially in social geography, of race as a socially constructed category, and of the recursive relationship between race and racism. Jackson’s work was strongly influenced by that of British sociologist Robert Miles, whose work is a direct translation of that of Colette Guillaumin, a French antiracist, feminist theorist who had also been a member of the UNESCO panel convened by Ashley Montagu. Although these two strands of antiracist thought in geography did not intertwine much at the time, they were both tied firmly to poststructuralism. Geographers have been strongly engaged with that tradition ever since, influenced by the work of postcolonial scholars such as Frantz Fanon, Stuart Hall, and Paul Gilroy, to name a few.

The influence of poststructuralist thinking on geographical approaches to race and racism has been profound. In simple terms, it acknowledges that race is a human invention, albeit one with a long and complex history and geography; perhaps the most powerful concept ever used to sort human beings, positioning them in specific ways, and with specific worth, in the flow of Colonialism, decolonization, and neocolonialism, globalization, urbanization, and geopolitics.
Poststructuralism is based on the principle of anti-essentialism – that human bodies become racialized, through powerful discursive processes of racialization – laying to rest any notions of race as a predetermined, primordial, or precognitive category. This is not to say, as has often been said, that race is not real; the reality is the powerful production of race and the lived conditions of racism that result. Race may be a pernicious concept, but it is not an abstract one.

The importance of anti-essentialist definitions of race (like that of gender) cannot be understated. It is often remarked that race is only socially constructed or only a concept, as though Montagu’s most dangerous myth could be made to disappear by exposing its roots in the human imagination. But anti-essentialism needs to be understood dialectically as the discursive relationship between people and their environment (including other people), mediated by the production of concepts that are not, in themselves, free-floating signifiers. In other words, reality is socially constructed. Reality is the discursive relationship from which ideas cannot be extracted or abstracted, not something that is underlain, structured, or created by some other – more material – something. As a relationship, racialization is also recursive: it involves the dialectical construction of the self and other. This point is made by a number of poststructuralist scholars, particularly Frantz Fanon who wrote at length about the ways in which blackness is constituted by whiteness but acts back to constitute both the black and the white. Fanon was inspired by the dialectics of Jean-Paul Sartre, whose work is particularly compelling for geographers because of his insistence on the importance of situation as a recursive relationship, which he called “spatializing–spatialized” but which also might be understood as “racializing–racialized” (Kobayashi and Boyle 2014).

The process of racialization – and its product, race – is therefore a thoroughly material production (it involves, for example, labor markets, housing markets, and so on), and it is this very materiality that affords race its ontological status. Understanding its ontological status has shifted, however, from debates about what race is (recalling Enlightenment period debates about how many races there are, or what physical attributes mark particular races), to how it has been constructed and deployed in the world, in the creation of specific landscapes that are the Habitus, that is both the dwelling and the doing, of human beings. Race is thus, in simple terms, contextual.

If the process of racialization is contextual, it always occurs in place. This point more than any other has energized a new generation of geographers to take up the study of how the modern world is racialized in specific places or by connecting specific places. The production of racialized places is also deeply political, but racist geopolitics too need to be understood as permeating every aspect of humanity (McKittrick and Woods 2007). For if race is not a given, nor simply an expression of some external science of space, then it follows that geographers should study the process of racialization, the effects of racism, according to actual conditions in places and in reference to the ways in which race is embodied in place, which is pretty much what has happened since the 1990s.

Urban geographers still map the consequences of racialization, often in very useful ways and using the power of new spatial technologies, but they are more concerned with how racializations shift and change in context. The world is now more urbanized than it has ever been. Cities are not only larger but often much more multicultural. International migration occurs at a higher rate than ever before, and there are more refugees fleeing a range of circumstances including war,
environmental disasters, poverty, and oppression. The major movements of people occur both intranationally as people move from more rural to more urban areas, and internationally, where the movement still tends to be from the developing to the developed world. Many of the destinations are settler societies (the United States, Canada, Australia) where colonial patterns remain to condition the lives of newcomers. Many destinations are also to traditional colonizer societies (Germany, the United Kingdom), where difference breaks on a hard core of established national identities. The cities that have the highest levels of recent migration (e.g., Los Angeles, Toronto, Miami, Amsterdam, London, Sydney, Dubai) have the greatest potential to shift popular understanding of what constitutes difference, but they also have the potential to create new forms of residential segregation. One of the ironies of shifting population over the decades since decolonization is that the other is no longer at a distance, but in many places now lives in the metropole, in closer proximity not only to the colonizer but also to other groups of the colonized, with the potential to construct new hierarchies of difference.

But, if cities are changing as a result of recent processes of decolonization, globalization, and urbanization, they are also the result of long historical geographies of racialization, especially as a result of slavery and indentured labor. The urban map of the United States, for example, needs to be read through several epochal movements: the creation of plantation societies based on slavery; the post-Civil War Great Migration to the industrializing North; the devastating effects of deindustrialization and more recent financial crises. All of these processes are shot through with racialized violence that continues today in encounters between racialized youth and urban police forces, in the violence of the prison industrial complex (Gilmore 2007), and in racialized poverty, to name some of the most significant.

If race is socially constructed, then so too is whiteness, and an increasing number of geographers have undertaken Whiteness studies to show that black and white represent a dialectical relationship, two sides of a single disc, one constituting the other. Whiteness studies in geography have unveiled racialization as enacted, spatially and temporally, in the most pervasive of moral, ethical, emotional identity formations that create and sustain difference (e.g., Jackson 1998; Bonnett 2000; Saldanha 2007). Recently Bonds and Inwood (2015) have emphasized the extent to which white privilege is embedded not only in personal relations but also in the state projects that constitute colonialism, as important now as ever. The sociospatial dialectic, they note, how land use practices and histories connect to the legacies and permanence of these structures and to the development of new understandings of the ways in which genocide, land appropriation, and labor exploitation are tied to white supremacy and other structures of oppression. (Bonds and Inwood 2015, 15)

The dialectic of racialization also intersects with other identity-forming discourses in a complex world. Intersectionality varies spatially and temporally, even as one moves through one’s day from one place to another, and it may be contradictory. Geographers seek understanding of landscapes of contradiction along lines of gender, ethnicity, age, religion, or class (Dwyer and Bressey 2008). For example, the way that gender identity is constructed changes not only according to the context in which it is enacted or performed, but also in the course of an event. Similarly, racialized sexual identity is a precarious situation subject to cultural change. Nayak (2003) demonstrates that racism and antiracism are both place- and age-specific, and intersect with whiteness.
Recognition of intersectionality has been a particularly powerful mode of thinking on the part of antiracist, feminist geographers such as Ruth Gilmore, Minelle Mahtani, Patricia Noxolo, Linda Peake, and increasing numbers of others, whose work spans international contexts to show the many and specific ways in which feminists cannot afford to ignore the often different conditions in which gender is constructed. This work collectively emphasizes the importance of understanding the human body as racialized and gendered and classed and differently abled, in ways that shift according to place and context.

Geographers also face a major challenge in their continuing dialogue with the biological sciences. The biologist Stephen J. Gould warned decades ago about the problem of the “mis-measure of man [sic],” of assuming that the study of genetic traits could somehow be hived off from human social history. Preconceived notions of genetic traits as somehow bundled into races have influenced scientists to ignore the fact that the very concept of race is socially constructed. The exciting scientific work of the Human Genome Project has led to many advancements in understanding human bodies and their interrelatedness, and provides plenty of evidence that such bundles are a mismeasure. Yet, ironically, as scientists discover more about genetics, social scientists need to work just as hard to deconstruct the racialized concepts that tend to be reinserted into genetic science. As Catherine Nash (2015) shows, notions of sex, nation, and identity are easily remapped onto genetic patterns, reinforcing old ideas in new packaging.

Such remapping can even lead to new forms of environmentalism. Recent emphases, for example, on the “Anthropocene,” as the period in which humans have had the greatest impact on the surface of the Earth, allows for a comprehensive, and potentially dialectical, thinking about geography. But physical scientists studying human impact, which may be read as a reaction to “nature,” are not trained to study human experience (and vice versa for social scientists); there is a risk of the insertion of popular notions of genetic adaptability that reracialize encounters not only with the Earth itself but also between humans. Similarly, recent attention to the “more than human” that has swept human geography, while giving due significance to the fact that the Earth is not only human, also risks reducing the human to just one of the many actors that have agency, or to bundles of putatively natural traits abstracted from their geosocial context. Here, too, geographers need to exercise not only caution over reinforcing deterministic assumptions, but also careful analysis of the ontological status of their subjects, for most of the more-than-human scholarship to date tends to homogenize the human, ignoring the significance of racialization in mediating human actions.

A similar ontological risk occurs with the implementation of Nonrepresentational theory (NRT) in geography, intended to move beyond fixed images and texts to recognize the flowing and often ephemeral qualities of human experience, including the production of affect that cannot be essentialized. A number of geographers have suggested that “more than representational” is a more appropriate term (see Lorimer 2005) in order not to lose sight of the fact that, on the one hand, all bodily performance is representative; it occurs in structured discursive contexts. On the other hand, even the hardest, most fixed of representations, say a marble statue in a public place, is always in flux, drawn dialectically into the flow of representation that never stops, that shifts with the viewer, with the geopolitical context, with the constant process of interpretation. Perhaps, when it comes to racialization, it is better to talk about how different forms of representation, including
their material consequences (of course all forms of representation, even ephemeral gestures, are material), their durability, their structural legibility, and especially their power endowed by their enactment, put in place racialized consequences.

It is ironic that, notwithstanding the vast amount of scientific evidence that shows race to be nonscientific, coupled with the vast and increasing amount of social science evidence of the ways in which the idea of race has been constructed and is used to organize the world in powerful ways, race remains one of the most dominant but least understood ideas in today’s world. W.E.B. Du Bois named the color line as the greatest problem of the twentieth century. Ashley Montagu named race as humanity’s most dangerous myth. Decades of critical race studies, and the evidence that antiracist geographers can see in the landscapes that surround them, have shown both scholars to be remarkably, and unfortunately, prescient. Bodies become themselves over time in complex ways, and the process of racialization represents a – and arguably the – dominant mode of becoming in the contemporary world. Geographical studies of race and racism have always been a reflection of their historico-geographical times, but too often they have been caught up in the dominant discourses of those times. Recent scholarship, however, provides a means of challenging dominant discourses in order to shift these most powerful of human creations in the direction of more just and democratic futures. Geographers have the ability to play a significant role in that project.

**SEE ALSO:** Antiracist geography; Environmentalism; Ethnicity; Feminist geography; Geopolitics; Globalization; Postcolonial geographies; Poststructuralism/poststructural geographies

---

**References**


Race, work, and employment

Beverley Mullings
Queen’s University, Canada

The nineteenth-century expansion of industrial capitalism is often the starting point of analyses of the relationship between race, work, and employment, but race and racial ideologies played a defining role in the division of labor as early as the fifteenth century, when an integrated capitalist world system began to emerge. The practice of race-making – assigning social meaning to specific physical traits in order to differentiate, separate, and exercise control over particular groups of humans – is as old as the history of humankind itself, but after the fifteenth century it became inextricable from the social, economic, and political forces that converged to support European colonial expansion.

The emergence of plantation modes of production based on slave labor marked a crucial period in the evolution of social relations and institutional practices that constructed racial categories in order to differentially value the work of particular laboring bodies. By dividing groups of humans into races based on beliefs about the traits, abilities, and qualities that each group shared, and conferring unequal privileges on them accordingly, this early form of capitalism established a global division of labor, and ways of thinking about labor rights, freedoms, and responsibilities that remain even today.

While liberal political economy scholars like Adam Smith (1993/1776) and John Stuart Mill (1848) had little to say about the relationship between race, work, and employment, their views on the relationship between unfree labor and capitalist expansion reflected the dominant racial ideologies of their time. Smith and Mill both viewed unfree labor as distinctly incompatible with a fully functioning and efficient capitalist system. They believed that societies that relied on unfree labor were unlikely to embrace technologies that would boost labor productivity and consequently market expansion. Thus plantation economies dependent on enslaved labor ultimately stifled capitalist expansion because slave owners were unlikely to invest in labor-saving technologies for fear of losses they would incur if they were not to maximize use of their human commodities. Implicit in Adam Smith’s theorizing was also the assumption that economies dependent on enslaved labor could not be efficient because the unfree were unlikely to work hard. He reasoned that where labor was unfree, workers derived no benefit from hard work or from the acquisition of skills associated with innovation. He therefore viewed unfree labor as inefficient and antithetical to the success of capitalist systems of accumulation because of the dampening effect that lack of freedom had on labor productivity.

It is only recently that scholars have begun to question Smith’s claims and the role that racial ideologies played in his characterization of slave labor. For example, Smith claimed that all the most important improvements, whether in machinery or in the arrangement and distribution of work that facilitate labor, were the discoveries of freedmen. Yet, historian Veront Satchell (2010) documented that many of the patents for innovations in Jamaica’s sugar mills between 1760 and 1830 were the inspiration of artisanal
slaves who worked within them. Examining the relationship between race thinking and the value of labor, Brass (2009) also argues that, although Smith never explicitly included slaves in the category of unproductive labor, his views on the relationship between unfreedom and idleness were already racially coded. In this respect, Smith’s inferences reflected views that were popularly held by scholars during the Enlightenment period, even before scientific forms of racism became fully embedded in social practice. Like Immanuel Kant, who viewed African and other racialized peoples as naturally inferior to whites, or David Hume, who reportedly declared there to be no free Negroes who had distinguished themselves with any particular skill, assumptions about the inefficiency of unfree labor were reflections of dominant racial ideologies that constructed the enslaved as predisposed to idleness in order to justify the forms of violence and superexploitation that they experienced.

The belief that racialized labor was synonymous with inferior labor can be traced to the influence of the philosopher Immanuel Kant, who believed that climate triggered permanent, irreversible, and heritable deviations that gave rise to the biological differences and traits that produced races. Thus Kant claimed: “A race, when once it has taken root and extinguished the other seeds, resists all further transformation because the character of race at one point became dominant in the generative power” (Kant, quoted in Larrimore 2009, 15). Bernasconi (2011) alludes to the extent to which the circulating ideas about race, of which Kant was a key progenitor, traveled through time and space. While Kant argued that the enslaved Africans in the Americas could not innovate and that progress through work was a permanent and inherited condition, he tended to support his views with deliberately sourced writings that cast black peoples in a negative light. Drawing on the writings of pro-slavery writers like James Tobin, who is believed to have claimed that, even when freed blacks were encountered in Britain or America, there was no instance where anyone pursued an occupation that could really be called work, Kant’s formulations on race set in motion a way of thinking about racialized bodies in relation to work that continues to be influential even today (see Bernasconi 2002).

Today the idea that unfree workers are inefficient because they lacked an incentive to work hard is recognized as an ideological fiction which, combined with emerging scientific constructions of race in the eighteenth century, served to justify the oppressions of European imperial expansion. But the link between race and human capacities, dispositions, and entitlements to work continues to haunt the discourses and practices that produce inequalities in work and employment.

**Geographical knowledge, race science, and labor market discrimination**

The geographical imagination played an important role in legitimating imperialism and the forms of knowledge about race and work that it produced. As a discipline, geography was widely regarded as the science of empire because of the rigorous way in which it produced detailed generalized accounts of the Earth and its peoples that were indispensable to imperial efforts to conquer and govern others. Geography offered a knowledge of the world that made it possible for the knower to pragmatically engage it. For example, geography’s development of mapping techniques to spatially delineate human environmental interactions within and between specific places functioned as a technology that provided colonizing powers with detailed knowledge of the overseas territories that they had acquired. Particularly with regard to its production of knowledge about the relationship between
climate, environment, and race, geography provided the theoretical frameworks needed to
dignify, as Said (1978) once observed, Europe’s
desire to appropriate geographic space. The ide-
oologies of white supremacy and imperial control
that geographers helped construct created what
Blaut (1993) described as the “colonizer’s model
of the world,” a model that supported a diffu-
sionist view of history that placed Europe and
Europeans at the center of all that represented
progress and modernity.

In the late 1800s, reflecting to the rapid
expansion of European colonialism, geographers
revived and extended many of the ideas about
race and human character that Kant had explored
over 100 years earlier. Geographers like Semple
(1911) and Huntington (1943) saw the biophys-
ical environment as playing an instrumental role
in shaping the psychological mindset of indi-
viduals, their behavior, and the attributes of the
societies they formed. Those in tropical climates
were given to laziness, relaxed attitudes, and
promiscuity, while groups in environments with
greater climatic variability were more deter-
mined and were driven by a strong work ethic.

By the early twentieth century, the popularity
of environmental determinist theories in geog-
raphy helped to establish the natural superiority
of those racialized as white to colonized others,
and consequently their “obligation” and special
capacity to govern them. Thus, even in the
1940s in the United States, when the Southern
states could no longer command the levels of
agricultural capital that they once had because of
crop disease and emigration, scholars continued
to draw on discourses of black incompetence and
lack of commitment to explain the crises within
these labor landscapes. Thus, as geographer J.S.
Gibson (1941) lamented, part of the reason for
the decline of Alabama’s agricultural belt, could
be attributed to the inability of its remaining
black population to facilitate change. For, as
he observed, lacking good judgment, a sense
of thrift, and a feeling of responsibility, black
farmers rarely attempted cattle raising on any
appreciable scale and, when working for others,
required close supervision.

While the ideas underpinning environment-
tal determinism were heavily critiqued and
eventually discredited, the belief in biological
races with sociopsychological characteristics that
reflected the environments within which they
held ancestral affiliations, remained an enduring
rationale for the systems of labor management
that emerged under colonial rule and persisted
until the 1960s. Roediger and Esch (2012)
document the way that labor management in
the United States at the turn of the twentieth
century drew on racial ideologies in order to play
one racialized group of workers against another.
They argue that the “racial” knowledge that
many slave masters and overseers claimed to have
quickly transformed into a set of discourses and
practices that linked whiteness to management.
They contend that it relied on divisions and
competition between racial groups, as well as
the duty to uplift racial others, as its primary
governing technologies. Similar racial ideologies
also existed across the British empire, providing
a justification for the forms of labor market
discrimination that became a part of standard
employment practice up until the 1960s. The
idea that racialized workers should never be
treated equally within labor markets, however,
masked the material nature of labor oppression
that ultimately drove down the overall value of
all workers by keeping wage floors low.

By the end of the nineteenth century, the idea
that the world could be divided into regions,
each populated by homogeneous racial groups
with particular temperaments and capacities to
work, no longer characterized the everyday labor
landscapes in which people worked. Migration,
war, and, importantly, the abolition of slavery
and indentureship brought ethnic groups into contact with each other in ways that had not been experienced before. In the United States, for example, the construction of the Transcontinental Railroad in the late 1800s provides a strong example of the way the strategy of racial competition was used to generate a workforce to undertake the grueling task of railroad construction at the lowest possible cost. Early efforts to generate a white labor force to build the railroads proved to be unsuccessful at the rates of pay, levels of productivity, and exposure to danger that companies like the Central Pacific Railway demanded. Immigrant Chinese workers, however, not only proved more efficient and productive, but were also, at least initially, considered more docile. They eventually displaced the white and mostly Irish workers who resisted the labor conditions associated with railroad construction.

Although Chinese workers emerged as a preferred labor force to whites, their labor was simultaneously devalued by their employers. Chinese workers were systematically paid less than their white counterparts, a practice that was, in large measure, made possible by the power of the circulating racial discourses that constructed them as a despicable group, a contempt that ultimately culminated in the restrictions placed on the immigration of Chinese labor by the US government in 1882 under the Chinese Exclusion Act. Though preferred for their reliability and skill, racialized Chinese workers were considered an expendable workforce whose abilities and rights did not merit recognition.

By the twentieth century, the abolition of slavery and large-scale mass migration opened up labor markets, particularly in the Global North, to broad and ethnically diverse populations, drawn from all over the space of empire. But widespread discriminatory employment practices ensured that the racial hierarchies that had been established over 300 years before could be maintained through the wage process. Racialized wage discrimination remained a largely uncontested practice up until the 1960s. It was an almost universal assumption that groups considered to be “racially inferior” and distasteful should be “taxed” for the privilege of working alongside racialized superiors. Lower rates of pay and the denial of benefits became the primary method of exacting this form of “tax.” Much of the logic behind racial wage discrimination was closely tied to deep-seated concerns about the levels of disaffection among working classes that would emerge if white workers were to work alongside workers considered to be their racial inferiors. But of equal concern was the fear that alliances between racialized and ethnic groups would ultimately destabilize labor control. The East St Louis race violence of 1917, which left 100–200 African Americans dead and an additional 6000 burned out of their homes, points to the deep threat that racialized workers posed to the racial division of labor. Despite that fact that there were labor shortages in the United States because of large numbers of potential workers engaged in active service during World War I, there was still violent outrage at the prospect of black workers from the South filling these vacancies. This was reflected in Marcus Garvey’s speech delivered after the event, in which he proclaimed that, despite the fact that for 300 years “the Negroes of America have given their life blood to make the Republic the first among the nations of the world,” they continued to be oppressed, lynched, burned, and butchered.

In Canada, discrimination against immigrant Chinese people not only took the form of wage discrimination, with workers building the Canadian Pacific Railway routinely earning half that of their white counterparts; once the railways were completed, it became manifest in
Canada’s immigration policies. Devised in 1885, the Chinese Immigration Act, also known as the Chinese Head Tax, imposed a $50 fee on each Chinese immigrant entering Canada. By 1903 this tax had risen to a maximum of $500 in an effort to deter further expansion of the Chinese labor force, and by 1923 it was replaced with an outright immigration ban on most Chinese migrants. By the end of the 1920s the power of the imperial racial imaginary to shape local and global spatial divisions of labor became acutely clear. Not only did race knowledge influence how employers determined the value placed on the labor of specific raced bodies, but even when the work performed was of superior quality it allowed states to maintain racial divisions of labor internationally through the regulation of the international migration of people altogether.

Throughout the 1900s geographers continued to lend legitimacy to these unequal practices through their efforts to establish the naturalness of discrimination. A case in point was the geographer Hereford Brooke George (1905), a fellow of New College, Oxford, and author of the influential textbook, *A Historical Geography of the British Empire*. In this book George’s casual reproduction of much of the circulating race knowledge about Chinese people in Australia supported the racial discrimination that ultimately led to the exclusion of Chinese immigrants from the United States, Australia, and later Canada. Though constructed as a careful observation of the character of different ethnic groups, his likening of Chinese immigration to a flood, and his belief that even those supporting socialist values would agree that “in a country inhabited by white European Christians the fewer Chinese the better,” served to reinforce established patterns of racial discrimination and exclusion (George 1905, 189).

Decolonization, the demise of race science, and the challenge of racism

By the 1960s the idea that the world was composed of biological races with distinct, inherited, and permanent physical and cultural characteristics had been thoroughly discredited. In 1950, at the end of World War II, the United Nations Educational, Scientific and Cultural Organization (UNESCO) produced the first of four statements on race that aimed to communicate what was scientifically known about the relationship between race and biology. The statements, signed by a number of leading researchers from the social and biological sciences were produced in the wake of the Nazi genocide of Jewish and Roma people (UNESCO 1967). They constituted a major step toward discrediting the social and political significance of race science, and contributed to a growing recognition of racism as a shameful sentiment to be disavowed. While the debunking of the scientific concept of race, and the moral condemnation of racism implicit in the UNESCO statements (and subsequent revisions), did not put an end to the production of racial inequalities among individuals or within institutions, the statements galvanized a whole generation of scholars across geography to devote their attention to the exploration of the role that race and racism played in shaping human relationships with environments.

Even though environmental determinism had lost its centrality within geography by the 1920s, it was not until the 1960s and 1970s that geographers began to study in earnest the relationship between race, work, and employment. Motivated by growing pressure to demonstrate its relevance as a distinct discipline with something to offer to an increasingly decolonized world economy, geography reinvented itself after World War II as a science concerned with producing facts and models to facilitate spatial
decision-making in the period. The quantitative turn, with its reliance on abstract theory building and determinist model building, had a profound effect on the study of race and work and employment because it provided geographers with a set of formal statistical techniques from which to draw scientific inferences about the spatial effects of racism – a shift that helped to reorient geographical scholarship away from its unsavory supporting role in the production of race science.

In the United States, the African American Civil Rights Movement was a catalyst in the number of studies that examined the effects of residential segregation on the quality of life of African Americans during that period. Geographers like Deskins (1972) and Morrill and Donaldson (1972) acknowledged the role that racism in labor markets played in the residential opportunities open to African Americans. While their work tended to focus more on the spatial and residential implications of racial discrimination for African Americans, and less explicitly on labor market discrimination, their writings represented a significant shift in the way that geographers engaged questions of race and racism in the discipline.

In Britain geographers were slower in their exploration of relationships between race, work, and employment. Ceri Peach (1966) represents one of the earliest of geographers to write about the relationship between race and racism in Britain, in the context of the postwar mass migration of West Indians and later South Asians to Britain. Like that of his American counterparts, his work in the 1970s focused on the relationship between race and spatial segregation and the spatial patterns of inequality that racism produced, an approach that offered an understanding of racism that focused primarily on its geography rather than its etiology. Thus, despite their clear break with the past, the pioneering studies examining the relationships between race and space between the 1960s and 1970s were limited to documenting patterns of racism rather than examining the economic, social, and political complexities of race as a social and institutional construction that shaped access to work and employment.

Marxist theory and the institutionalization of racism within work

The work of Bridget Leach stands out as a significant departure from the theoretical and methodological approaches to race and racism that were popular among geographers in the 1960s and 1970s. Reflecting the growing influence of Marxist approaches within the discipline, Leach’s short article published in 1974, entitled “Race, Problems and Geography,” was one of the earliest studies to examine race and racism in the context of capitalism and for that reason one of the first to challenge geographers to examine the particular geographies that racism produced in order to serve the needs of capital. Leach criticized geographers for being too narrowly focused on describing spatial patterns rather than the exploring the underlying structures that gave rise to these patterns. Reflecting the growing influence of Marxist theory within the discipline, Leach argued that geography was in danger of serving state and elite interests rather than the interests of racialized groups and working-class communities because scholars tended to problematize the spatial manifestations of class relationships in society rather than class relations themselves. She concluded therefore that the policies advocated by geographers would likely support only cosmetic rather than transformative change.

Leach’s views reflected a broad and growing dissatisfaction with the epistemological direction that geography had taken in the years after
World War II, a dissatisfaction that David Harvey eruditely described in an article received by the Transactions of the Institute of British Geographers, two months after Leach’s, entitled “What Kind of Geography for What Kind of Public Policy?” Harvey observed that geography had undergone an epistemological shift that reflected a desire to change its image from a nonacademic practical activity oriented toward the “technics and mechanics of the management of Empire” to a distinctive intellectual discipline oriented toward the “the technics and mechanics of urban, regional and environmental management” (1974, 20).

As geographers embraced radical approaches inspired by Marxism, their focus increasingly turned to the ways that capital actively produced space and the conflicts arising from the very different conceptions of economic fairness held by labor and capital. Throughout the 1970s and 1980s scholars documented the ways that capitalist systems incorporated labor and the forms of exploitation and inequality that emanated from this relationship. Geographers studied the unique ways that particular racialized groups were incorporated into capitalist markets and the particular forms of inequality that emanated within workplaces and labor markets and across the international division of labor. Influenced by a rapidly changing international division of labor, the feminization of factory production, and large-scale immigration in cities in North America and Europe, geographers examined how particular ethnic and racialized groups fared in a rapidly restructuring world economy.

Among economic geographers, for example, much attention was paid to the large differences in employment rates that existed between black and white populations in the United States, with many arguing that these differentials were the product of the growing mismatch between the spatial location of jobs and the spatial entrapment of African Americans in segregated inner-city neighborhoods long abandoned by industry. A number of feminist geographers also contributed to these spatial mismatch theories in the 1990s by demonstrating how gender and racial differences combined to doubly disadvantage particular racialized and gendered groups.

While spatial mismatch theories examined questions of racial discrimination, segregation, and the formation of an underclass, the methods of analysis were largely quantitative, with little interrogation of the strategies or motivations employed by racialized workers themselves to change their situation. While race was never totally excluded from studies of economic restructuring in the 1990s, it never occupied a central role in framing how the relationship between capital and labor was understood.

From labor geographies to embodied workers: new critical approaches to work and employment

Reflecting on the primary focus of scholarship during the 1980s and 1990s, Herod (1997) and Coe (2012) have both argued that, although much primacy was given to documenting the economic geographies at work in the restructuring of capital-labor relations, levels of analyses tended to examine labor as a largely abstract, aggregate, and disembodied class, with little agency or capacity to effect change. The geographies of labor produced during this period therefore left little space for examining the multidimensionality of labor – that is, as embodied workers engaged in paid and unpaid work who exercised agency even in the face of discriminatory and exploitative systems of work. This shift in focus – from labor as a factor of production understood from the perspective of capital, to workers as embodied
subjects understood from the perspective of workers – began to emerge in the late 1980s as scholars, many using feminist frameworks, began to explore the experiences of women in the context of economic restructuring.

One of the earliest examples of this shift in focus from a geography of labor, toward a “labor geography” (a phrase coined by Andrew Herod in the 1990s), could be seen in studies examining the incorporation of women into transnational factory production. Research by Lawson in the early 1990s, for example, explored the changing relationship between gender and work in Ecuador in the context of the export-oriented industrialization strategies adopted by neoliberalizing Latin American states (Lawson 1995). Feminist geographers, charting the shift from welfare to workfare, also studied emerging forms of regulation within geographically specific local labor markets and their impacts on women’s everyday lives. Pratt and Hanson’s book Gender, Work and Space (1995) was characteristic of the way feminist geographers took account of the embodied nature of work and the way sex-based occupational segregation affected differently how men and women workers ordered their daily lives. McDowell’s (1997) study of the work experiences of women in the UK banking sector is similarly notable for their attentiveness to embodiment and the ways that constructions of gender shaped the nature of work relationships. But, despite the incorporation of gender into geographical studies of work and employment, very few studies incorporated detailed analyses of race (for an early exception see Johnston-Anumonwo 2000). It would, in fact, be almost a decade before labor geography turned its attention to the relationship between social constructions of race and their embodied effects on spaces of work (Wright 2006; McDowell, Batnisky, and Dyer 2007).

The growing incorporation of critical race theory into geographical inquiry has opened up new terrains of inquiry into the relationship between race, work, and employment. Omi and Winant (1994) view race as a fluid, shifting, and yet pervasive social concept that is integral to the maintenance of a given social order. Shaped by ideologies that ascribe social and cultural meaning to specific physical differences between individuals and ethnic groups, racial categorization and exclusion continue to be sustained by a complex array of social, economic, and political structures that operate and interact among individuals and institutions at a range of scales. Thus, contrary to popular views of racism as the actions of unenlightened or troubled individuals, critical race theorists see racism as an ordinary and normal feature of contemporary society, and an integral aspect of all social practices and institutions, including those that govern work.

Critical race theorists like Omi and Winant apply useful frameworks to contemporary issues of work and employment to understanding how racial categories structure interactions in daily life. Much geographic scholarship documenting the relationship between race and employment in the 1960s viewed racism as having deep and intransigent roots in the very institutional structures of society. But studies that rely solely on institutionalized racism as an explanatory frame have not easily attended to the changing nature of racial regimes since that time. In particular, accounts embedded solely within the framework of institutionalized racism have not been able to adequately respond to the fact that, since the 1990s, particular racialized groups have experienced upward mobility, a fact, that has led some to assert that we now live in a “postracial” world. The growing “racial” diversity of populations, the reduction of “race” to “culture,” and the rise of “postracial” neoliberal discourses that view individualism and competition as the
main determinants of success, have all challenged scholars who study the relationship between race, work, and employment to develop new ways of understanding the spaces of work that changing racial regimes produce. Also important has been the challenge to understand how changing racial regimes are being contested by individuals and social groups and how, ultimately, the dynamic and fluid nature of the individual and collective social relations that constitute a given racial regime change across space and across scales.

A number of critical race scholars have begun to examine how racial regimes are negotiated and contested within spaces of work. Many geographers have used social theorist Judith Butler’s work on the social construction of sexuality and gender, and in particular her concepts of performance and performativity, to understand access to and mobility within work. Arguing that performativity must be understood not as a singular or deliberate “act,” but, rather as the reiterative and citational practice by which discourse produces the effects that it names, Butler (1993; 1997) offered geographers new opportunities to study the role of performativity in the production of workplaces. How performances of deference, docility, aggression, or care influence how individual bodies accommodate and resist the social constructions and relations of power that govern places of work and employment opened up new avenues in the study of work and employment. Yet few of these early studies explicitly foregrounded how race functioned as a performance that influenced access and mobility within workplaces.

Arguing that the workplace is an important site of racial construction, Carbado and Gulati (2003) argue that race is a micro and interpersonal dynamic which in workplaces is produced by the images of race that people of color project and how they are interpreted. For these scholars, race is a social construction that people of color perform in their daily negotiation of workspaces in order to shape how (especially white) people interpret their nonwhite identities. They argue that in modern workspaces, where racial discrimination is outlawed and where racial diversity confers institutional legitimacy, employers are likely to develop management techniques aimed at maximizing perceived efficiency gains of a homogeneous workplace culture. They also argue that a similar set of incentives exist for nonwhite employees who, in order to counter racial stereotypes, must demonstrate a willingness and capacity to assimilate. Carbado and Gulati argue that employers respond to homogeneity incentives by picking outsiders whose performance of a racial identity suggests that they will fit comfortably within workplaces that are overwhelming racially homogeneous. They therefore conclude that, without an understanding of the implicit racial norms that structure workplace interactions, and their relationship to workplace efficiency objectives and employer hiring practices, it will be impossible to understand why racially based distinctions continue to structure geographies of work and employment. Studies by McDowell, Batnitsky, and Dyer (2007) make similar observations about the ways that embodied constructions of ethnicity structures workplace access and mobility in the hotel industry in the United Kingdom. Noting associations between Indianness and servility made by hotel employers, coworkers, and guests, they argue that individuals who fail to perform these gender performances adequately experience constrained opportunities for recruitment and advancement.

While the relationship between racial performance and employment is a new and relatively underresearched area, understanding how the racial projects that define particular types of workspaces operate (both discursively and structurally) at specific economic moments offers important insights into the microgeographies at
work in an emerging global racial division of labor.

Back to the global: integrating the local and the global in studies of work and employment

Few scholars have studied in any real detail how social constructions of race shape and order the global division of labor. Indeed, this is a difficult claim to research given the very different meanings assigned to bodies within different racial formations. Most studies draw on largely economic factors such as labor costs to explain why global capital investors might prefer the labor of particular racial or ethnic groups over others. Yet, as scholars like Werner (2015), Roberts and Mahtani (2010), and Thomas and Clarke (2013) argue, current forms of global capitalism obscure the ways that race is embedded in the political rationalities that structure how individuals and whole populations come to be excluded from or included in global labor markets.

Particularly in neoliberalizing economies, where markets are assumed to be free from bias, how race and ethnicity become a meaningful way of determining the value and worth of workers is especially opaque. For, as Roberts and Mahtani (2010) point out, neoliberalism relies heavily on a moral discourse that locates lack of access or mobility in labor markets to individual failings to work hard or to follow the rules of labor markets. Citing the devaluation of immigrant credentials, they argue that racism is able to thrive within Canadian labor markets because immigrants are constructed as not quite “Canadian” and hence excludable from employment practices that claim to be market-based and meritocratic. A number of scholars argue that, in order to evaluate how race functions as a technology that facilitates the maximization of worker output and efficiency, we must examine the ways that workers accommodate, disrupt, and resist the racial scripts given to them. One case in point is the way that race intersects with gender and class to structure the immigration and temporary migrant worker policies brokered by the Canadian and Philippines government. As a number of scholars observe, even though the language of race never enters these policy domains, they nevertheless shape how Filipina women come to be incorporated into transnational care chains. As Parreñas (2001), Bakan and Stasiulis (1997), and Pratt (2004) have observed, whether constructed as celebrated heroines given the faithfulness of their remittances, or as docile, caring, and religious, Filipina women who migrate to metropolitan centers negotiate institutions at home and abroad that draw on specific elements of Filipina ethnic and national identity to construct them as a particular racial category of worker without recourse to the rights and protections open to other types of workers.

If race influences how spaces of work and employment are structured, and if space structures how embodied individuals come to be valued as workers, is it possible to discern how these co-constitutive practices produce a racial division of labor at a global scale? Thomas and Clarke (2013) suggest that the contours of an emerging racial division of labor can be discerned in the ways that older racial orders are entrenched in the emergence of a global working class that remains segmented along racial, gendered, ethnic, and national lines even though the language of race is not mobilized. Examining how race functions as a technology in the production of a low-waged unprotected workforce can therefore be discerned in seemingly disconnected economic processes such as the bilateral trade agreements struck between governments from the Global North and Global South, and also in the repercussions of arrangements
required by international donor institutions for the recovery of sovereign debt, a point that Werner’s (2015) study of racialization and global factory work in the Dominican Republic and Haiti, so poignantly makes.

Recent contributions inspired by new materialist theories draw our attention to the way that global racial regimes are producing new divisions of labor that rely primarily on female biological reproductive labor in ways that resemble the incorporation of Filipina women into transnational care production. Waldby and Cooper (2006), for example, make connections between the rise of medically assisted reproduction as a new global business and the outsourcing of onerous and risky clinical labor to poor populations in the Global South. Focusing on global markets for women’s oocytes (eggs), they argue that if we explore women’s contributions to the biotechnology industry as labor, we can begin to see the ways that the materiality of the female body has become integral to these highly profitable forms of industrial production that rely on the biological labor of living tissues and reproductive processes to produce new humans. Constructions of race have been integral to the profitability of the fertility industry, as they not only determine the market value of bodies but also the value of their reproductive capacities. In keeping with the observations of new materialist scholars, the racialized, segmented, and hierarchical nature of global fertility labor markets cannot be divorced from the globally circulating discourses of human worth that have become integral to the value attached to the reproductive capacities of the female body. In fact, as scholars like Winddance Twine (2011) and Waldby and Cooper (2006) argue, emerging global social reproductive labor regimes continue to be haunted by the social meanings based on specific physical traits that were used to differentiate, separate, and exercise colonial control over particular groups of humans.

To fully understand how race structures work, and employment globally, we must continue to track the ways in which the economic ordering of space generates new labor flows. We must also continue to examine how these macro-level trends are accommodated and contested within the microgeographies of work. In other words, in order to account for the way that race functions as a technology in the global division of labor, we must make connections between macro-level social relations and their manifestation in the institutional practices of the state and corporate sectors, as well as in the everyday social interactions that reinforce and contest workplace practices that produce raced subjects.

SEE ALSO: Care work; Colonialism, decolonization, and neocolonialism; Corporatization of race: an American case study; Cultures of work; Emotional labor; Feminist geography; Gender, work, and employment; Global factory; Labor geography; Labor market segmentation; Migrant division of labor; Neoliberalism; Power and development; Race and racism; Social reproduction; Unfree labor

References

RACE, WORK, AND EMPLOYMENT


The term “radar” is an acronym for “radio detection and ranging.” A radar transmits an electromagnetic wave and then receives the echoes to determine the properties of the targeted object, such as its range, location, speed, and altitude. Radar was first developed during World War II. Since then, various radar systems have been developed for different purposes. A good reference work is the *Radar Handbook* edited by M.I. Skolnik (2008).

**Synthetic aperture radar**

Radar technology was developed for remote sensing of the natural environment and objects within it. Radar echo is the interaction of electromagnetic waves with this environment, and might provide quantitative and qualitative information about it, such as the biomass of a vegetation canopy, soil moisture, rain rate, oceanic surface speed, oil spill on water, snow-pack depth, building height, and much more (Ulaby, Moore, and Fung 1986; Tsang and Kong 2000; Jin 1994).

One of most advanced radar technologies is the airborne and satellite-borne synthetic aperture radar (SAR) (in microwave region $\approx 1–15$ GHz) (Olivers and Quegan 1998; Cuming and Wong 2005; Maitre 2008).

Recently, airborne millimeter wave (e.g., 94 GHz) SAR has also been under study. It can produce two-dimensional images of a terrain surface with high spatial resolution. Resolution refers to the fineness of the detail that can be distinguished in a remotely sensed image, as one of the most important performance parameters of radar image quality. In most imaging radars, high resolution in the range direction is achieved by using the pulse compression technique. In this technique, a sequence of frequency modulated (FM) or “chirp” pulses of long duration is used and the received signals are suitably processed to yield fine range resolution. Today, most imaging radars employ the pulse compression technique. In order to obtain fine azimuth resolution, a technique known as “aperture synthesis” is adopted for modern airborne and spaceborne radars, which are called synthetic aperture radar (SAR).

One dimension in the SAR image is called range (or cross track) and is a measure of the “line-of-sight” distance from the radar to the target. Range measurement and resolution are achieved in SAR in the same manner as most other radars: range is determined by precisely measuring the time from transmission of a pulse to receipt of the echo from a target. Range resolution is determined by the transmitted pulse width, that is, narrow pulses yield fine range resolution.

The other dimension is called azimuth (or along track) and is perpendicular to range. It is the ability of SAR to produce relatively fine azimuth resolution that differentiates it from other radars. To obtain fine azimuth resolution, a physically large antenna is needed to focus the transmitted and received energy into a sharp beam. The sharpness of the beam defines the
azimuth resolution. To achieve high resolution, an antenna physically larger than can be practically carried by the platform is required. Radar can collect data while flying over a distance and then process the data as if it came from a physically long antenna, using the flight path of the platform to simulate a large antenna electronically. The distance the aircraft flies in synthesizing the antenna is known as the synthetic aperture. A narrow synthetic beamwidth results from the relatively long synthetic aperture, which yields finer resolution than is possible from a smaller physical antenna.

Today, digitizers and other data acquisition equipment can store data for offline processing or even process imagery in real time. As the object is moving either toward or away from the transmitter, there is a slight equivalent change in the frequency of the radio waves, caused by the Doppler effect. By utilizing the Doppler effect between the platform of the radar and the target, SAR can obtain high-resolution images with small antennae.

SAR systems take advantage of the long-range propagation characteristics of radar signals and the complex information processing capability of modern digital electronics to provide high-resolution two-dimensional imagery. SAR complements photographic and other optical imaging capabilities because of the minimum constraints on SAR of time of day and atmospheric conditions.

Polarimetric SAR

Electromagnetic (EM) fields are vectors, that is, linearly, circularly, or elliptically polarized. Any polarized EM wave can be expressed by dual linearly polarized (i.e. vertical and horizontal) waves. The SAR system that utilizes the polarized EM wave is called polarimetric SAR (PolSAR). For the incidence of a polarized wave from the PolSAR system, the EM fields, both co- and cross-polarized, received by the SAR system are obtained through a $2 \times 2$ complex scattering matrix. The scattering intensity, called the Stokes vector, is a four-dimensional real intensity, and the transform matrix between incident and scattering is a $4 \times 4$ Mueller matrix. PolSAR via these matrices may provide much more information about the observed objects and environment (Lee and Pottier 2009; Cloude 2010; Jin and Xu 2001).

The first spaceborne PolSAR was the Shuttle Imager Radar SIR-C, in 1994. Today PolSAR systems are a popular emerging technique with a high resolution of meters and decimeters.

Bistatic PolSAR

Conventional monostatic SAR has its transmitter and receiver installed on the same platform and collects the backscattered echoes from targets within the imaged area. During the past decade, bistatic SAR with a separate transmitter and receiver flying on different platforms has become of great interest, because this might significantly reduce vulnerability in military systems, enhance the response due to the large angle between incidence and scattering, present additional information, and so on (Chernyak 1998; Wills 1991).

Interferometric SAR

Interferometric SAR (InSAR) is a technique that compares the amplitude and phase signals received during one pass of the SAR platform over a specific geographic area with the amplitude and phase
signals received during a second pass of the platform over the same area but at a different time. The phase difference for each pixel in the resulting interferogram is a measure of the relative change in distance between the scatterer and the SAR antenna. With the collection of complex SAR data from slightly offset perspectives, the separation between these two observation points is termed the “baseline.” This baseline introduces for each point in the scene a slight range difference that results in a phase shift. From this phase shift, a digital elevation model (DEM) in three dimensions can be derived (Gini and Lombardini 2002; Zebker and Goodstein 1986).

The first InSAR system applied to Earth observation was reported in 1974. Since then, a series of satellite-borne SAR have been launched and provide the stable, well-defined orbits and short baselines necessary for InSAR generation of a consistent global DEM and local surface deformation with an unprecedented accuracy. InSAR has also been applied to the long-term study of earthquakes, volcanoes, and so on.

**Tomography SAR**

In the physical principle, SAR maps the three-dimensional reflectivity distribution from a scene of scatterers to be imaged onto the two-dimensional azimuth-range plane. Using several multipass acquisitions on the same scene, the aperture synthesis in the third elevation direction can be performed. It uses data stacks of several acquisitions from slightly different viewing angles (the elevation aperture) to invert or reconstruct the reflectivity function along the elevation direction by means of spectral analysis for every azimuth-range pixel. Hence, the three-dimensional SAR image can be obtained. This multibaseline interferometry with a number of tracks is called tomography SAR (TomoSAR). With SAR tomography, it is possible to separate the ground backscattering in a forest and to estimate height with only signal contributions. TomoSAR has also been applied to high-dimensional multitemporal information, such as seasonal inflation of buildings and deformation of buildings on the ground (Fornaro et al. 2012; Zhu and Bamler 2010).

**Inverse SAR**

Inverse SAR (ISAR) is analogous to conventional SAR, which generally refers to the case where the radar platform is moving while the target stays stationary. ISAR is used for scenarios when the radar is stationary and the targets are in motion, such as with airplanes, ships, tanks, and so on. Stationary radar collects the scattering data from the target at different look angles by utilizing the target’s movement. While the target is moving, the look angle of the target is assumed to be changing with respect to the radar line of sight (RLOS) axis in order to produce a successful ISAR image. This angular diversity in the ISAR dataset is used to resolve different points along the cross-range axis.

Additionally, the ISAR imaging procedure has some conceptual differences when compared to SAR imagery. Motion compensation is the first step in the ISAR image reconstruction chain. Several algorithms have been studied to accomplish motion compensation. After motion compensation, the received signal is processed to form the ISAR image. ISAR technology has been extensively applied to ATR (automatic target recognition). It is capable of displaying dominant characteristic scattering regions (hot points), that is, scattering centers on complex targets, such as satellites, spacecraft, airplanes, ground vehicles, and so on (Özdemir 2012; Fortuny 1998).
Radar sounding

In geophysical studies, such as studies of the ionosphere, land subsurface, and so on, the radar sounding system has been broadly applied. A radar sounder generates high-frequency (usually polarized, approximately MHz) radio waves to transmit and penetrate the underlying media (through approximately meters and kilometers). The radar echoes from the surface/subsurface inhomogeneity with different time delays may provide the structural characteristics of the media, for example, layer thickness, dielectric constant, inhomogeneity, and so on (Joe 2009).

Usually, the echoes directly from the surface and subsurface can be seen as signals, and the clutters from neighbor surface roughness and media inhomogeneity (scatterers) are seen as noise to disturb the signals. Analyzing the range of images from the radar echoes and clutters and then deriving their functional dependences on the media parameters can present structural information about underlying complex media. These technologies have been applied to planetary explorations, for example, of the lunar and Mars surface/subsurface (MARSIS, Mars Advanced Radar for Subsurface and Ionosphere Sounding; and SHARAD, Shallow Radar), detection of land mines and pipes (i.e., GPR, ground penetration radar), and so on (Mouginot et al. 2010; Picardi et al. 2005).

SEE ALSO: Ground-based radar; Microwave remote sensing; Synthetic aperture radar

References


Radical geography

Jenny Pickerill  
*University of Sheffield, UK*

Radical geography is an approach to geographic research that seeks to understand social and spatial problems, and to advocate solutions. Radical geography is not a coherent or unifying concept or methodology. There is great diversity in the research that would be considered “radical.” Radical geography is an approach developed in the 1970s and a contemporary ethos that is fundamental to many geographers’ work. It should be considered as a multiple, dynamic, and contested approach, for which there is no one definition. Instead there are radical geographies.

Radical geographers are interested in everyday lives: the lived experiences of members of society. They are interested in issues of relevance to everyday social life, such as access to safe and affordable food and housing, fair pay, educational opportunities, and basic health care, to name just a few. This emphasis reflects a desire for geographers to do research that is relevant to, and useful for, society.

There has been a specific focus on understanding spatialities of power, inequality, and oppression, which has required an understanding of the causes of such inequality and has led to research on power, neoliberalism, political structures, and corporate hegemony. It has been necessary to understand how neoliberalism and corporate power operate and to relate the everyday experiences of people to the broader structures that shape their lives.

There is an implicit left-wing progressive political agenda to much radical geography work, which prioritizes the oppressed, powerless, or marginalized in society. Such work often aims to transform the world through an emancipatory politics. The political agenda of this work has often been to advocate for a fairer and more equal society where everyone’s basic needs are met and everyone has an equal opportunity to participate. Consequently, radical geography has largely ignored research on elites, the powerful, or the wealthy, and issues that might be considered frivolous or of concern only to the privileged.

Once the causes of social problems have been identified, radical geographers take a normative approach. This means that they seek to offer solutions, alternatives, and possibilities for how the problem can be solved. A quest for social justice permeates much of radical geography. Radical geographers, who are often active participants in social justice campaigns, are also committed to building, supporting, and intellectualizing alternatives. In other words, radical geography is as much about offering alternatives as it is about understanding causes. Contemporary examples include Jane Wills’s work in supporting the campaign for a living wage in London and beyond, Gibson-Graham’s work on diverse economies, and Marc Purcell’s work on direct democracy.

A brief history and ideological foundations

Radical geography emerged in the early 1970s both in response to a changing social context and as a critique of the positivism of the earlier
RADICAL GEOGRAPHY

quantitative revolution in geography. The social context was a growing unease about social and economic inequality, poverty, racism, crime, sexism, and environmental issues, typified in dissent against the Vietnam War and support for antiracist and anticolonial movements. At the time the quantitative approaches and positivist epistemology of the 1950s and 1960s began to be considered as irrelevant to these social problems and as being inadequate in offering solutions or alternatives. Geographers wanted not only to identify social problems, but also to understand their cause and thus identify different possibilities. It was increasingly understood that merely identifying problems was not enough, and that geographers needed to be proactive in advocating change and to do so on the basis of clear theoretical and political foundations.

The shift to radical geography also represented a greater use of philosophical approaches other than spatial determinism to understand the world. Both Marxism and anarchism were employed as ideological underpinnings for the new radical geography. Crucially, these ideologies provide diverging and contested understandings of society, a debate that is ongoing and not limited to academia, least of all academic geography. The huge philosophical differences between Marxism and anarchism represent just one example of the proliferation of approaches that fall under the remit of radical geography, which, as a result, can differ widely and contradict each other. The study of these ideologies does not constitute radical geography: rather, it was the way in which these philosophies enabled geographers to identify action and alternatives to the current order that shaped radical geographical practices.

Such ideological underpinnings existed long before the 1970s, and these quite different ideological approaches have existed side by side within the discipline since the nineteenth century. Karl Marx and Friedrich Engels, although not formally geographers, provided a framework through which class relations and capitalism could be understood, and this became the basis of Marxist geography. Engels’s *The Condition of the Working Class in England in 1844* (1845) was a treatise on the poor urban conditions experienced by the working class. Geographers later took up this project, notably Karl Wittfogel (in Germany), who linked Marxism with geopolitics and materialism in a 1929 paper. In France, post–World War II geographers such as Pierre George and Jean Tricart developed Marxist spatial analysis.

The Marxist tradition was reignited by the work of David Harvey, whose *Social Justice and the City* (1973) advocated a Marxist approach to understanding the world. For Harvey this involved geographers engaging in both practical and theoretical work, which was reflected in his own engagement with political campaigns for social justice in Baltimore. Around the same time, Richard Peet was using political economy and Marxist geography to explore social justices, and the journal *Antipode: A Radical Journal of Geography* was established. These developments also built on the Detroit Geographical Expedition and Institute (1968) created by William Bunge. Bunge wanted to expand the processes and use of geographic knowledge, so he enrolled minority communities in the whole research process, beginning a radical geography approach that would later be developed into a participatory geography methodology.

This Marxist approach influenced radical geographers by offering an analysis of the world based on modes of production and by advocating normative practices. Marxist geography was interested in the dialectical relationships between humans and the modes of production that shape society: capital, labor, capitalism, the
state, class, and the market. These various modes of production were determined by Marxists to create the superstructure of society. As Derek Gregory (1978) pointed out, these ideas shaped the structure–agency debate; the idea that these structures determined society was challenged by Gregory’s assertion that humans had agency in the creation of structures. Agency varied across place, but led to the possibility that humans could challenge and change societies’ structures, thereby leading to social change.

Marxist geography focused on an analysis of modes of production as an explanation of social and environmental problems. It was interested in the inherent contradictions of capitalism. For example, early radical geographers linked environmentally destructive practices, such as old-growth forest logging, as originating in the structures of capitalism. For Marxist radical geographers, most social problems and injustices emanated from the workings of modes of production. Therefore, if the mechanics of production were changed, alternative outcomes would be possible for society and the environment.

The normative approach offered by Marxism encouraged geographers to be bolder in advocating solutions and alternatives. Normative research aims to produce work that has practical use for society. Normative work identifies actions to be taken and is assertive in its support for particular activities. Radical geographers use their research to justify and make interventions to shape certain futures. This radical approach was also mirrored in other social science disciplines, such as the emergence of radical sociology in the 1970s which also advocated human emancipation as its goal.

While Marxist geography became the dominant radical geography approach, anarchist geography proposed an alternative vision. Anarchist geography was established by Peter Kropotkin and Élisée Reclus, who were both formally trained as geographers but were also anarchist philosophers. Reclus’s 32–volume work, The Earth and Its Inhabitants: The Universal Geography (1876–1894) sought to abolish domination, including that of humans over the environment, by a merger between humanity and the Earth, creating equality between humans and nonhumans. Kropotkin published numerous articles in geography journals, pamphlets on anarchism, Fields, Factories and Workshops and The Conquest of Bread, and the hugely influential Mutual Aid: A Factor in Evolution (1902), in which he argued that cooperation rather than competition was the natural order of things. He advocated a life based on decentralization, localism, and self-provision. In 1855 he wrote “What Geography Ought to Be,” a radical anarchist agenda for the discipline.

Anarchism is a shifting set of beliefs adopted by a variety of groups which, at its core, rejects the need for rulers, domination, hierarchy, and any form of government. It provides a critique of power and hegemony and an alternative vision of liberation and autonomy. It was not until the 1970s, however, that geographers re-engaged with the work of anarchist theorists such as Proudhon, Kropotkin, Bakunin, and Goldman, to name just a few, to explore the potential for anarchism, find examples of anarchist practices in places, and advocate the adoption of anarchist principles. They also sought to understand the limitations of the state and how state and capital operate together as forms of control and containment. In particular, anarchist geographers such as Paul Chatterton have explored how the state seeks to limit autonomy and equality and the possibilities to carve out freedom nonetheless.

While anarchist geographers practice normative approaches to their work, much of their worldview contrasts sharply with that of Marxist geographers. Anarchist geographers have sought to illustrate the power and possibility of grassroots
radical politics and action. Anarchist scholars argue that each individual, rather than being passive in a hegemonic superstructure, has the power to change society, albeit most effectively as part of a mass mobilization. The social change advocated by anarchist geographers is one without leaders, government, or hierarchy and, crucially, is not reliant on a revolution (as in the Marxist lexicon). As a result, research by anarchist geographers has focused on analyzing the potential of many currently small-scale anarchist initiatives such as workers’ cooperatives, eco-villages, anarchist communities, alternative food networks, and online open publishing models. In each case, not all examples are overtly anarchist but many are imbued with anarchist principles.

Among other anarchists, Pierre-Joseph Proudhon offered a critique of property, Mikhail Bakunin critiqued the need for a state, Emma Goldman extended our understanding of domination beyond the state and advocated equal gender relations, and Colin Ward argued against planning restrictions and for more affordable and self-build housing. These anarchist principles include an emphasis on cooperation (mutual aid), direct democracy, prefigurative action, nonhierarchical decision-making, free movement, self-organization, and communalism (sharing). Mutual aid is the practice of helping one another and sharing skills, time, and resources freely. It is the belief that individualism and competitiveness are ultimately destructive in a society and that we could achieve much more if we worked together. Direct democracy and nonhierarchical decision-making mean that everyone should have an equal voice in the decisions made in society and that there are no leaders. Prefigurative action is the embodiment of acting now to make the world as we want it to be, rather than the Marxist approach, which advocates waiting for the revolution before we start to live differently.

In the 1970s geographers such as Richard Peet (1978), Myrna Brietbart, and Bob Galois began to use these ideas to call for a more anarchist radical geography. However, after a special issue of *Antipode*, there was little geographical engagement with anarchism until the 2000s. In the late 1990s interest in anarchist and autonomous anti-capitalist protests led to work by Paul Chatterton, Pierpaolo Mudu, Jill Fenton, and Jon Anderson. Central to all this work was an understanding that social transformation is a spatial project.

There were several challenges to anarchist and Marxist approaches to radical geography. Marxist geography was criticized for its overly deterministic view of the world, where individuals were assigned a passive role within a dominant and hegemonic superstructure. This metatheory was preoccupied with the relations between capital and class, while it ignored other dimensions that shape people’s lives, such as patriarchy.

The absence of attention to gender was countered by the emergence of feminist geography in the 1980s. At first, feminist work in geography identified and explored the absence of women in geographical scholarship and their institutional absence within universities. Such a critique exposed a masculinist tendency within academic geography that devalued women’s role in society and ignored their difference and oppression. At the same time, other scholars were becoming critical of the Marxist approach as a result of the changes in Eastern Europe and the disintegration of the Soviet Union, which for many undermined the possibilities of a Marxist future.

By the turn of the century, radical geography was more diverse and inclusive of a greater range of both philosophical approaches and research topics. It began to focus less on a Marxist critique of capitalism or an anarchist approach, and more on the diverse factors that shaped people’s lives, such as gender, sexuality, disability, and race. Marxism could not accommodate this diversity.
of identity politics. The postmodern critique of metanarratives led to a poststructuralist approach within human geography that persists. In the 1990s the terminology of radical geography fell out of favor and was replaced by scholars self-identifying as “critical human geographers.” The establishment of the Critical Geography email forum and the International Conference of Critical Geographers was motivated in part by a rejection of these metanarratives (particularly Marxism).

Radical geography did not entirely reshape the discipline in the same way that the quantification revolution had: its impact was far more partial and fractured. There is now concern that a lack of a clear framework of analysis, or worldview, and the shift toward poststructuralist geographies and the cultural turn, have limited the ability of radical geography to be normative and to advocate clear alternatives. At the same time, there is an ongoing threat from the neoliberal political project, reducing the possibilities and options for alternatives.

**Inequality, capitalism, and marginalization**

Radical geography is grounded in a basic concern for everyday lives and in identifying and making visible the geographical processes that shape and produce inequality, oppression, and marginalization. There are many radical geographers, but here three examples have been picked out to illustrate some of their intellectual and normative contributions to the founding dimension of the field—research on inequality. This work has tended to highlight the inequity of neoliberalism and its unjust consequences, and thus that capitalism is the structural cause of inequality.

The late Neil Smith’s work is exemplary of (Marxist) radical geographers’ engagement with issues of inequality, particularly in cities. His early work, *Uneven Development: Nature, Capital and the Production of Space* (1984), argued that capitalism produced uneven spatial development. He went on to critically identify gentrification as a process of capitalism that benefited from low urban land prices and capitalist speculation in *The New Urban Frontier: Gentrification and the Revanchist City* (1996). Smith developed the term “rent gap,” a concept used by anti-gentrification activists to help explain and fight against gentrifying processes. Indeed, much of his work was used politically and by activists, and Smith worked with many activists over his lifetime, especially the Harlem Tenants Association. He was always concerned with the importance of class to these processes of inequality. Later in his career he drew inspiration from the antiglobalization protests of the late 1990s and the revolutionary potential of the Occupy movement and the so-called Arab Spring. Smith became increasingly critical of the topical focus of contemporary geography, particularly the middle-class obsessions of anglophone geography, such as the concern with the community gardens in New York, rather than recognition of how social processes produced such a geographically uneven world (Smith 2005).

Doreen Massey (1984), another Marxist geographer, was also interested in uneven development and developed the theory of “spatial divisions of labor.” This approach posited that uneven development was a result of the different ways in which the capitalist economy engaged with space and spatial relationships. The outcomes were social inequalities that varied across space, creating a divide between rich and poor regions and between classes. For Massey, understanding space was crucial to understanding inequality and poverty, and once this was grasped it was then possible to identify solutions.
The final example is the work of Jane Wills, an expert on work and labor politics, who in 2005 worked with the campaign group London Citizens to develop the case for the living wage. Wills argued that, although Britain has a basic minimum wage, many people are in in-work poverty: despite being in full-time employment, they do not have enough money to cover their basic living costs (Wills and Linneker 2014). Using economic geography, Wills worked with community organizers to build a business case for a living wage, starting with the example of cleaners at Queen Mary University of London. Their aim was to reverse the trend for global outsourcing, which saw a reduction in the wages and benefits of staff but increasing profits for companies. This social justice work was expanded to a UK-wide campaign in 2011 through the Living Wage Foundation. Wills’s achievements not only illustrate the possibilities of change as a result of radical work, but also the importance of academics working with community organizers in a collaborative bottom-up process.

Activism and social change

Radical geographers’ links with political activists have extended beyond concerns with class and inequality into diverse geographical research of resistance and social change. Many radical scholars work with those who are campaigning, protesting, and instigating social change across a broad range of interests, including poverty, health, food, migration, environment, land rights, gender equality, queer activism, racism, and other social justice issues.

There are several dimensions and approaches to this work. Some scholars, like Paul Routledge, Kye Askins, Jenna Lloyd, Kelvin Mason, and Paul Chatterton (to name just a few), have sought to bring geographical concepts to the aid of activists, and to directly participate in activist struggles. For example, Paul Routledge (2003) has worked extensively with activist communities in India (among other places) on environmental issues and anticapitalist resistances, developing the concept of “spaces of convergence” to understand the global processes of antiglobalization protest. Jenna Lloyd (2014) has explored health activism and in other work critiqued the criminalization and militarization of migration.

As part of the Autonomous Geographies Collective, Paul Chatterton worked alongside Jenny Pickerill and Stuart Hodkinson to explore autonomous activists’ daily lives. This work involved active participation in British anticapitalist activism to understand and advocate for those spaces where there is a questioning of laws/social norms, and a desire to create non-/alter-capitalist, collective forms of politics, identity and citizenship. The collective’s work was able to identify the value of these contested, fractured, interstitial spaces of autonomy. In many ways this work built on the anarchist foundations of radical geography by exploring practices of mutual aid, solidarity, self-management, decentralized and voluntary organizing, and direct action. Chatterton went on to found and research the first affordable urban ecological cohousing community in Britain.

Others have taken a more discursive approach to researching activism, exploring the ways in which, for example, emotions are employed (Fernando J. Bosco) and solidarity expressed and practiced (Gavin Brown), and the gender politics of protest, violence, and militarization (Melissa W. Wright).

This type of research often involves difficult issues around ethics: protecting research participants, illegal activities, and power relations (between resourced privileged academics and underresourced marginalized activists) (Autonomous Geographies Collective 2010).
There remains a tension in much of geographers’ work between supporting and advocating diverse forms of activism and the intellectual need to critique and interrogate activist practices. There are numerous problems inherent in all forms of activism, most commonly around gender relations, measures of “success” (the global capitalist economy often prevails), the temporality of activism, the ghettoization of activist communities, and the time required for alternatives to be created; however, intellectual critique can be viewed as undermining the intentions of collective action and as politically unhelpful. Many radical geographers negotiate such tension by only offering constructive critique and limiting their intellectualization to academic journal publications.

**Participatory approaches**

In the past decade there has been a resurgence of interest in radical methodologies. Recently, radical geographers have advocated wider use of participatory methods. Participatory research methods start by asking groups outside the academy what research they think is required, how it should be conducted, and what outcomes they would like to see. Academics then work with these participants to help them conduct the research. It is a bottom-up approach to research and is based on the premise that research should benefit all, particularly those who have participated in it. This approach attempts to stretch the normative element of radical geography to thinking about who benefits from geographers’ work.

Participatory geography is concerned with empowering people to change their own lives (with the aid of researchers). This requires being proactively inclusive, with researchers seeking out those who might be excluded and taking a detailed approach, rather than categorizing people into groups. There tends to be a focus on disadvantaged groups and in challenging existing power relations. It is also a very reflective process that requires the researcher to self-critically reflect on their actions, learning, and positionality, and to value the expert knowledge of nonacademics. Key advocates of participatory geographies have included Rachel Pain, Sara Kindon, Mike Kesby, Kye Askins, Duncan Fuller, and Caitlin Cahill. In Britain, radical geographers’ support for participatory methods led to the establishment of a new Royal Geographical Society group, the Participatory Geographies Research Group (known as PyGyRG) in 2005 (though it formally became a research group only in 2009).

Participatory geography is simple in its premise – geographers should work with those they are researching and include them in every aspect of that work – but the approach simultaneously critiques contemporary academic practices and knowledge production. In valuing the contributions of nonacademics, participatory geography challenges the presumption that knowledge is best constructed by the academy, and raises questions about the purpose of university research. In the radical tradition, it also privileges work that is more likely to result in action, political change, and demonstrable improvements in society, rather than the highly theoretical esoteric work which speaks only to other academics.

**Grassroots solutions and alternatives**

Since the early years of radical geography, scholars have sought to suggest solutions and alternatives. These alternatives have tended to counter neoliberal logic with different value systems, based on, for example, environmentally sustainable, ethical, participatory, and/or communal processes. Three examples illustrate the
ways in which radical geographers have engaged with, and advocated for, grassroots solutions.

Gibson-Graham (Katherine Gibson and Julie Graham) (1996) developed the concept of diverse economies, which challenged conventional notions that the economy was made up only of formal commodity markets, waged and salaried labor, and capitalist enterprises focused on creating profit for owners or shareholders. Instead, Gibson-Graham proposed that the economy was made up of market, alternative market, and nonmarket transactions and that many of the most important goods and services are taken or exchanged outside the formal economy. This expansion of the economy celebrated its diversity and valued these different forms of transactions, labor remuneration (such as unpaid work), and forms of enterprise. In this approach, the neoliberal market was problematized as not being the main or dominant form of trade, and the potential of alternatives to support society was more visible.

In a similar way, Nik Heynen’s work on civil disobedience and protest in the United States celebrates the potential of activist practices. Using “Food Not Bombs” as one of his examples, Heynen (2010) not only identifies the contradictions of neoliberalism, but critically explores the alternative survival strategies proposed by collective action. Finally, Peter North (2007) advocates the use of alternative currencies while acknowledging their inherent limitations.

Beyond Anglo-American perspectives and sources of knowledge

Radical geographies have been most noticeable (though rarely dominant) in anglophonic geography. Although it has spread beyond North American and European geography, it has tended to be limited to the settler countries of Australia, New Zealand, and South Africa, in great part due to the dominance of the English language in radical geography (and human geography per se). At the same time there has been a lack of attention to the radical geographies of non-Anglo spaces of knowledge.

Manuel Aalbers (2004) identifies this dominance of English as resulting in the “creative destruction” of many non-English speakers’ work. In essence, the peer review process excludes work from beyond anglophone geography by requiring a certain written style and the referencing of key anglophone texts. This process excludes alternative knowledges.

Radical geographers have recognized the problems of such exclusions and the irony of radical geographers being complicit in these processes, given their apparent interest in alternative sources of radical ideas. Lawrence Berg, who established the open source radical journal ACME: An International E-Journal for Critical Geographies, attended to these problems by encouraging the submission and publication of articles to ACME in English, Spanish, Italian, French, or German, and by having a multilingual website. The journal Antipode has also started a project that will fund the translation of accepted articles.

Practicing what is preached

There remains a disjuncture for many geographers between the radical geographical scholarship they conduct, and their own elite positions within privileged educational institutions. For anarchist geographers who support the concept of taking prefigurative action, this can be particularly problematic.

Academic geographers, particularly those in the United States, the United Kingdom, and Australia, are experiencing the consequences of universities’ adoption of neoliberal approaches
to higher education. This is not a new dilemma; there have long been pressures on universities to commercialize and to serve the needs of industry and the political class rather than those of society. However, there are increasing pressures on academic practice, including national accounting measures of performance (such as Research Excellence Framework (REF) in the United Kingdom and the Excellence in Research for Australia (ERA)); profit-driven journal production; the casualization of staff (Purcell 2007); and reductions in grant funding. Such processes have tended to undervalue radical scholarship and participatory empirical work, and instead to valorize individual theoretical work that can be produced and published more quickly. The quantity of output has been valued over quality, and in struggling to secure a permanent faculty position geographers have made compromises. Yet radical geographers have complied with the research assessment processes, and many have excelled in it.

Many radical scholars have identified such pressures and fought to counter them. Neil Smith sought to challenge the “sausage factory” processes of academia. Katharyne Mitchell (2008) celebrated geographers’ scholar activism. Noel Castree, Don Mitchell, and Marc Purcell wrote about, and attempted to resist, the neoliberal assault on universities. In attempting to subvert profit-driven journal production, open access journals like ACME and Human Geography: A New Radical Journal have been established, and Antipode renegotiated its contractual relationship with publishers Wiley-Blackwell, using the additional income to set up a charitable foundation that funds radical geography research.

Recognition of radical geographers’ conflicted positions within neoliberal institutions has raised questions for academics about their power and compliance with such processes. Academics such as Marc Purcell have argued that we need to better challenge the political and social relations that shape academic practice. In other words, radical geographers have to resist neoliberalism within universities if they are to be taken seriously as researchers and activists when working against neoliberalism in society. Practicing what radical geographers preach is becoming increasingly difficult.

Future directions

It is doubtful whether radical geography has achieved its aims. Many social and environmental problems persist and are actually getting worse. Postpositivist theoretical achievements, unfortunately, pale into insignificance in an era of unconstrained inequality and a neoliberal privilege that remains unchecked despite attempts by democratic governments to impose some measures of control. For example, attempts to restrict the practice of awarding large bonuses to bankers in the City of London have largely failed. These right-wing times make radical scholarship harder. Some future possible directions include more diverse radical geographies, a feminist ethics, participatory methods, radical positivism, and understanding the elite.

Radical geographies are becoming more diverse, but they remain anglophone and steeped in either the Marxist or anarchist tradition. As journals and academics begin to operate in more languages, there are possibilities for further diversifying radical geography. Geographers are also increasingly questioning their understanding of knowledge, how knowledge is shaped, and how they represent the world, which opens up the possibility of other knowledges being accommodated. In terms of identity, it matters less whether scholars self-identify as radical geographers than that the practices – political and normative – continue
to spread and evolve. There is concern from academics such as the late Neil Smith that anglophone geography is becoming too conservative, and thus it is especially important to continue the debate on the role of politics in geography (Orzeck 2012).

Radical geography has, unfortunately, retained a certain masculinist edge, indicative perhaps of its Marxist heritage. Linda McDowell (1992, 56) originally identified this as being both “inside and outside ‘the project.’” Despite decades of feminist geography, Victoria Lawson (2009) argues that radical geography still ignores the feminist care ethics, an ethics that is central to human wellbeing and society. This marginalization of care, and gender and bodies, limits radical geographers in understanding inequalities, society, and the possibilities of alternatives to create new caring relations. Many academics still feel outside “the project,” and continued effort is needed to ensure that radical geographies are plural, dynamic, and open.

Although participatory methods are now being broadly accepted in human geography, there remains a need to stretch what a participatory ethics does. Issues of unequal power relations remain, and geographers need to explore ways of relinquishing ownership of the data and the final outcomes of research, a challenge that could be particularly difficult given the ongoing neoliberalism of universities, but working out ways in which participants retain control would be radical.

Wyly (2011) argues that this failure of radical geography to achieve social change can be countered by a form of radical strategic positivism. While it would be helpful to return to a notion of a universal truth, there is a need for radical geography at times to be able to generalize, to employ methods that others can replicate, and to have the appearance of impartiality and measurement. In part this is necessary because of an increasing lack of trust in the scientific process, as illustrated by climate change deniers, the rise of creationism teaching, and the belief that the free market provides equal opportunity of wealth for all. This new form of radical geography includes Sheppard’s (2001) notion of “insurgent quantitative practices,” using statistics to counter official use where they are all too often used to justify a neoliberal conservative agenda. This approach challenges how statistics are calculated, their levels of accuracy, conflicts of interest, and interpretations. There is also the use of constructive cartography, counter-mapping, and critical mapping, such as Sidaway’s (2010) maps of “black sites” and renditions used by the United States after the terrorist attacks of September 11, 2001, which employ positivist approaches but for radical ends. With these approaches it is possible to use radical methodological positivism to translate radical geographic ideas into progressive political agendas.

The dispossessed and marginalized in society have been researched at the cost of understanding the elite in society. Given the normative underpinnings of much radical geography scholarship, this is understandable; however, research to better understand elites, the privileged, and those with power would strengthen the radical geography project.

**SEE ALSO:** Anarchist geography; Critical geography; Feminist geography; Marxist geography

**References**


Further reading


Radiometric dating/techniques

Phillip L. Larson
Minnesota State University, USA

Ronald I. Dorn
Arizona State University, USA

Geographers both develop and utilize four general types of techniques to determine ages of Earth’s deposits and features: calendar ages, relative ages, numerical ages, and correlative ages. All of these techniques require the ability to measure changes over time. Biological systems can be measured by such approaches as counting annual tree rings. Chemical systems can be measured, such as changes that might occur in soils through the accumulation or depletion of minerals. Physical systems can also be measured through such constants as the decay of radioactive isotopes.

The most precise and accurate techniques assign calendar ages to events through historical record keeping (Figure 1), the use of annual tree rings (dendrogeomorphology) (Butler, Malanson, and Oelfke 1987), or counting annual layers (varve chronology) in lake sediment. The least precise dating method simply orders events in a relative sequence from oldest to youngest, such as soil development over time or a sequence of superimposed river changes (avulsions) (Figure 1). Although no specific age is assigned, relative ages are often used to detect problems with more precise techniques when results contradict a relative sequence. Numerical ages are assigned most often using radiometric techniques, such as radiocarbon dating, and typically have ± uncertainties based on measurement errors. Correlative ages derive from techniques that match an event that has a prior numerical age, such as tephrochronology that correlates the mineralogy and geochemistry of a particular volcanic eruption deposit to a volcanic ash deposit that already has a numerical age. Table 1 presents just a few of the dating methods used in physical geography.

The selection of the right dating technique for any particular project will usually hinge on a series of questions.

- What chronometric information do I need to answer the question?
- What material is available for dating?
- What method(s) are the most appropriate?
- What does the age represent?
- How do the precision, accuracy, and inherent uncertainties influence the ability of the age result to answer the research question?
- How much will the dating research cost?

Since it is not possible to review the entirety of the dating methods presented in Table 1, this entry instead illustrates two different case studies of landscapes where these key questions are being asked, and where a variety of dating methods have been (and are being) employed. One case study involves how geomorphologists are establishing ages of glacial and alluvial landforms to understand the Quaternary history of the Minnesota River Valley and its genetic relationship to proglacial Lake Agassiz. The other case study involves ongoing efforts to understand the glacial chronology of the Sierra Nevada of California. The idea behind presenting these case studies...
Figure 1  The Mississippi River just north of the Atchafalaya River undergoes avulsions frequently, where the relative sequence of events can be seen in the satellite image. (a) Map, 1944; (b) satellite image, September 2, 1999. The most precise dating technique, of assigning calendar ages, required precise record keeping (http://lmvmapping.erdc.usace.army.mil/index.htm). Image source: http://earthobservatory.nasa.gov/IO/TD/view.php?id=6887. Map courtesy the US Army Corps of Engineers, Landsat image by Robert Simmon, based on data from the UMD Global Land Cover Facility.
Table 1  Selected examples of dating methods used in physical geography.

<table>
<thead>
<tr>
<th>Type of age</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calendar</td>
<td>Dendrogeomorphology employs annual growth rings in trees to place a minimum age on the underlying landform. Its great accuracy facilitates use in analyzing erosion rates and alpine hazards such as avalanches on timescales of $10^0$–$10^3$ years (Butler, Malanson, and Oelke 1987).</td>
</tr>
<tr>
<td>Calendar</td>
<td>Sclerochronology studies changes in the accumulating tissues of such organisms as corals and mollusks. The annual growth rings in corals offer a powerful tool in climatic change and coastal geomorphic research on timescales of $10^5$–$10^4$ years.</td>
</tr>
<tr>
<td>Numerical</td>
<td>Radiocarbon measures the abundance of $^{14}$C in samples such as charcoal, wood, peat, shells, and carbon–containing minerals such as oxalate and calcite. The method assumes that the carbon-containing material stopped exchanging $^{14}$CO$_2$ with the atmosphere upon death or deposition. Radiocarbon has been used in almost every geomorphic setting including fluvial, colluvial, coastal, glacial, and eolian materials deposited during the last 40 000 years. Radiocarbon ages are typically translated into calendar ages through the use of calibrations.</td>
</tr>
<tr>
<td>Numerical</td>
<td>Optically stimulated luminescence (OSL) measures when sediments were last exposed to sunlight (e.g., transported and then buried). A light-sensitive signal that is “trapped” over time in crystal defects is released upon irradiation. This technique has seen an explosion of use, because of its utility in establishing ages on timescales of $10^0$–$10^5$ years for noncarbonate deposits such as eolian, fluvial, coastal, glacial, and colluvial settings.</td>
</tr>
<tr>
<td>Numerical</td>
<td>Cosmogenic surface exposure dating relies on cosmic rays interacting with rocks to create new nuclides over time. These include stable nuclides that just build up ($^3$He, $^{21}$Ne) and radionuclides that both build up and also decay at different half-lives ($^{10}$Be, $^{26}$Al, $^{36}$Cl). Through careful sampling and keeping track of the variables that alter build-up rates (e.g., altitude, latitude, burial history, surrounding topography), outputs include a combination of erosion rates and exposure ages, and combinations of nuclides can date surfaces on the timescale of $10^2$–$10^6$ years.</td>
</tr>
<tr>
<td>Numerical</td>
<td>Cosmogenic burial dating relies on the build-up of cosmogenic nuclides in rocks prior to deposition. After deposition, these nuclides decay at different rates, allowing the calculation a burial age on the timescale of $10^3$–$10^6$ years. This technique can estimate the ages of marine terrace, fluvial terrace, and even cave deposits previously undatable due to a lack of suitable materials (Granger and Musikar 2001).</td>
</tr>
<tr>
<td>Numerical</td>
<td>Uranium-series (U-series) dating typically uses a number of different decay pathways, but typically the $^{238}$U–$^{234}$U–$^{230}$Th–$^{226}$Ra system with half-lives of 245 000, 76 000, and 1600 kyr, respectively. Coral and speleothem deposits can yield very precise ages, but other materials are often not in truly closed systems typically applied to carbonates with timescales of $10^0$–$10^5$ years.</td>
</tr>
<tr>
<td>Numerical</td>
<td>K–Ar dating and the more precise $^{40}$Ar/$^{39}$Ar dating methods measure when liquid magma solidified through the decay of radioactive $^{40}$K into stable $^{40}$Ar. This method can determine the ages of lava flows and tephra on timescales of $10^3$–$10^8$ years.</td>
</tr>
</tbody>
</table>

(Continued opposite)
# Radiometric Dating/Techniques

<table>
<thead>
<tr>
<th>Type of age</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlated</td>
<td><em>Lichenometry</em> compares the growth of lichens (often the maximum diameter of particular species) on surfaces of known age (e.g., cemetery stones) to develop a growth curve for an area. This growth curve then allows age estimates for such features as glacial moraines, debris avalanches, and other features of cold dry climates over timescales of $10^1$–$10^5$ years.</td>
</tr>
<tr>
<td>Correlated</td>
<td><em>Amino acid racemization (AAR)</em> uses the protein residues in fossil carbonate that degrade through time-dependent reactions. While alive, organisms form L-configured amino acids (levorotatory) that re-equilibrate via the racemization reaction to the D-configuration (dextrorotatory). Although factors other than time (e.g., temperature, moisture, species) can also influence the D/L ratio, calibration via independent dating techniques and modeling the racemization factors makes it a powerful tool in understanding landforms on timescales of $10^0$–$10^7$ years.</td>
</tr>
<tr>
<td>Correlated</td>
<td><em>Tephrochronology</em> “fingerprints” the unique mineralogical and geochemical signature of volcanic ashes (tephra) – allowing the correlation of events over broad areas impacted by an ancient eruption. Since many ashes have precise radiometric ages, tephra events can be used to assign maximum ages for sediment deposited on tephra and minimum ages for the surface underneath the ash (Lowe 2011).</td>
</tr>
<tr>
<td>Correlated</td>
<td><em>Varnish microlaminations (VML)</em> correlate layers found within rock varnish to paleoclimatic events. Calibrated by radiocarbon, K–Ar, and cosmogenic ages, varnish microlaminations deliver millennial-scale minimum ages for the exposure of the underlying rock (Liu and Broecker 2013). This technique is limited to warm deserts, where rock varnish is biogeochemically stable. It can provide minimum ages for such landforms as alluvial fans, debris flows, and bare rock surfaces and petroglyphs on timescales of $10^2$–$10^5$ years.</td>
</tr>
<tr>
<td>Relative</td>
<td><em>Topographic position</em> provides physical geographers important guidance in understanding a relative sequence of events. Examples include: superimposition of a younger glacial moraine over an older moraine; a younger river course cutting off the meander scar of an older channel; lower/younger river (fluvial) terraces inset inside older terraces; younger marine terraces lower than older marine terraces; or a younger sand sheet burying a lower older dune deposit. Such relative sequences provide great power in checking the outcomes of correlated and numerical methods that can yield far more precise results – but if done incorrectly can generate inaccurate data.</td>
</tr>
</tbody>
</table>

is to provide the reader with a feel for the art behind the science of dating geographic features.

## Lake Agassiz Case Study

Episodic meltwater discharges from proglacial Lake Agassiz are a possible trigger for periods of abrupt Quaternary climatic change (Licciardi, Teller, and Clark 1999). Understanding the evolution of this ancient lake, the timing of spillway activation, and the processes associated with each spillway’s discharge event(s) has prompted several decades of research within and bordering the Lake Agassiz basin (Fisher 2003). Dating methods have played a crucial role in developing
our understanding of the timing and location of these episodic events – although uncertainty still exists (Fisher 2003).

The initial outlet of Lake Agassiz was to the south through the Big Stone Moraine (BSM) and Glacial River Warren – ultimately draining into the Mississippi River and Gulf of Mexico (Fisher 2003). The underfit nature and boulder strewn headwaters of the modern Minnesota River Valley (MNRV) reflect the catastrophic nature of the River Warren floods (Fisher 2003). In order to establish when the MNRV spillway was active, Fisher (2003) applied \(^{14}C\) dates from wood and seed fragments in two core locations in Big Stone Lake and the Browns Valley fan in the headwaters of the MNRV (Figure 2). He concluded that there were two major episodes of outflow here – the first ending approximately 10,800 \(^{14}C\) years BP (12,740 calibrated years BP) and the second ending approximately 9,400 \(^{14}C\) years BP. These periods of active flooding were separated by lacustrine deposits in the cores dating from 12,500 to 12,340 calibrated years BP, which reveal an approximate window of abandonment of the MNRV spillway. The initial abandonment of the MNRV spillway coincides with a period when the Agassiz outlet is thought to have shifted to lower elevations and likely emptied eastward into the great lakes via Lake Superior (Licciardi, Teller, and Clark 1999) – although some controversy exists about this hypothesis.

The resulting post-River Warren landscape of the MNRV has been a dynamic tale of erosion and fluvial response to catastrophic flooding that induced approximately 65 m of base level lowering. These floods carved out most of the modern MNRV. Archeological studies conducted through the Minnesota Department of Transportation collected \(^{14}C\) ages from organic materials in MNRV bottom alluvial fans and fluvial sediments beneath the fans. These measurements suggest that much of the MNRV incised close to its present depth by 10,400 \(^{14}C\) years BP (Fisher 2003).

Surprisingly, very little is known about the geomorphic history of the MNRV outside of a few studies. Johnson, Davis, and Pederson (1998) is the only modern study that attempted to analyze the morphology and sedimentology of fluvial terraces within the MNRV in order to reconstruct the valley’s history. They hypothesized a relative chronology where a major incision event linked to the initial catastrophic flood event from Lake Agassiz was followed by 22 m of infilling based on sand and gravel deposits with forest bedding up to 15 m in height (Figure 2). This period of aggradation was then followed by a brief “more stable” braided stream period and then another significant flood pulse resulting in incision. They concluded that terrace correlation would be impossible without “high-resolution dates” and further sedimentologic work on other terrace locations in the MNRV.

Several small tributaries of the Minnesota River are far better understood than the MNRV itself. Gran et al. (2013) used the Le Sueur River (Figure 2) as a case study to address the assertion that geomorphologists know very little about transient response in the fluvial system. The Le Sueur River is a transient system responding to approximately 65 m of base level lowering caused by MNRV incision. In order to address the river’s incision history and to compare with numerical modeling, optically stimulated luminescence (OSL) and \(^{14}C\) dating determined the age of fluvial terraces along the stream reach. Overbank alluvium was sampled for OSL dating while freshwater mollusks and gastropods were collected from terrace fills for \(^{14}C\) dating. The results of the study revealed that the Le Sueur River is a detachment-limited system where downstream bedload coarsening inhibits incision and transposes the knickpoint over a larger distance in the longitudinal profile.
Figure 2  Shaded relief map of southern Minnesota highlighting the Minnesota River Valley and important geographic locations discussed in the case study. U – Upham, N – Norcross, and H – Herman represent the three highest shorelines (from low to high) of Lake Agassiz described by Fisher (2003) and Lepper et al. (2007). BVF – Brown Valley Fan and BSL – Big Stone Lake represent the locations of the two core sites dated by Fisher (2003). LeS represents the approximate location of the Le Sueur River which was the focus of Gran et al.’s (2013) research. Kasota Pit is the location of Johnson Davis, and Pederson’s (1998) sedimentologic analysis of a terrace fill within the Minnesota River Valley.
Sierra Nevada glaciation case study

The Sierra Nevada Mountains in California remain a classic range for the study of glaciations starting with the use of topographic position to establish relative ages. The current state of understanding starts with $^{40}$Ar/$^{39}$Ar ages of basalts atop the range that record a preglacial plateau topography prior to 3.3 million years ago (Phillips, McIntosh, and Dunbar 2011). K–Ar and $^{40}$Ar/$^{39}$Ar dating of the Bishop Tuff (Sharp 1968) provided a minimum age for the most extensive glacial event, the Sherwin glacial till that rests under the 760,000-year-old supervolcano Bishop Tuff deposit.

As for glacial events younger than the Bishop Tuff, for many decades at the end of the last century, only relative dating techniques were available to order a host of different glacial moraines. However, it was not until the development and use of cosmogenic nuclide surface exposure dating that researchers began to get a handle on the true complexity of glacial sequences (Phillips et al. 2009). In spite of decades of grouping together moraines into broad age classes called “Tioga” and “Tahoe,” the precision and accuracy of cosmogenic methods continue to reveal that Sierra Nevada moraines reflect the complexity of millennial-scale climatic changes. Whereas all last glacial maxima moraines were once all grouped as Tioga, they can now be differentiated into multiple age groupings throughout Marine Isotope Stage 2, from 28 to 14 ka.

The timing of Holocene glaciation in the Sierra Nevada has been more difficult to understand. The chitin of insects has been preserved in sufficient quantity for $^{14}$C measurements along Bishop Creek in the central Sierra Nevada to establish at least four distinct glacial events at approximately 2700 calibrated years BP, approximately 1300 calibrated years BP, approximately 600 calibrated years BP, and approximately 150 calibrated years BP (Dorn 1996). Later, sediment cores along the nearby North Fork Big Pine Creek revealed a continuous record of sedimentation dated through $^{14}$C analyses of macrofossils. This research found similar glacial maxima at approximately 2800 calibrated years BP, approximately 700 calibrated years BP, and 250–270 calibrated years BP, in addition to other maxima at 2200 and 1600 calibrated years BP. Thus, ongoing research into understanding the ages of glacial moraines in the Sierra Nevada mirrors and lags behind our understanding of paleoclimatology in general – that millennial-scale high-magnitude changes in climate can imprint themselves throughout the geomorphic record globally, but the detection of these imprints is limited by the accuracy and precision of the techniques used to date landforms.

Table 1 and these two case studies illustrate that a great number of options exist for today’s geomorphologists. In both research contexts, the type of dating technique employed reflects a combination of the type of material available for dating and the expertise of the researchers investigating the problem. What remains true, however, is the ongoing need for chronometric information to advance our understanding of Earth’s landforms.

SEE ALSO: Applied geomorphology; Biogeomorphology; Fluvial depositional processes and landforms; Fluvial erosional processes and landforms; Geomorphic systems; Geomorphic thresholds; Glacial depositional processes and landforms; Glaciations; Glacier mass balance; Glaciers; Models in geomorphology; Paleoclimatology; Soils as relative-age dating tools


Rail transport: freight

Claude Comtois
University of Montreal, Canada

The geography of rail freight transport

Rail freight transport has played an important role in shaping the economic geography landscape. Global analysis of railway freight transport suggests a complex system of railway patterns and functions. First, several penetration axes can be found in Africa, Australia, Brazil, and Canada. These isolated railroads frequently run inland from a coastal port for the development of regions with resources. The mining cluster of the Mpumalanga region accounts for over 80% of South Africa’s coal production. Coal trains travel 588 km to the port of Richards Bay to be loaded on vessels bound overseas. Rail connection is used as a conveyor belt to ship products toward consumer markets. Rail freight traffic generated along these trade lanes can serve as a tool in opening up new regions, fostering settlements, and attracting investments, together with the emergence of intermediate centers along the penetration routes with tangential feeder services.

Second, railways can operate as a subsidiary of plants. Freight trains run on lines between factories to deliver supplies. As rail freight is well suited to mass production, plant railways facilitate the process of industrialization by carrying minerals, agricultural products, and building materials or chemicals to manufacturing centers for first-stage processing, fabrication, or assembly. Rail freight transport between a large number of interlocking individual firms has created favorable conditions for the development of large industrial complexes, as exemplified by the Ruhr basin in Germany, the Pampas in Argentina, and the Kanto plain in Japan. As traditional customers for the transport of bulk commodities on the input side, industrial enterprises also favor railways as the preferred transport mode for the carriage of any resulting waste and hazardous goods from their plants.

Third, railway pioneers realized the importance of linking water loading sites and inland dry ports with overland modes of transport. The rail transport of large volumes of different kinds of freight flows between transport terminals reinforces the complementarity between marine and land ports. The integration of terminals, transport corridors, engineering structures, and fleet illustrates the cumulative connections between infrastructure and trans-shipment function in the development of fully fledged logistics in freight services. At the port of Duisburg, Germany, the integration of rail infrastructures with barge transport has increased the service efficiency and activity of the fluvial port, reducing the cost of using the port, increasing the capacity of organizing logistics services, and improving the competitiveness of products exported to international markets. These linkages between capital-intensive transport infrastructures permit economies of scale, improve market access, increase fluidity of commercial exchanges, and strengthen industrial networks.

The fourth type of rail system consists of the linkages between continental landmasses. Railway carriers haul freight by means of owned or leased equipment. They operate partially over
RAIL TRANSPORT: FREIGHT

their owned tracks or negotiate a right of way owned by other freight railways. Continental railway lines can be found between Montreal and Los Angeles in North America, from Dalian to Moscow in Eurasia, across Southern Australia between Perth and Adelaide, and between Buenos Aires and Valparaiso in South America. The reduction of obstacles to international trade, including the establishment of economic blocs, has favored the introduction of rail land bridges. A rail land bridge refers to traffic transiting a continent bound to overseas destinations. This is the case of some Asian traffic bound for Europe. Cargo is shipped across the Pacific Ocean, carried by rail over North America, and transferred onto vessels for ocean sailing across the Atlantic.

The geographical distribution of rail freight transport suggests that rail corridors leading to and from important places of origin and destination are servicing the spatial distribution of production, procurement, and distribution of commodities. In addition, rail arterial links are contributing to the development of integrated intermodal transportation systems.

The capacity of rail freight transport

Further geographical expansion of rail freight transport must address issues of railway track standards, custom-made rolling stock, and applied information technologies. Since railways operate on specially prepared roadbeds, the number of railway tracks laid is a key determinant in rail freight flows, as trains can overtake each other only when moving on separate tracks. In regions where large metropolitan centers are spatially clustered, there is justification for promoting a higher utilization level of existing railway lines with double-tracking as transportation demand is concentrated, with large volumes of freight traffic over certain rail corridors. In sharp contrast, in areas where urban centers are relatively scattered, involving long distance and smaller traffic volume, the dominant consideration is to reduce construction costs with a single-track railway.

Rail freight transport is equally influenced by railway track geometry, which is produced mainly by the effects of heterogeneous landforms. This includes elevation, curvature, alignment, and track gauge. In Europe, different track gauges are restricting rail freight movement. The most important track gauge limitations within the European freight traffic are found between central Europe (1435 mm) and the Commonwealth of Independent States (1524 mm), and between central Europe and the Iberian peninsula (1676 mm). The competing conditions generated by liberalization processes have created a new dimension in railway track geography. In Europe, differences were established between rail infrastructure companies and rail transport companies. This suggests that the geographical expansion of rail freight transport is affected by nondiscriminating conditions for network access in terms of train path prices for the compensation of use. In North America, mergers and acquisitions led to a reduction in the number of Class I rail carriers from 11 in 1990 to 7 in 2013. The largest carriers in Canada are Canadian National and Canadian Pacific. In the United States, five companies qualify to be Class I railroads: BNSF Railway, CSX Transportation, Kansas City Southern Railway, Norfolk Southern Combined Railroad Subsidiaries, and Union Pacific Railroad. The commercial trend to reduce costs and offer better services pushed these rail carriers to concentrate on freight movement and to abandon part of their network. The process led to the emergence of new short-line railways. There are currently 30 regional and over 500 local railways operating in the United States and Canada. The benefits
of transporting different commodities between numerous pairs of origins and destinations prompted the private sector to focus on freight traffic. The North American railway passenger services are essentially public-owned and subsidized. These public companies operate almost entirely over private freight railway tracks.

The quality of railway tracks also affects the functionality of railway freight traffic capacity. There are various quality track standards. These pertain to the properties of steel rails in terms of chemical composition, the weight of rail expressed in kilograms per meter, and rail length as indicator of number of welded joints and dilatation spaces. Arguably, all these components of railway track structure determine the speed of trains, the axle load of wagons, locomotive traction, the density of freight traffic, and resistance to hazardous environmental events. Further, the implementation of longer trains, more cars per train, and heavier cars through high-quality infrastructure are concomitant with increasing yard capacity in permitting additional train blocks and higher network fluidity. Following the growth in traffic at the port of Los Angeles and Long Beach, the rail company Burlington Northern–Santa Fe has built a new intermodal terminal 30 km inland where double-stack trains are assembled with a view to distributing containers across the United States. Rail access between the port and the railway yards was impeded by single railway tracks involving 200 level crossings through dense urban neighborhoods. In 2001 a plan was approved to construct a dedicated 32 km rail corridor between the port and the inland rail terminals. The Alameda corridor is capable of handling 100 trains per day, reducing daily truck traffic by over 20,000 movements. Deficits in the standardization of railway networks require major investments in crossing loops, passing tracks, loading sidings, and marshaling yards to expand rail freight activities.

The volume of freight traffic that can be carried by railways is a function of rolling stock. Freight trains are loaded and unloaded at different lineside locations such as mining facilities, power plants, industrial sites, and grain elevators. The consignment of these commodities over long distances necessitates the availability of different types of freight cars adapted for the movement of a host of goods. The most common railroad cars used for carrying cargo are open wagons for minerals and other bulk cargo, flatcars for large break-bulk loads, ventilated box wagons for carrying livestock, tank cars for the movement of liquid bulk, chemical, and gas. In addition, the integration of railways within the global containerized intermodal system has led to flatcars designed for containers or semitrailers. The provision of suitable freight loading space is leading to purpose-built wagons. Developed to answer the transport needs of certain lines of business and to ensure closer ties between consignor and consignee, these include fully enclosed wagons for perishable, moisture-susceptible, or weather-sensitive goods.

The liberalization of the rail transport market has given rise to a multitude of new rail operators with a growing need for locomotives and freight wagons. The integration of rail freight transport in the modern logistics of merchandise and manufactures requires a flexible response to many customer-specific requirements. High investment costs and the maintenance of regular and adequate fleet utilization associated with a wide spectrum of rolling stock are significant barriers to market access of railway carriers with expansion plans. Accordingly, strategies to develop rail freight movement include service cooperation in terms of rail vehicle pools for both locomotives and freight wagons. This flexibility in the provision of vehicles allows railway operators to design railway fleets to freight
RAIL TRANSPORT: FREIGHT

shipments and secure greater market shares in the contested freight sector.

Information and communication technologies are key factors in increasing the market share of freight railways. Rail freight capacity is determined by the number of train movements along a railway line between numerous stopping points. As rail freight traffic becomes larger in scale with varied destinations, rail lines are affected by competing usage requirements and by unbalanced rail freight transport demand. Computer-assisted systems for sequencing trains are defining the rights of use in the form of available train paths on hinterland routes. Information technologies applied to rail freight transport include the remote control of vehicles and facilities, and commands for the operation of railway infrastructures. Railway technology development also encompasses railway equipment automation for shunting, freight handling, and trans-shipment operations. Adaptations and changes in rail freight transport are realized through integrated consideration of flows of information, fluctuating freight demand of commodities, and actual use of railway infrastructure materials, rolling stock, and rail services. Advantages have been derived from appropriate interfaces for exchanging information between rail freight service providers, shippers, and consignees in support of commercial transactions. The introduction of operational information systems permits the efficient organization of the complex process involved in the delivery of freight supplies. Arguably, direct access to complex databases, online planning, and the possibility of simulation pose enormous challenges for rail freight operators. Computer-aided design and transport softwares are increasingly being used alongside geographical information system in the planning, design, construction, and maintenance of rail freight infrastructure. Geographic information systems are playing a significant role in the continual planning and control of the transport processes on the basis of a corresponding application of network representation, modeling, and multivariate statistical analysis. Examples include shortest transport times and favorable turnaround times for freight wagons, balance in network flow, railway facilities location, and land-use planning.

The future of rail freight transport

The promising development of freight railways in capturing new business markets rests on the combination of rail transport, high-speed freight trains, and city logistics. The globalization of trade networks and resultant modal distribution between regions has led to the growth of container traffic. The overwhelming primacy of containers has become an essential feature of the commercial integration of the world’s regions. A key factor in the evolution of hinterland connections with container trade is the role of combined transport. The framework comprises a road or barge segment to the terminal, cargo handling at the shipping terminal, the rail journey, cargo handling at the receiving terminal, and fluvial or truck delivery to the customer. Each mode takes over the transport service segment that is economically optimal for it. The objective is to expand single-mode transport services beyond the single-mode network to achieve greater geographical coverage and involvement in long-distance freight traffic. Rail terminal facilities adjacent to major rivers and highways are offering high-frequency rail, road, and barge connections to and from deep-sea container terminals. The emergence of trimodal terminals along the Mississippi, the Rhine, the Yangtze, and the Saint Lawrence rivers allows shippers to develop clusters of services in the assembly, collection, shipment, and distribution
of freight with a view to offering door-to-door services over extended geographical market areas. The complexity of the services offered at these terminals means that long-term customer relations are important to the railways as a basis for establishing important linkages in the logistics chain.

The advent of just-in-time deliveries and the trend toward small- consignment and short-delivery deadlines in manufacturing processes have influenced the scope and spatial structure of freight traffic. Cargo potential includes perishable goods, express courier, and pre-fabricated components and parts involving substantial just-in-time manufacturing processes. The introduction of the high-speed train from southeast Paris to Lyon has led to the development of customized high-speed swap bodies to meet the technical performance requirements of very fast trains. SNCF, the French national railway, has built trans-shipment stations at the railway stations parallel to the tracks used for passenger transport. Freight is loaded through sliding doors located in the middle of these specially designed vehicles. High-speed freight train movement avoids burdening the highway networks with further lorry traffic and reduces the burden to airlines from airport gate delays. The development of high-speed rail corridors thus gives railways an opportunity to win market shares from road and air transport in long-distance freight movement. The trend toward miniaturization of shipped batches of forwarded goods can be fulfilled through the flexibility, punctuality, and reliability offered by high-speed trains.

Transportation is at the core of many urban problems and challenges in terms of sustainability. Freight railways can advance city logistics while fostering long-term protection of the environment. Cities encompass firms producing and offering goods. Transportation companies are moving these goods using a transportation system in which goods and passengers compete on the same arterial network. A basic rule for the profitability of rail shipments is that large quantities of goods be hauled on a regular basis between two points. In Dresden, the Volkswagen Group has entered into partnership with the federal state of Saxony and the local transport operator for the movement of parts and materials from a freight handling center to a new factory using the tram lines used to transport passengers. Appropriate facilities in terms of sidings are provided at both ends to prevent scheduled services from being disrupted. The system has a design capacity allowing 40 minutes’ headway between freight trams 21 hours a day and meets all requirements in terms of traffic lights and stopping points. The objective is to use tramways to feed automobile production manufacture near the city center from a logistics platform accessible by rail and motorway. The Dresden CarGoTram is an ecological and cost-effective transport alternative, carrying three times the load of a single lorry while offering savings in both labor and energy costs.

Rail freight transport contributes to the performance of complex logistics systems. As rail freight transport moves into the future, there will be additional pressure on rail operations that may result from security and environmental considerations. It is up to geographers to provide insights to minimize the impacts while maximizing the opportunities as rail freight transport closely mirrors changes in local, regional, and global transport geography.

SEE ALSO: Accessibility, in transportation planning; City logistics; Containerization; Industrial geography; Regional development models; Regional development policies; Transport networks; Urban geography
RAIL TRANSPORT: FREIGHT

Further reading

Rail transport is a means of conveying freight and passengers by moving wheeled vehicles on rail tracks. In a broad sense, rail transport also includes some new forms of guided land transport, such as monorail and maglev trains that run on electromagnetic guideways. Passenger rail has two main forms: first, conventional intercity rail and high-speed rail (HSR) that connect major urban areas; second, commuter rail, light rail, metros, and trams that mainly provide services within and near urban areas. Conventional intercity rail generally runs along densely populated corridors between cities and supports long-distance passenger travel; HSR, as a new kind of intercity rail, has more advanced technology, higher speed, and better service than conventional intercity rail; commuter rail, also called regional rail or suburban rail, mainly operates between a city center and its suburban or surrounding areas; light rail and metros, which have high capacity and speed, play important roles in urban transport, especially for passengers’ daily commuter travel; trams, as an old passenger rail form, are facing challenges of modernization. In accordance with the different services of these rail forms, their speed, capacity, frequency, and right of way also show some differences. For example, commuter rail usually has higher speed and lower frequency than light rail or metro, but covers a shorter distance with fewer stops than intercity rail; light rail does not have fully grade-separated right of way as do intercity rail, commuter rail, and metro, and it has lower capacity; metro is an electric urban rapid transit with relatively high frequency of service. Nevertheless, all these forms of passenger rail transport provide different kinds of travel services for passengers.

From a geographical perspective, there is great regional disparity in terms of passenger rail transport. The world’s total passenger-kilometers by rail transport reached 2867 billion in 2012. Table 1 shows the regional shares of Europe, the Russian Federation, Africa, America, and Asia, Oceania, and the Middle East and their growth rates from 2007 to 2012 (UIC 2013). It shows that the passenger rail traffic is highly concentrated in the region of Asia, Oceania, and the Middle East. Europe has the second-largest passenger-kilometer share but its growth rate is much lower than that of Asia. Although the passenger rail traffic share is very low in America, it has had a very high growth rate recently. Both the Russian Federation and Africa have low passenger travel shares and negative growth rates. Moreover, railways in Asia and Oceania are intensively used for passenger travel. Asia, Oceania, and the Middle East accounted for 76% of world passenger travel with only 23% of the world’s total (including both passenger and freight) railway length. Europe accounted for 17% of world passenger travel with 26% of the world’s railway length. America only accounted for 1% of world passenger travel with 36% of the world’s railway length, reflecting the dominance of freight transport in the American railway network (UIC 2013).

The great regional disparity of passenger rail transport could be due to several reasons. First,
for short distances, automobiles are faster and more convenient than rail transport. For long distances, air transport is much faster than most rail transport. So, in areas where automobiles and air transport are available and well developed, such as the United States, passengers tend to choose these other transport modes instead of railways. In much of the developing world, such as China and India, railways are still dominant in passenger travel. Second, the operational costs of railways are high and it is only economically feasible if the passenger number reaches a large volume. Facing the keen competition of automobiles and air transport, railways often experience financial difficulties. Although different kinds of strategies were adopted for lowering railway costs, such as closing unprofitable lines, increasing travel speed, adopting new and more efficient fuels, and providing air conditioning, night berths, and restaurant services (O’Dell and Richards, 1971), the decline of railways in many countries was not reversed. Third, the construction of railways requires relatively level ground because of its engineering characteristics and so highways and airports are more economical in regions without level ground topography (Sharma 2007). Nevertheless, in some regions with high passenger demand or undeveloped areas, rail transport was often the most secure and efficient means of passenger travel.

The technology of rail transport has constantly improved over time. As early as the early nineteenth century, steam trains had been successfully developed in England. Although electric trains, trams, and rapid transit systems had been introduced in the 1880s, it was not until World War II that the motive power of rail transport largely changed from steam to diesel and electricity. In the 1960s, HSR technology was first introduced to passenger travel in Japan. Over time, HSR was built in many other countries, such as France, Germany, Spain, the United Kingdom, Italy, and China. Compared with conventional railways, modern railways can accommodate passengers’ mobility needs with higher speed. Urban rail transit systems now support massive passenger travel in many metropolitan areas around the world.

A central issue related to rail transport in geography is its important role in sustainable transport, which covers the environmental, social, and economic aspects. In terms of the environmental aspect, rail transport has relatively less impact than automobiles and air transport per passenger-kilometer carried. For instance, studies show that the carbon dioxide emissions per passenger-kilometer of railways in cities are generally lower than other transport modes, except buses (Loo and Li 2012). Although the absolute level of emissions of HSR are higher than conventional railways, they can have lower emission intensity by having higher load factor, fewer stops, aerodynamic train design, and lower carbon content in electricity supply. Also, rail transport can offer higher passenger capacity than automobiles with less noise and air pollution,

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Regional shares and growth rates of passenger-kilometers by rail transport (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Africa</td>
</tr>
<tr>
<td>Regional share</td>
<td>1</td>
</tr>
<tr>
<td>Growth rate (2007–2012)</td>
<td>−20</td>
</tr>
</tbody>
</table>

Source: UIC 2013.
smaller land uptake, and lower energy consumption. However, there are still many challenges for making railways more environmentally friendly, including shifting the demand from automobiles and air transport to rail without generating new or induced demand, increasing the load factor of rail transport, and implementing operational and technological measures to improve the energy efficiency of rail operation (Givoni, Brand, and Watkiss 2009). In the future, as the technologies of clean energy and trains further develop, rail transport may play an even more important role in promoting the environmental sustainability of passenger transport.

In terms of the social aspect, rail transport performs better in safety, social equity, and traffic congestion alleviation. Rail transport is generally safer than other land transport modes because trains usually operate on exclusive tracks. In Europe, the risk of fatality per billion passenger-kilometers for rail transport from 2004 to 2011 was only 0.15, much lower than that of cars (3.8 fatalities per billion passenger-kilometers) (Allianz pro Schiene 2013). Also, compared with private vehicles, rail transport has the characteristics of public transport, which is more affordable to socially disadvantaged groups. Within cities, rail transport can often improve accessibility to different kinds of social activities for the poor and disabled who cannot afford a private vehicle or who cannot drive. In terms of employment, railways can generate different types of jobs, from senior managers, professionals, and engineers to operations staff, for local people and broaden the range of accessible employment for the socially disadvantaged population. At the regional level, rail transport is often built in the hope of improving accessibility for peripheral regions and providing them with more opportunities, thus promoting regional integration. In addition, rail transport with high speed, capacity, frequency, and reliability could effectively relieve the problems of traffic congestion in large metropolitan cities.

In terms of the economic aspect, the main contribution of railways to development is through indirect impacts on regional economies. On the one hand, the accessibility of the region along the railway may be improved by the connection of railways and employment, industry, and trade. It is a strategy of the World Bank to promote regional development and reduce poverty by lending to support railway projects. On the other hand, the investment of railways may stimulate land development and business activities around railway stations. The transit oriented development (TOD) model, which emphasizes mixed land use around a transit stop, gives potential opportunities to integrate railway development and land use. Accordingly, railway stations are no longer seen only as access nodes to railway systems but also as places with many other activities such as residential, commercial, and industrial services (Reusser et al. 2008).

Although rail transport has been recognized as a sustainable transport mode, rail transport projects are financially risky investments in light of high capital and operating costs. The construction of railway systems requires heavy capital investment, that is, planning and land costs, infrastructure building costs, and superstructure costs; and there is no guarantee of sound financial returns in the short term because of the high maintenance and operating costs, not taking into account external costs. Hence, planning and investment in rail transport (passenger) often receive much attention and are the focus of debate. For instance, it is argued that there should be socioeconomic benchmarks for evaluating cities for building or extending metro systems, such as population size and income level (Loo and Cheng 2010). For other railway systems, careful planning and feasible studies to
RAIL TRANSPORT: PASSENGER

contrast benefits and costs should be undertaken before the construction.

Moreover, project management and effective operational management policies are important. Considering the large investment and operating costs, traditional railway regulations were often based on public ownership and operation. However, government control of railways is increasingly being viewed as inefficient, quasi-monopolistic, and unresponsive to the market. Many countries aim to improve the efficiency and competitiveness of their passenger rail system by reforming railway governance: some have completely separated infrastructure management from rail operations and placed it in a state-owned organization, such as in Sweden and Britain; some have retained infrastructure and train operations within the same holding company but as separate divisions, such as in Germany, Italy, and Greece; some have separated infrastructure management from operations while still maintaining an operating company monopoly that retains control of many functions, such as in France (Nash 2008). With more empirical evidence, the discussion on railway governance will remain a key area in rail research.

Recently, the twenty-first century has witnessed a revival of rail transport for passenger travel. The reasons are manifold. First, the negative environmental, economic, and social impacts of automobiles and air transport, such as local air pollution and traffic congestion, as well as global warming, have raised worldwide concerns as the passenger traffic volume increases (Shaw and Docherty 2009). Second, the technological improvement of rail transport makes it not only environmentally sustainable but also more competitive with automobiles and air transport because of the higher speed and greater comfort. Third, since the possibility of alternative solutions for sustainable transport (such as improvement to telecommunications making transport less necessary and the large-scale use of alternative fuel vehicles) will not be realized in the near future, rail transport has been seen as a promising strategy for achieving sustainable transport.

In the last two decades, there has been large-scale construction and planning of HSR and urban rapid rail transit systems around the world. Both HSR and urban rapid rail transit have high capacity and high frequency. Shorter travel times and an improved level of service of HSR have led to changes in modal share along the HSR route (Givoni 2006). Experiences in Europe show that HSR can substitute short-haul air transport for passenger travel. However, whether the urban rapid rail transit systems can bring passengers from automobiles back to public transit is still under debate and there may be induced travel demand due to new rail transport development. Nevertheless, modern rail transport could have some impact on people’s travel behavior and slow down the rate of motorization in developing countries with currently low reliance on private vehicles.

Moreover, fundamental geographical changes are happening. Shorter travel times and increased accessibility of cities by HSR cause compression of time and space and drive change in economic development and social interaction between cities. Meanwhile, the impact of HSR on regional land development can also be great. However, compared with cities with stations in the HSR network, those that are bypassed would probably lose some opportunities. The impact of urban rail transit may not be as wide as that of HSR, but its impact on restructuring urban form could also be significant. The livability of communities along the metro or light railway may be improved and more employment opportunities and services may concentrate near the stations. However, property rents and living costs may also be increased.
Taking a wider perspective, the integration between rail transport and other transport modes and the integration between rail transport and land use have also received more attention from researchers. The former aims to promote the accessibility of railway stations and improve the connection between the railway and other transport modes. The key concern is to enhance the travel experience of passengers and, hence, enable the railway to fulfill its potential role in supporting sustainable transport. The latter considers the coordination between rail transport supply and travel demand depending on land use. The focus is to minimize unnecessary journeys through creating compact activity nodes and to increase the convenience and positive impacts of rail transport as a passenger transport mode.

There has been some progress towards better integration related to rail transport around the world. This includes the building of highly integrated transport hubs, the integration of land-use planning with rail transport planning, and the establishment of common information and ticket-booking systems for different kinds of transport modes. Yet, there is still a long way to go before we realize the vision of seamless integration and more efforts are needed in both infrastructural construction and institutional support.

In the near future, rail transport will remain important in contributing to sustainable transport. Transport geography, the discipline focusing on the network, movement, context, and impact of transport systems, will continue to contribute to the decision-making, impact analysis, and institutional reform of rail transport and its integration with other transport modes and land use.

SEE ALSO: Accessibility, in transportation planning; Sustainable transport; Transport geography; Urban transit

References


Realism, critical

Kevin R. Cox
Ohio State University, USA

Critical realism (CR) originated in significant part as a critique of positivism. The philosopher Roy Bhaskar was central to the project, and its entry into geography would be largely through the work of Andrew Sayer, at the beginning of the 1980s. The positivism that had been at the heart of the spatial-quantitative revolution of the 1960s had already come under attack at the hands of Marxist and humanistic geographers. The contribution of CR to that critique and its contributions to method in human geography, though, would be significantly different.

In contrast to Marxism’s totalizing account of the world and its critique of positivism, CR was more pluralistic. It represented a critique not just of spatial-quantitative work but also of the Marxist geography of the 1970s. For humanistic geography the search for laws ran aground on the inherent indeterminacy of human agency. CR recognized the force of this claim while keen to preserve a crucial role for structure. On both these grounds, therefore, it sat comfortably within an emerging society-and-space interest, anxious about what Thrift would call “jumbo Marxism” and about the question of structure and agency.

That it had a positive impact in human geography is without doubt. It was drawn on extensively by people of more radical persuasion such as Simon Duncan, Michael Storper, and Richard Walker. Doreen Massey’s work on spatial divisions of labor was seen to fall within the critical realist approach even while she never articulated her method as such. Significantly, Duncan, Massey, and Sayer would all be on the editorial board of the new journal Society and Space as it took off in 1983. Voices of disquiet were few and far between but David Harvey was notable among them. At its most ambitious, CR was presented as a method for a radical geography, and, according to some, Marx was a critical realist before his time. This now seems quite inaccurate (Cox 2013) and Harvey’s presentiments justified.

What was at stake was a philosophy of science and, as such, ontological claims were fundamental. For critical realists the world consists of the events we experience, the causal mechanisms responsible for them, and the structures of relations that in turn entail those mechanisms. Similar events can be generated by quite different mechanisms; their conditions of existence in structures of relations therefore vary. Structures of relations entail causal powers or causal susceptibilities – characteristic ways of being affected. The enabling of these causal powers or susceptibilities then results in the world of events.

Given the way in which spatial-quantitative geography and positivism in general hewed to a Humean concept of causation in which constant conjunctions of events were a necessary though not sufficient condition for claims about causation, CR appeared to offer powerful critical leverage. The immediate targets were the empirical regularities at the heart of the spatial-quantitative work. Conjunctions of events had their conditions in structures of relations that could vary significantly from one conjunction to another. Something like the distance-decay relation as it applied to migration provides a case in point. For movements at a given distance...
from a point of origin, the structures of relations necessary for them to have occurred can clearly be very, very different: a corporate transfer within the structure of a firm’s spatial division of labor is not at all the same as the case of someone who in the context of a military posting marries someone local and decides to stay. Distance relations might be part of the explanation—the way firm branching tends to be intra—rather than international—but only in a highly mediated way and as something that had to enter into socio spatial rather than simply spatial relations.

Even where a constant conjunction seemed to have some rationale, where it might seem that correlation was not just necessary but also sufficient to impute cause, again it could have other conditions that had to apply if it was to obtain. For example, the way in which, in the United Kingdom, the differential availability of central government employment subsidies seemed to even out unemployment rates in the 1960s attracted attention (Sayer 1986). Subsidies were made available for so-called development areas, which coincided with areas of relatively high unemployment—erstwhile coalmining or shipbuilding areas in particular. But clearly any success that the policy enjoyed depended on still other conditions. Almost invariably, the new locations were of branch plants, suggesting the significance of new spatial divisions of labor of firms. Labor processes requiring limited industrial skills, as in light assembly work, were also high on the list of activities subject to relocation in this way, but this in turn suggested the importance of processes of deskilling that allowed this process to occur, a process entailed, in turn, by the capitalist accumulation process. There was also the fact of labor shortages in existing industrial centers. This suggested that the location of branch plants in the depressed areas would have occurred regardless of the existence of government subsidies. In short, causal structures seemed much more complicated and diverse than those that had been embraced hitherto. One had to look beyond constant conjunctions of events to the structures of social relations that were their condition.

Structures of relations were the sphere of the necessary: what had to be present in order for an event to occur. But whether or not it did occur was a contingent matter—contingent on the presence of yet other conditions. A firm with a spatial division of labor of the parts-process form might have the ability to relocate low-skill production, but whether it did might depend on labor market conditions: how much its profit lines were being squeezed by the wages of semi- and unskilled workers. In contrast to the sphere of necessity defined by structures of social relations, events had a contingent status.

This argument raises the crucial question as to how structures of relations were to be identified. A useful contrast drawn by Sayer (1984) was between what he called “chaotic conceptions” and “rational abstractions.” Chaotic conceptions brought together conditions, practices, and objects that were related to one another regardless of the contingent or necessary nature of those relations. As such it was impossible to impute causal properties that necessarily inhered in them. Rational abstractions, on the other hand, brought together the necessarily or internally related, as in a division of labor. Each part depends on the others and on some sort of coordination mechanism. The result is that those participating in a division of labor find that they have new causal properties, not least an ability to specialize.

Rational abstractions could be identified on the basis of a variety of evidence. Intensive interviewing can go some way in identifying the necessary conditions that agents are aware of and which help them make sense of their actions. Documentary evidence suggesting causal mechanisms and their conditions could be important.
and even empirical regularities when drawn on critically. In any case, the typical approach was to work back from the world of events to what was conditioning them: a process defined as retroduction and in contrast to the inductive approach of positivism. Retroduced structures then stood to be revised in the wake of new research.

The interest in CR in geography generated a number of notable contributions. Critical work was to the fore and some of it involved new empirical investigations (England 1993). In other cases attention focused on particular structures of social relations such as patriarchy and the division of labor and their geographic implications (Foord and Gregson 1986; Cox and Wood 1997). In retrospect its impact on human geography might seem to have been short lived, eclipsed by the interest in the various posts that occupied the center ground of the field from the late 1980s on. Little is now heard of it in geography. Even so, there were some lasting effects: the language of necessity and contingency crept into geographical analysis, though not always accurately. More significant was the way in which, through its emphasis on preconditions for events and the utter contingency of the latter, it lent impetus to a heightened historical awareness in explanation.

SEE ALSO: Marxist geography; Quantitative methodologies; Spatial social networks; Structuration theory

References


Further reading

Recreation

Anne-Marie d’Hauteserre
University of Waikato, New Zealand

Even though recreation research was earnestly pursued by geographers as early as the 1930s, “recreation geography has never been valued … because it is believed it is impossible to be serious about individuals and groups having fun” (Butler 2004: 151). The specialty has lacked prestige in academia so, until recently, there was a paucity of material covered and relatively little research was undertaken by geographers compared to those in other disciplines. And yet leisure is considered a human right under the United Nations’ Universal Declaration of Human Rights. While geographers have discussed leisure as a force that allows individuals to consider and reflect on the values and realities that are missing in the activities of daily life, and is thus an essential element of personal development and civilization, little debate has emerged about what constitutes the field for geographers.

Recreation is an activity done for enjoyment when one is not working, and it has been emphasized that “recreation is an activity voluntarily undertaken primarily for pleasure and satisfaction during leisure time” (Patmore 1983: 3). The word comes from late Middle English (in the sense of “mental or spiritual consolation”) via Old French from the Latin recreatio(n-), from recreare, “create again, renew.” These roots help explain why recreation has also been considered to be useful for “recharging the battery” so that workers can be more productive. It is important to remember, though, that leisure time is a social construction and that such definitions apply primarily to Western societies, where work and free time (leisure) have long been distinct for some. Feminist literature has demonstrated, however, that women (and also others) often devote much time to unpaid work during their free time, and that historically leisure has been a male prerogative. Leisure and recreation, moreover, have become popular activities only since the 1900s, after workers gained the right to paid “free time” (holidays and weekends). China, for example, only recently provided three weeks of “leisure” per year, to entice the Chinese to travel and discover their own country. The increase in leisure and recreation has coincided over time with increased longevity and, for many, with decreased hours spent on physical and economic survival.

For geographers, recreation is a dynamic phenomenon that provides meaning through which people can construct their identity, as it creates a sense of belonging, enabling friendship, community building, and empowerment through its performance. Recreational activities can be communal or solitary, active or passive, outdoors or indoors, healthy or harmful, and useful for or detrimental to society. Many recreational activities are organized by public institutions, voluntary group-work agencies, private groups supported by membership fees, and commercial enterprises, and are thus considered as formal because they occur at specific times and sites, whereas informal activities tend to be self-organized and are fluid in relation to times and sites. Geographers have also examined the increasing consumption and commodification of recreation and underlined that a resource base (in quantity, quality, and extent) is a precondition for many recreational activities. These can be
formal (e.g., cultural resources such as libraries, museums, and theaters) or based on nature (e.g., parks, forests, farms), and operate as private, commercial, public, or professional ventures. Some facilities can be capital-intensive (e.g., opera houses, ice rinks) or land-intensive (golf courses), or both (sports stadiums), creating financial barriers to their use.

Geographers have underlined the complex relationship between the desire to partake in recreation and the opportunities (such as the existence of resources – the economic, social, and cultural means – and the physical abilities) to do so (Hall and Page 2014). Although motivation to participate might include the need for excitement through the challenging nature of the activity and the use of high-level or special skills, leading to a peak experience, in other cases only relaxing, undemanding, and low arousal activities are sought. A continuous spectrum lies between the two extremes, with different consequences on the environment or other resources. Geographers have also considered how creative expression figures in recreational practices.

Recently, geographers investigating recreation – a process that involves complex social and individual relations and contestations – have emphasized how space, itself socially determined, constructs opportunities for and access to recreation. Geographers have noted the rapidly changing nature of time and space, as in time–space compression, and how this phenomenon has extended the space of opportunities for recreation away from the home just as the home now provides a greater variety of recreational outlets. Recreationists themselves have also become the subject of studies by geographers in terms of their embodied and socialized practices and their multisensory sensitivities. Such studies have revealed a multiplicity of behaviors as well as emotions, affect, and identity constructions based on ethnicity, age, and other social dimensions. Emotions, for example, have much influence on the performance of a learned skill, a characteristic of many recreational activities.

Even though geographers agree that leisure and recreation are considered a human right, they have pointed to obstacles and constraints that prevent their practice. There is a large range of these: an individual’s habitus often determines the specific activity pursued (e.g., attending opera versus playing informal rugby); income influences mobility and accessibility; seasonality dictates the availability of resources (as in skiing); personal constraints may include conflicts with other family desires, lack of time, and health concerns. Geographers have shown how the use of space for recreation is constrained, for example, for women, who may find it difficult to join formal recreational activities as their leisure time is often divided into small blocks not always at convenient times. Some children find their public recreation spaces reduced because of fear for their personal safety. Use of certain spaces, according to some geographers, has also been curtailed by gender and ethnicity: for example, they may avoid peri-urban woodlands and even urban parks through fear of sexual violence and racial attacks.

The impact of recreational activities on natural resources has been a long-standing area of interest in geography, as has the planning of leisure and recreation. Recreation was also long considered by geographers to be significant only in rural areas, as indicated by numerous studies on forests and parks as sites for recreation. Geographers have demonstrated that recreation that occurs outdoors can impact the environment, although its practitioners can also learn to appreciate nature better. Golf courses require water that is then no longer available for other activities in more arid areas, rendering that activity unsustainable in those locations. Certain geographic spaces such as parks, beaches, or the countryside
are essential venues for many recreational activities where active forms of these are becoming more popular. But then questions of access (on the basis of income, race, religion, gender, etc.) arise, or of conflict when practices (on the same basis) clash or incompatible activities take place in the same space.

Early geographic studies of recreation were descriptive in nature, branching out into simple explanatory quantitative studies. Some used quasi-mathematical approaches such as Christaller’s central place theory, applying economic principles to recreation and its spatial patterns. Those who consider recreation as a basic human need, however, have tended toward classification, categorization, and modeling to explain its societal origins. Today, the methods used are very similar to those employed in other areas of geography as authors embrace the turn to cultural, embodied, and emotional theorization and the nonrepresentational aspects of recreation.

Recreation overlaps with leisure and tourism as many aspects of all three converge in personal experience and occur during free time outside of work commitments. Tourism has become a popular global leisure activity, and many tourists are specifically attracted by recreational offerings and often practice recreational activities while visiting a destination. Natural sites and the recreation opportunities they offer have also become an important element of nature-based tourism, and ecotourism in particular.

SEE ALSO: Environment and everyday life; Everyday geographies; Human rights; Leisure geographies; Public space; Tourism

References


Further reading

Reflexivity

Farhana Sultana

Syracuse University, USA

Reflexivity is practicing critical consciousness of one’s location, actions, and power relations during a research process. It involves locating one’s self and positionality both epistemologically and socially vis-à-vis research participants and research process. Reflecting upon and assessing how one is positioned can enable more conscientious and ethical engagement, be it with people, institutions, or texts. Reflexivity requires researchers to continually reassess how they themselves influence the research process and people’s reactions and responses. Useful insights are gained through this process that can then further enrich and inform the research process. Increasing interest in and application of reflexivity by geographers, particularly feminist geographers, has produced a sizable scholarship on the topic in recent years (Rose 1993; Moss 2002). Reflexivity has become established as an important component of research. Being reflexive can be a transformative process whereby the researcher is not only aware of but accounts for power relations, changing subjectivities, and overall effects over time and space. These issues are increasingly becoming critically important in research methods, especially field methods, in human geography that involves qualitative research. It is also important throughout one’s scholarship, which means being critical of not just the research process, but also about one’s writing and reading.

Since field-based research is often unpredictable, an approach that is flexible and reflects on the process continually is able to respond to complexities and contingencies. While this may be contrary to fixed research objectives and goals, or institutionally regulated research permissions, many scholars find the need to navigate such constraints and challenges regularly. A reflexive research process enables scholars to assess and adjust thereby ensuring that pursuits are fruitful and meaningful. While they are mindful of their own location in the process, scholars are also able to respond to situations and relations as they arise with greater awareness and conscience. This is different from maintaining predetermined goals and detached methods that are common in positivistic research approaches. Reflexivity is often contrasted with modernist research that claims to be “objective” whereby the power relations between researcher and researched are erased. There is also often problematic extractive knowledge production, and concerns over gaze, objectification, and reductionism are not directly addressed. To be self-reflexive of the research process is to be cognizant of the situatedness of all knowledge and to recognize that knowledge is partial and always produced from vantage points that are shifting and involve power relations.

Situated knowledge aims for locally “objective” knowledge that is recognized to be partial and context-driven; it does not speak to any absolute truth and thus opposes detachment and modernistic notions of objectivity. Situated knowledge comes as a critique of both logical positivism and standpoint theory. Standpoint theory argues that voices of the marginalized, especially women, should be prioritized and thereby privileges epistemic positions (Harding 2004). Situated knowledge emerged out of poststructuralist engagements in feminist
methodology and espouses that no position is permanent and that all knowledge is partial (Haraway 1988). Situated knowledge can thus be produced through the practice of reflexivity, whereby one’s location and positionality become part of the knowledge production process and evaluation of research outcomes, as well as being mindful of and accounting for the contingencies and partialities of all knowledge produced. Reflexivity is a research method of situating oneself and analyzing these complexities and challenges as well as the unequal social relations that infuse academic endeavors. Embracing situated knowledge means being critically conscious that power relations can produce “others” in the research process and that researchers are themselves part of the process that produces such differences. Self-reflexivity is thus a mindful way to situate the researcher themselves in this very process, to acknowledge the embodiments and scales involved, and to locate the researcher along a range of issues that can be social, political, economic, cultural, religious, or geopolitical. This acceptance of the location of variously situated subjects and the relational aspects of knowledge production thereby leads to an awakening toward the negation of epistemologies of all-knowing at all times or any absolute truths. Reflexivity thus is a critique of the God-trick (or the view from nowhere) (Haraway 1988) as it grounds the researcher and the research process in bodies, places, spaces, and scales. Embodied subjectivities of the researcher and all those who participate, or not (whether by choice or not), thus become central to the critical analysis of any research process and its outcomes.

There are many interpretations and uses of reflexivity and numerous personal and professional concerns that come with it. Scholars grappling with reflexivity are generally committed to maintaining a critical eye to the pitfalls and challenges while attempting to apply reflexivity in productive and meaningful ways. Impacts of the researcher on the research situation and shifting identities become part of the accounting process in research where reflexivity enables researchers to respond to and analyze different aspects of fieldwork accordingly. Reflexivity requires not only assessing and accounting for the researcher’s multiple locations across a series of intersections (race, class, gender, dis/ability, sexuality, nationality, etc.) in relation to others in the research process, it also raises questions of research ethics. Ethical research places importance on being conscientious, reflexive, and morally tuned. The intrinsic importance of critical thinking and ethical engagement are foundational to reflexivity as well as detailed methodological accounts and careful representation.

Since reflexive researchers attempt to demonstrate that research is about social construction and not purely representational, the process before, during, and after research becomes important. Social construction of experiences and interpretations are thus explicitly accounted for. Trying not to speak “for” others, but enabling collective construction of research data and output is often central in reflexive research. Reflexivity can foster rigorous research and encourage critical engagement with one’s methodological choices and outcomes. As such, thorough and rigorous documentation and evaluation of methods and techniques become necessary as well as multiple forms of capturing experiences, narratives, and voices. How people and places are documented and constructed become critical moments of being reflexive about texts and writing. This sheds light on silences and absences as well as any hyper-visibility or overstating of any particular issue at the expense of other issues in the research. By sharing a reflexive research process, the researcher opens it up for reinterpretation and scrutiny by others. Readers are able to judge
and assess methods, content, findings, and the production of knowledge. This public nature of reflexivity can be empowering as well as can open up further conversations and analyses. Validity of data produced or claims made can be gauged against the process through which knowledge was produced.

Being reflexive can also be a political act. This often means a conviction to making a difference in the lives of those involved in the research process rather than simply producing academic outcomes. Such political overtures of reflexivity can be complex and protracted. While being self-conscious does not necessarily rectify problems of representation and power, and the process may or may not result in meaningful research for the peoples and places involved, it has the potential to be so. What becomes important in undertaking reflexive research is to critically account for the research process historically, socially, and spatially. This enables upfront analyses of issues such as racialization, gendering, post/coloniality, geopolitics, and power relations that are intersectional. Being critically mindful of colonial legacies, imperialism and empire, politics of development, controversies of globalization, and other broader cross-scalar and historical issues enables researchers in “other” places to locate themselves on a broader canvas that then enables more comprehensive understanding of situations and contexts. Reflexivity thus should not just be about interpersonal relations but also about a range of cross-cutting and cross-scalar issues that constitute the researcher’s subjectivity (Sultana 2007).

Despite reflexivity having become a key aspect of qualitative research processes, concerns have been raised along multiple lines. While an unreflexive research process would not be encouraged or espoused by most human geographers, emerging concerns raise important questions about theorizing and practicing reflexivity. One concern is that it can become a superficial apology for being of the dominant group that usually has authority and power. While acknowledging being privileged through gender, race, class, ability, nationality, or any other axes of difference may signal a reflexive consciousness, a simple acknowledgment does not necessarily address issues of power and privilege (e.g., noting the researcher as white, male, able-bodied, American or British academic – the largest demographic within the discipline of geography – does not necessarily change anything). This is particularly troublesome when the voices, opinions, and realities of the less powerful or “others” are subsumed into the dominant narrative. This can end up reifying differences and reinforcing dominance as it reasserts attention to the researcher.

Refocusing on the privilege of the researcher has thus become a concern in overly subjective and insufficiently analytical narratives. If this actively shifts conversations and relations of power then this emplacing of the researcher may be more productive; however, it can also tend to silence and marginalize others while reinforcing the privilege of the researcher, however conscientious and ethical the desire may be. This dilemma has been a topic of debate for quite some time among feminist and postcolonial scholars (England 1994; Wolf 1996; Bondi 2003; Domosh 2003). Thus, reflexivity has the risk of being apolitical. It can also obfuscate analytical purchase of a research process if it is mired in auto-ethnographical minutiae. This may end up reinforcing the position of dominance of the researcher thereby absolving them from meaningful engagement, or end up reproducing an authoritative voice that actually can silence others. Reflexive acts can be misused to justify a research process or writing thereby negating the actual goal of reflexivity.
Another area of concern has arisen out of the growing number of publications by researchers who reflect upon their research experience in a way that can range from navel-gazing to tortured autobiographies. This self-indulgence can deflect attention to the positionality of the researcher without sufficiently accounting for a reflexive research process or contributing to debates around reflexivity. It has become increasingly common for geographers in recent years to produce such publications, some with greater success in moving scholarship forward than others. Merely recounting one’s research experience does not necessarily contribute to debates around reflexivity or demonstrate reflexivity. But the reflexive turn in human geography has resulted in scholars seeking acknowledgment of their research process through production of post facto auto-narratives that often lack self-examination or a transformation of the research process.

Beyond the problematic uptake of reflexivity and concomitant proliferation of publications, what is also of concern is that enforcing reflexive accounts of research may result in surveillance and domination by more powerful scholars, especially senior colleagues over junior colleagues. Unequal power relations outside of the research process can thus impinge on the extent and usefulness of reflexivity for emergent scholars and students as consequences can have deleterious impacts on the trajectory of the scholar. This raises some dilemmas for rising scholars who may be in precarious positions or liminal roles. Thus, reflexivity can be a double-edged methodological and scholarly device in that there is increasing pressure to demonstrate one’s reflexivity while at the same time it can have problematic outcomes for the scholar involved. Beyond academics, reflexive research and writing can also pose risks to research participants if sensitive material is inadvertently disclosed or hinted at. The emotional labor involved in reflexive practices can also be a discouragement to those who wish to employ it. Furthermore, the partiality of reflexivity should be acknowledged in that it may not be possible to be fully reflexive. There are thus many ongoing challenges to practicing reflexivity.

However, reflexivity can be very productive of political acts and collaborations and is generally espoused in order to make a difference not only to the research process but beyond it (Rose 1997; Bondi 2002). The simultaneous nature of being reflexive and living in the moment, and adapting and responding to insights generated, can produce useful information beyond the research project and foster relationships that can be generative. Thus, reflexivity does not have to be a self-indulgent or cathartic act, but rather must move beyond methodological tools to improve one’s research and the outcomes of the research process toward more meaningful impacts in the long run. For instance, it can encourage instances or events to be reinterpreted and rethought and thus generate new insights on hidden aspects of a situation that may be fruitful to explore or challenge. It can also have the potential to be politically and socially relevant when research data and experience are carefully and consciously analyzed and processed. Action research that engages the researched as researchers is increasingly being pursued in order to address concerns of privilege, voice, representation, and power. Reflexivity can be the basis of engaged activist scholarship and not just a corrective measure in the research process. Radical vulnerability and radical empathy can be cultivated out of reflexive praxis and thereby enable scholars to become scholar-activists that foreground politics, context, and situations. Collaborative endeavors can emerge from reflexive research and writing and new avenues of collectivization, solidarity, and empowerment.
can result, for instance, around issues of social justice (e.g., Sangtin Writers and Nagar 2006). Such formations and relationships can only be sustained when all participants engage in reflexivity through the process and thereafter.

Practicing reflexivity can be complex and challenging, and there are limits to the insights gained from it. One of the key challenges involves sustaining reflexive relationships and research processes over time, space, and life cycles. Another is the location of this practice within the academy amid growing pressures to produce large volumes of research output and publications quickly. The pitfalls of the extensive need of continuity of engagement and connection can discourage academics from engaging in such scholarship. Similarly, the commitments of time, energy, and taking risks can also discourage research participants from being involved as they may have other priorities of survival and life. As a result, while reflexivity can generate insights and foster relationships to address inequities not only in the research process but also in the wider world, the logistics of undertaking and sustaining reflexive work that has the potential to be politically and socially relevant may be curtailed by a range of factors. Constraints may be overcome depending on the context and people involved, and offer great potential for transformative politics, but continue to pose challenges that may be difficult to resolve in increasingly neoliberalized academic contexts where rapid publication outputs are often valued over other forms of scholarship and engagement.

Despite such challenges, reflexivity is a critically important aspect of research and an important skill in which students and scholars should be better trained so that they can incorporate such practices in their own work. Reflexive research and praxis needs greater engagement and acknowledgment within academia and beyond.

**SEE ALSO:** Feminist methodologies; Fieldwork in human geography; Positionality; Power; Subjectivity

**References**


Refugees are individuals or groups who have been forced to leave their home and have sought refuge across an international boundary. Refugees pose a series of challenges to nation-states in terms of classification, accommodation, and the human rights of those outside the formal state system. In this way, the study of refugees sheds light upon the nature of citizenship in contemporary geopolitics. While refugees form one specific group of forced migrants, their reception, representation, and treatment is often conflated with other migrant groups and with concerns over immigration and border security. A strict legal definition of a refugee is outlined in Article 1 of the 1951 Convention Relating to the Status of Refugees, which states that a refugee is a “person who owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion, is outside the country of his nationality, and is unable to or, owing to such fear, is unwilling to avail himself of the protection of that country, or who, not having a nationality and being outside the country of his former habitual residence as a result of such events, is unable or, owing to such fear, is unwilling to return to it.”

The legal status that defines refugees distinguishes them from other categories of migrant. In particular, asylum seekers are those who have applied for refugee status but are awaiting a decision on this claim. The forced nature of displacement for refugees and asylum seekers differentiates them from other forms of international migrant, such as economic migrants who travel in order to work. Many refugees travel alongside other migrants and increasingly geographers have argued that distinctions between refugees who flee persecution and migrants who seek to escape systemic poverty are unhelpful as distinctions between “forced” and “voluntary” migration are increasingly blurred.

Refugee histories

The term refugee was first applied to Protestant Huguenots who fled France in 1685. It was not until the twentieth century, however, that refugee populations became more significant. The events of World War I, the Russian Revolution, and the collapse of the Austro-Hungarian empire produced a displacement of people across Europe on a scale not seen previously. This led to the appointment of the first High Commissioner for Refugees by the League of Nations in 1921. The powers of the commissioner were strictly limited and an initial focus on refugees diminished as states focused on concerns over economic and political stability throughout the 1920s and 1930s.

Following World War II, the 1951 United Nations Convention Relating to the Status of Refugees was established as a means to define international obligations to refugees and to provide a definition of the refugee. The Convention was limited to those persons fleeing events before January 1, 1951 within Europe, thus marking a concern to address European groups displaced as a result of World War II. In 1967 the Convention
was amended though a Protocol that removed these geographical and temporal limits and gave the Convention universal coverage.

The 1951 Convention established two central principles that have shaped governmental approaches to refugees. First, Article 31 of the Convention prohibits states from penalizing refugees for illegal entry into a state when the purpose of their entry is to claim asylum. Asylum seekers and refugees are thus distinguished from “illegal” immigrants as their actions in entering a state are not defined as illegal. Second, the obligation of “non-refoulement” (literally to “push back” or “return”) requires that a state may not return refugees to any territory where their lives would be threatened. This obligation seeks to ensure that states are not able to remove refugees to conflict zones or to the sites of persecution they have fled. In order to manage refugee protection, the 1951 Convention saw the establishment of the Office of the United Nations High Commissioner for Refugees (UNHCR) who acts to coordinate resettlement programs and refugee provision.

The question of who qualifies as a refugee remains a contested issue, however, with nation-states often using the humanitarian associations of refugee status for political gain. For example, during the Cold War refugee status was used for ideological purposes, with capitalist states offering refuge to Soviet dissidents as a means to critique state socialism. While the definition of the refugee has not altered since 1951, it has been critiqued for its Eurocentric nature and its focus upon individual forms of violence and persecution. Asylum seekers must provide evidence of specific instances of persecution in order to gain refugee status and must demonstrate that they would face significant risks if returned to their countries of origin. As such, more systemic or generalized forms of violence and persecution may be disqualified.

For example, refugee groups displaced by civil, ethnic, and communal unrest, such as those fleeing the uprisings of the “Arab Spring,” are often unable to demonstrate the individual level of persecution required to meet refugee assessments on a case-by-case basis. Similarly, the need to cross an international border in order to qualify as a refugee has been criticized, as displacement through persecution may not necessarily lead to cross-border mobility. It is for this reason that after sustained pressure, UNHCR began to recognize the category of “internally displaced persons.” This category refers to those “who have been forced or obliged to flee or to leave their homes or places of habitual residence, in particular as a result of or in order to avoid the effects of armed conflict, situations of generalized violence, violations of human rights or natural or human-made disasters, and who have not crossed an internationally recognized State border.” The definition offered here is broader than that of the refugee, reflecting situations of generalized violence; however, it is important to note that UNHCR can only offer a set of “principles” for the protection of internally displaced persons and cannot compel states to enforce them.

Geographers have been critical of the focus on the national scale within refugee studies and of the inscription of national boundaries in refugee law, arguing that refugee status may serve to uncritically reproduce the authority and exclusions of the state (Mountz et al. 2002). Citing the complexity of migration within globalization, many have argued that other scales and sites may be more appropriate to deal with issues of migration, be that in terms of international flows of variegated mobility, cross-border networks of policy and practice, or the connections and social relations of world cities. Increasingly UNHCR are themselves publicly recognizing that the complexity of global mobility
undermines distinctions between refugees and other forms of migrant. As the reasons for individuals fleeing become ever more difficult to untangle and adjudicate, critical geographers and political scientists are challenging the practical and political value of refugee classification. For example, Nyers (2006) argues that the history of refugee institutions serves to construct refugees as temporary fixtures within global politics, as individuals or groups marked by conditions of crisis, and posing a challenge to the territorial assumptions of the state. Refugees in this sense exist as the binary opposite of the citizen, with the solutions proposed to refugee “crises” being to turn refugees into citizens – either through resettlement, integration, or repatriation. All of these solutions uphold the position of the state as the guarantor of rights and political belonging, and position the citizen as the “normal” and aspirational model of political identity.

Refugee resettlement

One of the highest profile “solutions” to refugee displacement is through international resettlement. Resettlement denotes a process in which preselected groups of refugees are assisted in moving to a new country by charities, NGOs, state agencies, and UNHCR. This process may be led by individual states through international agreements or by UNHCR in partnership with particular states. In recent years, the largest resettlement programs have been run by the United States, Canada, and Australia. In each case, governments coordinate resettlement programs with UNHCR in order to identify annual quotas for refugee resettlement and to negotiate the conditions under which they will accept refugees. Conditions include the prescreening of potential refugee groups to assess their needs, their ability to integrate, their level of vulnerability, and the urgency of their resettlement requirement. UNHCR view resettlement as one of the few “durable solutions” for refugee groups who would otherwise be confined to indefinite periods in refugee camps.

Resettlement involves a high level of scrutiny, from both “host” governments and UNHCR, and on arrival many resettlement schemes provide a high level of support. For example, in the United Kingdom the Gateway Protection Programme resettles an annual quota of refugees across the country. On arrival these groups are provided with housing support, a cultural orientation program, language training, and support to move into education or employment. Resettlement is thus a significant commitment from a receiving country as they are asked to provide both legal and physical protection and to aid refugees in becoming naturalized citizens. This contrasts with asylum seekers and other forms of irregular and regular migrants, who are rarely provided with such extensive support and are often restricted from accessing pathways to citizenship.

Despite offering a “durable solution,” refugee resettlement remains rare. UNHCR suggest that they are only able to put forward for resettlement around 1% of the refugees who fall within their remit, and not all of these are selected by “host” governments. The highly conditional nature of such programs has led geographers to be critical of their implementation and of the messages they convey about the politics of humanitarianism. In many respects, resettlement programs represent idealized forms of state managed response to refugees. Most often connected to a language of state “hospitality,” resettlement programs garner significant media attention, and in many cases public goodwill, not least because governments are keen to portray resettled refugees as safe, secure, and prescreened individuals who arrive in regulated numbers. By contrast, refugees who
arrive through other migratory routes or who are awaiting decisions on asylum claims may be presented as less “worthy” of a “hospitality” that is strictly limited.

Refugee governance

Refugee resettlement represents one means of managing refugee populations, yet this is far from the most common mode of refugee governance. Rather, three forms of spatial and social governance can be seen as conditioning approaches to refugees worldwide.

The first of these is through the construction of refugee camps. Refugee camps are constructed by governments or NGOs as a means to temporarily accommodate refugees, although in many cases such sites become semipermanent and have been argued to take on urban characteristics in their most long-standing forms. From a humanitarian perspective camps may be problematic in terms of the spread of disease, difficulties in securing camps against violence, and their precarious nature as informal and temporary settlements. At the same time, they offer opportunities to manage refugee populations through the coordination of resources, health care, and support. In his ethnographic work in the camps of northeast Kenya, Michel Agier (2010) has argued that it is through these forms of humanitarian provision that camps have become semipermanent structures, with their own systems of power, social hierarchies, and norms of everyday life. It is not only humanitarian action that enables the maintenance of camps, however; the policies of many states that seek to spatially segregate refugees from the wider population enable camps to be maintained as designated spaces for refugee groups. Providing such a space not only allows for the containment of the refugee population, it also works to prevent refugee integration and to reinforce the temporary nature of refugee identity. In discussing such spaces, geographers have been influenced by Giorgio Agamben’s notion of the “camp” as a spatial expression of sovereign abandonment, arguing that within camp spaces we might witness the inclusive exclusions that lie at the heart of contemporary geopolitics.

The significance of refugee camps can be seen in contrasting these spaces with recent discussions of urban refugees in the Global South. Despite refugees having long been attracted to urban environments, in part due to the opportunities for informal employment and anonymity, it was not until 1995 that UNHCR first reported on urban refugees. This report compared urban refugees unfavorably to encamped individuals who were presented as the “model” refugee. Only in 2002 did UNHCR recognize the need to deal with urban refugees in a more nuanced manner, highlighting the specific issues they may face, such as a lack of formal documents, everyday harassment and exploitation, the insecurity of informal housing, and fears of detention. Despite this, the priority of both governments and NGOs has been to place refugees in camps. Urban refugees pose a series of challenges for states as they are less easily distinguished from other marginalized groups and harder to regulate. At the same time, the self-directed, autonomous, and often more mobile nature of urban refugees challenges idealized notions of the displaced as helpless and dependent subjects to be fixed to specific camp locations.

The second form of governance arises through the mechanisms of refugee classification that are associated with asylum procedures across Europe, North America, and Australia. A range of governmental practices have been examined by geographers as means to control asylum seekers within countries while decisions on refugee status are taken. These include
practices of dispersal evident in countries such as Denmark, Germany, Sweden, and the United Kingdom, where asylum seekers are accommodated in specific locations so as to “spread the burden” on services. Significant restrictions and conditions apply to such accommodation, including the need to report to immigration authorities on a regular basis and constraints are placed on the ability of asylum seekers to seek employment.

The refugee determination process also produces distinct legal geographies. For, although the 1951 Convention establishes a single definition of the refugee, the interpretation of the grounds for such status differs markedly between states. For example, in 2011, recognition rates across the European Union (EU) for Afghan asylum seekers varied widely from 33% in Austria to 4% in Greece. Similarly, some countries currently implement a system of “safe country” lists, which identify countries where governments believe persecution does not take place and therefore asylum applications are readily refused from such countries. Furthermore, across the EU, the Dublin II agreement means that those seeking asylum may only do so in one member country, and if it can be shown that asylum seekers have traveled through a “safe country” to reach their destination they are returned to this first “safe country.” This serves to relocate asylum seekers from northern to southern Europe, and increases asylum claims in countries, such as Greece, with lower refugee recognition rates. At the same time, the history of refugee determination also illustrates a series of exceptional events where states have established resettlement schemes that go beyond the 1951 convention. For example, in 1999 a range of European countries resettled those fleeing the conflict in Kosovo, and in 2012 Norway, Denmark, and the Netherlands resettled Bhutanese refugees following landslides and flooding in Nepal.

Alongside moves to relocate asylum claims has been an international growth in practices of deportation through which states forcibly repatriate individuals whose claims for asylum have been rejected. Over the last decade, many states have used geopolitical influence to encourage countries of origin to accept deported migrants and to extend the scope of deportation agreements. Deportations to countries such as Afghanistan and Iraq are controversial given the obligation of non-refoulement, yet many governments in the Global North utilize deportation as both a border enforcement mechanism and as a symbolic measure of deterrence against asylum seeking.

Finally, the governance of refugees is intimately tied to discussions of border security. As refugees and asylum seekers have increasingly been conflated with “illegal” migrants, so have many states sought to extend their border control mechanisms and screening practices to extra-territorial spaces. Governments in Europe, North America, and Australia have all taken steps to enforce border controls at ports of origin for many refugees. At the same time, the growing use of detention centers in remote locations, such as the Pacific islands, the Australian outback, and remote military bases, have been highlighted as means to keep asylum and refugee groups isolated from advocacy movements and public visibility. These practices of offshore processing and detention seek to govern those awaiting refugee status at a remove from the territory of the state and the social life of citizens.

Throughout these different mechanisms of governance, practices of social and spatial isolation are notable in dealing with refugees as unwanted and unwelcome migrants. As a result, it can be argued that what unites strategies of refugee management is a desire to manage refugees in the interests of the nation-state,
rather, necessarily, than in the interests of refugees.

Refugees as political actors

While much work has explored how refugees may be produced by state violence and excluded through state practices, recent studies have emphasized the political agency of refugees. Refugee camps have thus been argued to represent spaces for the formation of community and solidarity, such that political possibilities arise in these environments despite their exclusionary nature. Similarly, recent examinations of anti-deportation campaigning and No Borders activism have highlighted how refugees may make claims upon the state despite lacking citizenship rights, working alongside citizens who view immigration controls as inherently racist forms of division (Nyers 2006). Studies of sanctuary movements in Europe, Canada, and the United States are also notable in advocating for the rights of irregular migrants and in seeking to shift public opinion on refugees (Darling 2010). While at times risking the reproduction of an image of the “needy” humanitarian victim, such movements have shown that issues of refugee support and rights are not wholly situated on the terrain of an exclusionary politics which views refugees as a threat to the stability and integrity of the nation-state.

Refugee experiences

Informing these studies has been a rich vein of work focusing upon the experiential aspects of refugee identity. Here studies have explored how state discourses of humanitarianism and border control are translated into prosaic practices which condition the lives of asylum seekers and refugees across the Global North, and the Global South. Links with studies of transnationalism and social networks have also been forged in order to trace the connections that sustain everyday life for refugees (Koser 1997). For example, Hyndman and McLean (2006) examine the experiences of Acehnese refugees resettled in Vancouver, and highlight the challenges they face in creating livelihoods despite their legal status within Canada. Taken together, this body of work demonstrates how geographers have begun to explore the social and cultural dynamics of refugee identity, the everyday conditions in which refugee governance is enacted and the multiple networks that contribute to the formation of “home” in conditions of exile.

Summary

Geographers’ discussions of refugees have been diverse and wide ranging, encompassing the spatial governance of refugees, the production of the refugee as a legal and political category, and the sociospatial experiences that accompany such a category. Refugees and asylum seekers are fundamental products of the nation-state system and, as such, are likely to remain central actors in debates over migration, citizenship, and mobility. Looking forward, the question of how refugees are defined and what limits are placed on their ability to access rights may be a key arena in which geographers can contribute. For example, a rapidly growing body of work is concerned with “environmental” refugees who are not covered under the terms of the 1951 Convention, but who may be forced to migrate through natural disasters and longer-term environmental degradation (Gill 2010). Exploring the complex
interactions between human and environmental processes that lead to displacement offers one avenue through which geographers may highlight the limits of current definitions of the refugee. It is this critical and spatially informed lens that geographers have brought to the study of refugees, and it is through such work that they have sought to support and advance claims for social and spatial justice.

SEE ALSO: Borders, boundaries, and borderlands; Citizenship; Migration: forced; Mobility gaps; Political geography

References


Further reading

Regional and interregional trade in producer services

William B. Beyers
University of Washington, USA

This entry begins with a discussion of early concepts regarding trade in producer services. The characteristics of this trade are then reviewed from multiple perspectives, including the role of trade in producer services revealed using various statistical sources and a review of surveys of producer services trade. The entry concludes with selected data from input–output models that document the growing importance of the demand for and trade in producer services.

Greenfield (1966) pioneered a vision for what we now accept as producer services – activities whose primary clients are businesses and governments, as opposed to households. The classic view of services as being perishable, being consumed in the moment of their supply, was challenged in Greenfield’s work, opening up the notion that services could be consumed (traded) other than at the time of their production. Greenfield conceptualized “semidurable” and “durable” services, extending concepts introduced by Kuznets as a part of his pioneering work on national accounts. Since Greenfield’s work the modes by which the outputs of producer services can be traded are very different. Information and communications technologies were much less developed 50 years ago; the advent of networked computers to store information combined with telecommunications technologies now allow interactions in modes not even imagined in 1966. Entire categories of producer service industries that are accepted as routine today – such as provision of cloud computing services or investment services guided by modern portfolio theory – did not exist when Greenfield conducted his research. Text, data, and graphic files were certainly created at that time, but their storage and transmission (over media such as the Internet) has subsequently helped to facilitate the growth and development of producer services trade. We can anticipate future changes in producer services supply and demand in ways that are unpredictable at the time of this writing.

Trade is defined as a sale between a buyer and a seller, and from a geographical perspective we can classify where this trade takes place. The simplest dichotomy is between trade that takes place between clients and suppliers locally or in close geographic proximity, and nonlocal trade that takes place over some greater distance. Such interregional trade, however defined, can be defined in a local–nonlocal dichotomy, or within a multiregional framework with many regions of supply and demand.

Unlike trade in goods, which has been tracked through customs and international trade regulations, and reported in great detail (such as through the World Institute for Strategic Economic Research), trade in services has not been the subject of government statistical programs in most countries. Output or sales of services has been measured through series such as the Economic Census in the United States, but the geography of these sales is undocumented. National programs for the measurement of trade between industries are reported in input–output models, but these are typically one-region as opposed to interregional models, and are rarely explicitly international models. Data on the
geography of trade in services have come from two broad sources: ad hoc surveys and regional input–output models that have an export measure of sales. One of the highest priority topics for the expansion of official statistics from national and international government agencies is the development of robust statistics on trade in services. Given the dominance of services employment and output in the global economy, this is a major oversight in measurement programs that needs to be remedied. The reasons include the need for basic documentation of the structure of regional and national economies and the role or value that this type of data has for economic development programs.

It is useful to look back on how the notion of trade in services was treated in early literature on the regional economic base. The classic economic base model divides economies into two markets: export and local. Before direct measures of regional trade were undertaken, authors classified industries as exogenous or localized, but struggled with this dichotomy recognizing that some industries were “mixed.” A paper by Polzin is an example of how services trade was treated in this early literature: “Trade and service activities constitute the bulk of the mixed industries and the planner/analyst has little to lose by ignoring them in the construction of the exogenous sector” (Polzin 1973, 231). The statistical basis for this statement is not explained and it was not until surveys of the markets of service industries were conducted in the mid-1980s that it became evident that this conclusion was erroneous.

The unfolding of perspectives on producer services trade

In a discussion of types of services, Daniels noted that geographic markets are likely to be very different across them. Discussing markets for a greengrocer versus a consultant providing advice on the purchase of telecommunications facilities he wrote: “It is conceivable … that the specialist consultant has a market which extends throughout a country, either because no other organization is able to provide the service or because there is insufficient demand to justify the existence of more than one such firm” (Daniels 1982, 31). Daniels wrestled with location principles guiding the greengrocer, and the producer services firm, noting that central-place model location principles were less likely to apply to location choices of producer service specialists. Without mentioning the geography of trade explicitly, Daniels implied the potential for trade, especially in the face of the development of information technologies.

In 1985 Daniels extended the argument just made, with some of the first estimates of the magnitude of trade in services:

During the last decade trade in services has become one of the most rapidly growing and changing sections of international trade. It has been estimated that services now account for one third of world trade amounting to a massive $550 000 million in 1980. About 20 per cent of the world total is accounted for by the USA, with France and Britain in second and third place respectively (9.4% and 9.2% of world trade in invisibles). Until recently most of the services crossing international boundaries were ancillary to, or in support of, the trade of goods (such as the financial services which permit trade in goods, transportation or consulting services). Now services are traded in their own right: advertising, market research and management consulting, telecommunications services, computer services, specialized financial services in banking and insurance and business services. The spread of multinational enterprises has encouraged the growth of international trade in services, even though some of the transactions will be between the constituent parts of the same enterprise. Unfortunately there is a shortage of reliable statistics on international trade.
in services; it remains difficult to measure and analyse trade volumes, patterns and composition. (Daniels 1985, 35–36)

Ochel and Wegner focused on trade in services in Europe and explored the transformation of markets in the context of the “new economy”:

The internationalization of markets is facilitated by the growing tradability of services. Advances in communications technology and in information storage and processing have made it possible to produce services in one place and to consume them somewhere else and to produce services at one point in time and to consume them at a later date. Any service product that can be reduced to electronically coded pieces of information can theoretically be delivered to any point in the world with great reliability, at relatively little cost, and with no time-lag. Of course, there are many difficulties in implementing such an information system. (Ochel and Wegner 1987, 62)

Their argument seems to ignore types of services where face-to-face communications is important in delivering the service.

Improvements in technology were also important such that:

Trade in producer services is also growing because the cost of “transporting” services is falling over time. The long-term rise of trade in goods is in part due to innovations in transportation technology: navigation, shipbuilding, railways, trucking, and so forth. Improvements in communications technology (telephone, fax, television, courier services, etc.) have had a similar impact on trade in services. To this must be added improvements in the transportation of persons, lowering travel costs for “labor embodied” services. From this follows increasing interregional specialization with respect to services, bringing into play questions of comparative advantage, where services come to play an increasingly critical role in the definition of the economic base of regions. (Bailly et al. 1992, 25) This was later confirmed in a survey of producer service firms that identified increased use of digital modes of delivering services, but no reduction in the need for face-to-face meetings with clients. Bailly et al. (1992) also examined the channels of interregional trade in services, distinguishing between direct service trade, intrafirm trade in services, intrafirm service trade by primary sector and manufacturing firms, trade in services as a result of trade in goods, links with factor flows, labor-embodied services, and service flows versus factor flows. They argued that intrafirm flows are important and can be tracked using survey methods and that survey data on interfirm trade in services underestimates total trade flows.

Daniels returned to the issue of trade in producer services, observing that: “In order for services to become prominent actors in the world economy they must possess attributes that enable them to be exchanged across geographic space” (Daniels 1993, 25). He noted that some services are not tradable, and presented a taxonomy for service firm internationalization with regard to service multinational enterprises, international trade and foreign direct investment in services, the global system of cities, and the internationalization of services and the restructuring of cities. He also speculated about the desirability of service-dominated economies, the inevitability of the globalization of services, the unfulfilled potential of telecommunications, and whether IT advances will impact interaction:

Innovations in the applications of information and telecommunications technology are therefore set to continue. They are worth noting because their ultimate effect, threatened for years but yet to be manifested, may be to replace a global service economy dependent upon intensive transactional flows (people, information, electronic messages, etc.) between a small number of large, highly centralized control-points with a more dispersed
telecommunications-dependent network of transactions. Flows of human capital will be far less significant though still necessary for the conduct of high-level negotiations and information exchange. (Daniels 1993, 175–176)

He went on to speculate about the role of telework and the possibility of business cycles in advanced services.

Another evaluation of the means of exporting services was offered by Reif and colleagues.

There are different ways of selling services across borders. Direct selling is by no means the primary method for residents of one country to provide services to residents of another. … International Consulting. International consulting, also termed transient service exports, occurs when a US businessperson renders a service to a foreign client on a short-term basis. … Direct Exporting. Because the value of some services is embedded in the tangible products associated with these services, the export of a service sometimes can be shipped much like a manufactured good. Examples of this type of service export include books, pharmaceuticals and software. … Telematic Trading. Telematic trade is another means of exporting a service and one which involves the transfer of knowledge and information abroad via telecommunications. … Royalties and Licensing Agreements. According to the Bureau of Economic Analysis, royalties and licensing fees consist of receipts and payments for the use of patented techniques, processes, formulas, and other intangible property rights used in goods production, as well as copyrights, trademarks, franchises, rights to broadcast live events, and other intangible rights. … Franchising. A franchise is a business arrangement in which an established business gives an independent party (the franchisee) the rights to use its brand names or trademarks, and transfers knowledge which is essential to running the business. … Providing services to foreign visitors. This export can be simply defined as a service performed and completed in the United States and sold directly to a foreigner while in the United States. (Reif et al. 1997, 5)

A recent assessment by the EU of the delivery of business services speaks to some of the issues surrounding the expansion of trade in services, including producer services. Four pathways to internationalization are identified and these could also be interpreted as pathways for inter-regional trade in a huge internal market such as within the United States. These pathways are as shown below.

Mode 1 Cross-border supply, which occurs when suppliers of services in one country supply services to consumers in another country without either supplier or consumer moving into the territory of the other.

Mode 2 Consumption abroad, which refers to the process by which a consumer resident in one country moves to another country to obtain a service.

Mode 3 Commercial presence occurs when enterprises in an economy supply services internationally through the activities of foreign affiliates.

Mode 4 Presence of natural persons describes the process by which an individual moves to the consumer’s country to provide a service, whether on his or her own behalf or on behalf of his or her employer. (European Commission 2014)

There are no data included in this EU report to illustrate the relative importance of these pathways.

Broader arguments regarding trade in services

The emphasis in the preceding section of this entry has been on the growing role of inter-regional trade in services, but it is important
to recognize that local demands for producer services remain important. This can be measured using a two-region model inspired by interregional input–output models of trade in all economic activities with explicit producer service trade modules linked to intraregional and interregional demand (Beyers 1990). Figure 1 illustrates this two-region system; in each region there is a block of local final demand (personal consumption expenditures, investment, and government) and a block of local income which is linked to the local production system (producer services and the transformative, distributive, retail, not-for-profit, and mainly consumer services sectors). Exports and imports are shown for all sectors in this fully enclosed system and it was argued that, on the basis of these connections, if demands rose or fell in one region this would lead to interregional consequences across all elements of the multiregional system.

Using data for the Bureau of Economic Analysis (BEA) economic areas in the United States, Beyers used the system illustrated in Figure 1 to estimate producer services trade using location quotients. Beyers concluded:

The growth of producer services employment in a given region should be seen to be a function of the growth of producer services exports and exports of other key industrial sectors. These changing export demands in turn lead to local direct requirements for producer services and other inputs, and they directly create local income. These direct requirements lead to a chain of indirect effects, which have associated with them additional demands for producer services and other sectors that also create additional income. Expenditure of locally created income leads through the consumption and government expenditure process to additional indirect and induced effects, including local final demands for producer services. … In regions with a strongly expanding export base, this set of linkages appears to have led to strong growth in producer services, and the opposite has been the case in regions with a declining or slow growing export base. … Over time, the importance of these interregional connectivities may be increasing in the producer services as businesses are able to extend market areas through the use of advanced information processing and telecommunications technologies. (Beyers 1991, 167)

Returning to trade in services in 2007, Daniels reminded us of the traditional argument for service production and consumption to be co-located because of the “non-tradable” nature of many services. However, he argued that this restricted view has been modified by the advent of information and communications technologies (ICT), including the associated increases in service sector productivity. These forces have led to services globalization with four dimensions: flows, networks, frameworks, and brands and trademarks. Daniels offers a portrait of trade in these dimensions benchmarked against the year 2000; the system of flows is similar to Figure 1, but reports estimates of inter- and intraregional flow (Daniels 2007). This rich diagram documents strong intraregional and export trade in Europe, the Americas, and Asia and Oceania. Implicitly, imports are captured in the system of trade flows documented in Figure 2. The size of the boxes in this diagram is proportional to total exports across the four regions, divided into intraregional and interregional trade. The high intraregional percentages for Europe reflect the large number of relatively small countries with large European markets. The export shares for each regional block add to 100%.

Daniels mentions a number of indicators of these dimensions, such as the availability of infrastructure and institutions allowing the four modes of supply identified above to operate for trade and foreign direct investment (FDI)
(for example, growth in services share of FDI and structural change within OECD countries in the composition of services FDI, especially the strong growth in the share of business services and transportation and communication services, and decreases in the share of trade, financial intermediation, and other services (subject to strong cyclicality)). Another indicator is a new international division of labor allowing integration of production systems across national boundaries (such as the offshoring of telephone-based support systems to places such as India by United States or European service firms), particularly by multinational corporations (MNCs) that threatens services job creation in the United States or Europe.

Producer services trade: survey research

Beginning in the 1980s there was a burst of survey research on the markets of producer service businesses. It is fair to say that this stream of research was stimulated by the lack of industry detail in official statistics and in regional input–output models that presented data for producer services in a highly aggregate manner. It is worth noting that there has not been a comparable body of research undertaken since the turn of the century, so it is not clear whether and how the conclusions of this earlier research may be altered by current data that show continuing increases in the magnitude of trade in producer services.

Beyers and Alvine reported on a large-scale survey of the markets of 2200 producer service establishments in the Puget Sound region of Washington State that found half of them had sales of at least 10% outside the local area (Beyers and Alvine 1985). Of the establishments that met this criterion the average percentage of external sales was 55%, with the overall level of export markets about 35% when merged with the localized establishments. There was no correlation between firm size and export market percentage although a large cohort of establishments that had experienced growth in their export percentage also expected an expansion of export share in the near future. Major variations in export propensities were reported by industry with accounting having a low export market percentage while other industries, such as research and development, reported a very strong export market percentage.

The channels of trade in producer services, including (i) direct service exports, (ii) intrafirm service exports by producer services firms, and (iii) intrafirm service exports by manufacturing firms were surveyed by Coffey and Polèse. They reported that
With the exception of engineering consultants, it appears that direct sales to other regions or nations are not a major element of total receipts. Direct sales for most sectors appear to be highly sensitive to distance or to other barriers. Language may act as a barrier, but other obstacles to the direct “transport” of producer services may also exist: local purchasing policies, professional licensing requirements, work permits, and so forth. Such constraints are often associated with sovereign states and thus seen as barriers to international trade, but their impact may

![Figure 2](image-url)  

*Figure 2* World and intraregional shares of services exports, 2000 (Daniels 2007, 106). Reproduced with permission from Edward Elgar.
also be considerable in federal states and other decentralized nations. Local purchasing policies and the provincial licensing of professions are traditional practices in Canada. Given such barriers and the high communications costs (for outputs), producer-service firms will be tempted to establish local offices to penetrate outside markets, the second channel by which producer services are exported. (Coffey and Polèse 1987, 600–601)

This last channel has a value of about half that of direct exports. Table 1 reports data on markets of Montreal producer service firms from this survey.

A survey of Vancouver, British Columbia, producer service firms measured their markets by sector served (primarily sales to other service industries) as well as their geographic reach (Ley and Hutton 1987). The study found 51.4% of firms had greater than 75% of their sales in the Vancouver area, 20.4% had at least 25% of sales elsewhere in British Columbia, 20.9% had at least 10% sales elsewhere in Canada, and 13.3% had sales of at least 10% elsewhere in the world (percentages do not add to 100% because of the way in which this survey was phrased). The respondents also thought that Vancouver would be a major source of growth (68%), along with the rest of British Columbia (20%). This contrasts with later research that showed a sample of firms anticipating much more growth in markets outside Vancouver (Davis and Hutton 1991). That paper also explored the idea that producer services were becoming increasingly tradable. This draws from Davis’s work on an input–output model of the Vancouver regional economy from the mid-1980s that showed strong exports of business services from the Vancouver area (44.7% local markets, 44.2% to other British Columbian, 7.2% to other Canadian, 1.5% to the United States, and 2.4% to the rest of the world). A related study in 1990 reported higher levels of exports from the Vancouver region for engineering, management consulting, and computer services businesses, while advertising was estimated to have strong local markets. This survey also asked firms about expected changes in markets, finding that they expected a larger share of exports: exactly what the Beyers and Alvine study reported in 1985.

Further corroboration was provided by surveys of the export markets of producer service firms in Edmonton, Alberta which showed an average of 36% in service exports – very similar to Beyers and Alvine’s documentation of trade in producer services in the Puget Sound region (Michalak and Fairbairn 1988). There were large variations across industries in export propensities; accountants largely served local markets, while engineering firms had the majority of their markets outside Edmonton. In findings similar to Coffey and Polèse, foreign markets were modest compared to nonlocal Canadian markets. Michalak and Fairbairn concluded, first, that a peripheral city such as Edmonton did not have strong producer services links to the Canadian economic core (Toronto/Montreal), implying a weak role for the core–periphery model in explaining markets for producer services. Second, the data suggested that the Alberta region was developing independently of the Canadian economic core. Third, they found that services exports were to places unrelated to their size. Fourth, they found market links with services and government and, finally, that a successful regional economy depended upon an adequate supply of producer services as inputs to the production process in ways that made their clients/users competitive.

A comparative study of producer services in Vancouver, Montreal, and the Puget Sound region generated findings that were remarkably similar with the Puget Sound region and Vancouver (Coffey and McRae 1989). A second important observation from the comparative
Table 1 Markets of Montreal producer service establishments.

<table>
<thead>
<tr>
<th>Service Type</th>
<th>% Montreal</th>
<th>% Other Quebec</th>
<th>% Other Canada</th>
<th>% Outside Canada</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management consultants</td>
<td>75.0%</td>
<td>12.2%</td>
<td>4.9%</td>
<td>8.1%</td>
</tr>
<tr>
<td>Advertising agencies</td>
<td>69.2%</td>
<td>6.2%</td>
<td>18.2%</td>
<td>6.5%</td>
</tr>
<tr>
<td>Computer consultants</td>
<td>66.2%</td>
<td>21.9%</td>
<td>10.7%</td>
<td>1.1%</td>
</tr>
<tr>
<td>Engineering consultants</td>
<td>57.5%</td>
<td>0%</td>
<td>16.0%</td>
<td>26.5%</td>
</tr>
</tbody>
</table>

Data on market orientation concerns the relative importance of service sector exports when allocated to the three categories of nonlocal sales. The Vancouver and the Seattle-based studies reported a very similar share of sales to customers located in British Columbia or Washington State respectively. However, Seattle area services exported a much higher percentage (p. 225) of the value of their output to buyers in other US states compared to the equivalent figure for Vancouver producers; at 5% this was even less than the 10% share of sales to the rest of Canada outside British Columbia reported by Ley and Hutton (1987). It is interesting to note, however, that Vancouver area service firms relied more heavily on international sales than their Seattle-based counterparts – which had international service revenue of only 3%, while the two Vancouver studies estimated an equivalent figure of 7% and 12% (Coffey and McRae 1989, 107–108). In relation to Montreal, a further dimension was identified by Coffey and McRae (1989) who reported a stronger share of markets within Quebec, implying an important role for language in mediating the geography of producer service markets. Their work also confirmed the reality that small firms are just as able to gain export markets as large firms. This was also a key finding in the Puget Sound region studies, although the Canadian studies reported somewhat weaker results on this score.

While the early research in the Puget Sound region was followed by several studies in and around Canadian cities, the next US study centered on Chicago did not appear until the mid-1990s (Esparza and Krmenec 1994). This study was partly driven by the proposition that major world cities strongly interact with each other and producer services transactions are part of that process. Esparza and Krmenec conducted a telephone survey aimed at documenting this type of connectivity, reporting average percentages of local sales for a sample, and a frequency distribution of percentage local sales as a share of the number of firms (this sample does not appear to be weighted by size of firm or size of sales). The number of linkages of individual establishments to other cities and by type of market was documented and showed that manufacturing was not the dominant market of producer services except for advertising firms. The linkages of the firms were mapped, revealing a two-tier system:

The results indicate that the supply of business services takes place within a two-tiered organization of cities. The first tier, which involves interaction among world-class cities, is insensitive to distance as an impediment to interaction. This confirms other research which argues that ‘gateway cities’ dominate interaction in the contemporary global urban systems. The second tier is organized within the US system of cities and more closely resembles a hierarchical structure. In this instance, distance appears to mitigate interaction over longer distances. (Esparza and Krmenec 1994, 44)

In the late 1990s it was recognized that trade in producer services would likely be influenced by considerations such as the adaptive behavior
of firms (Beyers and Lindahl 1997a). Using survey data focusing on changes in services sold, changes in client types, and changes in export markets, Beyers and Lindahl tested the Ansoff model of adaptive behavior with the results clearly indicating that producer service firms engaged in adaptive behavior had higher growth in sales than those that were not adaptive.

The behavior of producer services located in rural areas, rather than the city regions, was also a potential source of variation in local versus nonlocal transactions. Based on an analysis of a national sample of producer service firms in rural counties in the United States, Beyers and Lindahl (1997b) found that 43% of the firms were strongly engaged in selling their services in nonlocal markets, with most being niche market players. These firms were categorized as Lone Eagles (proprietors) or High Fliers (strongly exporting firms with employees). Quality of life, cost concerns, and proximity to the location of the firm’s headquarters were primary location considerations, along with other personal factors. Those working for export-oriented firms often traveled to meet with their clients and generated higher revenues per worker compared to locally oriented firms.

Finally, the results from eight surveys of producer services markets, some of which are referenced in this entry, were usefully summarized by Illeris. He reported that local markets varied from about 32% to 71% of sales, with an average of 49%. Put another way, the Illeris compilation indicates that just over 50% of producer-service markets were nonlocal or export (Illeris 1990, 146).

**Producer services trade: evidence from input–output models**

Useful insights on the scale and geography of trade in producer services are available from studies and reports that rely on a mix of primary and secondary data. An alternative source of data on the use of and trade in producer services is data from input–output models. In the United States these models are estimated annually for the national economy, and every five years at a more detailed industry level in “benchmark” tables. These models document the value of purchases of producer services by industries as well as households, governments, investors, and on export accounts.

Anne Carter reported on the increasing importance of producer services in a landmark analysis of structural change in the American economy utilizing US input–output models. Advertising was the fastest growing portion of the business service sector in the time periods analyzed (1939–1947 and 1947–1958) and, citing Greenfield’s work, Carter wrestled with the reasons for this growth. While there were “costs of coordinating a growing volume of specialized transactions,” externalization (the shift of provision of services internally in a sector to their external procurement) is also important: “Only part of the growth of the service sectors represents actual increases in the total volume of services performed. The rest is the transfer of service functions from firms and establishments primarily engaged in manufacturing and other product–oriented activities to specialized service sectors” (Carter 1970, 65). The data on this trend were sketchy, but she noted increasing contracting out for research and development services, for banks to increasingly perform record-keeping and information-storage functions that were formerly in the province of bookkeeping departments of customer establishments. Similarly, insurance companies assumed an increasing proportion of risks formerly taken by the insured companies themselves. Some of these services, such as workmen’s compensation, are mandatory. Specialized record-keeping and computer services have been growing rapidly.
Such business services range from very sophisticated data retrieval schemes to assisting small businesses in coping with tax regulations. To a greater and greater extent exploratory and trouble-shooting research on engineering and business problems have been delegated to consulting organizations. Advertising work is more centralized in specialized agencies. Complex equipment, including cars and trucks, is beginning to be rented more frequently than being bought outright. (Carter 1970, 65, 68)

Apart from the national level, input–output models at the regional level can also provide data on the demand for and trade in producer services. In practice such models are rare but a set of eight models stretching from 1963 through 2007 are available for Washington State. When combined with data from the US national input–output models, it is possible to compile a very useful portrait of changing levels of purchases and exports of producer services (Figure 3). FIRE (finance, insurance, and real estate) and business services are reported as a composite in Figure 3. Transportation services appear to have been a fairly constant share of inputs over the history of the Washington State input–output models. In contrast, the share of purchases accounted for by wholesale and retail trade has risen over the 1963 to 2007 time period.

Figure 4 shows a clear upward trend in the economy-wide purchase of producer services as a share of output in the United States. It also shows Washington State purchases as a share of output over time, along with the difference between national purchases and Washington regional purchases (these are interpreted here as the implicit imported share of producer services.
Figure 4  Producer service purchases as a share of output.

Figure 5  Implicit imports of producer services to Washington State businesses.

Figure 4 is derived from the history of the Washington and US input–output models and shows the percentage of total purchases accounted for by producer services from 1963 to 2007. (The author has estimated values for some of these years as detail was not sought in the surveys for these purchases.) The large increase in the 1987 to 1997 time period is undoubtedly related to the shift from the Standard Industrial Classification System (SIC) system to the North American Industry Classification System (NAICS), in which some activities were reclassified from industries such as manufacturing to producer services. One example of this type of reclassification is headquarters activity, which was classified as “administrative and auxiliary” employment in each broad class of industry (such as manufacturing or construction) in the SIC system, but transferred to “management of companies” in the NAICS system. The Washington businesses). These estimates are approximate, since the industry mix in Washington differs from the United States, and patterns of purchase within particular industries may differ between Washington State and the same industry elsewhere in the United States.
Regional and interregional trade in producer services

State regional requirement is the sum of values reported in Figure 3 for FIRE and Business Services.

An increase in the share of production costs accounted for by producer services over time is clearly evident in Figure 3. There is a split in this figure between Washington State production as measured in the Washington input–output models and between imports from the rest of the United States, implied by US input–output models. Estimates of the share of imports suggest that they have diminished somewhat over the time period included in these figures, suggesting that Washington producers have supplied a somewhat larger share of total demand for producer services over time (Figure 5).

The Washington input–output models also allow estimates of exports as a share of total sales over time. These have been calculated from the series of Washington input–output models (Figure 6). Several of the early Washington models did not disaggregate services, such that producer services could be identified, but estimates of exports are possible for the years for which these data are available. Two sets of measures are included: a narrow definition only includes FIRE and business services, while a broader definition also includes transportation services and communications. These export measures are presented at three levels – domestic exports, domestic exports plus sales to the federal government, and domestic and foreign exports, including sales to the federal government. There is a clear overall upward trend in exports of producer services (narrow definition) with important growth in foreign exports. This graph shows very strong growth in estimates of traded producer services between 1967 and 1987, and then relative stability in the export share since that date. Sales to the federal government as a share of output have declined and domestic exports as a share of output (not absolute value) are smaller while the share of foreign exports has risen.

The general lack of interregional trade data for producer services in the United States needs to be emphasized. The growth in exports documented in many surveys undertaken during the 1980s and 1990s is mirrored in the input–output data. This suggests that the inferences made here about imports to Washington businesses of producer services are also probably correct.

Distinguishing between exports of services to Canadian provinces and those abroad, Stabler and Howe also used input–output models for 1974 and 1979. Direct exports of services from Manitoba increased from 24.37% to 42.61%, in Saskatchewan they increased from 7.1% to 27.82%, in Alberta they increased from 22.59% to 23.8%, and in British Columbia they increased from 21.55% to 30.57%. These data showed a strong increase in the share of exports to the rest of the world compared to places within Canada between 1974 and 1979 and the share of total economic activity in the provinces related to all exports also rose. Stabler and Howe concluded that: “At the margin, service exports appear to have accounted for between at least one quarter and perhaps as much as one-half of that portion of real growth which could be attributed to an increase in exports from these economies” (Stabler and Howe 1988, 313)

Input–output data complement surveys of producer services markets. The overall use of producer services as an input in production processes has grown, with expansions of supply within regions, at the same time as exports (and imports) of producer services have grown in magnitude and as a share of overall regional exports and imports.

See also: Demand and supply for producer services
Figure 6  Alternative estimates of producer services exports, Washington State.

References


Regional definition and classification

Lewis Dijkstra
Hugo Poelman
European Commission, Directorate General for Regional and Urban Policy

Regions and areas can be defined for scientific, political, or administrative purposes. At present geographical information can be collected either directly for regions or, especially as a result of recent developments in GIS and remote sensing technologies, for grid cells. In the latter case, data can be aggregated to coincide with scientifically, politically, or administratively defined regions and areas of different types and sizes.

In the European Union (EU), the establishment of a regional policy at the start of the 1970s gave rise to the need for standardized regional classifications for the analysis of regional problems, policy design and implementation, and the compilation of regional statistics. At that stage, Eurostat, in collaboration with other European Commission directorates-general and member states, decided to achieve this objective through the establishment of the Nomenclature des Unités Territoriales pour la Statistique (NUTS) or, in English, Nomenclature of Statistical Territorial Units.

The current NUTS system is a three-level hierarchical classification of regions in which each member state is subdivided into a whole number of NUTS 1 regions, each of which is in turn subdivided into a whole number of NUTS 2 regions and each NUTS 2 region is subdivided into a whole number of NUTS 3 regions. The NUTS EC Regulation (1059/2003) lays down the requirements for these regions and how and when they can be changed.

As EU policy increased its interest in areas not derivable from these three levels, and especially in smaller territories (mountainous areas, less favored agricultural areas, coastal zones, and deprived urban areas), smaller NUTS 4 and NUTS 5 regions were also identified.

At present NUTS 4 and 5 areas are classed as Local Administrative Units level 1 and 2 (hereafter LAU1 and LAU2). Each member state has to supply Eurostat annually with lists of LAUs by NUTS 3 region (Dunford 2010, 2011). The NUTS regulation, however, does not cover the LAUs. As a result, they can be more heterogeneous than the NUTS regions which have to fulfill certain statistical requirements.

A problem with administrative or political boundaries is that they do not necessarily coincide with the features of important geographical distributions. Territorial units should ideally reflect the geographical distribution of the phenomena under investigation. In relation to many of the issues dealt with in EU Cohesion Policy, functional areas and, in particular, travel-to-work areas (TTWAs) would make more scientific and policy sense. Also, the harmonized application of rules for defining functional regions would ensure the international comparability of the regions in individual member states. For this and other reasons the EU has embarked on innovative work to establish new regional classifications.
REGIONAL DEFINITION AND CLASSIFICATION

Here, two local and two regional classifications are described based on definitions which rely on the population grid, which can make local and regional definitions and classifications more accurate. The four classifications are interlinked at both the conceptual and the methodological level. The classifications are presented in pairs as the strongest links are between the local and the regional variant of the classifications.

The first pair consists of two classifications that cover the entire territory and create three categories along the urban-rural continuum: the urban-rural regional typology and the (local) degree of urbanization. Both these classifications are based on the population grid. The definitions of the rural area and the rural region are identical with the first applied at the local (municipal or local administrative unit level 2) level and the second to NUTS 3 regions. The definitions of the other two classes are not identical, but both capture the increasing levels of urbanity.

The second pair consists of two classifications that focus exclusively on functional urban economic areas: the EU-OECD city and commuting zone (local) definition and the metro region. As a result, this pair does not cover the entire territory. The EU-OECD city definition captures all the cities with an urban center of 50,000 inhabitants and their commuting zone. A city and its commuting zone together form a functional urban area (FUA). The metro region definition captures only the FUAs of at least 250,000 inhabitants at the NUTS 3 level. In other words, the metro regions only capture a subset of the cities.

The two local classifications are also linked. The EU-OECD city definition is identical to the highest degree of urbanization. The two regional typologies, however, are not directly linked.

First the grid concepts used in the four classification are explained. Afterwards the classifications are explained.

Why use the population grid?

Traditionally, classifications of territories were primarily determined by population density of a municipality (what is referred to as “local administrative units at level 2” by Eurostat) and sometimes population size. For example, the definition of a rural area in the OECD urban-rural typology is a municipality with a population density below 150 inhabitants per km².

This leads to a number of distortions. For example, some municipalities with a very small area and a very small population are identified as urban (or nonrural). For example, Aldea de Trujillo in Spain has a population of only 439 inhabitants, but its surface is so small that it still exceeds the 150 inhabitants per km² and is thus classified as urban.

Some municipalities which include a big city will have a low density if the area covered by the municipality is large enough. So, for example, the municipalities of Badajoz and Cáceres in Spain and Uppsala in Sweden all have a density below 150 inhabitants per km² and are thus classified as rural. All three are, in fact, cities with a population of more than 150,000 inhabitants (see below).

These distortions arise due to the differences in the area of a municipality. The benefit of the population grid is that each grid cell has the same size. (Because the Earth is a sphere and not flat, the grid cells are not exactly the same size, but the differences are minimal compared to the variation in the sizes of municipalities.) See, for example, Figure 1 which shows the population density of a few municipalities in the Netherlands and the density of the population grid. The two circles show two clusters of high-population-density grid cells. In the map with the density of the municipalities, the second cluster is invisible because it is averaged out by the larger size of the area of that municipality.
What grid concepts are used?

The three grid concepts are created using 1 km² grid cells (and clusters thereof). The concepts are based on three criteria: population density, contiguity (see below), and population size.

The three types of grid cells or clusters in the typology are the following.

1 Urban centers or high-density clusters: clusters of contiguous (excluding diagonals) grid cells of 1 km² with a density of at least 1500 inhabitants per km² and a minimum population of 50,000 after gap filling.
2 Urban clusters: clusters of contiguous (including diagonals) grid cells of 1 km² with a density of at least 300 inhabitants per km² and a minimum population of 5000.
3 Rural grid cells: grid cells outside high-density clusters and urban clusters.

Contiguity and gap filling

To determine population size, the grid cells need to be grouped in clusters. The methods presented here use three different rules for contiguity to create clusters:

1 Contiguous including diagonals (used for urban clusters): if the central square (grid cell) in Figure 2 is above the density threshold, it will be grouped with each of the other surrounding eight grid cells that exceed the density threshold.
2 Contiguous excluding diagonals (used for urban centers): if the central square in Figure 2 is above the density threshold, it will be grouped with each of the four cells directly above, below, or next to the central square that also exceed the density threshold. This means that cells numbered

---

**Figure 1** Population density in municipalities and grid cells.
Figure 2  Grid cells and contiguity rules.

2, 4, 5, and 7 can be included in the same cluster. Cells numbered 1, 3, 6, and 8 cannot as they have a diagonal connection.

3 Gap filling using the majority rule (used for urban centers): the goal for the high-density clusters is to identify urban centers without any gaps. Therefore, enclaves need to be filled. If the central square in Figure 2 is not, in its own right, a part of a high-density cluster, it will be added to a high-density cluster if five or more of the eight surrounding cells (therefore including diagonals) belong to a single high-density cluster. This rule is applied iteratively until no more cells can be added.

Figure 3 shows the results of this method applied to a 2006 population grid for Europe. To show sufficient detail a zoom on most of the Benelux and parts of the United Kingdom, France, and Germany is produced. The same zoom level will be used to demonstrate the two local and regional classifications, which at Europe-wide level would be difficult to identify on a small map.

Urban–rural continuum

From these three grid concepts, two linked and compatible classifications have been created. The first is a local definition (degree of urbanization) and the second a regional one (urban–rural regional typology). Both use the population shares by type of cluster to classify municipalities or regions. Each municipality or region belongs to one and only one class. So both classifications cover the entire territory without overlaps or omissions.

Degree of urbanization

Depending on the share of the population living in the different types of cluster, a municipality is classified into one of the three degrees of urbanization using the rules below.

1 Cities (or densely populated areas): at least 50% of the population lives in one or more high-density clusters.

2 Towns and suburbs (or intermediate density areas): at least 50% of the population lives in one or more urban clusters, but the municipality is not defined as a city. This criterion is sometimes expressed, as less than 50% of the population lives in rural grid cells and less than 50% lives in high-density clusters. The results are identical.

3 Rural areas (or thinly populated areas): more than 50% of the population lives in rural grid cells.

In Figure 4, the main cities can easily be identified. The smaller settlements and fringes of the cities are identified as towns and suburbs. The rural areas cover over half of the area (in km²), but less than half of the population.
The degree of urbanization is used in all European statistical household surveys such as the labor force survey (LFS) and statistics on income and living conditions (SILC).

Urban–rural regional typology

This classification is a variant of the OECD urban–rural regional typology, which relies on data at the municipal level to classify regions. This method uses the share of population living in rural grid cells to classify each region (NUTS level 3) into one of the following three groups.

1. Predominantly urban regions/urban regions: the population living in rural grid cells is less than 20% of the total population.

2. Intermediate regions: the population living in rural grid cells is between 20% and 50% of the total population.

3. Predominantly rural regions/rural regions: the population living in rural grid cells is 50% or more of the total population.

In a last step, the size of the cities (based on the definition below) in the region is considered.

- A region classified as predominantly rural by the criteria above becomes intermediate if it contains a city of more than 200 000 inhabitants representing at least 25% of the regional population.

Figure 3  Types of cluster.
REGIONAL DEFINITION AND CLASSIFICATION

Figure 4  Degree of urbanization.

- A region classified as intermediate by the criteria above becomes predominantly urban if it contains a city of more than 500,000 inhabitants representing at least 25% of the regional population.

This classification captures the rural regions well, but the urban regions cannot always be easily linked to a single dominant city (see Figure 5).

The urban–rural regional typology is mainly used to monitor rural development based on figures from regional accounts, which provide indicators such as gross domestic product, gross value added, and employment per sector.

A functional definition of cities and metropolitan regions

Based on the urban centers and commuting flows, two linked classifications have been set up. The first is applied at the municipal level and the second at the regional level. They have been applied to the entire territory of the EU, Iceland, Norway, and Switzerland at the local and regional level.

As opposed to the degree of urbanization and the urban–rural regional typology which rely on population density and size, the two typologies presented in this entry add a functional
Regional Definition and Classification

Figure 5 Urban–rural regional typology.

dimension, the flows of people commuting to work in a city.

Functional urban area: a city and its commuting zone

Definition of a city

The new city definition works in four basic steps and is based on the presence of an “urban center,” a new spatial concept based on high-density population grid cells.

1 All grid cells with a density of 1500 inhabitants per km² or more are selected (Figure 6a).

2 The contiguous (i) high-density cells are then clustered, gaps (ii) are filled, and only the clusters with at least a population of 50,000 inhabitants (6b) are kept as an “urban center.”

3 All the municipalities (local administrative units level 2 (LAU2)) with at least half their population inside the urban center are selected as candidates to become part of the city (Figure 6c).

4 The city is defined ensuring that (Figure 6d) (i) there is a link to the political level, (ii) at least 50% of the population lives in an urban center, and (iii) at least 75% of the population of the urban center lives in a city.
In most cases, as, for example, in Graz (in Austria), the last step is not necessary as the city normally consists of a single municipality that covers the entire urban center and the vast majority of the city’s residents live in that urban center. For 33 urban centers stretching far beyond the city, a “greater city” level was created to improve international comparability.

To ensure that the above definition identified all urban centers, national statistical authorities were consulted and minor adjustments were made where needed and in a way which is consistent with this overall approach. In other words, urban centers with 50,000 inhabitants which this methodology failed to detect were added. In some countries the population grid was a “disaggregation” grid. This means typically that the population of the municipality was distributed across the grid cells using land-use data. This tends to underestimate the population of urban centers. In other cases, steep slopes or fjords mean that the methodology does not detect urban centers.

Identification of a commuting zone

Once all cities have been defined, a commuting zone can be identified based upon commuting patterns using the following steps:

- if 15% of employed persons living in one city work in another city, these cities are treated as a single city (Figure 7a);
- all municipalities with at least 15% of their employed residents working in a city are identified (Figure 7b);
- municipalities surrounded by a single functional area are included and noncontiguous municipalities are dropped (Figure 7c).

The FUA consists of the city and its commuting zone. This method creates mostly cities with a single commuting zone, but in some cases it also leads to multiple cities to be included in a single commuting zone as in the case of the commuting zones in the Ruhrgebiet (Ruhr region), London, and Paris (see Figure 8).

Figure 6 How to define an urban center and a city: the example of Graz, Austria.
Figure 7  How to define a functional urban area; the example of Genoa, Italy.
REGIONAL DEFINITION AND CLASSIFICATION

Figure 8 EU-OECD functional urban area definition.

A typology of metro(politan) regions

Metro regions are NUTS 3 approximations of the functional urban areas (city and commuting zones) of 250,000 or more inhabitants following the definition described above. Each metro region consists of one or more NUTS 3 regions and is named after the principal functional urban area inside its boundaries (see Figure 9).

The typology distinguishes three types of metro regions:

1. capital metro regions;
2. second-tier metro regions;
3. smaller metro regions.

The capital metro region includes the national capital. Second-tier metro regions are the group of largest cities in a country, excluding the capital. For this purpose, a fixed population threshold could not be used. As a result, a natural break served the purpose of distinguishing the second-tier from the smaller metro regions. The regions which do not belong to a metro region are simply called nonmetro regions. This typology can be simplified even further by grouping all individual metro regions together into metro regions.

The boundaries of a functional urban area do not necessarily coincide with those of NUTS level 3 regions. Therefore, NUTS level 3 regions in which at least 50% of the regional population
lives inside a given functional urban area were selected as the components of the metro region related to that functional urban area. In some cases, the NUTS level 3 approximation of the functional urban area is very good. In other cases, the metro region may be larger or smaller than the functional urban area. Each functional urban area is represented by at least one NUTS level 3 region, even if that NUTS level 3 region has less than 50% of its population inside the functional urban area.

Links between the city definition and the degree of urbanization

The first building block of the city definition described above is the urban center and this is identical to the one used in the degree of urbanization typology. As a result, the city (or densely populated area) as defined in the degree of urbanization is identical to the city definition described here. The two maps show the two local typologies for the Benelux area.
The difference arises in the second building block. The city definition identifies contiguous areas which have strong commuting flows. These are the commuting zones (see Figure 3).

The degree of urbanization identifies towns and suburbs (or intermediate density areas) and rural areas (or thinly populated areas) based on population density. As a result, these two categories partially overlap with the commuting zones. The towns and suburbs category will occur both inside commuting zones (in this case they are more likely to be suburbs) and outside (in this case they are towns). Rural areas fall primarily outside commuting zones, but some rural areas have developed a strong commuting relationship with a nearby city and thus can also be found in some commuting zones.

No link between metro regions and the urban–rural regional typology

The two local level typologies have one type in common, namely cities, whereas the two regional typologies each have different types. The typology of metro regions divides NUTS level 3 regions into metro and nonmetro regions while the urban–rural typology divides NUTS level 3 regions into predominantly urban, intermediate and predominantly rural regions.

Despite the absence of an identical type or class of region, the two regional typologies are quite similar:

- most urban regions are metro regions and vice versa;
- most rural regions are nonmetro and vice versa;
- intermediate regions are split between metro and nonmetro regions.

Figure 4 shows the classification of regions in the Benelux area and contains examples of the most common links between these two regional typologies. The differences arise from three main sources: a different logic, different size thresholds, and a different number of classes.

The different logic behind the two typologies can be described as morphological and functional. The urban–rural typology depends more on the population size and density, which in turn is determined by urban form. So this is a variant of the morphological definition. The typology of metro regions relies on the presence of an urban center and of functional economic ties to this center.

The two typologies use quite different size thresholds. Metro regions are related to cities plus commuting zones of at least 250,000 inhabitants. Urban regions represent urban centers of 50,000 or more inhabitants (which also define cities) and/or urban clusters of 5000 or more inhabitants (which also define towns and suburbs).

The urban–rural typology has three types of region, while the typology of metro regions has only two. Due to these differences, some urban regions will become nonmetro regions because the city and its commuting zone are too small (or there is no city). Some rural regions can become part of a metro region if they have strong commuting links to a city in that or a neighboring region.

Taking advantage of a new tool, the population grid, the two regional and the two local typologies were developed to provide a better insight into the development of different types of regions and areas, which is particularly relevant for both Cohesion Policy and rural development. In addition to using these typologies for statistical and analytical purposes, they can also be used for policy. For example, the degree of urbanization is used to report Cohesion Policy expenditure, to identify areas eligible for innovative urban actions and to identify which fund should be used to finance local roads. These typologies have more
potential policy uses. For example, the functional urban areas could be used to facilitate policy design and implementation using more functional geographies. As these typologies become more widely used and understood, it will be interesting to see how their policy use evolves.

SEE ALSO: Economic geography; Human geography; Rural/urban divide; Urban geography

References


Further reading


Policies for regional development are practiced in many countries and rely on spatial differentiation or some form of spatial preference: they may apply to only a limited subset of a country’s regions, or to all of them but in a differentiated way. They aim to stimulate development in order to reduce socioeconomic disparities and inequalities between regions or parts of regions by deploying public finance in a spatially redistributive fashion, often complemented by fiscal, regulatory, or other measures. They have evolved from simple tools for local area regeneration to become geographically based instruments which seek to stimulate socioeconomic growth. Importantly, they have characteristics that make them especially useful in globalized markets and single currency zones, since they constitute a mechanism for supporting areas that are less able to benefit from economic integration and globalization.

A common form of regional development policy applies to a subset of regions within a given country with the objective of reducing politically unacceptable, often persistent, territorial disparities in socioeconomic development. The disparities may be measured, and the areas defined, by such indicators as levels or rates of growth of income and of employment or unemployment, or by demographic changes like depopulation. Typically, policy aimed at a subset of lagging regions allocates public finance in a preferential way to the targeted area in order to subsidize and attract investment, improve the business environment (e.g., through infrastructure provision), and expand the education and training of workers; the amount and intensity of public finance deployed is graduated according to the area’s socioeconomic ranking, with aid being most intense in the worst-off places. This form of regional policy has been extensively operated in Europe, in what is termed its “problem regions.” The regions concerned were genuinely backward (as in the dual-economy case of postwar Italy), or undergoing profound restructuring (such as the industrial north of the United Kingdom, the Ruhr district of Germany, or northwestern or northeastern France), or suffering rural underdevelopment and out migration (such as the northern margins of Europe).

When regional development policies apply to all regions, the overall objective tends to be to harness their underused or spare capacities in order to increase their growth rates. Such policies aim to prompt regions to better understand and mobilize their endogenous potential and to better integrate their actions and resources with the actions and resources of other levels of government deployed in the region and the private sector. The result of this all-region approach tends to be a mosaic of socioeconomic growth strategies and programs adapted by policymakers to different geographical contexts. Such regional development policies are thus instruments for wider, macro-economic growth delivered through multilevel governance systems. In its most comprehensive form, regional development policy can adopt a dual approach, combining an intensive targeting of resources...
in worst-off regions and areas with an approach that is inclusive of all regions.

This entry uses policy practiced by the European Union – which in 2015 spanned 28 member countries, 272 regions, and 507 million people – as a vehicle to illustrate the history of regional development policy as a policy idea and to identify its basic components. Internal regional development policy in the European Union has evolved into an example of the dual approach which concentrates resources on a subset of the most lagging regions, while also seeking to help all regions to invest more effectively and to better organize themselves for growth. The EU experience illustrates some of the difficulties and inhibiting factors in the design and practice of regional policy. Similar policies are practiced in many other places; for example, the antipoverty policy described in China’s Twelfth Five-Year Plan is, in effect, a regional development policy, and policies are also implemented in Australia, Canada, Japan, Korea, Norway, Mexico, and South Africa, to name only a few. The entry concludes by briefly examining a number of issues regularly highlighted by the practical implementation of regional development policies.

While the essential nature of regional development policy is easy to grasp, an inspection of the policies in practice shows that a precise definition may be cumbersome. For example, regional or area development policies may be operated by levels of administration other than national governments; the territorial expression may not be limited to regions but be fixed on groups of regions, or cities, or districts; concerns about capital absorption may weaken a government’s resolve to concentrate its regional development funding on areas of greatest need. Also, the labels attached to regional development policies may vary. In the European Union, for example, “regional policy,” “cohesion policy,” “urban policy,” “rural development policy,” “neighbourhood policy,” and “development and cooperation policy” are all forms of regional development policy. Indeed, much of the inspiration for a government’s “internal” regional policies comes from the development literature and certain kinds of policy for “external” development aid. The term “regional development policies” is, therefore, best understood as referring to a genus containing several species of geographically based development policies.

This difficulty in precisely defining regional development policies is, in part, a reflection that they are first and foremost policies, invented and designed by politicians and policymakers to meet changing political necessities in a variety of administrative circumstances. As policies, in addition to being open to theoretical and empirical examination, they are ideological constructions and are subject to ideological and administrative criticism. Also, regional development studies itself occupies an academic space that combines the disciplines of geography, regional economics, political economy, politics, planning, and sociology, among others, and produces a rich variety of insights and policy recommendations (something evident in publications of the European Union, the International Monetary Fund, the OECD, the World Bank, and the journals of the Regional Studies Association).

Regional development policy and geography

The evolution of regional development policy has accompanied a growing realization and recognition of the importance of the spatial dimension in economic policymaking. A region derives its socioeconomic personality from its
geographical position and climate, its assets and resources (minerals, landscape, population, businesses, infrastructure endowment, technological capacity, culture), and its interaction with the forces of competition and policy. In the twenty-first century, this interaction is largely determined by the region’s capacity to exploit its assets and resources in a globalized market economy. Theories of growth suggest that regions and major urban agglomerations grow fastest when they provide an environment conducive to business investment, attract and concentrate population, and foster wealth-producing economies of scale (often in the form of business clusters). Key elements in the business environment include natural advantage, availability of well-trained workers and pools of creativity or local knowledge, availability of public goods (efficient and reliable basic services, universities, laboratories), and viable supply chains offering a wider variety of suppliers. This concentration enhances growth, thus causing further concentration and, in some cases, the emergence of growth poles or hub regions, frequently leading to a very uneven settlement pattern with wide socioeconomic disparities across the national territory.

Empirical studies point to the disproportionately large contribution that a small proportion of growth poles or hub regions make to national aggregate growth – 8% of regions (27) in OECD countries contributed about one-third of OECD growth over roughly three business cycles between 1995 and 2007. From this, policymakers have been tempted to draw the conclusion that socioeconomic growth and poverty reduction can be achieved by simply stimulating agglomeration and economic concentration.

However, the OECD’s analysis also shows that the remaining regions in aggregate have a substantial potential to contribute to national growth; for example, in the same period in six OECD countries, nonhub regions with an income level below the national average grew more quickly than the national average and contributed a third to national aggregate growth. The key to balanced development and faster growth, therefore, appears to lie in prompting market forces to achieve twin objectives: to continue to provoke the concentration of resources in growth poles and hub regions and, in addition, to foster growth in places beyond these growth poles and hubs, because this will achieve a better overall aggregate performance and contribute to the reduction of social and spatial inequalities.

OECD and other empirical data caution against treating the spatial unevenness of socioeconomic conditions as inevitable, and suggest a warning against policies, especially development policies, which are spatially blind (e.g., sectoral development policies). This is all the more important since the immediate economic benefits of concentration may be offset by longer-term disadvantages to business (congestion, higher transport costs), to working people (spatial mismatch between job vacancies and people searching for employment, housing shortages), and to society as a whole (health costs, longer commuting times, pollution and adverse impacts on global warming), many of which are not captured quantitatively in national and regional accounting.

Indeed, the theoretical suggestion that the externalities of geographical concentration begin at some point to produce overall disadvantages for the economy, and for society as a whole, is borne out by empirical studies which, by their nature, take space and time into account. They demonstrate that factor markets (capital, labor, state of technology) do not work with equal efficiency in all places or at all times and that they may require long adjustment times, be affected by factor rigidity,
REGIONAL DEVELOPMENT POLICIES

or take inadequate account of externalities like pollution. Differences in population density and physical and political geography (e.g., distance from markets; sparse population of mountains, islands, or border regions; legacy of old and declining industry; insecurity due to political instability or criminality) reinforce market imperfections and can slow the rate of socioeconomic development.

In addition, public policies, such as macroeconomic policy, do not have an even impact throughout a national territory: while the economies of some regions may be overheating and require one set of policy measures, other regions may have surplus resources and be suffering depressed economic activity and require different policy measures. The result is that, even where there are relatively limited institutional barriers to internal migration, regional differentials in income, employment, and unemployment may persist for generations.

Academic work that investigates the interaction between economic growth and efforts to reduce inequality through policies of fiscal redistribution indicates that such policies produce benefits. It shows that policymakers are mistaken if they believe that fiscal redistribution slows growth or that inequality will take care of itself in a growing economy, a belief frequently expressed as “trickle down” or in the idea that “a rising tide lifts all boats.” Not only does inequality produce its own undesirable social effects and political consequences, but it may constitute a “drag” on growth, causing it to be slower, because resources are left underutilized, and less sustainable, because it foments discontent, undermines political consensus, and contributes to economic instability to the detriment of investment.

These conclusions from theoretical work and empirical data chime with experience in the political world, where policymakers see compelling reasons for action. The effects of unevenly distributed urbanization and rural depopulation lead to the abandonment of valuable social capital in areas that become less populated and to higher costs of social provision; private capital “trickles up” to big cities rather than investing in poorer areas and thus alleviating some of the public cost; the economies and labor markets of poorer regions are more vulnerable to the economic cycle, presenting challenges to social protection systems; single-currency zones contain areas with disparate economic conditions, yet policymakers are denied economic levers such as exchange rate or interest rate changes with which to seek a remedy.

Such real-world events demand a political response through policy management and intervention. Practical experience shows that fiscal redistribution systems can go a long way to offset inequalities in service provision: for example, in education and health. But they cannot directly address the challenge of improving a region’s (or a nation’s) socioeconomic personality by prompting its factor markets to work better in order to enhance its competitiveness and to increase its stock of modern, productive, physical, and human capital at a regional level.

A regional development policy that is pursued in an integrated way to raise capital investment and labor quality, and to increase innovative and technological capacity, is also necessary.

Regional development policy permits national policymakers to target interventions on places where they consider them to be necessary, and to vary their intensity, thus supporting the goals of macro-economic policy, saving public finance, and avoiding risks of market disturbance. Regional policy may be especially useful in places where successive privatizations have reduced the quantity of public goods and thus depleted the range of possible public investments, adding
further difficulty for governments wishing to raise effective demand.

The potential for authorities to offset disparities through regional development policy is thus supported by academic work. Actual practice has revealed, however, that this is not a simple task. Intervention can be counterproductive through bad timing or by being plain wrong. Further, national, regional, and local administrations may not have sufficient capacity or the know-how to devise and implement development strategies. One of the challenges of regional development policy, then, is to ensure that all regions (better-off as well as lagging) have the administrative capability to design and manage the realization of the investment programs, projects, and actions that will lift their region’s socioeconomic growth performance. Administrations must develop the capacity to frame investment and social measures to stimulate markets to mobilize regional assets, increase employment, reduce unemployment, and delay the onset of increased costs due to overagglomeration. Their programs must aim to make factor markets function so as to diminish socioeconomic disparities and accelerate the growth rate of regions, which, in turn, will imply a gain in national economic efficiency. The administrative effort also needs to make optimal use of its human and financial resources and, once its goals are achieved, to scale back public expenditure.

The evolution of regional development policy

Approaches to regional development policy can be considered as reflecting different evolutionary stages. At its simplest, the policy may aim at the development of a small local area, as opposed to a region, and use public financial support to attract private investment to the area to be assisted and to subsidize public and private investment there (e.g., the siting of Olympic facilities or other flagship projects in run-down city districts). This large project approach focuses on physical capital and seeks to kick-start market activity through a one-off capital injection or through the attraction of a big propulsive industry, for example through foreign direct investment (FDI). It may be more suited to administrative capacity in some lagging regions because it is easier to implement. It is also simpler to organize politically because it may require only a single decision. Enterprise zones, with specific tax and regulatory requirements, are a particular approach to regional development but are suitable only for some regions.

Empirical research indicates, however, that the simplest forms of the large project approach to regional development are insufficient as instruments for socioeconomic development. Although large-scale infrastructure projects have the effect of boosting effective demand in a locality during construction, and of improving the connectivity of goods and labor markets, research shows that regional development also depends on the interplay between physical capital, human capital, and the business environment (linkages between firms, knowledge spillovers, the rate of innovation), and that improvements in worker education and training are important enabling factors. Likewise, history has shown that the benefits of attracting big enterprise investment to locations away from their traditional supply chains are often ephemeral.

When the ambition of regional development policy is enlarged to target whole regions, it aims to promote growth by improving a region’s competitiveness and giving equal weight to physical investments and support for people entering and progressing in employment. In this form, policies tend toward a longer-term and more integrated or programmed approach; seek a balance between environmental and
development concerns; mix public and private finance; and emphasize the importance of performance-oriented quantification and evidence-based policymaking. These multi-year development programs are closer to development engineering: they start from an analysis of the market breakdowns in the regional economy in question and then seek to correct them over the medium to long term through systematic actions embedded into integrated development strategies or plans. They require administrative efficiency, a broader coalition of partners, and longer-term political support.

Regional policies of this more evolved type have a special relationship with other policies – sometimes requiring their support, sometimes providing it. In order to work effectively, they need to be accompanied by a policy, or code, to prevent subsidy wars between territories; a public finance system with a time reach that goes beyond the annual cycle of national budgets and, if possible, lasts the duration of the electoral cycle; and an independent audit system to monitor the policy’s financial management. Since human capital is an important factor in national economic growth and regional development, more evolved regional development policies also need to be accompanied by geographically targeted social inclusion and labor market policies in support of employment creation. At the same time, regional policies can act as a vehicle to help other policies achieve their objectives. For example, since regional policies aim at development, their role is critically important in support of policies to combat global warming, pollution, and the indiscriminate use of resources. A local stimulus to effective demand in a lagging region can be provided as support for business and, in addition, can be conditional on energy efficiency. Income support for marginal farming areas can be channeled as ecological incentives to farmers. In such ways, regional interventions can be engineered to provide double benefits.

Regional development policy’s interrelationship with other policies imposes coordination requirements on government to ensure that the policy mix works effectively: top-down and bottom-up coordination between different levels of government needs to be accompanied by coordination between administrations on the same level as well as with the private sector. Effective coordination enables regional policy to better achieve its goals and to increase the quality, or added value, of public expenditure, by ensuring that each dollar spent may serve multiple objectives and gain double benefits.

The regional development policy that began in the European Union (then the European Economic Community) in 1973 illustrates a policy operating in a subset of regions and based on investment in large-scale public goods. Its main aim was to support the European Union’s social market economy by awarding subsidies principally for publicly owned infrastructure investments in transport, power production and distribution, telecoms, water and wastewater treatment. Its implementation system was rudimentary. It had one financial instrument to support investments (the European Regional Development Fund). Its financial resources were allocated between EU member countries by political negotiation on the basis of an allocation key, which took only partial account of the population of each country’s lagging (or depressed, or worst-off) areas. Projects for subsidy were first selected by member countries in line with their overall regional development plans, with final selection at EU level. The construction of projects was the responsibility of the member country according to its own timetable; subsidies were paid as a proportion of construction costs and were restricted to increases in infrastructure capacity. The policy was, in effect, a “one size
fits all regions” approach, and the effect of its financial redistribution on regional development was correspondingly weak.

EU policy changed little until 1988 but, during this first period, the European Union became aware that the policy’s shortcomings reduced its potential benefit. In particular, the policy favored national as opposed to regional priorities and needs. National governments used EU finance as a substitute for their own spending, thus denying the intended increase in regional investment. Project selection did not take sufficient account of employment creation, and there were long delays between the selection of projects and the start of work, which denied the policy an impact on the ground. Importantly, it had no mechanism for evaluating the economic impact and outcomes of the projects it subsidized. In search of greater efficiency and in parallel with its large project approach, the European Union began to experiment with a more comprehensive implementation system, focusing directly on regions rather than member countries, on collections of investments rather than single projects, and on developing a policy cycle as opposed to a project cycle.

Instead of basing itself on a loose connection between place and physical investment, the policy cycle began to knit together geographic, economic, and social aspects of regional development and to move toward a coherent whole. The emphasis on public goods and public finance remained, as did the emphasis on increased infrastructure capacity, but the starting point was now a regional development program that integrated different investment projects and operated over longer, multiyear time periods. These experimental actions began to subsidize endogenous or “softer” investments (e.g., advice and consultancy support for small and medium-sized businesses, training for entrepreneurs). They allocated EU finance to the region and promoted decision-making at the regional level as to which projects should be selected for subsidy. They encouraged the creation of program management teams within regional and city administrations. The experiments were given a sharp stimulus by the enlargement of the European Union to include Spain and Portugal and the demand that regional development policy allow them to compete more effectively in the European single market, while also helping to counter the competitive challenge that their entry itself posed to agriculture in the Mediterranean regions of France, Italy, and Greece.

At the same time, changes in the wider economy, the practices of government, and the nature of the European Union itself were making reform essential. The EU economy was undergoing a structural, as opposed to a cyclical, crisis, with the sharp decline of coal, steel, shipbuilding, and textiles causing high and persistent rates of unemployment that differed sharply between regions. The disappearance of the Soviet Union affected the defense sector. Governments were preparing to reduce their role as economic producers by outsourcing to private companies many of the services and production processes for which they had been responsible, making it inefficient for regional development policy to operate only in support of the shrinking public sector. The European Union’s program to reduce internal trade barriers was expected to create competitive challenges for many regional economies and to accelerate changes in economic structures – and the same would later be true for the creation of the single-currency eurozone.

By 1988 EU regional policy was ripe for reform and the previous experiments offered a template. It remained unchanged as a development policy that targeted a subset of regions but was relaunched as a decentralized policy with new objectives, a new implementation system, and a radical change in the categories of projects.
REGIONAL DEVELOPMENT POLICIES

to be subsidized. Thus, it integrated another financial instrument (the European Social Fund) to subsidize the development of a region’s human capital, and insisted on the integration of physical and human capital investments. It opened the possibility of publicly financed subsidies for privately owned projects. It devolved implementation to regional partnerships that brought together representatives of national, regional, and local government, business, and organizations from civil society in a framework that became known as multilevel governance. One of its biggest innovations was the policy cycle which firmly anchored policy to the region to be assisted, so that economic geography became the paramount consideration.

The new approach demanded new capabilities and changed attitudes from policy administrators, and created tensions between layers of government. A good example of the need for policy clarification was the notion of regional partnership, which differed according to whether a country had a federal structure with regional self-governance, whether central government administration was carried out through regional offices of the state, or whether the only regional administration was a grouping of local authorities. In addition, there were different approaches to the inclusion of nongovernment bodies in the partnerships, with some governments including employers but excluding trade unions on ideological grounds. It required several years and further policy adjustments before the reforms of 1988 were fully translated into practice.

EU regional policy has now evolved further to combine its targeting of poorer regions with a socioeconomic growth policy for all regions, thus adopting the dual approach. The policy cycle has developed, seeking, among other things, to ensure that the policy, while fully decentralized in operation, is still coherent with EU and member country development objectives. To achieve this coherence, regions implement the policy within a three-cornered framework: they are allocated a sum of EU finance, with the amounts varying according to the regions’ socioeconomic ranking; they plan and implement their programs within an EU set of development guidelines (thematic objectives) which aim to ensure a measure of concentration throughout the European Union on key investments; and they comply with a national indication of which of these key EU investments has the greatest relevance for the overall development of their member country. Within this framework, the regional partnership analyzes the region’s development potential and the market breakdowns that are impeding its progress, agrees on priorities for socioeconomic action, and designs the region’s medium-term development strategy and development programs, incorporating quantified objectives and milestones. It implements the strategy through the medium term (seven years), selecting projects and actions for subsidy. Evaluation is built in to the policy cycle, creating a loop back to objectives, with care being taken to assign effects to particular causes, and making possible results-oriented management. This cycle (which incorporates the principle of subsidiarity) is an integral part of the European Union’s geographically based instrument for socioeconomic growth. In addition, the policy is increasing efforts to ensure that the necessary administrative and executive capacities exist in the regions (and, if needed, at national level), in order to cope with the requirements of implementing the project cycle.

The components of a geographically based instrument for growth

Whether regional development policies operate in a subset of regions in a given country or as a
geographically based instrument for growth in all of them, whether they base themselves around large, publicly subsidized flagship projects or multiyear development programs, they can be assessed according to four broad components: the policy’s objectives, which are intimately connected to the method of selecting regions and areas for assistance; its system of implementation; its dedicated budget; and its content (the types of actions it subsidizes). A fifth consideration is the regional development policy’s dependence on accompanying policies.

The objective of regional development policy is, briefly stated, to improve the wellbeing of citizens by stimulating regional development by, in part, redistributing public finance between regions. A first requirement in setting the objective is to decide whether it is to be achieved by concentrating on reducing regional disparities (which implies that the policy is for a subset of badly-off regions), or by improving overall socioeconomic growth (which implies that the policy covers all regions), or a combination of both (which implies varying the intensity of aid to favor the more problematic areas, however defined, while supporting all regions). The objective may be expressed in a number of ways and, over time, it has tended to become more elaborate. “Inclusive” or “pro-poor” growth objectives tend to embrace the idea of seeking to raise income (GDP) while achieving a more even income distribution among the population, perhaps in addition to seeking to achieve other social goals. “Smart” growth refers to improvements in the technological and innovation content of the economy and levels of productivity or rate of productivity gain.

Global objectives take on a rather different nature the more they are transformed into real action on the ground through the implementation system. There is a hierarchy and sequencing of objectives: for example, where the overall objective may be to raise employment, the strategy may seek to deliver this through selected subobjectives, or “priorities,” such as a higher rate of employment for women. This, in turn, may make it necessary to set an objective for the creation of increased capacity for child care, which may require the objective of establishing more training places for child carers.

Where policy involves performance-based management, each objective in the hierarchy needs to be expressed in quantitative terms and this relies on the availability of regional datasets. Further precision is gained by specifying a time period in which the objective will be achieved. This allows management to be monitored against milestones during the policy’s implementation period. The quantification of objectives permits the evaluation of impacts, outcomes, and comparisons between regions and countries, making it possible to improve the policy’s efficiency and enabling improvements in the added value of its spending. Subregional actions are an important feature of regional policies since administrative regions may not be functional regions in the economic sense, or may contain pockets of deep-seated socioeconomic problems that also make them suitable targets for regional policy. However, the availability of subregional socioeconomic statistical data tends to be limited. For regional policies implemented at supranational level, achieving consistency in targeting subregions across countries can be a particular challenge not just because of the limited availability of data but also because of lack of its comparability.

There is, therefore, a close link between the clarity of the policy’s objectives, the availability of quantified baseline data, policy evaluation, and overall efficiency, with the availability of regional and subregional data acting as a governing condition on the policy’s speed of evolution. But, even when data of sufficient quality are available
to policy managers, experience shows that the quantification of objectives, the selection of suitable indicators, and the setting of targets are difficult and uncertain tasks. It is not uncommon, for example, for high targets to be established as political gestures, only to creep toward more modest levels as evaluation deadlines approach.

EU regional policy can be said to have objectives for the long and medium term, one concentrating on a subset of regions and one for all regions. Its long-term objective, set out in the EU Treaty, is “to promote economic, social and territorial cohesion,” where “territorial cohesion” can be taken to mean reducing the backwardness of least favored regions, including rural areas, and the disparities between the levels of development of all regions. The medium-term objective is today expressed as helping the European Union achieve its 2020 goal of “smart, sustainable and inclusive growth.” These objectives are broken down into subsidiary objectives, relying on quantification in the operational phase. Thus, EU regional policy concentrates its redistributive effort and budget on the subset of poorest regions, defined in statistical terms as those having a gross domestic product (GDP) per head below 75% of the EU average. For these regions, it can be said that the implicit objective is one of increasing their GDP per head to 75% of the EU average, if possible within a seven-year period. In principle, it seeks to achieve this by varying the intensity of its financial support per head of population in line with a region’s distance below the 75% threshold; the further below the threshold, the greater the intensity (additionally, the intensity is modulated with the level of national GDP and the level of unemployment). But, for some of the very poorest regions, the results imply an annual transfer of investment funds that is considered as a matter of political judgment to be greater than their national economies can absorb.

The size of the seven-year transfer is therefore capped in terms of a percentage of the country’s GDP, and the redistributive effort in such cases is reduced. The objective of achieving 75% of EU GDP, which implies that EU regional policy can help the incomes of the poorest regions to grow consistently faster than the EU average, is not often achieved in one seven-year period, although the trend has been one of convergence (a trend reversed, albeit slightly, by the post-2008 financial and economic crises). In practice the poorest regions have tended to experience more than one (seven-year) planning period in which they receive maximum support, although its intensity diminishes as they approach the 75% income threshold. Today, the European Union concentrates over 70% of its resources on the poorest regions, which contain some 25% of population, leaving 30% for two other groups of EU regions: those with a GDP per head between 75% and 90% of the EU average, and the remainder above 90%.

A system of implementation, sometimes called the delivery system, is necessary for regional development policy and incorporates the policy cycle. This is true for policies aimed at a subset of regions in a given country as well as for those operating in all regions or for those based on individual projects or on multiyear programs. The implementation system itself has its own objectives for helping to maximize the policy’s effectiveness. Thus it may integrate wider considerations (environmental protection, action against climate change) into decision-making; increase the policy’s financial impact by concentrating resources on some types of activity; and ensure a timely impact on the ground by incorporating a discipline for project implementation. In the more evolved and more comprehensive examples of regional development policy, the region’s contribution to the policy cycle consists of an overall strategy for a region’s development,
which analyses the region’s potential and weaknesses; the creation of development programs, or action plans, which establish detailed objectives with budgetary envelopes; and the realization of physical or human capital investments which further the strategy’s objectives. Evaluation is integrated as a feedback loop, first assessing strategy, then monitoring implementation, and finally assessing overall outcomes and impacts.

A distinctive feature of the more evolved forms of regional development policy tends to be their decentralized implementation and decision-making. This brings together the various layers of government in a form of multilevel governance, with each layer taking the decisions most appropriate to its knowledge and responsibilities (the principle of subsidiarity) and integrating policy actions at different territorial levels. Multilevel governance is enriched by the establishment in each region or target area of a partnership that contains nongovernmental socioeconomic actors (civil society) as well as representatives of the national, regional, and local layers of government. The partnerships organize and carry out the region’s contribution to the policy cycle. It is the partnerships that analyze socioeconomic conditions in their regions, develop the programs and action strategies that determine priorities, set detailed and quantified objectives to make comparative assessments of projects and actions which are applying for public subsidy, select those which promise the best added value, and monitor and evaluate the results of their implementation.

Decentralization and partnership strengthen regional development policy’s role in giving spatial awareness to macro-economic policymaking. They bring into the development operation socioeconomic actors with a close understanding of local conditions and geography; interlock and reinforce the coherence of their objectives; enable integration of development strategies; and, importantly, help to build and maintain a strong consensus around the development objectives. Experience has shown, however, that they also bring complications that may challenge the administrative know-how and capacity, as well as the mutual trust, of the partnerships and layers of government. These challenges have to be overcome by, for example, ensuring the compatibility of partners in the face of difficulties; working together; deploying the partnership to best effect; maintaining coherence; and avoiding decisions that are in effect merely the lowest-cost denominator of different possibilities. It is therefore important that the objectives of the region’s development strategy are widely agreed and that decision-making procedures are transparent. Decentralized decisions must remain consistent with centrally determined development objectives, and partnerships must be rigorous in selecting the most suitable investments and actions for public support.

EU regional policy is continually encountering and seeking to solve difficulties in its detailed implementation. It has introduced three important modifications to reinforce its overall coherence. Two of these are intended to strengthen its performance-based management and a third its balance between physical and human capital investments. It has established guidelines (thematic objectives) to aid the prioritization of physical and human capital investments. At first, guidelines were offered on a voluntary basis, but they are now more binding since experience has demonstrated the difficulty of ensuring that they are respected. The present attitude is that aid is conditional in relation to agreed objectives, and it is backed up by a compulsory system of performance assessment, measuring progress against quantified performance milestones and targets established at the strategy stage of the policy cycle. The third modification concerned the particular EU difficulty
that member countries were allocated insufficient EU funding for investments in human resource development, despite evidence that a region’s human capital is critical for its rate of development. This has been addressed by placing a greater obligation on member countries to maintain their EU funding for human capital above a minimum level. Faced with an implementation system that grows in complexity, and recognizing the importance of administrative capacity for effective policy delivery, the European Union allocates part of its regional policy funding for the modernization and reform of civil administrations and for the further education and training of regional policy administrators.

The dedicated budget of regional development policy is the medium of its spatial and social redistributive function, and is the root of the coordination and consensus that are necessary for its implementation. The availability of public finance incentivizes socioeconomic actors in targeted regions and subregions to support the policy’s actions, and encourages them to coordinate actions at different levels of administration and between the public and private sectors and civil society. Small budgets require implementation to be more concentrated, both in space and in practical content, in order to be effective. The budget is deployed through the policy’s financial instruments, and the choice of instruments reflects the degree of integration and targeting that policymakers seek between physical and human capital investments and between the different geographical spaces to be covered by the policy (regions, cities, rural areas, etc.). Where there are several instruments, policymakers must find ways to delineate their responsibilities and to allocate available public finance between them.

EU regional policy has five principal financial instruments: the European Regional Development Fund (ERDF), which concentrates on all types of investment project in all types of places; the European Cohesion Fund (CF), which operates as an additional investment fund in the poorest member countries (as opposed to regions) for transport and environmental infrastructure; the European Social Fund (ESF), which concentrates on investments in human capital, social inclusion, and the functioning of labor market institutions, again throughout the territory of the European Union; the European Agricultural Fund for Rural Development (EAFRD) for the socioeconomic diversification of rural areas; and the European Fund for Fisheries and Maritime Affairs (EMFF), which is mostly deployed in coastal areas. The five funds represent the European Union’s determination to concentrate its redistribution of public finance in low-income areas and to integrate subsidy actions in favor of physical and human capital. But, at the same time, they illustrate the difficulty in mobilizing and integrating policies with different spatial references behind a single unifying vision of regional needs (the funds overlap in scope). The allocation of public finance between the instruments has no single set of criteria and is achieved through a mixture of objective, regulatory, and political factors, with historical allocations having considerable weight.

The budget of regional development policy is usually composed of public and private finance, with public finance being the instrument through which government pursues its redistributive and regional development objectives. Public funding levers private capital and thus increases the financial mass being used for regional development. In its simplest form, public–private leverage is achieved through either using public finance as a loan fund (although loan instruments are less effective in times of economic depression or low growth) or by subsidizing only a proportion of a project’s cost, with private capital (from
business, philanthropic, charitable, or voluntary sources) financing the remainder. This presents governments with the possibility of a spatial incentive: fixing a higher proportion of subsidy for projects or actions that it considers important but that are sited in less attractive areas or that appear to carry greater risk and are, therefore, less likely to be carried out by the private sector within a given time. Governments can go further to leverage private capital by focusing public finance on projects or actions that they think will trigger a wider private sector response. Examples include the creation of business sites to attract inward business investment; the construction of access roads to open up land for private industrial development; and the provision of incubator places to encourage the creation of new businesses. The potential for public–private leverage is greatly magnified in Europe by the process of privatization, the transfer of public goods and services into private ownership.

An important requirement – known in the European Union as the additionality principle – is that national, regional, and local governments do not use private capital attracted to the regional development effort as a means to displace public finance that may have been intended for the target area, since this counteracts the increase in the financial resource available for the target area’s socioeconomic development.

Preferences for the practical content of regional development policy can be captured in the question “What sort of projects and actions should regional development policy encourage in order to enhance the socioeconomic personality of a given region?” The answer is linked to both the policy’s objectives and analysis of how the region’s assets and resources can be made more productive.

For a regional policy restricted to the subsidization of public goods, financial support will be confined to projects undertaken by the public sector. For a policy aiming to attract inward investment (FDI), it will take the form of a range of actions to make the region attractive for target businesses. For a policy that aims to improve the competitiveness and the rate of socioeconomic development of a region by better use of its endogenous potential, the policy will be ready to subsidize a wide range of physical or human investments, whether financed by public or private capital. The overall effect of a region’s projects and actions may be to alter some aspects of its socioeconomic personality.

Discussions on the policy’s practical content have provided important insights into economic development and the timing and range of projects or actions that regions may consider for subsidy. But they rarely provide compelling reasons for one type of investment over another, since factors that are peculiar to the region usually predominate in any regionally designed strategy. The general view is that sustainable regional development is the function of a simultaneous combination of investments that will improve the socioeconomic benefits derived from the region’s resources. It will improve human capital (education, training, social inclusion, equality, reform of labor market institutions, and social provision); the endowment of basic infrastructure (transport, power, environmental, communication); technological capacity (linkages between research and industry and between businesses, promotion of innovation); the socioeconomic environment in which business operates (preparation of business parks and incubators, provision of facilities to support start-up and growing businesses, loan and other schemes for capital availability, and support for inward investment); and organization (multilevel governance). One recurrent idea to assist the economic development of local areas, towns, or cities within regions is the adoption of local exchange trading systems (LETS), or
REGIONAL DEVELOPMENT POLICIES

local currencies, in attempts to build on the experience of the Wörgl, Austria, stamped currency of 1932–1933.

Given the importance of human capital in regional development, even policies with a narrowly defined capital project objective will nonetheless contain measures to promote education, vocational training, and social cohesion, offering subsidies to projects or actions that aim to improve the ability of people to enter the labor market (e.g., actions to promote equality in the labor force; prevocational training; language education; employment mentoring).

The nature of regional policy’s practical content has been the source of theoretical debate summed up by the idea of efficiency versus equality. The efficiency argument points to the economies of scale, or agglomeration, that companies derive from being in cities, and argues that investments in favor of concentration represent a superior value for public support. The equity argument urges the reduction of regional disparities, placing importance on investments in human capital and on pro-poor aspects of growth, as well as physical investments in basic and business infrastructure. The debate echoes the theoretical opposition between the benefits of concentration and its negative externalities, and sets the two notions in opposition to one another. The modern approach that underpins the more evolved forms of regional development policy promotes their complementarity, seeking a balance, determined region by region, between investments in human and physical capital.

EU regional policy subsidizes investments in physical and human capital in both the private and public sectors, including investments in productive capacity, with subsidy rates varying in accordance with project types and a project’s revenue generation. One of the difficulties public administrations encounter with the large project approach to regional development or an approach that seeks to favor a particular sector is to detect, and avoid, those that will not prove viable on a socioeconomic basis (“white elephants”). This difficulty is avoided when policy concentrates its attention on improving the overall climate for investment and leaves business to make its own investment decisions.

Regional development policy does not stand alone, and even a simple project-based policy depends for its success on a number of accompanying policies. When considered as a geographically based instrument for economic growth, regional development policy stands within the framework of macro-economic policies and their pursuit of the objectives of growth, trade balance, and exchange rate and price stability. As a financial instrument that targets spending on precisely defined areas and social groups, it offers the possibility to mitigate some of the unwanted regional effects (e.g., increased disparities) of certain phases of macro-economic policy. Indeed, for political leaders regional policies can be perceived as necessary in order to achieve national consensus in favor of other government actions, for example austerity policies. In the European Union, the macro-economic impact of regional policy is an important consideration for countries that receive large annual investment resource transfers (close to €10 billion a year in the case of Poland for the seven years 2014–2020, which is equal to a large part of the country’s investment budget).

All types of regional development policy require a regulatory framework for the avoidance of subsidy wars and to ensure fairness in the policy’s implementation. In regional development policies that limit themselves to the public sector, the objective must be to avoid subsidy wars between regions or countries. In policies that cover both the public and the private sectors and which grant public finance to private investments, competition requirements between
businesses become important to ensure that public subsidy does not distort market conditions. EU regional policy is implemented within a detailed regulatory framework that monitors subsidy rates to both public and private investors, and that has the power to force the repayment of subsidies that are proved to distort market conditions.

Multi-annuality of the national budget is necessary to underpin the multiyear regional development programs that are an important component of more evolved regional policies. It is the multiyear nature of programming that assures regional policy of its medium-term relevance; distinguishes it as an instrument which helps to manage structural change, as well as contributing to the management of economic cycles; provides a framework in which the partnerships of multilevel governance can plan ahead and create project pipelines; and insulates it against too frequent political and budget changes.

All government policies require efficient and sound financial management and this requires independent audit. For a decentralized regional policy where the management of public finance is devolved to a large number of socioeconomic actors, the importance of independent audit is critical. Without it, management systems atrophy; errors begin to occur; and the possibility of fraud and corruption increases. But a proportionate balance between the requirements of financial management and the amount of public finance involved in projects or programs may be difficult to achieve. Political factors may make this even more difficult.

Issues in regional development policy

Regional policy is a volatile field. It is subject to intense debate about its efficiency, effectiveness, and utility. Within individual countries, its importance within the policy framework waxes and wanes with changes in government ideology. Nevertheless, almost all governments recognize the need for policy instruments to address spatial disparities, either because of a commitment to territorial equity in living standards and employment or because of the concern that all regions should strive to contribute to national economic growth. A number of policy and technical issues arise regularly in the design and implementation of regional development policies, whether this is undertaken by central government alone or through a multilevel governance system involving regional and local actors. Seven are mentioned here.

Regional development policies rely on effective organization, and the quality of governance is an essential factor in their implementation. Their decentralized operation emphasizes governance quality because it brings extra layers of administration into the heart of policy implementation, granting more autonomy and wider responsibility for public expenditure. The belief that the quality of a region’s governance increases with improvements in the region’s socioeconomic development – that greater wealth makes possible better governance – is not born out by experience. Instead, studies suggest that low-quality governance is a barrier to growth and that good governance is a prerequisite since it alone can provide and guarantee the stable social and business environment that encourages development.

Among the factors that affect the quality of regional governance are the legal system, which must entrench the rule of law, equality before the law, and nondiscrimination; government administrations and their personnel, who must be free of corruption; and government systems, which must be capable of technological adaptation (e.g., systems of e-government) and open to organizational innovation. The ability and
capacity of administrators and their procedures is a determinant in the take-up of financial resources and ideas flowing from other policies, civil society, and the private sector; it is critical to the coherence, quality, and delivery of development strategies.

These considerations indicate a tension within regional development policies between the desire to ensure the best results, which frequently leads to elaborate and complicated procedures, and the need for bureaucratic simplicity. There is no lasting tradeoff between the two, and a condition for the success of policy is a constant effort to help administrators improve their know-how and capacity and to encourage bureaucratic innovation.

The essential task of regional development policy is to intervene in the working of regional economies, the need for action being strongest in the poorest regions or areas. This sometimes brings the policy into conflict with free-market ideologies that argue that regional economies function best when intervention is kept to a minimum, or even when there is no intervention. Such arguments rarely take account of the fact that there is a geographical impact and market effect of almost all government spending (e.g., road building; siting of military bases; awarding of research contracts). The proponents of spatially blind policies may ignore that agglomerations are often treated preferentially in the allocation of public spending (e.g., financial support for local transit systems). The counter-argument is that regional development policy is market-oriented and seeks to overcome market failures and to stimulate markets to work more effectively, reducing the scope and scale of intervention as market mechanisms prove capable of improving the socioeconomic performance of targeted regions.

Concerns about market intervention may spill over into arguments about aspects of regional policy’s implementation or delivery system. Calls for “flexibility” in implementation, for example, may disguise attempts to escape from a centrally determined policy focus. Such discussions risk treating the implementation system as bureaucracy or red tape (or, in environmental terms, green tape) and overlooking its role as an essential component of policy and as a determinant of its success.

Beyond such red tape arguments are important issues of balance. For example, experience shows the difficulty of achieving a stable balance between the empowerment of regions and the desire to ensure that they implement policy within an agreed framework; between the stringency of controls against bad financial management and the possibility for regions to introduce bureaucratic innovations in implementation; between objective financial allocations and a region’s capacity to design and execute sufficient projects to absorb funding within preset deadlines; between evaluations that serve a political purpose and those that aim to improve program performance; or, more widely, between economic development and the responsibilities of environmental and climate protection. Balance is difficult to achieve and maintain, and issues frequently resurface. The only remedy appears to be to treat administrative and executive capacity-building as an integral element of regional development policy, putting it on an equal footing with, for example, infrastructure provision or the retraining of workers.

A close inspection of regional development policy as a macro-economic instrument shows that it is not easy to place it in a classification that draws distinctions between structural and countercyclical or between supply-side and demand-side instruments. Its spatial and strategic foundation and its medium-term nature make it an apt instrument for guiding and managing structural change. The immediate availability of
its investment programs, containing a pipeline of mature projects and social and training actions ready to go, gives it the capacity to boost effective demand and to provide an effective local stimulus, especially in poorer regions where the policy's redistributive effect is greatest. Its concern with targeting market breakdowns in particular geographical areas to overcome impediments to the supply of factors of production suggests that regional policy is inherently a supply-side instrument. But its task of injecting resources into regional economies gives it the attributes of a demand management instrument, whose effectiveness is enhanced by its geographical targeting and its regional partnerships' knowledge of local conditions and socio-economic requirements. In fact, its distinguishing feature as a macro-economic instrument is what might be termed its "hands-on" nature: where other macro-economic policies are spatially blind (in that they operate at a distance from socio-economic realities), regional development policy is geographical, space-based, and anchored in actual conditions. If other macro-economic policies shape the overall economic and social environment in which economic activities may take place in a country (or group of countries), regional development policies add spatial and financial precision that improves overall political efficiency.

Whatever the geographical scope of the policy (the subset of regions with the weakest economies or all regions), there remains the question of what constitutes a region. An effective region from the point of view of regional policy tends to be one with a critical mass of economic opportunities, which provides a suitable context for integrated, cross-sectoral economic development within its boundaries and combines a certain level of decentralized responsibility in the field of investment policy. In the European Union, the reference region most commonly used has an average population of 1.7 million people which is sufficiently large in most cases to take account of commuting lines and travel-to-work areas. EU regions differ widely in the extent of their policy roles and their level of responsibility for investment – some are also (small) member countries and so have at their disposal the levers of macro-economic policy.

Cities typically meet the requirements for decentralized policy implementation. They contribute substantially to national development but are generally unable to ensure the development of the region as a whole, while their own needs may not be adequately reflected in a development policy targeted at broader territories. Wrapped up in this issue is the difference between administrative and functional regions. The boundaries of administrative regions may not coincide with those of functional regions (e.g., both national and regional boundaries may cut employment centers from their travel-to-work areas) or they may be drawn so as to include dissimilar geographical areas (e.g., island groups, a coastal region including islands). This may also affect the usefulness of regional datasets. In implementation, there are devices for overcoming this, such as allowing support to be granted to projects outside the administrative boundary of the supported region, provided the benefits accrue entirely or substantially to it (e.g., support for the training of the target region's workers in a training institution located outside the target area).

The decentralization of regional policy is important in resolving economic difficulties caused by administrative boundaries. Decentralization allows regions to tailor investments to take account of their own spatial context and of the relationship between their cities and towns and their hinterlands, as well as their relationship with neighboring regions. But the issue of the ideal geographical framework for regional
development policy, both spatial and political, remains. The only available solutions tend to be pragmatic ones, working with what exists while making adaptations to the implementation system in an effort to ensure that policy objectives are attained.

The data available to regional policy are traditionally some regional measure of income, unemployment, and employment, with income being measured through GDP, which, because it measures economic activity in the market, treats investment and consumption in the same way. With these data the policy can seek to quantify its objectives and measure its progress in both economic and social terms. Policymakers accept that there are limits to convergence – all regions may benefit from regional development policy, but not all regions can reach the same levels of development – and this is leading to more nuanced expressions of the objectives of regional development policies, which express concern for the quality of the growth and are reflected in terms like “inclusive” or “pro-poor” growth. This, in turn, calls for a wider demand for regional datasets (gender, labor market, education and training, health) and the creation of new ones that can capture income distribution and quantify “quality” objectives and progress – which are necessary when regional development policies are concerned with such things as impacts on the environment, biodiversity, and resource depletion and incorporate performance assessment. Academic innovations are responding to this development with the construction of “happiness” or “capabilities” indices to measure wellbeing, regional indices of the disposable incomes of households, and “risk of poverty” indices to capture not only progress toward social inclusion but also its durability. However, data are expensive and the wish list of regional data tends to exceed what the authorities can afford, hence the persistent reference over many decades to the same basic (but still meaningful) sets of income and employment data.

The evaluation of regional development programs and actions, and, ultimately, policy, requires data that are capable of expressing its intended and actual results. Thus the practice of evaluation is intimately connected to the availability of regional datasets and creates new challenges: for example, lags in data compilation reduce their usefulness in steering regional development programs; region-wide data may not be able to capture the area-specific results of the projects and actions carried out as part of regional policy. But the difficulty of matching their requirements with the availability of regional data is only one of the problems that have to be overcome in the evaluation of regional development programs. The list would include such factors as the need to ensure that evaluation and baseline indicators are built into program design; that intended objectives are set with sufficient precision to permit the choice of suitable indicators and targets; that quantifiable proxies exist wherever regional data cannot be the appropriate measure; that data is collected regularly; that other factors in the economy that influence the policy’s results (especially when the desired objectives can only be appreciated over the longer term) can be properly taken into account; and that results can be properly interpreted. Overshadowing these questions, however, is the question of whether evaluation is a political or a technical function. The essential task of evaluation goes beyond monitoring and observing to judge the performance of policy. This indicates that questions relating to evaluation reflect deliberate policy concerns – the deliberate policy choices that are made and the extent to which policy objectives are achieved.

Evidence-based evaluations have provided the information to demonstrate the transferability of development know-how from one
REGIONAL DEVELOPMENT POLICIES

region or subregion to another, avoiding the need to reinvent the ideas behind successful projects and initiatives and speeding up policy implementation and impact on the ground. Thus, it is widely accepted that regions can quickly gain from others the knowledge and ability to improve their administration and the development of projects. This is all the more important given the increasing complexity of successful regional development programs as they move beyond basic investments such as transport infrastructure to other types of projects, such as technology parks to spin out the results of university research; business networks to accelerate the rate of innovation in a region’s businesses; social inclusion activities to help long-term unemployed people regain a footing in the labor market; or localized business loan schemes to offset the reluctance of banks to lend at acceptable rates to start-up businesses in risky areas. Encouraging the transfer of proven ideas is an integral part of the implementation of regional development policy, shortening the time lag between political decisions and their impact on the ground.

Conclusion

Regional development policy provides a simple and immediate way to tackle pockets of regional or urban poverty. In its more evolved form, when applied to all regions in a given country, it combines geography, politics, and economics to become an instrument of macro-economic management that is oriented toward the wider spread of socioeconomic development and contributes to its acceleration. It offers a solution to several policy conundrums. It makes it possible for governments to intervene in their national economies in order to respond to structural economic problems without compromising the operation of markets; its spatial targeting makes it possible to raise effective demand in worst-off regions without disproportionate increases in national debt; its investment focus helps to resolve the efficiency versus equity dilemma; its technical nature allows policymakers to sidestep ideological issues and contributes to a consensus in policymaking. Its implementation may be complex, owing to its multiple objectives, but it creates the possibility for evidence-based, spatially targeted economic management in the context of globalization and within single-currency areas, whether nation-states or multistate zones, and can be especially useful in cases where governments wish to boost effective demand but where the range of potential public investments has been reduced by privatization.

SEE ALSO: Cities and development; Cross-border, transnational, and interregional cooperation; Economic development zones; Human capital; Local development; Multilevel governance; Political economy and regional development; Regional definition and classification; Regional development models; Regional inequalities; Urban renewal

Further reading


REGIONAL DEVELOPMENT POLICIES


The word region derives from the Latin word region, which referred to a realm or kingdom. In contemporary geography the word is used to refer to a delimited part of the surface of the Earth. In geography, three types of region are identified: areas that share certain common characteristics (homogenous regions); areas that are functionally interdependent (functional regions); and areas that are subject to economic, political or administrative dominion or authority (jurisdictional or administrative regions). Although the word is often considered to refer to a part of a nation, it can also denote national and supranational territorial entities. The word model refers either to a stylized representation of a particular historical or geographical pattern of development or to a simplified (abstract) conceptual representation of the characteristics and determinants of real world situations. In the former case, the aim is to identify the variety of forms that development can assume. In the latter case, the aim is to replicate or explain real world situations and to identify underlying causal mechanisms. The word development usually refers to the act of mobilizing and creating resources to make an area and its inhabitants more productive of things that are useful (use values), that in market economies they can exchange for other useful things (exchange values), and that can improve their capabilities and the lives of people who visit an area (human development). A distinction is sometimes made between development of an area and development in it (as a result of external intervention). The idea of development is also associated with the idea that places and their inhabitants can pass from lower to higher stages of organization. Development is, therefore, a process in which people are at one and the same time the actors and the beneficiaries.

Development is usually measured by calculating how much additional wealth (gross domestic product (GDP) or gross value added (GVA)) is created in a particular area in a particular year or how much income accrues to an area’s inhabitants (gross national income (GNI)). Models of regional development are concerned either with explaining such increases in output and income or with explaining the actions and improvements that bring them about. From a human development perspective, wealth creation is, however, a means to an end rather than an end in itself (Seers 1969; Sen 1999). Sen, for example, suggests that development involves expanding the real freedoms that people enjoy: development should increase human capabilities, defined as the freedom to achieve valuable beings and doings (themselves called functionings). These functionings include good health, adequate nourishment, access to educational services, happiness, self-respect, and security. An important implication is that models of regional development should address questions relating to the social and territorial distribution of wealth and income as well as of growth. Moreover, methods of measuring development should consider changes in the stock of wealth rather

---

The International Encyclopedia of Geography. 
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston. 
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd. 
DOI: 10.1002/9781118786352.wbieg0650
than flows of income and should employ wider definitions of development that embrace not just productive capital but also human, social, institutional, and natural capital.

In a proximate sense, an area’s wealth, productivity, and income depend on the interaction of four sets of factors (Figure 1). The first is its natural (first nature) and created (second nature) resource endowments (natural resources, population, skills, infrastructures, technologies, etc.). An area’s resource endowment is a cause of development, specialization, and the division of labor. The second is (i) the degree of mobilization of an area’s resources and human potential and (ii) the effectiveness, efficiency, and productivity with which they are used. In turn, these drivers depend on the capacities, strategies, and performance of economic actors, the social relations and institutional structures in which action occurs, and the wider national and international context and rules of the game. The third is the extent to which an area’s inhabitants command and control the resources of others, the degree of unequal exchange, and the extent of transfers of wealth. The fourth derives from the fact that resources are created, conserved, and reproduced in and through development and as a result of policy interventions, so that resource endowments are not just a cause but also a consequence of development. Development processes

![Figure 1](image-url)
are, therefore, cumulative, endogenous, and path-dependent, and can involve self-reinforcing virtual and vicious circles.

Models of regional development also deal with the evolution of regional disparities and the mechanisms that shape them. Existing models can be distinguished according to whether regional differences in GDP, GVA or GNI per capita are expected to get smaller (convergence or catch-up) or wider (divergence or falling-behind). In the recent period of hegemony of neoliberal economic ideas and programs, models that predict catch-up have predominated. The expectation was that convergence would result from the alignment of technologies, lifestyles, consumption patterns, and government policies, in part as a consequence of the discipline exercised by global financial markets and dominant economic and political interests (see Dunford and Greco (2006) and Dunford (2009) on which this entry draws).

**Neoclassical convergence models**

Neoclassical growth, regional development, and trade models all rest on the idea that regional development depends on an area’s resource endowments. Ricardian and Heckscher–Ohlin trade models treat resource endowments as exogenous. Solow-type growth models treat population growth and technical progress as exogenous, while capital resources depend on the rate of saving and the investment/accumulation of physical capital, although the impact of additional physical capital on growth is itself assumed to dwindle to zero. Neoclassical regional development models explicitly allow for the mobility of capital and labor and their impact on initial exogenous regional resource endowments, increasing human resources in areas of net in-migration, and capital in areas of net inward investment.

All three groups of models pay no attention to social relations and institutional performance. All assume that market mechanisms will automatically ensure that resources are employed fully and efficiently. These models all rest, however, on (unrealistic) assumptions of: (i) constant returns to scale, implying that costs of production do not change as the volume of output changes; (ii) the related assumption of perfect competition, with its implication that the economy is made up of small enterprises none of which can affect market prices; and (iii) an instantaneous re-employment of resources, implying that no resources remain involuntarily unemployed. Externalities, market failure, information asymmetries, and coordination failures (where an absence of investment by one enterprise makes it profitable for another to invest), which in practice play an important role in shaping regional development trajectories, are assumed not to exist and not to matter.

The Solow growth model and associated growth accounting exercises were developed as an alternative to Keynesian Harrod-Domar (HD) models. HD models attributed national and regional growth to the savings rate (on which capital investment depended) and the productivity of capital. Clearly, savings and investment are necessary, although not sufficient conditions for growth, which will also depend on human capital and institutions. These Keynesian models also suggested that actual growth could fall short of the rate required to ensure full employment, and identified instability as a possible feature of capitalist economies. The Solow model was radically different. The model itself rested on the idea that regional economic performance depends on (i) growth in the volume of capital and labor inputs, of which an area’s resource endowment is made up, and
(ii) increases in productivity, which depend on technical change. It predicts, however, that the growth path of a country or region will converge on a steady state. As long as labor is perfectly mobile (or regional rates of population growth are identical) and technological know-how is public, this steady state growth rate is the same for every country or region, while, in this steady state, per capita variables such as per capita output do not vary. The existence of similar conditions in the economies under consideration would, therefore, result in a convergence in per capita income. The reason why is that economies that are less developed are expected to grow/approach the steady state faster than economies that are more developed, while, in the steady state, per capita variables do not change. Empirically this claim is associated with the $\beta$ convergence hypothesis (Barro and Sala i Martin 1992).

Two factors underpin the expectation that economic growth will be faster in less developed economies. The first is the assumption of diminishing returns to capital and constant returns to scale. At any given level of technological development, these conditions imply that: (i) increases in per capita output decline as the amount of capital per worker increases; (ii) increases in productivity cease once a steady-state/equilibrium amount of capital per worker is reached; (iii) increases in per capita output associated with incremental investments are greatest in areas/enterprises that are the least mechanized; and (iv) countries that differ only in initial levels of per capita GDP will converge on the same level of GDP per capita. The second factor is the view that technology is a public good, available at no cost to everyone, and that the diffusion of technology and knowledge from advanced to less developed enterprises and areas will close technology and productivity gaps. Technology can only diffuse, of course, if there is no shortage of capital to invest.

Neoclassical models of regional economic development suggest that the flow of capital to low-income areas and of labor to high-income areas will result in an equalization of income per head. The movement of labor from low- to high-wage areas will raise wages in the former and reduce them in the latter. Capital, conversely, will move in the opposite direction, as the rate of profit is expected to be higher in less developed areas where wages are lower (Figure 2a), as will knowledge (Figure 2d). Movement will continue, it is argued, until there are no further differences in rates of return in different areas. Capital in-migration will be the greater the smaller the out-movement of workers and the smaller the rise in local wages. A number of recent cases indicating an interdependence between exporting skilled workers and importing capital in the industries in which they develop capabilities and acquire technological know-how place a question mark over this conclusion, while the assumption of constant returns to scale is crucial, as in its absence the net flow of capital may favor more developed areas (Figure 2b).

Neoclassical Heckscher–Ohlin factor proportions explanations of specialization and trade predict an equalization of factor prices independently of factor mobility. These models suggest that areas specialize in activities that make intensive use of resources that are relatively abundant. Areas with abundant supplies of skilled workers specialize, for example, in skill-intensive activities. Areas with abundant land resources per capita will specialize in land-intensive activities. Increases in the demand for relatively abundant resources and reduced demand for relatively scarce resources will raise the incomes of the former and reduce the incomes of the latter (Dunford et al. 2014), ultimately leading to factor price equalization.
Circular and cumulative causation

Models of circular and cumulative causation (CCC) differ significantly from the neoclassical tradition. These models identify positive/negative feedback mechanisms and emphasize the endogeneity of factors, such as resource endowments, considered exogenous in neoclassical models. CCC models suggest that comparative development depends on the relative weight of unequalizing, centripetal forces of attraction and suction and equalizing, centrifugal forces of diffusion.

Myrdal called the former backwash effects. An example is the way in which net flows of capital and labor are directed towards more rather than less developed areas (Figure 2b). As his aim was to explain increasing economic inequality between developed and underdeveloped areas, these centripetal forces received particular attention. Myrdal recognized, however, that there were spread (trickle-down) effects that worked to equalize development. Myrdal also
REGIONAL DEVELOPMENT MODELS

paid considerable attention to institutional factors and argued for active policy intervention to promote greater equality which, he thought, would contribute to greater economic growth. Net flows of public expenditure from developed to less developed areas designed to counteract the unequalizing effects of market mechanisms are an important example (Figure 2c).

Another strand of CCC models is associated with Kaldor, who drew on Young’s exposition of the possibilities of exploiting economies of scale in manufacturing and Verdoon’s assertions concerning the role of manufacturing as a driver of growth. The core idea was that economies of scale (assumed not to exist in the neoclassical tradition at that time) are a fundamental characteristic of economic life. Kaldor argued that national and regional growth was export-led. Export growth depends on the efficiency wage (the ratio of real wages and productivity). Increased exports of manufactured goods imply increased output. As a result of economies of scale, external economies/spillover effects, and complementarities, greater output implies increased productivity. Greater productivity improves competitiveness, contributing to a virtuous spiral. In the United Kingdom this model led to the introduction of wage subsidies in less developed areas in the shape of the regional employment premium (1967–1978).

The emphasis on export demand is indicative of a second major difference. Neoclassical growth models are supply-led, concentrate on the expansion of potential output on the supply side of the economy, and assume that aggregate demand will equal potential output. Kaldorian models are Keynesian models of demand-led growth. In these models, aggregate supply is adjusted to accommodate demand-led changes in actual output. Adjustment occurs through changes in capacity utilization and/or induced changes in accumulation, migration, and technical change.

Theories of modernization, dependency, and unequal exchange

Modernization theories have informed analyses of regional development. These theories examine the emergence of the modes of economic and social organization and life that first appeared in Europe and were subsequently extended to other parts of the world. An early example was Rostow’s model of stages of (neo-American) economic growth, in which all areas are considered to occupy different positions on the same universal development path (traditional society; the preconditions for take-off; take-off; the drive to maturity; and the age of high mass-consumption). Contemporary examples include neoliberal models of globalization and development that suggest that less developed areas simply find themselves at an earlier stage along a single development path already charted by the most developed capitalist economies in the world. Evident problems with these models are numerous and include, on the one hand, the way in which the emergence of economically advanced areas changes the conditions in which latecomers develop, and, on the other hand, the environmental unsustainability of the ways of life associated with existing concepts of modernity.

An influential East Asian variant of modernization theory was the flying geese paradigm. This model suggests that East Asian catch-up involved the emergence of a dominant growth center (Japan) that subsequently acted as the leader of a hierarchical group of followers that included a second tier comprising the Asian Tiger Economies, a third tier comprising Malaysia, Thailand, Indonesia, and most other members...
of the Association of Southeast Asian Nations (ASEAN), and a fourth group made up of China and Vietnam. Drawing on dynamic versions of trade theory, the idea was that the emerging economies in East Asia pursued, one after another: (i) a product cycle sequence involving successively the import of modern manufactured items, domestic production, export, and, finally, re-imports once production was moved offshore; (ii) an industrial sequence involving a movement from lower to higher value added activities that saw these economies specialize in a succession of supposedly increasingly sophisticated industries (textiles, chemicals, iron and steel, motor vehicles, and electronic products); and (iii) a transfer of products and industries from countries that were more advanced to countries occupying lower positions in the hierarchy.

Theories of dualism deal with structural asymmetries that perpetuate inequality. Developed especially in studies of economically under-developed areas, these theories are a variant of modernization theory that emphasized the existence of sharp differences in the organization, technological sophistication, degree of development, and goals of actors in the modern and traditional sectors of underdeveloped areas and envisaged a progressive expansion of the former at the expense of the latter. A classic example is the Lewis model, which posited the existence of disguised unemployment in the traditional agricultural sector/rural areas so that the transfer of surplus labor from the agricultural sector/areas to the modern sector/more developed areas would leave agricultural output unchanged yet increase the output of the modern sector.

In practice, rural–urban migration was far from smooth and varied widely from one country to another, with cases where: agricultural output declined (implying that the marginal productivity of farmers who left was not equal to zero); urban growth was insufficiently fast for the relatively capital-intensive modern sector to generate sufficient employment, leading to a rapid increase in informality and urban unemployment; profits were not re-invested domestically; and rural incomes were not large enough to provide a market for manufactured goods. To explain these phenomena neo–Marxist accounts of the articulation of modes of production interpreted dualism as a corollary of the rise of a money economy and industrial (as opposed to an earlier commercial) capitalism in less developed countries in which precapitalist subsistence economies or petty commodity production predominated.

Other models of regional development in underdeveloped countries identified vicious circles of underdevelopment. In many areas, for example, there are capital shortages due to insufficient savings, low productivity due to capital shortages, low real incomes due to low productivity, and insufficient savings due to low real incomes. To break out of this and other vicious circles and ensure a progressive expansion of the modern sectors of underdeveloped regional and national economies, a wide range of strategies were designed. Most targeted investment. Models of balanced growth proposed a series of complementary investments capable of providing market demand for each other (Nurkse) perhaps with the added impulse provided by a big push (Rosenstein–Rodan), and active development planning in part to overcome market coordination failures. Models of unbalanced growth (Hirschman) propose a concentration of investment in a few small projects whose backward and forward linkages sequentially create shortages, bottlenecks, and opportunities for further profitable investment. Growth pole models (Perroux) were a variant of an unbalanced growth strategy adapted to
the development plan approach that Hirschman opposed. For Perroux a growth pole is an enterprise or industry that has propulsive effects on related activities through the creation of external and agglomeration economies and forward and backward linkages, while for Boudeville the concept also applied to areas of concentrated development (growth centers and perhaps to industrial districts).

In the 1950s to 1980s, a more critical set of perspectives emerged in the shape of a renewed interest in (internal) colonialism and imperialism on the one hand and the emergence of theories of dependency and world systems on the other. These models contested the idea that all areas were following similar development paths and argued that the causes of relative underdevelopment were to be found not so much in the internal conditions in less developed areas as in the impacts of their dependence on economically more advanced/industrialized areas. Although economically underdeveloped areas did have distinctive characteristics, the characteristics considered were a result of relationships of dependence that created structural deformations, center-periphery relationships, and class divisions between externally-oriented elites and marginalized people within economically underdeveloped areas.

The original thesis derived from the argument of Singer and Prebisch, that the application of the principle of comparative advantage in the global economic system condemned some underdeveloped countries to a specialization in and export of raw materials and agricultural products and the import of manufactured goods from more developed countries. Contrary to the expectation that productivity growth in manufacturing would reduce the relative price of manufactured goods, Singer and Prebisch argued that the prices of rich-area manufactured items increased relative to the prices of the underdeveloped area exports. As a result, this division of labor created a self-perpetuating structure of development and underdevelopment, in which poor countries imported high value added goods and exported low value added goods, so that export earnings were never sufficient to pay for imports. If an increase in the volume of exports from less developed areas was more than offset by deteriorating terms of trade, a further possibility was that of immiserizing growth.

Theories of unequal exchange offered more radical interpretations of the drivers of the terms of trade, arguing that wages are relatively low in economically underdeveloped countries and that rates of profit do not differ (with capital actually flowing to economically-developed areas), so that the merchandise produced in underdeveloped areas is sold at relatively low prices: a worker in a rich area can, therefore, purchase for a few hours of his/her work the product of an entire day’s work of a worker in an underdeveloped area. This proposition was extremely controversial for it suggested that, instead of capitalists exploiting workers, workers in rich areas exploited workers in less developed areas: the former make the latter pay for their high wages through unequal exchange.

The dependency thesis was interpreted more radically with the rise of radical dependency and world systems approaches to explain the underdevelopment of dependent/peripheral/satellite and of semi-peripheral countries by their political and economic subordination to the interests of dominant/center/metropolitan economies and their domestic allies. Underdevelopment was viewed as radically different from nondevelopment or an absence of development on the grounds that the resources of peripheral economies are actively used rather than not used, although in ways that create internal
rural–metropolitan imbalances as underdeveloped economies are organized to satisfy the needs of their export sectors, the enclaves in which they are located, their rich externally-oriented elites, and metropolitan economies.

In the 1980s and 1990s a neoliberal counterrevolution, the Washington consensus and the augmented Washington consensus, saw the emphasis switch to the Solow-type neoclassical models considered earlier. Development, it was argued, should be market-led as markets were considered efficient; collective assets should be privatized; countries and regions should be opened up to global flows of capital and money; and government intervention should be minimized and largely confined to macro-economic stabilization, the protection of private and intellectual property rights and the creation of markets where they are missing (due to externalities for example) or imperfect. And if these measures did not yield economic progress the reason (as this discourse was fundamentally rationalist and deductive) lay not in the empirical invalidity of the neoliberal model but in the absence of good governance and institutions capable of creating the required conditions.

In regional development studies the abysmal growth record of the emerging economies that adopted these measures saw, however, the re-emergence of some earlier themes to deal with low returns to private investment and the cost and availability of finance. Sachs called for greater foreign aid, and a big push involving simultaneous increases in the capital stock (public, human and private) to overcome low returns to capital. The idea was that less developed areas face certain obstacles due to a lack of complements for modern industries, an inability to secure increasing returns to scale since initial capital levels were too small, an absence of higher savings rates channeled into the formal financial and banking system to provide further funds for investment as income increased, or increases in population growth due to higher incomes. The aim was to ensure that returns to capital increase and the economy shifts from low output per unit of capital to high output as the marginal productivity of capital increases, and that the marginal productivity of labor in the urban industrial sector increases relative to marginal productivity in agriculture.

Models of endogenous development I

Theories of dependency led to policies for infant–industry protectionism, import substitution, and autocentric, self-centered and locally-controlled development. In regional development studies two factors underpinned this change of emphasis. The first was the impact of the economic crises of the 1970s on exogenous or externally-controlled investments in many economically underdeveloped areas. The second was the wave of productive decentralization that these crises precipitated, the subsequent identification of industrial districts (especially in southern Europe) as drivers of regional economic development, and the yet later observation of the dynamics of new technology industries.

In many cases, areas of industrial growth were dense concentrations of interdependent small- and medium-sized firms (SMEs) in a single sector and in auxiliary industries and services. These districts were interpreted as the result of two sets of forces. The first comprised economic forces including: (i) scale economies that result from a high degree of specialization and division of labor; (ii) external economies that arise from the existence of shared infrastructures, services, and information; and (iii) the availability of special skills and the pooling of the workforce. The second included the interactions between the economic and social system that generated a
REGIONAL DEVELOPMENT MODELS

social atmosphere and communities of firms and people conducive to industrial development. These considerations opened the door to models dealing with the social, cultural, political, and institutional foundations of the district model, including analyses of social norms and values, political subcultures, associationalism, good governance, institutional density and performance, conventions, trust, social capital, and entrepreneurship.

These theoretical developments were associated with some important shifts in the models that dominated regional development strategy thinking. Self-reliance, autocentric growth and development from below were advocated as an alternative to development from above. New engines of regional economic growth were identified: instead of the consumer and intermediate goods sectors that had dominated growth, attention switched in developed countries to investments in human capital, information and communications technologies (ICT), and the commodification of knowledge to produce informational goods. More generally, emphasis was placed on innovation and learning, investment in research and development, the diffusion of knowledge and their social, cultural, and institutional determinants.

These ideas led, in due course, to approaches to regional development studies whose starting point was a conception not of a world divided into regional economies but of the firm and its environment or, at a meso scale, of global value chains and global production networks spread across the surface of the Earth. Firms were conceived as complex organizations driven by market competition and by a quest to expand and accumulate wealth, and also as a part of an ordered system of firms that makes up an increasingly globalized economy. Competition and the drive by individual firms to accumulate on the one hand and their interactions with other firms and institutions on the other leads to a constant equalization (competition and copying) and differentiation (innovation) of the conditions of production and exchange: the evolution of technologies, processes and products, and the characteristics of places create profit and growth opportunities and are associated with the rise, restructuring, and decline of growth centers and growth poles, constantly reshaping the geography of regional development.

Models of endogenous growth II

These developments in geography were accompanied by the development in economics of new neoclassical models of endogenous growth. Although these models were supply-led, some conclusions were similar to those of demand-led Keynesian cumulative causation models. Associated principally with Rohmer and Lucas, endogenous growth models assume nondiminishing, or, in the limit, nonzero marginal returns to the accumulation of physical capital, human capital, and knowledge. Nondiminishing returns arise because factor accumulation is associated with endogenous technical progress: the development of successive generations of equipment and goods and services are associated with learning, cumulative improvements in human knowledge and technology, and cumulative increases in human skill and know-how. At first these improvements were considered to be spillovers that led to increasing returns to scale at the level of the economy but constant returns at the level of the firm. As a result, economists could continue to assume that firms operated as if they inhabited a world of perfect competition. Subsequently, allowance was also made for economies of scale at the level of the firm and for imperfect competition. This change reflected a final recognition that firms
that carry out organized research and development can recoup the costs incurred by securing a temporary monopoly and restricting temporarily the diffusion of resulting innovations.

Growth can, therefore, stem from societal spillovers associated with the general advance of knowledge, and from the results of partially appropriable research and development. An important feature of this approach is the reconceptualization of the nature of knowledge. First, knowledge itself is a nonrival good that can potentially be used simultaneously by different people, and over and over again by the same people at close to zero marginal cost. Second, knowledge is a partially excludable good protected by patents and intellectual property rights and not a public good as assumed in Solow-type models. Third, investments that increase human knowledge yield increasing returns at the level of the firm for two reasons: (i) enterprises that manage to limit the diffusion of new ideas can spread the fixed costs of developing new products and technologies over a large volume of sales, yielding strong decreases in average costs as output increases; (ii) investments that increase knowledge nonetheless generate strong externalities – since knowledge cannot be perfectly patented or kept secret (and is therefore only a partially excludable good): once it is known that something can be done, other enterprises can seek to replicate it, so that new knowledge has a positive effect on their production possibilities.

Applied to comparative development studies, this movement away from models of constant returns to scale and perfect competition to models of increasing returns and imperfect competition can imply cumulative causation and divergence rather than convergence, as new investments in places and enterprises that are already advanced and associated improvements in knowledge will lead to a continuation of investment and growth and constantly renew development gaps, helping to explain why economically less developed areas fail to catch up.

A striking application of these ideas is found in Quah’s prediction of the emergence of twin peaks (with relatively high frequencies of high and low incomes and a relatively low frequency of middle incomes) in the distribution of regional income/household income. Quah associates this trend with weightless (superstar) economies in which value is embodied in immaterial things. On the supply side of weightless economies, ideas are themselves commodities. The immaterial goods that result are distinct in that they do not have to be transferred from one person to another but are simply replicated or copied at a marginal cost that is close to zero (and where, if equipment is required to use these goods, the upfront costs are small). On the demand side there are two further factors. First, the market for many immaterial goods is extremely large and often global in character (lifting the constraints on the division of labor posed by the extent of the market). Second, consumers often prefer to purchase the products/services of famous people rather than those of people who are less famous but whose talents, skills, and abilities may differ only marginally from the most famous. The combination of a very high level of demand for the goods/services offered by superstars and their very low costs of reproduction creates very high incomes in a world of winners and losers and wide inequalities. These mechanisms clearly have an impact on territorial development as immaterial goods and services are produced somewhere just as the people whose incomes are derived from them must live and work somewhere. Quah suggests that this twin-peak model may be applicable to regional economies.

The concept of endogeneity was important for another reason. Most models of regional development deal with determinate dynamic systems
REGIONAL DEVELOPMENT MODELS

which move towards a final state determined by exogenous data. In economic geography a great deal of recent attention has been paid to evolutionary models that deal with indeterminate dynamic systems in which outcomes are path-dependent. Analyses of indeterminate dynamic systems lead to models of regional development in which the current growth of an area is influenced by its own past record. The models examined earlier do point to generic causal mechanisms. Regional growth is, however, historically-contingent, requiring analysis also in terms of concepts such as cumulative causation, lock in, hysteresis, and evolutionary change.

Economic geography models

Explaining regional development implies explaining the territorial division of labor (who does what, where, and when, what rewards they receive, and in what relationships they stand to other people and economic activities in other places) and the constantly evolving resource endowments on which it depends.

In the past, geographical analyses of the territorial division of labor concentrated on abstract models of economic landscapes (von Thünen, Weber, Hotelling, Christaller, and Lösch). In the 1970s, Hymer, Lipietz and Massey developed conceptually informed classifications of the functional roles of regional economies in wider national and international divisions of labor. In the 1980s and 1990s a diverse array of conceptual models were advanced to characterize geographical systems of production, explain their structure, and account for their relative economic performance. Spatial categories examined included: industrial districts; areas of specialized production, local productive systems, and system areas; localized industrial systems; clusters; new industrial spaces; technopoles and science parks comprising innovative enterprises, research centers, and universities; worlds of production; milieux innovateurs; regional innovation systems; and learning regions. To explain these spatial structures and their reciprocal impact on development, three main groups of causal mechanisms were examined: (i) resource endowments; (ii) industrial organization, industrial strategies, externalities, and transaction costs; and (iii) innovation, systems of innovation, knowledge, individual and collective learning and creativity.

As a result, however, of interpretations of ideas associated with theories of regulation, institutional economics, or evolutionary sociology, these mechanisms were seen as operating in a series of specific historical contexts and of comprising a range of historical development models.

Another important recent development is the new geographical economics, which seeks to explain the emergence of economic landscapes (characterized by spatial agglomeration, regional specialization, and core–periphery structures) and the regional resource endowments that were taken as the starting point for many earlier models of trade and specialization in particular. Models are developed to examine the location of profit maximizing firms often in a world of increasing returns and imperfect competition, drawing on analyses of the increasing returns, forward- and backward-linkages, external economies, endogenous growth, product differentiation, external diseconomies, and so on to generate imaginary economic landscapes (see Figure 3 and Dunford and Greco, 2006).

Conclusions

In this entry most attention has been paid to economic models of regional development. These economic models afford important insights,
Dispersed rural population, yet scale economies lead to relatively few production sites for industrial goods: how few depends on the trade-off between scale economies and transport (iceberg effect) costs.

Differential transportation costs lead to relative growth of increasing returns activities and population in transportation hubs.

Specialization, division of labor and co-location near suppliers and customers. External economies also encourage proximity.

Endogenous growth/accumulation and competition differentially increase/decrease size of economic activities. Dispersion due to diseconomies/reduced transportation/factor costs.

Mobility reinforces agglomeration as does market access: suppliers of differentiated products in dense markets.

**Figure 3** Explaining economic landscapes. Source: Dunford and Greco (2006). Reproduced by permission of John Wiley & Sons, Ltd.
leading, in particular, to the identification of two sets of mechanisms. The first is a set of centrifugal forces, of which one of the most important is investment in infrastructure, skills, and economic activities and the transfer of technological and organizational knowledge to less developed areas or less advanced enterprises. Another centrifugal force arises if stagnation and relative decline occur in relatively developed areas. These mechanisms result in a tendency for an equalization of the conditions of production and exchange and of levels of development. The second is a set of centripetal forces. Centripetal forces arise when enterprises and areas that are developed create new sources of competitive advantage through, in particular, further investments in infrastructure, plant, equipment, knowledge, and skills. These mechanisms lead to a constant differentiation of the conditions of production and exchange and of the characteristics of regional economies. The value of these models is that they identify these and other equalizing and unequalizing mechanisms. The relative performance of different regional economies depends on the relative weight of these two sets of forces.

At any moment in time, existing resource endowments matter a great deal. Second nature resource endowments are, however, a consequence of earlier stages in historical processes of development: what happens at each point in time depends in part on what has happened in the past, and while each step can usually be explained in retrospect, it is normally impossible to predict what will happen next, as it depends on choices that cannot be anticipated with any degree of certainty. As a result, regional development is endogenous, cumulative, and path-dependent, with distinct possibilities of the emergence of virtual and vicious circles. The recent growth of interest in evolutionary models of regional development capable of dealing with indeterminate dynamic systems, in which current growth and transformation are influenced by the past record of development and outcomes are path-dependent, is, therefore, a significant step forward.

The relative weight of equalizing and unequalizing forces can change over the course of time. As a result, the evolutions of regional economies are associated with varying degrees of instability. Of the models considered it was Harrod-Domar and Keynesian-type demand-side models that predicted instability. Orthodox neoclassical models consider market systems as self-equilibrating and were ironically given an impetus by the relative stability of managed economies in the postwar Golden Age in Europe, North America, and western outposts.

Marxist and regulation approaches view capitalist development as involving phases of stable growth punctuated by periods of crisis. These crises are often important turning points in historical processes of regional development and in the development of ideas. Stable growth derives from the emergence of a sequence of new development models often centered on fundamental transformations of the preceding economic and social order. These new development models take shape in phases of crisis when older socioeconomic orders fail economically and are rejected politically and socially. As development models succeed one another, so do ways of thinking about regional development. A case in point is the shift from state-led developmentalism in the 1950s and 1960s to market-led approaches in the 1980s and 1990s.

The western financial crisis that erupted in 2007–2008 will mark yet another turning point for two reasons. The first is that it was simply another step in what was a progressive slowdown in the growth of developed regional and national economies. Alongside the relative stagnation of developed economies a number
of emerging economies have sustained high rates of economic growth over relatively long periods of time, radically altering the map of comparative development. The second is that many of these rising powers are associated with distinctive systems of political, social, and economic organization and policy regimes.

In this entry some reference has been made to social relations and to political, institutional, and cultural mechanisms, capacities, and performance as well as to the wider national and international economic and geopolitical context. Generally speaking, political and cultural mechanisms are not considered in modeling approaches designed to identify a small number of economically important factors seen as the main drivers of regional development. Economic mechanisms are, however, always grounded in particular political, institutional, and cultural contexts. Ever since the collapse of the Soviet Union, governments in developed countries have sought politically, diplomatically, and militarily to promote not just markets but also representative democracy. An examination of comparative regional economic performance indicates, however, that many areas that have achieved remarkable progress have on the one hand not embraced this political model and on the other are characterized by strong governance capacity. In the years ahead working out the implications of these trends and examining, in particular, the role of governance capacity and different social models will play an important role in studies of regional development.

SEE ALSO: Cores and peripheries; Cumulative causation, endogenous growth, and regional development; Local development; Regional inequalities; Trade and regional development

References


Further reading


Regional economic integration

Gianmarco I.P. Ottaviano
London School of Economics and Political Science, UK

Regional economic integration is the process of creating a unified market for goods, services, and factors of production (capital and labor) across the borders of countries belonging to the same geographical area. In the unified market, goods and services are freely traded, factors of production are freely mobile, and taxation and regulation are harmonized, so that economic activities face the same conditions both between and within countries. It differs, therefore, from national integration which involves the creation of a common framework of law and a common set of duties and obligations applicable to all the citizens of a country.

Real-world agreements differ in the extent to which the countries involved push the goal of economic integration along various economic dimensions. In a preferential trade area (PTA) member countries give one another preferential access to their markets without completely removing custom tariffs. In a free trade area (FTA), member countries agree to abolish custom tariffs among themselves while independently choosing custom tariffs with respect to the rest of the world. In a custom union countries also agree on a common external tariff (CET). The adoption of a common currency creates a monetary union, while the free mobility of factors transforms a PTA into a common market, and the combination of a customs union with a common market gives rise to an economic union. In a fiscal union member countries agree on a common fiscal and budgetary policy. The full harmonization of economic policies and the introduction of supranational institutions lead to a political union.

Examples of economic and monetary unions are the Caribbean Single Market and Economy (CSME) and European Union (EU). La Communauté Économique et Monétaire de l’Afrique Centrale (Economic and Monetary Community of Central Africa, CEMAC) and Union Économique et Monétaire Ouest Afrique (Economic and Monetary Union of West Africa, UEMOA) are customs and monetary unions. The European Economic Area (EEA) and European Free Trade Association (EFTA) in Europe are common markets. Custom unions include Comunidad Andina (Andean Community, CAN) and Mercado Común del Sur (Common Market of the South, MERCOSUR) in South America, the Eurasian Economic Community Customs Union (CUBKR) and European Customs Union (EUCU) in Eurasia, and the East African Community (EAC) and Southern African Customs Union (SACU) in Africa. The Association of Southeast Asian Nations Free Trade Area (AFTA), Gulf Cooperation Council (GCC), and South Asia Free Trade Agreement (SAFTA) in Asia; the Central European Free Trade Agreement (CEFTA) in Europe; the Commonwealth of Independent States Free Trade Area (CISFTA) in Eurasia; the Common Market for Eastern and Southern Africa (COMESA) in Africa; the Greater Arab Free Trade Area (GAFTA) in Africa and Asia; the North American Free Trade Agreement (NAFTA) in North America; the Sistema de la
Integración Centroamericana (Central American Integration System, SICA) in Central America; and the Trans-Pacific Partnership (TPP) in Asia, North and South America, and Oceania are all free trade areas. Even when they span different continents, these examples have a clear regional connotation as they typically involve countries that identify themselves as belonging to specific geographical areas. For instance, even though TPP countries span three different continents, they nonetheless feel attached to a common Asia–Pacific region.

Cutting through the complexity of these agreements, from the point of view of neoclassical economic geography/geographical economics, the most salient distinction between them is whether they seek to promote the cross-border mobility of goods (and services) only or also the cross-border mobility of capital and labor. The reason is that different provisions on what is allowed to move across borders may have different implications in terms of the induced spatial distribution of economic activities.

When goods (but not production factors) are allowed to move across borders and markets are perfectly competitive (i.e., production faces constant returns to scale and neither buyers nor sellers control goods prices because, for each good, there are several buyers and sellers), global economic integration shapes the global economic landscape by fostering the specialization of countries across sectors according to the relative efficiency of their production processes, driven by different technologies (comparative advantage) or different relative availability of factors that different sectors use with different relative intensity (factor proportions). As buyers are allowed to source from the most efficient sellers whatever countries they belong to, the latter expand and the overall efficiency of production increases. This gain is diluted in the case of regional economic integration. Freer goods mobility between member countries still fosters their specialization and trade according to comparative advantage and factor availability (trade creation). It also alters, however, the assessment of these relative characteristics with respect to nonmember countries, leading buyers to source from the most efficient sellers among their regional partners rather than from the world’s most efficient sellers (trade diversion).

When markets are imperfectly competitive (i.e., production faces increasing returns to scale and sellers control the prices because there are only a few of them for each good or they supply differentiated goods), regional economic integration has additional effects on the spatial distribution of economic activities. Lower barriers between member countries create a regional market that attracts producers from nonmember countries. This relocation of sellers (market size effect) is good for buyers in the integrating area but bad for buyers outside it. It also implies trade creation and trade diversion as before: regional integration pushes buyers in the member countries to source disproportionately from sellers based in the integrating area. If comparative advantage and factor proportions are controlled for, accrued efficiency gains are driven by the exploitation of economies of scale in the integrated area. Forgone efficiency gains are, instead, due to the fact that economies of scale would be better exploited in a larger globally integrated market.

The market-size effect is potenitated in the presence of value chains that serve consumers through a sequence of intermediate and final production stages. To see how that happens, one can consider a simple example in which final production uses only intermediate inputs, intermediate production employs only labor, workers are the only source of final demand, and labor is geographically immobile. By lowering the barriers to goods trade between member
countries, regional integration again enlarges the size of the market inside the integrating area. This makes that market more attractive for sellers, engendering self-reinforcing effects on the location of intermediate and final producers. Specifically, if we choose intermediate producers as first movers, for the sake of argument, their inflow into the integrating area increases its labor demand and intermediate supply. As a result of excess demand and supply, respectively, wages rise and intermediate prices fall. This is bad for intermediate producers in the area (market crowding effect) but it is good for final suppliers, which experience falling production costs and higher demand by richer workers. In turn, this makes the market of the integrating area even more attractive for final producers. As they enter its market, the expansion of final production feeds back into stronger intermediate demand so that intermediate suppliers also benefit (market expansion effect). When the market-expansion effect dominates the market-crowding effect, both final and intermediate producers end up being agglomerated inside the integrating area. This mechanism of cumulative causation between upstream and downstream producers’ location decisions can generate persistent asymmetries in the spatial distribution of economic activities, even between insider and outsider countries that are very similar in terms of comparative advantage and factor proportions. An analogous story can be traced using final rather than intermediate producers as first movers.

Cumulative causation is more likely to materialize in the presence of labor mobility and capital accumulation. The reason is that both tend to reinforce the market expansion effect through the additional income they generate. Rising intermediate production in the integrating area increases labor demand and wages in member countries, attracting workers from nonmember countries. As workers move, final demand in the member countries increases, generating an additional cycle of cumulative causation between firms’ and workers’ location decisions. At the same time, the expansion of intermediate and final markets increases profits, fostering investment in capital accumulation.

Regional integration may affect the distribution of economic activities not only between member and nonmember countries, but also within member countries as market crowding and market expansion effects operate inside the integrating area. When market crowding dominates, the economic geography of the integrating area tracks the spatial distribution of comparative advantage and factor proportions. When the market expansion effect dominates, cumulative causation sets in, leading to the possible emergence of a core–periphery pattern of location whereby economic activities concentrate in a small subset of member countries.

When the economic landscape is shaped by the location decisions of people and firms in imperfectly competitive markets, it is generally inefficient. There is, therefore, room for efficiency-enhancing regional policies, for instance, to contrast the emergence of inefficient core–periphery patterns through taxes and subsidies. Even when core–periphery patterns are efficient, regional policy could still be used on equity grounds to redistribute the associated benefits through transfers. It is the need of additional policy levers that calls for deeper regional economic integration.

Another issue of policy relevance is the extent to which regional economic integration paves the way to or stands in the way of global economic integration, which is preferred as it does not cause any trade diversion. Different theoretical mechanisms have been identified through which regional provisions may work as “building blocks” or “stumbling blocks” in the construction of an integrated global economy.
REGиональное интегрирование

Империальное доказательство в пользу какого-либо из этих взглядов, однако, неубедительно.

СИЕАЛ: Международное деление труда; Региональные модели развития; Региональные неравенства

Дальнейшее чтение

Бэлдин, Р.Э., Р. Форслид, П. Мартин, и др. 2003. Экономическая география и общественная политика. Принстон, Нью-Джерси: Принстонский университет.


Regional finance

Sheila Dow
University of Stirling, UK
University of Victoria, Canada

“Regional finance” refers to research in the interface between economic geography and economics concerning the relationship between financial variables and real economic development. It has tended to be a small interface since the role of finance has traditionally been regarded as peripheral to economic geography, while spatial concerns have traditionally been regarded as peripheral to economics. Indeed, it is now being argued by some within economic geography that, even if the spatial dimension of finance had been relevant in the past, it is no longer relevant. Nevertheless, the view persists in economic geography that finance continues to be important for regions (Garretsen, Kitson, and Martin 2009 and the following papers).

Within economics, any regional finance analysis needs to be understood in terms of the approach taken to economics more generally (Dow and Rodriguez Fuentes 1997). The mainstream approach is based on the view that money and finance are not fundamental to real economic development. Rather money simply determines the rate of inflation, that is, money is “neutral” (Patinkin 2008) and finance affects growth only insofar as it is more or less efficient. That market inefficiency may nevertheless be widespread opens up the possibility of finance having real effects on regional economies. Such inefficiency is the focus of the New Keynesian stream of mainstream economics (Dixon 2008). This latter approach is to be distinguished from the nonmainstream Post-Keynesian approach, which does not have a neutral-money, full-information benchmark, but rather analyzes substantive and enduring interrelationships between real economic conditions and the financial sector, which have a regional dimension (King 2008; Dow 1987). The concern in the Post-Keynesian literature, rather than being with market efficiency, is that money and finance should be economically functional.

How the regional character of national economies and of the national financial sector is regarded therefore depends on theoretical perspective. One of the main focuses of the literature has been the regional segmentation of the national financial system. This segmentation is understood in mainstream economics as a source of inefficiency, but for Post-Keynesians it is seen potentially as a protection for regional economies. But the nature of the regional segmentation of finance has evolved over the decades as the financial sector has undergone substantial changes. These changes have included the growth of nonbank financial intermediation within national economies, the consequent widening of the range of functions of banks in order to compete, the evolution of international financial arrangements which affect the domestic institutional structure, as in the European Union, and the globalization of finance.

A second major focus has been the spatial impact of national monetary policy, where the nature of such policy has also undergone major changes. Since interest in financial variables, and money in particular, had been fairly limited within regional economics and economic geography, much of the literature draws on the
money-macroeconomic analysis of European Monetary Union (EMU). The mainstream neutral-money view in this literature focuses on the relationship between changes in the official interest rate and inflation within a currency union with free flow of capital. The Post-Keynesian view rather considers differential effects of interest rate changes and central bank signaling on both the cost and availability of regional credit, as well as on expenditure plans and the demand for credit. While the first only allows for a real regional impact of monetary policy as a result of market inefficiency, the second sees a real impact as the norm.

The significance of the theoretical approach extends to the data themselves. Where regional data are unavailable, a common practice is to impute regional monetary aggregates from regional gross domestic product (GDP) shares, building the assumption of money neutrality into the data. In contrast to this mainstream focus on relations between aggregates, the Post-Keynesian approach focuses more on disaggregated financial and economic data and, in their absence, a range of different types of evidence (Rodriguez Fuentes 2006, 58–59). More generally, the field has been able to develop more where regional data are available, such as in the United States, Canada, Spain, and Italy. The best dataset now is probably the Brazilian one developed at the Center for Regional Planning and Development (CEDEPLAR) of the Federal University of Minas Gerais under the leadership of Marco Crocco. (An example of the application of these data to the Post-Keynesian theory of regional finance is set out in Crocco et al. 2012.) The situation is circular. Data are collected where there is more spatial awareness and they can be employed in analysis to inform and test regional finance theory, encouraging further data collection. Where a lack of interest in spatial matters has limited data collection, the limitations this poses for research mean that any lack of interest is reinforced.

A third focus of the literature has been debated as to whether the financial sector generally promotes economic convergence or divergence. The premise of the single European financial market was that it would promote convergence, while the globalization of finance more generally was encouraging the idea that spatial segmentation of finance was ceasing to be a factor in regional economic development. But then the 1997 Southeast Asia crisis raised questions about the continuing, if not increasing, potential for financial liberalization, and particularly free capital flows, to have real economic consequences. The banking crisis that began in 2007 then focused attention more generally on the relationship between the financial sector and the real economy and the consequences of limited availability of bank credit for business.

The analysis of this recent crisis provides a good example of the way in which the three theoretical approaches outlined above address a financial development which clearly has had real effects. The standard mainstream approach sees the crisis as a temporary departure from equilibrium, the result of misguided efforts by the monetary authorities to keep interest rates below their long-run equilibrium level resulting in excessive money creation. The New Keynesian analysis has focused on a range of other distortions of market forces which are more systemic. Many of these sources of market inefficiency were associated with the development of new products (derivatives) and practices (securitization) and come under the category of asymmetric information, although there is also a focus on state protection of failing banks. For both approaches, financial instability and its real consequences can be reduced by increased financial efficiency. In contrast, the Post-Keynesian analysis starts from the view that economies are inherently
unstable and that the financial sector is a major contributor to this instability. Financial fragility and the instability it brings about are exacerbated by financialization, such that financial inefficiency is advocated as a means of reducing the scope for fragility. (Hence the proposal for a financial transactions tax, for example.)

The spatial dimension of the real effects of the crisis is evident in the differing experiences of different economies, whether independent national economies or national economies within a currency union as in the EMU. It has been a persistent outcome of the crisis that there has been divergence in particular between EMU economies along standard center-periphery lines, with a cumulative process of capital outflows from peripheral economies at a time of rising public sector borrowing needs. Relief in the form of European credit was dependent on the introduction of fiscal austerity. From a mainstream perspective, this policy was designed to restore equilibrium through a process of convergence with the aid of private sector capital inflows. From a Post-Keynesian perspective, however, austerity policies have promoted divergence, causing recession, worsening the public sector finances and productivity, and further discouraging private capital inflows. This experience in the EMU demonstrates that there is considerable scope for analyzing the spatial effects of the crisis also at the subnational level. Here, too it is important whether capital flows are equilibrating (promoting convergence) or disequilibrating (promoting divergence).

In what follows, we explore the three topics noted above in turn, starting with the spatial structure of finance and proceeding to discuss the regional impact of monetary policy. The final section focuses on the ultimate question of the real spatial effects of finance. In the process we consider the implications of choice of theoretical approach.

The spatial structure of the financial sector

The histories of different national financial sectors, and in particular of banking, show varying spatial configurations. Awareness of the significance of such configurations for regional economies was perhaps most evident in the United States, where early banking legislation was designed to ensure spatial dispersion of banking; there was a political concern that the rest of the US economy might suffer from the financial strength of New York if the banking system were free to concentrate there. Indeed, the modern literature on regional finance first emerged in the United States in the 1950s, focusing on the relative economic performance of states, given national financial conditions and financial flows between states. But the monetarist premise of much of this literature was that capital flowed freely within and between states at the interest rate set by national monetary policy, such that it was real economic conditions rather than the regional configuration of the banking system that determined regional money flows.

The emergence and persistence of financial centers and their capacity to retain and build on their dominance of financial markets was one focus of the banking history literature. But alongside this enduring phenomenon there have been more complex spatial developments in financial structure. Since the second half of the twentieth century, while banking remains at the core of the financial sector (providing reserves and a payments system for the rest of the sector), other types of institution have increased dramatically in scale and importance. The process of deregulation, which started in the 1970s, allowed retail banks to expand into (spatially centralized) nonbank activities, such as insurance and investment banking. Increasing
competition within the financial sector and the increasing volatility in financial markets in that period threatened the survival of smaller institutions, encouraging a wave of mergers and acquisitions, leading to increased concentration in finance around financial centers. The focus of financial activity was shifting from the traditional economic functionality of providing a means of payment and business and household credit towards activity in speculative financial markets. This process was widely supported on the grounds that it fostered financial market efficiency by increasing the depth of financial markets and reducing spreads, as well as breaking down local monopolies.

These processes were particularly evident in Europe, with the emergence of the single European market in finance and the single European currency. The aim was to remove national barriers to capital flows, increasing allocative and cost efficiency. Financial liberalization also occurred at the global level, again with a view to increased efficiency, encouraging more free flow of capital between national economies. At the same time, information technological advance greatly enhanced the capacity to trade internationally in a range of instruments and markets, accelerating the globalization of finance. The presumption was that there would be general benefit from cheaper and better financial services and the more free distribution of credit according to expected returns. With some qualifications in terms of the fallout of the financial crisis, this is seen by some to indicate the end of geography.

While financial liberalization has led to concentration in the financial sector rather than increased competitiveness, nevertheless some regional segmentation persists. Large businesses tend to conduct their finances within the financial center; in the case of the largest businesses their finances are conducted within their own treasury departments. In any case these businesses have access to capital markets as well as to banks. But small and medium-sized businesses (SMEs) are more dependent on bank credit which remains primarily nationally, if not locally, based. So the issues of the potential for a national banking system to have differential effects on regional development remain.

Where there are nationwide banks, distance (in its various forms) of borrower from lender is shown to influence the availability and cost of credit, to the disadvantage of peripheral regions (Porteous 1995). For SMEs in peripheral regions, the information on which judgements about risk are based is particularly weak for national banks whose credit decisions are made centrally according to uniform conventional criteria. New Keynesians argue therefore that banks will ration credit in peripheral regions. Post-Keynesians agree that local financial institutions benefit from a stronger local knowledge base and thus capacity to assess risk and monitor borrowers is stronger than for national banks. But rather than being concerned with concealment of risk, Post-Keynesians are concerned that uncertainty precludes any notion of objective risk; risk assessment is a matter for judgement, which is socially conditioned. Further, when these judgements are based on weak knowledge, the result may be not just too little credit, but also possibly too much. Weak knowledge may thus explain overenthusiastic supply of credit to peripheral regions at times to finance projects (often natural resource-based) which hold the prospect of unusual profits. By the same token, such capital inflows can be readily reversed when the banks’ confidence in the projects weakens, contributing to increased instability for these economies (Chick and Dow 1988).

Further, Chick and Dow (1988) argue that credit availability and cost in peripheral regions will be influenced by local banks being at an
earlier stage of development than banks in the financial center. They have less easy access to wholesale markets than large national banks and a deposit base vulnerable to swings in local economic conditions. Thus interest rates in local banks will be less advantageous for depositors than those in national banks. Local banks also tend to have simpler portfolios than the universal banks in the financial center, and are thus disadvantaged by the increasingly strict regulatory requirements being introduced in the wake of the crisis, such as higher capital requirements. Local banks therefore are more constrained in their access to funding and competitively weaker than large national banks.

It is therefore of critical importance for peripheral regions how the banking system is structured: the perception of risk attached to lending in peripheral regions varies between national and local banks. But attitude to risk may also vary over time, as expressed by the liquidity preference of national and local banks with respect to peripheral regions. Even with a national branch banking system, it cannot be assumed that there is an elastic supply of credit to all regions (even at a markup over the national rate) if national banks’ liquidity preference means an unwillingness to extend credit in peripheral economies (Dow 1982).

The regional impact of monetary policy

While the literature on the regional impact of monetary policy has also been consistent with differences in theoretical approach (see Rodriguez Fuentes and Dow 2003; Rodriguez Fuentes 2006), the context for analysis of the regional impact of monetary policy has also evolved. In particular, the current discourse on monetary policy in the wake of the crisis has shifted attention from the neutral-money view of monetary policy as a means of inflation control to the view that monetary policy can have real consequences, at least in the short run. The fact that the early supply of liquidity averted a much worse crisis, however, lends support to a more lasting real effect of central bank actions.

The focus of the mainstream regional impact of monetary policy literature has been on the different regional responses to changes in a uniform national interest rate. As in monetarist international macroeconomics, the presumption is that, where money flows freely between economies with fixed (or no) exchange rates, monetary policy has no lasting real impact (regionally or nationally). Nevertheless, some allowance is made for compositional real effects where there are regional differences in interest elasticity of demand for credit, with an emphasis on regional economic structure. Thus, for example, a region whose economy was less diversified would respond differently to a monetary policy shock than a diversified region. Regional differences in financial conditions have also been addressed as a source of a differential regional impact of monetary policy in the mainstream literature. Thus regional differences in levels of wealth and in portfolio composition (based on exogenous preferences) will affect the way in which an interest rate change is transmitted to each region (Miller 1978; Goodhart 1989). For example, a higher level of household mortgage indebtedness in one region makes it more vulnerable to changes in the official interest rate than other regions. More recently, attention has shifted to the institutional environment as a factor in the transmission of monetary policy, for example exploring differences in the relative significance of the bank lending channel and the balance sheet channel between different European economies. While financial market segmentation may affect the regional transmission of monetary policy, this possibility is lessened to
the extent that financial market liberalization has eroded this segmentation.

The Post-Keynesian literature has probed behind the empirical relationship between the official interest rate and regional economies, following up two particular strands of reasoning. First, there may be unsatisfied demand for credit in some regions and excessive volumes of credit in others, while the long experience of financial vulnerability in peripheral regions may even discourage the emergence of credit demand. But if the starting point is not, as in the mainstream approach, an equilibrium position, it is hard to identify statistically the effect of national monetary policy. Second, monetary policy is understood in broader terms than changes in the official rate. Not only may markups on the official rate vary regionally in line with different levels of confidence in different regional economies, but what is regarded as creditworthy demand will vary regionally, affecting credit availability. More generally, liquidity preference and risk assessment, which can vary systematically by region, will be influenced by what is increasingly regarded as a central tool of monetary policy: the communication of central bank analysis.

A central feature of the Post-Keynesian approach is that credit creation is endogenous, that is, it is under the control of the banks, albeit influenced by the central bank. Since the central bank enforces its official interest rate by managing the supply of liquidity to the banks, the banks can create credit confident that the supply of reserves to back new credit is assured. If credit creation in any region is weak, this is more a reflection of the banks’ unwillingness to lend in that region than it is of monetary policy. This puts the spotlight on banks’ credit policies and the strength of the knowledge on which those policies are based. In contrast, according to the mainstream approach, monetary policy establishes a money supply (indirectly through the official interest rate) which the banks then distribute among the regions according to relative creditworthiness. If credit levels are low in peripheral regions, according to this view, it is due to real factors, that is, that regional productivity is low, or to information asymmetry.

The replacement of national monetary policy with a common EMU monetary policy provides a useful case study. One of the rationales for European Monetary Union was not only that it would reduce asymmetric shocks for each member economy, but that it would also remove the balance of payments constraint resulting from any shocks because of the greater freedom of capital flows within a currency union. The interest rate set by the European Central Bank would determine a common rate throughout the EMU. Membership of the EMU would be conditional on convergence in terms of financial variables, which the EMU would then sustain. But the conclusion that the EMU would promote continued convergence is based on mainstream theory which understands market behavior as equilibrating, with labor and capital flowing to maintain common returns across the EMU. The Post-Keynesian approach, however, allows for the possibility that the decline of a region’s sector resulting in labor outflows can if anything discourage capital inflows, driving the economy further from equilibrium; without an independent currency the only option for the region is to contract demand. If an asymmetric shock depresses expectations of returns on capital, then outflows of capital (lower bank lending) in depressed member economies would be encouraged, promoting economic divergence. If changes in interest rates due to monetary policy further disadvantage peripheral regions, then the forces for divergence would be compounded.
Finance and regional development

Regional finance clearly concerns more than just the regional impact of monetary policy, important though that is. There is now a much better understanding, supported by statistical evidence, of the relationship between finance and growth. But again difference of theoretical perspective is crucial. The mainstream literature focuses on the argument that economic growth is served by increased financial efficiency, and that this should be the focus of policy. The more efficient the financial sector, the closer is money to neutrality. The literature on finance and regional development within the Post-Keynesian approach focuses rather on the economic functionality of the financial sector in supporting real economic activity, given that money is not neutral (see Rodriguez Fuentes 2013 and the following papers). The regional character of the relationship between the financial and real sectors arises from the regional composition of the financial sector and from the regional pattern of economic activity, wealth levels, and financial behavior, all of which account for the non-neutrality of money (Chick and Dow 1988).

We have seen that the regional structure of finance is important for the regional pattern of credit creation. But for Post-Keynesians the liquidity preference, not only of banks but also of households and firms within peripheral economies, also plays a part. Whatever the level of expected returns in a region, if confidence in expectations is low, then there will be high liquidity preference in the form of low willingness to lend, borrow, or spend. While these may vary in the short run, the problem for peripheral regions is that there may be a persistently high level of liquidity preference. Unlike the exogenous portfolio preferences of the mainstream approach, Post-Keynesian liquidity preference is endogenous to economic processes. In order to understand financial behavior in peripheral regions, we need to consider the implications of long experience of vulnerability. Peripheral regions are more vulnerable because wealth levels tend to be lower than in the central region and also because there is less assured access to credit. So any economic instability may cause cash flow problems and difficulty servicing debt, further contributing to a normally conservative assessment of default risk. Peripheral regional economies tend to be more unstable than central regions, with dependency on a smaller number of sectors. Where these include raw material sectors, peripheral economies may go through boom periods. The weak knowledge base of the banks can allow large capital inflows which are then subject to sudden reversals.

The reasonable portfolio strategy is then for peripheral regions to have relatively high liquidity preference, not just as a short-run response to cyclical instability, but as a secular response to continuing higher levels of uncertainty and vulnerability to credit constraints. Several studies have found evidence to support this thesis (Porteous 1995; Amado 1997; Rodriguez Fuentes 2006). High liquidity preference takes the form of reluctance to commit to capital investment, reluctance to take on debt, and allocation of wealth to assets issued in the financial center as the most liquid available. Thus, for example, peripheral economies in Europe have been buying foreign (mainly German) sovereign debt, at the same time as limiting availability of credit to local companies. The effect of high liquidity preference is thus weak growth in peripheral regions, weak demand for credit, and higher capital outflows.

It is a common refrain from national banks that their low credit creation in peripheral regions is due to weak levels of demand that is creditworthy. But both may be the outcome of high liquidity preference. Creditworthiness
is a matter of judgement for the banks, where a poor knowledge base is compounded by the consequences of high liquidity preference. The credit decision is performative. By constraining credit availability in peripheral regions, national banks create the conditions that justify their cautious assessment of regional risk. More generally, the centralization of finance leads to the centralization of economic activity.

Concluding reflections

We have seen that the Post-Keynesian literature on regional finance has postulated a range of causal mechanisms which explain the aggregate relationships on which so much else of the literature focuses. These relationships imply that the regional dimension of financial structures and behavior are an important element in generating the regional effects of monetary policy and financial developments more generally. Since structures and their regional dimensions evolve over time and vary from one economy to another, analysis of regional finance in any region is conditional on the relevant economic and financial context. Thus, for example, the way in which financial sectors in different countries have responded to the crisis has differed because of differences in the initial financial structure and in policy, with differing consequences for the regional pattern of economic growth.

Since structure is amenable to policy influence, there is therefore scope for governments to promote a financial structure that is economically functional in general and addressed to the different needs of different regions in particular. Thus support for smaller local banks with more sound knowledge of local conditions would ensure a more robust provision of credit in peripheral regions; such support could take the form of subsidies, or lower capital requirements, for example. Such measures are required to counter the forces for divergence which are encouraged by financial liberalization. These measures would, in particular, increase confidence in and knowledge about peripheral economies and thus increase the willingness to both demand and supply credit to finance real investment.

But such measures would, from a mainstream perspective, damage financial efficiency. The mainstream approach rules out real effects of regional finance except in the form of financial inefficiency so that regional economies are best served by focusing on increasing financial efficiency, which also promotes economic convergence; the assumptions and structure of the theoretical framework and the formulation of evidence all serve to produce this conclusion. The Post-Keynesian approach rather adopts a theoretical framework, and seeks evidence reflecting the priority of economic functionality, with a view to promoting more, and more even, economic development. Choice of theoretical perspective is fundamental to policy direction in the field of regional finance, as in any field.

Acknowledgments: Comments and suggestions from Alexander Dow, Mick Dunford, and Carlos Rodriguez Fuentes are gratefully acknowledged.

SEE ALSO: Finance and development; Financial geography; Regional economic integration; Regulation/deregulation

References


Further reading


The region is, and always will be, a central concept in geography. So, thinking about regions can tell us a lot about the history of the discipline called geography. Before this though, the perennial question must be addressed: What is a region? In this entry, a region is considered as a “temporary permanence,” not fixed and absolute, but something held stable at different points in time and for different purposes. There is, then, no single reading of a region. Regions are multiple entities and geographical analysis should seek to uncover the process and practices of region-making.

Regions can have administrative territorial shape through the drawing of borders and boundaries and the making of state spaces. Regions can also delimit economic forms relating to particular functions, for example the spaces of exchange and trading such that we can talk of discrete regional economies in global networks. Regions can also act as symbolic shapes – they can become cultural identifiers and sources of belonging, and senses of ownership for political expression and mobilization become interesting objects of analysis thereafter. One helpful way of categorizing these distinctions is to think about the administrative and economic as “regional spaces” and the latter (cultural and political) as “spaces of regionalism” (Jones and MacLeod 2004). Again, these distinctions are never out there waiting to be found; they are forged historically through political and policy struggles involving state and non-state forces. This allows geographers to talk not about “regional geography,” which automatically conjures up fixity and rigidity, but about a “geography of regions” – the process and practices of making and remaking regions through “institutionalization” as Anssi Paasi (1996) puts it. The remaining discussion unpacks this, using four chronological phases, to reveal how geographers have grappled with regional geography. There is a need “to go back” to go “forward,” as Thrift (1994) has put it.

An era of traditional regional geography existed during the long nineteenth century. Here, the region was treated absolute: as an independent backdrop for, first, conducting geographical inquiry and, second, developing thereafter a world of different regional geographical types. Books such as Britain and the British Seas (1902) by Halford Mackinder made assumptions about the interconnected nature of the human and the physical. This regional geography was an ethnographical geography of successive phases and component elements, with (in this case) the physical environment as thus analyzed. The work of Andrew Herbertson (1905) developed “climatic regions” from this, and Hebert J. Fleure offered “human zones” conditioned, rather than entirely governed, by physical circumstances in his Human Geography in Western Europe (1919).

The backlash to this came from the likes of Richard Hartshone, whose The Nature of Geography (1939), contra Herbertson and Fleure, claimed that none of their regional geographies “alone” covered a significant number of features to offer an adequate background to regional study, or to provide a tentative system for organizing all our regional knowledge of
the world. A regional geography of elements (meso-worlds of human–physical interactions) was proposed to produce an elaborate regional geography of “areal differentiation” – the division of the Earth surface by observations that different areas of the world are somehow self-contained and where, extending the work of Paul Vidal de la Blache, relationships between people and land created a distinctive geography. The purpose of this regional geography was to understand what these differences are and how they are related. Books such as *One World Divided* (1964) by Preston James summarized the long results of this thinking through offering different areal differentiations.

Regional science, assisted by the power of computation and the desire to study connected variables as part of modernist social sciences dabbling with theories of physics (such as gravity models), logically followed from this. Walter Isard’s *Methods of Regional Analysis* (1960) suggested that any general theory of the region, whether this be in relation to locational analysis of industry or population, must be supplemented by techniques of regional analysis, which are operational techniques and which yield estimates of basic magnitudes for the space economy for both the proper understanding of social problems and policy formulation. Isard’s work sought to make and model the internal connections between things that flowed through and were seen to actively make regions, albeit in a discrete and bounded way, isolated from external drivers behind the capitalist space economy.

This concern with the structural and strategic forces acting in and on regions could be felt by those critiquing regional science and offering, through regional studies and largely influenced by Marxism, alternatives based around the dynamics and contradictions of capitalism. Books by Doreen Massey such as *The Anatomy of Job Loss* (1982) and the 1984 classic *Spatial Divisions of Labour* presented regions as the medium and outcome of economic and political struggles around the geography of production and class struggle. “Rounds of investment,” phases of capital accumulation dealt out like playing cards, as Derek Gregory (1989) put it, acted as waves on a beach, constantly making and remaking regions in geography as a distinctive economic geography of “localities.”

This emphasis on capital versus the regions, combined with the emphasis on place as a historically contingent process, meant that the mid- to late 1980s was ripe for thinking about a new regional geography. Whereas traditional regional geography talked about fixed and bounded territorial and process entities, new regional geography wanted to study regions as produced and transformed through various forms of agency. Anne Gilbert’s classic (1988) summary of this position drew attention to regions as varying local responses to capitalist processes; regions as the focus of cultural identification (or senses of place); and regions as the medium for social interaction or meeting places for human agency and social structures. The studies of localities in the United Kingdom, funded by the Economic and Social Research Council (ESRC), offered an economic and social window into this, although they often generated more heat than light. Authors such as Nigel Thrift took this further and laid down four challenges for regions in geography: how distinctive meanings are produced and contested; the changing forms of regions in space and time; the relations between people and nature and the deconstruction of landscape; and representing and writing about regions.

Globalization discourses and the politically charged nature of economic development, especially in Europe, overtook these debates, and in the early 1990s geographers and other social scientists were talking about the *new*
regionalism. This regional geography saw regions, not nations per se, as the crucibles for both economic success and democratic legitimacy. A rescaling of economic and social life was being witnessed, according to protagonists, who drew much attention to regional things occurring in Emilia-Romagna (Italy), Baden-Württemberg (Germany), Silicon Valley (United States), the Rhône-Alpes (France), and Tokyo (Japan). Michael Storper’s *The Regional World* (1997) and Michael Keating’s *The New Regionalism in Western Europe* (1998) summarized these trends and speculated variously on their wider applicability.

In short, regional geography was being read through a new economic geography of globalization and regional development and a new political geography of territorial politics and public policy communities. Critics such as Gordon MacLeod (2001) sought to draw attention to three shortcomings of this literature: (i) a “soft institutionalism” (where institutions are seen as the explanation for economic success, as opposed to looking at the historical geography of institutions); (ii) “thin political economy” (the absence of any analysis of the state and the political economy of territory, scale, and region-making); and (iii) “distorted policy” (the inability of some scholars to examine path dependency, the problems of policy transfer and policy borrowing, and how case studies are used to develop both academic and political knowledge.

We are currently witnessing a phase called the new new regional geography. One aspect of this does not and cannot believe in regions in geography. A group of scholars, following Ash Amin and Nigel Thrift (the latter somewhat strangely, as he was a previous advocate of regions in geography) and others advocating a “thinking space-relationally” approach, envisage a world without regions, where rhizomatic flows and networks dominate and help to promote a new politics of place based on “proquinity” (the state of being close to someone or something, achieved through the “folding” of existence, and made possible by globalization and technological empowerment). Three significant points follow from this position (see Amin 2004): regions have no automatic promise of territorial integrity since they are made through the spatiality of relational connectivity; there is a questionable assumption that a regional territory can then be defined and made manageable; and any attempt to fix spatial identities through policy initiatives will be characterized by oversimplification and an inability to capture the ever-changing character of regions. This approach has, in turn, produced a backlash, with scholars seeking to bring regions, territories, and geographical scales back onto the agenda. Their argument is not to completely drop regions but to think about regional geography differently.

Part of the problem is the way the debate is posed, both historically and now, as if the networked worlds of those advocating relational space stand in direct opposition to territories, and, vice versa, those advocating a network position will also find that these are not without regional geography anchors. Geographers are, therefore, seeking a way of reconciling regional geography within a territorial/relational world. In the new new regional geography, where regions remain temporary permanences, the languages of “bundling,” “assemblage,” “method assemblages,” “spatiotemporal fixing,” “com- possibility,” “phase space,” “new localities,” “lengthening of networks,” and so on are all being offered as various ways of examining “region-making.”

A paper by US geographers Alexander Murphy and John O’Loughlin (2009) outlines the framework and mandate needed for an emerging new new regional geography as a tool for understanding rapid and profound changes in
the contemporary world. The authors provide recommendations for, and note the benefits of, a strengthening of regional research and training in geography while, in their case study, mapping out an expanded role for Eurasian geography and economics in disseminating new regional research whose scope extends beyond disciplinary boundaries to embrace current public and political debate. In this sense, then, the important problems that human geographers face today can still be found grounded in the practice of doing regional geography.

**SEE ALSO:** Areal differentiation (or chorology); Regional definition and classification; Regional planning: the resilience of an imperative; Regional political movements; Regionalism

**References**


**Further reading**


Regional inequalities

Michael Dunford
Institute of Geographic Sciences and Natural Resources Research (IGSNRR), Chinese Academy of Sciences

All regions and all places are unique and unrepeatabe, as indeed are all natural and social phenomena. Generalizations can nonetheless be made and theories developed by making abstractions that identify characteristics that regions have in common, such as types of soil, settlement types, or income levels. In scientific approaches to geography these abstractions are the starting point for the making of generalizations and the development of theories, which are themselves the starting point for a process of thought which moves in the opposite direction, from abstract theories to concrete realities, in which all of the details set aside in the process of abstraction are reintegrated such that particular regions emerge as a synthesis of many determinations. In this approach to geography, inequality is used not to refer to an absence of sameness but to differences in a narrower subset of characteristics of different places.

The literature on regional inequalities deals with a large range of indicators representing the wealth, income, employment, life expectancy, health, and education of the inhabitants of different places. Most of this literature focuses on income inequalities, although attempts to explain them involve consideration of possible economic, demographic, political, and cultural differences that may cause income inequalities. In this entry, most attention will be paid to inequalities in output, income, and expenditure, their evolution, and some of the factors that underlie them. Attention will also be paid to the identification of regions and regional classifications and to indicators of inequality.

Territorial units

Essentially, there are three ways in which regions can be defined. Homogeneous regions are regions that are made up of areas that share certain classified characteristics (climate type, soil type, population characteristics, etc.). Analytical or functional regions are regions that are made up of places that are interdependent, such as a travel-to-work area comprising employment centers and the places where the people working in these centers reside. Administrative regions are regions established to manage certain activities and include supranational, national, and subnational administrative entities. Subnational examples include local government districts, health districts, and so on, which are usually defined according to the size of the population considered necessary to carry out these tasks efficiently, perhaps respecting historical legacies or cultural distinctions. Most statistics are collected for administrative areas, though in some cases subnational administrative areas do not make functional sense as they separate areas that are strongly interdependent, as when centers of employment and wealth creation are separated from the places where the people who create that wealth reside.

An example of a regional classification used in regional analysis is the European Union’s (EU) Nomenclature of Units for Territorial...
REGIONAL INEQUALITIES

Statistics (NUTS). The NUTS system is at present a four-level hierarchical classification: each member state (NUTS 0) is divided into a whole number of NUTS 1 regions; each NUTS 1 region is in turn subdivided into a whole number of NUTS 2 regions; and so on. Generally speaking, the first three NUTS levels correspond to the main regional divisions in each member state: Länder and Kreise in Germany; régions and départements in France; Comunidades autónomas and provincias in Spain; and regioni and provincie in Italy. To secure a degree of international comparability and, more specifically, to try to ensure that the areas lie within certain size limits (3–7 million people in NUTS 1 areas, 800 000 to 3 million in NUTS 2 areas, and 150 000–800 000 in NUTS 3 areas) a third administrative level is identified for each member state: NUTS 1 for France, Italy, Greece, and Spain; NUTS 2 for Germany; and NUTS 3 for Belgium. At a more detailed level, there are districts, municipalities, and communes, although these levels are not yet subject to the formal NUTS regulation.

The difficulty is that the national administrative and statistical regional divisions that are the foundation stones of the NUTS classification are not derived from a common set of criteria and in some cases make no functional sense whatever. In spite of the functional interdependence of Inner and Outer London, Inner London is, for example, a NUTS 2 area. Although an attempt is made to identify areas that are comparable in economic size, the size limits are guidelines rather than strict constraints. Consequently, the NUTS classification does not provide a harmonized set of regional units. As a result, the 2011 gross domestic product at purchasing power standards of Level 2 regions varied from 542,049 million euros in Île-de-France to 974 million in Åland in Finland. The population of Åland was just 28,150.

Another eight NUTS 2 areas had fewer than 250,000 inhabitants. At the other end of the scale Île-de-France had 11.8 million inhabitants, Lombardia 9.9 million, and Andalucia 8.2 million.

In spite of these difficulties, the system is used for three important purposes. The first is the collection, development, and harmonization of regional statistics. The second is the analysis of regional economies and societies. The third is the framing of EU regional policies: eligibility for Objective 1 Structural Fund assistance is established at the NUTS 2 level, while eligibility under other priority Objectives is mainly established at the NUTS 3 level.

Indicators

Gross domestic product (GDP) per inhabitant is the indicator most often used to measure national and regional inequalities. A community’s GDP is a measure of aggregate value added. As the new wealth created in production is distributed to those who participate in economic life, GDP is also a measure of primary incomes (profits, interest, rents, and wages) accruing to those who contribute to a region’s economic activities, although the income that results does not necessarily accrue to a region’s inhabitants: where the inhabitants of other regions have property rights in a region, there is an outflow of income, as there is if the human capital of other regions is used locally. GDP estimates do not include the value of those goods and services that people produce for their own use. This omission is significant as self-provisioning, which depends on the resources a region’s inhabitants control and can use, varies significantly from one place to another, especially at a global scale. Conversely, the activities of the informal and hidden economies are included,
at least insofar as government statisticians are able to estimate their magnitude. The primary distribution of income is subsequently modified by state-administered redistribution. The result is the secondary distribution of income. These secondary incomes are saved and/or spent, and thereby permit the establishment of claims over the goods and services produced. Measurement of these expenditures (conventionally divided into consumer and government expenditure, investment/savings, and net exports) offers a third way of measuring GDP. (The existence of inconsistencies between these three estimates helps in the identification of undeclared incomes.)

The use of GDP is, however, problematic. If the aim is to measure wellbeing, wealth (stocks) should be measured as well as income (flows), and the concept of capital should embrace all assets that contribute to long-term welfare increases. Total wealth includes tangible productive capital, intangible capital (human capital, institutional capital, social capital, net overseas financial assets), and natural capital, while the flows include not just the income flows measured in GDP estimates but also, for example, changes in the availability of the environmental services provided by natural capital. In this world, externalities can render individual consumer preferences nonadditive, as increases in private consumption can reduce the quality of the environmental services that enter consumer preferences schedules.

Inequalities in regional GDP per head can be measured in two ways. Measurements in a common international currency indicate the international value of the output of regional economies. The money value of regional output shows (i) what the output of the exposed sector can be sold for and can command on international markets, and (ii) what the output of the protected sector can command directly on regional and indirectly on external markets. If a region’s external trade is in equilibrium, this indicator reflects the “quality” of its goods and services and its competitive strength. Measurements in purchasing power standards (PPS) make allowance for differences in the prices of goods and services in different areas: the quantity of goods and services that a given amount of an international currency will exchange for is greater in a low-cost than in a high-cost region. The PPS measure allows for these differences in purchasing power and is therefore an index of the volume of goods and services that different local economies produce and the volume of output that their economic activities can command or exchange for in their own area. This second measure is an indicator of differences in living standards. A related distinction can be made between qualitative and quantitative growth: increases in the value a region’s goods and services can command on international markets, and quantitative increases in the volume of goods and services commanded domestically.

GDP per head and its growth can be split into productivity and employment rate elements, as indicated by the identities in equations (1) and (2), where $G()$ denotes a growth rate.

\[
\text{GDP per head} = \frac{\text{gdp}}{\text{population}}
\]

\[
\text{GDP per head} = \frac{\text{gdp}}{\text{population}} + \frac{\text{employment}}{\text{population}}
\]

\[
\text{GDP per head} = \frac{\text{gdp}}{\text{population}} + \frac{\text{hours worked}}{\text{population}}
\]

\[
\text{GDP per head} = \frac{\text{gdp}}{\text{population}} + \frac{\text{hours worked}}{\text{population}}
\]

\[
\text{GDP per head} = \frac{\text{gdp}}{\text{population}} + \frac{\text{hours worked}}{\text{population}}
\]

\[
\text{GDP per head} = \frac{\text{gdp}}{\text{population}} + \frac{\text{hours worked}}{\text{population}}
\]
This partitioning of GDP per head into elements that reflect the productive performance of regional economies (apparent productivity) and some of the features of their labor markets (hours worked and employment rates) is widely used as an initial step in the identification of the determinants of regional inequalities.

**Measures of inequality**

The simplest measures of inequality are the standard deviation, the mean absolute deviation, the Gini coefficient, and the Theil index. In the case of the first three measures, unweighted and weighted variants are used. An unweighted indicator gives equal weight to the average per capita income of each regional economy. A weighted indicator makes allowance for the fact that the population of regional economies varies. In this case, the average income of each regional economy is weighted by its share of the population. Measures of social, as opposed to regional, inequality deal with the incomes of individual households or individuals, each of which is weighted equally, instead of regional averages.

The definitions of the most important of these indicators are set out in the following equations. If \( p_i \) denotes the population of the \( i \)th region \((i = 1, 2, \ldots, n)\), \( y_i \) denotes its per capita GDP, and \( q_i \) denotes its GDP, the population-weighted standard deviation (WSD) is given by equation (3) and the weighted absolute deviation is given by equation (4).

\[
WSD = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \frac{p_i}{\sum_{i=1}^{n} p_i}}
\]  

\[
WAD = \sum_{i=1}^{n} |y_i - \bar{y}| \frac{p_i}{\sum_{i=1}^{n} p_i}
\]  

To facilitate comparisons, the weighted standard deviation and the weighted absolute deviations are expressed as percentages of the mean to give a coefficient of variation (CV).

The Gini coefficient (GC) is given by equation (5).

\[
GC = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j|
\]  

As with the indicators based on the mean, the deviations from which the Gini coefficient is calculated can be weighted by the product of the shares of the total population in each pair of regions.

Finally, the Theil coefficient (TC) is given by equation (6).

\[
TC = \sum_{i=1}^{n} \frac{q_i}{\sum_{j=1}^{n} q_i} \log_e \left( \frac{\sum_{j=1}^{n} q_i}{\sum_{j=1}^{n} p_i} \right)
\]  

Derived from the concept of entropy developed in information theory to measure the amount of information in a random event, the Theil coefficient can be interpreted as “the expected information content of the indirect message which transforms the population shares as prior probabilities into the income shares as posterior probabilities,” and as a measure of the gap between the share of the population that
each individual or group accounts for and the share of income it receives.

Territorial units and the measurement of inequality

In analyzing regional inequality, it is important to recognize that measured inequality increases with the degree of spatial disaggregation. Suppose that a country is divided into 16 areas (A1, A2, …, D4) with similar populations but different levels of GDP per head, and that these areas are grouped first into four and then into two regions (see Figure 1a and Dunford 1993). The standard deviation expressed as a percentage of the mean decreases from 38.5% (16 areas) to 10.6% (4 areas: A1..B2, A3..B4, C1..D2 and C3..D4) and 3.22% (2 areas: A1..B4 and C1..D4). It is important to note, however, that the choice of regional boundaries can affect the result. If in Figure 1a the 16 areas are divided horizontally rather than vertically into two groups (A1..D2 and A3..D4), the indicator falls to 9.67% instead of 3.22%. Alternatively, if four areas are identified in the manner indicated in Figure 1b, the coefficient of variation will equal 24.7%.

Measured regional disparities depend, therefore, not just on the degree of spatial concentration of economic activities, but also on the regional division of the country: the number of areas and the choice of boundaries affect the measure of disparity, just as the delimitation of electoral districts shapes the outcome of elections. Clearly, the ideal solution is to use functional economic areas which combine places of work with corresponding places of residence, though disparities between politically identified areas are significant as determinants of the resources over which different communities can exercise political leverage.

The dynamics of regional inequalities: convergence or divergence

In 1955, Kuznets suggested that there is an inverted U-shaped relationship between inequality and income per head. As an economy starts to develop, and as people move from agricultural to industrial jobs and from rural to urban areas, inequality increases. As income per capita increases further, this relationship is reversed and inequality declines. For Kuznets, the technical reason for the increase in inequality was that the transition to an industrial society was the saving of unskilled labor. In 1965, Williamson extended this argument to regional inequalities, and argued that regional inequalities will at first increase and subsequently decrease as the level of economic development increases. Williamson argued that the evolution of regional inequalities was a result of a set of spillovers associated with different stages of economic development.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>

(b)

Figure 1  Measured inequality and regional division.
REGIONAL INEQUALITIES

development, including migration, capital flows, government policies, and interregional trade. As evidence, Williamson relied on time series data for the first half of the twentieth century, which showed that regional inequalities increased in less developed countries and decreased in more developed countries, and on cross-sectional data, which showed that regional disparities are greater in less developed than in more developed countries. Williamson concluded that regional income inequalities are a natural consequence of economic development and that any attempts to reduce them might inhibit development.

More recent work on the dynamics of regional inequalities has involved investigation of the convergence hypothesis (Figure 2). The initial idea derived from the expectation that economies characterized by similar conditions would converge on a similar steady state rate of growth and similar per capita incomes, with economies that are less developed growing faster than economies that are more developed. Empirically this claim is associated with the $\beta$ convergence hypothesis, which states that the slope parameter, $\left(1-e^{\beta t}\right)/t$, of regressions of per capita GDP growth in $t$ years, $y$, on the logarithm of initial GDP per head, $y_0$, will be negative. It is important to note that this exercise is statistically questionable, as it involves the estimation of an equation in which initial income appears on the two sides, once with a positive sign and once with a negative sign. In these circumstances a negative regression coefficient is at least in part a result of correlating a variable with itself.

Two factors underpin the expectation of faster growth of less developed economies. The first is the existence of diminishing returns to capital and constant returns to scale. At any level of technological development, these two conditions imply that increases in per capita output get smaller as the amount of capital per worker, $K/L$, increases (since areas with low $K/L$ ratios grow faster than areas with high $K/L$ ratios), that increases in productivity cease once a steady-state/equilibrium amount of capital per worker is reached, that increases in per capita output associated with incremental investments are greatest in areas/enterprises that are the least mechanized, and that countries that differ only in initial levels of per capita GDP will converge on the same level of GDP per capita. The second is the view that technology is a public good available at no cost to everyone, and that the diffusion of technology and of knowledge from advanced to less developed enterprises and areas will close technology and productivity gaps.

Empirical convergence hypothesis research showed that there were differences in regional

![Figure 2](image-url)  
**Figure 2** Modeling convergence.
steady state growth rates and permanent differences in income per head. As a result, emphasis was placed on theories of conditional and club convergence that predict cross-sectional heterogeneity as a result of differences in economic conditions. Club convergence theories attribute differences in performance to the fact that economies fall into different groups/clubs because of differing initial conditions. Conditional convergence theories attribute differences to the cross-area heterogeneity of control variables such as rates of accumulation or the quality of modes of governance. In this much more restricted context, the convergence that does occur is very slow: economies converge at only about 2% per year on their own steady state as determined by structural characteristics and starting conditions.

A second strand of empirical research concentrates more on the evolution of the dispersion of regional incomes (sometimes called θ convergence). In this case the unweighted and weighted indicators identified earlier are used.

A third strand of empirical research involves distribution dynamics. This approach involves the application of Markov chain techniques (using either discrete transition probability matrices or continuous stochastic kernels) to examine the evolution of entire income distributions. These methods have indicated complex changes in the dynamics of relative income per head, including divergence, the persistence of inequalities, movements within the distribution, and polarization. Modeling the evolution of the distribution of regional and individual income with these techniques, Quah (1996) found, for example, an empirical tendency for the emergence of twin peak distributions with high frequencies for relatively high and relatively low incomes and relatively small frequencies for incomes in the middle of the distribution.

With the exception of the third, these approaches are associated with the theoretical view that convergence is an expected outcome of market mechanisms. As this entry will show, however, divergence as well as convergence can occur. In the neoclassical tradition it is common to argue that the absence of convergence is due to factors that prevent markets from working. In many respects, this type of claim is extraordinary: if the model’s predictions are not realized, it is the world that is wrong and not the model. There are, however, alternative models of circular and cumulative causation (Myrdal 1957; Kaldor 1972) and new economic geography models (Fujita, Krugman, and Venables 1999) that do not predict convergence. According to these models, comparative development depends on the relative weight of unequalizing centripetal forces of attraction and suction and equalizing centrifugal forces of diffusion. Myrdal, for example, called the former “backwash effects.” An example is the way in which net flows of capital and labor are directed towards more rather than less developed areas. As his aim was to explain increasing economic inequality between developed and underdeveloped areas, these forces received particular attention. Myrdal recognized, however, that there were also spread (trickle-down) effects that worked to equalize development. In addition, he emphasized the importance of institutional factors in shaping development trends and argued for active policy intervention in order to promote greater equality, which he thought would also contribute to greater economic growth. An example of such public action is provided by net flows of public expenditure put in place specifically to counteract the unequalizing effects of market mechanisms.

Another strand of circular and cumulative causation models is rooted in the work of Kaldor, who drew on the earlier work of Young. At the center of this approach was the idea that
Regional Inequalities

Economies of scale were a fundamental characteristic of economic life (yet were assumed not to exist in the neoclassical tradition). Essentially, Kaldor argued that national and regional growth was export-led. The growth of exports depends on the efficiency wage (the ratio of real wages and productivity). Increased exports of manufactured goods imply increased output. The consequent increase in output implies, as a result of economies of scale, external economies, or spillover effects and complementarities, increased productivity. Next, increased productivity improves competitiveness, contributing to a virtuous spiral.

The emphasis on export demand is indicative of a second major difference. Neoclassical growth models are supply-led. These models concentrate on the expansion of potential output on the supply side of the economy, and assume that aggregate demand will equal potential output. Kaldorian models are Keynesian models of demand-led growth in which aggregate supply is adjusted to accommodate demand-led changes in actual output through changes in capacity utilization and/or induced changes in accumulation, migration, and technical change.

Global inequalities

Measured in terms of per capita GDP in 2013 Geary–Khamis (EKS) dollars (converted to 2013 price level with updated 2005 EKS PPS) (Figure 3), the world remained extremely unevenly developed, with high levels of per capita income in North America, Australia, and New Zealand ($51,575), Western Europe ($36,974), Japan ($38,004), and the Asian Tiger economies ($38,655). Next came Russia ($18,596) and the rest of Eastern Europe ($12,579), though there were very wide gaps between the central European Union member

Figure 3  Total population, GDP per capita, and total GDP. The column width of each block represents the population of the country or region. The area of each block represents the total GDP. Modified from The Conference Board (2014).
states and members of the Commonwealth of Independent States. In Africa, per capita GDP stood at just $2747. From being the poorest country in the world in 1949, lying well behind Africa and India, China ($11 106) was an upper middle-income country, had overtaken Brazil ($10 248), and was not far behind the rest of Latin America ($12 810). India ($3 681) still had relatively low levels of income per head. Due to the large size of its population, the Chinese economy was extremely large, as the surface of the areas in Figure 3 shows: in 2013 its total GDP at PPP stood at $15.1 trillion compared with $15.4 trillion for Western Europe and $19.5 trillion for North Americana and Oceania.

These disparities are a result of trends in comparative development mainly in the last 250 years. At the dawn of the Industrial Revolution, the gap in per capita income between Western Europe and India, Africa, or China was probably no more than 30%. With the Industrial Revolution, however, the situation changed radically. In 1870, the per capita income of the wealthiest countries was already 11 times that of the poorest. By 1995, the wealthiest were more than 50 times as rich as the poorest. In the recent past, however, many of the relatively low per capita income parts of Asia have experienced relatively fast economic growth (Figure 4). In most of the continental zones identified in Figure 4, growth rates were faster in the 1960–1973 Golden Age than in subsequent cycles. In the case of Western Europe, growth rates in each subsequent cycle were one-half or less of Golden Age growth rates. The USA, Canada, Australia, and New Zealand saw output grow relatively faster than Western Europe, although their per capita growth rates were at best the same as those of Western Europe, as higher GDP growth rates were accompanied by stronger demographic growth. Communist Eastern Europe also saw strong growth in 1960–1973. Subsequent growth

![Graph showing average annual GDP growth over several successive economic cycles.](image)

**Figure 4** Average annual rates of GDP growth over several successive economic cycles. Modified from The Conference Board (2014).
REGIONAL INEQUALITIES

slowdowns saw the collapse of communism and opened the path to rapid transitions to capitalism. The consequences of these transitions for output were little short of catastrophic. In 1989–1997, output declined at an average of 4.5% per year. In 1997 output still stood at just 68% of its 1989 level. Foregone output over these years was massive. In these transition economies growth subsequently picked up, yet in 2013 output stood at a mere 128% of its 1989 level. More recent growth has depended to a significant extent on net capital inflows which contributed to unsustainable credit-driven growth. This record compares particularly unfavorably with that of China, which chose a fundamentally different development path to the European ex-communist states: instead of shock therapy, which Václav Havel sought to justify with the claim that “you can’t cross a chasm in two small steps,” the Chinese chose a gradual and experimental approach to “reform and opening up” captured in Deng Xiaoping’s aphorism that, to paraphrase, “the way to cross a river is step by step, feeling for the stones as you go.” The remarkable growth of China at an average of 8.7% per year since 1980 is one of the reasons for the growth of Asia at 6.2, 5.1, 5.4, and 6.2% per year in the four cycles from 1960 until 2013, with the growth of Tiger economies slowing down during the 1990s. As for the other parts of the world, high growth rates in 1960–1973 gave way to much slower growth in 1973–1989, especially in the economies of Latin America and Africa. Just as the differences in the performance of European transition economies and China reflect in part different development choices, so do the contrasts between Asia and Latin American and African economies, as the latter implemented the Washington Consensus and the enhanced Washington Consensus (which required that the original goals of stabilization, liberalization, and privatization be accompanied by governance reforms and country ownership), whereas countries such as Japan and the Four Tigers, and later on China, India, and Vietnam, violated virtually all of the rules of neoliberalism.

As a result of the recent rapid growth of very populous countries in Asia, the great divergence set in motion by the Industrial Revolution is starting to give way to a period of global convergence, especially as growth has picked up as a result of recent changes in the political complexion and the economic strategies of Latin American economies and the natural-resource-driven growth of Russia and of Africa. If these trends continue, there will be a significant increase in the relative economic weight of emerging economies, especially in Asia, at the expense of the economically advanced parts of the world.

National and regional inequalities in Western Europe

Figure 5 records per capita GDP at PPS relative to the average in 2011 for the NUTS 2 regions in the EU27 (27 EU member states) and in Norway, Switzerland, Croatia, Turkey, and the former Yugoslav Republic of Macedonia. Even within the EU the gaps are very wide as a result of a geopolitically driven process of enlargement. The addition of countries such as Croatia, the former Yugoslav Republic of Macedonia, and Turkey will make them still wider.

In Figure 5 countries are ordered according to their GDP per head from Luxembourg (264.7%) and Norway (184.9%) to Turkey (35.7%), with national GDP per head represented by red diamonds. The extreme values differ by a factor of more than 15.7. Some of the figures should, however, be treated with caution. In the case of Ireland, GDP far exceeds GNP: the incomes that accrue to people who live in Ireland are some 17% less than the income created in Ireland.
The reason for this discrepancy is that much of the wealth created in Ireland is produced by multinational companies that withdraw profits and that engage in transfer pricing arrangements. Transfer pricing results in an overstatement of the wealth created in Ireland.

For each country, regional GDP per head figures are recorded in the columns as circles. In most countries, and in this group of countries as a whole, there are wide regional disparities. At one extreme lie Inner London (319.0%), Luxembourg (264.7%), Brussels (220.6%), Hamburg (201.2%), Oslo (188.1%), Bratislava (184.9%), Groningen (181.0%), and Île-de-France (181.0%), although some of these figures must be treated with extreme caution as the areas concerned are areas of substantial net inward commuting. At the other end of the spectrum were Severozapaden in Bulgaria (28.6%), Nord-Est in Romania (28.6%), and five regions in Turkey, of which the lowest was Mus, Bitlis, Hakkari (20.2%).

**Trends in European national inequalities**

Empirical evidence derived from research employing the indicators introduced earlier

---

**Figure 5** National and regional inequalities in the EU27 plus Croatia, the former Yugoslav Republic of Macedonia, and Turkey. Data modified from Eurostat (2014).
places a question mark over the Williamson hypothesis; the reason is that in many developed economies regional inequalities, which did indeed diminish in the 1960s and 1970s, started to increase again in the 1980s, 1990s, and 2000s. What is more, in these countries the rates of economic growth were faster in the years of diminishing regional inequality than in the subsequent years of widening regional inequality.

Figure 6 plots the evolution of disparities between West European and East European countries from 1950 until 2013. The chart indicates that there was strong convergence among West European countries from 1950 until the early 1970s. In this period, often referred to as a postwar Golden Age, European economies experienced unprecedentedly fast economic growth: growth was associated with increases in territorial (and social) equality, while the movement in the direction of greater equality made positive contributions to European growth through, for example, its effect on the expansion of markets. After 1971, convergence slowed, and in the early 1980s divergence occurred. After a period of relatively slow convergence, driven by the relatively fast growth of the economically less developed European Union Cohesion Countries, divergence occurred.

A partitioning of West European disparities to identify the roles of productivity and employment rates is particularly informative (Figure 7). As the figure shows, apparent productivity has converged initially quite rapidly and subsequently at progressively slower rates. The main driver of changing trends in inequality is disparities in employment rates, which increased in the late 1970s and early 1990s after the mid-1970s economic crisis of Fordism and again after the
2007 financial crisis and with the subsequent debt and Eurozone crises.

In Central and Eastern Europe, on the other hand, the transition to capitalism saw sharp increases in disparities until the end of the 1990s, after which recovery from the dramatic transitional recession spread to other countries (Figure 7). Figure 7 identifies the trends in GDP growth of some of the Central, East European, and Central Asian transition economies since 1989, and demonstrates the remarkable differential impact of economic transition on GDP (see Dunford 2005). The Visegrad Four (Czech Republic, Hungary, Poland, and Slovakia), Slovenia, and the former German Democratic Republic (GDR) regained their former levels of GDP after sharp slumps and substantial foregone income. Poland was the first to regain its pre-transition GDP in 1996, followed by Slovenia in 1998. Of the rest, Albania regained its former level of GDP in 2000, and two Central Asian republics (Turkmenistan and Uzbekistan) did in 2001. In the case of Russia, the drop in output was about 42%. This fall was far steeper than the one registered by the United States during the Great Depression of the early 1930s. At the root of the collapse was a decline in income and demand, combined with an overvaluation of the rouble. Only after the 1998 financial crisis, devaluation, and the subsequent increase in oil prices did growth pick up, and only in 2007 did Russia reach its 1989 GDP. The Ukraine in 2011 stood at 66.7% of its 1989 level of GDP.

**Italian regional evolutions**

In the case of Italy’s regional economies there was a sharp decline in disparities in 1951–1953 after which territorial inequalities increased in the 1950s. In 1960, the degree of inequality was close to its 1951 level. In 1960–1975 the situation changed dramatically. For 15 years, coinciding in part with Italy’s economic miracle, there was strong catch-up as the less developed parts of Italy closed the gap on the more developed. After 1975 there was a reversal in the trend. Overall there was a clear increase in inequality. A not insignificant part of the relative improvement that had occurred after 1960 was reversed in 1975 and after 1983. Since the mid-1990s, conversely, there has been a renewed phase of convergence (Dunford and Greco 2006).

The recent evolution of Italian regional inequalities is a result of opposite trends in regional productivity and employment rates. As Figure 8 shows, productivity differentials have tended to decline. In 1972–1996 employment rate differentials generally moved in the opposite direction. Clearly, it is not productivity differentials that account for the recent overall increase in Italian regional inequality, as they have tended to diminish, though they remain a significant determinant of differences in the levels of regional development. At the root of the inequality increases up to 1996 were very sharp increases in employment rate variations.

As these two examples show, there are relatively sustained periods in which unequalizing tendencies outweigh equalizing tendencies and vice versa, where these empirical outcomes also reflect, of course, the impact of government policies designed to reduce development differences.

A second important feature of empirical trends in relative development is the persistence of territorial inequalities. Figure 9 records the evolution of Italy’s regional economies relative to the EU15 (15 Members of the EU before the accession of formerly communist East European countries) average. Among other things, it reveals a high degree of inertia at the top and bottom end of the distribution; the top regional economies in 1951 remained for the most part...
Figure 8  Trends in Italian regional inequality: disparities in regional GDP per resident, apparent productivity, and employment rates, 1951–2013. Modified from I.Stat (2014).
Figure 9  Growth of Italy’s regional economies relative to the EU15 average, 1951–2013. Data modified from I.Stat (2014).
at the top end of the distribution in 2013, just as the regional economies at the bottom end tended to remain at the bottom, although it also indicates some striking changes in the relative position of regional economies in the middle of the distribution.

**Chinese regional disparities**

The existence of phases of increasing and diminishing inequalities is also a characteristic of the extraordinary growth of China since the establishment of the new China in 1949, and especially since the start of reform and opening-up after 1978. Figure 10 records trends in disparities in per capita GDP between China’s 22 provinces, four municipalities, and five autonomous regions (excluding therefore the two Special Administrative Regions of Hong Kong and Macau, and Taiwan).

Generally speaking, disparities were relatively high in the planned economy period due mainly to the concentration of industrial growth in Northeast China (Figure 10). Disparities diminished from the mid-1970s and with the first phase of reform and opening-up but then increased strongly from 1990 until about 2004 with the rapid growth of east coast cities and coastal provinces. The end of the 1990s and the start of the new millennium saw, however, the establishment of government programs for the development of Western China, the restructuring of the Northeast, and the uplifting of Central China.

**Figure 10** Evolution of GDP per capita of China’s four main territorial divisions. Modified from National Bureau of Statistics (2014) (http://www.stats.gov.cn/).
China, which with the impact of the downturn in Chinese export markets after the Western financial crisis saw a reversal of trends and a decline in particular in the coastal–inland gap.

China’s regional disparities are, moreover, closely associated with disparities between rural and urban areas: as Figure 11 shows, increases and decreases in regional disparities were accompanied more or less by corresponding trends in the gap between rural and urban areas. These evolutions reflect the close relationship between regional and rural–urban disparities and phases in Chinese economic and political development. The early years of communist rule marked by land reform were, for example, associated with a low and steady degree of rural–urban inequality. In 1957, China embarked on the Great Leap Forward: while investment in heavy industries increased sharply, an agricultural crisis gave way to the Great Famine. In these years inequalities increased, peaking in 1960. From 1960 until 1967 inequalities diminished. 1966 saw the start of the Cultural Revolution which began a phase of increasing inequality, with a peak in 1976. 1978 marked a major turning point. An initial phase of rural reform saw inequalities continue to decline. After 1984, in a period of political decentralization,
increased foreign trade, and foreign direct investment, provincial, urban–rural, and inland–coastal inequalities all increased sharply. Overall inequality was relatively small when the development model was favorable to agriculture and the countryside.

Conclusions

In this entry most attention has been paid to regional variations in wealth creation. The evolving geography of wealth creation is underpinned by the evolving geography of economic activities and the dynamics of labor markets and population distribution. These drivers are themselves shaped not just by economic mechanisms but also by institutional factors and political choices (including, for example, specific policies designed to create a more equal distribution of people and work). Each of these drivers is associated with equalizing and unequalizing tendencies. Whether regional inequalities increase or decrease depends on the relative weight of these two sets of forces and, more generally, on the underlying model of development. Inequalities are, however, not just a result of the geography of wealth creation. Income redistribution and transfers towards economically disadvantaged regions are often substantial so that variations in income and consumption are often smaller than variations in wealth creation. Health, education, and collective services are often entitlements and as such are distributed far more equally than other economic activities.

SEE ALSO: Cumulative causation, endogenous growth, and regional development; Regional definition and classification; Regional development models; Uneven regional development

References


REGIONAL INEQUALITIES


Further reading


Regional planning in China is a new sphere of activity compared with Western countries. It dates from the 1980s when it focused on “economic cooperation zones,” such as the “planning economic cooperation zone along the Yellow River” (Liu et al. 2011). Since 2000, regional planning in China has been developed rapidly on the basis of learning from Western countries’ successful regional governance experiences.

Regional planning in China refers specifically to an integrative arrangement for development issues in an area (Yang 2010). The core components of regional planning in China include the general strategic development orientation of the region, regional development targets, regional industrial structure, overall arrangement of cities and towns, infrastructure construction arrangements, and so on. Generally speaking, regional planning has served as a method of assisting with the realization of the goals of national development. More specifically, breaking the economic shackles of the administrative division of the country, optimizing the development of regional geographical space, and advancing sustainable regional development are the three goals of regional planning in China (Liu et al. 2011).

According to differences in geographical scale, regional planning in China can be divided into national level, cross-provincial, provincial, and city level. National-level regional planning comprises regional plans that have national significance and are officially approved by the State Council, and include the Northeast Region Revitalization Plan. Regional planning at a cross-provincial level deals with planning regions that cross provincial boundaries of which regional development planning in the Chengdu-Chongqing Economic Zone is an example.

In China, regional plans are compiled and authorized by three different ministries. The National Development and Reform Commission is in charge of the Major Function-Oriented Zone that is incorporated into China’s twelfth Five-Year Plan (Fan et al. 2012) and most cross-provincial regional planning. The Major Function-Oriented Zone strategy, for example, is designed to foster a development model that does not emphasize economic growth at the expense of resources and the environment, by identifying the unique functions that each region should perform in the light of its characteristics, conditions, assets, and requirements. It classifies regions into four categories: development-restricted zones for maintaining food and ecological safety; development-optimized zones for transforming the current model of economic development; development-prioritized zones for accelerating industrialization and urbanization; and development-prohibited zones for protecting the natural and cultural heritage.

The Ministry of Land and Resources is responsible for national- and provincial-level territorial planning that focuses on spatial organization and territorial security (Dunford and Liu 2015). The Ministry of Housing and Urban-Rural Development takes charge of urban system planning, urban agglomeration planning, and
REGIONAL PLANNING: CHINA

metropolitan region planning. In recent years, territorial planning, major function-oriented zoning, and urban agglomeration planning have become major regional planning research foci in China. However, there has not been a uniform regional planning system in China up to now.

SEE ALSO: Economic geography; Geography in higher education; Public policy

References


Further reading

Regional planning: the resilience of an imperative

Richard J. Nunes
University of Reading, UK

What is regional planning?

A strength of the idea of regional planning, and why its importance has featured in several seminal texts on planning thought for nearly a century, can be attributed to the fact that this form of planning practice has not been constrained by a single concept of “the region.” Regional planning can address any single issue or interrelated set of supraurban issues that arises, which affected communities may wish to engage. This general understanding of regional planning thus often results in varying or oftentimes overlapping constructs of a region. In other words, regions can be defined in both prescriptive and descriptive terms while addressing a regional problem and aiding other forms and/or levels of planning practice. These distinctions have been and continue to be essential to ongoing debates over the importance of regional planning and the meaning of “the region.” Benton MacKaye is among one of the earliest known writers to ask what regional planning is. MacKaye claims that “regional planning is best defined by splitting the term in two. What is planning? and what is region?” (1940, 350). The region, according to him, is “more than an area, it is an area or seat of movement” (MacKaye 1940, 350; emphasis added). MacKaye’s stress on movement equates the region to a “sphere” or space of flows, from the flow of water in a watershed to the flow of commodities (such as milk or financial services) and population. In other words, if the planner were to concern herself with the flow of a river, the range of that water-flow or its watershed would be the region to be planned – not an area of land restricted to the boundaries of an administrative jurisdiction.

However, MacKaye also knew all too well that both of these notions were “still in the making. They involve not merely what the dictionary states but what it is that planners really mean [by the region, by regional planning]” (MacKaye 1940, 349). Sure enough, notable scholars have periodically revisited this question and reassessed the idea of the region, and altogether this has amounted to a resilience of the “regional imperative”:

By defining regional planning as being most commonly a process arising from tensions and gaps within systems of governance, it [the regional imperative] will always be with us. So much of regional planning arises because of cross-boundary issues and tensions inevitable with any pattern of governance, regardless of whether or not it matches geographical regions. (Wannop 1995, 403)

This resilience of the regional imperative will be explored throughout the following reflection on regional planning. Before elaborating on any definitions, it is best first to situate regional planning within a framework of what planning is. A discussion of the abundant individual definitions of planning is well beyond the scope of this entry. Nevertheless, there are qualities of planning practice that do feature in different types of planning, from identifying a problem, projecting its
future socioeconomic and environmental ramifications, and generating or evaluating a set of alternative courses of action in the form of a policy statement, strategy, or definitive plan. The practice of planning is not, however, restricted to its regulatory fold or its formal processes. There is also “a presumption that planning is being done” in grassroots community-led development activities, as well as in the work of government agencies and of private sector engagement in urban regeneration and subnational economic governance:

Primarily a way of thinking about social and economic problems, planning is oriented predominantly toward the future, is deeply concerned with the relation of goals to collective decisions, and strives for comprehensiveness in policy and program. Wherever these modes of thought are applied, there is a presumption that planning is being done. (Friedmann 1963, 169)

As already mentioned, what constitutes the practice of regional planning, and what it means to plan at the regional level, have been the subject of debate over several decades in academic and policy circles alike. This period has been marked by distinct shifts in regional planning thought and practice, which have sought to balance the principles of regional planning on the one hand, and the intra- and interlevel efficiency demands of supraurban governance on the other. Glasson (1974) outlines four sets of overlapping distinctions of general planning practice against which regional planning and its associated complexities and cross-boundary tensions can be understood, if not defined.

The first draws a distinction between physical and economic planning. Whereas the former has its origins in land use and development vis-à-vis direct regulatory controls, the latter operates through market mechanisms to address concerns with the economic structure of an area. The second distinction is between allocative and innovative planning, which are divided according to function and area of concern respectively. Allocative planning is concerned with ensuring the efficiencies of an existing system, such as housing delivery, in accordance with changing policies. Innovative planning moves beyond efficiency measures toward the betterment of the system as a whole, such as affordable low-carbon housing developments. The third and related distinction concerns single or multi-objective planning. The previous housing example illustrates this distinction whereby the delivery of affordable low-carbon housing can be explicitly evaluated or measured against the multiple objectives of a carbon budget within a local climate-change strategy, and affordable housing targets across a period of projected housing supply need. The final distinction, between indicative and imperative planning, relates to the method of implementation. Indicative planning is advisory in nature, whereas imperative planning works from legally binding directives.

These interwoven distinctions offer a framework for comprehending the complexities of what constitutes regional planning as a form of planning practice. Following this planning typology, Glasson defines regional planning as “both physical and economic planning. Some regional plans may be purely allocative, but the majority includes certain innovative elements … In addition, regional planning is invariably multi-objective, but the method of implementation may vary greatly” (1974, 21; emphasis original). Regional planning is a geopolitical activity, and therefore it cannot be detached from issues of governance and democratic engagement. However, perhaps more importantly, this definition broadly maintains that regional planning practice equates to a presumption that planning is done both within and/or outside the regulatory fold of planning policies, plans, and politics. In other words, regional planning practice is not restricted
to formal regulatory processes over a designated physical area. MacKaye’s early writings underpin this view.

As mentioned earlier, these definitions reflect ongoing academic considerations of the region as a space of flows. MacKaye defines regional planning as a “comprehensive ordering or visualization of the possible or potential movement, activity, or flow (from sources onward) of waters, commodities, or population, within a defined area or sphere, for the purpose of laying therein the physical basis for the ‘good life’ or optimum human living” (1940, 351; emphasis added). Where Glasson’s definition sets out a framework for evaluating regional planning practice and its associated complexities and cross-boundary tensions, MacKaye leaves us with the originating principle behind a presumption of what it means to practice regional planning.

Principal dimensions of regional planning and their interrelations

Regional planning practice consists of the formulation and articulation of local and national objectives into a strategically guided ordering of activities or interventions in a supraurban space. It is this space of interactions that gives rise to regional planning, regardless of whether it constitutes a formalized set of activities. As Wannop (1995, 403) rightly points out, “regional planning arises because of cross-boundary issues and tensions inevitable with any pattern of governance, regardless of whether or not it matches geographical regions.” Regional planning, as a form of planning practice, often has a specific method of procedure attributed to it (procedural planning theory), which is usually initiated by the state. Regional planning will also draw on a specific concept of development (regional planning doctrine) and a variety of theories from the social and environmental sciences (substantive theory in regional planning), notwithstanding competing notions of what constitutes the planned region or territory. Altogether, the planning process, concepts of development, and substantive theories are underpinned by ideological assumptions. These assumptions are set within existing socio-economic, political, and spatial organizations of societies, in turn shaping the contents of regional planning and determining its outcomes. In other words, these ideological assumptions shape why we plan, and to a lesser extent how we go about doing it. In Territory and Function: The Evolution of Regional Planning, Friedmann and Weaver (1979) carefully outlined these points in what largely remains to this day a concise overview of the principal dimensions of regional planning and its interrelations (Figure 1). The next section provides a concise overview of paradigmatic shifts in regional planning doctrine, charting its early advancement through to its subsequent fall and rise, and recent calls for its reconfiguration or transformation.

Epochs in the evolution of regional planning doctrine, 1925 to present

The French geographer Paul Vidal de la Blache provides a useful starting point for a journey into the transformation of regional planning doctrine. Vidal was concerned with everyday life (genres de vie) in a region and its links to socio-economic, ideological, and cognitive dimensions of practices. This Vidalienne tradition of research through regional monographs lies at the roots of what may typically be referred to as the studies of regionalism today.

The narrative of regional planning thought is one of a long and rich history of transformation. It begins at a distinct break in the advancement of regional planning thought and practice in the late...
1970s: “The broad field of development studies of which regional planning is a part, is currently in the throes of a profound transformation of its own which is rendering much of the received planning doctrine obsolete” (Friedmann and Weaver 1979, 2–3). Prior to this “fall” in regional planning was one the most academically rich and experimental periods in the evolution of regional planning doctrine. Fast-forward from this breaking point nearly 40 years, and regional planning is once again the focus of global debate (Hadjimichalis and Hudson 2014; Jones and Paasi 2013). During this period of nearly a century, the presumption of what it means to practice regional planning evolves into a tension between degrees of territorial and functional integration. Table 1 outlines four distinct epochs in this evolution, adapting and expanding on the work of Friedmann and Weaver (1979) to accommodate subsequent developments in regional planning thought and practice.

The overview that follows is structured slightly differently. The first and second sections discuss the early advancement of regional planning thought and practice. The first stresses the territorial integration of natural capital and economic activity within a designated geographical region, followed by later thinking that increasingly concerned itself with the economic globalization of goods and services and its functional integration with regional (economic) development. The third section considers the departure from much of this received regional planning doctrine until the mid-1970s before its resurgence in the 1990s. A final closing reflection addresses recent calls for a repoliticization of this resurgent urban regional doctrine of innovation, competitiveness, and creativity in the face of global crises experienced locally in cities worldwide.

**Territorial integration**

The early thinking, nominally referred to as “territorial integration,” was geared toward physical needs such as the integrated delivery of environmental conservation, housing, jobs, and critical infrastructure, and was thus geographically bounded by them. This resource development doctrine, internationally regarded as “comprehensive river basin development,” was propagated by the work of the Tennessee Valley Authority (TVA) in the United States (Figure 2). As the practical idealism of regional planners confronted the mundane concerns of attracting private sector investment and guiding future economic growth, their thinking increasingly turned to functional urban regions and their integration with the economic globalization of goods and services. Examples of territorial
Table 1  Epochs in the evolution of regional planning doctrine, 1925 to present.

<table>
<thead>
<tr>
<th>Epoch</th>
<th>Description</th>
<th>Integration Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1925–1935</td>
<td>Utopian planning: biosynthesis and a new culture; cultural regionalism</td>
<td>Territorial integration</td>
</tr>
<tr>
<td></td>
<td>Practical idealism: comprehensive river basis development</td>
<td></td>
</tr>
<tr>
<td>1950–1975</td>
<td>Spatial systems planning I:</td>
<td>Functional integration</td>
</tr>
<tr>
<td></td>
<td>1 spatial development in newly industrializing countries (growth centers);</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 backward regions in industrially advanced countries.</td>
<td></td>
</tr>
<tr>
<td>1975–2000</td>
<td>Selective regional closure: Metropolitan growth and the rise of the “global city-region”</td>
<td>Fall and rise of regional planning</td>
</tr>
<tr>
<td>2000–2015</td>
<td>Spatial systems planning II (information systems + social systems planning – smart cities/regions):</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 high-tech urban development in newly industrializing countries;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 retrofitting city-regions in industrially advanced countries.</td>
<td></td>
</tr>
<tr>
<td>2015 to present</td>
<td>“Redistribution recognition dilemma”</td>
<td>Repoliticization of regional planning doctrine</td>
</tr>
</tbody>
</table>

Source: Adapted from Friedmann (1979, 8); reproduced by permission of John Friedmann.

Integration today are enshrined in watershed- and ecosystems-based approaches, which are increasingly finding functional rural–urban linkages to multiple urban systems, such as the nexus of food, energy, and water.

**Functional integration**

Regional economic development is no longer (since the 1950s) preconceived as bounded to a physical locality. A scientific base that underpinned theories of urbanization, industrial location, and inter-city ties supported this new regional planning doctrine (also known as “regional science”), leaving regional planning practice to follow as a newly established “scientific endeavor.” This transformation in regional planning doctrine was bifurcated. One dimension of these regional policies stressed the spatial organization of urbanization and industrialization through “growth centers” in newly industrializing countries by means of public subsidies to private enterprise and the use of public investment as the main instrument of spatial policy. Peri-urban or hinterland development (including rural development) would be incorporated through the centripetal forces of economic diffusion from core to periphery. The second dimension stressed uneven development in industrially advanced nations,
or “backward” or “less developed” regions. Regional policy with regard to the latter regions remained ambivalent and disproportionately focused on core regions, and the emphasis on the regional increasingly merged with the concept of metropolitan planning. Much of the regional planning of this period was associated with the creation of regions for the coordination of state–local actions. In Europe, this involved European-level strategic frameworks and directives down to national strategic frameworks at the member state level and its delivery through local plans. The French *amenagement du territoire* and German land-use planning traditions embodied these policy developments, later challenging the presumptions of what it meant to *practice* regional planning among the newly acceded member states of Central and Eastern Europe in 2004 and 2007 (Adams, Giancarlo, and Nunes 2011). Local government reform in England in the 1990s similarly reflected the same, including the creation of government offices for the regions (1994), regional chambers (1998–1999), regional development agencies (1998–1999), the Regional Coordination Unit at the Office of the Deputy Prime Minister (ODPM) (2000), and the elected London Assembly and Mayor (2000).

**Fall and rise of regional planning**

Public uneasiness surrounding events such as the student protests of 1968 and the 1973 oil and economic crisis inadvertently spread to the planning profession. “Growth centres did not
grow, backward regions did not flourish, poverty continued to accumulate in cities, inequalities remained engrained as deeply as ever in the landscape” (Friedmann and Weaver 1979, 7). In the years that followed and for nearly two decades, there was a marked disaffection with regional economic development and planning, before a resurgence of interest in regional development policy. This disaffection was part of an ideological shift as planning was seen as interventionist, opposing market-led approaches to development shaped by planning. The resurgence that followed in the 1990s and 2000s included a re-engagement of the functional integration of economic growth and development through a renewed interest in the role of relational assets in global economic restructuring processes and the emergence of transborder regional economies (Krugman 1991; Storper 1997). It was a process that led to the creation of planning regions in Europe, for example, and the institutionalization of European spatial planning (Waterhout 2008).

This period also coincided with growing experimentation in the practice of regional planning and policy, a rethinking of the notion of regions and a re-examination of the idea of territorial development. Principal examples of the application of this thinking include the European Commission’s (1999) European Spatial Development Perspective and the World Bank’s World Development Report 2009: Reshaping Economic Geography (Gill 2009), both of which captured academic attention on a global level. Notwithstanding the critiques of their apolitical stance on development, both reports frame nearly a decade of political and academic thought on the spatial organization of commodities and labor flows within city-regional systems and across polycentric megaregions, and its integration with the restructuring of the global space economy (cf. the “European global-macro region” in Pain, Richard, and Van Hamme 2014; and the “global city-region” in Scott et al. 2001).

Repoliticization of integration

Where the previous period ended and gave rise to the fourth and last of the epochs in the evolution of regional planning is unclear. However, there has been an emergence of the use of information and communication technologies (ICTs), which are ubiquitously integrated into the design and planning of new cities and the ecological retrofit of others. This trend in smart city-regionalism reinforces the idea of the region as a space of flows. It also echoes the aforementioned ideas of regional science at the tail end of the advancement of regional planning thought and practice (1950–1975), while at the same time engaging decentralized processes of data gathering (e.g., citizen science). Smart city-regionalism has accompanied an enthusiastic growth in the suite of microtechnologies, apps, and urban-scale operating systems. These technological advances may well see the emergence of new norms and routines which reframe the presumptions of regional planning practice. The continuing stress on the space of flows of human labor, environmental resources, and financial investment is evident in this regional doctrine. Yet this emergent doctrine, the practice of which has been most evident in its incorporation within spatial planning processes and its contribution to urban management, sits alongside enduring questions of uneven development uneasy.

It is manifest in new high-tech developments and informal settlements within rapidly industrializing “global regions,” and in the retrofit of global city-regions within advanced industrialized economies. Examples include newly built cities such as Songdo (South Korea), and the retrofit of others such as Rio de Janeiro (Brazil). In some respects, this urban response is not far
removed from the systems approach to urban and regional planning (McLoughlin 1969) where planning practice is explicitly more rational, autonomous and, “scientific,” or value-free, ignoring the politics of planning, sources of change, and conflicts between actors. It is closely aligned with global financial investment in new ICT markets or with the commercialization of new ICT products and services, and with the marketization of associated city-regional planning services. Its practice invokes visions of urban utopias and intelligent urban responses to global urban and environmental change, but in what sense are its motivations addressing a “regional problem” (Massey 1979)?

The apparent global race to delivering on smart, or “intelligent,” cities partly reflects the global capital investments in advanced producer services in ICT, the attraction of reduced expenditure on public services, and a desire to re-empower underresourced public authorities. It is an example of resurgent urban regional innovation, competitiveness, and creativity in the face of global crises experienced locally in cities. It echoes the “Third Way thinking” that Hadjimichalis and Hudson (2014) attribute to the schools of thought in new economic geography and new regionalism in the 1990s and early 2000s. Both schools championed the entrepreneurialism, competitiveness, and labor flexibilities of “model regions” in Europe and in other global regions. Under the economic pressures of regional divestment and new global competition that paralleled political economic adjustments to the formation of the eurozone, new theoretical approaches and local and regional development policies were developed around “learning regions,” “regional innovation systems,” social capital, trust, and reciprocity. The ways in which the schools theoretically responded was (at best) de-politicized at a time when what was needed was a frontal attack against neoliberalism. It is unclear as to whether this de-politicization was deliberate or an inadvertent and unintended effect because policy implementation based on these theories is blind to their effects on socio-spatial inequality. (Hadjimichalis and Hudson 2014, 212)

In Territory and Function: The Evolution of Regional Planning, Friedmann and Weaver (1979, 2–3) foreshadow the fall of regional planning with the bold claim that “the broad field of development studies, of which regional planning is a part, is currently in the throes of a profound transformation of its own which is rendering much of the received planning doctrine obsolete.” Nearly 40 years later, in a critical retrospective look at the rise of a regional planning doctrine at the turn of the century, Hadjimichalis and Hudson (2014, 208) point to a time (now) that “is ripe for a paradigm shift in theory and that this should involve a reconsideration of earlier theoretical approaches that fell out of fashion for a variety of intellectual and political reasons, and of current radical social movements.” While both sets of authors suggest a break with regional doctrine, the latter is aimed at a repoliticization of the consensus that has come to characterize regional planning doctrine at the turn of the century and onto which the ethos of smart city–regionalism has been grafted. Hadjimichalis and Hudson (2014, 215) argue for a double paradigm shift: back to earlier political economy paradigms that fell out of political and intellectual fashion, although without repeating the mistakes of the past, particularly those related to clientelism and bureaucratic statism; the other is a step forward integrating lessons from emancipatory grassroots social movements and social struggles across Europe and beyond.

Against these paradigmatic shifts in regional planning doctrine – from its early advancement,
through its subsequent fall and rise, to recent calls for its reconfiguration or transformation, the regional imperative persists: “By defining regional planning as being most commonly a process arising from tensions and gaps within systems of governance, it [the regional imperative] will always be with us” (Wannop 1995, 403). Likewise, “the search for some timeless and universal best planning practice is a quest for the Holy Grail” (McLoughlin 1992, 283).

**Future directions for regional planning research and practice**

Following Hadjimichalis and Hudson (2014), future directions for research and practice could consider urban and regional planning under global crisis. Crisis rhetoric has been couched in the modish take-up of delivering resilience to global risks to infrastructure, economic growth, and health and wellbeing. These challenges or crises are often attributed to population growth, global finance, terrorism, coastal flooding, food, energy and water shortages, affordable housing, and pandemics. Reference to such crises as the “new normal” suggests that cities, regions, and nations alike may be in a constant state of struggle or conflict with processes of global environmental change.

On the one hand, historical-materialist considerations link such crisis scenarios to societal modes of production, or a society’s ability to produce and reproduce the means of its own existence, including the institutionalization of environmental injustices as manifested in class struggle and different ways of thinking which are reflected in contemporary planned economic activity. This context subjects the state apparatus of traditional planning practice to institutional pressures of social movements, civic contestation, and an ever more pronounced “legitimacy crisis” (Habermas 1976). On the other hand, the permanent place of crisis and conflict is variably embraced by businesses, communities, and individuals, yet aimed at accepting and redirecting everyday experiences of crisis positively.

However, whereas an acceptance of crisis as a constant in everyday urban life may prompt positive collective responses from businesses, communities, and individuals to global crisis, the extent to which these “resilience dividends” (Rodin 2014) are able to address the structural challenges of uneven geographical development and sociospatial and environmental justice becomes ever more pressing.

Future regional planning research and practice would help to advance new conceptualizations of the interconnections between insurgencies and radical-subversive planning, cities and global crisis. Potential themes would include but not be limited to (i) the limits of traditional planning practice and/or cultural change and flexibility in planning under global crises; (ii) the legitimacy of urban regional planning; (iii) theories of change and stability in planning thought, positioning planning as critical theory and praxis globally; and (iv) the place(s) for radical-subversive planning beyond the boundaries of professional planning practice and planning laws and regulations.

**SEE ALSO:** Multilevel governance; Polycentricity; Regional definition and classification; Regional development policies; Regional geography; Regional planning: China; River basin management and development; Uneven regional development

**References**


Further reading


Regional political movements

Michael Keating
University of Aberdeen, UK

Nationalization thesis

Modern political science has usually downplayed the importance of territory as a feature of political life. It has tended to a state-based teleology, based on the idea that history is a progress towards national integration in which function, politics, and institutions align on the same territorial basis. Territory is seen as an alternative basis for politics from class, sector or function, so as these last become important, territory must become less so. This applies to party politics and electoral competition. According to the “nationalization thesis,” social structures, value systems, and loyalties are homogenized across national space and political alignments and party systems follow. Lipset and Rokkan (1967) showed how party systems and electoral cleavages in Western Europe configured along national lines as borders were closed and politics turned inward. These cleavages were then “frozen,” in spite of social and economic change. Later, however, Rokkan (1999) noted that national integration was often incomplete, with older cleavages surviving due to the uneven impact of economic and cultural integration. Since then, interest in the politics of territory has increased, with a substantial literature on regions and stateless nations. Yet Caramani (2004) has argued, with detailed evidence, that nationalization continued through the twentieth century. So, is territorial politics becoming more or less important? To answer this, it is necessary to disentangle the strands of the argument, distinguishing convergence of values from convergence of political behavior, and taking into account the role of parties themselves in shaping political space. The assumption that territory must give way to other forms of social and political differentiation must be relaxed. Instead, politics refracts and shapes these other factors in multiple ways.

The first point to note is that voting and party competition may converge even in the absence of social homogenization. Parties seeking power at national level will build state-wide coalitions, which may require compromises to meet particular regional conditions. Indeed, Caramani (2004) shows that homogenization of voting patterns occurred in the nineteenth and early twentieth century (for the left), before social homogenization. Some examples can illustrate this. The Italian Communist Party in the 1950s advanced in the south by gestures to land reform, peasant ownership, and autonomy, which it had previously rejected. In early twentieth century Wales, Labour took over the nonconformist and Welsh-speaking ethos from the Liberals. French Socialists penetrated Brittany in the 1970s by playing down the anticlericalism that marked the left in the south of France. Scottish conservatives forged an alliance of urban protestant populism, the urban middle bourgeoisie, and a deferential rural vote, quite different from English conservatism. Scandinavian agrarian parties enlarged their bases to become centre parties (Caramani 2004). Left-wing parties in Catalonia, the Basque...
REGIONAL POLITICAL MOVEMENTS

Country, Scotland, and the Hapsburg territories synthesized class and nationality politics in different ways.

Second, voting behavior can diverge in the absence of differentiation in values. Just as nationalization does not always indicate a deeper social homogenization, so the re-emergence of territorial politics might not stem from a reversal of homogenizing tendencies at the deeper level of values or social structures. Indeed, as with the nation, the region may be a space for the realization of universal values, whether of liberalism, the market, or welfare.

There is evidence that values converged across European territories from the mid-twentieth century, mainly as a result of secularization (Chauvel 2002). On the other hand, some territorial identities, value systems, and social practices persist over time. They have been detected across Europe (Todd 1990), in France (Le Bras 1995), in Poland (Baldersheim and Swianiewicz 2004), and in Italy. These are not mere legacies of premodern societies but have shaped modernization itself. It is known that nineteenth-century modernization was accompanied by a rediscovery of territory (Keating 1988) and a re-invention of tradition to make sense of it. Similarly, in the late twentieth and early twenty-first century, territorial identities and cultures have been rediscovered and modernized.

Late nineteenth century Europe saw an explosion of regional and nationality movements, with a diverse agenda. Some stressed defense of the church against the secularizing state. Some spoke for local communities against new forms of capitalist production and the peripheralization of former trading regions as the state imposed new boundaries. Others were based on language and others again supported a traditional social order against centralized government. The late twentieth century saw the rise of a new regionalism (Keating 1998), with a different agenda. One issue again concerned language and its social usage. Another driver was economic restructuring and competitive regionalism as states lost control of their spatial economies and “spatial Keynesianism” reached its limits. There were issues of welfare state restructuring and the displacement of older systems of territorial representation by modernizing states. European integration has further undermined state capacity for territorial management and opened up a new level of politics. This has posed a new agenda for territorial politics.

Economic disparities have provoked two types of territorial opposition. There is a “revolt of the poor” in regions that are marginalized in national economies and unsatisfied with the response. More important in recent years is a “revolt of the rich,” in which movements in wealthy regions complain about the burden of national solidarity. These distributional conflicts have become more serious in the context of the opening of national markets, which weakens the capacity of states to manage their territorial economies and puts an emphasis on interregional competition.

So the relationship between party politics and territorial societies is rather complex. Party competition is molded by territory but parties themselves both shape and respond to local demands (Caramani 2004). A territorial dimension has emerged to party systems and elections in the United Kingdom, Italy, Belgium, Germany, and Spain (Swenden and Maddens 2009). State-wide parties in territorially differentiated societies are caught in the middle, alternating between trying to homogenize politics and accommodating local differences. In Belgium, they have fallen apart altogether into territorial-linguistic components while in Italy, Spain, and the United Kingdom they manage the tensions more or less successfully.
Imagining regions

Regions cannot be taken for granted as political spaces but are discursively constructed by political actors themselves. This includes the definition of boundaries, both territorial and social, and the criteria of inclusion. Regions are, however, notoriously loosely defined spaces, without the clear citizenship-based definitions that states can use. At the popular level, conceptions of who belongs to the territory vary and are often based on overlapping categories, including birth, ancestry, assimilation, and will, without a clear hierarchy. Not only do the criteria vary but so does the way they are used. So language may be used as a marker of exclusion or a mechanism for inclusion and integration. Social citizenship may be extended to all residents, or restricted. Integration policies may also be linked to the needs of region or nation-building. So Scottish nationalists are inclusive and pro-immigration for economic and demographic reasons and to legitimize their project. Catalanists accept that integrating incomers is a necessity if the national project is to survive. By contrast, Italy’s Lega Nord has defined Padania against a series of “others,” be they southern Italians, non-European migrants, or muslims. Basque nationalism has moved from the criterion of bloodlines to a more inclusive definition of identity. Some movements in boundary regions seek to distance themselves from neighboring nationalisms by constructing a multi-ethnic or multicultural identity (Stjepanovic 2012).

The construction of the region often involves the attribution of specific values or stereotypical traits. Bavarian regionalists merge tradition and modernity as laptop und lederhosen. Catalanists stress the consensual and pactist traditions of their society, rather than the social conflict that marked early twentieth century Catalonia. Tuscan region-builders exalt the cooperative traditions of the region, the place where the catholic and communist traditions meet together around themes of civil society and subsidiarity. Tuscany and Lombardy vaunt their models of policy-making and service delivery as distinctive. Scots recycle myths of egalitarianism, while in Wales the dominant theme is community. In Brittany, shared identity and cooperation are taken as the markers of local society. In this way, values that are in themselves universal and transferrable can be pressed into the service of a particularist project.

Like nation-builders before them, region-builders need to link past, present and future, pressing history into use and giving themselves a modernist teleology, to combat the older, archaic visions of regional space. They have bought into the new economic regionalism, which presents the region as the space, par excellence, for economic modernization and dynamism. This, too, is projected into the past, as in Flemish region-building from the 1990s and as in Catalonia and the Basque Country (Ibarretxe 2010). The region is credited with the civic virtue that underpins social capital and with an ingrained entrepreneurialism (Keating, Loughlin, and Deschouwer 2003). In another search for legitimacy, regions are vested with democratic and liberal traditions. So in Brandenburg, official region-building has sought to reinterpret the Prussian legacy, rediscovering a liberal tradition. Traditions of popular rule and of pactism in Catalonia and the Basque Country have been opposed to Castilian authoritarianism. Neotraditionalist cultural movements in France and Spain are linked to ecology and opposition to global capitalism.

Territorial definitions of the region are contested. Catalonia and the Basque Country can be defined expansively, based on cultural and linguistic criteria (the Països Catalans and Euskal Herria) or administrative ones, based on their
REGIONAL POLITICAL MOVEMENTS

respective autonomous communities. Modern Flanders is constructed using linguistic criteria and historic materials to unite previously disparate historic units (Kerremans 1997). Bavarian regionalists have added a wider area into the historic community. The Lega Nord has invented Padania, which stretches from the far north of Italy to somewhere around the center. In practice it and its component leagues have opportunistically used many different spaces, tapping into localist and provincial identities often more than regional identities. Progressive political forces have deployed an alternative concept of North East Italy, drawing on European new regionalist ideas but sometimes invoking the Venetian Empire. As the Europe of the Regions idea developed, some regionalists claimed that their territory was quintessentially European. Some regions are imagined as cross-frontier zones. The concept of Middle Europe (Mitteleuropa) is deployed because of its territorial vagueness combined with its positive historical resonance as an alternative to the nationalizing state. Discursive constructions build on pre-state identities, often interpreted anachronistically to suggest precocious forms of statehood.

Regionalisms may challenge the state and merge into counternationalist movements. The difference between regionalism and nationalism is not a descriptive or even analytical one but has a strong normative content. Political movements that describe themselves as nationalist are making a claim for self-determination, possibly as a separate state but sometimes on a federal basis. At the start of the twentieth century, Catalan regionalists transformed into nationalists but not (mostly) separatists. Flemish regionalists sometimes call themselves nationalists and sometimes not. Some movements oscillate between defining their communities as nations or regions, depending on the context, as in Belgium, France, and Italy. These semantic games have a deeper purpose, in shaping the movement to fit with legitimating principles and, in the case of those consistently defining themselves as nationalists, putting themselves on the same normative plane as established nation states. Other regionalisms are deeply loyal to the state but to a particular form of it. So Bavarian regionalism has mostly been about a way of being German; French regionalism opposes the “Jacobin” conception of the state, not the French nation.

Territorial parties

Territorial parties vary both on the left–right and the autonomy axes. Regionalism in Germany and Austria has been associated with the political right, as it was in France until the 1960s when it was taken up by the new left and environmentalists. In Italy, it is historically associated with liberalism, although the left has blown hot and cold on it and Christian Democracy has also had a regionalist wing. Since the 1990s it has found its strongest expression in the rightist populism of the Lega Nord, which itself has to accommodate a poujadist, small business element and a working class vote that expanded rapidly over its years of expansion. Flemish nationalism tends to the right, with a weaker social democratic strand. In the Basque Country, Catalonia, and Galicia, regional nationalism covers the political spectrum except for the extreme right, as that is associated with Spanish nationalism. Nationalism in Scotland historically leans to the center-left but with a center-right tendency always present. Some territorial movements have produced catch-all parties, spanning the range of class and ideological appeals, and positing a common territorial interest.

An essential context for regional political movements since the 1990s has been the changing European and international order, which
may provide new opportunities for territorial autonomy outside the state framework. For separatists, European integration lowers the cost of secession, providing an external support system and dealing with externalities like market access, security, and borders. Others have embraced the Europe of the Regions, a concept that developed during the 1990s, seeking a role as a third level of government within a new political order. Some have been more radical and gone for “post-sovereignty,” arguing that traditional notions of sovereignty are outdated in modern Europe, where authority is shared at multiple levels.

Parties whose own values are closer to those of the European project, rooted in Christian democracy, social democracy or liberalism, have found it easier to use Europe than have extreme right, extreme left, or xenophobic parties. The Lega Nord was initially in favor of Europe, at one point proposing that northern Italy could join the Euro while the south could keep the lira. When it entered government with the post-Fascist Allianza Nazionale, however, it was expelled from the European Free Alliance of regionalist, nationalist, and Green parties. It then turned virulently anti-European. The Vlaams Belang (formerly Vlaams Blok), a xenophobic party of the populist right, was always hostile to the European project.

Adaptation to post-sovereignty ideology is easier in places that possess a historical legacy of mixed sovereignty and interdependence. Catalan nationalism harks back to the kingdom of Aragon, a confederation of Catalonia, Aragon, Valencia, and the Balearic Islands, within the kingdom of Spain, and to Catalonia’s history as a Mediterranean trading nation. Basque nationalists point to the fueros, historic rights of the Basque provinces within the Kingdom of Castile and then Spain, as a rationale for shared sovereignty with Spain as an alternative to the radical separatism of their founder Sabino Arana. In the early 2000s, first minister Juan José Ibarretxe elaborated a plan to make the Basque Country a “freely associated state,” linked with Spain and within a European framework (Keating and Bray 2006). Moderate Flemish nationalists place their project within a European support system, seeing the growth of Europe and of Flemish self-government as complementary.

The Europe of the Regions movement peaked during the 1990s but then stalled and made little progress in the Lisbon Treaty of 2009. This failure produced some disillusionment with Europe as a means to find a middle way and encouraged territorial parties to move towards independence pure and simple (Elias 2008), for example, in Plaid Cymru (Party of Wales) and among Flemish, Basque, and Catalan nationalists. When it comes to specifying their proposals, however, these movements tend to come back to varieties of sovereignty-association and partnership. Even the Scottish National Party, which is in favor of full independence, links this to the maintenance of various unions with the United Kingdom, including a currency union. In a similar way, Quebec nationalists have veered away from versions of sovereignty-association and back again.

Regional parties have been able to exploit the erosion of support for mainstream parties and crises of national party systems. The fall of Italy’s partitocrazia in the early 1990s provided an opportunity for the Lega Nord to revive the territorial dimension of politics. Gradual erosion of Britain’s two-party duopoly allowed for the emergence of the Scottish National Party and Plaid Cymru. Constitutional crises in Belgium led to the successive fragmentation of all the parties on linguistic lines. German unification left a serious territorial cleavage, with the eastern Länder giving disproportionate support to the Links Partei (Left Party), based on the old East
REGIONAL POLITICAL MOVEMENTS

German communists (Hough and Koß 2009). In some regions, such as the Aaland Islands, Greenland, and Northern Ireland, the state-wide party system has never taken root in the first place.

State-wide parties have adapted more or less effectively to the new territorial challenge. Conservative parties can draw on ideological reserves of traditionalism and suspicion of the state. Christian democrats can use the ideas of subsidiarity and autonomy, which are part of their tradition, but must reconcile these with state nationalism, also part of their ideological portfolio. Liberal parties often had local origins, in networks of local notables, and harbor suspicions of the centralized state, but they also have centralist traditions. Neoliberalism might, in principle, seem compatible with regional autonomy but, in practice, neoliberal parties have not been notable regionalists, probably because the region is seen as yet another layer of interventionist government. Social democratic parties have both centralist and decentralist traditions. In the postwar years of the Keynesian welfare state, they favored centralized economic management, top-down regional policies, and spatial redistribution of welfare. The left saw regionalism and nationalism as a distraction from the class struggle. Since the 1970s, social democrats have rediscovered regionalism and localism, and used them in their advance in local government to secure power at the center. Their decentralist zeal has not always survived their return to power at the national level. Trades unions and the left have been drawn into local struggles over plant closures. Social democrats have had to respond to challenges from territorial parties as well as the Green movement, which is instinctively decentralist.

So the territorial challenge has increased stress within state-wide parties. Some have adopted regional structures within their own organization. In others, the regional wing has a federal or confederal relationship with their state counterparts, as with the Catalan socialists and the Bavarian Christian Social Union. The Spanish Partido Popular had for some years to defer to the Unión del Pueblo Navarro as the representative of conservatism in Navarre. The Basque socialists have absorbed an older nationalist party. State-wide parties have given more or less autonomy to their territorial branches, provoking some considerable tensions in Spain and Britain, while in France and Italy the parties remain a centralizing force.

Regional political arenas

The electoral nationalization thesis has been linked to the concept of “second order elections,” in which voters use European or local and regional elections to reward or punishment national governments rather than voting on issues appropriate to that level (Reif and Schmitt 1980; Hough and Jeffery 2006). In Germany, France, Spain, and Italy, national opposition parties usually perform well in regional elections, which are then taken as a mid-term verdict on the national government. On the other hand, in Scotland, Wales, the Basque Country, and Catalonia, local nationalists do better in devolved than in state-wide elections, because they are seen to stand up for the territorial interest (Liñeira 2011). In France and Italy, where the electoral systems favor pre-electoral alliances, slightly different coalitions are put together for regional elections. In Spain outside the historic regions, there has been a proliferation of regional parties competing only in their own autonomous communities, and the regional frame of reference has strengthened over time.

There has been some personalization of party politics at the regional level. National parties have had to live with dissident but successful local
politicians, such as Ken Livingstone in London in 1999 or Francisco Alvarez Casco in Asturias in 2011. In Spain and Germany, party “barons” have carved out a margin of independence and have become brokers in state-level politics and candidates for national office. Local office is an important asset for French national politicians, as city mayor or, since the 1980s, as president of a regional council. Such leaders, lacking automatic sectoral or class support, will usually make a “catch-all” appeal, claiming to embody the interests the region, over and above sectional concerns. Promises of economic growth and competitive regionalism are effective ways to unite the territory, while pitching it against rivals. Such appeals both draw on, and serve to build, a territorial identity, while forcing other parties onto the same agenda.

Multilevel systems give parties the opportunity to play a differentiated political game by forming different alliances or coalitions at each level. The interdependence of the levels in Belgium and the existence of single political class (albeit linguistically divided) induced politicians there to maintain congruent coalitions, although this has begun to break down with the rise of the nationalist New Flemish Alliance. The main German parties experimented with all possible coalition combinations at the two levels (Detterbeck and Jeffery 2009). In Spain, territorial parties have an advantage because the electoral system is based on the provincial lists, favoring territorially concentrated ones. This means that territory and partisanship are always combined in the forming of governments. Non-state-wide parties have given parliamentary support to minority Spanish governments in return for concessions on autonomy and reciprocal support for minority governments in their home territory. At the same time, they try to balance out their alliances by playing off the main Spanish parties, which has often meant different parliamentary coalitions at the two levels. In France and Italy, regional coalitions have been formed with different mixes of parties but they respect the basic right–left division, with no governments bridging these poles. Devolution in Scotland and Wales has allowed for a variety of coalition arrangements centered on the Labour Party, with the Conservatives being the only significant party not to have participated in devolved government.

Rescaling politics

The emergence of regional political movements is part of a wider process in which functional systems, identities, political mobilization, and institutions are migrating to new territorial levels above and below the state (Keating 2013). These trends are sometimes portrayed as an outcome of economic imperatives and functional logic (Alesina and Spoloare 2003). In fact, such trends provide the context but do not yield a definitive outcome. It takes political movements to use these opportunities to create new political spaces and power bases. There are centrifugal trends as regional parties use regions as power bases and these, in turn, generate institutional interests in further autonomy. Yet these are balanced by centripetal forces as regional politics is played back into the national arena. The economic crisis has had contrasting effects. On the one hand, it has accentuated competitive regionalism as systems of fiscal equalization come under strain and movements in wealthier regions emphasize autonomy and the need to keep their own taxes. On the other hand, poorer regions look more to the center. So as Catalonia and Flanders press for more fiscal powers, their poorer compatriots seek to conserve the national framework and, in the case of some Spanish autonomous communities, even want
REGIONAL POLITICAL MOVEMENTS

to hand powers back. Similar territorial strains are observable in Germany, albeit without the element of separatism. Territory now frames debates around the size of the state, austerity, and welfare reform and will remain an integral part both of state and European politics.

SEE ALSO: Local development; Multilevel governance; National and regional integration; Regional definition and classification

References

REGIONAL POLITICAL MOVEMENTS


Regional science

Jean-Claude Thill
University of North Carolina at Charlotte, USA

Defining regional science and the essence of the region

The region has been a core concept of geography, and of many of its branches, since times that even precede its formal establishment as its own field of study. In geography, the region is usually seen as a land area with certain attributes and characteristics that differentiate it from others that are spatially separate. It has long been a unifying concept that used to define the very nature of geography in the minds of some scholars of the discipline, who saw it as the patent expression of the unique relationship between land and the physical environment on the one hand, and people and their agencies on the other. Today, the concept of region remains prominent in geography as it is widely used as an operational means to classify and categorize places based on the shared properties they exhibit. Regional geography itself is the study of places and spaces that constitute regions due to the unique characteristics of their human and natural elements.

Like geography, regional science studies the region. Yet, while regional science and geography share some common intellectual roots and remain closely intertwined fields of study, their modes of inquiry and the specific objects of their studies are different but largely complementary. Regional science is a social science that is concerned with the socioeconomic workings of human societies within the context of a region. The region is a geographic territory, but in the words of the founding father of this field of study, Walter Isard, it “is not merely an arbitrarily demarcated area; rather it is an area that is meaningful because of one or more problems associated with it which we as regional scientists want to examine and help solve” (1975, 1).

As a territorial construct, the meaningfulness of a region is not tied to an arbitrary size, but its size stems from the territorial scope of the problem at hand. For instance, the Washington, DC, metropolitan region defines the proper geographic confines for the understanding of traffic congestion in Washington, DC, in terms of the effects of infrastructure investment strategies, regional governance structures, commuter sheds, and land-use planning strategies. On the other hand, the dearth of trade between Southeast Asian nations is best analyzed for this set of countries as a whole to underscore their deficient fiscal and monetary policies. Finally, a small Midwestern town is the proper regional context in which to assess options for workforce retraining and business strategies after the loss of a major employer. Hence the designation of the region is problem-driven; it is fluid and changing according to the evolution of social, economic, and political conditions, and the study of its longitudinal transformation may itself be of interest as it reflects the evolution of structuring processes of the region. Importantly, the region can reflect and capture multiple scales. When drilling down to the basic agents that are collectively driving the dynamics of the region (such as the citizen, the customer, the patient, the worker, or the firm
or plant), regional science can recognize and analyze the individual locations where the agents are sited. These locations can be regarded as the most elemental regions. Socioeconomic structures exhibited by these regions, as well as their organization into systems, also define the object of regional science.

Each region is supported, and in a sense defined, by a distinctive socioeconomic organization that regulates its workings and its dynamics. Equally important, regions entertain relationships with others to form larger ensembles, or systems, that intersect in different and complex ways, including hierarchically. Sociocultural agencies display these organizational tendencies and private businesses have their own territorial footprint, both on the demand side and on the supply side. A tenet of regional science is that the region and its socioeconomic complexion result from the cumulative interaction between its own human, social, and natural resources and initiatives (local and intraregional processes), how these processes operate across the boundaries of regions (interregional processes), and how they relate with larger regional ensembles or the rest of the world (the global) (hierarchical or multiscalar processes). Some of these systemic processes may be institutionally driven (top down), while others emerge from the bottom up. It has been argued therefore that complexity theory is well suited to frame the dynamics of urban and regional systems (Batty 2005). When an entire regional system is studied as consisting of multiple territorial units, and these units belong to distinct national entities, the policy and political dimensions (including financial matters, national security, and so on) may take precedence over the socioeconomic considerations that are typically at the forefront of regional science research.

Distance is a key feature of socioeconomic structures in a regional context. As such it is also one of the foundational concepts of regional science. Distance is important to regions because it measures the dissipative effect on socioeconomic processes that operate between places and regional territories that are not co-located geographically. While ideas and information are exchanged between individuals and organizations at rates that depend on their own logic, a customary mitigating intervening factor is distance (or some other measure of effectual impedance of the geographic space) (Brown 1981). Social interactions that manage to survive over greater distance are usually looser than those between nearby nodes of a social network, even when this network entirely operates in a virtual space, as social media do. Social and institutional controls are weaker over areas remote from the social core of a region. Similarly, the magnitude of tangible flows of commodities and people (e.g., commuter flows) strongly responds to transportation or shipping costs, which vary in direct relation to distance.

From this brief introduction to the essence of regional science, it may be construed that geography and regional science are in fact one and the same. Their shared intellectual heritage is undeniable and many geography scholars are also scholars of regional science. However, regional science was established to articulate an interdisciplinary perspective on socioeconomic structures within regions and spaces. As such, regional science benefits from contributions from economics, sociology, engineering, planning, and other disciplinary fields. Points of departure between geography and regional science are more than a matter of detail as will become clear through the rest of this entry. Interestingly, geography’s own quantitative revolution occurred in the same years that the essence of regional science was crystallizing in the minds of a few key intellectual leaders (incidentally, some participating in both movements, such
as Edward Ullman, Michael Dacey, William Garrison, William Warntz, and others). In sum, regional science and geography, along with geographic information science, share a common core of concepts, frameworks, and methods, which are collectively subsumed by the umbrella term “spatial sciences.” It should be pointed out that spatial science is the name that was originally envisioned before regional science was settled on.

A brief history of regional science

Regional science was formally born in the United States in the 1950s at a very auspicious time. Coming out of World War II, the United States was positioned to take advantage of incipient economic growth opportunities stemming from new technological innovations, the opening of new markets, financial strength, and more proactive public initiatives in cities, transportation, social welfare, job training, and the energy sector. It became clear to young economist Walter Isard and others that established economic orthodoxy was ill-suited to handle these challenges as it was still wedded to the principle of socioeconomic systems devoid of any spatial extent and spatial context such as transportation and communication costs, and that it failed to recognize that the fortunes and misfortunes of human communities are tied to their internal social and economic logic, to their intrinsic factor endowment, and to the efficiency of interregional networks within national and international contexts. Thus, regional science emerged squarely as a product of its time.

The scientific roots of regional science can in fact be traced much further back in time, namely to the 1826 treatise The Isolated State, written by German economist Johann Heinrich von Thünen. This book provides an analytical argument for the existence of a pattern of crops and land uses according to distance factors, yields, and production costs through the mediation of the land rent. Thünen and his notable followers Wilhelm Launhardt, Alfred Weber, Walter Christaller, and August Lösch established the so-called German school of spatial economics or location theory, which addresses the factors affecting the locational choice of economic activity, including industrial plants, agricultural crops, urban land uses, cities, and so on. Ponsard’s (1958) account of the early history of regional science also recognizes the path-breaking contributions of Swedish economist Tord Palander and American economist Edgar M. Hoover to location theory in these pre-formative years.

The launch of the Regional Science Association (subsequently renamed Regional Science Association International (RSAI) in recognition of its broad international scope) in 1954 was instrumental in the emergence of regional science as a new social science distinct from economics, geography, and sociology. Today, RSAI has an active membership that exceeds 4500 scientists and practitioners distributed over all world regions, and continues to expand its reach in developing countries of Asia, Africa, and Latin America. The scholarship of regional science is primarily published in several core journals, including Papers in Regional Science, Journal of Regional Science, Regional Science Policy & Practice, Regional Science and Urban Economics, and Annals of Regional Science, as well as a number of allied journals such as Geographical Analysis, Growth and Change, Environment and Planning A, and Computers, Environment and Urban Systems. Owing to their extensive research on urban and regional socioeconomic systems, recent Nobel Prize Laureates Paul Krugman (2008) and Daniel McFadden (2000) have left lasting impacts on fundamental theories and methods of regional science.
Regional science in perspective

Regional science resolutely espouses a scientific approach to conceptualize research questions in their regional context and to articulate analytical frameworks that enable researchers to identify and evaluate business, planning, or policy solutions. Like neoclassical economic theory, regional science is based on a strong axiomatic logic that explicitly recognizes the role of various socioeconomic actors, often predicated on the assumption of full and perfect information dissemination, and economic rationality. In this framework, new residents seeking housing are expected to optimize their personal utility which is derived from the characteristics of residential options, once their contributory value to the potential resident has been suitably assessed. A broad set of characteristics (including locational and distance factors) would typically be considered as causal predictors in the individual decision process, and allowance is made for the inability of the analyst to observe or measure every characteristic of the choice situation by resorting to random utility models. Similarly, firms are assumed to be driven by profit maximization imperatives to structure production and distribution, including decisions on what to produce, where to produce, where to source inputs from, and what markets to serve from each facility. This scientific approach has produced an extensive corpus of theories that aim to understand socioeconomic processes at the regional scale and their interactions across regions, ranging from urban real estate markets to emission reduction through public transportation systems, and from innovation-driven economic growth to local taxation systems that establish the equilibrium self-sorting of human communities. Socioeconomic processes at the core of regional analysis often follow a behavioral paradigm that gives rise to the sorts of assumptions noted above.

Since its inception, regional science has been data-centric, aiming to empirically validate the theories and operational models developed analytically. Many primary data sources are available for this purpose, often resulting from governmental data collection initiatives. The decennial Population and Housing Census in the United States and similar initiatives in other countries have supplied critical data at fine levels of geographic granularity (census block, block group, or tract) to test spatial theories of spatial organization of cities such as social area analysis, gentrification and urban filtering theories, theories of real estate markets, transportation theories of urban commuting behavior, regional demographic models, and many others. Economic and labor censuses and annual surveys are staples of economic development models and theories such as the regional economic convergence theory and regional and interregional input-output models, theories on firm creation and innovation, and theories about entrepreneurial environments, human capital (workforce training, education), and metropolitan economic performance. In fact, the considerable demands that spatial data collection, manipulation, and management typically place on regional scientists have made them earlier adopters of geographic information systems (GIS) and technologies necessary to handle such data (Anselin and Getis 1992). Not only did these systems engender greater efficiency in the conduct of regional analysis, but the enhanced capability to do geospatial data processing afforded by GIS enriched the representation of spatial relationships among economic agents (like the substitution of economic and time impedances for simple contiguity-based spatial weight matrices (Gatrell 1984)) and ultimately increased the explanatory and forecasting power of regional models.
As pointed out earlier in this entry, Walter Isard’s drive to stamp a new theoretical paradigm of socioeconomic structures based on the concept of region was prompted by the pressing need to better understand the differential ability of place-based communities to cope with fast unfolding waves of social and economic restructuring in the post-World War II era. Regional science was born as a problem-driven discipline, and has remained so to this day. Early successes at alleviating social and economic hardship on communities at the local or more regional scale, and at formulating path-breaking public policies that leverage local human or natural resources and emerging market opportunities at broader geographic scales, have demonstrated the policy relevance of regional science. The same theories and methods of regional science have also been praised in business practice. In contrast to other perspectives such as regional development or planning, regional science brings to regional policy and practice a “sense of place” (Bolton and Jensen 1995) in conjunction with an understanding of how the place or region maintains socioeconomic relationships with others within a broader economic space through networks or functional flows. Thus, efficacious policies or business decisions account for both local and more global conditions. Edward Ullman (1954), a geographer and earlier advocate of the regional science perspective, relentlessly promoted the radical position that spatial organization and geography are created by processes of socioeconomic interaction between places and regions. Today’s regional practices and technologies have internalized this view.

Core research themes and theories

While the subject matter of regional science consists of socioeconomic realities as they occur in territories that collectively form a space, different research streams have classically focused on different topical areas, often blending different theoretical frameworks and calling for specific methods of analysis. We will briefly touch on five such research traditions in this entry and readers can find a more detailed overview in Fischer and Nijkamp (2014) and Mulligan (2014).

Environmental and resource analysis

The physical environment and natural resources are an integral part of the makeup of a region. As early as the 1960s, the inextricable interdependency between socioeconomic systems and ecological and other natural systems was recognized and explicitly incorporated in models of regional analysis, most notably in input-output models and computable general equilibrium models of regions and systems of regions (find more detail on these models in the “Methodological legacy” section later in this entry). Isard (1975) referred to them as “Combined Economic-Ecologic Systems,” which three decades later would be recast as “Coupled Human and Natural Systems” (CHANS) in other fields of study. The principle is to extend the framework of economic accounting of inputs and outputs of a region, which typically include wages, final consumption by households and government entities, and sales and purchases of intermediate products and services by businesses, to encompass ecologic commodities from and to the environment. A number of such commodities are already included in economic accounts to the extent that they are priced properly at the market. This approach is critically important in case of nonmarket transactions, which are so common in the context of the environment. It is related to a desire to evaluate unpriced effects when the market is patently failing and externalities exist. This would be the
case of pollution of the atmosphere, streams, and soils, resource depletion (i.e., forests and other ground cover, surface and subsurface water reserves, etc.), and others. It should be pointed out that externalities often exhibit a spatial dependency as spillover effects dissipate with distance from the source; environmental externalities manifest themselves within the territorial confines of regions, yet they can also affect interregional relationships. The combined economic-ecologic systems approach is well suited to the study of interdependencies at multiple spatial granularities, such as the localized impacts of global and interregional trade on CO₂ emissions and therefore climate change. It has been central to the articulation of the scientific discourse on resource-based limits to growth and of the principles of sustainable urban development.

In related research, regional science has studied the extent of the validity of a localized relationship between economic development and environmental quality. The environmental Kuznets curve postulates the existence of an inverted U-shaped relationship where environmental degradation tends to worsen as modern economic growth occurs, up to a point where the course of development enhances environmental quality. Evidence in this respect has been mixed, depending on the environmental metric, the region, the geographic scale, and the modalities of interregional commerce. The regional policy aspects of the environmental Kuznets curve are considerable as it may offer an expedient rationalization for delaying the adoption of tighter environmental protection standards.

Following conventional economic development wisdom, natural resources are regarded as a blessing to well-endowed regions. After all, abundant coal deposits were instrumental in the success of the Industrial Revolution in narrowly circumscribed regions of the United Kingdom and Belgium in the eighteenth and nineteenth centuries. Yet the idea that natural resources may instead be an economic curse gained traction in the 1990s. Widespread evidence now exists in China, Africa, the Middle East, and other parts of the world that regions with natural resource abundance experience poor economic growth. This points to two significant insights. First, space and the region matter for the understanding of socioeconomic structures. Second, these structures are complex and multifaceted and cannot be boiled down to environmental determinism.

Location analysis

As pointed out earlier, regional science grew out of the nineteenth- and early twentieth-century traditions of location theory. This theory underscores the role of transportation costs, along with other costs related to shipping from sources to production sites and then on to markets (transaction costs), on the locational equilibrium of economic activities. Although Cairncross predicted the death of distance in the 1990s, distance remains a significant force in locational decisions at the local, regional, and global scales (McCann and Shefer 2004), yet under notably different modalities than before. Nodal access to communication networks and mobility systems for human, social, and financial capital (such as airports) has emerged as a dominant locational factor in postmodern economies.

By explicitly recognizing the existence of transaction costs associated with travel and shipping across the geographic space, and the monotonicity of these costs with distance, location theory marks a dramatic departure from conventional economic analysis, which is largely framed in terms of perfect competition. In fact, location theory has demonstrated that, because of the geographic space, imperfect competition prevails and spatial competition often boils down
to spatial monopolies or oligopolies (Thisse 2011). Markets are intrinsically spatial and their geography exhibits many of the properties of other geographic phenomena, particularly spatial dependence.

Since its inception, location theory has expanded from explaining the spatial organization of single economic activities to explaining why firms in various industries and households co-locate in the geographic space. Christaller’s central place theory and the many extensions of it that have sprung up since the 1930s to account for the spacing, number, and functional complexity of urban centers within a certain territory have been among the influential theories of modern spatial sciences. Explanations for co-location or agglomeration are multiple and include factors from exogenous first-nature elements studied extensively in geography (political geography, physical geography) to endogenous second-nature factors. The latter include customer behaviors (e.g., multipurpose shopping), as well as supply-side considerations such as localization and urbanization economies that materialize only where some degree of agglomeration is achieved. In all these cases, geographic proximity is the catalyst to generate savings through travel cost reductions, knowledge spillovers, more effective labor markets, and other means (Rosenthal and Strange 2001). These are precisely the triggers of the emergence of regional growth poles (Isard 1975), which exploit the input-output linkages between industries. Growth pole theory and its more modern version, industrial cluster theory, continue to shape regional economic development strategies across the globe.

**Regional analysis**

Regional analysis focuses on three fundamental questions about regional economic growth (i.e., the rise in economic wellbeing of populations of a region). It seeks to understand the processes and factors of regional economic growth, the longitudinal fluctuations in the rates of growth exhibited by regions (e.g., what may lead to a region’s decline), and finally the geographic concentration of growth in rather confined territories (e.g., the spikiness of economic wellbeing across the world). Factors of regional economic growth can be sorted into elements that are internal to the region itself, such as natural resource endowment, capital formation, and labor or human capital, which foster endogenous growth, on the one hand, and factors related to the commercial relationships entertained by the region with other regions, on the other hand.

Early research in regional analysis was in line with economic base theory, which distinguishes basic industries, those exporting from the region and bringing wealth from outside, from nonbasic industries, whose function is to support basic industries locally. The split of total employment between basic and nonbasic is what determines the multiplier effect in the regional economy, and thus the rate of growth. Input-output analysis (Isard 1975) extends economic base theory by explicitly recognizing the interdependencies between industrial sectors of the regional economy through matrices of aggregate transactions in the economy, along with estimates of imports and final demands. As a structural macro-economic equilibrium model of the economic system, it is a regional accounting system that simulates the cycling of money through the economy over the short term and the long term. When multiple regions are distinguished, transactions between regions as well as within regions are recorded. Input-output models have fixed prices for factor inputs, which precludes the estimation of relative price elasticity of substitution between pairs of factor inputs and restricts their usefulness for policy analysis. The
more recent computable general equilibrium (CGE) framework generalizes input-output analysis by allowing price variation among inputs; the relative price elasticity of substitution is estimated econometrically. The input-output and CGE frameworks both allow estimation of the direct, indirect, and induced effects of an exogenous stimulus (e.g., new investment, change in fiscal regime, change in transportation costs, etc.) within and between regions, while the dynamic characteristic of the latter enables an emphasis on technology changes over the long run (up to 100 years). Both regional accounting frameworks are in line with the more theoretically grounded new economic geography model. Particularly noteworthy is the common emphasis on transportation costs and on increasing returns to scale as factors structuring the socioeconomic landscape of regional systems.

A complementary perspective on regional economic growth has privileged the importance of labor-related factors of production. Under the Fordist paradigm of industrial organization and socioeconomic production and reproduction, industrialization through mass production is the benchmark of economic efficacy. Given the highly standardized tasks expected of labor in this context, the flexible availability of labor trumps other considerations. Conversely, in postmodern knowledge-based economies, human capital and social capital held in the region constitute the most prized assets for regional economic success. Workers must possess the capabilities to learn and adopt new processes, organizational forms, and technologies. In addition, entrepreneurship and new knowledge creation through innovation and invention are prime conditions in order to stay ahead of the competition, which is decreasingly localized. Research has shown that regional competitiveness and lasting economic success are enhanced by investment to expand human capital (from secondary education to college, and beyond), strategies to nurture socially and culturally diverse societies, and the formation of creative class districts. Creative milieux have been found to exhibit higher quality of life for the enjoyment of residents and to be conducive to more effective knowledge spillovers through formal and informal social networks.

Urban analysis

The urban phenomenon occupies a special place in regional science scholarship as it usually structures a region in and of itself according to its own complex socioeconomic logic, while inserting urban regions in a broader functional framework spanning countries and continents. The modern history of humanity is increasingly framed by urban life and over 50% of the world population now lives in urban regions. Systems of cities frame social organizations, whether they are regarded as market places as in Christaller’s central place theory, or as nodes of thick trade and financial flows, housing critical command and control business functions, as in Sassen’s global city networks theory (2011). Urban regions have arguably become the focal points of cultural, social, and scientific creativity and innovation. Armed with the theoretical paradigm of the new economic geography, regional science is in a position to explain the very existence of cities on the basis of second-nature principles (Fujita and Thisse 2002), namely the circular and cumulative growth inherent to locales exposed to increasing returns brought by agglomeration. This opens the question of economic efficiency of cities and the possible existence of an optimal city size, which remains hotly debated today.

Internally, the city has been conceptualized as space over which various socioeconomic agents compete for the highest and best use in relation to requirements of access to the city
center. As discussed by William Alonso, location theory explains the concentric sorting of land uses regulated by the bidding between agents who trade off land rent and transportation costs. The emerging structure of the polycentric city has also been studied in location theory, and persistent spatial differentiations in land rent, quality of life, and access to labor markets and other urban amenities continue to challenge our understanding based on single theories such as urban gentrification or spatial filtering.

The growing concern garnered by sustainable human development in contemporary societies has brought closer scrutiny on cities and urban regions as they epitomize excessive consumerism, loss of ecosystem services, and debauchery of energy use. Regional scientists have been actively involved in studying the relationship between urban form and sustainability and the policies and practices that may enhance urban sustainability. In particular, restating the inseparability of land use and transportation, new integrated models rooted in micro-behavioral principles and aimed at providing urban residents with the requisite accessibility to employment and other urban amenities without sacrificing social and economic wellbeing have become available to inform the policymaking process in urban regions.

Transportation and mobility analysis

The scientific discourse on socioeconomic structures at the scale of the region would be rather stale and incomplete without attention to the ability, and occasionally the lack thereof, to physically connect places separated by some distance from one another. Without transportation and mobility, regions do not trade, stores do not sell, workers do not commute. Transportation, along with the modern technological offshoot of communication, enables interactions and flows of commodities, ideas, and people. It is important to stress, however, that transportation differs from all other factors of regional organization of socioeconomic systems in one critical respect: the benefits afforded by transportation in enabling mobility and creating accessibility are network effects stemming from the connectedness of the infrastructure.

The ease of transportation influences decisions made by households and businesses. Nodes on a transportation system as well as at the interface between distinct modal networks (e.g., seaports or airports, train hubs) experience disproportionately good accessibility, which makes them desirable locations for households and businesses. It is also in this context that location theory explains that the long-term trend towards lower transportation costs has facilitated interregional trade, expanded market areas, and promoted regional economic specialization as agglomeration forces have seen their importance grow in relative terms. A politically contentious topic is whether investment in transportation infrastructure promotes economic development. While macro-economically the evidence is there to support such an assertion, the conclusion is much more nuanced at the regional scale. While under certain circumstances, transportation infrastructure investment has high regional multiplier effects, in others it may facilitate outside competition and put local businesses and workers at risk, and ultimately lead to a spiral of job losses and economic decline.

Regional science has been at the forefront of the study of urban transportation. This includes the assessment of land rent effects of highway and public transportation systems, on the supply side, and the estimation of the modal trip generation under alternative scenarios (including modal shifts in commute trips), on the demand side. The externalities stemming from imbalances...
between the demand and supply of transportation services in the form of traffic congestion have been a focal point of attention in regional science. Network-based equilibrium models have served to assess the efficiency and social and spatial equity implications of creating a market for such distortions through a regime of congestion pricing.

Methodological legacy

Being problem driven and policy oriented, regional science has contributed a number of techniques of data analysis and modeling to the social sciences, and to spatial sciences in particular. The following list introduces a selection of some of the most meaningful of these contributions. Several of them were tangentially mentioned earlier in this entry. A more complete overview can be found in Fischer and Nijkamp (2014).

Input-output and computable general equilibrium (CGE) modeling – These related modeling frameworks are macro-economic models based on the principle of general equilibrium to capture the interdependencies of economic sectors of a region (in the case of a single region model), or among regions (in the case of a multiregion model). They serve to evaluate and forecast economic impacts of various social, economic, trade, and environmental policies. Input-output and CGE simulations may involve thousands of equations and are computationally intensive. With CGE models, the relative price elasticities of substitution are estimated econometrically.

Spatial econometrics – Regional socioeconomic systems entail multiple entities with explicit geographic positions. Methods of spatial econometrics entail confirmatory models that account explicitly for spatial dependence in system variables, which enables regional scientists to test for distance-based diffusion and spillover effects, such as knowledge spillovers. Another operational concept of spatial econometrics is that of spatial heterogeneity, which relates to the nonstationarity of causal relationships across the regional systems according to the ambient socioeconomic context.

Spatial interaction models – Spatial interaction models form a family of models designed to quantify the magnitude of flows of interaction between individual agents (e.g., households and stores) or aggregates of agents (e.g., countries, counties) in different geographic positions. Flows are expressed as a function of intrinsic characteristics of interacting agents and of their spatial separation (e.g., economic distance). Many spatial interaction models are nonlinear and recent extensions account for spatial dependence issues stemming from the spatial layout of interacting agents.

Spatial optimization – Normative models of operations research seek to optimize decision variables in the context of scarce resources. In regional science, optimization models may serve to benchmark decisions and proposals on the basis of objectively quantifiable metrics. Alternatively, they are core elements of computer-based spatial (regional and multiregional) decision support systems used by various public and business organizations. For instance, they may identify the allocation of scarce land resources to multiple urban uses to achieve policy goals (e.g., accommodate a 20% population and 25% employment growth over a decade), enhance the spatial efficiency of a school system by selecting school locations and capacities that meet projected enrollment growth across an urban area, and design routes of hazardous material shipments so as to minimize risk to populations. Such problems
are well addressed by linear and nonlinear programming methods.

Recent methodological advances in the study of regional systems have been marked by two dominant emerging trends. The first is the multiscalarity of the analysis, in that its explicit purpose has been to better apprehend the diversity of factors that interface with regional socioeconomic structures. These would range from the individual and household attitudes, preferences, and choices, to the contextual and institutional effects at the level of the neighborhood, city region, country, and beyond. Multilevel modeling allows the effects to be sorted out (thus also capturing spatial dependence effects) and complex covariance structures to be accounted for. The second trend is the reliance on increasingly computationally intensive approaches. The greater availability of data at diverse levels of granularity and the advances in computing and information processing have opened new opportunities in spatial data analytics that the regional science community has embraced. As a result, microsimulation techniques and agent-based models (ABM) have emerged as viable analytical vehicles for research in regional science. The convergence of better data, better analytical and computational methods, and more comprehensive theories of regional socioeconomic structures may be the dawn of a new era for operational regional models, building on earlier generations of integrated transportation and land-use models.

SEE ALSO: Local statistics and place-based analysis; Regional definition and classification; Representation; Representation: geographic systems; Spatial concepts; Spatial econometrics

References


Regional unemployment and regional labor markets

Nik Theodore
University of Illinois at Chicago, USA

Regional restructuring and employment change

Capitalist economies undergo the continual creation and destruction of jobs, leading to rising and falling levels of unemployment. New firms start up and existing firms expand, while old facilities are shuttered and some companies go out of business. These dynamics are exacerbated by cyclical swings in the economy, such as during recessions, but they cannot be reduced to the volatility of the business cycle. Rather, they are endemic to capitalism – and they produce profoundly uneven impacts across places, regions, territories, and scales.

Since the 1970s, economic geographers have developed and refined a distinctive approach to the study of regional labor markets and regional unemployment. The 1970s and 1980s were a period of tremendous economic uncertainty, and economic geographers in the United Kingdom and North America in particular, along with researchers in allied fields, sought to grapple with the dislocations that were gripping older industrial regions. Plant closures, mass layoffs, capital flight, and strike actions were widespread, clear signs that the established economic order was being radically unsettled. The tools of mainstream economics seemed to be increasingly unsuitable for the analysis of these phenomena; the discipline’s fixation with mathematical models and detached theorizing had exceeded its explanatory limits. A new approach was needed.

The analytical and methodological approaches developed by economic geographers, which evolved in contradistinction to the precepts of neoclassical economics, conceived of labor markets as being socially embedded, shaped by institutional forces and structured by power relations. A “market” like no other, the labor market does not adjust mechanistically, as described in economics textbooks, with price changes leading to the smooth balancing of supply and demand. Neither is labor simply just another “factor input” to the production process, one whose price – wages – will rise and fall depending on its availability. As David Harvey (1989, 19) put it, “The buying and selling of labor power deserves special scrutiny. Unlike other commodities, labor power has to go home every night and reproduce itself before coming back to work the next morning.” Economic geographers have in fact subjected capital-labor relations to sustained scrutiny, developing robust subfields such as labor geography, industrial geography, and global value chain analysis to analyze the shifting contours of capitalist economies.

This work has been guided by the maxim that economic restructuring – as well as labor market adjustment – is an inherently sociospatial process. Doreen Massey’s (1995, 3–4, italics in original) account in *Spatial Divisions of Labor: Social Structures and the Geography of Production*, originally published in the mid-1980s, succinctly summarized the evolving view from within economic geography: “New spatial divisions of labour (forms of uneven development) are
thorough re-workings of the social relations which construct economic space . . . . They are more than just new patterns of employment, a kind of geographical re-shuffling of the same old pack of cards. They represent whole new sets of relations between activities in different places, new spatial forms of social organisation, new dimensions of inequality and new relations of dominance and dependence. Each new spatial division of labour represents a real and thorough spatial restructuring. It marks a new form of regional problem; and more basically it marks not a new re-organisation of relations in space, but the creation of a new space.”

The differences in the approaches to the study of labor market change between economic geography and neoclassical economics perhaps could not be more stark. In contrast to the aspatial approach to studying the functioning of regional economies and labor markets found in neoclassical economics, which posits the notion that capitalist markets are self-regulating, that they tend towards a state of equilibrium, and that they lead to regional convergence over the long run, the emerging geographical perspective saw the dynamics of capitalist growth as perpetually disequilibrated and inherently prone to producing new patterns of uneven geographical development. Orthodox economic theory assumes that (i) the fundamental economic activity is exchange through competitive markets; (ii) exchange in free and unfettered markets will lead to the optimal allocation of resources; and (iii) the natural state of economic systems is a stable equilibrium. Geographical perspectives are founded on a critique of these assumptions.

Writing as the British economy was in the throes of a deep recession in the early 1980s, Doreen Massey and Richard Meegan’s (1982) *Anatomy of Job Loss: The How, Why and Where of Employment Decline* stressed that deindustrialization (the systematic reduction of industrial capacity in an area) produces geographically uneven patterns of decline. In articulating a geographical approach to the study of economic restructuring and employment loss, they urged researchers to probe beneath the broad statistical depictions of economic change that are commonly used to describe economic conditions – such as unemployment statistics, counts of business closures, and aggregate measures of income loss – and instead interrogate the underlying causes of the observed patterns. Based on research on more than 30 industries, Massey and Meegan identified three distinct forms of production reorganization that can result in substantial job losses: intensification, technical change, and rationalization. Intensification refers to a type of reorganization that aims to increase labor productivity without making major investments in capital and equipment. Intensification can be pursued through a variety of means, such as speedup, piecework, and the introduction of new forms of teamwork on the shop floor. Technical change occurs when significant investments are made in capital and equipment that result in alterations in the machines, techniques, and processes used in production. Rationalization refers to disinvestment that results in a loss of production capacity, usually due to a (relative) lack of profitability in an industry. “Rationalization … involves complete or partial plant closure, the scrapping of capital equipment and cutbacks in the labour force” (Massey and Meegan 1982, 87).

Crucially, these three causal drivers of employment loss differentially impact regional economies. Intensification, through tactics such as speedup, results in employment losses in situ; its impacts are concentrated in the places where production is currently located, and it does not imply a shift in the location of jobs. Rationalization, on the other hand, may result in changes in the distribution of employment if production
is selectively eliminated within a given industry, or if firms consolidate employment at certain facilities while at the same time others are shuttered. For regions that are reliant on a single industry or firm, or even those whose economic base is in a small number of industries, the impacts of rationalization can be devastating if plants close or if jobs are relocated to distant facilities, thereby gutting the local employment base. Finally, because technical change occurs through new investments in productive capacity, unless there is a commensurate increase in total output, these investments will lead to employment declines, either through the closing of less technologically advanced facilities or through in situ employment losses.

One of Massey and Meegan’s (1982, 88, 89, italics in original) key insights was that economic crises, ruptures in industrial production systems, and the decline of entire industries all are endemic to capitalist economies and not the result of exogenous shocks: “[E]xcess capacity … is continuously produced by the very nature of accumulation. […] There comes a point where too much capital is chasing too little profit. […] If accumulation is to continue then the capital structure must be reorganized.” In other words, the industry will be a target of restructuring, potentially with far-ranging effects on regional employment.

In *The Capitalist Imperative: Territory, Technology, and Industrial Growth*, Michael Storper and Richard Walker (1989) subsequently took up the question of how the nature of industrial restructuring impacts regional economies and the geographic location of economic activity. Like Massey and Megan, they sought to make “space … an active variable in the social system,” emphasizing the instability of cities and regions over time as processes of capitalist restructuring continually rearrange the geographical location of production as capitalists search for profit (9, 10). So too did they conceive of economic crises and spatial disequilibrium as being internal to capitalist economies, as opposed to being the products of exogenous shocks, as typically is portrayed in the orthodox economics literature.

Storper and Walker sought to advance a theory of capitalist growth and restructuring with a clear focus on their spatial expressions, one that, along with the penetrating analyses of David Harvey (1982, 1989), marked a further, forceful challenge to neoclassical economics. They placed industrial production and technical change at the heart of their analysis (as did Massey and Meegan (1982) and Massey (1995), though Storper and Walker as well as Harvey (1989) and Massey (1995) also emphasized the path-altering role of class conflict within capital–labor relations and its effects on industrial location), thereby displacing the primacy of exchange that is found within orthodox economics. They also challenged assumptions concerning the optimal allocation of resources that is said to occur through market mechanisms. And they argued, contra mainstream economics, that “industrialization, as linked to capitalist competition and the dynamics of capital accumulation, … generates a highly disequilibrated form of growth that repeatedly unhinges the existing economic order” (Storper and Walker 1989, 8). Centers of industry shift, driven by the “coercive laws of competition [that] force capitalists to search out superior technologies and locations,” producing a restless economic landscape that is shaped by cycles of boom and bust, of overaccumulation and devaluation (Harvey 1989, 136, 143). For these reasons, “equilibrium [in the sense posited by neoclassical economics] is always just out of reach” (Storper and Walker 1989, 38).

Storper and Walker then turned their attention to processes of territorial development. Like Massey (1995), they argued that through the
varied pathways of restructuring that are pursued within capitalist economies, industries produce their own geographies – in other words, they produce regions. “[I]ndustries create regional resources and not the other way around,” Storper and Walker (1989, 96) asserted, and it is in this sense that regions are constructed through socioeconomic processes (such as through flows of commodities, the mobilization of labor, and the exercise of power). Economic regions are not, in other words, naturally occurring areas that precede the actions of capital. Rather, “The spatial distribution of employment … can be interpreted as the outcome of the way in which production is organized over space” (Massey 1995, 65). But more than this, “‘space’ is not a passive surface on to which the relations of production are mapped, nor yet simply a negative constraint (in the sense, for instance, of distance to be crossed). […] In relation to production, spatial form and spatial strategy can be an active element of accumulation” (66).

Moreover, in this era of increasing international economic integration, leading firms have considerable latitude in selecting production locations and (re)organizing production over space, the scope of which has been greatly expanded with the emergence of globalizing production networks and their distended supply chains (Coe and Yeung 2015). Industries can make, remake, and mobilize space in the pursuit of profits. For example, successful industries are able to expand into formerly peripheral territories in the search for profits and market share, and lead firms are able to orchestrate supplier networks that span national borders and continental divides, often in the search for low labor costs. These processes in turn create greater regional differentiation – in labor costs, factors of production, power geometries, and so on – rather than the regional convergence posited by neoclassical economics. In short, they produce and rework patterns of uneven geographical development.

Regional unemployment

The processes of economic restructuring analyzed by economic geographers can produce upheavals in regional and local economies. Industrial reorganization; capital flight; layoffs due to intensification, technical change, and rationalization; and plant closures can undermine the employment base of regional economies, leading to economic decline and mass job loss. In studying these effects, economists commonly distinguish between three types of unemployment: frictional, structural, and cyclical. Frictional unemployment refers to the unemployment that occurs when job seekers move between posts, when students leave school to join the labor force, when workers switch careers, or when job seekers relocate to new areas. It arises because workers and employers have incomplete labor market information, creating time lags for workers who are trying to identify and secure job opportunities, and for employers who are trying to fill job vacancies. Employers must search among pools of job seekers to find an appropriate hire, and this takes time. Similarly, job seekers who are recently out of work usually must participate in multiple job interviews before accepting a position. Frictional unemployment is considered to be a feature of job-search and recruitment processes, and to some degree it is always present in the economy. Because it is taken as a sign of a dynamic labor market, and because it usually is short term, frictional unemployment is typically regarded as beneficial for the economy. It is a side effect associated with labor market dynamism and occupational mobility.

Structural unemployment is a second form of unemployment. It arises when there is a
mismatch between the type and skill requirements of job vacancies in the labor market and the qualifications of the unemployed. Under these circumstances, unemployed workers are unable to easily fill available vacancies. The mismatch might be the result of a shift in the jobs base caused by technological change which renders skill sets obsolete; the ratification of trade agreements that facilitate the relocation of industries, leaving former employees without jobs and without the prospect of fully utilizing their skills and experience; or an economy’s reliance on declining industries. Structural unemployment is a form of long-term unemployment, and it disproportionately impacts regions where economic restructuring has been greatest. The loss of manufacturing jobs in certain regional economies, for example, leaves former employees without work and unable to capitalize on skills that are no longer in high demand. Labor market adjustment and reductions in unemployment in the short to medium term are stymied because the supply and demand for certain job skills and qualifications are grossly out of balance. In depressed regions especially, less skilled and less educated job seekers may be greatly affected by structural unemployment if sectoral changes remove the jobs for which they might qualify from the local labor market. Though far less severe a problem, seasonal unemployment may also be regarded as a form of structural unemployment because economic activity in the local economy changes over the course of the year, leaving certain workers without appropriate employment. Some economists, however, consider seasonal unemployment a form of frictional unemployment because even if seasonal workers find alternative employment during the off-season there typically is a time lag before a job is obtained.

Cyclical unemployment, sometimes known as demand-deficient unemployment, is a third form of unemployment. It is associated with economic downturns – slowdowns and recessions – that occur as the economy contracts. Cyclical unemployment is caused by sharp declines in demand and it is a sign that aggregate demand for goods and services is insufficient to generate jobs for all those who want to work. When consumer and business demand falls, work orders are reduced, business revenues decline, and businesses respond by laying off workers in order to reduce costs and maintain profit levels. Unemployment rates can rise rapidly if demand plummets because spending across the entire economy can be affected, simultaneously impacting multiple industries. Under the assumptions of neoclassical economics, cyclical unemployment should be a short-term phenomenon. As the ranks of the unemployed swell, downward pressures on wages mount, making it possible for employers to hire on new workers. Keynesian economists, however, hold a different view. They contend that a rise in unemployment deprives the economy of consumer spending, creating a negative multiplier effect as this loss in spending leads to further job cuts and a corresponding rise in unemployment. Under this scenario, cyclical unemployment will be associated with high and sustained unemployment levels over longer time horizons.

During recessions, the composition of overall unemployment shifts. Cyclical unemployment rises because businesses lay off workers in response to falling demand. Frictional unemployment tends to fall because workers are reluctant to quit their jobs, fearful that they will have difficulty finding a new one. At the same time, there is a danger that structural unemployment will set in as firms feel the pressure to restructure production systems, integrate new technologies, or close their doors due to declining demand. Under these circumstances, the structurally unemployed could face extended
periods without work because their skills are no longer demanded by employers. When regional economies are beset by high levels of structural unemployment firms may be reluctant to bear the costs of retraining displaced workers, contributing to the mismatch between job seekers’ skills and the vacancies available in the labor market.

The long-run impacts of structural unemployment can be seen clearly in the coalfield areas of Britain, where employment in the mining industry plummeting from 220,000 jobs in 1985 to just 7,000 in 2004 (Beatty, Fothergill, and Powell 2007). Massive job losses in export-oriented industries (such as mining) set in motion additional employment declines in affected regions. As the multiplier effects of industrial decline took hold in these areas, the shocks of pit closures rippled through local economies, robbing them of income and spending that supports a wide range of economic activity. Employment losses led to declines in local consumer spending, which in turn undermined the locally serving sectors of the economy, which in turn led to further declines in local consumer spending, and so on. Rising unemployment followed, and when unchecked it led to rising long-term unemployment. The local economies of these areas were, by any measure, in decline (Beatty, Fothergill, and Powell 2007; Pike, Dawley, and Tomaney 2010; Theodore 2007). With little employer demand for their specific skill sets, and with local economies facing entrenched job deficits, miners found themselves structurally unemployed, most of them for the long term.

Employment statistics divide the labor market into three groups. The employed, the unemployed, and those who are not in the labor force. The unemployment rate typically counts only those job seekers who are actively seeking work. Therefore, especially during periods of high unemployment, official statistics tend to undercount the extent of joblessness in the economy. Job seekers are considered to be experiencing short-term unemployment if they have been out of work for fewer than 27 weeks, and they are considered to be experiencing long-term unemployment if the jobless spell is 27 weeks or more. As overall unemployment increases in a labor market so too does the incidence of long-term unemployment. During economic downturns, unemployment rises, increasing the number of job seekers at a time when vacancies are scarce. When job seekers experience extended periods out of work, employers grow more reluctant to hire them, reading lengthy unemployment spells as a sign of a lack of labor market qualifications or problems with their “employability.” Under these circumstances, weeks of unemployment can turn into months, or even longer. Some job seekers will stop their active search for work, frustrated by their lack of success in finding a job. They are referred to as discouraged workers. Workers may become discouraged when there is a shortage of job vacancies in their local labor market, or when they face employment discrimination on the basis of race, ethnicity, nationality, gender, religion, disability, or sexual orientation. Discouraged workers, who sometimes are referred to as the hidden unemployed, are not included in official unemployment statistics.

The 2007–2009 Great Recession provides a revealing window into various dimensions of unemployment. The recession had devastating, though socially and spatially uneven, impacts on job seekers in the United States. The economy shed more than 8.5 million jobs from the recession’s onset to the bottom of the cycle, losses that were disproportionately located in the manufacturing and construction sectors. As 2009 came to a close, approaching 16 million job seekers were officially unemployed, and
reports of worker discouragement dominated news headlines. Geographically, unemployment rates varied greatly state by state, ranging from 3.9% in North Dakota to 15.9% in Nevada. To make matters worse, the effects of the recession lingered. Approximately two years after the recession had been officially declared over, unemployment durations were at their historic highs, as was the share of the unemployed who were classified as long-term unemployed. Protracted periods of unemployment raised new concerns that deep structural unemployment was affecting large parts of the country, undermining the future job prospects of laid-off workers and weakening the job market for new entrants, particularly young people.

The social distribution of unemployment

In the mid-1990s, questions regarding the social structuration of labor markets, including which social groups bear disproportionate burdens of economic restructuring and job loss, were taken up in earnest by economic geographers. The publication of Gender, Work and Space by Susan Hanson and Geraldine Pratt (1995) and Work-Place: The Social Regulation of Labor Markets by Jamie Peck (1996) marked a theoretical, conceptual, and empirical advance in geographical analyses of regional labor markets. Building on the work of Massey, Harvey, and Storper and Walker, along with research by heterodox economists, Hanson and Pratt as well as Peck sought to “spatialize” heterodox labor market theories, in particular labor market segmentation theory. Labor market segmentation theory had been developed by economists who were dissatisfied with the predominant view within neoclassical economics that labor markets behaved like commodity markets. In opposition to theories in neoclassical economics that viewed the job market as a single, unified market in which (i) opportunities are ordered according to skills and other forms of human capital and (ii) wages correspond to human capital attributes, labor market segmentation theory proposed that the job market is comprised of noncompeting subgroups. In the broadest terms, the labor market could be conceived as having a “primary segment” where higher status, more secure, and better paid jobs could be found, and a “secondary segment” comprised of lower status, insecure, and poorly paid jobs. The allocation of workers to jobs, according to the theory, was largely based on ascribed characteristics, such as race and gender.

Economic geographers placed the local labor market at the heart of their analysis, conceiving it as far more than merely a catchment area or commuting shed for workers and employers, thereby elevating claims of its theoretical status. Peck (1996, 95, italics in original), for example, states that “it can be argued that all labor markets are ‘locally constituted’… . Although a broadly similar set of causal processes underpins the operation of every local labor market, each local labor market is unique in that it reflects a unique intersection of those processes. While all local labor markets are shaped by, say, gendered domestic labor or ethnic stratification and marginalization, ‘generic’ tendencies do not have universally even outcomes. […] Further, local labor market differences are magnified by the fact that these processes are not operating on a tabula rasa, but in their realization are affected by the inherited social, economic, and institutional geographies of the labor market.”

Building on labor market segmentation theory, these geographers sought to interrogate various forms of socioeconomic disadvantage and how these might be produced and reproduced through the operation of local labor markets. The labor market is not seen as a neutral medium
of exchange that balances adjustments in supply and demand. Instead, it is socially structured and structuring, consigning members of disadvantaged groups, such as racial minorities, to jobs in the secondary sector. There they endure low pay, little upward mobility, and frequent spells of unemployment.

The existence of a secondary sector provides a number of benefits to employers. For large and stable companies, access to workers in the secondary sector, for instance through subcontracting chains, is a way to boost profits by outsourcing low value-added work to lower cost suppliers. For more marginal companies, tapping labor pools in the secondary segments of the job market can be a core workforce strategy, and the difference between surviving in a competitive marketplace and going out of business. In both scenarios, firms rely on secondary labor markets to, among other things, achieve numerical flexibility. They hire workers when there is a need to increase output, and they lay them off when demand is sluggish. This means that job losses will be disproportionately concentrated within the secondary sector and unemployment will disproportionately affect workers who tend to be consigned to job opportunities there, in the process reinforcing their labor market disadvantage.

SEE ALSO: Deindustrialization; Gender, work, and employment; Industrial restructuring; Labor market; Labor market segmentation; Restructuring; Surplus labor; Unemployment and “underclass”; Uneven regional development

References


Further reading


Regionalism

John Tomaney
University College London, UK

Regionalism is typically understood by human geographers as the identification and mobilization of divisions formed in relation to subnational, socioeconomic, cultural, or political territories. Regionalism can refer to efforts to organize government action at the subnational scale (such as “metropolitan regionalism” in the United States). In cultural terms, regionalism can refer to ideas of belonging and attachment to a particular territory expressed, for instance, through art, music, food, and literature. Political regionalism denotes the mobilization of actors around territorial claims. If regions are social constructions, regionalism is the means by which they are fabricated by administrators, intellectuals, and politicians. But regionalism is also reproduced in the practices of everyday life, for instance, through the use of dialect and accent, in media discourse, and in distinctive social practices. Regionalism is contingent and constantly being formed and reformed, negotiated and manipulated, promoted and contested. (A less common usage, used to describe macroeconomic and political blocs mainly by political scientists, is not considered in this contribution.)

Regionalism typically presupposes a regional identity. Regional identities are not pre-given cultural phenomena, but are constituted through cultural and political ideas, social movements, and modes of governance. Regionalism is a means of articulating the relationship between people and place. Regionalism is a narrative activity concerned with defining and acting upon boundaries – region-making – through the mobilization of political, cultural, and socioeconomic claims (Paasi 1991). For instance, interpretation of the past is a key element in the political and cultural construction of a region, often involving the search for a “usable past” as a guide to regionalist praxis.

Regionalism is a performativediscourse that seeks to achieve legitimacy for definitions in cultural, political, popular, and official understandings. Regionalism is also concerned with giving meaning to bounded material and symbolic worlds in an effort to create intersubjective meanings. Regionalism is frequently linked with “topophilia,” that is, to affective bonds between people and place. Regionalism in art and literature often gives expression to these processes, creating imaginative boundaries around particular territories exhibited, for instance, through a concern with dialect and landscape, which are intimately linked through topological nomenclature. Contemporary regionalism is often closely related to rising ecological awareness and the growing interest in more diverse and sustainable forms of living. The defense of regional cultures can occur alongside struggles to preserve biodiversity. On occasion, ecological concerns and cultural claims have proved the basis for political regionalism, although there is no necessary relationship between these factors. A shared regional identity, in both its cognitive and affective dimensions, does not have inevitable political consequences, but depends on whether this identity is used as a framework for perception and judgment of interests (Keating 1998).

Regionalists are at times criticized for emphasizing a topographical understanding.
of cultural and political geography. Regionalism tends to imply a territorial understanding of the production of space. Critics maintain that contemporary life does not take place in bounded territories, but occurs relationally in networks stretched through and beyond regions (Amin 2004). But the persistence of regionalism as a cultural and political form reaffirms the importance of the making (and unmaking) of territorial boundaries. There is a strongly normative dimension to the critique of regionalism, which is seen as tending toward chauvinism and exclusion of “the Other.” Regionalism, moreover, is criticized for offering a common geographical origin as a basis for solidarity in a way that obscures intra-regional inequalities. Escaping the particularist claim of territorial identities, such as regionalism, is a theme of those who advocate international solidarities and global citizenship.

Such normative criticism begs the question of why regionalism persists. The contrast between relational and territorial understandings may be unhelpful in providing an answer. The topographical and topological are perhaps seen best as being in tension. The appeal of regionalism may lie in its offer of moral particularity as a means to negotiate global networks. Regionalism may provide a means of politically and culturally “being at home in the world”; a sense of being grounded in a particular society while being open to the gains that arise from greater global integration. Regionalism may represent the search for rootedness and belonging and the valorization of local distinctiveness in the face of the universalizing claims of globalization (Tomaney 2013).

In practice, states are often compelled to accommodate regions in their political structures. Regional actors influence patterns of state formation and nation-building projects. The process of territorial integration is uneven, contested, and always incomplete. The resurgence of regionalism, which scholars identified from the 1980s, reflected a mix of old and new conditions and forces. The “new regionalism” signifies the reconfiguration of the nation-state in the face of global economic integration, on the one hand, and the assertion of regional identities on the other, resulting in an uneven process of political and administrative decentralization which can be identified across the globe. Regionalism developed after World War II in many countries as a reaction to processes of centralization and nationalization in economic and cultural life. The forms of regionalism that emerged from the 1960s were often based around language divisions (Flanders, Catalunya, Quebec). In some cases, regions have made secessionist claims. The line between regionalism and nationalism can be blurred. Wales, Brittany, and Bavaria can claim national identities within the context of larger states. The “new regionalism” became more widespread in the face of economic restructuring in the 1970s and 1980s. Modernization strategies of states were concerned with eliminating regional inequalities and integrating peripheral regions into national economies. Centrally directed regional policies and the regionalization of state structures were used to achieve these ends. But defining economic management in territorial terms fueled the growth of regionalism as actors mobilized around these institutions to make economic, cultural, and political demands. The new regionalism has become firmly founded on claims that economic restructuring is best managed by effective regional institutions (Rodríguez-Pose and Sandall 2008). These arguments, though, have been often most vigorously made by regions with strong regional identities. Indeed, it is claimed that such identities facilitate economic development through processes of “untraded interdependencies” at the regional scale, although the evidence for this is inconclusive. But, the growth of regional institutions
to promote economic development often has encouraged democratic regionalism – making regional institutions accountable to the populations over which they have jurisdiction.

The new regionalism has dominated contemporary debates about the salience of the region as a focus for economic, cultural, and political action and has been a fruitful site for the investigation of contemporary social change. It provides a powerful and resonant account of regional change since the end of the 1970s and one which has appealed to policymakers, as well as scholars. Critics contend that the new regionalism canonizes the region as a functional space for economic planning and political governance. The region is presented as having a special efficacy in the governance of the flexible, post-Fordist economy. But, according to the critics, such a viewpoint requires an overemphasis on the dynamism of an unrepresentative group of high profile urban agglomerations, where competitive advantage is embedded in civil and political networks characterized by “institutional thickness.” Such examples do not represent the “leading edge” of global change, but unique and irreproducible cases. Moreover, a focus on the internal characteristics of these places leads to a neglect of their position in wider economic divisions of labor and political systems and leads to little more than the elaboration of a set of stylized “success stories,” which obscure understanding of broader patterns of regional change and owe less to their explanatory purchase and more to their instrumental appeal to regional political elites searching for narratives to justify their actions in an era of enhanced territorial competition (MacLeod 2001). The simpler versions of new regionalism certainly underplay the complexities of the contemporary regionalization of business and governance and the changing role of the state in relation to these. By fetishizing the region we can overlook the contradictory nature of the reterritorialization of the state and the extent to which states and regions are entwined in multilevel systems of governance.

The power of regionalist narratives, however, cannot be ignored. The narrative construction of the contemporary regional world does seem to include a cultural and political reawakening of regional consciousness. Regionalism, therefore, remains a key concept in human geography. Practically, it continues to shape culture and politics, influencing patterns of state formation and nation-building and framing patterns of socio-economic development, albeit in diverse ways.

SEE ALSO: Borders, boundaries, and borderlands; Cosmopolitanism; Identity;

References


Regulation

Martin Jones
University of Sheffield, UK

Regulation – dictionary-defined as the action or process of regulating or being regulated, replete with associated categories of adjustment, control, management, balancing, setting, synchronization, modulation, tuning, and so on – is a geographical concern. Working on the interface between economic and political geography, regulation is an important category of geographical analysis, and it certainly warrants not just an entry here but also a sustained program of research inquiry and engagement with policy and practitioner communities.

The entry point for thinking about regulation is that mediation is needed in our economies and societies. Contra the thought processes of neoclassical economics and the assumptions of equilibrium and optimality, everything else is neither equal nor without visible and invisible hands. The economy, for instance, is “not an unmediated outcome of universal and trans-historical processes operating across a featureless” (Peck 2000, 61) and stateless isotropic plane, as perceived by the likes of J.H. von Thünen and followers in land economy modeling. The economy is a function of highly complex, messy, and geographically grounded processes in places and regions. In contrast to the somewhat “abstract, orderly, and mathematically regularized world envisaged by orthodox economics” and exhibited during the long tradition of location theory in geography, certainly from the turn of the last century up until the 1970s, in which synchronous processes and stylized facts hold sway, the contemporary geographer is reminded that the world is more concrete, institutionally occupied, sometimes cluttered, and above all geographically uneven (Peck 2000, 61). Contemporary economy and society is an unsettled dynamic, requiring different forms of intervention to ensure stability and coherence, and the various trends in terms of both academic analysis of economy–society–space interactions and societal challenges have been covered brilliantly over the years in the various editions of *Location in Space* and *Modern Western Society* by Peter Dicken and Peter Lloyd. Two distinctive branches of regulatory thought in geography have emerged: first, notions of “real” regulation; and, second, regulation theory and the regulation approach.

Given this concern for dynamisms and interventions, spatial processes and administrative practices offer one window through which to view regulation. Gordon Clark’s research has been considerable in this field and has enabled geography to consider “real” regulation where administration and space are not separable categories but are defined by interactions and relationships, which in turn require institutional-oriented, fine-grained case study research (Clark 1992). This tradition started with a theory of “local autonomy” offered in the early 1980s as a way of capturing institutional and relative geographical power vis-à-vis the different tiers, or scales, of the state, in this case local and state-level governmental relationships. A series of typologies were offered – type 1 (initiative and immunity), type 2 (initiative and no immunity), type 3 (no initiative and immunity), type 4 (no initiative and no immunity), and so on – to capture ideal interactive regulatory situations.
This typology was later put to work empirically on innovative research on the geographies of the capitalist state and the field of legal geographies. In the latter, regulation is considered deeply in relation to local autonomy, set within the context of the rise of US federalism and the shortcomings of standard empirical analyses of urban governance and policies. Clark’s geographical analysis in *Judges and the Cities* (1985) brings into play the fluid and evolving regulatory roles played by courts in giving expression to social heterogeneity and pluralistic values, thus revealing an uneven and evolving legal hierarchy landscape with both strong and weak forms of local autonomy.

“Real” regulation has been continued by Clark through later work on pension fund capitalism, and a more recent paper teases out US state-level differences in the provision and uptake of financial services. The authors are interested in the variegated geographies of US state and local public employee retirement systems (PERS), which utilize various models of contracting out. These differ both within and across states according to the roles played by investment management agreements (IMAs). Clark and Monk (2014) focus on the Illinois area PERS to highlight not just the processes of contracting out through IMAs but also the detailed forms that these regulatory contracts take and how precontract screening of potential suppliers effectively “sterilizes” investment manager services. This microgeography of regulatory processes and practices helps to produce and reproduce an “institutional ecology” of “real” regulation (Clark and Monk 2014).

A different and more embedded analysis of regulation has been offered by geographers through the lens of regulation theory and the regulation approach. This approach takes as its starting point the decline of the postwar Fordist model of growth and development that occupied North America and Western Europe until the 1970s, and the subsequent emergence and rise of new industrial spaces, based on niche areas of economic activity and underpinned and supported by new state and society regulatory arrangements. This Fordist regime of accumulation, that is a macro-economically coherent phase of capitalist development and its decline, required a particular form of analysis that could go beyond both neoclassical economics (noted earlier) and applied structural Marxism’s stress on indicative economic planning (popular in France at the time). What became known as the “regulation school” offered an analysis of socio-economic change that tried to understand the importance of rules, norms, and conventions at a number of spatial scales (local, regional, national, and supranational) in the regulation and mediation of capitalism. Regulationists go beyond “real” regulation and explore the regulation of economic life in its broadest sense, acknowledging that capitalist development does not possess its own self-limiting mechanisms or follow an exclusive economic logic. Regulationists instead argue that socially embedded institutions and their networks are critical to the endurance of capitalism, despite contradictions and crisis tendencies. This analysis is not restricted to Fordism, or any suggestions of post-Fordism; the regulation approach, as it developed in the social sciences through the 1990s and beyond, has been concerned with analyzing the institutional infrastructures around and through which capitalism proceeds. This infrastructure varies within and between nation-states, according to the different sets of economic, social, and political circumstances. Geography matters to this reading of regulation.

A good example of how geographers have used the regulation approach can be found in the work of Mark Goodwin and Joe Painter, who have looked at how the local state and
urban politics can act as sites of regulation. Goodwin’s research has focused on the changing institutional forms and functions of the local state and how these can act as both “agent and obstacle” to regulation. Building on this point, Goodwin discussed the ways in which regulation, its codes, and decision-making procedures, occur not in a national vacuum but through sub-national state agencies, which are often the very medium through which regulatory practices are interpreted and delivered. Goodwin and Painter (1996) accordingly explore the notion of regulation as process, where the institutional forms and functions of the state are seen as being associated not only with stability but also with managing fluidity, flux, and change, which are constituted geographically. This approach emphasizes the ebb and flow of regulatory processes across time and space by using a modified version of regulation theory that can explore the plethora of new institutions emerging in the local state. New institutional spaces have incorporated a wider variety of civil society and business sector groups, as opposed to the representational forces of labor, capital, and the state, into the regulatory frameworks of capitalism, as processes of depoliticization have been operating under a neoliberal regulatory regime of capitalism. In the area of economic development, these agencies have delivered specific rather than multifunctional policy remits, often also through smaller territories than those of elected local government.

Painter and Goodwin (2000) have deployed the term “regulatory capacity” to assess the effectiveness of such new institutional regulatory arrangements. Their case study research in Sunderland in the northeast of England claims that there is little evidence of new institutional state forms and their functions providing the necessary mechanisms for stabilizing what regulationists have called a “new mode of regulation.” Sunderland, in contrast, has a “deficit in local regulatory capacity” and some state forms and functions are “clearly counter-regulatory” (Painter and Goodwin 2000). This last point has been further explored by others, who suggest that state forms and functions are modified to regulate and deal with policy problems created by previous rounds of state intervention. Crisis of crisis-management situations can, in some circumstances, raise the issue of political legitimacy, ultimately threatening the state’s regulatory mode of operation. Jones (2010) has charted the rise of an “impedimenta state,” where the state and its regulation has become a form of bureaucratic ponderousness, with administrative inertia and modes of negative coordination. The state has baggage and is increasingly bag-like in its appearance, actions, and reactions. The trouble with bags, though, is that they demand to be filled, and impedimenta seem to replicate rapidly to create congestion and confusion.

SEE ALSO: Legal geography; Local state; Regulation/deregulation; State, the

References


Further reading

Economic regulation is the political control and monitoring of activities involving the production and exchange of commodities. Commodities are goods and services that can be put up for sale, including land, money, and labor power. Regulators are vested specifically with the power to determine—often but not exclusively through legislative mechanisms—entry and exit to geographically defined markets; price floors and ceilings; subsidies for and the taxation of producers and consumers within these markets; mergers and acquisitions that may affect market entry or exit; and items that can be imported and exported from these markets. Entwined with economic regulations are social regulations such as traffic rules, pollution guidelines, and educational standards. These regulations enable social cohesion and, by extension, the smooth functioning of commodity exchange. In many respects, socioeconomic regulation is the de facto antithesis to ideal-typical “free markets”: producers have to adjust their strategies in relation to regulatory constraints in different locations and industrial sectors in order to continue operations, while consumers’ purchasing decisions are shaped directly by pricing controls and taxes. Viewed across a landscape of variegated socioeconomic regulations that are determined by geography, sector, and competing interest groups, the notion of ceteris paribus will inevitably remain an assumption in economic studies.

Deregulation refers to the removal of pre-existing restrictions on economic practices. It became a major political focal point in the hitherto highly regulated US economy after then president Richard Nixon resigned in 1974. His successor, Gerald Ford, almost immediately worked at removing constraints to production and trade in major industries such as banking, transport, and energy. A process of competitive deregulation ensued and became pronounced after individual state governments within the United States were allowed to deregulate industries and, consequently, trigger wholesale changes in other states without affecting existing regulations in those states. For instance, following the 1978 Supreme Court decision to permit an operating bank to export the usury law of its home state nationwide, the relatively deregulated South Dakota became a highly attractive location for banks to headquarter their US credit card operations. The action of one state engendered significant impacts on credit card use—and in turn effective demand—across the whole of the United States without changing existing regulatory constraints in other states. Deregulatory pressures soon extended to the United Kingdom—facilitated in large part by the election of Margaret Thatcher—and, to a lesser extent, Western Europe through the 1980s. Taken together, this process unraveled what was known as the Fordist–Keynesian mode of capital accumulation (more on this shortly). Increased competition impelled firms from these advanced capitalist economies to seek alternative locations of production and target markets in order to secure growing rates of profits. Viewed
REGULATION/DEREGULATION

in tandem with regulatory shifts that enabled economic liberalization in developing countries, deregulation in the capitalist “first world” arguably underpinned the new international division of labor.

At the superficial level, deregulation appears to be diametrically opposed to regulation because it removes constraints that were previously imposed on producers and/or consumers. Yet regulation and deregulation are not binary constructs in reality: the latter is an extension of the former. Regulation is necessary to establish rules for economic transactions and to ensure that laws are followed; deregulation not only entails the prior existence of regulation but is in itself an act of regulation. This is because the removal of regulations entails the a priori power of state agencies to determine how and to what extent an industry or market is to be regulated. For instance, the Airline Deregulation Act which jump-started deregulation in the US air transport industry had to be granted by the pre-existing legislative powers of the US government. And, starkly contrasting Nixon’s penchant for regulation, Gerald Ford created a “great hullabaloo” about deregulation after taking over the presidency (see Crain 2009, ch. 16). That this “hullabaloo” could set in motion deregulatory processes that were subsequently deepened by Jimmy Carter and Ronald Reagan was due to one a priori power – the political authority invested in the American presidency.

In a counterpoint to market fundamentalism, an ideology that espouses the intrinsic superiority of self-regulatory markets in the organization of socioeconomic relations, institutional economists have shown how producers and consumers do not always favor deregulation. Producers often form coalitions with other producers to lobby the state for the development and/or revision of regulations. This underscores the constitutive role of regulation in the process of capital accumulation. Since the 1970s, economists have characterized the way in which well-organized groups manipulate state power to capture rents at the expense of less organized entities as the private interest theory of regulation (also known as economic theory). Contrasting this is the public interest theory of regulation, a consumer-oriented (and, some would argue, welfarist) view of regulation as an intervention to correct or pre-empt market failure. A market is deemed to have failed when the quantity of a product demanded by consumers, such as adequate and clean drinking water, is higher than the quantity suppliers are willing to produce. The demand–supply gap – which, in neoclassical economic parlance, reflects a state of disequilibrium – is attributed to the lack of specific incentives that can be corrected by regulators in the name of social welfare. Despite their differences, the private and public interest theories share an underlying assumption that regulators are self-interested political maximizers who are susceptible to the influence of interest groups (either producers or consumers) during the regulatory process. The challenge for researchers, therefore, is to identify and/or predict the variables that enable regulators to maximize their interests in ways that simultaneously enable producers and/or consumers to attain their personal objectives.

Perhaps more interestingly, deregulation triggers and functions on the basis of new regulations. In an intriguing study termed The Utopia of Rules, the anthropologist David Graeber (2015) shows how the move toward more extensive market-like rule has not engendered a reduction of regulations; in fact, new regulations have emerged in order to support market transactions. This study not only explains the enlargement of the state bureaucracy but also the expanded pool of professionals – predominantly consultants and...
lawyers – who now advise producers and consumers how to negotiate these regulations. The enforcement of these regulations is not simply the remit of state agencies; it involves a chain of actors ranging from telephone advisers, front-counter sales staff, and third-party agents such as debt collectors or mortgage advisers. It is in this regard that the entwinement between deregulation and regulation becomes positive – deregulation may be a signal of tension in the existing regulatory framework, but, contrary to arguments (still) advanced by pro-market scholars and think tanks, it does not exemplify a zero-sum process that leads to an overall reduction of regulation.

**Economic-geographical engagements**

The private and public interest theories – and, more broadly, theories of regulation inspired by rational choice – did not strongly influence economic-geographical scholarship. Perhaps the closest economic geographers came to economists’ conceptualization of regulation/deregulation is the two-pronged approach to (i) map where deregulation has taken place, and then (ii) assess whether geographical shifts engendered greater socioeconomic “efficiency.” Another related approach is to evaluate the cost impacts of deregulation on the locational choices of producers. An example of this engagement, which takes deregulation as an isolated phenomenon to be studied (i.e., as not entwined dynamically with regulation), can be seen in the study by Goetz and Sutton (1997) on airline deregulation in the United States, in which deregulation is found to have led to the consolidation of the domestic airline industry and the formation of key domestic hubs and international gateway cities. At the global scale, Bowen (2002) shows how the deregulation of the global airline industry led to the relocation of air gateways which, in turn, further isolated specific parts of the world. Terms such as “spatial effects,” “spatial transfers,” and “core–periphery” were characteristic of this research stream, which overlaps the genres of spatial and regional science. Unsurprisingly, these studies largely construe space and place as passive containers of regulatory shifts; their primary goal is to map and evaluate the implications of emerging spatial patterns in firm-level responses to the removal of pre-existing restrictions in specific industries.

Economic-geographical research on deregulatory shifts in advanced capitalist economies were more strongly influenced by regulation theory (RT). RT was developed in France as an attempt to theorize changes in Fordist Keynesianism with reference to how these changes are shaped by and in turn impact the overall social structure. Fordism refers to a particular sociospatial division of labor that facilitates the production of standardized goods. Keynesianism refers to a state that actively manages aggregate demand and establishes wage relations to facilitate mass consumption. Taken together, Fordism–Keynesianism was a form of state-managed and privately led economic production that viewed social agents as more than simple labor power. Workers were viewed as consumption agents at the same time, and ensuring that workers were given the ability to consume was fundamental to the management of aggregate demand. In itself a response to the Great Depression of the 1930s, Fordism–Keynesianism experienced its golden era in North America and Western Europe in the two decades that followed World War II. It then encountered crisis in the 1970s.

Why this crisis came into being and how it undermined extant socioeconomic regulations became a primary research focal point of RT. To Michel Aglietta (1976), whose book *A Theory of Capitalist Regulation* was arguably the pioneering work on RT, the primary motivation for the
development of RT was a dissatisfaction with the (still) obdurate focus on “homogeneity” in two prevailing economic “sciences,” neoclassical economics and Marxism: “The neo-classical theory inspired by liberalism, which amounts to a representation of the system as a pure economy in a natural state of equilibrium, stretches the postulate of homogeneity to its very limits,” writes Aglietta, while

the Marxist view of the economy remains strongly homogeneous because capitalism is supposed to move in accordance with general laws which lead to its overthrow, whatever the nature of the society in which it develops. Furthermore, the overthrow of capitalism heralds the coming of a transparent and homogeneous system of perfect planning. (1998, 42)

RT regards as problematic the “naturalness” of the “pure economy” and the “general laws” of capital accumulation in framing politico-economic evolution. It makes the presupposition that there is no such thing as a pure, self-regulating economy that will always reach a state of equilibrium; if “equilibrium” is or appears to be reached, it is always a temporary period of compromise and consensus that is mediated by social institutions or agents. It also presupposes that the so-called general laws of capitalist accumulation would not engender a prophetic movement toward socialism or communism because society is somehow able to negotiate — or perhaps “tolerate” would be the more precise term — the problems engendered by capitalist exploitation. It was this emphasis on heterogeneous institutional formations that made RT such an attractive conceptual framework for examining the primary phenomenon confronting advanced capitalist economies between the mid-1970s and early 1990s — the previously mentioned pressures to dismantle the Fordist–Keynesian social compact.

Focusing on what Adam Tickell and Jamie Peck (1995, 357–358) term the “structural coupling” between the accumulation system (AS) and mode of social regulation (MSR), economic geographers drew on RT to understand what made individual economies work coherently across an uneven socioeconomic landscape. To follow Mick Dunford (1990, 305–306), an accumulation system is defined as “a systematic organisation of production, income distribution, exchange of the social product, and consumption.” The AS comprises (i) organization of production and workers’ relationship to the means of production; (ii) speed of capital circulation and valorization (yielding of its value); (iii) management principles; (iv) distribution of value (e.g., good wage); (v) composition of demand; and (vi) manner of articulation with noncapitalist forms. This system is in turn supported by noneconomic or extra-economic institutions (the state, trade unions, industrial associations, schools and universities, etc.) to constitute the MSR. As Mick Dunford (1990, 306) puts it, the MSR “defines the rules of the game … [and] guides and stabilises the process of accumulation.” The MSR examines the determination of (i) the wage relation (or wage nexus); (ii) forms of competition; (iii) monetary and financial regulation; (iv) the state and governance; and (v) the international trading regime.

The successful coupling of the accumulation system and the MSR in turn produces a regime of accumulation. At one level, this coupling follows the tradition in political economy by examining how economic surplus is produced, distributed, and accumulated within a class-divided society. A key difference is that it does not privilege the capital or the state as possessing the determining role in the last instance. Management of the wage relation is fundamental to the socioeconomic regulatory framework, and it is when this management
is transferred from the state to profit-seeking firms that aggregate demand changes. The incorporation of the MSR in the analysis of regulatory shifts differentiates RT practitioners from those incorporating rational choice economic theories to explain or predict (de)regulation. Crucially, there is no deregulation from an RT standpoint. Rather, regulations are strategically reconfigured – and this reconfiguration includes the removal of pre-existing restrictions – or their repurposing. Accumulation regimes are assumed to contain intrinsic tendencies for overheating and would thus require rolling series of economic and noneconomic regulatory interventions in anticipation of, or to ameliorate, these tendencies. For instance, it was because of overheating that the Nixon administration intensified industrial regulation across the United States during the late 1960s and early 1970s.

The primary concern for researchers working with RT, broadly referred to as regulationalists, is to ascertain and explain the institutional means by which this overheating occurred. RT identifies three types of crises that could potentially destabilize regimes of accumulation:

1. microcrises – affecting individual units of capital, e.g., the unprecedented £6.4 billion loss declared by the British retail chain Tesco in 2015;
2. conjunctural crises – linked to cyclical downturns, e.g., short-term decline in tourist inflows due to uncertain political developments, like the Bangkok riots in 2010;
3. structural crises – affecting the entire economy, usually framed within the nation-state, e.g., the Greek and Irish governments’ inability to service debts after the 2008 global financial crisis, which led to bailout packages from the European Union and the International Monetary Fund.

Since its “importation” by economic geographers, RT has been instrumental in demonstrating the constitutive role of non- or extra-economic institutions in the multiscalar (de)regulatory shifts toward an uncertain future known as post-Fordism. In attempting to understand a distinct economic-geographical transformation – the new international division of labor (NIDL) – Alain Lipietz (1982; 1987) critically evaluates whether the regime of accumulation known as Fordism can be exported to the global scale. Building on his analysis on the central tenets of RT, he argues that the system of capitalism is being expanded geographically because the “crisis” of the Fordist accumulation system (e.g., falling rate of profits, falling productivity, reduced demand) can be successfully deferred overseas. This success is contingent on an uneven reregulation of labor–capital relations in the former Fordist heartlands. Comprehensive defeats of labor unions in the United States and the United Kingdom (led by Ronald Reagan and Margaret Thatcher) led, at one end of the spectrum, to neo-Taylorism, which is effectively a fragile balance characterized by flexible employment conditions. At the other end, collective renegotiation in the Scandinavian economies led to a regime known as Kalmarism, which retains many of the characteristics of an ideal-typical Fordist–Keynesian prototype. Lipietz (1997) subsequently situated his conceptual framing through categorizing the effects of this rebalancing act – or, using Tickell and Peck’s term, “structural recoupling” – in Western Europe and North America.

This focus on structural recoupling fundamentally eschews the presupposition of or empirical focus on a deregulated end state. While Lipietz is not the only analyst to examine the relationship between Fordism, NIDL, and post-Fordist restructuring, his works in the 1980s established the agenda through which to
comprehend Fordist crises and how these crises are resolved and/or deferred through multiscalar reregulation. His central and highly prescient insight is that NIDL was never an autonomous process; nor did it lead to the absolute reduction of regulations in advanced capitalist economies. Rather, its geographical reach was contingent on how institutional reconfigurations in previously Fordist economies could lead – for capitalists and politicians, at least – to the export of the crisis overseas, as exemplified by deindustrialization in the United States and the United Kingdom, or trigger in situ industrial restructuring, which occurred in Scandinavia and, to a lesser extent, the former West Germany. These insights endure in contemporary analyses of variegated neoliberalization, uneven economic-geographical development, and what has emerged as a major phenomenon: the emergence of the Chinese political economy as a central site of rapid but highly uneven socioeconomic transformations.

For one, it offers a grounded twist to David Harvey’s (2010) recent conclusion that capital never overcomes its crisis tendencies but merely moves them around: some regimes of accumulation are more enduring than others, and, for capital to do the moving, certain regulatory preconditions must be established (more on this shortly).

Economic geographers’ sensitivity to regulatory changes in and through different spatial scales led to a dissatisfaction with the state-centric focus of RT. The key issue of debate during the 1990s was the possibility that different scales of socioeconomic regulation could occur within the same national space, and that RT must respond to the issues surrounding uneven development (Peck and Tickell 1995; Low 1995). Put another way, the constitutive role of geographical scale in RT is a key “missing link” (Tickell and Peck 1992; see also Collinge 1999) in RT-oriented work. This critique helped drive subsequent economic-geographical research on the tensions between regulation and deregulation, especially work on understanding the rescaling of capital accumulation in and through city-regions. Interestingly, the growing influence of RT in economic geography corresponded with an “institutional turn” within economic geography in the 1990s. This was spearheaded by Ash Amin and Nigel Thrift’s (1994) well-discussed work on the constitutive role of “institutional thickness” on regional economic development in Western Europe. To Amin and Thrift (1994, 15), institutions “mobilize the region with speed and efficiency,” and an institutionally “thick” region is characterized by:

1. an extensive institutional presence, from governmental bodies to banks and educational establishments;
2. intense interactions between these institutions;
3. well-defined structures of domination (i.e., unequal power relations are defined and accepted by the different institutions); and
4. a collective geographical imagination, which leads to a sense of shared interest in regional stability and growth.

Amin and Thrift’s (1994) highly influential, agenda-setting work addresses some of the key concerns that were further elaborated in economic-geographical research from the mid-1990s to the present, namely the importance of noneconomic support for workers (a key tenet of research on “sticky” industrial districts/clusters); the importance of social relations (a key focus of the relational economic geography debate in the early to mid-2000s); unequal power distribution (still ongoing in studies of production networks); and, through the focus on geographical imaginations, discourse, and identity (revived recently by Allen Scott through his work on cognitive-cultural metropolises). Throughout
the 1990s, they proceeded to advocate a socio-economic approach to understanding economic development in Europe. While not directly addressing regulation and deregulation, this approach led to economic-geographical research that underscores the constitutive role of socio-economic institutions in economic practices. Notable works include Michael Storper’s (1997, 27) work on the “holy trinity” of economic production (technology, organizations, and territories); Philip Cooke and Kevin Morgan’s (1999) formulation of the “associational economy”; and Henderson et al.’s (2002) global production networks (GPN) framework. Indeed, Amin and Thrift (2000) have so taken to the causal strength of MSRs that they actually considered dispensing with concepts from economics in toto. Viewed collectively, this new research wave was instrumental in shaping what Bob Jessop (1997, 514) terms the three new “innovations” in RT:

- **space**: the focus on the politics and constitutive roles of scale;
- **government and governance**: power relations and the shift in state–market engagement through the formation of new ruling coalitions involving both public and private actors and institutions;
- **discourse and identity**: geographical imaginations, discursive and practical performativity.

The incorporation of these three innovations within the RT framework underscores how regulatory scales are not simply constructed by capital. The expansion of Chinese economic influence at the global scale is instructive in this regard: through its fixed foreign exchange regime and one-party political system, the Chinese government operates at a variety of geographical scales to contour and react to the demands of capital. To make strategic use of its foreign reserves (accumulated as a result of its exchange rate regime), a sovereign wealth fund (the China Investment Corporation) was set up to explore investment opportunities at the global scale. State-owned enterprises (SoEs) are also encouraged to “go abroad” (zouchuqu). In turn, China’s outward FDI – which surged from around US$3 billion in 2002 to US$116 billion in 2014 – places the country within the world’s top five national sources of FDI (Lim 2010). This globally expansionist strategy, however, is an anomaly for a developing country where the per capita income is approximately a sixth of that of the United States. An outcome of a strategic macroeconomic policy (fixed exchange rate regime), the sudden growth in outward direct investments from China, exemplifies the direct impact of national-level accumulation regimes on global capital flows. That this expanding influence was and remains predicated on the prior willingness of the Chinese state to de- and reregulate the national economy in 1978 points again to the dynamic entwinement between regulation and deregulation.

**Conceptualizing the regulation–deregulation dynamic: future directions**

Despite its theoretical promise, the importance of RT gradually faded since the mid-2000s. In the late 1980s Allen Scott, David Harvey, and Michael Storper led the way in arguing that a new accumulation regime characterized by flexibility has emerged. This triggered a series of heated debates, during which Meric Gertler (1988) and John Lovering (1990) rebutted the claims that post-Fordism has become a byword for any flexible form of production. The fundamental problem, as Ray Hudson (2006, 174) puts it cogently, is that RT portrays the transition from one regime of accumulation to another as
a “seamless” process that overlooks the “messy complexity of the historical geographies of capitalist economies and their overdetermination as a result of struggles between social actors over (inter alia) the extent and form of state involvement in economy and society.” Because RT could not offer any predictions on what post-Fordism is or could be, Peck and Tickell (1995, 36) presciently predicted that RT could go down “the dustbin of Fordism.” Against the contemporary backdrop of transnational production and financial networks, very few economic geographers talk about post-Fordist restructuring anymore. And this means few – if any – economic geographers now draw on even the spatially reformulated RT to analyze and assess contemporary problems of (de)regulation.

Crucially, in view of the severely reduced faith in neoclassical economics following the devastating 2008 global financial crisis, this might mean missing a critical moment to seize the theoretical and even policy initiative in driving future discussions on economic regulation. For one, deregulation is no longer postulated as a utopic end state of economic efficiency. Indeed, Alan Greenspan, one of the surviving economic advisers who helped Gerald Ford launch the aforementioned deregulatory wave across the United States, even publicly acknowledged that he might have been wrong in his convictions about the self-regulatory market all along. What this acknowledgment amounts to is not simply a reminder that economic regulation remains important (RT-inspired economic-geographical research has long demonstrated this as a fact); it points to financial deregulation in multiple economies (e.g., United States, United Kingdom, Iceland, Ireland) as a specific regulatory failure that plunged multiple social formations into a financial crisis from which they are still trying to recover. It is to this second point that economic geographers are well poised to respond.

In spite of this opportunity, however, Peck and Theodore (2007, 762) identify a paucity of “institutional theories of the uneven development of capitalism” within economic geography. This observation points to the continued and urgent need to understand capital accumulation as a regulated process, deeply embedded within and simultaneously cutting across multiple sociopolitical structures. Perhaps more important is the continuing importance of the state – the primary scale of reference for theories on regulation – in determining and implementing legal parameters that determine the geographies of production, work, and consumption. Christophers’s (2016) work on the “leveling” effects of laws regulating competition and monopolies in the United States and the United Kingdom exemplifies the contemporary relevance of a state-focused framework that critically engages with and contributes fresh insights on RT. There is much to build, then, from the regulationist approach, notably in two areas of contemporary economic-geographical research: (i) variegated neoliberalization and (ii) primitive accumulation.

Expressing a dissatisfaction with the constraining effects of preconceived economic categories on understanding extant neoliberalization processes, Brenner, Peck, and Theodore (2010, 207) argue that “the variegation of neoliberalization processes involves a series of constitutively uneven, if cumulatively transformative, shifts in the qualitative form of regulatory uneven development itself.” This approach effectively addresses and builds on Hudson’s (2006) concern that RT only offers a vision of a clean break from one production system (Fordism) to another (an undefined post–Fordism). To Brenner, Peck, and Theodore (2010), the impacts of specific neoliberal projects – characterized by an emphasis on economic deregulation and the corresponding formation of a “small” state – need not dramatically transform regulations over an
entire economy over a fixed period of time. Rather, each round of reregulation tends to be responses to the contradictions of earlier rounds of market-oriented deregulation, and this plays out differently within and between state space. The bigger picture of regulation and deregulation can thus be determined only after studying these transformative shifts over an undetermined period of time. Put differently, a coherent and defined accumulation regime may appear only after a series of discontinuous and capricious shifts toward market-like rule are aggregated. Sometimes, the accumulation of these changes and shifts may be limitless and there may be no coherent system. Neoliberalization remains variegated across the global economy precisely because the regulation–deregulation process is not a zero-sum one; in spite of deregulatory pressures, states continue to regulate economic practices in order to attain non-economic objectives, while workers remain active in fighting to secure better terms of employment. The goal of future research is to ascertain and evaluate these dynamic responses to neoliberalization.

Concomitant with the global expansion of neoliberal logics is primitive accumulation, or what Harvey (2003) terms “accumulation by dispossession.” The central question of contemporary capitalism remains why a small group of people is able to seize an inordinately huge amount of resources to start businesses and begin exploiting fellow human beings (through employer–employee relations). In Indonesia, for instance, 3% of the population owns and arguably continues to appropriate 70% of the economy’s value. In the 1990s the Russian state allowed a few oligarchs to purchase “shares vouchers” from ordinary citizens and effectively sold off a sizable portion of the state’s wealth. The primary consequence of primitive accumulation is to disconnect a large pool of people from resources that would allow for self-determination; this in turn impels these people to put on sale their labor power in labor markets. How the primitive accumulative process is effected, the extent to which this process is institutionally mediated, and how the affected population are persuaded to accept “proletarianization” are key research questions that address and build on the regulation–deregulation entwinement. Primitive accumulation may appear to be an unintended outcome of deregulation, but it may also be a deliberate calculation to transpose social agents into abstract labor power. That primitive accumulation is allowed to happen at such a large scale suggests there could be global regulatory failure. Identifying the specific (de)regulatory mechanisms that facilitate primitive accumulation in both developed and developing economies would go a significant way to understanding the multidimensional relationship between labor exploitation, surplus extraction, and crony capitalism.

SEE ALSO: Corporations and the law; Economic geography; Global production networks; Governance and development; International division of labor; Labor geography; Labor market; Neoliberalism; Power and development

References


Further reading


Relational assets

John Pickles
University of North Carolina at Chapel Hill, USA

“Relational assets” refer to the intangible, behavioral, and normative practices that influence interfirm behavior and firm and regional competitiveness. The concept was originally developed in two main contexts. In business and management theory, the idea emerged in response to the growth of multinational corporations and global sourcing practices, and the tendency to overlook the ways in which even multinational corporations and firms depend on acknowledged, unacknowledged, and untradable assets. The concept thus sought to deepen traditional economic and regional analyses focused on the factor endowments of emerging multinational corporations and international business activities by refocusing on the social and cultural practices that affect competitiveness.

Dunning (2002) identified three specific kinds of income-generating assets that shape the economic behavior of firms and regions: assets that are unique to specific firms; assets that are external to the firm but on which its performance depends; and assets that relate to the broader institutional and cultural contexts on which the first two depend. Historically, the relative importance of, and form taken by, such assets has changed. In agrarian economies, land and property ownership were crucial in shaping social relations, economic performance, and geographical outcomes. In industrial economies, the physical and financial assets of a firm and the institutions that protected financial property, regulated social and market norms, and managed labor markets and consumption norms were important. In contemporary globalizing economies, attitudinal attributes such as values, honesty, trust, cultural sensitivity, and reciprocity operating across spatial scales; firm-level attributes of reputation, experience, and openness to innovation; and characteristics of particular countries such as the rule of law, transparency, freedom from corruption, and the upholding of human rights are all crucial to economic performance (Dunning 2002, 476). For Dunning, firms with a greater number and intensity of linkages, such as multinational enterprises, tend to have more relational assets than those with fewer linkages, such as locally focused firms. The result has been a refocusing of the analysis of economic performance on the ways in which such cultural, regional, and institutional assets nurture or inhibit firm and regional competitiveness.

Regional geographers have complemented this business argument by turning to institutional analyses that focus on the ways in which regional economies can be understood as “stocks of relational assets” that rely on intangible institutional and cultural practices (such as trust and reciprocity). While Dunning focused on the growing complexity of relational assets in internationalizing businesses, geographers have shown how different regions enable different kinds and densities of relational assets, most notably in works dealing with Silicon Valley, the “Third Italy,” and metropolitan learning regions. For Storper (1995), the metropolitan economy possesses particular locational advantages in terms of the number, intensity, and diversity of linkages, networks, and relations they offer. Regional
RELATIONAL ASSETS

capabilities, information, and knowledge foster practices of conversation and coordination which enhance the activities of individual economic actors, both individual and collective, and intensify the effects of participation in regional economies to foster economic accumulation (Storper 1995, 3–4). Here:

the territorial element of the holy trinity (technologies, organization, territories) needs refocusing, from the geography of input–output relations – industrial complexes and spatial divisions of labor – and the economics of proximity in traded linkages, to the geography of untraded interdependencies and the dialectics of proximity and distance in them. (Storper 1997, 43)

A variety of related insights have emerged from this notion. Instead of “the region” being understood as a series of fixed material investments and technologies, the idea of relational assets has been important in the development of notions of institutional thickness, untraded interdependencies, learning regions, innovation regions, association economies, and perhaps most recently, from evolutionary economic geography, local buzz. For Henry Yeung, this marks a significant turn in the study of regional economies, where “a relational approach to regional development seeks to identify complex relational geometry comprising local and non-local actors, tangible and intangible assets, formal and informal institutional structures, and their interactive power relations” (2005, 48; emphasis original). In recent years geographers, anthropologists, economic sociologists, and cultural theorists have expanded this relational ontology by focusing on diverse and alternative modalities of power, economy, and life through their engagements with the importance of context. Yet, as Peck (2005, 133) warns us, when turning to issues relating to the social, spatial, and scalar constitution of economic systems, identities, processes, and development paths, we all too often leave “only fuzzily defined and un-theorized ‘context’ in the background.”

More recent efforts have engaged with this problematic of context. For Doreen Massey (2005, 110), the turn to concepts such as context and embeddedness are a recognition of “contemporaneous coexistence and becoming,” and for Larry Grossberg (2006, 4) what we mean by context must always be understood as “a complex, overdetermined and contingent unity.” From this same perspective, Pickles (2012) asks how an understanding of context and contingency help to develop a conjunctural analysis of economic and regional change. The result has been a rich array of investigations that show the many ways in which these relational assets are activated in firm and regional economies, and how social and spatial networks and linkages matter. By focusing on such networks, forms of access, information, and trust, economic sociologists, economic geographers, and regional economists have identified different kinds of networks and forms of governance – between firms in a particular industry, between buyers and suppliers, between one sector and others, and between manufacturing and service firms and venture capitalists and investment banks. Some have focused on the ways in which network resources obtained from prior alliances provide differential access and referral benefits, and also produce capacities that increase the proclivity of firms to form further alliances (e.g., Pickles and Smith 2015). Others have focused on the ways in which state and private actors have formed regional partnerships to enhance regional economic performance, integrate once fragmented production systems, and expand their role in regional trade systems.

SEE ALSO: Economic geography; Industrial districts; Industrial geography; Local embeddedness
References


Further reading


Since the 1970s there has been a lively scholarly exchange, principally among Anglo-American geographers, over whether and how academic geographers should seek to shape policy and influence public debates over pressing social, political, and environmental issues. These exchanges are commonly termed the “relevance debates.”

Since its institutionalization as a formal discipline, some academic geographers have had considerable influence in the public arena. In the early twentieth century, for example, Halford Mackinder advanced geopolitical ideas that influenced the orientation and practice of foreign policy in the United Kingdom and beyond. In 1919, Isaiah Bowman accompanied US President Woodrow Wilson to the Paris Peace Conference, where he headed the US scientific delegation that contributed to the redrawing of the political map of Europe. In the middle of the twentieth century, Gilbert White pioneered an approach to floodplain management that continues to be influential to this day.

Geography’s Quantitative Revolution opened up opportunities for academic geographers to contribute to transportation and urban planning initiatives, and to offer input on a variety of decisions with a locational dimension. By the turn of the 1970s, however, little geographic work, whether quantitatively oriented or not, spoke to the social upheavals, international conflicts, and growing environmental activism of the time. Against this backdrop, a number of prominent geographers began to draw attention to what they saw as geography’s lack of relevance. They argued that geographic theory had become abstracted from pressing real-world problems, and they called on geographers to undertake work aimed at solving those problems (e.g., Chisholm 1971). Because a particularly sharp debate over this matter emerged in Anglo-American geographical circles, the term relevance debates has come to be primarily associated with academic exchanges over questions of relevance that have occurred among North American and United Kingdom-based geographers.

In the wake of the growing calls for relevance at the beginning of the 1970s, debates quickly emerged about what constitutes relevance and what the goals of relevant research should be. Some argued that geographers should actively look for ways to contribute to the policy-making process, whereas others contended that a policy-driven geographical research agenda would subvert the discipline’s potential to contribute to an understanding of underlying social forces or challenge existing sociopolitical norms. In a particularly widely referenced article entitled “What Kind of Geography for What Kind of Public Policy,” David Harvey (1974) argued that most geographic efforts to inform public policy involved playing into an unjust status quo. Harvey did not argue against relevancy; instead he made the case for an approach to relevance rooted in the effort to counter the structural inequalities, racism, and ethnocentricism of the corporate state. For those sympathetic to Harvey’s perspective, the historical lineage of relevancy in the discipline of geography ran less through Mackinder,
Bowman, and White, than it did through the likes of Élisée Reclus, Peter Kropotkin, and (much later) Bill Bunge.

Geography’s relevance debate in the 1970s reflected fundamentally different views of the discipline’s social role. For some, that role lay in the contributions geography could make to promoting efficiencies and improving public policy; for others the discipline’s mission was to pave the way for the development of alternatives to unjust structures and arrangements. Those in the former camp sought to draw attention to the social costs of foregoing diagnosis and action in favor of what they saw as an abstract, potentially undemocratic revolutionary agenda. Those in the latter camp focused on such issues as the intellectual constraints that can accompany cooperation with policymakers.

During the last two decades of the twentieth century, debates over relevance took place against the backdrop of a growing body of geographical work that spoke directly to such issues as environmental change, urban restructuring, human responses to hazards, and planning choices with a locational component (the latter spurred in significant part by the rapid development and proliferation of geographic information systems (GIS)). As geographical scholarship with a clearly discernable sociopolitical dimension became more common, discussions of relevance came to focus increasingly on the practices and ethics that are appropriate to academic engagement with social, political, and environmental issues. A growing number of commentators called on geographers working with those in the policy realm to be explicit about their positional-ity and its influence on their work. Some argued that academics engaged in policy work should explicitly acknowledge their own values and beliefs, recognize their lack of independence, be willing to critique the status quo, and promote public accessibility to their work product. Others called on geographers to be attuned to ethical dilemmas that can accompany policy work, and be prepared for conflicts that might arise between them and policymakers with different agendas and priorities.

As the twentieth century gave way to the twenty-first, relevance came to be an expected feature of most geographic scholarship. Social and institutional changes were behind this development. The US National Science Foundation and other funding agencies began requiring applicants to explain in detail the larger social impacts of their projects. At the same time, the rapid growth in the use of GIS in the private and public sectors opened up rapidly expanding opportunities for geographers to address applied problems. Moreover, cutbacks in funding for higher education meant that geographers increasingly needed to justify their academic positions in terms of “real world” consequences and look for funding support from sponsors with an applied agenda.

An avowed commitment to relevance now runs through much geographical scholarship, but relevance debates are still very much in evidence. They center around three key issues: the long-standing question of what constitutes relevance, the obstacles geography faces in becoming more relevant, and the types of practices that are most likely to lead to constructive applied engagements. Turning to the first of these issues, a number of geographers have sought to broaden the concept of relevance beyond the policy realm. Kevin Ward (2006), for example, has argued that there are different types of geographies with an explicit relevancy component: activist, participatory, and policy-oriented. He and others have encouraged an expanded conception of the audiences geographical work should aim to reach. Borrowing from Michael Burawoy’s (2005) writings on “public sociologies,” the term “public geographies” has
gained increasing acceptance – referring to a variety of different types of audiences (publics) that geographic scholarship might address. The embrace of public geographies speaks to the need for geographers to recognize that there are important audiences for research that extend beyond social and political elites.

There is also a body of work highlighting the limited ways in which geography is having an impact beyond the academy and seeking to explain that state of affairs. Scholarship in this vein has drawn attention to the virtual absence of geography in major public debates over social and political developments, and the concomitant paucity of geographers who are public intellectuals. To explain this state of affairs, Castree and Sparke (2000) have highlighted some of the institutional barriers to public scholarship, Murphy (2013) has pointed to the hesitancy of geographers to engage grand narratives, and a number of other commentators have pointed to a dominant approach to writing that is inaccessible to the general public. Some commentaries in the latter vein have argued that geography’s post-modern turn had directed attention away from the types of grounded, empirical, and politically committed work that can influence policy. The latter claim has generated considerable controversy, even as it has drawn attention to a key contemporary element of the relevancy debate: is much geographical work in fact relevant, and if not, why not?

A final pillar of recent commentaries on relevance focuses particularly on the policy realm and concerns the advantages and pitfalls of conducting different types of policy-oriented research. Beaumont, Loopmans, and Uitermark (2005), for example, examine ways in which policy-oriented researchers can avoid clientalistic relationships with sponsors and retain a commitment to traditional academic standards in the face of social engagement. Interventions in this vein reflect a relevance debate that has moved from an initial focus on whether geographers should focus on relevance to an emphasis on how relevance is understood, and how it can be most effectively, and ethically, practiced.

**SEE ALSO:** Geography in higher education; Participatory geographies; Public policy

**References**


**Further reading**

The study of religion in human geography is a vibrant subfield of social and cultural geography that has grown significantly since being labeled a depleted and marginal field in the early 1990s. This growth is evidenced by the publication of key texts (Stump 2008) and edited collections (Hopkins, Kong, and Olson 2013). Geographies of religion have developed as a result of a range of different influences and interdisciplinary debates. Some of the earliest work in this area was shaped by the Berkeley School of cultural geography, with work on spatial patterns arising from religious influences and the impact that religion has in shaping the physical landscape. Much of the more recent work in this area was shaped by the Berkeley School of cultural geography, with work on spatial patterns arising from religious influences and the impact that religion has in shaping the physical landscape. Much of the more recent work in this area was shaped by the Berkeley School of cultural geography, with work on spatial patterns arising from religious influences and the impact that religion has in shaping the physical landscape. More recently, geographies of religion have diversified further through engagement with neighboring disciplines and different subfields of geography. For example, Knott (2005) offers a rich conceptual tapestry for understanding the spatial location of religion through her work, which draws on debates within sociology of religion and religious studies, and Brace, Bailey, and Harvey (2006) have proposed a framework for understanding historical geographies of religion. Moreover, researchers interested in political geographies and geopolitics as well as economic geographies have also started to explore what happens when debates about religion and faith intervene into landscapes of politics and finance. Within social and cultural geographies, an important focus of work about religion is on the mapping, measuring, and monitoring of patterns of religious affiliation and experiences of ethnic and religious residential segregation. This work emerged from long-standing debates within social geographies about the politics and spatiality of race and ethnicity. Research in this area has centered on mapping changes in levels of ethnic and religious segregation over time, as well as on the spatial distribution of congregations or places of worship, such as megachurches, and has tended to focus on minority religious groups in contexts where they represent a diasporic or migrant population. Research in this area has also looked at the existence of specific religious communities or areas – be these actual or imagined – such as the ummah (the global Islamic community) or the Jewish eruv and the political contestations associated with such spatial manifestations of belief. Some of this work has been useful for policymakers as well as academics, given that it has contributed to explaining patterns of segregation and has offered insights into changing population patterns over time.

Strongly related to the urban theme of these debates, geographers studying religion have also explored the construction and contestation of religious built environments and architectural styles in different contexts. Work here has focused on the spatial practices, social exclusions,
and contested politics associated with the location and use of religious buildings and places of worship. These often connect to debates about multiculturalism and citizenship, but have also included engagements with architectural research and discussions about urban design and planning. An important strand of this work has charted the racist and exclusionary ways in which the location and development of mosques, particularly in Muslim minority contexts, are resisted by national governments and local communities. This work has tended to look at officially sacred locations – such as mosques, churches, synagogues, and temples – although there is now more research emerging that explores how issues of religion and matters of faith come into play in locations beyond the officially sacred.

An interest in the cultural politics of identity among social and cultural geographers in the 1990s led to a series of important developments in geographies of religion, as debates within feminist geographies, children’s geographies, and political geographies and geopolitics emerged as significant strands of research. Work has drawn attention to the complex manifestations of religious identities and how these are performed by people of varying faith backgrounds in different geographical locations and contexts. In particular, such research has been sensitive to the intersection of religious identities with other social identities such as those related to gender, race and ethnicity, class, and sexuality.

Drawing on debates within feminist geographies and geographical studies of gender relations, a significant body of scholarship within geographical studies of religion focuses on the ways in which the politics and practices of religion and spirituality are intertwined with gendered expectations and relations. In particular, exploring the identities, practices, and attitudes of Muslim women in different contexts has been a key aspect of research (e.g., Gökariksel 2012). Work has focused on the contested politics of the Muslim headscarf in diverse contexts such as Turkey, Australia, and the United States, challenging dominant representations of Muslim women as oppressed and marginalized, and exploring their engagements with different private, public, and institutional contexts. Partly in response to this, an interest in Muslim masculinities has emerged, with research on the contested intersections of Islam and masculinity, as well as the spatial practices and exclusionary behaviors of Muslim men in the mosque, in public spaces, and at home.

Parallel with the development of gendered geographies of religion, the introduction into the discipline of children’s geographies and debates about young people and family life has created a rich set of insights into the religious lives of children and young people. Researchers have explored a diverse range of themes including young people’s engagements with religion through different (religious and secular) institutions such as schools; and the role of local and national geographies in shaping young people’s faith. Connected to debates about gender and home, research has explored the ways that young people’s religious identities and faith practices are shaped in complex ways by intergenerational relations (Hopkins, Kong, and Olson 2013). Research has also centered on different institutional contexts, including, for example, how young people who are members of the Scout movement in the United Kingdom engage with everyday informal expressions of “worship” that take place outside of formal religious space, and Christian students’ experiences of university life in England.

Debates about citizenship and rights to public space informed by political geographies, alongside the emergence of research about religious geopolitics, have played an important
role in shaping geographical scholarship about religion. An important example of work drawing on discussions of citizenship has been Nagel and Staeheli’s (2008) research about how Arab activists in the United States respond to political debates about integration and their complex positioning in relation to their place of settlement. Researchers within the subfield of geopolitics have also started to explore the role of religion, particularly with reference to Christianity and evangelicalism in the United States, and there are signs that this interest is now expanding to include other faiths and spiritual practices. A number of significant international (geo)political events (most notably 9/11, but also the bombings in Bali in 2002 and 2005, in Madrid in 2004, in London in 2005, and in Mumbai in 2008), as well as other national and local political occurrences, have played a highly significant role in shaping the everyday lives of people of faith, and this remains an important focus and contextual background for geographers researching religion.

Connecting with literatures about alternative spiritualities within sociology of religion and arising from a desire to engage with the unofficially sacred, a small strand of work within geographies of religion has focused on different forms of spiritual practice. Alternative spiritualities tend to refer to spiritual practices that distance themselves from mainstream traditional and institutionalized forms of religion. Many participants who engage in alternative forms of spirituality resist the institutionalized structures that they feel dominate mainstream systems of faith and belief. One example of work in this area is research with New Age spiritual seekers which has shown how everyday objects and actions are given spiritual meaning: this work suggests that geographers of religion could usefully open up the nature of the everyday to further scrutiny in order to understand more about the relationship between the sacred and the profane.

Interrelated with the debates in political geographies, and coupled with urban themes and an interest in the role of faith-based organizations, postsecularism is one of the most recent issues to emerge within geographies of religion. Connected with long-term discussions about secularization theory within the sociology of religion, research about postsecularism challenges assumptions about complete secularism and instead explores the ways in which religious actors and institutions have re-emerged so that there is a coexistence of religion and secularism. Olson et al. (2013) have recently offered a retheorization of the postsecular by suggesting that a focus on embodiment and transcendence challenges the solidity of traditional religious categories and their association with clearly defined spaces, thereby opening up the sacred and profane to deeper analysis.

Debates about affect and emotion have also recently found their way into discussions about religion, faith, and spirituality, partly as a result of debates about postsecularism and the blurring of boundaries between the religious and nonreligious. Holloway (2013, 205) observes that being spiritual or religious is “enacted through many different registers of emotion and affect – from quiet contemplation to ecstatic expression, from serious reflection to peaceful assurance, from joy to humbleness, from anger to awe,” and focuses on the spaces associated with ritual in order to explore these. Discussions about religion, affect, and emotion are likely to be an important feature of future research about geographies of religion. Furthermore, appreciations of the emotional landscapes of faith and belief have also provided a valuable context for the development of research about deathscapes and places of mourning. In particular, research has explored how geographies of death and bereavement are central to
understanding different patterns and practices of belief.

Aside from researchers, who seek to map, monitor, and measure patterns of religious affiliation and settlement, research in this field tends to rely on qualitative methods that engage with the experiences, feelings, and values of people who identify with particular religious or spiritual groups. Geographers of religion regularly employ methods such as focus groups, interviews, ethnography and participant observation, archive research, and participatory action research in their fieldwork. In addition to the complex ethical and methodological issues involved in researching geographies of religion, the issue of the positionalities of researchers in relation to matters of personal faith has been explored in research.

In a recent review of geographies of religion, Kong (2010) observed the significant advancement of the field since the early 1990s, commenting in particular on different sites of religious practice, different sensuous geographies, different constituents of population, different religions, and different scales of analysis. That being said, although geographical studies of religion have become increasingly diverse over the past couple of decades, most research still focuses on the cultural identities and religious practices of Muslims, with less work about Christianity and Judaism and less still about the other world religions or indeed about minority or alternative forms of religion or spirituality. Furthermore, geographies of religion tend to look at particular scales of analysis, especially those relating to cities, neighborhoods, and religious buildings, with some work showing appreciation of the multiple scales on which matters of religion come into play, from the micro- to the macro-geographical. The inclusion of debates about transnationalism, religious conversion, and suburbanization, coupled with a continued focus on the body and processes of embodiment, are useful reminders of the importance of considering the multiscalar significance of religion on social, cultural, and political relations.

SEE ALSO: Cultural politics; Landscape; Race and racism; Spiritual geographies

References


Olson, Elizabeth, Peter Hopkins, Rachel Pain, and Giselle Vincett. 2013. “Retheorizing the Postsecular Present: Embodiment, Spatial Transcendence, and Challenges to Authenticity among Young

Remittances

Philip F. Kelly
York University, Canada

Remittances are generally understood to be financial transfers sent by overseas migrants to households in their place of origin. With increases in the flows of both temporary and permanent migrants around the world, remittances have grown dramatically in the twenty-first century. While individual transfers are usually small, in aggregate they represent flows of resources that can have a significant impact. As a result, the movement, measurement, and consequences of remittances have been closely scrutinized by both development institutions and critical researchers.

In 2013 total migrant remittances to developing countries amounted to $414 billion, and it is these flows that have attracted the most attention. They represent nearly three times the aggregate value of official development assistance, and more than the total movement of private debt financing and portfolio equity investment to developing countries. While the largest total flows of remittances go to more populous countries, such as China, India, and the Philippines, in some smaller economies they represent a very substantial share of the country’s domestic economy. In Haiti, Lesotho, and Nepal, for example, remittances are worth well over 20% of their economy’s gross domestic product.

Remittances are, however, notoriously difficult to measure accurately. For example, money carried back home may be recorded as tourist spending rather than as migrant remittances; and, where gifts are sent instead of cash, they may represent an invisible transfer of wealth or may be recorded as imports. Some estimates suggest that the real magnitude of remittances may be twice the levels that are formally recorded. Exactly where remittances are coming from may also be ambiguous. A remittance agency or an international bank may route money through a third country before it reaches its destination, thereby distorting official records of remittance flows. In the case of systems such as honor-based hawala networks, which are widely used in South Asia and the Middle East, no money may flow across borders at all. And, of course, not all money transfers are remittances from migrants. Counting remittances is not, therefore, a straightforward process.

While remittances are difficult to count, there are nevertheless some precise international accounting standards that define them. Until recently, the definition combined “migrant workers’ remittances,” “compensation of non-resident employees,” and “capital transfers by migrants” (IMF 2009). With the development of a new set of standards in 2009 (known as BPM6), the International Monetary Fund redefined remittances so that the relationships involved, the reasons for sending money, and the duration of migration are no longer considered (IMF 2009). This formal definition now includes:

1 “personal transfers”: including all cash and in-kind transfers between individuals or households (not limited to migrant workers);

2 “compensation of employees” (minus taxes, social contributions, transport, and travel related to employment): referring to the
income of cross-border, seasonal, and other short-term workers who work in an economy where they are nonresident, and the income of resident workers employed by a nonresident entity (such as an embassy or a company based abroad);
3 “capital transfers” of assets between households from one country to another.

Together, these three items constitute “personal remittances.” Beyond personal remittances, international transfers of social benefits (e.g., pension payments) are added to give an amount referred to as “total remittances.” Finally, transfers to “non-profit institutions serving households” (NPISHs) are added to calculate “total remittances and transfers,” thereby including donations from humanitarian agencies, governments, or other organizations to charitable organizations in another country (IMF 2009).

There has been increasing recognition of the importance of remittances among international financial institutions and national governments for two reasons. First, it is clear that, although migrant remittances have always been a feature of international financial flows, they have increased significantly in the twenty-first century. In 1990 total global inflows of remittances were counted at $64 billion (in current dollars). This grew to $124 billion by 2000, and by 2013 total remittance flows had reached $550 billion.

A second reason for increased attention is that remittances are being seen by national and international policymakers as a promising source of development finance and welfare enhancement. In this respect, remittances are presented as having several advantages. First, as they are sent directly to households, they are less prone to capture by corrupt practices or political agendas. Remittances also provide some degree of immunity from the vagaries of business cycles and economic crises, especially in countries with migrant workers placed in very diverse locations and sectors around the world. Some analysts have also suggested that remittances are correlated with welfare gains such as lower infant mortality rates, higher life expectancy, better access to health services, and more equitable income redistribution. While remittances help a country’s balance of payments and strengthen its currency, they do not necessarily enhance a government’s fiscal position. Some governments, however, have sought the direct assistance of overseas migrants (and their descendants) by selling diaspora bonds, which provide low interest loans to a national treasury. In this way, migrant remittances can be used to invest in infrastructure and services. For all of these reasons, there has been a sense of optimism in relation to the developmental promise of migration, and much of the technical and policy discussion has therefore related to ways in which remittance transfers can be made easier and cheaper (World Bank 2006).

At the same time, critical research has drawn attention to the negative implications of a remittance-based development strategy. Some of these arguments concern the wider consequences of migration. For example, migrants with precarious residency status are more vulnerable to exploitation or abuse in their sites of employment around the world than workers with full citizenship rights. Remittances may also represent relations of subjugation for migrants, either because they are repaying creditors back home, or because they are fulfilling gendered expectations to support family members. There are also the social and psychological consequences for children who are left behind when parents migrate to work. The loss of highly qualified human resources – brain drain – has also long been a concern for sending nations. More broadly, it has been suggested that a migration-based development strategy is often a
way for governments to abdicate their responsibility to foster domestic economic prosperity and provide adequate public services.

Even the supposed benefits of remittances may have significant unaccounted costs. First, there is a long-standing debate over whether remittances lead to productive investment or just fuel immediate consumption – meaning that returning workers, and their communities, are no better off when they return than when they left. Second, several studies have shown that beyond the human capital lost to migration, the stay-at-home recipients of remittances may also change their labor market participation behavior (thereby withdrawing even more human capital). Third, it is clear in most contexts that the poorest segments of society generally do not have the resources to migrate and so remittances may actually increase income inequality in sending communities, as nonmigrants are left behind, in more ways than one. Fourth, remittances can have inflationary impacts on the costs of land, housing, schooling, and other everyday expenses in sending areas, thereby rendering migration a necessity rather than an option. These factors have all contributed to what some commentators have described as a new pessimism related to migration and development – in reaction to the optimism noted earlier (Gamlen 2014).

While migrating and remitting are, at one level, individual and familial decisions, they are also very much shaped by states. In some cases, states have actively promoted and marketed their human resources in global labor markets – the Philippines being the best-known example. Even where labor export policies have been less developed, governments have actively cultivated their overseas diasporas as investors, bond purchasers, returnees, and philanthropists (Mullings (2011) provides a Jamaican example). In some cases, governments have legislatively mandated remittances (e.g., the compulsory deferred pay program for miners from Lesotho in South Africa).

Migration and remittance processes have also attracted the attention of nongovernmental organizations and advocacy groups. Among international Filipino migrant groups, “zero remittances days” have been opportunities to express opposition to government practices and policies. Some organizations have also attempted to certify “ethical” remittance agencies that charge fair fees and to advocate boycotting those that are exploitative (www.remit4change.net). Activism around remittances has also taken the form of community-based economic initiatives in which remittances are harnessed for collective rather than private benefit. An example is the development of cooperative enterprises using remittance capital in the Philippines (Gibson–Graham 2005). Remittances are perhaps most effective where collectives such as hometown associations finance the creation of clean water, health care, and education infrastructure.

While the economic significance of remittances attracts most attention, a great deal more than money is usually being transmitted. To those sending and receiving them, remittances may symbolize the obligations and power relations inherent in parenthood, marriage, filial piety, debt repayment, entrepreneurship, and many other social and cultural relations. Similarly, the impact of remittances may go beyond the economic sphere. Remittances may rework power relations (based on class and gender, for example) in the places where they are received, and may have significant impacts on the cultural landscape in both rural and urban areas (Lopez 2015).

SEE ALSO: Domestic workers; International division of labor; Migrant labor; Migration: international
References


Rent gap

Jean-Paul D. Addie  
*University College London, UK*

Rent gap theory is a Marxian explanation of gentrification introduced by Neil Smith in 1979. The rent gap denotes the disparity between the actual ground rent being capitalized on a plot of land and the potential ground rent that could be realized if the site were developed to its “highest and best” use. The theory refers to the value of land (separate from improvements made on it) as appropriated through economic transactions in the form of ground rent. Potential and capitalized ground rents align immediately following the development of a site since land is employed in an optimal manner and intensity. Yet over time, the depreciation of capital fixed in the built environment and shifts in the social or physical condition of the surrounding area may prompt changes in a given location’s highest and best use. As the proportion of ground rent able to be capitalized under current uses diminishes, landlords divest in or abandon their buildings to avoid losses arising from rebuilding or maintenance costs. When capitalized ground rent falls sufficiently below potential ground rent, renewed opportunities for profit-making challenge rates of return available elsewhere and provide incentives for capital to flow back into devalorized inner-city neighborhoods. According to Smith, rent gaps represent a historical discrepancy arising from uneven patterns of investment and disinvestment in the built environment. They are a structural product of capitalist land markets that provide the necessary economic conditions to catalyze processes of revalorization, rehabilitation, and renewal, including gentrification.

The theoretical and empirical validity of rent gaps have been broadly contested, notably in a series of debates between Smith and his critics in the 1980s and 1990s. Chris Hamnett (1991) criticized rent gap theory for dismissing the role of individual agents in shaping gentrification, and reducing demographic factors and structural changes in (postindustrial) employment to consumption preferences. Feminist scholars further pointed to the need to supplant single causal mechanisms with a diversity of processes to explain inner-city redevelopment. Critics, including David Ley and Steven Bourassa, were more critical of Smith’s conceptual foundation, claiming that the rent gap lacked antecedents in Marxist analysis of land rent and failed to offer any significant insights relative to neoclassical land economics. In response, the theory’s proponents asserted that the rent gap concept should not be reduced to a simple deterministic economic model, but rather needs to be contextualized within a more general theory of uneven development (Clark 1995). Smith (1996) refined his account of the rent gap to address the intertwined cultural, political, and economic processes that unfolded along “gentrification frontiers.” His reformulations stressed that housing and other preferences are socially and collectively constructed and expressed by real social individuals. The rent gap does not then determine property development, but reflects ongoing social and political struggles over the appropriation of value from the built environment within capitalist space economies.
REN T G A P

SEE ALSO: Gentrification; Marxist geography; Urban uneven development

References


Representation: 3-D

Sisi Zlatanova
Delft University of Technology, Netherlands

Approaches to 3-D representations have been independently developed in two broad domains: the real world and the design world (Zlatanova and Prosperi 2006). Modeling of real-world objects has developed quite differently with respect to the objects of interest and the associated application domain. Real-world objects above the surface, such as buildings, streets, terrain, rivers, and so on, have been the conventional focus of geographic information systems and represented as 2-D objects (vector or raster) for many years (Longley et al. 2005). Subsurface modeling of geological and geotechnical phenomena has long recognized the need for 3-D representations (voxels and 3-D surfaces). The design world, that is, computer aided design (CAD), computer graphics, architecture engineering and construction (AEC), and building information modeling (BIM), on the other hand, has always dealt with 3-D representations and, naturally, many of the developments are currently being investigated for applications to represent the real world.

Several tendencies can be observed, as follows.

• The previously sharp boundary between real-world and design-world applications is diminishing. All systems, regardless of their origins and foci (GIS, database management systems, BIM, AEC, computer graphics, serious games, virtual and augmented reality), attempt to extend their core functionality to handle 3-D geographical data.
• The demand for 3-D applications is increasing (Billen et al. 2014). Cities grow and the dependencies between buildings, transportation, infrastructure, green and water areas, and so on are becoming more elaborate, which cannot be solved on 2-D maps. Numerous examples demonstrate the demand for 3-D, such as large civil engineering works (Tegtmeier et al. 2014), urban planning, environmental monitoring, and utility management (Figure 1).
• The importance of integrating data from different environments (indoor/outdoor, above/below), applications (cadaster, urban planning, water management), and domains (GIS, BIM) is becoming evident, which results in developing and investigating new approaches and frameworks for integrated data modeling, management, and visualization.
• Technological developments are moving fast: sensors decrease in size and weight and can be mounted on light platforms (such as drones). 3-D sensor measurements are becoming more accurate, cheaper, and faster. 3-D representations of the real world are closely related to the real-world phenomena being modeled. The interest in geosciences has traditionally been directed toward real-world objects with spatial extents, a.k.a. spatial objects. A spatial object refers to a real-world object with geometric and thematic (semantic) characteristics (Aronoff 1995; Longley et al. 2005). Lately, authors have drawn atten-
tion to a third group of characteristics, that is, radiometric characteristics of spatial objects, which have to reflect surface properties of the object. A further distinction is made between real-world objects with respect to discernible (determined) or indiscernible (undetermined) boundaries (Raper 1989). Usually, discernible objects are artificially constructed objects, such as buildings, bridges, and streets, and indiscernible objects are nature objects, such as geological formations.

Another aspect that influences the 3-D representations is the digital visual appearance of the 3-D object. In contrast to 2-D, 3-D visualizations require a certain level of realism, so that users are able to understand the model and the intended message. Digital 3-D models have been traditionally a topic of research and investigation of computer graphics and solid modeling (Mäntylä 1988). One of the main purposes of the 3-D modeling is to achieve maximum similarity to reality. The model may be visualized on a screen as an image or sequence of images (video or animation). The word “scene” is used to denote the rendered imagery on the screen. Several key components are necessary to achieve readable 3-D images on the screen or, in other words, to create realistic scenes: geometry, illumination, shading, texture, and camera position (or view) (Figure 2). These components attempt to capture features of real 3-D objects and represent them in accordance with human perception. 3-D geometry determines the position, shape, and size of the objects. Illumination and shading models control lighting from artificial light sources and corresponding reflections on the surfaces. Colors and textures aim to represent surface properties of the objects. Textures can be artificial or obtained from real-world images. The camera position specifies the location of the user inside the model and orients the model accordingly. The readability of the scene depends on the components used to render the model or part of it on the screen. In this respect, three basic techniques can be distinguished, that is, points, “empty” polygons (wire frame), and shaded (textured) polygons, containing information about the surface texture (or radiometric properties).

For digital representations, a description of real-world objects requires a priori clarification.
Representations: 3-D

of several aspects: (i) spatial subdivision; (ii) geometric primitives to define position/orientation, shape, and size of objects; (iii) levels of detail; (iv) appearance (realism, radiometric properties); (v) topology (spatial relationships); (vi) thematic semantics and attributes; and, last but not least, (vii) operations to be performed on the model.

Conventionally, two approaches to space abstraction are utilized in geospatial data modeling, that is, field oriented and object oriented (Worboys 1995). The field-oriented approach assumes complete subdivision of the space into smaller, often regular partitions, for example, a pixel. In the object-oriented approach, the space is “an empty container” and all the objects are placed (embedded) in it, that is, a house in a 3-D model. Both approaches have advantages and disadvantages and are appropriate for different applications. While the field-oriented approach is better suited for the representation of continuous phenomena, for example, elevation fields and rainfalls, the object-oriented approach represents better discrete phenomena, for example, buildings and roads.

The spatial subdivision and the geometric primitives have a large impact on the final data structure. Traditionally, within GIS, vector and raster approaches are most used. The CAD/AEC/BIM domain distinguishes between larger varieties of 3-D geometric representations, such as boundary representations (B-reps), meshes, constructive solid geometry (CSG), voxels, octrees, nonuniform rational basis splines (NURBS), and so on (Mäntylä 1988). However, they can be also subdivided into two large groups: 3-D vector and 3-D raster representations. These representations have been analyzed for 3-D GIS as well (Latuada 2006; Abdul-Rahman and Pilouk 2008).

3-D vector representations

This group of representations comprises approaches that describe the boundaries of objects. The used primitives are 0-D (points), 1-D (curves), and 2-D (surfaces) (Figure 3). Three large groups of representations can be further distinguished: boundary representations (B-reps), meshes (fishnet or raster representations), and CSG with parametric and freeform curves and surfaces, such as nonuniform rational basis spline (NURBS).

Boundary representation

Boundary representation (B-rep) is a relatively straightforward approach because the visible surfaces of the geographic objects are modeled. This approach comprises most of the data collection approaches, which also measure properties of visible surfaces. Therefore, it has been widely used, applied, and investigated for modeling of real-world phenomena above the surface (built and natural objects). However B-reps can result in very large datasets (in case of high resolution) and very complex data structures if validity and consistency has to be ensured. Many 3-D datasets modeled only with geometric data structures reveal many inconsistencies, which obstruct visualization and performing complicated spatial operations (Ledoux, Arroyo Ohori, and Meijers 2001).
B-reps have been largely employed for geometric representation in GIS. Each primitive is described by low-dimensional primitives subject to certain rules. For example, a curve is described by two points, a surface is described by three or more curves, and a solid is described by three or more surfaces. The primitives are utilized to describe objects such as buildings and streets. The objects are usually embedded in the 3-D space, that is, full subdivision of space is assumed rarely. A number of rules are applied to the geometric primitives and to the way they are combined to define the geographic object. Examples of such rules are: a curve (also called an arc) can be restricted to be the straight line between the two points or curves are not allowed to intersect. In the case of intersections, the curve has to be split. The two points of a line must not have the same coordinates. A surface (polygon) can be restricted to be composed of only three straight lines (i.e., triangle), or otherwise restrictions need to be imposed to ensure that it is planar. The order of the curves has to be specified to be able to define the orientation of the surface. The three points of a triangle must not be on one line. A polyhedron (a 3-D object) must have the following characteristics (Arens, Stoter, and van Oosterom 2005): (i) flatness – it consists of only planar polygons on the surface; (ii) all surface polygons together constitute only one 2-manifold, so that the polyhedron bounds only one volume; (iii) simplicity – tunnels are allowed but no inner rings, a line has only two points and all polygons must have area; and (iv) orientability – inside and outside must be specified. Ledoux, Arroyo Ohori, and Meijers (2014) provided tools for checking and repairing polygons.

The rules and relations between the primitives can be specified in a geometrical or topological data structure (see the examples of 3-D geometrical data structures shown later). The data structures take account of the space subdivision, define primitives, apply geometric rules, and explicitly store (or not) relationships. The difference between the two types of structure is that geometrical data structure coordinates define locations of individual primitives, while topology data structure maintains the spatial relationships among the primitives (Bilen and Zlatanova 2003). The geometric representation facilitates spatial indexing and metric operations and, hence, is commonly used in spatial data management. Topological structures, on the other hand, ensure object validity on editing and data manipulation as well as reducing data duplication. Development in boundary representation with rules for geometrical and topological structures is being extended to 3-D data models and data handling (Breunig and Zlatanova 2011).

The Open Geospatial Consortium (OGC; www.opengeospatial.org) has developed two standards for the representation of geometrical structures: abstract specifications (Topic 1 Feature Geometry, also ISO 19107) and implementation standards. The abstract specifications provide a guidance to what geometric primitives should be used for representing geographical features (Figure 4). The implementation specifications for simple feature access (SFA) are described for three different platforms: structural query language (SQL), common object request broker architecture (CORBA), and object linking and
Figure 4  OGC geometry package (ISO 19107, Feature Geometry; www.iso.org). Reproduced by permission of OGC.
embedding (OLE)/component objects model (COM). In 1999, the first implementations of the structural query language/spatial feature standards (SQL/SFS) became available; these marked a significant step forward in maturing spatial database management systems (DBMS). Currently, almost all commercial DBMS platforms support spatial data types, that is, Oracle, PostGIS, Ingres, MonetDB, Informix, IBM DB2, as most of them adopted OGC standards. Figure 5 shows a SQL geometric feature hierarchy of simple primitive points, curves, and surfaces according to OGC SQL/SFS. As can be observed, a much richer set of geometric primitives, including 3-D shapes (solid, cylinder, sphere, spline), is proposed in the abstract specifications (Figure 4).

Following the implementation specifications for SQL, two ways of storing geometries in DBMS can be distinguished: using the polygon and multipolygon data types (Figure 6). In the first case, a 3-D object will have as many rows as the number of polygons that constitute the 3-D object. The table for volumetric objects represented by tetrahedrons will be simpler and can be normalized. It can be organized as a table with five columns: one for the ID of the tetrahedron and four for the composing polygons (triangles). In general, such an organization can be seen as a partial topological model, as the 3-D object can be defined in reference to the composing polygons, which can be organized in a separate table.

The query shown in Box 1 retrieves the geometry of Polyhedron 1, Face 1 as stored in PostGIS. When multipolygons are of the data type to represent 3-D features, a 3-D object is stored in one row with its object ID and geometry (represented by one multipolygon). This case allows management of only one table, which will be composed of two columns: the ID of the object and the GEOM (an attribute field for geometry) for the spatial data type multipolygon. An apparent advantage of the 3-D multipolygon approach is the one-to-one correspondence between a record in a relational table and an object.

Presently, only a few DBMS platforms (e.g., Oracle Spatial, PostGIS) include a volumetric data type. The challenges to the type of volumetric objects, its validity, and the functions that have to be performed on it are numerous. A simple volumetric object can be represented by polyhedron, triangulated polyhedron, and tetrahedron, all of which can be realized with provided spatial data types of polygons and multipolygons (Zlatanova 2006). Moreover, there is no practical difference in the implementation of the polyhedron and triangulated polyhedron, since a separate triangle data type does not exist. Tetrahedrons would allow for a slightly simpler representation than polyhedrons or triangulated polyhedrons, as a tetrahedron has only four triangles.

User-defined spatial data types can be implemented using different approaches from the simple SQL create data type statement, to more complex implementations, using a procedural language (PL), Java, C++, and so on. The common drawback of such an implementation is that native spatial functionality (operations and indexing) of DBMS cannot be applied to user-defined spatial data types. Moreover, the user-defined spatial data types cannot be stored in the same column of the natively supported spatial data types. Visualization for front-end applications would be possible only by developing individual connections. User-defined spatial data types, nevertheless, are very useful for prototyping of new concepts. For examples Arens, Stoter, and van Oosterom (2005) showed implementation of polyhedron data type. Penninga and van Oosterom (2008) proposed an implementation of a tetrahedron data type.
**Box 1** Retrieval of the geometry of Polyhedron 1, Face

```sql
select asewkt(geom) from polyhedron1 where idb=1 and idf=1;
asewkt
POLYGON((172.578 290.563 7.647,
174.891 288.281 7.573, 178.625 288.438 7.552, 181.016 291.156 7.574,
180.75 294.625 7.586, 178 296.938 7.582, 174.563 296.781 7.586,
172.234 293.906 7.65, 172.578 290.563 7.647))
(1 row)
Time: 0.782 ms
```

**Constructive solid geometry**

Constructive solid geometry (CSG) has been seen by many as a modification of vector representation (Latuada 2006) because it utilizes parametrized primitives, such as cube, sphere, cylinder, pyramid, and so on, which are essentially described by their outer shell. Boolean operations (union, intersection, difference, complement, and cutting) may be performed on the selected initial primitives until the desired shape of the object results. A CSG tree is constructed with primitive solids at the leaves, operations at the inner nodes, and the final object at the top level. This approach to 3-D representation

---

**Figure 5** SQL geometry feature hierarchy (OGC Implementation Specification for Geographic Information, Part 2: SQL Option; www.opengeospatial.org/standards/sfs). Reproduced by permission of OGC.
Figure 6  Examples of storing of polyhedron/tetrahedron. Left: Multiple polygons (each polygon is a face) are used to represent a 3-D polyhedron (same idb) with coordinates in different lengths for each polygon (geom1, geom2, etc.); hence, each polygon with a string of $x$, $y$ coordinates is stored in one row in the DBMS. Center: When all faces are triangles, all polygons have three pairs of $x$ and $y$ coordinates; therefore, all polygons can be stored in an array of coordinate sets of three for all triangles in one row in the DBMS. Right: one multipolygon is used to represent a polyhedron; each geometrical object (e.g., geom1, geom2, etc.) corresponds to a set of multiple polygons.

has many advantages: the primitives are simple volumetric objects; the operators, such as drilling, cutting and gluing, are applicable to the CSG tree; and volumes can be easily computed. This modeling approach has been widely used in computer-aided manufacturing to construct machine components, for example. The CSG approach has also been used in software design and, specifically, in BIM (Revit, SketchUp). Many building components (e.g., walls and windows) are structured following the CSG concept.

CSG is of limited use to modeling of real world phenomena. While CSG is easy for design, the modeling of real-world objects is quite tricky and not intuitive. The shape of a real-world object has to be approximated and subdivided into predefined primitives. Only regularly shaped objects, such as buildings, might require less effort (e.g., CityGML LOD1 and LOD2). However, representation of a whole city will result in a complex CSG tree. A positive aspect of this way of modeling is that all objects must be represented by solids as in reality. Real-world
objects that are currently modeled by 2-D primitives (streets, digital terrain models) or even 1-D primitives (cables and pipes) can naturally evolve to 3-D objects. Research and development are needed to establish appropriate data management of CSG trees in DBMS. Alternatively, robust operations are needed for conversion of CSG trees to B-reps, and the data types currently available in DBMS (Figure 7). As mentioned previously, the OGC abstract specifications suggest that parametric shapes be considered for implementation, but whether the CSG tree should be stored in the database remains a question.

The first step towards utilizing CSG for modeling geographical objects is introducing parametric shapes such as cylinders, cones, and spheres. Such a representation is sufficient for performing numerous analysis but faces challenges when the depth has to be taken into consideration or they have to be visualized in 3-D environments, for example, in utility works. Three-dimensional lines alone may not be sufficient for 3-D visualization because they lack the volumetric appearance that is required to produce depth perception. Several approaches have been proposed in the literature. One approach is to model pipelines as 3-D lines in the database and only create their 3-D volumetric representation for visualization. Another approach is to store the centerlines and geometry of pipeline networks with parametric representation or their real 3-D representation. While all approaches have advantages and disadvantages, a smooth transition to manage real 3-D pipeline systems depends on the availability of parametric data types and corresponding spatial operations (Döner et al. 2011). Experiments with 3-D representation of pipelines have revealed that many intersection errors can easily be detected (Figure 8).

Figure 8  3-D visualization of underground pipes: intersection of two pipe segments (left) and pipes beneath the surface in the vicinity of silo foundations (right).

Freeform curves and surfaces

Another approach to 3-D representation is utilizing freeform curves and surfaces. There are several methods to represent freeform curves and surfaces. Bézier, B-spline, and NURBS methods are among the most commonly used in practice. They are all represented by parametric functions (Piegl and Tiller 1997).
Parametric functions have several advantages over triangulated and polygonal surfaces (meshes): parametric functions have more degrees of freedom to model shapes than predefined shapes; points on a curve or surface can be evaluated numerically and accurately; and they allow for more compact representations than meshes. The simplest of the three methods to represent a freeform curve is a Bézier curve. Its shape is defined by a sequence of \( n + 1 \) control points \( P_i (i = 0 \ldots n) \) in 3-D space. A Bézier surface is, similarly, defined by a grid of \( (n + 1) \times (m + 1) \) control points \( P_{ij} (i = 0 \ldots n, j = 0 \ldots m) \). To be able to keep the degree of the curves low (3 or 4) the objects are usually modeled by Bézier patches. Such an approach cannot ensure the smoothness at the edge of two patches. B-splines overcome this limitation, because the degree of the curve can be defined independently from the number of control points. A B-spline curve of degree \( d \), is defined by a sequence of \( n + 1 \) control points \( P_i (i = 0 \ldots n) \) in 3-D space, and a knot vector of \( m + 1 \) knots such that \( m = n + d + 1 \). Though Bézier and B-splines are widely used representations, the most popular method currently for representing freeform shapes is the NURBS method (Figure 9, right). The main difference between NURBS and B-splines is that the control points of a NURBS have weights that give extra degrees of freedom in modeling curves and surfaces. The important properties of NURBS curves include the following.

- A NURBS curve is a piecewise rational polynomial curve and has the same continuity conditions at knots as a B-spline curve.
- NURBS curves are projective invariant, that is, affine and projective transformations can be applied to the control points.
- NURBS curves can represent conic sections, such as circles and ellipses; that is, they can be used to represent parametric shapes.
- NURBS curves are, just like B-splines curves, locally modifiable and contained within the convex hull of their control points.

OGC has recognized the importance of freeform curves and surfaces and has included them in the abstract specifications, but currently no GIS software or DBMS supports them. They are mostly applied in the design phase. When the construction is built, the surfaces are modeled as

![Figure 9](image-url) A pavilion modeled with NURBS: building (left) and 3-D model (right).
simple surfaces and stored as meshes (utilizing the available data types) in the database. Zlatanova, Pu, and Bronsvoort (2006) demonstrated that NURBS can easily be stored, visualized, and analyzed at DBMS level (Figure 9). A range of questions have to be investigated. For example, the validation rules for freeform curves and surfaces have to be further specified. The present validation functions follow the mathematical definitions of freeform curves and surfaces. The function AnyIntersect, for example, has been implemented using the convex hulls of the control points of two shapes to identify intersections. Opportunities remain for improved accuracy in determining intersections of 3-D shapes. NURBS have been used in AEC and for approximation of underground geological formations. A broader implementation in GIS will improve 3-D GIS functions and utilities.

3-D raster representation

Another large group of 3-D representations is based on exhaustive enumerations: voxels, polyhedrons, tetrahedral, octrees, and so on (Figure 10). This approach assumes complete subdivision of space into equally shaped (voxels) or irregular shapes (tetrahedrons, polyhedrons) 3-D units. Spatial occupancy approach is appropriate for continuous phenomena, mostly natural objects such as geological formations, marine and atmosphere phenomena. As rasters in 2-D representation, voxels are popular in 3-D. Each voxel has distinct properties with which 3-D objects can be created based on a function of properties such as salinity, temperature, or pollution parameters.

Voxel representation has the advantage of using a simple primitive and simple data structure. The space subdivision schema can be optimized using octree and binary decompositions. Such schemas allow for flexibility of voxel sizes to account for local or regional homogeneity. In addition, voxel representation offers a convenient data framework for computing, such as calculating volume, intersections, and neighborhood analysis. However, they can result in high volumes of data, which need efficient management and indexing. Furthermore, the surface of the voxelized object can be rough, and hence cannot be textured with photo-images (as widely applied in B-reps). Tetrahedrons and polyhedrons are alternatives for better modeling of variant surfaces or boundary objects, albeit at the cost of trading the regularity of voxel data structures.

As mentioned previously, many natural phenomena have been modeled with 3-D rasters (i.e., voxels). Figure 11 shows the 3-D subsurface model of The Netherlands in a voxel representation. It is a nation-wide model covering 41 000 km² and the upper 500–1000 m of the subsurface. The size of the voxels is 100 × 100 × 0.5 m³. The information recorded per voxel describes stratigraphical unit, lithology, grain-size and hydrology, physical and chemical properties as well as associated uncertainty. This dataset is used for a number of applications, such as groundwater and pollution management, land subsidence studies, and infrastructural issues.

Voxels are commonly used to facilitate the processing of point clouds. Voxelization processes reduce the size of the point clouds and, at the same time, introduce regularization for ease of analysis. For example, Figure 12 shows a voxelized point cloud for the Anne Frank tree in Amsterdam, The Netherlands. The coloring of the tree (right) is based on the number of points per voxel. The higher the number of points, the greater the chance is that the voxel represents leaves. The voxels with the highest number of points are, hence, colored green. The result appears a good match of reality.
In recent years, voxel representations have attracted increasing attention due to their simplicity and flexibility in creating representations of different resolutions. Artificially-constructed objects, such as buildings, streets, and bridges, are increasingly being modeled with voxels in science, engineering, and entertainment. For example, Minecraft, the game, originally designed for children, is currently being investigated for serious applications in planning and public participation. The game has two modes: survival and building. The building mode allows a complete world to be designed with blocks of $1 \times 1 \times 1$ m$^3$. Figure 13 portrays a Minecraft model of the old city of Delft, The Netherlands.

Voxels can be created from existing 3-D vector models as well. However, 3-D vector-raster conversion requires accurate representations of geometry, topology, and semantics. Geometrically, voxelization algorithms should be able to deal with points, curves, surfaces, and solids. The semantics of the objects depends on...
the applied semantic model, such as the most commonly used CityGML. Figure 14 illustrates a complete voxelization of a building. Voxels are created from a 3-D vector of the model and the semantics of walls, floors, windows, stairs, and air is preserved.

The neighborhood relations in the vector domain are represented by the boundaries that completely enclose the objects. Since voxels do not represent boundaries explicitly, the closure of an object is determined by its neighboring voxels. The neighboring voxels give an indication of connectivity. Each voxel has six neighboring faces, 12 neighboring edges or eight neighboring vertices (Rosenfeld 1981). This results in six faces, 18 edges and faces, and 26 faces and edges and vertices connectivity in 3-D.

The type of connectivity relates to the shape of the voxelized object as well as the topological relationships between the objects. Figure 15 illustrates four cases in which the connectivity
may lead to disconnecting or penetrating objects in 2-D raster: 8-connectivity is favorable for crossing lines (a), but not for touching between a line and a polygon (b); on the contrary, 4-connectivity is favorable for touching between a line and a polygon (d), but may lead to discontinuity of line in case of crossing lines (c).

One of the best approaches for vector to voxel conversion is the topological method (Laine 2013). The method is intersection-based, which sets a voxel at an intersection target. A voxel’s intersection target is a spatial subset of the cubical space occupied by the voxel such as the quadrilaterals bisecting the voxel along its three axes. Intersection targets are chosen based on the desired connectivity level and the dimensionality of the input vector object. Since the proposed method is intersection-based, computational efficiency would be an issue.

3-D semantics

The word “semantics” is used to describe the meaning of things. Semantics can have different interpretations: as opposite to geometry (vector/raster), as a formal connection to reality (nature-oriented) or connection to usage (value-oriented) (Billen et al. 2014). In the text here, semantics is discussed in the context of reality, that is, the meaning of data objects representing geographical objects in an application.
model. Application models can emphasize different aspects of a geographical object, which can lead to associating a real-world object with different semantics. For example, a building can be associated with a building, settlement, construction, house, office, and so on in different models. The semantics used in 3-D city modeling is compatible with the semantics used for 2-D topographic maps. Few thematic semantic models for 3-D cities exist. A common understanding is that buildings and terrain objects are the most important features to describe. Therefore, the following are top-level object classes: buildings, streets, green areas, public areas and terrain surface or terrain, vegetation and built forms.

Models can be purely semantic (definitions and relations between the objects on the basis of a theme, e.g., geography) or might be related to spatial, topological, and appearance properties. For example, the vocabulary or taxonomy referring to spatial data types (e.g., polygon, multipolygon, solid) as well as topological relationships (intersect, overlap, meet) is also a kind of semantics. Many ongoing standardization initiatives (i.e., OGC, ISO) define semantics of geographical objects with respect to their spatial, topological, appearance, and thematic properties (geometric representation, time, accuracy). Examples of such standards are spatial schema (ISO 19107:2003, 2003), temporal schema (ISO 19108:2002, 2002), quality principles (ISO 19113:2002, 2002), quality evaluation procedures (ISO 19114:2003, 2003), metadata (ISO 19115:2003, 2003), schema for coverage geometry and functions (ISO 19123:2005, 2005), and the XML geometry encoding GML (ISO19136).

In 3-D modeling, semantics is commonly used to stress the meaning of the objects and not the meaning of geometric, topological, and appearance constructors. A number of semantics models have been defined for 3-D city modeling. Townontology is an example of ontology that deals with geographical (named physical), socioeconomical, and mental objects in an urban system with part-of relationships. It is possible to observe the first two levels of physical objects taxonomy (Caglioni 2006).

The INSPIRE “generic conceptual model” (http://inspire.ec.europa.eu/index.cfm/pageid/2) is another example of defining the semantics of objects with considerations of different application models called themes. INSPIRE has 34 different spatial data themes, which cover natural, built, administrative, and environmental objects. The nine themes (“Annex I,” completed by the end of 2009) has an elaborated semantics but is not specifically related to 3-D. One exception is the theme Cadastral Parcels, which refers to 3-D cadastral objects. Annex II (e.g., Elevation and Geology) and Annex III (e.g., Soil, Buildings, Atmospheric conditions, Oceanographic geographical features, and Energy resources) themes are being specified to contain more explicit reference to 3-D aspects. For example, the draft data specification on energy resources (INSPIRE TWG ER) mentions 2-D, 2.5-D, or 3-D geometries.

One of the most prominent semantic models for 3-D geographic objects is CityGML (Gröger and Plümer 2012). Work on CityGML was initiated by Special Interest Group 3-D, Germany, in 2002 and in 2008 it was adopted as a OGC standard. Since that time, many cities in Germany, France, the Netherlands, Switzerland, Turkey, Japan, and Qatar have been using and experimenting with CityGML to build their spatial data infrastructure or city models. Vendors such as ESRI, Bentley Systems, Autodesk, and FME provide import/export modules. A large number of 3-D viewers are also available: KIT Karlsruhe, BS Contact from Bitmanagement
Software, Aristoteles from the University of Bonn, LandXplorer of Autodesk, and so on.

CityGML maintains information about spatial, semantic, appearance, and topological properties of the objects. Semantics is defined for sites (buildings, tunnels, and bridges), transportation facilities, water bodies, digital terrain models, vegetation (areas, volumes, and solitary objects with vegetation classification), land-use type, city furniture, and generic city objects. The geometric representation is typical boundary representation, which follows the GML 3-D geometry model (based on ISO 19107 Spatial Schema). CityGML is also a multiscale model with five predefined levels of detail (LOD): LOD0 – regional landscape; LOD1 – city and region; LOD2 – city districts and projects; LOD3 – architectural models (outside) and landmarks; and LOD4 – architectural models (interior). The LODs have definitions referring to semantic and spatial properties of objects. In this respect, they can be seen as a vocabulary for LODs. The definition of buildings is most developed and tested among all definitions (Figures 16 and 17). The semantic hierarchy goes to windows, doors, and furniture. The buildings (or parts of buildings) can be represented by multiple geometries (solids or multisurfaces), depending on the purpose of the application. If individual texturing is applied for each surface, multisurface is recommendable, otherwise solid should be more appropriate.

As illustrated in Figure 17, the buildings LOD1 are “buildings”; LOD2 semantics is extended with “ground surface,” “wall surface,” and “roof surface;” LOD3 adds “window” and “door”; LOD4 includes “room,” “ceiling surface,” “interior wall surface,” “floor surface,” “closure surface,” “door,” “window,” “building furniture,” and “building installation.” Depending on the LOD, a voxel of a given size can have several different semantic tags, such as (building, roof), (building, room), (building, window), and so on. For example, the “wall surface” might need to be replaced with the notion of a wall between two rooms. The “room” is then represented by all voxels that can be placed between the walls.

As mentioned previously, a geographic object can be associated with different semantics. Here examples are given with the 3-D model of a building interior. It is differently modeled in three international standards: CityGML, Industrial Foundation Classes (IFC), and IndoorGML.

In CityGML, the building interior is represented in a similar way to the outdoor concept; that is, the visible parts of the interior are represented. WallSurface, CeilingSurface, InteriorWallSurface, and FloorSurface are identified per room and are represented by surfaces. A building is composed of Rooms; a room is composed of the surfaces representing the visible part of walls, ceiling, floor, doors, and windows. However, there is no explicit notation for story floors. IFC and IndoorGML take different approaches.

IFC is the mainstream standard of building information modeling (i.e., ISO PAS 16739, 2005). In contrast to CityGML, IFC centers on the key building elements (or construction objects), including beams (IfcBeam), columns (IfcColumn), walls (IfcWall), slabs (IfcSlab), and stairs (IfcStair) (Figure 18). A building is composed of story floors. Some building elements (i.e., walls, slabs) may contain openings (i.e., voids or holes). All objects are represented by constructive solid geometry or sweeping primitives. Boundary representation is used very rarely for individual elements such as stairs. The spatial relations are preserved in the form of spatial hierarchies with (spatial) relationship classes. IFC models a wall as a solid (physical wall) between two rooms. Therefore, the notation for a room does not exist. IfcSpace was introduced recently to represent the volume occupied by a space.
Figure 16  UML diagram of CityGML’s building model. Courtesy OGC CityGML. Reproduced by permission of OGC.
IndoorGML aims to provide a semantic, geometric, and navigation model for indoor navigation systems. IndoorGML takes an alternative view to IFC in the duality graph of indoor spaces and considers that spaces (instead of walls and other solid objects) are the most important building elements, which are used to automatically derive network for navigation. The interior of a building is subdivided into appropriate spaces, with respect to the user’s navigation profile (e.g., visitor, woman, or modes of movement) and the environmental characteristics (e.g., lights, crowds, renovation) of the interior. This subdivision can vary: some spaces can be united in larger units, for example, to represent areas which are not of interest for a visitor, or can be subdivided into smaller units to indicate some functional areas as “coffee corner” or “registration area.” The spaces are volumetric objects, conceptually very similar to ifcSpace of IFC. Regardless of how the spaces are defined, semantically they are distinguishable only with respect to navigation purpose into TransitionSpace, ConnectionSpace, GeneralSpace, and AnchorSpace (Figure 19). The spaces should not overlap. The connection space/connection boundary (i.e., doors) is used to derive the logical or metric network.

Indoor empty space (free air in a room) can be modeled as GenericSpace or TransitionSpace (in IndoorGML), ifcSpace (in IFC) and Room (in CityGML LOD2). The three representations differ geometrically and topologically and should be taken into consideration when modeling interiors. CityGML, IFC, and IndoorGML are also intended for different purposes. CityGML is intended to represent the real world as built,
IFC is intended predominantly for design, and IndoorGML is dedicated to a specific application (i.e., indoor navigation). Therefore, the semantics differs, which influences also the geometric representation of the topological relationships between objects.
3-D topology

Topological data structures are often referred to as B-reps. Topology is one of the mechanisms to describe relationships between spatial objects. Models utilizing the topological properties of spatial objects are usually called topological models and are considered best suited to maintain consistency and validity of objects as well as for performing certain type of operations. As mentioned previously, B-reps require strict rules to maintain consistency of data. They depend on space partitioning, the types and number of primitives, the explicitly stored and derived relationships, and the constructive rules. While in 2-D, several topological data structures have been developed and widely used (wing edged, wheel-chain), the 3-D data structures vary greatly and no 3-D topological structure has been implemented. OGC suggests a very general concept of topological data structures but no implementation specifications. The research in the last three decades has contributed to a number of 3-D topological data structures, such as the 3-D formal data structures (3-DFDS), unified data model (UMD), simplified spatial structure (SSS), tetrahedral network (TEN), and combinatorial maps. All data structures have advantages and disadvantages with respect to certain applications. Compared to geometry, topological structures are slower in metric operations, such as computations of area, length, and volume, or for visualization because the coordinates are stored with the nodes. Each metric operation will invoke a search of low-dimensional objects until the nodes with the coordinates are reached.

3-DFDS is the first data structure that aims at storing real-world objects and their spatial relationships. The model consists of three fundamental levels: feature (related to a thematic class), four elementary objects (point, line, body, and surface), and four primitives (node, arc, face, and edge). The model follows the concept of full space subdivision into nonoverlapping objects. Several relationships, that is, node-on-face, arc-on-face, node-in-body, and arc-in-body are explicitly stored. A number of rules specify the validity of objects and relationships. The data structure has been tested in research and has exhibited complexity that is difficult to maintain. For example, the faces must be planar, and the explicit relationships should be properly organized.

The tetrahedral network (TEN) was introduced initially by Pilouk in 1996 (Abdul-Rahman and Pilouk 2008) and later modified by Penninga in 2008. TEN has four constructive objects (tetrahedron, triangle, edge, and node). An ARC table stores arc-node relationships; a TRIANGLE table contains tetrahedron-triangle-edge links. A body object is composed of tetrahedrons, a surface object of triangles, a line object of arcs, and a point object of nodes. The general rule for creating the model requires that each node is a part of an arc, each arc is a part of a triangle, and each triangle is a part of a tetrahedron. Singularities are not permitted. This data structure simplifies significantly the 3-D representations, assuming that the world is partitioned into tetrahedrons and each object (its boundary) is embedded in the tetrahedronization. The first implementation of this data structure is in a relational data table. A later implementation of Penninga’s TEN model considers mixed geometry/topology description, keeping the coordinates of the primitives as attributes (Penninga and Oosterom, 2008). This representation speeds up the search and query of objects and allows spatial indexing to be applied.

The above mentioned models maintain all primitives, that is, node, edge/arc, face, and body. To improve model performance, several researchers suggest skipping maintenance of arc. Two data structures, the simplified spatial structure (SSS) and the urban data model
(UDM) have clearly demonstrated that search and visualization operations are significantly expedited. While SSS allows polyhedral representations, UDM assumes only triangles to represent the boundaries of all objects. These two data structures are quite similar to the most data structures used for meshes in CAD (e.g., Rhinorous). Data structures without explicit maintenance of arcs (while they still can be implicitly derived) will result in some deficiencies for operations on arc (e.g., “shortest path” operations). A number of alternatives have been developed to overcome the deficiencies to metric operations. However, the tendency is the same, that is, improving the performance of topological structures by either reducing the number of primitives or storing the coordinates as attributes of the primitives. More models are discussed in Zlatanova, Rahman, and Shi (2004).

No one data structure can account for the wide variety of 3-D objects and, hence, cannot be used for all kind of 3-D applications. The topological structures depend on many different factors: dimension of the embedding space (2-D, 2.5-D, or 3-D, time), used topological primitives (node, edge/arc, face, or body), and explicit or implicit relationships, topological rules (crossing, on, or in). Oosterom et al. (2002) proposed special encoding of the topological data structures similar to geographical coordinate systems; an example of such a coding is shown in Table 1. Each topological structure obtains a three-digit code (topological type) and six parameters, that is, dimension, primitives used, topological tables, explicit relationships, number of tables, and rules, which aim at providing all needed metadata. For example, TIN has a code 221, which means is 2-D and all the information can be organized in two relational tables.

Hybrid representations

As shown already, the 3-D geometry can be vector based or raster based but it could be combinations of the two. The interest in hybrid (vector-voxel) representation is becoming increasingly important. Seamless conversions between the two will bring benefits to the type of operations that can be performed. Voxel representations are favorable for volume computations and k-nearest neighbor operations. Vector representation are better for fast, realistic visualization, although much research is completed currently on visualization of (colored) point clouds. Figure 20 illustrates voxel and vector representation of the same building including interior rasterization. The voxel representation is obtained from the vector (CityGML LOD4), preserving the semantics of the original object. The semantics is added as an attribute to the description of each voxel. In this way, it is possible to select and visualize all the voxels with a tag “wall” (Figure 20, left).

The voxel representation allows for quick volume computations (rooms, walls) and various space-based simulations, for example, 3-D routing, wind/air quality analysis. The vector representation is more convincing for walkthrough, surface computations. The general performance can be improved if a hybrid model is used. Operations can be executed at the representation, which requires less computational complexity. The conversion between the two representations can be performed on the fly if parallel maintenance is not beneficial. Such a flexible voxelization allows work at various resolutions. Figure 21 illustrates the vector model and derived 3-D rasters with different voxel size. It clearly shows that objects with crisp boundaries (e.g., the house) need very high resolution to achieve acceptable visual appearance. In contrast, the shape of objects with undefined boundaries,
<table>
<thead>
<tr>
<th>Topol. type</th>
<th>Dimension</th>
<th>Primitives used</th>
<th>Topological tables</th>
<th>Explicit Relationships</th>
<th>All tables</th>
<th>Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIN</td>
<td>2D</td>
<td>Node, edge</td>
<td>node, edge</td>
<td>no</td>
<td>2</td>
<td>Planar Partition</td>
</tr>
<tr>
<td>Wing-edge</td>
<td>2D</td>
<td>Edge, face</td>
<td>edge, face</td>
<td>no</td>
<td>2</td>
<td>Planar Partition</td>
</tr>
<tr>
<td>Wheel (chain)</td>
<td>2D</td>
<td>Edge, face</td>
<td>edge, face</td>
<td>no</td>
<td>2</td>
<td>Planar Partition</td>
</tr>
<tr>
<td>3DFDS</td>
<td>3D</td>
<td>node, arc, edge, face</td>
<td>arc, edge, face</td>
<td>node-on-face node-in-volume arc-partof-line arc-on-face arc-in-volume</td>
<td>8</td>
<td>Space Partition</td>
</tr>
<tr>
<td>TEN</td>
<td>3D</td>
<td>node, arc, triangle, tetrahedron</td>
<td>arc, triangle, tetrahedron</td>
<td>tri-partof-surf arc-partof-line</td>
<td>5</td>
<td>Space Partition</td>
</tr>
<tr>
<td>Cell-tuple</td>
<td>3D</td>
<td>0-cell, 1-cell 2-cell, 3-cell</td>
<td>cells</td>
<td>no</td>
<td>1</td>
<td>Space Partition</td>
</tr>
<tr>
<td>SSS</td>
<td>3D</td>
<td>Node, face</td>
<td>face, line surface, volume</td>
<td>node-in-volume face-in-volume</td>
<td>6</td>
<td>Space Partition</td>
</tr>
</tbody>
</table>

Source: Oosterom et al. 2002.
such as trees, could be better represented. Some tiny objects (such as the windmills behind the house on the vector model) will disappear.

Current status and future research

This entry has provided an overview of 3-D representation and models developed to account for 3-D geometry, semantics, and topology. Presently, there is no single 3-D model that can support all applications. The models are still being adapted, extended or newly designed to address specific application needs or for general purposes (e.g., developing of a national 3-D topographic standard). This process affects all components of the 3-D representations in a variety of combinations. Most commonly, the semantics is adapted and the geometry is kept unchanged, such as extensions of CityGML and IFC (Tegtmeijer et al. 2014; Hijazi, Ehlers, and Zlatanova 2012; Isikdag, Zlatanova, and Underwood 2013; van de Brink, Stoter, and Zlatanova 2013). Geometric representation might be also changed but this is mostly for temporary purposes, such as integrating B-reps and 3-D raster. Very often 3-D topologically-structured data are converted to nontopological data but the semantics is preserved. Nevertheless, a strict internationally accepted mechanism for using, adapting, and creating new concepts for digital representation of the real world is needed. Figure 22 schematically illustrates linking concepts from different models. To facilitate this process, definitions, validity rules, and relations between objects in individual model have to be further clarified and strengthened in the description of the standards.

The discussion about which component of 3-D representation should be considered primary has been going on for decades. The most commonly accepted vision is that semantics is the very critical component, followed by geometry and topology. A general workflow for 3-D data modeling is outlined here.

- Semantic definition: investigate which objects are needed, including the required resolution, accuracy, and visualization; apply existing definitions and vocabulary as much as possible; introduce new semantics specifications only when no suitable semantics is available.
- Geometric definition: decide which representation is more appropriate and, if needed, apply hybrid models or on-the-fly conversions; consider which representations
are supported by DBMS. GIS-based software vendors and mainstream DBMS do not support parametric representations, CAD/AEC-based software vendors rely on them heavily.

- Topological definition: decide which extended 3-D topological data structure is needed; investigate options for building topology to check validity of objects or perform certain operations only; consider computational geometry libraries such as CGAL (the Computational Geometry Algorithms Library), which provide robust 3-D operations on polyhedra (e.g., Nef-polyhedra) (http://www.cgal.org/)

To conclude, 3-D modeling is entering a new phase that requires more attention and cooperation from different disciplines than ever before. The ultimate goal is to create a set of different core 3-D representations and provide tools for mapping, integration, and exchange. The first steps in this direction are being made by the many international organizations that created CityGML and INSPIRE, CityGML (OGC) and IFC (BuildingSMART), and Web3-D. Researchers face many new challenges in developing data structures, theoretical frameworks, and algorithms. New domains and cross-domains are emerging as well. A typical example is 3-D indoor modeling. Research in support of modeling interiors has been active for over thirty years but it has been mostly for design purposes. Governmental and commercial enterprises as well as individuals are beginning to apply indoor models in their business processes and applications. Is the technology developed for the outdoor world readily applicable for indoor? 3-D representations should be advanced to allow seamless transitions between indoor and outdoor spaces and between different application domains.
Acknowledgment

The author is grateful to May Yuan for valuable comments and edits.

SEE ALSO: Data model, object-oriented; Data structure, raster; Data structure, vector; Ontology: domain applications; Representation: geographic systems; Representation: indoor spaces; Spatial database; Topological relations

References


Further reading

Representation, including the presentation of spatial patterns in maps and diagrams, is a core concept in the discipline of geography’s history and is particularly central to the history of cartography, which has long struggled with questions of how and in what ways to represent the world. This means that representation is also one of the most contested concepts in the discipline. Geographers, cartographers, and GIScientists employ very different theoretical and methodological approaches to the organization and meanings of geographic data, the representation of patterns and processes across geographic spaces, and the analysis of cartographic practice. All of these occur within larger conversations about the history of ideas in the discipline. Today, geographers do not employ one definition of representation, but instead offer up multiple ways of knowing and thinking about representations of space and place. This is particularly true within cartography, which remains wedded, on the one hand, to a scientific approach to representation, and, on the other hand, offers numerous challenges to that same scientific cartographic practice. Given the complexity of how representation is discussed in the discipline, this review sketches a variegated history of this concept and examines the multiple trajectories that define representation in the discipline today (Edney 2012).

Scientific cartography

In Western cartographic traditions, maps have a long history as both scientific representations and expressions of power. The emergence of modern cartography in Europe between the fifteenth and nineteenth centuries served the needs of exploration, state formation, and colonization (Edney 1997). Increasingly centralized governments used maps, as representations of territories, to organize, tax, and control both space and subjects. At the same time, through their creation and use, maps helped people living within these territories identify themselves as subjects, citizens, and, eventually, members of nations. During the same period, emerging capital and states sponsored expeditions that mapped trade routes and “new” lands for colonization. As technologies, practices, and representations, these maps — politically, culturally, and scientifically — wrote the world into the known and, in many cases, served as artistically enticing advertisements to attract investors and colonists.

Cartography became integral within geography as it emerged as a formal discipline in Western academies in the late 1800s. During the quantitative revolution of the 1950s and 1960s, academic cartographers established their craft as a spatial science with maps as both the products and subjects of scientific investigations. To that end, cartographers theorized maps as communication devices.
More specifically, Arthur Robinson and Barbara Petchenick, in 1976, described maps as representations of the milieu — a term chosen to encompass the cartographer’s environment in addition to the meaning of place. They and other academic cartographers drew from structural linguists and psychologists to theorize that maps’ accurate topological representation of reality make them more fundamental to human experience than language. As other cartographers formalized the communication model further, maps became defined as scientifically created presentations communicating a representation of reality formed within the cognitive realm of the cartographer to the cognitive realm of the map-reader (Morrison 1976).

As a result of the cartographic communication model, academic cartographers relied on methods borrowed from behavioral psychology to measure the effectiveness of different symbol sizes, colors, and typefaces in communicating information between the cartographer’s and map-readers’ cognitive realms (Gilmartin 1981). Examples of this work included controlled experimentation that tested the accuracy of one’s map-reading skills based on the use of different symbolic systems (Board 1978). In these experiments, both author and reader were treated as individuals that had the capacity to cognitively read and map the world around them.

In contrast, but still working with a scientific approach, is the work of Alan MacEachren, who argued that communication models should be replaced by models based on “visualization” (MacEachren and Ganter 1990). As developed more fully below, maps as visualizations focus on representing unknowns rather than the known geographies of communication models. The model of representation that underwrote both the communication and visualization approaches abstracted map symbols and other elements from the social processes involved in a map’s production and consumption.

Critical cartography

While scientific geography faced increasing criticism from humanistic geography, feminist geography, and Marxist geography in the 1970s and early 1980s, most academic cartographers continued to operate within scientific communication or visualization models to incorporate digital technologies — including nascent geographic information systems (GIS). By the late 1980s, a few voices began questioning the assumptions about how maps represented reality that undergirded scientific cartography. J.B. Harley (1989, 1990) and Denis Wood (1992) adapted Michel Foucault’s understanding of power and, in Harley’s case, a relatively narrow version of Jacques Derrida’s deconstruction to retheorize the relationship between the map, the cartographer, and the worlds they represent.

Harley’s efforts were intimately tied to a much broader trend in the discipline of geography, which was in the midst of a “crisis of representation.” Informed, in particular, by the development of a new cultural geography, which called into question the discipline’s role in the reproduction of systems of power and authority, a group of geographers began to question how and for whom geographic representations worked. Harley, working within this wider conversation, called into question maps as representations of the “real world,” instead suggesting that maps, like all geographic representations, were better theorized as part of a wider cultural politics. This move aligned with the discipline’s own interest in interrogating its history and historical geographies more generally, from the role geographic representations played in the building of empires to the ways in which geographic representations obfuscated subaltern experiences of space and place.

Denis Wood (1992, 2010), in slight contrast, built his critique of the limits of scientific
cartography out of an engagement with “mundane” cartographies, both examining maps that were often marginal to cartographic discussion because of their apparent lack of scientific rigor and producing maps that built on mundane experience. Wood’s work pushed the boundaries of what cartographers might consider as objects of their scholarly analysis as well as what sorts of data and representations geographers might use and produce.

The critical cartography that emerged out of the disciplinary debates of the late 1980s and early 1990s called on practitioners to look at the margins of the map and beyond. It focused attention on the ways a map’s symbols, titles, legends, pictures, and descriptive texts represent both absences and presences and posited that these, in turn, are representations of social power. Critical cartographers argued that maps and map-makers need to be understood within the complex social and political contexts of their production. As such, maps, like landscapes, film, and other representations, are defined as both parts and products of social discourses.

As some human geographers turned toward poststructuralism to challenge any reliance on fixed and external referents in theories of representation, some geographers challenged Harley’s work for positioning social power as such a referent. Sparke (1995, 4), for example, argues that understanding a map requires the critic to constantly go back and “look for other ways in which the map and what is supposed to lie outside of it – power relations, interpretation, the ‘real’ world, and so forth – might be still more complexly inter-related.” Indeed, poststructural geography sought to unhinge the notion that binaries, such as reality and representation or production and consumption, can be theorized as distinct concepts. Offered instead, was a constitutive understanding of binaries as coproducive.

GIS’s renewed scientific representations

While critical cartographers were radically retheorizing maps as representations imbued with social power, geographers, computer scientists, and others developed GIS – databases and automated tools designed to function as scientific models of reality. Technological advances in geodesy, remote sensing, and surveying increased the accuracy of maps’ topological representation of reality. New developments in database structures and algorithms were developed to represent geographic information digitally. The process required a formal logic describing the relationship of reality to the digitally coded data representing that reality as well as to graphic representations produced from those data (Brodeur et al. 2003; Schuurman 2006).

As in the communication model of cartography, in the emerging geographic information science (GIScience) reality was first represented by cognitive models of reality – although the cognitive science underlying these models involved
a better understanding of the neurological processes involved in perception and cognition than the behavioral psychology cartographers borrowed in the 1960s and 1970s (see MacEachren 1995). These cognitive models then required formalization into conceptual representations that can be coded. The actual coding of these concepts comprised the next step resulting in a digital database, which was used to produce spatial concepts that could be represented in a digital map. Schuurman (2006, 729) summarized this process by stating that in GIS, “representations of space are expressed conceptually, programmed in code, represented graphically, and then become the basis for environmental and other decision making.”

Some of the limitations of this formal and functional method to represent reality were acknowledged by GIS developers who, as early as the late 1970s, noted that there were severe constraints in the ways digital data models could represent how people experience and organize “space” (see Schuurman 2006). This is, in part, due to the limitations of traditional cartographic data models which often serve as the conceptual representations of space described above. Yet, these internal critiques had little impact on the ways GIScience conceptualized representation before the late 1990s.

By that time, GIS and GIScience were displacing cartography from geographic education. In addition, as maps became digital they became more flexible and ephemeral. And, increasingly, map authors were not classically trained cartographers. These factors led to the rejection of the idea that maps are only communication devices. Instead, Alan MacEachren (1990, 1995) and others theorized maps as scientific geovisualizations that can be employed in different ways depending on the purpose and context of their production and use.

In his model of cartography, MacEachren acknowledged that a map designed to represent known geographic information to a general public who are expected to interact with its content as traditional map-readers can still be usefully understood as a form of scientific communication. However, he argued that increasingly maps are not designed or used in this way. His model allowed for maps to represent previously unknown geographic information to a small number of people who interact with it in ways that may profoundly change this representation. This is a map as a geovisualization; it is a scientific tool that allows map-makers and users – often the same people – to represent spatial patterns in order to first explore, then describe, and then to explain them. Like the communication model that preceded it, however, maps and the mostly quantitative, topological data they represented remained fixed in a functionalist scientific epistemology. And, as operationalized through GIS, maps were often thought of as examples of developments in visualization technologies.

Critical GIS, participatory GIS, and qualitative GIS

External critiques of GIScience emerged within human geography in the early to mid-1990s. Early criticisms characterized GIS as a dangerous return to the overly abstract quantitative geography of the 1960s and described the ways GIS practitioners represented data and reality as undertheorized. Writing from perspectives informed by feminism, Marxism, and poststructuralism, these critics chastised GIS developers for assuming that their models could objectively represent reality and/or for utilizing a Cartesian epistemology that is inadequate for representing both human and natural phenomena. Others argued that GIS represented and reproduced
space as a system of control and surveillance in ways that served the needs of capital and the state. In summary, they characterized GIS as a representational process that abstracted data, and the visualizations based on those data, from the social relations they both represent and reproduce. In addition, such critics argued that the technological costs and expertise needed to work with GIS excluded and subjugated already marginalized populations (Pickles 1995; Sheppard 1995).

Many GIScientists questioned the practicality of these critiques and argued that the critics did not understand the abstractions and formalizations undergirding their discipline (Openshaw 1997). Yet, some researchers working within GIS recognized that few in the field had conceptualized their work in a social context (Schuurman 1999; Kwan 2002). Lamenting the “antiseptic” nature of GIScience and worrying that the critiques leveled by human geographers in the 1990s had been ineffective, Nadine Schuurman, Mei-Po Kwan, and others outlined a critical GIS that works to incorporate issues raised by critical human geography into GIS using language understood by GIScience (Schuurman and Kwan 2004). Questions of representations were especially central to two streams of critical GIS research.

Public-Participation GIS (PPGIS) developed independently and in parallel to the external critiques described above. Focused on on-the-ground, practical research, practitioners of PPGIS engaged with underprivileged social groups to develop GIS meeting communities’ needs and help people in particular places visualize the spaces they share and hope to change. Central to the development of PPGIS were questions concerning the authorship and ownership of both data and maps representing these communities (Elwood 2006). In other words, in a participatory GIS, the representation’s subjects are the same as the representation’s authors and, as such, are understood to control how and by whom the representation is used and interpreted. Such concerns connect the ways representation is utilized with PPGIS with methodological literature in human geography and with poststructural theorizations of maps in critical cartography.

Qualitative GIS emerged, mostly after 2000, in response both to calls for a critical GIS and to meet the needs of researchers and communities working with data representing situated and negotiated knowledge about space and place (see Jung and Elwood 2010). Rooted in grounded theory, feminist, or mixed method approaches, qualitative GIS requires the production of formal codes and techniques that permit existing GIS software – typically designed to represent quantitative and topological data – to represent qualitative data, such as interview transcripts, photographs, or audio recordings expressing more experiential understandings of space and place. At the same time, such representations must be in a form permitting the use of qualitative analytical methods that allow researchers to explore the “contradictions, commonalities, and nuances of data that are rich in contextual and process-based detail” (Elwood and Cope 2009, 4). Finally, researchers utilizing qualitative GIS techniques must grapple reflexively with the ways in which the data and visualizations with which they work both represent and reproduce relationships of unequal power between, for example, researchers and subjects.

Nonrepresentational theory and cartography

Within critical cartography and critical GIS, the poststructuralist turn presaged the interjection of nonrepresentational theory (NRT) into understanding map-making and map use. As Joe
GERLACH (2014, 26) summarizes, NRT is “at once a critique of the epistemological certitude of representational, and by the same token cartographic, reasoning, while also affirming a geography of things happening.” Building from the work of Nigel Thrift (1996), NRTs argue that the focus on maps, landscapes, and other representations inadequately captures the multisensory and emotional practices and experiences of ordinary lives. For many geographers utilizing NRT, the goal is to theorize affect or to acknowledge that “there is no stable human or nonhuman existence because so much of what happens in the word, what moves, what becomes, from day to day and throughout spaces, does so before it passes the threshold of cognitive recognition” (Gerlach 2014, 27).

As applied to maps, NRT foregrounds emergence, practice, and performance rather than the representation as a fixed object. Nigel Thrift calls maps “their own practitioners” (2007, 58), meaning that they are always parts of what they represent and that their producers and consumers are not separable from either each other or from maps as emergent, practiced, and performed.

Rob Kitchin and Martin Dodge (2007) develop this line of argument more fully. They argue against a map as either an objective, scientific representation derived from reality through cognitive processes or a representation of social power. Instead, a map – or mapping – is simply of-the-moment, contingent on how it is practiced in any particular context. Using a map derived from Irish census data as an example, Kitchin and Dodge state that the lines, colors, and other marks only take the form of, and are understood as, a map through mapping practices. These practices are learned “knowledge and skills [that] (re)make the ink into a map and this occurs every time they are engaged with – the set of points, lines and areas is recognized as a map; it is interpreted, translated and made to do work in the work” (Kitchin and Dodge 2007, 335). Thus, a map does not shape how we think about space and, thereby, does not make space. It is a co-constitutive production between inscription, individual, and world; a production that is constantly in motion, always seeking to be an immutable mobile, a particular representation that carries its relationship to reality and to its producers and consumers with it wherever it is engaged. In this way, a map is always ontogenetic; it is emergent.

From this point of view, the map is gone. Only mapping remains. Cartography becomes a “science of practices not representations” focused on how the acts or performances of mapping are produced and how different mappings do work in the world (Kitchin and Dodge 2007, 342). The study of how certain mappings appear to be ontologically secure is another focus in this radical rethinking of the subdiscipline.

A nonrepresentational cartography does have points of attachment with other theories of representation in cartography and GIS. As a science of practices, it engages with the formal logic of digitally coded geovisualizations – entities called into being as parts of scientific research rather than as finalized and objective representations of known spaces. And, like Del Casino and Hanna’s (2006) map space, Kitchin and Dodge’s mapping cuts through the binaries between representation and reality, production and consumption, representation and practice, and ontology and epistemology. While others criticize nonrepresentational cartographic theory as asocial and, therefore, useless for understanding how sociospatial power is deployed and reproduced, Thrift (2011) offers a possible way forward, arguing that maps, as processes/practices, are now inhabitable. Inhabitable maps, such as those found on Google Earth or in video games, are engineered to produce “defined experiences which can be commoditized” (Thrift 2011, 9).
As such maps or mappings can become means of achieving social assent. Thrift’s approach does not presuppose unequal social relationships or hegemony, but allows for mapping practices, theorized as emerging in particular spatial and temporal contexts, to assert ontological security as a means to enforce a particular sociospatial order.

The democratic potential of a new cartography

Professional geographers are not the only people to develop and distribute geographic representations of space in map form. The emergence of new technologies that allow a variety of users to collect and map spatial data has facilitated a much broader world of maps. The authority of the cartographer is challenged by these developments, as is the capacity of the discipline to manage the representational politics of maps. One such avenue of challenge has developed from the practice of “voluntary mapping,” which allows people to either build their own maps or represent themselves in map spaces, from websites such as Wikimapia to organizations mapping data in real time and space to focus their political agendas (Elwood 2008). Websites, such as Google Maps, have also redefined the term “map maker” by allowing millions of people to tag and upload data to a Google Map database. These technological changes, a function of the development of Web 2.0, have increased the capacity of everyday users to engage in mapping practices. According to Michael Goodchild (quoted in Thrift 2011), “we are, in effect, back to the days of the 1500s when it was possible for someone – a cartographer – but with no qualifications whatsoever, and no authority, to produce a map which led, in effect to the naming of America.” This new way of mapping challenges those traditional “map authors” – cartographers, GIS professionals, and massive data organizations, such as the US Census Bureau – who have long controlled the organization and representation of much spatial data.

On the other side of the everyday cartographer is the development of massive datasets situated within a new era of big data. Resting at the heart of big data analysis is a commitment to data visualization, which includes the use of mapping technologies. Some of these projects rely on voluntary data to help destabilize normative assumptions found in historical mapping projects, such as the US Census. In other cases, visualization projects are gathering data points from around the globe to investigate sociospatial networks, be it within transportation systems, urban greenspace systems, or criminal trafficking patterns. The turn toward big data also calls into question the traditions of positivist quantitative representations, which rely on historical variables, such as fixed locations and generic social and economic variables. In this new era, crowdsourced data provides a different cut at how we might map diverse spaces and processes (González-Bailón 2013). At the same time, big data representations make claims to “real data” in “real time” in ways that may prove equally problematic in relation to earlier scientific approaches to geographic representations, which, for example, seemed to marginalize the histories that undergird these data (Barnes 2013).

Conclusion

Cartography is a core subdiscipline through which the debates around representation in geography have taken place. At the heart of these debates are the questions of what it means to represent the world and whether we can even do that work appropriately. On the one hand, the model of mapping as a science remains central to
the discipline, finding its outlet in a revitalized visualization approach to cartography and the formalistic approach to modeling in GIScience. On the other hand, there is a constant re theorization of maps and their relationships to both the spaces they represent and the people who simultaneously are their makers, users, and subjects. While many of the applications of critical, poststructural, and NRTs to cartography and GIS are performed by geographers outside of these twinned subdisciplines, an increasing number of scholars claiming the identities of GIScients and cartographers are seeking ways to “re-wire” geospatial technologies to work with and represent alternative ways space and place are experienced and known.

Such developments, and the explosion of voluntary mapping and big data applications, mean that cartography cannot be contained by disciplinary boundaries that seek to delineate its meaning and practice, suggesting that maps are more than objects of geographic use and inquiry. Whether theorized as representations, visualizations, or practices, maps may be better imagined as art and not science. This can be seen, for example, in projects designed to challenge familiar geographic sensibilities by distorting space through maps. Projects have developed that are much more playful with spatial data, producing new representational mapping practices that are not beholden to the conventions of a communication or visualization model. These new cartographies are less interested in the theoretical or ontological questions found in critical cartography, on the one hand, and geovisualization and GIS, on the other.

Despite the explosion of maps, visualizations, and other cartographic representations, it is clear that the fields of cartography in particular, and geography more generally, will continue to support a wide range of representational theories and practices. What this suggests is that the concept of representation – of how we can represent the world – will remain highly contested, even as some seek to remove representation from geography’s conceptual toolkit.

**SEE ALSO:** Cartography; history; Critical GIScience; Feminist geography; Geographic information science; New cultural geography; Nonrepresentational theory; Poststructuralism/poststructural geographies; Public-participation GIS; Qualitative GIS; Representation; Visualization

**References**


Further reading


In the early days of geographic information systems (GIS), the “field” view of geographic reality prevailed. A field answers the question “What is here?” It gives a value for each possible location on a surface (ideally, a continuous surface). The alternative view in geographic data modeling is the “entity” view. It answers the question “Where is this object?” An entity-based data model sees entities (or objects) as independent from each other and assigns them a link to their location in space (Couclelis 1992).

The field/entity dichotomy usually adopted in GIS has its roots in philosophical ground. In Aristotle’s view, void space does not exist: bodies, with their extension, determine the existence of space; therefore, a parallel can be drawn with the entity view. In the earlier view of Plato and the Pythagoreans, space was a more abstract concept (geometrical figures); here a parallel can be drawn with the field view.

Several data structures, raster or vector, were defined to represent geographic data models in the limited memory of computers. It is commonplace to think of rasters as the most immediate data structure to represent a field, and of vectors to represent objects. This is, however, not always the case and can lead to misunderstandings. There exist vector data structures (such as triangulated irregular networks (TIN) and contour lines) that represent fields.

Likewise, objects can be represented as rasters (it is common practice in remote sensing and object-based image analysis (OBIA)).

Nonetheless, for representing entities, vector data is the prime choice to assure flexibility and scalability. From the early vector-based GIS, single objects could easily be represented as points, lines, or regions, but it was not so straightforward to represent collections of objects and their relationships. The latter issue produced a range of different data models categorized as spaghetti models, network models, and topological models. While spaghetti models were based on collections of independently represented lines, network models and topological models aimed to explicitly represent topological relationships among entities and avoid redundancy or duplication. Basically, network and topological models were graph-like structures composed of nodes and edges. Topological models in early vector GIS were associated to the idea of single-value maps or planar enforcement allowing only one object at a given location. Single-value maps can represent spatial objects of a single type (e.g., administrative boundaries or a road network). Hence, spatial data is organized in several layers, each one showing a single-value map. At the same time, topological models experienced challenges in representing objects of complex shapes, such as regions with holes (a region with a hole cannot be modeled as a face of a graph embedded in the plane) or complex lines (when two roads intersect, the roads must be split to obey the single-value map principle).

Topological models, even if they were vector-based representations of entities, could not capture the complexity of real geographic fea-
tured and failed to find an adequate representation for them. From the 1980s, a new data modeling paradigm emerged in computer science, the so-called object-oriented model (Atkinson et al. 1992). While in the antecedent relational data model information was typically fragmented over several tables, the object-oriented model considers objects of arbitrary complexity, which can be constructed via the repetition of abstract modeling primitives of classification, aggregation, and generalization. The object-oriented approach was introduced to GIS during the 1990s, when the use of the term “complex object” in spatial databases was also observed. According to Goodchild (Goodchild 1992), complex objects are incompatible with the field view of reality. Complex objects are the true realization of the entity-based view in geographic data modeling. Complex objects are able to represent both the complex internal structure of entities (e.g., the presence of holes or subparts), and the complex relationships between them (e.g., topological relations of crossing and overlapping).

The object-oriented paradigm is now well established in information technology. Complex geographic objects have been included in spatial data type specifications defined by the Open Geospatial Consortium (OGC). The adopted definition of data types representing complex objects is described in the Simple Features Interface Standard (SFS) (OGC 2006). Such data types are the basis for all interoperability standards for geographic information exchange, such as the Geography Markup Language (GML) (OGC 2007), the Web Feature Service (WFS) (OGC 2005), and the language for supporting geographic objects in the semantic web, GeoSPARQL (OGC 2012). These data types are rooted in the mathematical foundation of point-set topology.

Definitions of complex objects in point-set topology

A mathematical definition of complex objects is usually expressed in point-set topology. A topological space is generally described as a set of arbitrary elements (points) in which a concept of continuity is specified. A continuous mapping transforms a topological space $X$ to a subset of a topological space $Y$ subject to the preservation of the neighborhood relations between mapped points. Continuous mappings are also known as homeomorphisms or topological transformations and include translation, rotation, and scaling. Topological relations remain invariant under topological transformations. The study of topological relations between objects also depends on the embedding space, which is assumed to be $\mathbb{R}^2$.

In the following, the definitions of simple and complex regions, lines, and points are introduced. All the definitions were originally published in (Clementini and Di Felice 1996). Simple regions are equivalent to a disk and simple lines have just two end points. Complex regions have separations and holes, complex lines have separations, more than two end points, and possibly self-intersections, and complex points are a set of points. Complex objects are far more common than simple ones in geographic reality. The definitions are based on concepts of open sets, closed sets, and continuity of point-set topology.

Let $A$ be a two-dimensional point-set:

- the interior of $A$, $A'$, is defined as the union of all open sets contained in $A$;
- the closure of $A$, $\overline{A}$, is defined as the intersection of all closed sets containing $A$. If $A$ is a closed set, then $\overline{A} = A$;
- the boundary of a closed set $A$, $\partial A$, is defined as the difference between its closure and its interior;
the exterior of a closed set \( A \), \( A^- \), is defined as \( A^- = (\mathbb{R}^2 - A) \).

Throughout the entry, it is assumed that regions are regular closed sets.

**Definition 1** A point-set \( A \) is called regular closed if, and only if, \( A = \overline{A} \).

Regular closed point-sets are usually assumed in spatial data modeling because they allow particular cases to be got rid of, such as those shown in Figure 1a–c. Nonregular point-sets can undergo a regularization process to become regular (Figure 1d–f).

The notion of complex regions relies on the topological notions of connectedness and component. A *separation* of \( A \) is a pair of disjoint nonempty open sets \( A_1 \) and \( A_2 \) whose union is \( A \). \( A \) is connected if there exists no separation of \( A \); otherwise it is disconnected and \( A_1 \) and \( A_2 \) are called the *components* of \( A \). Separations and components refer to interiors of regions, while spatial features are represented as closed sets, in order to permit that a region also contains its boundary. Simple regions are defined as follows:

**Definition 2** A simple region is a regular closed (nonempty) two-dimensional point-set \( A \) with connected interior and connected exterior.

A simple region is illustrated in Figure 2a. A region with disconnected interior (Figure 2b), a region with disconnected exterior (Figure 2c), and a region with spike (Figure 2d) are not valid simple regions.

Now the above definition can be extended by taking into account holes. The exterior of a region with holes has separations. Separations of the exterior imply that there exists one *outer exterior* (unbounded set) and \( n > 0 \) *inner exteriors* (bounded sets). The outer exterior will be denoted by \( A_0^- \) and the inner exteriors by \( A_1^- \ldots A_n^- \). Their union makes up the entire exterior, that is, \( A^- = \bigcup_{i=0}^{n} A_i^- \).

**Definition 3** A region with holes is a regular closed (nonempty) two-dimensional point-set \( A \)

---

**Figure 1** Examples of nonregular closed point-sets (a–c) and their respective regularizations (d–f).
According to Definition 3, the case in Figure 3a is allowed, while the cases of disconnected interior (Figure 3b) and of nonregular closed point-set (Figure 3c) are not allowed. Therefore, in a region with holes, the closures of any two different exterior components cannot touch along a line, but they are allowed to touch at some points. Formally, their intersections must be empty or equal to a finite set of points \( \{p_1 \ldots p_k\} \):

\[
\forall i, j = 0 \ldots n, i \neq j : \\
(\overline{A_i} \cap \overline{A_j} = \emptyset) \lor (\overline{A_i} \cap \overline{A_j} = \{p_1 \ldots p_k\})
\]

Now, composite regions can be defined:

**Definition 4** A composite region is a regular closed (nonempty) two-dimensional point-set \( A \) with disconnected interior.

If the interior of \( A, A^\circ \), is disconnected, the components \( A_1 \ldots A_n \) of \( A \) can be indicated as the closure of the corresponding components of \( A^\circ \). For the components of \( A \), it holds:

1. each \( A_i \) is either a simple region or a region with holes;
2. \( A_i^\circ \cap A_j^\circ = \emptyset, \forall i \neq j \);
3. \( (\partial A_i \cap \partial A_j = \emptyset) \lor (\partial A_i \cap \partial A_j = \{p_1 \ldots p_k\}) \).

In other words, each component is either disjoint from the other ones or touches other components in single points: it cannot touch other components along a one-dimensional piece of boundary. Figure 4 shows an example of a possible composite region.

Now, the definitions of lines can be illustrated. Intuitively, a simple line in the plane corresponds to a single pencil stroke obtained by not passing twice on the same point. To define a simple line in point-set topology, a homeomorphism must
be considered from a one-dimensional interval to the plane requiring that the mapped points are distinct, as follows:

**Definition 5** A simple line is a closed (non-empty) one-dimensional point-set, $L$, defined as the image of a continuous mapping $f : [0, 1] \rightarrow \mathbb{R}^2$, such that $f(t_i) \neq f(t_j)$, $\forall t_i, t_j \in [0, 1], t_i \neq t_j$.

Figure 5a shows a simple line. The mappings of the endpoints of the interval 0 and 1 through $f$ are the two endpoints of $L$. These two endpoints form the boundary of a simple line in the plane; that is, $\partial L = \{f(0), f(1)\}$.

By relaxing the constraint of distinct points in the continuous mapping, the notion of line with self-intersections is arrived at. Self-intersections can involve either the endpoints or the interior of the line. Figure 5b–d shows three lines with self-intersections: the boundary of a line with self-intersections can be composed of 2, 1, or 0 end points.

Complex lines cannot be obtained by a single pencil stroke. They are composed of several simple lines or lines with self-intersections that can be disconnected or connected each other (see Figure 6).

**Definition 6** Let $f_1, f_2, \ldots, f_n$ be continuous mappings from the interval $[0, 1]$ to the plane. For this entry, a complex line is any closed (nonempty) one-dimensional point-set, $L$, defined as the union of the image of the functions $f_1, f_2, \ldots, f_n$.

Figure 5  A simple line (a) and lines with self-intersections (b–d).

The boundary of a complex line $L$ can be built subtracting from the set of endpoints of single lines $f_i$ all the endpoints that touch other lines.

The closure of a line, $\overline{L}$, is the set of all points of $L$, end points included; therefore $\overline{L} = L$. The interior of a line, $L^\circ$, is the set difference between the closure of the line and its boundary: $L^\circ = \overline{L} - \partial L = L - \partial L$. The exterior of a line, $L^-$, is the set difference between the embedding space and the closure of the line: $L^- = (\mathbb{R}^2 - \overline{L})$.

**Definition 7** A point is a (nonempty) zero-dimensional point-set, $P$, consisting of only one element.

As a generalization of the previous concept, the notion of complex points is introduced as follows.

**Definition 8** A complex point is a (nonempty) zero-dimensional point-set, $P$, consisting of a finite number of distinct elements.

The boundary of a complex point $P$ is empty (i.e., $\partial P = \emptyset$); as a consequence, the interior of a complex point $P$ is equal to the union of all the elements in $P$ (i.e., $P^\circ = \bigcup_{i=1}^n P_i$).
Other dimensions in conceptual modeling of complex objects

If facing the conceptual modeling of a complex geographic object, various components are involved in its definition. Not all of them exist at the same time nor might there be the actual need to consider all of them in the model. For this entry, the following possible dimensions are considered to be taken into account.

- **Compound definition of the spatial component.** The actual geometry of a geographic object might not be described with a single spatial component as described in the point-set topology approach. Therefore, a compound geometry, that is, an aggregate of various geometric types, might be needed.

- **Multiple representations.** A single spatial representation might not sufficiently represent a given geographic object. A road could be represented as a polyline at a small scale and as a region at a larger scale. Therefore, a complex object might allow the coexistence of several geometrical representations at the same time, that is, multiple representations.

- **Semantic and spatial properties.** Besides the spatial component, complex geographic objects possess a semantic component. Complex object representations include thematic attributes, functional requirements, and various types of constraints.

- **Temporal component.** A geographic object is rarely seen as time-independent. Time relates in various ways to geographic object definitions. For example, trajectories (of cars moving in town) cannot be modeled with a separate temporal component, but they exist in both space and time. In other cases, for example, a lake that cyclically shrinks and expands its boundary due to rainy seasons may associate the temporal component by adding time stamps to the spatial component.

- **Uncertain boundaries.** The point-set topology approach produces models with precise boundaries (also known as crisp boundaries), which are always an approximation of reality at a given scale. Geographic objects in a variable measure all possess a degree of uncertainty in boundary definition. Even human-made objects such as buildings would not have crisp boundaries at a sufficiently large scale. Natural objects such as mountains and islands have a greater degree of fuzziness than buildings. While currently not much implemented, models considering fuzzy or broad boundaries could significantly improve geographic modeling.

- **Aggregations.** Often complex objects are made of other objects, which can exist independently. A mall is made up of several shops. A building is made up of walls, windows, and stairs. Object-oriented modeling does not directly include such components inside the definition of bigger objects: it includes references (or “pointers”) to other objects, which therefore conserve their own identities.

- **Complex relationships.** Complex objects interact with other objects in various ways. A city “is the capital of” a country. A river “flows into” another river. In many cases, such complex relationships hide a spatial relationship between their geometric representations (an “inside” in the case of city-country and “touch” in the case of river-river). A considerable part of geographic data modeling, therefore, may be devoted to the description of complex many-to-many relationships between geographic objects.
**SEE ALSO:** Data model, object-oriented; Data structure: spatial data on the web; Open Geospatial Consortium standards; Spatial database; Topological relations

**References**


**Further reading**


**REPRESENTATION: COMPLEX OBJECTS**


A key theme in the scholarship on dynamic complex systems is that the system as a whole takes on characteristics and functions that are qualitatively different from (i.e., is greater than) the components, processes, and interactions that comprise the system (i.e., the sum of the parts). The idea that something unique exists at the system level that cannot be identified simply by enumerating, describing, and understanding its parts (i.e., reductionism) is not new in geography, and a holistic approach to understanding system dynamics has characterized the discipline from its earliest days (e.g., von Humboldt’s (1845, 4) “perception of unity of plan amidst eternally recurring variety of form” in Cosmos).

In the early 1900s, geographer Vidal de la Blache postulated that interactions among entities were the key to the emergence and evolution of system properties. He described processes on the Earth’s surface that were interrelated and affected others (such as “localized ocean currents, geological traits, and climatological attributes”) and wrote that these processes were interdependent and together formed an organism, a whole (Archer 1993, 501). Later, the Gaia hypothesis (Lovelock 1979) and the whole field of Earth system modeling (Falkowski et al. 2000) formalized the notion of the organic evolution of interconnected Earth systems.

Fascinated with the whole as an organism, Vidal de la Blache described parallels between the formation of biological organisms and the emergence of social, political, and industrial organisms from society (Archer 1993). He noted that emerging organisms are dependent on and influenced by the interactions among the animate entities that form their base. He also noted that geographical boundaries often unite and bond certain animate entities together through their interdependence and interactions (Vidal de la Blache, 1926, 165 – cited in Archer 1993, 502). In the case of humans, where interactions have increased over time as a result of improved communication and transportation, larger social organisms have emerged (Archer 1993).

Vidal de la Blache (1898, 102 – cited in Archer 1993, 502) described relationships of interaction between “two principal cycles, that of the inanimate and that of the animate world,” which corresponds to the typical division of phenomena into human and physical geography. The inanimate cycle refers primarily to deterministic relationships and interactions among geological, hydrological, climatological phenomena, and processes of nature; while geographically animate cycles are primarily nondeterministic and are concerned with living phenomena, such as humans, that have the capability to self-organize and adapt (Archer 1993).

Cutting across the traditional domains of geography as it does, it could be argued that the representation of dynamic complex systems is a necessary component of spatial reasoning and analysis. Based on Tobler’s (1970) first
law of geography, which explicitly incorporates spatial interactions, that is, “All things are related, but nearby things are more related than distant things,” spatial analysis is the study of emergent spatial patterns due to changes in the underlying properties of the local environment (i.e., first-order variation) or the interaction among observations (i.e., second-order variation). Spatial analysis in geography, then, seeks to understand complex systems in a way that explicitly incorporates space. Because spatial patterns emerge and evolve over time, understanding geographical processes requires a spatial analysis that is dynamic and can adapt to a changing environment (first- or second-order variation) to describe and compare patterns over time due to behavioral and interaction processes.

Studying dynamic complex systems requires conceptualizing and operationalizing approaches for representing relationships and interactions in space, over time, and at multiple scales that can produce emergent and aggregate patterns. This entry reviews these issues in representing dynamic complex systems in geographic research and identifies gaps and potential future directions for geographic research.

**Complex systems**

A complex system can be defined as a system that comprises many heterogeneous interacting parts that, when aggregated, produce system-level properties for which analytical solutions are intractable and cannot be easily deduced from the actions of the parts. Flocking behavior by birds is an example of a phenomenon that illustrates the behavior of complex systems. The movement of the flock and any one bird are difficult to predict. However, the pattern of flocking behavior can be reproduced by simulating individual birds that abide by three simple rules (i.e., follow the average heading and velocity of the flock, maintain space so as not to bump into other birds, and move toward the center of the flock) (Reynolds 1987). This representation of a complex system provides an existence proof (Waldrop 1990), demonstrating that representation of dynamic complex systems can provide understanding of the processes by which individual components of a system interact with each other to produce observed spatial and temporal dynamics.

Complex systems can be nonadaptive or adaptive. Self-organization of landslide dynamics, in which adding grains of sand to the pile affects the frequency and magnitude of landslides (Bak, Tang, and Wiesenfeld 1987), is an example of a nonadaptive system. Here, the grains of sand do not have the ability to change their behaviors or their goal. In some cases, however, complex systems are adaptive (termed complex adaptive systems or CAS), in which the system components are able to adapt (Axelrod and Cohen 2000) and change their goals and their behaviors depending on their state, the states of their neighbors, or the state of the system (i.e., autonomous agents) (Iglesias, Garijo, and Gonzalez 1999). These complex adaptive systems are more resilient to external perturbations and may change their structure to improve the functionality of the system components or agents and/or the system as a whole. Most geographic research representing dynamic complex systems incorporates both adaptive and nonadaptive components.

Interactions among nonadaptive or adaptive components of a system can produce feedback effects which result in systems that can exhibit complex spatial and temporal structure, including nonlinear dynamics and threshold effects, that are sensitive to initial conditions and may hold path dependencies, and that can produce multiple equilibria (i.e., the system does not necessarily
converge on a single equilibrium outcome. To represent these feedbacks requires understanding and describing cross-scale dynamics in complex systems, which involves understanding the micro-level processes that form system behaviors and the macro-patterns that emerge from those behaviors.

**Representing dynamic complex systems**

Geographic information systems (GIS) and spatial analysis approaches have been successfully used to represent, describe, and analyze the spatial macro-patterns of dynamic complex systems. However, until recently, these approaches have been limited in their ability to represent detailed knowledge of process and behavior, the drivers of dynamic interactions among features in a landscape, and how features change over time and through interaction with other features and the environment.

Instead, a variety of other approaches have been used to represent and model the processes inherent in dynamic complex systems. In some cases space–time statistical models have been constructed to represent behavior and change across a landscape (e.g., land-cover change or disease spread). Additionally, mathematical models, such as ordinary differential and partial differential equations, can be used to capture the dynamics and interactions of dynamic complex systems. However, mathematical models typically have difficulty representing heterogeneity among system components and the micro-level behaviors of those components.

Arguably, the preferred approach for modeling complex systems is the representation of system components, be they locations on a grid or mobile decision-making entities (e.g., agents), in a simulation model (Torrens and Benenson 2005). Cellular automata (CA) describe the former type of models, with fixed topological relationships, and agent-based models (in the social sciences) and individual-based models (in ecology) describe the latter case of entities with an evolving topological structure (hereafter both are referred to as agent-based models or ABMs). These models are able to capture most of the primary characteristics that describe the micro-level processes driving dynamics in complex systems, such as agent heterogeneity (Holland 1995), interaction (Anderson 1972), adaptation and learning, and agent–environment feedbacks, among others. The ABM approach is appealing because it allows for a one-to-one matching of real-world actors to virtual-agent entities that can facilitate model parameterization, calibration, validation, and communication of model results with system stakeholders.

Next, how the various components of dynamic complex systems are typically represented are described. While many of these representations can be utilized across different modeling approaches, the focus is specifically on CA and ABM as the key tools for representing these dynamics.

**Representing space**

The representation of geographic space in a dynamic complex system provides a measurable environment within which system dynamics evolve, through changes in the location of dynamic entities (i.e., absolute space) and in their topological relationships (i.e., relative space). Traditional concepts from geographic information science (GIScience) serve well to describe these representations. Absolute space is defined by a mathematical coordinate system, and typically a projection, that maps cartographic or modeled entities to locations on Earth. The dominant data models for representing the space of a dynamic
complex system are raster and vector data models, which roughly but not perfectly correspond to field- and object-based conceptions (Bian 2000) of spatial variability, respectively. The coordinate system provides information about the location of entities (e.g., cells or objects), about which there is usually also some attribute information (i.e., categories or quantities) that is either fixed to a location (in the field view) or fixed to an entity whose location can change (in the object view). Relative space can be defined through explicit description of spatial connections between entities, through concepts of adjacency or proximity, or through nonspatial connections that represent known social, hierarchical, or network-based interactions.

Cells, features and agents

The entities that implement the dynamics in these models can be either agents (in ABMs) or cells (in CA). “Agent” is a term adopted from economics (Holland 1995) and is typically used to represent real-world actors (e.g., individuals, firms, governance entities) (Rounsevell, Robinson, and Murray-Rust 2012). While agents can be reactive or deliberative (Doran 2001), agents can be thought of as any object (defined as a software object) that has the ability to satisfy internal goals or objectives through actions and decisions based on a set of internal rules or strategies (Iglesias, Garijo, and Gonzalez 1999). An agent may represent a single entity, such as a blood cell, or a collection of agents, such as a firm or organization.

The only real distinction between an agent in an ABM and a cell in a CA is that agents need not have a fixed location, or any location for that matter (as in the case, for example, of agents that represent institutional rule setting), whereas cells in a CA are explicitly fixed to a geographical location. This distinction is important when it comes to understanding the relationship between the entities in the model (agents or cells) and their referents in the real world, and has parallels in the object/field distinctions in GIS. The key advantage of agents is that they can be more flexibly associated with individuals or groups of people, animals, or other mobile entities, whereas cells in a CA are limited to represent particular locations or areas as the basic entities in the model.

In the object-based (or vector) data model, a feature represents entities in the real world. Real-world features, like a person, a road, a coastline, or an area of specific soil type, can be represented virtually in a GIS as points, lines, or polygons. Features can be defined as simple and independent, such that they have no relationship with neighboring features (adjacent or distance based), or complex, with defined and distinct relationships with other features around them (Longley et al. 2001). Features have attributes that describe their properties. As spatiotemporal information is increasingly incorporated in to GIS, changes to features and their attributes are represented explicitly, but they do not have explicit representations of behavior (i.e., descriptions of how those changes occur).

Interactions between agents and geographic features have been described through four specific relationships: identity, causal, temporal, and topological relationships (Brown et al. 2005). Identity relationships are one-to-one or one-to-many matches between the agents, which have explicit descriptions of behavior, and the geographic features. The agents implement change and the GIS manages geographic location and feature attribute data. In the causal relationship, the agents alter features in the landscape but may have no physical presence as a geographic feature. For example, they may alter attribute values of a polygon, such as water or soil quality. In a temporal relationship, change in feature locations or attribute values...
may be implemented through the creation of new time-stamped data layers or attributes that record the history of processes created by an ABM. Lastly, in a topological relationship agent movement is influenced or constrained by the topological rules given to a set of spatial features (e.g., exclusion of movement or spatial associations (e.g., features within proximity or view).

The field-based (e.g., raster) data model tessellates a space with nonoverlapping entities and assigns one and only one entity (e.g., cell) to each location. Attributes of these locations are then recorded. Because of the fixed nature of the locations, the spatial topology among entities is relatively easy to ascertain, simplifying definitions of relative location.

**Representing processes and behaviors**

Possible approaches to representing spatially explicit processes in a dynamic complex system can be partly understood by exploring descriptions from fluid dynamics, specifically the Eulerian and Lagrangian views. The Eulerian view focuses on change in state or attributes at a location and is well represented by field-based models of space. CA models implicitly use the Eulerian view, in which they describe change of state at a location as a function of the characteristics at that location, characteristics at neighboring locations, and some rule or algorithm implemented iteratively (and usually uniformly) at all locations. While rules describe the conditions under which the attributes at a location change, and there is no explicit representation of entities, fundamental work on CA dynamics has shown that some sequences of changes in interacting neighborhoods of locations can produce complex behaviors that mimic movement of an entity across a space, such as a fire or diffusion front (Yassemi, Dragićević, and Schmidt 2008).

The Lagrangian view offers an alternative perspective on spatial processes that focuses on modeling movement of entities through space, and corresponds to the object representation of spatial entities. Descriptions of processes are tied to entities and the processes themselves describe the movement of the entities (i.e., change in location), rather than change in attributes. An example of the Lagrangian view implemented in an ABM is a predator–prey relationship, whereby individual predators moving across a landscape search for prey and the prey attempt to avoid detection and move away from predators. In these models, movement, and the ever changing relative spatial locations of entities, is key to the evolution of system properties.

In practice, many ABMs of geographical systems include representations of both movement and change, combining aspects of Eulerian and Lagrangian dynamics. When the representation of the dynamic complex system increases in complexity it becomes more likely that these two perspectives intersect as the modeler requires some processes to be represented using an Eulerian approach (e.g., soil erosion) and other processes (e.g., farmer decision-making) to be represented using a Lagrangian approach (Parker et al. 2003). Additionally, it is possible to model the same dynamics using either perspective; the choice of approach will generally result in focusing attention on different aspects of the phenomena being modeled. Relating these perspectives back to the work of Vidal de la Blache, the inanimate processes can be thought of as primarily using an Eulerian perspective, whereas the animate processes relate most directly to a Lagrangian perspective.

As the value of the Lagrangian perspective has become apparent, and its computation more tractable, more examples of its use have become available. For example, Torrens (2014) provides a counterexample of inanimate processes, that
REPRESENTATION: DYNAMIC COMPLEX SYSTEMS

is, buildings collapsing in an earthquake, represented from a Lagrangian perspective. However, typically, what processes are represented (e.g., change or movement), and how, depends partly on the types of observation data that are available as well as how the structure and results of the model are communicated to researchers trained in a specific domain or to the system stakeholders.

Time

The representation of dynamic complex systems requires that entities interact and change across space and over time. The representation of time and the inclusion of temporal dynamics in geographic analysis have been ongoing research topics within GIScience (Kwan and Neutens 2014). While GIS and remotely sensed data can increasingly support the representation of change and movement in geographic data, the complex behaviors and processes driving change over time are typically represented in the context of spatial models.

A classic approach to representing time in geographic information, and one that continues to influence the field, is Hägerstrand’s time geography, whereby he mapped the movement of individuals across space in the x and y plane and the movement through time over the z-axis. This approach has been used not only to trace the activities of individual agents in ABMs but also to assess the collective behaviors of individuals to improve evacuation procedures (Torrens 2012), solve n–p complete routing problems (i.e., the traveling salesman problem), and understand activity chaining and accessibility in spatial analysis (Miller 1999). In these examples, time geography has proved valuable, but in many cases it merely provides an alternative approach to data visualization and analysis.

Early geographic research that involved representing the dynamics of complex systems and explicitly incorporated time and process started with cell-based (or CA-based) representations of change processes, for example in Hägerstrand’s (1967) work on innovation diffusion and Tobler’s (1970) work on city growth. Agent-based approaches began to be used to represent movement, along with change, in geography in the 1990s (Deadman and Gimblett 1994; Westervelt and Hopkins 1999). These spatially explicit CA and ABMs represented both logical and physical time (Ziegler 2000). Logical time is the operationalization and virtual representation and meaning of time intervals or increments used in a model. For CA and ABMs time progresses in discrete steps over regular time intervals, for example, a day, month, or year.

Discrete-event simulations use an alternative approach for representing time. Models of this type do not have a regular interval of time progressing over the course of a model run, but increment time based on the occurrence of discrete events, such that time appears to jump between events. Discrete-event simulations tend to have processes occur in a specified sequence and have irregular temporal resolutions that are often finer than models employing regular time steps. For example, in a model of agricultural farming agents, all farmers might plant crops, then fertilize, then apply pesticide, then harvest, all while seeking opportunities to pursue off-farm labor. As each of these processes occurs, a different amount of time is advanced and recorded.

Physical time refers to the real-world temporal period the model is representing. In some cases these may map one-to-one with logical time. For example, a CA or ABM of farmer decision-making might use an annual time step because it maps directly onto the annual decisions of farm actors in the real world. With the discrete-event simulation example, planting may take 20 days,
initial crop growth may take 30 days, fertilizing may take 7 days, and so on, such that all the activities on the farm comprise one year.

Issues related to time are particularly important in how representations of dynamic complex systems are implemented in geography. A key challenge in representing time in geographic analysis has to do with assumptions of temporal stationarity in modeling dynamic processes. Models of processes and dynamics developed and parameterized with reference to observations taken over a particular time interval will provide poor predictions of conditions in different time intervals (e.g., in the future). The degree of error in these predictions will be a function of how much the processes and dynamics represented in the model differ from the constantly evolving observed system, the degree of nonstationary in the system of study. This issue will continue to limit predictability of complex geographical systems.

The historical challenge for geographic science to formally link temporal and spatial elements of geographical systems (Harvey 1968) is being solved. The adoption and integration of concepts and methods from complexity science in the study of spatial phenomena has recently enabled the linking of temporal and spatial elements through explicit representation of the processes of change and movement. The iterative nature of spatial simulation permits the execution of processes, behaviors, and changes, and provides the ability to conduct virtual experiments and perform novel scenario analyses that acknowledge the potential uncertainties and nonstationarities that can affect prediction.

Scale

The choice of what scale to use in geographic analysis (i.e., the “scale problem”) has always been difficult and, in many cases, the basis for the choice still remains “intuitively rather than empirically based” (Harvey 1968). The intuitive aspect is often associated with constraints imposed by preferred analytical tools and approaches, whereas the empirical aspect has often been constrained by available data. Representing the temporal and spatial aspects of dynamic complex systems also presents challenges associated with choosing an appropriate scale, involving tradeoffs between specificity and generality.

While the reliance on aggregate data to infer micro-level components and their behavior, known as the ecological fallacy, has been a problem in geographical research in the past, the ability to simulate the process generating temporal changes in spatial patterns can alleviate these difficulties by allowing for the representation of cross-scale processes (and analysis) to occur. Furthermore, the modeling tools now available (e.g., CA, ABM) permit the representation of dynamic complex systems at multiple scales, and simulation experiments that can help determine the scale at which a particular process contributes greatest to a specific spatial pattern. In many cases these models are combined with empirical analyses that determine, for example, the spatial distance over which interaction occurs using semi-variograms or incremental spatial association statistics (Getis and Ord 1992; Goodchild 1986). Collectively, traditional spatial analysis and contemporary complexity modeling approaches are able to illustrate how behaviors and outcomes of the actions of micro-level agents can give rise to macro-level (i.e., aggregate system level) patterns.

Aggregation

The approach in complexity science of growing or generating macro-level behavior based
on micro-level processes and the interactions among agents (Epstein 1999) provides an important cross-scale bridge. In this case, the “scale problem” exists primarily as a linguistic delineation of scale as referred to the areal extent viewed by the observer, or a particular level of resolution within a representation. Any number of features or agents may be modeled at the lowest level of representation, and any number of features or agents may also be parts of groups or organizations that exist at potentially any aggregate level of analysis.

This approach provides a solution to the problem of determining the scale at which a particular process contributes most to a spatial pattern. Using contemporary modeling approaches, a process may be modeled at a given scale to determine if it creates a specific (existing) spatial pattern. These patterns, in turn, serve as a test of the ability of the model to represent actual processes operating in the world. The better patterns represented by the model match or reproduce observed patterns in data, the more evidence there is that the modeled processes generating the patterns are correct (Grimm et al. 2005).

The observable system properties and system dynamics of a complex system grown through simulation are the result of the aggregation of individual behaviors, interactions, and resulting feedbacks within the system. This aggregation is sometimes referred to in complexity science as emergence (Johnson 2001), whereby a macro-level outcome emerges from the micro-level processes, sometimes unexpectedly and in ways that could not be intuited from knowledge of the system components. However, because intuition can be trained and expectations updated, a preferable term in the conduct of scientific approaches to complex systems is aggregation.

In representing dynamic complex systems, two types of aggregation can be modeled. The first type of aggregation may be referred to as outcome aggregation, whereby individual actions (i.e., local processes) and their outcomes derive system-level patterns and measurable outcomes (i.e., regional-level trends). The field of land-use science, for example, has embraced this type of aggregation and has used ABMs to demonstrate how individual landowner decision-making and actions can give rise to regional-level land-use and land-cover change patterns (Bell and Irwin 2002).

The second type of aggregation, which could be termed structural aggregation, is the aggregation of objects or agents into meta- or super objects or agents (Holland 1995). These meta-agents are collections of agents whereby the individual actions of agents work together to complete a high-level goal. Representations of structural aggregation in dynamic complex systems are rarely used in geographic science, but this is a direction of opportunity and future research, for example, having individual voters comprising the system-level properties of governance and directing governance agent policy development that, when implemented, affects the subsequent voting of the individuals. Whereas outcome aggregation is unidirectional in its effects, structural aggregation is bidirectional.

Generally, for either of the two types of aggregation to exist, the processes by which cells or agents change or move within the system must be described in some formal way (e.g., equations or algorithms). Empirically describing the actors within the system and representing their dynamic behavior, situated among other actors and system processes and how they interact, is difficult but has been a focus of geographic research (Robinson et al. 2007). Vidal de la Blache used biological methods to try to move geography from a descriptive discipline to one that was more experimental and able to explain
how localized phenomena were interrelated and how they self-organized into a greater whole (Archer 1993). In transcribing Vidal de la Blache’s “greater whole,” sometimes referred to as an organism, Archer (1993) used the common phrase of complexity science that “the whole is greater than the sum of its parts.” This may be interpreted as a macro- or higher-level behavior that occurs as a result of lower-level interactions and behavior that do not explicitly lend themselves to the formation of the whole. Therefore, if the rules and behaviors of the agents and interacting entities were “added-up” they would not equal or produce the same macro-level outcome as that which emerges from the system left unto itself. It has been the case that this outcome aggregation (a.k.a. emergence) has fallen under the definition of organicism, where the organism is made up of acting and interacting components that yield an organism with autonomous behavior that differs completely from those components of which it is made.

Summary

Modern social, health, environmental, and political challenges require geographic science to synthesize traditional and contemporary geographic theory, methods, and data to study the processes driving observed spatial patterns. Key to solving this challenge is the representation of dynamic complex systems, which includes both representations of the dynamics of spatial patterns and explicit representation and simulation of the processes. Coupling of spatial models that explicitly represent processes of change and movement, with spatial data representing the state of system components and aggregate spatial patterns, provides a powerful platform for exploring and evaluating the mechanisms by which spatial patterns change. Such models, then, provide a tool for testing the possible implications of physical or social interventions within the system that might be proposed for solving some of humanities toughest problems.

At first glance it may appear that representation of dynamic complex systems requires both a holistic approach and a comprehensive representation of the system of study. A misguided expectation of comprehensiveness can overwhelm researchers both in terms of the intensity of data requirements and the large amount of time needed to represent dynamic complex systems. However, successful scientific research investigating dynamic complex systems and representing them in computational models often starts with simple existence-proof models and systematically incorporates data and adds complexity over time. Research that is able to represent dynamic complex systems with levels of detail that span a spectrum, from simple existence proofs to high-fidelity models, is able to employ a suite of models to gain understanding and make scientific advancements.

Furthermore, while high-fidelity models may be able to encapsulate a large number of relevant processes driving a system of study and potentially provide for greater realism, they are often difficult to validate against historical outcomes, generalize across space and time, and communicate to stakeholders and policymakers. In contrast, a parsimonious model (i.e., a simplification of reality designed to improve understanding) can be a more effective tool for communicating process behaviors and outcomes and can provide a more efficient platform for understanding general processes and identifying new data that need to be collected, deriving new theory and thought experiments, and testing new policies or interventions. “All models are wrong, only some are useful” (Box and Draper 1987, 424) and the tradeoff between realism and generality is a critical element determining their usefulness.
This entry has shown that the representation of a dynamic complex system requires attention to the multiscale relationships and interactions among spatial entities over space and across time, and to the spatial patterns that emerge through those interactions through aggregation. Explicit representation of system components and outcomes provides an opportunity for geographic science to capitalize on the holistic approach to scientific inquiry that reaches back to its origins and make significant spatially explicit advances in science and society.

**SEE ALSO:** Agent-based modeling; Cellular automata; Geocomputation; Geography and the study of human–environment relations; Microsimulation; Spatial analysis; Spatial social networks; Time geography and space–time prism; Tobler’s first law of geography

**References**


While we all have a certain idea of what “space” is because we constantly interact with it, most of us would be hard pressed to give a definition. The nature of space has been debated for a long time, and no clear and uncontroversial definition exists because different disciplines (e.g., mathematics, physics, and geography) have different definitions, and these are context-related. However, we know that humans conceptualize space with two contrasting views: the object and the field views. The former considers space as being “empty” and populated with discrete entities embedded in space and having their own properties. We usually use this view when we reason about the spatial arrangement of the buildings and streets in our neighborhood, or about the pieces of furniture in our house. The latter view considers space as being continuous, and every location in space has a certain property (there is something at every location); entities are formed by clusters of properties. We usually use this view for geographical phenomena such as the amount of rainfall in a given province, or the temperature of a body of water.

When we want to represent and store a certain part of space in a computer, that is, to build a spatial data model and realize it with a data structure, the field-view approach is much more problematic than its counterpart. The causes of this are many. First, a field is intangible and is not part of our intuitive knowledge. It is easy for us to see and describe entities such as churches and rivers, but, although we can imagine a field, it is somewhat of an abstract concept. Second, fields are by definition continuous, while computers are discrete machines. Third, there is much confusion among users because for years the geographic information system (GIS) world has reduced each view to one of the two spatial models: “object = vector model” and “field = raster model.” Goodchild (1992), among others, explains that this standpoint is simply false as both views can be stored with either spatial model. For example, the elevation of a terrain can be represented with a raster or with a set of triangles, and buildings in a city with polygons or with a binary raster.

This entry begins with the definition of a field, as usually used in GIScience. It should be noticed that fields as found in GIS-related disciplines are not restricted to two dimensions, and we consider fields of three and higher dimensions (for spatial-temporal modeling). Different strategies to represent higher-dimensional fields in a computer are described under “Strategies to represent fields in a computer,” below. These strategies take into account the fact that while geographical fields are continuous, we first have to collect discrete samples to study them. The section “Samples + global interpolation rules” presents different interpolation methods that can be used to reconstruct the continuity of a field from its set of samples. The section “Piecewise spatial models” presents different spatial models to discretize a field into a finite number of parts; these spatial models can be based on the samples or be arbitrarily subdividing the space. For each of the spatial models, potential data structures for an efficient implementation in a computer are briefly discussed.
What is a field?

The definition usually used in GIScience is borrowed and adapted from physics. Physicists in the nineteenth century developed the concept of a force field to model the gravitational force, where a force (a vector with an orientation and a length) has a value everywhere in space, and changes from location to location. For most GIS applications, the vector assigned to each point is replaced by a scalar value, and we obtain scalar fields; unless otherwise explicitly stated, in the following all fields are scalar.

A field is a model of the spatial variation of an attribute $a$ over a spatial domain; we assume this domain to be $\mathbb{R}^d$, the $d$-dimensional Euclidean space. It is modeled by a function mapping one point $p$ in $\mathbb{R}^d$ to the value of $a$, thus

$$a = f(p)$$

The function can theoretically have any number of independent variables (i.e., the spatial domain can have any dimensions), but in the context of geographical phenomena the function is usually bivariate $(x,y)$ or trivariate $(x,y,z)$.

Bivariate functions are commonly used in GIS to model the elevation of a terrain. However, it should be noticed that the dependent variable $a$ (the elevation in this case) can have one and only one value for each location $(x,y)$. This restricts the real-world cases that can be modeled because, as shown in Figure 1, vertical surfaces (e.g., walls of a building if we model all human-made objects with the terrain to construct a digital surface model), overfolds (e.g., the balcony of a house), and caves are impossible to represent. Modeling the elevation (or other attributes) with a bivariate function is sometimes referred to as “2.5D GIS”; we model the function with a surface (i.e., topologically a two-dimensional (2-D) object) but we embed this surface in $\mathbb{R}^3$.

Geographical phenomena tend to vary over time, and, as a consequence, we obtain dynamic fields. If we assume that time is continuous and linear, it is possible to incorporate it as an extra dimension, perpendicular to the spatial ones. It becomes a new independent variable, and if we already had a trivariate function $(x,y,z)$ then it becomes a quadrivariate one $(x,y,z,t)$. A dynamic field representing the salinity of a body of water over time is thus modeled in $\mathbb{R}^4$.

The value of the attribute $a$ can be measured according to different scales, and depending on the scale used different types of fields will be obtained. We distinguish between two types of field.

**Continuous fields**: the scale of measurement is a real number $R$. Most fields studied in the geosciences fall into this category: temperature, precipitation, or salinity are examples because they can be measured precisely. The measurement scales “ratio” and “interval” are of this type.

**Discrete fields**: the possible values of an attribute $a$ are simply labels. An example is a map of Europe where each location contains the name of the country. The labels can also be

![Figure 1](image)

Top: a profile view of a terrain, with one tree and one house with a balcony. Bottom: the profile view if the elevation is a dependent variable of the $(x,y)$; notice that the walls of the house and the sides of the tree are not fully vertical (dotted lines).
ordered; for example, a certain region can be categorized from 1 to 5 according to its suitability for producing wine, 1 being poor, and 5 excellent. These are respectively the scales “nominal” and “ordinal.” It should be observed that in the former case the only comparison possible between two values is whether they are the same or not; in the latter case, operations such as “greater than” and “less than” can be made, but no arithmetic operations are possible. Discrete fields act also as objects because each region can be modeled as a unique object.

Notice that the terms “continuous” and “discrete” can be misleading because in mathematics continuity has a slightly different meaning. Here they refer only to the scale of measurement, and not to the spatial continuity of a field. Indeed, both types of fields are spatially continuous, since they are represented by a function and there exists a single and unique value at every location in space; this is called a \( C^0 \) function in mathematics. For the representation of fields, it is, however, often desirable to have a function for which the first or second derivative is possible everywhere; such functions are respectively referred to as \( C^1 \) and \( C^2 \).

**Strategies to represent fields in a computer**

The representation of a field in a computer faces many problems. First, fields are continuous functions, and, by contrast, computers are discrete machines. Fields must therefore be discretized, that is, broken into finite parts. Second, in practice it is usually impossible to measure continuous phenomena everywhere, and we have to resort to collecting samples at some finite locations and reconstructing fields from these samples. The discretization task therefore begins at the acquisition phase, and is affected by the acquisition tools and techniques. This fact is aggravated for fields as found in GIS-related disciplines because, unlike disciplines like medicine or engineering, we seldom have direct access to the whole object of interest. Indeed, to collect samples in the ground we must dig holes or use other devices (e.g., ultrasound penetrating the ground); underwater samples are collected by instruments moved vertically under a boat, or by automated vehicles; and samples of the atmosphere are collected by devices attached to balloons or aircraft. Moreover, because of the way they are collected, geoscientific datasets often have a highly sparse and anisotropic distribution: as shown in Figure 2, the distribution can be, for instance, dense vertically (with a sample every 2 m in that real-world case) but extremely sparse horizontally (water columns are located at about 35 km from each other). Even if a sensor is used to collect samples, the result (e.g., an image with pixels) is not a complete representation since each pixel usually averages the value of the studied phenomenon over the pixel area, or each pixel represents the value for one location inside the pixel. Third, another peculiar property of geoscientific datasets is that the samples they contain represent the spatial variability of a given attribute only at time \( t \), and many geographical phenomena in geosciences change and evolve relatively quickly over time.

**What is needed to represent a field?**

To represent a field in a computer, two things are needed.

1 A set of samples that were collected to study the field; for example, these can be the elevation of a terrain obtained with airbone laserscanning or the humidity of
REPRESENTATION: FIELDS

Figure 2 An oceanographic dataset in the Bering Sea in which samples are distributed along water columns. Each red point represents a (vertical) water column, where samples are collected every 2 m.

the soil collected with a certain device. This set forms the “ground truth” of the field, and its importance is such that it has been dubbed the meta-field by Kemp and Věkovskí (1998).

2 A set of rules to obtain one and only one value for the attribute studied at any location, in other words, to reconstruct the continuity of the phenomenon studied from the discrete samples. This operation is referred to as spatial interpolation.

The section “Samples + global interpolation rules,” below, presents the first strategy: storing the original samples with the parameters of the global spatial interpolation method that is best suited to the distribution of the samples and their accuracy. Notice that this strategy permits us to reconstruct the continuity of a field from the samples by calculating the value of the attribute, but that this value is not persistently stored in the computer.

The section “Piecewise spatial models” presents an alternative strategy: the spatial interpolation function used is piecewise (instead of being global). That is, the field is tessellated, or partitioned, into several pieces, and for each of these we assign an interpolation function describing the spatial variation. Piecewise models imply that a supporting data structure is constructed, and stored, to represent the tessellation (it is, however, possible to construct on-the-fly some well-known structures only when they are needed since they have formal rules). Some of these tessellations partition arbitrarily the space, while some are based on the set of samples that were collected.

Incomplete representations

In the GIS literature, different representations for 2-D fields are often listed. For instance, these six are common:

1. regularly spaced sample points;
2. irregularly spaced sample points;
3. contour lines;
4. rectangular cells (raster);
5. triangulated irregular networks (TIN);
6. planar partition with arbitrary polygons.

While these are by no means exhaustive, they represent how a typical practitioner would store her 2-D fields. Theoretically, these representations all generalize to three and higher dimensions (these are discussed in “Piecewise spatial models,” below).

Although the first three representations are commonly used, they are incomplete: the set of rules to reconstruct the continuity of a field at unsampled locations is not explicitly given. These should therefore not be considered valid representations of a field. It should be noticed that while finding a spatial interpolation function suitable for the first case is simple (a few are discussed below), for the third case it is a complex task. Indeed, Dakowicz and Gold (2003) demonstrate that using simple rules (nearest-neighbor for instance) yields fields that are not realistic and
have bad slope, which is in practice problematic for several applications. Obtaining good slope from contour lines is possible, but is in practice a complex process.

Samples + global interpolation rules

We consider as “samples” any data that were collected to study a phenomenon. In practice, these data can take different forms.

1. A set of scattered points.
2. A set of lines. The most common case is a set of contour lines coming from a topographic map.
3. A set of scattered polygons (or polyhedra or polytopes) to which one value is attached. Although this case is theoretically possible, in practice this is very rare. It is nevertheless always possible to discretize each primitive into a set of points, and then perform interpolation.

Observe that it is also possible to collect directly a raster image, coming from remote sensing, where the value of each pixel represents the temperature of the sea, for instance. The value is the average of all the values within the pixel, and this case is considered as being already tessellated (the pixels) and is discussed in the next section.

Interpolation methods are rather difficult to categorize because they are based on different paradigms, and some methods fall into more than one category. No attempt is made in this entry to introduce categories; the reader is referred to the entries Interpolation: inverse-distance weighting and Interpolation: kriging.

The following are a few examples of “global” interpolation methods, that is, the same function is applied everywhere to a subset of (regularly or irregularly spaced) sample points (and the function is not piecewise, as in the section “Piecewise spatial models”). Observe that these methods can all be used in three and higher dimensions.

Nearest neighbor: the simplest interpolation method. The value of an attribute at location \( x \) is assumed to be equal to the attribute of the nearest sample point.

Inverse-distance weighting (IDW): as described in the entry Interpolation: inverse-distance weighting, it requires different parameters to define which points are involved in the interpolation at a given location, and also the power must be defined. The parameter to define the points is usually a searching radius, which can be used in higher dimensions.

Kriging: as described in the entry Interpolation: kriging, while the modeling of a dataset is a difficult and time-consuming task, the output of the modeling process (a function characterizing the dependence between the attributes of any two samples that are at a given distance from each other) can be simply stored as a string.

Notice that while it is theoretically possible to find a polynomial function of high degree for any set of real-world samples, in practice this is very rarely used because, as the degree of the polynomial increases, so can the error at unsampled locations. That is, the error at the samples is zero, but between two samples the function returns a value that can be far higher or lower than the two values of the samples (known as the Runge’s phenomenon). Piecewise functions formed by many polynomial functions, known as splines, are usually preferred and used in GIS (see Mitasova and Mitas 1993 for instance).

Storing explicitly, the interpolation method is efficient because in practice only a few parameters have to be stored. For instance, in the case
of IDW, it suffices to store the search radius and the power used (two numbers). Finding the appropriate values for interpolation parameters is a difficult and time-consuming task, as the user must have a good understanding of the spatial distribution of the samples, and of the details of the method. A vivid example is kriging, with which experienced users can obtain very good results, but which also leaves newcomers clueless with its many parameters and options. Using kriging with the appropriate parameters leads to results having statistically minimum variance; however, simply using the default values for the parameters will most likely lead to unreliable results. Thus, it is better to let specialists perform the modeling of a dataset and find the best interpolation method (and the parameters to use), and store only the results of that modeling. The practitioners can then reconstruct the field without having to find the optimal parameters, provided they have access to software to perform the interpolation.

Piecewise spatial models

In this section different spatial models needed to represent tessellations of space (this space can be in our case two-, three-, or four-dimensional) are presented and discussed. While the same spatial models can in theory be used in any dimensions, their implementations with a specific data structure for the three- and higher dimensional cases are more complex than the 2-D case, and for this reason, potential data structures are also briefly discussed.

Once the \( d \)-dimensional space covered by a field is tessellated, the field function becomes piecewise: to each \( d \)-dimensional element of the tessellation – which is named a cell in the following – a function is assigned describing the spatial variation in its interior. This function is usually a simple mathematical function:

- a constant function: the value of the modeled attribute is constant within one cell, for example to represent a discrete field;
- a linear function;
- a higher-order function.

In general, we classify the tessellations of space into three categories (as shown in Figure 3): regular, hierarchical, and irregular.

Regular tessellations

In regular tessellations all the cells have the same shape and size. The most common regular tessellation by far in GIS is the grid (or raster representation), in which the cells are squares in 2D (usually called pixels, a portmanteau of “picture” and “element,” as an analogy to digital images). However, while they are not common in practice, other regular shapes are possible, such as hexagons or triangles. Regular tessellations generalize to higher dimensions: for instance, in 3D, instead of squares we have cubes (called voxels; “volume” and “element”), and tesseracts in 4D.

Observe that a regular tessellation often arbitrarily tessellates the space covered by the field without taking into consideration the objects embedded in it (the samples) or the nature of the phenomenon modeled. This is in contrast with irregular tessellation in which, most of the time, the cells constructed are constrained by the samples. Kemp (1993) states that “[a 2-D grid] requires us to enforce a structure on reality rather than allowing reality to suggest a more appropriate structure for our analysis.” In practice this means that, unless a regular tessellation is obtained from an image sensor (remote sensing or photogrammetry), we can
Figure 3  Types of tessellations and one example for each in two and three dimensions.
assume that it was constructed from a set of samples by using spatial interpolation. Converting sample points to cells is not optimal because the original samples, which could be meaningful points such as the summits, valleys, or ridges of a terrain, are not necessarily present in the resulting grid.

Concrete example: a 2-D grid. A 2-D grid, stored for instance with the GeoTIFF format, is thus a piecewise representation of a 2-D field: a regular tessellation where each cell has a constant function. The value assigned to each cell is either an estimation previously obtained by spatial interpolation, or, in case of aerial imagery, the average values inside that cell. However, for a given grid, it is usually unclear if the value of a cell is for its center, or for one of its vertices (and if the latter is the case, for which one). Different formats have different rules, and converting a field represented with one format to another one (while retaining the same cell resolution and spatial extent) can shift the value from the center to the top-left corner for instance.

Concrete example: a 3-D/4-D grid. As is the case for 2-D grids, 3-D/4-D grids are piecewise representations, and usually a constant function is used for each cell. Several implementations of a voxel-based GIS exist, notably the open-source system GRASS (Geographic Resources Analysis Support System; http://grass.itc.it) in which diverse operations such as interpolation and visualization are possible. Also, most of the earlier attempts at building 3-D GISs were made using voxels. Four-dimensional grids can be stored, among other ways, in the netCDF format (http://www.unidata.ucar.edu/software/netcdf/). It can actually be used for d-dimensional grids, with different spacing for different dimensions. They are binary and spatially structured, which means that parts of a dataset can be efficiently retrieved and processed.

While the function assigned to each cell is often a constant, other functions can be used. Notice that a linear function is impossible (unless, for instance, in 2D, the four vertices of the cells are coplanar), and instead bilinear interpolation is used. The method, which can be seen as an extension of linear interpolation, performs linear interpolation in one dimension (e.g., along the x axis), and then in the other dimension (y axis). The result is independent of the order in which the axes are used, and can be generalized to higher dimensions.

The wide popularity of regular tessellations in 2-D and 3-D GIS applications is probably due to simplicity and to the fact that they permit us to easily integrate 2-D remote sensing images and fields. Indeed, a grid is naturally stored in a computer as an array (each cell is addressed by its position in the array, and only the value of the cell is stored), and thus the spatial relationships between cells are implicit. This is true for any dimensions, thus, contrary to other tessellations, grids are very easy to generalize to higher dimensions. The algorithms to analyze and manipulate (Boolean operations such as intersection or union) are also straightforwardly implemented in a computer. On the other hand, grids also suffer problems. First, the size of a grid can become massive for data at a fine resolution; this problem gets worse in higher dimensions. Second, grids scale badly and are not rotationally invariant, that is, if the coordinate reference system used is changed, then the grid needs to be reconstructed to obtain regular cells whose boundaries are parallel to the axes of the reference system. To assign a new value to the transformed cells, spatial interpolation is needed, which is often performed not by reusing the original samples, but by using the values of the
neighboring cells. Unfortunately, each time a grid is transformed its information is degraded because what is used is not the original samples but interpolated values.

Hierarchical tessellations

Hierarchical tessellations attempt to reduce the number of cells in a tessellation by merging the neighboring cells having the same value (thus yielding cells of different sizes). While both regular and irregular tessellations can be hierarchical, in the context of the representation of fields, the former is more relevant and is often used in practice. Irregular hierarchical tessellations, usually triangulations, are used for obtaining multi-resolution models (De Floriani and Magillo 2002), or act as a spatial index for efficiently accessing triangles. Their use as a support to construct a piecewise function is not necessary since, as shown in the section “Piecewise spatial models,” irregular tessellations have cells of different sizes and shapes. Consequently, only hierarchical regular tessellations are discussed in the following.

A commonly used hierarchical structure in two dimensions is the quadtree, which is a generic term for a family of tessellations that recursively subdivide the plane into four quadrants. As is the case for grids, quadtrees are relatively easily implemented in a computer because they are trees in which each node has exactly four children, if any. Different variants of quadtrees exist, but the most relevant type in the context of this entry is the region quadtree. As shown in Figure 3, it indexes and merges adjacent grid cells to save space by recursively subdividing the space into four squares of equal size, until every square contains one homogeneous region (based on the attribute of every grid cell, the color in the figure). A region quadtree offers a more adaptive subdivision of space than a regular tessellation because smaller cells are present only when needed (to capture more details).

The region quadtree generalizes to higher dimensions. For instance, in three dimensions, the space is recursively subdivided into eight octants, and thus the tree is called an octree (each node has eight children). In four dimensions, space is recursively subdivided into 16 cells (each one of them being a tesseract).

Examples of the use of regular hierarchical tessellations in GIS for representing fields are more numerous in 3D than in 2D, probably because the space issue (i.e., space used to store a field in a computer) is more critical. However, it is possible that a GIS stores and handles internally 2-D grids as region quadtrees – the user does not need to know as no information is lost in the conversion, and important space and processing gains can be achieved.

The shortcomings of regular hierarchical tessellations are similar to those of regular tessellations: the rotation and scaling operations are difficult to handle. The main advantage of using them – saving memory space – is present only when there is spatial coherence between cells having the same attribute value, that is, when they are clustered together. Indeed, the size of a quadtree is not dependent on the number of cells but on their distribution. The quadtree of a 2-D grid having no two adjacent cells with the same value contains the same number of cells as the grid, and its size would most likely be worse because of the overhead to manage the tree.

Irregular tessellations

The cells of an irregular tessellation can be of any shape and size, and they usually “follow” – or are constrained by – the sample points that were collected to study the field, albeit this is not a requirement. Subdividing the space based on the samples has the main advantage of producing a tessellation
that is *adaptive* to the distribution of the samples (and thus to the complexity of the phenomenon studied). The subdivision is potentially better than that obtained with quadtree/octree (which subdivide arbitrarily the space without any considerations for the samples).

The most known examples of the use of irregular tessellations in GIS for representing fields are arguably the choropleth map, the triangulated irregular network, and the Voronoi diagram.

A choropleth map subdivides the plane into arbitrary polygons, and a constant function is used within each polygon. They are mostly used to model and visualize discrete fields, such as densities, rates, or proportions. Figure 4 is an example that shows the percentage of people above 65 years of age per neighborhood in a part of the Netherlands. Observe that since the field modeled is discrete, using other functions (linear or quadratic) would not make sense. Operations that are possible with continuous fields, for example interpolation, are also impossible.

The generalization to higher dimensions of choropleth maps, that is, of an irregular tessellation of space into cells of different shapes and size, is possible. In three dimensions, the cells are polyhedra, and these are used, among other areas, in geological applications where one type of rock is attached to each polyhedron. In four dimensions, the cells are polytopes, and can be conceptualized as polyhedra modeled over time.

The storage of a discrete field such as a choropleth is done with one of the data structures described in the entry Data structure, vector, as this is simply a set of polygons to which one attribute value is attached. For the three- and four-dimensional cases, more advanced data structures have to be used. One example is the structure *generalized maps*, as first introduced in Lienhardt (1994). It is capable of representing a wide class of objects in arbitrary dimensions. Practically, it also has the advantage of having been implemented in higher dimensions, and of being used in different systems in three dimensions (e.g., it is used in RING (Research for Integrative Numerical Geology; www.ring-team.org) for geological modeling and in Moka (http://moka-modeller.sourceforge.net/) for geometric modeling).

As shown in Figure 5, a triangulated irregular network (TIN) refers to an irregular tessellation of a 2-D field into nonoverlapping triangles (whose vertices are formed by three sample points), and to the use of a linear interpolation function for each triangle. While in theory any 2-D field can be modeled with a TIN, it is in practice mostly used for representing the elevation of an area because the lifting of each sample point to its elevation creates a surface, embedded in three dimensions, approximating the morphology of the terrain. The value of the attribute of a field at an unsampled location $x$ is obtained by linearly interpolating on the plane passing through the three vertices of the triangle containing $x$. TINs are the most popular

![Figure 4](image-url)
alternatives to 2-D grids for modeling elevation; both representations have advantages and disadvantages.

Given a set of points in the plane, there are several ways to construct a triangulation. A TIN usually refers – although this is not a strict requirement – to a Delaunay triangulation (DT) of the set $S$ of sample points projected on the 2-D plane. A triangle respects the Delaunay criterion if the unique circle on which its three vertices lie does not contain any other point in the set $S$ (see Figure 5). This also means that, among all the possible triangulations of a set $S$ of points, the DT maximizes the minimum angle (the max–min property) and minimizes the maximum circumradii; in other words, it creates triangles that are as equilateral as possible. This property is useful in applications where interpolation is necessary (such as the reconstruction of a field) since with long and skinny triangles the estimation of the field attribute at location $x$ by linear interpolation in the triangle will use samples that can be far away from $x$. In a DT the three vertices of the triangle containing $x$ will most likely be close to, and around, $x$.

A TIN in which a linear interpolation function is used yields a $C^0$ piecewise representation, that is, it is a continuous function but at the edges of the triangles the first derivative is not possible. Akima (1978) shows the advantages of using higher-order functions in each region of a TIN, to construct a $C^1$ or $C^2$ field. In other words, if elevation is the attribute modeled, the slope of the terrain will be “good,” which is important for several applications, for example flood modeling.

Storing a TIN can be done with any of the data structures described in the entry Data structure, vector since a TIN is formed by a set of triangles. However, more efficient data structures, in terms of storage space and in terms of complexity, exist. The simplest of these, the triangle-based data structure, considers the triangle as being its atom and stores each triangle with three pointers to its vertices and three pointers to its adjacent triangle; it can be easily implemented in an array. Shewchuk (1997) reports the implementation of his program to compute the DT to be nearly twice as fast when this structure is used instead of a generic structure for polygons. It is being used in several packages, notably by the open-source project CGAL (Computational Geometry Algorithms Library; www.cgal.org). It is also possible to see a TIN as a graph and to compress that graph, see for instance Blandford et al. (2005). The data structure is more complex to construct, but can take five times less memory than the alternatives mentioned above. It should, however, be noted that, from a practical point of view, if the triangulation used is Delaunay, it is easier to store only the sample points and reconstruct on the fly the DT when needed since the DT is unique for a given set of points (if no four or more points are cocircular). This is conceptually equivalent to the solution described in the section “Samples + global interpolation rules,” and it is known that the DT of a set of points can be reconstructed very efficiently (Snoeyink and van Kreveld 1997).
The generalization to three dimensions of the Delaunay triangulation is the Delaunay tetrahedralization: each triangle becomes a tetrahedron that satisfies the empty circumsphere rule. However, obtaining nicely shaped tetrahedra that can be used for linear interpolation of a field is more difficult than finding good triangles because the max–min property of 2-D Delaunay triangles does not generalize to three dimensions (nor to higher dimensions). A 3-D DT can indeed contain some tetrahedra, called slivers, whose four vertices are almost coplanar, and thus have a volume of nearly zero. (Observe that slivers do not have 2-D counterparts.) For the reconstruction of a 3-D field with linear interpolation inside a tetrahedron, these tetrahedra are not optimal and should be removed. This is unfortunately a rather convoluted operation. However, it should be said that in most cases Delaunay tetrahedra have a more “round” shape than arbitrary tetrahedra.

The DT generalizes to four and higher dimensions; the circumball around each simplex (the simplest element in a given dimension: vertex, edge, triangle, tetrahedron, etc.) is empty of any other points. It can be constructed, among other ways, with the algorithm and the implementation described in Barber, Dobkin, and Huhdanpaa (1996).

Storing a 3-D or 4-D Delaunay triangulation is simpler than storing higher-dimensional arbitrary subdivisions of space, and can be done with a straightforward generalization of the triangle-based data structure: the atom is a $d$-dimensional simplex having $d+1$ pointers to its vertices and $d+1$ pointers to its neighbors. However, as is the case for the cases in 2D and 3D, the reconstruction on the fly is a simpler alternative in practice.

Another example of an irregular tessellation is the Voronoi diagram (VD) of a set of points in a $d$-dimensional space, where each point is mapped in a one-to-one way to a cell representing the space closer to the point than to any other in the set. Figure 6 shows an example, and more details about the properties of the structure are discussed in the entry Spatial modeling: Voronoi diagrams.

The VD is the geometric dual structure of the Delaunay triangulation (DT), and that in any dimensions. In brief, it means that both structures represent the same thing, but from a different viewpoint. The sample points collected to study a field are not on the boundary of the cells of a VD, but are rather the points generating each Voronoi cell (they are “in the middle”). As is the case with the DT or other irregular tessellations, perhaps the main advantage of the VD for representing fields is that it permits us to preserve the original samples. The main advantage of the VD over the DT is that, for the three- and higher dimensional cases, the shape of the cells is always “good,” that is, even if slivers are present in the DT, the Voronoi cells will still be “relatively spherical” (which is an advantage for interpolation).

If a constant function is assigned to each Voronoi cell, the VD permits us to elegantly represent discrete fields. Indeed, if sample points were collected to study a discrete field, it makes little sense to use spatial interpolation functions such as IDW; the VD offers a natural way to reconstruct the field. To know the value at a location $x$, one simply has to find the Voronoi cell containing $x$ and query its attribute value. Observe that this is conceptually equivalent to

Figure 6 The two-dimensional Voronoi diagram. The cell in gray represents the space closer to the point in the middle of the gray cell than to any other point.
storing the sample points and using a nearest neighbor function, as explained in the section “Samples + global interpolation rules.”

To reconstruct a continuous field from a set of samples, more elaborate functions are needed since the VD creates discontinuities at the border of each cell. An interesting one is the natural neighbor interpolation method (Sibson 1981; Gold 1989) because it has been shown by different researchers to have many advantages over other methods when the distribution of samples is highly anisotropic: it behaves consistently whatever the distribution of the samples is, and it is an automatic method that does not require user-defined parameters. The method is entirely based on the VD for both selecting the samples involved in the interpolation process, and to assign them an importance. The neighbors used in an interpolation are selected using the adjacency relationships of the VD, which results in the selection of neighbors that both surround and are close to \( x \), the interpolation location. A field reconstructed with the natural neighbor function will be smooth and continuous everywhere (\( C^1 \)), except at the samples themselves. It is possible to use that method in any dimensions (see Ledoux and Gold 2004).

Since the VD and the DT are dual structures, the knowledge of one implies the knowledge of the other. In other words, if one has only one structure, she can always extract the other one. Because it is easier, from an algorithmic and data structure point of view, to manage simplices over arbitrary polytopes (they have a constant number of vertices and neighbors), constructing and manipulating a VD by operating only on its dual structure is simpler and usually preferred. In the context of representing fields, there is therefore no need to store a VD. When the VD is needed, it is extracted on the fly from the DT. And since the DT is rarely stored in a computer as far as the modeling of fields is concerned, neither is the VD.

**In brief**

The choice of a spatial model to represent geographical phenomena is a fundamental issue of spatial data handling, and unfortunately one that is more than often overlooked by GIS practitioners. For a given task, practitioners tend to simply use the spatial models, and data structures, available in commercial GISs, without assessing the consequences of their choice. This is particularly problematic for the representation of fields as found in GIS-related disciplines because, regardless of a few notable models developed in academia, the only solutions available are grid-based models, and this is especially true in dimensions higher than two. The popularity of grids is probably due to their conceptual simplicity and to the fact that they are easily and naturally stored in computers. However, as was demonstrated in this entry, they have many shortcomings, both conceptually and technically. They tessellate arbitrarily the space without taking into consideration the objects embedded in that space (or the phenomenon being studied), which yields an unnatural discretization. Moreover, grids need to be resampled each time the cartographic projection used is modified, and each resampling degrades the information stored in the file, and can lead to errors and misinterpretations during the analysis.

The alternative representations presented in this entry are often more complex to construct and store in a computer, but practitioners should consider them because they would often permit them to better represent the phenomena they are studying. Irregular tessellations indeed permit us to obtain a subdivision that is based on the samples collected (which are preserved, by opposition to grids where they are “lost”), and the cells constructed are adaptive to the spatial distribution of samples, which is crucial when
dealing with highly anisotropic distributions such as the ones found in the geosciences.

SEE ALSO: Data structure, vector; Digital elevation model and digital surface model; Interpolation: inverse-distance weighting; Interpolation: kriging; Spatial modeling: Voronoi diagrams

References


Further reading


Flow is the collective movement of energy, matter, or related concepts. Energy and matter flow includes electromagnetic fields, gravity fields, gases, liquids and solids, and self-changing organized matter in the form of biological entities. Matter flow is affected by the local energy field, that is, gravity or electromagnetism. Likewise, some energy flows, for example wind, are mediated by mass flow. Flow can also be used in a very conceptual way, such as flow of genes or flow of ideas. Biological entities, ranging in size from blue whales to viruses, can act as a conduit for genes of their own or genes from infectious agents such as parasites, bacteria, or viruses for larger organisms down to virophage infections for viruses. Biological entities also can drive flows of learned memes, ranging from simple behaviors to complex concepts. Nevertheless, in geographic domains, flows commonly refer to the continuous movement of water, air masses, lava, population, and traffic, as well as a diversity of dynamic fields in physical and human environments.

Flow can be tangible or intangible. Intangible conceptual flow, such as memes, requires carriers to transmit the concepts to other entities. The carriers interact in physical space or cyberspace mediating the flows of information, impalpable commodities, or concepts. The total flow is the result of both the changes in positions of entities that act as carriers and communication among entities. Flows may or may not be subject to rules of conservation: energy, matter, financial instruments, or commodities should account for everything throughout the courses of flow; yet, information communication or social constructs may experience gains or losses in flow.

Flow is represented either continuously or discretely. Physical fluids, for example atmospheric and surface water, are composed of innumerable individual constituents with viscosity and plasticity to move as a continuous whole. Representation of physical fluid is most common as two-dimensional or three-dimensional grids of data fields. In contrast, flows of people or other discrete entities are usually treated as a collective whole (e.g., people to population and vehicles to traffic) that moves in defined paths between end points. A common representation for discrete flows (e.g., a migration) is in the form of interactions between the two end points, using both arcs and nodes to show the magnitude of flows (e.g., the volume of traffic) from point to point or in an interaction matrix to enumerate the flow magnitude from origins to destinations.

Discrete models of arcs and nodes are popular in map representation of geographic flow. This tendency has continued in geographic information systems (GIS). An early geographical flow map was compiled by Henry Drury Harness (Robinson 1955; Figure 1) for railroad traffic analysis in the Atlas to Accompany the Second Report of the Commissioners Appointed to Consider and Recommend a General System of Railways for Ireland in 1837. Travel was shown as the actual rail path traveled between nodes of locations. Harness used a dasymetric method to calculate population densities for areas around the nodes.

The flow maps of Harness were simplified when Charles Joseph Minard produced the map of “Napoleon’s Russian campaign of 1812”
in 1861. The path information in the maps of Harness were removed to provide space for other variables. Tufte and Graves-Morris in their 1983 book *The Visual Display of Quantitative Information* highlighted the six variables captured in the map, with the two most important being line width to portray the number of troops and nodes to portray skillfully significant events in space and time. Algorithms for creating Sankey maps have been developed to create unique flow maps by selecting relationships and aggregation levels for nodes. The System for Automated Graphics and Explanation (SAGE) created a dynamic “March on Moscow” Sankey map by selecting relations.
among space, time, temperature, and troop strength (Friendly 2002). In addition, other research resulted in algorithms that enabled dynamic aggregation of nodes to simplify navigating the map and enable zooming in and out (Riehmann, Hanfler, and Froehlich 2005).

Tobler (1981) considered both discrete and continuous representations of flows in modeling geographic movement. Discrete flow maps illustrate flows using arrows with widths to show directions and magnitudes. The arrows do not fill the corresponding regions but connect a sequence of objects on the map. Continuous flow maps use vector fields that fill the region to represent continuous flow patterns that are defined at all locations. Vector fields are associated with physics and meteorology where they represent the flow of fluid through space. The vector field requires the formation of streaklines to determine where a fluid parcel travels with the passing of time. A continuous flow map of migration using streaklines has a very different appearance from that of a discrete map with weighted lines of thickness representing the movement of groups of people. Expanding upon Tobler’s work, discrete flow representation has been used by human geographers to model human migration flows, commuting patterns, traffic flows, commodity flows, and information flows. The idea of flows is instrumental to Paul Krugman’s economies of scale that produced flows of goods between regions as a result of production concentration.

The idea of flows is also common in the study of physical geography. Physical geographers use continuous flow representations to portray near-surface ocean currents in coastal waters, wind direction on meteorological maps, and glacial movement on geomorphic maps. Continuous data is often aggregated to create a discrete representation. Aggregation simplifies the data representation and analysis while providing feature abstraction that lessens the generalization required for map legibility. Extending flow maps beyond representing aggregated matter flow described in external data sources requires the ability to model dynamic systems within GIS. Cellular automata and vector raster operators have been integrated with GIS in an effort to represent and map dynamic systems.

Besides meteorology and oceanography, hydrology research also involves many kinds of flows, such as two-dimensional groundwater flow, runoff routing, and particle tracking in streams. Different perspectives of dynamics complicate the ways in which hydrological flows should be represented in GIS (Gold 2006), especially the need to simultaneously accommodate both Lagrangian (following a particle) and Eulerian (particles passing a cross-section) views of flows. Nevertheless, the current GIS framework for hydrological flow modeling starts with a hydrologically conditioned digital elevation model (i.e., a digital elevation model adjusted with known streams, water bodies, and other notable hydrological features) to calculate slope, flow directions, and flow accumulation. Flow direction is determined as being from the focal cell to its neighbor with the steepest slope. Flow accumulation is determined by the sum of the total land units (cells) along the path of steepest slope up to the point of interest. Results commonly are summarized into a vector representation, with nodes representing intersections and arcs representing flow paths. Often, underground flows have only origin and destination connections determined, with no path information.

Flows of concepts, ideas, and beliefs influence behaviors, ideology, and culture. Behavioral memes flow through generations and beyond and eventually contribute to the formation of social norms and cultures. Cultural memes, or singular cultural concepts, are transmitted as flows by humans who can consciously or accidentally
modify them. Flows of cultural memes can be traced back to the group of origin and can displace a region’s original culture. Nevertheless, cultural flow may be resistant to change as it is an intragenerational and intergenerational transmission of ideas that are realized in modes of cultural behavior. Political resistance can modify a cultural flow by importing cultural flow from a different region to show resistance to local control.

Flow representation is shaped by the need to abstract and generalize the idea that some quality and quantity may vary along with a movement or displacement between two locations. The flows between discrete nodes represent spatial interactions between sources and destinations. Conventional flow matrices do not record the travel paths but merely the start and end of the aggregated objects. Creating a flow map from the interaction matrix is possible but is of limited use. Further simplification can be achieved by using the gravity model to define interactions between all possible pairs. Gravity models are likely the most widely used spatial interaction representation.

GIS does not model flows well due to a lack of a suitable spatiotemporal representation to capture the transition of varying quality and quantity along the paths of movement. Temporal representation solutions have utilized domain-specific ad hoc conceptualizations to enumerate and characterize the relationships and rules. Yuan (2001) reviewed the five main approaches to temporal GIS modeling and found that all build on the scalar data of location, objects, and events. This focus makes it difficult to represent dynamic fields based on data derived from partial differential equation models of physical processes such as water-flows, sediment transport, pollution and waste diffusion, and cyclones. Many real-world spatiotemporal phenomena go beyond the representation of object-oriented or event-based data models. Mapping these phenomena is a challenge due to the lack of GIS continuous flow representation.

SEE ALSO: Data model, moving objects; Data model, object-oriented; Representation: fields; Representation: time

References


Further reading

A map is a scale model of a geographic area. Maps work because they clearly present the geographic layers of interest and follow cartographic conventions that everyone recognizes. Motorists easily navigate road maps with highways, residential roads, and dirt roads identifiable by their common line symbols. Contour lines effectively visualize slope angle by bunching up along steep mountainsides and spreading wide on the plains. The extent of a watershed on a surface can be found by deriving ridgelines around the stream that drains an area.

A map portrays geographic information through a series of thematic overlays – streets, buildings, water, terrain, and imagery. A GIS displays a map by rendering a collection of thematic overlays. The entities in each thematic overlay (points, lines, polygons, imagery, surfaces) are drawn with symbols representing attributes such as feature types, heights, and quantities.

A geographic database is a collection of datasets containing the data within each thematic overlay. Geographic databases store sets of features by extending tables in a relational database with shape fields for point, line, or polygon geometries. These are called feature classes. Nonspatial entities that are related to features are stored as rows with attributes in a table. Imagery and continuous data are stored with efficient raster data structures scalable to large data volumes. Surfaces are represented with triangulated meshes built from many sampled points with elevation values. Upon these core datasets – tables, feature classes, rasters, and surfaces – higher spatial structures can be overlaid to provide a framework for sophisticated modeling of natural and constructed systems.

Representing discrete entities with simple features

When geographers compile maps, they make many decisions about how to represent features on a map. While geographic reality is infinitely complex, they build finite models of features in a geographic database, detailed enough to create maps and perform geographic processing tasks, yet simple enough to keep data modeling and collection costs manageable.

The spatial types of geographic features drawn on maps can be sorted by dimension – points, lines, and areas – and are described by annotations on the map. Point features on a map represent geographic entities that are too small to be depicted as lines or polygons, or locations at which a dimensional measurement is made. Line features on a map represent geographic entities that are too narrow to depict as areas at the map scale of compilation. Line features are also used for computed lines, such as survey traverses and graticules or isolines of a given attribute, contour lines being an example. Sets of connected lines form networks, such as roads and streams.

Polygon features on a map enclose areas that represent the shape and location of homogeneous areas, such as lakes, forest stands, and soil types. Polygon features are also the foundation for divisions of land, ranging from country
boundaries to land parcels. A parcel fabric is a system in which every point within a jurisdiction is covered by exactly one polygon.

Sets of point, line, and polygon features are represented in distinct feature classes which are grouped by thematic content such as transportation, hydrography, and so on. As with other database tables, every row in a feature class has a common set of attributes. Point, line, and polygon features are the components of the simple feature information model. Features, as represented in a geographic database, have several prime qualities: they occupy a geometric shape in space, they have attributes stored in a table, they have rules to enhance data integrity, they have associations with other features and nonspatial objects, and the simple feature information model can be extended in a variety of ways for rich data modeling.

A feature class is a collection of features. It is persisted in a geographic database as a relational table with a special field for shapes. Each feature shape has one or a set of \( x, y \) or \( x, y, z \) coordinates which is spatially referenced on the Earth’s surface through a coordinate system. Line features can also have measurement values for calibrating route systems. Line and polygon features are made of segments and each segment can (depending on the GIS software implementation) be a straight line, circular arc, elliptical arc, or Bézier curve. Features in a feature class are tied to locations on Earth through a coordinate system, which defines the map projection used and its parameters. All features in a feature class have the same spatial reference.

A feature class has a set of attributes for each feature in a table. Fields define feature attributes and can be numeric, textual, or descriptive. Some fields are predefined; most are user defined. Attributes can store rich content such as photographs as blobs (binary large objects). Features in feature classes and rows in tables can have attribute relationships established between them through a common field called a key. A key is an attribute of a table that uniquely identifies each record in that table. Many attributes represent measurements. Types of measurements are commonly classified into four levels: nominal, ordinal, interval, and ratio, with progressively more functional sets of operations. Operations on measurements range from equality and inequality to mathematical functions. These levels of measurement have been formalized by Stanley Smith Stevens in his landmark 1946 paper “On the Theory of Scales of Measurement.”

Nominal data values are used to classify, identify, and categorize data. The value is arbitrary and represents a quality, not a quantity. Values are considered labels with no comparative meaning. The only valid comparisons among nominal data values are equality and inequality. Maps that present nominal data values include soil maps, land-use maps, parcel code maps, and countries of the world. Field types used for nominal data are integer and text.

Ordinal data values determine the rank of an entity versus other entities. As well as equality and inequality comparisons, ordinal data values allow greater-than and less-than comparisons, but no mathematical functions such as addition or multiplication. Maps that display ordinal data values include street maps symbolized by road type and maps of marketing territories shaded by sales ranking. Field types preferred for ordinal data are integer, but in practice, text is used as well.

Interval data values represent a measurement on a calibrated scale, but with an arbitrary zero point, like the Fahrenheit temperature scale. Interval data values support addition and subtraction. Multiplication and division are performed on differences of values. Because the zero point is arbitrary, negative values can be used. Maps that depict interval data values include temperature isoline maps and contour elevation...
maps. Appropriate field types are floating-point values and dates, with values sometimes rounded to short and long integers.

*Ratio data values* represent a measure on a scale with a fixed and meaningful zero point. Length, area, and the Kelvin temperature scale are examples of ratio data. Ratio data directly supports multiplication and divisions and negative values are usually not allowed. Maps that visualize ratio data values include weather maps with barometric pressure and rainfall, and chemical analyte concentration maps. The best field types for ratio data are floating-point values, but sometimes values are rounded to short and long integers.

To validate the accuracy of attribute data collection, each field of a feature class can have a *domain*, which is a numeric range or list of valid values. Fields can also be set with a default value when an object is created.

Features have associated behavior through rules. Rules enforce data integrity within and among datasets. Rules apply to feature shapes, attributes, and relationships. You use attribute rules to apply a domain to a field, relationship rules to constrain valid cardinality in relationships, topology rules for spatial relationships such as touches, contains, and overlaps, and network rules for validating which features can be connected. Attribute domains are rules that describe the legal values of a field type, providing a method for enforcing data integrity. Attribute domains are used to constrain the values allowed in any particular attribute for a table or feature class. A domain is a set or range of acceptable attribute values. Whenever a domain is associated with an attribute field, only the values within that domain are valid for the field.

All geographic objects have some relationship to other objects. Features can be related on-the-fly through joins on keys in tables. When necessary, you can define *relationship classes* among geographic objects in different feature classes and tables, such as the relationship between a house and its owner. Relationship classes can store many-to-many relationships between features in separate feature classes. Relationship classes help ensure data integrity. For example, the deletion or modification of one feature could delete or alter a related feature.

Spatial relationships can be discovered or defined among features. Spatial joins discover associations among features using spatial relationships such as “within,” “.touches,” or “overlap.” Spatial rules can also ensure that features maintain spatial integrity such as “not intersecting” or “not having gaps.” Higher-level structures built on the simple feature information model can also manage spatial relationships among features. For example, a network model can be used to find features that are adjoining, upstream, or downstream. A parcel fabric can find neighboring parcels or parcels within a jurisdictional area.

**Representing connected linear systems with networks**

Networks are simple. They are comprised of two fundamental components, edges and junctions. Examples of edges are streets, transmission lines, pipes, and stream reaches. Examples of junctions are street intersections, fuses, switches, and the confluence of stream reaches. Edges connect at junctions, and the flow from one edge – automobiles, electrons, and water – is transferred to another edge.

To analyze networks, geographers use a branch of mathematics called *graph theory*. Graph theory keenly interests us because it provides rigorous theorems concerning properties of networks and algorithms to solve network-traversing problems. For example, one set of related algorithms for solving a problem in graph theory known as the Chinese Postman Problem provides optimized
solutions for many applications such as postal delivery, street sweeping, bus passenger pickup and drop-off, and garbage collection.

To understand the geographic application of graph theory, contrast it with a surface representation. A surface can be considered an infinite set of points on which you can sample continuous varying phenomena, such as elevation, precipitation, temperature, soil moisture, and spectral reflectance at many wavelengths. But other phenomena are not continuous across a surface. People, manufactured goods, water, energy, communications, and commodities are transported in a constrained way, channelized by streets, railroads, pipes, fiber optic lines, and shipping routes. The set of paths and where they join is a network, a subset of the infinite points on a surface.

Graphs can represent systems that move along natural or constructed linear networks. For example, the interstate highway system can be abstracted to a set of edges representing sections of the interstate highway between interchanges and junctions representing the interchanges at cities. A graph is the set of edges that connect at junctions. Two junctions span an edge. A junction can connect to one or many edges. A graph can be directed or undirected. (A directed graph is also called a digraph.) Directed graphs model systems where each edge has a fixed direction of flow, such as a river network flowing downstream inside hydrologic channels. Undirected graphs model systems with no preferred direction of flow, such as street networks which predominantly allow travel in both directions.

A graph can be planar or nonplanar. Planar graphs connect on a two-dimensional plane at junctions, without any crossing edges. Nonplanar graphs have edges that cross. On a street network, nonplanar graphs can model a bridge over a road. A graph can be cyclic or acyclic. A cycle (or circuit) is a set of connected edges that eventually return to a junction (in the case of directed graphs, a further condition is that the edges line up in flow order to close a cycle). A graph without cycles is called a tree graph (or acyclic graph) in graph theory. Examples are local area communication networks and river systems. A graph with cycles is called a cyclic graph. Application examples are streets and water utilities.

Whether a graph is directed or undirected and cyclic or acyclic are important graph properties when planning the application of GIS network models. These types of graphs are optimally stored with different internal data structures and have separate classes of algorithms for network analysis.

Representing undirected networks with network datasets

A type of network model, the network dataset, represents undirected networks, particularly transportation networks. Network datasets encapsulate decades of GIS modeling experience on optimizing street datasets for network analysis. You can see this in the network dataset’s attribute model representing cost impedances, connectivity rules and groups, hierarchy, elevation fields, one-way restrictions, and turns. This model enables sophisticated analysis such as optimal routing and allocation, even on multimodal systems such as combined street, subway, and pedestrian traffic. A network dataset in a geographic database can have zero or more junction sources (point feature classes), zero or more edge sources (line feature classes), and zero or more turn sources (line feature classes with turn attributes).

Turns are essential for modeling transportation networks. An undirected network allows vehicles to travel freely in any direction. In reality, our street networks have many restrictions on left turns, right turns, and U-turns. Turns
model these restrictions. Turns are also useful for assigning additional travel costs on a maneuver. If making a turn between two major streets is constricted and slow, a turn can be added with that additional travel time.

Network elements are discovered from features when a network dataset is built by finding geometric coincidences of points, line end points, and line vertices. The network elements and connectivity information are stored in a graph, a set of element and index tables inside a network dataset.

Edges are network elements that connect to junctions. Edges are the links over which resources flow. Each edge has exactly two junctions. Junctions connect edges and facilitate navigation. A junction may be connected to one or more edges. Turns record information about a sequence of two or more connected edges. Turns model turn restrictions such as no left turns or turn impedances.

Network elements do not have any geometry but combine to form a nonplanar network graph. Streets generally meet at surface level, so the transportation system is mostly planar. When roadways split elevations, such as at bridges or tunnels, this can be modeled by using nonplanar features for the overlapping lines or planar features with elevation fields marking the elevation levels at both ends of a line. While network datasets accommodate nonplanar networks, much commercial street data is mostly or completely planar.

The idea of connectivity groups is used to model multimodal systems, such as combined street and subway systems. Elements in different connectivity groups cannot connect except at shared junctions such as subway stations. For junctions, a connectivity policy sets whether a point for a junction must be located at an end point or whether it can be at any mid-span vertex along a line. For edges, a connectivity policy determines whether junctions are added when vertices from different lines coincide or whether junctions are added only at line end points.

When creating a network dataset, one can identify special attributes from network sources with information about hierarchy (road classification such as highway, major road, minor road levels), elevations (setting the physical levels of roadways, like stories in a building), one-way or vehicular restrictions (such as no trucks or no pedestrians), and cost of travel time or distance traveled. With variations on implementation, commercial data vendors include these and other attributes on street networks.

Hierarchy in a network of datasets is used to optimize compute performance and to find the likely quickest routes. Sometimes a path found may not be the shortest in distance but the shortest in travel time.

**Locating features on networks with linear referencing systems**

Linear systems such as roads, streams, pipelines, and railroads have many attributes that are defined at discrete locations or sections and these attributes often change with time. In a geographic database, these complex systems may be represented with a simple model called linear referencing. This model is based on routes, which are line features with a defined measurement system. Linear referencing extends line feature classes for modeling events along routes and provides a framework for dynamically managing, displaying, and analyzing all the types of information that can occur at points and sections along routes.

Linear referencing is about locating features on routes. A route can be any linear feature such as a highway, river, pipeline, or GPS track. With linear referencing, features are located on routes by measurements made relative to a starting location.
instead of an x, y location. Measurements most often represent distances, but can also represent other values such as time.

Organizations such as highway departments, government agencies, and pipeline companies manage their asset inventory and incident data along linear networks by measurement systems such as mileposting. Examples are finding an accident at highway milepost 12.4, a navigational buoy at river kilometer 61, or a pumping station at 40 km from the start of a pipeline.

Linear referencing provides two key benefits: it is a natural way to work with linear features and incidents that are recorded using measurements such as distance and time and it allows the modeling of many overlapping attributes on a route without splitting features. Linear referencing is widely used in transportation agencies to manage highways and streets for assessing pavement conditions; maintaining, managing, and valuing assets such as traffic signs and signals, guard rails, toll booths, and loop detectors; organizing bridge management information; and reviewing and coordinating construction projects.

The principal challenge of managing information along route systems is how to handle the many types of attributes that occur along a common set of lines. Along a highway route, a department of transportation manages diverse attribute data such as number of lanes, pavement material, speed limit, and roadway quality. A section of a route that represents a speed limit value rarely coincides with another segment of the same route modeling a certain pavement condition, roadway quality, or number of lanes. Simple line features to represent these attributes on a route must be split whenever an attribute changes. But with the linear referencing model, the underlying line features do not have to be split when an attribute such as pavement quality changes.

Another benefit of linear referencing is that it follows a natural method for location along routes. It is easier for maintenance crews to locate themselves using odometers and signs rather than interpolating coordinates from a map. While GPS receivers now make it easier to find coordinate-based locations in the field, these organizations continue to use linear referencing because field personnel find it natural to use odometers and signage.

Routes are a special type of line features that have measure values, also known as m-values. M-values are used to designate relative positions along routes for linear referencing. In addition to (x, y) or (x, y, z) coordinate values that all features have, measure values (called m-values) are stored at each vertex of a route feature. Measure values can represent any unit of measurement you choose, including miles, kilometers, meters, and feet for distances, or hours, minutes, and seconds for time intervals. Measure values are independent of the coordinate system of a feature class, however, it is not uncommon for the measure unit to be the same units as the line feature class’s coordinate system.

Point events, such as accident locations, and line events, such as pavement conditions along a section of highway, can be stored in event tables with a unique route identifier, measure values, and other attributes of events.

Representing imagery and continuous phenomena with rasters

Rasters are a sampling of one or many attributes of continuous phenomena on a rectangular array of equally sized cells. These cells, when taken as a whole, capture imagery, thematic, surface, or picture data. Raster data can represent everything from qualities of a land surface, such as rainfall, elevation, temperature, or vegetation, to satellite images, scanned maps, and photographs. The raster data model is very simple but supports
the modeling of a wide variety of geographic phenomena.

While the structure of raster data is very simple, it is exceptionally useful for a wide range of applications. Within a GIS, the uses of raster data fall in four main categories: rasters as base maps to provide a photographic background display, rasters as surface maps to enable spatial analysis and relief mapping, rasters as thematic maps to present analytic results through derived classifications, and rasters as feature attributes to link a place with a photograph, document, or drawing.

The most common use of raster data in a GIS is as a background display for other map layers. Images drawn below other layers provide the map user with confidence that map layers are spatially aligned and represent real objects. Three main sources for raster imagery are satellite imagery, orthophotos from aerial photography, and scanned maps.

Imagery, such as aerial photography or some satellite imagery, provides an easily recognizable view of the Earth's surface. People trust imagery because it is an actual view of the Earth. Much of the information about our environment – such as vegetative and soil type, distribution of pollutants, extent of urban development – is best detected using imagery. A key advantage of imagery is it can be the most current source of geographic data in a GIS.

Many satellites and aerial sensors are constantly collecting imagery over parts of the Earth every day. Some types of imagery, such as weather, are available in real time. Imagery can also provide three-dimensional information. Stereo pairs of images provide topographic information and height information for building and other features.

A raster can comprise thematic data derived from analyzing other data. A common analysis application is classifying a satellite image by land-cover categories. This activity groups the values of multispectral data into classes (such as vegetation type) and assigns a categorical value. Thematic maps can also result from combining data from various sources, such as vector, raster, and terrain.

Rasters model the surface of the Earth with cell values for elevation. These rasters are known as digital elevation models (DEM), digital terrain models (DTM), and digital surface models (DSM). Elevations are just one of the surfaces that can be modeled with rasters. Other continuous surfaces modeled with rasters include rainfall, temperature, contaminant level, groundwater level, noise, gravity, and population density. Continuous surfaces can be captured through sensors such as radiometers, derived from geostatistical analysis, or calculated from a geoprocessing model.

Many continuous surfaces change over time. Contaminant plumes spread from a source, precipitation falls with the passing of a storm, and temperature is a field continuously varying in space and time. Rasters with time-variant surface values can be stored as separate raster datasets for each point in time.

The cell size determines the resolution of raster data. Features identifiable in a raster cannot be located more precisely than the resolution of the raster. The size chosen for a grid cell of a study area depends upon the data resolution required for the most detailed analysis. The cell must be small enough to capture the required detail, but large enough so that computer storage and analysis can be performed efficiently. Cell resolution also determines the largest map scale at which this data can be properly displayed.

The value associated with a cell defines the class, group, category, or measure at the cell position. Cell values are numbers, either integer or floating point. Cell values can represent a measured or derived quality of a location, such as reflectance, density, precipitation, or elevation.
These values are either integer for reflectance or floating point for many other measurements.

Cell values can also represent categorical values such as land classification, which may be calculated from a geoprocessing model with many input geographic datasets. These values are always integer numbers. Cells can also have a no data value to represent the absence of data. While each cell has a single value and a color has three components (red, green, and blue values), color is associated with a raster most commonly by using raster bands for RGB composite display, or in the case of thematic maps, colormaps that associate an RGB color with a thematic index value.

Representing surfaces with terrain datasets and digital elevation models

The surface of the Earth is modeled in a geographic database with terrain datasets. Terrain datasets can handle many millions and even billions of points with x, y, z values and produce realistic scenes of a landscape. Terrain datasets are well suited for careful shaping of features such as roads and rivers. Terrain datasets can generate elevation rasters to support surface analysis such as stream and watershed delineation and improved cartographic display such as elevation gradient and hillshading combined, as shown on this map.

Terrain datasets can be in the form of triangulated irregular networks (TINs). The TIN representation models a surface from a set of points from which triangles are formed, or triangulated. Triangles are made from three points that occur at irregular locations. Each triangle stores topological information about its neighboring triangles, thus forming a network.

The edges of TINs form contiguous, nonoverlapping triangular facets and can be used to capture the position of linear features that play an important role in a surface, such as ridgelines or stream courses. Because nodes can be placed irregularly over a surface, TINs can be sampled more densely in areas where a surface is highly variable or where more detail is desired and a lower sample density in areas that are less variable.

A terrain dataset may be a multiresolution, TIN-based surface built from measurements stored as features in a geographic database. Terrain datasets are made from data sources such as stereo-captured photogrammetric features and mass point collections of 3-D data such as LiDAR, sonar, and bathymetry. Terrain datasets in the geographic database are designed to handle voluminous sets of point data from many sources. Terrain datasets can also integrate other 3-D data, such as lake boundaries and breaklines for edges of roads that are collected with GPS receivers and other methods.

Feature integrity with topology rules

Topology rules assure accurate topological relationships among features. Topology rules can be stored in a geographic database as one or more relationships that define how the features in one or more feature classes relate to each other, including shared geometry, containment, and so on. Topology rules have long been a key GIS requirement for data management and integrity. In general, a topological data model manages spatial relationships by representing spatial objects (point, line, and area features) as an underlying graph of topological primitives – nodes, faces, and edges. These primitives, together with their relationships to one another and to the features whose boundaries they represent, are defined by representing the feature geometries in a planar graph of topological elements.

Topology rules are used to ensure data quality of the spatial relationships and to aid in data
Topology rules are also used for analyzing spatial relationships in many situations, such as dissolving the boundaries between adjacent polygons with the same attribute values or traversing a network of the elements in a topology graph. Topology can also be used to model how the geometry from a number of feature classes can be integrated. Some refer to this as vertical integration of feature classes. Features can share geometry within a topological association, such that area features can share boundaries (polygon topology) and line features can share end points (edge-node topology).

Managing cadastres with parcel fabrics

A parcel fabric stores a continuous surface of connected parcels or parcel network. Parcels are made up of polygon features, line features, and point features. Parcel polygons are defined by lines, which store bearing and distance measurements. These measurements should ideally or eventually match recorded bearings and distances from the record of survey or plan. Each parcel has its own set of lines, which means there are two lines that define the common boundary between adjacent parcels.

Parcel lines have end points, which are common between adjacent parcels. Common points between parcels establish connectivity and maintain internal topological integrity in the network. Parcel points store x, y, z coordinates which locate the parcel on the ground. Line points are assigned to parcel points that lie on an adjacent parcel boundary but do not split the boundary. Line points constrain the parcel point to lie on its adjacent parcel line. Control points are assigned to parcel points when running a fabric adjustment. Control points constrain and pin down the parcel point in the adjustment.

Connection lines can be used to link isolated blocks of parcels to each other or link parcel corners to control points. Because each and every parcel is either linked or connected, a seamless network of connected parcel boundaries, or parcel fabric, is formed.

Finding places with locators

Locating a place by address is the most common geographic task that people engage in. Every organization has databases with tables containing addresses for customers and facilities. With a GIS, an address locator is designed to locate addresses on a map in a geographic database. An address locator contains reference data and methods for interpreting addresses. The reference data is most often a street network containing street names and address ranges for each street segment. The methods in an address locator encapsulate the addressing rules for a specific area or country. Geocoding is the process of transforming a description of a location — such as a pair of coordinates, an address, or a name of a place — to a location on the Earth’s surface. The resulting locations are output as geographic features with attributes, which can be used for mapping or spatial analysis.

There are various kinds of locations through geocoding. The types of locations include points of interest or names from a gazetteer, like mountains, bridges, and stores; coordinates based on latitude and longitude or other reference systems, such as the Military Grid Reference System (MGRS) or the US National Grid system; and addresses, which can come in a variety of styles and formats, including street intersections, house numbers with street names, and postal codes.

Addresses are the most common form of geographic information and are an essential
data source for many geographic information systems. Practically every residential and commercial location has an address. Ideally, each address is unambiguously associated with a unique location, but in practice, addresses have several characteristics that present challenges for automating their placement on a map. Some of the issues with finding locations for addresses are errors and misspellings in data entry, variations in address elements (“Avenue,” “Ave,” and “Av” may be found in a street address), variations in how houses are numbered (such as the rule of odd-numbered addresses on one side of a road and even-numbered addresses on the other side), changes in street names over time, and primary street names versus alternate street names (such as “Route 66” or “Main Street”). In a GIS, all of the geocoding rules and geographic information necessary to find an address is stored in the geographic database as an address locator. An address locator encapsulates the rules for interpreting addresses (such as “odd house numbers are on the right side of a street”), lists of standard street components and their abbreviations (such as “Avenue,” “Ave,” “Boulevard,” “Blvd,” and so on), plus the reference map data needed to find addresses.

Address locators are designed to handle the many possible variations of a valid address. If incorrect or insufficient information is provided in an address, the suspect address matches can be manually inspected and used to refine the address locator rules or improve the reference data.

While address formats vary considerably around the world, every address consists of several address elements, presented in a particular address format recognized within that region. Address locators are built to recognize the address formats and types of address elements that are particular to a region or country.

**SEE ALSO:** Data model, object-oriented; Data structure, raster; Data structure, vector; Spatial database; Spatial feature classes; Topological relations

**Further reading**


Representation: indoor spaces

Imad Afyouni
Wadi Makkah, Umm Al Qura University, Saudi Arabia

Cyril Ray
Christophe Claramunt
Naval Academy Research Institute, France

The range of geographical information system applications is progressively evolving from large outdoor environments to small-scale indoor environments. A new generation of opportunities in indoor environments has recently emerged. These can take advantage of low-level sensor data at the local scale, as well as events and activities that can be inferred from mobile phones or sensor-based data. At the root of the development of indoor space applications, one of the preliminary research questions is closely related to the availability of an appropriate spatial data model and data management facilities. Such spatial data models generally share several levels of abstraction as well as discrete and continuous data layouts. At the normative level, the development of spatial data standards surely will help and favor the development of exploratory applications. Next, and although indoor spatial data models offer unprecedented opportunities, several research challenges are still to be addressed. Indoor spaces are by their nature complex systems, as the phenomena that happen range from moving objects acting and behaving within these spaces (humans, artificial or virtual objects), to the distribution of intelligent sensors, and the delivery of collaborative and context-oriented services. Overall, the integration, storage, management, search, visualization, and analysis functionality in space and time can be considered as complex data integration and manipulation issues. Moreover, designing a comprehensive spatial data model for indoor environments requires the integration and combination of data from diverse sources in addition to the user’s context. The role of contextual dimensions is primal and also offers additional opportunities, as sensor-based data and collective social data provide additional knowledge-based opportunities.

While presenting a survey of current spatial data models developed for indoor spaces, this entry also discusses a number of research challenges related to the integration of context-aware data models, continuous query processing, as well as semantic location-based services. The potential of indoor spaces is illustrated by a concept of an extended context model, as well as several open challenges for context-aware queries closely related to navigation services. A concept of semantically annotating users’ trajectories is also presented. Several application perspectives associated with the future of indoor-based systems from academic and industrial points of view are highlighted.

The remainder of this entry is organized as follows: current indoor spatial data models are briefly reviewed; the data management and query processing techniques closely related to the manipulation of indoor moving objects are addressed; application perspectives are discussed.
Indoor spatial data models

Indoor spatial data models are important foundations for the development of human- or robot-based navigation and the delivery of context-aware services. Such models should help to locate static and moving objects in indoor spaces, and support users interactions and the different levels of behavior that emerge. Indeed, a better understanding of location information and the relationships that might exist among spatial entities, either acting or located in the environment, should be taken into account. As shown in Figure 1, many components that contribute to the design of a context-aware system should be reflected by the spatial model underneath. Therefore, the integration of an indoor spatial model into a context-aware system implies considering the environment as a dynamic system that should represent:

- the features that populate the environment: a feature can refer to either a person (i.e., mobile users and other social entities/human beings that are located in the vicinity and are of interest) or an object of interest. (An object of interest (OOI) may be either communicating or not, mobile or static, physical or virtual, attractive or repellent. Examples of such OOIs are sensors, exits, etc.).
- their spatial properties, that is, locations of the objects of interest and the spatial relationships that relate them.
- the actions that emerge from them (i.e., physical interactions and communications).

Moreover, specific properties of the spatial data model retained for the representation of an indoor space influence the manipulation, visualization, and computational capabilities of the system.

Two main approaches have been so far considered for indoor space models: symbolic and geometric spatial models (Afyouni, Ray, and Claramunt 2012a). Geometric-based approaches model an indoor space using a series of geometrical primitives, that is, points, lines, areas, and even volumes for 3-D representations. Symbolic-based approaches retain a sort of qualitative representation rather based on some points of interest (e.g., room, corridor, etc.). While geometric models can efficiently integrate metric properties to provide accurate location and distance information, symbolic models maintain an abstract view of space where the structural properties of an indoor space are more appropriately represented. Over the past few years several indoor spatial data models have been developed and applied to different domains (e.g., robotics, GIS, ubiquitous computing) with often a combination of geometric and symbolic approaches (Walton and Worboys 2012). Such approaches favor the application of indoor spaces to a large spectrum of applications, as suggested by the spatial semantic hierarchy (SSH) as initially introduced in Kuipers...
This provides a support for advanced functionalities such as human reasoning and robot-centered activities, and even sensor-based interactions within the environment.

A brief summary of different modeling approaches is given here, along with a discussion of their use from an application perspective.

**Geometric-based approaches**

Amongst geometric approaches, grid-based models decompose space into regular cells and can efficiently integrate metric properties, thus allowing precise locations, direction information, and distances. Free-space tessellation and Voronoi-based diagrams aim at providing an adaptive decomposition of space that is suitable to exactly represent the complexity of the environment being studied. Free-space tessellations are less suitable for localization but they are more compact (Mekni 2010). Boundary-based models represent the obstacles’ boundaries of an indoor space with sequences of primitive geometries such as points, lines, curves, and so on. Boundary-based models are less suitable for navigational services but provide accurate location data. Overall, geometric models require an integration of semantic annotations to achieve a higher degree of location- and context-awareness.

**Symbolic approaches**

Symbolic approaches have frequently attempted to model indoor environments using topological-based structures, graphs by capturing the connectivity and reachability between spatial units, and hierarchies (Jensen, Lu, and Yang 2009). Symbolic models are generally less accurate but context-awareness is easier to achieve as it favors human-recognizable descriptions. Thanks to their hierarchical structure, set-based models achieve a good level of efficiency and flexibility, but lack topological relationships such as connectedness (Becker and Durr 2005). Graph-based models are widely used in applications at a coarse-grained level of abstraction thanks to their richness and variety (Jensen, Lu, and Yang 2009). The major shortcoming of symbolic models is the lack of geometric details on entities and places represented in space.

**Application perspectives**

Symbolic-based approaches are often preferred, from an application perspective, over conventional geometric-based approaches and have been recently used in many application scenarios because they can capture the semantics of entities and places represented in an indoor space. In particular, graph-based and semantic models constitute the most common approaches used, so far, in many application areas ranging from emergency management and safety control in microscale environments to indoor context-aware navigation services, and especially those adapted to users with special needs. Applications that aim at providing intelligent emergency responses mainly employ simple place graphs, which capture topological relations between structural entities, because they are more concerned with network-based models that allow nearest or optimal exits to be discovered. In contrast, applications that support contextual elements, such as user preferences and capabilities, tend to favor semantically-enriched data models, either by designing an ontology-based model or by employing a hybrid model that combines a graph-based with a semantic model of space.
Towards context-aware indoor spatial data models

On top of an indoor spatial data model, several service-oriented and efficiency-related requirements can be developed (Afyouni, Ray, and Claramunt 2012a). The delivery of appropriate services in an indoor space implies that not only are static and mobile objects represented in their environment but that the semantics of the environment and context-based properties are also to taken into account. This is a mandatory requirement to deliver advanced services and navigation facilities to users.

Different dimensions still have to be considered when integrating context-awareness with an indoor spatial data model. There is clearly no generic answer for the choice and the modeling of contextual dimensions. While some dimensions are common to many context-aware services, others might be application-dependent. Such contextual dimensions can be gathered through different ways, using physical as well as virtual sensors. Physical sensors embedded in mobile devices provide basic raw data, and other virtual sensors (agenda, profile, social network activity, etc.) can also help an indoor system to track the user context.

Context-based spatial data model

Most of the principles presented in this entry have inspired the development of a hierarchical and context-dependent multigranular indoor spatial data model that provides a flexible representation of an indoor space (Afyouni et al. 2012b). The modeling approach developed integrates different levels of granularity and considers several contextual dimensions that emerge from space and time as well as user profiles. It supports continuous- and discrete-based location-dependent queries, and has been implemented by a series of algorithms that support navigation and range queries (Afyouni et al. 2013a). Context dimensions considered can be classified as follows (Figure 2).

- The environmental context comprising the spatiotemporal dimensions, as well as other environmental parameters (light, temperature, etc.). The spatial dimension includes the location, direction, and orientation in an indoor space. A user trajectory is another important element of the data representation, as well as another context dimension to consider. A trajectory and more generally movement patterns can be represented by different modalities, for example, a series of locations and/or directions as well as qualitative representations. The temporal dimension depicts the time stamp for a given event or activity within an absolute time reference or using high-level time and date indicators (e.g., morning). This temporal dimension is incorporated in most of the monitored and scheduled events in a context-aware system.

- The user-centered context includes user activities and his/her personal context. The user activity is an important context dimension that reflects the task or process performed by the user when acting in an indoor space. Taking into account such activities provides a step forward to develop truly ambient intelligent systems. For instance, a user activity recognition technique based on accelerometer data can be used. Such a system can identify different activities, such as running, walking, going upstairs, going downstairs, sitting, standing, in an elevator, and so on, and considers them while re-evaluating the answer to a given query.
- The personal context includes the user profile that is made of static as well as dynamic properties (gender, age, interests, physical capabilities, access privileges, etc.). This information can be acquired by different means, either explicitly or implicitly. For instance, an overview of the sensors embedded in mobile phones and their potential as a computing platform to acquire user parameters is reviewed in Lane et al. (2010). Dynamic properties can be also considered on the fly to take into account real-time user’s behaviors, which, in turn, can be broadcast to other users when necessary and possible. In fact, user-derived contexts are of valuable importance for the development of recommender systems in indoor spaces.

Ideally, a contextual system can infer some predominant user interests and provide him/her with relevant recommendations. But users’ activities and interactions between them are still to be considered; this leads to a research challenge, as tracking these behaviors is far from straightforward. Interaction between users stresses the need for a better consideration of effective models of collaborative usage. A collaborative usage model should classify users into groups or communities, within which they share a common characteristic or interest. A user can belong to a community either explicitly by a subscription-based mechanism or implicitly if he/she has a common characteristic or property. On the one hand, subscription-based communities encompass users that share a given interest or are involved in a similar practice (i.e., group of people working together and constantly conducting solutions in a given shared field). On the other hand, implicitly generated communities depend either on the current location of users or on their trajectories (i.e., if they have similar spatiotemporal patterns) or on their behaviors in a similar context. This collaborative usage model derives similar behaviors and interests, so that the system can act proactively by notifying users of certain events or adapting certain query answers. Push-based services can thereafter be designed based on such communities.

One of the peculiarities of indoor spatial data models relies in its large scale, closely related to real-time activities as humans acting in such environments behave in a fast mode. The availability of data at the right time is then a crucial issue. Based on data gathered from fixed and/or mobile sensors and on user-generated content (i.e., information shared by users to the system or within communities to notify certain events), an indoor system should be able to manage events in real-time and even filter the ones that are relevant for a given user. Some examples of real-time services to deliver are that people who share a common interest or common context should be notified. Depending on the relevance of such events, an indoor-based system might either let the user decide whether to integrate this event into his/her relevant queries, or simply ignore it. Some types of events should be directly handled by the system itself (in case of emergency or other important events). Overall, these contextual dimensions represent the user-centric context and should be taken into account by an indoor system in real time in order to completely incorporate context-awareness into intelligent navigation services.

Towards a unified indoor/outdoor data model

Another crucial research challenge is to investigate how a contextually-enriched semantic location and semantic trajectory of moving objects can improve the quality and adaptability
Figure 2  Taxonomy of contextual dimensions.
of navigation services and LBSs in mixed outdoor/indoor environments. As already suggested in Worboys (2011), designing a generic and unified data model for a mixed indoor/outdoor environment is a mandatory development for the delivery of intelligent-based context-aware systems. Bringing these two spaces together into one generic model provides users with a consistent and seamless navigation between built and natural environments. Different issues need to be addressed at the conceptual, functional, technical, and application levels. Particularly, differences and similarities between both spaces should be clearly identified. For instance, in contrast with outdoor environments, a hierarchical structure is typical in an indoor environment. Furthermore, small-scale indoor spaces encompass some specific properties and different levels of granularity, when compared with large-scale open spaces. Also, interactions among humans and/or other objects in a small-scale space are much more frequent than in a large-scale one, which is usually beyond the range of humans’ physical interactions. Although research on unified indoor/outdoor models is still at an early stage, some researchers have already started to address this challenge (Xu and Giiting 2011). A generic data model for representing the movement of a moving object (i.e., indoor and outdoor movements), where the roads, streets, and rooms are considered as constituting entities, has been developed (Xu and Giiting 2011). An indoor graph has been designed based on specific data types to support indoor trip planning at the room level, while additional data types have been provided for representing moving points in a public transportation network (i.e., buses, trains, and underground trains). However, this model does not represent objects’ movements at a fine-grained level and does not support context-awareness.

Data management techniques for indoor moving objects

Many mobile applications need to incorporate a mechanism for the continuous processing of location-dependent queries over moving objects on top of indoor space models. Examples of such services include continuous crowd monitoring within a given area, location-based alerts (e.g., continuously send E-coupons to all customers within 200 meters of my store), crowd notification in an emergency situation, and location-based friend finder (e.g., 1 – “let me know if I am near a restaurant while any of my friends are there;” and 2 – “if I continue moving towards this direction, which will be my closest restaurants in the next 10 minutes?”).

A continuous query remains active over a period of time and has to be continuously evaluated during that period as the objects of interest are moving. Efficient processing of such queries is a complex task in location-based services due to its expensive consumption of memory and computational resources. A straightforward approach to deal with this kind of query would be to periodically re-compute the best answer from scratch for all queries upon arrival of every delay update. However, such a naive approach performs poorly in highly dynamic environments, where the locations of both the query point and the target objects change over time.

Different architectures, indexing strategies, and update methods have been proposed to deal with location-dependent queries over moving objects (Prabhakar et al. 2002). These approaches can be mainly classified into two categories: (i) approaches for continuous query processing in moving object databases; (ii) approaches for continuous processing of spatiotemporal data streams. It should be noted that most of those strategies have been developed for outdoor
environments. A few works regarding architectures for mobile data management in indoor environments are presented in Jeng (2009) and Afyouni et al. (2013a). There is also a lack of approaches that incorporate user preferences and other contextual dimensions in continuous location-dependent query processing. Consequently, there is a growing need for approaches that provide adaptive and incremental query processing techniques to enhance efficiency of location-dependent queries over moving objects (Mokbel et al. 2005). On the one hand, adaptive query processing addresses the problems of unpredictable costs and dynamic data by using runtime feedback in order to tune query execution in a way that provides appropriate answers with better response time or memory utilization. On the other hand, an incremental execution paradigm deals with continuous query processing by reusing information from previous searches to speed up current searches. Therefore, new answers for series of similar search problems can be computed faster than by re-evaluating each query independently from scratch (Sun, Yeoh, and Koenig 2009). Both adaptive and incremental execution paradigms are required in order to achieve sufficiently flexible and more efficient continuous query processing.

A few research studies have addressed the specific requirements necessary to manage and query indoor moving objects (Jensen, Lu, and Yang 2009). For instance, a framework for managing and querying indoor moving objects has been proposed in Xie, Lu, and Pedersen (2013). This approach suggests one manage indoor geometries (i.e., spatial entities), indoor topologies (i.e., relationships among spatial entities and between space and moving objects), as

Figure 3  Context-aware navigation system architecture.
well as indoor moving objects (modeled with uncertainty regions around), with a composite index scheme in order to support efficient distance-aware queries. However, only static query points are considered and this approach did not discuss how to incrementally re-evaluate those queries.

Over the past few years, several research studies have also discussed the integration of some contextual dimensions in query processing (Mokbel and Levandoski 2009). Users’ preferences have traditionally been exploited in query personalization to better anticipate their needs and customize their experience. In order to deliver personalized query answering in mobile environments, context- and preference-aware query processing techniques are increasingly required. A query processing strategy should provide sufficient flexibility to location-based services. However, there is a lack of approaches addressing context-awareness in continuous location-dependent query processing.

Towards context-aware query processing

A continuous query processing framework should take several contextual dimensions into account. A comprehensive approach should be designed to support the extended context model previously described. Such a system should deal with a highly dynamic environment in which objects are moving continuously and their environment is dynamically changing. One important direction to consider in future work is the complete integration of the context model into the query processing engine. A generic context-aware system architecture is presented in Figure 3. A spatiotemporal continuous query execution framework should first be used, such as the one proposed in (Afyouni et al. 2013a). This framework manages indoor moving objects and suggests continuous path and range queries over a hierarchical indoor spatial data model. This framework can then be linked to a user context manager that should take into account live context from mobile sensors, as well as from social and personal data, and also historical context in order to help enrich the user navigation experience.

This implies the development of a “cost model for evaluating context-aware queries” that provides not only a quantitative value as a result of a cost function but also a semantic indicator that evaluates the quality, usefulness, and uncertainty of query results. The main challenge in developing a performing cost model is to take into account the different parameters of the context while associating different quantitative scores and semantic indications to each parameter according to the corresponding user. This means that, for instance, a user with special needs has different constraints with different impedances than a user with complete physical capabilities. Taking all these factors into account when designing a cost model for context-aware queries is challenging. The main goal is to provide a context-aware query processor that avoids modifying the database engine with the addition of each new context parameter. Instead, an extensible query engine that is general enough to support any kind of context should be designed.

When considering context-aware query representation, a taxonomy of spatiotemporal operators that can be combined with conventional query operators has been proposed in a related work (Afyouni et al. 2013b). Several spatiotemporal operators that can execute standard as well as navigation-related queries have been introduced. To fully integrate the proposed context model, additional “preference-and context-aware query operators” are still to be developed in order to take into account explicit preferences and other users’ constraints and
Representation: Indoor Spaces

wishes. At the performance level, this will help to compute cost functions for the continuous processing of location-dependent queries.

Semantic annotation of heterogeneous trajectories

As another possible direction for further research, semantic and location-based services that combine users’ trajectories with contextual data can favor the development of advanced intelligent services. Not only will current movements be taken into account but so, too, will other user trajectories as well as additional semantic information that will help to deliver user-oriented services. In such a collaborative environment, active objects in the environment, which can be classified as repellent, attractive or neutral, as well as events will also be part of the system environment. Users should be given the opportunity to annotate these points of interest (POIs) and then contribute to a sort of volunteer indoor system (Mata, Claramunt, and Juarez 2011). Such a system should classify those POIs as either attractive, repellent or neutral regarding other users depending on their context. Users might also share some common interests and might follow some similar patterns from either semantic (e.g., shop recommendations, etc.) or spatiotemporal points of view. Overall, semantic annotations of locations, trajectories, and objects of interest in the environment together with typical movement patterns open new research directions for intelligent and collaborative location-based services in indoor spaces.

These research directions also require the development of benchmarks and evaluation methodologies for the assessment of scalable and distributed techniques for real-time processing of contextual data in mobile and indoor environments. Furthermore, the search for a unified indoor/outdoor design (discussed in the section “Towards context-aware indoor spatial data models”) would require the study and development of efficient algorithms and systems that will support scalable data management and information integration for spatial, temporal, and semantic big data at the city scale. A new paradigm for such large scale management is emerging and new techniques to store, share, query, and analyze these big datasets is surely required.

Application perspectives

Application perspectives are very large, from the development of interactive systems for built environments, to additional professional- and user-oriented services. The numerous application types range from mobile location-aware services to context-aware recommender systems. New classes of applications, which can take advantage of both low-level sensor data and high-level events, context, and activities inferred from mobile phone sensor data, are being explored not only in academic and industrial research laboratories but also within large corporations such as Google and Apple as well as start-up companies.

Not only are the application perspectives promising, but also in terms of business as the range of possibilities will surely open many opportunities. In a report published by ABI Research (http://www.abiresearch.com) on “indoor location smartphone applications,” it has been noted that the indoor location market broke one billion application downloads by 2016. It has been also emphasized that the future adoption of a variety of indoor location technologies will be considered across a range of different application categories, such as retail, navigation, environmental monitoring, location-aware social networking, health
and wellbeing, personal tracking, while also enhancing services such as advertising, ambient intelligence, augmented reality, and local search.

Conclusions

This entry has introduced recent advances and research challenges closely related to the development of indoor space applications and context-aware indoor navigation systems. Current spatial data models, data management, and query processing techniques developed for indoor spaces have been surveyed. It has been shown that the representation of indoor spaces implies different modeling abstractions, from static to dynamic features. A series of recommendations has been given for designing and developing indoor space applications. Amongst many research issues still opened, current challenges on modeling, querying, and indexing moving objects, as well as location-dependent queries have been discussed. In particular, this entry has shown the close connection between indoor spaces and the contextual dimension. It is believed that a close consideration of indoor spaces and context-aware systems should open a series of application avenues worth exploring by the GIScience research community.

SEE ALSO: Data model, moving objects; Geographic information system; Location-based services; Spatial database

References


**REPRESENTATION: INDOOR SPACES**


**Further reading**

The world is a dynamic place. Processes affecting our natural and human environments are highly complex and interrelated, operating at multiple temporal and spatial scales. Gaining understanding of these processes, and thus also of past and potential change, requires study of how properties of entities and their locations vary through time and how entities interact in cause-and-effect relationships.

It is only by the inclusion of the temporal component in the observed data or model that change can be represented, cause and effect relationships derived from the world model, and ultimately an understanding arrived at of the true nature and structure of the process involved. Change is normally described as an event or collection of events – something of significance that happens. For the purpose of spatiotemporal data representation, this is most frequently conceptualized as a change in state of one or more locations, entities, or both. Change relating to entities or locations can be sudden or gradual. Depending on the temporal resolution, events can be represented as having duration – extending over a temporal interval – or as having only a single temporal location. This is analogous to spatial representations where, depending on the resolution of the representation, an entity or spatial attribute can be represented as occurring either at a single point location or extending over an area. Change can also be a continual and gradual process, such as the income level associated with a particular county or groundwater level in a particular aquifer. For such instances of gradual and continual change, a change event is recorded at the time when the amount of accumulated change since the last recorded change is considered to be significant or by some other domain-specific rule. Events relating to entities have additional considerations. Entities can appear, move, expand, shrink, change shape, split, merge with another entity, and potentially disappear.

The study of temporal, as well as spatial, representation has a long history dating back at least to the ancient Greeks. Within geography, an early example of the explicit inclusion of time in the representation and analysis of space–time dynamics is Clark’s work on changing agricultural patterns in Nova Scotia (1959, 1962), Berry proposed a “geographic matrix” (1964) in which he integrated the temporal dimension within a general framework for classifying the multiple approaches for geographic analysis. Hägerstrand’s “time geography” (1967) and his well-known models of diffusion and spatial processes emphasized the temporal dimension. Going forward, other geographers, including Getis and Boots (1978) as well as Bennet (1979), incorporated time in their models.

Temporal and space–time representation from philosophical, cognitive, and formal perspectives has been a research focus in GIScience since the late 1980s (Armstrong 1988; Langran and Chrisman 1988). Interest in temporal, as well as spatial, analysis in geographic information science was spurred by the availability of remotely sensed satellite data as well as the exponential accumulation of other geographic observations over successive points in time. These data, along with advancing computer hardware and software...
Representational considerations relating to intrinsic properties of time versus space

Everything everywhere is inexorably progressing forward in time. Nothing can stop traveling through time in the experiential world. This is only possible in the sense of a historical retrospective. Because of this, we also usually think of time as being unidirectional, although we can usually go backwards in space. Things can also move through space at varying velocities, whereas everything progresses through time at the same rate (the theory of relativity aside, since it does not apply to the perceived everyday world). This can have obvious implications with respect to the varying measurement scales used.

From the standpoint of data collection and storage, we also normally look at space as being finite. We speak in terms of “the study area,” for example. Time, on the other hand, is open-ended and the temporal scope of observational data is continually expanding if one attempts to represent the ever-new “present” as well as past reality. One of the great potentials in the reality of spatiotemporal representations is the ability to search backward from an event or world state to its origins, thus using this technique as a tool to develop models of observed world phenomena.

For purposes of measurement, both space and time are normally broken down into discrete units. Space and time can both be divided into units of uniform or variable length, although the temporal units are necessarily different from those for space. For example, feet or meters cannot normally be used to measure temporal duration. Variable intervals of time normally separate events. Besides the events, the intervening intervals also have significance. Length of time, or duration, is divided into units of seconds, minutes, and days, moons, reigns, or other units suitable for the given data.

The segmentation of time into discrete units of measurement brings up the question of how small these units should be, particularly in measuring evolutionary change. This is directly analogous to spatial resolution and the modifiable areal unit problem (MAUP). And as is the case for...
spatial resolution, temporal resolution (or hierarchy of resolutions) must always be related to the scale of the phenomenon represented and to the intended problems or questions being addressed.

Extensions of traditional entity-based and location-based representations

Initially, representation of dynamics in observational data in GIS was accomplished by extending the familiar location-based and entity-based representations already used in GIS. This solution was pragmatic in that it employed familiar approaches, and it also fits with the dual philosophical views of space and time. Location-based representations, usually in grid or raster form, correspond to the Newtonian perspective of space. In this view, space is the container of objects and exists independent of those objects. In this view, time also exists independent of objects. Space and time are continuous; they are the backdrop upon which the dynamics of physical objects can be measured. In contrast, entity-based representations, most commonly in vector form, correspond to the Aristotelian view of space, where place is a property of objects. In this view, time is also a property of objects. Space and time are a system of relations between material objects. As such they are also viewed as discrete in nature.

The first space–time representation used in GIS is known as the snapshot approach. This representation utilizes sequential images, or “sequent snapshots,” of values for a given variable over a given area for a known point in time. These snapshots are conceptually equivalent to slices extracted from Hägerstrand’s continuous space–time cube. Snapshots also do not need to be evenly spaced in time.

Sequent snapshots are easy to implement utilizing the standard raster and vector data models available in nontemporal GIS by tricking the software. Instead of representing a set of thematic layers (elevation, vegetation, soil type, etc.) (Peuquet 2006), each spatially registered “layer” represents the state at a given moment in time for one of these themes. This approach is conceptually straightforward, and the state for any recorded spatiotemporal location can be easily retrieved. This approach can also utilize the standard map operations available in GIS for static data. Temporal change is easily calculated as the difference between two snapshots.

Recognizing the importance of inferring spatiotemporal change when the data are captured in snapshot form, as in satellite and other geosensor data, Worboys and Duckham (2006) derived a set of local transition rules used to constrain temporal inference for continually evolving regions and to detect higher-level events. McIntosh and Yuan (2005) provided a dynamic time warping (DTW) method to assess similarity of events and derive common characteristics of higher-level events using the example of precipitation data to derive similarity of storms.

There are, however, inherent drawbacks to the sequent snapshot approach, as detailed by Chrisman (1998). Since the actual changes that occurred between recorded snapshots are not explicitly stored and must be interpolated, it is possible that a critical change occurred between the two times. In this case, the stored observational data provide a misleading portrayal of the process or processes taking place. Another drawback is that values for the entire area (the state of all locations or entities at the time represented by the snapshot) are recorded, whether or not a particular location or entity had changed. This redundancy greatly inflates the total volume of stored data.

An alternative approach, to overcome the redundancy problem of the snapshot approach, was introduced by Langran (1992). In this
“base-state-with-amendments” representation, only the changes known as specific times are recorded. A third conceptual scheme is the space–time composite, originally described in Langran and Chrisman (1988). This can be thought of as the result of stacking all successive changes on top of one another as virtual transparencies, flattening the space–time cube into a single, cumulative depiction.

Implementing the base-state-with-amendments can be accomplished in either grid or vector form, as is the case with the snapshot approach. A grid-based implementation records a complete snapshot as the first in the sequence, and then grids representing subsequent times only record changes, noting the new values in the appropriate coordinate (row, column) location. A grid version of the space–time composite scheme was mentioned by Langran (1992) and also documented for use in electronic circuitry design analysis (Fujimoto 1990). Each location in the grid contains a list of timestamped changes. Each change list has a variable length, with each successive change added to the top of the list for that location. The “present” world state for the entire area is easily retrieved by referencing the top of the list for all locations.

The vector implementation is similar, only it is the changes to entities instead of locations that are stored in each successive time “layer.” The base state consists of all objects, or entities, as they are configured at the starting time (i.e., the original state) as points, lines, and polygons. Changes to point entities are noted as single $xy$ coordinates, and to line and polygon entities as strings of $xy$ coordinates representing “change vectors.” At each time “layer,” new and obsolete locations are noted for specific entities as they appear, disappear, grow, shrink, and move. These successive increments are stored in this model by noting only the timestamped “change vectors.” Similar to the grid representation, the “present” world state for any or all entities can be easily retrieved by referencing the most recent change vectors. Although the temporal integrity of individual spatial objects is maintained, the resulting representation for any real-world data quickly becomes unwieldy as individual map elements evolve through successive changes at various points in time. This is particularly true for situations where the typology of connected map vectors also changes as features come into and go out of existence as well as changing size and shape.

Following on this work, Worboys (1994) proposed an extension of the vector model where each object in space has a collection of boundary states through time. This work included specification of relationships from which a set of query operations could be defined. Raza and Kainz (1999) proposed an extension to Worboy’s model, but separated out space, time, and attributes as distinct elements. A different approach was proposed by Yick and colleagues (2002) that utilized the Voronoi spatial model to record topological relationships in an initial state and a set of operations to express changes, keeping only the changes in a log file.

Time-based representations

All the extended and vector models described above incorporate the temporal dimension in some way while retaining their functional organizational basis. They thereby also retain their relative functional advantages and disadvantages. All vector models are entity-based in the sense that all locational and temporal information, as well as other types of attribute information, is stored relative to specific geographic/cartographic entities, and their topologically defined components (lines and nodes). In other words, geographic/cartographic entities serve as the basic conceptual element
and organizational basis of the representation. Conversely, the grid model, or more generally any tessellation model, can be said to be location-based, since all other information is stored relative to specific locations. These two models correspond to the dual entity-based and field-based conceptual approaches, respectively, as discussed by Couclelis (1992).

Because of these two fundamentally different orderings of stored information, the vector model can be used most effectively to store information and perform tasks and queries relative to spatial entities, including determining topological relationships between them over space, but not nearly as effectively to store information and perform tasks relating to a specific location or set of locations. Conversely, a grid or tessellation model can much more effectively perform tasks relative to specific locations and sets of locations (Peuquet 1988).

In the time-based approach, location in time becomes the primary organizational basis for recording change. The sequence of events through time, representing the spatiotemporal manifestation of some process, is noted via a timeline, that is, a line through the single dimension of time instead of the two-dimensional (2-D) surface over space. Such a timeline, then, represents an ordered progression through time of known changes from some known starting time to some other known, later, point in time. An early example of this, proposed by Peuquet and Duan (1995), was the event-based spatiotemporal data model (ESTDM).

Each location in time along the timeline (with its temporal location noted as \( t_0, t_1, \ldots, t_n \)) can have associated with it a particular set of locations and entities in space–time that changed (or were observed as having changed) at that particular time and the new value to which they changed. Recording only times when change occurs, as opposed to a complete timeline that contains an entry for every “tick of the clock” at a given temporal resolution, avoids storing redundant data. In other words, the changes relating to times are explicitly stored, and only the changes. This top level of information where a time of change is recorded is called the “event list.” (An event is defined as an instant in time when a change occurs, that is, change in some property or attribute that can be denoted as such for some entity, location, or set of entities or locations.)

In the time-based approach, the time associated with each change (i.e., each event) is stored in increasing order from the initial “world state” at time \( t_0 \) (e.g., January 21, 1952) to the latest recorded change at time \( t_n \). These may be recorded at any desired temporal resolution. For most phenomena, the length of any temporal interval (i.e., temporal distance between \( t_{i-1} \) and \( t_i \)) and any other such interval will typically be unequal. Associated with each \( t_i \) are the changes that occurred between \( t_{i-1} \) and \( t_i \). The only exception to this is that the entire base map or starting world state must be stored with the first recorded time, \( t_0 \). The changes associated with any \( t_i \) may also be extensive, affecting a large geographic area and many entities within it, or perhaps may be only a single location or entity. The timeline is compacted in that temporal locations where no change occurs are not recorded. It is assumed that each event list and associated changes relate to a single thematic domain (e.g., land use or population). Each stored event can relate to a location or set of locations, or to a set of entities or set of entities.

This type of representation has the unique advantage of facilitating time-based queries (e.g., retrieve all events that occurred between May 1, 2016, and September 15, 2016). Adding new events as time progresses is also straightforward; they are simply added to the end of the timeline. A similar timeline approach, with associated temporal topology and query operators, was
described by Claramunt and Thériault (1995). An alternative approach was suggested by Frank (1994). In Frank’s approach, the timeline is strictly ordinal where the precise dates of events are not known. Such a limitation is encountered in various application domains such as archaeology and ecology. Nevertheless, a strictly ordinal timeline facilitates the comparison of entity histories and parallel sequences of events.

Recognizing that events can be viewed hierarchically at differing levels of detail, Yuan (2001) proposed a hierarchical conceptual structure where an event has duration, with a starting and ending time. Within this time span, an event is composed of multiple changes in space and time. Mountrakis and colleagues (2002) instead proposed dealing with differing levels of detail in event data through the use of aggregation operators.

In parallel with these efforts on general spatiotemporal representation schemes, work began to appear to provide structured methodologies for deriving models tailored to specific applications. Raper and Livingstone (1995) proposed OOgeomorph—an object-oriented technique for spatiotemporal representation of geomorphologic processes. Wachowicz (1999) demonstrated how object-oriented design methods can be used for modeling spatiotemporal data in a wide range of applications including boundary record maintenance and analyzing the spread of disease and the environmental impacts of climate change. There was also a significant amount of work on structured methodologies for spatiotemporal representation within the database management systems (DBMS) community. These included work by Parent and her colleagues (1999), who proposed a spatiotemporal modeling technique called MADS (modeling of application data with spatiotemporal features) for bridging the application-view/implementation-view gap. Tryfona and colleagues (1999, 2003) described an extension of the entity-relationship (ER) database design technique, which they called STER (spatiotemporal ER), and an extension of UML (Unified Modeling Language) elements to directly support design of spatiotemporal data models. Their STER technique utilizes a set of very basic ontological principles dealing with the unique aspects of the space–time domain— that properties and interrelationships of entities can be defined on the basis of either discrete or continuous change with respect to spatial change (i.e., movement, change of size, change of shape) and/or nonspatial change (e.g., change in temperature). Chomicki and Revesz (1999) proposed constraint databases as a common language for facilitating interoperability of spatiotemporal data models.

The space–time cube

The space–time cube was first employed as a spatiotemporal representation for geographic data in the early 1970s by Hägerstrand (1970) in his work on what became known as time geography. He used it as a convenient graphic device for showing movement of individuals in a “space–time path” and how people interact in space–time. A related concept was the “space–time prism,” which encompasses the extent of potential movement within given space–time constraints. The base of this three-dimensional (3-D) diagram represents continuous, planar space, and the vertical dimension of the cube represents continuous time. Advancement of time geography, however, was hampered by the very limited capacity of computers of the day as well as difficulties in collecting data. Both rendering of the space–time cube representation and data collection/processing were predominantly manual processes, and therefore also tedious and time-consuming.
As mentioned earlier, the renewal of interest in spatiotemporal analysis and representation in the late 1980s was coincident with the development of advanced storage and processing capabilities of modern hardware and advanced networking and software technology, as well as with large amounts of continuous spatiotemporal data becoming available. This spurred renewed interest in Hägerstrand’s work and what has been called the “new” time geography (Miller 2004). The space–time cube has been employed recently in transportation research as an analytical modeling tool, extending and refining the space–time path and prism concepts (Shaw and Yu 2009; Ellegård and Svedin 2012; Song et al. 2015).

Stefanidis and colleagues (2004) developed an extension of the space–time cube they called the ST Helix for the analysis and management of large motion imagery datasets. This model incorporates both movement and entity deformation within the same representation. A central “spine” models the spatiotemporal trajectory and protruding prongs express expansion/contraction and deformation of the entity’s outline at specific time instances.

Kwan’s work on the analysis of disaggregate activity-travel behavior (2000) focused on the use of the space–time cube as a 3-D visualization to facilitate exploratory analysis of space–time interrelationships of human activity and travel patterns. While modern computing capabilities allow interactive pan, zoom, rotation, walk-throughs, and fly-bys, the use of the space–time cube as a visualization tool is inherently limited in its capability to display more than very modest amounts of data since individual movement traces quickly become obscured in a visual tangle of lines.

Manipulations such as 3-D rotation and fly-bys can help overcome the scaling problem to some extent, but these can also be perceptually disorienting. The space–time cube also assumes data are continuous over both space and time. This can be particularly problematic with historical reconstructions. In Nakaya’s recent work (2013) focusing on three example uses of the space–time cube, he argued that its support as a visual representation should be supported in the next generation of geographic information systems. While acknowledging the problems with this representation as a visual tool, he suggested some possible strategies.

A number of visualization researchers have explored ways to overcome the problems associated with scalability, disorientation, and missing data. Kraak developed a multiview environment, with dynamic linking, that includes 2-D and attribute views as well as a close-up view of the space–time cube that assist the user in orientation and navigation (2003). Another notable effort investigated use of the space–time cube to explore events represented as dots or circles (Gatalsky, Andrienko and Andrienko 2004). Here, interactive temporal filtering and zooming were employed with the 3-D display linked to a (2-D) map.

Combined approaches for spatiotemporal representation

Associating temporal information with individual entities provides a means of recording entity histories, and thereby allows histories of entities and types of entities to be easily traced and compared. Similarly, associating temporal information with locations allows the history of individual locations and sets of locations to be traced and compared. Since these different types of representation provide differing perspectives on the data, each facilitates a different type of query.

Worboys and Hornsby (2004) showed how the entity–based paradigm could be extended
to incorporate event information and proposed a unified modeling approach for both events and entities. Earlier work recognized that location-based and entity-based representations should be considered complementary in a unified, dual field-based and entity-based conceptual framework (Peuquet 1988; Couclelis 1992). Based on this, Peuquet (1994) proposed that all three – time-based, location-based, and entity-based representations – need to be considered complementary and interdependent in a triad representational framework. A similar framework was proposed in an application context in Yuan (1994). This triad framework corresponds to human cognition in that people store and use knowledge about the world via interdependent location-based, entity-based, and time-based perspectives (Mennis, Peuquet, and Qian 2000; Peuquet 2002). Yang and Claramunt (2003) extended a combined representation based on locational, temporal, and thematic domains to allow for the representation of continuous change.

Van de Weghe and colleagues (2014) introduce a continuous spatiotemporal model (CSTM) in order to address the need for representations that can support multiscale analysis, since spatiotemporal processes commonly occur at multiple, interrelated scales. This has been a significant and long-standing issue, since analysis at a single scale can miss important patterns and relationships and thus present an incomplete, and perhaps misleading, portrayal of a phenomenon. While multiscale spatial and temporal representations are not new – as an example, Hornsby and Egenhofer (2002) presented a method to describe moving entities over multiple spatial and temporal resolutions utilizing lifelines within the space–time cube – the CSTM model provides a solution for multiscale representation of more complex and much larger volumes of data. The spatiotemporal hierarchy in the CSTM representation utilizes a tree structure based on the spatiotemporal evolution element (stevel). The stevel is composed of four variables: pixel location, spatial resolution, temporal interval, and temporal resolution.

Temporal and spatiotemporal query operations

If we loosely regard the location-based and entity-based views in the dual framework as storage of information relating to where and what, respectively, the time-based view in the triad representational framework provides storage of information relating to when. In the dual framework, the two complimentary representations together are able to answer the two basic types of static spatial questions. These questions in their most general form can be stated as follows.

1. Using the location-based view – given a location or set of locations, describe what things are there (where → what).
2. Using the entity-based view – given an entity or set of entities, describe their location (what → where).

For each of these, the given information represents the search keys, or elements that must be retrieved in order to provide the desired associated information. Extending this to include a temporally based view, there are then three basic types of questions, as follows:

1. when + where → what: describe what entities are present in a given location or set of locations at a given time or set of times;
2. when + what → where: describe the location or set of locations occupied at a given entity or set of entities at a given time or set of times;
where + what → when: describe which times a given entity or set of entities occupied a given location or set of locations.

Relationships, whether temporal, locational, or entity-oriented, are viewed as operators on stored observational values. As operators, they enable the user to combine data in an ad hoc manner and to combine relationships in order to express, and thereby search for, more complex patterns. In a manner similar to the way that a spatial distribution pattern can be random, uniform, or clustered, a temporal distribution can be chaotic, steady-state, or cyclic.

Utilizing a time-based representation, all temporal relationships can be divided into three distinct classes: (i) linear metrics and topology, (ii) Boolean operators, and (iii) generalization. Temporal relationships between and among entities specifically allow the examination of cause-and-effect as well as overall patterns (i.e., temporal rhythms). In other words, they are the key element that allows hypotheses regarding process to be derived and/or verified from stored observational data.

The first class of temporal relationships comprises those that operate along a single, continuous, linear timeline, or event vector, at a given temporal scale. These include metric and topological relations. Given that time is one-dimensional, there is only a single temporal metric: temporal distance. Temporal distance is the measure of the length of the interval between any two given points in time. This metric can represent the length of a single event (i.e., its duration), the length of time between two events, or the duration of some continuous “world state.”

Since intervals between events are subject to variability in the recording and fineness of measurement, it is also useful to be able to express relationships between events strictly in an ordering sense. In terms of real-world analysis, this allows for examination at different temporal rates or timescales that may have influenced different processes. The study of temporal topology has a long history in a number of fields, including philosophy, mathematics, linguistics, and artificial intelligence (Allen 1984; Shoham 1988; Davis 1990). These temporal topology relationships are: before/after, equals (coincident), meets, overlaps, during, starts, and ends. Note that each of these seven basic relative interval relationships has an inverse. Since $X = Y$ and $Y = X$ is an identity, there are 13 possible temporal topological relationships in all.

The second class of temporal relationships comprises those that operate specifically to combine separate event vectors (e.g., fire history and weather history). These are the temporal overlay relationships that act as Boolean set operators (intersection, union, negation) to associate the overall concurrence of different event types over the entire length of a given time interval.

The third class of temporal relationships is temporal generalization. Operators in this class associate specific events at various levels in the temporal scale hierarchy and are thus derived from the fourth characteristic. The method of generalization is dependent on the metric used (e.g., four seasons, 12 months, or 365 days (usually) each equal one year). It may also vary from scale to scale. It is temporal generalization that allows the rhythms of biological, cultural, or other types of phenomena that operate at different temporal scales to be compared.

In addition to temporal relations that operate on the time-based view, there are also temporal relations that operate on the locational-based view, and still others that operate on the entity-based view. The most apparent location-based temporal relationship is a location-based temporal overlay, utilizing Boolean functions to compare changes (or similarities) at the same set of locations over a series of different times in
a snapshot representation. This corresponds to how overlay is defined for a series of different conceptual layers.

A typology of entity-based temporal relations was presented in the work by Claramunt and Thériault (1995). These were further explored in an application analyzing dynamics of precipitation patterns in a projected climate change scenario (Bothwell and Yuan 2011). Claramunt and Thériault enumerate separate sets of temporal relationships for the evolution of a single entity, relationships between entities, and evolution of spatial structures involving two or more entities, as follows.

1. Evolution of a single entity: (a) existential changes: appearance, disappearance; (b) transformation: expansion, contraction, deformation; (c) movement: displacement, rotation.

2. Relationships between entities: (a) replacement processes: succession, permutation; (b) diffusion processes: production, reproduction, transmission.

3. Evolution of spatial structures involving two or more entities: (a) restructuring: split, join, reallocation.

Diffusion processes involve a transfer of characteristics between spatial entities. Production involves the creation of new entities by the behavior of one or more entities of differing types. Reproduction is the creation of new entities of the same type, and transmission is the modification of the characteristics of one or more entities due to the influence of one or more other entities. These diffusion relationships are truly important in space–time analysis as cause–and–effect relationships. Renolen (2000) defined seven entity-based temporal relations: creation, alteration, destruction, reincarnation, split, merge, and reallocation. These are similar to those given by Claramunt and Thériault but without the diffusion process relations. Hornsby and Egenhofer (2000) derived a set of 18 identity-based change operations in what they termed the identity-based change (IBC) model. Their approach highlighted the minimum elements necessary for reasoning about identity, as well as the ordering and co–occurrence of identity states. This work was recently expanded by Yi and colleagues (2014) to account for identity states of entities that are components of a composite structure. Worboys (2005) explored entity-based temporal relationships using algebraic approaches. This work was later extended to represent reasoning about complex entity changes using trees and structure-preserving mappings between them (Jiang and Worboys 2009).

From world models to process models

The location-, entity-, and time-based models discussed above are all designed to represent observational data in a world model. This conforms to the fact that many spatial and spatiotemporal analysis techniques used in GIScience, and geography more broadly, are observation-driven methods. The development of understanding of a phenomenon, however, involves a significant amount of theory-driven investigation. This is true for both science and everyday cognition. Theory–driven investigation in science is the domain of modeling and simulation. Well-known examples for modeling geographic processes include UrbanSim for simulating urban systems (Borning, Waddell, and Förster 2008) and IBIS (Integrated Biosphere Simulator) for vegetation dynamics (Foley, Kucharik, and Polzlin 2005). These tend to be stand-alone and focused on a highly specific context because of the complexity of the process being modeled.
Goodchild has stated that GIScience, however, has been focused on the study of form rather than process – of how the world looks rather than how the world works (2004). Thus, while vast amounts of spatiotemporal data can currently be stored, accessed, manipulated, and analyzed, what can actually be learned from observational data in terms of deeper understanding of the phenomenon under study relies completely upon, and is therefore also limited by, the expertise, experience, and memories of the specific people studying the data at a given time. This can be problematic, particularly in an emergency situation when insight and an actionable response are quickly needed. Process models, on the other hand, are designed to represent how a particular phenomenon works – the mechanisms involved that effect change. As such, this is higher-level, derived information involving the interactions among entities that are interlinked, with complex chains of interactions.

Raper and Livingstone (1995), as part of their argument for object-oriented data representation within the context of environmental phenomena, argued that process models and data models should be linked, thereby extending the triad approach described above for multirepresentation of data of the what/where/when scheme into a what/where/when + how framework. While direct linkages of current, complex simulation models with spatiotemporal data models as known in GIScience would be over-reach as a first step, linkage to a simpler structure for stored, higher-level rules of association and known cause-and-effect would allow two significant enhancements to current capabilities (Peuquet 2009). First, it would provide a common and perhaps negotiated set of derived views on which collaborative analysis could be based. Second, it would allow information or insights captured as the results of analyses to be explicitly stored and potentially used to inform subsequent analyses.

Yi and colleagues (2014) recently proposed such a composite structure. Their framework consists of a hierarchy of representations. The bottom of the hierarchy consists of what they call static structures: the shapes, locations, and characteristics of entities derived from a series of remote sensing images viewed as observational snapshots. At the mid-level of the framework are stored sequences of footprints for individual entities during the entire time span of the observational data assembled from the static structures. The top level contains what they call scenarios, which are composite structures representing interactions among the entities and groups of entities. They expand on Hornsby and Egenhofer’s IBC model (2000) to describe the changes and interactions of individual entities and composite structures.

While Yi and colleagues were able to build a composite structure, deriving the higher-level information required a significant amount of human input to interpret algorithmic results. Deriving higher-level information needed to build a process model from the large and complex data currently available requires a more automated approach. This brings up two well-known problems related to temporal reasoning, which have become known as the frame problem and the ramification problem (Shanahan 1997; Papadakis, Antoniou and Plexousakis 2006). Both of these are a direct result of the necessary incompleteness of stored world observational information. The frame problem is the problem of knowing what world states and other events are not changed by a given event or action. For example, migration of workers does not change their gender. The complementary problem is the ramification problem. This is the problem of predicting how one entity or location state will automatically change when another does due to some relationship connecting them.
REPRESENTATION: TIME

One solution to both of these problems is to explicitly enumerate the specific entity/location states that are or are not affected, but such exhaustive enumeration can have obvious drawbacks. The other solution is to incorporate generalized rules regarding what types of entity/location states may or may not be affected by a given event type. In either case, it is necessary to be able to separately and independently reference events, locations, and entities in order to reference these relationships and store this higher-level information within the specific type of element to which the information pertains. Brown and colleagues (2005) proposed the use of agent-based models for storing rules of how locations, entities, and events interact within a tightly coupled process—a multirepresentation data model environment. As part of this, they also discussed four key relationships that affect how the world model and process model interact: identity, causal, temporal, and topological. Torrens and Benenson (2005) proposed a similar approach that combined agent-based models with cellular automata. That same year, Reitsma and Albrecht (2005) proposed an alternative approach for representing process that tries to take the integration of behavioral rules a step further. Based on their critique that the rules in agent-based and cellular automata approaches typically represent the relationships only between the current and future states of cells or agents, they extend process representation to include the set of rules that govern interactions with other processes.

The problem remains of how to derive these rules and relationships. But also the question arises as to whether these are among the various types of associations that describe the process model that we are trying to uncover to begin with. This gets down to the ultimate goal of extending a world model, represented as locations, entities, and events, to a process model that can be represented as sets of quantified relationships and constraints.

Certainly, one way to get around the issue of needing a priori rules is spatiotemporal analysis in the form of exploration, or what Tukey (1977) called exploratory data analysis (EDA). This entails simply looking at the data in various ways visually and statistically to see what patterns and associations the human analyst can perceive. The visual aspects of this have been termed “geovisual analytics” within the GIScience community (Andrienko et al. 2007), and has gained significant attention for exploring and gaining insight relating to spatiotemporal processes (Beard 2008; Andrienko and Andrienko 2006; Chen, MacEachren, and Guo 2008; Kraak 2014). Both visual and statistical approaches, and their combination, as approaches to EDA are becoming a set of increasingly valuable tools, particularly in the current era of big data, where the volume of observational data about the Earth, once in short supply, now exceeds the human capacity to absorb all information contained within it without help from computers.

Cognitive science holds some insights into a method that would also allow computers to assist in deriving these relationships from complex observational data based on how humans learn such relationships in everyday life so that they can subsequently be verified and formalized in process models. Pazzani (1991) described a process that uses a theory of causality as background knowledge. The process, which he called theory-driven learning (TDL), hypothesizes causal relationships consistent with observed data and a general theory of causality.

Theories of causality are able to assist the learning process and allow cause-and-effect associations to be made more quickly than by a simple process of repeated observation to yield a conclusion, as well as allowing the connection to be made between a specific type of event...
and a specific world state change with causal links. It is these theories of causality that are gradually accumulated and refined, thereby progressively achieving greater accuracy. They are inductively learned over a series of temporally based observations comparing before-after states. This inductive learning also is based on a few elemental causal constraints. The elemental causal constraints that influence the acquisition of causal relationships from the observed world as discussed by Pazzani include the following.

- **Regularity** – An effect must regularly and repeatedly occur with its cause.
- **Temporal order** – An effect must happen after the cause.
- **Temporal contiguity** – An effect normally happens very soon after the cause. The more temporal separation, the less likelihood that there is a cause–effect relationship.
- **Spatial contiguity** – An effect normally happens in contact with (or very near) the cause. The more spatial separation, the less likelihood that there is a cause–effect relationship.

It is important to note that regularity and temporal and spatial contiguity are not absolute requirements, but provide generalized guidelines. Spatial continuity applies as a causal constraint only two processes involving spatial interaction. There are many processes involving social interaction that may not have a direct spatial component, particularly in the current era of electronic communication. All such constraints, however defined for a particular context, function to focus attention on the potentially relevant features and to establish chains of cause–effect relationships. They also help to ascertain the true cause of an event, such as being able to distinguish whether a tree fell down because its trunk had been sawn through rather than as a result of some other independent event that may have happened at the same time (e.g., a bird landing in the tree).

The combination of a large amount of complex observational data with TDL allows explanations and predictions to be made on the basis of relatively few selected factors, but also for the greater amount of information to be tapped into when exceptional conditions arise. Specific known or presumed causal relationships can be stored with the entities as causes/caused-by links. Causal constraints and the other rules regarding how entities, locations, and times can (and cannot) interact can also be stored.

### Issues and challenges

Significant progress has been made since the early 1990s in the development of spatiotemporal representations that go beyond extensions of traditional, static spatial data models. Along with evolution of an underlying theory of space–time representation, semantically driven representations and operators are being developed that seem more “natural” to the human user, and at the same time take advantage of modern computing capabilities. Two issues, in particular, remain as relatively unaddressed that have unique aspects with respect to temporal data, as follows.

### Missing and erroneous data

While we are currently in an era of continuous and fairly exhaustive spatiotemporal data collection with the use of geosensor technology, many historical databases also are being developed from archival sources that provide a wealth of information on past world states and events not previously accessible in any variety or volume (Knowles 2000; Knowles and Hillier 2008). These historical archives often contain gaps of missing information from some long-past time.
that cannot be revisited, and the past states or events often cannot be interpolated with any degree of certainty.

There is, however, a significant body of literature on how to evaluate and/or correct error and uncertainty in data, mostly generated within the image processing community. Some of these methodologies have been utilized to improve the accuracy of transitions noted from one time to another using erroneous maps, in effect distinguishing actual change from noise (Pontius and Li 2010).

As pointed out by Bol (2013), place names provide an essential link between the historic and the geographic. Data including tax records, population, and economic activities are linked to a place, yet place names are dynamic. Many of the primary sources of geographic names do not include time as an attribute of place. This includes those provided by the US government, including the Geographic Names Information System for place names within the United States and the GEOnet Names Server (NGS) for place names outside of the United States. The largest nongovernmental gazetteer, Geonames, also does not include time as an attribute.

One strategy for addressing the problem of missing and erroneous data is information fusion. As related digital data are accumulated, related elements can be compared to help validate elements and “fill in the blanks.” Information fusion is the process of integrating information from diverse sources and is a fairly standard process for business databases. Duckham and Worboys (2005) provided one example where this approach has already been applied to spatial information data.

Multiple times and alternative histories

Spatiotemporal representations developed within GIScience are designed to handle time with respect to events as they are reported to have occurred. But representation of the temporal dimension is not always that simple. It has long been recognized in the database community that for temporal databases there needs to be two kinds of times—“valid time” and “database time” (Snodgrass 1986). These relate to the time the event happened in the real world, thus often called “world time,” and the time when information relating to that event was entered into the database, respectively. The distinction between the two, and the ability to record both, is important in business applications, particularly banking. How to effectively represent both types of times within spatiotemporal databases was discussed as an important issue for applications such as cadastral systems early on (Al-Taha 1992; Worboys 1994). And as these and other researchers have noted, these two times can be noted as attributes in an entity-based representation, with the world time and database time for each boundary change as an example.

This idea of multiple time attributes can be extended. In the case of a spatiotemporal database for an archaeological dig, there can be the need to store multiple kinds of times: the date an artifact was found, the date the artifact was entered into the database, the date attributed to the entity as the time of its fabrication, and the date attributed to the stratum in which the artifact was found. But there are other kinds of time where representation is not as simple. The representation of multiple times to represent alternative histories or alternative futures is still problematic. This is an important issue in that the capability for such representation is important for the study of process, looking both backward in what are potential explanations as precursor conditions for observed events, and forward from observed events in the realm of prediction. This represents a critical link between observations.
and abstract process models, either mathematical or rule-based, as simulations.

SEE ALSO: Data model, event-oriented; Data model, moving objects; Representation: dynamic complex systems; Representation: fields; Representation: flows; Representation: trajectories; Time geography and space–time prism

References


A trajectory is the spatiotemporal footprint that an object leaves behind when its movement is tracked. As tracking technologies get better and cheaper, trajectory data advance to an important type of geographic information representing the movement of people, animals, goods, or any object that moves in geographic space. Even though all trajectories capture the spatiotemporal footprint of mobile objects, they do so in various ways and, hence, come in many diverse forms. They often come as lists of time-stamped positions recorded by a GPS receiver, for instance, fitted to a collar attached to a deer or a migratory bird in an ecological study (Figure 1a). Another typical form of trajectory results when connecting the sequence of positions where a mobile phone user made a call or connected to the Internet (Figure 1b). Yet another type of trajectory emerges when lining up the locations of a commuter’s check-ins into the electronic ticketing system of an urban transit network (Figure 1c). The interest in movement is omnipresent in geography; hence trajectories of a most diverse and wide range of objects are recorded, with examples including migrants or commuters in human geography, cars, truck or taxi fleets, ships, airplanes, or containers in economic geography, or tornadoes, icebergs, or even individual rocks in a rock slide in physical geography.

Major dimensions of representing movement trajectories

For a better understanding of the dynamic processes producing the observed movement, the respective trajectories must be abstracted and conceptually modeled such that they can be represented and then analyzed in computer systems (such as GIS or spatial databases). The way in which a specific movement process can be represented as trajectories results from the interplay of four important dimensions:

- the characteristics of the underlying (geographic) space embedding the movement;
- the characteristics of the movement process itself;
- the perspective of observation;
- the semantic aggregation level that is of interest in a given study.

In the following, the above four angles are illuminated. For every angle important properties are discussed and respective categorizations are proposed.

The choice of trajectory representation matters because the representation used has implications on suitable techniques for analysis. The representation rules which techniques can be used, for example, for assessing the similarity or dissimilarity of trajectories, or the way in which collective movement patterns can be formalized and algorithmically detected. For example, geometry-based measures from computational geometry might be suitable for assessing the similarity of line-shaped GPS trajectories, whereas
the similarity of mobile phone call logs might be better studied using sequence rule mining borrowed from molecular genetics.

Properties of the movement space

It is widely acknowledged that the representation of movement involves both the spatial and temporal characteristics of the trajectories, but crucially also the structure and properties of the space embedding that movement (Andrienko et al. 2008). In the following paragraphs key characteristics of the most commonly used spaces embedding movement are investigated.

Continuous versus discrete movement spaces

A first important distinction can be made between continuous and discrete movement spaces. In some contexts, the underlying movement space is best represented as a continuous Euclidean space. For instance, the trajectories of birds flying across the sea or a deer roaming through the woods might be perceived as sequences of fixes that can express any \( x \)- and \( y \)-coordinate tuple in a Euclidean space (Figure 2a). By contrast, the space embedding phone call histories from GSM networks is typically modeled as discrete Voronoi cell-shaped neighborhoods around antenna tower locations. Such a trajectory will then consequently be modeled as a sequence of visited Voronoi cells (Figure 2b, sequence \( h-i-j-f-g \)). Similarly, when cars are tracked in between cordons located at the nodes of a street network, their trajectories take the form of a sequence of discrete nodes and edges (see the networks in Figures 2b and 2c).

Conceptual models and data structures

The common GIS field and entity conceptual space models offer a second useful categorization for movement spaces. Deer in the Yellowstone National Park might be represented as moving over a field-based digital elevation model (DEM) (Figure 2a a.2). When their movement is simulated using an agent-based model (ABM), the underlying movement space will most likely also be represented using a grid-based raster data structure. By contrast, the same deer can also be represented as moving across an entity-based land-use layer consisting of habitat polygons enclosed by street and water body networks (Figure 2a a.3). Clearly, an entity-based
Figure 2 Movement spaces, movement traces, from varying observation perspectives: (a) unconstrained and limited movement in a Euclidean space, also modeled using field versus entity conceptual data models, observed from the Lagrangian perspective; (b) network-based movement observed from the Eulerian perspective, where the moving object passes by static beacons, here GSM antenna locations; (c) again network-based movement observed from the Eulerian perspective, where the objects are located when they pass through ticket gates.

Movement space is also a discrete space. Whereas animals are seemingly free to roam around, most human movement activity adheres to transportation infrastructure that is best represented using network data structures consisting of edges and nodes. Hikers walk on trails, cars drive along streets, and even airplanes follow clearly defined air traffic networks (Figures 2b and 2c).

Irrespective of the use of field, entity, or network movement spaces, GIS offers a wide range of concepts for modeling the properties of movement spaces that influence or control the movement embedded in those spaces. For instance, using a network movement space allows for modeling one-way streets or speed limitations (Figure 2b). Alternatively, field-based movement spaces can come useful when movement across space must be allocated anisotropic costs.

**Dimensionality of movement spaces** In most cases movement is represented in 2-D spaces. A position fix is then given with the coordinate tuple \((x, y)\) or \((\text{latitude}, \text{longitude})\), typically time-stamped with a time \(t\). However, in some cases the movement space can be simplified to a 1-D linear space. When movement is modeled over a digital elevation model (DEM), the dimensionality of the space becomes 2.5-D, with the pseudo-third dimension being the unique elevation value per location. Whereas such a 2.5-D space does not allow for representing vertical movement, a true 3-D space would allow exactly this (Ware, Arsenault, and Plumlee 2006). Here, position fixes come as tuples of \((x, y, z, t)\); with multiple \(z\)-values for a unique \(x, y\) position). Also, time \(t\) can be perceived as the third dimension \((2D + T)\). Such a movement in a space–time cube goes back to Hägerstrand’s time geography (Hägerstrand 1970) and is still popular especially for visualizing movement (Andrienko et al. 2010). The temporal dimension is different since it is directed and permanently progressing, hence movement trajectories in space–time cubes take the form of upwards-directed lifelines.
Properties of the movement

The movement processes studied in geography are very diverse, and so are the properties of the respective trajectories. Just as with the embedding movement space, choosing the appropriate representation for the movement trajectory itself is a modeling process, and hence requires abstraction focusing only on certain aspects of the movement process whilst ignoring others. So, identifying the relevant characteristics of the studied movement is a pivotal task when choosing the suitable trajectory representation.

Unconstrained versus limited versus constrained/restricted

The most crucial fact is whether or not the objects are free to move wherever they want, or whether their movement is constrained in space. Again, a bird can fly wherever it wants. Its movement is unconstrained. Some moving objects are somewhat limited in their freedom to move by boundaries that cannot be crossed. Examples include sheep on their paddock, shoppers limited by the walls of a mall, or football players that at least during the game typically do not leave the pitch (Figure 2a a.1). Yet other moving objects are totally constrained by their movement spaces. Trains must follow their tracks and cars (mostly) use roads. A frequent combination of constrained movement is with network movement spaces, where the moving objects can only move along the edges of the transportation network (Figures 2b and 2c).

Whether or not movement is constrained by its embedding movement space critically influences the respective trajectory representation. Unconstrained moving objects are typically represented by lists of time-stamped positions \((x, y, t)\). By contrast, network-bound movement can be represented as a time-stamped sequence of the identifiers (ids) of visited edges and nodes (Richter, Schmid, and Laube 2012). Also continuous movement can exploit network structures: the position of a train on a leg of the transport network can be represented as a fraction of the already traversed portion of edge divided by the remaining portion \((v:w)\) in Figure 2c).

Uniform versus intermittent, periodic

Another characteristic of the studied movement that can be exploited for representing trajectories is the rhythm or “pulsing” of the movement. Some objects adhere to a rather uniform or permanent form of mobility. Examples could include migrating animals. By contrast, other objects express intermittent or periodic movement. Here, stationary phases are typically intermitted by bursts of movement. Commuters show such intermittent movement: most of the time they are at home or at work, with these stationary phases only being disrupted by regular or irregular movement bursts to the mall or the gym. This can become relevant when deciding how to track such different movement forms. It may not be feasible to record large volumes of redundant GPS fixes of an object that is most often stationary. Here, a trajectory representation could be chosen that models steps or moves between stationary points of interest. Clearly, uniform and intermittent movement can also characterize distinct phases or segments of the same moving object. A bird’s trajectories can express phases of uniform migration but also phases of intermitted movement bursts when, for example, feeding.

Active or passive

Finally, movement can also be categorized into active and passive forms. In active movement forms, the moving objects are typically self-propelled, and often express purposeful and goal-directed movement. Salmon returning to spawning grounds show very much active and goal-directed movement. By contrast, jellyfish could be carried away passively by an ocean stream. Other examples for passive movement include particles in an air stream or survey pegs...
on a downhill creeping landslide. This last distinction has perhaps little direct implication for the choice of a suitable trajectory representation, but it is closely related to the Eulerian and Lagrangian perspectives of observations that are discussed next.

**Perspective of observation**

Movement must be observed before it can be studied or analyzed. Just as conventional static spatial variables can be sampled in different ways, so can movement be observed adhering to rather different observation perspectives. Again, depending on the perspective of observation, the very same movement can be represented using rather different concepts. The perspective of observation is clearly very much related to the tracking technology that is used: GPS data are quintessentially different from mobile phone call data records and e-ticketing check-in histories. Many movement studies exploit third party data, where the choice of a tracking system with all its implications lies beyond the reach of the data analyst. However, when a movement study is designed from scratch, it is well worth considering the implications of the tracking system on the representation of the trajectories to be captured.

**Lagrangian versus Eulerian perspective** Movement can be perceived from two different perspectives, inspired by the Eulerian and Lagrangian representation of fluid dynamics in physics (Turchin 1998). The Lagrangian perspective considers changes in a moving object’s position. Here, the moving object produces a sequence of position fixes as it moves across space (Figure 2a). GPS tracking is the classic example for the Lagrangian perspective, where at a given sampling rate the GPS receiver measures the object’s position as a tuple of the form \((x, y, z, t)\). By contrast, the Eulerian perspective describes movement as changes in position of moving objects relative to known, fixed locations in space. Movement is perceived as a flux of objects passing by check-points or cordons. Eulerian trajectories are produced by groupe spécial mobile (GSM; a global system for mobile communication) antennas that monitor passing-by mobile phone users (Figure 2b), radio frequency identification (RFID) tag readers that monitor the movement of shopping trolleys, or checkpoints and traffic gantries that monitor the flow of passing cars (Figure 2c). The key difference to the Lagrangian perspective is that here the location of the checkpoints is known and fixed, and the movement is represented in the form of time-stamped sequences of IDs of passing objects. Recent developments in GSM and mobile information and communication technologies (ICT) promote the latter perspective, as more and more applications track moving objects in cordon-structured systems.

**Sampling regime** A conventional study of static geographic variables requires setting a sampling regime; for example, choosing between a systematic sampling raster and scattered focus areas. The very same holds for sampling positions of moving objects, just that here the temporal dimension takes the role of the frame of reference for the sampling. Some movement studies require that the positions of the moving objects be permanently monitored. A GPS tracking device with a preset sampling frequency can deliver just such continuous sampling. Even though a GPS receiver can deliver continuous tracking, where battery power and communication bandwidth are limited (as, for instance, with GPS collars attached to animals in costly capture-release procedures), noncontinuous sampling regimes are used. In movement ecology, where the study of animal behavior is key, fine-grained bursts are often interspersed with longer intervals with much sparser sampling. Another strategy for avoiding
unnecessary depletion of sensor resources is offered by *event-based* sampling. When it is known that the movement performed by the studied actor is of the periodic type, the sampling regime can exploit this knowledge and only record the moving object’s position in the event of a position or behavior change. Finally, in many Eulerian tracking systems, the sequence of position fixes is governed by the times when a moving object happens to pass a check-point. Hence, the samples will come in *irregularly*, depending not only on the movement behavior of the actor but also on the arrangement and density of check-points. Event-based sampling and irregular sampling are closely related to the question whether the observer can enforce a position fix to be taken or not. This distinction is discussed in the next paragraph.

*Active or passive tracking* The distinction between active and passive tracking systems has an influence on how movement is appropriately represented. In *active* tracking systems the control on when and how often position fixes are taken lies in the hands of the observer. For example, when tracking users of mobile phones, an active tracking system can trigger the localization by presetting a sampling frequency or by remotely “pinging” the phone. By contrast, in *passive* tracking systems, the observer has no control over when and how often positions are captured. Again when tracking mobile phone users, in a passive setting the mobile phone’s location will only be logged when the user makes a call or accesses the Internet (so-called call data records or CDR). Clearly, the latter tracking procedure results in trajectory data of much lower quality. However, whereas active tracking is typically limited to rather small numbers of tracked individuals in the low hundreds, the sample sizes possible with passive tracking are almost limitless (Ahas et al. 2009).

*Off-line or real-time* The perspective on movement and how it is represented is furthermore depending on whether the objects must be followed in *real-time* or in an *off-line* mode when the movement is over. For example, air traffic control or a taxi fleet management requires real-time tracking, whereas sports scene analysis can easily wait until the game is over and done with. With respect to choosing a suitable trajectory representation, the key difference lies in the fact that for off-line tracking the movement is over and known in its entirety, whereas representation for real-time tracking must accommodate for an incoming stream of location information packages that immediately after their arrival are already outdated as the object has moved on in the meantime. The database community has addressed such issues and put forward concepts explicitly tailored for storing, managing, and querying moving object data in databases.

Most prominently, Güting and Schneider (2005) introduce the moving object spatiotemporal (MOST) data model for moving object databases (MOD). MOST explicitly aims at representing the current and even future movement of mobile objects in motion *right now*. The key ingredients of the MOST model are so-called dynamic attributes. A dynamic attribute, $A$, changes its value with time automatically, and consists of $A.value$, $A.updatetime$, and $A.function$. For instance, the permanently changing position of a moving car can then be described as pos: $(x: \text{dynamic real}, \ y: \text{dynamic real})$, whereas other car attributes are represented using conventional data types (licence_plate: string). The current position of the car is then represented as an extrapolation from the last known location $A.value$ based on $A.function$ and the progressed time. The challenge lies in finding a good balance between updating costs and acceptable imprecision of the estimated positions. MOST is closely linked to a query language called
future temporal logic (FTL), which allows for expressing queries about future developments. Clearly, extrapolating the position of a moving object into the near future requires a notion of uncertainty. Güting and Schneider therefore argue that a trajectory can also be viewed as a motion plan for the future, introducing uncertain trajectories with an associated bound on the uncertainty of the time-dependent position. When compared against a polygonal query region, uncertain trajectories then allow for uncertainty-based spatio-temporal predicates, such as, for example, “SometimesPossiblyInside” or “DefinitelyAlwaysInside.”

**Obfuscation for safeguarding privacy** Fine-grained individual trajectories are inherently sensitive information. Tracking data may reveal where a person lives, where the person works, and places that person frequently visits. Even though more and more users of mobile ICT devices want to benefit from the various location-based services available, most users do it with severe reservations and acknowledging that these services compromise their personal privacy. The crux with tracking data is that conventional anonymization techniques fail, because personal points of interest such as the home and work address serve as accessible quasi-identifiers allowing for a re-identification of beforehand anonymized records. It is here that GIScience offers a perspective on movement that explicitly aims at safeguarding the user’s privacy. The key idea is to blur the focus of observation when following a moving individual. By definition, obfuscation is the process of deliberately degrading the quality of personal location information (Duckham and Kulik 2005). For many location-based services it may be sufficient to reveal only the current street block a user is in; the precise position of the user is hidden away from the service provider. For example, Figure 2b reveals only that the object is in cell $i$, but obfuscates its precise location. An obfuscated trajectory representation can be based on any combination of the familiar spatial data quality measures – inaccuracy, imprecision, or vagueness. An app searching for the closest petroleum station could hence reveal the user’s position as 250 m off the actual position (inaccuracy), within a neighborhood of radius 500 m around the actual position (imprecision), or as somewhere in between intersections $M$ and $N$ (vagueness).

**Semantic level of interest**

The semantic level of interest constitutes the final dimension having a profound influence on how movement trajectories are represented. Just as static geographic variables are often aggregated for analysis (e.g., household socioeconomic data is aggregated to neighborhood or district level), so movement data are often studied at a wide range of spatiotemporal aggregation levels (Laube et al. 2007). The spectrum ranges from the instantaneous level (fix, vector) through to interval aggregations (move, segment) up to global aggregations (convex polygon and field-based aggregations (Figure 3)).

**Fix** The most primitive spatial expression of movement comes in the form of a time-stamped position, a $\text{fix} \ (x, y, (z), t)$. A fix captures where a moving object was at a given time. This level of aggregation offers itself to studies where the fixes are attributed by some spatial variable, for example, the land-use type the map pins are sitting on. The resulting distributions are then fed into descriptive statistics.

**Vector** Strictly speaking, the above fix is not really a representation of movement, but rather only of static presence. The dynamic notion of the instantaneous level comes in the form of a vector, where the punctual measurement not
only features the position but crucially also has a movement direction (azimuth) and a speed. Note, even though from a conceptual point of view, a vector can indeed be sampled at an infinitesimal short interval, the actual computation of direction and speed typically requires at least two consecutive fixes or even longer intervals.

**Move** Neither the fix nor the vector has a spatial extent. By contrast, a move is defined as the displacement between two consecutive stopping points (Turchin 1998), hence having a spatial and temporal extent. A move can then be represented as a straight-line connector between two consecutive fixes or vectors. A move also has a direction, which can be different from the directions of its constituting vectors. Moves can span very different spatial extents and temporal intervals; they can, for example, represent the displacement of a butterfly moving from one flower to another or residential displacement in geospatial lifelines spanning decades.

**Segment** A segment represents the movement of an object over more than two consecutive fixes. As a geometric primitive it is a polyline, which can self-intersect. Most often segments emerge from a segmentation process, where a list of fixes is divided into meaningful segments according to some quantitative rules. A first natural notion of a segment emerges when tracking data are split into stops and connecting movement segments. Stops can formally be defined in many different ways, for instance, using a lower speed threshold. Further natural segmentation criteria are time of day (morning, noon, afternoon, evening) or the four seasons. Segmentation aims for aggregating several consecutive fixes, vectors, or moves into segments, which are homogeneous with respect to their geometric characteristics (e.g., sinuosity), semantics (e.g., travel mode walk or drive), or even in terms of some properties of contextual information (e.g., on highway or minor road). In some contexts such segmentation is called “semantic enrichment” of trajectories and can be considered a first analysis step (Baglioni et al. 2009). Others explicitly require trajectories to be semantically structured (i.e., segmented) in the first place, and hence to be explicitly separated from raw tracking data (Bogorny et al. 2014).

**Areal aggregations** Finally, in some application contexts the individual moves are less important, but what matters is the global movement footprint, summarized, say, for an entire day, a migration cycle, or even a lifetime. To achieve such a global level of aggregation, all relevant fixes can be aggregated to form an areal representation of the trajectory. GIScience offers a range of concepts for representing the spatial extent of a point set. The minimum bounding rectangle (MBR, also minimum bounding box) is simple and very easy to compute. A little more effort is required for computing the convex hull, or the minimum convex polygon (MCP). Both approaches share the simplicity of resulting in a vector structure that can conveniently be used for subsequent spatial analysis. Given their easy accessibility in many computing environments, field-based density...
Maps or heatmaps are popular representations for the global aggregation level. These typically come as some form of kernel density estimation (KDE), where for every location in a field the fix density is interpolated. “Hotspots” of space-use can then be interpreted as important places for the behavior of the observed individual (e.g., the home and office locations for people, or nest sites or feeding grounds for animals).

It is important to note that MBR, MCP, and KDE all ignore the temporal sequence with which a location was visited in space. It could also be said that they are strictly “spatial” representations of movement, as they only consider spatial proximity of the fixes but not their temporal proximity. However, when studying movement it may very well matter if visits to the same location are temporally clustered or, by contrast, temporally scattered, separated by long temporal gaps. Addressing this issue, density representations were developed that respect the sequence and timing of fixes. For example, the concept of “time-geographic density estimation” extends on traditional KDE taking into account a maximal speed value for modeling the potential path area (Downs 2010). The Brownian bridge movement model (BBMM) goes one step further and has become very popular in animal movement ecology. The BBMM is a stochastic model of movement, where the probability of occurrence in an area is conditioned on the position, sequence, and elapsed time between fixes (Horne et al. 2007). When comparing KDE density surfaces with Brownian bridge based ones, it becomes evident that the latter are much better at representing the pathways traveled in between hotspots.

The road ahead

The arrival of inherently spatiotemporal movement trajectories has challenged GIScience and the spatial database community, both adhering to a rather static view of the world. Initially, using the available concepts for representing movement was logical, but somewhat limited. Polylines trajectories, convex hulls or kernel density maps helped in acknowledging the issues but their static nature remains a mismatch to inherently spatiotemporal movement. A movement trajectory is just simply more than just its static footprint polylines – it involves sequence and rhythm. Clearly, there is progress. Examples include the Brownian bridge movement model (BBMM) explicitly focusing on sequence and timing of the fixes. Or then there is the moving object spatiotemporal (MOST) data model for moving object databases allowing for handling current and near-future positions for real-time trajectory management. The rather sophisticated nature of these two successful examples underlines that there is more complexity to movement data that could be presumed from seemingly simple stringing up of some position fixes. Further challenges arise from the fact that the characteristics and the sheer volume of available movement data undergo dramatic changes. Whereas initial movement analysis studies investigated relatively small datasets of, say, dozens of collared animals, passive mobile phone tracking can easily result in millions of tracked commuters, producing an endless stream of fine-grained tracking data. Up-lifting movement analysis into the era of big and open data requires the theory of trajectory representation to address the following challenges.

- Privacy. Safeguarding individual privacy remains a pivotal issue when tracking people. Even though the increasing pervasion of information and communication technologies is likely to change people’s attitudes towards privacy, it is neither to be expected nor to be wished for that the users’ fundamental disapproval of service providers...
monitoring their activities will go away. Trajectory representations are required that at the same time allow for providing location-based services, but put the control over sensitive movement data back into the hands of the users.

- *Multisensor data.* More and more researchers agree that the behaviors producing movement data cannot be fully understood from only studying the spatiotemporal movement trajectories alone. For that reason, many studies aim for semantically enriching position data with additional concurrent sensory data. This starts with obvious accelerometer and gyrometer data that are easily available from mobile ICT devices, but can also include more specific sensors, tracking, for example, physiological variables of tracked animals (e.g., body temperature, heart rate).

- *Dealing with notorious uncertainty.* Most trajectory representations are based in one way or another on simplifying assumptions that are knowingly incorrect. For instance, many movement analysis techniques are based on representations of linear moves between crisp and precise fixes. This is in stark contrast to our actual knowledge about an object’s location that is most often inaccurate, imprecise, and also vague. Another known aspect of uncertainty lies in what happens in between observed fixes. It is clear that a straight-line connector between two consecutive fixes is a massive simplification and rarely represents where the moving object actually was. BBMM go into a right direction accounting for a stochastic probability of presence. However, when comparing to the simplistic straight-line link, BBMM are computation intensive and laborious in handling. New trajectory representations seek approaches to handle a detailed notion of uncertainty, and also allow this uncertainty information to be carried along into the subsequent analysis process.

- *Conflicting information.* Further representations are required to accommodate for conflicting information about the positions of tracked individuals. In many cases it is clear that a tracked object was moving along a street, but its positions were recorded next to the street due to the tracking system’s measurement inaccuracy. Although there are sophisticated map-matching algorithms that help in relocating obviously incorrect fixes onto the street, such alteration of the raw signal inevitably distorts the geometric properties of the resulting trajectory, and analyzing geometry-based, derived trajectory parameters such as speed or turning angle from the altered geometry can become inappropriate. Novel trajectory representations should be envisaged to handle uncertainty and conflicting information depending on the analysis task at hand.

**SEE ALSO:** Geolocation services; Location-based services; Privacy, personal privacy; Representation: time; Spatiotemporal analysis; Time geography and space–time prism; Trajectories: analysis; Transport geography

**References**


Data.” In Mobility, Data Mining and Privacy, edited by F. Giannotti and D. Pedreschi, 15–38. Berlin: Springer.

Further reading

In its broadest sense, representation refers to both a process and a product. The process of representation involves describing, depicting, or portraying something. The product may take a variety of forms, including written, visual, aural, or oral forms, and also products that may be material, symbolic, or both. While representation has a variety of meanings, recent academic work in geography has argued that the term is used in a limited way within the discipline, leading to claims of a “crisis of representation,” and resulting shifts in the theoretical understanding of representation. However, a lasting effect of the recent focus on representation has been a more nuanced understanding of its relationship with power and politics, which continues to be influential within contemporary academic geography.

Representation in geography

Representation has long been at the center of geography as a body of knowledge and as an academic discipline. The origins of the word “geography” – geo, meaning Earth, and graphia, meaning writing – suggest this centrality of representation. “Earth writing” involves the processes of description, depiction, and portrayal. For early geographers, this process happened through oral and written accounts and through images, particularly maps. These maps were visual representations that described and portrayed known, lesser-known, and unknown places and peoples. They served to disseminate knowledge, to assert control and ownership over territory, and to signify civilization and advancement. Maps were not the only way in which early Earth writing occurred, but they are one of the most lasting forms. They clearly show representation both as a product and as a process. The product is the tangible map, while the process is the art and science of translating three dimensions into two and of claiming knowledge and ownership of the territory represented by the map. As geography developed as an academic discipline, geographers began to prioritize a more scientific form of Earth writing. This empiricist emphasis allowed geographers to identify as objective observers and documenters of reality. Representation as process was subjugated to representation as product, where the product served a very particular role. In this more scientific approach to geography, representations were valued for their mimetic quality – in other words, they were believed to accurately resemble “reality.”

While representation has been an important part of geography since its early beginnings, in the late twentieth century it began to take a more prominent position in the discipline. This emphasis emerged in particular from the work of humanistic geographers and cultural geographers. Humanistic geographers were concerned with the ways in which the quantitative revolution in geography had resulted in particular and narrow representations of human undertakings. They argued for a more expansive set of ways of depicting and understanding human activities, motivations, and beliefs. In this way, their critique focused on the limited representations of
human geographies within academic geography in the 1960s and 1970s. These critiques were developed and expanded by “new” cultural geographers such as Denis Cosgrove and Peter Jackson. Cultural geography, one of the longest-established branches of academic human geography, had traditionally placed an emphasis on landscape as its central object of study. New cultural geographers argued that the way in which landscape was understood was too limited. They noted that traditional studies of landscape focused on physical artifacts – for example, cabins, fences, and boundaries (Cosgrove and Jackson 1987). This emphasis on landscape as the tangible, material product of a particular cultural group at a particular point in time ignored or marginalized the broader processes through which landscapes were created, maintained, and altered. Cosgrove and Jackson argued that landscapes needed to be understood as both product and process. In order to develop this perspective, they described the landscape as a text, and sought to interpret this text using methods of literary critique and analysis. They noted the role of humanistic geography in suggesting this potential approach, but wanted to extend the work of humanistic geographers in order to consider landscapes as “configurations of symbols and signs” (Cosgrove and Jackson 1987, 96). Kong suggested that this approach saw the “landscape as a text to be read as a social document” using interpretive rather than morphological methods (Kong 1997, 178). The description of landscape as text brought questions of representation to the heart of cultural geography from the 1980s onward. This approach transformed the study of landscape, but it also extended beyond cultural geography in a significant way. The “cultural turn” compelled other subdisciplines in geography, such as cartography, social geography, political geography, historical geography, and economic geography, to ask important questions about representation. In many cases, such questions involved significant challenges to dominant epistemologies and methodologies within these subdisciplines, such as recent efforts to develop a postcolonial economic geography. In some instances, the meaning of representation has remained stubbornly resistant to challenge. Within electoral geography, for example, the term “representation” continues to be associated with the practice of democracy and with measuring the extent to which elections may be described as fair and proportional.

Earlier work on representation, particularly within humanistic and cultural geography, tended to focus on landscape and place. As interest in representation gathered momentum, the relationship between representation and space came under scrutiny. One approach suggested that space could be understood in the manner of a text within literary theory, positing space, particularly social space, as a text that has authors and is interpreted by readers. This approach was similar to the treatment of landscape as text in new cultural geography. An alternative approach argues that space and representations are mutually constitutive, that representations help to construct space, and that space in turn shapes and challenges representations. The second approach was influenced by the work of French theorist Henri Lefebvre. In his discussion of how space is produced, he distinguished three different components, which include “representations of space,” “representational spaces,” and “spatial practice” (Lefebvre 1991). Lefebvre described representations of space as verbal signs to represent the dominant forms of space in a society. In contrast, he described representational spaces as the nonverbal signs and symbols of the inhabitants who lived in, rather than designed, planned, and managed, a particular space. Lefebvre’s conceptual triad, which linked
representation and space together, has been used by geographers to discuss a wide range of topics, from cartography to urban unrest.

The crisis of representation

The “crisis of representation,” in geography and in the academy more generally, is a challenge to the belief that representations as products are accurate portrayals of that which they purport to depict. It draws from the work of theorists such as Michel Foucault, Roland Barthes, and Jacques Derrida who, in a variety of ways, showed how the representations we create and use are incomplete and contested and always subject to challenge and change. These theorists focused primarily on language, showing how the meanings of words are unstable, uncertain, and contestable, despite a widespread belief that words accurately portray both tangible and intangible things in a way that permits a shared understanding. Foucault introduced the concept of discourse, which he defined as a system of representation or as groups of statements about a particular topic at a particular time. Barthes drew on semiotics to highlight the rupture between the signifier (e.g., a word) and the signified (that which the word purports to represent). From Derrida, geographers imported the practice of deconstruction, which involves challenging specific positions, particularly by looking for what got displaced in the process of securing those positions (Doel 2005). While different in focus and emphasis, the work of these and other theorists has served to challenge the focus on representation as a product, and instead to direct attention to representation as a (partial) process. Geographers have drawn inspiration from this work in a variety of ways.

The crisis of representation emerged in studies of cartography from the 1980s onward. In his plea for an “epistemological shift” in the nature of cartography, J.B. Harley drew inspiration from Foucault and Derrida to challenge the idea of maps as “objective” or “scientific” (Harley 1989). Instead, Harley argued for the use of deconstruction in order to better understand the power of maps. His argument was further developed by Denis Wood and John Fels, whose deconstruction of the power of maps showed the ways in which maps use systems of signification to serve interests and to mask those interests. As Wood later commented, he began to see maps as (partial) texts and narratives. “Only cartography’s general reference map pretends otherwise, has the hubris to present the world, you know … as it really is, as if to say Europe, … the real deal, now and forever” (Wood 2013, 29). This led to a recognition of maps as social constructions and as practices (Kitchin and Dodge 2007).

The work of feminist geographers also contributed to a critique of representation in geography. Early work challenged the belief that the “human” in human geography was inclusive, pointing out the various ways in which women’s lives and experiences were not considered or included in geographical research. Later, Gillian Rose went further, arguing that geography was masculinist because it claimed to be universal while ignoring women’s experiences. She suggested that masculinism had shaped what kinds of research and knowledge were validated, how research was carried out, and how research findings were presented within geography (Rose 1993). Rose included new cultural geographers in her critique, claiming that attempts to construct landscape as text employed masculinist ways of seeing and knowing and served to support the idea of an all-seeing and all-knowing authority. As feminist geographies developed, they placed considerable emphasis on both representations and representational processes. In particular, the work of feminist geographers
showed the silences and absences at the heart of the construction of geographic knowledge. This broader approach has been used to highlight other exclusions, for example in relation to issues such as race, sexuality, and disability. Feminist geographers also directed attention to the power relations involved in processes of representation, and instigated a more reflexive, self-consciously subjective approach to knowledge generation and knowledge dissemination.

The identification of a crisis of representation in geography also resulted in attempts to broaden the methods and means of representation. Geographers have engaged with and experimented with alternative forms of representation, including art, photographs, literature, music, film, documentary, and dance. In some instances, this has resulted in the recognition of different forms of representation, such as fiction or comic books, as sources of geographic knowledge. In other instances, it has involved alternative approaches to research, for example using visual methodologies. It has also involved experiments with different forms of representation, for example using dance, although this has more recently been described as “more than” or “non”-representational geography.

The crisis of representation was initially associated with postmodern or poststructural perspectives within geography. Critics of postmodern and poststructural approaches often claimed that these approaches resulted in a relativism that made it difficult to make judgments or identify different levels of significance or importance. This critique persists. In recent years, non-representational geographers have developed alternative critiques of representation as understood in geography. In particular, they associate representation with social constructivism, and argue for a broader understanding of the term that focuses on practices and on relational analyses (Anderson and Harrison 2010). In these ways, a crisis of representation continues, although the reasons for this crisis vary. The scrutiny of representation, and the insistence on representation as a process as well as a product, have had an important impact on geography.

**Representation, power, and politics**

The renewed interest in representation that began in the late twentieth century has had a lasting impact on the discipline of geography. In particular, the focus on the process of representation – on who speaks for whom, and in what form – has resulted in a re-examination of the history of the discipline, and an interrogation of its means of knowledge construction and its methods of research.

The history of geography came under particular scrutiny from the 1970s onward, from geographers concerned with the ways in which representations of the discipline’s history obscured its links to broader questions of power and politics. Brian Hudson (1977) highlighted the intimate link between the growing institutionalization of geography as a subject in European universities, and European imperial and colonial practices. Hudson’s was an early intervention, but later interventions made more use of the work of Edward Said, particularly *Orientalism* (1978). *Orientalism* uses an approach known as colonial discourse analysis, which involves analyzing discourses produced by and about the colonial period as a means of examining how colonial power operated. Said examined the discourses (particularly texts and images) that the West (or Occident) produced of the East (or Orient) in the imperial and colonial periods. He showed how the Orient was effectively a European invention, albeit an invention with material consequences. The invented Orient served as the other for Western European powers, and also
provided a justification for the ongoing practices of imperialism and colonialism. Inspired by Said, geographers have since examined the ways in which colonized or marginalized peoples and places were and continue to be portrayed, drawing attention to the assumptions that underpinned geographic discourses during the colonial period and to the absences in these discourses, such as the accounts of local people or women. In these ways, geographers interested in the history of the discipline have interrogated the relationship between colonial geographies as particular forms of representation and broader structures of power. This is sometimes described as a postcolonial approach. Postcolonial approaches are concerned with representation as both process and product, seeking to show the colonial power relations at work in the processes of producing and interpreting particular representations.

Postcolonial approaches have also been influential – together with feminist and poststructuralist approaches – in raising questions about the power relations that influence the construction of geographic knowledge as a form of representation. Each body of work has a slightly different focus. Feminist theory highlights the masculinist nature of knowledge construction in geography. Postcolonial theory shows the extent to which geographic representations were (and still sometimes are) Eurocentric and excluded the perspectives of those who were colonized. Poststructuralism challenges metanarratives, which claim to explain the world in singular ways, such as Marxist geographies, and also challenges concepts of objectivity, rationality, and truth. Despite these different emphases, these approaches share a concern with the processes through which geographic representations are constructed and validated. The concerns raised created the possibility for constructing geographic knowledge in new ways, for example, indigenous geography, which has emerged since the 1990s. It is influenced by postcolonial and poststructural thought in particular, and has a number of priorities. Indigenous geography aims to create a space for research that engages with indigenous issues, some of which are distinct to indigenous communities and some of which are shared with a wider population. It seeks the “decolonization” of geography as an academic discipline, which includes recognizing the very different forms of geographic knowledge produced by indigenous peoples. It also demands research by and with, rather than on, indigenous peoples (Shaw, Herman, and Dobbs 2007).

The focus on questions of power and politics in relation to representation has also influenced how research is conducted within geography. In particular, it has challenged the idea of researchers as objective, and it has shown the need for the process of research to be more participatory and engaged. Feminist geographers have been to the forefront of these developments, and their work has directed attention to the politics and power relations present in research with marginalized, minority, or less powerful groups of people. This careful scrutiny of how representations are constructed, portrayed, and received is now a feature of much qualitative, and some quantitative, research within human geography.

Conclusion

At the start of the twenty-first century, questions of representation no longer receive as much attention within the discipline of geography as they previously did. In many ways, the critiques of representation that developed as a result of the crisis of representation have become mainstream within the discipline. As a consequence, representations are now generally understood as processes as well as products, and both process
and product are interrogated when representations are researched and analyzed. Instead, attention has turned to the limits of representation, particularly in relation to embodied or affective experiences that are preverbal. However, these nonrepresentational or more than representational geographies continue to rely heavily on text and images in order to make their arguments, albeit from a perspective that is cognizant of the power relations, politics, and incompleteness that mark any representation.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Critical geography; Cultural geography; Cultural studies; Cultural turn; Feminist geography; Feminist methodologies; Film; Gender; Humanistic geography; Indigenous knowledge; Nonrepresentational theory; Postcolonial geographies; Postmodernity; Poststructuralism/poststructural geographies

References


Marxist feminists in the 1970s discussed the role of reproduction and the contribution of women’s unpaid work to capitalist accumulation. In doing so they challenged the idea, commonly held by Marxists, that unpaid labor was unproductive labor as if value was only generated through waged labor in the public domain. Although there were various formulations, intense debates took place around domestic labor and wages for housework. The emphasis was very much on labor benefiting the male breadwinner with no attention paid to childcare and rearing or to the use of paid labor in the household itself, issues that would subsequently become significant in studies of reproduction. By the end of the decade, many feminists concluded that the discussion was too simplistic and narrowly focused on the role of women's domestic work in relation to the needs of capitalism and that it did not adequately explain women’s subordination in industrial capitalist, socialist, or “third world” societies.

In gaining a more complex understanding of women’s subordination in society, three different categories of reproduction were outlined (Mackintosh 1981, 11–12). The first concerned the reproduction of labor, which includes the production of people, involving their biological production, care, and socialization of children, as well as maintenance of adults throughout their lives and over generations. A second, narrower meaning covered human reproduction centering on marriage and kinship and thus issues of the social organization of fertility and sexuality, leading to the argument that it was women’s domestic responsibilities and the sexual division of labor in the household that weakened their participation in the wage economy. The third meaning of social reproduction referred very broadly to the ways in which the relations of production are recreated and perpetuated.

Historically, dimensions of social reproduction, such as childcare and responsibility for dependents, including the ill and the old, were understood as being played out differently in middle- and working-class families, as well as with changing gender relations and greater public and state responsibility for social reproduction in the twentieth century. Racial differences in studies of social reproduction, on the other hand, were not considered (Glenn 1992). In the United States, racialized and ethnic minority women, such as Mexican, Chinese, and black Americans, had historically been assigned distinct responsibilities for reproductive labor and performed the most menial tasks, whether in the household or outside of it. In the course of the twentieth century, there was a shift for these women from working in households to performing low-paid reproductive work in institutional sites, for example, as nursing aides (Glenn 1992).

By the 1990s, as migrant women increasingly came to fill the growing demand for reproductive labor in North America, Europe, and Southeast and East Asia, global analyses of an international division of reproductive labor were developed (Truong 1996). Recourse to paid domestic work was seen to be due to the inadequacy or withdrawal of welfare services, the increasing participation of women in paid work, and the inflexibility of the sexual division of labor in the household.
Although the use of the concept of social reproduction had declined since the 1980s, often displaced by other terms such as “care” (Kofman 2012), the assault on public provision and declining support for social reproduction and work security has led to a renewed interest in the concept and fresh directions of study. Feminist geographers have focused on the new forms of governance or social contract (new constitutionalism) and how local-global relations have impacted on the struggle over reproduction and distribution of risk, characteristic of neoliberalism (Bakker and Silvey 2008). Unlike previous analyses, where production and reproduction were treated largely as separate spheres, the two are conceptualized as structured practices that unfold in dialectical relation with each other, mutually constitutive and in tension. Yet while some have sought to dissolve the boundaries between work and home and leisure or life’s work as the new “contemporary habitus ... of permanently mobilized bodies in new kinds of technologies of power” (Mitchell, Marston, and Katz 2003, 429), others have questioned to what extent this nexus applies beyond the world of the professional middle classes and to the working class and immigrants with precarious lives.

Social reproduction involves a range of activities for its maintenance and reworking. Two key sites, where the needs of capital accumulation and social reproduction come together, are the family and the state (Bakker and Silvey 2008). While marketization, privatization, and informalization have strongly characterized trends in contemporary social reproduction, the state sets the parameters within which it occurs and may support or offload the burdens and costs of reproduction onto the family. Through its family and social policies, the state may sustain the social reproduction of the middle classes and the deserving working poor while withdrawing support from the increasingly demonized undeserving poor (unemployed, disabled, homeless, migrants), who are forced to turn to charitable provision such as food banks. Thus the crisis of social reproduction, although more acute in the Global South, is also affecting large numbers in the Global North who face inadequate access to basic necessities such as food, heating, and shelter. In many African countries the crisis is further exacerbated by the prevalence of HIV/AIDS (human immunodeficiency virus/acquired immune deficiency syndrome), high rates of mortality among younger populations, and orphaned children, which place even more demands on women’s time and care-giving. The crisis in reproduction has been defined as a critical gap between outflows – domestic, affective and reproductive – and the inflows that sustain health and wellbeing. It is suggested that resources required and their absence could be measured using qualitative and quantitative methods, such as time-use surveys, which would provide the tools for recognizing the value and costs of domestic work, as has been recognized recently in Venezuela, where mothers are now eligible for pensions.

Of course, individuals and households deploy different means to counter depletion. In order to ensure the social reproduction of their households, women are working longer hours and individuals and families are migrating, both within countries and internationally. Hence, social reproduction is also undertaken across borders, resulting in the international transfer of physical and emotional labor. Social reproduction offers us a lens through which to examine the ways in which inequalities between women and men, classes, and racialized populations, are reproduced, intensified, and transformed. To these social divisions we need to add legal status, which stratifies the rights between citizens and noncitizens to social reproduction through
conditionalities regulating transfer payments and access to provision, such as childcare, education, and health.

Over the past 15 years, studies of social reproduction have sought to bring together the variety of everyday activities and resources required to sustain individuals, households, and communities and policies promoted by states and international organizations. However, this endeavor also demonstrates the difficulty of holding the tension between an emphasis on governance and structural processes, on the one hand, and on the other an engagement with struggles over social reproduction in terms of class, gender, race, and citizenship at the everyday level in communities, within states, and across borders.

SEE ALSO: Agroforestry; Care work; Globalization; Inequality; Labor geography; Marxist geography; Migrant labor; Neoliberalism

References


Resampling, raster

A-Xing Zhu
University of Wisconsin, USA

Resampling is a process of transferring an existing raster dataset to a new raster dataset of the same data type by computing a new pixel value from the values of the corresponding pixels in the original dataset. There are two typical scenarios for which resampling of a raster is required (Tobler 1988). The first one is the mismatch of pixel resolution between two raster datasets that are to be overlaid (map algebra) for geographic analysis (Figure 1). In Figure 1, raster A has a higher spatial resolution than raster B. Direct overlay of these two raster datasets is not possible because each pixel in raster B corresponds to many pixels in raster A. Raster A must be first resampled to the same resolution as that of raster B. Furthermore, resolution mismatch often results in scale incompatibility, which leads to modeling bias. Thus, incompatible datasets need to be made compatible, and one of the techniques to do so is resampling (Zhu 2008).

The second scenario is when a raster dataset is converted into a different coordinate system (Figure 2), which often occurs in remote sensing image processing (Lillesand, Kiefer, and Chipman 2007; Richards 2012). Under this scenario, after coordinate transformation (registration) the pixels in the transformed raster no longer have the “squared” shape of the regular raster (Figure 2b). This is problematic for the raster data model because under this model the shape of a pixel is implicit, which means that the coordinates of the pixel shape are not explicitly stored and the shape is assumed around the center of the pixel for a given resolution. If we simply assume the pixel shape from the center, then the area represented by this pixel is no longer the same as it was before (see the insert in Figure 2). In addition, the center of the pixel in the new raster may not exactly match the center of the pixel in the original raster for a number of reasons, including different pixel size and different starting coordinates of the new raster (Tobler 1988). As a result, the grid of the new raster and the grid of the transformed raster may line up in the way shown in Figure 2c. So, to populate the new raster dataset, one needs to perform resampling.

There are several methods for determining the pixel values in the new raster from those in the original raster, including the nearest neighbor, majority, and averaging methods, and bilinear and cubic convolution interpolation. Nearest neighbor method selects the value of the pixel in the original raster whose center is nearest to the center of the pixel in the new raster (Figure 3). The nearest neighbor approach does not change the values of pixels from the original raster, thus the pixels in the new raster always have the same values in the original raster. For example, the value of 10 in the original raster will always be the value of 10 in the new raster. It will never be 10.5 or 11. Due to the fact that the values of the pixels in the
RESAMPLING, RASTER

Figure 1  Resampling to correct the resolution mismatch between two raster datasets for overlay.

new raster remain the same, the nearest neighbor approach is a good choice in resampling for nominal or ordinal data, where each value represents a class, member, or classification (categorical data such as a land-use, soil, or forest type).

The majority method takes the modal value (the most frequently occurring value) from the original raster within the neighborhood defined by the center of the pixel in the new raster. This method is used for resampling of categorical data (such as land-use, soil, or forest type). It is often used when resampling a fine resolution raster to a coarser raster, in which the neighborhood is often defined to be the pixel size of the coarser raster (Figure 4). The averaging method is very similar to the majority method except that it takes an average of the pixel values from the original raster in the neighborhood and it is used for resampling of interval or ratio data types (such as temperature, rainfall). The values in the new raster will be different from those in the original raster due to averaging.

Bilinear interpolation uses the values of the four pixels from the original raster whose centers are nearest to the center of the pixel in the new raster to determine the value of the pixel in the new raster (Figure 5). The new value for the pixel in the new raster is a weighted average of these four values, adjusted to account for their distance from the center of the pixel in the new raster. Due to the fact that the value for the pixel in the new raster is a weighted average of the original values, bilinear interpolation produces a value which does not necessarily
exist in the original raster. For example, the original raster may just contain all integers, but with bilinear interpolation the new raster could contain values with decimal points such as 10.2. Since the values for the pixels in the new raster are calculated based on the relative position and the values of the pixels in the original raster, bilinear interpolation is preferred for data that are continuous in nature, such as elevation, slope gradient, pollution level, and soil organic matter content. *Cubic convolution interpolation* is similar to bilinear interpolation except that the weighted average is calculated from the 16 nearest pixel centers and their values (Figure 6).

In summary, the values produced from resampling are an approximation, even if the values in the original raster are observations, because resampling is essentially a spatial interpolation process, which estimates the new values from the values in the original raster. In this sense, raster resampling often results in loss of accuracy.

**SEE ALSO:** Data structure, raster; Map algebra; Spatial resolution
Figure 3  The nearest neighbor method. The pixel value in the new raster (the red grid) is the value of the pixel in the original raster whose center is nearest to the center of the pixel in the new raster.

Figure 4  The majority method. The pixel value in the new raster (the red grid) is the value that occurs most frequently in the original raster within the neighborhood as defined by the coarser pixel.

Figure 5  Bilinear interpolation. The pixel value in the new raster (the red grid) is a weighted average of the values of the four pixels from the original raster whose centers are nearest to the center of the pixel in the new raster.

Figure 6  Cubic convolution interpolation. The pixel value in the new raster (the red grid) is a weighted average of the values of the 16 pixels from the original raster whose centers are nearest to the center of the pixel in the new raster.
References


Residential mobility

Si-ming Li
Hong Kong Baptist University, China

Residential mobility encompasses residential moves of all kinds; however, given that longer distance moves are usually referred to as migration, the study of residential mobility is often restricted to moves within a metropolitan area. Ever since the publication of Peter Rossi’s landmark text *Why Families Move*, the proposition that residential mobility is a spatial process through which households adjust housing consumption to changing needs and circumstances has caught the attention of generations of geographers, sociologists, and economists alike (Clark and Dieleman 1996). The main hypothesis is that moving is costly in both monetary and psychological terms, and a move to restore the equation between housing needs and consumption will be considered only if the stress that the household has to endure by remaining in the same residence is beyond a certain triggering point. But whether a move will actually be conducted depends on whether an appropriate vacancy is located after a more or less exhaustive search. The notion of family life cycle, which is closely related to the housing needs and aspirations of a household and which correlates with the householder’s position in the job ladder, assumes particular importance. In the United States and other developed countries, it has often been argued that households move from renting an apartment in the central city to owning a suburban home in anticipation of the arrival of offspring to the family. From then onwards, the propensity to undertake further moves declines significantly. However, with children moving out to form their own households and with declining physical strength of both spouses, a move back to apartment living is not unusual for older families (Clark and Dieleman 1996).

Over the past several decades, in Western societies in particular, the family institution has been subject to major challenges. Unregistered and shorter term partnership is now an accepted part of life, and divorces and remarriages are becoming frequent. Such major life events are almost invariably accompanied by residential moves. Correspondingly, recent studies on residential mobility focus on the life course of individuals rather than households and analyze moves in association with complex partnership trajectories as well as the timing and sequencing of careers, such as education and jobs that parallel household and housing careers. The analysis of parallel careers often incorporates interregional and international moves and blurs the distinction between intra-urban and longer distance migration (Michielin and Mulder 2008). Perhaps paradoxically, relatively fluid partnership could have resulted in greater reliance on the extended family for personal support, be it financial, psychological, or social, especially at time of crisis. This in turn has major implications for residential decisions and change. Indeed, residential decisions are highly contextual, contingent upon the myriad of events that bear on an individual’s life course, including developments external to the individual and the family. For instance, in the United Kingdom, World War II delayed new family formation and hence home purchase. On the other hand, the postwar baby boom fueled the demand for suburban homes, resulting
RESIDENTIAL MOBILITY

in accelerated suburbanization (Michielin and Mulder 2008).

From an economic perspective, barriers to residential mobility are tantamount to barriers to attaining Pareto optimality, or the state that the utility of an individual cannot be improved without causing at least one other individual to be worse off. The obvious policy implication then is to find ways to reduce the costs involved in the housing transaction, including the cost of house search. Of particular importance is the reduction of the information asymmetry between sellers and buyers by providing more adequate and readily available and accurate information. In the United States, where housing access is largely market based, residential moves take place quite frequently. An oft-quoted figure is that one fifth of households move every year, although the residential mobility rate appears to be declining. The high mobility rate is often seen by economists as an indicator of market efficiency, although this does not mean that constant mobility is desirable for individuals, families, or the communities to which they contribute.

Until recently, in many European countries a high proportion of households have resided in social housing (Clark and Dieleman 1996). For such households, policies and regulations governing access to and management of social housing to a significant extent delineate the housing opportunity set. Gatekeepers also play a key role in structuring residential moves. In Britain in the 1980s, under the “right-to-buy” policy of Margaret Thatcher local governments sold the bulk of council housing to sitting tenants at heavy discounts, a major purpose of which was to free up mobility constraints. Yet, homeownership inhibits mobility. Also, because of the concentration of young, unemployed, and single-parent households, the mobility rate in the residual social rental sector is likely to be higher than in privatized council housing.

Massive transfers of ownership of public-sector housing to sitting tenants have also taken place in China and other former centrally planned economies. In the Chinese case this has been accompanied by unprecedented urban expansion and proliferation of sprawling suburban housing estates. Simultaneously, there have been wholesale redevelopments of inner-city neighborhoods, resulting in forced relocation particularly of older families to remote and unfamiliar suburban locations. A switch to market-based housing provision led to an increase in the residential mobility rate in Chinese cities in the late 1990s and early 2000s (Li 2005). However, the attainment of homeownership by large proportions of urban dwellers would have dampened further mobility. Residential decisions depend on prior migration experiences. Over the past two to three decades in China hundreds of millions of rural migrants have flocked to the major metropolises. Yet, the hukou or household registration system denies migrants permanent residential status in the city of current domicile; in many respects they are forever transients and subject to discrimination of all kinds, particularly in the job and housing markets. Split families are common – children are often left behind in the home village to be taken care of by the grandparents while the parents are in constant search for job and shelter in the host city. The residential mobility rate among the migrant population is high. However, this high mobility is unlikely to signify their capability to respond proactively to changing spatial opportunities over the urban area.

Clearly, high residential mobility might have nothing to do with economic efficiency. Furthermore, recent developments, more specifically the global financial crisis triggered by the collapse of the subprime mortgage market in the United States in 2007 – 2008, call into question
the unfettered faith in the market underscoring neoliberal reforms. Deregulation of the financial sector and the consequential financialization of the global economy have substantially amplified economic and housing market cycles. In what has been called the Great Recession, banking giants such as Citi were saved only with the heavy injection of taxpayer money. Individual households whose wealth evaporated overnight were not so fortunate, however. Millions of American households had their homes foreclosed and were forced to relocate. Others who would like to move to a more affordable residence because of unemployment or declining incomes were spatially trapped because the proceeds from selling the existing home could not cover the outstanding mortgage.

Within an urban area, residential location and relocation reflect differential spatial growth, especially between the central city and suburbs and sociospatial differentiation over the metropolitan area. In the United States and other countries characterized by fragmented metropolitan governance and local governments’ dependence on the property tax as the major source of income, the out-migration of high-income families to the suburbs and the consequential filtering down of central city homes and neighborhoods was a major cause of the fiscal crisis of the state, particularly the local state, in the 1970s. The central city–suburb schism is closely related to the notion of spatial mismatch between housing and labor market outcomes. The extent of spatial mismatch is particularly acute in US cities where African Americans are trapped in deteriorating inner-city neighborhoods amidst rapid suburbanization of employment. Ironically, the “flight to the suburb” by whites also provides opportunities for expropriations of rent gap, that is, the difference in rent under the current dilapidated conditions and potential rent upon refurbishment, by real estate developers through the gentrification of selected inner-city neighborhoods into citadels of high-income professionals, who are attracted to the city’s entertainment and cultural amenities.

Residential mobility is also related to neighborhood stability and change in some other ways. For instance, within a given neighborhood a high mobility rate is indicative of the lack of neighborhood satisfaction as well as community cohesiveness and attachment (Clark and Ledwith 2006). Lively engagement among long-time residents is often praised as a way of life that policymakers should try to preserve. Urban renewal projects in the United States in the 1960s and 1970s and more recently in China displaced hundreds of thousands of households from inner-city neighborhoods; invariably, this resulted in the severance of long-established ties and instantaneous evaporation of valuable social capital and aggravated the plight of the socially deprived. However, the absence of residential mobility does not necessarily imply health for the community. Instead, it could as well mean neighborhood aging and declining vitality. The decline and even absence of young families in a neighborhood means that schools and playgrounds are rendered redundant. At the same time, the migration of young families to outlying districts where good schools are lacking might necessitate unpleasant long-distance commutes for young pupils. Moreover, the inability to move out of crime-stricken poor inner-city neighborhoods contributes to the development of a culture of poverty and hence the persistence of inequality across generations. Experiments have been carried out in the United States to provide opportunities to low-income black families to move to higher status suburbs in order to generate gains in social integration, employment, and earnings for more marginalized populations (Rosenbaum, Reynolds, and Deluca 2002).
RESIDENTIAL MOBILITY

Echoing the experimental findings is evidence of increasing fluidity in residential patterning among minorities in American cities. Of course, these social engineering programs carry risks and ethical questions about the choices and agency of participants; those involved could also confront immense challenges, not least the continuing prejudice against ethnic minorities in new neighborhoods.

The study of residential mobility has occupied an important place in urban geography and cognate social science disciplines. Much of the literature is related to the matching of households with divergent needs and wants to housing of various kinds and in different locations within a metropolitan area. The recent adoption of the life-course approach has opened up a whole new front on residential mobility research, which contextualizes residential moves to events not only pertaining specifically to the individual and his or her family but also to developments that have taken place in the broader geographical and historical setting. The life-course approach to some extent has blurred the distinction between the study of intra-urban and longer distance migration.

SEE ALSO: Housing; Life course; Migration: internal; Neighborhood; Residential segregation; Urban geography; Urban redevelopment

References


Further reading

Residential segregation

Steven R. Holloway
University of Georgia, USA

Residential segregation occurs when people who identify, or are identified, with different social groups live in different sections of cities. Segregation characterizes contemporary cities and has a long historical precedent. Ancient Sumerian city-states were structured by a rigid social hierarchy reflected in distinct social orderings of urban space. The hierarchy's highest tiers occupied the city's central districts, often associated with the seat of military-religious power. Cities in Medieval Europe were socially organized on the basis of craft – producers of similar goods lived in the same areas of the city. Cities of early modern Europe pioneered another innovation in residential segregation – the ethno-religious ghetto. Jews were needed in Venice, Italy, by Christian merchants to handle financing (i.e., money-lending) prohibited by the Catholic Church, and in 1516 were forced to live in a city district, surrounded by canals, that had formerly been a foundry. The practice of forcing Jews to live in their own districts, often surrounded by walls with gates locked at night, spread throughout many other European cities where Jews lived in significant numbers, and persisted well into the middle of the twentieth century.

Urban residential segregation should be understood as a verb – a set of actions perpetrated on one identified group of people by others who possess more power (e.g., Christian Venetians segregated Jews into the ghetto; African Americans were segregated into racial ghettos of industrial cities; South Africans were segregated into apartheid settlements). The sociospatial patterning of the city that resulted highlights segregation as a noun. Throughout urban history, segregation has existed as both verb and noun – as a power-laden process and as a map-able pattern.

Global perspectives on residential segregation

Carl Nightingale (2012, 4–5) powerfully examines the techniques of segregation that were developed across time and space to support colonial European efforts in which he includes the United States and Canada. He “asks us … to consider not only the idea that residential color lines have proliferated across the world but also … that such movements to segregate cities spread because they were interconnected.” Those responsible for creating and maintaining urban racial residential segregation were successful because they developed savvy institutional connections in a wide range of governments and in the modern capitalist real estate industry. Critically, they were able to spread ideas and tactics through global networks of intellectual exchange that included books, lecture circuits, international conferences, and a suite of professional journals, newspapers, and popular magazines. The key insight is that racial residential segregation over the last several centuries is global in scope and yet reflects the specificities of each colonial society in which it became embedded.

We have seen over the last several decades a renewed period of economic globalization that affects residential segregation. General trends that characterize globalization include
interconnected financial markets that circulate at ever-increasing speeds, downscaled governance necessitated by federal devolution, deepening social polarization marked by increasing economic inequality, and increased competitiveness efforts between cities as they struggle to survive. The spatial imprint of globalization on cities is seen in the emergence of new citadels that cater to global financial elites and the development of exclusionary enclaves, often associated with invigorated gentrification. These features are offset by growing immigrant enclaves, old and new (often now found in the suburbs), transformed working-class neighborhoods devastated by the global financial crises, and new concentrations of poverty in older suburbs. Many of these spaces are segregated in new ways, including most prominently by class. Ironically, competitiveness rhetoric fetishizes diversity in urban spaces, using terms like “social mix” and encouraging of alternative sexuality (visible gay and lesbian) presences. Creative city proponents have long argued that cities can improve their competitiveness in the global arena by remaking themselves to be more attractive to folks who work in knowledge- and ideas-based industries. There is now concern that such creative class populations are increasingly living in segregated neighborhoods. Concerns are also growing about increased levels of economic segregation, fueled by both the isolation of the poor and wealthy from other segments of society.

Racial segregation in the United States

US cities, especially those in the industrial heartland, became extremely segregated by race in the middle decades of the twentieth century. While the causes are complex and dependent on conditions in the specific city, several themes recur. First, the great migration of blacks from the South brought millions of migrants, in pre- and post-World War II phases, to northern and Midwestern cities, mostly in inner-city neighborhoods that became almost completely segregated by race.

Second, segregation was imposed upon African American migrants as they faced extreme and often violent forms of systemic discrimination, including their explicit exclusion from residential areas where whites lived. The means of discrimination were both direct and indirect and simultaneously personal and institutional. Private actions, discriminatory market processes, and explicit governmental policy produced segregation. The result was extreme: whites and blacks lived completely separate lives. Many cities where black migrants were most segregated had also housed innumerable European immigrants in the decades just before and just after the turn of twentieth century. While immigrants also faced discrimination and lived in segregated neighborhoods, the magnitude of segregation experienced by black migrants far exceeded that experienced by European immigrants.

Third, because of multilayered discrimination experienced in labor and housing markets, many black residents of segregated neighborhoods were also poor – African American neighborhoods like Chicago’s Bronzeville, Cleveland’s Fairfax, New York’s Harlem, and Atlanta’s Sweet Auburn housed great concentrations of poor migrants. Later, after the ravages of early post-World War II urban renewal and highway building programs, racial ghettos were often spatially anchored by isolated and dense public housing. Neighborhoods proximate to the older black ghettos transitioned amidst great interracial conflict and anti-black violence because the black population of the ghettos swelled during the second wave of the great migration. White flight describes the process whereby most or all white residents would move en masse
away from neighborhoods when black residents would move in. Whites seeking to leave were greatly assisted by energized suburbanization and the housing construction boom of the 1950s and 1960s. Newly developed areas used a wide array of tactics to stem the expansion of the ghettos and maintain white segregation of the suburbs.

Civil rights struggles in the 1950s and 1960s altered the nature of discrimination faced by African Americans in cities across the United States. Some of the most explicit tools used by financial institutions, city leaders, and governments at multiple levels were either ruled illegal or rendered so by new laws that purported to guarantee the housing and economic rights of African Americans. By the end of the 1970s, black Americans began moving out of institutionalized ghettos and into nearby and suburban communities. Racial panic followed and processes of white flight increased in intensity.

Causes of segregation in the United States

Vigorous debate swirls around the causes of residential segregation, especially after the enactment of civil rights laws in the 1960s and 1970s. To what degree does racial residential segregation remain a cause for substantial social concern? If people simply live where they feel most comfortable and can afford, then should it matter if different groups choose to live in different parts of the city? Is not this natural? It is useful to break this down into facets of the debate.

First, housing is very expensive and costs vary widely across neighborhoods. Groups whose members have fewer economic resources will, on the average, concentrate in less expensive neighborhoods. While income and wealth differentials account for some portion of observed patterns, racial segregation levels exceed what we would expect based on economic differentials alone.

Second, the question of choice partly depends on the myth that housing markets function like abstract economic models of efficiency and perfect information. If housing markets function as predicted by normative economic theory, then residential segregation simply reflects rational choices made by informed housing consumers. This notion avers that all groups, majority and minority, prefer to live in neighborhoods with other members of their group. A neighborhood’s population can shift rapidly from majority to minority once the minority share of the population exceeds a “tipping point,” somewhere between 10 and 20%. This shift represents only the preferences of current and potential residents for an acceptable neighborhood racial composition. If the minority share gets “too high” for the majority group, current residents will move out and potential residents will select different neighborhoods.

Research demonstrates that while most social groups show some own-group neighborhood preference, only the preferences of the majority matter. The Multi-City Study of Urban Inequality (MCSUI) demonstrates that blacks typically prefer racially integrated neighborhoods, whereas whites typically prefer neighborhoods with no more than a trivial share of minorities. Thus, racial residential segregation can result only from the preferences and choices of white residents, even when minority groups do not prefer segregated outcomes.

Third, questions about discrimination and racism remain central to causal debates. The historical literature clearly shows that a wide range of discriminatory actions produced deep racial residential segregation: explicit and implicit,
RESIDENTIAL SEGREGATION

private and public, individual and collective. While few scholars would argue that segregation was not produced and imposed, historically, by racism and discrimination, some argue discriminatory processes no longer characterize contemporary cities. To what extent does discrimination continue to exist and, if present, to what extent does it produce residential segregation? Several studies demonstrate that minorities more frequently experience disparate treatment in housing market transactions. These studies use a carefully matched set of potential renters and homebuyers who respond to advertised housing offerings. While the magnitude of disparate treatment has declined since the 1970s, black and Hispanic testers are treated less favorably than white testers.

Minority consumers accuse financial institutions of ongoing unfair treatment. Until the late 1960s, mortgage-lending institutions openly practiced the spatial discrimination known as redlining, where they would either refuse to issue loans to minority borrowers for homes in minority neighborhoods or they would do so with extremely harsh financial terms. Even after key civil rights legislation outlawed these practices in the late 1960s and 1970s, lenders systematically deprived minority borrowers and minority neighborhoods of capital. In an intensely ironic twist, lenders began to predatorily target minority borrowers and their neighborhoods with new, evermore complex loan products designed to be sold quickly on the secondary mortgage market. These loans were designed, in part, to circulate quickly in global financial markets without regard to the fate of borrowers and their neighborhoods. Such predatory lending ramped up through the late 1990s and middle 2000s. When the speculative housing price bubble of the early 2000s burst and the global economy stalled, housing values plummeted and borrowers saddled with unrealistic loans defaulted in large numbers. The effect was to strip equity and wealth from vulnerable communities. Mortgage defaults both contributed to and resulted from the globalized financial crisis. Predatory lending was both made possible and intensified by racial residential segregation. The mortgage default crisis was concentrated in and had its most pernicious effects in segregated minority neighborhoods.

Impacts of segregation

Residential segregation creates a variety of impacts. Groups that voluntarily segregate do so for perceived benefits. For example, wealthy whites segregate because exclusivity is perceived to maintain property values and social homogeneity. Immigrants who congregate in gateway neighborhoods might benefit from co-ethnic social institutions, cultural amenities, and economic opportunities.

When segregation is something imposed, however, impacts are more detrimental. During the time of Jim Crow laws, residential segregation was a key feature of the broad set of social exclusions imposed on African Americans that also included churches, schools, parks, swimming pools, public transit, and almost every other aspect of life. These exclusions were designed to preserve racial and class privileges enjoyed by the majority. Even after Jim Crow was challenged and partially dismantled in response to civil rights activism, exclusions proved to be deeply persistent and black residents of segregated neighborhoods understood residential segregation to be central to their experiences. In response to the riots experienced by several large US cities in the late 1960s, prominent commissions were established to explore the causes of the violence. President Johnson appointed the Kerner Commission
in 1967, and California Governor Pat Brown appointed the McConne Commission in 1965 after the Watts rebellion. Both commissions reported that residential segregation, enforced through widespread discrimination, was a key feature of the multifaceted social and economic disadvantages that produced deep anger and frustration among residents. Violence was often spurred by racist policing, which saw echoes in the Rodney King riots in 1992 and the Ferguson and Baltimore riots of 2015.

Economist Yinger (1997) computed the financial and economic cost to those who experience housing market discrimination – a discrimination tax. Because discriminatory behavior is found at every stage of the housing search process, and because it reduces information about potential housing relocation choices and diminishes choice, minority housing consumers endure higher transaction costs, higher housing prices, higher interest rates, and less favorable loan terms. He estimates that discrimination might cost an average minority consumer almost $8000 (in 2015 dollars) and might cost society more than $12 billion (in 2015 dollars).

Sociologists Massey and Denton (1989) argue that residential segregation remains the structural lynchpin of racial oppression, provocatively titling their 1993 book *American Apartheid*. They contend that desperate social conditions experienced in impoverished central-city neighborhoods are created and actively maintained by ongoing racial discrimination and residential segregation. Scholars examine what impact living in deeply disadvantaged and segregated neighborhoods has, especially for kids. These interests converged with concerns that high-density public housing was not working. Scholars argue that disadvantaged neighborhoods adversely affect residents: (i) by disrupting socialization and mechanisms of informal social control; (ii) by limiting the development of effective social networks; (iii) by stigmatizing residents, especially in the labor market; (iv) because institutional resources are severely limited; and (v) because they are spatially inaccessible.

Sampson (2012) capped several decades of intense Chicago-based empirical research with a book in which he demonstrates that Chicago’s disadvantaged neighborhoods themselves are persistent and that they impart a number of adverse impacts on residents. Sharkey (2013) capitalized on unique data that allowed him to track people over time and argued that the adverse impacts of growing up in disadvantaged neighborhoods last over the lifetime. A major policy experiment called moving to opportunity (MTO) initially provided disappointing results (Briggs, Popkin, and Goering 2010). The experiment allowed some residents of disadvantaged public housing projects to move into better neighborhoods. Even though the first major assessments of this experiment did not confirm the idea that changing neighborhoods improves life chances, a major recent reassessment of this project reaches a more optimistic conclusion (Chetty, Hendren, and Katz 2015). Considerable benefits accrued to the children who relocated out of the worst of the public housing projects into better neighborhoods, even though these benefits did not appear in the short term.

Measuring segregation

Quantitative measures of segregation were developed soon after World War II as racialized segregation increased dramatically. The index of dissimilarity (equation 1), the most commonly used index, measures the degree to which two groups’ members are unevenly distributed across a set of spatial units.
where $X$ and $Y$ are the total number in each group in the city in which there are $J$ spatial units, often census tracts. The number in each tract is given by $x_j$ and $y_j$. The index ranges from 0 to 1 and represents the share of one group that would have to relocate to different tracts to equalize the groups’ distribution across tracts. The index of dissimilarity remains the most widely used measure of residential segregation, despite deep critique, because of its computational simplicity and ease of interpretation.

The second most common measure of residential segregation is based on a suite of exposure/isolation measures, known as “$P$-star” indices (equation 2).

where $y_j$, and $X$ are defined as before and $t_j$ is the total number of residents in each tract. The indices range from 0 to 1 and have several useful characteristics. First, they convey the average neighborhood racial composition for a group, or, equivalently, the chance that a group $X$ member can encounter a member of group $Y$ in their neighborhood. Second, whereas the index of dissimilarity is symmetric (meaning that group $X$’s dissimilarity from group $Y$ is the same as group $Y$’s dissimilarity from group $X$), the exposure indices are asymmetric and reflect each group’s share of the overall population. We can thus observe the degree to which two groups share neighborhood space. For example, if the neighborhood residential exposure of group $X$ to group $Y$ is 0.67, we can say that two-thirds of the people encountered in the neighborhood of the average group $X$ member would be from group $Y$. This will reflect both the overall racial composition of the city’s population and the degree to which the groups are segregated among its constituent spatial units. Conversely, the neighborhood residential exposure of group $Y$ members to group $X$ will not be 0.67 – it will be much lower.

We can compute a suite of $P$-star exposure values that capture the entire range of groups present in the city, including the neighborhood residential exposure of a group to other members of the same group.

The result tells us, on average, what share of people encountered in the average neighborhood are from the same group – in other words, how residually isolated are members of the group. The isolation version of the index is most commonly used in comparison with the index of dissimilarity because each of these indices tells something a little different about the degree of segregation between two groups.

Residential segregation has five primary measurable dimensions. In addition to evenness (most commonly measured with the index of dissimilarity) and exposure/isolation (most commonly measured with the $P$-star indices), centralization captures the degree to which a group lives in a city’s inner-most neighborhoods. Clustering describes the degree to which members of a group live in neighborhoods that are proximate to one another as opposed to scattered around the city, and concentration describes the degree to which a group lives in a minimal amount of urban space. A group that simultaneously experiences high levels on most or all dimensions of segregation is hypersegregated. Only African Americans are subjected to hypersegregation in significant numbers even though Hispanic segregation recently increased to the level of hypersegregation in a few metropolitan areas.
Massey and Tannen (2015) found in the most recent update of this work that while the number of hypersegregated metropolitan areas declined by half from 40 to 21 since 1970, the areas that remain hypersegregated are among the nation’s largest and house more than one-third of all metropolitan blacks.

Recent trends in residential segregation

Economists Glaeser and Vigdor (2012) argued in their provocatively titled report “The End of the Segregated Century” that racial residential segregation levels have declined enough over recent decades that we should end our concern with segregation as a social problem and de-emphasize its presumed links with broader racial inequality. Figure 1 shows the basic trend that they reported. Black/non-black segregation levels have indeed declined since their peak in 1970 as measured by both dissimilarity and isolation indices. They interpret the decline as a result of the ending of government sanctioned discrimination, which allowed easier access to housing finance capital. Two migration patterns resulted. One, blacks followed broad regional migration patterns from older Rust Belt cities to newer Sun Belt cities where segregation is generally lower. Two, blacks moved from inner-city ghettos to suburbs. They also argue that gentrification and immigration contributed to declining segregation and note that racial economic inequality has not declined along with segregation.

Glaeser and Vigdor’s optimistic interpretation has been challenged by prominent segregation scholars. John Logan (2013; Logan and Stults 2011) also reports that black dissimilarity and isolation declined, though the decline was more modest in the metropolitan regions where blacks are more likely to live. Three reasons capture Logan’s less optimistic interpretation.

One, black—white segregation remains relatively high in absolute terms; it will take several more decades before it declines to a level comparable with other minority groups. Two, the old industrial cities of the manufacturing belt persistently have the highest levels of racial residential segregation for blacks. In these areas, whites rarely move into black neighborhoods and seem to be actively moving out of ethnically and racially diverse neighborhoods, which does not augur well for the future. Three, blacks and Hispanics still live in neighborhoods with much higher levels of social and economic disadvantage, even when controlling for household affluence; the most affluent black households live in neighborhoods more disadvantaged than the least affluent white households.

Table 1 shows average segregation levels using dissimilarity and isolation indices for blacks, Asians, and Hispanics for large metropolitan areas (over 1000000 in population by 2010). These data show declines in segregation as measured by both indices for blacks over the last two decades – even though the magnitude of decline seems to be a bit more modest than argued by Glaeser and Vigdor. Hispanic dissimilarity
RESIDENTIAL SEGREGATION

Table 1  Average segregation levels for large metropolitan areas. Computed from segregation values obtained from the US 2010 project, Spatial Structures in the Social Sciences, Brown University. © John Logan.

<table>
<thead>
<tr>
<th></th>
<th>Dissimilarity</th>
<th>Isolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>57.1</td>
<td>61.1</td>
</tr>
<tr>
<td>Hispanic</td>
<td>45.9</td>
<td>46.4</td>
</tr>
<tr>
<td>Asian</td>
<td>39.4</td>
<td>39.7</td>
</tr>
</tbody>
</table>

increased during the 1990s and Asian dissimilarity remained constant. Isolation increased for both groups, especially Hispanics, as their share of the population increased due in part to immigration.

The categories “Hispanic” and “Asian” each include too many ethnic and national-origin backgrounds to be deeply meaningful. Logan and Turner (2013) examined Hispanic segregation by origin group. The US Census asks respondents to indicate their “Hispanic, Latino, or Spanish” origin separately from race, and requires them to report their specific origin using terms that are interpreted as the nation–state from which they or their ancestors originated (e.g., “Chinese,” “Filipino,” and “Vietnamese,” among others). The Asian population has grown rapidly over the last two decades and is now as large as the Hispanic population was in 1990. While Asians are overall less segregated from whites than are Hispanics or blacks, Vietnamese are the most segregated Asian group and their dissimilarity index values are higher than Hispanics as a group and almost as high as blacks in 2010. Japanese are the least segregated Asian group, in part reflecting the fact that most are US born. Importantly, Asian residential segregation is not associated with socioeconomic disadvantage in the same way that it is for blacks and Hispanic groups – in fact, all of the Asian groups live in neighborhoods where the typical household income is higher than it is for whites, with the magnitude of this neighborhood advantage greatest in suburban areas.

International studies of residential segregation highlight both different processes and different patterns. European cities did not experience the large waves of immigration in the late nineteenth and early twentieth centuries that so strongly shaped US cities. Thus, concepts of race, region, and ethnicity vary across national context, as do data collection and classification schemes. In
RESIDENTIAL SEGREGATION

the late twentieth century and early twenty-first century, however, guest workers, labor migrants, ex-colonials, and refugees increasingly moved to European cities. As a result, many major European cities, including London, Amsterdam, Frankfurt, Brussels, Paris, Lisbon, and Stockholm (among many others), have neighborhoods perceived as homogeneously “ethnic.” In London, for example, there are neighborhoods associated with Pakistani-, Bangladeshi-, or Indian-background residents. In Paris, ethnic neighborhoods are often populated by North-African immigrants. Scholars of segregation in European cities typically use the terminology of “ethnicity” rather than “race” and generally agree that residential segregation broadly characterizes European cities. Levels of ethnic segregation in European cities are lower than black–white segregation in US cities, although comparative research does not reflect the continued decline of black–white segregation during the 2000s. Several UK cities, including Birmingham, Bristol, London, and Oldham (Manchester), show very high levels of residential segregation, comparable to black–white segregation in US industrial cities, for Bangladeshi- and Pakistani-background populations. Catney (2016) shows, however, that residential segregation is declining in UK cities for most non-white groups, including Bangladeshis (outside of inner London) and Pakistanis.

Economic inequality has deepened dramatically over the last several decades. Buffeted by neoliberal globalization, and increasingly abandoned and/or disciplined by governments keen on rolling back the social safety net, the economic fortunes of the working and middle classes have deteriorated. At the same time, the fortunes of the wealthiest segments of society have improved dramatically. Economic residential segregation has increased accordingly. Scholars and policymakers increasingly express concern with increasing levels and new forms of residential segregation by class. Using census data, Bischoff and Reardon (2014) demonstrate a broad increase in segregation by income from 1990 to the present. They find that both the poor and the wealthy have experienced increased residential segregation. The global financial scandal and subsequent economic dislocation hit hard in many working-class and lower middle-class neighborhoods, while at the same time decades of building highly exclusive, often gated communities remained in place.

The space, scale, and geography of segregation

Wong (1997, 2002) and others (e.g., Reardon and O’Sullivan 2004) demonstrate that standard segregation measures inadequately reflect the spatial arrangement of neighborhoods within the city. They have developed an array of spatially sensitive alternatives to the index of dissimilarity and the P-star indices. Some scholars advocate the use of spatial statistics such as Morgan’s I or Geary’s C. These approaches generally require the use of GIS software and are more labor intensive to implement broadly than are the traditional measures.

Scale constitutes another strain of geographically relevant segregation research. Research shows that groups may be more or less segregated at different scales of aggregation within the city. For example, the US Census Bureau defines block groups as substantially smaller (≈1200 people) than census tracts (≈4000 people). Segregation indices measured using block group-level data will generally be higher than the same indices measured using census tracts. Historical research that attempts to measure residential segregation in the early twentieth century or late nineteenth century might use election ward data because the US Census Bureau had
not yet started defining and releasing data for census tracts. Since election wards are larger than census tracts, measured levels of segregation will generally be lower than if segregation had been measured using spatial units the size of census tracts. This issue can be generalized in the sense that our understanding of the magnitude of residential segregation depends on the scale at which we view it. While some groups are more segregated in smaller aggregates than they are in larger aggregates, other groups will show different levels of segregation across levels of aggregation. Scale constitutes an ongoing area of contemporary segregation research of interest not just to geographers, but also scholars in sociology, demography, political science, urban planning, and other disciplines. Lee et al. (2008) and Reardon et al. (2008), for example, develop a detailed methodology that uses very small census block data (there are more than 11,000,000 blocks in the United States and Puerto Rico and there are an average of 39 blocks per block group), sophisticated spatial analysis, and GIS, which demonstrates that residential segregation for any given city varies across scales in what they call a segregation profile.

Recent research recognizes the increasing complexity of urban residential landscapes in the face of growing income inequality and new patterns of ethnic and racial diversity. While several multi-group measures of residential segregation have been proposed, scholars have not yet settled on a single approach. As an alternative to the increasingly complex array of indices, several scholars are using categorical approaches, either alone or in combination with single-value indices. Johnston, Poulsen, and Forrest (2003) developed a classificatory scheme, which they argue is globally generalizable, that centers on the dominance of white settler populations. Other neighborhood types are then identified based on an array of racial composition thresholds. Holloway, Wright, and Ellis (2012) developed a categorization system to facilitate effective cartographic representation (mixedmetro.com) based on a summary index of racial and ethnic diversity augmented by easily interpreted thresholds of compositional dominance. They explored and depicted the ways in which increasing diversity and persistent residential segregation enfold one another in complex and city-specific ways.

Social media provide a large source of data that can be utilized for segregation research and cartographic representation. Shelton, Poorthuis, and Zook (2015) used geotagged Twitter data to critically examine and map Louisville, Kentucky’s racial divide. They caution, however, that “big data” present many technical and conceptual challenges to which researchers need to attend.

**SEE ALSO:** Ghetto; Urban geography; Urban mosaic

**References**


Resilience and human geography

Geoff A. Wilson
Plymouth University, UK

The notion of “resilience” has begun to influence research by human geographers, and human geographers, in turn, have begun to shape thinking about the meaning and significance of resilience. This entry outlines the scientific genesis of the concept, and then focuses specifically on contributions to resilience research by social scientists and human geographers. The entry demonstrates that resilience – much like “sustainability” – remains a relatively ill-defined and elusive term, and that different disciplines and subdisciplines use the word in often different and contradictory ways. The entry concludes by suggesting that human geographers are particularly well placed to contribute to both further clarification and improved practical application of the term.

The genesis of resilience research

The notion of “resilience” (from the Latin “resilire,” meaning to leap back, rebound, or recoil) has undergone profound changes in meaning, scientific relevance, and practical applicability over several decades. From the beginning of the twentieth century, the first scientific reference to resilience occurred in engineering, where resilience of materials or buildings was understood as an ability to withstand stress. In the early 1970s the notion of resilience was incorporated into ecology, where seminal work by ecologist Chris Holling (1973) suggested that the term could be used as a scientifically robust concept to highlight how able an ecosystem was to cope with stress and respond to disturbances (ecological resilience), whereby the vitality of a system was not seen to be defined as the ability to reach a “climax” but rather in terms of evolutionary change. Resilience in this sense was seen as an emergent property – a relative attribute characterized by responses to disturbances which can only be assessed by looking at changes in a system over time. Ecologists suggested that qualities of resilience are evident in the notion of adaptive capacity, which is generally used to analyze how a system does, or does not, respond to endogenous and exogenous changes and is defined as the ability of a system to adjust to change, moderate the effects of that change, and cope with a disturbance. A system characterized by diversity, potential for change (level of redundancy in the system), and connectedness (feedbacks, flexibility) usually has better adaptive capacity.

Resilience research between the late 1970s and the 1990s began to focus on an extended resilience definition strongly influenced by the theory of complex adaptive systems, which argues that ecological resilience concepts can be applied to human systems under the umbrella of social-ecological resilience. Brand and Jax (2007) argued that this marked the beginning of the notion of resilience as a “boundary object” between the natural and social sciences. An important building block of this research was Gunderson and Holling’s (2002) seminal concept of “panarchy,” which took fast/slow resilience dynamics and cross-scale interactions and interdependencies into account, and provided a
heuristic model of four nested adaptive cycles of resilience emphasizing cross-scale interplay ranging from periods of exponential exchange, to periods of growing stasis and rigidity, periods of readjustment and collapse, and periods of reorganization and renewal. The Resilience Alliance based in Stockholm (Sweden) – an influential international consortium of 15 research groups and institutes from many disciplines who collaborate to explore the resilience of social-ecological systems (SESs) – has since expanded this work and has suggested several methods and approaches to analyze the resilience of interlinked social-ecological systems (Resilience Alliance 2007).

Although SES approaches implicitly implied that social drivers of change and adaptation were important, their ecological or “scientific” approach was increasingly criticized for neglecting key social science processes such as power, agency, politics, and highly complex stakeholder interpretations of what makes a system resilient. The result was a lively debate in resilience research during the 1990s and 2000s that laid the foundation for current, more critical, and social science-informed resilience thinking. Authors such as Adger (2000), Wilson (2012a), and Brown (2014) highlighted that resilience of human systems could not be explained through linear processes of change, that memory in social systems plays a different role from that played in natural systems, and that human behavior affecting resilience can be irrational, unpredictable, and highly dynamic. The original natural science-influenced definition saw resilience as the capacity of a system to absorb disturbance and reorganize while undergoing change and still retain essentially the same function, structure, identity, and feedbacks, whereby resilience is measured by the size of the displacement the system can tolerate and still return to a state where a given function can be maintained. This began to change into a more malleable definition that included notions of uncertainty and unpredictability for assessing the resilience of human systems. The approach used in the social-ecological resilience framework was increasingly criticized for relying too heavily on the idea that complex system behaviors in multivariate space will typically fall within a stability domain around an attractor – behaviors that may not necessarily be true for the resilience of human systems. As a result, many argued that “measuring” social-ecological resilience is challenging, particularly because in order to clarify features that contribute to it, institutional and organizational processes must be understood as carefully as ecological ones. As a result, since the late 1990s the notion of “social resilience” has gained importance, especially as it focuses attention on the resilience of human systems with their intricate social, economic, and political processes that shape the resilience or vulnerability of systems and communities. A particularly important strand of research has been the focus on “community resilience” with a strong geographical focus on human communities (villages, towns, urban areas) and their ability to cope with disturbances – areas where human geographers have been particularly prominent.

Human geography, social science, and resilience research

As with other contributions to the International Encyclopedia of Geography, identifying clearly defined human geography contributions to a specific field of enquiry is difficult. This is because definitions of who and what “human geography” entails are becoming increasingly blurred. As the notion of resilience straddles several disciplinary and subdisciplinary divides (see above), inevitably some of the most interesting
research on resilience is done at the intersection between disciplines by researchers who often do not identify themselves with specific disciplines. In addition, while many of the contributors to debates on resilience are based in geography departments, they may not necessarily have been trained as human geographers (e.g., many political ecologists, political scientists, historians, sociologists, etc.). In the following, therefore, the boundaries of what constitutes human geography are seen as highly permeable, and reference is also made to non-geographers (i.e., “social scientists” more generally) who have contributed to resilience research by emphasizing human geography-specific issues of space, place, and the complexity of human systems.

Geographers have played a pivotal role in shaping resilience research since the 1990s, with seminal contributions by Adger (2000) on “social resilience,” Martin (2012) on “regional economic resilience,” Wilson (2012a) on “community resilience,” a recent intervention by Brown (2014) on conceptual issues of resilience from a human geography perspective, and a critical study by Rigg and Oven (2015) on notions of “liberal resilience.” These and other studies by social scientists have particularly criticized notions of linearity and “measurable” resilience responses assumed to be equally present in both social and ecological systems. As Adger (2000, 347) highlighted, “it is not clear whether resilient ecosystems enable resilient communities.” This was echoed by Davidson (2010), who argued that the panarchy model by Gunderson and Holling (2002; see above) implies that in the absence of disturbances, systems will tend toward increasing complexity, but that several problems arise with the transfer of this model to social systems.

The application of the social-ecological resilience framework to social systems, therefore, requires improved articulation of the multiple relationships between complexity and disturbance in a less deterministic manner than is afforded by ecological systems. As a result “the means by which social systems can be readily cast into the ecological terms encompassed by the resilience framework must be critically examined” (Davidson 2010, 1141). Thus, Davidson further argued, while the structural complexity of both ecological and social systems can be conceived of in similar terms, the feedback processes associated with each are incomparable: social systems are unique in that the tendencies toward complexity, and the responses of individual organisms to those levels of complexity, are defined not solely by structural variables but, crucially, also by agency. In this sense, many social scientists and human geographers argue that the four nested adaptive cycles suggested in the deterministic ecosystems-oriented panarchy cycle may not occur in a linear fashion in human systems, but simultaneously with strongly overlapping and, at times, contradictory behaviors associated with less predictable human agency. Reviewing social science and human geography studies on resilience, Brown (2014, 109) thus concluded that “the transference of ideas about ecological systems to the social realm is viewed as highly problematic.”

The result of these criticisms has been an emergent third strand of resilience research focused on the resilience of human systems and communities, referred to as social resilience. This has been shaped by both social scientists and human geographers (e.g., Adger 2000; Brand and Jax 2007; Davidson 2010). The social resilience focus is based on actor-oriented approaches and on social, political, and economic processes that better reflect power and scale imbalances that shape the resilience of human systems. In this sense, social resilience is an important component of the circumstances under which individuals and social groups adapt to change.
The focus on *community resilience* emphasized by several human geography resilience researchers (e.g., Wilson 2014a, 2014b, 2015) is a specific subset of research on social resilience defined “as the ability of groups or communities to cope with external stresses and disturbances as a result of social, political and environmental change” (Adger 2000, 347). It therefore concerns the ability of a social system to respond to and recover from disasters, and includes those inherent conditions that allow the system to absorb impacts and cope with an event, as well as postevent, adaptive processes that facilitate the ability of the social system to reorganize, change, and learn in response to a threat. Social resilience is, thus, about the necessity of human systems to learn to manage by change, and implies that uncertainty and surprise are part of the game. Social/community resilience can be preventative (avoiding poor outcomes by developing coping strategies), or it may facilitate recovery after a traumatic event or catastrophe. It also considers whether the worst-off stakeholder group (i.e., economically or politically weak groups within a community) can recover from a disturbance without necessarily reducing the wellbeing of any other community-based institution or individuals.

Critics of the SES approach to resilience research have particularly highlighted that recovery and adaptive capacity in human systems take on very different pathways from those in ecological systems, with notions of social capital and memory, institutional capacity, power, agency, governance, path dependencies, and “lock-ins” particularly important – notions that suggest that returning to an “original”/“pre-disturbance” state is most often not the most appropriate recovery strategy in human systems. There is, therefore, a growing recognition that the persistence of certain social and economic conditions hampers positive long-term resilience, but too often empirical work is still interpreting resilience in the narrow sense of return time and recovery, thereby possibly missing the broader use of the concept. As a result, Brown (2014, 109) emphasized from a human geography perspective the “depoliticized” nature of most resilience research and “the failure to recognize resilience as socially contingent, rarely addressing the question of ‘resilience for whom?’” and “that resilience fails to take account of politics and power relations.” A focus on social resilience, thus, suggests that rather than viewing humans as outside or apart from ecosystems, humans should be viewed as agents of change acting within SESs. In other words, persistence, adaptability, and transformability take very different forms and pathways in ecological versus human systems, with the critical literature increasingly suggesting that human systems can never be *fully* resilient (Wilson 2013) but that *multifunctional* human systems have a tendency to be more resilient (Wilson 2010). Brand and Jax (2007, 12), thus, argued that “the original ecological dimension of resilience is about to vanish.”

The notion of social resilience, therefore, assesses how communities develop adaptive capacity to respond to endogenous disturbances (e.g., political upheaval, revolutions) or exogenous threats (e.g., climate change, volcanic eruptions, flooding). As a result, human geographers have highlighted the complexity of forces affecting communities, from both human and natural causes operating across space and time. Wilson (2012a), for example, highlighted that such forces are referred to in various ways in the literature, including “threats,” “stresses,” “shocks,” “perturbations,” “disasters,” “hazards,” “disruptions,” “disturbances,” and so on, whereby the relatively neutral notion of *disturbances* is beginning to gain ground based on the argument that the resilience of human systems is affected by both anthropogenic (human-induced) and natural disturbances (Figure 1).
Not all disturbances shown in Figure 1 are necessarily negative for community resilience. Disturbances associated with globalization processes are particularly open to debate (Wilson 2012b), and some processes, such as technological change, can of course also positively influence community resilience. The disturbances shown in Figure 1 can have both internal (i.e., within the community; e.g., local pollution) and external (i.e., outside the community; e.g., hurricanes, wars) causes, and include both sudden catastrophic disturbances (e.g., earthquakes) and slow-onset disturbances such as climate change or shifts in global trade. The complexity of disturbances affecting communities suggests that communities

Figure 1  Examples of anthropogenic and natural disturbances affecting community resilience. Source: Wilson 2012a. Reproduced with permission of Taylor & Francis.
RESILIENCE AND HUMAN GEOGRAPHY

are never “stable,” but that they are continuously and simultaneously affected by several disturbances at any point in time. It is argued, therefore, that communities can never reach “maximum” resilience levels but can only strive towards the highest level of resilience achievable within community-specific contexts.

Although understanding human–environment interactions continues to form a crucial component of these approaches, human geography research on social resilience is often based on a “bottom-up” approach predicated on understanding human drivers and indicators of resilience, of which human–environment interactions are only one of many components. There is, therefore, an increasing need to question resilience theorists for a lack of attention to power relations, politics, and culture. As a result, social resilience and human geography research focuses more on the importance of politics, power, and socioeconomic, psychological, and moral parameters than “traditional” social-ecological resilience research.

There are obvious differences between the more ecologically informed notions of social-ecological resilience and those attempting to understand social resilience related, in particular, to issues of linearity/nonlinearity. A key issue of nonlinearity, for example, relates to the assumption that human systems can never return to their original state after disturbance (which ecosystems may be able to do) due to social learning processes and social memory. In human systems, therefore, adaptations can be both anticipatory and reactive, and depending on their degree of spontaneity they can be ad hoc or planned. The capacity to cope with nonlinearities or other forms of surprise and uncertainty, therefore, requires an openness to learning, an acceptance of the inevitability of change, and the ability to treat interventions as experiments. Adaptation in the context of human dimensions, thus, refers to a process, action, or outcome in a system (household, community, group, sector, region, country) in order for the system to better cope with, manage, or adjust to disturbances. This notion of “better coping” highlights that social learning implies human adjustment processes that propel the post-disturbance system to a different (sometimes “better”) state. Human geographers argue that this has crucial implications for the “management” of community resilience, as the goal for adaptive capacity after a disturbance should often not be to attempt to reinstate the original state but to use learning processes and social memory as a basis for the creation of a qualitatively different, more resilient community (Adger 2000; Wilson 2012a; Brown 2014). Thus, while technological adaptation to a disturbance is usually seen as an easier process (e.g., engineering solutions towards sea level rise), human attitudinal and behavioral adaptation is much more complex and usually takes much longer (e.g., changing recognition that there is a “problem”).

Social resilience is, therefore, about preemptive change which sees resilience as a desirable state, rather than simply a process to avoid disturbances. This is exemplified by Davidson’s (2010, 1146) suggestion that “rather than directing our attention primarily to identifying and prescribing conditions of sustainability, the [social] resilience framework directs our attention to information flows and cycles of change, exploring how our current institutions and connecting structures are likely to respond to disturbance, and how we can prepare for those outcomes.” The notion of social resilience is, thus, strongly focused on human learning pathways and how these may affect the resilience of communities. Resilience in this view is both an outcome, especially when linked to improved adaptive capacity of communities, and a process linked to dynamic changes over time associated
with community learning and the willingness of communities to take responsibility and control of their development pathways. It is this importance of human agency which defines an additional conceptual layer not present in ecological systems, and consequently not reflected in ecological theories of resilience – that is, human agency distinguishes social systems from ecosystems. While the ecological resilience literature has focused more on the ability of systems to return to function after a disturbance, social resilience is, therefore, about seeing disturbances as an opportunity for change and development. The focus on social resilience, therefore, shifts attention from controlling change in presumably stable community systems to managing the capacity of dynamic communities to cope with, adapt to, and shape change.

In essence, the notion of social resilience is about understanding a “positive” quality of a community under investigation, not dissimilar to the notion of “strong sustainability.” As a result, understanding social resilience is more closely associated with normative judgments about “good resilience” and “bad vulnerability.” The close association of resilience with the ability of a human system to absorb impacts/disturbance and to reorganize into a fully functioning (but qualitatively different) system is, therefore, closely associated with the notion of “positive” quality, while community vulnerability – usually used to describe exposure and sensitivity of a human system not able to cope with risks, hazards, and slow or catastrophic change – is generally associated with “negative” quality. A broad concept of resilience, therefore, includes normative dimensions. In contrast to ecological systems where resilience can be seen as an ideal linear end-point, some human systems, such as “resilient malign dictatorships” (i.e., a resilient system that maintains a malign dictator), or “resilient malign ideologies” (where ideologies (e.g., Nazism) can be resilient to change), can be resilient in a “negative” context for the wellbeing of communities/societies (i.e., “bad resilience” = vulnerability). Researchers involved with social resilience research increasingly, therefore, avoid the use of terms such as “recovery” and prefer concepts of “renewal,” “regeneration,” and “reorganization” following disturbance.

The latter point highlights that, from a social resilience/human geography perspective, resilience/vulnerability are, therefore, intricately linked to notions of “good” or “bad” community development pathways. Although several concepts of the interlinkages between resilience and vulnerability have been used, many human geographers argue that resilience/vulnerability can be expressed as a simple spectrum (Figure 2) (Adger 2000; Wilson 2012a; Brown 2014). Extreme ends of a spectrum are usually most easily conceptualized. For example, few would contest that the complete disappearance of a community due to destruction of the livelihood base (e.g., depletion of soils leading to desertification), or a malign dictatorship that has negative implications for the wellbeing of most of its citizens, can be seen as a “bad” development pathway associated with extremely vulnerable communities. This highlights that, ultimately, normative moral judgments based on specific conceptualizations of social resilience are generally rooted in measures of human survival – for example, complete disappearance of a community.

Figure 2 Resilience and vulnerability as opposite ends of a spectrum. Source: Wilson 2012a. Reproduced with permission of Taylor & Francis.
community means no resilience. However, these normative assumptions become more blurred as one moves along the spectrum. As Figure 3 suggests, it is, therefore, more difficult to find a common understanding of “good” development pathways associated with resilient communities after a disturbance or “transitional rupture” in community development pathways. Indeed, any notion of quality (such as resilience/vulnerability) is relational and, therefore, always subjective – in other words, different individuals and stakeholder groups will view quality in different ways. Finding a commonly or cross-stakeholder and cross-culturally acceptable definition of a strongly resilient community is, therefore, almost impossible, although most would agree that the complete disappearance/destruction of a human community is clearly associated with extreme vulnerability (i.e., pathway c in Figure 3).

**Conclusions**

This entry has critically discussed the notion of “resilience” and how social scientists and human geographers have contributed to recent research on resilience. The discussion has shown that resilience remains a relatively ill-defined and elusive term, and that different disciplines and subdisciplines use the term in often different and contradictory ways. This confusion in definitions and concepts of resilience has been criticized by many commentators, including several human geographers (e.g., Adger 2000; Wilson 2012a; Brown 2014). Indeed, most human geographers would probably agree with both Miller et al.’s (2010, 6) critique that the “tendency for each [resilience] community to redefine other communities’ terms in their own language, invariably situating their own
RESILIENCE AND HUMAN GEOGRAPHY

concepts on a higher ground and the other’s as a derivative, may have impeded collaboration,” and with Brand and Jax’s (2007, 12) caution that “the meaning of resilience [becomes] increasingly vague and unspecified … [and] that both conceptual clarity and practical relevance of resilience are at stake.” This underlines, therefore, that philosophical, moral, and epistemological differences in natural-science- and social-science-led resilience research need to be acknowledged and that it is unlikely at this juncture in resilience research that an overarching framework for understanding “resilience” can be found in the near future. Resilience will, therefore, continue to be a Janus-faced “boundary object” (Brand and Jax 2007). Indeed, with inclusion of ever more subdisciplines into the broad arena of resilience research, further fragmentation is most likely.

This discussion has nonetheless highlighted that disciplines such as human geography are ideally situated to make major contributions to resilience research in future, due to human geography’s strategic location at the intersection between several existing tensions regarding resilience research, in particular by contributing towards “socializing” and “politicizing” resilience research (e.g., by highlighting the crucial importance of power and politics within complex actor spaces in resilience research; cf. Brown 2014); by promoting resilience as a normative term with “real” potential for informing policy-making and changing attitudes, action, and behavior on the ground (cf. Wilson 2012a); and by further clarifying conceptual differences between “sustainability” and “resilience” (cf. Adger 2000; Davidson 2010).

As Adger (2000), Wilson (2014a), and Brown (2014) highlighted, it is key to analyze the nature of agency and governance as important drivers for resilience issues in a politically nuanced way and through hybrid and mixed methodologies that also allow investigation of “deep” sociopolitical processes, and that acknowledge the non-linear and heterogeneous nature of resilience pathways. In particular, researchers embedded in the social-ecological resilience paradigm need to acknowledge more that human systems can never “bounce back” to a pre-disturbance state, as social memory and learning processes mean that new resilience pathways emerge that may build on previous pathways and memories that are often also qualitatively different and may include unanticipated “surprise adaptations.” Thus, “the challenge from a [social] resilience perspective is to learn to live with change and develop the capacity to deal with it instead of trying to block it out” (Miller et al. 2010, 5). The social resilience approach, therefore, also acknowledges that no human system can ever be fully resilient, and that some components of the system (e.g., a weak stakeholder group) will still be vulnerable. It is in these resilience research arenas, in particular, that fruitful opportunities emerge for human geographers, as they are able more than any other discipline to provide a critical bridge between discussions on how space and scale, governance, power, and complex socioeconomic processes affect the resilience of human systems. Nonetheless, synergies are beginning to emerge with regard to responses to shocks and disturbances, interactions of slow and rapid changes (both in human and natural systems), and system and actor dynamics. What is particularly encouraging from a human geography perspective on resilience is that studies are increasingly acknowledging the importance of hitherto neglected forces and processes affecting resilience linked to human geography-related notions of power, space/place, agency, and governance.

SEE ALSO: Climate adaptation/mitigation; Climate and societal impacts; Climate change
Resilience and Human Geography

adaptation and social transformation; Community; Community-based natural resource management; Environment and migration; Environmental (in)security; Environmental science and society; Social-ecological transformation; Social vulnerability and environmental hazards

References


Further reading

The resource curse has deep historical roots, including the consequences of Spanish exploitation of New World gold and silver mines that evoke echoes of King Midas. Recent interest arose from the failure of many oil-exporting countries to apply their large windfall revenues from the 1974–1981 oil booms to sustainable economic development. Country case studies revealed mostly adverse outcomes, which prompted a surge of statistical studies triggered by the pioneering work of Sachs and Warner (1995). So far, however, the statistical studies have failed to provide a definitive explanation of the resource curse. Rather, they have expanded the range of explanations from an initial focus on economic causes to institutions and policy. Some scholars now query the existence of a resource curse (Brunnschweiler 2008). Certainly, the causes and symptoms of the resource curse are more complex than many researchers initially thought. This entry compares the competing explanations; identifies systematic differences in natural resource characteristics and outcomes; notes how outcomes vary with shifts in development policy; and conceptualizes the resource curse as part of a broader rent curse. Future research can build on this more nuanced context to model systematic variations in the impact of natural resources on economic development.

Economists expect resource-rich economies to grow faster than resource-poor economies if their governments apply the natural resource rent (the surplus revenue after deducting all production costs, including a risk-related return on investment from total revenue) to sustain a higher rate of investment and greater imports of the goods required to build the infrastructure of a modern economy. This appears to have been the case in many countries prior to the 1970s when the per capita GDP of the resource-rich economies as a group was 50% higher than that of the resource-poor economies (Auty 2001, 5). The sharp increase in the global incidence of the resource curse that emerged during the commodity price volatility of 1973–1985 therefore presents a paradox that persisted until the late 1990s when the growth rate of the resource-rich economies improved (Table 1). But by then the mean per capita income of the resource-poor economies exceeded that of the resource-rich ones. This waxing and then waning of resource curse effects at the global level suggests the curse is not a deterministic law, as many statistical studies have sought to prove, but varies in its intensity through time.

Sachs and Warner (1995) offer an economic explanation for the resource curse, which is rooted in Dutch disease effects (deindustrialization caused by large inflows of foreign currency from booming commodity exports).
### Table 1  Per capita GDP growth 1960–2010, by country resource endowment (%/year).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Resource-poor</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>2.4</td>
<td>3.7</td>
<td>4.7</td>
<td>3.4</td>
</tr>
<tr>
<td>Small</td>
<td>3.5</td>
<td>1.8</td>
<td>2.4</td>
<td>2.4</td>
</tr>
<tr>
<td><strong>Resource-rich</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td>2.7</td>
<td>0.7</td>
<td>1.9</td>
<td>2.4</td>
</tr>
<tr>
<td>Small, non-mineral</td>
<td>1.6</td>
<td>0.7</td>
<td>0.9</td>
<td>1.9</td>
</tr>
<tr>
<td>Small, hard mineral</td>
<td>2.2</td>
<td>0.1</td>
<td>−0.4</td>
<td>2.1</td>
</tr>
<tr>
<td>Small, oil exporter</td>
<td>4.0</td>
<td>2.3</td>
<td>−0.7</td>
<td>1.8</td>
</tr>
<tr>
<td>All countries</td>
<td>2.7</td>
<td>1.6</td>
<td>1.5</td>
<td>2.2</td>
</tr>
</tbody>
</table>

*Resource-poor = 1970 cropland/head < 0.3 ha
†Large = 1970 GDP > $7 billion (proxy for domestic market size)

Source: Derived from World Bank (2013). See Auty (2001, 4) for composition of country endowment groups.

Dutch disease effects occur when booming commodity revenue strengthens the real exchange rate, which reduces the competitiveness of the non-booming tradeable sectors (agriculture and manufacturing) so that they contract. When commodity prices eventually fall, the economy may have less capacity to generate wealth than before the boom occurred because the adjustment by firms to the exchange rate is asymmetric: tradeable activity is rapidly destroyed in booms but revives much more slowly when a boom abates (Gelb and Associates 1988). Sachs and Warner show that in order to counter the employment–diminishing effects of Dutch disease, most governments closed their trade policy, especially when their dependence on primary product exports was greater. This policy is counter-productive, however, because protective trade policies repress competitive markets so the economy becomes less flexible and more vulnerable to a growth collapse. Such growth collapses are triggered by sudden falls in commodity prices/revenue, like those the world economy experienced through 1973–1985. They damaged those resource-rich economies where governments had stoked Dutch disease effects by expending the revenue from their natural resources too quickly.

For geographers, the economic explanation for the resource curse appeared to herald a new variant of environmental determinism. Yet resource-rich Botswana, Malaysia, and Indonesia, along with earlier resource-abundant developers like Australia, Canada, and the United States, are anomalies that sustained rapid per capita GDP growth. Their experience nurtured institutional explanations of the resource curse led by Acemoglu, Johnson, and Robinson (2001) who argue that the quality of institutions (the formal rules governing political and economic exchange) is more important than natural resources per se in determining whether resources are a blessing or a curse. They identify as detrimental to economic growth
those extractive (colonial) institutions associated with tropical lands that were too unhealthy for significant European settlement. Managers of plantations and mines in the tropics promoted extractive institutions to export goods and send revenues to owners overseas while squeezing worker income and local expenditure (Isham et al. 2005). They further constrained economic diversification by repressing the local production of rival commodities to limit competition for land and labor. In contrast, small-scale farmers in the resource-rich temperate regions of the Americas and Oceania created inclusive institutions that were more conducive to local economic development (Baldwin 1956). As incomes rose, the farmers paid taxes to develop roads, expand education, and maintain law and order for all citizens, facilitating incremental gains in productivity that boosted farm incomes whose local expenditure helped diversify the regional economy.

Yet Glaeser et al. (2004) relegate institutions to secondary status in promoting economic growth. They find that institutions improve as a consequence of rising incomes but do not cause that rise, which they attribute to increased education and choice of economic policy. In fact, institutions in low-income economies are highly flexible. They bend to accommodate elite demands for personal enrichment rather than constrain those demands, which effective resource-driven development requires. Moreover, North et al. (2009) argue that the first priority of many governments in developing countries is to apply rent to appease elite groups that are deemed capable of violence, so that political motives override economic ones in the allocation of natural resource rent.

Consistent with this outcome, Schlumberger (2008) identifies a patrimonial form of capitalism in oil-exporting countries (with wider applicability), the principal feature being the use of informal institutions to bypass formal ones to benefit the politically connected elite. The World Bank (2009) provides an illustration. It reports that almost two-thirds of a sample of Middle East and North African civil servants view the private sector as rent extracting and corrupt, being dominated by a few well-connected firms that sustain quasi-monopolies by barring competitors, bribing civil servants, lobbying for privilege, and underreporting earnings. Such a system undermines long-term growth.

In fact, the economic, institutional, and political explanations are not mutually exclusive and may interact to cause the resource curse. Windfall revenues from natural resources trigger contests among the elite to extract rent, that exert political pressure for governments to expend rent too rapidly in the domestic economy, which undermines macro-economic management (Torvik 2002). This policy failure triggers Dutch disease effects that retard the competitive diversification of the economy; impede institutional maturation; and hinder shifts toward political pluralism. But policy failure is not inevitable: the risk is higher where the rent is concentrated on a handful of economic agents, as with modern mining, as well as in the presence of strongly interventionist government policies.

Development outcomes from dispersed and concentrated resource rent

Natural resource rent can be extracted from minerals, cropland, forests, and landscapes, but the statistical research has focused on mineral rent. This is partly because the World Bank provides annual estimates of mineral and hydrocarbon rent for most countries from 1970 but not for crops – data for this only became available recently and at irregular intervals because of the high preparation costs. GDP data are available, however, and Table 1 indicates that the hard
mineral exporters performed worst through the phase of commodity price volatility associated with the oil booms of 1973–1985. The oil-exporting countries saw their growth slow down through 1973–1985 and then collapse. By comparison, small crop-exporting countries fared somewhat better and might have performed better still with less government intrusion.

Systematic differences between the rent streams explain the varied development outcomes. Crop rent tends to be smaller relative to GDP than mineral rent (Auty 2001, 131), and also less volatile, which facilitates macro management (Van der Ploeg and Poelhekke 2009). In addition, crop rent is more dispersed across economic agents than mineral rent (Table 2, columns 1 and 2), which renders it harder for the elite to expropriate and more likely to be deployed efficiently.

Mineral rent streams tend to be singularly large, volatile and concentrated on governments. They are large because most mining is capital-intensive and subject to economies of scale, rendering both mines and their rent streams large as a proportion of the GDP of the typical host economy. Small mineral economies secure the highest rent in relation to GDP and the oil exporters the most rent of all (Table 1), so they are especially prone to resource curse effects. In a sample of six oil-exporting countries, Gelb and Associates (1988) estimated that the rent windfalls during the 1974–1978 and 1979–1981 oil booms ranged from 10% of non-mining GDP per annum in Venezuela to over one-third of non-mining GDP in tiny Trinidad and Tobago. Some Gulf oil exporters had rent/non-mining GDP ratios higher still.

The rent stream of mineral economies is also volatile because mineral markets adjust less flexibly than soft commodity (crop) markets. New mining capacity requires large investments with lead times of five to ten years, so capacity is difficult to match to demand. Shortages boost prices but the investment to meet the extra demand takes several years to construct, by which time conditions may have changed and created a glut. Yet low prices do not encourage temporary mine closure to match demand and supply because mining companies continue to operate as long as revenues cover operating costs and a modest contribution to capital recovery.

These characteristics of mining also concentrate its domestic economic stimulus (linkage) on taxation, and therefore on the government. For example, backward linkages from mining are for specialized inputs that are most cheaply imported from established suppliers. Moreover, further processing prior to export (forward linkage) is most efficiently undertaken overseas rather than at a remote mine or oil well unless processing entails significant weight loss. Finally, the capital-intensive nature of mining creates little employment, so local expenditure by workers (final demand linkage) is limited. Hence, the principal economic contribution of mining to the host economy is taxation (fiscal linkage), which governments receive but often struggle to expend effectively (Lal 1983).

The economic policies required to manage large and volatile revenue streams are well understood and include a revenue stabilization fund to smooth the rate of domestic absorption and fiscal rules to ensure the rate is sustainable. However, many governments find it difficult to resist political pressure to boost expenditure through booms and to avoid cuts through downswings. Consequently, domestic rent absorption is invariably over-rapid and intensifies Dutch disease effects and rent expropriation by elite groups. This retards the diversification of the economy that is required to establish the productivity-driven growth upon which long-term prosperity depends. Instead, governments rely on higher rents to raise incomes. For example, the oil price required by Saudi Arabia and Russia to balance...
their budgets increased four-fold between 2004 and 2012. 

Research suggests that governments not only expend revenue too quickly but also favor consumption over investment. They boost consumption inefficiently, however, by reducing taxes for the wealthy, subsidizing basic commodities like food and energy, and over expanding the bureaucracy and protected infant industry so as to mask unemployment. Prior to the economic reforms demanded by international financial institutions (IFIs) in the late 1980s and 1990s, governments in resource-rich countries frequently pursued strongly interventionist policies. They invested windfall revenue in state-owned enterprises that proved inefficient and required subsidies (Gelb and Associates 1988, 94–123). Many public expenditure programs degenerated into conduits for enriching the elite or for delivering political support. Far from promoting the competitive diversification of the economy, governments over expanded social entitlements, which they struggled to reduce when revenues fell. This, along with subsidies to loss-making state enterprises, boosted public debt to levels that required IFI-backed assistance for economic reform.

Table 2  Rent stream properties and predicted political economy impacts, by rent source*.

<table>
<thead>
<tr>
<th>Rent stream properties</th>
<th>Diffuse natural resource rent</th>
<th>Concentrated natural resource rent</th>
<th>Regulatory rent</th>
<th>Remittances (labor rent)</th>
<th>Foreign aid (geopolitical rent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale (% GDP)</td>
<td>5–15</td>
<td>8–20+</td>
<td>5–20+</td>
<td>2–10</td>
<td>2–10+</td>
</tr>
<tr>
<td>Degree of rent concentration</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Volatility (standard deviation)</td>
<td>Moderate</td>
<td>High</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Low</td>
</tr>
<tr>
<td>Potential rent impacts: economic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investment efficiency</td>
<td>High</td>
<td>Falling</td>
<td>Low</td>
<td>High</td>
<td>Equivocal</td>
</tr>
<tr>
<td>GDP growth</td>
<td>Rapid</td>
<td>Decelerating</td>
<td>Decelerating</td>
<td>Moderate</td>
<td>Equivocal</td>
</tr>
<tr>
<td>Dutch disease effects</td>
<td>High</td>
<td>High</td>
<td>Moderate</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>Market repression and corruption</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Equivocal</td>
</tr>
<tr>
<td>Potential rent impacts: political</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-reliant social capital</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Moderate</td>
<td>Equivocal</td>
</tr>
<tr>
<td>Proliferation of social groups</td>
<td>High</td>
<td>Constrained</td>
<td>Constrained</td>
<td>High</td>
<td>Equivocal</td>
</tr>
</tbody>
</table>

*Most economies generate more than one rent stream.
In contrast, the smaller, less volatile and dispersed rent streams from small-scale farming lower the risk of policy failure. Windfall revenue that is scattered across many economic agents tends to be more beneficial for economic growth than rent spent by governments (Isham et al. 2005). Freed of political pressures, small-scale farmers are more likely than governments to save more of their windfall revenue and thereby smooth its absorption into the economy, limiting Dutch disease effects (Bevan, Collier, and Gunning 1987). Small-scale farmers also allocate revenue more judiciously between consumption and investment than governments, and their investment is more efficient because it reflects economic goals rather than political ones.

Baldwin (1956) illustrates successful agriculture-driven economic diversification with reference to the nineteenth-century US Midwest. There small-scale farms made modest incremental investments that steadily boosted productivity and rural incomes. This expanded demand for locally manufactured farm implements and household goods. They also paid taxes for education, infrastructure, and public administration that conferred social benefits that were widely shared. The dispersed linkages of Midwest farms built a prosperous, diversified economy. In contrast, as institutional scholars have argued, the Southern plantation economy’s large-scale crop processing system minimized local incomes and tax revenues, cementing dependence on the export commodity and perpetuating a skewed income distribution and mono-crop economy.

East Asian case studies suggest the dispersed economic linkages of small farms can sustain rapid GDP growth at 6–7% per annum. Unfortunately, through the 1960s to the 1980s, government policy in many developing countries viewed agriculture not as a growth engine but as a supply of cheap labor, cheap food and rent, which was exploited to subsidize urban industry. Many governments in sub-Saharan Africa abused their crop marketing boards to extract rent from farmers by paying them far less than the world price for their crops (Krueger 1992, 58–73). Some governments went further and extracted part of the return to capital and labor from farmers as well as the rent. In effect, government rent extraction transformed the beneficially dispersed cash crop linkages into concentrated linkages like those in mining. This transferred economic resources from competitive rural activity to subsidized urban activity to appease vocal political groups. The policy was self-defeating because it eroded farm incentives, which encouraged more farmers to migrate to cities, where they demanded subsidized jobs.

To summarize, concentrated rent flows increase the resource curse risk for small economies, especially mineral exporters, by feeding contests among the elite to extract rent for immediate personal enrichment at the expense of efficient investment for long-term economic growth. The resulting over-rapid rent expenditure ignites Dutch disease, that causes growth collapses in the absence of economic reform, which rent recipients resist. However, the smaller dispersed rent streams from small-scale cash crops confer lower risk of resource curse effects because they are harder for the elite to expropriate, and farmers apply them effectively, provided governments resist political pressure to convert dispersed rent into concentrated rent.

Development policy varies the incidence of resource curse cases

include data from the 1990s and 2000s and are more likely to generate a positive association between resource abundance and economic growth, leading some to reject the very existence of a resource curse (Brunnschweiler 2008). The inclusion of recent statistics increases the weighting of the post-1997 years of economic recovery relative to those impacted by the 1973–1985 growth collapses (Table 1).

The growth recovery coincided with a major change in economic policy (Van der Ploeg and Poelhekke 2009) when IFI assistance with debt management was made conditional on the adoption of policies to stabilize the economy and dilute interventionist government policies. The IFI stance reversed an earlier global shift toward interventionist policies after the Great Depression and World War II, when some economists advised developing country governments to override markets to promote industrial diversification and reduce dependence on the advanced economies. Many developing country governments duly closed their trade policy through the 1950s, 1960s, and 1970s to protect “infant” industry from competition with imports from advanced economies until it was efficient enough to survive unaided. Infant industry maturation rates need, however, to be no more than five to seven years if the costs of infant support are to be recouped from the benefits of the successful industry. In practice, industry maturation rates spanned decades.

The Achilles heel of industrial policy is policy capture by the prime beneficiaries, the owners and workers operating the plants, and the civil servants awarding the licenses (Altenburg 2013). Import protection blunts incentives to compete, so industrial policy risks degenerating into a vehicle for extracting rent and bestowing political patronage, usually at the expense of domestic consumers (including farmers and unsubsidized manufacturers) who pay prices above world levels, often for inferior goods. Moreover, infant factories import inputs but do not generate foreign exchange to pay for them because they are uncompetitive, so their operation widens the trade deficit and can boost public sector debt. The infant industry policy caused stop–go macro-economic crises, notably in Latin America, which pioneered the strategy.

Research suggests the rise and then decline in the global incidence of growth collapses associated with the resource curse reflect, respectively, interventionist government policies in many resource-rich economies through the 1960s and 1970s, and then economic recovery aided by IFI-backed reforms. Case studies reveal that over the same period, many resource-poor economies grew rapidly by investing to export labor-intensive manufactured goods in line with their comparative advantage, which lay in cheap labor due to the paucity of natural resources. The resulting competitive industrialization development trajectory rapidly exhausted surplus rural labor. This beneficially pushed up wages and forced firms to contain rising costs by boosting worker productivity through rapid diversification into skill-intensive and capital-intensive manufacturing. Examples include the resource-poor Asian Dragons (Hong Kong, Singapore, South Korea, and Taiwan), Mauritius, and post-Mao China but also resource-rich Malaysia and Indonesia. Malaysia and Indonesia shared large rural populations with low incomes that were courted by rebels and so posed a political risk to the leadership. The elite responded by promoting rural reform and green revolution techniques to ensure that their rent extraction did not prevent development benefits from diffusing to the poor, an outcome that helped avoid growth collapses.
The resource curse within a broader rent curse

Overall, resource-poor developing economies were more successful than resource-rich ones during 1973–1997 (Table 1), but there are anomalies that add further complexity to the resource curse. Some resource-poor economies, like those of the Sahel, experienced growth collapses. An important reason is rent streams other than natural resource rent: namely foreign aid, worker remittances, and regulatory rent. Table 2 compares the features of these revenue streams with those of the two basic natural resource rent streams (diffuse and concentrated). Each rent source can be 10–20% of GDP or more, potentially taking the total rent within an economy to between one-fifth to one third of GDP and more. Mauritania, for example, generated total rent of one-third to two-fifths of GDP per annum for over forty years after independence. Such windfall revenue is detached from its source and is up for grabs, so it is not surprising that contests for rent among the elite profoundly impact the political economy. Yet many statistical studies of the resource curse ignore these extra rent streams so they are, at best, impaired by background noise and, at worst, flawed in their conclusions.

Rajan and Subramanian (2011) confirm that domestic expenditure of foreign aid within the public sector can replicate features associated with a natural resource commodity boom, triggering Dutch disease effects that stifle competitive manufacturing. Moreover, foreign aid may be stolen by the elite for personal enrichment by biasing public expenditure toward targeted goods such as construction projects that are awarded to political cronies, rather than toward universal public goods such as education and health care that benefit everyone, supporter and foe alike. Nevertheless, overall development outcomes from flows of aid are equivocal (Table 2) because of variations in the rigor of aid supervision: donors became more discriminating in allocating aid in the 1990s.

“Regulatory” rent arises from manipulating relative prices to favor one group over another. It is frequently generated by closing trade policy to expand the scope for state intervention to create rent by changing relative prices. By its nature, regulatory rent is concentrated, which facilitates its theft by the elite (Table 2). Since economic reform shrinks scope for rent extraction by expanding competitive markets, it is often opposed by elite groups.

Finally, wage remittances are a diffuse form of rent with beneficial effects for very low-income countries because they boost domestic consumption, often of the poorest, and fund local investment where financial systems are underdeveloped (Giuliano and Ruiz-Arranz 2009). These positive impacts on economic growth attenuate, however, because as incomes rise remittances weaken work incentives, and financial intermediation improves. Moreover, as with most windfall revenues, large remittance flows can depress GDP growth through Dutch disease effects (Rajan and Subramanian 2011).

Further research on the resource curse

The literature on the resource curse has expanded rapidly since the 1980s and this review has of necessity focused on key issues, to the exclusion of more specialized debates, such as: the rentier state theory of autocratic governance (Ross 2001); the impact of resources on violence (Collier and Hoeffler 2004); and whether the appropriate measure of resource reliance is the export share or GNP share (Van der Ploeg 2011, 380).
The causes and manifestations of the resource curse are more complex than initially expected so it is not surprising that statistical analysis has struggled to demonstrate conclusively whether a resource curse exists and what may cause it. Future research might combine the insights of statistical and case study research to construct and test stylized models of the interaction of economic, institutional, and political factors with critical variables such as the scale of the resource rent, its volatility, and its distribution across economic agents. For example, the successful low-rent economies appear to follow a competitive industrialization development trajectory rooted in the incentive that low rents confer on the elite to grow the economy (and their own wealth) by promoting economic efficiency. This aligns the economy with its comparative advantage in labor-intensive manufactured exports. It proved remarkably effective at propelling economic diversification and absorbing labor to sustain rapid and egalitarian growth in per capita GDP with strengthening institutions and shifts, albeit lagged, toward political pluralism.

In contrast, high rent frequently incentivizes the elite to compete for rent for immediate self-enrichment at the expense of promoting long-term economic growth. This set of incentives locks the economy into a staple-trap development trajectory that retards competitive industrialization. It perpetuates surplus labor and prompts government intervention with subsidies from the resource sector to create urban jobs that markets would not otherwise support. The subsidies over burden and weaken the resource sector, eventually triggering a growth collapse that is protracted because rent recipients resist reform. Further research might therefore also seek practical strategies to incentivize the elite in high-rent economies to prioritize growth over rent extraction, perhaps by distributing most rent on a per capita basis or adapting the dual-track reform strategies of Malaysia and Mauritius.

SEE ALSO: Comparative advantage; Dependency theory; Economic development zones; Environmental determinism; Growth theory; Industrial linkage; Natural resources; Political geography; Resources and development; States and development

References


Resource extraction

Gavin Bridge
Durham University, UK

Resource extraction refers to the separation and removal of natural resources from their immediate context. Extraction is widely interpreted as a physical process of displacement by which “gifts of nature” are removed from the Earth and transformed into useful raw materials: for example, the abstraction of water from rivers, the pumping of oil from underground formations, or the mining of coal, copper, and tin. In wresting materials from the Earth and mobilizing them into the economy, resource extraction exemplifies a classic “primary sector” activity. It is a process made epistemologically tractable primarily via the engineering and physical sciences: it is through hydrology, geophysics, and economic geology, for example, that raw materials are cleaved from their surroundings, enumerated in ways that enable their capture and insertion into markets, and are put to work in the service of trade and industry.

However, a more specific understanding of extraction interprets it as a political-economic process of expanding value and/or power via the identification, capture, and control of extractable materials. Because resources are not only materials but also arrangements of property, their extraction implies relationships of social power: the ability of a group (e.g., tribe, class, nation, corporation) to control land and reserve it against competing uses and users for extractive purposes; to marshal human labor and energy resources to the task of separating and moving materials; to achieve and control access to markets to realize value from the products of extraction; and to repress, legitimate, or otherwise contain the political tensions involved in controlling and appropriating natural resources. In short, resource extraction produces not only raw materials but also social-spatial distributions of power and geographically uneven forms of economic development. As such, extraction’s rationales and effects may be disclosed through critical social science.

A political-economic or political-ecological interpretation of extraction highlights how the reproduction of social power is at stake in the displacement and mobilization of raw materials. This is both a more specific and more expansive definition than physical removal, as it allows a wide range of natural resource gathering activities to be interpreted as extraction, beyond the iconic examples of vertical displacement associated with mining or other subterranean resources. The hunting and trapping of animals (e.g., the fur trade, or ocean fishing) or the gathering of plant material (e.g., rubber tapping) are also examples of extractive practices in that they involve the capture, control, and relocation of nonproduced materials in ways that generate value. The nineteenth century agricultural chemist, Justus von Liebig – whose analysis of nutrient cycling underpins Marx’s notion of socioecological metabolism – highlighted agricultural production as a process of nutrient export that impoverished the countryside by extracting its “mineral elements” and re-depositing them as “solid and fluid excrements” in the city. Extraction also occurs without the physical export of materials: the generation of electrical power via photovoltaic
RESOURCE EXTRACTION

cells and wind or water turbines, for example, is a form of resource extraction based on locating, controlling, and converting into electrical energy a particular concentration of productive forces (areas of reliable wind flow or insolation).

A political-economic interpretation of extraction foregrounds the question of what is being extracted, beyond physical materials. For Marxist analyses, an answer lies in the surplus value created by human labor. Resource extraction, in this sense, is no different to other forms of economic activity under capitalism which is an inherently extractive mode of production: although wages are represented as fair exchange, they do not fully compensate workers for their labor time as some of the value added by the power of labor (the “surplus” beyond that for which workers are paid) is appropriated by capitalists as profit. However, resource extraction is different from other forms of economy because the materials being extracted, and subsequently worked upon, pre-exist human labor: they are, in the language of classical political economy, “gifts of nature” as “the fundamental values in lumber, in minerals, oil, fish, and so forth, are predominantly in the good itself rather than in the labor incorporated in it” (Bunker 1984, 1054). The origin of value in natural resources thus turns out to be a surprisingly challenging question, revealing the limits of both orthodox and Marxist accounts of political economy.

In response, researchers have turned to the field of ecological economics which has sought, in different ways, to understand the relationship between the biophysical qualities of materials and the cultural metric of value. A common interpretation reduces value to physical properties, interpreting extraction as the capture and relocation of “natural values” in the form of low entropy (i.e., highly ordered) material forms. The energy density of crude oil, for example, arises from the processes of oil formation that effectively condense solar energy flows (received over extensive geographical areas and very long periods of time) into a material form with a very high capacity to do work: once brought to the surface, this capacity can be utilized (via combustion) to again compress the experience of space and time, in the form of enhanced mobility and speed. In a similar way, trade in agricultural goods like wheat, sugar, or cotton is a form of “time–space appropriation” as the calorific and nutrient content of these crops have been concentrated and upgraded by place-based processes of photosynthesis and nutrient cycling (Hornborg 2006). Such analysis enables trade to be understood as a form of unequal ecological exchange structuring uneven geographical development, in which highly ordered forms of raw materials are drawn from extractive peripheries of the world economy and consumed within the industrial core.

More recent research takes up the insights of ecological economics but aims for a nonreductionist account of resource extraction and trade, one that does not conflate the social category of value with physical measures of matter and entropy. By understanding industrial production as a process that simultaneously creates greater consumer value and greater entropy (see e.g., the pioneering work of Georgescu-Roegen 1971) it is possible to show how “the market exchange of finished industrial products for fuels and raw materials will inexorably reward the dissipation of natural resources in core areas of the world-system with economic access to more resources to dissipate … economic growth means that the more resources the core has dissipated today, the more new resources it will be able to dissipate tomorrow” (Hornborg 2015). In sum, contemporary approaches to the political economy and ecology of extraction decenter physical processes of material displacement and
identify how extraction sustains broader projects of capital circulation and/or the reproduction of social power (Huber 2013; Mitchell 2011).

Classifying the diversity of resource extraction

Resource extraction encompasses a bewildering diversity of practices, organizational forms, and materials, from birds’ nests and Brazil nuts, and coal and copper, to rubber, rutile, sulfur, seeds, and shrimp. Resources are extracted in all physical states (solids, liquids, and gases), and from all parts of the Earth’s system (atmosphere, lithosphere, hydrosphere, biosphere). Extractive economies involve diverse labor practices (hunting, cultivating, gathering, digging), and physical and chemical techniques (selection, concentration, disaggregation, transformation). Accordingly, extraction takes a diversity of organizational forms. Groundwater resources, for example, are abstracted for irrigation by private famers using gas-powered pumps in the western United States, mobilized from the Nubian aquifer by the Libyan state to supply coastal development, and drawn by hand from communal “singing wells” of the Borana region of Ethiopia. Coal is picked from slag heaps and waste piles as an informal source of energy the world over, but it is also mined deep underground with different degrees of technological assistance, and extracted in giant bucket-wheel excavators by state firms in Greece.

Organizing the diversity of extraction

Extraction’s diversity can be organized by reference to three broad criteria: the potential of the resource base to renew itself, the rationales underpinning extraction, and the industrial dynamics through which raw materials are appropriated. The first criterion centers on the timescale over which extracted materials will regenerate, where the primary differentiation is between stock (nonrenewable) or flow (renewable) resources. Stock resources exist in finite supply: removing these resources diminishes the physical stock available for future use, and so all nonrenewable resources are exhaustible by definition. Extractive resources that have significant regenerative capacity are classified as flow resources or renewables, although they too can be exhausted if consumption exceeds the rate of regeneration: this category of exhaustible renewables includes groundwater, fish, game species, and soil resources, all of which can be “mined” to depletion/extinction. Because a threshold exists beyond which further consumption exceeds the capacity of the resource to regenerate, exhaustible renewables are also known as “critical zone” resources. Many contemporary environmental challenges relate to the failure to manage critical zone resources effectively by maintaining rates of extraction within the boundaries of biophysical systems. The capacity of many biotic and abiotic resources to renew themselves is dependent not only on natural regeneration times but also on levels of social investment. Studies of traditional agriculture, for example, show human labor playing a vital role in maintaining soil fertility, and contemporary agro-industrial practices invest large energy and financial resources (in the form of synthetic fertilizers, or pumping for irrigation or drainage) to maintain the utility of the soil resource and offset the effects of nutrient depletion, soil erosion, and soil degradation. Similarly, development of an extensive infrastructure for collecting and recycling aluminum would enable society to treat this metal more like a flow resource than a stock. Advocates of a “closed loop” approach to material demands seek to expand secondary recovery of anthropogenic material flows in a bid to lessen the demand on primary forms of resource extraction.
RESOURCE EXTRACTION

Approaches to “mining the technosphere,” for example, highlight the potential for recovering materials accumulated in above-ground stocks, such as waste piles, materials in use, and a range of “hibernating” materials currently incorporated into abandoned urban infrastructures (Johansson et al. 2013). For this reason, it is helpful to think of a continuum of extractable resources running from nonrenewable stock resources which are consumed by use (such as fossil fuels) to the infinitely renewable resources of solar energy, tidal power, and wind.

The second criterion centers on a distinction between, on the one hand, extraction for subsistence and the reproduction of livelihoods and, on the other, intensive extraction on a scale that massively exceeds local consumption and is driven by a logic of export for commercial gain. Guha (2006, 233) differentiates *ecosystem people* from *omnivores* as a way of expressing the “basic and massive asymmetry” between these two resource extraction rationales. The terms capture two contrasting scales of “resource catchment” and suggest there is an important difference in the meaning and significance of extraction in each case. Development, Guha argues, is a state-assisted project to channel an ever-growing volume of extractive resources toward urban and rural omnivores. Moreover, it is a project that comes at the expense of ecosystem people, who have seen their subsistence extractive activities curtailed and replaced by large-scale forestry, dam schemes, mining, and agricultural plantations. In a similar way, Gudynas (2013) has coined the term *extractivismo* to reflect a particular modality of extraction that has dominated the history of Latin America: large-scale or high intensity resource extraction, geared for export as either raw material or with minimal processing. Coined initially in the context of the large-scale mining projects that materialized in parts of the Andes during the 1990s, *extractivismo* can equally apply to plantation agriculture, forestry, or fisheries. Importantly, these analyses are not a critique of extraction per se but of the sociopolitical context within which it occurs. There are parallels here with the challenge laid down by Howitt (2001) to find ways of rethinking resource extraction outside of a “colonial” frame in which extraction is directed by and serves nonlocal interests.

The third criterion recognizes that all extraction is a process of “primitive accumulation” through which economic value is captured via the appropriation of raw materials that, in significant ways, are produced independent of human labor. However, the industrial dynamics involved in generating this flux of materials take two broadly different forms. In some instances value is generated by the simple expedient of expanding the scale of appropriation: that is, by laying claim to a larger proportion of the Earth’s stock of resources and biogeochemical flows. The geographical extension of the “extractive frontier” and the subsequent development of property rights that license extraction of fish, timber, and mineral resources, for example, are illustrative moments of this “formal subsumption of nature” (Boyd, Prudham, and Schurman 2001). However, there are also instances where extraction involves laying hold of biogeochemical processes more directly, in order to harness and improve their productivity. The extraction of liquid oil from bituminous sands, for example, requires capital-intensive interventions in heating and hydrogenation to “upgrade” the raw material to a synthetic equivalent of crude. Similarly, the application of bacterial mining techniques (e.g., for the extraction of copper and gold) requires optimizing biological conditions for the reproduction of thermophillic bacteria able to liberate the target element from sulfide ores. Extraction in these cases involves a partial capitalization of the ecological conditions
RESOURCE EXTRACTION

required to generate the resource, with technology and the application of energy inputs substituting for the deep-time “geological” processes upon which conventional forms of oil and mineral extraction rely. This “real subsumption of nature” effectively “squeezes” or tailors natural production processes to accelerate raw material production or maximize particularly valued traits. The distinction between formal and real subsumption as modes of extraction is neatly illustrated by the difference between extracting timber from plantations or via old-growth forestry or the harvesting of fish via aquaculture or ocean fishing.

The distinctiveness of extraction

Juxtaposing extraction to industrial production highlights its distinctiveness as a mode of economy (Bunker 1984). First, materials extracted from the Earth are “fictitious commodities” (Polanyi 1944): we exchange them like industrial commodities but, unlike cell phones, chairs, or automobiles, they cannot be fully produced because societies are unable to replicate the geological, hydrothermal, and deep-time processes of which they are made. Second, extractive economies are marked by a low ratio of capital and labor relative to the value of the final product. This is strikingly illustrated by alluvial diamonds (which may be picked by hand from gravel with limited capital or labor), shallow oil and gas wells, or highly prized animal parts like ivory, although it is a general feature of all extraction including large mines, trawler vessels, dams, and offshore oil platforms. A corollary is that a large portion of the capital invested in extraction is tied up in export infrastructure tailored to handling a particular resource, which can be difficult to adapt to alternative demands once extraction is over. Third, the heterogeneous character of most natural resources means that economies of scale in extraction do not work the same way as in manufacturing. Scaling up often involves drawing upon resources located deeper or more remote from the initial point of extraction or working with lower quality materials. As a consequence, an increase in the volume of throughput tends not to reduce unit costs in the same way as in manufacturing, unless it is accompanied by changes in extractive technology or labor regime.

Fourth, the state frequently has a more central role in resource extraction than in industrial production. This is not only because many “strategic” resources are reserved to the state rather than being held in private hands, but also because the ability to access a resource is often more important than nominal patterns of ownership. The capacity of the state to authorize, enforce, and legitimate extraction is significant because resource development frequently overturns existing norms around land/resource ownership and access. Fifth, practices of extraction are shaped to a greater degree by the encounter with physical materials than are those of industrial production. The biophysical capacities and affordances of materials influence the forms of property, class relations, labor regimes, and technologies adopted for their extraction. Long-term extractive economies like the Amazon or California/Nevada illustrate a progression through several distinctive “modes of extraction,” with social relations laid down in one mode shaping the next (Bunker 1985; Brechin 1999).

The scale and temporality of resource extraction

Resource extraction is a primal pursuit, evolving over time from relatively simple processes of gathering pre-formed materials to the application of greater quantities of labor and fossil energies to locate and liberate materials of social
value. The recovery of nuggets and flecks of gold from placer deposits via physical panning techniques, for example, has over time been supplemented with high-pressure hydraulic mining, the development of deep mining techniques, and chemical processing to extract gold from highly disseminated (i.e., ultra-low grade) and refractory reserves. In a similar way, as the extraction of fish from offshore waters has shifted progressively further offshore and to deeper marine environments, the energy inputs required to extract a given unit of resource have increased. In short, many resources have become progressively more difficult to extract over time, although this “depletion” effect has, for the most part, been offset by new technologies of exploration and extraction, the availability of low-cost energy inputs, and forms of regulation that have facilitated access to new “frontiers” of extraction.

Since the early nineteenth century, a progressively growing torrent of raw materials has been drawn from the Earth and directed toward the industrializing heartlands of the world economy, exemplified by an increasingly significant long-distance trade in bulk resources like coal, copper, and cotton. For many resources the rate of extraction has increased markedly in the second half of the twentieth century. This “Great Acceleration” has driven a material intensification of the economy in terms of absolute volumes/tonnages of resources extracted from the planet’s biogeophysical systems. The most intensely extracted materials – sand and gravel – are now equal in volume to physical flows associated with fluvial systems. These flows, together with the very large volumes of waste materials moved in mining, mean that humans rival wind and water as terra-forming geomorphological agents. Anthropogenic flows of elements like carbon, mercury, sulfur, and nitrogen, which are linked in different ways to extractive practices, now exceed background biogeochemical flows.

While some resources (like sperm whales or guano), around which entire extractive industries were built, have ceased to be regarded as commercial resources, for the most part the volume and range of extracted materials has grown. This massive scale and diversity of resource extraction is a signal feature of the Anthropocene.

Several other temporalities of resource extraction are overlaid on this secular trend of an expanding volume. Extractive economies are characterized by cycles of boom and bust, in which capital and labor are deployed around rises in the price of raw materials, only to see new investment marooned and workforces laid off when prices collapse. This recurring characteristic in the development of extractive economies is closely tied to the capital intensity of extraction and the fungible nature of many extractive commodities. From the late 1990s onwards, for example, prices for many raw materials were driven upwards by high and sustained rates of resource-intensive economic growth in Asia. Projections of further demand growth attracted investment into the extractive sector, leading to the development of new mines, forestry plantations, agricultural projects, and the infrastructure for resource export. This “super-cycle” of rising prices and new investment was sustained for around a decade by relatively high rates of growth in emerging markets and growing concern about physical scarcity constraints (e.g., debates concerning peak oil; Bridge 2010).

The cyclical character of raw material prices means that extractive investment tends to wax and wane over time, although this does not always have a straightforward geographical expression. The history of mining investment includes periods of geographical extensification, when mining capital scours the globe for new extractive opportunities. During the 1990s, for example, mining capital underwent a profound process of geographical extensification with a
series of large new mines opened in Latin America, South East Asia, and Africa and which gave rise to a new “geopolitical ecology” of extractive exports (Bridge 2004; Bebbington and Bury 2013, 16). Extension of the extractive frontier in this period was enabled by the adoption in many resource-rich countries of national policies of economic liberalization, often required by international financial institutions as part of a package of “structural adjustments” in the wake of the 1980s debt crisis. Extractive histories also include moments of intensification, characterized by processes of “in-situ” restructuring that center on reducing production costs via technological change and renegotiating labor conditions (Bridge 2000; Ó hUallacháin and Matthews 1996). Finally, there is the temporality of abandonment around resource extraction when the process of “bust” involves more than a temporary withdrawal of capital and labor. While the cessation of resource extraction is often represented as an orderly process of closure and remediation, in which land is returned to a semi-natural state, extractive projects frequently have an “afterlife” that confounds the neat temporality of an extractive “life-cycle” from discovery to closure. Sandlos and Keeling (2013), for example, have shown how the long-term environmental impacts of extractive projects beyond closure – from the accumulations of toxic waste material to the challenges of acid mine drainage – lend resource extraction a “zombie” quality: long after production has ceased legacies of extraction continue to haunt communities.

Governing resource extraction

Resource extraction is a target of several different national and international policy concerns, from financial transparency and revenue management to sustainable livelihoods, global biodiversity, and the self-determination of indigenous peoples. Extraction has become a lightning rod for discontent with economic globalization and mainstream models of development. The elevation of practices of resource extraction to formal policy concerns reflects, in part, their politicization over many years by communities at the sharp end of resource development. An uneasy alliance of activists and social movements from the Global North and South have, for different reasons, challenged prevailing corporate and state practices around resource extraction. Often disadvantaged by proposals to mobilize natural resources in the interests of “national” economic growth, these groups have sought to reform the extractive sector via direct action, new national legislation and/or recourse to international agreements. The diversity of these concerns, the plurality of spaces and scales in which extractive activities are being challenged, and the modalities of struggle around extraction (including both resistance and adaptation to resource development, as well as forms of “neo-extractivism” that promote resource extraction for progressive social ends), suggest that the governance of resource extraction is being profoundly re-negotiated. Four significant trajectories of concern are discussed here that also reflect the evolution of geography’s engagement with resource extraction: from habitat conservation to livelihoods, from optimal rates of extraction to informed consent, from enclave economies to contested territorial imaginaries, and from “Dutch Disease” to transparency.

From habitat conservation to livelihoods

The impacts of extractive activities on the biological productivity and quality of environmental systems are widely documented. Air, water, and soil pollution, the bio-accumulation of metals and toxins mobilized during extraction, and the impacts of extraction on pastoral, forest, and
aquatic ecosystems are long-standing concerns of environmental policy. While the land area dedicated to extractive industries like mining and quarrying is relatively small, these activities generate “environmental multipliers” via (i) extractive infrastructure, such as access roads, which accelerates the circulation of capital and labor (research on Amazonian deforestation, for example, demonstrates how logging and mining impacts regional in-migration, settlement and forest conversion), and (ii) systemic impacts to biogeochemical cycles. The mining and combustion of coal, for example, significantly modify the cycling of mercury, sulfur, and carbon.

A concern with the disruptive effect of extraction on “nature” – expressed in efforts to improve the environmental performance of extractive industries or in research documenting resource extraction as a driver of land-cover change and/or species loss (Swenson et al. 2011) – has been supplemented over time with a growing interest in how extraction affects already-existing livelihoods. While this is not a new concern – some of the earliest examples of “environmental” legislation around resource extraction centered on conflicts between miners and farmers over the effects of metal smelting and waste disposal on agriculture (Kelley 1959) – it is increasingly acknowledged that policy challenges associated with resource extraction extend beyond the conservation of biodiversity or the management of physical material flows. Placing livelihoods at the center of analysis, rather than nature, reveals how resource extraction is also a process of dispossession, because of the particular ways in which extraction intersects with pre-existing economies. Dispossession may take place via the introduction of new rules on access to land (e.g., for small-scale mining, grazing, or forest products) or to water resources (important for watering livestock, irrigating crops or for catching fish); the assertion of formal property boundaries that cross-cut customary trade routes; or the sterilization of biotic marine resources via the dumping of waste materials or the construction of port facilities. Extractive projects can affect ongoing livelihoods by introducing and overlaying new forms of property (e.g., mine claims, forest concessions) on established rights and entitlements. Extractive property rights also are “technologies of rule” that enable land to be held speculatively, as an option for future development: they introduce logics, forms of discipline, and uncertainties for other land users prior to – and even in the absence of – raw material extraction (Peluso and Vandergeest 2001).

From optimal rates of extraction to informed consent

If extraction is understood primarily as a physical flow of raw materials, a central policy concern is the rate at which materials should be taken from the Earth and consumed within economies. How one should allocate scarce resources over time centers on the balance between extraction and resource conservation, and is one of resource economics’ classic questions. The so-called Hotelling rule, for example, provides a guide to how fast exhaustible resources like oil, tin, or gold should be extracted in order to maximize its value over time (Hotelling 1931). The desire to establish a socially optimal pace of resource development reflected a growing concern in the early twentieth century – articulated in the “conservation ethic” of Gifford Pinchot, the first Chief of the US Forest Service, among other leading figures in an emerging resource conservation movement – that the accelerating pace of timber cutting, oil field development, and minerals extraction would be ruinous for future generations. In a similar way, mid-century models of “sustainable yield” provided a guide
for the management of renewable resources like fish or groundwater.

The conservation–extraction dilemma frames resource decision-making as a technical question of how to optimize revenue from a fixed or variable stock. In so doing, it reserves decision-making for an elite group of scientifically trained resource managers. More recently, these traditional concerns with resource optimization have been supplemented with a broader set of questions about how scientific knowledge is mobilized in the interests of economic and political power; and about who should make critical decisions about whether, where, when, and how resource extraction takes place. At issue here are two related questions: the administrative scale at which decisions should be made (i.e., how “local” objections to resource extraction articulate with wider regional or national objectives); and the right of affected communities to exercise free, prior and informed consent (FPIC) over developments that may impact their wellbeing. The call for FPIC and its underpinning principle of self-determination have become a key demand of nongovernmental organizations affected by resource extraction as a way of addressing the historic disempowerment of indigenous and other marginalized social groups. The principle is outlined in both the United Nations Declaration on the Rights of Indigenous Peoples and in the International Labor Organization Convention 169. Its implementation has gone hand in hand with a growing use of community referenda that, whether sanctioned by the central state or not, provide a way for a community to register its consent (or lack thereof) to resource development proposals. Indeed, harnessing administrative scales below that of the national state has proven an effective strategy for communities seeking to challenge national institutions that promote extraction: between 2003 and 2008, for example, 7 out of 23 provinces in Argentina challenged the central state by banning open-pit metal mining (Walter and Martinez-Alier 2010). Others have noted how community referenda allow those opposing resource extraction to redistribute “the potential for empowerment” by linking locally scaled political and moral claims with organizational structures at national and international scales (Haarstad and Fløysand 2007, 289; Perreault 2006).

From enclave economies to contested territorial imaginaries

Resource extraction’s territorial form has been a long-standing concern for those seeking to harness resource extraction for national and regional development. The relative weakness of “forward” and “backward” linkages and a general orientation toward the export of raw materials mean that resource extraction has a tendency to produce enclave forms of development that are more strongly tied to external markets and centers of finance than to the host community or nation. The truncated territorial forms of development associated with extraction and raw material export occupied an earlier generation of scholarship grounded in the economic history of resource exporting regions (e.g., Innis 1956; North 1955; Bunker 1985). More recently, the imaginary of the “enclave” has been recycled to describe a particularly neoliberal modality of extractive development associated with both a minimal build-up of social infrastructure and flexible or “variegated” forms of sovereignty designed to facilitate the expansion of extractive capital (Ferguson 2006; Emel, Huber, and Makene 2011). In seeking to understand the link between resource extraction and territorial form, recent work focuses less on economic linkages and more on the ways in which historic and contemporary struggles over
RESOURCE EXTRACTION

extractive resources give shape to questions of citizenship, community, and identity (Perreault 2013; Valdivia 2008; Watts 2004).

From revenue management to transparency

Resource extraction poses a number of challenges for revenue management that belie the apparent simplicity of converting natural resource riches into social wealth. The scale of resource revenues can be difficult to absorb in states with limited infrastructure or consumption potential. The volatility of revenues over time, and the ways in which extractive exports interact with other forms of economic activity, can pose additional obstacles to achieving broad-based forms of growth that are economically sustainable. Historically these challenges have been framed as the problem of “Dutch Disease” (named after the counter-intuitive effects on the Dutch economy of developing Europe’s largest gas field in the 1960s), in which the revenues from extraction drive an appreciation of the domestic currency and a corresponding loss of competitiveness for manufacturing exports (Auty 2002). In this view, large-scale resource extraction and export can drive a process of de-industrialization in economies where manufacturing is already well-established. In economies with limited manufacturing and where the state is the principal recipient of large revenues from extraction – such as in the case of oil in Venezuela and Saudi Arabia or natural gas in Qatar – the state’s “rentier” character creates challenges not only for economic management but also political accountability. The dominance of resource rents in the state’s revenue stream, and the correspondingly limited role of general taxation, lends the rentier state a high degree of autonomy from its population. Such autonomy frequently expresses itself in authoritarian forms of government and political patronage.

These well-rehearsed policy concerns regarding the under-performance of extractive economies (Ross 2001), the “resource curse,” and the supposed “paradox of plenty” (Karl 1997) have been supplemented over the last decade or so by a growing recognition of the need for greater transparency and accountability in the management of resources, payments, and revenues associated with large-scale resource extraction. Calls for greater disclosure of the payments made by extractive companies to governments, from development-oriented nongovernmental organizations (NGOs) like Oxfam, Global Witness, and Publish What You Pay, contributed to establishment of the Extractive Industries Transparency Initiative (EITI) in 2002. This created a series of international standards, practices, and validation procedures that center on disclosure by host governments and investing corporations. High-profile projects like the Baku-Tblisi-Ceyhan oil pipeline from the Caspian to the Mediterranean confirm transparency’s role as political norm in an emerging liberal mode of resource governance. The politics of transparency go beyond placing information in the public domain, demanding the creation of publics able to bear witness to its operation and effects (Barry 2013).

Conclusion

During the first decade of the twenty-first century a commodity “super-cycle” helped push resource extraction to the center of public and scholarly attention. The intensification of logging, fishing, mining, and other extractive activities, and their geographical extension, created for those at the sharp end of extraction a series of ecological and political conditions
that have enabled popular mobilization around extractive projects. One result is a rich body of accumulated knowledge among activist networks and social movements that increasingly shapes corporate and state strategies for resource extraction. At the same time, overlapping concerns about looming resource scarcity, habitat conservation, the anthropogenic flux of carbon, dispossession of indigenous peoples, and the particular challenges of sustaining broad-based development from extractive economies have ensured resource extraction is a focal point for a range of national and international policies. The governance of extractive resources, then, is changing in significant ways, although the mobilization of raw materials from the Earth continues to grow in scale and complexity. Scholarly work on resource extraction has responded to these developments. On the one hand, the physical science base for understanding the scale of raw material extraction and its environmental and human health effects has grown considerably. On the other, critical social science analyses of resource extraction have identified how the political, economic, and environmental significance of resource extraction extends well beyond the physical mobilization of raw materials.

SEE ALSO: Energy resources and use; Mining and mineral resources; Natural resources development; Natural resources and human conflict; Resource curse; Resources and development; Uneven regional development

References


Further reading


The relationships between natural resources and development are profoundly important, but contingent and contradictory. Indeed, the evolutionary paths of different resource industries, and even different areas with similar resources, are remarkably diverse, and debates over the role of resources in development rage around the world. For some nations, resources have been an evolutionary stepladder for high incomes and diversified economies, while others seem cursed and trapped by their resource dependency. In Canada, with its long tradition of resource exploitation and of its theorization, these debates continue to engage different (optimistic and pessimistic) approaches (Gunton 2003). More evocatively, Canada’s Athabasca heavy oil deposits became both a vital national growth pole and national target over its environmental consequences and a polarization of views that fuels an oil pipeline extension controversy in the United States, where energy security and competition are also issues. Meanwhile from local perspectives, resource communities around the world grapple with the myriad problems associated with booms and busts while all levels of government seek approaches to parlay resource riches into long-run, place-based development (Markey, Halseth, and Mansen 2012).

The goals of resource exploitation have long focused on regional and national economic development and economic growth, conventionally measured by increases in per capita GDP (gross domestic product). In tandem, resource exports have long been emphasized as a driver of development for regional peripheries. Indeed, for new resource-based economic spaces, resource policy was industrial policy that privileged the commodification of nature and the industrial values of resources. In recent decades, however, this thinking has been questioned, even turned on its head, and criticized for its failure to cope with distributional inequities, quality of life, and environmental considerations. In this rethinking, the nonindustrial values of resources should define development rather than vice versa. The globalization of resource conflicts is a visceral indicator that the relationships between resources and development are in a transformative phase.

This entry’s objectives are to summarize the problematical implications of resource exploitation for economic diversification, and to discuss the implications of resource conflicts for development, particularly with respect to resource peripheries. The review is necessarily selective of the vast literature on these and related themes, within and beyond economic geography. That said, the entry focuses on fundamental perspectives and broaches new concerns. Economic geography’s long-standing interest in the human use of nature has been reinvigorated in recent decades from neoclassical, institutional, and Marxist perspectives, and from their offshoots.
such as political ecology, sustainable development, and environmental economic geography. This entry reflects an institutional perspective on environmental economic geography (Patchell and Hayter 2013).

In terms of format, the study begins with the interlocking concepts of resource cycles, governance, and multifunctionality. Second, we examine the role of resources as inputs to industrialization with respect to their diverse impacts on resource peripheries. Third, resource conflict and resolution are discussed in relation to a rethinking of the meaning of development.

**Key dilemmas of resource-based development: governance, multifunctionality, and cycles**

Relationships between resources and development are problematical. Their meanings and values are socially driven and judgmental, and vary over time and space and between people at particular times and places. Always intimately local, resource–development relationships have become global in scope, “officially” signaled by the United Nations (Brundtland 1987) following earlier alarms of environmental destruction. Resource and development relations are complicated because: resources are viewed as birthrights whose governance raises questions about public interest and the flow of benefits into the future; resource use offers multifunctional (industrial and nonindustrial) values that are often incompatible; resource cycles based on industrial exploitation generate significant positive and negative externalities; and – because resources are distributed according to nature’s endowment – they are often remote from where people need or want them. These themes are interrelated.

Resource endowments are derived from nature but are culturally defined. Social evaluations of nature typically evoke deeply rooted attachments and the bias that resources should be used or designated to enhance broadly based “local” (national, regional, community) development goals. Given that air, water, and food are necessities for life itself, the wise use of resources in general typically invokes reference to the public interest. This challenge of wise use became manifestly greater following the Industrial Revolution, which began in the late eighteenth century and heralded sustained increases in economic growth, population, and demand for resources. In tandem, the governments of leading nations prioritized economic growth through the spread of market institutions, industrialization, and trade. In this pursuit of the public interest as the wealth of nations, the privatization of resource ownership and/or control became widespread, often at the expense of common property resources. The search for productivity and new markets in turn stimulated processes of discovery and innovation that created new or diversified existing resource use. Indeed, even if it is not often recognized in contemporary literature, innovation, with its high-tech focus, has been a significant, often sophisticated impulse in the resource sector, creating new products while opening up new resource spaces and extending existing ones. In the case of oil, for example, the opening up of conventional and unconventional sources in deep sea and Arctic locations, and of oil sands and shale oil deposits, has depended on important technological improvements, as well as generating increased (financial and environmental) risks.

Yet, significant government involvement, inevitably justified as in the public interest, has proceeded, inter alia, with the privatization of resource exploitation (Bridge 2013). Governments remain vitally entangled in resource management, stimulated by their role as resource owners, by society’s demand for the wise use
of nature, by the need to adjudicate between competing diverse values, and by the social and economic externalities generated by industrial exploitation. Government stipulation of conditions on resource use has long-lasting effects. Among New World economies, for example, settlement was encouraged by the assumption of terra nullius and incentives to immigrants and business through stakeholder rights and land grant schemes that were deemed vital to economic development via resource exploitation. The recent explosion of oil shale fracking activity in the United States is based on subsurface rights awarded decades ago under privately owned farmland allocated even earlier. In the case of Canadian forests, past government decisions to retain public control over forests, even as they encouraged their private use, now help to stimulate broader debates over their wise use.

In recent decades, resource privatization is often criticized as part of a neoliberal agenda, not least in the context of water, the quintessential public good (Bakker 2010). Yet, when resource ownership and control have been privatized, governments have sought, perhaps increasingly so, to regulate on myriad issues in areas such as health, safety, and the environment. Even in resource peripheries with strong commitments to free enterprise and free trade, neoliberal attempts to privatize and deregulate resources to empower markets at the expense of community interests have limits (Hayter and Barnes 2012). Further, government interventions over resources are frequently pragmatic, reflect diverse motives, and include critical questions pertaining to water access (Agnew 2011). Interestingly, the strongly pro-market Canadian oil industry began requesting a stronger, more transparent commitment to the environment from the previous (pre-2016) recalcitrant federal government as a precondition for a stable business environment.

Moreover, resource privatization or nationalization can be reversed. As one interesting historical precedent, the privately developed British coalfields were nationalized for decades (1945–1987) before being privatized again. Less equivocally, the nationalization of oil resources has been a widespread trend around the world, beginning in the 1950s and 1960s, with state-owned companies now having larger reserves and output than established Western-based multinational corporations (MNCs), which remain larger in terms of market value. Moreover, to sustain biodiversity and offset global warming, national governments have agreed to the global expansion of conservation areas, many focused on Northern boreal forest regions (Affolderbach, Clapp, and Hayter 2012), in which private rights, for example those held by forest companies, have been removed or substantially modified. In many resource peripheries the empowerment of aboriginal rights is further complicating established notions of property rights.

In practice, resource governance features complex bundles of property rights, obligations, and limits on resource extraction whether the resources are fixed or mobile, land-based or beneath the surface. These resource rights define resource regimes that vary from place to place and refer to (formal) de jure and (informal) de facto rules of governance regarding resource use and transfer. Indeed, it is difficult to generalize about appropriate forms of resource property rights from a development perspective with one exception: open access resources that lack governance are widely condemned for encouraging overexploitation and stimulating negative externalities whether in the form of congestion on a beach, global climate warming, or fisheries depletion in open oceans.
**RESOURCES AND DEVELOPMENT**

**Resource cycles**

In response to industry’s escalating materials needs, export-driven resource exploitation has been pushed to all parts of the globe. In that process peripheral regions and communities have been created and subject to patterns of development and decline shaped by resource cycle dynamics. Resource depletion is an obvious reason for decline, but competition from other places and substitution by new materials and technologies also threaten resource-based development. While the distinction between renewable and nonrenewable resources is a basic dichotomy, defined by the implications of depletion and renewal, the sustainability of the former is not necessarily inevitable, or likely. If finite mineral deposits will inevitably be depleted, fish and forest resources similarly experience resource (S-shaped) cycles of discovery, boom, bust and collapse, or rationalization. In practice, the evolutionary trajectories of resource cycles vary. Typically reliant on distant, specialized markets, resource cycles are punctuated by short-term business fluctuations that in turn complicate longer-term trends and understanding of fundamental turning points. Indeed, the trajectories of resource industries during lengthy periods of postmaturity can be geographically and temporally uneven, with highly varying entry and exit rates of firms among subindustries, as illustrated by British Columbia’s forest sector (Edenhoffer and Hayter 2013).

Initiation and expansion of resource exploitation occurs through many configurations of firms with different size, networking, and ownership characteristics. Irrespective of these configurations, all firms must respond to market demands and with repercussions for the viability of production facilities and communities. Although most closely interrelated with the community, small local firms are the most fragile in terms of dependence on distant markets because they often lack the capacity to ride out market fluctuations or to diversify. Larger firms may have such capacity and may be able to switch resource locations. In either case, given concerns about community or regional welfare, governments typically play a large role in resource development.

The developmental role of governments varies during resource cycles. In the early stages and the opening of new resource spaces, governments facilitate growth by providing attractive governance conditions, low taxation or royalty payments, incentives for exploration and testing, massive investments in infrastructure, and support for research and development. In the nineteenth century, in Europe and the United States, local wealth and relative proximity often allowed the private funding of transportation for access to resources. In Canada, with a sparse population spread over vast distances, governments have needed to fund and often control transportation networks up to the present (Gunton 2003). Canadian governments also became increasingly involved in providing social or community infrastructure. Early resource towns were literally often company towns where the corporation provided housing, services, and local administration. After 1945, however, new resource towns with their own municipal powers were planned by senior levels of government, beginning when Kitimat in a remote part of British Columbia’s northwest coast was chosen as a location for an aluminum smelter.

In some resource sectors, especially where producers are small-scale, such as agriculture and forestry, governments have long played key roles in developing and transferring technology through networks of specialized laboratory operations. Recently, governments have been closely engaged with the private sector in developing renewable energy technologies, for example in...
In the late stages of resource cycles government attention becomes more focused on problems of decline, especially in regions and communities that remain resource-dependent. Declining coal mining and depressed rural areas in Europe and North America gave birth to the original regional problem areas. They stimulated the introduction of regional development policies in the 1930s and expansion in the 1950s. In practice, the challenges of rejuvenating resource communities in these regions have proven enduring while in recent decades the number of declining resource towns has increased. In the decline stage, the established community often adds local perspectives on development plans that are “placed-based,” highly varied, and made in the context of changing social structures (Markey, Halseth, and Mansen 2012). To pre-empt the building of communities only to see them painfully closed, companies and governments have developed a strategy of using fly-in fly-out workforces for resource projects in new spaces. Proponents are spared significant sunk costs and future policy challenges while providing companies with employment (albeit high-wage) flexibility.

Resource exports and regional (community) development

Can resource exploitation generate long-term development? In evolutionary explanations of capitalist development over the last 250 years (Patchell and Hayter 2013) resource inputs have been recognized as significant to, but not a primary causal factor of, industrialization. Rather, “paradigmatic” transformations evolve through interdependent institutional developments in science, technology, society, politics, and the markets that are favorable to entrepreneurship and innovation.

That said, resources nevertheless played a role in the development of industrial cores, if this is not always appreciated. For example, referring to a hearth of the Industrial Revolution, Potter and Watts (2010) interpret the evolutionary trajectory of Sheffield’s metalworking activities (steel, cutlery, and tools) as the quintessential example of localized Marshallian (positive and negative) externalities in the form of local skilled labor pools, suppliers, and knowledge spillovers. In their model, without reference to dates or periods (or institutions), “Embryonic agglomeration … starts with a mutation process … in which Geography is crucial” (Potter and Watts 2010, 421–422), by which they mean various knowledge-based advantages that accrue to pioneering new and growing firms from clustering together. However, Sheffield’s case also reveals the importance of local access to resources during the embryonic stage that culminated in Benjamin Huntsman’s innovation of crucible steel in the 1760s. Thus Sheffield’s dispersed metalworking sites enjoyed close access to several key industrial resources (charcoal, coal, iron, clays, sandstones, millstone grit and limestone, plus local agriculture), while the rivers and streams that crossed the city once powered the largest concentration of water wheels in Europe (and were organized along common property lines). Huntsman himself relocated to Sheffield to access these resources and conduct his experiments in relative secrecy.

Elsewhere, the evolution of North American industrial agglomerations pointed to resource access as part of their “initial advantages” that generated the continent’s industrial heartland. In Europe, Ruhr Valley coal deposits helped establish Germany’s industrial prowess. China’s steel mills and heavy industry are located near the coal and mineral deposits of the northern
RESOURCES AND DEVELOPMENT

provinces. In contrast, major steel and heavy industry locations in Korea and Japan relied on imported raw materials. Now, whether local supplies have never existed or have been exhausted, core regions and cities see resources as inputs to diversified ranges of manufacturing goods that can be imported. For new resource-based economic spaces, however, resource exports are their rationale and the specter of declining resource cycles, however implicit and vague, has stimulated resource peripheries to contemplate how to use resource exports to diversify and achieve self-sustaining growth (Auty 2001).

Resource periphery diversification

Understanding the role of resource exports in regional and urban diversification is sensitive to scale and context. At root, however, the implications of export-based resource exploitation for local economic development and diversification has long been conceptualized as a process of local linkage development within poor and rich country contexts and in densely populated and empty resource spaces (Baldwin 1956). Within geography, mesoscale studies have elaborated these linkage-based models, for example, to explain the early development of California (Walker 2001) and recent restructurings in British Columbia (Gunton 2003).

In broad perspective, the “virtuous” diversification of new resource-based economic spaces is rooted in the creation of backward, forward, fiscal, and final demand linkages. Backward linkages are investments in activities that provide inputs to the resource sector such as machinery, components, materials, wholesaling functions, research and development and engineering, decision-making and related services, and more broadly to the establishment of transportation and community infrastructure. Forward linkages add value to resource sector outputs (while generating further backward linkages); fiscal linkages are payments to governments from the resource sector; and final demand linkages are created by the demands by (growing populations of) local consumers for goods and services. Geographically, successful linkage diversification around export-based resource sectors is associated with urbanization and the generation of core–periphery relations, especially at national and regional scales. Potentially, related linkage development and Marshallian-type localization economies (Potter and Watts 2010) can stimulate more broadly based urbanization economies that help cities and regions diversify in new, creative directions and related localization economies.

California’s experience illustrates the extent to which urban and regional diversification can draw on and move well beyond resource dependence (Walker 2001). Meanwhile, if questions can be raised about sustainability, the growth of Calgary and Edmonton, Canada’s most dynamic metropolitan areas of recent decades, reflects the continuing urbanizing power of resource-based linkages.

In general, resource periphery diversification is related to integration into rich market economies, and is shaped by the establishment of local entrepreneurial and labor pools, by government policies for stimulating resource cycles, and by their resource endowments. Thus, the ability to diversify and urbanize around a resource base has geographic foundations, most obviously in the richness, scope, and variety of resources, their costs of exploitation, and access to markets. Extensive, high-quality, diverse resources facilitate economies of scale and scope, rival investments, cumulative advantages, stronger justifications for infrastructural investments, new resource discoveries, and intersectoral innovation, while growing technological expertise can create processing efficiencies that offset higher costs of exploiting less accessible, lower-quality
resources. Innovative activity has also often been stimulated by locally distinct resource conditions (Walker 2001).

Regional and national contexts and situations further shape urban and regional diversification. Among rich Western economies, for example, Canadian and Scandinavian approaches to export-driven forest exploitation evolved differently. The latter emphasized forest privatization, domestic companies, and in-house research and development (R&D) investments that have collectively stimulated stronger export-oriented forward (e.g., furniture) and backward linkages (e.g., machinery). In Canada, forests are publicly owned by provincial governments, foreign direct investment (FDI) is significant, government R&D is important, and linked developments are less export-oriented. In the oil sector, while Norway exerts control through a state-owned company (SOC), Statoil, and taxes it heavily (78%) to fund a national heritage fund (over US$900 billion in 2014), Alberta welcomes foreign ownership and prefers low taxes and a stronger role for the private sector in reinvesting profits. Its heritage fund of US$16.7 billion in 2014 is relatively small. If Alberta’s oil industry’s backward linkages are considerable, provincially and nationally, its forward and fiscal linkages are modest, as illustrated by debates to add refining capacity prior to export.

In part, the Albertan reticence to demand larger oil rents reflects geographically ingrained geopolitical realism. Alberta’s political and business community has been committed to North American “continentalism” which casts Canada as a resource supplier to the United States as a major market and hegemonic power. This role counsels against government interference over profits or industrial organization. However, Canada’s resource exports, for such a large country, are remarkably dependent on US consumption, rendering it vulnerable to US trade protectionism and policy shifts.

Resource enclaves and policy debates

Local linkage generation around resource exports is not automatic but can be arrested. Resource enclaves are those peripheral regions and communities that struggle to escape from narrowly defined resource dependency, and whose linkages are largely generated elsewhere. Enclaves include mining operations, agricultural plantations, logging camps, forest product complexes, and oil fields that are characterized by remoteness, isolation, control by distant decision-makers, reliance on imported technology, and exports of low-value commodities to be processed in cores. Enclaves involve the branch plants of MNCs (or government agencies), and their viability in particular locations is closely tied to resource cycle dynamics and exploitation, and also to the discretionary powers of MNCs (or governments) that have location options. Enclave operations are mandated by organizational dicta with regards to scale, commodity mix, and markets, with any expansion or restructuring decided by head offices and related to broader organizational strategies. Vertically integrated MNCs may further enjoy discretionary powers over pricing such as imposing low internal transfer prices to reduce payments of taxes in peripheral regions.

The challenges to the long-run diversification of enclaves are geographically widespread. Thus the underlying dilemmas are referred to as a staple trap in Canada, addictive behavior among American resource communities, and a resource curse among poor countries (Auty 2001) – metaphors that readily connect with the functional, cognitive, and political lock-ins of path-dependent behavior. There is considerable debate on the interrelated causative factors
that conspire to restrict enclave diversification. These may be summarized broadly in terms of geographic, institutional, and boom and bust perspectives. In this context, geography implies site (endowment) and situational factors that shape resource cycles. Markets have enormous centripetal tendencies, and the ability of resource towns and peripheries to diversify may be limited by their small size, isolation, increasing costs, resource exhaustion, vulnerability to competition, and possibilities of substituting other inputs. However, whether or not resource peripheries inevitably suffer declining terms of trade is debated.

“Staple traps” and “resource addiction” are virtual synonyms with respect to institutional constraints on resource diversification. The terms describe narrowly based competencies, capital and human inertias, and conservative attitudes among corporations, governments, and communities that mutually reinforce resource dependence. Thus resource-based corporations tend to be powerfully locked into established strategies, constrained by highly specialized managerial expertise, immobile resources, dedicated machinery, and the expense of large-scale investments to modernize operations and maintain market share. Further, the implications of flexible mass production in resource industries are variable, and often not an option. It is worth recalling that the local role of MNCs in resource exploitation has long been criticized, not only within radical literatures such as dependency theory but also in Baldwin’s (1956) neoclassical analysis. In Baldwin’s comparison of the production functions of MNCs with local entrepreneurs, the resource plantations of the former were less likely to diversify local economies than local pools of entrepreneurial risk-takers who reinvest profits locally, purchase locally, and stimulate external economies and more accessible infrastructure. Nevertheless, FDI in resource exploitation is an important influence, increasingly originating within developing countries. Resource-based and global production networks are typically orchestrated by large organizations with influence over the location of value-adding activities.

Governments can perpetuate resource dependency, not only by investing in dedicated infrastructure and incentives, but also through bureaucracies and politics ingrained in the primacy of resource exports as the growth mechanism. As noted, resource peripheries tend to be strong supporters of free trade agendas even when trade agreements counter their interests. If poor countries have become more skeptical in this regard, domestic political conditions can pose powerful limits on resource diversification. In Southeast Asia, for example, the patron–client model of governance has facilitated deforestation and log exports as much as the behavior of Japanese MNCs. More widely, corruption and political instability in poor countries directly and indirectly stimulate tendencies to exploit resources quickly while dissuading investments diversify economies.

With respect to how the boom and bust of resource cycles thwart local diversification, the resource curse, or Dutch disease, has received much attention. Originally developed in the context of oil exports from the Netherlands, the curse now mainly refers to poor resource-specialized countries whose booming resource exports increase local currency values to render imports cheaper than local supplies (Auty 2001). Further, the ability of resource sectors to pay high wages and high prices for land and other local inputs can crowd out other endeavors, including at the resource-town level (Argent 2014). Even without the Dutch disease, business cycles can reduce local diversification as local governments and labor unions are uninterested in attracting alternative businesses in boom times
and unable or unwilling to seek alternatives during recessions. Unfortunately, when permanent downsizing occurs resource towns are frequently unprepared, and start contemplating diversification only when financial resources have become limited, labor pools specialized, and competition for alternative opportunities considerable. How rates of resource cycle growth relate to diversification is an interesting question. If too fast growth implies shortened cycles and Dutch disease effects, too slow growth may fail to generate positive externalities and attract investment.

Redefining resources: conflict or compromise

The multifunctionality of resources has increasingly imposed highly contested choices on developmental paths. The crux of the issue is whether conventional resource exploitation can coexist with increasing demands for ecological preservation. The first perspective emphasizes the jobs and rents directly generated by resource exploitation, and the second seeks to control exploitation while promoting alternative (nonindustrial) resource values with different consequences for economic development. Economic benefits such as ecosystem services or ecotourism have been demonstrated, but they differ fundamentally from those generated by the resource industry. Moreover, the empowerment of aboriginal rights is complicating property rights in resource peripheries around the world, from poor developing countries to rich economies such as Australia, Canada, and New Zealand. Aboriginal rights invoke self-determination and self-governance as aboriginal peoples seek their own industrial and ecological goals.

Stakeholder diversity and analysis

In practice, resource conflicts feature diverse, internally differentiated groups and evolving alliances between governments, labor, local and international businesses, community economic leaders, local and international nongovernmental organizations (NGOs), aboriginal peoples, and distant consumers. The conflicts involve political processes, civil disobedience, litigation and court challenges, intensive public relation campaigns, and what Clapp and Mortenson (2011) refer to as “adversarial science.” For so-called conflict minerals, confrontations are violent, involving myriad cultural allegiances and armed groups, official and otherwise, and development goals are illusory (Le Billon and Cervantes 2009). In general, even in violent places, environmental and cultural arguments seek to systemically rethink the resource–development nexus and the global role of peripheries. If these debates focus on resource peripheries they crisscross national borders and reflect a search for (new) forms of development that incorporate new global as well as local imperatives.

Resource conflicts have a long history but have escalated within contemporary globalization processes, driven by society’s heightened ecological and cultural concerns, debates over whether free trade is fair, and emergent community and business interests. New information technologies have provided the key infrastructure to mobilize opposition to industrial interests and to shape public opinion. Resource conflicts are widespread and contingent. Hundreds of forestry conflicts have occurred around the world, for example. In one case, British Columbia’s durable “war in the woods” comprises trade, environmental, and aboriginal conflicts, all conjointly sparked in the severe recession of the early 1980s (Hayter 2003). The main protagonists – US trade protectionists, environmental nongovernmental organizations (ENGOs), and Aboriginal
peoples – have different interests but have cooperated to oppose vested industrial interests, and to stimulate profound changes in the rules of the game for British Columbia’s forest industry.

Analytically, stakeholder models have been used to conceptually frame resource conflicts and to identify relevant stakeholders and their diverse powers, economic and noneconomic motivations, global–local connections, and relationships to one another (Affolderbach, Clapp, and Hayter 2012; Hayter and Barnes 2012). Predominantly, stakeholder models have been elaborated in democratic, pluralist market economies that tolerate social action and criticism, and where protesting stakeholders are striving for legal and political acknowledgment of the resource values they promote. In the United States, for example, the idea of a public trust doctrine for natural resources such as oceans, shorelines, air, and (some) land, has been pioneered since 1970 when Michigan was the first state to legally adopt this principle, with other states following. This doctrine itself is part of the trend toward the adoption of environmental legislation at the state scale, hitherto most effective at national and local levels, which has directly and indirectly imposed conditions on or prevented industrial resource use. Aboriginal protests around the world are similarly driven to have aboriginal rights recognized and enforced (Mander and Tauli-Corpuz 2006).

Ecological and cultural imperatives driving resource conflicts are hard to reconcile with export commodity interests that are often allied with geopolitical concerns for supply security and stability. Indeed, conventional cost–benefit frameworks, and market solutions have limited potentials for resolving such conflicts. As an alternative, stakeholder approaches emphasize the importance of learning and bargaining processes among stakeholders in specific contexts. Bargaining (and learning) can evolve from confrontation when participants recognize that their powers are limited by the desire for social legitimacy and when some common ground or currency is identified – which in forestry can literally mean land (Afforderbach, Clapp, and Hayter 2012). Learning (and bargaining) can be enhanced by exchanging adversarial for cooperative science, for example, by the creation of boundary organizations that help coordinate scientific expertise from different sources into an acceptable form such as a multifaceted geographical information system (GIS). Boundary organizations also point to the role of social innovations in conflict resolution.

In this regard, the idea of a social license to operate (SLO) is an important initiative that has gathered momentum since the 1990s, especially in mining. SLOs, a form of corporate social responsibility and recommended by the World Bank, emphasize dialogue with community and environmental interests to ensure environmental sustainability and equitable flows of local benefits. SLOs encourage senior governments and corporations to incorporate all community and environmental interests in negotiations for new projects from the beginning to build trust and achieve less political risk to business in exchange for sustainable community development. Counterparts to SLOs are the various forms of small-scale community developments that have evolved in poor peripheries, in part as elaborations of intermediate and appropriate technology themes.

Emergent processes and values

Given extraordinary geographic variations, the relationships between resources and development are being transformed in terms of decision-making processes, purpose, and outcomes. Decisions over resource exploitation were long dominated by industrial interests, especially
large corporations and big government, often with the tacit support of labor. Increasingly, however, resource decision-making is pluralistic and contested, and engages multiple institutional perspectives that crisscross geographic scales. In tandem, environmental and cultural impact studies have become routine parts of these processes in many peripheries. Moreover, the “emerging geographies” (Moorcroft and Adams 2014) of these decisions are privileging environmental and cultural resource values and the local redistribution of benefits and rents (Argent 2014) to a greater degree than in the past. The global expansion of conservation areas, the associated designation of ecologically and culturally based world heritage sites, the remapping of resource peripheries, and the spread of environmental certification and SLOs reflect this rethinking and reregulation of resource peripheries (Affolderbach, Clapp, and Hayter 2012; Hayter 2003; Zimmerer 2006). The increased emphasis on tourism, ecotourism, and ecosystem services are related trends that offer employment alternatives to the industrialization of resources.

Moreover, pressures on resource peripheries to give greater priority to environmental values are strengthening, stimulated by the recognition that nature’s ecosystem benefits are huge; Costanza et al. (2014) have estimated that ecosystem services amounted to US$142.7 trillion globally in 2014, bigger than global GDP, and calculations can be made at regional levels as well. Such estimates are controversial, not least within the environmental movement, and the benefits identified apply abstractly to “everyone,” while incorporating intangible development goals related to quality of life considerations and moral obligations. Nevertheless calculations of ecosystem services are escalating concerns over global climate change that is both altering biotic resource endowments and raising the intensity of debates over resource use, especially with respect to energy (Bridge et al. 2012). With the major exception of the Montreal Protocol, however, global governance of climate change has proven elusive. It may be that these negotiations focus too much on the energy sector itself and need to shift to emphasizing MNCs in end product industries without vested ties to conventional energy sources. In the meantime, an important question is whether continued resource conflicts lead to desired long-run outcomes. Conflict may be necessary to provoke necessary changes in resource–development relationships but conflict solutions require consensus and involve forms of global social cooperation that are diverse and complementary.

Conclusion

In recent decades a profound shift has occurred in social attitudes to resources. Resource peripheries are being asked by a host of diverse local and global interests to serve an increasingly wide range of economic, cultural, and environmental benefits. No longer are they seen as just sites for specialized (but marginal) inputs drawn from peripheral places to supply industrial cores, which are readily substituted and cast aside, whether or not communities develop around them. Rather, according to the green paradigm or sustainability transition models, resource territories need to sustain values that link the local with the global in environmental, cultural, and quality of life as well as economic terms. In understanding this transformation in resource–development relationships, geography, with its inherent appreciation of variation and the need to integrate economic, political, social, and environmental perspectives, has an important role to play.
RESOURCES AND DEVELOPMENT

SEE ALSO: Community-based natural resource management; Conservation and capitalism; Construction of nature; Cores and peripheries; Corporate environmental responsibility; Development; Ecological footprint; Economic geography; Ecosystem services; Energy resources and use; Environment and development; Environment and resources, political economy of; Environmental degradation; Environmental governance; Environmental impact assessment; Environmentalism, grassroots; Geography and the study of human–environment relations; Neoliberalism and the environment; Political ecology; Regional development models; Regional development policies; Resource curse; Resource extraction; Water conflicts

References


Restructuring

Elliot Siemiatycki
York University, Canada

Restructuring is one of the most important, and contentious, terms in urban, economic, and labor geography. The dictionary defines restructuring straightforwardly as “the process of organizing a system or business in a different way.” Yet this term is shrouded in confusion and ambiguity because of the rather important word embedded within it. Restructuring is a direct extension of the word “structure,” which in this context is defined as “the arrangement or interrelation of all the parts of a whole.” This complex relationship between structure and restructuring implies that a system or business has an initial form which, for some reason, undergoes a change. Indeed, as John Lovering commented in one of the key contributions to the geography literature on this topic, the term restructuring “suggests a qualitative change from one state, or pattern of organization, to another” (Lovering 1989, 218).

Restructuring happens in many different realms of life and is studied in academic disciplines ranging from biology to psychology and sociology. In all cases, the study of restructuring is about the causes and consequences of change for a given system or structure. The focus of this entry is on the way in which restructuring has been treated by human geographers. For the most part, restructuring has been examined as it relates to the economic sphere and with particular attention to the changing locations of economic activities. Three major questions are the focus of this entry: (i) what are the different types of economic restructuring; (ii) how have geographers studied these different types of restructuring; and (iii) what do geographers think causes restructuring of systems and organizations.

Types of restructuring

There are four main types of economic restructuring which refer to different, though often related, processes. The terms used here to describe these different types of restructuring are descriptive terms that have sometimes, though not exclusively, been used in the literature to examine the various forms of economic restructuring.

Sectoral restructuring refers to the changing structure of a given economy in terms of its key sectors. Economists have long divided economies into different activities, though perhaps the first formal example was British economist Colin Clark’s three-sector hypothesis (Clark 1951). The primary sector consists of agriculture and resource extraction. The secondary sector consists of those activities, namely construction and manufacturing, through which raw materials are transformed into products that can be consumed, such as bread and cars.

The tertiary sector is a broad term that is usually synonymous with the service sector. Since this last sector includes highly dissimilar services, such as babysitting and banking, in more recent times economists have identified a quaternary sector that involves finance, business services (including law, accounting, management consulting, etc.), and public administration which help to create markets. This leaves the tertiary sector as
RESTRUCTURING

consisting of transportation and infrastructural services, wholesale trade and distribution, and retail services.

This framework enables economic thinkers to compare the structures of different economies across space and time. Sectoral restructuring, then, is about the transition points when a given economy (at whatever scale) shifts from one dominant economic sector to another. Many economists follow Clark (1951) in arguing that economic growth naturally involves the evolution of economies between these three sectors. According to the economist Michael Porter (1990), David Ricardo’s classical treatment of comparative advantage is not a static outcome but rather economic entities ranging from the corporation to the city to the nation-state often need to change their economic orientations to suit new environmental, technological, or market realities. Economic geographers have played a key role in detailing these sectoral shifts and have developed alternative theories of how and why urban, regional, and national economies evolve over time.

The second form of economic restructuring surveyed here is industry restructuring. As will be seen, it is important to note the differences between an economic sector and an industry. A sector is a broader category that tries to compartmentalize various economic activities based on the type of production being carried out. An industry, on the other hand, refers to the many parts of producing a single good or service, and thus an industry often crosses the four main sectors. For example, the agricultural industry (or agribusiness) involves agriculture, food processing, food distribution, food retail, and the financing of all these activities.

Industry restructuring refers to the changing business strategies, locational decisions, market trends, and technologies that shape the production process for a certain good or service. Economic geographers have been particularly attentive to this form of restructuring, as it enables a closer view of how the owners of capital make locational investment decisions (Bradbury 1985). Often in the early literature the terms sectoral and industrial restructuring were conflated, but geographers were really concerned about the fate of certain regions that were facing difficult economic times.

The third type of economic restructuring is closely related to the previous two as it underscores the ways in which cities are often disproportionately impacted by economic change. Urban restructuring refers to the changing structure of the city in terms of its physical and social morphology. Morphology is a term that comes from the biological sciences; it describes the relative shape, size, and anatomy of a cell. Similar to how biologists might talk about cell morphology in terms of the location of the nucleus relative to the mitochondria, geographers have discussed the relative locations of different physical and social parts of the city.

Geographers have developed different models of urban morphology to understand where certain economic activities are located, where various social classes or ethnic groups tend to live, how natural resources may impact urban growth, and the relationship between the city and its peripheries. Often these models have used an exemplar of a certain form of urban morphology: Chicago as the industrial city, Los Angeles as the postmodern city, New York or Tokyo as the global city. These models have been the subject of serious debate and disagreement, but they have been useful in that they provide a comparative framework to understand the different form and functions of cities around the world.

Again, however, the past half century has been one in which the structure, feel, and function of many cities has changed significantly. Some cities have grown considerably, while others have
withered. Some cities are defined by the same basic industries while many others have seen their economic sectors and industries change. Urban restructuring has been a key process that contemporary economic and urban geographers have tracked across space and time.

The fourth, and final, type of economic restructuring is one that has received relatively less attention from economic geographers. Corporate restructuring refers to the changing legal, organizational, and operational functions of a corporation. Like the other forms of economic restructuring discussed thus far, corporate restructuring involves a change to the basic structure of a corporation or company. Often, corporate restructurings take place when companies are in a distressed position, such as mounting debts or bankruptcy proceedings. Other forms of corporate restructuring include mergers and acquisitions, market repositioning or market exit. In his studies of corporate restructuring in the food retail business, Neil Wrigley (1999, 288) acknowledged that the field of economic geography “has traditionally underemphasized types of restructuring that involve transformations of the capital structure of firms.”

While corporate restructuring has been studied in more depth by business and legal scholars, economic geographers have made important contributions to the literature. Gordon Clark intimated the increasing “financialization” of corporate strategy, where operations are valued less for their productive capabilities and more for their asset potentials (Clark and Wójcik 2007). Social commentators today have extended that very point to argue that markets have become inherently more volatile and unstable due to the rise of what has been termed “vulture capitalism,” an economy in which corporate restructuring and strategic liquidation of assets becomes endemic. As “financialization” becomes an ever more important theme in economic geography research, it can safely be assumed that corporate restructuring will become a more prominent part of disciplinary debates.

Now that the different types of economic restructuring have been detailed, the ways in which geographers have studied these processes can be examined.

Economic geographies of restructuring

Restructuring has been studied in different ways by geographers, reflecting the contemporary diversity of the field. First and foremost, it has been one of the main political and economic phenomena studied by urban and economic geographers since the late 1970s. This is the use of the term implied by the above discussion on the different types of restructuring. At the same time, restructuring has also been conceived by some as an “approach” to economic geography research (Lovering 1989). In this application of the term, the restructuring approach refers to the complicated emergence of Marxist thinking in economic geography during the 1980s. To be sure, these two different dimensions of restructuring were intimately related in that the rise (and ultimate decline) of the restructuring approach was connected to the academic interest in restructuring as an economic phenomenon.

The first major studies of restructuring as an economic phenomenon by geographers were focused on sectoral and industrial restructuring occurring in the Global North. Beginning in the 1970s, many regions that had been dominated by manufacturing activity began to see job losses and plant closures (Massey 1984; Peet 1983). Perhaps the most important term related to restructuring at this time was deindustrialization (Bluestone and Harrison 1982). Economic geographers of the 1970s and 1980s were preoccupied with explaining why traditionally dominant industrial
Restructuring

regions were losing their primary and secondary sector activities such as mining, shipbuilding or automobile assembly. Manufacturing, especially, had been the engine of industrialization for countless Western nations and yet very quickly the stability of economic, social, and political structures was being compromised.

By the 1980s, a specifically urban variant of these issues was starting to take focus in geography and this would mark an important contribution to the literature. Ever since David Ricardo used the examples of cloth and wine production in England and Portugal to explain his notion of comparative advantage, economists have mostly focused on restructuring at the scale of the nation-state. Yet it was becoming clear that city-regions were differentially impacted by restructuring forces depending not only on their dominant industries but also on other factors such as demographics, rates of unionization, and institutional capacity to attract capital investment (Soja, Morales, and Wolff 1983). While “industrial” cities such as Detroit and Manchester may have been hemorrhaging jobs, so-called postindustrial cities such as Los Angeles and London were assuming new positions of power in a hyper-globalizing economy.

What would come to be known as the “restructuring approach” really emerged out of these issues in the mid-1980s as economic geographers obsessed not only over explaining the problems of Western “industrial” cities but also what might be done about them. The “restructuring approach” was developed by Marxist economic geographers that were involved in the “locality debates” of the 1980s (Cooke 1989). The locality debates refer to a set of intense and highly contentious studies of the impact that increasing global competition might be having on local economies.

In the context of widespread economic upheaval and conservative political sentiments in the United States and United Kingdom, which viewed industrial workers as spoiled and entitled, scholars argued that capitalist economies were prone to regular bouts of restructuring as profit levels declined and new markets opened up. Instead of a momentary lapse in economic equilibrium, this moment of economic crisis was seen as an inherent part of the capitalist system.

Analytically, much of this research focused on the inner workings of multinational corporations (MNCs) and their locational decisions. Massey’s (1984) conception of the “spatial division of labor” identified firm-level decisions as being strategically constructed based on the potential profitability of different locations. With the rise of information and communications technology, it became technically feasible for the most profit-maximizing worker at each stage of production to be used no matter what part of the world they inhabited. The division of labor for a product or service was moving beyond the confines of a single workplace and was increasingly becoming part of a “global supply chain” (Sturgeon, Van Biesebroeck, and Gereffi 2008). In some instances, workers were protected from globalization based on their skills, tradition of innovation, or local rootedness of employment clusters. But in many cases, especially for those workers in the secondary sector, there was very little protection against manufacturing workers in the newly industrializing countries of East Asia who could do similar work at a fraction of the cost.

As radical economic geographers were developing a “restructuring approach” to understand urban industrial decline in the Western world, other economic geographers were examining economic restructuring and rapid industrialization at the nation-state scale in emerging economies of East Asia, Central and South America, Eastern Europe, and parts of Africa.
(Dicken 1998). For Dicken and other geographers interested in the new international division of labor (NIDL) it was not a coincidence that the moment of sectoral restructuring and deindustrialization in the 1980s, which so concerned Marxist economic geographers, corresponded with the rise of the Asian “tiger” economies. Faced with the challenge of competing on cost, many Western countries and city-regions shifted their economic orientations away from the secondary sector and toward the tertiary and quaternary sectors where they had an apparent advantage in “knowledge work.” Cities in the Western world have repurposed industrial-era factories and homes to fit a reconstructed “creative city” defined by knowledge work in office towers and creative districts, leisure opportunities in retail and entertainment districts, and an emphasis on urban design and architecture.

According to scholars of the “restructuring approach,” the transformation of sectors and industries as well as national and urban economies was one of the main costs of a dynamic capitalist system driven by the profit motive and competition between workers in different places. Depending on the nature and history of local and national economies, certain places would be disproportionately deprived by broader processes of economic restructuring while some places would be left alone or even benefit from restructuring. Beyond these very basic conclusions, however, scholars could agree on very little. More strict Marxists saw the restructuring approach as it related to the locality debates as a retreat from theory building and a return to descriptive empiricism in economic geography. For others, the policy implications of trying to engineer local economic revitalization plans at the urban scale were problematic (Lovering 1989).

More than two decades after these debates took place, what has become clear is that different forms of restructuring have had a dramatic impact on national, regional, and urban economic fortunes. The restructuring of places has certainly not stopped since the locality debates withered in the 1990s, and, if anything, the rise of China, India, Russia, Brazil, and other developing economies means that restructuring is happening on an ever larger scale and impacting greater numbers of people. It is more urgent now than ever before to understand the root causes of economic restructuring.

The causes of restructuring

Why places and companies change their economic orientations is perhaps the oldest and most important question in economics. According to mainstream economic thinking, there are two fundamental mechanisms through which restructuring occurs: Ricardo’s comparative advantage explains restructuring of the space economy and Schumpeter’s creative destruction explains how companies try to profit from the introduction of new goods, services, and production processes.

Some geographers, such as Dicken (1998), note that places tend to follow their own comparative advantage to improve economic prosperity in ways originally described by David Ricardo. Poorer countries can excel by extracting their natural resources or by developing manufacturing sectors that enable their citizens to contribute to the overall production of goods and services. This model of development has shown its benefits most clearly in East Asia, where economic performance and living standards have improved dramatically since the 1970s. Yet as Dicken recognizes, the industrialization process provides no guarantee of economic success. Natural
resource markets are highly unstable and prone to rent-seeking behavior by elites, contributing to what has been termed the “resource curse.” Similarly, many countries in Latin America, the Caribbean, and Africa have attempted to establish manufacturing sectors with very little success, due in part to excessive competition from other industrializing countries.

For Marxist economic geographers, however, an understanding of the structure of an economy must come before restructuring can be evaluated. Theories of social structure have come to be defined by notions of uneven development that work at a variety of different scales. Smith (1984) used the metaphor of a “see-saw movement of capital investment” to explain geographically uneven development. Corporate investment decisions are nothing more than strategic attempts to discipline workers and reduce labor costs in order to find the location that will produce the highest profits. But this is a restless process because, in time, workers tend to demand their fair share of profits and this means that capital often goes back to the places it previously abandoned. There is always the hope of economic redemption in a capitalist system because places and people are most attractive when they are at their most desperate.

The mainstream and Marxist view of economic restructuring may appear to be very different but, in actuality, they are describing the same processes from very different ideological perspectives. Both approaches see competition between places and companies as the key mechanism that causes different forms of restructuring. Mainstream market economists view this mechanism as the source of efficient resource allocation and technological innovation. When goods and services are produced more cheaply, it not only lowers prices but it frees up capital to be invested in new projects, products, and services. This is how our quality of life has improved so dramatically over the past few centuries according to many economists and economic historians.

If mainstream economists tend to view economic phenomena in the aggregate, Marxist and other radical commentators try to expose the localized economic pain brought on by restructuring. The decline of industrial cities, the rise in income inequality across all Organisation for Economic Co-operation and Development (OECD) nations, the persistent poverty in various parts of the world (including Western countries), the unequal access to health services, education, safe streets, stable jobs, and a clean environment; all of these are signs that economic restructuring is not a benign or beneficial process. Like most issues in the social sciences, individual views about economic restructuring differ based on personal values and priorities as well as the evidence which is seen as most relevant.

Conclusion: structure and restructuring

Geographers have made important contributions to the understanding of economic restructuring in its various forms. The particular interest of geographers has been to detail the movement of economic activities across space and the impacts of changing trade relationships between different places associated with globalization. The past half-century has been one of immense economic transformation. Most countries have seen a dramatic shift in the goods and services they produce as well as the products available for local consumption. Within most countries, there is significant diversification of economic activities between different city-regions. Indeed, the fastest growing economies of the twenty-first century are at once postindustrial, industrial, and agrarian societies. In China, for example, coastal cities such as Shenzhen and Shanghai represent
the postindustrial economic future while industrial activity moves further inland and Western China focuses on primary activities such as resource extraction and agriculture. Yet despite these dramatic shifts in the global space economy, perhaps the most important question that remains about economic restructuring today is in the temporal, rather than spatial, sphere. How has restructuring changed over time and what might this suggest about the future geographies of economic development?

The social, economic, political, and technological bases for the changing speed of economic relations are little understood. When David Ricardo was writing about comparative advantage in the 1800s, he used examples of national economies producing wine and cloth. A country that had natural advantages in the way of soil and growing seasons could reasonably expect to have a comparative advantage that would last long enough to produce durable trade relations. The notion of economic restructuring, in any of its forms, hinges on there having been an initial structure that can be expected to last long enough to inform rational economic decision-making. If that structure is subject to almost constant change, restructuring ceases to be an exceptional moment when places or entities shift their economic orientations and, instead, becomes the new normal of disorienting economic relations.

There are very good reasons to believe that the early decades of the twenty-first century represent just such an era of constant, complex, and confusing economic restructuring. Despite supposedly radical improvements in information, data analysis techniques, and communications technology, scholars in economics, geography, and urban studies continue to have a hard time making sense of a fast-changing world. Economists failed to prevent, or indeed even to identify, one of the largest financial crises in modern history. Western nations are becoming increasingly unequal, undermining the long held notion of the Kuznets curve in development economics (that inequality decreases as countries industrialize). Predicting economic outcomes is as hard as ever.

The source of this disorienting change in the nature of restructuring is simple: never before in the history of human relations have so many people, in so many different places, had the ability to engage in economic activity that is so fundamentally connected. Market-based economic theories are based on the elegant idea that everyone is ultimately better off if they specialize in producing the good or service they know best and then trading with people to consume a diverse basket of goods. Specialize and trade. If we are entering a period of pronounced economic restructuring driven by new markets, new technologies, and intensifying competition for jobs and investment, geography’s penchant for studying places from close to the ground may be the only way of truly understanding structural economic change.

**SEE ALSO:** Corporate financialization; Creative cities; Deindustrialization; International division of labor; Uneven regional development; Urban geography

**References**


Rights, labor

Jamie Doucette  
*University of Manchester, UK*

Labor rights are sets of normative and applicable legal rights as well as presumed geographically and politically universal human rights. They may be exercised by workers individually and on a collective basis. Individually exercised labor rights often pertain to the terms of employment contract, the length of the working day, protection against unemployment, the right to equal pay for equal work, and the right not to be subject to forced labor. However, depending on the jurisdiction, such rights can extended to include other rights pertaining to the length of the work week, daily breaks, paid vacations, among other items. Labor rights that are collectively exercised often involve the right to freedom of association such as the right to establish unions, the right to collective bargaining, and the right to strike (which is considered to be the most contentious and often least guaranteed labor right in various national contexts).

Most but not all of these rights have been encoded in a number of international declarations such as the United Nations’ Universal Declaration of Human Rights (Articles 23 and 24) and the International Labour Organization’s Declaration of the Fundamental Principles and Rights at Work. More recently, the United Nations Global Compact (2013) has incorporated the content of the ILO’s declaration to recognize four broad principles concerning labor rights:

- Principle 3: Businesses should uphold the freedom of association and the effective recognition of the right to collective bargaining;
- Principle 4: the elimination of all forms of forced and compulsory labour;
- Principle 5: the effective abolition of child labor; and
- Principle 6: the elimination of discrimination in respect of employment and occupation.

These principles are considered to be universally applicable; the UN Global Compact (2013) states that “consensus now exists that all countries, regardless of level of economic development, cultural values, or ratifications of the relevant ILO Conventions, have an obligation to respect, promote, and realize these fundamental principles and rights.” Nonetheless, the compact depends on the voluntary compliance of member countries for the enforcement of these rights, which means that in practice there are many hurdles that obstruct the effective exercise of labor rights.

The spatial configuration of labor rights

While considered universally applicable, in practice labor rights tend to be unevenly observed and enforced by nation-states, transnational corporations, and other actors involved in the configuration of labor relations. Many countries act to restrict a number of basic labor rights in practice, such as the right to strike, particularly for public service workers but periodically for workers in other sectors deemed to be key priorities for national development as well. Furthermore, international organizations such as the ILO refuse to object to national laws.
or regulation withholding the right to strike for the police and armed forces as well as for “essential services,” which can often be loosely interpreted. While nation-states may strongly recognize labor rights in their national laws and labor standards, the reality is that core labor rights are often not enforced and, quite often, all too easily suspended in practice. Scholars thus distinguish between de jure and de facto labor rights: between rights that are recognized in principle and those exercised in practice. This conceptual distinction opens up the study of labor rights to an analysis of how legal and exercised labor rights tend to be distributed in an uneven and spatially variegated manner.

Legal scholars and geographers have pointed out that the representation of labor rights as universal human rights is a contentious topic, as rights discourse can often mask the collective or class-based nature of effective human rights within capitalist economies where collective labor rights are unevenly exercised under the law relative to individual property rights. Geographers have also shown how the courts sometimes regulate the spatial performances of such rights in an uneven manner. For example, Mitchell (2003, 42–80) examines how in the early twentieth century American courts came to regard forms of workers’ speech such as picketing and public assembly as acts of violence that disrupted employers’ private property rights; thus, numerous restrictions were placed on the ability of workers to use public space during industrial disputes in ways that restricted their use of the right to strike. While these restrictions eventually shifted from the formal denial of rights to a language of formal protection that regarded picketing as a legal form of speech, Mitchell shows how the court, through the language of law, has remained an instrument that is used to contain workers’ speech in the name of wider social order. This is often accomplished by prioritizing the rights of employers to have their workplaces and the public spaces that surround them free from disruption by pickets or other forms of collective action. In this sense, by treating both employers’ and workers’ speech as equal, the courts ultimately favor more powerful and property-based voices that benefit from the existing “orderliness” of the public realm over the marginalized voices that need to speak in public space. For workers that don’t have recourse to private property, speech takes on the characteristics of a disruption of a supposedly consensual ordering (the right of employer to freely appropriate an individual’s labor power). This logic of the securitization of public space from the collective action of workers can be witnessed in many contexts where strikes and the right of the individual to withdraw labor are easily made illegal and where workers can be held individually responsible for “economic damages” caused by strike activity and other forms of collective action. Furthermore, many governments raise elaborate legal hurdles to the collective exercise of labor rights such as complex rules for union accreditation, strike votes, and mandatory arbitration and/or “cooling-off” periods. Nonetheless, for Mitchell and many others it makes little sense to abdicate the language of rights; rather they see the potential of continuing to exercise the discourse of rights in order to bring relations of domination and subordination into greater visibility as part of wider campaigns for social justice.

Labor rights in a global context

The ability to claim labor rights has important geographical dimensions as labor rights are often decided through social struggles that are deeply spatialized, and involve contextual dimensions of particular places, territories, networks, and other
spatial registers. Increasingly, the study of labor rights involves an examination of the unevenness of institutional labor rights and the ways in which the very configuration of geographical space can affect the rights of workers either as a tool for the control and containment of the agency of workers or as a site for exercising and demanding rights. As Kelly (2002, 405) argues, the precise combination or configuration of such practices itself varies across space, reflecting diverse historical–geographical contexts. Furthermore, the extension of labor rights has often hinged on the importance of place-based social struggles and legal victories. Both the spatial variability of rights and contextual histories surrounding the ability to claim rights has important implications for the strategic assessment of the appropriate scale at which the rights and entitlements of collective labor should be sought and negotiated; however, a strategic focus on scale should not neglect the ways in which the exercise of labor rights involves connections across multiple scales, from the scale of the body to the scale of the global economy. Feminist geographers have been particularly sensitive to the importance of such interscalar relations in the configuration of labor and human rights.

Increasingly, scholars have been interested in the changing global context of labor rights, particularly alongside processes of globalization, and have focused their attention on labor rights and workers’ agency within global production networks and global value chains (Hess and Coe 2013). Campaigns for labor rights can be quite diverse in these spatially extended networks, with some struggles targeting the brand recognition of leading firms, using both adversarial and cooperative campaigns. Such campaigns have led to the increasing popularity of corporate social responsibility (CSR) codes of conduct as well as international framework agreements (IFAs) (Wills 2002). Nonetheless, scholars have pointed out that, while the compliance-based model of the UN Global Compact and many CSR codes of conduct may improve labor standards, they do not always improve workers’ rights surrounding freedom of association and collective bargaining. Furthermore, critics have pointed out that many CSR codes and compliance-based regulations seeking fair labor standards tend to naturalize relations between wage labor and private-property, leaving intact exploitative class-based mechanisms within capitalism. In contrast, IFAs tend to allow more room for unions to establish the terms of good conduct in a way that endorses workers’ rights and supports union organizing efforts across the company and, by extension, geographical space, often with the help of global union federations. Furthermore, IFAs also create the conditions for workers to upscale their organizing efforts from the local to the global, by providing an extra lever that can be used during local labor struggles.

Similar to global union federations and trade union representatives in international organizations like the ILO and OECD (Organisation for Economic Co-operation and Development), IFAs can be interpreted as an important manifestation of labor internationalism, one that is complemented by a range of contemporary consumer-oriented campaigns involving labor rights and fair labor standards such as Jubilee 2000, Trade Justice Movement, Students against Sweatshops, Clean Clothes Campaign, Worker Rights Consortium, and Women Working Worldwide, to name but a few prominent examples. These prominent campaigns notwithstanding, there is a complex politics surrounding the ability of different forms of labor internationalism to secure workers’ rights. During the Cold War, nationalist and protectionist forms of labor internationalism (sometimes labeled “labor imperialism”) often prevailed and a number of northern trade unions were criticized for aiding
in the repression of the rights of workers in developing countries through their participation in Cold War foreign policies. Based on a critique of such conservative forms of internationalism, in recent years initiatives such as the Southern Initiative on Global Trade Union Rights (SIGTUR) have sought to strengthen workers’ rights at an international scale through promoting the development of democratic and independent trade unions.

SIGTUR regards an alliance of Southern unions as its critical constituency and embraces social movement unionism as a strategy to “bridge space” between social movements and trade unions. Furthermore, the initiative’s Southern identity denotes a political experience of exploitation and subordination in the global economy rather than an explicitly geographically bound network of unions (Lambert and Webster 2001); though, in practice, SIGTUR key participants have been located predominantly in Asia, Africa, and Oceania. In many ways, strategies such as SIGTUR have arisen in part from the uneven territorial configuration of labor rights, particularly in contexts where labor movements have become involved in wider social movements for democracy. The success and failure of movements such as SIGTUR and other attempts at social movement unionism as strategies of exercising labor rights are, in many ways, related to the ability of trade unions to involve popular social movements and ongoing democracy struggles in their campaigns. The articulation of labor rights in relation to democratic struggles for freedom and equality is particularly important in the context of Southern unionism owing to the ways in which these movements coalesced in many developing countries during the past decades. At the same time, labor activists have stressed that social movement unionism is a relevant strategy for workers and social movements in subaltern positions in the Global North as well.

**Territory, mobility, and labor rights**

Beyond a focus on the global and/or international scale of struggles over labor rights, scholars have also enriched the analysis of labor rights by focusing on other spatial registers including the territorial configuration of labor rights and, in particular, the ways in which contemporary processes such as labor migration and spatially selective zoning policies have often involved a disaggregation of labor rights from the national-territorial context in which they have been won. For instance, Ong (2006) examines how Asian states have used various “zoning technologies” such as export-processing zones, special administrative regions, and special economic zones to effectively reterritorialize spaces within and beyond the national territory, creating exceptions to normalized administrative practices in order to attract or produce specific economic activities. Many of these zones offer exceptions on national labor standards and many have included prohibitions on collective bargaining and strike action. Furthermore, Ngai (2005) has examined how special economic zones combine with other forms of territorial zoning, such as China’s household registry or hukou system, uneven development, and constructions of gender subjectivity/identity to produce relations of power and difference that deprive female, rural migrant workforces in export-oriented production zones of the effective capacity to exercise their labor rights.

Shifting focus from territorial strategies to the influence of mobility on labor rights, other scholars have examined how both documented and undocumented labor migration can lead to the scattering of citizenship rights across different jurisdictions in ways that weaken worker abilities to claim their rights and entitlements across place. As Pratt (2012) shows, guestworker programs such as Canada’s Live-in Caregiver
Program violate workers’ rights by paying below market rates, limiting the labor market mobility of foreign workers and subjecting them to perpetual vulnerability to deportation. Similarly, the rise of business migration programs and other entrepreneurial migration strategies has contributed to the fragmentation of rights across jurisdictions inasmuch as such programs involve an uneven extension of rights. In this case mobility becomes a political process in which entrepreneurial migrants are often able to take advantage of significant, differential citizenship rights that are not often extended to labor migrants owing to the more temporary and often more restricted nature of their work permit. In this case, the migrants’ deportability, experienced as either a lack or fragmentation of effective citizenship and labor rights, becomes a key aspect in the appropriation of their labor power within a global market in migrant labor arbitrage. This market has, in part, been successfully produced through the incorporation of a systematic fragmentation of rights into the political order of territorially delimited nation-states. Migrant workers are thus denied labor rights and made to be “indispensably disposable” through their status as deportable labor (De Genova 2010, 47). Furthermore, global regulations seeking to embed migrant mobility into a more substantive set of effective rights for migrant workers such as the United Nation’s International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families remain limited to date, inasmuch as such protections remain unratified by most migrant-receiving countries, particularly in North America and Western Europe.

Finally, while labor rights are often constrained through multiple geographical dimensions, the attempt to exercise effective labor rights in practice has the potential to produce new geographical relations between places, scales, networks, and bodies. While a geographical analysis of labor rights can bring attention to the forces that constrain workers’ agency, it should also seek to show how the actual practice of labor rights, and whatever contestations or “disruptions” such exercise may bring, have the potential to create important changes to existing geographies of rights and regulations not only in particular places, but also within the wider networks, territories, and practices of mobility that exist within the global economy.

SEE ALSO: Citizenship; Exploitation; Feminist geography; Human rights; Labor geography; Labor migration; Migrant division of labor; Migrant labor; Territory and territoriality; Trade unions; Unionism, community

References


Riparian ecosystems

Jacob Bendix
Syracuse University, USA

The term “riparian” is generally used to denote terrestrial organisms or locations that are close enough to rivers to be affected by them, so riparian ecosystems are essentially linear phenomena, distinguished by their proximity to the rivers that they parallel. This proximity allows for multiple interactions between the river and the nearby vegetation, which distinguishes riparian ecosystems from the surrounding landscape.

The means by which rivers affect the adjacent ecosystems fall into three categories: through the direct biological consequences of floods, including plant damage or mortality and propagule transport; through the indirect consequences of floods, including substrate modification and creation of colonization sites; and through the high water tables found close to rivers. All of these factors combine to influence the composition and population structure of riparian plant communities. Very often there are feedbacks involved, as riparian vegetation, where dense, impedes flows, reducing the river’s velocity and promoting sedimentation.

The effects of floods may be seen not just in the destruction of plants in the flood’s path but also in the absence of species that are more vulnerable to flooding. For example, one of the most obvious means by which floods affect vegetation is for plants to be broken or uprooted and washed away by high-energy events. In some cases, this will be evident from the absence of any plant cover following a flood, but in other instances the influence may be more subtle, as plant cover remains but is limited to species with adaptations that allow them to survive the flooding. Such adaptations may include deep root systems, coupled with either large impervious stems or flexible ones that allow the plant to bend rather than breaking in the face of the floods energy.

Flood energy may also remove riparian plants by washing away the substrate in which they are rooted. Thus, the distribution of plants may be influenced not only by their own characteristics but by the characteristics of their substrate, in terms of its resistance to erosion. Plants rooted in cobbles and boulders may survive a flood that would have scoured away gravel or sand.

In the United States, much of the research on riparian ecosystems has focused on rivers in which floods generally occur every 1–2 years and last for only hours or days. There are, however, many low-gradient rivers in which floods occur annually and may last for several weeks or months. These can be found on the coastal plain of the United States, but such rivers are perhaps best exemplified in the lowland portions of the Amazon watershed. Flood energy is often less of an issue in these low-gradient rivers; however, the extended hydroperiod of these floods creates a different set of stresses, as prolonged inundation imposes anoxic conditions in the soil. Again, there is a winnowing effect on species composition. The more frequently a given site is subjected to long-term flooding, the more its composition will be dominated by species that are adapted to survive in saturated soils.

Floods may also have a direct impact on plants by transporting their seeds, either downstream or across the floodplain. For some riparian species,
RIPARIAN ECOSYSTEMS

the timing of seed production is apparently linked to the hydrological regime. Dispersal may also occur when plant fragments are transported to new sites allowing for vegetative reproduction.

The effects of floods are also felt indirectly, because the sediment they transport and deposit serves as the primary substrate in which riparian vegetation grows. Because the texture of sediments deposited at any given point reflect the energy the river had at the moment that material was deposited, substrate texture is directly tied to the details of flood history. As outlined, substrate texture may affect vulnerability to flood scour. Perhaps more importantly, texture exerts a strong control on water availability. This is particularly important for germination and seedling survival, so substrate texture may control the sites at which some species become established.

For plants in a given location, water availability reflects not just soil texture, but also the height of the water table. It is almost axiomatic that the water table is high near river channels, and depending on the environment this may be either a boon or a constraint to the riparian vegetation. In dry regions, the high water table is a resource that typically allows denser and more productive growth, and a different species mix, near the river than in the surrounding landscape. Conversely, in humid regions, the shallow water table may be a limiting factor, as it excludes deeper-rooted species.

These factors work in combination to determine the distribution and composition of riparian vegetation. In some river settings, the hydrologic and sedimentary characteristics of particular landforms may be consistent enough that they support distinctive vegetation. Thus, there may be particular species that typically occur on point bars, on the active channel bank, on floodplains, and so on. This is not universally the case, however; in many rivers the correspondence between landforms and hydrogeomorphic variables is less clear.

The energy of floods (and thus also the amount and texture of sediment they deposit) is strongly influenced by the hydraulic roughness, or frictional resistance, that they encounter. Riparian vegetation forms a key component of this roughness, as dense riparian thickets can very substantially reduce flood velocities. This means that complex feedbacks exist between the influence of floods on the vegetation, and the influence of the vegetation on floods. Part of the complexity derives from temporal change in the vegetation. The roughness provided by a given patch of vegetation will depend not just on its species composition (and associated growth forms), but also on its age, which may, in turn, depend on the interval since the last major flood.

The influences and interactions outlined here are all distinct from the factors affecting vegetation outside the riparian zone. As a consequence, riparian vegetation is usually compositionally distinct from the surrounding landscape, and often structurally distinct as well (as where riparian gallery forests contrast with surrounding shrubs in arid regions).

The riparian zone is also an important habitat for a wide range of animal species, which are drawn by the structural and compositional details of the vegetation as well as the proximity to water. Some, especially amphibians, are entirely limited to riparian habitat, while others depend on it at least some of the time for food or nesting sites.

SEE ALSO: Biogeomorphology; Disturbance in biogeography; Fluvial depositional processes and landforms; Rivers and streams
Further reading


River basin management and development

François Molle
Institut de Recherche pour le Développement (IRD), France
International Water Management Institute (IWMI), Egypt

River basins in historical context

In common language, the concept of the river basin pertains to the field of physical geography, and is well established in secondary classrooms: river (or drainage) basins are extents of land that drains all streams and rainfall toward the same terminus, generally a river or the sea, or sometimes an inland water body. River basins are also often called catchments in British English, while watershed, which in American English designates smaller basins of a few thousand square kilometers, refers to the line dividing two river basins. River basins which drain to an inland water body are called endorheic basins, and form large areas of Central Asia and desert regions, like the Sahara or the Arabian Peninsula. River basins can also be seen as nested “rainfall collectors,” with small tributaries converging to larger rivers.

Although there are indications of sophisticated knowledge of both river systems and the hydrological cycle by the Chinese as early as the third century BC (with a clear description of how vapor generates clouds and clouds rivers), and despite the refined hydraulic skills of ancient civilizations and later of the Romans and the Arabs, the conceptualization of hydrology remained limited. Land and water resources were in general exploited through piecemeal projects destined for channeling, lifting, storing, poldering, or diverting water in places deemed suitable, based on the characteristics of the land, the understanding of the flow regime, and the available technology.

The clear articulation of the notion of the river basin was probably hindered by difficulties in comprehending the hydrological cycle, most particularly the origin of springs and why and how rivers were flowing despite the absence of rainfall for long periods of time. In 1674 Pierre Perrault, a French geographer, wrote the treatise De l’origine des fontaines (On the Origin of Springs), which established a crude water balance of the upper Seine river basin, where he compared the river discharge with “the rainwater that falls around its bed,” a calculation which would later be extended by Mariotte to the flow of the Seine through Paris. In 1752, Philippe Buache, a French cartographer attached to the court of King Louis XV, published an essay attempting to describe the structure of continents based on the study of mountain ranges, streams, and rivers which defined a river basin as “the set of all the slopes on which fall the waters that converge to a same river or creek.” His theory was rapidly taken up by Gatterer in Germany who improved it and made it the basis of a theory of the division of the world into lands and regions.

Industrialization paralleled by scientific and technological development would subsequently project the river basin as the locus of the human conquest of nature. Ambitious national water projects including irrigation schemes or
RIVER BASIN MANAGEMENT AND DEVELOPMENT

ydropower generation were often discursively rooted in the promotion of the river basin as a natural unit for planning development or organizing societies. The river basin concept, beyond its alleged naturalness, thus came to embody a number of ideologies and was instrumentalized by different constituencies. In the late nineteenth century the concept of a river basin nurtured utopias and political struggles concerning the relationships between central and local power in several countries. In France, the concept was captured by political interests to serve as a weapon against the revolution and centralization, and was supported by the landowning aristocracy who sought to re-establish the pre-eminence of the “local.” In Spain, the *regeneracionismo* movement embraced the river basin as a “natural unit” that signaled a natural and harmonious order that was in contrast to the traditional political and administrative divisions inherited from the past; it was used against the traditional landowning elite. In the United States, John Wesley Powell advocated the establishment of self-determined “commonwealths,” independent of both capitalists and bureaucrats, organized along hydrographic basins and based on natural resources rather than on the prevailing township and county system.

Although its relation with the basin scale was not always prominent because early developments occurred in the context of a relative abundance of water, irrigation development became the center of water resource development during the second half of the nineteenth century, with enthusiastic private investments in places such as the western United States, Australia, and India often meeting with bankruptcy and calling for public intervention. As a result of this financial failure and given various national objectives, the states stepped in and endorsed the role of (large-scale) developers of water resources. Imbued with the fresh legitimacy of technical marvels and the presumably unlimited power of science, inspired by the colonial deeds of the British in India and the Sudan, the Dutch in Indonesia, and the French in Vietnam, hydraulic bureaucracies were created to take up the challenges of flood protection, large-scale public irrigation, and hydropower generation. These bureaucracies took as their motto the promise that not a single drop of water should reach the sea without being put to work for the benefit of humanity: the “hydraulic mission” was born (Molle 2007).

The beginning of the twentieth century was thus marked by the creation of many hydraulic bureaucracies in different parts of the world. These include the US Bureau of Reclamation (1902), the National Irrigation Commission in Mexico (1926), the Department of Canals in Siam (1902; now the Royal Irrigation Department), the General Directorate of Public Works in Turkey (1914; now DSI), the Inspeetoria de Obras Contra as Secas in Brazil (1909; later DNOCS), and the strengthening of many of the corps of engineers that had been created in the eighteenth and nineteenth century in European countries, as well as colonial irrigation administrations in Indonesia, Egypt, and India. But the mission of reducing flood damage or irrigating fields was soon to be enlarged with hydropower and the task of generating electricity, fueled by progress in technology in high dam construction, turbines, and high-voltage transmission lines at the beginning of the twentieth century.

These technological innovations and the many “missions” entrusted to hydraulic bureaucracies were pooled and came to be associated with river basin planning and management, as epitomized in the 1930s by the advent of the Tennessee Valley Authority (TVA), launched by Roosevelt in the aftermath of the economic crisis. River basin development was taken to new heights by combining the concept of *unified development*
(the damming of all the streams of a given river basin to bring the river under total control), the benefits of multipurpose dams (hydropower, flood protection, transportation, irrigation, and other uses), and the idea of regional development (associating water development with other interventions such as reforestation, production of fertilizers, industrial development, etc.). These ideas were soon expanded to the Columbia Basin in the United States which was to become the “battery” of the west coast, with the first concrete for the grandiose Grand Coulee Dam poured in December 1935. Similar large-scale projects and planned development were floated in Stalin’s Russia, in a political and parallel contest, where technology, mechanization, and large-scale centralized planning and production processes were part and parcel of a vision of what Josephson (1995) called a “supremely rational society,” which would plan massive hydropower plants and canals (e.g., the infamous White Sea–Baltic Canal), as well as “domesticate” rivers like the Volga.

In the postwar period of the 1950s and 1960s, which was marked by the need for reconstruction and to grow food for a world in shambles, grandiose water resources development schemes were soon in high demand. In the United States, the Bureau of Reclamation and the United States Army Corps of Engineers engaged in a sweeping damming of the country’s rivers (Reisner 1993/1986) and envisaged megaprojects like the transfer of water from Alaska to Mexico. In the Soviet Union, electricity production, and how it would transform society and the economy (a vision long nurtured by Lenin), received much attention from Stalin who launched the Volta Project – epitomized by the Kuibyshef dam – and the 1948 Plan for the Transformation of Nature. This plan and the later Siberian river reversal project to divert water to Central Asia were echoed by what Shapiro termed “Mao’s war against Nature” and its major flood control, canal and hydropower projects in the 1950s. In Spain, Franco undertook the relentless construction of 800 dams (and irrigation schemes) as a way of legitimizing his power and buying support from rural elites. Hydropower development soared in countries like Canada, Norway, and Sweden, where rivers were, in the words of Jakobsson, “industrialized.”

In the developing countries, particularly newly independent ones, elites and governments enthusiastically embraced the icons of modernity and development epitomized by large-scale irrigation schemes and dams – in India famously referred to by Nehru as “the temples of modern India.” In that, they were supported by either Western countries or the Soviet Union, which had both economic and geopolitical interests in fueling this postcolonial hydraulic mission. Massive investments – most pre-eminently in dams and irrigation systems – in countries with potential rural instability were thought by the United States to be the best defense against the spread of communism. It was in this context that the TVA, marketed in particular by the prophetic tone of Lilienthal’s book TVA: Democracy on the March (1944), was to become both an icon of modernity and development and a major asset of US overseas development and diplomacy (Ekbladh 2002): in a matter of years the TVA would become the “grand-daddy of all regional development projects,” embodying the social engineering drive that Scott (1998) has termed “high modernism,” and inspiring a multitude of TVA-like projects in river basins such as the Jordan, Danube, and Mekong, and in countries as diverse as Sri Lanka, Afghanistan, Colombia, China, and South Africa. This further spurred the creation of national water bureaucracies entrusted with river basin planning and the construction of hydropower dams, reservoirs, and canal networks for irrigation.
The four decades following World War II witnessed massive investments in reservoirs (large dams increased globally from 5000 in 1950 to around 50 000 at present, while irrigated areas doubled from 140 million ha to 280 million ha). All the ingredients of nineteenth-century scientism, hubris, and utopias were alive and well and the hydraulic mission was in full swing, predicated on an ideology of the domination of nature, where “conquering,” “harnessing,” and “taming” the wilderness were touted as part of a civilizing mission, and rooted in the conviction that water flowing to the ocean was a waste that called for infrastructure to capture and manage the resource in each river basin.

Beyond promises to feed the masses, raise rural income, or meet energy requirements, the development of public irrigation and associated dams was central to Cold War geopolitics as well as to wider national state policies, whether it was to settle nomads (as in the Middle East and Tunisia); to provide jobs to returning servicemen after the two world wars (as in Australia and South Africa); to break up haciendas and colonize them, with farmers practicing “revolutionary irrigation” (Mexico); or to strategically occupy land (as in the United States, Israel, and Sri Lanka). As a result, the hydraulic mission era was characterized by a massive injection of public money in all countries, with associated subsidies and political favors.

In industrialized countries the hydraulic mission started to lose momentum in the early 1980s, with the growing recognition of associated social and environmental costs, and also with the decreasing availability of suitable dam and irrigation sites. A similar trend was observed 15 years later in developing countries, although infrastructure development remains largely unabated in a number of countries. This change was due to the rise of environmental concerns (priority shifting toward water quality and environmental sustainability), the public costs of such water resources development, and increasing criticism from affected groups and the civil society at large. Challenges to conventional river basin development also resulted from the shift from government to governance, whereby water issues came to be considered as societal questions requiring participation from, or co-management with, concerned stakeholders. In the early 1990s these concerns were reflected in the Dublin Principles and the formulation of integrated water resources management (IWRM) approaches, which were later formalized by the European Union in its Water Framework Directive. The directive sanctioned the river basin as the appropriate unit for managing water, partly under the influence of some national models (e.g., France and Spain), and partly as an embodiment of the promotion of basin-centered management by mainstream international organizations. But the appeal of river basins as an organizing principle also came from its “naturalness,” as the locus of hydrological processes where the integration of water-related issues, as well as the participation of stakeholders, should be facilitated. The promotion of the river basin as a universal unit for water management has triggered wide discussion and scholarly debates from various disciplines.

Current research agenda

Basin management, modeling, and hydrology

River basin management is a subject of predilection for modelers. Hydrologists and engineers have developed numerous models to reproduce hydrological processes and to study and optimize the allocation of resources. Models, long limited to the study of surface water, have grown in sophistication and now increasingly include groundwater modules coupled with surface
water. However, it often remains difficult to appreciate, for example, time lags and two-way interactions between surface water and ground-water resources, how changes in land use, in particular deforestation, alter runoff and ground-water recharge. Likewise, actual management rules of dams and irrigation schemes, and more generally the change in actors’ behaviors at all levels in the face of extreme events, are difficult to model. A growing body of literature is also addressing the evolution of supply and demand in river basins under varied scenarios of climate change. Economists have developed their own models to optimize the economic efficiency of resource allocation within river basins. Despite growing sophistication, these models generally work at a high level of aggregation and are often unable to reproduce local dynamics and to capture the complexity of the interaction between physical and human systems.

The question of efficiency in water use within river basins has also been the subject of many works that have emphasized the concept of river basin efficiency, as opposed to local user or system efficiency. They have shown how local “inefficiencies” associated with leaky canals, reservoir spills, return flows from irrigation, or other system “losses” are often the primary source of water for other users or for ecosystems within the basin. Successive reuses of water across nested scales greatly complexify water balances and accounting, and introduce additional and intertwined questions about changes in water quality and energy costs. Although this important work has critical implications for the concepts of efficiency and allocation, it is not always well understood, and reasoning based on conventional point-of-use efficiencies often prevails, partly because of adherence to classic engineering conceptions but also because it provides easy justification for investments in water-saving technologies.

River basin overbuilding and trajectories

Long-term interactions between societies and their river basin environment are sometimes described by the term “river basin trajectories,” which examines human efforts to assess, capture, convey, store, share, and use available water resources in the face of changing physical and social environments, as well as how the distribution of decisional and discursive power results in specific water regimes, with particular patterns of distribution of costs, benefit, and risk across space, time, and social groups. Other concepts and approaches within the field of political ecology, such as the hydrosocial cycle or the socioecological concept of the waterscape, although seldom applied to river basins as such, also emphasize relations of power behind the manipulation of the water cycle.

One commonplace aspect of a basin trajectory is basin closure, which occurs when the quantity of water abstracted from the river or the aquifer is so high that it can no longer ensure the supply to downstream users, dilute pollution, control salinity intrusion, flush sediments, or sustain healthy ecosystems at the river mouth (or terminus). This imbalance can manifest itself only during a few dry months (where the basin is said to be closing), or almost permanently (where the basin is said to be closed). Closure and resulting scarcity can sometimes occur in sub-basins or small catchments, while the larger basin remains open. Rivers hardly reaching the sea, or contracting lakes, are the most visible signs of basin closure, as exemplified by the Colorado or Yellow rivers, the Aral Sea, and the Dead Sea. In some cases, like in the Lerma–Chapala Basin (Mexico), overabstraction of groundwater, and excessive surface water withdrawals can lead to water depletion exceeding annual renewable water.

The natural interconnectedness of ecosystems and users across a river basin increases with
basin closure. As the amount of available water is insufficient to dampen or absorb fluctuations in supply and demand, or to dilute salt and pollutants, conflicts and negative externalities increase, posing increasing challenges to regulation and management. What particular actors do at a particular point in space and time (e.g., digging farm ponds, tapping groundwater, harvesting water, lining canals, changing cropping patterns, or irrigation techniques) modifies the circulation of water, salts, sediments, and biota, disrupting the environmental processes and human activities associated with the prevailing water regime. The lack of data on, or knowledge of, both hydrological processes and actors’ behaviors often makes it difficult to comprehend, evaluate, or anticipate how the water cycle is altered and what positive and negative social and economic externalities are produced. Externalities travel across space and time and sociopolitical categories of stakeholders. They amount to a constant redistribution of costs and benefits along lines of power that eventually tend to determine who are the winners and the losers between diverse stakeholders. Third-party impacts must be regulated, with the state usually playing a critical role.

Basin closure is generated by the overbuilding of river basins, a socially constructed process of overextension of the water abstraction capacity, in general for irrigation. The process is fundamentally driven by the vested interests of politicians, water bureaucrats, private construction firms, and development banks and the powerful incentives they face in sustaining water resources development. Overbuilding is also caused by regional politics and issues of equity, whereby differences in relative wealth between regions are used by poorer ones to claim for hydraulic (and other) investments even if hydrologic and economic conditions should discourage them. In other settings, between federal states (e.g., India) or between nations (e.g., Nile), it is commonplace to see a rush toward infrastructure development in order to lay claim to or to support a prior claim on the shared resource. Supply augmentation options are more attractive to decision-makers because they avoid politically costly reductions in use or reallocation, but are often adopted at the expense of the public purse and environment preservation, with opportunistic, optimistic hydrologic or economic hypotheses that have to be paid for later: overallocated resources, managers having to tap reservoirs’ security stocks, helplessness in the face of aquifer overexploitation, and the necessity of reallocating water between sectors by fiat or stealth, in general toward municipal and industrial users and to the detriment of agriculture, the environment, or weaker constituencies.

The overdevelopment of water use infrastructure, principally irrigation schemes, generates water scarcity “mechanically.” When most available resources are committed, little slack remains in the stock and fluxes of the river basin, and any disruption caused by hydrologic variability (compounded by climate change) or mismanagement (notably the tendency to overallocate the resource) generates crises that are opportunistically seized by politicians or interest groups to further develop supply. Politicians then often “naturalize” water scarcity and “securitize” the debate by framing and justifying their responses and policies under the cloak of national or food security or other overriding metaphysicifications. In both discursive and material ways, scarcity is manufactured.

Critique of the river basin scale and boundaries

River basins are promoted as the integrative locus of human uses and the environment, the scale at which the resource can be efficiently
managed and allocated, and where participation of actors with a stake in the basin’s water should naturally occur. It is therefore not surprising that river basins have been associated with IWRM and promoted as one of its cardinal “best practices.” They have also been justified by the alleged necessity to improve “spatial fit,” that is, the matching of resources boundary and institutional regimes governing them. A growing scholarship has challenged this prescription. From a management point of view, it has been pointed out that river basin boundaries may not be relevant, for example in the case of small islands, deltas, arid areas, floodplains, and coastal areas; that surface and groundwater interactions need to be considered; that aquifer systems are often noncoterminous with river basins; and that interbasin transfers are also frequent and demand consideration of an expanded scale.

From the government’s point of view – with a focus on the structure of power and processes of decision-making – river basin boundaries are problematic in different ways (Norman, Cook, and Cohen 2015). The accountability and legitimacy of organizations or policies based on river boundaries can be weak, and may generate conflicts with the existing layers of sectoral or political administration and agencies.

More crucially, river basins are affected by social or ecological processes which unfold at different scales and spatial units (e.g., climate change, invasive spaces, etc., which therefore have different “problemsheds”), and basin regulations or management practices may intersect in sometimes conflicting ways with other “policysheds,” or geographical units in which policies (e.g., on land-use planning, reforestation, urban or industrial development, spatial conservation, or protection areas) are implemented. Thus, there has been a recognition that many drivers and consequences of river basin dynamics can be located outside the basin, where solutions to local problems may also lie. This recognition in particular speaks to the relationships between food production, water, and energy, and has triggered calls for integrating policy thinking at a higher level, through what is commonly referred to as a “nexus” approach.

Hydrologic boundaries are a starting point but often beg to be pragmatically adjusted to particular geographic, administrative, and political realities. Mechanisms to harmonize policies, resolve conflicts of prerogatives, and ensure participation in the coordination of multiple levels of organization and administrative layers need to be established. Coordinating existing state and nonstate actors may therefore be the primary goal, rather than creating a new basin organization. Eventually, the selection of boundaries for water or environmental management, whether of the river basin or otherwise, is a political choice.

Ecosystemic approach and environmental management

Intensive river basin development has resulted in major ecological changes. Dams have radically altered the natural flow regime of rivers and the circulation of sediments; large-scale irrigation schemes have withdrawn and depleted large amounts of water; cities, industries, and agriculture have injected massive amounts of chemicals and pollutants. Although hydraulic infrastructures have provided huge benefits in terms of flood control, energy and food production, or navigation, the dramatic alteration of existing hydrologic regimes in terms of quantity, quality, and timing have often undermined or destroyed rich ecosystems, together with the elaborate human uses that had developed around them.

Dams, irrigation, and pumping schemes have led to the loss of springs or wetlands, to the gradual disappearance and contamination of
terminal lakes or seas, and to the many benefits associated with floods (source of nutrients, recharge of aquifers, support of wetlands and biodiversity, flood recession agriculture, reproduction of fisheries, etc.), which have been severely curtailed.

Not only has development, in most cases, resulted in a shift of benefits from the local population to other, often urban, populations, but the overall economic assessment has sometimes been negative, the loss of ecosystem services and existing productivity used being higher than the benefits generated by the investments, as the cases of the Hadejia’ Jama’a floodplain in Nigeria and the Kafue Flats in Zambia famously illustrate.

The necessity to view a river basin as a continuum of interconnected ecosystems in order to understand how changes in one part of a basin affect both water availability and environmental health in other parts of the basin has spurred the development of an “ecosystem approach,” defined by the Convention on Biological Diversity as a strategy for the integrated management of land, water, and living resources that promotes conservation and sustainable use in an equitable way. It has contributed to raising people’s awareness about the diversity of services obtained by people from ecosystems, either directly (fresh water, food, fuel, fiber, genetic resources, recreation, aesthetic experiences, spiritual enrichment, etc.), or indirectly (air quality maintenance, climate regulation, erosion control, regulation of human diseases, water purification, etc.), and has substantially enriched the conception of river basins and the understanding of their management.

This concept has also spurred work in the field of economics, with the development of methodologies for valuing ecosystem services to reveal the hidden costs of interventions and contest cost–benefit analyses and feasibility studies that routinely justify projects by ignoring their negative environmental externalities. They also argue that higher water prices could encourage conservation (thus increasing river flows) and have developed the concept of payments for environmental services.

Environmentalists have also promoted the notion of environmental flow, defined as the flow regime required to ensure the maintenance of particular environmental functions in a river ecosystem. Although the scientific determination of these environmental flows is problematic, and although these flows are often the result of negotiations between different interest groups, claiming a share of the available water for the environment has contributed to the political recognition of environmental requirements, and influenced major policy shifts and regulations such as the European Water Directive Framework. It even gave way to more radical claims, as illustrated by the movement for the removal of dams, which symbolically heralds a nascent paradigm shift.

Basin governance or management models: river basin organizations

The international promotion of the river basin as the natural or adequate scale for water resources management has contributed to the creation and spread of river basin organizations (RBOs). “River basin organization” is a generic term for organizations that come under a variety of names, including agencies, committees, commissions, authorities, associations, administrations, directorates, councils, hydraulic confederations, boards, and trusts. If the diversity of those denominations is partly due to the approximate translation into English of local administration names, it also reflects the historical pathways of the different basins and does make clear from the start that those organizations may sometimes be so different that grouping them under a single
category might actually be misleading: basins may be managed without RBOs, and some RBOs have a very narrow mandate that does not amount to basin management.

Discriminating factors first include the size of the basin (both the problems and the solutions, and how different stakeholders may contribute to them, vary greatly between small watersheds and international river basins) and the mandate of the organization: an RBO may be responsible for any combination of tasks that include construction, maintenance and management of infrastructures, development of basin master plans, allocating water or administrating rights, monitoring and collecting hydrologic or water quality data, law enforcement, fee collection, and promotion of public participation and awareness. But, from a governance point of view, an RBO can be typified, first, by its vertical integration within the state administration and, second, by its horizontal integration with nonstate actors.

Within the state, an RBO can be given all-encompassing powers that include planning, construction, and management, as well as regulation, in which case it will often be an autonomous authority, with prerogatives that override those of line agencies. But it can also be entrusted with a more limited mandates, in which case it will often be located under a particular ministry or department. The idea of integration has also promoted the concept of regulation, where an RBO is often supposed to define the rules of the game (like water quality standards, maximum aquifer withdrawals, user fees, and water rights) by which different sectors, users, and governmental agencies must abide. Unsurprisingly, the creation of a layer of governance at the basin-level results in the redistribution of bureaucratic power and often generates tensions or conflicts with other segments of the bureaucracy, as well as with local administrations. Because of the political difficulties of reshuffling prerogatives, RBOs often end up layered on top of existing institutions rather than replacing or complementing them. The development of a regulatory RBO, often located under the new and weak ministry of environment, is often not well accepted, especially if it threatens sectoral vested interests associated with the planning and construction of infrastructure or with unchecked pollution.

The degree of horizontal integration indicates how nonstate actors, such as representatives from the agricultural or industrial sectors, environmental nongovernmental organizations, and various civic groups, are contributing to decision-making. Nonstate actors can be called on to participate in different ways, from just consultation or participation in basin councils or platforms to representation in executive boards and decision-making, to being the driving force of RBOs that are partly independent from the state (which will be more common in small watersheds).

Integration is often taken as a justification for centralizing decision-making power and internalizing decision-making in a powerful organization, with the frequent risk of combining regulation and operation, and limiting accountability. However, it can be used to promote polycentric governance, where both different levels/scales and the views and interests of state/nonstate entities are expected to be harmonized, with the risks of stalemate by fragmentation of decision-making power and high transaction costs. These patterns of vertical and horizontal integrations define various forms of governance that must be further analyzed and characterized, for example by looking at their efficiency in terms of delivering sustainable or equitable management, the way their legitimacy is built and affirmed, their degree of accountability to society, and their capacity to be financially self-sustaining.
Collaborative arrangements for river basin governance are growing, as a result of the failure or limitations shown by models of decision-making restricted to state bureaucracy and experts, of the increasing demand from the private sector, interest groups, and civil society to have a seat at the table, but also of the new emphasis on, and the interest of the public in, restoring environmental quality in line with new values and uses, such as recreation and aesthetics. These factors have been at the root of the surge in the 1990s of the US watershed management movement, which includes over 1000 watershed experiences in collaborative planning, whereby local stakeholders decide the issues and discuss their options, with some technical assistance and funding from both federal and state agencies.

The mandate and prerogatives of an RBO may evolve with time, as a reflection of changes in the problems and challenges, in societal values, and in state–civil society relationships, and of political changes. More generally, these changes refer to the continuous adjustment of governance frameworks to ever-changing context and challenges, a need advocated by scholarship on adaptive management and governance.

The recent work of geographers on the politics of scale and processes of rescaling addresses the social production of scale and its impact on the distribution of power. Here the issue is how actors gain or lose influence as a result of authority being reconfigured around new spatial levels or by virtue of their own ability to work across different scales or levels.

Transboundary basins and hydro-hegemony

Scholarship on transboundary river basins examines how water is managed in the 263 basins that cross national boundaries. They represent 60% of total river flows and 45% of the Earth’s land surface, while being home to about 40% of the world’s population. A first focus is on legal issues, including the 1997 UN Convention on the Law of the Non-navigational Uses of International Watercourses, with its two main principles of “equitable and reasonable use” and the obligation not to cause “significant harm” to neighbors, treaties between two or more countries sharing a river basin, or other institutional arrangements for transboundary river basins on issues such as pollution and navigation (as with the Rhine and the Danube), water allocation (the Indus and the Nile), and joint management (the Mekong and the Senegal).

Despite such arrangements, binding agreements are rare or nonexistent; transboundary management organizations are only given limited power; mechanisms for monitoring and enforcement are lacking; hydrological data remain secretive; and historical political relations between neighboring countries, as well as strong sentiments of territorial sovereignty, make it difficult to ensure equitable and environmentally friendly management.

One way forward has been to respond to the lack of public involvement in interstate water agreements by developing river basin councils, platforms, and forums in which water user representatives discuss plans and allocation issues within a country, like in the Zambezi basin, where a basin strategy has been developed with the active involvement of stakeholder groups in all eight riparian countries. Another way has been to develop the concept of benefit sharing, whereby the stalemate in negotiations over water allocation is overcome by introducing other benefits related either to the use of water (e.g., sharing the benefits of hydropower between countries) or to other issues such as trade.

Yet, the topic remains a favorite of political scientists, who have in particular developed the concept of hydro-hegemony to describe
relations of power between countries sharing a same river basin.

**Future research directions**

This brief review of ongoing scholarship about river basin development and management points to a few questions and areas of research that deserve further inquiry. Modeling approaches and stochastic hydrologic models are crucial to providing information on management options and associated levels of risk, but the study of extreme events must be paralleled by an understanding of policy and political processes that are difficult to model. Although often limited by the availability and quality of data, hydrologic modeling needs to refine the representation of the interaction between surface water and groundwater, and to better take into consideration water quality issues.

Given the baffling diversity of physical river basin environments, the combinations of problems faced, and the multilayered institutional arrangements, more effort should be put into understanding the relationship between the nature of water governance regimes and their effectiveness. More elaborate typologies of river basin organizations and other institutional arrangements should consider the wider historical, social, and political contexts in which these governance regimes emerge and evolve, and provide insight into the administrative configurations which should be favored in a particular context.

These typologies also require a more nuanced understanding of bureaucratic dynamics and reforms, in particular a closer look at the structure of incentives available to different actors and organizations, when the structure of decision-making power is reconfigured to address issues of basin-level environmental management.

More work is needed on the societal and political drivers of river basin closure, as a means of challenging discourses that naturalize water scarcity and water crises or frame them in Malthusian terms. The preference for supply augmentation and capital-intensive solutions results in the generation and compounding of water crises, and is therefore self-sustaining; this preference must be explained by analyzing the social distribution of costs and benefits attached to different policy responses, in particular the convergence of the interests of politicians, bureaucrats, and interest groups.

Urban studies focusing on water and sanitation issues have produced a substantial body of scholarship in the field of critical geography. Insights from urban studies need to better fertilize, and to be integrated into, studies on river basin dynamics and governance, and vice versa. More generally, studies of river basins provide an opportunity for multidisciplinary work integrating approaches from hydrology, economics, human geography, and policy studies, among others.

Multisectoral or nexus approaches also have the potential to refine understanding of the systemic complexity of resource use and economic activities, although it is unclear whether they can substantially influence sectoral policies and practices.

Last, the transfer, operationalization, and adaptation of river basin-based management or governance reforms in different contexts must be accompanied and scrutinized. Critical work on the ideologies, interests, and social mechanisms that sustain the reproduction and dissemination of particular practices and policy models is needed. Europeanization, for example, and the application of the European Water Directive Framework provide an opportunity to assess the confrontation of uniform policy guidelines.
with the diversity of environmental and human contexts.

**SEE ALSO:** Governance and development; Irrigation; Political ecology; Water resources and hydrological management

**References**


**Further reading**

River ice and ice jams

Étienne Boucher
Université du Québec à Montréal, Canada

Ice covers are distinct features of cold-environment rivers. From the formation to the breakup of ice covers, a wide array of key biological, hydrological, and environmental components of hydrosystems can be affected. More particularly, the transition from an ice-covered regime to an open-water regime is a crucial moment in the hydrological year of cold-environment rivers. During this period and when conditions are favorable, severe ice jams can occur that may potentially threaten riparian communities and infrastructures, and transform the fluvial landscape in the long term.

Ice cover formation and breakup dynamics are both climate-sensitive phenomena, meaning that ongoing climate changes are susceptible to having an important effect on the timing, magnitude, and frequency of extreme ice-induced events (Beltaos and Burrell 2003). However, despite the ubiquity of ice-covered rivers in cold regions of the world and the sensitivity of river ice processes to climate change, scientific knowledge relating to ice cover formation and breakup remains limited, due to the complexity of the physical mechanisms involved and the difficulty of studying ice properties, especially at breakup. Here, a review of the most important processes associated with river ice formation and breakup is presented, along with a description of hydrological, biological, and geomorphological impacts.

River ice formation

River ice covers can form at any location where the air temperature drops below 0 °C for at least a month. This area extends north of 30° N in North America and north of 25° N in Eurasia (Prowse 1995). The length of the ice season therefore varies with the latitude, with longer ice seasons in high-latitude and high-altitude environments.

River ice forms because, under frigid weather, water loses heat to the atmosphere. Ice cover formation must therefore be seen as a thermally dependent phenomenon. Water phase changes, however, do not strictly relate to absolute temperature but are strongly linked to the presence of nucleating agents (or freezing nuclei). Freezing nuclei are microscopic particles (e.g., dust, bacteria, ash, spores, etc.) present in all natural freshwaters and that favor the crystallization of water molecules into ice.

Typically, river ice first forms in tranquil flow sections. The freeze-up of turbulent sections happens later in the cold season, when persistent, frigid weather become more common. Ice crystallization initially occurs in low-flow sections adjacent to the riverbanks, and tends to grow laterally, from the sides to the center of the stream. This results in the formation of laterally growing ice strips often referred to as border ice. If the growth of border ice is uninterrupted, strips eventually coalesce, leading to the complete and rapid closing of the river reach.

In the central section of a watercourse, however, turbulence is often more important, impeding (and often stopping) the lateral
expansion of border ice, even if air temperature
drops significantly below freezing point. Tur-
bulence causes the rapid mixing of the water
column, which in turn favors the replacement
of cooled water from the surface by less cool
waters from the bottom. These specific turbulent
flow conditions, along with the presence of a
strong (negative) heat gradient between water
and the atmosphere, result in the supercooling
of the water and in the formation of frazil ice.
Frazil crystals are small (0.1–5 mm) subrounded
ice particles that are transported downstream
by streamflow. Once formed, frazil ice crystals
typically grow in size and present more angular
shapes as the crystallization process advances.

Frazil ice crystals can be seen as “buoyant
sediment.” This is because the analysis of frazil
ice dynamics is somehow quite similar to the
study of sediment origin, transport, and fate in
rivers and generally consists in the identification
of frazil production zones, the description and
modeling of transport mechanisms, and the
characterization of deposition contexts. In that
sense, frazil ice has very important physical
characteristics that need to be underlined.

First, frazil crystals are formed in turbulent
reaches where, as mentioned earlier, surface
water loses significant amounts of heat to the
atmosphere. For that reason, frazil ice produc-
tion typically (but not exclusively) occurs during
cold, windy, and cloudless nights associated with
important heat losses to the atmosphere.

Second, frazil ice crystals are buoyant and can
be transported downstream. In fact, when they
form in the initial stages of supercooling, frazil
crystals present a buoyancy that is comparable
to that of water and thus can be transported
even at small flows. For that reason, frazil crystals
can be transported far (downstream) from the
zones in which they were initially formed. Once
produced and incorporated into the water col-
umn, frazil crystals usually agglomerate within
hours, forming flocs that rapidly grow in size.
Frazil flocs can be described as relatively loose
agglomerations of frazil ice crystals. If the air
temperature keeps decreasing, frazil flocs become
continually exposed at the water’s surface and
gradually freeze, solidify, and consolidate as
sheets. The proportion of ice to water aug-
ments within frazil agglomerations, and as a
consequence, they become even more buoyant
(in fact, more buoyant than freshwater). For
that reason, frazil ice flocs become less and less
affected by turbulent waters and easily float
downstream. Floating ice flocs often result in
the formation of ice pans (often referred to as
pancake ice). Ice pans are circular agglomerations
of frazil flocs that drift downstream and that can
be easily recognized by their upturned edge.

Third, frazil ice crystals are sticky and can
adhere to rough underwater surfaces (Ettema
2002). Rough underwater surfaces are numer-
ous: aquatic plants, irregular gravel beds and
banks, and the ice cover itself. Since supercooled
waters and frazil flocs are affected by turbu-
lent flow, they can be transported downwards
towards the river bed where they agglomerate
(or “anchor”) to form anchor ice (often called
bottom-fast ice). Anchor ice corresponds to a
relatively flat accumulation of frazil crystals in
submerged portions of the river bed. At breakup,
anchor ice might be mobilized by the flood wave,
carrying and transporting lag deposits on its way.
Another surface where frazil flocs tend to accu-
mulate is beneath a static ice cover. Ice cover
bottom presents an irregular topography against
which frazil ice crystals agglomerate. The sus-
tained agglomeration of frazil flocs under a dense,
thermal (black) ice cover results in the construc-
tion of a hanging dam. Hanging dams grow
from the surface in a downward direction, and
frazil ice tends to be highly consolidated near the
black ice surface and looser near the frazil/water
interface. Rapid hanging dam growth may cause
important flow constriction, and often result in abnormal ice thicknesses ranging from a few tens of centimeters to a few meters. Hanging dams tend to form in the vicinity of slope breaks, where frazil ice is produced massively in favorable hydroclimatic conditions.

In very cold environments, river ice can sometimes freeze to (or close to) the bottom of the channel, due to the length of the cold season, or following important frazil ice accumulations. The result is a flow occlusion that can be significant enough to prevent the normal evacuation of water. In such conditions, water can sometimes find its way to the top of the ice cover, and refreeze there. This event is often referred to as an aufeis (icing or nalled), and produces a typical upward-growing ice cover where layers of black ice (corresponding to successive icings) alternate with snow deposited between aufeis events.

River ice breakup

River ice breakup is a crucial moment in the hydrological year. It corresponds to the moment (usually spring or late winter in the northern Hemisphere) where rivers transit from an ice-covered regime to an open-water flow. This has drastic hydrological and environmental consequences. River ice breakup is a complex and multivariate progression of events that is best described by a succession of thermal and mechanical processes through which river ice decays, degrades, deconsolidates, fractures, and ultimately, gets washed away.

Breakup occurs when factors and processes that tend to maintain the ice cover in place (cold temperatures, snow precipitation, etc.) become less important than those that contribute to its thermal or mechanical degradation. Thermal processes control the rate at which the ice cover melts in situ, while mechanical processes are mainly hydrodynamic forces that contribute to fracture and break the ice cover into floating rafts, before its complete thermal degradation.

Thermal breakups

Degradation and decay can be defined as a loss of mechanical strength and integrity of the ice cover (Figure 1a). Speed and timing of ice cover degradation are driven by two major variables: incoming solar radiation and water-to-ice or air-to-ice heat transfers. Of the two variables, radiative processes have the most important influence on river ice cover degradation and decay (Prowse 1995). The influence of short-wave incoming solar radiation augments through time, particularly through changes in surface albedo. In early breakup stages, the ice cover is often entirely covered by snow, so a considerable amount of energy is returned directly to the atmosphere. However, as the ice cover decays, surface albedo decreases with superficial ponding, therefore accelerating ice cover degradation. The speed of this process and the amount of water ponding at the surface sometimes depend not only on the amount of incoming radiation, but also on the quantity of fresh snow covering black ice.

Radiation influences surface melt, but river ice cover also melts from the bottom as water temperature rises. Most of the heat transferred to the water at the time of breakup originates from the friction caused by the movement against the ice cover and/or the channel. This effect amplifies with increasing velocities due to rapid snowmelt.

The speed and timing of river ice cover degradation by hydrothermal processes is likely to vary with river ice thickness. At locations where frazil accumulations have been important during the antecedent winter, hydrothermal processes may need more time to degrade the multiple and often imbricated layers of black ice and consolidated frazil ice. By contrast, in river
sections where black ice dominates, ice decay is much faster. It is therefore common to observe an important spatial irregularity in the timing of thermal breakup in river reaches that are located next to one another, owing to nonuniform ice thicknesses.

Thermal breakup in rivers can be compared to the process of breakup in lakes (see Lake ice). The progressive thinning of the ice is accompanied by an increased water content and a general loss of structural integrity, characterized by a progressive reduction in ice density and strength that results from the augmentation of ice porosity. Degradation by hydrometeorological processes advances to a point where the ice cover weakens enough to lose most of its original mechanical properties. Ice cover remnants can then be easily washed away by the flood.

Mechanical breakups: onset and triggering

In particular hydrometeorological contexts, thermal degradation of the ice cover is not completed
when the spring flood wave propagates downstream. As a consequence, at the time of breakup, the ice cover is mechanically unaltered (i.e., intact or undegraded), but the hydraulic action of the flood is such that the ice cover is rapidly fractured and forced to move downstream (Figure 1b). This critical situation is referred to as a mechanical breakup and announces the formation of ice jams.

Mechanical breakups occur when hydraulic forces provided by the flood wave are greater than the forces that tend to maintain the ice cover in place. It is generally admitted that the strength, resistance, and density of the ice cover are properties that are “inherited” from the succession of cold events that prevailed during the preceding winter (Beltaos 2003). In other words, the ice cover’s resistance is maximal just prior to breakup, but its strength progressively (and inevitably) weakens under the effect of hydrothermal degradation. On the other hand, increasing velocities and water levels during flood may force the ice cover to lift and break into rafts. The amount of hydrodynamic force required to fracture the ice cover during a mechanical breakup episode depends on multiple factors such as ice cover thickness, water temperature, and most importantly, the state of degradation and decay of the ice cover (Beltaos 2003). Thus, the timing of the flood is crucial for the triggering of a mechanical breakup. It can be assumed that very large and early floods are required to fracture the ice cover at the very beginning of the breakup season, because the ice cover presents a full structural integrity. However, as the breakup season advances, fewer and fewer important flood events are required in order to break the ice covers into rafts. With the progression of thermal decay, increasingly smaller flood events are required to wash away the ice cover.

Hydroclimatic conditions and thresholds associated with the triggering of mechanical breakups in rivers are complex and often site-specific, although some generalities can be identified. In most cases, both flooding and ice conditions are important.

Floods volumes do not need to be extreme to trigger a mechanical breakup. On the contrary, mechanical breakups are usually much more sensitive to the timing of the flood. Early and rapidly rising floods are more likely to fracture the mechanically intact ice cover. In cold regions where snow precipitations are important, rain-on-snow events occurring early in the season may provide sufficient energy to the flow to fracture the ice cover before its degradation by thermal processes. This phenomenon is amplified in northward flowing (large) hydrosystems. In such environments, the southernmost portion of the river melts first, while the northernmost portion of the river ice cover remains intact and undegraded. Floodwave propagation (northwards) augments the risk of mechanical breakups in the lower portion of the watershed.

While the timing of the flood is crucial, mechanical properties of the ice cover are also important. The thicker, denser, and more resistant the ice cover is at the time of breakup, the greater will be the probability of a mechanical breakup. Thus, all variables that prevent the ice cover from melting, and that control the mechanical properties of the ice cover, are important. Overall, the length of the frost season usually has a significant effect. Colder frost seasons may lead to thicker and denser ice covers and may also result in an abundant frazil production which accelerates the formation of hanging dams. Ultimately, abundant snow precipitation, especially in weeks preceding breakup, may slow down the thermal degradation of the ice cover.
RIVER ICE AND ICE JAMS

Mechanical breakups: ice jam localization

Ice jams may be defined as in-channel accumulations of ice that prevent the normal evacuation of streamflow. More formally, jams occur where ice discharge exceeds transport capacity. They can occur at freeze-up, but are most common at breakup.

Ice transport capacity is not uniform along a watercourse. Numerous obstacles exist along a watercourse that prevent the free circulation of ice rafts in the downstream direction. Channel constrictions associated with a marked reduction in velocity and width are clear examples of sites that can lead to significant congestion. Channel constrictions may be natural or man-made. Bridge piers are the most common man-made obstacles. Natural constrictions are often associated with the presence of geological obstacles (e.g., rock outcrops) that force the passage of water into a bottleneck. The presence of island as well as sharp meander bends can also be identified as possible obstacles for the free circulation of the ice. Finally, the ice cover itself, owing to the nonuniform patterns of decay along a river, may be seen as an obstacle for the circulation of ice rafts drifting downstream. Depending on the thickness of the “ice wall” (e.g., hanging dams) and as a function of water velocities and ice discharges from upstream, the magnitude of the blockage by the ice cover might vary considerably between years. Ultimately, ice jamming sites tend to be located where a combination of the above-mentioned obstacles is found.

Hydrological impacts of ice jams

One of the most perceptible threats associated with the occurrence of ice jams is flooding. Flooding occurs upstream of the jam when floodwater passage is significantly obstructed by the presence of an ice dam. Ice-induced flooding is considered extreme in many ways when compared to floods produced in open-water conditions.

The first reason why ice-induced floods are considered extreme is their amplitude. Ice jam floods often reach much higher levels than those produced in open-water conditions, at comparable discharge. This effect is mostly attributable to the resistance exerted by the ice cover and the jam itself. The second reason relates to the speed of flood rise. In fact, ice-induced water level rises are much more rapid and sudden than those observed during open-water conditions. Actually, in minutes, waters close to freezing point can rise by a few meters and threaten riparian communities and infrastructures. The menace of a rapidly rising ice flood must be interpreted in light of the difficulty to predict both timing and location, due to the complexity of processes involved and also because of the lack of systematic observations on breakup dynamics in most rivers. Thus, the development of modeling tools that would serve as a basis for the deployment of warning systems is a goal that, even today, seems far from being achieved.

Geoecological impacts of river ice jams

Breakup ice jams are channel-forming events. They have major and long-term impacts on channel geomorphology and ecology and are associated with a distinct signature in fluvial landscapes. This signature exists because ice is an abrasive agent that combines with flood waters and significantly augments the erosional capacity of the flood at breakup. The type and extent of ice-induced erosional features and landforms, as well as their long-term maintenance in the landscape, depend on many factors such as the characteristics of the ice-flood regime (frequency,
magnitude, type of ice event) and the nature of bed and bank material available for erosion.

Influences of ice processes on bed forms and morphology are sometimes difficult to measure because they occur underwater. Ice runs and jams can nevertheless significantly scour the riverbed through a variety of processes. Since, under the jam, speeds tend to increase to accommodate for friction losses, flow may promptly become more competent and transport a larger amount of sediments. This typically occurs under thick hanging dams where flow is diverted downwards towards the bed. Another process affecting bed morphology is the removal of anchor ice at breakup. Anchor ice strongly attaches to riverbed sediments during the winter but when breakup arrives, rafts may detach from the river bottom and transport sediments glued to the floating anchor ice rafts. Ultimately, in meandering streams, river ice jam floods may cause channels to avulse (Smith and Pearce 2002), that is, to shift position abruptly while abandoning the ancient channel. This phenomenon, however, remains poorly documented.

The most obvious effect resulting from ice jams and ice runs is bank gouging and abrading (Figure 1c). In sandy rivers, lateral erosion that occurs during ice jams and runs typically results in the formation of two-level ice-scoured riverbank morphologies that consist in (i) a steep erosional talus located just above the bankfull stage and, (ii) depending on the height of valley walls, an aggradation surface covered with fresh alluvia deposited during ice-induced floods. The erosional talus results from the mechanical abrasion by thick and dense black ice rafts that are pushed, during ice runs, against the recently unfrozen and thus weakened riverbanks. Erosion typically occurs slightly higher than the bankfull level that is constructed by regular open-water floods. Contrarily, vertical aggradation during ice-induced floods occurs on top of ice terraces, at levels rarely attained by typical open-water events.

Whether or not this two-level ice-scoured morphology can possibly be constructed during a single event or, alternatively, is the product of recurrent ice floods that maintain the erosional feature in place for decades, remains difficult to verify, owing to the lack of long-term and large-scale data on river ice processes. Studies from dendrochronological investigations of river ice processes (e.g., Boucher, Bégin, and Arseneault 2009) nevertheless associate the occurrence of such lateral erosion features with river reaches that are commonly affected by ice jams. Other studies have even proposed that cold-environment rivers may, in the long-term, adjust in terms of both size and shape to local ice conditions. For example, in several Albertan rivers where ice jams occur relatively frequently, channels were possibly enlarged (at bankfull stage) by a factor of 2–3, by comparison to channels where ice effects were negligible (Smith 1979).

Where riverbank material is coarser, ice-induced erosion will often take the form of boulder pavements/ridges or bechevniks. Boulder pavements and ridges form where heavy ice push events occur. Coarse sediments are pushed by ice rafts against the (often steep) valley walls, resulting in a subparallel alignment imbrication of cobbles and boulders, forming either an extended flat erosional surface (pavement) or a bow-like structure (ridge). Fine sediments are usually rapidly washed away from these surfaces during open-water floods. Bechevniks are quite similar but also include overbank deposition on the upper portion of the river bank. Bechevniks form by the erosion of the riverbed and the subsequent transport and deposition, by ice rafts, of eroded, unconsolidated coarse sediments on the upper part of the riverbank, at heights that are much higher than the bankfull stage.
RIVER ICE AND ICE JAMS

Such features were named by Hamelin (1979) in reference to the strip of land adjacent to several Russian rivers, where boats were towed manually.

Finally, ice-induced lateral erosion is a dominant process in the riparian landscape because it has the potential to deeply transform riparian ecosystems (Scrimgeour et al. 1994). Impacts on the vegetation are multiple, and the responses, also, depend on the species affected and the resilience of the ecosystem. In most cold-environment rivers that are commonly affected by ice jams and runs, vegetation found on riverbanks is typically composed of erosion-tolerant species. Such species (e.g., shrubs such as Alnus rugosa in northeastern North America) vegetatively reproduce quite efficiently and may survive even the most abrasive ice-scouring events. For that reason, erosion-tolerant vegetation is often found immediately adjacent to the riverbed in cold-environment rivers. At slightly higher levels, vegetation becomes less resistant to flooding and erosion, producing a recognizable vertical stratification of riparian vegetation. At this level, vegetation is often scarred and damaged (trunks, branches) during high-intensity events, but will not be completely destroyed unless extreme ice push conditions are encountered. In rare occasions though, severe ice push may destroy the ecosystem (Figure 1d) and interrupt forest succession, a situation that ultimately results in an important rejuvenation and sometimes replacement of the original vegetation.

SEE ALSO: Lake ice; Rivers and streams

References


Rivers and river basin management

Naho Mirumachi
King’s College London, UK

Rivers are a major source of fresh water. They play an important role for livelihoods and economic development, and the river basin is vital to both terrestrial and freshwater ecosystems. River basins also represent a political space where different interests of water use and allocation are negotiated, reflecting conflict and cooperation between various stakeholders. River basin management is often bound by the hydrological unit but with varying purposes, including engineering the river flows and conserving ecosystems. River basin management is mediated through legal principles, agreements, river basin organizations, and other institutions at multiple spatial scales. The river basin becomes the context in which ideologies of nature and modes of environmental governance are played out.

Developing the river and basin closure

Broadly speaking, rivers have historically been the object of control and exploitation by societies attempting to utilize the water resources and to prevent hazards such as flooding. River basin management has been centralized to roll out large-scale irrigation schemes and to develop dams. Often described as the hydraulic mission, the state organizes engineering projects so that the benefits of abstracting water resources from the river are maximized. Irrigation projects utilizing dams and canals are iconic of the hydraulic mission, embodying a masculine representation of approaches to river basin management (Zwarteveen 2008). Water transfers between and within basins provide the means of supplying water resources where needed, reconfiguring the river flow. For example, in the American West engineering interventions have transformed an arid landscape. The South–North Water Transfer Project in China is another major scheme to address water scarcity in the northern regions.

This mode of river basin management is also an expression of modernity, employing technology to subject nature. The river becomes the means through which economic development is achieved and a utilitarian view of water resources is prioritized. Moreover, the mix of centralized, large-scale engineering with ideals of modernity enables political control over land and peoples, as can be seen in the example of Spain during the fascist regime (Swyngedouw 1999). The river and river basin can be geographies where political ideals are projected and made to symbolize state power.

However, relying on infrastructure to manage the river can pose unique temporal problems. While dams can benefit economic development, these infrastructures do have a life span and become obsolete after decades in operation. Nowadays, countries that have long achieved their hydraulic mission face the issue of decommissioning dams. In the United States, consideration of the removal of dams has come late: dam removal procedures were only formalized in 1994 after prompting from the Congress and civil society (Doyle et al. 2003). The problem of removing dams touches upon not only issues
RIVERS AND RIVER BASIN MANAGEMENT

of hazard risk and safety from deterioration but also impacts on ecosystems as a result of removal, and its socioeconomics costs.

This extensive and intensive river basin management has led to situations of basin closure. Basin closure occurs when the river flow is not sufficient in quantity and quality for downstream uses and ecosystems, and does not function to control saline intrusion in delta areas and to transport sediments (Molle et al. 2007). The Colorado and Yellow rivers, in addition to the Aral Sea are considered to have experienced basin closure; the Jordan, Indus, and the Murray–Darling rivers are nearing closure. The increase of rivers facing closure has meant that 1.4 billion people lived in closed basins at the turn of this millennium (Falkenmark and Molden 2008). In situations of basin closure, not only is water scarce but also the health of ecosystems is at risk. This has implications for livelihoods and poverty, especially rivers in the Global South where many communities are directly reliant on resources from the river basin.

Assessing environmental flows to address basin closure

To ensure water is allocated in a way that does not cause tradeoffs with the environment, different methods to calculate environmental flow requirements have been developed. The concept of environmental flows explicitly considers water for the environment, not just in terms of quantity but also the quality and timing of available water within the river basin (Smakhtin 2008). Reallocation of water from one use to another can be informed by knowledge of environmental flows. Assessing environmental flows as part of river basin management may change the rationale of controlling the river for maximum utility to one based on sustainability principles.

Environmental flow assessments have been utilized to facilitate river basin planning in the Murray–Darling River in Australia. Intensification of irrigation, based on a hydraulic mission of storing and transferring water, has reduced the quantity and also quality of water resources in the river basin (Figure 1). In order to curb extraction and also restore the environmental health of this basin, river basin planning has focused on increasing environmental flows. However, in many river basins around the world, the concept of environmental flows has yet to be fully embedded in planning and management. This process requires not only robust science to assess the requirements, costs, and benefits of environmental flows but also institutional changes to river basin management. In developing countries where water policies may be weak, the institutional setting may not enable reallocation water from existing uses to the maintenance of environmental health based on environmental flows. Even in the case of Murray–Darling River basin, the different governments of New South Wales, Victoria, Australian Capital Territory, Queensland, and South Australia lacked a clear policy for implementing environmental flows despite the recognition of their importance (Pittock and Finlayson 2011).

Integrated water resources management (IWRM)

Pollution, watershed degradation, and scarcity of water resources have prompted serious consideration about the sustainability of water resources and how river basin management can reflect it. In this regard, policy innovation in river basin management in the last twenty years cannot be detached from the debates of sustainable development. Water resources feature prominently in documents such as the Agenda 21, which was
the output of the Earth Summit in 1992, the first major global forum that called for concerted action for sustainable development. In Agenda 21, the scope of water resources management was not limited to the mere prevention of pollution or conservation. It also included the holistic treatment of multiple water uses to sustain lives and livelihoods. Consequently, managing water for health and sanitation, agriculture, energy, ecosystems, and other purposes within a river basin was understood to require an integrated approach that is distinct from management under the hydraulic mission.

Integrated water resources management (IWRM) has been promoted widely since the 1990s. It reflects the discourse of sustainable development and is defined as “a process which promotes the coordinated development and management of water, land and related resources, in order to maximize the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems” (GWP 2000, 22). IWRM is based upon three key principles. The first principle is economic efficiency to guide water use. Underlying this notion is the assumption that water resources are already scarce and will face increasing demand. Equity between water users is a second principle that is to inform water access. The third principle considers the environmental and ecological sustainability so that present water use is not at the expense of future uses. The river basin, defined by its hydrological basin, is seen as the optimal unit to practice integration and coordination under IWRM. IWRM at the river basin level allows the consideration of river use with other water resources, such as groundwater and coastal water resources.

In Latin America, Brazil has been the forerunner in national-level efforts to establish
IWRM strategies. An example of efforts beyond national governments to adopt IWRM is the European Water Framework Directive. As a regional policy, the framework directive requires each of the member states to make provisions for water resources planning at the river basin level. IWRM can also be applied to international transboundary river basins, as can be seen in the example of the Niger River, shared between ten countries. The riparian governments recognized IWRM in a statement on the “Clear and Shared Vision” of the river basin, which aims to achieve sustainable development by 2025.

IWRM attempts to better coordinate water use by society. Equity is considered from the perspective of water use by humans only, excluding ecosystems. Sustainability is also understood as a means to ensuring continued water use for societies. This strong emphasis on human water use is underscored by a set of global principles established for sustainable development, widely known as the Dublin Principles. These principles, derived from the Dublin Statement on Water and Sustainable Development established in 1992 to inform Agenda 21, focus on human agency, highlighting the importance of public participation for decision-making and recognizing the role of gender in managing water.

IWRM often requires major water policy reform as it rejects the traditional top-down decision-making process as the only way to manage river basins. The role of the government is to contribute to an “enabling environment” where public participation is the basis of decision-making (GWP 2000). This means that businesses and civil society also play an important role in IWRM. IWRM encourages economic instruments such as pricing, tariffs, and water markets to be introduced for efficiency. Policy and legislation need to be revised for efficient water use and also to underscore how river basin management impinges on other sectors such as land and energy. These rules and instruments need to be practiced on the ground by implementing agencies that may not necessarily be mandated or equipped with the appropriate resources to accommodate these changes towards IWRM. Capacity building has been important in facilitating the transition towards IWRM. International organizations such as the Global Water Partnership have set up capacity building initiatives and programs. Overseas development aid has also been directed towards achieving IWRM in developing countries.

While IWRM has been mainstreamed in many government policies, there are many critiques. IWRM is ambitious, with the goal of balancing water use and sustainability, to the extent that it has been criticized for being unrealistic. The three principles are broad so as to be applicable to various river basins with differing hydrological, ecological, and socioeconomic conditions. However, this risks IWRM becoming a framework that has little utility for practices on the ground. IWRM also assumes that public participation is the best way forward for decision-making. Introducing public participation may be difficult in cultures where bottom-up decision-making has not been practiced. In addition, it does not necessarily mean that inclusion of all stakeholders is ensured or in a way that is meaningful, and actual changes to water management may be hard to come by. The debate on IWRM calls into question broader questions of water governance, on the values over the water resources and environment that is being contested, and also the process of decision-making when different rationalities, interests, and motives exist across a range of stakeholders.
International transboundary river basin management as a context for conflict and cooperation

Developing the river for sustainable development calls for river basin management that considers improvement to efficiency of use, including recycling water, reducing demand, and reallocation of water resources. The process of changing existing patterns of allocation and identifying rights to water use is inherently a political process. River basin management is imbued with power relations between various stakeholders: both conflict and cooperation play out in negotiating and implementing river basin management (Mirumachi 2015).

The management of international transboundary river basins is particularly noteworthy, as there are concerns that rivers can be the source of tension between states. There are 276 river basins that are shared by two or more states around the world. These shared rivers can form political boundaries between states or traverse through territories of sovereign states. As rivers change course, boundary disputes may arise. Another problem is often described as potential water war, pointing to the risks of water scarcity. International transboundary river basins are unique in that they pose problems of balancing sovereignty of states with other principles such as sustainable water use.

Especially in a current world order where there is no supranational structure to police state behavior, cooperation over the international transboundary river basin is up to the goodwill of states. Nonetheless, historically, there have not been any incidents of militarized conflict over water resources between states sharing rivers. Rather, joint river management is common practice that can be expressed in the form of bilateral and multilateral agreements and treaties. These agreements can be project-based, such as the bilateral treaty between Lesotho and South Africa in the Orange–Senqu River basin concerning a water transfer scheme. Agreements can also be broad frameworks, as in the case of the Lower Mekong river basin between Laos, Thailand, Cambodia, and Vietnam concerning sustainable development. Often, river basin organizations (RBOs) are set up to facilitate decision-making and as agencies for policy development, like the Niger Basin Authority in West Africa. The International Boundary and Water Commission between the United States and Mexico is responsible for the management of the Rio Grande and Colorado rivers. This RBO has evolved in its mandate over time, originally established to deal with river border issues but now including water and sanitation as well.

River basin management facilitated with agreements and RBOs has seen some diffusion of global norms (Conca 2006). The United Nations Convention on the Law of the Non-Navigational Uses of International Watercourses (UN Watercourses Convention) is a source of reference for global principles such as equitable and reasonable utilization, prevention of significant harm, and prior notification of the water resources development plan to other states. This convention is buttressed by the doctrine of limited territorial sovereignty, which sets forth entitlements to states for the utilization of shared waters so long as it does not cause undue harm or deprive other basin states of their rights (McCaffrey 2007). As a framework convention, the UN Watercourses Convention provides a set of principles that is to be localized to help guide practice for individual basins. For example, with a high concentration of transboundary river basins in the Southern African region, member states of the Southern African Development Community (SADC) have relied heavily on the UN Watercourses Convention text to draw up
the Revised Protocol on Shared Watercourses in the SADC signed in 2000.

However, the extent of cooperation over river basin management requires scrutiny. While there was a remarkable increase in the number of RBOs established around the world after 1950, there are still only 61 organizations. Moreover, they tend to have limited authority and there are cases of overlaps between organizations within a basin (Lautze et al. 2013). Research has focused on the institutional design of river basins to identify how effectiveness can be improved. It has often been suggested that the function of data and information sharing could enhance cooperation over the management of shared rivers. Many RBOs have the task to generate scientific studies and baseline reports of the environmental state of the basin and to facilitate the exchange of information and data, such as the International Commission on the Protection of the Danube River. However, it has been found that rather than exchanging actual quantitative data, indirect mechanisms of data and information sharing such as prior notification are much more widely used between basin states. Data and information exchange can be costly for states; the reluctance can also be attributed to data of the river and water resources utilization considered as a matter of national security (Gerlak, Lautze, and Giordano 2011).

Negotiating river basin management underscores the role of power and power asymmetry between states. The concept of hydro-hegemony demonstrates that decisions of water allocation are subject to the ways in which both hard and soft power are exercised to control and capture water resources, with implications to equity of outcomes (Zeitoun and Warner 2006). The stronger basin state, or the hydro-hegemonic state, may capture water resources by physically building infrastructure to store and divert water, or by ensuring compliance to water sharing arrangements through negotiation and other discursive means. This means that even if RBOs and other river basin management institutions exist, they may not ensure equity, especially for those with less power. The insights on power highlight that geography and riparian position are not the sole determining factor. In other words, while upstream basin states may be privileged to have better geographical access to water resources, downstream states can exercise hydro-hegemony to secure preferable allocation. This is the case of the Nile River basin where Egypt has historically been able to influence water allocation so that it can access a majority of the water resources despite its being the downstream state.

National and local institutions for river basin management

Conflict over water resources in basins at the subnational level can be even more intense than at the international level, with actual instances of violence and death. The Water Conflict Chronology published by the Pacific Institute provides a useful list of violent conflict over water (Pacific Institute 2016). River basin management is compounded by issues of water rights, especially in cases where they are not necessarily formalized in legal entitlements or are complex, tied up with issues of land ownership. Water users’ associations (WUAs) are often set up to ensure rights are recognized to groups of stakeholders.

Moreover, disputes may not necessarily occur between local water users. The issue of “land and water grabbing” shows how actors from outside of the river basin also play a major role in managing water. Land grabbing, or acquisition of lands by both domestic and international investors, reflects the competition for land largely
for agriculture, helped by the globalization of investment and food commodity trade. Land grabbing has implications for water resources because access to land also offers access to rivers and aquifers, not to mention soil moisture necessary for agricultural productivity. The changes to land and thus water entitlements can disrupt the livelihoods and local practice of land and water resources management.

In order to avoid and resolve conflict and to ensure sustainable river basin practices, it is assumed that decentralization of decision-making at the “lowest appropriate level” is necessary, as exemplified in IWRM (GWP 2000). Various institutions have been tested to foster collective action between stakeholders with disparate interests and motivations at the local level. In addition to WUAs, catchment management committees are common mechanisms that enable decision-making at a more localized level than at the larger, basin-wide level that can be led by state actors through national policy and legislation. Multistakeholder platforms (MSPs) refer to the different types of fora that facilitate participatory decision-making.

However, the effectiveness of devolved decision-making has yielded mixed results, showing that it is not a panacea to river basin management. South Africa presents an interesting example where catchment management agencies (CMAs) have been set up as statutory organizations to implement IWRM. These CMAs work in tandem with WUAs made up of actors such as irrigation boards (who may have already organized as a stakeholder group before the formal introduction of WUAs) and other stakeholders such as small-scale farmers. The experience in the Komati basin showed that these new institutions do not necessarily change the power asymmetries between stakeholders and may even risk further marginalizing those with less power (Waalewjin, Wester, and van Straaten 2005). Devolving decision-making is particularly challenging for river basins in developing countries. In these situations, formal institutions recognized by legislation and governmental policy may coexist with informal institutions that have been established over time, based on traditional and cultural practices. In this regard, river basin management entails more than just changes to policy but also to the culture of decision-making.

Consequently, there are scholarly debates about the variety and plurality of decision-making rather than a narrow consideration of river basin management techniques per se; these discussions ultimately touch on the issue of water governance. IWRM has largely drawn criticism because it presents a generalized river basin management model and does not offer ways to deal with the complexity of water governance. There are attempts to understand the polycentric nature of water governance as multiple and nested loci of authority and decision-making. River basin management would thus be informed through a range of stakeholders and experts that can provide tailored solutions for different scales of management (Lankford and Hepworth 2010). Moreover, multiple institutions may interact in a way that they cannot be designed a priori. The implication is that river basin management needs to recognize the ecological and social conditions through which institutions influence each other in a process of institutional bricolage (Cleaver 2012).

The limitations of IWRM have brought about discussions of how river basin management can accommodate the changing nature of socioecological processes that throw up issues of complexity and uncertainty. Collaborative water governance is suggested as one way of taking into account different worldviews, knowledge, and interests of different stakeholders so that they
RIVERS AND RIVER BASIN MANAGEMENT

engage in problem-solving jointly. Public participation is not a given solution in this case, but as something that needs to be developed through time and with trust between stakeholders. Adaptive water resources management has also been suggested as a way to address complexity through social learning. River basin management is thus shaped by the various governance mechanisms put into place and no longer a mere technical issue dealt with by the state alone.

Managing beyond the river

River basin management increasingly requires the consideration of impacts to and from issues beyond the hydrological unit, such as those relating to international food trade, regional energy demand, and global climate change. For example, understanding how demands for energy drive hydropower development is useful and provides a more comprehensive picture of the ways in which water use and allocation can be regulated. This means seeking solutions beyond the river basin. Globalization has also facilitated the expansion of international food trade. Climate-proofing the water sector requires a renewed look at the ways in which river basins are managed. In this regard, considering the river basin as a fixed unit of management may not be helpful anymore. Adaptive and collaborative water governance emphasizes the multiscalar networks of power and knowledge that go beyond the spatial boundary of the hydrological river basin. It can be argued that the river basin is not a fixed spatial entity but a sociopolitical arena in which environmental governance is established and challenged.

SEE ALSO: Dams; Environmental governance; Riparian ecosystems; River basin management and development; Rivers and streams; Sustainable development; Water conflicts; Water resources and hydrological management

References


Rivers and streams

Edgardo M. Latrubesse
Edward Park
University of Texas at Austin, USA

Definition

“Stream” is a general term used to define water bodies that flow downslope due to gravity through a natural open channel. The channel will contain water for at least part of the year. Streams with a significant flow and volume of water are generally considered rivers. Rivers and streams shape the emerged surface of the Earth and ideally reach the ocean, which is the general base level for geomorphologic systems. In other cases, they end inland due to climate (water scarcity), diversions and use by humans, or even geologic controls that impose local or regional base levels (interior basins).

Runoff, drainage basins, and channels

When rain or snow fall onto the Earth, part of this precipitation seeps into the ground to replenish groundwater (infiltration) while another part returns to the atmosphere through evapotranspiration. The remaining water starts to flow on the surface as runoff. The amount of runoff depends on the amount, intensity, and temperature of the rainfall as well as on biophysical factors such as vegetation cover and terrain characteristics such as elevation (rock and soil types), slope characteristics (steepness and length), and landscape modifications produced by human land use. Additional sources, such as melting ice and snow, trigger overland flows which, together with groundwater flowing beneath the surface (throughflow), are the sources for streams and rivers.

Surface water has the capacity to produce erosion and to transport sediments. Runoff tends to concentrate in self-made channels that can be: shaped by the erosion of the terrain surface in a self-building channels network; formed by the alluvial materials that their own system transports and deposits; or built by a combination of both processes. Thus, a channel is a landform outlined by its bed and banks where water, sediments, and solutes are flowing through. Channel development is controlled by the lithological substrate and both water and sediment flows, which results in a variety of channel geometries in nature. They can be relatively stable or unstable features as they can constantly be modified by erosional and aggradational processes. Channels adjust and evolve over time and a wide range of channel types (discussed in the section Channel patterns and floodplains) are found in nature. During flooding periods, when water flow exceeds the channel capacity, flooding waters overtop the channel and diffuse across the surrounding terrain.

Drainage basins include both the streams and rivers that convey the water and the land surfaces from which water drains into the channels. A drainage basin is separated from adjacent basins by a drainage divide, which acts as a geographical barrier such as a ridge, hill, or mountain. Other terms that are used to describe a drainage basin are “catchment,” “catchment basin,” “river basin,” “water basin,” and “watershed.” In a
RIVERS AND STREAMS

technical sense, a watershed refers to a divide that separates one drainage area from another, but in the United States and Canada, the term is often used to mean a drainage basin or catchment area itself. Drainage networks are present in different types of patterns and the channel arrangements depend mainly on geological factors (lithology and structure). Depending on the scale and geologic and geomorphologic complexity of the basins, not only one but several styles may be present. Thus, a drainage basin is formed by a hydrographic channel network that carries runoff and sediments produced by erosion in the drainage basin. Except under certain circumstances, all watercourses in a hydrographic network must be able to flow without interruption from their headwaters (sources) to their mouths (outlets).

The drainage network can end in a river, lake, reservoir, estuary, wetland, sea, ocean, or some flat dry interior basins. The basins can be also classified in function of either their connections or their lack of connection to the ocean. Exorheic basins are those where the drainage systems can reach the ocean while endorheic basins are inland basins with interior drainage.

The scales of the basins are highly variable. These range from small networks of rills with ephemeral or intermittent channels generated by erosion during a single storm event on agricultural land, to a huge and incredibly interconnected channel network of more than 6 000 000 km² such as the Amazon basin. The role of large rivers draining continents is very relevant. For example, the 10 largest fluvial basins of the world (Amazon, Congo, Nile, Mississippi, La Plata, Obi, Yenisei, Lena, Niger, and Amur) are responsible for draining nearly 21.5% of the emerged land of our planet.

However, nearly 20% of all land drains to endorheic basins such as lakes or interior seas. In endorheic basins evaporation is the primary factor of water loss and the water is frequently more saline than the oceans. The largest endorheic basins are in the interior of Asia, such as the Caspian Sea and the Aral Sea basins. Other large basins include the interiors of Australia, the Great Basin in the United States, the Chad basin in Africa, and some basins in Mexico and the Andes, such as Salar de Uyuni and Titicaca basins, among others.

Rivers as conveyors of water and sediment

A drainage basin is the basic unit of the fluvial system. Fluvial systems are open systems, where energy and materials are exchanged in the fluvial basin. The major inputs of the fluvial system are water, sediment particles, and solutes (solid load and dissolved load).

Sediment load consists of particles of mineral and rocks that can be transported as bedload, rolling, sliding, or saltating along the bed, or wash load, which consist of the finest particles transported in suspension by the water flow. Dissolved load consists of ions in solution derived mainly from weathered rocks and, secondarily, of organic materials from soils and other sources of organic matter such as organic litter.

In fluvial geomorphology, the fluvial system is normally described as a process-response system, where the interrelations and adjustments between the geomorphological processes and forms take place through continuous feedback between them. The fluvial system can be divided into ideal-schematic reaches or zones, each operating as a process-response subsystem with its own inputs. Certain processes dominate within each reach or zone. The zone of erosion is located at the headwater regions where most of the sediment originates through erosion processes, supplying sediment to the channel network of
the drainage basin. Sediment moves downslope through the channel network towards the *zone of transport*. The *zone of transport* acts rather like a conveyor belt for water, sediment, and solutes. As the river approaches the ocean or a local base level, sediment transport energy is greatly reduced mainly due to the decline of channel gradient in the *zone of deposition*. In other cases, discharge simply decreases downstream, as can happen in arid regions.

In perennial rivers, the drainage area and water discharge typically increase downstream while slope or gradient generally decrease. This means that slope is normally steeper in the upper reaches and becomes gentle downstream, and that discharge is commonly inversely proportional to the slope. For that reason, the longitudinal profile of a river typically exhibits a shape like a smooth concave curve. In the lower reaches close to the base level, the longitudinal profile is almost flat. The ideal longitudinal profile can be broken in knickpoints by natural factors such as structural controls (hard rocks on the river bed, faults) where rapids or falls can develop, or by human interventions such as dams.

**Channel patterns and floodplains**

Channel form adjustment is largely a function of the water and sediment supplies. Channel adjustments are produced in the channel dimensions (width, depth) and slope, as well as in planform. Along a given reach, channel adjustment can be constrained within certain boundaries or *boundary conditions* that are imposed by local conditions such as valley confinement, valley slope, channel substrate, and riparian vegetation.

Channels can be thoroughly classified as alluvial or bedrock. *Bedrock channels* are cut directly into the underlying bedrock, while *alluvial channels* are formed in alluvium that has been stored in the valley floor by fluvial processes. Channel patterns of bedrock rivers are strongly influenced by the resistant nature of their substrate (structural controls and underlying geological rocks). Bedrock rivers may represent single or multiple channels. Single straight reaches often follow a fault, fractures, or a lithological contact. Multiple channel bedrock rivers are formed by several individual channels cutting their own courses to flow around bedrock “bars.”

In alluvial channels, banks and beds are almost exclusively formed by alluvial sediments. Because water flow and sediment supply change through time, alluvial channels are continuously adjusting the channel shapes through erosional and depositional processes. The alluvial channels are diverse in their three-dimensional (planform and cross-section) style as consequences of the variability of the alluvium property, vegetation covers, and valley settings.

Classification of alluvial channel patterns as a continuum from straight to meandering to braided rivers has been adapted widely in fluvial geomorphology and only slightly changed ever since. For decades, meandering and braided patterns have been the focus of fluvial research. However, anabranching multichannel patterns dominate the largest rivers of the world (Latrubesse 2008). In this context, the most reasonable classification of alluvial channels is considered to be (i) straight, (ii) braided, (iii) meandering, and (iv) anabranching patterns.

*Straight rivers* are a single channel system without significant bends and represent low sinuosity (ratio of channel length to valley length), with an index typically below 1.1. This channel type rarely persists over a long term in an alluvial reach.

*Braided rivers* occur where the flow divides the channel into a series of braided bars by
RIVERS AND STREAMS

accumulating sediment. Braided rivers are bedload-dominated systems characterized by unstable networks of channels, ephemeral bars, and limited riparian vegetation (Ashmore 2013). They possess a relatively high slope in relation to the available discharge and an abundant proportion of bedload in relation to the total load. Mid-channel bars are typically covered by water during the bankfull discharge stage, acting as a single channel system.

Meandering rivers wander across a floodplain forming a series of bends, with a sinuosity index typically above 1.3. The water flow patterns through meanders encourage erosion and undercutting of banks on the outside of bends while deposition and the formation of point bars happen in the inner bends, generating asymmetric cross-sections of sequences of pools and riffles. With river processes, meander bends tend to grow outward followed by chute cutoff and neck cutoff, leaving oxbow lakes and scroll bars on floodplains.

Anabranching rivers consist of multiple channels that are clearly separated at bankfull discharge stage by vegetated semipermanent alluvial islands excised from existing floodplains or formed by within-channel accretions. These rivers occupy a wide range of environments, from subarctic to the tropics. Anabranching rivers construct semipermanent multiple channel systems to concentrate streamflow and maximize bedload transport by constantly adjusting the optimum channel width-to-depth ratio under the conditions where there is little or no opportunity to increase gradient (Nanson 2013). Branches of the anabranching rivers can have a tendency to meander or braid; however, they are generally restricted to short distances laterally from the main channel. The particular importance of the anabranching channel pattern has been emphasized after the observation that Earth’s large alluvial rivers (typically defined as mean annual discharge $>1000$ m$^3$s$^{-1}$ and having very low slopes) predominantly follow the anabranching channel pattern (Latrubesse 2008). Although the anabranching pattern is the ultimate planform adjustment of large rivers, small to mid-sized anabranching rivers displaying various styles can also be found extensively in different climatic and geological regions around the world, indicating the representative nature of the anabranching channel pattern in the fluvial system of our planet.

In alluvial rivers, floodplains are a fundamental component. Floodplains can be defined from hydrological or geomorphological points of view. Floodplains are flat depositional surfaces adjacent to the active channel that are constructed by river processes and built with alluvial sediments. This means that the floodplain is related to the present-day dynamic. However, older units of alluvial materials can be frequently flooded. Because a fluvial system is a natural system with a history, it can be considered that some recent alluvial units of Holocene age that are frequently flooded can be also included as part of the active floodplain.

Engineers usually apply a hydrological definition for floodplains, where it is demarcated as a flat surface supporting river floods with some particular recurrence interval of flooding (10–100 years). The 100-year recurrence criterion has been thoroughly used by urban planning and insurance companies to delimit the areas of flood risk. However, the delimitation of that area is not necessarily a well-delimited natural unit as 100-year floods can spread on different kinds of lithology and landscape components. Floodplains are very important in the fluvial system because they support a variety of environments and habitats, act as an erosional–depositional buffer area during extreme events, provide aquifer infiltration and agricultural land for humans, and can be a sink of sediments and pollutants.
Rivers and Streams

The role of rivers in shaping Earth

Rivers continuously reshape Earth’s surface, mainly through erosion or deposition in sedimentary sinks (sedimentary basins) where alluvial plains form. Depositional environments include floodplain, alluvial fans and megafans, deltas, and complex areas with rivers and swamps in a variety of tectonic settings. However, on Earth’s surface the balance is in favor of erosion, as huge amounts of sediments eroded from the emerged lands reach the oceans each year. Erosion depends on the regional climate, geology, topography, and land cover. The variations of river discharge through time are defined as the river or hydrological regime. River regimes are typically linked to the regional climate and, secondarily, to the catchment characteristics (topography, lithology). However, large basins can spread in a variety of climates and on diverse landscapes.

In erosional processes, about \(1.316\times10^6\) m\(^3\) s\(^{-1}\) of surface waters shape the Earth. Of this total, the Amazon River alone is responsible for approximately 17%. Erosion by rivers annually produces approximately 19 billion tons of sediments, 3.8 billion tons of dissolved solids (Milliman and Farnsworth 2011), and probably nearly \(0.9\) Pg of carbon (1 Pg = \(10^{15}\) g) (Aufdenkampe et al. 2011) that are transferred to the oceans. The proportion of sediments eroded by streams and rivers stored in the continents and not reaching the oceans is still a matter for analysis. Small rivers draining mountain areas such as the islands of Southeast Asia, the Himalayas, the Andes, New Zealand, and others are responsible for producing the major part of the sediments. Other highly productive areas are regions with highly susceptible sediments and rocks (such as the Loess Plateau in China) or small catchments in semiarid Mediterranean climates.

Rivers and humans

Rivers and streams have been used as a source of water in various ways, such as for waterways, as a source of food and water for human consumption, for irrigation, for recreation, and also as a major energy resource through hydroelectric power generation. For these reasons, many of the world’s big cities have been developed close to rivers. Rivers have been altered by humans since the early stages of civilization. Humans produce direct and indirect impacts on rivers and creeks. Changes in land cover and urbanization of drainage areas produce some of the common indirect impacts, while channelization, dams that regulate flows and trap sediments, flow diversions, dredging, mining, and bank stabilization produce common direct impacts. The impacts produced on rivers by humans in the twentieth and twenty-first centuries are dramatic and >45 000 dams above 15 m retain 96 500 km\(^3\) of water, or about 15% of the total annual global river runoff (Nilsson et al. 2005). Hundreds of new dams are planned for construction, mainly in countries of Asia, Africa, and Latin America. Currently approximately 18% of the total electric energy is produced by hydroelectric power.

Rivers, ecosystems, and biodiversity

River systems are like corridors that sustain a diverse mosaic of landscape units. The floodplain mosaics and channel-related elements (islands, fluvial bars, beds, and banks) can hold different types of riparian vegetation (alluvial forests, marshes, meadows), lakes and floodplain channels, alluvial aquifers, and are the habitats for a diverse fauna of aquatic, mixed terrestrial-aquatic, and even groundwater communities. Through hydrological and geomorphological processes, fluvial systems support
and control ecosystems of channel and riparian zones of the floodplain. Abiotic processes, such as floods and erosional and depositional bank processes, constitute a dynamic variable for biological processes (e.g., migration, predation, and competition) while the mosaic of landforms, sediments, and alluvial soils provide sustenance to the alluvial vegetal communities. Biological communities inhabiting fluvial systems include a rich diversity of plants and animals. Plant diversity is greater in tropical rivers than in other regions. More than 900 flood-tolerant species were recorded in the Amazon (Wittmann et al. 2006). Plant species can often be rooted to the channel substrate or can float on the water surface.

Fauna communities inhabiting fluvial ecosystems include reptiles, amphibians, waterfowls, and mammals, such as freshwater dolphins, manatees, beavers, hippos, capybaras, and many others, as well as a large variety of insects. Microorganisms play substantial roles in primary production by storing chemical energy through photosynthesis, decomposition of organic materials, and nutrient recycling.

Invertebrates include an array of benthic communities such as freshwater shrimps, worms, and insect larvae. Types of plants and animals inhabiting fluvial environments depend on geographical locations, because ecosystem functions are characterized by the energy and flows of the channel.

Although freshwater habitats represent around 0.01% of the world’s water and cover only about 0.8% of the Earth’s surface, they support at least 100,000 species (Dudgeon et al. 2006), approximately 40% of global fish diversity (over 10,000 fish species), and one-quarter of global vertebrate diversity. When amphibians, aquatic reptiles (crocodiles, turtles), and mammals (otters, river dolphins, platypus) are added to this, in total it is estimated that one-third of all vertebrate species are confined to fresh water. The richest fish diversity is found in the Amazon basin, where a single river can sustain more than 1000 species of fish. With 801 species, the United States ranks seventh in the world, after Brazil, Venezuela, Indonesia, China, Democratic Republic of the Congo, and Peru, for the greatest variety of fish species.

However, due to the impact of dams, pollution, and other human interventions, fluvial biodiversity is in a state of crisis. It is estimated that at least 10,000–20,000 freshwater species are extinct or at risk as a result.

SEE ALSO: Groundwater; Hydrologic cycle; Riparian ecosystems; Rivers and river basin management; Soil erosion and conservation; Stream ecosystems; Surface water; Water and climate change; Water resources and hydrological management; Water: urban; Watersheds

References


**Further reading**

Road transport

William P. Anderson
University of Windsor, Canada

Road transport (alternatively called “highway transportation”) has become the dominant mode for personal transportation and the fastest-growing mode for freight transportation in many parts of the world. In the United States, automobiles and light trucks account for between 85% and 90% of passenger miles traveled and are the principal means of travel to work for over 85% of employees. (Hereafter “cars” is used to refer to the aggregate of automobiles and light trucks used primarily for personal transportation.) While the railways still move slightly more ton-miles of freight than intercity trucks, the value of the goods shipped in trucks is much higher (Bureau of Transportation Statistics 2013, tables 1-40, 1-41, 1-49, 1-58). Despite the greater use of transit in the European Union, cars still account for about 83% of domestic passenger miles and trucks have a much higher share of domestic freight than railways (Eurostat 2014). However, road shares are somewhat lower for intercountry personal and freight transport in the European Union. While the share of road transportation in most of the developing world is lower than in the United States or Europe, it is growing quickly. It is estimated that by 2035 China will have more cars on the road than the United States.

The rise of road transport has been spurred by a combination of technological innovations and supportive investments in public infrastructure. While the first self-propelled steam-powered vehicles date back to the eighteenth century, the commercial production of four-wheeled automobiles with internal combustion engines began in the late nineteenth century. Initially automobiles were just for the very wealthy, but the application of production line methods, pioneered by the Ford Motor Company in the United States, led to lower prices and much broader market penetration by the late 1920s. Following the Great Depression and World War II, US automobile sales exploded in the 1950s, almost quadrupling their prewar value by 1955; truck sales grew at a similar pace. The combined US sales of cars and trucks peaked at almost 17.5 million in 2005, with total motor vehicle registrations reaching over 240 million – more than the number of licensed drivers (US Census Bureau 2012, table 1098).

A salient characteristic of road transportation is that, while most vehicles are owned by individual households or firms, the infrastructure they travel on is provided by the public sector. This is in contrast to rail, where a few large firms own both the vehicles and the infrastructure. In the first half of the twentieth century the network of paved roads in the United States and most other countries was very limited, especially in rural areas. This limited the scope of road transport, especially for intercity trips. The dominance of road transport was made possible by policies of massive public investment into networks of paved roads. The networks are made up of a hierarchy of road types, with freeways (limited access, no signalized intersections) at the top followed by...
ROAD TRANSPORT

arterials (major roads with intersections), collectors (which feed traffic to arterials), and local roads for access to individual addresses. Local roads generally account for about two-thirds of the network mileage, but carry only about 10% to 30% of the vehicle miles traveled. However, this dense network for local access endows road transportation with one of its greatest advantages over other modes: the ability to provide point-to-point passenger or freight transportation in a single vehicle to a very large number of origins and destinations.

The creation of the Interstate Highway System, beginning in the 1950s, created a network of freeways connecting all major US population centers. At the same time, massive residential and commercial development in the suburbs required the creation of millions of miles of local access roads, along with collector and arterial roads to complete the hierarchical network. This was perhaps the largest and fastest transformation of the built environment via infrastructure development in human history. It had profound impacts on the private economy, on the public sector, and on land-use patterns.

The need to provide road infrastructure places major fiscal burdens on all levels of government. In some cases, dedicated funding mechanisms have been instituted. For example, in the United States a federal gasoline tax is paid into a separate trust fund administered by the Federal Highway Administration and distributed to the states for freeway and major arterial projects. Roads at lower levels of the hierarchy are often the responsibility of municipal governments, who may impose impact fees on new residential and commercial developments to offset the cost of road and other infrastructure required to service them. In many cases, however, such special funding arrangements are not sufficient, so road construction and maintenance must be funded from general public revenues.

Spatial impacts of road transport

The rising dominance of road transportation has had a profound effect on spatial patterns of settlement and economic activity, especially in metropolitan areas. Geographers have long recognized that new transportation technologies affect urban form, as for example when the introduction of electric streetcars enabled the creation of suburbs beyond walking distance from the central business district (CBD) (Borchert 1967). The substitution of the private car for public transportation took this process a step further by undermining the role of the CBD in urban spatial structure. While cars are good at accessing a broad range of dispersed locations, they are not good at delivering large numbers of workers to a compact district because they require too much space for parking. Cities that have maintained strong CBDs (New York, Chicago, Toronto, London), therefore, generally have higher transit shares because rail transit systems are best suited to delivering workers to high-rise office buildings. In more dispersed cities, employers may locate in peripheral industrial parks, preferably situated close to highways, which are easily accessed by car and can provide cheap parking. Shopping and other personal services may choose similar sites. Furthermore, employees who commute by car are free to choose between a much greater range of residential settings, as long as they are on local roads that connect to the broader road network.

The design of highway systems has also served to undermine the CBD and reduce densities. Urban rail systems, as well as early urban highway systems, tend to have hub and spoke structures that create uniquely high accessibility at the central node. The construction of ring roads (known as “beltways” in the United States) that run around the periphery of the urban area, connecting the spokes of the initial network,
creates secondary points of high accessibility where highways intersect, and generally reduces the dominant position of the CBD (Anderson, Kanaroglou, and Miller 1996). Overall, the rise of road transportation is a principal driver in the process of urban sprawl. Furthermore, highways connecting metropolitan areas tend to support the spread of commercial and residential growth along corridors comprising several cities and the rural areas between them.

For large cities in most parts of the world, the dominance of road transportation has led to increasing urban traffic congestion. Congestion may be viewed as a common resource problem. As there is generally no restriction on the use of the public highway network, the number of cars on the road tends to exceed the number that can be accommodated without overcrowding. Addition of new capacity often fails to solve this problem because the availability of new or expanded roads tends to increase the volume of travel, so that congestion relief is short-lived (Downs 1992).

The dominance of road transport is almost as great for intercity trips as for urban trips. In the United States almost 90% of all trips of over 50 miles are made in cars. The number of passenger miles on intercity passenger trains, by contrast, fell by more than 60% between 1960 and 2010. While airplane passenger miles increased eighteenfold over the same period, they still account for only about 7% of trips of over 50 miles. However, because airplane trips tend to be longer, they account for 40% of passenger miles, compared to 55% for cars (Bureau of Transportation Statistics 2013, tables 1-40, 1-41, 1-42).

Road freight

The decision to move goods by road (truck) rather than by rail or water is affected by the relative cost for shipments of different distances. Total transport costs may be decomposed into terminal costs, which are the cost of loading and unloading goods at either end of a trip, and line-haul costs, which are the cost of moving the goods between origin and destination. As Figure 1 shows, terminal costs are higher for rail and water transportation, which require specialized terminal infrastructure, but lower for road, which does not. Because road transportation uses more labor and energy per mile, however, its line-haul cost is higher. Since line-haul costs vary directly with distance, the terminal cost advantage of road transportation is offset by higher line-haul costs for trips of
sufficient length. This partly explains why the average distance for US domestic rail shipments in 2007 was 728 miles, while the average for truck shipments was only 206 miles (Bureau of Transportation Statistics 2010). Airfreight costs are higher over all distances than the three modes shown in the figure, so its use is justified only by the fact that it is faster.

The dominance of road transportation also reflects its wide availability. Since most spatial markets in the United States do not have access to inland waterways, for example, water transportation represents only a small share of domestic freight movement. Reliance on trucks is also reinforced by the low-density urban form associated with dependence on cars for personal transportation. Since most retail stores, warehouses, distribution centers, and factories are located in dispersed patterns around highways, direct delivery of freight by some mode of transport other than truck is often not possible. Shipments moved by rail have to be transferred to trucks at the rail terminal in order to reach their final destinations, and the cost of this transfer offsets the line-haul cost advantages of rail freight. For very long distances, however, intermodal service, where goods are shipped in containers that can be cheaply transferred between marine, rail, and road vehicles, is the best option. This kind of service is especially common for the delivery of internationally sourced goods that are moved by rail from marine ports to inland ports, where they are transferred to trucks for final delivery.

Environmental impacts

A major concern with the growing dominance of road transportation is its poor environmental performance relative to other modes. In the United States, for example, energy use per passenger mile for light-duty highway vehicles is about 2.4 times as great as for intercity passenger rail and even 50% higher than for domestic air travel, although it is only 16% higher than for transit buses. Furthermore, the prospects for long-term reduction in energy intensity is not good, given that this value actually increased by 8.6% between 2000 and 2011 (Bureau of Transportation Statistics 2013, table 4-20). The main reason for the increase is that, despite technological improvements in engine efficiency, consumers are opting for more powerful light trucks (especially sport utility vehicles) over the sedans that formerly dominated the market for private cars. The impact of Americans’ preference for large cars is reflected in the fact that US energy use per vehicle kilometer is about twice the equivalent measure for the European Union (Global Fuel Economy Initiative 2011, figure ES1).

Despite growing energy consumption by cars and trucks, their contribution to urban air pollution problems has actually declined. Emissions of the precursor pollutants to ground-level ozone – principally volatile organic compounds and oxides of nitrogen – have declined dramatically, largely as a consequence of advances in emissions control technologies mandated by federal regulation. As a result, the number of days in which urban ozone levels exceed government standards has fallen.
by one-half or more in many American cities. However, the broader environmental problem of greenhouse gas emissions from cars and trucks keeps growing in pace with energy consumption. Over the period from 1990 to 2009 – during which industrial emissions of CO₂ declined by 17% – emissions from the transportation sector increased by 17% (Bureau of Transportation Statistics 2013, table 4-53). Since cars and trucks generate most of those emissions, America’s dependence on road transportation and its failure to make progress in vehicle energy efficiency are among the greatest impediments to progress on greenhouse gas emissions.

Of even greater concern are the environmental impacts that are likely to occur as cities in the fastest-growing economies achieve levels of car ownership and use similar to those found in the United States and other Western countries. Typically, countries experience rather sudden growth in car ownership when per capita incomes reach a critical level of around US$7000. Many cities in Asia and Latin America are currently reaching that level, and others in Africa are likely to do the same in the next few decades. Because of the lack of environmental standards, and the widespread use of motorbikes (which are especially intensive producers of ozone precursor pollutants), major cities such as Beijing, Mexico City, and New Delhi have pollution levels high enough to cause serious health problems for most residents, and some smaller cities in Pakistan, Iran, and other countries are even worse. Furthermore, the growth in car and truck transportation in such cities generates greenhouse gas emissions that are sufficient to more than offset the reductions achieved in higher-income countries.

SEE ALSO: Environmental impact assessment; Global climate change; Infrastructure; Transport geography; Transport networks; Transportation history; Transportation and land use; Urban geography

References


Further reading

Romania: Asociatia Profesionala a Geografilor din Romania (APGR) (Professional Association of the Romanian Geographers)

Founded: 2008
Location of headquarters: Bucharest
Website: www.apgr.eu
Membership: 290 (as of November 15, 2013)
President: Ioan Ianos
Contact: ianos50@apgr.eu

Description and purpose

APGR aims to increase knowledge about regional change in Romania, the impact of some extreme events and to promote the scientific communication related to the problems encountered in regional planning.

APGR wants to promote the multiscale approach on space among decision-makers at all levels to increase geography’s general visibility. APGR acts through its central office and seven regional divisions, each of which decides how best to get involved in community support. APGR believes that geographers have the power to influence decision-makers and other professionals to have more respect for space and its place in regional and community planning.

Journals or major publication series

Geograful (Geographer). www.apgr.eu/geograful.php

Current activities or projects


Submitted by Ioan Ianos
Romania: Institutul de Geografie al Academiei Române (IGAR) (Institute of Geography of the Romanian Academy)

Founded: 1944
Location of headquarters: Bucharest
Website: www.geoinst.ro
Membership: 60 (as of December 31, 2013)
Director: Dan B˘alteanu
Contact: igar@geoinst.ro

Description and purpose

The Institute of Geography (IGAR) is the main institution of fundamental and applied geographical research in Romania. It has two branches in Ia¸si and Cluj-Napoca and carries out integrated research in physical geography, human geography, environmental geography and geographic information system (GIS) focusing on investigating the relationships between the components of the physical and biotic environment and of the social and economic milieu, as well as their spatial distribution in connection with global environmental change. Its expertise in studying man–environment relations (integrated studies of natural and technological hazards; studies of community vulnerability to extreme events) has been the focus of several national and international projects.

Journals or major publication series


Current activities or projects

The members of the institute participate in a variety of national and international interdisciplinary programs and hold membership in several IGU commissions. IGAR is involved in projects financed by the Romanian Academy, such as the Environmental Geographical Atlas, integrated studies of the Lower Danube River and of the Carpathian Mountains; studies of the settlement system in Romania in the context of sustainable development; and in international projects (Framework Programs 6 and 7 of the EU).

Brief history

The Institute of Geography was established in 1944 by the Ministry of Culture and National Heritage on the initiative of Professor Vintila Mihailescu. The then Geographical Research Institute of Romania, as it was named, had two branches, one in Cluj, the other in Ia¸si. In the beginning it was staffed by members with an interest in research, but without professional training. After 1952, the Institute trained its own specialists in various fields of geography. Between 1958 and 1974, it functioned as the research institute of the Romanian Academy, one year later (1974–1975) it moved to the Ministry of Education and Instruction. In 1975, its administration passed to Bucharest University. As of February 7, 1990, it returned to the Romanian Academy. Today, the Institute’s scientific activity is coordinated by the Section of Geonomic Sciences of the Romanian Academy.
ROMANIA

The Institute of Geography sponsors activities at PhD level and has six doctoral supervisors. IGAR annually hosts an International Summer School on Natural Hazards and Sustainable Development at its research station located in the Carpathian Mountains.

Submitted by Dan Băleanu
Description and purpose

The Societatea de Geografie din România (SGR), one of the oldest geographical associations in the world, was founded to bring together geographers, educate them, and help them develop scientific activities. It is the leading association of Romanian geographers and collaborates with other professional associations (geomorphology, human geography, limnology, and professional geographers). Its purpose is to develop geography, mainly among teachers of secondary and high schools and establish relations with geographical societies from abroad. The society has a management board and coordinators at county level.

Journals or major publication series


Current activities or projects

The SGR continues the traditions of Romanian geography developing several educational programs with a focus on conferences, symposia, and roundtables in various regions of the country; maintains contacts and participates in meetings of professional societies, especially those from neighboring countries; organizes various educational activities (summer schools and different training courses); and publishes scientific research results in two journals.

Brief history

In 1875 the SGR was established, with the king of Romania as its president. This act was the result of geographical activities developed within the framework of the Romanian Academy (founded in 1886). The society had three sections: mathematical geography, physical geography, and ethnology. In 1884, the bulletin of the SGR (*Buletinul Societății de Geografie din România*) was sent to 72 foreign geographical societies. Between 1886 and 1896, 32 geographical and historical dictionaries of Romania’s counties and the *Great Geographical Dictionary of Romania* (1900) were published. Before the communist regime came to power in Romania, the society was coordinated by Professor Simion Mehedinți, the founder of modern Romanian geography and author of the volume *Terra: Introduction to Geography as Science* (still relevant in terms of global human pressure on the environment). In 1912, the SGR was renamed the Romanian Royal Society of Geography until 1947, when the communist power was instituted. From 1949
ROMANIA

to 1968, the name was the Society of Natural Sciences and Geography. From 1968 to 1989, and then after the fall of communism in 1989, it has been known by its current name.

Submitted by Dan Băleanu
Routing and navigation

Kai-Florian Richter
University of Zurich, Switzerland

Navigation is a coordinated and goal-directed movement of one’s self through the environment (Montello 2005). It combines two different skills: locomotion, which is proximally coordinated movement, and wayfinding, which is distally coordinated planning.

Routing is the process of path selection in a network. In a geographic context, it refers to determining a path from an origin to a destination through a transportation network, such as a street network, a subway system, or a system of corridors in a building. While in principle navigators themselves can and do perform routing, that is, plan the path, routing is typically considered an activity that comes from the outside, that is, it is delivered by some external aid (another human or a computational system).

Navigation is sometimes considered to be a technical activity executed only by experts, such as pilots or ship captains. But in fact, it is a fundamental process in the daily lives of humans and nonhuman animals, and a crucial ability for mobile robots (machines). Without navigation, we literally would not get anywhere. This entry will focus on human navigation and path selection, and algorithmic routing to support human navigation.

Human navigation

Navigation consists of two components: locomotion and wayfinding. Locomotion is used to move the body in a coordinated fashion through an environment, that is, without bumping into obstacles or tripping over one’s feet. It is especially relevant for the proximal surrounds, that is, the environment that is directly accessible to sensory and motor systems (Montello 2005). Generally, locomotion solves immediate behavioral problems, such as standing sure-footed, obstacle avoidance, and moving in the right local direction. There are various ways of performing locomotion (walking, running, crawling, etc.) and modes of locomotion (on foot, on a horse, by bike, driving a car, riding a bus, etc.). These different modes of locomotion have a large influence on how we perceive the environment and, therefore, how we acquire and process environmental information.

Wayfinding, on the other hand, is a goal-directed and planned activity. It is about getting around an environment in an efficient way (Montello 2005). Being a planning act, it requires higher cognitive functions and much more conscious thought than locomotion. In most cases, wayfinding involves reaching a specific destination in an environment, which is often not in the proximal surrounds but rather somewhere not directly accessible to the navigator’s senses. Successful wayfinding requires memory, that is, knowledge about the destination’s location and a way leading to it, or external aids, such as maps, a human guide, or a navigation system, that direct the way. Wayfinding also addresses behavioral problems; however, they are less immediate (to an extent) but of a decision-making nature, such as choosing a route or planning a multistop trip.

The term “wayfinding” was introduced by Kevin Lynch (1960) in his seminal book The Image of the City (actually as ‘way-finding’).
ROUTING AND NAVIGATION

Lynch had then already looked back at a long list of previous research, but dismissed the idea of seeing our ability to find our way as being some “mystic ‘instinct’, but rather as the consistent use and organization of definite sensory cues from the external environment” (1960, 3).

The distinction between coordinated movement in the proximal surrounds (locomotion) and the efficient getting to distal locations (wayfinding) already imposes some kind of hierarchy in the processes involved in navigation. And viewing wayfinding as decision-making, the planning process itself has a hierarchical nature. Once a destination has been chosen, intermediate locations will be identified that need to be reached successively in order to get to the final destination. Reaching these intermediate locations may again require more fine-grained planning, for example, even more intermediate locations, or changes in transportation mode, or switches between goal-directed and search behavior (see below). Often, these hierarchical plans will not be fully fleshed out before embarking on the journey to the destination. This may happen because not everything is known beforehand and some of the required information needs to be retrieved from the environment directly once in situ. Alternatively, some unforeseen changes may occur while traveling through the environment, for example road blockages or canceled public transport, that require finding an alternative route. Or some possible pathways (shortcuts), which are known in principle, may only be remembered when passing them in the environment, which will lead to updates of the plan.

For example, reaching a specific office on a university campus first requires getting to that campus, which may be a known route. Once on campus, the navigator will look for signage or a map pointing to the building in which the office is located. Once the building is identified, a route to that building from the current location will be planned and executed, which may involve reaching some in-between locations, for example, the central square on campus. Upon entering the building, information on how rooms are laid out in the building will be sought, along with discovering on which floor the office in question is located. Assuming the office is located on the fourth floor, the next task will be to find an elevator and take that up to the floor in question. Leaving the elevator, the corridor in which the office is located needs to be searched, which may either involve looking for more signage or employing some exploration strategy.

The decision-making involved in such a wayfinding process makes use of different levels of spatial knowledge. This knowledge changes over time. Repeated navigation in an environment will lead to changes in the structure of spatial memory about that environment. Initially, when new to an area, a navigator will have no knowledge about this specific environment; however, from life experience there will be some background knowledge about how environments of that kind are generally structured and how they function. After a little exposure, landmark knowledge will emerge. This kind of knowledge describes the state where a navigator is able to recognize certain geographic objects and places in the environment, having seen them before. But connecting these places, that is, directly navigating from one to another, is still difficult, as the routes are not (fully) known. Acquiring this route knowledge will be the next step, which happens after even more exposure and repeated travel in the environment. The final qualitative change in spatial knowledge comes once the environment is so well known that navigators can detect shortcuts between places and connect places and routes to form new routes that previously have not been traveled.
This model of spatial knowledge acquisition by Siegel and White (1975) simplifies and idealizes aspects of learning environments – as any model does. For example, it assumes that navigators never consult external aids or ask other people for assistance, which would accelerate knowledge acquisition. In general, the separation between the three stages, landmark knowledge, route knowledge, and survey knowledge, is not as clear-cut as the model postulates. It has been shown that some form of survey knowledge, namely metric distance information, is acquired early on in learning an environment. Still, this model provides a useful taxonomy of different types of spatial knowledge that is the foundation for a range of other research in spatial learning, and spatial cognition more generally.

To sum up this far, in a nutshell the wayfinding process can be described as getting to a location at which a decision has to be made (a so-called decision point), taking that decision, and then moving on to the next decision point. Locomotion keeps us on track in-between. Still, not every wayfinding process is the same. Consider the differences between commuting to work, taking a pleasure walk through the local woods, finding an office in a university building, or getting to a hotel from an airport. These wayfinding tasks differ in the amount of attention they require, the knowledge needed to successfully execute them, the importance of making the right decisions, and the need for external aids.

Several classifications – or taxonomies – of wayfinding and its different tasks have been proposed over the years. A very first broad distinction is the separation discussed above of locomotion as the process of moving around in the immediate surroundings, and wayfinding as the process of planning to move to distal targets. A further broad distinction can be made between unaided and aided wayfinding. Unaided wayfinding either requires at least partial knowledge about the environment under consideration, or it becomes a search task where navigators try to locate their destination. Aided wayfinding essentially only requires the correct execution of the instructions provided, which may be more or less cognitively demanding depending on the kind of aid used.

Allen (1999) defined three different wayfinding tasks. He distinguishes between exploratory navigation, travel to a familiar destination, and travel to a novel destination. We experience exploratory navigation, for example, when relocating to a new place to live and then exploring the neighborhood to familiarize ourselves with the surrounds. Our daily commute is a typical example of traveling to a familiar destination, and finding the way from the airport to our hotel at our summer holiday destination is a typical example of traveling to a novel destination.

Wiener, Büchner, and Hölscher (2009) introduced a more elaborate taxonomy of wayfinding tasks. Their taxonomy specifically takes into account different levels of required knowledge and task constraints. They focus on unaided wayfinding; aided wayfinding is subsumed as a single high-level item of their taxonomy. In unaided wayfinding, their first differentiation is between undirected and directed wayfinding. Undirected wayfinding occurs if there is no specific destination, for example, if taking a one-hour pleasure walk where the aim is to get some exercise but it does not really matter where one goes. If a navigator has survey knowledge about the environment, such a kind of wayfinding may be considered an actual pleasure walk; if no survey knowledge exists, it may rather be considered to be exploration or roaming.

Directed wayfinding, on the other hand, has a destination, that is, the aim is to get to some specific location. If this destination is unknown, search behavior needs to be evoked in order to get there.
It is conceivable that the exact destination is unknown even though the environment is well known (the navigator has survey knowledge). This may be the case if looking for a friend who is supposed to be in some downtown bar. In this case, it is possible to execute an informed search. If one has no knowledge about the environment (neither destination, nor route, nor survey knowledge), the person is likely to embark on some uniformed search quest for the destination.

If the destination is known, a target approximation behavior can be used. If the route to the destination is known, this behavior becomes relatively simple, as simple route following will get the navigator to the destination (e.g., the daily commute). If the exact route is not known, this route needs to be found first. If the navigator does not possess survey knowledge of the area, this will be a proper path search behavior. If survey knowledge is present, finding the route is a matter of path planning – (at least partially) previously untraveled but known path-segments need to be connected such that a route emerges between origin and destination.

Many of these tasks are now supported by devices in our cars or on mobile devices, which guide our way using digital maps and/or voice instructions. Their algorithmic foundations and communication principles will be discussed in the remainder of this entry.

Routing and routing algorithms

Routing is the process of path selection in a network. Routing determines which parts of the network have to be traversed in which order to get from some origin to a destination. Several criteria may drive the routing process. Typically, the shortest (in terms of distance) or fastest (in terms of time) route is being sought. However, other criteria that may be selected include the most scenic route, the safest route, the route easiest to describe, or the route that avoids specific parts of the network altogether.

Routing usually has to account for static and dynamic aspects of moving through a network. In other words, it is important to distinguish between the features physically present in an environment independent of any navigation and their role in the process of navigation. Klippel (2003) introduced the concepts of structure and function in wayfinding. With structure, he refers to an environment’s physical features. The structural level describes a static configuration of these features. Function denotes the relation of these structural elements to actions performed in the environment. The functional level demarcates those features relevant for a navigation action. It describes a dynamic navigation situation.

This distinction between structure and function is reflected in routing by differentiating between path and route (Klippel 2003). A path is a linear, unbounded feature in the environment upon which travel occurs. A route represents a behavioral pattern. It is directed and bounded by having an origin and a destination. A route demarcates a path, that is, it determines those parts of a network that are traversed during a trip. These parts are called path-segments. The points where path-segments meet are termed branching points. The paths, that is, branching points and path-segments, form a path-network, a graph-like structure, which reflects the geometric layout of the paths in an environment. Branching points correspond to nodes and path-segments to edges in the network. From a functional perspective, route-segments become relevant, which correspond to those path-segments demarcated by a route. The point where two route-segments meet is termed decision point. At a decision point, a navigator needs to decide on the further
direction to take. Decision points correspond to branching points on the structural level.

Simple routing ignores the dynamic aspects of navigation and adopts a static view of the network. It focuses on the structural level. The two best-known path search algorithms are Dijkstra’s shortest path and A*. They operate on a graph representation of the network, whereas the vertices (the set $V$) represent branching points and the edges (the set $E$) the path-segments. Both algorithms require that the edges of the network be annotated with a “cost” greater than zero of traversing this edge (e.g., geometric distance or travel time). In a nutshell, starting at an origin node, Dijkstra’s shortest path algorithm selects in each step that node that can be reached from any of the already visited nodes with the lowest costs. It stores for every node the cheapest, that is, shortest or fastest, path found so far for reaching this node from the origin. Therefore, in the end, for each node the shortest path to the origin is known. The A* algorithm uses a heuristic to speed up the search; it performs a best-first search. Similar to Dijkstra’s algorithm, A* stores for each visited node the costs to reach this node from the origin. Different from Dijkstra, it uses an admissible heuristic to estimate the distance to the goal from the current node. “Admissible” here means that the algorithms must not overestimate the distance to the goal. With these prerequisites fulfilled, it can be implemented efficiently (i.e., it runs faster than Dijkstra’s algorithm) and is guaranteed to find the shortest (“cheapest”) path.

From a dynamic perspective, the focus is on direction changes at decision points, that is, turning from one route-segment onto another at decision points. These turns may differ in their costs, for example, because of differences between turns in the need to slow down, or in terms of the cognitive complexity of understanding what to do. In terms of the graph used for routing, modeling these turns involves two edges and the vertex at which they meet. If cost functions for path search are to account for the dynamics of navigation, the basic network representation of a graph $G = (V, E)$ will not work because efficient algorithms assign a single cost value to each edge. Therefore, to capture the dynamics of navigation the complete line graph is used as a representation of the network (Winter 2002). The complete line graph is the graph $G' = (E', \varepsilon)$, where $E'$ is the set of edges in $G$ – though direction of edges is ignored. $\varepsilon$ is the set of vertices in $E$ that share their “middle” vertex, that is, $\varepsilon = \{(v_i, v_j), (v_j, v_k) \in E \times E\}$. That is, in the complete line graph, each possible turning action is represented by an individual edge, whereas nodes in the complete line graph represent the edges of the original graph. By traversing an edge from one node to another in the complete line graph, in the underlying original graph the said turning from one route-segment (edge) onto another via a decision point (vertex) is manifested.

Over the years, several routing algorithms have been developed that account for cognitive aspects of navigation. Duckham and Kulik (2003) presented “simplest paths,” an algorithm that accounts for the cognitive complexity of correctly turning at a decision point. Different types of intersections are classified according to complexity of describing a turning action at this intersection. This classification determines the costs for the routing. That is, the underlying cost function applies to the complexity of navigation decisions rather than to travel distance or time. “Most reliable paths” (Haque, Kulik, and Klippel 2007) minimize the chance of choosing a wrong turn at a decision point. The fundamental assumption is that navigation instructions (e.g., “turn right”) are generally easy to understand, but they may result in wrong decisions being made if they are not unambiguous, that is, if there are several possibilities to turn right
at a decision point. The algorithm calculates an unreliability measure for each turn. This simply counts the number of turns that are instruction equivalent – all turns that are described by the same linguistic variable. Another algorithm, the “landmark spider” (Caduff and Timpf 2005), routes navigators through a network along local landmarks located at decision points. Costs are determined by a landmark’s salience, its orientation with respect to the navigator, and its distance to the route-segment. The “landmark spider” algorithm calculates the “clearest” path, that is, that path that leads a navigator along the most suitable landmarks.

Richter and Duckham (2008) extended “simplest paths.” Their “simplest instruction paths” algorithm also finds the route associated with the lowest costs in terms of instruction complexity. The algorithm is based on Dijkstra’s original algorithm, but additionally uses optimization criteria that reflect dynamic aspects of navigation and direction giving, and avoid ambiguity in instructions. When selecting the node with the currently lowest costs, the algorithm spreads all the instructions with which that node has been reached forward through the graph. This step checks whether neighboring nodes can be reached with the same instructions. Spreading of instructions through the graph corresponds to traversing several decision points with a single instruction (as in “follow the river until the roundabout”). In the routing process, this reduces costs as the number of instructions that need to be communicated gets reduced.

Routing services and route instructions

Routing services employ additional strategies beyond the path search algorithms (Dijkstra or A*) described above. As most of them operate on a national or even world-wide scale, path search would be much too slow – in fact computationally infeasible – if every street was treated the same. Therefore, the street network is organized into a hierarchy to expedite search. The environment is divided into several regions. Search on local streets is only performed within these regions. If the path requires leaving a region, connections are only made via a network of major streets. This can be further nested. For example, when calculating a path from a residential address in one city to a residential address in another city, routing services would first find a path from the local residential streets to the major streets of the city. Next, along these major streets a connection to the system of highways is calculated, which is then used to get path search (close) to the second city. Here, search is reversed. From the highway system, an entry to the network of the city’s major streets is determined, along which the destination’s region is reached. Finally, within the destination region a path to the actual destination is calculated using the network of local streets. For such path search strategy to work, it is important that each network level has some dedicated entry (and exit) points to each region. For the highway system, this is straightforward as dedicated entries and exits (the ramps) already exist. For the other network systems, for each region these points need to be defined. The more points are defined, the closer the calculated path gets to the actual shortest path, but also calculation will get slower as more nodes need to be taken into account. Thus, a challenge for routing services is to find the right balance between accuracy and calculation speed.

Most conventional routing systems only account for distance and travel time in their calculation. They ignore cognitive factors as they have been outlined for some of the algorithms discussed above. Many of the navigation systems commercially available today are also able to
dynamically adapt to changes in the environment. This not only refers to their ability to calculate a new path on the fly if the navigator deliberately or accidentally decides to ignore an instruction. More importantly, they also account for network load, namely congestions or slow moving traffic, in calculating a path. This information may be received via radio transmission (as the driver would from traffic news on the radio) or be inferred from behavioral (movement) data of other users using the same system. Users’ movement behavior may be sent back to a central server, which then updates the network status. If several users move only slowly along some part of the network, there is most likely some congestion through which other users should not be routed.

In communicating the calculated routes, navigation systems also rely on distance and time. Typically, they provide information about the distance and length of time a street has to be traveled on before the next turn. Additionally, they name the street into which a navigator has to turn. A typical instruction would be “turn left into Beach Road and stay on it for 748 m (3 min).” In dynamic situations, that is, if the system travels with the user, the approaching turn would be announced as well, for example, in an instruction such as “in 100 m turn left,” which then gets repeated (with a different distance) once or twice when getting closer to the turn (“in 50 m turn left”; “turn left now”).

Human communication of route information uses different principles. People make extensive references to landmarks, that is, salient objects in the environment that (they believe) are easy for others to identify. Landmarks are fundamental in our mental representation of an environment, as discussed above in the context of the Siegel and White (1975) model. Humans use landmarks as anchor points in their spatial knowledge and arrange other spatial objects relative to these landmarks.

Given the high prominence of landmarks in human spatial memory, not surprisingly they feature prominently in human route instructions as well. Landmarks serve several purposes in such instructions. Importantly, they identify decision points. In other words, landmarks anchor wayfinding actions in space. They indicate when and where a specific action needs to be executed (“turn left when at the church”). Further, landmarks can be used to indicate whether a wayfinding error has been made (“you have gone too far if you see the church”), or indeed that the correct route is still being followed (“along the way you will pass a church”). Any object that stands out from the environment (in the given context) may serve as a landmark. An object may stand out visually (the red house among all white houses), structurally (the house located at the major intersection), or culturally (the old town house).

Overall, the following principles of good route directions have been identified in cognitive and linguistic research (Denis 1997; Lovelace, Hegarty, and Montello 1999).

- The principle of spatiotemporal order: route instructions should mention locations and actions in the same order as they will be experienced when traveling along the described route.
- The principle of referential discrimancy: as already mentioned, landmarks should be mentioned at points where decisions have to be made in order to link the action to the event of detecting the landmark.
- The principle of mutual knowledge: when describing a route to somebody else, the instruction giver should choose objects and delimiters describing topological, directional, and distance relations according to the
ROUTING AND NAVIGATION

communication context, in particular what the instruction receiver can be expected to know. Specifically, this also includes referring to objects (landmarks) such that they become identifiable (e.g., “the big red house” instead of “St Mary’s College” when talking to someone unfamiliar with an environment).

Furthermore, human route instructions usually have more linguistic variations than computationally generated ones. Navigation systems tend to have a single template for instructions, whereas people use general discourse principles also in route instructions, even if the focus is strongly on delivering relevant information. In particular, they employ a strategy termed “spatial chunking” (or aggregation), which combines several local instructions into a single higher-level instruction that subsumes those local instructions (e.g., “turn left at the third intersection”; “take the third exit at the second roundabout”). The intermediate local instructions are left implicit, but are also not needed for someone to follow the described route successfully.

More generally, people are well able to adapt the level of granularity in their instructions to the context, particularly the complexity of wayfinding situations or the knowledge about a route they can assume their communication partner to have. For example, significantly more information is needed to reach the train station from some distal origin than is required to travel by train between cities. Or people usually do not require detailed instructions for getting to the highway from their home location, given that they will have traveled that route several times before. But they may then need detailed instructions for taking the right exit and navigating to an unknown address in a foreign city. Human instruction givers consider all these with ease (even though there are still misunderstandings, of course). Navigation systems, on the other hand, always use the same level of details for instructing their users. They always deliver detailed turn-by-turn instructions and, indeed, always use the same instruction structure for all cases.

Current developments

The discrepancy between the principles of good human route directions and the way routing services instruct their users still sees a range of active research – even if it has been researched for more than a decade now. Recent developments with respect to user-generated content and “big data” have spurred researchers to seek these new data sources for improving automatically generated route directions. One direction of approaches exploits the newly available vast data resources (e.g., photo sharing repositories, panorama views of streets, user-contributed mapping approaches) to mine for information that may support navigators. For example, using methods of machine learning and geographic information retrieval, some approaches aim at identifying landmarks in an environment based on how many photographs of a particular location have been contributed and by whom. Others use methods of image processing to extract shop signs from street panorama views and then present them to navigators in a highlighted fashion to ease orientation at intersections.

Another direction of research looks at a dedicated collection of information relevant for navigation using principles and methods of user-generated content. This may be a direct approach, that is, setting up a system for explicitly asking users to contribute landmark information (“which object in your surround is a landmark and why?”). Or it may use a more game-like character where such systems are supposed to entertain their users while the game setup
requires them to contribute relevant information (e.g., photographs of salient geographic objects) as a byproduct. This byproduct may then be exploited to augment navigation instructions, possibly employing techniques similar to those used in the first direction of research, albeit with less noise in the data as the collection process has been more controlled.

Given the universal access to outdoor positioning via the Global Positioning System (GPS) and the many location-based services that can be built on the availability of a user’s position, there is increasing interest in also transferring such services to indoor spaces. However, for indoor spaces, no such universal positioning infrastructure exists, and GPS simply does not work indoors as it requires line of sight to the satellites. Different approaches have been developed for determining a user’s (or device’s) position indoors; the most promising at the moment seems to exploit existing wireless network infrastructure. By dividing an indoor space into a regular grid and then mapping the signal characteristics received from different antennas at each grid point (so-called fingerprinting) it becomes possible later on to calculate the most likely position by matching current signal readings with the stored ones. An object’s position can be determined by the grid point associated with most similar signal characteristics from the wireless access points.

On the assumption that indoor positioning will become available eventually, researchers further look into the integration of outdoor and indoor services by, for example, enabling a seamless routing from some indoor location to an outdoor location – or another indoor location with the route crossing through parts of outdoor space. A particular challenge here is to unambiguously determine whether a user is actually indoors or outdoors, which is most challenging in the transition spaces where signals from both indoor and outdoor positioning systems may be received. Recent work has shown that while users in their navigation can easily cope with being mispositioned by several meters, they consider services as highly unreliable that accidentally position them as being outdoors while they are actually indoors. This is true even if the geometric error is not greater (or even smaller) than a mispositioning error that still positions them in their actual environment. The semantic error is seen as much more severe than the geometric error, which is important to consider when designing such indoor/outdoor positioning services.

Finally, the increasing use and ubiquity of navigation services affects spatial knowledge acquisition. Using a navigation service to guide to a destination has consequences similar to the backseat driver effect – we hardly learn anything about the traveled route, let alone the environment in which we navigate. This is not surprising. One of the main ideas of navigation systems is to largely take over the wayfinding part of the navigation process. Users are still required to locomote (e.g., not to bump into other cars on the road), but any planning is done by the system. The users only need to execute one local instruction after another (“turn left here,” “turn right next” …) without getting any overview on what the overall plan is. That is, these systems generally fail to provide proper spatial context. In combination with the discrepancies between human direction giving and the way such systems instruct their users, it becomes difficult for navigation system users to remain oriented towards the destination (or origin for that matter) and to remember the route later on. This may not seem like a problem since the system is there to support; however, if that system breaks down, people may be unable to recover and would be essentially lost. Also, the lack of context results in users feeling insecure in the navigation
ROUTING AND NAVIGATION

process and dependent on the device. Research has identified all these issues; the next step now would be to design navigation assistance that at the same time supports users and lowers cognitive demands of the navigation process, but still provides a context and allows users to learn about the environment during a navigation.

SEE ALSO: Behavioral geography; Cognition and spatial behavior; Geocomputation; Graph theory; Information technology and mobility; Location-based services; Spatial thinking, cognition, and learning; Transport networks

References


Further reading


Rural citizenship

Richard Yarwood
Plymouth University, UK

“Citizenship,” like “rurality,” is a highly contested term. It has widely been used to describe a person’s relationship with a nation-state and, in particular, the rights and duties that are associated with it (Smith 2000). Of late, this idea has been challenged by geographers who have pointed to the importance of spaces above and below the nation-state in the formation and practice of citizenship (Desforges, Jones, and Woods 2005; Yarwood 2014). The concept of transnationalism recognizes that the practice of citizenship may cross national boundaries and engage citizens with political and cultural processes at a global level. At the same time, localities provide an important context for engagements such as voting in local elections, writing to councilors, volunteering to provide local services, staging protests, or simply living out daily life as a citizen. Citizenship is therefore fluid and multiscalar and much more than just a person’s relationship with his or her nation-state. Anderson et al. (2008, 34) contend that:

Citizenship is increasingly organized and contested through a variety of non-state as well as state institutions. This extends citizenship in the cultural sphere, to describe people’s senses of belonging in relation to places and people, near and far; senses of responsibility for the ways in which these relations are shaped; and a sense of how individual and collective action helps to shape the world in which we live.

Investigations of citizenship have tended to focus on urban areas, perhaps reflecting that its etymology refers to the inhabitants of cities. Yet, emerging research has suggested that distinctive forms of citizenship are becoming associated with the countryside and deserve closer scrutiny.

It is widely acknowledged that rurality does not shape social relations per se. Distinct forms of economic development and political conflict, together with different ways of imagining rural space, influence how citizenship is imagined, contested, and performed in rural places. Although significant differences exist within and between rural spaces in the majority and minority world, it is possible to discern a “global countryside” that has common characteristics. These include the presence of:

- globalized commodity chains and agri-food systems;
- the growth of transnational corporate investment and networks;
- the supply and employment of migrant labor;
- flows of global tourists;
- non-national property investment;
- the commodification of nature;
- large-scale exploitation of primary resources;
- social polarization;
- new sites of political authority;
- political contest. (Woods 2011)

Halfacree (2007) argues that rural space has three facets. It is simultaneously a locality that reflects the outcome of productive and consumptive economic activities; it is represented, for example through the much contested rural idyll; and it is something that is played out and given
meaning through the performance of everyday lives. Significantly, political contest means that these three elements do not always sit easily with one another, with the result that rural space may be disjointed or chaotic in nature. These three aspects of rurality have the potential to shape, and to be shaped by, different practices of citizenship, as the following sections reveal.

The imagined countryside and citizenship

As Halfacree’s (2007) model recognizes, social constructions of rurality have significant bearings on rural society. Hegemonic views of the countryside have been enrolled into discourses of citizenship and national identity. Heritage and folk traditions have been appropriated to evoke the idea that a nation is somehow more authentic if it has rural roots. This is evident in museums that link imagined folk cultures with nationhood and in folk songs that associate rural landscapes and people with national identity. In Sweden, for example, rural folklore was used to instill a sense of national identity in the early twentieth century (Buttimer and Mels 2006). By contrast, those unable or unwilling to appreciate these hegemonic views of the countryside were positioned as “anti-citizens” (Matless 1996). The active exploration and understanding of the countryside was seen as important in developing these forms of citizenship. In the 1930s the Scottish Youth Hostel Association sought to develop a sense of national identity by encouraging working-class youths to engage physically and bodily with the Highlands (Lorimer 1997).

Indeed, the countryside has often been viewed as a training ground for citizenship. One contemporary example is provided by the Duke of Edinburgh’s Award, a voluntary scheme undertaken by young people in Commonwealth countries. Participants are required to undertake an “unaccompanied, self-reliant expedition with an agreed aim” in a rural or “wild” setting. The countryside is seen as a testing ground in which young people perform skills that are deemed to make them good citizens, including teamwork, leadership, self-sufficiency, fitness, inquiry, resolve, and confidence. Organizations such as the Scouts, the United Kingdom’s National Citizen Service, and the annual Ten Tors Challenge in Dartmoor National Park (UK) also draw on rurality to test and shape future citizens through camps, residential projects, and challenging walks respectively (Yarwood 2014). Rural settings are seen to provide opportunities for citizenship to be embodied and performed, although there is an expectation that the skills learned will then be applied in everyday (perhaps urban) settings.

These kinds of practices reproduce dominant views of rurality and certain expectations of citizenship. At the same time, hegemonic and conservative visions of the countryside combine to exclude some groups of people from full participation in society. Thus, the discourses of heritage and citizenship discussed earlier often imply a white history, contributing to a sense that rural space is white space. Equally, indigenous people, such as Native Americans or Indigenous Australians, are curiously absent from both the imagination and the reality of rural space. Indeed, Aboriginal Australians were not granted full citizenship until 1968 and, until then, were only allowed limited access to rural towns.

Painter and Philo (1995) state that, if people cannot be present in public spaces without feeling “out of place,” it is hard for them to consider themselves full citizens at all. While these issues are not confined to rural areas, they are nevertheless exacerbated in rural settings because of greater visibility, the hegemonic imagination of rural space, and a lack of support.
services. Women are also expected to conform to particular gender roles, especially in farming; gay people may hide their sexuality because of conservative values in their society; racial and ethnic minorities may feel isolated; nomadic lifestyles may be illegal; young people may be barred from public space; and disabled groups may find it harder to access rural places. At the same time, rural areas can also be seen as a place to which others can be banished: rural places have been used to house prisoners, asylum-seekers, the mentally ill, or indigenous people, who are kept out of sight and mind in remote reservations and institutions.

The next section examines how the language of rights and duties has been deployed to understand and resolve some of these issues.

Rural localities, rights, and duties

There are significant differences in the standard of living between urban and rural places. According to the United Nations, 71.6% of rural people at a global scale live in extreme poverty, including 1801 billion who live on less than $2 a day and 1010 million on less than $1.25 a day. In the United States, the most persistently poor counties are nonmetropolitan; in Australia infant mortality rates in remote communities (12 per 1000) are significantly higher than in major metropolitan areas (6 per 1000). Tonts and Larsen (2002, 135) frame the differences between urban and rural areas in the language of human rights: “as governments withdraw, or fail to provide, certain services and infrastructure the human rights of rural people are diminished.” By implication, rural people are unable to achieve full citizenship because they are unable to access the welfare rights afforded to their urban counterparts.

In some countries, this reflects a form of local rather than national citizenship (Smart and Smart 2001). In postwar China, for example, there was a formal divide between urban and rural hukou. In the countryside welfare was place-specific, whereas urban welfare was based on particular enterprises. This has meant that citizens only have access to welfare in specific parts of the country, which limits their ability to travel and seek work. Outside their home areas they have been treated as second class citizens and tolerated only if the state did not need to provide for them. This has limited their ability to travel not only to urban areas but also to more prosperous rural ones too. The situation is similar to the experiences of international migrants seeking work outside their own country.

The example illustrates Cresswell’s (2009) assertion that one has to be mobile to be a citizen. In the West the development of national systems of welfare untied people from their home localities by offering welfare based on universal rights rather than a reliance on local charity. Yet poor or nonexistent transport networks render many people living in rural areas, especially the old, young, poor, disabled, and women, into immobile, semi-citizens trapped by rural localities. Cresswell (2009) argues that citizenship relies on “prosthetic” materials, such as shops, services, employment, and transport, to achieve full social and welfare rights. The daily trek for clean water or the closure of a local post office suggests that many rural citizens lack the supports needed to enable them to participate fully as citizens of their wider society.

There have been various efforts to develop rural places that have had important implications for rural citizenship. Forms of endogenous development have been associated with the “modernization” of rural places. These include state-led (or quasi-autonomous) development agencies that may not be directly accountable
RURAL CITIZENSHIP

to local people, as well as forms of private capital such as food-processing plants of global corporations that are powerful by virtue of their position as monopolistic employers. In terms of citizenship, exogenous development is frequently associated with the imposition of new forms of political authority that cut across and restrict existing networks of governance. These center on economic productivity rather than social and political equality, reinforcing existing structures of inequality. Thus, efforts to modernize rural China have improved per capita incomes and led to a boom in consumer spending but, at the same time, have contributed to a growing gap between country and city.

In an effort to counter these effects, more endogenous forms of development have been encouraged that rely on forms of active citizenship that emphasize the duty of citizens to contribute to their localities. Citizens are increasingly required to fill gaps left behind from the neoliberal rollback of the state by, for example, running their own services and working as a community to supplement state provision. There are three reasons why this form of development has been favored in rural areas. First, rural areas have been more likely to suffer from the withdrawal of state services (as witnessed by the decline and closure of public services), and are therefore more likely to rely on citizen action to fill gaps in state provision. Second, there has been a long-standing obligation, evidenced in many countryside policies, that rural areas should provide for their own needs. Examples include community-run shops, voluntary policing, locally built housing, and health care. Finally, rural areas are perhaps better placed to engage in this form of local participation. The lowest tier of formal government, such as parish councils in England or communes in France, are found in rural places, perhaps offering greater opportunities for citizens in rural areas to engage with government than their urban counterparts. Many rural policies have encouraged partnerships between the state and private and voluntary sectors, which offer further opportunity for citizenship engagement in local decision-making and action. The European Union’s LEADER program is one such example that has not only encouraged local action but a form of transnational citizenship that links rural localities to other places in the wider EU supra-state.

Yet, rural communities are far from autonomous and local action in them is usually scrutinized and managed by government agencies, especially where it draws on state funding. Local organizations act as a proxy for government and, rather than empowering communities, these schemes simply aid the rollback of the state.

Furthermore, the idea of community is frequently used to impose unity and obscure diversity beneath a banner of communal identity. Notions of community can exclude as well as include, and often imply a rather bounded, insular view of rural space that seems oblivious to the significance of outside connections. Often “community views” are those of the elite or wealthy: powerful farming interests still dominate local politics in some places, and in others the interests of new rural elites are to the fore. In South Africa, McEwan (2005) has argued that established gender roles made it difficult for women to participate in consultation exercises, rendering the practice of citizenship “a meaningless concept.”

Marginal/third space

Although policies of active citizenship have failed to transform the countryside profoundly, rural places can offer space for new, more radical, forms of citizenship to emerge. Reimagined rural places have provided spaces for new utopian communities to emerge that are based
on faith, gender, green politics, political extremism, nomadism, or a desire to live sustainably (Halfacree 2007). These have their own forms of membership, structures of decision-making, and, by implication, forms of communitarian citizenship that seek to disengage their members from the state. Although these groups strive toward new forms of citizenship, they are prone to disintegration as a result of internal tensions or state legislation to counter them. As the next section explores, people have been more successful when they have adopted transnational rather than isolationist stances.

**Transnational ruralities**

One of the characteristics of the global countryside has been a “depeasantization” of rural places (Woods 2011) by neocolonial, exogenous, and exploitative forms of transnational capitalism. This has led to landlessness, loss of rights, and the suppression of local cultures, and contributed to migration from rural places to urban ones or, more significantly, across borders to work (legally or illegally) in spaces of primary production. At best, these denizen workers have few rights and can be subject to exploitation or even slavery. Despite this, many countries have focused on tightening their borders and placing ever more stringent requirements on migrants who have sought to gain citizenship. Such actions remind us that de jure notions of citizenship are still closely regulated by nation-states.

At the same time, transnational actions have been launched to support those marginalized by global capitalism. The campaign for fair trade emerged in the 1980s to connect Western, urban consumers more closely with “distant,” “other” producers of food in the majority of developing countries. The movement seeks to develop non-exploitative trading relations by paying producers a guaranteed price to ensure the sustainable production of crops and by investing a social premium in social, environmental, and economic projects. By acting as consumer-citizens, those in the West are encouraged to use their purchasing power not only to make personal ethical decisions but also to support a politics of change. These types of transnational coalitions have the potential to empower the most excluded rural citizens. Thus, cooperatives of female artisans have not only used transnational opportunities to develop trade, but also provided an important and alternative platform for local women’s voices.

Depeasantization has also prompted local resistance and the emergence of transnational networks aimed at empowering poor rural populations. In South America, peasant movements have successfully mobilized indigenous communities and identities to address common concerns. Building on social networks left in place by prior rounds of political and religious organizing, indigenous groups have used unions, churches, nongovernmental organizations, and even state networks to mobilize across communities in order to demand rights and resources (Yashar 1998). As well as linking local sites of resistance, networks have, crucially, been used to foster transnational support.

One of the most prolific networks has been the Zapatistas, a Mayan resistance movement from Chiapas, Mexico, which emerged in response to unfair trade, exogenous exploitation of resources, and the loss of power and land. The movement gained international support through effective use of the Internet and collaboration with activists at a global scale. Another example is the Via Campesina (International Peasant Movement), which was formed in Belgium in 1993 to defend small-scale agriculture against corporate and transnational companies. It aims to bring together “peasants, small and medium-size farmers, landless people, women farmers, indigenous
RURAL CITIZENSHIP

people, migrants and agricultural workers from around the world” and claims to have 164 organizations in 73 countries representing 200 million farmers. These forms of new social movements are autonomous, pluralistic, and transnational, occasionally crystallizing in particular (and often urban) protest sites. Their actions represent a form of transnationalism that is concerned with global rather than national citizenship.

Conclusions

This entry has used the lens of citizenship to examine a range of actions in rural areas. On the basis of this evidence, it is possible to draw two broad conclusions. First, citizenship is beneficial to rural studies and, second, better understandings of citizenship can be gained by a focus on rural places.

Within rural geography, Paul Cloke (2006, 26) has argued that there is a need for “theoretical hybridization which can combine, for example, the concerns of the cultural turn with those of political and economic materialism.” Thinking more closely about rural citizenship is one way of fulfilling this call. Citizenship is concerned with understanding how broader political structures shape, and are shaped, by wider changes in society. At the same time, it is concerned with individual identity and performance. It offers a chance to bridge the personal and performative aspects of the cultural turn with the structural and institutional foci of political and social geography within various fluid spaces and places. As Susan Smith (2000, 83) argues, the concept of citizenship “marks a point of contact between social, cultural and political geography.”

Whether there is a distinct form of rural citizenship is open to debate, and reflects the way in which rurality is conceptualized. Using Halfacree’s (2007) model reveals that rural localities have been subject to distinctive but differentiated forms of social, economic, and political restructuring that, on the one hand, are leading to a “global countryside” with common characteristics but, on the other, are producing very different experiences of rurality. Nevertheless, these wider structural changes provide the context for citizenship action (or inaction) in rural places. Social constructions of rurality have also been deployed to fix the identity of and to mobilize citizens, be it “country people” in the United Kingdom or landless campesinos in South America. Thus people who consider themselves “rural people” may be coerced to engage with a variety of issues and rights that are broadly associated with the countryside. Recognizing the diversity of identities and actions under the banner of “rural” contributes to understandings of citizenship as multilayered and fluid. Rural citizenship is also performed in a variety of different ways. These range from overtly political actions, perhaps campaigning for rural issues, to more everyday performances required by rural citizens simply trying to live out their lives in rural societies. Closer investigations of citizenship therefore have the potential to improve understanding of rural areas.

A closer focus on rural places can benefit understandings of citizenship. Rural citizenship in the West has often been associated with rather parochial concerns and small-scale disputes concerning the impact of development on the rural setting (Woods 2011). Very often these debates revolve around different ways in which rurality is represented (idyll or productivist workplace, for example), which in turn reflect changes in the social structure of a locality. Too often rurality has been associated with “community” and, as a result, has been rather inward-looking and concerned only with local places.

Yet, as this entry has shown, rurality in the developing world has the potential to frame more
radical transformative forms of citizenship. As the example of the Zapatistas shows, rural space, often considered peripheral, offers a site for radical, transformative actions that have the potential to jump scales to impact on wider society. Transnational rural actions represent an attempt to develop a global civic society and, with it, a citizenry that challenges the conventional association of citizenship with the nation-state. It is perhaps significant that campaigns such as fairtrade are rural campaigns, aimed at supporting and transforming the lives of people in rural places. Although urban areas often provide the setting for rural protests (the Zapatistas, for example, first occupied cities in Chiapas), it is from and within rural places that some of the potentially most transformative citizen actions are occurring. Far from being peripheral to citizenship, rural places have the potential to develop truly radical forms of citizenship.

SEE ALSO: Citizenship; Environmental citizenship; Globalization and rural areas; Rural geography; Rural policy and politics; Rural society in the Global North

References


Further reading


Rural geography

Paul Milbourne
Cardiff University, UK

The development of rural geography

Rural geography has emerged as an increasingly important subdiscipline of human geography during the last two or three decades. Rural geographers have engaged productively with critical theory and methodological innovation to make important contributions to human geography. “The rural” has also emerged as a more significant arena of investigation for other human geographers, particularly in relation to research and publications on nature and food. This increasing interest in rural geographies has occurred during a period of rapid urbanization in global terms, with 54% of the world’s population now residing in cities and this proportion predicted to increase to 66% by 2050 (United Nations 2014). While these two trends would at first sight appear contradictory, the growth in the urban population has been largely concentrated in countries of the Global South whereas rural geography has almost exclusively been concerned with rural spaces, issues, and processes in the Global North.

Using the Journal of Rural Studies as a proxy measure of recent scholarship in rural geography, we find that 64% of articles published between 2012 and 2014 were written by authors from Europe, 17% by those in North America, and 16% were produced by scholars based in Australia or New Zealand. Only 2% of authors contributing articles were from Asia, 1% from South America, and no articles were published by academics in Africa. Although urbanization processes are apparent in Global North countries, these have been much less pronounced than in the Global South. There have also been strong counter-urban processes evident in many of these countries with rural areas witnessing economic growth and population increases resulting from movements of urban residents to rural regions. In addition, rural spaces are increasingly being drawn into national and international debates surrounding climate change and the security of food, water, and energy supplies. Consequently, rural spaces have continued to attract considerable research attention.

A more detailed reading of these bibliometric data reveals another spatial peculiarity – that of the significance of the United Kingdom to rural geography, with 31% of all articles published in the Journal of Rural Studies in 2012–2014 written by scholars from the United Kingdom. This figure is equivalent to the total output of authors from all other countries in Europe, more than double the output of Australian scholarship, and almost triple the number of articles published by US academics. Indeed, when we explore the development of rural geography, and particularly that relating to recent theoretical and methodological advances, we are largely talking about the changing face of UK rural geography. Elsewhere, rural scholarship has been associated with other social science disciplines, which has perhaps taken away some of the impact of rural geography. In the United States and some European countries, for example, rural sociology and agricultural economics have largely provided the institutional settings for rural research, while

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg1023
in many countries of the Global South rural scholarship has predominantly been undertaken within development studies.

How one measures the significance of a sub-discipline is a question open to debate. Certainly, rural geography has an institutional base within human geography: for example, it has a presence within the main geographical associations in the United Kingdom and the United States; and rural geography can be found within the courses of many geography degree courses in the Global North. As with most subdisciplines, though, the nature and influence of rural geography has waxed and waned over the years. It is generally accepted that rural scholarship occupied a prominent position within human geography for much of the first half of the twentieth century as regional and ethnographic studies contained a strong rural component. As these accounts of regions and places became marginalized by human geography's increasing engagement with spatial modeling, so rural geography lost its prominent position within the discipline. Not only did the burgeoning quantitative geography lose touch with individual places, but spatial modelers constructed urban space as offering more opportunity to practice their newly acquired skills, allowing them to more easily package people, firms, and services into neat geographical units for the purposes of spatial analysis. That said, the translation into English of Christaller's central place theory and Von Thunen's agricultural land-use model in the 1960s did provide a rural dimension to quantitative geography.

The emergence of humanistic and structural approaches within human geography in the 1970s did little to raise the profile or change the nature of rural scholarship. There was some engagement with humanistic geography as rural places and landscapes were explored in relation to their position within art and literature, but this work involved those based outside rural geography. In relation to structuralist geography, it can be suggested that there was a double neglect of rural spaces, with rural scholars largely eschewing Marxist theory and the new generation of radical geographers using the city as a case study of the crises in and of capitalist systems in the 1970s. Indeed, throughout the 1970s, rural geographers seemed to identify more with long-established empirical themes in rural planning, rural sociology, and agricultural economics – such as land use, services, community, and agriculture – than to emerging geographical approaches.

The early 1970s did mark a major milestone in the development of rural geography. In the seminal text on the subject, Clout (1972) sought to assemble what he saw as the key components of rural geography. His book included material on rural population change, the urbanization of the countryside, rural land use, economic restructuring in rural areas, agriculture, rural transport, and integrated rural management. While noting the decline of rural geography, Clout did feel able to point to signs of a resurgence of interest in rural scholarship, evidenced by increased numbers of rural geographers in UK universities, an expanding policy based research agenda, and more rural geographers employed by the planning profession and rural policy organizations. Unfortunately, not only are recent theoretical developments in human geography largely absent from the book, but Clout appears to celebrate the applied and policy relevant nature of rural geography and its close relations with rural planning.

By the early 1980s, calls were being made, in the United Kingdom at least, for more sophisticated forms of rural scholarship that could engage with emerging theoretical and methodological approaches in human geography and cognate disciplines (see Cloke 1980). The establishment of the Rural Economy and Society Study Group...
(RESSG) in 1979, partly as a reaction to the uncritical and narrowly focused research agenda of the Agricultural Geography Study Group of the Institute of British Geographers, provided an important platform for the development of theoretically informed research on political economy, feminism, localism, and welfare (see Bradley and Lowe 1984; Lowe, Bradley, and Wright 1986; Marsden, Whatmore, and Lowe 1990). The work of the RESSG together with a small number of other rural geographers did much to shape the future directions of the subdiscipline, encouraging rural scholars to engage critically with theory and method, and to detach themselves from the worlds of planning and policy. By the mid-1980s, the first book on rural social geography had emerged (Phillips and Williams 1984), which applied theoretically driven work in urban social geography, particularly concerning political economy and managerialism, to rural geography. By the end of the decade, not only had political economy become a widely accepted theoretical lens with which to view rural restructuring, but these more critical accounts were beginning to appear in mainstream geographical publications (see Cloke 1989).

The 1990s witnessed the “coming of age” of rural geography, with increased engagement on the part of rural researchers with emerging theoretical and methodological approaches in human geography, leading to rural geography being taken more seriously by the geography community as a whole. Political economy, particularly relating to agricultural restructuring, remained an important focus for rural scholarship but it was the emergence of the so-called cultural turn within human geography that provided the spark that ignited rural geography. Prompted by Philo’s (1992) review essay on the neglect of “others” within rural geography, a growing number of rural geographers began to explore cultural differences, identities and marginalization processes in the context of rural places. What followed were influential debates about the role of power and culture in processes of exclusion and marginalization (see Murdoch and Pratt 1993) as well as a series of case studies of various groups, including women, young people, travelers, people living in poverty, gay men and lesbians, and minority ethnic groups, who had been largely hidden from the gaze of rural geographers (see Cloke and Little 1997; Milbourne 1997).

The changing relations between nature and society also began to attract increased attention from rural geographers. Ideas of sustainability were utilized by some to explore existing tensions and potential synergies between economy, society and environment in rural areas. Others engaged with actor network theory to highlight the complex relationships between human and nonhuman actors in the rural environment. Still others pointed to the commodification and exploitation of rural nature within the capitalist system. The politics and cultures of rural nature were also subject to academic scrutiny. Research highlighted the significance of nature within dominant discourses of rurality, and how such discourses had been embedded within certain national planning systems and utilized by campaign groups to contest rural development. In addition, place-based studies pointed to the importance of nature to the experience of rural living and how the natural attributes of and nature-based activities in rural areas could cut across social class divisions (see Bell 1994; Whatmore 2002).

In more recent years, rural geography has continued to flourish, deepening and broadening its theoretical, methodological, and empirical interests and making important contributions to critical debates in human geography. Rural scholars have made important interventions in relation to political economy and neoliberalism, including research on the agri-food system,
globalized ruralities, workfare, and new governance arrangements. Critical attention has also been given to how emerging themes in cultural geography, such as consumption, embodiment, performativity, nonrepresentational theory, and mobilities, are played out in rural settings. However, it has been food that has emerged as the most significant research theme with rural geography during the last few years. Returning to the *Journal of Rural Studies*, 13 of the 25 most downloaded articles in the first three months of 2015 were on this subject. In a sense rural geography would appear to have gone full circle, reconnecting with early research agendas that were dominated by agriculture, but the more recent focus on food has been different in three senses. First, food has been explored in much broader and critical terms, with attention given to processes of production, regulation, and consumption, alternative as well as conventional food systems, and issues of food poverty, justice, and security. Second, food research has blurred the boundaries between rural and urban spaces, with some of the most exciting developments concerning alternative food systems in the city. Third, and connected to this last point, recent work on food has been undertaken by rural and other geographers.

**Questions of rural definition**

Rural geographers have had a keen interest in, and some would say an unhealthy obsession with, defining the spatial limits of their field of study. If this were a review of urban geography then little would need to be said about how researchers define urban spaces and places. Cities are seen as clearly delineated and possessing neat boundaries that lend themselves to precise statistical definition. By contrast, the question of how “the rural” is defined has emerged as a major concern for rural geographers. Early attempts to define rural space were very much focused on its physical attributes and functions, including such things as extensive usage of land, low population densities, dominant forms of economic production, and distance from urban centers.

Recognition of the changing physical and social composition of rural space together with engagements with postmodern geographies in the early 1990s led rural scholars to approach “the rural” less as a functional entity and more as a space that is socially and culturally constructed by multiple actors. While the idea of “the village of the mind” had been in circulation for many decades, attention was given to how meanings of rurality are constructed and circulated through works of art and literature, advertising, mass media publications, political and policy discourse, and direct experiences of rural spaces. It is now widely recognized that any particular rural place will comprise multiple social spaces, as different users of rural spaces experience and understand rurality in a variety of ways, some of which may conflict with each other.

Drawing on Lefebvre’s writings on the production of space, Halfacree (1993) proposed that “the rural” is most appropriately understood in relation to three components – its sign (rurality), its signification (meanings associated with rurality), and its referent (rural space) – with the relations between these components becoming more complex as more diverse meanings are attached to rurality. Important questions have also been asked about who has the right to define rurality. With conventional definitions devised by academic and policy actors, it is argued that rural geography has tended to marginalize “lay” discourses of rurality and there have been calls for the development of more sophisticated and sensitive accounts of rurality that can navigate a course between academic, policy, and lay discourse.
The contours of contemporary rural geography

Shifting demographics and sociocultural recomposition

Interest in demographics and sociocultural change has long been evident within rural geography. For much of the twentieth century, academic (and policy) attention was paid to rural depopulation given that this was the dominant form of demographic change in countries of the Global North. Associated with job losses and low levels of pay, particularly in primary industries, this outflow of (mainly young) people became a potent symbol of rural decline, both “on the ground” and within human geography. From about the 1970s, this situation began to change in some countries, with rural researchers in the United States and United Kingdom identifying what has come to be termed counter-urbanization (see Boyle and Halfacree 1998). Initially involving rural places located close to major cities and then rippling out to remoter rural areas, these urban to rural movements of people have received considerable scrutiny from rural scholars over recent decades. Research has highlighted the spatial, temporal, and social characteristics of rural migration as well as the motivations behind relocations. Hidden within this meta-narrative of rural population growth, though, are rather complex geographies of migration within rural regions, with overall population growth often masking pockets of depopulation, particularly in remoter areas. In addition, many rural regions continue to record net outmovements of young people.

There has been particular interest in the social and cultural dimensions of these population shifts. Migratory processes have been awarded a great deal of symbolic value, acting as a proxy for the vitality of rural regions and places, and being drawn into debates about the meanings of rurality. Attention has been given to the ways in which migrants have reshaped rural society and altered many of its cultural norms. In the United Kingdom there has been a clear class dimension to inward migration with urban middle-class groups gentrifying rural housing markets and displacing low, and, in some places, middle-income groups from their home communities (see Phillips 2002). The middle classes have also imported particular cultural constructions of rurality, most notably those bound up with rural idyll and landscape aesthetics, which are often at odds with the realities of the working countryside. In other countries the situation is more complex as rural population growth has been associated with the migration of different class groupings. Indeed, there is evidence of lower income groups relocating to declining rural regions in order to take advantage of low-cost housing opportunities.

Linked to this last point, attention has been given to the movement of international economic migrants to rural places. In the United States, work has highlighted how the rapid growth in the Latino rural population is offsetting population loss in declining rural regions. In the United Kingdom, economic migration has involved workers from eastern and central European countries, following the enlargement of the European Union. Unlike traditional types of economic migration, large numbers of these migrant workers have bypassed cities and moved directly to towns and rural areas. In economic terms, these types of migration are reinforcing the low wage nature of many rural economies as migrants gravitate toward low skill and low wage agricultural and service sector employment. In other ways these migrants are challenging established cultural norms through the import of new forms of culture and cultural interaction into rural places that have little history of multiculturalism.
RURAL GEOGRAPHY

In many countries of the Global South, rural depopulation continues to represent the dominant demographic trend as significant numbers of rural residents relocate to cities in search of employment and improved quality of life. The consequences of these movements are significant for both rural and urban areas in several ways. First, cities have become increasingly hybrid spaces as rural and urban cultures and cultural practices coexist within these expanding urban spaces. A second consequence is that rural migrants often find themselves in vulnerable situations, occupying poor quality housing, in insecure and low paid employment and, in some countries, excluded from the welfare rights enjoyed by urban residents. Third, these movements have impacted detrimentally on the places and people that migrants have left behind, draining rural regions of their young and economically active populations.

Another important feature of changing rural demographics is aging. Although research has largely concentrated on countries of the Global North, it remains the case that, across the world, aging represents a significant trend impacting on rural places. The main reason for this is perhaps obvious from reading previous sections of this entry. As younger groups leave rural areas, so they alter their population structures. Added to this, in countries of the Global North the attraction of rural places to retirees in metropolitan areas has accentuated the rural aging process. When we look at statistics on rural demographics in different countries, it is clear that older people comprise a much larger proportion of the population in rural areas than in cities and that rural places are aging at a faster rate. There are both positive and negative consequences that flow from this aging of the rural population. In terms of the former, older people provide an important community resource, with their ability and willingness to act as volunteers enabling local services to be sustained. In other ways, though, large and increasing numbers of older people impose additional pressures on the local welfare provision in rural areas, particularly in terms of health, social care, and public transport services (see Skinner and Hanlon 2015).

Recently, researchers have begun to reposition rural migration within the more critical context of mobilities. It is suggested that too much attention has been placed on uni-directional, longer distance, and permanent relocations between urban and rural places, which has simplified the complexity of rural population movements. Not only does demographic change involve movements to, from, and through rural places, but mobility represents a significant shaper of rural places and rurality. Within the mobilities literature, the city is viewed as the archetypal space of hypermobility whereas rural places tend to be discussed in terms of stability and stasis as anchor points that provide a sense of tradition and belonging. The reality of rural living, though, is that mobility has become an increasingly important prerequisite for remaining in or moving to a rural place, particularly when many services and retail facilities have ceased to be provided locally. However, some of the taken-for-granted aspects of physical and virtual mobility in the city, such as public transport, pavements/sidewalks, high speed broadband and 4G mobile phone technologies, are more difficult to access in rural areas, which means that being mobile presents a set of challenges to rural residents. Beyond the everyday realities of mobility, engagement with the mobilities literature also permits new ways of interpreting rural places – seeing them less as static and bounded entities and more as spaces that are continually being made and remade by complex flows of people as well as nonhuman actors, objects, and images (see Milbourne and Kitchen 2014).
Rural economies

For much of the twentieth century research on the rural economy was focused on primary industries and agriculture in particular. Indeed, for a long period rural geography was synonymous with agricultural geography. In the United Kingdom, for example, the Agricultural Geography Study Group of the Institute of British Geographers only renamed itself as the Rural Geography Study Group in 1982. Furthermore, in some countries rural research continues to be institutionalized within departments of agricultural economics. It is now widely recognized that the rural economies of countries of the Global North have witnessed a transition from agriculture and other primary sector industries toward the manufacturing and services sectors during the last few decades. In some cases, this transition has been caused by the decline of and crisis in the primary sector resulting from global economic competition, which has created problems for particular (remote) regions largely reliant on resource industries, such as mining, forestry, or fishing for employment. In others, the primary sector has declined in relative terms as manufacturing and service sector firms have increasingly constructed rural locations as attractive spaces for investment.

Notwithstanding these economic trends, it should be recognized that rural areas remain important in terms of natural resource industries. In some countries, forestry and the production of timber and pulp remain significant components of the rural and national economy. In others, major opencast mining activity continues in wilderness areas. Rural spaces also supply essential natural resources to urban populations. Upland lakes and reservoirs provide the bulk of drinking water to cities. In terms of energy production, nuclear and hydroelectric power production tends to take place in rural regions, and wind and solar farm developments are being increasingly located in rural spaces. Attempts have also been made to use natural resources for the purposes of creating sustainable forms of economic development in rural areas. Labeled by some as the eco-economy, the aim has been to create synergistic relations between economic development and the environment, through such things as eco-tourism and green energy businesses.

It is claimed that many rural labor markets provide mainly low skill and low wage employment opportunities. One consequence of this is that young people with qualifications and skills are required to migrate to urban areas in order to secure appropriate work. Another is that rural labor markets are characterized by high levels of underemployment, with people in jobs for which they are over-qualified; in-work poverty resulting from low rates of pay; and multiple job holding, which is how many rural workers secure a sufficient income. The fact that rural labor markets are dominated by small and medium enterprises has implications for securing employment in rural areas. With these companies utilizing largely informal methods of recruitment, social standing in the community can have a significant influence on people’s ability to secure work. In some rural places, the public sector represents a significant employer, providing jobs in education, local government, and health care. Such employment tends to provide improved working conditions and levels of pay than those associated with the private sector, although the introduction of austerity policies in many Global North countries has led to public sector cuts and reductions in local government employment.

While agriculture now comprises a small proportion of the rural workforce in many countries, it has retained its prominent position within rural policy discourse and rural research. For example, the European Commission’s rural
development program remains largely focused on agri-environmental initiatives rather than dealing with structural obstacles facing the rural economy. In terms of research, rural geographers have continued to dedicate considerable attention to agriculture and in recent years have focused on the changing composition of the agri-food sector and its role within global food security. In relation to the first of these themes, it is generally agreed that the second half of the twentieth century was associated with the industrialization and globalization of the agri-food system as farm sizes increased and food production and processing were scaled up and operated as capitalist enterprises. Accompanying these processes were new forms of state regulation of agriculture as national governments introduced policies to ensure a sufficient supply of food at affordable prices for their populations.

By the turn of the twenty-first century, some researchers began to identify a postproductive turn in the agri-food system, involving a shift away from industrial production methods and a broadening of the role of farming to encompass environmental and social considerations. Key characteristics of this new form of agriculture included less intensive modes of production, the reduced use of industrial chemicals, the diversification of farm businesses, the development of eco-system services to benefit rural landscapes and environments, and the production of niche commodities, such as regional foodstuffs. Others have questioned the significance of this transition, arguing that postproductivism remains a niche activity that should be positioned within rather than as having displaced the dominant industrial agri-food regime (see Evans, Morris, and Winter 2002). Furthermore, it is argued that the emergence of postproductivist modes of agriculture has been associated with only a small number of countries in the Global North, most notably those located in the European Union where the Common Agricultural Policy has sought to encourage these new forms of agriculture.

More recently, multifunctionalism has been proposed as a more appropriate lens through which to understand changing modes of agricultural production (Wilson 2007). Within multifunctionalism farming is constructed as making a range of contributions to the rural economy, rural society, and rural environment, in relation to food production, conservation, biodiversity, and economic support for peripheral rural areas, with the specific mix of functions provided varying across different rural regions and types of agriculture. Multifunctionalism has been viewed by some as an alternative to dominant industrial regimes of agriculture. Others have claimed that it represents merely another twist in the neoliberalization of the global agri-food system and a further attempt to commodify the (positive) impacts of farming on landscape and nature.

Elements of postproductivism and multifunctionalism have also been discussed in relation to forestry. Research in North America and the United Kingdom has highlighted how state agencies and private corporations are reconstructing industrial forestry in more-than-productivist terms. Important here has been the emergence of community forestry, whereby forestry is viewed as providing environmental and social benefits to local populations as well as contributing to wellbeing and social inclusion policy agendas. This connects with broader discourses of multifunctional forestry that place value on forests and trees in noncommodity terms, for example, in relation to the social goods they provide and their contribution to environmental improvements and landscape restoration.

Returning to agricultural research, food security has emerged as an important theme in rural geography in recent years. The growth in the global population has exerted new pressures
on the agri-food system and climate change is threatening the viability of agriculture in particular regions of the world. In addition, we have witnessed dramatic fluctuations in food prices resulting from the devastating impacts of extreme weather conditions, such as floods and droughts, on the production of food, which has led to food riots in certain countries. Critical attention has been directed to neoliberal responses to these “crises” and particularly those that involve the application of technical fixes. One such fix has been the development of genetically modified organisms (GMOs) that are capable of producing crops more resistant to disease and the impacts of climate change. More recently, academic attention has shifted to another proposed technical solution to food security problems, that of sustainable intensification, which promises to increase crop yields from the existing agricultural land resource without damaging the environment through the use of new biotechnologies. Research has focused on the political economy of these developments and, in particular, the position of multinational corporations in promoting technical fixes that maintain the integrity of the industrial agri-food regime.

As with other aspects of rural geography, the focus of much of this agricultural research has been on the Global North. References have been made to Africa, Asia, and South America, but largely in relation to how the global agri-food system relies on the import of cheap foodstuffs from these regions. The practice of “land grabbing” is also beginning to receive scrutiny from rural scholars. Involving national governments and major corporations purchasing or leasing large tracts of rural land in developing countries in order to supply food to their domestic markets, land grabbing has been constructed as a new phase of neo-colonialism, involving the displacement of people from their lands, the downgrading of farmers to laborers, and the imposition of new forms of industrial production out of balance with the environment. Indeed, it has been suggested that the solution to global problems of food security lies not through the expansion of industrial agriculture but through the promotion of local, small-scale, and sustainable farming that the agri-food corporations are actively seeking to destroy.

New food geographies

Food has emerged as one of the most exciting areas of rural geography. Moving beyond discussions of the conventional agri-food system, food researchers have engaged with alternative processes of food production, exchange, and consumption as well as the different scales and spaces bound up with the food system. While this interest in food has done much to raise the profile of rural geography, it is has also blurred its subdisciplinary boundaries in three senses. First, food research has been increasingly associated with geographers working beyond the rural subdiscipline. Second, food scholars are located in a range of disciplinary areas, including human geography, economics, sociology, anthropology, and cultural studies, and there now exists a loosely defined body of research labeled food studies. Third, the focus on food has complicated the boundaries between rural and urban spaces.

In relation to this spatial complexity, we have witnessed increased research interest in urban food systems. Food has always been a key part of the fabric of the city, grown in domestic gardens and allotments, displayed and sold in markets and shops, and publicly consumed in restaurants and cafes. More recently, food has entered the city through other mechanisms. Farmers’ markets have emerged in many cities to allow the direct exchange of foodstuffs between producers and consumers, and, in so doing, have developed
new relations between urban residents and rural farmers. However, it has been through new forms of collective growing that food has really entered into urban spaces. Whether labeled as urban agriculture, city farming, or community gardening, there has been an exponential growth in projects that are using the cultivation of plants to transform (largely abandoned) urban spaces and bring about environmental improvement, political empowerment, and sociocultural change. Of course, in many countries of the Global South, agriculture has never left the city. Indeed, it is claimed that agriculture in and around cities in the developing world plays an important role within the conventional food system, delivering significant amounts of locally sourced food to urban populations.

Food is also being increasingly discussed in relation to justice. The agricultural and food processing sectors of the Global North have long been associated with unequal power relations, low wages, and the exploitation of the workforce, with migrant labor and tied accommodation utilized to maintain unfavorable working conditions in rural areas. In the context of the Global South, hunger represents a potent symbol of injustice and absolute poverty in rural areas and has led to a series of high profile interventions by the United Nations and charitable organizations. Food poverty has also been on the increase in countries of the Global North since the economic downturn of the mid-2000s, with a dramatic increase in the number of food banks reported in countries such as the United Kingdom. Issues of justice have entered the global food system in other ways. The ethics of distant responsibility toward farmers in Global South countries has been transformed from an issue of minority interest to a mainstream element of mass consumption through fair trade initiatives. Justice has also taken on a more-than-human dimension with increased academic interest in animal welfare themes, including welfare friendly farming practices and cruelty toward wild animals associated with traditional hunting activities in rural areas.

The growing interest in food geographies has been particularly associated with food consumption and the cultures of food. Food has become part of the cultural fabric of contemporary society and bound up with our personal identities. While recognizing the growth in food poverty, food banks, and low-cost supermarkets in Global North countries, where we shop for our food, which restaurants we frequent, and what foods we eat now reflect our cultural tastes and choices more than our nutritional needs. In some countries, food has exploded on to our television screens through cookery programs and celebratory chefs have become powerful agents in the food system. In some countries, rural areas have witnessed the rise of the gastro-pub, providing high-end cuisine utilizing locally and ethically sourced foods. Food festivals are now a feature of many rural towns, bringing together food producers, retailers, chefs, and consumers to celebrate local and regional produce. In addition, some market towns are actively creating clusters of niche food shops and restaurants to market themselves as places of food, while others have become known for their resistance to the industrial food system through, for example, their promotion of “slow food” and their opposition to proposed developments of supermarkets and fast food restaurants.

Rural politics, governance, and conflict

There has been a long tradition of work in rural geography on rural politics. Early research focused on how the historical control of local government by rural elites – such as landowners, farmers, and owners of industry – led to the development of paternalistic and antiwelfarist political cultures and policy discourse in many
rural areas, elements of which continue to cast a shadow over contemporary rural politics. Engagements with theories of class then followed as rural researchers sought to make sense of processes of social change and conflict. In the United Kingdom, for example, conflicts surrounding demographic change in rural areas previously discussed in terms of disputes between “newcomers” and “locals” began to be reframed in relation to economic class-based conflicts as the interests of in-moving middle-class groups clashed with those of the more established working classes. Others pointed to the formation of new alliances between existing rural elites and in-moving middle-class groups as each attempted to protect their vested interests from development and, particularly, the provision of new housing. Employing more sophisticated accounts of class theory, rural scholars began to give attention to intra-class conflicts between established and newly arrived middle-class groups in rural places as well as domestic property class theory to provide new understandings of conflicts bound up with the production, regulation, and consumption of rural housing.

The shifting system of rural governance has attracted the interest of researchers during the last couple of decades. Some have examined the changing regulatory framework of the global agri-food system, highlighting shifting sets of scalar politics bound up with local, national, and supranational forms of regulation. Reflecting broader developments in political geography, other researchers have considered the changing relationship between the state, market, and civil society in rural areas (see Goodwin 1998). Of particular interest has been the transition from a system of government to one of governance within national and rural policy, involving the enactment of a wider range of nongovernmental organizations into the governance of rural areas. This has led to more complex policymaking processes with new networks and partnerships delivering an increasing array of rural services, from economic development to social welfare, and health to housing. It is claimed that there has been a “hollowing out” of the local state with many of its core functions transferred to other agencies to the extent that it now represents more an enabler than a provider of key services. In a sense, these changes have had a more limited impact in rural areas given that the rural local state has been less interventionist than in its urban counterpart for the reasons mentioned previously. Indeed, the provision of rural services has long involved third-sector agencies, sometimes working in partnership with the local state. Nevertheless, these shifts in governance structures have produced more complex geographies of public service delivery in rural areas as increased reliance has been placed on local coalitions of state and nonstate actors.

In recent years there has also been a switch in research emphasis from rural politics to the politics of “the rural” as rural issues have become increasingly embroiled in national political debates (see Woods 2005). At the core of these debates is the question of who has the right to regulate rural spaces and places. This question of regulation has been bound up with a series of scalar conflicts that have pitched local rural interests against national regulatory systems. In many countries, rural space is regulated in ways that protect particular landscapes and environments in the national interest – as designated national parks, wilderness areas, or spaces of ecological significance. Such designations essentially construct rural spaces as sites of national consumption – as spaces of spectacle, leisure, and pleasure. However, some rural spaces continue to be constructed by national government in productivist terms as resource spaces that need to be exploited in the national interest to provide such things as food, timber,
RURAL GEOGRAPHY

minerals, energy, and drinking water. These different constructions and uses of rural space are clearly prone to conflict, particularly in areas where resource extraction and special landscape designation coexist. What is also apparent within both types of designation is a scalar tension between local and national interests, with “the rural” viewed principally as a space that exists to meet the needs of nation.

Three examples of this scalar tension can be presented. First, national planning policies devised to protect the landscape and environmental attributes of rural spaces tend to restrict local economic and housing development and work against the needs of rural communities. These restrictions reduce employment and housing opportunities for working-class residents in rural places, fueling the out-migration of young people and perpetuating the low-wage economy. In the United Kingdom there is also evidence that campaign groups and middle-class migrants utilize natural discourses of rurality to contest new housing development in rural localities. A second example centers on the introduction of new forms of environmental regulation that conflict with traditional norms of rural living and working. For example, the global recognition of the ecological significance of the ancient temperate rainforests in British Columbia, Canada has led to a series of high profile debates and conflicts over the environmental impacts of industrial forestry involving the government, forest companies, loggers, environmental campaigners, forest communities, and First Nation people. Across the Atlantic, the proposed introduction of legislation to outlaw hunting with dogs in the United Kingdom in the early 2000s resulted in a significant campaign of opposition coordinated by the Countryside Alliance – an organization established to represent hunting interests. While hunting was eventually banned, the Alliance’s campaign constructed the government’s actions as a national imposition of urban values on local rural ways of life. Third, the expansion of renewable energy production by national governments in order to meet climate change targets has resulted in considerable debate about the location of wind and solar farms in rural areas. These farms have not only been constructed by opponents as industrial developments that are detrimental to rural ecosystems and landscapes of rural areas, but as exploiting local rural resources to satisfy national and urban needs.

Rural welfare

Welfare represents a rather neglected area of rural geography. While rural social problems were discussed for much of the twentieth century, it was not until the late 1980s and early 1990s that the first studies of welfare were conducted (see Milbourne 2004; Tickamyer 2006). Based on analyses of official statistics in the United States and bespoke surveys of rural households in the United Kingdom, this work highlighted the significant presence of poverty in rural places. Indeed, research in the United States pointed to higher levels of poverty in rural areas than in metropolitan centers as well as the persistence of poverty in certain rural regions. Notwithstanding this statistical visibility, poverty remains largely hidden within the physical, political, and sociocultural fabric of rural space. Not only are the poor scattered across the rural landscape but some of the taken for granted sites and sights of poverty, such as social housing estates and welfare support centers, are largely absent in rural places. In addition, the social profile of the rural poor cuts against dominant (urban) discourses of poverty in terms of economic activity, household composition, and ethnicity.

Also apparent are some distinctive cultural attributes of poverty in rural areas. It would appear to be the distance from rather than
dependency on the welfare state that forms a striking feature of rural poverty. In part this relates to the less developed nature of state welfare provision in rural places but also to local cultures of “making do” and antipathy toward state-based welfarism. Indeed, the presence of poverty in rural areas is often denied by local residents and by local people living on low income in particular. Research has referred to moral discourses of rurality, whereby material hardship is downplayed, self-sufficiency is promoted, and welfare dependency very much despised, which leads to those on low income internalizing their financial problems and relying on informal systems of support provided by family and friends. A further consequence of these cultural norms is that rural poverty tends to be hidden from the gaze of state agencies, which perpetuates its neglect.

Housing problems and homelessness have also emerged as important welfare themes within rural geography. Although rural housing needs have been widely discussed for several decades, homelessness only emerged as a research topic in the late 1990s. Research on rural housing in the United Kingdom has highlighted processes of unequal housing competition as affluent ex-urban migrants buy into rural housing markets, the role of the planning system in restricting the supply of new housing, and the limited provision of social housing, which collectively have created sets of housing problems for low income households in rural areas. More recently, attempts have been made by researchers in the United Kingdom and United States to relabel these types of housing problems as homelessness. Questioning why homelessness tends to be discussed as an urban problem, a small number of researchers have explored the rural dimensions of homelessness. As with poverty, the hiddenness of homelessness in rural places emerges as a significant theme, with dispersed settlement patterns combining with the absence of support services for homeless groups to remove the presence of homeless people from the dominant gaze of academics and politicians. Homeless people are also constructed as “out of place” in rural spaces; they are forced to camouflage their presence in prime rural spaces and journey to urban places to secure vital support services. The spatialized provision of welfare for the homeless in rural places also leads many to relocate to proximate urban settlements, reinforcing the hiddenness of rural homelessness (see Milbourne and Cloke 2006).

Critical attention has been given to the impacts of national programs of welfare reform in rural places. Again, it has been in the United States and United Kingdom where rural researchers have been most active. In these countries, the principle of universal welfare support has been gradually undermined during recent decades through a transition from a welfare state to a welfare society, whereby state responsibilities for welfare provision have been devolved to a broader range of nonstate organizations, local communities, and individuals and through a shift from a welfare state to a workfare state, involving a move from universal welfare entitlement toward labor market inclusion as the central plank of welfare policy.

Four themes arising from rural research on welfare reform in these two countries are worthy of mention. First, and most significant, workfare appears to be a blunter policy instrument in rural places given that low paid work represents a more significant problem than unemployment. Second, the particularities of rural labor markets and the structural underpinnings of rural poverty have undermined the key policy goal of labor market integration: lower wages, increased distances to employment centers, and the limited availability of training, public transport, and childcare in rural areas present significant barriers to work. Third, although workfare has produced significant reductions in rural welfare
RURAL GEOGRAPHY caseloads, welfare recipients have tended to move to low paid jobs, meaning that levels of rural poverty have remained largely constant. Fourth, the devolution of welfare powers and responsibilities has encouraged the development of new place-based approaches to labor market inclusion that are better able to deal with local obstacles to work. However, welfare devolution has been utilized by local elites to reinforce local antiwelfare discourse and maintain unequal power relations in rural places (see Milbourne 2010; Pickering et al. 2006).

Future rural geographies

Rural geography is now much better aligned with key theoretical and conceptual developments in human geography than it was in previous decades. Rural geographers have made important contributions to human geography in terms of publications and there is a sense that in relation to certain research areas, such as food, rural scholars are leading rather than following research in other areas of the discipline. Looking to the future, it is possible to identify emerging themes as well as existing gaps in rural geographical scholarship that could usefully warrant increased attention. The first concerns the geographical coverage of rural research. Rural geography remains too narrowly focused on a small number of countries in the Global North and the United Kingdom in particular. While this does not mean that important rural research has not been undertaken in the Global South through other disciplines – most notably development studies – it suggests that rural geography lacks the global reach and relevance of other subdisciplines of human geography. Broadening its spatial outlook will no doubt be unsettling for some within rural geography as the spatial core of scholarship will inevitably shift.

For example, poverty and development themes would increase in significance; the dominant narratives of change would be those concerning rural depopulation, declining rural regions and enforced economic migration; and work on agri-food would include more coverage of small-scale and subsistence farming in rural areas as well as the global significance of agriculture in and around cities.

It is likely that boundaries between “the rural” and “the urban” will become increasingly blurred in future years. As processes of urbanization accelerate in Global South countries, so many more rural residents will relocate to urban areas, and cities will emerge as hybrid spaces combining urban and rural cultures and lifestyles. Although such migratory processes are evident in northern countries, they are largely restricted to younger age groups, and it is processes of counter-urbanization that will continue to dominate spatial demographics, with the movement of ex-urban groups to rural places continuing to complicate the meanings attached to rurality. It is also the case that the rise of urban agriculture in northern and southern world contexts is leading to new forms of rural–urban hybridity. What follows from these trends is that some of the divisions between rural and other subdisciplinary areas of human geography will become less distinct. The themes of migration and food, together with the global challenges of climate change and water and energy security, will necessitate multiple and complex geographical approaches; they will also require more productive working relations between human and physical geographers, as well as social, physical, and natural scientists.

Linked to these points is the increasing globalization of rural areas. International migration will continue to alter the sociocultural composition of many rural places, establishing multiculturalism as a key part of the fabric of rural towns and
providing new possibilities for rural living. Rural labor markets will be drawn more into global economic systems and regulatory frameworks, with rising numbers of goods as well as people flowing across national boundaries. The growth in international tourism will open up new rural regions to larger numbers of visitors in search of “authentic” rural places and cultures. Improvements in digital communications technologies will also provide new virtual possibilities, opening up (remote) rural places to wider worlds and creating new options for social networking, home working, economic development, and the provision of key services in rural areas.

The global challenges of climate change, population growth, and the security of food, energy, and water supplies will undoubtedly continue to impact on rural communities, economies, and environments. Rural spaces will be required to provide greater amounts of renewable sources of energy through wind farms, solar projects, and hydroelectric schemes. These energy projects will generate further debate and conflict concerning the ways in which rural areas are constructed as spaces of development, protection and/or consumption. Important here will be the increased flow of natural resources between rural and urban spaces, which will require the development of new forms of infrastructure, such as roads, pylons, and pipes. In terms of food security, it is likely that the neoliberal agri-food regime will continue to dominate, adopting new bio-technological fixes to produce increasing amounts of food in ever more efficient ways. However, there are signs that other approaches to the production, exchange, and consumption of food are now challenging the dominant agri-food model. Community supported agriculture, collective growing in cities, ethical consumption, and the localization of the food network are examples of current alternatives that may well emerge as mainstream aspects of, and research on, the food system in future years.

Moving toward the social research agenda, population aging will represent a more important rural theme in the future. Reflecting the general aging of the population in most developed countries, but also age-related processes of population movements to and from rural areas, older people will comprise a larger proportion of the rural population in future years. While this aging of rural society will no doubt provide social benefits, such as increased levels of volunteering and social support, it is likely to generate some significant challenges in relation to the provision of social welfare and health care in rural places. Welfare as a more general theme will also require increased attention from rural geographers. Inequality, injustice, and poverty remain rather marginalized research issues within rural geography even though many countries of the Global North have witnessed economic downturns, recessions, and rising levels of poverty during recent years and poverty remains a significant feature of rural life in the Global South. The neoliberal response to these problems has been austerity with dramatic reductions in public spending, cuts to welfare programs, and the withdrawal of the state from key areas of welfare provision evident in northern countries. With many rural places already poorly provided for in terms of welfare services, it is likely that austerity will hit vulnerable groups and welfare support agencies particularly hard in rural communities. Researchers need to engage more with these themes to ensure that the poor and the vulnerable in rural areas are positioned more prominently within rural geography.

SEE ALSO: Environmental governance; Environmental planning; Environmental policy; Globalization and rural areas
References


Further reading


The changing governance of rural places has long been of interest to geographers. To begin with, agriculture is in most cases a dominant land use, and of all sectors in industrialized economies it has one of the longest and most controversial histories of state intervention and influence (Potter 1998; Sheingate 2001). However, attempts by governments in jurisdictions like the United States and the European Union to reform the (very substantial) subsidies offered to farmers in response to mounting budgetary pressures and a shifting global market context have proved controversial and been subject to challenge. Geographers and rural social scientists have helped explain why this is and have documented the powerful impact that farm subsidies have had (and continue to have) both on the way agriculture is practiced and on how rural land is used and managed (see, e.g., Winter 1996). They have also contributed to politically charged debates about the need for agricultural policy reform, specifically for “greening” farm support in order to provide better incentives for environmental management on private lands. Forestry policy in most industrialized countries constitutes a separate domain but, traditionally, it has often shared with agriculture a heavy state-bureaucratic emphasis. In European Union countries such as Finland and Sweden, for instance, strong state involvement in forest establishment and woodland management dates back to the early twentieth century. Here, too, geographers have been at the forefront of analysis and critique.

But rural policy covers more than the governance of agriculture and forestry as productive sectors and as the basis of rural livelihoods. A second major theme of rural policy and politics has been how best to respond to economic pressures and accommodate social change while ensuring that the public goods of rural landscapes, habitats, and wildlife are safeguarded in the postindustrial countryside. Rural areas have been changing socially as well as physically in the postwar period, with profound shifts in the geographical distribution of economic activity and populations. This has increased development pressures while at the same time creating new demands from more pluralistic rural communities for better control of farm pollution and improved countryside access and recreation, among other things. Geographers have documented these trends and reflected on what they might mean for a countryside that is as much a lived-in space as it is a site of production or consumption. What might be called “rural environmental policy” has a complex history of institutional adaptation and change, with early interventions to designate or reserve land in order to protect natural assets giving way more recently to more politically ambitious attempts to integrate rural areas into broader debates about resilience and sustainability. In an era of climate change, rural areas are now poised to be key sites of adaptation and mitigation, but it remains unclear how far this will require continued governance by the state. New, arguably more neoliberal, models of conservation are emerging, and geographers are reflecting on what the emergence of “ecosystem
services” thinking in particular could mean for rural policy as traditionally conceived.

Agricultural policy and the rural politics of decoupling

Rural geographers have made major contributions to our understanding of agricultural restructuring and change (see, e.g., Ilbery and Bowler 1998). They have also tried to explain the apparent ability of the farm lobby and wider agribusiness networks to resist reform at a time when deregulation means that the state has been withdrawing from most other sectors of the economy (Potter and Lobley 2004). Dating from the Depression years of the 1930s, and continuing more or less intact ever since, a state assistance paradigm has asserted a large role for governments in subsidizing agricultural production and underwriting farming incomes. The earliest farm policies, such as those introduced in the United States as part of Roosevelt’s New Deal, involved offering marginalized farmers guaranteed prices for their output. After World War II, government subsidies were needed to bring about agricultural recovery here and throughout Europe. This food security emphasis soon gave way during the 1960s and 1970s to modernization and intensification, with governments intervening to offer incentives for farmers to expand output in order to achieve, first, self-sufficiency and then to exploit the scope for exporting their output to overseas markets.

These strictly productivist motives have often been coupled with a concern to maintain family farming as a cultural asset and are one of the defining features of rural space. Family farms are seen as cornerstones of rural society and guardians of the rural environment and thus deserving of state support. Policy scientists have long been interested in the role of ideas, social framings, and mental paradigms within the policy process. There has been increasing recognition of the way certain historical constructions of understanding can frame and stabilize a policy field, often for long periods, and also mounting scholarly interest in the scope for new ideas to emerge as recruiting devices for policy change and even wholesale reform (Stogstad 1998). In relation to agricultural policy, geographers have pointed to the way in which state support has for a long time been justified publicly by certain taken-for-granted notions about the nature and constitution of the agricultural industry and the privileged status of farmers in supplying public goods such as food security and environmental quality. These understandings have proved to be powerful defensive weapons in the subsequent battle to reform farm support in an increasingly globalized context. In the European Union, for instance, policymakers struggled throughout the 1980s and 1990s to reform an increasingly expensive and trade distorting Common Agricultural Policy (CAP). Despite a well-rehearsed economic efficiency case for reform, they have found it difficult to agree any significant reduction in the overall scale of public subsidy because of resistance from a farm lobby that has a deeply vested interest in maintaining the status quo.

The nature of this resistance and the political battles that have ensued, not just in Europe but also in the United States and various other developed countries, have been studied by scholars of many disciplines, among them geographers. In Wilson’s (1977) classic analysis, agriculture is a defining example of “client or corporatist politics,” in which a powerful and well-organized set of interest groups has colonized government departments and agencies in order to defend their (often very generous) state entitlements. Their success, as has been noted, is reflected in the extraordinary persistence of the state
welfare model to the present day. That being said, the emergence of a global agri-food system in which stateless transnational corporations (TNCs) have growing control over the production, processing, and marketing of food through increasingly integrated supply chains has brought new players onto the scene. These actors take a very different view of state support and have, to some extent, driven a wedge between the large-scale agribusinesses that are able to compete on world markets and the more traditional family farming concerns that have been the main focus of policymakers’ concern. Concerned to see the opening up of markets rather than their continuing state-led regulation, agribusiness interests have campaigned for the liberalization of agricultural policy, organizations representing many of these TNCs such as the Confederation of Food and Drink Industries of the EU and the Food and Drink Federation lobbying within the World Trade Organization (WTO) to promote freer world trade. A campaign to decouple the subsidies farmers receive from their production decisions in order to make sure that they do not give domestic producers an unfair trade advantage over other competitors has continued to gain support internationally. With traction from trading nations such as Australia and Brazil, anxious to gain improved access for their own producers to European and US domestic markets, this agenda has continued to gain ground within the WTO. Policymakers in the European Union have finally abandoned the trade-distorting price supports that for decades made up the lion’s share of farm support, putting a system of still expensive but less trade-distorting direct payments in their place. Nevertheless, little progress has been made in reducing the overall scale of support, and agricultural policy remains an object of exceptionalist fascination for many rural geographers.

Conservation and the neoliberal turn

The protection of public goods such as biodiversity, landscape, and open access that are closely associated with rural land has been a major rural policy concern throughout the twentieth century, beginning with the first national park and other protected area designations in the United States and in most EU member states. As Adams, Hodge, and Sandbrook (2014) observe, the dominant motive has been “territorialization” – the designation and setting aside of land in order to protect it from development and change. For a long period the focus of conservation policy in countries such as the United Kingdom has been on how best to manage national parks and various other types of designated or reserved land. Not until the 1980s did conservation policy begin to address how to protect nature outside these protected areas, largely driven by concerns about the impact of agricultural intensification on farmed landscapes in places such as Western Europe. Geographers were at the forefront of debates at this time about the need both to restrain intensification and to better integrate agricultural and environmental policy domains (see, e.g., Cox, Lowe, and Winter 1986).

Agri-environmental policy (AEP) can be traced back to the early 1980s, a time when the state project for agricultural modernization in industrialized countries was coming under attack from critics of the heavy domestic budgetary costs of farm support. Pressures toward budgetary retrenchment and the market liberalization described earlier coincided with the emergence of an environmentalist and social critique of modern agriculture (Potter 1998). While explicitly critical of the state’s role in subsidizing agriculture, sustainability advocates bolstered support for continued state engagement by arguing for a redirection of, rather than a net reduction in, the farm spend, albeit in favor of rural
development and agri-environmental policy objectives. In rural political terms, this was attractive to farm interests anxious to maintain overall levels of state support, and it made possible a new and, in retrospect, durable consensus between farming, conservation, and state bureaucracies.

Subsequent research and commentary by geographers and others has demonstrated how the resulting agri-environmental policy programs profoundly reflect the historical-institutional context in which they were made (Sheingate 2001; Potter 1998). A major source of funding for environmental management on private lands, these programs stand as classic examples of the second-best, politically compromised, arrangements that policymakers and those who lobby them invent in order to address a range of policy problems — in this case rural environmental degradation, the support of farmers’ incomes, and the overproduction of agricultural commodities. Indeed, from the beginning, agri-environmental payments to farmers have served purposes other than conservation, and their wide scope and the tradeoffs they embody are important explanations of their durability. For example, the Conservation Reserve Program (CRP), the most expensive federal AEP program in the United States, enrolling over 10% of all farmed land, serves both environmental conservation and supply control goals while contributing significantly to farm incomes. As a policy field, AEP is state-centered and characterized by formal bureaucratic procedures, but it has been maintained politically by a diverse set of nonstate actors who have contributed to the renegotiation of policies and administrative practices in response to changing external and internal policy conditions. Closely interconnected AEP networks have developed in these countries. Actors such as the National Farmers’ Union in the United Kingdom and the American Farm Bureau Federation in the United States have worked with commodity groups and environmental NGOs and rural advocacy organizations to shape the creation and evolution of agri-environmental policy. Subsequent adjustments and policy redesigns have been the result of negotiations by this relatively stable and closed set of elites.

While the political foundations of AEP over the past 25 years have exhibited tremendous stability, there have been significant changes in specific policy aims and administrative arrangements. Over time, new ecological risks and new frames for articulating the public interest in agriculture and rural landscapes have entered the picture. The tendency has been to move up in scale from field-level conservation objectives (e.g., soil erosion control) to incorporate whole farm (e.g., water quality) and landscape-level management (e.g., biodiversity). These reforms have added new layers of justifications and administrative controls onto the existing state-centered framework without destabilizing the basic institutional arrangements and political constellations. The recent increased advocacy of offering farmers and other land managers payments for ecosystem services (PES), a neoliberal policy idea which emphasizes the scope for nonstate actors to get involved in the contracting of environmental goods from farmers and other land managers, however, has been seen by some commentators as posing a challenge to the settled state-bureaucratic way of doing things (Potter and Wolf 2014). PES is part of a wider shift in discourses about nature that has resulted from the development of “ecosystem service” thinking, a set of ideas which emphasize the flow of benefits that nature provides and which benefit human wellbeing. As Fish, Winter, and Lobley (2014) have argued, the currency of this idea has grown steadily in the interdisciplinary resource management literature and its influence is now being felt in the rural sphere.
Geographers like Higgins, Dibden, and Cocklin (2012) in Australia and Wynne-Jones (2012) in the United Kingdom point to new lines of rural geographical inquiry that are needed to better understand what the resulting model of hybrid governance will mean for policy actors, local communities, and the use of land.

Climate change and the new governance of rural areas

The emergence of climate change as a new public policy justification for environmental management on farms, and on rural land more generally, nevertheless adds significantly to the external and internal pressures for change. This is especially true if the requirement to bring agriculture and rural areas within the carbon economy means further extending the environmental remit of farming and farmers. While there are currently no specific international targets for the agriculture or forestry sectors, the Intergovernmental Panel on Climate Change (IPCC) has identified the maintenance of extensive grassland systems, improved livestock manure management, and more controlled fertilizer applications as key climate mitigation measures. This and other assessments are leading to a further reappraisal of agriculture’s environmental footprint and to the emergence of a new research and policy agenda around climate adaptation, carbon sequestration, bioenergy production, and the need to achieve greenhouse gas (GHG) emission reductions. At the same time, growing concerns about food security (an agenda that is partly driven by worries about the impact of climate change on future agricultural productivity) means that this must all be achieved without compromising productivity or output.

The resulting “new land use debate” (Winter and Lobley 2009), centered around the idea of “sustainable intensification,” is one that geographers and others are still struggling to understand. For instance, in the European Union, climate change imperatives are behind moves to establish a new soils directive and to strengthen biodiversity protection, while larger strategic questions concerning how best to steer the future management of land remain firmly within the purview of the CAP and the huge financial resources at its disposal. In the United States, the most recent Farm Bill, the 2008 Food, Conservation and Energy Act, empowers the US Department of Agriculture (USDA) to facilitate farmer, rancher, and forest landowner participation in carbon mitigation activities. But, rather than offering direct subsidies for the deintensification of cropping systems to mitigate GHG emissions, for example to reduce reliance on inorganic nitrogen fertilizer inputs, the dominant response in the United States has been public and private investment in corn ethanol production, a development that enjoyed an initial wave of broad support but was then found to have unintended consequences and highly questionable carbon mitigation benefits (Wolf 2014). The carbon and food security debates, taken together, strengthen the case for more integrated public policymaking, posing larger policy questions such as: What is the best balance between extensive and intensive modes of land use? How can investments in carbon mitigation contribute to water quality and biodiversity conservation goals? How far should climate change policies mesh with efforts to address food and energy security? Geographers will remain deeply involved, both in the research that is required to address these global challenges and in the policy debates and public discourses that follow.

SEE ALSO: Climate adaptation/mitigation; Climate change policy; Nature conservation; Neoliberalism and the environment; Public
RURAL POLICY AND POLITICS

policy; Rural geography; Rural society in the Global North

References


Rural society in the Global North

Lisa Mathis Butler Harrington
Kansas State University, USA

Although the world’s rural population is now less than half the total, and proportionately even smaller in the more developed countries of the “Global North,” rural residents work to provide the food and natural resources upon which the world’s population relies, and have held an important place in the historical development of today’s most influential countries. The vast majority of ecosystem services supporting human populations and global society are in rural areas, including the many natural resources utilized directly and indirectly in local to global economies. The characteristics and situations of people living in rural areas, including villages and small towns located in rural regions, are highly variable around the globe. There is great variability even in the part of the world known as the “Global North,” particularly those countries that are economically well-off and that frequently have their historical development closely linked to Western Europe (e.g., European countries, the United States, Canada, Australia, New Zealand), but also including other countries that are considered more developed, such as Japan. There is no universally accepted listing of countries of the Global North, but these countries overlap with those identified as “developed,” “First World,” and “advanced economies.” The most clearly connected group of countries belonging to the Global North would be the Organisation for Economic Co-operation and Development (OECD) member states.

The more developed Global North countries tend to differentiate urban and rural environments more than is done in the Global South, although specific definitions of what is rural vary among countries and among differing governmental agencies and applications within individual countries. Policy and statistics-oriented definitions generally focus on rural as “not urban” (with definition of urban coming first) or are based on low population densities and settlements below a certain threshold of population size. Even when consistent definitions are applied there are variations in proportions of populations living in rural areas and in rural population densities. General considerations regarding rural society include key economic activities and configurations; connections to the landscape; characteristics and change in rural populations, particularly regarding age, education, and migration; and availability of social services and connections to urban areas.

Resources, landscapes, and economics

The characteristics of rural societies are linked to the resources, physical situations (especially accessibility), and economic activities of their locations. Rural people often are considered to be more closely connected to and cognizant of the environment and natural processes than people living in the (generally dominant) urban society. Rural areas are more economically diverse than is frequently perceived, however; the most commonly associated activity with rural areas is farming or agriculture, although
there are several other focal activities depending on resources available in rural regions. The farming identification may be based upon the most widely practiced rural activity – and the basis for family and national population support – in Europe and other areas for centuries and in European-colonized countries (e.g., the United States and Australia) at least since Europeans first arrived and often dating to pre-Columbian social groups. Many rural regions are resource-dependent; that is, most of the income for the area is derived from a specific set of natural resources, such as forests, mineral resources, fish, scenery (for recreation and tourism), or agricultural resources (land, soil, and climate). The interactions of rural people with their environments lead to varying landscapes, from open rangeland to dense forests, from seemingly continuous flat wheat fields to small communities in hilly patchworks of farm fields and woods, and from strip mines to coastal fishing villages. The creation of interdependence of rural society and local landscapes, including variations in physical accessibility, has led to the development of some social variability among rural residents connected to their differing livelihoods even when they are clearly also a part of a larger national society.

In cases of dependency on a single resource and a dominance of one type of livelihood/economic product, rural regions often undergo boom-and-bust periods. When a given resource is in demand and prices are high the local economy appears to be quite healthy, but when prices fall, resources are depleted, or conditions limit resource availability, the local area may face a great deal of stress and even population loss. Boom-and-bust economic oscillations have historically been clear with respect to mining districts. However, other types of resource-dependent areas also can display sudden declines. In the cases of some renewable resources, overuse can cause collapse and a loss of resiliency. For example, the Canadian North Atlantic cod fishery collapsed in about 1991 because the cod population could not sustain heavy fishing pressures, fueled partly by changes in fishing technology. Following the collapse, fishing-dependent communities in the Canadian Maritime provinces were depopulated as residents were forced to move elsewhere for employment. In forested areas, employment also can be cyclical, depending on the level of timber harvest and amount of growth since the last harvest.

A useful categorization of core economic activities in rural regions has been developed by the US Department of Agriculture’s Economic Research Service (USDA ERS). The ERS identifies rural counties as farming-dependent, mining-dependent, manufacturing-dependent, federal/state government-dependent, or services-dependent based on the proportion of earnings coming from the specified source; nonspecialized counties are those that do not meet the threshold for dominance by any one of the specified types (USDA ERS 2012). The ERS typology is incomplete, however, as it does not include some other important activities in rural areas, including fisheries, forestry/timber production, and tourism. There also should be recognition of seasonality in many activities and employment as well as supplementary livelihood activities like hunting (which may be either a supplementary household food provisioning activity or primarily a recreational activity, depending in part on income and social position). In some regions, government employment is significant particularly where much of the land base is in public ownership. Transfers of payments – that is, government disbursements to particular classes of citizens through retirement or disability programs and government support for social services (health and education), for
example — can be of high importance in retirement destination areas and in low-opportunity regions with significant indigenous or minority populations.

Even in areas that have traditionally been dominated by farming activities, there have been recent shifts to what often are known as pluriactivity and to multifunctional agriculture. In the case of pluriactivity, farmers have generally taken up additional income-earning activities, such as vacation or holiday rentals, farm-based retail sales, and tourism-based events. This is most common on small farms where the land base no longer can produce at the scale expected to support a family under modern conditions, particularly in parts of Europe and in regions of countries like the United States and Canada where settlement happened early on relatively small tracts of land. In addition to farm-based pluriactivity, one or more members of the farm household often will have off-farm employment in another industry such as services or manufacturing. In the case of multifunctional agriculture, which may be connected to but is not the same as pluriactivity, there is explicit recognition that agricultural land and activities provide goods and services beyond those traditionally considered. That is, in addition to food and fiber (e.g., wool and cotton) production, farmland may provide social benefits and ecosystem services including aesthetically attractive scenery, biodiversity enhancement, and control of the movement of water and other materials. Government policies often explicitly recognize these broader benefits of lands used for agriculture, forest, or other cover types to local and broader scales of society.

Another shift in approaches to the use of natural resources has been the partial move toward “postproductive” agriculture and forests by some landowners, rooted in changes in social attitudes toward the environment and health. By this, authors have generally meant a shift from the highly industrialized/high input versions of production to agriculture, forests, and land management focused on alternative production (e.g., organic farming) and greater recognition of alternative values or benefits from land resources, including recreation and other ecosystem services such as wildlife habitat provision and aesthetic enjoyment. These alternative approaches, often combined with multifunctional land management, contrast and are in tension with continued “productivist” (industrialized and profit-maximizing) approaches which continue to dominate many areas of the Global North (see, e.g., Wilson 2001; Evans, Morris, and Winter 2002; Mather, Hill, and Nijnik 2006).

Social considerations

In addition to numbers and densities of rural residents, natural resources and related livelihoods, and economic characteristics, aspects of rural society include other demographic characteristics, cultural practices, religious and political tendencies, and sense of place. Oftentimes rural society is associated in public perceptions and popular culture (e.g., film) with traditionalism and higher levels of religiosity, self-reliance, and low educational attainment, although any such generalizations must be recognized as oversimplifications and often not reflective of rural society in a given locale (and certainly not reflective of all individuals living in rural areas). In contrast to some of the negative views of rural inhabitants, in many countries the perceived independence and simpler ways of life closer to the Earth also are highly valued and even romanticized, creating a dichotomy in views of rural society.

Rural society has seen significant relative and total population change since about 1900. For
some rural areas in the United States, populations peaked before the Great Depression and have been in near-continuous decline since the 1930s. Continued population decline in rural areas, both in terms of absolute numbers and in terms of proportions of total national populations, has been widespread (Table 1), but countries varied from less than 5% to 50% rural in 2010. It is important to note, however, that definitions of rurality for statistical descriptions vary by country. For most countries, a majority of national territory is considered rural.

Table 1 Proportion of population living in rural areas in selected countries of the Global North.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>16.5</td>
<td>14.7</td>
<td>14.1</td>
<td>14.2</td>
<td>14.5</td>
<td>14.6</td>
<td>13.9</td>
<td>12.8</td>
<td>11.8</td>
<td>11.0</td>
</tr>
<tr>
<td>Austria</td>
<td>35.0</td>
<td>34.7</td>
<td>34.7</td>
<td>34.6</td>
<td>34.4</td>
<td>34.2</td>
<td>34.2</td>
<td>34.2</td>
<td>33.5</td>
<td>32.5</td>
</tr>
<tr>
<td>Belgium</td>
<td>6.9</td>
<td>6.2</td>
<td>5.5</td>
<td>4.6</td>
<td>4.1</td>
<td>3.6</td>
<td>3.2</td>
<td>2.9</td>
<td>2.7</td>
<td>2.5</td>
</tr>
<tr>
<td>Canada</td>
<td>27.1</td>
<td>24.3</td>
<td>24.4</td>
<td>24.3</td>
<td>23.6</td>
<td>23.4</td>
<td>22.3</td>
<td>20.5</td>
<td>19.9</td>
<td>19.4</td>
</tr>
<tr>
<td>France</td>
<td>32.9</td>
<td>28.9</td>
<td>27.1</td>
<td>26.7</td>
<td>26.4</td>
<td>25.9</td>
<td>25.1</td>
<td>23.1</td>
<td>18.4</td>
<td>14.8</td>
</tr>
<tr>
<td>Germany</td>
<td>28.0</td>
<td>27.7</td>
<td>27.4</td>
<td>27.2</td>
<td>27.3</td>
<td>26.9</td>
<td>26.7</td>
<td>26.9</td>
<td>26.6</td>
<td>26.2</td>
</tr>
<tr>
<td>Greece</td>
<td>52.5</td>
<td>47.5</td>
<td>44.7</td>
<td>42.3</td>
<td>41.6</td>
<td>41.2</td>
<td>40.7</td>
<td>40.3</td>
<td>39.7</td>
<td>38.8</td>
</tr>
<tr>
<td>Hungary</td>
<td>42.0</td>
<td>39.9</td>
<td>37.8</td>
<td>35.8</td>
<td>34.9</td>
<td>34.2</td>
<td>34.8</td>
<td>35.4</td>
<td>33.6</td>
<td>31.0</td>
</tr>
<tr>
<td>Ireland</td>
<td>51.3</td>
<td>48.3</td>
<td>46.4</td>
<td>44.7</td>
<td>43.7</td>
<td>43.7</td>
<td>43.1</td>
<td>42.1</td>
<td>40.9</td>
<td>39.5</td>
</tr>
<tr>
<td>Italy</td>
<td>38.2</td>
<td>35.7</td>
<td>34.4</td>
<td>33.4</td>
<td>33.2</td>
<td>33.3</td>
<td>33.1</td>
<td>32.8</td>
<td>32.4</td>
<td>31.8</td>
</tr>
<tr>
<td>Japan</td>
<td>32.1</td>
<td>28.1</td>
<td>24.3</td>
<td>23.8</td>
<td>23.3</td>
<td>22.7</td>
<td>22.4</td>
<td>21.4</td>
<td>14.0</td>
<td>9.5</td>
</tr>
<tr>
<td>South Korea</td>
<td>67.6</td>
<td>59.3</td>
<td>52.0</td>
<td>43.3</td>
<td>35.1</td>
<td>26.2</td>
<td>21.8</td>
<td>20.4</td>
<td>18.7</td>
<td>17.1</td>
</tr>
<tr>
<td>Lithuania</td>
<td>55.6</td>
<td>50.4</td>
<td>44.3</td>
<td>38.8</td>
<td>35.0</td>
<td>32.4</td>
<td>32.7</td>
<td>33.0</td>
<td>33.4</td>
<td>33.0</td>
</tr>
<tr>
<td>Netherlands</td>
<td>39.2</td>
<td>38.3</td>
<td>36.8</td>
<td>35.3</td>
<td>33.3</td>
<td>31.3</td>
<td>27.2</td>
<td>23.2</td>
<td>19.8</td>
<td>17.3</td>
</tr>
<tr>
<td>New Zealand</td>
<td>21.1</td>
<td>18.9</td>
<td>17.2</td>
<td>16.6</td>
<td>16.3</td>
<td>15.3</td>
<td>14.7</td>
<td>14.3</td>
<td>13.9</td>
<td>13.8</td>
</tr>
<tr>
<td>Norway</td>
<td>42.4</td>
<td>34.6</td>
<td>31.8</td>
<td>29.5</td>
<td>28.7</td>
<td>28.0</td>
<td>26.2</td>
<td>23.9</td>
<td>22.5</td>
<td>20.9</td>
</tr>
<tr>
<td>Slovenia</td>
<td>67.5</td>
<td>63.0</td>
<td>57.6</td>
<td>52.0</td>
<td>50.4</td>
<td>49.6</td>
<td>49.4</td>
<td>49.2</td>
<td>49.5</td>
<td>50.0</td>
</tr>
<tr>
<td>Sweden</td>
<td>22.9</td>
<td>19.0</td>
<td>17.3</td>
<td>16.9</td>
<td>16.9</td>
<td>16.9</td>
<td>16.2</td>
<td>16.0</td>
<td>15.7</td>
<td>14.9</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>22.2</td>
<td>22.9</td>
<td>22.3</td>
<td>21.5</td>
<td>21.6</td>
<td>21.9</td>
<td>21.6</td>
<td>21.3</td>
<td>21.0</td>
<td>20.5</td>
</tr>
<tr>
<td>United States</td>
<td>28.1</td>
<td>26.4</td>
<td>26.3</td>
<td>26.3</td>
<td>25.5</td>
<td>24.7</td>
<td>22.8</td>
<td>20.9</td>
<td>19.3</td>
<td>17.9</td>
</tr>
</tbody>
</table>

Source: Data obtained from World Bank (2014). Definitions of rural areas vary by country.

Much of the rural population decline has been a result of changing technology, and economic trends have contributed to outmigration from rural to urban areas. Mechanization and technological changes led to a need for fewer workers on farms and ranches or stations and in other rural industries during the twentieth century. These changes have merged with broader socioeconomic trends toward urbanization and creation of urban-focused economic growth. Landholdings have often increased in size with shifts in technology and needs to produce more to provide sufficient income while rural products...
fail to increase in value at the same rate as many other goods and services. Corporate ownership of land and facilities connected to food and timber production has increased, although in some cases this is in the form of family corporations. In other cases, very large multinational businesses have come to control large parts of the rural economy, and rural regions have come to be closely tied into the globalized economy.

The characteristics of rural residents also have changed over time. In many areas, loss of population has been particularly connected to departure of the young to gain education and take up employment in more urban locales. This has been associated with an aging of the local population, and a form of “brain drain” and loss of social capital. A sort of vicious cycle of population loss and loss of services has become common, with closure of schools, grocery stores, and other services leading to further stress. Reductions in post offices and postal services have been a growing concern in Australia, the United States, and other countries, and a lack of sufficient medical facilities, especially given aging rural populations, is another widespread concern. As a means to battle population loss, some rural communities and small towns in rural regions of the United States and Canada have created land giveaways (free to a very nominal cost) to try to attract new residents.

Population loss is not the only picture for rural places, however. After decades of rural population loss, the 1970s brought a “rural turnaround” or period of counterurbanization to parts of North America, Australia, and Western Europe. The period of movement from cities back to rural areas was particularly dramatic for a period of 10–15 years. Shifts back to population loss in some areas followed the mini-boom. In other areas, retirement in-migrants and second homeowners are playing a role in maintaining or increasing rural populations. With greater opportunities for telecommuting, high amenity value regions also attract new residents who are able to rely on technology for their work connections, as long as sufficiently high speed Internet connections are available. The picture is mixed when it comes to movement of retirees and higher income new residents to rural locations. Oftentimes they bring differing attitudes and expectations than long-time residents. New residents often expect a level of local services that local governments cannot economically provide, and may not have an understanding or appreciation of traditional ways of life – the odors and sounds associated with farming, for example, may be objectionable to new residents, even though the landscape created by farming may have attracted them to the area in the first place. Differences may lead to conflict in local decision-making, and may cause resentment if new residents gain too much power too quickly.

Adding another dimension to changes in local populations and culture, with potential conflict, is the in-migration of ethnic minorities into rural communities. For example, in parts of the United States, Hispanic immigrants have come to be a significant part of local populations since the 1990s. This type of change can cause social strains where there is prejudice, a lack of understanding, or simply a stress on services like a need for multilingual education. In some communities, however, an influx of immigrants, even if ethnically dissimilar to the original inhabitants, is seen as an opportunity for survival of places in danger of dwindling to near disappearance. Rural locales in Europe also have seen the stress of shifting ethnicities or nationalities as movement follows economic shifts, with, for example, Poles and Brazilians moving into and out of Ireland depending on relative economic and employment conditions.

Pockets, and even entire regions, of rural poverty also exist in parts of the Global North as well as in the developing world. In many countries there is a greater incidence of poverty in rural regions than in urban areas although there are variations in poverty rates among countries and among national subregions (Figure 1). The nonmetropolitan US South, for example, shows a much higher rate of rural poverty than other regions, and far higher poverty levels than in southern cities.

Several conditions affect the economic status of rural society, beyond resource demand and availability as mentioned above. The European Commission (EC) (2008) noted that several common characteristics of rural areas (particularly those that are less well-off) contribute to poverty, including their demography, remoteness, education levels, and labor market. The EC also noted that negative situations regarding these characteristics may result in “vicious circles” that reinforce conditions of poverty. For many rural areas, having few young people and age distributions tilted toward the elderly contributes to low incomes and to further out-migration of younger residents. Rural areas may be physically remote from more societally central locations, particularly in mountainous and semi-arid regions, and may also have the difficulties of poor infrastructure regarding transportation networks as well as health care and communications technology. Low educational levels are common in rural areas. This in combination with lack of accessibility can lead to few employment options, contributing to local low incomes/poverty and more out-migration for jobs and education. Demographic, educational, employment, income, services, and accessibility conditions contribute to common – but not universal – differences between rural and urban areas, sometimes known as the rural–urban (or urban–rural) divide. Rural areas more accessible to cities generally are in much better situations regarding opportunities and economic status.

Policy and programs

Policies and government organizations to address rural conditions and policy administration in the Global North are generally separated from those oriented toward urban conditions. Some policies are oriented specifically around economic activities, such as agricultural or timber production, or environmental and natural resource considerations, such as prevention of excessive soil erosion. Additional policies are more directly aimed at addressing social concerns. For consideration of US federal policy, particularly expenditures, the USDA ERS (2012) has several overlapping categories for rural counties. Most are oriented toward rural social problems: housing stress, low education, low employment, persistent poverty and persistent child poverty, population loss, nonmetropolitan recreation, and retirement destination. The latter two are
less problematic although they can be associated with particular social concerns as well (including income and age-related needs). The types of issues identified for the United States are similar to those identified for Europe, and several are connected to poverty conditions. The US federal government has paid particular attention to conditions of rural society in Appalachia for decades. Administration-initiated programs include loans and services directed toward rural small businesses, support for rural health care, development of infrastructure to provide broadband coverage in rural areas, and support of rural public education through technology deployment. Other policies and programs have been established through legislation, often as parts of appropriations acts. Legislation also requires payments to local governments from revenues to federally owned property or “payments in lieu of taxes” to offset the loss of local tax dollars due to reduced private property. Such income may help to support schools and other social services. In addition to national-level policy and programs, individual states develop their own approaches to rural concerns.

European policy addressing rural regions includes the European Union’s major Common Agricultural Policy (CAP) focused particularly on farming and countryside protection, but other policies exist to address other rural concerns. Rural development is a focal area of policy for the European Union (EU) as part of the CAP. Member countries create their own national or subnational rural development programs and submit these to the European Commission (EC) for approval. Expenditures go toward such things as measures for rural quality of life, economic diversification, environmental and countryside improvements, and improvements to agricultural and forestry competitiveness. Some EU expenditures go toward sustainable development of villages and improved information and communications technology as well as promotion of adaptation of buildings to new uses and of new economic activities. Local areas often pursue development through tourism, “green” energy, and other alternatives for new income.

Rural environments and people will continue to be important to provisioning larger populations and supporting ecosystem services connected to planetary operations, even while the rural percentage of the Earth’s and Global North’s human population is likely to continue shrinking. Rural society is important to supporting transitions toward greater sustainability via its traditional provisioning activities combined with environmental management and stewardship.

SEE ALSO: Agricultural environments; Globalization and rural areas; Landscape; Migration: internal; Natural resources; Resources and development; Rural geography; Rural/urban divide

References


Further reading


Rural/urban divide

Katherine V. Gough
Loughborough University, UK

Although “rural” and “urban” are widely used terms, they are difficult to define. Their distinguishing features are becoming increasingly blurred, and there are multiple links between them across all scales. Consequently terms such as “rural–urban linkages,” “rural–urban interactions,” and “rural–urban dynamics” have become commonplace. These discussions most frequently take place in, but are not restricted to, a Global South context.

National definitions of “urban” abound and are composed of a range of factors used either independently or in combination, such as population size, population density, occupation, land use, infrastructure, and officially designated boundaries. Even in instances where the same factor is used, such as population size, the numerical value can vary greatly, making any cross-national comparisons highly problematic. “Rural” is often defined as the residual, that is, areas located outside of cities and towns. The term “peri-urban” refers to the transition zone between urban and rural areas typically located on the outskirts of cities.

Although urban and rural are defined as being different places, in practice the distinction between the two is far from clear-cut. There are many similarities in terms of the nature of the activities that take place in what are seen as rural and urban places, as well as numerous linkages that bind the two together. Rural areas are typically characterized as being primarily agriculturally based. While many people resident in rural areas still engage in agricultural production for at least part of the year, and often identify as farmers, farming is not necessarily their only or their major activity. Many rural residents engage in nonfarm activities and/or work elsewhere for part of the year. Similarly, urban residents may undertake agricultural activities either by moving temporarily to rural areas for work at times of high labor demand, or by engaging in agricultural production within the urban area where they are resident. Urban agriculture is widespread in many cities in the Global South as urban residents cultivate unoccupied plots of land, growing food for home consumption and sale. Households in some sub-Saharan African cities have been found to derive between 25% and 40% of their annual income from urban agriculture, with urban and peri-urban areas contributing 30% to 60% of food supplies such as vegetables, thus making an important contribution to food security (see Lwasa et al. 2014).

Linkages between rural and urban areas across space include flows of people, finance, commodities, and information. There is a constant movement of people between rural and urban areas in both directions and over differing scales and time periods. With increasing competition over scarce resources in rural areas, individuals and households move to urban areas in search of better opportunities as well as to gain access to services such as education and health. Many urban dwellers maintain strong links with their rural roots, returning to their home villages for short visits or, especially where conditions in urban areas have deteriorated, for longer periods. Numerous households straddle the rural/urban divide, creating mul-
RURAL/URBAN DIVIDE

tispatial households where members reside in rural and urban areas in an attempt to maximize incomes and minimize risk. Movement between rural and urban areas also takes place for social and cultural reasons, such as marriage and in order to satisfy the desire to experience other places and demonstrate the ability to survive in differing contexts.

Remittances sent by migrants are a key aspect of financial flows between rural and urban areas. Whereas these flows have typically been from urban to rural areas, with increasing economic hardship in the former the reverse is also occurring. Goods, especially agricultural produce, are another form of remittance sent between household members living in rural and urban areas. Most commodity flows, however, are for commercial reasons, regardless of the direction. While radio, television, and newspapers remain important for the spreading of knowledge, flows of information have been revolutionized by the rapid increase in mobile phone usage. In multi-spatial households, access to a mobile phone has transformed flows of information between household members, enabling them to have much more frequent communication. As well as facilitating contact for social and emotional reasons, mobile phones are also widely used for business purposes, for example to gain knowledge about prices and markets, and for financial transactions. These multiple types of flows, which illustrate how rural and urban areas are inextricably interlinked, are central to achieving balanced economic growth and reducing the vulnerability of the poor.

There has been renewed interest since around the mid-1990s in the role of rural–urban linkages in economic development and in how these links are affected by processes operating at global, national, and local scales (Lynch 2005). Starting at the global level, the liberalization of trade has resulted in a greater availability of imported goods, affecting consumption and production patterns in rural and urban areas, with local producers often being outcompeted. Trade in export crops tends to be controlled by international companies, which often bypass local urban centers for processing and marketing. At the national level, structural adjustment and economic liberalization policies have resulted in the reduction of subsidies to agricultural inputs, which has affected the incomes of small-scale farmers with limited capital, while the retrenchment of workers in the formal sector has deepened financial insecurity in urban centers. With the introduction of user fees for many services, and the rising cost of food, many households have increased their cash incomes through employment diversification, including nonfarm occupations and urban agriculture, and by migrating. At the local level, the nature and intensity of rural–urban linkages depend on access to agricultural land and farming systems, urban employment opportunities, population density and distribution, and roads and transport networks, all of which are shaped by local government policy. Furthermore, the nature and role of rural–urban linkages in contributing to livelihoods vary according to the wealth and status, gender, age, and ethnicity of household members.

As Tacoli (2003, 3) argues, the notion of a rural/urban divide “has become a misleading metaphor, one that oversimplifies and even distorts the realities.” While there is no denying the significant variations between rural and urban contexts, the increasing convergence of activities and the numerous links between the two make any dividing line imprecise. This has important consequences not only for research but also for policymakers, who tend to be split into separate rural and urban camps in their mandate and through their training. While the structure of local government means that responsibility for development planning often falls into areas that
are designated as urban or rural, even when they form part of a single metropolitan area, it is essential to take into account the multiple linkages and similarities between rural and urban areas when planning for the future.

SEE ALSO: Cities and development; Development; Migration: internal; Population mobility and regional development; Rural geography; Rural policy and politics; Urban geography; Urban planning: human dynamics

References


Further reading


Description and purpose

The Institute of Geography of the Russian Academy of Sciences is the oldest and the largest Russian multidisciplinary research center in physical and human geography coordinating geographical studies in all Russia. Its main research fields are the evolution of natural environment and land natural resources, natural cryogenic systems, geographical problems of land use and environment, interaction between environment and society, sustainable development, urban and rural geography, spatial patterns in Russia and post-Soviet countries, borders studies, political and cultural geography, GIS technologies, and mapping and remote sensing.

Current activities or projects

Apart from fundamental research and theoretical studies, the institute carries out applied studies, is involved in the work on environmental bills, and the monitoring and expertise of large projects of urban and regional development. It published the first maps and reports on the state of environment in the USSR, Russia, and several CIS countries. The institute has worked on a number of projects sponsored by UNEP, World Bank, and different EU programs. It developed the GIS currently used by the Russian Ministry of Emergencies and other governmental institutions. The institute has annually about 40–50 PhD students and hosts 50 visiting scholars.

Brief history

The institute was established in Petrograd in 1918, soon after the 1917 revolution, on the initiative of D.N. Anuchin who, in 1915, was the first to suggest the creation of a specialized geographical research center. The demand of Russia’s growing economy in the early twentieth century for geographic studies could be satisfied neither by the Imperial Geographic Society, involved mainly in expeditions to remote regions, nor by university geography. Systematic work was needed for the geographic description of the country and its newly
RUSSIA

reclaimed regions, such as the Arctic, Siberia, the Far East, and Central Asia. The institute’s work was based on the achievements of Russian geographical schools of the late nineteenth–early twentieth century’s academic legacy of V.V. Dokuchaev, P.P. Semenov-Tyan-Shanskii, A.I. Voeikov, and others. The institute was first called the Commission for the Study of Natural Productive Forces (CSNPF). Ten years later it was renamed as the Geographical Department of CNPF, which in 1934 became the Institute of Geomorphology. In the same year it moved to Moscow and adopted the name “Institute of Physical Geography” and since 1936 bears its present name – the Institute of Geography of the Academy of Sciences of the USSR (now Russian Academy of Sciences).

Submitted by Vladimir Kolossov
Russia: Vserossiyskaya obshchestvennaya organizatsiya (Russian Geographical Society (RGS))

Founded: 1845
Location of headquarters: Moscow
Website: www.rgo.ru
Membership: 14894 (as of December 31, 2014)
President: Sergei Kuzhugetovich Shoigu
Contact: rgo@rgo.ru

Description and purpose

RGS is a nonprofit organization and does not receive state funding. The main goal of the society is to consolidate the power and ambitions of the Russian population in the study and promotion of national geography. The president of Russia, Vladimir Putin, is the chairman of the society.

The society has members in Russia and abroad. Regional offices are present in all 85 Russian regions. The society currently focuses its activities on expeditions and studies, education and grant programs, environmental protection, and publishing.

Journals or major publication series

Izvestiia Russkogo geograficheskogo obshchestva
Lyod i sneg (Ice and Snow), in collaboration with the Institute of Geography RAS. http://ice-snow.igras.ru
Voprosi geografii

Current activities or projects

The society currently organizes expeditions and studies, education and grant programs, environmental protection, and publishing of its journals.

Major projects are: the Russian Geographical Society’s Award ceremony, archaeological and geographical expedition “Kyzyl-Kuragino,” the Russian Geographical Society Festival, the International Arctic Forum “The Arctic: Territory of Dialogue,” Arctic floating university, Far-Eastern floating university, summer youth school, the Geography Olympiad, national youth gathering, protection of endangered species, Arctic cleanup program, undersea exploration, technological platform (Technologies for Sustainable Ecological Development (TP)), lectures in Moscow and Saint-Petersburg, and “Great Russian Travelers” in libraries throughout Russia.

Brief history

The Russian Geographical Society was established by Nicholas I in 1845. The original idea for creating the society came from Admiral Litke, the mentor of the great Duke (Kniaz) Konstantin, who later became the first chairman of the society. Great explorers (Litke, Krusenshtern, Wrangel, Rickord), members of the St Petersburg Science Academy (Ber, Struve, Gelmersen, Keppen), great army leaders (Berg, Vronchenko, Muraviev), as well as outstanding Russian linguists (Dal, Odoevsky) were among those who established the society.

The society has never ceased operation since it was founded. Many great people led the society during its history, from the dukes to
famous explorers, travelers, and state leaders. RGS greatly contributed to the exploration of the European part of Russia, the Urals, Siberia, the Far East, the Caucasus, Iran, India, New Guinea, and the Polar territories as well as other lands. The society was traditionally very close to the navy and marine expeditions.

During the twentieth century many famous polar explorers, such as Papanin, Treshnikov, and Schmidt, were attached to the society. In Soviet times, the society promoted geographical knowledge by establishing special commissions, councils, and the auditorium.

Submitted by Vladimir Kolossov
Rust Belt cities

David Fasenfest
Wayne State University, USA

Traditional manufacturing regions across Pennsylvania, Ohio, Michigan, and Indiana in the United States, in the Ruhr Valley in Germany and the industrial centers in northern England and Japan were signified by prospering cities. These were home to industries like automobile production in Detroit and steel in Pittsburgh, and were exemplified by a growth in employment and a rapid increase in their populations. By the mid-1970s, a combination of shifting global production, technological change, and more recently an intensification of global competition has resulted in the hollowing out of these traditional smokestack industries. Cities that were once the center of growth became known as Rust Belt cities, reflecting disinvestment, population loss, economic decline, and bleak prospects for recovery. These old industrial cities experienced a decline in their urban infrastructure, losses in their local tax base, and a steady outmigration of residents unable to find work. Over the second half of the twentieth century cities like Detroit, Youngstown, and Pittsburgh lost over half of their populations.

Rust Belt cities are making efforts to return through a range of activities and decisions about restructuring and repositioning themselves for the twenty-first century. Many are now witnessing a growth in their downtown populations, especially among 18–24 year olds, as they become a mecca for the arts. Abandoned warehouses and factories afford cheap homes and studio spaces for artists and sculptors, and coffee houses and experimental chefs who focus on local food (the locavore movements and the new urban agrarianism) grown in and around urban spaces are imparting hipness to once declining cities. Some of the most distressed buildings and industrial spaces now have the cachet of “ruin porn” as they adorn the pages of photography books on coffee tables. The success stories and benefits of living in these resurrecting cities are promoted by sites like Rustbelt Living (http://rustbeltliving.com/) and rustwire.com (http://rustwire.com/).

Other strategies include embracing regionalism; building mass transit to connect these city centers to their surrounding areas; reversing job sprawl by bringing jobs back into the cities; repurposing abandoned structures to shift old to new industries with new jobs in areas like biotechnology, nanotechnology, and other green activities. Indeed, some of the smaller Rust Belt cities are in the forefront of creating environmentally conscious communities. Finally, these cities are attracting immigrants, who bring with them an entrepreneurial spirit, create local businesses, and improve services, and in many cities they have begun to reverse the decades-long population decline.

SEE ALSO: Deindustrialization; Firm migration; Globalization; Industrial restructuring

Further reading

RUST BELT CITIES


Description and purpose

The SGC strives to encourage study and research on national and regional topics as they relate to geography in particular, and to cooperate in the development of geography in general. As an advisory board to the national government in geographical matters, it is always ready to provide the information requested and to advise in consultations related to different aspects of the national geographic reality. Similarly, the society tries to attend and solve geographic queries submitted by the general public. It also promotes studies and geographic research through open contests and by means of awards granted to distinguished students in upper-level courses.

Journals or major publication series

Sociedad Geográfica de Colombia, Bulletin Geoventana (newsletter)

Current activities or projects

The society currently has several activities, for example, the organization of ExpoGeografica, an academic event designed to provide opportunities for interaction and exchange of knowledge on products and services of geographic nature as well as to promote the science and praxis of geography, study, and research in geographic problems; and the organization of the National Center of Geographic documentation “Rafael Convers Pinzon.” The society is also in charge of assorted publications; courses on GIS; virtual geography courses; digitization of Agustin Codazzi’s original manuscripts; and other activities, including seminars, lectures, periodic conferences, and research.

Brief history

The first attempt at creating a geographical society in Colombia is credited to Carlos Holguin, when he was ambassador to Spain. Writing in 1893 to Manuel Ponce de Leon, then chair of the Colombian Society of Engineers, Holguin urged him to promote in Bogotá “the organization of a Geographic Society like the Spanish one.” Nevertheless, that initiative was delayed by the climate of political unrest that eventually led the country in 1895 to generalized civilian
COLOMBIA

strife (Thousand Days’ War). After peace was achieved, the national government issued a decree establishing the Sociedad Geografica de Colombia, which was confirmed by Mr. Jose Manuel Marroquin, vice-president of the republic, who inaugurated the society in an official meeting convened at the National Observatory. The event on August 20, 1903 was attended by distinguished members of the government, the army, intellectuals, and members of learned societies. During its more than one hundred years of institutional life, the SGC has been chaired by major researchers and writers related to geography, such as Julio Garavito Armero, Francisco Javier Vergara y Velasco, Jorge Alvarez Lleras, Belisario Ruiz Wilches, Julio Carrizosa Umaña, Alvaro Valencia Tovar, Julio Londoño, and Manuel Jose Forero, among others.

Submitted by Hector F. Rucinque
A “safe haven” refers to a geographic area designated as a place where people might find protection and freedom from violence, harm, persecution, and harassment. The phrase enjoys some recognition in international law (Koh 1994) and has historical connotations associated with concepts of refuge, asylum, shelter, and the welcoming of strangers. Religious institutions and government organizations have a long history of welcoming strangers and providing safe haven that well pre-dates the international systems, laws, and institutional structures put into place in the mid-twentieth century to govern human mobility.

Within geography, “safe haven” has been used in the context of migration and refugee flows and debates about the provision of protection in geographically designated areas. Scholars writing in the subfields of migration and refugee studies have used the term, often reflecting critically on geopolitical projects associated with its use (Chimni 1995; Hyndman 2003; Han 2013).

Contemporary uses and understandings of safe haven are broad and vary by context. Safe haven may refer, for example, to geographical areas such as shelters for those fleeing violence, whether domestic abuse at home or migrants and asylum-seekers in flight from conflict or persecution experienced in their hometown or home country. The safe haven may be as small as a building or a camp, or the designation of safe cities. These examples highlight the notion that a safe haven is a geographic area that operates simultaneously across a number of scales, connecting the intimate space of the body to household, local, domestic, and global economies and politics.

Safe haven for those in flight due to displacement or forced migration involves what the United Nations High Commissioner for Refugees calls “internally displaced persons” (people displaced from home, but still located within the boundaries of their country of origin) or those displaced who crossed an international border during their journey. At the point of crossing an international border, people may ask that government for safe haven in the form of protected legal status, either according to the 1951 UN Convention relating to the status of refugees and its 1967 Protocol, or some other legal designation that grants a person temporary or permanent status in this new “host” country. Those in search of refuge or asylum may register at ports of entry, in refugee camps, or be resettled from safe zones elsewhere.

Related to safe haven, but distinct, is the term “safety zone.” Chimni traces the notion of safety zones created in the Geneva Conventions as multilateral arrangements wherein “individuals fleeing danger can seek safety within their own country” (1995, 825). As Chimni notes, however, post-Cold War geopolitics make safe havens less about protection than they might appear.

Safe haven also references historical moments on the international stage when states have devoted resources to the designation of particular locations (such as refugee camps, cities, or bases) as protected. Since the 1990s, the politics surrounding asylum-seeking has supported the closing down of routes to safe haven, thus shrinking...
SAFE HAVEN

protection and spaces of safe haven through aggressive policing (often offshore) in the form of border enforcement and discursive othering.

A key issue is that there are always contradictions associated with designation of a territory as a “safe haven.” Projects to provide protection contain unintended consequences, underlying objectives, and more safety for some than for others. Safe havens “other” at the same time as they protect, through their spatial and temporal constitution of inside and outside.

While “safe haven” is a status that is well established in international law (Koh 1994), the purpose of this designation does not always correspond with its original intent. In the international realm of migration politics, for example, projects called “safe havens” may in fact be a way of controlling human mobility and inhibiting access to rights accrued when people reach sovereign territory. Thus, the geopolitics of asylum and actions carried out by nation-states in the name of safe haven often thwart the very protection they purportedly establish. As Hyndman (2003, 167) notes, in comparing United Nations-designated “safe spaces” to protect displaced civilians beginning in the 1990s, “Some safe havens and protected areas are safer than others.” Safe havens will prove safer for some than for others based on context, identity, history, and the particularities of the conflict and associated violence (Chimni 1995). Others will belie underlying projects to contain people and problems (Hyndman 2003, 168).

One such example involves enforcement activity of the United States in the Caribbean in efforts to control migration by sea in the 1980s and early 1990s. This was a period of mass migration by sea when Cuban and Haitian nationals (among others) embarked on boat journeys to the United States to seek asylum. During this time, Panama and Honduras agreed to provide temporary shelter for those intercepted by the United States at sea, and the United States arranged “safe haven” camps on US bases in Antigua, Dominica, St Lucia, Suriname, and the Turks and Caicos Islands (Koh 1994, 154). The most notorious safe haven camp in the region was on the US naval base in Guantánamo Bay, Cuba. There, President George H.W. Bush offered “safe haven” to Haitian nationals who fled the military coup against President Jean-Bertrand Aristide. This practice continued under President Bill Clinton’s administration. Haitian and Cuban asylum-seekers were held there after being intercepted at sea by US patrols. While narrated as a humanitarian response, detention was “indefinite,” and the only way out for most detained people was not a path to asylum in the United States but “voluntary” repatriation, or return home. The Guantánamo Bay Naval Base later became notorious under President George W. Bush’s administration for its continued use of indefinite detention, this time of foreign nationals arrested during the “war on terror.”

Offshore enforcement renders complex the ability of individuals to seek asylum, since they are often denied rights associated with landing on sovereign territory. Therefore, the US safe havens of the 1990s to house those seeking asylum but intercepted at sea offer an example of the contradictory and geopolticized uses of the term “safe haven.” Another similar pattern involves the use of islands designated as sites of humanitarian rescue but used as platforms for offshoring asylum through detention and deportation off the coasts of North America, Europe, and Australia. Camps set up in the name of humanitarian rescue can quickly become sites where racialized forms of exclusion are carried out and where people are denied the protections of human rights, access to legal representation, and due process. Military bases used as safe havens, for example, are often places where sexual violence threatens women on- and off-base.
Projects enacted in the name of protection through legislation also have other underlying political agendas. Sanger (2006) argues, for instance, that infant safe haven laws developed in the early 2000s in the United States were ostensibly designed to decriminalize mothers’ abandonment of their children: by leaving infants in a safe haven, the infant would be cared for by the state, rather than harmed. In practice, Sanger suggests, these functioned as part of US culture wars related to protecting infant life and working toward overturning Roe v. Wade, which allowed for the performance of abortions.

**SEE ALSO:** Migration: forced; Migration: internal; Migration: international; Refugees; Safe space; Violence

**References**


The term “safe space” has long been in operation to reference a physical location or shelter where people go to be safe from threats of physical harm. While the term has a lengthy association with sanctuary in safe spaces such as churches (Coyle 2004, 67), it has more recent associations with specific groups under threat and in search of safe haven. The term gained currency in the second wave of the women’s movement, in which “safe space” is used to refer to places dedicated to occupation by survivors of domestic violence and rape, and for demarcating places of safety in expression. More recently, this last usage has become commonly associated with queer, gay, lesbian, bisexual, and transgender people seeking space dedicated to freedom from homophobia, transphobia, and other forms of discrimination and violence. Particularly prevalent in educational institutions, “safe space” stickers featuring rainbows or pink triangles were a common sight on university campuses in North America in the 1990s and 2000s. Located on doors and bulletin boards of campus spaces, these indicated a space occupied by people supportive of LGBT communities, students, and staff.

The term is highly geographical as it dedicates space, assigns meaning, and draws parameters around particular locations. Accordingly, it has been used in various ways by geographers. Until recently, most of this usage remained casual; scholars did not study the genealogy of the term so much as simply reference it in their writing. In the 1980s and 1990s geographers made occasional reference to safe space and discussed safety more generally, particularly in feminist and gendered geographies of women and children and their perceptions of spaces deemed safe or unsafe (see Pain 2000).

In the 1990s and 2000s the term gained currency in geographies of gender, sexuality, and sexual identity in the discipline. Safe space operated across a range of scales in this literature, often referring to specific locations or gay neighborhoods within cities where LGBT individuals created positive spaces, whether communal or commercial, public or private. Borrowing from queer scholarship and activism, these scholars operationalized the term but generally used it loosely without necessarily defining or delimiting its meaning. As feminist and poststructural scholarship attended increasingly to the body and emotions, new uses of “safe space” emerged. The term arose in relation to emotional geographies of fear, for example (Pain 2000), and of the body (Coyle 2004). These approaches generated novel understandings of safe space. Coyle (2004), for example, conceptualizes safe space as domestic space where women suffering somatic symptoms of environmental illness are able to stabilize their health through the elimination of toxic chemical substances found in other places. The term is also used more widely in the subdiscipline of health geographies with reference to toxic environments for people with environmental, chronic illness, and other conditions, who may be highly susceptible to certain kinds of environments and require safe space that is healthy (Coyle 2004).
**SAFE SPACE**

Safe space is often discussed in relation to spaces of fear (of crime and physical violence), and debates recur about whether safety is more likely to be found in public or private space. While many associate danger with public space, much violence against women occurs in the private space of home (Pain 2000, 373). “Safe space” is also sometimes used interchangeably with “safe haven,” particularly by political geographers working on displacement and refugee issues in sites of conflict where people might seek refuge from physical violence. Physical scientists have also used the term in relation to climate change and planning safe space to sustain human life (Roestone Collective 2014, 1346–1347).

Current scholarship on safe space in the discipline continues in the fields of queer and feminist geography. Recent literature demonstrates that sites deemed safe spaces that may simultaneously protect and exclude or otherwise cause harm. The Roestone Collective (2014) published an essay that reviews uses of the term and suggests its reconceptualization. Observing that safe space has been defined statically in relation to imagined and constructed “unsafe” spaces, they advocate for “safe space as relational work” (2014, 1347–1348) that involves negotiation of differences. They review safe spaces deemed separatist (e.g., Take Back the Night) and inclusive (e.g., classrooms) and find some to be classist, racist, and sexist (Roestone Collective 2014, 1348). Put another way, safe space is paradoxical space (Roestone Collective 2014, 1356). The Roestone Collective (2014, 1360) concludes that “cultivation of safe space is a way of practicing social justice that recognizes, emphasizes, and in some ways encourages social difference.”

Beyond the discipline, scholars are also arguing that – like all spaces – safe spaces are racialized and classed, and more welcoming to some than others. In a recent book called *Safe Space*, Hanhardt (2013) traces LGBT activism surrounding safety in gay neighborhoods. She links these grassroots movements to neoliberal policing initiatives that drew on privatization, white supremacist imaginaries, and constructions of blackness in relation to LGBT communities, which proved devastatingly divisive in the racialized and classed exclusions that characterize urban redevelopment. Hanhardt complicates and advances debates about and historical records of safe space by showing how constructions of safety are embroiled in neoliberal policing projects that create safety for some (privileged people) at the expense of intensifying precarity for others marked by age, race, class, gender, and sexual identity. She reveals safe spaces as racialized, hegemonic constructions of exclusivity and privilege.

Like many other geographical locations designed for purpose of protection, safe space is a double-edged sword, or paradoxical in the Roestone Collective’s terms: it involves the likelihood of social exclusion at the same time as it is designed with the intention of protecting.

**SEE ALSO:** Safe haven; Sexualities; Violence

**References**


Saudi Arabia: Aj-jam’aiya Aj-joğrafiya as-Sa‘ūdiyya (Saudi Geographical Society)

Founded: 1985
Location of headquarters: Riyadh
Website: www.saudigs.org
Membership: 227 (as of December 31, 2013)
Chairman: Mohammed Shawqi Ibrahim Makki
Contact: sgs@ksu.edu.sa, makki14@live.com

Description and purpose

The Saudi Geographical Society at King Saud University was established to link national expertise in geography, whether in education, research, or training in various areas of geography and technology, so as to contribute to scientific progress in the discipline and to promote cooperation with other institutions at the national and international level. The SGS sponsors research, seminars, panel discussions, public lectures, scientific publishing, training courses, joint scientific projects, and specialized consulting to serve the nation and its development. The SGS is also keen to promote understanding of the role of geography and its applications in balancing environmental and cultural heritage in national development programs at all levels in the framework of the Islamic faith and within the context of Arab and Islamic heritage, as well as projects developed by members of the SGS and those aimed at helping the employment of new geographers.

The vision of the society focuses on the development of human culture and the environment. Thus, SGS seeks to play a leading role in deepening and expanding the geographical knowledge of specialists and members of the community in general, while enhancing communication with educational institutions. It also seeks to promote professional, research, and training within and outside the kingdom within the guidelines of Islamic concepts; to serve as an influential voice for the local community in the kingdom; and interact with the scientific world at large.

Journals or major publication series


Current activities or projects

The SGS has undertaken a number of research activities including forums and conferences attended by prominent geographical research figures from all over the kingdom. Such activities have contributed greatly in enriching the role of the society and making it a more effective research institution.

The board of directors applied to register the society with the International Geographical Union (IGU), with the approval of the king, so that the Kingdom of Saudi Arabia could be represented by the Saudi Geographical Society as a member to the IGU. The union approved this membership in November 2000 and we hope that this will help to highlight the society’s work globally and contribute to
communication between Saudi geographers and other geographers around the world.

The Saudi Geographical Society contributed to the establishment of the Gulf Geographic Society and has helped support this new organization. Members of the board of directors call upon senior officials of the state to introduce them to society activities and projects and to indicate the extent to which these projects need moral and financial support.

The Third Forum of Arab Geographers of the Saudi Geographic Society, entitled “The Major Cities in the Arab World,” was held at the King Saud University in Riyadh. The SGS has coordinated the event along with the Yemen Geographical Society and the Egyptian Geographical Society, which hosted the first and second meetings of Arab geographers.

Brief history

The establishment of the Saudi Geographical Society was approved during the 11th session of the King Saud University Research Council held on March 13, 1983. King Saud University became the sponsor and headquarters for the society, although it has the option to establish, if required, additional branches in other places.

The statute was amended within the guidelines regulating research societies operating in Saudi universities, which was ratified by the Higher Education Council by virtue of its decision No. 10/15/1420, and the endorsement by the custodian of the two holy mosques, chairman of the Higher Education Council, of the council’s decision made during its 284th session held on June 18, 2000. The first general congress meeting of the Saudi Geographical Society took place on December 21, 1984. Prior to the establishment of the Saudi Geographical Society, there were a number of small geographical associations operating within the geography departments of the Saudi universities. The society has published considerable research and has a number of projects underway or in process.

Submitted by Mohammad S. Makki
Scale and anti-scale

John Paul Jones III
University of Arizona, USA

The measurement of scale is central to understanding how maps represent the Earth’s surface and in popular usage is still the most common meaning attached to the term. Beginning in the 1980s, however, the term “scale,” especially as it was used by human geographers, was dramatically broadened from its traditional definition as a measure of cartographic transformation. During this period, scale became one of the pillars of social theoretic reflection in geography, and for many geographers it became perhaps the central contribution of the spatial perspective to the social sciences more generally. The arguments were simple and appealing. If, as was being widely acknowledged at the time, the operations of social processes are to a great extent dependent upon the larger sociocultural, historical, and geographical contexts in which they are found, then might not these processes function differently, or even be sorted according to, various scalar contexts, such as the local, regional, national, and global levels? If yes, then geographers could argue that scale is foundational to social explanation more generally. What is more, if scalar contexts are the result of social processes, rather than being simply given as set categories of analysis, then social and spatial explanation would need to go hand-in-hand. This pairing offers compelling reasons why “geography matters,” or should matter, to the broader social sciences. Not only were many human geographers attracted to theories of scale in geography, so too were many theoretically-inclined social scientists outside the discipline.

Scale theory arose in the wake of, and is sometimes viewed as a component of, the sociospatial dialectic, an influential theory that maintains that social processes simultaneously produce and are produced by space, including all aspects of geography, from the forms and organization of our natural and built environments, the material, symbolic, and ideological aspects of place, and the everyday geographies forged through human lived experience. Scale is one concept within such dialectical thinking. Yet, at the same time that scale was being developed as part of spatial dialectics, the critical wing of the discipline was undergoing other shifts; in particular, away from what was perceived to be a strict materialist, economistic, and structural Marxism towards a more avowedly poststructuralist (née postmodern) body of theory. The ideas geographers brought to the concept of scale were affected by these transitions. Whereas initially theorists viewed scale as the product of capitalist social relations, later, under the sway of poststructuralism’s less centralized and more discursively-inflected forms of power, the boundaries and operations of scale became more uncertain. Indeed, under poststructuralism’s “crisis of representation,” questions arose as to whether scale is, in fact, an ontological bedrock of social space or merely an epistemological framework that we impute to space to help provide order and meaning. The local-to-global scaffolding upon which scale was initially theorized was challenged by such nonhierarchical theories as found in Gilles Deleuze’s and Félix Guattari’s philosophy and Bruno Latour’s actor-network theory. These issues are at the
center of a series of works that have made scalar and anti-scalar theories a much debated area in the 2000s.

**Scale as cartographic measure**

In its traditional, cartographic meaning, scale is defined as the ratio of the distance on a map relative to that same distance on the Earth’s surface. By convention, the first number in the ratio is the map’s unit (=1) distance, whereas the second number is the corresponding (i.e., same metric) distance on the Earth’s surface. So, for example, a map ratio of 1:100000 indicates that every, say, inch of map distance shows 100000 inches of surface distance, or roughly 1.6 miles (100000 inches/5280 feet/12 inches = 1.578 miles). This is a common ratio used by the United States Geological Survey for topographic mapping at an intermediate scale.

Map scales are distinguished as large or small based on the size of the ratio of the two numbers. A common USGS map at 1:24000 is said to be relatively large scale, owing to the size of the resulting number, relative to a map with a scale of 1:250000, which has a much lower value and is, therefore, referred to as smaller in scale. A point of common confusion is that since larger scale maps reduce the Earth’s surface to a lesser extent, they therefore show less surface area, while small scale maps are needed to show large areas. A 1:24000 map enables viewers to identify urban or rural features such as streams, roadways, and land use patterns, while maps at smaller ratios, such as the commonly used 1:100000 or 1:250000 maps, are used to discern the shapes of river basins and state political boundaries. The zoom function in digital mapping has made these fixed scales much less relevant. Seasoned map enthusiasts can tell you what scale a USGS map is by simply glancing at it, but we now commonly select our scales in a continuous fashion based on our needs by simply scrolling a mouse.

**Theorizing scale**

It was Peter Taylor who first offered a social theory of scale. In his 1982 article he makes two arguments. First, he maintains that political economy should be the theoretical basis for the subdiscipline of political geography. Second, he argues that scale should be the grounds for spatializing this body of theory. (Taken together, it is not surprising that, after Taylor, theories of scale have seen most of their influence in political and economic geography.) Inter alia, Taylor critiques several treatments of scale then extant in political geography, finding them to be drawn from commonsense understandings instead of rigorous theorizing. His avowedly materialist theory of scale derives from Immanuel Wallerstein’s world-systems approach, which Taylor affirms but identifies as deficient in its implicit horizontal approach to space (i.e., as areal extension rather than in terms of spatial levels). In its place, he proposes a “political economy of scale” with three vertically conceived levels: “the scale of reality (global), the scale of ideology (state) and the scale of experience (urban)” (Taylor 1982, 24). Here reality refers to the facts of capitalist accumulation at the global scale, while ideology implies a view of the state as an institutional apparatus “whose purpose is to simply separate experience [the urban scale] from reality” (Taylor 1982, 24). Crucially for Taylor, the global political economy is foundational, yet every scale is relational:

Hence we do not propose three processes operating at three scales but simply a single manifestation of capitalist accumulation within which the arrangement of three scales is functionally important. For instance, the needs of accumulation will be experienced
locally (e.g. closure of a hospital) and justified nationally (e.g. to promote national solvency) for the ultimate benefits organized globally (e.g. by multi-national corporations paying less tax). (Taylor 1982, 24)

The second important contribution to scale comes from Neil Smith in his 1984 book, Uneven Development. Throughout, Smith delivers an impressive presentation and unpacking of a wide range of concepts related to the “dialectic of [spatial] differentiation and equalization” (Smith 1984, 135) under capitalism. By the time we get to his chapter on scale and the “see-saw” of capital (Smith 1984, 131–154), we know that scales do not simply exist as given levels, but are at the heart of uneven development under capitalism: “Capital inherits a geographical world that is already differentiated into complex spatial patterns. As the landscape falls under the sway of capital (and becomes increasingly functional for it …), these patterns are grouped into an increasingly systematic hierarchy of spatial scales” (Smith 1984, 135). One of Smith’s advances over Taylor’s, it could be argued, lies in his departure from what he calls Wallerstein’s “perspective of exchange space” (Smith 1984, 176, footnote 7). For Smith, scale is produced by capital, while capital itself becomes bound to that spatial configuration:

I think it is possible to use the dialectic of differentiation and equalization to derive the actual spatial scales produced by capital, and to show that the result of uneven development is simultaneously more complex and [simpler] than a mosaic. There is little doubt about the impossibility of a spatial fix for the internal contradictions of capital, but in the doomed attempt to realize this spatial fix, capital achieves a degree of spatial fixity organized into identifiably separate scales of social activity. (Smith 1984, 135)

Smith agrees with Taylor that scales exist at the urban, nation-state, and global levels, but he is reluctant to assign to these levels a specific and allied social process (e.g., ideology to the nation-state, experience to the urban). Rather, as he shows in subsections devoted to each level, capitalism produces scales as part-and-parcel of its endless shifting between equilibrium and disequilibrium. It does so through the “spatial fixes” – that is, centralization, expansion, and jumping into new spaces altogether – that are inherent in solving capitalism’s fundamental contradictions. But however fixed these scales are made, they are subject to change, and it is through the continual determination and internal differentiation of spatial scale that the uneven development of capitalism is organized. The vital point here is not simply to take these spatial scales as given, no matter how self-evident they appear, but to understand the origins, determination and inner coherence and differentiation of each scale as already contained within the structure of capital. (Smith 1984, 136)

In the two decades following these two important works, the discipline saw many extensions and refinements to the concept of scale. These took several interrelated lines of analysis. The first was a focus on the “politics of scale” – a result of the recognition that, while capital might attempt to produce space in its own image, scales could also be strategized and fought over. Among those who made important contributions in this area was John Agnew, who demonstrated that Italian political parties organize around spatial categories of the local, regional, and national so as to “define the geographical scales that channel and limit their political horizons” (Agnew 1997, 118). Similarly, Byron Miller showed how social movements, such as peace activists in Massachusetts, deploy different scalar strategies as political opportunities suggest themselves (Miller 1994). More generally, as Erik Swyngedouw puts it:
Geographical configurations as a set of interacting and nested scales (the “gestalt of scale”) become produced as temporary stand-offs in a perpetual transformative, and on occasion transgressive, social–spatial power struggle. These struggles change the importance and role of certain geographical scales, reassert the importance of others, and sometimes create entirely new significant scales, but – most importantly – these scale redefinitions alter and express changes in the geometry of social power by strengthening power and control by some while disempowering others. (Swyngedouw 1997, 169)

Swyngedouw’s distributed notion of power is consistent with a second line of analysis that developed in the scale literature, namely the loosening of strict scalar boundaries (e.g., urban, nation state, global) and the broader integration of vertical scalar productions with horizontally conceived social networks. Neil Smith, for example, continued his elaboration of malleable scales, coining the terms “scale jumping” to describe how political power established at one scale can be expanded to another, and “scale bending,” the process by which “entrenched assumptions about what kinds of social activities fit properly at which scales are being systematically challenged and upset” (Smith 2004, 193). Less rigid forms of scalar thinking also emerged as geographers came to see that scales and networks together might provide a more potent descriptive and explanatory framework. For example, Kevin Cox (1998) proposed that scales are contingent on the “networks of association” upon which different capitals and local states depend. As networks become more global in reach, they begin to stretch across scales. The necessity of relaxing otherwise rigid understandings of scale was, for Neil Brenner, a result of globalization, which he defined in terms of scale: “a reconfiguration and re-territorialization of superimposed spatial scales, and not as a mono-directional implosion of global forces into sub-global realms” (Brenner 1997, 159). He explicitly links scales and networks, offering that:

Scales evolve relationally within tangled hierarchies and dispersed interscalar networks. The meaning, function, history and dynamics of any one geographical scale can only be grasped relationally, in terms of upwards, downwards and sideways links to other geographical scales situated within tangled scalar hierarchies and dispersed interscalar networks. ... Each geographical scale is constituted through its historically evolving positionality within a larger relations grid of vertically “stretched” and horizontally “dispersed” sociospatial processes, relations and interdependencies. (Brenner 2001, 605–606, emphasis in original)

Helga Leitner is another theorist of scale who responds to globalization – particularly of the political variety – by integrating vertical scales with horizontal networks. She writes:

transnational networks represent new modes of coordination and governance, a new politics of horizontal relations that also has a distinct spatiality. Whereas the spatiality of a politics of scale is associated with vertical relations among nested territorially defined political entities, by contrast, networks span space rather than covering it, transgressing the boundaries that separate and define these political entities. (Leitner 2004, 237)

A third development in the scale literature comes from glances downward – downward even from the urban scale – to the scale of the home, the street, and the body. In what remains the most cited paper on scale, Sallie Marston (2000) deploys the concept of social reproduction to illustrate how most scale theory has been: “largely unresponsive to questions of difference in human agents and how power relations outside the relations of capital and labor might also influence scale-making” (Marston 2000, 238). In response, she situates the home as the scale at which the
everyday relations of patriarchy, racism, and citizenship connect to wider scales. Marston demonstrates her argument through a description of the expanding consciousness and political roles of US women in the nineteenth and early twentieth centuries:

greater political empowerment proceeded from the reconstitution and reclamation of the social geography of daily life. A discourse about women as “female citizens” operated among and between scales from the household out to the globe and provided these subjects with a consciousness that enabled a particular negotiation of patriarchal subordination and began a gender transformation of the public sphere through a reconstitution of the private sphere of the home. In short, the home was utilized as a scale of social and political identity formation that eventually enabled American middle-class urban women to extend their influence beyond the home to other scales of social life. (Marston 2000, 235)

In what might be said to be geography’s first skirmish over scale, Neil Brenner reacted to Marston’s expansive theorization. His aim in responding was to:

contribute to the development of an approach to sociospatial theory in which the specifically scalar dimensions of social spatiality – in contradistinction to its many other dimensions, such as localization, place-making, territorialization, spatial distanciation, the formation of spatial networks, the production of environment/nature and so forth – may be adequately recognized and theorized. (Brenner 2001, 593)

In arguing for “a more precise and hence analytically narrower conception of geographical scale” (Brenner 2001, 593), Brenner’s reply – which includes eleven methodological hypotheses – was only strategically directed at Marston, for, as elaborated above, since its initial tracing by Taylor and Smith, the concept of scale has been progressively widened through a series of theoretical interventions. These have made the concept of scale both more processually inclusive (e.g., through a focus on the politics of scale and the inclusion of social reproduction and environmental processes in scale theory) and, relatedly, more spatially complex (e.g., by including horizontal networks and households). In offering that scale should be tied to “an explicit causal argument linking the substantive social content of the spatial unit in question to its embeddedness or positionality within a broader scalar hierarchy” (Brenner 2001, 600, emphasis in original), Brenner sought to limit scale to “relations of hierarchization and rehierarchization among vertically differentiated spatial units”, such that they can be distinguished “from other forms of sociospatial structuration” (Brenner 2001, 603).

Marston’s reply to Brenner, co-authored with Smith (Marston and Smith 2001), concedes the importance of more analytic precision around scale, but concludes that Brenner will not find the tools for it by maintaining boundaries between scalar production and the wider social production of space (à la Lefebvre):

scale is a produced societal metric that differentiates space; it is not space per se. Yet “geographical scale” is not simply a “hierarchically ordered system” placed over pre-existing space, however much that hierarchical ordering may itself be fluid. Rather the production of scale is integral to the production of space, all the way down. Scaled social processes pupate specific productions of space while the production of space generates distinct structures of geographical scale. The process is highly fluid and dynamic, its social authorship broad-based, and the scale of the household (or the home) is integral to this process. So too, we contend, is the scale of the body. (Marston and Smith 2001, 615–616)

In addition to their theoretical response – which is centrally pointed to the question of what kind
of space scale is—Marston and Smith also criticize Brenner for an “inability to see the theoretical relevance of the social reproduction argument” (Marston and Smith 2001, 617); they maintain that it is “arbitrary that the home is relegated to a ‘place’ or ‘arena,’ while the state gets to be a multifaceted ‘scale’” (Marston and Smith 2001, 618). “Future historical research may yet reveal the household to be a ‘stable background structure’ in all of this,” they reply, “but the smart money will be wagered elsewhere” (Marston and Smith 2001, 618).

Anti-scale theorizing

As suggested above, scale might not merit a simplistic oppositional trope—“anti-scale”—for it has long proved to be a multifaceted, evolving, and contested concept, even among its proponents. Nonetheless, it is clear that some time near the end of the century geographers began to reflect more critically on the analytics of scale—as witnessed in the debate between Brenner and Marston and Smith—but also, and more fundamentally, on the very question of whether scale adds value to the geographic lexicon. For example, in addition to her questioning of the primacy of capital, labor, and the state in the social construction of scale, Marston’s seminal intervention presaged her and others’ later work under the anti-scale label when she openly entertained the “the rejection of scale as an ontologically given category” (Marston 2000, 220). This point echoes an earlier but less widely-known essay by Katherine Jones (1998, 28), who may have been the first to sharpen the ontological and epistemological distinctions of scale when she wrote: “[W]e may be best served by approaching scale not as an ontological structure which ‘exists’, but as an epistemological one—a way of knowing or apprehending.” Writing in response to Kevin Cox (1998), who, like Brenner (2001), was concerned to distinguish between scales and areas, such as localities, Jones suggested that the politics of scale may be little more than a special case of the politics of representation:

If scale is a representational practice deployed by participants in struggles, a practice situated within a community of producers and readers who actively negotiate and construct it, then what is its ontological status? Does scale exist beyond that community as a fundamental structure of the world, or is it a mode for apprehending the world that is tied to a particular historical/geographical context? Furthermore, does it make any difference whether we see scale as a fundamental ontological category, or as an epistemology, and if it does make a difference, then what is that difference? I would argue that scale is an epistemological category, rather than an ontological one, and that the difference is an important one. (Jones 1998, 27)

The most extensive critique of scale to date came in a 2005 article by Sallie Marston, John Paul Jones III, and Keith Woodward (2005). Unlike Brenner (2001), Marston and her colleagues saw no possibility of achieving analytic specificity in the concept, which they claimed was ontologically and, hence, causally deficient. They illustrated their point by claiming that the primary difference between the horizontal

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Spatial associations of the horizontal and vertical.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horizontal geographies</td>
<td>Vertical geographies</td>
</tr>
<tr>
<td>Network</td>
<td>Scaffold</td>
</tr>
<tr>
<td>Extensive</td>
<td>Layered</td>
</tr>
<tr>
<td>Horizon</td>
<td>Summit</td>
</tr>
<tr>
<td>Distance</td>
<td>Elevation</td>
</tr>
<tr>
<td>Milieu</td>
<td>Dominion</td>
</tr>
<tr>
<td>Dispersed</td>
<td>Stacked</td>
</tr>
</tbody>
</table>

geographies of networks and the vertical geographies of scales is the reflexive position (i.e., epistemology) of researchers’ spatial imaginaries (Table 1), and not scale’s ontological – and therefore casually grounded – foundation. The resulting confusion leads, they argue, to conceptual gymnastics that are increasingly detached from the concrete social spaces of everyday practice: “[O]ne encounters … ‘structures’ not at some level once removed, ‘up there’ in a vertical imaginary, but on the ground, in practice, the result of marking territories horizontally through boundaries and enclosures, documents and rules, enforcing agents and their authoritative resources” (Marston, Jones, and Woodward 2005, 420). Scale theorizing, they continue, has so infiltrated – and been affected by – discourses of globalization that “over the past twenty years, political and economic geographers have tended toward macro pronouncements that assigned the global more causal force, assumed it to be more orderly … and, by implication, relegated [the local] to the status of the case study” (Marston, Jones, and Woodward 2005, 421). Such “globe talk” cordons off, if not eviscerates, agency and resistance relative to detached, but somehow powerful forces such as “global capitalism,” “national social formations,” and the like. These conceptual apparatuses diminish the epistemological, methodological, and political insights of feminists and other theorists of the everyday (Gibson-Graham 2002). Critically for them, these problems will not be solved by replacing one transcendent account of spatiality (scale) with another (networks):

Network-based horizontality does avoid some of the problems [of scale], but in reviewing this literature we see significant evidence of ‘flowsterism’: the idea that people, phenomena and processes somehow fly above the stickiness of space in an atmosphere of frictionless fluidity. We find, moreover, the same tendency to spatial abstraction in the horizontal view that we also criticize in the vertical one, with scattering lines of flows now standing as transcendental counterparts to layers of nested territories. (Jones, Woodward, and Marston 2007, 265)

In their 2005 and later articles, these anti-scale critics draw on various conceptual resources, including Deleuze and Guattari, Latour, Spinoza, and philosopher Theodore Schatzki, to construct a spatial ontology that replaces both scalar and network theorizing with a “flat,” or “site,” ontology, which they propose to consist of:

immanent (self-organizing) event-spaces dynamically composed of bodies, doings and sayings. Sites are differentiated and differentiating, unfolding singularities that are not only dynamic, but also “hang together” through the congealments and blockages of force relations. The “actuality” of any site is always poised for compositional variation – subject to reorganizations and disorganizations—as its inexhaustible “virtuality” or potential continually rearticulates itself. (Jones, Woodward, and Marston 2007, 265)

Site ontology thus stands as an alternative to the structural imperatives of scalar spatiality, avoiding the latter’s transcendent logics, predetermined spatial frames, and axiomatic strategies that, Marston and her colleagues maintain, identify research problems such that they can be “solved for scale” (Marston, Jones, and Woodward 2005, 426).

Conclusion

In 2006 and 2007 the journal Transactions of the Institute of British Geographers published a handful of replies – some spiritedly critical, others sympathetic – to “Human Geography without Scale” (Marston, Jones, and Woodward 2005), and the authors penned a lengthy rejoinder (Jones, Woodward, and Marston 2007).
SCALE AND ANTI-SCALE

The exchange has become known as the discipline’s “scale debates.” Since then, a number of other commentators have offered additional perspectives, leaving the current thinking on scale falling into two camps: (i) those who have reaffirmed scale’s ontological status (sometimes in combination with network theories), and as part of this view, continue to champion it as a foundational component of social space (Jessop, Brenner, and Jones 2008; Leitner, Sheppard, and Sziarto 2008); and (ii) those in geography and elsewhere who have criticized not only structuralist approaches to social relations, but also the attendant and conformant structuralisms of hierarchical spatiality, and who are thus supportive of developing ontological alternatives to scalar thought (Escobar 2007; Isin 2007).

Regardless of their position with respect to scale’s ontological status, both groups are generally united in their agreement on a third position, namely, that the scalar imaginary – that is, scale as epistemology – plays a significant role in social determination (Moore 2008; Kaiser and Nikiforova 2008). Perhaps here it bears emphasizing that no scale critic ever discounted the fact that people’s geographic imaginaries are often socially constructed in scalar terms. This is especially so in the Western tradition, where spatial order remains a powerful legacy of Cartesian thought. Likewise, no scale critic ever dismissed the fact that scale talk – that is, scale deployed as a discourse – is political, and thus regularly fought over by social actors. In this sense, anti-scale theorists continue the project of elaborating the “politics of scale,” even while they resolve to replace the concept with others more consistent with a poststructuralist ontology, such as the assemblage or the site. And so it might be said that the singularly important and as yet unanswered question about scale is a larger, disciplinary one: What is this thing we call “space,” and what added value does scale provide in our efforts to define it?

SEE ALSO: Critical geography; Globalization; Glocalization; Marxist geography; Ontology: theoretical perspectives; Scale; Social constructionism; Uneven regional development

References


Further reading


Scale

Jingxiong Zhang
Wuhan University, China

The concepts

As widely understood, geographic information is georeferenced information (GI) about spatial entities (e.g., buildings and rivers) and distributions (e.g., terrain elevation and land cover), and is becoming increasingly important for modern society. The increasing sophistication of spatial information technology, proliferation of spatial data and information products, and the development of geographic information science have prompted greater research efforts on some of the fundamental issues surrounding consumption of GI, including that of scale. Indeed, as an essential dimension of geospatial research and applications, scale is becoming increasingly acute as an issue due to accumulation of spatial data (with inconsistencies in geometry, attributes, and semantics), the ease of processing datasets of different scales in a computerized environment, and a lack of readily available tools to handle multiple scales and the associated challenges for interoperability (Zhang, Atkinson, and Goodchild 2014).

Definitions of scale include geographic scale, operational scale, measurement scale, and cartographic scale. As reviewed by Lam and Quattrochi (1992) and Zhang, Atkinson, and Goodchild (2014), geographic scale is also called observational scale, referring to the size or spatial extent of the study. The operational scale refers to the spatial (or temporal) extent at which certain processes operate in the environment and is also known as process scale or intrinsic scale (Blöschl and Sivapalan 1995). The measurement scale is commonly called (spatial) resolution and is commonly indicated by the sampling intervals or pixel sizes in remote sensing images. The process scale of a geographic phenomenon needs to be quantified to determine both the geographic scale and the measurement scale of observed patterns resulting from the underlying process. Cartographic scale refers to the ratio between the length measurements on a map and the actual measurements on the ground; it is also known as map scale. It refers to data representation, unlike geographic scale, process scale, and measurement scale, which refer to data characteristics (Lam and Quattrochi 1992).

An important concept related to (measurement) scale is data support. Data supports generally refer to the shapes and geometric dimensions (i.e., length, area or volume) associated with individual data. While certain entities or variables can be measured or evaluated at point-like locations (e.g., bench marks in land surveying), many others are sampled over units of positive finite geometric dimensions (e.g., sample plots in forest surveys and pixels in images). Data and geoprocessing over finite supports are common in geography, such as areal (polygonal) data of land use and demography, analysis over areal units, or inference about one set of areal units from another, as is the case for lattice data (Saveliev, Mukharamova, and Zuur 2007).

Discussion about data support may also be framed in relation to geospatial conceptualization or data modeling, for which there are two complementary models: entity models and fields. In the former, the real world is conceived
of as populated by discrete entities or objects \( ID(x, \ attr) \), in which position is indicated by \( x \) and attributes (both qualitative and quantitative) by \( attr \), while, for the latter, the real world consists of layers of single-valued functions (either continuous variables \( Z(x) \) or discrete variables \( C(x) \)) defined over space. In both object- and field-based representations, \( x \) should be viewed as data support rather than a mere notion of position, thus indicating the multiplicity of scales, as \( x \) can be different-sized units implying different spatial resolution.

Although scales of geographic phenomena, measurement, and analyses are required to be commensurate, there are often scale discrepancies among them due to the complexity of scale dependencies of geographic phenomena and differences in their data support. This indicates the importance of relating and communicating information across scales, which is facilitated by change of scale (i.e., scaling). For example, scaling is involved when predictions need to be made at a scale that is different from the scale at which data are acquired. There are two kinds of scaling in terms of the relationships between the scale of the existing data (or analysis) and the required scale: (i) scaling up or upscaling, which is information translation from finer resolution (smaller supports) to coarser resolution (larger supports), and (ii) scaling down or downscaling, which is to translate information from coarser to finer resolution, as discussed by Meentemeyer (1989), and Blöschl and Sivapalan (1995). Changing the support (or scale) of a variable creates a new variable, related to the original one, but with different statistical and spatial properties. This is a key idea in research on change of scale, as elaborated later.

In summary, scale is an essential metric for understanding information mechanisms in geographic information gathering, processing, and dissemination. Research on scale issues should focus on models for characterizing scale in spatial data and the underlying geographic patterns and processes, and methods for changing scale. Equally important are related tool developments to assist consumers of geographic information.

**Modeling scale**

There are various techniques to measure scale in a given dataset and to model scale in the underlying data-generating processes. Below, some of the scale descriptors and models will be reviewed, including fractals, local variance, spectral analysis, and wavelets, with spatial statistics (i.e., geostatistics and spatial autocorrelation analysis) being introduced as the framework for scale modeling in the second and third subsections. Such scale descriptors are useful for determination of appropriate scales for data collection, information extraction, and spatial analysis.

**Various scale descriptors**

Many linear features with complex detail (e.g., coastlines) are said to be statistically *self-similar*, because they appear to possess the same statistical properties over a range of scales. Statistical self-similarity is a well-known manifestation of fractals (Mandelbrot 1967), which have long been studied for their applicability in geography (Lam and Quattrochi 1992). For a curve that is not recursively generated, its fractal dimension can be estimated by measuring the length of the curve using a series of step sizes and performing a linear regression. Suppose the measured lengths \( L(\Delta) \) of such geographic curves are related to the measurement scale \( \Delta \) via a function of the form \( L(\Delta) = k\Delta^{1-D_{fractal}} \). We can see the linearity of a plot of the quantity \( L(\Delta) \)
versus scale $\Delta$ on a log-log graph: $\log L(\Delta) = \log k + (1 - D_{fractal}) \log \Delta$, where the slope of the regression line is $1 - D_{fractal}$, while $D_{fractal}$ indicates the fractal dimension (Mandelbrot 1967). A relatively smooth (irregular) curve should have a fractal dimension close to 1 (2). Fractal analysis can be performed not only with vector data but also raster data, such as digital elevation models (DEMs) and remotely sensed images.

For analysis of scales in remote sensing images, Woodcock and Strahler (1987) propose the local variance method, which can be used for quantifying the relationship between the size of objects in a scene and the spatial resolution of sensors, and thus useful for determining a suitable measurement scale for surveys. By this method, the mean value of the variance of a moving window (say 3 by 3 pixels) over the entire image is computed and plotted as a function of spatial resolution. Generally, the peak appears when the pixel size (i.e., spatial resolution) approximately matches the size of the objects, with multiple local-variance peaks indicating multiple scales of variation in the scene (Woodcock and Strahler 1987; Zhang, Atkinson, and Goodchild 2014). It is important to note that interpretability of scale in the underlying scene based on local variance is dependent on the image’s global variance. The local variance method has been used widely for multiscale analysis in various applications.

There are also spectral analysis methods that may be explored for analysis of scale. By (discrete) Fourier transforms, a signal (e.g., a terrain profile (1-D) or an imaged scene (2-D)) can be transformed into its equivalent in frequency space, and expressed as a linear combination of oscillating functions with different frequencies via inverse transforms. Fourier transform can be used to construct a Fourier spectrum indicating magnitudes and phases of the components at different frequencies in a signal. Pike and Rozema (1975) derived the spectral density function via the discrete Fourier transform of an autocovariance function estimated for a terrain profile, with the resulting spectrum depicting the functional relation between the variance associated with different-sized terrain undulations and topographic wavelengths.

As reviewed by Dale and Mah (1998), the wavelet transform method uses a localized function in time or space (to adjust and shift the wavelet’s size and location) for signal analysis at multiple resolutions. The wavelet variance can be calculated on the basis of the wavelet transforms and used as an indicator to quantify the spatial scale and as a suitable tool to study multiscale characteristics of spatial data and the underlying processes.

Theoretically, scale is related to the structure, extent, and strength of spatial dependence inherent in many spatial processes and patterns. A statistical measure of the similarity of properties over space is called spatial autocorrelation, as reviewed by Goodchild (1986). In the next two subsections, geostatistics and spatial autocorrelation are discussed for scale modeling in geostatistical and lattice data, respectively.

Geostatistical models of scale

Geostatistics is the mathematics for quantifying and utilizing spatial dependence in spatial analysis based on the theory of regionalized variables, which states that the spatial variation for $Z$ at $x$ can be viewed as a regionalized variable $Z(x)$, and is the sum of a deterministic mean, a spatially correlated residual, and uncorrelated noise (Journel and Huijbregts 1978; Webster and Oliver 2007). A related concept is the so-called random field (RF), which refers to a set of regionalized variables $\{Z(x), x \in D\}$ over a problem domain $D$ (which is discretized by $x$ as in the notion $\{Z(x), x \in D\}$). In geostatistics,
covariance functions and variograms are used as tools to quantify spatial dependence, in terms of correlation and dissimilarity, respectively, in regionalized variables, say $Z$, which exhibit both deterministic and stochastic characteristics as mentioned above. For a RF $Z$, its variogram is defined as half of the variance of the difference between $Z$ values at two locations, $x$ and $x+h$, which are separated by lag $h$ (the distance and direction of separation), in the domain $D$ (under the intrinsic hypothesis):

$$
\gamma(h) = \frac{1}{2}E[(Z(x+h) - Z(x))^2]
$$

A variogram model is shown in Figure 1.

There are three named parameters of a typical variogram model: nugget, sill, and range, which are represented by $c_0$, $c_0 + c_1$, and $a$, respectively, (Figure 1). These parameters can be used to characterize spatial dependence and, based on this, the scale. The nugget is the discontinuity of the variogram at the origin, due to uncorrelated noise (measurement error) and/or spatial structures operating at scales finer than the spacing of observations. The sill (if one exists, as in the case of bounded variability) is the value that the variogram reaches (asymptotically sometimes) as the lag distance increases, equaling the variance of the underlying RF. The range (if one exists, corresponding to the existence of a sill) is the lag distance at which the variogram reaches a sill, indicating the process scale over the domain, as also described by Zhang, Atkinson, and Goodchild (2014).

The scale model (e.g., a variogram model) of a geographic distribution is usually derived from sample data. As these sample data are often of finite support themselves, the resultant experimental or sample variograms are, thus, regularized versions of the underlying point-support variograms. These point-support variogram models are prerequisite for calculation of regularized versions over any finite supports and cross-scale data analyses. For this, variogram deconvolution or deregularization, as an inverse of variogram regularization, is required to estimate point-support variograms from finite-support variograms (Goovaerts 2008).

In addition to spatial dependence (as a general pattern), spatial heterogeneity is also characteristic of geographic distributions and entities. Spatial heterogeneity implies nonstationarity in spatial structure of an underlying RF (Sampson and Guttorp 1992), necessitating use of local models of scale to accommodate nonstationary structural pattern in the RF. For local modeling of scales, first-order (i.e., mean) and second-order (covariance) moments can be dealt with separately or in combination. Nonstationarity of the mean may be handled by the intrinsic assumption, by decomposing the RF into a local mean and a residual, or by partitioning a domain into homogeneous subdomains and modeling locally varying means within these subdomains. To handle local variation in other statistics (e.g., variance, spatial covariance, and spatial cross-covariance), it is sensible to adopt such a partitioning and local modeling strategy. It
amounts to restricting the stationarity in second order moments of the RF locally, thus resulting in local variogram models, for instance. There is another approach to nonstationary structural modeling, which works directly over the entire problem domain without partitioning it at the first place. An example for this approach is a spatial deformation approach (Sampson and Guttorp 1992): the coordinate space is transformed so that the covariance structure becomes stationary and isotropic in the transformed space.

In principle, many of the methods for variogram modeling can be applied, with proper modification, to categorical variables. This is often done through indicator transforms, resulting in indicator auto- and cross-variograms for quantifying spatial dependence in area-class occurrences and distributions. Methods for modeling spatial structure in area classes can be usefully addressed on the basis of phase space (Goodchild, Zhang, and Kyriakidis 2009), which is characterized by the process covariates that are believed to have shaped the area-class distribution observed. Also, to analyze multiple RFs involved in a problem domain, it is important to extend models of univariate regionalization to those of co-regionalization.

Lattice data and scale models

As reviewed above, variograms are versatile in being able to characterize RFs of both continuous and categorical variables, and flexible for global and local quantification of scale. In the literature, geostatistical modeling has implicitly assumed continuous indexing of locations over space. In addition to this kind of “geostatistical data,” there is another type of data known as lattice data, that is, data that are aggregated over finite areal units. For this kind of data, alternative modeling approaches are necessary and complement the geostatistical ones, as explained below.

In terms of data support, spatial data sampled from fields (of both continuous and categorical variables) can be called geostatistical data (over point or quasi-point support) and lattice data (over finite areal support). Geostatistical data result from sampling a field over a continuous domain $D$, where sample locations can be continuously indexed. A variogram quantifies the degree and range of spatial dependence in a particular dataset (and the underlying RF, say $\{Z(x), x \in D\}$ if the variogram is point–supported) and can be used as a model of scale, as reviewed in the preceding subsection. Unlike geostatistical data, lattice data result from aggregation of RF $Z$ over finite areal units (with prescribed tessellation schemes). This subsection discusses lattice data and their scale models, based on measures of spatial autocorrelation, for which neighborhood relationships (Goodchild 1986) play a central role in data analysis and scale characterization.

For lattice data, the number of lattice units imposed on a domain is countable (and usually finite), as the lattice units are themselves of finite areal extents. Census data and forest inventory over different mapping units are examples of lattice data (see Saveliev, Mukharamova, and Zuur (2007) for further examples). Lattice data can be regular or irregular in terms of tessellation (as exemplified by remotely sensed images and county-level cropland inventories, respectively), discrete or continuous in terms of scale of measurement (e.g., land use and population data), as in the case of geostatistical data.

Spatial autocorrelation in lattice data provides the basis for describing their scales. A key element in measuring and understanding spatial autocorrelation is to describe the relationship between the similarity of values of a field variable and the distance separating them, as described by Goodchild (1986). Quantification of spatial
autocorrelation in lattice data is usually based on Moran’s $I$ and Geary’s $c$, which can be used to create correlograms as models of spatial autocorrelation. In lattice data analysis and scale modeling, neighborhood relationships (spatial adjacency) rather than Euclidean distance alone are important for the prediction of relevant quantities over space.

As for geostatistical data, local measures of spatial autocorrelation, such as those proposed by Anselin (1995), may be explored for lattice data, resulting in local models of spatial autocorrelation and those of scale. Such local models are more useful and valuable than global ones in being able to capture spatial variability (i.e., nonstationarity) of scales over the extent of a problem domain $D$. Local indicators of spatial autocorrelation can be built upon both Moran’s $I$ and Geary’s $c$ (Anselin 1995), to accommodate spatial heterogeneity in the strength of spatial autocorrelation in $D$, and thus constructing local models of scales for lattice data.

Lattice data can be sensibly classified in terms of their suitability for geostatistical descriptions. There are lattice data, such as remotely sensed data (regularly shaped) and areal data sampled or aggregated over naturally regionalized zones (with bona fide boundaries). Such data can be analyzed with respect to spatial dependence by using geostatistics through use of finite blocks (as opposed to points), and are called Type I lattice data. On the other hand, lattice data not directly amenable to geostatistical handling are called Type II lattice data, such as those collected over arbitrarily demarcated reporting zones (with fiat boundaries). This is because, in Type II lattice data, lags over which spatial autocorrelation is quantified are not based on the Euclidean metric but on spatial adjacency, which is often not conformal to Euclidean distance. Therefore, for Type II lattice data, spatial statistics, such as Moran’s $I$ and Geary’s $c$ indices, become particularly useful for analyzing the spatial autocorrelation and spatial scale inherent in the underlying fields, as discussed also by Zhang, Atkinson, and Goodchild (2014).

In the light of the classification of lattice data above, regularization and its inverse, deregularization, of spatial statistics relevant to scale modeling can be discussed accordingly. As mentioned in the previous subsection, variogram regularization is the process of quantifying variograms over finite support from point-support ones, while deregularization is the inverse process. For Type I lattice data, because their scale description can be based on geostatistics as reasoned above, methods for regularization and deregularization of variograms can be applied straightforwardly. For Type II lattice data, regularization of spatial statistics should be based on proper weighting of spatial weight matrices in the aggregation of the underlying $Z$ fields, leading to regularized correlograms. However, the process of deregularization is far more complicated due to the difficulty in deconvoluting lower-order spatial adjacency from its higher-order version (Zhang, Atkinson, and Goodchild 2014), requiring further research in this regard.

Both global and local measures of spatial autocorrelation have been used for describing spatial variability of lattice-sampled fields. Proper use of correlograms provides valuable information about the ranges and strengths of spatial autocorrelation in the underlying fields and can thus be used to elucidate models of scale concerning the lattice processes and the data drawn from them.

### Scaling up and down

As already mentioned, there exists scale discrepancy in multisourced spatial data, geoprocessing algorithms, spatial models, and applications, necessitating change of scale to enable cross-scale information communication. In remote sensing,
for instance, it is often necessary to upscale finer resolution data (e.g., RapidEye image-based reference data with 5 m by 5 m support) to coarser resolution data (e.g., Landsat TM images with 30 m by 30 m pixels) for comparison or validation purposes.

Different methods have been developed to transfer data and model outputs from a finer to a coarser scale (upscaling or aggregation) or from a coarser to a finer scale (downscaling or disaggregation). Fractal methods may be used to extend the statistical properties (i.e., moments) of a variable from one scale to another. Area-based weighting (e.g., for deriving coarse-resolution data from fine-resolution data based on the area extents of different land cover types within the coarse-resolution pixels) is based on the law of conservation of energy or matter, and often applied for scale changes in remote sensing.

Variograms and autocorrelation functions are designed for modeling scales in geostatistical and lattice data, respectively, as discussed in the preceding section. Accordingly, the next two subsections will describe methods for scaling of geostatistical data and lattice data, respectively. For the latter, the focus will be on regularly gridded regions, such as various images or raster data.

Geostatistical methods for scaling

We consider upscaling first. Upscaling is increasing the size of data support or coarsening the corresponding spatial resolution. The choice of method to use for upscaling depends on the characteristics of the finer resolution data. When the smaller-support data are sparsely (and usually irregularly) distributed, certain interpolation approaches may be performed to predict the values of a variable $Z$ over larger supports (denoted $Z_v$) within which the data of smaller support fall. For this, block kriging may be applied, because it is optimal in the sense of minimum prediction variance and unbiasedness (relative to the spatial statistical models assumed for the underlying RF) and because it enables prediction over the larger supports automatically as a result of linear kriging equation systems. In addition to kriging prediction itself, location-specific kriging variance can be obtained, providing useful information on prediction uncertainty (Journel and Huijbregts 1978). When finer resolution data are on regular supports (i.e., regular lattice data, such as remote sensing images) and abundant, upscaling may be performed by: (i) averaging approaches (i.e., averaging finer resolution data falling within coarser resolution cells); (ii) other statistics, such as median, minimum, and maximum, as averaging is equivalent to taking mean values; or (iii) some more sophisticated procedures that handle sensor properties (e.g., point-spread functions) (Justice et al. 1989). This is further discussed in the context of gridded lattice data in the next subsection.

For downscaling (i.e., to refine the spatial resolution of an existing dataset), block-to-point kriging (or area-to-point kriging in 2-D cases where source data and locations being predicted are defined on area and point supports, respectively) may be employed (Kyriakidis 2004), providing a potential solution to downscaling in both geostatistical and lattice data. It is helpful to conceptualize downscaling as larger blocks-to-smaller blocks interpolation, because the source data and prediction locations are often defined on finite supports (larger and smaller, respectively). Thus, area-to-point kriging mentioned above can be seen as a special case of larger blocks-to-smaller blocks interpolation.

Geostatistical inverse modeling is another option for downscaling (Kitanidis and Vomvoris 1983). As a Bayesian approach, it is based on the principle of combining prior information with
information latent in measurements (or observational data), where the prior information concerns spatial structure (i.e., mean and spatial (or temporal) correlation) prescribed for an underlying distribution say $Z$ (Kitanidis and Vomvoris 1983). Clearly, geostatistical inverse modeling provides links to the data-generation processes, and will be a useful extension to kriging-based downscaling mentioned previously.

Methods for scaling gridded lattice data

As discussed previously, geostatistical scaling is particularly useful where the data to be scaled are sparse with respect to the density of locations to be predicted, because it makes use of spatial structural information in addition to the measurements themselves for accuracy in scaling. Besides, uncertainty related to spatial interpolation (downscaling) and averaging (upsampling) can be quantified and used as valuable information for balancing between accuracy and resolution in down- or upscaled data products. This subsection addresses the issue of scaling of lattice data, in particular, those of regular supports.

Consider upscaling first. Although the geostatistical methods can be employed for scaling lattice data, in particular, Type I lattice data that include regularly gridded data and areal data partitioned with bona fide boundaries, other methods may be explored for scaling gridded data (i.e., images). For upscaling, statistical approaches can be used to aggregate fine-grained data into coarse-grained data through certain statistical operators, such as mean, median, mode, maximum, and minimum; mode can be used for aggregating categorical data. In addition to statistical aggregation mentioned above that proceeds as a kind of sampling from smaller-support cell values within the bounds of larger-support cells, there are mechanistic approaches that are based on the imaging mechanisms pertaining to image acquisition and thus better suited for coarser-resolution image or sensor simulations. For this, convolution kernels, based on proper parameterization of sensor point spread functions (Justice et al. 1989), need to be prescribed.

For downscaling, deconvolution may be used. Deconvolution aims to reverse the effects of convolution on data (e.g., blurring and noise in imaging), although it may or may not refine the spatial resolution of the input datasets. For resolution refinement, super-resolution mapping may be explored through various techniques, such as the combined use of multiple observed images of coarser spatial resolution. Mechanistic downscaling of gridded data may be developed from the geostatistical inverse modeling approaches described previously. Both statistical and mechanistic approaches to downscaling of gridded data are useful and should be explored in combination.

It can be asserted that further developments are likely to be seen in the integration of sensor/imaging models, statistical image analysis, and geostatistics. Thus, on one hand, the physical models of imaging and sensor parameters can be used for geostatistical scaling of gridded data, as in geostatistical inverse modeling. On the other hand, the characteristics of image data (and other gridded data) should be utilized in geostatistical scaling and to enhance the efficiency of computing with gridded data (Zhang, Atkinson, and Goodchild 2014).

As shown above, for Type I lattice data, geostatistical scaling (both up and down) may be implemented. For Type II lattice data, however, methods oriented to geostatistical data and gridded data are seldom applicable in a straightforward manner. There may be merits in enriching the semantics of Type II lattice data by use of covariates that are regionalized and correlated with the variable being studied (Zhang,
Atkinson, and Goodchild 2014). Hierarchical models may also be used for scaling Type II lattice data, especially those hierarchically structured.

Process-pattern strategies and information-theoretic perspectives

This section highlights the importance of process-pattern strategies for addressing scale issues not only in categorical fields but also in objects. The process-pattern strategies will be shown as useful extensions of spatial statistical approaches to modeling scale and scaling. This is followed by some discussions about information-theoretic perspectives in spatial data, analysis, and scale-related research.

Undoubtedly, geostatistical and spatial autocorrelation models are valuable tools for characterizing scales. Local models can accommodate nonstationarity in spatial structures and, hence, scales. Cross-variable correlation should also be incorporated in scale modeling, in particular, that concerning multivariate distributions. In addition to these general modeling approaches, alternatives that are suited specifically for certain geographic distributions are worth exploring. For example, representative elementary areas (reviewed by Blöschl and Sivapalan (1995)) and various ruggedness indices, which are indicative of spatial scale, can be used for modeling scales in terrain variables.

The process-pattern paradigms are valuable for investigating scales in area classes. Area-class maps depict distributions of land cover, soil types, ecoregions (Bailey 2014), and other thematic classes. For modeling scales, there are various landscape spatial pattern indices (Gustafson 1998), which can be seen as a manifestation of the scales of the underlying controlling factors (Bailey 2014). Conventional methods for scaling area-class maps are statistically based: for upscaling, statistical operators, such as majority rule, averaging, and random selection, may be used; for downscaling, the methods include estimating, at fine resolution, areal extents of individual classes based on coarse-resolution data (either discrete classification or class proportions) and fine-resolution prediction of class labels from coarse-resolution data and covariates. Discriminant space (a renaming of phase space) (Goodchild, Zhang, and Kyriakidis 2009) was proposed as a conceptual framework for area-class mapping. It provides the basis to construct a process-pattern (environmental-species) strategy, which should be exploited for both upscaling and downscaling of area-class information: the scaling of categorical information is accomplished through scaling of class-controlling covariates, which span the discriminant space whereby area classes are defined.

The process-based strategy can be sensibly extended to investigate how scales in objects (in particular, the bona fide type) should be described and how their scaling may be more effectively implemented than otherwise. For instance, some terrain derivatives, such as flow paths, can be studied with respect to scale and scaling in the light of the process-pattern paradigm elaborated here.

As discussed above, the process view is valuable as to how terrain properties are derived from elevation data, how land cover information is extracted from class-controlling factors or covariates, and how objects are delineated from fields. The integration of spatial statistics and process models then provides a coherent framework for scale modeling and scale change.

Information theory offers yet another language for addressing scale-related issues. Entropy is a measure of the uncertainty associated with a random variable (RV), and quantifies the amount of information (number of bits) required on average to describe the RV. Let $S$ be a discrete RV
with probability mass function \( p(s) = \Pr\{S = s\} \), 
\( x \in \mathbb{S} \), \( \mathbb{S} \) being the set of states \( S \) can take. The entropy \( H(S) \) of \( S \) is defined by:

\[
H(S) = E[I(s)] = E \left[ \log_2 \frac{1}{p(s)} \right] = -\sum_{i=1}^{n} p(s_i) \log_2 p(s_i)
\]  

(2)

where \( I(s) \) denotes the entropy contribution of an individual message \( s \) (also known as the self-information of \( s \)), and \( E \) indicates expectation, with \( n \) being the number of states for \( S \) in the set \( \mathbb{S} \) (Cover and Thomas 2006). We also define the conditional entropy of a RV, say \( S \), given another, say \( Z \), as the expected value of the entropies of the conditional distributions: \( H(S|Z) \). Joint entropy \( H(S, Z) \) can also defined for a pair of discrete RVs \( (S, Z) \) with a joint distribution \( p(s, z) \). Joint entropy can be seen as an extension to univariate entropy and can be extended to modeling the entropy, measured by the so-called entropy rate, of sequences of RVs (i.e., stochastic processes). Mutual information is a measure of the amount of information that one RV contains about another, and can quantify the actual amount of information conveyed by a measured or estimated value (say spot height) about the true value. The mutual information between two RVs, say \( S \) and \( Z \), is often denoted as \( I(S, Z) \), and can be calculated as the reduction in the uncertainty of \( S \) due to the knowledge of \( Z \), i.e., \( I(S, Z) = H(S) - H(S|Z) \) (Cover and Thomas 2006).

Consider an image \( \text{IMG} \) of \( m \times n \) pixels, with each taking on one of the \( K \) discrete values (indicating the image’s radiometric resolution). If we view \( \text{IMG} \) as a noiseless communication channel, there will be \( K^{mn} \) distinguishable images that can be encoded and transmitted, with the amount of information conveyed by \( \text{IMG} \) being \( \log K^{mn} = mn \log K \). It is easy to see that, for an image acquired over a given study area, the finer its spatial resolution, the more pixels there will be in it, thus increasing its information capacity. However, inter-pixel spatial dependence in the image suggests that there exist informational redundancy among its \( m \times n \) pixels \( Z = \{Z_1, Z_2, \ldots, Z_{mn}\} \) and that the relationships between information content and resolution are rarely linear or simple (Zhang, Atkinson, and Goodchild 2014). Spatial statistics will be useful for accounting for effect of spatial dependence on entropy, say through entropic modeling of coarsened or refined pixel sequences \( H(Z_1', Z_2', \ldots) \). We can also examine the effect of image spatial resolution upon the mutual information conveyed by pixel measurements about the underlying spatial distributions.

Information content in images and maps should be quantified as part of performance analysis and map design. This may be done by taking an image as a whole or, more sensibly, locally, just as statistical quantities related to scale are often modeled locally due to nonstationarity in spatial distributions and their spatial dependence strengths and ranges. This indicates that spatial variation of information granularity, as often observed, should be accommodated in representations. We can accomplish this by using differently sized rasters depending on the density of local information in the dataset. Hierarchical data models are valuable constructs for geographic representations adaptive to local variability (information amount) in the underlying forms and processes.

As hinted at previously, spatial dependence will have complex effects upon the informational aspects of the processes underlying image acquisition and analysis. As process scale is inherently related to the characteristics of spatial dependence, it is possible to look into the relations between informational and scale indices. Journel and Deutsch (1993) described
a measure of spatial entropy based on bivariate probabilities quantified over lags. It is basically the joint entropy associated with a set of joint probabilities corresponding to pairs of intervals of values discretizing the joint distribution of the underlying RF $Z$ at two locations separated by a lag $h$. This bivariate joint entropy can be standardized as relative entropy, which ranges between 0 and 1. When a particular class of $z$ values persists over larger (shorter) distance $h$, relative entropy will be increasing at a slower (faster) rate, implying a stronger (weaker) spatial order for prediction at unsampled locations (Journel and Deutsch 1993). The similarity between standardized variograms (to sills) and relative spatial entropy suggests that we may well be able to use informational measures to model and interpret scale in spatial processes and forms.

Despite ever-increasing computing speed and storage capacity, data compression remains important for efficiency in spatial information gathering and processing, especially in this so-called big data era. Research on information granularity (not merely data amount) and its scale dependence suggests that optimality in spatial sampling and geoprocessing can be achieved in the light of (classic) information theory and, more importantly, the recent developments in compressed sensing (or compressive sampling, shortly known as CS) (Candès, Romberg, and Tao 2006). CS capitalizes on the sparsity of the signals being sensed and utilizes the informational efficiency of CS sampling matrices and signal reconstruction based on convex optimization, which provides the solution to an underdetermined system of linear equations even if it is sampled at a rate lower than the Nyquist sampling rate (Candès, Romberg, and Tao 2006). Information theory will be useful for exploring scene sparsity and determining optimal sampling rates in CS applications.

Prospects

This section summarizes methodology regarding scale and scaling. For this, some prospective views are offered on how scale-related research may be integrated in data collection (e.g., spatial sampling), data conflation, and uncertainty characterization.

Measurement scales are determined by the sampling framework. The sampling framework can itself be divided into the spatial coverage of the sample (defined by the sampling scheme, sampling density, and sample size) and support of each individual observation (Skøien and Blöschl 2006). As sampling can be uniform or non-uniform, the discussion about sampling density can be organized accordingly. For uniform sampling, the Shannon sampling theorem states that a sampling rate twice the maximum frequency of a band-limited signal (i.e., a spatial distribution and entity) is sufficient for exact recovery of the signal being sampled. The optimum sampling density for remote sensing (typically seen as a uniform sampling technique) should be determined based on the spectra of the underlying signals, although this is seldom pursued in practice. For spatial sampling undertaken at irregular nodes or on non-uniform grids, the literature on non-uniform sampling is growing and may shed light on how optimality in spatial sampling should be defined and what are the major factors in sampling implementation. For the classic sampling (either uniform or not) theorems, infinite sample support in space and time is assumed. However, spatial variation in sampling density and local definition of effective data supports are commonly encountered in practice. Further research must account for the interactions between sampling rate (density) and finite data support in determining the optimum sampling protocols for spatial applications. Moreover, determination of measurement scale over a
nonstationary domain is challenging. Obviously, there will be no single optimal measurement scale for such a domain, but measurement scales can be made to adapt to the nonstationarity in process scale therein. Webster and Oliver (2007) discussed how to design a network of sample sites (data support) so that multiple scales in the underlying RF are captured in a variogram. As mentioned in the preceding section, the CS paradigm should also be explored for optimum accuracy and efficiency in spatial sampling and mapping.

Sophistication of spatial information technology has led to spatial data proliferation. Remote sensors are providing unprecedented amounts of geographic data at increasingly fine spatial, spectral, and temporal resolutions, while various surveys are collecting data at finer spatial detail. However, despite these, our ability to maximize information potentials from multisource data lags behind, as it is hardly easy to analyze and integrate multisource data (often implying different accuracies, resolutions, semantics, or classification schemes). Techniques for data conflation must be developed that can handle scale mismatch and semantic inhomogeneity in multisource geospatial data effectively and efficiently, facilitating information fusion. Conflation of heterogeneous spatial information sources will also benefit from information-theoretic perspectives for modeling and assessing the sources and the fusion products.

Uncertainty is inherent in spatial information and analysis, and should thus be characterized properly. Components of uncertainty-related research include error description (accuracy assessment), error propagation, and management, with the former two being the focus here. Error description can take advantage of validation procedures, which are becoming increasingly integrated in operational remote sensing and spatial information production. It is also important to assess the impacts of errors upon spatial analysis and applications. This is known as error propagation. While analytical approaches based on variance and covariance propagation implicitly assume homogenized scale for model inputs and, if required, model parameters, geostatistical simulation is more robust to scale discrepancies, and should be explored in any scale-dependent modeling of uncertainty and applications. In addition to the conventional methods in which realizations are generated over grid nodes of supports conformal to those of the conditional data (implicitly, points or quasi-points), cross-scale simulation methods are particularly useful in the context of multiscale data and applications, as translating information about error margins is important.

Acknowledgment

The work underlying this entry is supported by National Natural Science Foundation of China (grant no. 41471375).

SEE ALSO: Fractal analysis; Geographic information science; Geostatistics; Interpolation: areal; Interpolation: kriging; Representation; Spatial resolution; Uncertainty

References


Further reading


Scarcity poses a paradox. If, according to the Merriam-Webster Dictionary, scarcity means “the quality or state of being scarce; especially: want of provisions for the support of life,” then presumably actual scarcity has grown ever less common over the past 200 years, as industrial capitalism has expanded global supplies of food and other material goods to unprecedented levels in both absolute and per capita terms. But, during the same period, the idea of scarcity has grown increasingly pervasive in both obvious and obscure ways. Modern economics takes scarcity as its fundamental theoretical premise and raison d'être, and major strands of environmentalism likewise find motivation in the idea that natural resources are not only finite but also increasingly scarce. The prospect of imminent scarcities of food, energy, water, or other key resources has been a leitmotif of modernity, and chronic economic inequality and poverty have left significant numbers of people lacking basic necessities, apparent victims of scarcity. It is as though society cannot produce abundance without constantly thinking about—and in some places producing—its opposite.

Although the English word dates back to the fourteenth century, scarcity’s prominence as a modern concept coincides with the ascendance of market economies in northwestern Europe in the latter half of the 1700s. Scarcity was associated first and foremost with food, and especially with the political threats that high food prices could pose to the old and declining feudal regimes of the time. Occasional food shortages and famines were, of course, much older phenomena; what was new was the extensive mediation of the food supply by intra- and interstate commerce and the dynamics of supply and demand that this entailed. High prices could no longer so easily be blamed on climatic vagaries or divine misfortune if hoarding, speculation, monopoly, and trade were potential alternative culprits.

In the 10 years beginning with 1795, a series of spikes in the price of grains in United Kingdom prompted a flurry of polemical speeches, pamphlets, and treatises dedicated expressly to the topic of scarcity. Some viewed the deficiency of grain as a natural result of unfavorable weather, basing their arguments on personal observations of harvests in various parts of the kingdom. Others blamed farming methods and the quality of various soils, or referred to data on imports and exports and pointed to disruptions in interstate trade caused by the Napoleonic Wars. But, in the absence of comprehensive agricultural statistics—both for the United Kingdom and for countries whence grain was imported—actual supplies of food were difficult to judge, and the causes of price increases were subject to speculation and debate. Merchants and middlemen were widely accused of hoarding, monopoly, and speculation to create the appearance of (and to profit from) scarcity, regardless of real supplies. Some extended this criticism to landowners who were consolidating their holdings into fewer and larger farms, which critics viewed as more prone to monopoly and more vulnerable to unfavorable weather than smaller farms.

While nominally about scarce food supplies, the debates were more fundamentally concerned...
SCARCITY

with how government should relate to the market economy. Two other pressing issues lent great urgency to this question: the specter of the French Revolution and the emergence of widespread poverty among commoners, many of them recently displaced into urban centers by the Enclosure Acts. In a time of growing national and imperial prosperity, the masses of impoverished laborers were a paradoxical and ominous novelty, and elites feared that high prices – whatever their cause – might result in riots, rebellion, or revolution. In Karl Polanyi’s (1957/1944, 111) words, “When the significance of poverty was realized, the stage was set for the nineteenth century.”

Adam Smith’s Wealth of Nations provided the intellectual ammunition for those, such as Edmund Burke, who argued that free trade and private property were the best – indeed, the only – way to solve imbalances between producers and consumers of food. “Labour is a commodity like every other, and rises or falls according to the demand. This is in the nature of things,” Burke (1800/1795, 6) insisted. Middlemen “are to be left to their free course; and the more they make, and the richer they are, and the more largely they deal, the better for both the farmer and the consumer” (Burke 1800/1795, 24). By this logic, government measures to cap prices or ban hoarding would only make the problem worse by preventing market signals from operating effectively. “Of all things, an indiscreet tampering with the trade of provisions is the most dangerous, and it is always worst in the time when men are most disposed to it: – that is, in the time of scarcity” (Burke 1800/1795, 1).

Thomas Malthus weighed in with An Investigation of the Cause of the Present High Price of Provisions (1800), echoing many of Burke’s positions in defense of landowners and middlemen and in opposition to government intervention. Malthus saw bad weather as responsible for short-term declines in agricultural output, but he argued that the price increases were out of proportion to these shortfalls. Monopoly was not the cause of this disparity, however: in Malthus’s view, the market for food was too large and had too many players for anyone to monopolize it. Rather, he diagnosed the problem as stemming from the Poor Laws, which since 1795 had guaranteed supplements to wages at levels pegged to the price of bread (Polanyi 1957):

the attempt in most parts of the kingdom to increase the parish allowances in proportion to the price of corn … is, comparatively speaking, the sole cause, which has occasioned the price of provisions to rise so much higher than the degree of scarcity would seem to warrant. (Malthus 1800, 4–5)

Giving the poor money simply added to the effective demand for food without increasing supply, further raising prices, even if it did also help diminish suffering by spreading the impact across a larger number of people. The only real solutions were increased production and imports, both of which would come about in response to higher prices; in the shorter term, large farmers and middlemen were to be thanked for raising “the corn to that price which excluded a sufficient number from their usual consumption, to enable the supply to last throughout the year” (Malthus 1800, 15).

At the end of his pamphlet, Malthus shifted the question from food prices to the growing dependence of United Kingdom on food imports. Asserting (without evidence) that domestic agricultural production could not have “gone backwards” in the preceding 20 years, he deduced that “the present inability of the country to support its inhabitants” could only be due “to the increase of population” (Malthus 1800, 27). He referred readers to his Essay on the Principle of Population, originally published in 1798, which he was in the process of almost completely rewriting for the more famous second edition of 1803.
Two points about Malthus’s *Essay* are key to understanding its outsized influence on subsequent thinking about scarcity. The first is that it was fundamentally a moral argument about the nature of good and evil, virtue and vice – what LeMahieu (1979) termed a “theology of scarcity.” Although Malthus dampened his theological claims in the second (and subsequent) editions, he continued to view humans as naturally “inert, sluggish, and averse from labour” (Malthus 2004/1798, 115–116), and he considered moral and physical deprivation to be necessary, divinely sanctioned spurs to virtuous activity. High food prices were therefore actually a good thing, or at least served as a kind of medicine whose effectiveness required an unregulated market economy:

There are some disorders, which, though they scarcely admit of a cure, or even of any considerable mitigation, are still capable of being made greatly worse. In such misfortunes it is of great importance to know the desperate nature of the disease. The next step to the alleviation of pain, is the bearing it with composure, and not aggravating it by impatience and irritation. (Malthus 1800, 1)

LeMahieu (1979, 474) concludes: “Here at last was the Protestant ethic and the spirit of capitalism.”

The *Essay*’s second key contribution was interpreting scarcity as ultimately a function of human population – this was the underlying “disease.” Malthus defined his principle in terms of food: “the power of population is indefinitely greater than the power of the Earth to produce subsistence for man” (Malthus 2004/1798, 19). But neither the reductive circularity of his logic (Engels famously pointed out that even two people would constitute overpopulation by Malthus’s argument), nor the empirical fact that food production has indeed increased more rapidly than population since he wrote, has prevented Malthus’s followers from seeing virtually any environmental or natural resources-related problem as a function of population. Neo-Malthusianism can be understood as a body of thought that ignores the moralism and political context of Malthus’s arguments in order to take – and mistake – his principle of population as a “scientific” insight based on “natural” laws.

By the end of the nineteenth century, the meaning of scarcity had shifted from anomalous and transient episodes of shortage (people spoke of “a scarcity” as a period of time) to a normal, universal condition of human life and activity. The leading proponents of this shift were economists such as William Stanley Jevons, Leon Walras, and Carl Menger, who developed marginal utility theory on the basis of “what has come to be known as the scarcity postulate, an assumption of the universality of the condition of scarcity that at once gives neoclassical economics its focus and provides the legitimacy of its claim to science” (Xenos 1989, 68). (In his study *The Coal Question*, Jevons also developed the famous paradox that bears his name: When per unit costs are reduced, gains in the efficiency of use of a resource will increase, not reduce, its overall use. Scarcity thus becomes self-reinforcing.) Lionel Robbins (1984/1932, 15) summarized the postulate succinctly: “Scarcity of means to satisfy ends of varying importance is an almost ubiquitous condition of human behaviour. Here, then, is the unity of subject of Economic Science, the forms assumed by human behaviour in disposing of scarce means.”

In contrast to the political economy of Smith, Marx, and Mill, neoclassical economics was unconcerned with wealth and value. The new science understood human needs as “constructed solely out of the individual’s preferences, without any trace of social determination” (Xenos 1989, 70), expressed through individual calculations
of economizing self-interest. In one sense this reduced economics to the narrow world of commodity exchange, with scarcity as its fundamental premise. Anything that was abundantly available to all was not worth economizing, and therefore irrelevant to the discipline:

Economics is the science which studies human behaviour as a relationship between ends and scarce means which have alternative uses … It does not attempt to pick out certain kinds of behaviour, but focuses attention on a particular aspect of behaviour, the form imposed by the influence of scarcity. (Robbins 1984/1932, 16–17)

But this formalism also contained the potential for almost infinite expansion as scarcity came to characterize – or could be understood to characterize – a proliferating diversity of phenomena. “Economics is entirely neutral between ends … in so far as the achievement of any end is dependent on scarce means, it is germane to the preoccupations of the economist” (Robbins 1984/1932, 24; emphasis in the original). Even exchange was “subsidiary to the main fact of scarcity” (Robbins 1984/1932, 20): an individual’s time, for example, could be economized in choices about how to “spend” it, provided time were made or perceived to be scarce. Modern macroeconomics also rests on the scarcity postulate, applied at the scale of aggregate output, which is constrained by diminishing returns to capital and labor (Cobb and Douglas 1928; Solow 1956).

Scarcity in neoclassical economics naturalizes more than just a particular model of rationality: “It also universalizes a particular set of institutions – property and markets – which are deemed to be natural results of scarcity” (Xenos 1989, 72). Focusing on exchange-value (price) to the exclusion of use-value affords neoclassical economics the apparent objectivity of quantitative methods based on the universal abstraction of Homo economicus, but this very tractability presupposes that individuals always already appropriate property through markets. The question that preoccupied Burke and Malthus – the extent and legitimacy of a free market, with minimal government interference – vanishes by conceptual fiat under the postulated universal condition of scarcity.

The realization and extension of markets – and therefore the applicability of economic science understood in this way – is of course a historically and geographically specific process attended by contentious political struggles. Markets are not pregiven by nature but must be produced, and the state necessarily plays a major role in that production. It is in this role that Michel Foucault (2007) identified the origins of modern governmental power – government based on an “apparatus of security” rather than on a juridical-disciplinary system – which he traced to debates about food scarcity in eighteenth-century France. There, as in United Kingdom, older measures to keep food prices low began to backfire by inhibiting production and trade; what was needed was policies to foster and manage circulation. According to Foucault (2007, 41–42), a profound shift in the meaning of government pivoted on scarcity:

It means allowing prices to rise where their tendency is to rise. We allow the phenomenon of dearness-scarcity to be produced and develop on such and such a market, on a whole series of markets, and this phenomenon, this reality which we have allowed to develop, will itself entail precisely its own self-curbing and self-regulation. So there will no longer be any scarcity in general, on condition that for a whole series of people, in a whole series of markets, there was some scarcity, some dearness, some difficulty in buying wheat, and consequently some hunger, and it may well be that some people die of hunger after all. But by letting these people die of hunger one will be able to make scarcity a chimera and prevent it occurring in this massive
scarcity operates as a principle of rule and of personal conduct. Indeed, one might say that scarcity represents a sort of precondition for the modern, an epistemological principle on which our lives are built. It operates as part of a powerful discursive formation of the modern world. (Watts 2005, 99)

The Enclosure Acts are the most famous instance of what Marx termed “primitive” or original accumulation, which can be understood as the making scarce of the means of production (including land) and subsistence (including food) through commodification (including the commodification of labor, which brings scarcity to bear on time itself). But as Foucault suggests, the realization of scarcity encompasses much more. Scholars have shown that primitive accumulation is an ongoing process by which the commodity form is imposed on an ever-growing array of inputs and outputs. As Kloppenburg (1988) shows in the case of seeds, scarcity need not be based on actual physical dearth: all that is required is that private property claims be established to some good, and that others be compelled by law, technology, or other circumstance to recognize those claims and purchase that good in the market.

Actual physical dearth can also be produced by the very market processes that benefit from it, and the neoclassical economists’ narrow focus on price plays an instrumental role in this dynamic. Resources that are abundant or freely available, and therefore not economized, can nonetheless be brought into production for exchange, directly as raw materials or indirectly as conditions of production (e.g., sinks for pollution), and these actions can cumulatively degrade or reduce the resources in question to the point that they become scarce at local or larger scales. Or, short of actual scarcity, such actions can prompt measures to enclose the resources and assign property rights to them, effectively turning them into economic goods subject to the logic of scarcity. Such measures are often taken under the sign of the “tragedy of the commons” thesis, according to which the absence of property rights, rather than the market forces that motivate overexploitation, is the cause of environmental degradation.

Neoclassical economic theorizing about market effects on unenclosed (uneconomized) resources has often employed fisheries (Gordon 1954) and pastures (Coase 1960) as exemplars and touchstones from which to draw conclusions for broader application. In a more recent variant, economists understand free or nonpriced “ecosystem services” as “externalities” that should be “internalized” into market signals through the assignment of prices. Costanza et al. (2014), for example, calculate that the value of the world’s ecosystem services...
SCARCITY

dwarfs official global economic output, and the Millennium Ecosystem Assessment prescribes internalization as the solution to addressing global environmental problems (even though it attributes most of those problems to industrialization and globalization). The related concept of “natural capital” equates the ecological with the economic: nature becomes capital, and vice versa, and the conservation of one becomes identical with the conservation of the other. But as Robertson (2006) has shown, the practices of measurement and abstraction necessary to commodify nature are riddled with weaknesses and may well make matters worse.

Neo-Malthusianism arose in parallel to, but distinct from, the consolidation of neoclassical economics in the mid-twentieth century. Broadly speaking, neo-Malthusians focused on the use-values of the natural world, beginning from the axiomatic observation that the Earth is finite, and deducing that a growing human population must inevitably run up against the limits thereof. Progressive era conservation and the emerging science of ecology, coupled with the catastrophic spectacle of the Dust Bowl, provided intellectual and political traction for these views, often expounded (then and now) in the rubric of carrying capacity (Sayre 2008). In Road to Survival, William Vogt (1948) defined carrying capacity as the ratio of “biotic potential” to “environmental resistance” – more encompassing, but otherwise analogous to Malthus’s pairing of population growth and agricultural production. Early neo-Malthusians were also strongly influenced by eugenics, which subsequent proponents euphemized and transmuted into theories of modernization and development. During the Cold War, fears that population growth and resulting food shortages would favor communism in the developing world helped motivate both the Green Revolution (to modernize agriculture and raise yields) and international family planning programs (including coercive sterilization campaigns in places like India). Paul Ehrlich’s famous book The Population Bomb (1968), commissioned and published by the Sierra Club, helped consolidate population as a defining concern of emergent environmentalism. Economists countered with studies that showed declining real costs for natural resources based on technological innovation and substitution; Barnett and Morse (1963), for example, could not find evidence of increasing scarcity in US natural resources except in the case of forestry.

In the 1970s, neo-Malthusians turned to systems analysis and computer-based modeling to assess the prospects for continued economic growth at a global scale. The most famous result was The Limits to Growth (Meadows et al. 1972), which predicted the collapse of the modern world economy within a century. Although more sophisticated than Malthus’s principle of population, the Limits model ultimately reflected the same mathematical disparity between arithmetic and geometric (or exponential) growth. The book provoked raging debates, in which mainstream economists attacked the model on technical grounds while insisting that economic progress would solve environmental problems and extend the limits of available natural resources. Growing prosperity, they argued, would also lead to growing concern about environmental quality and political momentum to protect it – the so-called environmental Kuznets curve. In response, the original Limits authors twice refined and updated their model, reaching broadly similar conclusions as before. Others have built similar models to calculate the “ecological footprint” of human activities, calibrated by the number of planets identical to Earth that would be needed to support those activities indefinitely; values over 1.0 were reached, they claim, in the late twentieth century (Wackernagel and Rees 1996) – a conclusion that critics
might understandably see as proof that the model must be wrong.

The dominant debates about scarcity today cleave along the same lines as they did circa 1970: *either* economic growth has already exceeded (or will inevitably exceed) the limits of a finite Earth, *or* human ingenuity, motivated by market competition and opportunity, will extend those limits indefinitely. What both sides of the debate share, perhaps unwittingly, is an obsession with scarcity as the unquestioned lens through which to examine modern society. Lost are the voices of those who have challenged scarcity itself, whether as a flawed basis for economic reasoning and policy (e.g., Leon Keyserling, who helped design the New Deal) or as an anachronism rendered obsolete by the miracles of modern technology (e.g., ecological anarchist Murray Bookchin). Equally invisible are the views of the man who designed the original *Limits to Growth* model, pioneer systems analyst Jay Forrester (1971) of Massachusetts Institute of Technology, who concluded in a separate monograph that industrialization was the root cause of both population growth and environmental problems, and that restraining economic development might therefore be the best solution. Demographers have concluded that there is no fixed carrying capacity of Earth for humans; their projections indicate, with significant confidence, that the human population will stabilize this century for reasons unrelated to food supplies. Yet mainstream debates continue to focus on “feeding 9 billion by 2050,” as though alleviating hunger were a matter of increased production rather than more equitable distribution.

In this light, it is supremely ironic that the emerging field of behavioral economics, conducted squarely within the neoclassical tradition, has produced clinical evidence that the perception of scarcity can interfere with rational thought. A large body of research indicates that, whether real or imagined, scarcity makes people “less insightful, less forward-thinking, less controlled” (Mullainathan and Shafir 2013, 14). The idea of scarcity causes people to misapprehend their interests and misallocate their attention and resources, leading to self-perpetuating “scarcity traps”: the poor become poorer, the busy become busier, and so on. This is presented not as a threat to the fundamental premise of neoclassical economics, however, nor as a critique of the society that scarcity thinking has helped to produce, but as a useful insight that “sheds new light on how we might go about managing our scarcity” (Mullainathan and Shafir 2013, 15). Scarcity is taken not as a postulate, but as an unavoidable, ubiquitous, universal reality of the modern world.

The most prominent challenge to scarcity thinking was mounted by Karl Marx, who insisted that any shortage – whether of food, land, income, or jobs – could only be understood in relational terms, and that the root cause of imbalances in capitalist society was not scarcity but overabundance or surplus. If there were too many people relative to available jobs, it was not due to any “natural” human propensity to breed, but to excessive amounts of capital that had been amassed and transformed into machinery that displaced laborers. The resulting oversupply of laborers – the industrial reserve army of the unemployed – depressed wages, rendering the poor vulnerable to high prices for food and other necessities. Even for capitalists, the threat of overproduction – which can cause prices to collapse and profits to evaporate – has been a greater threat than scarcity in the history of capitalist natural resources extraction, as Huber (2011) has shown for the case of oil. Perhaps the path to addressing modern scarcity begins by thinking about its opposite.
SCARCITY

SEE ALSO: Commodification of nature; Environment and resources, political economy of; Environment and the state; Environmental degradation; Environmental valuation; Environmentality and green governmentality; Famine; Food security; Governmentality; Modernity; Natural resources; Neoliberalism and the environment; Population growth; Population and natural resources; Poverty; Property and environment

References


Science and technology parks

Susan M. Walcott
University of North Carolina at Greensboro, USA

Definition and context

Science and technology parks (STPs) occupy a piece of land configured to promote economic activities of integrated firms, infrastructure, and amenities supporting technology-intensive production and commercialization in innovative enterprises, technology-based start-ups, research institutes, and universities (Bellavista and Sanz 2009). The key geographical feature of these developments lies in their combination of territoriality and functionality, which is targeted to promote a particular type of technology-based economic activity (Asheim 2000). Firms are attracted to these clusters due to their supply of a sufficiently skilled and abundant labor pool, and access to input, output, and assembly infrastructure, along with spaces organized to promote information exchange. The latter can take place in food and drink areas designed to mix workers informally, and/or more formally in lectures and events providing training and networking opportunities. Advanced and developing countries utilize STPs’ spatial agglomeration for a “catch up” technology leap, creating technopole “learning regions” to build local and global knowledge relationships while efficiently concentrating scarce fiscal and human capital in discrete places (Castells and Hall 1994). University “spill-overs” of research and personnel increase the absorptive capacity of companies, promoting codified and tacit knowledge transfer by increasing their proximity.

Physically, STPs appear similar to research centers. Glass buildings and landscaped campuses usually sprout in outlying rather than central city urban locations due to their relative newness and the amount of land assembled for amenities as well as core business functions. The “synergy of amenities, accessibility and agglomeration factors found in large urban regions” (Malecki 1997, 270) indicates that metropolitan scale urbanization as well as localization economies come into play. These characteristics hold whether in advanced countries in North America and Western Europe or in rapidly developing, relatively advanced regions such as East Asia and India. Regional economic development motivated creation of the first STP – North Carolina’s still flourishing and growing Research Triangle Park – in 1959 as a joint government-university-private enterprise. This motivation and underlying triad of contributing interests continues to fuel the spread of STPs around the world where the incentives and supports are strong enough to feed the needs of this particular type of development geared to generate innovation-based profits for corporate and territorial interests.

Unlike endogenously generated regions that loosely include related clusters of high technology companies such as in northern California’s “Silicon Valley” or Boston’s “Route 128,” STPs comprise territorially distinct and specifically configured real estate. High technology-based companies and business incubators (BIs) constitute the core components. Due to the relative inexperience of their occupants, BIs
Science and Technology Parks

rely particularly closely on cooperative networks for training and support. Although research and development (R&D) personnel frequently feel that they are intelligent enough to manage the business end of their innovative spin-offs, STP personnel feel it greatly beneficial to include training in business best practices and encourage teamwork along with other specialized skills for building long-term successful companies.

Particularly in developing countries where attractive amenities such as housing, transportation, retail, entertainment, and education facilities tailored to a globally competitive “knowledge worker” labor pool are in short supply, the STP can include them within its boundaries or cooperate with other interested parties to construct them nearby. Amenities include low-density settings near to prestige research facilities, thus promoting intellectual interaction. More amenities may be introduced in areas where they less frequently occur, such as country clubs, tennis courts, and international schools. This variety of elements and levels of regional development provides some flexibility in definition, but several attributes remain constant. The following sections will consider the aspects and impacts of territorial boundaries, regional development (at various scales), public–private partnership, university technology transfer, birth and growth of innovative companies, and global linkages.

Industrial parks and their STP variant inspired a number of terms as part of post-Fordist (postmanufacturing-focused) “new industrial spaces” forming new “high-tech locations” as an “innovative milieu” on the high-value-adding end of a production chain network (Park 1996). More simply, Henry Ford’s model of a large factory complex where all supplies entered and a finished product exited became eclipsed by a model prioritizing greater variety of products for a market demanding faster and more customized production. This model combined with increased technological sophistication of production tools and methods, drawing on computer-aided elements for both design and manufacturing. The large corporation dissolved into a more “horizontal” organization utilizing several competing but allied suppliers that were often small and medium-sized enterprises (SMEs). Locally clustered concentrations enhanced communication and inspection that increased accuracy and speed while decreasing the “friction of distance.” However, local specializations and widespread telecommunication technologies have enabled global extension of production networks, so STPs draw from local research strengths at nearby institutions as well as featuring multinational companies (MNCs) drawn by these specialties and a perceived desirability of being in or associated with that geographic location. MNCs coordinate basic and applied research with long-distance product development and production, marketing, distribution, service, and financial functions in multiple locations, with the former frequently taking place in an STP and related activities at nearby or headquarter locations.

Castells and Hall (1994) listed three motivations for establishing science parks: reinstitutionalization based on greater infusions of science and technology (S&T), regional development, and synergy creation through promotion of technology transfer to businesses from research institutions. The conscious creation of spatial proximity among these players in the vicinity of an STP is designed to promote communication interaction among humans to create an atmosphere and structural framework conducive to promoting and sustaining constant innovation (Scott and Storper 1987; Castells and Hall 1994). This virtuous cycle requires a suitable physical
infrastructure, institutional development, systemic actor networks, and universities with specific research linked to product in order to foster and embed social relations in a particular place.

Networks and the embeddedness of human and corporate participants are key features in the success of an STP’s economic development mission (Park 1996). These networks consist of multiple dimensions. On the human level, trust-building connections are key attributes for information sharing. In both developed and developing countries, STP managers need to be well connected to government representatives because of the nature and extent of government ties regarding funding and regulations. The first major business in the first American STP, for example, was a major government contractor. The need for ties to multiple government levels became evident, for example, in the case of one Chinese STP that was a product of two national governments but was undercut by the local and provincial government, which did not have a stake in the STPs’ success and started their own rival STP zone. On the micro level, companies benefit best from proximity to similar companies within an STP when managers have good relations with other company managers, and when workers circulate among companies and transmit information that improves their employer’s performance. The density of employment opportunities helps to attract mobile “kangaroo” knowledge workers whose job-hopping corporate cross-fertilization tends to raise the level of innovation within the region.

National governments play an indispensable role in the initial establishment of an STP due to the amount of capital required to assemble the land, physical and intellectual infrastructure, and amenity inducements necessary. Local authorities and finally businesses established within the STP become more important over time. Complex collaboration networks develop to coordinate research, development, production, financing, and other services needed at various stages for businesses in the STP. Corporate or government funds, for example, can be provided to research university laboratories, professors, and graduate students to work on projects targeted for development or commercialization. University technology transfer offices frequently provide assistance with expensive and time-consuming patent applications filed by their employees. The degree of expertise in sophisticated institutions such as Stanford, Massachusetts Institute of Technology, and UK Oxbridge universities supports the success of their related regional STPs. The more cooperative and interlinked the participants, the “fuzzier” the institutional boundaries and the greater the likelihood of successful ventures.

Major universities in more advanced developing countries such as China often have allied STPs supported by their employees and alumni, with students given time off to develop their ideas in special arrangements outside of academic requirements. STPs supply savings in sunk costs by offering occupants otherwise expensive components and services such as computers and printers, telecommunications linkages, training sessions, and opportunities for information exchange with others in the same facility and/or mentors who may come from successful companies related to their area. As Park (1996) observes, “Collaboration among small innovative firms – as well as between large and small firms – can change the local industrial culture and may be the means for regaining regional competitive advantage” (486).

The idea was, and still is, to provide a technical, logistical, administrative, and financial infrastructure to nurture new enterprises so they can eventually survive in an increasingly competitive market. STP funding generally comes from five
sources: universities (including bank borrowing), local authorities, government development agencies, private-sector institutions, and the tenant enterprises themselves. Governments devote considerable resources to science parks as policy instruments aimed at promoting research based on industrial and innovative activity. Proximity promotes the ability to fund, staff, and share resources with outside information sources such as higher education institutes. Science parks provide a nurturing cluster of elements acting as innovation enclaves, populated by groups of related enterprises located in one geographical region or centered at a nation’s science-based park. The clustering and interchange process among industries in the cluster works best when the industries involved are geographically concentrated.

Changes over time

Early observers of STP processes varied from skeptical to enthusiastic adopters (Castells and Hall 1994; Park 1996; Malecki 1997). Examinations of individual STPs and of the underlying concept are far from unanimous in their assessments of success or desirability. Massey, Quintas, and Wield (1992) see STPs largely as land speculation built on wishful “high-tech fantasies”: build it, they will come — then what? Contemporary observers in developing countries are more universally sanguine about the STP strategy favored by economically interventionist governments. At present the strategy of creating a science park to concentrate companies involved in high-technology ventures is globally utilized in developed and higher-level developing countries. Established examples of major parks from East Asia (Japan, Taiwan, Korea, and China) to North America, throughout Western Europe, and in developing regions in Malaysia and Iran have all been studied for their relative successes, strategies, and possible policy applications. Studies indicate that the success of university-linked science parks in the United States and Europe provided global models for next-tier developing countries such as Kuwait, Brazil, Russia, Israel, India, Taiwan, and China seeking to derive economic benefits by building their own. STPs spurring significant growth in newly developed Asian countries include Japan’s Tsukuba, Korea’s Daedeok, Taiwan’s Hsinchu Science Park, and several in China, most notably Zhongguancun in Beijing and Zhangjiang in the newly developed Pudong district of Shanghai, with a high linkage to public policy and governmental support. In several of these cases new urban areas grew up to support the growth of the STP on the outskirts of older cities.

Korea initially established science parks in Taejeon and Kwangju, urban centers that were products of government industrial and regional policies encouraging population density away from the capital city of Seoul in the 1970s and 1980s. Highly educated citizens were particularly targeted to move to developments such as Daedeok Innopolis. These areas acted as an export generator, with R&D centers in the science park drawing largely on global linkages. National government incentives included tax exemptions, desirable infrastructure, cheap land, and financing. These bore fruit as spin-off firms related to research institutions appeared since the 1990s in Daeduck Science Town and nearby Taejeon city. Signs of STP success also lay in increasing linkages and firms of various sizes, resulting from a mixture of post-Fordist corporate deverticalization and national government policy (Park 1996).

Japan created the technopolis complex of Tsukuba as part of a government policy to promote business links between industries and universities. Under the auspices of the national
government’s ministry of international trade and industry, R&D infrastructure along with lifestyle amenities created a landscaped oasis in a mountain setting for high-tech industries employing young graduates from across the country. Planning started in the 1960s for a new city that began to rise in the 1970s and flourished in the next decade with research laboratories, universities, housing, and related civic functions. Tsukuba in turn served as a model for and fed employees to Taiwan’s first successful STP in Hsinchu, on former rural rice paddy land lying several hours outside the capital city of Taipei.

Built by the government in the early 1980s, Hsinchu picked up unemployed Japanese R&D talent to staff numerous high-tech companies including semiconductor manufacturers (Taiwan Semiconductor Manufacturing Company Ltd., United Microelectronics Corporation), and computer, telecommunication, and optoelectronics firms. Hsinchu combined science-themed cultural amenities, proximity to powerhouse universities, and links to Silicon Valley in the United States through alumni and personal networks, reversing previous “brain drain” activity by luring Taiwanese back to rebuild their country of birth in settings similar to those they were accustomed to overseas. Science parks in Taiwan have incubated six major industrial clusters: integrated circuits, computers and peripherals, telecommunications, optoelectronics industry, precision machinery, and biotechnology. As key actuators, science parks play an important role in stimulating high-technology industry by Taiwanese companies, both in Taiwan and in mainland Chinese STPs and special “cross-Straits” parks.

Special areas that in China are variously called “pioneer,” “gazelle,” or “returnee” parks within larger STPs provide incubators, office space, and related facilities for high-technology firms started by Chinese who formerly studied abroad and now seek to utilize local employees for ventures based on lessons learned abroad. Due to the level and nature of skills involved, these returnee entrepreneurs often particularly seek to employ others like themselves who studied at overseas universities. This development is clearly a valuable return on the investment of basic training in the home country, in some cases, and increasingly, including state scholarships to study abroad. The reversal of “brain drain” seen in the past provides valuable new talent with managerial and technology knowledge. Employment advantages of the overseas educated are decreasing as domestic universities raise their level of training, but STPs remain key sites for high-tech economic activity. They are now linked in a national organization so they can share mutually reinforcing strategies and experiences.

Future directions in theory, research, and methodology

A third-generation science park (3GSP) structure is detected as having evolved from a first stage focused on pushing innovative technology with a beneficial regional economic impact. The second stage followed market-led directions and contributed at a national level of economic benefit. The focus of a 3GSP was seen to be at a global level of innovation generation and diffusion along with new management processes that are more sustainable and human-oriented (Kakko and Inkinen 2009). In line with this emphasis, questionnaire and interview methodologies continue to be extensively utilized as in earlier studies for assessing the workings and impact of STPs.

Major challenges shaping the elements of a particular STP include creating sufficiently attractive elements to draw in the type of creative, innovative idea-generating talent desired. Social network analysis, focusing on the particular
structure of human networks involved in information creation and exchange, is increasingly applied to examine the workings of STPs in their role of promoting collaboration among corporate, public-sector, and research entities (Kakko and Inkinen 2009). A new research thrust increasingly focuses on policies promoting these creative/innovative interactions leading to development of ideas and commercialized products. Analysis utilizes approaches from organizational studies, economic development, social geography, urban theories, and educational and human resource studies to account for variables such as culture, climate, systems, resources, skills, and leadership. This focuses particularly on the actions of skilled human actors facilitating flows of knowledge exchange through STP structures (Bellavista and Sanz 2009).

SEE ALSO: Accessibility, in transportation planning; Economic geography; Industrial districts; Industrial location theory; Innovation and regional development

References


Further reading

Sea ice formation

Freshwater typically freezes at 0°C; however, because seawater contains a high concentration of salt (seawater with a salinity of 34 contains 34 g of salt per kilogram of water) it does not freeze at the surface until the water temperature falls below −1.86°C. Once it starts to freeze, ice crystals form (called frazil ice) and rise to the ocean surface to form slicks of accumulated crystals that are termed “grease ice.” In turbulent waters the ice crystals accumulate to form loosely aggregated discs, “ice pancakes,” 5–10 cm in diameter. These grow larger becoming pancakes 20–50 cm thick that can eventually reach several meters in diameter. They freeze together and eventually a closed ice cover forms. Under quiescent conditions, the ice pancakes do not form but instead the ice crystals consolidate into uniform sheets of ice, called “nilas ice.” As temperatures decrease further, both consolidated pancake ice and nilas ice thicken by the accumulation of more frazil ice crystals, but also by the growth of “columnar” ice. This is ice that is made up of vertically elongated crystals that can reach lengths of tens of centimeters.

Other mechanisms by which pack ice thickens include winds and/or water currents causing ice floes to converge. When ice floes collide under great pressure the edges break into boulders of ice and the rubble collects in ridges of ice called “pressure ridges.” These jumbled up blocks of ice can, in extreme cases, extend down to 50 m under the water (keels) and tens of meters into the air (sails). Such substantial sails and keels clearly are important in determining the effects of currents and wind on the subsequent movement and transport of an ice field. Conversely, winds are a very effective force for breaking open fields of pack ice. As the ice is forced apart, the water underneath is exposed forming a “lead.” Depending on the speed of the wind, and the strength of the ice floes, leads can be anything from a few meters wide to several kilometers.

As ice crystals form, the dissolved salts in the seawater are expelled and collect as brines between the ice crystals in a network of pores and “brine channels” that permeate the ice matrix. The brine channels vary from a few micrometers to several millimeters in diameter, the size of which is determined by temperature: the amount of ice in relation to the pore/channel volume increases proportionally with decreasing temperature. Additionally, as temperatures decrease, the salinity of the brines will also increase; for example, for ice made from water of a salinity of 34, at −6°C the corresponding salinity of the brine will be 100 and at −10°C it
will be 145. As ice gets older, drainage of brines by gravity gradually results in salt being lost from the ice. This is one of the main mechanisms by which cold, high-salinity water masses are formed, which subsequently sink to the ocean depths and partly drive large-scale global ocean exchange.

Sea ice and Arctic Ocean oceanography

Arctic oceanography and the role that sea ice plays in setting the structure and transforming the water have their formal origins in the nineteenth century. It was the Norwegian explorer and scientist Fridtjof Nansen, during the arctic drift of the *Fram* (1893–1896), who described the characteristic layers of water with different physical properties in the ocean, and the possible interactions between the ocean and the ice. Access to the central Arctic Ocean by ship has historically been extremely challenging and so the sea ice itself has been exploited as an ideal platform for trans-arctic sampling as the icedrifts across the Arctic Basin. The first of these manned scientific stations drifted 2850 km between May 21, 1937, and February 19, 1938. Since then there have been around 40 such drift stations established, the most recent ending in June 2013. During the course of the drift, a program of sampling for oceanographic and meteorological parameters has been conducted which has significantly added to our understanding of arctic oceanography. Ice drift stations have also been set up to tackle particular research challenges, an example being the Surface Heat Budget of the Arctic Ice Station experiment (SHEBA), which aimed to measure the rate at which heat is exchanged between the ocean, the sea ice, and the atmosphere throughout the polar year. The year-long study was supported by the Canadian Coast Guard vessel *Des Groseilliers* which drifted over 2000 km between October 2, 1997, and October 11, 1998.

Sea ice is also used to support autonomous technologies for sampling the seasonally changing oceanic conditions of the Arctic. The current generation of ice-tethered platforms (ITPs) are able to make measurements in the upper 750 m of the ocean using a package of sensors that are able to move up and down a fixed line suspended underneath the ice. Typically, measurements of temperature, salinity, and depth are made at a resolution of 25 cm and the data are sent through satellite links back to the shore in near real time whilst the profiler drifts in the ice for up to three years.

The Arctic Ocean is essentially landlocked with rather limited routes for exchange and communication with the rest of the global oceans. The deepest connection is Fram Strait, which is greater than 3000 m; there is also a shallow entrance at the Bering Straits and further connection through the network of straits and sounds that is formed by the Canadian archipelago. The surface circulation patterns in the Arctic Ocean are dominated by two wind-driven currents: the Beaufort Gyre over the Canada Basin and the Transpolar Drift, which flows from the Siberian coast across the North Pole and exits through the western Fram Strait (Figure 1a). The major inflow to the Arctic is warm and saline Atlantic Water entering via the Barents Sea or through the eastern Fram Strait forming a warm boundary current that encircles the Arctic Basin. There is also a warm and fresh inflow to the Arctic Ocean from the Pacific that brings a third of the freshwater into the Arctic and is a significant source of inorganic nutrients.

The surface of the Arctic Ocean is covered with a seasonal sea ice cover that forms each winter, reaching its maximum extent in March before diminishing through melt and
Figure 1  (a) The Arctic. Blue arrows represent cold currents and red arrows the inflowing Atlantic Water. BG, the Beaufort Gyre; TD, the Transpolar Drift; AW, Atlantic Water. The drift of Fridtjof Nansen’s ship the *Fram* is shown in Green. Darker gray shows the mean maximum winter ice extent from 1979 to 2007 (February) and the lighter gray the minimum summer ice extent (September). (b) The average seasonal cycle of arctic sea ice extent against date calculated from the satellite record from 1981 to 2010; the gray shading either side of the line is the standard deviation.
export to reach a minimum extent in September. The winter maximum extent is typically 15 million km$^2$ whilst the summer minimum is around 5 million km$^2$ (Figure 1b) with a currently observed record minimum in 2012 of 3.4 million km$^2$. The extent of arctic summer sea ice cover is undergoing a rapid decline, decreasing at an average rate of around 14% per decade and some commentators have predicted an ice-free summer Arctic Ocean within a few decades.

The impact of sea ice on arctic oceanography is essentially through three principal mechanisms: (i) the release of dense, salt-enriched brine during ice formation, (ii) the export or release of freshwater during melt, and (iii) the insulating effect of the ice between the atmosphere and the ocean. Other impacts include enhanced stirring and mixing of the water when the ice and ocean move relative to each other. It is important to stress that the structure and properties of the ocean also impact on the formation and retention of sea ice. Consequently, they must be considered as a closely coupled system with multiple feedbacks.

One of the most important oceanic structures in the Arctic Ocean that is intimately related to sea ice is found at depths between 50 and 200 m and is the boundary between cold and fresh polar surface waters and the warm and saline Atlantic Water. It is characterized by a rapid increase in ocean salinity (a halocline) whilst the water temperature remains close to the freezing point. As noted above, it was Fridtjof Nansen who observed that beneath the halocline “this warmer and more strongly saline water must clearly originate from the warmer current of the Atlantic Ocean,” so the halocline is a steep gradient in water density and acts as a barrier preventing the warm Atlantic Water from mixing up into the surface water and causing enhanced sea ice melt.

The formation of the halocline has been linked to the presence of extensive shallow shelves around the Arctic Ocean. Sea ice formation in these areas can produce large volumes of cold and dense brine that flow across the shelf and enter the Arctic Ocean below the polar surface waters and replenish the cold saline waters of the halocline. Another, arguably more likely mechanism is that of sea ice formation inducing brine release and initiating convection in the surface waters. The cold and rather saline water produced sits on top of the Atlantic Water which is then further capped in the summer through sea ice melt. As this seasonal cycle continues the halocline layer becomes fully formed.

A critical measure for the stability of the remaining sea ice cover is the strength of the halocline, measured by the salinity gradient between the surface and the Atlantic waters. This strength of the halocline has been observed to vary between years and geographically. Weakening of the halocline has been linked with the way that fresh water is distributed across the shelves. Any weakening has the potential to allow Atlantic Water to mix more effectively into the surface waters and this enhanced ocean heat flux can reduce the winter growth rates of sea ice and thus the total ice formation during a growth season. Therefore, the formation and presence of the arctic halocline are intimately linked to the formation and retention of sea ice.

One process by which the sea ice can strongly impact the ocean is the formation of polynyas, which in Russian means “ice hole.” They can vary in area from tens to thousands of square kilometers and usually reoccur at specific geographic locations, implying that the mechanism of their formation is related to regional factors. Whilst polynyas can be formed through enhanced ocean heat fluxes causing thinning and melting of the ice cover (termed a sensible heat
Sea ice, ice drift, and oceanic circulation

Polynyas (or latent heat polynyas) can also be formed through a highly dynamic wind-drive process. Often these polynyas are located near the coast with offshore winds pushing ice away from the coast to maintain a region of open water, and there are many such localized polynyas along the arctic coastline. The surface of the water is maintained at freezing point, and as latent heat is lost to the atmosphere ice begins to reform only to be advected away from the coast by the wind. As this process continues, the release of brine from the forming sea ice causes the salinity of the shallow coastal waters to increase while the ocean-to-atmosphere heat fluxes can be many orders of magnitude greater than elsewhere in the polar ocean. The high salt fluxes to the ocean created in these so-called ice factories can ultimately lead to the formation of dense waters that can feed the halocline layer of the Arctic Ocean.

Sea ice in the Arctic goes through an extensive melt phase in the summer when it is warmed from the surface and from beneath. Energy balances at the interface between the atmosphere and the ocean determine when melting will commence, but once started it can establish very effective feedback mechanisms that enhance the melt rates. Surface melting of arctic sea ice creates relatively dark melt ponds which absorb solar energy, greatly increasing the melt rates. As ice becomes mechanically weaker and breaks up, the exposed ocean absorbs significant solar energy at the surface causing melting from the bottom of the ice such that sea ice thickness has been observed to decrease by over 2 m in a summer season. Freshwater from melting sea ice can be exported through the Fram Strait and Canadian Archipelago or retained within the biggest store of freshwater in the Arctic – the Beaufort Gyre.

As summer sea ice continues to retreat, the area of open water and the duration for which there is open water will increase. This will change the way in which ice can impact the Arctic Ocean. The loss of summer sea ice removes the insulating properties of the ice such that there will be increasing heat content of the surface waters by autumn. The effect is a delay in the onset of freezing as ocean heat is removed to the atmosphere. Once freezing starts, there is now a greater area of open water for ice formation and thus the generation of brine enriched waters is occurring over a much more extensive area both over the shelves and over the deep ocean. The delay in onset of freezing leads to a shorter freezing period and thus the ice at the end of the following winter is thinner and more susceptible to breaking up, exposing the ocean surface to solar heating. Therefore the fate of sea ice is inextricably linked to its relationship with the oceanography of the Arctic Ocean.

Sea ice and Southern Ocean oceanography

The Antarctic has been relatively isolated from the rest of the planet since the continent first separated from South America about 23 million years ago. At this point a circum-continent ocean current called the Antarctic Circumpolar Current (ACC) was established and the temperatures on the continent fell until it soon became ice covered – both on land, and with a very large seasonal sea ice cover. In complete contrast to the Arctic, there has never been an indigenous population such as the Inuit learning to live with and exploit the sea ice environment. As a consequence, the region was untouched until recorded history tells us Captain Cook discovered the Southern Ocean sea ice at 66.5°S in January 1773. During the next 100 years many sealing and whaling ships had visited the region, although it took until the late twentieth
SEA ICE, ICE DRIFT, AND OCEANIC CIRCULATION

century for even the continent’s coastline to be comprehensively mapped. It remains a fact that the Southern Ocean is extremely under-sampled in both oceanography and sea ice when compared with the work conducted in the Arctic Ocean. Furthermore, there are still large regions devoid of any data. In this section we will discuss the background oceanography of the Southern Ocean before considering the sea ice and its impact on the underlying waters.

The Antarctic is a central continental land mass surrounded by deep oceans, and Figure 2a shows a schematic diagram of the ocean circulation and extremes of the summer and winter sea ice extent. The isolation of Antarctica and the global wind field formed the Antarctic Circumpolar Current (ACC), and it is the largest continuous ocean current on the planet, circulating clockwise around the continent. The ACC is delineated by four circum-Antarctic full-depth ocean fronts (boundaries that can be clearly identified in temperature and salinity). The northern boundary of the ACC that is most pronounced in temperature and salinity of the surface waters is called the Sub-Antarctic Front (shown in Figure 2a and labeled SAF). The surface temperatures of waters to the south of this front are typically in the range 5 to 7°C. Further south is the Polar Front (PF) which has been historically termed “the Antarctic Convergence,” and surface temperatures again fall, this time to approximately 2 to 3°C. The SAF and PF can sometimes be relatively close, particularly when passing through the constriction between South America and the Antarctic Peninsula, which is called Drake Passage. The other two fronts are much further to the south. They are the Southern Antarctic Circumpolar Current Front (SACCF) and finally the southern boundary of the Antarctic Circumpolar Current (SBACC). Whilst there is much evidence to connect the SACCF with biological processes, it is the SBACC which is the easier to identify and its mean location is shown in Figure 2a as a dashed line. The flow of the ACC is large and in Drake Passage is typically around $100 \times 10^6$ m³ s⁻¹ between the SAF and SBACC.

South of the ACC there are two large closed circulations called gyres (marked on Figure 2a). These are the Weddell Gyre and the Ross Gyre, and both circulate in a clockwise sense. There is no equivalent to the arctic halocline in the Antarctic such that there is less restriction on vertical exchanges of water. Consequently, the gyres are globally important for the production of some of the densest waters in the global ocean – the so-called bottom waters. Finally, south of the ACC and the gyres there is evidence of an anti-clockwise coastal current close to the continent through for example the drift of icebergs breaking from the land ice of the continent. But as data are so lacking close to the continent we cannot say whether this current is continuous all the way around the continent.

The sea ice in the Southern Ocean, as in the Arctic, has a strong relationship with the underlying ocean circulation. The light gray shading in Figure 2a shows the mean minimum sea ice extent in February and at the height of summer, and the dark gray in Figure 2a shows the maximum sea ice extent in September. Overarching all of these general features of the sea ice, we are severely constrained by the lack of data. This is particularly striking when contrasted with the historical data from the Arctic collected from the drifting ice stations, ships, and many tens of thousands of kilometers of sea ice thickness data collected by nuclear submarines since 1958. In the context of sea ice extent, this has not been too much of an issue since the satellite record has produced high quality data since 1981. But for sea ice thickness the only data sources going back any period of time have been field experiments,
Figure 2  (a) The Southern Ocean. The location of the two ocean fronts that delineate the Antarctic Circumpolar Current (ACC) are labeled SAF for the Sub-Antarctic Front and SBACC for the Southern Boundary of the ACC. Arrows represent the schematic circulation direction. The darker gray shows the mean maximum winter ice extent from 1979 to 2007 (September), and the lighter gray the minimum summer ice extent (February). (b) The average seasonal cycle of Antarctic sea ice extent against date calculated from the satellite record from 1981 to 2010; the gray shading either side of the line is the standard deviation.
SEA ICE, ICE DRIFT, AND OCEANIC CIRCULATION

observations by ship collated through the Antarctic Sea Ice Processes and Climate (ASPeCt) program, and a few under-sea moorings equipped with upward looking sonar. The only drifting ice station comparable to the arctic efforts was Ice Station Weddell in 1992 (ISW-1) which followed a track through the Weddell Sea similar to that of Shackleton’s doomed Endurance in 1915.

Figure 2b shows the annual cycle of the mean Antarctic sea ice extent based on daily data from the period 1981–2010, and the gray shading either side of the line is ±1 standard deviation. The incredible range shown in the extent is striking. The minimum sea ice extent is in late February and is approximately $2.9 \times 10^6$ km$^2$; at this time one can see in Figure 2a that the ice is limited to relatively close to the continent and in the Ross and Weddell Seas. In September the extent has increased to approximately $18.6 \times 10^6$ km$^2$ and the northern limit in many locations around Antarctica is close to the SBACC. Where the northern limit of the sea ice extent diverges widely from the SBACC may be due to a relatively poor determination of the mean location of the SBACC due to data limitations, and it is likely not well fixed. The range of sea ice extent over the course of a year is approximately $15.7 \times 10^6$ km$^2$ and the sea ice expands by more than five times the minimum extent between summer and winter. At its maximum, sea ice in the Southern Ocean covers an area larger than the size of the Antarctic continent. This has enormous implications for the thickness of the sea ice: clearly most of it is going to be less than a year old and the only multi-year sea ice (as a consequence much thicker when unrafted) would be close to the coast and the gyres. It is also clear that there is an asymmetry in the seasonal cycle of Antarctic ice extent. Figure 2b shows that the sea ice grows at a steady rate from February through to July, then a much slower rate from July through to the beginning of September, followed by a very rapid sea ice retreat from the middle of September through to February.

Satellite images show that the steady sea growth due to the changing seasons extends outward from the continent. There is relatively high snowfall compared with the Arctic, and surface flooding of the sea ice and formation of snow ice (where sea water permeates into snow and freezes) has been widely observed. In contrast, as the sun returns the ice decays in much of the dark gray region in Figure 2a. The effect of this slow sustained growth versus rapid and distributed retreat is seen in the ocean. The slow growth and continued salinity input increase the density of the surface waters, which, through currently not well-defined processes in the gyres and on the continental shelves, creates water dense enough to sink and form bottom water, which then contributes to the global ocean circulation. The retreat causes a considerable input of low salinity water into the surface waters that are then mixed into the surface layers by wind activity.

Although there have been many regional changes in the distribution of sea ice extent over the length of the satellite record, there is a particularly significant retreat to the west of the Antarctic Peninsula. However, over the duration of the satellite record there has been a slight positive trend in Antarctic sea ice extent. Our understanding of the exact processes by which water density is increased due to sea ice formation is currently limited by a lack of data on water properties, but we expect this to be addressed in the coming decade through innovative techniques such as deploying hydrographic sensors on different species of Antarctic seals. The lack of comprehensive sea ice thickness data will be partially addressed by altimeter-equipped satellites such as ICESat and CryoSat.
Sea ice as a biome

When ice forms, the accumulation of ice crystals is effectively inoculated by a diverse group of organisms commonly found in the smallest size classes of the plankton in the waters from which the ice is formed. Most notably, these include bacteria, unicellular algae, ciliates, small species of copepods, flatworms, and nematodes. Many of these organisms have highly adaptable physiologies that enable them to survive in the brine channels in the ice and/or on the ice surfaces on the undersides of ice floes. In the open oceans these organisms experience rather buffered abiotic conditions in terms of temperature and salinity. In contrast, within the ice they experience vastly changing salinities in the brines and highly variable temperatures that can fall to $-20^\circ$C. Despite the extreme habitat, these ice dwelling organisms can thrive to such an extent that the ice, rather than being colorless, is frequently seen to be strikingly coffee-colored. This is due to the photosynthetic pigments contained within the unicellular diatoms that frequently dominate the sea ice communities in all ice-covered oceans and seas.

In turn, the biology contained within the sea ice is a key source of food for other marine organisms at times when food in the water column is scarce. One of the best examples is the Southern Ocean krill (*Euphausia superba*), a voracious grazer that normally eats plankton. In ice-covered periods the sea ice organisms provide the only food for many populations of krill that have been observed harvesting the algal-rich diet from the undersides of ice floes. The importance of this for the overall Southern Ocean ecosystem is highlighted by the fact that krill is in turn the major food source for squid, fish, seals, birds, and whales in these waters.

Although the study of how organisms are able to survive in the ice and their ecology continues to be an engrossing area of research, in recent years the role of ice in influencing the biogeochemistry of the surface waters of oceans and seas has emerged as a distinctive discipline. For many years it was presumed that a solid cover of ice on the surface of the ocean would effectively seal off the water from the atmosphere. It is now clear that this is not the case and during the process of ice formation, consolidation, and melt there is considerable exchange of gases such as CO$_2$ between the ice and atmosphere and the ice and water. Partly, this is a result of the biological processes in the ice (photosynthesis and respiration), although more recently the influence of cryogenically induced chemical processes has received increased attention.

On ice melt, some of the biology contained within the ice has been shown to inoculate extensive blooms of phytoplankton at the ice edges. Naturally, these are important for the overall biological productivity of these seasonally ice-covered ecosystems. However, a large percentage of the ice production is thought to simply sink to the ocean depths and if it reaches the ocean floor it provides a valuable food source for the organisms living there.

Although scientists have been venturing to the frozen oceans and collecting data for several hundred years now, it is only in the past 50 years or so, with the advent of specialized research platforms, that there have been the internationally coordinated expeditions that have had the resources and opportunities to elucidate the profound implications of sea ice formation, not just for the oceanography of polar oceans and seas, but also for the seasonal dynamics of the biology living in these waters and underlying sediments.

**SEE ALSO:** Antarctica; Climate change and sea ice; Earth system science; Oceanic circulation; Oceans and climate; Oceans and seas: physical geography
SEA ICE, ICE DRIFT, AND OCEANIC CIRCULATION

Further reading


Sea level rise

Stephen P. Leatherman
Stephen B. Leatherman
Florida International University, USA

Coastal zones have been the wellspring of civilization since early times because of the ease of water-based transportation. About 50% of the world’s population lives in this critical interface between land and water, and 13 of the world’s 20 largest cities are located on the coast. Increasing populations and development are placing significant stresses on coastal resources; rising sea level is already causing land loss, putting humanity on a collision course (Figure 1).

Changing sea levels have been the driver of shorelines over geologic time. Tens of millions of years ago, the seas had advanced several hundred kilometers farther inland than today along the US East Coast. These ancient shorelines are marked by the “falls” in the river systems, and many of the major cities and state capitals are located there (e.g., Washington, DC, Baltimore, Maryland, Richmond, Virginia, and Raleigh, North Carolina). This fall line is the furthest upstream position for navigable waters and delimits the inland edge of the coastal plain.

If all the water tied up in the world’s glaciers and ice caps were to melt, sea levels would rise about 70 m, which would be devastating on a global basis. Consider that the Mall in Washington, DC, is little more than 3 m above today’s sea level and the Potomac River there is tidally driven.

Sea levels have also been much lower in the “recent” geologic past, approaching 100 m below present levels approximately 18 000 years ago. Coastal resorts, such as Atlantic City, New Jersey, and Virginia Beach, Virginia, would have been located several hundred kilometers landward of the shoreline at that time as the seas receded to the shelf edge, exposing the wide continental shelf. The Chesapeake Bay and associated salt marshes did not exist during the last Ice Age, and the Susquehanna River was a rapidly flowing river that reached all the way to the shelf edge.

Unlike many other manifestations of climate change, sea level rise is already a problem. While global sea level has only risen about 20 cm in the past century (Douglas 1991; Church et al. 2010), there is clear physical evidence of impacts manifested by coastal land loss on a worldwide basis. Erosion is the physical removal of material at a shore due to wave action. Leatherman, Zhang, and Douglas (2000) showed that beaches on the US Atlantic Coast, not affected by inlets or engineering modifications, erode at about 100 times the rate of sea level rise on average. They analyzed a large database of historical shoreline positions and long-term sea level change data obtained from tide gauges.

Over 70% of the world’s sandy beaches are presently eroding (Bird 1985), and the percentage increases to 80–90% for the better-studied and documented US sandy coasts. While many factors contribute to shoreline recession, sea level rise is considered the underlying factor accounting for nearly ubiquitous coastal retreat (Leatherman, Zhang, and Douglas 2000). Erosion of beaches and bluffs is a major worldwide problem. This land loss has enormous economic impacts; for example, some of the most expensive real estate in the United States is beachfront properties. Examples include Atlantic
City, New Jersey, Ocean City, Maryland, Myrtle Beach, South Carolina, Miami Beach, Florida, and Galveston, Texas, to name just a few of the cities on the beach.

Sea level rise also affects low-lying areas through inundation. Inundation is the permanent submersion of land as the sea level rises. This is not to be confused with erosion, which is the physical removal of sediment from the shore. The combination of inundation and erosion together is termed “coastal retreat.” Inundation is evident in many low-lying marshy areas, which can cause the marshes to die and disintegrate.

Sea level rise is the most certain consequence of global warming and has many adverse effects on the world’s coastal zone. Rising sea levels erode shores by increasing offshore and longshore loss of sediment, directly inundating (or submerging) marshes and other low-lying lands, increasing the salinity of estuaries and aquifers, raising coastal water tables, and exacerbating coastal flooding and storm damage. Beach erosion along developed coasts is the most obvious problem of rising sea level. However, sea level rise causes extreme disruption to humans living on small islands and river deltas. Millions of people will be displaced when large river deltas such as in Bangladesh become inundated; people living on small islands have no place to retreat, and this is especially true of island nations. While the potential impacts of sea level rise itself may not always be the largest threat to coastal environments and human habitability, in conjunction with other stresses they can become a serious problem for coastal
societies where the resilience of natural coastal systems has been reduced.

In addition to problems of erosion and inundation, hurricanes may become more intense due to global warming. Greater losses would occur from stronger storms because damage is proportional to wind speed cubed. In northern climates such as that of Alaska, retreat of sea ice is further amplifying coastal damage by allowing winter storms to pound and erode the land that has served as the home of Inuit and other indigenous groups for many millennia; so much erosion is occurring that villages have to be relocated inland (Arctic Climate Impact Assessment 2004).

Global (eustatic) sea level rise during the past century of about 2 mm annually is partly attributed to thermal expansion of the oceans and melting of small glaciers. About 25% remains unaccounted for, but the most likely candidates are the Antarctic and Greenland ice sheets (IPCC 2013).

The International Panel on Climate Change (IPCC) is the leading international body for the assessment of climate change; the organization was established by the United Nations Environmental Program and the World Meteorological Organization. The IPCC Fifth Assessment Report (IPCC 2013) projects that the global sea level will rise by 26–63 cm by the end of the twenty-first century (Figure 2). While unlikely, a sea level rise of greater than 1 m by 2100 is possible based on several IPCC models.

Sea level rise, which is supported by good historical data, is the primary climate indicator and provides constraints on global climate models. Sea level rise can be considered the “dipstick” of global warming. It is a much better indicator of climate than weather patterns, which are constantly changing. Sea level rise is derived from the long records of tide gauges, some of which extend back more than a century. Tide gauges are available at many of the major ports in the Northern Hemisphere and sporadically elsewhere.

Tide gauges provide the longest historical records of sea level rise and are therefore the best data sources available to determine a trend. The longest continuous tide gauge record in the United States is located in San Francisco and began in 1854, but the vast majority of tide gauge records are much shorter (Douglas 1991). Variability, both interannually and over longer periods, as well as vertical land movements, can severely distort a trend calculation.

In recent decades satellite altimeters have been used to measure sea level rise. Church and White (2006) utilized such data to show an acceleration of sea level rise by as much as 50% since 1930. Other authors (e.g., Ray and Douglas 2011) did not find this acceleration; therefore, more study is required. While tide gauges have a high noise-to-signal ratio, satellite altimetry data, which are nearly global in coverage, can provide more accurate measurements of global sea level rise and hence sooner detect any acceleration of the rate.

In addition to inundation of low-lying areas and the erosion of beaches and bluffs, sea level rise also impacts the coastal zone through saltwater intrusion and higher water tables. Saltwater intrusion occurs when saltwater is pushed into a freshwater aquifer due to a higher water head, and can drastically decrease the amount of available fresh drinking water. This problem is greatly intensified when over-pumping of aquifers occurs.

The Marshall Islands in the Pacific Ocean has a drinking water problem. This is true for many small island states such as the Maldives in the Indian Ocean. The Marshall Islands only has 1–2 m of relief. As the sea level rises, the prism of freshwater under the island is drastically reduced by saltwater intrusion. In fact, the residents of some small coral reef atolls in the Marshall
Islands have to scoop the top 5 cm of the fresh water head or else they will encounter salty water (Leatherman 1997).

Higher water tables are also a problem for reasons other than saltwater intrusion. Higher water tables cause issues with drainage in low-lying coastal cities as there is not enough freeboard for the hydraulic head to cause rapid runoff. This effect is evident in cities, such as South Beach in Miami, Florida. During king tides (e.g., extreme spring high tides), Alton Road, which is the major throughway on the bay side of South Beach, experiences saltwater flooding. In fact, South Beach is undertaking a multimillion dollar project to replumb its drainage system to prevent nuisance occurrences of intrusion of salty water on roadways. Other areas affected by higher water tables include Charleston, South Carolina, and the Galveston Bay area of Texas.

Subsidence causes higher sea level rise relative to surrounding areas. This lowering of the land elevation is due to compaction of underground sediments or the collapse of the forebulge area in response to glacial activity (Douglas, Kearney, and Leatherman 2001). In order for subsidence to occur, the sediments need to be compactible.
such as clay or organic matter. Subsidence can occur from natural or anthropogenic causes, and is a major problem in deltaic areas.

Anthropogenic subsidence is due to the extraction of subsurface fluids such as oil, water, and natural gas. Such activities, especially over-pumping of groundwater, have resulted in inundation of parts of San Jacinto National Monument in Houston, Texas. This monument is an important historical area—it was the famous battleground where General Sam Houston defeated Antonio López de Santa Anna’s Mexican army for control of Texas. Due to the severe inundation experienced here, the Houston–Galveston Subsidence Control District has banned groundwater mining, and residents are now required to obtain water from rivers, which is purified through treatment to make it potable.

Deltas are one of the world’s most vulnerable areas to sea level rise in terms of endangering vast numbers of people, especially during the advent of coastal storms. There are millions of people living on deltas who are vulnerable to natural and human-induced subsidence, global sea level rise, and large coastal storms, especially hurricanes, cyclones, and typhoons.

Deltas are made of rich deposits of sediment forming flat, low-lying land. Human settlements are easily built on these flat lands and are prime agricultural areas. Salt marshes also abound. Deltas are prone to erosion and inundation in response to rising sea level. New Orleans is the only major US city located on a delta, which has been built over the last few thousand years next to the Mississippi River that drains the mid-section of the country. The original city was built on the natural levee of the delta, but further expansion with increasing population led to the draining of wetlands and development of areas below sea level. Hurricane Katrina in 2005 graphically demonstrated their vulnerability when the levees were breached, flooding 80% of New Orleans. Nearly 600 people drowned in the flood waters, making it one of the worst natural disasters in US history.

Breaching of the New Orleans levees is not a rare event; in fact, it has happened more than 20 times in the past (Colten 2005). Hurricane Betsy in 1965 also flooded 80% of New Orleans, and the US Army Corps of Engineers constructed much stronger fortifications following this devastating event and pledged that the levees would never fail again. It has been said that there are two kinds of levees—those that have failed and those that will fail in the future.

New Orleans exhibits the highest relative rate of sea level rise in the United States—it amounts to 1 cm per year. Eighty percent of this rise is due to natural and anthropogenic subsidence with the remaining 20% being the result of global sea level rise. New Orleans is also in the track of major hurricanes from the Gulf of Mexico, and Louisiana is the US erosion “hot spot” with beach erosion rates of up to 15 m per year.

Tokyo is another vulnerable city built on a delta, and almost half of this capital city has been built below sea level (termed a “below zero city” by the Japanese). In addition to vulnerability from typhoons, Japan also has the additional coastal hazard of tsunamis. Another Japanese city that is partly below sea level is Yokohama, which has the same vulnerabilities as Tokyo. The devastating effects of the 2011 earthquake and resulting tsunami that struck Japan demonstrate the vulnerability of all low-lying areas, especially cities near or below sea level.

Venice, Italy, is built on the Po River delta and is experiencing major issues with sea level rise, both from local subsidence due to oil and gas withdrawal and from global sea level rise. Venice was originally built on a marsh during medieval times and was once the world’s largest trading port. It is a major tourist attraction, with the famous Piazza San Marco dominated by the great
church of St Mark. The problem is that Venice is experiencing a high level of relative sea level rise and is frequently flooded during winter storm events termed “acqua alta” or “high water” (Figure 3). This salt water flooding is destroying the wooden doors and even the mortar of the brick and stone buildings of the famous churches and monuments such as the bell tower.

One proposed solution to prevent flooding in Venice is called the MOSE (Modulo Sperimentale Elettromeccanico, Experimental Electromechanical Module) project, modeled after the Thames Barrier that protects London from North Sea storm surges. This coastal engineering approach involves sealing off the three inlets to Venice lagoon using a retractable dam system. These massive structures would be able to rise out of the water and close the inlets during acqua alta events to prevent flooding of the city, but remain open at all other times to allow flushing of the lagoon. This tidal flushing is critical as Venice has no sewage treatment system and relies upon water movement to remove pollutants from the lagoon enclosing the city.

Figure 3 An “acqua alta” event at Venice, Italy, causing saltwater flooding during winter storms of residences that is also damaging iconic landmarks and historical buildings. Photo ZioNicco, courtesy of Wikimedia Commons.
Implementation of the MOSE plan has proceeded slowly largely because of the high cost of construction. Also, the moveable dams would have to be in the “up” position and hence the inlets closed at all times to prevent flooding if sea level were to rise just 30 cm, which would prevent flushing of the bay. Venice is considered one of the most romantic cities in the world and a popular tourist destination, but this would not remain true if the canals were grossly polluted.

There are many ramifications of rising sea level, and the Chesapeake Bay can serve as a microcosm of sea level rise impacts. This area experiences a wide range of impacts from sea level rise. The long shoreline (e.g., nearly 10,000 km) encompassing the Chesapeake Bay is subject to coastal erosion. Local tide gauges indicate local sea level has risen 0.35 m in the past century, which is about double the global rate during this time period.

One of the consequences of slowly rising sea level has been the diminution and loss of small islands in Chesapeake Bay due to long-term erosion and submergence (Leatherman et al. 1995). These vanishing lands, such as Smith Island, named after the great British explorer Captain John Smith, are also experiencing saltwater intrusion and higher water tables. The cause of this high relative sea level rise is a global sea level rise of 20 cm from the melting of glaciers and thermal expansion of water (e.g., eustatic sea level rise), plus local subsidence from geologic factors (e.g., collapse of the glacially induced forebulge) and probably also the withdrawal of groundwater for agricultural purposes (Douglas, Kearney, and Leatherman 2001).

The western shore of Maryland is a cliff shoreline composed primarily of sand, silt, and dewatered clay termed “marl.” This shoreline is experiencing land loss, albeit at a fairly slow rate, at only centimeters per year, because of the erosion-resistant nature of the marine marl. The cliff erodes in chunks from wave action at its base, exposing a plethora of fossils and making it one of the best places to find sharks teeth in the world. Many houses and even roads have fallen over the 20 to 30 m-high cliff into the water as erosion has progressed over the past century. Point Lookout, located where the Potomac River meets the main stem of the Chesapeake Bay, is one of the most scenic spots on the bay. This area has lost 50% of its land in the past century (Leatherman et al. 1995).

The bluffs of the eastern shore contrast sharply with the flat and low-lying western shore. Erosion does occur on the bay edge, but the primary response to rising sea level is the submergence (e.g., inundation) of land and drowning of salt marshes. Between 1938 and 1988, one third of the salt marshes on the eastern shore have been lost due to sea level rise (Leatherman et al. 1995). These salt marshes are very important for the ecology of the Chesapeake Bay, being a base of the food chain and primary source of nourishment for many organisms, including economically important oysters and blue crabs.

Coastal marshes can sustain themselves during a slow rate of sea level rise. These marshes accrete vertically through sediment input and/or biomass production. However, when sea level rise reaches a certain threshold value, these marshes can no longer sustain themselves and become inundated, resulting in massive land loss. These land loss events can occur over wide areas within a few decades as the individual plants become waterlogged and sulfates build up in the soil layer beyond their tolerance levels (Stevenson, Ward, and Kearney 1986).

Island communities in the Chesapeake Bay have been dramatically affected by rising sea level. These islands have been depopulating because of severe erosion and inundation that leaves less room to live and farm as well as bringing...
SEA LEVEL RISE

the problem of saltwater intrusion. The highest points on these bay islands rarely exceed 2 m.

Poplar Island in the Chesapeake Bay is an example of a bay island community that is rapidly disappearing. The first recorded sale of Poplar Island in the late 1670s indicated an area of 560 ha (Leatherman et al. 1995). In 1848, the island was accurately mapped by the US Coastal Survey at 300 ha. In the early 1900s, the town on the island had a general store, post office, school, church, and sawmill. Erosion was becoming such a major problem that the island was completely deserted in the 1930s. Poplar Island, at an erosion rate of 4 m per year, was expected to completely disappear by the year 2000. However, the US Army Corps of Engineers saved the island through implementation of a restoration project to help protect endangered black ducks.

Perhaps of all coastal areas the most vulnerable to global climate change and sea level rise are small island nations (Leatherman 1997). These areas have limited resources and little to no means of preventing the impacts of sea level rise. Small island nations are characterized by their size, limited range of natural resources, susceptibility to natural hazards, limited biological diversity, great distance from other markets, extensive land/sea interface per unit area, extreme openness of economies, low resilience of a subsistence economy, and narrow range of skills and lack of educated specialists (UN General Assembly 1992). These characteristics make island nations very vulnerable to external stresses, particularly sea level rise impacts.

Coral reef atolls are oceanic islands that are volcanic in origin. A coral reef atoll begins as an extinct, subsiding volcano. As the volcano subsides, coral growth forms a fringing reef. These coral reef atolls are very low-lying and susceptible to sea level rise. The Maldives is an island nation in the Indian Ocean that consists of 26 coral reef atolls. Sea level rise is already causing extreme problems in the Maldives. Erosion of sandy beaches is impacting the tourism industry, and therefore the Maldives economy. Additionally, sea level rise is resulting in saltwater intrusion and loss of fresh drinking water.

There are two responses to sea level rise: mitigation and adaptation. Mitigation must be undertaken on a global scale to reduce the rate of sea level rise, while adaptation can be accomplished on a local or national level. Mitigation can slow the rate of sea level rise, which would reduce the impacts. The goal of mitigation is to reduce the sea level rise to a rate at which the impacts can be reasonably accommodated through adaptation.

Hurricanes are historically known to force retreat because a catastrophe is sometimes necessary to convince coastal inhabitants to abandon the area. It would be wiser to anticipate a certain rate of beach loss and prohibit construction within a designated distance from the shore. Adaptation to sea level rise can be one of three approaches: planned retreat, accommodation, or protection. Planned retreat (Figure 4) involves moving development and infrastructure landward as coastal erosion encroaches. This is the best method along little-developed coasts, but in highly developed areas, such as Miami Beach, retreat is not economically or politically viable.

Accommodation is another method to respond to sea level rise. Accommodation consists of building elevated homes to adjust to projected future water levels. Houses built on strong pilings and elevated high enough can survive coastal storms; the surge will flow underneath the structure, causing little damage during these episodic events. This response method is best combined with retreat or else the structure will still be vulnerable to coastal erosion on sandy beaches.

Protection is the method of using soft or hard engineering approaches to protect the land from a rising sea. Stabilizing the shoreline with seawalls was once a common method of
responding to sea level rise, storms, and coastal erosion. Seawalls are massive concrete structures built to protect upland areas that are developed with houses and infrastructure. Seawalls are costly, and eventually they need to be replaced or protected with rock riprap or concrete tetrapods in response to progressive beach erosion. A concrete seawall can successfully protect the land behind the seawall, but the beach will erode. The seawall constructed on Galveston Island, Texas, following the Great Hurricane of 1900, prevented erosional loss of urban development and stopped overtopping of storm surges, but the recreational beach was lost long ago.

Beach nourishment is now considered a much better solution than hard stabilization, but it can be more expensive. Nourishment through the introduction of new sand to build wide beaches protects buildings and infrastructure by serving as a buffer to storm waves and surge. In addition, the recreational beach is maintained for both residents and tourists to enjoy. However, beach nourishment only sets back the erosion clock, and additional sand is required over time as it only treats the symptoms (e.g., erosion) and does not cure the disease (e.g., rising sea level).

Sea level rise is clearly affecting the world’s coast, and the impacts will be exacerbated in response to accelerated sea level rise. The most vulnerable landforms are deltas and small island nations. Millions of people reside on low-lying river deltas, some of which are in the poorest countries in the world, such as Bangladesh, Egypt, and Nigeria. These countries, with large populations and few resources, do not have the

Figure 4  A house on the Outer Banks of North Carolina in the process of being moved inland due to coastal erosion. Photo Jimmy Luby. Reproduced by permission.
SEA LEVEL RISE

ability to undertake large-scale projects to mitigate sea level rise impacts. However, sea level rise affects all coasts and is a large economic burden even in developed countries. For example, the Chesapeake Bay in the United States has more than 10,000 km of developed shoreline. The cost of shore protection for this area as a whole is prohibitive so that only highly developed areas such as Annapolis, Maryland, can achieve a high enough cost-to-benefit ratio to merit major stabilization efforts.

Unfortunately, there is no ultimate answer to prevent sea level rise. Reduced greenhouse gas emissions can help slow down the rate, but the sea level is still rising. A global agreement to reduce emissions does not seem possible as China, India, and other major developing countries are emitting more carbon dioxide each year. Sea level rise is a very slow process, and the best solution is to plan ahead for its consequences. Unfortunately, the public does not typically plan for sea level rise until it is too late, and politicians are much better at responding to a sudden crisis (e.g., hurricane landfall) than to a slow one (e.g., rising water level).

SEE ALSO: Climate change, concept of; Coastal zones; Oceans and seas: human geography; Oceans and seas: physical geography

References


Security

Ian Klinke
University of Oxford, UK

In its most basic form, security is the absence of threat, whether real or perceived. Throughout history, different issues, objects, and polities have been perceived as threatening and it is therefore difficult if not impossible to uncover a transhistorical and transcultural essence to security. Security has become a ubiquitous term in modern societies, governing highly different social contexts from national defense to cyberspace, and from airports to suburban crime statistics. Security is often invoked by states and their elites to foster support for exceptional measures, such as the restriction of basic human rights, targeted killings, or military invasions. Any call for such measures depends upon a certain level of insecurity or fear in society to seem pertinent and legitimate. The shaping and sometimes manipulation of this threat perception is key to the functioning of security discourses. Yet, despite its omnipresence and omnipotence, security is also elusive as a concept and a practice. It has become so interwoven with the social fabric that it is often difficult to locate or define.

Debates on security have traditionally been situated in security studies, a subdiscipline of international relations. Human geographers and those working on the margins of the discipline have made a number of important interventions into these debates, particularly by rethinking the question of security in an explicitly spatial register. Inspired by the end of the Cold War and the new security architectures that emerged with it, much of this work has challenged security as an expert-level realm of military knowledge. These transformations initially saw the emergence of critical geopolitics, an approach that investigates the role of problematic mental and cartographic maps in global security and human security, a paradigm that has placed the individual at the center of debates on international security. In the late 1990s, geographers started to discuss the displacement of security practices by a deterritorialized logic of risk. More recent research has tried to understand the development of security practices in a biopolitical age, tracing how they transform when states seek to govern populations through the attributes of life itself. In doing so, they have drawn attention to security dynamics in a number of areas that have received less attention in security studies: borders and territory, urban counter-terrorism and resilience, energy and climate change, and finally biosecurity.

Approaching security

Research in critical geopolitics emerged in the late 1980s and early 1990s to understand the role geopolitical knowledge had played in mediating between political elites, the military-industrial complex, and the wider public during the Cold War. It has approached geopolitics as a security discourse that surfaced in imperial Britain and traveled through Nazi ideology and Cold War strategy and into the unipolar world order and the “global war on terror.” Written by and for the state, geopolitical discourse naturalizes insecurity, conflict, and war as part of the human condition and suggests that balance of power considerations rather than ethical
SECURITY

concerns should be a guide to the pursuit of (national) security. Drawing on postcolonial theory, poststructuralism and feminism, critical geopolitics has helped to unearth the often banal and gendered identity dimensions involved in security discourses (Dalby 1990). In doing so, it has investigated the ways in which threats are constructed as emerging from far away, supposedly exotic and dangerous spaces, like the Soviet Union or the Middle East. Key to this endeavor is the insight that the constant articulation of insecurity is in fact the precondition for the state’s very existence. It follows that states should be understood not as security providers in a straightforward way, but as sources of global instability. Critical geopolitics has remained a vibrant field that continues to produce research on a wide array of security issues, including terrorism, humanitarianism, and global health.

Human security is a politically explicit approach to security that has tried to shift the referent object (that which is to be secured) from the nation-state to the scale of the individual. Drawing on a wide variety of ideological sources, including liberalism and feminism, human security was enshrined in the 1994 United Nations Human Development Report and has remained a powerful paradigm both inside and outside of academia. Part of its success has followed from the broadness if not the vagueness of human security, a syntagm that has variously included the freedom from poverty, hunger, diseases, pollution, torture, and forms of political oppression. Feminist geographers have stressed that human security makes visible vulnerable groups that do not pop up on the radar of narrowly state-centric and military conceptions of security. It is through human security that the politics of the family and of the gendered body can emerge to challenge hegemonic and gendered constructions of global security. More recently, there has been concern among geographers that the concept of human security is losing its emancipatory edge by increasingly speaking truth for rather than to power. Indeed, human security has the potential to be conscripted to power as a legitimation for humanitarian interventions into sovereign states that are deemed particularly hazardous to human security, as happened in the 1999 NATO bombing campaign on Yugoslavia.

By the end of the Cold War a number of incidents, including the Chernobyl disaster, had made it apparent that the world was facing new and human-made threats, which were neither of a military nature, nor could they be countered at a territorial level, but which were nevertheless commonly framed as security issues. In the mid-to late 1990s, these transformations prompted work around the concept of securitization. Securitization tries to capture the dynamic process by which previously merely political issues are turned into security issues (Buzan, Wæver, and De Wilde 1998). Securitization occurs when agents utter “security” and successfully move an issue from the realm of normal politics into that of panic politics in order to authorize exceptional measures that were previously legally or ethically impossible. In order for a securitization to be successful, these agents (elites, bureaucracies, scientists, companies, NGOs) make a case that although the survival of some vital referent object (the state, the environment, the economy, an identity) is threatened and a point of no return is about to approach, the issue is also resolvable. They will call for exceptional measures to be implemented at global, regional, national, or local scales, depending on the way a threat is framed as a security issue. Securitization has remained a useful tool for geographical research, even though critics have held that the concept lacks an explicitly political agenda and remains too analytically static to capture the complex interplay of twenty-first-century globalization and geopolitics.
At the same time as the concept of securitization started to gain academic impetus, other researchers suggested that the broadening of the security agenda to encompass nonmilitary threats should best be read as a new paradigm—risk (Ó Tuathail 1998). They proposed that as an industrial society organized around scarcity was making way for a wealthy society geared toward the management of hazards, the logic of security was slowly being replaced by one of risk. The government of risk, which is seen to have taken a foothold in Anglo-American societies in particular, challenges the logic of security in a number of ways. First, threats such as environmental disasters and terrorist attacks are now seen as deterritorialized and therefore no longer constrained by a particular territorial state. Moreover, rather than emanating from a foreign and threatening external space, they emerge from the midst of modern society. Second, such hazards are no longer immediate but long term, constituting threats to survival only in the distant future. Finally and most importantly, a state that operates in a modality of risk tries to manage hazards rather than removing them. It construes threats as being of such a nature that they cannot simply be eliminated but must instead be controlled and minimized through a complex set of precautionary governmental practices. In sum, this calculation of risk is not so much targeted at the causes of insecurity but at governing the future through predictive modeling. While earlier security discourses strove to eliminate insecurity, the logic of risk tolerates low levels of insecurity while trying to improve the resilience of modern society at a distance. As deterritorialized threats are unlikely to disappear in the foreseeable future, policy can only aim to reduce their frequency and impact. Under the government of risk, previously exceptional measures are made permanent and unexceptional, thus becoming normalized, everyday, and often unnoticeable.

This invisibility of security links up the literature on risk with that of securitization. Arguably, the calculation of risk becomes the dominant governmental rationality in a society that is already saturated with the logic of securitization. While the defenders of the concept of securitization understand security as a cycle in which issues become both securitized and desecuritized (moved off the security agenda), others have been more skeptical, arguing that security is a logic that subsumes democratic life through the progressive normalization of exceptional measures. This becomes clearest in a biopolitical age, which can be seen to emerge when the modern state starts to intervene in its population’s life, its health, welfare, productivity, and even sexuality. This biopolitical mode of power is closely connected to the rise of the welfare state and the idea of social security. As theorists of biopolitics point out, however, there is a dark side to a form of power in which the preservation of life may legitimate death. This often happens when states establish legally exceptional spaces through emergency laws that allow violence to be inflicted on bodies that have their rights as political subjects removed from them, as witnessed in the global war on terror (Gregory 2007). Geographers have been particularly visible in tracing these transformations in five areas: border, urban, energy, environmental security, and biosecurity.

Security issues

The border is often seen to have originated with modern territoriality and the creation of sharp outer boundaries that distinguish one sovereign state from another. Borders have been
securitized for a number of reasons, including military conflicts, but only in the cause of the twentieth century have they become governed through technologies such as fences, watchtowers, passports, and computers. Today the main purpose of borders is to mediate between two of society’s fundamental desires: mobility and security. Borders increasingly function as membranes that filter wanted mobility (tourism, business) from the unwanted (“illegal,” criminal, or terrorist migration). In the biopolitical age this is done through techniques of racial profiling, biometric data surveying, and the import of military technology at border crossings (Amoore 2006). There is a clear identity dimension to border security in its materialization and expression of anti-immigrant discourses that are home to the most advanced capitalist societies. The terrorist attacks of 9/11 led to an upsurge of border securitization, most visibly at airports. Although these were initially met by public outrage, they were subsequently normalized and increasingly seen as a necessary part of travel.

Another and closely related issue that geographers have explored is the securitization of urban space. In response to a perceived vulnerability to global terrorism, modern cities have seen a number of profound changes to their socio-material fabric, including gated communities, mobile checkpoints, and the explosion of CCTV technology. In the same way as the biopolitical border tries to filter wanted from unwanted migration, the resilient city tries to separate out “good” from “bad” urban circulation. The reconfiguration of the built environment according to principles of counterterrorism is no recent trend and many of its policies were previously tried out in urban warzones and divided cities. Although they are fueled by discourses of insecurity that rely on traditional inside/outside binaries, these transformations of urban space are also integrated into the calculation of risk and the way cities market themselves to investors as being secure. In the mid- to late 2000s, cities in the Global North increasingly became the target of policies that tried to create defensible space in ways that were inconspicuous. Working at the design stage, this invisible fortress urbanism introduced discreet but permanent fortifications into urban life, from concrete flower pots to collapsible pavements, all of which tried to provide security without a climate of fear (Coaffee, O’Hare, and Hawkesworth 2009). These debates have drawn attention to the way in which new security architectures manifest themselves materially in the urban landscape.

In recent years, journalists, scholars, and policymakers have self-evidently treated energy as a security issue. Yet, despite the prominent place that energy has had within twentieth-century geopolitical writing, military strategy, and logistics, it is especially since the 1970s and even more so since the 2000s that energy supply has become framed as a core matter of national security in many states of the Global North. This is due to a perceived scarcity of hydrocarbons, a concentration of the remaining fuels in states that are typically thought of as unstable or undemocratic, and the fear of piracy along key sea routes. Human geography has done much to interrogate the imaginative geographies behind this concern with natural resources and to problematize deterministic readings of global energy politics as a “scramble for resources.” Whether energy is securitized is very much a question of the local context. The 2003 US-led invasion of Iraq was not justified by a securitization of Iraqi oil but had to rely on other sources of public legitimacy. Nevertheless, despite this absence, the securitization of energy was very visible at the level of everyday practices, as research on the sport utility vehicle in the United States has shown (Campbell 2005).
Debates on the environment in the 1990s asked whether it qualified as a security issue and in what ways. While in the early 2000s insights from critical security studies and critical geopolitics were imported into the debate to unsettle some of its underlying assumptions (Dalby 2002), discussions in the late 2000s turned to questions of causality, calculation, and prediction as well as to the ethics of framing the environment as a security issue. In doing so, these discussions focused less on nuclear hazards and increasingly on the issue of human-induced global warming. Environmental security therefore tends to deal with the impact that drought, flooding, storms, and the melting of Arctic sea ice have begun to have on the security of humans and nonhumans. Particular emphasis has been placed on resilient infrastructure, which often determines whether natural disasters turn catastrophic. Similarly to energy security, much of the media and policy debate on climate change has focused on whether global warming will increase the likelihood of armed conflict in the future. Scholarly research has cautioned against overly deterministic arguments that do not take into consideration the complex ways in which climate change impacts on conflict landscapes. Crucially, it has shown the analytical limits of the concept of securitization in accounting for environmental politics, questioning whether the environment is best served by securitization and the logic of national security, geopolitics, and militarization that tends to accompany it.

The final and perhaps most recent area of concern is biosecurity. Biosecurity deals with the protection of human from nonhuman life, particularly pandemics, which have in the past included the plague, HIV/AIDS or the H1N1 virus. Human geographers have paid particular attention to the way in which global health security has operated in an unequal neoliberal political economy and tends to be scripted by the states of the Global North. Debate has focused on international nodes like airports and seaports at which governments seek to prevent the spread of diseases through emergency measures. It has also interrogated the (often racist) geographical imaginations that operate behind these policies. Pandemics are often constructed in similar ways to terrorism. On the one hand, they are framed geopolitically as emerging from far away and threatening places in the Global South or the Far East that need to be contained, controlled, or disciplined. On the other hand, they are treated as externalities of modern global mobility that emerge from the very target of biopolitical governance – the population. This split is particularly visible in debates around bioterrorism. Crucially, biosecurity is a governing rationality that seeks to anticipate the evolution of biological forms that are yet to emerge (Hinchcliffe and Bingham 2008). It therefore treats these organisms as dynamic threats that cannot be eliminated but only managed in the way that other risks are. While this governing of the future is long term, the outbreak of pandemics can be sudden and therefore potentially rather easy to securitize. The geographical literature is undecided as to how successful the securitization of pandemics like HIV/AIDS has been in the past, and skeptical as to whether it is ethically desirable.

**SEE ALSO:** Biopolitics; Biosecurity; Borders, boundaries, and borderlands; Energy security; Environmental (in)security; Environmental risk analysis; Urban politics

**References**

SECURITY


Sensor networks, the sensor web, and the Internet of Things

Steve H.L. Liang
University of Calgary, Canada

Neil Gross’s article, “The Earth Will Don an Electronic Skin,” provides a compelling explanation of the sensor networks and sensor web vision: “In the next century, planet Earth will don an electronic skin. It will use the Internet as a scaffold to support and transmit sensations. This skin is already being stitched together. It consists of millions of embedded electronic measuring devices: thermostats, pressure gauges, pollution detectors, cameras, microphones, glucose sensors, EKGs, electroencephalographs. These will probe and monitor cities, endangered species, the atmosphere, our ships, highway traffic, fleets of trucks, our conversations, our bodies–even our dreams.” Figure 1 illustrates the concept of the sensor web as an electronic skin of planet Earth.

The scope of this entry is to provide an architectural overview and research challenges of the sensor networks and sensor web. As sensor networks and sensor web technology become more widely available, they are becoming essential tools for all GIScientists and geographers to collect, analyze, and share observation data. This entry presents the architectural components involved in an end-to-end sensor web system, from the sensor nodes in the field to the interoperable sensor web services on the web. Design tradeoffs and research challenges are also introduced.

The remainder of the entry is organized as follows: In the section “Sensors,” an overview is first given of the sensor network technologies, with a focus on the wireless sensor networks (WSN). At the end of the section, the sensor networks long tail issue is introduced leading to a discussion of the need for a sensor web infrastructure to capture the sensor networks long tail. In the section “Sensor web,” the OGC sensor web architecture is described and the open problems from a big sensor data management perspective are presented. The next section presents the next generation sensor web: the Internet of Things. This is followed by a conclusion.

Sensor networks

Sensors

A sensor is a device that is capable of observing a phenomenon and returning an observed value. A sensor detects and responds to physical stimuli, such as movement, light, heat, and so on. Today various physical, chemical, and biological properties can be measured and monitored by in situ sensors or remote sensors. Sensors are everywhere and include satellite-based sensors providing multispectral information about the Earth’s surface (imagery, land cover, vegetation indices), airborne sensors for detailed imagery but also for laser scans of physical or human-constructed structures (LiDAR), and sensors near, on, or under the Earth’s surface measuring anything from physical characteristics (pressure, temperature, humidity) and phenomena (wind, rain, earthquakes), to the tracking of animals, vehicles, and people.
Technical advances are allowing sensors to be smaller, lighter, and more energy efficient.

**Sensor networks or networked sensors**

A sensor network (SN) is a computer network consisting of spatially distributed sensors with the purpose of cooperatively monitoring physical and environmental conditions. In addition to being comprised of one or more sensors, each node in a sensor network is typically equipped with a communication device which incorporates means of transmitting data, from wires to cellular phones to microwave radios. Figure 2 shows a spectrum of sensor network devices, from large, expensive satellites, to tiny, cheap RFID (radio-frequency identification) tags.

Today, it is feasible and economically viable to deploy a large number of sensor networks to continuously monitor our environment. In fact, many sensor networks, including both in situ and remote, have been built and deployed over the past decade. The combination of these heterogeneous sensing systems can provide enormous amounts of timely, comprehensive, continuous, and multiresolution observations for applications, including environmental monitoring, underwater monitoring, habitat monitoring, and battlefield and disaster management. Examples of large-scale operational sensor networks include the regional-scale NEPTUNE (North-East Pacific Time-series Undersea Networked Experiments) ocean observatory network monitoring long-term changes at water depths ranging from 17 to 2660 m and the global-scale Argos network of buoys measuring temperature and salinity of the world’s oceans.

Sensor networks can also be mounted on mobile platforms including, for example, aerial tramways, robots, unmanned aerial systems, and satellites.

**Wireless sensor networks**

Recently, a new breed of sensor networks, called wireless sensor networks (WSN), is seamlessly immersing itself into our daily lives. A WSN is a network of sensors, each with an embedded processing unit and a wireless communication device. Each sensor is placed into the physical world and interacts with its environment. New technologies, such as microelectromechanical systems (MEMS), allow the WSN nodes to be constructed at the micro- or even nano-scale, which results in greater portability and flexibility than is otherwise possible.

Worthy of note, not all sensor networks are wireless sensor networks and many references do not differentiate between the two. This entry considers any sensor network that consists of sensors with networking capabilities, and they can use wire, wireless, or even underwater acoustic communication protocols. Wireless sensor networks means the WSN nodes that are described in this section.

A WSN node consists of four basic components as shown in Figure 3: a sensing unit, a processing unit, a transceiver unit, and a power unit. Many applications require additional
modules such as a location-finding unit and a power generator.

Sensor units usually have two subunits: sensors and analog-to-digital converters (ADCs). The analog signals produced by the sensors based on the observed physical phenomenon need to be converted to digital signals by the ADC, and then sent to the processing unit. The processing unit is programmable and provides computation, storage, and bidirectional communication with other nodes in the system. The processing unit interfaces with the sensing unit, performs basic signal processing (e.g., simple translations based on a sensor’s calibration curve), and disseminates the data according to the objective of the sensing task. Individual WSN nodes use the transceiver unit to communicate and coordinate with one another, as it is not feasible to bring direct Internet connectivity to each WSN node—the Internet communication module is too costly and it consumes too much power. The WSN nodes form a multihop network by forwarding each other’s messages, which can greatly extend the coverage of the monitoring system.

Through the exchange of messages, the WSN network can also perform in-network processing, for example, reporting the average temperature across a region or monitoring spatial change over time (Duckham 2012). This flexible communication structure allows us to
produce a network that performs decentralized sensing and computing tasks. Such in-network processing capability is a key differentiator between traditional data logger–based monitoring systems and WSN–based systems. Interested readers should refer to Duckham (2012), which provides an in-depth foundation of the methodologies and algorithms of such decentralized computing in a geospatial context.

Each WSN node needs a WSN operation system (OS) to manage the above hardware components. Subject to the resource constraints of typical WSN nodes, these nodes cannot use traditional OSs, such as embedded Linux. They require very specific operation system designs on architecture, programming model, scheduling, memory management and protection, communication protocols, resource sharing, and support for real-time applications. Existing WSN OSs include TinyOS, Contiki, MANTIS, and so on. Farooq and Kunz (2011) provide a comprehensive survey on the current state of the art in WSN OSs.

When considering deploying a WSN for applications, it is important to know the necessary components in the system. Figure 4 shows a typical tiered architecture of a WSN–based monitoring system, and this architecture has been widely deployed in field applications (Mainwaring et al. 2002).

The lowest level consists of the WSN nodes that are able to perform general purpose computing and networking in addition to sensing tasks. These small, battery–powered nodes are placed in areas of interest and embedded in the environment. Because they are placed very close to the observed phenomenon, these nodes can often use small and inexpensive off–the–shelf sensors. These WSN nodes as a whole can achieve high spatial resolution through dense deployment. Compared with traditional data–logger approaches, which use a few high–quality sensors with sophisticated signal processing, a WSN–based monitoring system provides high fault–tolerance against occlusions and component failures.

![Figure 4](image-url) A tiered architecture of a WSN–based monitoring system.
Ultimately, data from each WSN node needs to be propagated to the Internet. Each local wireless sensor network has at least a gateway, which often has more computing, storage, and energy capability (e.g., more storage space for caching the historical data from the WSN nodes and larger battery and solar panels). The gateway can communicate with the local wireless sensor network and provides connectivity to the transit network. Depending on the environmental constraints, each transit network design may have different characteristics with respect to expected fault-tolerance, energy efficiency, cost, real-time requirements, and manageability. For example, the transit network may use a single-hop wireless communication protocol to reach a long range or adopt a series of WSN nodes in a path from the gateway to a base station. Finally, the base station connects to sensor web service servers to place the sensor measurements in interoperable, reusable, and machine-readable web service interfaces and data formats.

When choosing a wireless sensor network for field deployment, the following list of performance metrics needs to be considered (Tilak, Abu-Ghazaleh, and Heinzelman 2002). In fact, these performance metrics are also the main research challenges in developing and deploying wireless sensor networks.

- **Energy efficiency/system lifetime.** Nodes in wireless sensor networks are powered by batteries and in some cases with rechargeable batteries and solar panels. As a result, these nodes must be energy efficient to maximize system lifetime. The lifetime of a single WSN node is in fact not the measure of the system lifetime, as a WSN is a cooperative sensor network. A WSN’s system lifetime can be measured by generic parameters, such as the time until half of the nodes in the system die, or by application-directed metrics, such as when the network stops providing the application with the desired information about the phenomena.

- **Latency.** Latency is defined as the time duration between the time instant when the observation happens and the time instant when the observation result is delivered to the interested user or application. Due to the high energy costs of the wireless communications between the nodes, there is a tradeoff between latency and energy efficiency.

- **Sampling frequency.** Obtaining accurate information is the primary objective of a sensing task. Higher sampling frequency can improve the accuracy because it reduces the need for interpolations between discrete data points. However, there is a tradeoff between sampling frequency, latency, and energy efficiency. A WSN should be adaptive so that the intended application obtains the desired accuracy with acceptable delay and with minimal energy consumption. For example, a sensing task can achieve the same objective by either requesting more frequent data dissemination from the same sensor nodes or distributing data dissemination from more sensor nodes with a lower sampling rate.

- **Fault-tolerance.** A WSN node may fail due to surrounding physical conditions (e.g., in a wildfire scenario) or when its batteries run out. It may be difficult or not economically feasible to replace existing nodes. A WSN must be fault-tolerant such that noncatastrophic failures can be hidden from the sensing task. Fault tolerance can be defined as the ability to sustain sensor network functionalities without any interruption due to sensor node failures. One effective strategy to achieve fault-tolerance is data replication. However, replicating data in different nodes...
requires energy. There is again a tradeoff between data replication and energy efficiency. A data replication strategy should be sensing-task-specific. For example, WSN nodes can only share and replicate higher priority data and neglect other data.

- **Scalability.** Depending on the application, the number of WSN nodes may reach an extreme value of millions. When the number of nodes in the network grows, the communication frequency between the nodes may also grow, thus consuming more energy and reducing system lifetime. For large-scale networks, it is likely that localizing interactions through a structured topology and aggregation will improve scalability.

Eventually, the observations collected by the sensor networks need to be sent to a data repository via the Internet so that they can be used by applications. However, from a data management perspective, we find that the so-called long tail phenomenon exists in sensor networks. A long tail distribution means a large number of occurrences are far from the “head” or the central part of the distribution (hence the long tail). Anderson (2006) popularized the concept of the long tail in his book *The Long Tail: Why the Future of Business Is Selling Less of More*, in which he mentioned Amazon.com as an example of a business applying this strategy in the online bookstore business. For example, before the establishment of online bookstores customers could only purchase books from physical bookstores. However, physical stores tend to concentrate on offering popular book titles, that is, focusing only on the head of the long tail. Less popular books are commonly unavailable from physical bookstores. Customers may not be aware of books that they might be interested in because they cannot see them on the shelves in their local bookstores. Such a long tail has been found in a variety of businesses. Similarly, sensor networks also exhibit such long tail properties.

Today most researchers rely on the sensor data portals hosted by a few large-scale government and research institutes (e.g., World Meteorological Organization (WMO), Global Earth Observation System of Systems (GEOSS), National Oceanic and Atmospheric Administration (NOAA), National Aeronautics and Space Administration (NASA), and Environment Canada) to search for and download sensor measurement data. The number of these portals is relatively small, but these portals provide substantial amounts of continuous and long-term sensor measurements. These sensor data are quality controlled, calibrated, indexed, stored, and published through the web. We can call these sensors and their data the “head” in a sensor networks long tail graph, and the head of sensor networks is currently accessible to the researchers. However, the number of local-scale sensor networks, such as the WSN nodes deployed and operated by individual scientists, has grown exponentially in recent years. These sensors/data are not accessible to most researchers. Heidorn (2008), in his article “Shedding Light on the Dark Data in the Long Tail of Science,” described this particular troublesome class of “Dark Data.” They are not indexed nor stored at the centralized sensor data portal, so they become nearly invisible to most scientists and other potential users. These sensors and data are more difficult to find and less frequently reused than those in the head of sensor networks. As a result, they are more likely to remain underutilized and eventually become lost. Consequently, there is a strong need to integrate these heterogeneous sensor networks into a coherent and interoperable global observation system, and such a system is termed the “sensor web.”
Sensor web

While sensor networks are primarily concerned with interconnecting the physical sensor nodes (e.g., the networking and in-network processing aspects of a monitoring system), the sensor web is mainly concerned with interconnecting the observation data and applications (e.g., interoperability and innovation-in-assembly). However, integrating diverse sensors into a coherent and interoperable observation system is complicated. Sensor network nodes are computers in the field. The design and deployment of the sensor networks are constrained by the application and the physical environment. Different environmental conditions in the field represent different challenges (e.g., remote location of the sensors, power constraints, extreme weather conditions, etc.). Different challenges lead to different designs and tradeoffs. As a result, these specialized sensor network systems have become islands of information systems. A coherent information infrastructure is needed to treat sensors in an interoperable, platform-independent, and uniform way.

The concept of the sensor web reflects such a kind of infrastructure for sharing, finding, and accessing sensors and their data across different applications. It hides the heterogeneous sensor hardware and communication protocols from the applications built on top of it. Similar to the World Wide Web (WWW), which acts essentially as a worldwide computer, the sensor web can be considered as a worldwide sensor. In the sensor web, users, applications, and sensors can access, as a single unit, vast amounts of data and processing power from a large number of widely distributed, heterogeneous WSNs or individual networked sensors.

Some references refer to this system as the worldwide sensor web, because the connection among these sensor networks is typically achieved via the WWW. It is worth noting that the concept of the sensor web is evolving, as evidenced by the wide range of sensor web definitions (Gibbons et al. 2003; Liang, Croitoru, and Tao 2005), and some references do not differentiate between the sensor networks and the sensor web. The key differentiator between the two is that while sensor networks research is mainly concerned with the networking and in-network processing aspects of sensing systems, sensor web research is mainly concerned with the data management and data interoperability aspects of sensing systems.

Open Geospatial Consortium Sensor Web Enablement (OGC SWE)

Similar to many global-scale information infrastructures, such as the WWW based on the World Wide Web Consortium (W3C) open standards, the sensor web also requires open standards to fulfill the sensor web vision. The Open Geospatial Consortium (OGC) Sensor Web Enablement (SWE) standards are the fundamental building blocks for the worldwide sensor web. The first version of the SWE specifications was approved as standards between 2006 and 2007 and the second version was approved between 2012 and 2013. They offer the following functionalities.

- Description of the observations collected by sensors to enable further processing.
- Description of sensor metadata including properties and behavior of sensors, as well as correlating reliability and accuracy of collected measurements.
- Access to observations and sensor metadata based on standardized data encodings and appropriate query and filter mechanisms.
- Tasking of sensors for the acquisition of observations.
To provide the above-mentioned functionalities, OGC SWE consists of two layers of standards. Each layer of standards deals with a “level of interoperability” issue. The two layers are: (i) SWE information models, and (ii) SWE service interfaces. The SWE service interfaces provide a standard way for sensor networks to access and exchange resources over the web, while the SWE information models provide a standard way to describe the information exchanged between the sensor networks. The two layers of the SWE framework are detailed below.

The purpose of the SWE information model is to allow applications to understand and use sensor networks data.

In SWE information models, the sensor web consists of three elements: (i) Observation, (ii) Sensor, and (iii) Phenomenon. It seems obvious that these are three different things; however, people often tend to use them interchangeably. The concepts underlying these entities are defined by OGC as follows:

1. **Observation**: an act of observing a phenomenon, with the goal of producing an estimate of the value of the phenomenon.
2. **Sensor**: an entity capable of observing a phenomenon and returning an estimated property value of the phenomenon.
3. **Phenomenon**: an event or physical property that can be observed or measured.

The relationship between these can be described using the following description: **Observations Use Sensors to Observe Phenomena**. Figure 5 illustrates the relationship between the three basic entities of SWE information models.

With the major entities and relationships identified, OGC then defines corresponding standard information models and encodings for the above three major sensor web entities.

The ISO/OGC observation and measurement (O&M) is the core of the sensor web, and it provides a conceptual model to represent data of spatiotemporal observations. Moreover, Sensor Model Language (SensorML) provides a generic model and XML encodings to describe both in situ and remote sensors and related processes. Regarding phenomenon-based models, since there exist ontological definitions for phenomena in different domains, SWE leverages the existing standard efforts, such as the SWEET ontologies for earth sciences (https://sweet.jpl.nasa.gov/) or the IOOS ontologies for marine sciences (https://ioos.noaa.gov/).

In addition to O&M and SensorML, there is the SWE Common standard describing common and basic data types used throughout the SWE framework. As SWE Common is primarily an implementation standard describing syntaxes and encodings, interested readers can refer to the OGC SWE Common 2.0 standard specification.

In the OGC O&M standard an observation is defined as an event of observing a phenomenon, with the goal of producing an estimate of the value of the phenomenon. O&M is a standard conceptual model and physical encoding for observations and measurements.

One misconception about the sensor web is that we tend to think “sensors” are the major interest of the sensor web. The name “sensor” web easily misleads us to view it from a...
sensor-oriented perspective. However, here it is argued that it is in fact the “observations” that are the major focus of the sensor web. Most of the sensor web applications are more concerned with observations than with the sensors that make the observations. Therefore, a standard model for the observations, such as O&M, is in fact the core of the sensor web. Figure 6 shows the basic observation model of O&M.

In the O&M model, an observation should at least have the following six mandatory properties. Each property is necessary in order to make an observation useful.

1 **Feature of interest.** Feature of interest is an identifiable object that is associated with the observed phenomenon. An observation results in a value being assigned to a phenomenon. The phenomenon is a property of a feature, the latter being the feature of interest of the observation. Feature of interest is a critical concept that allows O&M to be a generic model accommodating observations made by both in situ and remote sensors.

2 **Phenomenon/observed property.** Phenomenon is the observable property of the physical entities measured and quantified by the observation. Examples of a phenomenon can be temperature, humidity, or chemical concentration.

3 **Procedure.** The observation uses a procedure, which is often a sensor but may be a computing model or a simulator. For example, a thermometer can be the procedure for performing a temperature observation.

4 **Results.** An observation result is an estimate of the value of a phenomenon. This is the goal of an observation. A result is not only a value. It consists of two parts: (i) result value, and (ii) unit of measure. For example, 15°C

---

**Figure 6** Basic observation model of O&M. Bröring et al. 2011. Reproduced from MDPI.
is an instance of an observation result, while 15 is the result value and °C is the unit of measure.

5 Time. There are three temporal attributes for observations. An observation has a result time referring to the time when the observation’s result became available. Then an observation has a phenomenon time referring to the time instant or period when the result applies to the property of the feature of interest. For some observations, their result time is identical to the phenomenon time. However, there are important cases where they differ. Take water quality monitoring as an example. Where a measurement is made on a water sample in a laboratory, the phenomenon time is the time the water sample was retrieved from a river or a lake, while the result time is the time the laboratory procedure was applied. In some cases, an observation may have a valid time describing the time period during which the result is intended to be used. Valid time is commonly used in forecasting applications.

6 Location. An observation does not have an inherent location property. An observation’s principal location of interest is usually associated with the ultimate feature of interest but sometimes is associated with the observation procedure, according to the specific scenario.

The object-based model (or the ISO OGC feature model) and the field-based model (or the ISO OGC coverage model) are two common models of geographical information. O&M provides an additional view, an observation-based view of the geographical information. It is worth explaining the relationships and differences between these models or views.

ISO 19109 describes the term “feature” as a “fundamental unit of geographic information,” and in many GIScience literatures, features are called objects. The general feature model (GFM) presented in ISO 19101 and ISO 19109 defines a feature type in terms of its characteristic set of properties, including attributes, association roles, and behaviors, as well as generalization and specialization relationships, and constraints. The feature model presents a viewpoint of the world in terms of the set of discrete identifiable objects that occupy it.

ISO 19123 describes the coverage model as a principal alternative model for geographic information, and in much GIScience literature coverages are referenced to as fields. This viewpoint presents a perspective of the world that is property focused, highlighting the variation of a property within the (spatiotemporal) domain of interest. These viewpoints are not exclusive, and both are used in analysis and modeling.

ISO 19156 describes observations concerned with the data collection event. The purpose of an observation event is to assign a value to a property of a feature. The results of a collection of observations of the different properties on a feature can provide a complete description of the feature (or object) instance. Alternatively, the results of a collection of observations of a specific property on a collection of different features provide a discrete coverage (or field). The other properties of the observation are the important metadata (such as time, unit of measurement, quality, etc.) concerning the estimation of the value(s) of a property on a feature of interest.

In particular, observations are concerned with properties (e.g., temperature, humidity, status of on/off) whose values are determined with an identifiable procedure (e.g., a sensor), in which there is an uncertainty in the estimated result. The metadata provided by the observation instance are necessary in order to use the observation result for further analysis.

In addition, we can use the example data table of Figure 7 to illustrate the relationships among
the three models or viewpoints. The table is a dataset comprising values of a set of properties (e.g., Value\textsubscript{n}m) at a set of locations (e.g., x\textsubscript{n}, y\textsubscript{n}). A row of the table offers the properties of a single location and this row can be considered as a representation of a feature (or an object). A column of the table provides the variation of a specific property across the set of locations and this column can be considered as a representation of a coverage (or a field). A single cell in the table provides the value of a single property on a single feature. In the case that the value is the result of an observation event, this cell and the associated metadata form an observation of the O&M model.

A sensor is capable of observing a phenomenon and returning an estimated value of the phenomenon. SensorML is both a conceptual model and a physical encoding for sensors. SensorML provides a framework within which the geometric, dynamic, and observational characteristics of sensors and sensor systems can be defined.

There is a great variety of sensor types, and SensorML established a common framework to accommodate these different types of sensors. The key concept of the framework is the abstract model of processes as the atomic unit for modeling different types of sensors. SensorML provides the following functionalities:

1. Supporting the discovery of sensors by providing a means for encoding sensor metadata.
2. Providing information that can be used to interpret and analyze data produced by the sensor (e.g., the parameters of the sensor calibration).
3. Allowing the description of post-processing steps applied to sensor data for data reconstruction in an a priori process of data creation.

In the SensorML model, a process (a sensor is a special case of process) comprises the following components.

1. **Inputs, outputs, and parameters.** A process takes one or more inputs and, based on parameters of a methodology, generates one or more outputs. Inputs and outputs represent “ports” where outside processes exchange data with this process. Parameters can be considered as a special type of input, which is specifically required by the methodology of the process.

2. **Position.** Positions provide the location and orientation of a process at some discrete moment in time. Positional information is critical for the use of sensor observations. Positional information includes not only location but also orientation. Orientation is especially important for remote sensors. For example, knowing in what direction a traffic camera sensor is facing is vital in order to make use of the images taken by the camera. For mobile sensors, positional information is temporal in nature because positional information changes over time when the sensor moves.

3. **Metadata group.** The purposes of the metadata group are (i) to support discovery of resources, (ii) to qualify process results, and (iii) to assist human understanding. These metadata include identifiers, classifiers, constraints, capabilities, properties, contacts, documentation sources, and history.
The OGC Sensor Observation Service (SOS) provides standardized access to observations and sensor metadata of all sensor systems. Sensors can be characterized in at least three different ways: (i) proximity to the phenomenon being sensed: that is, remote versus in situ; (ii) mobility: that is, fixed versus mobile; and (iii) duration/lifetime of the sensor: that is, temporary/disposable versus permanent. SOS's extensible interface is able to hide the heterogeneity of the different sensors above. SOS provides two major functionalities: (i) it fetches observations according to spatiotemporal queries, and (ii) it describes the sensor according to spatiotemporal queries. SOS is not just a web service for users to download observations. SOS can filter out observations according to constrained queries. The query for observations or sensors can include one or many of the following constraints: (i) a specific sensor that reports the observations; (ii) specific time period(s) for which observations may be requested (can be real-time, historical, or even forecasting data); (iii) a specific phenomenon that is being observed by the sensor; (iv) a specific geographical region that contains the sensor; and (v) a specific geographical region that covers the feature of interest of the observations. Detailed operations and schemas can be found in the OGC SOS specification. Bröring et al. (2011) offer a comprehensive review of the second version of the OGC SWE specification (including SOS 2.0) with a focus on the improvements from version 1.0 to version 2.0.

The difference between SOS and WFS (Web Feature Service) needs further clarification because the two share several similarities and there are existing implementations that use WFS to serve sensor observations. The major difference between the two is that while WFS uses a very flexible modeling approach to encompass a real-world entity, SOS defines a focused common model for sensors, observations, and phenomena. WFS uses Geography Markup Language (GML) application schemas to define the specific properties of each type of feature. Clients of WFS in a particular domain must have a priori knowledge of the application schemas used in that domain. In contrast, SOS models are not domain-specific, and the client does not need a priori knowledge of domain-specific application schemas. For applications concerned with sensor observations, using SOS improves the applications' interoperability.

The purpose of Sensor Planning Service (SPS) is to provide a standard web service interface for collection assets and for the support system that surrounds them. A collection asset is an available means of collecting information. Collection assets include, but are not limited to, sensors or actuators. It can also be a simulation system, or any other information-gathering asset. SPS provides three major functionalities: (i) to assist in making task feasibility plans, (ii) to execute feasible task requests for collection assets, such as a sensor or sensor networks, and (iii) to query where the measured data can be accessible after the completion of the task. A typical use case for an SPS is updating a sensor's sampling frequency or the flight path of an unmanned aerial vehicle.

Organizations and research projects around the world are building the worldwide sensor web by establishing large-scale sensing systems with the SWE standards. The number of successful deployments demonstrates the maturity and suitability of the SWE standards in real-world applications. In 2013, the UK Ministry of Defence (MoD) funded an assessment of the maturity of implementations of OGC SWE standards (Percivall 2013). Bröring et al. (2011) and Percivall (2013) both offer a nonexhaustive list of the SWE applications and research projects around the world.
Sensor web research challenges

The following discussion highlights some open challenges and future work in the area of the sensor web. Corcho and García-Castro (2010) offer their view on the research challenges from a semantic sensor web perspective. Bröring et al. (2011) discuss some open research challenges from an interoperability perspective. Below, this entry will focus on some open problems in this research field from a big sensor data point of view.

The term “big data” is used to describe datasets that are too large or too complex to manage and process with traditional database management systems (DBMS) and data processing applications. The most widely recognized model for big data is the 3Vs model. The 3Vs model defines big data as data that are large in volume, velocity, and variety. And each “V” would pose different data management challenges. Sensor data in fact are considered a typical big data source with all of the 3Vs data challenges. Below the 3Vs model is used as a framework to discuss the sensor web research challenges.

**Volume:** One characteristic of big data is the large data volume, which can mean large size or large number of data records. While the data volume in social media is enormous, some even predict the size of sensor data will be 10–20 times that of social media. As the number of sensors and the size of the sensor data grow exponentially, how to store, transmit, process, analyze, and visualize sensor web data in large volumes is one of the major challenges. New big data architectures have been proposed and developed, including the MapReduce framework, Apache Hadoop and related projects (Shvachko et al. 2010), the Lambda architecture, and Storm (Marz and Warren 2015), and so on. However, these new frameworks are not specifically designed for spatiotemporal datasets. As the “spatial is special” argument is especially valid in the sensor web context, specialized spatial algorithms and architectures need to be proposed and developed.

**Velocity:** The velocity characteristic refers to the rate at which data are produced. Unlike the human participants in social media, the sensors in the sensor web can produce data in very high frequencies as long as they have a sufficient power supply. For example, Boeing jet engines produce 10 terabytes of sensor data every 30 minutes during flight. As a result, efficiently processing the high-velocity sensor data streams becomes very challenging, especially if we consider the geospatial nature of the sensor data. Many existing geospatial algorithms and architectures need to be updated in order to deal with the high volume combined with the velocity of the spatiotemporal sensor data streams. For example, the traditional request/response client-server architecture and DBMS may not be suitable for handling the high-velocity sensor streams. The publish/subscribe architecture and data stream management system (DSMS) is an alternative emerging data processing solution worth further investigations. Huang (2014) proposed GeoPubSubHub, a publish/subscribe architecture for the sensor web. The core of the GeoPubSubHub is the aggregated hierarchical spatial model (AHS) model, which is a new algorithm based on the DE-9IM designed specifically to efficiently determine topological relationships between continuous incoming sensor data streams and a large number of predefined queries (i.e., subscriptions). Huang scaled up the AHS model by developing the distributed AHS model based on MapReduce, a typical big data management methodology. The AHS model and the distributed AHS model are
examples of updating existing geospatial algorithms and architectures to handle the big data volume and velocity challenges.

**Variety:** The variety challenge includes non-aligned data structures, inconsistent data semantics, and incompatible data formats. In general, the variety characteristic refers to the differences between data records. In the context of the sensor web, the OGC SWE interoperability framework has been demonstrated to be useful and practical for addressing the sensor web variety issues, especially on observation exchange. However, while SWE provides a high-level semantic and syntactic interoperability framework, additional semantic annotations and domain-focus models are still required in order to improve the interoperability further. For example, Sheth, Henson, and Sahoo (2008) proposed that the semantic sensor web investigate the role of semantic annotation, ontologies, and reasoning to improve sensor discovery and sensor integration. The global hydrology communities demonstrated a great example by working collaboratively and defining domain-focus models for groundwater, surface water, and water quality observations based on the SWE framework. The result of the collaboration is the OGC WaterML standard and the OGC SOS hydrology profile (Valentine, Taylor, and Zaslavsky 2012). WaterML 2.0 was accepted as an official OGC standard in September 2012, then endorsed by the US Federal Geographic Data Committee (FGDC), and has been recommended for adoption as a joint WMO and ISO standard. Once officially adopted by WMO Commission for Hydrology, the hydrology communities around the world will have a truly global and interoperable framework on which to build a worldwide hydrology sensor web. More similar efforts are required in order to fulfill the worldwide sensor web vision.

For any large-scale distributed system (e.g., the WWW), both communication and data management fall back to the problem of resource discovery. Similarly, the sensor web needs an efficient and effective discovery mechanism, such as a registry, a search engine, or a recommendation engine, specifically for the sensor web. Jirka, Bröring, and Stasch (2009) developed a discovery mechanism in the sensor web, proposing two components, namely Sensor Instance Registry (SIR) and Sensor Observable Registry (SOR), offering mechanisms to search for sensors, exploit basic semantic relationships, harvest sensor metadata, and integrate sensor discovery into already existing catalogs. While the traditional centralized registry approach may be suitable for relatively static geospatial datasets and sensor networks, it may not be suitable for the always changing sensor networks and their data. New approaches need to be explored in order to address the dynamic and transient nature of sensor networks. For example, Liang and Huang (2013) proposed a hybrid peer-to-peer (P2P) architecture for sensor web research discovery, in which every sensor web service also serves as part of the sensor web service discovery infrastructure (i.e., a peer node). These nodes operate on a cooperative model, where each peer leverages the others’ available resources (i.e., central processing unit, storage, bandwidth, etc.) for mutual benefits. The decentralized design enhances the scalability of the sensor web to accommodate a very large number of sensors and users because it is more reliable due to the fact that it has no single point of failure and it achieves better load balancing by distributing the system load.

In contrast to the traditional top-down taxonomy-based registry approach, crowd-sourced folksonomy-based recommendation engines
recently emerged as an innovative approach to address the resource discovery issues in large networks. Some real-world examples include Netflix recommending movies, Delicious recommending URL bookmarks, and Flickr recommending images. A recommendation engine’s collaborative filtering framework can aid users to deal with the petabytes of data that are generated by the sensor networks. For example Rezel and Liang (2011) developed SWE-FE (folksonomy extension), a folksonomy-based recommendation system for the sensor web. SWE-FE is a collaborative tagging system that exploits the geospatial information associated with three key components: tags, resources, and users. SWE-FE is able to recommend new sensor networks and datasets according to the similarity of a user to the rest of the users in the sensor web system. In order for sensor web users to find relevant, fresh, and complete search results about their sensors and data of interest, new approaches that differ from traditional registries are needed.

The Internet of Things – the next-generation sensor web

In the past 50 years, information technology (IT) has twice radically reshaped every organization and business. During the 1960s and 1970s, the advent of computers brought the first wave of IT-driven transformation which automated individual activities in the value chain, from order processing and bill paying to computer-aided design and manufacturing resource planning. The birth of geographical information systems (GIS) is a typical result of the first wave of IT-driven transformation. In the 1980s and 1990s, the rise of the ubiquitous and inexpensive Internet unleashed the second wave of IT-driven transformation. The second wave broke the barrier of geography and enabled coordination and integration of the individual activities automated in the first wave. In a GIS context, web-based GIS, virtual globes, spatial data infrastructure (SDI), location-based services (LBS), and location-based social networks are some examples of the results of the second wave of IT-driven transformation. Following the advent of computers and Internet, the Internet of Things (IoT) is projected to be the third wave that will radically transform the world again (Porter and Heppelmann 2014).

In the near future, millions to billions of small sensors and actuators will be embedded in real-world objects and connected to the Internet forming the IoT. The basic premise of the IoT is that everyday objects or devices will be able to sense their environment, collect information, and communicate and interact with each other. By capturing this collective observation data, it will be possible for sensor-gathered IoT information to be accessed by multiple applications on multiple devices from any physical location. By embedding real-world sensor-based devices in our environment, the IoT is opening the door to limitless exciting possibilities.

Although the IoT is a new term, it shares a similar vision with the sensor web. Gubbi et al. (2013) provide comprehensive surveys and reviews of the IoT. As the IoT can be seen as an evolution of sensor networks and the sensor web, the IoT and the sensor web architecture stacks are very similar, if not identical. While OGC SWE has been demonstrated to be very effective for complicated sensor web applications, its complexity and the computational overhead may not be suitable for resource-constrained IoT devices. For example, Jazayeri, Huang, and Liang (2012) provide a comprehensive study and identify that the OGC SWE standards are too heavyweight and complicated for the resource-constrained IoT devices.

In June 2012, a new OGC Standards Working Group (SWG) was formed called the Sensor
Web Interface for Internet of Things (SWE-IoT) SWG; later the name was changed to the SensorThings API (application program interface) Standards Working Group. The OGC SensorThings API (Liang, Huang, and Khalafbeigi 2015) provides an open and unified way to interconnect IoT devices, observation data, and applications over the web. The SensorThings API is based on the comprehensive OGC SWE standards. The core of the SensorThings API is the ISO O&M model. The SensorThings API sensing profile is based on the OGC SOS and the tasking profile is based on the OGC SPS. The main difference between the SensorThings API and the existing SWE standards is that the SensorThings API is designed specifically for the resource-constrained IoT devices and the IoT application developers. As a result, the SensorThings API adopts the REST (representational state transfer) principle, the efficient JSON (JavaScript Object Notation) encoding, and the flexible OASIS (Open Artwork System Interchange Standard) OData (Open Data) protocol and URL conventions.

Conclusion

In 1991, Mark Weiser described his vision of ubiquitous computing: “The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.” The author believes this vision is also applicable to describing the sensor networks, sensor web, and IoT. Today, we give little thought to the modern electrical grid and what our lives would be like without it. Tomorrow we will give little thought to the IoT, which will have grown to impact every aspect of our lives. Sensor networks will weave themselves into everyday life until they are indistinguishable from it.

SEE ALSO: Big data; Location-based services; Service-oriented architecture; Spatial sampling; Traffic sensors; Web-mapping services

References


Internet of Things and wireless sensor network

The Internet of Things (IoT) was built on the concept of ubiquitous computing (Weiser 1993; Sakamura 1987; Abowd and Mynatt 2000; Hightower and Borriello 2001), or pervasive computing, that has been proposed for more than 30 years. IoT refers to the interconnection of a uniquely identifiable embedded computing device within the current or next generation Internet infrastructure (Atzori, Iera, and Morabito 2010; McFarlane et al. 2003; Abdelwahab et al. 2014). The Internet of Everything (IoE), proposed in 2013 (Cisco 2013), refers to bringing together networked connections of people, processes, data, and things. Figure 1 shows the architecture of a ubiquitous system.

Sensing and wireless communication are two important components of the IoT. A variety of mechanical, thermal, biological, chemical, optical, and magnetic sensors may be applied to measure properties of the environment. These sensors could work together for specific applications (Figure 2).

Wireless communications include wireless WAN (wide area network), wireless MAN (metropolitan area network), wireless LAN (local area network), wireless PAN (personal area network), and wireless BAN (body area network). Wireless communication standards are developed and maintained by the Institute of Electrical and Electronic Engineers (IEEE) 802 LAN/MAN Standards Committee (http://ieee802.org/).

As an important part of the IoT, wireless sensor networks (WSNs) have gained worldwide attention since the 1990s (Akyildiz et al. 2002; Yick, Mukherjee, and Ghosal 2008; Sohrabi et al. 2000; Akyildiz and Kasimoglu 2004; Shiro 2006). With the advancement of micro-electro-mechanical systems (MEMS) and wireless communication, low-power, low-cost WSN nodes equipped with multifunctional sensors, a processor, memory, power supply, radio, and actuators can be developed. Figure 3 shows the components of a WSN node.

Research on WSNs includes (i) the network platform, (ii) communication protocol, and (iii) network services, as discussed in the next section. The design of a WSN is highly dependent on the application; it must consider such factors as the environment, objectives, cost, hardware, and system constraints. WSN applications are discussed in a later section.

Sensor network platform

A WSN platform includes sensor node software, hardware, interface, and sensors (Figure 4). It is important to integrate multifunctional sensors on a WSN platform and process raw data with limited resources. The operating system and protocol stack are designed to support...
these platforms. Figure 5 shows the component of a WSN platform, which was applied in Great Duck Island, Maine, USA, for habitat monitoring (Polastre et al. 2004).

WSN standards define the functions and protocols for sensor nodes to interface with a variety of networks including IEEE 802.15.4 (IEEE 2004), ZigBee (www.zigbee.org/), WirelessHART (Chen, Nixon, and Mok 2010), ISA100.11 (www.isa100wci.org/), 6LoWPAN (Hui and Culler 2008), and Z-wave (www.z-wave.com/).

Communication protocol
A reliable and energy-efficient protocol stack is important for various WSN applications. The relationship among applications, services, communication protocols, and platforms, is shown in Figure 6. The functions of each layer are shown in Table 1 (Yick, Mukherjee, and Ghosal 2008). Zigbee is a representative and widely used
Network services

Network services of WSNs include localization, synchronization, coverage, compression and aggregation, and security (Yick, Mukherjee, and Ghosal 2008). They enhance the performance of the network in terms of power, task distribution, and resource usage.

For localization, a global positioning system (GPS), beacon or anchor nodes, and proximity-based localization methods could be deployed for the sensor nodes. Time synchronization is important for routing and power conservation. The time synchronization protocols aim to accurately estimate time uncertainty and synchronize each node’s local clock in the network. Determining sensor coverage is important for evaluating its effectiveness. The quality of monitoring in a WSN is dependent on the application. Data compression techniques involve compressing the size of the data before transmission. Decompression of the data occurs at the base station. For data aggregation, data are collected from multiple sensors and combined together to transmit to the base station. The techniques address the issue of energy, robustness, scalability, accuracy, and efficiency, aiming to reduce communication cost and increase reliability of data transmission. Wireless nodes broadcast their messages to the medium, so the security issue must be addressed in WSN.

WSN applications

As an important part of the IoT, WSNs have many applications for environment/Earth sensing, health-care monitoring, target tracking, and so on. A WSN application system includes a server, sensor nodes, network protocols, and a database.


<table>
<thead>
<tr>
<th>Layer</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transport layer</td>
<td>Ensure the reliability and quality of data at the source and the sink</td>
</tr>
<tr>
<td>Network layer</td>
<td>Handle the routing of data across the network from source to the destination</td>
</tr>
<tr>
<td>Data link layer</td>
<td>Ensure the data transfer between two nodes that share the same link</td>
</tr>
<tr>
<td>Physical layer</td>
<td>Provide an interface for transmitting bit streams over the physical communication medium</td>
</tr>
<tr>
<td>Cross-layer</td>
<td>The protocol stack is treated as a whole system and not individual layers; layers share information from the system</td>
</tr>
</tbody>
</table>

IoT architectures and applications in the world

In the United States, in 1993, the Defense Advanced Research Projects Agency (DARPA) supported a WINS (wireless integrated network sensors) project (Vardhan et al. 2000) at the University of California Los Angeles (UCLA). In 1998, DARPA supported the SensIT project (Kumar and Shepherd 2001) in more than 20 research institutes, to develop software for distributed WSNs. In 1999, the University of California Berkeley started the “Smart Dust” research project (Kahn, Katz, and Pister 1999).

Key research groups in the United States include the WEBS (wireless embedded system) laboratory at Berkeley (http://smote.cs.berkeley.edu/motescpelescope/), the Center for Embedded Networked Sensing at UCLA (http://research.cens.ucla.edu/), and the Ubicomp laboratory at the University of Washington (http://ubicomplab.cs.washington.edu/).

In the past 10 years, SmartGrid (http://www.nist.gov/smartgrid/), using information technology (IT) to deliver electricity in efficient, reliable, and secure ways, has become an important IT strategy. IBM outlined a new agenda for building a “Smarter Planet” in 2008 (www.ibm.com/smarterplanet/).

EPCglobal proposes a structure and worldwide standard for electronic product code (EPC) technology (www.gs1.org/epcglobal). Auto-ID laboratories under EPCGlobal is the leading IoT research network of academic laboratories (http://autoidlabs.org/).

A number of companies are working on IoT, including MEAS, Honeywell, Zebra, Symbol, Magellan GPS, HP, IBM, Cisco, AT&T, Oracle, Microsoft, and Intel. Representative WSN companies include Dust networks (acquired by Linear Technology in 2011), Archrock (acquired by Cisco in 2010), and Crossbow (acquired by Moog in 2011).

The main IoT application areas include smart grids, health-care monitoring, smart homes, environment monitoring, intelligent transportation systems (ITS), and the military.

In the European Union (EU), the eEurope strategy (eEurope: an information society for all) was proposed in 2000 (EUR-Lex 2005), followed by the i2010 strategy (i2010: information society and the media working towards growth and jobs) in 2005 (EUR-Lex 2009).

IoT projects supported by the EU Seventh Framework Programme (FP7) include IOT-A (Internet of Things-Architecture) (www.iot-a.eu/), SENSEI (integrating the physical with the digital world of the network of the future) (www.ict-sensei.org/), CASAGRAS (coordination and
support action for global RFID-related activities and standardization) and BRIDGE (building radio frequency identification solutions for the global environment) (www.bridge-project.eu/).

Many standardization organizations are working on IoT, including ITU-T, ETSI M2M TC, ISO/IEC JTC1 WG7/SC27, ISO TC104 SC4 WG2, and CEN/CENELEC/ETSI ITSTC.

A number of companies are working on IoT in the EU, including Ericsson, Nokia Siemens Networks, Alcatel-Lucent, and Vodafone. A representative WSN company is Libelium. The main IoT application areas include smart grids, smart homes, ITS, health-care monitoring, environment monitoring, industrial monitoring, and logistics.

In Asia, Japan proposed strategies including e-Japan (Electronic Japan 2001), u-Japan (MIC 2004), and i-Japan 2015 (i-Japan 2009). A number of research projects have been developed including CUBIQ (cross ubiquitous platform) (Dempo and Yoshida 2010), TRON (the real time operating-system nucleus) (www.tron.org/), Live-E! (live earth) (www.live-e.org/), WIDE (widely integrated distributed environment) (www.wide.ad.jp/), and GUTP (Green University of Tokyo) (www.gutp.jp/).

A number of companies and institutes are working on IoT in Japan, such as Hitachi, NEC, NTT DoCoMo, KDDI, NICT, and NRI. The main IoT application areas include smart grids, smart homes, ITS, health-care monitoring, environment monitoring, and earthquake prediction.

Korea proposed the u-Korea Master Plan strategy and the u-IT839 strategy in 2006, to promote the development of Internet Protocol version 6 (IPv6), ubiquitous sensor networks (USN), and radio frequency identification (RFID), aiming to achieve the first ubiquitous society. Four u-IT plans – u-City, telematics services, u-IT cluster, and u-Home – are proposed. A number of companies and institutes are working on IoT in Korea, such as ETRI, Samsung Electronics, LG Electronics, and KT.

China proposed a sensing China strategy in 2009, and a Smart City strategy in 2012. A number of companies and institutes are working on IoT in China, including the Chinese Academy of Sciences, China Mobile, China Telecom, China Unicom, Huawei, and ZTE. The main IoT application areas include smart city, ITS, smart homes, smart agriculture, health care, and environment monitoring.

Canada, India, Russia, Brazil, South Africa, Australia, Singapore, Malaysia, and other countries have started IoT projects in the application areas of smart city, ITS, smart agriculture, smart grid, health care, and oil exploration.

Field observation research network

Field observation research refers to monitoring and indicator measurement in the field for scientific research across many disciplines, including ecology, earth and atmospheric sciences, and ocean sciences. Through long-term observation, valuable raw data can be accumulated for original scientific discovery, new technology development, experiments, and demonstrations. Field stations are usually far from human settlements, so it is necessary to build field observation research networks for data acquisition and transmission.

Currently, there are many ecological research networks, such as the international long-term ecological research network (ILTER) (Aronova, Baker, and Oreskes 2010), national ecological observatory network (NEON) (Hobbie et al. 2003; Johnson et al. 2010), and Chinese ecosystem research network (CERN) (Fu et al. 2010). The system infrastructure includes components such as data acquisition, data transmission, data management, and application.
SENSOR NETWORKS, WIRELESS

(Figure 7). The Open Geospatial Consortium (OGC) sensor web enablement (SWE) (Botts et al. 2008) defined a suite of web service interfaces and communication protocols for the heterogeneity of sensor communications.

Exemplar environment monitoring applications

Environment and habitat monitoring are important application areas of sensor networks (Mainwaring et al. 2002).

Volcanic monitoring (Allen et al. 2006) provides a science-centric evaluation by a 19-day sensor network deployment at an active volcano on Volcan Reventador in northern Ecuador. The system used 16 sensors to continuously collect seismic and acoustic data and transmit reliable data to the base station.

GreenOrbs (Liu et al. 2013) is a consistently operating sensor network system deployed for forest surveillance. With up to 330 nodes deployed in the wild, GreenOrbs provides various data, including temperature, humidity, illumination, and carbon dioxide titer, for forest observation applications.

The Live E! project (Esaki and Sunahara 2006) is an open research consortium founded by the WIDE project and the IPv6 Promotion Council, Japan. This project is aimed at exploring the platform to share digital information related to the Earth and the living environment for disaster monitoring, reduction, and recovery. Live E! has accommodated more than 100 stations using a low-cost weather sensor network.

The ZebraNet (Zhang et al. 2004) system is a mobile WSN for tracking animal migration. It is composed of sensor nodes built into the zebra’s collar. GPS data are sent in multiple hops across zebras to the base station. Six to ten zebras were collared during the test in central Kenya. A dataset was collected to understand their migration pattern.

VigilNet (He et al. 2009) is designed to support long-term military surveillance using a sensor network consisting of 200 nodes in an outdoor environment. The complete system is designed to scale to at least 1000 XSM motes and cover a minimum of \(100 \times 1000 \text{m}^2\) to ensure operational applicability.

There are many research projects using WSN for environmental monitoring and target tracking. SenseScope (Barrenetxea et al. 2008) is a real-world deployment for environmental monitoring that took place on a glacier in Switzerland, consisting of about 100 sensor nodes. Trio (Dutta et al. 2006) is a sensor network deployment consisting of 557 solar-powered motes for multitarget tracking. MacroScope (Tolle et al. 2005) is a sensor network for monitoring the microclimate of a redwood tree. The experiment recorded 44 days in the life of a 70-meter-tall redwood tree, at a density of every five minutes in time and every two meters in space. ExScal (Arora, Ramnath, and Ertin 2005) is an attempt to deploy a sensor network at extreme scales. The system consists of about 1000 sensor nodes and 200 backbone nodes in a 13 km \(\times\) 300 m remote area in Florida, USA. An underwater WSN (Detweiler et al. 2010) has been used to understand water dynamics and their impact on the global environment.

Sensor network and integrated Earth observation network

Geospatial information science and technology (geomatics, geoinformatics) is the discipline of gathering, storing, processing, and delivering geographic information. This discipline includes tools and techniques used for land surveying, remote sensing, cartography, geographic information systems (GIS), global navigation satellite
Figure 7  Field observation research network.

systems (GPS, GLONASS, Galileo, BeiDou), photogrammetry, geography, and other fields involving mapping the Earth. Google Earth is changing the way the world is seen.

The integrated Earth observation system is the frontier research area of geomatics; it comprises the measurement of air, water, and land based on a network of sensors stationed on the ground, in the air, or in space, to improve the understanding of the changing Earth and interactions among various components on Earth. The global earth observation system of systems (GEOSS) (www.epa.gov/geoss/), promoted by the group on earth observations (GEO), connects the producers of environmental data and decision-support tools with the end users to enhance the relevance of Earth observations to global issues.

With the development of IT and communication technologies, the integration of Earth observation networks includes smart sensor networks, intelligent data-processing systems based on cloud computing, grid GIS, and intelligent geoservice agents (Li and Shao 2009).
SENSOR NETWORKS, WIRELESS

IoT is becoming the most important information technology that connects distributed, embedded microsystems in the physical world with the cyber world through networks. As an important part of IoT, core WSN information technologies include sensing, computing, and communication. WSNs will have more real-world applications in the IoT and IoE fields in the future.

SEE ALSO: Sensor networks, the sensor web, and the Internet of Things

References


SENSOR NETWORKS, WIRELESS


Unlike many terms and concepts in geography, “sequent occupance” has a clear genealogy. Derwent Whittlesey (1929) coined the term and proposed the methodology.

He declared that human occupance of an area, similar to biota, carries the seeds of its own transformation. Given this, it would be the task of geographers to discern the various stages of occupance, and thereby to provide regional or chorographic studies with a structure amenable to generalization. Although cast as an approach to historical geography, the concept was adopted and adapted mostly by cultural and regional geographers. Whittlesey’s initial training at Beloit College and the University of Chicago was in history, punctuated by service in World War I. Returning from the war, he completed his PhD in geography at Chicago and joined the geography faculty there, before moving to Harvard in 1928. Like Carl Sauer’s (1925) “The Morphology of Landscape,” published soon after Sauer’s arrival at Berkeley, Whittlesey’s “Sequent Occupance” might be seen as a “habilitation” exercise – providing a programmatic statement and a preview of his epistemological orientation. He failed, however, to demonstrate its utility with directed studies of his own. That task was left to other geographers and a generation or two of graduate students for which the concept and method offered an attractive and tractable approach to geographical change in bounded places or regions over time. Marvin Mikesell (1976), in his essay “The Rise and Decline of ‘Sequent Occupance’,” suggests that the concept enjoyed a decade or two of popularity and patronage before sliding into senescence by the 1950s coincident with the eclipse of regionalism as geography’s key paradigm and with Richard Hartshorne’s earlier advocacy of an ahistorical chorology. Pace Mikesell’s somewhat ironic title (see References), the concept’s own occupancy in geography and neighboring fields has had a longer if not stronger presence than is often assumed.

Whittlesey’s move from Chicago to Harvard meant a shift from a human geography-oriented department to one where Davison physiography and ontography still held sway. In proposing a model that held that human occupance of an area, similar to biotic phenomena, carries within itself the seed of its own transformation, Whittlesey appeared to many to be following the logic and logistics of Davis’s stage-managed erosion cycle. Further he held that the succession of stages of human occupance establishes the genetics of each stage in terms of its predecessor. Thus it followed that the life history of each discloses the inevitability of the transformation from stage to stage. He did, however, suggest that change could come from exogenous as well as endogenous causes, whether from the inherent character of a particular mode of occupance, or from extreme physical events (earthquakes, windstorms, tidal waves, floods, volcanic eruptions, landslides, insect or other biological pests, pestilence, and so on) or human agency (boundary shifts, political upheavals, legal changes, population movements, communication innovations). But rather than narrowly viewing sequent occupance as a variant of the Davison model, or an opportunistic entrée into his new milieu, one might look to
SEQUENT OCCUPANCE

his Midwestern schooling for inspiration or at least incubation. Most commentaries on sequent occupancy point to similarities with Frederick Jackson Turner’s Frontier Thesis. Whittlesey and Turner not only shared middle-class small-town Midwestern origins and careers capped at Harvard, but Turner’s historical vision was among the optics with which Whittlesey focused his earlier historical studies. While their stage models of historical change invite comparison, the pioneering work of the Chicago biological ecologists probably had the larger influence. Human ecology as formulated by Chicago geographers J. Paul Goode and Harlan Barrows, or the derivative and lateral development of the urban theories by the Chicago sociologists Robert Park and Ernest Burgess, owe initial impetus to their cross-campus colleagues studying ecological plant succession. Whittlesey himself notes that the analogy between sequent occupancy in chorology and plant succession in botany should be apparent to all. But he goes on to say that human occupancy involves greater complexity, as the botanist is dealing with a single subject – plant associations, and the main agents of transformation are limited to three factors – climate, soils, and human disturbance. With human occupancy, the successional history is not only mediated by changes wrought by biophysical factors, but also by those stemming from correspondingly complex cultural forms.

Sauer’s “Morphology of Landscape” (1925) was a resounding repudiation of the environmental or geographical determinist current in geography that had flowed through the discipline’s history since its inceptions in Classical Antiquity, but had welled up in the late nineteenth century and first two decades of the twentieth. Sauer’s “Morphology” also provided an opening or avenue for historically directed chorology or regional studies, but this was not to be its principal legacy. Whittlesey’s aim at offering chorology conceptual and methodological underpinnings was more direct, and more realized. With the discrediting of the environmental determinist conceit, geography was deprived of its reigning nomothetic or generalizing approach. Whittlesey recognized that chorology, left to its own logic, would be endlessly idiographic – rendering regional descriptions at any and all scales. No matter how artful or erudite these constructions might turn out to be, in aggregate, the summation would be simply that. Instead, Whittlesey proposed that his scheme could coax an order out of chorology that would lend itself to generalization. By identifying the sequential occupation stages evident in different regional histories, geographers could then begin to build a quasi-science of chorology.

To illustrate the concept and method, Whittlesey chose a small district in northern New England rather than his native Midwest, no doubt a nod to his recent relocation to New England. He does not specify its precise location, but does give its areal extent – 15 square miles, thus at the scale of a “micro-region.” While in graduate school at Chicago, Whittlesey participated in the Midwestern geographical summer field schools directed by Robert S. Platt, advocate of microgeographical research. Work at this scale was deemed manageable for either teams or individual researchers. Whittlesey’s sample district’s features was described as having a massif of high hills, mantled by bouldery podsols, with deep winter snow and well watered by frequent summer rains. The climate consisted of long, cold winters and cool summers with only occasional hot days. The vegetation cover was a tangle of pioneer, mixed deciduous and coniferous second-growth forest. A few grassy lanes penetrated the plot, allowing cattle grazing. Whittlesey suggests that finer microgeographical discrimination was possible, but would not alter its exclusively descriptive character. To
go beyond, posing the description in sequence with preceding modes of human occupance and positing the probable subsequent mode(s) would allow for genetic correlations. Whittlesey detected three distinct stages and predicted a fourth. The prior stage had featured farming and associated land use: plowed fields on gentle slopes, pasture land on steeper slopes, sugar maple selection on rougher terrain and soils, and wood lots in remaining spaces. Evidence for this regime could be found in both field and archive in the form of relic walls, paths, and roads, house foundations, and written accounts. Prior to European agriculture, Whittlesey posited virgin mixed forest with only a few migrant Indians pursuing a hunting and gathering existence. Whittlesey foresaw a future reoccupation of the district by nonresident owners cutting it periodically for pulpwood or possibly lumber. From this example of sequential occupancy, Whittlesey drew a confident conclusion as to its internal logic and by extension its seemingly universal applicability. He asserted that each generation of human occupance is linked to its forbearer and to its offspring, with individuality expressed through mutations in some elements of its natural and cultural characteristics.

The reception by geographers was far from universal acceptance, but it did attract adherents, especially among geography students writing theses and dissertations within the larger regionalist or chorological framework. Perhaps the most widely cited example was Preston James’s (1929) Blackstone Valley study. This publication actually preceded Whittlesey’s paper, but it has been drafted as exemplar. Afterward, the center of gravity of these studies moved west. Darrell Davis (1930) published on the succession of human activities in the Kentucky Mountains, and Stanley Dodge (1931) the following year on sequent occupance of an Illinois prairie. Perhaps the doyen of this genre was Alfred Meyer (1934), who wrote his dissertation under Dodge on the Kankakee Marsh in northern Indiana and Illinois. Meyer spent the next two decades publishing a series of studies on this region using the sequent occupancy approach. Whittlesey mentee Edward Ackerman (1941) demonstrated sequent occupance’s potential in the urban sphere. Subsequently, a somewhat independent current developed in urban studies and the planning arena. To date, there has been no complete accounting of the number of studies that Whittlesey’s paper inspired, nor of their locational distribution. But in the spirit of Whittlesey’s concept, there seem to have been three distinct stages: (i) a scattering of geographical studies prior to 1929 (sans the precise terminology) that anticipated Whittlesey’s directives, (ii) a collection of studies during the 1930s and 1940s spawned by his manifesto, tailing off in the 1950s, and (iii) residual studies to the present day that may or may not be directly derivative of Whittlesey’s proposal.

**SEE ALSO:** Human ecology; Regional geography

**References**


SEQUENT OCCUPANCE


Serbia: Geografski institut “Jovan Cvijić” Srpske akademije nauka i umetnosti (Geographical Institute “Jovan Cvijić” of the Serbian Academy of Science and Arts)

Founded: 1947
Location of headquarters: Belgrade
Website: www.gi.sanu.ac.rs
Membership: 58 (as of December 31, 2013)
Executive director: Milan Radovanović
Contact: m.radovanovic@gi.sanu.ac.rs

Description and purpose

The institute is a unit of the Serbian Academy of Sciences and Arts in Belgrade. Its mission is to organize systematic scientific research in geography, including the study of landscape phenomena, objects, forms, relationships, conditions, and processes within the geographical environment. The institute focuses on the landscapes of Serbia as well as the broader territories of the Balkan Peninsula, south and central Europe, and the world.

Journals or major publication series

Other editions and joint publications. www.gi.sanu.ac.rs/en/publications/other_editions/other_editions.html

Current activities or projects

The scientific research of the institute is performed by way of engaging, individually as well as collectively, the services of all the collaborators of the institute with a view to realizing scientific projects. The results of this scientific work are published in institute, academy, national, and international periodic and special publications. Research results are presented at scientific conferences, congresses, symposia, seminars, expeditions, and other gatherings that contribute to scientific advances.

Brief history

The institute was founded on May 31, 1947, within the framework of the Serbian Academy of Sciences and Arts. One of the founders was the world-renowned scientist Milutin Milanković. It has been operating since 1961 under the name of the Geographical Institute “Jovan Cvijić” of the Serbian Academy of Sciences and Arts. Jovan Cvijić was one of the greatest Serbian scientists, the leading geographer, a founder of modern Serbian and Yugoslav geography, an influential intellectual, and a great national figure in Serbia and Yugoslavia at the beginning of the twentieth century. The first director of the Institute was Petar S. Jovanović, a member of the Academy of Sciences and Arts.

The institute’s fundamental interest is in basic, applied, and developing scientific research in
SERBIA

physical geography, the geography of population and settlements, social geography, regional geography, and cartography. To support its scientific programs, the institute is organized into departments that are subdivided into separate divisions.

In addition, the institute has an administrative organization which offers professional services.

Submitted by Milan Radovanović
Serbia: Srpskog Geografsko Društvo (SGD) (Serbian Geographical Society)

Founded: 1910  
Location of headquarters: Belgrade  
Website: www.sgd.org.rs  
Membership: 1500 (as of December 31, 2013)  
President: Stevan M. Stanković  
Contact: geofiz40@gmail.com

Description and purpose

The Serbian Geographical Society is a professional organization whose mission is to bring together scientific, professional, and educational activities of geographers, scientists, university professors and assistants, and professors of secondary and primary schools in Serbia. The main objective is to develop field and theoretical research in geography, study events and processes in all geographic disciplines and develop subjects at different levels of education, improve methodological and didactic principles of teaching at all levels of education, publish textbooks, manuals, scientific, professional, popular, and methodical-pedagogical works, organize scientific gatherings, congresses of Serbian geographers, regular annual seminars (10–12 lectures), and regular monthly meetings of the teaching department and the scientific department. Organization of competitions in geography for pupils in the final grades of primary school, collaboration with corresponding organizations and popularization of geographical knowledge in a variety of media on different occasions are also relevant.

Journals or major publication series

Special issue of the Serbian Geographical Society (in Serbian)  

Current activities or projects

The SGD organizes 10–20 expert and scientific lectures for its members every year. It carries out 4–7 excursions in Serbia and abroad where geographical issues of the visited areas are discussed. Once a year accredited seminars for teachers of geography of secondary and elementary schools are organized. On that occasion, eight to twelve lectures are presented and discussions follow. Members of the society are encouraged to prepare papers on a regular basis for publishing in journals for which financial support is provided by the Ministry of Education, Science and Technological Development of the Republic of Serbia. Organization of regional and state competitions for students in grades 7 and 8 of primary schools is of utmost importance. More than 1300 participants take part and many of them continue to the highest state-level competitions depending on their success. Every year, on Jovan Cvijić’s birthday (the founder of modern geography in Serbia and a world-renowned scientist), the Serbian Geographical Society organizes a exhibition, The Days of Cvijić, when representatives of 13 participating schools from Serbia and the
Republic of Srpska with the name of Jovan Cvijić take part. The event includes expert lectures, experience sharing, socializing, and field work.

Brief history

The SGD was founded in 1910 on the initiative of Jovan Cvijić and his associates, and was associated with the Faculty of Geography at the University of Belgrade. It was on this occasion that the rules of the society were established and the publication of the journal started. It is the oldest geographical publication in the Balkans, which is still regularly published in four volumes per year. During both world wars, due to the bombardment, destruction, and burning of the existing archives, the assets, library, and unsold copies of the books were destroyed. The main task of SGD is the scientific and professional work and popularization of geographical knowledge in all geographic disciplines and teaching subjects. The society has established successful cooperation with institutions in Serbia and abroad. Through the exchange of publications it has created a rich library. In 1935, it established the Medal of Jovan Cvijić as the most important recognition to individuals and institutions in Serbia and abroad for their outstanding contribution to the development of geography. The society has a board comprised of 11 members. The head of the board is the president who has assistants for teaching and research. Activities take place through the work of the teaching department and the scientific department. The SGD has branches in several major cities in Serbia. In 2010 the SGD celebrated a century of existence and operation by publishing a commemorative issue.

Submitted by Stevan M. Stanković
The first geographic information systems were built as tightly coupled whole systems. This coupled and closed system architecture made it easy to develop relatively simple domain-focused applications. But when more functions were added to these systems, and especially with the widespread proliferation of computer networks, it became necessary to decouple the monolithic system architecture into a componentized framework to improve the management of system design and development functions. At that stage, system architectures started to evolve into structure-oriented, object-oriented, and component-oriented frameworks. For example, to further improve software componentization and make it correspond more closely to our conceptual recognition of the world, object-oriented system architectures were proposed and became popular for a few decades. Since the beginning of the 1990s, numerous software components have been developed and put into operation to provide valuable data-processing capabilities. These operational components are of value when they can be integrated and used to build new computer systems. Even more value can be added when components are further built into interoperable services based on standards defined by user communities as well as by national and international standard bodies. Service-oriented architecture (SOA) has emerged and has served as an important model for system design and development since the mid-2000s (Erl 2005). SOA has been advanced as a set of methodologies and principles that guide the design and development of new computer systems by leveraging existing computing services and managing their development for future reuse in an interoperable fashion.

The geographic information science (GIScience) community recognized the importance of SOA for information system integration, and it was initially adopted to build the National Spatial Data Infrastructure (NSDI) by the Federal Geographic Data Committee (FGDC) and to support the reuse of geographic information assets by adopting a set of geospatial standards defined by the Open Geospatial Consortium (OGC), such as the OGC suite of web services for sharing geospatial information assets. Many recent developments in GIScience have benefited from this service-oriented approach of reusing and interoperating existing geographic information systems. For example, the Google Map and Google Earth systems can be easily integrated into new geographic information system (GIS) applications through the use of relatively simple scripting programs. This service-oriented design enables developers to leverage the vast amount of geographic data and mature functionalities of Google (such as geocoding, routing, IP-to-address conversion, and geographic pattern visualization) in a rapid fashion, thus reducing the system–development life cycle from years to weeks or days, with high reliability and scalability.

This entry introduces SOA from the aspects of key supporting technologies, popular examples, and future developments in the GIScience domain.
Key technologies

The reuse and interoperability of information systems has been an elusive goal of system designers for several decades. This goal became achievable with the development of SOA. SOA evolution has been enabled by several technological developments since the mid-1990s, including software design technology, computer networks, standardization, and services.

Software design technologies

The first computers were created to carry out numerical calculations to meet complex production requirements. This focused capability was supported using sequential software design and engineering processes. In the 1960s computers began to be used for many business functions, and these general-purpose information systems required much more complex functionalities. To deal with the complexity of software development, multiple software design approaches were proposed in the 1960s and 1970s; the most notable ones (Shaw and Garlan 1996) are structure-oriented (SO), object-oriented (OO), and component-oriented (CO).

The SO software design approach was adopted to improve code clarity and quality, and to reduce development time by using subroutines and block structures as well as for and while loops. This structured approach to system design and development was very useful when software assets were limited, and the approach was used for several decades, achieving particularly widespread use in the 1980s to design business systems.

Seeing the need for modeling software components that would correspond to real-world objects and to design software in a more “natural” fashion, OO design and development emerged in the late 1960s and became popular in the 1990s for the design and development of software systems. Because it is close to human perceptions of problem domains, the OO approach remains a very popular approach for software design and development.

The CO software design and development approach was also developed in 1960s when general purpose, and thus complex, software systems became common. The CO approach divides an information system into different functional components and ensures that each component is fully documented and tested; complexity is hidden from end users in this approach (Heineman and Councill 2001). The CO approach is popular for software development because of several important characteristics: (i) single modules can be developed and used multiple times; (ii) noncontext-specific development maximizes reuse; (iii) modules can be combined with other components to form complex systems; (iv) complexity is encapsulated, thereby leaving external interfaces open to current and future users; and (v) independent units can be flexibly deployed and tested. These characteristics were eventually developed into the service-oriented software design and development approach.

Computer networks

Computer networks were introduced in the late 1950s and early 1960s to share data between computers and facilities, though they operated initially at a very low speed (at the rate of kilobits per second or kbps). In the 1980s the speed was improved to megabits per second (mbps) with Ethernet. In the 1990s the speed was increased to 100 Mbps, and at the end of 1990s the network speed was increased to gigabits per second (Gbps) with new fiber optic cables and fast switches. This development is best represented by the Internet, the global interconnected Internet protocol (IP) system.
In 1960s the US Department of Defense sponsored the ARPANET to connect computing facilities across different organizations in the 48 US continental states using interconnecting transmission control protocol (TCP)/IP networks. In 1982 the TCP/IP protocols were standardized, and it then became much easier to connect to the Internet. In the 1980s and early 1990s the Internet expanded quickly to Europe, Asia, and other continents. The late 1990s and the first decade of the 2000s saw enormous increases in content delivered through the Internet. This fast-growing use of content exchange was especially accelerated with the World Wide Web (WWW).

WWW refers to the system of interlinked hypertext documents accessed via the Internet; it has its origins in the late 1980s at CERN, the European Organization for Nuclear Research. In the WWW, a web browser (the end-user application) can universally display all web pages provided by different web servers by following the Hypertext Transfer Protocol (HTTP) and the Hypertext Markup Language (HTML) standards. This seamless integration of web content made it extremely easy to publish documents in a worldwide environment.

The original format of serving static web pages through HTTP and HTML was termed Web 1.0. After the success of static content delivery, developers began to share multimedia through the web, including animations, videos, and others. Web 2.0 emerged in 2004 for end users of the web to contribute content through tools such as wiki, Twitter, Facebook, and blogs. The emerging Web 3.0 is envisioned to have intelligent capabilities, such as the ability to understand the meaning of text and to use that meaning to conduct searches and access information and other resources available on the Internet.

The continuous evolution of computer networks has provided convenient channels for end users to freely share information and processing capabilities through web services. The broadening of web capability from desktops and laptops to portable devices such as smartphones and tablet computers has made it easy for people to access and contribute content through web services.

**Standardization**

The rapid development of software technology and computer networks has been enabled by the standardization of system interfaces, from hardware switches and plugs (such as RS-232 and HDMI) to software and information coding (such as HTTP and HTML). These standards are developed by professional organizations, such as the Institute of Electrical and Electronics Engineers (IEEE), in collaboration with the International Organization for Standardization (ISO). In the GIScience domain, many GIS application successes have benefited from the standards developed by organizations such as the FGDC, OGC, and ISO (Alameh 2002).

In the early 1990s the FGDC was established to build the NSDI for the United States. Soon, it was recognized that standards and interoperability are the core enabling technology for sharing both information content held across different agencies and services built and operated by different organizations. OGC was established at a similar time to look at how to standardize geospatial information systems. During its first few years of operation, OGC laid out a framework for various aspects of standardization, from projections, geographic coordinate systems, data content, and data encoding, to data and information services. In the early 2000s, the focus of OGC has shifted to standardizing the format of information encoding (such as Geography Markup Language (GML)) and system interfaces as service contracts.
among different services (such as the OGC Web Services).

OGC Web Services include the well-known Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS), Catalog Service for the Web (CSW), Web Processing Service (WPS), and Sensor Web Services, among others. With these services standardized and approved by the American National Standardization Institute (ANSI) as national standards and by ISO as international standards, OGC is focusing now on how standards could facilitate the sharing of geographic information services to support the integration of new applications to address challenges ranging from global climate change and natural hazards to emergency response. A primary focus is placed on its well-known OGC Web Services (OWS) Initiatives and by leveraging the latest technologies such as cloud computing.

Services

Based on advances in software technologies, computer networks, and standards, a variety of services are being developed to provide content or functional capabilities to a wide variety of end users. These services (Erl 2005) are:

- platform-independent: based on a set of standards to enable the interoperability between different services across different platforms and languages;
- loosely coupled: minimum amount of interdependencies between services;
- autonomous: service owns its own implementation logic;
- abstract: services hide their internal logic except through the interfaces that are defined by the service contract;
- service contractual: the interface between services is based on communication and documented agreements or standards;
- reusable: the abstraction and autonomy minimizes interactions among services, thereby improving reuse;
- composable: services can be assembled from other existing services;
- stateless: a minimum amount of activity-specific information is stored;
- discoverable: services can be discovered by their users through a registry or its interface.

These characteristics make it easy to compose new applications by using existing assets of data, information, and processing services. The next section provides examples of how SOA is utilized in the cutting-edge development of geoportals, mashups, and geoplatforms to facilitate the sharing of services across distributed environments.

Examples

Recent developments in GIScience have benefited from the evolution of OGC Web Services and the SOA approach in developing applications. SOA is made up of the three parts of service, catalog, and application (Yang and Tao 2006): services are different types of systems operated by different organizations; catalogs are where organizations register their services so that others can search and find them; applications are where the services are bound to support a specific end-user function. Figure 1 shows a typical OGC SOA for the services publish, find, and bind (OGC, 2003). The bottom of the figure shows different types of data, portrayal, and processing services that can be published into a catalog service. The catalog can be searched by applications to find specific services. Then the services can be bound to form new applications, such as data discovery, map viewing, and others.
The OGC web service architecture adopts SOA concepts, including service, catalog, and application, in the functions of publish, find, and bind. Publish, find, and bind are supported using different types of encodings (adapted from OGC 2003).

Spatial web portal

A spatial web portal (SWP), also known as a geoportal, is based on SOA to build integrated geospatial information applications for data discovery, access, and integration. Figure 2 shows the architecture of the NASA Earth Science Gateway (ESG), which is a typical SWP and was one of the milestones of NSDI, along with the Geospatial One Stop (GOS) portal. The figure shows that many services are built and operated at the bottom with facilitating services, spatial web services, and maintenance services. These services are easily accessed by the spatial web portal through SOA, and the spatial web portals interact with each other to support different types of applications with publish, search, view, and visualization services. The interoperable services can be leveraged easily without much centralized development.

Mashups

Mashups have become a major application development approach for GIS in the mid-2000s because SOA enables rapid and reliable leveraging of existing GIS services (Batty et al. 2010). For example, homesdatabase.com easily integrates Microsoft Bing Maps as a background map with a minimal level of scripting development. In a similar vein, Google Maps easily integrates
Figure 2 The spatial web portal, supporting the Geospatial One Stop, geoportals, and many other portals in the past decade, adopted the SOA approach to leverage existing GIS services.

near real-time traffic and Global Positioning System (GPS) information to show driving conditions through scripting to ingest services from relevant systems. Figure 3 shows the architecture for the development of an emergency response system using existing services for the Loma Linda University Medical Center (LLUMC) Advanced Emergency GIS (AEGIS).

GeoPlatform

GeoPlatform (www.geoplatform.gov) is the latest development of the NSDI. The initiative is targeted to be a collaborative environment that leverages SOA and the latest advancement of Web 2.0 to facilitate community application building by combining existing geospatial assets from different agencies, companies, academia, and other organizations. The system is built on top of a cloud computing infrastructure and is linked to existing geospatial information services through interoperable interfaces. This SOA approach can easily piece together data and services from different resources to form an application that addresses a specific problem. For example, Figure 4 shows a map that integrates different types of natural hazards information from different organizations such as the National Oceanic and Atmospheric Administration (NOAA). Microsoft Bing was adopted as the background map and Esri’s GIS services were used through the SOA architecture. With this final product in hand, decision-makers are able to understand what natural hazards are present across the country without needing to care about system integration details. Most important of all, this functionality can be quickly achieved in GeoPlatform, and live feeds from operational services would be able to keep the system up to date.
Figure 3  Leveraging SOA, GIServices (represented by all the boxes) can be accessed through standardized protocols including HTTP/SOAP, TCP/IP, XML, and other scripting-level development techniques to easily integrate data from a variety of resources in a relatively fast fashion to support the emergency response of the Loma Linda University Medical Center.

New developments

Cloud computing

Cloud computing is a new computing paradigm that is being used to provide computing resources (virtualized hardware, platform, software, functional components, and others) as a service. This new computing paradigm continues to benefit from SOA service standardization. Cloud computing provides five valuable computing characteristics – on-demand self-service; elastic; pooled computing; measured services; and broadband access – which have helped to popularize the use of SOA. However, it should be noted that cloud computing is focused on sharing computing capabilities and SOA is focused on sharing services; they benefit each other. Spatial cloud computing is proposed to address how the GIScience domain could benefit from and help to shape cloud computing (Yang and Huang 2013).

Big data

Big data are emerging as a new opportunity and challenge for scientific discovery. Historical and new datasets are accumulating at geographically distributed locations. The processing capabilities, as well as the information demands for mining big data, are also distributed. It is very
likely that SOA will play a driving role in the development of solutions to convert big data into information and knowledge by improving the discoverability, accessibility, and usability of massive data streams.

VGI or social media
Volunteered geographic information (VGI) has become a popular way to amass user-contributed content which can be used to support decision-making and provide information that was heretofore unavailable. The collection, management, processing, and serving of VGI have also benefited from SOA; the general public can freely and easily contribute information and data through publicly accessible services.

Crowdsourcing or volunteered computing
Many traditional computing problems are shipped out to the general public to use idle computing cycles (Anderson et al. 2002). The support for this type of computing and the collection of computing results are provided in an SOA environment. The evolution of this approach has also helped in the development of popular worldwide service systems such as Google Maps and Microsoft Bing Maps.
Citizen as sensors and the Internet of Things

The popularity of SOA and the ability to connect anything to the Internet make it very convenient for users (people, devices, sensors) to contribute information in an automated fashion. This makes it possible for citizens and things to easily serve as sensors by contributing their situational awareness and information to the overall computing infrastructure. The infrastructure in turn can be mined to support better decision-making with new information and knowledge.

Discussion and conclusion

SOA as an architecture and methodology for developing GIS applications has been firmly established during the past decades. An increasing number of applications are being built using SOA practices. The SOA approach is also being driven by the demand for agile and scalable information and solution developments, such as for NSDI, geopertals, the GeoPlatform, and the NSDI Clearinghouse, as well as mashups. The development life cycle for integrating different services and applications has been reduced from years to months, weeks, even days and hours. Systems ranging from GIScience infrastructure to emergency response can rely on an unprecedented availability of information and computing infrastructure.

The SOA evolution has supported major GIScience advances across several GIS frontiers, such as spatial cloud computing (Yang and Huang 2013), big data, VGI, crowdsourcing (Goodchild and Glennon 2010), volunteered computing (Anderson et al. 2002), citizens as sensors, and the Internet of Things. Given its maturity and verified efficiency for GIS application development, it is expected that SOA will continue to play a significant role in GIScience and geography in the coming decades.

Acknowledgments

This publication is partially supported by the NSF EarthCube Program (ICER-1343759), CyberPolar Program (PLR-1349259), and I/UCRC Program (IIP-1338925), by the Federal Geographic Data Committee, and by NASA’s Center for Climate Simulations.

SEE ALSO: Cloud computing; CyberGIS; Mapping cyberspace; Mapping mashups; Open Geospatial Consortium standards; Web-mapping services

References

SERVICE-ORIENTED ARCHITECTURE


During the past several decades, geographers have begun to consider seriously how human sexualities are both shaped by and expressed through the places in which they live and encounter others. Prior to the 1980s, human geography had typically treated human sexuality as innate, fixed, and presumably heterosexual. Despite some early attention to sexuality in psychoanalytic strands of the social sciences, a mid-twentieth-century quantitative revolution favoring measurable phenomena had made Western geographers more concerned with, for example, fertility patterns in countries rather than the gendered and sexualized experiences or family formations that produced those patterns. Progress toward more critical, flexible understandings of sexualities has since been somewhat gradual and uneven. Geographers’ understandings of sexuality now encompass lesbian, gay, bisexual, transgender, queer, and intersex (LGBTQI) populations, and often emphasize a “queer” politics that blurs the lines between sexual and other social categories. At the same time, much of their work remains uneven both geographically and socially, focusing mostly on gay and lesbian sexualities in the cities of the Global North, and for many geographers around the world it remains a nonsubject.

Many of the first interventions into geographies of non-normative sexualities in the late 1970s and 1980s simply described uses of urban space among sexual identity “groups” (e.g., gay men) and still treated sexuality as a fixed characteristic. Many of these studies were also urban economic in orientation, identifying how spatial concentrations of residences and businesses among gay men in cities such as New Orleans formed the basis of political voting blocs and complementarities within a gay consumer market. More recent work has focused on how these areas in North American and European cities have been developed, branded, and visibly marketed to attract a broadening based of consumers, including heterosexual people, while excluding others such as gays and lesbians who are working-class or transgender individuals. Meanwhile, work on lesbian communities has revealed a more diffuse, socially networked (as opposed to spatially concentrated) form of kinship (Peake 1993). Work on younger gay men and lesbians has also examined the role of more specific processes (e.g., migrating away from the family home to the “big city”) in sexualized life transitions, such as coming out or developing a gay or lesbian identity. Over time, research on sexuality and space became influenced increasingly by feminist and critical approaches in Euro-American geography that acknowledged sexualities as one of the multiple, overlapping forms of social subjectivity that mediate how individuals experience the places they inhabit.

The narrower body of work on urban-focused “gay and lesbian geography” is now joined by work on sexual identities in a wider variety of populations (including heterosexual people) and on the inscription of sexuality in places through ongoing processes of performance, evaluation, and regulation. Performativity refers to the ways in which individual representations of sexualities
SEXUALITIES

(e.g., dress, corporeal movements, and speech) are shaped by dominant, often repetitive, discourses. While performances that align with these scripts might elicit positive responses such as love, affection, and social affirmation, those that do not might be met with violence or exclusion. In this way, individuals become the actors that perform “acceptable” discourses to make them a lived reality. Since sexualities – at least as they are interpreted societally – are characterized by particular sets of gendered norms (e.g., “masculine” confidence and aggression vs “feminine” fragility) that can be altered or concealed, they may be performed differently in different areas: a gay-identified man might walk more “butch” in certain areas where he expects to be threatened or harassed for being perceived as gay. Similarly, research on urban nightlife has shown that codes of behavior (e.g., self-segregation of men and women into different spaces of bars, perceived conventions that men must approach women or buy drinks for them) are reinforced by expectations that those behaviors are given in those spaces. Consequently, they may disempower some individuals (heterosexual women) by casting them as sexualized objects that are meant to be sought out or solicited by others (heterosexual men).

Research on the regulation of sexuality has also been fruitful in outlining how sexualities are inscribed in particular places. Some work has looked at the regulation of female sexualities, which were historically thought of as needing to be “cocooned” in domestic, procreative heterosexual relationships. Historical work on British imperialism, for example, demonstrates how control over sexuality, through dress codes and curfews for single women, was used to create “proper” female subjects (and prevent prostitution by “fallen women”) within a colonial empire. More recently, the study of how women’s sexualities are regulated has somewhat shifted to the city scale, again with particular emphasis on their location in public versus private spheres. Some work on female sex work has shown that local and national governments’ attempts to regulate it, positioning the sex worker as harmful or morally corrupted, may push it further underground and into more geographically marginal areas, threatening the safety of workers. Conversely, attempts to bring sex work into the public sphere through legalization and containment (e.g., in urban red light districts) can encourage a sense of public ownership over individuals’ sex work, leading to moral panic over “powerless” sex workers being trafficked and, in turn, police interventions that may leave sex workers without economic livelihoods. In these ways, sexualities also became understood as fundamental to how individuals and the state relate to one another. Through a process of sexual citizenship, individuals are granted moral and social recognition for acting in a way that underpins the values that states use to construct themselves (e.g., as economically viable, safe, and family-oriented). Sexualized bodies therefore make places by providing the basis for social norms and policies, but also solidify (or resist) those norms by enacting (or rejecting) them.

For gay men and lesbians, regulation stemmed from the categorization of same-sex attraction into a single population category (i.e., homosexuals) at the turn of the nineteenth century, the subsequent classification of homosexuality as a mental illness as outlined (until 1978) in the American Psychiatric Association’s Diagnostic and Statistical Manual of Mental Disorders, and a criminal offense (until 2003 in many US states). The classification and regulation of same-sex attraction as “deviant” was evident in a number of historical campaigns. The US and Canadian governments, for example, sought to root out sexual non-normativity from the civil service by expelling thousands of gay-identified or
gay-suspected employees in the 1950s and 1960s, often on the premise that homosexuality was tantamount to other deviant affiliations, such as communism. In the 1970s and 1980s police raids and subsequent closures of urban bath houses, carried out in the name of public decency and public health, removed sexual and social spaces that were crucial to many in the community amid otherwise heteronormative and heterosexist environments. More recently, some urban gay communities have experienced the “de-gaying” of historically established gay space through gentrification and residential development intended for straight, middle-class inhabitants.

The study of geographies of sexualities has also been extended into a wider range of places and scales beyond the urban center. Work on sexuality and space has often positioned metropolitan areas as the socially progressive places in which sexual and gender norms are challenged, and from which important notions of “gay culture” and “queer culture” diffuse (Knopp and Brown 2003). Rural and suburban areas, in contrast, have often been assumed to be profoundly heteronormative places where coupled relationships and gendered work/home divides most in line with the Euro-American state are valued and sexual non-normativity – in the form of being queer, or even heterosexual and single – is excluded. Since the 1990s, however, many geographers of sexualities have challenged these notions. Some have looked explicitly at events (e.g., gay- and lesbian-friendly festivals and retreats) or processes (e.g., return migrations of gay men and lesbians) that temporarily or permanently transform rural areas. Others have examined more intimate spaces (e.g., gay men’s suburban homes) that act as community spaces and are marked as visibly gay or queer through subtle uses of art and design referencing places and events relevant to the gay experience. Still other work, however, has suggested that rurality is not necessarily something to be transformed or overcome but a consciously chosen milieu that facilitates its own forms of kinship among sexual “others.”

More theoretical strands of sexuality and space, particularly those concerned with queer theory, call into question the concepts of “abnormal” and “other” themselves. Queer theory offers a challenge to heteronormativity (the norms that make heterosexuality right in the first place) and the associated systems of patriarchy and of othering of those who do not conform. It not only imagines sexuality as a flexible continuum rather than a set of categories, but posits queerness as both a sexual non-normative identity and a political disposition that actively rejects the normalization of certain sexual identities and the notion of fixed identities themselves. Part of this work is concerned with critiquing the ways in which some sexual identities have become accepted and even co-opted into heteronormative “mainstream” society. The concept of homonormativity, for example, suggests that gay and lesbian citizens, as long as they engage in coupledom, monogamy, full-time work, homeownership, and consumption habits that build the economy, can transcend their previous position at the margins of society. Figures such as the white, middle-class lesbian couple and the well-traveled “global gay” become assimilated while other less acceptable sexualities (e.g., nonmonogamous, working-class, and trans sexualities) are pushed to the margins in the same way that gays and lesbian had been within heteronormative society (Duggan 2002). These so-called homonormative subjects have also been described as complicit in the systems traditionally used to oppress them. Since they now partake in recently granted marriage and adoption rights, and other practices regarded by the state and society as “responsible” (e.g., homeownership), their actions are thought to marginalize those...
queer people who are unable to participate in these practices as a result of laws, discrimination, or financial limitations.

Other work influenced by queer theory asks why certain sexualities are considered normal in the first place. Recent work on polyamory, for example, has suggested that, rather than combating portrayals of polyamory as personal underdevelopment, immaturity, or self-absorption, scholars should question why “traditional” romantic love and monogamy were ever considered morally or ethically superior. Even framing polyamory as an acceptable, private personal choice, then, prevents it from being understood as another way of loving that need not be valued differently. Yet other work rejects treating sexualities as categories or stand-alone attributes. Recent work on intersectionality has suggested that sexualities cannot truly be separated from other individual identities and characteristics such as “race,” and the term “intersection” erroneously assumes that they are separate to begin with. Although more traditional streams of the social sciences suggest that discrimination experienced by queer individuals who are also working-class and/or nonwhite is additive (e.g., “double discrimination”), others have suggested that discrimination can only be understood through lived experience and encounters with different places (Valentine 2007). In these encounters, one subjectivity may be favored over another to negotiate a new place or scenario. In others, a more privileged subjectivity (maleness) can undo a more marginalized one (gay sexuality).

In the past decade, work on global sexualities has further destabilized ideas about Euro-American categories of sexualities. Many scholars have noted that the ideas animating much of sexuality and space studies, such as urbanity, rurality, liberalism, conservatism, and “coming out,” are profoundly Western and white. In addition, the places and events (e.g., San Francisco’s Castro, the Stonewall Rebellion in New York City) discussed in so-called gay and lesbian history tend to be located in North America and Europe. Work on South and East Asia, in contrast, shows that same-sex attracted communities are constructed flexibly alongside – or even as part of – religious, nonurban, and deeply familial communities that would be described as “traditional” in the West. By the same token, work on countries such as Brazil and Thailand has shown acceptance for same-sex affinities involving drag queens and other gender-bending performers, but less so for men that present as typically male (in Western countries, the opposite is often true). The study of global sexualities has broadened not only in its geographic scope but also in its scale: increasing mobility of individuals and information means that sexualities are also informed by cross-cultural encounters. A gay Filipino man who moves to the United States, for example, may find that the more effeminate performance of a bakla identity that would be accepted or even celebrated in the Philippines is marginalized and excluded in New York City. In contrast, Filipina domestic workers working in Hong Kong who have heterosexual identities at home may find that it is common within their specific community to have masculine female lesbian partners while working away.

The globalization of sexuality is now also increasingly seen in concepts of development and nation-building that have gained transnational popularity. There is, for example, an increasing association of sexual diversity with cosmopolitanism and the “creative city” notion popularized by Richard Florida. In this model, cities that are welcoming to sexually non-normative individuals are thought to also support artistic enterprise, innovation, and diversity in general, making them viable centers of tourism and capital.
accumulation. Consequently, countries such as Singapore have overturned strict antigay policies and begun holding pride festivals, even though the lived experience of being gay or lesbian is still subject to less strenuous but equally present regulation. Despite welcoming sexual diversity in theory, many jurisdictions have also tried to suppress the sexualization of urban space. The Giuliani administration in 1990s New York City, for example, removed sex work, sex shops, cinemas, and other visibly sexualized spaces from Times Square to secure the city’s place as a global tourist destination. At the national scale, non-normative sexualities have been employed in what Puar (2007) calls “homonationalist” discourses of nationalism, terrorism, and counterterrorism. In these scripts, governments and media outlets in countries such as Israel and the United States may portray national acceptance for non-normative sexualities as evidence of their own political liberalism even as they pursue aggressive military actions in Muslim countries. In European countries, such as the Netherlands and France, similar scripts have been used as justification for assimilationist integration policies that presume homophobia and intolerance among new, often Muslim, immigrants.

Some have argued that the recent emphases on sexual identity, the state, and its regulatory capabilities have obscured some of the practices and emotions associated with sexualities. At the same time, work in several areas is both returning to the particularities of sexualized everyday lives and extending sexualities work beyond the realm of cultural geography. Work on love within emotional geography, for example, has shown that – despite its status as an ideal in the eyes of the state – it can actually be a competitive and pressure-filled endeavor for single people, especially living in certain settings. Similarly, work in migration and population geography has shown how mobilities for gay men and lesbians are influenced by distinct historical contexts, institutions, and personal transitions they encounter across the life course (Lewis 2014). Geographers working on health and sexuality, a field long dominated by psychology, epidemiology, and other quantitative health sciences, have sought more to engage with the actual contexts (e.g., the clinic, the “gay scene”) in which elements of sexualities – including sex itself – might influence health outcomes (see Del Casino 2007). Sexualities therefore retain solid (and growing) intellectual purchase within geography into the second decade of the new millennium.

SEE ALSO: Bodies and embodiment; Cultural geography; Feminist geography; Gender; Intersectionality

References


SEXUALITIES


Further reading


Congressional districts, changing urban form, geomorphologic processes, and ecological fragmentation are geographic topics that draw from the conceptualization of shape. Shape is broadly defined as the geometric outline or boundary of an object that, internally, is typically homogeneous in nature. Some shape definitions regard only the outside outline as the shape while others recognize that interior variations modify shape. For example, some regard a solid object and a hollow object as having the same shape, and others do not. Mathematicians formally define shape as what is left after the properties of translation, rotation, and scale are removed (Kendall 1984). With this basic principle, shape measures describe or define an object’s shape qualitatively or quantitatively to assess whether two objects are the same or similar in shape. Qualitative descriptions use nominal values to describe shape. This works effectively with simple regular shapes such as circles, squares, and triangles. Objects with irregular and complex shapes are more challenging to assign to a specific name. Quantitative methods for both simple and complex shapes rely on assigning a number or an equation to describe an object’s shape. The field of statistics has a structured research program to quantify shape with methods relevant to fields including medical research, image analysis, and geography.

These conceptual and formal definitions of shape flow intuitively into geographical definitions of shape. Geographic objects are those regions on the surface of the Earth that are conceptually homogeneous, such as political, municipal, and property boundaries, geology type delineations, or clusters of events such as crime or disease instances. As in the formal mathematical definition of shape, two geographic objects are the same shape even when they are in two different locations, when they are rotated by any amount, and when they are a different size (Figure 1). With some exceptions, geographic objects tend to have complex shapes rather than simple mathematical shapes such as circles, squares, and triangles. Shape plays a fundamental role in geography through map projections, geographic representation, shape analysis, and several application areas.

Map projections

Shape is one of the fundamental properties either preserved or distorted in a map projection. A map projection is the transformation of the three-dimensional spherical coordinates of the Earth’s surface (e.g., latitude and longitude) to two-dimensional Euclidean coordinates on a flat plane. This process that creates representations of the Earth’s surface features results in distortion or preservation of one or more geometric properties such as area, shape, or direction. A map that preserves shape is known as a conformal map; it preserves the angles between intersecting lines in two-dimensional
SHAPE

Figure 1  Geographic objects with the same shape with variations in translation (left), translation and rotation (center), and translation and scale (right).

space. By preserving angles, the shape of surface features (e.g., boundaries of land masses) look basically the same as they do on a globe with distortion increasing away from the standard point or line.

Common conformal map projections include the well-known Mercator projection and the Lambert conformal conic projection (Figure 2). By examining the Tissot circles, Figure 2 shows that shape is preserved but area is distorted in both projections. The types of maps best suited for a conformal projection, important even in a digital age, include navigational charts, meteorological maps, and cadastral or parcel maps.

Geographic representation

Thematic maps, which display the spatial distribution of a specific subject matter, use point, line, and area symbols to represent the type or quantity of features. Area-based thematic maps, including area-class maps, categorical maps, nominal maps, and dasymetric maps, all use fill symbols within the boundary of a geographic object or line symbols along the boundary to represent the variations in the theme of the map. In other words, the shape of the geographic object influences the way the map is displayed. One distinction between the different types of area-based thematic maps is how the boundaries are derived (Chrisman 2002). In area-class maps, for example, the themes (or attributes) are controlled and the geographic boundaries are measured. Land-use maps fall into this category because the land-use categories – the attributes – are defined in advance and the boundaries are derived to maximize homogeneity within the boundaries. Choropleth maps, however, are different because in these types of maps the geographic boundaries are controlled and the attributes are measured. In census-based mapping, for example, census tract boundaries are predefined and the statistical data (e.g., population density or income levels) are measured.

Regardless of whether the attributes or boundaries are controlled or measured, the object’s shape depends on measurement scale. Measurement scale on the attributes refers to the level of detail and the number of categories represented. A finer measurement scale results in a classification system with more categories, affecting the shape, size, and number of geographic objects. Similarly, measurement scale on the boundaries also impacts geographic objects. A larger scale map will have more detail in the geographic object’s boundary and, more than likely, more geographic objects in the mapped
Figure 2  Mercator (top) and Lambert conformal conic (bottom) projections showing the preservation of shape through Tissot’s indicatrix.
SHAPE

region. So while comparing shape remains invariant under scale, how real-world entities become geographic objects (thus becoming a shape) depends on both attribute and spatial scale.

Geographic shape analysis

Geographic shape analysis is a systematic process that uses the qualities of a geographic object’s boundary or outline to gain insight into spatial and temporal processes. The qualities of shape are described qualitatively and quantitatively. Qualitatively, an object such as Italy could be described as having a “boot-like” shape. Perhaps the most pervasive use of nominal terminology in geography is geomorphic forms such as deltas (e.g., bird-foot delta) and drainage systems (e.g., dendritic). However, the shape of geographic objects is often difficult to describe qualitatively so quantitative measures are used when greater precision is needed, the comparison of similar shapes is needed, or there are reasons to believe that shape reflects important underlying processes.

Quantitative shape descriptors involve a mathematical metric or equation to characterize geographic objects with a simple or complex reproducible representation. Simple metrics, such as most compactness measures, are often indices, with results between 1.0 and 0. Shapes closer to 1.0 are more compact and closer to a circle (the most compact shape). Figure 3 illustrates the results of two such measures, the iso-perimeter quotient (IPQ) and the normalized moment of inertia (NMI), on three different shapes (Li, Goodchild, and Church 2013). IPQ uses an object’s perimeter as a parameter in the compactness measure, and so the results are much influenced by the convolution of an object’s boundary. Therefore, although the left and center objects in Figure 3 appear similar in shape, their compactness indices are quite different. The far right object has a convoluted shape and therefore receives a very low score by IPQ. The NMI, on the other hand, relies on the variability of the boundary’s distance from a central point rather than area-to-perimeter ratio to determine compactness. In addition to compactness measures, other simple shape measures include elongation.

![Figure 3](image-url) Comparison of IPQ and NMI compactness shape measures for three different shapes.
measures, direct comparison with a standard shape, and dispersion methods (Wentz 2000).

Complex shape measures result in an equation or set of numerical indices to describe an object’s shape. Fractals for example, widely used for simulation and analysis across many fields, use the principles of self-similarity to examine the relationship between scale and geographic boundaries. The resulting index indicates the complexity of the boundary at a given scale. Fractals have been used to quantify the dynamics of urban form, coastline processes, and boundaries between ecological zones (see Xu et al. 2014; Frost et al. 2005). Fourier harmonic analysis, which also focuses on the complexity of the outside boundary of a geographic object, provides a method for quantifying shape that is reversible. That is, the shape of the object can be reconstructed from the mathematical harmonic analysis (Moellerling and Rayner 1981; Akamatsu, Takayama, and Ikeda 2012). The method uses the eigenvalues and eigenvectors from a discrete Fourier transformation as indices to quantify spatial form.

Shape is conceptually and empirically incorporated into geographic studies and analysis to support spatial pattern analysis, to support image classification, and to correlate geographic form and process. Spatial patterns emerge through the composition and configuration of shapes on the landscape, which gives rise to questions on how and why these patterns occurred where they did and how they change over time (Williams and Wentz 2008). Recognizing the regularity of shapes in the landscape, algorithms to classify satellite images use shape metrics to improve classification results by extracting features with known shape parameters (Jiao, Liu, and Li 2012). Finally, geographic theory suggests that human and physical landscapes contain structures and forms that are strongly influenced by the underlying processes that gave rise to those forms (Walsh, Butler, and Malanson 1998; Bunge 1966). Shape analysis provides a fundamental mechanism to correlate form and process.

**SEE ALSO:** Data model, object-oriented; Fractal analysis; Point pattern analysis; Spatial analysis

### References


In September 1956 a conference was held in Alba, Italy, that brought together avant-garde revolutionary groups from all over Europe. The conference created an accord between three of the attending groups, the Lettrist International (led by Guy Debord), the Scandinavian and German based Movement for an Imaginist Bauhaus, and the London Psychogeographical Association, to disband and reform as a single organization committed to the revolutionary critique and subversion of contemporary forms of alienation. This organization was called the Situationist International (hereafter SI). Although membership of the SI never rose above 70, through its journals (especially the Paris-based Internationale Situationniste) and agitational role within the French student revolts of the late 1960s, the group managed to attract a level of recognition and notoriety, particularly within the French far left, out of all proportion to its size.

The two theorists who dominated the movement were Guy Debord and, to a lesser extent, Raoul Vaneigem. Debord and Vaneigem’s most influential works, respectively, The Society of the Spectacle and The Revolution of Everyday Life, were both first published in 1967 and were intended to complement each other (the first English translations appeared in the early 1970s). Both texts express what can be seen as situationism’s unique theoretical contribution, a rebooting of Marxism for the late twentieth century; more specifically, the development of a critique of “the spectacle,” the term Debord and Vaneigem gave to the massively reified, appearance-based, nature of contemporary life. Debord describing the onset of spectacular society, suggested:

The first phase of the domination of the economy over social life brought into the definition of all human realisation the obvious degradation of being and having. The present phase of total occupation of social life by the accumulated results of the economy leads to a generalised sliding of having into appearing, from which all actual “having” must draw its immediate prestige and its ultimate function. (Debord 1977, thesis 17)

It is no coincidence that the favorite social metaphor employed by situationist authors was of that archetypal modern cultural product, the motion picture. It seemed to them that all society had turned into one giant movie spectacular, with a transfixed and willing audience. The situationists insisted that Marx’s (Marx and Engels 1970, 96) dictum that “men make their own history … but they do not make it under circumstances chosen by themselves” can only be understood in the second half of the twentieth century through an appreciation of the centrality and totality of the “image society.” “The spectacle,” wrote Debord (1977, thesis 34), “is capital to such a degree of accumulation that it becomes an image.”

After the events of May 1968 and a series of acrimonious expulsions, however, the SI lost momentum and dissolved itself in 1972. Although described by Berger (1975, 600) as “one of the most lucid and pure political
formulations of the 1960s,” until the 1980s the situationists tended to be dismissed both by orthodox radical opinion and by the mainstream press as fanatics whose extravagant rhetoric exceeded their political or intellectual significance. The situationists were brought back into fashion in the 1980s, in part because of the rise of postmodernism, a current that in both intellectual and artistic life encouraged an interest in the transgressive, playful, and everyday as forms of political activity (Bonnett 1989).

Situationism is unique among radical theories in placing geography at the center of its ideas and activities. Place and space were at the heart of their investigations. In part, this was a result of their suspicion of modern town planning, which they regarded as reactionary and politically pacifying. “The whole field of urban planning” suggested Vaneigem and Kotanyi (1981, 66) should be understood “as the organization of participation in something in which it is impossible to participate.” What is required, proposed the situationists, is a new form of geographical investigation that can enable the revolutionary reappropriation of the landscape. “Proletarian revolution is the critique of human geography through which individuals and communities have to create places and events suitable for the appropriation, no longer just of their labor, but of their total history. In this game’s changing space, and in the freely chosen variations in the game’s rules, the autonomy of place can be rediscovered” (Debord 1977, thesis 178).

The techniques involved in this form of reappropriation revolved around the idea that, through the instinctual exploration of the emotional contours of one’s environment, playful and anti-authoritarian places and journeys can be discovered or created. This process was supposed to enable the construction of revolutionary “situations,” hence the appellation “situationists.” What the SI called their “psychogeographical” investigations adopted the practice of the “dérive” (literally translated, “drift”) that involved an unstructured wandering through the landscape, allowing oneself to be drawn consciously and unconsciously toward those sites and movements that heighten one’s experience of place and disrupt the pacifying banality of everyday life as it is lived in “the society of the spectacle.”

The situationists proposed that new maps expressing psychogeographical possibilities and explorations should be drawn up and used as blueprints for a situationist town or “utopian center.” The visionary nature of these proposals was most fully articulated by Chcthegov (1981), who looked forward to the division of cities into zones of emotional intensity (for example, the Bizarre Quarter, Sinister Quarter, Noble Quarter) through which people would be in perpetual dérive. Other, slightly less extraordinary schemes suggested the setting up of sites of psychogeographical experimentation that could “function as bridgeheads for a new invasion of everyday life” (SI, quoted in Barrot 1987, 24).

As Marcus (1989, 170) notes in Lipstick Traces, a work that traces the creative links that stretch from Dada to the situationists to punk rock, for Debord and his coworkers, “to pursue the dérive” was “to give yourself up to the promises of the city, and then to find them wanting – to drift through the city, allowing its signs to divert, to ‘detourn’ your steps, and then to divert those signs yourself, forcing them to give up routes that never existed before.” The dérive, Marcus continues, was supposed to contribute to the development of a revolutionary “way of life that anyone could understand and anyone could use.”

The practical activity that emerged from the theories of the SI ranges from seemingly inconsequential rambles around European cities to
SITUATIONISTS/SITUATIONIST GEOGRAPHY

relatively rigorous and well-documented experimental activity. In either case, however, the primary intention was not to release the unconscious or to admire the delights of “chance,” as it had been for the surrealists, but to think about and engage physically with the possibilities of creating a new and radically stimulating kind of city.

Over the past three decades situationist geography has developed a scholarly following (see Bonnett 1989; Pinder 2005). Much of this attention is focused on the radical nature of “everyday space” and the “suppression of the street” augured by modern traffic and modern planning. Outside of the academy, the situationists have had an even greater influence. The early 1990s witnessed the birth of local psychogeography groups across the United States and Europe. From the late 1990s activity passed to the arts community and subsequently, to “urban explorers,” who investigated the hidden and exciting within the seemingly ordinary world of the everyday city.

SEE ALSO: Phenomenology; Radical geography; Urban geography

References


Further reading

The concept of skill is important to many disciplinary fields in the social sciences including economics, sociology, urban studies, labor studies, and human geography. It has also been closely linked to some of the most significant debates in the social sciences over the past half-century including the “deskilling debates” of the 1970s and 1980s regarding the growing levels of income inequality and ideas emanating from human geography about creativity and innovation. An apparently simple idea, the concept of skill, on reflection, proves to be rather more complex and perhaps even controversial. Not only are there definitional ambiguities related to skill but the way in which the concept is used in academic and mainstream discourses has invited much confusion. In this entry, skill is treated mostly in relation to economic activities rather than social development or educational outcomes, though these latter issues are often closely linked to economic outcomes.

After a brief section detailing the definitional issues related to skill – much of this work having been carried out by human development scientists and psychologists – the first major section of this entry will examine how skill has been understood in economics and sociology. It has been in these two fields that the bulk of social scientific theorizing on skill has been undertaken. What is most interesting about connecting some of the definitional issues studied by scientists with the research carried out by social scientists is how the work of the former intimates or foreshadows the conclusions drawn by the latter. This is particularly the case with respect to the perceived scientific distinction between cognitive skills and motor skills, a distinction that has filtered into social science as a way of distinguishing “high-skilled” from “low-skilled” workers. One of the most prominent and elegant theories of skill has been provided by economists in the form of human capital theory as a way of explaining divergent labor market outcomes based on the skills and qualifications of individual workers. Many sociologists have been highly critical of human capital theory, arguing that notions of skill are often socially constructed to fulfill particular social and economic goals. This is most evident in the vast literature from within industrial and labor sociology which identifies the skill levels and attributes of workers in various occupations over time.

Following the description of the ways in which skill has been understood by economists and sociologists, the second major section focuses on how the concept of skill, and debates from other fields, have been translated into human geography. Since the 1970s, skill has been examined in a variety of different ways by geographers, mostly within the subdiscipline of economic geography. Since the political economist Adam Smith developed the idea of the “division of labor,” scholars have been interested in the ways in which workers of different skill levels are integrated together in the production process. In many ways, economic geographers and regional scientists have been particularly attentive to issues of skill because one of the key goals of these fields has been to map how divisions of labor are organized and managed across space.
This is related to perhaps the broadest question at the heart of economic geography: What explains why some regions succeed economically while others fail? Since the mid-1990s, renewed interest in this question has spawned important debates where skill – either explicitly or implicitly – is at the center of attention. Many urban and regional economies in OECD (Organisation for Economic Co-operation and Development) countries are increasingly being defined by the spatial and social polarization between “high-skilled” knowledge workers and “low-skilled” service workers in personal care, retail, and food services. Skill, creativity, innovation, and economic growth have never been distributed evenly across space. The questions which currently animate much of economic geography revolve around the study of exactly which factors explain the clustering of so-called skilled knowledge workers, whether such clustering provides a good model of urban economic growth, and whether cities can be designed to attract knowledge workers. Beneath the surface of these debates, however, reside foundational issues long studied by social scientists about what skill is, who is considered skilled, and what functions these definitions serve.

**Definitional issues and the science of skill development**

The dictionary definition of skill reveals some of the complexity and ambiguities of the concept. Each of the subdefinitions from the dictionary has at its core the notion of competence or expertise: “the ability to do something well,” according to the *Oxford English Dictionary*. Should these abilities, however, be considered natural gifts or as having been acquired through training? Do these abilities belong in the realm of physical dexterity and coordination or of knowledge and intellectual activity? In general use, “skill” often refers to both mental and physical processes and is understood as the result of both training and genetic predisposition. Scholars in the human sciences – especially those interested in child development – have studied the ways in which humans acquire skills, and their conclusions are very relevant to the study of skill by social scientists.

Scientists tend to distinguish between motor development, which governs physical skills, and cognitive development, which covers mental processing skills. The traditional view among scientists has been that there is an important distinction between these two forms of development, not only in the skills they produce but also in the processes which initiate such skill development. As prominent neuroscientist Adele Diamond has stated, “motor development and cognitive development have been studied separately” and “have generally been viewed as independent phenomena, although occurring in the same organism over the same time period” (Diamond 2000, 44). The scientific rationale for such a distinction has been based in the historical understanding of different locations of brain activity for motor and cognitive skills. Motor skills have been thought to originate in the cerebellum and to develop earlier in childhood. In contrast, cognitive skills have been linked to the prefrontal cortex, which does not fully develop until later in adolescence. This simple mind/body dichotomy has been incredibly impactful because it supports the view of motor skills as basic and cognitive skills as more advanced.

Yet new research using medical technologies such as functional neuroimaging is beginning to show the errors of this traditional view. Not only are motor and cognitive skills intimately related in terms of brain activity, but certain advanced (or “fine”) motor skills take equally long to develop in the human brain. Neither
As Diamond argues, “Motor development and cognitive development may be much more interrelated than has been previously appreciated. Indeed, they may be fundamentally intertwined” (2000, 44). Since the early 2000s, significant progress has been made in the scientific understanding of skill development and the neurological processes which support such development. The mind/body dichotomy is increasingly being supplanted by integrative understandings of how motor and cognitive development are mutually reinforcing. This type of thinking has not fully diffused into the social sciences where there is still a strong assumption of high skills being related to cognitive abilities and lower skills being related to physical or emotional tasks.

The economics and sociology of skill

In a classic article on the subject entitled “What Is Skill?” sociologist Paul Attewell (1990) details many of the challenges related to the study of skill in economies. Most importantly, he outlines four main approaches through which skill has been studied in the social sciences: positivism, ethnomethodology, Marxism, and Weberian social constructionism. The first, and perhaps most prominent, approach is the positivist view, which is most closely related to human capital theory in economics. Human capital theory was established by the economist Gary Becker to account for the knowledge and skills that an individual has gained through education, job training, or work experience. According to neoclassical economics, human capital can be considered either as an investment in training made by an individual or as skills that add value in the production process. The amount of human capital – or skill – that each worker has is seen as directly proportional to the amount of wages that can be earned: “Wages are seen either as a return on an individual’s investment in human capital or as determined by the marginal utility of the skill attained” (Attewell 1990, 425).

Given this elegant argument that an individual’s investment in their own skills will result in higher productivity and ought therefore to be rewarded with higher wages, the key question is how human capital should be measured by economists. Human capital is usually measured as the sum of years spent in formal education or job training, as well as years of job experience. This approach is beneficial in that it provides a clear and consistent way of measuring human capital across occupational segments. In general, economists have consistently shown that workers with greater levels of human capital are rewarded with higher wages and participate more fully in the labor market. Yet, many sociological theorists have been critical of human capital theory because of the rather tenuous links between educational achievement, skill acquisition, job performance, and income. As Attewell (1990, 426) remarks, human capital theory:

(a) … abstracts from such phenomena as credentialism (whether college degrees really provide job skills or function as a rationing device); (b) the impact of occupational monopolies (unions and professional and state certifications); (c) the relationship of sex-typing of occupations to their perceived skill; and (d) the effects of ongoing gluts of certain skills and shortages of others on the social and monetary valuation of those skills. From a sociological viewpoint, therefore, human capital represents a shortcut around studies of skill and income attainment by assuming a fairly direct equivalence between education, skill, and reward.

There are two other issues that arise in relation to human capital theory. First, a bias can clearly be detected between skills that are considered cognitive and those that are considered motor
development. That years of schooling are the key measure through which levels of human capital are determined suggests that cognitive skills are more valuable than motor skills, social skills, or emotional skills. Yet a growing body of research from various different academic disciplines is showing that work that has traditionally been considered “low-skill” in fact requires high levels of skill that are not visible in human capital measures.

One of the original lines of analysis of this sort came from sociologists in the field of ethnomethodology, the second approach detailed by Attewell (1990). For scholars such as Kusterer (1978), the level of skill of individuals was something that must be studied in extraordinary detail rather than something to be assumed from level of education. A core idea of ethnomethodology is that all human activity is rather complex and is oriented toward fitting into the existing social order. In the work sphere, this means that scholars would expect there to be high levels of skill in occupations that have traditionally been considered unskilled. Kusterer (1978) was at the forefront of this type of research which involved participant observation and in-depth interviews. Studying machine operators and store clerks, Kusterer found that most skill is invisible to outsiders. Such researchers argue that the most skilled workers are those that make their jobs look routine, irrespective of their levels of human capital, since they have mastered all the nuances of their work.

More recent research from within sociology and labor studies has extended these basic ideas to study the key skill capacities of so-called unskilled service workers, who are a growing part of many urban economies. Despite their relatively low levels of human capital, many service workers have important physical, social, and emotional skills that they utilize in their work (Hampson and Junor 2010). This point is related to the second major problematic with human capital theory. The view of work in human capital theory is one in which each individual can be effortlessly distinguished for the value that they have brought to the production process. As more types of work become team-oriented, however, it is possible that individual performance related to cognitive skills becomes less important than the social and emotional skills involved with teamwork. More fundamentally, we might ask whether skill is best treated as an attribute of a person or whether the ability to use certain types of skill are largely determined by the nature of the job itself (which is often not controlled by the worker).

This distinction becomes important when considering how Marxist scholars have thought about skill. Marx himself announced early in *Capital* that he would treat every type of labor as “simple unskilled labor.” This is representative not only of the rising importance of industrial labor at the time of writing, but, more crucially, of the importance Marx placed on the ownership over the means of production in a capitalist economy. Where Marx does discuss skilled labor, he argues that the higher cost of skilled labor is required to pay for investment in skills training. Marx, in some ways, prefigures human capital theory (Attewell 1990). Many later Marxist scholars would develop more sophisticated ways of thinking about skill. The most famous of these was Braverman (1974) who developed his deskill thesis in line with Marx’s notion of alienated labor. For Braverman, the imperatives of capitalist competition and the falling rate of profit necessitated the constant search on the part of the owners of capital for efficiencies in the production process.

According to this view, the related rise of Fordist mass production on the assembly line and Frederick Winslow Taylor’s techniques of “scientific management” were an attempt to wrest
control of the production process away from skilled workers and toward unskilled workers who could be paid less. Extending the division of labor may not essentially change the skilled character of workers but it does reduce the task complexity of jobs and therefore affords reductions in the cost of labor. In this sense, the job itself is first deskilled before the worker’s previously valuable skill becomes outdated. Any job in which technology and extended divisions of labor can be applied might be at risk of deskilling.

Braverman’s 1974 book *Labour and Monopoly Capital* spawned intense debate across the social sciences between those who reported deskilling occurring in workplaces and those who saw problems with both the theory and the empirics used in its support (Attewell 1990). This debate, which not coincidentally corresponded with an age of deindustrialization and stagnation in many Western economies, was never fully resolved. There was, however, a sense that with the emergence of the postindustrial and “knowledge-based” economy in the 1990s, the worry over deskilling had subsided. Some semiskilled industrial jobs had been lost to deskilling and offshoring but many new skilled jobs were emerging, sometimes in entirely new economic sectors.

In an interesting twist, recent research by labor economists is suggesting that job routinization (a process resembling deskilling) is becoming endemic with the ever greater use of information and communication technologies (Autor, Levy, and Murnane 2003). This push for routinized work has not stopped at the occupational level of unskilled and semiskilled workers but instead has rapidly breached so-called skilled job sectors, threatening the job security and wages of professional workers in law, accounting, engineering, science, and software development (Blinder 2006). The outcome of these processes is that many OECD countries are currently experiencing significant polarization in the skill requirements of jobs between low-wage service workers (some of whom are economic migrants) and those professional workers who still have the opportunity to fully utilize their knowledge and creativity (Autor 2010).

This final point about the declining protection of skilled professional workers from competition and routinization is related to the last approach to understanding skill as outlined by Attewell (1990). For those in the Weberian tradition of sociology, the real task is not to identify the best way of defining skill in the economic sphere once and for all, but rather to appreciate that human capital often has very little to do with economic standing. This approach has two interrelated branches. Some scholars focused on how market factors of supply and demand, rather than skill composition, determine wages to argue that there are moments when relatively low-skilled work can be paid handsomely. The other branch focused on ideological factors to argue that the idea of skill is a vehicle used by certain groups to promote or maintain their own power and prestige. This notion of the social construction of skill is related to the market factors approach because the ideology of skill is often used to improve supply and demand conditions in certain occupational segments. In Weber’s own writing, occupational groups are considered as another social group interested in securing its own goals (often at the expense of others). The clearest way in which they do this is by restricting access to the social group so as to reduce the number of competitors for supplying such skills in the labor market. Medieval guilds, craft unions, and modern-day credentials work on this basic principle, often sanctioned by state bodies.

If the income level and prestige of various jobs is mostly a product of social manipulation by those that fill the occupation, what does this suggest about the actual skills that are needed
to complete certain worker tasks? Among neo-Weberian sociologists some have taken a very strong stance that the social construction of skill has successfully created occupational power even in the evident absence of task complexity. Others have taken a weaker stance that some “real” skills are necessary to carry out work tasks but that these do not fully explain the different standing of occupational groupings. What we have seen, then, is that over the past half-century economists and sociologists have been engaged in serious debates about the meaning and significance of skill in economic activity.

Skill in economic geography

Economic geography has become an immensely diversified field of study since the 1980s, with a plethora of coexisting theoretical, methodological, and empirical approaches. This is a dynamic academic discipline which perhaps complements its topic of study: an ever-changing global economic system. Moreover, as economic geography has widened its scope of analysis, it has come into closer contact with related academic fields including economics, economic sociology, urban studies, labor studies and international political economy. Yet, despite the changing directions and orientation of economic geography, the main objective of the field has remained much the same: to trace in real time the changes that have been occurring in the spatial dynamics of economic activity.

The notion of skill has been central to these efforts. What is most interesting is the changing way in which skill has been studied and used in contemporary economic geography. Indeed, it is possible to track the real-time changes in global economic activity alongside the particular interests of key economic geography research. When this is done, it becomes clear that skill has been at the heart of some of the most important debates in the field while also being one of the key topics on which economic geographers remain divided and fragmented.

In the 1970s and 1980s many Western economies experienced plant closures, jobs losses, and a general decline in their industrial sectors. Economic geographers made key contributions to understanding these phenomena. Massey (1984) argued that growing international competition among jurisdictions for industrial activity – largely carried out by so-called unskilled and semiskilled workers – was leading companies to relocate production in search of cheaper labor costs. The labor arbitrage strategies of multinational firms were seen as a key element in explaining the deindustrialization of Western industrial regions during the 1980s. Relatedly, economic geographers developed theories of “uneven development” to show how the owners of capital leveraged their spatial mobility in order to find the most profitable locations for production (Smith 1984). Capital flight not only shaped levels of economic development across space but also disciplined workers through threats of plant relocation.

As a result of deindustrialization, the proportion of overall employment in manufacturing jobs declined considerably across most Western countries (Pilat et al. 2006). During this time, while many economic geographers focused their research on the challenges of economic restructuring, others began discussing the apparent rise of postindustrial or post-Fordist economies (Scott 1988). The basic argument was that newly industrializing countries in the Global South would have a comparative advantage in production that required unskilled or semiskilled workers, while Western economies had an advantage in the production of knowledge and information which was an increasingly important component of economic growth. Economic geographers
across various subfields have appreciated that the fundamental feature of the global economy is the way in which different places are interconnected through trade relations (Dicken 1998). It is not possible to have a place defined by knowledge production unless that place is connected to resource, agricultural, and industrial peripheries somewhere else. Indeed, this is precisely the shape that the global economy has taken.

Since the 1990s, as this postindustrial economy came to fruition to varying extents across space and time, much contemporary economic geography has been concerned with the nature and spatial dynamics of the “knowledge economy” or the “learning region” (Morgan 1997). Whether it is acknowledged or not, this is largely about skill. There is a strong assumption in much of this research that successful economic places – global cities, creative cities, and so on – are defined by their highly skilled workforces. A key area of interest has thus been about cities, since city-regions have been the main site through which knowledge economies defined by innovation, research and development, business services, and cultural production are organized. Economists have demonstrated that workers with high levels of cognitive skill – or human capital – generally earn higher incomes and are increasingly clustering in “skilled cities” (Glaeser and Saiz 2003). As a result, many economic geographers have studied the conditions under which such knowledge economies grow, though there has been an interest in labor migration and the challenges facing unskilled workers as well.

Most commentators agree that knowledge economies are defined by the agglomeration, or clustering, of innovative people and firms in a specific location. These agglomeration economies provide mutual benefits to firms and skilled workers. Firms benefit from reduced transactions costs, a larger labor pool, and the potential for knowledge spillovers that potentially generate innovation. At the same time, workers gain from access to a greater number of firms, as well as the opportunity to develop social networks that can yield employment or entrepreneurial opportunities. The unanswered question remains how these agglomeration economies are catalyzed and to what extent cities can themselves establish the conditions to achieve such firm clustering.

Within economic geography there has been a recent debate about the mechanism through which agglomeration is initiated. Richard Florida (2002) argues that workers are the instigators. Cities that offer the best and greatest number of amenities will attract the best and most talented workers, which will in turn draw innovative and profitable firms. In contrast, Michael Storper and Allen Scott (2009) suggest that firms come first. Talented workers are attracted to a city not because of its amenities but because of jobs that have been created by firms already there. While the exact mechanism remains somewhat unclear, what is evident is that so-called skilled and unskilled workers are increasingly residing in different types of cities (Glaeser and Saiz 2003).

This polarization of economic opportunities across space is leading to new geographies of uneven development. At the urban scale, the polarization of income groups within the city is increasingly problematic for social and economic reasons (Wilkinson and Pickett 2009). At the same time, however, not all successful regional economies in Western countries are knowledge-based and this perhaps provides some hope for regions not defined by knowledge activities. Likewise, not all cities that have aimed to transition to a knowledge economy have succeeded, and this reminds us that the same imperatives of competition which impacted unskilled workers in the 1980s may now be impacting skilled workers as newly developing countries such as China, India, and Brazil increasingly
prepare for their “knowledge economy” future. In both the United States and Canada, some of the fastest-growing regional economies, which are providing high wages, are defined by resource extraction activity. In Germany, manufacturing has remained an important part of employment and economic growth (Pilat et al. 2006). What economic geography tells us about skill, above all else, is that there is nothing static about how certain skills are valued over time and across space in a globalizing economy.

**Conclusion**

There are many ways to understand skill in the social sciences and it is a term that is mobilized in various ways to suit the interests of different social groups. Most scholars agree that skill represents an endowment of capabilities that enable an individual to contribute to the production of valuable goods and services. Where differences of opinion emerge are with how different sets of skills are defined, measured, and rewarded.

This entry began with a discussion of how skill as a human development trait is currently being understood by scientists. It appears that the scientific consensus is moving to a view of skill as a multifaceted and complex human development process instead of previous dichotomous conceptions of physical skills as rudimentary and cognitive skills as more advanced. This new orientation toward skill raises important questions about how the categories of “skilled” and “unskilled” workers have long been activated in the economic sphere. A scientific view of skill in the workplace would question why a marketing analyst with a university degree is seen to be more skilled than a personal care worker or a flight attendant. Indeed, as has been shown, this view is emerging across the academic social sciences through research by economists, sociologists, economic anthropologists, and geographers.

The more fundamental question remains, therefore, on what basis economic participation should be rewarded. Income inequality has been rising dramatically across most OECD countries since the 1980s (IMF 2014), puncturing the smooth Kuznets curve view that inequality declines as economies develop. This growing disparity in incomes and wealth is not only a threat to social cohesion (Wilkinson and Pickett 2009), but it very likely represents a significant dampener on economic growth as workers struggle to consume (IMF 2014).

Beyond academia, issues of skill in the workplace are increasingly becoming fodder for film and TV franchises. One of the most popular new shows on North American television over the past few years has been *Undercover Boss*. Each week, the CEO of a major company takes on the guise of a training employee to see how the business operates “on the ground.” The most memorable moments of each episode are reliably when the CEO cannot perform the task being taught to them by the employee mentor. The viewer is constantly asked to reconsider which one of these two people is the “skilled” worker. Yet, at the end of the show, the stunt is revealed, the CEO learns important lessons about how to improve their business, the employee-mentors are surprised with a substantial gift, and there is a lot of hugging and crying. Needless to say, the broader world of work does not reflect these understandings of how much skill the “unskilled” worker actually has.

As scientific evidence is increasingly suggesting, skill is not an exclusive cognitive category for those who have received a certain level of education. There are multiple forms of skill that workers perform during the process of production. Skills have been recognized and rewarded differently over time and space.
Geographers have made important and lasting contributions to social science debates by highlighting the locational dynamics surrounding economic skills. We can only assume that there is more research to be done as the nature and operation of skill shifts in a fast-changing economic world.

SEE ALSO: Creative class; Deindustrialization; Deskilling; International division of labor; Knowledge-based economy; Labor migration

References


Description and purpose

The Institute of Geography is a scientific center of the Slovak Academy of Sciences involved with research in several branches of geography. The mission of the institute is to carry out basic research into spatial structures and linkages existing between the natural and socioeconomic systems with special regard for Slovakia. The institute accomplishes its goals through its journals, postgraduate educational programs, research grants and contracts with academic institutions and government agencies, and participation in numerous national and international conferences. Currently the institute consists of three scientific departments: Physical Geography, Geomorphology and Natural Hazards; Human and Regional Geography; and Geo-Information Sciences.

Journals or major publication series


Current activities or projects

The institute actively collaborates with partner universities and academic institutions in Slovakia and abroad. The institute participates in activities of numerous international organizations (e.g., IGU, ICA, IUCN, IALE, and MaB). The IG actively participated in preparation of the Regional Conference of the IGU (Prague 1994) and the 2nd EUGEO Congress (Bratislava 2009), and in the edition of Population Atlas of Slovakia (2006) along with the monograph Demo-Geographical Analysis of Slovakia presenting changes in Slovakia after 1989. Participation of the institute in the international projects financially supported by the EU: CORINE land cover 2006, PANet2010, INFRAREGTUR, and VITAL LANDSCAPES.

Brief history

The institute was founded in 1943 as the Geographical Institute of the Slovak Academy of Sciences and Arts. Apart from geographical
research, studies in sociology, ethnology, and demography were pursued in the institute as well. The second stage of the institute’s development (1953–1963) started with the founding of the Slovak Academy of Sciences (SAS). At the beginning the focus was in tasks dedicated to the physical-geographical themes, later the asymmetry was corrected as additional themes were added. The institute has used its present name since 1963. The third stage (1964–1989) involved tasks imposed by the governmental plan for the basic research, including work synthesizing research on landscape and environment. Results of the research were incorporated in the monumental work, the *Atlas of the SSR* (1980). The institute entered the fourth stage of its history after 1989 when the new political and socioeconomic situation were reflected in its different forms of organized research-scientific activities. Future opportunities are defined in four research clusters: structure and dynamics of natural landscape, hazards and risks; landscape changes explored by application of remote sensing data and the GIS; sustainability and quality of life in changing environments; society in flux, spatial disparities; and local and regional development.

Submitted by Vladimír Ira
Slovakia: Slovenská geografická spoločnosť pri Slovenskej akadémii vied (SGS) (Slovak Geographical Society at the Slovak Academy of Sciences)

Founded: 1955
Location of headquarters: Bratislava
Website: www.sgs.sav.sk
Membership: 290 (as of January 1, 2014)
President: René Matlovič
Contact: rene.matlovič@unipo.sk

Description and purpose

The SGS was founded to support geographic education and professionalization and the dissemination of geographic research and knowledge in Slovakia. The members are mostly scholars, teachers, researchers, and other professionals in Slovakia. It has six regional branches (Bratislava, Prešov, Nitra, Banská Bystrica, Košice, and Ružomberok) covering three areas of geography (theoretical geography, applied geography, and geographic education). The society organizes a congress every 4 years as well as various other events (conferences, field trips, workshops, lectures, exhibitions, and expeditions). SGS is a member of EUGEO (Association of Geographical Societies in Europe).

Journals or major publication series

Spravodaj SGS

Submitted by René Matlovič
Slovenia: Geografski inštitut Antona Melika ZRC SAZU (Anton Melik Geographical Institute ZRC SAZU)

Founded: 1946  
Location of headquarters: Ljubljana  
Website: http://giam.zrc-sazu.si/sl#v  
Membership: 36 (as of 2014)  
Director: Drago Perko  
Contact: drago@zrc-sazu.si

Description and purpose

The Geografski inštitut was founded in 1946 by the Slovenian Academy of Sciences and Arts. In 1976 it was named after Slovenia’s greatest geographer, academy member Dr Anton Melik (1890–1966), who served as the institute’s first director. Since 1981, the institute has been one of the members of the Scientific Research Center of the Slovenian Academy of Sciences and Arts. In 2002 the Institute for Geography (established in 1962) and the Geographical Museum of Slovenia (established in 1946) joined the institute.

Journals or major publication series

*Acta geographica Slovenica: Geografski zbornik.*  
http://ojs.zrc-sazu.si/ags  
*Geografija Slovenije.*  
http://giam.zrc-sazu.si/sl/publikacije/geografija-slovenije#v  
*Georitem.*  
http://giam.zrc-sazu.si/sl/publikacije/georitem#v  
*GIS v Sloveniji.*  
http://giam.zrc-sazu.si/sl/publikacije/gis-v-sloveniji#v  
*Naravne nesreče.*  
http://giam.zrc-sazu.si/sl/nns#v  
*Regionalni razvoj.*  
http://giam.zrc-sazu.si/sl/publikacije/regionalni-razvoj#v

Current activities or projects

The institute participates in numerous projects in Slovenia and abroad, organizes academic conferences, trains junior researchers, and participates in professional exchanges. In the past ten years, the institute’s research team has published over 3000 bibliographic units and made over 500 presentations at conferences in Slovenia and abroad.

Most of the institute’s research work derives from its long-term research program Geografija Slovenije (Geography of Slovenia). In 2013, the research work supplemented seven basic and seven applied national projects, mainly funded by the Slovenia Research Agency, as well as twelve international European projects.

Brief history

The institute’s main task has always been to conduct basic and applied geographical research on Slovenia and its landscapes as well as to prepare basic geographical texts on Slovenia. Since Slovenia gained independence the institute has prepared a large variety of basic geographical works on Slovenia, including national, world, school, and census atlases, a dictionary of geographical terminology, a lexicon of Slovenian places, and a regional and general monograph.

Submitted by Drago Perko
Slovenia: Zvezi geografov Slovenije (Association of Slovenian Geographers)

Founded: 1922
Location of headquarters: Ljubljana
Website: http://zgs.zrc-sazu.si/sl-si/domov.aspx
A slum is a squalid settlement of substandard housing and living conditions where inhabitants lack fundamental resources like shelter, job opportunities, water, and food. While conditions vary from place to place, most slums are heavily and densely populated, have a substandard housing stock, contain a large pool of underemployed and unemployed workers, and lack reliable utilities and services such as clean water, electricity, sewers, and waste collection. In addition to being underserviced, most slums do not contain or have easy access to public institutions such as hospitals and schools.

The first large slums originated during the rapid migration, industrialization, and urbanization processes that occurred during the Industrial Revolution in Europe and the United States. New York City is considered to have had the first slum in the late 1700s, Five Points, in what is today’s Little Italy in Manhattan. Five Points housed newly arrived freed slaves and immigrants of Irish, Italian, and Chinese origin who came to New York City for work opportunities and the prospects of a better life. Five Points was situated near the center of the urban area close to numerous large factories, and was known for its filthy, disease-ridden conditions caused by trash accumulation, a poor sanitation system, and polluting industries that occupied the city center. Slums in the twenty-first century are predominantly found in urban areas of underdeveloped and developing counties.

Today, slums and slum residents are widely stigmatized as being socially dysfunctional, carriers of antisocietal norms, and characterized by an absence of morals. News and media portrayals of slums disproportionately invoke realities of crime, gang violence, drug abuse, and individual social malaise. Yet, while such outside perspectives may see disorder and aberrance, life in slums is typically organized and replete with nourishing social networks, engaged leadership, and strong communication channels. Lost in the common caricaturing and stereotyping of people and place is the reality of people practicing superior survival skills, resourcefulness, and adaptability to the challenging circumstance of difficult everyday conditions.

Slums of the world

Slums exist across the world in developed, developing, and underdeveloped countries. UN–Habitat (2012) estimates that the proportion of urban people living in slums in developing countries decreased from 39% to 32% between 2000 and 2010. In 2012 the proportion of urban population living in slums was highest in sub-Saharan Africa (61.7%), followed by southern Asia (35%), Southeast Asia (31%), East Asia (28.2%), West Asia (24.6%), Oceania (24.1%), Latin America and the Caribbean (23.5%), and North Africa (13.3%). Compared to 1990, these regions tended to experience losses in proportion of urban dwellers living in slums, e.g., South Asia (−22.2 percentage points), North Africa (−21.1 percentage points), Southeast Asia (−18.5 percentage points), East Asia (−15.5 percentage points), Latin America and the Caribbean.
(-10.2 percentage points), and sub-Saharan Africa (-8.3 percentage points). Oceania saw no change between 1990 and 2012 and only West Asia saw an increase (2.1 percentage points).

**Migration and the reproduction of slums**

Migration from rural to urban environments for economic opportunities, better schools, and a higher quality of life has been a major force in sustaining slums. With populations in cities swelling, local labor markets have often been unable to provide decent-paying jobs and suitable shelter for those that desire them. Formal and informal mechanisms, such as realtor steering, zoning, and discriminatory house buying and renting practices have often systematically concentrated these populations in slums. Moreover, technological change displacing workers from agricultural settings has exacerbated rural to urban migration, visibly expressed as the increasing substitution of capital for labor in the agricultural process across the globe. Thus, many countries have had a declining share of agricultural contribution to GDP from 1961 to 2015, including Turkey (-44.8 percentage points), Bangladesh (-42.1 percentage points), Niger (-37.2 percentage points), Indonesia (-32.0 percentage points), China (-31.2 percentage points), Sudan (-26.8 percentage points), Malaysia (-26.6 percentage points), India (-24.8 percentage points), Ecuador (-23.6 percentage points), Ghana (-19.0 percentage points), Pakistan (-19.9 percentage points), and Brazil (-13.9 percentage points).

**Slum characteristics**

A slum is typically established at the urban edge of a city on undesirable public land. As urban boundaries expand over time, it is common to see slums that were once on the periphery of a city become integrated into its urban morphology. The exact location of a slum on the urban edge depends on a variety of factors, including locations of employment, availability of land, and layout of transportation systems. Slum settlements often begin with impermanent tents or shelters before slow investment associated with longer tenure creates more permanent settlements with more durable materials. The availability of water is of vital importance to sustain life and is strongly considered in settlement location. Accessibility to work outside of a slum causes settlers to locate on the edge of an urban area or near rail infrastructure. Settlers may choose to locate on land viewed as undesirable because it poses little threat of being seized by government. These considerations often mean that slums occupy environmentally dangerous areas such as floodplains and earthquake fault lines, making them susceptible to natural and human-made disasters (Wisner *et al*. 2003).

Migrants who settle on land that will become a slum seldom have the legal authority to occupy the land. The land may belong to another individual or the state, or may be nonparceled property with unidentified ownership. Even in cases where a government does not address the settlement of a land which the inhabitants do not own, the settlers do not obtain the property rights, making them susceptible to displacement (Agbola and Agunbiade, 2009). In the absence of formal land tenure and regulation of land use and building codes, an informal group may sell the informal right to establish residency in the slum and dictate where building construction can occur. Having paid to occupy a slum often gives residents a feeling of true community; however, this landownership is not legally binding or recognized by governments. Having legal land tenure
and recognition by government provides benefits such as an improved ability for residents to obtain financing, the willingness of government to establish or upgrade infrastructure within a slum, and integration into the larger urban fabric with roads and infrastructure. Governments may be reluctant to recognize slums because of the costs associated with providing basic services and assumptions that official recognition and service provision may cause further slum formation.

Housing structures in slums are often poorly built and structurally unsound to endure extreme climates or weather conditions. The materials used to build shelter increase the instability as the building materials used are inexpensive, lightweight and capable of being carried long distances, and are easily maneuverable for construction without the aid of heavy machinery. The structural soundness and quality of materials vary widely from place to place and are dependent on the expertise of builders and the availability of materials in the region. Construction materials include durable materials such as brick, concrete, wood, metal, and less durable materials such as straw, mud, and wood and metal scraps held together by ropes. Overcrowding is an issue in dwellings, as a single structure with one or two rooms may house a family and sometimes more inhabitants. Inadequate space requires inhabitants to utilize one room for sleeping, cooking, and recreation.

Slum dwellers participate in both the formal and informal economy both within and outside the confines of a slum. High unemployment rates persist in slums as a result of limited employment opportunities and a skills mismatch between slum residents and formal employment opportunities. The industries and occupations that comprise the informal economy may be remarkably similar to the formal economy, the only difference being that the informal enterprise is unregulated and untaxed by the government. Workers in this setting may be exploited as a result of the lack of legal oversight into employment conditions. Informal economies in cities of developing countries may be larger than the formal economy and encompass entire industries such as manufacturing. Dharavi in Mumbai, India, has a $1 billion a year informal economy which includes a recycling industry that processes 80% of Mumbai’s waste (Yardley 2011).

While slums are situated in different contexts within cities, UN-Habitat (2008) provides a typology for three distinct types of slums that can be seen in cities across the world. The first is a slum city or cities with high slum prevalence. These slums exist in cities and countries where poverty is endemic, and adequate shelter, infrastructure, and protection from environmental hazards are often nonexistent for much of the population. This type of slum is most prevalent in countries in sub-Saharan Africa. A second type is an isolated slum environment that exists in a city where there is a large nonslum area which is easily distinguishable from concentrated slum areas. Large capital cities in Latin America are clear examples of this form; economically prosperous areas are typically in close proximity, sometimes separated only by a wall. A third type of slum, representing poverty at the geographic margins, occurs in cities and countries with low slum prevalence. In these cities and countries, low-income households are often most visibly marginalized and denied the goods and resources that a city and region contains. These slums are most common sub-Saharan Africa, Latin America, and the Caribbean, and Asia.

Efforts to eliminate slums

Attempts to eliminate slums have focused on removal by demolition, guided relocation of
SLUM

residents, and upgrading with essential infrastructure. Slum removal has been justified by legal means in instances when slum residents do not have the legal right to occupy their land. This strategy is widely viewed as shortsighted as it does not address the causes or the likely perpetuation of slum conditions. This strategy forces slum residents, who need a place to live, to find similar affordable conditions elsewhere. Slum clearance is a common intervention strategy by governments, and is common when slums occupy valuable land near city centers.

Slum relocating typically seeks to guide residents to areas with improved living conditions. Housing in these areas, often rural or semirural settings, may be provided free or at low prices to encourage occupancy. While this strategy reduces slum occupancy, it often provides spatial disadvantages to former slum occupants who are displaced from job opportunities and not provided with transportation to cities centers. Opposition to relocation by slum residents may occur when residential improvements are not coupled with job accessibility and employment opportunities. Slum residents who sign up for housing relocation programs may not understand how suburban living conditions will affect them and, as a result, are worse off after relocation. This approach also ignores the fact that many very poor areas have viable community links that are disrupted by relocation.

Slum upgrading aims to provide infrastructure and physical improvements to address poor living standards, environmental hazards, and the disconnection of slums from surrounding urban areas. In addition to vitally important goods and services such as clean water, electricity, and waste collection, slum upgrading may establish schools, community centers, or paved roads, and may connect mass transit networks to slums, providing access to jobs and amenities outside of the slum. Upgrading may not involve the construction of new dwellings but may provide housing finance options for residents to obtain construction loans. Even in instances where no finance options are present, housing upgrades are believed to be encouraged by residents as a result of the supplementary infrastructure upgrades (Turner and Fichter 1972). Most importantly, residents are most likely to invest in housing when a deed is granted for legal occupancy. The amount a slum dweller is willing to pay toward legal housing typically exceeds what a government is willing to contribute and is seen as a less beneficial investment compared to infrastructure installation.

SEE ALSO: Ghetto; Poverty; Residential segregation; Unemployment and “underclass”; Uneven regional development

References


Further reading


Snow and ice avalanches

Christophe Corona
Centre National de la Recherche Scientifique (CNRS), France

Markus Stoffel
University of Bern and University of Geneva, Switzerland

An avalanche is a rapid flow of snow and/or ice with a volume of more than 100 m$^3$ down a steep slope (30–45°) with a path length of more than 50 m. Snow and ice avalanches are natural processes occurring in many parts of the world, and they are a risk where their occurrence interferes with humans and their infrastructure. The term “avalanche” implies the whole process of movement, from release in the starting zone to displacement (flow) through the path and deposition in the runout zone (Rudolf-Miklau et al. 2015). According to the American Avalanche Association, the worldwide number of avalanches falling annually is on the order of 10$^6$. Throughout history, avalanches have endangered human life in mountainous areas, and thereby exerted major control on the development of settlements and infrastructure. Over the past century, the combination of urban sprawl in mountain ranges across Europe and North America with the growing demand for mobility and recreational activities has increased avalanche risk significantly (Rudolf-Miklau et al. 2015). The growth in winter sports, for example, has led to a significant increase in avalanche deaths over the past few decades, with an average of 250 ski tourers or free-riders being killed worldwide every year (Ancey 2006). In a society with ever-increasing avalanche risk and safety expectations, the risk acceptance of modern societies has been decreasing in equal measure over recent decades.

Substantial efforts have therefore been deployed over the past decades to enhance understanding of avalanche processes. This entry summarizes the key features of snow and ice avalanches, details the main approaches used to describe snow avalanche motion and related hazards, and summarizes recent findings on avalanche triggering, with a special emphasis on meteorological factors in the current context of global warming.

Classification and formation of ice and snow avalanches

Ice avalanches have frequently been responsible for large disasters in ice-covered mountains. According to McClung and Schärer (1993), they result from the calving (breaking off) of ice masses or seracs. Ice avalanches typically occur at positions where glacier ice flows over a steep slope or a cliff and/or as a result of the breakoff of a large piece of glacier (massive events), thus resulting in a tensile failure of the ice which is usually caused by a combination of internal flow (creep) and sliding of the glacier over its bed.

Most catastrophic snow avalanches, by contrast, follow a different basic principle where gravity at the top of the slope exceeds the binding force holding the snow together and thus results in a weakening within the snowpack. A solid slab of the surface layer of snow can then push its way across the underlying layer, and so result...
SNOW AND ICE AVALANCHES

in a snow avalanche (Ancey 2006). The triggering of an avalanche can occur as a result of (i) localized rapid near-surface loading by, for example, people or explosives (the latter being called artificial triggering), (ii) gradual uniform loading due to precipitation or other factors, or (iii) a no-loading situation that changes snowpack properties, for example, surface warming (called natural triggering or spontaneous release) (Schweizer 2003). Schematically, two types of release exist. Loose snow avalanches start from a point in a relatively cohesionless surface layer. They then spread out laterally by pushing and incorporating more snow. Snow of low cohesion may be either dry or wet, but the important point with respect to water content is that wet loose snow avalanches can be much more massive than dry avalanches. Snow slab avalanches, by contrast, involve the release of a cohesive slab over an extended plane of weakness. The failure usually develops first along a sliding surface, and then propagates throughout the upper layers across a crack perpendicular to the downward direction.

Avalanche motion and related hazards

The motion of ice avalanches has been described as analogous to rockfalls: ice chunks roll and bounce along the surface enveloped by a powder cloud. By contrast, and for practical reasons, the movement of snow avalanches has been characterized in more detail, taking into account the velocity, flow height, and mass density of the avalanche snow. In addition, the maximum runout distance, defined as the point of farthest reach of debris, and impact pressure, a relevant criterion of the avalanche movement in hazard mapping or planning of mitigation measures, have been quantified. Data from field experiments from various avalanche test sites as well as numerical modeling have exemplified that the total snow mass of an avalanche, mainly controlled by entrainment, clearly defines runout distance. Similarly, avalanche front velocity, a common measure to estimate avalanche intensity (as it can be linked with expected impact pressures), scales with the square root of the total fall height of the avalanche front (Schweizer 2014).

Schematically, three limiting cases of avalanche dynamics exist and they depend on the form of motion rather than the quality of the material. Within these three limited cases, discussed below, a wide variety of avalanches can be found; they are referred to as mixed-type avalanches.

Laminar type flow avalanches are characterized by a high-density core at the bottom where particle sizes are defined by the state of the snowpack. Fresh dry snow tends to form small granules (2–3 mm in diameter) whereas old snow which has undergone several cycles of metamorphosis consists of ice grains of typically 5–10 mm in diameter. Flow depth does not generally exceed a few meters, and trajectory of the flow is dictated by topography. The typical mean velocity of laminar type flows ranges from 5 to 25 m s⁻¹ but velocity profiles typically differ in each section of the avalanche. The highest velocities of the flow section can be observed at the avalanche front (head), whereas the velocity profile of the normal flow (i.e., the movement recorded perpendicular to the topography) will be disordered with heavily fluctuating velocities. The avalanche body, by contrast, is characterized by a typical shear velocity profile, whereby velocity increases with increasing height above the ground. The avalanche tail, however, is characterized in many cases by weak or missing shear in the flow, such that snow flows in a block on a very thin shear layer or glides directly on the substratum. On average, the density of laminar type flows is fairly high, generally ranging from 150 to 500 kg m⁻³. These avalanches can cause extensive damage because of the large
SNOW AND ICE AVALANCHES

Snow masses involved in the flow in spite of their comparably low speed.

Powder-snow avalanches are large-scale turbidity currents descending slopes at high velocities, in which most snow particles are suspended in the ambient air by turbulence. These clouds can reach 100 m in height and very high front velocities in the order of 100 m s⁻¹. They grow continuously but have fairly low average densities ranging from 4 to 25 kg m⁻³. The front velocities are smaller than core velocities due to air resistance. The mass coming from behind to hit the front is severely decelerated and therefore slower than the avalanche core. Relief has usually only a weak influence on aerial flow. Typically, for snow depth, mean velocity, and mean density, the orders of magnitude usually observed during events are 10–100 m, 50–100 m s⁻¹, and 5–50 kg m⁻³, respectively. These often spectacular avalanches occur only under certain meteorological conditions, namely after abundant fresh snowfalls, as well as in the presence of cold, dry, and weakly cohesive snow on strong slopes. The danger generated by the powder layer of an avalanche is comparable to squalls. They mainly result from powder pressure on the upwind side and suction forces of the aerosol. These devastating pressures (up to 10 bars) are able to uproot forest stands, to destroy buildings, to tear off roofs, to cross valley bottoms, and even to climb up the opposite slope during their propagation stage. The loads caused by powder avalanches are dependent on the density and velocity of the aerosol (Rudolf-Miklau et al. 2015).

Wet snow avalanches have a low flow velocity – with the exception of very steep avalanche paths – and a clear structure of clumps. While agreement exists that the high density of wet snow avalanches can develop very high pressures, appropriate tools to calculate pressures or runout distances are still largely lacking today (Rudolf-Miklau et al. 2015). This is partly due to the fact that wet snow avalanches show special deposit patterns with extremely long and unpredictable runout directions. Slush flows represent a special form of avalanche found especially at high latitudes; they occur as a result of the rapid onset of snowmelt in spring as the return of sunlight adds intense radiation to snowpacks that have been previously subjected to strong temperature gradients. Slush flows consist of a suspension of snow dissolved in water and their rheology has been described as a hybrid of a liquid and granular constitutive law (Rudolf-Miklau et al. 2015). Slush flows can occur on low-gradient slopes and have even been observed at terrain inclinations of 12° or less. Densities of slush-flow deposits can exceed 100 kg m⁻³ due to the combination of water, ice, snow, and entrained rock or earth material. Such densities imply high damage potentials and class them among the most destructive avalanche types (McClung and Schaeerer 1993).

Fixed triggering parameters

The avalanche release process still remains largely unknown, mainly because of complex interactions between terrain, snowpack, and meteorological conditions leading to avalanche release (Schweizer 2003). Yet, one can distinguish between fixed parameters related to the avalanche path and varying parameters which are generally related to weather conditions (Ancey 2006).

Terrain characteristics are the only constant factor in snow avalanche triggering over time, and slopes with an inclination of >30° are generally required for snow avalanches to form. It is also recognized that the starting zone is usually the steepest segment of the avalanche path, and that starting zones are often located above timberline, where avalanches break away, accelerate down the slope, and pick up additional material as they move. Avalanches tend to
SNOW AND ICE AVALANCHES

occur more frequently on surfaces with concave starting zones (both in terms of profile and in the longitudinal direction). From the starting zone, the avalanche moves into the track (20–30°) where velocity tends to remain more or less constant and little or no additional snow is added to the moving avalanche. The runout zone is an area with an average inclination of less than 10°, which is insufficient for further movement of snow. Two main path configurations are particularly prone to produce devastating avalanches: (i) large bowl-shaped starting zones in which the stress increases until these sections can no longer support themselves and break loose; (ii) paths consisting of several gullies that feed into a common track farther down.

Terrain roughness also influences avalanche formation by hindering the formation of continuous, weak snowpack layers (Schweizer 2003). A minimal snow depth, of 0.3–1 m, for example, is needed to smooth out most terrain roughness. Vegetation, by contrast, has mixed effects on avalanche release. Dense forest stands (>200 trees ha⁻¹) in the starting zone increase friction coefficients between the snowpack and the substratum, thus limiting the triggering of snow avalanches. In addition, the interception of snow — compared to nonwooded areas — can be between 30 and 70% greater on such slopes, depending on snowfall, snow volume, tree species, air temperature, and wind conditions. However, the protective effect of trees and forest stands usually decreases downslope and under extreme conditions. Forests cannot stop large avalanches that have released above the forest limit, as their braking effect remains insignificant. Conversely, unmanaged pasture land and the existence of shrubs have negative effects on the stability of snowpack since they form ideal surfaces for snow gliding (Rudolf-Miklau et al. 2015).

Finally, the orientation of slopes with respect to the sun and dominant winds has a strong influence on the day-to-day stability of the snowpack. For instance, in winter, shady slopes receive little incoming radiation and conversely lose heat through longwave radiation. It is generally observed that for these slopes, snowpack will be cold and will tend to develop weak layers (i.e., faceted crystals, depth hoar). In late winter and spring, temperature increase will enhance stability of the snowpack on shady slopes and, at the same time, instability on sunny slopes (Ancey 2006).

Meteorological triggers

The collapse of glacier seracs in icefalls and related ice avalanches are usually caused by glacier motion (or less usually earthquakes). As a consequence, actual occurrences of ice avalanches are not usually related to weather conditions even though meltwater has been demonstrated to often play a role in their release (McClung and Schaerer 1993). Yet, since glacier sliding is known to be faster in spring or summer when more basal water is available, more events have been shown to occur during these seasons.

By contrast, most of the factors contributing to snow avalanching are related either to the strength or to the load of snow and its variation over time. In addition to terrain, Schweizer (2003) describes four essential factors leading to the release of snow avalanches: precipitation (snowfall), wind, temperature, and snowpack stratigraphy.

Among the varying factors intervening in avalanche release, experience clearly shows that in most cases snow avalanches result from changes in weather conditions. Snowfall has been demonstrated to be one of the most important forecasting parameters for large snow avalanche events. In European mountain ranges, heavy snowfalls with a total precipitation exceeding 1 m within three days have been considered
critical for the initiation of extreme avalanches; about 30–50 cm is seen as a threshold for naturally released avalanches in general. However, even with large accumulations of new snow, the combined release probability of a group of avalanche paths is frequently <50%, underlining that the newly deposited snow depth alone is not sufficient to explain avalanche activity. The precipitation rate or loading rate can also strongly influence the critical balance between stress and strength. Snow within the snow cover behaves like an elastic, yet viscoplastic, material that can absorb slow load changes. However, under rapid load (2.5 cm h⁻¹), the weak layer below the storm snow layer might not gain strength sufficiently quickly. In such cases, the response of the snowpack becomes elastic and the snowpack is much more likely to break. In addition, rainfall plays a complex role in snow metamorphism. Generally, for dry snow, a small increase in the liquid water content does not significantly affect mechanical properties of snow. However, heavy rain induces a rapid and noticeable increase in liquid water content, which results in a drop in the shear stress strength. This situation leads to widespread avalanche activity (wet snow avalanches).

In addition, redistribution of snow by wind is a major feature of mountain snowpack and essential for avalanche formation. For loose, unbounded snow, the typical threshold wind speed is 5 m s⁻¹ at which snow is picked up from the surface by turbulent eddies. For a dense bounded snow cover, winds greater than 25 m s⁻¹ are necessary to produce blowing snow (McClung and Schaerer 1993). In mountain environments, snow redistribution by wind is uneven and can significantly influence snowpack stability. Variations in wind speed and subsequent snowdrift form layers of different density or hardness, thereby creating stress concentrations within the layered snowpack, accelerate snow metamorphism, and form cornices on the lee side of ridge crests, whose collapses may trigger snow avalanches as well. When discussing the role of snowdrift in avalanche formation, the influence of wind is very diverse, and can either consolidate snow (compacting and rounding snow crystals) or, by contrast, weaken it.

Apart from precipitation and wind loading, rapid increases in air temperature and/or solar radiation are commonly considered as additional meteorological factors contributing to avalanche formation, particularly in situations without loading. Wet loose snow avalanches are thus usually triggered by intense melting resulting from snowpack warming. The effects of temperature on slab avalanches are more complex, as the cohesion of wet snow causes changes in air temperature to primarily affect the top surface layers (20–30 cm). Since thermal conductivity of snow is low, the energy added to snowpack by sensible heat will only travel slowly from the snow surface to the layers beneath. With increasing temperature in the near-surface layers, stiffness will decrease – in other words, deformation will increase, both in the slope parallel and in the vertical direction (Schweizer 2003). Indeed, recent studies have shown that surface warming can increase the propensity for crack propagation, provided that pre-existing weaknesses exist and that considerable energy inputs are available (Schweizer 2014).

Finally, snowpack stratigraphy is recognized as a key contributing factor to avalanching as the complex structure of an individual snowpack will affect its stability, and therefore its propensity to fail catastrophically in an avalanche. Any snowpack will develop a layered structure during its formation, which will in turn evolve through time. Each layer is formed by the deposition of falling snow, the redistribution of windblown snow, or a combination of these two processes. The initial snowpack will subsequently be modified through compaction and thermal metamorphic processes, thus resulting
SNOW AND ICE AVALANCHES

in the gradual densification of the snowpack. Surface melting, percolation, and, under certain conditions, refreezing of meltwaters further modify snowpack. When comparing stable with unstable stratigraphy, the differences in grain size and hardness between the weak and the adjacent layers of unstable profiles are significantly larger than those of stable profiles. Temperature gradient within the snowpack has been found to be decisive for the metamorphosis processes and avalanche formation as well. If snowpack is subject to strong temperature gradients, the resulting movement of water vapor between grains at different temperatures will drive kinetic growth metamorphism (with grain growth rates in the order of 0.1 mm day$^{-1}$ for a 150 K m$^{-1}$ gradient), thereby producing characteristic faceted grains known as surface hoar and depth hoar. These are among the most critically weak layers. Together with their edgy shapes, formed at the beginning and then developing into jar-like depth hoar, they are at the origin of a majority of slab avalanche incidents.

Snow avalanches and climate change

Statistical relations between snow avalanche activity, snow, and weather drivers are often significant and may help forecasting in operational contexts. At larger spatiotemporal scales, links with atmospheric circulation patterns exist as well. Being phenomena controlled by climate, snow and ice avalanches are not only expected to be affected by changes in atmospheric conditions, but also by climatic change.

Several ground-based studies recently demonstrated a decrease in snow precipitation phase, snow depth, snow cover duration, or snow water equivalent in the Alpine space (e.g., Laternser and Schneebele 2003). On the other hand, however, only very few studies discussed the impacts of ongoing climatic change on actual avalanche activity. Over long timescales, records derived from lichenometry, lake sediments, tree rings, or pollen counts have shown an increase in avalanche frequency in northern and western Europe during the cold phases of the Little Ice Age (Corona et al. 2012). Yet, these studies relied on rather limited datasets of relatively low resolution and were limited to exceptional study sites where depositional sequences have been conserved. Over shorter timescales, based on continuous systematic records, Laternser and Schneebele (2002) did not, by contrast, observe any significant changes in avalanche activity between 1950 and 2000 across Switzerland. On the other hand, thanks to the exploitation of systematic regional avalanche records from the French Alps, Eckert et al. (2009) highlighted significant changes in avalanche activity since the 1950s, with a clear temporal trend toward changes in avalanche runout altitudes which were, in turn, well correlated with climatic variables. Owing to the lack of continuous and multicentennial avalanche chronicles, quantitative prognostics on the future evolution of avalanche activity (e.g., to the end of the twenty-first century) remain extremely scarce and often speculative. The current state of knowledge, still rather qualitative, predicts an ongoing increase in the proportion of wet loose snow avalanches compared to dry snow avalanches, as well as a shift in their respective timing, which is in good correlation with field observations of snow cover wetting at smaller scales.

SEE ALSO: Avalanche meteorology; Climate change, concept of; Climatology; Glaciers; Global environmental change: human dimensions; Patterned ground; Periglacial processes and landforms; Snow; Snow cover; Snow cover changes
References


Further reading

Snow cover changes

Ross D. Brown  
*Environment and Climate Change Canada*

**Background**

Seasonal snow covers a large fraction of the Northern Hemisphere (NH) land area for periods ranging from several weeks at lower latitudes to eight to ten months in the Arctic. Snow has unique physical properties (high reflectivity, low thermal conductivity, and low surface roughness) that together with storage of water over the winter period play important roles in climate and hydrology. For example, snow insulates soils, modifies underlying ice growth rates (on lakes and sea ice), and modifies energy and water budgets. Snow also provides protected winter habitat for plants and animals and supports a wide range of human activities such as overland transportation in the Arctic, skiing, tourism, and water storage for hydroelectricity production and fresh water supply. The storage and release of water from winter snowpack are especially critical in semiarid regions of the world where mountain snowpacks are a major freshwater source. Global warming is a particular threat in these areas due to high and growing demands on the existing water supply and strong temperature sensitivity of the fraction of precipitation falling as snow.

The accumulation of snow on the ground (to form a seasonal snow cover) requires snowfall (precipitation forming at air temperatures below freezing) and a negative energy budget that precludes snow from melting. These conditions occur for various periods of time during winter over most of the NH land area north of about 40°N and in mountainous regions at lower latitudes, for example, Himalayas and Atlas Mountains (Figure 1a). Outside Antarctica, seasonal snow cover in the Southern Hemisphere is mainly confined to mountainous regions of South America and New Zealand with smaller localized areas of snow cover in South Africa and Australia. Seasonal snow cover over NH lands accounts for about 98% of the global terrestrial snow cover extent (Armstrong and Brun 2008) and exhibits a seasonal range of more than 40 million km² from the minimum in August to the maximum in January (Table 1). Interannual variability in snow cover extent (SCE) is largest in the snow cover onset (October–November) and melt (May–June) periods and over mid-latitudes and coastal mountain regions where winter air temperatures are relatively close to freezing (Figure 1b). Snow cover in these regions is particularly sensitive to warming air temperatures.

The duration of snow on the ground (SCD) is closely linked to air temperature which explains the strongly zonal character of SCD in Figure 1a and the longer SCD at higher elevations. How much snow accumulates on the ground is initially governed by geography. Latitude, elevation, proximity to storm tracks and moisture sources are key factors controlling snowfall amount and air temperature that together define the length of the snow cover season and the maximum amount of snow that accumulates. Once on the ground, local factors such as land cover, vegetation type, exposure, surface roughness, topography, and interactions with surface wind determine how much snowfall is lost to sublimation (from the...
Snow Cover Changes

Figure 1  (a) Mean annual duration of snow on the ground (days) from the IMS 24 km daily snow cover analysis (Helfrich et al. 2007) over snow seasons 1998/1999 to 2012/2013. (b) Interannual variability in snow cover duration (days) over the 15 snow seasons.

Table 1  Mean and standard deviation in NH snow cover extent (million km$^2$) observed by visible satellite data over the period 1972–2012 from the NOAA Climate Data Record (CDR) dataset maintained at Rutgers University (http://climate.rutgers.edu/snowcover/).

<table>
<thead>
<tr>
<th></th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>46.2</td>
<td>45.2</td>
<td>39.8</td>
<td>30.0</td>
<td>19.1</td>
<td>9.6</td>
<td>3.9</td>
<td>2.9</td>
<td>5.2</td>
<td>17.5</td>
<td>33.2</td>
<td>42.8</td>
</tr>
<tr>
<td>SD</td>
<td>1.6</td>
<td>1.9</td>
<td>1.7</td>
<td>1.6</td>
<td>1.8</td>
<td>2.2</td>
<td>1.0</td>
<td>0.6</td>
<td>0.9</td>
<td>2.5</td>
<td>2.1</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Snow surface, from blowing snow, and from the vegetation canopy) and how much accumulates at a particular location. Local site characteristics also play a major role in snow melt processes (e.g., topography and vegetation cover influence local advection, shading, and longwave radiation fluxes). The reader is referred to Armstrong and Brun (2008) for a detailed discussion of snow processes and their physics.

The effect of geography on mean annual maximum snow accumulation over the NH is demonstrated in Figure 2a. The largest accumulations are found in maritime mountain regions with large snowfall amounts related to
proximity to storm tracks and moisture sources and orographic effects on precipitation and temperature. These same regions are also characterized by large interannual variability in snow accumulation linked to atmospheric variability in air temperature and precipitation (Figure 2b). The North Atlantic Oscillation is an important driver of annual variations of snow cover in Europe while snow cover variability over western North America is linked to various modes of Pacific-driven climate variability including El Niño–La Niña, the Pacific Decadal Oscillation and the Pacific-North America pattern. Accumulations are much lower over the relatively colder and drier continental interiors and Arctic regions, and also around the southern margins of the continental snow line where snow cover is ephemeral in nature. Moisture penetration into the Arctic via Baffin Bay and the Norwegian and Barents Sea results in higher snow accumulations in these regions of the Arctic.

Observed snow cover changes

While the average snow climate of a region is initially governed by the abovementioned geographical factors, the seasonal evolution of a snow cover and its physical properties (e.g., depth, density, snow water equivalent (SWE),
SNOW COVER CHANGES

albedo, stratigraphy, crystal structure, thermal conductivity, and liquid water content) are the result of a complex set of processes and landscape interactions driven by the local climate (incoming solar and longwave energy, wind, air temperature, humidity, cloud cover, and precipitation). Changes in any of these climate drivers, or changes in the landscape through enhanced shrub growth for example, will influence the snow cover. Atmospheric aerosols can also influence snow cover: the deposition of black carbon particles (“soot”) on the snow surface reduces the reflectivity of the snow and contributes to enhanced spring snowmelt. There are thus many potential drivers of snow cover change.

A number of surface and satellite sources of snow information are available for monitoring snow cover; these cover different periods of time and different spatial scales and have particular characteristics that need to be taken into account when interpreting observed trends (see Armstrong and Brun 2008, ch. 5, for a review of snow observing systems). Surface observations of daily snow depth are available over much of the NH from approximately 1950 but the observations are biased to lower elevations and lower latitudes with large data gaps in low populated regions such as northern Canada. Point measurements of snow depths are usually made at open sites which may not be representative of the prevailing snow cover (note that Russia takes depth measurements in open and vegetated areas). For example, measurement made in open locations in the Arctic will not register the effect of any Arctic green-up and increased shrubiness which increases the snow trapping potential of the landscape and modifies the spring melt energetics.

Major advances have been made in the ability to monitor snow cover from space over the period since approximately 2000 (see recent review by Frei et al. 2012) but this period is too short for establishing trends. The satellite-derived NOAA snow cover climate data record (CDR; Estilow and Robinson 2013) provides weekly snow cover extent information over the NH at approximately 200 km resolution since November 1966. This dataset is suitable for continental-scale monitoring of snow cover extent during the winter season but there are homogeneity issues related to changes in mapping methods and satellite resolution and coverage over time. Efforts have been made to mitigate these effects but there are known inconsistencies in mapping snow over mountain regions and in the snow cover onset period. In the past, attempts to monitor and map global SWE from space using passive microwave had limited success as the retrievals did not work well in many landscape areas (dense forest, complex terrain, and deeper snowpacks). The recent GlobSnow passive microwave SWE dataset (Takala et al. 2011) overcomes some of these limitations by incorporating surface observations in the satellite retrieval process. GlobSnow SWE data are available daily over the nonmountainous NH at a 25 km resolution from 1979.

Information on historical variability in snow cover can also be obtained from reanalysis datasets that are essentially reconstructions of past weather using an atmospheric circulation model and an ensemble of historical meteorological observations (e.g., surface observations, upper air soundings, satellite data) in a data assimilation framework. Snow cover information can be generated from data assimilation of available surface snow depth observations (e.g., ERA-interim reanalysis) or reconstructed using a land surface scheme coupled to the atmospheric model (e.g., MERRA reanalysis). Another approach is to use the output from a reanalysis to drive a detailed snowpack model. This approach was used by Liston and Hiemstra (2011) to reconstruct snow cover at a 10-km resolution over the Arctic for the period from 1979 to 2009. Great care
must be taken when using snow output from reanalyses as there can be discontinuities created from changing data streams (e.g., new sources of satellite data) as well as important biases in temperature and precipitation from the atmospheric model.

The spatial and temporal variability in NH monthly SCE and SWE over the period from 1972 and 1980 respectively are shown in Figures 3 and 4 from the NOAA CDR and GlobSnow datasets. The period prior to 1982 had above average snow cover over the Arctic.

**Figure 3** Evolution of NH monthly snow cover extent over the period 1972–2015 from the NOAA CDR dataset. Values are expressed as departures (%) from the 1971–2000 mean.

**Figure 4** Evolution of NH October to May monthly SWE over 1980–2014 from the GlobSnow V2 dataset. Values are expressed in units of standard deviation with respect to a 1981–2010 reference period; for example, a value of 2 for a month indicates the SWE was two standard deviations greater than the monthly average for 1981–2010. Note the different start periods in Figure 3 and Figure 4.

**Figure 5** (a) Projected change (%) in annual SCD from 8 CMIP5 models for the RCP8.5 emission scenario for the 2080 period (2070–2099 minus 1970–1999). (b) Same as top left panel for annual maximum monthly SWE from 11 CMIP5 models. (c) Mean expected year of significant change in annual SCD for the RCP8.5 emission scenario. (d) Same as bottom left panel for SWEmax.
Figure 5
while the period since approximately 2000 has been characterized by progressively earlier spring melt over northern latitudes. SWE was above average over most of the NH in the decade from 1988 to 1998 with below average values dominating since 2005. The recent marked reduction in high-latitude spring snow cover is consistent with enhanced snow albedo feedbacks at this time of year. Brown and Robinson (2011) were able to extend the NOAA NH spring SCE record back to the early 1920s with surface snow observations which showed that most of the twentieth century change in NH spring snow cover has occurred in the period since 1980. A synthesis of trends in snow cover from surface-based observations in the 5th IPCC Assessment (Vaughan et al. 2013) concluded that the majority of sites have experienced declining snow cover with the largest reductions at locations where winter temperatures were closer to freezing. In these regions, particularly those with high precipitation amounts like the coastal mountain ranges of North America and Norway, warmer temperature result in the combined impacts of a shorter snow season and less snowfall. Some regions of northern Russia show increases in maximum snow depth in spite of trends toward a shorter snow cover season (Callaghan et al. 2012). This apparent contradiction is consistent with climate model projections and is a response to increasing snowfall. Brown and Mote (2009) showed that the response of snow cover to concurrent trends of warming and increasing precipitation depends on the snow cover variable, climate region, and elevation. Start/end dates and snow cover duration are closely linked to air temperature and exhibit rapid reductions in response to warming. However, the response of annual maximum accumulation depends on the interplay between a decreasing length of the accumulation season and increasing snowfall. Winter warming also has implications for snowpack properties; winter thaw events and rain-on-snow events have the potential to generate a denser snowpack with ice layers that can hinder ungulate grazing. An increased frequency of ice layers within the snowpack has been observed in northern Sweden (Callaghan et al. 2012).

Projected changes in snow cover

Climate model projections of future changes in snow cover (Räisänen 2007; Brown and Mote 2009; Brutel-Vuilmet, Ménégoz, and Krinner 2013) provide a picture of widespread significant reductions in SCD and maximum SWE (SWE-max) over most of the NH mid-latitudes (Figure 5, top panels), with the largest changes around the southern margins of the continental snow line and the previously mentioned temperate mountain zones of North America and Europe. The models also project significant increases in maximum SWE over high latitudes although there is much greater model scatter in the models for SWE-max than SCD, particularly over Eurasia. Some idea of the spatial pattern in the timing of the projected changes can be obtained from the expected year of emergence (EYE) of the climate change signal following De Elía, Biner, and Frigon (2013) (Figure 5, bottom panels). On average, EYE is approximately 30 years earlier for annual SCD than annual SWE-max with the climate change signal emerging earliest over the western margins of each continent and latest over Siberia and the Tibetan Plateau. Between-model consistency in EYE is much lower over the last mentioned regions particularly for SWE-max. Large scatter in the strength of snow albedo feedbacks in climate models (Qu and Hall 2013) is likely contributing to this uncertainty, and recent evaluations (Derksen and Brown 2012; Brutel-Vuilmet,
SNOW COVER CHANGES

Ménégoz, and Krinner 2013) show that climate models underestimate the response of NH spring SCE to observed warming. There is evidence of improvements in the representation of snow cover in climate models but important sources of uncertainty remain related to snow albedo feedbacks, precipitation bias (models overestimate precipitation over high latitudes), and inadequate treatment of landscape-dependent snow processes such as blowing snow, snow–canopy interactions, and sub-grid scale advection of energy during snow melt.

Acknowledgments

The author gratefully acknowledges helpful comments and feedback provided by a valued colleague, Dr Chris Derksen, from the Climate Research Division of Environment and Climate Change Canada.

SEE ALSO: Snow; Snow cover

References

Snow cover

David A. Robinson
Rutgers University, USA

Snow cover plays an important role in many of the processes within the Earth’s climate system (Barry 2002). The presence of a snow cover has been shown to affect boundary layer climate variables such as atmospheric temperature, moisture, sensible and latent heat fluxes, and radiative fluxes (Flanner et al. 2011). The amount of snow cover and its subsequent ablation is also important to the hydrology of a given region, including streamflow (Yang et al. 2003). The influence of snow cover on a continental scale is clearly linked to other geophysical and biophysical systems (Callaghan et al. 2011). In turn, the presence or state of snow is influenced by weather, climate, topography, proximity to water bodies, and humankind. Likewise, snow cover, or the lack thereof, is of considerable importance for decision-makers in transportation, recreation, utility, and other sectors of local to global economies.

Snow cover characteristics

Snow covers approximately 30% of the Earth’s land surface on a seasonal basis, with additional coverage over alpine glaciers, polar ice sheets, and sea ice. High reflectivity associated with fresh snow affects the energy balance of a region by increasing surface albedo and thermal emissivity, decreasing shortwave radiation absorbed and consequently increasing the amount of infrared radiation lost to the atmosphere, and reducing surface air temperature. Snow also acts as a thermal insulator due to its low thermal conductivity, resulting in decreased sensible heat flux from the ground into the lower atmosphere. Energy involved in the warming and melting of a snowpack serves as a sink for latent heat, thus making snow cover a major factor in modulating climate variability and change. Snow cover also plays a key role in the hydrologic cycle by acting as the frozen storage term in the water balance.

Snow ablation is a major contributor to streamflow, soil moisture, and groundwater supplies. Approximately one-sixth of the world’s population lives within snowmelt-dominated regions that encompass much of the industrialized world and account for approximately one-quarter of the global gross domestic product (Barnett, Adam, and Lettenmaier 2005). At locations where snow cover is ephemeral, the lack of a single seasonal melt, or lack of consistency in volume or timing, has made forecasting the release of water from the snowpack a formidable challenge, leading to harmful environmental and societal consequences. These consequences may be in the form of snowmelt-induced floods, lack of streamflow in snowmelt-fed rivers, or transport of pollutants or excess nutrients in rapid snowmelt events. These events are also a significant risk to human life. For instance, excluding flash flooding, snowmelt floods result in roughly one-third of all flood-related fatalities in the United States, despite the fact that large populations in the southern United States are not at risk from these floods.

A number of studies have shown that the distribution of snow cover has an influence on low-frequency atmospheric circulation patterns
SNOW COVER

(Rydzik and Desai 2014). Variations in snow cover may be associated with sea surface temperature patterns or involve the vertical propagation of stationary waves within the troposphere and lower stratosphere. This can result in zonal wind and subsequent latitudinal temperature shifts, changing sea level pressure patterns, and shifts in storm track activity. These influences apply to both above and below average snow cover extents, and of course extent changes are associated with these relationships that feed off one another. Determination of cause and effect is often elusive, but without question snow cover is not a passive participant.

The presence or absence of snow cover also plays a role in short-term weather anomalies, most notably departures of surface air temperatures. An anomalous covering of the ground with snow may result in temperatures far lower than might have been achieved (or forecast) in the absence of snow (or knowledge of the presence of snow in a forecast model). Likewise, anomalous warmth may be amplified at a given time and location where snow cover is expected climatologically but is absent. This may result in a positive feedback and greater warmth, much as an extensive cover may enhance cooling.

Monitoring snow cover

Recent years have seen the availability of more accurate and complete information on the spatial extent and physical state of snow. Mature and well-documented climate data records (CDR) of snow on regular grids are available for empirical and modeling studies. This is leading to a better understanding of the variability of snowfall and snow cover on annual to decadal scales, of cryosphere–climate interactions, and of the role snow may play in regional and global climate change. However, observing snowfall or snow cover continues to present a variety of challenges. Given the remoteness of much of the cryosphere, the small, scattered human populations in regions with extensive snow cover, and the logistical constraints associated with in situ sampling, satellite data provide the most comprehensive source of information to generate snow and sea ice data records.

Ground–based data on snow cover are not without worth. While relatively sparse outside of the lower elevations of the middle latitudes, they include daily station observations of snow depth made at a single point. Less frequent and far fewer measurements are taken along snow courses that may extend for multiple kilometers. Data on the amount of water in a snow pack (snow water equivalent (SWE)) are gathered at very few point sites, along snow courses, and by snow pillows that measure snow mass in some mountainous regions. In situ observations serve a useful role as ground truth for satellite-derived estimates of snow depth and SWE. They are critically important in alpine regions where the snowpack serves as a reservoir to fulfill water needs elsewhere and remote sensing cannot provide adequate SWE information. Of course, prior to the advent of the satellite era, in situ observations were the only means of identifying snow cover variability and change.

Visible and passive microwave sensors on geostationary and polar orbiting satellites record information used to monitor snow cover on regional to hemispheric scales. Snow cover extent (SCE) is best identified on visible imagery by recognizing characteristic textured surface features and brightness (Figure 1). Shortcomings include the inability to detect cover when solar illumination is low or when skies are cloudy. In the figure, characteristic cloud textures and a priori knowledge of what the landscape looks like under snow cover (e.g., the most often unfrozen New York Finger Lakes are not seen) assist in determining that the white shades in
the image toward the west (middle left) are thick clouds, while thin cirrus partially obscure snow-free ground at the southern end of the Delmarva Peninsula (lower center). In addition to clouds obscuring the surface, trees protruding through the snow pack may mask the presence of snow cover from the satellite vantage point. Once analysts determine that skies are clear,
they rely on characteristic patterns or the spatial texture of the snow-covered landscape to reach a determination of snow presence or absence. For instance, deciding whether the densely wooded Adirondacks of northern New York and the Pinelands of southern New Jersey are snow covered on February 12, 2010, relies in part on the brightly snow-covered lands surrounding them. The broad valley running generally from southwest to northeast in the lower left quarter of the image is largely forest free, thus bright snow cover is seen. To the northwest of the valley the wooded ridge tops and open valleys of the geologically folded Valley and Ridge province are quite apparent when snow covered. Utilizing a priori knowledge of snow-covered lands is important where the darker areas in the Hudson and Connecticut River valleys and in the Boston area suggest that snow cover is thin and patchy, as they are much brighter when deeply covered. However, at best, visible imagery only provides a qualitative estimate of pack depth.

The recognition of snow using microwave techniques results from differences in the emissivity of snow-covered and snow-free surfaces across several different frequency ranges. Information on water equivalent can be obtained, although not of the accuracy necessary for climatologic studies. Clouds and low solar illumination are not problems when using microwave data to chart snow cover; however, there are difficulties in identifying shallow or wet snow.

Weekly snow files of SCE produced from visual analyses of visible satellite imagery are available from late 1966 through May 1999. The 1972 to present data were mapped by National Oceanic and Atmospheric Administration (NOAA) meteorologists, while the 1966–1971 data were reanalyzed at the Rutgers Global Snow Lab. The maps are digitized in binary format to a 128×128 pixel grid over the entire Northern Hemisphere. In June 1999, the weekly map series was replaced with the daily NOAA/National Ice Center Interactive Multisensor Snow and Ice Mapping System (IMS) product. These maps are digitized to a 1024×1024 grid (even higher resolution products have been generated as the IMS system has recently been upgraded to a third version). A routine was developed to reduce these daily maps to a weekly product at the former lower resolution to facilitate multidecadal assessments. This data product has undergone several stages of reanalysis and recently efforts culminated in raising this environmental satellite data record to CDR status. This involved several steps to reduce inconsistencies over time and improve continuity between the historical weekly record and the more recent IMS-derived snow maps. Thus this CDR now constitutes the lengthiest satellite-derived dataset of any environmental variable (Estilow, Young, and Robinson 2015).

Another snow mapping effort employs MODIS (moderate-resolution imaging spectroradiometer) satellite visible data to recognize snow-covered continental surfaces. The snow product is generated in an automated manner, employing an algorithm to filter snow from clouds. Daily, eight-day, and monthly composite snow maps displaying fractional snow cover are available at 500 m and 0.05° resolution. For modelers, and others interested in a coarser-resolution product, a 0.25° resolution product is also available. A cloud-gap filled (CGF) snow cover fraction product provides daily cloud-free views of global snow cover (Hall et al. 2010). The CGF maps allow the user to select how many days they wish to look back in time to see a snow decision for each grid cell.

Satellite monitoring of snow cover using the microwave portion of the electromagnetic spectrum dates back to the late 1970s. While active radar or scatterometer approaches have
SNOW COVER

been developed, they do not have the tempo-
ral or spatial coverage of passive methods that
measure emitted radiation with a radiometer.
Microwave emission from a layer of snow over
a ground medium consists of emission by the
snow volume and emission by the underlying
ground. Both contributions are governed by
the transmission and reflection properties of the
air-snow and snow-ground interfaces and by the
absorption/emission and scattering properties of
the snow layer, as well as by a myriad of physical
parameters that affect the emission. As an elec-
tromagnetic wave emitted from the underlying
surface propagates through a snowpack, it is
scattered in all directions by the randomly spaced
snow particles. As the snowpack grows deeper,
there is more loss of radiation due to scattering,
and the emission of the snowpack is reduced.
The deeper the snow the more ice crystals
or grains are available to scatter the upwelling
microwave energy; because of this relationship
between microwave energy and number of snow
grains, it is possible to estimate the depth and
SWE from microwave emission.

Microwave measurements can be made through
darkness and can often be made through non-
convective clouds. While there is proven utility
of microwave data for snow cover studies, there
are significant limitations as well. If the snow
cover is thin it may not be detected as snow.
Wet snow is indistinguishable from snow-free
soil. Vegetation, especially forest cover, causes
further complications. SCE can be processed
using a land-cover decision tree that uses mul-
tiple frequencies or through empirically derived
multispectral algorithms. Algorithms also permit
estimates of SWE or depth (using empirical
algorithms based on estimates of snow density)
and snowmelt occurrence.

Snow frequently covers sea ice in the Arctic
and Southern Oceans. However, particularly in
the Arctic, snow melts atop sea ice, though in
some areas it remains well into the summer. The
onset of snowmelt over sea ice can be deter-
mined using multichannel satellite microwave
data. Algorithms can detect when liquid water
begins to be found within the snowpack, but
once the pack is wet, the signal emerging from
an empirically derived algorithm is no different
from that for snow-free ice, thus the actual
disappearance of snow over the ice cannot
be determined (Drobot and Anderson 2001).
To date, there has been no sustained effort to
monitor snow atop sea ice using visible imagery.
The condition of snow cover over the Green-
land ice sheet is monitored using microwave
satellite data using several methods. This includes
the approach of Mote (2007) that uses a radiative
transfer model to estimate the microwave bright-
ness temperatures (TBs) associated with a snow-
pack containing 1% liquid water by volume. The
modeled TBs are used as threshold values and
compared to observed radiances. Much as over
sea ice, only the presence of liquid water within
the snowpack can be determined. The limited
areas of the Greenland sheet that lose snow cover
during the summer are not mapped on a regular
basis. Liquid water within the snowpack over the
Antarctic ice sheet, let alone the complete loss
of snow cover, is exceedingly limited in area.

Continental snow extent

The considerable interannual variability of snow
makes long-term datasets a necessity when
investigating means, trends, low frequency
events, or interactions of snow with other
climatic elements. The standard 30-year climatic
normal used for variables such as temperature
and precipitation may not be of a sufficient
length for realistic cryospheric means. Unfortu-
nately, consistent decades-long data are scarce.
Satellite-derived snow maps for Northern Hemi-
sphere lands have been available since the late
SNOW COVER

1960s. Automated maps of SCE over Southern Hemisphere lands have recently begun to be generated using visible or microwave satellite data. With time, they will provide useful information for the development of climatologies. Station data exist for approximately the past century, though they are of limited scope early on and, as mentioned previously, even today there is limited coverage. Issues with the collection, quality control, archiving, and synthesis of snow data, be it in situ or satellite derived, must be considered before employing the data in long-term climatologic studies.

Utilizing observations from the NOAA snow product, the geographical extent of snow cover in the Northern Hemisphere is found to vary from a maximum of $46.8 \times 10^6$ km$^2$ in January and February, to a minimum of $3.6 \times 10^6$ km$^2$ in August. Between 60 and 65% of winter (December–February) snow cover is located in Eurasia (Figure 2). The vast majority of Russia and Canada are snow covered throughout at least 80% of the winter. A rather pronounced winter decline in SCE frequency occurs to the south of the Canadian-US border and across the middle latitudes of Eurasia. Areas further to the south, such as the Himalayas, Turkish mountains, European Alps and US Rockies, are also more frequently snow covered than nearby lands. The southern advance of snow cover in fall begins in northern Siberia and the Canadian Archipelago. Far less than half of the season is snow covered elsewhere. Spring melt progresses poleward, with extents averaging about 20% greater than fall in most regions. Summer cover is confined to lower elevations above the Arctic Circle and mountains further south in June and only on the Greenland ice sheet and mountain glaciers in July–August.

In the Southern Hemisphere, most of the seasonal snow cover is located in the Andes of South America and on occasion the higher-elevation grasslands toward the south of this continent. Elsewhere, snow cover is found in the New Zealand Alps and less commonly over higher elevations in Australia and even South Africa.

Over terrestrial surfaces, total spring and summer snow cover has been considerably less extensive since the late 1980s in comparison with the satellite era and likely with earlier twentieth-century conditions (Brown and Robinson 2011), but with regional variations, such as increased fall and winter snow cover in Siberia. Along with the recent decline of arctic sea ice, this loss of terrestrial snow cover is one of the better-known and more striking examples of global climate change.

Snow cover was more extensive in the first half of the satellite record than in the past decade (Figure 3). Between 1972 and 1985, annual means of snow extent fluctuated around a mean of 25.9 million km$^2$. An abrupt transition occurred in 1986 and 1987, and since then the mean annual extent has been 24.2 million km$^2$. Monthly anomalies from the long-term mean are most often less than 3 million km$^2$; however, on occasion they range to 4 and 5 million km$^2$, with October 1976 having a positive anomaly of over 8 million km$^2$.

Recent decreases in snow extent are large during the spring and summer, while winter and fall extents show no statistically significant change (Figure 4). The tendency toward less late-season cover in recent years begins in February. During seven of the first 15 years of record, February snow extent exceeded the January value. This has occurred only once in the past 12 years.

Station records for portions of the Northern Hemisphere continents suggest that spring and early summer snow extents in the past decade may be at their lowest values of the century. Otherwise, winter season extents in the latter portion of this century exceed those of earlier years. Annual and decadal fluctuations
are embedded in this upward trend, and, at least over portions of central North America, the variability of snow cover duration has also increased throughout the century. The increase in snow cover duration in this region is accompanied by statistically significant increases in seasonal snowfall. In recent decades, snowfall has also been heavier to the lee of the North American Great Lakes than earlier in the century. These findings are in line with observations from Canada and the former Soviet Union, with all areas appearing to be part of a trend towards increased winter precipitation over the middle latitudes lands in the Northern Hemisphere.

Figure 2  Seasonal frequency of snow cover (mean percent of days) over Northern Hemisphere lands, including Greenland. Values are based on an evaluation of gridded NOAA snow charts from 1981 to 2010. Source: Rutgers Global Snow Lab (GSL), http://snowcover.org.
Figure 3 Twelve-month running mean departures (from 1981 to 2010 mean) of snow cover extent over Northern Hemisphere lands between 1967 and 2014. SCE is gleaned from NOAA snow maps and converted to a climate data record at the Rutgers Global Snow Lab. Source: Rutgers Global Snow Lab (GSL).

Modeling snow cover extent

General circulation models (GCM) simulate the seasonal timing and the relative spatial patterns of continental-scale snow cover extent fairly well. However, there is a tendency to overestimate the rate of ablation during spring, and significant between-model variability is found at various spatial scales (Frei et al. 2005). GCM predictions of future changes in snow and sea ice cover under different global warming scenarios show a continued overall decrease in snow and sea ice extent (Brutel-Vuilmet, Menegoz, and Krinner 2013). This is hardly surprising, as ultimately with temperatures more often climbing above the freezing point in many areas that presently experience snow, snowfall and SCE will be greatly diminished. However, before this occurs, and in some cold regions well after the planet has warmed considerably, winter snowfall and snow depth may increase. All of these changes are likely to have major impacts on humans and the environment, including amplifying global warming.

Conclusion

Snow plays a critical role in the climate system and will continue to be an important climate variable to monitor when assessing climate change. However, there remain many
unanswered questions regarding the variability of snow and the long-term representativeness of recent empirical studies of cryosphere–climate interactions, such as links between snow extent and monsoon strength. Additional information is needed over space and time, including better information regarding snow cover atop sea ice and ice sheets in both hemispheres. Historic station datasets of snowfall and snow cover remain to be collected and analyzed. Recent efforts to integrate visible and microwave satellite and station data must continue for past intervals and future years. Studies need to examine variations in snow water equivalent, and work begun recently to improve model simulations of snow must continue.

With increasing levels of greenhouse gases expected to result in warmer conditions over the coming decades, the coverage of snow across continental landmasses will likely decrease, which in turn should enhance the warming. Snow cover should become established later in the fall, and melt sooner in the spring. Areas where snow cover is ephemeral throughout the winter will see the ground covered less often. Likewise, it is possible that warming will be accompanied by an increased flux of moisture into the high latitudes, resulting in a deeper
SNOW COVER

winter snow pack in some regions, though probably in a shortened snow season.

**SEE ALSO:** Climatology; Cryosphere: remote sensing; Cryosphere studies: history; Polar climates; Snow; Snow cover changes

References


Further reading


Snow

Martin Schneebeli
WSL Institute for Snow and Avalanche Research
SLF, Switzerland

Snow is a unique matter in nature. It symbolizes coldness for humans, but is the only widespread high-temperature material on Earth. It modifies the landscape dramatically, has a huge effect on the biosphere and strong interactions with traditional and modern human culture, especially as a water resource. Snow crystals also became a symbol of natural uniqueness and beauty in the past century. At the same time, snow avalanches are an important natural hazard in mountainous areas. This makes “snow” a unique, complex phenomenon.

Snow is formed in the atmosphere by direct crystallization from supercooled water vapor and sub-micrometer sized water droplets. These minuscule droplets usually contain nuclei; foreign material, for example, as dust from soils, plants, and fires. Snow crystals can rarely homogenously nucleate, by spontaneous freezing at a temperature below about $-40^\circ$C, but typically by heterogeneous nucleation. Heterogeneous nucleation occurs at higher temperatures, and is mediated by nuclei. The immense variety of snow crystals stems from the fact that the growth rate of the basal and prism plane of the hexagonal ice crystal depends on temperature and excess vapor density (supersaturation). Several transitions between preferential growth of the prism plane ($a$-axis of the crystal) and basal plane ($c$-axis) occur over the temperature range 0 to $-40^\circ$C and a supersaturation between 0–0.3 g m$^{-3}$. As the individual snow crystal travels through the clouds, the crystal encounters continuously different conditions, creating the many different combinations of plates, dendrites, needles, and columns, until it is too heavy and falls to the ground. Typically, the formation of snow crystals in the atmosphere takes around half an hour. The size of the falling crystals is between about 10 μm and 10 mm. The very tiny crystals are called “diamond dust,” and are most often observed in inland Antarctica, the largest crystals appear in areas with optimal conditions for snow crystal formation, as on the south coast of the Great Lakes or in Hokkaido.

As soon as snow is on the ground, and often even before, diminution and sintering begin. Diminution of the original crystals is caused by wind action in the air or close to the ground. Under windy conditions, only a very small percentage of undamaged crystals survive until they settle on the ground. Fallen new snow is often redistributed again, especially in windy areas (Arctic, Siberia, Antarctica, alpine ridges), which causes further diminution to small particles. As soon as the snow is settled, the single ice particles, comparable to sand, start sintering. This process starts within seconds, even at low temperatures. Sintering causes the snowpack to stop being granular; instead a solid frozen bond connects all crystals (grains). Snow on the ground must be imaged as a fully connected three-dimensional structure of mostly monocrystalline ice particles.

Metamorphism is the process of recrystallization. Snow being a high-temperature material, very close to the melting point, recrystallizes easily. A high-temperature material is defined as the absolute temperature higher than 60% of

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0182
SNOW

the melting temperature, which is for snow at about $-110^\circ$ C. Metamorphism and compaction (settling) are the key processes acting on snow on the ground. Snow on the ground is a monomineralic rock. Snow as a rock is unusual with regard to that it exists at a high homologous temperature but at very low external pressure. The combination of high vapor pressure of water and ice and high homologous temperature causes a very rapid recrystallization. A typical arctic snowpack may completely recrystallize within a few days (Pinzer, Schneebeli, and Kaempfer 2012). Earlier held views suggested a relatively slow change of the structural evolution of the snowpack. The rapid structural changes directly affect the optical, mechanical, and other physical properties. Depositional and metamorphic processes lead to a complex stratigraphy. The complex stratigraphy is a key component in avalanche formation, especially the formation of the weak layer. The seasonal and annual patterns in the stratigraphy are important for ice core interpretation, as they can be recognized after hundreds and thousands of years. The structure of snow is described by the shape and size of the snow grains. The most basic snow types are new snow, rounded snow, faceted snow, and depth hoar (Fierz et al. 2009). All these snow types show transitions, and an objective determination is often difficult. Substantial effort is going into a more quantitative description of the snow structure as physical models are in need of a more precise quantification. The density of snow varies between 50–550 kg m$^{-3}$. The upper value is the threshold density to firn. A typical alpine snowpack has a density of about 250 kg m$^{-3}$, a tundra snowpack of about 350 kg m$^{-3}$, as wind packs snow particles denser. The specific surface area of snow is about 40 m$^2$ m$^{-3}$, and as low as 1 m$^2$ m$^{-3}$ for refrozen coarse-grained wet snow. Mechanical strength of snow is mainly determined by density, specific surface area, and time of sintering. Due to its high homologous temperature, mechanical properties can change within hours. This is important for the formation of avalanches, trafficability, and snow sports.

Snow on the ground has one of the highest albedo of all natural surfaces, up to 90% of solar radiation is reflected (other examples are dried salt lakes). Typically, snow increases the albedo by 30–50%. This property changes the radiation balance dramatically, and is the main reason that snow cover is relevant to climate. The high albedo of snow is caused by two physical properties. First, ice is one of the least light-adsorbing materials in the visible spectral region. Second, the small size of snow particles, which is commonly measured as specific surface area or optical diameter, causes multiple scattering. However, the influence of snow albedo is somewhat limited due to the polar night in the Arctic and, to lesser degree, in Antarctica. Snow albedo can be substantially reduced by impurities, most effectively by black carbon. Dust can have similar effects, although the radiative forcing is much smaller by mass, it is known to strongly influence snowmelt in mountains. Recent research shows that snow impurities affect snowmelt in many alpine mountains and in arctic regions.

The thermal conductivity of snow varies between almost air (0.026 W m$^{-1}$ K$^{-1}$) and 0.6 W m$^{-1}$ K$^{-1}$. A typical snowpack has a thermal conductivity of 0.2–0.3 W m$^{-1}$ K$^{-1}$. Snow is therefore in most cases not a very good thermal insulation, but thermally similar to dry soils. However, snow is a still a very important insulation for flora and fauna, as the ground heat flux often maintains a 0$^\circ$-boundary at the soil–snow interface. Snow is also very important to reduce temperature fluctuations, which is a very important effect. The large latent heat stored in snow causes a substantial delay in the warming of the soil during melt.
The snow cover in the Northern Hemisphere extends over about $45 \times 10^6$ km$^2$ during winter maximum, and is less than $1 \times 10^6$ km$^2$ during summer. Snow cover extent is measured by remote sensing. As there is no visible light during polar night, the emissivity in the microwave spectrum (brightness temperature) is used during winter. These microwave observations form one of the longest data series in remote sensing. On a global scale, the year-to-year variation of the maximal extent is about 10%, and there is a small trend of about 1% reduction per decade since the 1970s. In the western United States the duration of the snow cover decreases by 2–3 days per decade (Maurer 2007). The mass of snow deposited every winter is of immense importance for agriculture, human water consumption, and electric hydropower production. The mass of snow is usually expressed as snow water equivalent (SWE). It is defined as equivalent height of water stored as snow. The density of snow is highly variable, but can be estimated based on the snow climatology (Sturm et al. 2010). Remote sensing methods to reliably estimate SWE are still in development, and especially difficult over mountains. Recent advances are made using airborne laser scanning techniques, which are now routinely applied in some watersheds in the United States. Most of the snowpack in the Northern Hemisphere is seasonal, and only a very small fraction perennial, as on the Greenland ice sheet. The snow cover in the Southern Hemisphere is mostly perennial, and is concentrated in Antarctica. Perennial snowpacks can exist at densities below 550 kg m$^{-3}$ for up to about 20 years, undergoing long-term metamorphism.

Snow has a very strong effect on growth and hydrology of forests. Snow is easily intercepted on branches, temporarily stored, and released later. Interception and radiation can have very different effects on total water stored. In the northern boreal zone, sublimation of the intercepted snow leads in general to a smaller SWE than in open land. However, forests in Mediterranean mountains often store more snow on the ground compared to open area, due to the shading of the trees and reduced radiation. This complex relationship is of interest as the increasing demand of water resources by competing uses intensifies. Management practices can significantly contribute to the optimization of forest function. Forests are also the best and cheapest way to prevent and protect from snow avalanches. Forests prevent the formation of snow avalanches by stabilizing the snowpack. The main effects are the interruption of weak layers by unloading of intercepted snow, the reduced formation of weak layers by the higher snow surface temperature under trees compared to the open field and reduced redistribution by wind. Although forests cannot stop a fully developed avalanche, they are able to stop or reduce run-out of smaller avalanches.

Snow is the key element for winter tourism. Winter tourism is a multibillion economy, and in many mountainous areas the key income. The availability of sufficient snow for snow sport is crucial to attracting a large number of tourists. The increasing number of alternatives to winter tourism has made the guaranteed availability of ski slopes a necessity. Spurred by the development of technical snow making (“artificial snow”) in the United States, all snow resorts worldwide are now routinely equipped with sufficient snow making capabilities to guarantee snow. The snow making technology developed to more energy efficient systems, some even using only the available water pressure. Specific reservoirs for snowmaking lower the pressure on other water resources. Typically, 30 cm snow depth during 100 days is considered as the minimal natural snow depth to run a ski resort economically.
SNOW

SEE ALSO: Precipitation; Recreation; Snow cover; Snow cover changes; Snow and ice avalanches; Water and climate change; Water resources and hydrological management

References


Further Reading

Social capital

Kenji Tsutsumi
Osaka University, Japan

What is social capital?

Social capital refers to the mutual trust, social norms, and supportive networks in a society or community. The idea stems from notions of reciprocity, and the concept has a strong relationship to several kinds of social matters. The ideas of trust and reciprocity are fundamental, and include a personal, strategic trust, which expects something to be returned by others (such as club goods), and a general, moral trust, in a wider or bigger society (such as public goods). According to Ronald Burt (Lin, Cook, and Burt 2008, 32), “[S]ocial structure is a kind of capital that can create for certain individuals or groups a competitive advantage in pursuing their ends. Better connected people enjoy higher returns.” A quantum leap from personalized trust up to a generalized trust may be vital for economic development. Though controversial, the concept is of particular interest to geographers, as it implies that it is necessary to look beyond a purely deterministic account of economic relations.

Geography and social capital

Under the current conditions of economic depression in advanced societies, various regional revitalization measures have been adopted for local implementation. Recently, some economic geographers have advocated that social regeneration, which can be achieved through the solution of social problems on a regional basis, will finally lead to regional economic regeneration, instead of giving subsidies from local government to the private sector. In such social regeneration schemes, various forms of private network or local social capital play important roles. Rather than offering “top-down” solutions, in this model the government should be a mediator to help enhance and capitalize on the underlying strengths in local neighborhoods.

Social capital is also of interest to health geographers; indeed, the relationship between well-being and social capital is a field that has aroused active discussion. For example, in some developed countries, a strong correlation is observed between a high elderly employment rate and low medical bill per capita for elderly people. Generally speaking, the higher the employment rate becomes, the more active citizens’ movements become, with more abundant social capital. Accordingly, in an aging society, social capital is of great significance when it comes to health policy and the cost of care. It has been seen not only to mobilize the power of the elderly, but also can have benefits when it comes to longevity and the quality of life afforded to people of the silver age.

Another important attribute that geographers bring to the study of social capital concerns the issue of scale. Societies and networks are formed across many different scales, spatial and temporal. Wide–narrow and long–short relationships offer a world of ontological settings. For geographers, social capital needs to be considered at multiple scales. Moreover, even in a virtual world, people can build networks with social capital. This
suggests a more spatially sensitive account of social capital, where it can be considered virtual or real, simple or complicated, wide or narrow. Social capital is formed, used, and activated in networks. One’s social capital may relate to several kinds of network: relationships among humans, resources, goods, information, and so on. John Urry’s concept of social scape is helpful when thinking about the problem. Social scape is a network that connects certain relationships with a spatial and temporal situation. It is a dynamic and structuralized entity, reconstructing dimensions of time and space (Urry 2000). In this spatial application, social capital may give shape to the social scape and generate change.

Histories of social capital

There is a long history of research into social capital. A number of writers see antecedents in the work of Aristotle, John Dewey, or Glen Loury, or in the excellent concept of public responsibility offered by Jane Jacobs in the 1960s. But few people directly defined or referred to social capital before Robert Putnam (2002), whose work is discussed in the following section.

In the nineteenth century, before the American Civil War, Alexis-Charles-Henri Clérel de Tocqueville, a French politician and philosopher, wrote that the democracy of America had been formed upon a social capital led by local community churches. Early in the twentieth century, Lyda Hanifan was perhaps one of the first people to suggest that a local school can be a center of local social capital in each community. Putnam has often cited Hanifan’s words as a simple and understandable definition of social capital. As for the “capital” of “social capital,” Hanifan did not refer to real estate, or to personal property, or to cold cash. Rather, he mentioned “goodwill, fellowship, mutual sympathy and social intercourse among a group of individuals and families who make up a social unit, the rural community, whose logical center is the school” (Hanifan 1916, 130).

Hanifan, a state supervisor of rural schools in West Virginia, would have liked to establish schools as public centers in communities. Thanks to the efforts of local teachers and neighbors, some schools acted as community centers, forming social capital through community surveys, community center meetings, agricultural fairs, school exhibitions, community histories, school attendance, evening classes, lecture courses, school libraries, and school athletics. This was a vision from the time of World War I, but even now social capital is said to have a strong influence on educational outcomes. For instance, it can be seen that in a region where the social capital index is high, rates of school violence and dropout are low.

In the early part of the twentieth century, particularly in the United States, modern capitalism developed on a drastically bigger scale, alongside social issues concerning to society and the economy. Some settings in American stories, such as An American Tragedy (1925), drew people’s attention to practical social phenomena, which form the basis of much of the famous works of the Chicago School of Sociology. The age of American pragmatism and practicalism was the embryonic stage of social capital, as captured by the sophisticated expression of Hanifan, Dewey, and others. Also, American sociology, rural or urban, had expanded into all the fields of community research, within which social ties and their changes became one of the central issues in the discipline.

Since the 1960s, Pierre Bourdieu, a French sociologist, addressed the concept of capital in a series of works on economic, social, cultural, and symbolic capital (Bourdieu 1983). Among these, social capital refers to a variety of kinds
of valued relations, some more important than others. In the United States, the sociologist James Samuel Coleman has referred to three forms of social capital: obligations and expectations, information channels, and social norms. He placed social capital as a resource of action and thought through which a rational action paradigm may be related to social structure, connecting a network of human capital with social capital (Coleman 1988).

In the twenty-first century, a few researchers have started to measure a part of social capital (Grootaert and van Bastelaer 2002). Some of them have belonged to the World Bank and have made much of measuring social capital to consult about regional development policy in developing countries. The Social Capital Thematic Group of the World Bank has developed two measurement tools: the social capital assessment tools (SOCAT) and the social capital integrated questionnaire (SOCAP IQ). There is some criticism of such measuring methods, because as a matter of course quantitative research limits treatment of qualitative data, but improvements have been made to these tools for use in the field in developing countries.

The revitalization of social capital

Since the late 1990s, the term social capital has increased in popularity. In his book Bowling Alone, Robert Putnam provides examples of research outcomes in the United States, claiming that in states with high social capital, schools work better, children watch less television, and violent crime is reduced (Putnam 2000, 299–310). Some scholars have criticized the vagueness of the definition, suggesting that social capital should be a social resource. In particular, it is important to distinguish between social capital, a social network, and social ties. But despite these shortcomings, the seed of social capital sown by Putnam has grown up and generated a lot of fruit for discussion and research. Putnam emphasizes that democracy should be sustained by the horizontal and spontaneous activities of citizens. Activating people towards more cooperative forms of behavior may contribute to social efficiency, as social capital. The concept can be a good platform to discuss life, intersubjectivity, and public utility in a society. For example, a sense of wellbeing, satisfaction, and altruistic behavior that cannot be explained by economic rules alone may be built upon the foundation of social capital.

Putnam has also referred to a distinction among types of social organizational activity. Bonding and bridging refer to relations of internal and external networks respectively. For example, although a little bit simplified, an organization in a premodern, traditional rural community (or in a Gemeinschaft) is a bonding network, and a bridging network is one among socially heterogeneous or inter-independent groups (or among Gesellschafts). In other words, groups with a high level of bonding capital tend to be self-reliant, with close links within their networks. In contrast, those with a high level of bridging capital have the potential to make links across networks and, therefore, have more extensive connections and are not wholly reliant upon a single node or community.

According to Putnam’s argument, recent decades have shown declining trends in the power of traditional groups such as unions, parties, and churches, to accumulate social capital. But some new groups have emerged, including sports groups, environmental groups, new social movements, and so on. Naturally, the attributes of actors of each group have changed. Social capital is mentioned constantly in relation to communities. In many works, a community is a matrix of social capital. Some of
SOCIAL CAPITAL

the essential points in the matrix are presented in Box 1. Regarding social capital in a community, attention should be paid to its membership, purpose, property, accessibility, connection, energy, and functioning. Activities that develop social capital would work in a very complicated structure, which can be dissected from several perspectives, scales, and levels. The basic unit for social capital is an individual. A group of people form social relations or networks with certain purposes, goals, or motivations. This bonding perhaps leads to further bonding, or bridging to another level. At each phase, or stage, of shaping social capital, important elements can be seen. After Lin, Cook, and Burt (2008), the utility of social capital also differs by setting.

The limits of social capital

A number of works have addressed relations between social capital and economic phenomena. For example, a large accumulation of social capital reduces transaction costs, resulting in a positive impact on economic growth. In economics, income inequality is due to differences in preferences for labor and differences in the abilities of individual persons. The variation among individuals is not a major problem discretely, but from the point of view of social capital, the expansion of economic disparity would lead to damage to social capital, which then generates further disparity. Inequality of endowments of social capital seems to lead to a widening of economic disparities.

Sometimes social capital is damaged by economic disparity, crime, or violence. And one form of social capital may itself sometimes resist other forms of social capital. Very strong social capital with tight inner ties in a bonding-type community will result in a restriction of the freedom of the members, and the group will be restrictive and exclude others. Social capital is, therefore, often considered in connection to external relations. Thus, the dark face of social capital can be seen: exclusiveness, external diseconomy, and so on.

Regarding disparity, developments in information technology are likely to create fewer winners and more losers, with social fragmentation and bipolarization among classes. If this occurs, young people tend to be NEETs (not in education, employment, or training). In EU countries, in addition, immigrants experience social exclusion, as a result of low social capital, and public policy should be developed to avoid the isolation of young or immigrant people, leading to social inclusion and social protection.

Social capital in a society has a major impact on the efficiency of the government. As government itself tends more and more to outsource its own work, the question arises as to what kinds of social capital and networks the government can construct. The government has an obligation not only to promote and maintain the external establishment of social capital, but to generate and capitalize new social capital by itself.

In today’s information society, some people believe that television and computer games have damaged social capital because they have deprived people of time and the chance to establish and maintain social relationships. But the Internet and SMS text messaging can also expand social networks, albeit in different ways. People can create virtual communities that are free from the bonds of place. They might also strengthen social capital of the bonding type, even among family members and good friends. In an enterprise, for example, an in-house local area network (LAN) and a cloud computer system are very efficient at bridging sections; therefore, in the information society bonding
SOCIAL CAPITAL

Box 1 Important elements of social capital in a community.

- Member or membership: Who is the member? (Age, sex, occupation, status, ability, consciousness, motivation, etc.) A hierarchy of members is also an essential factor.
- Purpose, goal or motivation: What is the logic behind establishing social relationships for social capital?
- Property of social capital: What are the resources, rules and norms of social capital?
- Boundary or range: How broad is the group to which social capital is extended?
- Accessibility: Is there a degree of freedom or exclusiveness to use the social capital, and how can people access it?
- Connection: How does one unit of social capital connect to others (further bonding, bridging, linking, etc.)?
- Energy: How does the social capital get energy, or how is it reinforced? (Adding budget, members, trustworthiness, etc.)
- Interscale or interlevel functioning: How does the social capital work on different scales or levels?

Source: Tsutsumi 2011, 136 (modified by the author)

and bridging are still possible, and are becoming more complex with globalization.

There is also a problem of government and research perspectives on social capital creating a dualism: “top down versus bottom up.” The roles of national and local institutions in restoring American community need to be complementary; neither can solve the problem alone (Putnam 2000, 413). But in some countries, such as Japan, at the beginning of the twenty-first century, most scholars who had introduced the concept of social capital were economists, with a strong top-down perspective, and a tendency to speak of the nation state, but not of local smaller communities. However, there were also regional sociologists who had been more accustomed to a bottom-up perspective, making much of social ties or social relations in a community. Furthermore, a traditional sociologist would say that social capital is not new at all and terms such as social ties and social relations have been around for more than 100 years, so they may avoid using the term “social capital.”

Conclusion

Although on first consideration Putnam’s definition of social capital is relatively simple, it is, in fact, more complex, and subsequent research has been broad and deep. Putnam’s definition is constructed from structural, collective, behavioral, and cognitive perspectives, because it is formed from functional and structural, or physiological and morphological, points of view.

Although social capital can be formed even in a virtual or cyber world, it is still a very convenient conceptual tool for the study of community, regional education, regional identity, regional welfare, social networks, and social relations in space. In other words, it is sometimes rooted in a place, so social capital has a certain geographical character. Hanifan (1916) saw a school or a church in a community as a center for social capital, and Putnam (2000) also made regional disparity clear on a scale of states. Social capital consists of several
kinds of function that are rooted in a place or a space within a regional scale system. And social capital can be researched or analyzed geographically in the context of such topics as regional living functions and regional differentiation (Tsutsumi 2011).

In shrinking societies, local social capital has the advantage of supporting the regional daily life of residents at varied points: education, medical service, welfare, shopping, housekeeping, security, and so on. Sometimes religion, sports and recreation help people form a network of social capital. On the other hand, in developing societies a network with vernacular social capital can act as a conduit for industrialization and modernization.

SEE ALSO: Community; Community-based natural resource management; Cultural capital; Regional inequalities

References


Social constructionism

Szu-yun Hsu
University of British Columbia, Canada

Tsung-yi Michelle Huang
National Taiwan University

Social constructionism and its critique

Social constructionism – a trend of theoretical thinking about the social construction of reality, and synonymous with the “cultural turn” in the social sciences and humanities in the 1980s – became popular in human geography in the early 1990s. The roots of social constructionism lie in a number of philosophical movements re-examining what used to be regarded as “common sense” or “reality,” as well as the means by which we perceive it. Its precedents can be traced back to the nineteenth century, when criticism of teleology and metaphysical ontology by post-Kantian continental philosophers – from Hegel and Marx, to Sartre and Nietzsche – eventually contributed to the rise of poststructuralist scholarship in the late twentieth century; American pragmatism, pioneered by theorists such as John Dewey, also engendered constructivist philosophies aimed at critically reviewing traditional epistemology in regard to the foundation of knowledge, everyday experiences, meaning-making, and communication.

Social constructionism has developed into an umbrella concept over time, covering a wide range of philosophical trends spanning poststructuralism, postmodernism, hermeneutics, and phenomenology, and thus defies easy definition. Nonetheless, social constructionism has the goal of accentuating differences, situatedness, and positionality in knowledge construction; specifically, it refutes the universal theories and essentialist and foundationalist categories that engaged in stereotyping and social marginalization. It is now widely acknowledged, for instance, that there is no inherent “truth” to gender and the assigned social roles defined by it. Categories of race and ethnicity have also been deconstructed, with whiteness being unraveled as formed in opposition to the racial Other. The complex intersections between always multiple and situational identities have undergone critical examination as well. In geography, the ways in which place, landscape, and nature are socially constructed and contested have been extensively interrogated. This intellectual development of social constructionism has contributed to the emergence of identity politics, the politics of recognition, and the celebration of cultural diversity, which have been posited as politically emancipatory and progressive.

Despite its growing prevalence in the social sciences, social constructionism has aroused heated debate regarding its theoretical validity and its political and moral desirability. This is especially the case for the realist scholarship initiated by Roy Bhaskar. First, its nonessentialist ontology and its emphasis on language has been criticized for being haunted by philosophical idealism, for if everything is socially constructed, how can things independent of mind, such as rock, or sensations, such as bodily pain, be accounted for? However, criticism of this kind usually resorts to a binary reading of mind and matter, which itself is contentious. Second, social
constructionism is often considered theoretically flawed, for its inherent difficulty lies in identifying the causes of social change. The charge of relativism is also apt: if reality is entirely subject to construction from different social positions, it then becomes impossible to identify the victims of social oppression. Moreover, its associated identity politics may hint at hermeneutic emancipation without substantial political actions. Such apprehensions led the geographer Sayer (1997, 445) to contend that social constructionism is “unsatisfactory with regard to ontology, epistemology and its conceptualization of nature and society and is as open to reactionary interpretations as it is to progressive ones.”

Commenting from the position of critical realism, Sayer further distinguished a strong form of social constructionism from a weak one; the former deems reality and social life as nothing more than what we know about them, and is therefore more aligned with a postmodernist philosophical stance, while the latter posits that there is a reality that can be known but that there are different, always partial and often competing, ways of recognizing it. In this formulation, Sayer argues that critical realism and social constructionism are not incommensurable, for critical realism shares some common ground with weak social constructionism in its view of relational and constantly evolving essences, concerns for historical contingencies, and heedfulness to differences.

Discourse theories in social constructionism

The development and critiques of social constructionism have contributed to multifarious evaluations and contentions, mostly in relation to epistemological and ontological questions. Scholarship indebted to Foucault’s theoretical framework, or poststructuralism in general, addresses the critique of social constructionism by highlighting the notion of discourse and its roles in social processes. Instead of viewing language as a mere reflection of the “real world,” discourse theories emphasize how linguistic structures produce meaning for social practices and constitute “reality.” It is noted, however, that there are divergent approaches to discourse, ranging through critical discourse analysis proposed by critical realism scholarship, Laclau and Mouffe’s discourse theory, to Butler’s performativity theory. These discourse-focused approaches, while all rejecting a postmodern take on the world as constituted solely by free-floating symbols, diverge not only in their definition of discourse, but also in their understanding of power, the relation between discourse and materiality, the formation of subjectivity, and whether/where potential space for resistance can be identified. These differences have contributed to their distinct ways of addressing social constructionism.

Asserting the importance of incorporating a cultural dimension into the tradition of political economy scholarship, critical discourse analysis (CDA), developed by Fairclough, and cultural political economy (CPE), advocated by Jessop, among others, provide a theoretical supplement to social constructionism from within critical realism. While these approaches acknowledge the constitutive role of discourse and language in the processes of social change, they insist on the distinction between discourse and nondiscursive social realms – where institutional arrangements or economic forces belong – and on their dialectical interplay. As such, it is believed that only through scrutinizing the interrelation between discourse and extra-discourse can social causalities be accounted for, while at the same time avoiding the quagmire of political relativism inherent in social constructionism. Methodologically, both CDA’s and CPE’s take
on discourse emphasizes semiotics and regards the intersubjective production of meanings as their major site of analysis. The philosophical outlook of critical realism and the methodologies stemming from it have had great implications in economic geography, which is now in search of analytical approaches to the study of the discourses of economic imaginaries and the intersections between governance, identity, and the restructuring of political economic forces.

Grounded in Gramsci’s theory of hegemony, Laclau and Mouffe (2001) use discourse theory as a way to rework the Marxist framework of base–superstructure as well as subject–object relations. In contrast to critical realism’s preservation of the discourse/nondiscourse demarcation, Laclau and Mouffe assert that the totalizing nature of discourse includes all forms of social practices. In their theoretical framework, “reality” is constituted through endless discursive struggles until a fixed meaning system – so-called hegemony – is reached, which makes social constructionism a question of hegemonic construction. Since there is no pregiven reality that can be apprehended prior to discursive practices, and it is impossible to tell exactly where discourse stops and the nondiscursive realm starts, they contend that the line between discourse and nondiscourse should not be drawn analytically but empirically. It is because of the nonpredetermined nature of subject positions, existing in the non-eliminable gaps between ideas and materiality, that room for radical democratic politics opens up. This theoretical approach has been popular among political geographers in their inquiries into popular culture, popularism, social movements, and resistance in the face of hegemonic dominance operating on various scales and in various locales.

A counterargument against the critique of social constructionism as idealist and politically relativist is addressed by Butler. Resorting to Foucault’s notion of discourse as power effects, Butler’s approach delimits the role of human will in language and foregrounds the historicity of discourse. Moreover, Butler shows that the material/discourse dichotomy – a position underpinning the charge that social constructionism is overly idealist – is exactly the premise upon which power operates and is normalized. Butler posits construction as a constructive constraint that demarcates, differentiates, and renders certain matters real, yet unintelligible – the very logic that sees gender as socially constructed also insists on the existence of sex as nothing more than “real.” It is through the production of such pre-discursiveness prior to human thoughts, for example, that heteronormativity is naturalized. To resolve the structure/agency dichotomy, Butler further draws on the notion of performativity – a reiterative, yet unstable, process from which subject positions emerge and are constantly (re)formed (Butler 1993). Butler’s take on discourse and performativity seeks to overcome the mind–body divide and question the superiority of human will, themes which surface in much of the critique of social constructionism. Performativity has been taken up widely by geographers in relation to studies of gender, sexuality, ethnicity, race, place, and space. Nonetheless, the notion of performativity is also cautioned against by geographers for being unable to attend to the historical and geographical embeddedness of subject formation.

Social constructionism in geography

This critique has been furthered at both ontological and epistemological levels by the flourishing scholarship challenging human centrism through decentering discourse, reason, and the human being at large. Scholars have attempted to discard the residual Cartesianism that has haunted modern human knowledge, and have looked
SOCIAL CONSTRUCTIONISM

to philosophies such as phenomenology and Deleuze and Guattari’s theories as a way out of the deadlocked debate between realism and social constructionism. The emergence in geography of nonrepresentational theory (NRT) and the wide adoption of actor-network theory (ANT) are two ways in which the discipline has contributed to this critique.

Nonrepresentational theory, a project launched by the geographer Nigel Thrift (2007), among others, in the 1990s, marks a departure from a representation– (or discourse–) focused approach to social constructionism. Also known as “more-than-representation” theory, NRT focuses on the embodied and relational practices through which a performed subject arises – a concern shared with performativity scholarship. Yet, instead of chasing meaning–making practices (and the unavoidable accentuation of language and reason), NRT highlights unintentional, precognitive, or noncognitive affects and emotions generated from encounters between hybrid entities – be they bodies, things, or ideas – that bring forth the social. The tenets of NRT attend to the lifeworld as it exists in the interrelational actions between such entities, and thus is always “becoming,” “emerging,” and “destabilizing,” forming spaces of in-betweenness.

Likewise, drawing on the studies of laboratory practices and technology, sociologists led by John Law and Bruno Latour formulated ANT as a nonanthropocentric approach to debunk human mastery that sets up the one-dimensional “tug of war” between realism and social constructionism (Latour 1992). While stressing the socially constructed aspect of science, Latour refutes the ability to explain nature in terms of social construction. ANT is neither a return to the realist account for the “out-thereness” awaiting discovery nor a return to its interpretation by human agents. Instead, ANT rejects the human being as the only actor with the power to contribute to the formation of the social – either through discursive articulation or through material interaction. It holds to the ontology of emergence, where different entities (or so-called actants) – humans, animals, things, and discourses, for instance – are constantly enrolled into a heterogeneous network of assemblage. Although not exclusively a geographical theory, ANT is widely deployed within the discipline, resonating with many geographers’ concerns about spatiality, socio-nature, and connectivity.

Social construction of scale

The vibrant ontological and epistemological discussions surrounding social constructionism take place both within and outside of human geography scholarship, and have largely reshaped the ways in which the core concepts of the discipline – scale, for instance – are theorized. Scale, a way of looking at, thinking through, and acting upon variegated social, economic, and political processes, once regarded as fundamental to the discipline by many, since the 1990s has been questioned for its self-evident nature. Despite differences in conceptualizing the social production of scale, studies range through (re)scaling processes of capitalism to the scale-making and scale-jumping strategies deployed by various actors.

Nonetheless, the scale literature, as such, has been critiqued for being unable to eschew the limitations of a hierarchical ontology. Against this backdrop, Marston, Jones, and Woodward (2005) proposed a provocative alternative approach, calling for a “flat ontology” to replace any transcendental understanding of spatial organization. With the abandonment of scale, “site ontology” – always emergent and only existing in the interactive processes between humans and non-humans – is favored for its theoretical openness.
and political radicalness. Yet such a proclamation has received a diverse range of evaluations and criticism. While many wonder whether site ontology parallels the ontological ground of ANT or that of phenomenology, others are concerned with the political implications of such a theoretical practice, criticizing it for falling short of addressing concerns of power and politics.

While it is generally agreed that scale is socially constructed and therefore dynamic and relational in nature, it remains unresolved as to whether scale denotes a complex set of ontological structures through which social relations are organized and contested, or whether scale should be understood as a language of spatial politics that produces its own empirically determined “scale effect.” Viewpoints over which approach possesses more radical potentiality in opening up space for resistance are also in contention and resemble the ongoing struggle, straddling critical realism, poststructuralism, and much of the scholarly work deriving from, yet radically modifying, social constructionism.

Developing themes within and beyond social constructionism in geography

There has been a sea change in the research agenda of human geography, with a move towards unconventional themes triggered by the recent developments in social constructionist literature. The subfields of economic geography, political geography, and political ecology, among others, are undergoing considerable revision through approaches adapted from, but not limited to, those reviewed above.

In economic geography, the economy and the market have been widely viewed as having logics and spaces not only indifferent to culture, but also independent of how we understand or study them. This doctrine did not change much when economic geography shifted its attention from neoclassic economic theories to that of Marxist political economy; however, the ontological and epistemological premises of such a theoretical orientation have met with critical examination from social constructionism. The constructive ways in which economic imaginaries and discourses, such as “globalization” and “development,” contribute to the organization of material, institutional, and social relations is critically investigated by poststructuralist economic geographers. Unpacking capitalism, Gibson-Graham’s (1996) work presents a critical inquiry into the politics of knowing and urges us to recognize the real existing diverse economic practices discursively excluded from orthodox political economy’s theorization. Informed by both performativity theory and ANT, Barnes (2001) further delineates how the discipline of economic geography and its subject of study came into being in the situated messiness comprising a hybrid network of theories, textbooks, institutions, and professionals.

Social constructionism has also found an extensive influence in the recent development of political geography, whose traditional theoretical foundation – be it notions of states, sovereignty, borders, or territory – has been fundamentally reshaped. It has directly contributed to the burgeoning critical geopolitics literature aimed at unraveling the geopolitical reasoning and discursive practices of statecraft and the world politics associated with them. A critical interrogation of Cold War rhetoric demonstrated the importance of discourse in constituting the nature of war and military manipulation. Fruitful work on border studies has also revealed the socially constructed nature of modern state boundaries and the violence associated with border practices. Working with emotional geographers, recent work in political geography has begun to apply affect and NRT to scrutinize issues from a geopolitics
SOCIAL CONSTRUCTIONISM

of fear to nationalist aspirations, showing that the nature of politics is not restricted to discursive struggles, but also embodies feelings and emotions.

The reworking of social constructionism is also manifest in the realm of recapturing the nature−culture relationship, where the conventional wisdom of the social construction of nature is being critically revisited in distinctive ways. The poststructuralist approach not only challenges the dualisms between human and nature upon which the notion and the existence of “nature” are formed, but goes further to depict the mutual construction between the two. The development of posthuman scholarship further discards the anthropocentricity of a constructionist approach and turns to heed the hybrid practices of matters such as animals, water, and disease, manifesting the wide application of ANT in the subfields of animal studies, resource geography, and health geography. Another theoretical intervention into the social/nature question can be found in Haraway’s (1991) interrogation of technoscience and the very nature of life. Her construction of the “cyborg” as “a cybernetic organism, a hybrid of machine and organism, a creature of social reality as well as a creature of fiction” (149) breaks down the boundaries between animals and humans, animal-humans (organisms) and machines, and the physical and nonphysical, by which the inquiry of political ecology is fundamentally modified.

SEE ALSO: Actor-network theory; Cultural turn; Discourse; Nonrepresentational theory; Poststructuralism/poststructural geographies; Realism, critical

References


Further reading

Social geography

Vincent J. Del Casino Jr
University of Arizona, USA

Social geography is a subdiscipline with its roots in a number of different intellectual traditions and places. It has developed into what it is today as a result of the institutional and theoretical contexts in which it emerged, flourishing in some national contexts and barely recognized in others. The question of what constitutes social geography is still debated, as scholars consider what it means to study social life geographically. Broadly speaking, social geography focuses on the relationship between society and space, with a particular emphasis on issues related to social identity, nature, relevance, and justice. Social geographic research also focuses on the spatialities of difference and inequality and has links to the question of “who gets what, where, and how” (Smith 1994). With diverse histories that stretch from European anarchist traditions in the early nineteenth century to the social relevancy debates and the quantitative revolution that took hold of geography in the 1960s, and later the feminist, postcolonial, antiracist, and Marxist geographies that emerged in the 1970s across the globe, social geography today is a mix that draws on a range of theoretical and methodological approaches to include more recent interventions from poststructuralism and postmodernism.

Historical origins of social geography

Social geography’s historical roots are challenging to trace. Some claim that sociologists in France were the first to use the term “social geography,” while others trace its origins specifically to the work of the French-born scholar and anarchist, Élisée Reclus, both in the late nineteenth century. In reality, social geography has its roots in French, German, and British geography and has developed different trajectories in the United States, Canada, New Zealand, and Australia. The history of social geography also parallels the development of the wider discipline across Asia, Latin America, and Africa, where social geography has a complex relationship with, for example, the system of apartheid in South Africa or international collaborations across Southeast Asia. Rob Kitchin’s (2007) project, Mapping Worlds, illustrates some of this breadth, even as he acknowledges the challenge of internationalizing social geographic thought in relation to the hegemony of Anglo-American traditions, a concern that John Eyles (1986) also noted in Social Geography in International Perspective. Instead of tracing the history of social geography through its own diverse geographies by country (see reports in Social & Cultural Geography), it is better to approach its history through a framework that highlights both the inclusions and the exclusions that mark its development over time.

Social geography in the early to mid-twentieth century

In the early twentieth century, geographers had to contend with a significant amount of theoretical and methodological change, as evolutionary thinking began to permeate not only the natural sciences but the social sciences as well. For some geographers, the integration of theories of natural selection and other components of
evolutionary thought led to environmental determinism and theories of social evolution whereby climate and other physical factors were said to determine social structures, the success of and hierarchy among groups of people in particular places, and the spatial organization and distribution of different ethnic groups, races, and societies. Some of these theories were applied to overtly racist understandings of social development, while others were picked up and appropriated by eugenics movements as well as used in late nineteenth-century anthropogeography and twentieth-century geopolitics. For geographers interested in explaining the spatial organization of societies and their distributions over time, the larger question was the relationship between nature and society, the natural and the social worlds. Ontologically speaking, Western thought had, in the nineteenth century, begun to develop a fairly clear delineation between nature and society – the natural world was out there and organized by societies to meet human needs. Environmental determinism inverted the relationship, regarding nature as a determining factor in social development and organization. Although the directional arrow was changed, the ontological division between nature and society remained.

Frederich Ratzel’s *Anthropogeographie*, which was integral to German geopolitical theory, was based on social evolutionary theories of human development and spatial organization. Ratzel’s influence extended to other parts of Europe and to the United States, where his student, Ellen Churchill Semple, applied his thinking in her famous *Influences of the Geographic Environment*. As Semple argues, “man is a product of the earth’s surface,” and, as such, geographers must understand “the dominance of different geographic factors at different periods” (1911, 1, 31). Semple’s theories took as a starting point a global environmental system, which was mediated by local physiographic factors. Her theorization was not simple or straightforward but suggested that the “geographical environment, through the persistence of its influence, acquires peculiar significance … in the shifting fate of races” (Semple 1911, 6). This environmental basis of racial difference emerged as a critical aspect of early environmental determinism, which is also found in the work of Ellsworth Huntington (1915), whose *Civilization and Climate* made direct linkages between the spatial organization of racial differences and climate. By 1930, C.C. Huntington and Fred Carlson had tempered the argument about environmental determinism in *Environmental Basis of Social Geography*, in which they outline two disciplinary subareas: physical geography and social geography. Huntington and Carlson suggest that “it must not be supposed … that environment alone can furnish all of this explanation of ‘social causation’” (1930, 44) even as they insist that “location is a geographical factor of fundamental importance. It largely discloses the nature of man’s environment, and profoundly influences the character of his activities” (Huntington and Carlson 1930, 63). This tension – between (different degrees of) sensitivity regarding the complex drivers of social organization and the determining position of the environment – informs the question “How do we describe and explain the spatial distribution and organization of social life?” Despite Huntington and Carlson’s attempts to elevate social geography to the human social science approach in geography, in the United States in particular, cultural geography and regional geography held much more import, as social geography remained marginal to the discipline. This was not the case in other countries, such as the United Kingdom, France, and Germany, where social geography built on environmental traditions and also expanded other social theoretical work to explore the spatial organization of societies.
Despite its marginal status in US geography, social geography changed throughout the early to mid-twentieth century, as questions of the spatial organization and distribution of societies animated social science in Europe and North America. European social geography was strongly influenced by the work of French geographers, including Paul Vidal de la Blache, whose “science of place” as deep regional description was picked up by French and British social geographers during the early twentieth century. Many social geographers borrowed from Vidal de la Blache’s approach, focusing on the material aspects of social life, although he acknowledged the importance of studying social networks and norms as well. By the mid-twentieth century, geographers in the United Kingdom and a growing group of urban geographers in the United States were beginning to draw more inspiration from the field of sociology and the work of scholars, such as Park, Burgess, and McKenzie (1967/1925) who famously employed ecological models to study the city and its social differentiation. Their text, *The City*, strongly influenced the development of urban geography and social geographies of the city more generally by focusing attention on issues of racial and ethnic segregation and criminal patterns, as well as how cities were spatially organized over time. Park, Burgess, and McKenzie also theorized models of change in the city, how neighborhoods developed new characters and populations. In the United States, these ideas found their way into quantitative urban analyses of cities and regional development. In fact, the development of social geography in a number of countries, such as Germany, relied on deepening conversations between sociology and geography.

At the same time that social geography began to grow in North American and European contexts, many nongeographical scholars, such as W.E.B. DuBois and Jane Addams, were asking social geographic questions. DuBois’s *The Philadelphia Negro* paralleled the scope of Park, Burgess, and McKenzie’s work, although those scholars never recognized DuBois’s contributions (Sibley 1995). As David Sibley (1995, 154, cited in Del Casino 2009, 39) argues, “DuBois eschewed Social Darwinism, recognizing that it provided a justification for the oppression of black people.” DuBois’s qualitative inquiry into the lived experiences of black people in Philadelphia was seen as marginal to the scientific work of other scholars at the time. And yet, DuBois’s scholarship has more affinity with the work of many current social geographers, who often use qualitative methodologies to study race and racism. Jane Addams’s scholarship has also been marginal to the broader cannon of social geography and sociology (Sibley 1995). But, her research, which was part of a larger Chicago project called Hull House, applied sociological and geographical methodologies to the study of urban housing discrimination. Like DuBois, Addams and her Hull House colleagues deployed their research toward applied, political ends, which included the reorganization of urban space to address poverty and racism. Their work did not fit neatly into the objectivist approach of Park, Burgess, and other urban sociologists and geographers.

From today’s perspective, the research of DuBois and Addams is easily recognizable as social geography, but it is not surprising that their work remained marginal to the historical writing of social geography as the twentieth century moved forward. The quantitative revolution further reinforced this marginalization, as geographers conceptualized their field as a social science similar to economics. The turn toward a spatial scientific geography, which included the incorporation of hypothetical and deductive approaches as well as statistical methodologies and the use of large datasets, was applied generously to social geographic questions in the
1960s and 1970s. Social geographers began to analyze the spatial organization of racial and ethnic groups, crime patterns, segregation, and access to services, for example, asking how distance from a certain place or the agglomeration of particular activities or people could be scientifically analyzed to explain how and why they appear as they do. The goal was to model and map the patterns of difference among groups and peoples, and to predict how those patterns change over time. Richard Morrill (1965) examined the US ghetto and how its spatial organization could be analyzed through a scientific approach, modeling the variables that have driven the growth and decline of these spaces. Morrill’s work attempts to model diffusion and suggests that the racist practices of white urban dwellers could partially explain how neighborhoods change in relation to their social organization. Put another way, Morrill’s research presupposes other urban social geographic work on “white flight” and the suburbanization of extra-urban areas in the 1960s. Morrill suggests that “a real reduction in ghettoization implies governmental, not a voluntary, regulation of the urban land and housing market” (1965, 361). In taking up the question of race – in this case black and white populations in Seattle – Morrill argues that social geographic research can and should be applied to address social problems which, he suggests, have a spatial dimension.

Social geographers in the mid-twentieth century also drew on other social theories, particularly those emerging from humanism, to think about the relationship between human lived experience and place. Humanistic philosophies recenter social geography on subjective experience and qualitative methodologies. David Ley pushed social geography to pick up on humanist approaches and examine the “taken-for-granted realm of experience” and the social realm of “shared meanings” ((1977, 509, 505). Ley theoretically positions places as sites of both shared meaning and lived experience, the latter differentiated by one’s subjective relationship to that place. These approaches demand a new type of fieldwork involving conversation, description, and deep connections between researchers and research subjects. The shift away from the distanced perspective of spatial science strongly influenced social geography moving forward, particularly as new social theories entered the scene and questions shifted toward explorations of different people’s lifeworlds. This research mapped onto critical social geography in the 1970s and 1980s, which sought to reimagine the relevancy of social geography to the world in which that scholarship was being produced.

Social geography’s critical turns

As social geography began to take up questions of social difference and inequality, social geographers turned their attention to the spatial organization of society and the ways in which that organization benefited some populations and not others. That work did not always take up deeper social issues, such as the history of racism, sexism, or classism that could explain the everyday social geographies of living in the city, for example. And, when social geographic work did take up those questions, the larger field of geography often ignored it. It was the turn to a more critical social geography informed by the wider social changes taking place at the time that brought social geographers into conversation with Marxism, feminism, cultural studies, and European continental philosophy as well as antiracist and postcolonial theory. Bill Bunge’s (2011/1971) Fitzgerald: Geography of a Revolution engaged presumptions about the spatial organization of difference and inequality through a focus on one neighborhood, Fitzgerald, in Detroit. Bunge’s analysis of how people grow
up in a ghetto demonstrates that the spatial organization of the city is quite fluid. As children turn into youth, spaces that were once available to them are closed off by racialized politics and violence, marginalizing male black youth in particular. A politics of racialized and class fear structures the spatial organization of Fitzgerald and the state response to the neighborhood, producing material barriers in the landscape that limited the social mobility of people living there. In many ways, Bunge’s *Fitzgerald* asked very different questions from Morrill’s study of the ghetto, even as both were rethinking how cities might be better organized to address inequalities. Bunge worked from the ground up to ask what it felt like to live in a ghetto, and what structural constraints were found in that space.

David Harvey’s (1973) *Social Justice and the City*, more than Bunge’s work at the time, marked a distinct break from a spatial scientific social geography, as he took up the question of the spatial inequalities that are part and parcel of capitalism. His class-based analysis suggested that social and spatial differences in the city are structurally embedded in capitalist social relations. Put another way, Harvey pointed out that a spatial science of the city, which focuses on social groups, activities, and places as points in space that can be analytically evaluated through a scientific model, fails to account for the fact that “the city has to be regarded as a functioning totality within which everything is related to everything else” (1973, 303). The social relations of capitalism are the ways through which people and places relate and are organized. The development of a more socially just social geography is paralleled by the question of social geography’s relevance to everyday social problems. Relevance thus emerges as a topic in parallel to the work of Harvey from social geographers in the discipline with interests in applied geography. However, the approach to relevance also prompts the question of for whom relevance was being measured and defined.

The Marxist turn in social geography also ushered in a retheorization of the relationship between space and society. Drawing from Henri Lefebvre, a French social theorist, geographers began to theorize space as a social relation, and the ontological distance between space and society collapsed. Edward Soja coined the term “sociospatial dialectic” and Derek Gregory developed the concept of spatiality to explain this new understanding of the space–society relation, suggesting that, for example, “all organized space will be seen as rooted in a social origin and filled with social meaning”; filled with social meaning, space “is a structure created by society” (Soja 1980, 210, cited in Del Casino 2009, 45). Space is an active agent in the structuring of social relations, as spaces produce and reinforce capitalism and by extension other oppressive forces, such as patriarchy or heterosexism. Theorized this way, space is not a backdrop to social relations. Rather, space is always already social, as it helps constitute difference and maintain inequalities (see Figure 1). Social geographers suggest that the organization of space reinforces the division between “races,” articulates the boundaries between classes, and maintains the historical marginalization of gendered and queer bodies. It is also through protesting such boundaries and divides that dominant forces are challenged. The application of this retheorization of space cannot be understated; space is also a social construction, constituted by the same relations that reinforce race, ethnicity, gender, sexuality, and class. Social geographers thus began to question the representational politics of power and authority that produce spatialities of difference. As geographers expanded their intellectual reach, the realm of the symbolic and the importance of symbolic spaces to the constitution of social life and sociality became more important. This is where
the boundaries between cultural geography and social geography became less clear, as culture was theorized as a social relation but also as a political process through which social categories and differences became sedimented in space.

The cultural revolution of the 1970s, which marked the world’s rapid period of decolonization and the feminist, queer, and antiracist movements in many parts of the world, affected the types of social geography being written. Feminist geographers criticized spatial science’s study of Homo economicus (economic man) as a clearly masculinized subject. Feminist geography developed the notion that...
geographers must analyze how gender relates to the experience of living in particular places. From there, feminist geography asked inherently social questions about the spatial organization of gendered lives and how those spaces both reinforced gendered roles and norms and offered opportunities to challenge patriarchy and sexism. Early feminist geographers focused attention most closely on women’s lives, while feminist theory began to push the question not only toward women’s roles and experiences but also to the broader sociospatial constitution of gendered relations. Drawing from a wide range of theories, feminist geography quickly engaged social questions related to women’s inequality, heterosexism, and the hegemony of masculinist models of the spatial organization of societies, as well as the gendered body. By the 1980s, the cultural revolutions taking place in many societies had infiltrated the discipline of geography, although feminist geography and the social geographies of gender remained rather marginal to a wider disciplinary canon. By the mid-1980s social geography had begun to take on a very different and more diverse sensibility, in large part based on the influence of feminist geography.

The expansion of postcolonialism, a function of the massive wave of decolonization that took place across the globe in the 1960s and 1970s, also raised questions and produced challenges for social geography. The postcolonial scholarship of Edward Said and Frantz Fanon illustrates how the colonial imagination extends well beyond the specific geographies of colonial occupation. Both argue that the binary logics of Western thought – for example, white/black, colonizer/colonized, West/East – are hierarchically structured and problematically reinforce relations of difference and inequality. This work has had a significant impact on geography, and the cultural production of the East as an object of Western desire and control has been taken up in critical social geography, particularly by those interested in the relationship between spatial practice, thought, and identity. Perhaps most interesting is Said’s theorization of an internalized Orientalism, whereby former colonized elites appropriate Western cultural representations of an idealized society, and deploy that within their own national contexts. Social geographers have built on this critique to suggest that modern-day oppression and violence are situated within much wider power relations that include the Orientalist thinking of hegemonic global authority and geopolitics. Postcolonial theories thus help social geographers explore how concepts of race and racialization are situated within global geographic imaginaries and colonial and neocolonial projects. Postcolonial work has also inspired social geographic scholarship as it is taken up directly in revolutions and in inspiring anticolonial efforts. Steve Pile (1997) draws on Fanon to re-examine French colonial occupation and the spatialities of conquest as an incomplete process. Indeed, as Pile suggests, women’s experiences of the Algerian revolution demonstrate that “relationships of confinement, passing, boundedness, unbinding, birth are profoundly spatial as much as they are also temporal. In occupying space, women changed that space” (1997, 22). In short, the social geographies of the city have been altered both by the oppressive occupation of French colonialism and by the everyday practices of Algerian women who “changed the rules governing gender and sexual relationships as well as those imposed by colonial administrators”; as a result, the French occupation was “never more than partial” (Pile 1997, 22, 23).

The emergence of subaltern studies in India also challenges the broader canon of Western social theoretical thought, affecting the ways in which social geographers imagine their research
and the relationship between researched and researcher. Indian postcolonial feminist scholars, such as Gayatri Chakravorty Spivak and Chandra Talpade Mohanty, have had a dramatic impact on how social geographers think about their work. Their challenges to Western feminism suggest that the constitution of “third world women” as an object of Western discourse is problematic, an issue that social geographers have taken up through research on the historical construction of the gendered subjects of colonialism as well as through more recent considerations of how racialized, postcolonial subjects are constituted through the practices of modern-day development work. That said the postcolonial turn within India had limited impact on Indian geography in the 1970s and 1980s, where much social geographic literature focused on South Asia as a region and not as a site of knowledge production in its own right.

The challenges to Western social theory have pushed many social geographers to rethink their own theorization of the relationships between race and space. The examination of race in geography has been further inspired by black feminism, particularly the work of bell hooks and (the novels of) Toni Morrison, both of whom build on the notion that the binary logics of black and white break down when one examines the constitutive nature of identity (McKittrick 2006). Put another way, social geographers should theorize racialized spaces as productive of social relations of difference, while also calling into question the project of racism as a partial one that fails to invest space fully with a totalizing narrative. Social geographic theories inform and are informed by theories developing from antiracist and queer philosophies, as geographers ask questions about the relationship between racial formations and segregation, as well as from other forms of difference around, for example, sexuality and space, in ways that take up the question of space as a social relation of power.

Social geographies in the twenty-first century

The development of social geography during the early twenty-first century has taken place along a multiplicity of trajectories informed by the histories discussed so far. Today, social geography is a theoretically and methodologically pluralistic subdiscipline that draws inspiration from spatial science and applied geography while focusing on the social geographies of inequality and building on the critical cultural revolutions that have demanded social geographic research focus on the constitution of radical difference. To offer an approximation of what social geography is today, it is important to recognize both its pluralism and its contextualization and also to appreciate how certain practices serve to delimit the subdiscipline, cutting off areas of difference (Peake 2011). The challenge of thinking about social geography this way is that it highlights that geography has its own colonial roots and has served as a vehicle through which empire has been justified. There is, however, a radical tradition to social geography that should not be underestimated, and even as social geography remains vested in certain traditions (spatial science) or places (UK-based social geography), it is also a subdiscipline with a core focus on inequality and difference, including areas such as children’s geographies, geographies of aging, and geographies of religion. This range suggests that social geography can be a site of radical alterity, and provides geographers with the space to interrogate relations of power and authority.
Spacing inequalities

Social geography has remained invested in the question (initially adopted by welfare geography) of “who gets what, how, and where” (Smith 1994) through a body of scholarship interested in investigating both the distributions of spatial inequalities and the factors that limit accessibility to needed resources, be those health care, food, or green spaces, to name a few. Much of this research is applied, and looks at addressing sociospatial problems. Drawing from both the critical and spatial scientific traditions of social geography, questions related to access and accessibility to services, the relationship between crime and violence, gentrification and the politics of difference, health and wellbeing, and food deserts remain important. In these areas, social geography overlaps with other subdisciplines, including urban geography, population geography, and health geography. In the application of geographic methodologies to the study of social problems tied to inequality, social geographers have taken up questions that challenge societies to recognize the relationship between where one lives and what resources one has to manage that life.

The study of homelessness, a core concern for social geography, investigates the complex relationships between housing availability, economic conditions, social exclusionary politics, and other spatial relationships that affect accessibility to needed resources. Marginalized from everyday political practice, homeless populations are rarely considered in policy relating to urban revitalization projects, which gentrify and marginalize homeless peoples. Nor are homeless populations considered when environmental mitigation strategies designed to protect populations from natural hazards are constructed.

Social geographers are interested in questions regarding the relationship between homelessness and public transportation accessibility. Challenging the geographical perception that homeless people are spatially bounded in particular parts of the city, social geographic research suggests that homeless peoples have a sophisticated pattern of mobility as they seek out services and opportunities. Neither is it only physical accessibility to public transportation that is a critical component for understanding how and in what ways homeless people move through the city. Discrimination and stigmatization are aggravating factors that limit the capacity of homeless people to take advantage of various transportation networks.

The notion that the sociospatial processes of inequality mediate the everyday landscapes of social services is an important point of analysis in social geographic research on inequality. This work is not only limited to traditional areas of social geographic inquiry, such as health care, but addresses new areas of inquiry, including Internet accessibility. Social geographers have long been interested in questions of “digital differentiation,” and the relationship between Internet resource accessibility and social inequality. In food desert research, social geographers map and analyze the relationship between high-quality food resources and the social organization of urban and rural space. In some places, which have seen declines in broadband speed availability, supermarkets serving diverse foods are also limited. Social geographers are particularly interested in interrogating how food deserts emerge out of the structures of inequality that are produced through racism and classism. This tradition is paralleled by the work of other social geographers that call into question how inequalities are tied to health disparities between different populations. This line of research is particularly pressing for the spread of both infectious diseases such as Ebola, HIV/AIDS, polio, and dengue, and chronic diseases such as diabetes, heart disease, and asthma. Social vulnerability
Social geography relies heavily on social geographic approaches to investigate the potentialities of population health, for example, in a wide range of contexts from natural disasters (e.g., the relationship between race and flooding in the wake of Hurricane Katrina in New Orleans) to the spread of HIV among commercial sex workers in Southeast Asia.

Social justice

With a focus on inequality and difference, it is not surprising that social geographers have also asked questions about social justice and equity. Articulating the contradictions inherent in the spatial logics of social differentiation, social geographers have pursued more socially just geographies through their scholarship and activism. David Smith's *Geography and Social Justice* (1994) mounts a case for the role that geography must play in building societies based on the principles of justice. Smith's analysis is quite complex, but at its core rests the notion that “People who have lost their place, for one reason or another, must be provided with or find another. There is no question about it. People need it. They just do” (1994, 152). A socially just geography must take into consideration not just the material elements of place but also the politics of inclusion and equity. This principle situates a much larger tradition in social geography, which takes up social justice by providing alternative models and theories for how to reorganize space to be more socially just. Smith is not the only geographer to take up this question, of course, as many have pursued the theoretical and practical implications of a socially just geography in the context of concerns about ethics, morality, economy, and politics.

Social geographers' interest in questions of social justice has been applied to many subjects, including rights to the city, rights to education, rights for green space, rights of transportation accessibility, and rights of equity regardless of disability, sexuality, gender, and race/ethnicity. Undergirding geographies of social justice is the question of how to constitute spaces that provide for difference and allocate resources in ways that are not based on the economic principles of capital. Such research is inherently activist in nature, as social geographers become aligned with struggles that push against the tendency to treat space as a neutral backdrop to social relations. Put more directly, social geographers have argued that quality housing, safe spaces, and resource accessibility, for example, must be mapped onto a broader ethics of care and hospitality. This raises ethical questions about how to create spaces of inclusion that match an ethical concern for a multisituated subject that is not based on the individual responsibility to care for one's self but on the sociospatial responsibility to build just and equitable societies. A social geography of activism is thus concerned with questions such as the right to public space and how we reimagine the spatial organization of societies to parallel our ethical and moral concerns for social justice. This is quite challenging, as public space is often constituted as a space of fear and thus as a site of necessary state control. What can and cannot happen in public space is a reflection of a much larger set of sociopolitical processes that social geographers have long been interested in interrogating. From the ways in which certain sexual minorities are marginalized through the practice of limiting spaces of social connectivity to the policies and practices that limit public speech and gathering, social geographers are concerned with how rights are limited by the spatial practices of state and corporate actors.

Methodologically, social geographers have taken up the question of social justice not only through the study of injustice but also through their own practices. Participatory action
research (PAR), for example, is a methodological approach that de-centers the researcher and situates the research process as a collaborative engagement, which includes the so-called research subject or research informant. As Mrs. C. Kinpaisby-Hill suggests, “spaces normally reserved for data collection are turned into arenas for learning and action, and the ethic that respondents should be left unaltered is replaced with an ethic that research can directly challenge injustice and facilitate social change” (2011, 220). PAR is not only a challenge to the question of what makes a socially just geography, but also to the very practice of social geographers, who tend to focus on the research process as much less dialogic. These practices also challenge the assumptions of institutions, which tend to privilege objectivist science and the publication of scholarly research in a narrowly defined set of academic journals. Thinking about research as a different sort of process, one in which academics and activists engage in challenging inequality through the research process, constitutes a new agenda for social geography: “Instead, the agenda for the ‘communiversity’ envisages a two-way interaction between scholars and communities, where communities (local and global) may define and conduct geographical research with academics, but based on their own priorities” (Mrs. C. Kinpaisby-Hill 2011, 221).

Social movements

Social geographers have also focused attention on the emergence of social movements, which are best understood as mobilizations of political action that challenge the hegemony of sociocultural and political-economic elites. Geographically, social movements not only take place somewhere, they are also always about the rights to space. From the Black Consciousness Movement, which began in South Africa in the 1960s to challenge apartheid, to the Chipko Movement in India from the 1970s, which gave rise to the practice of tree hugging to stop logging, to the protests in Seattle against the World Trade Organization in 1999, to the Arab Spring that began in 2010, to the Occupy Wall Street movement of 2011, all social movements are struggles over space and the right to use space in ways that provide the opportunity to reallocate resources or promote difference. Social geographers are interested both in how social movements work and in how they operate spatially through the development of social networks that facilitate flows of ideas and practices. Social movements scholarship within social geography has tended to focus on urban movements, although social movements research more broadly draws from a wide range of theoretical and empirical examples, particularly from movements located outside North American and European locations. Social geographers also focus on what some call new social movements, which have built on the traditions of earlier movements that emphasized resources and struggles against capital, to develop arguments for broader rights, related to sexuality, gender, race, or ethnicity. Social geographers investigate not only how these networks emerge and function, often through encompassing numerous localities into wider networks, but also how they are mediated through different spaces. How social movements play out is always tied to the spatial politics of the movement itself. Social movements scholarship also calls into question not only how movements are organized across and through different spatialities but also how they are organized in relation to spatial imaginaries. The challenge for social movements is tied to the politics that undergird them, which insists, at some level, that the movement articulate its goals and therefore its constituencies. This demand sometimes puts such movements into...
conflict with themselves, as a movement in one place might have to manage multiple goals and objectives. Routledge, Cumbers, and Nativel (2007) highlight these tensions in their analysis of global social movements, which are trying to produce horizontal connectivities that foster autonomous possibilities in various localities. At the same time, these movements often rely on “key organisers (those we term ‘imagineers’) and key events in producing the network” (Routledge, Cumbers, and Nativel 2007, 2577). These concerns point to the ways in which social movements function in relation to the complex networks of power that sometimes operate in and through them. What becomes clear through the study of social movements is that, like all sociospatial processes, social movements have the potential to sediment binary logics of inclusion and exclusion, insider and outsider, either through direct intention or through a reliance on a politics of verticality that limits the autonomous potential of various aspects of a movement. Barker (2012) powerfully argues that the Occupy Wall Street movement in North America excluded indigenous peoples through, on the one hand, a deployment of settler myths of indigenous peoples and their activism, and, on the other hand, a thorough dissociation of the movement from the long-term occupation of indigenous spaces. And yet, as Barker points out, “Indigenous peoples are willing to share the struggle, but only to the point that the ultimate goal of liberation from colonial power, expressed through state and capital, is achieved” (2012, 332).

Much social movements research focuses on left-leaning, liberal activism, eschewing the movements that are tied to right-leaning, conservative political agendas. And yet, such movements are powerful political forces within the modern-day landscape, as they create sociospatial networks that permeate everyday life. Carolyn Gallaher (2003) turns her attention to the emergence of the Patriot Movement in the United States to suggest that these movements have their basis in a complex set of identity and economic politics that should not be dismissed. Rather, through the study of such movements, it is possible to disentangle the complex geographical imaginations that animate such movements, with their interests in separatism and isolation, from others whom they see as degrading their own social, political, and economic positions.

Space, representation, and identity

Perhaps one of the biggest challenges to the integrity of social geography as a subdiscipline has been the alignment of the discipline with poststructuralism. Poststructuralism, at its simplest level, suggests that there is no predetermined structure (or hierarchy) to social relations or, for that matter, spatial relations. The earliest intersections of social geography and poststructuralist thought pushed the discipline toward the question of how, why, and in what ways space is represented. Put another way, geographers called into question some of its own core practices, including that of cartography, which was questioned for its authorial approach to representing the world. For social geographers, the question of what social life means spatially was reimagined to suggest that spatialities were only temporary formations enshrined through social practices and representations. To this end, social geographers turned heavily to textual analysis that they began to apply to a myriad number of geographic objects, including maps, landscapes, and film, to name a few. Thought this way, space as representation is also material in that it mediates the flows of people, ideas, and objects. Del Casino and Hanna (2000) suggest, for example, that spatial representations, such as maps, are possible only because of the performances that
people bring to them. Suggesting that map and space are not separate – one representing the other – Del Casino and Hanna argue for a theorization of

a map space [that] is not bound by the margins of the paper on which it is printed, but is inscribed with meaning through its intertextual linkages with other texts and spaces. In addition, map spaces are sites through which we can examine the processes of identity construction, and the historically and spatially contingent social relationships that constitute identity categories. (2000, 30)

This move, toward a much more performative sense of space and of identity, followed the work of scholars who suggested that categories of identity – gender, race, sexuality, class – were not embedded in a priori structures but were made real only through repeated performances of them. Identity thought this way is also a performance that demands certain practices to make real its meanings and organizations.

Following the work of Michel Foucault, social geographers also began to theorize power as a central object of analysis, and scholars began to ask how certain forms of authority are enacted through the organization and construction of space. Disciplinary logics, what one should and should not do in various spaces, were theorized as projects of the state and of social elites, who seek to reinforce and naturalize their power through people’s own acquiescence to various spatial forms of authority. Social geographers began to investigate this internalization of power through scholarship that reimagines space as epistemological. Thought of as such, spatial imaginaries – how we view the world as organized through discreet or relational categories of identity, for example – become a way of mapping difference (see Jones 1998). Spaces do not exist prior to our ways of knowing and understanding them but are socially constructed through our relationships with them. Foucault’s work has profoundly impacted on the discipline, as the categories of politics and geopolitics have been called into question as foundational concepts. Instead, Crampton and Elden (2007, 7) argue that, for Foucault, “the art of government is less about geopolitics (territorial gain and retention) and more deeply geographical … [The] population thus had to be known in its spatial dispersion, giving rise not only to statistics but to a new form of cartography.” Picking up this thread of analysis, social geographers have pushed the discipline to question its inherent assumptions around territory as fixed and naturalized and also have called into question the relationship between social and spatial categories, such as citizen and noncitizen. Crossing over with questions relevant to political geography, then, social geographers have analyzed the concept of citizenship through the lens of poststructuralist philosophy by interrogating the grounds on which the boundaries of the “citizen” are marked by the everyday politics of class, race, sexuality, and other forms of identity.

This poststructuralist turn was largely a critique based in epistemology in the sense that knowledge of the world cannot exist prior to human engagement of it. Rather, knowledge produces the ways through which we understand space and identity. This is why the focus on representation and representational politics became such a powerful question for social geographers, some of whom reimagined the ontological basis of their arguments about sociospatial structures of capitalism and power. Power operates not through predetermined core structures, but through relationalities – including economic systems – that are always constitutive of the so-called other that power seeks to disable. Core concepts in the discipline, such as scale and sovereignty, have also been called into question for the ways in which they are deployed to
produce difference and violently to mark out spaces of inclusion and exclusion.

Gender and sexuality

Social geographers have long been interested in questions of the spatialities of gender and sexuality, most recently investigating the assumptions that all human beings experience the world as if they are of one gender or sexuality. Asking to what extent spaces are differently constituted in relation to gender, for example, social geographers also challenged the hegemony of the adult subject as a totalizing narrative of lived experience, leading to the development of the subfield of children’s geographies. Research on gender and sexuality draws from the work on theories of social reproduction and finds commonality in the examination of home spaces as well as spaces of caring. Social reproduction theory suggests that the spaces of waged work are not the only ones that support the expansion of capitalist social relations or patriarchy, for example. In interrogating home spaces as sites of unwaged domestic labor, which is vital to the reinforcement of certain sociocultural and political-economic processes, social geographers have offered up analyses that demonstrate the complex relations of home and work and opened up questions related to how one’s experience of the life course is contingent on the sociospatial processes of not only age but also race, gender, sexuality, and class. This work thus builds on feminist geography’s interest in exploring the everyday gendered experiences of waged and unwaged labor.

In this way, feminist geography has much more in common with social geography than, say, cultural geography, which has often focused on social power and authority instead of the positionalities of women, queer subjects, and people of color. Gender is taken up in social geography to include the analysis of gender as more than a binary of men and women, but as an ongoing sociospatial process of gendering and constituting femininities and masculinities. Indeed, the scholarship on gender and geography includes not only the study of women’s experiences but also the gendered dynamics of relations that inculcate spaces of femininity and masculinity in ways that are sedimented in relations of power. That said, gender is not a constant force across space but is intimately interwoven in space. Put another way, the constitution of what is normalized as appropriate gendered performance is developed through the social geographies in which those performances are being enacted. The violence that structures “normal” gendered performance is thus always a social geographic process, one that can be investigated through an analysis of everyday practice. In interesting ways, this approach has opened up social geography to the broad study of masculinities as well, as social geographers investigate the spatial formations that reify certain forms of masculinity and privilege those forms over others.

All of this work has forced social geographers to think about the body and body politics as an integral component of their social geographies. Historically, social geographers studied the body not as an object unto itself, but rather as an external marker of social identity that could be mapped in and through space. But the body is a complex space as well; it is constituted through relations of power and authority in ways similar to other spaces. And, yet, the study of bodies and bodily processes has not always been the subject of social geography. It is queer geographies and the study of sexualities and geography that have expanded the study of bodies and body politics to the area of sexual politics, opening up questions as to how bodies and other spaces are heteronormatively constituted and possibly queered through the practices of radical sexual
alterity. While *Geographies of Sexualities* (Brown, Lim, and Brown 2007) built on the questions offered through sexuality studies in the discipline and provided concrete empirical examinations of queer bodies and spaces, other social geographers have extended these arguments, calling into question the construction of queer urban and rural bodies, spaces, and networks. The literature on/in queer geographies also challenges the notion that queer space is simply an alternative to heterosexual space, suggesting that the study of sexualities is much more than simply mapping out the spaces of queer collectivity, but rather includes the interrogation of sexualities and sexual politics more generally, including processes of the heterosexualisation and queering of space. This work has also pointed to the complex relationship between sexualities and other forms of bodily difference and has noted that the reorganization of queer studies of the city, in particular, has moved away from queer urban spaces as empirical objects of analysis to the examination of embodied queer experience.

Race and racialization

The study of race and racialization is also a key inquiry in the field of social geography, although these categories of scholarship have been differentially taken up across the subdiscipline. The development of critical race studies in the discipline, in particular, calls into question the historical lack of race studies in human geography writ large. Thus, despite the existence of a number of leading scholars in the discipline, including Audrey Kobayashi (2012) and Ruthie Gilmore (2007), who have long interrogated not only the discipline’s problematic racial imaginary but also the complex relationships between race, space, and everyday lived experience, there remains a problematic relationship between social geography and the study of race. As Minelle Mahtani (2014, 365) recently suggested, “Simply acknowledging the whiteness of the discipline does nothing to eradicate it.” In this regard, critical race scholars in the discipline have struggled with both racial geographies as a subject of analysis and the disciplinary context in which those racial geographies take place. For critical race scholars in geography, it is not possible to separate out the two. Following on from Katherine McKittrick’s (2006) powerful argument, Mahtani goes on to aver that far too often black bodies are simplified, related to occupying inevitably black spaces, resulting in dire and dour geographical determinism (what would Semple say?), or else reduced to a solely corporeal identity based on the colour of their skin, reflecting an almost obsessive focus on phenotype and flesh, where body is understood as the only useful black geographical scale. (Mahtani 2014, 362)

The study of race suggests that there are new ways to imagine a social geography of difference, which operates not from the social construction of so-called black or white bodies (Mahtani 2014) but, rather, through a more robust analysis of the everyday material violence that marks out spaces of difference for people of color. This includes a critical interrogation of whiteness as well as the intersections of race, racism, and other forms of subjectivity, including the postcolonial. This is where social geographers draw from the rich array of critical race studies scholarship (e.g., black feminism) and postcolonial theory (e.g., subaltern studies) to interrogate both the representational and the material violences that are associated with the processes of racialization, a historical-geographic process that is not produced merely through social constructions of categories, such as black, white, or brown, but through epistemic and material violences that are constituted through the deployment of racial difference.
SOCIAL GEOGRAPHY

One important argument that critical race scholars contribute to the work of social geography is the notion that representations are material; they do not simply circulate out there in the world but rather are inculcated in the real worlds of racialized bodies and spaces. This is an argument that Audrey Kobayashi (2012) makes in her recent review of the field of race studies in geography, suggesting that the concern within race studies is not whether representations have material effects, but rather that a postrepresentational world, which is suggested to be postracial, minimizes the ongoing racial violence (both material and representational) that marginalizes people of color. As Divya Tolia-Kelly (2010, 364) puts it: “Questions of race, identity and ‘other’ bodies fuel contemporary lived environments and world politics. When foregrounded, these questions can unravel a myriad of coordinates, yet to be visited.” These coordinates include the ways in which racialized processes and race as lived experience mediate the everyday lives of people of color. These concerns have been taken up in the study of immigration politics in the discipline, as social geographers investigate the relationships between racialization, for example, and anti-immigration politics. Helga Leitner (2012, 836), whose work on white residents’ relationships to new immigrant populations in the rural upper Midwest of the United States, suggests that “Cultural and racial differences were not only remarked on but evaluated as inferior.” Leitner also shows how whiteness is not a singular category but, rather, intimately related to class position. While almost all white participants in the study illustrated concern around immigration and degradation of white, rural (American) values, the ways in which racism appeared in their practice was differentiated along class lines, with upper middle-class professionals hiding their racialized politics in rhetoric reminiscent of postracial discourse. This suggests that the social geographies of race are complicated by the intersectionalities between race and other subject positions, including class, gender, ethnicity, and sexuality.

Posthuman social geographies

Social geography has changed and evolved over time to take up new questions and challenges, including ones related to a more-than-human social geography. What role, for example, do nonhuman subjects have in the study of social geographic processes? Is there a way to reimagine the nature/society binary in social geography in line with current theories of the nonhuman and questions of affect or emotion? Humans are, after all, embroiled in all sorts of relations with nonhuman subjects that have a real world impact on the organization of social geographies. The materiality of the urban environment, for example, has long been examined as a medium through which humans constitute their social worlds. But, the turn toward posthuman social geographies is more than simply a recognition of the nonhuman; it is about investigating the complex hybrid geographies through which humans and nonhumans interact. It is also about the affective moments in which humans co-create their worlds with, for example, animals, that social geography, as a purely human endeavor does not call into question. Indeed, the research in which human–animal relations are taken up provides an opportunity for social geographers to think about how humans are enrolled in a range of social relations through their interactions and affective relations with animals. Humans not only form relationships with their pets, but their pets call on them to form new relations with others – both human and nonhuman. Heidi Nast (2006, 304) takes up this question in her discussion of the evolution of “dominance–affection–love,” suggesting that
“post-industrial isolations and narcissisms have made pets into screens onto which all sorts of human needs, desires, and investments can be and are being projected, such projections [are] part and parcel with larger sociospatially uneven processes of wealth accumulation and investment.”

Recent theorizations of the posthuman in social geography, however, have extended the question of whether animals and other nonhuman subjects are reflections on which the human is presented, to ask how a more-than-human social geography is enacted through relationships between various subjects, both human and nonhuman. To this end, some social geographers have turned to actor-network theory (ANT) or nonrepresentational theory, two bodies of literature that, in different ways, ask how social geographies of the more-than-human are enacted through the enrollment of human subjects in different relations or forces. These philosophies, which argue for a rematerialization of social geography, are paralleled by the work of geographers interested in site ontology, which is best understood as a theoretical approach that “emphasizes the immanent, material connection between bodies and unfolding, situated practices” (Woodward, Jones, and Marston 2010, 273). Ian Graham Shaw (2012, 620) builds on this approach and picks up the philosophies of ANT and nonrepresentational theory, along with site ontologies, to suggest that “A world is a constellation of objects.” This constellation of objects is not evenly distributed but, rather, forms assemblages through geo-events. It is the unfolding possibilities that a more-than-human or posthuman social geography can interrogate that make this new line of inquiry particularly provocative for social geographers interested in how human sociospatialities are always already constituted through their broad range of relations with the nonhuman world. One way in which this inquiry is currently being investigated is through the concept of the Anthropocene.

Conclusion

Social geography is an incomplete project that has been built on through a wide range of theoretical and methodological approaches. Today’s social geography is itself an assemblage of multiple approaches and empirical subjects that stretch the expanse of the discipline. There are likely to be many future social geographies that pull on and push against the historical evolution of the subdiscipline. This suggests that it is difficult to clearly map a singular path for a set of traditions that is marked as much by its differences as it is by its conjunctures. What is clear, however, is that social geography as a subdiscipline will continue to mark out its intellectual trajectories that map relations of inequality and difference in provocative and important ways.

SEE ALSO: Aging; Class; Commodification of nature; Gender; Identity; Inequality; Intersectionality; Postcolonial geographies; Poststructuralism/poststructural geographies; Race and racism; Sexualities; Social justice; Social movements; Social reproduction; Spatiality

References


SOCIAL GEOGRAPHY


Further reading


Social justice

Jenna M. Loyd
University of Wisconsin–Milwaukee, USA

The domain of social justice theorizing and practice is social, political, and economic institutions and processes. These institutions and processes structure the production and distribution of resources and create the conditions for individual and collective action. Thus, social justice theorizing and claims-making address issues of distribution, democracy and power, and procedures for decision-making. Social justice theorists also consider how these institutions and processes reproduce group oppression and domination materially and discursively, thereby maintaining unjust constraints on self-development and self-determination. Movements for social justice, thus, necessarily challenge dominant cultural norms and knowledge hierarchies.

Social justice has a central and deeply contested place in contemporary political theory, political mobilization, and social movement organizing. Conservative critics such as Hayek contest the concept of social justice, arguing that only individual action can be judged as just or unjust and that state action to redistribute social goods will inevitably impede individual freedom. Among proponents of the concept, debates center on questions of universality and human nature, and the interrelations among different forms of power and injustice. Some argue that universal norms of justice and humanity can be defined. Decolonial, feminist, and antiracist critiques of the apparent universality of concepts such as justice, liberty, and humanity have led others to abandon the pursuit of a normative definition of human well-being; however, recognizing a world of differences and distances among humans has not meant the end of social justice, but rather its pursuit through a range of ethical and political registers.

Social justice, moreover, is spatialized. Over the past 30 years, the discipline has shown that social and environmental goods and dangers are unevenly distributed across space, how socio-economic processes produce such unevenness and can create oppressive and violent places and spatial relations, and theorized how just geographies might be created. Geographers have contended that theorizing the production of unjust and oppressive geographies is necessarily related to collective efforts to transform their constituent spaces and power relations. For these reasons, social justice is at the center of debates over the relationship between theory and practice, the relevance of the discipline, responsibility, and caring.

Intellectual and historical context

Social justice – as a domain of political theory and organizing – emerged in the late eighteenth and early nineteenth centuries as part of revolutionary upheavals in the Caribbean, Europe, and North America against slavery, land enclosures, and monarchial rule. Abolitionist, organized labor, and socialist movements challenged existing terms of freedom, rights, and humanity. The development of different conceptualizations of social justice from this period onwards has been closely tied to the formation of, and contests over, state social and welfare institutions. Liberal political theories of justice emphasizing
individual rights have supported a limited state while communitarian principles of solidarity have been used to advocate a greater role for the state in society, which in principle, better fosters individual wellbeing. In the twentieth and twenty-first century, mobilizations for freedom among oppressed social groups have again fundamentally expanded terms of rights, sovereignty, belonging, citizenship, and humanity.

Liberation movements, relevance, and unjust geographies

Geography’s involvement with social justice can be traced to the late nineteenth century. Anarchist geographers Peter Kropotkin and Élisée Reclus developed theories for dismantling hierarchies of power and organizing more egalitarian ways of living. Kropotkin wrote against the dominant terms of free market individualism and social Darwinism, and sought to provide a scientific basis for mutual aid. Reclus developed a social and ecological ethics premised on reciprocity between humanity and the Earth. Arizona Baber, founder of the Chicago Geographical Society, was committed to anti-imperialism, environmental conservation, and actively supported women’s suffrage in Puerto Rico.

Despite these important early contributions, it was not until the late 1960s and early 1970s that social justice gained a prominent place within geography. The context was one of worldwide anti-colonial struggles; movements for freedom from racial oppression and for self-determination; and antiwar, environmental, gender, and sexual liberation movements. Humanist, radical, and critical geographers criticized the field’s inability to explain pressing social and political crises such as racial segregation, poverty, and ecological destruction. Radical geographers further questioned the positivist paradigm undergirding geographic knowledge production and the discipline’s tacit support for colonialism and international hierarchies. They argued instead that scholars should research and explain the roots of social problems and actively work to change them.

Marxist theory increasingly became central to many of the theoretical developments in radical geography, as evident in Antipode: A Journal of Radical Geography (established in 1969) and Harvey’s (1973) Social Justice and the City. Harvey built on Rawls’s (1971) A Theory of Justice, which remains a central text in contemporary debates over social justice. Rawls’s two-fold theory advances a universal principle of equal rights to basic goods and liberties and a “difference principle” whereby social and economic inequalities should be arranged to benefit the least advantaged. Harvey theorized “territorial social justice” as “a just distribution justly arrived at” (1973, 116–117), a conceptualization that remained closely tied to distributive models of social justice.

Radical and critical geographers in the 1970s also drew on and developed diverse threads of anarchism, feminism, and critical social theory. Theorizing domination and oppressive social relations was a main thread in this work. Blaut argued that the white, European claim to universal knowledge was better understood as an “ethnoscience” that erases other forms of knowledge and undergirds the violence of Western imperialism. Feminist geographers examined gendered geographies of everyday life, gendered class relationships, the gendered and capitalist division of public and private space, and gendered inequities in the discipline, launching various feminist groups within the professional organizations of geography, such as the Geographical Perspectives on Women (GPOW) group established in 1979 in the Association of American Geographers. Richard Peet, in the early pages of Antipode, proposed that Kropotkin’s
anarcho-communism should form the basis of radical geographic thought, and Myrna Breitbart moreover organized an issue of the journal dedicated to anarchism and the environment.

Social justice and difference

The dominance of modernist, Western teleologies of history and theorizations of society was undermined from organized challenges to structural racism, colonialism, and heterosexism. Into the mid-1980s to 1990s, these intellectual currents overlapped, and frequently were conflated, with developments of postmodern and poststructural theory. Some on the left dismissed new social movements for being concerned merely with superficial “identity politics” and not with broader social justice. For some, postmodern and poststructural emphases on fluidity, multiplicity, and social constructionism undermined the possibility for commonality or shared purpose. For others, multiple identities and identifications opened possibilities for coalition-building rather than assuming a common basis of interest (or community or nation), which easily could elide difference and power differentials within social movements. Antiracist, feminist, and queer theories sometimes engaged directly with each other, and other times not, but all would pose challenges to the whiteness, heteronormativity, and masculine dominance of the discipline.

Philosopher Iris Marion Young’s work pushed the focus of social justice beyond that of distribution to one that starts from the “concepts of domination and oppression” – exploitation, marginalization, powerlessness, cultural imperialism, and systematic violence – through which inequitable group differences are produced and maintained (Young 1990, 16). Young has been influential in human geography, in part because she insists on understanding social justice in specific historical and social contexts. Harvey engaged Young’s work to reconceptualize social justice as the “just production of just geographical differences” (Harvey 1996, 5).

But how has difference been theorized? Debates in geography in the 1980s over gender and class tended to ignore racism, and into the 1990s, studies of gender and sexuality also tended to analyze these categories separately from one another. Such debates often were tied with reconceptualizations of the “subject” of history (in Marxist terms: the proletariat, peasantry, and lumpenproletariat). There was, however, a growing appreciation for the co-construction of categories of difference. Concepts of intersectionality indebted to women of color feminists, the idea of articulation attributed to Stuart Hall, and poststructuralist ideas of assemblage of complex systems provided different ways of understanding the co-construction of difference. Together, they showed how the complex construction of identities and subjectivities is material and tied to political-economic processes. Race, for example, would come to be understood as specifically constructed through ideas of femininity, masculinity, and sexuality.

Far from being immaterial, unreal, or secondary to fundamental social change (commonly read as class and class struggle), theorists of race, gender, and sexuality have demonstrated how such relations of power and difference are intertwined or articulate with political-economic structures and processes. Space, not just identity and subjectivity, is also produced through multiple, interacting social relations. Differently positioned groups thus contest (and reproduce) dominant geographies in distinct, if interrelated, ways. The political implication is that conceptualizations of social justice that ignore how race, gender, or sexuality constitute class or citizenship will inevitably reproduce social and spatial injustices. This understanding changes
SOCIAL JUSTICE

both the subjects and spaces of social change. For example, feminist interventions into debates over capitalism showed how social reproduction is necessary work for capitalist accumulation, and therefore, identified spaces of the home, education, and collective care as fundamental sites of political emancipation.

Antiracist, feminist, postcolonial, and queer theorists have vividly demonstrated how challenging dominant epistemologies is an imperative aspect of liberatory struggles. Such challenges involve both deconstruction and ideological critique and the creation of transformative geographic imaginations and forms of political praxis through subaltern, queer, and insurgent epistemologies. The field of black geographies, for example, has been established through efforts to undermine the symbolic and structural violence of dominant white epistemologies to offer a fundamentally different way of thinking about freedom, space, and what it means to be human. Decolonization has also emerged as an important theoretical strand and practice that seeks to undermine ongoing colonial practices and epistemologies. Strategies of decolonial practice involve recognition and use of Indigenous epistememes as a way of unsettling Eurocentric claims to knowledge and sovereignty.

Postcolonial and feminist theories in particular have emphasized the principle that social change must be made by oppressed people with strategies informed by their experiences and analyses. Ruth Wilson Gilmore argues that “centering attention on those most vulnerable to the fatal couplings of power and difference signified by racism … will [enable geographers to] develop richer analyses of how it is that radical activism might most productively exploit crisis for liberatory ends” (2002, 22). This principle has significant implications for thinking about coalition-building, differential positions of power, and representation. Spivak, for example, argued that while categories of difference are socially constructed and not natural, sometimes it was necessary to deploy a strategic essentialism in the course of making political claims. This idea has been widely debated, perhaps most forcefully by women of color feminists who contend that such moves enforce rather than transform hierarchies within oppressed groups.

Environmental justice and the production of nature

A key contribution to thinking about social justice has come through theorizing not merely uneven distribution of goods across space, but the production of space. Neil Smith’s Uneven Development (2008/1984) offers a Marxist theorization of the production of nature wherein the capitalist production of space is a dialectical process that equalizes and differentiates space, creating unevenly developed geographies across spatial scales. The “abolition of uneven development,” he argues, “is one of the first conditions of communal life” (2008/1984, 204).

Marxist theorizations of human–nature relations were central to the development of explicit links between environmental and social justice concerns. The field of political ecology emerged in the 1980s and brought Marxist (and later poststructuralist and postcolonial) theories to questions of nature, society, and change. The field has focused on the historical, social, and institutional relationships shaping the environment and development projects primarily in the Global South. Rather than attributing issues such as environmental hazards, famine, or land degradation to premodern or uninformed practices, founding studies in the field showed how local agricultural and resource extraction practices are tied to specific histories of colonialism, state development projects, and the international
political economy. Feminist political ecology emerged in the 1990s to theorize gendered relations of knowledge production, development practices, and decision-making; and to advance feminist practices and strategies for change.

Environmental justice centrally concerns the distribution of societal harms across space, procedural justice, difference, and recognition. The term environmental justice emerged in the late 1980s in the United States out of a nationally networked movement to challenge environmental racism. The term, first defined in a 1987 United Church of Christ report, refers to the disproportionate exposure of nonwhite people to toxins and pollution (Pulido 2000). The centrality of questions of racism for the US movement for environmental justice has generated debates over the interrelations between race and class, the role of state regulation, and epistemology. Dominant understandings of racism tend to frame the distribution of environmental harms as the result of discrete acts. Yet such liberal individualist paradigms ignore histories of colonialism and racism that have produced uneven geographies of harm. Thus, Laura Pulido’s intervention reframes questions of distribution into the uneven production of goods and harms by conceptualizing racism “as a sociospatial relation” that links together “industrial zones, affluent suburbs, working-class suburbs, and downtown areas, all of which are racialized” spaces (2000, 13).

“Think globally, act locally”: debates over the scale and scope of justice

Anticolonial movements and the rapid restructuring of the global economy and nation-state provide the context for theorizing the scope and scale of justice. The politics of scale involves attention to how different kinds of political-economic spaces are produced and unevenly related, and the discursive strategies whereby different scales are constructed as “higher” or more powerful. Existing nation-state structures and forms of political affiliation (e.g., citizenship) often do not provide the means for political action. The traditional confinement of politics to the national scale or to formal citizenship in turn has led to a proliferation of political action, particularly since the late twentieth century, including the growth of the “global justice” movements and subaltern transnational political networks. It also has fueled theorizing on how to understand the relations between people and different places, including efforts to redefine terms of solidarity, responsibility, citizenship and political affiliation, and sovereignty at a range of geographic scales.

Harvey’s reconceptualization of international solidarity offers that “theoretical practice must be constructed as a continuous dialectic between the militant particularism of lived lives and a struggle to achieve sufficient critical distance and detachment to formulate global ambitions” (1996, 44, emphasis in original). Feminist theorists in particular have contended that this conceptualization of lived as local and global as abstract reproduces discourses of globalization as all-powerful and localities as weak. Doreen Massey, for example, contends that the global is also concrete. She theorizes places as relational constellations of power relations that actively reconstitute “localities” as well as “the global.” While different social groups and individuals are differentially situated in relation to socioeconomic processes, conceptualizing space as relational and continually being made opens the possibility for “a more configurational and outwardlooking stance” toward political responsibility (2006, 94).

From the early 1990s to mid-2000s, geographers engaged in a robust series of discussions over theorizing geographic scale. A principal
SOCIAL JUSTICE

question concerned how situated political actors organizing in one kind of place (e.g., in the household, in the city) can link their issues both to other places and across scales, and thereby foster avenues for broader political mobilization and policy change. Feminist interventions have been fundamental to the so-called scale debate, particularly by demonstrating how attention to the politics of the body and social reproduction transforms understandings of justice and politics. Geographers drawing on poststructuralist theory, however, have contended that vertical imaginations of scale may limit progressive political organizing and have turned to flat ontologies, horizontality, rhizomatic networks, and assemblages as alternative conceptualizations for complex political terrains.

The urban question, spatial justice, and right to the city

Lefebvre’s work on everyday life in capitalist society and the production of space offered a conceptualization of how class relations are built into cities, and in turn how class struggle is also about city life itself. The concept of spatial justice is a dialectical one wherein spatiality is understood as the contradictory outcome and object of social struggle. Thus, spatial analyses of injustice can link the economic and social conditions of different groups to interventions in the exploitative and oppressive production of space. This resolutely practice-oriented conceptualization of spatial justice has been developed most explicitly since the early 2000s, at the same time as the flourishing of political organizing (e.g., Shack/Slum Dwellers International and the US-based Right to the City Alliance) and theorizing under the banner of right to the city.

Principles of communitarianism, direct democracy, and deliberative communication converge on urban public space, as an ideal and material space. For Young, “because city life is a being together of strangers, diverse and overlapping neighbors, social justice cannot issue from the institution of an Enlightenment universal public” (1990, 240). Rather than ignoring group power differentials, she instead explores the “implications of a politics of difference by envisioning an ideal of city life as a being together of strangers in openness to group differences” (1990, 256). Marxist urban geographers have explored how in practice bourgeois private property relations underscore the constitution of public space as a site for public political debate. Feminist geographers, moreover, have illustrated how access to and the creation of public space are gendered. Feminist, queer, and anarchist scholarship has also shown how relations of power and difference intersect to structure public and private space, with state and state-sanctioned violence frequently deployed to create and sustain particular inclusions and exclusions. In drawing attention to these interrelations and to systemic violence, this scholarship works to theorize new possible forms of politics as alternatives to representative and deliberative democracy.

Questions of normativity and postfoundationalism

Debates over scale and difference have dovetailed with debates since the 2000s over the necessity and desirability of normative thinking and the development of postfoundational concepts of justice. These concepts have been elaborated in terms of ethics, morals, and emotion. In the wake of postmodern, postcolonial, and feminist critiques of universal truth claims and a common humanity, normative reasoning was also deeply in question. Yet given the widely shared view
that all knowledge construction is partial and political, many geographers with a social justice commitment have advocated that researchers make explicit the terms of their normative thinking. Among the sources for normative ethics are religion, animal–human ethics, and the capabilities approach to moral philosophy developed by Amartya Sen and Martha Nussbaum.

In contrast, Barnett argues that foundational definitions of values or justice do not necessarily provide the means for judging a situation as just or unjust. He suggests that analyses begin “not so much from a clear-sighted definition of justice but from widely shared intuitions of injustice” (2010, 248, emphasis in original). He further suggests that theorizing and political mobilization around concepts like right to the city, spatial justice, and environmental justice exemplify forms of practical reasoning through which “normative concepts of justice are worldly, emerging from situated conflicts” (Barnett 2010, 248). Geographers have drawn on postcolonial theory to develop non-Eurocentric ethics and alternative geographical imaginations of connection and responsibility. Barnett also observes that Massey’s conceptualization of geographies of responsibility resonates with Young’s idea of shared responsibility that is “distributed across complex networks of causality and agency” (Barnett 2010, 252).

Attention to emotion, affect, and performativity has opened another avenue through which geographers are theorizing ethics. For some, the embodiment of encounters with other people suggests an ethics that emerges from conviviality and solidarity, whereas for others such encounters must still provide the means for challenging structural inequities. Ethics of care also may draw on anarchist praxis. It is feminist geographies of care though that have been pivotal to this ethical project, beginning as Lawson writes, “from the centrality of care work and care relations to our lives and societies. Care ethics begins with a social ontology of connection: foregrounding social relationships of mutuality and trust (rather than dependence)” (2007, 3). While affect has been identified as an important element of social justice organizing, feminist scholars have warned about unjust circuits of power that may be reproduced through appeals to not only fear, but also outrage and empathy.

Questions of practice and relevance over time

Geography, like other disciplines, has long debated the relevance of its scholarship to explaining pressing social and environmental issues. Concerns of relevance in the discipline have revolved around its influence in political and scholarly debates, applicability in policy or practical application, and teaching. Questions of relevance are tied to a vibrant debate about the roles of scholars in social change, including as public intellectuals or scholar-activists. Some maintain a sharp distinction between the academy and social change, arguing that the best way for academics to engage in social change is through their scholarship and teaching. Others counter that academic efforts must be tightly tied and accountable to groups engaging in activism and political movement building. While these positions hold different understandings of the spaces of knowledge production, ultimately, these debates suggest that being a teacher and researcher entails active deliberation over ethics and ethical practice.

Practical examples of efforts to grapple with these issues are numerous. The Detroit Geographical Expedition and Institute was an experimental research and activist project co-founded in 1968 by geographer Bill Bunge and Gwendolyn Warren, an African American community
leader. It was intended to study and intervene in race and class injustice in urban space and continues to inform scholar activism and participatory research. Feminist, antiracist, and postcolonial movements and theory have developed transformative research methods and programs and have emphasized the values of self-reflexivity and obligations for ethical research and political practice. To name just one such method, participatory action research (PAR) has emerged out of efforts to shift power to (or build power among) marginalized groups, in part through creating more space for these groups to reshape research agendas and dominant representations as part of their mobilization for social change. Other efforts to foster just academic practice in the past two decades have included challenges to whiteness and heteronormativity in the discipline, organizing to prevent cuts to education funding and improve working and learning conditions, and campaigns to end the discipline’s active and complicit involvement in military activities. Most recently, mental health and the academy has entered onto the agenda of geographers.

Current emphases and future areas of research and practice

Conceptualizations of social justice have changed significantly from the focus on distribution in the early 1970s to include processes of group domination and multivalent strategies for social change, which both include and reject a role for the state. It is perhaps ironic that as conceptualizations of social justice have broadened, there has been a steady political effort to minimize the state’s role in classic (re)distributional concerns. The reconfiguration of the state role in social provisioning and regulation of the economy in turn have deepened material inequities at multiple geographic scales around the world. These reconfigurations, however, have been the grounds through which social justice continues to be reconceptualized and vibrantly lived.

Theorizing and transforming systemic relations of violence remains a pressing issue in geography, albeit addressed more among racialized geographers than the white majority. Harold Rose’s 1978 presidential address to the Association of American Geographers called on the discipline to pay closer attention to black interpersonal violence in US cities. He challenged (still frequent) explanations positing a black subculture of violence, and proposed instead that geographers investigate the interrelationships among growing economic despair, homicide, and stressful urban environments. Katherine McKittrick argues that ending racial violence will entail challenging modes of research and representation that “reproduce the axiomatic frame of survival, wherein the suffering body and the dying have always been the racial Other to the white Western-liberated human norm” (2011, 954). Responding to this challenge invites reflexivity about (who engages in certain spaces of) geographic knowledge production, and more so, imaginations for creating different ethical spaces of encounter.

The practical necessity of transforming systemic and interlocking relations of violence also employs the energies of geographers. The protest chant, “No justice, no peace,” which is commonly heard at US political demonstrations, is an example of practical social justice theorizing that deconstructs how dominant definitions of violence and state institutions of justice obscure the everyday workings of racial and class oppression. Challenging dominant epistemologies of violence and state security practices will continue to be part of efforts to create justice with peace. Wars, imperial occupations, and state-centric definitions and practices of security not only exert direct violence on people’s lives,
but also have enduring and wide-ranging effects, including on land use, political institutions, and trauma. Grassroots efforts to end violence and heal from interpersonal and intergenerational trauma are places where much is to be learned (and supported) about building a socially just world. Here, feminist and postcolonial attention to the symbolic violence of the gaze, appropriation of suffering, and the circulation of affect are deeply relevant. Moreover, feminist and queer interventions in social reproduction and care ethics will continue to generate possibilities for living otherwise.

Finally, global climate change, and its uneven ecological and social effects, represents a pressing issue that is fueling a movement for climate justice and is generating new policy, geopolitical, and theoretical terrains. The politics of disaster, emergency preparedness, and “resilience” are areas where environmental, social, and political-economic forces converge. Thus, bridging fields of knowledge will be imperative to developing just and ethical responses in the era of the Anthropocene. While global climate change is urgent, the danger of a politics of crisis is that it will reproduce existing hierarchies and forms of exclusion. This danger makes efforts to decolonize knowledge about environmental change and humanity all the more important.

SEE ALSO: Anarchist geography; Critical geography; Difference; Environmental (in)justice; Feminist geography; Peace; Poverty; Queer geographies; Race and racism; Radical geography; Sexualities; Violence

References


Further reading


Social movements

Byron Miller
University of Calgary, Canada

Social movements are typically a contentious form of social and political resistance based in collective action. Collective action necessarily requires that people work together in a cooperative fashion at the same time that they express opposition to other groups and forces. While considerable debate has centered on the question of whether “the political” is primarily antagonistic or associative, social movements are clearly both.

At the heart of traditional social movement theory is the free-rider question: why would rational, self-interested individuals participate in collective action when each individual’s impact on large-scale collective action is negligible and the benefits of collective action are public and free? Rational choice theory suggests that collective action cannot occur without “selective incentives,” that is, individual rewards for participation, yet clearly people come together to act collectively without such incentives. They do so because they form social bonds that cannot be reduced to individual self-interest. Accordingly, their motivations cannot be so reduced. Social bonds can take a variety of forms: shared social identity, affective bonds with others, attachments to place or region, or shared understandings and values. How social bonds come to be formed in the mobilization of social movements, as well as the strategies social movements employ in contention with other political actors, are fundamentally geographical questions.

Social bonds in place and region

The formation of social bonds of one sort or another is a prerequisite for all social movements. Social bonds are commonly, although not exclusively, formed through everyday lived experience in place-based social interaction. Pred (1981), drawing on the work of Hägerstrand, argues that social interaction and opportunities for bonding are shaped by daily time–space “paths and projects” that are usually structured by organizations and institutions. As people go about their daily time–space routines they participate in an internal–external dialectic in which experience and intention influence corporeal action and vice versa. Pred’s work points to the critical importance of place in the amassing of “mental (or internal) experiences that are fundamental to the shaping of … values, perceptions, attitudes, capabilities, preferences, and conscious or subconscious motivations and hence … goals and intentions (only some of which will be realized)” (Pred 1981, 242–243; italics in original). Such experiences, values, and understandings form the basis from which actors, including social movement actors, make choices.

Place, of course, cannot be reduced to the sum total of individual time–space paths. In Agnew’s (1987, 28) multidimensional definition, place consists of:

- locale, the settings in which social relations are constituted (these can be informal or institutional);
- location, the geographical area encompassing the settings for social interaction as defined by social and economic processes operating at a wider scale; and
- sense of place, the local “structure of feeling.” Or, by way of example, home, work, school, church and so
on form nodes around which human activities circulate and which in toto can create a sense of place, both geographically and socially. Place, therefore, refers to discrete if “elastic” areas in which settings for the constitution of social relations are located and with which people can identify. The “paths” and “projects” of everyday life, to use the language of … time geography, provide the practical “glue” for place in these three senses.

Place, in other words, must be understood as constructed by actors as they go about their diverse activities, and by diverse institutions, including the state. The “elasticity” of place is critical to its constitution. The social relations of place may be geographically extensive, even global. Such extensive relations mean that place is not necessarily inward-looking and exclusionary. On the contrary, as Massey has argued, place can be outward-looking and inclusionary – qualities that may be critical to the mobilization of geographically extensive social movements. At the same time, through recurring activity patterns and discourses, regional identities, conceived as bounded regions in our imaginations, are forged. These identities do not exist independently of the state and indeed may be reinforced and promoted by institutions of the state. Moreover, multiple and differently scaled regional identities commonly emerge, creating a rich palette of geographical identities and imaginations that social movements can draw from when trying to construct a sense of “we-ness” and solidarity necessary for mobilization.

Cities as a space of mobilization

While social movements can and do arise in any context where grievances combine with social interaction and collective consciousness, cities are especially powerful incubators of movement mobilization (Nicholls 2009; Miller and Nicholls 2013). The importance of cities to social movement mobilization has to do with much more than sheer population concentration. Cities are both territorial structures with characteristics specific to them that may give rise to place-specific grievances and relational places of interaction that facilitate not only the dynamic construction of identities and discourses, but also the sharing of information and resources.

In Castells’s (1983) classic analysis, “urban social movements” are seen to stem from three distinctly urban dynamics: (i) conflicts around the city as use-value rather than exchange value; (ii) conflicts around neighborhood identity, cultural autonomy, and communication; and (iii) conflicts around territorial (neighborhood) self-management. These conflicts are experienced at the urban or neighborhood scale and in large measure have their origins in place-specific structural and political arrangements. Such “collective consumption” movements are largely defensive, seeking to defend or reclaim urban “living space” from processes of commodification and state control. They represent reactive utopias that strive for place-based solutions to problems that are increasingly driven by global processes – what Castells terms the “global space of flows.” The geographical specificity of urban social movements raises critical questions of both the origins of the grievances they address, that is, do the problems urban movements address actually originate from processes operating at an urban scale and is urban-scale mobilization the most effective way to address their grievances? These questions lie at the core of Raymond Williams’s notion of “militant particularism” – later elaborated upon by David Harvey – which explores whether and how movements might shift from a focus on the deprivations and oppressions of everyday lived experience to a basis for, or articulation with, broader-scale systemic movements.
The tendency for most significant movements to emerge from urban contexts – which can range from large cities to more modest settlements – stems from the nature of social interaction in urban places. Cities are places where people concentrate in numbers sufficient to support the social institutions and practices necessary to produce the affective bonds, solidarity, and shared resources needed for social movement mobilization. Cities’ social diversity, moreover, provides a variety of norms, ideas, ways of life, and resources that, when mobilized, can foster diverse and creative strategies of resistance. The paths and projects that structure everyday social interaction, moreover, produce two critical types of social bonds or “ties”: strong ties of intense bonding and common identity, creating the sense of “we-ness” required for collective action; and weak ties of infrequent interaction and passing familiarity that nonetheless facilitate the sharing of information among different social groups (Granovetter 1983). Extensive weak-tie networks facilitate the sharing of information among groups and provide the basis for alliance-building – a necessity for most effective movements. Building broad-based and well-resourced movements inevitably requires the formation of alliances; opportunities to do so are, in the early stages of mobilization, most readily available in cities.

Geographies of systemic grievances

While many social movements appear to be specific to particular cities, regions, or countries, they often stem from broader systemic processes and state policies. As Peck has noted, cities are “where austerity bites,” but rarely does the imposition of austerity measures stem, in the first instance, from municipal policy agendas. Rather, uneven capitalist development creates uneven geographies of economic deprivation and wealth on urban, regional, and global scales, and state policies are applied to manage these circumstances. Since the 1970s, state policies have been increasingly neoliberal and emanate from scales “above” the cities and regions that suffer under them, including the global, for example, policies of the European Union (EU), International Monetary Fund (IMF), World Trade Organization (WTO), World Bank, and so on. Several of the most prominent movements of the early twenty-first century have opposed globalization and unfair trade, involving a variety of civil society and labor organizations from the Global South and North; austerity policies, as illustrated by Los Indignados/Podemos and Syriza; and inequality and increasing polarization associated with globalization and neoliberalism, most notably Occupy. The conditions that give rise to social movements are by no means limited to the economic realm, however. The Arab Spring focused on democracy and justice, although worsening economic circumstances and food shortages played a significant role in the movement. Immigrant rights and civil rights movements focus primarily on issues of recognition, albeit often with underlying economic objectives. The feminist movement focuses primarily on issues of recognition and voice; these of course have distributional implications as well. And environmental movements, such as the movement to halt climate change, are concerned with the integrity of nonhuman natural systems, often with an eye to the impacts degradation of these systems have on human wellbeing. In short, the systemic forces that give rise to the grievances that social movements express are diverse and complex, not only in the nature of the issues they address, but in their geographies. The geographies of systems, however, may be very different from the experiential geographies of those who suffer under these systems. In other
words, the lifeworld identities and values around which social movements mobilize have their own geographies which only partially overlap with the geographies of systems. Reconciling the two, that is, building movements that are sufficiently extensive to address systemic oppression and dysfunction, is “one of the most intransigent and paradoxical problems facing social movements” (Miller 2000, 67).

Framing and the geographies of mobilization

While economic systems, state systems, and ecosystems have material effects shaping the circumstances of everyday life in diverse lifeworlds, how those effects are conceived and experienced in everyday life cannot be “read off” their materiality. As Lefebvre has argued, and many others have subsequently elaborated, understanding the production of social space requires attention not only to its materiality but also to its representations and to lived, embodied experiences. Lefebvre’s trialectic of spatial practice (perceived space), representations of space (conceived space), and representational spaces (lived space) points to the importance not only of materiality but also of cultural and ideological frameworks and the lived, embodied experiences of material circumstances that are interpreted and understood through geographically specific lenses of culture and ideology. That material circumstances do not “speak for themselves” is illustrated through innumerable examples of hegemonic bargains that mute resistance, the discounting of future wellbeing in favor of a more comfortable and less contentious present, and ideologies of individualism and meritocracy that impede collective action. Cultural and ideological frameworks have their own geographies, varying by place and region. As Cresswell (1996, 8) explains, “space and place are used to structure a normative landscape – the way in which ideas about what is just, right, and appropriate are transmitted through space and place. Something may be appropriate here but not there.” These geographically variable normative landscapes have at least two consequences for social movement mobilization: (i) issue frames designed to effect mobilization in one place may be ineffective in others; (ii) strategies of protest and resistance that may be effective in one place may fail to be effective in others. Geographies of culture and ideology matter.

Given that places and regions “provide historically contingent but durable ‘schemes of perception’” (Cresswell 1996, 16), the development of place and region-specific “collective action frames” is a necessity for most social movements. Collective action frames are discourses that reference a variety of ideas, concepts, facts, norms, and ideologies around which collective action can be mobilized. They are critical to motivating individuals to act collectively, as well as to build broader alliances. Attention to place and region-specific normative landscapes and material circumstances are critical to the facilitation of collective action in those places and regions; in other places and regions where norms, ideologies, and material circumstances differ, substantially different collective action frames may be required. “Place-based collective action frames” (Martin 2003) perform three key functions: motivation, diagnosis, and prognosis. Motivational frames define the collectivity to be mobilized and values that spur it to action such as justice, rights, safety, equity, opportunity, and so on; diagnostic frames present analyses of the causes of the problems the collectivity faces; and prognostic frames lay out a course of action to be taken that may lead to a solution. Together, place-based collective action frames play a critical role in
defining grievances, moving individuals in particular places and regions to act collectively, and, in the best of circumstances, building alliances among groups that might not otherwise be allied. The creative process of collective action framing, moreover, is itself discursive and dynamic and not merely a response to a static cultural and ideological landscape. Norms, ideologies, and identities are not only drawn upon, but also shaped and defined in the process. In other words, processes of social movement mobilization are not merely reactive, but can also serve to produce new cultural and ideological landscapes.

The creation of effective place-based collective action frames, in the context of extensive strong-tie and weak-tie networks, is critical to social movement mobilization in “spaces of dependence.” Cox (1998, 2) defines spaces of dependence as the spaces of “those more or less localized social relations upon which we depend for the realization of essential interests and for which there are no substitutes elsewhere; they define place-specific conditions for our material wellbeing and our sense of significance. These spaces are inserted in broader sets of relationships of a more global character.” To secure the conditions of wellbeing in everyday life we must increasingly address the broader relationships of an ever more interconnected and global world, what Cox calls “spaces of engagement.”

Networked mobilization in a global world

The central dilemma of modern social movement mobilization is how to move from relationships built in place to engage relationships and power dynamics that may operate at much broader scales. The notion that the grievances that arise in localized realms may need to be addressed at broader scales is not new. More than a decade before Castells’s (1983) analysis of specifically urban social movements, Lefebvre (2003/1970) argued that the problems that confront urban residents can no longer be analyzed in terms of independent cities and urban agglomerations because urbanization processes had become global. Two years prior, when coining the term “the right to the city,” Lefebvre argued that the central problem facing society was the conflict between abstract space – based in the processes of commodification and bureaucratization – and social space – the space of use-values produced through the complex social interactions of everyday life. Noteworthy is the fact that Lefebvre did not conceptualize the problems of the city in discrete, bounded, territorial terms, but rather relationally, and in terms that suggest geographically extensive processes.

While Lefebvre’s relational understanding of social and political conflict presaged contemporary geographical approaches, the reconciliation of social movement mobilization with systemic processes remains problematic. While many movements strive to be global, they must be attuned to place-specific circumstances and identities. Analyses of social movements increasingly focus on networks as a means to understand how mobilization in diverse places might be joined up.

Network analysis has long played an important role in social movement research. Granovetter’s work during the 1980s on strong and weak ties represents one body of work among many examining the formation and significance of networks in localized contexts. What differs in the contemporary context is the role of information communication technologies, social media, and global travel, allowing networks to expand their reach dramatically, seemingly without geographical limits. The development of geographically extensive networks has facilitated dialogue and alliance-building on a global scale, allowing the
antiglobalization movement to challenge the forces of capitalist globalization wherever they operate (Routledge 2003; Featherstone 2008).

The building of extensive social movement networks is, however, fraught with tension. In an increasingly global world, identity construction has become increasingly complex. No longer are collective identities built primarily on the basis of frequent local interaction, although this remains one of several mechanisms of identity construction and social bonding, producing traditional constructions of tightly knit, place-based communities. In a world of global communication and global travel, cosmopolitan identities and bonds have become more common among global elite who live their lives in transnational circuits of interaction. While cosmopolitans’ experiences and perspectives may be grounded in diverse places around the globe, fostering tolerance and broad cultural competence, they may also be divorced from the places and experiences of the considerably less mobile and privileged. At the opposite end of the spectrum are individuals who have relatively little social interaction, producing very parochial and individualist identities. Significantly, virtually everyone in modern societies is now connected to electronic and print media, making discourses, cultural practices, and a variety of information from one’s own city, region, and country readily available. Access to, and production of, this information is often promoted by states, business interests, and cultural institutions, giving rise to place- and territorially based “imagined communities,” often without the production of the strong or even weak ties developed in traditional place-based communities. In the context of this increasingly complex landscape of social identities, contemporary movements attempt to mobilize and build alliances.

Contemporary social movements are increasingly focused on building networks (Routledge 2003; Featherstone 2008; Nicholls, Miller, and Beaumont 2013), recognizing the geographical diversity of material circumstances, interests, and identities that must be addressed in mobilization and alliance-building efforts. Central to contemporary mobilization efforts is recognition that networks are not free-floating spaces of open access, but are contextually embedded in places, regions, countries, institutions, and so on (Nicholls 2009; Nicholls, Miller, and Beaumont 2013). Networks are not only spaces of inclusion; they are also spaces of exclusion. They connect some, while leaving others disconnected. Indeed, the employment of spatial technologies of disconnection is common among states and other institutional actors that seek to undermine social movements (see Miller 2013).

Effective contemporary movements must not only actively strive to include diverse actors and groups that share the movement’s goals, movement activists must also recognize their own positionality and address power differentials within networks. Attempts to build broad inclusive movements are rarely simple or straightforward but are, rather, processes drawing on a multitude of actors, resources, and institutions. In the early stages, movements tend to grow through pre-existing relational ties as individuals, groups, and institutions reach out to potential allies. Such network-building efforts often involve “jumping” or “shifting” scales, as growing movements recognize the advantages of drawing from a broader resource base, assembling more voices, or working within state political opportunity structures that may be more favorable. Growing beyond networks of pre-existing relational ties is more difficult and typically involves the work of “brokers” or “imagineers” (Nicholls, Miller, and Beaumont 2013) who can help to build relationships among individuals, groups, and institutions that previously had no connection. While these new relationships may be extremely important, they may also be fragile.
and unstable because social integration and relationships of trust are weakly developed and the circumstances and positions from which the participants come may be extremely different. Under such conditions, encounter, discussion, and negotiation in “convergence spaces” become vitally important. According to Routledge (2003, 345; italics in original), convergence spaces: (i) “comprise diverse social movements that articulate collective visions, to generate sufficient common ground to generate a politics of solidarity, that is, multiscalar collective action”; (ii) “facilitate uneven processes of facilitation and interaction. The diverse groups and movements that converge in such spaces enact a practical politics consisting of at least five processes: communication, information sharing, solidarity, coordination and resource mobilization”; and (iii) “facilitate multiscalar political action by participant movements. Social movements engaged in grounded material struggles, and articulating place-specific concerns, also actively participate in forging a globalizing network of such struggles.” Convergence spaces can take a variety of forms: conferences, forums, joint campaigns, and ad hoc meetings. Critical is that they provide opportunities for discussion and the building of shared understandings, solidarity, and, in some cases, shared identity. Social movements, in other words, do not encounter the political landscape as fixed, but are themselves constitutive of it.

While convergence processes are an important component of geographically extensive movements that bring activists together from diverse places, regions, and circumstances, there is no guarantee they will actually converge on common objectives and strategies. Under such circumstances, multiscalar organizational structures and strategies are frequently employed, providing a common broad-scale framework and substantial autonomy and flexibility to pursue place, region, and country-specific strategies and campaigns. As Routledge, Cumbers, and Nativel argue (see Nicholls, Miller, and Beau-Dumont 2013), “multi-scalar networks of support and solidarity” recognize the diversity of place- and region-specific cultures, ideologies, and circumstances and facilitate broad coordination, providing an alternative to universalist strategies that may neglect geographical difference. Nonetheless, there is no guarantee that compatible strategies will be pursued across multiscalar networks. Indeed, many movements, such as the Right to the City movement, have been plagued by incompatible framings and objectives that have undermined a coherent understanding of the movement (Mayer 2009). Such problems are likely unavoidable, and a relatively small price to pay, for movements that seek to respect the social, cultural, and geographical diversity of the world in which we live.

**SEE ALSO:** Civil society; Class; Community; Cultural politics; Difference; Environmental movements and protest; Globalization; Identity; Inequality; Place; Political geography; Regional political movements; Representation; Scale; Social capital; State, the; Urban politics

**References**


Cresswell, Tim. 1996. *In Place/Out of Place*. Minneapolis: University of Minnesota Press.

SO-CIAL MOVEMENTS


Social reproduction

Cindi Katz
City University of New York, USA

Social reproduction encompasses the daily and long-term reproduction of the means of production, the labor power to make them work, and the social relations that hold them in place. Under capitalism, social reproduction entails the production and reproduction of the capitalist and laboring classes as such, the means and arenas of capital circulation, and the physical and discursive conditions to maintain the production and reproduction process over time and across space. It includes both the reproduction of capitalist social relations as a whole, along with the sphere of everyday life, which may appear outside of the relationship between waged labor and capital but is wholly entwined with it. In other words, social reproduction includes the production and reproduction of a differentiated labor force and the cultural forms and practices that at once maintain these differences and make them common sense. These distinctions may be constructed as sectoral or related to skill, education, and training, but they mobilize differences associated with gender, race, sexuality, and nation even as they produce them. The material social practices associated with social reproduction build upon, reproduce, and reshape inequalities that enable and enhance capital accumulation, racial capitalism, and patriarchy among other means of uneven development.

Social reproduction is at once the “fleshy, messy” stuff of everyday life and the structured practices that unfold in dialectical relation to production. These material social practices take place across geographic and temporal scales, encompassing daily and generational relations that are as much intimate as global in their effects and charge (Marston 2004). The form and content of social reproduction, as much as the means of its provision, are forged through struggle, making it a contradictory realm of both the continuation of capitalist social relations and the possibility of their transformation. Geographers have been attentive to geographical variations in what constitutes social reproduction (e.g., Cravey 2003; Laurie, Andolina, and Radcliffe 2004; Mullings 2009; Brickell and Yeoh 2014; Meehan and Strauss 2015; Kofman and Raghuram 2015), and marked the environment as part of the means of production, thereby recognizing environmental care and rehabilitation as part of social reproduction (Di Chiro 2008; Marks 2015).

Production and social reproduction are historically as well as geographically contingent. What is required of a labor force – and what is considered socially necessary labor – varies by time and place, as do the work and other resources required to reproduce the labor force (or an individual worker) on a daily, monthly, yearly, and generational basis. What may be required of a worker in New York City will differ from what is required for the daily reproduction of a worker in Guangzhou. Among other things, the cost of living will differ, the clothes one is expected to wear may be different, educational requirements will vary, and so on. What is considered necessary for the production of any given labor force is a social and political economic question, and changes with and alongside the shifting ground of capitalist social relations in any given time and place. This dialectical relationship between
reproduction and production is the outcome of both intra- and interclass struggle, and encompasses such things as education, health care, child care, and so-called benefits packages, or what is conceived of more broadly as the “social wage.” What constitutes the social wage is historically and geographically contingent, as is who is considered responsible for the provision of its components: the individual, the household, the state, capital, or the private institutions associated with “civil society.” Within these arenas the means of accomplishing social reproduction varies as well – in the household labor may be provided by family members or employees such as child minders, housekeepers, cooks, tutors, and the like, or through the purchase of such things as prepared meals, laundry services, child-care services, home health-care services, and so on. Individuals may do the work of self-care on their own or may avail themselves of therapists, sex workers, and personal groomers, among a growing array of services available for those who can afford it. The state may provide for social reproduction directly through public education, social housing, publicly funded health care, police and fire departments, highways, or public utilities. It may also provide for social reproduction through tax relief and policies geared to the provision of social welfare, child care, infrastructural development, and unemployment insurance. The mix of what is entailed in social reproduction and how it is accomplished and paid for by whom is a social, economic, and historical geographical question.

At another scale entirely, globalized capitalist production brings geographically disparate labor pools in more direct competition with each other, and can exert a homogenizing influence on the conditions of production and social reproduction. Through the expansion of capitalist social relations and the loosening of barriers to investment and trade, foundry workers in Ohio are put into competition with industrial laborers in Gujarat, where the costs of social reproduction – and thus wages – are lower. Likewise, when production shifts to places with fewer environmental regulations or their lax enforcement, the lower costs of maintaining and reproducing the conditions of production enhance capital accumulation. In these ways and others social reproduction is at the heart of capital accumulation, which is predicated on uneven geographical development and labor differentiation. The latter relies upon the reproduction of inequality through constructions of difference around race, gender, class, sexuality, and location. Thus, while the globalization of capitalist production can have a tendency to homogenize the conditions of production and social reproduction, the nature of uneven geographical development in combination with the rootedness of so much of what is entailed in social reproduction means that as production moves to areas of cheaper labor and lower environmental costs, the responsibility for and costs of social reproduction in the settings left behind or threatened with disinvestment are offloaded from capital to other sites such as the state, nongovernmental organizations, the household and family, and the individual. While this situation may leave those who are “stuck in place” having to cover more and more of the costs of their own social reproduction, it also suggests an arena ripe for organizing labor to ensure that any “homogenization” of the conditions and relations of social reproduction pushes the costs back to capital.

Social reproduction is a term often associated with Marxist political economy. Marx wrote about reproduction at the scale of the individual laborer and capitalist – what he called simple reproduction – in Capital, Volume 1, and of the reproduction of capitalist relations as a whole, which he referred to as expanded reproduction,
in *Capital*, Volume 2. Throughout his discussions of the contingent, historically grounded, and contested processes of reproduction, however, Marx’s focus was on capitalist production and on the waged labor–capital relationship. Unwaged reproductive labor, the domestic sphere, or what is now often referred to as caring work, do not figure significantly in Marx’s writings, to say nothing of social relations of difference apart from class or the other realms of social reproduction such as those associated with the production and exchange of knowledge and the sphere of everyday life. Marx associated these cultural forms and material social practices with the “superstructure,” which rested on and shaped the forces and relations of production or the “base,” and his interest was largely in the latter. Many Marxists continue to consider the realm of social reproduction as external to the creation and circulation of value, which theorists of social reproduction, many of whom are Marxist-feminists, see as limiting their analysis and praxis (Dalla Costa and James 1972; Federici 2004; Katz, Marston, and Mitchell 2015).

Henri Lefebvre is a notable exception to this tendency. While most geographers are familiar with his work on the production of space, Lefebvre focused three volumes on the “critique of everyday life.” Lefebvre understood the cultural forms and practices of everyday life in all of their mundane triviality and “colonization” by consumerism as “critical” material social practices wherein the possibility for rupture was immanent in the routine and quotidian. He framed everyday life as the site in which all activities were encompassed and encountered in all of their variation and contestation, creating a “common ground” of practice, social engagement, desire, and possibility. Likewise in geography, Allan Pred’s (1981) consideration of time geography as a way to examine the individual in relation to a social formation producing and reproducing one another as they are structuring and structured by time and space offered a specifically geographical understanding of social reproduction and the paths and projects of everyday life. Pred’s work deepened the analysis of everyday practice with attention to its necessary temporal and spatial constraints. In this effort, he was attentive to the intricate social realms – intimate and more global – of any “individual path” in articulation with a myriad of “institutional projects” so that the production and social reproduction of everyday life encompassed biography formation intertwined with the constitution of social institutions and their associated structural practices. Other Marxist theorists with an appreciation for everyday life and its critical possibilities include Pierre Bourdieu whose notion of the habitus is a way of understanding the durable socially structured systems that at once produce social life and are its outcome. Implicit but often difficult to imagine in the ongoing interrelated “structuring structures” of Bourdieu’s model is the possibility of social transformation – the incremental and more dramatic shifts that inhere in and produce everyday life and its myriad institutions and practices. Indeed, Bourdieu wrote incisively about how schools function to teach the skills and inculcate the discipline appropriate to participating in a capitalist workplace while at the same time naturalizing the sorting they perform in a classed and socially divided society. In a more lively and culturally rich vein, Paul Willis’s landmark book, *Learning to Labor* (1977), addressed how working class boys in Britain encounter the class-bound hierarchies and disciplinary practices of public education, and engage them in ways that at once express their interests, disdain, and desires and consign them to a future on the shop floor. Willis’s work, which is exemplary of the Birmingham cultural studies tradition, demonstrated brilliantly how social reproduction is an indeterminant process
of structurally constrained but active “choices” and “decisions” by social actors who resist the narratives of conventional schooling to make themselves otherwise. His work shows the concrete and intimate ways that people make themselves and their paths in and against the social formations in which they live, and how their everyday material social practices help to constitute those very social formations and the social relations associated with them.

By looking concretely at how social reproduction was accomplished historically through everyday practice in and against structuring structures across geographic scale, this work complemented and added immeasurably to Marxist understanding of social reproduction as at once the reproduction of the laborer and of the class relationship from the point of view of an individual worker or firm and the reproduction of capitalist relations as a whole, because it brought the cultural forms and practices through which social reproduction is accomplished – or not – to the fore. It remained to others, however, to look more closely at the labor involved in ensuring social reproduction – simple and expanded. The contributions of Marxist-feminists have been crucial in these realms. One of the most pivotal and inspiring texts was a 1972 manifesto by Mariarosa Dalla Costa and Selma James, which laid out in bold terms how the naturalized, unending, and unrecognized work of women in the home produced the commodity most central to capitalism – labor power. Withholding this labor, they argued, had the capacity to destroy capitalist (and patriarchal) relations of production and subvert conventional hegemonic social relations of production and reproduction (see Federici 1975).

Such analyses of the work of social reproduction as the gendered and largely unwaged work of reproducing the labor force expose the multiple layers of structured exploitation that women face in the heteronormative nuclear family. Working for a share of their husband’s wages – what economists describe as the “family wage” – women providing reproductive labor in the home are doubly alienated and doubly exploited. While many marriages and familial relationships are rooted in love and sustained by care and mutual aid, it remains that what is constructed – and may be experienced – as a consensual familial relationship is at the same time a structured set of relationships that extract value from the body of the housewife in the form of reproductive labor. While the fruits of her labor benefit and are enjoyed by her husband or partner and children, the value produced by domestic labor is tapped by capitalists in the course of the production process and contributes substantially to capital accumulation. Drawing an analogy to the terminology of racial capitalism, we might understand this relationship as gendered capitalism. Both forms of capitalism, which are everywhere entwined, enable and hide the deeper exploitation of women and people of color in distinct, overlapping, and intersecting ways.

Another revealing aspect of domestic labor has been that the increased participation of women in the wage labor force did not result in the transformation of their reproductive roles as many had imagined. Women working outside the home – which is by now commonplace and for many has always been the case – generally work a “double shift”; their labor outside the home (often in the caring professions) is preceded and followed by their reproductive labor – cooking, cleaning, shopping, laundering, child-rearing, and so on – at home. In each case, she produces value for the owners of the means of production directly in the workplace and indirectly in the home through producing and reproducing the labor power of herself, her husband or partner, and children.
Starting in the 1970s and picking up steam in the 1980s, feminist political economists and others provided rich analyses of these social relations of production and reproduction and the extraordinary, but commonly invisibilized, labor involved in producing the worker on a daily basis and over generations. Focusing on the domestic sphere, they documented the unwaged work of child care, food preparation, cleaning, laundering, scheduling, and other aspects of what is now referred to as “care work” and “affective labor,” and analyzed its contribution to the relations of production under capitalism and capital accumulation more generally. Looking at the articulation between capitalism and patriarchy, feminists examined how the naturalized labor of women over the life course produced labor power – a commodity sold to capitalists in the workplace – and thereby contributed to the extraction of surplus value by capitalists. For some theorists this situation spurred a movement to compensate domestic labor – the Wages for Housework movement, which was meant not as an end point in itself, but as a “political perspective” (Federici 1975). That is, the radical intent of the Wages for Housework movement was that it would force recognition of the role of women and domestic labor in producing uncompensated but necessary value for capital. Without ready workers able to sell their labor power to capitalists there could be no production and no extraction of surplus value from the worker. Cheapening and devaluing domestic labor enhance labor exploitation. The struggles associated with the Wages for Housework movement were seen as revolutionary in their capacity to undo the naturalized assumptions and practices of patriarchy, heteronormativity, and capitalism, along with the differentiated social positions and roles associated with them. Wages for Housework was, then, a call to women to organize with other women against the isolation and individuation of household labor, and thereby to “subvert the community,” and undermine capital accumulation.

Other feminist responses to the social reproduction work of women in the household and elsewhere had a less revolutionary imaginary perhaps, but include the enduring radical aspiration to socialize domestic labor through collective arrangements for its accomplishment, including for example shared kitchens, communal living arrangements, and collective child-care spaces. Still other approaches have advocated policies that recognize the economic contributions of caring work and provide public support for the value of that work, and the more liberal-minded reworking of gendered relations of production in the household in a more individualized manner. Some imagined that technological advances in the home would reduce women’s work, but while these technologies altered the nature of domestic work, most of them did little to reduce the time spent on it, and perhaps worse, served to elaborate on and extend its expectations (Strasser 1982).

Of course the domestic work of social reproduction is not just gendered, it is raced and classed, and often involves stretched geographies of care involving the labor of women from the Global South and poorer countries working in the homes of people in wealthier parts of the world. As Angela Davis (1981) and many others have noted, for many black women, housework and reproductive work in general, have long been performed outside of the home initially under conditions of slavery and then for depressed wages often under informal conditions, which do not provide benefits or security (Glenn 2012). Likewise, the intimate relations of social reproduction are often quite global. So-called nanny chains or the extended geographies of care work more generally involve women of color from poorer nations migrating to wealthier places to
SOCIAL REPRODUCTION

care for the children, homes, elderly, and health needs of people there. Among other things, and crucially, their migration is enabled by the labor of their extended families who take care of these needs – covering the work of social reproduction – in the migrants’ home environments, and by immigration policies that foster these sorts of exchanges to try and ensure that such migration remains temporary and that family members and communities in the sending countries receive remittances (Pratt 2012; Parreñas 2015). And yet the necessity of social reproduction and its accomplishment one way or another – by extended family in one place and hired labor in another – is at the heart of these policies and the demands that call them forth. These policies and practices rely on, as they maintain and reproduce, uneven development and the social relations of production and reproduction associated with racialized and gendered class difference. Whether temporary or permanent, labor migration represents a transfer of wealth between nations. Labor is after all a form of variable capital, and when it is produced, reproduced, and nurtured in migrants’ country of origin, it represents a transfer of capital between nations no less significant for being embodied than flows of financial capital or commodities produced cheaply.

While there are many critical insights in this work on the labor practices and exchanges of value in social reproduction, there are three themes in particular that have profound implications for the study of social reproduction as both central to capitalism and as the stuff of everyday life. The first theme is recognizing the home and family, along with other institutions such as schools and community groups, as sets of relationships structured to draw value out of unwaged work, to build consent around these structured relationships, and/or to contain and manage the contradictions of production and reproduction. Dalla Costa refers to these institutions and the sphere of social reproduction as the “community,” arguing that the community functions as a sort of “social factory,” producing value for capitalism while appearing to be outside of production. But along with this key role, the home and the sphere of social reproduction more generally are a key arena where the contradictions and crises of capitalism are managed. Analyses of social reproduction can reveal the unremarkable cultural forms and practices through which the social relations of capitalism and other structured relations of domination are managed, mediated, hidden, and reproduced, but also and importantly, can be recognized, remade, and resisted.

The second theme is the centrality of race, gender, age, and sexuality to capital accumulation. As suggested above, unwaged and other caring work, differentially exploited, globally sourced, and intimately deployed, is naturalized through institutions such as the family and the household, the nursery, and the school, and with them, the gendered, age-graded, and racialized divisions of labor that comprise and animate them. As long as these relations are naturalized their structural underpinnings and outcomes are obscured, as are their intricate and important connections to class exploitation and capital accumulation. As George Caffentzis (1999, 162) notes, “The problem with Marxist theory was that it could only explain the reproduction of the capitalist-waged-worker relation. But the revolutionary subjects of the ’60s were mostly unwaged.” Today these subjects would include the unemployed, the incarcerated, students, peasants in the Global South, those engaged in the informal economy, and housewives, among others. These populations have been recognized as a sort of safety valve for capital accumulation by comprising what Marx understood as a “labor reserve army,” but also through their profitable
and strategically stabilizing containment, management, disciplining, and distracting. Unwaged workers, then, are very much part of the labor force and are differentiated through the production and reproduction of race, gender, age, and sexuality. Critical attention to social reproduction can illuminate the ways these differences are produced, are structured in relation to class, and are sustaining to capitalism. Their recognition as such would highlight the potential for altering them along with the social relations that hold them in place in and through the material social practices of social reproduction.

Finally, the Marxist feminist position on value production through reproductive labor has important ramifications in terms of locating contradiction and possibility in capitalist crisis. “Once we see the community as a productive center and thus a center of subversion, the whole perspective for generalized struggle and revolutionary organization is re-opened” (Dalla Costa and James 1972, 13; emphasis in original). As suggested above, organized labor, the factory worker on strike, and waged workers more generally can no longer be constructed as the only revolutionary subjects – or even the most likely. While some Marxist theorists continue to question what constitutes social reproduction under capitalism, and whether or not certain types of nonwaged labor and extra-market relations produce value, it is clear that theorizing social reproduction beyond the reproduction of the waged labor–capital relation brings with it a new way of imagining – and organizing – the working class, and opens up new arenas for mobilizing resistance to various relations of domination and exploitation.

Across the board it is important to remember that social reproduction is always contested, and plays out according to concrete historical and geographical conditions that are shaped through struggle. Social reproduction is always already fraught with tension and possibility; its “fleshy, messy” practices are essential to the production of value through the wage labor–capital relation, but at the same time can be reservoirs of possibility and potential change as well as self-care and reconstitution, which themselves can be a form of resilience. Alongside and through these practices, however, capital is accumulated through the looting of value produced in this sphere of everyday life. This dialectical relationship adds to the potency of social reproduction, but also its perils.

This dialectical relationship undergirds and sustains the capitalist mode of production. By definition if social reproduction is in crisis so too is capitalism. Periodic crisis is in fact part of uneven development, and often resolved through the so-called spatial fix, although there are many other means of crisis resolution depending in part upon its nature, location, extent, and historical geography. Crises of accumulation occur, for instance, when excess capital pools and cannot be productively employed, or productive capacity is idled for lack of demand in existing markets or lack of a labor force adequate to the task. In other words, when there is too much time between the production of commodities, including labor power, and their sale, or if there is no way to expand markets geographically or to intensify production, then capital accumulation falters and may be in crisis. While such crises may be ameliorated through the material social forms and practices associated with social reproduction such as investments in education or infrastructure, they may provoke a crisis in social reproduction as well.

A crisis of social reproduction occurs when existing social, political economic, or environmental conditions and relations can no longer be reproduced, for example, if commodities once created cannot be brought to market, if the time between production and consumption is too long, if flows of capital cannot be invested in
productive activity, or the material forces and grounds of production are degraded. Likewise, a crisis of social reproduction occurs if the labor force cannot be reproduced in a given time and place or find means to labor productively in a given setting. Such crises may be resolved by labor migration, but can also lead to a large number of people looking for employment and finding none as has become common in many parts of the world. The burgeoning of such populations itself threatens the stability of capital accumulation. One way that surplus labor is “managed” is through containment strategies of various kinds – the coercive structures of state violence and incarceration are the most pernicious, but so too is the structural and systemic neglect of social welfare, health care, housing, and education which fail populations excessed by production and accumulation crises subjecting them to harm and early death. In another realm, the military and militias absorb many young people with few other options. The attenuation of higher education both in the ways it can draw young people with few employment options or because the credentials required for various positions have been expanded is another means of managing the contradictions of capitalist accumulation and the crises that accompany them.

These examples suggest some of the ways social reproduction encompasses the structured material and discursive practices and institutions that hold capitalist social relations together, and can manage their inherent contradictions through state policy and practice, the affective labor of the domestic sphere, and the social work associated with civil society. These institutions and practices are wide-ranging, and include those associated with education, the judiciary, the military, the “nonprofit industrial complex,” and cultural governance, among others. These contradictions, discontinuities, and possibilities for social transformation are inherent in capitalist production, but also in the reproduction of capitalist social relations. The centrality of reproduction to capitalism and what its work consists of, then, are critically important political questions, which are often sidestepped in conventional Marxist praxis and analysis, but which have played a growing role in more recent social movements such as “occupy,” various “mothers’” movements, and social activism around “commoning,” for example.

As is well known, capitalist production requires a class of people who own the means of production and a class of people who do not own the means of production, and are free to sell their labor power as a commodity to those who do. In Volume 1 of *Capital* Marx explains the violent historical practices of primitive accumulation, which alienated masses of people from the means of production, “freeing” them to sell their labor power to capitalists who were taking ownership of the means of production, and leaving them with little choice for economic survival but to do so. Primitive accumulation, as the name implies, was understood as “the original sin of capitalism,” outside of and prior to capitalist social reproduction and production, and the relationship between the violence and coercion of original accumulation to production remains a historical one. But there is a parallel here with the everyday work of social reproduction, and the coercion, violence, displacement, and dispossession associated with practices akin to primitive accumulation remain an integral part of capitalist relations in the present. The historical and ongoing role of primitive accumulation, or what David Harvey (2003) has dubbed “accumulation by dispossession,” relates directly to social reproduction in as much as primitive accumulation describes the process of looting the “outside” of capitalist social relations: nonmonetized or noncapitalist material practices – including the
“fleshy, messy stuff of daily life” and material social practices and relationships outside of the waged labor–capital relationship.

The confluence of primitive accumulation and social reproduction is a cornerstone in the scholarship of Silvia Federici (2004), who locates the historical primitive accumulation that preceded European capitalism in women’s bodies and material practices. In her book *Caliban and the Witch*, Federici argues that it was the subjugation of women and the forcible separation of “productive” and “reproductive” labor that allowed for the concentration of wealth necessary for the creation of capital. Tracing the development of capitalism through the history of reproductive labor, Federici argues that “primitive accumulation has been a universal process in every phase of capitalist development” (2004, 16). Jennifer Morgan’s (2004) historical study of reproductive labor under slavery similarly locates the site of primitive accumulation in women’s bodies. African slaves and their descendants, Morgan shows, performed both productive work and reproductive work by reproducing the labor force daily and generationally. Under slavery, women produced both the commodity and the labor force, which were one and the same. The wealth that was accumulated in slaveholding economies and then put into motion as industrial capital was extracted from black and brown women’s bodies. Social reproduction and reproductive labor, while commonly naturalized as outside of production and capitalist social relations, are central to the production of value both through the production process, which requires the reproduction of labor, and through accumulation through dispossession, which extracts value from noncapitalist spheres, from racialized and gendered bodies, from the material practices of reproduction, from the environment, and from the future in the form of credit.

Social reproduction is the material social practices through which people reproduce themselves on a daily and generational basis and through which the material bases and social relations of capital are maintained, invigorated, and renewed. These material practices are also spatial practices. Social reproduction encompasses the stuff of everyday life, is grounded in the production and consumption of use-values in the everyday environment, and is embodied. But beyond the realm of the everyday, social reproduction involves the reproduction of the conditions of production, which among other things includes the physical environment – the built forms and spaces of production and the environmental resources and physical conditions that enable ongoing production.

Social reproduction has political economic, cultural, and environmental aspects. As a set of structured material practices, it must be “held together” through consent and the production of commonsense understandings of its compass and the nature of its provision. But social reproduction also refers to the making and maintenance of the forces and material grounds of production. Much of the work of political ecologists essentially addresses social reproduction especially at its confluence with ongoing primitive accumulation in underdeveloped and postcolonial settings, although this work is rarely understood by these researchers or others as of a piece with social reproduction. But, much of the early work in political ecology was concerned with environmental destruction resulting from increasing pressures brought to bear on the social reproduction of peasant and rural social formations as they were incorporated, directly and/or indirectly, into a global capitalist system. A focus on social reproduction offers a theoretical and practical framework that brings together Marxist, feminist, and antiracist work
on the social relations of production with environmentalist work focused on the deterioration of environmental resources and landscapes of labor, and can enhance the political horizons of both concerns. Likewise, work focused on environmental justice is directly connected to social reproduction in its concerns with the uneven environmental conditions in which people carry out their everyday lives and children come of age. Political organizing around social reproduction and environmental justice has found common ground and made common cause in many places – rural and urban, Global North and South – around the stark and growing issues of the uneven distribution of waste and value at all scales. In these ways and others, social reproduction is a key concept of geography, and its analysis offers scholars and practitioners powerful means of understanding the everyday environment and the social relations that hold it in place and might change it.

SEE ALSO: Care work; Caregiving; Domestic spaces; Domestic workers; Environment and everyday life; Everyday geographies; Feminist geography; Gender, work, and employment; Habitus; Home; Labor migration; Life course; Marxist geography; Patriarchy; Race and racism; Work–life balance

References


Further reading


Social resilience and environmental hazards

Hugh Deeming
Northumbria University, UK

The etymology of the word “resilience” has been traced from its use in ancient Greece, via the reign of the English king Henry VIII, to its modern multidisciplinary applications across the disciplines of materials science, systems theory (e.g., ecological and social-ecological), psychology, disaster risk reduction, and others (Alexander 2013). The consistent interpretation, which tracks through much of this usage, is the idea of “bouncing back,” that is, there has been a tendency for resilience to be associated with the concept of equilibrium, or rather the return to equilibrium after perturbation. This idea of resilience as an indicator of a type of system plasticity has been referred to as “engineering resilience.” In recent decades, however, other academic interpretations have diverged from the idea that resilience can be considered as a factor in the maintenance of some sort of quantifiable stasis and, rather, propose a more flexible situation where resilience allows a system to maintain not equilibrium, but functional persistence. From an ecological perspective, resilience is viewed as an attribute that allows systems to operate within any one of a number of alternative stability regimes, with each regime determined by prevailing conditions/stressors. For example, a lake can either function in a clear or turbid condition depending on environmental factors. In terms of understanding a social system’s ability to persist, use of the concept has included the focused investigation of three characteristics: coping (resistance), self-organization, and learning from experience (adaptation). It has been suggested that social resilience, understood through this lens, depends on complex interactions between ecological, social, economic, infrastructural, and institutional components, which if effective can allow the system to respond to predictable disturbances, as well as to unforeseen shocks.

Resilience has become a focus of attention as societies have come to realize the unpredictable complexity of nature and the operation of social and ecological systems within it. Rather than vulnerability, which has been associated with a pessimistic focus on a system’s potential for harm, resilience has emerged over recent years as a positive, forward-looking system attribute. As a result of the attention paid to it, the concept of social resilience has developed through both analytical approaches, which concentrate on measurement, and normative approaches that are more metaphorical than empirical. These focus on judgments and prescriptions over subjectively aspired-to system conditions, rather than the isolation and quantification of specific variables. This duality has led to significant criticism, wherein the term has been accused of becoming so malleable in its application as to leave it without a distinct meaning. However, further attempts at better definition have included its separation from the related concept of transformation, that is, this separation has reinforced the perspective that the principal attributes of a resilient social system are inherent in its capacity to maintain its function, rather than being a driver of transition or transformation to a fundamentally different regime. It is this
resilience for functional persistence that has been identified as a problem for resilience-based strategies. This is because there may be risk involved in maintaining regimes that have been socially inequitable from the outset, for example, states of poverty and autocratic government that have proved to be highly resistant to change. From this perspective, it could be argued that resilience can be viewed as a constraint, in that it limits a society’s adaptive potential to actions that maintain the status quo ante. From this perspective, resilience can be viewed in terms of its ability to disguise and perpetuate the negative aspects of path dependence. However, given the multiple usage of the concept, it is important to take this as an opportunity to review how some of the principal approaches to understanding social resilience, and particularly social resilience to environmental hazards, have developed over recent decades.

One of the earlier modern uses of the concept was in the field of psychology in the 1940s and 1950s, where researchers initially regarded resilience as a personality trait. However, further work on the role of external conditions in defining childhood resilience led to a hypothesis that resiliency was developed through social processes rather than being a predetermined characteristic, that is, personal resilience became a function of social interaction. This view was useful because it allowed individuals’ resilience to adversity to be scaled up to higher social levels. From this perspective, mental health research has highlighted the importance of access to social support, which reduces mental health problems following disaster.

Returning to the idea of system persistence, Wildavsky (1988) developed his theory of organizations by distinguishing between the concepts of resilience and anticipation. He argued that anticipation focuses on avoiding and mitigating known and specific hazards, whereas resilience is inherent in an organization’s “capacity to cope with unanticipated dangers after they have become manifest, learning to bounce back” (Wildavsky 1988, 77). This understanding is useful in that it supports the idea that general resilience is an important mediator of the impacts associated with surprise events. Viewed in these terms, it is perhaps a little disheartening to find that a particularly powerful determinant of an individual’s perception and self-efficacy in the face of hazards is their personal experience of those hazards. While such a finding supports approaches that propose learning as a principal determinant of resilience, it is also limiting in that, if known hazards are those that people engage with best, the challenge of encouraging engagement with the unanticipated is obvious. In terms of risk perception, this tendency is referred to as “preparing for the last disaster.” Much civil protection doctrine is constructed around what are effectively anticipation-only approaches, which could be said to de-emphasize the fact that surprise events may require innovative and flexible responses rather than the pre-scripted actions.

One of the most widely recognized approaches in the development of the resilience concept is that of social-ecological systems (SES). Holling (1973) was arguably the first to incorporate resilience into the discipline of ecological systems science in a way that challenged the notion of systems operating at or around equilibrium. He suggested that systems constantly operate in a transient state. He used the idea of the “adaptive cycle” (i.e., growth or exploitation; conservation; collapse or release; and reorganization) to describe the way ecosystems develop and persist over time (even though some constituents of that system may experience extinction, to the benefit of others). He illustrated how this model could be used to identify how ostensibly closed systems, such as lakes, could be seen to respond to external...
(principally anthropogenic) perturbations (e.g., eutrophication, overfishing), through regime persistence or, alternatively, through crossing thresholds into alternative stability regimes (e.g., from clear to turbid), while maintaining their primary function (i.e., of being a lake). He referred to these alternative stability regimes as “domains of attraction.” The term “social-ecological system” evolved from this approach, in order to further emphasize the role of humans in nature. Berkes and Folke (1998) stressed that the separation between social and ecological systems is effectively artificial and arbitrary and suggested that social and ecological dimensions be attributed equal weight in any analysis.

SES theory incorporates issues of scale through the concept of panarchy. Panarchy describes how system persistence is determined by the interrelationships between the slow adaptive cycling experienced at a system’s macrolevel and the more rapid cycling that occurs at the meso and micro (i.e., local) scales. As an example, panarchy can be imagined as the feedback processes that operate through the cascading social scales of county, town, neighborhood, and household.

The persistence or destruction of nonhuman systems will ultimately be determined by the environmental stressors to which it is exposed (whether natural or anthropogenic). However, the key determinant of human agency is intention. Through their intent, humans give effect to their capacities for learning and technological progress. Therefore, human social systems can be seen to have a greater capacity for adaptation (i.e., the collective capacity of the human actors in the system to manage their resilience, whether this is done effectively or not). Adaptability has, therefore, been proposed as a crucial social component in the regulation of SES. This is despite the fact that adaptive behavior and self-organization within the system may be intentional or unintentional in relation to its effect on the system or interrelated systems. Social actions should also be understood to have the potential to push the system closer to, or over, thresholds into different stability regimes or even toward collapse. Such a process, wherein resilience is reduced rather than increased by actions taken, is referred to as maladaptation. Maladaptive actions are, however, often difficult to differentiate from effective actions, because their results may still be viewed as successful by some observers.

Actions or circumstances that push a social system beyond the tenable boundaries of its normal functioning (i.e., beyond its domain of attraction) can take a system beyond resilience and into transformation, creating a fundamentally new system with radically different ecological, economic, or social (including political) characteristics. Pelling (2010) discusses transformation and suggests that it may be triggered by any economic, political, or environmental shock. In terms of social contracts, such events can precipitate perceptions of manifest failure in what are perceived by the public to be institutions of security against these shocks. For example, sea level rise has been associated with the dispersal, and effective destruction, of many small communities built on eroding coastlines, after decisions to cease erosion prevention work have been introduced top-down from higher political levels.

Incorporating these considerations and the principle that resilience involves an element of anticipation and planning for the future, the concept of social resilience has gained increasing recognition in the literature. Adger, for example, has defined social resilience as “the ability of groups or communities to cope with external stresses and disturbances as a result of social, political and environmental change” (2000, 347). However, this conceptualization of resilience as a social capacity for delivering system persistence has also crossed from the academic/empirical sphere into that of governance and practice. In
places it has morphed into the more normative concept of “community resilience.” While apparently compelling, the attraction of applying a social metric such as “community” to any understanding of system persistence is challenging, not least because of the lack of an adequate definition of “community” as any sort of uniformly coherent social unit. For example, the consideration of community in terms of exposure to environmental hazards could legitimately focus on the concept of place communities (e.g., populations resident in a seismic zone). However, in wider resilience terms, the risk-mitigation and recovery-enabling resources available to different individuals and groups within any place-based population would likely depend much more on their connection to other, less geographically dependent, forms of community, for example interest, identity, support, and circumstance. Furthermore, communities of practice – which can comprise practitioners as well as engaged publics, whose collective intent is to mitigate disaster impacts – are increasingly utilizing modern social media in order to connect and share experiences and to advocate for disaster risk reduction causes across the globe.

In terms of understanding organizations as formalized social structures, it has been suggested that anticipation strategies again provide the most popular focus for understanding their capacity to survive disruptions. However, such approaches are also acknowledged as offering the temptation for these organizations to develop resistance and efficiency strategies that could actually undermine rather than enhance their resilience to the unexpected. Weick and Sutcliffe (2007) propose that high-reliability organizations (HROs) (e.g., nuclear power plants) are less likely to degrade their resilience in this way because they operate on a set of key principles that enhance their ability to respond dynamically to known, as well as new and emerging, hazards. These are a preoccupation with failure; a reluctance to simplify; sensitivity to operations; a commitment to resilience; and deference to expertise (i.e., rather than rank).

Despite its association with anticipation, the capacity of a social system to resist hazard effects is a fundamental factor in increasing resilience through the reduction of harm, particularly from high-probability hazards. For example, the benefits of technologies such as flood barriers, or legislative measures such as seismic building codes, have in reality reduced the impacts of predictable hazards. However, resilience is also associated with an element of recovery from impact (i.e., the bouncing back). Therefore, a social system’s ability to substitute resource elements (e.g., information networks, human or electronic), to act with resourcefulness and to do so with rapidity, have also been identified as fundamental in defining that system’s resilience capacity. In an increasingly interconnected world, the fact that nobody can have complete knowledge about future hazard effects further highlights the importance of understanding resilience in terms of enhancing these attributes. Not least because of the increasing risk that hazard effects can become linked or compounded in unanticipated ways (e.g., the impact of the 2011 floods in Thailand on the global availability of computer hard disk drives), Furthermore, the attributes that ultimately enable effective response to surprise events to be achieved with dynamism and flexibility may not necessarily be those that are normally associated with civil protection. It has been proposed, for example, that social resilience be defined as “a process linking a set of adaptive capacities to a positive trajectory of functioning and adaptation after a disturbance” (Norris et al. 2008, 131). From this perspective, social resilience is no longer considered as something that can be achieved
simply by, for example, building flood barriers. Rather, it should be seen as something that encompasses a more complex approach that nurtures a social system’s ability to create and to maintain social wellbeing through the development of an ability to dynamically link its own capacities for economic development, social capital, information and communication, and community competence.

This idea of wellbeing as a desirable manifestation of effective adaptation to a new postdisturbance environment is useful, because it equates resilience with collective health and quality of life. If a population has sufficient dynamically linked capacities, then hazard perturbations do not have to turn into disasters. Such thinking emphasizes the importance of acknowledging the importance of any work undertaken to increase social equity and to reduce vulnerability, not just in relation to predictable hazards, but more generally in terms of mitigating the root causes, dynamic pressures, and dangerous conditions that create all forms of social vulnerability. This perspective resonates with some of the critiques of SES and organizational theory, in that it brings power relations to the fore. Understanding resilience as a social process that is manifested in sustainable collective wellbeing and an ability to rally resources to deal with “known” hazards as well as unanticipated shocks, reduces the temptation to externalize chronic social inequities from the equation (i.e., where vulnerabilities not directly related to predictable hazards are displaced as a draw on others’ budgets rather than on one’s own).

SES theory in particular has been criticized for downplaying the contested nature of social interactions (e.g., the resilience of whom and at whose expense greater resilience comes) and the “wicked” complexity of society–nature dependencies (e.g., the fact that it is often the most vulnerable populations who are forced to live in the most hazardous environments). This argument proposes that these inherent social complexities and inequalities reveal that, unlike nonhuman ecological systems, not all human social systems can self-organize and adapt in any sort of optimal way, that is, there are usually winners and losers.

This complexity and the degree of latency involved, which blurs the predictability of how any particular population will actually react to a hazard-related disturbance, means that quantifying social resilience remains problematic. However, the evolving understanding of the many factors involved has allowed first attempts to be made at developing baseline indicators by which to measure and monitor the disaster resilience of places (Cutter, Burton, and Emrich 2010).

When considering community resilience in the context of natural hazards, a traditional view would be to focus on the interface between the hazard/disaster and the narrowly constrained emergency response, preparedness, and mitigation activities that have become the remit of civil protection practitioners (i.e., the emergency services). However, the focus has evolved from the simple questioning of whether a society would be more resilient if a perfect warning system, flood control measure, or response capability could be designed. It has been proposed that social resilience encompasses much broader issues. The lenses through which resilience has been viewed have revealed that social systems require a much broader and more complex frame through which to understand their resilience and how it is generated, regenerated, maintained, degraded, or lost, than can be provided within even the quite advanced sociotechnical hazard and risk management systems. Even with the best will, social resilience is not something that can be created purely through altruistic intent or policy directive. Understanding social resilience to environmental hazards increasingly requires a
recognition of the exposed populations’ broad entitlements and access to a range of capital resources, as well as the capability for them to be deployed and used effectively in ways that allow those people to “cope with a hazardous event or trend or disturbance, responding or reorganizing in ways that maintain their essential function, identity, and structure, while also maintaining the capacity for adaptation, learning, and transformation” (Field et al. 2014, 5).

SEE ALSO: Climate change adaptation and social transformation; Community; Social capital; Social vulnerability and environmental hazards

References


Social vulnerability and environmental hazards

Terry Cannon
Institute of Development Studies, UK

Vulnerability and social vulnerability

The term “vulnerability” is widely used in general and in a range of academic disciplines including psychology, social work, ecology, gender studies, sociology, anthropology, and geography. It has many definitions but is often used in a lax and undefined way, especially by the public, media organizations (including non-governmental organizations), and governments. Like the term “sustainability,” it is in danger of being used uncritically and of losing its validity as a basis for rigorous analysis. This is especially the case when it is used to convey ideas of victimhood or incapacity without reference to any particular threat.

“Vulnerability” is a term that is used to guide significant amounts of spending in welfare and health (including psychiatry) and in relation to environmental hazards. Because it is employed across disciplines, its academic use needs to be rigorous to counter its potential to mislead when used by media and politicians, who may find it convenient to use the term imprecisely. The primary difference between rigorous and lax usage concerns causation and attribution, and the need to specify:

- what has caused a particular entity (person, household, etc.) to be liable to suffer harm;
- what the specific threat is that is defined in relation to this liability to suffer harm.

This is especially important in regard to environmental hazards, and the modified term that is widely used in that context, “social vulnerability.” It is the causation aspect that tempts politicians to refer to natural disasters, because this enables them to avoid acknowledging that there are social factors that cause vulnerability over which they have some control.

Table 1 provides some of the more widely used definitions of vulnerability in relation to natural hazards; they do not necessarily refer to social vulnerability, as this tends to be taken for granted when authors refer to people and human systems. A general definition of vulnerability is the potential of an entity to suffer harm, which involves the characteristics of someone or something that may lead them to suffer as a result of a particular event or process. This is common to all the definitions given in Table 1. In terms of causation, it is therefore necessary not only to understand why the event can have harmful effects (e.g., the potential harm that can come from floodwaters, storm impacts, earthquake shaking), but also what causes the conditions that lead to possible harm (social processes that generate vulnerability).

Social vulnerability is a modification of the concept that emphasizes two aspects:

1. the objects of interest or the entity that exhibits vulnerability are people and the social system; and
2. the causes of vulnerability are to be found in social systems (i.e., the economic, political, and cultural aspects of society).
SOCIAL VULNERABILITY AND ENVIRONMENTAL HAZARDS

Table 1  Some widely used or referenced definitions of vulnerability (including social vulnerability) in the context of natural hazards.

<table>
<thead>
<tr>
<th>Source</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergovernmental Panel on Climate Change (IPCC 1997)</td>
<td>“Vulnerability is defined as the extent to which a natural or social system is susceptible to sustaining damage from climate change. Vulnerability is a function of the sensitivity of a system to changes in climate and the ability to adapt the system to changes in climate. Under this framework, a highly vulnerable system would be one that is highly sensitive to modest changes in climate.”</td>
</tr>
<tr>
<td>UN International Strategy for Disaster Reduction (UNISDR 2009)</td>
<td>“The characteristics and circumstances of a community, system or asset that make it susceptible to the damaging effects of a hazard.”</td>
</tr>
<tr>
<td>Wisner et al. (2004, 11)</td>
<td>“By vulnerability we mean the characteristics of a person or group in terms of their capacity to anticipate, cope with, resist and recover from the impact of a natural hazard. It involves a combination of factors that determine the degree to which someone's life and livelihood are put at risk by a discrete and identifiable event in nature or in society.”</td>
</tr>
<tr>
<td>Bohle, Downing, and Watts (1994, 37)</td>
<td>“Vulnerability is best described as an aggregate measure of human welfare that integrates environmental, social, economic and political exposure to a range of potential harmful perturbations. Vulnerability is a multi-layered and multidimensional social space defined by the determinate, political, economic and institutional capabilities of people in specific places at specific times.”</td>
</tr>
</tbody>
</table>

Note: See also weADAPT 2011; Adger 2006.

It is thus distinct from vulnerability, where the focus is on physical structures: although these structures are part of the social system, it has a different entry point and different policy connotations. Social vulnerability is therefore definable as the likelihood that people and the way they live will be harmed by a natural hazard because of the influence of economic, political, and social factors. In this definition, vulnerability is relevant when a hazard occurs: one can be vulnerable only in relation to a particular type of risk. This precision avoids the lax usage that ascribes vulnerability to poverty or marginality. While many poor people may be vulnerable to a flood, they may not be badly affected by an earthquake (poor people’s houses may be too flimsy to cause death or injury if they collapse).

Vulnerability should therefore be defined in relation to a particular hazard. The potential for harm is different depending on the type of hazard, the sort of damage it can do, and the particular characteristic (of a person or entity) affected. Vulnerability is therefore “activated” only in the context of a particular hazard. The likelihood of harm is a consequence of already pre-existing conditions that are socially constructed. However, there is an alternative view that considers vulnerability to be a characteristic that is independent of any particular hazard. This approach may be useful in identifying large cohorts of people who share general characteristics that they are likely to be harmed by a broad range of hazards, although it runs the risk of becoming unspecific, for instance, when the term vulnerability is correlated with poverty.

The widespread correlation of poverty with vulnerability should not imply that there is a simple causal relationship. Poverty is often
invoked in relation to disasters as part of humanitarian arguments against inequality and injustice and how these can be linked (as causal factors) with disasters. But it can be misleading, as is the common assumption that most disaster victims are poor. This is largely a function of the high proportion of poor people in many less developed countries: when it strikes, a hazard is likely to find many poor people exposed to it. Another related statement is that poor countries are “more vulnerable” because they are the ones that experience most (or the worst) disasters. Poverty may generally be a good predictor of negative impacts of a hazard, but it is entirely possible for a disaster to affect predominantly middle-income and even rich people, depending on their actual characteristics of vulnerability. For example, there are cases where a high proportion of death and injury caused by earthquakes has been suffered by the middle class, as in Izmir (Turkey) in 1999. This is typically a result of the people living in badly (and corruptly) constructed buildings.

Environmental hazards

These are defined here as hazards caused by processes of the natural environment. Other problems are often described as environmental hazards, but these are generally derived from human sources such as pollution (of soil, air, or water) which then affect or become embedded in the natural environment. Because these are caused by human action (industry, agriculture, transportation, mining, nuclear and other energy sources) they are not included here. This discussion of social vulnerability is related in the strict sense to environmental hazards that arise from geotectonic and climatological processes. However, environmental damage (e.g., air pollution, including greenhouse gas emissions, deforestation, built environment, and urban heat island effects) can intensify some natural hazards. This is especially evident with global warming and the causal link with extreme climate events (IPCC 2012).

Social vulnerability as a reflection of socioeconomic differences

As with earthquakes, each type of hazard needs to be assessed according to how particular characteristics of people and human systems may or may not be affected by it. Vulnerability varies across different categories of people according to the way economic, political, and social factors affect them. Gender is one significant differentiating factor that leads to higher and lower levels of vulnerability. In many parts of the world gender relations may make women more vulnerable than men to specific hazards. For example, in Bangladesh, traditional gendered behavior means that women are reluctant to evacuate to safety when cyclone warnings are issued unless they have the permission of their husband or a male relative. For this and other reasons (e.g., being responsible for the care of children and elderly relatives, and cultural prohibitions on learning to swim), mortality among women is often much higher. Another aspect of gender relations may make men more vulnerable, as for example when gendered culture is expressed as “macho” behavior. In some countries some men appear to be willing to drive into floodwater, resulting in higher male mortality rates (admittedly with a much smaller sample size). There is also some evidence in developed countries that husbands may experience more significant mental disturbance after floods, because of their sense of failure to protect their family and home. So social vulnerability in relation to gender behavior can have what appear to be opposite
SOCIAL VULNERABILITY AND ENVIRONMENTAL HAZARDS

effects, depending on the hazard and the cultural context.

Others key drivers of social vulnerability include class, caste, ethnicity, age, and (dis)ability (Wisner et al. 2004). Each of these factors may predispose particular people to be more or less likely to be harmed in relation to a specified hazard. The key analytical value of social vulnerability is that it enables disasters that are triggered by environmental hazards to be understood in relation to people's different characteristics as determined by social factors and processes. A key factor is the economic context. A poor person may have a weak immune system because of inadequate nutrition (often as a result of poverty). In a hazard such as a flood, where diseases transmitted through fecal contamination of water may be a danger, this can be a characteristic that increases vulnerability. Having a low-income livelihood may mean that a household cannot afford to improve its house to withstand a hazard, or that they can only pay low or no rent, and so live in a dangerous location (perhaps a hillslope in a city). However, having higher income is no guarantee that people will spend what it takes to live more safely, even when they are well aware of the hazards they face.

Political factors affect vulnerability in two ways. First, the power relations of a society determine the allocation of assets and production resources (e.g., land tenure systems in many countries are highly unequal and may be class- or caste-based), the income distribution and allocation of services (health and education), and welfare (unemployment pay and pensions). This can clearly affect aspects of vulnerability that are related to poverty, gender, ethnicity (discrimination in allocation of assets, income, services, and welfare) and disability. Many studies of the impact of Hurricane Katrina in 2005 argue that these factors were very relevant in the disaster. Resource allocation is also a very significant component of vulnerability in many less developed countries. Second, the type of power system affects the level of resources available to, and the efforts made to prepare for (and to respond to), disasters. For example, a government that restricts or represses civil society is likely to have a reduced level of disaster preparedness because NGOs and other organizations cannot advocate for risk reduction and are restricted from organizing their own activities. This political connection with vulnerability has been identified, for example, in Myanmar (in relation to Cyclone Nargis in 2008), China (in regard to several disasters and disease outbreaks), North Korea (famine), and the United States (Hurricane Katrina). A related political factor involves the freedom of the media, and the extent to which it can act without censorship. This link was famously made by Drèze and Sen (1989) in relation to famine, in a comparison of India and China: in the former, a relatively free media can act as a watchdog for the formation of food crises, while in China hunger and other disasters are concealed and those who might reveal them suppressed.

As the popularity of the perception of disasters as being socially constructed has grown in recent decades, a number of critiques and embellishments of the vulnerability concept have emerged. In the 1980s the notion of vulnerability was challenged on the basis that it led to a perception of people (especially in developing countries) as victims (Anderson and Woodrow 1998/1987). This critique argues that it is essential to acknowledge (and support) people’s own capacity in dealing with hazards. However, there is potential for confusion here: capacity can logically be seen as the opposite end of a vulnerability spectrum. Therefore people may have capacities in certain areas of their ability to withstand hazards but not in others. This approach was accompanied by a widespread
introduction, mainly by nongovernmental organizations (NGOs) and the International Federation of Red Cross and Red Crescent Societies (IFRC), of participatory vulnerability and capacity assessments that are designed to involve people at grassroots in reducing their own vulnerability.

A linked critique argues that the vulnerability concept arose among Western academics and ascribes to people in developing countries a “colonized” characteristic in which the concept helps fulfill power relations (e.g., Bankoff 2003). In disaster situations in developing countries, people are defined as vulnerable in order to provide justification for Western agencies (e.g., international organizations, development banks, and NGOs) maintaining a power hierarchy that privileges the developed countries as benefactors. This critique argues that greater allowance must be made for the people affected by hazards to exercise their own capacities. It also acknowledges people’s own culture and priorities rather than those of the outsiders who claim to be helping. This is part of a wider critique of disaster risk reduction organizations for giving priority to their own goals (“intensive risks”) when many of the risks that people face in developing countries are related to everyday (“development”) problems (e.g., of health, education, nutrition, crime) – what have been called “extensive risks.” This also shows that analysis of social vulnerability has mainly emphasized economic and political factors and excluded culture and people’s own notions of risk (IFRC 2014). Moreover, it relates to the fact that many people voluntarily expose themselves to hazards (e.g., in floodplains, volcanic slopes, coasts, and fault zones), often using cultural justifications (such as explaining the causes of disasters in religious terms or responding to them in a fatalistic way).

Measuring vulnerability and using vulnerability analysis

Vulnerability analysis, which refers to social vulnerability in the context of natural hazards, describes a range of tools developed in the research and the practice of organizations that conduct disaster risk reduction (DRR) activities. It emerged in recent decades, and is used mainly by NGOs and international organizations to assess different levels of vulnerability and their causes. In most cases, vulnerability analysis involves efforts to disaggregate people into different levels of vulnerability and to assess their likely harm in relation to natural hazards. Generically these assessments have been termed “community vulnerability assessments” (CRA) when they are conducted at the local level and are aimed at supporting DRR at the local level (often termed “community-based disaster risk reduction,” or CBDRR). A CRA is typically based on participatory methods (such as participatory rapid/rural appraisal, or PRA), and is usually claimed to involve the “community” in designing and implementing DRR activities. Most large NGOs, the Red Cross/Red Crescent, and international organizations have their own (very similar) versions of CRA (see IFRC 2006 for an example). There are more recent initiatives (mainly by large NGOs and some international organizations) to use similar methods for assessing vulnerability to climate change in order to inform and carry out community-based adaptation (CBA).

These approaches and activities tend to be based on an idealized notion of community, which lacks analysis of the causal factors (especially systems of power) that are involved in generating vulnerability. In other words, they ascribe to a normally very heterogeneous set of people in a location a collective characteristic, such as “vulnerable community.” This is partly in
tacit recognition of the inherent difficulty of changing power relations when engaged in short-term DRR or CBA projects (which are normally of three to five years’ duration). CRA methodologies assess key areas of social vulnerability, including livelihoods and their asset base, income levels and other economic variables, environmental hazards, and institutional linkages within and beyond the “community.” There is a tendency to sidestep or play down issues such as unequal access to land and other resources, and low income from poor livelihoods. These are typical vulnerability factors that cannot easily be addressed because of power relations and the short-term involvement of the outside organization. Analysis of local vulnerability that is actually caused by power systems is often absent, including assessment of land tenure, elite capture, and problems of participatory methods. This is also partly a function of how DRR and CBA are implemented: programs, projects, and their budgets are spatially restricted and focus on localities at the expense of understanding networks of power and systems of economic and political determinants of vulnerability.

At the macro (national and international) scales, there are other, more quantitative, initiatives to assess vulnerability to natural hazards and/or climate change. These tend to make estimates or use secondary information to assess the numbers of people exposed to relevant hazards of a specified territory, the type and potential intensity of those hazards, the likely effects on different components of the physical structures of that territory (e.g., infrastructure, built environment, health and education facilities, economic assets), and the degree to which mitigation and preparedness measures are in place for the relevant hazards. A recent approach is the World Risk Index (UNU–EHS 2014); others have included a “hot spot” approach that identifies locations that are “vulnerable” to particular extreme events (e.g., World Bank 2005).

What is the opposite of social vulnerability?

If social vulnerability is reduced, what is there instead? Is it capacity, capability, or resilience? This issue forms the basis for a great deal of discussion. One of the first challenges to the use of the term “vulnerability” came from people and organizations that considered it conveyed the idea of victimhood and portrayed vulnerable people as having little capacity of their own. As a result of this critique, CRA and vulnerability analysis claimed to respect and measure capacities (sometimes capabilities) rather than vulnerability alone. While the concern over labeling is understandable, it produces confusion as to whether vulnerability and capacity are conceptually different, or whether they lie on a scalar continuum with vulnerability (likelihood of being harmed) at one end and capacity (less likelihood of being harmed) at the other. Where the two are regarded as distinct, it becomes conceptually difficult to determine the basis of the difference between them.

More recently, the issue has been further complicated by the rise of the use of the term “resilience,” and more specifically of “social resilience.” Like “vulnerability,” “resilience” has origins in a number of disciplines (especially psychology, engineering, and ecology) and a very large literature both about the conceptualization involved and its application in policy. It has become standard terminology for a number of important international organizations and some aid donors (notably the UK Department for International Development, or DFID). It is
SOCIAL VULNERABILITY AND ENVIRONMENTAL HAZARDS

also incorporated into much of the academic and policy discussion about climate change. Whether resilience constitutes the basis for a separate conceptual and analytic approach is a significant debate. The term is becoming widespread, despite concerns that it is often used to avoid the need to determine causation. There is a danger that using the term resilience suppresses the need to ask why people need resilience and what has prevented them from being resilient already. There is considerable discussion of this, especially in regard to the conceptual difficulties of playing down vulnerability in favor of resilience in the social ecological systems framework (Béné et al. 2012; Béné 2013; Cannon and Muller-Mahn 2010; Miller et al. 2010).

Conclusion

As with many concepts in the social sciences, social vulnerability is located within debates and competing notions of causation (of disasters) and the relative significance of different types of attribution, framings (of the way in which hazards, social factors, and processes and exposure to risks interact), analytical traditions (geography, ecology, sociology), and political bias (the validity of social construction). The current rise of resilience as a means of framing preparedness for both disaster and climate change is likely to continue, mainly because of its adoption by donors and governments, and its influence on sources of funding. The main problem of whatever conceptual framework is the degree to which it allows for analysis of causation: of the reasons why people are harmed by hazards and climate change. If resilience as an approach can protect that necessity, it may become valid and have the potential to integrate development (improvement of human wellbeing and reduced vulnerability to risks of all types), climate change adaptation, and disaster risk reduction (through preparedness and prevention). Until 2015 these three areas have been dealt with internationally by three entirely separate United Nations processes: the Millennium Development Goals (MDGs), the Kyoto Protocol (under the United Nations Framework Convention on Climate Change, or UNFCCC), and the Hyogo Framework for Action (HFA). In 2015 each of these was revised. HFA has been replaced by the Sendai Framework for Disaster Risk Reduction, following the Third UN World Conference on Disaster Risk Reduction at Sendai. The MDGs have been replaced by the Sustainable Development Goals, and the Twenty-first session of the Conference of the Parties of the UNFCCC led to a new climate change agreement at the Paris meeting in December. Although there are signs of greater integration across these areas, they are still regarded by many as weak and inadequate. For geography and its academic and research contribution, there is a danger that political and funding preferences of donors and governments will continue to channel efforts into a parallel set of separate activities, despite the significant existing research that argues for integration across the three areas. Whether this is done in relation to a concept of social vulnerability or of resilience matters less than ensuring that the analysis is integrated and that it insists on understanding the social causation of vulnerability.

SEE ALSO: Climate change policy; Community; Construction of nature; Environment and development; Environment and resources, political economy of; Environmental hazards; Environmental risk analysis; Geomorphic hazards; Global climate change; Global environmental change: human dimensions; Natural hazards and disasters;
SOCIAL VULNERABILITY AND ENVIRONMENTAL HAZARDS

Power and development; Social resilience and environmental hazards; Urban resilience; Vulnerability; Weather, extreme

References


Social-ecological transformation

Ulrich Brand
University of Vienna, Austria

Markus Wissen
Berlin School of Economics and Law, Germany

Social-ecological transformation is an umbrella term which describes political, socioeconomic, and cultural shifts resulting from attempts to address the socioecological crisis. Under this conceptual and epistemic heading, such terms as green, great or social-ecological transition, great or societal transformation, green economy, and sociotechnical transition have increasingly come into use. Their goal is to provide a comprehensive understanding of current global environmental change and to contribute to a social and political strategy for dealing with the crisis. Research programs for the social sciences have been oriented accordingly (Hackmann and St Clair 2012).

The concepts and related debates have gained specific importance, first, due to the increasing acknowledgment that existing sectoralized and top-down forms of global or regional environmental management have failed and, second, in view of the “multiple crises” of the financial system, the economy, nature, energy provision, and food. There is a strong consensus in the literature that profound societal changes will be required in order to get a grip on these multiple crises. In the context of this consensus, however, the socioecological crisis is approached from the positions of divergent normative interests and different theoretical perspectives, with the result that a variety of different and even contrasting analyses and proposals concerning the ways out of the crisis have emerged (Brand et al. 2013).

The following noncomprehensive overview first addresses important flagship reports issued mainly by political organizations and think tanks. These are seen as indicators of a shift of perception within state apparatuses on various spatial scales and as an attempt to shape the corridor of social-ecological transformation. Second, key aspects are presented of the academic debate which have to some extent influenced these political flagship reports, but have also raised more fundamental theoretical and political questions. In the outlook, conclusions are drawn regarding a critical perspective on social-ecological transformation.

Social-ecological transformation in flagship reports

1 The concept of a green economy, which puts forward the claim of being able to overcome the economic and ecological crisis, was promoted before and around the Rio + 20 Conference in June 2012. The United Nations Environment Programme (UNEP) started its Green Economy Initiative in 2008. In 2011, it issued a comprehensive report in which it identified a “widespread disillusionment with our prevailing economic paradigm, a sense of fatigue emanating from the many concurrent crises and market failures experienced during the very first decade of the new millennium, including especially the financial and economic crisis of 2008. But at the same time, we have
seen increasing evidence of a way forward, a new economic paradigm – one in which material wealth is not delivered perforce at the expense of growing environmental risks, ecological scarcities and social disparities” (UNEP 2011a, 1).

In line with the United Nations Development Programme (UNDP), the United Nations Department of Economic and Social Affairs (UN DESA) argues for a great green technological transformation, with a scaling up of clean technologies, waste reduction, and sustainable agriculture. In its report with the same title, it states that a green economy “embodies the promise of a new development paradigm, whose application has the potential to ensure the preservation of the Earth’s ecosystem along new economic growth pathways while contributing at the same time to poverty reduction” (UN DESA 2011, v).

The European Commission has developed a plan for sustainable growth: the promotion of a resource-light, ecological, and competitive economy. In a communication of September 2011, the commission stated the necessity to fundamentally transform the European economy within the time span of one generation. It saw reducing resource use and increasing resource efficiency as key mechanisms for coping with environmental problems and resource shortages and, at the same time, strengthening European competitiveness (European Commission 2011; for a similar approach see the green growth strategy of the OECD 2011).

An example for an initiative at the national level is a report by the German Federal Government’s Advisory Council on Global Change (WBGU) entitled “Social Contract for Sustainability.” Its plea for a great transformation (WBGU 2011) specifically refers to Karl Polanyi’s concept in which he explains the passage to industrial capitalism during the nineteenth century, in order to emphasize the magnitude of the socioecological challenge. One point of departure is the assumed global transformation of values toward a sensitization for ecological questions (WBGU 2011). In order to promote and strengthen this transformation, the report states that a new “global social contract” (WBGU 2011, 8, 276) is needed. Central to realizing the great transformation, along with the transformation of values, are “pioneers of change” (such as innovative individuals, NGOs, and companies in all sectors of society and the economy) and a “proactive state” (WBGU 2011, 203). The latter is to create an adequate framework for change agents and to promote required innovations.

The common denominator of these reports and strategy papers is that they consider economic growth desirable, necessary, and capable of reconciliation with the environment. They express, first, a belief, akin to that which prevailed at the beginning of the sustainable development discourse in the early 1990s, that comprehensive win–win situations can be created; and second, firm trust in the existing political and economic institutions and elites, which they see as both able and willing to guide this process.

The shortcomings to particularly focus on are: first, the concepts argue for strong regulatory frameworks and thus neglect the prevailing power relations. In the current crisis, regulatory frameworks tend to develop in an authoritarian direction in order to secure access to resources for particular countries or regions (Lander 2011). Moreover, the economy which is to be politically regulated is mainly understood as the formal capitalist market economy. Accordingly, gender perspectives and their focus on social
reproduction and reproductive work are largely absent in the debate about a green economy (Deutscher Frauenrat 2011).

Second, whereas a relative decoupling of economic growth from resource use and environmental impact can be observed in many advanced capitalist economies, it is far from clear how the necessary absolute reduction might be achieved within the paradigm of economic growth. Since 2008, strategies for coping with the multiple crises have not gone hand in hand with the reorientation of production and consumption patterns designed to promote sustainability. They thus face the danger that improvements in resource efficiency may be overcompensated by the quantity of resources consumed in a growing economy (UNEP 2011b).

Third, neoliberal open-market policies and fierce competition have led to deindustrialization in many countries of the Global South. What is reasonable from a neoclassical perspective is that production that takes place where it is economically most efficient has pushed many countries into the new/old strategy of resource extractivism (Lang and Mokrani 2013). In most countries in Latin America, even in Brazil and Mexico, this seems to be the only viable development strategy capable of alleviating poverty. And it is the flip side of the green economy, since the rare Earth metals needed for green high-tech products mostly come from the countries of the south.

Finally, in addition to the universalization of the Western model of production, globalization implies what can be called an “imperial mode of living” (Brand and Wissen 2013). The logic of globalized liberal markets is reflected in the everyday practices in which access to cheap and often unsustainably produced commodities and labor power are normalized. Currently, this logic is being universalized among the upper and middle classes of newly industrialized countries. The social relations underlying the imperial mode of living, and possible ways for overcoming them, have hardly been challenged at all during this crisis and have been insufficiently reflected in the reports cited above.

The academic debate about social-ecological transformation

Aside from the political-strategic debate, there is also an academic debate over social-ecological transformation. Most of the approaches here have a longer history. Nevertheless, like the political debate, the academic one has gained momentum in the context of the current multiple crises. The various contributions emphasize, first, that socioeconomic, political, and cultural changes have to go beyond incremental steps and toward particular policy fields, such as climate change or biodiversity policies. Second, transformation is understood as a manifold nonlinear process, since it deals with dynamic, multidimensional, and complex systems as well as potential tipping points. Third, it is acknowledged that technical innovation is necessary, but not sufficient, while social innovation is central to social-ecological transformation (Brand et al. 2013).

Within this framework, several approaches can be distinguished, as follows.

1 The concept of social metabolism/socioecological transition developed in the interdisciplinary context of the Vienna Institute of Social Ecology. It conceptualizes social-ecological transformation from the perspective of the use of energy and material. The history of humankind is understood as a succession of “sociometabolic regimes” that differ in their energy sources and in the “colonization” of nature. Hunters and gatherers, the first sociometabolic regime, relied on the solar energy stored in the plants and animals available on their territory, but did not
social-ecological transformation

systematically influence the reproduction of these resources. By contrast, agrarian societies, while still relying mainly on solar energy, started to systematically intervene in nature – to “colonize” it – in order to enhance its productivity.

Industrial societies, the thus far last stage of human development, invented increasingly sophisticated and, at the same time, destructive forms of the colonization of nature. Even more important, they replaced reproducing biomass as the main source of energy with fossil fuels, which now, given the prevailing patterns of production and consumption in the Global North and the fact that two thirds of the world’s population are currently in transition from the agrarian to the industrial regime, is imminently threatened with exhaustion. Given further that the remaining fossil resources cannot be burned without exacerbating global warming, the social metabolism of industrial society today faces a fundamental crisis (Haberl et al. 2011). The current pattern can thus be understood as a “structural exhaustion of opportunities” (Fischer-Kowalski 2011) of the existing sociometabolic regime. It has given rise to a social-ecological transformation which may either take on the form of a catastrophic break with industrial metabolism or, if the patterns of energy generation and consumption are radically changed, may result in a new, sustainable sociometabolic regime.

Whereas the concept of social metabolism is mainly concerned with the physical basis of the social-ecological transformation, transition research and management focus on societal and institutional aspects as well as on technological and social innovation. Furthermore, their temporal and thematic scope is significantly narrower than that of the social metabolism/socioecological transition approach. Starting from the analysis of concrete transition processes in such sectors as energy and agriculture, transition research has developed a “multilevel concept” of major societal shifts toward sustainability (Verbong and Geels 2010), according to which transitions often originate in societal “niches,” then spread to the level of “regimes” (institutional structures), and finally contribute to transforming “landscapes” (the overall social, political, economic, and cultural setting). Radical innovation is considered to take place primarily in niches, while at the mesolevel of regimes, changes occur more incrementally because of path dependencies and lock-in processes. The interplay of the three levels is key to sustainability transitions which are understood as “long-term, multidimensional, and fundamental transformation processes through which established sociotechnical systems shift to more sustainable modes of production and consumption” (Markard, Raven, and Truffer 2012, 956). Transitions can be the results of evolutions and/or of clear-cut goals.

Transition management aims to utilize the findings of transition research in order to inform and shape the governance of concrete reform processes (Kemp, Loorbach, and Rotmans 2007). Collaboration between actors and learning processes is fundamental. Governance can influence cultures and discourses, actors and structures, as well as innovations in order to accelerate and trigger transitions. However, command and control strategies are not possible owing to the complexity and uncertainty of transition processes. The transition management approach sees its role in the debate
on social-ecological transformation as providing methodologies which are of practical and policy relevance (Wesely et al. 2014).

Transition research and management has been criticized by proponents of approaches in the tradition of practice theories (see Røpke 2009 for an overview). According to these theories, there is a producer and manager bias in transition management. Consumers and complex configurations of everyday life are treated more or less as external to the system of innovation. However, according to Shove and Walker, they are a constitutive part of it. Neglecting them conceptually is like an “act of writing to an audience that might not be listening” (Shove and Walker 2008, 1012). The crucial concept for a better understanding of the causes of the socioecological crisis, and for discerning possible ways out of it, is that of social practice. It refers to shared behavioral routines constituted by sets of interconnected elements: the social and political institutions that facilitate them, the sociotechnical configurations, such as the physical infrastructures, that enable them, the available knowledge and prevailing symbolic orientations that, consciously or not, guide them, and the forms of power that are inscribed in all these elements (Spaargaren 2011).

Because of its habitualized character, an environmentally detrimental social practice, such as driving a car, is only to a very limited extent accessible to intentional steering and management or to consciousness-raising campaigns. This makes social-ecological transformation a far more complicated process than it is assumed to be in transition research and management. It is a process which cannot be influenced from any preferential entry point, but which has to address the various elements which constitute social practices (Shove and Walker 2010). Overcoming automobility, to take this example, would require an understanding of driving, not only as a form of movement with the intent of relatively rapidly overcoming a distance, but also as an issue which has to be addressed in terms of the prevailing and power-laden conceptions of progress, freedom, and masculinity and their institutional and infrastructural manifestations.

The central motives and arguments in the context of the degrowth debate maintain that the orientation toward economic growth as the crucial point of reference of economic policy and as an indicator of prosperity and quality of life no longer holds (Kallis 2011). The issue of the suitability, or lack thereof, of markets as a mechanism for dealing with ecological and social problems is another core point of commonality. Some authors argue for an internalization of ecological and social costs; others go further and add that more structural changes as well as a decolonization of economics and of our minds from the domination of economism, and a move toward a different collective imagery, are the preconditions for meaningful change. Degrowth is “a multifaceted political project that aspires to mobilize support for a change of direction, at the macro-level of economic and political institutions, and at the micro-level of personal values and aspirations. Income and material comfort is to be reduced for many along the way, but the goal is that this is not experienced as welfare loss” (Kallis 2011, 878). Normative principles, such as cooperation and social justice, are being reintroduced, while social movements are seen as the major subjects of change. Many contributions to the debate
do not focus so much on crises or secular trends of diminishing growth rates in highly industrialized countries. Rather, they propose a “voluntary, smooth and equitable transition to a regime of lower production and consumption” (Schneider, Kallis, and Martinez-Alier 2010, 511). Degrowth is thought of as a conscious societal process based on a change of values.

Contributions from a critical geography/political ecology perspective (Robbins 2004; Perreault, McCarthy, and Bridge 2015) differ from the approaches described above in focusing more explicitly on issues of power and domination. Where the social metabolism approach addresses mainly physical materiality, political ecology also takes into account the materiality of social structures. In political ecology, the terrains and processes of governance, which transition management tries to shape in order to facilitate sustainability transitions, are less understood as solutions than as part of the problem. Like practice theories, political ecology focuses on the reproduction of social relations, but also addresses the contradictions inherent in them and, unlike the degrowth debate, the political ecology perspective sees economic growth as a social relation intrinsically linked to societal domination that reproduces social structures.

From a political ecology perspective, nature is societally – that is, socioeconomically, culturally, and politically/institutionally – produced and appropriated. The focus is not on “the environment,” but rather on the social forms of the appropriation of nature: that is, the forms in which such basic social needs as food and housing, mobility, communications, and health and reproduction are satisfied. This is not to deny the material peculiarities of biophysical processes for they are, under certain circumstances, no longer reproducible, but they are shaped by society. And conversely, the materiality of nature shapes societal processes. Importantly, the production of scale is considered crucial in transforming the conditions of access to natural resources and reshaping societal nature relations.

Political ecology argues that the metabolism of human society with nature, which is essentially mediated by labor, assumes a particular form in capitalist society: the production of use values for the sake of exchange value and/or profit; a hierarchy between capital and wage labor as well as other forms of labor; and, moreover, the development of a modern state separated from the capitalist economy and the class relationships. The dynamics of the capitalist economy consists of the valorization of human labor power and of nature.

Therefore, a crucial assumption regarding social-ecological transformation is that in modern capitalist societies, change takes place continually. “The bourgeoisie cannot exist without constantly revolutionizing the instruments of production, and thereby the relations of production, and with them the whole relations of society. Conservation of the old modes of production in unaltered form, was, on the contrary, the first condition of existence for all earlier industrial classes. Constant revolutionizing of production, uninterrupted disturbance of all social conditions, everlasting uncertainty and agitation distinguish the bourgeois epoch from all earlier ones” (Marx and Engels 1998/1848, 243). The decisive question is what kind of logic of transformation is to predominate.

Outlook

For a theoretically adequate understanding of transformation, it is useful to link political ecology with critical political economy and social
theory, especially critical state and hegemony theory. In so doing it can be shown that capitalist societies, with their tendencies to destroy their own material foundations, can in certain ways develop stabilizing forms of the societal appropriation of nature. The societal regulation of interaction with nature is possible and does in fact occur; herein lies a central dynamic of politics (Görg 2011). The regulation of societal nature relations does not imply abolition of the largely destructive forms of appropriation of nature. However, the destruction of nature will not necessarily become an urgent problem for overall capitalist development, since dangerous negative impacts can be spatially externalized and temporarily postponed. This can be seen in climate change, many effects of which will occur in the future; those that are indeed manifested in the present usually occur in more vulnerable, peripheral places. Crises will particularly occur at the local and regional levels – or are already occurring there today. However, that fact does not necessarily call the fundamental structures and developmental dynamics of capitalism into question. With regard to a possible scarcity of resources, we can also see that in the interplay of fears of global scarcity and local valorization strategies, the regulation of societal nature relations today means new exploration for tar sands, fracking for natural gas in slate formations, energy crops which involve the control and utilization of land, or a partial switch to solar energy. Insights into the changing forms of capitalist regulation help to understand the direction of capitalist development, for example, toward a selective greening of capitalism.

The thus enhanced critical concept of transformation focuses on complex societal and social-ecological relations and, in particular, on their dominant development dynamics. Moreover, it focuses on structures and processes by means of which society organizes its material foundations, including its metabolism with nature – socioeconomically, politically, culturally, and subjectively.

Such an analysis would consider the structure and power of sustainability discourses (Brand 2010) and the tendencies toward the “neo-liberalization of nature” (Castree 2008), that is, the shifting politico-economic and sociocultural dynamics of the appropriation of elements of nature. And it would acknowledge the still powerful structures, interests, and instruments of financial market capitalism. It would ascertain that in spite of all tendencies pointing toward greater sustainability, the state and the international political institutional system have tended to reinforce the dominant conditions and developments. The term “imperial way of living” (Brand and Wissen 2013) identifies a determining factor why very little is happening politically, along with such other factors as power strategies, including repression of criticism and alternatives, and political co-optation.

Again, this has political-strategic implications. First, research into social-ecological transformation needs to consider and evaluate the various strategies and possibilities for dealing with the multiple crises, that is, business-as-usual or more authoritarian alternatives, an imperial deepening of global fragmentation, social-democratic steering at various spatial levels, or more radical-democratic alternatives.

Second, analyzing hegemony, capitalist regulation, and its social forms means considering how the corridor of both top-down and bottom-up alternatives tends to be systematically narrowed down to a form of capitalist ecological modernization. It remains to be seen whether projects like the greening of the economy or green capitalism will be potentially capable of ushering in a new accumulation dynamic by changing the energy and resource base.
The question of a democratic shaping of society and of societal nature relations would appear crucial. That implies the democratic control of resource use, but also of the manifold processes of production and consumption. This is an important research perspective to determine what the already existing democratic forms of resource control are, which struggles will be necessary in order to generalize them, and how they are to be stabilized institutionally. It must also be determined which demands can be made in a comprehensive sense for the democratic structuring of society’s interaction with nature and to what extent the concrete strategies for a green economy or a green new deal have a supportive effect or not. Taking the perspectives presented into account, it would be necessary to evaluate whether, and to what extent, a “passive revolution” in the form of an eco-capitalist modernization might take place in response to the multiple crises and how it could be addressed from an emancipatory perspective.

Acknowledgment

The authors wish to thank Phil Hill for his excellent language editing.

SEE ALSO: Commodification of nature; Consumption; Democracy; Political ecology; Sustainable development

References


The concept of socio-nature is used to argue that “society” and “nature” should not be analyzed in abstraction from one another but, rather, that they are inseparable and always already coproductive. It has been promoted as offering a radically reimagined political ecology, rooted in historical materialism but engaging with the processes of production that lead to heterogeneous material and discursive human and nonhuman entanglements. An extension and critique of the Lefebvrian notion of the production of nature, it represents an attempt to create dialogue between Marxist approaches and poststructural approaches such as actor-network theory. Theorists arguing for the concept of socio-nature reject Marxist contentions that nature is the substratum to society, while also contending that nature is not merely a social construction. Instead, focusing on historically and geographically situated materiality helps researchers to see that socio-natural entities are never fixed but are always in a process of becoming. The concept has been applied to diverse contexts, such as local food systems, gardens, and the politics of water resource management, and has been most influential within political ecology.

Erik Swyngedouw developed the notion of socio-nature in the mid-1990s from his readings of, in particular, Karl Marx, Henri Lefebvre, David Harvey, Donna Haraway, and Bruno Latour. Arguing that “society and nature, representation and being” are inseparable from each other in the city (1996, 66), he was critical of much research on the political economy of cities for its tendency to sideline environmental issues. Concurrently, he contended that the existing theoretical vocabulary was insufficient to develop a critical and transformative reading of cities. Although Marx had argued that social development has natural foundations, and that social relations transform the natural environment, so-called natural processes were still being characterized as external to the social. It has been argued that this focus on social transformations of nature might imply that “nature” is “subsumed under social control” (Swyngedouw 1999, 446) and that the separation of social and natural relations hides the processes that produce hybrids. In contrast, Lefebvre’s notion of the production of nature embodies material, representational, and symbolic practices; while insisting on the materiality of socio-nature’s component elements, Lefebvre argues that their “content can be approached only through the excavation of the metabolism of their becomings” (Swyngedouw 1999, 447). As such, he emphasizes process and flux as opposed to ontological essentialism. Swyngedouw’s significant theoretical intervention was to draw this Marxist literature together with the poststructuralist writings of authors such as Latour and Haraway to analyze how cities are produced as “hybrids” or “cyborgs.” Latour (1993) had famously argued in We Have Never Been Modern that “nature” and “society” are merely labels – the outcomes of processes of “purification” associated with modernity, and that this could be studied empirically by tracing the networks (involving human and non-human actants) through which “quasi-objects” or “hybrids” had emerged. Authors who have
employed the concept of socio-nature, however, argue that it is not sufficient to merely follow actants as they come to be enrolled in networks, as this could hide the historical-geographical processes through which they were produced. While there is nothing intrinsically ahistorical about actor-network theory, it is often associated with ethnographic work that traces networks in-the-becoming rather than after the event. Instead, proponents of socio-nature argue that Latour’s approach of following actants could be combined fruitfully with Lefebvre’s focus on temporalities. In such ways, the concept has been most influential as a dual critique of research on political economy that views “nature” as a mere backdrop to “society,” and of broadly poststructural research that has emphasized processes of becoming in the present, rather than histories of the processes of production that underlie them.

Through its roots in Marxist theory, work on socio-nature builds on the two related notions of “processes of production” and “metabolism,” which are explored in turn here. The focus on processes of production puts labor at the heart of analyses of socio-nature. In one sense, this is very similar to Neil Smith’s work on the production of nature, wherein nature is materially transformed through capitalist relations rather than preceding them. Here, a socio-natural analysis highlights the social in what might appear ostensibly natural. Beyond this, though, analyses emphasize the entangled “networks of power” that enroll and reconstitute “parts of nature” (Swyngedouw 2007, 10). The question, in other words, is not simply about what labor was involved in producing particular socio-natures, but also over the emergence of sociopolitical struggles, the mobilization of elite groups and concurrent “discursive and material enrolements of nature” (Swyngedouw 2007, 10). For instance, adopting this approach in a study of the genetic modification of plants would lead to a focus not only on the appearance of new plant forms. Instead, such an analysis might examine the uneven relations in the plants’ production, for instance contrasting the wealth and power of large multinational companies that research, develop, and patent seeds with the relative lack of power of small-scale African farmers. Such research might also examine the discourses that surround and entwine these socio-natures, exploring their changing meanings, values, and uses. In this sense, a genetically modified plant is never just a plant but it is an embodiment of historical-geographical processes that are always in flux and transience.

Metabolism was at the heart of Marx’s (1970, 283) conceptualization of labor, which he viewed as “a process between man and nature, a process by which man, through his own actions, mediates, regulates, and controls the metabolism between himself and nature.” However, Swyngedouw (2006, 109) argues that much Marxist analysis during the twentieth century sidelined these society–nature connections, viewing labor as merely social and, through this, ignored material metabolic relationships and partially blinded social sciences to questions of political ecology. He therefore promoted a re-engagement with the metaphor of metabolism, viewing metabolic circulation as “the socially mediated process of environmental – including technological – transformation and trans-configuration, through which all manner of ‘agents’ are mobilized, attached, collectivized, and networked” (Swyngedouw 2006, 113). Processes of metabolic circulation lead to the production of heterogeneous assemblages. While some authors view these assemblages as in many ways similar to Latour’s “hybrids,” those arguing for socio-nature claim there are fundamental differences. First, they contend that “hybrid” is suggestive of a mere mixing of ontologically separate or pure things, whereas metabolic circulation stresses that the constitution of cities
is intrinsically heterogeneous. Second, they are critical of Latour for being dismissive of the question of power; in actor-network theory, “power” is held and negotiated through networks and might become apparent through the research process, while analyses of socio-nature might take uneven power relationships as their starting point. While heavily influenced by actor-network theory, therefore, socio-nature cannot be viewed as an easy reconciliation of historical materialism with poststructuralism, but rather as an ongoing critical dialogue.

While offering a lens through which to view society–nature relations in general, work on socio-nature has tended to focus on specific projects, issues, or places. Through this geographical specificity has emerged a concern about the relationship between socio-nature and scale. Following Neil Smith’s (1984) contention that the process of producing scale can involve producing nature, Swyngedouw (2007, 11) maintains that “scale is not ontologically given, but socio-environmentally mobilized through socio-spatial power struggles.” By conceptualizing scale as emerging through relational processes, rather than as pre-given, socio-nature is again influenced by poststructuralists such as Latour, who critiques polarized ontologies of local versus global. In Swyngedouw’s reading, the scalar constitution of sociospatial relations forms through the forging of relational networks, which exist in dialogue with produced territorial configurations. In other words, socio-nature is not only an attempt to move away from artificial social or natural abstraction, but to see metabolic circulation as intrinsically embedded in a scalar politics. He grounds these ideas in his examination of changes in the Spanish waterscape under the Fascist political regime of General Franco. Under Franco’s direction, over 600 dams were built in Spain, leading to a dramatic shift in the country’s hydrological landscape. In this research, Swyngedouw shows how the techno-natural arrangements of rivers and dams produced new socio-natures; indeed, he contends that it is not possible to understand any form of “social” change in Spain without concurrently understanding changes to the hydrological process. They are, in other words, indistinguishable from each other. Central to these socio-natural changes was a particular scalar politics – also intrinsically bound up in metabolic circulations. This research shows how Franco’s physical reorganization of the country’s hydrological networks went beyond the practical goals of developing a reliable water supply and producing hydro-electricity and rested on: (i) a desire to eradicate regionalism; and (ii) a need to expand its geopolitical outlook in order to attract support and financial aid from the United States for the construction of dams and watercourses. In relation to the former, the material connections between once separate river basins, and the interbasin transfer of water, were designed in part to both integrate national territory and produce a coherent national cultural identity. In relation to the latter, for the hydrological transformation to take place, it was not possible for Spain to maintain a resolutely nationalist outlook. In this example, it becomes apparent that the metabolic circulations that produce socio-natures concurrently disturb essentialized notions of scale.

Literature on socio-nature often refers to the closely related notion of techno-nature, with some authors using the terms interchangeably. There is, in other words, some conceptual fluidity around the term. In some cases, “techno-nature” stands as shorthand for antipurification – an acknowledgment that nature is increasingly transformed by the interference of technology. Others argue that the concept has scope beyond this, suggesting that it denotes not only a quasi-object but also an ideology. Research on socio-nature has tended to employ
SOCIO-NATURE

the notion of techno-nature more specifically to refer to the enrollment of technologies by elite groups for the mobilization of particular socio-natures. Here, technologies provide some of the tools through which socio-natures are transformed, while also embodying particular socio-natural relations.

Research associated with socio-nature has tended to focus on its conceptual development and its practical applications, rather than on methodological implications and innovation. As a result, studies have drawn heavily on methodological approaches that developed alongside some of socio-nature’s theoretical influences, such as actor-network theory. While much research adopts a broadly similar “follow the actor” approach to that suggested by Latour, and while authors such as Swyngedouw acknowledge the agency of nonhumans, the focus of socio-nature analyses is often more explicitly on human actors. However, the concurrent and continuing focus on historical materialism frequently leads to detailed historical empirical research, documenting decisions around particular schemes, and tracing the uneven power struggles that lead to particular developments being rejected or achieving success. Swyngedouw’s examination of the (re)production of the Spanish waterscape offers a particularly clear example of such a historical focus. At the center of his analysis of these socio-natures in Spain is a focus on power. First, he is interested in how elite groups came to be in a position that allows them to reconfigure techno-natural arrangements. Through this, he emphasizes that Spanish society has organized itself partly in response to its desire to control water. Second, Swyngedouw explores the ways in which the relations bound up in socio-natures shift when there are changes to discourses around water, uses of water, management structures around water, or the materiality of water itself. In producing this analysis, he produces a forensic reconstruction of the events surrounding the waterscape’s gradual reconfiguration, concurrently detailing the physical and political (in its broadest sense) changes.

Socio-nature research is not only based on archival work, however. First, it increasingly draws on interviews and ethnography. Alkon (2013), for example, has examined how advocates of organic agriculture acknowledged the food it generates as the product of labor, but are likely to celebrate the role of farmers while ignoring the role of other (often immigrant) farmworkers. Rather than attempting to reveal a political truth, Alkon instead emphasizes the emergence of situated socio-natures and examines the ways in which labor practices and the agency of socio-natural entities are selectively rendered visible or invisible. Such work steers future research on socio-nature to explore the processes through which purification occurs, especially through the hiding of the labor process in socio-natural entities, rather than attempting to (merely) reveal the hybrid entanglements of socio-natures. Second, recent work on socio-natures has more explicitly engaged with flows and movements (the materiality of metabolic circulations). Linton and Budds (2013), for instance, emphasize the agency of water and point toward a fuller recognition of the political as not only encompassing but also potentially being driven by the nonhuman. This research has stressed water’s “unruliness,” which can result both in efforts to regulate and order it, and in its active disruption of social power structures. These authors have developed the related notion of the hydrosocial cycle to further highlight the inseparability of water’s social and physical dimensions. While emphasizing the intrinsically political nature of water, the concept’s resemblance to the widely known hydrological cycle also emphasizes the physical power, flow, and changing materiality of water.
This emphasis on the agential role of water itself addresses criticisms of work on socio-nature as being ontologically imbalanced toward the “social.”

In spite of the conceptual advancements associated with “socio-nature,” various criticisms have been leveled at its application to empirical contexts. First, while the concept stresses the demolition of nature/society dichotomies, addressed through the study of “materially discursive practices” (Swyngedouw 1999, 447), some commentators have criticized a tendency to focus on how quasi-objects such as water embody human political tensions and human social inequalities. In contrast, it might be argued that socio-nature retains many of the central arguments of actor-network theory but develops these in a theoretically coherent manner that responds to wider criticisms claiming that actor-network theory produces analysis that is more descriptive than transformative. A second criticism leveled at socio-nature is that it “redraws the very binary it seeks to upturn, naturalizing either the natural or social, while treating the other as something to be demonstrated through research” (Alkon 2013, 667). In other words, much of the applied research associated with socio-nature begins either with something ostensibly natural (and demonstrates how it is also social) or with something ostensibly social (and demonstrates how it is also natural). So, while focusing conceptually on historically and geographically situated heterogeneous assemblages, the concept’s application has (ironically, given its criticism of Latour’s hybrids) been critiqued for a tendency to either reproduce binaries or place greater emphasis on one “side” than the other.

**SEE ALSO:** Actor-network theory; Hybridity; Nature; Political ecology; Production of nature; Scale

**References**


The solid phase of soils contains both organic and mineral matter. Except for the small percentage of world soils composed of organic materials, most soils are dominated by mineral matter. Soil minerals provide structural support and the initial intergrain porosity necessary for adequate permeability and water storage. Soil minerals also provide a major contribution to soil reaction and mineral nutrition to plants. They function geochemically to buffer drainage waters, absorb and exchange nutrients and pollutants, and even function biologically as a home to microscopic organisms. Soil minerals are useful pedologically because they greatly impact soil development pathways and processes. Thus, they can be used to provide evidence of past soil-forming processes or used to determine the provenance of regolith (soil parent materials). The significance of soil minerals to the character of soils is widely recognized by land managers, who may use colloquial terms for the soil minerals, and by pedologists, who recognize mineralogy at various levels in soil taxonomic systems. Both the USDA’s Soil Taxonomy (USDA 1999) and the FAO’s World Reference Base for Soil Resources (FAO 2015) use soil mineralogy at various levels, ranging from the highest to lowest levels of the systems.

Soil minerals can be either primary or secondary. Primary minerals are those that originate in a rock-forming environment and are often thermodynamically unstable in the soil system and, thus, eventually become altered by chemical weathering. Secondary minerals are those minerals that form in the soil environment and are, therefore, more thermodynamically stable in the environmental conditions found in soils. Nonetheless, it is important to note that changing environmental conditions through time may lead to unstable secondary soil minerals as well.

Mineral groups present in soils

Minerals are classified by their chemical composition and crystalline structure. Seven major groups of minerals are recognized in nature, all of which may occur in soils, but their prevalence in any soil geographic region is a function of factors such as the origin of the regolith the soils have formed in (parent material), climate control on mineral formation or degradation, and the age of the soil, which impacts how weathered the soil mineral suite will be. The mineral groups described here occur over large areas of the globe or are important to understanding soil formation and behavior.

Halides, sulfates, and other very soluble minerals are composed of elements ionically bonded as salts. Common soil minerals in this group are halite (NaCl) and gypsum (CaSO₄•2H₂O), which can occur in parent materials, but more typically form in the soil environment under dry climatic conditions. In soil classification, the importance of these minerals to formation and use of the soils is recognized by salic and gypsic diagnostic horizons in Soil Taxonomy, and by reference soil groups (RSGs) such as Gypsisols in the World Reference Base (WRB). Soil management issues arise when using these soils for irrigated agriculture, because these salts...
are often toxic to plant growth or can cause nutrient uptake imbalances. Drainage of salty irrigation waters may also cause ecosystem stress in receiving basins.

The most common sulfide mineral in soils is pyrite (FeS₂), which may occur in geological deposits, but also forms in wet coastal settings such as tidal flats. Oxidation of pyrite leads to the formation of sulfuric acid, which produces extremely acid conditions (pH < 4) in soils, severely limiting plant growth. Such soils are known as acid sulfate soils. Geographically, these soils occur in low-relief marine coastal regions and present challenges to agricultural and building development, largely because the sulfates become more pronounced after the soils are drained and aerated.

Carbonate minerals such as calcite (CaCO₃) and dolomite ((CaMg(CO₃)₂) are extremely widespread and important soil minerals, and are geographically diverse in their occurrence. Carbonate minerals occur as primary minerals in many Pleistocene-aged, glacial and eolian deposits of mid-latitudes. In humid climates, these minerals are leached from soils by acidic, percolating soil waters. The depth of dissolution and leaching of these minerals acts to delineate the boundary between genetic soil horizons and the (C horizon) parent material. Calcite also occurs as a secondary mineral in soils of semi-arid and arid regions, forming in situ or accumulating as wind blown dust and washed into the soil by rain. Forms of secondary calcite accumulation in soils have been studied quite extensively, enabling their use as calibrated and relative dating features in soils in the southwestern United States and elsewhere. Classic studies have recognized up to six stages of morphological expression of secondary carbonate, ranging from accumulation of minor masses and pendants to entire horizons completely engulfed and cemented to the point that additional deposits of secondary calcite accumulate in a laminar fashion on top of the cemented horizon (Gile, Peterson, and Grossman 1966; Machette 1985). These studies show that, for many soils, the accumulation of secondary calcite and carbonates is time-dependent and thus related to the age of the land surface. Evaporation of capillary water associated with shallow groundwater is another mechanism of calcite accumulation that has historically led to confusion in soil geomorphic interpretations. The occurrence of secondary calcite in various forms is important to land use and soil classification. Soil Taxonomy recognizes calcic and petrocalcic diagnostic horizons, which impact classification at several taxonomic levels, while WRB recognizes the RSG Calciols and the qualifiers calcic and calcareous.

With few exceptions, calcite is the dominant secondary carbonate mineral in soils, making secondary carbonates in soils a significant carbon sink in the terrestrial carbon cycle. Estimates are varied and dependent on calculation methods, but secondary carbonate minerals are believed to account for approximately 38% of total carbon stored in soils.

Oxide minerals are those with a metallic cation bonded with the oxide anion (O²⁻). Oxide minerals comprise a diverse group. They include residual primary and secondary minerals of low solubility, as well as secondary minerals that form and dissolve with changes in soil redox (oxidation–reduction) conditions. The most stable oxide minerals in soils are those formed with the cations titanium and zirconium, namely the minerals rutile and zircon. While only comprising a small fraction of most soils, their extreme low solubility implies that they can accumulate residually in soils that are experiencing chemical weathering. Therefore, they can be used as a stable index mineral to assess parent material uniformity, the degree of chemical weathering when compared to the parent regolith, and to
SOIL AND CLAY MINERALOGY

determine the volumetric changes (collapse or expansion) that soils have experienced during pedogenesis.

Iron oxides are probably the most important oxide minerals in soils. Iron oxide minerals refers to what is more correctly named iron oxyhydroxides and hydroxides, which denotes their structures as consisting of various crystalline and poorly crystalline forms that have hydroxide groups and water molecules as part of their atomic structure. Structural differences produce different physical and chemical properties of the iron oxides and are used as the basis for classifying and recognizing the various iron oxide mineral types. A thorough treatment of iron oxides is not possible here, but recognizing that most red, brown, and yellow colors in subsoils are a function of the type and crystallinity of iron oxides present is critical. They are a very important pigment, especially in subsoil horizons, enabling their occurrence and red-brown colors to be used in environmental assessments of wetland hydrology and for on-site waste disposal. A lack of strong subsoil pigmentation is typically attributed to the lack of appreciable quantities of iron oxides and is thus used to determine that soil redox chemistry has crossed a threshold and enabled the iron minerals to be chemically reduced. Because iron has such a strong affinity for oxidation, the reduction of iron connotes environmental conditions that are limiting the presence of oxygen, while also providing a source of electrons for reduction. Typical conditions for limiting oxygen, and thus creating anoxic conditions in soils, include soil saturation or compaction, while the source of electrons to cause reduction is from microbial respiration. Iron oxides are known to adsorb phosphorus, thus limiting ecosystem productivity and agricultural development; they can adsorb heavy metal pollutants as well.

Manganese is a related transition metal that behaves similarly to iron in redox reactions, but is not considered as important because it is less abundant and is less important as a pigment. Nevertheless, in certain soil systems manganese oxide minerals can be important in the formation of black and bluish pigments, and their presence as coatings in soils provides information about soil redox and/or pH conditions that allow mobilization and precipitation of manganese.

The aluminum hydroxide gibbsite, Al(OH)₃, is the most common secondary mineral that forms in soil environments. Gibbsite exhibits extreme low solubility and is most important in soils that have experienced severe, long-term weathering, such as some tropical soils. Gibbsite and other less-frequently occurring aluminum oxide minerals are not strong pigments and do not behave like iron and manganese, so have received less attention in the literature. Poorly crystalline and amorphous aluminum compounds are important in many soils, and are used in recognizing certain kinds of soil horizons, but are not considered further here.

Taxonomically, oxide minerals in soils are also very important. They are used to define mineralogy classes of soils at even the highest levels of classification, for example, the soil order Oxisols in Soil Taxonomy and the RSG Ferralsols. In addition, pigmentation produced by the presence/absence of oxide minerals is important in recognizing lower levels of taxonomy, such as aquic subgroups in many soil orders, along with the qualifiers in WRB.

Silicate minerals are the largest and likely the most important group of soil minerals. Silicate minerals have as their basic structural organization the silica tetrahedron, formed through the bonding of one silicon atom with four oxygen atoms. Silicate minerals are divided into groups based on how many oxygen atoms
SOIL AND CLAY MINERALOGY

Figure 1  Ball and stick representations of silicate mineral structures, arranged according to increasing sharing of O\(^{2-}\) at the corners of silica tetrahedra: (a) neosilicates, or isolated tetrahedra; (b) sorosilicates, which share one corner of each tetrahedron; (c) cyclosilicates, which share two corners of each tetrahedron; (d) single-chain inosilicates, which share two corners of each tetrahedron; (e) double-chain inosilicates, which share two or three corners of each tetrahedron; (f) phyllosilicates, or layer silicates, which share three corners of each tetrahedron, all in the same plane; (g) tectosilicates, such as quartz, which share all four corners of each tetrahedron in a 3-D network. Source: Adapted with permission from Schaetzl and Thompson (2015). Cambridge University Press, Cambridge.

are shared between tetrahedra, the geometrical arrangement of the linked tetrahedra, and the type and number of other cations bonded into the structure (Figure 1). In general, a greater degree of oxygen sharing leads to greater stability of the minerals and, thus, increased resistance to weathering. Quartz is the dominant silicate mineral in many soils of the world, reflecting its stability and resistance to weathering, which results from a dense network of shared silicon and oxygen bonds. Its resistance to weathering allows quartz to accumulate as a residual product in many highly weathered soils. Like some of the oxide minerals discussed previously, the stability of quartz allows it to be used as a stable index mineral, in comparison to other, more weatherable, silicate minerals.

Many silicates are primary minerals that originated in rock-forming environments, for example, quartz. However, there are some notable examples of secondary silicate minerals that have formed in soils proper. The latter minerals are important to interpreting pedogenic environments or paleoenvironments of
soil formation. The most important secondary silicates are phyllosilicate minerals and opal. Opal in soils is most commonly the result of biogenic formation of phytoliths in plant cells and is used as an indicator of the type of vegetation growing on the soils. Grasses and sedges tend to be the most prolific producers of biogenic opal and, in combination with the study of phytolith forms, can be used to interpret the occurrence of vegetation communities during soil formation.

Perhaps the most important group of silicate minerals in soils is the phyllosilicate minerals, which are also called layer- or sheet-silicate minerals. The terms layer or sheet silicates are commonly used because the structural units of the minerals run in long layers, or sheets, rather than more complicated 3-D crystalline networks typical of other silicate minerals such as quartz. To complicate matters, most phyllosilicates in soils occur in the clay-size fraction (e.g., <2 μm diameter), so the term “clay minerals” is applied to these minerals, even though many other mineral groups also occur in the clay fraction. The reason for the strong interest in these minerals is that they possess a negative electrical charge that has tremendous environmental importance, namely the ability to attract and retain cationic substances, referred to as the cation exchange capacity (CEC). The negative electrical charge is produced by the substitution of cations of lower charge into the structure (e.g., substituting Al\(^{3+}\) for Si\(^{4+}\) in the tetrahedron), resulting in a crystal structure with a negative charge imbalance. The CEC of a soil is of utmost importance to agricultural productivity through the retention of plant nutrients that behave as cations.

Phyllosilicates in soils may be primary or secondary, but it is the secondary mineral types, or clay minerals, that receive the most attention in soil science, because they form in the soil environment. Thus, they reflect a variety of soil-forming factors and environmental conditions, such as duration of weathering and soil solution chemistry as influenced by plant uptake, acid leaching, or drainage water concentration. The most common primary phyllosilicates are chlorite, muscovite, and biotite mica, which are common in igneous and metamorphic rocks, while muscovite mica is also a common constituent in sedimentary rocks. Serpentine is a primary phyllosilicate mineral in soils that originates from ultramafic volcanic rocks. While not widely distributed, where serpentine occurs the soils are often toxic to plants because of low concentrations of ecosystem-essential elements and high levels of toxic metals.

Phyllosilicates are classified on the basis of structural and compositional criteria. The first criterion is how the layers of cations are structurally coordinated with oxygen atoms. The cations are organized in sheets of linked tetrahedra or octahedra. In an octahedron, the cations are bonded with six oxygen atoms, which forms an eight-sided structure. Silicon is the main cation of tetrahedral layers and aluminum is the most common cation in octahedral layers, but these cations are readily substituted, which creates a negative layer charge and may define different mineral species. The broadest grouping of phyllosilicates is as 1:1 or 2:1 minerals (Figure 2). The 1:1 minerals have one layer of cations coordinated as tetrahedra and one layer as octahedra. Kaolinite is the most common 1:1 mineral. The 2:1 minerals have two layers of tetrahedra that are separated by an octahedrally coordinated layer of cations. Illite, vermiculite, and smectite group minerals are all common 2:1 minerals. Other criteria related to crystal structure composition or arrangement are also used by mineralogists to further define mineral species within these groups.

A critical feature used to define different minerals and their change through weathering is the nature of any material in the interlayers between the 1:1 or 2:1 sheets. For example, the interlayers
SOIL AND CLAY MINERALOGY

![Diagram of common soil phyllosilicates](image)

Figure 2  Schematic diagram of common soil phyllosilicates showing tetrahedral and octahedral sheets and type of interlayer materials. From left: 1:1 minerals kaolinite and serpentine; halloysite with interlayer water; 2:1 mica group with interlayer potassium residing in a cavity in the tetrahedral sheet that bonds the layers together; vermiculite with hydrated interlayer cations such as Mg\(^{2+}\); smectite group with expanded interlayers and hydrated cations; chlorite (2:1:1) with a continuous interlayer of octahedrally coordinated cations; hydroxy-interlayered vermiculite or smectite with clusters of octahedrally coordinated cations, commonly Al\(^{3+}\). Source: Adapted with permission from Schaetzl and Thompson (2015). Cambridge University Press, Cambridge.

of muscovite and illite have potassium (K\(^{+}\)) ions that act to bond the 2:1 layers together. Loss of interlayer potassium from these minerals results in greater charge imbalance and greater CEC of the resulting minerals, unless satisfied by another substance. The nature of the interlayer is very important to the identification and behavior of certain clay minerals. For example, smectite group minerals primarily have water in the interlayer position, and the drying and re-wetting of the interlayer causes large shrink–swell volumetric changes that affect land use.

Another example is the infilling of the interlayer by relatively insoluble aluminum hydroxide layers, producing minerals known as hydroxy-interlayered vermiculite (HIV) or smectite (HIS), which tend to reflect acidic, soil-forming environments of humid climates.

In the soil environment, secondary phyllosilicates form via the transformation of existing phyllosilicates, alteration of primary minerals (mostly silicate minerals), or by crystallization from soil solution components. Specific mechanisms of formation are not well understood because of the fine size of clay minerals, but recent innovations in high-resolution transmission electron microscopy have yielded new evidence. Some mechanisms of mineral formation include (i) the removal of cations from the crystal structure, as in illite losing interlayer potassium to become vermiculite; (ii) the hydrolysis of silicate minerals with synthesis of phyllosilicate minerals at the interface, as for instance smectite forming from basic primary minerals and, with more total hydrolysis removing even silica, kaolinite forming from feldspars or other 2:1 phyllosilicates; and (iii) the formation of hydroxy polymers that fill the interlayers of 2:1 clays, such as vermiculite.

Geographically, the occurrence of secondary clay minerals has been demonstrated to be related to climate and landscape stability (Folkoff and Meentemeyer 1987). In humid tropical and subtropical climates, low–activity clays such as kaolinite dominate soils, especially in older soils (Figure 3). Smectite or other 2:1 clays are common in tropical or subtropical soils that occur in
climates less conducive to leaching or that have soil-forming conditions that retain basic cations or supply inputs of fresh, base-cation-rich, parent materials, such as from volcanic ash. In high latitudes and the humid mid-latitudes that are dominated by Pleistocene glacial and periglacial deposits, the clay minerals are typically mixed, containing 1:1 and 2:1 minerals in variable concentrations that reflect inheritance from parent materials. Only the upper soil horizons show appreciable infilling with hydroxy interlayers to form HIV/HIS or degradation of illite to form vermiculite or smectite. In arid regions, where weathering is typically minimal due to the
dry conditions, secondary phyllosilicates largely reflect parent material inheritance. However, experimental and field evidence suggests potassium ions may refill the interlayer space of some high-charge clay minerals such as smectite, thus producing the widespread occurrence of illite in arid-region soils (mica in Figure 4) (Wilson 1999).

Soil Taxonomy and WRB both recognize the importance of clay minerals. For instance, the order Ultisols and the RSGs Lixisols, Acrisols, and Nitosols are groups with low-activity clays, namely kaolinite, with low amounts of charge. Low available charge tends to limit agricultural or ecosystem productivity and is typically interpreted to be the result of long-term weathering. A prevalence of smectite, with its shrink–swell potential, is the hallmark of the Vertisols in both taxonomic systems. The shrink–swell activity in Vertisols causes notable microlandforms, known as gilgai, and also presents hazards for building development. Other soil orders and RSGs are not determined by specific content or types of clay minerals present, but it is important to note that most intensive agricultural regions of the world have clay fractions dominated by mixtures of 2:1 minerals with high CEC.

SEE ALSO: Quaternary geomorphology and landscapes; Soil mass balance; Soils as relative-age dating tools; Soils of the United States; Soils and weathering

References


Further reading


Soil biology and organisms

David C. Coleman
University of Georgia, USA

Soil microbes and invertebrates and their roles in soil ecosystems

Soils contain living (biomass) and dead (necromass) organic matter, with varying amounts of gases and liquids, all within a structural matrix of mineral particles. The interactions between the various geological, hydrological, and atmospheric facets overlap with those of the biosphere, leading to the union of all, overlapping in the pedosphere. Soils, in addition to the three geometric dimensions, are also greatly influenced by time, the fourth dimension, over which the physicochemical and biological processes occur.

Plants and soil biota (microbes such as bacteria and archaea), and fungi and fauna (mostly invertebrates, including protozoa, nematodes, small arthropods, and worms) are of vital importance to soil formation and pedogenic processes. Plants introduce organic substrates into soils as products of growth, senescence, and death, while soil microbes and invertebrates utilize these organic substrates as an energy source and in the process contribute their own organic substrates through consumption, growth, and death. For microbes, the dominant constituents include microbial cell wall residues. For fungi, the main constituents are extracellular polysaccharides, carbohydrates, and amino sugars (glucosamine, galactosamine, and muramic acid). The diversity of archaea and bacteria is as yet little known, but with the use of new massive parallel sequencing devices, we are beginning to develop an appreciation of the very large biodiversity of these organisms as well.

Soil fauna may also be characterized by their presence/absence in the soil, or by their microhabitat utilization (Figure 1). Transient species, exemplified by the ladybird beetle, hibernate in the soil but otherwise live aboveground in the plant stratum. Gnats (Diptera) are temporary residents of the soil; the adult stages live aboveground. Their eggs are laid in the soil and their larvae feed on decomposing organic debris. In some soil situations, dipteran larvae are important scavengers. Cutworms are only temporary soil residents; their larva feed on seedlings by night. Some nematodes that parasitize insects and beetles spend all or part of their life cycle in soil. Periodic residents spend their life histories below ground, with adults, such as the velvet mites, emerging only to reproduce. The soil food webs are further linked to aboveground systems, making trophic analyses much more complicated than in either subsystem alone. Even permanent residents of the soil may be adapted to life at various depths within it, depending on their daily or annual life cycles.

Among the microarthropods, such as mites and collembola, the latter are examples of permanent soil residents. The morphology of collembolans reveals their adaptation for life in different soil strata. Species that dwell on the soil surface or in the litter layer may be large, pigmented, and equipped with long antennae and a well-developed jumping apparatus (furcula). Alternatively, collembolans living within mineral soil tend to be smaller, with unpigmented, elongated bodies, and they possess a much reduced furcula.
A generalized classification of soil biota by length illustrates a commonly used device for separating the soil fauna into size classes based on body length: microfauna, mesofauna, macrofauna, and megafauna (Figure 2). This classification encompasses the range from smallest to largest fauna, that is, from circa 1–2 μm for the microflagellates to >2 m for giant Australian earthworms. Body width of the fauna is related to their microhabitats. The different kinds of fauna tend to occupy different habitats. For example, the microfauna (protozoa, rotifers, tardigrades, nematodes) inhabit water films. The mesofauna, such as mites and collembolans, occupy existing air-filled pore spaces and are largely restricted to these kinds of existing spaces. The macrofauna, for example pocket gophers, have the ability to create their own spaces through their burrowing activities, and like the megafauna, can have large influences on gross soil structure.

Methods for studying these faunal groups are therefore mostly size-dependent. The macrofauna may be sampled as field collections, often by hand sorting, and populations of individuals are usually measured. There is considerable gradation in the classification based on body width. The smaller mesofauna overlap with the size range of the microfauna, and so forth.

The vast range of body sizes among the soil fauna emphasizes their effects on soil processes at a range of spatial scales. Three levels of participation may be suggested: (i) “Ecosystem engineers,” such as earthworms, termites, or ants, alter the physical structure of the soil itself, influencing rates of nutrient and energy flow, (ii) “Litter transformers,” the microarthropods, fragment decomposing litter and improve its availability to microbes, and (iii) “Micro-food webs,” which include the microbial groups and their direct microfaunal predators (nematodes and protozoans); these fauna are all key players in nitrogen and phosphorus remineralization, particularly in root rhizospheres and decomposing plant material. These three levels operate on different spatial and temporal scales.
Fauna affect soils both biochemically and biomechanically. Although most of the discussion below focuses on biochemical effects [(ii) and (iii)], it is important to note that mega- and macrofauna are key components of soil development from a biomechanical perspective (i). The larger fauna are particularly important as biomechanical instruments, moving soil particles around via a process called bioturbation (Johnson 1990; Johnson, Domier, and Johnson 2005). Bioturbation is an important slope process, assisting in the creep of materials downslope, and it is also a key component of the genesis of certain soil types. Recent data have shown that soils undergo very different genetic pathways and their morphology changes rapidly, when bioturbators such as earthworms are introduced.

**Interactions of soil organisms and organic matter; aggregate formation and breakdown**

Decomposing organic matter is the main source of food for the soil fauna; hence they are attracted to it. As organic matter decomposes, however, its fate and influence on soil biota are far from being linear or additive (see entry Soils and the carbon cycle). Some forms of organic matter contribute to the stabilization of other forms, and the collective contributions of biota to the process lead to counterintuitive results. Nowhere is this more evident than in the emerging understanding of the role of soil microbes and fauna in the formation and degradation of soil aggregates, the importance of consortia of microbes to the
decomposition of complex organic compounds, and the relationship between changes in the flow of nutrients through food webs and the rates of mineralization.

Soil aggregates are composed of sand, silt, and clay particles that are bound together by soil organic matter (SOM) and inorganic compounds. Aggregates are hierarchically classified into water-stable size classes based on their solubility. The smaller size classes include a silt and clay fraction (<53 μm) and microaggregates (53–250 μm). These smaller factions combine with coarse particulate organic matter (CPOM, >250 μm) to form small macroaggregates (250–2000 μm), which in turn all combine with organic matter in the form of free light fractions (particulate organic matter outside of aggregates (POMLF) and residues (intra-aggregate particulate organic matter, or iPOM, 250–2000 μm) to form large macroaggregates (>2000 μm). The organic components originate largely from plants in their original forms, surface litter, structural materials, and roots and root products (sloughed cells, exudates, and mucigels). These compounds are further degraded by microbes and invertebrates, resulting in the primary binding agents that are mostly microbial in origin (Six et al. 2004); it is these components that function best in the formation of aggregates.

**Faunal feeding effects; the phenomenon of “optimal grazing”**

Apart from the transformation and direct inputs of organic substrates, soil invertebrates indirectly affect the decomposition of organic substrates by enhancing the activities of soil microbes through grazing, comminution, and dissemination (Coleman, Crossley, and Hendrix 2004). Moderate levels of consumption of microbes by protozoa and invertebrates can stimulate further microbial growth through the principle of *optimal grazing*. Assuming that microbes exhibit logistic growth tied to available resources with a carrying capacity K, microbial growth rates are maximized (“optimal”) at population densities of K/2. Higher growth rates translate to increased utilization and transformation of organic substrates and increased production of microbial-derived organic substrates. Given that invertebrates consume more nitrogen than they require for growth (i.e., invertebrates are net mineralizers of nitrogen), enhanced microbial growth through grazing and subsequent excretion leads to increased nitrogen availability for plants. When this process plays itself out within the rooting zone of plants, a positive feedback can occur wherein plants and the soil food web feed off one another’s excess and waste. Under this scenario, plants exude excess photosynthate in the form of labile carbon substrates from their roots, which in turn are utilized by microbes and subsequently as food for microbivorous invertebrates, and these microbes then release nitrogenous waste that is utilized by plants.

Many invertebrates feed directly on plant materials and organic substrates. The fragmentation or comminution of these materials by soil fauna enhances their decomposition. For plant structural materials, comminution increases the surface areas of the materials and exposes cytoplasm, thereby enabling greater access by microbes. Decomposition is further accelerated as the feeding activity often results in the translocation of nitrogen from the soil to the substrate in the form of fecal material and through fungal hyphae. Grazing by invertebrates disseminates microbes from one organic source to another as many microbes adhere to invertebrate exoskeletons and cuticles and survive passage through their digestive tracts.
Towards an integrative framework for soil and soil organisms interactions

Soil fauna exist in food webs containing several trophic levels (Moore and de Ruiter 2012). Some are herbivores, since they feed directly on roots of living plants, but most are microbivores, subsisting upon dead plant matter or living microbes associated with it, or a combination of the two. Still others are carnivores, parasites, or top predators. Analyses of soil food webs have emphasized the numbers of the various organisms, as well as their traits, and trophic resources. The structure of these food webs is extremely complex, with many “missing (poorly described or unknown) links,” although tools such as stable isotopes are revealing C and N transfer through trophic and species levels within the soil food web (Crotty et al. 2012).

A framework that links soil microbes and invertebrates directly to pedogenic processes, as described above, is beginning to emerge. Soil food webs and the differences in microbial and invertebrate life forms within them offer a starting point to connecting the components of the detrital food web to soil pedogenesis and SOM dynamics (Coleman, Crossley, and Hendrix 2004). Coleman et al. earlier identified a bacterial-based fast cycle and fungal-based slow cycle within soils. Subsequent studies revealed that these fast cycles and slow cycles were borne from the trophic interactions of detritus to bacteria and their consumers, and from detritus to fungi and their consumers. These dominant trophic pathways, or “energy channels,” are ubiquitous across ecosystem types and grounded in the basic architecture of soil food webs and their structural stability. This complexity can be condensed into the dominant pathways beginning with pools of detritus or SOM that differ in quality. These pools would serve as the primary energy sources for a suite of bacteria and fungi, each of which is consumed by a host of microbial consumers and predators. Metabolic wastes and by-products that cycle back as energy sources and binding agents would be factored in, much as C and N are in the current generation of models. This approach preserves the basic premise of material transformations that occur in the soil carbon models and material transfers that occur in food web models in a way that provides a common currency.

In more arid regions, for example large areas of the Sahel, Australia, and the arid regions of North America, the keystone roles of termites in transporting and concentrating both organic and inorganic nutrients are of similar magnitude to that of earthworms in more mesic habitats (Whitford 2000), and should be borne in mind when considering faunal effects on SOM on a global scale.

Changes in land use and impacts on detrital food webs

The impacts of changes in land-use practices are of increasing interest to ecologists worldwide. With changes from agroecosystems to pasture lands and longer-term succession into forested ecosystems come significant changes in the detrital food web. Thus the more frequently disturbed tilled agricultural system tends to be more bacterial-dominated versus the fungal-dominated no-tillage or pasture land (Wardle et al. 2004). Land abandonment by humans is another form of land-use change. Soil communities respond more slowly to changes occurring with land abandonment than in the communities aboveground. Slow development of the belowground community may be an important factor controlling ecosystem services and the outputs (goods) provided by these restored ecosystems (Wardle et al. 2004).
Roles of soil fauna in ecosystems worldwide

Global experiments and syntheses have continued to address and quantify the role of soil fauna in ecosystem processes and, in particular, have led to increased evidence for their contribution to C cycling. Global multisite experiments have shown that soil fauna are key regulators of decomposition rates at a range of scales, even at biome and global ones. García-Palacios et al. (2013) conducted a meta-analysis on 440 litterbag case studies across 129 sites to assess how climate, litter quality, and soil invertebrates affect decomposition. This analysis showed fauna were responsible for approximately 27% average enhancement of litter decomposition across global and biome scales.

Information about the role of soil fauna in food webs and how it will respond to environmental change and influence aboveground processes has advanced considerably. Loss of species due to varying management practices, such as erosion, pollution, urbanization, and the resulting effects on ecosystem function and services, are becoming widely recognized. These resulting changes are related to larger issues of biodiversity loss, desertification, and elevated greenhouse gas concentrations, drawing international attention to the importance of soils, and in particular, to soil biodiversity. We are now able to answer basic questions about soil fauna, including such topics as: (i) What are the roles of cosmopolitan versus endemic fauna species? (ii) What is the local and global biogeographic distribution and range of faunal species? (iii) What are the factors influencing the distribution of key species? and (iv) How will the loss of species affect ecosystem functioning (Wall et al. 2012)? Food web ecology, with its emphasis on community assembly and disassembly, has the potential to answer these questions.

Conclusions

Studies of the distribution, abundance, and roles of soil fauna and their associated bacteria, archaea, and fungi are poised on the threshold of new discoveries. Some of this is the result of new techniques in genomics, aided by an increasing awareness of the importance of aboveground and belowground linkages in soil ecology. The combined influence of the importance of soil food webs and the need for a greater understanding of the as yet unknown extent of biodiversity of soil organisms worldwide makes the future decade or so bright with promise for the rising generation of researchers.

SEE ALSO: Biodiversity; Soils and the carbon cycle

References


**Further reading**

Soil development: numerical indices

Daniela Sauer
University of Göttingen, Germany

Background and definition of soil indices

Studies of soil development generally aim at improving the understanding of pedogenic processes – both their character and their rate, as well as their outcomes – over time. In many cases, it is useful to know not only which processes have taken place but also the degree to which these processes have proceeded over time, which may be assessed by various quantitative or semi-quantitative indices of soil development. Such information may then be used, for example, to (i) determine the chronological order of formation of soils on different land surfaces and, hence, estimate the ages of the respective land surfaces, (ii) compare pedogenesis in different climates, parent materials, or landscape positions, or (iii) compare paleosol development within a paleosol–sediment sequence.

Indices of soil development can be derived from field data, laboratory data, or both. Field data may be quantitative (such as horizon or profile thickness) or they may be recorded in classes (e.g., size of aggregates, abundance of coatings). Easily obtainable field data (e.g., soil color) may reflect important pedogenic processes (e.g., hematite formation or processes related to certain redox conditions) that are not always easily measured in the laboratory. Therefore, efforts at transforming ordinal field data into numerical values have been undertaken.

General assumptions of soil development indices include that (i) they should be based on data that reflect clearly defined and well-understood soil-forming processes, (ii) they should normally utilize data obtained from field observation, or from laboratory analyses that can be carried out in any ordinarily equipped soil laboratory, so as to ensure their potentially wide applicability and possible use in comparative studies, and (iii) they should work in all kinds of environmental conditions, the environmental boundary conditions of their applicability need to be defined, or they need to be adjustable to different environmental conditions. Most soil development indices have been derived to ascertain the intensity of weathering and pedogenic transformation of a soil parent material. Soil thickness may or may not be included in a soil development index.

Field data suitable for constructing soil development indices

Field data utilized in soil development indices generally should reflect soil-forming processes that proceed over time, leading to predictable and well-understood changes in soil morphology. The latter include soil texture (reflecting clay formation, clay illuviation, and perhaps eolian silt accumulation), clay films or
SOIL DEVELOPMENT: NUMERICAL INDICES

argillans (indicating clay illuviation), rubification or reddening (indicating the formation of iron oxides, especially hematite), melanization or darkening (indicating organic matter accumulation), soil (profile or horizon, eluvial or illuvial zone) thickness, dry/moist consistence (mainly reflecting clay formation and illuviation), soil structure, and abundance of pedogenic carbonates and other illuvial materials (gypsum, salts, etc.), as well as their location within the profile.

Importance of the timescale considered

Some soil properties exhibit their main changes during certain periods of soil development, rather than progressively throughout the period of soil formation. For example, soil structure formation and melanization are dominant processes in young soils, on timescales of decades to a few millennia. On the other hand, such young soils exhibit little rubification and rarely any clay films. These properties usually develop in older soils, depending on climate and parent material. Therefore, indices that combine various soil properties, each of which may have their main changes in different phases of soil development, are particularly useful for studying soil chronosequences that extend well back, into the Pleistocene. Soil chronosequences comprising only Pleistocene soils may be better examined using properties that exhibit appreciable changes within the longer time span covered by the chronosequence (e.g., carbonate accumulation, rubification), but which exclude properties such as soil structure and melanization that form quickly and then do not change appreciably over the longer timescale. The latter may be used for assessing Holocene soil development, for which, in turn, rubification, for example, cannot be used.

General numerical soil development indices based on field data

Redness rating (Torrent, Schwertmann, and Schulze 1980)

Redness rating (RR) has proved useful in numerous studies of Pleistocene soil chronosequences, especially in Mediterranean climates, where rubification is a major pedogenic process (Figure 1). It is calculated as: 

\[
(10 - H) \times C/V,
\]

where \( H \) is the numerical value preceding YR in the Munsell hue, and \( V \) and \( C \) represent value and chroma. RR is not useful for environments or time spans where the Munsell hue is 10YR, as in most older and well-developed soils, thus resulting in RR = 0. For example, most soils in cold or temperate climates have a Munsell hue of 10YR. Progressing rubification may also be masked by subsequent eolian or volcanic additions to soils.

Figure 1  Maximum redness rating (RR) in B horizons of three soil chronosequences in Mediterranean climates. Black quadrats = San Joaquin Valley, California (Harden 1982); white quadrats = Sacramento Valley, California (Busacca 1987); crosses = central western Spain (Dorronsoro and Alonso 1994).
Soil age (ka)

Field morphology rating scale (Bilzi and Ciolkosz 1977)

This rating scale has been developed for soils in humid-temperate climates. It can be used for (i) evaluating the degree of pedogenesis by comparing soil horizons to C horizons, and/or (ii) evaluating distinctness of horizons by comparing the properties of adjacent horizons. The comparisons are done by allocating one point per unit difference between the two horizons (e.g., for textural class, grade of structure, amount of clay films). The points are then summed over the profile.

Profile development index (Harden 1982)

The profile development index (PDI) has become the most widely used soil development index based on field data (Figure 2). It has been successfully applied to soils in a wide variety of environments, although most of this work has been done in Mediterranean and humid-temperate climates. The procedure is as follows.

1. The properties of each soil horizon and the parent material are described in the field.
2. Each considered property is evaluated by assigning 10 points per class crossing between the parent material and the soil horizon.
3. The values obtained in step 2 are normalized to a scale from 0 to 1. The normalized morphological property (NMP) values of all considered properties are summed up for each horizon.
4. The result of step 3 is divided by the number of included properties to calculate the morphological horizon index (MHI).
5. The MHI is multiplied by horizon thickness.
6. The results of step 5 for each horizon are summed up through the profile to calculate the PDI.

The number and kind of soil properties that are included in the index calculation can be adapted to the environmental conditions and processes that have taken place in the soils under investigation.

To include or not to include profile thickness?

Because of the way that step 6 is defined, the numbers obtained in the profile index will include the increase in profile thickness with progressive soil development in the index. This approach may or may not be applicable to all soils. That is, when choosing a method to calculate a profile index, one has to decide whether soil thickness should be included. An index that includes soil thickness may increase in a soil chronosequence even though the horizon property values remain constant. Therefore, if one wants to identify solely the intensity of pedogenesis, excluding profile thickness, one may compare, for example, the maximum

![Figure 2](image_url)
SOIL DEVELOPMENT: NUMERICAL INDICES

NMP or MHI occurring in each pedon of a chronosequence or, after step 6 of the PDI calculation, divide the index by profile thickness. Alternatively, to include soil thickness in the index, one may artificially adjust all profiles to the depth of the thickest soil, either by assigning an index of zero to each pedon for the part between its soil depth and the soil depth of the thickest soil, or by increasing the thickness of the lowermost exposed soil horizon of each pedon so that all pedons have the same thickness (Birkeland 1984).

Numerical soil development indices based on laboratory data

Indices related to the processes of clay formation and illuviation

Clay accumulation in B horizons is a common process in various kinds of soil environments, including particularly Mediterranean, humid-temperate, and savanna ecosystems. It can be assessed through the clay accumulation index (CAI) proposed by Levine and Ciolkosz (1983), which is calculated as: CAI = Σ [(Bc – Cc) * T], where Bc = clay content of B horizon (%), Cc = clay content of C horizon (%), and T = B horizon thickness (cm). Progressive clay translocation from E to Bt horizons is often calculated as the ratio of clay content in Bt horizon (%)/clay content in E horizon (%). This ratio is applicable only where the present E horizon is the source of additional clay in the Bt horizon. This may not be the case where, for example, the original eluvial horizon has been replaced by a younger deposit, or where the primary clay content of the parent materials of the E and Bt horizons differ.

Indices related to pedogenic iron oxides

Formation of, and pedogenic processes associated with, iron oxides (such as crystallization, remobilization, and translocation) occurs in almost all kinds of environments. Various Fe ratios are commonly used for assessing these processes. For example, the Fe_d/Fe_t ratio (Figure 3) reflects the proportion of total iron (Fe_t, obtained, for example, by x-ray fluorescence analysis or combustion with hydrofluoric acid) that has been transformed into iron in pedogenic iron compounds (Fe_d, obtained by sodium citrate-dithionite extraction), whereas the Fe_o/Fe_d ratio characterizes the proportion of iron in amorphous and poorly crystalline iron oxides and hydroxides (Fe_o = Fe extracted by acid ammonium oxalate), as compared to the total amount of iron in pedogenic iron compounds (Fe_d) (Torrent, Schwertmann, and Schulze 1980). Other commonly applied iron ratios include the (Fe_d–Fe_o)/Fe_t ratio, characterizing the proportion of iron in well-crystallized pedogenic iron oxides (Fe_d–Fe_o) as compared to total iron (Fe_t), and the ratio Fe_d in B horizon(s)/Fe_d in the C horizon.

Indices reflecting progressive silicate weathering and leaching of mobile elements

In many environments where there is an excess of precipitation over evapotranspiration, water percolates through the profile at times during the year, and leaching occurs. Leaching removes soluble substances from the soil system. As soils of this kind get increasingly weathered, ratios of insoluble to soluble substances increase, making such ratios valuable as indices of weathering, pedogenesis, and leaching, and, hence, soil age. Commonly applied indices of this group include the molar ratios of SiO_2/Al_2O_3 or (Fe_2O_3 + Al_2O_3)/SiO_2, reflecting the processes
of desilication (removal of Si) and formation of kaolinite, gibbsite, and iron oxides from the relatively enriched Al and Fe. Ratios of mobile elements released during silicate weathering (e.g., Na, K) to an immobile element (mostly Ti or Zr) are often used as well (in the form both of elements and of oxides). There is some discussion about whether Ti is completely immobile or whether it can be co-illuviated with clay. Also Al is not immobile in all soils; it is readily translocated during podzolization and clay illuviation. Chittleborough (1991) thus suggested a weathering ratio (WR) based on Zr as an immobile element: \[ \text{WR} = \frac{(\text{CaO} + \text{MgO} + \text{Na}_2\text{O})/\text{ZrO}_2}{\text{Al}} \].

Among a number of similar weathering indices developed by geochemists, the one that has been most widely adopted in pedology is the chemical index of alteration (CIA) (Nesbitt and Young 1982). It is calculated from the molar proportions of feldspar constituents as: \[ \text{CIA} = \frac{\text{Al}_2\text{O}_3/((\text{Al}_2\text{O}_3 + \text{CaO}^* + \text{Na}_2\text{O} + \text{K}_2\text{O}))}{100}, \] where \( \text{CaO}^* \) is Ca after subtraction of Ca in carbonates and apatite. Sauer et al. (2010) preferred a weathering index based on the molar element ratio \( (\text{WI}_{\text{MER}}) \) for use in pedology, calculated as: \( \text{WI}_{\text{MER}} = (\text{Ca}^* + \text{Mg} + \text{K} + \text{Na})/\text{Al} \), where \( \text{Ca}^* \) is Ca after subtraction of Ca in calcium carbonates, because this ratio reflects most directly the progressing release and leaching of bases from silicate weathering during pedogenesis. Both indices should be calculated without potassium \( (\text{CIA} - \text{K}, \text{WI}_{\text{MER}} - \text{K}) \) when applied to paleosols in which postburial alteration may include metasomatic additions of potassium. See Duzgoren-Aydin, Aydin, and Malpas (2002) or Schaetzl and Thompson (2015) for reviews of chemical weathering indices.

**Index of profile anisotropy (Walker and Green 1976) and modified index of profile anisotropy (Birkeland 1984)**

These two indices are based on the concept that anisotropy of soil properties increases with time, as occurs in most soils, except for soils characterized by strong pedoturbation (e.g., steppe soils or Gelisols). The index of profile anisotropy (IPA) is calculated as: \( \text{IPA} = D \times 100/M \), where \( M \) is the weighted mean value of a soil property in the profile and \( D \) is the mean deviation of the horizon data of that property from the weighted mean value. The modified index of profile anisotropy (mIPA) uses the parent material data \( (\text{PM}) \) instead of the weighted mean value \( M \): \( \text{mIPA} = D/PM \). It is determined as follows.

1. The deviation of each horizon from the parent material with regard to the respective property is calculated and multiplied by horizon thickness.
2. All profiles are artificially adjusted to the thickness of the thickest soil by increasing the thickness of the lowest horizon.
The sum of horizon deviation data is divided by this thickness.

**Soil-specific numerical soil development indices**

The indices introduced above may be applied to soils in a comparatively wide range of environments. In addition, another family of indices is designed to reflect specific soil-forming processes that take place in specific environments. For example, pedogenesis in many desert environments is commonly described in stages of secondary carbonate accumulation (Gile and Grossman 1979). A quantitative numerical index, based on the amounts of secondary carbonates in a profile (g cm\(^{-2}\)), was introduced by Machette (1985). Amounts of secondary carbonates (cs) in each horizon (g cm\(^{-2}\)) are calculated as 
\[
cs = ct - cp
\]
where \(ct\) is the horizon’s present total amount of carbonates (g cm\(^{-2}\)) and \(cp\) is its initial amount of carbonates (g cm\(^{-2}\)); \(ct\) is calculated from present carbonate contents (g CaCO\(_3\) per 100 g soil), bulk density (g cm\(^{-3}\)), and horizon thickness (cm), whereas \(cp\) is calculated from the calcium carbonate content in the parent material and reconstructed initial bulk density (g cm\(^{-3}\)) and horizon thickness (cm). The amounts of secondary carbonates of all horizons are then summed.

Muir and Logan (1982) introduced an eluviation/illuviation coefficient (EIC) that had originally been developed by Rode (1935) for describing the degree of podzolization. It is calculated as in equation 1, where \(S_{hi}\) and \(X_{hi}\) are the percentages of the soil constituent of interest \(S\) and an internal index element \(X\) in the illuvial soil horizon, and \(S_{oi}\) and \(X_{oi}\) are the original percentages of the same constituents, obtained from the parent material. The EIC is converted into a percentage (by multiplying by 100), a negative EIC indicating a percentage loss, a positive EIC indicating a percentage gain of the soil constituent \(S\) in the respective soil horizon, as compared to the parent material.

\[
\text{EIC} = \frac{S_{hi}/X_{hi}}{S_{oi}/X_{oi}} - 1
\]  

An index of podzolization based on soil color and horizonation was introduced by Schaetzl and Mokma (1988), calculated as: 
\[
\text{POD} = \Sigma (\Delta V \times 2^{\Delta H})
\]
where \(\Delta V\) is the difference in color value (moist) between the E horizon and B horizon (or subhorizon), and \(\Delta H\) is the number of Munsell hue pages by which the E and B horizons differ. Thus, \(\Delta V\) is multiplied by \(2^0 = 1\) if the horizons have same hue, by \(2^1 = 2\) if there is one page difference, by \(2^2 = 4\) if there are two pages difference, and so on. The resulting differences between the E horizon and each B subhorizon are summed. If several E subhorizons are present, the one with the highest Munsell value is used.

**Conclusions**

Numerical indices of soil development have been applied successfully in numerous studies to assess the degree of pedogenesis, for example, in soils of different ages or on different positions in the landscape. They have thus helped to improve the understanding of the influence of the various soil-forming factors on pedogenesis. Moreover, they have been used for establishing chronologies of land surfaces or cycles of alternating sedimentation and soil formation, thus contributing to the reconstruction of landscape evolution. Successful application of soil development indices requires some knowledge about various factors that may affect the indices in the soils under study. For example, the material exposed in the C horizon may not be the same parent material. 
that the soil horizons above have developed in. In fact, sedimentary discontinuities in soil profiles are not often addressed in the literature but are nevertheless rather common in nature (Schaetzl 1998). In addition, soils that have developed over long time spans have usually experienced major climatic changes (e.g., glacial/interglacial cycles) and geomorphological processes including erosion and redeposition (Sauer et al. 2010), and have possibly received additions of dust and/or volcanic ash. Such processes need to be taken into account when applying soil development indices to older soils.

**SEE ALSO:** Paleosols; Soils in geomorphic research; Soils as relative-age dating tools; Soils and weathering

**References**


SOIL DEVELOPMENT: NUMERICAL INDICES


Further reading


Soil erosion and conservation

Tom Vanwalleghem
University of Cordoba, Spain

History, definitions, and importance of soil erosion

Soil erosion is the process by which soil particles are detached, transported, and deposited by water, wind, gravity, or biota. Soil erosion, as part of the geological cycle, is a natural process. However, anthropogenic disturbance of the landscape, mainly by replacing the natural vegetation cover with agricultural crops as well as by tilling and disturbing the soil, leads to an increase of erosion rates by a factor of 10–100 (Montgomery 2007). This type of accelerated erosion generates an imbalance with the rates by which soils are formed, generally resulting in a net loss of soil, resulting in decline of soil fertility and, potentially, the total loss of soil cover. Also, the increased sediment and nutrient load to downslope water courses generates problems such as deteriorating water quality and damage to infrastructure, for example, the infilling of reservoirs behind dams. The field of soil conservation encompasses the efforts of humans to correct this imbalance and, ideally, reduce soil erosion to long-term geological rates, or even to put practices in place that will encourage soil deepening and increased fertility. Soil conservation practices include changes in soil management and the implantation of engineering structures. Thus, their proper implementation requires an understanding of erosion processes (Morgan 2004).

The first evidence of human-induced soil erosion can be dated back as early as the Mesolithic Period, as hunter-gatherers cleared natural forest vegetation by fire (Dotterweich 2013). In general, important and widespread periods of erosion in history can be found with the arrival of the first farmers; this association holds throughout the world. In areas with a long land-use history such as the Mediterranean, clear signs indicating the advanced degradation status of the landscape, such as heavily truncated soils, are visible throughout the landscape. Although within a complex mix with many other factors, such as drought, socioeconomic change, or soil salinization due to irrigation, the decline of some ancient civilizations, such as the Maya or Easter Island’s, has been linked mainly to soil erosion.

Each of the driving forces for soil erosion – water, wind, gravity, and/or biota – varies greatly in space and time. Thus, the resulting detachment, transportation, and deposition of soils are therefore controlled by different variables at different places. Anthropogenic perturbation of the erosion cycle, usually called accelerated erosion, has been most pronounced on agricultural land and can lead to rapid land degradation. Some of the areas that are currently most affected include Africa south of the equator, Indo-China, Myanmar, Malaysia, Indonesia, South China, Northern Australia, the Pampas, and swaths of the high-latitude forest belt in North America (Bai et al. 2008). Areas with a long history of land degradation, such as the drylands around the Mediterranean, have largely been stripped of their soil cover and have now often become more stable landscapes. Typical
soils in such areas are shallow and with a high rock fragment cover.

Current global estimates indicate that among the different erosion processes, water and tillage erosion are the most important on agricultural land, resulting in a global flux of sediments of, respectively, 28 Pg year\(^{-1}\) and 5 Pg year\(^{-1}\). About 2 Pg year\(^{-1}\) of sediment is mobilized by wind erosion (Quinton et al. 2010). Tillage erosion refers to the direct translocation of soil by tillage operations, and is discussed in more detail below.

Although mass movements have been widely studied by geomorphologists and are important processes of local sediment generation in mountainous catchments, they have generally been considered negligible in the larger context of soil erosion. Sediment budgets from cosmogenic nuclides seem to suggest that their importance to the global sediment budget is limited (Willenbring, Codilean, and McElroy 2013). Currently, no global estimates exist for erosion by mass movements on agricultural lands. This discussion, therefore, mainly deals with water and tillage erosion, dominant on agricultural land, although wind erosion is also briefly discussed because of its local importance. For further details on the other erosion processes, the reader is referred to geomorphological literature, for example Anderson and Anderson (2010).

**Water erosion**

Water erosion processes are generally subdivided into splash, sheet or interrill, rill, and gully erosion. Splash erosion refers to the direct impact of raindrops on soils, which breaks up soil aggregates and transports the individual soil particles away from the impact zone. Sheet erosion then refers to the erosion by the runoff water that flows as a laminar or turbulent sheet downslope. The effects of sheet flow are complex and encompass both the transport of splashed soil particles, erosion by the sheet flow itself if the flow’s shear stresses exceed the soil’s resistance, and a partial protection of the underlying soil from raindrop impact. Rill and gully erosion occur further downslope, where sheet flow converges and, using the increased stream power, erodes a distinct and often deep channel or gully (Figure 1). Most prediction models use a critical drainage area \((A)\) and slope \((S)\) threshold relation of the form \(S = aA^b\), with site-specific parameters \(a\) and \(b\), to delimit the areas in the landscape where rills and gullies are expected to form. The subdivision between rills and gullies is generally based on size, using a critical cross section of 0.929 m\(^2\) (1 ft\(^2\)) to distinguish them (gullies are larger).

Water erosion is strongly scale-dependent because of the interactions between these different erosion processes, each active in a different area of the landscape, and sediment deposition, which also occurs at distinct locations in the landscape, such at the foot of hill slopes or in low relief valleys. These complexities make measurement of soil erosion, and especially the spatial extrapolation of results, a difficult task. Erosion can be measured directly by determining the elevation changes at the soil surface with techniques ranging from reference markers, such as erosion pins, to repeated terrestrial LiDAR measurements. Various types of tracers have also been used to track soil redistribution on the landscape over a range of timescales. Fallout radionuclides, such as \(^{137}\)Cs, \(^{210}\)Pb, or \(^{7}\)Be, which fall out of the atmosphere at known rates and accumulate in soils, especially near the surface, allow for the averaging soil erosion and deposition rates over timescales ranging from days to 100 years (Guzmán et al. 2013). In addition to these nuclides, other tracers, such as rare Earth or magnetic iron oxides, have been artificially incorporated to the soil – a type of erosion...
experiment. After a rainfall event or after several years, the determination of the concentration of such tracers allows the experimenter to calculate erosion extent and intensity. The most widely used method of determining the extent of soil erosion, however, is the direct measurement of sediment production in catchments or plots of a known size. This method is also known as universal soil loss equation (USLE). It also provides for direct comparisons of soil erosion under different management systems or conservation practices.

Sparked by the 1930s dust bowl, the United States Soil Conservation Service (currently the Natural Resources Conservation Service) initiated the most significant soil erosion field plot measurement program to date. Currently, plot measurements have been collected worldwide as a result of individual research efforts. Available data now comprise >10,000 plot years. Montgomery (2007) reviewed these soil erosion field plot data under conventional agriculture and reported median and mean soil erosion rates of 18.4 and 48.6 ton ha$^{-1}$ year$^{-1}$, respectively, far exceeding rates of soil formation. Soil conservation programs generally consider soil losses between 5 and 12 ton ha$^{-1}$ year$^{-1}$ to
SOIL EROSION AND CONSERVATION

be tolerable, because they can be matched by natural pedogenic processes. Ongoing research on the spatial and temporal variability of soil formation rates will allow setting more precise and site-specific tolerable thresholds.

The various soil erosion prediction models that currently exist can broadly be categorized as empirical or physically based. The choice of model depends mainly on the objective: the first group is geared towards prediction (forecasting) while the latter type of model is used best for explaining (back-casting). The most widely used model is the universal soil loss equation (USLE) and its subsequent modifications (Renard et al. 1991). Based on a statistical fit of the available plot data described above, the USLE model is considered an empirical model. It predicts the average annual soil loss \( A \), ton ha\(^{-1}\) year\(^{-1}\), as a function of five factors: rainfall erosivity \( R \), soil erodibility \( K \), slope length and gradient \( LS \), soil cover \( C \), and management \( P \):

\[
A = R \times K \times LS \times C \times P
\]

For example, to calculate erosion in an olive orchard in Cordoba, Southern Spain, the \( R \) factor calculated from local climate records is 850 MJ mm ha\(^{-1}\) h\(^{-1}\). The soils are Vertisols with a \( K \)-factor equal to 0.047 \( \text{t ha}^{-1} \text{MJ}^{-1} \text{mm}^{-1} \). On a typical field of 180 m long with an 8% slope, the \( LS \) factor is 2.47 and typical \( C \)-values for this crop are around 0.34. Without additional conservation practices such as terraces, the \( P \) factor is equal to one. The mean annual soil loss can then be calculated for this case as 850 \( \times \) 0.047 \( \times \) 2.47 \( \times \) 0.34 \( \times \) 1 = 33.6 ton ha\(^{-1}\) year\(^{-1}\).

The various physically-based soil erosion models are largely based on the concept of sediment transport capacity. Most draw heavily from generalized sediment transport equations originally developed for fluvial conditions, such as ANSWERS, CREAMS, GUEST, LISEM, PESERA, or WEPP (Merritt, Letcher, and Jake-man 2003). Finally, models such as WaTEM/SEDEM apply a hybrid approach, combining USLE-based sediment production with posterior sediment routing, using the sediment transport capacity concept.

Most of these models do not include gully erosion. In spite of some valuable efforts, such as the (revised) ephemeral gully erosion model, at present no widely validated model exists for predicting gully erosion rates.

**Tillage erosion**

Tillage erosion or translocation is the loss and accumulation of soil by tillage operations such as plowing or cultivation (Van Oost et al. 2006). Soil moved during a downslope tillage operation is not fully compensated by the corresponding upslope tillage operation so that, overall, tillage results in net downslope soil transport. A general equation describing the tillage erosion or deposition rate \( E \) is:

\[
E = k_{til} \frac{\partial^2 h}{\partial x^2}
\]

Tillage translocation rates are thus controlled by slope curvature in the direction of tillage (with \( h \) = elevation) and a tillage transport coefficient, \( k_{til} \), that incorporates the various effects related to the characteristics of the tillage operation itself (number of passes per year, direction, speed and depth of tillage, and the tillage implement used). Finally, soil bulk density effects are included in \( k_{til} \).

Field measurements have shown that tillage transport coefficients are an order of magnitude larger for mechanized moldboard plowing, as compared to noninversion animal-powered tillage, which is still common in some developing countries. These data explain why, historically,
water erosion rates have been much higher than tillage erosion rates. Presently, though, water and tillage erosion processes are often equally important, as intensive mechanization is now widespread. However, because the travel distance of soil particles due to tillage erosion is limited to the field scales, tillage erosion commonly results in the formation of benches or hedges at downslope field borders. Although it can exert an important control on the spatial variability of soil properties, tillage erosion does not contribute directly to sediment export from a catchment because the almost all of the eroded soil stays within the field. Some research has pointed to positive feedback effects between water and tillage erosion in areas, where gullies are regularly filled in by tillage and easily form again each year (Gordon et al. 2008).

Wind erosion

Wind erosion, the initiation, transport, and deposition of soil by wind, is an important process in certain areas – known as hotspots. These areas include sandy, peaty, and loess soil types in large fields without barriers, and arid areas that support little vegetation cover (Nordstrom and Hotta 2004). These hotspots are mostly in the Northern countries of Europe, the Great Plains of the United States, the Sahel of West Africa, large parts of Australia, northern Kazakhstan, and Inner Mongolia in Asia. The erosion risk in these places depends on factors such as vegetation cover, soil roughness, and moisture, all of which vary temporally.

Wind erosion occurs when the wind velocities exceed the threshold friction velocity to move soil particles. It often is focused during distinct periods of the year, usually when critical values are lower due to dry conditions and low soil aggregation. Apart from soil surface conditions, the amount of total erosion will also depend on wind energy and duration. The best prevention is offered by windbreaks and measures that increase surface cover or create aggregates that resist entrainment. Most of these measures are similar to conservation measures geared toward water erosion.

From the previous sections, it is clear that both water and tillage erosion processes have shaped our agricultural landscapes and soils significantly. The interaction between soil formation and soil erosion along a slope is part of the catena concept. For this reason, truncated soils are widespread on hillslopes in many agricultural fields (Figure 2). As soils are the main terrestrial stock of carbon and nutrients, soil erosion has also led to important lateral fluxes of associated soil components, such as carbon, nitrogen, and phosphorus. Current estimates calculate the agricultural erosion flux of carbon to be 0.5 Pg C year\(^{-1}\) and estimates of the fluxes of nitrogen and phosphorus are estimated to be 23–42 Tg year\(^{-1}\) and 2.1–3.9 Tg year\(^{-1}\), respectively. For comparison, the amounts of nitrogen and phosphorus added through chemical fertilizer are 112 and 18 Tg year\(^{-1}\). This implies that soil conservation, or its absence, especially in developing countries where fertilizers are not readily accessible to farmers, can have an important impact on the nutrient budget. Soil erosion processes also play a crucial role in the global carbon flux in other ways; the burial of eroded soil and the dynamic replacement of carbon on eroded sites leads to long-term carbon storage in actively eroding agricultural landscapes (Quinton et al. 2010).

Soil conservation: definition and principles

Soil conservation is aimed at obtaining a maximum level of production from a given area of
Figure 2  Soil erosion is evident on this hillslope in France. The whiter areas on the shoulder slope are eroded, allowing the bedrock to show through. Photo by R. Schaetzl.

land, mainly by maintaining soil losses below a threshold level (Morgan 2004). Soil and water conservation usually go hand in hand, as promoting infiltration and reducing on-site runoff is paramount to reducing soil loss. More recently, soil biodiversity and other ecosystem functions have also received attention within the framework of soil conservation (Palm et al. 2014).

There are two main approaches to soil conservation: (i) agricultural measures and soil management, aimed at reducing on-site runoff and sediment generation, and (ii) mechanical measures, aimed mainly at reducing off-site effects. When designing soil conservation programs preference should always be given to agronomic treatments, although mechanical measures can be locally indispensable, for example in gully restoration or for farming in steep terrain with terraces.

Soil conservation has a long history, possibly beginning with agricultural terracing at around 6000 BCE in the Middle East. From here, the practice spread globally. In fact, ancient management systems have been used at many locations around the globe, as templates for developing new conservation strategies, for example,
Soil erosion and conservation research emerged in the late eighteenth century (Dotterweich 2013) and currently most countries have active research and extension programs.

With respect to soil conservation, agronomic practices and soil management strategies are based on (i) enhancing infiltration rates to reduce overland flow, and (ii) covering the soil to protect it from raindrop impact. Because plant roots contribute to soil cohesion and reduce sediment production by overland flow, maintaining a cover crop on the soil is often an important part of soil conservation management practices. The simplest way of assuring this is to grow crops continuously, year-round. Growing cover crops in between the main crop cycle(s) assures adequate soil protection during times when the soil would otherwise be bare and prone to erosion. In addition, these crops improve soil structure and fertility when incorporated into the soil. Crop rotation is another way of lowering long-term soil erosion losses, by growing soil-protecting crops of legumes or grasses after row crops such as wheat or maize, which allow more soil erosion because of the bare, interrow areas. Intercropping or multiple cropping, for example maize–legume, is another attractive option to increase canopy coverage and reduce soil erosion. Strip cropping – alternate strips of protection-effective crops with row crops – can be used on areas of moderate slope (Figure 3). High-density planting is a way of achieving a more complete soil cover in a monoculture system, and has been shown to be effective to control gully and rill initiation, even if limited to the concentrated flow zones. A more dense plant cover can also be simulated by mulching, which is the covering of soil with plant residue such as straw, maize stalks, or chipped pruned branches (Figure 4). Mulching is particularly useful in arid

Figure 3 Strip cropping used to break up long sloping fields, in combination with a detention pond for control of storm runoff and sediment (left corner). Photo by R. Schaetzl.
areas where rainfall does not permit growing crops year-round. In cold climates it is also used, but mainly to protect plants from low temperatures. Finally, severely degraded areas such as gullies or mine spills should be revegetated artificially in order to obtain a protective plant cover quickly.

Soil management is aimed at improving soil structure and fertility. This outcome can be achieved by improving the soil’s organic matter content or by applying soil conditioners, such as gypsum in sodic soils or lime in acidic soils. Conventional tillage is still the most common soil management practice globally, as tillage is an essential agronomic management technique that increases surface roughness, albeit temporarily, prepares seed beds, controls weeds, and enhances infiltration. However, the benefits of tillage can easily be offset by compaction under intense traffic or during wet conditions. In contrast, conservation tillage aims at maximizing soil cover by leaving surface plant residue and minimizing soil disturbance (Figure 4). It includes practices such as no-till, minimum tillage, ridge tillage, and direct drilling.

Mechanical soil conservation requires engineering techniques or structures, and often includes surface contouring, terraces, check dams, grassed waterways, grass buffer strips, and the use of various kinds of geotextiles. Its main
objectives are to modify the soil slope, to provide a controlled way of conveying runoff water from slopes to waterways, or to detain sediments before they enter the river courses.

Because of the high costs of soil conservation to farmers, for example, for terrace building or switching to new machinery adapted for direct drilling, a change to more sustainable agricultural practices is often supported by policymakers. Growing public concern about agricultural sustainability and off-site effects of soil erosion has resulted in a set of policy measures that, in many countries, have already contributed to lower erosion rates. In Europe, policies are a mixture of mandatory measures, for example, under the Common Agriculture Policy farmers have to comply with certain standard management requirements in order to receive direct government payments, and incentive-based measures, in which farmers that comply with so-called agro-environmental climate measures receive additional subsidies. The United States has similar conservation programs under the Conservation Reserve Program aimed at applying conservation measures on highly erodible land in exchange for financial and other benefits.

**SEE ALSO:** Environmental degradation; Eolian erosional processes and landforms; Fluvial erosional processes and landforms; Land degradation; Mass movement processes and landforms; Soil fertility and management; Soils on slopes: catenas

**References**


SOIL EROSION AND CONSERVATION


Further reading

Soil fertility and management

Cynthia A. Grant
Agriculture and Agri-Food, Canada

Soil fertility is the quality of a soil that allows it to provide an adequate amount and balance of nutrients to support healthy plant growth, while maintaining overall soil quality and environmental sustainability. It is a function of the amount and form of nutrients present, their movement to plant roots, and their availability for uptake and utilization by plants. A fertile soil contains adequate soil organic matter, lies within a pH range that supports growth, and has a vibrant community of micro-organisms that participate in nutrient cycling. It lacks constraints that limit the ability of the plant to accumulate nutrients, such as compaction, poor aeration, toxic materials, and excess salts. In managed agriculture or forest ecosystems, soil fertility is often defined in terms of economic productivity, while in natural ecosystems it is related to the capacity of a soil to support a healthy natural plant community.

Historical development of soil fertility

The importance of soil fertility was recognized by early civilizations in Asia, Europe, India, and the Americas that developed agricultural practices at locations where the soil was initially fertile. Soil fertility, however, must be maintained and/or improved for sustainable crop production and the decline of many cultures was linked to soil degradation. In contrast, in Egypt, soil fertility was replenished each year by sediments deposited by flooding of the Nile River, allowing the development of its advanced civilization. In Asia, animal and human manures were used as nutrient sources and soil conditioners thousands of years BCE. In the Americas, active soil fertility management may also have been in place several thousand years BCE, as indicated by the development of the Terra Preta soils that were formed by large additions of charcoal, human excrement, and food waste; these substances greatly improved the fertility of the normally infertile Amazon soils. During Greek and Roman times, many of our current soil fertility management practices had already been adopted, including crop rotation and use of fallow, as well as the addition of soil amendments such as animal manures, green manures, compost, and liming materials. More detailed understanding of plant nutrition came in 1840 when Justus von Liebig published his book Organic Chemistry in Its Application to Agriculture and Physiology, in which he outlined the concept that plants derive carbon, oxygen, and hydrogen from the air and water, but access most of the other nutrients they require from the soil. He developed the “Law of the Minimum” proposing that plant growth will be constrained if one of the essential nutrients is deficient, even if all the other nutrients are present in adequate amounts (Figure 1). Since that time, essential nutrients for plant growth have been identified and factors contributing to soil fertility characterized. In 1913, Victor E. Shelford expanded the concept of limiting factors by proposing the “Law of Tolerance” suggesting that the growth of an organism will be limited if environmental factors, including nutrients, are outside of the
SOIL FERTILITY AND MANAGEMENT

Maximum plant growth is determined by the most limiting factor

Figure 1 Liebig’s “Law of the Minimum” for crop growth can be illustrated by a barrel. In this case, sulfur (S) is the most limiting nutrient. Figure adapted from Saskatchewan Ministry of Agriculture.

minimum or maximum tolerance range for that organism.

Nutrients required for plant growth

Plants require at least 16 essential elements for growth. As von Liebig determined, carbon (C), hydrogen (H), and oxygen (O) are provided from the air and water, while the other 13 are normally taken up from the soil. The three primary macronutrients, nitrogen (N), phosphorus (P), and potassium (K), are required in the largest amounts, while sulfur (S), calcium (Ca), and magnesium (Mg) are considered secondary macronutrients. Although the secondary nutrients are required in relatively large amounts, they are less commonly limiting than the primary nutrients and so are less frequently applied onto agricultural land. The remaining nutrients, the micronutrients iron (Fe), manganese (Mn), zinc (Zn), copper (Cu), boron (B), chlorine (Cl), and molybdenum (Mo), are needed in only very small amounts by plants. In addition, nickel (Ni) has been shown to be necessary for legumes and cereal crops, and is considered by many sources to be a seventeenth essential element. Cobalt (Co) is also needed for some crops, including legumes that require it for nitrogen fixation. Silicon (Si) can be beneficial for certain crops such as rice and sugarcane, but is not essential for growth. Plants may differ in the specific amount of each nutrient required, and each plant has an optimum nutrient range to support growth and a minimum requirement below which growth and/or reproduction is restricted. As Shelford observed, excess amounts of nutrients may also pose a problem, because of direct toxicity or through interference with the uptake and/or use of other essential nutrients.

Factors affecting soil fertility

Soil fertility depends on complex and often little-understood interactions among the chemical, biological, and physical properties of the soil. Soil is formed very slowly by physical and chemical processes, modulated by living organisms, acting on geological parent material, and is affected by climate and topography. Most of the soil’s initial mineral nutrient supply was provided by the decomposition or “weathering” of minerals in the parent material. Plant-available nitrogen is added to the soil when lightning converts atmospheric dinitrogen (N₂) gas into reactive forms, which then enters the soil by direct deposition or in precipitation. Plant-available nitrogen is also added to soils by biological fixation carried out by symbiotic or free-living, nitrogen-fixing bacteria. Atmospheric deposition also adds small amounts of sulfur and chlorine to soils. The size and nature of the mineral particles produced in the soil-forming process affect both the release and the retention
SOIL FERTILITY AND MANAGEMENT

of nutrients. For example, fine-textured materials such as clays can better release and retain nutrients than can coarser materials such as sand. The physical and chemical characteristics of the initial parent material will have a major influence on soil fertility, but as soils develop over time both natural and anthropogenic processes modify the initial fertility.

**Physical** aspects affecting soil fertility include texture, structure, and water-holding capacity. Soil texture is the proportion of sand, silt, and clay in the soil, while soil structure refers to the aggregation of the particles, affecting the soil porosity. Structure and texture interact to affect the ratio of water- to air-filled pores within the soil (plant roots respire and require oxygen), water movement in the soil, plant root growth, and the soil’s capacity to retain and release water and nutrients. Soil structure becomes more important as the yield potential of agricultural crops increases because many crops, especially spring-sown crops, have to produce a root system quickly to obtain the nutrients they require for optimum growth.

**Chemical** characteristics affecting soil fertility include pH, reduction–oxidation potential, cation exchange capacity (CEC), base saturation, carbonate content, salinity, and content of nutrients and potentially toxic compounds. Availability of both nutrients and toxic compounds are affected by soil pH and reduction–oxidation potential, as well as their reactions with soil organic matter, soil mineral particles, and one another.

**Biological** factors affecting soil fertility include the actions of many types of micro-organisms, micro-, meso-, and macrofauna, as well as plant root activity. The living soil microbial community plays a critical role in the development of soil from the original parent material and in the fertility of the soil after it has formed. Some organisms form symbiotic relations with plants; for example, *Rhizobium* bacteria form nodules on the roots of leguminous plants where they fix atmospheric nitrogen into forms that are transferred to and used by the plant. Similarly, mycorrhizal fungi have hyphae that enter the root of a host plant and also extend into the soil around the root. The hyphal system accesses a larger volume of soil than the roots alone, thus increasing the uptake of nutrients such as phosphorus and some micronutrients, a portion of which is released into the root cells for the plant to use. Such symbiotic associations, however, are not a “free service” to the host plant because the bacterium or fungus has to be supplied with energy in the form of carbon compounds by the host.

The types and abundance of organisms within soil are very dynamic. Populations can respond rapidly to changes in moisture, temperature, and available food sources. For example, microbial activity increases when a readily available food source is added, such as when crop residues are incorporated into the soil. Very importantly, the microbial breakdown (mineralization) of added organic material releases nutrients in plant-available forms, leading to the recycling of nitrate, phosphate, and other important plant nutrients. However, subsequent microbial activity may also form nitrate (nitrification) and convert the nitrate into gaseous nitrogen forms (denitrification) such as dinitrogen (N₂), nitrous oxide (N₂O, a greenhouse gas), and nitrogen oxide (NO₂) that are released to the atmosphere.

One of the end products of the microbial decomposition of raw organic matter is soil organic matter (or humus), which is the highly stable organic matter fraction remaining after most of the residue has decomposed (Soil Science Society of America 2014). The many beneficial properties of organic matter contribute to the fertility of the soil. Micro-organisms utilize organic matter as an energy source and can break down added crop residues, animal wastes,
biomass, and composts. The rate of breakdown often depends on the proportion of the different compounds in the organic material as carbohydrates are decomposed readily by bacteria, while lignin is broken down by basidiomycete fungi. The carbon to nitrogen ratio (C:N ratio) of the added material is also important, with decomposition proceeding more rapidly when the C:N ratio of the residue is narrow. In addition, the decomposition of residues with a narrow C:N ratio, such as alfalfa, releases plant-available nitrogen whereas decomposition of residues with a wide C:N ratio, such as cereal straw, removes plant-available nitrogen from the soil, at least during the initial stages of decomposition. Soil organic matter also plays a number of critically important roles in soil quality. Not only does it contain nutrients, but it can also adsorb nutrients and absorb large amounts of water. Thus, it helps soil to retain a reservoir of both nutrients and moisture for plant growth. Very importantly, soil organic matter and organic compounds produced during decomposition of added organic matter assist in the formation and stabilization of aggregates. The production of stable aggregates improves soil structure and tilth, which aids root growth, aeration, water infiltration and drainage, and resistance to compaction and erosion. Good soil structure provides an environment conducive for root growth and microbial activity.

Fertility depends on the available rather than the total nutrient content of the soil. Plant-available nutrients are the nutrient forms that the plant can absorb and utilize. Only a small fraction of the total nutrient content of a soil is immediately available for uptake by roots from the soil solution, the water surrounding the roots and soil particles. The concentration of soluble nutrients that are available for crop uptake in the soil solution is usually too small to meet the total requirement of a harvested crop. Consequently, the concentration of most nutrients in the soil solution must be replenished many times throughout the growing season from the large reserves of plant-available nutrients that are (i) adsorbed to the soil surfaces; (ii) in soil minerals or organic matter; or (iii) in soluble forms that can move rapidly into the soil solution.

The surfaces of soil organic matter and clay-sized mineral particles have both negative and positive charges enabling them to retain positively and negatively charged ions, respectively, through adsorption (the retention of atoms, molecules, or ions on the surface of solids by chemical or physical bonding) so as to maintain electrical neutrality. Plant nutrients are generally present in the soil as ions, either positively charged cations such as Ca$^{2+}$, K$^+$, NH$_4^+$, or negatively charged anions such as H$_2$PO$_4^-$, or SO$_4^{2-}$, and are retained in soil by being held on the adsorption sites. The CEC is the total amount of exchangeable cations that a soil can adsorb, or hold, on the negative exchange sites. Adsorbed cations can be readily exchanged with cations present in the soil solution. Cations that are not held on the exchange sites can move in soil water below the rooting zone of plants, so a high CEC can protect cations from leaching loss, but can also limit their mobility. Soils with a high CEC can hold more cations, which can replace those in the soil solution when they become depleted by root uptake. Therefore, soils with a higher CEC generally have more inherent fertility and greater productivity than do low CEC soils. Most of the soil’s CEC is permanent, while some is pH-dependent. A pH-dependent charge develops because, as soil pH increases, the concentration of H$^+$ ions in the soil solution decreases and that of OH$^-$ in the soil solution increases. In response, the H$^+$ ions from the hydroxyl and water groups at the edges of the clay colloids or the surfaces of oxides or soil organic matter can move into the soil solution or be neutralized by the OH$^-$ ions, increasing the
negative charge on the particle edge or surface. Conversely, in very acid soils (pH less than \( \approx 5.5 \)), kaolinite (a type of clay mineral) and iron and aluminum oxides in soil may develop positively charged sites that can hold anions, helping in the retention of phosphates and sulfates.

Nutrients in the soil solution are in equilibrium with the nutrients adsorbed on cation and anion exchange sites and those present in easily dissolved solid forms. This equilibrium is shifted when plants absorb nutrients from the soil solution, causing nutrients to be released into the soil solution (Figure 2). The ability of the soil to replenish the nutrients in the soil solution from the nutrients in the adsorbed or solid phase in the soil is the soil’s nutrient buffering capacity. Fertile soils have a high buffering capacity and are not depleted of nutrients as rapidly as less fertile soils.

Nutrients taken up by roots must be present in the soil solution or held on sites adjacent to the root. Nutrients can reach the sites on roots where absorption takes place through interception, mass flow, and diffusion. Interception occurs when the root grows and comes into direct contact with the nutrient, but roots normally contact less than 1% of the soil surfaces in the rooting zone. Therefore, only a small proportion of the nutrients present in the soil can be accessed by interception. Mass flow is another mechanism by which plant roots access soil nutrients. Mass flow is the movement of dissolved nutrients in the soil water, through the plant uptake of water from the soil pores. This convective flow of water to the root is driven by water uptake by the plant and its movement up through the plant in the transpiration stream. Mass flow is important for highly soluble nutrients, because the concentration of soluble nutrients in the soil solution, and hence the amount of nutrient moving to the plant root, will be relatively high. However, for sparingly soluble nutrients, such as phosphorus or potassium, movement to the root via mass flow will be small. Mass flow can also lead to leaching of nutrients below the plant rooting zone. Diffusion is the third mechanism of nutrient movement to plant roots. Diffusion is the movement of an ion through a water film from an area of high concentration to an area of low concentration. This form of movement is much slower than mass flow. Diffusion is important for moving low-concentration nutrients to the root surface; however, because of the slow rate of diffusion, the limited movement of nutrients by this pathway poses a substantial challenge for the plant. Therefore, a high density of roots is important to ensure that nutrients such as potassium and phosphorus that have low concentrations in the soil solution can be accessed.

Soil fertility will therefore be affected not only by the amount of nutrient present in the soil but also by factors affecting the ability of the plant to access the nutrient. Important contributors to plant nutrient uptake include factors affecting
root growth and function (e.g., compaction and temperature), factors affecting the supply and form of nutrient (e.g., pH and aeration), and factors affecting the mobility of nutrients (e.g., soil moisture and temperature).

Soil degradation

Soil degradation refers to the deterioration of soil properties and functions that lead to a decrease in soil quality, making it less suitable for purposes such as crop production. Soil degradation often leads to reductions in soil fertility. It has many causes, including nutrient depletion, soil erosion, compaction, salinization, acidification, and loss of organic matter.

Nutrient depletion from soils can occur when the nutrients removed in the harvested crop are not replaced. The risk may be minor with micronutrients if the soil reserve is very large and the amount removed by the crop is very small. However, with macronutrients such as nitrogen and phosphorus, nutrient removal is much greater, so there is greater risk of soil depletion. Nutrient depletion is a major concern in tropical soils, where the nutrient content of the soil is often inherently low and increasing population pressure has led to intensification of production, high rates of nutrient removal, and limited return of nutrients. Conversely, in developed regions, particularly in regions with intensive livestock production, excess accumulation of nutrients can occur if nutrient input is greater than removal. Buildup of nutrients in the soil can increase the risk of off-site movement and contamination of air and water.

Another major cause of soil degradation is erosion, where topsoil is removed by the action of wind or water, removing with it valuable nutrients and organic matter. Topsoil loss reduces soil fertility, impairs tilth, and can lead to environmental problems when nutrients are deposited into water bodies. Soil formation is a very slow process and topsoil loss from erosion cannot be easily corrected.

Degradation due to soil compaction can reduce soil fertility by limiting plant rooting, aeration, and water movement into and within the soil. Compaction can occur naturally, or through pressure caused by machine traffic.

Salinization is the accumulation of salts, generally occurring when saline water evaporates from the soil surface, leaving behind soluble salts. Salinity can develop from the use of poorly managed or low-quality irrigation water, where salts added in the water are retained in the rooting zone and not leached from the profile. It can also occur if drainage patterns are disrupted, allowing an elevated water table to bring salts to the surface. Similarly, under dryland conditions saline groundwater from a shallow water table can rise to the surface by capillary action. Dryland salinity is accentuated when water use by plants does not utilize the incoming precipitation, such as when deep-rooted trees are cleared and replaced by annual crops or when summerfallow is included in the rotation. The unused precipitation increases recharge to the groundwater, causing water tables to rise, leading to more evaporation and accumulation of salts in the rooting zone. High concentrations of salts in the soil reduce a plant’s ability to absorb water and nutrients and may cause direct toxicity.

Acidification is the reduction in soil pH. It is a natural part of the soil formation (pedogenesis), occurring when basic cations are removed from the soil by leaching. Acidification also results from nitrogen reactions in the soil, whether from applications of fertilizers or growth of legume crops. Over time, acidification can interfere with crop growth and nutrient availability.
Soil fertility management

The basic principle of soil fertility management is to adopt practices to allow economically optimum crop yields with minimum adverse environmental impacts. Long-term sustainability requires that nutrient losses and exports from the systems are balanced with nutrient input and recycling, and that the physical, chemical, and biological properties of the soil are not compromised over time (Figure 3).

Nutrient availability is maintained by applications of organic or inorganic fertilizers and, if the soil is acidic, by liming. An international program called the 4R Nutrient Stewardship (International Plant Nutrition Institute 2012) has been developed by the fertilizer industry to promote sustainable nutrient management. This involves providing the right source of nutrients to the crop at the right rate and the right time using the right method of application so that as many of the added nutrients as possible are utilized by the crop. Defining the “right” approach to each of these management factors involves the economic, social, and environmental dimensions of stakeholder goals. In 4R Nutrient Stewardship, managers consider these goals while selecting best management practices. Best management practices must be developed in relation to site-specific factors such as soil and climatic conditions, crop type, economic factors, and the agronomic practices of the producer.

Plant nutrients can be added as both organic manures and synthetic fertilizers. Synthetic fertilizers tend to contain a limited number of nutrients at a higher and strictly regulated concentration, as compared to organic manures. Most synthetic fertilizers provide nutrients in a soluble form that is quickly available for crop uptake. Organic manures, such as barn or farmyard manure, biosolids, crop residues, fishmeal, and other by-products, are normally lower in nutrient content than synthetic fertilizers, so a larger volume of material must be applied to meet crop nutrient demands, increasing transportation costs. The nutrient composition and availability of any type of organic manure will tend to vary from batch to batch, making it difficult to calculate the amount to apply to achieve a given rate of nutrient addition. Also, the balance of nutrients in organic fertilizers rarely matches crop requirements, leading to either over- or under-fertilizing with some nutrients. Plants absorb nutrients primarily in the inorganic form, so organic manures must decompose, usually by microbial activity, to release these inorganic ions before roots can take up the nutrients. As a result, the rate of nutrient release from organic manures is generally slower and less predictable than that from synthetic fertilizers. However, organic manures often contain a range of nutrients and are an excellent source of organic matter that can contribute both to the nutrient supply and to the physical and biological properties of the soil.

Application of adequate amounts of nutrients is essential to replace nutrients removed in the harvested crop, so as to maintain crop yields and to avoid nutrient depletion and soil degradation. However, environmental problems can also occur with excessive or poorly managed fertilizer inputs, whether from organic or synthetic sources. Emissions of reactive nitrogen to the atmosphere can harm the ecosystem and human health, by contributing to acidification, greenhouse gas production, eutrophication, formation of ground-level ozone and particulate matter in the atmosphere, and loss of biodiversity. Phosphorus movement to fresh surface water is a major cause of eutrophication. Trace elements such as cadmium that are naturally
Figure 3  Overview of nutrient behavior. Dashed lines represent nutrient gains or losses in the soil system; solid lines represent internal transformations within the soil system. Sustainable crop production requires that nutrient losses and exports in crop products are balanced by nutrient imports and recycling. Diagram courtesy of Dr Don Flaten, University of Manitoba.

Present in phosphate rock and phosphorus fertilizers can accumulate in the soil, leading to long-term soil degradation and human health concerns. Carbon dioxide emissions from the large amounts of fossil fuel used in fertilizer production and transport can also contribute to climate change. Therefore, proper soil fertility management is important not only to ensure sustainable food production over time, but also to prevent negative direct and indirect environmental effects. Nutrient management practices that match supply to the crop demand
closely, in terms of both rate and timing of the applications, can improve nutrient use by crops and reduce potential negative environmental impacts.

The importance of soil fertility management in ensuring economic crop production and avoiding negative environmental impact has helped to encourage the development of improved fertilizer products and practices. Enhanced-efficiency fertilizers utilize either physical coatings or chemical additives to slow their release into the soil solution or delay their chemical transformation in order to reduce losses and improve efficiency of crop use. Site-specific management techniques can be used to vary fertilizer application rates within the field, so that rates are matched more closely to crop requirements. Better methods of soil testing and in-field plant nutrient determination can more accurately determine whether nutrient applications are required. New genetic techniques are also being adopted to develop crops that are more efficient at accessing and using nutrients.

Soil fertility management is not just a matter of applying the correct amount of nutrients in either synthetic or organic form. An integrated soil fertility management strategy combines techniques to maximize nutrient use efficiency with practices to improve water use as well as overall agricultural productivity and sustainability. Local sources of crop residues, animal manures, biosolids, and municipal wastewater nutrients should be returned to the soil to help close the nutrient cycle by returning nutrients removed from the soil back to the soil–plant system. Yield constraints should be minimized by control of weeds, pests, and diseases. Use of crop cultivars with a high yield potential, a diversified rotation, and optimum seeding and tillage practices should be used to optimize crop yields and take the greatest advantage of nutrient and water resources.

Conservation agriculture practices, such as reduced tillage, should also be adopted to reduce the risk of soil degradation and to maintain long-term soil fertility. Such practices retain crop residues on the soil surface to protect it from erosion and to allow more water to infiltrate, rather than run off or evaporate. Conservation agriculture also aims to increase soil organic matter content by retaining more organic matter in the system. Other important soil fertility management practices include diversified crop rotations, production of legume crops, maintenance of soil cover with cover crops and mulches, retention of crop residues, use of available organic manures, including composts and green manures, and efficient use of inorganic fertilizers.

Conclusions

Soil fertility provides the base for a healthy and vibrant plant community. Improved understanding of the complex interactions between the physical, chemical, and biological factors that influence nutrient dynamics is critical to the development of effective nutrient management strategies for economically optimum production in managed ecosystems, while also minimizing negative environmental effects. As the world’s population grows, soil fertility management will play an increasingly important role in ensuring food security and the long-term sustainability of our natural and agricultural ecosystems.

SEE ALSO: Soil biology and organisms; Soil and clay mineralogy; Soil erosion and conservation; Soils and the carbon cycle; Soils on slopes: catenas; Soils and weathering
SOIL FERTILITY AND MANAGEMENT

References


Further reading


Soil mapping and maps

Matthew C. Bromley  
US Department of Agriculture, Natural Resources Conservation Service, USA

Bradley A. Miller  
Iowa State University, USA  
Leibniz Center for Agricultural Landscape Research (ZALF), Germany

A soil map portrays information on the locations and properties of soils. There are two main components to a soil map: spatial and tabular. The spatial component is the map itself, consisting of polygons (or raster cells) that delineate soil types as map units. Map units are cartographic entities that help simplify the map by grouping similar delineations of soils. Each map unit is symbolized to correspond to a legend. The tabular component of a soil map is a summary of the range of physical and chemical properties for each soil type along with interpretive information based on those properties, for example, capabilities or limitations for crops, construction, and sanitary facilities. Today, the tabular data are typically stored in a soils information system. For example, in the United States this system is called NASIS or the National Soils Information System.

The process of soil mapping includes gathering information about the soil landscape and communicating that information in a format suitable for the end user. Part of that process involves synthesizing the available and known information and condensing it to meet the requirements of the map. Classification is a means for simplifying the complex soil landscape into easily communicated concepts and units. Even though the soil landscape is complex, practical limitations almost always restrict direct observations to a small fraction of the landscape. Because of this limitation in sampling, the soil mapper must have a systematic means for extending observations taken at a single point to a greater area. This is the essence of many forms of mapping.

The characteristics of a soil at a single location, that is, a pedon, can be described using a set of established taxonomic definitions for relating what is observed (e.g., World Reference Base or the US Soil Taxonomy). A set of pedons can be grouped together by similar characteristics and described by defined taxonomic units called soil series, taxadjuncts, or phases. However, as more pedons are grouped together, the heterogeneity of the spatial unit is likely to increase. Synthesizing this information is the challenge of the soil mapper. To successfully manage this challenge, the soil mapper must have a deep understanding of natural processes on the landscape and the ability to organize the information in a useful manner.

History and purpose of soil maps

Beginning at least by the fourth century, soil maps have been linked to the valuation of land, which during early times was mostly dependent upon soil productivity. By the late nineteenth century, more emphasis was placed on the soil maps as a source of information for determining crop suitability and best management practices (Krupenikov 1993). In the early twentieth century, interest in soil maps grew beyond the agricultural sector. Soil maps were able to...
SOIL MAPPING AND MAPS

provide valuable information for residential and commercial development, locating sources of aggregate, sand, clay, and other building materials, and helped guide land conservation efforts. Interpretative (derivative) maps created from soil maps can spatially represent a variety of properties such as erodibility and corrosivity, and suitability for land uses such as for building construction, septic tank drain fields, wildlife habitat, and road construction. Today, as environmental issues rise in importance and as recognition of soil’s role in ecosystem services increases (e.g., water quality, carbon sequestration), new demands are being made from soil maps.

Although regions of the world were described by the characteristics of their soil by the ancients, the earliest known linking of soil attributes to a map was for cadastral purposes. These maps were focused on local areas, and map units were based on property boundaries rather than the actual distributions of the soil properties. Soil maps that would be widely published required accurate base maps onto which the thematic information could be drawn. The base maps were needed to provide location references for transferring observations to the map. The increased availability of accurate topographic maps at the turn of the nineteenth century coincided with the dawn of soil mapping as we know it today (Miller and Schaetzl 2014). As an example, when the US Soil Survey began in 1899, a lack of available topographic maps meant that soil mappers had to rely upon plat maps for spatial reference, or had to create their own base maps with a plane table and alidade (Lapham 1949). These early surveys were performed by crews camping and traveling on horse, sampling the soil at regular spatial intervals. Soil maps produced by this method had a cartographic scale of 1:63,360 and generally had less than 10 map units. As better base maps became available, soil maps became more accurate and detailed. Efficiency also increased, because more preliminary work could be done to guide field sampling. The advent of aerial photography for map making greatly enhanced soil mapping efforts by providing a better spatial reference, thereby improving mapping consistency (Bushnell 1929). The availability of aerial photographs as base maps also improved efficiency, because they included useful parameters for predicting soil properties (e.g., vegetation, topography). Today, the soils in almost all of the United States (more than 8 million km² of soil maps) have been mapped at least once, mostly at a map scale between 1:24,000 and 1:12,000. A majority of these maps were done on a county or multicounty basis and were drawn on aerial imagery base maps.

The progress of soil mapping in other parts of the world has varied, largely dependent upon government investment and landscape accessibility. Nonetheless, work has already been done to merge the various soil mapping efforts from around the world. Today, work is being done to leverage new technologies to produce an improved digital soil map of the world (Sanchez et al. 2009). A global map that is more accurate and contains more soil information is beneficial for addressing issues such as global climate change and global food supplies. At the same time, there has been a resurgence in interest for soil mapping at national levels as well, but with a greater emphasis on using remote sensing and geostatistical techniques to reduce the amount of field work required to create the map.

Fundamental concepts in soil mapping

Soil mappers rely on a certain degree of spatial autocorrelation in order to group areas on the map as homogenous delineations. It is also helpful to recognize that each delineation is not completely unique. In areas where the
original parent material is similar and has been subsequently modified by similar environmental conditions, soils in them are expected to have developed similar soil properties. This concept is expressed as the state factors of soil formation, which soil mappers use to predict the distribution of soil properties and to group similar soils. These state factors are identified as climate, organisms, relief, parent material, and time (Dokuchaev 1967/1883). Observing soil properties at every location on the landscape is a practical impossibility, for example, a mapper would need to make approximately 10,000 observations with a standard bucket auger in order to sample even 1% of the soil in a half hectare of land. Therefore, when all five state factors are relatively uniform, the soil mapper can make assumptions about the similarity of the surrounding soils to the soil directly observed. Similarly, by spatial association, the soil mapper can make a similar assumption about other parts of the landscape that also have the same state factors. This relationship between soils with similar state factors allows for multiple delineations to be described together as a single map unit. Thus a soil map with hundreds of delineations can use a fairly simple legend. Map units are also linked to tables or databases that provide information about each map unit. These tables add many dimensions to the soil map as the information about soil properties, and derived interpretations, have infinite possibilities.

To make determinations about the uniformity of the state factors across the landscape, soil mappers rely on a combination of field observations and a variety of base maps. One of the most valuable base maps currently used in soil mapping is the aerial photograph. These base maps not only provide spatial reference but also provide information about vegetation and topography – two of the most important state factors at the local level. Further, topography often provides indications of parent material, and parent material is generally related to time. Also, vegetation communities often develop in response to the climate, topography, and parent material. Therefore, aerial photographs provide soil mappers with a wealth of information for identifying areas expected to have similar soils.

The soil mapper draws map units that are as homogenous as possible, under the constraints of map scale, available information, and time. In practice, it is difficult to delineate perfectly homogenous map units, and thus, certain compromises must be made in the translation of the real world’s complexity to the map representation. In actuality, the spatial variability of the soil landscape makes it impossible to draw a completely accurate soil map. It is the goal of the soil mapper to produce the most accurate map possible for the intended purpose, within the constraints of the resources available.

Although soil mappers strive to gather as much information as possible for the mapping area, limitations in base map resolution, accessibility, and time are a constant reality. Map scale, which determines the minimum size delineation of the map, strongly influences soil mappers’ decisions for dividing an area into multiple delineations, or grouping similar soils into delineations. Because there is sometimes the need to group soils with known differences, there are different types of map units, with differing degrees of cartographic purity. When delineations can be made, such that the majority of the area meets the definition of the map unit (i.e., soil series), the map unit is considered to be a consociation. In cases where there is more variation in soil types than can be delineated separately, but there is a predictable pattern of the component soils, then the map unit is considered either a complex or an association. The term complex is used for map scales larger than 1:24,000, and the term association is used for smaller map scales. All of these map
units are named by their constituent soils. When the variations in soil types do not have a spatial pattern, but the soils can be grouped because their differentiation is not important for the purpose of the map due to some overriding property such as flooding, then the map units are referred to as undifferentiated groups. In all of these cases, any dissimilar inclusions in the map units are noted, along with their respective composition percentages. These minor components may be important in differentiating between two map units which have the same major component, but which occur on two different landforms. For example, the hypothetical Alpha series may be mapped as a coarse-loamy deposit on both till plains and outwash plains, but the map units on the two different landforms may have very different minor components.

Soil maps can differ in intensity, as determined by the intended uses of the finished product. Lower intensity mapping, such as may be used in forests, requires fewer observations per land area and results in a less detailed map. This lower level of detail is reflected in the scale of the map, the number of soil map units delineated, and the sizes of the delineations. Higher intensity mapping is typically performed in areas where more specific soil information is needed, as in agricultural landscapes. The US Soil Survey recognizes five orders of soil mapping intensity (Table 1).

Producing the soil map

Soil mapping starts in the office. Before heading to the field to start digging, reference materials including aerial photographs, digital elevation models and maps of bedrock geology, landforms, vegetation, and topography are assembled and studied. Before digital elevation models became common, stereoscopes and stereo-pairs of aerial photographs were used to evaluate slope steepness and shape. Using these data, soil patterns begin to emerge before a single hole has been dug. Depending on the area being surveyed, different attributes of the various base layers become important in attempting to predict soil patterns. In this way, the soil mapper develops a preliminary mental model for how soils are distributed in the area. This model may or may not eventually be recorded in the form of a block diagram or other graphical representation (Figure 1). The soil mapper then applies this initial model to delineate potential differences across the soil landscape, either by manually drawing them on aerial photographs or by sketching them digitally using a geographic information system (GIS). Thus, preliminary work done in the office greatly increases the efficacy of fieldwork. However, subsequent field investigations remain critical for validating, or recalibrating when needed, the mental model developed by the soil mapper (Figure 2).

After the preliminary linework has been drawn, the soil mapper begins to test and apply their mental model in the field. To do this, they plan traverses to verify expected soil patterns and to explore areas where more information is needed. The term traverse is used to refer to this stage of the mapping process when preliminary models are being tested. Most traverses are designed to cross the landscape in a way that allows the soil scientist to see an entire toposequence of soils – soils on the uplands to the lowlands – and those in between. Sometimes, just getting to the survey area is a challenge, requiring kilometers of hiking or boating, using trucks, all-terrain vehicles, and other modes of transportation. In more remote areas, overnight camping is necessary. As they walk across the land, soil mappers dig or auger holes where changes in soil characteristics are likely. These changes could be variations in vegetation, landscape position,
Table 1  Different intensities of soil surveys (modified from Soil Survey Staff 1993).

<table>
<thead>
<tr>
<th>Soil mapping intensity</th>
<th>Example purpose</th>
<th>Field procedures</th>
<th>Minimum size delineation (ha)*</th>
<th>Typical map units</th>
<th>Appropriate map scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st order</td>
<td>Experimental plots or individual building sites.</td>
<td>Soils are identified by transecting or traversing. Soil boundaries are observed throughout their length. Remotely sensed data are used as an aid in boundary delineation.</td>
<td>1 or less</td>
<td>Mostly consociations; a few complexes and miscellaneous areas possible.</td>
<td>1:15 840 or larger</td>
</tr>
<tr>
<td>2nd order</td>
<td>General agriculture, urban planning.</td>
<td>Soils are identified using field observations and remotely sensed data. Boundaries are ideally verified at closely spaced intervals.</td>
<td>0.6 to 4</td>
<td>Consociations and complexes; few undifferentiated groups.</td>
<td>1:12 000 to 1:31 680</td>
</tr>
<tr>
<td>3rd order</td>
<td>Range or community planning.</td>
<td>Soil boundaries are plotted by observation and interpretation of remotely sensed data. They are verified by traversing representative area and by some transects.</td>
<td>1.6 to 16</td>
<td>Mostly associations, some consociations and undifferentiated groups.</td>
<td>1:20 000 to 1:63 360</td>
</tr>
<tr>
<td>4th order</td>
<td>General soil information for broad statements concerning land-use potential and general land management.</td>
<td>Soil boundaries are plotted by interpretation of remotely sensed data. They are verified by traversing representative areas and by some transects.</td>
<td>16 to 252</td>
<td>Mostly associations, some consociations and undifferentiated groups.</td>
<td>1:63 360 to 1:250 000</td>
</tr>
<tr>
<td>5th order</td>
<td>Regional planning, selections of areas for more intensive study.</td>
<td>Soil patterns and composition of map units are determined by mapping representative ideas and like areas by interpretation of remotely sensed data. They are verified by occasional onsite investigation or by traversing.</td>
<td>252 to 4000</td>
<td>Associations; a few consociations and undifferentiated groups possible.</td>
<td>1:250 000 to 1:1 000 000 or smaller</td>
</tr>
</tbody>
</table>

*This is about the smallest delineation allowable for readable soil maps produced in a paper format. In practice, the minimum size delineations are generally larger than the minimum size shown.*
Figure 1  Example of a block diagram, which communicates the soil mapper’s broad-scale conceptualization of the soil landscape (Minger 1991). Although such diagrams communicate much information about the soil landscape, they do not fully capture all the parameters that a soil mapper integrates to produce the final soil map.

slope, or other observable differences that may or may not have been evident from the base maps examined in the office. It is important that these field observations are tied to specific locations. Today, soil mappers track their location using global positioning system (GPS) units.

Field observations are typically made using hand tools such as spades, bucket augers, and probes (Figure 2b). When practical, motorized tools such as truck-mounted probes and backhoes can also be used. Terrain, accessibility, and soil types will determine which tools are best suited for the job. As the soil mapper excavates the soil profile, they record changes such as color, texture, density, rock fragment content, as well as concentrations of salts and other minerals. They also look for a direct observation of the water table and/or redoximorphic features that may indicate a seasonally high water table. Altogether, this information will help the soil mapper to classify the different soil types as defined by soil series. A soil series is the finest level of classification and akin to classifying a plant or animal to the species level. Soil series are represented by a range of soil properties linked to soils that share the same parent material and state factors. Although concepts of soil series are defined by the state factors, classification of a soil into a series is done by specified ranges of soil properties to avoid issues with interpretation.
Soil mapping is a process that includes diligent work both in the office and in the field. The process (a) begins with research and careful examination of available base maps, which (b) guides investigations in the field, and (c) is completed in the office using GIS software.

Once the mapper finishes with fieldwork, they return to the office to summarize their findings and adjust their preliminary maps. As observations are accumulated, the soil mapper’s mental model is calibrated, strengthened, and refined. Eventually, emerging patterns form the basis for a working legend, or list of soil series found in the survey area. A key can then be developed to help the mapper more quickly identify the appropriate soil series at future field sites, based on site conditions and important soil characteristics. Once the commonly occurring soil series have been identified, representative sites are chosen and a representative soil is described in more detail. At this point, samples can be taken for laboratory analysis of soil chemical and physical properties that are not observable in the field. After all the map units have been delineated and soil types have been defined, transects (making soil observations at regular intervals) may again be performed to quantify map unit composition.

In the past, soil maps were commonly hand-drawn on aerial imagery, compiled to orthophotography, and then digitized. Modern soil mapping methods encourage linework to be recorded in a digital format from the very beginning, eliminating several steps that can potentially introduce error into the final product. The modern digitizing process usually involves “heads-up” editing, which uses a combination of GIS reference layers and field observations. Digital map finishing, including the addition of other data layers such as roads, streams, and political boundaries, is usually performed by cartographers before final publication of the map, in order to give the map reader a way to relate the soil boundaries to key geographic landmarks (USDA-NRCS 2014a). In the past, soil maps were published as hardcopy books. More recently, those maps have been digitized and published on CDs/DVDs or made available on the Internet as either a downloadable product or interactive maps (Figure 3). The benefit of online soil maps is that information for large areas across political boundaries can be organized in one location and made accessible to a general audience. The technology of online GIS is
Figure 3  Screenshot from the US Web Soil Survey (USDA-NRCS 2014b). This soil map is in an online GIS that allows users to select an area of interest, access map unit information stored in the database, and download the data in a variety of formats.
developing rapidly and will continue to change how we interact with soil maps in the future.

Future of soil maps

Soil maps will continue to be updated and improved, especially as to the level of map detail and accuracy, with some expansion into previously unmapped areas. Future soil mapping will continue to deal with the constraints of time and money that can be spent on a single project. However, soils will be more accurately described as more time is spent assessing soil properties with field and laboratory testing. Many current soil maps can be improved with further observations and GIS modeling techniques, using remote sensing and digital elevation models.

A major change for soil maps in the future will be the shift from polygon to raster data models in a digital environment. The raster data model is gaining in popularity because of its efficiency at storing large amounts of information. Data storage efficiency is important for covering larger areas with more detailed information. Because of this, raster data models are better suited for applications such as precision agriculture and environmental modeling. There is a push in the United States to create raster versions of the current vector-based soil maps as a first step in making digital soil maps available. In other parts of the world, where soil mapping in the field has not been as intensive, new digital techniques are being emphasized to develop soil maps that more fully utilize the resolution capabilities of the raster data model.

In addition to the changes in the production process and formatting, the purposes of soil maps continue to evolve and expand. Highly urbanized areas were often initially dismissed when soil surveys of the more rural surrounding areas were taking place because of their lack of agricultural potential. In recent years, the demand for soils information in these areas has increased; major cities like Berlin, London, New York, and Detroit have been surveyed or such surveys are in progress. Urban soil maps are being used for a variety of city planning purposes. Due to the unique nature of urban environments, soil classification systems have recently been amended to accommodate urban soils.

Conclusions

Despite the fact that we live in a rapidly evolving world of technology, knowledge of the very ground beneath our feet will remain an important part of human life. Soil maps communicate knowledge about soil resources, which is essential for sustainable management and conservation of soil along with the other natural resources that interact with it. Although the monumental efforts of mainly the past 100 years have produced an excellent repository of soil maps, especially in the United States, these maps need to be regularly updated to incorporate current understanding of soil science, utilize modern technologies, and reflect changes that have occurred in the landscape.

SEE ALSO: Digital soil mapping and pedometrics; Soils on slopes: catenas; Soils of urban and human-impacted landscapes

References

SOIL MAPPING AND MAPS


Further reading


Soil mass balance

Kyungsoo Yoo
University of Minnesota, USA

Soils form at the intersection of the lithosphere, biosphere, atmosphere, and hydrosphere. They are comprised of a solid phase, consisting of minerals and organic matter, a liquid phase consisting of water and dissolved constituents, and a gaseous phase. Quantitatively assessing the fluxes (i.e., moving masses) of materials within and across the boundaries of a soil body, and the coupled changes in the total inventory and distribution of soil constituents, defines the soil mass balance approach (Figure 1). This is a key endeavor in understanding soil genesis and the relationship of soils to terrestrial ecosystems. Defining a soil system is synonymous with delineating its boundary, a task that is limited only by the researcher’s ideas and goals. The system may be a volume of soil material, various depth intervals, a horizon or a set of horizons, or the entire soil profile. As much as a mass balance can be constructed for soil systems of varying spatial scales like these, it can also be constructed across different time scales. For example, a flux that is crucial in describing soil mass balance in days to months may be less significant when addressing soil genesis over time periods of thousands to millions of years, but may be key to effective soil managements. Therefore, choosing the most appropriate fluxes for the soil constituent of interest depends upon the temporal and spatial scales of the research question.

Simonson (1959) described a soil’s material composition as a result of four main suites of pedogenic processes: additions, losses, translocations, and transformations. This view has functioned as a fundamental precept for the soil mass balance approach in understanding soil genesis and biogeochemistry for decades. For instance, research on the coupled processes of eluviation and illuviation has sought to close the mass balance of iron and clay by describing vertical translocations and transformations, and their association with organic and mineral materials. Similarly, hydrochemists measure cations in soil water exiting the soil profile and use the data to infer biogeochemical changes in soils by utilizing mass balance concepts. Similar approaches can be done on watershed scales. Additionally, concerns about soil erosion as a threat to soil sustainability begin from the consideration of soil mass balance.

Widely used soil mass balance approaches

Moving beyond concepts and theory to define explicitly quantitative treatments of soil mass balance is an integral part of current research on soil genesis and biogeochemistry. For the soil constituents modeled by mass balances, analytical instruments must first be used to determine their concentrations in units of mass per mass or mass per volume. This approach provides a dilemma when the research question focuses on the abundance of a constituent, for example clay or iron. Because concentration is a relative measure, a soil constituent’s concentration may increase despite some of the constituent being lost from the soil, due to greater losses of other soil constituents. For example, the concentration of iron may be greater in a soil than in its parent...
SOIL MASS BALANCE

Input flux of $j$

$I_j$

Output flux of $j$

$O_j$

Figure 1 A soil system exchanging its constituent, $j$, with its surrounding environment. The mass change (i.e., $\Delta M_j$) in the soil system, if its transformations to other constituents do not occur, is equivalent to the mass difference between the input and output fluxes (i.e., $I_j - O_j$).

By calculating mass fluxes, the extent of cumulative losses or additions of various soil constituents during soil formation can be determined. When the calculations are made for individual horizons within a soil profile, the vertical distributions of the fluxes inform us of the long-term translocations of soil constituents within the profile. These horizon-based fluxes may then be integrated for an entire soil profile to characterize soil genesis in terms of the evolution of soils from parent materials with associated mass fluxes. By applying this approach to soils with well-constrained soil-forming factors and of known age, researchers can develop a highly quantitative understanding of the environmental controls on soil-forming processes, and on rates and pathways of pedogenesis. The geochemical mass balance approach therefore offers a powerful means to quantify the processes depicted in Simonson’s (1959) conceptual view of soil formation.

Wide application of the geochemical mass balance to characterizing elemental behaviors in soil formation is partly facilitated by the routine procedure of obtaining oxide concentrations in

where $h_p$ is the thickness of parent material that has been converted to a soil horizon with the thickness of $h_w$, $m_{j,\text{flux}}$ is the mass loss (negative) or gain (positive) of constituent, $j$, during the formation of the soil horizon [kg m$^{-2}$], $C_{j,w}$ is the mass fraction of the constituent, $j$, in the soil horizon [kg kg$^{-1}$], $C_{i,p}$ is the mass fraction of $i$ in the parent material [kg kg$^{-1}$]; $\rho_p$ is the bulk density of parent material [kg m$^{-3}$]; $\rho_w$ is the bulk density of the soil horizon [kg m$^{-3}$]. From the equations 1 and 2, the following equation 3 is derived for calculating $m_{j,\text{flux}}$ in the soil horizon:

$$m_{j,\text{flux}} = h_w \rho_w \left( C_{j,w} - \frac{C_{i,w}}{C_{i,p}} C_{i,p} \right)$$ (3)
soil and geological materials. The mass balance approach has been also modified for situations where aeolian inputs are significant, or for soils that are continuously rejuvenated by erosion. Although the application of the quantitative mass balance has been largely for elements with pedologic or ecological significance, such as silicon, iron, aluminum, calcium, phosphorous, and potassium, the method has the potential to be applied to quantifying the long-term inputs of elements acquired from contaminants, or from agricultural fertilizer or other kinds of soil amendments.

In contrast to the geochemical mass balance model that has been largely used for soils on geomorphically stable landscapes, a physical mass balance model has been developed for actively eroding landscapes. In eroding landscapes, soils function as a critical component of landscape evolution because they are the location where physical and chemical weathering are most active. A *geomorphic soil mass balance* relates a soil to a topographic location on a hillslope, and the soil's mass is constrained by (i) colluvial inputs (from upslope) and outputs (to lower hillslope areas) as well as (ii) soil production from the underlying geological materials (Figure 2).

\[
\frac{\partial h}{\partial t} = -\nabla \cdot Q - \rho_r \frac{\partial e}{\partial t}
\]  

(4)

where \( h \) is soil thickness (m), \( Q \) is the volume of colluvial flux crossing a unit contour line (m³ m⁻¹ yr⁻¹), \( e \) is the elevation of soil to bedrock boundary (m), \( \rho \) is bulk density (kg m⁻³), and the subscripts \( s \) and \( r \) represent soil and bedrock, respectively. The elevation change at the soil-bedrock boundary (i.e., \( -\frac{\partial e}{\partial t} \) in equation 4) is frequently referred to as *soil production*, as it is equivalent to the material flux from the bedrock to the overlying regolith and soil. The soil erosion rate is calculated as the difference between colluvial outputs and inputs, and soil losses via erosion are compensated by soil production in a steady-state system. When landscape morphology is a primary concern, this volumetric representation of colluvial flux and soil production provides a strong theoretical foundation for geomorphological research.

This simple mass balance approach has been central not only to numerically modeling landscape evolution but also in motivating researchers to seek key feedbacks in geomorphic systems. This approach has a long conceptual history in geomorphology, dating back to G.K. Gilbert (1909). A soil geographer, Donald Johnson (Johnson, Domier, and Johnson 2005), independently proposed the surprisingly similar concept of dynamic denudation, where physical interactions between soil organisms and properties of soils and geological materials, via bioturbation, were emphasized as responsible for producing, mixing, and eroding soils and thus forming soil and landscape morphology. It is, however, the recent widespread application of radioisotope dating tools that has assisted the expansion of the application of the model. Among these dating tools, the most widely used include cosmogenic radionuclides such as \(^{10}\)Be.
SOIL MASS BALANCE

\( t_{1/2} = 1.3 \text{ million years} \) that accumulates as atmospheric fallout in soils. With isotopic dating tools such as this, researchers have begun to constrain the turnover time of the soils on eroding landscapes. The outcomes, when combined with this mass balance, facilitate the quantification of functional relationships between soil production rates and soil thickness (i.e., the soil production function) or the relationship between colluvial flux and topographic attributes such as slope gradient and curvature.

The previous two mass balance approaches have focused primarily on geologically inherited materials in soils. Soil carbon mass balances have developed largely independently, but their importance has grown sharply because of the significance of soils in the global carbon cycle. Soil organic carbon storage has been modeled as the balance between plant carbon inputs and losses via the microbial decomposition of organic matter. One well-accepted view is that decomposition of organic matter follows first order kinetics, such that decomposition loss rates increase in direct proportion to the amount of organic matter present in soils. This concept is expressed in this mass balance:

\[
\frac{dC}{dt} = I - kC \tag{5}
\]

where \( C \) is the soil organic carbon storage \([\text{kg m}^{-2}]\), \( t \) is time \([\text{yr}]\), \( I \) is the plant carbon input rate \([\text{kg m}^{-2} \text{ yr}^{-1}]\), and \( k \) is the decomposition rate \([\text{yr}^{-1}]\). One of the direct implications of this model is that the greatest accumulation rates of carbon are observed in young soils over the course of soil formation, a point that has been confirmed in field observations using a soil chronosequence (Schlesinger 1990). Despite the simplicity, this model offers an explanation of the global pattern of soil carbon storage as a balance of net primary productivity and decomposition (Amundson 2001).

Over the past decades, primarily driven by the societal needs to better constrain the soil carbon pool, the soil carbon mass balance has become increasingly sophisticated, while its basic conceptual structure as written in equation 5 has been maintained. New developments include but are not limited to: (i) diversifying pathways of carbon fluxes, (ii) the division of the soil carbon pool into subpools with differing susceptibilities to decomposition, and (iii) the incorporation of a rapidly evolving mechanistic understanding of environmental controls on decomposition rates.

Criticisms and future of soil mass balance approaches

The above mass balance models, despite sharing soil systems as a common object, have been advanced by different scientific communities, each with unique research interests and agendas. One direct consequence of this is the differing – often described as conflicting – views of the boundaries of soil systems. For example, soil scientists, who are familiar with pedogenic processes operating within a three-dimensional continuum from soil to parent material on stable landforms, may find difficulty in adopting the term “soil production,” which geomorphologists use to refer to mass fluxes across the boundaries between bedrock and physically perturbed colluvial soil on actively eroding landforms. Likewise, researchers studying soil carbon have largely focused on balancing carbon in the topmost layers of soil profiles that are accessible and thus easy to replicate, while more recent data demonstrate the globally significant presence of subsoil and deep subsurface carbon. One might criticize these discrepancies as limitations of the mass balance approach. However, the ability to identify these poorly quantified zones of soil and critical zone research has been made possible due to existing mass balances. Soil mass balance
SOIL MASS BALANCE

has been therefore, in hindsight, an essential tool to advance our understanding of soil genesis and biogeochemistry.

New, important findings in soil genesis are being made increasingly at the interfaces of two or three mutually excluding mass balances, or by merging them. The unification of geochemical and geomorphic soil mass balances will allow for the expansion of quantitative assessments of soil biogeochemistry to rugged or marginal landscapes that are under increasing pressure for urban and agricultural uses, especially in developing countries. The same integration may also benefit geomorphologists in better understanding the biogeochemical impacts of landscape evolution, such as the long-postulated hypothesis that orogenic activities have contributed to global cooling events in the geologic past. Lastly, on the issue of the global carbon cycle, emerging interests in the nonbiological fluxes of carbon transported via colluvial and alluvial sediments, and via inland waters, call for the integration of all three solid phases (geochemical, geomorphic, and biological) in a soil mass balance. Therefore, the development and application of increasingly complex soil mass balance models is an integral part of ongoing research which should produce novel insights into issues related to soil sustainability, landscape evolution, the global carbon cycle, and human health.

SEE ALSO: Soil development: conceptual and theoretical models; Soils and the carbon cycle; Soils in geomorphic research; Soils on slopes: catenas; Soils and weathering

References


Further reading


Soil taxonomy and soil classification

Craig A. Ditzler
Jon Hempel
USDA NRCS National Soil Survey Center, USA

Overview of the system

Soil taxonomy replaced the classification system of the United States that was published in 1938 (Baldwin, Kellogg, and Thorpe 1938). There were several reasons for the change. Classes in the 1938 system focused on their central concept, but the limits of the classes were not well defined, making it difficult to classify soils that had properties intermediate between two classes. Class definitions were based on the prevailing notions of soil genesis at the time, but where genesis was unknown or in dispute, soils could not be classified consistently. Classes were defined on native conditions, meaning that some cultivated soils could not be easily classified. Due to these and other problems, it was determined that an entirely new system was needed.

In response, soil taxonomy was developed by the United States Department of Agriculture (USDA) in the 1970s, and has been updated several times. The effort was initially led by Dr Guy D. Smith from the 1950s to the early 1970s. Successive drafts of the system were developed and field tested within the United States and other countries. It had several specific objectives (Smith 1963). It needed to organize the several tens of thousands of kinds of soils known to exist around the world into useful classes whose members share key properties considered important for understanding the soil’s response to use and management. Also, the classes were designed to reflect, but not be directly defined by, soil genetic processes because processes cannot be observed directly. Very importantly, the classification system was designed to fulfill the practical goal of supporting soil-mapping projects of the National Cooperative Soil Survey (NCSS) in the United States and similar efforts in other countries.

In achieving these goals, soil taxonomy has organized our knowledge about soils and improved our ability to communicate about them. It has helped us to understand relationships between different kinds of soils. It has enhanced our ability to obtain information through experimentation and observation on a few selected soils and to extend and apply that knowledge to other soils in the same or similar classes.

The system was first published in 1975 as Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys (Soil Survey Staff 1975). A second edition was published 24 years latter (Soil Survey Staff 1999). Since then, several revised versions of the taxonomic keys have been published and are available online (Soil Survey Staff 2015a). Each successive version reflects the expanding knowledge of soils around the world.

Grouping soils into classes, by its very nature, reflects the intended purpose and goals of whoever may be designing the classes. For example, an ecologist might group soils by their natural ecological types – tropical rainforest soils, prairie soils, hardwood forest soils, desert soils, and so on. A surficial geologist might group soils based on their parent material – glacial till soils, lacustrine sediment soils, weathered granitic rock
SOIL TAXONOMY AND SOIL CLASSIFICATION

soils, eolian sand soils, and so on. An engineer might group soils according to their particle-size distribution – sandy soils, gravelly soils, clayey soils, and so on. There are as many ways to classify soils as there are purposes for doing so, and thus many soil classifications systems exist around the world.

Soil taxonomy is a morphogenetic system; it is based on the morphology and inferred genetic history of the soil. Soil scientists studying the origin and characteristics of soils (pedologists) consider soil as a three-dimensional natural body. Soils form a continuum extending both laterally across the landscape, and to some depth (ranging from less than one to several meters). Successive layers or soil horizons are formed in the soil due to natural soil-forming processes, acting over time, on an initial parent material.

How soils form and why we can map them

The morphogenetic view of soil was first proposed by the Russian scientist V.V. Dokuchaev in the late nineteenth century. It was expanded upon in the United States by scientists such as Coffey, Marbut, and Whitney in the early twentieth century. As knowledge of soil genesis grew, earlier concepts of soil as simply the weathering product of geologic materials were expanded to recognize that soil is the result of the interaction of five fundamental factors of soil formation: parent material, climate, organisms, relief, and time. Parent material is present at the starting point of the soil formation process. Examples include alluvium, wind-blown sands, glacial deposits, weathered bedrock of various types, colluvium, thick deposits of organic matter (as in bogs), and many others. Parent material provides the initial physical and chemical properties to the soil. Climate, especially precipitation and temperature, is an active soil-forming factor that regulates the intensity of parent material weathering and translocation of materials within soils. Freeze/thaw cycles promote the physical breakdown of mineral particles. Temperature regulates many biological and chemical weathering and transformation processes. Precipitation supplies water to dissolve and translocate mineral and organic constituents from the surface to deeper soil layers, and affects the kind and amount of vegetation that can be supported. Organisms, which include microbes, earthworms, insects, burrowing animals, plants, and humans, also affect soil development. They contribute to soil mixing, cycling of nutrients, and add organic material to the soil. Relief (landscape position) further influences the rate and intensity of soil formation locally. Low-lying, concave positions receive runoff water and sediment from higher positions, thus increasing the water available to move into and through the soil. Higher, convex positions tend to shed water and sediment, thus lessening the amount of water entering and passing through the soil. Topographic aspect (the compass direction that a slope faces) is an important component of the relief factor. In the Northern Hemisphere, south- and west-facing slopes receive the most direct solar radiation and tend to be warmer and dryer than north- and east-facing slopes.

The longer these soil-forming factors act upon the soil, the more the original parent material is changed. Gradually, the soil comes to reflect the state factors that interacted to form it, until in some soils a state of equilibrium with its environment is achieved. Some soils have been forming for relatively short time periods, such as those developed in recent eolian or alluvial deposits. Others are thousands of years old, such as those formed in materials deposited by the last continental glaciers. Soils on very stable landform surfaces that have remained relatively unchanged
for long periods of geologic history are extremely old and weathered.

The five factors of soil formation interact to control the processes responsible for soil formation. These processes include physical and chemical weathering, biological activity, chemical reactions, additions of new material, translocation of constituents within the soil, and leaching of material out of the soil. The effect of these ongoing processes is expressed by the development of distinct layers (soil horizons) that can be observed in many soil profiles (Figure 1). These horizons differ from one another in any number of properties such as thickness, color, texture, structure, salinity, pH, fertility, mineralogy, consistency, organic matter content, and so on. Soil horizons are the morphological expression of the interaction of the factors of soil formation at that location, and form the basis for soil classification.

A fundamental tenet of soil science states that where the factors of soil formation are the same (or similar), the soil horizons of that soil body will be the same (or similar). As a consequence, soil bodies tend to occur in repeating patterns in the landscape, because the state factors occur in predictable, repeating patterns. If one understands the local factors of soil formation, the kind of soils in the area can be understood, predicted, and mapped (Figure 2). The fact that soils form predictable natural bodies makes mapping their spatial distribution and geographic location possible, and at a variety of scales (Hudson 1992). In the United States, the NCSS has been conducting soil mapping for over 100 years, and maps are available for most of the United States (http://websoilsurvey.nrcs.usda.gov/). Soil surveys have also been conducted in many other countries. As more and more soils were identified, the need for a classification system to organize this knowledge was recognized.

### Diagnostic horizons in soil taxonomy

A key feature of soil taxonomy is the use of diagnostic horizons and features to define taxonomic classes. They are defined based on a combination of morphological characteristics that can be observed in the field, as well as other properties that must be measured in a laboratory. Soil taxonomy includes definitions for 66 different kinds of diagnostic horizons and features. Diagnostic horizons at the soil surface are called epipedons (Gr. epi, over, and pedon, soil). Other diagnostic
horizons are simply called subsurface horizons. Two examples (one epipedon and one subsurface horizon) are provided here to illustrate this concept. Many of the world’s grasslands have a thick, dark surface horizon rich in organic matter and high in native fertility. It is formed as a consequence of abundant inputs of organic matter to the soil by deep-rooted grasses with high belowground biomass, coupled with the mixing activity of fauna. This diagnostic surface horizon is a mollic epipedon (L. mollis, soft) (Figure 3). Some arid and semiarid soils have a subsoil horizon rich in calcium carbonate. Typically this horizon forms as a result of water dissolving calcium carbonate in the surface horizon and carrying it downward as it percolates into the soil. The water only moves a relatively short distance into the subsoil before being taken up by plant roots, or moving back up, toward the surface, due to evaporative demand. Left behind in the subsurface are any dissolved minerals that the water contained, such as calcium carbonate, and these accumulate over time. This diagnostic subsoil horizon is a calcic horizon (L. calais, lime) (Figure 4).
The thick, dark surface horizon of this soil from Wisconsin is a diagnostic horizon known as a mollic epipedon. Photo by Dr Randy Schaetzl, USDA Natural Resources Conservation Service.

The use of diagnostic horizons and features that are defined by measurable and observable soil properties and that infer pedogenesis is an effective strategy, first used by soil taxonomy, to integrate pedogenic theory into soil classification. An important advantage is that it does not require prior knowledge of the genetic history of the soil. Rather, one needs only to identify the kinds of diagnostic horizons and features that are present based on the properties of the soil.

Structure of the system

Soil taxonomy is a hierarchical system with six categories, or levels: order, suborder, great group, subgroup, family, and series. The highest category – the order – has 12 entries. All of the soils known in the world belong to one of these 12 soil orders. As one moves progressively lower in the system, from order to series, the number of classes within each category becomes more numerous and more narrowly defined. Also,
because members of each class share all of the properties required for the class above it, the number of characteristics used to differentiate the classes within each category increases in the lower classes.

To classify a soil, one starts at the beginning of the key, with soil orders, and works through the keys progressively. As soon as the listed criteria are met at a particular level, the soil is placed in that class and then the criteria for the next level are considered. To illustrate the hierarchical nature of the system, consider the Spodosol order. As currently documented in the United States it consists of 5 suborders, 22 great groups, 121 subgroups, 425 families, and 747 soil series.

The characteristics of the classes within each category are defined so as to reflect the major soil-forming processes, or at least those that have exerted significant control on the formation and development of the soil. Soil orders are separated based on the presence of properties resulting from major soil-forming processes, often operating over large geographic areas. Suborders are defined based on the presence of properties that exert major control over currently active soil-forming processes. These commonly consist of dynamic climatic controls such as the soil moisture or temperature regime. Great groups are defined based on the presence of properties that exert additional controls over the development of the soil in its current environment. These are often static properties such as the presence of a specific diagnostic horizon or feature. Subgroups are defined by the presence of properties that either reflect a relationship to similar soils in other categories, or are unique properties not recognized in any other class. Soil families consist of properties that are particularly important to the use and management of the soil. These include capacity factors such as soil texture or depth, as well as intensity factors such as soil temperature or inherent ability to hold and exchange nutrient cations. Finally, soil series are separated by narrowing the range of important properties so as to reflect important use and management considerations and to facilitate identification in the field while making soil maps.

Naming conventions

Soil taxonomy uses a unique mnemonic system of nomenclature for naming the classes (Heller 1963). The class names therefore are very effective at communicating the overall soil characteristics to those trained in its use. At each level (above the soil family) short formative elements, derived mostly from Greek or Latin roots, are combined in specific ways to form the names of each class.

Soil order names all begin with a syllable from the root word used to represent the order, and end with the syllable sol (L. solum, soil). The beginning and ending syllables are connected with the vowel i (Table 1). For example, the order name Aridisol is based on the Latin word aridus (dry). This soil order contains many of the world’s dryland soils. The name Histosol is based on the Greek word histos (tissue) and consists of soils formed in thick deposits of decaying organic material. The other orders are named in a similar fashion. For the names of the soil classes below the order level, the names end with a short formative element syllable derived from the Latin, Greek, or other word used for the order (e.g., id for Aridisols, ist for Histosols, etc.) as shown in the second column of Table 1.

Suborder names consist of two syllables, for example Aquolls or Fibrists. The first syllable is connotative of either a major soil property or the moisture or temperature regime of the soil. The second syllable is the soil order’s formative element. For example a soil belonging to the Calcids suborder is an Aridisol (denoted by the
Table 1  Soil order names, their formative elements, and a general description of each order.

<table>
<thead>
<tr>
<th>Soil order</th>
<th>Below order level, names end with</th>
<th>Formative element derivation</th>
<th>General description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alfisols</td>
<td>-alf</td>
<td>Meaningless syllable from early soil term “pedalfer”</td>
<td>Naturally fertile soils with high base-cation saturation and a clay-enriched subsoil horizon.</td>
</tr>
<tr>
<td>Andisols</td>
<td>-and</td>
<td>Modified from “ando.” J. An, dark, do, soil</td>
<td>Relatively young soils, mostly of volcanic origin, characterized by minerals with poorly organized crystalline structure.</td>
</tr>
<tr>
<td>Aridisols</td>
<td>-id</td>
<td>L. aridus, dry.</td>
<td>Soils of dry climatic regions.</td>
</tr>
<tr>
<td>Entisols</td>
<td>-ent</td>
<td>Meaningless symbol, sounds like “recent”</td>
<td>Young soils with little or no soil profile development.</td>
</tr>
<tr>
<td>Gelisols</td>
<td>-el</td>
<td>L. gelare, to freeze.</td>
<td>Soils of cold regions with permafrost in the subsoil.</td>
</tr>
<tr>
<td>Histosols</td>
<td>-ist</td>
<td>Gr. histos, tissue</td>
<td>Soils formed in decaying organic material.</td>
</tr>
<tr>
<td>Inceptisols</td>
<td>-ept</td>
<td>L. inceptum, beginning</td>
<td>Youthful soils with a weak, but noticeable, degree of profile development.</td>
</tr>
<tr>
<td>Mollisols</td>
<td>-oll</td>
<td>L. mollis, soft</td>
<td>Very dark-colored, naturally very fertile soils of grasslands.</td>
</tr>
<tr>
<td>Oxisols</td>
<td>-ox</td>
<td>Fr. oxide, oxide</td>
<td>Highly weathered tropical soils with low natural fertility.</td>
</tr>
<tr>
<td>Spodosols</td>
<td>-od</td>
<td>Gr. spodos, wood ash</td>
<td>Acidic soils with low fertility and accumulations of soluble organic matter and iron/aluminum oxides in the subsoil.</td>
</tr>
<tr>
<td>Ultisols</td>
<td>-ult</td>
<td>L. ultimus, last</td>
<td>Soils with low base-cation saturation and a clay-enriched subsoil horizon.</td>
</tr>
<tr>
<td>Vertisols</td>
<td>-ert</td>
<td>L. verto, turn</td>
<td>Very clayey soils that shrink and crack when dry, and expand when wet.</td>
</tr>
</tbody>
</table>

Fr., French; Gr., Greek; J., Japanese; L., Latin.

second syllable id) that has a calcic diagnostic horizon in the subsoil (denoted by the first syllable calc). Table 2 lists some of the formative elements used in suborder names. Currently soil taxonomy recognizes 68 suborders.

Great group names are constructed by adding one (or sometimes two) formative elements to the beginning of the name of the suborder, for example Haplorthods or Endoaquolls. They consist of three or four syllables. The formative elements describe the presence of (or sometimes lack of) additional important soil properties that exert further control on the soil-forming process. For example, a soil in the Petrocalcids great group belongs to the Aridisols order, and Calcids suborder. Importantly, though, it has a petrocalcic horizon (cemented calcic), so it is placed in the Petrocalcids great group. Table 3 lists some of the formative elements used in
### Abbreviated list of some formative elements for suborders and their connotations.

<table>
<thead>
<tr>
<th>Formative element</th>
<th>Derivation</th>
<th>Soil property connotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aqu</td>
<td>L. <em>aqua</em>, water</td>
<td>Saturation with water, wet soils.</td>
</tr>
<tr>
<td>Calc</td>
<td>L. <em>calcis</em>, lime</td>
<td>Has a horizon rich in calcium.</td>
</tr>
<tr>
<td>Cry</td>
<td>Gr. <em>kryos</em>, cold</td>
<td>Cold soil temperatures.</td>
</tr>
<tr>
<td>Fibr</td>
<td>L. <em>fibra</em>, fiber</td>
<td>Slightly decomposed organic matter.</td>
</tr>
<tr>
<td>Fluv</td>
<td>L. <em>fluvius</em>, river</td>
<td>Formed in alluvial (river) materials.</td>
</tr>
<tr>
<td>Gyps</td>
<td>L. <em>gypsum</em>, gypsum</td>
<td>Has a horizon rich in gypsum.</td>
</tr>
<tr>
<td>Hem</td>
<td>Gr. <em>hemi</em>, half</td>
<td>Moderately decomposed organic matter.</td>
</tr>
<tr>
<td>Hum</td>
<td>L. <em>humus</em>, earth</td>
<td>High organic matter (humus) content.</td>
</tr>
<tr>
<td>Orth</td>
<td>Gr. <em>orthos</em>, true</td>
<td>Common morphology.</td>
</tr>
<tr>
<td>Psamm</td>
<td>Gr. <em>psammos</em>, sand</td>
<td>Sandy textured.</td>
</tr>
<tr>
<td>Sal</td>
<td>L. <em>sal</em>, salt</td>
<td>Has a horizon rich in salts.</td>
</tr>
<tr>
<td>Torr</td>
<td>L. <em>torridus</em>, hot and dry</td>
<td>Hot and dry (torric) moisture regime.</td>
</tr>
<tr>
<td>Turb</td>
<td>L. <em>turbidis</em>, disturbed</td>
<td>Intense churning by frost (cryoturbation).</td>
</tr>
<tr>
<td>Ud</td>
<td>L. <em>udus</em>, humid</td>
<td>Udic moisture regime, with well-distributed rainfall.</td>
</tr>
<tr>
<td>Ust</td>
<td>L. <em>ustus</em>, burnt</td>
<td>Ustic moisture regime, with somewhat limited moisture and a dry summer season.</td>
</tr>
<tr>
<td>Vitr</td>
<td>L. <em>vitrum</em>, glass</td>
<td>Contains volcanic glass.</td>
</tr>
<tr>
<td>Xer</td>
<td>Gr. <em>xeros</em>, dry</td>
<td>Soils of Mediterranean climates: cool, moist winter and a dry summer.</td>
</tr>
</tbody>
</table>

Great group names. Currently soil taxonomy recognizes 337 great groups.

Subgroup names consist of the great group name, proceeded by an adjective term (ending with *ic*) describing some important soil feature that is either shared by another class (these subgroups are intergrades), or that is unique to the subgroup (these subgroups are extragrades). The adjective descriptor is a separate word, not joined to the great group name, for example *Lithic Dys-trudepts* or *Aeric Haplaquods*. In addition, one subgroup within each great group is named with the adjective “Typic,” for the central concept of that group, which may, or may not, be the most extensive. For example, a soil belonging to the *Aquic Petrocalcids* subgroup is a member of the *Petrocalcids* great group and has a water table within 100 cm. Table 4 lists some of the formative elements used in subgroup names. Currently soil taxonomy recognizes 2264 subgroups.

Soil family names consist of the subgroup preceded by multiple terms indicating properties that are important to the use and management of the soils. Family names usually include three to five terms, all used to describe conditions such as particle-size characteristics, mineralogy, soil temperature, cation exchange activity, calcareous and reaction classes, rooting depth, rupture resistance, or classes of coatings and cracks. The rules for assigning these terms are described in soil taxonomy. Currently in the United States, 10,895 taxonomic families have
been recognized. Many additional soil families have been recognized in other countries as well.

Soil series make up the lowest categorical level of soil taxonomy. Their definitions include characteristics such as kind and arrangement of soil horizons, physical properties such as color, texture, and consistence, and chemical properties such as pH, salinity, and so on. Many soil series were recognized prior to the development of soil taxonomy, and they were used to help define the classes when soil taxonomy was developed. Series names do not include the mnemonic terms described above. Rather they are assigned short names derived mostly from local places or physiographic features near where the soil occurs, for example Miami, Clinton, or Kalkaska. In this way, soil series names are easy to pronounce and remember locally. They do not, however, contain the information supplied within the terms used in the higher categories, and are therefore only effective for communicating with people who are familiar with the specific soil series. Currently in the United States, 22,253 soil series have been recognized, and several new ones are defined each year, as new information surfaces during mapping operations. Many additional soil series have been defined in other countries. Descriptions of all current US soil series can be accessed online (Soil Survey Staff 2015b).

Two examples of soil names at the various class levels are shown in Table 5. The Tonuco soil series is found in the deserts of southern New Mexico. It is a member of the Aridsoil order. The presence of a calcic horizon places it in the Calcixerdermic suborder. Cementation of the subsoil places

### Table 3
Abbreviated list of some formative elements for great groups and their connotations.

<table>
<thead>
<tr>
<th>Formative element</th>
<th>Derivation</th>
<th>Soil property connotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acr</td>
<td>Gr. <em>akros</em>, at the end</td>
<td>Extremely weathered soil.</td>
</tr>
<tr>
<td>Dystr (Dys)</td>
<td>Gr. <em>dys</em>, ill, infertile</td>
<td>Low base-cation saturation.</td>
</tr>
<tr>
<td>Endo</td>
<td>Gr. <em>endo</em>, within</td>
<td>Groundwater saturates the whole soil.</td>
</tr>
<tr>
<td>Epi</td>
<td>Gr. <em>epi</em>, on, above</td>
<td>Groundwater perched on an impervious layer, but unsaturated below.</td>
</tr>
<tr>
<td>Eutr</td>
<td>Gr. <em>eu</em>, good, fertile</td>
<td>High base-cation saturation.</td>
</tr>
<tr>
<td>Frag</td>
<td>L. <em>fragilis</em>, brittle</td>
<td>Has a fragipan.</td>
</tr>
<tr>
<td>Gloss</td>
<td>Gr. <em>glossa</em>, tongue</td>
<td>Has a glossic horizon.</td>
</tr>
<tr>
<td>Hapl</td>
<td>Gr. <em>haplos</em>, simple</td>
<td>Minimal horizon development.</td>
</tr>
<tr>
<td>Kand</td>
<td>Modified from <em>kandite</em></td>
<td>Abundant low-activity clay minerals.</td>
</tr>
<tr>
<td>Moll</td>
<td>L. <em>mollis</em>, soft</td>
<td>Has a mollic (humus-rich, high pH) surface horizon.</td>
</tr>
<tr>
<td>Natr</td>
<td>L. <em>natrium</em>, sodium</td>
<td>Has a horizon rich in sodium.</td>
</tr>
<tr>
<td>Petro</td>
<td>Gr. <em>petra</em>, rock</td>
<td>Has a cemented horizon.</td>
</tr>
<tr>
<td>Plinth</td>
<td>Gr. <em>plinthos</em>, brick</td>
<td>Presence of plinthite.</td>
</tr>
<tr>
<td>Quartzi</td>
<td>Ger. <em>Quarz</em>, quartz</td>
<td>High quartz content.</td>
</tr>
<tr>
<td>Sapr</td>
<td>Gr. <em>sapros</em>, rotten</td>
<td>Highly decomposed organic matter.</td>
</tr>
<tr>
<td>Verm</td>
<td>L. <em>vermes</em>, worm</td>
<td>Worm-riddled, or mixed by soil fauna.</td>
</tr>
</tbody>
</table>
SOIL TAXONOMY AND SOIL CLASSIFICATION

Table 4  Abbreviated list of some formative elements for subgroups and their connotations.

<table>
<thead>
<tr>
<th>Formative element</th>
<th>Derivation</th>
<th>Soil property connotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abruptic</td>
<td>L. <em>abruptum</em>, torn off</td>
<td>Has an abrupt change in texture.</td>
</tr>
<tr>
<td>Aeric</td>
<td>Gr. <em>aeris</em>, air</td>
<td>Aeration is implied, the soil is drier than expected.</td>
</tr>
<tr>
<td>Arenic</td>
<td>L. <em>arena</em>, sand</td>
<td>Moderately thick, sandy, surface horizon.</td>
</tr>
<tr>
<td>Cumulic</td>
<td>L. <em>cumulus</em>, heap</td>
<td>Over-thickened surface horizon.</td>
</tr>
<tr>
<td>Grossarenic</td>
<td>L. <em>grossic</em>, thick; and arenic, sand</td>
<td>Thick, sandy, surface horizon.</td>
</tr>
<tr>
<td>Leptic</td>
<td>Gr. <em>leptos</em>, thin</td>
<td>A thin soil.</td>
</tr>
<tr>
<td>Limnic</td>
<td>Gr. <em>limne</em>, lake</td>
<td>Has a limnic layer of lake sediment.</td>
</tr>
<tr>
<td>Lithic</td>
<td>Gr. <em>lithos</em>, stone</td>
<td>Has a shallow contact with bedrock.</td>
</tr>
<tr>
<td>Oxyaquic</td>
<td>Modified from <em>oxygen</em>, and L. <em>aqua</em>, water</td>
<td>Wetness, but with oxygenated water.</td>
</tr>
<tr>
<td>Petronodic</td>
<td>Gr. <em>petra</em>, rock; and L. <em>nodulus</em>, little knot</td>
<td>Contains concretions or nodules (petronodes).</td>
</tr>
<tr>
<td>Sulf</td>
<td>L. <em>sulfur</em>, sulfur</td>
<td>Has a sulfuric horizon or high content of oxidizable sulfur.</td>
</tr>
<tr>
<td>Terric</td>
<td>L. <em>terra</em>, earth</td>
<td>Organic soils with underlying mineral soil material.</td>
</tr>
<tr>
<td>Typic</td>
<td>Modified from <em>typical</em></td>
<td>Either the central concept or simply the last subgroup listed in the classification key.</td>
</tr>
</tbody>
</table>

Table 5  Examples of taxonomic names used in Soil Taxonomy.

<table>
<thead>
<tr>
<th>Order</th>
<th>Suborder</th>
<th>Great group</th>
<th>Subgroup</th>
<th>Family</th>
<th>Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aridisol</td>
<td>Calcid</td>
<td>Petrocalcids</td>
<td>Typic Petrocalcids</td>
<td>Sandy, mixed, thermic, shallow</td>
<td>Tonuco</td>
</tr>
<tr>
<td>Histosol</td>
<td>Saprist</td>
<td>Sulphisaprist</td>
<td>Terric Sulphisaprist</td>
<td>Loamy, mixed, euic, thermic</td>
<td>Rappahannock</td>
</tr>
</tbody>
</table>

It is in the Petrocalcid great group. It is a member of the Typic Petrocalcids subgroup because it has no features found in any other subgroups of Petrocalcids. The family terms indicate the soil has sandy textures, mixed lithologies, warm soil temperatures, and a shallow rooting depth. The Rappahannock series is a soil formed in moderately thick organic deposits within tidal marshes of coastal Virginia. The thick organic soil materials place it in the Histosol order, and the high degree of decomposition of these materials places it in the Saprist suborder. Its high content of oxidizable sulfur compounds places it in the Sulphisaprist great group. The presence of mineral materials below the organic material makes it a Terric Sulphisaprist. The family terms indicate...
that the underlying mineral material is loamy and has mixed lithologies. The soil also has relatively high pH (euic) and warm soil temperature.

**General descriptions of the soil orders**

The hierarchical nature of soil taxonomy lends itself well to the production of maps at various scales. Soil orders, as the highest and most generalized level, are effectively shown on small-scale maps, such as those at global or continental scales (Figure 5). Progressively lower categories (suborders, great groups, etc.) can be depicted on progressively larger-scale maps, such as those of nations, regions, watersheds, and eventually farms, fields, and research plots. A brief description of the geographic distribution and environmental setting of each of the 12 soil orders follows, in alphabetical order.

**Alfisols** are naturally fertile soils with high base-cation saturation levels and a clay-enriched subsoil horizon. They form in a wide range of parent materials and occur under broad environmental conditions ranging from tropical to boreal, although most form on parent materials rich in calcium-bearing minerals. Alfisols develop primarily (though not exclusively) under forest vegetation. Translocation of clay and other weathering products from the upper horizons, and their subsequent accumulation in the subsoil, are important processes in Alfisols. Soil-forming processes are in relative balance so that base-cations such as calcium, magnesium, and potassium are supplied to the soil through weathering; the processes of leaching are not sufficiently intense to remove them from the soil before plants can utilize and recycle them. Thus, the soils remain fertile. Alfisols are extensive worldwide, occurring on every continent except Antarctica. Globally, they occupy about 10% of the ice-free land area.

**Andisols** are relatively young soils that are mostly of volcanic origin. A few Andisols have formed in nonvolcanic parent materials, particularly in cool, moist environments with relatively high levels of soil organic matter. The dominant pedogenic processes in Andisols are weathering and mineral transformation (with little or no translocation) of the volcanic alumino-silicate minerals (often glassy) to form short-range-order (poorly crystalline) minerals, such as allophane, imogolite, and ferrihydrite. Most Andisols are highly productive soils, in part because their unique mineral constituents provide for high water-holding capacities. Periodic volcanic eruptions provide fresh tephra that resupplies nutrients and maintains a favorable fertility status. One negative aspect of their unique mineral makeup is a propensity to adsorb phosphorus strongly, making it unavailable to plants. Andisols occur in a wide range of environments but occupy only about 1% of the ice-free land area globally, mostly concentrated along the Pacific Ring of Fire, an area of frequent volcanic activity.

**Aridisols** occur throughout the world in dry and desert environments. Intense or prolonged climatic dryness is their dominant feature. Evaporation exceeds precipitation in these environments, thus limiting the amount of moisture available to support plants. Because little or no water moves completely through the soil profile, salts such as calcium carbonate, sodium, gypsum, or sodium chloride often accumulate in some part of the soil profile. Globally, Aridisols occupy about 12% of the ice-free land area, in both hot and cold deserts.

**Entisols** have little or no soil profile development, and are often found on geologically young landscapes. They occur in a wide range of environments in settings such as steep, actively eroding slopes, floodplains that receive frequent deposits of alluvium, and shifting sand dunes. Consequently, there has been little opportunity
Soil orders

- Alfisols
- Andisols
- Aridisols
- Entisols
- Gelisols
- Histosols
- Inceptisols
- Mollisols
- Oxisols
- Spodosols
- Ultisols
- Vertisols

Rocky land
Shifting sand
Ice/glacier

Figure 5 Global distribution of the soil orders. Map courtesy of USDA Natural Resources Conservation Service.
SOIL TAXONOMY AND SOIL CLASSIFICATION

for soil-forming processes to develop many soil horizons. Many Entisols are quite productive agriculturally, in part because they are minimally weathered and thus retain large amounts of primary minerals. Globally, Entisols occupy about 16% of the ice-free land area.

_Gelisols_ are soils of cold landscapes, with permafrost in the subsoil. Extreme cold is their dominant feature, with cryoturbation an important physical process in many of these soils. Unique “patterned ground” land surface features are characteristic of some areas. These include sorted and unsorted rock circles and stripes, large polygonal surface patterns, ice wedges, frost boils, pingos, and more. Gelisols occur in both the northern and southern circumpolar high latitude regions and are found in tundra, taiga, and boreal forest environments. They are important globally because they store large amounts of carbon as peat, often cryoturbated deep into the subsoil. Globally, Gelisols occupy about 9% of the ice-free land area.

_Histosols_ form in thick masses of decaying organic material, and usually on very wet sites. Their organic layers vary in color, botanical origin, amount of mineral soil material mixed in, degree of decomposition, pH, and other properties. Histosols are found in the tropics, temperate areas, and boreal forests. They form in environments where rates of inputs of organic matter exceed losses due to decomposition. Generally these are cool, wet environments where precipitation exceeds evaporation, or areas with persistently high water tables. Therefore, Histosols are common in low-lying areas with waterlogged, anaerobic conditions. Globally, Histosols occupy about 1% of the ice-free land area.

_Inceptisols_ are youthful soils with a limited, but noticeable, degree of profile development. As a group, they have profiles exhibiting more development than the Entisols, but are less developed than any of the other orders. Inceptisols tend to be found on relatively young geomorphic surfaces that are sufficiently stable (and old) to have resulted in weak but noticeable profile development. Common settings include upland slopes, stream terraces, and recently glaciated landscapes. Inceptisols are found in wide-ranging environments from the tropics to the tundra and from semiarid to humid climates. Globally, Inceptisols occupy about 17% of the ice-free land area.

_Mollisols_ are dark-colored, naturally very fertile soils, which are found primarily in grasslands and savannas. The thick, dark surface horizon is rich in organic matter, which accumulates as the result of inputs from decaying roots, especially short, mid, and tall grasses common to prairie and steppe areas. There is a high belowground biomass production and tight biocycling of nutrients. Their naturally high native fertility is due to nutrients distributed throughout their profile, and to their high organic matter contents and excellent soil structure. Globally, Mollisols are predominantly in temperate grasslands at mid-latitudes. They occupy about 7% of the ice-free land area.

_Oxisols_ are highly weathered tropical soils with low natural fertility. Most Oxisols are on old landscapes that have undergone multiple cycles of weathering, erosion, redeposition, and renewed weathering, leaving only highly resistant minerals that retain very little native fertility. The few reserves of soil nutrients are concentrated mostly in the living and dead plant tissues at the site and these are quickly lost upon clearing and cultivation. Most Oxisols are naturally forested, but many are now being cleared for agriculture, which is possible with consistent inputs of chemical fertilizers. Globally, Oxisols occupy about 8% of the ice-free land area, mostly at latitudes between the Tropics of Cancer and Capricorn.

_Spodosols_ are coarse-textured, acidic soils with low fertility and accumulations of soluble organic
SOIL TAXONOMY AND SOIL CLASSIFICATION

matter and iron/aluminum oxides in the subsoil horizon. They are found mostly in forested areas with cool, moist climates in the humid boreal zones, but also occur on wet sites in tropical or subtropical areas. The genesis of these soils is driven by the production of organic acids from decaying leaf litter, which form organometallic complexes with iron and aluminum compounds released by weathering. These complexes are translocated downward in the profile and accumulate in the subsoil. Most Spodosols are minimally productive as agricultural soils, but many sustain thriving forests. Globally, Spodosols occupy about 3% of the ice-free land area.

_Ultisols_ are acidic soils with low base-cation saturation values and a clay-enriched subsoil horizon. As such, they have only low to moderate native fertility. They are found mostly in warm humid climates under forest vegetation where long-term leaching of nutrients from the profile has exceeded the rate of nutrient release from mineral weathering. Plant nutrients are typically concentrated in the upper part of the profile (due to biocycling) and decrease with depth. With adequate fertilization, Ultisols can be made to be quite productive. Globally, Ultisols occupy about 8% of the ice-free land area.

_Vertisols_ are dark-to-black, very clayey soils that develop deep cracks in a dry season. They are dominated by clay minerals (smectites) that swell when moistened and dramatically shrink when dry, forming the deep cracks. They are very sticky and plastic when wet and very firm and hard when dry. Most Vertisols are in grasslands or savannas, but a few are in forest, or even desert shrub. Commonly, Vertisols form from parent materials that have neutral pH or are calcareous, such as sandstone, shale, chalk, or limestone. Some Vertisols have developed a unique pattern of microrelief known as gilgai – an intricate pattern of micro-highs and lows repeating over scales of around 3–10 m. Gilgai forms because of the mixing that occurs in these soils as they swell when wet. Many Vertisols sustain highly productive agricultural systems. Globally, Vertisols occupy about 2% of the ice-free land area.

SEE ALSO: Digital soil mapping and pedometrics; Soil mapping and maps; Soils of cold and permafrost-affected landscapes; Soils of desert landscapes; Soils of humid mid-latitude landscapes; Soils of tropical landscapes; Soils of the United States

References


Further reading


Soil water

Tim Beach
Jonathan M. Flood
University of Texas at Austin, USA

This entry reviews the types and roles of soil water, its movement, its environmental contexts, and what is known about its trends and geography.

Soil water in the hydrologic cycle

More water exists in soils than in rivers, the atmosphere, life forms, and swamps and marshes put together. If we include permafrost, then soil water is the third-largest reservoir of the Earth’s fresh water after glaciers and lakes. The quantity of soil water, like all types of water, varies tremendously around the planet due to all the factors of the hydrologic cycle. Much of this soil water is the main source of water for ecosystems above and below ground and for crops and pasture lands (Shiklomanov 1993).

Soil water exists at the nexus of the critical zone within the hydrosphere, atmosphere, biosphere, and lithosphere, and is at the leading edge of the major hydrologic processes of precipitation, infiltration, storage, evaporation, and transpiration. Its presence or absence influences planetary albedo, humidity, and temperature, which in turn impact global energy and water circulation.

Soil water types

We usually define soil water by the force that holds it to the surface. Soil water has four general types: structural water, gravitational water, capillary water, and hygroscopic water. Structural water exists within the bonds of minerals such as gypsum (CaSO₄·2H₂O) with its two water molecules, as opposed to its waterless cousin, anhydrite (CaSO₄). Soil water potential, or soil moisture tension, is the force, measured in bars, that holds water to a surface (Figure 1).

In nonstructural soil water, water sticks to surfaces by adhesion and cohesion. Adhesion is the strong electrical attraction of water molecules to a surface, such as a straw or clays or humus in a soil pore, in this case a micropore, <50 nm. This hygroscopic water, the thin layer of water molecules, is held too tightly to a surface (at ca. −15 to −31 bars or atmospheres) for plants to use, but labs routinely remove it by heating. Cohesion refers to the water molecules attached to other water molecules by positive charges of hydrogen atoms attracted to the negative charges of oxygen atoms. Gravitational water is not held with enough force (−1/3 bar) to overcome gravity and thus flows out of a soil after a few days. This water exists in so-called macropores and, if it does not drain, may cause plants to wilt by blocking oxygen flow.

Capillary is plant-available water held from −15 to −1/3 bars by cohesion that resists the pull of gravity but not the pull from plant roots (Figure 1). It is thus the source for...
SOIL WATER

Figure 1 Soil water types.

most ecosystems and crops, and many aspects of environments govern how much capillary water is available for plants: rainfall and evapotranspiration, water table depth, soil textures, and the network of mycorrhizae and root systems. Soils tend to be moist where precipitation exceeds evapotranspiration, but this varies by soil texture and water table depth and height of the capillary fringe above the water table.

Four states of water in soils are also referred to by scientists: saturation, field capacity, wilting point, and hygroscopic (discussed above) (Figure 1). Saturation refers to the point when water fills soil macropores and moves out by gravity unless held in place by a high water table, for example in wetlands. Field capacity refers to the maximum plant extractable soil water held by cohesion. Wilting point occurs when soil water is held by adhesion as hygroscopic water and plant roots cannot overcome the tension with which water molecules attach to each other. Thus, as with too much water, too little water causes plants to wilt.

Soil texture is an important determinant of water holding capacity: clay holds more water than sands and silts, but more of this water is too tightly held by adhesion to the clay surface. Likewise, sand texture holds soil water too loosely and much of this water is in macropores where water drains by gravity. Loam soils, especially silt loams, have the highest level of plant available water, one of the reasons loamy textured soils are more productive. For example, the water holding capacity of sand is about $0.075 \text{ cm cm}^{-1}$, clay is about $0.12 \text{ cm cm}^{-1}$, and silt loam is about $0.2 \text{ cm cm}^{-1}$. Water holding
capacity increases for each texture with more organic matter. For example, silt loams increase to 0.24 cm cm$^{-1}$ with more than 3% organic matter, and Histosol soils (dominated by organic matter) may have a water holding capacity of up to 0.65 cm cm$^{-1}$ (USDA 2016a).

Soil structure also affects soil water flow, and soil horizons with granular structure (the most macropores) allow the fastest flow. The rate of flow generally decreases from granular to blocky, to columnar, to massive structures. This means in many soils that gravity water in the soil will flow through granular or blocky layer but perch above massive and clay layers such as Bt horizons. Thus rainfall may saturate the soil pores and flow may occur as throughflow in soil pores under gravity and as Dunne flow or Hortonian flow on the soil surface.

Soil moisture regimes

The United States Department of Agriculture (USDA) has classified the world into soil moisture regimes – a useful shorthand for understanding average soil water conditions. These include aquic, udic, ustic, xeric, and aridic divisions from wettest to driest, which along with soil temperature regimes provide a general framework and a base for understanding global-scale environmental change problems. For example, regions with ustic and aridic regimes have ephemeral water and high evapotranspiration rates, which allow us to map and upscale problems such as water stress and salinization (USDA 2016b).

Movement

Soil water can move in any direction depending on gravity, capillarity, roots, porosity, permeability, texture, and structure. The main pathways into soils are infiltration through raindrops, fog drip, upward flow from capillary fringe or rising groundwater tables, and runoff. Stems of plants are important to infiltration in stem flow or directing water to roots, and thus soil water may have a patchy distribution around rhizospheres. Percolation is the downward transit of water in soils, and a key parameter of soil usage for engineering land uses. A common classification of soils is the soil permeability class, ranging from very rapid to very slow and from more than 25 cm hr$^{-1}$ to less than 0.13 cm hr$^{-1}$. Scientists estimate water movement through soil with some derivation of the Richards’ equation (Pachepsky, Timlin, and Rawls 2003), a modified version of Darcy’s law (groundwater). The key variables in this equation are pressure head (water engineering) and the unsaturated hydraulic conductivity.

Hydraulic conductivity measures the ability of soil or other geologic mediums to transmit water along a hydraulic gradient. Conductivity is primarily governed by the intrinsic permeability of a material – symbolized as $K$ in equations – and to a lesser extent by the density, viscosity, and distribution of the passing fluid. The USDA and other state and federal agencies document and make available hydraulic conductivity classes for soils and geologic strata. These run from highly conductive coarse sands to nearly impermeable fine clays and fragipans. Conductivity classifications help determine best land-use and irrigation strategies (USGS 2016c).

Another salient soil water classification is the natural drainage class, or simply drainage class. Drainage class refers to the moisture condition of the soil in its natural state over the course of a year. The classification scheme categorizes the degree, frequency, and duration of soil wetness over time. Seven classes exist, determined by soil color, water table characteristics, soil texture and morphology, and the soil’s position in the landscape. The seven classes are: excessively
drained, somewhat excessively drained, well drained, moderately well drained, somewhat poorly drained, poorly drained, and very poorly drained (USGS 2016c).

Environmental contexts

Geography

There are many dimensions to the global distribution of soil water; one useful dimension is the distribution of plant available soil water per year. The factors of plant available soil water are inputs, outputs, and storage (Dunne and Wilmott 1996). Despite multiple variables, the main variable that stands out on a global map is precipitation minus evapotranspiration. The highest soil moisture occurs in the tropical rainforest zones due to high input and output, the lowest occurs in subtropical deserts due to low input and high output, and the second-highest occurs in mid-latitude forests and taiga due to low input and very low output as well as storage in Histosol soils (Figure 2).

In 2014, The Jet Propulsion Laboratory (JPL) deployed a new satellite, called SMAP (Soil Moisture Active Passive), with soil moisture sensors for the top 5 cm of soil (http://smap.jpl.nasa.gov). This system will improve our knowledge of the planet’s upper soil moisture at a finer resolution and provide critical information concerning the scale and intensity of emerging flood or drought events.
**Soil color**

Soil moisture leaves visible clues in terms of either dark colors associated with high soil organic matter and thus arrested decomposition from water saturation and redoximorphic features, or mottles or color variations in soil associated with wetness. Iron oxides in well-drained soils make them yellow or reddish. Wet soils have reduced colors (gray, blue, green) of whole horizons or mottling with distinct or prominent contrasts in reduced and oxidized colors. The in situ reduced color soils will begin to oxidize with exposure and water removal in half an hour or less. Soils that go through wetting and drying cycles will have redox concentrations and depletions, meaning areas redder with \( \text{Fe}^{3+} \) and grayer with \( \text{Fe}^{2+} \). The reduced zones may develop iron masses, concentrations, or iron and manganese nodules, depending on time and specific environments.

**Soil catena**

Soil water is one of the key characteristics that vary along soil catenas. Soil water characteristics on catenas depend on texture, structure, infiltration, local water tables, and the steepness of the catena. Generally, however, well-drained soils occur on the steep shoulder and backslopes followed by crests, and less well-drained soils occur in depressions and foot and toe slopes associated with water flow convergence or high water tables.

**Soil water and society**

Soil water is a critical element in agricultural productivity at every scale. The riverine cradles of civilization along the banks of the Tigris, Euphrates, and Nile had annual pulses of nutrients and restorative soil water for the vast rainless stretches of the fertile river valleys. This may have also led to the demise of Mesopotamia because salinization accompanied irrigation there. Agriculture still uses 70% of global freshwater consumption, mainly to increase soil water through irrigation, and this comes at the cost of salinization in dry regions (UNESCO 2012).

Altering soil water content by way of irrigation has been important to agriculture for millennia but tended to rely on the redirection of stream energy or animal power to distribute water into soils. In the past half-century, fossil fuel-powered pumps have been used to remove groundwater from tube wells and have brought prosperity to marginal agricultural regions. Groundwater pumping tends to positively impact societies in the short term, but negative impacts on regional hydrology often create serious water shortage or water quality problems over the long term (Konikow and Kendy 2005). Over-pumping to increase soil water for higher agricultural yields depletes aquifers and often impacts regional hydrology that supports soil moisture away from the places where water is applied. More efficient irrigation applicators such as drip tube systems, traditional agricultural technologies such as braided terracing, and regionally appropriate cropping and landscaping decisions may reduce some deleterious effects of human–soil water interactions (Postel et al. 2001).

**Soil water trends**

Lastly, soil degradation in all of its forms affects soil water. This is true in terms of soil pollution affecting water quality directly from contaminants attached to eroded soil particles, or indirectly from salt water intrusion or inundation in low spots from irrigation runoff. The most obvious impact on soil water is the steady loss of soil through erosion of A horizons, which often have better soil moisture holding capacity, and exposure of B horizons, which often have lower
SOIL WATER

water holding capacity and infiltration. Soil erosion continues to be high around the world (Montgomery 2007), and will likely increase with climate change (Nearing, Pruski, and O'neal 2004). The combination of continued high accelerated soil erosion and global warming makes severe decreases in soil moisture more probable in many places (IPCC 2016, 184–185). Indeed, the whole range of global warming scenarios produces decreased soil water, because of evaporation in winter and spring from both increased temperatures and reduced snow cover (IPCC 2016, 166; UNDESA 214).

SEE ALSO: Agriculture; Aquifers; Ecosystem services; Environmental degradation; Geomorphic systems; Soil erosion and conservation; Water budget; Wetlands hydrology

References


Further reading


Soils represent the largest terrestrial stock of carbon in the global carbon cycle, with more than 2300 Pg (Petagrams or $10^{15}$ g = billion metric tons) of carbon stored in soil organic matter (SOM) and 950 Pg held as carbonates in the upper 3 m. This entry focuses only on SOM because it is the biologically active fraction of soil carbon that exchanges with the atmosphere through natural and anthropogenic processes. The amount of SOC found in soils globally is considerably larger than the carbon stored in the living biomass or atmosphere (Table 1). Globally, inputs of carbon (C) to soils are roughly balanced by losses via decomposition, but because the atmospheric C exchange with soil is so large ($\approx 60$ Pg year$^{-1}$) the potential for changes in SOC and its influence on atmospheric carbon dioxide ($CO_2$) are large.

Historically, soils have contributed to planetary cooling by sequestering atmospheric C into SOM. Warming of the atmosphere in recent decades, however, may reverse this trend and contribute to further atmospheric warming: as soils warm, decomposition of SOM is enhanced if moisture does not become limiting, releasing more CO$_2$ to the atmosphere (IPCC 2013). Moreover, changes in land use have led to large releases of CO$_2$ from soils. Although plant growth and C uptake can also occur in response to warming, even small perturbations to the soil can have major ramifications for atmospheric CO$_2$ levels.

Why carbon cycling in soils is studied

Soils represent the largest terrestrial stock of carbon in the global carbon cycle, with more than 2300 Pg (Petagrams or $10^{15}$ g = billion metric tons) of carbon stored in soil organic matter (SOM) and 950 Pg held as carbonates in the upper 3 m. This entry focuses only on SOM because it is the biologically active fraction of soil carbon that exchanges with the atmosphere through natural and anthropogenic processes. The amount of SOC found in soils globally is considerably larger than the carbon stored in the living biomass or atmosphere (Table 1). Globally, inputs of carbon (C) to soils are roughly balanced by losses via decomposition, but because the atmospheric C exchange with soil is so large ($\approx 60$ Pg year$^{-1}$) the potential for changes in SOC and its influence on atmospheric carbon dioxide ($CO_2$) are large.

Historically, soils have contributed to planetary cooling by sequestering atmospheric C into SOM. Warming of the atmosphere in recent decades, however, may reverse this trend and contribute to further atmospheric warming: as soils warm, decomposition of SOM is enhanced if moisture does not become limiting, releasing more CO$_2$ to the atmosphere (IPCC 2013). Moreover, changes in land use have led to large releases of CO$_2$ from soils. Although plant growth and C uptake can also occur in response to warming, even small perturbations to the soil can have major ramifications for atmospheric CO$_2$ levels.
SOILS AND THE CARBON CYCLE

Table 1  Approximate magnitude of organic C pools and exchange rates of soils.

<table>
<thead>
<tr>
<th></th>
<th>Soil (arctic areas, other areas)</th>
<th>Land plants</th>
<th>Atmosphere</th>
<th>Surface ocean</th>
</tr>
</thead>
<tbody>
<tr>
<td>C pools (10^{15} g)</td>
<td>&gt;2300</td>
<td>550</td>
<td>800</td>
<td>900</td>
</tr>
<tr>
<td>C exchange with</td>
<td>+61, −60</td>
<td>+61, −60</td>
<td>+9</td>
<td>+79, −80</td>
</tr>
<tr>
<td>atmosphere (10^{15} g year^{-1})</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Positive exchange refers to net release to the atmosphere per year; negative refers to C uptake onto land or ocean; imbalance in exchange indicates a sink (to land/ocean) or source (to atmosphere) on an annual basis. Residual terrestrial sink of 2.5 Pg year^{-1} is rounded to 2 and partitioned equally between soil and plants. Atmospheric release represents both fossil fuel emissions and effects of land-use change. Intergovernmental Panel for Climate Change, assessment AR5, and soil pools of Jobbágy and Jackson (2000) which do not include permafrost C pools (see Figure 2).

Table 2  Examples of disturbances and the attributes that impact carbon exchange in soils.

<table>
<thead>
<tr>
<th>Insect outbreak</th>
<th>Harvest</th>
<th>Wildfire</th>
<th>Modern agriculture</th>
<th>Erosion</th>
<th>Deposition</th>
<th>Urbanization</th>
<th>Climate change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>Aboveground</td>
<td>Above, shallow</td>
<td>Above to deep</td>
<td>Above to deep</td>
<td>Above to deep</td>
<td>Above to deep</td>
<td>Above to deep</td>
</tr>
<tr>
<td>Duration of event, year</td>
<td>Decadal</td>
<td>Annual</td>
<td>Event</td>
<td>Event to seasonal</td>
<td>Event to seasonal</td>
<td>&lt;Decadal</td>
<td>&gt;Decadal</td>
</tr>
<tr>
<td>Return interval</td>
<td>Decade to century</td>
<td>Decadal</td>
<td>&lt;Decadal</td>
<td>&lt;Decadal</td>
<td>Annual</td>
<td>Variable</td>
<td>Variable</td>
</tr>
<tr>
<td>Areal extent per event</td>
<td>Local to regional</td>
<td>Local</td>
<td>Local</td>
<td>Local</td>
<td>Local</td>
<td>Local to regional</td>
<td>Global</td>
</tr>
</tbody>
</table>

Recently, much attention has been paid to improved land-use and land management strategies that may promote C sequestration and offset an estimated 5–15% of annual anthropogenic C emissions. Improving SOC levels in soils has been suggested to be a win–win policy, because in addition to the likely climate mitigation benefits of carbon sequestration, SOC enhances soil fertility, helps reduce soil erosion by improving structure and enhancing porosity, and promotes diverse soil food webs that through decomposition and retention facilitate better nutrient cycling and plant productivity.

SOM also plays a major role in the formation of stable aggregates that improve soil structure and enhance water-holding capacity, gas exchange, and root growth. Through its reactive chemical structure, SOM also acts as a major buffer for toxins and heavy metals.

Assessments and accounting of SOC

A full accounting of SOC embodies interactions with climatic and anthropogenic changes on immediate timescales, but these interactions are
tempered by longer-term factors involving ecosystems and landscape properties such as topography and geology. An assessment of SOC stocks (amounts) and fluxes (in both gaseous and dissolved phases) can be gained from at least three main types of scientific approaches: (i) field and lab-based studies of variations in SOC over gradients in time, topography, mineralogy, vegetation, and climate—such gradients allow mathematical equations to be derived from SOC-landscape associations; (ii) inventory-based studies such as soil taxonomic mapping or soil-vegetation mapping—these maps allow spatial distributions of SOC to be linked to assessments of stocks and fluxes; and (iii) model-based assessments that are constructed from fundamental principles of land–atmosphere C exchange processes and are grounded in lab and field data. All three approaches are key to the accurate assessment of C fates under climate and anthropogenic change.

How carbon cycles through soils

SOC pools exist in a dynamic balance between inputs of C from vegetation and losses that are dominated by decomposition. Plants fix CO₂ through photosynthesis, and the majority of C inputs to the soil occur when fresh plant components senesce or die and this C is transferred to soil through litterfall and roots. Some C-rich photosynthate is immediately transferred to the soil via root exudates and mycorrhizal associations. Soil fauna and micro-organisms then begin to process and decompose the fresh organic material, thus C contained in these plant materials will either be incorporated into microbial biomass or be respired as CO₂ (or as methane (CH₄), depending on the reduction/oxidation status of the soil). Newly formed microbial biomass in turn represents a fresh pool of material for further decomposition as the microbial population turns over. The rate at which SOM decomposition proceeds will be strongly dependent upon the climatic and physiochemical conditions of the soil. Most photosynthate-C exchanges rapidly with the atmosphere upon decomposition, resulting in large fluxes of CO₂ relative to the small amounts transferred to more stable SOC pools. Although decomposition is the primary way that SOC is lost from soils, other processes such as combustion during fire, erosion, and deep leaching can represent quantitatively important fluxes of C. Research seeks to define the nature of these stable pools and the numerous processes that stabilize C by retarding the rate of decomposition (Figure 1) relative to the rate of input of new C.

Inputs of C to soil—plant productivity sets an absolute upper limit to potential C inputs to the soil system. For this reason, in many areas, SOC stocks are correlated with primary productivity, although only part of the total carbon that gets captured by plants becomes biomass; on average, about half of all assimilated carbon is respired back to the atmosphere as CO₂. In soils, the proportion of net primary productivity (NPP) allocated to belowground (roots) versus aboveground plant biomass varies widely, from <1 up to >80% across global biomes. Belowground, the depth distribution of roots declines precipitously in all biomes, with the uppermost 20 cm of soils receiving about 60–80% of all root inputs of C. Belowground, C can enter the soil directly through roots, as they die and decompose, through their exudates, or indirectly through C flow to mycorrhizal fungi and the subsequent turnover of the fungal hyphae. The tight coupling between root distribution and SOC distribution with depth in many soils is often cited as evidence for the importance of root inputs in maintaining SOC stocks of mineral soils.
SOILS AND THE CARBON CYCLE

Figure 1  The regulation of C cycling pathways through soils. Inputs that occur via annual net primary production (≈40–70 Pg C year\(^{-1}\) globally) can be lost before entering the soil. After C enters the soil via leaf and root litter contributions, it can be lost to heterotrophic respiration (Rh) (≈40–70 Pg C year\(^{-1}\)) or can be stabilized by energetic, physical, or chemical processes (≈2–10 Pg C year\(^{-1}\)) through the regulation of decomposition.

Aboveground plant material takes a more indirect route in its transfer to SOC. Before senescence or litterfall, biomass (and its C) can be lost to photooxidation, in situ herbivory, combustion, or crop removal, depending on environmental conditions such as shading, allelopathy, moisture, or harvest patterns, respectively (Figure 1). In agricultural systems, for example, harvest can remove significant portions of aboveground material, leaving roots as the main source of SOC. At the other extreme, in cold, excessively wet climates where decomposition processes are retarded, aboveground litterfall will often accumulate on the soil surface as organic (O) horizons, with accumulation rates fast enough to be measurable over annual to decadal timescales. Conversely, vertical mixing by biological or physical processes can bring SOC to the surface, or these processes can physically transfer new SOC into deeper soil horizons, as
exemplified by freeze–thaw processes in Gelisols, and by the action of earthworms or other soil fauna. Leaching and downward transport of dissolved organic matter and subsequent sorption onto mineral surfaces is also a major mechanism of vertical transport of SOC in soils.

*Decomposition of C in soil* is the main pathway for C loss from soils. And yet it is also one of the critical ecological processes for maintaining life on this planet. In natural terrestrial ecosystems, nutrient recycling via decomposition of SOM accounts for the vast majority of plant-available macro- and micronutrients in the soil. Decomposition is the net result of three concurrent processes: comminution or fragmentation of larger plant detritus by soil fauna, leaching of soluble compounds, and microbial catabolism. Annually, most newly fixed plant substrate returns to the atmosphere via respiration and microbial metabolism, especially in mature ecosystems, where the SOC fluxes and stores may be in a steady state.

*Stabilization* of SOC can be thought of as any process that acts to slow the decomposition process (Figure 1). Conversely, destabilization processes act to enhance decomposition. SOC stabilization can undergo several phases of progression and regression. Therefore the long-term fate of SOC is reflected in the SOC stocks, but over shorter timescales net fluxes between the soil and atmosphere can be either positive or negative.

In some soils, particularly in peats and permafrost soils, biological energetics play a key role in decomposition. Here, certain plant-derived substrates such as lignin and some paraffinic macromolecules selectively accumulate during the initial phases of decomposition until more labile substrates are exhausted. Thus, litter can take months to decades to fully decay. Water saturation, anoxia, and cold temperatures contribute to the buildup of highly decomposed humic and dissolved counterparts, with other soil properties (e.g., porosity) and environments (e.g., depth) influencing decomposition as well (Marschner et al. 2008).

Many soil processes regulate decomposition of SOM, and the ultimate fate of C represents the net effect of these regulation processes over time and the conditions under which the soil formed. Regulation processes that govern SOC decomposition and stabilization may operate on somewhat different timescales, simultaneously in different parts of the soil, or interactively over the course of soil formation. They can be broadly grouped into three categories: (i) biological energetics, (ii) physical accessibility of substrate to microbes and their enzymes, and (iii) mineral interactions of SOC. Enhanced regulation promotes C storage in soils, while limited regulation promotes decomposition and C loss (Figure 1). Persistence or accumulation of SOC is evidence for strongly regulated systems, but, given enough time in the absence of exogenic processes such as erosion, most (e.g., >99%) natural and xenobiotic compounds will be fully mineralized to CO₂ or other inorganic forms. As a result, on the timescale of millennia, only a fraction of a percent of carbon captured by NPP becomes stabilized as SOC (Figure 1).

### Stocks and pools of SOC

SOC stock refers to the weight of C within a given volume for a given soil depth (e.g., kg C m⁻² stored in the upper 1 m of soil). Dynamics of SOC, which are important for understanding land–atmosphere exchange, include assessments of C flux between land and atmosphere. These fluxes, naturally, affect C stocks. Turnover refers to the rate at which C cycles or is replaced by new C in the soil. Mean residence time is the average time that a C molecule might reside in the soil and is in most cases equal to the
SOILS AND THE CARBON CYCLE

turnover time. For example, a turnover time or mean residence time of 10 years suggests that most of the carbon in that soil is replaced over a decade. In reality, SOC is composed of a diverse range of compounds that vary in their inherent degradability and susceptibility to stabilization by physiochemical processes. As such, most major SOC models use a number of conceptual pools of C that have characteristic turnover times. In mathematical models, each pool has a separate fractional coefficient and a $C_{stock}$ that when multiplied represent the C flux from that pool. The sum of each product for each pool equals the decomposition flux from the soil.

The simplest conceptual framework for SOC utilizes the steady-state geochemical model in which inputs and outputs are assumed to be equal, and a mass balance is achieved. The steady-state model is justified by observations that rates of C inputs and outputs to and from soils, as measured by flux rates, are much larger than rates at which C is stabilized by soil. This assumption is typical of mature soils and ecosystems. In very young systems, however, rates of input typically exceed rates of output, thus C is sequestered. In cases of disturbance (discussed below), outputs can exceed inputs where C is lost from the system and over time can re-accumulate SOC or adjust to a new state.

Decomposition rates can be assessed for soil profiles, soil samples, soil horizons, or specific depth increments. Estimates of sizes and decomposition rates for pools can be based on fractionations using chemical extractions (McKeague, Brydon, and Miles 1970) or physical means (Christensen 1992). Carbon isotope ($^{13}C$ and $^{14}C$) measurements on soils and soil fractions are particularly helpful method, used to constrain sizes and rates of C pools (Balesdent, Wagner, and Mariotti 1988) even where measurements provide only partial information on pool sizes and turnover.

Spatial distributions of SOC stocks are important because they reflect important variations in plant productivity, nutrient status, and elemental concentrations. SOC stocks reflect the influence of pedogenesis on C dynamics over long time periods that include past changes in climate, vegetation, erosion, and so on. One way to assess SOC dynamics of such soils but at larger spatial scales would be to use spatial information of the five soil-forming factors (climate, organisms, relief, parent material, and time) as well as human impacts (Amundson and Jenny 1991). Changes in SOC contents along such gradients in temperature, slope, biome type, or elevation, which represents gradients of several environmental factors, have improved our understanding of controls over SOC. Nonetheless, few studies have provided multifaceted predictive equations for SOC and soil factors, probably because of the complexity of factors that control SOC. Stocks of SOC can be highly variable even over small distances if the environmental factors which influence C inputs, decomposition, and stabilization also vary at the same distances/gradations. In addition to inputs, decomposition and stabilization, lateral losses (such as to cropping, wildfire, or erosion) or gains (such as downslope sedimentation) can be as important to C stocks assessments as spatial variations in plant production and decomposition. Because of these circumstances, varying methods for assessing stocks have been applied across different spatial scales.

Maps of SOC stocks have been constructed from maps of soils and from biome or ecosystem association maps. Geospatial analysis of soil profile data has also been used in which multivariate data are used to represent spatial variations in SOC typically at regional scales. Classification or taxonomy-based SOC stock assessments relate key soil-forming processes to SOC stabilization (Batjes 1996). And yet, key uncertainties exist for a full accounting of the SOC budget at any
Spatial scale. First, SOC stocks develop over long, often millennial, timescales, whereas SOC fluxes are dominated by more recent and shallower processes. Second, soil maps often delineate complexes of several different soil types within a discrete mapping unit (polygon). Such complexes can reflect significant variations in properties, but the exact spatial location of any given soil within the map unit remains unknown. SOC maps or assessments have also been constructed for various biomes or ecosystems (Figure 2). Uncertainties in this approach arise from mismatches between current and past vegetation communities over the timescale of soil development. Last, data-intensive geospatial techniques have been applied to investigate and model SOC variations at local to regional scales (Viscarra Rossel et al. 2014). In instances where such approaches to mapping SOC have been compared, interesting differences emerge. Because of recent advances in global-scale geospatial mapping of key environmental variables, digital soil mapping applications are also feasible at global scales (Sanchez et al. 2009). Importantly, spatial variation of SOC stocks caused by human land use and management are often underestimated, as there is a mismatch in spatial scales between land management practices and the scales at which SOC stocks are normally mapped.

Figure 2  Mean SOC storage in soils, by depth (kg C m$^{-2}$) and according to biome. Based on data from Jobbágy and Jackson (2000), Table 3, and on data for permafrost regions from Hugelius et al. (2013) (n for tundra permafrost = 200; n for boreal forest permafrost = 324). Standard deviations about the mean were similar to the reported mean values for each depth increment, which reflect the heterogeneous nature of soils and SOC within the biomes.
Table 3  Summary of major management options for sequestering C in agricultural soils.

<table>
<thead>
<tr>
<th>Management</th>
<th>SOC benefit*</th>
<th>Confidence level†</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Shifts within an existing agricultural system</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Cropping and mixed systems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i. Maximizing productivity‡</td>
<td>0/+</td>
<td>L</td>
<td>Yield increases do not necessarily translate to increased C return to soil; potential trade off between increased C return to soil and increased decomposition rates</td>
</tr>
<tr>
<td>ii. Stubble retention</td>
<td>+</td>
<td>M</td>
<td>Greater C return to the soil should increase SOC stocks</td>
</tr>
<tr>
<td>iii. Tillage practices§</td>
<td>0/+</td>
<td>M</td>
<td>Reduced till has shown little SOC benefit; direct drill reduces erosion and destruction of soil structure; however, surface residues can decompose with only a minor contribution to soil</td>
</tr>
<tr>
<td>iv. Improved crop rotation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v. Soil amendments of C#</td>
<td>+++/++++</td>
<td>H</td>
<td>Direct input of C, often in a stable form, into the soil; additional stimulation of plant productivity</td>
</tr>
<tr>
<td>b. Pastoral systems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i. Increased productivity‡</td>
<td>0/+</td>
<td>L</td>
<td>Potential trade off between increased C return to soil and increased decomposition rates</td>
</tr>
<tr>
<td>ii. Rotational grazing</td>
<td>+</td>
<td>L</td>
<td>Increased productivity, inc. root turnover and residue incorporation by trampling but lacking evidence</td>
</tr>
<tr>
<td>iii. Shift to perennials</td>
<td>++</td>
<td>M</td>
<td>Plants can utilize water throughout the year; increased belowground allocation but few studies to date</td>
</tr>
<tr>
<td>c. Plantation forests</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i. Site improvements**</td>
<td>-/0/+</td>
<td>L/M</td>
<td>Complex interaction between increased aboveground productivity and microbial processing of C; drainage generally leads to losses due to increased mineralization</td>
</tr>
</tbody>
</table>

(Continued opposite)
Table 3  (Continued)

<table>
<thead>
<tr>
<th>Management</th>
<th>SOC benefit</th>
<th>Confidence level</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>ii. Lengthened rotations</td>
<td>+</td>
<td>M</td>
<td>Greater time between harvest disturbances</td>
</tr>
<tr>
<td>iii. Amendments from surplus wood†</td>
<td>++++/+++</td>
<td>L/M</td>
<td>Direct input of stable C into the soil; however, untested at large scales and implementation costs may be a barrier</td>
</tr>
</tbody>
</table>

Shift to different agricultural system

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Conventional to organic farming system</td>
<td>0/+++/++</td>
<td>L</td>
<td>Likely highly variable depending on the specifics of the organic system (i.e., manuring, cover crops, etc.)</td>
</tr>
<tr>
<td>b. Cropping to pasture system</td>
<td>+/++</td>
<td>M</td>
<td>Generally greater C return to soil in pasture systems; will likely depend greatly upon the specifics of the switch</td>
</tr>
<tr>
<td>c. Retirement of land and restoration of degraded land</td>
<td>++++/+++</td>
<td>H</td>
<td>Entire annual production is now returned to soil; active management to replant native species often results in large C gains</td>
</tr>
</tbody>
</table>

For full C accounting, C costs of transportation and application must be considered.

* Qualitative assessment of the SOC sequestration potential of a given management practice (0 = nil, + = low, ++ = moderate, +++ = high).
† Qualitative assessment of the confidence in this estimate of sequestration potential based on both theoretical and evidentiary lines (L = low, M = medium, H = high).
‡ Examples: irrigation and fertilization to overcome water and nutrient limitations, maximizing water- and nutrient-use efficiency.
§ Shift from conventional/full-inversion tillage to reduced tillage or direct drilling.
|| Examples: replacement of fallow with cover crop, increased proportion of pasture phases, pasture cropping where a cereal is direct drilled into a perennial pasture.
# Examples: manure, compost, biosolids, biochar.
** Examples: fertilization, drainage, liming.
†† Direct burial of wood products or conversion to biochar (Lenton 2010).
**SOILS AND THE CARBON CYCLE**

*Depth variations in SOC stocks* are particularly important because surface zones are where most disturbances occur. In most soils, C storage is highest in the upper profile, owing to proximity of plant detritus, and generally declines with depth. However, the high bulk densities and relatively slow turnover of deep soil C account for large masses of deep soil C. The decline in SOC concentration with depth is most pronounced in well-drained soils because decomposition is normally high throughout the profile while C inputs decline exponentially with depth. In poorly drained soils, the high soil water content has a strong influence on the depth profile of SOC because of moisture controls over gas diffusion, root production, and decomposition. In the deeper profile of all soils, C stabilization is influenced by soil mineralogy, aggregation processes, and factors that influence soil morphology. Over centennial to millennial timescales, processes such as cryoturbation (soil movements from frost heave that bury SOC), repeated sedimentation to the soil surface, or peat accrual can lead to the accumulation of very large SOC stocks. For example, in the circumpolar permafrost region, peatland SOC storage in Histosols (peats) and Histels (permafrost peats) often exceeds 100 kg C m$^{-2}$ (see Figure 2). Although the majority of the world’s peatlands are situated in the cold climates of the Northern Hemisphere (Gorham 1991), there are also substantial stocks of SOC in tropical peatlands, especially in Southeast Asia (Page, Rieley, and Banks 2011).

**Decomposition and turnover of SOC**

Variations in SOC turnover are as important as variations in SOC stocks because turnover is a gauge for biogeochemical reactivity and the potential for land–atmosphere C exchange. Fast turnover reflects a rapid cycling through the plant–soil–atmosphere cycle; slow turnover reflects the regulation of decomposition by soil processes and therefore a sequestration of C relative to fast cycling (Figure 1). Turnover of SOC typically declines or slows with depth in a manner similar to the depth distribution of SOC stocks (Masiello *et al.* 2004). The decline in turnover rates with depth can be linked to three mechanisms of C stabilization. When viewed as regulators of decomposition, these mechanisms can promote or diminish C storage and may dominate stabilization in different parts of the soil or over different timescales (Figure 1). In shallow soils, where C stocks and turnover are driven by water and nutrient cycling, plant and microbe-based energetics tend to regulate decomposition. Organic horizons at the surface typically reflect the chemistries of plant substrates and the decomposition pathways dictated by redox reactions, combustion history, and nutrient stoichiometries that regulate plant–soil interactions. In thicker organic soils, redox and moisture status persist as controls over decomposition (Maltby and Immirzi 1993); thus peatlands act as “hotspots” for C stabilization in any biome or ecosystem. In the upper profile of many mineral soils, such as in A horizons and transitional B horizons, SOC stabilization is typically dominated by aggregation (Six *et al.* 2002). Deeper in the profile, the accumulation of mineral weathering products creates conditions favorable for the stabilization of SOC (Torn *et al.* 1997) via mineral sorption and precipitation reactions.

Despite the large stocks and advanced ages of SOC in many deep soils (Masiello *et al.* 2004), deep SOC also can be reactive and dynamic, especially under conditions that disrupt the mechanisms of stabilization. Fast-cycling C, which has important implications for C cycling under climatic and anthropogenic disturbances, has been detected at depth by several investigators (e.g., see citations in Trumbore 2009).
The role of disturbance

Initially, many types of disturbance result in a net C release from ecosystems and soils, typically because oxidative decomposition or combustion can act on newly exposed organic matter and because disturbances temporarily cut off plant inputs to the soil. Left to recover, SOC stocks can re-accumulate according to the balance of plant inputs and decomposers if there is no additional factor retarding decomposition such as poor drainage (Figure 3).

*Generalized models of disturbance* could greatly help to improve assessments of SOC. Wildfire, cropping, insect outbreaks, earthworm invasions, exposure or burial by urbanization, sedimentation, erosion, thermo-erosion, and ground subsidence are all examples of disturbance events that change SOC stocks and flux, and which may be large or pervasive enough to significantly impact the land–atmosphere C exchange. Some processes, such as deep burns or permafrost degradation, can destabilize SOC stored at depth. Because deep C typically had been immobilized for centuries or millennia, its destabilization and release reflect significant perturbations to the steady-state models so often used in SOC assessments.

Several attributes of disturbance are important to the dynamics and fate of SOC. The depth of disturbance into the soil is important because of the depth dependency of inputs, stocks, and turnover times. In general, disturbance to vegetation and shallow soils are likely to have immediate impacts on the net SOC storage (emissions versus sequestration) but the impacts are usually short-lived (Richter et al. 1999) because of the rapid C cycling that occurs near the soil surface. In contrast, disturbances that impact deeper SOC stocks will have longer-lasting effects. The duration of the disturbance event also matters, as does the severity of the event relative to SOC or the soil properties that stabilize/destabilize SOC. The return interval of the disturbance event relative to input and turnover will dictate the C storage at intermediate to longer timescales. Last, the area impacted by the event is multiplied by the change in storage. Thus, it is paramount to assessing the impact of disturbance on atmospheric C budgets. Accounting for such
attributes of disturbance provides a framework for understanding impacts on SOC fate.

This link between SOC turnover rate and “recovery” period of the stock relates to the concept of steady state and to depth profiles of declining inputs and turnover times. In many assessments, disturbance causes sudden C loss, while recovery occurs over longer time periods (Sanderman and Baldock 2010). Deep disturbances and prolonged or pervasive disturbances that impact soils at depth are receiving more attention because the net impacts to SOC and land–atmosphere exchange are likely longer lasting owing to smaller inputs and slower turnover times in deep soils.

**Anthropogenic disturbances and management of SOC**

Today, 38% of the Earth’s land area is under agricultural production. Such widespread land-use changes have had a dramatic effect on the quantity and quality (C pools and nutrient provision) of SOC. Most soils under native vegetation have SOC levels that have evolved into a quasi-steady state. However, when land is cleared for agriculture, SOC levels normally drop by 30–50% from preclearance levels. Globally, it has been estimated that agriculture has resulted in the loss of 78 Pg C or about 19% of the anthropogenic C emission since 1850, with most this loss being attributed to the change in land use resulting in accelerated decomposition of SOC. The decomposition process results in a net release of CO₂, further accelerating potential global warming.

The reasons for this near universal decline in SOC following land clearing are twofold. First, natural C inputs to the soil decline. In a production system, a large fraction of aboveground plant material is removed in harvest or ingested by grazers. Most crops have also been bred to maximize C allocation to harvestable plant parts. Second, losses of SOC to decomposition and erosion typically increase owing to stimulation of microbial activity by fertilizer addition, irrigation, or tillage; relative to native plants, the labile nature of crop residues makes them more readily decomposable. The only exceptions to the near universal decline in SOC are in regions where there are severe constraints to plant growth and the implementation of agriculture brings management to alleviate these constraints. Examples include irrigation in water-limited regions and phosphorous fertilizer applications in ancient, highly weathered soils. In such circumstances, primary productivity is increased under agricultural management, overcoming the reduced C return and accelerated loss mechanisms listed above.

Agroforestry represents another major anthropogenic perturbation to the terrestrial C cycle. In general, soil C stocks in plantation forests are under influences similar to those in other agricultural practices. Soil C sequestration in agroforests is often delayed, however, due to the longer timescale of forest growth relative to crops. The initial SOC losses due to site preparation are compensated for after a couple of decades of growth and increasing C inputs to the soil.

There has recently been considerable interest in halting and even reversing the decline in SOC in order to improve soil health, while also sequestering more CO₂ from the atmosphere. In general, any of several crop management practices can result in increased inputs and/or reduced losses of SOC (Table 1; see also Lenton 2010). In pasture systems, improved grazing management and optimized nutrient management have been shown to increase SOC retention. Meta-analyses of field trial data indicate that local climate and soil properties will be important determinants of whether or not any
given management practice will improve SOC stocks. Importantly, these sequestration rates will diminish within a few decades as SOC contents achieve a new equilibrium. Given the recent estimates of SOC sequestration (Lenton 2010) and the benefits of SOC sequestration for soil health, increased research, efficacy, and verification are likely to occur.

SEE ALSO: Biomes; Climate change and permafrost; Ecoregions; Ecosystem; Ecosystem services; Global climate models; Soil fertility and management; Soil mass balance; Soils of cold and permafrost-affected landscapes; Soils of urban and human-impacted landscapes; Soils and weathering

References


SOILS AND THE CARBON CYCLE


Weathering is the physical, chemical, and/or biochemical breakdown of rocks and minerals at the interface between the lithosphere and the atmosphere (Bridge and Demicco 2008). It is the starting point of pedogenesis because it produces regolith, which is defined as loose, unconsolidated deposits at the Earth’s surface (Alexander 2013). Regolith originates from hard bedrock via weathering and biological processes. Regolith can have one of two origins: residual regolith (or residuum) which is formed in situ as bedrock weathers, or transported regolith which is transported to a site by wind, water, gravity, and so on (Schaetzl and Thompson 2014). The part of the regolith in which a soil develops is called its parent material. Rocks are weathered along fluid and gaseous pathways such as fractures, joints, faults, and so on. Weathered rocks are not necessarily covered by soil. Weathering is driven by the energy of the sun and involves exchanges between lithosphere, hydrosphere, atmosphere, and biosphere at low temperatures and under mainly subaerial conditions.

Weathering has traditionally been divided into physical (or mechanical) weathering and chemical weathering, but it is becoming increasingly apparent that biochemical weathering is also important (Velde and Meunier 2008). In reality, physical, chemical, and biochemical weathering processes work together. Alteration of the chemical or mineral composition of Earth materials does not occur during physical weathering, only a physical breakage of rocks into smaller pieces. Breakup of rocks and minerals create more surface area for chemical weathering, which operates mainly on the available surface area of minerals. Chemical and biochemical weathering both alter the chemical and mineralogical composition of rocks and regolith, producing new solid materials (e.g., clay minerals) and dissolved substances such as ions. Many of the dissolved solutes exit the soil system in groundwater or are biocycled by vegetation (Bridge and Demicco 2008).

During weathering, the fabric of the parent rock changes progressively from unaltered through to the most weathered zones (Figure 1). Weathered rock that retains much of its geologic fabric is called saprolite. Taylor and Eggleton (2001, 157–159) considered that regolith is composed of saprock and saprolite, where saprock is the altered coherent rock and saprolite is the altered soft rock. Together, saprock and saprolite form residual regolith, or residuum (Taylor and Eggleton 2001, 1–3), and above the saprolite, the altered soft rock is referred to as the pedolith (Figure 1).

Velde and Meunier (2008) proposed new definitions for altered rock facies based on physical properties that are easily observable in the field, that is, conservation or destruction of the primary rock structure and physical resistance (coherent, friable, soft, clay-enriched). Weathering profiles developed under temperate climate exhibit two altered rock types: coherent rock which may be sampled as self-sustained pieces that preserve the original rock texture and friable rock, which when sampled becomes gravel. According to these observations, one can distinguish three facies:
Physical weathering

Bedrock on the surface may weather, or disintegrate, physically from the stresses of heating and cooling, wetting and drying, and/or freezing and thawing. Physical weathering (PW) is the term used to describe the disintegration of rocks and minerals into smaller pieces. The main physical and mechanical processes of weathering are the *abrasion* and *comminution* of rocks and minerals as a result of erosive, expansive, and compressive forces and hydration along grain boundaries (Retallack 2001). Abrasion is the rubbing of rocks against each other, whereas comminution refers to the physical breakup of rock by faulting as a result of tectonic movement. Abrasion and comminution can occur simultaneously. They involve rocks rubbing against one another, chipping and breaking off small pieces.

Rocks deep below the Earth’s surface support the weight of the overlying column of rock. Lithostatic pressures can cause rocks to expand in the direction of least compressive stress, producing joints parallel to the principal stress direction (Bridge and Demicco 2008). As a result, all rocks contain fractures. This release of lithostatic pressure – often termed *unloading* – is most often due to the removal of overlying material by erosion.

Physical weathering acts on these joints to pry rocks apart and to break them into smaller pieces. Rocks can be physically weathered by *wedging* – prying them apart at cracks. The wedges can be formed by either ice or salt crystals. One of the main drivers of PW at the Earth’s surface is the *freeze–thaw* cycle of water and the associated growth of ice crystals. Wedging occurs as the result of the expansion of water as it is converted to ice. Cracks filled with water are forced further apart when the temperature drops below freezing. The splitting process in freeze–thaw produces fresh, angular fragments of rock (see Figure 2). The growth of salt minerals

**Figure 1** Classification of a typical weathering profile. Adapted from Migoń 2010. Reproduced by permission of The Geological Society.
in fractures and pores exerts increased pressure, analogous to ice. Additionally, the hydration of anhydrous salts (such as anhydrite to gypsum) within rocks (Bridge and Demicco 2008) can cause wedging and break off angular slabs or blocks of rock. These mechanisms are commonly cited as the main PW processes in arid regions, where extremes of temperature and moisture occur, and where many soils contain salts.

Other mechanisms of PW include the breakdown of rocks due to differential thermal expansion, and the contraction of individual rock-forming minerals during diurnal temperature changes. Plants and animals also have physical effects during weathering. A large number of animals make tunnels through hard rock and under the ground. Surface rocks also get crushed when animals tread on them or dig into them. Even earthworms make tunnels in the ground which leads to the breakdown of stones and rocks over a period of time. Plant roots at times grow deep into the cracks present in rock. As the plant roots grow thicker and deeper, rocks may begin to break. Both pressure release and wedging break up rocks and minerals, creating more surface area for chemical weathering (see Figure 2).

### Chemical weathering

Chemical weathering (CW) is the decomposition of rock materials by a series of chemical reactions. In most instances, the term biogeochemical weathering is more appropriate, as biota often interact chemically with the geological material to weather the minerals (Singer and Munns 2002). Chemical weathering generally occurs whenever water comes into contact with primary minerals (McBride 1994). The main processes of CW are solution, hydration, hydrolysis, oxidation, reduction, carbonation, and other reactions associated with inorganic and organic acids (Osman 2013). All forms of chemical weathering tend to promote the decomposition of minerals in rocks to a less resistant and more stable form, commonly a clay mineral (see Figure 2).

All minerals have two formation categories: primary and secondary. Primary minerals are formed at high temperatures and pressures through geologic processes. Usually, they form as molten materials (lava, magma) cool and harden. Consequently, they are unstable under atmospheric conditions. The stability of primary minerals (resistance against weathering) generally follows the Bowen reaction series, as those formed under high pressure and temperature tend to be less stable under atmospheric conditions. Secondary minerals form under atmospheric conditions (such as those found in soils) but may be unstable if conditions change from those of their formation. Primary minerals tend to release their ionic constituents as they weather; this can lead to the “new” formation (i.e., precipitation or neoformation) of secondary minerals. Secondary minerals can also be formed from the solid-state transition of a primary mineral, for example, mica→vermiculite. Many forms of CW also release cations, anions, and salts to the soil or regolith system; these may later recombine to form other secondary minerals.

The alteration processes that comprise CW occur between water (and its constituent dissolved substances, particularly acids) and minerals, operating at the mineral surface. Physical weathering, therefore, sets the stage for CW by increasing the surface area of minerals. Minerals are not ordinarily soluble in water, but in soils and regolith, acids, bases, and dissolved organic substances enhance their solubility. Most minerals in igneous and metamorphic rocks formed at around 1000°C and at many thousands of kilopascal pressure, under chemically reducing conditions. However, where they are exposed at the Earth’s surface at approximately 15–25°C, at
SOILS AND WEATHERING

Figure 2 An exposure of bedrock, showing the effects of physical and chemical weathering.

Table 1 Solubility of ions depends on their ionic potential.

<table>
<thead>
<tr>
<th>Ionic potential</th>
<th>Examples</th>
<th>Solubility/mobility</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤3.0</td>
<td>K⁺, Na⁺, Ca²⁺, Mg²⁺, Fe²⁺</td>
<td>Soluble, mobile</td>
</tr>
<tr>
<td>3.0–9.5</td>
<td>Fe³⁺, Al³⁺</td>
<td>Soluble under acidic conditions, mobility depends on pH</td>
</tr>
<tr>
<td>≥9.5</td>
<td>H₃SiO₃⁻</td>
<td>Forms oxi-anion complexes, solubility depends on pH, soluble around pH 7</td>
</tr>
</tbody>
</table>

*Ionic potential is defined as charge/ionic radius.

100 kPa pressure, and with abundant free oxygen dissolved in meteoric surface waters, primary minerals readily break down chemically (Bridge and Demicco 2008).

Dissolution occurs when rocks and/or minerals are dissolved by water. Dissolution of the most common ions of the Earth’s crust is largely determined by the pH of the soil solution (Figure 3). The dissolved material, typically as ions, is transported away in water or biocycled, leaving a void in the rock. Commonly, this kind of loss is what makes saprolite weaker than its parent rock, but because some minerals remain unweathered, the saprolite retains the structure of the parent rock. The readiness of a given ion for dissolution primarily depends on the ratio of its charge and radius (ionic potential, Table 1).
Hydration refers to the association of water molecules with minerals (Buol et al. 2011). During the hydration process, water molecules form a hydration shell around the ions of dissociated salts. Hydration primarily affects the mineral surfaces and (broken) edges of minerals, but in some cases it may dissolve the entire mineral (congruent dissolution).

Carbonation is a very active and common CW process, involving carbonic acid, and affecting mainly carbonate types of rocks (e.g., limestone, dolostone). Carbonic acid is naturally produced by dissolution of CO₂ in water. Carbon dioxide is available from the atmosphere and exists at high levels in regolith, because it is added as a product of microbial and root respiration. Thus, most carbonation processes are biochemical in nature. Carbonic acid is capable of dissolving many different types of minerals, bringing them into solution. The reaction for the dissolution of CaCO₃ is:

$$\text{CaCO}_3 + 2\text{H}_2\text{CO}_3 = 2\text{HCO}_3^- + \text{Ca}^{2+}$$

Carbonation can have an etching effect upon some rocks, especially limestone. The removal of cement that holds sand particles together leads to this disintegration.

Hydrolysis (splitting of the water molecules) is probably the most effective chemical weathering process. Hydrolysis occurs when primary minerals react with water, in the form of H⁺ and \(\text{OH}^-\) ions, to form other products. Hydrolysis is the reaction of acid (H⁺) with a cation-rich mineral grain to produce a clay mineral and various kinds of free cations. As a result, clay minerals accumulate in the soil and the cations are released, to later be leached or taken up by plants in solution (Retallack 2001). In hydrolysis, hydrogen ions (H⁺) in the water molecules replace other ions in the minerals. Due to this replacement, the mineral becomes susceptible to decomposition by further hydrolysis or other chemical reactions. Feldspars, the most common minerals in most rocks, are particularly prone to hydrolysis, reacting with water and free hydrogen ions to form secondary minerals. Hydrolysis is more rapid in silicate minerals characterized by well-developed cleavage (e.g., feldspars, pyroxenes) and is less effective in quartz. An example for hydrolysis (under acidic conditions) is shown by the following equation that describes the weathering of albite:

$$\text{NaAlSi}_3\text{O}_8 + 4\text{H}_2\text{O} + 4\text{H}^+ = 3\text{Si(OH)}_4 + \text{Na}^+ + \text{Al}^{3+}$$

Under less acidic conditions (pH 7) and long weathering, kaolinite may be formed, most typically in Ultisols:
When weathering further proceeds and Si(OH)$_4$ is progressively removed from the upper soil horizons, kaolinite disintegrates and loses silica (desilication), forming gibbsite (Figure 4). In this manner, reddish Oxisols with high gibbsite content develop. Al$^{3+}$ may also directly react with water, forming gibbsite. This process would be typical for soils of udic and perudic soil moisture regimes under tropical climates.

\[
\begin{align*}
\text{Al}^{3+} + \text{Si(OH)}_4 + \frac{1}{2} \text{H}_2\text{O} &= 3\text{H}^+ + \frac{1}{2} \text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4 \\
\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4 + 5\text{H}_2\text{O} &= \text{Al(OH)}_3 + 2\text{Si(OH)}_4 \\
\text{Al}^{3+} + 3\text{H}_2\text{O} &= 3\text{H}^+ + \text{Al(OH)}_3
\end{align*}
\]

Commonly, the silicic acid that is released by hydrolysis precipitates to form amorphous clay minerals like allophane and imogolite, and other, more structured secondary phyllosilicates (e.g., smectites; Figure 4). Short-range order amorphous silicates are often the first weathering

**Figure 4** Weathering pathways for the common rock-forming minerals of the Earth’s crust. Adapted from Mason 1966.
product on volcanic parent materials, and are primarily found in Andisols. Halloysite (Al$_2$Si$_2$O$_5$(OH)$_4$·2H$_2$O) may also appear as a weathering product of volcanic ash (as the end product of the allophane-imogolite-halloysite sequence; Figure 4).

At higher pH and Mg concentrations, montmorillonite may be formed by the resilication of kaolinite (Jenny 1980) due to the process called neoformation, which involves the dissolution of kaolinite and its subsequent reprecipitation in the form of montmorillonite (Figure 4).

Oxidation reactions are those in which an element suffers electron loss when forming a compound (Retallack 2001). Oxidation is the process of addition of oxygen to minerals. The absorption is usually from O$_2$ dissolved in groundwater and that present in atmosphere. Oxidation is more active in the presence of moisture, resulting in hydrated oxides (e.g., minerals containing Fe and Mg). Oxidation in minerals occurs when oxygen reacts with iron-bearing minerals to form iron oxide minerals (iron rusting), for example, goethite.

$$\text{FeOOH} + e^- + 3\text{H}^+ = \text{Fe}^{2+} + 2\text{H}_2\text{O}$$

The latter minerals give rocks red or orange colorations, indicative of oxidation and aerated conditions.

The environment in which soils form may also be reducing, usually when waterlogged conditions exist. Reduction is important to weathering in such soils, because many cations are more soluble when reduced than when oxidized (Singer and Munns 2002). A common process in soils is the vertical fluctuation of the groundwater table which changes the location of the boundary between the phreatic (saturated) and vadose (aerated) zones. The location of the groundwater table and the level of water saturation control the pH and Eh (redox potential) conditions of the subsoil and determine the oxidation forms of several ions, including iron, manganese, and arsenic. The changes between reducing and oxidizing conditions will result in the electron transfer reactions between ions like Fe$^{2+}$ and Fe$^{3+}$, Mn$^{2+}$ and Mn$^{4+}$, and As$^{3+}$ and As$^{5+}$.

Ion exchange refers to two different processes in the crystal lattice of clay minerals. The first process is called the isomorphic substitution and results in the appearance of permanent charges in the clay minerals due to the replacement of Si$^{4+}$ by Al$^{3+}$ or Al$^{3+}$ by Fe$^{2+}$. The second type of ion exchange occurs in the interlayer of the 2:1 clay minerals, primarily in micas. This type of chemical weathering process is typical under acidic conditions where a large concentration of H$^+$ is available for ion exchange. With the removal of more than 50% of the interlayer K$^+$ (and replaced by H$^+$), the sheet alignment in micas is lost and distortion of the lattice may happen; the mineral eventually collapses and weathers. With the complete removal of the interlayer K$^+$, vermiculite and smectite-type layered silicates are formed.

Organic acids also play important roles in weathering. These acids can either be released directly from the organism or be derived as byproducts of organic decomposition. Chelation is a process in which metal ions are usually bonded to an organic compound (ligand). These organic complexes usually have aromatic (ring-like) structures which incorporate the metal ion. Once formed, this structure is stable and may persist in the soil (Buol et al. 2011). Because of chelation, otherwise insoluble metal ions become soluble and may subsequently be removed from the soil. This process is typical in Spodosols, where metal oxides (sesquioxides, such as Al$_2$O$_3$ and Fe$_2$O$_3$) are removed from the A-horizon in percolating water, forming A$_2$ or E-horizons (Schaeztl and Thompson 2014). These substances then accumulate in the deeper Bs or Bsh horizons. In other cases, however,
organic ligands are stabilized by the metal ion, resulting in the relative accumulation of the organic-metal complexes in the upper soil horizons. An example of the ring-like structure of a chelate would be a complex between an oxalate anion and a metal ion:

\[
\text{M(H}_2\text{O)}_6^{n+} + \text{OOC}^- \text{COO}^- = ((\text{H}_2\text{O})_4\text{M} - (\text{OOC})_2)^{(n-2)+} + 2\text{H}_2\text{O}
\]

Any one rock or mineral, however, weathers at different rates in different climatic conditions. Chemical weathering is slow in cold and dry areas. In the humid tropics, not only is the rate of weathering rapid (especially chemical weathering), but also the weathering products are lost quickly. In nature, physical and chemical weathering operate individually, or concurrently. In most cases, one accelerates the other.

**SEE ALSO:** Lithologic discontinuities in soils; Paleosols; Soil biology and organisms; Soil and clay mineralogy; Weathering processes and landforms

**References**


**Further reading**


Soils as relative-age dating tools

H.W. Markewich  
Milan J. Pavich  
US Geological Survey, USA  
Douglas A. Wysocki  
National Soil Survey Center, USA

Many geomorphic investigations rely on obtaining or inferring the relative ages of soils across a landscape, or among several landscapes. Relative ages (e.g., A < B < C, etc.) provide information about landscape age and stability, and allow for better understanding of pedogenic and other surficial processes since the initiation of soil formation, often called time zero. Combined with geomorphic process information (e.g., erosion rates) or numerical dating of parent materials for specific pedons, relative-age information provides information critical to decipher landscape evolution and environmental change. Understanding processes and mechanisms involved in soil development within a specific landscape enables the reconstruction of its pedogenic history, including pedogenic development rates. A quantitative pedogenic history in turn provides for objective prediction of future soil landscape changes. As both anthropogenic land alterations and global climate impacts occur, pedogenic predictive models become increasingly important for long-range land use planning, especially for agricultural sustainability and water availability.

Soil development and the state of the soil

Soil formation, or pedogenesis, is the transformation of raw parent material into a soil body. Jenny (1941) stated that, “Soils are those portions of the solid crust of the Earth the properties of which vary with soil-forming factors …” The state of a soil is quantifiable by five soil formation factors, which Jenny listed as climate (cl), organisms (o), topography (r), parent material (p), and time (t). The factors are quasi-independent variables that describe the state of the soil system. Jenny expressed soils as a function of these factors in a fundamental equation: 

\[ S = f(cl, o, r, p, t, \ldots) \]

where S denotes a soil or a soil property, the factors (in parentheses) are a sequence of variables or state factors, and the dots (\ldots) represent unstated, minor factors, for example, human influence or dust additions. If all state factors are known, or at least held constant, the equation can be rewritten as 

\[ S = f(t, k, \ldots) \]

where k is the total influence of factors other than t, and k is held constant. To estimate soil formation rates and soil or geomorphic surface age, we can compare soil and weathering profile properties as functions of time, or as proxies, and compare soils or landforms of different age. In such a comparison, as the equation shows, we select for soil age (time) but endeavor to hold other soil forming factors (k) constant.

Time, like the other state factors, is an independent variable. Conceptually, it is useful to define a starting point in time, so we define here time zero as the point in time that soil-forming processes begin to act upon unaltered parent material. The goal of most age studies – in soils
SOILS AS RELATIVE-AGE DATING TOOLS

(or in geomorphology) – is to estimate time zero of soil (or landform) formation.

In his fundamental state factor equation, Jenny demonstrated that soil properties change mathematically as a function of time, but also that properties co-vary as a function of the other factors. Fortunately, we can minimize or hold constant factors other than time in such comparisons, thereby clarifying the relationships between soil properties and the soil-forming duration (time).

Traditionally, and for relative-age dating purposes, we consider soil development as progressive and unidirectional, with profile thickness and horizonation (vertical anisotropy) developing steadily and increasing with time. We recognize that soils may regress or develop episodically (e.g., due to climate variations, erosion, or mixing processes), but for general soil–time comparisons, these aspects of pedogenesis must, necessarily, be regarded as secondary. The proven concept of directional pathways in soil development enables soil-to-soil comparisons via examination of how their morphologies and properties change as a function of age. Water availability and flow direction exert strong influences on soil morphology. We, therefore, constrain such comparisons to soils of equivalent drainage and landscape position. Moderately well and well drained soils generally exhibit the best progressive development and yield the most useful comparisons for dating purposes.

The soil/landform complex

Within a selected landscape, distinctive soils with shared characteristics are associated with (form on) equally distinctive landforms or morphologic units. Soils and landforms that display consistent geomorphic association conjoin space and time and represent a soil/landform complex. Such a complex has formed within a constrained time. Soil/landform complex examples include fluvial terraces and glacial moraines, as well as alluvial fans, pediments, sand dunes, emergent marine barriers, and landslide deposits. Because soil/landform complexes can have internal morphology differences, equitable comparisons as a function of time require that soil/landform complexes represent similar morphologic components within a landscape. For example, we cannot equitably compare footslope soils to those on summits.

An archetypal soil/landform complex would be an array of soils developed in alluvium on stream terraces that occur at differing elevations in a valley (Figure 1). Each terrace and its associated soils represent a soil/landform complex.

When comparing soils among two or more soil/landform complexes within the valley, it is important that the parent material be compositionally similar and that the soils have developed in similar topographic positions, and have undergone minimum erosion and (or) surface accretion. That is, we endeavor to hold the state factors other than time constant. When comparing very young soils, it is also advantageous to select soils that have developed under similar vegetation. For this river valley example, each soil/landform complex started developing at a different time (each soil has a different time zero), but all the soils have remained subaerially exposed and continued to develop under the same climate. By controlling or minimizing differences among the other four soil forming factors, we can prudently assume that the terrace soils differ primarily due to age (time since initiation of soil formation). This approach allows the researcher to evaluate the influence of time, or age, on soil development.

Geomorphic principles (cross-cutting relationships and ascendancy) provide an independent means to define relative terrace ages. Terrace
SOILS AS RELATIVE-AGE DATING TOOLS

Qt3
age: >700 ka to 1 Ma
pendon thickness: >2.3 m;
argillic horizon thickness: >2 m
clay maximum: ≥45%;
clay mass: >130 g cm⁻²;

Qt2
age: ≈500–700 ka
pendon thickness: >1.6 m;
argillic horizon thickness: ≈1.3–2.0 m
clay maximum: ≈36%;
clay mass: >80 g cm⁻²;

Qt1
age: ≈200–500 ka
pendon thickness: >1.6 m;
argillic horizon thickness: ≈0.4–1.3 m
clay maximum: ≈36%;
clay mass: ≈20–80 g cm⁻²;

Figure 1  Schematic diagram of a Quaternary soil chronosequence composed of three soil/landform complexes developed in alluvial terraces in the Chesapeake Bay area of the Atlantic Coastal Plain of the United States.
SOILS AS RELATIVE-AGE DATING TOOLS

Age increases with elevation. We know which terraces are older and which are younger, and this knowledge allows us to evaluate soil development on older versus younger terraces. We can apply this soil information to other landforms and evaluate their age using data from their soils as well.

The soil chronosequence

A sequence of soils developed on similar but different aged landforms in a given terrain defines a soil chronosequence. Simply stated, a soil chronosequence (Figure 1) is an array of soils in a specified geographic area, developed in similar parent materials, on similar landforms, and in similar topographic positions; these soils differ in their degree of profile development due predominantly to differences in age, or time since initiation of soil formation. As noted by Stevens and Walker (1970), the study of soil chronosequences is essentially examining time as a soil forming factor. Soil chronosequences can occur in all terrains. The fundamental unifying condition is that landforms or geomorphic surfaces of different age have formed in a region with uniform climate and parent materials. Typical examples are adjacent glacial moraines, or a sequence of fluvial terraces. Fluvial terrace chronosequences occur across various geologic terrains and climatic regimes, making them the most commonly studied chronosequence type. Soil chronosequences, however, also occur in other landscape systems, such as those produced by volcanic activity (adjacent basalt flows), eolian processes (stable and reactivated sand dunes), and crustal deformation (uplift) and (or) eustatic sea level changes (stepped marine terraces and shorelines).

Vreeken (1975) designated several chronosequences types, with post-incisive and pre-incisive chronosequences being the two most common. A post-incisive chronosequence consists of soils that began forming at different times on deposits of different age, and which have all remained exposed at the surface (Figure 1). They include soils developed on progressively younger surfaces, are the least ambiguous to interpret, and, hence, the most commonly investigated. A pre-incisive chronosequence includes soils that developed during the same time in one deposit, some of which subsequently undergo burial, but at different times. This chronosequence type may also include a soil that remains at the ground surface. An example of this chronosequence would be a soil array developed on a land surface that through time is partially buried by intermittent eolian activity. When the burying sediment is thicker than some critical level, soil formation ceases in the buried soil. The buried soils retain their original time zero, but development time ends at burial. The oldest soil in this pre-incisive chronosequence remains unburied. The unburied soil has the same time zero as the nearby buried soils but development continues to the present. The post-incisive chronosequence that includes soils developed on progressively younger surfaces, is the least ambiguous to interpret, and is routinely the most investigated.

Vreeken (1975) describes two additional chronosequences: (i) time-transgressive without historical overlap and (ii) time-transgressive with historical overlap. A time-transgressive without historical overlap chronosequence is a vertical soil sequence formed in, and separated by, parent materials. Soil development ceased upon subsequent burial(s). An example is a thick loess sequence with inclusive paleosols (Figure 2). Individual loess units and their associated paleosols represent climate cycles (e.g., dry vs wet climate). Such a chronosequence involves a succession of soil-forming environments; it does
not represent an age developmental sequence. To evaluate soil development rates through time in this chronosequence type requires other age controls. A time-transgressive with historical overlap chronosequence is a vertical or lateral soil sequence that developed on surfaces of differing age, which undergo burial at different times. Although a common chronosequence type, the pedogenic complexity and episodic history of pre-incisive sequences complicates chronosequence studies.

Soil age, soil development, and weathering

Soil chronosequences can occur on surfaces that have wide age ranges. Soils in most glaciated regions are late Pleistocene or younger in age. Unglaciated regions, however, can have soils and surfaces that span the entire Quaternary Epoch, and possibly the late Neogene. Many studies that use soils for determining relative age are in landscapes that are 100 000–1 000 000 years in age (middle-to-early Pleistocene) (Figure 1). In geologically recent terrains, for example, areas of late Quaternary and Holocene volcanism, and in anthropogenic terrains such as spoil piles in old mining areas and razed-building sites in urban landscapes, soil age commonly ranges from a several thousand years to as little as a few decades. Regardless of setting, soil formation depends upon land surface stability (no erosion or subsequent deposition). In other words, soil formation cannot begin until the geomorphic surface upon which the soils form is stable, that is, no longer eroding or accumulating new sediment.

It is important to distinguish among three interrelated terms – soil age, soil development, and weathering status. We define soil age as the amount of time that has passed since the soil began to form (time zero). Explicitly, it is
the time elapsed since the land surface became stable – when deposition of the parent material ceased – and processes related to weathering and soil formation began. Soil age generally is expressed in years before present. Soil development is measured or evaluated by the number, type, and thickness of genetic soil horizons. It encompasses any of a number of soil characteristics observable at the time of study. In general, the greater the number of horizons, the thicker the horizons, and the greater the differences between horizons, the more developed, and hence the older, the soil is assumed to be. Soil formation rates vary as a function of the kinds and intensities of pedogenic processes that are operative, and these are, in turn, a function of the climatic setting, parent material composition, and biotic and topographic influences. Consequently, soils with equal development are not necessarily equal in age or vice versa, that is, they may simply have been subject to different rates of soil formation, due to their environmental setting. Weathering status reflects the mineral composition of a soil. Chemical weathering sequentially removes and/or concentrates minerals in soils, depending on their weathering resistance. In general, when iron and aluminum oxides and kaolinite dominate as weathering products in soils, they indicate a more advanced weathering state. Soils dominated by smectites, mica, and feldspars, which are less stable in well-drained, slightly acidic surface environments, represent a less weathered state. Because parent material composition determines the initial weathering state for a soil, mineral composition alone is not a consistent age indicator. That is, some young soils may have inherited minerals in their parent material that generally occur in older, highly weathered soils. The soil here is young (recent time zero) but its weathering status is similar to that of older soils. Thus, one often compares soil–weathering status in the profile to that of the parent material to avoid this complication.

In chronosequence studies, a single numeric age can serve as an anchor that guides further relative age assessment based on soil development. Numeric soil age refers to the absolute age of a soil or a geomorphic surface. We often express numeric age as in years ago or years before present. We can obtain a maximum soil and associated landform age by determining the age of the parent material in which the soil has formed. The numeric age of the soil parent material can be determined, or sometimes bracketed, by using any of several techniques. The most common are radiometric techniques, such as radiocarbon analysis, and luminescence techniques, such as optically stimulated luminescence. An alternative approach for soils formed in pre-Pleistocene parent materials is to identify age distinctive fossils, thus providing a maximum biostratigraphic age for the parent material. We can use radiocarbon analyses of the stable soil organic-matter fraction to date soils younger than about 40 kyr. In arid and semi-arid environments, we can employ the radiometric techniques of radiocarbon and isotopic uranium and thorium disequilibrium analyses to date pedogenic calcium carbonate; this method is particularly useful for soils >40 kyr.

In most environments, determining the soil’s numeric age is more difficult than determining the parent material age. Meteoric $^{10}$Be soil residence time analyses, however, provide a promising solution to this dilemma. Such analyses give an estimate of the minimum time required for a soil to form (Pavich et al. 1986) and are particularly useful in humid environments, where little organic matter or carbonate accumulates in the soil or parent material. In these environments, $^{10}$Be soil residence time analyses is one of the few methods available for determining numeric soil age, especially for soils >40 kyr.
In summary, there exist a number of numeric dating methods to obtain absolute soil or sediment ages; this list continues to grow. In many soil studies, however, dateable material is lacking, the techniques are costly, and many rely on strict and often unobtainable assumptions, which is why we turn to relative-age dating techniques.

**Relative ages based on soil chronofunctions**

Relative-age dating establishes the chronological sequence of events, that is, which soil formed first, second, and so on, or which soil is younger or older. Relative-age dating does not provide the exact soil or parent material age, for example, how long ago in years a soil or geomorphic surface formed. Relative-age dating linked with chronofunction formation rates can estimate the magnitude of the age differences between two or more soils. By geomorphic principles (ascendancy), it can be established that geomorphic surface A is older than surface B. We can use quantified soil development differences to additionally show that surface A is slightly older than surface B, or is three times older, or ten times older. This added degree of relative-age distinction allows for more detailed age discrimination. For example, in three adjacent moraines it can be determined whether they resulted from three distinct advances (OIS 2, 4, and 6) or from a single major advance with three short pulses/surges (all OIS 2).

Relative-age dating relies on observable relationships and properties that yield a general chronological rank. Moreover, relative-age dating creates a chronologic framework that gives greater relevance to measured numeric ages, when they are available. We can assign relative ages to soils and (or) soil/landform complexes based on (i) geomorphic and stratigraphic principles (i.e., stepped river terraces and floodplains or nested glacial moraines), (ii) degree of morphologic soil development (horizon thickness, number, and distinctness), and/or (iii) differences in physical and chemical soil properties and weathering, as compared to the parent material. Soil morphology and development, as quantified by numeric age or constrained by relative geomorphic age, permit judicious estimates of the soil formation time, which is also known as the soil-forming interval. A key example is the accumulation of pedogenic carbonate (Bk, Bkk, and Bkkm horizons) in arid regions. Detailed field soil observations (Gile, Peterson, and Grossman 1966), coupled with age and geomorphic surface studies (Gile and Grossman 1979, and other investigations), defined a six-stage morphogenetic sequence that conceptualizes this process. This sequence is widely used to infer relative soil and landform age in arid settings where age measurements and geomorphic constraints are lacking or unknown.

Many soil properties (e.g., chemical depletion and clay content) commonly exhibit uni-directional change over time. Because of this, we can use the progressive soil development encompassed in a soil chronosequence to mathematically express soil formation as a function of time, that is, a soil chronofunction. Establishing a chronofunction requires a chronosequence with bracketed ages (geomorphic principles) or known ages (numeric techniques). When we plot soil data as a time function, it often is expressed as best-fit regression model (Figures 3 and 4). The function type (linear, logarithmic, exponential, etc.) varies, depending on the soil property measured and its pattern of change through time (Figure 4). In a chronofunction, soil properties represent dependent variables; time is the independent variable. The chronofunction is a mathematical relationship that quantifies the change in the observed soil or soil properties...
through time. Its slope defines the rate of change for that soil or soil property. Consequently, we can use chronofunctions to calculate development rates for measured soil properties. Figure 3 shows B horizon clay formation rates for glacial tills in different climate settings. It is assumed that soil properties change in a systematic way (linearly, logarithmically, etc.) with time (as shown by Figures 4 and 5) and that they can be measured in all of the soils for which comparisons are being made. In general, many soil properties change logarithmically through time, such as horizon and solum (A and B horizon) thickness, clay mass, bulk density, rubification (reddening), and weathering depth.

After chronofunctions are established and their soil–age relationships verified, we can then use them to estimate soil and landform ages for surfaces and places where such ages are otherwise unknown. We accomplish this by reversing the function to predict formation time or age from the given soil properties. Of equal import, a
SOILS AS RELATIVE-AGE DATING TOOLS

Figure 4  (a) Linear chronofunction developed using chemical indices ($\text{Fe}_2\text{O}_3 + \text{Al}_2\text{O}_3/\text{SiO}_2$) from soil B horizons with the maximum clay content. The soils developed on river terraces. (b) Logarithmic chronofunction for $\text{Fe}_2\text{O}_3 + \text{Al}_2\text{O}_3/\text{SiO}_2$ plotted against ages of soils developed in Atlantic Coastal Plain and outer Piedmont sedimentary deposits of the eastern United States. (Modified from Markewich and Pavich 1991, Figure 5b.)

chronofunction can also be used to determine how long it would take to develop selected types of profile, horizon, or property characteristics. Multiple measured ages, compiled for several soil/landform complexes, can theoretically facilitate the most rigorous and reliable calculations of soil and (or) landform development.

Soil properties that have the best potential for chronofunction formulation and, hence, relative-age assessment, are those that display change with depth through time. Examples include field-measured attributes for various horizons (e.g., thickness, color, texture, clay films, structure, consistence, pH, thickness and characteristics of organic and inorganic carbon accumulation, salt accumulation, etc.) and, where appropriate, properties of the weathering profile (weathering/oxidation depth, mineral abundances, and color) relative to the parent material. Laboratory-measured soil properties that change with depth also work well in this type of application. Examples include content of clay, carbon (inorganic and organic), specific major and minor elements (i.e., Al, Ca, Fe, Na, P, Rb, Si, Ti, Zr), and ratios thereof. These properties (field or laboratory), individually or in combination, are used for intersoil comparisons and serve as dependent variables in chronofunction formulations. When compared to the same properties in the unaltered parent
material, they allow for the comparison of one soil/landform complex to another within the same terrain.

Soil development indices

Because direct comparison of multiple and complex field and laboratory data across soils in a chronosequence can be complex, confusing and/or difficult, researchers often use the data to derive more integrative measures, known as soil development indices. Indices allow for simpler, yet quantitative, comparisons of soils of different ages. Importantly, soil development indices capture the complexity of soil development in a single number. This simplifies comparisons within or across chronosequences, when such indices are plotted against time to develop a chronofunction. A soil development index consists of a comparison of selected properties, usually by horizon, to that in the parent material or other selected control. This index may be expressed as a number or as a ratio.

Numerous soil morphological and chemical indices exist to characterize weathering and soil profiles. The focus here is on indices most applicable for soils developed in similar parent materials. Soil indices can be formulated for a single soil property or for a combination of soil properties, for one horizon or a combination of horizons, or for the entire soil profile (a soil development index). Commonly used soil profile indices include several based on soil morphology: (i) the index of Bilzi and Ciolkosz (1977), a rating scale that uses soil...
properties to assess genetic horizon development strength relative to the C-horizon; (ii) the soil development index (Harden, 1982), which relates up to eight field soil properties (clay films, texture plus wet consistence, rubification (hue and chroma), structure, dry consistence, moist consistence, color value, and pH) to parent material properties; and (iii) the POD index of Schaetzl and Mokma (1988) for spodic (podzol) soils, which relates soil development to the number and colors of soil subhorizons. Other indices relate laboratory-measured soil properties (particle size, elemental chemistry, etc.) for either the entire soil profile and (or) parent material. One example is the podzolization index of Duchaufour and Souchier (1978), which is a measure of the degree of chemical differentiation within a podzol profile. Indices developed for the whole soil include the index of profile anisotropy (IPA) (Walker and Green 1976), which relates soil-property variation of one horizon to the weighted mean of that property in the soil profile, and the modified IPA (Birkland 1984), which is similar to the IPA but takes into account horizon variations with respect to the parent material. We use horizon-to-horizon and horizon-to-whole-soil comparisons to develop a soil morphology index that relates field properties of the solum to those of the C horizon. Other properties helpful for estimating relative soil age include solum (A and B horizon) and argillic horizon thickness, and elemental ratios such as SiO₂/TiO₂ and Zr/Rb, which indicate the degree of soil weathering. Other elemental ratios employed for relative soil age estimation, particularly in warm–humid to tropical environments, include the ratio of K₂O + MgO/TiO₂ and the ratio of Fe₂O₃ + Al₂O₃/SiO₂ to clay mass and to time (Markewich and Pavich 1991) (Figure 4b).

Some indices are simple to calculate. A good example is the Duchaufour and Souchier (1978) podzolization index. Aluminum contents in the A, B, and C horizons are plotted and the values for the A and C horizons are connected on a graph, using a straight line from the maximum aluminum content in the B horizon to the y-axis. The amount of aluminum represented by the intersection of the two lines is divided into the maximum aluminum content to give the podzolization index, Kₐ. Figure 6 is modified from the author's 1978 publication and graphically represents the technique.

Some indices, such as the Harden index of soil development and the Bilzi–Ciolkosz index require a large suite of soil property data and are more complicated to calculate. The advantage is that more data goes into the final index values and, hence, the indices may be more robust. Figure 7 outlines the steps for determining the soil development index for a soil profile. We often determine which soil property or combination of soil properties is the best for a particular soil by measuring and comparing as many of the pedogenically significant soil properties as is practical. We then compare the calculated index values to age ranges, if known, either for the parent material, or for the soils.

By synthesizing results of a large number of chronosequences in a number of environments, we can compare soil forming rates across climates and parent materials. Bockheim (1980) developed chronofunctions for 32 chronosequences from seven climatic regions using linear and nonlinear models. Graphical comparisons of the chronofunctions illustrate how soil properties develop at dramatically different rates, depending mainly upon climatic regime (Figure 3) or parent material.

As chronosequence studies have progressed, research goals have expanded to include complex and interrelated challenges, such as (i) understanding how and why soil development rates change through time, (ii) whether rate
changes are due to the soil or a soil property attaining a steady-state, (iii) whether thresholds in soil development that mark significant decreases or increases in the rate of development exist, and (iv) the reasons or drivers for this type of threshold, such as the formation of impermeable horizons, climatic changes, such as increase in temperature or precipitation, or anthropogenic effects, such as deforestation.

Regardless of the specific goals, the soil chronosequence studies offer promise for greatly advancing our knowledge of the pathways and rates of soil and soil property transformations through time. The importance of this knowledge cannot be overstated. Soils are the underpinning of all terrestrial ecosystems. They function as biodiversity refuges, filters for natural and recycled waters, sinks for nutrients, carbon, and other greenhouse gases, and form the basis for our agroeconomy. Understanding the length of time necessary for soils to develop, and their relative ages, in different environmental settings is essential for model development in a world of increasing climate change with increasing acreages of human-modified soils and parent materials.

SEE ALSO: Soil development: conceptual and theoretical models; Soil development: numerical indices; Soils and weathering; Soils in geomorphic research; Soil mass balance
SOILS AS RELATIVE-AGE DATING TOOLS

Describe the soil profile

Assess the parent material

Quantify each field property

Normalize each quantified property

Comparison of soil properties

Calculation of the soil development index

Sum the normalized properties for each horizon

Divide each normalized property value by the total number of quantified properties. This gives the “horizon index”

Multiply the value determined in step (4) or (6) by the thickness of the horizon

Sum the horizon products determined in step (7) through the soil profile. This gives the “profile development index.”

Example for calculations in step (3)
Assign 10 points for each shift from parent material toward the finer, the more plastic, and the more sticky soil texture classes. Such as, if a horizon is one class finer in texture than the parent material than texture is assigned 10 points. If the horizon stickiness is two classes greater than the parent material, stickiness is assigned 20 points. Similarly, If horizon plasticity is one class more than the parent material, plasticity is assigned 10 points. This gives an overall texture value of 40 points.

Methods for normalizing (step 4) soil property values calculated in step (3) differs by property or combination of properties, but in general the process for step (4) is to divide by the maximum possible value of the point scale for that property. This produces a scale ranging from 0 to 1.

Figure 7 Outline of the steps required to determine the Harden soil development index. (Modified from Harden 1982, Figure 2.)

References

SOILS AS RELATIVE-AGE DATING TOOLS


Further reading

Soils in archaeological research

Vance T. Holliday
University of Arizona, USA

Soils are a potential source of much information in archaeological studies at all scales, from specific archaeological sites (or even parts of sites) to broad regions. Soils are literally a part of the physical stage upon which humans evolved. Soils are not only integral components of most natural landscapes, they are also significant components of cultural landscapes. Beyond physically supporting humans and their endeavors, however, soils are also indicators of the nature and history of the physical and human landscape. They record the impact of human activity, they are a source of food and fuel, and they reflect the environment and record the passage of time. Soils also affect the nature of the cultural record left to archaeologists because they may encase archaeological materials and archaeological sites. They are a “reservoir” for artifacts and other traces of human activity.

Soil forming processes also are an important component of site formation processes. Pedogenesis influences which artifacts, features, and environmental indicators (floral, faunal, and geological) are destroyed, which are preserved, and their degree of preservation. Two realms of soil science are most directly applied in archaeology. Pedology deals with soils as three-dimensional bodies, and thus is intimately related to the landscape and other aspects of the environment. Soil chemistry is concerned with the chemical properties of soils and thus provides a means of detecting past human impacts on soils.

Pedology in archaeology

Several components of pedology and the allied subdiscipline of soil geomorphology are widely used in archaeology, in particular in the subdiscipline of geoarchaeology. These components can be grouped into soil surveys, soil stratigraphy, soil geomorphology, and site formation processes.

Soil surveys in archaeology

With proper interpretation, a considerable amount of information can be gleaned from soil surveys for use in archaeological and soil-geomorphic investigations (Scudder, Foss, and Collins 1996; Holliday 2004). Surveys can provide a good base map and useful aerial photographs. They can also provide valuable information about surficial geology (Figure 1). Perhaps most significantly, soil surveys have the potential to aid in understanding landscape evolution and estimating the ages of various parts of the landscape (Figure 1). This is useful in predicting presence of archaeological sites and their age.

Most archaeologists are aware of soil maps as a useful resource, but their usefulness may be overestimated or misunderstood. For example, the purpose and preparation of soil surveys impose limits on their utility in archaeology. They are not designed for use as guides to local geology or geomorphology or for reconstructing prehistoric vegetation patterns. Instead surveys reflect mainly the local bedrock, sediments, and...
landforms. The degree to which the surveys accurately depict soil-geomorphic relations (see below) will vary tremendously from survey to survey, depending on the size of the features of interest, the area, and the training and experience of the mappers. During the course of the survey there will generally be little emphasis on or data gathering concerning the origin or historical development of the soils. Further, the interpretation of soils at a specific site can be problematic because soil surveys represent generalizations. Nonetheless, careful reading of descriptions and map legends for soils depicted on maps can provide general indications of geoarchaeologically important characteristics such as depth to bedrock, site wetness, landform type, and degree of soil development (see below).

Soils as stratigraphic markers

The unique physical and chemical properties that distinguish soils from sediments make soils quite useful for stratigraphic subdivision and correlation. Soils have thus been commonly used as stratigraphic markers in archaeological research (Mandel and Bettis 2001; Holliday 2004) (Table 1). However, the nature of soils also necessitates the exercise of a certain amount of caution in their use in stratigraphy.

The presence of a soil in a stratigraphic sequence marks the passage of some amount of time with no or very little erosion or sedimentation. However, the parent material for that soil, which is as old as or older than the soil, may have been deposited virtually instantaneously. If the soil in such a sequence is buried, the contact

**Figure 1** Soil map of the Harappa area, Pakistan (based on Amundson and Pendall 1991; Belcher and Belcher 2000). A soil survey around this famous Bronze Age site aided in reconstructing the evolution of the Holocene landscape and in understanding the relationship of the site to the Ravi River. The site was constructed on the oldest alluvial terrace (unit 20) but also adjacent to a channel (Rc) active during occupation. The pattern of stream meanders is clearly apparent around the site (soil units Rc, Sc, 16, 17, 18, 19).
Table 1  Stratigraphy of the “Loessic Paleolithic” in southern Tajikistan (modified from Holliday 2004, Table 6.4). Buried soils were key stratigraphic markers throughout the region, used to correlate multiple Paleolithic occupations.

<table>
<thead>
<tr>
<th>Soil number</th>
<th>Soil type*</th>
<th>Pebble tradition</th>
<th>Blade tradition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sites</td>
<td>Industries</td>
</tr>
<tr>
<td>1</td>
<td>Serozem</td>
<td></td>
<td>Tutkaul III</td>
</tr>
<tr>
<td></td>
<td>(Orthent? Cambid?)</td>
<td></td>
<td>Shougnou I</td>
</tr>
<tr>
<td>2</td>
<td>Lt. Cinnamon</td>
<td></td>
<td>Shougnou II</td>
</tr>
<tr>
<td></td>
<td>(Ustochrept)</td>
<td></td>
<td>Shougnou III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Shougnou IV</td>
</tr>
<tr>
<td>3</td>
<td>Lt. Cinnamon</td>
<td>Khonako II (?)</td>
<td>Flakes debris, pebble tools technology</td>
</tr>
<tr>
<td></td>
<td>(Ustochrept)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Brown</td>
<td></td>
<td>Lakhuti III</td>
</tr>
<tr>
<td></td>
<td>(Inceptisol?Alfisol?)</td>
<td></td>
<td>Nonclear with crude flakes, debris, pebble tools technology</td>
</tr>
<tr>
<td>5</td>
<td>Brown</td>
<td></td>
<td>Lakhuti I</td>
</tr>
<tr>
<td></td>
<td>(Inceptisol?Alfisol?)</td>
<td></td>
<td>Evolved local pebble tools technology with a few Levallois elements</td>
</tr>
<tr>
<td>6</td>
<td>Brown</td>
<td>Karatau I</td>
<td>Local pebble tools technology without systematic flaking</td>
</tr>
<tr>
<td></td>
<td>(Inceptisol?Alfisol?)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Brown</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Brown</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B†</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M†</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Kuldara</td>
<td>Short flakes, pebble techniques without choppers</td>
<td>Donguttuo</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Soil type listed by Ranov (1995, Table 1) is indicated, plus likely soil taxonomy equivalent in ( ).
† B/M indicates position of the Brunhes–Matuyama magnetostratigraphic boundary.
of the top of the soil with the sediments that bury it marks the gap in time. This means that cultural material found at the top (in or near the soil) and bottom (deep below the soil) of even a thick deposit may be nearly contemporaneous. In addition, the artifacts near the top of the unit may be considerably older than an artifact immediately above, on the paleo-surface. Moreover, artifacts found on the paleo-surface may well represent a mixture of cultural material left on the surface during the entire period of soil formation. The artifacts and occupations are also subject to mixing by biological and geological activities, as well as stratigraphic compression. Bioturbation, in particular, is well known as an agent by which artifacts on the surface can quickly become buried (Johnson 1990; Van Nest 2002) (see below). The surface of a buried soil (or “paleosol”) may be a zone likely to contain archaeological material because it represents a stable landscape that persisted for some amount of time. The contact between the buried soil and the overlying sediment, however, may be time-transgressive because burial of the surface may take place at different times in different places or continuously over a long period of time.

Soils can be useful in identifying various sedimentary strata, but they should never be considered strata themselves or referred to as “strata” or “layers,” or otherwise treated like geological deposits (Holliday 2004). A soil profile is imprinted over and into geologic deposits through time, generally from the surface of the parent material downward, and, therefore, the boundaries between soil horizons often have no relationship to geological layering. Depending on the slope of the surface associated with the soil and variations in the permeability of the parent material, soil horizons can crosscut depositional layering.

For many of the same reasons, pedogenesis also can obscure sedimentary features and contacts. One of the criteria often used to differentiate between a B horizon and a C horizon is whether bedding in the parent material has been destroyed, but evidence of bedding can persist in a B horizon if the original beds included lenses of gravel. Otherwise, to document the sedimentology of parent material obscured by pedogenesis will require laboratory analyses. However, the original particle-size distribution of the parent material can also be modified by pedogenesis. Significant amounts of illuvial (postdepositional) clay can accumulate in soils, which can affect sedimentological interpretations.

Soil geomorphology

The “state factor” approach to soil genesis is the theoretical framework for much of pedology and is widely applied in soil geomorphology (Jenny 1941; Birkeland 1999; Holliday 2004; Schaetzl and Anderson 2005). Jenny (1941) defined the factors of soil formation (the external factors that drive soil forming processes) as climate, organisms (flora and fauna), relief (or landscape setting), parent material, and time, often written as the equation

\[ S = f(d, o, r, p, t \ldots) \]

where the upper case \( S \) is the whole soil. This equation defines the state of the soil as a function of the five factors (the “state factors”) and other, unspecified factors of local or minor importance (\( \ldots \)) (e.g., dust fall, groundwater fluctuations). The equation can be solved by studying the variation in a soil as a function of one factor, keeping the others constant or accounted for. For example, variation in soils due to differences in climate could be studied by keeping all factors except climate constant. Birkeland (1999) applied this approach to studying the evolution of soils. This approach also has applications in archaeology. The factors of soil formation that
are generally of most concern in archaeology are time and climate. Specifically, this includes using soils as indicators of age, past environments (climate and vegetation), and environmental change.

The concept that some time must elapse before a soil can form is probably one of the most significant soil-related aspects of archaeology. The presence of a soil in an archaeological site, or in any stratigraphic sequence, as noted above, indicates that there has been a significant period of landscape stability, during which there has been relatively little or no erosion or deposition. In many situations, such as alluvial or eolian depositional environments, deposition can occur in as little as a matter of days, years, or decades. Soils almost invariably take longer to form, usually at least 100 or several hundred years, commonly thousands of years.

The degree of development of a soil or specific pedologic features in a soil can be used as a relative indicator of the amount of time elapsed since deposition of parent material and the beginning of soil formation. In some situations, such data can also provide a rough indication of a soil’s numerical age. In a situation where there are a number of soils, either buried in a stratigraphic sequence or exposed at the surface in an array of surfaces and where the influence of the other factors can be considered negligible, held constant, or otherwise accounted for, a chronosequence can be defined. A chronosequence is where the soils with stronger profile development can be considered older than those that are less developed (Figure 2). Pedologic features that are time-dependent include overall profile morphology as determined by soil indices, horizon and profile thicknesses, contents of illuvial clay, calcium carbonate (CaCO₃), and humus, reddening of the B horizon, alteration or formation of certain clay minerals, and alteration or translocation of certain forms of iron, aluminum, and phosphorous (Birkeland 1999) (Figure 2). In an archaeological site with a chronosequence and also producing time-diagnostic artifacts, radiocarbon ages, or some other form of numerical age control, rates of soil formation can be determined. Such data can
SOILS IN ARCHAEOLOGICAL RESEARCH

then be applied in other sites in similar situations to provide an age estimate for natural or cultural deposits.

The goal of some archaeological research is the reconstruction of the climate during the time that a site or region was occupied, or the detection of climate change over a long period of occupation or relative to abandonment. Soils probably have limited applications for reconstructing climates in archaeological investigations, however (Birkeland 1999; Holliday 2004). In general, soils are not sensitive to discrete or even subtle climate changes which may be culturally significant. Such changes can be more easily detected from plant or animal remains than from soils. Furthermore, climate changes in the Holocene, the time period that most American archaeologists deal with, were often of insufficient magnitude to be detectable in the pedologic record. Furthermore, soil properties related to the time factor are often difficult to separate from those related to climate. Some pedologic properties, such as reddening of the B horizon, can be time- or climate-dependent. Following a climate change, some time is also required for the soil properties to reflect the new climate and, in any event, properties related to the previous climate may persist.

In some situations, however, soils have proven useful in providing general paleoclimatic and paleoenvironmental data. The properties of soils that seem to best reflect those climatic parameters include overall profile morphology, organic matter content, and depth of leaching. The last attribute is a function not only of the amount of CaCO₃ and more soluble salts in the soil but also of soil age and climate (Birkeland 1999; Holliday 2004; Schaetzl and Anderson 2005).

Environmental changes both directly and indirectly linked to climate have produced striking changes in the morphology of some Holocene soils in valleys of the Great Plains, including archaeological contexts. High water tables in the final millennia of the Pleistocene into the early Holocene produced a variety of buried soils. These soils have dark colors, suggestive of high organic matter contents (Figure 3). Gleying immediately below the surface horizons indicate a high water table at some times of the year (Reider 1990; Holliday 2004; Haynes 2008). Soils that overlie them, which probably date to the middle and late Holocene, formed in sandy parent materials with eolian and/or alluvial origins. The latter soils are well drained, with combinations of Bw, Bt, and Bk morphologies depending on their age, all indicative of well-drained conditions likely due to lower water tables, plus filling of the valleys with clastic sediments (Figure 3).

Information about archaeological environments gained from study of factors of organisms, relief, and parent material has only limited application (Birkeland 1999; Holliday 2004). Plant and animal distribution is so closely linked to climate that sorting out the effects of each is often difficult. The best results achieved along these lines have been vegetation reconstructions in restricted areas where the regional climate can be assumed to have been constant over time. Most of this work has focused on studies of the ecotone between forested and unforested (including both prairie and tundra) regions. Alfisols and Spodosols with E horizons will typically form under forested conditions, whereas Mollisols, Alfisols without E horizons, and Inceptisols are common in the nonforested areas. Because the forest boundary fluctuated through time as climate changed, the forest soils left their imprint in the subsequent nonforested soils. Mapping and dating of buried and altered forest and tundra soils also allowed for a reconstruction of shifts in the forest-tundra boundary throughout the Holocene in Canada, and similar investigations have been carried out.
SOILS IN ARCHAEOLOGICAL RESEARCH

along the ecotone of the forests of the Midwestern United States and the prairies of the Great Plains, and in archaeological contexts along the flanks of the Rocky Mountains (Holliday 2004).

Soils also can vary as the parent material or relief changes. These variations can influence soil profile morphology, making the tracing of a particular soil or using a soil as a stratigraphic marker difficult. This problem could be especially significant when working on a large archaeological site or over a larger area such as a drainage basin and where only limited natural or artificial exposures are available. On the timescale of the Holocene, probably one of the most significant influences that parent material can have on a soil is in water movement. Coarser-textured soils allow much faster percolation than do fine-grained soils. Water movement over and through soils is strongly influenced by slope position, that is, along a catena (Figure 3). Soils near the summits of slopes are better drained, but receive less effective moisture (due to runoff) than soils at the foot of slopes. Soils in upslope positions will also tend to have material constantly removed from their surfaces by erosion, whereas the soils at the foot of the slope will tend to accumulate materials eroded from upslope. The morphology of a soil can therefore vary markedly along a slope. This is an important consideration when using soils to date landscapes or when identifying or tracing buried soils. The concept of variation in soil morphology as a function of landscape setting can be quite useful in the reconstruction of landscapes and local environments, however. This principle is particularly applicable to buried soils, which represent buried landscapes (Figure 3).

Site formation process

The physical, biological, and chemical processes associated with pedogenesis can move, alter, and/or destroy archaeological materials and features (Johnson and Watson-Stegner 1990; Thorson 1990; Holliday 2004). For example, the shrinking and swelling that typifies Vertisols

Figure 3 Generalized catenary relationships of soil-stratigraphic units at the Lubbock Lake site, Texas. Source: Reproduced from Holliday 2004. © Oxford University Press: New York.
SOILS IN ARCHAEOLOGICAL RESEARCH

Figure 4  A sketch of a soil profile from Nabke, Guatemala, illustrating a buried A horizon disrupted and distorted by shrink-swell processes in a Vertisol. Based on Jacob 1995.

can result in movement of artifacts (up or down, or both) within the profile, which can disrupt the stratigraphic and cultural contexts of a site (Figure 4). Similarly, freeze-thaw and related processes of solifluction can move artifacts laterally or vertically and in some cases overturn a site. These processes can have subtle but significant impacts. Rooting, tree uprooting, and burrowing by biota can dramatically move artifacts and disrupt stratigraphy. Chemical transformations such as humification, podzolization, oxidation-reduction, and calcification all alter the chemical state of the soil and as a result can destroy anthropogenic plant and animal remains and metals.

Agriculture and Anthrosols

The development of agriculture has had pervasive physical and chemical effects on soils. Certainly no other preindustrial human activity has touched so much of the Earth’s surface as has agriculture. At its most basic level, farming is the replacement of some natural vegetation community with some sort of artificial vegetation, that is, an agroecosystem. This relatively simple and straightforward activity has a host of far-reaching effects on the landscape and on soils. The original plant cover can be partially or completely removed, leaving the ground bare for at least some part of the year and subject to erosion by water or wind. Cultivation loosens the soil, and the hooves of domesticated animals loosen or compact it. Devegetation alters soil water relationships, and can even affect groundwater. Plowing, excavation of irrigation ditches, and construction of terraced fields all physically disturb soils as well. Devegetation, additions of various kinds of plant residues (from burning and cropping), and fertilizers can all alter soil chemistry and nutrient balance. Changes in groundwater conditions can drastically affect the soil-forming environment. An elevated water table as well as irrigation also induces salinization if salts are present. Often, long-term irrigation leads to elevated water table levels. If this water contains measurable amounts of salts (as most irrigation waters do), salinization of surface soils can result.

The unique morphological (macro- and micro-) and chemical characteristics of soils provide an excellent backdrop against which past agricultural activities may be identified and measured (Courty, Goldberg, and Macphail 1989; Holliday 2004). The physical signatures of agriculture in soils are related to the disruption of the lateral continuity of, and vertical gradations between, soil horizons. These disruptions result
largely from plowing and construction of ditches and furrows. Probably the most obvious initial effect of farming is mixing of the upper solum by plowing. This process is widely recognized today in the identification of an “Ap” (plow-zone) horizon. Many taxonomic systems also recognize other soil horizons associated with long-term agriculture, such as the pluggage horizon (long-term manuring), the agric horizon (illuvial materials that have accumulated below a plowed zone), the anthropic horizon (see below) (Soil Survey Staff 1999), and Anthrosols (see below) (IUSS 2006).

At microscopic and molecular scales, effects on soils due to human activity are generally much more subtle than are physical impacts. The former usually require laboratory analyses for identification. Microscopically, the effects of agriculture include evidence for rapid infiltration of coarse-grained illuvial coatings in soils, caused by deforestation, and poorly sorted mineral coatings and infillings of charcoal and soil organic matter (SOM), due to farming. Among agricultural and pre-agricultural groups, anthropogenic chemical additions to the soil came primarily from human refuse and waste, burials, and the products of animal husbandry in barns, pens, and on livestock paths, in addition to more traditional soil fertilizers (Eidt 1984). Metallurgy and later industrialization introduced a much broader spectrum of chemicals and compounds to soils, such as heavy metals and hydrocarbons. The most common elements added to soils by human activity are carbon, nitrogen, phosphorus, and calcium, with lesser amounts of potassium, magnesium, sulfur, copper, and zinc. The most common chemical compound added to soils by humans in agricultural and pre-agricultural societies, and one that is also easily recognizable in the field, is SOM. Discard of organic waste (either in middens or as fertilizer) can add significant amounts of organic matter to the soil surface. Further, additional SOM can be produced and added to the soil by stimulation of soil biota and above-ground biomass subsequent to human activity due to more favorable nutrient conditions often associated with anthropogenic changes. These are notable characteristics of the Anthrosols described below.

Anthropogenic chemicals and chemical compounds can, in theory, be used as indicators of past human activity (Eidt 1984; Holliday 2004). Interpretations must be carefully made, however, because most can be removed from soil more or less readily by leaching, oxidation, reduction, or plant uptake. The proportional relationships of certain ions in soils has also been investigated archaeologically. Soil pH, for example, has some sensitivity to anthropogenic inputs. The concentration of cations in the soil strongly influences pH. Prolonged or more intense occupations tend to release more cations to the soil; therefore, pH tends to be higher under longer or more dense or more intensely occupied sites.

Phosphorus (P), in its common form as phosphate, is the only compound that is stable and generally immobile in soils (Eidt 1984; Holliday and Gartner 2007). Phosphorus added to the soil as organic products or inorganic compounds quickly bonds with iron, aluminum, or calcium ions to form relatively stable chemical compounds of inorganic phosphate minerals. In most soils, removal of these compounds cannot be stimulated by normal oxidation, reduction, or leaching processes, as is true of other elements. Anthropogenic P therefore accumulates at the site of deposition and is thus a sensitive and persistent indicator of human activity (Eidt 1984; Holliday and Gartner 2007). High levels of soil P content is the basis for the “anthropic” horizon (Soil Survey Staff 1999) and part of the defining characteristics of Anthrosols, which are those soils heavily modified by human activity (IUSS 2006).
Sources of anthropogenic phosphorus include human and animal waste, refuse (derived from bone, meat, fish, and plants), burials, and manure. With prolonged occupation, the accumulation of anthropogenic P in soils can become quite large (by orders of magnitude), in comparison to the contents of natural P in soils. Another factor which makes P suitable for geoarchaeological study is that anthropogenic P can persist in the pH range of most soils. Under acidic condition, P combines with iron and aluminum, whereas under basic conditions, P combines with calcium. Consequently, soil P analysis can and is being used successfully in a wide variety of archaeological contexts to indicate human input such as fertilization of fields, and in the study of household activity such as cooking (Eidt 1984; Holliday and Gartner 2007).

Anthrosols vary widely in their physical and chemical characteristics. The most obvious characteristic of Anthrosols is the presence of archaeological debris such as bone and charcoal, that is, Anthrosols tend to be associated with middens. Other physical features of Anthrosols include abrupt, smooth boundaries between horizons or layers, abrupt, laterally discontinuous layers, and dark matrix colors (low value and chroma) extending to depths greater than in natural soils in the area (Figure 5). These greater depths are usually due to artificial upbuilding of the soil surface by processes such as additions of significant amounts of manure in Plaggen soils (see below). Chemical signatures of Anthrosols include higher-than-expected values of organic matter relative to natural soils and, in particular, phosphate (see above). Anthrosols may exhibit some form of pedogenic alteration, although such alteration is relatively minor in many instances due to minimal anthropogenic additions.

The general group of soils identified as Anthrosols can include a wide array of soils, but three types have been described and discussed at some length: Plaggen, Dark Earths, and Terra Preta. Various other kinds of middens may also qualify as Anthrosols. Plaggen soils are most common on the sandy landscapes of The Netherlands, Germany, and Belgium, but similar soils are reported from other parts of northern Europe and Great Britain, Crete, Peru, and New Zealand. In Europe they developed in the Middle Ages as manure, forest litter, heather, and grass sod was removed from stables and strewn on fields. This mixture increased water-holding capacity and also deepened the plowzone, thus minimizing crop failure.

Dark Earths are common in cities throughout much of Europe. They are called Urbic Anthrosols by the IUSS (2006). Dark Earth is a term applied to dark-colored, seemingly homogeneous urban deposits with relatively high organic carbon content, elevated levels of phosphate, and abundant midden debris. In many ways they can be considered anthropogenic sediments rather than soil, but they have undergone
surface weathering and are typically considered a soil. In Britain these soils are linked to late-or post-Roman, Saxon, Viking, Medieval, and perhaps post-Medieval occupation.

The *terra preta do Índio* (“black earth of the Indian”) or simply *Terra Preta* soil of the Amazon Basin and neighboring regions are well-drained soils characterized by the presence of a thick black, or dark grey, topsoil which contains artifacts (Sombroek et al. 2002; Woods et al. 2009) (Figure 5). They are found adjacent to waterways on older terraces and interior uplands. In all settings the dark Terra Preta horizon contrasts strongly with underlying subsoils which are red to yellow Ultisols, Oxisols, Spodosols, and eutrophic Oxisols. The classic black Terra Preta and associated midden debris represent household or near-household trash dumps, but the more ubiquitous dark brown variety (sometimes called *Terra Mulata*), largely devoid of artifacts or other obvious human debris, may represent agricultural soils modified by repeated mulching and burning as means of increasing the soil nutrient status.

**SEE ALSO:** Agriculture; Geoarchaeology; Paleoecology; Paleosols; Soil erosion and conservation; Soil mapping and maps; Soils in geomorphic research; Soils as relative-age dating tools; Soils on slopes: catenas; Soils of urban and human-impacted landscapes

**References**


SOILS IN ARCHAEOLOGICAL RESEARCH


Soils in geomorphic research

Douglas A. Wysocki
Philip J. Schoeneberger
National Soil Survey Center, USA

Soils: the integrator of space, time, and process

Geomorphic research requires the integration of space (land features) and process (physical, chemical, biological) through time. Given that the Earth’s land surface has evolved over decadal to millennial timescales, the question arises as to how one successfully integrates space and time. One approach is to evaluate natural entities that encompass space and time during their genesis, and which show progressive development that varies predictably as a function of time. Soils are one such entity. Soil forms a continuum across the land surface at the interface of atmospheric, biological, and geological systems. Soil horizons and morphological characteristics result from, and are uniquely dependent upon, the processes that operate at this interface. Soils, landforms, and surficial sediments (or rocks), therefore, together comprise three-dimensional (3-D) systems that coevolve through the balance and interaction of physical and chemical weathering versus erosion and deposition processes. It is important to note that the chemical and kinetic energy that drives landscape evolution and soil formation derives from water that infiltrates into or flows over soils or surficial sediments.

Knowledge of soils thus also yields information about physical processes, near-surface hydrology, and age (time) in a given landscape.

Two key geomorphic applications stem from the soil space–time linkage: (i) soil landscape models based on geomorphic concepts that permit meaningful segregation of the soil continuum into natural soil units at various scales (e.g., hillslope, regional, and continental), and (ii) fundamental geomorphic principles (e.g., the geomorphic surface) combined with soil patterns, that allow the interpretation of geomorphic evolution or history for a soil, a landform, or a land area.

Soils: definition, development, and geomorphic function

Before addressing the key geomorphic applications in detail, two terms, “landscape” and “soil,” will be defined. In this entry, landscape is referred to as a group or assemblage of spatially related landforms observable from a given vantage point. Soil is a complex entity that has multiple definitions. For this discussion, soil is herein defined as a natural, 3-D body with definable boundaries that occurs on the land surface and is composed of solids (mineral and organic), liquids (water), and gases. It is characterized by horizons that are distinguishable from the initial material from which it formed, and that have formed from physical, biological, and chemical processes in the surface environment (Soil Survey Staff 2014). Most soils support or have the capacity to support higher plants.
Simonson (1959) explained soil formation through four interacting suites of processes: additions, removals, translocations, and transformations. Note that the four process suites are applicable to the formation of soil horizons, such as the dispersion and downward transfer of clay from an A or E horizon to form a clay-rich Bt (argillic) horizon, and to a soil body as a whole. The same suite of processes driven largely by geomorphic forces applies on landscape scales as well. On a holistic basis, additions, removals, translocations, and transformations create and modify landforms, sediments, and soils from microscopic to landscape scales. Consider, for example, material eroded from the flank of a hillslope. It is transferred and added at the base of a slope as colluvium or alluvium. This sediment either thickens existing soils or becomes parent material for another cycle of soil formation. Upslope, where this removal has occurred, the soil is thinned, and thus fresh parent material below is brought into the soil profile and begins to weather. This example shows how the soils on the two landscape positions are linked.

Another eloquent means to explain soil patterns is the factor (or factorial) approach, which was initially developed by Russian soil scientists in the 1800s, and later espoused by Hans Jenny (1941), a soil scientist at the University of California, Berkeley. This model describes soil as a function of five interacting state factors: climate, biological influences (organisms), topography or relief, parent material, and time. It is often stated as $S = f(Cl, o, t, p, t)$. This model has been widely applied to explain soil formation and distribution in a qualitative or semi quantitative manner. The model can be solved quantitatively, but this requires precise input values for the five factors. Further, the five factors are interdependent and change through time. Stated in another way, the model purports that a soil or its distribution can be explained after all its factors are adequately known. The difficulty lies in deciphering how the factors have interacted through time, that is, landscape processes modify topography and parent material as soils form, and climate and vegetation change through time. Soil development itself, which proceeds with time, can affect the factors; for example, subsurface pans (duripan, fragipan) impact water-flow, and have greater erosion resistance than adjacent horizons. These characteristics in turn can be a controlling influence on local relief or topography. Soil development in this case alters topography, and vice versa.

The important attributes in this entry’s soil definition are their spatial (3-D) form, the presence of horizons, and the differentiation of soil from its initial (or underlying) material. Consider the soil profile in Figure 1; the two-dimensional (2-D) exposure reveals several readily observable soil horizons and sediment relationships. Plants
grow in this soil. The upper boundary between soil and atmosphere is readily apparent. The lower boundary is recognizable, although it may require scientific judgment to make a precise distinction between the solum (genetically formed soil horizons that include all the A and B horizons) and the parent material or substrate. Less altered sediment is visible at depth. The boundary between the solum and the substrate in Figure 1 is gradual. The profile has a distinct A, Bw, Bk, 2BC, 2C horizon sequence. The upper horizons formed in loess, whereas the lower ones are in alluvium, which explains why gravel is present in the 2BC and 2C horizons, but is absent in the upper horizons.

A soil profile/pedon is an observational unit. Pedons and profiles have sufficient size to display the complete horizon sequence, often down to and including the parent material. Soil profiles or pedons are defined as being of a convenient size to describe, sample, and classify the soil. Soil horizons visible in a 2-D profile obviously also extend in three dimensions. The pedon is the 3-D conceptual equivalent of a soil profile. Nonetheless, both the profile and the pedon lack one crucial attribute. They lack naturally defined lateral boundaries, which are important aspects of our soil definition. No evident lateral boundaries occur in the profile presented in Figure 1. Lateral boundaries are not readily definable at this observational scale.

Geomorphic processes operate on a spatial scale larger than a pedon or profile. A profile or pedon does not have naturally defined lateral boundaries. A geomorphic entity should have a common set of properties that occupies a naturally defined land area. We must consider an entity larger than a profile to observe lateral soil change or boundaries. To that end, Figure 2 displays a broader landscape view of the soil in Figure 1. The right foreground shows the A, Bw, Bk, 2BC, and 2C horizons of the soil profile shown in Figure 1. These horizons, particularly the Bk, are readily traceable across the exposure. Note that the Bk thins and thickens in a repeating pattern. Field observation shows that the Bk horizon is thin under topographic highs, but it is also thin under topographic lows. The Bk horizon, as well as other horizons, changes thickness predictably at prescribed landscape locations. It is important to note that the changes in soil horizon thicknesses are not random. When soil properties correlate to landscape position, the delineation of soil areas (bodies) that possess relatively more uniform characteristics than the soil continuum at large is possible. This is the essence of soil mapping and soil survey. Soil surveys are maps that depict such soil delineations, and these are the most useful soil individuals for geomorphic research and studies. Soil surveyors delimit boundaries between soils by mapping out the observable and predictable differences in soil morphology including horizon type and thickness, soil color, texture, and structure across a landscape. In an ideal world, these soil boundaries correspond to measurable and observable properties of the surface landscape. For example, field observation confirms that the soil pattern in Figure 2 has formed due to the influence of repeating, shallow, closed depressions on the land surface. The depressions are scoured into Miocene-age alluvium of the Ash Hollow formation of the Ogallala group below. The depressions likely formed initially as deflation basins (caused by wind erosion). Pleistocene (Peoria) and Holocene (Bignell) loess and slopewash subsequently filled the depressions. The Bk horizons are thickest at the depression margins because evaporative discharge from the surface draws carbonate-rich groundwater up from below and concentrates CaCO₃ in the soil at these locations.

As can be seen, the soils on this landscape record a part of their geomorphic history. This
example illustrates the fact that to decipher the geomorphic history of a soil landscape requires systematic field observation and the application of some fundamental concepts of geomorphology and soil science.

The soil landscape

The state factor approach of Jenny (1941) and the process-systems approach of Simonson (1959) are important conceptual models for understanding soil formation and distribution. Neither approach, however, defines functional entities in the soil–landscape continuum. Therefore, to do that, this entry now turns to the catena concept.

First proposed by Milne (1935), the catena is a fundamental geomorphic concept that explains soil differences at the hillslope scale. A catena is a repeating sequence of soils that occurs from the top of the hillslope to the adjacent valley bottom. The catena concept includes both surficial stratigraphy and internal hillslope structure or lithology. Two main types of catenas exist (Figure 3). The first (Figure 3a) involves soils underlain by a single parent material or by similar bedrock. Despite this initially uniform parent material, the hillslope soils commonly display sequential and predictable change along the slope, usually resulting in spatially distinct soil types. Red-brown soils occur on the hilltop, where conditions are highly oxidizing.
Figure 3  Soil landscape relationships, illustrative of the catena concept.

Downslope, the soils become wetter and progressively more reduced, thereby displaying increasing amounts of patchy iron depletions and segregations. The soils also display textural variation across the hillslope. On the hilltop, soils (in this example) have formed directly in weathered granite residuum. Soils below the hilltop have formed in transported sediment that becomes finer textured downslope. Milne (1935) attributed the type of repeating soil sequence that embodies catenas to systematic patterns of (i) internal water-flow, (ii) chemical and physical weathering of bedrock, (iii) lateral sediment transport, and (iv) translocation of materials.
across and beneath the soil surface. In sum, the hillslope and soil sequence have evolved coevally through interrelated processes. Note that both the hillslope and the catena are geomorphic entities. The pattern of soils observable across the hillslope, therefore, is both a soil development sequence and a geomorphic record of landscape evolution.

In the second type of catena (Figure 3b), the hillslope contains several different kinds of bedrock, or parent material, rather than having been formed all in one uniform type of sediment. Nonetheless, the hillslope also displays a similar sequence of distinct soils. The soils display similar red-to-gray color sequencing downslope, and show downslope sediment sorting. Variations in bedrock downslope, however, lead to increased soil complexity, as the soil textural variation results from both lateral transport and changes in bedrock lithology. (Lateral sediment transport occurs in this catena, as in all catenas.) Because of lateral sediment transport down the slope, the surficial sediments form a drape on the landscape that is not exactly coincident with the underlying rock strata.

As can be seen from these two examples, the catena concept includes both surficial stratigraphy and internal hillslope structure or lithology. The catena is both a soil landscape and a geomorphic model or system that links slope position and parent material to soils. Lateral transfer and sorting contribute to the sequential soil variation across the hillslope. Erosion and sedimentation processes, driven by relief and water movement, redistribute sediment across the hillslope, creating subtle lateral differences in soils and soil parent materials. Recall that both parent material and topography are factors in Jenny’s (1941) model of soil formation. These examples of the sequential changes in soil morphology across a landscape show how soils and landscape evolution on hillslopes function across both time and space.

### Hillslope positions

Using the catena concept, correlations between soil characteristics and slope position can be used to link soils to specific landscape positions. This association involves a conceptual model that explains and communicates soil distribution and variation across space. To operationalize such a model, however, we must first define the various slope positions from slope geometry, independent of soil observation. Slope positions are distinguishable by shape, particularly in profile view and hillslope location, and both affect water-flow (surface and subsurface), erosion, and sediment transport and deposition.

Five slope elements are commonly defined (Ruhe 1975). From the top of the slope to the bottom, they are summit, shoulder, backslope, footslope, and toeslope (Figure 4). Summits are the nearly level, uppermost slope position. Summit soils generally exhibit the greatest degree of soil development in a given landscape, unless the summit is narrow and convex, in which case erosion may dominate over soil formation. Shoulders have a convex shape and occur just downslope from the summit. The shoulder slope is subject to greater erosion than the summit. Soils here tend to be similar to summit soils, but have much thinner profiles, especially if a caprock exists. Below the shoulder is the steepest, and generally linear, slope position, the backslope. Surface runoff, erosion, and sediment transport commonly reach peak values at this position. As a result, backslope soils display less development (thinner or absent horizons) and are generally, but not always, shallower to bedrock or parent material than are summit or shoulder soils. The slope position below the backslope, the footslope, has a concave slope curvature and a much lower slope gradient. Concordant with decreased slope gradient is a decrease in sediment transport capacity and thus surface
runoff. Thus, sediment is usually deposited here, either as colluvium or as slopewash. Note that this position delimits the transition from erosion to deposition on the hillslope. Footslope soils tend to be wetter and display lesser development, mainly because they accrete upward over time due to additions of sediment to their surfaces. This stands in contrast to higher slope positions, where erosion dominates over depositional processes. The footslope merges downslope with the toeslope, which has a slightly concave or even linear slope curvature. Low-energy alluvial processes such as slopewash or overbank deposition from adjacent channels dominate on toeslopes. Toeslope soils therefore accrete upward, are wetter, display limited horizonation, and are often younger than are soils on summit and shoulder positions. Buried soils are common on footslopes and toeslopes. These five terms are a simple, eloquent set of terms to communicate slope position and soil occurrence.

The overall slope curvature from summit to toeslope provides important visual clues about the internal structure of a hillslope. Hillslopes that are strongly convex from the summit through backslope profile occur in selected geomorphic settings, for example, where erosion-resistant bedrock underlies the hillslopes. Soils on these slopes tend to be shallow and usually display limited horizon development. Similar convex slopes, however, commonly occur in erodible material that can maintain a steep angle of repose. Slopes developed on thick loess are prime examples. Hillslopes with slightly concave backslopes and footslopes generally occur where the underlying material is relatively erodible, such as weakly resistant bedrock or unconsolidated sediments. Soils on these slopes, in contrast to bedrock-controlled slopes, are thicker and display greater horizon development. The concave portion of a slope, that is, footslopes and toeslopes, display a decreasing slope gradient. Both potential energy and overland flow velocity decrease at the slope inflection from a linear to a concave slope, leading to accumulations of colluvium or sediment derived from slopewash. Soils on these slopes often form by cumulic processes and usually have thick, poorly developed

Figure 4  Soil landscape relationships according to hillslope position. Modified from Ruhe (1975).
horizons, when compared to soils on higher hillslope positions.

Another soil-forming factor that changes across a hillslope is time, or soil age. Figure 5 shows hillslope modification by backwearing through four stages. At stage 1, the summit is stable and older than the narrow shoulder and backslope, while the footslope and toeslope receive sediment eroded mainly from the backslope and therefore are equivalent in age. In stage 2, the hillslope has retreated, but the summit remains stable and the shoulder is broader. At stage 3, the encroaching slopes have narrowed the divide and the hilltop is formed by coalesced shoulders. Erosion has consumed the summit. The hilltop is no longer a geomorphically stable position. The age of the hilltop slope now equates to that of the other slope positions and to the time (period) of backwearing. Stage 4 shows the narrow remnant of the original hillslope. These hillslope age relationships are important concepts that will be discussed in more detail below.

**Geomorphic components**

Erosion that drives mass transfer across slopes results primarily from hydraulic (e.g., raindrop impact, dispersion) and fluvial (i.e., detachment and sediment transport by running water) processes. Thus, water movement over the soil surface and the way a hillslope sheds or concentrates water as surface flow are significant geomorphic concerns. Recognizing this, Ruhe and Walker (1968) developed their concept of a 3-D hillslope. Put another way, they took the catena concept into the third dimension by describing and examining slopes in plan view, as well as along the profile (the 2-D catena approach). The components of their 3-D hillslope model are the head slope, side slope, and nose slope (Figure 6). Recent work by Schoeneberger and Wysocki (2012) added two additional components: the interfluve and the base slope (Figure 6).

The interfluve, a 3-D equivalent to a summit, is the uppermost landscape area; it is a stable component of the hillslope system, being less subject to erosion than many lower components. Wide interfluves are summits and contain the most well-developed soils. Below the interfluve, a hillslope undergoes both backwearing and headward stream incision, commonly focused along overland flow pathways. Head slopes have convergent flow paths (when considered in a downslope direction), side slopes have parallel flow paths, and nose slopes have divergent flow paths. Flow direction, therefore, concentrates water in the head slope, especially its lower part, which in turn is more likely to initiate rill and channel formation. Soils here are commonly the wettest of any slope position. Below the nose slope, side slope, and head slope is the base slope (Figure 6), where sediment deposition dominates over erosion and sediment transport. Base slopes are roughly equivalent to foot- and toeslopes. A colluvial apron or wedge of sediment commonly forms on the base slope. This transported material can range from coarse debris to finer sediments.
sorted by slopewash (nonchannel, overland flow) processes. On long, high-relief hillslopes or mountain slopes, soils on side slopes and base slopes often form in colluvial sediments that are in slow, gravity-driven transport down the slope. Slopewash occurs here, but is not the dominant process. The base slope does not typically include stratified alluvium associated with channel deposition. Distal base slope sediments, however, may grade into or interfinger with alluvial deposits.

Unless masked by changes in parent material, some degree of lateral sorting downslope is present on most slopes, even long, mountainous ones. Soil profiles on side slopes thus vary greatly in thickness, depending on the rate of erosion (high or low, respectively), the magnitude of the slope gradient, and the extent to which the bedrock is resistant to weathering and erosion. In many landscapes, the traditional assumption of prevalence of shallow soils over hard rock on hillsides and mountainsides is overly simplistic.

**Hydrology, hillslopes, and soils**

The catena concept encompasses both lateral transport of water and sediment, and soil water/wetness conditions along a slope. These are all contributing processes to soil formation, hence, landscape, and slope evolution. Overland or surface flow is the main driver of lateral sediment transport along slopes. Soil water results from infiltration at the soil surface. Soil formation depends upon such infiltration and percolation processes, as soil water is the matrix for chemical and geochemical transformations. Moreover, water-flow within and through soils transports soluble and colloidal constituents, leading to the development of soil horizons. Water-flow on level terrain primarily has a vertical, downward path into and through soil horizons. On hillslopes, even those with low slope gradient, however, lateral soil water-flow, also called throughflow, can occur. Throughflow is dominantly downslope, flowing beneath and roughly parallel to the surface. Figure 7 depicts three such
flow types and directions that occur on a head slope. Water as surface runoff (Hortonian overland flow) converges toward the thalweg (lowest portion of the slope, along the length of a valley).

Water that infiltrates has an initially vertical flow path. Vertical hydraulic conductivity through a soil generally decreases with depth, however, because of increasing bulk density, changes in size and kind of soil structure, and increasing clay content. Because of such changes with depth, soil horizons become preferential conduits for lateral flow (throughflow) at depth, along slopes. In a head slope, subsurface flow converges toward the thalweg, although its flow rate is much slower than overland flow, and water can take days to weeks to reach an equivalent point. This differential results in an accumulation of subsurface water in the lower parts of the head slope, temporarily forming a saturated zone there (Figure 7). The boundary of this saturated zone migrates slowly upslope during wet periods. It is important to note that throughflow may occur under both saturated and unsaturated conditions. Saturated flow occurs under positive pore pressure. Unsaturated flow is a function of matric potential.

Figure 7 Surface and subsurface soil hydrology in a head slope. The A, B, and C horizons of the soil are shown.
Historically, research on soil erosion has de-emphasized the contribution of such internal water-flow to soil formation and landscape evolution. Soil hydrology studies in recent years (e.g., Arndt and Richardson 1989 is this recent) have reintroduced and reinforced the importance of throughflow and have emphasized the need for a 3-D approach that accounts for lateral, divergent, and convergent throughflow pathways.

Soil landscape hydrology

As should be obvious by now, in the soil landscape, a number of key soil and geomorphic attributes control or greatly influence water-flow. Some have been mentioned already, such as parent material, slope and stratigraphic relationships (Figure 3), and soil horizonation and land surface shape (Figure 7). Water-flow can be predominantly downward, lateral, or even upward (toward the surface), depending upon the landscape and how it affects flow dynamics. The influence of the land surface on water-flow, and thus soil occurrence, also depends on climate, specifically precipitation quantity and intensity.

Consider three soil landscapes (Figure 8), all underlain by medium-to fine-grained glacial deposits and all containing closed depressions. These landscapes occur across a humid to dry subhumid climatic gradient (400–875 ml mean annual precipitation). In humid central Iowa (Figure 8a), water recharge from precipitation occurs at topographic highs and flows downward and laterally toward topographic lows, discharging at depression edges and concentrating CaCO₃ there due to evaporation from the soil surface. Water-flow at the basin center is downward, toward the water table. Due to the water-flow system, soil B horizons differ markedly along the slope. At the topographic highs, soils have a Bw horizon, at the depression edge a Bkg forms, and in depression centers soils have Bg horizons. This soil landscape relationship exists at most medium to large depressions, regardless of elevation and location within the landscape.

In eastern North Dakota (Figure 8b), under a moist subhumid climate, recharge occurs under
SOILS IN GEOMORPHIC RESEARCH

depressions that are situated in high landscape positions. Within depressions in intermediate positions, water-flow is dominantly lateral, resulting in the formation of a flow through wetland. Depressions that are at a low relative position are groundwater discharge sites. This flow pattern results in distinctly different soils at the various depressions. In the upland depressions, soils have a Btg (clay increase and gleying) horizon, due to their long-standing wet conditions but downward-flowing water (and clay). The intermediate depressional soils have Bg horizons (gleying). Soils in the low-lying depressions have Bkgz (carbonate, gleying, salts) horizons because precipitation and percolation are insufficient to remove soluble salts from the soil; therefore, both carbonates and salts accumulate at the discharge sites.

Under the dry, subhumid climate of southern Saskatchewan (Figure 8c), the landscape exhibits a still different macroscale flow pattern. The flow path under depressions is more localized because the regional water table is at greater depths. Regardless of landscape position, all depressions are recharge sites where percolating water flows mainly downward. Evaporative discharge dominates on adjacent topographic highs because of the limited percolation depth. Thus, soils in the depressions have Btg or Bwg horizons, whereas soils on the topographic highs have Bkz horizons.

Soil: a geomorphic entity

The catena concept establishes the link between space and time. Soils, parent materials, and surface age change along a catena (hillslope) due to sediment transport, erosion, and internal water-flow. The kinetic energy that drives landscape (geomorphic) evolution is expended in sediment transport and hillslope backwearing via slopewash and fluvial erosion. Water that enters the soil surface, in contrast, drives physical and chemical soil development across the slope. In low-relief landscapes, erosional modification is minimal and internal water-flow is the dominant driver of soil formation and variation. Landscapes such as lake plains and dissected coastal plains are examples. Erosion and slope development are limited in these landscapes but soils change in a predictable manner. The input energy is expended mainly as chemical and geochemical weathering, rather than as kinetic transport. Soluble and colloidal transfers occur within the landscape during soil formation, that is, in pedogenesis. The soil change across the landscapes represents geomorphic evolution. The geomorphic history of the landscape is recorded in the soils, rather than as erosional modifications.

An example of such a landscape is the Coastal Plain of North Carolina. Figure 9 displays a stratigraphic and soil cross-section of this region. Major streams are confined to inset valleys, and the adjacent upland is a broad, undissected divide. The general stratigraphic relationship is that coarser-grained, permeable deposits overlie finer-grained, less permeable materials. Stream downcutting and headward incision occurs only along slopes directly adjacent to the main valleys.

Despite the low relief, soils display distinct differences based upon distance from the low relief, divide centers. The major soils at the divide margin are within the Orangeburg and Norfolk series, which have yellowish-red to yellowish Bt horizons, respectively. The seasonal high water table is deep. Toward the divide center, however, the seasonal water table comes increasingly closer to the surface. The pedological result is a sequence of soils – Goldsboro, Lynchburg, Rains, and Pantego – that become progressively more poorly drained and have grayer Bt horizons as the divide center is approached. As can be seen here, landscape hydrology is the dominant
soil-forming influence. Water table depths are a function of both stratigraphy (strata that differ in hydraulic conductivity) and the limited local relief (long flow paths and low hydraulic gradient).

The soil sequence shown in Figure 9 extends from the divide center to a stream valley. The divide is stable; little or no erosion has modified the divide. Surface age is the same across the divide. The soils across the divide, given the material age, have resulted from hundreds of thousands to perhaps a million years of landscape evolution. The soils themselves are geomorphic products that occupy predictable locations. Following parent material emplacement, slope processes have had minimal effect on landscape evolution. In these respects, the soil sequence does not fully meet the classic catena definition.

The earlier discussion illustrates that hillslope profile terms provide key descriptive terms for defining soil position. The question arises as to what hillslope terms apply to the distinct soils in

**Figure 9.** Soil landscape and hydrology relationships in the low-relief, middle coastal plain North Carolina, USA. Modified from Daniels and Gamble (1967).

**Figure 9.** The divide is a hillslope summit. Additional soil landscape terms are needed to describe the divide positions on such landscapes.

### Soil landscape age assessment

Soil-forming processes occur at different rates and intensities through time. As a result, soils that have similar amounts of horizon development may not necessarily be the same age, and vice versa. To unravel the time–space association inherent in soil landscapes requires a conceptual means to assess time or soil age, independent of soil development. The hillslope stability and age relationships alluded to (above) underscore a set of fundamental geomorphic concepts or principles that allow for the relative age assessment of soils and surfaces across the landscape.

The first such concept this entry will discuss is the principle of superposition (Daniels, Gamble, and Cady 1971), which specifies the deposition sequence of sedimentary rocks, unconsolidated
SOILS IN GEOMORPHIC RESEARCH

sedimentary deposits, and extrusive igneous rocks. Sediment can only be deposited atop pre-existing materials. Therefore, younger sediments invariably overlie older deposits. Observe the strata depicted in Figure 10a. By the principle of superposition, Beds 1 through 4 represent a time sequence: Bed 2 is younger than Bed 1, Bed 3 is younger than Bed 2, and so on. Even if the absolute age differences of the beds are unknown, their relative ages are inferable by using this principle. The principle of superposition is not complex, but field application requires observations at key locations and consistent identification of units, and often is operationalized using data taken from drill cores.

Bed 4 in Figure 10a demonstrates another important detail. Note that it is topographically lower than most of Bed 3. Bed 4, however, stratigraphically overlies Bed 3 at the footslope. This stratigraphic context determines their relative age relationship. As illustrated here, there is often a limited area within which one can make field observations for age determinations in this manner. In Figure 10b, Bed 5 overlies both Bed 1 and Bed 2 and thus must be younger than either bed. Bed 5 does not stratigraphically overlie Bed 4 at any location. This raises the question as to whether it is younger or older than Bed 4. Further observation indicates that Bed 5 is inset below Bed 4. Therefore, Bed 4 confined the depositional processes responsible for Bed 5’s emplacement. An erosion event or cycle must have cut into Bed 4 before Bed 5 was deposited. This cross-cutting relationship, in and of itself, establishes that Bed 4 is older than Bed 5, and represents another important geomorphic concept – the principle of cross-cutting relationships. An event or feature that cuts into a deposit must be younger than that deposit.

An erosional (downcutting) event appears to have produced a hillslope that descends to

Bed 5. This raises the question as to the age of the hillslope. Slope processes would shed sediment onto the top of Bed 5, as would alluvial deposition from up valley. The top of Bed 5 and the hillslope are equivalent in age and both are younger than Bed 4. The hillslope and top of Bed 5 together represent a definable land surface area formed during a given time period, primarily by erosional and depositional processes (e.g., stream incision, hillslope retreat, pedimentation). This land area constitutes a geomorphic surface, which is a mappable land

Figure 10 Geomorphic surfaces in a soil landscape, illustrating various principles of geomorphology. See text for details.
SOILS IN GEOMORPHIC RESEARCH

area of uniform age. Note that the geomorphic surface includes two components – one erosional the other depositional. Nevertheless, these two components are time equivalent. An erosion cycle or event cross-cuts the pre-existing soils, slopes, and substrates. The erosion event and the geomorphic surface produced are younger than any entity (soil, surface, or depositional unit) that is cross-cut. In Figure 10a, soil 1 on Geomorphic Surface A is older than soil 2 or 3, which are equivalent in age and occur on the erosional and depositional components of Surface B, respectively. Despite being the same age, soil 2 and soil 3 differ because the substrates and components are contrasting.

Figure 10b depicts a second downcutting cycle and subsequent deposition of younger sediment below Surface B. A new geomorphic surface (C) forms; it has both an erosional and a depositional component. The landscape now has three geomorphic surfaces, each of which is underlain by different materials and each of which contains distinct soils.

The landscape depicted in Figure 10 corresponds to a set of adjacent stream terraces. As illustrated here, landscape evolution commonly proceeds episodically through time, which creates a series of stepped or faceted geomorphic surfaces along interfluves. As you ascend this landscape, the geomorphic surfaces increase in age. This case illustrates the principle of ascendancy and descendancy, which states that an erosion surface is younger than the youngest deposit or surface that it cuts across or truncates. By this principle, Surface C is the youngest, Surface B is intermediate in age, and Surface A is the oldest. From this example, one can also infer that an erosional surface is the same age as the alluvium to which it descends. This example illustrates the principle of descendancy. Based upon these principles, several time relationships of erosion surfaces can be summarized. Any erosion surface (i) is younger than the youngest material, landform, or soil it bevels, (ii) is younger than any erosion surface or erosional remnants that stand above it, (iii) is the same age as the alluvium to which it descends, (iv) is the same age or older than any deposits that lie on it, (v) is older than deposits in valleys below it, and (vi) is older than any erosion surface that occurs below it. We can infer the relative age of a hillslope (erosional surface) in the field by determining the alluvial level that it grades downward to in a smooth, concave profile.

The temporal and spatial relationships portrayed in Figure 10 apply to landscape evolution by fluvial erosion in general. Now consider the development of an arid, tectonic basin, as exists in the Basin and Range of the western United States (Figure 11). Crustal extension and faulting that began in the early Miocene drives long-term basin evolution. The geomorphic/stratigraphic/soil record is a result of successive fluvial downcutting cycles. In the initial downcutting, the oldest alluvial fans form along the valley flank. During a period of stability, erosion continues to modify the fan surfaces. As shown in Figure 11, the oldest geomorphic surface (Surface 1) eventually comes to consist of rounded erosional (fan) remnants termed “ballenas,” which are underlain by alluvial fan sediments. Surface 2, below the ballenas, consists of a fan piedmont with rounded finger ridges along the distal front. This surface is more complete than Surface 1, but is nonetheless dissected to some degree. Surface 3, which is cut into Surface 2, consists of several laterally adjacent landforms. Inset fans occur in the upper valley margin. The inset fans grade to a coalesced set of fan skirts that merge down valley with an alluvial flat. Each of these landforms results from lateral sorting and alluvial deposition through a specific time. Despite the lateral changes in sediment, distinctly different landforms, and suites of soils

15
that occur across Surface 3, it is age equivalent throughout its extent.

The question arises as to where on this landscape the “oldest” soil is found, in terms of relative age. The answer is Surface 3. Because hillslope processes have rounded the ballena summits (Figure 11a), the best-preserved and oldest soils will occur along the central ballena ridge, on summit positions. Surface 2 is less modified by secondary erosion than Surface 3; soils on Surface 2 are younger, but may be better preserved because of less secondary surface modification. Soils on Surface 3 are the youngest in this landscape. On this surface, soil characteristics are a function of sediment sorting. Soil patterns here have resulted from the integration of sediments, landforms, age, and, even in this environment, hydrology, as poorly drained soils may occur in the alluvial flat due to runon and throughflow. Note that the age sequence of geomorphic surfaces follows the principle of ascendency, that is, the geomorphic surfaces shown are a sequence of inset depositional surfaces of decreasing age.

Figure 11b depicts an erosional cycle caused by axial stream downcutting that has resulted in an incised floodplain with tributary valleys graded to it. This cycle yields geomorphic Surface 4, a depositional surface that again has distinct landforms, sediments, and soils. Its erosional component, which is not shown, is the adjacent mountain front. No scale is shown in Figure 11; the basin size can range from hundreds to thousands of square kilometers. The geomorphic concepts and the application of them remain the same regardless of basin size (Gile and Grossman 1979).

From the above discussion, it is emphasized that geomorphic surfaces are a spatial and chronological framework that applies to landscape evolution and soil distribution, and at a number of scales. A given surface may have multiple soils, but they will have a common duration of soil development. Soils will also display increasing (or at least, predictable) development from the youngest to the oldest surface in a landscape. Field identification of geomorphic surfaces is based on stratigraphic and geomorphic principles. Recurring soil patterns linked to a geomorphic surface may be used to delineate that surface in a similar landscape.

**Soil: a geomorphic record**

As stated at the outset, soils integrate across space and time. Figure 12 displays a dual parent material soil that contains a petrocalcic horizon. Field examination and description provide the following horizon sequence: A, Bk, 2Btk, 2Bkm. This soil’s horizon sequence, including the two substrates, forms a record of the site’s geomorphic history. From the horizon sequence, one can infer that a soil with a petrocalcic horizon formed in older alluvium. The presence of the petrocalcic horizon is noteworthy, as it requires a significant period to form (Gile, Peterson, and Grossman 1966). Younger alluvium later buried the soil. New soil horizons then formed in the fresh alluvium. The result is a composite or “two-storied” soil with a lithologic discontinuity separating the two parent materials. Soil formation is time transgressive and the soil in Figure 12, therefore, has a rich geomorphic history that can be inferred from the horizon sequence and stratigraphic units present.

Let us examine more closely the geomorphic history of the pedon in Figure 12. This pedon is on an upper fan piedmont similar to that shown in Figure 11. Detailed field study (Gile and Grossman 1979) has formally recognized and named the geomorphic surface on which the petrocalcic horizon exits as Jornada 1, with an approximate age of 500,000 years. The younger
deposit is a Holocene (≈7 ka) aged deposit, the Organ alluvium. The soil and geomorphic record here includes two deposition cycles, 500,000 years of landscape evolution, at least two periods of soil formation (likely more), and the associated climatic shifts that occurred during that time interval. One can derive much from the soil profile alone, but placing the profile into its proper stratigraphic, geomorphic, and hydrologic context gives a more complete perspective of its geomorphic history. In turn, subsequent soil observations based on the geomorphic and time context inherent in this profile have greater merit because the degree of soil formation is linked to the substrates, ages, and geomorphic position.

Summary

Soils serve as the matrix for biological and geochemical processes at or near the Earth’s surface. The cycling of key elements, as well as important trace metals, occurs within them. Further, soils form an absorbent but porous and permeable layer that mediates water infiltration and flow within and across landscapes. Soils provide a short-term water reservoir that sustains plant and animal life in terrestrial ecosystems for days to weeks. They are the medium we depend upon for much of the world’s food and fiber, and for filtering our waste products.

The quantification of soil formation through time and the associated links to geomorphic processes outlined above permit both a spatial and a temporal understanding of soils on landscapes. This entry has articulated several important principles in its discussion, which include the catena, hillslope and geomorphic components, soil landscape hydrology, and geomorphic surfaces.

By use of a soil geomorphic approach, an understanding is gained of how landscapes evolve through time. Where and why specific soils...
occur within a landscape can also be determined. This knowledge has some important applications. First, soils defined and recognized as geomorphic entities are a functional unit that has a common process, physical properties, and ecosystem function. To manage the soil landscape for long-term sustainability, functional entities at geomorphic and other scales must be identified and understood. Geomorphic entities define landscape scale units. Second, a geomorphic approach provides the chronological framework to understand soil and landscape change through time. This knowledge allows for the reasonable prediction of future soil conditions that will occur due to climate change and/or human influence. The scientific and societal value of such knowledge is important because our sustained existence depends on our soils.

**SEE ALSO:** Soil development: conceptual and theoretical models; Soil development: numerical indices; Soil mass balance; Soils as relative-age dating tools; Soils on slopes: catenas; Soils and weathering
References


Further reading


Soils of cold and permafrost-affected landscapes

Charles Tarnocai
Agriculture and Agri-Food Canada

James G. Bockheim
University of Wisconsin–Madison, USA

Cold soils occupy approximately one-third of the world’s landmass. Because most of this area is dominated by permafrost, the genesis of these soils is also strongly affected by cryogenic processes. As a result, these soils contain large amounts of ice and they sequester and store large amounts of frozen organic materials, rendering them a huge storehouse of carbon in particular. Because of these unique properties, these soils could release this stored carbon if the climate were to warm, further enhancing atmospheric warming. Therefore, these soils do and will play an increasingly important role in both the global carbon cycle and global climate change.

In addition, they support both the vast arctic tundra and the northern circumpolar boreal forest, which are still in their natural states.

“Cold soils” is a loosely defined term. Some authors define cold soils as those soils having mean soil temperatures lower than 8°C at 50 cm depth (Rieger 1983). Others consider cold soils to be all soils in the permafrost region. These would then include soils in the southern part of this region, in which seasonal frost extends below the topsoil and stays frozen for several months during the winter. These soils are not frozen during the rest of the year. Regardless of which definitions are used, three major groups of soils occur in the Cold Soil Zone. The first group are the permafrost-affected soils. These soils are dominated by cryogenic processes and have permafrost in their control section (0–2 m). The second group of soils are those having deep permafrost. They are also affected by cryogenic processes. Although they have no permafrost within their control section, permafrost does occur at depth and may have existed closer to the surface during colder periods in the past. These soils are extremely dynamic and can change into permafrost-affected soils, depending on the shift in their thermal regime, often because of environmental or climate changes. The third group of soils are the seasonally frozen soils. These soils contain no permafrost, even below the control section. These soils are seasonally frozen below the topsoil during the winter, and they thaw during the following spring. Regardless, they are also affected by seasonal cryogenic processes (e.g., frost heave and ice lensing). Many of these soils contained permafrost immediately following recession of continental ice sheets.

Most of the cold soils occurring in the arctic, subarctic, and boreal soil climate regions are characterized by low temperatures and moderate amounts of precipitation. The arctic soil climate region has a mean annual soil temperature (MAST) of ≤−7°C. Mean daily temperatures above 0°C occur only during the short summers. The precipitation totals are generally low and precipitation occurs mostly in the form of snow. All of the arctic soil climate region is underlain by continuous permafrost. The subarctic soil climate region has a MAST between −7°C and...
<2°C. This area receives a moderate amount of precipitation, but most of it is in the form of snow. Permafrost is widespread in this region. Although the boreal soil climate region has somewhat warmer summer temperatures, the winters are still long and cold. The MAST for this region is between 2°C and 8°C. Although the southern portion of the boreal soil climate region is free of permafrost, the northern part contains discontinuous or sporadic permafrost (Figure 1). The Antarctic soil climate region has a MAST of <−10°C, and the mean annual precipitation is less than 50 mm per year (water equivalent).

Soil parent material in these regions is highly variable and is one of the main contributors to the wide variety of soils. Frost-shattered rocks and stony glacial deposits dominate large parts of these regions. However, fine textured sediments occur in lacustrine basins and along the seashores. Organic (peat) deposits are common, especially in the southern Arctic and the subarctic and boreal regions. The vegetation in the Arctic is partially vegetated tundra in the north and continuously vegetated tundra in the south. Open coniferous forests dominate the subarctic. The arctic tree line forms a natural boundary between the Arctic and subarctic regions. Lastly, a closed canopy conifer forest to aspen parkland and grasslands are associated with the boreal region. In Antarctica, vegetation is either absent or limited to isolated patches of moss, lichens, and algae. Higher plants (grasses) exist along the western Antarctic Peninsula.

Soil development and properties

A variety of soil-forming processes have influenced the soils in these areas. The development of seasonally frozen soils is dominated by soil-forming processes such as podzolization, brunification, luvification, and gleysation. Podzolized soils are usually sandy and have a B horizon that is enriched in Fe, Al, and organic carbon compounds. Many of these soils are classified as Spodosols. Brunification results in soils with red and red-brown B horizons, due to in situ weathering, and minimal translocation of soluble materials into these horizons. These are weakly developed soils that are commonly classified as Inceptisols. Luvification is the translocation of clay from the upper, lighter colored, horizon into the B horizon (argillic horizon). These soils classify as Alfisols, or if weakly developed, Inceptisols. Wetness causes gleysic processes and reduced conditions in many lowland soils. These soils, in most cases, are poorly drained with a high water table and so are classified within aquic suborders, or in Canada, as Gleysols.

Permafrost-affected soils and, to some degree, the soils with deep permafrost, are dominated by cryogenic processes. The mechanisms driving these cryogenic processes are the increase in volume when water freezes, turning it into ice, and the migration of water from warm to cold regions in the soil, feeding and growing the ice bodies. These processes increase the ice volume in the soil, resulting in differential frost heave, a form of mixing called cryoturbation. This process is eventually responsible for formation of patterned ground and cryogenic microstructures, sorting of coarse fragments, and burial of organic matter.

The most common soil features resulting from cryoturbation are irregular or broken soil horizons, organic matter incorporated into lower soil horizons, oriented stones, and cryogenic structures in forms such as platy, blocky, granular, amorphous, or massive soil structure. Massive structures develop as a result of cryostatic pressure caused by desiccation during freeze-back. At this time, two freezing fronts, one from the surface and the other from the permafrost table, compress the soil material, resulting in desiccation,
SOILS OF COLD AND PERMAFROST-AFFECTED LANDSCAPES

Permafrost distribution

- Continuous (91–100%)
- Discontinuous (51–90%)
- Sporadic (10–50%)
- Isolated patches (<10%)
- Treeline

Figure 1  Northern circumpolar permafrost map (Tarnocai et al. 2009) derived from information from Brown et al. (1997). Reproduced by permission of John Wiley & Sons, Ltd.

Compaction, and movement of materials within the soils. Cryodesiccation is the predominant process in hyperarid soils of Antarctica.

Another important characteristic of permafrost-affected soils is that most of the soil water occurs in the form of ice lenses, ice crystals, ice wedges,
SOILS OF COLD AND PERMAFROST-AFFECTED LANDSCAPES

and massive ground ice. Soil texture is one of the factors controlling ice build-up. Because they hold more water in general, fine-textured soils generally contain a higher content of ice in the form of ice crystals, ice lenses, and thick ground ice. This high ice content is due to the fine (clayey) soil materials that provide an avenue for the translocation of water from warm to cold. On the other hand, regardless of the texture or type of materials, ice wedges can occur in both frozen mineral and organic soils.

These processes and the resulting pedogenic features are very common in permafrost-affected soils of the northern circumpolar region, but they are less common in soils of Antarctica. Due to the dry, cold climate, soils in the Antarctic region are commonly associated with dry permafrost. In most cases, these soils lack ice-cemented permafrost because of the very low moisture contents, and few cryogenic features are found in them. For example, in the Transantarctic Mountain areas the soil moisture content is only <1–5% during the short summer. In the Antarctic Peninsula, however, the soils are slightly moister and may have some organic matter accumulations on the surface.

Unique thermal characteristics separate permafrost-affected soils from all other soils. In summer, the upper boundary of the permafrost is usually present 30–60 cm below the surface. Due to this perennially frozen layer, these soils have a steep vertical temperature gradient in summer. The temperature gradient can also be large if these soils are associated with patterned ground. For example, in permafrost-affected soils associated with earth hummocks in the Mackenzie River Valley, the soil temperature at the apex of the hummock can decrease from 14°C at the surface to 0°C at 60 cm depth (permafrost table) during the summer months. Horizontally, soil temperatures at the same depth under the interhummock depression, about 1 m away from the center, can be 4°C cooler.

Permafrost-affected soils underlain by a permafrost layer have an active layer that thaws during the summer and freezes in winter. As mentioned above, the active layer is generally between 30 and 60 cm thick, depending on the thickness of the surface organic layer, the soil texture, and the vegetation cover. The permafrost underlying the active layer is also very dynamic, especially in permafrost-affected mineral soils. Changes in vegetation cover because of forest and tundra fires, disturbance of the soil surface, and climate warming, can all lead to increases in the depth of thaw. As a result, the permafrost table (top of the permafrost) moves to greater depths. For example, after forest fires, the active layer depth could increase threefold in thickness, in part because the darker soil surface absorbs more solar energy. During this period, cryoturbated soil features and other cryogenic features can develop in this thicker active layer. Re-establishment of vegetation and the surface organic layer, which usually takes a century or more, results in a gradual re-establishment of the permafrost table to its original, shallower depth. Therefore, the cryogenic properties developed in the former, deeper active layer are preserved and are now frozen in the near-surface permafrost.

Peatlands associated with Histels (permanently frozen organic soils) and Histosols (unfrozen organic soils), which are common soils in the northern circumpolar regions (arctic, subarctic, and boreal), cover large areas in both Canada and Russia. These peatlands develop in wet areas as raw biomass (leaves, stems, mosses, and other plant materials) that is deposited on the surface. These organic materials are then restricted by decomposition due to low temperatures, saturated conditions, and low pH. This gradual build-up process produces 2–2.5 m-thick peat
deposits that developed during the past several thousand years.

Peatlands in the southern boreal region are composed of various types of fens, bogs, and swamps. These peatlands are not affected by permafrost. Further north, the occurrence of permafrost-affected peatlands has gradually increased. Most of the peatlands in the northern boreal and subarctic regions are perennially frozen peat plateaus, polygonal peat plateaus, and palsa. In the arctic region, all of the peatlands are composed of perennially frozen lowland polygons.

**Soil classification**

Until the 1970s, permafrost-affected soils in the various soil classifications were handled as northern extensions of the southern soils. Soil scientists from the United States and Russia divided these soils on the basis of soil-forming processes such as gleyzation, podsolization, brunification, salinization, and peat accumulation. Cryogenic processes and features were not considered to be part of the genesis of these soils. Canadian soil scientists were the first to recognize cryogenic processes as the dominant soil-forming process in permafrost-affected soils and, based on this concept, the Cryosol soil order was developed and incorporated into the 1978 edition of the Canadian System of Soil Classification. This new Canadian concept relating to permafrost-affected soils slowly gained acceptance and, by the early 1990s, the Gelisol order was developed and incorporated into the US System of Soil Taxonomy. Shortly after the development of the Gelisol order, a similar development took place when the Cryosol soil reference group was established to classify permafrost-affected soils in the World Reference Base for Soil Resources (WRB).

In the Canadian system, the Cryosol soil order includes both mineral and organic soils that have permafrost within 1–2 m depth (Soil Classification Working Group 1998). This soil order was further subdivided into great groups of Turbic (cryoturbated soils), Static (non-cryoturbated soils), and Organic (soils developed on >40 cm of peat) Cryosols. These great groups were further subdivided based on the soil development (chemistry of the B horizon, gleying, surface organic layers, and lack of soil development). The Organic Cryosol great group is subdivided on the basis of the thickness and rate of decomposition of the organic horizons.

The Gelisol order in the Soil Taxonomy also contains mineral and organic permafrost-affected soils, all with permafrost occurring within 1–2 m depth. Gelisols are subdivided into three suborders: Turbels (cryoturbated mineral soils), Orthels (non-cryoturbated mineral soils), and Histels (organic soils). Further subdivisions of these suborders are based on the presence of morphologies that may indicate processes such as gleying, podzolization, and salinization, and on the presence of ground ice. The Histel suborder is subdivided into great groups, based on the degree of decomposition of the organic materials. Glacic and Ruptic-Histic subgroups are associated with Gelisols containing massive ground ice (Glacic), while Ruptic-Histic subgroups are associated with variable thicknesses of organic matter in the subsoil. Anhydrous (meaning extremely dry) great groups were added to accommodate hyperarid Antarctic soils.

The Cryosol soil reference group in the WRB system includes only the mineral permafrost-affected soils that have permafrost within 0–1 m depth (IUSS Working Group WRB 2015). Permafrost-affected organic soils are not included in the WRB Cryosol group of
soils. They are included in the lower level of cryic units, in the Histosol soil reference group. Soils with permafrost within 1–2 m depth are included as lower case Gelic units in most of the soil reference groups.

The Russian soil classification system follows the genetic approach. This classification recognizes one, relatively small, group of Cryozems in the high-level group of soils associated with cryogenic soil features, but permafrost, as a thermal condition, does not play any role in the Russian soil classification system.

The classification of seasonally frozen soils was developed during the mid-twentieth century and the last version was published in Soil Taxonomy (Soil Survey Staff 1999). Most of the seasonally frozen soils in the cold soils region are one of the following.

- Inceptisols: These soils have a weakly developed subsoil and a light-colored surface soil horizon.
- Spodosols: These are mineral soils that are dominated by podzolization processes. They are commonly associated with an upper eluvial (leached) E horizon and lower illuvial B horizon.
- Histosols: These soils developed on organic materials (peat) are derived mainly from plants and contain 17% or more organic carbon.
- Entisols: These are mineral soils that have minimal pedogenic alterations, implying that their chemical and physical properties are similar to the associated parent materials.
- Mollisols: These are mineral soils with thick, dark surface horizons resulting from the incorporation of organic carbon into the mineral soil, mainly by roots. These soils are moderately rich in basic cations and clay.

Distribution of cold and permafrost-affected soils

In the early and mid-twentieth century, national soil maps were prepared in order to determine the distribution of these soils. These early soil maps used local classification systems and various mapping approaches. In Canada, permafrost-affected soils were mapped as cryic versions of nonpermafrost soils. Most of them were mapped as Cryic Regosols. Similarly, on the Food and Agriculture Organisation (FAO) Global Soil Map, permafrost-affected soils were mapped as gelic versions of the major soil groups – for example, Gelic Histosols, Gelic Gleysols, Gelic Cambisols – but most of the permafrost-affected soils were mapped as Gelic Regosols (IUSS Working Group WRB 2015). The soil map for the entire area of the former USSR was prepared by Fridland (2015), Naumov (1993), and Uspanov (1976). Following the Russian concept, only the soils strongly affected by cryogenic processes were mapped as permafrost-affected soils. On the Russian soil maps, however, these soils contained only a very small portion of the soils affected by permafrost.

The concept of Cryosols was developed in Canada during the 1970s and the Gelisol Order was established later in the United States. During the 1980s, national soil maps were prepared using the new concept in order to recognize the cryogenic properties in permafrost-affected soils, which distinguish these soils from those found in nonpermafrost regions. The availability of more robust geographic information system (GIS) mapping capabilities during the late 1990s led to the development of soil databases that covered the entire northern circumpolar soil region (NCSR). In order to develop these large, unified databases for the entire NCSR, all of the various national classifications needed to be integrated so that the soil maps could be
converted to the US System of Soil Taxonomy. Using these new, unified soil databases, soil maps (Figure 2) and other derivative maps were prepared for the entire NCSR. Based on these soil databases, Turbels were shown to be the dominant soils in the circumpolar permafrost region, covering approximately $6454 \times 10^3$ km$^2$.

The second most common soils were Histels, which cover approximately $2714 \times 10^3$ km$^2$, followed by Orthels, which cover approximately $1051 \times 10^3$ km$^2$.

Seasonally frozen soils are patchy in the discontinuous permafrost region but they cover the entire soil area below the southern boundary of the permafrost region. The most common soils in the nonpermafrost area are Inceptisols, Spodosols, Histosols, Entisols, and Mollisols.

**Global significance**

The global importance of soils in the polar regions was only recently recognized, when the role they play in global climate change was discovered. It is now widely accepted that climate warming has been, and will continue to be, much greater in both the northern and the southern circumpolar regions than in other parts of the globe. This warming is already in effect (Turner, Overland, and Walsh 2007; Barrie and Kochtubajda 2008), although the magnitude is variable from region to region.

Perennially frozen soils in the northern circumpolar region contain the single largest component of the terrestrial carbon pool. Tarnocai et al. (2009) estimated that the permafrost-affected soils in the northern circumpolar region contain approximately $1672 \text{ Gt}$ of organic carbon with approximately 88% occurring in permafrost-affected soils and deposits. This $1672 \text{ Gt}$ of organic carbon accounts for approximately 50% of the total below-ground global organic soil carbon. This total is greater than the amount of carbon stored in the atmosphere or the biosphere. Only the oceans store a greater amount of carbon than do these soils, and the oceanic carbon is much more difficult to release to the atmosphere.

Very little information is available for fluxes in both unfrozen and perennially frozen cold soils. Unfrozen cold mineral soils in the boreal forest of North America sequester about 60–100 g C per m$^2$ per year. The slow rate of decomposition in these soils results in relatively rapid organic matter accumulation. This organic matter, however, is affected by wildfires that can completely remove it. The only deep carbon that survives is that which is derived mainly from roots. Therefore, the carbon pools in these unfrozen mineral soils are low, and the turnover time for this carbon is only 100–1600 years.

Carbon dynamics in cryoturbated mineral soils (Turbels) are very different. Cryoturbation translocates the organic matter deposited on the soil surface into deeper soil horizons. This cryoturbated organic matter is then unaffected by wildfires and, due to the cold soil temperatures, can be stored for thousands of years. It has been estimated that, in Canada, about 9 million tons of organic carbon are added annually to these types of cryoturbated mineral soils.

Vegetation on Histels and Histosols annually deposits raw organic material (litter) on the soil surface. About 25% of this deposited material will decompose during the following years, but the rest will be incorporated into the organic soils as peat. Researchers working in the Mackenzie River Valley of Canada found that, in the sporadic permafrost zone, the mean carbon
Figure 2  This map shows the distribution of permafrost-affected soils (Turbels, Orthels, and Histels), deep permafrost soils, and seasonally frozen soils in the northern circumpolar area (Tarnocai et al. 2009). Reproduced by permission of John Wiley & Sons, Ltd.
accumulation rates for unfrozen bogs (Histosols) over the past 100 years were 88.6 ± 4.4 g C per m² per year. In the discontinuous permafrost zone, they found that the mean carbon accumulation rate during the past 120 years was 13.31 ± 2.20 g C per m² per year. It was also estimated that organic soils in the boreal region accumulate approximately 9.8 Mt of organic carbon annually. Based on these estimates, the Turbels, Histels, and Histosols together in Canada accumulate (sequester) approximately >18.8 Mt of organic carbon annually. Thus, these soils play a very important role not only for storing large amounts of carbon, but also for removing large amounts of carbon annually from the atmosphere and for safely storing it for thousands of years.

Effects of climate change

Temperature and ice content are probably the most important parameters for determining the vulnerability of permafrost-affected soils to climate change. The temperatures of the permafrost layers have increased in most permafrost regions during the past several decades. This increase has been most noticeable at the southern limit of the discontinuous permafrost zone, where the temperature of the permafrost is just slightly below 0°C (warm permafrost). The increase in ground temperatures here has resulted in permafrost degradation in the form of thermokarst development, deeper active layers, and surface subsidence. The surface subsidence is especially critical because this indicates that there is thawing of the ice-rich, near-surface permafrost. Thawing of the ice-rich permafrost not only affects the surrounding environment, but has already been influencing the stability of the human-constructed infrastructures, as well as the livelihood of the people living in the north.

The thawing of ground ice due to climate warming results in severe landscape changes. These changes come in the form of detachment slides, thermokarsting, retrogressive flow slides, and hydrological changes. In addition, the perennally frozen soils contain a huge amount of water in the form of ice. When this ice melts, the release of this additional water will affect the hydrology from local to global scales.

The thawing of frozen soils results in landscape changes that are generally of local affect. Because these soils contain huge amounts of organic carbon, the release of this carbon could affect, and probably will affect, global climate. It has been estimated that the amount of carbon released from the thawing of perennially frozen mineral and organic soils by 2100 will be between 50 Pg C and 250 Pg C. Estimates vary as to how much of this carbon will be released in the form of CO₂ and CH₄ and how much these greenhouse gases will increase future warming. It is clear, however, that removal of large amounts of soil carbon from permafrost-affected soils may also reduce or stop carbon sequestration in these soils. So, the impact of climate change has a double effect on these soils not only by releasing the stored carbon, but also by stopping the sequestering of carbon from the atmosphere.

SEE ALSO: Antarctica; Climate change and permafrost; Pedoturbation; Permafrost: definition and extent; Soil taxonomy and soil classification; Soils and the carbon cycle

References

SOILS OF COLD AND PERMAFROST-AFFECTED LANDSCAPES


Further reading

Soils of desert landscapes

Leslie D. McFadden

University of New Mexico, USA

The world’s warm arid and semiarid regions occupy at least 40% of the Earth’s land surface. Soils here strongly reflect the limited precipitation and the generally sparse vegetation found. Soil scientists characterize such areas as having aridic or torric soil moisture regimes with soil water deficits several months long (Soil Survey Staff 1999). Well-developed soils that form in this moisture regime are most commonly classified as Aridisols. However, Entisols (weakly developed soils) dominate many desert landscapes, and Inceptisols, Mollisols, Alfisols, and Vertisols may also occur in marginal areas (e.g., cooler “high deserts” or slightly moister areas).

Globally, desert areas are increasing, due in part to anthropogenic activities that produce “desertification.” These trends are accentuated by global warming.

Water availability is the most critical, ecologically significant factor in desert biomes. The types and spatial distribution of soils strongly influence vegetation in desert landscapes, mainly through their influence on rates and magnitude of infiltration, water percolation depths, soil water storage, and evaporation. Understanding how future climate changes will affect the Earth’s deserts requires an appreciation of the nature of desert soils.

Classification and morphological characteristics of desert soils

Desert soils are characterized by a vertical sequence of horizons that typically exhibit master A, B, and C horizons (Figure 1) (Birke-land 1999; Schaetzl and Anderson 2015). Many properties of A and B soil horizons of desert soils resemble those of soils of more humid regions. A fundamental requirement in the classification of all Aridisols, however, is that they must presently experience an aridic soil moisture regime. Also, unlike soils of the Entisol order that lack subsurface “diagnostic horizons,” soils in the Aridisol order must have certain diagnostic B horizons which display specific properties that serve as the basis for soil classification. For example, the presence of translocated, or “illuvial,” clay is signified in descriptions of soils by adding the lower case “t” after B (i.e., Bt). To qualify as an argillic horizon, however, certain values (e.g., clay content, presence of translocated clay, horizon thickness) must be met. Argillic horizons are common to soils of both arid and humid regions. It is the key horizon to the Aridisol suborder Argids (Figure 1); the “Arg” prefix is derived from “argillic,” and the “id” suffix is derived from “aridic.” This same scheme is generally used to derive other soil order and suborder names.

The calcic horizon, which is key to the Aridisol suborder Calcids (Figure 1), is defined mainly on the basis of accumulations of calcite (CaCO₃) (hereafter referred to as soil carbonate). The
SOILS OF DESERT LANDSCAPES

Figure 1 A desert soil formed in mid-Holocene-aged alluvial deposits in the Chihuahuan Desert, southern New Mexico (Gile, Hawley, and Grossman 1981). The presence of an argillic (Bt) horizon in this soil, as well as the aridic moisture regime of this area, classifies it as an “Argid.” Note the white coatings of soil carbonate on gravel clasts in the Bk (calcic) horizon, which exhibits Stage II morphology.
Scale on tape graduated in decimeters. Photo by L.D. McFadden.

lower case “q” (e.g., Bq). The cambic horizon, which is key to the Aridisol suborder Cambids, is a subsurface horizon that shows some, usually weak, evidence of obliteration of parent material structure or other inherited features, color change, and/or loss of parent material calcite. B horizons qualify as Cambic if they have experienced some measurable amount of soil development but fail to qualify as another type of diagnostic B horizon. Development of these morphological features is indicated by “w” (e.g., Bw). Other diagnostic soil horizons in Aridisols include gypsic and salic horizons, which are dominated by, respectively, sulfate and chloride salts. Soils with these horizons are most typically found in the most arid and warmest deserts, including “hyperarid” regions where very little or no rain falls.

Low contents of organic matter (e.g., humus) and pale colors are typical of most desert soils, with the exception of commonly thin biological surface crusts that form in some desert settings. These characteristics reflect the commonly sparse vegetation mantle in deserts and low production of organic matter by biota. Sustained high soil temperatures cause rapid oxidation and loss of much of the organic matter that ends up in the soil. Many thin, pale A horizons of desert qualify as Ochric horizons. In some semiarid grasslands, however, where organic matter production is higher, soils have A horizons that qualify as mollic epipedons.

In drier and “stony” desert regions, many A horizons contain little organic matter, but also typically have abundant vesicular pores throughout (Figure 2). Other typical attributes of such horizons include blocky and/or prismatic structure, subordinate platy structure, fine, often silty textures, and accumulations of calcium carbonate and clay, often present as oriented laminae. Commonly referred to as a “vesicular A (Av)
horizon,” this horizon is recognized only informally. Av horizons are usually found immediately below desert pavements, which are typically a one-pebble and/or cobble thick surface layer of generally closely packed, subangular to sub-rounded gravels (McFadden 2013). The surfaces of desert pavement gravels are commonly thinly coated (tens to a hundred microns) with a dark layer referred to as rock varnish. The varnish is composed mainly of iron and manganese oxides, clay minerals, and silica that accumulate, and hence thicken and darken, with time.

Some desert soils contain indurated subsurface diagnostic horizons such as duripans or petrocalcic horizons (Figure 3). Some of the latter exhibit zones of laminated, nearly pure calcite. In some cases, these subsurface horizons are more than a meter thick. One of the best examples is the impressively thick petrocalcic horizon associated with an ancient soil on the Mormon Mesa, in southern Nevada (Figure 3). This soil represents several overlapping, buried petrocalcic horizons that have been “welded” together during the past 5 Ma (Brock and Buck 2009). Other indurated horizons include those composed of cemented gypsum (petrogypsic horizon) or soluble salts (petrosalic horizon).

**Formation of desert soils**

The accumulation of soil carbonate, gypsum, or more soluble salts in desert soils mainly reflects the relatively shallow depth of leaching associated with limited precipitation. Periodically high soil temperatures also promote relatively high evapotranspiration values and losses of soil water. Consequently, a substantial amount of dissolved anions in soil water, such as carbonate and sulfate, are not flushed completely from the soil, forcing them to precipitate as minerals such as calcite and gypsum in the B horizon (Figure 1). Also, the limited supply of soil water and low concentration of organic matter in the
SOILS OF DESERT LANDSCAPES

A horizon diminish the overall magnitude of chemical weathering in desert soils.

CO₂ in the soil atmosphere strongly influences carbonate solubility and, therefore, affects the depth at which it accumulates. Equation (1) shows the factors in the desert soil environment that drive the solubility and precipitation of calcite where $K_{\text{calc}}$ is the calcite solubility product, $K_1$, $K_{\text{calc}}$, and $K_{\text{CO}_2}$ are dissociation constants in the carbonate system, and $\gamma_{\text{Ca}^{2+}} + \gamma_{\text{HCO}_3^-}$ are the activity coefficients of $\text{Ca}^{2+}$ and $\text{HCO}_3^-$. This equation is valid for the system $\text{CaCO}_3$-$\text{H}_2\text{O}$-$\text{CO}_2$, assuming that activities are equal to concentrations and $\text{pH} < 9$.

$$m^3\text{Ca}^{2+} = \frac{(p\text{CO}_2 K_1 K_{\text{calc}} K_{\text{CO}_2})}{(4 K_2 \gamma_{\text{Ca}^{2+}} + \gamma^2_{\text{HCO}_3^-})}$$

Concentrations of CO₂ in desert soils are often elevated well above atmospheric values, leading to increased calcite solubility. The elevated levels of soil CO₂ result mainly from CO₂ production by plant roots, and/or microbial respiration (soil respiration) associated with soil organic matter. In deserts that lack vascular plants and, hence, have relatively limited soil microbial populations, calcium carbonate dissolution and transport are inhibited. As a result, more soluble salts that only require small amounts of soil water for their dissolution and transport, such as gypsum, dominate B horizon development.

Studies of the micromorphology, mineralogy, and chemical composition of desert soils have shed light on other important soil-forming processes in deserts. For example, because much of the clay in desert soil B horizons commonly is composed of smectite, a clay mineral with shrink–swell tendencies, some desert soils expand and contract as moisture contents change. The small size and surface charge of smectite clay particles favor their translocation in suspension, and eventual accumulation in illuvial clay-rich B horizons. The ability of smectite to absorb large amounts of water is conducive to large volume increases and soil expansion during the rainy season, followed by contraction when the soil dries. These wetting–drying (expansion–contraction or swelling–shrinking) cycles, coupled with progressive increases in soil B horizon clay content, ultimately favor development of strongly expressed soil structure (e.g., subangular blocky or prismatic soil peds) as soil clay contents increase.

Smectite clays also play a key role in the formation of vesicular A horizons (McFadden et al. 1998). During sustained rainfall and infiltration events, some of the percolating water is absorbed by smectite, causing the minerals to expand. Then, when the soil dries, dehydration of smectite causes contraction and the development of an extensive system of polygonal cracks and, commonly, prismatic structure. Subsequent rainfall causes initially rapid infiltration through these cracks. Thus, the percolating water moves deeply into the soil, and because this water carries suspended clay, fine particles and solutes are translocated below the vesicular A horizon, potentially also promoting the development of Bt and Btk horizons. Periodic leaching also dissolves and removes some of the calcium carbonate near the margins of the blocks or prisms of the vesicular A horizon. The fine texture and strong matric potential of the vesicular horizon, however, draws some soil water into ped interiors. This process also likely drives the formation of vesicles, entrapping air in the horizon because the ped margins are sealed. Soil air expands as the soil warms and dries, which may cause expansion of voids, and hence, vesicle formation. The advance of the wetting front may also increase soil pore pressure in the sealed horizon, and has been suggested as another mechanism for vesicle formation (Dietze et al. 2012). With the passage of time, the vesicular horizon can thicken to 10 cm or more, and
SOILS OF DESERT LANDSCAPES

becomes increasingly clay and silt enriched and locally reddened, exhibits development of different stages of vesical formation and destruction, and is characterized by the accumulation of illuvial clay.

The development of a vesicular A horizon over initially highly permeable parent materials strongly affects soil development that follows (McFadden et al. 1998). For example, if a source of calcite is present in surface materials (e.g., limestone), the hydrologic behavior of the A horizon enables solutes derived by partial dissolution of these calcareous materials to precipitate as layers of calcite on the sides of large surface clasts that lie immediately below the soil surface. Upwardly diffusing CO₂ in the soil prevents precipitation of soil carbonate on the bottoms of embedded surface clasts that exist at slightly greater depths. This pattern of soil carbonate formation results in the development of “carbonate collars” on the near-surface clasts (McFadden et al. 1998). In this way, near-surface soil carbonate can accumulate well above the usually observed depth of carbonate accumulation in the B horizons of desert soils.

Many important discoveries concerning desert soil-forming processes have also emerged from studies of soil chronosequences. A chronosequence is a suite of soils that differ primarily in age (Jenny 1941). In essence, soil chronosequence studies involve soils associated with series of geomorphic surfaces of different ages, but with uniform parent materials and in a similar climate. Many are located on alluvial terraces, alluvial fans, eolian sheet dunes, loess sheets, and topographically smooth lava flows. Chronosequences enable investigations of how soils develop over thousands to hundreds of thousands of years.

One of the best and well-known examples of chronosequence-based research in a desert landscape is that of Leland Gile and his colleagues, associated with the Desert Project (Gile, Hawley, and Grossman 1981; McFadden, 2013). Conducted mainly from the 1960s through the early 1980s in southern New Mexico, their research showed that on alluvial fan and terrace surfaces ranging from late Holocene through Middle Pleistocene in age (0–790 ka), soils on older surfaces become predictably thicker and morphologically better developed. Increasing B horizon clay contents and reddening also are observed, but the most strongly time-dependent soil property is soil carbonate accumulation. A very small amount of soil carbonate is visible as early as only a few centuries, occurring mainly as coatings on the surface of gravel clasts or as small filaments in fine-textured parent materials (Figure 1). Carbonate accumulations increase predictably in older soils, and in most cases a petrocalcic (Bkm) horizon forms after approximately 100,000 years. The horizon continues to thicken for several hundred thousand years more. In addition to the complete cementation, or “plugging,” of soil materials by soil carbonate, older petrocalcic horizons also exhibit zones of nearly pure, laminated soil carbonate, typically at the top of the plugged parts of the horizon (Figure 4).

Using these observations, Desert Project scientists defined four time-dependent stages of soil carbonate morphology (Stages I–IV) (Figure 4). Subsequent studies suggested the addition of two more stages, V and VI, for soils that exhibit advanced soil carbonate morphologies. These last two stages extended the time range covered by this model to several million years (Figure 4). The use of this terminology has been nearly universally adopted by scientists in studies that involve carbonate-enriched desert soils.

Desert Project scientists (and others) have concluded that most soil carbonate reflects the incorporation and pedogenic alteration/translocation of eolian (wind-deposited) dust, rich in carbonate minerals, that becomes incorporated
into the soil via percolating water. Initially, this conclusion was inferred because of the rapid rate of soil carbonate accumulation in soils with parent materials with no, or only small amounts of, calcium-bearing minerals. These parent materials also showed little evidence of chemical weathering of the Ca-bearing minerals. The presence of substantial calcite in dust collected by dust traps since these studies were done has provided additional evidence supporting the role of dust in the formation of carbonate-rich desert soils. Calcium dissolved in rainwater provides
another “external” source of soil carbonate, but because the dissolved Ca is derived by dissolution of Ca-bearing dust in the troposphere, it also ultimately represents a dust source.

Recent research continues to demonstrate the eolian origin of many materials (Reynolds et al. 2006; McFadden 2013). Continent-scale, long-term, regional studies of dust flux and composition in the western United States, using specially designed dust traps, have shown that dust influx and composition readily account for the rapid accumulation of carbonates, sulfates, silicates, and silica in soils – the building blocks of calcic, gypsic, argillic, and duripan horizons. Other studies have used modeling and chemical mass balance approaches, as well as data on soil mineralogy and the isotopic composition of dust and soil carbonate to arrive at the same conclusion – dust is an important driver of desert soil genesis (Derry and Chadwick 2007).

Because the rate and magnitude of soil carbonate accumulation in many desert soils are primarily controlled by dust composition and influx, high influx rates can accelerate calcic horizon and Bt horizon development. In desert environments, the accumulation of dust-derived soil materials in soils is such a common and widespread process because of deflation and eolian sediment transport from extensive, barren surfaces (e.g., playas), sparse vegetation, frequently dry soil conditions, and seasonally high wind velocities. Suspended dust may be transported to high altitudes and carried great distances. It is delivered to the soils either by rainfall that scrubs the dust from the atmosphere or by entrapment that is favored by the effects of rough surfaces of gravelly alluvial fans, volcanic flows, or sparsely vegetated shrublands. Infiltrating water translocates this entrapped dust into the soil.

Studies of strongly dust-influenced soils associated with desert pavements in the Cima Volcanic Field of the Mojave Desert of southern California show that their development is markedly different from that proscribed by traditional models of soil profile development in humid regions (McFadden 2013) (Figure 5a). According to that model, A and B horizons develop mainly through biogeochemical weathering that entails loss of solutes by leaching. Progressive surface lowering of mineral soils is the inevitable consequence of this net loss of mass, despite some mass gains due to organic

![Figure 5](a) Classical soil development sequence for a desert environment. White areas in the Bk horizon represent segregated, illuvial soil carbonate. Modified from McFadden (2013) with permission of the Geological Society of America. (b) Similar diagram to Figure 5a, but in this instance the soil has formed by time-dependent accretion and inflationary profile (AIP) strain-dominated development. Light-brown, irregular shapes on soil surface could represent either rubble formed by initial physical weathering of bedrock (R), or coarse clasts in the parent material. Note how the soil surface is stable in Figure 5a, but grows upward here, due to accretion of dust. Modified from McFadden (2013) with permission of the Geological Society of America.
matter accumulation in O and/or A horizons, and gains in voids by bioturbation. Soils that develop beneath desert pavements, including the vesicular A and Bt horizons, however, form mainly through the process of *cumulative* soil development. This process is driven by the gradual additions, or accretion, of eolian materials. This material accumulates in the A and B horizons beneath the pavement and is variously altered by subsequent soil-forming processes (e.g., weathering of calcium-bearing minerals in dust and subsequent precipitation as soil carbonate). The accretion causes soil inflation driven by the cyclic generation of new soil void space via hydration–contraction processes. *Accretionary and inflationary profile (AIP)* development (Figure 5b) ultimately drives the “upward growth” desert pavements (McFadden 2013).

Maintaining a balance between dust entrapment and soil incorporation is necessary for continued AIP and the concomitant upward growth of an evolving desert pavement. This type of desert pavement formation profoundly contrasts with other models that entail losses of fine material and surface lowering thorough deflation and/or surface erosion, leading to concentration of a coarse gravel lag (the pavement layer).

Soil chronosequence studies have shown that AIP and desert pavement formation are favored on certain types of landforms (e.g., alluvial fans), mainly because they provide weathering–resistant rocks (Figure 6). AIP is also an important mode of soil profile development in semiarid regions where desert pavements cannot form and vesicular horizons are weakly developed or absent. Increased plant density, however, also promotes rates of dust entrapment necessary to favor AIP.

Besides hydration–contraction, other physical weathering processes play key roles in weathering and soil development in deserts. One is salt weathering, a process responsible for the disintegration of large subsurface gravel clasts. It has also been suggested as an important process driving the weathering of large surface clasts and ultimately the formation of desert pavements. Recent studies, however, have indicated that another weathering process is responsible for the initial fracturing of surface rocks. They show that stresses produced by nonuniform rock surface insolation are responsible for crack formation in surface rocks (McFadden et al. 2005; Eppes et al. 2010). The impact of insolation is demonstrated by a statistically significant pattern of generally north–south crack orientations, which can only reflect the systematic heating and cooling of the rock as it is subjected to directionally changing, diurnal insolation cycles (Figure 7). The subsequent introduction of salts or other eolian materials into these cracks may then favor additional crack extension and widening.

*Figure 6* Extensive areas of desert pavement on a piedmont in the western Andes Mountains of central Peru. Note the pale-colored boulders on the unpaved road across the pavement. Photo by L.D. McFadden.
Recent research that demonstrates the preferred orientations of cracks on surface rocks in Martian deserts, and modeling studies of asteroids and regolith on asteroids, have shown that insolation weathering is also an important weathering process on these surfaces as well (Abernathy, Eppes, and Willis 2013).

**Desert soils and plant communities**

With few exceptions, most deserts are vegetated, although the vegetative cover is typically sparse and often strongly spatially aggregated. The characteristics of the vegetation communities of deserts reflect climatic regimes ranging from hyperarid climates that support virtually no vascular plants, to semiarid climates that support denser communities such as piñon-juniper woodlands or shortgrass prairie communities. The character of desert biota reflects the overall aridity and unpredictability of precipitation, leading to low biotic productivity and, surprisingly, relatively high species diversity. Desert plants also must be adapted to climatic extremes. Obviously, water availability is the most critical factor limiting ecosystem processes in deserts. Consequently, soil properties that strongly influence water infiltration, available water-holding capacity, and overall soil water balance strongly influence most desert plants. For example, common desert shrubs of the Mojave and Sonoran Deserts, in the United States, such as creosotebush (*Larrea tridentata*) (one of the most spatially extensive and long-lived plants), are uncommon on soils with well-developed Bt horizons, because other plants such as white bursage (*Ambrosia dumosa*) have shallower root systems adapted to quickly utilizing soil water commonly held in the upper part of the profile. Another example is shown by the evolution of desert pavements and soils. Many vascular plants
SOILS OF DESERT LANDSCAPES

can recruit and become more easily established on recently deposited, well-drained gravelly alluvium. With increasing age, however, as surficial weathering and hydrologic processes progressively concentrate increasingly darkened surfaces of clasts above a vesicular horizon, the combination of high surface temperatures, high evapotranspiration rates, and fine soil textures inhibit the colonization and persistence of perennial vascular plants on the desert pavement soils. Variability in pavement packing, clast size, surface relief, and local burial of pavements by younger alluvial or eolian deposits may, however, influence the development of some hydrologically significant soil properties, locally increasing runoff and infiltration, thereby enabling plant recruitment and persistence (McAuliffe, Hamerlynck, and Eppes 2007).

Studies of desert soil carbonate show that its stable isotopic composition strongly reflects the photosynthetic pathway utilized by plants growing there (Sharp 2007). This relationship exists because of photosynthesis-driven isotopic fractionation in plant tissues; the $\delta^{13}C$ value in plant tissues is substantially decreased compared to its atmospheric value of approximately $-7$‰. The $\delta^{13}C$ value of plants that utilize the C$_3$ photosynthetic pathway, which characterizes woodland trees and high-latitude grasses, is $-23$ to $-33$‰. The $\delta^{13}C$ value of plants that utilize the C$_4$ photosynthetic pathway, used by many warm-season perennial grasses and a few shrub species, such as the saltbushes (Atriplex spp.), is $-16$ to $-9$‰. Desert soil-respired CO$_2$ thus possesses either a C3 or a C4 isotopic signature. The ratio of carbon isotopes in the soil carbonate precipitated in isotopic equilibrium with CO$_2$ molecules with these different isotopic values reflects a C$_3$ or C$_4$ value, or more typically, a value that reflects a composite value of a population of both C$_3$ and C$_4$ plants. Recent research focusing on incipient soil carbonate in soils associated with diverse plant communities in the northern Chihuahuan Desert of New Mexico (Breecker, Sharp, and McFadden 2009) revealed the relatively close relationship between plant types and the $\delta^{13}C$ values in the soil carbonate (Figure 8). Such studies, and studies of other stable isotopes in soil materials, elucidate the processes largely governing the timing of carbonate movement and soil carbonate precipitation. Such research shows, for example, that soil carbonate may not necessarily form annually during the growing season as many soil scientists had long believed. Instead, much soil carbonate may accumulate principally every few to several years after an especially strong drought.

Desert soils and climatic change

Worldwide, patterns of climate and its effect on soil water balance are closely associated with variations in the dominant types and density of plants. Consequently, climate variation in deserts also is associated with morphological variability. For example, climosequence studies show that as annual precipitation increases, the depth at which soil carbonate accumulates commonly increases. In transitional areas (desert to humid regions of ≥50 cm annual precipitation), carbonate may accumulate at depths greater than 100 cm, or soils there may be thoroughly leached and hence noncalcic. In marked contrast, in the driest and warmest deserts (e.g., <8 mm annual precipitation), accumulated gypsum or other soluble salts typically predominate shallow B horizon development, and soil carbonate accumulation is inhibited (Amit and Yaalon 1996).

During the Quaternary period, Earth’s climate changed substantially, varying between generally warmer interglacial (IG) periods, such as that represented by the Holocene epoch, and generally colder glacial (G) periods, such as the most
Figure 8  Profiles of carbon isotope values of soil and those predicted from measured soil-gas values, based on analysis of soils associated with four different biomes in arid and semi-arid areas of the northern Chihuahuan Desert, New Mexico. The key for all plots is shown in panel c. Results from both disturbed and undisturbed soil are shown at 92 cm for the Great Basin Shrubland. Dashed horizontal lines show the depth below which measured and predicted values can be directly compared. Equilibrium conditions were most closely achieved on May 4, 2008, and the profiles for this date are highlighted by enlarged symbols. VPDB – Vienna PeeDee Belemnite. Adapted from Breecker, Sharp, and McFadden (2009) with permission of the Geological Society of America.

Recent ice age of the late Pleistocene epoch. Results of these studies indicate that average annual temperatures change by as much as 4–5 °C from a glacial to an interglacial period, and that these changes were accompanied by substantial increases in precipitation in most desert areas. These types of dramatic climatic changes strongly affected the plant communities of deserts. For example, in the southwestern United States, areas now occupied by desert shrub vegetation were occupied by piñon-juniper woodlands during the most recent glacial period, equivalent to a shift in the average elevation of these biomes by 1000–1400 m. Global changes in climate during the Quaternary period were also characterized by other important environmental changes, such as strong variations in dust flux. Considering the strong relationships between
climate, vegetation, and dust on pedogenesis, it is easy to understand why these climatic changes also substantially influenced desert soil formation. From the 1960s through the early 1980s, Desert Project researchers, working in New Mexico, concluded that at least some of the features observed in pre-Holocene soils reflected these influences, rather than simply reflecting time-dependent changes. For example, they attributed the accumulation of desert carbonate in argillic horizons to a shift to drier climate and shallower depths of clay accumulation within a formerly noncalcic B horizon.

Desert pavement-forming processes also may have been inhibited during the cooler and effectively wetter climates of the last glacial period. Paleobotanical features such as the presence of microtopographic traces of former occupation by perennial plants (“plant scars”) on some desert pavements in the US Southwest likely reflect the shift from G to IG climate and the associated vegetation contraction. However, some older Pleistocene pavements and soils in some high desert areas seem to have persisted through the last glacial period. Possibly the influences of some features, such as dark pavements conductive to elevated soil temperatures and evapotranspiration rates, were sufficiently unfavorable to the recruitment and sustained colonization by plants, even during glacial periods. The presence of well-developed vesicular horizons, through their strong influence on soil hydrology, also can substantially decrease the depth of soil water movement and soil carbonate accumulation (McFadden et al. 1998). As a result, some desert pavements persisted through glacial periods and soil development continued (McFadden 2013).

Numerical modeling studies of soil carbonate accumulation can be used to study processes controlling its accumulation, and can help explain how Quaternary climate changes might have affected rates and depths of carbonate accumulation, as affected by changes in dust influx, soil water balance, plant canopy cover, evapotranspiration and soil respiration (McFadden et al. 1998). For example, it has been long hypothesized that G to IG climate changes should have caused a significant decrease in the average depth of soil carbonate accumulation, potentially resulting in a bimodal profile of soil carbonate concentrations with depth. Field studies of late Pleistocene desert soils show that such a pattern is common, and results of numerical modeling provide strong support for this hypothesis. Shifts in biotic communities of deserts, and, in many cases, the relative proportions of C_3 and C_4 plants, are also recorded in the isotopic compositions of soil carbonate in late Pleistocene and Holocene desert soils.

Desert soils and landform evolution

Chronosequence studies demonstrate the interdependent relationships between soils and associated desert landforms. They show, for example, that dust entrapment and accumulation on stable geomorphic surfaces promote the development of well-formed desert pavements and soil profiles in as little as a few thousand to a few tens of thousands of years (Wells et al. 1995). Once formed, a desert pavement can persist for hundreds of thousands or even several million years (Figure 6). Sustained pavement stability and dust accumulation rates promote continued AIP, and therefore also the continued subsurface accumulation of clay-sized soil carbonate, silica, and/or other salts, and leading to, respectively, increasingly thicker argillic horizons, petrocalcic horizons, and so on. Soils that thicken by AIP, however, cannot continue to thicken indefinitely. Eventually the destabilization of geomorphic surfaces is promoted by the presence of increasingly thicker, clay-enriched B horizons, which act to decrease...
soil infiltration rates and lead to increases in the proportion of surface runoff. Thus, erosion potential is increased. Erosion of most of the Bt horizon and decreased depths to a plugged and essentially impermeable subsurface horizon also would substantially decrease translocation of soil water to great depth, thereby favoring increased runoff. The progressive development of a network of drainage lines and channel incision on older desert piedmonts locally steepens surface gradients, further increasing erosion potential.

Once exposed by erosion, massively cemented, indurated soil horizons such as a petrocalcic horizon get exposed at the surface. These horizons are very resistant to weathering and erosion, enabling them to persist for millions of years (Figure 3). Accordingly, such horizons have been regarded as essentially “irreversible.” Studies in Australia indicate that some duripans can also remain as landscape elements for tens of millions of years. On some piedmonts in the southwestern United States, slow dissolution of the uppermost part of a petrocalcic horizon beneath a thin B horizon, combined with continuous channel incision, ultimately results in a convex-up surface form, known as a *ballena*. In other circumstances, however, rapid, deep channel incision has initiated instead cliff retreat enabled by the presence of the thick, indurated, locally silica-rich petrocalcic soil in southern Nevada (see Figure 3).

Much of the landscape in desert regions is hilly or mountainous. These surfaces typically generate considerably more runoff and thus have high erosion potential. Steeper, high-relief hillslopes also favor mass movements, such as landslides and debris flows. Together, these circumstances do not favor soil formation. Also, the limited magnitude of chemical weathering in deserts often means that weathering of hillslope materials and bedrock occurs mainly by physical weathering, leading to the production of mobile regolith (colluvium). This material can then accumulate as talus or mass movement debris at the bases of steep bedrock slopes. Steep, bedrock-dominated hillslopes in the arid and semiarid regions of the western United States are very commonly associated with well-cemented sandstones, welded ignimbrites, massive limestones, basalt flows, or other rock types (Figure 9a).

Interestingly, the typical hillslopes associated with marine mudrock-dominated formations such as the Chinle Group and the Mancos Shale in the American Southwest exhibit a very different hillslope form, with generally lower slope gradients and numerous rills and gullies (Figure 9a). Referred to as badlands, the associated sedimentary rocks lack the shear strength to support steep, high cliffs. Even minimal weathering can help to reduce the strength of mudrock. The high expandable clay content of these rocks also favors shrink–swell weathering and some creep-driven downslope transport. Saturation of the mudrocks during prolonged rainfall, however, can effectively seal the surface layer, substantially increasing runoff and erosion. The commonly high salt content of mudrocks also significantly inhibits colonization by all but very salt-tolerant plants, which also increases erosion potential. As a consequence of rapid erosion and slope retreat, as is the case for steep bedrock hillslopes, badlands usually lack soils.

In a famous paper, Gilbert (1880) argued that relatively steep, bedrock-dominated hillslopes form when the rate of regolith formation is exceeded by the rate of the downslope transport of weathered material. This situation is described as a *detachment-limited* hillslope. Such a slope typically exhibits variously developed soils (Figure 9a). Gilbert also proposed that, in a more arid climate where weathering processes are weak, detachment-limited hillslopes would be the most common hillslope type, particularly on rock types resistant to weathering, such as basalt. Recognizing the existence of smooth,
SOILS OF DESERT LANDSCAPES

soil-mantled hillslopes, Gilbert observed that they were generally prevalent at high elevations; the bedrock-dominated, detachment-limited hillslopes were at lower elevations. Gilbert hypothesized that the smooth form and commonly continuous vegetation mantles reflected higher magnitudes of weathering in a more humid climate. In these circumstances, formation of weathered material outpaces the rate of its erosion. The vegetation cover intercepts precipitation, limits erosion, and favors infiltration, all of which enhance the weathering of bedrock. The main form of downslope transport on such hillslopes is by abiotic and biotic-mediated creep processes, which lead to the formation of smooth hillslopes that become increasingly steeper downslope, accommodating transport of additional weathered material. These soil-mantled hillslopes are referred to as transport-limited (Figure 9b).

Transport-limited hillslopes are present but somewhat uncommon in desert regions (Figures 9b and 10). Soil geomorphic studies show that desert hillslopes are excellent dust traps, especially on the favorably oriented, windward aspects. Dust entrapped in bedrock cracks can increase the magnitude of physical weathering, thereby increasing the production of colluvium. In certain situations, dust entrapment may enable the formation of isolated areas of soil development that exhibit vesicular A, Bt, and Bk horizons. Locally continuous, thick soil mantles have formed, however, on some of the lower, piñon-juniper-covered foothill slopes of the granite-dominated Sandia Mountains (Figure 9b) (Persico et al. 2011). The slightly moister semiarid climate, generally windward orientation, and ample dust supply support a moderate vegetation cover here, which favors soil development. Most of the foothills of the Sandia Mountains exhibit bedrock-dominated corestone topography, implying that another factor also must locally favor soil development. In this area, the granitic rocks locally exhibit fine-grained dikes or small bodies composed of aplite, a rock that is resistant to chemical weathering. Physical weathering of aplite produces angular, joint-bound blocks that range from a single to a few decimeters in diameter. Unlike the much finer granitic grus, blocky aplite is far more resistant to erosion, accumulating instead as colluvium. Because relatively large, erosionally resistant bodies of the aplite eventually form the summit areas of hillslopes, aplitic colluvium buries granitic bedrock on lower hillslopes areas. The angular colluvial cover effectively entraps dust, and much of the dust is washed into the granite-mantling colluvium. Thus, despite the relatively steep hillslope gradient, remarkably thick, reddened Bt horizons have formed on these hillslopes, mainly by AIP, likely over tens of thousands of years.

Piñon-juniper woodland-mantled hillslopes also occur in other semiarid regions of the southwestern United States, including many areas of the Colorado Plateau. Some of these hillslopes are also smooth and have continuous soil mantles (Figure 10). These soils

Figure 9  (a) Detachment-limited hillslopes formed on Jurassic sedimentary rocks in the Grand Staircase National Monument, southern Utah. The uppermost, steep cliff-dominated hillslopes have formed in relatively well-cemented sandstones. The rilled (drainage lines) badlands that dominate the lower portions of these hillslopes have formed on mudrocks. Photo by L. D. McFadden. (b) Transport-limited hillslopes in the foothills of the Sandia Mountains, central New Mexico. Corestone-dominated, largely detachment-limited hillslopes are shown in the left foreground. Reprinted from Persico et al. 2011. © 2011, with permission from Elsevier.
SOILS OF DESERT LANDSCAPES

Figure 9

(a) Soil-mantled, transport-limited "Aplite hillslopes"

(b)
Figure 10  Steep, transport-limited hillslopes (S) with a southwest-to-west, relatively warm aspect contrast with the soil-mantled, transport-limited hillslopes (N) with an effectively more humid northern aspect formed on Jurassic sandstones, as seen in northeastern Arizona. Recent erosion on the steep, detachment-limited hillslopes on the left side of this image has removed all but a few isolated remnants of soil mantle, exposing bedrock. Decreases in the number of piñon and in the area of soil cover in the transitional, ecotonal region between these two hillslope areas are taken as evidence of ongoing hillslope transformation, likely occurring in response to climate change(s). Note the emerging, large sandstone concretions (C) and living roots (R) of cliffrose shrubs exposed by very recent soil erosion. Photo by L.D. McFadden.

are, however, relatively thin and have weakly developed A–C soil profiles. In northeast Arizona, these hillslopes have formed on regionally extensive sandstones of Jurassic age. Some Jurassic sandstones are cemented by clay, rather than calcite and silica cement. A significant fraction of the clay is smectite, and its weathering by cyclic hydration–contraction is conducive to the formation of a continuous, but weakly developed, soil mantle. The continuous vegetation canopy inhibits erosion, and consequently downslope sediment movement occurs by soil creep. Rapid soil creep and the lack of large, erosionally resistant rock fragments on the hillslopes minimize entrapment and accumulation of dust and, therefore, also the formation of thick, well-developed soils by AIP (McFadden 2013).

Smooth soil and vegetation-mantled hillslopes in this region of the southwestern United States are most prevalent on areas of northern aspect. Hillslopes with a warmer, effectively drier, southern aspect are generally steeper, with a much higher proportion of exposed bedrock (see Figure 10). Differences in topographically controlled, aspect-related microclimate diminish the magnitude of weathering and minimize soil formation on slopes with southern aspects, favoring the evolution of bedrock-dominated,
detachment–limited hillslopes. Aspect-associated
differences in average soil temperature are, how-
ever, not especially large, demonstrating the
sensitivity of these hillslope systems to relatively
small variations in topoclimate. In spatially
extensive hilly landscapes in the southwestern
United States associated with largely uncon-
solidated, noncohesive substrates, topoclimatic
variation also favors aspect-related difference in
hillslope form and vegetation cover. In these
circumstances, the exposure of the underly-
ing, noncohesive alluvium on southern aspect
hillslopes, rather than bedrock, causes the for-
mation of surface runoff-dominated hillslopes
with lower gradients than the soil-mantled,
north-aspect hillslopes (McAuliffe et al. 2014).
Hillslope soil, vegetation, and dendrologic
studies in the ecotonal areas between the
piñon-dominated northern and shrub-dominated
southern aspect hillslopes in the Colorado
Plateau study area show that the transformation
of soil-mantled, transport-limited hillslopes to
bedrock-dominated, detachment-limited hill-
slopes is very rapid (McAuliffe et al. 2014).
The transformation is achieved on timescales of
a few thousand years, or even centuries. The
change in hillslope form and soil–slope relations
is probably triggered by substantial reductions
in intercanopy grass cover caused by sustained
drought. Subsequent wetter climate, the reduced soil cover limits re-recruitment
and vegetation recovery. These hillslopes reflect
a landscape system that is highly sensitive to
changes in vegetation cover and soil thickness.
When a hillslope system threshold is reached,
subsequent erosion rates greatly increase, and
the result is the irreversible transformation of
hillslope vegetation communities, soils, and
hillslope form (Figure 10). Many well-known
areas in the Colorado Plateau with picturesque,
colorful sandstone weathering landforms are
likely the consequence of these types of process
interactions (Figure 11).

Conclusions

Soil geomorphic research over the last few
decades has revealed the critical roles played
by dust, parent materials, and vegetation in
the formation and evolution of desert soils and
associated landforms. On geomorphically stable
surfaces, prolonged dust entrapment and its
subsurface incorporation promote the formation
of thick, well-developed soils by AIP, over
long time spans. In drier deserts with very
sparse vegetation, these soils are commonly
associated with desert pavements and vesicular
A horizons. Thick well-developed soils also
form under dust-trapping plant communities
in semiarid regions. Even on geomorphically
unstable, steeper hillslopes, dust entrapped on
hillslopes associated with rock types that produce
dust entrapping colluvium, or the presence of
weakly cemented sandstone, is conducive to soil
development.

Understanding the relationships among the
diverse soil types and landforms in deserts
requires an appreciation of the sensitivity of
plants and certain rock types to climate. Whereas
some soils and landforms evolve slowly despite
having been subjected to major climate changes,
others have changed in profound ways. In
some desert settings, the rapidity of changes in
response to even relatively minor recent and
ongoing changes in climate is remarkable, and
such responses to predicted future global climate
changes are all but certain.

SEE ALSO: Biogeomorphology; Climate
change and biogeography; Geomorphic
thresholds; Holocene; Quaternary
geomorphology and landscapes; Soils in geomorphic research; Soils on slopes: catenas; Soils and weathering

References


Further reading


Soils of humid mid-latitude landscapes

Darrell G. Schulze
Purdue University, USA

Soil properties on a global scale are often discussed in terms of general soil regions, with climate being a defining characteristic, for example, “temperate soils” versus “tropical soils.” Although there are clear distinctions between soils in various climatic regions of the world that warrant their grouping into broad classes with similar soil properties and functions, only rarely are these classes formally defined. Therefore, to facilitate discussion and comparison of “humid mid-latitude soils” with respect to other soils of the world, the four global soil regions are first defined using criteria based on the soil temperature and soil moisture regimes of Soil Taxonomy (Soil Survey Staff 1999). These four groups, namely cold soils, arid soils, humid tropical soils, and humid mid-latitude soils, parallel entries with similar names in this encyclopedia.

Areas delineating each of the above groups (Figure 1a) were created utilizing maps of global soil temperature regimes (USDA 1997a) and global soil moisture regimes (USDA 1997b). Four mutually exclusive classes can be delineated according to the following criteria:

- Cold soils consist of all soils in the gelic, pergelic, and hypergelic soil temperature regimes. These are very cold soils that either contain permafrost, or have mean annual soil temperatures \( \leq 1^\circ C \) at 50 cm depth.
- Arid soils are all soils that occur in aridic soil moisture regimes. These soils are dry throughout the soil profile for more than half the year when soil temperature at 50 cm is >5°C, and moist for <90 consecutive days per year when the soil temperature is >8°C. In other words, they are dry during the warm growing season. They support xerophytic vegetation in their natural state, and must be irrigated if they are used for crop production.
- Humid tropical soils are soils that occur in iso-temperature regimes, that is, areas in which the mean summer and mean winter soil temperatures at 50 cm differ by <6°C. These soils have no distinct “cold season.” Most are warm throughout the year, although the few that occur at high altitudes, such as those on Mounts Kilimanjaro and Kenya in Africa or in the high Andes in South America, can be cooler. Note that soils already grouped with cold and arid soils are excluded from this category.
- Humid mid-latitude soils consist of all soils that are not previously defined as within the cold, arid, or humid tropical soil groups. Thus, they are a diverse group that can have xeric, ustic, udic, or perudic soil moisture regimes, and cryic, frigid, mesic, thermic, hyperthermic, and megathermic soil temperature regimes (Table 1).

In general terms, humid mid-latitude soils are those soils that (i) have sufficient moisture in normal years to support some sort of agriculture (as opposed to only grazing or pasture) without irrigation, (ii) have distinct seasonal temperature differences between summer and winter, with
SOILS OF HUMID MID-LATITUDE LANDSCAPES

plant growth limited by cool temperatures in winter, and (iii) are warm for long enough periods of time during the growing season to allow agricultural crops to reach maturity. As thus defined, humid mid-latitude soils comprise 27.4% of the Earth's ice-free land area (Table 2, Figure 1a and 1b), and most, 83%, occur in the northern hemisphere. More than half of the Earth’s population, 58.6% in the year 2000 (CIESIN and CIAT, 2005b) (Table 1; compare Figure 1a or 1b with Figure 1c), lives on and is supported by humid mid-latitude soils.

Most humid mid-latitude soils occur on various kinds of geologically young parent materials. Such sediment is usually rich in the weatherable minerals that impact the soils that

---

**Figure 1** (a) General global soil regions. (b) Distribution of humid mid-latitude soil orders. (c) Population density in the year 2000. (d) Fraction of land area used for agriculture. (e) Nitrogen fertilizer use. Data sources: (a) derived by combining USDA (1997a) and USDA (1997b) as described in the text; (b) global soil orders map (USDA 2005) clipped to the humid mid-latitude region in 1a; (c) CIESIN and CIAT (2005a); (d) Ramankutty et al. (2008, 2010); (e) Potter et al. (2010, 2011a). *Continued opposite.*
Figure 1  Continued
formed on them, and provides nutrients for the ecosystems and crops that these soils support. There are a number of reasons for the geologically young landscapes on which many humid mid-latitude soils occur. First, large areas in the Northern Hemisphere were glaciated during the Pleistocene epoch. Glacial ice obliterated the pre-existing landscapes, eroding uplands and filling valleys so that smooth, gently rolling plains remained after the glaciers receded. The sediments deposited by these glaciers were derived mainly from fresh bedrock, explaining

Table 1 Descriptions of soil moisture and temperature regimes used in *Soil Taxonomy*.

<table>
<thead>
<tr>
<th>Moisture regime</th>
<th>Description*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aridic</td>
<td>Soil moisture regime of arid and semiarid climates where there is insufficient soil moisture for crop production without irrigation.</td>
</tr>
<tr>
<td>Ustic</td>
<td>Soil moisture regime of semiarid climates where soil moisture is limited but available for portions of the cropping season.</td>
</tr>
<tr>
<td>Udic</td>
<td>Soil moisture regime of humid regions with seasonally well distributed precipitation throughout the cropping season.</td>
</tr>
<tr>
<td>Perudic</td>
<td>Soil moisture regime of areas of very high rainfall in which precipitation exceeds evapotranspiration during every month.</td>
</tr>
<tr>
<td>Xeric</td>
<td>Soil moisture regime of Mediterranean climates with cool, moist winters and warm, dry summers. Crops must rely on stored soil moisture during the growing season or irrigation may be needed.</td>
</tr>
<tr>
<td>Aquic</td>
<td>A soil moisture regime that occurs when the soil is periodically saturated with water due to a high water table for long enough periods of time that microorganisms deplete dissolved oxygen and the soil becomes anaerobic. This situation results in distinctive soil color patterns that are used to identify this moisture regime. The aquic soil moisture regime can occur in any climate.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temperature regime</th>
<th>Description†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gelic</td>
<td>Too cold for almost all crops. Very cold (≤1°C) mean annual soil temperature.</td>
</tr>
<tr>
<td>Cryic</td>
<td>Too cold for almost all crops. Cold (&lt;8°C) mean annual soil temperature and also cold in the summer.</td>
</tr>
<tr>
<td>Frigid</td>
<td>Too cold for most crops except those adapted to cold weather such as wheat, oats, hay, or barley. Cold (&lt;8°C) mean annual soil temperature, but summers are warmer than the cryic regime. Distinct seasonal difference in soil temperatures (&gt;6°C difference between mean summer and mean winter soil temperatures).</td>
</tr>
<tr>
<td>Mesic</td>
<td>Suited to crops such as corn, soybeans, and wheat. Moderate (8–&lt;15°C) mean annual soil temperature. Distinct seasonal difference in soil temperatures.</td>
</tr>
<tr>
<td>Thermic</td>
<td>Suited to crops such as cotton that require warm weather. Warm (15–&lt;22°C) mean annual soil temperature. Distinct seasonal difference in soil temperatures.</td>
</tr>
<tr>
<td>Hyperthermic‡</td>
<td>Suited to freeze-intolerant crops such as citrus. Hot (22–&lt;29°C) mean annual soil temperature. Distinct seasonal difference in soil temperatures.</td>
</tr>
</tbody>
</table>
Table 1 (Continued)

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Megathermic‡</td>
</tr>
<tr>
<td>Isomesic, isothermic, isohyperthermic, isomegathermic‡</td>
</tr>
</tbody>
</table>

*Soil moisture regimes are defined on the basis of soil moisture, which is determined by the amount and distribution of rainfall and the temperature for all the moisture regimes except the aquatic regime, which is defined by soil morphological characteristics. The full definitions of soil moisture regimes and soil temperature regimes in *Soil Taxonomy* are lengthy and quantitative (Soil Survey Staff, 1999, 2014). The descriptions here are very generalized and based on Soil Survey Staff (2015).

†Further subdivided into hypergelic (≤10°C), pergelic (−4°C to −10°C), and subgelic (+1°C to −4°C) according to the mean annual soil temperature at lower levels of *Soil Taxonomy* (Soil Survey Staff, 2014).

‡The upper limit of the hyperthermic regime and the megathermic and isomegathermic regimes are defined by Eswaran et al. (1997).

Table 2 Surface area, total population, fraction of land in cropland, and mean nitrogen and phosphorus fertilization rate for the general soil regions shown in Figure 1.

<table>
<thead>
<tr>
<th>Region</th>
<th>Area (km²)</th>
<th>Global fraction of the Earth’s ice-free area (%)</th>
<th>Total population (x1000)</th>
<th>Fraction of total world population (%)</th>
<th>Fraction of area in cropland</th>
<th>Mean N fertilization rate (kg ha⁻¹)</th>
<th>Mean P fertilization rate (kg ha⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cold soils</td>
<td>21,786</td>
<td>16.7</td>
<td>24,868</td>
<td>0.4</td>
<td>0.004</td>
<td>0.1</td>
<td>0.004</td>
</tr>
<tr>
<td>Arid soils</td>
<td>38,294</td>
<td>29.3</td>
<td>635,684</td>
<td>10.7</td>
<td>0.068</td>
<td>0.45</td>
<td>0.068</td>
</tr>
<tr>
<td>Humid tropical soils</td>
<td>34,800</td>
<td>26.6</td>
<td>1,790,138</td>
<td>30.2</td>
<td>0.123</td>
<td>0.58</td>
<td>0.123</td>
</tr>
<tr>
<td>Humid mid-latitude soils</td>
<td>35,748</td>
<td>27.4</td>
<td>3,468,679</td>
<td>58.6</td>
<td>0.225</td>
<td>2.68</td>
<td>0.225</td>
</tr>
</tbody>
</table>


why they are so rich in unweathered (primary) minerals. The soils on these formerly ice covered areas are no older than Pleistocene age (less than about 1.6 million years) and many are no older than 20,000 years. Pleistocene glaciers also impacted soils at considerable distances from the glacial margin. Meltwater carried freshly ground sediment from the glacial margin, depositing it as sandy and gravelly sediment called glacial outwash. Wind blowing across the dry floodplains in the winter picked up silt that was then deposited over large areas of the surrounding uplands as a deposit called loess. Loess often covers entire landscapes with deposits several meters thick. Even when thin, however, the loess adds fresh, unweathered minerals to older,
SOILS OF HUMID MID-LATITUDE LANDSCAPES

more weathered soils, adding divalent cations such as calcium (Ca\(^{2+}\)) and magnesium (Mg\(^{2+}\)), as well as silica and other elements, and initiating a new cycle of weathering and soil formation. Other areas of humid mid-latitude soils occur in tectonically active areas where erosion from mountainous areas provides fresh sediment to the surrounding lowlands. Finally, in some areas, volcanic activity provides fresh material, such as volcanic ash, for soil formation. For this reason, many of the older, more weathered soils like Ultisols and Oxisols are relatively uncommon across the humid mid-latitudes.

Types of humid mid-latitude soils

Soils of the humid mid-latitudes have been intensively studied; thus, it is no accident that they are represented prominently in all of the world's major soil classification systems. Eleven of the 12 soil orders of Soil Taxonomy occur in (or even dominate) humid temperate regions (Figure 1b, Table 3). Each order is summarized briefly here, in order of generally increasing pedogenic development.

Entisols are weakly developed and, hence, often very young soils. They have an ochric epipedon, which is a light colored and often thin surface horizon. Entisols lack subsurface horizons that have been significantly altered by soil forming processes. They are common on steep, eroding surfaces, as well as on floodplains and other areas that have only recently been exposed to pedogenic processes. They also include some very sandy soils and soils formed in parent materials that are resistant to weathering. Entisols make up 12.4% of the land area of humid mid-latitude soils (Table 3) and are distributed widely such that large contiguous areas do not stand out on small-scale maps (Figure 1b).

Like Entisols, Inceptisols usually have an ochric epipedon, but are slightly better developed in the lower profile. Inceptisols have a

<table>
<thead>
<tr>
<th>Soil order</th>
<th>Area (km(^2) ×1000)</th>
<th>Percentage of Earth’s land area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inceptisols</td>
<td>8178</td>
<td>22.9</td>
</tr>
<tr>
<td>Alfisols</td>
<td>7603</td>
<td>21.3</td>
</tr>
<tr>
<td>Mollisols</td>
<td>5152</td>
<td>14.4</td>
</tr>
<tr>
<td>Entisols</td>
<td>4408</td>
<td>12.4</td>
</tr>
<tr>
<td>Ultisols</td>
<td>3394</td>
<td>9.5</td>
</tr>
<tr>
<td>Spodosols</td>
<td>3283</td>
<td>9.2</td>
</tr>
<tr>
<td>Vertisols</td>
<td>1010</td>
<td>2.8</td>
</tr>
<tr>
<td>Histosols</td>
<td>901</td>
<td>2.5</td>
</tr>
<tr>
<td>Aridisols</td>
<td>702</td>
<td>2.0</td>
</tr>
<tr>
<td>Rocky Land*</td>
<td>442</td>
<td>1.2</td>
</tr>
<tr>
<td>Andisols</td>
<td>423</td>
<td>1.2</td>
</tr>
<tr>
<td>Oxisols</td>
<td>193</td>
<td>0.5</td>
</tr>
<tr>
<td>Gelisols</td>
<td>0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

*Includes small areas of shifting sands.
cambic horizon, which is a subsurface horizon that shows evidence of chemical or physical alteration, but lack increases in clay content or changes in other properties that define more developed subsurface horizons. Because of their weakly expressed horizons, both Inceptisols and Entisols resemble their parent materials more closely than most of the other, more pedogenically developed, soil orders. Many Inceptisols are on steeply sloping landscapes. They are widely distributed throughout the area of humid mid-latitude soils (Figure 1b) and are particularly common in mountainous areas (Figure 2a). They make up the largest area of humid mid-latitude soils (Table 3) and are particularly common in western Europe, the northeastern United States, and eastern China (Figure 1b).

Alfisols have an ochric epipedon overlying an argillic horizon, which is a subsurface horizon that has higher clay content than the horizons above it, along with evidence that clay has moved into the horizon from above. Alfisols have high content of base cations (Ca$^{2+}$, Mg$^{2+}$, K$^+$, and Na$^+$) and are defined as having >35% base cation saturation at the bottom of the profile. Alfisols make up 21.3% of the area of humid mid-latitude soils (Table 3). Because it takes several thousand years for an argillic horizon to form, Alfisols are found on landscapes that have been stable for at least that long (i.e., late Pleistocene) and in parent materials that contain ample amounts of clay and other primary minerals such as quartz, feldspar, and mica. Many Alfisols have formed under deciduous broadleaf or mixed coniferous–deciduous forest, although they occur under a variety of other vegetation types as well. Large areas of Alfisols occur south of the Great Lakes in North America, along a wide swath across northern Europe and northeastern Asia, and in Australia (Figure 1b). Alfisols are generally high in natural fertility and are responsive to good management. As a result, they are used widely for agriculture and forestry. A representative profile of an Alfisol is shown in Figure 2b.

Spodosols are often coarse-textured soils that have an ochric epipedon overlying a spodic horizon, which is a subsurface horizon with an accumulation of organic matter, usually also containing large amounts of poorly crystalline iron and aluminum oxides and hydroxides. An albic horizon, which is a bleached, sandy, light-colored horizon, often occurs between the ochric epipedon and the spodic horizon (Figure 2c). The most extensive areas of Spodosols occur in cryic and frigid temperature regimes along the northern edge of the humid mid-latitudes (Figure 1b), although smaller areas occur throughout the world, where they are usually restricted to coarse textured parent materials and wet sites. Spodosols are the typical soils of the northern boreal and mixed coniferous–deciduous forests and are extensive in southeastern Canada and on the Scandinavian Peninsula and adjoining areas of Russia, as well as in the northern Great Lakes region of the United States (Figure 1b).

Ultisols have an ochric epipedon overlying an argillic horizon or a kandic horizon. The kandic horizon is similar to an argillic horizon but contains clays that have lower cation exchange capacities (CEC). In contrast to Alfisols, Ultisols are more acidic and have <35% base saturation at the bottom of the soil profile. Ultisols are deeply weathered soils, usually associated with acidic parent materials; they occur on old and stable land surfaces. In areas of crystalline bedrock, the thick and deep soil profile may be underlain by several meters of saprolite above the hard bedrock. Because of the high content of low CEC clay minerals, Ultisols have low inherent fertility and most plant nutrients occur in the surface horizons due to biocycling. These soils usually support forest vegetation if they
SOILS OF HUMID MID-LATITUDE LANDSCAPES

Figure 2 Photographs of four soil orders and of an agricultural landscape representative of the humid mid-latitudes. The depth scales are in centimeters and photographs (a)–(d) have been sized so that the profile depths scales are similar for each photo. (a) An Inceptisol in the Carpathian Mountains of Ukraine near the Slovakia–Poland–Ukraine border. The top of the scale is at the interface of the approximately 10 cm thick O horizon consisting of leaf litter in varying stages of decay and the underlying mineral soil. An ochric epipedon occurs from 0 to approximately 25 cm, followed by a cambic horizon from approximately 25 to 110 cm. Unweathered rock occurs below approximately 110 cm. The stones in the cambic horizon are typical of soils in mountainous landscapes. Its great group classification as a Lithic Dystrudept indicates the soil occurs in an udic moisture regime, is well drained with no indication of a seasonal high water table, is acid throughout, and is shallow to bedrock.
Figure 2  (b) An Alfisol from eastern Lower Michigan, USA. This soil has an ochric epipedon approximately 15 cm thick, overlying an argillic horizon from approximately 15 to 50 cm. Unweathered, calcareous glacial till occurs below approximately 50 cm. Its great group classification as an Inceptic Hapludalf indicates that the soil is in a udic moisture regime, is well drained with no indication of a seasonal high water table, and is less developed than most Hapludalfs. (c) A Spodosol from eastern Upper Michigan, USA. This soil has a dark ochric epipedon from 0 to approximately 15 cm, an albic horizon from approximately 15 to 40 cm, a spodic horizon from approximately 40 to 70 cm, and unweathered, sandy parent material below approximately 70 cm. Its great group classification as a Typic Durorthod indicates that it is a typical, well drained Spodosol with no evidence of a seasonal high water table, but the spodic horizon is indurated. (d) A Mollisol from Illinois, USA. This soil has a dark mollic epipedon up to approximately 60 cm thick overlying a gray argillic horizon. This is the Drummer soil series typical of the wet prairies of Illinois and neighboring states in the Midwestern USA. Its classification in the Typic Endoaquolls great group indicates that it has a seasonal high water table, which is also reflected in the standing water at the bottom of this profile. These soils are usually tile drained and cropped extensively to corn and soybean. (e) A landscape in northwestern Tippecanoe County, IN, USA that is typical of many agricultural areas of the Midwestern USA. Center of the photograph is at latitude N 40.479°, longitude W 86.985°, WGS 84, and the view is to the northeast. The area is at the interface between soils formed under prairie and soils formed under forest. Wet Mollisols, as in Figure 2e, occur on the nearly level glacial till plain in the central part of the photo, while better drained Alfisols, as in Figure 2b, occur in the bottom and right side of the figure along drainage ways and more sloping areas. Rectangular wooded areas are remnants of the forest that covered parts of this scene prior to European settlement. Crops are predominantly corn and soybeans. Sources: 2a, 2b, and 2c reproduced from http://geo.msu.edu/extra/soilprofiles/; 2d reproduced from Natural Resources Conservation Service, http://www.nrcs.usda.gov/wps/portal/nrcs/detail/il/soils/?cid=nrcs141p2_030652; 2e Schulze, 2015. Reproduced by permission of D.G. Schulze.

have not been cultivated. Ultisols are common in the southeastern United States and across southeastern Asia (Figure 1b). They make up slightly less than 10% of the area of humid mid-latitude soils (Table 3).

Mollisols are defined by the presence of a mollic epipedon, which is a thick, dark surface horizon (Figure 2d) that is high in organic matter and base cations. In addition to the mollic epipedon, base saturation must be ≥50% throughout the profile. Because they are classified mainly on the basis of the surface (A) horizon, Mollisols can have any of a variety of other subsurface horizons, including an argillic horizon or a calcic horizon, which is a horizon that contains an accumulation of calcium carbonate. Mollisols are the typical soils of grasslands. Thus, they are extensive on the Great Plains of North America and the Steppes of western Asia, where they occur between the Alfisols in the more humid climates to the east or north (Figure 2e), and the Aridisols of the more arid climates to the west or south (Figure 1b). Mollisols are also extensive in the Pampas of South America (Figure 1b). They have high natural fertility and are so extensively cultivated that in many areas few examples of their native grassland vegetation remain.

Vertisols are very clayey soils that shrink and swell upon drying and wetting. They have deep wide cracks that open to the surface when the soil is dry, but close when the soil wets up, usually in the wet season. These soils show evidence of shrinking and swelling throughout the profile. Many have dark surface horizons high in organic matter, although this is not a
SOILS OF HUMID MID-LATITUDE LANDSCAPES

defining characteristic. Vertisols form in neutral to alkaline parent materials that are either clayey initially, such as clayey sedimentary rocks or clayey alluvium, or that weather to clayey soils, such as mafic igneous rocks and volcanic ash. The shrinking and swelling of Vertisols damages houses, roads, and other types of infrastructure. Most Vertisols supported grasslands prior to cultivation and many remain under pasture today; when cultivated they are difficult to manage because the smectitic clays are sticky when wet and hard when dry. Vertisols make up about 3% of the land area of the humid mid-latitude soils (Table 3). The largest areas of Vertisols are on the Gulf Coastal Plain of North America, the Indian subcontinent, and eastern Australia, with smaller areas occurring in southern South America, southern Africa, and China (Figure 1b).

Histosols are soils that consist of approximately 50 cm or more of organic soil materials, such as peat and muck. They form in areas where organic matter accumulation exceeds organic matter oxidation. This situation occurs in areas that are saturated with water throughout much of the year, which is particularly common in cool, humid climates where precipitation frequently exceeds evapotranspiration. Histosols are most extensive along the northern fringes of the humid mid-latitudes, where they are associated with Spodosols in Canada and on the Scandinavian Peninsula (Figure 1b). Most of these areas of Histosols are in their natural state. Smaller areas of Histosols occur scattered throughout the upper Great Lakes Region of the United States, northern Europe, and western Russia, where they occur in association with Alfisols, and some Histosols occur in wetlands along the Atlantic and Gulf Coasts of the United States. Many of these areas have been drained for agriculture or forestry, while some are harvested for horticultural potting materials or mined for heating or electricity generation. Histosols store large amounts of reduced carbon and their use usually results in oxidation and release of carbon into the atmosphere.

Andisols are soils that have formed in volcanic materials such as tephra. Rapid weathering of the volcanic materials like glassy ash, pumice, and cinders results in the formation of poorly crystallized, high surface area, short-range-ordered minerals such as allophane, imogolite, and ferrihydrite that impart unique chemical and physical properties to these soils. For example, some Andisols may hold up to 200% or more water on a dry weight basis, and have a high capacity to fix phosphorus added as fertilizer. Nevertheless, Andisols that occur in the humid mid-latitude area are generally highly productive soils for agriculture and forestry. Andisols occur in volcanically active areas throughout the world at the edges of tectonic plate boundaries. They are common at various locations around the Pacific Rim, in Iceland, and in parts of Italy, but, because they make up only 1.2% of the humid mid-latitude soils (Table 3), they consist of only a few, barely visible areas on small-scale maps (Figure 1b).

Aridisols are the typical soils of arid regions. Their presence as up to 2% of the area of the humid mid-latitude soils (Table 3, Figure 1b) may be the result of cartographic error. Aridisols, however, can occur outside of arid climates if they contain a salic horizon, which is a horizon that is high in soluble salts. The dissolved salts in the soil water result in osmotic stress to plants like that of otherwise drier soils.

Oxisols are the world’s most highly weathered soils. They have low amounts of weatherable minerals and low inherent fertility and are typically associated with old, stable land surfaces in tropical climates. They are defined by their chemistry and mineralogy, not their climate. Thus, some Oxisols in the southern Brazilian states of Santa Catarina and Rio
Grande do Sul (Figure 1b) occur within the humid mid-latitudes as defined above, but they make up only 0.5% of the total area (Table 3).

Mineralogy of humid mid-latitude soils

Except for Ultisols and Oxisols, which together make up 10% of the humid mid-latitude soil area, the soil orders that cover the remaining 90% of the humid mid-latitude area (Table 3) are generally rich in weatherable primary minerals. These minerals are inherited either directly from the weathering of the igneous and metamorphic rocks themselves, or after previous cycles of erosion and re-deposition by water, wind, or ice.

The great majority of these weatherable primary minerals are the common minerals that make up most igneous and metamorphic rocks, for example, olivine, pyroxene, amphibole, mica, and feldspar group minerals, along with quartz. These minerals formed at high temperatures, and often at high pressures, either by the cooling and crystallization of a magma, or by metamorphism, usually deep below the Earth’s surface. They are far from thermodynamic equilibrium in the aqueous soil environment at the Earth’s surface. Thus, they slowly decompose and release their constituent elements to the soil. Primary minerals occur mainly in the sand (2 mm–50 μm) and silt (50–2 μm) size fractions. The majority are aluminosilicate minerals made up mainly of oxygen, aluminum, and silicon, with the remaining naturally occurring elements occurring in lesser quantities. As primary minerals weather, they release most of the nutrient elements (Ca, Mg, K, P, S, Fe, Mn, B, Zn, Cu, Cl, Co, Mo, and Ni) needed for plant growth, as well as the additional essential elements required by animals. Thus, in general, the presence of weatherable minerals is associated with young and productive soils, particularly when additional nutrient inputs are minimal.

The mineralogy of the clay fraction (<2 μm) also plays a critical role in defining the properties of humid mid-latitude soils and, with some exceptions, most humid mid-latitude soils have clay fractions with high surface areas and, thus, high chemical activity potentials. In all soils, clay minerals form from the elements released by the weathering of primary minerals. The majority of the clay minerals present in many soils, however, reflect the complex history of the parent materials from which the soils formed, as much as they do the current pedological conditions. Nevertheless, some generalizations are possible.

First, phyllosilicate clay minerals tend to be the most abundant secondary minerals in soils of the humid mid-latitude regions. The 2:1 phyllosilicates, represented mainly by smectite, vermiculite, clay-sized mica or illite, and various intergrades between these end member minerals, tend to be the most abundant clay minerals in Alfisols, Mollisols, Inceptisols, Entisols, and Vertisols. These 2:1 phyllosilicates have a net negative charge on their layers that is balanced by divalent and monovalent cations, predominately Ca$^{2+}$, Mg$^{2+}$, K$^+$, and Na$^+$, along with smaller amounts of other cations, that are loosely held and can be exchanged for other cations. The high surface area and net negative charge of 2:1 phyllosilicates results in the high cation exchange capacities (or nutrient holding capacity) that are common in humid mid-latitude soils and explains why these soils are highly productive. Alfisols, Mollisols, Inceptisols, and Entisols often have a mixed clay mineralogy in which no one phyllosilicate mineral predominates. Vertisols, on the other hand, are characterized by high contents of smectitic clay. The swelling and shrinking of the smectitic clay at the molecular level as it wets and dries imparts the macroscopic properties
SOILS OF HUMID MID-LATITUDE LANDSCAPES

that define the Vertisol order, namely, slickensides and deep, wide cracks when the soil is dry.

The 1:1 phyllosilicates, mainly kaolinite and halloysite, make up the other major group of phyllosilicate minerals. Kaolinite and halloysite have very low charge on their layers and, as a result, they have very low cation exchange capacities and little ability to hold nutrient elements against leaching. Thus, soils dominated by these types of clay minerals often lose fertility over time. The clay fraction of both Ultisols and Oxisols are dominantly kaolinite, and thus, both soil orders are inherently less fertile than those rich in 2:1 phyllosilicate minerals.

Poorly crystalline and short-range-ordered minerals such as the minerals allophane, imogolite, and ferrihydrite occur in Andisols. These minerals account, in part, for the low bulk densities, high organic matter contents, high water holding capacities on a dry weight basis, high phosphorus fixation capacities, low cation exchange capacities, and other properties of Andisols that make them unique.

Human impacts on humid mid-latitude soils

Human land use activities are transforming large portions of the Earth’s surface, often at the expense of the environment (Foley et al. 2005). Such impacts can be locally dramatic, such as when surface mining for coal completely transforms a landscape and begins a new cycle of soil formation, or they can be less obvious but perhaps more significant, such as when excess nitrogen and phosphorus from agricultural areas results in eutrophication of lakes and marine environments. Human pressure on the global environment shows no signs of abating. World agricultural production will need to roughly double in the future to keep pace with population growth, changes in diet, and increased use of bioenergy (Foley et al. 2011), placing increasing pressure on the environment in general, and soils in particular.

Agricultural land uses occupy about 38% of the Earth’s land surface, with about 12% in croplands and 26% in pastures, making agriculture the largest use of land globally (Foley et al. 2011). Almost one-quarter (22.5%) of the land area of humid mid-latitude soils is in cropland, while 12.3% of the land area of tropical soils, and 6.8% of the land area of arid soils, is cropped (Table 2, Figure 1b). Additional expansion of agriculture by extending croplands and pastures to new areas still in natural ecosystems will have serious impacts in reducing biodiversity, increasing greenhouse gas emissions, and depleting ecosystem services. Most of the recent expansion and resulting environmental impacts have occurred in tropical areas, while adding relatively little to the global food supply (Foley et al. 2011). Rockström et al. (2009a, 2009b) suggest that croplands should occupy no more than 15% of the Earth’s ice-free land. Most of the recent gains in agricultural production have been achieved through intensification of agriculture by managing lands to be more productive by using fertilizers, irrigation, and biocides, and by mechanization and improved management practices (Foley et al. 2011). Much of this intensification has occurred on soils of the humid mid-latitudes, and this trend is likely to continue.

The amount of fertilizer applied to crops is one measure of how intensively soils are managed for agriculture. Using this metric, humid mid-latitude soils clearly stand out (Figure 1e). The mean rate of nitrogen fertilizer application for humid mid-latitude soils is 2.68 kg ha⁻¹, roughly five times the mean fertilizer application rates for humid tropical soils and arid soils (Table 2). Rates for phosphorus fertilizer
show similar trends (Table 2). Regions with the most highly fertilized soils include eastern China, northern India, Western Europe, and the Midwestern United States. These high rates of fertilization result in high crop yields, but not without environmental costs. Loss of nitrogen and phosphorus through runoff or leaching to surface or groundwater results in the eutrophication of lakes and marine ecosystems, with serious human and ecological consequences.

Agriculture is the largest source of excess phosphorus and nitrogen in waterways and coastal areas (Foley et al. 2005). Rockström et al. (2009a, 2009b) argue that the flow of new reactive nitrogen into the environment has already greatly exceeded its “planetary boundary,” and suggested that this flow should be limited to 25% of its current value. As a start, targeting “hotspots” of low fertilizer use efficiency could significantly reduce nutrient pollution from intensive agriculture. Roughly 32% of the global surplus of nitrogen and 40% of the global surplus of phosphorus come from only 10% of the world’s croplands (Foley et al. 2011), and most of these losses occur from soils in the humid mid-latitudes. The challenge will be to develop and implement sustainable agricultural systems that reduce nitrogen and phosphorus runoff and leaching from agricultural soils, while still maintaining high levels of production.

Conclusions

Defined by the broad forces of geology that made most humid mid-latitude soils geologically young and inherently fertile, and climate that has provided the moisture and temperature conditions conducive to agriculture, soils of the humid mid-latitudes have supported human civilizations for thousands of years. Because of the great variety of parent materials and climates across the humid mid-latitudes, soils here display remarkable diversity. The challenge in the future will be to sustainably manage these soil resources so that they continue to provide the food, fiber, fuel, shelter and other resources for a still growing population, while preserving and enhancing their ability to provide equally essential ecosystem services.

SEE ALSO: Ecosystem services; Food security; Land degradation; Land-use/cover change and climate; Population and natural resources; Soil and clay mineralogy; Soil erosion and conservation; Soil fertility and management; Soils of cold and permafrost-affected landscapes; Soils of desert landscapes; Soils of tropical landscapes; Soils and weathering; Sustainable development

References


SOILS OF HUMID MID-LATITUDE LANDSCAPES


Further reading


Soils of mountainous landscapes

Markus Egli  
University of Zurich, Switzerland

Jérôme Poulenard  
University of Savoy, France

Characteristics of mountainous landscapes

Mountain areas cover, depending on which definition is applied, roughly between 22 and 27% of the Earth’s total land area (Grabherr and Messerli 2011). A common definition of mountains is highlands and elevated ice shields above 2500 m, along with hilly areas below this altitude. Mountain areas play an essential role for life on this planet, but they also have vulnerable ecosystems that are complex and fragile. Soils are intimately linked to these ecosystems and consequently exhibit similar traits. This vulnerability – or the rate and direction of change – is fundamentally dependent on the composition of the parent material in which the soils are developing, as affected by the five soil-forming factors. Although not new, the principle that soils are a function of independent state factors such as time, climate, topography, vegetation, and parent material has become widely accepted, and it highlights and helps us to understand soil formation and the problems of soil wastage better (Figure 1).

Mountain landscapes are generally associated with relatively steep slopes that have a high energy potential. Furthermore, mountains frequently exhibit a high variability of forms. Morphodynamic processes related to the high-relief energy create specific habitats such as screes, landslides, and avalanches that further contribute to the richness of species and landscape elements. Owing to the great differences in environmental conditions within relatively short distances, mountain areas are often hotspots of biodiversity. The wide variety of areas at different elevations gives rise to the coexistence of several major climate zones and biomes from the mountain base to the summit. These altitudinal vegetation belts (orobiomes) that follow the climate zones, show the types, habits, and strategies of plants that are adapted to a particular water and energy balance (Burga, Klötzli, and Grabherr 2004). The following altitudinal sequence can usually be observed in many mountains: footslope (planar-collineate), montane, alpine, nival, or glacial zones (Figure 2). Along such an altitudinal sequence, the soil characteristics are clearly determined by the changing climate and vegetation.

Only geologically young (i.e., formed during the alpine orogeny; 100 Myr BP to present) areas actually occur as high mountains. Examples include the Himalayas, the Rocky Mountains, the Andes, the New Zealand Alps, and the Alps (Burga, Klötzli, and Grabherr 2004). A large part of the variscan (ca. 400–280 Myr BP) and Caledonian (ca. 570–390 Myr BP) mountains were eroded and smoothed during the Paleozoic and Mesozoic eras. Examples of the variscan high mountains are the Scottish highlands, the northern Appalachians, southeastern Australia, Newfoundland, eastern Greenland, and so on. Examples of mountains produced by the Caledonian orogeny are the Urals, southern Appalachians, north-eastern Australian mountains, and many others (Burga, Klötzli, and Grabherr 2004). The alpine orogeny occurred during the Cretaceous and Tertiary periods and is still ongoing in many places. These mountains reach relatively high altitudes, have steep slopes, and have a high potential for erosion. Landslides,
rock falls, debris flows, or avalanches are typical geomorphic processes in young, high mountains. Clearly, the age and type of the mountains has a strong influence on their ecosystems and soils.

A number of climate characteristics in mountain areas reflect regional or even local, rather than global, patterns. This climatic setup complicates explanations of mountain ecosystems and soils. Consequently, a typical mountain climate does not exist (Burga, Klötzli, and Grabherr 2004). Nonetheless, some generalities about mountain climates can be provided. With increasing altitude, temperature decreases in all mountain systems. The lapse rate of atmospheric temperature varies in general between 0.4°C in the more continental climate regions and 0.8°C per 100 m in the more oceanic regions. A decrease of 0.6°C per 100 m seems to be the best estimate for both temperate and tropical mountains. Mountain belts also experience considerable fluctuations in temperature between day and night, with frequent night frost at high elevation sites in tropical as well as in temperate zone mountains. The snow regime and distribution helps to prevent soil and plant exposure to these low temperature extremes. Winds also impact mountain climates because of their potential cooling, warming, and desiccating effects.

Figure 1 Model of the process interplay among mountain soils, in relation to soil formation processes and factors of soil formation with a given parent material.
Soils of Mountainous Landscapes

The amount of precipitation and its annual distribution are important for plant growth. The most common observation in mountain areas is an elevational increase in precipitation. However, in some mountain areas, precipitation tends to increase with elevation, but then to decrease above a mid-altitude maximum (e.g., in some subtropical areas such as Tenerife). Several investigations (e.g., Egli, Norton, and Dahms 2014) showed that the presence of moisture plays a crucial role in soil formation even in cold environments. Periodic extreme moist situations in soils may occur because of snowmelt, which has the same effect as heavy rainfall and results in a rapid passage of soil water. This flush of water produces much increased rates of leaching in soils and regolith.

Mountain soils: processes, evolution, major characteristics, and threats

Slope processes strongly influence the fate and evolution of mountain soils. The mass balance of a soil is the result of the relationship between soil material production and erosion (Figure 3). For all soils, changes in soil material mass can be expressed by (Egli, Norton, and Dahms 2014):

\[
\frac{\partial (\rho_w h_w)}{\partial t} = \rho_r Q_{UA} - \nabla \cdot (\rho_w \vec{Q_D})
\]

\[
= \phi - \nabla q_e - \nabla q_w
\]

(1)

where \(\rho =\) bulk density (\(w:\) soil; \(r:\) parent material), \(h_w =\) soil material thickness, \(Q_{UA} =\) soil material mass production rate per unit area, \(Q_D =\) denudation mass flux rate per unit area,
SOILS OF MOUNTAINOUS LANDSCAPES

Figure 3 Schematic overview of the evolution of a soil profile. Symbol $e$ refers to the interface between bedrock (or parent material) and soil; $h$ is soil thickness. The transformation rate is given by $P_{\text{soil}}$. $A$, $E$, and $W$ correspond to atmospheric input, erosion, and chemical weathering (solute output), respectively. Reproduced from Egli, Norton, and Dahms, 2014. © Elsevier.

$\phi = $ soil production rate, $q_e = $ soil erosion rate per unit area, and $q_w = $ chemical weathering rate per unit area. Due to steep slopes and in many cases extreme climate conditions, mountain soils may experience particularly high rates of erosion and chemical weathering and, consequently, production rates.

In mountain soils, chemical and physical weathering and mineral transformation contribute to progressive soil development, whereas erosion and deposition of unweathered materials lead to regression. In general, erosion leads to a rejuvenation of the soil surface by exposing less-weathered minerals, enhancing chemical weathering. Changes in soil-forming factors in mountain soils, across space, might be very abrupt, either by catastrophic and natural events or by the influence of humans (land-use change, intensification of agriculture). Thus, on a spatial level, mountain soils often show a high variability within very short distances. Soil evolution is, thus, best regarded as discontinuous over time and conceptualized by “progressive” or “regressive” process phases (Johnson and Watson-Stegner 1987). Because of the sloping nature of mountains, their soils are prone to lateral fluxes of water, matter, and solutes between soils. This tendency leads to strong dynamics in the spatial extension of soil domains.

The active nature of erosion and accumulation gives rise to the existence of (i) thin and constantly rejuvenated soils in the erosion zone and (ii) thick soils in the accumulation areas. Such events may lead to the successions of paleosols in the latter (Figure 4).

Due to geomorphologic activities and the minimal human impact, many mountain landscapes offer unique opportunities to study and model soil development pathways and rates. Chronosequences are particularly useful for estimating rates of soil formation and weathering (e.g., Föllmi et al. 2009) as well as for ecosystem development.

Because of their young age and high rock fragment content (Figure 4), mountain soils have

Figure 4 (a) A soil showing several buried paleosols, in the Anterne catchment, French Alps, 2100 m above sea level (photo J. Poulenard). (b) Typical soil in the Italian Alps, having a relatively high content of rock fragments (Entic Podzol (Episkeletic)). Photo reproduced with permission from M. Egli.
long been considered weakly developed. To many, weathering seemed to be almost nonex-
istent. However, recent studies have shown that soil formation and weathering rates in moun-
tainous areas – even in cold alpine regions – can be very rapid. Studies in mountain regions
have demonstrated that young soils weather intensively, even in areas of continuous seasonal
snowpack (Egli, Norton, and Dahms 2014; Figure 5).

Data on physical erosion and weathering
rates in mountainous regions have become
more widely available through cosmogenic
nuclide techniques (e.g., Riebe, Kirchner, and
Finkel 2004). Surveys in mountain areas using
cosmogenic nuclides ($^{10}$Be) and the chemical
depletion factor have indicated that physical ero-
sion and chemical weathering are, to a certain
extent, coupled (Riebe, Kirchner, and Finkel
2004; Dixon, Heimsath, and Amundson 2009;
Figure 6). Beryllium-10–derived estimates of
soil production and erosion rates show that
soil production rates increase with increasing
catchment-averaged erosion rates, a relationship
that enhances soil-cover persistence. A transition
to landslide-dominated erosion in steeper, more
rapidly eroding catchments results in thinner,
patchier soils and rockier topography.

Although subject to controversy, there are
increasing amounts of evidence that soil mois-
ture (rather than temperature) controls the
rate of soil formation in mountains, and that
this relationship holds even in cold alpine
environments (see e.g., Egli, Norton, and
Dahms 2014). As a consequence of microcli-
matic differences, more water is available in
soils with northern exposures (in the north-
ern hemisphere). Due to cooler conditions at
north-facing sites, evapotranspiration is lower
there and as a result soils may dry out more
slowly. This leads to a higher amount of water
percolating the soil and, subsequently, a higher
net export of elements. Weathering rates of
silicates depend on the action of water (together
with dissolved acids and organic and inorganic
ligands) on minerals. At north-facing alpine
sites more organic ligands seem to be present in
the percolating water, and these stimulate the
leaching of elements from soils and enhance
mineral dissolution and transformations. Con-
sequently, soil formation proceeds faster and
weathering processes are often more intense
on north-exposed slopes, even in cold alpine
environments.

Similarly, soil biological properties are influ-
enced by aspect. Although soil biota usually
show complex responses to environmental
settings, several differences in soil biological
properties can be related to local thermal dif-
ferences, caused by differences in aspect. Ascher
et al. (2012), among others, found that species
richness of microannelid assemblages was higher
under warmer conditions (south exposure; lower altitude) in alpine areas and that micro-
bial biomass (total phospholipid fatty acid) was
higher at cooler sites. Lower decomposition rates
were detected at cooler sites resulting in a lower respiratory loss and accumulations of weakly decomposed organic materials.

Organic matter is the most active and vital component of these soils because of its role in chemical, physical, and biological processes. Very high carbon stocks and sequestration rates can be found, even in (young) mountain areas (Figure 7); common carbon stocks values for the Alps range from 20 to 25 kg C m$^{-2}$. But these partially high stocks are under threat due to climate and – often increasing – land-use changes. The sequestration rates of organic C and N in soils of the European Alps, the Rocky Mountains (Wind River Range) and many other areas can reach very high values in young soils (up to 10–100 g m$^{-2}$ year$^{-1}$ for C and 0.5–5.5 g m$^{-2}$ year$^{-1}$ for N), whereas in old soils ($>20$ kyr) sequestration rates are several orders of magnitude lower ($<1$ g C m$^{-2}$ year$^{-1}$ and 0.01 g N m$^{-2}$ year$^{-1}$, respectively).

It is known from palynological and pedoanthracological studies that the altitudinal belts of vegetation in the mountains have shown considerable variations during the Holocene for reasons related to climate changes and/or in connection with human activities and land-use change. Studies of paleosols (Figure 4), soil charcoal, or lacustrine sedimentary archives (Mourier et al. 2010) have confirmed the succession of progressive (especially during the first part of the Holocene; Figure 8) and regressive phases of soil genesis phases in these mountain areas. In the European Alps, the tipping point between the two phases seems to be between 5000 and 2500 year BP and coincides with a stronger
human impact (agricultural use) on mountain environments.

In large parts of the mountainous landscape, grazing is the main traditional, human-induced, land-use activity. The distribution of pastural herds during periods of summer has significantly impacted the nutrient reserves in the soil and the spatial distribution of soil properties. There are, thus, very strong spatial variations in nutrient stocks in soils (predominantly P and N but also K, Ca and Mg) across mountains. High nutrient stocks are encountered when the contributions from animal droppings are higher than can be exported. In some areas, this situation can even lead to a dominance of nitrophilous plant species when the conditions are extreme. In areas where

Figure 7  Average carbon and sequestration rates (over the time considered of each single profile) in alpine areas. Reproduced from Egli et al. 2012. © Elsevier.
Figure 8 A conceptual soil evolution model (upper), showing progressive and regressive phases (Johnson and Watson-Stegner 1987), as compared to the dominant soil evolution pathway in the subalpine belt during the Holocene (lower). The model was developed by the author, from lacustrine sediment data. Created by author, using data from Johnson and Watson-Stegner 1987.
the output of nutrients (due to grazing) is higher than the input, nutrients are depleted from the soils. As a consequence, strong acidification occurs in these soils. Such spatial differences are much more pronounced in naturally acidic environments (crystalline rocks) than in environments having calcareous parent materials.

Mountain soils play a major role in the functioning and conservation of unique ecosystems and the hydrology of mountain belts. Thus, a change in alpine soil properties due to land use can raise concerns about their fertility and associated services. Traditional, human-made alpine pasture near the tree line has been common in mountainous regions for at least 7000 years. In parts of Europe, the abandonment of such pastures is starting to affect soil dynamics. In many other parts of the world, mountain soils are subjected to increasing grazing pressure, which leads to a rapid degradation and erosion. Acid deposition is another important threat to mountain soils in regions with silicate parent material and low buffering capacity. High mountain ecosystems are furthermore considered to be particularly sensitive to climate change. Overall warming and associated changes in precipitation patterns and snow cover will distinctly influence mountain vegetation as well. Climate change can influence both carbon balance (C mineralization versus C sequestration) and mineral balance (weathering versus erosion), although its effect is still not clear. Lastly, heavy rainfall events (due to global warming) will likely become more frequent, and erosion of mountain soils is likely to be enhanced in the future.

**SEE ALSO:** Earth system science; Quaternary geomorphology and landscapes; Soil erosion and conservation; Soil mass balance; Soils of cold and permafrost-affected landscapes; Soils on slopes: catenas; Soils and weathering

### References


SOILS OF MOUNTAINOUS LANDSCAPES


Further reading

Soils of the United States

Patrick J. Drohan
Pennsylvania State University, USA

The diversity of US soils (Figure 1) is arguably one of the nation’s most important assets. This is especially so given the commodities produced from soil and the influence of soils on the nation’s social, economic, and political history. The genesis of soil is the result of many surficial processes, some acting independently but many working synergistically, with erosion/deposition, biocycling, pedogenesis, and, recently, the actions of humans being key drivers of soil development and change.

Understanding soil distributions necessitates a larger and more holistic understanding of landscape evolution through time. Soil genesis, or pedogenesis, is typically first dependent on the parent material from which soil forms. The resulting soil horizons reflect the degree that other soil-forming factors (climate, time, topography, and organisms; Jenny 1941) and processes (additions, removals, translocations, and transformations; Simonson 1959) have had on the parent material. These concepts are used widely by academics and within the US Soil Survey program. They emphasize dominant differences in the state factors of soil, and the resulting effects that these state factors have had on soil morphology (color, texture, structure, etc.; see Figure 2).

A hierarchical classification system, Soil Taxonomy (Soil Survey Staff 1999), uses our accumulated knowledge of pedogenic processes to place the various soil morphologies into categories with similar pedogenic backgrounds (Table 1). Although all soil-forming processes are potentially active anywhere on Earth, their effect on pedogenesis can be limited by the complex interaction of the soil-forming factors. For example, translocation of soil materials downward in the profile is greatly limited in Gelisols, in comparison to other soil orders, owing to the cold climate where Gelisols occur. However, such processes are very common in many other soil orders.

Since 1894, when the United States Department of Agriculture (USDA) created the Division of Agricultural Soils, the United States has led the world in mapping and interpreting soils, in soil conservation, and in improving agricultural productivity based on soil science. Nearly all US soils have been mapped in considerable detail, and hundreds of thousands of soil samples have been analyzed by the federal government, land-grant universities, and other academic institutions. This effort has resulted in greater understanding of soil-forming processes and how they (and the resulting soils) vary across the landscape.

The Soil Survey was developed from a need to better support and guide agriculture, and thus in most instances their data products include soil information derived from a depth of only a meter or two. In addition, the complex nature of soil genesis has often been simplified so that the products can be effectively used by conservation agencies and the public. However, the reality of soil genesis is that virtually all soils are far thicker than depicted in the Soil Survey, and most soils are polygenetic in origin, having resulted from multiple periods and processes of soil formation.
Figure 1  Dominant soil orders in the United States (http://www.nrcs.usda.gov/Internet/FSE_MEDIA/stelprdb1237749.pdf).
Table 1  Soil orders of US Soil Taxonomy: their origin and defining characteristics. The orders are listed in their order of classification; the most limiting for agricultural use is first (Gelisols) and the least last (Entisols) (Soil Survey Staff 1999).

<table>
<thead>
<tr>
<th>Order</th>
<th>Characteristics</th>
<th>Formation details</th>
<th>Land-use trend</th>
<th>US land area (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gelisols</td>
<td>Permafrost within 100 cm, often with frost churning</td>
<td>Northern and southern latitudes; weakly developed</td>
<td>Ecological habitat; construction and development</td>
<td>6.65</td>
</tr>
<tr>
<td>Histosols</td>
<td>Soils formed in decaying organic materials ≥40 cm thick</td>
<td>Very wet areas, bogs, and swamps</td>
<td>Ecological habitat; agricultural production if drained</td>
<td>1.45</td>
</tr>
<tr>
<td>Spodosols</td>
<td>Soil with a subsoil horizon with an illuvial accumulation of organic matter, associated with aluminum and usually iron</td>
<td>Sandy, acid soils in either hot or wet (Florida) or cool and moist (Michigan) climates</td>
<td>Ecological habitat; forests; agricultural production; construction and development</td>
<td>6.83</td>
</tr>
<tr>
<td>Andisols</td>
<td>Relatively young soils, mostly of volcanic origin</td>
<td>Near volcanic landforms and on volcanic parent materials</td>
<td>Agricultural production with and without irrigation; ecological habitat; construction and development</td>
<td>1.63</td>
</tr>
<tr>
<td>Oxisols</td>
<td>Highly weathered soils with few primary minerals remaining</td>
<td>Tropical regions (hot and moist); on stable upland landscapes</td>
<td>Agricultural production; ecological habitat; construction and development</td>
<td>0.01</td>
</tr>
<tr>
<td>Vertisols</td>
<td>Clayey soils that contract and crack in the dry season and expand and swell in the wet season</td>
<td>Basic or calcareous parent materials rich in clays on stable landscape positions</td>
<td>Agricultural production with and without irrigation; construction and development</td>
<td>1.91</td>
</tr>
<tr>
<td>Aridisols</td>
<td>Soils of dry climates and deserts</td>
<td>Stable landscape positions; very low rainfall; carbonates and salts often accumulate in the subsoil</td>
<td>Ecological habitat; construction and development; rangeland; agricultural production</td>
<td>9.02</td>
</tr>
<tr>
<td>Ultisols</td>
<td>Soils with low amounts of base cations and a clay-enriched subsoil</td>
<td>Stable landscapes; formation from base-poor parent materials or as result of long-term weathering; periods of wet and dry conditions needed</td>
<td>Agricultural production; forests; construction and development; ecological habitat</td>
<td>9.95</td>
</tr>
</tbody>
</table>
Regional trends in soil characteristics across the United States

The size of the United States and its complex topography can make conveying soil patterns based solely on abiotic variables such as temperature or precipitation challenging. Nonetheless, the effect of climate on soil is described in part by the Soil Survey via soil moisture and temperature regimes (Figure 3; Figure 4). These data are also used at various categorical levels in Soil Taxonomy and infer conditions affecting soil development and agricultural practices. Soil moisture regimes describe classes of soil moisture conditions, whereas soil temperature regimes describe classes of soil temperature.

As one moves from the colder and wetter regions in northeastern North America (Figure 2) to the warmer, drier southwest deserts (Figure 4), the effects of precipitation and temperature on vegetation and soil morphology change. Climatic conditions limit aboveground vegetation in northern latitudes; hence, permanently frozen soils with weak development (Gelisols) are common (Table 1). Gelisol subsoil horizons may be frozen (Cf horizon) or mixed by freeze–thaw processes (Bwj horizon). Farther south, coniferous vegetation becomes common, in addition to young, sandy soils
formed in sediments deposited by Pleistocene glaciers. Cold temperatures and precipitation totals that exceed evapotranspirative demand in these soils often result in rapid, deep leaching and eluviation of soluble substances and clay minerals (Figure 2). Excessive leaching in an environment with few base cations (Ca, Mg, etc.) often leads to the development of acidic soils known as Spodosols (Table 1). Spodosols have a subsoil horizon with high contents of organic matter, aluminum, and often iron (Bs or Bhs horizons). Inceptisols (Table 1) can be common in such areas too, but they often occur on more unstable landscapes, on resistant parent materials, or where soil formation is somehow limited. The limited time of soil development results in weak color changes or soil structure development (Bw horizons) relative to parent materials (Figure 1).

Moving toward the east central United States, many soils had more time to develop on the older landscapes. Hence, subsoil clay and iron accumulations are greater (Figure 2). Warmer temperatures have increased the rates of chemical reactions, aiding weathering reactions and producing more clay minerals. Precipitation is less than at the northeast end of the transect; this difference in precipitation results in dry
periods of the year that allow for the stabilization of weathering products transported via soil water. Stabilized weathering products accumulate in the subsoil in the B horizon as secondary silicates and iron/aluminum oxides, that is, clay minerals (Bt or Btx horizons). Clay-rich, often reddish-colored, soils that have not been extensively weathered are Alfisols. Where leaching has been excessive or the parent material was low in base cations to begin with, Ultisols form (Table 1).

In the Midwestern United States, near the 100th meridian, precipitation totals drop further (Figure 4). Vegetation changes from forest to mostly grassland. The high belowground organic matter production from grassland environments that preceded modern agriculture has resulted in thick, dark A horizons rich in organic matter (Figure 2), typical of Mollisols (Table 1). Because precipitation can still exceed evapotranspiration, especially in the eastern parts of the region, carbonates (Bk horizons) and salts do not persist in the upper soil profile. Instead, they are leached to the subsoil or to the water table. However, west of the 100th meridian, decreasing amounts of precipitation result in carbonates becoming more common at increasingly shallow depths.
Figure 4  (a) Soil moisture and (b) soil temperature regimes in the contiguous United States (see SCS 1994 for an explanation of classes; see also Newhall and Berdanier 1996).
SOILS OF THE UNITED STATES

Across the far southwestern section of the transect (Figure 2), soil moisture deficits are common, as the climate becomes increasingly desert-like. Exceptions occur at higher altitudes, which are cooler and wetter. Entisols, Aridisols, and Mollisols are common orders across this section of the transect (Figure 1; Table 1). Relative to other portions of the United States, the limited precipitation and associated leaching result in high concentrations of salt and/or carbonate precipitates at the surface and in the subsoil (Bkk or Bkm horizons). Soils here can also contain substantial amounts of ash from past volcanic eruptions, which can persist in a relatively unweathered state because of the dry conditions. Vegetation in the southwestern United States has evolved to tolerate dry conditions and/or soils dominated by salts and carbonates. Where vegetative production is high, organic matter can accumulate in the soils, but not to the extent one finds in the central United States, under grassland or even forest. In this area, like many others, increasing altitude often results in a sequence of soils and vegetation similar to moving north in latitude.

Landscape evolution and its influence on the distribution of US soils

We begin by examining the distribution of lithologies and geomorphic surfaces on which soils form, using the United States Geological Survey (USGS) physiographic divisions and provinces as a guide (Fenneman 1928; Fenneman and Johnson 1946) (Figure 5). The states of Alaska and Hawai‘i will be examined separately.

Appalachian Highlands division

The Appalachian Highlands (Figure 5; Table 2) are dominated by sedimentary rock from
Pennsylvania south across the division, and crystalline rock from New England through the Piedmont and Blue Ridge provinces (Fenneman 1928). The division’s northern end (roughly bisecting Pennsylvania and the upper half of New Jersey) has experienced numerous glaciations, with the effects of the last glaciation still evident (Ciolkosz et al. 1989). North of the glacial boundary, till, outwash, and loess are common parent materials, and many soils are wet owing to high water tables. Moving south along the region’s major tributaries that were fed by Pleistocene glaciers (the Susquehanna and Delaware rivers, for example), outwash and loess become the most common parent materials. Virtually all backslopes and footslopes in the division are mantled by mass weathering deposits known as colluvium (Ciolkosz et al. 1979). The division’s soil moisture regime is mostly typic udic with some aquic areas in lowlands. The soil temperature regime is thermic in the south, mesic in the center of the division, and frigid in the north (Figure 4).

Throughout the Adirondacks and New England (Figure 5), the effects of glaciation are ubiquitous. Glacial sediments are thus the primary parent material. An exception is in the Adirondacks where, mainly due to glacial erosion, igneous and metamorphic rocks of the Canadian Shield are exposed (Fenneman 1928). Spodosols are common in these provinces, especially on sandier sediments. Inceptisols occur on terraces, steeper slopes, resistant parent materials, or where bedrock is shallow. Entisols occur on the least stable landscapes (floodplains, very steep slopes) or very sandy parent materials. Histosols are common in very wet areas such as bogs and swamps. Forests cover much of the region, but pasture and some row crop agriculture are other land uses (USDA–NRCS 2006).

Across the Appalachian Plateau, Blue Ridge, and Piedmont provinces (Figure 5), Inceptisols, Ultisols, and Alfisols are the most common soil orders. Inceptisols occur on backslopes in colluvium and along floodplains, that is, on young landscapes. Ultisols can form on stable backslopes and footslopes, or in valley locations. Ultisols in this area form on base cation-poor, or very weathered, parent materials. Alfisols have also formed in lowlands, but from parent materials that are base cation-rich, such as limestone or calcareous shale, weakly metamorphosed or crystalline rocks, and some glacial tills in the north. Spodosols have formed on stable Appalachian Plateau landforms in sandy soils, and where precipitation drives deep leaching in porous materials. Timber production is extensive on the Appalachian Plateau, but there is also a mineral resource extraction economy and some row crop and pasture agriculture. Across the Blue Ridge and Piedmont provinces, row crop and pasture agriculture are more extensive. Recreation is a common land use in both provinces (USDA–NRCS 2006).

Atlantic Plain division

Along the Atlantic Plain (Figure 5; Table 2), soils have formed largely in sediments eroded off the continent. However, along the Mississippi River, thick and extensive loess deposits can be found. The soil moisture regime inland is typic udic, but aquic conditions occur near coastal areas, along some major rivers, and on low relief areas of uplands (Figure 4). The soil temperature regime is mesic in the north of the division, thermic in the south central, and hyperthermic in the south (Figure 4). Because many soils have poor drainage, organic materials accumulate, resulting in the formation of Histosols. Entisols occur on young deposits along floodplains, and on unstable beach deposits. Adjacent areas of higher elevation have a cover of Inceptisols and Mollisols. Mollisols, derived under former

SOILS OF THE UNITED STATES
Table 2  Area of conterminous soil orders by US Geological Survey physiographic provinces.

<table>
<thead>
<tr>
<th>Division</th>
<th>Province</th>
<th>Alfisols</th>
<th>Entisols</th>
<th>Histosols</th>
<th>Inceptisols</th>
<th>Mollisols</th>
<th>Spodosols</th>
<th>Andisols</th>
<th>Ultisols</th>
<th>Aridisols</th>
<th>Vertisols</th>
<th>Province soil extent (km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appalachian Highlands</td>
<td>Adirondacks</td>
<td>1.3</td>
<td>0.3</td>
<td>0.4</td>
<td>9.5</td>
<td>0.0</td>
<td>88.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>25283</td>
</tr>
<tr>
<td>Appalachian Highlands</td>
<td>Appalachian Plateau</td>
<td>24.3</td>
<td>0.8</td>
<td>0.1</td>
<td>35.1</td>
<td>0.0</td>
<td>0.8</td>
<td>0.0</td>
<td>38.9</td>
<td>0.0</td>
<td>0.0</td>
<td>292007</td>
</tr>
<tr>
<td>Appalachian Highlands</td>
<td>Blue Ridge</td>
<td>2.1</td>
<td>0.0</td>
<td>0.0</td>
<td>58.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>39.4</td>
<td>0.0</td>
<td>0.0</td>
<td>46635</td>
</tr>
<tr>
<td>Appalachian Highlands</td>
<td>New England</td>
<td>0.9</td>
<td>4.4</td>
<td>0.5</td>
<td>33.5</td>
<td>0.0</td>
<td>59.9</td>
<td>0.0</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>168068</td>
</tr>
<tr>
<td>Appalachian Highlands</td>
<td>Piedmont</td>
<td>14.0</td>
<td>0.6</td>
<td>0.0</td>
<td>5.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>79.6</td>
<td>0.0</td>
<td>0.0</td>
<td>199572</td>
</tr>
<tr>
<td>Appalachian Highlands</td>
<td>St Lawrence Valley</td>
<td>16.1</td>
<td>1.7</td>
<td>3.2</td>
<td>39.1</td>
<td>0.0</td>
<td>39.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>9724</td>
</tr>
<tr>
<td>Appalachian Highlands</td>
<td>Valley Ridge</td>
<td>12.5</td>
<td>1.7</td>
<td>0.1</td>
<td>46.8</td>
<td>0.0</td>
<td>0.2</td>
<td>0.0</td>
<td>38.6</td>
<td>0.0</td>
<td>0.0</td>
<td>120011</td>
</tr>
<tr>
<td>Atlantic Plain</td>
<td>Coastal Plain</td>
<td>25.4</td>
<td>7.9</td>
<td>3.4</td>
<td>6.5</td>
<td>4.0</td>
<td>5.0</td>
<td>0.0</td>
<td>37.9</td>
<td>0.1</td>
<td>9.8</td>
<td>1103050</td>
</tr>
<tr>
<td>Interior Highlands</td>
<td>Ouachita</td>
<td>17.0</td>
<td>4.7</td>
<td>0.0</td>
<td>0.0</td>
<td>1.6</td>
<td>0.0</td>
<td>0.0</td>
<td>76.1</td>
<td>0.0</td>
<td>0.6</td>
<td>53074</td>
</tr>
<tr>
<td>Interior Highlands</td>
<td>Ozark Plateaus</td>
<td>64.0</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>2.9</td>
<td>0.0</td>
<td>0.0</td>
<td>32.1</td>
<td>0.0</td>
<td>0.0</td>
<td>125868</td>
</tr>
<tr>
<td>Interior Plains</td>
<td>Central Lowland</td>
<td>35.1</td>
<td>5.2</td>
<td>2.5</td>
<td>3.0</td>
<td>50.3</td>
<td>2.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.1</td>
<td>141863</td>
</tr>
<tr>
<td>Interior Plains</td>
<td>Great Plains</td>
<td>8.9</td>
<td>23.4</td>
<td>0.0</td>
<td>5.3</td>
<td>50.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>9.2</td>
<td>2.2</td>
<td>1379319</td>
</tr>
<tr>
<td>Interior Plains</td>
<td>Interior Low Plateaus</td>
<td>58.9</td>
<td>1.2</td>
<td>0.0</td>
<td>10.3</td>
<td>1.4</td>
<td>0.0</td>
<td>0.0</td>
<td>28.2</td>
<td>0.0</td>
<td>0.0</td>
<td>1363754</td>
</tr>
<tr>
<td>Intermontane Plateaus</td>
<td>Basin and Range</td>
<td>3.0</td>
<td>22.7</td>
<td>0.0</td>
<td>3.3</td>
<td>19.5</td>
<td>0.0</td>
<td>0.7</td>
<td>0.0</td>
<td>50.2</td>
<td>0.6</td>
<td>888889</td>
</tr>
<tr>
<td>Intermontane Plateaus</td>
<td>Colorado Plateau</td>
<td>12.7</td>
<td>36.4</td>
<td>0.0</td>
<td>2.6</td>
<td>19.4</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>28.5</td>
<td>0.3</td>
<td>333082</td>
</tr>
<tr>
<td>Intermontane Plateaus</td>
<td>Columbia Plateau</td>
<td>2.3</td>
<td>3.1</td>
<td>0.0</td>
<td>1.3</td>
<td>61.7</td>
<td>0.0</td>
<td>5.5</td>
<td>0.0</td>
<td>26.0</td>
<td>0.1</td>
<td>281062</td>
</tr>
<tr>
<td>Laurentian Upland</td>
<td>Superior Upland</td>
<td>31.5</td>
<td>2.3</td>
<td>7.8</td>
<td>18.3</td>
<td>0.1</td>
<td>40.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>109119</td>
</tr>
<tr>
<td>Pacific Mountain System</td>
<td>Cascade Sierra Mountains</td>
<td>13.9</td>
<td>7.0</td>
<td>0.3</td>
<td>27.0</td>
<td>13.9</td>
<td>9.5</td>
<td>21.9</td>
<td>5.7</td>
<td>0.8</td>
<td>0.2</td>
<td>166871</td>
</tr>
<tr>
<td>Pacific Mountain System</td>
<td>Lower Californian</td>
<td>40.4</td>
<td>32.0</td>
<td>0.0</td>
<td>0.3</td>
<td>24.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>2.2</td>
<td>7655</td>
</tr>
<tr>
<td>Pacific Mountain System</td>
<td>Pacific Border</td>
<td>20.6</td>
<td>12.6</td>
<td>0.3</td>
<td>27.6</td>
<td>19.6</td>
<td>1.7</td>
<td>6.9</td>
<td>4.3</td>
<td>2.8</td>
<td>3.6</td>
<td>279943</td>
</tr>
<tr>
<td>Rocky Mountain System</td>
<td>Middle Rocky Mountains</td>
<td>9.5</td>
<td>22.9</td>
<td>0.0</td>
<td>17.5</td>
<td>44.0</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>5.7</td>
<td>0.0</td>
<td>125099</td>
</tr>
<tr>
<td>Rocky Mountain System</td>
<td>Northern Rocky Mountains</td>
<td>10.3</td>
<td>4.2</td>
<td>0.0</td>
<td>40.8</td>
<td>29.8</td>
<td>0.0</td>
<td>12.3</td>
<td>0.0</td>
<td>2.6</td>
<td>0.0</td>
<td>254475</td>
</tr>
<tr>
<td>Rocky Mountain System</td>
<td>Southern Rocky Mountains</td>
<td>31.3</td>
<td>3.9</td>
<td>0.0</td>
<td>20.4</td>
<td>31.5</td>
<td>0.3</td>
<td>0.1</td>
<td>0.0</td>
<td>12.4</td>
<td>0.0</td>
<td>117721</td>
</tr>
<tr>
<td>Rocky Mountain System</td>
<td>Wyoming Basin</td>
<td>2.6</td>
<td>45.3</td>
<td>0.0</td>
<td>1.7</td>
<td>14.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>35.9</td>
<td>0.0</td>
<td>101895</td>
</tr>
</tbody>
</table>

Source: STATSGO soil data provided by Sharon Waltman, USDA-NRCS.
soils of the united states

grasslands, are most extensive across Florida, and currently produce a wide variety of agricultural crops. The division’s Spodosols occur in very poorly drained to moderately well-drained conditions, and dominantly in sandy but wet parent materials (Harris 2000). Where sediments are less sandy and time has allowed for more weathering, clay minerals have accumulated in B horizons. Such soils fall within the Alfisol and Ultisol orders and often occur on terraces or wet areas of floodplains or depressions. Ultisols occur inland on stable landscapes that are old and weathered, or on base cation-poor parent materials. Alfisols form on base cation-rich sedimentary rocks, provided the mineralogy does not have the shrink–swell capability of Vertisols. Vertisols are common in the southwest of the division, where clay-rich, calcareous parent materials occur, commonly derived from sedimentary rock or unconsolidated sediments. The mineralogy of these parent materials imparts a high shrink–swell capacity to the soils, as they dry and wet, respectively. Where soils are not too wet, crop production is diverse (corn, sorghum, cotton, bean, rice, hay, truck crops, and citrus). Rangeland is present in the southwestern parts of the division, and timber production is extensive in the south and southeast (USDA-NRCS 2006).

Interior Highlands division

The Interior Highlands (Figure 5; Table 2) is a remnant of the Ouchita–Ozark Highlands; the northern end of the division is the Ozark Plateaus province and the southern end is the Ouchita province. The division’s topography originates from the dissection of an old plateau, resulting in slight to steep relief. Carbonate and sedimentary rock (sandstone and shale) predominate. Although soils are often residual and, hence, relate closely to their local lithology, soils in the northern Ozark Plateaus province often have formed in 1–5 m loess that was derived from the Missouri and Mississippi rivers or their major tributaries (Bettis et al. 2003).

The division as a whole has a typic udic soil moisture regime, and in general spans the boundary between mesic and thermic soil temperature regimes (Figure 4). Across the Interior Highlands, approximately half of soils are Alfisols (most in the Ozark Plateaus province) and half Ultisols. Alfisols and Ultisols have formed on stable landscapes underlain by limestone and cherty limestone. Mollisols occur in the western or southwestern sections, on terraces or broad open meadows within forests. Deciduous forest predominates here, but coniferous forest can be found at high elevations. Agriculture is most successful where topography is least limiting; hence, pasture (versus cropland) is the more common agricultural practice (USDA-NRCS 2006).

Interior Plains division

The Interior Plains (Figure 5; Table 2) form an extensive portion of the conterminous United States, including the Great Plains, Central Lowlands, and the Interior Low Plateaus. The division as a whole has low relief (except near river valleys) and consists largely of sedimentary rocks derived from sediments eroded off the Rocky Mountains to the west, and the Appalachian Mountains and Ozark–Ouachita Mountains to the east and south, respectively (Fenneman 1928). The Interior Low Plateaus has the most varied relief, with steep to gently rolling topography. Because it was never glaciated, its topography is reflective of differential erosion of the sedimentary rocks. The Great Plains province has been extensively glaciated in the very north, as was the upper half of the Central Lowland province. As a result of glaciation, the topography of these two provinces is flatter than
SOILS OF THE UNITED STATES

the Interior Low Plateaus. Loess is an extensive parent material across the division (Bettis et al. 2003; Schaezl and Attig 2013). Great Plains loess deposits can be >20 m thick in Nebraska, along the Iowa–Nebraska border, and along the Mississippi River and its tributaries, but across the Central Lowland and Interior Low Plateaus, these deposits typically are <5 m thick. Across central and northwest Nebraska is the Sand Hills region. This unique area features Quaternary sand dunes derived from the underlying bedrock.

The division’s soil moisture regime varies from typic udic in the east to ustic aridic in the west. The temperature regime ranges from frigid in the north, mesic in the center, to thermic in the south (Figure 4). Mollisols are associated with stable landscape positions where prairie ecosystems existed prior to European settlement. Alfisols have formed on stable landscape positions, typically on sites of former forest and where parent materials are base cation-rich. Ultisols occur on the older soils of the Interior Low Plateaus on stable landforms. Entisols occur where soil-forming materials are young, for example, on floodplains and areas of steep slope. Inceptisols are common on young but stable landscapes. Aridisols occur in the Great Plains province where landforms are stable and the soil moisture regime is aridic. Agriculture is the division’s primary economy and consists of diverse cropland and grassland (USDA-NRCS 2006). Native prairie grassland is much less extensive today than prior to European settlement. Mineral, oil, and gas production are extensive in various parts of the division’s northern and southern ends.

Intermontane Plateau division

The rugged Intermontane Plateau division (Figure 5; Table 2) consists of the Basin and Range province, and the Colorado (Arizona, New Mexico, Colorado, and Utah) and Columbia (Oregon, Washington, and southern Idaho) Plateaus provinces. Topography across the Basin and Range and Colorado Plateaus provinces is the product of tectonic uplift and subsequent erosion, while the Columbia Plateaus province is derived from extensive lava flows. Catastrophic water release from Pleistocene glacial lakes dramatically shaped some of the division’s topography and parent materials. Examples include the breach of Glacial Lake Bonneville (northern Utah), which drained northwest into the proximity of the present-day Snake River of southern Idaho, and that of Glacial Lake Missoula in western Montana, which drained into Glacial Lake Columbia out of northern Idaho into eastern Washington and northern Oregon. This complex geomorphic and geologic history has resulted in soils forming from a variety of materials.

The Columbia Plateau province consists of broad areas of flood basalt, bordered to the west by the Cascade Mountains and to the east by the Rocky Mountains. Most upland soils are shallow to bedrock. Valley alluvial sediments are deep and can be composed of volcanic, metamorphic, and igneous sediment derived from the weathering of the surrounding mountain systems (Fenneman 1928). Loess is extensive along the major river valleys and tributaries of the Columbia River, especially in the Palouse and Nez Perce Prairie areas to the east of the province, and along the Snake River to the south. Glaciofluvial and glaciolacustrine parent materials can also be found. The soil moisture regime is dominantly xeric to xeric aridic, with some small pockets of aquic and typic udic. The soil temperature regime is mostly mesic, but frigid to cryic at higher elevations (Figure 4). Mollisols are common along floodplains and terraces, but most common at high elevations under
SOILS OF THE UNITED STATES

Tree and shrub communities. Aridisols occur on stable landscapes at low elevations where the soil moisture regime is aridic. Smaller extents of Alfisols, Entisols, Inceptisols, and Andisols occur. Alfisols occur at elevations above Aridisols and below that of higher-altitude Mollisols. Although Andisols are not extensive, volcanic ash occurs in varying amounts across the province’s soils. Agriculture is a mix of grazing and cropland, and forest and mineral/gas extraction are also important. Recreation is an important use of forested areas (USDA-NRCS 2006).

Rugged, deeply incised river valleys dominate the Colorado Plateau province. A classic example is the Colorado River’s Grand Canyon. Although sedimentary rocks predominate, volcanic rock associated with lava flows is also common (Fenneman 1928). Like much of the western United States, the province’s soil moisture and temperature regimes vary with altitude. The driest soil moisture regime, typic aridic, is found at low elevations, with soil moisture increasing with elevation (Figure 4). The soil temperature regime is mesic at low elevations and frigid to cryic at higher elevations (Figure 4). Entisols predominate across young geomorphic landscapes but extensive areas of Aridisols are common on stable landscape positions, given the aridic soil moisture regime. Mollisols are found at higher elevations under tree, shrub, and some grass cover. Alfisols occur on stable landscape positions (Table 2) and at higher elevations than Aridisols. The occurrence of Alfisols is, in many places, attributed to a past, wetter period of soil formation, during the Pleistocene. Under the present drier conditions, many of these soils have accumulated calcium carbonate in their B horizons. Much of the region is rangeland used for cattle grazing. At high elevations, forest and shrub cover are more common. Much of the land is used for timber, oil and gas production, and recreation (USDA-NRCS 2006).

The Basin and Range province consists of rugged topography with dramatic changes in elevation, with north–south trending mountains (some with summits >3200 m) and broad valleys of similar orientation. Although sedimentary rocks predominate, volcanic materials are common (Fenneman 1928). During the Pleistocene, a vast system of lakes existed across the province; Utah’s Great Salt Lake is an example of a modern remnant of the former Pleistocene Lake Bonneville. Today, many of these former lakes are dry or ephemeral lakebeds called playas. The province’s basins consist largely of alluvium derived from the adjacent mountains, but in playas, salty lake sediment is found, often covered with isolated sand dunes. Soil development is strongly tied to the age of the landform (Peterson 1981); thus older landforms have older, better developed soils with increasingly rich CaCO3 or Si subsoil accumulations. Soil morphology frequently reflects the former, usually wetter, soil-forming environment.

The Basin and Range province supports several types of deserts; their differentiation is based on differences in moisture and/or temperature, topography, and resulting flora and fauna. Deserts of the province’s southern end consist of the Mojave, Sonoran, and Chihuahuan, and across the province’s central and northern area is the Great Basin Desert. The Basin and Range province’s current soil moisture and temperature regimes vary with altitude. The driest soil moisture regime, typic aridic, is found at low elevations, with increasingly wetter conditions at higher elevations (Figure 4). The soil temperature regime is hyperthermic and thermic at the lowest elevations in the Mojave Desert of southern Nevada and California, and mesic at low elevations in the province’s center. Frigid and cryic soil temperature regimes occur at higher elevations, which along with the additional moisture provided at altitude support
SOILS OF THE UNITED STATES

forest cover more typical of northern locations in Montana, Wyoming, or Idaho. Aridisols make up nearly half of the province’s soils and occur on stable, older landscape positions. Some of the oldest soils in the United States are Aridisols found in the Mojave Desert, on Mormon Mesa near Mesquite, Nevada (Brock and Buck 2009). Entisols are common on unstable, recently deposited sediments across steep slopes, or on floodplains. Mollisols can be found at higher elevations (see Tables 3 and 4, below), where soil moisture is greater and cooler temperatures persist, but they can occur along river systems on terraces or floodplains as well. Alfisols are largely a product of the Pleistocene and occur at higher elevations. Due to the current dry climate, calcium carbonate accumulations are common in the upper B horizons of some Alfisols. At high elevations, Alfisols may currently be forming (Elliott and Drohan 2009). Inceptisols occur at higher elevations or along major river systems on terraces, but lack the marked clay accumulation in B horizons that are common to Alfisols. The province’s Vertisol extent is limited to stable landscapes and base cation-rich parent materials. At low elevation, rangeland for grazing is prevalent, along with oil and gas production, recreation, and some row and truck crop agriculture. At high elevations, forest and shrub cover are more common, fueling timber extraction and recreational land uses (USDA-NRCS 2006).

Laurentian Upland division

The Laurentian Upland division (Figure 5; Table 2) is a low relief area that spans northern Wisconsin and Minnesota; its sole province is the Superior Upland. Nearly the entire division was glaciated during the Pleistocene, and thus most parent materials are glacially derived, unless bedrock is exposed. The area is the southernmost extension of the Canadian Shield, which consists mainly of igneous and metamorphic rocks (Fenneman 1928). Glaciation filled in many former bedrock valleys, resulting in a low relief, rolling landscape. Glacial geomorphology (outwash plains, drumlins, moraines, etc.) and associated parent materials (till, outwash, lacustrine deposits, loess, etc.) set the stage for soil formation across the division, and thus soils that have formed in multiple, stacked, parent materials are common (Schaetzl 1998). The generally level terrain has led to the formation of many areas with wet soils, commonly in bogs and swamps. The division’s soil moisture regime (Figure 4) is mainly typic udic, but aquic conditions are found in lowlands. The soil temperature regime (Figure 4) is frigid. Spodosols are common in the division and form in a number of sandy to loamy parent materials (e.g., alluvium, loess, outwash). Alfisols form across a range of parent materials, textures, and drainage classes, with the one common feature being a stable landscape that has allowed enough time for development of clay accumulations in the B horizon. Unless the parent material consists of finer materials to begin with (e.g., silt-sized loess or glaciolacustrine materials), the clay contents

<table>
<thead>
<tr>
<th>Soil order</th>
<th>Area (km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gelisols</td>
<td>338 256</td>
</tr>
<tr>
<td>Spodosols</td>
<td>329 952</td>
</tr>
<tr>
<td>Inceptisols</td>
<td>269 060</td>
</tr>
<tr>
<td>Entisols</td>
<td>46 358</td>
</tr>
<tr>
<td>Histosols</td>
<td>45 703</td>
</tr>
<tr>
<td>Andisols</td>
<td>34 725</td>
</tr>
<tr>
<td>Mollisols</td>
<td>12 549</td>
</tr>
</tbody>
</table>

Table 4  Area of the six major Hawaiian Islands occupied by each soil order (%).

<table>
<thead>
<tr>
<th>Soil order</th>
<th>Hawaiian Islands (%)</th>
<th>Area</th>
<th>km²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entisols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaua'i</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O'ahu</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moloka'i</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lāna'i</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maui</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hawai'i</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>157</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Histosols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inceptisols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td></td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>39</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>Mollisols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td></td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>663</td>
<td></td>
<td>663</td>
</tr>
<tr>
<td>Spodosols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Andisols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td></td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>3312</td>
<td></td>
<td>3312</td>
</tr>
<tr>
<td>Ultisols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>363</td>
<td></td>
<td>363</td>
</tr>
<tr>
<td>Aridisols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>279</td>
<td></td>
<td>279</td>
</tr>
<tr>
<td>Vertisols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>138</td>
<td></td>
<td>138</td>
</tr>
<tr>
<td>Oxisols</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>44</td>
<td></td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td></td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>39</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>48</td>
<td></td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>876</td>
<td></td>
<td>876</td>
</tr>
</tbody>
</table>


of soils in this division are lower than in Alfisols typical of the midwestern or eastern United States. Entisols have typically formed in sandy glaciofluvial deposits on floodplains. Inceptisols can occur in loamy deposits on ground moraines or end moraines, but also in sandy deposits on terraces (Table 2). Histosols are found in bogs and swamps (Table 2). Forestry and recreation are the most extensive land uses across the region (USDA–NRCS 2006).

Pacific Mountain System division

The Pacific Mountain System (Figure 5; Table 2) has extensive relief. Tectonic and volcanic activity continues to shape this landscape. Three provinces encompass the division: the Pacific Border (coastal and running north to south), the Cascade–Sierra Nevada Mountains (inland and running north–south), and Lower California in the far south. The Cascade Mountains consist of volcanic rocks, whereas granitic rocks dominate the Sierra Nevada Mountains. Alpine glaciers existed in the Sierra Nevada Mountains during the Pleistocene, with Yosemite National Park being the most prominent example. Glaciofluvial deposits and loess can be found along major rivers, especially the Columbia River. The Pacific Border province (the largest of the three) is very diverse. Along its northern end, in the Oregon Coast Range, volcanic rocks dominate, but sedimentary rocks and marine terraces occur along the coast, and glacial parent materials occur widely around Puget Sound. In the south central part of the province are the California Coast Ranges (generally north–south trending mountains) and the Los Angeles Range (east–west trending mountains). Soils in coastal areas can be formed in weathered sedimentary rocks and/or on marine terraces. Coastal valleys are often filled with unconsolidated coastal sediments or sediments derived from the adjacent mountains. Lithologies are a mix of metamorphic, granitic, and sedimentary rocks (Fenneman 1928). Many soils near areas of volcanic activity have volcanic ash parent materials, and soils near the Columbia River in the northern end of the division can be derived from loess or have some loess influence.

The division’s soil moisture and temperature regimes are complex and generally drier to the
SOILS OF THE UNITED STATES

south, with hot and dry summers throughout (Figure 4). Soil diversity reflects the division’s complicated physiographic history (Table 2). Entisols are on the youngest landscapes and steep slopes. Inceptisols are most common where landscapes are relatively young, for example, on fluvial or colluvial sediments. Mollisols occur in valleys on the coastal side of the division’s mountain ranges. Andisols are found in areas of recent volcanic activity, most commonly in the Cascade Sierra province. Alfisols are far more common than Ultisols and occur on a wide range of base cation-rich parent materials. Ultisols occur mostly in the northern end of the division, in colluvium or residuum derived from sedimentary and basic igneous rocks. Volcanic ash is found in most soils, and most recently came from the eruption of Mount Mazama (presently Crater Lake, Oregon), nearly 7700 years ago.

The division’s soil moisture and temperature regimes are highly diverse and vary greatly across even short distances. Higher-elevation areas tend to be cooler and wetter, and lower-elevation areas drier and hotter (Figure 4). Mollisols are most common where former grasslands existed at lower elevations, or at higher elevation under shrub and tree communities. Inceptisols are extensive across the division’s steeper but stable landscapes, especially in the Northern Rocky Mountains province where clay accumulation has not been extensive. Entisols form on young or unstable landscapes, such as on the steep slopes and floodplains. Alfisols form on base cation-rich parent materials across stable landscapes, and Aridisols are found at the lowest elevations on stable landscapes, but in an aridic soil moisture regime. The division’s soils are frequently influenced by volcanic ash, but not necessarily enough to always result in Andisols. Andisols form where volcanic materials predominate, as in the Northern Rocky Mountains. Some Spodosols occur on sandy parent materials where precipitation is substantial. Rangeland is extensive and hay and row crops can be supported, often with irrigation. Forested areas are important for timber and for recreation (USDA-NRCS 2006).

Rocky Mountain System division

The Rocky Mountain System division (Figure 5; Table 2) is composed of steep, formerly glaciated mountains, often consisting of sedimentary rocks; however, granitic (especially volcanic rocks) and metamorphic rocks are most common. Alluvial sediments and a variety of glacial deposits (lacustrine, till, glaciofluvial, etc.) can also be locally important parent materials. Volcanic ash is found in most soils, and most recently came from the eruption of Mount Mazama (presently Crater Lake, Oregon), nearly 7700 years ago.

The division’s soil moisture and temperature regimes are highly diverse and vary greatly across

Alaska

Alaska has a highly diverse geomorphic history, with both active glaciation and volcanism. Pleistocene glaciers covered many, but not all, areas of Alaska (Fenneman 1928; USDA-NRCS 2006). Where erosion and/or the burial of glacial deposits has occurred, colluvium and alluvium are the main parent materials. Glacial deposits that persist are varied and include lacustrine and glaciofluvial sediment, loess, and till. Holocene volcanic activity has contributed much ash to the landscape, and Holocene wind action has reworked dry stream sediments across the landscape. Glaciers still persist
in Alaska, and permafrost can occur but is typically not continuous. Across the coldest areas, periglacial features, for example pingos, gelifluction lobes, and patterned ground, are common.

Soil patterns across Alaska (Table 3) are perhaps best explained by patterns of temperature, length of winter (Walker et al. 2004), landscape stability, and vegetative production. Gelisols dominate in the far north, where soil processes are severely limited by cold conditions. Permafrost can cause significant cryoturbation, resulting in convoluted subsoil horizons. Where temperatures are not so low and vegetation production is not limited, organic matter can accumulate in soils, leading to the formation of organic-matter-rich Gelisols. Central Alaska’s complex soil temperature regime results in an intermixture of Gelisols, Inceptisols, and Spodosols. Although Gelisols occur in the coldest areas, where permafrost is absent but landscapes are quite young and/or precipitation is limited, Inceptisols dominate. Where precipitation is ample and temperatures support forest across sandy parent materials, Spodosols dominate. Where landscapes are persistently wet and organic matter production is high, Histosols can be found; the colder temperatures help reduce organic matter decomposition. Volcanic activity in southern Alaska can result in Andisols where volcanic ash is dominant. However, where volcanic ash is less prevalent and frozen soil conditions are absent, high precipitation totals, coniferous vegetation, and sandy parent materials result in the formation of Spodosols. Agriculture is diverse but constrained owing to cold temperatures, high precipitation, or terrain. Pasture or hay production is common, but timber production is predominant (USDA-NRCS 2006).

Hawai’i

Hawai’i’s volcanic islands have some of the nation’s most fascinating soils and soil patterns (Figure 1). Despite its small size, Hawai’i contains 10 of the 12 soil orders (Deenik and McClellan 2007). Given that the parent material of the islands is a constant (basalt), climate and topography are the main drivers of pedogenesis. The youngest island, Hawai’i (ca. 500,000 years old), has substantial areas of nonsoil (bare bedrock) owing to recent volcanic activity. Across all the islands, rainfall is highest on the windward – north and northeast – sides, with much less precipitation on the leeward sides. Precipitation also varies with altitude and is typically lower on the coast and increases with elevation (USDA-NRCS 2006).

Andisols are the most common soil order in Hawai’i (Table 4) and form near former volcanic vents. These soils are most common on the younger Hawaiian islands of Hawai’i and Maui (Macdonald, Abbott, and Peterson 1983). Windward or lower elevation areas on several islands sustain higher organic matter production, and if supported by low to moderate precipitation, produce Mollisols. Interestingly, the island with the smallest extent of Mollisols is Hawai’i. Could this suggest a minimum time period for Mollisol establishment? Wet or wet and cool areas of the islands can support Histosols, and the greater the precipitation the more acidic the Histosols tend to be. Many Histosols occur on the island of Hawai’i, over shallow lava flows that perch water. Oxisols can form either on low-elevation dry areas, or on high-elevation wet areas. Oxisols are more common on the older islands. The amount of rainfall determines the amount soil development in Oxisols, and thus, the level of fertility. Older soils are generally less fertile. Entisols are common in coastal areas and develop often from coral limestone. They also are found in drier areas where alluvium is
The soils of the United States are one of the country’s most important resources. It depends on them for food, water filtration, fiber production for clothing and manufacturing, wastewater filtration, and ecosystem support. An increased need for greater crop production and the development of physical infrastructure to meet a rapidly growing human population places substantial pressure on soils worldwide. Although anthropogenic disturbance has not been specifically separated out in this entry as a driver of soil genesis, it is often a dominant soil-forming factor; human actions often result in a change in soil character. A substantial challenge facing the United States will be improved conservation measures to prevent further degradation of soil properties, so as to meet the needs of an ever-growing population.

**SEE ALSO:** Agricultural environments; Agricultural geography; Environmental history; Environmental management; Ethnophysiography; Food security; Soil mapping and maps; Soil taxonomy and soil classification; Soils and weathering; Spatial concepts

**References**


Soils of tropical landscapes

S.W. Buol
North Carolina State University, USA

All soils perform functions critical to shaping both natural and human ecosystems. They provide a rooting medium to physically support plants. Soils acquire temperature and natural moisture characteristics closely related to the climatic conditions at that location. In addition to carbon, oxygen, and hydrogen acquired from air and water, soils provide the other essential nutrient elements required for physiological functions of plants and animals. Most of these elements, except nitrogen, are derived from minerals in the soil and taken up as inorganic ions. Although plants obtain nitrogen from the soil, it is derived from the air by either micro-organisms symbiotically associated with leguminous plants or free living in the soil. It is available to plants as inorganic ions in the soil as the micro-organisms and plant residues decompose. Many tropical soils contain few nutrient bearing minerals.

Contrasts between tropical and temperate latitude soils: what makes tropical soils unique?

Minimal seasonal variation in soil temperatures is universal throughout soils in tropical latitudes. Only a 6°C or less temperature difference exists between the average soil temperature of June, July, and August and the average temperature of December, January, and February for almost all soils in the tropics (USDA Natural Resources Conservation Service 1999). This nearly uniform soil temperature regime contrasts with temperate latitudes where natural vegetation and agronomic practices have to accommodate much colder winter temperatures and utilize warm summer months for crop production. At the highest elevations in the tropics where mean annual soil temperatures are below 10°C the growing of almost all food crops is precluded. At somewhat lower elevations with mean annual soil temperatures of below about 15°C plants are subject to sporadic freezing, limiting production of many food crops and certain natural species.

A significant proportion, but not all, of the soils in the tropics are formed in materials that have been on the land surface for longer expanses of geologic time than in temperate latitudes. During the extensive exposure to water enriched with organic acids, and repeated cycles of erosion and deposition, almost all of the more soluble minerals have decomposed and their by-products removed, leaving behind soil parent materials enriched in quartz, iron and aluminum oxides, and kaolinite clay. Soils formed in such materials are nearly devoid in minerals bearing elements necessary for plant growth.

Another less commonly recognized feature of most soils in the tropics is their lack of silt-sized particles, so prevalent in the loess-derived soils in the Midwestern United States and in Eastern Europe. Silt-rich soils retain approximately 0.2 cm of plant available water per centimeter of soil depth, whereas silt-poor soils retain half that much. As a result, plants growing on silt-poor tropical soils are more adversely affected by droughts during the growing season.
Soils of Tropical Landscapes

Soil moisture regimes in tropical landscapes

Weather conditions vary from year to year in the tropics, but the climatic probability of rainfall dictates the probable amount and duration of plant available water in the soil. Probable soil moisture is a characteristic of individual soils because each soil has a fixed geographic location and topographic position in the landscape. Soils are thus classified by the rainfall and potential evapotranspiration conditions at their location in normal years, as well as the seasonal depth to the ground water table (USDA Natural Resources Conservation Service 1999). Aquic soil moisture regimes identify soils that are periodically saturated with water; engineered drainage is required for growing most food crops. Aquic conditions are present in localized landscape depressions within upland areas and near rivers and lakes. Aridic soil moisture regimes identify soils with less than three consecutive months of adequate plant available soil water, making the growth of most food crops without irrigation impossible. Ustic soil moisture regimes identify soils where normal rainfall patterns provide plant available soil water for the growing of one or two food crops annually. Most of these areas have a single reliable wet season extending from three to nine months each year, although notable contrasts are present in parts of eastern Africa where two short growing periods, separated by a dry season, occur. Udic soil moisture regimes identify soils that annually have less than three months that are too dry for growing food crops; two crops are usually possible each year. Perudic soil moisture regimes identify soils where precipitation exceeds potential evapotranspiration each month of the year. Lack of a dry season in which to harvest grain limits food production on these soils to fruits and nuts and human populations are low (see Van Wambke 1981, 1982, and 1985 for soil moisture regime location maps).

Soil chemical, physical, and mineral properties of soils in the tropics

The mineralogical, chemical, and physical properties of soils are in large part determined by the geologic composition of materials in which they form. Perhaps the most common generalization of soils in the tropics centers on the characteristics of Oxisols (USDA Natural Resources Conservation Service 1999) or ferrasols (IUSS Working Group WRB 2006). Oxisols, sometimes known by other historical names such as latosols and laterites, dominate landscapes in about 24% of the tropics (Van Wambke 1991, 343). Oxisols are loamy or finer textured, physically friable soils of low bulk density. They contain mainly quartz, kaolinite, gibbsite, and iron oxide minerals. Most are various shades of yellow or red in color and contain only sparse amounts of minerals containing elements necessary for plant growth. They have nearly uniform texture with depth and a strong granular structure that permits rapid percolation. Most are on level to rolling landscapes with slopes of less than 10%.

Oxisols are present in all soil moisture regimes in the tropics. Extensive areas of Oxisols are present in central Brazil and the lower portion of the Amazon basin. They are also extensive in the central portion of the Congo basin and other interior basins in Africa, parts of tropical Asia, and northern Australia. In South America, most are formed in nutrient poor sediments, often several meters in thickness deposited during the warm, humid climate of the Cretaceous period and early Cenozoic era (Orme 2007) (Figure 1).

Most Oxisols are some of the least fertile soils in the tropics (Buol 2007). Being formed in largely inert parent materials, these soils have
few morphological features indicative of present climatic conditions (USDA Natural Resources Conservation Service 1999; Buol et al. 2011). However, not all Oxisols are naturally infertile. Some, formed in ancient sediments derived from basic rocks like basalt, have a high base cation saturation percentage; in moist areas these soils usually support lush forest vegetation. Many of these more naturally fertile Oxisols have been cultivated for food crop production. The least fertile Oxisols naturally support only savanna vegetation so deficient in calcium and phosphorus that grazing mammals cannot maintain proper bone structure and it is impossible to grow food crops without additions of externally derived sources of plant essential nutrients. Historically these areas have been almost devoid of human populations. However, in recent years one of the most extensive areas of the least fertile Oxisols in central Brazil, known as the Cerrado, has experienced extensive agricultural development utilizing modern methods of fertilization and mechanization (Lopes 1996). Initially large amounts of phosphorus are required to overcome the phosphorus fixation capacity of these soils. However, after a one-time capital investment in phosphorus and lime is made, agricultural production is sustainable and profitable with annual increments of fertilizer. Then, annual fertilizer needs to replace nutrients removed by the grain harvest, and lime to combat acidity, are no greater than in other soils. The remarkable commercial agricultural development ongoing in Brazil (Figure 2) attests to the huge potential for increased human food production in South America, Africa, and elsewhere in the udi and ustic soil moisture regimes of the tropics. Commercial market infrastructure is necessary for modern agriculture to fully utilize these soils.

Throughout the tropics, soils formed on sediments containing nominal amounts of weatherable minerals, or directly from underlying rock and on slopes less than about 15%, have loamy to sandy topsoil overlying finer textured subsoil horizons. These soils classify as Alfisols and Ultisols. On slopes greater than about 15%, most soils are either shallow or have nearly uniform texture with depth. These soils classify as Entisols and Inceptisols and are very similar to many such soils in temperate latitudes. Chemical differences among these soils are determined by the composition of the parent material. Ultisols and relatively infertile Inceptisols and Entisols are formed in acidic parent materials. Alfisols and more fertile Inceptisols are formed over basic rocks or sediments. In udi and ustic soil moisture regimes these soils are most often naturally forested. Here indigenous humans commonly practice slash-and-burn, that is, shifting cultivation.

**Indigenous land uses in the tropics**

Slash-and-burn land use practices consist of cutting, drying, and burning native vegetation to volatilize the organic carbon, thereby converting...
The plant essential elements contained to plant available inorganic forms. Slash-and-burn practices locally differ in detail but all provide the high concentrations of nutrients necessary for food crops. Most food crops grow in less than 100 days in the tropics, requiring much more rapid uptake of nutrients than do native species. Burning also kills weed seeds near the soil surface. After one to three crops are harvested, the fields are abandoned and slow-growing native plants invade. It takes about 20 years, or more, in the least fertile soils for enough natural regrowth to accumulate sufficient plant-essential elements to nourish a reasonable food crop after again cutting and burning the vegetation. In many remote areas indigenous humans continue to utilize shifting cultivation for subsistence food crop production (Figure 3). In tropical Asia and elsewhere, thicker soils on steep slopes are often terraced for water management and fertilized for sustained growing of food crops.

Aquic soil moisture regime (seasonally saturated) Inceptisols and Entisols are present in major delta areas and smaller areas in river flood plains throughout the tropics. Proximity to river transportation and the possibilities for irrigation have attracted the development of civilizations in these locations. Frequent flooding is a hazard but the renewal of essential nutrients (obtained from the eroded topsoil deposited during floods) maintains soil fertility. Human intervention has largely removed native vegetation from these areas, and paddy rice production is extensive. Many of these areas, especially in Asia, have reliable infrastructure, and farmers supplement the natural soil fertility with commercial fertilizers to support the genetic potential of improved cultivars.
Thick, clayey textured soils composed primarily of smectite clay (Vertisols) on nearly level landscapes are rather extensive in India and Sudan, with smaller areas scattered throughout the tropics in old lake beds. These soils expand when wet and shrink when dry, forming deep, wide cracks. Most Vertisols have natural savanna ecosystems, and are chemically fertile. Thus, despite the undesirable physical properties, humans have used them extensively for food production and livestock pasture. If they can be kept moist via irrigation, they are good sites for paddy rice.

Andisols, soils formed in volcanic ash, comprise only a small portion of the tropics. Most Vertisols have natural savanna ecosystems, and are chemically fertile. Thus, despite the undesirable physical properties, humans have used them extensively for food production and livestock pasture. If they can be kept moist via irrigation, they are good sites for paddy rice.

Organic soils (Histosols) are present in small, level, swamp areas throughout the tropics. The most extensive areas of Histosols are on the north coast of Sumatra and in parts of Borneo. Most areas have largely been avoided by indigenous peoples, except for paddy rice growing. Engineered drainage is needed for cultivation of most other crops.
SOILS OF TROPICAL LANDSCAPES

Sandy soils with a subsoil (B) horizon of amorphous materials composed of organic matter and aluminum (Spodosols) are of little agronomic value. Such soils are scattered in small areas throughout the Amazon Basin, with a rather extensive area present in the Rio Negro watershed. In Africa an area of southwestern Zambia is dominated by Spodosols.

Soils with an aridic soil moisture regime, mainly Aridisols but also some Oxisols, Andisols, Entisols, and Vertisols, receive insufficient rainfall to grow most food crops without irrigation. Only low intensity grazing is possible. The most extensive aridic soil moisture regimes in the tropics are south and east of the Sahara Desert in Africa, the Arabian Peninsula, and the west coast of South America in Peru and northern Chile.

Summary

Minimal seasonal soil temperature variation is the only characteristic common to all soils in tropical latitudes. The majority of soils in the tropics have other physical, chemical, and mineralogical properties comparable to many soils in temperate latitudes. Only one major order of soils, Oxisols, is not found in temperate latitudes; they occupy only about 24% of the tropical land area. Viable technologies that greatly increase food production in the tropics are available. Lack of infrastructure to implement these technologies, not insurmountable soil properties, relegates farmers to low yielding subsistence farming in many tropical areas.

SEE ALSO: Agricultural environments; Agricultural geography; Soil taxonomy and soil classification; Soils of desert landscapes; Soils of humid mid-latitude landscapes; Tropical geography; Tropical geomorphology

References


Further reading


Soils of urban and human-impacted landscapes

Jeffrey L. Howard  
Wayne State University, USA

W. Lee Daniels  
Virginia Polytechnic Institute and State University, USA

Many soils of urban and human-impacted landscapes have been so modified by human activity that the original soil remains only as a buried profile, or has been so drastically changed that it is no longer recognizable. Often these soils have unique characteristics that lie outside the range of indigenous soil series. Such soils have been labeled by various adjectival terms, for example “anthropomorphic,” “anthropic,” and “anthropogenic.” To facilitate discussion, the term “Anthrosol” is used informally here for any soil formed or profoundly modified by human activity (IUSS Working Group 2006). Anthrosols may be formed from human-transported materials (anthropogenic deposits) or human-altered natural parent materials.

In this entry, we recognize four distinct types of Anthrosols based on geographical setting and historical context: (i) agricultural, (ii) archaeological, (iii) mine-related, and (iv) urban. Agricultural Anthrosols differ from plowed and other soils normally found in agricultural settings in that they have been modified so profoundly by human activity that they require reclassification as a different type of soil. These changes are usually caused by long-term additions of soil amendments (manure, fertilizer, etc.), or by a fundamental change in soil drainage brought on by irrigation. Agricultural Anthrosols can be classified into two distinct subtypes: (i) Plaggenic, and (ii) Anthraquic. Plaggenic Anthrosols are characterized by an artificially overthickened A horizon called a plaggen in Europe, or a plaggen epipedon in the United States (Soil Survey Staff 2010). A plaggen epipedon is a surface layer ≥50 cm thick produced by long-term additions of manure or other organic materials. It often contains artifacts (objects of anthropogenic origin) such as charcoal, bone, brick, and pottery, and may contain included chunks of subsoil churned up by digging or tilling. In northwestern Europe, where plaggen are widespread, it was common practice during Medieval times to spread forest litter, heath-turf, or grass sod on the floors of barns and stables. These materials, along with the droppings compacted by earthmoving equipment, and are sometimes formed in parent materials that were imported from offsite. Urban and archaeological Anthrosols are unique in that they often contain abundant artifacts (objects of anthropogenic origin) which may be comprised of substances not found in nature. Anthrosols are also associated with suburban settings, streets, highways, and utility line or pipeline construction areas.

Agricultural Anthrosols

Agricultural Anthrosols differ from plowed and other soils normally found in agricultural settings in that they have been modified so profoundly by human activity that they require reclassification as a different type of soil. These changes are usually caused by long-term additions of soil amendments (manure, fertilizer, etc.), or by a fundamental change in soil drainage brought on by irrigation. Agricultural Anthrosols can be classified into two distinct subtypes: (i) Plaggenic, and (ii) Anthraquic. Plaggenic Anthrosols are characterized by an artificially overthickened A horizon called a plaggen in Europe, or a plaggen epipedon in the United States (Soil Survey Staff 2010). A plaggen epipedon is a surface layer ≥50 cm thick produced by long-term additions of manure or other organic materials. It often contains artifacts (objects of anthropogenic origin) such as charcoal, bone, brick, and pottery, and may contain included chunks of subsoil churned up by digging or tilling. In northwestern Europe, where plaggen are widespread, it was common practice during Medieval times to spread forest litter, heath-turf, or grass sod on the floors of barns and stables. These materials, along with the droppings
of livestock, were then collected and strewn onto sandy fields under cultivation. European plaggen may be 1 m or more in thickness as a result of many centuries of these types of activities. Plaggenic Anthrosols are common in England, Ireland, Holland, Belgium, and Germany.

Anthraquic Anthrosols are a special kind of aquic soil in that they are affected by excess water introduced by flood irrigation, often associated with paddy rice cultivation. Anthraquic soils are affected by continuous or periodic saturation for three months or more resulting in chemical reduction. They are often characterized by a tilled surface layer with redoximorphic features lying directly above an impermeable horizon (Soil Survey Staff 2010). During a wetting event, \( \text{Fe}^{2+} \) may be released to solution by protonation and then adsorbed onto the exchange sites of clay minerals and organic matter. Basic cations are desorbed and undergo hydrolysis generating \( \text{OH}^- \) anions. During a subsequent drying event, sorbed \( \text{Fe}^{2+} \) is oxidized to form ferrihydrite and \( \text{H}^+ \). This cyclic process was referred to as ferrolysis by Brinkman (1977), who thought that excess \( \text{H}^+ \) generated by this process could eventually cause dissolution of clay and a reduction of cation exchange capacity in these soils.

Well drained sandy agricultural Anthrosols also may be characterized by an agric horizon, that is, an anthropogenic illuvial horizon lying directly beneath a plowed A horizon. An agric horizon is \( \geq 10 \) cm thick with illuvial humus, silt, and clay comprising \( \geq 5\% \) of the horizon by volume. It is typically characterized by abundant earthworm burrows (krotovina) and clay lamellae (Soil Survey Staff 2010). This horizon develops because the plowing disturbance facilitates translocation of fine particles into the zone immediately underlying the plow layer. Over time, illuvial particles form lamellae, or coatings in worm burrows and on ped surfaces. Examples of agric horizons are found in Belgium, and in India on the floodplain of the Ganges River.

Archaeological Anthrosols

Archaeological Anthrosols are developed on agricultural, habitation, ceremonial, and burial sites. They are commonly found as buried soils in tells and burial mounds. Archaeological Anthrosols differ from native soils because they have been altered drastically by additions of organic matter derived from plant and animal wastes, or charcoal, along with mineral material in the form of ash, bone, and shell. Four types of archaeological Anthrosols can be distinguished (Holliday 2004): (i) Plaggenic, (ii) Dark Earth, (iii) Terra Preta, and (iv) Midden. As noted above, plaggen result from the intentional addition of manure and other wastes, although they also frequently contain archaeological artifacts dating back to the tenth century or earlier. “Dark Earth” Anthrosols are common throughout Europe, and in Britain they may date back to the late Roman Empire. They are referred to as “urbic Anthrosols” in the World Reference Base classification (IUSS Working Group 2006). Dark Earth Anthrosols are somewhat similar to plaggen in that they are an overthickened A horizon, perhaps 70 cm thick or more, containing brick, mortar, pottery, bone, shell, and other artifacts. However, they are morphologically more homogeneous than plaggen, and Dark Earth Anthrosols are thought to represent former occupation sites (Holliday 2004). Terra Preta Anthrosols are also thick, black, artifact-bearing topsoils which have been artificially enriched in organic matter. These soils are characteristic of the Brazilian Amazon basin, where the thick, black topsoil is in striking contrast to the underlying reddened Ultisol horizons. Their origin appears
to be complex. Scientists today are unsure whether Terra Preta were formed intentionally, unintentionally, or both (Lima et al. 2002). A midden is an organic-rich topsoil formed in the classic sense by dumping of “kitchen” wastes at archaeological sites. Kitchen middens were the inspiration for the anthropic epipedon as originally defined in Soil Taxonomy. A midden is basically a mollic epipedon containing elevated levels of phosphorous, commonly due to the incorporation of bones or organic substances that are rich in phosphorous. Shell middens are common in coastal, estuarine, or lacustrine settings, which are in proximity to sources of shellfish. Burned rock middens are common in Texas. They are essentially an anthropic epipedon containing charcoal, and burned or fire-cracked rock representing a former fire pit.

Mine-related Anthrosols

Mine-related Anthrosols differ from native soils in that they are formed on anthropogenic (human-created) landforms, whose surface and subsurface hydrological characteristics have been drastically altered. These soils also have formed from parent materials that have been artificially sized by blasting and mixed together in an unnatural way by mechanized equipment. Mine-related Anthrosols are particularly widespread in areas surface-mined for coal and aggregate (sand and gravel), but large land areas have also been disturbed by metal and aggregate mining and associated waste dumps.

Of all mine-related Anthrosols, Appalachian (United States) coal mine spoils, and resultant topsoil-substitute mine soils, have been studied most extensively. They differ from natural parent materials because they are often characterized by a considerable amount (20–40%) of soil-sized (<2 mm) material coupled with a large proportion of hard rock fragments. They tend to be highly compacted by earthmoving equipment, and highly variable in both physical and chemical properties (Haering, Daniels, and Galbraith 2005). In contrast, large areas of coal mine soils in the central United States have been reconstructed from pre-existing agricultural soil materials and, as a result, more closely resemble their native counterparts in texture and bulk soil chemistry.

Coal and associated carbonaceous shale often contain pyrite, possibly in great abundance. In an oxidizing environment, pyrite weathers rapidly to generate sulfuric acid, which can in turn drastically lower the pH value of the soil, as well as solubilize iron and various toxic trace metals within it. Mine-related Anthrosols formed from pyritic parent materials are typically characterized by a sulfuric horizon, that is, a horizon ≥15 cm thick comprised of mineral or soil material with a pH ≤3.5. If calcareous parent materials are present, sulfuric acid can react with calcite to produce gypsum. Coal mining also generates a large volume of coal-processing wastes (coal refuse) that are usually much coarser, more acidic, and more carbonaceous than other mine spoils. Pyrite and toxic trace metals are often concentrated in coal refuse. Mining of sand and gravel, and minerals containing valuable trace metals (e.g., titanium, zirconium), has disturbed large areas of land in various parts of the world. A wide range of hard-rock mining activities for metals, aggregate, and carbonate rock has also generated large areas of waste rock dumps and/or tailings that contrast strongly with native parent materials, resulting in a wide variety of mine-related Anthrosols with unique properties.

Urban Anthrosols

All soils in an urban setting are affected by environmental changes of anthropogenic origin
SOILS OF URBAN AND HUMAN-IMPACTED LANDSCAPES

(i.e., anthropogenic soil-forming processes) such as elevated soil temperature resulting from the urban heat island effect, deposition of wind-blown soot and coal ash, runoff containing deicing salt, acid rain deposition, and so on. However, urban Anthrosols differ from urban native soils because they have been modified so profoundly by human activity that they require reclassification as a different type of soil, or the native soil remains only as a buried profile. Urban Anthrosols often contain large volumes of artifacts, and are sometimes characterized by layers of concrete, asphaltic pavement, and so on, which seal off part or all of the soil from the effects of more “natural” soil-forming processes. To facilitate discussion, urban Anthrosols are classified into two types depending on whether they were formed in: (A) human-altered material (modified native parent materials), or (B) human-transported material (anthropogenic deposits). Human-altered parent materials are formed in place either by deep mixing, excavation and replacement from a single pedon, or truncation and removal of the surface soil. Human-transported parent materials are formed by excavation of material from one pedon and mixing with materials from other pedons, or by moving material horizontally onto a pedon from other sources, usually with the aid of hand tools or mechanized equipment.

Type A urban Anthrosols are characterized by a profile similar to that of native soil, but with a thicker, darker A horizon created by human additions of organic substances other than manure. They often contain artifacts (Figure 1a) and show some evidence of human disturbance by digging. This type of anthropogenic A horizon may or may not have elevated phosphorous levels (from animal bones), but otherwise usually meets the current requirements of an anthropic epipedon (Soil Survey Staff 2010). The darkening of the topsoil may be the result of human additions of organic substances (char, charcoal, coal ash, etc.), or perhaps soot and coal-ash contamination, as a legacy of fossil fuel combustion. Artifacts are typically household items or trash.

Type B urban Anthrosols are characterized by a weakly developed soil profile (Figure 1b) developed in an anthropogenic deposit, often overlying buried native soil (Figure 1c), or resting directly on native parent materials. The anthropogenic deposit is typically the result of construction or building demolition. The parent materials are comprised of an artificial mixture of artifacts and earth materials (rock, sediment, soil, etc.), often imported from offsite. The anthropogenic deposit is usually massive, highly compacted, artifact-rich, lithologically heterogeneous, and laterally discontinuous. Such deposits may be comprised of a single massive layer, or multiple layers, and occasionally have a complex microstratigraphy. Artifacts in such soils are often waste building materials (wood, glass, brick, mortar, concrete, plaster, nails, etc.) or coal-related wastes (soot, fly ash, cinders). They may be comprised, at least in part, of artificial substances (e.g., coal tar, plastic) not usually found in rocks and native soils (Figure 1d). They can range in size from gravel to sand or finer, and may represent waste materials simply left onsite, or comminuted by the demolition process (Howard, Dubay, and Daniels 2013), and then incorporated into earthy backfill materials during grading. Glass and glass-bearing artifacts (brick, coal cinders and ash, steel-making slag) can be expected to have good preservation potential based on their resistance to chemical weathering. There is also a high probability of preserving organic artifacts comprised of bituminous materials or black carbon (coal, asphalt, black carbon, tar) because they are relatively recalcitrant to biodegradation.

Urban Anthrosols which have experienced excessive compaction are usually characterized by greatly reduced infiltration, excessive dryness, and possibly anaerobic conditions. A perched
Figure 1  Urban Anthrosols at former demolition sites in Detroit, Michigan: (a) Type A soil characterized by anthropic epipedon formed in human-altered material; (b) Type B soil 24 years old with weakly developed A horizon formed in human-transported material; (c) Type B soil 68 years old with human-transported material overlying buried native soil; (d) Type B soil comprised of coal-related wastes overlying buried native soil. Scales in cm. Pick handle is 67 cm long. ^A denotes anthropogenic horizon; u, artifacts; b, buried; d, densic; g, gleyed; k, calcic; w, cambic. Note krotovina filled with earthworm casts and calcic horizon formed by leaching of carbonate derived from weathered mortar. Source: Adapted from Howard, Dubay, and Daniels 2013. © Elsevier.

water table (episaturation) is common, but the hydraulic conductivity of the underlying restrictive layer may also be significantly enhanced by the presence of vertical worm burrows. Urban Anthrosols are structureless and may contain elevated levels of iron oxide and manganese oxide and carbonate as a result of the presence of weathered ferruginous and calcareous artifacts. The organic fraction often contains a mixture of soil organic matter (produced by humification), black carbon (e.g., charcoal, soot, and coal ash), and various types of carbonaceous substances derived from geogenic materials (e.g., coal, coked coal, coal tar, asphalt, carbonaceous shale). Soil pH and exchangeable calcium levels are typically elevated as a result of weathered calcareous artifacts. Exchangeable sodium levels may also be elevated in urban soils which have received excessive runoff from pavement carrying de-icing salts. Measured carbon/nitrogen ratios and pH values show significant decreases, and soil organic matter contents substantial increases, in urban soils over the course of about a century of soil development in Detroit, Michigan (Howard, Dubay, and Daniels 2013). Microbial biomass appears to be very low initially in anthropogenic fills, but rebounds with increasing humification and nitrogen accumulation such that microbiological activity in urban soils may be comparable to that of native soils in less than 100 years. Anthrosols similar to those termed “urban” for this discussion may be present in suburban and even possibly in rural geographical settings.

Genesis and classification of Anthrosols

Plaggenic Anthrosols (and certain archaeological Anthrosols) are known to have required as much as 600 years or more to form, and there are Anthraquic paddy rice soils in southeastern Asia which have formed as a result of 1000 years of cultivation. However, an A horizon is usually the first feature to develop during
pedogenesis. Weakly developed A horizons have been observed to develop relatively rapidly (about 15–30 years) in mine-related (Haer- ing, Daniels, and Galbraith 2005) and urban (Howard and Olszewska 2011; Howard, Dubay, and Daniels 2013) Anthrosols. In contrast, except for some weak signs of structure, B horizons require centuries to form. Thus, the typical Anthrosol is characterized by an A/C profile. Where compaction by earthmoving equipment has occurred, mine-related and urban Anthrosols often have a densic (Cd) horizon which cannot be penetrated easily by plant roots and has a bulk density \( \geq 2.8 \text{ g/cm}^3 \). Mortar and other calcareous artifacts in A horizons may show significant weathering within a few decades and, accompanied by leaching, may form a calcic (Ck) horizon at depth. Despite excessive compaction, krotovina (burrows) produced by earthworms such as *Lumbricus terrestris* may become prevalent in urban Anthrosols, especially at sites that have remained vacant for decades.

Anthrosols first appeared on the revised legend of the Food and Agriculture Organization–United Nations Educational, Scientific and Cultural Organization soil map of the world in 1990, and are still recognized as a major grouping in the soil classification systems of some countries (IUSS Working Group 2006). In the case of agricultural Anthrosols, human influence is mainly restricted to surface layer, and several categories of “anthric horizon” have been recognized. Anthric horizons are generally similar to mollic or plagggen epipedons in the US Soil Taxonomy system. Early efforts to apply the US system of Soil Taxonomy to urban soils showed that they were typically classified as Entisols, but because of their unique properties, they did not conform to the central concepts of existing Entisol suborders. Several early proposals for modifying Soil Taxonomy to account for disturbed soils were eventually presented and widely debated. It was proposed that all eastern US mine soils could be classified as various subgroups (e.g., Calcic) of a new Great Group called Udispolents, emphasizing their “mine spoil” components. Much wider revisions at the suborder level (Garbents for garbage, Urbents for urban soils, etc.) and lower categories were also proposed. The International Committee for Anthropogenic Soils was first convened in 1995 to help formulate revisions to Soil Taxonomy pertaining to the classification of Anthrosols. Although much discussion followed about the possible creation of an Anthrosol Order, it now appears Anthrosols will be classified as Anthropic Subgroups of existing Great Groups. Efforts are still underway to update the definitions of buried soils, artifacts, human-altered and human-transported parent materials, the anthropic epipedon, and other anthropogenic features.

Historically, county-level soil surveys in the United States mapped urban and human-impacted soils simply as “made-land” or “urban-land” before the 1970s, and as various subgroups of Udorthents in more modern surveys. However, in recent years urban land areas have been delineated as newly created anthropogenic soil series, or as soil complexes comprised of an existing native soil series and undifferentiated urban land. A problem commonly encountered during urban soil mapping is that anthropogenic landscapes have often been leveled, and thus the standard approach of using slope to delineate genetic map units is ineffective. Another problem encountered in the mapping of urban soils is that the high level of variability and the presence of abundant artifacts make it necessary to obtain an inordinately large number of auger borings. As a result, remote sensing techniques using geophysical methods (e.g., electrical conductivity mapping) may be more effective for mapping Anthrosols.
Reclamation and revitalization of Anthrosols

Reclamation of coal mine-related Anthrosols in the United States is now mandated by the 1977 Surface Mining Control and Reclamation Act. One requirement of this law is that mine operators post bond monies which are not returned until the disturbed land has been revegetated. Hence, a great deal of research has been carried out in the United States with the objective of developing cost-effective strip mine reclamation techniques. Similarly, revitalization of urban Anthrosols has become a target of study because large tracts of vacant urban land produced by building demolition have attracted considerable interest for urban gardening and agriculture, green infrastructure (e.g., for groundwater recharge or wildlife habitat), and as a repository for wastewater from combined sewer overflow. Unfortunately, mine-related Anthrosols and Type B urban Anthrosols are rendered biologically unproductive, and have severely limited infiltration and hydraulic conductivity rates, as a result of excavation and backfilling processes.

Mine-related Anthrosols are strongly influenced by local geological and soil conditions, coupled with mining and reclamation methods (Figure 2). Revitalization of these soils is commonly limited by excessive potential acidity (from acid mine drainage), compaction, high rock fragment contents, low water-holding capacities, low plant-available nitrogen and phosphorous contents, and saline-sodic conditions. Coal-processing wastes are particularly difficult to reclaim because of acidity, and large piles of coal refuse are prone to mass movements (Daniels and Stewart 2000). Assuming that issues with pyrite are managed via application of appropriate acid-base accounting procedures, application of organic waste materials (sewage sludge, biochar, etc.) is still the most common approach for strip mine reclamation.

Demolition site soils are in need of revitalization because of low organic matter contents, compaction, excessive artifact contents, and various forms of other contamination, particularly by lead. Reconditioning techniques for urban soils include promoting air, water, and nutrient infiltration into the soil in order to enhance biological activity, application of amendments to alter soil chemistry, addition of compost, and removal of artifacts and tilling to reduce bulk densities (USEPA 2011). In some cases, it may be advantageous to leave calcareous and ferruginous artifacts in the soil because their weathering products act as immobilizing agents for certain toxic trace metals (Howard and Olszewska 2011).

Conclusions

Anthrosols are widely distributed in agricultural, archaeological, mine-related, and urban geographical settings. They are an important natural and historical resource, and have become a topic of considerable interest for geologists, geographers, and soil scientists over the last few decades. The fact that Anthrosols have characteristics outside the range of native soil series has important implications for their use and management. Hence, more study is needed of anthropogenic processes, anthroscapes, and techniques better suited for the mapping and classification of Anthrosols. As reclamation techniques improve, mine-related and urban Anthrosols may eventually reach their full potential as a repurposed natural resource for agricultural and other types of land use. Many earth scientists now believe that we are living in a new geological epoch, the Anthropocene, based on the fact that human activity has altered the Earth’s surface to such an extent that the uppermost part of the geological
Figure 2  Drastically disturbed landscape in southwestern Virginia where strip-mining for coal is taking place. Note heterogeneous texture of mine spoil ranging in size from fine sand to boulders.

column is profoundly different from that of the Pleistocene and Holocene epochs. Hence, the societal need for a better understanding of Anthrosols is expected to continue growing in importance because the rates of world population growth, and associated land disturbance by humans, are increasing exponentially.

SEE ALSO: Anthropocene and planetary boundaries; Anthropogeography; Environment and urbanization; Soil taxonomy and soil classification; Soils in archaeological research; Urban redevelopment

References


**SOILS OF URBAN AND HUMAN-IMPACTED LANDSCAPES**


**Further reading**


Wetland soils – also known as hydric soils – are found wherever the combination of physiography, hydrology, and climate allow for prolonged saturation. Anaerobic (oxygen-depleted) conditions alter biogeochemical processes, favoring chemically reducing conditions and giving rise to a diagnostic suite of redoximorphic features in soils. Wetland soils include those that have developed in both mineral and organic materials. Some wetland soils are well understood, including peatlands, coastal soils, and paddy soils; others have often fallen into the gap between soil science and wetland science due to their ephemeral nature, such as vernal pools and prairie potholes. For this reason, these types of wetland soils are more poorly understood. Saturation in wetland soils may be permanent, temporary or seasonal, but any soil where saturation is persistent enough for anaerobic conditions to develop in the rooting zone can be described as hydric (Richardson and Vepraskas 2001). Sometimes wetland soils are defined in terms of their ability to support hydrophilic vegetation, but given the ubiquity of land-use change, wetlands that have retained their native vegetation are increasingly rare and therefore the hydric soil concept was developed to recognize the longer-term imprint of hydrology on soil properties.

Classification

In mineral soils, hydric soil characteristics are recognized at different levels in different systems of soil classification: as a moisture regime in United States Department of Agriculture (USDA) soil taxonomy (Aquic suborders), at the order or subgroup level in the Canadian System of Soil Classification (Gleysol order or Gleyed subgroups), and as major groups in the World Reference Base system (Stagnosols, Planosols, Fluvisols, Gleysols). Soils dominated by decomposed plant residues or peaty material (>12–18% organic carbon) can classify as Histosols (USDA order, World Reference Base group) or Organic soils (Canadian System order). However, not all of these soils develop under saturated conditions, and thus, organic soils alone do not universally indicate a hydric soil system.

Hydrology

Typically, hydric soils develop in environments of imperfect to poor drainage, leading to saturation. The source and duration of saturation varies; in some areas, hydric soils develop due to a humid climate or proximity to persistent water bodies (lakes, rivers, or oceans). In others, an occluding or slowly permeable subsurface layer (bedrock or heavy clay) prevents adequate drainage following spring snowmelt or precipitation events. This situation is especially common on level landscapes with minimal run-on or runoff. On sloping landscapes, saturation is often associated
with specific slope positions, with hydric soils occupying the lowest positions on the landscape. For example, even in semiarid regions, like the prairie pothole region of North America, hydric soils may develop in the bottoms of depressions. This region is also an example of how groundwater can play a major role in hydric soil formation. Groundwater discharge from unconfined aquifers, which are common in the glacial deposits throughout the region, can be an important water source in some wetlands (Bedard-Haughn 2011). The upward gradient of groundwater movement contributes to saturated conditions near the surface and prevents drainage of surface waters.

The source and duration of saturation, along with the biophysical environment, affect the properties of the soil. For example, where the dominant direction of water movement is downward through the profile (recharge wetland; Figure 1) the profile will have more typical horizon development and may include evidence of downward transfer of clays, organic matter, or other compounds (eluviation-illuviation). In contrast, profiles dominated by upward movement of groundwater (discharge zones; Figure 1) may have weakly developed horizons and evidence of secondary salt deposition, if these compounds are present in the groundwater. Where plant residue inputs are sufficiently high and saturation is persistent, organic soils are likely to develop regardless of the dominant groundwater flow.

Redox processes and redoximorphic features

Hydric soils are characterized by properties and processes that reflect their prolonged saturation. These features develop via oxidation-reduction (redox) reactions that occur when oxygen is depleted from the soil. The soil microbial community (i.e., bacteria and fungi) decomposes (oxidizes) organic residues from flora and fauna. Heterotrophic micro-organisms use organic tissues as a carbon source for respiration; these tissues release electrons that are then available for reducing reactions. Under aerobic conditions, oxygen \( \text{O}_2 \) is the preferred electron acceptor (O is reduced), but under the anaerobic conditions in wetland soils, an alternative electron acceptor (Table 1) must be reduced for microbial respiration to continue (Richardson and Vepraskas 2001). Some reducing reactions give rise to observable redoximorphic features; others are less visible, but are nonetheless important wetland processes. The most obvious expressions of redox reactions in wetland soils are (i) accumulation of soil organic matter, (ii) iron and manganese masses and/or depletions, and, in some cases, (iii) smell of hydrogen sulfide \( \text{H}_2\text{S} \) from sulfate reduction (Richardson and Vepraskas 2001).

Organic matter accumulation in wetland environments is due to slowed decomposition under saturated conditions; some of these sites are also colder, which further slows organic matter decomposition. Oxygen dissolved in the pore water is rapidly reduced; after it is depleted, microbial decomposition slows dramatically. The resulting accumulation of organic carbon occurs as layers of peaty or mucky material, or as accumulations of dark-colored organic material (humus) within a mineral horizon. Organic matter accumulation is not diagnostic of reducing conditions because it can also accumulate in highly productive, well-drained soils; nonetheless, it is a common feature of wetland soils. Consequently, wetlands are a major global carbon sink. For example, peatlands, which cover only 3% of the terrestrial surface, store up to 33% of global soil carbon. Freshwater mineral soil wetlands are estimated at about 10% of the peatland storage (global estimate: 462 Pg
in peatlands versus 46 Pg in freshwater mineral soil wetlands), but are still a substantial pool and vulnerable to loss under changing land use and climate (Bridgham et al. 2006).

The most common redoximorphic features that are used to delineate and differentiate hydric soils include masses and/or depletions of iron and manganese. Under moderately reducing conditions, these elements serve as electron acceptors for facultative anaerobes, and thus, manganese and iron oxides are reduced (from Fe$^{3+}$ to Fe$^{2+}$ and Mn$^{4+}$ to Mn$^{2+}$). Their reduced forms are more mobile in soil water and tend to impart a greyish or low chroma (≤2 in the Munsell color system) to the soil matrix. Low chromas develop when Fe$^{2+}$ and Mn$^{2+}$ are present in the soil (reduced matrix) or when manganese and iron have been removed from the soil, such as along root channels (redox depletions). This dull, often blue-grey, color can be described as “gleyed.” Typically, the reduced iron will move within the profile and be reoxidized when the soil dries out, forming mottles, which are rust-colored oxidized masses, pore linings, or nodules (Bedard-Haughn...
SOILS OF WET AND HYDRIC LANDSCAPES

Where manganese is present, the mottles may be black. Expression of redoximorphic features will vary with mineralogy, organic matter content, and texture, but in most cases they provide key morphologic information about the wetness status of a soil. For that reason, the location, size, and color of redox features in soils is central to their classification as to drainage class or wetness status.

Under more strongly reducing conditions, sulfate (SO$_4^{2-}$) can be reduced to S$^-$ or H$_2$S, giving rise to the “rotten egg” smell of hydrogen sulfide. In North America, sulfate reduction is commonly associated with sulfate-rich (oceanic) coastal systems, but is also widespread in areas of the Great Plains where the subsurface glacial till is rich in pyrite, which contributes sulfates to the groundwater as it oxidizes.

Two other key wetland redox processes include reduction of nitrate (denitrification) and carbon dioxide (methanogenesis). Both have been much studied in recent years because their products include potent greenhouse gases: nitrous oxide (N$_2$O) and methane (CH$_4$). Denitrification is a stepwise process, whereby nitrate (NO$_3^-$) is reduced first to nitrite (NO$_2^-$), then to gaseous N$_2$O, and finally to N$_2$. Under saturated conditions, the reaction proceeds all the way to N$_2$. Where some oxygen is present within the soil (60–80% water-filled pore space), there is greater potential for N$_2$O to be released to the atmosphere. In contrast, methanogenesis occurs under the most strongly reducing conditions, where no other electron acceptors are available. Methane production may also be limited by the availability of labile organic substrates for decomposition and by the activity of methanotrophic (methane-consuming) bacteria.

Theoretically, redox reactions would proceed in sequence, reflecting their respective redox potentials (Eh), with O$_2$ being reduced first and CO$_2$ reduced last (Table 1). Consequently, the presence of soil nitrate may inhibit iron reduction because nitrate reducers are able to outcompete iron reducers for electrons. Similarly, high levels of sulfate have been shown to inhibit methanogenesis in both coastal soils and those prairie potholes affected by sulfate-rich groundwater. However, microscale variability in water and oxygen contents, along with other biogeochemical properties (concentrations of alternate electron acceptors, location of organic matter and soil microbes), implies that different redox reactions can occur simultaneously over short distances.

### Land use

When drained, wetland soils, characterized by high levels of organic matter and nutrients (returned to the soil by plant residues and accumulated from upslope soils by erosion), are excellent for agriculture. Both peatland and mineral wetland soils are commonly drained for production (for example, Dutch peatlands and widespread drainage of tropical peatlands for palm oil production). However, this practice is known to result in rapid carbon loss (i.e., greenhouse gas emissions) and subsidence via soil compaction (Verhoeven and Setter 2010). Up to 70–90% of freshwater mineral soil wetlands throughout North America have already been drained, contributing substantially to greenhouse gas emissions, downstream nutrient loading, and lake eutrophication. However, not all agriculture results in wetland loss. Globally, paddy soils for rice production cover nearly 155 million hectares (Kogel-Knabner et al. 2010). These soils may not have been hydric prior to rice production, but now function as hydric soils due to the saturation maintained throughout most of the growing season. Wetlands are also created for regulatory requirements (i.e., to compensate for wetland loss via development or resource extraction).
and pollution attenuation, including mitigating excess nitrogen (via denitrification), phosphorus, pesticides, pathogens, and trace metals (O’Geen et al. 2010). Created wetlands of this kind differ in origin, of course, but when carefully developed and maintained can perform many of the same ecosystem functions as their natural counterparts.

**Current and future research**

Current research emphasizes measuring and modeling (i) the effects of climate and land-use change on wetland soils, and (ii) the effects of wetland soil management on climate change and water quality. There is also an emphasis on determining economic value for the ecosystem services provided by wetlands and wetland soils, such as habitat, flood mitigation, and improved water quality. Food and water security are intertwined in hydric soils. Wetlands provide some of the most productive agricultural soils. However, the destruction of wetlands for agriculture alters local and regional hydrology, potentially increasing the risk of downstream flooding and/or contamination with agricultural pollutants. One of the challenges faced in the study of wetland soils is the lack of a detailed inventory of wetlands or hydric soils, particularly in developing nations and sparsely populated regions. Where inventories do exist, many fail to take into account smaller (<1 km²) or temporary wetlands, because these are not as readily captured with common remote sensing technologies, nor are they included at the scale of many soil maps. Going forward, efforts to (i) model the effects of land-use and climate change on wetland soils and (ii) make sound long-term land management decisions will require a high-resolution baseline inventory and a detailed assessment of ecosystem services provided by wetland soils.

**SEE ALSO:** Land-use/cover change and climate; Soil water; Soils on slopes: catenas; Wetlands hydrology

**References**


Soils on slopes: catenas

Peter C. Almond
James L. Moir
Lincoln University, New Zealand

Hillslopes are a pervasive feature of Earth’s surface. The proportions of land with slopes greater than 5°, 15°, 20°, and 30° are 25%, 9.4%, 5.9%, and 1.8%, respectively (Larsen, Montgomery, and Greenberg, 2014). Soils on hillslopes thus represent a significant resource for producing food and fiber for a growing global population and provide other important ecosystem services.

Most hillslopes on Earth result from the interaction of (i) rock uplift resulting from crustal deformation, combined with (ii) the downcutting response of agents of denudation, for example, rivers and glaciers. Once formed, hillslopes act as both the sources and conveyors of sediment, focusing that sediment and water into valley floors. Soils on hillslopes serve an important hydrological function by allowing precipitation, which would otherwise travel rapidly overland on the inclined surface, to infiltrate and to be transmitted to streams via subsurface flow paths. This phenomenon slows the flux of water to streams, thereby reducing flood peaks, enhancing water quality, maintaining baseflow, and, overall, fostering stream habitat quality.

Globally, and over geological timescales, soils on hillslopes have probably contributed disproportionately to the development of Earth’s climate. Primordial Earth’s atmosphere had very high concentrations of CO₂, which, being a greenhouse gas, sustained hothouse climates unsuitable for most life. Weathering of silicate minerals by carbonation (in soils) has sequestered much of the atmosphere’s original CO₂ as biogenic carbonate minerals (limestones, dolomites) within the oceans. Hillslope soils are especially important because rejuvenation of soil by erosion supplies fresh soil minerals to maintain these C-consuming weathering reactions. In contrast, flat landscapes with no erosion and minimal deposition become cloaked in thick soils and chemically altered rock (regolith) that is more completely exhausted of weatherable minerals. The rejuvenating effect of erosion on hillslope soils also extends to depositional landforms on valley floors and piedmonts. Relatively unweathered sediment from hillslopes is distributed onto fans and floodplains and may be remobilized to form drapes of windblown sediment (loess) over surrounding landscapes. The unweathered minerals in soils rejuvenated by erosion and deposition also contain large pools of biogeochemically important elements.

The persistence of soils on hillslopes

Whether or not hillslopes are soil-mantled depends on the balance between denudation (chemical and physical erosion) and the rate of conversion of the underlying substrate (by physical disintegration) into mobile (colluvial) materials (regolith and soil). The rate of conversion of bedrock to regolith, in the geomorphological literature, is referred to as soil production (Figure 1). In general, where soil can be produced faster than the denudation rate, slopes are described as being transport limited and will be soil-mantled. At first inspection, this condition would appear to create a progressively (and
unlimited) thickening soil mantle. However, an internal soil feedback occurs whereby increasing soil thickness is associated with declining soil production rates. The soil thickens until it forms at a rate that matches denudation, at which point a steady state prevails. Two forms of the relationship between soil production rate and soil thickness (the soil production function) have been proposed. The bulk of empirical evidence from measurement of cosmogenically produced beryllium-10 at the bedrock/soil interface supports an exponentially declining soil production rate with increasing soil thickness (Heimsath et al. 1997). Maximum soil production rates (at soil depth = 0) vary two orders of magnitude from 0.03 to 2.5 mm year$^{-1}$ depending on a range of factors including rock type, climate, and the intensity of biological disturbance from vegetation and burrowing animals. On gentle slopes where soil is thick, the rate at which the chemical weathering front advances downward may exceed the soil production rate. Then, the colluvial soil is underlain by essentially in-place but chemically weathered rock (saprolite).

Where denudation occurs faster than the maximum soil production rate, slopes are described as weathering (or detachment) limited and the soil mantle tends to be patchy and thin. However, even where the denudation rate exceeds the maximum potential soil production rate, it is still possible for large parts of hilly or mountainous landscapes to be soil-mantled. In the Southern Alps of New Zealand, for example.
Zealand, for example, rock uplift and denudation rates reach 10 mm year$^{-1}$ and are as much as four times faster than the maximum soil production rate of 2.5 mm year$^{-1}$ (the fastest measured), yet slopes largely remain soil-mantled. In this landscape, mass movements (landslides) are the dominant denudation process but their recurrence at any one location has a return period long enough for soils to form at this high rate.

**Hillslope-soil models and the soil catena**

The emphasis on research of soils on hillslopes varies among the different branches of the earth sciences. Geomorphologists have been interested in soil-mantled hillslopes from the perspectives of landscape evolution, delivery of sediment and dissolved load to rivers, and interactions of hillslope soil weathering with climate. Hydrologists are interested in the properties and spatial variability of soils on hillslopes with respect to flow paths and dissolved load of water transmitted to surface and groundwater. Soil scientists’ interests include understanding soil variability on slopes, for applications in soil survey, soil erosion, and soil fertility studies. There is considerable overlap among the disciplines; however, the need for models to explain and predict soil variability on hillslopes is common to all.

Although geomorphologists have incorporated soil production as a mechanism of denudation in mathematical hillslope and landscape evolution models, the depiction of soil has generally been limited to the single descriptive variable of soil thickness. Models, often qualitative in nature but incorporating more complete descriptions of soil (vertical and lateral differentiation of properties and horizons), have been developed and used primarily within the realm of soil science. All models take as their fundamental premise the downslope movement of sediment, water, and solutes in response to gravitational force.

Geoffrey Milne, working in East Africa in the 1930s, proposed the “soil catena” concept to embody the interrelationship between soils and hillslope form and process. Brown (2005), quoting Milne, wrote that, “The name itself ‘was intended to serve as a mnemonic, the succession of different soils corresponding to the links in a hanging chain’ in a progression from one hilltop to the next.” The catena concept was advanced as a practical response to a soil survey problem: meeting the demands of agricultural advisors who needed fine-scale resolution of the soil pattern on the one hand, and of cartographers, who needed regional scale soil information for incorporation in a world soil map, on the other (Milne and East African Agricultural Research Station 1936). Although in its conception the catena was an expedient for soil survey and was directed to illustrating and predicting the occurrence of different soil taxa, it explicitly included the role of processes: “Soil differences are brought about by differences of drainage patterns, combined with some differential reassortment of eroded material and the accumulation at lower levels of soil constituents chemically leached from higher up the slope” (Milne 1935). Milne also recognized that the effects of parent material would be apparent where a slope cut across different underlying lithologies.

A complementary notion by Polynov (1935) evolved at about the same time in the Russian literature. Polynov posited that element redistribution on hillslopes occurs according to their inherent mobility and that soils on hillslopes were linked by the lateral migration of chemical species for a single geochemical landscape, according to the sequence of mobility: $\text{Cl} > \text{SO}_4 > \text{Ca} > \text{Na} > \text{Mg} > \text{K} > \text{Si} > \text{Fe} > \text{Al}$.

The two concepts have since merged as reflected by Yaalon, Brenner, and Koyumdjisky’s
SOILS ON SLOPES: CATENAS

(1974) metaphor of the sequence of soils along a catena behaving as a chromatographic column for elements released by weathering and translocated by lateral water movement. The catena has thus become the conceptual framework in which both (i) relationships between soil taxa and elements of hillslopes and (ii) the continuous variation of soil properties in relation to slope position and form are expressed.

Hillslope hydrology

Net precipitation on hillslopes, resulting from the balance of total precipitation minus interception and evapotranspiration, moves within hillslopes either as overland flow, throughflow, or percolates into the substrate to recharge groundwater (Figure 2). Throughflow is shallow subsurface downslope flow; it occurs in saturated subsoil zones, which, because they parallel the sloping ground surface, create a pressure head gradient that drives the flow. Saturation of the soil is commonly caused by contrasts in hydraulic conductivity among horizons. Water moving as throughflow carries with it mobile (dissolved and dispersed) soil constituents and has the potential to differentiate soils in a downslope direction, just as vertical percolation of water promotes translocation, producing eluvial and illuvial zones in soils. This tendency is particularly well expressed in some dryland areas where salts released by chemical weathering of rocks in upslope positions are translocated downslope and precipitate there as salty encrustations resulting in salic soils in lowlands.

Overland flow (runoff) results when the rate of net precipitation exceeds the soils’ infiltration capacity, whereupon it is referred to as infiltration-excess or Hortonian overland flow. Overland flow may also occur because the soil is saturated (saturation overland flow) (Figure 3).

Hillslope processes and form

Hillslopes come in myriad forms, but they result from a finite set of processes, some of which have distinctive geomorphic signatures. Hillslopes most commonly connect the ridges and valleys and form drainage basins. In soil-mantled landscapes of the humid temperate regions, the ridge tops are convex upward. The convexity arises because soil creep is the dominant soil transport process on ridgetops. Soil creep is a slow shuffling of soil material downslope at a rate dependent only on local gradient. It results from soil disturbance processes such as rainsplash, bioturbation, or cryoturbation. The convex form of ridge tops is a response to the need for an increasing creep rate in a downslope direction to accommodate the increasing flux of soil generated by soil production from the slope above. On the gentle slopes of ridge tops where creep rate is directly proportional to gradient, convexity is directly proportional to erosion rate. The crest and shoulder slope elements (Figure 2 and Figure 4) that make up the convex ridgetop are usually dry parts of the landscape, because there is limited contributing area and an increasing downslope angle favors acceleration of throughflow.

A backslope element may be found between the shoulder and the footslope below (Figure 2). The backslope is the steepest part of the slope and is usually straight in form. On long slopes, this form arises because the rate of soil creep increases rapidly and non-linearly once slopes increase to about 20°. Hence, the downslope-increasing soil flux can be accommodated with imperceptible increases in hillslope angle and the slope appears linear. On the backslope, erosion rate and curvature are decoupled. Linear backslopes may also be found where slopes are steep enough (>26°) that soils and even bedrock reach a threshold of stability and undergo episodic mass
movement by processes such as landslides or earthflows (*Threshold slopes*). Alternatively, backslopes may simply be short transitions between convex shoulders and concave footslopes.

A non-linearly increasing flux rate with slope angle does not apply to throughflow, however, and at the point at which flow accumulation overwhelms the transmissivity of the soil, soil saturation and saturation overland flow can occur.

At this point, soil transport involves sheetwash, and the rate of transport is no longer dependent on the local slope angle alone. The transport capacity of overland flow grows non-linearly with increasing upslope contributing area and hence further down the slope lower gradients are sufficient to accommodate the soil flux, and slopes take on a concave upward form. This slope element is called the footslope (Figure 2),
SOILS ON SLOPES: CATENAS

Figure 3  Saturation overland flow in shallow soils over schist bedrock, central South Island, New Zealand (photo: P. Almond).

and being at the base of the slope it is often a zone of net sediment deposition. Being a zone of slowing throughflow flux and because of interactions with groundwater, footslopes are commonly wet. Soils are characterized by high and seasonally variable water tables, low redox potentials, and redoximorphic features (gleying) (Figure 4). Below the footslope, toeslopes mark

Figure 4  Typical catenary variation along a convex–concave slope transect in a superhumid temperate climate, western South Island, New Zealand. Note the transition to the concave footslope coincides with soil thickening and the increase of gray soil colors, which indicate prolonged soil wetness (photo: P. Almond).
the interaction of hillslopes with floodplains and soils here may be very wet and complex in their layering. Toeslopes and even footslopes may be absent if fluvial erosion is efficient.

A proper consideration of water and sediment movement on slopes must also consider their three-dimensional (3-D) configuration. Concave or convex curvature of hillslopes in plan form cause convergence or divergence of flowlines, respectively, for sediment, water, and the dissolved and dispersed materials. Slopes of different 3-D form are consistently arranged across the hollows and noses that define zero-order drainage basins (erosional drainage basins with no, or only an ephemeral, channel) (Figure 5). Convergence or divergence of flow lines intensifies or dampens, respectively, the effects of water and material transport in the downslope direction. For example, the strong convergence in the concave planform hollow intensifies wetness and can form a zone of intense saturation known as a saturation wedge at the basin outlet. Alternatively, divergent noses tend to be the driest parts of the landscape. Convergent hollows also accumulate the greatest thickness of soil.

Figure 5  Zero-order drainage basins, northern South Island, New Zealand. Terracettes formed by animal tracking following contours and highlight areas of convex plan curvature and divergent flow lines (noses), and areas of convex plan curvature where flow lines are convergent (hollows). The area in red identifies the inferred saturated wedge. Image from Google Earth.
The soil catena concept encapsulates the systematic variation of soil properties along hillslopes. This variation occurs as a result of the interactions among hillslope, hydrological, and pedogenic processes as outlined above. However, the particular kinds of soils and the different properties formed along a soil catena also depend on the local state factors of parent material, organisms, climate, and time. For example, as precipitation increases progressively more mobile elements become differentiated spatially along the soil catena. In higher rainfall environments saturation of soils and gleying enhance the mobility of elements such as Fe and Mn. Mobility of metal ions, including Al$^{3+}$, can be further enhanced where acidifying vegetation supplies organic compounds that can form organic-metal complexes with higher solubility (Figure 4).

The notion of soil age on hillslopes has some nuances. Whereas it is meaningful to ascribe an age to a soil on a non-eroding geomorphic surface, colluvial soils on hillslopes are made of an ensemble of mineral particles each with its own history of detachment from the immobile substrate. The relevant concept is the soil residence time ($T_R$), which for a steady-state soil thickness is given by

$$T_R = \frac{\rho_s h}{\rho_r P_h}$$

where $\rho_s$ and $\rho_r$ are the bulk densities of the soil and substrate, respectively, and $P_h$ is the soil production rate corresponding to soil thickness $h$. For example, in the Oregon Coast range in northwestern United States, soils are commonly about 40 cm thick, the production rate at this soil thickness is about 0.01 cm year$^{-1}$, and the ratio of the soil and bedrock bulk densities is about 0.5. The residence time, which is equivalent to the time for the whole soil volume to be replaced by soil production, is about 2000 years. Correspondingly, soils are relatively weakly developed Inceptisols (Figure 6). In general, more rapidly eroding parts of hillslopes will have soils with shorter residence times and more slowly eroding parts will have longer residence times. These residence times are reflected in the age distribution of mineral particles in the soils and the extent of pedological (including chemical) differentiation. On rapidly eroding slopes with short soil residence times catenary differentiation may be poorly expressed (Yoo et al. 2011).

The effects of soil residence time on soil taxa and properties are overprinted and often dominated by the redistribution of dissolved and dispersed soil materials and changing weathering environments brought about by hillslope hydrology. One of the most detailed accounts of these effects is reported by Park and Burt (2002) in a study of a zero-order drainage basin in Somerset in the United Kingdom. Figure 7 shows the pattern of soils within the drainage basin, which is formed in Devonian sandstone and under acidifying shrub vegetation and a humid climate. They measured 32 soil properties from 69 pits across multiple soil depths over an area of about 2.5 ha.

The soil pattern is quasi-symmetrical with respect to the axis of the hollow. Stagnoluvic gley soils occur on the gently sloping interfluve where lateral drainage is negligible and vertical drainage is limited by slowly permeable horizons – these soils are strongly gleyed. They give way to stagnogleyic podzolic brown soils on the convex shoulder slope, where enhanced throughflow due to the convex curvature improves drainage and leaching. Stagnogley podzols form in the hollow as a result of convergent throughflow, intensifying leaching, and saturation above the low permeability, subsurface horizons. Lower in the hollow these soils give
way to lateral podzolic soils where an open network, gravelly, colluvial deposit, inferred to have formed as a result of past periglacial processes, permits deeper and more rapid percolation. The concave footslope in the lower hollow is characterized by springs and hence is dominated by strongly gleyed soils. Along the noses, where throughflow is divergent, a catenary sequence of soils lacking gley features occurs as podzols grade to podzolic brown soils to orthic brown soils. Gleyic phases of the orthic brown soils occur in the saturation wedge that forms at the base of the noses adjacent to the seepage soils.

It is clear that the distribution of soil taxa across the slopes of the drainage basin can be reconciled in a qualitative way with the inferred patterns of topographically determined throughflow. Quantitative analysis of the spatial variation of soil chemical properties also demonstrated catenary patterns. Soil properties relating to relatively mobile soil constituents such as exchangeable bases showed the strongest spatial patterns. Between 31% and 69% of spatial variability of total exchangeable bases (TEB) could be explained using multiple regression against morphometric terrain variables calculated from a digital elevation model of the
SOILS ON SLOPES: CATENAS

**Figure 7** Distribution of soils within a zero-order drainage basin, Quantock Hills, UK (after Park and Burt 2002).

Aboveground influences on catenary patterns

Although gravitational organization may be the most consistent effect in catenary redistribution of soil materials, it may not be the dominant one in some circumstances. In grazed pastoral hill country, nutrient elements move independently of gravity, powered by energy from photosynthesis. Grazing animals harvest nutrients throughout the landscape and deposit them in concentrated form as dung and urine in preferred congregation sites (Haynes and Williams 1993). As a result, some areas of the
landscape become enriched with nutrients while others become depleted.

Preferential stock congregation often occurs on the gentle slopes of ridge tops, particularly in steep, hill country. Typically, the animals will congregate in such areas (referred to as “campsites”) through the night and graze the slopes in daylight hours. Campsites often cover 15% or less of the total grazable land area (Haynes and Williams 1993).

Soil chemistry, both organic and inorganic, is profoundly influenced by animal grazing behavior. In general, animals tend to harvest nutrients from backslopes and footslopes as they graze the pasture and deposit them on stock camps and on the tracks on which they walk. The tracks mostly run horizontally around hills (Figure 5) making it easier for the animals to move around the landscape particularly on steep slopes. Using nutrient balance approaches researchers have demonstrated that large accumulations of nutrients occur on campsites and stock tracks and a corresponding depletion of fertility develops on the majority of the hillslopes. The nutrients moved in relatively large quantities in this way include nitrogen, potassium, and sulfur in urine and phosphorus and sulfur in dung. At these rates of movement, large quantities of nutrients are translocated from hillslopes to campsites and tracks (Gillingham, Syers, and Gregg 1980). Nutrient redistribution creates feedbacks involving pasture sward composition, N-fixation, nutrient cycling, and C-fixation rates, which tend to reinforce the heterogeneity of soil fertility.

The future for hillslope-soil studies

Recent mathematical process models of catenary sequences advanced by geomorphologists have made major contributions to our understanding of soil-mantled hillslopes. These studies have incorporated soil production and transport, mineral dissolution and soil mass balance, and carbon dynamics to provide new insights into chemical denudation, redistribution of biogeochemically important elements, and C sequestration on slopes as well as within soils on slopes. Developing similar models, which incorporate a more comprehensive suite of soil and aboveground ecosystem processes, is the next major challenge.

SEE ALSO: Digital soil mapping and pedometrics; Hillslopes; Hydrologic cycle; Landforms and physiography; LiDAR; Models in geomorphology; Quantitative geomorphology; Soil mapping and maps; Soil mass balance; Soil taxonomy and soil classification; Soil water; Soils and the carbon cycle; Soils in geomorphic research; Soils of mountainous landscapes; Soils and weathering

References


SOILS ON SLOPES: CATENAS


Further reading


Soils: salinization

Graciela Metternicht
University of New South Wales, Australia

Soil degradation is the decline in soil quality caused by its improper use, usually for agricultural, industrial, or urban purposes. Processes leading to soil degradation include chemical, physical, and biological actions and interactions that affect soil functions and productivity (Lal, Hall, and Miller 1989). Factors involved in soil degradation include natural and human-induced agents and dynamics that set in motion the processes leading to changes in soil properties and life-supporting attributes (Figure 1).

Physical processes of soil degradation lead to changes in the physical, mechanical, hydrological, and rheological properties of soils. Change in soil structure is, for example, an effect of physical degradation. Factors responsible for physical degradation include deforestation, intensive row cropping, and overgrazing. Reduction in soil organic matter is an example of a biological process of degradation that can be caused, for example, by intensive cropping, or surface erosion. Chemical degradation processes include (i) changes in soil’s chemical properties that regulate nutrient activity and capacity, which maintain a favorable balance among main nutrient elements, and (ii) the accumulation of chemical substances, possibly to toxic concentrations. Soil chemical degradation often results in depletion of major plant nutrients and/or accumulation of salts and heavy metals in concentrations detrimental or even toxic to plant growth. Main processes leading to soil chemical degradation include leaching, acidification, and salinization (Figure 1).

Salinization is a natural process resulting from migration and dissemination of water-soluble salts from a source area to an area originally free of salt. It can occur in any sediment, whether it be soil or rock-like material, as long as it is porous. Hereafter in this entry, salinization will be considered mainly as a human-induced process, sometimes called secondary salinization.

Salinization results from excessive accumulation of water-soluble salts such as chlorides, sulfates, and carbonates of sodium, magnesium, or calcium on the soil surface, in the subsoil, or in groundwater. Human-induced salinization increases salt concentrations in soils already affected by natural salinity, or leads to contamination of salt-free soils because of inadequate water and land management (e.g., using brackish water for irrigation, or overgrazing).

Salinization is a concern because salts – when found above a certain threshold level – hinder the growth of crops by limiting their ability to take up water. Excess salts have the same effect on plants as does drought. For this reason, although salt is “the savor of foods” it is said to be the scourge of agriculture (Figure 2).

Salinization can take any of several forms, and a rich nomenclature has developed to describe these forms and processes (Figure 3). Soils with a high concentration of salt in the root zone are called saline soils if the electrical conductivity (EC) of the saturation extracts is above 4 dS m$^{-1}$ at 25°C, and the sodium absorption ratio (SAR) is less than 13. Soils are called alkaline or sodic if they contain sodium salts capable of alkaline hydrolysis, the SAR exceeds 13, and the soil EC
SOILS: SALINIZATION

Figure 1  Various factors and processes of soil degradation. Created with information from Lal, Hall, and Miller (1989).

is below 4 dS m$^{-1}$. Soils are called saline-alkaline or saline-sodic if the EC and the SAR are above 4 dS m$^{-1}$ and 13, respectively.

Geography and global extent of salt-affected soils

Saline soils tend to dominate in arid and semiarid regions, where low rainfall and high evapotranspiration concentrate sodium (Na), magnesium (Mg), and calcium (Ca) salts in soils, mainly in the form of chlorides and sulfates. Sodic soils are found in semiarid and subhumid regions, where Na carbonate and bicarbonate dominate, and Na ions get adsorbed on the soil exchange complex. Salinization and alkalinization processes also manifest in some coastal regions where rainfall mixes with sea spray, affecting areas located relatively far away from the coastline, as is the case with inland salt deposits present in the Western Australian “wheat belt” (Hingston and Gailitis 1976). The intrusion of saltwater into freshwater aquifers in low-lying coastal areas is another process leading to soil salinity in some delta and estuary areas of countries such as Bangladesh and India. Figure 4 presents the dominant ions causing salinity and/or alkalinity, their dominant environment of formation, and the main effects they have on soils.

Statistics on the worldwide extent of salt-affected areas vary according to data sources, and the most recent inventories are out of date (dating back to 2000). Nonetheless, according to the Food and Agricultural Organization of the United Nations (FAO), the global surface area affected by salinity was 831 million ha in 2000, including 397 million ha of saline soils and 434 million ha of sodic soils. Other estimates (IIASA and FAO 2012) raise the area affected by excess salt
Figure 2 Examples of impacts of salinity on soils in a semiarid area of Bolivia (the Cochabamba valleys): (a) puffy crust containing soil aggregates, spots of thin salt crust, and salt-tolerant Chenopodiaceae; (b) sodium sulfate-rich crust broken by cattle or sheep trampling, leaving the underlying dispersed soil material exposed to wind erosion (c); (d) topsoil degradation from combined salinization and wind erosion; (e) salt crust maximized in an irrigated parcel, at the end of the dry season; (f) irrigated agricultural field with local salt concentration affecting alfalfa growth; (g) stunted barley in a salt-affected paddock in the southwest of Western Australia.

Accumulation to 1.39 billion ha; this is about 7% of the Earth’s surface, and equates to the size of a country like the United States or Canada, or 20 times the size of Spain. Besides naturally saline areas, the global assessment of human-induced soil degradation (GLASOD), carried out in the 1980s (Oldeman, Hakkeling, and Sombroek 1991), determined that about 76.6 million ha have been salinized in “modern times” as a consequence of human activities. This figure represents 3.9% of the 1.96 billion ha affected by human-induced soil degradation worldwide (Figure 5). About 58% of all salinized soils are located in arid or semiarid, irrigated areas. On average, 20% of the world’s irrigated lands are affected by salts (45 million ha of the 230 million ha irrigated worldwide). Furthermore, of 1.5 billion ha of dryland agriculture, 32 million ha (2.1%) are salt-affected, with varying degrees of human-induced degradation (FAO 2000).

Salinized soils have expanded in area since the inventories of the 1970s and 1980s, as newly affected areas most probably exceed the areas remediated by reclamation and rehabilitation efforts. At the national level, data on salinization are generally more reliable and sometimes more current. For instance, in 2001, Australia, a country with large areas of naturally occurring saline soils, reported 2.5 million ha of land salinized since the introduction of European farming practices. These data compare to the 0.9 million ha identified by the GLASOD project in the
SOILS: SALINIZATION

Chemical indicators:
EC < 4dS m^{-1} / SAR > 13 / pH < 8.5

Dominant anions:

Dominant cations:
Na: not more than half of the soluble cations.
Ca and Mg: considerable amounts.
K: not common.

Gypsum and lime present sometimes.

Other properties:
Organic matter dispersion and dissolution.
Clay deflocculation.
Columnar or prismatic structure.

Main effects on soil profile:
Changes in structure and decreases in permeability and porosity.
Changes in soil biological activity
pH increases beyond 9 or 10.

Saline soils

Chemical indicators:
EC > 4dS m^{-1} / SAR < 13 / pH < 8.5

Dominant anions:
Chlorides, sulfates, and bicarbonates. Carbonates: small amount.

Dominant cations:
Na: dominant.
K: sometimes (exch. and soluble).
Ca and Mg: small amounts. At high pH and in the presence of carbonates, Ca and Mg will precipitate.

Other properties:
Generally flocculated; permeability equal or higher than that of similar nonsaline soils.
White crusts on soil surface.

Main effects on soil profile:
Higher osmotic pressure.

Saline-alkaline (sodic) soils

Chemical indicators:
EC > 4dS m^{-1} / SAR > 13 / pH variable

Dominant anions/cations:
If excess of salts: appearance and properties similar to saline soils (i.e. pH < 8.5, particles remain flocculated).
If soluble salts are leached downwards: appearance and properties similar to sodic soils (i.e., pH > 8.5 and soil particles are dispersed).

Main effects on soil profile:
Soils become unsuitable for the entry and movement of water and for tillage.

Figure 3  Types of salt-affected soils and associated chemical and physical parameters. Created with information from Richards (1954).

late 1980s (Oldeman, Hakkeling, and Sombroek 1991). Projections for Australia conclude that salinization may affect a total of 17 million ha of valuable farmland by 2050 (National Land and Water Resources Audit 2000). For Australia, salinization has become one of the most costly forms of land degradation.

Salinity intertwines with other major soil degradation processes such as particle size dispersion, flocculation, and compaction, and as a result often leads to increased soil erosion. Degradation assessment and hazard prediction require that the causes of salinization be clearly identified; this topic is discussed later.

Salinization: causes and process

The primary minerals of igneous and metamorphic rocks, mainly silicates and alumino-silicates, are the original source of all salts found in nature. Upon weathering, primary minerals release cations and anions that combine to form a variety of salts such as chlorides, sulfates, carbonates, and bicarbonates of sodium, calcium, magnesium, and potassium, among others. The most soluble salts are removed as water traverses the terrain surface or percolates through the soil profile. Thus, salts commonly are translocated...
Figure 4  Geography and types of salt-affected soils and their effect on soils. Created with information from Szabolcs (1989).

Figure 5  Global assessment of the status of human-induced soil degradation (GLASOD) as of 1990. Chemical deterioration includes: salinization/alkalinization; acidification; pollution; and loss of nutrients/organic matter. Created with information from Oldeman, Hakkeling, and Sombroek (1991).
from their original areas to (i) deeper locations within the soil, and (ii) lower landscape positions. In both cases, they naturally contaminate soils or soil horizons that might have originally been salt-free. The onset of salinity causes salts to be redistributed in the soil profile according to their solubility under prevailing climatic, topographic, and hydrologic conditions, and increasingly under the influence of land use and management practices (Figures 6 and 7). More soluble salts are translocated to greater depths, and to lower-lying landscape positions. These factors affect the soil water balance and, therefore, the movement and accumulation of salts within the soil profile (Zinck and Metternicht 2009).

**Identification and classification of salt-affected soils**

Throughout the world, different approaches are followed to identify and classify salt-affected soils. In the 1960s, Russian soil scientists developed a system based on *salt types*, that is, the chloride, sulfate, and carbonate *anion ratios* present in the soil saturation extract (Plyusnin 1964). Using this classification, types of saline soils include sulfate (SO₄) soils, chloride-sulfate soils, sulfate-chloride soils, and chloride (Cl) soils, all according to their Cl/SO₄ ratio. The *World Reference Base for Soil Resources* (Sparr 1994) also follows an approach based on anion groupings, distinguishing six facies of salt-affected soils (Table 1).
Good practices

- A healthy tree cover uses groundwater reserves and evapotranspiration which keeps the water table at a safe depth.

- A vegetative cover together with minimal runoff ensures surface stability.

- The lower slopes of a well-timbered catchment permit a range of productive agricultural land uses.

- A low water table does not bring salts to the surface.

- Land degraded by saline seepage and affected by a high water table severely limits productive agricultural activity.

- Surface stream becomes saline through runoff from saline seepage and interception from the water table.

Poor practices

- A cleared catchment increases infiltration which in turn raises the water table. A minimal amount of moisture is transpired while an increase is experienced in surface runoff.

- A rising water table brings natural salts towards the surface, killing existing vegetative cover.

- A low water table does not bring salts to the surface. A rising water table brings natural salts toward the surface, killing existing vegetative cover.

- Land degraded by saline seepage and affected by a high water table severely limits productive agricultural activity.

- Surface stream becomes saline through runoff from saline seepage and interception from the water table.

Figure 7 Environmental and human factors controlling dryland salinity. Wyong Shire Council (2015).

Table 1 A soil salinity approach based on anion assemblages. Adapted from the World Reference Base for Soil Resources (Spaargaren 1994).

<table>
<thead>
<tr>
<th>Soil type</th>
<th>Facie</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chloride soils</td>
<td>Acid chloride soils</td>
<td>Cl &gt;&gt; SO₄ &gt; HCO₃, and Na &gt;&gt; Ca</td>
</tr>
<tr>
<td></td>
<td>Neutral chloride-sulfate soils</td>
<td>nearly neutral pH</td>
</tr>
<tr>
<td>Sulfate soils</td>
<td>Neutral sulfate soils</td>
<td>nearly neutral pH, Na &gt;&gt; Ca, and SO₄ &gt;&gt; HCO₃ &gt; Cl</td>
</tr>
<tr>
<td></td>
<td>Acid sulfate soils</td>
<td>very low pH (less than 3.5)</td>
</tr>
<tr>
<td>Carbonate soils</td>
<td>Alkaline bicarbonate-sulfate soils</td>
<td>pH &gt; 8.5, HCO₃ &gt; SO₄ &gt;&gt; Cl, and Na &gt; Ca</td>
</tr>
<tr>
<td></td>
<td>Strongly alkaline soils</td>
<td>pH &gt; 10, HCO₃ and CO₃ &gt;&gt; SO₄ &gt;&gt; Cl, and Na &gt;&gt; Ca</td>
</tr>
</tbody>
</table>

Alternatively, the classification system developed by the US Department of Agriculture Soil Salinity Laboratory (Richards 1954) makes no distinction between salt types and considers only the total salt level in the soil as the guide to its classification – in this case, estimated from the EC of the saturation extract, expressed in dS m⁻¹ at 25°C, and the exchangeable sodium percentage (ESP) or SAR. Using these data, soils are classified as saline, saline-alkaline, or alkaline, as presented in Table 1. In addition to the chemical and physical soil properties, Soil Taxonomy (IUSS Working Group WRB 2014) uses diagnostic horizons to further discriminate among salt-affected and similar soils, including the salic horizon with soluble salts and the
SOILS: SALINIZATION

natric horizon with high levels of exchangeable sodium, as well as gypsic and petrogypsic horizons, sulfuric horizons, and calcic and petrocalcic horizons.

Effects of salts on soil properties

The accumulation of soluble salts in soils produces changes in both physical and chemical properties, mainly affecting soil structure, permeability, and chemical balance. The specific effects of water-soluble salts on soil properties depend on environmental conditions as well as on the types of ion causing the salinity and/or alkalinity.

In agricultural areas, high levels of water-soluble salts are harmful to crops. However, determination of a salt percentage limit for agriculturally limiting saline and nonsaline soils is very difficult, because some crops are more tolerant than others to the effects of high salt content. Furthermore, not all salts are equally harmful: sodium salts, for instance, are more harmful to most crops than are calcium and magnesium salts (Table 2). The detrimental effect of adsorbed sodium on the physical properties of soils results in low water infiltration rates, low permeability rates for water and gases, and most importantly, poor soil structure or even loss of structure entirely. The degree to which the physical condition of the soil deteriorates in the presence of exchangeable sodium is a function of content and mineralogy of clay particles versus the electrolyte concentration of the soil solution.

Mapping and monitoring soil salinity distribution

The management of salt-affected areas requires information on salinity-alkalinity classes and their severity. Conventional ways of determining, and then tracking, changes in soil salinity are based on field observation and laboratory analysis, both of soils and of crops. Usual indicators of developing or increasing salinity include the invasion of salt-tolerant weeds, irregular, patchy patterns of crop growth, lack of plant vigor, and other indicators of crop stress that result from the direct effect of high osmotic pressures on plant growth. Other, nonbiological indicators of salinity problems that can be observed in the field include the presence of white crusts and efflorescences in saline soils and dark spots in alkaline soils, and the presence of white spots and streaks in both saline and alkaline soils. Where salt efflorescence is not visible, laboratory determinations of soil parameters such as pH, ESP, and SAR can detect early onsets of soil salinity. All of these salinity indicators and estimations are used in commercial dry farming, the management of irrigation schemes, and the reclamation and rehabilitation of salt-affected areas for agricultural production. They are best applicable at farm level and, in the case of laboratory determinations, might be too expensive for frequent replications, if the intent is to enable accurate monitoring of salinity changes. Remote sensing tools, therefore, offer complementary data, and often constitute a less costly, more versatile and timely option to conventional monitoring, especially at regional or smaller scales (Metternicht and Zinck 2003).

Table 2 Harmful (bold italics) and less harmful salts. Adapted from Plyusnin (1964).

<table>
<thead>
<tr>
<th>NaCl</th>
<th>Na₂SO₄</th>
<th>Na₂CO₃</th>
<th>NaHCO₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>MgCl₂</td>
<td>MgSO₄</td>
<td>MgCO₃</td>
<td>Mg(HCO₃)₂</td>
</tr>
<tr>
<td>CaCl₂</td>
<td>CaSO₄</td>
<td>CaCO₃</td>
<td>Ca(HCO₃)₂</td>
</tr>
</tbody>
</table>
Summary and conclusions

Soil degradation is the decline in soil quality caused by improper land use and management practices, usually for agricultural, pastoral, industrial, or urban purposes. It encompasses a wide range of physical, biological, and chemical deterioration. Currently, about 33% of the world’s land masses have been reported as highly or moderately degraded (FAO 2011).

Soil salinization is a specific form of chemical soil degradation, frequent in arid and semi-arid regions and low-lying coastal areas, in which it affects soil functioning and productivity. In agricultural areas, salinity is of concern when the accumulation of salt in the root zone is at a concentration that causes a reduction in yield, often accompanied by other changes in physical soil properties. Yield reductions occur when the crop is unable to extract sufficient water from the salty soil water solution. Salinization also affects other soil degradation processes such as particle size dispersion, flocculation, and compaction. It changes soil structure and the soil biological activity. Therefore, it is considered a major global environmental problem, and in countries such as Australia, salinization is the most costly form of land degradation.

SEE ALSO: Geographic information system; Land degradation; Land-use/cover change and climate; Landforms and physiography; Optical remote sensing; Radar remote sensing; Soil erosion and conservation; Soil mapping and maps; Soils of desert landscapes

References


**SOILS: SALINIZATION**


---

**Further reading**


Sonar

ZhaoBao Fang

Guangdong Society of Remote Sensing and Geographic Information Systems, China

Overview

Sonar is the acoustic equivalent of radar. Pulses of sound are used to probe the sea, and the echoes produced are then processed to extract information about the sea, its boundaries, and submerged objects. An alternative to sonar, known as passive sonar, focuses on similar goals by listening to the sounds radiated by underwater objects.

Sonar was originally the acronym derived from SOund Navigation And Ranging. Today, however, its meaning has far exceeded this original definition. Sonar is the use of sound waves propagated in water for underwater target detection, positioning and communication. It is one of the most important and widely used underwater acoustic techniques.

Sonar technology has been used for approximately 100 years. In 1906, Lewis Nixon invented the first sonar receiver to detect icebergs. During World War I, submarine detection became essential and, as a result, interest in sonar increased. The first sonar equipment made was passive receiving equipment, in which no signal was generated. In 1918, Britain and the United States established a sound transmission system. The system used sound waves that were transmitted and received using sonar equipment. The invention of the sound sensor and effective sound emission made it possible to create more advanced forms of sonar.

Sonar is currently the main technique used by naval forces for underwater surveillance, target detection, classification, positioning, and tracking. It is used for underwater communications and navigation, and by security vessels, anti-submarine aircraft, and anti-submarine helicopters in weapon systems and for tactical maneuvers. Sonar technology is also widely used for torpedo guidance, mine fuse and fish detection, offshore oil exploration, marine navigation, underwater operations, hydrological measurements, and seabed geology reconnaissance.

Structure and working principles of sonar

Sound waves are used to detect, observe, and measure objects in water, as they possess advantages over other detection methods that can detect only at very short distances in waters. The penetration of light in water is limited, even in clear water; objects can only be observed at a distance of tens of meters. Electromagnetic wave attenuation in water is too rapid; the shorter the wavelength, the greater is the loss, even with high-power, low-frequency electromagnetic waves, which only span dozens of meters. However, the attenuation of acoustic waves propagated in deep water means that it is possible to receive signals from the detonation of a few kilograms of explosive material at a depth of 2000 m. Additionally, low-frequency sound waves can penetrate the sea bottom through thousands of meters of stratum to obtain formation information. Measurements and observations made with electromagnetic waves
in water have yet to be proven more effective than those made with sound waves.

Generally speaking, sonar array devices, electronic cabinets (for analyzing the echo signal, processing and recognizing the electronic signals), and auxiliary equipment exist as three separate parts. In a geometric pattern array an acoustic transducer is usually spherical, cylindrical, flat, or an in-line column, with a receiving array, a transmitter array, or a single, combined receiver/transmitter array. Electronic cabinets generally have a bisected launch, receive, display, and control system. The auxiliary equipment comprises a power supply, a connecting cable, an underwater junction box, a repeater, and sonar array transmission control matched with the lifting and rotating, pitch, receiving, towing, dipping, release device, sonar dome, and so on.

The transducer — a device that converts energy from one form to another — is an important device in sonar, converting not only sound energy but also other forms of energy, such as mechanical, electric, and magnetic energy. For example, the piezoelectric sensor is a type of energy conversion sensor that can convert mechanical energy into electricity and electrical energy into mechanical energy. An electrical signal transmitter and a transducer (piezoelectric crystal) are commonly used to transform an electrical signal into a sound signal, which is then launched into water. If any submarine, mine or fish targets are present in water, sound is reflected back to the receivers as electrical signals, which are then processed to enlarge them and displayed on a screen or presented as an audio signal through headphones (Figures 1 and 2).

According to round-trip signal time, the distance to a target can be determined and the high and low tones used to evaluate the character of the target. For example, submarines possess a steel shell and, therefore, the echo is clear and has a long playback time. Echoes from deep fish are chaotic. If a target is in motion, the echo of a tone changes as a result of the relative motion of the observer and target.

**Animal sonar**

Many animals possess sonar. Bats emit a sound from their larynx 10–20 times per second. Bat ears receive ultrasonic pulses back as echoes. With the aid of this type of “sonar,” small insects and obstacles as small as a 0.1 mm thick wire can be detected. With the aid of sonar that has a range 40 meters greater than that of a bat’s ultrasound, insects such as moths can clearly hear to avoid attacks. Therefore, animals and humans can experience “sonar war.”

---

**Figure 1** Working principle of sonar.
The sonar sensitivity of dolphins is very high and can detect small objects, such as thin wire or nylon rope. Fish can detect a target from several hundred meters away, can cover the eyes and can move quickly through bamboo poles without running in to the poles. Dolphin sonar target recognition ability is very good; not only can dolphins identify different fish and distinguish yellow copper, aluminum, Bakelite, plastic and other materials, but also they can separate their voice echoes from the voice echoes of people and record and replay them as sound waves. Dolphin sonar anti-interference ability is also very impressive. If noise interference exists, dolphins can improve the sound intensity over the noise to make their own judgments. Dolphin sonar also has emotional expression ability; dolphins “talk” through their sonar system.

Many whales use sound to detect and communicate. It is at a much lower frequency than that of dolphins; their sound travels over a far greater distance. Other marine mammals, such as seals and sea lions, emit sound signals for detection as well.

Animals in the extreme dark depths of the ocean have been forced to adopt various means, including sonar, to search for prey and avoid being attacked. Animal sonar performance is far more advanced than modern human technology. Thus, solving the animal sonar mystery remains a research topic of modern sonar technology.

Russian researchers, using the principle of dolphin sonar, developed a new radar positioning method. They reported that compared with the traditional radar positioning method, their new strategy’s main difference is that it can simultaneously measure the distance of an object and its moving velocity or acceleration. With multiple radar transmitters, this method can be applied to better control the movement of a ship, making its docking capability more accurate at a large dock.

**Application and development of sonar technology**

Early submarines relied on periscopes for observation. However, a periscope can only observe the surface of a target, and observing an underwater target is pointless; therefore, there was a high accident rate in early submarines, which often hit reefs, mines, and other underwater submarines. During World War II, German submarines sank more than 100 ships.

Modern submarines are equipped with a variety of sonar technology. For example, a United States submarine that is equipped with sonar technologies for different uses can possess more than 15 types of sonar. Vessels with a sonar detection instrument can intercept and listen to an enemy’s sonar signal. If sonar is employed, a special password is used to assay the signal. Communication sonar can be used for communication by the vessels, while other sonar is responsible for navigation, ranging, vigilance, detection, measuring landforms, and so on.
SONAR

In many countries, in the areas near to naval ports and important straits, large sonar transducer arrays are often installed in the main channel and other places, and thousands of transducers on seafloor are controlled by computer control onshore. Accordingly, a submarine force can be found in a timely manner. Latent prevention early warning systems of this type were built as early as 1952 and are now in their fifth generation of development, in which alerts can be generated up to several hundreds of kilometers away.

In addition to fixed warning sonar systems, submarines can use airborne sonar for detection. A helicopter can lower a 100-meter-long cable to act as a cable hanging sonar. Sonar depth changes in seawater according to the fall or rise of a fuselage. A plane flying above the sea can undertake detection over a large area (dragging sonar).

A new airborne dipping sonar has been developed that is “wireless,” so does not need to connect with a cable and a plane. It is only 10 kg in weight and is dropped by anti-submarine aircraft into a predetermined area, where it floats on the surface of the sea. Anti-submarine aircraft can deploy many of these floating sonar, which have an antenna out of the water and a hydrophone submerged in the water. The hydrophone converts underwater acoustic signals into received electrical signals that are then sent out through the antenna. According to the signals received, the anti-submarine aircraft can determine the position of a submarine.

In the past 20 years, underwater sonar technologies with respect to positioning, target detection and recognition, communication, navigation, remote control, remote ocean floor sensing (bottom, formations, landforms, etc.), mineral, oil, and gas exploration and military developments have seen extremely widespread application.

Detecting sonar systems in the marine environment

Three research and application aspects exist for sonar in marine environments. The first application uses sonar for marine pollution detection. Using an acoustic backscattering instrument that records acoustic scattering intensity, the concentration of waste in the ocean can be measured. The second application concerns the design of ocean engineering structures, for which marine environment parameters, especially wave and current statistics, employ sonar systems for measurements. The principle here is to fix a sonar system at the bottom of the sea to record sound waves, launch sound waves in the direction of the bottom of the sea, and receive reflections from waves; thus, the propagation time of rolling waves of the relevant parameters can be detected. The third application is to use a low-frequency sound wave that can spread far in the ocean to measure water temperature change; this is similar to medical chromatography.

Marine surveying, mapping, and exploration

With the aid of the global positioning system (GPS), sonar systems, such as sounding sonar, side-scan sonar, sonar imaging, and so on, can generate an electronic chart, on-site, in real-time integrated with the ship’s navigation system, establishing an important technical foundation. Regarding acoustics, the Chinese Academy of Sciences, Institute of the East China Sea Station, National Ocean Technology Research Institute, and other units have been successful in profiling sea bottom sediment, the bottom of the sea, and geological structures of shallow and deep seas, as well as in engaging in seabed oil and gas resource exploration, underwater archaeology, port construction, oil pipeline and sea platform installation, and engineering. Engineering has
been widely used to equip underwater robots with many types of sonar.

**Integrated ship sonar system research**

As demonstrated during the Gulf War, ships and, especially, submarines play important roles in modern naval battles. When ships reach a certain depth, for underwater activities sonar is only effective as an external information tool. Additionally, from the perspective of anti-submarine warfare, submarine sonar represents the main counter. Because the future naval combat system is comprehensively integrated and distributed, an integrated ship sonar system configuration with new requirements has been proposed. The integrated ship sonar system uses information provided by sonar, including comprehensive processing, display, and control, to complete detection, location, tracking, and recognition processes.

Currently, the study of integrated ship sonar systems at home and abroad follow a number of development trends. The first is the continued development and research on towed linear sonar sensor arrays, multiple sensors, and perfecting the function of the complementary base sonar (Figure 3). Multisensor configuration implies that a high density of observable data can improve the credibility and accuracy of target data. The second development trend is the integration of a sonar system with data fusion technology. An integrated sonar system is configured with multiple sensors. The amount of information gathered from the various sensors is vast and must be filtered through target information fusion technology for dimension reduction to provide integrated sonar systems with optimal and unique target information sets. The third development trend is the study of underwater acoustic countermeasure systems for ships. Using computer simulation technology,
research vessel sonar, and acoustic confrontation, the demonstration and design phase before sonar development must be completed to assess whether its performance can meet the practical requirements, providing a theoretical basis for determining the acoustic equipment performance indicators. The underwater ship acoustic countermeasure simulation system is mainly composed of attackers, defenders, and public control, among which are the offensive submarine platforms and weapons (e.g., wire-guided torpedoes, remote control mines, etc.); defenders are composed of ship’s surface and defense weapons.

Underwater acoustic signal processing
and underwater target recognition research

Modern sonar and acoustic technology primarily focuses on research into underwater acoustic signal processing and target recognition because of their importance and complexity. Due to the complexity of the marine environment, underwater target identification has become increasingly complex. Because microelectronics, signal detection, processing, and computer technology are inseparable, modern sonar technology has become very advanced.

The main research directions are: (i) matching underwater acoustic signal processing problems and (ii) digital sonar design. Digital sonar has undergone six generations of development; digital signal processing technology, computer system structure, and microelectronics technology have been combined in digital sonar design. This technology is being extended to generate (iii) a new method of increasing the accuracy of underwater target recognition. Wavelet analysis and fractal theory are used for the extraction of target parameters, using a fuzzy inference neural network classifier design, which is a combination of neural network and expert system identification research.

Weapons and mine applications in water

Weapons used in water primarily comprise torpedoes and mines. Torpedoes are active offensive weapons, whereas traditional mines represent passive defensive weapons. Modern mines are
intelligently designed, having the ability to be disguised and hidden, as well as to attack. A torpedo can be launched after passive sonar has been used to search for and track targets; mines and especially modern mines (remote control mines) use a sonar system that relies on their own sonar detection and target identification.

Mine warfare is also closely related to sonar technology. It generally requires high-performance, high-precision equipment when searching for underwater bombs. Currently, side-scan sonar and depth-variable sonar are commonly used. The armed forces began with synthetic aperture sonar.

Special dolphin forces

Dolphins are an object investigation due to their striking ability of using “natural” sonar in water. Dolphins have the ability to distinguish obstacles and the nature of artificial objects from natural objects in the water. Dolphins can find the location of mines and signal divers to conduct demining work. Due to a dolphin’s search capabilities, special dolphin forces have been formed (Figure 4).

SEE ALSO: Digital elevation model and digital surface model; Maritime transport

Further reading

Geographic information systems (GIS) process both spatial and attribute data (also known as feature data). Their performance is greatly dependent on the representation of the data and the extent to which they can integrate the data. Ideally, this integration is done through the use of database management systems that are designed to deal with attribute data and, hence, must be modified to also handle spatial data. One of the main issues for spatial data is the volume of the data, which affects how they are stored and accessed, and what operations will be applied to them.

There are two types of operations. The first type of operation processes the data in their entirety, as is the case with operations such as map overlay and connected component labeling in the case of raster images. At times, the operations are applied to several datasets, in which case they should be applied in such a way that the same locations in the two datasets can be accessed at the same time. This means that the order in which the data are processed is of importance; this is usually achieved by sorting the data.

The second type of operation processes only a small subset of the data, as is the case when the data are accessed at random or the portion of the data that is processed satisfies some predicate, which can involve both its spatial and/or attribute components. An example is when browsing through the data looking for cities with a population greater than one million within 10 miles of rivers (Brabec and Samet 2007; Esperança and Samet 2002; Samet et al. 2003). The latter type of processing is important as its efficient execution (other than using brute force to examine every data item) brings into play the need to incorporate notions of sorting, as its use is a prerequisite for its efficient retrieval.

In this entry, the use of sorting to enable the efficient execution of operations on spatial data is discussed. The discussion focuses on the type of the spatial data, which in this case fall into raster and vector. The application of sorting to raster data, where the main motivation is the fact that the data are extremely voluminous and, thus, all of them cannot be fit into memory at once, is reviewed in the next section. Therefore, external storage must be used. In this case the motivation for the sorting is to enable efficient execution of operations. Later, the application of sorting to vector data is reviewed, although, as will be pointed out, the methods that are described are also applicable to raster data. Concluding remarks are drawn in the final section.

**Raster data**

The natural representation for raster data (i.e., 2-D images) is as an array of pixels, where the array serves as an access structure to the raster image. The problem is that even for moderately sized images the storage requirements are high, and thus the array does not always fit into memory. Hence, it is stored on disk in some order with the aid of a mapping from the multidimensional space (two in this case) to a 1-D space (i.e., $Z \times Z$ to $Z$), which reflects
the order in which the array elements (i.e., the pixels here) are stored. There are many ways of ordering the pixels that make up a raster image. The objective of the ordering is to provide a systematic way of processing all of the pixels with the property that every pixel in the image will be processed (i.e., none are missed). Any ordering that is devised must satisfy this property, as otherwise it is not useful. Such orderings are termed space-filling curves (Sagan 1994).

Some of the most important orderings for a 2-D space are illustrated in Figure 1 for an 8 × 8 portion of the space and are described briefly here. Of course, orderings can also be devised
for data of three dimensions and higher, but examples involving them are beyond the scope of this discussion. Choosing among the different orderings is not easy because each one has its advantages and disadvantages. A few of their desirable properties are reviewed here and which of the orderings satisfy them, and which do not, are indicated.

The mapping from the higher-dimensional space to the integers should be relatively simple, and likewise for the inverse mapping. This is the case for all but the Peano–Hilbert order (Figure 1d). For the Morton order (Figure 1c), the mapping is obtained by interleaving the binary representations of the coordinate values of the pixel in the grid. The result of the mapping (i.e., the number associated with each pixel) is known as its Morton number. The Gray order (Figure 1g) is obtained by applying a Gray code to the result of bit interleaving, and the double Gray order (Figure 1h) is obtained by applying a Gray code to the result of bit interleaving the Gray code of the binary representation of the coordinate values. The U order (Figure 1i) is obtained in a similar manner to the Morton order, except for an intermediate application of \( d - 1 \) “exclusive or” (\( \oplus \)) operations on the binary representation of selected combinations of the coordinate values prior to the application of bit interleaving. Thus, the difference in cost between the Morton order and the U order in \( d \) dimensions is just the performance of additional \( d - 1 \) “exclusive or” operations. This is in contrast to the Peano–Hilbert order, where the mapping and inverse mapping processes are considerably more complex.

The ordering should be stable. This means that the relative ordering of the individual locations is preserved when the resolution is doubled (e.g., when the size of the two-dimensional space in which the pixels are embedded grows from 8×8 to 16×16) or halved, assuming that the origin stays the same. The Morton, U, Gray, and double Gray orders are stable, while the row (Figure 1a), row-prime (Figure 1b), Cantor-diagonal (Figure 1e), and spiral (Figure 1f) orders are not stable. The Peano–Hilbert order is also not stable, as can be seen by its definition. In particular, in 2-D, the Peano–Hilbert order of resolution \( i + 1 \) (i.e., a \( 2^{i+1} \times 2^{i+1} \) image) is constructed by taking the Peano–Hilbert curve of resolution \( i \) and rotating the NW, NE, SE, and

![Figure 2](image)

Figure 2  Peano–Hilbert curves of resolution (a) 1, (b) 2, and (c) 3. Copyright H. Samet.
SORTING SPATIAL DATA

SW quadrants by 90° clockwise, 0°, 180°, and 90° counterclockwise, respectively. For example, Figure 2(a–c), gives the Peano–Hilbert curves of resolutions 1, 2, and 3, respectively.

Two pixels that are adjacent in the sense of sharing an edge or a side (also known as 4-adjacent) in space are neighbors along the curve defined by the ordering and vice versa. This property is impossible to satisfy for all pixels at all space sizes. However, for the row-prime, Peano–Hilbert, and spiral orders, every element is a 4-adjacent neighbor of the previous element in the sequence, while this is not the case for the other orders. This means that the row-prime, Peano–Hilbert, and spiral orders have a slightly higher degree of locality than the other orders.

The process of retrieving the neighbors of a pixel should be simple. Finally, the ordering should be admissible (Dillencourt, Samet, and Tamminen 1992) – at each pixel position in the ordering, at least one 4-adjacent neighbor in each of the lateral directions (i.e., horizontal and vertical) must have already been encountered. This is useful in several algorithms (e.g., connected component labeling). The row and Morton orders are admissible, while the Peano–Hilbert, U, Gray, and double Gray orders are not admissible. The row-prime, Cantor-diagonal, and spiral orders are admissible only if the direction of the 4-adjacent neighbors is permitted to vary from position to position along the curve. For example, for the row-prime order, at positions on odd rows, the previously encountered 4-adjacent neighbors are the western and northern neighbors, while at positions on even rows, they are the eastern and northern neighbors.

The row order (Figure 1a) is of special interest because its mapping function is the one most frequently used by the multidimensional array, which is the most common access structure for a raster image. Assuming that the first entry in the linear structure that is accessed by the mapping has an index value of 0, location \((a, b)\) in an \(8 \times 8\) collection of pixels is mapped to index value \(8b + a\). An alternative ordering that is also used in the multidimensional array is known as column order, where the difference from row order lies in the order in which the various dimensions are scanned. In the column order, the \(y\) coordinate value (row number) varies most rapidly; in the row order, the \(x\) coordinate value (column number) varies most rapidly. The one to choose is arbitrary, although the row order is preferred as it yields a lexicographic ordering when array references are of the form \(T[i, j]\), corresponding to an element in row \(i\) and column \(j\) of array \(T\) (i.e., at pixel address \((x, y) = (j, i)\)).

The Morton order has a long history, having been first mentioned in 1890 by Peano, and is often used. It is also known as a \(Z\) order and as an N order. The Peano–Hilbert order was first mentioned soon afterwards by Hilbert and a number of researchers.

The \(U\) order is a variant of the Morton order but also resembles the Peano–Hilbert order. The primitive shape is a “U,” which is the same as that of the Peano–Hilbert order. However, unlike the Peano–Hilbert order, and like the Morton order, the ordering is applied recursively with no rotation, thereby enabling it to be stable. The \(U\) order has a slight advantage over the Morton order in that more of the pixels that are adjacent (i.e., in the sense of a \((d - 1)\)-dimensional adjacency) along the curve are also neighbors in space. This is directly reflected in a lower average distance between two successive positions in the order, which can also be shown to be lower than that of the Gray and double Gray orders. However, the price of this is that, like the Peano–Hilbert order, the \(U\) order is also not admissible. Nevertheless, like the Morton order, the process of retrieving the neighbors of a location in space is simple when the space is ordered according to the \(U\) order.
At times, the number of pixels may be so large that application of statistical data compression techniques from coding theory may be worthwhile. However, the interest is in methods that exhibit progressiveness. Therefore, as results of processing the data are obtained, the partial results of the operation can be seen rather than having to wait until the operation is complete. The result is that we limit ourselves to the situation where many of the pixels have the same data values in which case we try to reduce the required storage by grouping them into blocks of identically valued constituent pixels. This is especially the case for binary images or other images where adjacent pixels often have the same value. For example, consider a crop map where each crop type is assigned a different numeric code or color.

There are two ways to group identically valued pixels into blocks. The first aggregates consecutive identically valued pixels into 1-D blocks and is termed a runlength encoding (Samet 2006). This is particularly useful in the case of the row order (Figure 1a), where the elements of the ordering are pairs of the form \((a, b)\) where \(a\) is the value representing color/type and \(b\) is the length of the block.

The second approach applies a 2-D aggregation into blocks where the sizes of the blocks are constrained to be powers of two and are located at specific positions. In particular, assuming an origin at the upper-left corner of the image, the coordinate values of the upper-left corner of each block (e.g., \((i, j)\) in two dimensions) of size \(2^i \times 2^j\) satisfy the property that \(i \mod 2^i = 0\) and \(j \mod 2^j = 0\). The resulting block decomposition is known as a region quadtree (Klinger 1971), which also serves as the basic data structure in the QUILT (Shaffer, Samet, and Nelson 1990) GIS and spatial library. Such a block decomposition is particularly useful in the case of the Morton, Peano–Hilbert, Gray, double Gray, and U orders (Figure 1). In this case, again, the elements of the ordering are pairs of the form \((a, b)\) where \(a\) is the color/type and \(b\) is the side length of the square block.

The result of the ordering is an effective linearization of the data. The orderings are adequate when the operations require that all elements in the dataset (i.e., image) be accessed from start to end. However, some operations require access to particular elements of the array (known as random access) and, therefore, performing a sequential scan from the start is very inefficient. In this case an access structure is useful and can be the 1-D array corresponding to the result of the mapping from 2-D to 1-D. However, this is more complex once a start is made grouping identically valued blocks and keeping them in the ordering, as the image elements corresponding to the elements in the ordering are no longer the same size. Thus, a tree-like access structure is re-sorted to, where associated with each element of the ordering is a number corresponding to its position in the ordering, and this is the number used to access the tree. The tree can be a binary search tree, B-tree, and so on. For example, in the case of the Morton ordering this number is formed concatenating the result of interleaving the binary representation of the \(x\) and \(y\) coordinate values of the pixel in the upper-left corner of the block (assuming an origin in the upper-left corner of the image) and the binary representation of the base two logarithm of the block’s side length. Traversing the tree in the order NW, NE, SW, and SE yields an ordering of the blocks in increasing order.

**Vector data**

The application of sorting to vector data has its roots in visualization applications in computer graphics. The earliest examples are the hidden-line and hidden-surface elimination
algorithms due to Warnock that recursively decompose the picture area into rectangles that are successively smaller while searching it for areas that are sufficiently simple to be displayed. The determination of what part of the picture area is hidden or not is equivalent to sorting the picture area with respect to the position of the viewer. The concept of “sorting” is used in two ways in the above hidden viewing algorithms. The first is the conventional one of ordering the visible part of the picture area with respect to the viewer. However, a different interpretation of “sorting” is given when the picture area is recursively decomposed into rectangles stopping when their complexity is reduced. In this case, the dictionary definition of “sorting” is being used, which is one of “putting in a certain place or rank according to kind, class, or nature”. Notice the absence of “ordering” in the definition.

Notwithstanding the above definition, sorting usually implies the existence of an ordering. Orderings are fine for 1-D data. However, they do not exist in two dimensions and higher. For example, suppose all of the cities in the United States are sorted by their distance from Dallas. This is fine for finding the closest city to Dallas, say with population greater than 200,000. However, the same ordering cannot be used to find the closest city to Miami, say with population greater than 200,000, without re-sorting the cities. In contrast, once the sort is made with respect to a reference point in one dimension (e.g., people by their height), the nearest value for any other reference point can be found; there is no need to re-sort the data.

The problem is that for two dimensions and higher, the notion of an ordering does not exist unless a dominance relation holds – that is, a point \( a = \{a_i | 1 \leq i \leq d\} \) is said to dominate a point \( b = \{b_i | 1 \leq i \leq d\} \) if \( a_i \leq b_i, \ 1 \leq i \leq d \). Thus, the only way to ensure the existence of an ordering is to linearize the data, as can be done, for example, using a space-filling curve as described in the previous section. The problem with such an approach is that the ordering is explicit. Instead, what is needed is an implicit ordering so that it is not necessary to re-sort the data when, for example in the sample query, the reference point for the query changes (e.g., from Dallas to Miami). Such an ordering is a natural byproduct when objects are sorted by spatial occupancy, and is the subject of the rest of this section.

The indexing methods that are based on sorting the spatial objects by spatial occupancy essentially decompose the underlying space from which the data are drawn into regions called buckets in the spirit of classical hashing methods with the difference that the spatial indexing methods preserve order. In other words, objects in close proximity should be placed in the same bucket or at least in buckets that are close to each other in the sense of the order in which they would be accessed (i.e., retrieved from secondary storage in case of a false hit, etc.).

There are two principal ways of implicitly sorting spatial data. The first distinguishes occupied from unoccupied space, while the second sorts the regions so that the number of spatial objects that they contain is within the same range.

The first method makes use of an object hierarchy that initially aggregates objects into groups based on their spatial proximity and then uses proximity to further aggregate the groups, thereby forming a hierarchy. Note that the resulting object hierarchy is not unique as it depends on how the objects were aggregated to form the hierarchy. Queries are facilitated by also associating a minimum bounding box of an appropriate shape (e.g., hyper-rectangle, sphere) with each object and group of objects, as this enables a quick way to test if a point can possibly lie within the area spanned by the object or group of objects. Negative answers mean that no further
processing is required for the object or group, while a positive answer means that further tests must be performed. Thus, the minimum bounding box serves to sort the space according to the occupation status. Data structures such as the R-tree (Guttman 1984) and the R*-tree (Beckmann et al. 1990) illustrate the use of this method. As an example of an R-tree, consider the collection of straight line segment objects given in Figure 3a shown embedded in a $4 \times 4$ grid. Figure 3b is an example of the object hierarchy induced by an R-tree for this collection. Figure 3c shows the spatial extent of the bounding rectangles of the nodes in Figure 3a, with heavy lines denoting the bounding rectangles corresponding to the leaf nodes, and broken lines denoting the bounding rectangles corresponding to the subtrees rooted at the nonleaf nodes.

The drawback of the object hierarchy approach is that the resulting hierarchy of bounding boxes leads to a nondisjoint decomposition of the underlying space. This means that if an object is not found in one search path starting at the root, then it does not mean that the object will not be found in another search path starting at the root. This is the case in Figure 3c when searching for the line segment object that contains Q. In particular, first nodes $R_1$ and $R_4$ are visited unsuccessfully, and thus there is the need to visit nodes $R_2$ and $R_5$ in order to find the correct line segment object $i$.

The second method is based on a recursive decomposition of the underlying space into disjoint blocks, so that a subset of the objects associated with each block satisfies some predetermined criterion. There are several ways to proceed. The first is to simply redefine the decomposition and aggregation associated with the object hierarchy method, so that the minimum bounding rectangles are decomposed into disjoint rectangles. This implicitly partitions the underlying objects that they bound. In this case, the partition of the underlying space is heavily dependent on the data and is said to be at arbitrary positions. The K-D-B-tree (Robinson 1981) and the R$^+$-tree (Sellis, Roussopoulos, and Faloutsos 1997) are examples of such an approach.

The second way is to partition the underlying space at fixed positions so that all resulting cells are of uniform size, which is the case when using the uniform grid, also the standard indexing method for maps. Figure 3a is an example of a $4 \times 4$ uniform grid in which a collection of straight line segments has been embedded. The drawback of the uniform grid is the possibility of a large number of empty or sparsely filled cells when the objects are not uniformly distributed. This is overcome by using a variable resolution representation such as one of the quadtree variants (Samet 2006), where the subset of the objects that are associated with the blocks are defined by placing an upper bound on the number of objects that can be associated with each block (termed a stopping condition for the recursive decomposition process).

The PM$_1$ quadtree (Samet and Webber 1985 (also the related PMR quadtree (Nelson and Samet 1987)) is an example of a variable resolution representation for a collection of straight line segment objects, such as the polygonal subdivision given in Figure 3a. In this case, the stopping condition of its decomposition rule stipulates that partitioning occurs as long as a block contains more than one line segment unless the line segments are all incident at the same vertex which is also in the same block (Figure 4). A similar representation has been devised for 3-D images (Ayala et al. 1985) where no block contains more than one face, edge, or vertex unless the faces all meet at the same vertex or are adjacent to the same edge all in the same block. In this example, the PM$_1$ quadtree represents the polygonal subdivision.
by its constituent objects, which are the line segments of arbitrary orientation. When the line segments are constrained to be orthogonal to the coordinate axes, the line quadtree (Samet and Webber 1984) uses a decomposition rule in terms of both the boundary and interior of the subdivision so that decomposition into blocks takes place until no boundary element passes through the interiors of the blocks.

In the case of regions, the quadtree decomposition rule halts the decomposition once the data in the block are uniform. It has been used primarily in two and three dimensions where they are known as octrees (Meagher 1982), as well as for surface data (Sivan and Samet 1992). In the case of point data, quadtrees enable the execution of many queries without having to perform explicit sorts of the data. They are

Figure 3  (a) Example collection of straight line segments embedded in a 4 × 4 grid; (b) object hierarchy for the R–tree corresponding to the objects in (a); (c) spatial extent of the minimum bounding rectangles corresponding to the object hierarchy in (b). Notice that the leaf nodes in (c) also store bounding rectangles, although this is only shown for the nonleaf nodes. Copyright H. Samet.
especially useful for determining the nearest object to a particular location (i.e., a “pick” operation in computer graphics). In this case, the underlying space is recursively decomposed into four congruent square blocks until each block contains no more than a predetermined number of points (a variant of a bucket PR quadtree) (Samet 2006). The advantage of the implicit sorting of the underlying space into the blocks is that the position of the subdivision lines vis-à-vis the position of the query object can be used to prune certain blocks from further consideration when executing the operation. Note that the same pruning properties also hold when the blocks are not congruent, as in the case of the point quadtree (Finkel and Bentley 1974; Samet 1980). Quadtree and their variants are to be distinguished from pyramids (Aref and Samet 1990; Tanimoto and Pavlidis 1975), which are multiresolution data structures.

The principal drawback of the disjoint method is that when the objects have extent (e.g., line segments, rectangles, and any other nonpoint objects), then an object may be associated with more than one block. This means that queries such as those that seek the length of all objects in a particular spatial region will have to remove duplicate objects before reporting the total length. Nevertheless, methods have been developed to avoid these duplicates by making use of the geometry of the type of the data that are being represented (Aref and Samet 1992) to avoid multiple reporting of the same object. Another approach known as the MX–CIF quadtree (Kedem 1982) represents each object just once by associating it with its minimum enclosing quadtree block. The loose quadtree (Ullrich 2000) and the cover fieldtree (Frank and Barrera 1989) take this approach further by expanding the size of the minimum enclosing quadtree blocks, so that the size of the expanded minimum enclosing quadtree block is not completely independent of the size of the object (Samet, Sankaranarayanan, and Auerbach 2013).

Note that the result of constraining the positions of the partitions means that there is a limit on the possible sizes of the resulting cells (e.g., a power of two in the case of a quadtree variant). However, this means that the underlying representation is good for operations between two different datasets (e.g., a spatial join) (Hoel and Samet 1995) often implemented as top–down tree traversals (Samet 1985), as their representations are in registration (i.e., it is easy to correlate occupied and unoccupied space in the two datasets, which is not easy when the positions of the partitions are not constrained, as is the case with methods rooted in representations based on an object hierarchy, even though the resulting decomposition of the underlying space is disjoint).
Concluding remarks

An overview of the utility of sorting spatial data for use in geographic information systems has been provided. A distinction has been made between the application to raster data and vector data. This is important, as in the case of raster data the sorting is based on ordering the actual data, while in the case of vector data the sorting is more in terms of the underlying space in which the data are embedded. Hence, the sorting is characterized as being explicit in the former and implicit in the latter. It has been pointed out that sorting is useful in finding nearest objects. In this case, the distance was measured in terms of “as the crow flies.” However, these methods are also useful in finding neighbors in road networks where the distance is defined along the network (Sankaranarayanan and Samet 2010). The sorting has also been used in a distributed peer-to-peer setting (Tanin et al. 2005).

Acknowledgement

This work was supported in part by the National Science Foundation under Grants IIS-10-18475, IIS-12-19023, and IIS-13-20791.

SEE ALSO: Data structure, raster; Data structure, vector; Indexing; Spatial database

References


Further reading


For more about the different raster orderings: as well as a comprehensive discussion of spatial data structures, see the discussion in Samet 2006.

For JAVA applets that illustrate a variety of spatial data structures including many of those describe here: [http://donar.umiacs.umd.edu/quadtree/index.html](http://donar.umiacs.umd.edu/quadtree/index.html) (accessed May 5, 2016).
“Soundscape” is both a popular term and an academic concept. As the former, it is used widely by the general public, and particularly by individuals working in sound-related sectors (such as the music and film industries) and related media. As the latter, it is increasingly used by researchers across the sciences and social sciences, either in passing as a useful descriptor or as a dedicated subject of study.

Soundscape is rarely narrowly or strictly defined, and thus is open to a wide variety of interpretations and understandings. In basic terms, it is a metaphor – a play on the word “landscape” – that denotes the way in which sound fills, and makes, environments in various ways and at various scales. The connection between landscape and soundscape is also more than just metaphorical. Because soundscapes unfold in environments – that is, across time and space – a landscape of some description necessarily has to exist for a soundscape to exist (even if it is very modest in scale, such as a small room, for example). Soundscape is often, but not exclusively, talked about in the context of musical sound. Soundscapes can, however, be composed of a range of nonmusical natural and human-made sounds – such as those found in natural environments (Matless 2005), outside in towns and cities (Raimbault and Dubois 2005), and within buildings. These sounds – including those from electronic devices, motor vehicles, human speech, construction work, wind, and rain – are of course the most common components of soundscapes, constituting a constant auditory backdrop to, and an essential part of, everyday life (albeit not always the most obvious, explicit, or entertaining).

A wide range of academic disciplines discuss soundscapes, including acoustic ecology (from where it “officially” originates), the sociology of music, musicology, ethnomusicology, cultural studies, urban planning, and, largely because of its spatial dimensions and character, human geography; specifically the subfields of music geography (Connell and Gibson 2003) and sonic geography (Matless 2005). It was, however, a seminar paper, itself named “Soundscape,” published in 1994 by Susan Smith, that made the first compelling case for the development and investigation of the concept in the parent discipline. Indeed, Smith (1994) argued that the (then) new cultural geography’s engagement with and reading of landscape had been rooted almost exclusively in the visual, neglecting how sound structures space and characterizes place. She claimed that scholars had looked at the world and had written about it without listening carefully to or hearing it. Smith posited that sound was inseparable from natural, built, and social landscapes and that, consequently, scholarship that does not consider sound is incomplete in a sensory sense. Moreover, rounding off her case, she argued that sound signifies place, enhances place’s spectacle, and reflects aspects of society in particular places. Since then, much research has been informed and influenced by Smith’s observations. In music geography the term “soundscape” emerged soon after, hand in hand with a change in emphasis whereby scholars began to look beyond the
SOUNDSCAPES

basic locational and distributional aspects of music (often involving quantitative research methods to map broad musical trends across the Earth’s surface) to consider the relationships and dynamics between music and places. This endeavor involved the use of qualitative research methods to unpack places as complex social and cultural phenomena, to understand how music is wrapped up with place experiences, meanings, attachments, representations, and identities, and to do justice to music itself; its complexity, richness, diversity, aesthetic qualities, and cultural and political language (Leyshon, Matless, and Revill 1998).

Underpinning the aforementioned tradition since the mid-1990s has been a social constructivist theoretical approach. This approach digs down to uncover meaning in, and attempts to represent and theorize, the nature of soundscape. Here a loose reading of soundscape has emerged, which does not always take it to be an immediate and momentary experience. On the one hand, research has considered the more fixed and territorial aspects of soundscapes, for example, the form and implications of musical “sound scenes” as they form in situ (Connell and Gibson 2003). This might be where particular types and genres of music become associated and entwined with places (such as particular towns or cities), or where particular forms of music become associated with particular types of urban spaces (such as subways or cafes). On the other hand, research has considered the fluidity and mobility of musical sound, including, for example, changes in musical styles across time and space (often at the international scale) and issues related to the authenticity and hybridity of the emerging forms (Connell and Gibson 2003). Research has also considered the economic, political, and cultural drivers, mechanisms, and mediators for such movements, often locating them within forms of nationalism, regionalism, and wider globalization processes (Connell and Gibson 2003). In these ways – through both fixing and fluidity – a soundscape can be a multiscaled phenomenon, by which sound takes a prominent yet not exclusive role in defining and making place.

Most recently in the new millennium, the idea of soundscape has been influenced by the emergence of nonrepresentational theory in human geography, which has taken the discipline far closer to the literal and popular understanding of soundscape as an immediate – often all-enveloping – experience and environment. In contrast to social constructivism, nonrepresentational theory places far greater emphasis on sound’s performative aspects, on what actually happens in time and space and on the momentum in how it takes place. Moreover, nonrepresentational theory brings into view the subtle and often unspoken dimensions involved in soundscapes, including how they relate intimately to the human body and embodiment (people’s senses, immediate responses, and expressions). In this research, the Deleuzian concept of affect has been drawn on to help articulate the role soundscapes play in how the world is experienced and acted on precognitively (its somatically registered “feeling states,” “vibes,” and “atmospheres”). Scholars have discussed how, through affect, music might assist people’s transitions from one experiential state of their body to another, potentially impacting on their capacities and wellbeing (in either positive or negative ways). They have also considered how, within soundscapes, affective feelings might slide into more conscious realms – such as attitudes, opinions, and emotions – which are even more clearly relational to the social world.

Although the consideration of musical soundscapes remains predominant in human geography, an ever-wider range of subjects is now being tackled by scholars taking both
social constructivist and nonrepresentational perspectives. Well-established subdisciplines have become involved in the debates. In rural geography, for example, the emphasis has turned to how contests over identity, aesthetics, and land use are played out with, and over, specific sounds as much as other factors, presenting moral geographies of appropriate conduct in these areas (Matless 2005). In urban geography it has been proposed that research move beyond thinking about urban noise as pollution and a problem and seek a more neutral starting point that accounts for other forms of sound (particularly subtle forms), subjectivity/differences in attitudes and experiences, and temporal change over different time frames (Raimbault and Dubois 2005). More specifically, there are emerging interests in particular facets of soundscapes, signifying a more sophisticated engagement with and development of the concept – such as voice and conversation, silence and stillness, art and movement, and therapeutics. Methodologically, as Smith (1994) highlighted, problems will always persist when studying soundscapes, in no small part because many sounds – unlike words and images – if not recorded, disappear. Moreover, writing tends to deaden soundscapes, the change of medium killing the essential energy, and soul of the real-time experience. Hence, researchers are gradually coming to grips with a range of methodological, representational, and theoretical challenges that will help them better understand and convey soundscapes.

SEE ALSO: Affect; Culture; Emotional geographies; Music; Popular culture

References


South Africa: Society of South African Geographers (SSAG)

Founded: 1994
Location of headquarters: Bloemfontein
Website: www.ssag.co.za
Membership: 735 (as of 2013)
President: Kevin Mearns
Contact: mearnkf@unisa.ac.za

Description and purpose

The Society of South African Geographers’ main objective is to advance the research and educational activities of all South African geographers by collectively representing the interests of South African geographers regionally, nationally, and internationally; encouraging and supporting high quality research and teaching in geography; providing a national geographic information resource for geographers and interested groups; and stimulating awareness of geographic and environmental matters through academic collaboration with other intellectual communities and accountable interaction with the public at large.

Journals or major publication series

South African Geographical Journal. www.tandfonline.com/loi/rsag20

Current activities or projects

The SSAG focuses on publishing and disseminating scholarly research results in the South African Geographical Journal and other occasional publications; and by organizing a biennial conference and an annual student conference hosted by geography departments in South African universities. We also aim to support projects that build geography beyond universities and to represent geographers on various official regional and national Department of Education and National Science Foundation groups, as well as internationally through links with the International Geographical Union. The society recognizes the achievements of its members through several types of awards, among them the Jubilee Bursary (for an Honors student), the Bronze Medal (for an outstanding Master’s thesis), the Gold Medal (for outstanding service to the geographical community in South Africa), and the Fellowship award (for outstanding and sustained scholarly contributions).

Brief history

The Society of South African Geographers is an organization that focuses on building geography and supporting the work of departments across South Africa, most of which are small in size. The society was founded in 1994, during a period of tremendous change in South Africa with the advent of democracy and the restructuring of the state, universities, and the education system after apartheid. In this contemporary period, our efforts have focused on building an inclusive and supportive context for scholarly work and teaching as well as for the building of geography in schools and the public more broadly. This work builds on close to a hundred years of disciplinary history in South Africa, with the first organization founded in 1917 (the South African Geographical Society – SAGS) dominated by English-speaking universities and academics and with the Society of Geographers
(SG) founded in parallel in 1957 and representative of Afrikaans-speaking geographers, teachers, and lecturers. Continuing through and reflective of the divisions and politics of the apartheid period, these societies merged in 1994 forming the present organization inclusive of all South African geographers across the country’s racial, ethnic, and linguistic divide, and a significant focus of the SSAG in its contemporary form.

Submitted by Sophie Oldfield
South Korea: Daehan Jiri Hakhoe (The Korean Geographical Society)

Founded: 1945
Location of headquarters: Seoul
Website: www.kgeography.or.kr
Membership: 1400 (as of December 31, 2012)
President: Bokyung Yang
Contact: ybk@sungshin.ac.kr

Description and purpose

The Korean Geographical Society was founded in September 1945 to promote the “advance-ment of geographic research” and has since been at the forefront of geographical research in South Korea. The society also serves as the Korean Committee of the International Geographical Union. The society has around 1400 individual members and more than a hundred member organizations.

Journals or major publication series


Current activities or projects

The society advances geography in academia, industry, research, and education through its journal, newsletter, research projects, annual meetings, and educational programs. The society publishes the *Jirihak*, a quarterly newsletter, and is committed to publishing scholarly empirical and theoretical research articles. The society conducts various research projects which are either under contract with government agencies or jointly with private sector organizations. Through the contract with the Ministry of Land, Infrastructure, and Transport, the society has published the *Korean Geography* and *National Atlas* series. The annual meetings are the largest events of the society and approximately 150 papers and presentations on geographic topics are delivered by 600 attendees. A wide array of research symposia, informative workshops, special events, field trips, exhibits, and paper and poster competitions for students are also prepared. The society runs the Geography Olympiad, which is designed to challenge Korea’s most talented students and reward them for their success. Each year, more than 4000 high school students representing their schools participate in the Olympiad on university campuses across the country.

Submitted by Jeong Rock Lee
Sovereignty

Ishan Ashutosh  
*Indiana University, USA*

Sovereignty has been traditionally defined as a political authority practicing ultimate power over a given territory and population. The sovereign has the monopoly of violence and the determination of right through the making of laws within its jurisdiction. Moreover, the sovereign produces spaces subject to its power, as associated with territory as well as a population who are united as the people in their subjection to the supremacy of the sovereign. There is an internal and external dimension to sovereignty, since in principle a sovereign is not subordinate to other political authorities and acts autonomously within its jurisdiction, yet sovereignty is predicated on its recognition as such by other sovereign powers. Today sovereignty is most commonly associated with the state and indeed, the concept acts as the primary basis of state legitimacy. Transformations of political authority over time and space, however, reveal that sovereignty is a problematic concept and one that is best seen as complex, contingent, and Janus-faced.

The distinct and often contradictory understandings and utilizations of sovereignty over the past five centuries reflect broader struggles over order, legibility, and the operation of power that shape geographic processes of politics and law. Embedded within the very concept of sovereignty are histories of conflict, domination, and resistance, although sovereignty is often granted an a priori status that takes the concept for granted. As the expression of absolute political authority, sovereignty remains critical in the range of activities and processes that comprise social and political life, including membership and belonging to a political community, the rights and obligations of subjects and citizens to political authority, the conducting of war, and the regulation of human migration. Over the past quarter century, international treaties and the appearance of supranational entities, alongside the rise of transnational capital and international human migration have seemingly signaled the withering away of state sovereignty. Such evaluations, however, have been shown to be premature and scholarship in geography has departed from the “end of sovereignty” accounts. Instead, geographers have provided grounded and theoretically rich insights into the spatial transformations of sovereignty and its intersection with other forms of political entities and organizations to reveal that the concept is characterized by a range of multiple and contradictory practices that characterize the diffusion of political authority.

**The longue durée of sovereignty: from the sixteenth century to the present day**

As a concept in Western philosophy and political organization, the emergence of sovereignty was as discontinuous as it was fractious and contested. Saskia Sassen (2006) has shown that in Europe during the mid-Medieval period power was shared between three forms of political organization: the church, feudalism, and empire. By the thirteenth century, sovereignty became an increasingly important principle for monarchical authority, and by the Tudor period had found representation in the split of the king’s two
bodies – the body natural and the body political (Kantorowicz 1957). On the one hand, the body natural was the physical body of the king, and thereby corruptible and mortal. On the other, the immortality of the body political attempted to portray the indissolubility of the sovereign as a godlike form of political power that resides above and in control of the finitude of human life.

In conventional accounts of sovereignty, its rise is owed to the gradual erosion of the Church’s political authority in favor of political power being located in the state. By the sixteenth century, the juridical conception of sovereignty based on rights, the duties of subjects, and the monarch became central to political conflict in Europe and were concomitantly elaborated on in debates on the nature of politics and the state. For the French political theorist Jean Bodin, the supremacy of the sovereign rested on a moral and legal basis and reveals the shift in thinking about the legitimacy of political power when contrasted with Niccolò Machiavelli’s earlier realist argument that political authority is simply predicated on the ability to rule and not on a moral or even a legal basis. In spite of such differences the political philosophies from the late sixteenth to the mid-seventeenth centuries highlight the shift in thinking about how political power is achieved, maintained, and legitimated.

The treaties that comprised the Peace of Westphalia (1648), designed to bring an end to the long-standing feuds and wars of major European political powers that still consisted of overlapping forms of political authority are frequently seen as the primary manifestation of sovereignty in territorially based states. Basic aspects of Dutch jurist Hugo Grotius’s political and legal philosophies, which stressed the importance of sovereignty as the ability to act without external interference, became a cornerstone in the treaties of Westphalia. In calling for singular political authorities over demarcated space in which a political power would have exclusive authority, the treaties enshrined a system of state sovereignty predicated on mutual recognition. Moreover, the interstate system ushered in by Westphalian sovereignty produced political authorities whose legitimacy was increasingly independent of papal command in contrast to the shared, overlapping, and conflicting political organizations that existed during the medieval period.

For the social contract theorists of the Enlightenment, juridical sovereignty would produce a political order that would progress society beyond the state of nature or a society without the protection of the sovereign that would provide order and security. Like the philosophies that underwrote the treaties of Westphalia, conceptions of sovereignty were cast in light of political conflict and colonial conquest that raised fundamental concerns over the legitimacy and status of sovereignty. The English Civil War, which pitted Charles I against Parliament, with the former arguing that political power was ordained from God and encompassed a duty to provide security to the people in exchange for their subjection, served as the immediate political context for Thomas Hobbes’s *Leviathan* (1968/1651). Hobbes depicted the pre-sovereign state of nature as one of constant war and fear that can only cease with a social contract between the people and the sovereign. In exchange for security, the people subsume their freedom in the state of nature to the sovereign, who is tasked with maintaining peace and order. The figure of the leviathan, a behemoth that stands above both territory and its subjects illustrates the commanding vision and authority of the sovereign.

While John Locke’s state of nature departed from the Hobbesian view, sovereignty was nevertheless needed in order to protect property. Much like Hobbes, Locke also argued that sovereignty rested on a contract with the people, rather than between God and the king. The shift
from divine right as the basis of monarchical power to the people as the basis of sovereignty. Made its largest pronouncement during this time in the decidedly anti-monarchical works of Jean Jacques Rousseau. For Rousseau, the sovereign is nothing other than a representation of the general will of the people and as such, Rousseau elaborated and expanded on the notion of popular sovereignty. The sovereign’s identity is equivalent to the people in their entirety and is therefore an undivided expression of the will of the people, a philosophical understanding of sovereignty that has been used for gaining legitimacy by democrats and totalitarians alike. Post-Westphalian philosophies of sovereignty revolved around a number of key issues that remain fundamental to understandings of contemporary sovereignty, most notably the transfer of sovereignty to the people and the ties between subject/citizens and sovereign protection.

To locate the concept of sovereignty as merely a narrative of European statehood, however, elevates that Westphalian moment to a mythic status and produces a history that itself mirrors sovereignty’s desire for legitimacy. Sovereignty was in fact founded on violence and not simply through the convening of a peace conference attended by representatives of the European powers. The consolidation of the political entity of Europe based on political equivalencies among sovereign powers and enshrined in the principle of reciprocity as the basis of Westphalian sovereignty, furthermore, elides how the philosophies and political treaties that characterize sovereignty were largely conjured beyond Europe. Given this context, sovereignty is best seen as a political ideal of supreme authority that also emerged through the violence and oppression of slavery and colonialism.

Though traditional accounts of the philosophies of sovereignty view the term as the political corollary of the broader Enlightenment project, the concept should not be severed from its development and application in colonial conquest and subjugation. Sovereignty is, in the first instance, constituted through projects of violence and colonialism. By the sixteenth century, European attempts to dominate much of the world proved to be the grounds under which the uneven application of sovereignty was utilized to dispossess non-European political entities. The universality of political organization was dismissed under the arrogation of European sovereignty over other lands that were in turn violently stripped of their sovereignty through claims of despotism, civilization, order, and peace. The Berlin Conference of 1884–1885, to provide one notable example, extended European sovereignty by stripping African political powers of their autonomy and organization in the name of European colonialism. In the colonial context, the violence of sovereignty becomes clear and is not merely an aberration of sovereignty. The ability of the sovereign to determine arbitrarily what constitutes right as a measure of its authority is achieved at the expense of its colonized subjects.

The logic of sovereignty was also expressed in anticolonial movements, which argued that their political and cultural sovereignty were oppressed and demanded recognition. Following formal decolonization, in many instances postcolonial nation-states are beset by fragmented sovereignty, in which more than one political organization attempts to act as the final authority, leading to displacement, pogroms, riots, and daily performances of violence in an attempt to mimic the supreme authority of sovereignty that was founded through the dispossessing acts of colonialism. Of course, the postcolonial context reveals a central problematic of sovereignty that extends well beyond the condition of oppressed states, by highlighting the fact that sovereignty is fundamentally predicated on violence and exclusion. Moreover, the violence of colonialism
SOVEREIGNTY

and sovereignty illuminates the variance of sovereignty in which some states practice greater sovereignty within their borders than others. The uneven application of sovereignty can be extended to the international arena in which contemporary colonialism and imperialism continue in the name of sovereignty as it shapes the politics of war and marginalization in the present day. The invasions of Afghanistan, Iraq, and drone strikes in Pakistan, for instance, are all acts carried out in the name of American sovereignty, but such displays of sovereign power violate the sovereignty of other states.

Geography and sovereignty: from authority over territory to authority over life and death

While sovereignty has long been a key concept in the discipline of geography, the focus has tended to be on the objects of sovereignty, such as population and territory, or on its institutional organization, as with the state. In the late nineteenth and early twentieth centuries, the emphasis of geopolitics on territory and natural resources during the age of high imperialism championed the control of space as central to national sovereignty as well as global political power. For instance, Halford Mackinder’s “Heartland Thesis” postulated that control of a particular geographic zone located in the central Eurasian landmass would lead to world domination, and thus a global sovereign. Similarly, Fredrich Ratzel’s call for expanded sovereignty for the purposes of lebensraum (living space) ensured that the discipline of geography would become a crucial site in knowledge production in the aid of empire building. Dominant geopolitical discourse based on a defense of national sovereignty would come to play the premier role in the exacerbation of inter-imperialist rivalries and in the first and second world wars.

Recent scholarship in political geography, especially in critical geopolitics, has called into question the realist model of sovereignty championed by classical political geography and has cast new light on the discipline’s role in promoting colonialism and imperialism by positioning sovereignty as a necessary spatial strategy in maintaining political control and order. Poststructural and postcolonial perspectives in particular have been used to reconceptualize sovereignty within the discipline as well as in the related fields of political science and international relations. Sovereignty is today understood as having been unmoored from the territorial nation-state and is the product of a range of political actors beyond the state. The territorial nation-state is no longer necessarily the locus of sovereignty, as de facto forms of sovereignty that include diasporic political organizations characterized by long-distance nationalism, global financial institutions that act with greater sovereignty than many states, and networks of public–private partnerships now also dominate the political landscape.

The disarticulation of sovereignty from its traditional referents has led to new research paradigms examining contemporary political power, including the movement of rights and obligations associated with citizenship beyond national territory and new forms of political authority predicated on a mixture of state and nonstate actors. Given the diversification of contemporary political authority, it is therefore necessary to distinguish between different types of sovereignty. Agnew (2009) identifies four distinct “sovereignty regimes” that are formed in relation to state authority and territoriality. These regimes consist of “classical” sovereignty based on the territorial state of the Westphalian model, the imperialist type in which economic control supplants territorial sovereignty, integrative
sovereignty composed of levels of political power beyond the state, and the globalist sovereignty regime in which ultimate political authority rests on the hegemony of the global sovereign over and often with the consent of the sovereigns of dominated states. In each of these ideal types, the fundamental aspects of sovereignty, such as authority, territory, and legitimacy are reconfigured and respatialized through the new international division of labor and transnational capital.

Current research on sovereignty centers on territoriality, or the use of territory by political actors for domination and subordination. The use of territory includes border enforcement (Coleman 2007), extra-territorial detention (Mountz 2011), and new practices of citizenship and belonging that are made possible through a dispersed network of sovereignty. The dispersal of sovereignty beyond national territory engages with not only the spatial distribution of sovereign power, but also on new forms of subject-making predicated on forms of inclusion and exclusion from the political community. In many cases, insights into the functioning of contemporary sovereignty stem from the work of German political theorist Carl Schmitt, who as a constitutional adviser to both the Weimar Republic and then to the National Socialists, viewed sovereignty in Hobbesian terms in which its legitimacy in a constant state of conflict rested on securing order. Schmitt’s (1985/1922, 5) formulation that the “sovereign is he who decides on the state of exception” has engendered a conception of sovereignty that focuses on the exclusions that political authority rests and relies on. For political philosopher Giorgio Agamben (1998), who also contends that sovereign power is manifest in the production of the exception, in the figure of homo sacer, or sacred man, who represents the politicization of bare life, it is the sovereign’s control over the physical wellbeing and very life of its citizens/subjects that grants the sovereign supremacy over all other forms of political authority. Agamben cites the concentration camp as the exemplary space of exception that grants sovereign power its authority. The camp also points to the fundamental paradox of sovereignty, namely, that the sovereign exists both inside and outside of the law. Agamben’s exclusive emphasis on state sovereignty and bare life has generated important geographical analyses of the spaces of exception that include the Guantanamo Bay detention camp, refugee camps, and prisons.

Another theoretical strand of contemporary research on sovereignty is indebted to Michel Foucault’s concepts of biopolitics and biopower (2003). Like those of Agamben, these terms are meant to focus on the ways in which natural life becomes central to political power, although the emphasis here is on the regulation of bodies and the production of the subject through disciplinary power. Most significantly, Foucault has urged us to look beyond juridical conceptions of sovereignty, which, it is argued, have constrained understandings of how modern political power functions. The anachronistic Hobbesian figure of the sovereign towering over territory and representing a centralized form of power above and apart from its subjects obscures the ways that subjects are produced through a dispersed disciplinary power. Biopolitics and biopower have refined analyses of the intersection of sovereignty alongside other forms of political authority that produce new geographies of political control and contestation.

Research in geography has interrogated sovereignty from multiple perspectives, as a spatial practice of political authority, as a discourse that enables and legitimizes political control, and as a means for domination over others. What unites these diverse strands is the challenge to think through the concept of sovereignty rather than dismissing it as having eroded under conditions of globalization. While sovereignty
SOVEREIGNTY

does not operate in its ideal form as theorized in
traditional accounts as the centralization of polit-
ic power in the supreme authority of the state,
the concept remains crucial to understanding the
geographies of political power and the produc-
tion of territory, the nation-state, and citizenship.

SEE ALSO: Geopolitics; Globalization;
Nation-state; Power; Security; State, the;
Territory and territoriality

References

and the Bare Life. Stanford, CA: Stanford University
Press.
Lanham, MD: Rowman & Littlefield.
Coleman, Mathew. 2007. “A Geopolitics of Engage-
ment: Neoliberalism, the War on Terrorism, and
the Reconfiguration of US Immigration Enforce-
New York: Picador.
Books. (Original work published in 1651.)
Kantorowicz, Ernst. 1957. The King’s Two Bodies: A
Study in Medieval Political Theology. Princeton, NJ:
Princeton University Press.
Archipelago: Detention, Haunting, and Asylum
on Islands.” Political Geography, 30: 118–128.
Sassen, Saskia. 2006. Territory, Authority, Rights: From
Medieval to Global Assemblages. Princeton, NJ:
Princeton University Press.
Schmitt, Carl. 1985. Political Theology: Four Chapters
on the Concept of Sovereignty. Chicago: University of
Chicago Press. (Original work published in 1922.)
The concept of space of exception has become particularly relevant in geography and the social sciences in the context of the “war on terror” launched by the Bush administration after the 9/11 terrorist attacks in New York and Washington, DC. The term is often used to describe an extraterritorial enclavic space in which the juridical order is intentionally suspended for a number of purposes, including the imprisonment of individuals via procedures dismissive of “the rule of law.” The academic literature adopting this concept has accordingly focused mainly on the most recent proliferation of camps, often described indeed as spaces of exception, and in particular on the paradigmatic case of the infamous Guantanamo Bay detention complex, located in an extraterritorial enclave in Cuba controlled by the US military, where suspected terrorists have been detained for years with essentially no legal protection and no formal arraignment. The US-led program of “extraordinary renditions” of suspected terrorists launched in the same period has also been sometimes defined as a globally connected archipelago of enclavic spaces of exception, based on detention centers located in a large number of countries around the world, not only kept largely secret in terms of their location and function, but also exempt from the effects of the ordinary legislation operative in those countries.

More broadly, the concept of space of exception is often recalled to qualify many diverse situations in which the “normal” juridical order is suspended (sometimes temporarily, sometimes permanently) in order to pursue specific military and police objectives (like indefinite detention, torture, and other unconventional treatment of detainees) beyond the rule of law, especially in reference to international legislation concerning prisoners and the respect of individual human rights.

While the arbitrary enforcement of exceptional rules to specific places and moments in time is certainly not new historically, the concept of space of exception discussed here may be understood as somehow the by-product of relatively recent academic debates focused on the parallel concept of state of exception. A key influence in these latter debates is that of Italian philosopher Giorgio Agamben, whose widely cited analysis of the principles of modern sovereignty is largely based on a critique of Carl Schmitt’s theory of sovereign exception. In his influential Political Theology, German conservative legal thinker Carl Schmitt (1998/1922) famously described the sovereign as “he who decides on the exception,” and who therefore has the capacity to determine, in case of emergency, the state of exception. While denouncing Schmitt’s affiliation with the Nazis, Agamben in both Homo Sacer (1998) and State of Exception (2005) draws on this definition to explain the workings of contemporary sovereign power via the regime of exception, in particular by making reference to Nazi biopolitics (and the Bush administration’s arbitrary approach to the rule of law after 9/11), but also, more generally, to the functioning of contemporary democratic societies.
modern totalitarianism can be defined as the establishment, by means of the state of exception, of a legal civil war that allows for the physical elimination not only of political adversaries but of entire categories of citizens who for some reason cannot be integrated into the political system. (Agamben 2005, 2)

Agamben’s work is particularly important for geography, since it shows how, to be implemented, the state of exception must happen somewhere and must be linked to specific spaces. These spaces, where the rule of law is suspended and where arbitrary power may be exercised on subjected individuals, are indeed described by many geographers as spaces of exception. Spaces of exception are often linked to the existence of broader geographies of exception, consisting of the networks and the overall relations of power that make these spaces of exception effective, and part of a specific set of political interventions normally justified by situations of emergency.

The concentration camp is often presented as a paradigmatic space of exception. Furthermore, a space of exception emerges every time sovereign power is enacted through legitimate exceptions to the “normal” rule, every time specific subjects are “banned” from ordinary legislation and the violence operated on them is exempt from sanction. In other words, a space of exception is created when, despite the existence of juridical order, a situation of perceived emergency strips an individual or a group of legal protection and, in some cases, even enables the killing of them without committing a crime. Agamben, again, has described this situation as one that reduces individuals to bare life (zoe), to biological bodies deprived of a political identity, and therefore entirely controlled by the sovereign power. The significant corpus of literature that has in different ways incorporated the notion of space of exception, including, among many others, the works of Butler (2006) and Mbembe (2003), suggests that the existence of spaces of exception is indeed fundamental to all biopolitical regimes, like the proliferation of camps and the realization of a greater geography of exception on the part of both the Nazis and Stalin’s Soviet Union have demonstrated.

In geography, the concept of space of exception has been initially popularized by Derek Gregory, especially in his works on Guantanamo and on the production of bare life in the global geographies of the war on terror, including the US interventions in Afghanistan and Iraq, but also the political situation in Palestine (Gregory 2004, 2006). Here Gregory explains that, while Agamben often refers to the state of exception as the space of exception, this spatiality must be understood as a performance, a doing, in order to show how the passages between “inside” and “outside,” law and violence, are effected. After these initial interventions, the concept of space of exception has then been used in geography not only for work on camps, prisons, and detention centers of all kinds, but also to describe special economic zones, borders, airports, refugee hospitality centers, and, more generally, all the sites where the suspension of the rule of law is enacted.

While Agamben goes as far as claiming that the camp has become the “paradigm” of our time, work in geography engaging with his theory of sovereign exception (sometimes even in critical terms) has attempted to link the contemporary geographies of exception to broader questions of geopolitics as well, showing how the creation of lawless spatialities has often become a strategic tool in waging war, or in managing the mobility of refugees or of individuals deprived of any juridical status (i.e., protection). Spaces of exception in fact sometimes also characterize the compounds offering temporary hospitality to illegal migrants landed on the northern Mediterranean coasts, or the detention
(centers) of individuals caught while transgressing the topographies of surveillance enforced along the United States–Mexico or Schengen borders.

Empirical work has illustrated how these spatialities of exception may operate on a permanent basis in cases like that of the Palestinian refugees hosted for decades in Lebanese camps, but may also emerge in sudden temporary spatiotemporal windows created by the ad hoc operations of CIA agents in countries such as Italy, where a local imam suspected of terrorist connections was kidnapped in Milan in 2003. Spaces of exception may be generated by the sudden intervention of the regular police as well, when conditions perceived as an emergency threat allow them to legally suspend the ordinary rule of law and, if deemed necessary, even to kill – as in the controversial episode that led to the death of a Brazilian electrician in the London Underground in July 2005, three weeks after London had suffered a series of lethal terrorist attacks directed at its transportation system.

Finally, much work has been preoccupied with the normalization of spaces of exception – like the Guantanamo detention complex – where the juridical order may be permanently suspended. The consolidation of countless spaces of exception around the world, together with the establishment of a broader US-led global geopolitics of exception in the post-9/11 age – the “drone war” patrolling and hitting “exceptional” targets being one case of the exceptional spatiality of war deserving particular attention – has therefore drawn lots of academic attention to this concept. This was also the response to much-needed new theorizations of power in order to understand the changing geo-biopolitical global conditions and the increasingly pervasive implementation of biometrical forms of control and policing intervention that affect our everyday lives, and in relation to which anyone can potentially be suddenly projected into a (spatial) regime of exception.

**SEE ALSO:** Biopolitics; Carceral geographies; Geopolitics; Political geography; Sovereignty

**References**


**Further reading**


Space

Nigel Thrift
*Schwarzman Scholars, USA*

We live in a world that is fundamentally diverse, a world of endless complexity, with billions of human inhabitants, animals, plants, and objects. Such a world cannot be reduced to one or even a few things, and that is why geography exists. Geography doesn’t just document the diversity and richness of the world. It doesn’t just celebrate it. It is *about* difference and what that means. It is about the connections in each moment that lead to change. Because things exist in space and as space, they do not – indeed, they cannot – stay the same. They move. They spread. They stumble. They stutter. They fracture and they corrugate. They widen their contacts. They multiply discrepancies. They do all kinds of things that would be impossible if all that we looked at was time, a dimension which is often wrongly depicted as though it were somehow of greater significance than space.

And that world’s diversity and richness are stunning in both scale and extent. Take the total mass weight of living things on the planet – the weight of life itself. It is reckoned to be 1.7 trillion tonnes, excluding bacteria. The overwhelming majority of this quantity is plant life – probably about 99%. The total human mass is about 350 million tonnes. Cows add in another 520 million tonnes. There are about 4200 million tonnes of fish and 2700 million tonnes of ants. Then bacteria weigh in at nearly the same as all other living things put together, at about 1.3 trillion tonnes (D’Efillipo and Ball 2013).

Now, take in the fact that there are estimated to be somewhere between 5 and 100 million different species of plants and animals that currently exist (with most experts fixing on around 8–30 million), many of which still remain to be described and archived (Maro et al. 2011).

Now, put all of this biomass and all of these species together in a vast network of different locations where they are mixed together as different ecologies that require different kinds of interdependence, ecologies which are themselves vital actors since through them new pressures to adapt and new combinations of life can come into existence, not least because all of the evidence suggests that these ecologies evolve towards more and more complexity.

Now stir human beings into the mix. They have had an enormous impact on the Earth’s surface, as geographers have known for a long time. We live in what is now called the Anthropocene: a permanent alteration of the geological record resulting from the actions of human beings. Climate change is perhaps the most obvious example of that impact. Climate change has been written about in many ways but still perhaps the most extraordinary figure that can be mustered up is that our global civilization is powered by around 13 terawatts of human-made energy. To give some points of comparison: the flux of energy from the center of the Earth is about 40 terawatts; the primary production of the biosphere via photosynthesis is 130 terawatts; and the Earth receives 170 000 terawatts from the sun (Morton 2009). The point is that human beings already act as an energy source that begins to compare with the energy generated by plate tectonics. But human impact on the Earth has
been expressed in many other ways too. This entry will mention just two of them.

Most particularly, human beings make objects: buildings, tools, machines, and all manner of general bric-a-brac. By one estimate, an average human being will come across something like 20,000 different objects in their lifetime. These objects now inhabit the world with us – as part of us. Then, human beings move things around – and around. These include objects, of course, but also plants and animals. Plants and animals can mix quite well without our help. Indeed, recent scientific studies suggest that epic journeys by plants and animals – so-called long-distance dispersal – may well have been a norm of life and a key driver in evolution: plant seeds were carried in the plumage of ocean-going birds, frogs and mammals as large as monkeys were cast this way and that by driftwood and icebergs, tiny spiders drifted hither and thither on storm winds, and so on. These journeys may have been flukes but their consequences were not (de Queiroz 2014).

All that said, it is clear that human beings have redistributed a large part of the Earth’s flora and fauna, sometimes gradually, sometimes through ecological blitzkriegs. They have reassembled large parts of the biosphere in ways which challenge the whole idea of “native” species (Thompson 2014). Thus, the photosynthetic machines we call plants have regularly swapped countries and continents. For example, Britain only had 35 native tree species a few thousand years ago but many hundreds of others have now been added to the roster. Animals have been even more spatially promiscuous. Take the humble rat. The rat has traveled with human beings wherever they have journeyed on the Earth’s surface. Indeed, so numerous and well adapted to almost any conditions are these fellow travelers that they might well claim to be the main beneficiares of human evolution (Zalasiewicz 2009).

The Pacific rat (Rattus exulans) has traveled all around the Pacific but its travels have been as nothing compared with the Norway rat (Rattus norvegicus), actually from China originally, which has become common almost everywhere, and especially in large cities such as New York. Or take the starling, introduced to North America by a Bronx pharmacist intent on bringing to New York every bird mentioned in the works of Shakespeare and now the continent’s most numerous avian population. Or take that living fossil and geographical adventurer, the ubiquitous cockroach. The average New Yorker is most likely to encounter the German cockroach, which probably originated in Southeast Asia, while the “American” cockroach is actually a native of Africa, and so it goes on.

But it doesn’t end there. We still don’t know a lot of what is in the world. Though the first age of exploration of the Earth may have ceased, we are now engaged in a second round. But this one depends on exploring our own social and technical creations as well, creations that we have made but which are not entirely within our control, some of which, indeed, could produce Doomsday in one way or another. These creations only underline that things never just add up in this world: they combine and multiply. There isn’t anything (including us) that is just one thing. There are no essences, just passages, as Bashō says. A new age of contingent mapping beckons, a mapping of “us” where what we are is no longer either certain or restricted to what we used to regard as the human: the individual human body and “mind.” In a sense, we are all geographers now trying to make sense of fragmented and therefore partial spaces in which everything is a part of something else – while also trying to cope with all of the problems the planet needs to resolve. To paraphrase the philosopher Wittgenstein, we must build the
boat that is supposed to carry us while we are already swimming in the water.

And finally, there is no reason to think that we are it, that human beings are the end point. We might keep our grip forever and go on to conquer the universe, or we might be the equally temporary equivalent of the formerly ubiquitous but now extinguished North American passenger pigeon to which we bear quite a few similarities, not least the fact that we too have descended on the Earth like a plague of locusts. We might end up like a phantom limb, remembered still but lost from the body of history – and geography – like the elephants we are so busily and criminally slaughtering. After all, let’s just remember our place in the scheme of things: “take away humans and the present world will … function quite happily as it did two hundred thousand years ago. Take away worms and insects, and things would start seriously to fall apart. Take away bacteria … and the viruses and the world would die” (Zalasiewicz 2009, 192).

**Knowing where we are in space**

Maps are important to geographers because geographers tend to think about the world as terrain to be crossed. Space is hard work. It has to be traveled. It has to be journeyed. It has to be molded. It has to be experienced. It has to be gripped in some way, even if that is with the lightest of touches. History is littered with examples of expeditions getting lost in space and grand navigational blunders leading to nautical and aeronautical disasters. But we don’t need to opt for such grandiose instances of spatial delinquency. People still get lost, even now in a world where technology seems to offer a permanent locational crutch, while many people simply fail to return home from their journey. Even today, journeying remains a stressful proposition, one which is, in many parts of the world, time-consuming and, in many countries, a major cause of death. For example, about 1.2 million people die on the roads each year globally – they do not arrive at their destination – but certain countries are far more dangerous to travel in than others: whereas countries such as Great Britain have an annual death rate of 2.75 per 100,000 inhabitants, others, such as Venezuela, are up in the thirties.

One of the first things human beings needed to know is where they were and where they were going, and as people began to make longer and longer journeys, the problems of fixing location and wayfinding became correspondingly greater. There are a lot of ways of knowing where is where and many of these do not require all the paraphernalia we have come to regard as imperative for finding that out. In the past, sophisticated means of wayfinding existed without any recourse to the battery of devices we now regard as a normal part of everyday life. Take the example of the sea. People sailed long distances without charts or any but the most basic of instruments by means of simple but effective techniques such as sun and star sightings, as well as by learning the habits of wind and water. These skills could be found equally amongst Inuit or Polynesian or Norse or Medieval navigators, although their exact use varied. Such skills could be extraordinarily sophisticated and lead to the ability to travel long distances with relative accuracy. For example, many Pacific Islanders used the rising and setting stars and a mental system which allowed them to think of their boat as static and the world as moving relative to the boat, as well as other maritime information such as ocean swells caused by islands. But as navigation improved and became susceptible to measurement and calculation, so it became possible to take both the old and the newly discovered navigational knowledge and insert it into devices
and machines. A whole history of these almost forgotten devices and machines exists, remarkable in its diversity, devices and machines which peppered the historical record as navigators attempted to work out basic problems such as the position of the sun and stars, elapsed time and, of course, longitude: rudimentary compasses, logs, sandglasses, cross-staffs and back-staffs, quadrants and sextants, watches and chronometers, all now replaced in a bonfire of space by satellite navigation (Glennie and Thrift 2009).

Often, early nautical travelers possessed only the most rudimentary of maps, pictures of coastal features and, if they were lucky, a simple chart. But, of course, maps – a part of the world represented on a paper or other surface – have become the pre-eminent way of finding where we are. Determining when the first identifiable map was created depends in part on what is counted as a map. If we take wayfinding maps as the gold standard (in that, unlike many other early maps they are less easy to contest as recognizably maps) then the earliest surviving map showing roads and other pathways on either land or water seems to be a papyrus map from ancient Egypt dating from 1160 BCE (Akerman and Karrow 2007), although there are earlier examples of ancient nonwayfinding Babylonian maps on clay tablets dating from 7000 BCE which show parcels of land as well as evidence of surveying techniques. But maps proliferated as the demand for them increased (as a result of the needs of the state (and especially colonization), trade, and a host of more mundane everyday concerns), as mapping technology became more and more sophisticated, and as it became possible to mass produce maps. Now, of course, we are surrounded by maps. Indeed, it is possible to argue that we have reached an era of mapping which is as consequential in its way as the great age of mapping that followed the initial European exploration of the globe. Maps are becoming sewn into the very fabric of everyday life through their rejuvenation as geographical information systems which are becoming the locational underpinning to just about everything and, in turn, have become one of the chief means of cultural experimentation.

Sometimes the history of maps and mapping is downplayed. But that history is, of course, one of the great human conceptual achievements born out of successive mathematical and geometrical innovations which gradually locked the world into the semblance of a grid on which, at least in theory, everything could be located from outside its location. They provided a new “grammar of the gaze” (Jacob 2006, 6).

There is a standard history of the idea of “space,” which is often assumed to run alongside this cartographic history, in which a “Newtonian” world view develops, based on a grid of coordinates which “without relation to anything external, remains similar and immovable” (Newton, cited in Jammer 1954, 33). It is a notion subsequently championed, although only in part, by philosophers such as Kant, for whom absolute space was an a priori form that organizes our cognition – an “intuition” – and was not in itself cognitive or conceptual at all. But it is also a notion which made its way into science in the shape of classical mechanics and even religion as evidence of the divine.

One of the problems with this view of the world is that it can be made to suggest that the Earth is held in a rigid, contained frame. But, beginning in the eighteenth century, another way of thinking about space began to take hold which privileged relative space in which space is no more than the relative location of things, a network of places, so that when motion is ascribed to a body rather than to its reference point that is simply an arbitrary convenience. This was the view of the philosopher and mathematician Leibniz. It prefigured Einstein’s
invention of relativity theory, which, in turn, was borne out of urgent practical necessity: the practical spatial challenges of the late nineteenth century provided the essential background to his theoretical breakthrough (Galison 2003). One of the challenges that engaged the younger Einstein was the spatial and temporal complexity of Europe’s burgeoning rail network. Only a century previously, the continent had had hundreds of time zones, and no universal system for synchronizing them. Given that local time could therefore vary from town to town, scheduling rail services was hard but also vital – not least to stop trains from colliding. Synchronized clocks, set by telegraph, were vital to determine longitude and therefore provide precise coordinates.

In other words, the history of ideas about space and practical spatial concerns can rarely be divorced. Again, as this example of spatial measurement and timetabling shows only too well, geographers as varied in their approach as Torsten Hägerstrand and Doreen Massey have come to understand that space and time cannot be easily separated out. Space always has temporal extension and vice versa: neither dimension is privileged over the other since both depend on each other for their expression. In turn, four consequences have unfolded.

To begin with, we have come to realize that there is not really such a thing as “empty” space. Quite the reverse: space is chock full of stuff. We can think of this fullness in several ways. One is as all the things that are there but of which we have as yet no conception, just momentary glimpses, what Bruno Latour calls “dark matter.” Another is as a gathering of matter like the glial cells in the brain. These were first thought of as being a little like wallpaper – as a necessary but generally unimportant background to neurons – but now we know that glial cells are themselves important elements of how the brain functions. Then, we can think of this fullness as all manner of geographies we can sense but cannot yet fully grasp: for example, according to quantum theory, in the same moment, the same atomic particle can be in two places far apart. Finally, we can think of that fullness as being like the light that arises from distant galaxies, echoes of events long past which ghost everything that we do.

Then, we understand to a much greater degree than formerly large expanses of time and space which in previous eras would have been much more difficult to conceive of. Think, for example, of our ability to peer back into the early history of the universe via telescopes like the Hubble or the Square Kilometer Array. But it is not just large expanses of time and space that have become visible. We have also become able to see and operate in small moments in time, with the result that new spaces have opened up in this realm too. Many examples of this adage can be found. For example, international finance often operates in the realm of seconds and even microseconds. But, equally, we can track the way in which information and trends make their way around the world in seconds, as in the periodic global Twitter storms that can surround the doings of celebrities – storms that they have often whipped up – or events such as the outcome of soccer matches. In other words, human spatiality has both stepped up and stepped down a level.

Then again, the link between time and space has become more visible through movement. As transport and communications technology improve so new perceptions of movement become possible. In the latest iteration of this process of boosted journeying, it becomes possible to see each step in the chain of movement of people and things and to adjust that movement whilst in movement: this logistical thinking has now become commonplace. But smartphones and tablets, global positioning,
SPACE

instant geographical search, and the panoply of other technological advances we now regard as indispensable, have only existed for a very short time within the historical record. Only 30 years ago, people still set out in cities like London with an A to Z gazetteer or map, a street finder first produced in the 1930s which became indispensable for a series of generations, just like the Thomas Street Guides that were so familiar in Los Angeles and other cities of the United States. Even armed with a street finder, drivers and walkers still navigated with considerable degrees of uncertainty, often roaming up and down streets searching for the right turn-off and then the right house number – and they couldn’t call anyone to find out if they were on track until the widespread adoption of handheld mobile phones (invented in 1973) in the 1990s. Equally, drivers couldn’t use anything like satellite navigation in their cars until the widespread adoption of these devices in the late 1990s and early 2000s, and walkers couldn’t use simple smartphone apps like Google Maps until the mid-2000s.

There are still relics of the old way of proceeding, even in the most technologically enabled cities. For example, in London, the black cab or taxi (from taximeter, the machine invented in 1891 that measured the distance and time to make a journey) drivers have to take “the Knowledge,” a test, first introduced in 1851, of their ability to find their way around the city without anything like a global positioning system (GPS). Not surprisingly, it usually takes two to four years to pass muster. To become an all-London taxi driver – a Green Badge holder – you need to master no fewer than 320 basic routes, all of the 25,000 streets that are scattered within the basic routes, and approximately 20,000 landmarks and places of public interest that are located within a 6-mile radius of Charing Cross. This is where the famous Blue Book is required, a handbook for budding taxi drivers which contains all of these routes and landmarks. Now there is pressure from GPS systems and apps like Uber and Hailo to automate this process of learning geography. In other words, London taxi drivers could go the way of other craft workers. Or maybe not: London’s traffic geography changes so often that ideal routes can switch in the blink of an eye and automated systems still find that recalibration hard to achieve.

Finally, the question of where we are in space and time has also expanded as aerial views have become a norm. Being able to see over a space is not a natural right: it had to be constructed. For a long time, that meant high vistas, towers, and the like. Even when the idea of a bird’s-eye view became engrained, it was not always easy to portray particular features such as the Grand Canyon. Then, along came balloons. A number of geographers have considered the early history of ballooning, precisely because it was the first time that aerial views became commonplace: as early as the French revolutionary era, for example, a group of “aérostiers” was established whose task it was to sketch the contours of a battlefield and communicate them to officers on the ground (Adey, Whitehead, and Williams 2013; McCormack 2014). Subsequently, the spread of aviation and the aerial photograph cemented the view from above – the view from falling up instead of down – as a normal way of seeing space. Now people all around the world think of the view from an aircraft window as somehow natural and rarely remember the assorted cloudscapes that they traverse. One of the most privileged moments the author can remember was flying into London on November 5 at a time when, by ancient custom, bonfires are lit and fireworks are set off. Another was flying past London when the cloud base was very low and seeing the taller buildings in London sticking up out of the cloud like jagged teeth.
But, at the same time as revealing this other often joyous way of seeing, it is likely that the view from above also prompted a greater emphasis on abstract spatial form, on diagrams of the world to which the world should conform, which has been responsible for some of the worst excesses of “top-down” state planning (Haffner 2013).

But, in the 1960s, another view from above became possible as photographs of the globe taken from space began to circulate, and specifically the Apollo 11 photographs of 1968 of a “spaceship” Earth hanging in the sky as taken from the moon (Cosgrove 2003). This view undoubtedly helped to stimulate a rise in ecological thinking about the fragility (and beauty) of our habitat. But now yet another view of the planet from above has become the norm: urban networks depicted at night as a weave of artificial light. Significantly, this latter-day view is constructed more as an assertion of a kind of erstwhile human supremacy: these networks of light are our new pyramids supposedly marking our supremacy across all of the Earth.

Knowing space

However precise and sophisticated our tools for mapping may be and however lofty the vantage point we are able to build, surveying where we are isn’t the same as knowing all there is to know about space, although you might think so from a lot of what you read. When we live in and encounter space, we are doing more than tracking its contours. It is what we are. It is the fiber of our being. It is both the folly and the grandeur of our nature. We are internal to it. All animals, including human beings, have particular dispositions towards space that are a part of what their bodies are. A nineteenth-century ecologist, Jakob von Uexküll, described the way in which animals live in different but intersecting life-worlds, a series of “Umwelten.” So a bee might only fix on just a few cues which go to make up its visual field – bright flowers, the sun, and so on – which are registered in relatively fixed corridors. A snowy owl has a quite different spatial footprint and correspondingly different spatial cues – wintering snowy owls spend 98% of their time sitting still during the day and probably do most of their hunting after dark using their extraordinarily acute sight and hearing; and they tend to roam extensively when they hunt (Strycker 2014). A human being might fix on many more but often markedly different spatial cues, not all of which are as discriminating as those of the bee or the owl. In other words, living beings see very different things or they see the same things very differently.

Human beings aren’t so different from animals, then, in that their spaces of existence reveal things selectively. At one level, our spatial disposition follows from the immense debt we owe to sensation in how we move and thereby think: we rely on all manner of physical forms of intelligence to make judgments, most of which exist at the interface between our bodies and the animated spaces within which we live. We touch things that manifest different temperatures and textures; we smell different odors; we see all kinds of colors; we lift or push against objects and so sense their weight and force, and so on. In turn, these simple sensations produce all kinds of unconscious effects up to and including moral judgments (Lobel 2014). At another level of spatial disposition, philosophers such as Heidegger concentrated their attention on how the world shows up — makes itself present — to us. Heidegger argued that we ordinarily encounter entities like tools as being for certain kinds of practices (hammering, cooking, writing, and so on). He asserted that we achieve our relationship with the world around us not through looking at
an entity such as the hammer, or through some
detached theoretical study of it, but rather by
becoming involved with and skillfully manipu-
latin it. In turn, this engaged, “ready-to-hand”
way that the world shows up to us means that we
do not have a straightforward spatial relation to
it of the kind that might be found using a GPS
device. We “dwell” in it in the same sense that
to dwell in a house is not merely to be located
inside it but to exist, or to belong, in a familiar
place. At one other level, spatial disposition
follows from the “atmosphere” with which
we surround ourselves, a bubble of attentive
involvement – of “inhabited orbs” (Sloterdijk
2013, 11) – in which life, thinking, and the
formation of these “spheres” amount to the
same thing. Living means building spheres, and
these spheres, made up of a scaffolding of bodies,
things, and signs, form an innate geometry of
knowing where and therefore what you are
which is both internal and external. In other
words, “we are in an outside that carries inner
worlds” (Sloterdijk 2013, 27): the scratches and
 scrapes of experience which make up what we
regard as our inner life are made by the way we
construct that outside.

The experience of space is a quality that we
all think we know well but it is one that remains
very difficult to describe. It is no surprise that
geographers have spent so much time trying to
do precisely that, usually by linking it to the
body and the web of practices within which
the body exists. This kind of space can never
be reduced to an aesthetic quality or simple
utility, though it might have such dimensions.
It is not so much observed as felt in the interac-
tion of body and landscape. Practices of living,
inhabiting, and thinking all have to take place
 somewhere, whether that somewhere is close
and snug or abstract and exposed, in a relation of
connectedness, and the notion of place provides
an orientation to thinking precisely what that
connectedness might be. Philosophers such as
Heidegger and Sloterdijk have often argued that
the problem of modern life is precisely that place
 is being reduced to an aesthetic or a utilitarian
dimension, so blocking out a certain kind of
freedom to be in place and of place. Whether
this sense of loss, apparently caused by the can-
nula of “modernity” draining color out of the
world, is a valid one or not, the fact remains
that this feeling has been a staple of Western
cultures, echoing down from the time when
the opposition between “culture” and “nature”
became established and nature somehow seemed
to move offshore. The sensuous character of life
was displaced.

What is clear is that the bodily experience
of space – the different sensations it engenders,
the different complexities it lends to life, the
different ways that it folds into how life and
its practices show up, its different character, if
you like – has varied through time. (Although I
suspect that some experiences are held in com-
mon, such as the aching void left in space when
someone close to us dies.) Let’s take just one
obvious example: different bodily experiences
of space in Ancient Greece and Rome. It is
a particularly appropriate example given that
Homer is often regarded as the founding father
of geography, while Greek geographical knowl-
dge is often thought to underlay the Roman
Empire, as evidenced by Strabo’s Geography. Just
as the Ancient Greeks and Romans had different
words for different colors than us, so they had
different means of apprehending space. So there
is no word in Ancient Greek that simply and
unproblematically maps onto what we now
think of as space, and there is a considerable
debate about how accurate Greek maps were.
Meanwhile, the Roman “spatium” meant stretch
or extent and again it is not equivalent to what
we think of as space. Equally, Greek and Roman
cities and sacred sites were run on different spatial principles, which provoked different bodily experiences of the world. Until recently, such insights have been difficult to articulate because of the tendency for archaeological excavation reports to be split by function, thus making it problematic to trace the linkages that produce that holistic sense of space. But now we have begun to see the range of spatial experience that was apparent in those times (Scott 2013); for example, towns and cities in the Roman Empire tended to use the template of Rome while Greek towns and cities were much more varied in style. The Greek agora was genuinely meant to be an open space (at least for certain classes of persons), whereas Roman public spaces were much more ordered in terms of which and what activities were allowed. Greek understanding of geography tended to be based on world maps, whereas the Roman understanding of geography was modeled on itineraries, and so on.

In other words, we need to be careful. As this brief excursion into the historical record shows, there is no reason to think that the world turns up the same for all people. The bodily experience of space varies on all kinds of dimensions, which are both physical and cultural, as numerous studies of gender and ethnicity have shown only too well. But such experiences can vary even more fundamentally. Take the case of those cultures that are often described as having an “animist” outlook – that is, cultures that ascribe nonhumans like animals and plants an interiority identical to their own. “We are a long way from the reassuring world of Being and existing beings, of primary and secondary qualities, of perennial archetypes and of knowledge as revelation. For all those weary of an overuniform world, that realization is surely a cause for a measure of rejoicing” (Descola 2013, 143). In animist cultures, animals and plants are “humans in disguise.” They are persons sometimes thought of as having their own cultures. Humans and nonhumans differ by virtue of the form of their bodies but the substance of their souls is counted as similar. So, for example, animals might be thought of as essentially human but they have been transformed into animal forms by means of the outward clothing of their skin. Or all beings are thought to be made from the same substances. The result is that human beings are joined by all manner of other humans – but humans who do not present a human appearance. The myths of animist people nearly always remember a time when humans and nonhumans were not differentiated, and shared “Umwelten.”

Then there is “totemism.” In this scheme of things, each person belongs to a group that bears the name of, and is linked to, a natural object such as an animal or a plant or an artefact, or even a natural phenomenon such as a river, that is associated with a set of moral, physical, and behavioral attributes, a living expression of certain qualities that transcend their own existence, rather than an individuality, with which it is possible to maintain a personal relationship (as in animism). Here we can see another notion of space entirely. The classical example is provided by the indigenous Australian “dreamtime,” a single word used to sum up many different kinds of experience. In this experience of experience, primeval beings emerged from the depths of the Earth from specific sites. Some embarked on journeys that left their marks on the landscape and behind them they strew many of the existing beings of today, all manner of humans, plants, and animals, each with its respective totemic affiliations. These primeval beings might be human in terms of their behavior, language, and social codes, but they also have the attributes of plants or animals and they are the origin of stocks of spirits laid down at the sites where they themselves disappeared. “Dreaming” isn’t just
ancient history, for when the beings departed they left a legacy of spirits, routes, and rituals that continue to affirm their presence in the present. The paths continue to thread their way through the landscape, rather like the maps of underground and subway systems with which we are now so familiar, and indigenous Australian peregrinations continually reaffirm them. They express a living relationship with the land, with “country,” understood as an animate entity: “The country may be mother or grandfather, which grows them up and is grown up by them. These kinship terms impose mutual responsibilities of caring and keeping upon the land and people.... For many indigenous Australians, person and place, or ‘country’ are virtually interchangeable” (Arthur, cited in Nicholls 2014, 4). Indeed, such cultural differences as these mean that different human beings may experience the basics of space quite differently. These differences, both biological and cultural in nature, can be quite extreme. Work in cognitive anthropology shows just how extreme. For example, the work of Steven Levinson (2003) on certain indigenous Australian and Mexican groups shows that, because they tend to use absolute rather than relative frames of reference, not only do they have very good senses of direction – they are able to point the right direction to a place without demur and past events are directionally anchored in great detail – but they genuinely perceive cardinals such as up and down and left and right in very different ways.

Western cultures often think of technology as the primary determinant of spatial perception. That is important, of course. When new technology of various kinds arrives, space can be perceived in ways that were not possible before. We have noted the case of maps and measurement, but there are many other examples of how technology changes spatial perception. Take the faculty of vision. The invention of spectacles is an obvious but often overlooked example. The invention of oil paints, crucial to the invention of impressionism, is another. A further example is provided by artificial lighting. Night time is no longer another country, dark and forbidding. Thanks to reliable and constant artificial illumination produced by large industrial power networks the space that we might inhabit, the “sphere” to use Sloterdijk’s term, has expanded its orbit. The process began in Britain, which became a gas-lit society when the first public street, Pall Mall in London, was lit with gas in 1807. Electric lighting arrived in 1878 but did not find its way into homes on a large scale until after World War I. Artificial light has now spread all over the world, producing new apprehensions of how and what can be seen (Otter 2008). But we should not overestimate this process of enlightenment: for many people in the world the dark is still dark. For example, in India in 2012, the largest electrical blackout in human history left 600 million people without power and therefore without light. But discussing the blackout, the economist Amartya Sen wryly noted that the media had neglected an important fact: “Two hundred million of those 600 million people never had any power at all.”

But our apprehension of space is not just about simple advances in technology. It also follows from accompanying shifts in how people perceive the world, each of which has a spatial dimension – what is considered to be inside or outside, what is considered to be public and private, what is considered to be full or empty, what is considered to be a whole or a part, what is considered to be general or particular, what is considered to be different or the same, what is considered to be near or far, what is considered to be small or large, what is considered to be center or periphery, what is considered to be where or nowhere, what is considered to be separate or juxtaposed, what is considered to
be still or moving. These shifts, which can be
domestic or global in scale, produce new spatial
orders, new placings, new spatial signposts. But
dimension is perhaps too weak a term. What is
meant is what is meant by life: the way things
are. These shifts can be very general cultural
currents, such as that European complex of arts,
sculpture, and building that became known as
the Renaissance. They can be artistic move-
ments such as Impressionism or Cubism. But
sometimes they stem from just a few small points
of change that become general.

A good example is provided by Pablo Picasso
(Clark 2013). Picasso pushed at the bounds of
space all through his career. He tried to identify
new ways of thinking about inside and outside,
which acknowledged the domestic interior that
had become a key feature of the inner and outer
of European bourgeois life and its representa-
tions in paintings, and increasingly challenged its
ascendancy. Thus, he tried to identify new organs
that linked bodies in different ways, both inside
and outside conventions of what counted as a
room. For example, he conjured up ways of mak-
ing sexuality surprising, unfamiliar – even fright-
ening – by draping bodies over and around spaces
so that intimate interiors could become loud
exteriors. In his later career, he also tried to ques-
tion what we mean by objects by constantly buf-
feting the common notions of proximity so as to
produce new “inhuman” mixtures. For example,
he disrupted the pure lines through which space
could flow unobstructed, the certainties of
geometry that allow us to place and contrast our-
selves and to adhere, and the pressure of surfaces
and boundaries produced by the containerized
nature of so much representation of room. In
doing so, Picasso operated on what might be
regarded as the “truth” of space, trying to paint
or sculpt what “really” existed. He provided
new kinds of reaching out, new kinds of limits,
new means of looking back, new animations
that have allowed us to produce a new visual
vocabulary of space: in other words, he produced
a new cultural “floor” to spatial perception.

The new geology: exploring the space
of the Anthropocene

Our technological and cultural perception of
locating and knowing space is changing, then,
as we lay down layer upon layer of infrastructure.
The world is threaded with roads and cables,
undergirded by pipes and tunnels, saturated by
wireless signals, loaded down with all kinds of
built infrastructure, and heated up by all manner
of energy sources. It is crisscrossed by airline
routes, lit up by the innumerable street and
other lights mentioned above, and shaken by all
manner of artificial sounds.

Geography was often thought to be an offshoot
of geology. But the question arises as to what
happens when human beings begin to change
the planet to such a degree that they start to
create their own geological epoch by building
one vast ants’ nest which extends both down into
the ground and up into the air. We might see
human beings as creating a new space, one with
its own atmosphere, its own dynamics – and
its own problems. This new era is creating a
distinct geology. When an alien explorer looks
back at the Earth from millions of years in the
future, she or he or it will be able to discern a
whole stratum based on the detritus of human
habitation, not just buildings but all manner of
objects and even clothes – which could well be
interpreted as external coverings periodically
discarded as they grow too tight to fit a growing
organism. Although human cities may not yet
be as impressive as biological cities, such as
coral reef, in their scale and longevity, or the
nesting sites of the now extinct passenger pigeon
mentioned earlier, each host of pigeons weighing
down the trees in their millions and covering a space as large as Manhattan (Strycker 2014), still humanity is catching up fast (Zalasiewicz 2009). Furthermore, human cities are giving rise to exactly the same spatial promiscuity with regard to minerals and metals as has already been noted with flora and fauna.

The geological record will also show other marks and, most particularly, climate change and mass extinctions. Humans have undoubtedly begun to change the chemical composition of the atmosphere. At the same time, humans are producing a mass extinction, on the same scale as the catastrophic extinctions caused by asteroid impacts, as they wipe out species directly or simply leave them marooned and so condemned to fade away. “It is estimated that one-third of all reef-building corals, a third of all freshwater mollusks, a third of sharks and rays, a quarter of all mammals, a fifth of all reptiles, and a sixth of all birds are headed toward oblivion” (Kolbert 2014, 17).

Then how we experience space begins to be mediated to the second or third degree by infrastructure. Take the case of sound. In the past, the soundscapes of cities left room for sounds emanating from natural sources. Church bells might produce a periodic cacophony, the noise of horse-drawn traffic might create a low roar at certain times of the day, and various shouts and bursts of music interspersed the daily round. But there were periods of quiet and many acoustic spaces in which sound still consisted of the wind and birdsong (Glennie and Thrift 2009). Now, things are very different. Places where the only sound is natural sound are few and far between— a British professor of acoustic engineering tried to map them. In the United Kingdom, you have to go far into a remote Northumbrian bog to find the quietest place. Cities are chock full of noises that form their own associative landscapes intensifying and fading as we move around—from the sound of aircraft overhead through the hum and honking of traffic to the snatches of music and conversation, each with its own arc (Gandy and Nilsen 2014). Meanwhile, the world has become so beset by noise that great tits are forced to sing faster and higher in the urban din, and robins have taken to singing at night when it is quieter. Shipping noise disturbs the breeding of whales and dolphins: whales are having to sing louder and louder—and in some cases have given up entirely (Cox 2013). Equally, there are radio quiet zones around a few large radio telescopes to prevent interference from radio waves. The main reason why we can’t get away from all this noise is infrastructure.

Infrastructure consists of all of those objects that allow human beings, cars and trucks, water, electricity, radio signals, information, and the like to flow from one place to another, to become mobile, to circulate. Mainly they consist of continuous conduits of one form or another but, increasingly, as wireless has become more common, these conduits have broadened out into signals transmitted from and received by masts, although the principle remains much the same.

The point is that the Earth’s surface has become an anthropic stratum hundreds of kilometers high and at least 3 km low that is a bit like a Swiss cheese in its makeup, a stratum through which pipes and cables crawl, under which tunnels and boreholes and mineshafts bore down into the Earth, on which all kinds of reservoirs and power sources hold sway, and over which airplanes and satellites and wireless signals fly back and forth.

So generic is this infrastructure that we take it for granted, at least until it goes wrong; think of the burst water main or the power cable brought down by an ice storm. Equally, we forget that it has its own history. We have already touched
upon systems of lighting with their pipes and cables. But take another example, this time from transport infrastructure: the humble elevator upon which tall buildings rely for sustenance. Although records of elevator design date back to Roman times, elevators (or “ascending rooms”) were first commonly used in the nineteenth century when hydraulic, steam, and electric power were all used as sources of lift. In 1852, Elisha Otis introduced the first safety elevator, which prevented the cab from falling if the cable suspending it broke. The first such passenger elevator was installed in 1857 in a building that still stands in the SoHo neighborhood of Manhattan. The first office building to have passenger elevators, the Equitable Life Building, was completed, again in New York City, in 1870, and prefigured the growth of the late-nineteenth-century commercial “skyscraper” (a term first used then) and the consequent transformation of the skyline of many cities. The adoption of elevators had many consequences, apart from creating recognizable “downtowns.” For example, before the use of elevators, most residential buildings were limited in height to about seven stories. The wealthy lived on the lower floors, while the poorer residents lived on the higher floors, and were thus required to climb many flights of stairs. As exemplified by the modern penthouse, the elevator reversed this stratification (Bernard 2014). Now, as buildings have become ever taller (with a tower 1 km high planned in Saudi Arabia), so elevators have become increasingly complex. Of course, they need to be faster (with speeds of up to 18 m/s) but they also need to include sky lobbies where passengers can change lifts to reach higher floors, and means of increasing capacity such as double-decker lifts or multiple cabins in the same lift shaft.

So ubiquitous is infrastructure such as elevators – New York has at least 64,000 according to the city’s buildings department – that this infrastructure has become an accepted feature of the landscape: airports, railway stations, service stations – and elevators – have all become stock scenes of human conduct, places where we expect certain classes of drama – meetings and farewells, temporary sojourns where it is possible to eat or refuel, places to gather and disperse, uneasy encounters and serendipitous meetings – to play out against a background that is made as predictable as it can be, from the identical architecture to the uniforms. A French anthropologist, Marc Augé (2009), once called them “non-places” but as a description that is very far off the mark. Rather, they are “by-places,” which depend on the journeys of people and things for and as their existence.

Transport infrastructure has all kinds of other effects. Most particularly, in allowing people to move, it allows all kinds of communities to spread out across the world. All of the diasporas that have arisen from population movement are striking in their extent. In large part, the history of countries such as Australia and the United States is the history of wave upon wave of diasporic communities. Migration has been a constant condition of human history but it has picked up in pace and volume since the nineteenth century as movement has become easier and less hazardous. All kinds of discontinuous diasporic spaces have been created, many based on simple economic incentive but some the result of concerted state and state-related action such as ethnic cleansing (Osterhammel 2014). But movement is also born out of other imperatives – such as tourism, religious observance (consider the hajj), and all manner of other motives.

Governing space

In most societies around the world we have come to believe that space can be governed or owned.
We have come to believe that space should bear our indelible mark, the mark of “territory” (Elden 2013). We do not belong to places, as the indigenous Australian thinks. No, they belong to us. We contain them rather than they contain us. The question arises as to why and how this container style of thinking, with all of its physical accoutrements, from fortresses to barbed wire, from passports to border posts, from fences to electronic surveillance, has come about, and how we have arrived at a world in which spaces are able to be stabilized sufficiently that when we draw lines around them, the lines not only stick but become important icons in their own right.

The answer is multiple, but one undoubted answer is through the gradual growth of the modern state. We should not exaggerate the powers of the state, of course, whether they consist of organized violence or bureaucratic practices or legal codes or shared norms or ideologies. Indeed, it could be argued that in many countries states didn’t do as much as has often been argued until relatively recently in the historical record and in large parts of the world they still don’t. Other structures, such as organized religion, have been powerful and have continued to exert heft. Part of the reason for this exaggeration of states’ powers (for example, the idea that states create one type of agent) comes from the tendency to portray states as unitary authorities bringing one space under the sway of one will, but that is an illusion. The production of state action is distributed among many authors and material artefacts in acts of collective governance that are always spread across many different types of space, each responding differently even as they add up to something greater than the sum of the parts. Even when the state is engaged in the greatest of great projects, which look as if they are all central diktat and have no local room for maneuver, whether they be the Pyramids or the Three Gorges Dam, the expertise needed to produce the project is never to be found only at the “center,” whether that center is a temple or an anonymous set of offices. It depends greatly upon the skills of the artisans and workers who are there on site. Indeed, without this initiative from below, not much will happen, and the conjuring act is to allow input from below while ensuring that power and direction remain at the upper levels. Much of the bureaucratic weight of empires and other state forms relies on getting this balancing act right through devices such as addresses, which exactly locate a destination, files and paperwork more generally, manuals, offices (with their double meaning) and waiting rooms, examinations, general notions of conduct and habit, and even styles of communication (from the flowery language of the courtier to the “plain” prose style of the civil servant), each of which shapes people and space to the state’s will by manipulating the efforts needed to access and exert power. Connecting and governing are two sides of the same coin (Joyce 2013).

Equally, that weight depends upon producing a space in which the state seems to be a constant companion. Ordering and governing space depends upon naming places, localities, topographical features, regions, and larger areas too, such as various kinds of world region. It depends upon delineating borders. It depends upon producing uniform and continuous spaces through the use of standardized units of measurement. So the whole profession of geography as it appeared in the nineteenth century, with its own research methods, terminology, career paths, academic departments, textbooks, and journals, depended upon a practice of exploration that was in part precisely concerned with what could often be the simultaneous location and naming of places, a kind of “worlding,” establishing new borders or asserting old ones and producing quotidian spaces of reckoning. In Britain, for example, what had been the practice of military
mapping of Scotland and then England gradually turned into a more general national mapping agency, the Ordnance Survey, founded in 1791 although called by this name in print only from 1810 (Hewitt 2010). In turn, its cartographic practices moved out into parts of the British Empire such as India. But, ironically perhaps, practices that were conceived with military means in mind were subsequently appropriated for civilian purposes, not least tourism and hiking, which are associated with freedom of movement rather than binding state grids. A similar process happened in countries such as the United States where state mapping began as a means of inventorying the rapid expansion of the nation, usually via a characteristic grid form, but became something less utilitarian as time went by.

All of this state activity had to be held together by practices of connection and, just as importantly, nonconnection such as secrecy: secrecy is a key feature of state power. States need to keep things out of sight (Galison 2003; Paglen 2010), and that means walling off spaces from public view or only allowing unequal access. This practice may be at odds with modern views of citizenship, which rest upon the idea of greater mutual visibility between state and citizens, but the fact remains that modern states keep a good deal secret, just as the states of yore did with their invisible ink and strings of secret intelligencers, tightly guarded documents, and the pouches of diplomatic couriers. The intricate patchwork of secret spaces expanded mightily in World War II and now includes defense facilities, many government buildings, and bunkers. It also includes unequal knowledge of individual citizens of the sort that is promoted by the apparatuses of files, constant electronic surveillance, and finally, the actual removal of knowledge from public view as promoted by the practice of classification. The net result can sometimes seem baroque: currently, around 4 million people hold some measure of clearance to view classified documents in the United States and 33 million documents were newly classified in 2001 – a figure which has almost certainly increased since then (Galison 2003).

Of course, different states hold such geographies of connection and nonconnection together differently. To begin with, just look at the evidence of maps. They often loudly declare not just that this is “my space” but also that a particular space is central – and other spaces are not. Over a long period of time, imperial era Chinese maps have been the perfect example of this statement (Hostetler 2001; Smith 2013). Imperial era Chinese maps tended to look and feel quite different from Western maps in that they were not gridded and devoted much more space to written text, given that in China written words counted above visual images in terms of the weight given to their representational authority. In these maps, China is always the pivot, and other lands are gathered around at a much smaller scale as a fringe of tributary relationships. When the Jesuits produced maps of China and the world from their perspective, they were attacked for dislodging the “Central Kingdom” from its rightful position at the center of things. Equally, the incentive to seek out knowledge of such foreign spaces seems to have been marginal since China was so clearly the important space. Thus, “Chinese mapmakers understood the utility and appeal of accurate measurement, and their colleagues in astronomy developed sophisticated instruments that made possible the projections and coordinate systems that Westerners associate with Ptolemaic cartography. But throughout most of the imperial era, they found no compelling reason to conceive of the world as spherical, nor did they see any merit in drawing all maps ‘to scale’” (Smith 2013, 87). China was where it was at.
This example may be an extreme case of the state’s spatial narcissism but it is hardly unusual. Wars, legal cases, and boundary disputes are all fought as though only some can have rights to a piece of space and others should be excluded—or even deleted from existence so that they can never make claims on the state again. So it is that between 1776 and the present the United States could seize about 1.5 billion acres from North America’s indigenous peoples, an area 25 times the size of the United Kingdom, with a consequential decline in the population of indigenous peoples. It is worth remembering that as late as 1750 these peoples had still constituted a majority of the North American population.

But it is important to understand that we live at what is probably the high water mark of a particular kind of state, the nation-state, a spatial form that can be dated to the Treaty of Westphalia of 1648, which both formalized and then normalized this particular kind of political geography, although it did not set it in train, as some would have it. But today political geography is driven by all manner of exceptions to the idea of a set of contiguous nation states legitimated by legal sovereignty over “their” space and by constitutions that abjure any other way of thinking, and which are able to claim the right to indulge in armed conflict with one another as a result. Indeed, this form is, if anything, becoming defined by the host of exceptions to it: all of the other institutional bodies whose spaces do not coincide. These spaces are legion and include those of the numerous international and subnational bodies, global cities such as London, New York, and Shanghai, various ecological enclaves, diasporic communities, global brands, special economic zones, massive computational networks that share files with abandon, logistical networks, the footprints of satellite surveillance, criminal networks and clans, financial markets with their server farms and special offshore domains, and goodness knows what else. This tendency has only been accelerated by the multiple networks of information technology, which allow all kinds of new addresses to come into being: land, then sea, then air, and now—cloud. The difference is that we have built this composite space of address and its forms of participation, via server farms and data centers, millions of miles of cable, the wireless frequency spectrum, and, by 2020, an estimated 1.9 terawatts of power (placing the cloud, if it were a nation-state, as the fifth largest consumer of electricity in the world). With it we have formed a world in which not only can we communicate but so can things, and not just with us but with each other, so heralding what might become a post-Anthropocene (Bratton 2014).

Perhaps we shouldn’t be too surprised by the increasing diversity of institutional spaces. Before the seventeenth century, millions of acres weren’t strictly owned by any one person or institution. Rather, they were subject to a patchwork of rights. Through history, people have held “rights of common” to carry out activities (to pasture animals, to fish, to take wood for fuel, to take sand and gravel, and so on) in and on particular pieces of space. Originally, these spaces were appurtenances to other privately held space but, over time, they have come to be understood as “public,” a word which has its own geographical history. Thus, the idea of a common was exported by the Puritans to North America: famous examples include New Haven Green in New Haven, Connecticut (the Puritans were said to have designed the green so that it was a large enough space to hold the number of people who they believed would be spared after the Second Coming of Christ: exactly 144,000), Cambridge Common, and Boston Common, all of which were converted into parks.

But the idea of ownership is not just the preserve of the state, as the examples of many spaces
held in common make clear. It also comes from the notion of an economy, in which the practices of exchange have become the be-all and end-all. Like the state, the economy, especially the global capitalist economy, isn’t a natural construction. It has arisen over many centuries from interactions that have become increasingly regularized by a network of practices, specialist laws, and regulations, by brute enforcement by empires, by colonial adventures, and now, by large multinational corporations that can span the world and create their own owned spaces, both in and out of any one state jurisdiction.

Trading space

If we turn to the economy, then, there is another way in which space has become objectified – as a traded commodity available for exchange, as “land” that can be “owned” and then “valued.” It is an obsession that is deeply rooted in history – some of the first known maps are probably records of land ownership – and has produced for millennia an uneasy equilibrium between those who believe in common land and those who believe in privately owned land. At least part of that debate comes from what seems to be a very human need to close off access, so as to make space “mine” and give it secure title (Linklater 2014). Notice the quotation marks around the words “land” and “owned” though: in these days people very often assume that one must mean the other, but there are still many instances of land that is owned in common or in trust and therefore cannot be trespassed upon. Thus, it has been estimated that nearly 15% of the Earth’s surface has been preserved for conservation in areas such as National Parks. In some countries – such as Scotland, for example – there are universal access rights to most land and inland water (although there are exceptions) and there are still common grazing rights in a number of areas of England such as the Lake District. But it is also clear that the majority of land and what is distributed over it is privately owned in some way and has an explicit economic value, especially if it is thought that the Earth has now effectively been domesticated for human use by the economy, and that “wild” nature is no more and can be replaced by human-designed ecosystems. The planet’s main purpose is to be friendly to humans by accumulating wealth through trade in commodities, including land, so that humans can be, well, more wealthy.

Everyone talks rather glibly about this process of trade and commodification as a “world economy,” but such an entity exists as a definite object only in the imagination and in theoretical texts: what actually exists are numerous overlapping spatial networks of people and things set up to allow predictable exchanges to take place on a continuous basis (Thrift et al. 2014), exchanges of consumer goods, money, and remittances, and pleasurable tourist experiences certainly, but also of slaves and trafficked women, international criminals and drug runners, and all manner of pollution.

These networks are vast historical accretions. Take the case of the seas. Nowadays trade has become a maritime infrastructure in its own right, an infrastructure dependent upon a forest of rules and regulations that range from the international law of the sea to the humblest health and safety inspections and regulations. Then there are 100 000 freighters on the seas at any one time carrying nearly everything we eat, wear, and work with, crewed by an itinerant population numbering in the hundreds of thousands (George 2013). Then there are the ports, full of specialized equipment ready to load and unload the containers, which are the basic grammar of maritime trade (Levinson 2006). Then, of course, there
is all the trade carried out by road, rail, and air, which cannot be mentioned in any detail here.

But this over-the-seas world of trade has not just magicked itself into existence. It has come about as a result of definite historical processes. To govern spaces such as the sea means being able to define, measure, recognize, and enforce boundaries and thus control movement by means such as cartographic, population, and navigational data, understood as tools and not just as a representation of power. To understand this trading space, we can return to the map – after all, the power of the map is its ability to qualify movement – and specifically to the famous Selden map of China, bequeathed to the Bodleian Library in Oxford in 1654 by an English lawyer, John Selden (Brook 2013; Batchelor 2014). For etched into the history of this unique 16 2/3-square-feet map of China we can also find the classic signifiers of bounded territory – the nation-state, the rule of law, and the corporation – and the consequent enhanced ability to control movement. At the time Selden owned the map, the British state was experimenting with new models of bureaucracy that would bind it closer together as a coherent space. The map was in the ownership of Selden, one of the pioneers of the project, to allow nations to own the sea through a dense skein of maritime law, an autonomous legal and mercantile framework that would allow maritime trade to flourish unencumbered, a project that had become vital as “ships in their tens of thousands were sailing from every port of Europe and Asia” (Brook 2013, xxii). Seas could be closed off – the title of one of Selden’s works was The Closed Sea – and taken in to a coherent space that could support the growing global ambitions of “empires,” but with the monarchy becoming the symbol of coherent state authority and not just a political settlement. Finally, there was the joint-stock company, the progenitor of which, significantly, was called the Cathay Company, founded by Sebastian Cabot in London (Cabot was the son of John Cabot, the Italian navigator and explorer whose discovery in 1497 of parts of North America under the commission of Henry VII of England is commonly held to have been the first European encounter with the mainland of North America since the Norse visits in the eleventh century). Significantly, Cabot had also partnered with Clement Adams to publish London’s first printed world map. The joint-stock company was not just a tool that allowed the risks of long-distance maritime trade to be taken on board, but also a means for producing fiscal and commercial autonomy from the burgeoning state and so establishing the mixed authority between state and economy that we now take so much for granted. Some of these ideas quite clearly came from an awareness of the sovereign boundary definition going on in Ming China, with its dependence on notions of counting population and on technical achievements such as the compass.

In turn, trade has become the means by which we understand the world. It is a part of how we are. That has been the case for a long time, of course. For example, Europe has been a trading zone since at least 9000 BCE. But instead of amber, soapstone, and shells brought overland or by sea and made into necklaces, bangles, and the like (Cunliffe 2011), our very bodies are now maps of global trade. Shirts, tee-shirts, and blouses from Thailand and China, coats from Malaysia, glasses and contact lenses from Germany, watches from Switzerland, shoes from the United States: each of us is a lesson in exchange, a world in miniature. We live in places chock full with stuff that has transited from other places and still contains those places’ echoes. This fact is underlined not just by the constantly changing retail landscape with its standard grammar of shops and malls and warehouses, but by much less imposing but
equally crucial intermediary locations such as Chungking Mansions in Hong Kong, where cheap goods are sold (Mathews 2011). That is before we get to the afterlife of all these things as they pass to other owners as second-hand goods, accumulate in garages and attics, or end up being recycled or cast aside in garbage dumps where they recreate value as waste – and create records for future archaeologists.

Indeed, this roundelay of trade has produced what some have called a junkyard planet (Minter 2013). Much but by no means all of what we produce goes into landfill. Cans, computers, and cars rarely end up in landfills. Indeed, automobiles are nearly 100% recycled. No surprise then that the global recycling industry has a turnover of $500 billion a year, roughly equal to the gross domestic product of Norway. It is an industry which forms a trading complex in its own right since it consists in large part of the complex movement and exchange of goods such as paper and cardboard, iron and steel, aluminum and copper, from countries such as the United States over the sea to countries such as China – reflecting exactly an imbalance in trade. The United States consumes and throws away much more than it manufactures. China, meanwhile, is both the largest exporter of new goods to the United States and the largest importer of US recyclables. In turn, all kinds of places are linked together.

They are linked in other ways too. For example, new goods are crucial but getting them or inventing them is no easy matter. Let’s return to the topic of secrets. When a new product or service is invented it immediately becomes a matter of how easy it is to steal it away from its originator in a complex geographical game of push-me-pull-you. For example, nowadays great attention is paid to Chinese economic espionage around the world, but it is worth remembering how active the United States was in illicitly appropriating mechanical and scientific innovations from the British in the nineteenth century, often with the encouragement of the US government (Ben-Atar 2006). Keeping ideas at home was a British national priority: Britain had strict laws against the export of machines, and skilled workers were banned from emigrating, so American industrial spies were lionized. The British were hardly blameless either. In 1848–1849, in one of the most successful acts of industrial espionage ever carried out, the East India Company, tiring of the Chinese monopoly on tea, which meant that one out of every five chests of tea manufactured in China was being purchased by Britain, sent the botanist Robert Fortune to China where, heavily disguised, he stole both the technique for processing tea leaves and a large collection of tea plants that were subsequently conveyed to India to be grown, thus breaking the Chinese stranglehold on the global tea market forever (Rose 2002).

But trade isn’t only the to and fro of what we regard as “material” goods. There is also trade in language and more generally in meanings – in signs that work across a series of different sensory registers (such as vision, sound, and movement) and different means of registering them (such as books or recordings or films) which are constantly mobile and changing – and although they may exist within predefined systems of meaning, such as textual or filmic conventions, they tend to overflow them. Take the case of language: we move about constantly so language is always in movement, constantly mutating: the “most fundamental fact about human communication is its variability” (Levinson 2003, 318). It can be moved about through the myriad interactions of “ordinary” conversation, with its encyclopedia of verbal and nonverbal cues, including particular ways of organizing and distributing spatial information, but equally it is moved about through all the means of expression that are currently
available: music, dance, all manner of texts, movies, and television, and latterly the Internet, with its new variants of communication. In turn, when signs are combined with bodies and emotions they produce deep and fast-flowing cultural currents that we still struggle to describe, even as we feel them as the compasses that set our behavior. These cultural currents wax and wane and join with other currents in all kinds of combinations. Take just the case of music (Byrne 2012). It’s a set of cultural baggage etched in sound — tempos, melodies, means of collaboration (from rock groups to orchestras), different kinds of instruments and contexts (such as bars and clubs, opera and concert halls, arenas, or outdoors), different ways of amplifying sounds (from cathedrals to microphones), and emotional reactions — which has traveled all around the world, just like any other traded goods, through a mixture of passion and economic calculation: geography and music are now inseparable. Part of the reason for music’s power as a set of signs is that it is closely allied to the makeup of particular spaces but can now also transcend them thanks to technologies of recording — the first sound recording was made in 1878 — which have themselves acted as means of producing new kinds of music while they have passed music around and around and made it possible to intersperse musical genres to an even greater extent through techniques such as sampling and the influence of all manner of new digital formats.

To conclude: cityspace

This is where cities come in, as the chief conductors of all the spatial transactions that define the human planet, from the to and fro of state intelligence through the vagaries of economic exchange to simply acting as founts of creativity. Rather late in the day to write about cities, you might think, in any account of space, but, then again, maybe not. Cities have always been seething multitudes, billions of interactions pulling together states and economies through great networks of influence and exchange. But now cities are growing in size and number to the point where the planetary condition is becoming the urban condition. Cities are becoming states and economies in their own right as a growing share of the world’s activity flows across national borders, both cementing and challenging a container view of space. That said, this pervasive urban condition is much too complicated to be described by concepts such as “planetary urbanization” — the variation in what cities are like and why they exist is too great for such simple-minded cut-throughs to say much at all — but what cities allow is for senses of connection to be both strengthened and subverted. Cities are not merely quantitative amplifiers but also, because of their great complexity, they are generators of qualitative change. They both hold things together and allow things to diverge: everything can become a stakeholder and so everything is put in relation and at stake (Bratton 2014). Cities are where questions and questioning chiefly emerge, and there are plenty of questions to be asked, many of them precisely concerned with how cities will survive in an era of climate change, population growth, energy deficits, lack of fresh water and clean air, and sometimes grotesque inequality: after all, so far, the Anthropocene has been a “filthy undertaking” (Vince 2014, 3).

At their worst, cities can be murderous entities casting people carelessly aside. But at their best, they will allow us to die another day by finding the answers we need to the questions we cannot avoid. The time will come when we realize that just having and owning means that we won't be heir to much at all. The planet will take it all back. Let's hope we can avoid such a bitter inheritance.
But to avoid that inheritance means that we must find new ways of sensing and representing space, ways that recognize that to love something, you do not have to govern it or own it or trade it. We can’t have it all. That’s a fact. It’s not just “our” planet. Mutuality needs to replace mastery. We need to become responsible earthlings before the Anthropocene becomes simply an obscene blot on the geological record (Ellsworth and Kruse 2013). Geography, the guardian of difference, can map out the long and no doubt painful road to what could be a shimmering redemption.

SEE ALSO: Economic geography; Map projections and coordinate systems; Place; State, the; Territory and territoriality

References


Spain: Asociación de Geógrafos Españoles (AGE) (Association of Spanish Geographers)

Founded: 1975
Location of headquarters: Madrid
Website: www.age-geografia.es
Membership: 950 (as of 2013)
President: Carmen Delgado
Contact: carmen.delgado@unican.es

Description and purpose

The Asociación de Geógrafos Españoles (AGE) is a group of geography teachers, researchers, and professionals whose main aim is to promote and advance geographical science and its applications in Spain. In addition, it aims to promote wider public knowledge and a greater understanding of geography. Since its foundation it has been mainly involved in promoting scientific meetings, coordinating the activities of Spanish geographers, and collaborating with other national and international associations. Given the wide variety of topics related to geography and ever greater thematic specialization, the association has set up fourteen working groups.

Current activities or projects

The association hosts the Congresos de Geógrafos Españoles (Congress of Spanish Geographers), a biennial congress, the most recent being held on October 23, 2013, in Palma de Mallorca. AGE also organizes annual meetings for the fourteen working groups as well as annual seminars on “Geography Teaching in Secondary Education,” the last one was held on July 6, 2013, in Getafe (Madrid).


Brief history

In the mid-1970s an emerging need for organizing scientific activities and to solve common matters encouraged academic geographers to found the Asociación de Geógrafos Españoles (AGE) in 1975 during the IVth National Geography Conference. The first president of the association was Professor Jesús García Fernández.

AGE has maintained relations with the Associação Portuguesa de Geógrafos (Association of Portuguese Geographers), which has resulted in the Iberian Colloquiums on Geography, held alternately in Spain and Portugal. AGE has also signed permanent agreements with geographical associations from other countries,
such as the Institute of British Geographers and L’Association Française pour le développement de la Geographie. AGE holds the presidency or the vice-presidency of the IGU Spanish Committee in rotation.

AGE’s administrative bodies comprise the general assembly, which is an annual meeting of all members, and the board of directors, which is in charge of implementing the program of activities put forward by the general assembly. The board of directors is elected for a four-year term and is partially re-elected every two years. The board of directors comprises the President, the secretary, the treasurer, and eight board members. The working groups are governed by specific committees.

Submitted by Javier Martin-Vide
Spain: Real Sociedad Geográfica (RSG) (Royal Geographical Society)

Founded: 1876
Location of headquarters: Madrid
Website: www.realsociedadgeografica.com/es/site/index.asp (Spanish), www.realsociedadgeografica.com/en/site/index.asp (English)
Membership: 250 (as of 2015)
President: Joaquín Bosque Maurel
Contact: secretaria@realsociedadgeografica.com

Description and purpose

The purpose of the Real Sociedad Geográfica is to promote the production and dissemination of geographical knowledge with particular attention to subjects of relevance to contemporary social, political, and economic life. The focus of studies is on Spanish lands and the peoples linked to Spain by language and other reasons. The society organizes and participates in conferences, courses, and workshops; undertakes state-funded research projects; publishes a journal, books, and other special publications; and has a major collection of books and maps.

Journals or major publication series


Current activities or projects

The primary activities of the society are courses, conferences, workshops, and exhibitions, which are either conceptual (about the nature of geography), territorial (about the different lands and nations), or about geographical tools (cartography, GIS, and remote sensing).

Boletín de la Real Sociedad Geográfica is the Society’s primary journal which has been published as far back as 1876 (except during the Spanish Civil War), being the oldest geography journal in Spain, with papers by eminent geographers, historians, and scientists. The society has also published around fifty books and other special publications relating to its activities.

The society has a large bibliographic and map collection which is stored at the Spanish National Library in Madrid including over 11,000 books, 12,700 booklets, 150 journals, and several thousand maps dating back to the mid-eighteenth century.

Brief history

The society was founded in 1876 as the Sociedad Geográfica de Madrid and renamed the Real Sociedad Geográfica in 1901. Among the first members were prominent people in the public and scientific sphere including politicians, military cartographers, engineers, and planners.

At that time, geography was a discipline of great political relevance offering a scientific and academic basis for international relations as well as for domestic development. Thus, the society started its activities immersed in a political climate of colonial expansion, national unifications, industrial imperialism, and also
urban and infrastructure developments. In particular the society was involved in exploration and discovery in Africa and Asia including Guinea, Morocco and the Sahara, and Eastern Africa.

Another chapter was the collaboration and merging with other scientific and exploration associations such as Sociedad Española para la Exploración de África and Sociedad Española de Africanistas y Colonialistas (later the Sociedad de Geografía Comercial), all of which were absorbed by the Sociedad Geográfica.

The society organized several important scientific meetings such as the Congreso Español de Geografía Colonial y Mercantil (1883) and the Congreso Geográfico Hispano Portugués Americano (1892, 1914, 1921), which included representatives from the geographical societies on the American continent.

Spain, represented by the Real Sociedad Geográfica (RSG), was one of the seven founding countries of IGU in 1922, which many other associations have joined right up until the current IGU’s Spanish Committee.

Submitted by María Luisa de Lázaro y Torres
Spain: Sociedad Geográfica Española (SGE) (Spanish Geographical Society)

Founded: 1997
Location of headquarters: Madrid
Website: www.sge.org
Membership: 1500 (as of December 31, 2013)
President: Juan José Herrera de la Muela
Contact: sge@sge.org

Description and purpose

The Spanish Geographical Society (SGE) is a nonprofit organization aimed at encouraging the rediscovery of Spanish history throughout the world by Spanish people, promoting geographical and social awareness of different peoples, creating an understanding of, and human contact with, different cultures and ways of life, and becoming the driving force behind projects that serve investigation, science, art, and the revitalization of geography and traveling as a personal discovery, using the means available in the twenty-first century. The society was appointed “Of Public Interest” by the corresponding government department in 2012.

Journals or major publication series

Several titles published on Spanish exploration across the world. http://www.sge.org/

Current activities or projects

The SGE works actively to organize conferences, courses, seminars, periodical gatherings, trips, cultural tours, slideshow projections, exhibitions, travel book releases, and other events open to the general public with a special preference for members.

Each year, the SGE gives out awards in eight categories related to its objectives.

The society maintains an editorial activity by publishing specially ordered books on subjects relating to exploration, travel, and geography. One of the primary concerns of the SGE is the recovery of ancient geographical books and expanding a comprehensive library and documents department at the society’s headquarter.

Brief history

The SGE was founded on October 12, 1997, through the initiative of a mixed group of people with a common interest in geography and traveling. It already comprises more than 1500 members and boasts the participation of geographers, travelers, writers, anthropologists, ecologists, biologists, people professionally involved with tourism, and many other people with an interest in the society’s goals as stated in its articles of association.

The SGE has its headquarters at the Instituto Geográfico Nacional in Madrid.

The SGE’s goals are to promote the study and understanding of our environment, to promote geographical knowledge by all available means, and, above all, to make public modern geographic science while also investigating
the adventurous side of Spanish geographic discoveries throughout history.

The society publishes its own magazine every four months in collaboration with some of the most prestigious travelers, geographers, and investigators in the country.

For the past 13 years the SGE has awarded different prizes (National, International, Enterprise, Trip of the Year, Image, Communication, Investigation, and Honorary Member) to several distinguished representatives of these fields in a ceremony which is attended by members of the SGE and by outstanding representatives of the cultural and social world and the media.

The society keeps in touch with different societies all over the world as well as Spanish geographers’ associations in order to encourage collaboration projects with university departments within this discipline. The SGE also collaborates with Spanish universities and will continue to increase contact with other groups of investigators (sociologists, biologists, city and town planners, etc.).

Submitted by Juan J. Herrera de la Muela
Spain: Societat Catalana de Geografia (SCG) (Catalan Society of Geography)

Founded: 1935
Location of headquarters: Barcelona
Website: http://scg.iec.cat
Membership: 429 (as of June 16, 2014)
President: Josep Oliveras-Samitier
Contact: scg@iec.cat

Description and purpose

The objectives of the SCG are to promote research in all geographic areas and specialties, to spread knowledge among the population, and to collect and publish the scholarly and popular works of its members and other interesting authors. The SCG gives preeminence to items that deal with the geography of Catalonia, but it is open to the dissemination of knowledge from other regions, geographical problems, and foreign authors. It is a member of the EUGEO (Association of European Geographical Societies) and of the Spanish Commission of the International Geographical Union (IGU). SCG members are mostly professional geographers.

Journals or major publication series


Current activities or projects

The SCG publishes, since 1984, a biannual journal of geography (Treballs de la Societat Catalana de Geografia). It also publishes a collection of books on topics in geography by ancient and modern authors as well as foreign and domestic authors. The SCG’s website gives information of the activities that are carried out, the history of the Catalanian geography, and work of people dedicated to it. The SCG maintains another website giving information on tools and resources for teaching geography (http://ensenyament-geografia.espais.iec.cat).

Submitted by Josep Oliveras-Samitier
Spatial analysis

Levi J. Wolf
Alan T. Murray
Arizona State University, USA

Spatial analysis is an expansive term that has been defined a number of different ways. Primarily, spatial analysis is the process of generating useful knowledge to answer questions about order, pattern, and structure inherent in spatial problems (Anselin, Murray, and Rey 2013). Structure and pattern arise from tangible questions emanating from spatial systems. Spatial analysis techniques seek to facilitate answering these questions through the application of formal mathematical, statistical, and/or computational methods. The domain has a long history of focusing on geographic issues and problems, where generating useful knowledge/insights can help policymakers, stakeholders, managers, and the broader scientific community.

By engaging problems that arise from geographic interactions, spatial analysis is faced with ever evolving questions and problems. This tightly couples spatial analysis techniques to empirical questions. Understood in this way, spatial analysis is the process of describing, characterizing, understanding, creating, and visualizing latent order, pattern, and structure associated with geographic phenomena. As the questions and problems change, so do the methods needed to address them. Spatial analysis therefore reflects many diverse perspectives, each with slightly different goals, concerns, and methods. However certain traits are common across spatial analysis domains and perspectives.

The most common of these traits are a consistent use of abstraction, mathematical expression(s), a quantitative or empirical focus, and an aim for broader spatial understanding.

Spatial analysis involves the use and application of statistical, mathematical, and/or computational methods that account for or exploit order, pattern, and structure, providing insights about how spatial processes work and may be managed. These methods and tools are widely used, both within geography and in other domains. As many problems across domains have some geographic component, spatial analysis is often an interdisciplinary point of contact between geographers, spatial scientists, and decision-makers. Spatial analysis provides unique, powerful, and incisive ways to generate useful knowledge because it involves the use of quantitative methods as part of a process.

Distinguishing traits

Spatial analysis is a unique domain that exists at the crossroads of many different disciplines and methodological perspectives. In addition, spatial analysis frequently offers one view out of many on a problem or question. While there are numerous domains that attempt to understand order in the world, spatial analysis is different in many ways. The questions that spatial analysis addresses are quite different from other types of aspatial questions. In addition, geographic data contain information that spatial analysis techniques can exploit in novel ways. Therefore, the way data are stored, analyzed, and displayed must reflect this information, as structure, order, and pattern inherent in geographic data are a central
Spatial analysis questions all share these common traits and confront similar difficulties. As spatial analysis perspectives are all united by a goal to provide useful knowledge about the order inherent in spatial data to answer questions, contextual factors shape the character of spatial analysis. Therefore, discussing these common themes, how they arise, and how they contribute to answering questions about spatial systems shows the unique geographic context of spatial analysis.

Spatial data by design reflects the fact that observations have geographic meaning. In particular, observations have geographic position on the surface of the Earth defined by coordinates in two or three dimensions. Thus, patterns in the proximity of observations frequently contain information about underlying spatial processes. However, there is frequently more information associated with each observation than its position alone. These “attributes,” essentially the traits associated with each observation, are important. If attributes are ignored, significant amounts of information are discarded. To handle the unique pairing of position and attribute information, spatial analysis must be sensitive to potential variation in patterns of both attribute(s) and location. Thus, spatial analysis techniques tend to be quite different from other analytical techniques because the data express information in a distinctly geographic way.

Spatial representation of observations is unique for geographic data. In spatial analysis, many different types of pattern may be present in data. However, some cannot be identified or visualized using traditional means such as histograms or distributional charts. The fact that the data contain spatial specificity also provides more information than just the nonspatial attributes alone, and requires unique systems of representation. Thus, storage, visualization, and mapping are quite important for spatial analysis. Often, many different “views” of data must be considered, each capturing a different aspect of the structure in the data. Spatial analysis techniques frequently rely on visualization and the use of multiple views, particularly when mapping, distinguishes spatial analysis from other domains.

Because spatial data are novel, demands on information systems that process spatial data are different from traditional database infrastructures. Because position and spatial configuration express special kinds of structure, reference frames for geographic data preserve geometric or topological relationships. As such, spatial analysis requires different technologies to handle and retrieve data. These systems are called geographic information systems (GIS). Because spatial analysis is often strongly coupled with explicit GIS functionality, discussion of spatial analysis is sometimes equated to GIS operations. This is an incomplete view. While spatial analysis may include GIS functionality, it goes beyond capabilities offered through the use of GIS. Basic tools in GIS for handling representation, storage, and display of spatial data are important, but come nowhere near to providing a complete picture of the breadth and depth of spatial analysis as a domain of inquiry.

Unique structures in spatial data abound (Anselin, Murray and Rey 2013; Murray 2010b). Common concerns focus on the structure of spatial dependence between observations’ attributes or their locations, the spatial heterogeneity of this dependence in space, and problems with the observational scale or framing of the data. Spatial dependence is a central concern for spatial analysis techniques. It manifests when the...
geographic location or measured properties of data rely on the spatial configuration of the data. That is, spatial dependence occurs when spatial relationships between observations change or affect the configuration, behavior, or value of observations themselves. Spatial heterogeneity manifests when the behavior of spatial dependence changes in space. As measuring and understanding the structure of spatial heterogeneity involves studying the spatial dependence of the data, it follows that understanding spatial heterogeneity also requires special treatment.

Handling frame and scale effects requires special attention as well. These effects stem from direct limitations on processes of observation, representation, and the conceptualization of spatial processes. Frame effects, in the context of spatial analysis, are encountered when changing the area of focus for a spatial question changes the results. This implies that all of the work done to characterize and explain pattern and order within the problem frame may not have external validity. Because spatial heterogeneity is often the rule rather than the exception, these issues challenge spatial analysis results. Similarly, scale effects manifest when a question’s scope or extent is either reduced or expanded, but stays centered at the same spot. These can be thought of, more generally, as a kind of frame effect, as the dependence on spatial frame is scaling uniformly rather than translating in space. In addition, this relates to well-known aggregation effects such as Simpson’s paradox and the modifiable areal unit problem, where the significance or character of structure in data changes significantly depending on how it is grouped.

With all of these concerns, it is clear that spatial problems are distinct from other types of problems. As such, spatial analysis has contributed significantly to the body of useful knowledge on dealing with geographic issues and problems. Spatial analysis is also well poised to continue doing so due to the unique traits of the encountered problems. Again, spatial analysis generates useful knowledge to answer geographic questions. It does so by making latent order and structure in spatial data clear through formal mathematical, statistical, or computational methods. Spatial analysis is distinct because spatial structure in geographic data is different: the ways which spatial problems are framed, measured, expressed, and solved are often more fundamental concerns than in other analysis perspectives.

### Analytical approaches

As mentioned, spatial dependence or heterogeneity effects, in addition to problems of scale and frame dependence, are novel issues characteristic of spatial analysis and are challenging to handle appropriately. In light of all of these concerns, spatial analysis perspectives are unique. The aim is to generate useful knowledge to answer questions involving spatial data. The large diversity in spatial analysis approaches is a response to the many needs of managers, stakeholders, policymakers, and scientists using geographic data. Spatial analysis is often strongly use-oriented, so abstractions that are useful for answering some questions may not be helpful for others. The varying perspectives in spatial analysis still aim to make inherent structure, pattern, or order apparent while generating useful knowledge, but often they seek to answer different questions or do so in different ways. It is important to note the many communities of practice that have developed in spatial analysis; their concerns often appear disparate. However, each perspective shares in the same aim common to spatial analysis discussed above, that of generating useful knowledge to answer questions about latent structure, pattern, or order in spatial data.
Focus on applied problems has long been central to spatial analysis. Commonly cited as one of the first cases of a coherent spatial analysis perspective, John Snow’s 1854 analysis of cholera outbreaks around the Broad Street pump in London is noteworthy (Snow 1855). By considering spatial dependence of cholera cases in relation to water pumping sites, Snow used spatial logic to advance the then-controversial germ theory of disease. In this vein, later spatial analytical efforts developed new tools by focusing primarily on derived useful knowledge: Von Thünen (1875) and spatial models of economic rents developed by studying urban land markets, Weber (1909) and location models of economic distribution structures, and McHarg (1969) and proximal relationships using overlay and map algebra for examining land-use systems. Later advancements made in the mid- to late twentieth century hinged on the same focus on tools for producing useful knowledge, and were empowered by the rise of digital computation. Computing access and pervasive data collection facilitated spatial analysis intertwined with GIS (Goodchild et al. 2000), but also into common day-to-day life experiences. This integration, particularly due to availability of inexpensive digital computing and abundant data, has driven development of quantitative tools for spatial analysis. Newer spatial analysis perspectives retain this overarching quantitative focus on spatial dependence, heterogeneity, and frame/scale effects, yet with more computing power, data, and nuance than before.

With this in mind, current spatial analysis perspectives seek to satisfy slightly different goals while generating useful knowledge. Modern spatial analysis has also retained the quantitative and computational traits developed during the latter twentieth century. Current techniques in spatial analysis retain the strong linkages to problem-based reasoning established throughout its history, and the development of new techniques often occurs when particularly difficult or poorly behaved real-world problems arise. Some of the more prominent spatial analysis methods include: exploratory spatial data analysis, spatial regression, spatial optimization, and geosimulation. These methods are discussed in some detail here.

**Exploratory spatial data analysis**

One popular perspective in spatial analysis is exploratory spatial data analysis (ESDA), emphasizing the discovery of latent trends. These approaches avoid imposing predictive hypotheses on data before analysis. Work applying ESDA approaches does not aim to confirm prior expectations about how a process behaves but interacts with the data in ways that generate new questions and provide grounds for new hypotheses. Discussion and application of ESDA can be found in Anselin and Bao (1997), Fotheringham, Brunsdon, and Charlton (2000), Murray et al. (2001), and Murray (2010b), among others. ESDA techniques characterize the geographic distribution of the data by searching for spatially interesting phenomena that violate expectations of spatial heterogeneity or independence. Frequently, ESDA includes inherently visual components, relying somewhat on natural pattern-seeking capabilities of the human mind to validate or initiate the application of methods or statistical tools. Visualization and methods to support searches for pattern or disparity in data are central concerns of ESDA approaches.

ESDA techniques are centered on examining spatial dependence in data and determining whether or not spatial dependence is heterogeneous in space. Without imposing particular models, ESDA techniques tend to focus on uncovering structure and order in data through examination of “neighborhoods” around observations. Measures of nearness
Spatial analysis

frequently incorporate aspects of contiguity, proximity, and spatial interaction. Based on this, ESDA techniques seek to uncover locales that are unusual with respect to general patterns of spatial dependence.

Depending on the type of spatial data involved, ESDA techniques can operate differently. The mechanics of individual techniques vary significantly depending on the spatial support. However, ESDA techniques often consider changes in spatial dependence over the entire problem space or look for specific locales that defy the broader observed trends. ESDA techniques often are flexible with respect to the scale of the processes involved. “Global” measures that characterize the behavior of the entire problem space and “local” measures that examine a subset of the geographic space both are prevalent. Regardless of the scale, visualization techniques that pair multiple views of the data in different formats provides a more flexible and incisive way to understand the patterns in the data across time, space, and scale.

Spatial regression

Spatial regression aims to describe the geographic structure and distribution of the data by focusing on modeling conditions of spatial dependence and heterogeneity using formal mathematical structures. Detailed discussions of the various approaches can be found in Anselin (2003, 2010), and Pace and LeSage (2009). Spatial regression models estimate the value of a given attribute in space using information about the relationships between attributes, with spatial dependence being one of many potential factors. By determining which factors are statistically significant in influencing the observed location or attributes of features, spatial regression provides incisive empirical models for understanding spatial variation.

Spatial regression techniques account for many different kinds of structure and pattern in geographic data. Because of sensitivity and flexibility, spatial regression techniques are used across many different problem domains and spatial data types. A common type of spatial regression is the spatial lag model. This model assumes that the value of an attribute under study is likely related to the value of its neighbors. By considering spatial variance as its own direct effect, this model specification is quite sensitive to spatial dependence between observations and their neighbors. However, another regression model type uses regularities in the error terms to structure estimations. These models, called spatial error models, suggest that the errors in the trait to be estimated vary systematically in space rather than the values themselves. Another kind of spatial regression model divides the problem area into distinct regions and estimates the values in question differently in each new subregion. These spatial regime regression strategies allow the analyst to vary the expected structure of spatial dependence across the problem space in a systematic way. Other techniques, such as conditional autoregressive models or simultaneous autoregressive models, allow spatial information to be incorporated into regression models, albeit through different conditioning processes. These techniques are frequently used, and play a central role in spatial regression.

Because of their flexibility and the diversity of model specifications, spatial regression can be applied to many diverse problem types. Spatial regression techniques provide valuable insight into the structure of spatial dependence and can take into account issues of modeling spatial heterogeneity. With spatial regression techniques, hypotheses concerning potential relationships between explanatory variables can be formulated and evaluated. This power is useful for examining spatial processes. As such, spatial regression
SPATIAL ANALYSIS

techniques provide a unique set of methods to predict and estimate attributes in the sample space and provide a unique contribution to answering spatial questions.

Spatial optimization

Spatial optimization shares spatial regression’s focus on modeling order and structure. But, instead of examining or specifying the order and structure inherent in data, optimization approaches tend to focus on modifying current systems and predicting how they will then behave. Optimization techniques have long been part of spatial analysis, and applications and discussions of common spatial optimization techniques are provided in Tong and Murray (2012), Murray (2010a), and Church and Murray (2009).

A few common concerns for spatial optimization are how objects are located, service and trade, organizational structure orientation or alignment, and system flow. Common applications of spatial optimization include location of emergency facilities, land-use problems, districting or assignment problems, and urban planning. Spatial optimization models are useful when three aspects of a problem can be identified: a goal or objective, a set of decisions to make in order to achieve the goal, and usually (but not always) a set of constraints that restricts decisions. Optimization models are then constructed from these pieces to reflect latent order in the spatial relationships or aspatial structures in the data.

Of course, relationships between components of a spatial optimization model may not be obvious. Dealing first with the goal, spatial optimization is most useful in cases where there is some desired outcome to be achieved. Often, policymakers or modelers designate these objectives in advance. They reflect concerns about the desired end-state of the model. Narrowly, objectives can be thought of as some desirable property that the analyst uses to discover a better structure or configuration of the system under study.

Second, some decisions about how to achieve the objective must be made. For example, in a production distribution system, spatial decisions could affect what warehouse supplies which stores, what roads are selected to ship material or finished goods, or where a factory should be located. Decisions are formally expressed in the problem structure. These “decision variables” are related to each other through the objective and through the constraints on decisions discussed later. The value of these decision variables expresses information about how the system must be configured in order to achieve the objective. The kinds and complexity of decisions that are modeled in the spatial optimization literature is incredibly diverse. This is because decision variables are designed to reflect real-world choices and are specified for a specific problem instance.

Finally, spatial optimization problems typically have some kind of constraints that prevent simple satisfaction of the objective. Often, there are limiting factors that restrict a system’s performance, particularly with respect to a given objective. For example, common constraints on the performance of businesses, consumers, governments, or city planners are budgets, regulations, or laws. Constraints reflect these factors. They also link the decision variables to limiting factors in meaningful ways. Constraints encode information about the problem, as aspects of the inherent spatial order of a system are identified and codified in mathematical language. Unconstrained problems do exist, though, and are sometimes quite difficult to solve. Aspatial factors frequently come into play in spatial problems, so constraints on spatial optimization models are also designed to incorporate these various concerns. However, in general, constraints incorporate both spatial and aspatial information.
into a problem, making spatial optimization techniques useful across many domains.

Once these three components are specified and put together into the optimization problem, appropriate solution techniques are applied to find a useful answer to the question modeled. Solution techniques can either be exact or heuristic. Exact solution techniques use the problem’s special mathematical properties to generate a solution that is guaranteed to be optimal. The fact that solutions found by exact methods carry this guarantee is useful, but exact techniques are often difficult to discover and slow to execute given the complexity of many problems. Heuristic strategies, however, are developed individually for different problem instances. These techniques typically find acceptable solutions quickly, but cannot demonstrate how good the solution is with respect to other, undiscovered solutions. Regardless of the specific tool, though, solution techniques serve to link the output of the optimization procedures to the original question posed by the spatial optimization model. Questions of spatial dependence and heterogeneity are captured in the construction of the model and affect their solutions. This focus on making decisions about some system while accounting for its spatial structure and relationships makes spatial optimization approaches quite distinct from other spatial analysis perspectives and heightens their ability to contribute useful knowledge to spatial questions.

Geosimulation

The final, broad perspective in spatial analysis discussed here is spatial simulation, or geosimulation. These techniques aim to provide insight into “what-if” questions about the future configuration of spatial systems. Discussions of geosimulation techniques and applications can be found in Beneson and Torrens (2004) and Ward, Murray, and Phinn (2003). Geosimulation is often most relevant when attempting to figure out what will happen given some set of decisions to be made or rules to be followed. These approaches aim to simulate the emergent behavior of complex geographic systems. Through modeling small-scale interactions between elements of the system, the behavior of the entire system can be understood. Geosimulation techniques focus on exploring hypothetical, “what-if” configurations of existing geographic systems. This aim provides a unique toolset to spatial analysis more generally: simulating spatial order given some current set of known behaviors provides the ability to examine potential scenarios that may occur given a particular course of action.

To construct geosimulations, analysts decide on hypotheses about how parts of a spatial system should behave given a certain set of suppositions. These rules describe the multiple states any individual component of the spatial system can occupy at some time. They also govern the transitions between states in time and control how elements affect their neighbors. Consensus in current scholarship drives the development of these rules. However, these suppositions can also be structured in order to examine possibilities on the research frontier, where no clear consensus about the states or behaviors of a system exists. In these cases, suppositions and their attendant rules allow analysts to examine many possible scenarios about the future given different rules of interaction. Once the data are codified and the rules of interaction are formalized, the simulation begins by using the rules to simulate a new configuration given some observed initial configuration. Simulated elements act according to simple rules, but can generate complex emergent patterns of spatial dependence and heterogeneity. This process of applying the rules to the dataset continues until
some kind of termination criteria, designated in advance, is reached.

Evaluating a geosimulation is, in some ways, trickier than evaluating the output of other spatial analysis tools. Using simulated data, the original hypothetical questions can be answered. However, it is difficult to validate the rules used to simulate new data, as they depend on domain knowledge about the system being simulated. When simulations are primed on different hypotheses about what drives the system’s behavior, differences between simulated scenarios can provide support for or against causal or descriptive hypotheses. Analogous to the hypotheses testing or generating present in other spatial analysis techniques, geosimulation techniques are varied in their aims to confirm hypotheses or generate new questions based on simulated output of potential scenarios. It is precisely these “what-if” cases where geosimulation provides powerful insight into the unique structure of the initial data, and how it may affect future derived data. Geosimulation’s utility in answering questions at the frontier of research, in situations where more traditional causal or predictive models are too immature or too rigid, serves a unique yet essential niche in spatial analysis’ pursuit of useful knowledge to answer spatial questions.

**Common themes**

Given these primary analytical thrusts in spatial analysis, recognizing the common themes between approaches helps to tie the domain together. Spatial analysis techniques are incisive and powerful in the face of unique geographic issues due in large part to a few traits common across many spatial analysis perspectives. By focusing more generally on order, pattern, and structure inherent in the phenomena around us, spatial analysis aims to generate useful knowledge that answers questions about geographic systems. The different perspectives on spatial analysis share common design choices that make it possible to achieve powerful, general insights. These general traits are abstraction, mathematical expression, a focus on quantitative or empirical questions, and a drive for broader spatial understandings. Essentially, these themes characterize the factors that differentiate spatial analysis methods from other perspectives. While these approaches have their merits, spatial analysis makes these tradeoffs to ensure unique, powerful insights into many different problem domains.

Abstraction is a long-used design technique to reduce complexity or improve clarity of complex problems or questions. In spatial analysis, abstraction is employed as a way of uncovering structure in data. By removing or ignoring smaller, confounding detail, it may be possible to understand the problem more clearly. Extraneous or unneeded information may be ignored and the analyst can focus only on the important, central aspects of a system’s structure. As a tool for designing inquiry, abstraction also works well with the other aspects of spatial analysis, namely mathematical expression and drive for general explanations. Thus, through intelligent application of abstraction, spatial analysis techniques generate strong and robust results.

Abstractions are integral to many different kinds of techniques, so spatial analysis’ use of them is not, by itself unique. However, abstraction in spatial analysis typically enables expression of the problem in mathematical language. This kind of abstraction is necessary to exploit special mathematical structures in spatial problems. By pairing the use of abstraction to remove unnecessary or irrelevant detail and mathematical expression to be explicit about the details remaining, structure and pattern in the important traits of a problem become clearer. Mathematical structure abounds in the systems
humans encounter. Whether by design or as a product of analysis, these structures often illustrate regularity in a problem. Because the mathematical characterizations of problems are often easier to express, understand, and control, these structures can be used to make complex geographic problems more accessible.

Because spatial analysis aims to answer questions that arise in dealing with spatial systems, it is empirical in focus. This means that spatial analysis is often concerned with the tangible, measurable, and quantifiable. However, this does not mean that spatial analysis fails to handle issues of uncertainty or vagueness. Many methods have been developed to deal with the unknown or unknowable factors that affect geographic questions. Because geographic questions frequently involve the limitations of formal expression, measurement, and quantitative tools, spatial analysis is acutely aware of where focus on known, observable factors is not sufficient. Spatial analysis, while frequently engaging issues of uncertainty is often quantitative and empirical.

Finally, approaches in spatial analysis all exploit order inherent in geographic data. By understanding and exploiting the unique mathematical structures of geographic data, broader understandings of how space and location affect geographic processes are possible. As such, spatial analysis techniques tend to exploit order common across many problems, rather than unique to a single instance. This means that the frontier in spatial analysis research is often split into two pursuits. One involves the efforts to develop stronger tools to hit stubborn special cases that resist more general solutions. The other deals with the examination and characterization of commonalities in spatial structure across problems. As such, spatial analysis work often moves freely between theoretical work on general structures and applied work on particular geographic problem instances.

Overall, these four traits are common across spatial analysis approaches. They describe the core of what differentiates spatial analysis from other geographic perspectives and other data analysis approaches. Each is tightly related to the other, as some design choices amplify the effectiveness of others. Abstraction helps mathematical expression remain concise and specific. Mathematical structures lend themselves to empirical or quantitative approaches. Geographic data, carrying the unique problems and possibilities associated with spatial information, is amenable to quantitative analysis when sufficiently and effectively abstracted from irrelevant detail. Together, a quantitative focus, use of efficient abstraction, mathematical expression and structure, and a search for broader spatial understandings of geographic processes drive the success of spatial analysis methods.

Future trends

As a domain, spatial analysis approaches have been useful across many problem domains throughout time. However, spatial analysis faces a few central challenges. Good overviews of spatial analysis’ potential challenges and opportunities exist in Goodchild et al. (2000), Anselin (2010), or Murray (2010b). However, techniques that handle uncertainty, intensify spatial analysis’ past interdisciplinary linkages, and are scalable over the pervasiveness and size of modern geographic data collection are often cited as central areas of development for spatial analysis.

With respect to the size and pervasiveness of geographic data, concerns about “big data” abound. Efficient computational techniques to manage the increasing size and scope of geographic data are needed. With computation getting cheaper and microelectronics getting smaller, data are recorded at astounding rates...
SPATIAL ANALYSIS

and with startling pervasiveness. Intensive data generation and its pervasive scope present unique challenges for spatial analysis. Discussions of big data reflect these concerns and spatial analysis will continue to be central to the analysis of data, regardless of its size. However, current analysis techniques sometimes are too slow when presented with large datasets, and systems of geographic data representation can break down when attempting to visualize extremely dense sets of observations. Developing better tools, computational methods, and visualization strategies will be central to emerging spatial analysis research.

As geographic data that are broader in scope and more in-depth are collected, the relevance of spatial analysis increases across disciplines. Many social problems and scientific questions concern objects with some kind of geographic representation: location is a ubiquitous trait. With better, wider coverage of geographic data, spatial analysis is well poised to answer questions about geographic systems in the future. Spatial analysis is certainly positioned to play a major role in future interdisciplinary science applications. Its historical role in tying together disparate domains in order to solve intractable problems in geographic contexts will likely continue as research focusing on commonalities in order, structure, and pattern continues. Across domains, calls for spatially integrated social science (Goodchild et al. 2000) or intensification of geographic methods in the natural sciences reinforce the centrality and importance of spatial analysis methods and, more generally, reasoning about geographic problems in a quantitative fashion.

Concerns about the size, pervasiveness, or speed of computational tools may prove irrelevant, though, if more fundamental issues about uncertainty are left unaddressed. Often, geographic measurements are uncertain. This can be due to measurement uncertainty, where the exact dimension, position, or trajectory of an object is difficult to observe reliably or precisely. It can also be due to representation issues, where the methods to represent an object’s geographic properties are not sufficiently descriptive or lack some dimension needed to fully understand the object. Uncertainty can also be caused by a lack of information about the problem itself. Concerns posed by the emergence of “big data,” essentially concerns about what to do with massive and pervasive datasets, are mirrored by questions about what to do in cases where the data is not “big” enough, and components needed to apply spatial analysis approaches are either missing or estimated. Handling uncertainty, particularly how it can affect core assumptions or suppositions needed for analysis, is a central concern for spatial analysis work going forward.

Another important trend is the proliferation of spatial analysis tools. Spatial analysis has traditionally been a domain reflecting concentrated technical skill, using powerful insights derived from mathematics to exploit and characterize latent order, pattern, and structure in geographic data. Many commercial GIS packages now contain basic spatial analysis functionality. Yet, many lack the appreciation and understanding of spatial analytical traditions more broadly and their relationship to central issues in statistics, mathematics, and issues of representation in geography. As such, development of spatial analysis tools focuses both on utilizing pervasive and cheap computation to provide new functionality, but also must make it easier to effectively apply this functionality. By improving CyberGIS and the quality of open spatial analytical software, the quality and accessibility of spatial analysis techniques will be improved. Improving the power and usefulness of spatial analysis techniques in conditions of pervasive data and uncertainty in data will certainly be difficult. Accomplishing this while remaining accessible to end users and
generalizable across many problem cases, though, will assuredly be quite the task.

SEE ALSO: Exploratory spatial data analysis; Geocomputation; Geographic information science; Geographic information system; Geostatistics; Measuring spatial dependence; Quantitative methodologies; Regional science; Spatial econometrics; Spatial optimization

References


Von Thunen, Johann. H. 1875. Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie [The Isolated State in Relation to Agriculture and the National Economy], vol I. Berlin: Henipel and Parey.


Spatial concepts

Karl Grossner
Stanford University, USA

The term “concept” has been defined in several ways. From the perspective of cognitive science, concepts are mental constructs about the nature of material and abstract things, and the relationships that obtain between them. The words we use in communication – and some would add in reasoning – refer to concepts, and their meanings are shared between individuals to a greater or lesser degree. There is little agreement among the philosophers, linguists, and psychologists who study the nature of concepts as to how such mental representations and lexical concepts precisely relate to language, thought, and activity.

An important branch of geographic information science (GIScience) studies such ontological issues from several perspectives, motivated by the need for better analytical software, better navigational devices, and better geographical education. Spatial information theory has emerged in the past two decades as a blending of interests from researchers in many fields, including geography, cognitive psychology, computer science, and linguistics. The biennial Conference on Spatial Information Theory (COSIT) has become an important meeting ground and publication venue for this multidisciplinary community of interest.

The creation of geographic information and knowledge follows from the gathering of observational data from human senses and mechanical sensors. Understanding the progression of representations involved at each stage of this process is important from both scientific and social theoretic points of view. Because geographic knowledge informs important policy and geodesign decisions, it is essential we understand the sources of error and bias in measurement and interpretation. The need for development of semantic reference systems as an essential complement to existing spatial and temporal reference systems has been described by Kuhn (2003) and has motivated considerable research activity on geographical ontologies.

Spatial concepts and spatial reasoning lie at the heart of geography as a multifaceted academic subject, and the professional practice of geography – one that began with the fundamentally spatial tasks of measuring Earth features and describing their spatial relations. Geography’s numerous subfields are often classified at the highest level in terms of conceptual or methodological duals: physical and human, quantitative and qualitative, space and place. As the intellectual traditions of geography have multiplied and grown more varied over two millennia, space and spatiality have remained integral to each. However, geographers’ conceptions of space are by no means uniform. There has, for example, been considerable debate on whether geography is a “spatial science.”

Space, space–time, and place

Broadly speaking, there are in geography two distinctive views of space: first, as a fundamental attribute of reality that is, with time, the mathematically describable context for natural phenomena; second, as a count noun standing for human conceptual constructs borne of
individual experience and societal factors such as power relations. Within the first view, there are distinctive conceptual frames: that of an absolute, Newtonian space as “container,” normally associated with Euclidean geometry, and that of relative space within which distance may be geometric or derived from other relations. The term “place” is often used to refer particularly to the human experience of space. Description affords analysis, and the concepts discussed below are by and large those enabling mathematical and formal-logical spatial analyses.

In modern physics, space and time are understood to be joined as the unbounded continuum of space–time, a four-dimensional reality wherein all phenomena are events. Certainly, to exist in a location with certain attributes is to exist there in that state for some period of time. However, those dynamics are not necessarily a focus in any given investigation. Other branches of science, including geography, have sought to develop four-dimensional representations, but have found it generally useful to hold time apart from the spatial dimensions in a “three plus one” model.

The spatial concepts enumerated here should be understood to have an important temporal dimension even where time is not mentioned explicitly.

**Universality and ubiquity**

Although geography is thoroughly spatial, the most fundamental of spatial concepts are by no means exclusive to geography. There are distinctive disciplinary perspectives on many, and some spatial terms (i.e., lexical concepts) can have multiple meanings. For example, surface normally refers to the physical bounds of an object, particle, or organism. In oceanography, surface topology is used to map currents and study climate. Materials scientists study the physical properties and chemical interactions of surfaces as interfaces between phases of matter. The calculations and design choices of structural engineers depend in part on surface stress properties of building materials. The term “surface” also refers to the mathematical description of a two-dimensional field, which is an important analytical construct in geography among many other fields. Finally, surface is also used to mean, “that which is superficial or most readily perceived” in any domain.

This example illustrates the way that many core spatial concepts have both a basis in physical reality and important metaphorical meanings in the physical and social sciences, in engineering and other design fields, and in the humanities. Johnson (1987) has argued convincingly that the ubiquity of metaphorical spatial conceptual language stems from the human bodily experience of physical interaction with the world, and proposed corresponding cognitive structures, termed image schemas, that represent for example the concepts and experience of containment, path, force, counterforce, and center-periphery.

**Spatial concept taxonomies**

There has been a growing impetus to enumerate spatial concepts and to organize them in one or more taxonomies, both within and outside of disciplinary perspectives. The motivations for this include (i) designing instruction in academic subjects such as geography, geoscience, chemistry, and engineering; (ii) designing training to improve spatial reasoning ability at all developmental and educational levels more generally; (iii) developing more usable software for professional practice, including geographic information systems (GIS); and (iv) making useful links between the above requirements
and basic research in cognitive psychology and education.

For education

Scientific instruction proceeds through grade levels with the introduction of concepts and conceptual frames of increasing complexity, seen for example in *Benchmarks for Science Literacy* (American Association for the Advancement of Science 1993). Hierarchies of complexity for spatial concepts in geography have been developed by at least two research groups. In one case (Golledge, Marsh, and Battersby 2008), five concept/task tiers are identified: primitive, simple, difficult, complicated, and complex; single examples of concepts for each of those levels are location, distance, adjacency, scale, and interpolation. In a second case (Jo and Bednarz 2009), three levels are identified: primitive, simple-spatial, and complex-spatial; single examples of concepts for these are, respectively, location, distance, and distribution.

For software design

Spatial analytical software and GIS enable the application of computational methods to research questions by means of specific models and algorithms. These are often instantiated as modular software programs (“tools”) applied individually or in sequence upon spatial data. Choosing the correct tool or tools to use for answering a particular question, from a library of hundreds, is problematic. This problem could be mitigated by organizing such tool libraries according to some spatial conceptual logic, preferably informed by cognitive studies and our improving understanding of hierarchical concept complexity. The structure for such organizing models could range from simple thesauri specifying broader, narrower, and related relations, to formal ontologies.

For improving spatial ability

There has been considerable psychological research on spatial cognitive ability in the past several decades, and there is now a common understanding that spatial ability is not monolithic, but rather multifaceted, making it more realistic to speak of multiple spatial abilities. That is, there are distinctive components involved, and individuals may be differentially competent at each. Recent research has shown that performance at some spatially demanding tasks can be improved by training and other interventions (e.g., Newcombe and Frick 2010).

A taxonomy adopted by cognitive psychologists (Newcombe and Shipley 2014) is based on two conceptual divisions. The first is between concepts concerning the internal structure of objects and those concerning objects’ locations in the world. The second is between static representations of such structure and position, and their dynamic nature and representation. This suggests four broad categories of spatial skills, which may be extended to concepts: intrinsic-static, intrinsic-dynamic, extrinsic-static, and extrinsic-dynamic (Figure 1). It should be noted that many spatial concepts do not fit neatly within a single category. For example, clusters are defined by spatial relationships within a reference frame, and have identity as objects unto themselves, with structural characteristics such as density.

Developing a universal taxonomy of spatial concepts has proven to be difficult, as spatial concepts seem to defy narrow classifications. This suggests multiple taxonomies will be useful, for different purposes. This entry does not attempt a strict ordering; rather, it enumerates a number of the fundamental spatial concepts in narrative fashion within the broad thematic groupings that follow.
Divisions of size and scale

The term “scale” has several related but distinct meanings. As a noun, it refers to a particular range of sizes or magnitudes. In an adjective phrase, scale is often used to mean relative or absolute size on one such scale, for example “large-scale” or “human scale.” The same entity may be considered large or small on different scales; even large ponds are small-scale water bodies.

When we say that each scientific, engineering, and artistic field is concerned with phenomena at some range of scales, we mean in some cases absolute size on the scale of all things, and in others the projective psychological scales of perception and representation (Montello 1993). Half a county can be viewed in the frame of an airplane window, and a two-dimensional projection of the entire Earth represented at a very small scale (a high ratio, or small representative fraction) to fit on a single sheet of paper. The “landscape” of a molecule can be scaled by a microscope to fill our field of vision.

There are disciplinary associations with particular scales, and disciplinary subdivisions can be scale-based. For example, distinct branches of physics concern phenomena at subatomic,
atomic, molecular, and astronomical scales. Chemistry concerns the dynamics of matter, principally at atomic and molecular scales. Industrial designers create tools and other mechanical devices at scales ranging from handheld objects to large vehicles. Architects design at the scale of buildings – single structures, ensembles of buildings, and the spaces within and between them. Many fine arts, including painting, sculpture, and dance, are intrinsically spatial, producing visual and/or kinetic forms; artistic scales in human terms can range from the figural (smaller than the human body) to vista (larger than the body but apprehended from a single location).

At the next larger scales we find geographers and planners, seeking scientific explanation and designing settlements and settlement systems in environmental space, which “is usually thought to require the integration of information over significant periods of time,” and geographical space, which “cannot be apprehended directly through locomotion; rather it must be learned via symbolic representations such as maps or models …” (Montello 1993, 315).

Fundamental spatial concepts

Given that no universally accepted taxonomy of spatial concepts has been developed to date, much less a formal ontology, the following geographically oriented groupings are presented. Definitional statements have been synthesized from geographic textbooks and dictionaries.

Spatial primitives

The most basic of spatial concepts are studied in classical physics, topology, and mereology (part–whole relations). These are primitives in a sense, from which we derive more complex concepts and principles. As noted earlier, many primitive concepts are used metaphorically in reasoning about spatiality as well.

Objects (bodies) in the universe are constituted by matter, giving them mass. Objects thus have physical extension and time-indexed spatial location. They are bounded – that is, there is a division between object and not-object – and they may be composed of smaller objects or have features considered as parts. Objects are subject to external forces, including from gravitational fields, and can move in space. That is, they change position with respect to each other or to a reference point, according to laws first described by Isaac Newton.

The relative position of objects (and their abstract representations as points, lines, and areas) can be described in the primitive non-metric terms of topology. These include concepts of containment and connection. Relevant cognate terms include adjacency and neighborhood.

Location

Arguably, the most fundamental of spatial concepts for geographers is location, and like many spatial terms, its definition is problematic. The terms location, position, and place are often used interchangeably, along with site and locale. They are “near-synonyms,” as each commonly appears in definitions of others. Each can be viewed as one kind of answer to the question, “where?” Geography seeks to explain the why of where, which requires first establishing location, as well as is practicable, by means of spatial descriptions. Geospatial data describe at minimum the location of some phenomenon and its class or category. As noted earlier, whereas all entities with spatial extent also have temporal extent, many geospatial data sets are either encompassed within a single temporal frame or divided into time-indexed “snapshots.”
SPATIAL CONCEPTS

Geographic locations in this broad sense can be represented mathematically in the so-called absolute terms of coordinates on a grid of two to four dimensions \((x, y, z, t)\), where \(t\) is time) having an arbitrary origin, or in relative terms of distance and orientation (or bearing) related to other locations. Relative locations are also frequently described in topological or otherwise nonmetric terms, including vague natural language such as “just north of.” Strictly speaking all locations are relative, whether to coordinates on an abstract grid and its possibly arbitrary origin, or to other entities. The reference origin of geographic coordinate systems is an estimated center of the Earth.

The location, position, site, or locale of a thing might be represented solely in mathematical terms or with additional attributes deemed essential to a particular investigator or perspective. For example, site and locale suggest important containment or environmental relationships.

In the most general terms, a place is a location about which we have something to say (Haggett 2001). However, the qualities of a place are understood by many geographers to include not only metrical location, but also (or alternatively) individual or collective human experiences there (Tuan 1977).

Distance

Euclidean distance is essential to metric descriptions of location and is thus a first-order concept one step removed from that primitive. Distances are the essential measure of spatial distributions (patterns), and therefore the basis for identification of clusters, spatial association, and centrality.

An important related concept for geographers, cost distance, is another example of the mutability of spatial concepts and their important value in reasoning more generally. Geographers have noted for many decades that metric distance is but one way of describing how far one thing is from another, effectively. The 12 miles between two locations across a plain and the 12 miles traversing a mountain have radically different meaning in human and economic terms of cost in physical effort and in elapsed time. Janelle (1968) developed the concept of time–space convergence to reflect the shrinking effect that transportation and communication technologies have upon the effective or functional distances of commerce and everyday life.

Representation

Our reasoning about spatial phenomena, and spatial concepts like their absolute or relative location, involves representation in several respects. When converting directly or remotely sensed data of counts and measurements into the mental, mathematical, physical, and graphical representations we call geographic information, we are reasoning about representations of concepts and not the phenomena themselves.

Two distinctive conceptual representations of space – those of objects and fields – inform representational strategies and methods. In simplest terms, the object view of space is that of “an inert container populated by objects” whereas the field view posits, “a ‘plenum’ characterized by a ubiquitous field” (Couclelis 1992, 70). These are analogous to a similar division in physics, and correspond to choices of vector and raster data formats in a GIS.

Several concepts pertaining to spatial representation derive from the object-field division and inform the digital maps geographers and cartographers make. We most commonly model physical objects geometrically in terms of points, lines, areas, and volumes. Fields are most often represented as matrices of uniformly sized rectangular cells corresponding to a particular spatial resolution, each with a single value for
a given attribute. Fields can also be derived from point data by means of spatial interpolation and represented in vector form as, for example, triangulated irregular networks (TINs).

One of the most important representational techniques for viewing and analyzing spatial data involves the overlay of spatially aligned datasets in order to merge or compare them visually and computationally.

Pattern and process; form and function

A cornerstone activity of both physical and human geographers is the identification of patterns of spatial distribution and their explanation in terms of the underlying processes that produce them. An analogous activity, more particular perhaps to physical geography, is the description of spatial form and its explanation in terms of processes, function, or human purpose. Spatial patterns concern the extrinsic spatial relations between objects of interest, including events as commonly understood (e.g., crime or earthquakes). In contrast, spatial form concerns intrinsic properties of objects and entities considered as objects, including indistinctly bounded features of a larger object (e.g., a mountain and the Earth), or collections of component objects themselves (e.g., wells in an oil field, or streams in a watershed). This flexible and scale-dependent conception of objects means the division between form and pattern is not crisp, as we have seen in several imprecise classifications discussed already.

A number of fundamental spatial concepts are common to the discovery and description of form and patterns; some of these have been mentioned previously. The form of a geographic object or feature is described in terms of its size and shape, as delimited by its boundary, and by its structure – the number and arrangement of component parts or features. Many geographic phenomena have a fractal nature. That is, their structure is self-similar at any scale of observation. We see this when viewing higher and higher magnifications of crystals, for example. Clouds, river networks, and coastlines are said to have fractal qualities.

Patterns in the spatial distribution of geographic phenomena can be reified as the human-defined analytical objects, regions and clusters. The concept of region is normally defined as an area having one or more characteristics distinguishing it from surrounding areas; however, regions can be purely spatial (e.g., northern) and are not always geographic (e.g., anterior). The identification of particular geographic regions can be controversial because their defining criteria – characteristics considered and quantitative thresholds – are often subjective. Like objects, regions are bounded, although regional bounds are more often indistinct, or fuzzy.

Clusters are spatial distributions defined by an abnormal concentration (i.e., high density) of some phenomenon within a study area. Although clusters are often abstracted to aggregations of points, the phenomena can be large geographic features as well, such as a cluster of islands. Clusters may define regions: for example, Tornado Alley is an informally bounded region of the United States where tornados have historically occurred most frequently.

Spatial context

Observations of natural phenomena occur within a reference frame, which forms the context for their description and analysis. The concepts of scale and granularity are closely related to reference frames. For example, what constitutes a cluster at one scale and within a particular reference frame can appear as an even or random distribution in another. The well-known modifiable areal unit problem arises from the
SPATIAL CONCEPTS

often times arbitrary choice of bounds used for aggregating point data.

Objects of study – whether things or occurrences – are almost always impacted by their surroundings, that is, their neighborhood, setting, or environment. Elements of relevant spatial context can include adjacent and nearby things, network connections at any distance, the bounds of a study area, and areal divisions within it. All are important factors influencing the results and interpretations of scientific studies.

Tobler’s “First Law of Geography” (1970) asserts that attributes of things that are near each other tend to be more similar than attributes of things that are far apart. Such similarity is assessed and confirmed by measures of spatial dependence and spatial autocorrelation. Because location and proximity are often omitted from statistical models, assumptions of independence for observational data are in many cases a fallacy.

Networks, connection, and interaction

Networks are the essential structural form of many human artifacts, from roads and railway systems to utilities infrastructure such as pipelines and power grids. Network structure appears throughout nature at many scales, for example in watersheds, circulatory systems, proteins, lightning, and neurons. Abstract network representations consisting of nodes and edges are an invaluable method of representing connectivity and interaction of all kinds. The analysis of social and professional networks has become a significant methodology in fields ranging from the history of science, to sociology, to business marketing. Nodes might correspond to individuals, groups, and publications or other products. Edges can correspond to friendship, co-authorship, affiliation, or production.

Spatial relations in all networks are at minimum topological ones of connectedness. If a network is geographically embedded, its nodes correspond to geographic locations. Its edges may as well, but with more or less resolution. That is, in some cases the geographical paths between nodes are either unknown or incidental. A common example of this is the representation of flows, where the attributes of interest are the source, target, class of substance, commodity, or activity and its aggregated magnitude for some period.

Many generalized spatial measures are relevant to all networks, whether geographically embedded and physical, or entirely abstract. These include size, density, connectedness, clustering coefficient, and node or edge centrality.

Spatial dynamics

Patterns of spatial distributions and of connectivity are intrinsically dynamic. They are the product of processes that transpire over time and their properties are time-dependent. The qualities, magnitude, and identity of many things in the world are in continual flux, and so a significant proportion of our scientific observations, measurements, and analyses seek to explain spatial change. Physical objects and Earth features change form, position, and orientation. Their identity can be changed by splitting and merging events, and by changes to essential attributes. A prime example is the changing shape, size, and nature of geographic features at the Earth’s poles, as influenced by climate. These types of spatial change extend to non-physical geographic features such as cities and countries, as well as to regions, which can be defined by any number of physical and social variables.

Concepts involving the spatial dynamics of movement are relevant in many fields and at most scales. Many have precise meanings in physics and chemistry, and alternate but similar or metaphorical meanings in other fields. In
physics, diffusion refers to a heat exchange process where a high concentration of a finite number of particles spreads throughout a solution in a random walk motion. In geography, diffusion can refer to the spread of a concept or practice from one or more locations to many more, but with some distinct differences: the paths taken by individuals are unlikely to be deemed random, and processes like dispersion and migration do not always entail some finite quantity dispersing elsewhere. Rather, geographic diffusion suggests probabilistic or deterministic dispersion and increasing magnitude. The term “flow” in the physical sciences refers to the continuous movement of matter (normally fluids) in a stream-like fashion. It is also used routinely and metaphorically in geography and many other fields in reference to nonmaterial things like ideas, and to nonfluids, such as commodities and currency in trade activity.

SEE ALSO: Geodesign; Ontology: theoretical perspectives; Place; Representation; Scale; Space; Spatial thinking, cognition, and learning; Spatiotemporal analysis; Topological relations

References


Further reading


Spatial context

Sara McLafferty
University of Illinois at Urbana–Champaign, USA

Spatial context refers to the geographic setting in which events and processes occur. It comprises nearby places and the environmental, social, and geographical characteristics of those places. For a person, the relevant spatial context might include the person’s home and the surrounding neighborhood, along with locations that are important in daily life, such as the locations of work, school, friends, and family. The spatial context for a city might include the state or country in which the city is located, nearby towns and suburbs, and natural features in the neighboring area. Thus, spatial context is a relational concept that reflects the interactions between an entity and its nearby environment. The term “spatial context” is often used interchangeably with terms like “geographical context,” “neighborhood context,” and “place;” however, terms like neighborhood context and place include an important perceptual and experiential dimension – how people experience their local settings. In contrast, spatial context emphasizes the spatial properties of the setting, including its geographic size and relative location, and the proximity and density of environmental characteristics.

Three elements make up the spatial context for a particular topic: (i) the entity – the person, organism, location, or object of interest; (ii) the geographic setting or environment in which the entity is situated; and (iii) a set of rules or processes that govern the relationships between entity and environment. In defining a person’s spatial context with respect to health-care access, the entity is the person, the relevant environment comprises the locations of health-care providers, and the processes linking the two consider issues such as the person’s ability and willingness to travel, access to transportation, and providers’ referral networks, services offered, and outreach efforts. These processes determine the size and geographical characteristics of the spatial context.

Interest in spatial context first appeared in geography in the mid-1960s, when economic and political geographers like Julian Wolpert and Kevin Cox investigated how the spatial settings in which people live influenced their decision-making on questions like which crops to plant and how to vote. Interest in spatial context also emerged among behavioral geographers whose work examined the important question of how the spatial distributions of retail and service opportunities affect people’s decisions about where to travel for everyday activities. At the same time, urban and transportation geographers began to analyze spatial context through research on the spatial and temporal characteristics of activity spaces, the spaces in which people’s daily activities unfold. However, these early efforts were constrained by limited computing resources and difficulties in managing and analyzing the complex geospatial data needed for measuring spatial context.

These barriers began to fall during the 1980s, when geographic information systems (GIS) provided, for the first time, tools for rapidly measuring and analyzing spatial contexts and for categorizing and updating their social and environmental characteristics. A host of innovative spatial–contextual methods and approaches grew out of the GIS revolution, and their development....
SPATIAL CONTEXT

continues apace. At the same time, researchers in fields spanning the natural, social, behavioral, and health sciences are embracing the concept of spatial context and recognizing its significance, conducting empirical research on the implications of spatial context for issues ranging from animal behavior to geopolitical conflict to obesity. The remainder of this entry emphasizes advances in spatial context methods, models, and applications since the 1980s.

Identifying spatial contexts

A wide range of methods have been developed for identifying and estimating spatial contexts and most methods can readily be implemented using geospatial data and GIS. Given a set of one or more discrete entities – for example, people, locations, residences, or zones – the goal is to identify a spatial context for each entity. Typically, the spatial context is estimated as a discrete zone extending around the entity; however, alternative approaches have also been developed.

One of the simplest measures of context is spatial membership: location inside or outside a geographic zone. For example, location within a specific kind of climate zone might affect the presence of particular animal or plant species; similarly, living in a particular state in the United States influences access to benefits from the Medicaid program. Determining spatial membership is straightforward in GIS using point-in-polygon operations.

Another class of methods focuses on contiguity – the topological relationships between spatial objects. Spatial context for a zone can be defined as the set of neighboring zones, zones that share a common border, a principle known as rook contiguity (Figure 1a). To include diagonal neighbors, that is, zones that share a common point, the queen contiguity rule is used. In some cases, the spatial context that is relevant for a zone may extend beyond adjacent zones, especially in places where the zones are small in size. These larger areas of influence can be modeled by using higher-order definitions of contiguity: second-order contiguity encompasses all zones that are contiguous to a first-order contiguous zone (Figure 1b). As the order of contiguity increases, the geographic size of the spatial context increases, implying that the geographical setting or zone of influence for a place extends over a wide area. Defining spatial context based on contiguity is straightforward when the spatial data of interest are recorded for zones (polygons). If the data are recorded for points, Theissen polygons can be used to convert the data from point to polygon format, and then identify spatial contexts based on polygon contiguity.

Distance-based methods are widely used in identifying spatial contexts. The spatial context for an entity is defined as all places located within a particular maximum buffer distance of the entity. For example, a one-kilometer buffer distance from an individual’s home is often used to define the spatial context that can be reached by walking (Figure 1c). Although many studies have relied on Euclidean distance for measuring spatial context, a growing trend is the use of network buffers, which involves computing distance or travel time along a transportation network. Spatial contexts identified this way are irregularly shaped, following the contours of the transportation network. More complex metrics that incorporate transportation access and cost have also been developed.

A crucial issue in using distance-based methods to identify spatial context is the choice of buffer distance. By determining the radius or maximal extent of the modeled spatial context, buffer distance critically affects the results of spatial contextual research investigations. Studies show significant changes in model results as the buffer distance varies (Spielman and Yoo 2009).
Moreover, researchers typically use the same buffer distance for an entire study area or study population; however, the relevant spatial context is likely to differ among entities based on their social, geographical, and economic positions. For example, evidence suggests that the distances and times people are willing and able to travel vary greatly based on social characteristics such as age, gender, income, and rural/urban location.

Distance and contiguity methods represent the spatial context for an entity as a discrete, bounded zone: places are designated as either inside or outside the spatial context and only places within are important. In reality, most spatial contexts have fuzzy boundaries that reflect varying influences across space. Geographers have turned to field-based methods to represent the fuzziness of spatial contexts. Kwan (2000) generated spatial contexts for individual people via kernel density estimation of their daily activity locations (Figure 2). Spatial context is depicted as a continuous surface of varying intensity, in which peaks represent locations where daily activities are concentrated and valleys areas with few or no daily activities. Similar approaches have been used in representing animal’s home ranges – their spatial contexts – based on locations recorded by GPS monitoring devices. Field representations of context have also been important in modeling the social and environmental characteristics of spatial contexts, such as: ethnic/racial segregation; spatial accessibility of jobs, services, and resources; and environmental quality. Representing these characteristics as continuous fields enables examination of how the intensity of risk, access, and exposure vary within spatial contexts.

There is growing interest in space–time contexts, contexts that change dynamically over time. Such contexts are critically relevant for representing the environments that people experience as they move about in everyday life and
throughout their life course. The uncertain geographic context problem highlights the fact that spatial contexts are uncertain, reflecting myriad individual decisions and contextual influences, and that contexts change dynamically over time (Kwan 2012), with critical implications for research design and research findings. Interest in space–time contexts dates back to the work of Hägerstrand on time geography, and Hägerstrand’s space–time prism concept has been used to represent contexts as they unfold over time (Miller 1991). Activity locations, and the pathways connecting them, comprise the space–time context. Similar approaches are employed by researchers studying environmental influences over the life course based on individual migration data. For all of these space–time approaches, it is important to incorporate accurate data about the environment at each relevant time and place – a challenging task.

Today, GPS–enabled cell phones and other devices can be used to track individual movements in real time, providing highly accurate representations of space–time context that can be used in studying topics like exposure to environmental hazards, access to healthy/unhealthy food environments, and risk and protective environments for substance use (Zenk et al.}
These technologies are quickly decreasing the need to model spatial and space–time contexts, providing a strong empirical foundation for retrospective contextual representation. However, privacy and confidentiality restrictions and the high cost of data acquisition often limit application of these approaches to relatively small samples, and methods for analyzing the vast and complex “big data” obtained from GPS–enabled devices are still in their infancy. Still, analyzing these “experienced” space–time contexts presents tremendous opportunities for research in areas ranging from health sciences to ecosystem dynamics to environmental risk and vulnerability.

Spatial context, spatial analysis, and modeling

The notion of spatial context lies at the heart of many recent advances in spatial analysis and modeling. Developments in local spatial statistical analysis and modeling rest on the fact that spatial context matters (Lloyd 2011). Local spatial analysis methods involve the calculation of statistics or models at each location (entity) within an area, and the computations are made within the entity’s spatial context. The local Moran’s I statistic for location j, for example, represents the well-known Moran’s I statistic computed within the local spatial context of j. Location j’s spatial context is modeled by a vector of spatial weights that define j’s proximity to other locations. Spatial weights matrices express contextual relationships among all locations within a study area, and these matrices are foundational for all local spatial analysis methods, including spatial and spatiotemporal hotspot detection, spatial filtering and kernel density estimation, and geographically weighted regression. Development and application of these methods grow out of the key recognition that processes vary over space in response to environmental, social, and geographical characteristics of local spatial contexts. Efforts at downscaling global climate models, for example, clearly acknowledge the significance of spatial context.

The multiscalar dimensions of spatial contexts are also being incorporated in spatial analysis and modeling. Multilevel models represent contextual influences that exist at multiple geographical scales, such as the household, census tract, state, and country. In hierarchical models, these scales are neatly nested, so that each lower-level unit falls completely within a corresponding upper-level unit. Efforts are underway to develop methods for modeling non-hierarchical, multilevel, spatial contextual effects, although overlap of spatial effects across levels poses conceptual and methodological challenges. Spatial multilevel models – those that incorporate spatial associations beyond those represented by hierarchical spatial units – are also attracting attention.

Development of context-aware models reflects heightened awareness of spatial context among spatial analysts and modelers. Context-aware models involve constant updating of model parameters and outcomes based on changing conditions at a particular time and place. Methods like agent-based modeling enable researchers to represent the dynamic relationships between individuals and their local contexts, with contextual characteristics affecting behaviors and those behaviors simultaneously affecting spatial contexts. These methods have been used effectively in modeling the spatiotemporal spread of disease, impacts of habitat change on endangered species, and other complex, dynamic phenomena.

An important, but less studied implication of spatial context for spatial analysis and modeling is its constraining effect on model results. For many empirical research studies, the spatial configuration of environmental and social factors
forms a template within which data collection and sampling take place. That template limits the kinds of relationships and patterns that can be observed. For example, in a city in which the residential locations of high- and low-income groups are evenly intermixed, it is impossible to locate services to geographically advantage one group over the other: inequity cannot be observed in such a spatial context. This point was recognized in the 1970s and 1980s, when behavioral geographers observed that people’s revealed spatial preferences are influenced by the spatial configuration of opportunities and when empirical studies demonstrated the effect of this spatial constraint on measures of statistical association. Recent work suggests that the topological properties of spatial contexts are related to the properties of spatial models and tests (Farber, Paez and Volz 2009).

Impacts of spatial context on social and environmental processes

The past two decades have seen an enormous expansion of empirical research on spatial contextual effects. A wide range of phenomena are influenced by spatial context, including social outcomes such as employment and incomes, crime, and health and well-being, as well as environmental phenomena such as animal behaviors, natural hazards, and vegetation types and distributions. Because all living things depend on the environment for their survival, all are strongly influenced by spatial context. The diversity of topics makes it difficult to summarize empirical findings on contextual effects. Evidence shows, for example, that attributes of geographical and ecological context strongly influence species–area relationships at multiple scales (Drakare, Lennon, and Hillebrand 2006). Studies of people’s employment outcomes in cities indicate that unemployment and incomes are strongly shaped by the local availability of jobs, while at the same time reflecting individual and household characteristics such as gender, age, and ethnicity/race (Preston and McLafferty 1999). Research in these diverse fields emphasizes that the influence of spatial context differs from place to place, at multiple scales, and among entities (e.g., people, species), reflecting the dynamic connections between them.

In understanding how spatial context affects a particular outcome variable, it is essential to think through the processes that link context and outcome, so that context can be modeled appropriately. For human health studies, at least three processes are thought to connect outcomes and spatial contexts: exposure to environmental agents, such as pathogens or toxins in air and water; access to local economic and social resources, including parks and recreational facilities, stores offering healthy foods, job opportunities, and social support; and access to health-care services (Macintyre, Ellaway and Cummins 2002). Each of these processes has different implications for defining spatial context: analyzing environmental exposure requires detailed data on people’s indoor and outdoor locations and the intensity of environmental agents at those times and places. Researchers are increasingly using mobile environmental monitoring devices and dense sensor networks to record the spatial contexts of people’s environmental exposures. In contrast, access to health care is affected by human decision-making and mobility, availability of services, and political and economic forces that affect the array of services available and people’s abilities to use them. In this case, researchers model spatial context based on potential access to health care and incorporate important interpersonal variations.

The dynamic and relational qualities of spatial contexts present challenges for empirical
research. Contexts affect entities, but entities simultaneously affect contexts, making it difficult to disentangle their effects. Moreover, people and animals actively select contexts based on personal preferences, resulting in the important issue of selection bias. Although methodologies exist for addressing selection bias, these methods may not be feasible for particular kinds of research problems. Finally, even when selection into particular contexts is not based on preferences, there are often strong social or environmental forces that channel people into particular contexts, sharply limiting the opportunity to observe all types of people in all geographic settings. In the United States, for example, racial residential segregation implies that African Americans only live in a very limited subset of residential contexts, which confounds our ability to distinguish the independent effects of race and segregation.

In summary, it is widely recognized that spatial context matters in fields spanning the social and environmental sciences. At the same time, developments in GIS, spatial analysis methods, and geospatial technologies have greatly advanced the ability to model spatial contexts and explore their complex effects. Already, fundamental shifts in spatial contextual analysis are being witnessed, including the increased monitoring of contexts in real time using GPS-enabled mobile devices and environmental sensing networks, and heightened attention to methodological challenges, such as the uncertain geographic context problem, selection bias, and the constraining effects of spatial contexts on model estimates. Understanding how spatial contexts both constrain and enable lives and livelihoods, and how those, in turn, shape spatial contexts, presents exciting opportunities for research across the social and environmental sciences.

SEE ALSO: Agent-based modeling; Buffers; Local statistics and place-based analysis; Neighborhood; Spatial analysis; Spatial filtering; Spatial organization and structure; Spatial weights; Spatiotemporal analysis; Uncertain geographic context problem

References


7
SPATIAL CONTEXT


Spatial crowdsourcing

Cyrus Shahabi
University of Southern California, USA

Smartphones are ubiquitous: we are witnessing an astonishing growth in mobile phone subscription that surpassed 4.35 billion worldwide at the end of 2009 (ABI Research 2010) and reached 6 billion in 2011, which is 87% of the world population. By 2013, Gartner, the world’s leading information technology research and advisory company, predicted that mobile phones would overtake PCs as the most common web access device worldwide. Meanwhile, the mobile phone’s bandwidth is constantly increasing: from 2.5G (up to 384 kbps) to 3G (up to 14.7 Mbps) and recently 4G (up to 100 Mbps) (Sauter 2011). Hence, the multiplication of the above two factors plus the constant progress and increase of smartphone’s sensors (e.g., video cameras) suggest an exponential growth in data collection and sharing by smartphones. That is, every person with a mobile phone can now act as a multimodal sensor collecting and sharing various types of high-fidelity spatiotemporal data instantaneously (e.g., picture, video, audio, location, time, speed, direction, and acceleration).

Exploiting this large volume of potential users and their movability, a new mechanism for efficient and scalable data collection has emerged: spatial crowdsourcing (Kazemi and Shahabi 2012). Spatial crowdsourcing requires workers (e.g., willing individuals) to perform a set of tasks by physically traveling to certain locations at particular times. Spatial crowdsourcing has applications in numerous domains such as journalism, tourism, intelligence, disaster response, environmental monitoring, and urban planning. To illustrate, consider a citizen-journalism scenario depicted in Figure 1, where a news agency (i.e., requester) is interested in collecting pictures and videos of demonstration areas from various locations of a city. With spatial crowdsourcing, the requester issues a query to a spatial crowdsourcing server (SC-server). Consequently, the SC-server distributes the query among the available workers in the vicinity of the events. Once the workers document their events with their mobile phones, the results are sent back to the requester.

Taxonomy

To start, a taxonomy is defined for spatial crowdsourcing (Figure 2). First, spatial crowdsourcing is classified based on workers’ motivation. Next, two modes of task publishing in spatial crowdsourcing are defined. Finally, the workers are classified into two groups based on whether or not they have constraints.

A major challenge in any crowdsourcing system is how to motivate people to participate. Four levels of worker motivation have been identified by Quinn and Bederson (2011), including pay, altruism, fun and implicit. To simplify, spatial crowdsourcing can be classified based on the motivation of the workers into two classes: reward-based and self-incentivized (Figure 2). With reward-based spatial crowdsourcing, every spatial task has a price (assigned by a requester) and workers will receive a certain reward for every spatial task they perform correctly. Examples of this class include Field Agent (www.fieldagent.
Figure 1  Ten locations of disasters. Ten tasks to be assigned (t1–t10) and three available workers in the immediate vicinity (w1–w3).

SPATIAL CROWDSOURCING

net/) and Gigwalk (http://gigwalk.com/). With self-incentivized spatial crowdsourcing, workers volunteer to perform the tasks or usually have other incentives rather than receiving a reward, such as documenting an event or promoting their cultural, political, or religious views. An example of this class is the University of California, Berkeley, traffic project (http://traffic.berkeley.edu/), in which more than 5000 users have voluntarily installed traffic software onto their phones and report traffic information.

Next, two task publishing modes in spatial crowdsourcing are defined: worker selected tasks (WST) and server assigned tasks (SAT). With the WST mode, the SC-server publishes the spatial tasks and online workers can choose any spatial task in their vicinity without the need to coordinate with the server. One advantage of the WST mode is that the workers do not need to reveal their locations to the SC-server since they can choose any arbitrary task in their vicinity autonomously. However, one drawback of this mode is that the server does not have any control over the allocation of spatial tasks. This may result in some spatial tasks never being assigned, while others are assigned redundantly. Another drawback of WST is that workers choose tasks based on their own objectives (e.g., choosing the k closest spatial tasks to minimize their travel cost), which may not result in a globally optimal
Spatial Crowdsourcing

Figure 2 A taxonomy of spatial crowdsourcing.

Assignment. An example of the WST mode is provided by Alt et al. (2010), in which the workers browse for available spatial tasks, and pick the ones in their neighborhood.

With the SAT mode, the SC-server does not publish the spatial tasks to the workers. Instead, any online worker sends his location to the SC-server. The SC-server, after receiving the locations of all online workers, assigns to every worker his nearby tasks. The advantage of SAT is that, unlike WST, the SC-server has the big picture and, therefore, can assign to every worker his or her nearby tasks while maximizing the overall task assignment. However, the drawback is that the workers should report their locations to the server for every assignment, which can pose a privacy threat. Examples of this mode of spatial crowdsourcing are provided by Kazemi and Shahabi (2012) and Kazemi and Shahabi (2011). In Kazemi and Shahabi (2011), a framework for small campaigns is proposed, where workers are assigned to their nearby sensing tasks.

Finally, in the case of SAT, we divide the workers into two groups based on whether or not they have constraints. With workers without constraints, the server has full flexibility on how tasks should be assigned to the workers. This means that workers only send their locations to the server, and the server assigns every spatial task to its nearby worker (Kazemi and Shahabi 2011). With workers with constraints, the server needs to satisfy the constraints while assigning the tasks. An example of spatial constraint is that every worker only accepts spatial tasks in a spatial region (i.e., his working region).

Task assignment

The main challenges of spatial crowdsourcing are due to the large-scale, ad hoc, and dynamic nature of the workers and tasks. First, to continuously match thousands of spatial crowdsourcing campaigns, where each campaign consists of many spatiotemporal tasks, with millions of workers, an SC-server must be able to run efficient task assignment strategies that can scale. Second, the task assignment must be performed frequently and in real time as new tasks and workers become available or as tasks are completed (or expire) and workers leave the system. Third, while in small campaigns the workers may be known and trusted, with spatial crowdsourcing the workers cannot always be trusted. In fact, some skeptics of crowdsourcing go as far as calling it a “garbage in, garbage out” system due to the issue of trust. Finally, individuals with mobile devices need to physically travel to the specified locations of interest. An adversary with access to these individuals’ whereabouts can infer sensitive details about a person (e.g., health status and political views); thus, protecting worker location privacy is an important concern in spatial crowdsourcing.

In the following it is shown that the combination of the abovementioned challenges renders the spatial crowdsourcing problem unique, where
SPATIAL CROWDSOURCING

the solutions proposed for similar problems cannot thoroughly address its task assignment problem.

The first class of similar problems is called volunteered geographic information (VGI) with the goal of generating geographical information provided voluntarily by individuals. Most real-world examples of VGI are limited to adding contents to a prebuilt map such as Wikimapia (www.wikimapia.org/), Google Map Maker (http://www.google.com/mapmaker/), or Open Street Map (www.openstreetmap.org/). As opposed to spatial crowdsourcing, with VGI applications there is no need for the worker to be present at any specific location in order to perform the task (even though being at a location can still improve the accuracy of the performed tasks; for example, in disaster regions the Red Cross may use Open Street Map because it is far more accurate than Google maps since people voluntarily report broken streets or bridges). As the name suggests, VGI is more self-incentivized, whereas in spatial crowdsourcing workers are assigned tasks by the SC-server.

Another class of similar problems, called participatory sensing, exploits mobile users to collect and share data using their sensor-equipped phones for a given campaign. Some applications of participatory sensing in the real world include traffic monitoring, urban planning, disaster response, and urban air pollution. Most studies on participatory sensing focus on small campaigns with a limited number of workers. However, with spatial crowdsourcing the focus is on devising a scalable, generic, and multi-purpose crowdsourcing framework, similar to Amazon Mechanical Turk, but spatial, where multiple campaigns can be handled simultaneously. Therefore, the main challenge with spatial crowdsourcing is to devise an efficient approach to assign tasks to workers given the large scale and dynamism of the environment.

Some recent studies in spatial matching do focus on efficiency and use the spatial features of the objects for more efficient assignment. Spatial matching is a one-to-one (or in some cases one-to-many) assignment between objects of two sets where the goal is to optimize over some aggregate (sum, max, etc.) function of the distance between matched objects. These studies assume that a global knowledge about the locations of all objects exists a priori and the challenge comes from the complexity of spatial matching. However, spatial crowdsourcing differs due to the dynamism of tasks and workers (i.e., tasks and workers come and go without our knowledge), and so the challenge is to perform the task assignment at a given instance in time with the goal of global optimization across all instances. Moreover, the fact that workers need to travel to task locations causes the landscape of the problem to change constantly. This will add another layer of dynamism to spatial crowdsourcing that renders it a unique problem.

One can consider the task assignment problem in spatial crowdsourcing as a matching problem between tasks and workers, which makes it similar to the classic matching problem. In particular, the online matching problem is the most relevant variation to spatial crowdsourcing as it captures the dynamism of tasks arriving at different times. However, once the number of tasks assigned to one worker is more than one, the online matching problem cannot capture the true cost of performing tasks. More specifically, the cost for a worker to perform a task mainly corresponds to the time it takes for him or her to travel to the location of the task. Thus the cost of a task is not a fixed value but is dependent on the user’s location before performing the task. The execution of a set of tasks is therefore the distance of the shortest path that starts from the worker’s current location and goes through the locations of all the assigned tasks. On the other hand, with online
matching the overall cost for one worker would be the sum of the distances between the worker and each assigned task.

Modeling the assignment cost as the shortest path visiting the location of multiple tasks brings another class of problems to attention. In this context the assignment problem in spatial crowdsourcing becomes similar to the traveling salesman problem (TSP) and the vehicle routing problem (VRP). The goal of VRP is to minimize the cost of delivering, using a fleet of vehicles, goods located at a central depot to customers who have placed orders for such goods. The online versions of both TSP and VRP have been studied to some extent where new locations to visit are revealed incrementally. Since there is only one salesman in the standard version of TSP, this entry focuses on VRP. Different variations of VRP have been studied, yet there are still some differences between task assignment in spatial crowdsourcing and these variations. In VRP, a server can pay a penalty and deny visiting a location; however, in spatial crowdsourcing the goal is to maximize the number of assigned tasks so the worker does not have the option of denying a task. Furthermore, with VRP, all servers start from the same depot whereas in spatial crowdsourcing each worker can have a different starting location. Moreover, with VRP there are a fixed number of servers whereas in spatial crowdsourcing the same type of dynamism for tasks can apply to the workers. That is, workers can be added (removed) to (from) the system at any time.

Now some preliminary approaches to solve the task assignment problem of spatial crowdsourcing are summarized. The first approach to task assignment problems is to simplify the issue by focusing only on a given instance of time and ignoring the dynamism of the environment. In this case, task assignment can be modeled as a matching problem, which can be reduced to the maximum flow problem (Kazemi and Shahabi 2012). In this study, the constraints of workers have been taken into consideration. Each worker has two constraints, including a spatial region \( R \) and the maximum number of acceptable tasks \( \text{max}T \) (see Figure 3). The spatial region \( R \) is the area represented by a rectangle, in which the worker can accept spatial tasks. The other constraint, \( \text{max}T \), is the maximum number of tasks that the worker is willing to perform at the given time instance. Therefore, the optimization goal is to maximize the number of assigned tasks while conforming to the constraints of the workers, referred to as maximum task assignment instance (MTA instance).

It has been shown (Kazemi and Shahabi 2012) that the MTA-instance problem is reducible to the maximum flow problem. Thereafter, one can use any algorithm that computes the maximum flow to solve the MTA-instance problem. One of the well-known techniques in computing the maximum flow is the Ford–Fulkerson algorithm.

Next, Kazemi and Shahabi (2012) extended the maximum task assignment (MTA) problem by taking into account the time dimension. That is, MTA considers multiple instances of task sets.
Spatial Crowdsourcing

and worker sets over time. Also, each task has an expiration time and thus needs to be completed before the deadline. Therefore, the goal of MTA is to maximize the total number of assigned tasks over all time instances. Solving MTA could be straightforward if the SC-server was clairvoyant, knowing what tasks and workers would arrive or leave in future. Three alternative solutions to the MTA problem have been proposed by Kazemi and Shahabi (2012). The first approach, called greedy (GR), follows the local optimization strategy by maximizing the task assignment at every time instance. The second approach, called least location entropy priority (LLEP), improves the overall task assignment by prioritizing spatial tasks located in worker-sparse areas (i.e., places with low location entropy). The assumption is that spatial tasks are more likely to be performed in the future if they are located in worker-dense areas. In this approach, each task is associated with a certain cost, which is the entropy of the task location. Accordingly, tasks with lower costs have a higher priority, since they have a smaller chance of being completed. Thus, the goal is to assign the maximum number of tasks during every instance of time while the total cost associated to the assigned tasks is the lowest. This problem is reducible to the minimum-cost maximum-flow problem, which can be solved by using the Ford–Fulkerson algorithm to compute the maximum flow and then minimizing the cost by applying linear programming. With spatial crowdsourcing, the travel distance of the workers is also important. The third approach, namely nearest neighbor priority (NNP), incorporates the travel distance of the workers into the task assignment by assigning higher priority to the tasks with lower travel distance. Therefore, a similar solution to that of LLEP but with a different cost function (i.e., travel cost) can be applied to solve this problem.

Trust

One challenge with spatial crowdsourcing is that the human workers are not equally trusted and hence it is important to verify the validity of the results provided by these workers. This problem was studied in Kazemi, Shahabi, and Chen (2013), where every worker was assumed to have a reputation score indicative of the probability that they would perform a task correctly. The concept of a reputation score is similar to any reputation system used by Internet markets such as eBay. Moreover, each spatial task is assumed to have a confidence value (i.e., a necessary criteria set for expected answers) which states that the answer to that spatial task is only accepted if its confidence is higher than a certain threshold. Hence, the problem is to maximize the number of assigned tasks while satisfying the confidence of every task, referred to as the maximum correct task assignment (MCTA) problem.

Figure 4 illustrates an example of a trustworthy spatial crowdsourcing system with a set of spatial tasks \( T = \{ t_1, \ldots, t_{10} \} \) and a set of workers \( W = \{ w_1, w_2, w_3 \} \). The desired level of confidence of the tasks and the reputation scores of the workers are shown in two different tables. An example of an assignment is to assign \( t_2 \) and \( t_3 \) to \( w_1 \), since they have a smaller chance of being completed. Thus, the goal is to assign the maximum number of tasks during every instance of time while the total cost associated to the assigned tasks is the lowest. This problem is reducible to the minimum-cost maximum-flow problem, which can be solved by using the Ford–Fulkerson algorithm to compute the maximum flow and then minimizing the cost by applying linear programming. With spatial crowdsourcing, the travel distance of the workers is also important. The third approach, namely nearest neighbor priority (NNP), incorporates the travel distance of the workers into the task assignment by assigning higher priority to the tasks with lower travel distance. Therefore, a similar solution to that of LLEP but with a different cost function (i.e., travel cost) can be applied to solve this problem.

However, there may be tasks whose confidence values are too high to be matched by the reputation of any single worker. Therefore, Kazemi, Shahabi, and Chen (2013) proposed having tasks performed redundantly by multiple workers. The idea of redundant collection is to increase the reputation score on the content. That is, multiple workers with different reputation scores can solve the same task so that their aggregated score
satisfies the required confidence for the task. Consequently, an exhaustive approach needs to compute the aggregate reputation score (using a typical decision fusion aggregation mechanism, such as voting) for all possible subsets of the workers, which renders the problem NP-hard (by reduction from the three-dimensional matching problem (Kazemi, Shahabi, and Chen 2013)). Therefore, a number of heuristics have been proposed to achieve close to optimal performance but with a cost close to that of a GR approach, by exploiting the spatial properties of the problem space.

**Privacy**

Another challenge with spatial crowdsourcing is to protect worker location privacy while still using an SC-server as a broker to assign tasks to workers. Location privacy has been studied before in the context of location-based services. Several solutions (Gruteser and Grunwald 2003; Ghinita et al. 2008; Mokbel, Chow, and Aref 2006) have been proposed to protect location-based queries, that is, given an individual’s location, find points of interest in the proximity without disclosing the actual coordinates. However, in spatial crowdsourcing, a worker’s location is no longer part of the query, but rather the result of a spatial query around the task location. This makes the problem more challenging, and previous solutions do not offer satisfactory results. In addition, some work considers queries on private locations in the context of outsourced databases, in which the data owner entity and the querying entity trust each other, with protection being offered only against intermediate service provider entities. This scenario does not apply in spatial crowdsourcing since there is no inherent trust relationship between requesters and workers.

A framework for location privacy in spatial crowdsourcing was proposed in To, Ghinita, and Shahabi (2014). With this framework, worker locations are first pooled together by the data owner (i.e., cell service provider, or CSP) and sanitized according to differential privacy (DP) (Qardaji, Yang, and Li 2013; Cormode et al. 2012). Thereafter, the SC-server only has access to the sanitized data. However, using DP techniques introduces two challenges. First, the SC-server must match workers to tasks using noisy data, which requires complex strategies to ensure effective task assignment. Worker location data are sanitized at the CSP using the private spatial decomposition (PSD) approach (Cormode et al. 2012). PSD is a sanitized spatial index, where each index node contains a noisy count of the workers rooted at that node. Particularly, this study devises a mechanism to create a worker PSD by extending the adaptive grid technique (Qardaji, Yang, and Li 2013). On top.
SPATIAL CROWDSOURCING

of the noisy data, to ensure that task assignment has a high success rate, the authors developed analytical models and task assignment strategies that consider task completion rate, worker travel distance, and system overhead. Second, by the nature of the DP protection model, fake entries may need to be created in the worker PSD. Thus the SC-server cannot directly contact workers, not even if pseudonyms are used, as establishing a network connection to an entity would allow the SC-server to learn whether an entry is real or not, and breach privacy. To address this challenge, a geocast mechanism was introduced for the task request dissemination. Geocast is a routing and addressing method, which is used to deliver information to a group of destinations in a network identified by their geographical locations. Once a PSD partition is identified by the analytical model outlined above, the task request is geocast to all the workers within that partition. Geocast introduces system overhead that needs to be carefully considered in the framework design.

Conclusion

Participatory sensing and crowdsourcing have each been shown to have major societal impacts in various areas: from disaster response when organized information collection systems fail, to citizen news dissemination when mainstream media are censored, to community data collection at large scales when centralized data collection approaches are deemed too expensive (e.g., Google map community), to human-reliant data processing tasks used in order to improve query answering in databases when either queries are subjective or information required to answer queries is missing. Spatial crowdsourcing is a natural next step to scale single-campaign, customized participatory sensing to the generic, multipurpose, scaled-up platform of crowdsourcing. Given the increasing popularity of mobile phones, their continuously increasing bandwidth, power, and sensory features, and the consistent drop in their price, the growth of spatial crowdsourcing is inevitable.

With crowdsourcing, there are many obvious social implications, for example, spatial crowdsourcing can be abused to collect illegal information. There are also orthogonal technical challenges that are not discussed here, for example, with “trust,” the quality of the collected data was not discussed – a worker may have a good reputation but his or her sensor may be broken or wrongly configured (e.g., flash turned off) and he or she may collect garbage. Alternatively, one may consider the expertise of the workers in order to assign them tasks. Note that a worker’s expertise is different and orthogonal to his or her reputation. Another interesting extension is to consider the allocation of rewards to tasks to incentivize workers. The challenge is how to assign a reward to each task in order to maximize the number of completed tasks.

In addition, this entry only focused on server-assigned approaches to spatial crowdsourcing where the server has all the information and matches workers to the tasks. Alternatively, we can consider a scenario where the server publishes all its tasks and workers autonomously select the tasks in their vicinity. This approach would result in local optimization of the task assignment problem because each worker acts alone and selfishly without considering the global objective. However, this approach has the advantage of being more privacy friendly as the workers do not need to share their locations with the server.

In sum, spatial crowdsourcing is a fruitful area of research with many challenging problems and a strong connection to real-world application scenarios benefiting from further research.
SEE ALSO: Big data; Volunteered geographic information; Volunteered geographic information: quality assurance

References


Further reading


SPATIAL CROWDSOURCING


Spatial data infrastructures

Max Craglia
European Commission Joint Research Centre, Italy

Background

A spatial data infrastructure (SDI) is an Internet-based facility to search, find, access, and share location-specific information. The term “spatial” is generic and may refer to any space, from the microscale to the cosmos. In practice, however, the term is often used as a synonym of “geographic” or “geospatial,” that is, space that refers to the earth’s surface, or near-surface. A widely used definition of an SDI is the one adopted in President Clinton’s Executive Order 12906 of 1994 that established the National SDI in the United States: a framework of “technology, policies, standards, and human resources necessary to acquire, access, store, distribute, and improve utilization of geospatial data” (Clinton 1994).

The development of SDIs started in the 1990s following the rapid diffusion of geographical information systems (GIS), and the increased availability of spatial data in digital format. The transition from mainframe computers and dedicated workstations (in the 1970s and 1980s) to desk-top personal computers (in the 1990s) made it possible for thousands of organizations across the world, in both public and private sectors, to start using GIS and to develop applications in many areas, such as urban planning and management, land management, environmental assessment, risk management, and logistics. Initially, each organization had to acquire hardware, software, and data either by digitizing paper maps or by buying digital data from third parties. Most of the use of these data was internal to each organization. As the use of the Internet also spread rapidly in the 1990s, it became possible to publish and share spatial data through this medium. This was a significant step forward but also exposed the limitations of uncoordinated developments in which each organization developed its own applications without consideration of the interoperability of systems or data, making reuse by other organizations very difficult and costly. SDIs were initially developed to address these limitations and to provide a framework for a more coordinated sharing of spatial data based on agreed standards.

Over the past 25 years, most countries in the world have embarked on the development of SDIs, initially driven by the public sector at a national level, but increasingly also at subnational and local levels. The private sector is also a key player, not only as provider of hardware and software (and sometimes data) but also as a user of SDIs and GIS to develop services and applications for both private and public sector clients, in what is often referred to as the location-based industry (or location-based services).

Key components

Most SDIs have developed along similar lines and have the following components.

- **Metadata**, which are structured descriptions of the datasets available through the SDI.
Typically, this also involves the adoption of an existing metadata standard or the definition of an agreed upon way to document available resources.

- A catalog of the metadata records and a set of Internet services used to search, discover, view, and access the data, which are held not in a centralized data store but in a distributed network of data providers. In the United States’ National SDI, this set of services is called a clearinghouse. In other SDIs, this term is not used but the essence is still the same: support for users to search and find what they are looking for, and for data providers to publish their data and related descriptions. Again standards are implemented to ensure that the services can be used effectively across multiple technological platforms.

- Core data sets or framework data are frequently used across multiple applications, and are prioritized in the SDI to be documented and made accessible. The United States’ National SDI, for example, identified seven framework themes: digital orthoimagery, geodetic control, elevation, transportation, hydrology, government boundaries, and cadastral or land ownership information. The European SDI discussed later is much more ambitious with 34 core data themes.

- A set of policies must govern data access and reuse. This is a key component of any SDI because the heterogeneity of policies related to data access and reuse has been, and often continues to be, a key stumbling block to users. In a few countries data can be used with no restrictions at all, in some it can be used with few restrictions (e.g., for non-commercial purposes), whilst in others data must be purchased at high cost and reuse is severely restricted. Compounding the problem of different policies across countries, and even organizations in the same country, is the fact that policies are sometimes not clear, discriminatory, or poorly developed. Users wishing to integrate data from multiple sources with different policies will then encounter the complex problem of restrictions stacking up and compounding each other, making use of the derived data very difficult.

- Coordination of the infrastructure is absolutely crucial, as the essence of an SDI is based on distributed sources of data (and services) from multiple contributors. To build an effective infrastructure it is necessary to get these multiple contributors to agree on standards, protocols, policies, and funding mechanisms to deploy and sustain the operations of the infrastructure. These agreements are by no means easy to forge and sustain, as each partner has to derive a benefit in the endeavor, and the benefit has to be perceived at all levels from senior management down to the technical staff who implement standards and protocols in their daily work. Underpinning the implementation of an SDI with a legal framework it sometimes is helpful to get the attention of senior management and have some leverage to support coordination. It is, however, not a sufficient condition to guarantee success.

A practical example from Europe

Whilst many countries in the world are building an SDI, one of the most complex examples is the Infrastructure for Spatial Information in Europe, INSPIRE, because this SDI, focused on environmental applications, spans 28 countries in 24 languages, and has a strong focus on harmonizing the data necessary to underpin environmental policy across 34 different thematic domains (EC 2007).
The purpose of INSPIRE is to support environmental policy and overcome major barriers that affect the availability and accessibility of geospatial data. These barriers include:

- inconsistencies in spatial data collection: spatial data are often missing, incomplete or, alternately, the same data are collected twice by different organizations;
- lack of or incomplete documentation of available spatial data;
- lack of compatibility among spatial data sets that cannot, therefore, be combined with others;
- incompatible SDI initiatives in Member States that often function only in isolation;
- cultural, institutional, financial, and legal barriers that prevent or delay the sharing of existing spatial data.

The key elements of the INSPIRE Directive that have been developed to overcome these barriers include:

- metadata to describe existing information resources, so that they can be more easily found and accessed;
- harmonization of key spatial data themes needed to support environmental policies in the European Union (EU);
- agreements on network services and technologies to allow discovery, display, downloading of information resources, and access to related services;
- policy agreements on sharing and access, including licensing and charging;
- coordination and monitoring mechanisms.

INSPIRE addresses 34 key spatial data themes organized in three groups (or Annexes to the Directive) reflecting different levels of expected harmonization, and a staged phasing (Box 1).

The legal framework of INSPIRE has two main levels. At the first, there is the INSPIRE Directive itself, which sets the objectives to be achieved and asks the EU Member States to pass their own national legislation establishing their SDIs. This mechanism of European plus national legislation allows each country to define its own way to achieve the agreed objective, taking into account its own institutional characteristics and history of development.

The drawback of having 28 different “flavors” of INSPIRE is that making the system work is undoubtedly more difficult. For this reason, the Directive envisages a second level of legislation, the implementing rules, which are more stringent because they must be implemented as is, and do not require follow-up national legislation. Therefore, INSPIRE envisages technical implementation rules in the form of formal legal acts for metadata, harmonization of spatial data and services, network services, data and service sharing policies, and monitoring and reporting measures to evaluate the extent of the Directive’s implementation and to assess its impact. Each of these regulations needs the approval of the Member States and of the European Parliament. The first implementing rules were approved in December 2008 for metadata, while the last set of implementing rules to specify the data models of Annex III data themes and for the interoperability of spatial data services were approved in 2013.

INSPIRE has several characteristics that make it particularly challenging. The most obvious is that it is an infrastructure built in 28 different countries using 24 languages. The requirements for multilingual services and interoperability among very different information systems and professional and cultural practices are, therefore, very demanding. This means, for example, that existing standards, where they exist, must be tested in real distributed and multilingual settings.
In the best scenario, all works well, but for a European-wide implementation there is a need to translate the standards and related guidelines into the relevant languages (the standards by the International Standards Organization, ISO, and the Open Geospatial Consortium, OGC, are typically in English only). In other instances, testing has demonstrated that the standards are insufficiently mature, or leave too much room for different interpretations, and thus require further definition or individual bridges to make different systems interoperate. This can be seen in the results of tests on distributed queries in catalogs that used the same specifications; the results identified a number of shortcomings that required the development of an adaptor for each catalog, which in a Europe-wide system with thousands of catalogs, would obviously not scale. In harder cases still, there are no available standards, and therefore they have to be created. This applies, for example, to the specifications required for the interoperability of spatial datasets and services, which is a central feature of INSPIRE.

To understand the context, it is worth remembering that each country in Europe has its own heritage and traditions, which also include different ways and methods for collecting environmental and geographic data, different traditions on how to analyze and visualize them, including different coordinate reference systems (sometimes more than one in each country), projections, and vertical reference systems. These different traditions mean that it is not enough for an SDI in Europe to help users find and access data; it is also necessary to understand
the meaning of what we are accessing to make appropriate use of it. This means, in turn, that it is necessary to develop not only translation tools to help overcome the “natural” language barriers but also agreed reference frameworks, classification systems and ontologies, data models, and schemas for each of the data themes shown in Box 1, against which the national data can be transformed or mapped. This is necessary because it is not possible to ask the Member States and their national and local organizations to re-engineer all their databases. Thus, the approach adopted is to develop agreed European models and systems of transformation (on-the-fly or batch), so that the level of interoperability necessary for key European applications can be achieved.

The approach sounds simple but putting it into practice and sustaining it over a long period is very complex, as it required seven years of work to develop an agreed methodology (the generic conceptual model) and tools, mobilize hundreds of experts in different domains, and deliver and test the specifications for the datasets falling in each of the themes identified in the Annexes of the Directive. A visit to the INSPIRE website (http://inspire.jrc.ec.europa.eu/index.cfm) in the data specifications’ sections demonstrates the large amount of work involved. There are thousands of pages of specifications and, not to be forgotten, tens of thousands of comments that had to be addressed individually during the stakeholder consultations.

The organizational model put in place to develop INSPIRE is one of its more interesting features, drawing significant attention also from outside Europe. In essence, it is a substantial exercise in public participation, the likes of which is most unusual in policy-making. From the outset it was recognized that for INSPIRE to be successful and overcome barriers to data access and use, it was necessary for the legislators, implementers, and practitioners in the Member States to come together and agree on a shared understanding of the problem, and of possible solutions. Therefore, an expert group with official representatives of all the Member States was established at the beginning of the process in 2001, together with working groups of experts in the fields of environmental policy and geographic information to formulate options and forge consensus. The INSPIRE proposal was subject to an extended impact assessment to identify potential costs and benefits, before it was opened for public consultation. The revised proposal was then debated by the Council and European Parliament over a three-year period before final adoption in 2007. This process in itself is a good example in democracy but the more interesting aspect is the way in which interested stakeholders continue to participate in ongoing activities required to develop the INSPIRE implementing rules (i.e., the follow-up legal acts and detailed technical guidance documents).

To organize this process, two mechanisms were put in place: the first is to engage the organizations at European, national, and sub-national levels that already have a formal legal mandate for the coordination, production, or use of geographic and environmental information (the so called legally mandated organizations or LMOs). The second mechanism aims to facilitate the self-organization of stakeholders, including spatial data providers and users from both the public and private sectors, into spatial data interest communities (SDICs) by region, societal sector, and thematic issue. Through these mechanisms, hundreds of organizations and thousands of experts across Europe have volunteered to contribute to the development of the technical specifications, and their testing.
The complexity of this participatory approach is certainly innovative, not only in relation to the developments of SDIs but also more generally to the formulation of public policy at the European level. The outcome produces both consensus-based policy and the development and maintenance of a network of stakeholders that make it possible to implement a more effective distributed European SDI. Participation is high, but it requires considerable effort. In particular, maintaining the momentum of the process, the commitment of hundreds of organizations, and thousands of experts, and building a relationship of trust across multiple levels of governance is a difficult challenge. Without this social network, however, it would not have been possible to develop INSPIRE. An interesting outcome of this process is that the network is now used by the Member States to help them implement their technical infrastructures, and to share expertise, funding, and technical components across multiple Member States.

The INSPIRE example is notable because it highlights a general lesson that applies to any SDI: SDIs are about data, technology, and standards but would not work without a strong social infrastructure underpinning the technical one. This social dimension will become even more important as SDIs coalesce into other data infrastructures in the public and private sectors (e.g., the recent open data initiatives taking shape around the world) and the general public becomes not just a user but also a producer of information and data, which then can be integrated with official data. Technology is rapidly shifting towards mobile devices and open applications, thus challenging the established patterns of relationships among the traditional public sector players (mapping agencies, geological agencies, land and property organizations), value-added integrators, and users. The increased availability of data from high-resolution satellites and sensor networks sending constant streams of terabytes of data per day, coupled with unstructured data from mobile devices and social networks, will change the way data and information are organized and processed, but geographical location will remain as one of the key organizing principles of data streams. In this respect, SDIs will continue to be important whatever shape they take in the future.

SEE ALSO: Data structure: spatial data on the web; Digital Earth; Geocomputation; Geographic information science; Geographic information system; Geoportals; Geospatial metadata; Open Geospatial Consortium standards; Spatial database; Virtual geographic environments

References


Further reading


Spatial database

Markus Schneider

University of Florida, USA

Definition

In various application disciplines, such as geography, the earth sciences, and geographical information systems (GIS), there is a need to manage large volumes of spatial or geometric data, that is, data related to space. Geographic data are a special category of geometric data. According to the old vector/raster debate, data related to space can be subdivided into spatial objects (vector view), which have features such as identity, location, extent, and relationships to other objects, and image data (raster view), which are available as digital bitmaps and can be stored, manipulated, and retrieved as discrete entities. Because of the fundamentally distinct nature and dichotomy of the two kinds of data, the requirements and techniques for dealing with spatial objects are rather different from those for dealing with raster images. This entry will only deal with spatial objects that are kept in spatial databases, not with images that are kept in image databases. In some sense one can say that spatial databases and image databases are complementary. A spatial database system (Rigaux, Scholl, and Voisard 2002; Shekhar and Chawla 2003) is a full-fledged database system with additional capabilities for handling spatial objects. A spatial database system should not be special purpose systems only. Since in practice spatial objects always occur in connection with nonspatial, alphanumerical data, a spatial database system should be a full-fledged database system with additional capabilities for handling spatial objects. Feature 2 stresses that special data types are needed for spatial data. In the two-dimensional (2-D) space, we distinguish point, line, and region objects, each of which have different properties. Feature 3 relates to special operations that are needed.
for processing spatial objects. Examples are the computation of a region object as the result of the intersection of two region objects, the calculation of the length of a line object, and the determination of the convex hull of a large number of point objects as a region object. Feature 4 deals with special Boolean functions that allow one to check topological properties or cardinal directions. For example, one can check whether a line representing a river is located with in a region representing a state, or whether the city district Brooklyn is southeast of the Central Park. Feature 5 requires a special query language that allows a user to pose spatial queries that are evaluated by the spatial database system. Feature 6 refers to the implementation of the aforesaid concepts and aims to achieve a high performance of geometric computation in a database context. This especially requires methods from computer science and computational geometry (de Berg et al. 2008; Preparata and Shamos 1993). Spatial join techniques support connecting spatial objects from different database tables through some spatial predicate in a more efficient way than combining all tuples from individual tables (Cartesian product) and filtering the combinations that fulfill the predicate. Spatial index structures (Samet 2006) avoid scanning the whole set of spatial objects in order to identify those objects that are located in a particular area. This entry will exclusively focus on the conceptual features in points 1 to 5 and neglect the implementation features in point 6.

Historical background

In the late 1970s, an interest arose in storing geometric data in databases in order to support spatial and geographic applications. The success and efficiency of relational database technology for standard alphanumerical applications, which is rooted in its simple data model, its structured query language (SQL), and its well-understood underlying theory, has led to many proposals to transfer this technology directly to geometric applications and to explicitly model the complex structure of spatial data in database tables. But this turned out to be a very unfavorable approach since the consequence is that the user conceives spatial data in tabular form, just the same as standard data, and that a spatial object is represented by several or even many tuples and possibly even spread over several databases tables. An example of such a relation schema is

$$\text{RelName}(id : \text{real}, x_1 : \text{real}, y_1 : \text{real}, x_2 : \text{real}, y_2 : \text{real}, \text{type} : \text{string}, <\text{other information}>)$$

where $x_1, y_1, x_2,$ and $y_2$ are the coordinates of a point or a line segment. The flag type indicates whether a tuple describes a point, a single line, a line segment of a line, or a line segment of a polygon. The value $id$ denotes the object identifier. A line object with several line segments has now to be represented by several tuples, each of which describes a single segment. One can show that an adequate representation of a polygon needs three tables.

This approach has revealed a number of fundamental drawbacks. Since all lines and polygons are decomposed into a set of line segments (tuples) scattered over one or more relations, a spatial object is not treated as an entity or unit but only corresponds to several tuples. This is different compared to values of standard data types such as integers or strings. A second drawback is that the approach forces the user to model complex spatial objects in flat, independent tables. Since the representation of spatial data occurs on a very low level and is exclusively based on standard domains such as integers, strings, and reals (while the user has originally intended to deal with points, lines, or polygons (regions)), an adequate treatment of spatial data is impeded. Although the facilities of the query language of a database management system (DBMS) are available, they
are only of limited use. Since such a language is based on standard domains and has no concept of spatial data types, it cannot provide and support any meaningful geometric operations.

**Spatial data types**

The numerous deficiencies of the approach of modeling spatial data as database tables have resulted in the assessment that this approach is unsuitable to manage spatial data in a clean and efficient manner and that a high-level view of spatial objects is essential. This has led to the design and implementation of special data types called spatial data types (Schneider 1997). Data types, in general, are a well-known concept in computer science (e.g., in programming languages or in database systems). A data type defines a set of homogeneous values and the allowable operations on those values. An example is a type integer representing the set of 32-bit integers and including operations such as addition, subtraction, and multiplication that can be performed on integers.

**Spatial data types** or geometric data types provide a fundamental abstraction for modeling the geometric structure of objects in space as well as their relationships, properties, and operations. They form the foundation of spatial databases. One speaks of spatial objects as values of spatial data types. Due to the inherently complex structure of spatial objects, spatial data types have been designed as abstract data types (ADTs). This enables them to (i) provide a high-level view of spatial objects since the internal data structure of a spatial object is invisible to the user, (ii) offer a high-level usage of spatial operations and predicates because the underlying geometric algorithms are hidden from the user and results of geometric computations and properties of spatial objects can only be obtained by calling these (abstract) operations, and (iii) be used as attribute data types in a database schema in the same way as standard data types such as integer or string. Spatial data types have a conceptually complex structure and can be defined for the 2-D space and the three-dimensional (3-D) space. But spatial databases nowadays mainly support 2-D spatial data. In the 2-D space we distinguish the spatial data types point, line, region, spatial network, and spatial partition. Note that different implementations (e.g., software packages) may adopt different terms for each of these spatial data types, but the underlying concepts are universal. If we are only interested in the locations of a spatial object, we use a point object. A point object represents a finite number of locations, that is, single points. For example, the locations of all lighthouses in Florida can be kept in a point object. A point object is called simple (Figure 1a) if it only contains a single point; otherwise, it is called complex (Figure 1d). If we are interested in the areal extent of a spatial object, we call it a region object. A region object consists of a finite number of disjoint components (except for common single boundary points) called faces, possibly with one or more disjoint holes (except for common single boundary points). Examples are an air-polluted zone or Italy with its mainland and offshore islands as faces and the Vatican as a hole in the mainland. A region object is called simple (Figure 1c) if it only contains a single face without holes; otherwise, it is called complex (Figure 1f). If we are interested in the linear extent of a spatial object, we call it a line object. A line object consists of a finite number of disjoint components called blocks, each of which contains a finite number of curves. Examples are the ramifications of the Nile Delta or the geometry of a street network. A line object is called simple (Figure 1b) if it consists of a single block with a single curve; otherwise, it is called complex (Figure 1e). The generality of the aforementioned spatial data
Spatial Database

Figure 1 Examples of a simple point object (a), a simple line object (b), a simple region object (c), a complex point object with six points (d), a complex line object with two blocks (e), and a complex region object with five faces and five holes (f).

types in terms of multicomponent spatial objects is important since only this property ensures that the spatial data types are closed under geometric set operations such as geometric intersection, geometric union, and geometric difference. That is, applying a geometric set operation to two spatial objects of the same spatial data type will lead to a resulting spatial object of the same type.

Even more complex spatial data types are spatial networks and spatial partitions as the essential components of maps. They represent spatial connectivity structures, that is, the individual spatial objects in them fulfill certain topological constraints toward each other. A spatial network (Figure 2a) can be viewed as a spatially embedded planar graph which consists of a set of point objects representing its nodes and a set of nonintersecting line objects describing the geometry of its edges. Examples are highways, rivers, public transportation systems, power lines, and phone lines. A spatial partition (Figure 2b) is a set of region objects fulfilling the topological constraints that any two regions either meet or are disjoint. The neighborhood relationship is of particular interest here since region objects may share common boundaries. Examples are states, school districts, crop fields, and land parcels. Both spatial networks and spatial partitions are not purely geometric structures since their components (line objects, region objects) are annotated with thematic data such as state name, unemployment rate, and parcel identifier.

Three-dimensional data types have so far not been implemented and provided in spatial databases. But increasingly geoscience applications are interested in 3-D data representations and geometric computations. They will require data types such as surface, for example, to model the shape of landscapes, and volume, for example, to represent urban areas, clouds, hurricanes, and temperature zones.

Spatial operations

Spatial objects as values of spatial data types are processed by spatial operations. A spatial operation is defined as a unary or binary function with spatial argument objects, that is, it takes one or two spatial objects as operands and returns either spatial objects or scalar values as results. Spatial operations are especially important for spatial query languages into which they are embedded and which allow new kinds of powerful queries
that were impossible before. In the following we provide an overview of the most important spatial operations and consider them in the 2-D space. They can be classified into the following categories: (i) spatial predicates returning Boolean values, (ii) spatial operations returning constituents of a spatial object, (iii) spatial operations returning numbers, (iv) spatial operations returning spatial objects, (v) spatial operations on sets of spatially referenced objects, and (vi) spatial selection and spatial join.

A spatial predicate is applied to two spatial objects and returns a Boolean value that represents the validity of a spatial relationship. A spatial relationship is a relationship between two spatial objects and characterizes their relative position to each other with respect to different aspects. They can be classified into the following subcategories: (i) topological relationships, (ii) directional relationships, (iii) spatial order and strict order relationships, (iv) metric relationships, and (v) fuzzy relationships. Their importance for spatial databases consists in their use as filter conditions in spatial joins and spatial selections.

Topological relationships such as \textit{intersect} and \textit{disjoint} describe purely qualitative properties of the relative positions of spatial objects toward each other. They rest on topological properties for their description and include concepts such as continuity, adjacency, overlapping, interior, boundary, connectivity, and inclusion. There is no dependence on a distance function, since topological relationships are not distance-preserving. An essential property is that they are preserved under continuous topological transformations including all affine transformations such as translation, rotation, and scaling. They enable the user, for example, to ask for land parcels that are \textit{adjacent} to a hazardous waste site, to explore the \textit{overlapping} of a river floodplain with a proposed highway network, or to determine for a collection of cities and states which city is located \textit{within} which state. They belong to the most formally investigated spatial relationships. The goal has been to formally determine sets of topological relationships that characterize all topological situations between two spatial objects and are mutually exclusive. The most well-known models for achieving this are the 9-intersection model (9IM) (Egenhofer and Franzosa 1991) and the Region-Connection-Calculus (RCC) (Cui, Cohn, and Randell 1993). Figure 3 shows the eight topological relationships \textit{equal}, \textit{disjoint}, \textit{coveredBy}, \textit{covers}, \textit{intersect} (\textit{overlap}, \textit{cut}), \textit{meet} (\textit{touch}, \textit{adjacent}, \textit{neighboring}), \textit{inside} (\textit{within}), and \textit{contains} between two simple region objects as they have been identified by both models.

Topological relationships have been investigated for all combinations of simple spatial objects as well as all combinations of complex spatial objects of the spatial data types \textit{point}, \textit{line}, and \textit{region} (Egenhofer and Franzosa 1991; Schneider and Behr 2006). Table 1 shows their numbers based on the 9-intersection model.

The large number of topological predicates for each combination of complex spatial data types makes it difficult for the user to distinguish, remember, and handle them. Topological cluster predicates (Schneider and Behr 2006) have been defined that group similar topological

![Figure 3](image)

**Figure 3** The eight topological relationships between two simple region objects.
Table 1  Numbers of topological relationships between two simple/complex spatial objects.

<table>
<thead>
<tr>
<th></th>
<th>Simple/complex point</th>
<th>Simple/complex line</th>
<th>Simple/complex region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple/complex point</td>
<td>2/5</td>
<td>3/14</td>
<td>3/7</td>
</tr>
<tr>
<td>Simple/complex line</td>
<td>3/14</td>
<td>33/82</td>
<td>19/43</td>
</tr>
<tr>
<td>Simple/complex region</td>
<td>3/7</td>
<td>19/43</td>
<td>8/33</td>
</tr>
</tbody>
</table>

relationships together based on user-defined and/or application-specific rules.

*Directional relationships* also describe purely qualitative properties and check the *cardinal directions* between two spatial objects. More precisely, we distinguish *absolute directional relationships* such as *north* and *southwest* with respect to a given reference or coordinate system from relative directional relationships such as *rightOf* and *leftOf*. Absolute directional relationships are especially important for geographical applications. Users may, for example, ask for all hurricanes that have ever been west of Florida, the general direction of the whale routes in the Gulf of St Lawrence, or all land parcels that are located to the west of the parcels of power stations. The background for the interest in the latter query could be that due to west winds these parcels are endangered by the hazardous influence of air pollution.

*Metric relationships* such as *inCircle* and *inWindow* use distance measurements in order to test if a spatial object lies within the scope of a predefined circle or rectangle, respectively.

*Spatial order and strict order relationships* are based on the definition of order and strict order. As a rule, each order relation has an inverse relationship. For instance, *behind* is a spatial order relationship based on the order of preference with the inverse relationship *inFrontOf*. Other examples are the pairs *above*/*below* and *over*/*under*. Also, the pair *inside*/*contains* can be regarded as a spatial order relationship.

*Fuzzy relationships* such as *nextTo*, *close*, *near*, or *far* are vague distance relationships. While users are able to cognitively interpret their meaning, it is difficult to formally define these predicates.

*Spatial operations returning constituents of a spatial object* allow the user to extract components of a spatial object that are conceptually known to the user but not directly accessible at the user’s level; they can only be accessed through operations. Examples are *xCoord* and *yCoord* to obtain the x-coordinate and y-coordinate of a single point as real numbers, *getSegment* to get access to a segment of a line or region object, and *getFace* to get access to a face component of a region object.

*Spatial operations returning numbers* comprise spatial operations which compute a metric property of a spatial object and return a number as a result. The operation *area* returns the area of a region object, and the operation *perimeter* yields its perimeter. The operation *length* calculates the total length of a line object. The operation *dist* determines the minimal Euclidean distance between any two spatial objects. The operation *diameter* computes the largest Euclidean distance between any of its components. The operation *direction* computes the angle between any two spatial objects as a real number in the range $0 \ldots 359°59’59”$. The results of all operations enumerated so far are real numbers. The operation *noOfComponents* (cardinality) yields the number of components (points, blocks, faces) of a point, line, or region object as an integer number.
Spatial operations returning spatial objects take one or two spatial objects as arguments and return single spatial objects as results. Two subcategories can be distinguished: (i) object construction operations which construct new objects from two or more existing objects, and (ii) object transformation operations which transform a single spatial object into a new spatial object.

Object construction operations as the first subcategory include geometric union, intersection, and difference operations and are applied to two point, two line, or two region objects. These operations fulfill closure properties. This means that if they are applied to two objects of the same spatial data type, they always yield an object of the same spatial data type. For example, to find out which part of a school district is exposed to an air-polluted area, we can geometrically intersect both region objects and will obtain a region object as a result. For the intersection of line objects we have the special case that they can also intersect in a collection of points, that is, in a point object. Two region objects might meet and share a collection of single boundary points represented as a point object. In both cases the operation commonPoints computes this point object. Two region objects may also meet and share a common boundary line. The operation commonBorder determines this common boundary as a line object. The operation intersection can also be applied to two objects of different spatial data types and yields an intersection object of the spatial data type of lower dimension.

The operation convexHull (Figure 4) constructs the convex hull of a point set (point object) and is defined as the smallest convex polygon enclosing all points. It is, for example, deployed to find the diameter of a point object, that is, the maximum distance of any two points of the point object. The diameter will always be the distance between two points of the convex hull.

The operation closest takes a complex point object $p$ and a reference point object $r$ as input values and yields a complex point object whose points are from $p$ and nearest to the points in $r$ (Figure 5).

The operation $mBB (box, mBR)$ computes the minimum bounding box (aka minimum bounding rectangle) that bounds a spatial object and is axis-parallel to the selected reference frame (coordinate system). The operation centroid computes the arithmetic mean, that is, average, position of all the points in the spatial object. The operation contour (aka boundary, border) produces the boundary line of a region object as a line object. Such a line object contains the outer cycles and hole cycles of all region object components (faces). The operation vertices (aka boundary, border) returns the end points of a line object as a point object. The operation choose produces an “arbitrary” point on or inside each component of a line object or a region object. This is useful when such an object is to be approximated by a single point for each component as a representative. The operation interior is applied to a line object, identifies all outer, encompassing
cyclic structures, and represents them as a region object. The effect is that all cyclic and acyclic structures within these outer cyclic structures are removed. For example, if $R$ is a region object, then the term $\text{interior(\text{contour}(R))}$ can be used to remove all holes in $R$. Another variant of the operation $\text{interior}$ takes a point object and a line object as operands and produces a region object that comprises a set of smallest faces where each face surrounds a point of the point object and has a boundary that is part of the line object. That is, if $R$ is a region object, then the term $\text{interior(choose}(R), \text{boundary}(R))$ is equal to $R$.

Object transformation operations as the second subcategory comprise operations such as $\text{extend}$, $\text{rotate}$, and $\text{translate}$. The operation $\text{extend}$ (also called $\text{buffer}$, $\text{buffer zoning}$, or $\text{region expansion}$) takes a spatial object $s$ and a real number $r$ as operands and creates a polygonal outer region which is a spatial expansion of $s$. Parameter $r$ indicates the expansion distance from $s$ (Figure 6). If $r$ is negative, a shrinking operation on a region $s$ is computed and leads to an inner region. The operation $\text{rotate}$ rotates a spatial object around a point; the operation $\text{translate}$ moves a spatial object by a defined vector.

Spatial operations on sets of spatially referenced objects support the manipulation of spatial connectivity structures such as spatial partitions, that is, subdivisions of space (such as school districts), and spatial networks, that is, spatially embedded graphs (such as road networks), which are associated with thematic data (such as the number of students or road names). As such they are very complex since, in contrast to the operations of the other categories, these operations have to cooperate with a DBMS data model (relational, object-oriented, complex object, etc.). They do not manipulate individual spatial objects but collections of spatial objects with special topological properties and annotated with thematic data. They take a set of spatially referenced objects as arguments and produce a new set of spatially referenced objects as a result.

The operation $\text{overlay}$ is one of the most frequently used spatial operations in a GIS and allows a user to transparently lay two partitions modeling different thematic topics on top of each other and to combine them geometrically.
Two variations of the operation overlay.

Figure 7 Two variations of the operation overlay.

into a new partition of disjoint or adjacent regions. Two different interpretations of the overlay mechanism are given in the literature (Figure 7). In the first interpretation (overlay₁), the resulting object set contains one new region object obtained as the intersection of a region object of the first partition with a region object of the second partition. Since the plane need not be completely covered by regions, it is possible that a region object of one partition does not intersect any region object of the other partition. In this case it will not be part of the resulting partition. In the second interpretation (overlay₂), also taken into account are those parts of region objects that do not intersect any region object of the other partition.

In this case it will not be part of the resulting partition. In the second interpretation (overlay₂), also taken into account are those parts of region objects that do not intersect any region object of the other partition.

The operation superimposition allows the spatial objects of a partition to be superimposed on another partition and parts of the other partition to be to covered and erased (Figure 8).

The operation fusion (merge, generalization) takes as input a spatial partition and a set of thematic (i.e., nonspatial) attributes that are a subset of those attributes associated with all region objects of the spatial partition. As output it creates a new spatial partition that geometrically merges the region objects of the input partition on the basis of the equality of the values of the reduced attribute set and that associates the corresponding values of the reduced attribute set to each obtained region object. Figure 9 shows a partition of districts with their land use. Each district is marked with an identifier \( d_i \) and its crop cultivation. The task is to compute the regions with the same land use. For this purpose, the set of attributes \{identifier, crop cultivation\} is reduced to the set \{crop cultivation\}. This has the effect that single districts with the same land use are merged by geometric union to a complex district possibly consisting of several faces. Further, neighboring districts with the same crop cultivation are geometrically merged by erasing the common boundary line between them. Each hatched area \( d_i \) on the left side of Figure 9 is part of an object describing a district. On the right side, after the application of the operation fusion, all areas belonging to the same group \( g_i \) form a single region object and are hatched in the same way. District boundaries are not distinguished any more. To be feasible, the fusion operation requires that region objects are defined as many-component objects with holes.

The operation cover (sum) yields a single region object as the geometric union of all region objects of a spatial partition (Figure 10).

The operation windowing allows one to retrieve those regions of a spatial partition whose intersection with a given (rectangular, polygonal, circular) window is not empty (Figure 11). An example of windowing is a query where the window is defined as a circle with center \( p \) and radius \( r \) and where we ask for all objects whose distance from a given point \( p \) is less than \( r \).

The operation clipping selects those parts of a partition which lie inside a given rectangular window (Figure 12).
The operation \textit{voronoi} is based on a well-known structure from computational geometry. For a given set $S$ of points in the plane, the Voronoi diagram (also called Thiessen diagram) associates with each point $p$ from $S$ the region consisting of those points of the plane that are closer to $p$ than to any other point in $S$ (Figure 13). The resulting set of regions forms a spatial partition. Usually, some regions of a Voronoi diagram are infinite, namely those belonging to the points on the convex hull of $S$. But if we assume a rectangular reference frame that surrounds all points in $S$, the operation \textit{voronoi} constructs the Voronoi diagram for the points inside this rectangle.

An important task of spatial predicates is to serve as filter conditions for \textit{spatial selection} and \textit{spatial join} operations. Let $S$ be a set of database
objects or database tuples, each of which contains a spatial object. A *spatial selection* filters out all those database objects of S whose spatial object fulfills a selection condition given either by a spatial predicate or by a comparison expression that contains a spatial operation with a numerical result as a constituent part. For example, having stored data about the US states, one could ask for all those states with an area less than 66000 square miles.

*Spatial join* operations represent database operations that combine two sets of database objects on the basis of the fulfilment of a spatial predicate; they construct a new database object as a concatenation of the two argument database objects. For example, given a set of city objects with their geometric point locations and a set of state objects with their regions, one could ask for all those city names and state names where the city location is located in the state region. This is an example of an *inside join*.

In the literature, the term “spatial join” is not used in a uniform way. Some authors equate the term with the spatial join intersection operation, that is, where the spatial predicate in the join condition is restricted to an intersect predicate. Other authors do not distinguish between spatial selection and spatial join operations. Another example is an *overlap join* where the filter condition checks the intersection of two region objects. An example is a query that finds out all counties that were affected by hurricane Katrina.

### Spatial query languages

Spatial query languages serve as a communication interface between the spatial database system and the user. By adding spatial data types, operations, and predicates, they are often extensions of SQL, which is the standard query language for relational databases. Spatial queries allow the user to get the spatial information of interest. A few examples will illustrate this. Consider the map of the 50 states of the United States. Each state has its thematic attributes, such as name and population, as well as a geometry which is described by its region. Such a region has an internal complexity since it can have holes (such as enclaves) and can consist of several components (such as mainland and islands). Cities have a name and a population and are geometrically represented as points, that is, we are here interested in their location and not so much in their extent. Let us assume the two relational database tables *states* and *cities* whose schemas are given as

**states**(sname: string, spop: integer, territory: region)
**cities**(cname: string, cpop: integer, loc: point)

Each table schema consists of attributes (such as *sname* or *loc*) that describe properties of database objects or tuples (such as *states* objects or *cities* objects), and each attribute has a data type such as *string* or *integer*. The two spatial data types *point* and *region* are used to represent the spatial objects of cities and states, respectively. They are attribute data types in the same way as the types

---

**Figure 13** The operation *voronoi*. 
**SPATIAL DATABASE**

*string* and *integer*. The difference is that spatial data types have an inherently complex structure. The internal structure is made deliberately invisible to a user and can only be accessed by high-level operations. A query could ask for all pairs of city names and state names where a city is located in a state. This can then be formulated as a spatial join.

```sql
select c1.sname, c2.sname, dist(c1.loc, c2.loc) as distanceBetweenCities
from cities c1, cities c2
where c1.cname <> c2.cname
```

This query uses the metric operator *dist* to compute the minimum Euclidean distance between two point objects. The next query determines all neighbor states of Colorado and is another example of a spatial join.

```sql
select s1.sname, s2.sname
from states s1, states s2
where s1.territory meets s2.territory and s2.sname = 'Colorado'
```

This query make use of the topological predicate *meets*. A further query lists all states with at least one neighboring state and sorts this list according to the number of neighboring states.

```sql
select s.sname, count(s1.sname) as noOfNeighbors
from states s, states s1
where s.territory meets s1.territory
order by count(noOfNeighbors)
```

This query first combines each state tuple *s* with each other state tuple *s1* if their territories are adjacent. Then it groups all resulting tuples that have the same state name *s.sname* together, uses the SQL aggregate function *count* to total its neighboring states *s1.sname*, and outputs the state names sorted in ascending order by the number of neighboring states. Our last query asks for those states that have only one neighboring state.

```sql
select s.sname
from states s, states s1
where s.territory meets s1.territory
having count(s1.sname) = 1
```

Here, by using the SQL *having* clause, only those groups that consist of only one tuple are admitted for further consideration. They describe states with only one neighboring state. The reader is invited to determine whether this query has a solution and, if so, what the solution is.

**SEE ALSO**: Data model, object-oriented; Data structure, vector; Geographic information science; Geographic information system; Indexing; Open Geospatial Consortium standards; Representation: complex objects; Sorting spatial data; Spatial feature classes; Topological relations
References

Spatial problems confronting policy, business, and planning arenas are often complex and “wicked” in nature, in that they are hard to specify and have no easy, satisfactory solutions (Jankowski et al. 1997). Yet, decisions still need to be made. This dilemma – decision-making in the face of uncertainty, ambiguity, and incomplete information – has formed the foundation for spatial decision-support systems (SDSS) research and development (Densham 1991). This dilemma has also guided the development of SDSS from workstation, expert-oriented, and single-user systems to web-enabled, multi-user systems that can be used in settings as diverse as participatory planning (Jankowski et al. 2006) to on-the-ground, real-time decision-making for land management (Arciniegas and Janssen 2012), routing, and many other applications. The rationale for ongoing SDSS development is exemplified by the estimate that 80% of data used by managers are spatially dependent, meaning that many high-stakes decisions have integral spatial components.
SDSS researchers and practitioners have decided to address have shaped the current status of SDSS research and understanding.

Background and history

Sugumaran and Degroote (2011) divided the history of SDSS into the three overlapping phases. In the 1980s, researchers and practitioners merged the notions and technologies of geographic information systems (GIS) and decision-support systems (DSS), and subsequently developed prototypes. During the 1990s, the SDSS community integrated new technologies into SDSS and applied the technology to new disciplines and in novel use cases, such as group and collaborative contexts. By the start of the 2000s, SDSS were being implemented more widely, particularly with the advent of web-based SDSS and the subsequent deployment on mobile technologies. While these phases are not definitively demarcated points in time, they can serve as a general frame of reference for considering the broad trends within SDSS research and development.

During the introductory phase of SDSS, researchers developed conceptual frameworks and subsequent prototypes for command-line desktop or workstation users. Patrick Mantey and Eric Carlson were likely the first to document an SDSS in the 1970s and 1980s, though they did not use the term “spatial decision-support system.” Their work focused on the geodata analysis and display system (GADS). Developed by IBM, GADS allowed noncomputer users to interact with spatial data. Another similar system, the generalized planning system (GPLAN), used spatial data in a decision-support system. The system allowed users to iteratively analyze and plan land use while considering the impacts on water quality. For the first time in 1985, Hopkins and Armstrong used the term “spatial decision-support systems” to describe an analytic and cartography framework, and by 1989, the terminology of spatial decision-support systems was becoming more widely used, particularly in the geographic and land information systems literature where prototype applications were being described. This last point illustrates the long connections between SDSS and land-use analysis and planning. Generally speaking, wider adoption was hindered during the introduction phase by expert knowledge requirements, and the high technology costs related to computation, memory, and data storage.

Computer hardware and software advances in the 1990s encouraged SDSS development, which spurred greater deployment for new use cases. Specifically, three things enabled SDSS expansion: (i) less expensive, higher-powered desktop computers, (ii) easy-to-use graphical user interfaces (GUIs), and (iii) more accessible software development environments. Because SDSS became easier to develop and use, they could be more widely researched and adopted. Commercial GIS packages such as the Environmental Systems Research Institute’s (ESRI) ArcInfo and ArcView, MapInfo, IDRISI (now known as TerrSet), TransCad, along with open source GIS such as Geographic Resource Analysis Support System (GRASS), played important roles in advancing the usability of GIS and encouraged the development of SDSS.

Accompanying these new software packages were easy-to-use application development languages such as Arc Macro Language (AML). These new languages enabled experts to more easily incorporate and create problem-specific models, and allowed researchers to address one of the core limitations of utilizing GIS as SDSS (Densham 1991). The proliferation of GUIs also enabled collaborative or group
SDSS by overcoming some of the limitations of single-user SDSS and command-line interfaces.

Theoretical developments in participatory planning processes and collaborative decision-making accompanied technical developments. For example, Jankowski et al. (1997) developed and implemented a group SDSS that utilized multicriteria decision modeling combined with consensus-building techniques. The general framework, called Spatial Group Choice, was originally used to support a habitat restoration project, but was also shown to be applicable to areas as diverse as healthcare and transportation system design. Looking ahead to the potential proliferation of SDSS, where experts would not always be able to guide use, researchers saw a need for automated expert-guided decision systems. Such systems, developed using expert knowledge, relied on extensive sets of if-then-else statements that would guide users through a problem context. As GIS and ICT advanced, and the Internet emerged through the late 1990s, researchers began turning their attention to the potential that existed for web-based SDSS, though the vision wouldn’t be fully realized until the 2000s (Sugumaran and Degroote 2011).

The number of SDSS publications grew substantially in the 2000s as GIS became more usable and increasingly applied to specific problem domains. Use of web-based SDSS enabled increased accessibility and cross-platform flexibility in addition to an efficient mechanism for data storage, administration, and distribution. Specific web-mapping platforms emerged in the late 1990s and early 2000s such as ESRI’s ArcIMS (Arc Internet Map Server), the University of Minnesota’s MapServer, and GeoMedia WebMap, though these platforms supported limited analysis capabilities. In many respects, the 10 years of web-based SDSS development from 2000 to 2010 mirrored the desktop developments of the 1990s. In terms of content, web-enabled SDSS often focused on environmental or conservation impacts of land use, and often employed elements of multicriteria decision analysis, but also included examples of identifying potential store locations and route finding. Web-enabled tools for spatial data storage, processing, and analysis could be built using many different development platforms, such as Extensible Markup Language (XML), Web Services Description Language (WSDL), Simple Object Access Protocol (SOAP), Universal Description, Discovery and Integration (UDDI), and ESRI’s ArcGIS. The rise in web-based services marked the beginning of service-based SDSS, and provided an opportunity for modular, problem-specific instantiations of user-friendly SDSS.

Currently, web-based and distributed SDSS continue to advance in their capacities for usability, mobility, and computational power. With the rise of cloud computing technologies, many SDSS follow a server-side data storage and processing architecture where users interact with data visualized through a web browser or lightweight application, and then send processing requests to a central server that can handle large volumes of data (Wang et al. 2013). New distributed server frameworks such as Spatial-Hadoop (http://spatialhadoop.cs.umn.edu) and other “big data” infrastructures are enabling users to “mash up” data and perform analysis through a web browser with processing occurring off-device (Wang et al. 2013). The trajectory of SDSS fits into the larger trajectory occurring within GIScience (i.e., CyberGIS; Wang et al. 2013) as demand increases for real-time, flexible spatial data analysis and visualization. Example use cases include the monitoring of and responding to social media for disaster relief. This move in GIScience corresponds with the general trend toward remote
monitoring and dynamic management of the “Internet of Things” that consists of critical infrastructures and services such as the power grid or emergency response systems, as well as consumer household products. Additionally, as the trajectories of planning and business continue toward collaborative processes, researchers are exploring how to maximize participation through new input/output devices such as large touch screens (Arciniegas and Janssen 2012).

SDSS as sociotechnical systems

Taking the broadest definition, an SDSS consists of a technical component (i.e., hardware and software) and a stakeholder component (i.e., users; Jankowski 2008). These two components form a sociotechnical system that can function in many different ways, and therefore, can be described in many different ways, which is aptly illustrated throughout the SDSS literature (Densham 1991; Jankowski 2008; Sugumaran and DeGroote 2011).

Technical components

The technical component of an SDSS combines the spatial capacities of geographic information systems with analysis, modeling, scenario evaluation, and other decision tools to support a user or group of users in a spatial decision-making processes. As data availability has improved, and technology has become less expensive and easier to use, more people in a greater diversity of domains have begun to use SDSS, which mirrors the general trajectory of GIS and ICT (Sugumaran and DeGroote 2011). From a hardware perspective, this trajectory has meant a shift from individual workstations and personal computers to collaborative touchscreens and mobile and cloud-based SDSS models (Arciniegas and Janssen 2012). Software has also undergone a revolution, transitioning from single-user software to web-based applications that run through Internet browsers supported by large-scale, cyberGIS infrastructures (Wang et al. 2013).

Spatial decision-support software components are typically quite diverse, and rely on various configurations, depending on the specific decision problem. The various software components for an SDSS can be (i) embedded, where the system components are hard-coded together and share common data structure, storage, and user interface, (ii) tightly coupled, where components share a common user interface, have separate data structures or storage, but may still be hard-coded together, or (iii) loosely coupled, where inputs and outputs are used from component to component, but may be transferred by an analyst or user (Jankowski 2008). Embedded and tightly coupled components can happen on individual systems, but are increasingly happening server-side through web portals.

The core of an SDSS is the database that houses vector, raster, and tabular forms of data. Depending on the scale of the SDSS implementation, databases can be housed on individual workstations or on a centralized server. If the data are housed on a centralized server, data processing can occur either client-side on users’ devices, or server-side. The latter is often preferable for web-based applications given the greater computing power of servers. Data interoperability has long been a major challenge within GIScience; however, new data synthesis efforts combined with easy-use portals could substantially decrease this problem. Examples of such projects include TerraPopulus (Minnesota Population Center 2013), the Integrated Public Use Microdata Series (iPUMS; Minnesota Population Center 2013), the Globe Project (globe.umbc.edu), NASA’s socioeconomic data center (SEDAC), DataOne (www.dataone.org), the University of
Wisconsin’s Center for Sustainability and the Global Environment (http://nelson.wisc.edu/sage), and commercial data warehouses such as ESRI’s Living Atlas of the World. However, having large amounts of data readily accessible creates new concerns surrounding data accuracy, uncertainty, and appropriate use, particularly if such data are being used in high-stakes decision processes (Sugumaran and DeGroote 2011).

Models used in SDSS can take many forms, ranging from simple land suitability models, to complex dynamic equilibrium, agent based, spatial analysis, or multicriteria optimization models. Simple distance measures, overlays, buffers, and land suitability modeling capabilities often come with common GIS packages (Jankowski 2008). Dynamic equilibrium and agent-based models often require add-on packages or extensive programming in GIS-specific languages in order to implement (Parker et al. 2003; Shook, Wang, and Tang 2013). However, additional modeling with GIS add-on packages is often a primary requirement of creating a fully functioning SDSS. Spatial analysis functionality encompasses a large range of spatial statistics applications, and can be used to explore empirical or modeled data. Multicriteria optimization modeling can involve all of the previously mentioned models, and seek a mathematically optimal outcome across multiple variables (Malczewski and Rinner 2015). Model selection will depend heavily on the specific purposes of the SDSS.

Visualization within SDSS mostly consists of two-dimensional maps and charts, but with new computing advancements, three-dimensional visualization is becoming more feasible. Visualization of data should enable users to simply and quickly weigh the tradeoffs and synergies for the relevant variables across the different decision situations (Jankowski 2008). One of the reported benefits of using maps as a central means to visualize data is that decisions can be made more quickly; however, decision performance can decrease as problem size and data dispersion increase and data aggregation decreases (Keenan 2006). As a result, experts are often more accurate than nonexperts in using maps for decision-making, which means that education and training are key in order for nonexperts to successfully use maps for decision-making.

User interfaces for SDSS have ranged from command-line and mouse and keyboard interfaces, to large touch tables for group collaborative work. While command-line and mouse and keyboard interfaces may be preferable for the expert working alone, they can present barriers to group interaction. As a result, large touch tables and screens have been deployed. In one example, Arciniegas and Janssen (2012) describe the use of a large touch-enabled screen laid flat as a table. The tool can sense which of four users is touching it, and allows for simultaneous usage. Such tools enable increased face-to-face collaboration and decision-making by removing the computer from the primary center of focus, but still provide a shared map interface.

Domains as diverse as transportation engineering, city and regional planning, water resources management, agriculture, and forestry have adopted SDSS, with users ranging from individuals to large groups across a variety of levels of technical expertise. New commercial tools, such as ESRI’s ArcGIS Online, CityEngine, and Geodesign tools, along with open source Javascript libraries, such as Leaflet and D3, are likely to continue to increase disciplinary use of SDSS as applications become easier to build, deploy, and maintain, and less expensive to rapidly develop. However, “big data” applications will likely remain time-, money-, and expertise-intensive, though with the advent of
modular and reproducible “big data” SDSS, this may not be the case.

Social component

Decision-making involves the selection from two or more potential options (Jankowski 2008). How individuals and groups move through the decision process has been theorized in various ways, each of which has strengths and weaknesses when considering the implications for SDSS technical and process design. Antunes et al. (2014) describe three complementary views of models for the decision-making process: sequential, dynamic, and continuous (for more information, see Decision analysis). Each theoretical model is likely to have tradeoffs and synergies, given a specific SDSS application context. Stakeholder expertise, as stated above, plays a significant role in what creates an effective SDSS for an application context. Regardless of the theoretical model that an SDSS will be designed with, iterative development and user testing will be required.

Design considerations

A major goal of human–computer interaction is anticipating user needs for information (Jankowski et al. 2006). In practice, doing so requires iterative development of an SDSS. Development processes are discussed at length in the human–computer interaction literature. There are many factors to consider when designing an SDSS, including software and hardware costs; the deployment hardware platform and the respective operating system; user or group technical abilities and their respective user interface needs; the availability, suitability, interoperability, and usability of different models; and model data requirements and types (Jankowski 2008). Determining what the “best” choices are in each of these categories will be entirely dependent upon the context in which the SDSS will be used, and the unique opportunities and limitations this context offers. Uran and Janssen (2003) reviewed the SDSS literature and asked a simple question: Why are some spatial decision-support systems not used? They guided their evaluation of various SDSS by five questions.

- How are alternative scenarios or decision pathways specified?
- How do users get from the start of a decision pathway to the finish?
- How is the output of the process presented?
- How is evaluation of the results supported?
- Does the SDSS function as it is supposed to?

To answer these questions, they compared across multiple case studies and found that SDSS provided little to no support for aspatial or spatial evaluation of the scenario outputs, which they think may have caused low levels of adoption in the Netherlands.

Sugumaran and Degroote (2011) provide six additional questions that are directly related to the five listed above and which they believe are often overlooked in the implementation of SDSS.

- What software system should be used? Which is best?
- What aspatial and spatial models would address the decision problem?
- Are the user interfaces easy to use? How should these be developed?
- How are outputs validated and evaluated?
- What technologies should be used?
- Do sufficient resources exist to create an SDSS?
Because of the very nature of ill-structured spatial decision problems, even if these questions are answered, the task of successfully implementing an SDSS and the accompanying workflow or stakeholder process is socially and technically complex. However, with advances in both open source and proprietary software, the complexity of technically designing SDSS is becoming simpler.

**SEE ALSO:** Cartographic modeling; Decision analysis; Geodesign; Map algebra; Participatory modeling; Public-participation GIS; Spatial weights

### References


SPATIAL DECISION-SUPPORT SYSTEM


Further reading

Spatial econometrics

Xinyue Ye
Kent State University, USA

The small volume entitled *Spatial Econometrics* by Paelinck and Klaassen (1979) is considered to be the first comprehensive attempt to set the stage for spatial econometrics. Anselin (2010) also stresses the importance of the seminal book on spatial analysis by Berry and Marble (1968), which outlines the quantitative revolution in geography. These efforts are followed by the inspiring work of Anselin (1988), which is the first book in spatial econometrics formally establishing the theoretical and methodological framework. It warrants notice that the methods illustrated in Anselin (1988) mainly deal with cross-sectional data. In the past three decades, spatial econometrics has witnessed marvelous development in terms of theoretical thoughts, applications, and software development beyond the methodological toolbox of applied econometrics (Anselin and Rey 2012).

According to spatial econometrics, any individual or collective decisions would rely on the decisions made in the neighboring spatial units, indicating that an explicit accounting for spatial effects is required (Arbia 2011). Spatial effects are classified into two major aspects, namely, spatial dependence and spatial heterogeneity. Spatial econometrics lies at the heart of research in regional economics and quantitative geography, which calls for a set of tools to model the interaction and measure the spillover effects among a group of spatial units with particular attention to location, spatial structure, and spatial processes (Anselin 2010). Spatial econometrics has provided the methodological foundation for regional and urban econometric models. By analyzing recent developments, Anselin (2010) argues that spatial econometrics is among the mainstream of applied econometrics and social science methods, reflecting a fast-growing need to incorporate spatial effects into operational urban and regional models through spatially explicit variables. This growth can also be attributed to synergies between applied economics and geographic information science (GIScience), as well as to the adoption of spatial econometric methods owing to a growing list of software (Goodchild et al. 2000). The ultimate success of spatial econometrics will be determined by the extent to which domain scientists can access, adopt, and utilize them in order to incorporate additional advances in research inquiry for specific research questions (Arbia 2011). In other words, wider scientific and open source communities lead to a brighter future for spatial econometrics and interdisciplinary collaboration (Anselin 2010). Anselin and Rey (2012) also argue that replication mechanisms, collaborative environments, and computing power can fundamentally change the qualitative and quantitative nature of spatial econometric work. Spatial econometrics studies spatial interdependence both in a spatial perspective and in a more general socioeconomic sense. It provides new theoretical perspectives to empirically verify models of social and spatial interaction, such as peer effects, neighborhood effects, spatial spillovers, and network effects (Anselin 2010).
Interestingly, the role of space and geography in socioeconomic dynamics is also evidenced by the World Bank Development Report of 2009 devoted to economic geography, and the 2008 Nobel Prize in Economic Sciences awarded to Paul Krugman, in part in recognition of his achievements on spatial economics and new economic geography. The last two decades have witnessed a growing literature of empirical studies on socioeconomic dynamics (Anselin and Rey 2012). Some dominating applications include, but are not limited to, income growth, regional convergence, industrial concentration, regional labor markets, land use, housing, voting behavior, and political trends. The availability of large georeferenced event datasets along with high-performance computing technology has raised fundamental challenges and opportunities for mainstream social science research on whether and how these new trends in data and technology can be utilized to collect thematic information and help understand socioeconomic dynamics, in order to enhance and advance the research in social, economic, and behavior sciences in general (Arbia 2011). Such large-scale data analytics will stimulate the development of new computational models. In turn, these newly developed methods are adopted in real-world practice, forming a positive feedback loop. Enabling such space–time analysis over cyber-infrastructure is particularly useful for spatial econometrics, where this positive feedback loop will allow researchers and practitioners to test their models faster and scale them to a larger dataset, therefore producing more policy-relevant results. Such research bridges emerging advanced computing infrastructure and computing technology with socioeconomic analysis, which is among the burgeoning efforts seeking the cross-fertilization of multiple fast-growing interdisciplinary communities (Anselin and Rey 2012).

Definition and adoption

Taking into the account the fact that the spatial perspective is generally ignored by econometricians, Paelinck and Klaassen (1979) delineate five fundamental principles to guide the formulation of spatial econometric models in their pioneering work. These five rules include the role of spatial interdependence, the asymmetry in spatial relations, the explanatory factors in other locations, the differentiation between ex post and ex ante interaction, and the explicit modeling of space and topology. In particular, these rules highlight the importance of a realistic specification of space in econometric models. Indeed, this seems to be a natural progression resulting from the increasing availability of empirical data with locational information. Spatial econometrics goes on to develop into an emerging paradigm for spatial modeling of socioeconomic data, but with limited adoption among econometricians and economic geographers. The situation in the late 1980s and early 1990s was that applied spatial econometric work was constrained by the lack of appropriate spatial data analytical software (Anselin 2010). Though a few methods had been developed by individuals, they were not generally available to the wider community. In addition, extensive documentation and tutorials were required to gain sufficient adoption of methodological development and empirical practice, which were missed at this stage. Anselin (1988, 8) compares spatial econometrics to standard econometrics, stating that spatial economics models should deal with “the specific spatial aspects of data and models in regional science that preclude a straightforward application of standard econometric methods.” In his book, Anselin stresses the estimation and specification testing methods. Specifically, the domain of spatial econometrics is defined as “the collection of
techniques that deal with the peculiarities caused by space in the statistical analysis of regional science models” (Anselin 1988, 7). To promote spatial econometrics studies, much effort has been spent on developing spatial econometric scripts or a set of routines in the context of existing commercial statistical software. The availability of spatial econometric software is a key driver in the adoption and dissemination of spatial econometrics throughout the social sciences (Anselin 2010). In contrast to the abundance of mainstream econometric packages and tools, this limitation has been considered a major impediment for the adoption of a spatial perspective in empirical work. SpaceStat is the first truly self-contained spatial econometric analysis tool, used to illustrate the estimation methods and test statistics in Anselin (1988) based on GAUSS routines. Hence, the approaches from Anselin (1988) have become empirically feasible. The commercial sector has not lagged behind. This has been very quickly followed by the commercial specialized package S+Spatialstats. Since version 9.2, the Environmental Systems Research Institute’s (ESRI) ArcGIS geographic information system (GIS) software has contained the functionality for spatial autocorrelation analysis and geographically weighted regression.

In the early definitions of spatial econometrics, the scope of the field is largely limited to urban and regional modeling, which fails to anticipate the enormous growth of geospatial technologies and data analytics toward both conceptual and empirical studies in many social sciences (Anselin 2010). Anselin (2006) removes the limiting context of urban and regional modeling by broadening the scope of spatial econometrics from the cross-sectional setting to the space–time domain as “a subset of econometric methods that is concerned with spatial aspects present in cross-sectional and space–time observations. Variables related to location, distance and arrangement (topology) are treated explicitly in model specification; estimation; diagnostic checking and prediction” (Anselin 2006, 902). A set of specialized toolboxes have been yielded, with a variety of flexibility and scalability, within the commercial software environment (such as Matlab and Stata), as freestanding software, and through open source ventures such as Python and R (Bivand, Pebesma, and Gómez-Rubio 2008; Anselin and Rey 2012). Free spatial econometric software GeoDa consists of a graphical user interface allowing exploratory spatial data analysis and spatial regression to be carried out. The history of the open source movement is much younger, but its impact on the GIS world is impressive through the creation of a powerful community of collaborators and peer review resources. The Open Source Geospatial Foundation (OSGeo) projects that support spatial data handling have a large developer community with extensive collaborative activities, possibly due to the wide audience and publicly adopted Open Geospatial Consortium (OGC) standards. In comparison, spatial econometric analysis is often field- and data-specific. Therefore, the analysis routines are often written by domain scientists with specific scientific questions in mind and are not of a collaborative nature. To date, only a handful of open source spatial econometrics tools exist. The R open source statistical programming framework contains the spdep (spatial dependence: weighting schemes, statistics, and models) and sphet (estimation of spatial autoregressive models with and without heteroskedastic innovations) packages among others (Bivand, Pebesma, and Gómez–Rubio 2008). The Python Spatial Analysis Library (PySAL) deals with a range of spatial analysis and spatial econometric codes in the open source Python language (Rey and Anselin 2007).
A space–time perspective has been increasingly relevant in the investigations of socioeconomic phenomena. More and more data available for spatial econometric research are characterized with space–time labels, such as historical census data provided by governments, environmental observations recorded by sensors, or volunteered geographic information through social media. With the increasing availability of spatiotemporal data, a gradual shift has been identified from the analysis of synchronic cross-sectional spatial data to that of diachronic spatial panel data (Arbia 2011). Paralleling this is the development of the theoretical foundation and methods dealing with panel data. There is also a crucial need to better understand the fundamental processes behind the spatial and space–time correlation.

The evolution of the definition and scope of spatial econometrics over time reflects a major migration of the field from the margins in applied urban and regional economic analysis to the mainstream social sciences (Bivand, Pebesma, and Gómez-Rubio 2008; Anselin and Rey 2012). Arguably, early work on theoretical and methodological advances in the 1970s and 1980s is mainly published in regional science and quantitative geography journals. The situation has changed dramatically. Spatial econometrics has become a recognized subfield in many handbooks of econometrics. In recent years, the interest in spatial analysis in general and spatial econometric modeling in particular has witnessed an almost exponential growth and a rapidly expanding application in increasingly diverse scientific disciplines (Goodchild et al. 2000; Bivand, Pebesma, and Gómez-Rubio 2008). Consequently, the related publications are wide-ranging and appear across a variety of scientific journals. The presence of spatial econometrics in these mainstream compendia has facilitated its interdisciplinary popularity and acceptance in both theoretical development and empirical practice (Arbia 2011). In addition, the open source movement in spatial econometrics blurs the line between a scholar’s role as software user and as collaborative producer (Anselin and Rey 2012).

**Spatial dependence, spatial heterogeneity, and spatial models**

Ignoring the spatial relationship among geographic observations is analogous to disregarding the ordering of time-series data. While similar to correlation in the time domain, the distinct nature of spatial dependence requires a specialized set of techniques. There is a fundamental difference between spatial series and time series due to the feedback and simultaneity originating from spatial interaction (Arbia 2011). Importantly, there is no straightforward extension of time-series methods to two-dimensional geographic space. In addition, the spatial relationship is more difficult to define because it needs to be derived from a specific ordering, such as measures of potential, distance-decay functions, and spatial arrangement. Furthermore, the spatial dependence may decrease to zero beyond a given threshold of distance that separates two spatial units. For instance, it is expected that the spillover effect will be stronger on closer neighbors, and that it will vanish as distance increases.

In general, spatial effects are present in two ways: spatial dependence and spatial heterogeneity (Arbia 2011). In spatial dependence, the structure of the correlation between random variables at different locations is determined by the relative position of the observations in geographic space. Spatial dependence is a direct result of Tobler’s first law of geography, which states that “everything is related to everything else, but near things are more related than distant things.” Anselin (1988) argues that
this spatial dependence is the rule rather than the exception, which conflicts with the usual assumption of independence among observations. Spatial dependence can also originate from measurement error. Spatial dependence exists if the cross-sectional correlation is nonzero, following a certain spatial ordering.

In the spatial lag model, the spatial autocorrelation appears in the dependent variable, indicating that the spatial units influence one another; this is also the case in time series when the lagged variable influences the contemporaneous one. By contrast, when the spatial autocorrelation appears in the error term the model is known as a spatial error model. It indicates that the neighbors are subject to the same random shock. Spatially lagged variables can be included for the explanatory variables, leading to so-called spatial cross-regressive models. In addition, different combinations of these models are possible, yielding a rich array of spatially explicit models such as the spatial Durbin model (Anselin 2010).

The Global Moran’s I and Local Moran’s I are applied to measure spatial autocorrelation. Global spatial autocorrelation analysis mainly concerns the spatial distribution characteristics of the values for some property of spatial units in the entire region, and it can analyze the overall regional spatial associations and spatial differences by the calculation of the Global Moran’s I. The formula is shown in equation (1), where

\[
I(d) = \frac{\sum_{i}^{n} \sum_{j \neq i}^{n} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{s^2 \sum_{i}^{n} \sum_{j \neq i}^{n} w_{ij}}
\]  

\(s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad x_i\)

is the observation at spatial unit \(i\), \(\bar{x}\) is the arithmetic mean value of \(x_i\), and \(w_{ij}\) is the spatial weighting value. The value of the Global Moran’s I ranges from −1 to 1. A positive value for \(I(d)\) indicates spatial clustering of similar values, whereas a negative value indicates spatial clustering of dissimilar values between a focal region and its neighbors.

The local spatial autocorrelation analysis deals with the distribution pattern of the values for some property of spatial units in the heterogeneous space, and it can measure the local spatial correlation between each focal region and its surrounding regions. Local Moran’s I was first proposed by Anselin (1995) as a local indicator of spatial association. The formula for the Local Moran’s I is shown in equation (2), where \(x_i\) is the observation at spatial unit \(i\), \(W\) is the spatial weight matrix, \(w_{ij}\) indicates the interaction between spatial unit \(i\) and \(j\), shown in equation (3). If \(I_i\) is significantly greater than 0, it indicates that spatial inequality between the given location \(i\) and its surrounding regions is significantly small, and vice versa.

\[
I_i = \frac{(x_i - \bar{x})}{s^2} \sum_{j}^{n} W_{ij} (x_j - \bar{x})
\]

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

The specification of spatial heterogeneity can be classified as either discrete heterogeneity or continuous heterogeneity. The former consists of a prespecified set of spatial regimes, while model coefficients and other parameters might vary across regimes. Continuous heterogeneity specifies how the regression coefficients change over space, either following a predetermined functional form such as the spatial expansion method, or as determined by the data through a local estimation process such as geographically weighted
regression. The statistical literature specifies spatial heterogeneity as a special case of random coefficient variation. The treatment of spatial heterogeneity initially focuses on further elaboration of the expansion method. However, the most important development in this respect is the advent of geographically weighted regression as a way to model parameter variability across space.

An important distinguishing characteristic of spatial econometrics is the need to explicitly deal with the spatial structure of the observations: their spatial arrangement, or topology. Spatially lagged variables are weighted averages of observations for the neighbors of the focal spatial observation. An important aspect is the definition of neighbors, which is typically carried out through specification of a spatial weights matrix, the key concept in this regard. This aspect requires the use of spatial data structures that do not fit within the architecture of the standard econometric software. This neighbor structure not only forms the basis for the specification of spatial dependence in the econometric model but also leads to a series of computational issues that set spatial econometrics apart from traditional approaches (Anselin 1988).

As an elegant conceptual structure expressing the spatial extent of the influence of the focal spatial unit, the spatial weights matrix $W$ is of dimension $n \times n$. Each nonzero element $W_{ij}$ indicates whether the locations $i$ and $j$ are neighbors. It conceptually defines the underlying relationship between units of observations. In addition, it can also give a measure of relative strength of the interaction between these two units, which is named the “power” of the influence. The spatial data structure needs to be accessed in order to derive or extract the adjacency structure of polygons, or to compute various distance metrics from point coordinates. The neighbor structure of the observations can be mainly defined in two ways. The first type (contiguity-based spatial weight matrix) identifies a relationship based on shared borders. The second type (distance-based spatial weight matrix) establishes a relationship based on the distance separating observations. The selection of a spatial weight matrix can be either ad hoc or data-driven. The purpose of involving a spatial weight matrix is to take the spatial effects into account in the spatial analysis and modeling. In most cases, the selection of the spatial weight matrix is conducted before running the model, usually based on past experience. The spatial weight matrix is first introduced to treat the correlation between geographical units. The aim is to define how close or remote neighbors depend on each other, which has been extended to economic, social, cultural, and demographic distance or neighborliness (Villano, Fleming, and Moss, 2016).

Regarding contiguity-based spatial weight matrices, there are three types of contiguity: rook contiguity, bishop contiguity, and queen contiguity. Rook contiguity means that two polygons share a common border, while bishop contiguity indicates that two polygons share a

![Figure 1](image_url)  
**Figure 1** Contiguity-based spatial weight matrix.
common node. Queen contiguity is combination of rook and bishop contiguity. Figure 1 demonstrates how rook, bishop, and queen contiguity are defined. Lattice 1 has two neighbors, Lattice 2 and Lattice 3, in the context of rook contiguity, while Lattice 4 is the only neighbor of Lattice 1 if bishop contiguity is applied. Under queen contiguity, Lattice 1 is contiguous to all three other lattices. Regarding the power of influence in these contiguity structures, each neighboring lattice exerts the same influence. Such spatial weight matrices are widely used in applications where the polygons contain the attributes.

The specification of the contiguity-based spatial weight matrix is defined as shown in equation (4).

\[
W_{ij} = \begin{cases} 
1, & \text{if } i \text{ and } j \text{ are neighbors} \\
0, & \text{if } i \text{ and } j \text{ are not neighbors}
\end{cases} 
\tag{4}
\]

where \( W_{ij} \) is the \( i,j \)th element of \( W \).

Distance-based spatial weight matrices are mainly used in the context of point data or where the use of polygon data is not appropriate. A typical specification of the relationship in the \( W \) matrix is shown in equation (5), where \( d_{ij} \) indicates the distance between two points \( i \) and \( j \). The term \( \theta \) represents the power of the influence, which can be considered as the transportation friction. It can demonstrate a variety of spatial decay effects when the parameter \( \theta \) is assigned different values. The term \( a \) is the threshold to define the extent of influence. The selection of such a value largely depends on the extent of the conceptualized neighborhood of a focal unit. After the selection of spatial weight matrix, row standardization is needed. Such a process will make each row sum to one.

\[
W_{ij} = \frac{1}{(d_{ij})^\theta}, \quad \text{if } i \neq j \text{ and } d_{ij} < a \\
W_{ij} = 0, \quad \text{if } i \neq j \text{ and } d_{ij} \geq a, \quad \text{or if } i = j \tag{5}
\]

The general spatial model follows (Anselin 1988), as shown in equation (6), where \( Y \) represents a \( n \times 1 \) vector of dependent variables, \( \rho \) is the coefficient of the spatially lagged dependent variable, \( X \) is an \( n \times k \) matrix of independent variables, and \( \beta \) is a \( k \times 1 \) vector of parameters associated with \( X \). The \( \epsilon \) term represents an \( n \times 1 \) vector of disturbances, \( \lambda \) is the coefficient of the spatial autoregressive structure for the disturbance \( \epsilon \), and \( \mu \) is an \( n \times 1 \) vector of disturbances. The disturbance \( \mu \) is normally distributed with a general diagonal covariance matrix \( \Omega \). The diagonal elements serve as a function of exogenous variables with a constant term. \( Z \) is an \( n \times r \) matrix of exogenous variables with the constant term. \( \alpha \) is an \( r \times 1 \) vector of coefficients related to the terms.

\[
\begin{align*}
Y &= \rho W Y + X \beta + \epsilon \\
\epsilon &= (I - \lambda W)^{-1} \mu \\
\mu &\sim N(0, \Omega) \\
\Omega_{ii} &= h_i(Z \alpha) \tag{6}
\end{align*}
\]

A series of spatial models can be derived from this general spatial model and are demonstrated in Table 1.

### Comparative space–time dynamics

Comparison is at the heart of human behavior and reasoning. Comparative analysis is a method used to observe and interpret the world, including spatial and temporal characteristics of places. A powerful analytical framework for identifying spatial econometrics research gaps and frontiers is fundamental to the comparative study of spatiotemporal phenomena throughout the social
Table 1  A taxonomy of spatial models.

<table>
<thead>
<tr>
<th>Parameter values</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ρ = 0, λ = 0, α = 0</td>
<td>classic linear regression model</td>
</tr>
<tr>
<td>λ = 0, α = 0</td>
<td>spatial autoregressive model (SAR)</td>
</tr>
<tr>
<td>ρ = 0, α = 0</td>
<td>spatial error model (SEM)</td>
</tr>
<tr>
<td>α = 0</td>
<td>general spatial model (SAC)</td>
</tr>
</tbody>
</table>

sciences. Increasingly easier development processes facilitate the explosion of related routines with powerful scripting language environments such as R and Python (Bivand, Pebesma and Gómez-Rubio 2008; Rey 2009). However, duplicated efforts exist and many critical gaps remain unexplored. Ye and Rey (2013) argue that methodological duplicates and gaps exist due to the lack of systematic exploration of the space–time dataset. The multiple dimensions and scales of socioeconomic dynamics pose numerous challenges for the application and evaluation of public policies in the comparative context. However, the definitions of the unit of analysis and the unit of observation should be distinguished before the structure of the space–time dataset is characterized.

In addition to their relationship in space, things near in time or in statistical distribution are more related than distant things. There is a growing consensus that all dimensions of the data should be handled together instead of them being treated independently (Anselin 2012). Hence, ignoring the interdependence across space, time, and statistical distribution leads to overlooking many possible interactions and might cause biases in the analytical results (Rey and Ye 2010). To reveal these relationships, the distributions of space, time, and attributes should be treated in the context in which a measurement is made, instead of specifying a single space or time as the context. The “distribution” in space (the dimension of space) refers to the spatial distribution of attributes while the “distribution” of attributes (the dimension of statistical distribution) implies the arrangement of attributes showing their observed or theoretical frequency of occurrence. In addition, the “distribution” of time (the dimension of time) signifies the temporal trend of attributes (Rey and Ye 2010).

Besides the dimensions in analyzing data, it is also important to recognize the issue of scales of data and analysis. Four scales are taken into consideration. The unit of analysis at the individual scale signifies the geographical location of an attribute (A1 in Table 2), the temporal label of an attribute (A5 in Table 2), or the rank of an attribute (A9 in Table 2). The unit of analysis at the local scale explores a group of units which is formed by the focal observation and its neighboring observations in one of these three dimensions. A focal state and its neighboring states, for example, can be considered as a unit of analysis from the perspective of the spatial dimension (distribution) at the local scale (A2 in Table 2). A focal year, the previous year, and the following year can be considered as a unit of analysis from the perspective of the temporal dimension at the local scale (A6 in Table 2). A focal rank and the two immediate higher/lower ranks can be considered as a unit of analysis from the perspective of the statistical dimension (distribution) at the local scale (A10 in Table 2).

A mesoscale analysis studies a group of entities which shares similar feature(s) in spatial, temporal, or statistical distributions. In other words, the local-scale analysis differs from the mesoscale analysis in how a subset of data is retrieved for analysis. The former emphasizes that the rest of the subset are “near things” to the focal element while the latter does not have this constraint. Hence, the latter usually has a
larger subset (larger in space and lengthier in time) as the unit of analysis than the former does. The spatial distribution of rich states, for example, can be considered as a unit of analysis from the perspective of the spatial dimension (distribution) at the mesoscale (A3 in Table 2). All the years since a policy has been implemented can be considered as a unit of analysis from the perspective of the temporal dimension at the mesoscale (A7 in Table 2). An income quartile can be considered as a unit of analysis from the perspective of the statistical dimension (distribution) at the mesoscale (A11 in Table 2).

The analysis at the global scale examines the distributions of all the regions, times, or attributes. Spatial distribution of all the incomes, for example, can be considered as a unit of analysis from the perspective of the spatial dimension (distribution) at the global scale (A4 in Table 2); all the years can be considered as a unit of analysis based on the temporal dimension at the global scale because space–time dynamics research is very sensitive to the selected starting and ending years (A8 in Table 2); the statistical distribution of all the incomes can be considered as a unit of analysis based on the statistical dimension (distribution) at the global scale (A12 in Table 2).

Limiting attention to only one of these dimensions or scales may result in a misguided or partial understanding of economic growth dynamics.

The unit of analysis is the major entity that is being analyzed in the research, while the unit of observation is the basic entity described by the data. The unit of analysis is the “what” that is being studied, and is designed by the researcher. Nevertheless, the unit of observation is determined by the way the dataset is collected, which cannot be fully controlled by the researcher. In most studies, the difference between the unit of analysis and the unit of observation is not emphasized, but it is important to recognize this difference for comparative space–time analysis. Because the unit of analysis involves the issues of scales and levels of aggregation of data, it is critical in designing data analysis tasks. Various spatial partition schemes often lead to different types of units of analysis, which in turn generate different perspectives of the same data. Hence, it is valuable to consider all possible perspectives before formulating research questions. All possible temporal configurations should also be considered. Monthly unemployment counts (the unit of observation), for instance, can be aggregated into quarterly or yearly periods.

Table 2  Examples for unit of analysis.

<table>
<thead>
<tr>
<th>Levels</th>
<th>Individual</th>
<th>Local</th>
<th>Meso</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributions</td>
<td>Spatial</td>
<td>California (A1)</td>
<td>California and its neighboring states (A2)</td>
<td>Spatial distribution of rich states (A3)</td>
</tr>
<tr>
<td>Statistical</td>
<td>No. 3 income (A9)</td>
<td>Nos 2, 3, and 4 incomes (A10)</td>
<td>The first income quartile (A11)</td>
<td>Statistical distribution of all the state incomes (A12)</td>
</tr>
</tbody>
</table>
Many types of units of analysis can be generated when both spatial and temporal partition schemes are considered. Unemployment issues, for instance, can be analyzed at the county level using monthly counts, or at the state level using yearly counts.

Most empirical studies are motivated by only a few well-defined research questions. Then, the researcher chooses the appropriate analytical methods while also obtaining the data needed for application of the methods. Only at the end of the process does the analyst interpret and evaluate the results. These traditional research methods and procedures, however, tend not to be very useful in systematically revealing patterns and trends in complicated space–time datasets. This is because the intermediate stages in a process are not observed or analyzed. Typically, the analyst has to get acquainted with the data before formulating novel questions. Rey and Ye (2010) develop a framework based on Table 2 that leads to a general task topology for socioeconomic data by integrating spatial, temporal, and statistical distributions at individual, local, meso-, and global scales. One research question can thus engender many follow-up research questions. This framework allows the behavior of a dynamic system to be reconstructed from a group of units of analysis. The key aspect of the work is to integrate the three dimensions of a space–time dataset into a four-scale environment. Taxonomy of methods can then be built by combining any two approaches at any two scales.

Big spatiotemporal data and computation

A growing research community of spatiotemporally integrated social science has emerged aimed at analyzing spatial patterns and geographical dynamics, and maximizing the potential of massive data to improve human wellbeing. Big spatiotemporal socioeconomic data will become increasingly available in the near future, allowing the possibility for individuals’ space–time economic behavior in space and time to be modeled and for the results of such models to be used to gain information about economic trends across spatial scales (Arbia 2011). For instance, the US Census Bureau’s Longitudinal Business Database provides annual records for every private sector establishment from 1976, which involves about 4 million establishments and 70 million employees each year. However, as the space–time data accumulate, the rich details of space–time complexity in social science remain largely unexplored because of many binding constraints for scientific advancement, such as the challenge of intensities of data computing and very large georeferenced dynamic databases (Anselin and Rey 2012). Computation has been playing a central role in the development of econometrics, in terms of both theory and applications (Bivand, Pebesma, and Gómez-Rubio 2008). In addition, the unique characteristics of spatial data bring extra challenges, such as data summarization, numerical optimization, simulation-based estimation, and nonparametric/semiparametric estimation (Anselin and Rey 2012). The following notions will become unrealistic in the big data setting: the standard sample–population paradigm, spatial stochastic process, spatial stationarity, and the statistical significance test (Anselin 2010). Many current space–time statistics and simulations can only deal with a limited amount of data due to computation constraints. Hence, more advanced computational techniques are needed to handle the complex large-scale space–time interactions through deploying modern accelerator technology and hybrid computer systems.

Cyberinfrastructure has enabled both physical and social scientists to collaboratively tackle sophisticated problems across various
domains. Many studies have been undertaken to empower space–time analytical models with high-performance computing techniques. These methods generally take advantage of multicore central processing units (CPUs), many-core graphical processing units (GPUs), or Intel® Many Integrated Core (MIC) architecture. In social science research, quite a few studies have been conducted to speed up agent-based modeling and land-use planning. The cyberinfrastructure poses challenges to the existing spatial econometric packages largely relying on the single-CPU realm. Anselin and Rey (2012) envisage a spatial econometrics workbench as a framework for supporting spatial econometric research in the cyberscience era, taking into account that cyberinfrastructure offers the potential of lowering the hardware and software costs facing the applied spatial econometrician. Timely analysis using spatial econometrics models will open up a rich empirical context for social sciences and policy interventions, and will be critical for quantifying the impact of socioeconomic policies.

However, there are no easy ways to directly equip these existing spatial econometric models, estimators, algorithms, and data structures with high-performance computing capacities in the cyberinfrastructure, because most of the analytical and processing functions need to be rethought and rewritten. There is a critical need to examine their particular computational requirements and how they may map to specific types of parallelization. At the same time, the characteristics of a problem determine to what extent and how conveniently a certain algorithm can be parallelized (Ye and Shi 2013). New algorithms have begun to use the new features in computational languages supporting parallelization (such as MatlabMPI, Stata/MP, and the parallel packages in R) in the context of rapidly changing computing technology, such as distributed computing, cloud computing, and the use of handheld devices (Anselin and Rey 2012). In addition to technical solutions of embracing high-performance computing, a major challenge is to deal with the conceptual questions raised by massive datasets and the complex dynamics underlying the existence of spatial interaction (Anselin and Rey 2012).

**SEE ALSO:** Economic geography; Geographic information science; Spatial analysis

**References**


SPATIAL ECONOMETRICS


Further reading

Spatial epidemiology
Daniel J. Exeter
The University of Auckland, New Zealand

Spatial epidemiology refers to research that analyzes and describes the geographical distribution of health and wellbeing, in relation to sociodemographic, environmental, and behavioral risk factors (Elliot and Wartenberg 2004). Epidemiologists aim to better understand the social and environmental patterns and processes of diseases. In addition, epidemiologists are particularly interested in the distribution of disease events over time, and investigate such patterns across large geographic regions such as administrative districts or counties. By contrast, many spatial epidemiologists focus their attention on small areas, whose populations are less than 5000 residents, in an attempt to identify the spatial distribution of disease, exposure to toxins, and health outcomes and disease clusters. Spatial epidemiologists maximize the use of population, environmental, and disease data represented as points or areas in their analysis.

There are many ways to classify the techniques used in spatial epidemiology. The most common, however, are in terms of broad applications: disease mapping; geographical correlation studies; and clusters, clustering, and surveillance, which are addressed in turn below. Methods used by conventional epidemiologists and statisticians involve performing analyses that make the assumption that results obtained for one area are independent of other areas in the study (this being the converse to Tobler's First Law of Geography, whereby geographic entities are all related, but entities in close proximity to each other are more related than distant entities; Tobler 1970). In the grouping of the spatial epidemiological approaches discussed below, there is also a general transition from the use of conventional epidemiological techniques (e.g., age-standardized rates) to more advanced approaches (e.g., local spatial clustering) that are designed to respect Tobler's First Law of Geography.

Disease mapping

Early examples of spatial epidemiology include John Snow’s dot maps that were developed to investigate the cholera outbreak in London (UK). Snow mapped every case of cholera at the household level upon the local street map and there were more cases in households that were serviced by a contaminated water source. This rather primitive dot map led to the discovery in the 1850s that cholera was a waterborne illness. Similarly, Alfred Haviland’s mortality maps of heart disease and cancer used routine data sources such as the mortality statistics and the census populations in the 1870s, extending the influential work of early pioneers such as William Farr. Today, disease maps represent a major stream of spatial epidemiological research, displaying standardized rates of disease to adjust for differences in population structure of residents within regions included on the map. Disease maps can rapidly highlight geographical patterns that would otherwise be hidden in a table, and are commonly used for hypothesis generating and as the first step in exploratory spatial data analysis. Disease mapping can also
be used to show results from statistical modeling, demonstrating whether disease rates in a given area are statistically significant compared to the average rate for the study area (i.e., the mean), or by grouping disease rates according to standard deviations from the mean. Moreover, the increased availability of commercial and open-sourced geographic information system (GIS) software has led to the rapid development of disease mapping at the local (e.g., small census areas or block-groups in the United States), regional (county), and national (state) or international scales.

Researchers mapping spatial patterns of disease for small areas are confronted with a number of challenges, including cartographic design, the modifiable area unit problem, and also the “small number” problem, where some places have no, or few, disease events and/or a low population at risk of contracting the disease. This is a substantial problem in places beyond the urban fringes, which raises the subsequent problem associated with thematic (choropleth) mapping: localities with smaller populations often have the largest areas on the map, thus distracting the user from potentially more important results in the more populous communities. Disease rates can be very unstable when calculated for small geographical areas or relatively rare diseases. Overcoming the small number problem may be achieved by increasing the size of the regions being mapped, for example, from census areas to electoral districts. Alternatively, spatial epidemiologists may choose a spatial smoothing process such as Kernel Density Estimation (KDE) or a statistical smoothing approach (e.g., Empirical Bayes Estimation) to reduce the extreme values. Kernel density estimation transforms point data (e.g., an outbreak of measles) to a continuous data surface of disease risk. Varying the size of the search area or bandwidth used in the KDE process results in different depictions of risk. Thus an ideal representation is a surface that demonstrates spatial variability that exists within the data with precision while removing extreme values or “noise” from the dataset. Empirical Bayes Estimation is used to smooth areal data (e.g., heart disease mortality rates by region) by removing some of the random error in the rates to produce smoothed relative risks of disease for each area, and has the greatest impact (i.e., data are smoothed the most) in areas with a low population at risk.

Geographic correlation studies

Disease maps can present a visual summary of disease patterns, but there is often a need to investigate the patterns further and, if possible, to identify potential causes or circumstances underlying those patterns. Ideally, we would have individual-level measurements of exposure to environmental determinants of health (e.g., exposure to air pollution) in addition to disease outcomes and individual risk factors (smoking, drinking) for all patients in a population, but such information is often difficult, impractical, or impossible to obtain for investigation. Therefore spatial epidemiologists compare the spatial distribution of disease with the geographical variations in environmental or lifestyle exposures. These studies are often referred to as geographic correlation studies and this literature aims to develop or test hypotheses regarding the etiology (causes) of disease. By way of example, the INTERSUN project (World Health Organization 2003) used a geographic correlation study to determine the relationship between solar ultraviolet (UV) irradiance and the consequences of UV radiation exposure in terms of skin, ocular, and immunological diseases.

The use of GIS in public health research and by various agencies in the past decade has contributed to the considerable growth in this
SPATIAL EPIDEMIOLOGY

field of spatial epidemiology. As a result, the measurement of exposure variables such as access to food outlets or other community resources has been used to explore associations with obesity and other chronic illnesses. Broadly, these studies have reported that with increased availability of health-promoting facilities within close proximity to an individual's home, the population is less likely to be obese. The widespread availability of digital information for populations and the ease with which individual-level information can be obtained via Global Positioning System (GPS) sensors and other smart technologies have also facilitated the use of multilevel modeling techniques to investigate geographic correlations between exposures and health outcomes.

Spatial clustering, disease clusters, and surveillance

In spatial epidemiology, a disease is said to be “clustered” if spatial variation in risk remains after controlling for known confounders (Pfeiffer et al. 2008). There is a multitude of cluster detection techniques available for both point and areal data, and these can be categorized according to the nature of the clusters identified. Global clustering methods identify the presence of a cluster in a study area, but do not identify the specific location(s) of the clusters. Rather, methods such as Ripley’s bivariate k-function use point data and provide a single value indicating the degree of clustering, which can then be tested for statistical significance. Local clustering methods are commonly referred to as hotspot analysis tools, and use point data to identify the location and geographical extent of one or more clusters in a study region. Focused-cluster detection is a form of local analysis in which local clustering techniques are used to detect clusters around a specific point, such as a cancer cluster around an industrial pollutant site. The Geographical Analysis Machine (GAM) was developed in the 1980s (Openshaw et al. 1987) and was used to seek clusters of leukemia in the United Kingdom. The GAM used a fine grid over the study area upon which circles of varying radii were generated. For each circle, the GAM calculated the observed and expected number of cases. Circles with higher-than-expected values were retained for visual inspection, but a significant limitation of the GAM was its multiple testing. Kulldorff’s spatial scan statistic, packaged as SaTScan™, extended the GAM to overcome the multiple-testing limitation and has been extensively used in many aspects of human and animal health research to detect local clusters of disease. Currently, SaTScan is the most common local clustering approach used in the spatial analysis of disease data published in epidemiology journals (Auchincloss et al. 2012), and can be used to investigate temporal, spatial, and spatiotemporal clusters. In addition, the space–time permutation model in SaTScan may also be used in public health offices as a surveillance tool. By linking into routine disease notification registers, the software can prospectively detect an elevated risk or occurrence of disease and trigger early warning surveillance systems.

Moran’s I and Geary’s C statistics are commonly used “global” clustering techniques for use with areal data. These statistics measure the spatial autocorrelation between two variables and use a matrix of nearest neighbors defined in terms of adjacency or distance. This matrix is used to ensure that closest neighbors are assigned greater weight than more distant neighbors. In a similar way to the k-function test for point data, these provide an indication of spatial autocorrelation for the dataset overall. Local clustering techniques for areal data include Getis and Ord G* statistic, and the local Moran’s I statistic is commonly referred to as Local Indicators of
Spatial Association (LISA), which detect clusters by calculating the association for a value (i.e., disease rate) for one area and values in its adjacent areas.

Regardless of the methods used, the quality of the data used in analyses is vital. Spatial epidemiologists must address the challenge of depicting spatial processes as accurately as possible, while ensuring that the confidentiality of patients is preserved. This concern is especially the case for spatial point-pattern analyses. Aggregation of the data from individual events to administrative units poses problems related to the scale and aggregation effects of the modifiable areal unit problem (MAUP), in addition to the risk of the ecological fallacy.

The recent growth of “big data” (a collection of data too large for storage in traditional databases or analysis using conventional data processing software) and the increasing availability of spatially referenced data present a bright future for spatial epidemiology. The development of electronic health records may provide opportunities for obtaining more accurate residential histories of (anonymized) patients and their lifestyle risk factors of disease. These may then be used to enhance explanatory variables in disease clustering and the validity of results from geographical correlation studies.

**SEE ALSO:** Big data; Cartographic design; Exploratory spatial data analysis; Measuring spatial dependence; Modifiable areal unit problem; Spatial analysis

**References**


**Further reading**


Spatial feature classes

E. Lynn Usery  
US Geological Survey, USA

Spatial feature classes are the basic result of human cognition and categorization of geographic phenomena. Categorization and classification are basic to human knowledge and understanding and form a basis of scientific investigation. For geographical phenomena, humans categorize the world into identifiable, discrete entities whether these entities occur naturally or are human-constructed. Spatial feature classes are defined as categories of geographic phenomena that are uniquely identifiable and distinct. Based on cognitive category theory, this categorization typically occurs cognitively at a most basic level (Rosch et al. 1976) and for geographic phenomena that level corresponds to features commonly represented on cartographic abstractions of geographic space or maps (Usery 1993). Thus, spatial feature classes, such as roads, streams, hills, cities, and structures, are a basic level of human cognition that cartographers have used for thousands of years in constructing maps that communicate effectively about geographical phenomena.

Digital representation of geographic phenomena also requires use of this basic level of abstraction and fortunately, as geographic information systems (GIS) were developed, the basic features traditionally used on maps were replicated as digital representations. It is the purpose of this entry to present the concepts and history of spatial feature classes, which form a part of the representation and theory of geographic information. The second section, “History,” provides a history of spatial feature classes from concepts of geographic regions and cartographic generalization processes (particularly selection) to the use of points, lines, and polygons as feature classes in commercial GIS software. The third section, “Standardization of spatial feature classes,” describes attempts at standardization of spatial feature classes based on the design of feature lists and feature attribute catalogs, and examines specific standards that were developed, some of which are still in use. The fourth section, “A theory of spatial feature classes,” documents research to develop a theory of geographical phenomena or spatial feature classes. The fifth section, “The rise of ontology in spatial feature classes and feature class semantics,” details the rise of ontological concerns in geographic information science and presents a basis for feature class semantics dependent on feature identity, attributes, and relationships to other features and provides a basis for the current (2016) geographic feature vocabularies being developed for the Semantic Web. A final section draws conclusions and posits a future of spatial feature classes in relation to changing technology and new approaches to geospatial data collection, storage, and distribution.

History

The concept of a class of spatial features spans the temporal range of geography, the earliest attempts being to define a basic set of entities or phenomena in geographic space. Dividing the Earth into areas of water and land develops a coarse set of spatial feature classes with members of oceans and continents, respectively. Further division leads to feature classes of smaller spatial extents such as seas, countries, mountains,
valleys, and others. Although such attempts at categorization of geographical phenomena date from earlier times, near the beginning of the twentieth century geographers made concerted efforts to define a set of natural geographic regions (that is, features) but because these regions were cognitively generated, much controversy ensued. Through the work on definition of geographic regions, geographers developed the concept of a region that is neither an object nor self-determining, but rather an intellectual concept, an entity for the purpose of thought, created by the selection of characteristics that are relevant to an area of interest and by the disregard of all characteristics considered to be irrelevant (Whittlesey 1954). The selection in the statement is equivalent to selection in the generalization process of creating a map at a particular scale, that is, one selects features of interest for a particular scale and disregards others. Since the term “feature,” as used in current geographic information science (GIScience) and as used in this entry, is defined as “a geographical entity and its object representation” or “a digital representation of a real world entity or an abstraction of the real world,” and possesses “a common set of attributes and relationships,” (Guptill et al. 1990), it corresponds exactly to the intellectual concept of a geographical region. Although regions are commonly thought of as areal phenomena, the strict definition allows point, line, area, volume, or pixel and voxel representations of the geometry as with features. Thus, the current concept of a feature class in geographic information science has long roots in geography through its association with the region concept. More importantly, feature classes are intellectually generated based on physical geographic characteristics and human cognition, which involves classification and category concepts as explored in cognitive psychology and GIScience (Usery 1993).

Since the early models of geographic phenomena and space were created from the map model, spatial feature classes abstracted basic geometry from that model in the form of points, lines, and polygons for geographic phenomena, and picture elements (pixels) for geographic space. Feature classes were represented either as layers of point, line, or area objects or as a raster layer of pixels. These are the fundamental processing components of GIS. Thus, for the first 30 years of GIS, these geometric elements formed the basis of spatial feature classes, and commercial software implementations of GIS termed these basic geometric elements as features. These geometric elements rarely corresponded to real geographic entities as cognized and named by humans. They provided a basis for geometric and analytic processing that supported multifarious applications, such as site location, land-use planning, corridor analysis, land management and monitoring, and many others.

Standardization of spatial feature classes

As the development of spatial databases became a primary activity associated with GIS, there was a need to catalog spatial feature classes as well as their associated attributes and relationships, although little was achieved on relationships for many years. Since primary mapping activities rested with national mapping organizations, such as the US Geological Survey (USGS), responsible for topographic mapping in the United States, the Defense Mapping Agency (DMA, now the National Geospatial-Intelligence Agency (NGA)), responsible for providing maps of the rest of the world to the US Department of Defense, the National Oceanic and Atmospheric Administration (NOAA), responsible for coastal mapping and charting, the US Census Bureau, responsible for maps to support the decennial
census, and others, in the 1970s these organiza-
tions began developing digital cartographic
databases primarily from existing maps as a
base to support GIS analysis and operations.
Key components of those databases were the
geographical entities selected and represented,
albeit as geometric elements of points, lines, and
areas, which became feature classes.

The early results of these efforts led to car-
tographic feature coding (Wood and Douglas
1984) with extensive lists of point, line, and area
objects annotated as geographic features. For
example, the USGS created Digital Line Graph
(DLG) data with more than 200 feature codes or
feature types that represent various classes of data
such as roads, railroads, streams, and others with a
seven-digit numeric code, the first three digits of
which are the major code for the class of feature,
and the last four digits a minor code denoting the
attributes associated with the feature. It should
be noted that this early use of numeric codes
rather than actual names was primarily due to the
technological limitations of computer program-
ing in languages, such as FORTRAN, that
supported numeric data processing much better
than alphabetic names. As a case in point, a road
symbolized on a USGS 1:24 000-scale topo-
graphic map as a “Primary route, class 1, symbol
undivided” is coded as 1700201. The 170 is the
major code for the class of roads and the 0201
is the minor code specifying the exact symbolic
attributes of the cartographic feature. Most map-
ing organizations developed catalogs of feature
codes and, with the associated geometric data
that were digitized from maps and annotated
with these codes, constructed base geographic
data layers for the GIS software available at
the time. Examples of such cartographic fea-
ture codes include the USGS DLG, the DMA
Feature Attribute Coding Standard, Canada’s
Department of Energy Mines and Resources
National Topographic Database, South Africa’s
Proposed Standard for Exchange of Digital
Geo-Referenced Data, and many others.

Because many of these catalogs of feature codes
were cartographic or geometric in nature and
lacked exact correspondence to real geographical
phenomena, researchers in GIS in the 1980s and
1990s began to develop alternative models of
spatial feature classes. The development of the
Topologically Integrated Geographic Encoding
and Referencing (TIGER) system by the Cen-
sus Bureau created one of the first systems to
code spatial feature classes, such as census tracts
and minor civil divisions, which correspond
to enumeration units rather than cartographic
representations or geometric elements. It was
also during this time that theoretical work
to determine human understanding of spatial
feature classes began.

The USGS developed Digital Line Graph–
Enhanced (DLG-E), where the enhancement
was representation of real-world spatial feature
classes (Guptill et al. 1990). A major part of the
DLG-E research effort was examining topo-
graphic maps and determining feature classes,
developing rigorous definitions of the spatial
feature classes, then associating the attributes and
relationships of those features to provide a set
of objects that could be used in GIS. The final
developed feature list for DLG-E appears in its
entirety in Guptill et al. (1990). Simultaneous to
the research around DLG-E, and with some of
the same researchers and a similar methodology
for defining features (Rugg 1988), the Spatial
Data Transfer Standard (SDTS) was being devel-
oped and promoted. SDTS was created from
the merger of the Federal Geographic Exchange
Format (FGEF), which was designed to support
exchange of geographical data among Federal
agencies and developed by the Federal Inter-
agency Coordinating Committee for Digital
Cartography, with the standard prepared by the
National Committee for Digital Cartographic
Spatial Feature Classes

Data Standards (NCDCDS). NCDCDS was led by the American Congress on Surveying and Mapping operating on grant funding from the USGS. The SDTS specification became a standard under International Standards Organization (ISO) 2011 and originally was published in *The American Cartographer*, in 1988, with an implementation specification published in *Cartography and Geographic Information Systems* (Morrison and Wortman 1992). The final version of SDTS is available from the USGS (2014). SDTS includes 202 different entity types (spatial feature classes) and accounts for 1203 included terms, which are alternate methods of naming these entity types. For example, “access way” and “alley” are included terms of the entity type Road.

Although significant resources had been expended to develop SDTS and define a common set of spatial feature classes, there was little use of the standard that had been developed. The complexity of SDTS was daunting to the casual user and dedicated professional GIS users opted for simpler de facto standards based on vendor domination. For example, by the late 1990s, the dominant transfer format for spatial data of vectors was Esri’s shapefile, originally developed as a display format for ArcView GIS. The primary transfer format for raster data was Tagged Interchange File Format (TIFF). Although SDTS failed to be adopted as a universal standard, its development led to the recognition of the need for a theory of geographic information in which spatial features held a prominent role. Research on such a theory had begun in the 1970s and 1980s and gained significant momentum in the 1990s with the recognition of GIScience as a separate discipline. As a discipline, GIScience needed its own subject matter and that subject matter is space carved into spatial feature classes. Using theory as a basis, the design and development of spatial feature classes moved toward ontology and the definition of a geospatial vocabulary for use in semantics processing, particularly for the Semantic Web.

A theory of spatial feature classes

Research in spatial data models and the fundamental basis of human cognition of spatial phenomena led the development of a theory of geographic phenomena and spatial feature classes (Mark 1993; Usery 1993). Berry (1964) developed a conceptual framework of a geographical matrix to account for all possible types of spatial analysis. In his matrix, the columns represent locations and the rows represent measurement of a variable or theme. Time was added as a perpendicular dimension to the rows and columns. With such a matrix, each cell value corresponds to a geographical fact, the occurrence of a single attribute at a single location at a specific time. This concept of a geographical fact corresponds to the description of a geoatom provided by Goodchild, Yuan, and Cova (2007) exploring a general theory of geographic representation in GIS. The basic dimensions of the geographical matrix and of spatial feature classes are space, theme, and time (Table 1), of which one of the three dimensions is fixed, a second is controlled, and the third is measured. This view is supported in GIS since most data sources are maps and maps usually fix the time, control the theme, and vary the spatial location. The concept of the geographical matrix and the analysis methods it presented became a cornerstone of GIS design, development, and implementation.

Goodchild (1987) examined spatial analysis concepts based on data models for point, line, area, and grid cell objects represented in GIS software packages at the time, and the types of analysis he defined can all be mapped to the types defined by Berry (1964). These observations
concerning the fundamental basis of geographic phenomena, and the possible types of analytical operations, led to a set of requirements for any theory of geographic phenomena and any implementation of such a theory in terms of analytical needs. These requirements are that any theory must account for geographical entities (features) and their attributes and relationships along the three dimensions, space, theme, and time, and allow for implementation of the types of analyses observed by Berry (1964) and Goodchild (1987).

Toward generating such a theory, defining spatial feature classes becomes a major consideration. Such classes are defined with attributes and relationships along the three dimensions of space, theme, and time (Table 1, Usery 1996). Yuan (1996) preferred the term “domain” to “dimension” and developed a three-domain model. Spatial feature classes must also be composed of one or more geographical facts or geoatoms. This theory, supported by work in cognitive psychology on category theory, in geography by work on geographic regions and spatial analysis, and in cartography by work on abstraction and generalization, provides the basis for the development of a multidimensional representation of geographic phenomena.

It has been shown that features commonly defined as mappable geographic entities fit category theory’s basic level of cognition for geographic phenomena (Rosch et al. 1976; Usery 1993; 1996) and that formalization of these entities in systems design can be achieved through interaction concepts such as metaphor and object semantics. Theoretical work on human cognition indicates the use of metaphor is an appropriate mechanism to relate geographic knowledge. Alternative conceptions of space, including Euclidean, physical, sensorimotor, perceptual, and cognitive, can be distinguished and represented as algebraic group structures. Interface design requires formalization of interaction concepts, one of which is metaphor. Metaphor domains are formalized by algebras, morphisms, and image schemas. Usery (1993) examined category theory as a basis of defining basic level geographic features which correspond to mappable features, such as roads, streams, hills, and valleys.

While components of a basic theory of geographic entities or features and human understanding through metaphor and algebraic concepts are well illustrated, the theory is incomplete and the path to implementation is not clear. One of the missing parts of the theory is how one feature relates to another (the relationships shown in Table 1). These relations are extremely important and often form the basis of geographic applications. Early in the study of computer-based representation of geographic phenomena, Freeman (1975) defined 13 different spatial relationships: left of, right of, beside, above, below, behind, in front of, near, far, touching, between, inside, and outside. Peuquet (1986) refined Freeman’s work and stated that all spatial relations

**Table 1** Attributes and relationships of fundamental geographic dimensions.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Attribute</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space</td>
<td>$\phi, \lambda, \zeta$, {Point, line, area, volume}, Pixel, voxel, …</td>
<td>Topology, direction, distance, …</td>
</tr>
<tr>
<td>Theme</td>
<td>Color, shape, size, pH, …</td>
<td>Topology, is_a, kind_of, part_of…</td>
</tr>
<tr>
<td>Time</td>
<td>Duration, date, period, …</td>
<td>Topology, was_a, will_be, …</td>
</tr>
</tbody>
</table>

SPATIAL FEATURE CLASSES

can be defined in terms of three primitives: direction, distance, and Boolean set operations (and, or, not). She further developed some implementation methods for these set operations using a raster GIS database. More recent work has focused on defining the fundamental bases of spatial relations and the operations allowed.

The thematic and temporal relationships of Table 1 are not addressed directly and receive attention only when spatial concepts, such as topology, can be applied to the thematic and temporal dimensions. The definition of geographic entities or features requires definition of all attributes and relationships of the entity. With the development of the World Wide Web and the Semantic Web, with its inherent semantic structure and logic and inference capabilities, the definition of spatial feature classes has become an exercise in ontology design, development, and engineering.

The rise of ontology in spatial feature classes and feature class semantics

Spatial feature classes are human conceptualizations and, whether occurring naturally, such as hills, valleys, streams, and mountains, or as a result of human engineering, such as roads and buildings, it is the human mind that gives meaning to the spatial feature class. Spatial feature classes are determined in a human cultural context and cannot be separated from human culture and associations. The study of spatial feature classes is an examination of what exists in reality and thus leads to a study of ontology of geographic information. GIScience researchers began to examine ontological concepts of geospatial phenomena in a quest to establish robust theoretical foundations for geographic information science. Whereas the early work with identifying spatial feature classes for use in GIS and the developments of feature lists and standards, such as SDTS, were really ontological efforts developed by data gathering organizations for operational use, the earliest conceptual and theoretical work on spatial feature classes identified specifically as ontology in the GIScience discipline occurred in the late 1990s. For example, Smith and Mark (2001) empirically tested subjects to identify geographic categories. Among their findings, the term geographic feature conveys concepts of mountains, hills, valleys, and other natural physical entities, whereas geographic object is most often associated with a map, globe, compass, or atlas. For human-engineered geographic features, such as roads, bridges, boundaries, and cities, the common association is with something shown on a map, rather than the term “geographic feature.”

Using ontology to construct a set of spatial feature classes has become the norm in GIScience. The proliferation of ontology tools and the existence of the Semantic Web and linked data activities since 2001 led to the use of ontology and semantics as methods for defining and accessing spatial feature classes. In the post-2001 timeframe, the Ordnance Survey developed an ontology which includes topographic and administrative spatial feature classes. The USGS has also developed an ontology for The National Map. The primary activity of Geospatial Vocabulary Camps (GeoVoCamps) is to crowsource or collaboratively build small ontologies or design patterns that are generic geographic features that can be reused. Examples include path, point-of-interest, and others. The focus of many of these efforts is to build a set of spatial feature classes with semantics that account for attributes, relations, and interactions of one spatial feature with another. It should be noted that these ontologies are domain ontologies of classes and rarely include instance data of actual geographic features in real-world locations. There also
are efforts to provide instance-level data and develop procedures for use and interaction of those instance data (for example, see Usery and Varanka 2012).

Conclusions

The concept of spatial feature classes is now an integral part of an ontology of geographic information. The design and development of lists of spatial feature classes is an ongoing effort and will continue as long as there is a need to define, identify, and interact with geographical entities in the real world. Defining classes allows assignment of members to those classes and provides a mechanism to use instance-level data with coordinate representation and the many levels of attributes and relations built into the spatial feature class ontology. A major effort today in defining spatial feature classes is occurring through crowdsourcing the effort through GeoVoCamps. With new sensors, including humans with geolocation devices, providing geographic data of many different kinds through crowdsourcing, the definition and assignment of data to spatial feature classes are being crowdsourced as well.

SEE ALSO: Cognition and spatial behavior; Data model, object-oriented; Fiat and de facto objects; Fuzzy classification and reasoning; Geospatial Semantic Web; Ontology: domain applications; Open Geospatial Consortium standards; Representation; Spatial concepts

References


SPATIAL FEATURE CLASSES


Spatial filtering
Victor Mesev
Florida State University, USA

Spatial filtering is a data frequency manipulation technique designed to improve geographical feature interpretation. An independent $n \times n$ moving window (filter, mask, or kernel) is used to suppress or enhance spatial frequencies between local neighboring data values of a separate target dataset. For example, high frequencies indicate rapid changes of values between neighboring cells or pixels (values that represent the edges of features, such as buildings, coastlines, or roads), while low spatial frequencies indicate more gradual changes between neighboring values (values that represent more homogeneous features, such as the areas within deciduous forests, wheat fields, or lakes). Spatial filters are also used to improve data quality by removing, or at least minimizing, striping and noise, which are unexplained variations in data frequency and are, on the whole, unrepresentative (or unexpected) of spatial patterns and trends. Spatial filtering is part of statistical or quantitative analysis in geography. Many types of spatial filters are available in current geographic information systems and image processing software, but two main groups of spatial filters are those used for (i) data manipulation of remotely sensed imagery, including digital photogrammetry, and (ii) data manipulation of spatial data in response to the effects of spatial autocorrelation.

Spatial filtering of remotely sensed image data is known as spatial convolution, and the objective is to improve feature interpretation by visually smoothing or sharpening the image. Convolution uses windows of typically $3 \times 3$, $5 \times 5$, $7 \times 7$, or $9 \times 9$ fixed arrays that multiply pixel brightness (or digital numbers) by moving systematically across the entire raster data image and altering the pixel corresponding to the central cell of the window (hence the use of odd numbers only). For example, when using a $3 \times 3$ window, each pixel in the target image is assigned a new number based on the coefficient of the nine cells in the window (Gao 2009):

$$DN_{OUT} = \frac{1}{W} \sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij} DN(i,j)_{IN}$$

where, $DN_{OUT}$ = convoluted output pixel value; $DN(i,j)_{IN}$ = pixel value in the input image at location $(i,j)$; $i$ ($i = 1, 2, 3$) = row index; $j$ ($j = 1, 2, 3$) = column index; $w_{ij}$ = value of the cell at location $(i,j)$ in the window; $W$ = sum of all window elements; and $d$ = window size (in this case 3).

There are essentially three types of spatial filters used in photogrammetry and remote sensing: low pass filters that de-emphasize the spatial frequencies of neighboring raster pixel values, high pass filters that enhance spatial frequencies of neighboring values, and edge enhancement filters that also sharpen the contrast between the spatial frequencies of neighboring pixel values, but in a vertical, horizontal, or diagonal direction (Jensen 2004; Gonzalez and Woods 2008). The images in Figure 1 illustrate the visual effects on a remotely sensed image by applying the mean low pass spatial filter, the high pass smoothing or sharpening the image. Convolution uses windows of typically $3 \times 3$, $5 \times 5$, $7 \times 7$, or $9 \times 9$ fixed arrays that multiply pixel brightness (or digital numbers) by moving systematically across the entire raster data image and altering the pixel corresponding to the central cell of the window (hence the use of odd numbers only). For example, when using a $3 \times 3$ window, each pixel in the target image is assigned a new number based on the coefficient of the nine cells in the window (Gao 2009):

$$DN_{OUT} = \frac{1}{W} \sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij} DN(i,j)_{IN}$$

where, $DN_{OUT}$ = convoluted output pixel value; $DN(i,j)_{IN}$ = pixel value in the input image at location $(i,j)$; $i$ ($i = 1, 2, 3$) = row index; $j$ ($j = 1, 2, 3$) = column index; $w_{ij}$ = value of the cell at location $(i,j)$ in the window; $W$ = sum of all window elements; and $d$ = window size (in this case 3).

There are essentially three types of spatial filters used in photogrammetry and remote sensing: low pass filters that de-emphasize the spatial frequencies of neighboring raster pixel values, high pass filters that enhance spatial frequencies of neighboring values, and edge enhancement filters that also sharpen the contrast between the spatial frequencies of neighboring pixel values, but in a vertical, horizontal, or diagonal direction (Jensen 2004; Gonzalez and Woods 2008). The images in Figure 1 illustrate the visual effects on a remotely sensed image by applying the mean low pass spatial filter, the high pass
Figure 1  Visual effects on a remotely sensed image by applying filters: (a) original image; (b) mean low pass spatial filter applied; (c) high pass Laplace edge enhancement filter applied; (d) Sobel linear edge enhancement filter (in the vertical direction) applied.  

Laplace edge enhancement filter, and the Sobel linear edge enhancement filter (in the vertical direction). The mean low pass filter has a window with an array of values all set to 1, or in this case set to $1/9$, which when multiplied with the central pixel brightness value results in a reduction in spatial frequencies of values between neighboring pixels. In contrast, the window for the high pass Laplace filter has 0 and $-1$ values surrounding a central value of 4. The effect is to sharpen spatial frequencies in the image data by computing the derivative from the gradient of the spatial frequencies. Occasionally, the directional alignment of geographic features is of interest, and using linear edge enhancements can accentuate the visual contrast of buildings and roads aligned in, say, a vertical direction. For that to happen, the Sobel filter may have values set to 1 in the left column of the window and values of $-1$ in the right column.

Spatial filtering is termed linear when the output image is a function of some weighted average (a linear combination) of pixel brightness values. This process of evaluating the weighted neighboring pixel values is called 2-D convolution filtering. Another linear spatial filter technique designed for remotely sensed image data is Fourier analysis, which statistically separates the image into its spatial frequency components (known as Fourier transforms) by a summation of a series of sinusoidal terms of varying spatial frequencies (for a statistical formulation of the Fourier transform see Jensen 2004). The technique is able to maintain fidelity when combining (or fusing) multiple images at a lower spatial resolution with the high spatial resolution properties of panchromatic-sharpened images, which has the effect of reducing the storage space necessary for the images (Civco and Witharana 2012). In contrast to convolution techniques, another means of applying linear
Spatial filtering to remotely sensed images is by correlation; this is used to measure the similarity between images (a technique known as pattern matching).

Linear spatial filters are far more amenable to numerical analysis, but nonlinear filters are able to restore image data values, reduce noise levels, and, unlike linear filters, do not simultaneously blur edges. The simplest nonlinear spatial filter is the moving median, which is used for removing striping noise in an image (including raster digital elevation models) by ranking the pixels in the neighborhood from lowest to highest, and then selecting the median value. Nonlinear filters are not additive; repeated application of a median filter is not equivalent to a single application of a median filter using a different size of window. By making repeated use of a nonlinear smoothing filter with a small window, it is sometimes possible to improve on noise reduction while at the same time retaining fine details that would be lost if a larger window was used. In terms of accentuating the contrast of edges using a nonlinear process, the minimum and maximum values in a window are typically used in a range filter, or as in a Roberts’s filter, the sum of the absolute differences of diagonally opposite pixel values (Gonzalez and Woods 2008).

In general statistical analysis, spatial filters are used to isolate the underlying spatial effects of geographical variables from the effects of temporal and spatial autocorrelation.

This second main group of spatial filtering techniques measures the lack of independence between adjacent observations across geographical space. Essentially, this is the measurement of the effects of spatial autocorrelation, either using Moran’s \( I \) correlograms to explore spatial patterns in data, or to verify the presence of latent predictors in a multiple regression at different scales. In addition, the Getis and Griffith approach uses an eigenfunction decomposition based on the geographic connectivity matrix used to compute a Moran’s \( I \) statistic (Getis and Griffith 2002). Eigenvector-based spatial filtering can be used to analyze geographical patterns in many different applications (Diniz-Filho and Bini 2005). In addition, numerous other techniques have attempted to incorporate the temporal component of data, as well as the spatial aspects, when studying stochastic error characteristics in a time-series process (Getis and Griffith 2002).

**SEE ALSO:** Data structure, raster; Discrete global grid systems; Geocomputation; Photogrammetry: 3-D from imagery; Quantitative methodologies; Resampling, raster; Spatial analysis; Spatial weights; Spatiotemporal analysis

**References**


Spatial interaction is a generic term for any form of movement or transmission over space that results from a decision-making process. Movement implies a physical relocation of an object or a person from one location to another, whereas transmission implies a nonphysical interaction such as communication. In both cases, a key concept of spatial interaction is that there is a transfer between an origin and a destination – in the case of movement, this transfer is a physical entity such as a good or a person; in the case of communication, the transfer is of information. To make this clearer, examples of spatial interaction include commuting flows; shopping trips; recreational visits; airline passenger movements; visiting relatives; attending sporting events, concerts, or religious services; migration; freight movements; going to school; telephone calls; and various forms of social networking involving location. In short, spatial interaction involves many of the important things we do as human beings. The only types of movement or communication that are not included within the definition of spatial interaction are those types of movement or communication that do not result from a choice, such as storm tracks, water-flows, the diffusion of pollution, forced migration, and the spread of forest fires. All of these involve movement of one type or another but the movement does not result from a spatial choice process.

Consequently, spatial interaction is arguably one of the most important topics within geography because it encompasses so many different and important types of human behavior. For instance, the types of questions that can be answered through an understanding of spatial interaction include the following.

1. Is there sufficient demand to establish a profitable air service between two cities? What frequency of air service will meet demand?
2. Is the investment in a new rail service, road system, or bridge cost effective given the expected demand?
3. What will be the effect on traffic patterns of building a new shopping center?
4. What will the effect on housing demand be throughout the city if a new employment center locates in place X?
5. From which schools does a university draw more (or fewer) students than would be expected given the locations and sizes of the schools?
6. Where is the optimal location for a new supermarket?
7. How sensitive are various types of movements to changes in travel costs?
8. What differences in travel behavior are there between different types of individuals? Does where you come from affect how far you are likely to travel?
9. Are we becoming less constrained by space over time?
SPATIAL INTERACTION

11 What makes some areas attractive (or unattractive) to migrants?
12 Would it be better to offer a specialist type of medical service in one hospital rather than another based on demand?
13 Where does the catchment area of one retailer end and that of a competitor begin?
14 How will the location of a new store affect demand at existing stores?
15 How many people can be expected to attend a mobile breast cancer screening service in a given location?

As Olsson states (1970, 233): “Under the umbrella of spatial interaction and distance decay, it has been possible to accommodate most model work in transportation, migration, commuting and diffusion, as well as significant aspects of location theory.”

Despite the diversity of types of spatial interaction, there are several tenets in common across all these areas.

1 A choice has to be made from a set of alternatives at different locations.
2 The choice is discrete, that is, only one alternative is selected.
3 The alternatives have fixed spatial locations.
4 There are many factors that affect this choice and these factors are context-dependent. That is, what determines a consumer’s choice of store will not be the same as what affects a migrant’s choice of destination.
5 However, regardless of what are the exact determinants of choice, no matter what the context, an individual somehow evaluates alternatives and trades off poor scores on one attribute with good scores on other attributes in order to select one particular alternative.

6 An important attribute of the choice process, common to all types of spatial interaction, is the spatial separation between an origin and a destination (often measured by distance but can also be thought of in terms of cost or time). Generally, we are more reluctant to interact with places farther away than we are with nearer places, all things being equal.

What do spatial interaction data look like?

Spatial interaction data denote the propensity of movement or communication between a set of origins (labeled $i$ here) and a set of destinations (labeled $j$ here), as depicted in Figure 1, where $M_{ij}$ represents the number of interactions or flows between $i$ and $j$. The propensity of flow could equally be denoted by a probability of movement between $i$ and $j$ representing the likely actions (decisions) of an individual, but for the sake of convenience here $M_{ij}$ is used as an aggregate number of interactions between an origin $i$ and a destination $j$.

In this situation there are $m$ origins and $n$ destinations so that we have an $m \times n$ interaction matrix. The sum of the interactions across a row, $O_i$, represents the total flow out of an origin and the sum of the interactions down a column, $D_j$, represents the total flow into a destination. The overall total interaction in the system is denoted by $M$ where:

$$M = \Sigma_i \Sigma_j M_{ij} = \Sigma_i O_i = \Sigma_j D_j$$

A spatial interaction matrix such as that in Figure 1 can be disaggregated in many ways, such as over time, over space, and by person type. For example, it is possible to have a separate matrix of shopping trips for car owners versus
Figure 1 A typical spatial interaction matrix.

non-car owners; for weekdays versus weekends; and for one city versus another.

Spatial interaction models

Geography can be defined as the study of spatial processes. In human and environmental geography, spatial processes are the actions and decisions that lead, at least in part, to the data that are measured to describe the social and physical environments. In order to better understand these processes, it is useful to formulate and calibrate models of the real world. In terms of spatial interaction, this is done through the formulation of spatial interaction models – mathematical models that relate the propensity of interaction between an origin \( i \) and a destination \( j \) to attributes of origin \( i \), attributes of destination \( j \), and some measure or measures of the separation between \( i \) and \( j \). The particular attributes of origin \( i \) and destination \( j \) deemed relevant to describing variations in interaction propensity depend on the type of interaction being examined, but, in general, people speak of the “push” and “pull” of places, or how the attributes of an origin would be those describing the “propulsiveness” of the place and the attributes of a destination would be those describing its “attractiveness.”

In terms of shopping behavior, for example, origin propulsiveness might be measured in part by population, income levels, and car ownership rates and destination attractiveness might be measured by store size, average price levels at a store, parking facilities, or other relevant factors. What constitutes relevant propulsiveness and attractiveness measures is something of an art form based on a mix of experience, guidance by the existing literature, common sense, and statistical testing.

A very simple spatial interaction model with only one origin propulsiveness attribute, \( v_i \), and one destination attractiveness attribute, \( w_j \), and one measure of spatial separation taken to be the distance between \( i \) and \( j \), \( d_{ij} \), is:

\[
M_{ij} = k v_i^\alpha w_j^\theta d_{ij}^\beta
\]

where \( k \) is often referred to as the intercept (a scaling parameter with no behavioral value and its value is not generally of much interest); \( \alpha \) indicates the sensitivity of the flow from \( i \) to \( j \) to the propulsiveness of origin \( i \) as measured by \( v_i \); \( \theta \) indicates the sensitivity of the flow from \( i \) to \( j \) to the attractiveness of destination \( j \) as measured by \( w_j \); and \( \beta \) indicates the sensitivity of the flow between \( i \) and \( j \) to the spatial separation between \( i \) and \( j \). This last parameter is sometimes referred to as “the disutility of distance” or “the friction of distance” or “distance decay.” It will almost always be negative, indicating that, other things being equal, as the distance between an origin and destination increases, interaction will decrease. That is, we are more likely to patronize a store which is 100 m away from our home than one which is 10 km away, other things being equal.

One very important and useful reason for calibrating a spatial interaction model is to obtain estimates of the parameters in the model to obtain information on the determinants of
interaction. For example, it might be interesting to examine how the parameter estimates vary by type of interaction, by the type of people interacting (i.e., calibrating separate models for different groups of people perhaps based on characteristics such as age, income, and gender), over space, and over time (either as transportation access and incomes vary over time or diurnal variations, such as during rush hour versus during regular traffic conditions). For example, Figure 2 depicts a situation in which $\beta$, the distance-decay parameter, is becoming less negative over time, indicating that distance is becoming less of a deterrent to interaction over time, presumably as transportation systems evolve and/or people’s ability to pay for transportation increases.

Arguably the distance-decay parameter, $\beta$, is the most interesting element of a spatial interaction model, as it measures the inherent ability of individuals to overcome the friction or distance (alternatively, the disutility of traveling). This parameter provides a means of quantifying variations in the ability to select from a wider set of destinations by individuals; for example, between those who own cars and those who do not own cars, or between younger migrants and elderly migrants, or between people with access to rapid transit and those without. Similarly, estimates of this parameter over time allow the degree to which traveling longer distances has become easier to be quantified, allowing the growth of cities, the decentralization of retailing away from city centers, and the growth of mega-malls. Indeed, the concept of distance decay is fundamental to understanding the evolution of urban areas, retail systems, patterns of migration, and most facets of regional and urban planning.

There are numerous examples of empirical applications in which variations in the parameter estimates obtained in the calibration of spatial interaction models have been reported; a useful overview of these can be found in Fotheringham and O’Kelly (1989). It is important to note that such comparisons are possible because the parameters of the model described above are power functions, and hence are scale invariant. For instance, in some studies an exponential function of distance has been employed so the model calibrated is then:

$$ M_{ij} = k v_i^\alpha w_j^\theta \exp(\beta d_{ij}) $$

In this case, for example, the estimate of $\beta$ depends on the units in which distance is measured.

The spatial interaction model given above is perhaps the simplest possible model that yields useful information on the processes generating the observed flows between origins and destinations. More complex forms are generated by (i) adding additional origin and destination attributes and (ii) by using more complex mathematical model forms. For instance, consider the investigation of shopping trips where the focus is on uncovering attributes of stores that make them attractive or unattractive to consumers. Here, it is possible to use the known volumes of outflows from each origin and concentrate solely on assigning these flows to the various destinations based on the characteristics of these destinations (including, importantly, how far each potential destination is from an origin). In
which case, the appropriate spatial interaction model has the form:

\[ M_{ij} = O_i w_j d_{ij}^{\beta} / \Sigma_j w_j d_{ij}^{\beta} \]

where \( O_i \) is the observed (known) outflow from origin \( i \) and the situation is restricted to where there is only one destination attractiveness attribute in the model, \( w_j \), for convenience, but it is recognized that further attributes with associated parameters can be easily added. This model is often referred to as a production-constrained spatial interaction model because the sum of the predicted outflows for each origin, \( \Sigma_j M_{ij} \), must equal the observed volume, \( O_i \). In essence, this model allocates a share of the total flows from an origin to each destination based on the relative attractiveness of each destination where relative attractiveness is defined by the term \( w_j d_{ij}^{\beta} / \Sigma_j w_j d_{ij}^{\beta} \). This model is hence particularly useful in retailing studies where the aim is to identify those store attributes that have a significant impact (both positive and negative) on consumer store choice. It can also be used to determine the optimal location of stores by finding the location which will maximize \( \Sigma_j M_{ij} \), the predicted total inflow into a destination. (Further retailing applications of this model can be found in Fotheringham 1988 and Fotheringham and Trew 1993.)

Alternatively, if the focus of the analysis is on analyzing residential choice, given that it is known where individuals work for the purposes of understanding housing demand, then the following spatial interaction is appropriate:

\[ M_{ij} = D_j v_i d_{ij}^{\beta} / \Sigma_i v_i d_{ij}^{\beta} \]

where \( D_j \) is the observed total flow into destination \( j \) and a share of this total is allocated to each origin based on the relative propulsiveness of each origin as measured by the term \( v_i d_{ij}^{\beta} / \Sigma_i v_i d_{ij}^{\beta} \). This model is often referred to as an attraction-constrained spatial interaction model because the sum of the predicted flows into each destination, \( \Sigma_j M_{ij} \), is constrained to equal the known total inflow into each destination, \( D_j \). Apart from identifying those features of an origin that lead to large outflows, a useful application of this model is in predicting housing demand given a change in the employment structure of an area such as the opening or closure of a large factory. By varying one or more values of \( D_j \), the effect of these changes on housing demand in each origin can be assessed by computing \( \Sigma_j M_{ij} \), the predicted total outflow from each origin.

A fourth type of spatial interaction is one in which the focus is to simply predict flows between origins and destinations given that the total outflows from each origin and the total inflows into each destination are known. Such a situation arises, for example, in urban transportation planning, where the total number of employees living in each neighborhood and the total number of jobs in each destination zone is known and the aim is to allocate flows from each origin to each destination under different scenarios. Essentially, this is a situation where the row and column totals of the matrix in Figure 1 are known but the cell counts are unknown. In this case, the appropriate model form is referred to as a doubly constrained spatial interaction model where

\[ M_{ij} = A_i O_i B_j D_j d_{ij}^{\beta} \]

where

\[ A_i = [\Sigma_j B_j D_j d_{ij}^{\beta}]^{-1} \]

and

\[ B_j = [\Sigma_i A_i O_i d_{ij}^{\beta}]^{-1} \]

The terms \( A_i \) and \( B_j \) are known as balancing factors because they ensure that the following two conditions are met:

\[ \Sigma_j M_{ij} = O_i \text{ for all } i \]
The former states that the sum of the predicted interactions from each origin will equal the known total outflow from each origin and the latter states that the sum of the predicted inflows into each destination will equal the known total inflow into each destination. Hence, the model is useful for assessing the impact of changing road conditions within a city. A change can be represented by an increase or decrease in the distance between $i$ and $j$, or between a set of origins and destinations, and the model can be used to predict a new flow matrix under this change. It is thus useful in transportation planning to assess the impacts of new transportation infrastructure such as building a rapid transit line or a new crossing over a river.

The four models just described can be derived from a common general form (Alonso 1978) and can be seen as four particular outcomes from an infinite number of possible models, as shown in Figure 3, which is derived from the formula:

$$M_{ij} = k v_i A_i^{\alpha-1} w_j B_j^{\gamma-1} d_{ij}^{\beta}$$

where

$$A_i = \sum_j w_j B_j^{\gamma-1} d_{ij}^{\beta}$$

and the destination-specific, attraction-constrained model is:

$$B_j = \sum_i v_i A_i^{\alpha-1} d_{ij}^{\beta}$$

So that:

- if $\alpha = 1$ and $\gamma = 1$, the model is unconstrained;
- if $\alpha = 0$ and $\gamma = 1$, the model is production-constrained;
- if $\alpha = 1$ and $\gamma = 0$, the model is attraction-constrained;
- if $\alpha = 0$ and $\gamma = 0$, the model is doubly constrained.

Consequently, it is possible to think about spatial interaction models derived from this general framework where the values of $\alpha$ and $\gamma$ are between zero and one. Such models can be considered as “quasi-constrained” or “relaxed” spatial interaction models, which may be useful in situations where the constraints on either or both the outflows from each origin or the inflows into each destination are not known with certainty (Fotheringham and Dignan 1984).

Finally, the form of spatial interaction models described thus far assumes that each model is calibrated using all the data from the full interaction matrix depicted in Figure 1. In such applications, one estimate of each parameter in the model is obtained reflecting general relationships across all origins and destinations. However, to obtain much more detailed information on the determinants of interaction, the models can be calibrated separately for each origin and each destination in the system so that the parameters are specific to each origin or to each destination. For example, an origin-specific, destination-attractiveness model has the form:

$$M_{ij} = O_i w_j^{\theta(i)} d_{ij}^{\beta(i)} / \sum_j w_j^{\theta(i)} d_{ij}^{\beta(i)}$$

and the destination-specific, attraction-constrained model is:
Theoretical foundations

There have been at least four attempts, of increasingly behavioral validity, to provide a theoretical basis for the form of spatial interaction models described above. These are described with an approximate evolutionary timescale in Figure 4.

The earliest form of spatial interaction was termed a “gravity” model because it was developed as an analogy to Newtonian laws of gravity that the force of gravity between two objects increased as the masses of the objects increased, and decreased as the distance between them increased. This simple analogy, part of a broader trend of importing concepts from physics into social science, termed social physics, served well to predict population movements between cities but was clearly devoid of any behavioral context save from a vague notion of distance having a frictional effect on movement.

In the 1970s, an attempt was made by Wilson (1975) to put spatial interaction on a firmer theoretical footing by demonstrating how the models could be developed by reference to the principles of entropy maximization – essentially finding the most likely set of origin–destination flow counts in Figure 1 given constraints on the row and column totals. However, although it was possible to derive spatial interaction models via this method from first principles, the behavioral foundations were really no stronger than in the previous social physics era. Indeed, the development of spatial interaction models through an entropy maximization framework could be viewed as an alternative type of social physics.

The behavioral foundation of spatial interaction models was given a boost by McFadden’s (1974) derivation of choice models via a utility maximizing framework. For the first time, models such as the production- and attraction-constrained models could be seen as the outcome of choice processes in which individuals selected one alternative from the set of all alternatives that maximized their utility. However, this development, although extremely insightful, was essentially aspatial in that it ignored the spatial arrangement of potential destinations and treated the choice of a destination in exactly the same manner as individuals chose a brand of coffee. It remained for Fotheringham (1983, 1986) to add a spatial component to McFadden’s choice axioms by recognizing that individuals faced with large numbers of alternative destinations arranged in space did not always, and indeed could not always, evaluate all the alternatives available to them to make a choice. An example would be
a person selecting a residence in a large city and being faced with thousands of possible houses or apartments. Consequently, Fotheringham recognized that people were likely to make spatial choices hierarchically, first selecting an area or spatial cluster of alternatives and then selecting one alternative from this subset having evaluated only the alternatives in this subset. This led to the development of the competing destinations model which has the form:

$$M_{ij} = O_{ij} w_j^\theta d_{ij}^\beta c_j^\varphi / \sum_j w_j^\theta d_{ij}^\beta c_j^\varphi$$

where $c_j$ is the degree of competition faced by destination $j$ from all other destinations and is generally defined in terms of some measure of the proximity of $j$ to all other destinations, preferably weighted by the attractiveness of each destination such that high values of $c_j$ reflect a destination in close proximity to other destinations and low values a destination that is relatively isolated. The parameter $\varphi$ indicates the effect of this competition on destination selection and is generally negative, indicating that as a destination has greater competition from nearby destinations, it attracts fewer flows, all things being equal.

**Summary**

The term spatial interaction encompasses many important components of social systems, including, but not limited to, commuting patterns and residential choice; shopping trips and the location of stores; the efficient and equitable location of facilities such as medical services; transportation demand; the impact of new services and facilities on existing demand; the analysis of social networking data such as twitter contacts; the analysis of communications; and the growth of cities. By reversing the modeling component, spatial interaction models have also been used to predict the locations of ancient settlements and to derive measures of the relative attractiveness of recreational areas and other destinations (Fotheringham et al. 2000). Indeed, it is difficult to comprehend properly the growth of cities and the distribution of population without an understanding of spatial interaction and its determinants.

Consequently, it can be argued that spatial interaction is one of, if not the most important subject areas in geography. Indeed, it can easily be argued that human geography would not exist without spatial interaction and distance decay. Hence, it is essential that geographers continue to conduct research that improves both the understanding of the processes leading to spatial interaction and the ability to model and predict patterns of spatial flows.

**SEE ALSO:** Accessibility, in transportation planning; Graph theory; Location-allocation analysis; Regional science; Spatial analysis; Spatial optimization; Topological relations; Transport geography; Transportation planning

**References**


Spatial modeling: Voronoi diagrams

Christopher Gold
*University of South W ales, UK*
*Jiaotong University, China*

In the context of this entry, a spatial model is a conceptual model of the space under consideration, and a data structure is the way it is stored in the computer. The Voronoi diagram is a particular, and very useful, spatial model.

The Voronoi diagram (also called a Dirichlet tessellation or a set of Thiessen polygons) is a partition of a metric space such that we associate all locations in that space with the closest object in that space, based on the specified metric. Thus each object is the generator of a cell or tile and the set of tiles covers the space or map. In the simplest case we are given a set of points \( S \) in the Euclidean plane, which are the Voronoi generators. Each generator \( s \) has a Voronoi cell \( V(s) \) consisting of all points closer to \( s \) than to any other generator. Voronoi cells can be defined for metrics other than Euclidean, such as the Manhattan distance, and for other surfaces than the plane, such as the sphere. Voronoi cells can also be defined by measuring distances to objects that are not points.

We will first discuss spatial models, then some appropriate data structures, and conclude with some applications of the Voronoi/Delaunay approach, to illustrate its utility and flexibility.

Objects and fields

The way we think of space varies considerably depending on the application, the type of data, and our own presuppositions. One well-accepted distinction is between a set of spatially located “objects” such as buildings and a continuously varying value – a “field” – such as temperature. Another useful distinction concerns our navigation within this world: are we a “plane” flying above, or a “boat” acting within this space and interacting with it – this would mostly, but not necessarily, apply to the “object” model.

A “field” is continuous, but not necessarily smooth: it could consist of a tessellation with constant values in each cell – “blocks.” An “object” model would have empty space between at least some of the tiles, and does not necessarily have any obvious spatial relationships between them.

A famous debate between Newton and Leibnitz concerned whether space was “real” and could be measured (Newton) or only existed as the spatial relationships between objects (Leibnitz). The first led to Cartesian coordinate systems and raster spatial models, and the second led to graph theory and networks. (We can of course have two-dimensional (2D) and three-dimensional (3D) Cartesian space – and other dimensions – but other metric spaces are possible, including “city block metric” where distance must be measured along the axes only.) We may also have spatial embedding, when a lower dimension structure may be embedded in a...
higher dimension – for example a terrain surface embedded in 3D space, a road embedded in 2D space, or a node embedded in a one-dimensional (1D) road. (All of these examples imply that the higher-dimensional space is not necessarily fully occupied: that is, an arbitrary query at any location in the bounding space will not always return a value.)

The simplest form of spatial relationship between a pair of objects is adjacency: if their boundaries touch (presupposing that only one object may occupy any particular location) then a relationship – an edge of a graph – may be recorded. In a field model the whole space – within some bounded region – is filled and all objects (tiles) have recorded neighbors. In an object model this will, at most, be only partially true, meaning that navigation within that model cannot be guaranteed.

The Voronoi spatial model attempts to remedy that deficiency: objects are expanded until all boundaries meet, “holes” in the space are filled in, and a complete set of spatial adjacency relationships are obtained: the object model becomes a field. As with fields, a response is available at any spatial location, and as with the object model, the “name” of the object and its spatial properties may be obtained directly (unlike traditional raster fields, where the object must be reconstructed from grid cell attributes). Only one change needs to be considered: instead of obtaining “the object at this location is …,” we get “the object closest to this location is ….” As we will see, this is often what we wanted to ask in the first place.

**Dominance**

In its simplest form, an object (often a point) dominates all spatial locations closer to it than to any other point: this is its dominance region – think perhaps of predators who protect their territory with greater ferocity closer to their lair; there will be a boundary between two territories at an equal distance from the lairs. This boundary is the perpendicular bisector of the line between these generators. For three generators – and three boundaries – there is a single location equidistant between the three, at the center of a circle touching the three generators, where the three boundaries meet. (For collinear generators and boundary cases it is often easier to place them on a sphere rather than a plane. In 3D the separating boundaries are planes, with four meeting at the circumcenter.) Adding additional generators produces a complete tessellation of the space, termed a Voronoi diagram (VD). The collection of the edges connecting generators sharing boundaries is termed the Delaunay triangulation (DT). (In the plane the DT forms a set of triangular regions except where four or more generators are cocircular, in which case that region may be arbitrarily subdivided into triangles without losing any of the useful properties.)

**Planar graphs, duality**

A graph consists of a set of nodes and connecting edges. The most common type is a “planar graph”: here the graph is embedded in the plane and no edges cross. (To be precise, no edges have to cross – taking a planar graph and pushing the interiors of some edges across other edges does not make it nonplanar – this is only a geometric change, and pure graphs have no geometric properties.) Graphs may be undirected or directed: undirected edges may be traveled in both directions, and directed edges are like “one-way streets.” Edges may also have weights: these describe the “cost” of traveling them, for example travel time on a road segment.
Planar graphs may also have faces (regions) defined by closed loops of edges. The Euler-Poincaré formula states that for a planar graph $V + F - E = 2 - 2G$, where $V =$ # vertices (nodes), $F =$ # faces (regions), $E =$ # edges (arcs) and $G =$ the genus (# holes in the manifold). The rules for “planar” graphs hold for closed surfaces (2-manifolds): a plane or sphere has $G = 0$, a doughnut has $G = 1$. The exterior of a graph is also a region: think about drawing the graph on a balloon, and looking at it from behind.

In order to walk around the edges of a face we need to know the “next” edge around each vertex, which adds “topology” to the original graph. If we know this we can navigate in the graph, for example around polygon faces (Figure 1). This is normally difficult when only given digitized edges, but easy when using the Voronoi diagram.

If we start with a set of regions (adjacent countries perhaps), put a “capital city” within each, and connect the capitals whenever two countries have a common boundary, we create a “dual” graph. This dual graph is also composed of vertices and edges: the vertices represent the original regions and the edges were defined to cross the original region boundaries. Itselves a planar graph, the dual graph has regions, which are the duals of the junctions of the boundaries of the original countries. Both the primal and the dual graphs follow Euler’s formula. The important point to note is that, in any number of dimensions, every entity has a matching (paired) entity in the dual – and vice versa, as the dual of the dual is the original graph. As both the primal and the dual structures are graphs, standard graph theory algorithms may be applied to traverse the graph efficiently, detect closed regions (polygons) if given the edge order, as in Figure 1, and estimate flow in networks. This is also true for Voronoi/Delaunay graphs, as Figure 2 shows.

Imagining that Figure 2 represents an elevation map, with values at the data points, it can be visualized in two ways. In the first way it is now common to represent terrain as a TIN – a triangulated irregular network. In this case the surface would be composed of a set of flat, sloping, triangular plates: along the boundary between a pair of triangles they share common elevations. Alternatively, each Voronoi cell could be considered as a block having the data point elevation.

**Figure 1** Ordering of edges around polygons, and around vertices.
There are many ways that triangles may be constructed from a set of data points. The method now almost universally accepted is to construct the DT, as it has some very valuable properties. First, the definition maximizes the minimum angles of the triangles, reducing the chances of having long skinny ones. Second, the solution is unique: no matter the order in which the points are inserted the resulting triangles are the same. (The exception to this is when there are four or more cocircular points: any set of interior triangle edges may then be chosen.)

The incremental algorithm is the most common approach, as it is simple and robust, although not the fastest. It can be very simply described: (i) create a boundary triangle large enough to contain the whole map; (ii) perform a “local search” to find the triangle enclosing each new data point (initially the boundary triangle); (iii) split this triangle into three, with edges connecting each of its vertices with the new point; (iv) test each pair of triangles having a common edge, and decide whether to “switch” this diagonal for the other alternative; (v) test all edges whose adjacent triangles may have been changed by all this, and switch if necessary; (vi) go back to step ii and repeat for each new data point (Figure 3) (Lawson 1977). A “local search” may be performed by testing each edge of some initial triangle against the desired location, moving to an adjacent triangle if the location is outside, and repeating the process until the location lies on the interior of each of the three edges (Gold, Charters, and Ramsden 1977).

Many tests could have been done at step iv but the one that guarantees that the minimum angle in the triangle pair is progressively maximized, and that a finite number of switches (averaging six for each point insertion) is required, is based on the Voronoi/Delaunay duality. Voronoi “bubbles” created by the dominance method meet at a triple junction, which is the center of a circle (the circumcircle) touching the associated triangle’s three vertices. If a new point is placed inside this circle then this center, and its associated triangle, will be destroyed. Thus for each triangle to be tested, the three exterior vertices (associated with its adjacent triangles) must be checked to see if they fall inside its circle (Shewchuk 1997). If they don’t the triangle is good, otherwise the common edge between the two triangles must be switched and the tests repeated. It can be shown that this converges rapidly. The Delaunay triangulation test is based on the underlying Voronoi model. This is a basic example of the importance of the interlinking of the primal and dual graphs; we will see many more.

Parenthetically, the circumcenter is the interior point furthest away from the three triangle vertices – so an examination of the circumradii of all the Voronoi vertices on the map will locate the largest available empty circle (Figure 4).
null
Spatial models

We previously imagined that the terrain model of the given elevation data points of Figure 2 could be expressed as a set of flat triangular plates. However, we could equally have imagined that it was represented by the set of Voronoi cells, each with the height of the generating data point: this gives us an alternative spatial model, which is equally useful – we have a “block” world, to go along with our previous “slope” world (Figure 5).

This is what, in this entry, we mean by a “spatial model.” The most common is the raster, a field model where the partition is based on the coordinate system, not the data distribution. There are others; for example, a typical choropleth map or a road network, where the relationships between the entity types may be different: in a road network the only connectivity is between individual segments, and the only navigation queries are from segment to segment, whereas the Voronoi approach is a “space-filling” model, where adjacency relations between all 2D cells are available.

The hierarchical Voronoi diagram

It is conventional knowledge that one of the disadvantages of the VD is that there is no hierarchical structure available, so that search of very large datasets is slow. Various kinds of searching may be performed (Figure 6). One solution can be found based on Christaller’s work (see the discussion in Gold and Angel 2006) where he examined town and village distributions and their respective hinterlands. He did not attempt to insert lower-level village zones of influence within the higher-level zones for towns. He inserted the center of the zone – the village location – instead. Clearly points can
easily be inserted within a town’s higher-level index zone and, similarly, lower-level points can be located within a village zone.

The hierarchical Voronoi problem can thus be solved by using high-level Voronoi cells as index cells for the next level of points, letting their cells be secondary indexes for the third level points, and so on. In searching for a particular location, a local search (as described above) is performed at the top level until the enclosing cell is found, along with a pointer to some point in the next level down (if any exist). This is repeated on this second level, and on the third level, and so on, until a closest point is found at the base of this search tree. This speeds up the theoretical time for the walk (and hence the whole incremental algorithm), and any desired index level may form the basis for separating the map “sheets” and storing them in separate portions of the hard drive.

2D Spatial data structures and the quad-edge

A “spatial data structure,” on the other hand, as opposed to a “spatial model,” is a technique to allow the representation of a particular spatial model within the computer. Since the Delaunay TIN terrain model is well known, a triangle-based data structure is often used: here a topological element consisting of a triangle, with pointers to the three vertices and to the three adjacent triangles, is often used (Gold, Charters, and Ramsden 1977). This, however, has several disadvantages: it cannot be used to navigate from one triangle to the next without checking which edge it has in common with the starting triangle; and it cannot directly represent the dual (Voronoi) structure. The first issue can best be resolved by using an edge-based structure such as the winged-edge or half-edge (Baumgart 1972), which permits direct pointer-based navigation from one element to the next throughout the whole map. The second issue may easily be achieved by using the quad-edge data structure (Guibas and Stolfi 1985), which combines two half-edges in the primal structure (e.g., the Delaunay) with two in the dual structure (e.g., the Voronoi). This is particularly valuable when both the primal and the dual spatial models are required for an application.

An early data structure for modeling 2-manifolds (e.g., closed surfaces in Computer Aided Design (CAD) systems) was the winged edge. Each data element is an edge, with four pointers to the adjacent edges in the structure, to permit navigation from one edge to the next. It has a particular problem, in that when moving to a new edge – perhaps round a face – it is not clear which way it is oriented. A vertex check is required in order to find which end of the new edge connects to the old. An improvement is the half-edge, where each edge is “split down the middle” and each half-edge has a vertex pointer, a “next” pointer to the following edge around the relevant face, and a pointer to the other associated half (Figure 7). This permits navigation without stopping to ask which end or face is required.

The quad-edge data structure incorporates the half-edge structures of both the primal (e.g., Delaunay) and the dual (e.g., Voronoi) into one. Figure 8 illustrates this: although it appears complex, it is merely one primal plus one dual half-edge pair linked by a “Rot” (R) pointer with four rotating values – one may navigate from the Delaunay structure to the Voronoi one simply by following R. (N is the pointer to the next edge around a vertex or face, V points to a vertex, and S – not shown – points to the other associated half-edge.) This is particularly
Applications of the Simple DT/VD

Single cell analysis

It has been found that, when the primary objective is the total volume of a resource, such as coal seams or rainfall, the sum of the volumes of the individual Voronoi cells provides a more reliable estimate than attempting to interpolate a smooth surface: assuming that the value at any location equals that of the nearest data point avoids many awkward conditions. An example is the thickness estimation of formations, such as coal, based on drill-hole observations. Rather than estimating the upper and lower surfaces and subtracting them, summing the thickness value times the cell area, for each cell, may give a more consistent overall volume estimate. The same holds true for total rainfall estimation based on rain gauges – as originally suggested in Thiessen (1911).

A related, and common, problem is that of point density mapping. This is usually attempted by counting circles – counting how many points fall within a given circle, and then moving it and trying again: a very small circle, or a very large one, would give implausible results. Noting that point density (points per unit area) is the reciprocal of the Voronoi cell area (area per unit point) greatly simplifies this process.

Context

It can be argued that one of the main problems of spatial analysis – the knowledge of the neighborhood of an object – can be effectively managed by using the Voronoi diagram. In this entry we call this the “context” of the object.

Thus the context – the neighborhood – of the point $P$ in the red cell in Figure 9 is the set of (green) Voronoi neighbors: any minor perturbation of $P$, or even its deletion, will only affect
the green cells. (Conversely, if \( P \) was not initially there, its insertion will only affect the green cells.) This structure resolves the “post-box” problem – to find the nearest point (post box) to any spatial location – as each Voronoi cell contains all the spatial locations closer to its generating point than to any other. This is also the basis for nearest-point interpolation: occasionally – especially in some remote sensing processes – a pixel should take the value of its nearest data point, rather than averaging a set of neighbors, which would blur the boundaries required for object recognition. The VD directly provides this function.

For the simple VD of point datasets the context may well be expressed as the DT: each Delaunay edge states that another vertex is a Voronoi neighbor to the central one. Thus vertices may be visited in any desired order by following the DT edges. One particular property is often useful: the minimum spanning tree (MST) of a set of points is a subset of the DT edges. The MST is the shortest graph (tree) that connects all the points – although there may be others that are equally short. In geographic terms, it would be the shortest total length of road that would connect all the villages.

The VD could, in theory, be created by growing “bubbles” at an equal rate from each of the data points until all spatial locations are filled. This models a large number of natural processes: examples are turtle shells, crocodile skins, mud and basalt cracks, and many more. Of clear geographical interest is the example of many types of forest growth: the trees try to crowd each other out, thus the Voronoi cell area approximates the crown size, which approximately estimates the tree volume.

This approach may also be useful for the extraction of the boundaries of clusters of unlabeled points – and may serve to give them cluster labels if required.

As previously indicated under “spatial models,” a terrain surface may be thought of as a DT, as a VD, or as both together. A prime example of this is runoff modeling, whether of a raster grid or of irregular Voronoi cells. Finite-difference modeling, where a quantity of water moves from each cell to its downhill neighbors during each iteration, is an excellent example of this: it needs both a “bucket” and a “slope.” The bucket, of some known area, serves to contain the water; the slope indicates the downhill velocity. But adjacent buckets have vertical boundaries! (see Figure 5). Thus, even for simple raster algorithms, there are primal and dual concepts involved – the square bucket, and a slope imagined to be between the centers of adjacent grid cells. The Voronoi runoff model is identical, except that the geometric structure of the Voronoi edges and the Delaunay edges must be explicitly preserved. Then, for each time step, the water transfer is calculated from the slope of the Delaunay edge, the area/volume of the Voronoi cell, and the common edge between an adjacent cell pair (Dakowicz and Gold 2007).
Labeled skeleton – grouped context

The Voronoi diagram clearly gives a good estimate of the context of an individual tree in a forest: its constraining neighbors are found directly. It is similarly useful for “labeled maps,” where each Delaunay vertex, or Voronoi cell, is given a label rather than a continuous variable such as elevation. Preliminary geological mapping is a good example: in many areas the geologist has to hunt to find identifiable outcrops that are not covered with overburden, and label them as to the rock type. As nothing is known in the covered regions, the best guess is the closest observed outcrop – hence the Voronoi cell can form the basis for the initial map-coloring exercise. This is still interpolation, although the variable being plotted is not continuous.

To improve the results of such geological maps, boundaries between cells with identical classifications may be suppressed. We refer to this as the “labeled skeleton” to distinguish it from the “geometric skeleton” described below, where the skeleton boundaries are derived without manual point labeling. Estimated geological boundaries are not required between outcrops of the same rock type, so a simple pruning of the Voronoi boundaries will give a preliminary geological map that may be amended by additional geological knowledge. Thus the original simple context becomes a grouped context, with a complex Voronoi boundary around each contiguous set of similarly labeled outcrop regions, as in Figure 10 (Gold 1999).

A similar approach has been found very useful in some major map digitizing projects, such as soil or forest maps and reports, where a large number of contiguous polygons need to be digitized and the underlying map is sufficiently complex that simple scanning will not suffice. In this case the cursor is “rolled” round the interior of each polygon (perhaps with the aid of a small central dot) and a polygon label assigned to each of the points (Figure 11). This is repeated for all polygons on the map. When discarding all Voronoi boundaries between points with the same label – between adjacent points on the polygon interior – the boundaries between the polygons remain – in many
cases with enough precision for the application. Efficient algorithms exist for this extraction and aggregation process (Gold, Nantel, and Yang 1996).

An interesting point is worth noting: because the original set of Voronoi boundaries is topologically connected, and because the same is true for contiguous aggregates of cells, the boundaries between the polygons are guaranteed to be topologically complete, even taking account of digitizing errors such as forgetting a polygon. This is particularly useful in the processing of scanned maps when using the geometric skeleton, as described below.

The geometric crust and skeleton

Before the widespread use of computers, Blum (1967) discussed a method for describing irregular (“biological”) shapes. Using optical defocusing techniques he showed that if “waves” are created — perhaps by synchronized vertical oscillation of a set of generators — the boundaries between the merging wave sets formed equidistant “skeletons” (also called medial axes) between them. For point generators these form the Voronoi diagrams described above. For complex shapes, such as polygon boundaries, the skeleton may be used to describe the polygon’s shape (see Figure 12). Note that both endo- and exoskeletons are produced. Skeleton points are equidistant from at least two closest points on the polygon boundary: in fact they are Voronoi vertices whose circumcircles touch the boundaries. In addition, he suggested that these circumradii could be used to assign heights to the skeleton points: a larger circle gives a greater height (Gold and Dakowicz 2005).

Polygon boundaries are commonly defined as a series of points. If the VD/DT is constructed from these, and they are sufficiently close, then a subset of the Voronoi edges will form the skeleton and a subset of the Delaunay edges will form the “crust” or connected boundary of the polygon. As primal/dual structures, each Voronoi edge will have a matching Delaunay edge, and vice versa. It can be shown that for each edge pair either the Delaunay edge will be part of the crust (and the Voronoi edge can be ignored), or the Voronoi edge will be part of the skeleton (and the Delaunay edge ignored). This can be determined by a simple test: create a circumcircle from the two Delaunay vertices and one of the Voronoi vertices — if the second Voronoi vertex is inside the circle then the Voronoi edge is part of the skeleton (and the Delaunay edge is ignored), otherwise the Delaunay edge is part of the crust (and the Voronoi edge is ignored) (see Figure 13; Gold and Snoeyink 2001).

We will illustrate this with a scanned outline of a maple leaf (Figure 14). The left-hand part shows all the Voronoi and Delaunay edges together; the central part shows the resulting crust and skeleton when the above rules are implemented — note that no Delaunay edges cross Voronoi ones, and vice versa. Note also that the skeleton suggests the locations of the veins within the leaf.

There is one obvious criticism: the skeleton is notably “hairy.” This is because small perturbations in the digitized boundary points produce triangles with small circles that are far away from the true skeleton. The answer is to perturb each boundary point until the tip of each hair retracts to its parent in the skeleton; there is no tolerance involved. The right-hand part shows the smoothed version (Gold and Thibault 2001). This process can also be applied to arbitrary curves or networks, as well as polygons, as illustrated by the river network in Figure 15.

This clearly gives the impression of delineating the watersheds around each river segment. Applying Blum’s height transform to the skeleton points — height equals distance from the
Figure 12  Blum’s skeletons of irregular shapes, and his height transform.
Figure 13  Crust/skeleton testing.
river – implies a 45° slope everywhere, although this may be scaled as desired: not a bad initial approximation. Figure 16 gives a perspective view of the resulting TIN model.

However, we may go further. Each of the data points along the river has its own Voronoi cell: the region whose rainfall would run off downhill and meet the river close to that point (see Figure 17a). In addition, the crust – a subset of the Delaunay edges – gives the connectivity of the river segments and thus the flow of the accumulated runoff. The right-hand figure gives a static view of the overall total flow: approximate, but useful as an initial estimate for flood-risk analysis, despite its large assumptions about geological and surface homogeneity (Gold and Dakowicz 2005).

A related problem concerns terrain interpolation from contours – an old, but still important, data source. A good TIN may be produced by triangulating points on adjacent contours, but problems arise if all three vertices fall on the same contour: “flat” horizontal triangles result, as in Figure 18.

This affects ridges, valleys, peaks, and pits, although common sense can often estimate the desired result. Here again the skeleton helps to fill in the blanks. Skeleton points between contour levels contribute little, but those between points on the same level indicate, as for rivers, the ridge or valley lines (Figure 19a). The problem is to estimate their heights: Blum’s height transform will not do, as there is no overall fixed slope, but we may assume constant slope within a particular valley segment, for example. If this assumption is reasonable, then we may again take advantage of the already-existing circumcircles centered on the Voronoi skeleton points. First estimate the local slope at a skeleton junction whose circumcircle touches two contour levels, and then propagate this slope along all of the valley’s skeleton points – if we know the slope then we can estimate the skeleton height based

Figure 14  VD/DT; crust/skeleton; smoothed crust/skeleton.
Scanned map processing

A final example of the use of the simple incremental VD/DT algorithm and the geometric skeleton is the processing of scanned maps: due to the probable presence of unclosed loops we cannot use the labeled skeleton. Instead, we perform edge detection on the map image, generating boundary points between “black” pixels formed within the interiors of the line artwork and “white” pixels elsewhere. We then construct the VD/DT, and the crust and skeleton, and smooth it. The results for simple text are shown in Figure 20.

The resulting endoskeletons, with their particular loops and branches, provide an initial stage in character recognition. The exoskeletons express the relationships between the characters. If we repeat the exercise for a scanned cadastral map the result is as shown in Figure 21.

This provides the framework for further detailed analysis (Gold 1997). First, the skeletons form the centerlines of the initial line work – and if this was topologically connected then so are they, so in most cases the topological digitizing of property boundaries is automatic. Second, individual skeletons of the “white” regions identify separate map components – in this case mostly buildings and text. Third, the skeletons show the adjacency relationships between these components: the placing of a building within a property, the association of individual characters to form a word or number, and the placing of these labels within a property or building. (Note that all skeleton segments have an associated, and perpendicular, Delaunay edge connecting the two components.) Other applications can easily be envisaged.

Interpolation: the dynamic algorithm

Many traditional surface interpolation methods are based on the weighted-average technique: some set of neighbors to the query point are obtained, some kind of weighting function is calculated (where the weight decreases with the...
metric distance between the neighboring point and the query point), and a weighted average of these neighbors is used to estimate the height of the query point. Various difficulties can arise, especially with supposedly precise data. The biggest of these is that the selection of the set of neighbors is not directly related to their spatial distribution, usually being based on the metric distances, and thus the weighting functions can not readily take account of this distribution: ideally neighbor weights should decrease to zero precisely when that point ceases to be selected as a neighbor, and should reach 100% when the query point and the neighbor coincide. This is not easy to achieve.

An alternate approach is to take account of the Voronoi neighbor relationships – where these include a newly inserted query point, as in Figure 22; the weights are the areas stolen from the original neighbors (Sibson interpolation, natural neighbor, or sometimes area-stealing: Sibson 1982; and Gold 1989). This has two advantages: weights decrease to zero precisely when a data point ceases to be a neighbor, and, as the query point approaches a neighbor, that neighbor’s contribution approaches 100% and the others’ decrease to zero. It can be shown that this approach is “smooth” throughout, except at the data points themselves.

This is really an example of an order-2 Voronoi diagram, where the “weights” are the areas of
those regions closest to the query point and second-closest to the original data point. It may be thought of as the hospital-districting problem: what the consequences would be if the central hospital was incapacitated by fire — or, conversely, if a new one was built at the query point (Okabe et al. 2000).

Figure 23 shows the results of using this interpolation approach on our contour dataset with added skeleton points: the surface is smooth except at the data points — an interpolated point can never have weighted neighbors that only come from one direction. This is also valuable for other anisotropically distributed datasets, such as airborne surveys, ships’ traverses, or, in vertical sections, oceanographic soundings and temperature profiles. In all these cases Sibson interpolation guarantees contributions from a
Often forgotten is that, for many real-world problems, the surface slope is even more important than the elevation – think of insolation or vegetation analysis. However, the above methods may be modified for this purpose by replacing the elevation at each data point by a (usually planar) function. This is then evaluated at the query point. This produces a nearly – but not quite – smooth surface overall (Dakowicz and Gold 2003). Interestingly, if the same approach is applied directly to the TIN model itself, using only planar interpolation, the results are surprisingly good (see Figure 24). (One empirical way to check the surface quality is to interpolate the heights of a random set of points, and use these as the basis for finite-difference runoff modeling, as described earlier: irregularities in the surface will show up as patches of water, or dry areas, at various time steps.)

In summary, interpolation first of all needs the definition of a meaningful set of neighbors, second, a weighting function based on the spatial point distribution, and third, slope estimates.

However, Sibson interpolation needs the addition, and deletion, of individual points; complete set of neighboring data points (Gold and Condal 1995).

[Figure 21 A scanned cadastral map and its topological relationships.]
The kinetic Voronoi diagram

If a Voronoi generator $P$ is moved slightly its Voronoi boundaries with its neighbors will be perturbed. If it is moved sufficiently far then one of these boundary segments will disappear as $P$ moves far enough away – or a new boundary will appear as $P$ moves towards a new neighbor. These are “topological events,” where the diagonal of a pair of adjacent Delaunay triangles is switched to the other configuration. The key idea is that the Delaunay condition – every triangle’s circumcircle must be empty – must be preserved. The moving point $P$ is tested to see if it enters the circle of an adjacent triangle immediately exterior to those with $P$ as a vertex – or, alternatively, if it leaves a circle formed by three contiguous neighboring points of $P$. If so, $P$ gains – or, alternatively, loses – a Delaunay neighbor and the process may be repeated as $P$ moves through the mesh. In Figure 27, $P$ moves towards $Q$ (going from $a$ to $b$) or else away from it (from $b$ to $a$). Care must be taken if two points collide, as all Voronoi generators must have distinct locations.

The kinetic VD has various applications. In Free-Lagrangefluid flow modeling each Voronoi cell represents a fixed “packet” of fluid, which is pushed around by external forces as well as by its neighbors. Thus a smaller cell area implies a greater cell height – for example of water. In another application – a marine geographic information system (GIS) – moving points represent ships, and fixed points the coastline, permitting navigation decisions based on adjacency (Figure 28; Goralski and Gold 2007).

Complex objects

When looking at the scanned map application of Figure 21 it is evident that we are using many
points to represent basic objects. Much work has been devoted to finding a way of working with other kinds of Voronoi generators. Practical algorithms for the line-segment Voronoi diagram, for example, have involved many years’ labor as there are many degenerate configurations where generators meet, nearly meet, or intersect (Held 2001; Gold and Dakowicz 2006). This is unfortunate, as it is, in practice, a very valuable tool. The algorithms involved are complex, and only one – based on the moving-point VD – will be outlined here. We will start by looking at an alternate model based on the DT.
point is preserved it will have all of the historical adjacency relationships of the generating point. If some point $e$ is split off from one data point, moved to another and then merged, we could preserve its track as a triangle edge, even if its properties were no longer Delaunay. From this we could construct, for example, the outline of a house using the “constrained Delaunay triangulation” (Figure 29).

This is used fairly frequently in GIS but it has a particular difficulty: the constrained edges have a dual role, as they must represent objects as well as express the relationships between objects (which the other triangle edges do). This can cause problems. An alternative, but more complex, approach is for these trailing edges to be entities in the map layer itself, rather than in the topology layer formed by the triangles. This requires that the data structure be changed, with the trailing edge being represented by two half lines: this allows the representation of the Voronoi region on each side of the line (Gold and Dakowicz 2006).

Figure 30 shows the Voronoi regions for a set of buildings, which are compound objects.
constructed of points and line segments. Each point, and each half-line, has its own cell, and these are connected by the dual triangulation (not drawn). It is therefore easy to assemble all the connected component parts of each building, as well as their interior cells. Note that the exterior cells together give the Voronoi region of the building as a whole, and the boundaries between the buildings form the skeleton – being the line of maximum distance between them, which is necessary for many navigation problems.

This approach demonstrates that the Voronoi diagram answers most questions about the spatial context of map objects. Worthy of note is that the fundamental query when pointing at a map – “What is here?” – is modified to “What is closest to here?” in the Voronoi approach. Clicking anywhere on the map returns the appropriate response.

Partitioning the map into these proximal regions converts a set of map objects into a “field” – a continuous coverage that always returns a valid response to the fundamental query. In addition, certain other queries may be greatly simplified: buffer zones, for example, can be trivially constructed. By definition they are the map portions within a specified distance of the closest generator. Since all map locations are already assigned to the closest generator, then the “clipped context” is those portions of each proximal region that fall within the specified distance (Figure 31).

It is worth noting that for some applications it is easier to estimate the Voronoi diagram in the...
Figure 32  A unified spatial model based on the VD/DT.
raster domain, and perform the above analyses in that mode. However, such approaches are only approximate – depending on the limitations of raster/vector conversion and the approximation of the Euclidian metric in raster space – and are not rotation invariant (Li, Chen, and Li 1999).

**A unified spatial model**

Throughout this survey there has been one theme: many aspects of 2D spatial analysis may be conveniently expressed in the VD/DT framework. Indeed, it could form the framework for a general spatial data structure that is appropriate for most of the usual geographical data types, and with a common underlying algorithmic basis. Figure 32 illustrates this unified spatial data model for GIS (Dakowicz and Gold 2012).

**Beyond 2D**

Finally it should be remembered that we really live in a 3D world. While the construction of 3D complex objects is still very much a research topic, the 3D point VD/DT is achievable, and is of considerable utility for 3D interpolation problems in such fields as geology and oceanography (Ledoux and Gold 2008). Going further afield, a quick search will discover applications at every scale, from the structure of molecules to the structure of the universe. It is clear that many more applications are waiting to be found.

**SEE ALSO:** Representation: complex objects; Representation: fields; Scale; Spatial sampling; Spatial tessellations

**References**


Spatial optimization

Arika Ligmann-Zielinska
Michigan State University, USA

Spatial optimization uses mathematics to find solutions to spatial decision problems that seek the best arrangement of locations under strictly defined conditions. It is a branch of spatial multicriteria decision analysis, which centers on designing a collection of spatial options and judging how close each of these options satisfies a given set of objectives (Malczewski 1999). Two broad types of problem are considered. First, the decision may involve choosing a location for a particular activity or facility. Second, the problem may focus on selecting an activity for given locations. Either way, space is the essential component of the decision problem, and the best solution offers the optimal design of space under given conditions. Spatial optimization has been used as a prescriptive (normative) tool in a wide variety of geographic applications, including land-use patterning for planning, locating public or private facilities, transit scheduling, moving goods efficiently, finding the shortest path, redistricting schools and voting areas, improving emergency response, locating landfills, designing parks and reserves, and finding the closest facilities, to name a few.

Formalization of spatial optimization concepts and methods in geography can be traced back to Garrison (1959), who applied linear programming to locating manufacturing plants and to solving transportation problems involving the movement of goods from supplying centers to receiving places. Church and ReVelle (1974) presented a general location-allocation model of maximum population coverage of selected services given limited number of facilities. Harvey, Hocking, and Brown (1974) applied a traveling-salesmen algorithm (which determines the shortest route among several locations) to geographic space structuring. Notable publications on spatial optimization include Haggett (1965), Scott (1971), Ghosh and Rushton (1987), Church (2002), Church and Murray (2008), and Tong and Murray (2012).

Any optimization problem consists of decision variables, objective functions, and constraints. The essential element that distinguishes spatial optimization from other optimization problems is the explicit representation of space in the decision variables and functions used to solve the problem. The general formulation of a spatial optimization problem is shown here. Given a set of geographic decision variables \( x \in X \) representing locations:

\[
\text{Optimize} ( f(x) )
\]

subject to a number of conditions \( n \) in the form:

\[
g_n(x) \leq c_n \quad (2)
\]

\[
g_n(x) \geq c_n \quad (3)
\]

\[
x \geq d \quad (4)
\]

where \( c_n \) is a limit value for the \( n \)-th condition, specified as an upper bound (inequality 2) or a lower bound (inequality 3).

Geographic decision variables represent locations that are most often defined using \((x, y)\) coordinate pairs. They can also be defined using three or
four dimensions, especially if the optimization involves time in addition to space (Shirabe 2004). Decision variables can be formalized using any data model, including vector points (Church, Scaparra, and Middleton 2004), lines (Harvey, Hocking, and Brown 1974), polygons (Williams and ReVelle 1996), or raster pixels (Ligmann-Zielinska, Church, and Jankowski 2008; Cromley and Hanink 1999). Function 4 puts technical constraints on the decision variables setting them within a specific range of real or integer values. Geographic decision variables are often binary (Williams and ReVelle 1996), where \( x_i = 1 \) indicates a selected site, and \( x_i = 0 \) signifies a rejected site. The variables can be discrete or continuous, with the latter offering an almost infinite number of possibilities to choose from. For example, a mobile phone tower site selection problem may involve a finite set of candidate locations, or may consider any point in the area of interest. Variables are weighted by coefficients that indicate their relative importance, such as the magnitude of impact, the importance associated with a site, the cost of traversal, or the benefits of selecting a particular site. In general, coefficients are derived from attributes associated with a site or computed based on a relationship between locations. Following the example, if a pair of terms \( 2x_j, 8x_k \) represents the cost of mobile tower installation for two different sites, location \( x_j \) is four times less expensive than location \( x_k \). Furthermore, if \( 200x_{lj}, 50x_{lk} \) stand for distances between an existing tower, l, and the proposed towers, j and k, respectively, tower k would be situated four times closer to l than j.

The term \( f(x) \) is called an objective function, which represents a statement about the desired direction of achievement of a particular decision goal. For example, if the problem is to find the least expensive location to build a new mobile phone tower, the direction of achievement (i.e., optimization) is to minimize the cost. However, if the decision-maker wants to identify a location that serves the largest number of mobile phone users, the direction of achievement is to maximize population coverage. Formally, assuming that the solution to the problem is represented as \( S \), maximization requires that \( S \geq f(x), \forall x \in X \), whereas minimization assumes that \( S \leq f(x), \forall x \in X \). A solution to the optimization problem consists of the optimal value of the objective function and the corresponding values of the decision variables. Due to the complex nature of geographic decision-making, the vast majority of applications use multiple objective functions that reflect the ecological, social, political, economic, legal, and demographic aspects of the problem. In the case of mobile phone tower siting, for example, there may be simultaneous interest in minimizing land acquisition costs and maximizing the coverage of mobile phone users. Multiple objectives are often conflicting, and it is rarely possible to find a solution that optimizes all of them. Instead, multiobjective problems produce a suite of optimal solutions that are noninferior, that is, they cannot be improved on one objective function without degrading another objective (Church and Murray 2008). Noninferior solutions represent tradeoffs among the conflicting objectives and the final choice is often made based on other intangible decision factors.

Constraints (inequalities 2 and 3) are conceptually equivalent to the objective functions because any decision criterion can be either an objective or a constraint. For example, the cost of tower installation can be minimized, or can be set to not exceed a certain threshold. What makes constraints different from the objective function is that they confine the solution space to feasible results, meaning that only the decision variable values satisfying all constraints are considered as potential solutions. Inequality 2 represents cost constraints where lower values are preferable, and
inequality 3 represents benefit constraints where higher values are considered preferable.

Decision criteria used to formulate objective functions and constraints may denote costs (due to activities including engineering, construction, maintenance, transportation), stocks and flows (associated with the amount of goods stored or moved between places), and surpluses (such as monetary and nonmonetary benefits), among others. Most notable, however, are those criteria that reflect topological and nontopological spatial relations, dependencies, and interactions, including distance, pattern, adjacency, connectivity, shape, overlap, containment, and space partition (Tong and Murray 2012; Shirabe 2005).

Addressing distance (proximity) is ubiquitous in spatial optimization applications, such as measuring the shortest path between police stations, the distance traveled from home to work, the fastest route between cities, or the minimum distance necessary to get to the nearest hospital. When operationalizing proximity, it is important to select appropriate distance metrics. The simplest are Euclidean (straight line) and Manhattan (grid) distance measures but, for practical purposes, complex networks of roads, pipelines, power lines, and so on are often utilized. Proximity is usually computed using physical measures (like distance in kilometers), but other types of metrics like travel time are also common. It may be formalized as an objective function (e.g., the shortest path) or as a constraint (e.g., for every residential block, a tornado shelter must be within 10-min drive). Public and private facility location as well as routing are representative applications of distance relations (Church and Murray 2008; Church, Scaparra, and Middleton 2004).

A different type of problem arises from an explicit evaluation of pattern – the spatial distribution of locations, which can be clustered, random, or dispersed. Following the mobile phone tower example, the decision may require the towers be spread relatively uniformly, so that the overall areal coverage of the service is maximized. Conversely, land acquisition for residential development may require clustering of parcels to minimize various infrastructure construction and maintenance costs. Pattern is measured using a variety of approaches borrowed from spatial statistics (average nearest neighbor, spatial autocorrelation, hot-spot analysis), geostatistics (density calculation), and fragmentation statistics (perimeter:area ratio, fractal dimension index). Classic applications include facility dispersion, land patterning, and backup facility siting (Church and Murray 2008).

Clustering often requires addressing issues of adjacency and connectivity. Adjacent or touching spatial objects, referred to as neighbors, share a common boundary. Connectivity (contiguity) is the capacity to move from one spatial object to another without interruption. Both characteristics necessitate a well-defined topology in the underlying data model, so that the geometric intersection of the spatial entities can be evaluated. Contiguity is required in network analysis, and measured as a path between network nodes connected through arcs. Contiguity is enforced using various approaches like evaluating network flows, using dual graphs, optimizing ratio between perimeter length and area of the selected polygons, or evaluating adjacency of spatial units (Ligmann-Zielinska, Church, and Jankowski 2008; Williams and ReVelle 1996; Aerts et al. 2003). A closely related spatial decision criterion is the geometric shape of the solution. Some reserve delineation applications (such as setting aside land for a specific animal or plant species) may require circle-like patches, whereas some habitat corridors may benefit from elongated configurations. Adjacency and connectivity are encountered in a wide array of problems, such as those involving urban growth and land-use planning, nature reserve design, right-of-way
SPATIAL OPTIMIZATION

corridors, transportation network planning, traffic flow management, and evacuation planning.

Other forms of spatial overlap, including intersection and containment, are also common in spatial optimization problems. Polygon intersection can be used in problems that involve service overlap. Critical real-time situations often require some redundancy in service provision (Curtin, Hayslett-McCall, and Qiu 2010). For example, for regions characterized by a high elderly population, some level of overlap could be enforced between two medical emergency service areas, in order to ensure a fast response during peak hours. Intersection can be operationalized using standard spatial analysis methods including point, line, and polygon-on-polygon overlay. Containment is present when one spatial object lies completely within another object. It is commonly used in applications that delineate service areas with focus on the maximum coverage of the target population (retail outlets, banks, hospitals) or in the siting of emergency facilities (Church and ReVelle 1974).

Spatial optimization has long been employed to problems of space partition for administrative, legislative, and other regionalization and zonation purposes (Garrison 1959; Church and Murray 2008). Representative examples include the design of school and voting districts. Partition of space is typically based on population counts (the number of students in a city or voters in a region) and requires no overlaps between districts.

Solution methods to spatial optimization problems are grouped into exact and heuristic categories (Church and Murray 2008). Exact methods evaluate all feasible options and always result in a unique optimal solution, or, in the context of multiobjective decision problems, a unique collection of noninferior solutions (Ligmann-Zielinska, Church, and Jankowski 2008; Aerts et al. 2003; Eastman et al. 1995; Church, Scaparra, and Middleton 2004). They use analytical methods and, therefore, guarantee that the identified solution is indeed the best for the specific mathematical problem. Common methods include enumeration, linear programming, and integer programming with branch and bound (Tong and Murray 2012). Enumeration is a complete search strategy where all feasible solutions are listed and evaluated in order to identify the best option. While enumeration is a simple approach, it becomes unfeasible for problems with a large number of discrete solutions, and impossible for problems with an infinite number of solutions. Linear programming uses a system of linear equations to calculate the optimal function value. While commonplace and relatively efficient, linear programming is of limited use for problems that cannot be formulated with linear functions. Integer programming with branch and bound is used in problems with discrete integer variables by using a divide-and-conquer approach, where the problem is split into smaller and more tractable subproblems. Unlike linear programming, integer programming with branch and bound does not easily handle large numbers of variables.

In contrast to exact approaches, heuristic methods involve rules-of-thumb and ad hoc procedures that result in near-optimal solutions. Since they can handle complex nonlinear functions with a large number of variables in a way that exact methods cannot, they are often an attractive alternative for solving spatial optimization problems. Common approaches include simulated annealing (Aerts and Heuvelink 2002) and genetic algorithms (Tong, Murray, and Xiao 2009). The major drawback of heuristics is that it does not guarantee a truly optimal solution, which may be acceptable in problems that do not exclusively optimize monetary objectives or involve irreducible uncertainties,
such as maximizing ecological suitability in a reserve selection problem or minimizing travel time for a transportation problem. Exact and heuristic methods have been compared in a number of studies (Williams and ReVelle 1996; Church and ReVelle 1974; Harvey, Hocking, and Brown 1974).

While most spatial optimization methods are derived from aspatial approaches, spatial optimization problems offer their own unique challenges, such as the mathematical abstraction of ill-defined spatial problems and integration of optimization with GIS (Tong and Murray 2012). With the ever-increasing complexity of the real-world problems, often defined by fuzzy spatial boundaries and complex measures of social and natural landscapes, it becomes more and more difficult to express variables, objectives, and constraints in a rigid mathematical format. Another problem lies in selecting the appropriate scale of analysis and the representation of the decision variables as points, lines, polygons or more complex entities. Other more conventional yet still important challenges include problems of accuracy and error, and various aspects of uncertainty in spatial data and measurement.

GIS plays a pivotal role in addressing spatial optimization problems (Malczewski 1999; Church and Murray 2008; Cromley and Hanink 1999). It serves as a repository of data used to initialize the variables, a geoprocessing engine that allows operationalization of the spatial properties and relationships of objective functions and constraints, and as a tool for visualizing the results. GIS integration challenges result from the constantly growing amount of spatial data and the higher resolution of the available datasets. As a consequence, the size of optimization problems is surging, requiring computationally efficient solution algorithms that work in real time and can simultaneously handle multiple objectives.

SEE ALSO: Geocomputation; Location-allocation analysis; Spatial analysis; Spatial decision-support system; Transportation and land use

References


Spatial organization and structure

Jiaoe Wang
Chinese Academy of Sciences

Spatial organization is a complicated concept which can be roughly matched to the whole object and subject of economic geography as a scientific discipline, and also figures in different geographic traditions across the world. The idea that geography is the study of spatial organization was emphasized by Edward J. Taaffe (1974). This view sits alongside human–land relations and area studies, and, along with earth science, these are the four geographic traditions suggested by William D. Pattison (1964). Apart from general geographic knowledge about where to locate, spatial organization helps people understand how to organize geographic space (in which space is seen as property or territory) to satisfy their needs and demands. Meanwhile, generalizations about spatial structure can be made due to the existence of similar spatial patterns or landscapes, enabling understanding, prediction, and a capacity to influence human behavior (Morrill 1974).

Spatial organization relates to many research themes examined at different scales, including the distribution of agricultural activities (such as the Corn Belt), industry layout (such as steel factory layout), services location (such as geography of retail shops), the distribution of human settlements, central place and urban systems, the distribution of infrastructure layout, movements of people, and land-use patterns in cities. Undoubtedly, the most interesting part for geographers is the analysis of the structure and organization of cities, as well as the development of urban systems, due to the high population densities and the existence of dynamic socioeconomic activities.

The spaces in which human activities take place are usually studied at three research scales: global, regional, and local. Meanwhile, space can also be considered abstractly as comprising three geometric elements: points, lines, and surfaces. As such, spatial organization patterns can be classified statically into three types: contiguous structures (including uniform regions and functional regions), discrete structures (such as settlements and network cities), and integrated structures that combine both. Contiguous structures are constituted by spatially continuous surfaces (areas or zones) with a functional center (functional regions) or on the basis of similarity in some respect (uniform regions). For example, urban regions and industrial regions are functional regions, but agricultural regions are typical of uniform regions. Discrete structures are constituted by network elements (mainly points and lines) that comprise, for example, transportation-economic complex axis and pole-axis systems. However, some geographic systems are usually a combination of contiguous and discrete structures. This kind of system is called an integrated structure. One example is the major function-oriented zone, proposed by Chinese economic geographers for promoting coordinated regional development and reshaping in an orderly way the spatial structure by regulating territorial functions. The major function-oriented zone includes four major functional categories: optimized development areas; key development areas; restricted development areas; and prohibited development areas. These four
area types are identified on the basis of their different major functions and roles in urbanization and industrialization, ecological construction, grain production, and protection of natural and cultural heritages.

Chronologically, the research of spatial organization and structure could be coarsely divided into four periods: (i) early development in geography (pre-1920s); (ii) emergence of spatial organization as a focus of geographical research (1920s–1940s); (iii) quantitative revolution (1950s–1970s); and (iv) new challenges since the 1980s.

Early development in geography (pre-1920s)

Geography has traditionally sought to examine the spatial organization of human activities and the principles that account for them. Immanuel Kant (1724–1804), a prominent philosopher, proposed that the geographical and historical sciences extend humans’ knowledge, in which geography focuses on space, while history deals with time. An outstanding German geographer, Alfred Hettner (1859–1941), stated that geography was a spatial or chorological science that studied the relationships of diverse phenomena in terms of space or areas. Richard Hartshorne (1899–1992), however, interpreted earlier research as making geography a largely descriptive and interpretative study of areal differentiation and an important part of area studies.

As a matter of fact, geographers have implicitly engaged practically in spatial thinking from an early stage. Two important activities are worth mentioning. One is the design of maps by cartographers to show the locations of villages, mountains, and forests. In order to record every-thing on the map as clearly as possible, symbols representing soil, climate, vegetation, and so on, were used with different sizes, frequencies, spacing, and directions expressed in an orderly spatial structure, even if cartographers themselves did not recognize it. As such, maps were used to express how humankind’s homelands were organized, but with little deep thinking. The other is the work done by historiographers, urban planners, and sociologists, who observed the unique patterns of city shapes as well as inner-city communities, and put forward some related concepts such as Scoria’s linear cities, Howard’s garden cities, and Galpin’s agricultural community.

During this period, German economist Johann Heinrich Von Thünen (1783–1850) and geographer Johann Georg Kohl (1808–1878) initiated the groundwork for location theory. In 1826, Von Thünen published The Isolated State, in which he presented a land-use model, with concentric rings of land use around a central market: an inner ring of intensive agriculture gave way successively to rings for forest, extensive agriculture, ranching, and waste (Haggett, Cliff, and Frey 1977). Von Thünen’s land-use model was closely related to the more recent bid rent theory developed to explain urban land use. Soon afterwards in the 1840s, Kohl developed representation of a settlement pattern from empirical observations, where the settlements were symmetrically arranged along corridors with hierarchies and scales in a planar graph, which looked like a tree. Kohl depicted the vertical and horizontal landscapes of a city, with a declining trend horizontally in the density of economic activities and population from the city center to the peripheries (Berry and Rees 1969) and vertically in the height of buildings, with different uses of the floors of multistory buildings.

Besides Von Thünen and Kohl, economists Wilhelm Launhardt (1832–1918) and Alfred Weber (1868–1958) separately made excellent progress analyzing the location of industries. In 1909, Weber published the Theory of the Location
of Industries, where he formulated a least cost theory which tries to explain and predict the location of industries at a macro scale. The point was to minimize the cost of transportation and labor by analyzing three factors: the material input, labor, and agglomeration economics. These three factors are usually regarded as mechanisms driving the formation of spatial structures.

Emergence of spatial organization as a focus of geographical research (1920s–1940s)

Most European and American countries underwent rapid urbanization in the nineteenth and/or early twentieth century due to the diffusion of successive industrial revolutions. On the one hand, urbanization implied a change of lifestyle and a redirection of research from the countryside to cities. On the other, human beings’ capacity to change the earth’s surface was greatly increased by the emergence of new technologies during this period. For example, the Fordist revolution based on an industrialized and standardized form of mass production rapidly pushed automobiles into people’s daily life (Knox and McCarthy 2005), greatly influencing urban patterns and structures. Great changes in urban areas gradually led to a decline of academic interest in environmental determinism, while socialists and geographers gradually became interested in how to organize space to promote rapid urban population growth.

City growth

Although human–land relations were advocated by a geographer, Harlan H. Barrows (1877–1960), the initial research on the spatial organization of human activities in cities was done by the Chicago School, and in particular by Robert E. Park (1864–1944), Rodericke D. McKenzie (1865–1941), and Ernest W. Burgess (1886–1966) from the perspective of human ecology.

Burgess proposed a descriptive urban land-use model in 1925 that divided cities into six concentric zones expanding from downtown to the suburbs: central business district (CBD), factory/industry zone, zone of transition, working-class housing zone, residential zone, and commuter zone. This model represented the tendency of a city to expand outwards and for land use to change as a result of competition and the growth and relocation of population with different levels of income and belonging to different social groups.

Homer Hoyt (1895–1984) modified the concentric zone model of city growth, presenting a sector model in 1939 through a study of residential areas in the North American context. In this model, transport had directional effects on land uses. As such, various socioeconomic groups tended to expand outward from the city center along railroads, highways, and other transportation arteries. Meanwhile, the city grew in wedge-shaped patterns or sectors emanating from the CBD and centered on major transportation routes.

Following Burgess and Hoyt’s monocentric city model, Harris and Ullman (1945) introduced a multiple nuclei model with more effective generalization of urban land use. Through case studies in London and Chicago, Harris and Ullman found that the spatial structure of large cities was organized with irregular sectors and multiple separate CBDs, which were shaped by different degrees of accessibility, land use, and regional specialization. Satellite settlements are typical examples and follow neither the concentric zone theory nor the sector theory.
Functional regions

Functional regions are usually organized around a center, or focal point, with the surrounding areas linked to that center through systems, associations, and activities. Actually, functional regions were first put forward in the 1920s, when McKenzie argued that the spatial distribution of human beings and institutions was a result of evolution and dynamic functional interrelationships, with all of the elements organized around centers or points of dominance. McKenzie identified the boundaries of US metropolitan regions through the analysis of daily newspaper circulation. At the global scale, the world evolved from fragmented places into a highly linked and centralized pattern during the 1920s and 1930s. Europe and the United States were the world centers and possessed large cities, while oriental and tropical areas were underdeveloped and located on the fringes of the world.

Meanwhile, geographers firstly focused on visualizing the landscapes of functional regions and describing their characteristics through field investigations. For example, a popular approach involved the preparation of atlases depicting land cover and land use and the way they changed. The mosaic of land-use patterns reflected the relationships between agricultural crops and natural conditions in geographic space. However, Robert S. Platt (1891–1964) emphasized the limitation of an atlas, which cannot reflect the relationships (e.g., retail, entertainment, trade, and migration) between different regions. Then the synthetic map appeared with an overlapping and cobweb-like complex structure, focusing on communications in a region (Hanson 1997). The identification of functional regions is a particularly important way of examining the world from a geographic point of view.

Settlements and central place theory

Settlements, as the space in which most human activities are carried out, are the main research field in studies of the spatial organization of regions. The first statement of the rank-size rule on city population was given by Felix Auerbach in 1913, and was written as:

\[ p_i = p_1 / i \]

where \( p_1 \) is the population size of the largest city, and \( p_i \) is the population size of the city ranked \( i \) in the series 1, 2 \( \ldots \) \( n \), in which all cities in a region are arranged in descending order by population (Haggett, Cliff, and Frey 1977). The rule did not receive much attention until the 1920s, when Alfred J. Lotka (1880–1949) restated it, and the 1940s, when it was popularized by the sociologist George K. Zipf (1902–1950). Today, the revised rule, Zipf’s law, a special case of Pareto distribution, is still widely used by geographers to describe how urban systems are organized.

An examination on the rankings of cities at the global scale was first completed by a geographer, Mark Jefferson (1863–1949), who compared the first-, second-, and third-largest cities by population size for 50 countries of the world and found that in most countries the population of the largest city was much higher than that of the next largest. This phenomenon is also known as the law of the primate city, which was defined by Jefferson as being at least twice as large as the next largest city and more than twice as significant. Today, both the rank-size rule and the law of the primate city are still important approaches for statistically analyzing the structure of urban systems.

Central place theory was proposed by Walter Christaller (1893–1969) in 1933 and is a milestone for studying the organization of human settlements. Christaller asserted that settlements simply functioned as central places providing
services to surrounding areas, and built the central place theory with a case study in southern Germany. Central places and their market areas form a hierarchy in which the demand and supply of goods and services of different kinds are (spatially) equilibrated. The larger the settlements are in size, the fewer in number they will be. With quite a lot of assumptions, Christaller deduced that settlements would tend to form a hexagonal lattice, which was the most efficient pattern to serve areas without any overlap. Soon after, in 1940, the German economist August Lösch (1906–1945) expanded on Christaller’s work. Unlike Christaller, whose system of central places began with the highest-order center and focused on maximizing consumer welfare, Lösch began with a system comprising the lowest-order (self-sufficient) farms and created an ideal consumer landscape where the need to travel for any good was minimized and profits were equalized. The central place theory proposed by Lösch was regularly distributed in a triangular-hexagonal pattern. After that, Harris and Ullman (1945) focused their research on investigating the influences of transportation and specialized function on the structure of central places. Central place theory provides an integrated explanation of the number, size, and location of human settlements, and is a comprehensive approach to understanding the spatial organization of urban systems (Hanson 1997).

Quantitative revolution (1950s–1970s)

Starting in the early 1950s, a quantitative revolution occurred in the social sciences as well as in its subdisciplines. Instead of focusing on what made places different, geographers looked for what they had in common and searched for spatial order or rules using deductive and inductive mathematical and quantitative methods (Isard 1956; Berry and Marble 1968; Haggett, Cliff, and Frey 1977). Emphasis was placed on the measurement and mathematical/statistical representation of all elements of the landscape, such as population, industry, market, trade, cities, and land use. A geographer, William Bunge, even insisted on using the science of geometry to interpret space. Most of this research concentrated on the analysis of spatial organization, the structure of cities, and the organization of systems of cities.

Urban structure

Urban structure is recognized as an integration of urban form and spatial interaction as well as organization rules (e.g., land rent). Urban areas are complex regions aggregating arrangements of land use, social groups, and economic activities. In the 1950s, social area analysis and factorial ecology were pioneered by urban ecologists to explain urban structure. These methods were used to examine residential differentiation in preindustrial and postindustrial urban areas and identified a city’s discrete divisions with their distinctive social traits in terms of three underlying dimensions (ethnic status, family status, and economic status) (Berry and Rees 1969). These studies showed, however, that urban structure varied from one city to another, indicating and intertwining of concentric zones and sectors with disturbed mosaic areas.

There are many models of population and land use developed to explain urban organization rules. Colin Clark proposed the first urban population density model in 1951, written as:

\[ y = Ae^{-bx} \]

where \( x \) represents the distance from the city center and \( y \) represents the density of the residential population, which time series and spatial
tests suggest usually follows an exponential function.

In 1964, William Alonso (1933–1999) put forward the bid rent model, which is an extension of the Von Thünen model. The Alonso model states that each land user (including commercial, industrial, and residential) will compete with one another bidding for land, and that land rent is a function of distance to the CBD. The bid rent curve of commercial land use is the steepest, followed by industrial and then residential land use. Based on Clark and Alonso’s work, in 1967 Edwin S. Mills (b. 1928) incorporated a transportation component into land-use models. In 1964, Ira S. Lowry drew on gravity principles in developing the Lowry model to examine the relationships between transportation and land use. This model opened the way to a large volume of work on the modeling of cities. In adapted forms this model is still in use today.

Systems of cities

Central place theory was first tested by John E. Brush (1919–2007) in the early 1950s with an empirical study in southwestern Wisconsin. Then Walter Isard (1956) improved the spatial and regional frameworks of economics through the development of a more adequate general theory of location (including central place theory) and space economy. Thereafter, geographers Brian J.L. Berry, Michael F. Dacey, and John B. Parr and economist Martin J. Beckmann, among others, carried out case studies to re-examine the statistical rules of central place theory and the rank-size rule. Megalopolis, a type of region playing a special role in the organization of urban systems with frequent and close connections, was defined by Jean Gottmann (1915–1994) on the basis of a study of metropolitan areas on the northeastern seaboard of the United States. In the 1960s, G. William Skinner (1925–2008) examined periodic markets and social structures in underdeveloped areas, such as in traditional rural China, and proposed a modification of central place theory.

Different from the static pattern and relations of central place theory, a four-phase model of transportation expansion and the growth of cities was proposed in 1963 by Edward J. Taaffe, Richard L. Morrill, and Peter R. Gould, with case studies in underdeveloped areas such as Western Africa. The first phase consists of a scattering of small ports and trading posts along the seacoast with limited hinterlands. In the second phase, hinterland transportation costs are reduced for certain ports with the emergence of major lines of penetration, and inland trade begins to concentrate in these ports. In the third phase, lateral interconnection begins to move out as feeder lines develop both from the ports and from nodes and subcenters along the penetration lines. In the fourth phase, a transport network and the urban system is formed with a set of high-priority linkages between the largest centers. This four-phase model describes the evolution of the transportation system as well as urban systems from a spatial organization perspective (Haggett, Cliff, and Frey 1977). In developed areas, the coastal transport route usually helps to integrate the port-cities into a megalopolis.

In 1970, James E. Vance (1925–1999) proposed a mercantile model for explaining the trading-settlement structure of regions between North America and Europe, in which transportation and trade roles shaped settlements. Designed to address the deficiencies of Christaller’s endogenetic central place theory, in which the growth of settlement systems derives from an increase in internal demand for traded goods, Vance’s mercantile model is exogenous, in that central place networks evolve in the context of externally based transportation-trading systems. Vance argued that long-distance trade ties resulted in the
formation of new settlements, initially serving as trading entrepôts or “points of attachment,” and that the number of settlements derived from political and transportation conditions. If successful over time, these new settlements became established wholesale centers with their own hinterlands. Subsequent empirical studies have confirmed that the mercantile model applies well to cases of trade-led expansion of settlements in North America as well as in the other parts of the world. Central place theory is suitable for explaining at all times and places a settlement structure centered on the provision of services in an endogenous, closed economic system. Once the system is opened up by long-distance trade or penetration from outside and there is exogenous growth, the mercantile model takes priority over the central place model in shaping the major lineaments of settlement.

New challenges since the 1980s

The economic development of major industrial countries of the world was heavily affected by two energy crises in the early and late 1970s and a related “Fordist” economic crisis. At that time a new international division of labor started to emerge as companies took advantage of developments in transportation and communication technologies as well as the fragmentation and locational flexibility of production, greatly influencing movements of people, commodities, and information. These distant and dynamic interactions in a network society gave rise to what Manual Castells called “spaces of flows,” as a way to “reconceptualize new forms of spatial arrangements under the new technological paradigm.” This concept led to a great deal of research re-examining the organization of space on different geographic scales, with cities in space (networks) as well as cities as places (regions) acting as the critical hubs and nodes in the space of flows.

Global system

At the global scale, transnational corporations began to relocate their production processes from developed countries to developing countries, searching for the cheapest locations to manufacture and assemble components, giving rise to global production or value chains. Commodities, services, financial capital, and technologies moved and diffused globally, and producer services, including business, legal, finance, insurance, and management services, developed very fast. Both the headquarters of transnational corporations and advanced producer services were inclined to concentrate in certain cities. A world city network accordingly emerged from the global space of flows, although world cities were recognized much earlier by Patrick Geddes (1854–1932) in research on Europe.

John Friedmann and Goetz Wolff (1982) first pointed out that the command-and-control functions of global cities were concentrated in corporate headquarters, from which global businesses were run. Then Friedmann (1986) identified a core–periphery pattern and divided world cities into four categories: primary core cities; primary semiperiphery cities; secondary core cities; and secondary semiperiphery cities. At a global scale, New York, Chicago, Los Angeles, Tokyo, London, and Paris belonged to the group of primary core cities. Saskia Sassen coined the term “global city” in 1991 and emphasized that the concentration of advanced producer services in large cities helped them to assume advantageous positions in the global economic system and dominate it. Although both Friedmann and Sassen identified New York, London, and Tokyo as the leading cities in the world economy, neither was quite clear
about the structure of the global system. Peter J. Taylor (2004) identified world cities as well as the world city system with the analysis of inter-city relations in networks. Until recently, the spatial structure of the global system has been evolving into a more uneven structure.

Regional system

Regional integration has resulted in an interlocking network and urban agglomerations as network cities have emerged. A network city evolves when two or more previously independent cities with potentially complementary functions strive to cooperate and achieve significant scope economies aided by fast and reliable corridors of transport and communications infrastructures. The economic area bounded by the cities of Paris, Milan, Munich, and Hamburg in the European Union has become an integrated region with network cities as well as world cities.

China, as a developing country, has seen major changes in its spatial structure over the past few decades. Following the start of the reform and opening-up policies in 1978, a “T”-shaped development model was proposed by economic geographers under the guidance of the pole-axis theory, where socioeconomic activities were mainly concentrated in the eastern coastal area and the Yangtze River area (Lu 1998). Since the late 1990s, the Chinese government has adopted a series of development strategies, such as the Western Development Strategy, Revitalizing the Old Northeast Industrial Base Plan, and the Rise of Central China Plan, in order to reduce inequalities in spatial development. In 2010, the Chinese central government released the Major Function-Oriented Zoning Plan as the guideline for optimizing the territorial development pattern. Under the plan, the whole country is reorganized into a complex and integrated regional system, and the urban system is distributed as “two vertical and three horizontal” axes with different hierarchies of urban agglomerations. The Yangtze River Delta, the Pearl River Delta, and the Jing-Jin-Ji area are undoubtedly the leading hierarchies of urban agglomerations.

Local system

At the local level, human behavior has its most obvious and significant spatial impacts on urban areas. Since the late 1970s, both counterurbanization in developed countries and urbanization in developing countries have progressed simultaneously and finally ended with the strengthening of large cities. In 2010, people living in cities accounted for 50% of the world’s population.

In the large cities in developed countries, people moved from urban areas to suburbs or rural areas as a reaction to overcrowding, traffic congestion, air pollution, and so on in the inner city, and the population of large cities dropped quickly. Meanwhile, with the prevalence of information and communication technologies, the urban form was splintered in terms of land use, and the galactic metropolis came into being, with a salient landscape – the growth of the edge city as a part of urban agglomerations. Since 2008, re-industrialization and the relocation of industries have appeared due to the impact of the global financial crisis. Mobility, interactive relations, and inclusive environments are making urban communities more diverse.

In most developing countries urbanism is in an upgrading stage, according to the urbanism theory proposed by Ray M. Northam in 1975. For example, 47% of China’s population lived in cities and towns in 2010. The equivalent figures in India and Egypt were 30% and 43%, respectively. With large influxes of population from rural areas to cities, local systems are diverse due to their development history (e.g., a colonial
past), culture (e.g., Islam), economy, and other factors. For example, the combined industrial and residential communities in Chinese cities have gradually disappeared due to the adoption of a market-oriented housing policy since the 1990s and a change of focus in city development from the inner city to the outer city, as in most Western countries. Today, however, many municipal governments in China have realized the importance of renewing the historical areas in the inner city in terms of cultural heritage conservation, modifying urban structures. In addition, the emergence of Desakota regions has been examined in many countries of Asia. The Desakota region is a region with an intense mixture of agricultural and nonagricultural activities that often surround a core city or stretch along corridors between large city cores (Knox and McCarthy 2005).

Spatial organization is one of the most important research themes in the science of geography and has played an increasingly important role in understanding and examining changes in the organization of human activities on the earth’s surface, particularly in urban structures, systems of cities, and functional regions at different scales. The patterns of spatial structures vary with different organizational models: contiguous structures, discrete structures, and integrated structures. The study of spatial organization and structure, however, is facing new challenges in seeking to understand how to organize geographic space to satisfy human beings’ needs and demands in an interdependent, uneven, and diverse world.

SEE ALSO: Central place theory; Human ecology; Polycentricity; Representation: geographic systems; Spatial analysis; Spatial concepts; Spatial context; Spatial interaction; Urban geography; Urban mosaic

References

SPATIAL ORGANIZATION AND STRUCTURE


Further reading


Spatial resolution

Peter M. Atkinson
University of Lancaster, UK
Queen’s University, Belfast, UK
University of Southampton, UK

This entry considers the topic of spatial resolution from a geography and environmental science perspective. As this is an encyclopaedia of geography, the entry is written at a general level for a wide audience, including the general public. For a more detailed consideration of the topic of spatial resolution, readers are referred to the book by Zhang, Atkinson, and Goodchild (2014).

Geography is inherently spatial in nature. Topics studied range from a spectrum of environmental phenomena and challenges, including understanding environmental processes through paleoecology, geomorphology, hydrology, biogeography, landscape ecology, and spatial epidemiology, through to social and economic analyses. In all cases, the focus is on the spatial environment, human spatial constructs, and the interplay between them (socioecological interactions). For certain subdisciplines this extends to interactions between humans, animals, and environment, for example in the fields of conservation biology and One Health. Increasingly the focus of geography is also being extended to the temporal dimension, without losing its spatial underpinnings, resulting in space–time geographical inquiry. Such space–time investigation has been a natural part of some subdisciplines for a substantial period for example paleoecology, but is only just being enabled in others.

In all of the fields of inquiry noted, there are two basic ways of knowing about the real world. The first and most fundamental is “measurement.” In the context of research, this can mean measurement followed by induction of a hypothesis that explains potentially some unexpected observation, and measurement designed within the frame of an experiment (and control) such that the difference in outcomes is a function only of the experiment. It should be obvious that here measurement has been placed into the inductive and deductive scientific research methodologies, respectively. Abstraction of ideas is also possible.

The second route to understanding is “modeling,” which involves the fitting of logical and mathematical constructs to data to generalize their information content, and also, through their application, to constrain what can possibly be true given the data and the model; that is, to predict. Here, the model, no matter how poorly, encapsulates our understanding of the processes in the real world (usually directly, but also indirectly, for example through statistical models), and differences between predictions and observations allow potential refinement of our understanding. The focus of this entry is measurement, but models are also invoked that capture and test our understanding of spatial resolution.

Spatial resolution may seem an unrelated concept to the topics discussed so far, but it will be seen that all measurement and all models are dependent on it. Spatial resolution is a fundamental concept in geography, in science, and in our everyday lives. How strange, then, that until recently it has rarely been considered. Spatial resolution is now introduced in the context of remote sensing.

Spatial resolution has gained interest as a subject of investigation in geography primarily as a result of the advent of remote sensing.
SPATIAL RESOLUTION (Atkinson and Tate 2000). Spatial resolution is a term associated with images, including remotely sensed images, and commonly it is used to refer to the pixel size, where the pixel is generally a grid cell in an image, is square, and, crucially, has a specific size represented as the length along one side of the square. The choice of spatial resolution in remotely sensed imagery is important because it determines the information content of the image. Thus, via the launch of early satellite sensor systems such as the Landsat multispectral scanning system (MSS) (with a spatial resolution of 79 m), the Landsat thematic mapper (TM) (with a spatial resolution of 30 m), and the National Oceanographic and Atmospheric Administration’s (NOAA) advanced very high resolution radiometer (AVHRR) (with a spatial resolution of 1.1 km), the term spatial resolution was popularized among scientists. More recently, digital cameras and online services such as Google Earth™, which allow rapid zooming in and out of geographical images and data, have brought the concept, and the meaning, of spatial resolution to the general public.

The two sensor systems described above (Landsat, 79 m and 30 m; AVHRR, 1.1 km) illustrate a key issue in the design of satellite sensors: a trade off has to be made between spatial resolution, spatial extent, and temporal revisit frequency. Landsat TM has a swath of about 185 km and a revisit frequency of 16 days. NOAA AVHRR has a swath of about 2600 km and a revisit frequency of once or twice per day. Thus, the former is rich spatially, whereas the latter is rich temporally. This constraint also means that spatial resolution is determined at the design stage with specific applications in mind; it is not an unconstrained choice made by users. From a user’s perspective this means that sampling choices are, in fact, usually a choice made from the available satellite sensors (of which there are now many), with spatial resolution crucial among the parameters defining the ultimate choice.

One of the key drivers that has opened up the possibility of space–time analysis in geography is the long time series of images now available from several long-term satellite sensor systems, such as Landsat and AVHRR (of the order of 40 years), and MODIS and MERIS (of the order of 12 years) (Wulder et al. 2008). Such data have very recently supported analysis of changes on the surface of the Earth and long-term monitoring. Interestingly, monitoring was the original objective of programs like Landsat, but it was not until a sufficient time series had been built up, and challenges in processing had been addressed, that analysis of change could really begin. Again, spatial resolution is a key parameter that needs to be selected in support of time-series analysis.

In the next section, remote sensing is compared with traditional field measurement and differences drawn out that help to illuminate the meaning of spatial resolution through its context.

The sampling framework

Traditional field measurement in geography and environmental science usually refers to point or pseudo-point measurement – for example, cores in paleoecology and soil science, or quadrats in biogeography and landscape ecology. Such field sampling is characterized by being expensive to acquire and expensive to process in laboratories. Moreover, it is also often difficult to acquire such data in the first place and sometimes impossible due to accessibility constraints, which can lead to biases in estimates.

The focus of sampling in field survey is to design a sampling framework focusing on several parameters (Table 1). Limiting the analysis to the spatial domain initially, the first thing that needs
to be defined is the spatial extent; this has a size, geometry, and orientation. It also needs to be anchored to a specific coordinate system via a spatial position (Table 1).

Once the spatial extent of interest is well defined, it is necessary to design the sampling: this requires definition of the size of sample \( n \) required, definition of a particular sampling scheme (e.g., random, stratified random, systematic square grid, systematic triangular grid – i.e., a specific geometry), and a decision about orientation. It is also necessary to fix the design relative to some coordinate system (Table 1).

Finally, although often neglected, there is a third component of the sampling framework: the support (Journel and Huijbregts 1978; Openshaw 1984; Atkinson and Tate 2000). This refers to the size, shape, and orientation of the space over which individual measurements are obtained. The support is a fundamental concept, just like the concept of spatial resolution, but note that the two are, in fact, different, a point that is returned to later. Once measurement has been made over the support, the measurement process cannot be reversed and there is a loss of information.

In field survey, the goal is often to choose a sample size \( n \) so that it provides sufficient precision (represented here by the standard error, SE) when estimating a mean through, for example,

\[
    n = \left( \frac{\sigma}{SE} \right)^2
\]

where \( \sigma \) is the sample standard deviation. However, this equation applies strictly only to independent data, which invalidates it for application to spatially varying environmental properties. Alternatively, increasingly the goal has been to provide a map through interpolation techniques such as kriging, in which case the goal of sampling design can be centered on achieving some specific function of the locally estimated prediction variance (referred to as the kriging variance). In the latter case, it has been shown that systematic schemes are most efficient, and especially more so than random sampling. The goal then is to select \( n \) such as to ensure a greater than some minimum level of kriging variance everywhere, given a systematic scheme. The use of covariate data to aid in mapping through regression approaches is now common and needs to be incorporated into sampling design.

The above efficiency of design is usually the focus of sampling in the environmental sciences, especially given the associated costs. However, in remote sensing the focus of sampling design, as has been seen already, is the spatial resolution. In remote sensing, then, once the study site extent is defined, the focus of sampling design effort is on the spatial resolution, which defines the entire sampling framework. This switch in thinking arises because the extent and spatial resolution are the only parameters remaining to be defined through the single choice of a particular sensor image, as complete coverage is provided within

| **Table 1** The parameters of the sampling framework, illustrated with a hypothetical sampling strategy. |

<table>
<thead>
<tr>
<th></th>
<th>Size (density)</th>
<th>Geometry</th>
<th>Orientation</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extent</td>
<td>100 km (NA)</td>
<td>Square</td>
<td>North-to-South</td>
<td>Top left corner at ((x, y))</td>
</tr>
<tr>
<td>Sampling design</td>
<td>( n = 10,000 ) (1 per 1 km(^2))</td>
<td>Systematic (square grid, evenly spaced)</td>
<td>North-to-South</td>
<td>First point at center of implicit top left grid cell</td>
</tr>
<tr>
<td>Support</td>
<td>10 m radius (NA)</td>
<td>Circular disc</td>
<td>NA</td>
<td>Center of implicit grid cells</td>
</tr>
</tbody>
</table>
a (commonly) square extent on a fixed raster grid scheme, \( n \) is defined by the extent and resolution, and the orientation is predetermined. The rationale for selecting a given image with specific spatial resolution is commonly to ensure that the desired information is captured in the image (i.e., sufficiently resolved).

The next section explores the meanings of support and spatial resolution, and the differences between them.

**Defining spatial resolution**

It is fairly obvious that the concept of spatial resolution in remote sensing overlaps directly with the concept of support in sampling design. However, the support and spatial resolution are not the same. Referring to Table 1, it can be seen that the concept of resolution relates partially to the support but also to what is defined here as the sampling design, including the sample density and sampling scheme, with orientation having relevance only if anisotropy is present in the property of interest. In fact, resolution refers to the smallest object that is resolvable in a dataset, and this arises as a function of both the support and the set of distances between observations which, themselves, are determined by the sampling design. The reason for the confounding of the concepts of support and resolution in remote sensing should now be obvious: complete cover ensures that the support size and spatial resolution are the same in magnitude. There is another reason that the two concepts are often confounded and that is that the cells on the retina of the human eye also provide complete coverage in the form of an image.

It is useful at this point to provide greater definition of the two fundamentally important concepts of support and spatial resolution. The support is important because it is always present, not only in all data but also in everything that we can ever know about the real world through our own senses. The support ensures that measurement is made as an integral of some underlying reality over a positive finite space. This integration is important as several properties are associated with it, including that it is a forward not inverse process and, in general terms, the concepts of the central limit theorem apply. All measurements have a support as it is impossible to measure on a point. Evidently, cores are made over a positive cylindrical volume in 3-D and quadrats lead to an average over a square space in 2-D. These issues are returned to later.

The spatial resolution, on the other hand, is important because it limits the information content of data. All information exists in the relations between data values, not in them. Thus, if presented with a single value, say, 65, it is strictly of zero information content on its own. It means strictly nothing. Of course, your brain will try to interpret the value in relation to values or information that you already hold. So you might try to guess that the value relates to age, for example, or perhaps the age of retirement in some countries. While this may be true, the information content does not lie in the value itself. It lies in the relation between the value and the context provided by your brain. Thus, spatial resolution, as a function of the support and the geographical distances between data, controls the information content of the data, which exists in the relations between them. Importantly, the nature of the relation between environmental data is commonly a function of geographical distance.

While the support can be thought of as an averaging process, stealing all information within the support and replacing it with a noninformative (on its own) single value, the spatial resolution determines the information content between data as a function of the support and the geographical distances between them. Thus, the support and spatial resolution act as a smoothing filter through which low frequency spatial
variation can pass, but through which higher frequency spatial variation does not (it gets averaged out). The information content of the measured data is, thus, a function of the combination of the support and spatial resolution, within the given spatial extent.

To understand the role of the support and spatial resolution it is necessary to consider a third property: spatial variation. While it is assumed that spatial variation exists, in reality this cannot be shown; spatial variation only ever exists in measurements or data. Consider a simple image of say 100 × 100 pixels each of 1 km on a side. The spatial variation (and potential information content) of the image lies in the relations between the data values (i.e., in the set of data values) that, themselves, can be defined as a function of the sampling framework and some underlying external reality. Note that while the former is known the latter is only an assumption (albeit one that all scientists embrace). This relation is general, and is captured by the equations (1) and (2):

\[ I_z = f(z_s(x)) \]  
\[ z_s(x) = f(\theta, R, \varepsilon_s(x)) \]

where, \( I_z \) represents the information content of \( z \), \( z_s(x) \) represents a variable defined at a vector \( x \) of locations in space on support \( s \), \( \theta \) represents the vector of parameters defining the sampling framework (Table 1), \( R \) represents some underlying reality which is not observable without sampling, and \( \varepsilon_s(x) \) is a vector of measurement errors.

To bring home the fundamental concepts, consider zooming in and out. This is, of course, now done readily for Earth viewing using services such as Google Earth. However, the concept is entirely general and applies also to our visual system directly. If we walk away from a particular scene, eventually we are no longer able to see components of it (e.g., ants, then bees, then birds, then mammals, etc.). We are comfortable with this zooming property as it is fundamental and always present. As we zoom out, more of the underlying variation in reality is obscured through the measurement process and the less there is left to provide information to the observer. We do not notice this generally because as the support increases (and resolution decreases) the extent also increases, meaning that we are always afforded a complete “scene”: the information lost within the coarsened support of the old scene is replaced by new information in the added part of the new scene gained through the increased spatial extent.

The most important point is that the spatial resolution (for imagery, also the support size) should be chosen in relation to the spatial variation that needs to be captured in the data. Generally, the spatial resolution needs to be fine enough to capture the important information. This simple message underpins attempts to define an appropriate spatial resolution for remotely sensed imagery (Woodcock and Strahler 1987; Atkinson and Curran 1995). In relation to this, it is worth remembering that the spatial variation that is filtered out by the support is of high (relative to the support) spatial frequency and that which remains between the data is of low spatial frequency; thus, the filtering is specific.

The above concept is general and can equally be applied to temporal variation. Just as humans require the aid of either microscopes or telescopes to “zoom in” to see beyond “human” spatial scales (i.e., spatial resolutions), human comprehension of the dynamics of fast processes such as a fly flapping its wings and slow processes such as a river changing its course on the earth’s surface requires transformation. This transformation is required in order to match the (in this case temporal) resolution to the temporal variation that needs to be captured in the data. In the former case, high frequency (i.e., fine temporal resolution) video allows the capture of wing beats, which when slowed down sufficiently are
SPATIAL RESOLUTION

comprehensible to humans. Likewise, long time series of remotely sensed imagery, as discussed in the introduction, are of sufficient temporal extent that, when speeded up, they allow comprehension of fundamental Earth processes for the first time, particularly through the power of data animation. Interestingly, this comprehension arguably comes because of the specific part of the human visual system that is devoted to movement, as discussed below. Gupta, Atkinson, and Carling (2013) provide an example for the River Ganga in India. Yet more impressive, Google Earth Engine is now bringing this type of animation to a wide audience.

Given the above discussion, it should be clear that the sampling (extent, support, sampling design) imposed by the sensing systems of the human brain fundamentally limits everything that we can know about reality. Interestingly, spatially different parts of the human visual system (which itself is a large part of the human brain) have different functions. Some brain regions are devoted to interpreting color, while some are devoted to interpreting shape, and others are devoted to interpreting movement. It is interesting to consider that such brain functions also fundamentally limit everything that we can know. If our brains are not evolved to sense something we will be forever ignorant of it, even if it exists externally. However, this is beyond the scope of the present discussion, which is focused on the effects of spatial resolution and the support.

It is now time to turn attention to how these basic concepts can be put to good use. The focus of the next section is on changing the support.

Operations

Upscaling and downscaling

The understanding characterized above has been used to develop statistical models that allow scale changes, or scaling, also referred to as “change of support” (Cressie 1996; Atkinson and Tate 2000). This operation falls into two classes: upscaling, which involves increasing the support size, and downscaling, which involves decreasing the support size. It is worth spending some time considering these goals in order to capture what is needed for their solution.

Upscaling is a forward process and, thus, it is tractable. It involves averaging or integration and it is, thus, similar in nature to measurement. Once a scaling function has been defined (which commonly is the simple linear average), upscaling is straightforward given the fine support data if complete coverage is provided. In remote sensing, upscaling is an important process, as it is frequently necessary to compare coarse spatial resolution imagery to sparse (pseudo-) point ground data for calibration and validation purposes. Where the pixel size is large (e.g., >1 km) the link between sparse point data and pixels can be tenuous. In these circumstances, it is common to use imagery of some fine spatial resolution (e.g., airborne imagery) to provide a bridge through which it is possible to scale from points to coarse pixels. The strategy is simple: (i) relate the ground data to their corresponding fine pixels directly through regression; (ii) average the fine spatial resolution image pixels within the coarse pixels; and (iii) relate the averaged fine spatial resolution image pixels to the coarse spatial resolution pixels directly through a second regression.

Downscaling, in contrast, is the opposite to upscaling. It involves attempting to disaggregate a single value into a series of values associated with smaller supports. It is an inverse process and is ill-posed, meaning that there are many possible solutions that will, when upscaled, give the same result. In fact, the condition of coherence, that the fine support solution when upscaled will equal the starting large support value, provides an important constraint on the downscaling
process. Solutions without further constraint are numerous and of little use. For example, taking imagery as an illustrative example, in the continuous case it is possible to imagine a fine spatial resolution image of noisy appearance with values changing haphazardly, constrained only by the coarse resolution image. All these solutions are valid under the coherence constraint. In the categorical case it is possible to imagine the colors of a fine spatial resolution image changing color haphazardly without spatial structure. Neither of these situations is particularly useful.

It turns out that information that can be used to further constrain the solution using optimization type approaches is available. The very fact that the images appear noisy means that they are not plausible solutions. They do not accord with our a priori expectations of what the fine spatial resolution image should look like. This kind of thinking led to two classes of solution to this problem. The first is based on Tobler’s first law of geography, that things near to each other tend to be more similar than things that are far apart, also referred to generally as spatial dependence. Since all phenomena on the earth’s surface are spatially dependent, at least at some scale (if they were not there would be no structure to the world at all, and all noise would be averaged out through the integration of the measurement process – a complete loss of information), it is possible to generate more realistic solutions by maximizing the spatial dependence in the realized fine spatial resolution data (Atkinson 2005).

The second solution is more general and involves characterizing our a priori expectations in some (usually spatial statistical) model. The goal of the downscaling process is then to match the character of geographical variation in the realized fine resolution data to the prior model. Examples of tools used to characterize the variation include the variogram (a two-point statistic that extends the concept of the variance to space by making (semi)variation a function of the vector of separation), the two-point histogram (which captures double the information in the variogram), and multiple-point statistics (which captures the full available information, but is less readily interpreted) (Strebelle 2002; Mariethoz, Renard, and Straubhaar 2010). To bring these ideas to life, imagine that the goal is to downscale an image of a suburban area from a spatial resolution of 30 m to 4 m. Prior expectation about the nature of the size and shape of buildings, as well as the pattern between buildings, will help to greatly constrain the solution.

The above solutions are interesting because the information content implied by the assumption of maximizing spatial dependence and some prior model is tiny compared to the potential information content of the downscaled images. This means that something fundamental has been captured about the nature of both the data and the scaling process.

As an alternative to optimization approaches, geostatistical models and related approaches have been used to downscale imagery and other data directly. Geostatistics is a branch of mathematics that deals with spatial data (Journel and Huijbregts 1978). Kriging involves prediction of unobserved locations as a linear weighted sum of neighboring data. A function known as the variogram is used to determine the weights applied to the neighboring data. This makes intuitive sense because the variogram captures the character of spatial variation and spatial dependence: locations that are near to the prediction are given more weight but, more than this, the exact weights are determined by the magnitude of dependence. However, it also makes statistical sense: by letting the data speak for themselves, the prediction is both optimal and unbiased under the constraint that it is a linear solution.

Kriging has been adapted to allow spatial prediction of continuous values defined on a support
SPATIAL RESOLUTION

that is smaller than that of the original data. This process is termed area-to-point kriging (ATPK) (Kyriakidis 2004; Atkinson, Pardo-Iguzquiza, and Chica-Olmo 2008; Goovaerts 2008). ATPK extends kriging to the downscaling case. It requires the variogram not only between the original data at the large support size but also between the small supports to be predicted along with the cross-variogram between the large and small supports. These can all be predicted as a function of the punctual or point variogram. Unfortunately, the punctual variogram is not measurable directly for the reasons elaborated above as a central message of this entry. Thus, it is necessary to estimate the punctual variogram initially through inverse methods. While this is an ill-posed problem, and many possible solutions will lead to the same observed variogram when convolved or regularized, solutions based on eliciting expert opinion have been proposed.

Scaling the model

Beyond upscaling and downscaling the data, geostatistics also provides a model of the scaling itself. This is conceptually important, as it helps to illuminate the nature of the geographical scaling process itself, and, therefore, it is reviewed briefly here.

In classical statistics, the variance of a group $s^2_G$ is defined as:

$$ s^2_G = \frac{s^2}{n} \tag{3} $$

where $s^2$ is the original sample variance of the original data and $n$ is the number of observations being grouped. Thus, the variance of any new set of data obtained by averaging subsets of the original data can be predicted as a simple function of the original variance and the group size. The potential for application of this equation to the measurement process should be clear: both deal with averaging. However, this equation does not apply to the geographical scaling problem because it assumes statistical independence between the original data, which clearly is not applicable to geographical data, which as already explained tend to be spatially dependent, at least at some scale. Thus, the equation needs to be replaced with an alternative that takes into account the statistical dependence between data that arise as a function of proximity in geographical space.

Using the geostatistical variogram it is possible to predict changes in the variance with change of support directly. First, an equation is introduced that describes the process of coarsening the support of the variogram itself.

$$ \gamma_v(h) = \tilde{\gamma}(\nu, \nu_h) - \tilde{\gamma}(\nu, \nu) \tag{4} $$

In words: the variogram $\gamma_v(h)$ (think variation, or even potential information), which relates the semivariance to lag vector of separation $h$ obtained on a support $\nu$, is defined as a function of the punctual (or point support) variogram $\tilde{\gamma}(\nu, \nu_h)$ integrated over two supports $\nu$ and $\nu_h$ of equal size separated by a given lag vector $h$, minus the double integral of the punctual variogram $\tilde{\gamma}(\nu, \nu)$ over the support $\nu$ overlapping itself without geographical separation (Clark 1977; Journel and Huijbregts 1978). Note that since all data are defined on a support $\nu$ of positive size then so too is any variogram $\gamma_v(h)$ estimated from observations.

Equation (4) means that it is possible to scale the model rather than the data; the variogram constitutes a fundamental parameter set defining a random function, which is central to geostatistical theory and geostatistical operations such as kriging. Thus, equation (4) has captured something fundamental about the measurement process, and the support, which affects all data.

As explained above, the punctual or point variogram can never be measured directly. Thus, it is
necessary to estimate the punctual variogram initially through inverse methods. While this is an ill-posed problem, this does not detract at all from the power of the regularization equation when applied in the forward sense.

Given the punctual variogram it is possible to estimate the dispersion or sample variance $s^2_v$, that would be observed on a support $v$ within a region of interest or extent $R$ using equation (5):

$$s^2_v = \tilde{\gamma}(R, R) - \tilde{\gamma}(v, v)$$

where $\tilde{\gamma}(R, R)$ is the integral of the punctual semivariance over the full domain of interest (the extent) and $\tilde{\gamma}(v, v)$ is the integral of the punctual semivariance over the support $v$, as above. This means that by replacing the classical equation with an alternative based on the variogram it is possible to predict the effect of changes in the support on the sample variance in geographical data. Importantly, if we upscale or “zoom out” we can predict precisely how the sample variance will decrease, where the classical equation fails. Since the variance represents the variation between data values, this means that as we zoom out we are able to capture the measurement process (and the loss of variation, and information, to within the support), and predict the variance (and, thus, potential information), remaining in the data. This scaling process is entirely general and affects everything that can be known about the real world.

Many other interesting tools have been used to explore the concepts of change of support. These include plots that capture the effect of scaling on data such as the scale variance, local variance, and dispersion or sample variance plotted against support size (or spatial resolution in remote sensing). Fractal concepts and wavelet analysis are also central concepts in discussions about scaling, but there is not space here to cover them. Readers are referred to Deidda (2000) for an example involving fractals in space–time.

Discussion

Having considered the change of support problem, attention is now returned to spatial resolution. As explained above, for images and the human visual system, the support and the spatial resolution change in tandem. This means that a discussion of change of support reveals simultaneously the effect of (i) the support (averaging out higher frequency spatial variation within the support) and (ii) spatial resolution (determining the variation and, therefore, potential information content of the data after averaging). However, this is not always the case. Sometimes, the support is very small in relation to the set of distances between the data (Table 1), for example, in field survey. In this case, the information content in the data is determined only partially by the support and more so by the set of vectors between data locations, relative to the underlying spatial variation.

At this point it is worth introducing the possibility of varying the spatial resolution (specifically, the set of distances between the data points) spatially. Since such “nonstationary” sampling schemes can be expensive it might be wondered why anyone would consider it. The answer relates to the simple goal of ensuring that the spatial resolution is sufficiently fine to reveal the variation of interest (in the observations). Spatial variation, or more precisely the character of spatial variation, is not generally uniform across space. A simple example is given by elevation, which commonly manifests as smooth, planar floodplain, rolling foothills, and rugged mountainous terrain; the variation in elevation varies locally. The spatial resolution required to capture such local variation must adapt to each local situation. This is exactly what happens when a digital elevation model is created using the triangular irregular network (TIN).
Future work is likely to focus on extending the change of support problem to the temporal dimension, as suggested by the opening paragraphs on the nature of geography. In particular, the emphasis of downscaling research is beginning to turn to the problem of downscaling in space and time simultaneously. This problem is both challenging and of real potential value in remote sensing (Zurita-Milla et al. 2009). For example, many satellite sensors can be combined with different spatial resolutions. However, the images with coarse spatial resolution, as suggested above, commonly have a fine revisit frequency, whereas the fine spatial resolution images tend to have a low revisit frequency. Thus, a natural space–time downscaling goal is to integrate these two sources of information such as to produce fine spatial resolution imagery with fine revisit frequency. The new Sentinel series of satellites is a case in point, providing data at different spatial resolutions and revisit frequencies simultaneously.

An interesting reflection can be made in relation to the tradeoff between resolution and revisit frequency as discussed previously. It is not feasible to monitor the whole world at very fine spatial resolution (e.g., submeter) and fine temporal frequency (e.g., minutes) because (i) the data volume would be enormous, (ii) the computational burden would be prohibitive, (iii) an individual human brain would not be capable of even looking at all the data, let alone analyzing it, and (iv) depending on the goal, most places are not changing at such high frequency anyway. Thus, it is interesting to consider the design of a system that would allow coarse spatial resolution monitoring combined with fine spatial resolution investigation where necessary, for example, using pointable sensors.

The human brain provides a guide in this case. A part of the human visual system is devoted to movement, and this is connected to the whole field of vision, which has a surprisingly coarse spatial resolution. (Try holding an unidentified playing card at arm’s length and slowly bring your arm around to in front of your eyes. See how long it takes you to identify the card without moving your eyes; you should recognize the card at about 2° from your direction of sight.) However, within this general field of view the human brain is excellent at spotting movement. Once movement is spotted the eyes quickly move to it and settle on it, thus bringing the very fine spatial resolution associated with our central vision to bear on the movement (the bit that allowed you to identify the playing card).

A similar system could be envisaged for monitoring changes on the surface of the Earth. While the association between our visual system and monitoring the Earth may not be obvious, it is worth noting that our visual system developed in this way over evolutionary time as a means of responding to potential threats. The earth’s natural environment has no such mechanism and yet, as a species, we depend utterly on it as a life support system. Thus, ever more sophisticated systems are likely to emerge to monitor our interactions with the environment in the present and recent times and the solution proposed above is potentially highly efficient.

Whatever the solution to monitoring our rapidly changing earth’s surface, understanding the nature of the data (what has been captured in the relations between the data and, more importantly, what has been lost through measurement), and specifically understanding the measurement and scaling processes implicitly present in such data, will be central in both designing appropriate surveillance systems and in interpreting correctly the results.

SEE ALSO: Critical spatial thinking; Fractal analysis; Geostatistics; Interpolation: areal; Interpolation: inverse-distance weighting;
Interpolation: kriging; Map generalization; Modifiable areal unit problem; Optical remote sensing; Representation: fields; Resampling, raster; Scale; Spatial analysis; Spatial sampling

References


Further reading

Samples are selected from populations – such as 100 of 100,000 persons, 20 soil samples from a park, 50 questionnaires at a school – because inspecting the entire population is either impossible, too time-consuming, or too expensive. To obtain meaningful conclusions, sample units should usually be random selections. Non-randomly selected samples may lead to biased estimation, although a biased sample may be remedied by some statistics.

In contrast with an exhaustive survey, a sampling survey is quick, inexpensive, and accurate if the sampling is well designed to represent the target population or if a small, high-quality dataset is used rather than a big dataset with uncertainty. Sampling techniques are thus widely used across a broad range and at various levels of environmental disciplines, such as agriculture, geology, soil science, ecology, oceanography, forestry, and meteorology, as well as socioeconomic disciplines, such as human geography, spatial econometrics, epidemiology, and spatial planning.

Spatial sampling draws a sample from a population with geographical or locational references. The sample is then run through an estimator to infer the parameters of the target population, for example the map, mean, or total values, or the coefficients of population regression. A good spatial sampling method chooses a small sample to infer parameters of population with high accuracy. The size of the sample and the accuracy of the sample estimation are dependent on the properties of the target population, the methods of sampling, and the methods of estimation. These three determinants form the spatial sampling trinity (Wang et al. 2012).

The target population refers to exhaustive observations of an entire study area. For example, a survey that enumerates birth defects in all villages in a country gives the target population of birth defects for the country. Sometimes, the process that generates the target population is of interest, and the observed target population is regarded as just one realization of the underlying process, called a superpopulation. For example, data about birth defects over many years contain almost all information relating to the risk of the disease in the study area, and can thus be regarded as a superpopulation. Sometimes a superpopulation is assumed and modeled rather than observed. Conventional sampling techniques usually ignore features in the target population or subpopulation (Cochran 1977). However, spatial autocorrelation (Tobler 1970; Haining 2003) and stratified heterogeneity (Goodchild and Haining 2004; Wang, Haining, and Cao 2010; Wang et al. 2016) are ubiquitous in geographically distributed target populations. Properly accounting for these properties can improve sampling efficiency (Griffith 2005).

The sample may be drawn randomly, systematically, or by stratifying the study area, as illustrated in Figure 1. These approaches may also be combined. For example, for a stratified systematic sample, the study area is first divided into strata according to the spatial heterogeneity of the target population, and the sample units are then drawn systematically from each of the strata. Another commonly used approach is
two-stage sampling: large units are selected in the first stage, and in the second stage a number of units are randomly selected in each large unit. Random sampling guarantees an unbiased estimation of the population, systematic sampling is more easily implemented in practice, and stratified sampling reduces the variance of estimation. A sample that is geographically random may not be random for a target population, such as prevalence of disease, because the latter is usually not randomly distributed geographically. The difference between the sampled population and the target population causes a biased estimate, which needs to be remedied.

The population or superpopulation can be estimated by both design- and model-based approaches (Wang et al. 2012). Design-based approaches handle data (Cochran 1977), that is, values at sampling sites are fixed while they may vary in different sites; a sample estimate varies with different samples. Model-based approaches handle random variables, that is, value at a sampling site is a variable, which is assumed to follow some probability distribution function (Matheron 1963). Table 1 presents a simple example to illustrate the relationship between the population, the superpopulation, design-based estimation, and model-based estimation.

Spatial sampling (S) and statistical inference (I) are usually connected, and in short written as spatial sampling, which consists of three steps:

\[ I - S - I \]

Step 1 (I₁) Homework before fieldwork. Decide a sample size \( n \) for a prior given a specific accuracy of estimation (or variance \( v \)), or estimate the accuracy (or variance \( v \)) of sample estimation for a prior given a specific sample size \( n \) (related to survey budget). See Table 1 for an example of the relationship between \( v \) and \( n \). The coefficients of dispersion variance \( \sigma^2 \), spatial autocorrelation \( r \) or \( C \), and strata in \( (v, n) \) may come from previous survey data, secondary data related to the target population, or pre-sampling.

Step 2 (S) Fieldwork. Distribute the sample over the target population in the field randomly, systematically, or in a stratified way (see Figure 1). If some secondary information related to and covering the target population is available, the sampling sites can be optimized by simulating different locations of the samples and choosing the one that minimizes the variance of an estimator (Hu and Wang 2011).

Step 3 (I₂) Homework again. Make a statistical inference of the population using the sample collected. The variance of an estimate may be reduced by taking into account
Table 1  Illustration of population, superpopulation, design-based, and model-based approaches.

<table>
<thead>
<tr>
<th></th>
<th>Population mean</th>
<th>Superpopulation mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>True value</td>
<td>( \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i )</td>
<td>( \mu = E\bar{Y} )</td>
</tr>
<tr>
<td>A design-based estimate:</td>
<td>( \bar{Y}<em>{\text{rand}} = \frac{1}{n} \sum</em>{i=1}^{n} Y_i )</td>
<td>( v_{\text{pop}}(\bar{Y}<em>{\text{rand}}) = v</em>{\text{superpop}}(\bar{Y}_{\text{rand}}) = \frac{\sigma^2}{n} )</td>
</tr>
<tr>
<td>simple average</td>
<td></td>
<td></td>
</tr>
<tr>
<td>variance for i.i.d. target</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( v_{\text{pop}}(\bar{Y}_{\text{rand}}) = \frac{\sigma^2}{n} (1 - r) )</td>
<td>( v_{\text{superpop}}(\bar{Y}_{\text{rand}}) = \frac{\sigma^2}{n} (1 + r) )</td>
<td></td>
</tr>
<tr>
<td>variance for autocorrelated target</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A model-based estimate:</td>
<td>( \bar{Y}<em>{BK} = \sum</em>{i=1}^{n} w_i Y_i ), to minimize ( v(\bar{Y}<em>{BK}) ) s.t. ( E\bar{Y}</em>{BK} = \bar{Y} )</td>
<td></td>
</tr>
<tr>
<td>block kriging</td>
<td></td>
<td></td>
</tr>
<tr>
<td>variance for i.i.d. target</td>
<td>( v_{\text{pop}}(\bar{Y}<em>{BK}) = v</em>{\text{superpop}}(\bar{Y}_{BK}) = \frac{\sigma^2}{n} ) (see Note)</td>
<td></td>
</tr>
<tr>
<td>variance for autocorrelated target</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( v_{\text{pop}}(\bar{Y}<em>{BK}) = v</em>{\text{superpop}}(\bar{Y}_{BK}) = C - \lambda )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>where ( C_{NN} = \frac{1}{N^2} \sum_{i,j} C_{ij}, C = \begin{bmatrix} C_{11} &amp; \cdots &amp; C_{1n} \ \vdots &amp; \ddots &amp; \vdots \ C_{n1} &amp; \cdots &amp; C_{nn} \end{bmatrix}, w = \begin{bmatrix} w_1 \ \vdots \ w_n \end{bmatrix}, \lambda = \begin{bmatrix} C_{1N} \ \vdots \ C_{nN} \end{bmatrix} )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notations: \( n \) and \( N \) denote the sizes of a sample and a population, respectively; subscripts \( \text{pop} \) and \( \text{superpop} \) stand for population and superpopulation, respectively; subscript \( \text{rand} \) denotes random; \( w \) denotes weight, which is estimated by the minimization of variance of an estimation, subject to some restrictions; \( \lambda \) is Lagrange constant to be estimated; \( E \) and \( v \) denote mathematical expectation and variance, respectively; \( r \) is Pearson correlation; \( \sigma^2 \) is dispersion variance; \( C \) denotes covariance; s.t. stands for subject to.

Note. When populations are i.i.d., spatial autocorrelation in the kriging family is null, so block kriging reduces to a simple random mean.

spatial autocorrelation and spatial stratified heterogeneity.

Sampling an independent and identically distributed population

A target population is independent and identically distributed (i.i.d.) if there are no major determinants, or if there are too many determinants with none dominating across the target population. In this case, a random sampling or a systematic sampling (see Figure 1), followed by a simple average of the sample, is an appropriate estimate of the population mean (Cochran 1977), and no better estimates can be made. Interpolation is not applicable if the population is i.i.d. and there are no explanatory variables, because there is no relationship between any two sites.

Spatial sampling an autocorrelated population

Spatial autocorrelation can be tested by Moran’s I statistic (Moran 1950), which allows a single...
Spatial sampling a stratified heterogeneous population

A stratum is an area where the population is relatively homogeneous and so can be represented by a small sample (Li et al. 2008). The spatial stratified nonhomogeneity can be diagnosed by a geographical detector (Wang et al. 2010, 2016). If a population is spatially stratified heterogeneity, a stratified sampling should be used (Cochran 1977), while mapping can be done by sandwich interpolation (Wang et al. 2013a), which “borrows samples from the same class” and maps both the area values and their variances.

Spatial sampling a target with both autocorrelation and stratified nonhomogeneity

1 If a sample is large enough that each stratum has at least two sample units, then the MSN (mean of surface with nonhomogeneity) can provide a BLUE estimate of the population mean (Wang, Christakos, and Hu 2009) by a summation of weighted sample values. The weights are calibrated to minimize the variance of the MSN estimation, subject to the sample mean being unbiased to the population mean. For mapping, sandwich-kriging is used: first, by removing the stratified means from the sample data to obtain residuals; second, by applying the sandwich mapping to the stratified data, and apply the kriging mapping to the residual data, respectively; and finally, by combining the sandwich map and the kriging map and adding the variances of the two maps together. In this way, the map and the variance of a target are obtained with both autocorrelation and stratification.

2 If the sample size is small and biased, B-shade (biased sentinel hospital area disease estimation) (Wang et al. 2011) can remedy the bias to achieve a BLUE estimate of the population using secondary information. The method does not assume second-order stationarity as with kriging; instead the ratio between sample mean and population mean of the secondary variable is used to remedy the sample bias.

3 If only a single-point sample is available over the stratified population, its observations are obviously biased to that of the stratified population. SPA (single-point area estimation) (Wang et al. 2013b) can remedy the bias of the single-point sample for the entire area giving a BLUE result, if there is secondary information related to the primary variable and covering the area. In SPA, the value at a single point is weighted as a function of the correlation of the values of the primary variable between the single point and the population, with the correlation of the values between the stations virtually distributed over the study area as the second variable. The parameters related to the primary population are estimated by those of the secondary attribute. The estimation error caused by the approximation is largely absorbed by regression to empirical data.
Spatial sampling of a spatiotemporal, multivariate, or spatiotemporal multivariate population

A copula-based model offers a solution if a joint probability density function is available (Marchant et al. 2013). Another solution is to weight the multiple populations into a single population, although this approach suffers from subjective weighting. A third solution is to sample randomly or systematically in space, because any sampling that is optimized to one time or to one variable may conflict with another time or variable. In the inference stage, various inference methods can be implemented flexibly at each time and to each variable when sample data are available.

SEE ALSO: Scale; Spatial analysis; Spatial tessellations; Spatial weights; Uncertain geographic context problem

References


Spatial social networks

Ming-Hsiang Tsou
Jiue-An Yang
San Diego State University, USA

Spatial social networks, an emerging topic across geographical information science (GIScience), computational social science, and social network analysis, visualize social networks by using real-world or abstract map coordinates. From a geographer’s perspective, spatial social networks represent the linkages between people’s communications (such as e-mails, telephones, social network services) and real-world events (elections, protests, parties, music concerts). This entry gives a brief introduction to social networks and spatial social networks, and highlights the key characteristics of spatial social networks with a real-world scenario (the 2012 US presidential election) in the Twitter platform.

Social networks

We live in a connected world. People interact with each other via various types of social networks. A social network is a communication structure of social connections, personal relationships, or group interactions within a local or global community. It is an important interdisciplinary research domain in social science, communication theory, computer science, and mathematics (Wasserman 1994). The study of social networks can be traced back to the famous epidemiological study by Dr John Snow (1855), who used the locations of cholera death records to examine the spatial pattern of disease and its connection to contaminated water sources in London during the 1800s. The spatialization of social networks can provide valuable information and knowledge discovery for understanding human activities, disease outbreaks, and public opinions.

Traditional social network analysis is a popular method in epidemiology, sociology, psychology, and social work that is implemented using surveys and personal interviews. The recent development of the Internet and wireless communications (such as short message services (SMS) and social network services (SNS) in smartphones) links millions of networks and billions of people together. These services and communication tools provide basic infrastructure for computer–supported social networks (CSSNs) (Wellman et al. 1996). Scientists now can trace, monitor, and analyze the dynamic change of social networks and social communications related to protests, political campaigns, and public opinions via social media and weblogs (Lazer et al. 2009). These research efforts can help us understand the diffusion of innovations (Hägerstrand 1967) and facilitate the discovery of knowledge in cyberspace and social media (Tsou and Leitner 2013).

Social media services (such as Twitter and Facebook) are one type of powerful communication platform and CSSN for idea exchange, personal networking (Lerman and Ghosh 2010), political opinions (An et al. 2011), and collective actions (Earl 2010). By using social media services, people can communicate and coordinate their activities geospatially and, to a significant degree, accomplish these social communication functions in near–real time. The impacts of
SPATIAL SOCIAL NETWORKS

these tools were vividly demonstrated in recent antigovernment protests, including the Arab Spring, Occupy Wall Street, and the Umbrella Revolution in Hong Kong.

Spatial social networks

Spatial social networks transform social networks into visual maps and information landscapes using real-world coordinate systems (latitudes and longitudes) or abstract coordinate systems. After such transformations, geographers can conduct space–time analysis, understand the diffusion of information, and identify spatial cluster patterns of network hubs and elements. One early pioneer in studying spatial social networks was Dr Torsten Hägerstrand. In his famous book (his doctoral dissertation), Innovation Diffusion as a Spatial Process (1967), Hägerstrand used computational methods (Monte Carlo simulation) to simulate the diffusion of technologies across social networks with an empirical analysis of cattle tuberculosis controls (inoculations) by using farm soil mapping, population density, and other map layers in Sweden. Hägerstrand created several simulation models to illustrate which geographic factors (such as population density, locations, and road networks) are influential during the innovation diffusion process by using probability models and statistical methods.

The diffusion of innovations became an important theory and concept in many research disciplines, including business marketing, health promotion, geography, and communication. Along with the popularity of weblogs and social media services, researchers have begun to study the processes of meme diffusion over social networks. A meme is a cultural idea or concept that can be transferred, replicated, and modified (mutated) in a way that suggests genetic processes as they operate in biological evolution. There have been extensive studies on how information (memes) diffuses and spreads in a social communication network, such as the follower network in Twitter and the friendship network in Facebook. In early studies, researchers used and refined the disease-propagation susceptible, infected, and recovered (SIR) model in epidemiology (Bailey 1975) to simulate and understand the information spreading process. Many recent studies considered the spread of a meme through a social network as an innovation diffusion process, where two fundamental models, the linear threshold (LT) model (Granovetter 1987) and the independent cascade (IC) model (Goldenberg, Libai, and Muller 2001), are frequently used.

Using Twitter to analyze spatial social networks: a case study involving the 2012 US presidential election

Established in 2006, Twitter has become one of the most popular online social networking services. Twitter users can post and broadcast short messages (restricted to 140 characters) along with pictures, URLs, or video clips to their “followers.” These short messages are called “tweets,” which are searchable by keywords, authors, and hashtags. Many social media platforms (Twitter, Instagram, Facebook, YouTube, and Flickr) provide application programming interfaces (APIs) to enable automatic data fetching/downloading processes by using Python or other programming languages. Twitter is the most popular platform for studying social network analysis due to the large numbers of users and tweets, the openness of its APIs, and the rich metadata associated with tweets and user profiles.

Developers can define different types of criteria (keywords or predefined regions) via the APIs and conduct spatial searches in social media. The
two major APIs for Twitter data retrieval are the Search API and the Streaming API. Programmers can use the Twitter Search API to conduct a spatial search for specific keywords within a circle by defining the center of the circle (the latitude and longitude) and its radius (typically miles or kilometers). The Search APIs will retrieve those tweets that are posted from the search area from 6–9 days before. This retrospective search function is very useful for monitoring unexpected events, such as earthquakes. On the other hand, the Twitter Streaming APIs can provide a live stream of new tweets (no historical tweets) by defining keywords or a region (defined by a bounding box). When the bounding box is used with the Streaming APIs, search results will only include geotagged tweets. There is a significant limitation in the (free) Twitter Streaming APIs because users can only download up to 1% of total tweets within the selected regions.

There are three types of location information directly associated with social media messages: geotagged locations provided by GPS-enabled devices, self-reported locations specified in user profiles, and place names mentioned in tweets. Geotagged locations are latitude and longitude pairs created by mobile devices with built-in GPS receivers, place names (gazetteers), or by Internet Protocol (IP) geolocation features. The self-reported location is specified by social media users and it can be changed by users at any time. The self-reported location can be a city name, a state name, or any text string that may be unrelated to a typical place name. Based on some previous research findings, between 0.7 and 6% of tweets are geotagged depending on different topics (Tsou et al. 2013).

Downloaded social media data are usually in unstructured data format and stored in No-SQL databases, such as MongoDB. No-SQL databases have better performance and flexibility for storing social media messages compared to traditional relational databases. Many programming tools, such as Python, R, and JavaScript, have useful libraries to help us analyze the contents of social media or to create search tools. Figure 1 is an example of Twitter GeoSearch Tools developed by the Center for Human Dynamics in the Mobile Age (HDMA) at San Diego State University. Each social media message is collected with detailed attributes including a user ID, creation time, text content, media content, and spatial locations, which includes geotagged coordinates or self-reported locations. Since each city (circle) may have a different population, 2010 census tracts were used to calculate the estimated population for each circle (Figure 1, right). The total number of tweets

![Image](image.png)

**Figure 1** The Twitter GeoSearch Tool using Search APIs to collect tweets from multiple US cities.
Figure 2 Social network analysis example: comparing the Twitter followers of Mitt Romney and Barack Obama in Ohio (combining the cities of Cleveland and Columbus) during the 2012 US presidential election.

collected from each city is normalized by this estimated population to reflect the tweeting rates in each area.

Figure 2 illustrates a social network analysis conducted during the 2012 US presidential election using the Twitter GeoSearch and network graph tools. Tweets were searched for using candidates’ full names (“Barack OR Obama” and “Mitt OR Romney”) to capture results that mention the two candidates in full name or first/last name only. Their followers in the cities of Cleveland and Columbus were then analyzed. The right side of the figure illustrates the social network among these Twitter accounts and their following structure. Romney supporters and their following paths were symbolized using red lines and the Obama supporters with blue lines.

In the graphs, it was found that many Romney supporters in Ohio were also following Obama, but only one Obama supporter in Ohio was following Romney (Figure 2, left).

In addition to social networks composed of social media users and followers, the interactions between users can also form other types of networks. The first is formed by the use of mentions in tweets. Mentions (@) are mostly used in Twitter when an author wishes to direct a tweet message to specific users. Retweet activity also forms another type of social network. A retweet is an echo to a tweet posted by another user. The action of user X retweeting a tweet authored by user Y can be conceptualized as a signal being diffused from user Y (the tweeter) and received by user X (the retweeter). However, whether
the retweeter adopts the idea expressed by the tweeter is unclear and requires further study. A list of tweeter–retweeter pairs and a list of author–direct tweet receiver pairs were created to visualize a social network from tweet messages. Figure 3 shows a mention network for New Hampshire and Wisconsin. The network visualization shows that Twitter users in New Hampshire appear to have fewer common mentions that form a sparse mention graph with many small clusters. The Wisconsin mention network has several large clusters and this may suggest that some users in the state received high attention during conversations about the two candidates.

Figure 4 demonstrates the retweet network of New Hampshire, Wisconsin, and Ohio. The smaller number of total tweets from New Hampshire can be seen in the graph when compared to the graph of Wisconsin or Ohio. Wisconsin has relatively fewer clusters than New Hampshire, which can indicate that there are fewer opinion leaders in the network. Based on the retweet network, most Twitter users in New Hampshire tend to have preferred information sources that are highly connected with each other. Ohio has the most connected retweet network, which can result in faster information diffusion in the network when compared to the other two states.

Figure 5 shows another interesting finding from the 2012 US presidential election (Tsou et al. 2013). The tweets collected in the 30 largest US cities are represented using pie chart maps to compare changes in tweet attention levels.
Figure 5  The comparison of tweet attention levels between “Romney” and “Obama” (a) before (October 24, 2012) and (b) after (November 1, 2012) Hurricane Sandy. Red: “Romney” related tweets; blue: “Obama” related tweets. Source: Tsou et al. 2013, 345. Reprinted with permission of Taylor & Francis Ltd.
before and after Hurricane Sandy. Hurricane Sandy caused severe damage to many cities along the US East Coast in October 2012, a few days before the 2012 election. This disaster event created a significant change of tweet attention levels between the two candidates. The size of the circle indicates the total numbers of tweets standardized by city population, with larger circles representing more tweets in the city. Comparing Figure 5a and 5b, the circles of East Coast cities (Washington DC, New York, and Boston) increased significantly after Hurricane Sandy. The attention levels between “Obama” (blue color) and “Romney” (red color) also changed. In New York City, the tweet percentage of Romney decreased from 56% (October 24, 2012) to 34% (November 1, 2012). Overall, 9 of the 30 cities changed from a majority Romney tweeting percentage to a majority Obama tweeting percentage, and many others increased the Obama percentage compared to the Romney percentage (Tsou et al. 2013).

Future developments and research challenges of spatial social network analysis

By tracking and analyzing the contents of social media across different social network services, researchers can reveal important social processes related to specific events (such as presidential elections or protests) and gain insight into the temporal and spatial relationships among messages and human behaviors. The digitization of social networks will provide valuable big data resources for studying human dynamics and is facilitating the emergence of a data-driven computational social science (Lazer et al. 2009). In the domain of GIScience, knowledge obtained from spatial social network analytics can not only help geographers verify existing social behavioral theories but can also contribute to problem-solving across a range of areas. For example, searching social media at municipal or regional levels can help detect disease outbreaks (flu or whooping cough), track natural disaster impacts, or predict political election outcomes. Geographers now can build a new research framework using spatial social media and networks to study human geography topics, such as population, transportation, urbanization, and social theories.

Many research challenges remain in spatial social network analysis. First, social media data samples do not represent everyone in the population. Social media is generally associated with younger generations; older adults have much lower usage. The 2013 Pew Research Center Survey (www.pewinternet.org/) suggests that over 90% of young Internet users (age 18–29) use social networking sites. For example, older teens and young adults are the heaviest Twitter users: approximately 74% of Twitter users are in the age range of 15–25 years (www.beevolve.com). Second, the percentage of geotagged social media data is small when compared to non-geotagged data: only 1–4% of tweets have GPS coordinates. Although 80% of tweets have city-level location info (user profiles), current geographic information system (GIS) software cannot effectively use this type of spatial information. Many geographers only use geotagged data in their spatiotemporal analysis. Third, protecting user privacy and locational privacy in social media is very important, but comprehensive privacy protection measures are difficult to implement. The privacy concerns and spatial information disclosure risks in social media are significant and relevant to users and their social connections. The vulnerability of spatial information disclosure increases significantly when social media users have large social networks (Tsou and Leitner 2013). Finally, different social media platforms require different types of data,...
SPATIAL SOCIAL NETWORKS

analysis methods. For example, Twitter is more open and allows users to access most of the contents in Twitter. Facebook focuses more on personal social networks that provide stricter ways to access the content only when users are already in the network or make their network public. In general, Twitter is more suitable for real-time monitoring (e.g., wildfires) while Facebook can be a great source for studying personal social networks.

SEE ALSO: Big data; Geovisualization of social media; Graph theory; Networks, social capital, and development

References


Further reading


Spatial tessellations

Atsuyuki Okabe
Aoyama Gakuin University, Japan
University of Tokyo, Japan

Given a space (e.g., an area on a plane), a spatial tessellation is defined as a set of subspaces (subareas) that are collectively exhaustive for the space (area) and mutually exclusive except for boundaries. In the real world, many kinds of spatial tessellations are observed, such as administrative districts of a country, postal zones covering a country, public school districts in a region, and so forth. To study geometrical characteristics of these real spatial tessellations, various conceptual spatial tessellations have been proposed in the related literature; in particular, geography, geology, ecology, meteorology, and geographic information science. Among these, two of the most frequently used spatial tessellations are Voronoi diagrams and Delaunay triangulations. In this entry, these two types of spatial tessellation are described in depth, together with some of the other types of spatial tessellation.

Ordinary Voronoi diagrams

Voronoi diagrams are named after a Ukrainian mathematician, Georgy Fedoseevich Voronoy (Georges Voronoï) (1868–1908), who specialized in number theory involving quadratic forms and tessellations. The same concept was developed by several other scholars almost at the same time; for example, a German mathematician, Peter Gustav Lejeune Dirichlet (1805–1859), who contributed to number theory as well as Fourier series, and an American meteorologist, Alfred H. Thiessen (1872–1956), who used a Voronoi diagram to compute accurate estimates of regional rainfall averages (Thiessen 1911). Voronoi diagrams are sometimes called Delichlet domains (Delaunay 1929a, 1929b) or Thiessen polygons (Whitney 1929). The history of the concept of Voronoi diagrams and its developers is described in detail in Okabe et al. (2000, Section 1.2).

To provide an intuitive definition of a Voronoi diagram, consider a set of cell phone antennas in a region, and assume that a cell phone at a given location communicates through the nearest antenna. Then each antenna has its own communicative service area, and the set of the resulting service areas forms a spatial tessellation, called a Voronoi diagram. To depict this, suppose that a set of points is placed on the Euclidean plane (the black circles in Figure 1), and assume that the number of the points in the set is more than one but finite and that no points in the point set coincide. Given this set, every location in the plane is assigned to the closest member (i.e., point) in the point set with respect to the Euclidean distances. If a location happens to be equally close to two or more members (points) in the point set (points on the line segments in Figure 1), the location is assigned to those members (points). As a result, the set of locations assigned to every member (point) in the point set forms its own region, referred to as a Voronoi region (a polygon in Figure 1). Because every location is assigned to at least one member (point) in the point set, the resulting Voronoi regions are collectively exhaustive in the plane. Furthermore, because two adjacent Voronoi regions overlap only on
their boundaries, the Voronoi regions in a set are mutually exclusive except for boundaries. The set of Voronoi regions associated with points in the point set is termed the Voronoi diagram generated by the point set (Voronoi, 1908).

To formalize this verbal definition mathematically, let \( P = \{ p_1, \ldots, p_n \} \) be a set of points on the 2-D Euclidean space \( \mathbb{R}^2 \), where \( p_i \neq p_j \) for \( i \neq j \). A point \( p_i \) is referred to as a generator (point), and \( P = \{ p_1, \ldots, p_n \} \) is referred to as a generator (point) set. Let \( d(p, p_i) \) be the Euclidean distance between an arbitrary point \( p \) and a generator point \( p_i \) on \( \mathbb{R}^2 \), and let \( V(p_i) \) be a set of points defined by equation 1

\[
V(p_i) = \{ p | d(p, p_i) \leq d(p, p_j), \quad j \neq i, \quad j = 1, \ldots, n \} \tag{1}
\]

and \( V(P) = \{ V(p_1), \ldots, V(p_n) \} \). The set \( V(p_i) \) is termed the Voronoi polygon associated with (or of) \( p_i \), and \( V(P) \) is termed the Voronoi diagram, or more specifically, the ordinary planar Voronoi diagram, generated by \( P \). This diagram is easily extended to an ordinary \( m \)-dimensional Voronoi diagram by replacing \( \mathbb{R}^2 \) with the \( m \)-dimensional Euclidean space \( \mathbb{R}^m \).

The ordinary Voronoi diagram is generalized with respect to a distance, a generator set, and a space (note that the ordinary Voronoi diagram is defined with Euclidean distance, a set of points, and Euclidean space). This implies that the possible number of generalizations is [the number of different kinds of distances] \( \times \) [that of generator sets] \( \times \) [that of spaces], which amounts to a great number (a variety of generalizations is introduced in Okabe et al. (2000)). To avoid too many generalizations, it is first supposed that a generator set and a space are fixed as a set of points and Euclidean space, respectively.

**Weighted Voronoi diagrams**

Consider a distance, \( d_{MW}(p, p_i) = \alpha_i d(p, p_i) \), where \( d(p, p_i) \) is the Euclidean distance between an arbitrary point \( p \) and a generator point \( p_i \) (\( \alpha_i > 0 \)), which is referred to as a multiplicatively weighed distance. Let \( V_{MW}(p_i) \) be a set of points defined by equation 2.

\[
V_{MW}(p_i) = \{ p | \alpha_i d(p, p_i) \leq \alpha_j d(p, p_j), \quad j \neq i, \quad j = 1, \ldots, n \} \tag{2}
\]

The resulting set \( V_{MW}(P) = \{ V_{MW}(p_1), \ldots, V_{MW}(p_n) \} \) is termed a multiplicatively weighted Voronoi diagram (Aurenhammer and Edelsbrunner, 1984). An example is shown in Figure 2.

Next, consider a distance, \( d_{AW}(p, p_i) = d(p, p_i) + \beta_i \), which is referred to as an additively weighed distance. Let \( V_{AW}(P) = \{ V_{AW}(p_1), \ldots, V_{AW}(p_n) \} \), where \( V_{AW}(p_i) \) is given by equation 3.

\[
V_{AW}(p_i) = \{ p | d(p, p_i) + \beta_i \leq d(p, p_j) + \beta_j, \quad j \neq i, \quad j = 1, \ldots, n \} \tag{3}
\]
The resulting diagram $V_{AW}(P)$ is termed an additively weighted Voronoi diagram (Aurenhammer 1988). Figure 3 depicts an example. An extension along this line entails a mixed distance, $d_{CW}(p, p_i) = \alpha_i d(p, p_i) + \beta_i$, called a compoundly weighted distance. The resulting Voronoi diagram is referred to as a compoundly weighted Voronoi diagram. A diagram closely related to an additively weighted Voronoi diagram is an (additively weighted) power diagram $V_{PW}(P)$ = \{ $V_{PW}(p_1)$, ..., $V_{PW}(p_n)$ \}, where $V_{PW}(p_i)$ is given by equation 4.

$$V_{AW}(p_i) = \left\{ p | d(p, p_i)^2 + \beta_i \leq d(p, p_j)^2 + \beta_j, \quad j \neq i, \quad j = 1, \ldots, n \right\} \quad (4)$$

$k$-th nearest-point Voronoi diagrams

In the ordinary Voronoi diagram, points are assigned to their “first nearest” generator points. In turn, the first nearest may be extended to the “second nearest”, “third nearest”, ..., and “$k$-th nearest”. Let $V_K(p_i)$ be the set of points $p$ whose $k$-th nearest generator point among $n$ generator points $P = \{ p_1, \ldots, p_n \}$ is $p_i$. The resulting set $V_K(P) = \{ V_K(p_1), \ldots, V_K(p_n) \}$ is termed the $k$-th nearest-point Voronoi diagram generated by $P$ (Dehne 1982). A second nearest-point Voronoi diagram is depicted in Figure 4. By definition, the first nearest-point Voronoi diagram is the ordinary Voronoi diagram.

A specific case of the $k$-th nearest-point Voronoi diagram is $k = n$, which is alternatively called the farthest-point Voronoi diagram generated by $P$, $V_F(P) = \{ V_F(p_1), \ldots, V_F(p_n) \}$, where $V_F(p_i)$ is given by equation 5.

$$V_F(p_i) = \left\{ p | d(p, p_i) \geq d(p, p_j), \quad j \neq i, \quad j = 1, \ldots, n \right\} \quad (5)$$

An example is shown in Figure 5.
Figure 4  A second nearest-point Voronoi diagram (note that the configuration of the generator points is the same as that of the ordinary Voronoi diagram in Figure 1; the color of each generator point corresponds to that of its second nearest-point Voronoi polygon) (provided by W. Morioka).

Voronoi diagrams for lines

In the above Voronoi diagrams, a generator set is a set of points. A generator set can be extended to a set of line segments, \( L = \{L_1, \ldots, L_n\} \), which may be straight line segments, curved line segments (including circles), or chains of connected line segments. To define Voronoi diagrams for such lines, let \( d_L(p, L_i) \) be the minimum distance between a point \( p \) on a plane \( \mathbb{R}^2 \) and a point on \( L_i \), and let \( V_L(P) = \{V_L(L_1), \ldots, V_L(L_n)\} \), where \( V_L(L_i) \) is given by equation 6.

\[
V_L(p_i) = \{p \mid d(p, L_i) \leq d(p, L_j), \quad j \neq i, \quad j = 1, \ldots, n\} \tag{6}
\]

The resulting diagram is called the line Voronoi diagram generated by \( L \) (Drysdale and Lee 1978). An example is depicted in Figure 6.

Figure 5  A farthest-point Voronoi diagram (note that the configuration of the generator points is the same as that of the ordinary Voronoi diagram in Figure 1; the color of each generator point corresponds to that of its farthest-point Voronoi polygon) (provided by W. Morioka).

Voronoi diagrams for areas

Noting that a closed chain forms an area, a generator set of lines may be extended to a set of areas. To be explicit, let \( A \) be a set of areas, \( A = \{A_1, \ldots, A_n\} \); let \( d_A(p, A_i) \) be the minimum distance between a point \( p \) on a plane \( \mathbb{R}^2 \) and a point on \( A_i \); and let \( V_A(A) = \{V_A(A_1), \ldots, V_A(A_n)\} \), where \( V_A(A_i) \) is given by equation 7.

\[
V_A(A_i) = \{p \mid d(p, A_i) \leq d(p, A_j), \quad j \neq i, \quad j = 1, \ldots, n\} \tag{7}
\]

The resulting diagram is called the area Voronoi diagram generated by \( A \) (Roos, 1989). Figure 7 illustrates an example of an area Voronoi diagram. Rappaport (1992) discusses in depth a special case in which an area is given by a disk.
So far, distances for defining Voronoi diagrams have been Euclidean distances or those defined in terms of Euclidean distances. Voronoi diagrams may also be defined with distances other than Euclidean distances. An extension of the Euclidean distance is given by:

\[
d_{L^p}(p, p_i) = \left[ \sum_{j=1}^{m} |x_j - x_{ij}|^p \right]^{\frac{1}{p}}
\]

where \((x_1, \ldots, x_m)\) and \((x_{i1}, \ldots, x_{im})\) are the Cartesian coordinates of points \(p\) and \(p_i\), respectively. The distance \(d_{L^2}(p, p_i)\) is termed the Minkowski distance. In a specific case where \(p = 2\), the distance is the Euclidean distance. When \(p = 1\) and \(m = 2\), the distance is written as \(d_{L^1}(p, p_i) = |x_1 - x_{i1}| + |x_2 - x_{i2}|\), called the Manhattan distance. The diagram \(V_{L^1}(P) = \{ V_{L^1}(p_1), \ldots, V_{L^1}(p_n) \}\), where \(V_{L^1}(p_i)\) is given by equation 8,

\[
V_{L^1}(p_i) = \{ p \left| |x_1 - x_{i1}| + |x_2 - x_{i2}| \leq |x_1 - x_{j1}| + |x_2 - x_{j2}|, \ j \neq i, \ j = 1, \ldots, n \} \quad (8)
\]

is termed the Manhattan Voronoi diagram generated by \(P\) (Carter, Chaiken, and Ignall 1972). An example is shown in Figure 8.

Besides the Minkowski distance, Voronoi diagrams can be defined in terms of many other distances. Examples are Voronoi diagrams defined with convex distance, elliptic distance, Karlsruhe distance, Hausdorff distance, boat-in-a-river distance, and so forth. For details, see Okabe et al. (2000, Section 3.7).

### Voronoi diagrams on spaces other than Euclidean space

The Voronoi diagrams above all assume Euclidean space, but they may be defined in other spaces. First, for instance, consider a network space, \(N\), such as a street network, on which points \(p_1, \ldots, p_n\) are placed, and let \(d_S(p, p_i)\) be the shortest-path distance between an arbitrary point \(p\) and a generator point \(p_i\) on \(N\).
Then the set \( V_N(P) = \{ V_N(p_1), \ldots, V_N(p_n) \} \), where \( V_N(p_i) \) is given by equation 9,

\[
V_N(p_i) = \{ p | d_S(p, p_j) \leq d_S(p, p_i), \quad j \neq i, \quad j = 1, \ldots, n \} \quad (9)
\]

forms a tessellation of \( N \), which is termed the (ordinary) network Voronoi diagram generated by \( P = \{ p_1, \ldots, p_n \} \) (Okabe and Kitamura 1996). An example is shown in Figure 9. This diagram can be extended to the weighted Voronoi diagram, the \( k \)-th nearest-point Voronoi diagram and the line Voronoi diagram defined on a network as the extensions of those defined earlier on a plane (for details, see Okabe and Sugihara (2012, Chapter 4)).

Other spaces that are fairly tractable for constructing Voronoi diagrams are cylinder surfaces, cone surfaces, and polyhedral surfaces, because those surfaces in \( \mathbb{R}^3 \) are developed into rectangles (e.g., the left-hand side figure in Figure 10), fan-shaped regions, and polygonal regions in \( \mathbb{R}^2 \), respectively (for details, see Okabe et al. (2000, Sections 3.7.7–3.7.9)). An example of a cylindrical Voronoi diagram is shown in Figure 10.

Spherical surfaces are often used for analyzing global phenomena, but the above development approach is not applicable because a sphere surface in \( \mathbb{R}^3 \) cannot be developed into \( \mathbb{R}^2 \). To define a Voronoi diagram on a sphere surface \( S \), consider a sphere with the unit radius centered at the origin and where generator points \( p_1, \ldots, p_n \) are placed on \( S \). In this case, the distance between points \( p = (x_1, x_2) \) and \( p_i = (x_{i1}, x_{i2}) \) on \( S \) is measured by the great circle distance \( d_{GC}(p, p_i) \); that is, the length of the lesser arc on the great circle (i.e., the circle centered at the center of \( S \)). Mathematically, the set \( V_{GC}(P) = \{ V_{GC}(p_1), \ldots, V_{GC}(p_n) \} \), where \( V_{GC}(p_i) \) is given by equation 10,

\[
V_{GC}(p_i) = \{ p | \arccos(x_{i1}x_{j1} + x_{i2}x_{j2}) \leq \arccos(x_{i1}x_{j1} + x_{i2}x_{j2}) \leq \pi, \quad j \neq i, \quad j = 1, \ldots, n \} \quad (10)
\]
forms a tessellation of $S$. The resulting tessellation is referred to as the spherical Voronoi diagram generated by $P$. An example is shown in Figure 11.

**Delaunay triangulations**

A tessellation closely related to the Voronoi diagram is the so-called Delaunay triangulation.

The triangulation is named after Boris Delaunay (1890–1980), a Russian mathematician. For a given ordinary Voronoi diagram where the number of generator points is three or more that are not on the same line, join all pairs of generator points whose Voronoi polygons share the common Voronoi edge by line segments. The resulting line segments form a tessellation, called a Delaunay triangulation. The Delaunay triangulation obtained from the Voronoi diagram in Figure 1 is illustrated in Figure 12. To state this definition mathematically, let $V(P) = \{ V(p_1), \ldots, V(p_n) \}$ be an ordinary Voronoi diagram generated by $P = \{ p_1, \ldots, p_n \}$, $3 \leq n < \infty$ such that the points in $P$ are not on the same line; let $q_1, \ldots, q_{n_v}$ be the set of Voronoi vertices of the Voronoi polygons in $V(P)$; and let $(x_{ij1}, x_{ij2})$, $j = 1, \ldots, k_i$ be the coordinates of the generator points whose Voronoi polygons share a vertex $q_i$, ($k_i$ is the number of such generator points). Let $D(q_i)$ be the set of points $(x_1, x_2)$ satisfying equation 11.
Spatial Tessellations

\[ D(q_i) = \left\{ (x_1, x_2) \mid (x_1, x_2) = \sum_{j=1}^{k_i} \lambda_j (x_{ij1}, x_{ij2}), \right\} \]

where \( \sum_{j=1}^{k_i} \lambda_j = 1, \lambda_j \geq 0, j = 1, \ldots, k_i \) (11)

and \( D(P) = \{ D(q_1), \ldots, D(q_{n_P}) \} \). The resulting set of triangles is termed the Delaunay triangulation spanning \( P \), which is a tessellation of the convex hull of \( P \) (i.e., the minimum polygon including all points in \( P \)) (Delaunay 1934). The Delaunay triangulation in \( \mathbb{R}^2 \) can be extended to that in a higher dimension \( \mathbb{R}^m \), \( m \geq 3 \), which is referred to as a Delaunay tetrahedrization in \( \mathbb{R}^3 \) and a Delaunay tessellation in \( \mathbb{R}^m \), \( m \geq 4 \).

Random line mosaics

In ecology, the Voronoi diagram generated by random generator points (the Poisson Voronoi diagram) is sometimes referred to as the random-sets mosaic (or just \( S \)-mosaic) (Pielou 1977). Contrasted to this mosaic is the so-called random-line mosaic. To be explicit, a line \( l \) on a plane \( \mathbb{R}^2 \) is written as \( x_1 \cos \theta + x_2 \sin \theta = h \), where \( (x_1, x_2) \) are the Cartesian coordinates, \( h \) is the length from the origin to the foot of the perpendicular line to \( l \), and \( \theta \) is the angle between the \( x_1 \) axis and the perpendicular line in counterclockwise. A random line in the circle with radius \( R \) centered at the origin is obtained from generating a random point \( (\theta, h) \) in the region \( 0 \leq \theta \leq 2\pi \) and \( 0 \leq h \leq R \). A set of independent random lines forms a tessellation, termed the random-line mosaic (or just \( L \)-mosaic). An example of a random-line mosaic is shown in Figure 13. The statistical properties of the \( L \)-mosaic and those of the \( S \)-mosaic are discussed in Pielou (1977, Chapter 12) and in Okabe et al. (2000, Chapter 5), respectively.

Tilings

A tiling, \( T = \{ T_1, \ldots, T_{n_T} \} \), is a tessellation where the number \( n_T \) is usually countable infinite. The origin of tilings is very old and dates back to at least Roman times. In statistical contexts, random tessellations, such as Poisson Voronoi diagrams and \( L \) mosaics, are basic tessellations. In contrast, a major concern of tilings is “regular” tessellations. When all the tiles \( T_i \) in \( T \) are the same size and shape, the tiling \( T \) is termed a monohedral tiling. In a specific case where \( T_i \) is a regular polygon, the monohedral tiling \( T \) is referred to as a regular tiling. There exist only three kinds of regular tilings: regular triangle tilings, square tilings, and regular hexagon tilings. These kinds of tilings are obtained from Voronoi diagrams where the generator points
are placed on regular hexagon lattices, square lattices and regular triangle lattices, respectively. When the tiles in $T$ consist of two types of tile, $T$ is called a dihedral tiling (Figure 14). In a similar manner, trihedral, quadhedral, ..., $n$-hedral tilings can be defined. In addition to these tilings, many other kinds of tilings have been proposed. For details, see Grünbaum and Shephard (1987).

### Computational methods for constructing tessellations

Because constructing tessellations usually requires heavy computation, efficient computational methods are important. Two major methods for efficiently constructing the ordinary planar Voronoi diagram are the incremental method and the divide-and-conquer method. In the former, given $n$ generator points, Step 1 is to construct the Voronoi diagram for two generator points. Step 2 is to construct the Voronoi diagram for three generator points using the Voronoi diagram resulting from Step 1. Step 3 is ... Step $n-1$ is to construct the Voronoi diagram for $n$ generator points using the Voronoi diagram for $n-1$ generator points resulting from Step $n-1$. Figure 15 illustrates that when using the Voronoi diagram for six generator points, the Voronoi diagram for seven generator points is constructed as in Figure 1.

In the divide-and-conquer method, $n$ generator points are indexed as $P = (p_1, \ldots, p_n)$ according to the increasing order of the $x$ coordinates.

#### Algorithm: divide-and-conquer (D&C) method (Okabe et al. 2000)

Step 1. If $n \leq 3$, construct the Voronoi diagram directly and go to Step 3.

Step 2. Otherwise, do the following.

2.1 Let $t$ be the integer part of $n/2$, and divide $P = (p_1, \ldots, p_n)$ into $P_L = (p_1, \ldots, p_t)$ and $P_R = (p_{t+1}, \ldots, p_n)$. 

Figure 14: A dihedral tiling (provided by W. Morioka).

Figure 15: The incremental method for constructing the Voronoi diagram for seven generator points (Figure 1) using the Voronoi diagram for six generator points (provided by W. Morioka).
2.2 Construct the Voronoi diagram $V_L$ for $P_L = (p_1, \ldots, p_t)$ by Algorithm D&C.

2.3 Construct the Voronoi diagram $V_R$ for $P_R = (p_{t+1}, \ldots, p_n)$ by Algorithm D&C.

2.4 Merge $V_L$ and $V_R$ into the Voronoi diagram $V$ for $P = (p_1, \ldots, p_n)$.


Figure 16 illustrates the final step where $V_L$ and $V_R$ are to be merged into $V$ shown in Figure 1. Concerning the implementation of the above two computational methods and that of other types of Voronoi diagram, see Okabe et al. (2000, Chapter 4), Okabe and Sugihara (2012) (Voronoi diagrams along networks), and Aurenhammer, Klein, and Lee (2013). Computation of the Delaunay triangulation may be achieved through the corresponding Voronoi diagram by joining generator points whose Voronoi polygons share the common Voronoi edge or through a “swapping procedure” (Okabe et al. 2000, Section 2.4; Aurenhammer, Klein, and Lee 2013). The $L$-mosaic consisting of $n$ lines on a disk centered at the origin with radius $R$ is computed by generating $n$ points $(\theta_i, h_i)$ independently according to the uniform distribution on the rectangle $0 \leq \theta_i \leq 2\pi$ and $0 \leq h_i \leq R$, $i = 1, \ldots, n$. Preparata and Shamos (1985) show basic computational techniques used in the above computations.

SEE ALSO: Digital elevation model and digital surface model; Distance; Geographic information system; Graph theory; Spatial analysis; Territory and territoriality; Thiessen polygons; Topological relations

References


Further reading


Spatial thinking is fundamental to learning geography at all educational levels and it underpins the practice of geography in many problem contexts, from applied to academic. Its value is not unique to geography, playing comparably significant roles in the learning and problem-solving practices of disciplines ranging from the sciences (e.g., astronomy, geology, and physics) to applied domains (e.g., architecture, design, and engineering).

Spatial thinking underpins many activities in the contemporary workplace, ranging from the decision-making by taxi cab dispatchers and air traffic controllers to the use of geographic information system (GIS) software in infrastructure management and urban planning. Spatial thinking also plays a significant role in everyday life, in situations such as a child being taught how to tie a shoelace, a parent following the text and graphic instructions for assembling a child’s bicycle, and parent and child together finding their way to a new friend’s house in the neighborhood.

Spatial thinking is a distinctive, universal, and powerful form of thinking on a par with, although perhaps not yet as well recognized and formalized as, mathematical or verbal thinking (Committee on Support for Thinking Spatially 2006). Although everyone can and does think spatially, there are significant individual differences in capacities and performances in relation to age, gender, training, experience, and context. Spatial thinking can be learned and should be taught at all levels in the formal education system because life in our spatial world is inconceivable without the aid of spatial thinking.

Spatial thinking defined and exemplified

The process of spatial thinking links particular forms of knowledge with cognitive operations that can manipulate this knowledge. It is a constructive amalgam of three things: an understanding of concepts of space, an understanding of the forms and uses of tools of representation, and the use of processes of reasoning (Committee on Support for Thinking Spatially 2006). Taken separately, the concepts of space, representation, and reasoning have long intellectual histories beyond the discipline of geography or any other single discipline. As in chemistry, however, an amalgam is a distinct entity resulting from the synergistic combination of elements.

The idea of spatial thinking as an amalgam is exemplified by one of the well-known classics of epidemiology and geography: Dr John Snow’s discovery in 1848 of the water-borne cause of the diffusion of cholera. Snow studied cholera at two spatial scales: in a neighborhood and city-wide. By mapping the incidence of cholera deaths in a small area of London, Snow saw a possible connection between the spatial pattern of deaths and the location of a particular water pump at the corner of Broad Street. This recognition led to a decision to remove the pump handle and to the subsequent decline in the incidence of cholera in the area around the now handleless water pump. In a second
city-wide analysis, Snow mapped the relation between spatial patterns of cholera deaths and water distribution. There were two companies providing water to London; one company drew its water from relatively uncontaminated inland subsurface wells and the other piped water directly from the heavily polluted and contaminated River Thames. Snow demonstrated the significant difference in the incidence of cholera deaths between areas served by these two water sources. Thus, Snow first connected cholera with water and then with polluted water.

Snow’s work exemplifies the synergies involved in spatial thinking, its role in seeing, believing, understanding, and communicating. By representing multiple datasets in map form, Snow reasoned from the spatial co-occurrence of two phenomena (water sources and cholera deaths) to the likelihood of a correlation between the two phenomena, and subsequently to an argument for a possible direction of a causal relation. Space, representation, and reasoning are inseparably integrated in spatial thinking.

Thinking in general

All three elements are critical for an understanding of spatial thinking. The concepts of space and representation traditionally have been central to the teaching and practice of geography. Although equally fundamental to the practice of geography, the process of reasoning has been relatively less well-articulated (Ishikawa 2013).

As is clear from the name of the concept, understanding space is fundamental to spatial thinking. Space provides the essential conceptual and analytic frameworks within which data about the properties of phenomena can be structured and related. The frameworks allow geographers to employ spatial concepts such as location, density, distribution, pattern, symmetry, isomorphism, continuity, orientation, reflection, rotation, translation, distance, scale, dimensionality, and autocorrelation to define the properties of phenomena and the relations between phenomena.

The key to the process of spatial thinking is spatialization, using the properties of space to represent data about the properties of phenomena that are spatial and, importantly, also those that are not inherently spatial in and of themselves. Phenomena that are nonspatial in nature can be assigned spatial locations based on their properties and the phenomena plotted or mapped in various spatial configurations. Degrees of proximity in the resulting space reflect the degrees of similarity between the phenomena. Patterns in the space reflect structural relations among the phenomena. The patterns can be described as random, as systematic (e.g., linear, grid-like, hierarchical), and as clustered. Pattern descriptions in turn can lead to inferences about the underlying processes generating the patterns (e.g., distance-decay effects, random walks, contagious versus hierarchical diffusion, autocorrelation, increasing or decreasing entropy, achieving stable equilibrium, etc.).

The representations resulting from the spatialization of data can be either internal and cognitive or external and physical in character, and they can be expressed in linguistic, mathematical, graphical, or concrete forms. These forms allow the spatialized data about phenomena to be stored and analyzed. Representations allow us to describe, explain, and communicate about the structure, operation, and function of the phenomena and their relations.

Reasoning processes provide the means by which the spatialized data can be manipulated, interpreted, and explained. The processes can involve making inferences on the basis of if-then analyses (hypothetical or counterfactual),
interpolation, extrapolation, analogies, and up- or down-scaling. Alternatively, by making assumptions about phenomena, relations, and processes, the operation and functioning of phenomena can be deduced.

Spatial thinking in geography

The general concepts of spatial thinking are tailored to the specific phenomena of study in a particular domain of knowledge. In the case of geography, spatial thinking is shaped by a focus on spatial patterns and processes on Earth’s surface, by a sensitivity to scalar relationships, and by an appreciation of contextual relations among people, places, and environments.

For geography, the space is that of Earth’s surface and areas immediately above and below that surface. The physical space of Earth is typically thought of in terms of dimensionality (two or three dimensions); as being described by geometrical primitives such as points, lines, areas, and volumes; as continuous or discontinuous; as bounded or infinite; as homo- or heterogeneous; as being viewed at scales ranging from local to the global, with discrete scalar steps (micro, meso, and macro) or continuous scale changes (zooming in or out); as being scaled by measures such as metric units (kilometers, hectares, etc.) or standard units (miles, acres, etc.); and as being indexed by coordinate systems (e.g., latitude and longitude).

Geographers study the links among people, places, and environments, and therefore they think of spaces beyond the physical space of Earth: there are social spaces (e.g., urban neighborhoods), economic spaces (e.g., the North American Free Trade Area (NAFTA)), political spaces (e.g., the European Union), emotional spaces (e.g., home and the area in which one grows up), and intellectual spaces (e.g., the English-speaking world). In all spaces, geographers identify places, discrete points and areas that can be characterized in multiple ways: as towns, forests, counties, and so on. Spaces are nested in scale; political spaces can range in scale from wards to counties to states to countries to global organizations such as the United Nations. Spaces are interdependent; for example, countries are defined by interlocking relationships among political, economic, social, and emotional spaces. Places take on characteristics of the spaces in which they exist; a town has a political status, a local economy, demographic characteristics, and an identity in the minds of its residents.

Representations of Earth space have been dominated by the visual sense through the medium of hard-copy cartographic maps, remotely sensed images, photographs, diagrams, and graphs, although electronic representations have become increasingly important if not yet dominant. Representations can appear in multiple ways: as small multiples, in static or dynamic forms, in stereo form, in black-and-white or color, and organized into atlases. The range of forms of representation has broadened significantly with the development of the field of geovisualization. One of the advantages of onscreen visualizations is the capacity for users to manipulate the properties and appearance of images online in real time.

Reasoning processes applied to Earth space can be deductive or inductive; they can involve interpolation or extrapolation; they can be based on statistical description or inference; and they can lead to descriptive or normative statements. These lists are by no means exhaustive but they illustrate the ways in which spatial thinking underpins and pervades the thinking processes of geographers.

However, even this description of the three concepts taken separately is insufficient to characterize the idea of spatial thinking in geography
in practice. Thinking is a complex, multi-stage process. Spatial thinking in geography involves understanding when, how, and why to transform a variable into, for example, a logarithmic form. It involves understanding the problems resulting from attempting to downscale the projections from a global climate model to a regional level. It involves knowing how to choose the appropriate map projection to express areal data versus travel distances on a world map. It involves using maps and graphs to plot the residuals from the statistical analysis of spatially distributed data. It involves understanding the effect of the modifiable areal unit problem (MAUP) on the aggregation of demographic data. Spatial thinking is inextricably interwoven into the intricacies of the everyday practice of thinking in geography. It is not a separate step to be applied. It is an integral part of doing geography.

The history of spatial thinking in geographical practice

Google’s Ngram Viewer graphs the frequency of use of words in an extensive corpus of digitized books. “Spatial thinking” has a very low frequency of use, reaching a high value of 0.000003215% in 2008 (the most recent data available). While the value was zero in 1944, there has been a significant increase in use since 2000. Despite this recent focus on spatial thinking, there is no reason to believe that it is a recent addition to the human cognitive repertoire. The “discovery” of spatial thinking is no more remarkable than the discovery made by Molière’s character Monsieur Jourdain, who remarked: “For more than forty years I have been speaking prose without knowing it.” On logical and evolutionary grounds, it is impossible to imagine humans functioning successfully without the capacity to think about the spaces they inhabit. Geography studies the interactions among the people, places, and environments that comprise Earth’s surface and thus of necessity it has always engaged in spatial thinking.

As a consequence, there is a long history of intellectual achievement in geography and the geological sciences based on spatial thinking: Eratosthenes’ remarkably accurate calculation of the circumference of Earth, William Playfair’s ingenious development of statistical graphic diagrams, Alfred Wegener’s provocative description of continental drift, Marie Tharp and Bruce Heezen’s meticulous mapping of the ocean floor that provided critical support for the idea of plate tectonics, Walter Christaller’s creative formulation of central place theory, the imaginative reconstruction by Henry Darby and colleagues of the geography of England based on entries about places in the Domesday book, and Richard Edes Harrison’s brilliant perspectival maps of Earth. To these outstanding achievements, one could add the current efforts to understand the past and, especially, the projected future impacts of climate change at spatial scales ranging from the global to the local and for timescales ranging from decades to centuries.

There have been equally significant practical cartographical achievements based on spatial thinking: John Speed’s seventeenth-century comprehensive set of maps of counties and plans of towns in England, the eighteenth-century Cassini family’s National Survey of France, the late eighteenth-century geological map of England and Wales produced by William Smith, and the nineteenth-century Survey of India’s Great Trigonometric Survey of the South Asian subcontinent.

The explorations of Earth’s surface are a testament to human resourcefulness and persistence and to the power of spatial thinking: the early nineteenth-century Corps of Discovery Expedition through the western United States
led by Meriwether Lewis and William Clark, the multiple expeditions to find the source of great rivers such as the Nile and the Amazon, the now-successful but earlier fruitless effort to find the Northwest Passage between the Atlantic and Pacific Oceans, and the remarkable tradition of South Pacific ocean sailing undertaken by the Puluwatan islanders in Micronesia (Gladwin 1970).

The history of spatial thinking in geographical theory

Within geography, although the term “spatial thinking” may have been either absent or rarely mentioned, there are many antecedents that relate directly to the current understanding of the concept. Intellectual autobiographies, such as The Geographer’s Art by Peter Haggett (1990), make clear the many roles that spatial thinking plays in the process of geographic discovery. Map reading and interpretation have long been valued and instructed skills in all branches of geography. Armin Lobeck’s (1956) Things Maps Don’t Tell Us exemplified the power of the informed geographic eye in going beyond the information given in simple maps and graphs. Fieldwork has always been central to geography, and geographers such as Charles Hunt and Peirce Lewis were incisive readers of the intricate relationships between pattern and process in physical and human landscapes, respectively. W.G.V. Balchin and Alice Coleman (1966) introduced the term “graphicacy” to complement the three skills of articulacy, literacy, and numeracy in specifying what should be taught in schools. To emphasize its importance, they referred to graphicacy as the fourth ace in the pack of fundamental skills. More recently, GIScience has become strongly linked to spatial thinking, with many of the geospatial tools providing support for the process of spatial thinking (Committee on Support for Thinking Spatially 2006).

The history of spatial thinking in psychological theory

In psychology, there are multiple concepts with similar etymological roots that overlap with spatial thinking: spatial ability, spatial skill, spatial intelligence, spatial literacy, spatial perception, and spatial cognition. Although the modifier “spatial” applies to these concepts, there is an important difference carried by the noun “thinking.”

Several of these overlapping concepts derive from the psychometric tradition which measures people’s traits or the ability to do various things: think, solve problems, understand others, perform various activities, and so on. The psychometric tradition has generated tests, especially of intelligence, and a long-standing debate as to whether intelligence is general or multifaceted. Both approaches to intelligence measurement include items based on an understanding of space. In the United States, the Scholastic Aptitude Test (SAT) and the Graduate Record Exam (GRE) contain items that involve an understanding of space. Gardner (1983) defined eight forms of intelligence of which one, visual-spatial, is clearly related to spatial thinking.

However, the psychometric approach to intelligence and concepts such as spatial ability and spatial skill are more restricted than spatial thinking. The noun, thinking, places a premium on the problem-solving process. Spatial thinking is less a property characterizing a person and more a characterization of the actions that a person performs in solving problems. It is the interaction among spatializing, representing, and reasoning that gives spatial thinking its power,
versatility, and applicability (Committee on Support for Thinking Spatially 2006).

**Approaches to spatial thinking in psychological research**

Within psychology, there are three approaches to spatial thinking. The first studies the nature, function, and operation of underlying cognitive processes. The second focuses on the development of spatial thinking over the life span. The third explores the nature, origins, and consequences of individual differences in spatial thinking. The differences among these approaches are ones of emphasis, not kind. Work on cognitive processes leads to questions about the origins of those skills and the extent to which they can be performed by all people of all ages. From the educational perspective, the approaches are valuable because they address the pedagogical question: What should be taught, why, when, and how?

A classic task in developmental psychology, the water-level task, illustrates the psychological approaches and the links to spatial thinking, especially from the perspective of geography. Developed by Jean Piaget and Barbel Inhelder (1967), the water-level task focuses on understanding the experiential links between a person and Earth space: the two universal invariant relations, horizontality and verticality, which result from gravitational forces.

In the task, people are presented with drawings of empty glasses tipped at various angles and asked to draw the level of the water if the glass were half full. The key is recognizing that the water level is invariant with respect to the horizontal, no matter at what angle the glass is tipped. The task is very challenging for young children, and children perform better with increasing age, although significant numbers of college students and adults do not understand the invariance. Results are similar if people are asked to select from drawings rather than drawing a line, suggesting that the mode of representation is not a major factor in understanding. Instructions about what happens to water in tilted glasses, physical demonstrations with water in glasses, and explicit instructions to remember key ideas all increase performance. Even with guidance, however, some people still do not demonstrate an understanding of invariant horizontality. Not only is there a significant age difference but there is also a sex difference, with men performing better, on average, than women (Liben 1991).

These results are typical outcomes from psychological research about spatial thinking. First, psychologists are ingenious in developing ecologically valid tests. For example, experiences with containers and liquids are part of everyday life. Second, hypothetical questions (“What would it look like if …?”) require reasoning about patterns, movements, and relations among objects in space. Third, understanding of space, representation, and reasoning increases with age in complex ways. An understanding of static relations, for example, develops earlier than an understanding of dynamics. Fourth, there are significant differences based on gender, experience, training, and context for many spatial reasoning challenges. Finally, space, representation, and reasoning are intellectually challenging concepts in science and everyday life. Escher’s and Penrose’s representations of impossible shapes and spatial relations are reminders that the eye and mind of a spatial thinker can be deceived, baffled, and amused simultaneously.

One of the key questions for teaching and learning spatial thinking is highlighted by attempts to explain the invariant horizontality of water. Statements about invariance and demonstrations of water’s invariance did improve the performance and, presumably, understanding.
of some people. In some instances, however, statements and demonstrations did not lead to improved performance. This raises the question of to what extent spatial thinking can be taught and to what extent teaching on a particular topic in one subject transfers to an understanding of similar topics in another subject.

Approaches to spatial thinking in education

As is sadly true for geography as a discipline, spatial thinking is not systematically taught in schools. The important qualifier is “systematically.” In school subjects such as mathematics, chemistry, earth science, and geography, spatial thinking is necessary and to some extent taught and practiced. However, its appearance in the curriculum is incidental to and contingent on other purposes. For example, in chemistry, understanding the implications of the periodic table of elements is in part an exercise in spatial thinking. Understanding conic sections in mathematics, understanding the relation between cross-sections and the processes of folding and faulting in earth science, and understanding the properties of different map projections in geography all require significant levels of spatial thinking. However, these exercises are neither sequenced in time nor coordinated across school subjects. Current instruction of spatial thinking is a hit-or-miss proposition.

At present, spatial thinking is not a stand-alone subject as is chemistry or mathematics; instead it permeates most if not all content-based disciplines. Therefore, spatial thinking does not have an obvious stand-alone presence in the curriculum. Although its role is minimal at present, spatial thinking could potentially be the functional equivalent of the concept of reading across the curriculum. Texts are introduced into as many school subjects as possible to foster an understanding of reading as an integral part of learning and thinking. In the United States, given the time constraints posed on the curriculum by the No Child Left Behind legislation (NCLB), finding room for spatial thinking in any subject will be a significant challenge, as will be the need for coordination across subjects. However, to the extent that mathematics and science are two key subjects of concern under NCLB, there is an opportunity to teach spatial thinking as part of mathematics and science instruction.

The explicit addition of spatial thinking was one of the major changes to the US National Geography Standards between the 1994 edition and the 2012 edition. As is the case with curricular statements for other school subjects, spatial thinking was implicit in many statements in the 1994 edition. In the 2012 edition (Heffron and Downs 2012), spatial thinking was explicitly and extensively used throughout many of the 18 content standards. Sinton et al. (2013) show how spatial thinking is aligned with seven of the content standards and suggest that it may underlie all 18 standards.

The goal of systematic instruction in spatial thinking is to foster a generation of students that is spatially literate. First, students must have the habit of mind to think spatially, knowing where, when, why, and how to do so. Second, they must practice spatial thinking in an informed manner. Being informed means having an across-the-board understanding of space, representations, and reasoning, and also a deep understanding of the integration of those concepts within one or more school subjects. Third, they can apply a critical stance to thinking spatially. This entails critical metacognitive awareness of the strengths and weaknesses of the approach and the ability to be articulate and convincing in using it (Committee on Support for Thinking Spatially 2006).
The future of spatial thinking in education

Spatial thinking is one of many interrelated forms of thinking: verbal, mathematical, metaphorical, statistical, and so on. Spatial thinking, for example, uses verbal, mathematical, and statistical approaches. In turn, mathematical thinking draws on notions of space and statistics, and can be represented verbally and graphically. Spatial thinking is important and distinctive but it is not unique. Human thought is multifaceted and flexible.

Spatial thinking is, however, likely to increase in practical importance. Given the increasing availability of “big datasets,” spatial thinking will be important in helping to make data relations visible, comprehensible, and communicable. Spatial thinking is inseparable from the pursuit of geographical understanding. Geography is changing as a function of new approaches, especially GIScience, and new technologies, especially GIS and geovisualization, and therefore spatial thinking is an essential skill to be fostered. Tools, especially geovisualization tools, will support spatial thinking. Those tools will be increasingly important in the workplace and in everyday life as geospatial technologies become commonplace in all forms of processing devices, from geovisualization in computers to global positioning systems (GPS) and GIS in smartphones (Downs 2014).

In the United States, the teaching of spatial thinking may become more significant at all educational levels but especially in the K–12 context. The first reason is concern about the inadequate performance of American students in the area of science, technology, engineering, and mathematics (STEM). In absolute terms, student performances are insufficient for meeting the workforce challenges of the twenty-first century, and in relative terms, the performance of US students lags significantly behind those of students in many other countries. Spatial thinking is an essential underpinning in STEM education (Newcombe 2013). The second reason is the increasing importance of critical thinking across the curriculum (Goodchild and Janelle 2010). The emphasis on critical thinking is based on the ability to reason well and the disposition to do so. Spatial thinking facilitates reasoning across the curriculum, and tools developed to support spatial thinking enable students to translate a disposition into action. The third reason is the increasing recognition of the importance of lifelong learning. Given the role of spatial thinking in STEM activities and the proliferation of rapidly evolving geospatial technologies in many careers, a baseline understanding of spatial thinking is critical for enabling people to engage in lifelong learning. In 2013, for the 25–29-year-old cohort, 90% had a high school diploma, 34% a bachelor’s degree or higher, and 7% a master’s degree or higher. Those numbers will increase and so too will the number of people already in the workforce who take online, distance education courses.

There are also reasons to believe that the teaching of spatial thinking must overcome constraints before it can become significant in education at all levels. First, there is a problem of transfer. This concerns the question of to what extent skills learned in one domain context transfer either to other parts of the same domain or to other domains. At present, there is insufficient evidence to be confident that transfer is possible to near (i.e., similar) contexts, or to far (i.e., different) contexts, and if transfer does occur, under what conditions. Second, there is a problem of developing a generation of trained teachers equipped with the materials and technologies with which to teach spatial thinking.

Within geography, there are signs that there is support for spatial thinking in multiple forms,
Spatial thinking is and has always been an integral and essential part of the fabric of geography. The same could be said for many other disciplines. At issue is the degree to which spatial thinking can be explicitly incorporated into the curricula for individual disciplines versus the extent to which it can play the equivalent role to that of reading as a skill that is essential across the curriculum. The discipline-based approach requires finding room for spatial thinking in each separate discipline, thus teaching it multiple times in an already crowded curriculum. Disciplines have some experience in teaching spatial thinking within a subject domain, but the question of transfer across disciplines remains unanswered. The across-the-curriculum teaching approach would be more efficient in terms of requirements for time, resources, and teacher training. It would be more likely to generate spatially literate generations of students. However, there are no models for a stand-alone spatial thinking curriculum and the transfer questions remain unanswered. Whatever the model adopted, spatial thinking will join the three Rs of the curriculum. That linkage, however, will depend upon research on the particular nature of spatial thinking processes in geography itself and on research on the process of near transfer within geography and far transfer to other subjects.

**SEE ALSO:** Critical spatial thinking; Geography education: primary and secondary; Geography education, workforce trends, twenty-first-century skills, and geographical capabilities; Geomorphic systems

**References**


Spatial weights

Daniel A. Griffith
University of Texas at Dallas, USA

With the advent of spatial autocorrelation analysis in the late 1960s, quantitative spatial scientists began conceptualizing geographic relationships in a variety of ways, one of which is described by Tobler’s first law of geography: “Everything is related to everything else, but near things are more related than distant things.” This statement alludes to the 2-D, two-directional nature of geographic relationships arising from the relative location of things. In doing so, it requires a conceptualization that captures this complicated relational structure. In addition, the locations in question almost always are not systematically distributed across a geographic landscape, although by being evenly spaced remotely sensed images are an important exception. Accordingly, the conceptualization needs to accommodate irregular spacing of points, and varying sizes and shapes of polygons constituting areal units. The statistics literature offers guidance for formulating this type of conceptualization with its entries concerning the use of weight functions to quantify the differential importance of observations in a calculation, perhaps to align the relative frequencies of observation types between a sample and its parent population, or to adjust for non-constant variance associated with observations. These weights are observation specific, with the default case being all observations having the same weight value. In a geographic context, these weights are specific to individual pairs of locations and, as such, are spatial weights (SW) that quantify the relative possibility or strength of a relationship between two given locations. Classical analysis assumes that these weights are zero, implying that geographic locations are independent. Contemporary spatial analysis assumes that they are non-negative, employing model specifications in which locations are correlated. The fundamental idea here is that phenomena present at a given location are related to the same category of phenomena present at its nearby surrounding locations. This geographic relationship can arise from either direct interaction among (e.g., air pollution spilling over into nearby locations), or common factors underlying (e.g., climate, soil fertility), geographically distributed phenomena. SW provide one way to represent the geographic structure giving rise to such geographic relationships.

SW specifications vary according to how geographic relationships are conceptualized. Popular criteria used for this purpose include: contiguity (e.g., common boundaries of polygons, urban system central place linkages), continuous distance (e.g., inverse, power inverse, negative exponential), nearest neighbor distances (e.g., the six closest), fixed distance band (e.g., all locations within a given proximity range), and spatial interaction (e.g., the number of migrating households). Figure 1 illustrates each of these conceptualizations. Because relative location relationships cast a geographic landscape as a set of \( n \) interrelated locations, its set of SW is organized and operated with as an \( n \times n \) matrix, where the row and column labels identify the locations, with common practice being that both matrix axes use the same ordering of the locations. In other words, each location has one row and one column in this matrix. The diagonal
Figure 1  Conceptual bases for operationalizing SW. (a) Rook’s adjacency for a regular square tessellation (black pixels are neighbors to the gray pixel). (b) Black lines denote rook’s adjacency for an irregular surface partitioning (part of Florida). (c) Inverse distance proportional to line weights connecting county centroids (Chicago metropolitan statistical area). (d) Queen’s adjacency for a regular square tessellation (black pixels are neighbors to the gray pixel). (e) Additional zero-length boundary queen’s adjacency for an irregular surface partitioning (part of Florida) denoted by a gray line. (f) Journey-to-work flows proportional to line weights connecting county centroids (Chicago metropolitan statistical area).

entries of this matrix (i.e., the cell for which a location’s row and column intersect) are zero; everything perfectly correlates with itself, which is uninformative and has nothing special to do with location. The off-diagonal cells contain the SW for their corresponding row and column locations. Concatenating the values of some geographic distribution by location into an $n \times 1$ vector permits it to be premultiplied by a spatial weights matrix (SWM). This multiplication yields an $n \times 1$ vector of the sum of surrounding values. The zeros in the diagonal eliminate the values themselves from their corresponding sums (i.e., 0 times any number is 0). A scatterplot of these pairs of values and their surrounding value sums furnishes the information necessary to study a global relationship across, or local relationships within, a geographic landscape. The former is the trend line in the scatterplot; the latter are geographically clustered deviations from the trend line.

SW are used extensively in spatial autocorrelation calculations, spatial autoregressive model
specifications, geostatistical semivariogram construction and kriging, and Moran eigenvector spatial filter construction.

**Principal types of spatial weight**

In the tradition of time series analysis, the initial SW were binary 0–1 values based upon contiguity: if two polygons are contiguous, then they are considered neighbors and are related, and hence their SW is one; otherwise, their SW is zero. Exploiting analogies with rook and queen pieces moving across a chess board, this contiguity may be defined in two different ways. The rook’s definition assigns an SW the value of one if two polygons share a nonzero length (i.e., line) common boundary; otherwise, the SW is zero; the rook moves from square to square in a straight line either horizontally or vertically. Implicitly, this is the definition used in the original Moran Coefficient and Geary Ratio calculations. The queen’s definition assigns an SW the value of one if two polygons share either a nonzero or a zero-length (i.e., point) common boundary; the queen moves from square to square in a straight line either horizontally, vertically, or diagonally. This is the default option in GeoBUGS, the Bayesian mapping module in WinBUGS. In most irregular surface partitioning situations (e.g., counties forming a state), frequently the queen’s definition increases the number of connections by no more than roughly 10%; but for remotely sensed images and other regular square lattice partitioning, this increase can approach 100%. The associated scatterplots describe values at locations as being a function of the sum of values at surrounding locations. Many researchers find this relational description unappealing, preferring to describe values at locations as being a function of the arithmetic average of surrounding values. To achieve this result, the SWM is row standardized: each one in a row is divided by its corresponding row sum. This is the most popular specification for spatial autoregressive models. However, GeoBUGS employs the 0–1 binary version for its autoregressive model specification. The 0–1 and row-standardized versions are limiting cases for a family of possible SW. If $n_i$ denotes the sum of the ones in row $i$ of an SWM, then the family members are given by $n_i^c$, $0 \leq c \leq 1$. If $c = 0$, then the SWM is based upon binary 0–1 values, commonly being denoted by $C$. If $c = 1$, then the SWM is row standardized, commonly being denoted by $W$. The exponent $c$ can be some intermediate value, tending to optimally control for topology induced heterogeneity at 0.5, and as such is referred to as the S-coding scheme (where S signifies stabilizing variance).

Not all georeference data are tagged to polygons. And not all neighborhoods are conceptualized in terms of contiguity. Because irregular surface partitionings often have a large range in polygon size, sometimes neighborhoods are defined in terms of polygon centroids (e.g., geometric center, spatial mean of population, principal administrative city); geometric center coordinates can be calculated easily in most geographic information systems (GISs), although multiple-part and concave polygons may require extra effort. As an aside, geographically distributed points can be converted to polygons by constructing their corresponding Thiessen polygon surface partitioning. SW can be constructed based upon polygon centroids. Binary, S-coding scheme or row-standardized SW can be defined for nearest neighbor centroids. Including a sufficient number of nearest neighbors avoids having the accompanying SWM represent a disjoint set of subregions. Most often, SWMs are specified so that a path can be traced from any given location to any other location in a geographic landscape by at least one sequence of nonzero SW; this is a desirable property of these matrices. Binary,
S-coding scheme or row-standardized SW can also be defined for centroids within a fixed distance band. This conceptualization is the basis for the queen’s definition of contiguity, which involves nonzero as well as zero length (i.e., point) boundary connections. The distance band should be large enough to avoid disjoint subsets, but small enough to avoid corrupting the localness of spatial relationships. Sometimes the (effective) range indicated by a semivariogram is used to determine the size of this distance band. One drawback of this specification is that interlocational relationships are not necessarily formulated to decline in strength with increasing distance.

Row-standardized SW can be created with continuous distance, too. One common specification is based upon inverse distance, \(1/d_{ij}\), for locations \(i\) and \(j\) and their intervening distance \(d_{ij}\). This weighting scheme discounts pairs of points with larger intervening distances. Rather than zeroes being assigned as SW, zeroes are approached asymptotically as distance increases to infinity. Now the diagonal entries of the SWM must be coded zero by construction. Once all of the inverse distances are calculated for the entries of an SWM, each entry is row standardized. This weighting scheme can be generalized (i.e., powered inverse distance) by attaching a positive exponent to \(d_{ij}\), resulting in \(1/d_{ij}^c\) for \(c > 0\). The exponent \(c\) cannot equal zero, or everything is depicted as being related to everything else, regardless of location; \(c = 1\) for the inverse distance specification. As \(c\) increases beyond one, the SW increasingly approximate those for the row-standardized binary SW specification; \(c\) may need to be as large as 6–7 to achieve this outcome. The negative exponential distance decay specification furnishes a third alternative. Here the SW are based upon \(e^{-c d_{ij}}\), where \(e\) is the base of the natural logarithms. As for the generalized inverse distance case, a value is needed for \(c\). The coefficient \(c\) may need to be as large as nearly three for these SW to approximate those for the row-standardized binary SW specification.

The magnitude of spatial interaction flows furnishes an additional basis for specifying SW. The number of goods, services, money, telephone calls, or people (e.g., journeys-to-work, to-shop, and to-recreation), say, flowing among locations can be inserted into an SWM. Next, the diagonal cell entries are coded with zero. Then the matrix is row standardized. This SW specification captures both a contiguity type of spatial relationships, and more distant spatial relationships attributable to geographic hierarchical effects. This latter component is missing from the preceding SWMs. If the locations are cities, then central place theory offers a hierarchical articulation of places that transcends mere contiguity. An alternative approach would be to specify two SWMs with the preceding specifications, one for contiguity and the other for hierarchy.

The conceptualization of SW has changed over time, primarily moving from the binary 0–1 version to the row-standardized version. Few examples appear in the literature involving the use of the S-coding scheme, generalized inverse distance, negative exponential distance decay, or multiple SWMs for a single geographic landscape. GISs often offer the option of employing rook/queen contiguity, fixed distance bands, or inverse distance/squared distance, or a user-defined set of weights (which could be any one of the other specifications). These packages also might furnish a metric other than Euclidean. A few more recent studies employ flows-based SW, a specification that replaces the nominal measurement scale binary row-standardized specification with a ratio scale measure of spatial linkage, although it is not without drawbacks (e.g., it can replace an exogenous with an endogenous SW specification).
Selected salient spatial weights controversies

Current emphases in SW research include the data analytic nature of, and somewhat arbitrariness associated with, SW specifications. The preceding section illustrates this perspective: SW can be specified in a number of different ways, some of which define a particular weight as zero while others define it as nonzero. Various critics argue that endogeneities of SW, arising from their definition in terms of specific phenomena or the choice of locations, are problematic. For example, if a response variable is a function of spatial interaction, and these flows are used to define the SW, correlation may well occur between the error terms associated with the SW and the disturbance terms in spatial autoregressive models. Or, changing geographic scale or resolution almost always results in changes in at least some of the SW.

Frequently, defining SW in terms of geographic distance for a functional geographic landscape avoids serious endogeneity complications. Rather than using phenomena to refine the specification of SW, percentage of common boundary lengths can be employed, as can hybrid versions combining such refinements with intercentroid distances, or changing the metric (e.g., Euclidean to network shortest paths). Meanwhile, because demarcating functional geographic landscapes can be impractical or nearly impossible, corrections for edge effects can be employed to ameliorate SW changes attributable to varying the size of a geographic landscape for a given resolution.

Sensitivity of spatial analysis results to the specification of SW may well be a myth (LeSage and Pace 2011). Incorporating some SW scheme achieves the biggest improvements in data analysis results, whether it is over- (i.e., has too many nonzero values) or under- (i.e., has too few nonzero values) specified. Identifying the correct set of nonzero SW makes little difference if most are correct, if no neighbor set is distorted in such a way that it contains most, if not all, of the other locations (Griffith and Lagona 1998). This feature is desirable because otherwise the SWM would be an unreasonable tool for analyzing relationships involving spatially correlated data. It also helps counter criticisms about the arbitrariness and ad hoc construction of SWMs. If spatial analysis results are robust in terms of an SWM specification, then they are not conditional on subjective decisions about its definition; accordingly, estimates and inferences are insensitive to small changes in SW specifications. However, to preserve mathematical statistical properties of estimators, modest underspecification seems preferable to overspecification, and using SW appears to increase the small sample size threshold from the conventional 30–100 to 60–200 for moderate positive spatial autocorrelation cases. As an aside, the automated procedure AMOEBA (a multidirectional optimum ecotope-based algorithm) has been developed to create SWMs in a more objective way by exploiting local spatial autocorrelation statistics.

Future directions in spatial weights research

Future SW research needs to address three prominent issues: SWM comparisons, integrating multiple spatial structures, and space–time specifications. The first issue concerns informative comparisons of SWMs. These comparisons are needed for a number of different definitions of the same set of SW, as well as for the addition and/or subtraction of various nonzero SW for a given definition. Because a set of SW can be organized as a matrix, mathematical properties of matrices furnish one criterion for comparison.
SPATIAL WEIGHTS

Any matrix can be characterized by mathematical quantities known as eigenfunctions, which relate to a number of its features (e.g., trace, determinant). An \( n \times n \) matrix has \( n \) such functions, each of which is a scalar paired with an \( n \times 1 \) vector. All of these functions are real for a symmetric matrix with real entries. Eigenfunctions are encountered in regression and multivariate statistical theory also. Eigenvalues of SWMs quantify distinct patterns of correlated subsets of locations. If no correlation exists amongst locations, then the eigenvalues equal the diagonal entries of the matrix; for SWMs, each would equal zero. For a row-standardized matrix, the largest eigenvalue in absolute value is one. If all SW are equal (i.e., all locations are related, and to the same degree), then the largest eigenvalue is one and the \( (n - 1) \) remaining eigenvalues are \(-1/(n - 1)\). If a zero spatial weight becomes nonzero, or vice versa, some of the eigenvalues change. Therefore, the rank ordered sets of eigenvalues furnish a means for determining the degree to which different SWMs differ. To maintain a standard measurement scale, an SWM whose largest eigenvalue is greater than one can have all of its eigenvalues divided by its largest eigenvalue. The important research question to be addressed here asks by how much sets of eigenvalues can differ without the SWMs rendering different statistical estimates and inferences.

A second issue is the use of multiple matrices. Defining SW in terms of flows in space may suffer from endogeneity complications. But such definitions capture both geographic contiguity and geographic hierarchical relationships. The two components are necessary for properly analyzing, for example, spatial diffusion phenomena. Spatial relationships channeling leaps across space can play an important role in understanding and properly analyzing the geographic distribution of a phenomenon. Much of the existing SW literature addresses contiguity relationships. Adding a second SWM for an underlying hierarchical structure complicates autoregressive estimation, although its specification is a straightforward extension of existing SW specifications. The mathematical statistics algebra for estimation and inference remains tractable. But calculations become considerably more numerically intensive. Moran eigenvector spatial filtering may furnish an efficient and effective way to handle this complication. This relatively new spatial statistical methodology employs the eigenvectors of the SWM to account for spatial autocorrelation latent in georeferenced data. The important research question to be addressed here asks how these two-SWM spatial statistical models can be implemented. Space–time research offers some hints about this.

A third issue, which relates to the preceding one because it involves multiple matrices, concerns space–time data, which are becoming more readily available. Two popular conceptualizations for such data are instantaneous and spatially lagged space–time correlation. Both cast an attribute value at a given location during time \( t \) as a function of that attribute’s value at that location during time \( t - 1 \). The former also casts an attribute value at a given location during time \( t \) as a function of its neighboring attribute values at time \( t \), whereas the latter casts this attribute value as a function of its neighboring attribute values at time \( t - 1 \). Geographic and temporal distances have different metrics, complicating specification of the overall correlation structure. But the SWM and the temporal weights matrix can be combined, and in efficient ways when space and time correlation components are separable. The important research questions to be addressed here ask how to test for space–time separability, and how space–time conceptualizations can incorporate nonseparable space–time correlation structures. Both Moran eigenvector
space–time filtering and space–time geostatistics already involve research focusing on this topic.

**SEE ALSO:** Geostatistics; Quantitative methodologies; Spatial analysis; Spatial concepts; Spatial context; Spatial econometrics

**References**


Further reading


Spatiality

Audrey Kobayashi

Queen’s University, Canada

The conceptual transit from space to spatiality is one of the most significant developments in recent geographical thought. Both space and spatiality are historical concepts that refer to the relationship between people and things, including other people, on the Earth’s surface. While spatiality is relatively recent, space is a historical concept. Like all human inventions, it has been subject to many interpretations. Space is one of the most – if not the most – important concepts in the discipline of geography. It is also the most difficult to define and possibly the most contested. For students of the subject it is a bewildering term, sometimes depicted as an absolute product of nature with a predetermined structure, sometimes as a metaphor that is as changeable as the imagination, sometimes as an abstract concept that defies either specificity or logic. Often it is used uncritically by geographers – as “space and place” – based on unquestioned assumptions that its use will be understood by readers. Such assumptions make understanding space even more difficult.

The major issue is ontological. Just what is this thing geographers call “space”? Is it, after all, a thing? Is it an assemblage or collection of things? Is it a conceptual principle for organizing things? Is it notional, relational, or rational? Does it exist and, if so, where and how? What does space look like? How does it work?

Geographers generally acknowledge that space is “no longer viewed as a fixed and absolute container within which the world proceeds. Rather, space is seen as a co-production of those proceedings, as a process in process” (Thrift 2008, 86). Space then is relational; it is fungible and never static; it is contradictory – both unstable and capable of irrepressible stabilization; it is a product of conflict; it is inherently political and powerful on the one hand, and capable of myriad redefinitions and reimaginings on the other. Many geographers would assert that it is the task of the discipline to understand how space has changed and is changing, treating space as an emergent condition of the world.

Doreen Massey (2005), in one of the major works on the topic of space, challenges geographers to let go of and to reimagine the conceptual categories of space and place. She sees the problem as one in which the inherent spatiality of the world has allowed humans to reify these notions. To contest such reification and the sociopolitical issues it creates, geographers need to refuse to think of space as a singular phenomenon, such as the Earth’s surface, on which the objects that fill the world are placed; to refuse to convene space into time; and to challenge the tendency to acknowledge notions of space or place in ways that convey power relations on earthly territories. Massey’s work has been incredibly influential, and has influenced a generation of geographers to think differently, as well as to meet the conceptual challenge posed by thinkers in other disciplines for whom space has in recent years become something of a fetish.

In the end, however, Massey still titles her book For Space. Spatiality, a condition of relationship between people and things, is somehow returned to space, and the reason for such nominalization is never made quite clear. The
book ends on a somewhat mysterious note (and maybe that is just what she intended) that recasts the taken-for-granted assumptions about space that the book has challenged so successfully. This entry on spatiality challenges the reader to give up on space – especially its conceptual mysticism – altogether, to cast off the comfort blanket of the term, to think instead of spatiality.

**Space: a linguistic paradox**

The term or concept of space varies drastically in different languages, and in many languages there is no usable translation. In English, the term “space” originated with the Latin word *spatium*, translated variously as “room,” “area,” “distance,” “extent,” or “duration.” The Old French term *espace* carries a similar meaning, conveying a sense of distance or interval, which has been taken up by numerous French philosophers in the twentieth century and has influenced recent poststructural thinking. The English term is especially fascinating because its connotations and everyday usages have changed over time with changing philosophies and technologies. During the fourteenth century, it was used in connection with marking territory or boundary, a common part of the discourse of colonial expansion. From the fifteenth century, it was incorporated into musical theory, and by the sixteenth century into typography (still used today, as the “space” between words). By the seventeenth century, it had become more abstract with the development of geometric theory, as that which comes between points and lines, a notion that extends into theories of mathematics and physics of the twentieth century as a metaphor, even more abstract, for that which does not exist except to differentiate points, which also do not exist as such. Also in the twentieth century, the term took on many aspects of modernity, as the void outside of Earth, which became the obsession of both science fiction and astronomy (“space” travel), as a psychological state (from “personal space” to “head space” to “spaced out”), to most recently “cyberspace,” a realm that is both familiar and ineffable to most of the world today.

In most of these popular understandings, space is either full or empty, that is, it is either something or nothing, either a physical thing that has dimensions, is measurable, and provides the substance of the world, or an abstract concept that allows us to think about the relationships between actual things, whatever they are. Space has monetary value (think of its usage in real estate sales), or political value (as a fundamental commodity in colonialism, for example). The myriad uses of the term in ordinary language reflect multiple meanings, conflicting ideologies, diverse cultural practices, and changing technologies. Space is thus a paradox, not only for intellectuals, but also for politicians, capitalists, and ordinary people, a contested “thing” that can only be understood in relation to other things. But its ontological status remains unclear. Geographers have even debated whether an ontology of space is needed, whether more than one ontology can exist simultaneously, and whether we should even think about the issue when there seems to be some sort of linguistic consensus about what we are talking about. How important is it for geographers to worry about the theoretical basis of this thing called space that is both abstract and concrete at the same time? Should they settle for conceptual pluralism?

**Geographical concepts of space as a thing**

For most of its history, the discipline of geography has been a sparring ground for debates over
the thingness of space. Until the mid-twentieth century, much of the discipline assumed, often uncritically, Newton’s "container" theory of space as something independent of the things that fill it. Leibnitz, in contrast, was a relativist, who wrote:

These Gentlemen maintain therefore, that Space is a real absolute Being. But this involves them in great difficulties. For such a Being must need be Eternal and Infinite. Hence some have believed it be God himself or one of his Attributes, his Immensity. But since Space consists of Parts, it is not a thing which can belong to God … I hold Space to be something merely relative, as Time is, that I hold to be an Order of Coexistences, as Time is an Order of Successions. For Space denotes, in Terms of Possibility, an Order of Things which exist at the same time, considered as existing together, without enquiring into their Manner of Existing. And when many Things are seen together, one perceives That Order of Things among themselves. … I have many Demonstrations to confute the Fancy of Those who take Space to be a Substance. (Leibnitz, in Clarke 1717; emphasis original)

This quotation provides the gist of a metaphysical argument that went on for centuries (and in some respects is still going on, although it has been long surpassed in the discipline of physics, where theories of relativity have been firmly established since Einstein’s work in the early twentieth century).

Perhaps none of the Enlightenment thinkers was more influential on the discipline than Immanuel Kant, who attempted to move past the absolute–relativist debate by depicting space as a sensory capacity. Well aware of the debate between Newton and Leibnitz, Kant first wavered between thinking of space as absolute and as relative, before settling on the original and revolutionary idea that space, like time, is a pregiven cognitive category that allows humans to perceive an outer reality in spatial terms. Both external reality and internal perception operated for Kant according to laws that consciousness could not itself violate. The project of the "Enlightenment" was to use human reason to discover and apply those laws to increase knowledge. Kant’s dualism between inner consciousness and the outer material world has been maintained in geographic understandings of space as both outer materiality and inner, or mental, perception.

Kant influenced the modern discipline of geography by positing the spatial as chorological, that is, as the organizing principle for things that occur in proximity, just as time is the organizing principle for events that occur in sequence. Chorological thinking dominated human geography during the first half of the twentieth century. Geographers often referred to things “in space,” and it was largely assumed that areal or regional differences are spatial. Geographers debated the nature and existence of regions; however, discussions of Space itself, or its ontological status, were of little expressed interest for most geographers until the mid-twentieth century.

With the spatial turn from the 1950s onward, interest focused explicitly on spatial theory. Studies of spatial pattern proliferated, and the term “the science of space” became common in geographical writing, so much so that some geographers have labeled thinking during this time as a “spatial fetish.” That fetish, nonetheless, departed from the container or absolute view of space to a relative understanding of the externally driven processes that create spatial patterns and relationships. David Harvey (1969, 209), among others, contended that “Space may well be the central concept on which geography as a discipline relies for its coherence.”

The “space” of the spatial turn was construed relatively, according to Leibnitz, not Newton, and in line with the thinking of relativity theory.
Jones (2009) claims that the four-dimensional concept that resulted depended on two assumptions: (i) that spatial order can be determined only through relations between things in space; and (ii) that spatial relationships are not fixed – they change over time. Regional science, then, depends by and large on a concept of space–time. It is also ironic, however, that, while much was written about discovering the laws of space, very few engaged in a discussion of its ontology. At issue was the status of spatial laws, not space itself, laws that could be observed empirically in predictable, measurable, mappable, and repeatable patterns on the Earth’s surface. At this time, the practice of urban planning expanded, and many of its practices responded to a burgeoning global population, a rapidly expanding network of roads and airways, and massive, unprecedented increases in the movement of people and things, not to mention the proliferation of weapons as technologies that could quickly and dramatically disrupt spatial patterns. To respond to changing spatial conditions and possibilities, practices of regulation and control were developed and strengthened in nearly every aspect of human governance. Geographers who could map such changes were in their disciplinary element.

**Overcoming dualism**

Controversy over space and place erupted in the 1970s as a correlate to controversies over the role of spatial science. Some Marxist geographers, including David Harvey, challenged the spatial turn on the grounds that the aim was to understand not spatial processes but the process of capitalism. For them capital produced space, rather than the other way around. Some humanists, particularly phenomenologists, rejected the concept of space altogether as abstract, and focused instead on place, as that which is filled by human experience. Yi-Fu Tuan (1977) attempted to get past the place/space dichotomy by depicting space as that which exists prior to place; in other words spaces become places as a result of human actions. In this interpretation, space is an abstract concept, since empty space could not exist as such. Many geographers began to write of “space and place,” although much of the work that uses this term, even up to the present, is extremely vague about why geographers study both, and about just what is meant by these two nouns or how they are distinguished from one another.

Marwyn Samuels (1978) penned one of the first in-depth discussions of human spatiality in the discipline of geography, creating not an ontology of space but a spatial ontology of human being. Samuels refuses to reify space and acknowledges that human beings have produced different kinds of spatiality, among which spatial science is but one. Using theologian Martin Buber’s existential notion of human relationship as setting the other at a distance in order to enter into relationship, he explores the complex and ironic situation of human alienation on the one hand, and yearning for relationship on the other, as the result of both the fundamental human condition of spatiality, and the many factors through which relationships (which include everything from direct human contact to the roads by which we overcome distance) are founded. He sees irony in the fact that, as human beings overcome distance technologically, they create new kinds of alienation.

As the tenets of poststructuralism began to infiltrate the discipline, a consensus had developed that there are various kinds, or modes, of space, and that human agency determines the qualities of space. Henry Lefebvre’s (1991/1974) monumental work *The Production of Space* was justified on the grounds that, notwithstanding volumes of work, theorists had failed to
establish a viable science of space, and that spatial scholarship consisted of “bits and pieces” of largely descriptive research without much analytical rigor. Lefebvre set out to address the dualism between mental or psychological space and space as it is experienced in the everyday. He posits that all spaces – abstract, absolute, or social – are historically constructed. They become, are organized, represented, and given meaning, by collective human action. He seeks to analyze “not things in space, but space itself, with a view to uncovering the social relationships embedded in it” (Lefebvre 1991/1974, 89). Space is therefore the totality, the every thing, of a delineated portion of the Earth. It can be a room, a street, a city, even a nation, whose qualities are determined by human habitation. Dialectically, it is both the product of human capacity and constituent of human experience. Lefebvre’s space is thus distinguished from that of the chorologist because he obviates the dichotomy of space and things. For him space is a project – in fact, it is the human project – and thus represents the projection of humanity on the world.

Lefebvre depicts three modes of spatiality: spatial practice, the material process of creating the world around us; representations of space, as the power-laden practice of regulating or structuring human behavior spatially, and finally spaces of representation, in which the human imagination works in various ways to create experiences that are not always open to objective capture. Human geographers have capitalized extensively on this triad in order to organize conceptually the seemingly endless ways of understanding spatial being. The most sophisticated expansion of the triad is by Edward Soja (1996), who deems Lefebvre’s modes first, second, and third space, and then projects the concept of third space into a complex journey across the heterotopic landscape of Los Angeles, an “Exopolis” across which the conditions of postmodernity are spread through imagination, metaphor, irony, paradox, and virtuality. In addition to providing a compelling account of an actualized landscape, Soja provides a healthy geographer’s critique of some of the excesses of spatial thinking in the work of continental poststructuralists.

The recent contributions of poststructuralist feminists in geography have particularly emphasized the complex and often conflict-laden nature of sociospatial experience, and the utilization of “paradoxical space” (Rose 1993) to capture the ways in which especially the spatially excluded imagine and act on the conditions of power and control through which the world has been structured, governed, and disciplined, and in the process made uneven and oppressive. Sue Ruddick (2009, 224) views “Society–Space” as an “immanent field upon which geographers have the potential to unbound, destabilize, re-assemble, and transform the concept of space not only as a geographical theory but as a way of being.” In her view, “space is not strictly a techne, nor a juridicopolitical boundary, but the manifestation of process itself, a kind of becoming … in a process whereby subjects’ capacities are invested through particular roles that they carry regardless of their location” (Ruddick, 2009, 224). In other words, we have reached a point of conceptual no return where we no longer study space itself.

Spatiality

Geographers have bashed away at the concept of space itself to the point where, as itself, it no longer carries much analytical utility, not simply because its meanings have become so diverse as to be almost meaningless (although perhaps they have), but because the discipline is more and more aware that it is precisely the spatial–social,
the becoming of humanity, rather than the fictional quality that encloses or regulates humanity, that is our major concern. Boyle and Kobayashi (2011) turn to Jean-Paul Sartre, whose work has been virtually ignored by geographers (at least in print). They take up Sartre’s refusal of colonialism, and his acknowledgment of the metropolitan anxieties created in a world in which colonial distance is being transformed, as a basis for understanding human existence as collective spatiality, a dialectic that Sartre refers to as spatializing–spatialized:

both a product of and modality through which particular formations of the colonial become possible, through particular forms of human relationship (spatializing–spatialized) and in particular landscapes ... We refer to these historicized phenomenologies as: spatiality as tyranny, spatiality as liberation and spatiality as incarceration. (Boyle and Kobayashi 2011, 420)

Spatiality is the condition, the form, the becoming of human, concretized through and as human relationship. In other words, the shift is from relative space to relational spatiality.

The social theorist Gilles Deleuze once remarked that it was the Sartrean dialectic that made poststructuralism in the twentieth century possible. This dialectic informed subsequent thinking in which knowledge, or discourse, is always and only dialectical reason/praxis, “always partial, necessarily contingent, purposive ... unbounded, and never separate from its concrete, material, historical, circumstances ... Praxis imposes on knowledge the synthetic engagement with the world (existence), as well as the possibility/necessity to transcend existence (becoming) through distantiation,” and therefore is never outside the dialectic either as an ideal (pure thought) or as an external structure (Kobayashi 1989, 170). For geographers, then, to understand spatiality, or distantiation, is fundamental to understanding the possibility of existence. The possibility of “history” becomes the possibility of geography: spatializing–spatialized.

Perhaps spatiality is thus, to use Ruddick’s (2009) term, human geography’s immanent field. To claim that the concept of space itself is therefore redundant is not, however, to substitute one abstraction for another. It is rather to recognize that the sociospatial – the only – world is the complex, interpretive and interpolating, interconnected, often paradoxical and contradictory, stuff of (human) relationship, in all its forms and modes. Spatiality is a condition of being, not a thing in itself. It is no longer possible – if it ever was – to isolate space from its sticky context, nor is it possible to isolate spatiality as a particular mode of being or historical moment, a particular place, landscape, or setting. Rather, geographers increasingly recognize and understand spatiality, and its possibilities, as a dialectical process.

SEE ALSO: Areal differentiation (or chorology); Critical geography; Existentialism; Feminist geography; Humanistic geography; Place; Radical geography; Space

References


Further reading

Hubbard, Phil, and Rob Kitchin, eds. 2011. Key Thinkers on Space and Place, 2nd edn. Los Angeles: SAGE.

Spatialization

Sara Irina Fabrikant
University of Zurich, Switzerland

Information access and dissemination

Efficient and effective access to and knowledge construction from massively growing spatial and nonspatial databases available online have become major bottlenecks for the rapidly evolving information society at large. The rise of the Internet since the 1990s, especially the rapid expansion of the World Wide Web, and the current growth of social-networking platforms brought about the construction of massive, multivariate and multimedia online databases (e.g., Wikipedia, OpenStreetMap, Flickr, etc.). These (increasingly user-generated) databases can include vast amounts of non-numerical, semistructured data (e.g., Wikipedia entries, blog posts, etc.) or nonstructured data (e.g., online books), and these data very often also contain geographically relevant content. Most geographic information systems, however, typically rely on numerical, tabular georeferenced information for space–time data analysis; the kind of numeric data that fit into a table, neatly organized in rows and columns. Analyzing large and complex alphanumeric datasets with traditional spatial analysis methods can become inadequate, as the data may be bound by a priori assumptions (e.g., dependent on a particular data distribution, or a sampling method, and moderate sample sizes, etc.). Due to their inherent complex, multidimensional nature, vast unstructured and semistructured datasets need to be transformed into meaningful chunks of information, and adapted to the limited information-processing and sense-making capacity of humans. With these information science-related developments mostly happening outside of geography, it became apparent to geographic information scientists that methods and approaches that geographers have been using for hundreds of years to model and visualize geographic phenomena could also be applied to the representation of any data record, object, phenomenon, or process exhibiting spatial characteristics (Fabrikant and Buttenfield 2001).

Many different types of spaces and spatial reference frames exist beyond geographical space that could be reorganized, analyzed, and explored with spatiotemporal methods, and mapped using well-established cartographic practices. For example, chemical structures that build up to the human genome, medical records that refer to human body space, or neuronal connections in the human brain are good examples of non-geographic spaces that lend themselves for spatial analysis and mapping. For example, Atkins (1995, 34) depicts the periodic table as a stepped undulating terrain of chemical properties. The higher the hills in this landscape of the chemical elements, the larger the atom diameters of the depicted chemical elements.

Spatialization offers the field of geography, which investigates space and spatial relations, opportunities to apply its rich spatial knowledge to noneographic, and even nonspatial, data domains (Couclelis 1998). One can also imagine abstract information worlds presented in map-like displays for visuospatial analysis that are not spatial at all. For example, real-time digital
SPATIALIZATION

stock transactions create a space of money flows, and digital interactions of people through online networking platforms form social networks in cyberspace, which could both be mapped in the form of transportation network maps. Songs played on online music sharing websites create a landscape of musical tastes that can in turn be depicted as a relief or terrain map in which musical tastes shared by many people pile up metaphorically to create mountains of popular music styles. Figure 1 depicts a two-dimensional (2-D) space of musical tastes and listening experiences sampled from members of the popular online music sharing website Last.fm. Listeners of this social music media website are invited to share personal information about the music they upload or listen to by tagging songs with additional text information. Based on users’ listening patterns, this information is then used by the system to suggest additional musical pieces that users might enjoy listening to.

The music spatialization in Figure 1 renders a meaningful spatial pattern of musical styles, where rock music pieces cluster in the western parts of this Last.fm landscape (Skupin, Biberstine, and Börner 2013), clearly separated from world music, classical music, and jazz, concentrated in the eastern parts of this relational world of musical tastes. Interestingly, funk music, located in the southeast of the space, seems to be seen as a musical transition zone between jazz and classical music.

Spatial metaphors

Spatialization takes advantage of people’s familiarity with space in everyday life to produce information spaces that are intuitive and internally coherent. Everyday experience of the real world involves visuospatial perception and memory, spatial reasoning, and communication about features and objects in the environment, their spatiotemporal and thematic attributes, and the relationships between these objects. For example, instead of presenting users with large abstract tables or long lists of queried items from an online archive in text format, information spatializations allow users to visually explore graphic displays of information by means of spatial metaphors, as if they were exploring a real landscape or a cartographic map of a real environment, allowing them to see and experience the “layout” of the information in a single view (Skupin and Fabrikant 2003).

Successful interface metaphors create explanatory theories for users interacting with a system (Kuhn 1996). That is, users can map the processes and relations of a real-world source domain (e.g., geographic space) onto the processes and relations of the target domain (e.g., an information space), allowing them to use their knowledge of familiar source domain operation to predict the unfamiliar, often abstract target domain operation. What is common to geographic and nongeographic spaces (e.g., body space) are fundamental spatial concepts such as location, distance, arrangement, hierarchy, scale, and so on. A fundamental assumption is that spatialized displays work because users can understand them intuitively (Wise et al. 1995). If this assumption is generally true, understanding the fundamentals of geographic space (the metaphor’s source domain) as understood by display users will help to construct cognitively adequate information displays based on meaningful spatial metaphors (target domains). Location and distance (i.e., arrangement) on the Earth’s surface are among the most fundamental geographic primitives, and both are reflected in the distance–similarity metaphor, that is, more similar items should be placed closer to one another in a spatialized display because closer items will be seen
by people as more similar than distant items. Montello et al. (2003) have called this principle the first law of cognitive geography, as a consistent result of a series of empirical studies with various types of spatialized displays. This research provides empirically validated evidence to support Gatrell’s (1983) proposal for using a broader conceptualization of space in geography, and thus to employ a variety of distance metrics beyond Euclidean geometry to analyze and visualize a multitude of spaces of interest to geographers. The current popularity of ordination methods utilized to reorganize and mine vast amounts of spatial and nonspatial data, for example, by using nonmetric multidimensional scaling, principal component analyses, and self-organizing map techniques in various subfields of geography (i.e., physical geography, remote sensing, and behavioral geography), is another good example
SPATIALIZATION

of the analytical power of the distance–similarity metaphor.

Metaphors are also successfully employed in everyday communication, because they encapsulate so-called image schemas that have been hypothesized to be at the core of human cognition. Image schemas are fundamental cognitive structures shaped by bodily interactions with the real world that allow basic human reasoning and understanding (Lakoff and Johnson 1980). For example, the spatial image schema MORE IS UP is the basis for understanding any type of linguistic, auditory, or graphical metaphor to communicate the magnitude or quantity of an item of interest. Examples of this image schema are in everyday language: “Her appreciation for beautiful maps keeps rising every day.” When one is pushing up a lever on a vintage analog stereo system this typically increases the volume of the played sound source. Also, in any statistical charting software, when creating a graph of analysis results, the higher the bar, the higher a point or a line in a graph, and the larger the slice of pie in a pie chart, by default, the greater the magnitude of the depicted data. The use of metaphors in user interfaces is also inspired by Gibson’s (1979) ecological theory of visual perception, which postulates that a human observer can intuitively grasp affordances in a perceived environment, resulting in appropriate human action and behavior in that environment. Hence, a perceived sound lever on the analog stereo system, or a depiction of a sound lever in a graphical user interface to play digital music, affords a user to push it up or down to change the sound volume. To be cognitively inspired, and thus intuitively understood, the appropriate mapping of more or less sound to yield the predicted result should in turn be based on the MORE IS UP image schema, with a higher/lower sound lever (source domain) for more/less sound (target domain), respectively. Similarly, on a mobile device, the spreading of two fingers, that is, the increase of distance between the thumb and the index finger, typically increases the size of objects perceived on the graphical user interface and the distances between them in the display, and also the level of graphic detail shown in the interface.

The graphical “desk in an office” interface to run a personal computer is another everyday example of the integrated application of spatial metaphors based on (spatial) image schemas and affordances that spatialize an abstract multidimensional digital data environment. This graphical user interface, initially developed by Xerox’s Palo Alto Research Center in the 1970s, and later popularized in the 1980s by Apple Inc. for its Macintosh computer, is still the dominant interface metaphor and user interface spatialization to this day. The 2-D view of a computer operating system as an office desk covered with writing tools, documents, and folders, in an office equipped with filing cabinets and a trash can (a table-top source domain), enables a user to visually create, process, store, and delete data in a digital filing system (target domain). Using spatial properties such as proximity (i.e., NEAR–FAR image schema), users typically regroup related files or applications, by putting them into a common folder (CONTAINER image schema). Consequently, hierarchies of folders can be created to simplify navigation through “data space” (VERTICAL ORIENTATION image schema). Deeper into the hierarchy, more detailed information about the data is revealed, thus relating to scale dependence in the real world (PART–WHOLE image schema). Moreover, by surmounting distance with the “drag and drop” option (SOURCE–PATH–DESTINATION image schema), we are able to perform actions within the computing environment, such as copying or deleting files. Files that have to be deleted are carried to a specific place on the
desktop, to be put into a trash can. Typically
the trash resides somewhere at the edges of
the desk, neither obstructing our working
environment, nor being too close to important
files (CENTER–PERIPHERY image schema).

Applying the distance–similarity metaphor
(i.e., NEAR–FAR image schema) to a spa-
tialized digital online library, one can couple
document locations with distance, and thus two
documents may be cross-referenced by a linear
connecting transect (i.e., SOURCE–PATH–
DESTINATION image schema). Items falling
along this transect may be characterized as being
more similar to one item (endpoint) or the other.
Documents within a given (radial) distance of a
central thematic location form clusters of related
information (i.e., CENTER–PERIPHERY).
Clusters may be nested hierarchically and build
up document ontologies. Introducing the con-
cept of scale, clusters can be explored at different
levels of detail. One level of detail provides an
overview of the entire information space. Other,
more detailed levels “zoom in” on a specific
theme or a specific document, and so on.

The power of metaphors in spatialization is
its strength, but can also be one of its greatest
weaknesses. It is important to realize that a
metaphor is only like the real thing, not the thing
itself. This means that a metaphor may include
some but not all characteristics of the mapped
source domain, and may in fact have additional
(magical) or counterintuitive properties in the
target domain. Consider a digital folder to store
files as mentioned above. The digital folder
(target domain) exhibits similar properties to
an analog manila folder (source domain) in that
“files” can be stored in it. However, the digital
folder cannot be bent, and files never fall out if
it gets too full. The digital folder also exhibits
“magical” powers in that it can hold many
hierarchically stacked folders, and potentially
store an infinite number of files (provided an
infinite amount of digital storage space is avail-
able). In fact, the manila folder might not even
be recognized as such outside of the United
States, because in many parts of the world the
cream-colored folder with tabs to write on is
not known or used at all.

Empirical research suggests that a continuous
landscape metaphor to represent document
collections is not as self-evident as information
designers seem to believe. Like visualization
designers, lay users reveal a similar naive under-
standing of geomorphological structures and
processes used as source domains, and thus have
difficulty interpreting and fully grasping the very
popular continuous terrain metaphor to depict a
news article collection (Wise et al. 1995). On the
one hand, the spatial metaphor might be taken
literally; viewers interpret depicted features as
islands, mountains, and so on, because of their
graphic appearance, even after they have been
specifically told that the display is an abstract
information space of new stories. On the other
hand, once the abstract news concept is applied
by users, it is mixed with naive conceptions
about landscape forms, because spatialization
users are not familiar with the true nature of
geomorphic processes (e.g., more information
is accumulated in the valley) and do not share
commonsense/naïve ideas about topography
suggested by spatialization designers (e.g., higher
mountain means more information) (Fabrikant,
Montello, and Mark 2010).

Spatialized views

Spatialization is related to geographic visu-
alization and geovisual analytics in the sense
that powerful interactive map–like displays and
interfaces are used to gain insights into massive
databases that are spatialized. These spatialized
SPATIALIZATION

views are typically the result of applying sophisticated data mining and exploratory multivariate data analysis approaches, including dimensionality reduction techniques (i.e., factor analysis, multidimensional scaling, principal component analysis, etc.) to large multivariate databases of interest. Such advanced multivariate statistical methods themselves are based on spatialization, that is, they reorganize and visualize massive and complex high-dimensional databases based on the distance–similarity metaphor (e.g., any type of mathematical ordination method), with the goal to uncover interesting patterns and gain an overview of latent relationships buried in the data, which in turn can be further analyzed statistically in more detail.

A spatialized view differs from ordinary data visualization and geographic visualization in that it may be treated as if it depicted geographic information. Armed with the arsenal of long-established cartographic visualization techniques, information spaces can be visualized in various ways, for example, as simple point maps, network maps, or continuous terrains. The more cognitively supportive the employed spatial metaphor, and the more perceptually salient its depiction, the more intuitive the spatialized view, and eventually the more effective the use of spatial analysis techniques for data exploration and knowledge construction. While technical developments and analytical innovations in spatialization research have continuously and rapidly advanced, theory development, empirical evaluations, and validations of design principles for spatialized views have received much less attention. Early empirical findings suggest that spatialized views should be based on sound spatial theory and principles, and adhere to cartographic design guidelines to be cognitively inspired, usable, and useful (Fabrikant and Skupin 2005). For example, empirical evidence shows that 2-D point and surface spatializations are equally effective as their 3-D equivalents. This is probably due to the fact humans rarely experience 3-D space directly (e.g., flying or diving), and thus human perception and cognition are closer to 2D than to 3D (Gibson 1979). The strength of a similarity relationship between spatialized items is most effectively depicted by connecting lines between items, and by varying line width, akin to showing the magnitudes of flows on quantitative flow maps. This is because the visual variable size is based on the MORE IS UP image schema, and the links between nodes on network maps encapsulate the SOURCE-PATH-DESTINATION image schema (Fabrikant et al. 2004). Groups of related items are shown either in clusters on point-display spatializations (Montello et al. 2003), similar to a dot density map (i.e., NEAR–FAR image schema), or as distinctly colored regions in region display spatializations (i.e., CONTAINER image schema), akin to area class or choropleth maps (Fabrikant, Montello, and Mark 2006).

Ongoing spatialization research focuses on how to effectively and efficiently depict multivariate, increasingly non-numeric, and nonspatial data stored in massive online databases, so as to provide aesthetically pleasing, perceptually salient, and cognitively supportive displays for effective and efficient sense-making and knowledge construction.

SEE ALSO: Mapping cyberspace; Space; Spatial concepts; Spatiality; Spatiotemporal analysis; User-centered design

References


Further reading


Spatiotemporal analysis

Li An
Stephen Crook
San Diego State University, USA

Despite being central to the study of many disciplines, spatiotemporal analysis has not been defined in a unified way. This encyclopedia entry aims to bring together the different uses of, and methodologies for, conducting spatiotemporal analysis in order to provide a more centralized understanding of the topic. To achieve this, we define spatiotemporal analysis as “the depiction, representation, visualization and tracking of changing location in space and time of a certain phenomenon or event of interest, which are often (yet not necessarily) connected to seeking understanding about the mechanisms behind such data of spatial locations and temporal stamps” (An et al. 2015). Spatiotemporal analysis is also referred to as spatio-temporal analysis, space time (sometimes space–time) analysis, spatial temporal (sometimes spatial-temporal) analysis, and the like. Following an introduction of its historical, intellectual context, this entry investigates the range of disciplines that contribute to spatiotemporal analysis and the specific methods employed for conducting it. To conclude the entry, the future directions of spatiotemporal analysis are presented.

Intellectual context

From early civilizations to modern times, space and time have been the two fundamental domains under which people have characterized events and phenomena of interest around them. Space and time, both abstract and often invisible, are conceptualized in a variety of ways. Space can be understood as “absolute space,” which can be characterized with a number of specific properties (Hinckfuss 1974). Under this view, space and time exist in their own right, which represents an object-independent framework. Space, along with time, is considered as a container within which all things or events of interest take place. Newton’s analysis of space and time followed this line of conceptualization. To describe the laws of motion in regard to the trajectories of moving objects in space and time, he used the languages of mathematics (e.g., geometry and calculus). In his absolute framework, the objects move and change their properties in space and time, but the framework or the container itself remains unchanged (see Peuquet 2002).

In an alternative conceptualization of the so-called relative space proposed by Leibniz, space is created differently (Cresswell 2013). In contrast to a pre-existing container in the absolute space concept, relative space is conceptualized to represent relative locations among objects. Along this line, Minkowski extended the traditional three-dimension \((x, y, z)\) geometry to include time as a fourth dimension, which forms the basis of the united, relativistic space–time concept. The relative view of space continued in its development and culminated in Einstein’s work on the theory of relativity. Research in physics and mathematics, in particular, has enriched the conceptualizations of space–time at very large and very small scales, such as those in electronics, mechanics, and cosmology. At the human and landscape scale,
conceptualizations of space–time analysis have been fertilized by the study of biology, ecology, hydrology, epidemiology, and geography (especially the subdisciplines of geographic information systems (GIS) and remote sensing).

Springing from the ideas of Kant and his fellow philosophers, earlier scholars largely held a dichotomous view of space and time, that is, that they are two separate key categories within which all activity occurs. Early geographers, such as Darwin, von Humboldt, and Ritter, focused their academic effort on depicting and understanding places, including their physical and human differences (Cresswell 2013). Under this view, space and time are not considered in tandem. For instance, regional geographer Hartshorne regarded that history aims to address changes in time, while geography addresses differences in space (Cresswell 2013). Early geographic models took a stance of either ignoring time or viewing it as a function of spatial variables such as distance or transportation costs (e.g., Von Thünen, Christaller; Cresswell 2013). Movement and transportation, which are concerned with time as much as space, can be “effectively studied in spatial terms” (Cresswell 2013). This means that movement is dictated by economics: supply and demand, least net effort, and travel costs (Cresswell 2013). Geographers, especially regional geographers, continued with this tradition until the so-called quantitative revolution in geography in the mid-twentieth century. The quantitative revolution has also witnessed the rise of spatial science within geography and the entry of geographers into the field of quantitative modeling. Since then spatiotemporal analysis has boomed in both quantity and quality (An et al. 2015), largely due to the advent of computers and advances in computing and analytical power.

Spatiotemporal analysis, as defined earlier, seeks to answer questions of both “when” and “where” (and, to some extent, “why” at the time or location) things occur. However, people use the words “when” and “where” in a variety of different ways (Couclelis 1999). The word “when” can be used in a relative sense (e.g., one event happens between two storms), or in an absolute sense that refers to time span (e.g., time duration of a storm) or clock time (e.g., at which time point a storm occurs). Similarly, space can be defined and used in varying ways. These ambiguities have led to a rich literature on the nature of space and time in physics and philosophy. In addition to the absolute/relative classification, Yuan, Nara, and Bothwell (2014) bring in another dichotomous classification of realism versus idealism, which refers to whether space–time, objects, or events are mind-independent (realism) or mind-dependent (idealism). The unifying principle underlying these conceptualizations and classifications is that space–time representation lays a foundation for subsequent spatiotemporal analysis methods as well as the corresponding results (see Yuan, Nara, and Bothwell 2014 for examples of this principle). Spatiotemporal analysis must recognize the intimate link between space and time, implying that changes of a phenomenon or object over either space or time would often inherently include a change in the other. This endorses the importance of putting space and time together to perform spatiotemporal (in contrast to spatial or temporal) analysis, and helps explain the exponential increase in the number of publications related to spatiotemporal analysis since the late 1940s (An et al. 2015).

Spatiotemporal analysis in geography

The absolute or relative conceptions of space–time structure can be found in spatiotemporal models in geography (Massey 1999). In an
SPATIOTEMPORAL ANALYSIS

absolute representation, researchers use fixed coordinate systems to represent the study site and events of interest, mark the changes in the associated variables, and explain or predict the pattern of change over time. Traditional geographic models represent the physical environment of interest in a two-dimensional space of grid or vector, without specific representation of time as a key dimension of concern. Largely ignoring time, such models are essentially spatial models. As time has moved on, the possibility of a space–time conceptualization has been explored by geographers. Early spatiotemporal analysis in geographic models often resorts to a GIS. Traditional GIS have elegantly represented space, but not so time (Peuquet and Duan 1995). The mainstream space–time representation in traditional GIS is through the snapshot model, where a spatiotemporally continuous world is shown at limited snapshots in time (Peuquet and Duan 1995). To address some difficulties faced by the snapshot model, the event-based spatiotemporal data model (ESTDM) by Peuquet and Duan (1995) allows organization of space–time data by time. Other data models, including the space–time composites model, the spatiotemporal object model, and the three-domain model, offer different tradeoffs between representing both space and time (An and Brown 2008).

On the other hand, a relatively new, object-oriented approach is arising in the geographic modeling and analysis arena. This approach leans more toward the relative representation of space and time. Under this approach, features on the Earth’s surface (e.g., land parcels, people, households) are conceptualized and represented as objects that are relatively independent entities, which may change their locational and nonlocational attributes over time – some of them may even have a certain level of intelligence and autonomy, make decisions, and/or adapt to the changing environment or the changes made by other entities (see the section for agent-based models below).

Spatiotemporal analysis techniques

This section sets out to give an overview of spatiotemporal analysis methods that have been used in geography, environmental sciences, and related disciplines. Each method will have a brief discussion of how it represents time and space, what type of data is required, what type of output it may yield, and whether it is primarily focused on prediction or explanation. These methods have been selected because they share several important features: they allow representation, visualization, or quantification of a system over space and through time in a dynamic way; they help in identifying mechanisms for observed changes in spatial patterns over time; and/or they can predict changes in temporal patterns by analyzing spatial data over time steps.

Time geography

Unlike early spatial science with its focus on place-based aggregations and generalizations, time geography has arisen attempting to integrate individual trajectories of human movements in both space and time (Hägerstrånd 1970). As a pioneer, Hägerstrånd (1970) emphasized the need “to have not only space coordinates but also time coordinates” because people exist both in a specific space and at a particular point in time. In his seminal work on space–time life paths, a horizontal plane represents position in space and a perpendicular axis represents time. Space–time prisms, often derived from space–time life paths, are a powerful tool to show movement patterns within an individual’s world, which are subject to physical and physiological constraints and also
SPATIOTEMPORAL ANALYSIS

conform to public and personal decision-making (Hägerstrand 1970).

Recent years have witnessed rapid improvement in geospatial information technology, paving the way for quickly collecting individual, georeferenced human activity data in large amounts or over large spatial extents (Kwan 2004). In parallel with this increasing data availability, a number of models and techniques are being developed to capture, represent, and analyze such data, including the ones focusing on the exploration, visualization, and generalization of large space–time trajectory datasets in the GIS software environment (Kwan 2004). Such models and rich datasets have also opened new perspectives and opportunities in geocomputation; for example, real-world accessibility measures can be developed and obtained by calculating the maximum travel distance of individuals subject to multiple constraints (Kwan 2004; Lenntorp 1976). Among these models and techniques, statistical methods have played an essential role in time geography, allowing quantitative analysis and comparison of the space–time trajectories of different individuals or groups. The statistical methods allow calculation of a number of different measures, including the Hausdorff distance, Fréchet distance, dynamic time warping algorithm, and longest common sequence algorithm. Also worthy of mention in time geography is space–time path clustering analysis, which classifies individuals sharing similar space–time paths into groups.

There is no doubt that the time geographic approach is particularly important in space–time analysis, because for individuals under investigation, both spatial coordinates and temporal coordinates are tracked down continuously and given equal weights in later data analysis. Nonetheless, it is a big challenge for time geographic researchers to address patterns and processes at multiple spatial and temporal scales. Currently, most activities captured in time geography research operate either at a small spatial scale or over short time periods, or both. Seldom have movements over larger distances and larger time scales, without sacrificing spatial or temporal resolution of data collection and analysis, been included and characterized well in this type of continuous tracking (Meentemeyer 1989).

Analysis of time-series spatial data

One category of spatiotemporal analysis methodologies consists of analysis of time-series spatial data using a number of spatial statistical and analytical techniques. Broadly, this category includes time-based descriptive statistics, statistics for clustering and dispersion, and establishment of correlations or causal relationships through regression analysis. Change through time is assessed either qualitatively or quantitatively, through the comparison of maps and graphs at different time steps, and by graphical output of statistical change over time. Below, several increasingly recognized methods are introduced, which are powerful at analyzing time-series spatial data and potentially unveil the mechanism behind such data – for nearly all these methods, spatiotemporal patterns are represented by the so-called snapshot data model, where space at one time is represented in a certain GIS layer (often raster unless otherwise specified) and time is represented at (and constrained by) the associated data collection frequency.

Exploratory space–time data analysis

This category of spatiotemporal analysis attempts to develop novel measures (including graphs) to show trends in both space and time. A number of space–time statistical techniques in this general category are included in commercial GIS. For
instance, ArcGIS has enabled the visualization of events in three dimensions, with the time dimension being displayed vertically, which is conceptually similar to the space–time paths in time geography. The Crimestat software has tests for spatiotemporal clustering, diffusion, and interaction (Levine 2004). The Space–Time Analysis of Regional Systems (STARS) package is designed to perform exploratory analysis of spatial data with numerous time points (Rey and Janikas 2006). STARS offers several important functions, including qualitatively displaying spatial data patterns over time, and calculating descriptive statistics such as global and local Moran’s I, and the Gini coefficient (Rey and Janikas 2006). Though spatial analytical and statistical software tools are not elaborated here, it is worth mentioning that software development is one of the rapidly improving areas for spatiotemporal analysis.

Identifying spatiotemporal clusters, for example epidemiologists identifying clusters of disease outbreaks, stands out as a very important type of applications. Many useful indices and/or tests have thus been developed for different purposes: (i) the Barton and David test, which aims to find if spatial patterns of events vary by temporal cluster; (ii) the Knox test, which focuses on finding whether events in one time–space window would differ from the expected amount in the same window given the total number and time range of all events; (iii) the Mantel Index, which helps find correlation between distance and time interval (Levine 2004); and (iv) many other indices, including frequency, duration, intensity of events, spatiotemporal covariance structures, and space–time hotspot indices (local indicators of spatial autocorrelation or LISA, bivariate LISA, Getis-Ord Gi*), which continue to be developed to assess the spatiotemporal patterns of the phenomenon or event of interest.

In addition to finding clusters, assessing changes in an individual feature’s locations in a GIS environment is another application domain of spatiotemporal analysis. Worthy of mention is the Spatio-temporal Moving Average and Correlated Walk Analysis by Levine (2004), which facilitates tracking of the mean location of a moving event/feature and thus helps prediction of its location in the future. All these methods conceptualize a spatially continuous world at limited snapshots in time (Peuquet and Duan 1995), and the snapshot GIS data model (often raster layers over multiple time points) underlies the corresponding data collection and analysis.

Spatial panel data analysis

Spatial panel data refer to time-series observations of a number of spatial units over time (Elhorst 2010), and often raster layers over multiple time points are data input. Spatial panel data analysis consists of two types of models. The first type includes dynamical models that predict the dependent variable at a given time by its value at the prior time and a set of independent variables using the so-called difference equation models. Though conceptually such models are ideal for spatiotemporal analysis, they are considered less compelling and more complex and thus not commonly employed in literature. The second type is related to two kinds of multilevel models (MLMs), that is, fixed effects and random effects models, where the multi-time observations for each individual unit (e.g., land parcels) are lower-level (level-1) data, and the individual units are higher-level (level-2) data. This data structure is the same as that in MLMs (see the section for MLMs below). The model for this type of panel data analysis is similar to an ordinary least squares (OLS) regression model in many aspects (Elhorst 2001; Petersen 2009).
SPATIOTEMPORAL ANALYSIS

What distinguishes the above spatial panel data analysis from regular OLS regression models is a spatial effect term that deals with the impact of neighboring spatial units on the unit of interest. Depending on how the space–time autocorrelation in data is to be addressed, researchers may use spatial lag models (when the dependent variable of interest is autocorrelated) or spatial error models (when residuals from OLS models are autocorrelated; Anselin, Le Gallo, and Jayet 2008; Elhorst 2010). Then essentially spatial panel data analysis has to extend the fixed (or random) effects models with a spatially lagged dependent variable or with a spatial error term. Therefore the challenges related to MLMs, for example the need for large and balanced samples and the inability to incorporate endogenous predictor variables, are also extant (see the section for MLMs below). Also, model estimation may be biased by the spatial dependence among observations, and this complication may need special attention, according to Anselin, Le Gallo, and Jayet (2008).

Markov chain modeling

This type of model aims to represent changing temporal dynamics and spatial patterns of spatial units (e.g., individual cells/pixels). With input being often raster layers over multiple (at least two) time points, Markov models often focus on predicting or projecting future spatiotemporal patterns (e.g., in terms of raster maps). All units have spatial coordinates, and their attributes can change over time. In a Markov process, the landscape type at a given location and certain time depends only on its previous type and a transition probability. Transition probabilities, often presented in a matrix form, are obtained through assessing historic conversions between transition types.

Markov chain models are logically simple and useful for exploring patterns of spatiotemporal changes within a relatively short time span. Nonetheless, they have a number of drawbacks that deserve the attention of spatiotemporal modelers (An and Brown 2008). Primarily, the assumptions of spatial independence and stationarity (in time and space) may not hold true in many applications (NRC 2014). In addition, Markov models focus explicitly on prediction, and provide little insight into causality or low-level processes that generate the observed spatiotemporal patterns (NRC 2014). In many studies, Markov chain models are made more useful when they are integrated with other modeling or analysis techniques (see the section for cellular automaton).

Bayesian spatiotemporal models

Bayesian statistical models have been adapted for use in spatiotemporal analysis. With input being often raster layers over multiple time points, Bayesian models often focus on predicting future spatiotemporal patterns (e.g., in terms of raster maps) or explaining observed patterns. In this type of model, a spatial dependent variable (with both spatial and temporal stamps) is described as a function of both time and a number of parameters (that are related to the specific site). According to prior knowledge (e.g., physical laws) about the phenomena or events of interest, a priori statistical distributions are used in order to impose constraints on a model that may have too many parameters, for example through a conditional autoregressive function. This results in a relatively small number of site-independent hyperparameters. Posterior distributions of site-specific parameters are used to update the parameters based on empirical spatiotemporal data, specifically by maximizing the probability of observing the empirical data.
through modifying parameter values. In determining these probabilities, researchers could use techniques such as the Markov Chain Monte Carlo method to generate samples for these posterior distributions. Weaknesses related to Bayesian spatiotemporal models include the fact that subjectivity might come into play when choosing either the posterior distribution or the a priori distribution, and also that modelers may have to assume some untested parametric distributions between space, time, the dependent variable, and the independent variables.

Survival analysis

Survival analysis is a methodology that successfully integrates both time and space in order to understand the contribution of different mechanisms to observed spatiotemporal patterns. While related to, and comparable to, more traditional static spatial analysis techniques such as logistic regression, it represents a more dynamic, integrated method for spatiotemporal analysis (An and Brown 2008; Wang et al. 2013). With input being raster or vector layers over multiple time points, survival analysis models are very useful for either prediction (e.g., in terms of raster maps) or explanation purposes. The technique has been adapted to land-change studies from a diverse range of earlier uses, including public health and demography (An and Brown 2008). While the name of the methodology is derived from investigations into mortality, it can easily describe the survival of a particular land unit, that is, not being changed to an unintended land use or land cover. Indeed, in recent years, a multitude of studies have used survival analysis to find drivers of land change (An and Brown 2008; Wang et al. 2013).

The hazard function is central to survival analysis. Based on theoretical assumptions or empirical event frequency and timing, the hazard function can be calculated as a measure of the risk that a change will occur at a given space and time point. Related to survival probabilities, the hazard function differs in its ability to either increase or decrease depending on the influence of explanatory variables over time. Survival analysis uses information at all of the time steps to calculate the hazards for each individual pixel over time. The hazard function is regressed against a set of predictor variables, including those time-dependent ones. Because it accounts for the dynamic values of these variables at different time steps, a key strength of survival analysis is that it allows varying contributions of each time-dependent predictor variable to the hazard over time. Survival analysis is also suited to accounting for censored data: if an event happens outside of or in between time steps in a set of time-series data, it can still be included in analysis and contribute to estimating the related model parameters.

Despite its strengths in incorporating time into statistical studies of change, survival analysis has a number of drawbacks. First, it is weak in handling continuous changes in a certain dependent variable because essentially it deals with events or qualitative changes that happen at specific time points. Second, and related to the first drawback, it depends on the use of subjective thresholds to specify whether change has occurred if the dependent variable is continuous. Percent of a land cover cannot be included in the corresponding dependent variable, which must be declared as one type or the other based on a cutoff value. Third, the hazards of different spatial units sometimes may require certain untestable assumptions to be made; for example, the hazards are constant within each period in the piecewise exponential model.
Multilevel models and latent trajectory models

Multilevel models (MLMs) and latent trajectory models (LTMs) are two types of relatively new models that contribute to spatiotemporal analysis. MLMs use data at multiple hierarchical levels, including individuals who are members of a group or neighborhood, in order to account for the clustering effect that may exist within individuals of the same higher-level grouping (Browne and Rasbash 2004). These clustering effects violate the assumption of independence in standard OLS regression, meaning that conclusions of statistical significance may be drawn incorrectly at the individual level. In various demonstrated case studies in health and education, outcomes at the individual level were found to no longer be significant when clustering due to higher-level group membership, such as classrooms or hospitals, was accounted for. Functionally, MLM differs from OLS regression by allowing variability in both the intercept and the coefficients of the model. For each intercept or coefficient, there is one random term that is allowed to change from entity to entity, and/or a fixed term that describes the influence of the higher-level grouping on each individual within the same group.

MLMs are used in spatiotemporal analysis because of their ability to investigate time-series spatial data by nesting multiple time measurements within each individual unit. In this case, time measurements would be the level-1 units, individual observations (which could be pixels) would be level-2 units, the neighborhood or other higher-level group would be level-3 units, and so on (Subramanian 2010). By structuring spatiotemporal analysis in this way, causal mechanisms that are operating at different levels over time can be revealed. Challenges in using the technique include the need for sufficiently large populations of higher-level groups in order to draw a sample of adequate size, as well as the related need for significant computing power to conduct analysis of such large datasets. In addition, MLM models do not allow endogeneity within the model, that is, they do not allow parameters of the model to function as predictors for other variables (Preacher et al. 2008).

LTMs are related to MLMs, and in some cases result in equivalent models for spatiotemporal analysis. This technique, however, is more explicitly aimed at analyzing longitudinal data in order to find the impact of time and a number of independent variables on a dependent variable (Guo and Hipp 2004). The input data of both LTMs and MLMs could be either vector or raster maps at multiple times, and both can be used to predict or explain spatiotemporal patterns. Using patterns of change (such as quadratic or linear, but not necessarily monotonous increase or decrease) in the response variable \( y \) as latent variables, LTMs are able to model change patterns in \( y \) as latent variables, and estimates complex causal relationships or plausible pathways among these change patterns and a set of independent variables (Preacher et al. 2008). Weaknesses of LTMs include the assumption that the dependent variable follows a mathematical function in time, when the reality may be more complex, and also the assumption that LTMs are limited in integrating hierarchical data in the way that MLMs do. By combining LTMs and MLMs into hybrid multilevel latent trajectory models, the strengths of both can be accentuated and the weaknesses minimized.

Spatiotemporal simulation techniques

The previous sections focus on statistical models that establish empirical relationships between drivers and outcomes in order to derive understanding about processes from observed patterns
Another category of spatiotemporal models is that of simulation-based ones that focus more explicitly on the processes through which changes develop. These rule-based models, including cellular automaton (CA) and agent-based modeling (ABM), concentrate on how systems function, and how lower-level processes result in emerging higher-level patterns.

**Cellular automaton**

CA models portray the spatial variable of interest in a contiguous array of cells (or pixels), and the cells can change between a certain number of predetermined states (e.g., land-use and land-cover types) as time passes (Parker et al. 2003). CA rules and parameters are set on a time step basis such that each snapshot outcome represents the system's status at a certain step. CA models can be used for either prediction or explanation purposes. In these models, change is dictated by transition rules, under which the change of state of a certain cell of interest is often dependent on the state of neighboring cells. For instance, one rule might state that an undeveloped cell that has four developed cells in its neighborhood may remain in its undeveloped state due to "crowding." Due to these simple rules, the states of all cells may change over each successive time step, resulting in a progression of new patterns. In empirical studies, change rules can become more complex. For example, one study assessing patterns of urban growth over time in the Bay Area of California formulated transition rules based on slope, roads, and amount of nearby development (Clarke, Hoppen, and Gaydos 1997).

CA models are often combined with other techniques such as Markov models, overviewed earlier. Markov analysis alone projects future change based on transition probabilities calculated from time-series data. Integrating CA with Markov models allows neighborhood interactions to be integrated so that the amount of change between any two types is derived from a Markov process, while change locations are prioritized based on CA rules. Together, they better address spatial dependence while predicting changes over time. Despite their utility in projecting change and investigating emergence, CA models have a number of drawbacks. Specifically, they are weak in making links between real-world, human decision-making and conversion rules, and they have a poor ability to project over long time periods (due partly to an inability to include feedbacks that operate across spatial, temporal, or organizational scales; NRC 2014).

**Agent-based models**

Agent-based models (ABMs), also called multi-agent systems and individual-based models, focus on lower-level processes that generate larger-scale patterns (Parker et al. 2003). Unlike a CA model that only deals with fixed pixels, ABMs use agents that are allowed to move in time and space. Its input often includes (i) at least one layer of raster data, ArcGIS shapefile, or ASCII file (increasingly vector data as well) that contains spatial information; (ii) data for agent attributes (e.g., an ASCII or Excel spreadsheet for agent information); and (iii) rules and parameters that make the actions or processes play out. ABM models can be used for either prediction or explanation purposes. The output of ABMs can be different graphs (e.g., population dynamics over time), maps (e.g., shapefiles, ASCII files), or numbers (e.g., An et al. 2005). Each agent contains a number of the attributes that characterize it, and behaviors that can cause changes to the agent itself, to other agents, and
Spatiotemporal Analysis

to the environment as time passes. As in a CA model, ABMs function based on a set of decision rules. As these decision rules play out over time steps and agents interact with other agents and the environment, they update the attributes of themselves, other agents, and the environment. Consequently, the system represented by the ABM will evolve over time.

When applied to human-environment systems, urban systems, or other geographical systems, agents can include individuals, households, neighborhoods, or other hierarchical groupings. The spatial domain in which the agents operate is often made up of GIS data layers that can include pixel-based or sometimes vector-based representations of the landscape. The representation of time is via time steps, or ticks. At each tick, which coincides with a real-world time interval such as a month or year, agents make decisions regarding themselves, other agents, and the environment. In a human-environment ABM, for example, these decisions can relate to things such as marriage, childbirth, the decision to collect fuelwood in a forest, or the decision to create a new household where there had previously been forest (An et al. 2005).

ABMs are especially suited for modeling complex systems, allowing the investigation of emergent properties, due to the fact that each agent has autonomy, intelligence, the ability to communicate and interact with other agents and the environment, and the ability to make informed decisions regarding the environment (Parker et al. 2003). These features allow ABMs to characterize further features of complexity such as feedbacks, nonlinearity, thresholds, time lags, and resilience. In doing this, they demonstrate strengths in integrating heterogeneous, multiscale data, including a diverse array of agent types at different hierarchical scales. Despite these strengths, challenges in ABM implementation include steep learning curves for new modelers, high data requirements, challenging verification methodologies due to path dependence, equifinality, and multifinality (Brown et al. 2005; NRC 2014).

Verification, validation, and other aspects in spatiotemporal analysis

All of the spatiotemporal methods above purport to have utility in understanding, describing, and/or predicting spatiotemporal patterns of interest. In order to show this utility, the degree to which models demonstrate agreement among their theoretical frameworks, model predictions, and real-world observations needs to be established. Verification refers to the task of making sure that the construction of the model (usually in computer code) reflects the design intentions of the modeler, and is achieved partly through debugging (NRC 2014) and a set of other techniques (An et al. 2005). Validation can jointly refer to “structural validation,” or demonstrating that model processes largely reflect or approximate real-world processes, and “outcome validation,” in which model outputs are compared to real-world data of the same outcome through, for example, map comparison if the outcomes are maps (NRC 2014). One of the most-used methods is to compile an error matrix and calculate a Kappa statistic based on a sample of points, tabulating cases of agreement and disagreement between types. However, other methods for outcome validation have arisen that allow more sophisticated comparison. These include fuzzy set approaches, multiple resolution map comparison, and the variant–invariant method (Brown et al. 2005; Pontius, Peethambaram, and Castella 2011). These methods are crucial in examining outcomes of spatiotemporal analysis.
Conclusion

The aforementioned models represent some, though not all, of the methodologies used in spatiotemporal analysis in geography, environmental studies, and related disciplines. Other techniques, including some geovisualization and land-change models, contain overlap with these spatiotemporal analysis techniques but are less explicitly focused on looking at changes over both space and time. Despite the strengths of the demonstrated methods in appropriately dealing with spatiotemporal systems, there are a number of future research directions and challenges that the discipline faces.

In traditional spatiotemporal analysis, there has always been a dominant focus on space. Most analysis techniques are built around GIS software packages that give a good representation of space, at the expense of offering an adequate representation of time. Indeed, in many applications of spatiotemporal analysis in geography, the decision regarding the number and resolution of time steps to include is based on convenience or data availability. While space can be represented in a continuous way, time steps are broken up and included without much thought as to what the most appropriate time span and/or temporal resolution might be (An and Brown 2008; Wang et al. 2013), which might cause important patterns and processes to be overlooked and thereby generate biased or even misleading conclusions.

Spatiotemporal analysis, with limited capability to handle temporal variability, has a bunch of other implications. First, changes in state are often limited to low temporal accuracy, as it can only be said that a change happens before or after a time step, or between two time steps. Survival analysis models offer one potential solution to this problem by appropriately dealing with censored data. Second, the common methods of storing data based on consecutive time steps may be inefficient in data processing and data storage when compared to time-based storage techniques that are organized by the times an event happens (Peuquet and Duan 1995). Finally, it has been noted that spatiotemporal models are generally conceptualized using an absolute conception of space and time. Relative conceptions of space and time, focusing on where and when events happen in relation to other events rather than in relation to fixed space time stamps (latitude/longitude, clock time), should be further developed. Fuzzy land-change methods are one methodology moving in a direction toward relative space–time conception and deserve more attention in the future (An et al. 2015).

The techniques discussed in this entry offer an introduction to the many methods used for spatiotemporal analysis in geography, environmental studies, and related disciplines. An important note is that many of these models can be combined to create hybrid models (NRC 2014). Some, such as CA/Markov models, or the use of statistical methods in parameterizing an ABM, have been demonstrated here, though other hybrid combinations exist. When undertaking spatiotemporal analysis, the selection of a proper model is key to characterizing the desired pattern or processes. Understanding the models listed above, each with their advantages and drawbacks, provides a starting point for conducting spatiotemporal analysis.

SEE ALSO: Agent-based modeling; Geographic information system; Spatial analysis; Time geography and space–time prism

References

SPATIOTEMPORAL ANALYSIS


Peuquet, Donna J., and Niu Duan. 1995. “An Event-Based Spatiotemporal Data Model (ESTDM) for


Spiritual geographies

Avril Maddrell
University of the West of England, UK

Spirituality concerns beliefs and associated practices that pertain to being in relation to self, others, the material world and, crucially, what is believed to be beyond that material world, whether that be a deity, a number of deities, angels, fairies, spirits, natural energies, or the occult. The “spiritual” has been defined as part of the non- or more-than-representational and posthuman arena: “that part of the virtual in which faith forms a significant part of the move beyond rationality and of the possibility of other-worldly dispositions” (Dewsbury and Cloke 2009, 696). The spiritual is part of the virtual or more-than-human, which informs lived epistemologies and ontologies, addressing what is beyond human limits. Spiritualities play an important role in the conceptual understanding of many individual and community relations to identity, embodiment, consumption, community, politics, material, and visual culture.

Common preconceptions represent spirituality as the opposite of religion, that is, religion is an institutionalized approach to regulated beliefs and practices, whereas spirituality relates to the beliefs and practices held by those outside the confines of a religious regimen, leading to descriptors such as “spirituality without religion,” “do-it-yourself,” or New Age spirituality (MacKian 2012). This dichotomy is an over-simplification. Rather, spirituality can coincide with, but can also operate independently of, religion. Like love, the spiritual is an umbrella concept, encompassing the inner prayer life of a devout Buddhist or Roman Catholic nun, a therapeutic reiki massage, or energy-seeking at Sedona or Stonehenge.

How then does this wide-ranging but significant set of phenomena relate to geography? On a global scale, spirituality is often represented as an essentialist Oriental/Occidental dichotomy of “natural” Eastern spirituality, versus spiritual paucity in the West. While the effects of the Enlightenment, decline in institutionalized religious practice in Western societies and other social changes have marginalized spirituality, spiritual beliefs and practices persist. Spiritual beliefs and associated practices have contributed to shaping historical and contemporary designations of sacred places, nodes, and routes in palimpsest landscapes across the globe. For example, India is etched with an ancient spiritual geography of Hindu shrines linked by pilgrimage paths, sacred rivers and crossing points (tirtha), which are deemed to be spiritual resources. Likewise, some perceive Europe – and indeed the world – as a mesh of leylines. Such sacred sites are frequently multivalent and can fluctuate in meaning and status; they are also performative, reflecting spiritual beliefs and practices which may be woven into everyday life and/or mark extraordinary moments in life and experience, such as birth, marriage, and death.

Hence spirituality can be practiced, marked, or can leave traces on a wide range of geographical landscapes and arenas: the body, the home, at places of religious practice, or sites associated with the life cycle. Recent themes emerging within geographical work on spiritualities have drawn particularly on the agenda and methods of cultural and, to a lesser extent,
SPIRITUAL GEOGRAPHIES

social geographies, and have built on work on historical geographies of belief, geographies of religion, and studies of sacred spaces to consider wider contemporary spiritual beliefs and practices marked and performed within urban and rural landscapes, therapeutic environments, and domestic arenas. This can be seen in studies of postsecular cities and suburbia, young people’s spiritualities, and spaces of wellbeing and retreat. Feminist and queer geographies have prompted engagement with the intersection of embodiment, gender, sexuality, and spiritualities; and the non- or more-than-representational turn within cultural geographies has underscored attention to performative spiritual practices such as meditation at domestic shrines, rituals at pagan sites and memorials as well as through yoga and the spiritual mobilities of many forms of pilgrimage. There is enormous scope for further geographical engagement with the spiritual, which will enrich understanding of particular places, practices, and people’s relation to space and the environment, such as migrants’ and older people’s spiritualities, spiritualities and art, and spirituality-informed consumption.

Qualitative methodologies appear best suited to studying experiential and embodied spiritual geographies but emerging work suggests that quantitative techniques, such as geographic information systems (GIS), effectively capture complex spiritual networks and are potentially powerful analytical tools for examining geographies associated with particular forms and practices of spirituality and their complex interrelation with other factors such as gender, ethnicity, socioeconomic class, politics, migration, and so on.

Anne Buttimer (1993, 3) has argued that “For each facet of humanness – rationality or irrationality, faith, emotion, artistic genius, or political prowess – there is a geography. For each geographical interpretation of the Earth, there are implicit assumptions about the meaning of humanness.” Spiritual geographies represent a specialist field in their own right, but attentiveness to the significance of the spiritual in the lives of many would enrich many wider geographical studies beyond the subdiscipline.

SEE ALSO: Cultural studies; Place; Religion

References


Further Reading

Sports

Paul Gilchrist
Kath Browne
University of Brighton, UK

Sports geographies deal with the social and spatial dynamics and impacts of sport. Research and scholarship have engaged with many established areas of geographical investigation – landscape, place, environment, nature, colonialism and imperialism, nationalism, globalization – and have intersected with and catalyzed the study of sport in other fields. Academic interest in sport is of course not confined to geography. In addition, the significance of space and place in understanding sport is recognized by other disciplines across the social sciences and the humanities, though many scholars working on sport in these areas and from within mainstream geography have attested to a concern over the marginality of the subject, where there is a tendency to dismiss or overlook sport as trivial or not central to the core concerns of a parent discipline. As a result, analyses of the spatialities of sport are to be found predominantly in the more developed multidisciplinary field of “sport studies.” There are few in geography departments who define themselves as “sports geographers.”

Histories of sports geographies

The origins of sports geographies reside in regional cartographic analysis of geographical variations in American sport undertaken by North American scholars in the 1960s and 1970s. There has been a long-standing interest in the geographic diffusion of sport based on factors such as distance, population density, transport infrastructure, the presence of cultural brokers, and forms of social capital at work in the development of sporting systems. Historians and political scientists have addressed spatial themes, for instance, the suppression of street football and other premodern games in industrializing towns and cities and the emergence of sporting venues and facilities through commercial entrepreneurship and working-class self-organization. Colonial and imperial histories of sport have added a wider scope for understanding the spatial trajectories of modern sport during a phase of internationalization throughout the nineteenth and twentieth centuries, directing scholars to investigate the coercive and collaborative forms of power at work between center and periphery in the transmission of and encounter with dominant cultural forms.

A major geographical theme has been to trace the diffusion of modern sport from its origins in folk games to nationally organized leagues and competitions toward its US-influenced globalized form and to sport mega-events played out in super-stadia to television audiences numbering in the billions. Soccer is viewed as the world’s most popular sport, a form of popular culture that transcends provincial boundaries. This claim may hold true for the rest of the world, but not for the United States where its exceptionalism to the spread of the global game has been noted by many authors.

Mapping of common or divergent features of sporting forms and practices across time and space has been organized around the concepts of “diffusion” and “adaptation.” Tomlinson and
Young (2011) note a narrative orthodoxy across sport studies that credits England as the cradle of modern sport able to successfully export sports such as football, rugby (both codes), and cricket through the military, trading, and ecclesiastical arms of the British Empire. Globalization is seen as crucial to understanding contemporary sporting landscapes, including sports stadia (see below). Although sports are rarely globalized in the sense of being in every country in the world in equal measure, the hegemony and prestige of particular sporting events, such as the Olympics and the soccer World Cup, create and are created through globalizing factors. These include transnational migrations, the movement of capital, and, perhaps most importantly for vast sporting audiences, globalized media. The media provide a key way in which sport travels internationally and sporting fixtures can be globally consumed. The global media–sports–culture complex (Rowe 1999) operates to globalize sporting practices, open new markets for sport and consumption, as well as often to reiterate core–periphery relations through the scouting for sport talent in areas where sport participation and achievement are not as lucrative. In other words, the globalization of sport can reiterate and reproduce global inequalities.

A homogenizing globalized view of sport is increasingly challenged by sports geographers and historians who have highlighted the importance of local conditions to processes of transmission, reception, and adoption, pinpointing more complex cultural flows to the sharing of new sporting practices and technical innovations. In this way globalized trends are seen to be created locally and their manifestations related to local contexts, what is termed glocalization. Diffusion narratives can also work for particular sports, in specific contexts, but do not account for sports that challenge the move from Global North to Global South, colonized to colonizer, for example, the popularity of Eastern martial arts in Europe and North America.

**Sportscapes and cities**

British sports geographer John Bale is a pioneer in the field. Bale’s work has been core to sports geographies, and he remains a key figure in a field that has few who would define themselves as “sports geographers.” He has coined the concept of “sportscapes” to identify the material and bounded nature of modern sport. Drawing upon Arjun Appadurai’s notion of “-scapes” to show the nature of global cultural flows, Bale delineates the spatial ordering and reproduction of sites of sport. Unlike its premodern forms where participation takes place in unconfined and unmarked spaces in an existing landscape – the traditional spaces for folk games and contests at fairs, feast days, and holidays – modern sport is “refined and ordered in both time and space.” “Sportscapes” are, Bale argues, increasingly “monocultural,” focused solely on sports rather than “multi-functional landscapes” (Bale 2003, 131). While “sportscapes” do not account for the whole of potential sites of sport the concept has alerted geographers to key and enduring trends which emanated in the nineteenth century toward the construction of sites of sport according to the modes of a scientific-rational orientation essential to the production of the controlled conditions necessary for “achievement sport.”

Sport is key to understanding historical and contemporary built environments. The concept of “sportscapes” is useful in helping to explain built environments. Evidence for sports practice can be seen in places far and wide across many cultures and civilizations, from the gladiatorial contests staged in arenas throughout the Roman Empire to the ballgames played in specially constructed courts across Mesoamerica to the
shooting galleries of Edo-period Japan as Samurai warriors showed off their skill with the bow. Archaeologists and historians have recorded the changing spatial parameters of sports and the variation of practices, forms, and institutions.

The study of new sporting practices, particularly sports classed as “extreme,” “action,” or “lifestyle” sports (Wheaton 2013) has contributed to a deeper consideration of urban dynamics and an understanding of domination and contestation. Skateboarding and parkour are two well-studied sports for understanding spatial appropriation and redefinition, with many studies heavily indebted to the spatial theories of Henri Lefebvre in terms of understanding the spatial tactics mobilized by participants in opposition to the normative regulation of the built environment and the omnipresence of capitalist relations. Parkour has been conceptualized as a critical spatial practice and urban intervention that ascribes new meanings to places as parkour practitioners – traceurs – extend mobilities in the city through viewing everyday material objects – benches, railing, fences, lampposts – as ripe for tactile engagements that generate new place-meanings. The poststructuralist writings of Gilles Deleuze and Félix Guattari have been employed to theorize the juxtaposition between the fluid movement of the traceur and the negotiation of “striated” urban space. While the conceptualization of parkour as a form of subversive urban politics can be overstressed, sport sociologists have claimed that in their hyper-awareness of the environment and urban landscape, participants exercise “post-sport” subjectivities defined by morally oriented, community-centered, green and anarchic physical cultural practices which can generate an aesthetic-spiritual realization of the self. The valorization of play/playful within a redefinition of sport as a physical cultural practice opens up possibilities for the study of other practices and forms (e.g., urban exploration) which contain aspects of the sporting in their standards of challenge, need for physical competency, and embrace of personal risk.

Stadia

The use of sport facilities for urban regeneration emerged as a specific public policy in North America in the 1970s and continues to feature in location decisions of stadia linked to sport franchising business models. The hosting of sport mega-events such as the Olympic Games and FIFA World Cup has led to the construction of new stadia and arenas for domestic spectators and international visitors keen to observe sporting action. Human geographers have documented the relationship between sport-led regeneration and urban and regional planning, which have provided empirically rich accounts of the rationales for the construction of stadia linked to strategies for local economic development, gentrifications, city image creation, and inward investment, and their subsequent legacies and impacts, including negative externalities. Bale (1995) explained that stadia were as much economic entities, appearing on business pages, as they were sporting venues. They are seen as boosting cities, markers of successful urbanities as well as “urban machismo and vibrancy.” Moreover, stadia attract inward investment and development because they are understood as providing “a much-wanted – or, at least, much perceived – multifunctional leisure facility” (Bale 1995, 13).

The human experience and meanings that accrue to stadia have been widely researched, drawing on Yi-Fu Tuan’s (1974) concept of “topophilia” to articulate the “sense of place” and affective qualities of place-attachment. Such work has challenged the attribution of the modern sport stadia as “placeless” entities.
The argument of placelessness suggested that sporting spaces are globalized to such an extent that they are no longer locally specific, and that sporting stadia have become homogenous in form and structure. However, Tuan’s concept of “topophilia” has been used to question stadia as placeless, drawing attention to the ways in which they are the site of multisensory experiences and collective imaginaries shared by consumers and spectators, rendering the space of the stadium meaningful. The importance of fan communities in constructing a “sense of place” through cultural expression and festive celebration and following stadium relocation has widened the appreciation of personal and community investment in a sporting space and the forms of agency involved in collectively producing a sense of belonging.

Everyday spaces of sports

Cultural geographers, sociologists, and political scientists have extensively illustrated how sports foster senses of collective belonging in terms of civic pride and regional and national identities. The predictability of the sporting calendar produces the recipe for expressions of what Billig (1995) terms “banal nationalism” as the identification with a sport team spills over into the everyday spaces of the home, street, shopping center, café, and bar. Sport then can be key to the performance of contemporary national identities and uniting diverse people and places in expressions of nationalistic pride. This unity creates imagined communities of people who may never meet, but share national identities and expressions through how they “support” “their” team.

Sportive nationalism is thus a mechanism for geographical, class, and ethnic identification as well as serving assimilationist policies successfully in places like the United States through baseball and Canada where Quebecois nationalism finds an expression in ice hockey. Sport has high symbolic capital in the campaigns for separatist nationalist movements (e.g., Barcelona/Catalonia), and entering a sports team to the Olympic Games as representative of the nation sends a strong symbolic message about the territorial legitimacy of the nation-state.

Bodies, difference, and emotions

Despite assertions of the recent inclusion of bodies into social theory, sporting bodies have long been explored as core to thinking through sporting places/spaces. The German sociologist and historian Henning Eichberg has been influential to sport geographers in tracing the evolution of the sporting landscape. Eichberg (1998) proposed a trialectic model of “body culture” to identify the spatial parameters of sport. Within this model sport is practiced under welfare-oriented guises concerned with health, fitness, and the recreation of the body, typically practiced in purpose-built halls and leisure centers as facilities for mass sport. A playful body culture accounts for a spatially less defined encounter with nature where the spontaneous and the sensual are not so strictly constrained. “Achievement sport”, the third in the model, covers the time/space constraints of rule-bound and record-oriented elite sport, with super-stadia the pinnacle for the display of sporting excellence. Eichberg’s (1986) thesis of an indoor–outdoor impulse in the spatial configuration of modern sport has influenced subsequent considerations of the changing forms of outdoor recreation and the withdrawal of some sports – climbing, surfing, snowboarding – from natural environments to purpose-built, commercialized, and confined indoor facilities.

Sport geographies have engaged with phenomenology, feminism, and poststructural theory to examine the (re)production of the
sporting body by material, performative, and discursive practices that occur spatially, and which emanate through human and nonhuman encounters. At a microscale, the sporting body which runs, jumps, swims, surfs, and climbs has been reported upon, detailing the cognitive and corporeal forms of knowing which are produced through the exercise of visceral skill and talent in a sporting space. The felt intensities of bodily movement, where environment is sensed through sporting performance and is inscribed upon the sporting body through tactile and kinaesthetic engagement, are pronounced in considerations of the socio–nature relationships underpinning outdoor sport. The emotional state of the sport participant in extreme sport cultures and spaces – where the demands of fear, anger, relief, freedom and ecstasy are confronted – has further complicated the analysis of gendered embodiment and subjectivities.

The broader cultural and ideological contexts of sporting bodies account for another strand of scholarly interest. Bodies have also been examined in terms of constraint, power relations, and regulation. Seeing sportscapes as a disciplinary space, scholars have been interested in the control of bodies through the banning and control over movement. Access and maintenance of exclusivity of landscapes for sporting and recreational purposes create a specific form of exclusion, not only from the physical spaces but also from the symbolic meanings of these spaces. Scholars have been encouraged to attend to the social construction of bodies by hegemonic power relations so as to cultivate more nuanced inquiries into the spatialities and materialities of gender, sexuality, and race. A well-developed literature has considered the body as a spatial scale within patriarchal sporting cultures, revealing the maintenance of corporeal ideals by consumerist social forces and forms of exclusion and oppression faced by divergent, transgressive, and less able “Others.” Queer geographers have highlighted the presence of the erotic and sensual as part of sports places, contested homophobia in sporting places, including stadia, and sought to critically interrogate heteronormative sporting spaces. Racial differences are also a key social difference through which sporting spaces are created. The racializations of bodies through sports include the ways in which white ethnicities in certain sports pass unnoticed while Black successes require explanation, often through genetics.

Access to both sporting cultures and sporting spaces is also classed in historically specific as well as spatial ways. Legal factors relating to property rights and patterns of land ownership can shape the ability to use space for sporting purposes. Various outdoor sports in the United Kingdom – canoeing and kayaking, climbing, mountain biking, surfing – continue to confront access issues linked to the defense of land rights, and conflicts over multi-uses of limited natural resources have been persistent aspects of the sporting experience as participants seek to exercise a right to enjoy the countryside for recreational purposes. Land ownership conflicts are also a feature of sport tourism development in Canada as the rights of First Nations come into conflict with economic development priorities over new ski resorts, while a global antigolf movement has been prominent in Mexico and elsewhere, showing the limits to the imposition of “sportscapes” on indigenous populations with differing development priorities.

Future directions

There are numerous productive avenues of investigation for sports geographies. These include: critical spatial theories which can address recent developments around affect/nonrepresentational theory where affect and materialities become
key in investigations of sporting places and experiences and to account further for the actants/agentive capacities of sporting landscapes in the production of the sporting experience; political ecology of sports tourism/sport-for-development, and the links to global social justice campaigns and movements; sport and virtual geographies, to account more deeply for the importance of new media technologies in extending sportscapes into private and virtual settings; and the greater employment of Geographic Information Systems to map sportscapes, sites, and places and to provide the researcher with greater evidence on the complex networks existing among players, coaches, fans, and consumers, accounting for the forms of transnational exchange and cultural encounter that can occur in and through sport and the various spaces it is practiced.

SEE ALSO: Affect; Class; Gender; Globalization; Landscape; Nonrepresentational theory; Place; Political ecology; Sexualities; Social justice; Space; Tourism

References


States differ in terms of their population and the size of the region upon which they seek to exercise their authority over social forces and events. The significance and power of states in world politics and world economy differ fundamentally. In scholarly literature, states may be categorized as “weak” or “strong” with regard to the legitimacy they enjoy among their population as well as with regard to their capacity to exercise territorial sovereignty. Histories, development trajectories, and institutional structures of states also vary significantly.

In everyday language the state is often understood as a subject or as a thing. But as a scholarly concept the state is hideously slippery and tends to escape all attempts at consistent definition. The modern state nonetheless marks a development from medieval to modern space: the homogenization of territorial authority, particular linearization of state borders, as well as more or less successful elimination of alternative and often non-territorial forms of political organization. The state affects the everyday lives of people across the globe and is constantly reinforced through ideas such as democracy and nation. Forecasts of the demise of the state have proven premature.

Theorizing the state

With very few exceptions, the land surface of the Earth is divided up into territories over which states claim authority. Again with very few exceptions, all human beings inhabiting the Earth hold a citizenship that makes them citizens of one state or another (or, in case of multiple citizenship, of several states). In traditional ways of thinking about the modern state, the state is defined by three central components: territory, population, and authority. Specifying the relationship between these components, the German sociologist Max Weber coined the formulation that the state “is a human community that (successfully) claims the monopoly of the legitimate use of physical force within a given territory” (1994/1919, 310–311; emphasis in original). For a variety of authors from a range of academic disciplines, this definition and its components have become the point of reference for theorizing the concept of the state and its multiple and changing forms.

Following Weber, the British sociologist Anthony Giddens (1985, 120) calls the (modern) state a “bordered power container” and describes it as a political organization whose rule is territorially organized and which is capable of using the means of physical violence to sustain that territorial rule. Accordingly, the state is the container of modernity and the power of the state is exercised almost uniformly throughout an exclusive territory. In this tradition, it has become common to understand the state as a social organization that is composed of a relatively coherent
matrix of institutions of executive and political control (regulative authority) and core personnel (elites) through which it exercises power over a particular population within its well-defined geographical space.

Taking issue with the Weberian tradition, some scholars have questioned whether the monopoly of legitimate violence can ever be actually achieved, whether other spatial forms such as networks and spatial scales are not at least as important for the state as is territory, and in what sense the people of a state actually form a “community.” Through such controversies, the theory of the state has both progressed and differentiated into rather specialized debates. Another important controversy revolves around the question whether the state is primarily defined by internal relationships, such as the capitalist economy, gender relations, nation building and nationalism, and the national legal systems or by its external relationships, mainly by the fact that states need to be recognized by other states in order to actually become a state.

While more traditional positions focus on the state as a subject or a thing possessing these components as properties, most authors today will agree that it is much more fruitful to understand the state as a process, that is, as a flexible entity defined by territory, population, power, and a relationship to other states which develops over time. In this regard, more materialist traditions emphasize the way in which the state, as a social relation, is the result of practices that obtains a certain fixity which in turn structures practices. More poststructuralist positions on the other hand emphasize the way in which the state is mainly the effect of practices, especially of mundane and discursive ones, and therefore not so much structuring but itself structured. Believing that both traditions need to be taken seriously and should be combined, a brief sketch follows before focusing on selected aspects of the state which are deemed to be of particular importance within geography.

The state as social relation

The Greek theoretician of the state, Nicos Poulantzas, coined the formulation that the state is “the specific material condensation of a relationship of forces among classes and class fractions” (1978, 129; emphasis in original). By this he means that in capitalist societies, social processes – the capitalist economy in particular – result in different groups (especially classes and class fractions) with different and partly antagonistic interests. To prevent these antagonisms from turning into upheaval, the state organizes social classes in a way that includes the interests of all relevant classes into the regulation of society without touching class society as such. Social relations are “condensed” in state apparatuses such as ministries and administrations in a way that grants influence to all relevant groups to pursue their interests. For example, in many Western countries the labor unions have a strong influence in the ministry of social affairs and in welfare administrations, whereas financial capital has strong influence in the ministry of finance, and environmentalist groups in the ministry of environment. For this arrangement to be possible, scholars in the materialist tradition emphasize the relative autonomy of the state vis-à-vis civil society. This line of argument goes back to the Soviet legal scholar Evgeny Pashukanis, who asked as early as 1924: “Why is the apparatus of state coercion created not as a private apparatus of the ruling class, but distinct from the latter in the form of an impersonal apparatus of public power distinct from society?” (1980/1924). The answers that Pashukanis, Poulantzas, the West German state derivation debate, and others in this tradition give is that this is the most efficient structure for capitalism to function and the
unintended result of struggles between classes and inherited forms of domination from the pre-capitalist period. The separation of the state from society is thus explained both structurally and historically, that is, as the result of practices and as a structuring feature of practices.

Building on this tradition, the British sociologist Bob Jessop has developed his strategic-relational approach to state power (e.g., Jessop 1990). In his conceptual frame, the state is first the site of strategies of groups, classes, and class factions, and as such more accessible to some social forces than others at a particular conjuncture according to the strategies they adopt to gain state power. Second, the state generates strategies. It is an institutional platform through which societal forces are able to mobilize strategies for accumulation and related projects. Third, the state is a product of strategies, a particular sum of currently existing organizational structures and modes of intervention which are conditioned by inherited strategies.

The state as effect

Following an examination of some arguments by Poulantzas, the British historical sociologist Philip Abrams argued that the processes and relations that make up the “palpable nexus of practice and institutional structure centred in government” (1988/1977, 82) do not result in the object “the state,” which is “an essentially imaginative construction” (pp. 75–76), but only in the pervasive idea of the state which legitimizes domination. Accordingly, the state is not considered as a thing but rather as a powerful spatial imaginary and a mystifying idea, an abstract-formal object. The state is thus an abstract phenomenon that comes into being as a structuration within political practice. In a similar vein, the French philosopher Michel Foucault (2007, 144) doubts that the state has a “unity, individuality, and rigorous functionality [or] importance. After all, maybe the state is only a composite reality and a mythicized abstraction whose importance is much less than we think. Maybe.” What these and other discussions of the state have in common is that they acknowledge the concrete institutions and practices of state institutions and discourses about “the state” but refuse to accept the actual existence of an entity or object called “the state.”

Along these lines, the British geographer Joe Painter (2006) discusses the “everyday prosaics of stateness,” that is, the mundane practices in which “statization” takes place. In such a view, the state does not exist independently of the everyday practices which make up the state. The state, in short, is a reification of social practices. Painter uses the concept of statization of social life as opposed to a separate sphere called “the state” in order to overcome the binary between the state and civil society as well as to describe “the intensification of the symbolic presence of the state across all kinds of social practices and relations” (Painter 2006, 758; emphasis in original). Accordingly, the statization of social life (and space) proceeds through and thus also depends on quotidian practices undertaken by thousands of citizens and state officials on a daily basis.

Geography and the state

The development of geographical thinking in Europe has always been closely associated with the state. The institutionalization of geography as a university discipline from the nineteenth century onwards took place concomitantly with and often in the service of the “nation-state.” Geographic knowledge was regarded as important for more practical questions such as the mapping of the nation-state’s territory and the drawing of external and internal boundaries or the gathering of useful information about
international commerce, colonialism, and war, as well as for more ideological ones such as the production of discourses that tied together the imagined community of the nation within its territory and state, especially for schools. While the actual importance of the contribution of geography and geographers toward these goals is disputed, the theoretical foundation of academic geography in the nineteenth century as well as major works of central figures of the discipline of that time is evidence of the close relationship between the discipline and the state.

Historians of the discipline regard the German tradition as an especially important, clear-cut, and problematic example of the close relationship between the discipline and the state. Carl Ritter (1779–1859), the first chair of geography in Germany (1820) and one of the founders of the academic discipline of geography based on environmental determinism, and others who further developed this tradition, such as Friedrich Ratzel (1844–1904) and Alfred Hettner (1859–1941), argued that states were “not only earth-bound, but also earth-determined” (Schultz 2013, 21). In this tradition of thought, people of nations with distinct cultures develop by dealing with distinct natural environments whose borders should, ideally, also form the borders of the state (Schultz 2013). In the twentieth century, this line of thought became one of the foundations of Geopolitik or Geopolitics, which is often associated with Adolf Hitler’s thinking and politics.

After World War II, the focus on the relationship between geography and the state changed. Within the national welfare states that were constructed between the 1940s and the 1980s in advanced capitalist states, academic geographers in Europe and beyond have explicitly contributed to the constitution of state spatiality. Consider, for instance, the influence the discipline provided in many European state contexts to the birth of spatial planning (sometimes termed as town and country, or regional, planning) as a “scientific” approach to organizing the space of the nation-state and to the emergence of regional planner as a profession.

The state has long been equated with the territory or society, which has conditioned research questions, and thereby results, in geographical scholarship by being the taken-for-granted object of research (“methodological nationalism”). It has formed a powerful and often taken-for-granted structural frame for disseminating knowledge in lecture halls and geography textbooks. Coupled with descriptiveness, this embedded statism, as Peter Taylor would have it, was the defining characteristic of most of the geography of the state prior to the 1970s (for an overview of theories of the state in geography prior to the 1980s, see Clark and Dear 1984).

**Geographies of the state**

The relationships between population, territory, authority, and state institutions are continuously changing and are underpinned by shifting power structures. Their relations have been widely discussed in geographical literature during the past twenty years. Geographers’ and other spatially oriented social scientists’ contribution to the state-related literature has been made through conceptual work on the interconnections between state power and state spaces. The key contention has been that the (capitalist) world is constituted spatially within societal processes and practices and that the state remains one of the key components of this spatiality. State spatiality is not only a passive outcome of societal process, but rather its constitutive dimension.

The rise of social theorizing in human geography in the 1970s resulted in challenging the taken-for-granted spatiality of the state in geography. Structural (for early contributions see Clark and Dear 1984) and poststructural
positions in particular have been central in this endeavor. The re-assessment of the geography of the state stems from the debate on capitalist globalization. This debate deals with whether the state is being “hollowed out,” “filled in,” or is under transformation in the age of growing “marketization.” The transformationalist argument to this debate is that though the state remains a key constituent of the world political map, the ways in which state power has been exercised have fluctuated over the past decades (Peck 2004). There is also geographical variation in this process: state transformation has taken place differently, say, in Western Europe and in North America, in the former Eastern Europe, and in the “developmental states” of East Asia.

The materialist (spatial) research on state transformation has from the 1990s been structured around a few concepts such as the competition state, the workfare state, the rescaling of the state, and the neoliberalization of the state. These literatures are predicated upon an idea that the state has not been hollowed out but rather has been under qualitative change in recent decades. Today, state power is exercised differently compared with the so-called Keynesian welfare state that was characterized by nationally scaled systems of redistribution, planning, reproduction, and so on.

One influential way to describe this transformation in state theory is Jessop’s (2002) distinction between the “Keynesian welfare national state” and the “Schumpeterian workfare post-national regime.” These terms are used to elucidate the state transformation that has occurred in the advanced capitalist economies over the past decades. The Keynesian welfare national state refers to the economic policies that clustered around the issue of fighting unemployment in a relatively closed economy. Here state intervention was intended to compensate for market failure. Moreover, it was a regime that sought to use the state to construct demand and create norms of mass consumption throughout state territory. The national was in this regime the primary scale in policymaking.

The “Schumpeterian regime,” which emerged in the 1980s, in turn, highlights innovation policies and policies of competitiveness as central public policies for coping with the purported “open economy.” Social policies become increasingly subordinated to economic imperatives and governance is invented to compensate for market failures. Scale-wise, this regime marks a relativization of scale. The Schumpeterian competition regime does not result in the evaporation of the national but rather it de-stabilizes its supremacy in policymaking (Jessop 2002).

Neoliberalization of the state

Neoliberalism, as a theory or ideology of market fundamentalism built on privatization, liberalization, deregulation, and neoliberalizations, as processes enforcing changes following these principles, were central to the way in which societies and states have been transformed in the last decades. In a broad sense, neoliberalization is a process of state restructuring under contemporary capitalism (Ward and England 2007). Scholars of neoliberal state transformation are concerned with the ways in which financialization, trade liberalization, boosting of internationalization, individualization, and the associated politics of entrepreneurialism, privatization, and increasing labor market flexibility have been articulated and enacted in policymaking. Here state power is understood as the primary catalyst of neoliberal globalization. The Italian political scientist Giandomenico Majone (1994) coined the concept of the regulatory state to highlight the fact that political re-regulation has been essential to the neoliberalization of the state. Neoliberal restructuring of the state has been possible only through projects in which
states engineer their own reform and in so doing rework their internal spaces and institutional matrix.

Not only is neoliberalism a multifaceted phenomenon that has been subject to mutation over time but there are also spatially variegated forms of neoliberalization. Given that the process is both historically and geographically contingent, it is critical to investigate the multiple and sometimes unexpected ways in which neoliberalism is domesticated in different contexts, through different sites and in particular conjunctures as well as how it coexists with other political rationalities. This is crucial for an understanding of how capitalism becomes sociospatially differentiated in different state contexts as inherited state structures and transnational policy processes get entangled.

Rescaling the state

Scholars interested in rescaling usually depart from the idea that state spatiality needs to be viewed as a complex expression of ongoing practices and associated processes of sociospatial regulation at various scales. Drawing from strategic-relational state theory, the American urban theorist Neil Brenner (2004) has made important contributions to the debate on the rescaling of the state. His synthesis on the “new state spaces” that emerged in the 1980s as a response to the crisis of Keynesian state strategies and associated North Atlantic Fordism is a careful empirical and theoretical work on the processes of state restructuring that have taken place in Western Europe during the past four decades. Following German theoretician of the state Joachim Hirsch, he contends that Western European states, through all sorts of policymaking and associated regulatory experiments, were increasingly becoming “competitive states” that foremost aim at restructuring society in the service of the competitiveness of national capital. In this process, one strategy has been the rising importance of subnational scales, that is, the local, metropolitan, and regional scale. Sociospatial transformation must thus be seen as a response to the crises of previous forms of regulation. Whereas some of the recent scaling activities of the state, which Brenner conceptualizes as rescaled competition state regimes, may facilitate economic growth, these attempts are also characterized by potential failures, uneven development, and sociospatial polarization.

Even though one who is writing from another welfare state context may question the argument that post-Keynesian states have abandoned the ideology of uniform territorial development altogether, and moved toward fostering the competitiveness of selected nodes in state space, Brenner’s approach (2004) is very insightful. The territorial state is understood here as an essential “geographical arena” and “agent” of the process often dubbed globalization rather than as a passive victim of it. One of the strengths of the rescaling approach is that it provides a vocabulary to analyze developments which take place today in, say, metropolitan governance as expressions of broader state rescaling processes which often aim at promoting territorial competitiveness rather than generating inter-territorial equalization. One may thus focus on the urban governance and urban policies to understand the state restructuring in the Organisation for Economic Co-operation and Development (OECD) sphere and beyond.

Spatial selectivity of the state

The rescaling of the state, which is motivated by neoliberal political rationality and associated regulatory experimentation and which marks the resurgence of the local and regional as sites of governance, indicates a growing “spatial selectivity.” Spatial selectivity refers to
the processes of spatial privileging and articulation in state strategies and related projects through which public policies become differentiated in state territory in order to target particular conceived scales such as the “European” or “global” (see Jones 1997). To illustrate, governments are investing lots of money to make new spaces, ranging from certain local transport infrastructures to various high tech areas, special economic zones, and convention centers in order to “go global.” Although these scaling attempts may be done for a variety of purported aims, they nonetheless accentuate the claims that local developments can be understood as components of global processes and that the “national” is actually one of the key sites for the “global.” The national features of state space can become de-nationalized through different kinds of spatial arrangements.

The spatial selectivity which manifests itself in different globalizing investments has often been motivated by policies of competiveness and other economic priorities. Spatial selectivity can be understood as resulting in “neoliberal exceptions,” that is, often highly economized spatial exceptions (such as “special economic zones”), which treat different segments of population differently in terms of rights and privileges (Ong 2006). The spatial selectivity of the state underscores the fact that state spatiality is never fixed. Rather, the spatiality of the state can be understood as a process of constant becoming which is underpinned by changing political rationalities and conditioned by political contestation, shifting power relations, and inherited spatial structures.

The everyday spaces of the state and nation: prosaics and embodiments

The state should not be understood as a “thing” which occupies a clearly separable sphere of the social whole. The state, conceived of as a coherent entity separate from society, has proved a remarkably elusive object of analysis. Geographers have sought to overcome the persistent state/civil society binary through inquiry into the everyday spaces of stateness/statehood. These everyday geographies point not only to the inescapably uncertain and potentially unexpected spatial outcomes of state-related actions but also to the potential spatial unevenness of state-effects (Painter 2006).

The everydayness of the state also alludes to the entanglement of “nation” and the “state.” For example, media, religion, education, and military are often examined in the context of national identity construction, indicating that the statization of people takes place through these “national” institutions and related practices. Accordingly, the nation is constantly connected to the state as an imagined collective actor in mundane symbolic practices such as festivities, commemorations, sporting events, food, shopping, and so on. The very foundation and functioning of conceived nation-states is dependent on these and many other mundane routines.

Over the past two decades, scholars drawing from different strands of feminism have significantly contributed to the spatial study of the state. They “embody” the nation-state, and demonstrate how the state is embodied in all sorts of social practices. The feminist scholarship is often in line with the previous arguments which challenge the conventional position of viewing the state as a coherent entity that exists independent of practices and configurations of the everyday.

Feminist scholars approach the state as a set of gendered and raced institutions and associated social relations and practices (Mitchell, Marston, and Katz 2003). They examine how gender and patriarchal relations make a difference to the state. The state is viewed as a form of potential
STATE, THE
domination – a complex mixture of social practices that are potentially spatially discriminating and marginalizing (Mountz 2004). The focus can be, for instance, on how particular gender identities are privileged in state strategies and how the structures of the state tend to reinforce gender biases.

Rather than treating the state and civil society as two distinct spheres, feminist scholars seek to make visible the fact that embedded power relations position individuals and groups of people in different hierarchical relations, with the state as an imagined collective actor. Some identities may become privileged and some others silenced or marginalized in the sphere of decision-making, planning, and political agenda setting. This is particularly the case in some policy sectors such as security policies. Feminist scholars thus destabilize the conceived coherence of the state by inquiring into the silenced voices and discussing, for instance, mundane aspects of security of the private (home) and the everyday spaces of the local. It is equally important to study “human security” related to issues like immigration or labor mobility rather than militarized and bounded versions of state security (Sharp 2007). The feminist approach thus challenges the militarization of states and attempts to develop a politics of security at the scale of the body.

The state/violence nexus

The modern state attempts to institutionalize and to stabilize a certain societal order in line with established class, gender, and race relations, thereby producing and continuously reproducing these very relations. In modern states one central method for effectuating this order is criminalization, that is, the defining of rules of allowed conduct in criminal law and their enforcement through the police, criminal courts, and the penal system – all in the name of security. As all these institutions, together often referred to as the criminal justice system, are state institutions, it is here that the state’s claim of the “monopoly of the legitimate use of physical force within a given territory” (Weber 1994/1919, 310–311, emphasis in original) or state power becomes most tangible.

On all stages of the process of criminalization a constant struggle occurs between social forces that aim at de-/criminalizing certain conduct. Think of the struggle to decriminalize homosexuality, or the status of the sans papiers (undocumented immigrants), feminist campaigns to decriminalize abortion and to criminalize domestic violence, or the struggle over what counts as legitimate social protest in public spaces. Also, a tension between fixed rules and regulations on the one hand and the flexibility of the discretion of individuals, for example, police officers or judges, can be observed.

Processes of criminalization also have a geography. The extent to which panhandling or demonstrations are tolerated or prohibited differs not only between states, but also within a territory depending on, for example, condensed relations of force within a city tending more toward “law and order” or more toward “social justice.” Whether a person is perceived as “out of place” and possibly dangerous based on her or his appearance is fundamentally geographical as well – and so is the police’s mental geography of possible dangers (Herbert 1997). While this mental geography and the discriminatory results of geographically uneven policing (“racial profiling”) were regarded as a problem during Fordism, with the neoliberalization of societal relations we can observe policing and police institutions increasingly turning to spatialized ideologies, technologies, strategies, and measures. Keeping the impoverished victims of neoliberalizations out of certain parts of the
city that are reserved for affluent residents, consumption, or higher education, for example, can be accomplished by criminalizing panhandling, loitering, or gathering. In such instances, state power is used in the service of particular interests and legitimized with security.

Investigating territorial and relational state spaces

The state can be understood in a number of ways. It is abstract yet concrete, a thing and a process, structuring practices and structured by practices, material and imagined, theorized and experienced. In recent human geography, these seemingly mutually exclusive characteristics have often been discussed under the rubric of “the relational” versus “the territorial,” with the relational being associated with poststructuralist, feminist, and network approaches, and the territorial with materialist and structural interpretations. Far too often, this has been presented as an ontological questions about how the world really is or how it should be (for some scholars, the issue of territoriality and relationality is a pressing political concern). This framing sometimes prompts superficial “articulations of change”: for instance, that the contemporary world is increasingly made up of networks rather than bounded state territories.

In contrast, the mutual exclusivity often constructed between these two perspectives needs to be overcome. Scholars have argued that social processes are more often than not relational and territorial at the same time, that the difference between the two is not ontological but practical, and that to understand the qualities of a phenomenon is a matter of ex post and empirical rather than a priori and theoretical engagement with it. Thus, network-related practices create state territory, but similarly territory-related practices create “state networks.” “Territory” and “network” are not separate or contradictory attributes of space.

An illustrative argument could be, for instance, that in spatial planning the concurrent presence of both territory and network is something unforeseen. Rather than taking the novelty of the coming together of territorial and relational as given, there is a need to engage in spatially and temporally sensitive scholarship that is interested in the intertwining of the relational and the territorial processes in different historical conjunctures (of modernity) and in different state contexts.

The challenge for geographical studies of the state is to examine the coming-together of the territorial and the relational in a variety of state-related social practices such as the metrics and indicators that measure competitiveness or creativity of places and states, education, policing, planning manuals, and so on. To meet this challenge requires empirical research on the coming-together of the interface of yet another binary, that of politics and economy, meaning that the distinction between the territorial and relational state spaces gets easily reinforced if economy is understood as inescapably relational (movement, networks, openness, inclusion, soft spaces, porous borders) and politics is conceived strictly as a matter of bounded territory (exclusion, spatial closure, hard spaces, fixity, borders, sovereignty).

Second, one of the possibilities for actually examining this coming together is to pay attention to the political rationalities that are in operation in state transformation and in the related process of policy formation. Such political rationalities can be investigated empirically, for instance, by examining various critiques of the territorial state. During the past two decades, such critiques have been aired by very different actors such as neoliberal globalization boosters, consultants promoting “sustained growth,” and
all kinds of urbanists and environmentalists. For instance, (global) city-centered arguments that governance based on territorial states should be replaced by city-network based governance and democracy (the so-called let the mayors rule the world thesis) have become increasingly salient, at least on both sides of the Atlantic.

Some have indeed been ready to suggest that intercity networks are rapidly taking over the interstate networks in global governance. In such a view, the territorial state is dysfunctional in contemporary world politics which is less based on command of territory than on capabilities to associate the state with “global networks.” This form of reasoning is often predicated on all sorts of “flow arguments” that highlight the economic and political significance of the so-called global cities as both critical circulatory infrastructures for growth and political command centers. In sum, the numerous critiques of the territorial state highlight the changing city/state relations as a nodal research topic for the geography of the state.

A third issue merits scholarly attention. The way in which state power is exercised through different kinds of assemblages/ensembles/apparatuses is an important topic for the geography of the state. Consider, for instance, the knowledge-based society ensemble as a form of government that seeks to govern life and spaces within states according to the conceived rules of the purported ubiquitous and borderless “knowledge-based economy,” which has been the dominant narrative of state transformation in the OECD sphere and beyond during the past two decades. In addition, an investigation of the different spatial imaginaries through which different kinds of assemblages/ensembles/apparatuses exercise power over life and spaces is an important topic that may provide insights on the progressive and regressive dimensions of the coming together of the territorial and relational spaces in contemporary state transformation.

Fourth and finally, examining the territorial-relational aspects of state transformation may open up new perspectives to understand the economization of the political and the associated processes of de-politicization. Again, there are many possibilities to study de-politicization. One may, for instance, study strategic (spatial) planning through an examination of the representations of space – such as maps. These representations may “sell” particular planning ideas as politically neutral devices and de-politicize neoliberal development within the state. In such a view, representations of space regulate thought and action by concealing motives, interests, and societal tensions. Another example of the study of de-politicization would be an inquiry into the role of different forms of management knowledge and the transfer and mutation of such knowledge in policymaking and in state spatial transformation more generally.

**SEE ALSO:** Civil society; Geopolitics; Nation-state; Power; Public policy; Scale; Sovereignty; Territory and territoriality

**References**


State-owned enterprise

F.Z.Y. Hu
Hong Kong Institute of Education, China

State-owned enterprises (SOEs), as defined in law and economics, refer to enterprises in which the right to use, to benefit from, and to transfer the assets within a firm’s organizational boundary accrue to the state in an abstract sense, albeit in reality this right is exercised by an ensemble of bureaucratic entities or government jurisdictions. The phenomenal rise of SOEs through large-scale nationalization programs during the decades following World War II was inseparable from disillusion with market fetishism, but the move to state ownership has taken on distinct forms and functions in different world regions (Toninelli 2000). In the developed world, SOEs were driven by the perceived need to correct market failure; in developing countries, SOEs more often served the purpose of post-independence governments seeking to achieve developmental goals. The significance of SOEs is particularly pronounced in former socialist economies, where they were the extension of the state redistributive regime to steer production, extract surplus, and organize society.

Since the 1980s, SOEs have become the subject of aggressive privatization (Megginson and Netter 2001). Neoliberal discourse ascribes the privatization of SOEs to potential efficiency gains resulting from a clearer definition of private ownership. Political-economy perspectives focus on the incentive structure of state bureaucrats and other related stakeholders in affecting the pace, form, and effect of ownership transformation in the state sector. While ownership matters for the growth and performance of SOEs, it is increasingly accepted that their functionality depends on prevailing political, social, and geographical conditions.

The conventional wisdom about SOEs, of a public versus private dichotomy, is challenged by the presence of variegated organizational forms taken by SOEs. Scholarly debate exists over the nature and implications of such hybrid or recombinant property forms. Some underscore the transitional nature of such intermediate property forms, positing their inevitable convergence on the conventional private ownership of liberal capitalism. Others, following a path-dependent treatment of SOEs, advocate the search for a theory of state capitalism as an alternative to the hegemonic project of neoliberalism.

The growth trajectory of China’s SOEs has presented itself as a rare and valuable way to interrogate the received wisdom on the necessity of getting the institution right in order to foster development (Whyte 2009). Whereas the recent rise of China’s SOEs on both the domestic and the global economic stage is interpreted by neoliberal proponents as hindering the growth of private entrepreneurship and long-term sustainable development, others acclaim state ownership of firms in the Chinese economy as the core feature of a successful alternative development model to the Washington Consensus, one with normative implications for the world’s developing countries.

SEE ALSO: Firms; Neoliberalism; State, the
STATE-OWNED ENTERPRISE

References


States and development

Pádraig Carmody
Trinity College Dublin, Ireland
University of Johannesburg, South Africa

States

There are a variety of definitions of what constitutes a “state” and also many theories about their purpose and nature. For Strange (1996) states are institutions to mediate and resolve conflicts. However, this arguably neglects how states initially, and the interstate system more generally, came into being, often through war. For Jessop (2002, 6), however, states are social relations. For him the state is “an ensemble of socially embedded, socially regularized and strategically selective institutions, organizations, social forces and activities organized around (or at least selectively involved in) making collectively binding decisions for an imagined political community.” The extent to which this applies to “weak” or “failed” states in parts of the Global South can, however, be debated.

States serve, or are often meant to serve, a variety of functions including the provision of security and public goods, management of the economy, and the reduction of inequalities generated by and through markets. However, different states have different emphases even if most states around the world are now capitalist, in the sense that they generally serve to promote economic growth and capital accumulation for reasons of legitimation, collective good, and sometimes the self-interest of officeholders.

Development

For mainstream analysts development is synonymous with economic growth, industrialization, and modernization. However, as noted previously, the term “development” is also highly contested, with some arguing that this ideology is in effect a Trojan horse for a neocolonial or imperial project. For others, development can be both a project and a process.

The “age of development” is sometimes traced to US president Harry Truman’s speech in 1949 when he argued that “We must embark on a bold new program for making the benefits of our scientific advances and industrial progress available for the improvement and growth of underdeveloped areas.” Despite the contested nature of the term, “development” has arguably become the central mission, rationale, and justification for the existence of states in the Global South. Indeed there has been a substantial amount written in recent decades on a purportedly new genus of states – developmental ones (Wade 2004; Amsden 1989).

The role of the state in development

For some scholars the nature and role of the state are the central issue in international development. For example, Acemoglu and Robinson (2012) argue that where colonial institutions have been more fully implanted the result has been more successful development. This perspective feeds into the debate in economics about whether institutions or geography matter more for developmental outcomes. However, this neglects the fact that institutions are
States and Development

geographical, in the sense that they are both associated with geographical territories and interscalar relations and help to constitute them. More recently, others have argued that it is policies, rather than institutions or physical geography, that matter most for economic development outcomes. However, this perspective might be questioned for its methodological nationalism and geographical theorization. Relations between places shape policies, as do the particular socioeconomic and physical geographies of nation-states. Nonetheless, the state remains a central ensemble of institutions and practices or an “actor” in the (under)development process.

In the immediate postcolonial period for Africa and much of Asia there was extensive optimism about the role that the state could play in development. This was, in part, based on the successes of postwar rebuilding in Europe under the auspices of Keynesian welfare states. Some of the most influential theories about the role of the state in development included those of Rosenstein-Rodan (1943), Hirschman (1978/1958), and Myrdal (1968). These different theories emphasized the importance of the state’s promotion of development through planning and the expansion of manufacturing, in particular, which was seen to be more technologically dynamic, to create higher value-added, and to have greater employment potential. Rosenstein-Rodan’s “big push” theory argued that both capital and consumer goods industries should be developed simultaneously to allow for complementarities, market creation, and economies of scale. Somewhat later, Thomas (1974) argued that there were “basic industries,” such as intermediate goods ones, which had the greatest potential for value addition and economic transformation.

The Latin American structuralist school of economics argued that the imperative for economic development was to engage in import substitution of manufactured imports from the United States and Western Europe in order to diversify and develop economies. While the raising of tariffs and import quotas and bans did protect domestic markets, it raised costs and was often import-intensive in terms of capital goods, contributing to debt accumulation. The raising of costs and the limitations of small domestic markets also meant that economies of scale were also often not realized, further reducing efficiency, unless complemented by measures to promote exports, as was the case in the successful newly industrializing countries of Asia. In some of the most successful newly industrialized countries of Asia, stepwise policies of primary import substitution were followed by primary export promotion and then secondary import substitution of more technologically advanced and capital intensive industries.

While the World Bank was willing to selectively support policies of import substitution until the early 1980s, under Robert McNamara that institution promoted state-led development strategies in the 1970s focusing on “basic needs,” such as water provision and integrated rural development. These twin foci were later to reappear in different guises in the emphasis on health and education in the World Bank and International Monetary Fund (IMF) Poverty Reduction Strategy Papers from the late 1990s and the Millennium Villages Project supported by the United Nations Development Program. The World Bank sometimes engaged in historical amnesia, with its staff critiquing projects such as the Morogoro shoe factory in Tanzania, which never operated at more than 5% of capacity and was meant to export shoes to Italy, even though the Bank itself had funded this project.

Both modernization theory and heterodox structuralist approaches placed a heavy emphasis
on the role of the state in development. Both shared an underlying assumption that states would be rational, Weberian actors which would promote development. However, with sometimes poor performance and the economic crises of the 1970s and 1980s, this optimistic view of the role of the state generally gave way to more sanguine, pessimistic, or restricted views of the role that these institutions could or should play in promoting development.

Rethinking the state’s role in development: “overdeveloped” and “neoliberal” views

After the initial postindependence optimism about the role which the state could play in guiding and promoting economic and social development, some social scientists began to question earlier, positive managerialist (Weberian) conceptions of the state. Neo-Marxian theories were developed which posited an alliance between domestic capitalist classes, bureaucratic elites, and international capital and great powers which perpetuated economic underdevelopment, or in the case of Brazil “dependent development.” For Hamza Alavi (1972) the underdevelopment of the domestic bourgeoisie in postcolonial societies resulted in an “overdeveloped” state which was relatively autonomous in relation to its own society but heavily influenced by the needs of, and responsive to, international capital. In relation to Africa, the “Kenyan debate” was about whether the state was primarily responsive to the domestic or international fraction of capital. Later work emphasized the importance of the domestically based, but “nonindigenous” bourgeoisie of Asian extraction.

While cynicism about the role of the state in the economy is often traced to the Reagan and Thatcher “revolutions” in the United States and United Kingdom respectively, and the intellectual currents associated with these and their precursors, there was also, as noted earlier, substantial left-wing criticism of actually existing states in the Global South. These concerns, from both left- and right-wing perspectives, were to come to a head with the Latin American debt crisis of the early 1980s.

The global economic turbulence of the 1970s and the abandonment of the Bretton Woods system of fixed exchange rates, the second oil crisis, and the Mexican debt default created conditions whereby many developing countries experienced economic crises and severe difficulty in accessing private international capital markets. Many countries were consequently forced to turn to the World Bank and IMF for balance-of-payments support. In exchange these international financial institutions demanded wide-ranging programs of economic liberalization be implemented. Fundamentally these, at least in the initial stages, were about reducing the role of the state in the economy, so that it would function as a “nightwatchman” over the market, which it was felt would drive development.

According to the World Bank, officials at the IMF and also a number of academics took the view that the economic crisis in much of the developing world in the 1980s was an outcome of inappropriate economic policies, which protected inefficient domestic industries and were biased against rural development and exports. It was argued that these policies were designed and perpetuated, not because of their inherent public merit, but because of the “rent-seeking” opportunities they afforded state elites and their cronies. In order to discipline the state and force “economic efficiency,” it was argued that privatization, trade liberalization, monetary tightening, and other policy reforms should be instituted. This neoliberalization was paradoxical
in that it was a process of the global deregulation of capital through the international regulation of states, and indirectly driven by the world’s most powerful states, such as the United States, through their voting weight and influence in the international financial institutions.

The impacts of programs of economic liberalization varied depending on the specific geographical political economies undertaking them. They were generally more “successful,” in their own terms, in countries in Southeast Asia for example, where there was a regional product cycle at work. However, where this was not operative, they often resulted in competitive displacement of manufacturing production and social immiseration, particularly in much of Latin America and Africa during the “lost decade(s)” of the 1980s and part of the 1990s.

State failure, violence, and underdevelopment

The programs of economic liberalization sponsored by the World Bank and IMF were known as “structural adjustment programs” (SAPs). While the reasons for their general failure in the developing world have been explored extensively in the literature, the effect on states varied depending on the economic context and the balance of social forces. In cases of previous state collapse, a “trough factor,” where in effect the only choices were continued disorder or state reconstruction, was sometimes observed. In some cases “strong states” emerged, as in Ghana and Ethiopia for example, which were able to take advantage of inflows of hard currency associated with structural adjustment to relieve foreign-exchange constraints on economic growth. In other cases however, the cutbacks in government expenditure associated with fiscal rectitude under SAPs resulted in the dissolution of the “patrimonial glue” which held some extant regimes together, fomenting disorder. In some instances these dynamics fed into the creation of transborder crisis complexes, for example in West and Central Africa.

From a rational choice perspective Bates (2008) argues that state rulers are “specialists in violence” who may choose to deploy it either to prey on wealth or to protect its creation. Despite conditions of often intense poverty and the incentives it creates for conflict, such as the lowering of opportunity costs for youth to engage in violence, African rulers tend to remain in power longer than those in other world regions. In part this is because of the way in which they are able to leverage external resources, such as aid or inflows of foreign currency from natural resources, into internal authority. In many developing countries, regime maintenance rather than economic and social development is the primary goal of state regimes, despite public pronouncements to the contrary. This often results in unstable social formations, however, as state hegemony is not secure.

Developmental and catalytic states

The general failure of structural adjustment to achieve economic transformation in the developing world led to increased academic interest in a variety of other state models, particularly those of the newly industrialized countries of East Asia (Amsden 1989; Wade 2004). It was argued that these states had succeeded by deliberately creating economic incentives which deviated from those of the “free” market – “getting prices wrong” (Amsden 1989), channeling finance to fast-growing industrial sectors, and adopting a variety of other policy interventions to deliberately steer industrialization. According to Amsden and
Wade, developmental states are characterized by insulated, meritocratic, yet embedded bureaucracies. Others characterize the social relations of developmental states as “embedded autonomy,” where the bureaucracy is embedded in and responsive to, but largely autonomous from, rather than captured by, societal interest groups (Evans 1995). According to Woo–Cumings (1999, 27), “such economic and political relationships often imply a corporatist framework, involving large economic groupings (like the keiretsu, corporate conglomerates in Japan, or the chaebol, their equivalent in South Korea) with which the state can coordinate and negotiate investment decisions.”

In an evolution of this literature Linda Weiss (1998) argued that states in Europe, East Asia, and elsewhere continue to play a vital role in the promotion of economic development. She coined a new category of states – catalytic ones – which facilitate and guide private sector development rather than being highly directive as in the developmental state model. In light of the revisionist thesis promoted by authors such as Amsden and Wade, and partly at the instigation of the Japanese government, the World Bank also began to revise its conception of the appropriate role of the state in the economy in the 1990s, arguing that states should match “capabilities” with “roles.” However, this ignored the possibility of how to build state capacities; one of the most important questions in international development. In practice, however, it has supported the work of, among others, the African Capacity Building Foundation, which works to develop national statistical or financial management capabilities, for example.

One of the key, and often neglected, factors in the emergence of developmental states is either an internal or external security threat which incentivizes state elites to develop the economy as a source of legitimation and funding stream for the military. In Africa the developmental states of Rwanda and Ethiopia are dominated by minority ethnic groups who have experienced genocide, for example. However, this state strength may be implicated in the perpetuation and deepening of state weakness in other countries. For example, the Rwandan government has repeatedly invaded the neighboring Democratic Republic of Congo and has been implicated in support for rebel groups in the east of the country. This is another important reason to avoid the “territorial trap” of methodological nationalism when considering state forms in the developing world.

There are also a variety of models in development not captured by the state/market dichotomy. For example, some scholars have also written about the role which civil society can play in the coproduction of (public) goods and services (Tendler 1997). For Tendler the key to effective, development-oriented governance may be “synergy” between different social forces and actors. Brazil has also been marked by state innovation at a national level through programs such as Zero Hunger. The Brazilian state can be characterized as a hybrid neoliberal–developmental or neostructuralist one, where macroeconomic policy follows economic orthodoxy but microeconomic and social policy is more interventionist.

Globalization: the rise or retreat of the state in development

The nature and role of the state in development is fundamentally related to globalization. However, globalization is not a static process but a dynamic one involving rapid evolution in the forms of sociospatial relations. The rise of China in the international political economy has reconfigured the nature of globalization, given the still strong
role of the state in the economy there and the prominent role which state-owned enterprises have played in that country’s “Go Out” policy to promote investment by its companies overseas.

There has been a debate about the nature of the Chinese state. It has been variously characterized as developmental, neoliberal, neo-Stalinist, market socialist, or fragmented authoritarian. Even as it has liberalized its economy domestically, the Chinese government has expanded its spatial reach overseas through its state-owned corporations, respatializing rather than reducing the power of the state. This new model of globalization, which is state-promoted, not only calls into question the ontological distinction between “states” and “corporations” as separate social forces, but also the posited opposition between the power of states and “globalization.” Most of the 89 Chinese corporations in the Global Fortune 500 are state-owned. One of the most influential works which questioned the separation or distinction between state and society is Migdal (2001).

While many analysts now agree that states have a vital role to play in promoting economic and social development, the extent to which they can do this depends on their particular configuration. A central axis of debate is the nature of the state: whether it is hard (has the ability to enforce its will) or soft (ineffective and reluctant to challenge powerful social forces) (Myrdal 1968). Other axes of debate center on whether or not states or societies are “strong” or “weak” in relation to each other. This approach also suffers from methodological nationalism, by not interrogating which social forces are designated by such terms. For example, authoritarian states may be “strong” in relation to their domestic societies but weak relative to international capital.

Actor-network theory provides a more nuanced characterization of states as assemblages of institutions, practices, actors, and artifacts. Under conditions of globalization such a perspective allows us to examine the constitution of particular states and how they may be differentially and selectively internationalized (made responsive to international forces), for example. Geographers have also recently theorized the spatiality of state networks and sovereignty using assemblage theory, arguing that, under conditions of globalization, with a shift from “government” to “governance,” a primary characteristic of states is that they possess “reach,” not “height” (Allen and Cochrane 2010). In this theory states are not “national” but operate through multisite power relations and assemblages, and consequently development outcomes depend on the nature of these interactions and the power capabilities of other actors.

Different world regions have given rise to different theorizations of the state. For example, in relation to Africa much has been written about neopatrimonial or rhizome states, which have patronage networks through society in a root-like fashion but fail to achieve state hegemony because of their inherently exclusionary nature. Other influential theories related to the African context include those of the “suspended” and “bifurcated” state. In the East Asian region the central theory in recent decades has been the developmental state. However, with the advent of network trade, where components for manufactured products are often sourced in multiple different countries and then assembled in a low-wage economy before being exported for consumption, some have questioned whether this model still applies in the original newly industrialized countries. Others have also questioned the opposition between patrimonialism and developmentalism, arguing that some states in Africa are characterized by developmental patrimonialism.
Despite widespread neoliberalization across much of the developing world, the role of the state in promoting economic diversification and improved living standards and the ways in which it may hamper or prevent these outcomes are still some of the most debated topics in the international development literature. The “rise” of China and the South, more generally, is creating a new macro-political region – a “South space” – which has greater policy latitude and autonomy from the international financial institutions of the World Bank and the IMF. This greater policy autonomy which countries across the Global South are experiencing may not fundamentally challenge neoliberalism, but it could lead to its renegotiation through resource nationalism, whereby resource-rich states capture greater shares of resource rents and revenues. Some refer to this conjunction of events as providing scope for neodevelopmentalism.

As noted at the beginning of the entry the terms “state” and “development” are highly contested. Globalization and the widespread shift from government to governance as a modality of authority and rule have troubled the notion of the state as an ontologically distinct actor (if it ever was). What constitutes development has also been debated and redefined, and looming climate disruption and ecological constraints raise the issue of the sustainability of carbon-intensive global industrialization. What constitutes development is arguably partly about values and what different people may consider to be worthwhile. New metrics, such as Gross National Happiness or the Happy Planet Index, suggest that less attention should be paid to traditional indicators of “development” such as increases in economic output and more to ecological and social sustainability.

SEE ALSO: Dependency theory; Development; Developmentalism; Globalization; Industrialization; State, the

References


Strategic essentialism

Raksha Pande
Newcastle University, UK

“Strategic essentialism” is a term coined by Gayatri Chakravorty Spivak, a postcolonial feminist philosopher and literary theorist who is currently University Professor at Columbia University. She is an eminent thinker in the fields of postcolonialism and subaltern studies. Strategic essentialism can be understood as a deconstructive strategy of representation that involves taking the risk of adopting an essentialist position with respect to identity categories, or as Spivak (1993, 3) calls them, “masterwords” (such as woman, worker, nation, or the subaltern), in order to mobilize a collective consciousness for achieving a set of chosen political ends. The genesis of the term follows her central academic pursuit of exploring the role of representation in subject constitution. Along with the notion of the subaltern, strategic essentialism has been hailed as one of the most significant theoretical contributions of subaltern studies. It has been adopted as a slogan for identity-based politics and remains an important and contested theme within feminism.

Strategic essentialism: epistemological origins

Spivak first uses the term in her analysis of the works of the Subaltern Studies Group, a Marxist history collective of Indian scholars engaged in reclaiming Indian history from its elite imperialist versions and moving towards a historiography that gives due credence to the previously unheard voices of the subaltern – the so-called underclass of Indian society, including the peasants, Dalits, and women. In reading the work of the group “against the grain” (Spivak 1988a, 10), she interrogates the representational tactics employed in the writing of history from below. She argues that in their attempt to restore the subaltern as the subject of their history, the Subaltern Studies Group can, at first glance, be seen as embracing an essentialist position, one that assumes a stable, foundational essence as characterizing the nature of the subaltern subject. Spivak uses deconstruction as her approach to critical analysis, and posits that there is no such thing as a stable subject but only “subject-effects” brought into existence through complex networks of politics, ideology, economics, history, sexuality, and language (1988a, 13). Consequently, the subject-effect of the subaltern is brought into being through the writing of their history by the well-meaning historian, and essentialism guides the epistemology through which this subject position is restored; however, she explains that this is not substantive/positivist essentialism, which homogenizes the subject as its essence (the subaltern studies collective is well versed in the nonfoundational poststructuralist philosophies), but a strategic one, which encourages “a strategic use of positivist essentialism in a scrupulously visible political interest” (Spivak 1988a, 13). For example, in representing the history of the peasant uprising
in pre-independence India, the historians can be seen as establishing a subject position, namely that of a collective subaltern or peasant consciousness, as the focus of their inquiry. In so doing, they strategically presume the stability of that subject position (akin to an essentialist approach) in the political interest of retrieving the hitherto unheard stories of the subaltern role in the Indian independence struggle and also in critiquing the elite histories that omit them from any account of the Indian past. This essentialism is strategic because it is self-conscious and is always attuned to its “constitutive paradox,” which evokes that the object of historiography is not reducible to the essence of subaltern (1988a, 13), but at the same time it also involves taking the risk of essentialism in order to have collective political purchase. Strategic essentialism also involves the methodological presupposition that the strategically essentialized subject position is marked by “a negative consciousness – the consciousness not of the being of the subaltern but of the oppressor.” Moreover, the strategic restoration of a unified or essentialized subaltern consciousness is done with the knowledge that in the task of representation the subaltern consciousness “is subject to the cathexis of the elite” and it is never fully recoverable beyond its discursive origins, since “the subaltern cannot appear without the thoughts of the elite” (Spivak 1988a, 11); or, in other words, the subaltern cannot speak (Spivak 1988b).

Strategic essentialism can be seen as the hallmark of a “negotiated postcolonial positionality” (Spivak 1990, 72), which involves strategically intervening in the constitutive discourses (for example essentialism, colonialism) of the identity positions inhabited as a colonial or marginalized subject. Homi Bhabha (1994, 6) has called this positionality an “in between” or “hybrid space,” which allows a more complex ontological basis for understanding the material and discursive aftermaths of colonialism and Western domination. The notion of hybrid space has been highly influential in postcolonial human geographical approaches calling for alternative geographical imaginaries that recognize the complex web of relations that characterize the interconnections between the Global North and Global South (Raghuram, Madge, and Noxolo 2009). This strategy also reveals the limits of Western poststructuralist critiques of the essentialized subject, where the author is either engaged in granting the “very expressive subjectivity” to the oppressed, which she or he critiques as being discursive, or failing that, arguing for a “total unrepresentability of the marginalized subject” (Spivak 1988a, 17). According to Spivak, neither approach leaves much room for political agency and change. Instead, strategically adhering to the essentialist notion of identity categories provides a “transactional” reading of the subject (1988a, 15) where a stable, unified essence for identity categories (contextual to practical exigencies) is accepted in exchange for collective political mobilization. In other words, Spivak (1990, 45) acknowledges the impossibility of escaping an essentialist understanding of the productions of our beings and consequently advocates that we recognize and strategically “carve out a representative essentialist position” in order to “do politics according to the old rules” while staying attuned to the dangers of universalism that arise from adopting such a position.

For Spivak, all historiography, and by extension all acts of representation, can be better conceptualized as strategy (Spivak 1988a, 15) – one that determines to what extent the author chooses to embrace the notion of a stable and unified subject, and to what political end. On a wider scale, she offers the strategic essentialism of the subaltern studies groups as a representational
strategic essentialism in human geography

Within human geography, strategic essentialism can be read in the works of feminist geographers. Whether it was the initial calls for gender equality in the academy or the demands for establishing gender relations, sexuality, family, home, and the so-called irrational subjective notions of the body and emotions as legitimate topics of geographical scholarship, feminist geographers have strategically employed their subjectivity as women to expose and critique the hegemony of masculinism in the discipline (Rose 1993). The works of the Women in Geography Study Group (WGSG), for instance, can be seen as a very successful example of strategic essentialism in practice where, through their insistence on telling women’s stories (and provisionally accepting the stability of the identity category “women”) in geography, they have helped to dismantle the elite masculinist traditions of academic geography.

Gillian Rose’s arguments in the path-breaking book *Feminism and Geography: The Limits of Geographic Knowledge* can be seen as the most cogent example of strategic essentialism in practice within geography. Although she does not use the term herself, her manifesto for an emancipatory geography, which empowers women in favor of the extant masculinist epistemology of the discipline, has all the hallmarks of a strategic approach to essentialism. In interrogating the nature–culture dualism, which has fed the masculinism of academic geography, her vocabulary is redolent of strategic essentialism, with references (1993, 78–85, 159) to “strategies of critique,” “other tactics,” and “master subjects” in suggesting what an alternative geographical imagination that allows for thinking space beyond dualisms may look like. Her proposal for a feminist politics based on the strategy
of “paradoxical space” – which “straddles the spaces of representation and unrepresentability” (Rose 1993, 154) – shares the central paradox of strategic essentialism, which posits that although essence as identity (woman) is being mobilized for political ends, the subject is heterogeneous and never reducible to her essence. Her “paradoxical geography” thus allows for understanding the subject of feminism as “that of both prisoner and exile” (Rose 1993, 155) – prisoner of the essentialist category “woman” but also exiled from that essentialism in insisting on the possibility of resistance, which in its emancipatory intent will lead to the dismantling of the stability and unity of that category. This “strategy of oscillation,” or strategic essentialism (Rose 1993, 152), between the categories of self and other, inside and outside, center and margins, allows for a more politically flexible and vigilant feminism and geography.

The “hybrid spaces” (Bhabha 1994, 6) that characterize the oscillation and blurring of sharp distinctions between the center and margins have also been the focus of postcolonial human geographical approaches to development. For example, in their critique of the neoliberal discourses of development, Raghuram, Madge, and Noxolo (2009, 5) offer “care” and “responsibility” as an alternative vocabulary to developmentalism. Their focus on discourses of care and responsibility, instead of the ethnocentric vocabulary of development and aid, is aimed at revealing the power relations underlying neoliberal approaches to development where the Global South is imagined as the developing margins with the developed Global North as the center. They call for transcending the margin–center distinction by recognizing the “postcolonial intimacies” and interdependencies that characterize the relationship between the Global North and Global South. This approach involves inhabiting a “negotiated postcolonial positionality” (Spivak 1990, 72), which involves inhabiting one’s location always in relation to the other – as a relationality that posits that “all of us are already implicated in each other’s ‘presents’ in complicated ways” (Raghuram, Madge, and Noxolo 2009, 9).

A paradoxical positioning with respect to identity categories such as “race” is also evident in the work of geographers examining the viscosity of the categorization of racial identities (powered by biological essentialism) while also accepting the nonfoundational and socially constructed nature of such identities (Tolia-Kelly and Crang 2010).

When she first “began to write as a feminist,” Spivak (1993, 17) supported thinking strategically in feminist practice by choosing essentialist discourses of the category “third world woman,” for example, in order to articulate a voice and representational tactic for those on the margins of Western bourgeois feminism. Such politics of location within feminism has been expressed in the works of feminist postcolonial geographers such as Richa Nagar (2002), who have attempted to clear “a representative space” (Spivak 1990, 46) for the “third world woman” by simultaneously deconstructing this subject position and showing how it is symptomatic of the cathexis of Western feminists, while in the same vein also using the identity category and location as a third world feminist to argue for a feminist politics of difference. Here, a critical awareness driven by strategic essentialism finds a way of using essentialism as a discourse for achieving the political ends of representation and voice without making an overall commitment to these concepts.

In spite of the fact that the term “strategic essentialism” has been enthusiastically embraced as a slogan for progressive identity politics both in the academy and in the world beyond it, its critical edge has been dulled over the years, so much so that Spivak (1993, 17) now rejects the term,
saying that for her “like most strategies … it has served its purpose” and her “feminism now takes a distance from that debate.” In its wide-ranging adoption, strategic essentialism has lost its deconstructive strain and at its best has become a reference for coming to terms with the inevitability of essentialism in feminist and postcolonial discourses. At its worst, essentialized understandings of cultural difference have now replaced essentialized understandings of gender, where strategic essentialism is used to signal the feminism of the other woman. This approach has also had the unwitting effect of fragmenting the feminist movement. The notion that “the personal is political,” Spivak argues, has been understood, at least in the US academy, as something like “only the personal is political” (Spivak 1993, 4, original emphasis), where strategic essentialism is understood as only being able to speak from one’s own ground, rather than strategically moving between the personal and the collective, depending on what the particular situation at hand demands.

The notion of strategy from which the term derives much of its transactional import has been forgotten while the term is adopted and adapted as another postcolonial theory of difference. For Spivak, strategic essentialism is a strategy, not a theory, most effectively employed as a provisional enterprise in conducting a deconstructive critique of the theoretical, the ultimate goal of which is: “Not to arrive at a point where one no longer says I, but at the point where it’s no longer of any importance whether one says I or not” (Spivak 1988a, 18).

SEE ALSO: Difference; Feminist geography; Identity; Postcolonial geographies; Subaltern

References


Further reading


STRATEGIC ESSENTIALISM


Ozone is a gas naturally present in the Earth’s atmosphere. Its molecule consists of three oxygen atoms (O$_3$). Ozone concentration is largest in the stratosphere, a region in the atmosphere at approximately 10–50 km, depending on latitude and season. The so-called ozone layer in the stratosphere contains about 90% of atmospheric ozone. Even in this region, ozone’s abundance is very small, not exceeding 8 to 10 molecules per million of air molecules. Yet, ozone is one of the key species in the atmosphere, since it protects life on Earth by filtering out damaging ultraviolet (UV) radiation from the sun. Ozone is the main atmospheric constituent that absorbs solar radiation in the UV-B band (280–320 nm), protecting living organisms from direct DNA (deoxyribonucleic acid) damage. Before the appearance of ozone in the Earth's atmosphere, living organisms could only survive in marine environments, and it was only after the ozone layer was formed about two billion years ago that life could expand at the surface.

While ozone in the high atmosphere is beneficial to life, it is also a powerful oxidant. Its oxidizing properties make it a potent respiratory hazard and pollutant. At ground level, it can be formed by chemical reactions between air pollutants emitted by vehicle exhaust and human activities. High concentrations of ozone (e.g., larger than 90 molecules per billion of air molecules) are toxic to people, animals, and plants. Ozone can thus be considered as good or bad for life on Earth, depending on its location in the atmosphere. The vertical distribution of ozone in the atmosphere is displayed in Figure 1.
At the end of the twentieth century, the discovery that man-made chemicals, referred to as ozone-depleting substances (ODSs), could destroy stratospheric ozone stirred international concerns about the possible impact of human activities on the Earth’s atmosphere at a global scale. Such a decrease of stratospheric ozone abundance would induce an increase in UV radiation at ground level, dangerous for human health. ODSs include various gases produced by the chemical industry, such as chlorofluorocarbons (CFCs) and halons (organic compounds containing bromine atoms). They were used as refrigerants, solvents, blowing agents for plastic foam manufacture, fire extinguishers, or aerosol propellants.

ODSs are very stable and, once released into the air, degrade very slowly. They can remain intact for many years until they ultimately reach the stratosphere through atmospheric transport processes. In the stratosphere, they are broken down by the more intense UV radiation at this altitude range, and release reactive chlorine and bromine molecules that are involved in the ozone destruction catalytic cycles. Ozone chemical loss is thus increased, which induces a thinning of the ozone layer.

The discovery of the so-called ozone hole in the mid-1980s (Farman, Gardiner, and Shanklin 1985) heightened the general concerns about the fate of the ozone layer. This phenomenon, which is still observed today, corresponds to an important decrease (larger than 40%) of ozone total column during the springtime over Antarctica. Satellite measurements show that the ozone hole covers an area larger than the Antarctic continent. At the time of discovery, this decrease was far larger than the natural year-to-year variation in monthly averaged total ozone. In situ measurements of ozone by balloon soundings also show a near complete disappearance of ozone molecules in the lower stratosphere, between 16 and 20 km, at an altitude range where ozone concentration is at its maximum, over Antarctica.

The elucidation of the specific mechanisms that lead to the rapid disappearance of ozone in the Antarctic lower stratosphere during the springtime took a few years after the discovery of the ozone hole. Aircraft and ground-based observation campaigns were carried out, which allowed the scientific community to identify the main culprits. It was shown that the destruction of polar ozone during austral spring required (i) increased concentrations of chlorine and bromine species in the stratosphere, (ii) very low temperatures in the polar stratosphere in winter, and (iii) the return of sunlight above Antarctica at the end of the winter.

The existence of very low temperatures in the austral stratosphere in winter is due to the formation of the polar vortex, a persistent and large-scale cyclone hovering over the pole. Strong westerly winds isolate the polar air masses, which then cool to temperatures lower than $-80^\circ$C. In this extreme environment, polar
Stratospheric and tropospheric ozone

Stratospheric clouds are formed and chemical reactions at the surface of cloud particles liberate active chlorine species, in much larger quantities than in any other stratospheric regions. When solar radiation reappears above the pole at the end of winter, it initiates rapid catalytic chemical cycles that destroy ozone at a rate of a few percent per day. By the end of September, total ozone content has decreased by more than 50–60% within the polar vortex. The polar stratosphere then warms gradually, stopping the ozone destruction processes, and the breakup of the polar vortex at the end of spring dilutes the poor ozone air masses into the whole Southern Hemisphere. Since the early 1980s, the ozone hole over Antarctica has become a recurrent seasonal feature of the austral climate. As an example, the average ozone hole in September 2014 is displayed in Figure 2. After a regular increase up to the end of the 1990s, the size of the ozone hole area has stabilized at around 20–25 million km².

In the Arctic, similar processes occur but seasonal ozone losses are much smaller, due to the higher temperature conditions in the arctic stratosphere in winter. The arctic polar vortex is less stable than its Antarctic counterpart, due to the different configuration of continents in both hemispheres, and is generally warmer. The largest arctic ozone depletion to date was observed in the winter 2010/2011. Exceptionally cold meteorological conditions throughout the whole winter led to a record total ozone loss of about 40% (Manney et al. 2011). At a global scale, evidence was also found of a thinning of the stratospheric ozone layer due to elevated ODS abundance. Ozone decrease was observed up to the end of the 1990s in the lower and higher stratosphere. According to the last scientific assessment of the ozone layer, current average ozone levels are about 3.5% lower than those measured before 1980 over the 60°S to 60°N latitude range. The decrease is largest in the Southern Hemisphere mid-latitudes with a reduction of 6%, compared to 3% in the Northern Hemisphere. Ozone reduction in the tropics is very small (WMO 2014).

In order to regulate the emission of ODSs, the Montreal Protocol on Substances that Deplete the Ozone Layer (MP) was signed on September 16, 1987, in Montreal, Canada. It entered into force in 1989 and was amended several times up to 2007. The MP is the first environmental treaty to reach global ratification (UNEP 2015). Due to its widespread implementation, it has been hailed as an example of exceptional international cooperation. The protocol and its successive amendments have led to the near complete phase out of the production and consumption of ODSs in developed countries. The main regulated compounds

Figure 2  Average total ozone in September 2014. Reproduced from NASA Ozone Watch.
Stratospheric and Tropospheric Ozone

are chlorofluorocarbons CFC-11 (CCl\textsubscript{3}F) and CFC-12 (CCl\textsubscript{2}F\textsubscript{2}), carbon tetrachloride (CCl\textsubscript{4}), halon 1211 (CF\textsubscript{2}BrCl), halon 1301 (CF\textsubscript{3}Br) and methylchloride (CH\textsubscript{3}Br). These products have been replaced by hydrochlorofluorocarbons (HCFCs), which are much less toxic for the ozone layer. HCFC molecules contain a hydrogen atom and are thus mostly dissociated in the troposphere. The hydrofluorocarbons (HFCs), substances that are harmless for the ozone layer, now replace HCFCs: unlike CFCs and HCFCs, HFCs do not contain chlorine that depletes the ozone layer. Thanks to the MP and its amendments, the abundance of ODSs in the atmosphere is now decreasing after a peak reached in the mid- to late 1990s, depending on latitude. Yet, due to their long lifetime in the atmosphere (e.g., 100 years in the case of CFC-12), the return to levels below the abundance observed in 1980 will not occur before 2040. The recovery of the ozone layer will thus follow this path, and ozone levels are projected to increase slowly in the course of the twenty-first century.

Stratospheric ozone depletion and climate change

Climate change and ozone depletion are separate yet intrinsically linked issues: climate change influences ozone recovery while ozone depletion has influenced climate change in the troposphere, mainly in the Southern Hemisphere. Increased greenhouse gas abundances induce a cooling of the stratosphere. This cooling is caused both by less infrared terrestrial radiation reaching the stratosphere as it is absorbed in the lower atmosphere and larger infrared emission to space by carbon dioxide in stratospheric thermal conditions, which cools the local environment. Ozone depletion itself has induced a cooling of the stratosphere, since the formation of ozone heats that region. In the polar stratosphere, the cooling enhances the formation of polar stratospheric clouds, thus reinforcing ozone destruction processes. In the higher stratosphere, the cooling slows down ozone destruction reactions and acts as a self-healing effect, which accelerates ozone recovery. Climate change can also affect the meridional circulation that naturally transports ozone from the tropics, where it is formed, to the higher latitude regions. Current climate model simulations indicate that climate change will accelerate this meridional circulation inducing a super-recovery of ozone at high latitude and a decrease of ozone in the tropics. On the other hand, ozone depletion has altered the surface climate, particularly in the Southern Hemisphere. The most notable effect is a southward shift of the jet stream in summer. In addition, winds above the austral ocean have intensified and contrasted temperature trends have been observed on the Antarctic continent in summer, with a warming across the peninsula and a cooling on the plateau (WMO 2014).

Tropospheric ozone

As in the stratosphere, tropospheric ozone is formed from the recombination of an oxygen atom with an oxygen molecule. But in the troposphere, the main source of oxygen atoms is the photolysis of nitrogen dioxide (NO\textsubscript{2}) by visible solar radiation (wavelength \(< 400\, \text{nm}\) since energetic UV radiation likely to dissociate oxygen molecules does not reach the troposphere. In the absence of nitrogen oxide, ozone is rapidly destroyed by deposition to the ground or via reactions involving water vapor or by UV radiation with wavelength larger than 320 nm.

The abundance of ozone throughout the troposphere is generally much lower than in the stratosphere, for example less than 100 molecules
per billion of air molecules (100 ppb). In the free troposphere (3–10 km altitude), large temporal variability of ozone is also observed but on the scale of a few days. This variability is due to dynamical processes that transport ozone from rich air regions such as the stratosphere or the polluted continental boundary layer. Long-range transport between continents also plays a role since tropospheric ozone has a lifetime of several days and can thus be transported away from its source regions.

Although ozone protects human health by filtering out the harmful solar radiation, it is also a strong oxidant, which causes problems with breathing. Ozone has been linked to tissue decay and cell damage by oxidation. Inhalation of ozone can exacerbate lung diseases such as emphysema, chronic bronchitis, and asthma. Ozone also affects vegetation through oxidative stress. It can cause damage in trees and decreases in agricultural yields. At ground level, ozone is produced by reactions involving nitrous oxides (created by combustion in industry and vehicles), sunlight, and volatile organic compounds (VOCs). VOCs are carbon-containing compounds that evaporate readily. They are emitted by a variety of sources that include gasoline vapors, chemical solvents, emissions from vegetation, and household cleaning products. Large concentrations of ozone can build up in summer during so-called photochemical smog episodes that are characterized by conditions of low wind and high levels of sunlight. Ozone is then considered “bad ozone,” in contrast to the “good ozone” in the higher atmosphere that protects life from UV radiation. In the presence of large ozone concentrations at ground level (e.g., 90 ppb), pollution alerts are issued reducing urban traffic or urging people with respiratory problems to remain indoors. Various environmental agencies have set programs for the reduction of nitrogen oxides and VOC emissions from vehicles and industry in order to improve air quality in urban areas.

During the twentieth century, tropospheric ozone levels increased substantially, particularly in the continental atmospheric boundary layer in the Northern Hemisphere (Parrish et al. 2012). This positive trend is attributed to increased emissions of nitrogen oxides and hydrocarbons by human activities. Present estimates from chemical transport models indicate that 60–80% of the tropospheric ozone burden results from chemical production in the troposphere while the remaining part is linked to air mass transport from the stratosphere. Uncertainty remains regarding the magnitude of increase from the preindustrial era since very few routine ozone measurements were made before the late 1970s. Early measurements showed values around 10–20 ppb prior to the 1950s over Europe compared to 40–50 ppb in the 1990s (Marenco et al. 1994). For the most recent decades the global picture from surface station observations is not uniform. Increases are largest from the 1970s to the mid-1980s and tend to slow down afterwards. Anthropogenic emissions of ozone precursors are mainly located in Europe, North America, and more recently East Asia. Certain regions have implemented emissions controls since the 1980s, leading to decreases in ozone precursor emissions. As a consequence, tropospheric ozone levels have stabilized and show signs of decrease at some sites in some seasons, particularly in summer.

Atmospheric ozone and climate

Changes in ozone can have a significant impact on climate because ozone absorbs both solar and terrestrial radiation, and plays a major role in determining the energy balance of the troposphere and the stratosphere. Ozone acts as
a greenhouse gas by absorbing longwave terrestrial radiation. The global average radiative forcing due to increases in tropospheric ozone since preindustrial times is estimated to have enhanced the anthropogenic greenhouse gas forcing by about 0.4 Wm$^{-2}$ compared to a total of 2.29 Wm$^{-2}$ (IPCC 2013). The effect of increased ozone levels is largest near the tropopause (the region between the troposphere and the stratosphere) due to the thermal conditions in this region. In contrast, stratospheric ozone depletion has induced a negative radiative forcing, resulting in a cooling of the surface. But this negative forcing is very small compared to the net positive radiative forcing caused by the emissions of ODSs, which are strong greenhouse gases. It was indeed shown that reductions in atmospheric ODS concentrations have benefited the global climate (Velders et al. 2007).

SEE ALSO: Climate change and health; Environmental policy; Environmental regulation; Global climate change; Polar climates; Temperature

References


Further reading


Stream ecosystems are organized within a catchment network that is characterized by unidirectional flows of water from source to mouth. Stream networks form spatially nested hierarchies, from catchment, stream segment, reach, pool/riffle to microhabitat. Each level of the hierarchy represents the environment for the habitat contained at the lower level, and maintains dynamism of habitat structure and ecological processes over particular spatiotemporal scales, from microhabitats that may reorganize over days to years, to whole-system changes that may take many thousands of years (Frissell et al. 1986).

Flows vary according to water inputs, which change temporally. Streams will typically follow broad annual flow regimes within particular climatic regions, but finer levels of flow variation are also important. Flow variation creates longitudinal (upstream–downstream), lateral (channel–floodplain) and hyporheic (channel–streambed) connectivity that is crucial for the ecological processes that structure stream ecosystems and their communities (Ward et al. 2001). A typical stream structure is given in Figure 1. Such connectivity between stream structural components increases exchanges and flows of biota (e.g., organisms) and abiot (e.g., sediment), and helps to disturb and reorganize physical habitat, increase habitat heterogeneity, disperse species, cycle nutrients, and oxygenate backwaters, among other processes. Less frequent increases in flow that greatly facilitate connectivity are often referred to as “flood pulses,” while more frequent (e.g., daily) minor fluctuations are termed “flow pulses.” Streams may be persistent (with continual, year-round flows), intermittent (with seasonal flows or dry periods), or ephemeral (with flows occurring only after temporary inputs such as rainfall in semiarid areas). “Streams” and “rivers” are governed by the same processes. The two terms are often used interchangeably, though “stream” may (though not universally) indicate a relatively small channel or system.

Streams are among the most biologically diverse ecosystems within a given region. Stream communities have evolved specialisms of physiology and life cycle to live in flowing water (lotic) conditions, which are characterized by hydrogeomorphological variability and dynamism. Each structural component of the river (Figure 1) has particular species and communities that have adapted to the typical conditions associated with the component, though there is much variability within communities. Benthic organisms live on or close to the streambed and may utilize the shelter provided by bed sediments and debris to avoid fast flows. Many organisms are free-floating in the water column, sometimes seeking sheltered areas for particular purposes such as spawning. Some organisms, particularly macroinvertebrates, may be found mainly on the surface of the water column. Organisms in the riparian zone or floodplain may be tolerant of temporary high-water conditions and may display a gradient of tolerance to fluvial disturbance, from primarily aquatic species adjacent to the channel to those that are largely terrestrial, with increasing distance or elevation from the...
STREAM ECOSYSTEMS

channel. Hydraulic conditions are important for structuring communities, with plants, for example, being found mainly in areas where flows are lower and disturbances reduced. Backwaters and other areas where hydraulic stress is reduced may display different communities to within the main channel.

The most abundant and diverse organisms in stream ecosystems are microbes, meiofauna, and macroinvertebrates, with plants and vertebrates being less common but often important in structuring communities and food webs (particularly fish). Communities change with longitudinal position along the river, reflecting change in flow conditions and the physical environment. This is termed the river continuum concept (RCC), and suggests that headwater (upper catchment) areas are dominated by “shredder” species that break down the large amount of coarse plant material that enters the channel from outside the stream (allochthonous). Middle reaches have an increase in internally produced (autochthonous) fine organic material, and communities are dominated by collector and grazer species, while downstream reaches have a high abundance of collectors. The RCC has been regarded as an oversimplification of longitudinal trends, with other highly variable factors such as channel hydraulics being particularly important in structuring the distribution of species.

Historically it was assumed that biota were primarily governed by the physical processes found within streams, and simply responded to the hydrological and geomorphological conditions present. More recently it has been recognized that a more reciprocal relationship between

Figure 1  Simplified cross-section of a stream, highlighting broad habitat types.
STREAM ECOSYSTEMS

organisms (particularly plants) and physical processes exists, with (for example) vegetation reinforcing banks and bar formations by trapping sediment and preventing erosion, and thereby influencing stream morphology. Some organisms may have particularly dramatic impacts on stream ecosystems, such as beavers (Castor spp.) being responsible for the creation of wood dams and areas of wetland habitat that have increased depths and reduced flows, which would not exist without their presence. Such organisms are termed “ecosystem engineers.” Streams have therefore been fundamental in highlighting biogeomorphological processes occurring within ecosystems.

Streams supply important ecosystem services, such as provision of food, water, energy, transportation, species habitat, nutrient cycling, recreation and spiritual wellbeing. Despite (or because of) this, they are among the most degraded ecosystems in the world (Dudgeon et al. 2006). Particularly significant impacts include regulation for hydropower, navigation, or irrigation; urbanization; industrial pollution; and colonization by alien invasive species. These impacts are especially problematic when they interrupt the connectivity of stream networks and compromise essential processes. As a result of extensive global degradation of stream ecosystems, there are significant efforts in many parts of the world to improve or restore their condition.

SEE ALSO: Biogeomorphology; Disturbance in biogeography; Fluvial depositional processes and landforms; Fluvial erosional processes and landforms; Freshwater resources: past, present, future; Hydrologic cycle; Rivers and streams; Riparian ecosystems; Surface water

References


Streams, gaining and losing

Woonsup Choi  
*University of Wisconsin–Milwaukee, USA*

Water moves both horizontally and vertically across the Earth, between the atmosphere, lithosphere, hydrosphere, and biosphere. Vertical moves include precipitation, evapotranspiration, infiltration, and percolation. Surface and subsurface flows can be viewed as horizontal moves. Surface flows occur according to the elevation gradient, and stream channels are conduits for water on the land surface. Naturally, with a cross-sectional view, a stream channel has lower elevation than its surroundings (Figure 1).

Water in stream channels is supplied from different sources, such as atmosphere (in the form of precipitation), snowpack (in the form of snowmelt), and subsurface storage (in the form of groundwater flow), depending on the season and geographic location.

Underneath the land surface, water is stored in unsaturated and saturated zones (Figure 1). Water in the saturated zone is often called groundwater, and an underground geologic layer saturated with groundwater is called an aquifer. A saturated zone that is not overlaid with an impermeable layer and thus interacts with the unsaturated zone (like in Figure 1) is called an unconfined aquifer. The unsaturated zone (also called the vadose zone) is located above the saturated zone, right underneath the land surface, and the boundary between the saturated and unsaturated zones is called the water table. The water table is an undulating surface; it fluctuates in response to recharge and discharge of water to and from the saturated zone. For example, it rises with rainfall, and falls with an extended period of no rainfall.

Water in the saturated zone flows in the direction of maximum hydraulic gradient. In other words, it flows in the direction where the water table falls most per unit distance. In general, the water table will slope toward lower elevations of the land surface, such as streams or lakes.

If the water table is coincident with the surface of the stream and the groundwater flows toward the stream (Figure 2), the stream is supplied with water from the saturated zone, and is called a gaining stream. It is also called an effluent stream when the saturated zone is used as a reference point. In other words, water “flows out” from the saturated zone to the stream. Gaining streams generally have increasing discharge as water flows downstream.

If the water table is not coincident with the surface of the stream or is coincident but lower than the stream surface, the stream loses water to the saturated zone, and is called a losing stream (Figure 3). In this case water “flows in” to the saturated zone from the stream and is also called an influent stream. Because water is lost to the saturated zone, stream discharge tends to decrease as water flows downstream in a losing stream.

The interaction between stream water and groundwater takes place in the hyporheic zone in which saturated sediments lie beneath the streambed and stream banks. Stream water flows through these short segments of its streambed and banks. The hyporheic zones have varying depths depending on the composition and size of the streambed, from several centimeters to more than one meter. This is where exchanges of water and material take place between stream...
STREAMS, GAINING AND LOSING

Figure 1  Water table is the boundary between two subsurface water storages, the unsaturated and saturated zones (Winter et al. 1998). Reproduced from US Geological Survey. © Department of the Interior/USGS.

Figure 2  A: Cross section of a gaining stream. B: Water table contour and groundwater flow off around a gaining stream (Winter et al. 1998). Reproduced from US Geological Survey. © Department of the Interior/USGS.

Figure 3  A: Cross section of a losing stream. B: Water table contour and groundwater flow off around a losing stream (Winter et al. 1998). Reproduced from US Geological Survey. © Department of the Interior/USGS.

Water flowing above and groundwater sitting below, and is receiving increasing recognition for its importance for aquatic organisms, such as fish during their spawning.

The definitions of gaining and losing streams have somewhat different nuances between references. In Dingman’s book (2002), a losing stream “is one in which discharge decreases downstream” (p. 344). Dingman’s book says the reason why the stream loses water downstream is because the water table is lower than the stream surface. The definition is sometimes considered as sinking streams in karst topography, which disappear underground as losing streams. On the
other hand, other sources define a losing stream as one that loses to the saturated zone (e.g., Cech 2010). The difference is whether the focus is given to stream discharge changing downstream or interaction between surface and groundwater, even though they are intertwined.

Gaining streams have water flowing year-round or for extended periods such as wet seasons. Losing streams tend to have flow only in response to water-input events like rainfall and otherwise have no flow because water is lost. Streams with year-round flow are called perennial streams, and streams with flow only during wet seasons are called intermittent streams. Streams that flow only in response to precipitation are called ephemeral streams.

Gaining and losing streams can occur in different reaches of the same river. Correspondingly, gaining and losing reaches can be distinctly defined. A stream encounters a variety of geologic characteristics, land covers, human activities, and even climates as it flows from headwater to outlet, thus gaining and losing reaches can appear alternatingly. For example, if a well is drilled near a stream, the water table will be lowered and the stream reach can become a losing reach. In some cases, streams switch between gaining and losing by season or during floods. Covino and McGlynn (2007) suggest that streams do not simply lose or gain over a particular reach, but the gain/loss is spatially and temporally dynamic. The switch between gaining and losing conditions is particularly variable in headwater areas where the location of start of streamflow can move up- or downstream, depending on snowmelt or rainfall events and groundwater recharge (Winter et al. 1998).

Because the flow between the streambed and the saturated zone occurs in opposite directions between gaining and losing streams, the pore size and content of fine particles in streambeds, thus hydraulic conductivities thereof, have different characteristics. Downward movement of water in a losing stream pulls fine particles to the bed, resulting in clogging, whereas upward movement of water in a gaining stream lifts fine particles and expands the pore space. Therefore, it is likely that the streambed hydraulic conductivity of a losing stream is smallest at the top layer of the streambed and increases with depth. On the other hand, that of a gaining stream is likely to be greatest at the top layer and to decrease with depth (Chen et al. 2013). These hydraulic conductivities were demonstrated in a field study (Chen et al. 2013) of the Platte River Valley in Nebraska.

The importance of the hyporheic zone lies not only in water exchange but also in material exchange. Contaminants released into a losing stream will result in contamination of both the stream and the underlying aquifer (Field 2006). Nutrient exchange also occurs in the hyporheic zone with opposite characteristics between gaining and losing reaches. Harner and Stanford (2003) found faster growth of cottonwood trees (Populus trichocarpa) in a gaining reach than in a losing reach in a floodplain as a result of water availability and nutrient (nitrogen and phosphorus) delivery from regional upwellings of hyporheic water. Because the saturated zone interacting with streams can underlie the land surface, the biochemical conditions in a given reach reflect interactions between water in the hyporheic zone and groundwater from distant sources (Winter et al. 1998).

Good examples of gaining streams include the Mississippi and Columbia Rivers in North America and the Amazon in South America, and many other rivers in humid climates. Losing streams are far less common than gaining streams and mostly found in arid climates (Strickler 2012). Losing streams are also common in regions of karst topography, along with sinking streams (Field 2006).
The Colorado River in North America is a good example of a losing stream. It begins in the Rocky Mountains, flows generally southwest, and drains to the Gulf of California in Mexico. It used to have abundant flow, discharging approximately 18500 million m³ each year into the Gulf of California (Flessa et al. 2013). However, it no longer regularly reaches the Gulf due to a range of reasons, such as dam building, overuse of water, and droughts (Zielinski 2010), and the flow crossing the border into Mexico today is only about 1850 million m³ each year (Flessa et al. 2013).

SEE ALSO: Base flow; Groundwater; Rivers and streams; Surface water; Water and climate change; Watersheds

References

Structuration theory was arguably the most influential of theoretical imports into human geography during the 1980s. Its influence and allure came as a direct result of a capacity ascribed to structuration theory to transcend the binary polarity between “structure” and “agency,” which at the time dominated theoretical discussions in social theory and in human geography. At the same time, it also marked a considerable incorporation and adaptation of geographic discourses and practices by non-geographers. The result was a two-way street of theoretical traffic between geographical discourses and associated practices on the one hand and social theory broadly construed on the other. The ensuing circulation of concepts helps explain the lasting appeal of structuration theory within geographical discourses as well as the continuing interest of social theorists in space and spatially-resonating concepts. As such, structuration theory was thus very much part of the “spatial turn” taking roots across the social and human sciences during the 1970s, the legacy of which is still very much present today.

In both the social sciences and human geography, structuration theory addressed a key issue that had vexed the work of both theory and practice, namely the question of how to relate the working of structural elements (state, companies, institutions) with those of individual or collective agents in a coherent manner without prioritizing one over the other. In social theory and philosophy, the problem was not unknown and had motivated the writings of many authors from Marx to Weber and beyond; however, it was the work of Anthony Giddens from the late 1970s onwards that provided a genuinely novel answer in not starting to theorize from either “agency” or “structure” and work towards a synthesis but by seeing both “agency” and “structure” as always already the product of a wider dialectical process that predates whatever concrete empirical shape either of these would take (Giddens 1984). It was at this pre-existing junction between “agency” and “structure” that the eponymous “structuration” took place: a formatting process that theorizes the mutual constitution of individual and society through institutionalized (and thus routinized) practices. Giddens characterized this process with reference to a quote from Marx’s *Eighteenth Brumaire of Louis Bonaparte*: “Men […] make history, but not in circumstances of their own choosing” (Giddens 1984, xxi). Mention of Marx furthermore allows us to state that, in a broader sense, structuration theory can be read as an extension of older epistemological questions that have been central to social theory at least since the nineteenth century. Marx’s elaborations on and crucial differentiation between what he called “formal” and “real” subsumption, itself related to Hegel’s dialectical philosophy, are perhaps the most obvious antecedents. Here as there, social transformation is theorized in a nuanced and locally resonating manner, shifting the analysis from one focused on abstract structures to their emergence and sustainability within concrete local settings and practices. In contemporary social theory, a close relative to the notion of “structuration” can thus be found in Pierre Bourdieu’s concept of “habitus.”
STRUCTURATION THEORY

It should hence come as no surprise that many geographers began to embrace structuration theory from the early 1980s onwards. Following a decade that was typified by theoretical responses to a perceived hegemony of “quantitative” forms of geography (which had developed in the aftermath of World War II), responses that by-and-large emerged from the two distinct philosophical traditions of Marxism and phenomenology, geographers, too, were eager to embrace nondualist modes of explanation (Gould and Strohmayer 2004). This willingness to part company with at the time established modes of explanation was motivated less by shortcomings identified within these existing theories (most commonly labelled “determinism” and “voluntarism”) (Thrift 1983, 25) but by a Zeitgeist best characterized by the cultural and “spatial” (or “linguistic”) turns then prevailing in the human sciences. Disaggregating these “turns” allows us to conceptualize the latter or “spatial” turn as directly motivating less hegemonic, “functionalist” or determinist theoretical narratives while the former or “cultural” turn implicated arguably more nuanced approaches to questions addressing issues of social relevance. Both were united, however, in their theoretically motivated rejection of stable, aspatial structures, most particularly of the dualist kind, of which the artificial but long-established separation between “structure” and “agency” was perhaps but the most obvious of expressions. In geography, the inability of the spatial science tradition conceptually to embrace nondeterminist approaches supplied a further easy target to unite towards (Gregory 1994, 119).

But the appeal issued by a decidedly nondualist approach only partly explains the popularity of structuration theory throughout the 1980s; equally important was the fact that structuration theory incorporated concepts and methods known to geographers into its explanatory frame. Chief among these must count the interest accorded to “time geography,” as developed by Torsten Hägerstrand in the 1960s, especially by Anthony Giddens. The visualization of “constraints” and “abilities” as combined in the form of time-geographic “life-paths” provided a welcome means to articulate the ever-present (and thus original) structuration of social action; for geographers, who were just then learning to appreciate the fact that other social and human sciences had discovered space, the nod to the work of Hägerstrand and others provided further incentive to engage with structuration theory. At any rate, the result was a more thoroughly aligned surface of theoretical possibilities where geographers conversed effortlessly with other social scientists.

Beyond academia and given its overall thrust, it is not surprising that the political corollary of structuration theory came in the form of a reformulated social democratic program in the 1990s and early 2000s: heralded initially as a “post-Marxist,” “third way” and supported by Anthony Giddens and the creation of Polity Press as a sympathetic publishing outlet, structuration theory proved to be amenable to reformulated progressive politics precisely because it was eminently flexible, adaptable even (Hannah and Strohmayer 1991). Of course, this involvement (or political use-value) can also be used to critique structuration theory as a rather “toothless” form of social theory – or, more pertinently, as asking as many questions as it could answer within its own stated epistemological frame.

Since the height of its visibility within human geography in the mid-to-late 1980s, structuration theory has lost considerable sheen: its explicit invocation, even by those who regularly trumpeted its virtues, has significantly diminished since the turn of the last millennium. In a less openly discernible manner, however, all geographers inhabit a world profoundly shaped
by structuration theory, a world characterized by an absence of attempts to explain significant social, cultural, or economic developments or events through recourse to theoretical explanations based singularly on either “agency” or “structure.” Virtually all key terminology that has since dominated (or more modestly sought to contribute to) theoretical discourses in the social sciences and in geography implicitly accepts the mandate bequeathed by structuration theory: to conceptualize its work and workings in an explicitly nondualist manner. Furthermore, the proliferation of “site-specific” material, performed and explicitly “scaled” forms of geographical knowledge since the 1980s similarly owes more than a passing nod to structuration theory given its insistence on the importance of “local” anchorage for processes of structuration. If larger social and historical processes only acquire momentum by being embedded within specific local configurations, research on the latter becomes of paramount importance not because it uncovers something specific about such sites themselves but because it contributes knowledge that is essential to any understanding of social processes at large while accepting that those social processes remain no less essentially different in dissimilar places or regions.

**SEE ALSO:** Cultural geography; Habitus; Marxist geography; Phenomenology

---

**References**


**Further reading**


“Subaltern” is an emotive, elastic, and elusive term and critical concept pertaining to issues of domination and subordination, voice and agency, and the domineering way in which Western discourses relate to other cultures and ways of knowing. Geographers have characterized “subaltern space” (Clayton 2011) and “subaltern geographies” (Jazeel 2014) as the media and outcome of relations of power and identity formation that split the world and its inhabitants into elite (or superordinate) and subaltern (or subordinate) sectors and locations, and that give the condition these processes fashion – subalternity – both general and specific qualities.

The term has two meanings: first, a person, group, or entity of subordinate status; and, second, a junior army officer (introduced into nineteenth-century European armies to convey the orders of military leaders to troops). The formative theorist of the subaltern, the Italian communist Antonio Gramsci, deployed the term in both ways. He used it to signal a concern with the most oppressed, disadvantaged, and voiceless sectors of society (he had in mind a largely peasant southern Italy), who, in his remarks on subaltern history in his *Prison Notebooks*, are the group that take the orders and “are always subject to the activity of ruling groups, even when they rebel and rise up” (Gramsci 1971, 55). He also used “subaltern” as a surrogate term (in part, to get his writing past prison censors) that flipped the military idea of the subaltern as a medium of communication and re-envisioned this subject as pivotal to a broader “war” over social power in fascist Italy. That war was over how charismatic leadership and dominant control in society were effected, and how, for Gramsci, this was achieved by consent as well as coercion (creating what he termed “hegemony”), with intellectuals playing a vital link between the elite and subaltern masses akin to the link role that a subaltern army officer plays in a military chain of command. In other words, Gramsci’s subaltern was a code word for resistance and the difficulties of realizing it.

Partly because of the unfinished quality of Gramsci’s treatment of the subaltern, he bequeathed the term a prolific and pliable afterlife. The subaltern has variously been conceived as an object of study, a form of agency and resistance, and a mode of consciousness and critique. It has been approached differently through modalities of gender, sexuality, race, culture, and socioeconomic differentiation (class and caste), and has different connections to debates about modernity, colonialism, historicism, nationalism, Eurocentrism, and now globalization (and related problematics of governance, aid, humanitarianism, cosmopolitanism, minority status, and risk and conflict). Questions of subalternity have been studied at different scales (with the subnational scale long privileged, but with local–global articulations, and a language of “networks” and “assemblages” currently in vogue), and taken on different inflections in different fields of theory and study (history, anthropology, geography, literary history, Marxism, feminism, postcolonial studies) and different parts of the world (especially South Asia and Latin America).
Since Gramsci, the term has been conferred on a wide range of groups – peasants, workers, the poor, women, indigenous peoples, the colonized, slaves, refugees, asylum-seekers, and religious and ethnic minorities – and is often used loosely (to the point of losing its critical efficacy) as a generic expression for oppression and disenfranchisement on the one hand, and for emergent and insurgent knowledges and practices on the other. The term is often equated with the project of rereading “history from below” and with politico-intellectual anxiety over who has the right and authority to speak for “the other”: however, it is usually reserved for non-Western, colonial, and developing-world configurations of domination and subordination, and has played an important role in radical global history and politics, naming the pursuit of a non-Western and nonindustrial revolutionary subject (other than “the proletariat”), and thus serving as a means of questioning the primacy of class struggle in Marx’s theory of history, and becoming linked to a family of terms that stem from anticolonial and communist-internationalist thought and praxis (“the people,” “the masses,” “the popular,” “the multitude”).

During the 1980s and 1990s Gramsci’s idea was popularized and reconfigured in relation to Indian history by Ranajit Guha and the “subaltern studies” group and book series he spearheaded. The founding aim of this group (of mainly Indian and British historians) was to challenge the elitist biases in the colonial, nationalist, and Marxist historiography of India – its failure to speak for India’s peasant masses. Guha used the terms “the people” and “subaltern classes” interchangeably, and argued that Indian nationalism (both pre- and post-Independence) was not simply the product of elite initiative. He contended that, while subaltern resistance to the British was not strong enough to generate a nationwide struggle for liberation, peasant consciousness and resistance were not backward or immature in the way that European Marxist historians had described preindustrial peasant rebellion in Europe. The subaltern studies group questioned the class determinism and evolutionism in Western Marxist historiography and found in Gramsci’s work a complex figure of subordination and resistance with no easy route to historical recognition or political representation. The peasant insurgent in colonial India had been written out of history (captured and denigrated by elite discourse, including that of the Communist Party of India). Guha invested the subaltern classes of India with political agency, and in a structuralist vein claimed that the (minority) elite and (mass) subaltern sectors of society ran “on parallel tracks over the same stretches of history, as mutually implied but opposed aspects of a pair of antagonistic consciousnesses” (Guha 1983, 11). Reading hundreds of cases of peasant resistance to the British between 1783 and 1947, the subaltern studies group identified a subaltern will and logic to resistance that could be excavated from colonial texts, recovered from Indian sources, and used to show that the peasant insurgent was not an ineffectual relic in a modern (industrial, colonial) world.

Moving now from how the subaltern is configured as a subject/class to how it serves as a category/epistemological figure, theorists underscore the relational qualities of the category. The term enjoins us to specify how subordinate groups are connected to power and “hegemony” in specific times and places. But, from Gramsci onward, the subaltern has usually been conceived as a bifurcated (and, as shall be seen, paradoxical) category that has been politicized both negatively, in terms of what it is not, and in positive and prospective terms, as an oppositional figure of creative resistance. The subaltern is a figure of both constraint and release – of what power and knowledge can and
cannot grasp and control, what power clasps in its
own image and making, and what it excludes and
casts to one side. The subaltern has been located
both outside (exterior to) and at the margins
of (but still inside) fields of power, and as both
separate from, and an effect of, power. Geogra-
phers are not the only scholars to examine how
this bifurcated configuration of the subaltern is
shaped in and through space – through politically
charged sociospatial processes of partitioning,
segregation, classification, surveillance, targeting,
exclusion, expulsion, and hierarchization oper-
ating at a range of scales (for instance, at border
patrols; in ethnic enclaves, refugee camps, slum
districts, war zones, migration flows, and national
and international citizenship strictures; and in
terms of unequal access to wealth and welfare).

There is a core tension in conceptual work
on the subaltern between what might be called
“interiorizing” and “exteriorizing” visions of
how the subaltern generates a politics of knowl-
dge that is geared to the question of how
hitherto ignored and disqualified knowledges and
ways of life might be approached and analyzed.

This exterior–interior split in the subaltern’s
location frames the question of whether the aim
of subaltern critique and resistance is to identify
and protect the subaltern’s exteriority, or to dis-
solve the subaltern’s interior marginality. Some
subaltern projects are underpinned by the desire
for autonomy, critical-intellectual investment
in seeing the subaltern as untouchable, and the
search for a subaltern history, space, and politics
that can be deemed a record and indicator of
these drives. Others are framed by the idea that
the subaltern is a marginal, incomplete figure
that can only ever be spoken for, and that cannot
simply be plucked from the past or present as a
full-blooded subject and used to turn the tables
on domination.

Conceptualizing subalternity in locational
terms – that is, as meaning different things in
different times, places, and projects – generates
the worry that the subaltern can be broken into
so many diverse and localized pieces that it loses
conceptual coherence, comparative utility, and
critical bite. Subaltern worlds and agendas can
become closed in on themselves – scattered,
disconnected, and incommensurable. The whole
idea of the subaltern is also disheartening in
that it names an abject space of domination and
a closed space of difference. Yet the subaltern
also names an anticipatory space of opposition
from which, it is hoped, subordination can be
assuaged, if not overturned. It is common to find
the subaltern invoked in this double sense: as a
space of denial and denigration, in which the
subaltern is assigned a place in society by elite
power and its norms, and granted an identity
that marks her or him as subordinate and adrift;
but also as an alternative and counterhegemonic
space in which the desire and ability to challenge
power are imagined and enacted. Geographers
view the subaltern in this twofold way, as both
a figure of oppression and a liminal space of
becoming and critical position “on the mar-
gins.” They thus enact the “classic fissure,” as
Timothy Brennan (2001, 164) sees it, in the sub-
altern literature, “between opposition as action
and opposition as a demand for recognition.”

Two examples to fill in this picture will need
to suffice. First, the image of the subaltern as
an essence (a constant and compensatory oppo-
sitional presence in history) cultivated by the
subaltern studies group has solicited extensive
debate. In being “understood as the essential
resistance of the voiceless,” Brennan (2001,
163–167) remarks, the subaltern is “shielded
from an imposed and always arrogant rational-
ism”: he is kept on the outside; but by the same
token, it is “a life privileged only insofar as it
remains subaltern.” Consequently, the subaltern
“would seem to represent a sacred refuge, a
dark space of revelation … [and an] occult
subalternity is conceived as theory’s deepest and most irreducible value.”

Second, Gayatri Chakravorty Spivak is perhaps the best-known theorist of the subaltern today, and the doubleness glossed over in the previous paragraph is captured in her characterization of subaltern space as “differential space” – on the one hand, naming a “position without identity” (albeit for her a profoundly gendered figure) and without a basis for action, naming a group “whose identity is its difference,” that is “cut off” from society and history, with no access to the structures of the state and citizenship; but, on the other hand, declaring a critical intent “to clear a space” in which dominant groups and privileged figures and institutions (including, of course, academics and universities) can learn how to listen to and work with subordinated people and groups (Spivak 2005, 476–479; also 1999, 272–308). Her 1988 essay “Can the Subaltern Speak?” is one of the most cited papers in the critical humanities of the past 50 years, and has had an important impact on geographers. Her core concern with what she terms “epistemic violence” (1999, 272–308) – the predilection of language and knowledge not only to “decipher” but also to “intervene,” “cage,” and “dominate” – fuels a staple concern that attempts to “speak for the subaltern” can be self-defeating. It is through this problem that geography students generally encounter the subaltern in textbooks (e.g., Sharp 2009, 110–120).

But subaltern themes cast a longer shadow over the discipline of geography, and have been broached in eclectic ways. They can be found in critical histories of geography that alight on the Eurocentric blinkers and conceits that imbue the discipline’s modern development and contemporary (chiefly Western) hegemonic forms. Geographers here use the subaltern as a code word for the need to decenter geography’s imperial maps of meaning and challenge derogatory images of otherness in both popular and academic geographical writing and representation. There are also subaltern threads to research on postcolonial and development geographies that endeavor to “learn from other regions” (to use David Slater’s expression), which work on spaces and networks of conflict, terror, risk, and security; and in radical (oppositional, activist, and participatory) geographies that are concerned with social movements and antisystemic projects. Finally, in a more vexed way, the image of the subaltern as space on the margin shadows geographers’ incessant search for novelty and variety: their cultivation of ever new “margins of study” and quest for unencumbered spaces in which new freedoms (at least of inquiry) might be found (albeit often through a surreptitious romancing of resistance).

Geographers have paid considerable attention to the historical and containment and constraint dimensions of the subaltern concept, although there is also a range of geographic research on subaltern dynamics today that is yielding an inventive geographical vocabulary, of “convergence spaces,” “translocal assemblages,” “subaltern networks,” and “subaltern cosmopolitanism,” and that advocates a subaltern politics of place and space, yet without lapsing into romanticized and reactionary shibboleths of nativism, and its geographical corollary, place-bound identity (see Jazeel 2014).

Cultural and historical geographers have a strong interest in the processes of subalternization at work in different imperial and colonial projects, discourses, practices, spaces, and networks – in plantations, native reserves, missionary activity, projects of mapping, naming, and scientific classification, and colonial practices of settlement and native dispossession. Subaltern themes also influence a wider range of work on modernity’s historical geographies. Modernity
generated new distances – between worker and product, producer and consumer, city and country, knowledge and the conditions of its making, viewer and object, self and other, and representation and reality – and its self-image as an ideology of “the new” involved the creation of specific locales and spatial markers that tied ideas of progress to those of distance and absence. Ulf Strohmayer (2003, 523) and other geographers have noted that the epistemological question of how European knowledges about “there” and “them” are produced “virtually disappears behind their claim to validity” – as detached and disinterested, rational and scientific, and civilizational, and so on, and in binary opposition to non-European knowledges, which are represented as backward and superstitious. Geographers have opened up how European “objectivity” hinged on the denial of coevalness (two-way encounters and flows of ideas and influences between metropole and colony), the promulgation of sweeping imaginative geographies (with their stereotypes about ignorance and savagery), and distinct spaces (such as the ship, laboratory, map, museum, learned society, and exhibition) where European authority was shaped and secured. The same interior–exterior tension regarding the subaltern applies to this Eurocentric imperial worldview. The West justifies colonization on the grounds that the beneficent colonizer seeks to civilize the benighted subaltern (“native” and “subaltern” are sometimes used interchangeably) by bringing “gifts” of science, civilization, and material and social progress (in some discourses it is the native subaltern who beseeches them). But, at the same time, if colonialism is to endure and remain legitimate in Western eyes it can never admit that it has completed its job; there is a constant need for a subaltern to be worked on by European reason and its colonial scripts. As a large body of postcolonial theory contends, the colonial subaltern becomes a conscript of civilization but is ultimately always denied access to modernity’s full liberating potential and told to follow orders in this Gramscian subaltern way.

Finally, and as much of the preceding discussion suggests, subaltern space may be construed as a paradoxical space. An elementary paradox is of inside and outside: the idea of the subaltern as separate, and immune from, yet only discernible in terms of, an organizing center of meaning, representation, and power. A potent criticism of much subaltern theory and scholarship is that the desire for a separate subaltern history and identity can only be expedited via those processes of connection and division that have brought putatively autonomous histories, identities (and geographies) into the elite–subaltern relationships that are now a source of consternation and contestation. This core paradox leads to a related one, of language and progress, acutely apparent today in subaltern struggles over land, cultural recognition, citizenship, and ecological and political security. The ability of subaltern groups to articulate their grievances is seen as a sign that they are no longer subaltern, for such grievances are recognized only if they are articulated using the language of hegemony. In other words, do subalterns have to abandon or devalue their languages and lifeworlds in order to undo their subalternity? This question has a pivotal bearing on debates about the remit and responsibilities of a postcolonial (and post-development) geography. A third paradox, of situated knowledge, has a wider bearing on how geographers approach issues of othering and subordination. How, in rendering the subaltern as profoundly “situational,” will the geographer account for the locational qualities of her or his own work, particularly as their discipline has shaped and rationalized Western dominance and fueled exoticism over the centuries? A final paradox, of impossibility, frames subaltern theory
more generally. If, by definition, subalternity connotes a lesser, or outside, position that cannot be unraveled either in theory or with hindsight, why should we want the term? Is the subaltern condemned to always be a subject and category of failure and despair? An awkward, stubborn question remains: Can the subaltern be placed in an unreachable realm of exteriority to power and yet be rendered inside a space of domination enough to provide a recognizable ground for action and politics?

SEE ALSO: Colonialism, decolonization, and neocolonialism; Difference; Power; Radical geography

References


Further reading

Subculture of poverty

David Wilson
University of Illinois at Urbana–Champaign, USA

The controversial theory of a subculture of poverty was first offered by University of Illinois anthropologist Oscar Lewis in his attempt to understand the persistence of poverty in rural and urban settings (particularly in Mexico, Puerto Rico, and New York City). His 1959 book, *Five Families: Mexican Case Studies in the Culture of Poverty*, was his initial foray into this intellectual project. The subculture of poverty theory asserts that the poor adapt to conditions of poverty and marginality by creating a culture or subculture adapted to those conditions that ensure some degree of normalcy and happiness. This culture sustains poverty by normalizing its supporting value orientations: helplessness, the search for immediate gratification, dependency, marginality, and powerlessness. This culture, to Lewis, becomes akin to a prison without walls: for those who bear these powerful feelings and lack the education to see beyond them, a kind of social entrapment follows. Structural forces, then, come to cage and isolate the poor, as these are ultimately self-created forces that bind and permeate a population.

The subculture of poverty theory has remained a widely accepted way to explain the persistence of poverty in the United States and beyond. It has been a dominant influence on US public policy, forming the basis for programs on how to redistribute wealth and resources to the poor since the 1960s. Its shift to the political center was fueled especially by Senator Daniel Patrick Moynihan’s (1965) report, *The Negro Family: The Case for National Action*. Moynihan stated that a pervasive “tangle of pathology” was responsible for the plight of America’s urban poor and that a particular group, the black poor, were most dramatically affected. The fundamental dynamic was the breakdown of the black family which was purportedly afflicted by the decline of the traditional male head of household. To Moynihan, the loss of a key parental influence, the eclipse of an important role model, and the evaporation of discipline transmitted across generations were most crucial. Moynihan argued that the origins of this deviant family structure lay in slavery, where the destruction of the “traditional” family “broke the will of the Negro people,” particularly black males. This sense of powerlessness led to, in effect, a culture of dependency.

Criticisms of the notion of the subculture of poverty have been widespread and virulent. A dominant criticism focuses on its very reading of poverty communities and people: the poor are cast as unduly passive and uncreative in their everyday lives. They seem unable to reflexively and creatively make decent lives for themselves, to accurately read how society operates, or to organize to make their lives and their communities better. Yet, it can also be posited that the poor are frequently adroit readers of their predicament and the broader society and that they only superficially appear as caught in a powerless vice of marginality. Beneath the surface are active beings who seek to remake the forces that engulf them. Following this reasoning, the subculture of poverty theory shifts the blame for poverty from social and economic conditions to the poor themselves. According to this view,
the poor, seemingly, need to be smarter, more responsible for shaping their everyday lives, and more actively involved in civic duties.

A second major and related criticism has focused on a reliance on the assumption that the behavior of the poor derives solely from preferred cultural values. That is, everyday realities, particularly underemployment, unemployment, criminal behavior, dropping out of school, and drug use are the result of behavior freely chosen by individuals. According to this view, these are the choices that people have made and are an accurate reflection of their preferences. Yet, to critics, such behavior may reflect something else, for example patterns of compensatory survival. Thus, people may act in ways that push legality to the limits, for example engage in shadow economies or participate in youth gangs, less because they desire to do so or are following cultural norms, but because there is little choice given their limited education and the job opportunities available to them. Therefore, “ghetto dwellers” may see themselves as having few real options about what they can do to survive materially and psychologically. In the final analysis, so-called ghetto behaviors are adaptive, not normative, and, given sufficient opportunities, individuals within the ghetto would welcome alternative options.

The subculture of poverty thesis continues to be used as a way of explaining the persistence of poverty, particularly its resiliency in urban environments. Among geographers, political scientists, sociologists, and policy analysts, the debate continues as to the core sources of poverty. While the popularity of this notion ebbs and flows over time, its remarkable durability speaks to the concept’s visceral appeal of pointing to the culpability of individuals for their own choices.

**SEE ALSO:** Neoliberalism; Poverty; Race and racism

**References**


**Further reading**


Subjectivity

Patricia Noxolo
University of Birmingham, UK

The term subjectivity has been used in geographical research in two closely interrelated ways: as the antithesis of objectivity in research, and as a way of locating the individual in their social and political context. The unifying factor might be understood most broadly as perspective: subjectivity describes the situated perspective of the individual in society and, more specifically, of the researcher and of the researched in the research encounter. The specific role of geography in theorizing subjectivity, then, is in thinking through the locatedness or the situatedness of the subject: what is the place from which the subject’s perspective comes?

This might seem a rather tortuous question, containing its own answer: subjectivity is the perspective of the subject. If the subject is understood as the Cartesian subject, the rational individual who can say with certainty, “I think therefore I am,” then subjectivity is a product of that rationality. Each human being looks out at the world from the fixed center of the thinking individual. In relation to research, by disciplined application of that rationality, combined with reliable and valid research methods, each individual can come to understand objective truths about the world surrounding them, objective truths that all other rational subjects can also come to understand. Subjectivity, then, is the perspective of thinking subjects who are located in a world of objects that their individual rationality, combined with their collective tools of enquiry, allows them to comprehend. Researchers are, in this limited sense, individually and collectively, masters and mistresses of all they survey.

Even within this Cartesian concept of the rational subject, however, there has always been an understanding that human beings are not simply rational brains. Human beings have emotions and sensory experiences that may radically change the way we see the world, making the perspectives of one subject sometimes very different to the perspectives of other subjects who may have different experiences. Taking this into account, subjectivity can be seen as much more individual than collective. The purpose of rigorous methodological techniques can be to minimize the effects of the subjective bias of individual researchers, often attempting to screen out opportunities for the researcher to interpret the data through experience or emotion. However, this screening out can rarely, if ever, be complete. Indeed, phenomenological approaches to research argue that the outcomes of research are ultimately entirely constrained by subjectivity: that is, that human subjects can only ever know that which we can perceive through the senses available to us. Subjectivity appears always as somewhat in tension with objectivity.

So it becomes clear that the place of the subject is not so easy to locate as it would seem. Even the rational, centered Cartesian subject is located within a body that has sensory and emotional experiences that are not completely governable by the subject. One might say that Cartesian subjectivity is bounded by the body and mind of the individual, but that these boundaries have always been recognized as porous and open to influences that the subject cannot control. Perhaps
as much in response to this internal pressure as in response to historical and academic change, theories of the subject and their subjectivity became much more complicated over the course of the twentieth and early twenty-first centuries, and subjectivity has become much more difficult to locate. Subjectivity has gradually become recognized as more diverse, more fragmented, and more relational, and all three of these qualities make it more open to spatial theory.

The global history of the twentieth century has been marked by an increasing recognition of diversity, for example, in terms of gender, ethnicity, and sexuality. The quest for equality for a range of worldviews and in a range of professional spheres has included recognition of diversity in academic research, both between researchers and between the researcher and the researched. Subjectivity has therefore begun to be understood as diverse – it is not located in a universal capacity for rationality, but is mediated by a range of political and social contexts. Although this insight leads to shorthand formulations of this diversity, such as “black female subjectivities,” this does not mean that individuals’ perspectives are necessarily or essentially determined by particular identity positions; however, it does mean that uniformity across all subjectivities should not and cannot be assumed.

The category of “experience” has often been deployed in understanding subjectivities as diverse. Experience, particularly as routed through feminist geographies, highlights that subjectivities are diverse because people are diversely located: people experience the world in different places and in different bodies. Among the many effects of being in different places is that “interpretive communities” give rise to particular assumptions and interpretations of how the world works (Livingstone 2005, 395) – that is to say, place matters in shaping subjectivities. Similarly, the different meanings inscribed on different bodies – their gendering, racialization, and so forth – make for lived experiences of the world that are both spatially and historically contingent. Bodies change their social and political meanings over time and over space, mediating our experiences, and these mediations make for different subjectivities. So, although this is not a question of essential difference (the same body can be interpreted and experienced differently at different times and in different places), it can be a question of radical difference in that the subjective experiences and perspectives of differently embodied people can be (but are not necessarily) radically different.

As researchers, this possibility of radical difference in subjectivity can push us well out of the comfortable zone of objectivity (in which “error” and bias can be eliminated through methodological rigor) into a more uncomfortable translation zone between different subjectivities, most notably those of the researcher and the researched. Questions of how the subjectivity of the researcher relates to the subjectivities of those being researched become particularly sharp in field-based research. Where development geographers, for example, relocate from well-heeled universities in richer countries in order to study communities in poorer countries elsewhere, the question of power becomes crucial: whose subjectivity is prioritized in research design, data collection, and interpretation? Some argue that it is always preferable to research within one’s own interpretive community for this reason, because the researcher and the researched will only really understand each other within a shared subjectivity (Tuhiwai Smith 1999). Others argue for the possibility of intersubjectivity through participatory approaches, in this way becoming a dialogic process across acknowledged difference demanding from the researcher an ongoing reflexive alertness to power differentials.
Combining these last two points conveys us to the second way in which subjectivity has changed over time: that is, the ways in which subjectivity is increasingly understood as fragmented. Those who argue that subjectivity needs to be shared in the research encounter also routinely acknowledge that there is always some difference within, as well as between, interpretive communities – broadly speaking, for example, women are differently raced and classed, black women have different incomes and sexualities, black lesbians are located in a range of countries with a range of juridical and political regimes, and so on. Intersectionality (Brah 1996) has been a useful tool in pushing understandings of diversity in subjectivity beyond the image of a range of separate interpretive communities to an understanding of subjectivities as internally diverse. In an analogy with recent understandings of places as meeting points or intersections, rather than as bounded parcels of space, individual subjectivity becomes the unstable meeting point of a range of aspects of identity that do not simply juxtapose or appear as a list (black, female, Global North, high-income, etc.), but that interact with and shape each other as experiences and perspectives. Each individual is engaged in an unpredictable dialogue between different aspects of their own subjectivity. At the same time their dialogue with the subjectivities of others is an engagement with both shared and contrasting aspects of others’ identities – understood as unbounded in this way, individual subjectivity is an intersection, with multiple connecting lines with other subjectivities, as well as multiple aspects of difference.

One implication of this fragmentation for the research encounter is that, while there may be radical differences in subjectivity that make it difficult to establish a dialogue between researcher and researched, there may equally be a range of unpredictable meeting points between the subjectivities of researcher and researched. These can shape the research encounter as a fragile, but nonetheless real, moment of intersubjective meaning-making, albeit with no guarantees for prolonged or replicable connection. For example, some have found that shared embodied experiences between some (though of course not all) women, such as pregnancy and childbirth, can form connecting lines that allow for shared understanding in some areas of research, but others have argued that even these apparently unmediated experiences can have so little purchase in the research design and execution, or can be so radically differentiated by diverse experiences of medical and community care or of maternal and political rights, that potential lines of identification are effectively erased (Noxolo 2009).

A second implication of the fragmentation of subjectivity is that the perspectival center of the subject has become radically unstable. There is no central characteristic of the multiple aspects of the researcher’s intersectional identity that necessarily adjudicates truth for the researcher; there is no necessarily universal rationality that steps in to define what is and what is not the right interpretation of the data. In this sense, from the so-called god-trick (Haraway 1988, 587) of comprehending everything everywhere (or at least wherever the right methodological tools were applied rigorously), the perspectival center of the researcher-subject now appears to be nowhere. Even where there is a concerted attempt to carry out the so-called goddess-trick (Rose 1997, 311) of reflexive vigilance toward the circulation of power in the situated research encounter in order to ensure intersubjective meaning-making, there is no centered researcher subjectivity to do this monitoring in an incontestable and coherent way – how much of the power relations in which they are themselves implicated will this researcher really be able to see in their vigilance, how will...
they interpret what they see, and, given their partiality, how will they ensure that this knowledge is deployed in ways that are valuable for both researcher and researched in all their multiplicity?

Partly as a response to subjectivity becoming more diverse and more fragmented, the third way in which understandings of subjectivity have changed is a shift in focus, which can be understood spatially as a shift away from locating the center of the subject toward understanding the spaces between subjectivities, or the spaces of relationality. The metaphor of intersection or meeting place conjures up images of lines of connection, but is less able to assist us in understanding the qualities of different kinds of relationship, or indeed the absence of connection. Work by a range of geographers on aspects of materiality and technological change has led to analogies for relationality that are more intimate and more textured than lines, analogies that allow researchers to reinterpret the research encounter as intersubjective relationality. For example, concepts such as the “economies of touch” (Ahmed 2000, 50) allow researchers to conceive relationality as fine-grained, with multiple subtle but powerful distinctions in quality, as are found in the different ways in which bodies relate one to another. These insights have a particular force for face-to-face or immediately embodied methods, such as interviews or participant observation, in which subtle nonverbal performances (e.g., gestures, handshakes, or the arrangement of the research space) can have profound intersubjective effects. Concepts such as the cyborg or the “technological unconscious” (Thrift 2004) allow researchers to conceive of relationality as dynamic and diverse multidirectional interactions and interpenetrations, as are found between human beings and the technologies with which many human beings increasingly share their lives. Such insights can allow a reappraisal of notions like “interviewer bias” (where the interviewee adapts their responses to meet the supposed desires of the interviewer) to allow greater openness to the mutuality of effects of and on the research encounter between researcher and researched. Finally, concepts such as networks and translations allow researchers to conceive of relationality as subtle but ongoing mutual transferences of matter and meaning, as are found between human and nonhuman actors or “actants” (Latour 1997). Again, such insights can allow the researcher to interpret the roles of both human and nonhuman contributors to the research space, recognizing the roles of physical environments (from landscapes to furniture) in the construction of the intersubjective encounter.

What comes out of the focus on the relationality between subjectivities, in the context of recognizing the diversity and fragmentation of subjectivities, is a reappraisal of the research encounter as a performance that needs to be interpreted with multiple actors playing roles that all need to be interpreted, including that of the researcher. Of course, there are no guarantees in this interpretation either, and multiple interpretations may be possible. This notion of the intersubjective research encounter as a performance to be interpreted is very different to that of the incontestably powerful researcher-subject who sits at the center of the research encounter and uses methodology to engineer the research objects in order to optimize the flow of unbiased data and its rational interpretation. This said, changing understandings of subjectivity do not of course remove researchers’ responsibility to carry out research in a methodologically careful and critical way. What they do call for is an equally careful and critical attention to the demands of intersubjectivity in the research design and in the production of the knowledge that comes from it.

SEE ALSO: Intersectionality
References

Suburbanization

Paul Watt
Birkbeck, University of London, UK

Suburbanization is a prominent feature of global urban development, even if there is considerable variation in what “suburbs” look like in architectural terms (Clapson and Hutchison 2010). The geographical, demographic, and sociological significance of suburbanization is especially marked in the United States, which in many ways can be described as a “suburban nation” (Jackson 1985). Suburbanization has massively expanded during the latter half of the twentieth century, with over 50% of the US population living in the suburbs of metropolitan areas by 2000 compared with only 23% in 1950 (Hanlon, Short, and Vicino 2010).

The suburbs and suburbia can be conceptualized in several ways, cultural as well as spatial. The focus here is geographical and revolves around the notion that suburbs are peripheral locations that lie beyond a city’s boundaries, but have strong economic and social connections to the city. Connections between the city and the suburb include employment and transport whereby many suburban residents commute into the central business district (CBD) on a daily basis for work, a prominent feature of Anglo-American suburbia. The suburbs are spatially in between the city and the countryside, and they are also intermediate in terms of population densities, having lower densities than cities but higher than rural areas. Within this overarching geographical framework, suburbanization is the concomitant sociospatial process whereby cities expand outwards beyond their original central areas. This suburban expansion involves the building of new homes for sale or rent, combined with residential mobility in which people leave the city in order to live in suburban homes and neighborhoods.

Up until the last twenty years, the dominant geographical and sociological picture of the suburbs was one of white, middle-class residential areas consisting of nuclear families living in owner-occupied houses, that is, what is often called the Anglo-American model of nineteenth- and twentieth-century suburbanization in the United States, United Kingdom, Canada, and Australia. In these societies, the spatial movement from the city to the suburbs is often associated with a process of social mobility whereby those moving out also move up the social hierarchy, for example, in terms of class, income, and housing, as seen in the notion that suburbia represents a “bourgeois utopia” (Fishman 1987). This centrifugal movement outwards and upward is a significant aspect of early/mid-twentieth century US and UK suburbanization which involved intertwined processes of middle-class formation and suburban homeownership (Jackson 1985; Clapson 2003). These Anglo-American residential suburbs were based upon – and simultaneously reinforced – a pronounced gender division of labor in which wives undertook domestic labor while husbands commuted to paid jobs in the CBD. Classic examples include the outer London suburbs consisting of rows of 1930s semi-detached houses, and the postwar US Levittowns such as that built on Long Island as a suburb of New York City. Culturally the Anglo-American suburbs involve a “suburban way of life” based on maintaining social order, reinforcing class, status, and gender distinctions,
as well as conforming to middle-class norms of respectability.

During the last two decades, “revisionist” suburban scholars have increasingly challenged this dominant representation of the suburbs. Thus attention is given to the distinctive characteristics of urban peripheries in the Global South, for example, the Brazilian favelas (Clapson and Hutchison 2010). Revisionists have importantly highlighted the socially diverse nature of the Anglo-American suburbs along ethnic, racial, and class lines throughout the twentieth century (Li 1998; Harris and Lewis 2001; Clapson 2003; Wiese 2004). Immigrants to the United States and Canada, for example, have lived and worked in the suburbs, as well as in the more famous urban “ghettos.” In his study of Chinese suburbanites in Los Angeles, Li (1998) introduced the concept of the “ethnoburb” in order to highlight the geographical significance of suburban ethnic enclaves. While African Americans were less suburbanized during the first half of the twentieth century than immigrants (Harris and Lewis 2001), this does not mean to say that there were no black suburbs either, as Wiese (2004) details in his history of US African American suburbanization.

Suburban diversity also encapsulates the existence of “industrial suburbs” which contained factories as well as homes. In class terms, this means that while male, middle-class suburbanites commuted into the CBD for work, blue-collar workers in the United States and Canada lived and worked in working-class suburbs dominated by factories. US and Canadian workers often owned their homes, partly as a result of self-building (Harris and Lewis 2001). In the United Kingdom, working-class suburbanization from the 1920s to the 1970s also resulted from the large-scale provision of suburban public housing estates by local authorities at below-market level rents (Clapson 2003; Watt 2007).

While white, middle-class residential suburbanization came to dominate the early post-war period in the United States (Jackson 1985), suburban diversification has accelerated. Recent demographic and economic trends have prompted Hanlon, Short, and Vicino (2010) to suggest that there is as much diversity between suburban areas in the United States as there is between the city and the suburb. Such suburban diversity also means that the axiomatic association between outwards and upward mobility via suburbanization requires qualification. This is especially the case given the academic and media concern with contemporary suburban poverty, notably in the inner suburbs of the United States, which has given rise to the negative term “slumburbia.”

Before examining explanations for suburbanization, a definitional caution is necessary. When cities had walls, as during the Middle Ages, the suburbs literally lay outside (under) the city walls. Modern European cities have expanded well beyond their medieval boundaries, however, which indicate how areas can change their geographical status over time in that “suburbs” can subsequently become urban. Dyos (1961) notes that Fleet Street, which is now considered to be in central London, was a designated suburb in the Middle Ages, but that the “suburban frontier” gradually moved south over the centuries so that by the mid-nineteenth century Camberwell, a mile and a half south of London Bridge, became a Victorian suburb. In the epilogue to his book, however, Dyos identifies most of early 1960s Camberwell as urban. In 2013, Camberwell is swallowed up in the inner south London borough of Southwark and is simply part of the metropolis. In the United States, Brooklyn was a suburb of Manhattan at the beginning of the nineteenth century but became the fourth largest city in the country by the century’s end, eventually to become incorporated into New
Suburbanization is thus in many ways a moving frontier.

Explaining suburbanization

Explanations for suburbanization encompass a wide range of causal factors. The Chicago School, and especially the work of E.W. Burgess on the concentric ring model of the city, provided an ecological explanatory framework for suburbanization using Chicago as an exemplar city. Suburbanization here is the end product of a supposed inevitable consequence of the middle- and upper-middle classes moving out of cities in search of more space and pristine environments. The Chicago School approach has been extremely influential in relation to understanding how suburbanization occurs, but critics have drawn attention to its US-centric bias, as well as a tendency to naturalize suburbanization as being an inevitable consequence of rising affluence, and hence underplaying processes of class and racial differentiation.

The development of transportation infrastructure that facilitates commuting has been shown to have a key role in suburbanization. Developments such as the railway, the streetcar, and later the suburban highway have all been associated with the creation and expansion of suburban residential settlements, notably around transport hubs. This can be seen in the association between the development of London’s suburbs and its expanding railway network and stations, first overground and later underground. Camberwell in southeast London, for example, became a railway suburb by the 1860s (Dyos 1961). The enlargement of the railway line between Liverpool Street and Ilford in Essex during the 1890s helped to facilitate the latter’s reputation and growth as a commuter suburb into the City of London; Ilford’s population subsequently increased sevenfold from 1891 to 1911 (Heller 2013).

Contra to the one-dimensional account given by the Chicago School, suburbanization has occurred in various different ways (Clapson and Hutchison 2010). In Anglo-American societies, notions of voluntarism and powerful anti-urban ideologies are prominent, even though neither can be regarded as universal. In such societies, two groups in particular have tended to dominate suburbanization. The first group comprises property developers and speculative house builders who produce suburban housing. In the United States this has typically taken the physical form of detached houses located in what are, by international standards, large yards (Jackson 1985). In the case of the United Kingdom, especially England, suburban housing has often taken the form of semidetached as well as detached houses, but with gardens that are much smaller than their US counterparts (Clapson 2003). The second group is middle-class suburban dwellers who consume housing, usually by taking out long-term loans (mortgages) in order to finance homeownership.

A key debate is the relative causal role played by the producers and consumers of suburban housing. Many mainstream scholars have located the roots of suburbanization in the middle classes exercising consumer preferences in the sense that they optimize their utility by living in large houses in low-density areas – in other words, that they prefer to move out of crowded inner-city homes and neighborhoods. Jackson (1985) discusses the prevalence of the US “suburban ideal,” while Clapson (2003) similarly highlights the “suburban aspiration” in the United States and England, including a combination of anti-urbanism with a desire to own a home and garden in a high-quality residential neighborhood.
Marxist geographers have alternatively sought to explain suburbanization with reference to the spatially uneven nature of capital accumulation. For Harvey (1989), the period of massive postwar suburban development in the United States was primarily driven by flows of surplus capital into what he terms the “secondary circuit of capital,” that is, investment in the built environment. This investment in suburban housing and infrastructure in turn fueled a huge demand for consumer goods in order to facilitate the American suburban lifestyle, notably household motorcars and domestic appliances. Harvey even goes as far as suggesting that suburbanization helped save postwar US capitalism. Thus for Harvey, pursuit of the American suburban ideal was as much a product of capitalist property developers, aided and abetted by advertising agencies and realtors who pushed the dream onto acquiescent US consumers.

Harvey has undoubtedly highlighted the vital contribution played by capitalist processes in the production of the suburbs, as Jackson (1985) also notes in his history of US suburbia. At the same time, Jackson gives more credence to the notion that there was a deep and widespread desire on the part of ordinary middle-class Americans to relocate to the suburbs. A similar desire has been identified on the part of African American suburban residents (Wiese 2004). In other words, capitalist developers were pushing at an open door since potential suburbanites all too willingly wanted to buy into the suburban ideal.

Clapson’s (2003) notion of the suburban aspiration appears to be both relatively benign and democratizing – everyone presumably aspires to live in a “better” home in a “quality” neighborhood. In reality such aspiration involved middle-class suburban inmovers putting miles between themselves and what they disparagingly regarded as “slums” and “slum dwellers” in the city. In other words, suburbanization involved processes of sociospatial class differentiation whereby there was a desire on the part of affluent groups to spatially distance themselves from working-class people and places that they looked down upon. Moving out was – and to some extent still is – a cultural signifier of moving up, even if the objective social distance traveled is in many cases negligible.

As Watt (2007) indicates in his research on London suburbanization, the move to a “better” neighborhood can also have racial as well as class connotations, since white out-movers sought to distance themselves from ethnically mixed inner-city areas. While the role played by race and ethnicity in UK suburbanization remains under-acknowledged, early postwar US suburbanization was a largely white phenomenon, and part of the rationale for this was middle-class whites leaving inner-city areas because they had large African American populations – “white flight” – a racist process of residential mobility that enhanced racial residential segregation (Jackson 1985). Middle-class African Americans also came to leave deindustrialized inner-city ghettos in increasing numbers since the 1970s, prompting some scholars to refer to processes of “black flight.”

The impression is sometimes given that Anglo-American suburbanization was and is simply a matter of market forces, an approach that underplays the historical and contemporary role of the state. In the United States, the state has facilitated the growth of the suburbs, for example, via federal funding for highways and providing tax breaks for home ownership (Jackson 1985). In the United Kingdom, the state has more obviously played a crucial role in working-class suburbanization, notably via local authorities building peripheral public housing estates (Watt 2007). The presence of such estates is also a prominent feature of suburban areas in other Western European societies, for example, the French banlieues.
which have increasingly become home to first and second generation North African immigrants from France's ex-colonies. Finally, if early twentieth-century suburbanization involved the formation of industrial suburbs, these were the result of industrial decentralization.

The above discussion has tended to treat the various explanations of suburbanization as if they are more or less discrete causal factors. In reality, however, these factors work together. One example can be seen in the development of the Essex suburb of Ilford near London from 1880 to 1914. As Heller (2013) notes, we must recognize not only the consumers of new suburban, owner-occupied houses – the growing numbers of male city clerical workers – but also four stakeholder groups involved in the production and marketing of Ilford as a respectable, middle-class suburban neighborhood. These groups are, first, the property developers who built and sold the houses, second, the Great Eastern Railway who doubled the commuter railway tracks, third, Ilford Urban District Council (IUDC) who provided public infrastructure and assisted in the provision of credit for homeownership, and last, Ilford's existing local population who facilitated the social integration of the newcomers. Importantly, Heller (2013) demonstrates the multiple interconnections between these stakeholder groups, for example, in relation to IUDC members who were themselves developers and hence profited from the building of new suburban housing estates.

Suburbanization in post-socialist societies

Rapid and extensive suburbanization has been one of the main themes of urban change in post-socialist societies since the 1990s. In comparison to Western capitalist societies, suburbanization is a relatively recent feature of societies in Central and Eastern Europe, one that has accelerated in the period since the collapse of communism.

In the case of Poland, suburbanization is prominent in the major cities of Warsaw, Krakow, Poznan, and Wroclaw. Polish suburbanization is often based around the core of existing old villages, rather than US-style large-scale real estate developments. In her research on the suburbs of Wroclaw, Kajdanek (2014) highlights how suburbanization involves purchasing or self-building new detached houses. The primary motivation is a desire for improved housing – in terms of size and facilities including gardens – after people have lived in blocks of flats, the predominant dwelling form in Polish cities. In Wroclaw, there is little sense of the anti-urban “suburban aspiration” that Clapson (2003) identifies in the United States and England. The Polish suburbs have a more pragmatic, housing-quality based appeal which facilitates a privatized familial lifestyle. Kajdanek (2014) also found little sense of neighborliness and community in the Wroclaw suburbs and indeed considerable tensions exist between the old villagers and the suburban newcomers who tend to keep very much to their own homes and families.

Post-suburbanization

It has been suggested that suburbanization is currently mutating along “post-suburban” lines. According to Phelps and Wu (2011), suburbs have simultaneously changed form (from mainly residential to mixed use) and exploded across city-regions, a process of entropic urban decentralizing and fragmentation which has created a “post-suburban world.” Suburban hinterlands are no longer mere residential dormitories for commuters to the CBD, but are increasingly sites of out-of-town shopping
SUBURBANIZATION

malls, industrial parks, office complexes, warehouse and distribution centers, airports, and other transport infrastructure. This putative post-suburbanization – variously associated with “Edge Cities,” the “In-between City,” “city-regions,” and so on (Phelps and Wu 2011) – has rendered the question of “where does the city end and the suburbs begin?” increasingly difficult to answer.

If post-suburbanization exists anywhere, it is in the United States and especially Los Angeles which represents the de-centered metropolis par excellence, as seen in the work of the Los Angeles School. However, evidence for post-suburbia has also been identified in European, East Asian, and South American city-regions.

SEE ALSO: Built environments; Chicago School; Class; Edge city; Housing; Los Angeles School; Marxist geography; Residential mobility; Residential segregation; Rural/urban divide

References

Superstate, supranational union

Julian Clark
University of Birmingham, UK

Historical and intellectual context

Superstate defines an organizing political-administrative entity constituted from multiple states, and possessing substantive or total authority over these states. Constituent states may be ethnically and linguistically varied. The notion of superstate has parallels with, but is distinct from, related concepts of superpower, hegemon, and empire. Historical examples of superstates include ancient China, India, and Persia. In the late nineteenth century, there was discussion about reorganizing the British Empire into an imperial federation, which would technically have been a superstate. More recently the Soviet Union, though constitutionally defined as union of republics, was de facto a superstate because of its major cultural, linguistic and ethnic diversity. As superstate has few definitive characteristics, however, the term is often used descriptively (for example, in terms of scale, to contrast with the nation state), or pejoratively (as in an undemocratic, authoritarian form of statehood).

In geography, superstate has a lengthy intellectual history. There is no definitive claim to its first use, though its antecedents lie in Friedrich Ratzel’s organic theory of the state. Existing geographical theories of superstate are poorly defined, with a lack of clarity over what exactly constitutes a superstate, or how they might be created. Early examples include Naumann’s (1915) advocacy of an Oberstaat (superstate) comprising Germany and Austria-Hungary with an associated free-trade area in central Europe; Kjellén’s (1916) assertion that competition between states would lead to the emergence of superstates in Europe, Africa, and Asia; and Mackinder’s (1904) territorial explanation of how a single dominant world power could emerge by seizing control of central and eastern Europe.

All three notions of a single dominant power underpinned Haushofer’s (1931) concept of “pan-regions,” defined in terms of their distinctive resource, political, and cultural attributes and extensive land area, sufficient to create their own autonomous economic space. Haushofer maintained pan-regions would be unified by a “pan-ideen”: a supranational conception or doctrine, creating regional coherence, unity, and distinctiveness. Instantiations of supranational doctrines during the early twentieth century included Pan-Slavism, Pan-Africanism, and Pan-Europeanism. Haushofer identified four potential pan-regions as the United States, Germany (“Eurafrika”), Russia, and Japan (the focus for a “Greater Asia Co-Prosperity Sphere”). Pan-regions would depend upon their constituent states developing economic and political alliances with core states and yielding to sanctions and major cultural designations enacted by these states.

Each of these geopolitical conceptualizations portray the construction of superstates as inherently geographical – through, variously, the control of strategically positioned territory;
the use of particular technologies to maintain flows of trade, commerce, and military control between states; and by the command of space for explicit geopolitical purposes. Thorndike (1942) provides one of the first popular descriptions of superstate derived from these formal geopolitical theories. By contrast, theories of superstate and supranational union in political science are better developed, the most influential of which are neofunctionalism and liberal intergovernmentalism.

By the mid-twentieth century, with the advent of the Cold War, the importance of superstate was largely supplanted by the notion of superpower. In political geography, superstate is now rarely a topic for explicit study. Recent debates about sovereignty and democratic deficit in domestic and international politics, however, particularly in Europe, have seen a modest revival of interest in the concept.

Major contemporary dimensions of superstate/supranational union

In European mass media, superstate is increasingly used to describe one possible outcome of European Union (EU) integration. Euroskeptics argue that the EU is an emergent superstate, with the term “European superstate” used to critique further supranational union and to imply the forced loss of national sovereignty and autonomy among EU member states, and the gradual erosion of their national identity. The term is often used, mistakenly, as a synonym for a “federal Europe” which, however, differs in that national identities are accommodated with partial sovereignty retained by states, with a new tier of authority above state level given exclusive competence in one or more policy areas. A more accurate expression would be “supranational Europe,” where sovereignty of constituent states would be distributed to European institutions such as the European Parliament, European Commission, and European Court of Justice, with member states retaining little or no autonomy.

Nonetheless, from the outset the EU foresaw a supranational rather than an intergovernmental method of political cooperation. Thus the preamble to the founding Treaty of European Union, the Treaty of Rome (1957), pledges “ever closer union” between the peoples of its member states and is often cited by Euroskeptics as indicative of the covert supranational goal of the “European project.” However, this and subsequent European treaties do not clarify what is meant by “ever closer union,” or whether it should be attained through supranational or intergovernmental means, leading to uncertainty among political scientists and political geographers over the EU’s status. In reality the prospects of a European superstate are remote, as the day-to-day processes of European integration are a complex admixture of intergovernmentalism (with growing importance attached to the European Council as decision-taker) and cooperative federalism. Moreover, there is widespread public opposition among European citizens to concentration of powers in a centralized bureaucracy and little attachment to a European identity; while national identity continues to be overwhelmingly the strongest form of collective affiliation.

It should be noted that there exists a more generalized historical notion of Europeanness spanning national borders which is pertinent to notions of a European superstate. Thus during the Renaissance, the concept of “Europe” was often invoked by humanists as the secular counterpart to Catholicism. Later, “Europe” was used by Voltaire to argue for a European federation based on the rule of law. In the early twentieth century, Richard Coudenhove-Kalergi (1923)
argued the need to recognize and engender Pan-Europeanism. Similarly in the late 1960s the French journalist Jean-Jacques Servan-Schreiber set out a case for the then European Economic Community becoming a superstate in the context of growing US power globally.

While some work has been done to describe the incompatibilities between superstate and the apparent multiple social, economic, and political constructions of Europe, very few studies have been conducted by geographers explicitly on the EU as a superstate. Hudson (2000) sets out three “visions” for the EU’s development in the twenty-first century, one as a superstate. Boedeltje and Van Houtum (2008) characterize a European superstate/suprastate as an “abduction” of European identity.

The fact that there is a wealth of work now on the EU but very little that systematically engages with superstate suggests this category, like superpower before it, fits the analytical categories of the Cold War. This assessment is reinforced by the rise of international organizations (such as the United Nations) and global social movements that undermine the rationale for superstates; the presence of nongovernmental organizations (NGOs) campaigning against state authoritarianism and corruption, such as Transparency International; and the extraordinary growth in importance of social media as a means of moderating citizen–state relations. Yet it should be noted seven of the sixteen Arab states still pledge the creation of a superstate (khilafah) in their constitutions. Al-Qaeda also claims an Islamic superstate as its ultimate aim, while the Salafi jihadist group Islamic State announced in June 2014 its intention to create a worldwide caliphate. Some philosophers and political scientists maintain adherence to the superstate concept. Habermas (2006), for example, has argued in favor of a European superstate to counterbalance perceived US hegemony, while Wendt (2003, 491) contends a “world state” is inevitable if global society is to avoid nuclear annihilation.

Fresh perspectives on understanding the state and the possible role of superstates are thus needed. Two possibilities present themselves. One is to build upon work in political geography on neoimperialism and the reemergence of empire and world states as a focus of study in political science to develop greater cross-disciplinary engagement on the superstate concept. Another is for critical geopolitical analysis of the wealth of fictional superstates in literature, often depicted as dystopian, as a means to explore more utopian possibilities that can enhance contemporary state interdependencies and combat growing authoritarianism and increased surveillance of citizens within states.

SEE ALSO: Empire; Geopolitics

References


SUPERSTATE, SUPRANATIONAL UNION


Supervised classification

Xiuping Jia
University of New South Wales at Canberra, Australia

Supervised classification is one of the machine learning approaches. While no training data are used in unsupervised classification, supervised classification relies on an analyst to define the classes that the data are classified into and provide the training data of each defined class. The output of the trained classifier with the training data is an assignment of a class label to each input data point. Several classifiers are widely used, including a minimum distance classifier, maximum likelihood classification, K-nearest neighbors, and support vector machines. When an image dataset is classified, a thematic map is generated from the class label of each pixel. Training data play the key role in the establishment of the discriminant rules among the classes. The quality and quantity of the training samples are critical issues.

While unsupervised classification generates data classes (clusters) itself based on the defined similarity criterion, the first step of supervised classification relies on the analyst to define the classes to classify. They are information classes and often associated with culture names, for example, lake, ocean, urban, and road. These definitions are direct in terms of meeting users’ interest and the requirement of specific applications. However, it often needs a powerful classifier to handle the complex and highly likely nonlinear class structures. The training of the classifier becomes challenging as well. Alternatively, it is better to define a class that is relatively pure in its constitution to minimize within-class data variation, such as clear water, healthy grass, and concrete car park. By doing this, a few advantages can be obtained: it is easier to capture and describe the unique characteristics of a class; the ability to separate classes from each other is higher and a relatively simple classifier is adequate. It is also important, in general, to define all the classes present in the dataset. For example, if an imaged scene consists of deep water, shallow water, four types of vegetation, and two types of soils, all the eight classes must be defined, even if the user is interested in shallow water and one type of vegetation only, or many pixels may be wrongly labeled as the shallow water and the type of vegetation defined. A postprocessing of class merging must then be performed before a final product is generated for the user.

The second step of supervised classification is for the analyst to provide training samples and corresponding data, which are called training data. Take remote sensing image classification as an example: training samples can be identified with local knowledge, by site visits or using photo-interpretation skills based on the understanding of the remote sensing data. Training data play the key role in establishing the discriminant rules among the classes. They are used directly or indirectly via class modeling and class signature extraction (depending on the classifier used) to recognize to which class a new datum belongs to. Therefore, the quality and quantity of the training samples are critical issues to ensure the classifier is appropriately trained. Basically, the training data of each class are expected to represent the characteristics of the complete class, that is, to cover the variations presented within
SUPERVISORY CLASSIFICATION

the defined class. In the case of remote sensing image classification, multiple training areas may be selected for each class to accommodate the spatial variations. To provide a full picture of each class data and ensure the class statistics can be estimated reliably, the more training samples, the better.

The third step is to select an appropriate classifier. There are several supervised classifiers. They are different in terms of whether or not they model each class data and, if modeled, how they model. A few typical ones are briefly explained here.

Parallelepiped (boxed) classifier: this classifier uses the minimum and maximum values of each attribute (also called measure, or feature) as a class signature. The data are labeled depending on which “box” the unknown datum falls in. This method is simple and fast, but can be sensitive to noise. Class overlapping can happen easily.

K-nearest neighbors (K-NN): this classifier uses the training data in the original form without modeling. An unknown datum is basically labeled based on the labels of its surrounding training data (neighbors) in the feature space by majority vote. The neighbors are first identified using a selected distance measure, for example, Euclidean distance. The distribution of the labels of those neighbors is used for unknown data classification. This method is direct in concept, but can be time consuming to implement.

Support vector machines (SVM): this is a binary classifier, which considers two categories (one class versus another class, or one class versus the rest of classes) at a time. The training data are used to generate a hyperplane in the feature space to separate the two categories with the maximum margin. The regions on each side of the hyperplane define the data space of the two categories, respectively. If the data cannot be separated linearly by a hyperplane in the original data coordinates, kernel SVM can be applied by mapping the data to a higher dimensional space. The data on the class boundaries are called support vectors and they are the critical data in determining the separation plane.

Minimum distance classifier: the mean value of each dimension is calculated using the training data of each class. A mean vector with N elements (N is the number of dimension of the data) is formed and used as the class signature. An unknown datum’s distance to each class mean vector is evaluated and the minimum distance indicates the class it belongs to. This method is insensitive to noise and the requirement for the number of training samples is not high, as only mean values are estimated. However, the difference in the data distribution from class to class is not taken into account.

Maximum likelihood classifier: this classifier considers the distribution of each class datum by modeling it as a normal distribution. The normal distribution parameters, mean vector and covariance matrix, are estimated using the training data. Bayes’s rule is then applied to find out the highest probability of the class that an unknown datum belongs to. This classifier works well if the class data can be reasonably assumed as a normal distribution and the parameters can be estimated reliably using the training data. Mathematically, the minimum number of training samples is N + 1, where N is the dimension of data, so that the covariance matrix has full rank and can be inverted. In practice, multiple times of N are required to avoid overtraining. When a classifier is overtrained, the training results are biased and are unable to be generalized to the data other than training data.

Considering the new big data era, N can be a high number. A very large number of training data is required to run supervised classification. Unfortunately, collecting training samples is an expensive procedure as human resources are
heavily involved. A semi-supervised approach (Chapelle, Schölkopf, and Zien 2006) has been developed that incorporates the use of unlabeled samples. In this approach, the initially trained classifier, using the original training dataset, is updated via self-learning using the information provided by the unlabeled data. An active learning approach (Tuia et al. 2011) addresses the problem through optimal experimental design to minimize the effort in training data collection. The classifier is initially trained using a small training set. Then a learning algorithm is applied to ask the analyst to provide the labels for the new most informative samples. In this way, the redundancy in training data collection can be reduced. Alternatively, feature mining can be applied to reduce the dimensionality of the data and, consequently, reduce the size of the training data set required (Jia, Kuo, and Crawford 2013).

The performance of a supervised classifier can be evaluated using another set of labeled data, namely testing data, which is independent from the training data. Similar to training data, testing data can be identified with local knowledge by site visits or using photo-interpretation skills based on the understanding of the remote sensing data. However, they are not used in the class modeling and classifier training. Instead, they are labeled by the classifier and the results are checked against their ground truth. The checking results are often given in tabular form (called a confusion table). Classification performance can be measured by producer accuracy, user accuracy, overall accuracy, and kappa coefficient (Congalton and Green 2008). A reasonable size of testing samples for each class is required to provide meaningful statistics.

SEE ALSO: Class; Machine learning; Spatial sampling; Unsupervised classification

References


Further reading


Surface water

L. Allan James
University of South Carolina, USA

This brief synthesis of nonsaline, surface water integrates both the physical and human elements of freshwater resources, that is, both hydrologic and water resources perspectives. The hydrology of surface water is concerned with the physical aspects of water, such as the hydrosphere and hydrologic cycle, water budgets, and runoff-generation processes. In a complementary manner, societal aspects of surface water resources management are concerned with interactions between water and humans that involve types of water use, water availability, scarcity, and equity, global sustainability, and demand management and reallocation. This holistic view of water and water resources management can be summed up under the rubric of integrated water resources planning and management. Management requires an appreciation and understanding of the physical resource but goes beyond that to embrace complex issues of water allocation, conveyance, forecasting demand, and planning for future availability. Growing water demands to irrigate food crops in order to provide for rapid population growth exemplify the dire necessity of getting both physical and human planning perspectives right and getting future supply and demand to meet in an equitable manner.

The study of surface water is commonly separated with a strong focus on either societal facets of water resources or physical, chemical, and biological science perspectives. Such a separation is artificial, however, and often leads to specialized treatment of isolated problems rather than an integrated approach to problems that recognizes interactions between systems. Familiarity with both physical and human phenomena provides water managers with an overview of how physical and social systems interact regarding surface water. This integration is highly geographic insofar as it mimics several geographic traditions, such as regional complexes, spatial analysis, and tensions between humans and the environment.

Constraints for brevity of this entry combined with the broad topical approach to surface water taken here mandate a summary synthesis rather than comprehensive coverage. Thus, several key elements of surface water are omitted entirely. For example, broad topics such as water quality, flood and drought risks, and legal constraints on use and protection of water resources are not covered. For treatment of these topics elsewhere in this volume, please consult the “see also” section at the end of this entry. Moreover, concepts of local water management are given less attention than global views, due to emerging issues of global change, climate variability, and the research and policy initiatives aimed at addressing these problems on a global scale. This article begins with a brief discussion of geographic traditions in the study of surface water. This is followed by the physical aspects of surface water and by resources management concepts.

Geographic traditions in studying surface water

Geographers have long been involved in the study of surface water. In the United States this history can be traced to the explorations...
and studies of John Wesley Powell, a founding member of the National Geographic Society, whose explorations of the western North American continent led him to recognize the critical importance of water resources to western settlement (Powell 1878). In the mid-nineteenth century, surface hydrology became a core concern of physical geographers in the United States and Great Britain through an extension of interest in fluvial geomorphology by scholars including Richard Chorley, George Dury, Reds Wolman, and many others. At the same time, the work of Gilbert White and his students spawned an era of research in two areas: water resources in developing nations and water resources hazards.

The increasing importance of spatial analysis, investigations of human interactions with the environment, climatology, and other traditional areas of geographic research led to a growing involvement of geographers in water science and management over the past several decades. The value of remote sensing mapping technologies coupled with spatially distributed hydrologic modeling brought geographers into mainstream scientific research and planning programs. A panoply of diverse hydrologic models emerged during this period for rainfall–runoff relationships, water yields, droughts, water use, and water quality. Geographers also bring a sophisticated recognition of the importance and complexity of human–land interactions from scientific, environmental, and humanistic perspectives. These perspectives are useful for modeling but also are essential to water resources management and planning. Geography was a mainstay of climatology and hydroclimatology throughout the twentieth century. The tight linkages between surface water and both climate and environmental sensitivity to human activities resulted in a rich tradition of hydrologic and water resources research in geography from a variety of viewpoints.

Representing freshwater resources

Water resources management is a 4-D problem that can be accounted for with simple mass balance methods. A complete accounting should consider water transfers in 3-D Euclidean space over time, which involves evaluations from both hydrosphere and hydrologic cycle perspectives with quantitative accounting methods.

Hydrosphere and hydrologic cycle

The location and amount of fresh water at the Earth’s surface relies on inputs and outputs that vary greatly with location and scale. The global water system consists of the hydrosphere, a snapshot of large-scale water budgets, and the hydrologic cycle, large-scale hydrologic processes and water-flows. The hydrosphere includes water stored in oceans, rivers, lakes, glaciers, soils, the atmosphere, the lithosphere, and the biosphere. From this static standpoint, most of the water on Earth is stored in oceans and is too saline for most human uses. Of the 3.5% of water that occurs on land, most is in glaciers or groundwater reservoirs. Nonsaline surface water that is not locked in ice is only a very small percentage of the total hydrosphere (0.01%) and only 0.28% of the nonoceanic water (Table 1).

Although these freshwater storage reservoirs are a critical resource, fluxes of water between them – integrated over time – are what governs these reservoirs and are what environmental and social systems depend on. Water is a renewable resource, so an appreciation of the hydrologic cycle with flow rates to and from the volumes of storage in the hydrosphere is a central concern of planning and management. By analogy, the volume of water stored in a short garden hose may seem negligible in comparison with the volume of water in a rain barrel. Over a period of time, however, flows through the hose may fill
Table 1  Fresh surface water in the hydrosphere.

<table>
<thead>
<tr>
<th></th>
<th>Volume (10^3 km^3)</th>
<th>% of total</th>
<th>% Without ocean</th>
<th>% Thawed surface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Earth water</td>
<td>1385 968</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lakes</td>
<td>91</td>
<td>0.0066</td>
<td>0.1900</td>
<td>67.4</td>
</tr>
<tr>
<td>Soil water</td>
<td>16.5</td>
<td>0.0012</td>
<td>0.0344</td>
<td>12.2</td>
</tr>
<tr>
<td>Atmosphere</td>
<td>12.9</td>
<td>0.0009</td>
<td>0.0269</td>
<td>9.6</td>
</tr>
<tr>
<td>Swamps</td>
<td>11.5</td>
<td>0.0008</td>
<td>0.0239</td>
<td>8.5</td>
</tr>
<tr>
<td>Rivers</td>
<td>2.1</td>
<td>0.0002</td>
<td>0.0044</td>
<td>1.6</td>
</tr>
<tr>
<td>Biomass</td>
<td>1.1</td>
<td>0.0001</td>
<td>0.0023</td>
<td>0.8</td>
</tr>
<tr>
<td>Total</td>
<td>134</td>
<td>0.0098</td>
<td>0.282</td>
<td>100</td>
</tr>
</tbody>
</table>


thousands of barrels. So it is that the dynamics of atmospheric fluxes and river flows may deplete or replenish the largest storage reservoirs and persistent withdrawals may disrupt balances in water resources systems. Flow rates and processes of the hydrologic cycle are important for maintaining the storage compartments of the hydrosphere.

The hydrologic cycle emphasizes fluxes and flows of water between the atmosphere, land, and water bodies, including precipitation, evapotranspiration, infiltration, and surface runoff. Water may move across the land surface as runoff and be stored in soils, biomass, rivers, ponds, lakes, reservoirs, and wetlands before returning to the atmosphere or to the sea. The large amounts of latent energy stored by evaporation and released by condensation, coupled with energy transfers due to convection and atmospheric circulation, result in strong links between climate and surface water at all scales.

Water budgets

Given the principle of conservation of mass, a water budget (also known as a water balance) can be used to determine water movements by accounting for all inputs, outputs, and changes in storage. For example, a simple water budget can express the primary pathways that precipitation (P) may take:

\[ P = ET + RO + I_{nf} + I_{nl} \pm dS \quad (1) \]

where \( ET \) is evapotranspiration, \( RO \) is surface runoff, \( I_{nf} \) is infiltration, \( I_{nl} \) is interception by vegetative matter, and \( dS \) is change in water stored in rivers, ponds, lakes, soil, groundwater, and so on. Change in storage may be positive if inputs are greater than losses or negative if less than losses. Budgets such as equation 1 can be used to compute a missing parameter by specifying all of the terms but one. Each of the terms can be expanded or redefined to account for different fluxes such as adding irrigation water to precipitation inputs, adding water withdrawals, or assuming a parameter is negligible. For practical management purposes, surface water is essentially incompressible, so water budgets can be expressed as masses or volumes. Alternatively, they may be expressed in units of depth over a given surface area, that is, as centimeters of precipitation, evapotranspiration, and runoff. Water budgets are often expressed at regular time intervals, such as hourly or monthly,
in order to track water through a system. Water budgets can be created at any scale. For example, a budget can be constructed for the effects of a storm or long-term fluxes on an individual water body such as a lake or a reservoir. The ending volume resulting from an initial volume and net inputs and outputs can be computed from the mass balance:

\[
V_2 = P - ET + I_{sfk} - O_{sfk} + I_{sub} - O_{sub} + V_1
\]

where \( I_{sfk} \) and \( O_{sfk} \) are surface inflows and outflows, \( I_{sub} \) and \( O_{sub} \) are subsurface inflows and outflows, and \( V_1 \) and \( V_2 \) initial and ending lake volumes. Water budgets are needed that are based on accurate assessments of inputs and outputs and changes through time.

At the global scale, climate is the overarching driver of water availability unless external water sources are available, such as an exotic river or independent groundwater sources. Long-term global water budgets are dominated by precipitation and evapotranspiration fluxes (Gleick 1993). Approximately two-thirds of global terrestrial precipitation is lost to evaporation. Global budgets representing average conditions factor out large spatial and temporal variations. Geographic variability is key to accurate assessments of water availability. As a first approximation, freshwater availability can be mapped by subtracting evapotranspiration from precipitation. Several international assessments of past and future global water availability have been made (Shiklomanov 2000; Shiklomanov and Rodda 2003). Ultimately, however, human activities must be added to these estimates in the form of water transfers, consumptive use, land use, and climate changes that alter evapotranspiration, infiltration, runoff, and storage.

### Surface runoff and flood generation processes

Surface water results primarily from precipitation that does not infiltrate into the soil and from groundwater seepage back to the surface. Pathways of water at the atmosphere–land boundary determine the rate and proportion of precipitation that is lost to abstraction and how much will run off as surface water. Human land-use activities have a tremendous effect on these processes, rates, and proportions of surface runoff generated.

### Water pathways at the hillslope scale

As rain falls or ice melts, surface water at the soil–atmosphere interface flows along identifiable pathways such as through the vegetation canopy, into the soil, or downslope. These paths include interception by vegetation, infiltration into and percolation through soils, surface runoff, and storage. Surface runoff may occur rapidly by Hortonian flow when precipitation intensity exceeds infiltration capacity or by typically lower-velocity saturated overland flows when soils are saturated. Much slower delivery may result from shallow subsurface flows within the unsaturated vadose zone of soils. Flow paths determine the timing of water delivery that governs flooding and the magnitude of soil erosion and pollution generation, which govern water quality. Abstraction processes of interception by vegetation and infiltration by soils reduce runoff so less of the precipitated water runs off on the surface. This abstraction tends to reduce flooding, erosion, and generation of pollution. The infiltration capacity of soils, that is, the maximum rate at which water enters the soil, governs abstraction and the generation of surface runoff. Infiltration diverts water from relatively rapid and erosive surface pathways to...
subsurface pathways that are relatively slow and nonerosive. Both interception and infiltration decrease greatly during storms as vegetation and soils become wet, so rates of initial abstraction are much greater than after it has been raining for some time. In fact, little of the initial rainfall runs off. Later phases of rainfall that generate runoff are referred to as the effective rainfall. Infiltration capacities vary with a number of natural factors, including geology, soils, precipitation characteristics, and vegetation, but they are greatly altered by human activities such as soil compaction and land use that changes land cover.

Human impacts and urbanization

Human manipulations of soil, vegetation, and land cover often result in substantial changes in infiltration capacities. These changes may result in dramatic hydrologic responses, including increased surface runoff, erosion, and generation of nonpoint source pollution. A key environmental consequence of urban land-use changes is exacerbated flooding due to a transition from gradual hydrologic response with relatively low flood peaks to amplified flood peaks that rise rapidly. Pavement, rooftops, storm sewers, vegetation clearance, and soil compaction that often accompany urbanization all contribute to magnified and accelerated runoff generation. In fact, dense urbanization causes increases in flood magnitudes to much greater multiples of flows in unaltered basins than are anticipated with even the worst-case climate change scenarios in noncoastal areas.

The percent area of impervious surfaces has become a common metric in urban hydrology that can be used to indicate not only the degree of increased risk of flooding but also the degree to which riparian habitats and ecosystems have been compromised (Schueler 1994; Walsh et al. 2005). Similarly, the degree to which a basin is drained by storm sewers drives hydrologic response. Computations of runoff for urban watersheds where stream-gauge data are not available should adjust estimates upwards for areas with high percentages of impervious surfaces and high drainage densities of storm sewers in order to compensate for greatly increased drainage efficiencies of those systems. To some degree these processes can be mitigated through the use of conventional storm detention or retention structures or the use of more spatially distributed low-impact development (LID) techniques, such as bioretention (rain gardens), tree-box filters, rain barrels, cisterns, permeable pavements, and green roofs.

Characteristics of different water uses

Water is used in a variety of sectors that may be generally classified as municipal, agricultural, industrial, thermoelectric cooling, and hydropower. At the global scale, water-use categories are defined broadly in a simple tripartite classification system of agricultural, municipal, and industrial, because more specific details cannot be distinguished consistently between nations. These classes can be applied to areas with sparse, incomplete data collected by a variety of methods or approximated by indirect methods such as global modeling. A fourth water-use category for reservoir evaporation losses is included in data developed by Shiklomanov (2000). Fundamental differences within and between water-use categories include variations in consumptive use, in-stream versus off-stream use, environmental impacts, cost, and requirements for water quality, including the ability to use saline water. Understanding water-use statistics is a surprisingly complex skill due to these variations in water requirements and outcomes. For example, in developed nations, the largest
SURFACE WATER

water-use sector may be hydropower production, which is an in-stream, nonconsumptive use. Hydropower production may have extensive environmental consequences and may alter water temperatures and the timing of daily releases. However, impacts of hydropower production on water availability tend to be relatively benign compared to agricultural water applications that are an off-stream, highly consumptive use. Municipal water use, which tends to be of the order of 10% of total use, requires high-quality water and generally has the ability to pay a higher price than other sectors. This use is not as consumptive as agricultural use but returned water requires extensive treatment. The dominant water-use sector in a region tends to be related to the level of economic development. Developed nations often use a high proportion of water for hydropower, thermoelectric cooling, and industrial applications, whereas the dominant water use in developing nations is usually irrigation for food production. The inflexible and highly consumptive nature of agricultural water use for feeding vulnerable local populations reduces flexibility in water use and has grave implications for growing water demands.

Global water scarcity and food security

Shortages of water call for quantitative indicators, that is, metrics of water availability. Available water refers to the amount of water normally accessible for human use. A shortage of available water differs from drought conditions that are relatively brief deviations below normal conditions. Low water availability in arid or semiarid regions is a long-term condition and persists during relatively wet years. In humid regions, a relatively large availability prevails even during droughts. Water shortages depend on both available water supplies and on demand.

Two common water shortage metrics that factor in demand are the Falkenmark index (water stress) and relative water demand (RWD). The Falkenmark index measures water availability per capita. A condition of water stress occurs when long-term per capita availability drops below 1700 m³ year⁻¹ (1230 gallons day⁻¹). Water scarcity occurs when per capita availability drops below 1000 m³ year⁻¹ (724 gallons day⁻¹); this condition is associated with severe health, economic development, and human wellbeing issues. Water shortages can also be expressed as the RWD or criticality ratio, defined as the ratio of demand to supply, that is, the ratio of water withdrawals to available water:

\[ RWD = \frac{\text{annual withdrawals}}{\text{available water}} \]  

Specific thresholds of RWD have been defined with moderate RWD setting in at 20% and severe RWD beginning at 40%. RWD can be defined over the long term, seasonally, or by sector. Water availability is strongly dependent on scale; that is, the amount of water available increases as a power function of the area of the basin or grid cell used for computation. Conversely, water scarcity and RWD are scale-independent because they are standardized by population and water availability, respectively (Perveen and James 2011).

Water stress and RWD are strongly interdependent insofar as demand is often related to population. Scatter around the predicted line in a plot of the two metrics for 26 global regions, however, reveals important differences between regions (Figure 1). African regions all plot below the line indicating that water shortage is driven more by population (per capita water availability) than by other demand factors generating withdrawals (RWD). Subsistence agriculture is prevalent in Africa with relatively small withdrawals by industry or thermoelectric cooling. Developed nations with high withdrawals for
industry or cooling water, such as in North America, plot above the line. Of the 26 regions, only northern Africa has critical water scarcity, although southern, central, and western Asia have extremely high RWD values and several regions have moderately high RWD values. The small number of regions with water scarcity in this plot is due in large part to regional averaging that dampens out local extreme shortages. Mapping at subregional scales would reveal many areas with critical water shortage conditions.

Two types of water can be identified (Falkenmark 1995). Green water is stored in soils, is relied upon by rain-fed farming, and is usually available for plant use. Most global food production is currently based on green water. Blue water is water stored in surface and groundwater bodies and can be transported and used for irrigation projects. Water resources management is largely concerned with storage, transport, and allocation of blue water. Global water demand for irrigation water is growing, however, in response to exponential population growth. The need to shift water to irrigation to meet growing food demands stems from the much greater efficiencies of irrigated food production over dry farming. Postel (1993) estimated that irrigated lands comprised only 16% of the farmland areas, yet they produced approximately one-third of the global food supply (Figure 2). Expansion of food production to feed growing populations represents a rapid increase in the demand for blue water for irrigation.

Global expansion of irrigated lands accelerated greatly after World War II, as mechanized farming and the use of chemical fertilizers proliferated and demand for food grew in response to population growth during the baby boom era (Figure 3). Details in the rate of growth of irrigation after 1960 can be inferred from the areal expansion of irrigation land. By far, Asia has the largest area of land under irrigation and irrigated lands in Asia more than doubled from 106 million hectares in 1961 to 229 million hectares in 2012 (Table 2). North America and Europe have the next-greatest areas of irrigated

Figure 2 Irrigated land can produce much more food than dry farmed land. The area of irrigated land is a modest proportion of global farmland (left). Yet, irrigated lands produce approximately one-third of the global food supply (right). Data from Postel (1993).
Global sustainability and water resources

Sustainability refers to rates of resource use that can be maintained for an extended period of time without substantial depletion of the resource or damage to social and environmental systems. This implies time periods that should be measured in centuries or longer. Violations of sustainability principles may lead to both intergenerational inequities and environmental damages, because shortages often generate serious socioeconomic and ecological damage. Global water supply and demand are presently out of balance and projections indicate that this imbalance is growing, which will continue to damage environmental and social systems. In other words, present water-use policies are not sustainable. Solutions to this situation are extremely difficult to find because the fundamental problem is rooted in population growth and the growing need for food production.

At the tenth Stockholm Water Symposium in August, 2000, acute water shortages in dozens of nations were recognized, and water shortages or worse were projected for up to two-thirds of the world’s population by 2020. In the interim, climate change and political unrest have complicated such forecasts, but shortages are coming to bear. Implications of these shortages vary with supply and demand as well as with social resiliency. In impoverished areas with severe water scarcity, extreme hardships may emerge. The traditional approach for meeting demand in times of water shortage has been to increase supplies, but this has shown rapidly decreasing effectiveness as sources of additional water have dwindled. Conservation of water resources,
that is, water demand management, must be an important element of the solution. Conservation will be difficult, however, due to increasing demands for food production, which requires a growing proportion of use for irrigation, a highly consumptive application. One strategy for increasing food production in wealthy countries with limited water supplies is to import food. Market exchange of commodities for food, known as using “virtual water,” is a practice that may be a viable alternative to increasing irrigation for nations that can afford it, but it is not a solution for the poor countries with severe water stress.

**Integrated water resources planning and management**

Complex interactions between environmental and social systems that interact with surface water—require a broad, integrated management perspective. An interdisciplinary approach to water resources management is needed that integrates economic, political, physical, and technical methods to achieve sustainable and ethical ends. Integrated approaches were advocated by Gilbert White (1961) and were still considered relatively new in the 1990s (Getches 1993). The concept of integrated water resources management (IWRM) is now commonplace, but in practice procedural norms are far from integrated. Integration requires coordination of diverse methodologies and disciplines as well as cooperation between regulators, institutions, and stakeholders across a spatially diverse landscape. It requires public and professional interactions that only arise from active attempts to engage diverse viewpoints in dialectic exchanges. Spatial integration of phenomenon is also an important element of IWRM. Locations of features and stakeholders in space are affected by adjacency, proximity, and distance, and these geographic factors should be considered essential.

Integration can greatly reduce redundant efforts between groups studying and managing water resources. It can also reduce errors generated when nonspecialists attempt to solve problems outside their areas of expertise by ensuring a diverse team approach. Efforts to integrate broad scientific, sociopolitical, and environmental perspectives initially encounter challenges due to communication and ideological differences. Once these barriers are overcome, however, integration results in improved efficiencies, as solutions are developed from a variety of perspectives that satisfy multiple objectives.

**SEE ALSO:** Freshwater resources: past, present, future; Irrigation; Mountain hydrology; Rivers and river basin management; Water budget; Water conservation; Water resources and hydrological management; Water security; Wetlands hydrology

**References**


SURFACE WATER

“Surplus labor” is a concept developed by Karl Marx in the mid-nineteenth century. Unlike classical economists of his day and ours, Marx’s labor theory of value explains economic relations through the lens of the production and distribution of surpluses. Surplus is something extra, something beyond that which is needed. Surplus labor refers to the labor that produces a value over and above what is necessary for the reproduction of an individual worker or workforce. Workers in societies of all kinds perform surplus labor, which is “extra” labor beyond what their immediate subsistence requires. Thus workers in all kinds of societies produce a surplus beyond what they need for their own survival. Those surpluses are distributed in a variety of ways in different contexts. For Marx, how and by whom that surplus is appropriated and distributed is the moment that distinguishes an economic process as capitalist, slave, communal, or feudal to name a few among many possible relations of production.

In order to consider some portion of labor “surplus,” it follows that some other portion of labor is considered “necessary.” Thus, it is helpful to divide the working day into two (see Table 2; see below). For Marx, necessary labor is the portion of the working day that produces the value of the means of subsistence. The value of what is necessary for subsistence is socially defined and continually negotiated. For example, workers in many places have fought to establish and raise minimum wages. We might consider the minimum wage of a given location, the socially recognized value of necessary labor. We could call this variable amount the survival payment or, in the language of Marx, the value of labor power. It is the amount workers need to pay for necessities such as housing, clothing, medical care, food, and transportation to and from work. It is the amount they need in order to show up to work the next day. Consider the example below.

A small business with one employee generates $200 in revenue per day. The employer pays the minimum wage set by law which in this fictitious example is $10 per hour. As depicted in Table 1, after the owner has covered his/her nonlabor inputs of $40 and labor inputs of $80, there is $80 left over. That remaining $80 is called “surplus.”

The $80 per day wage paid by the owner to the worker is the socially negotiated value that covers what the worker needs to show up for work the next day. “Necessary labor” is the labor required to produce the $80 value of the workers’ survival payment, the minimum wage. If the worker produces that value in four hours, they perform four hours of “necessary labor.” Labor performed above and beyond those four necessary hours is considered “surplus labor” which is in this accounting unpaid labor. The value produced during those surplus labor hours is called “surplus value” (“SV” in Table 2).

Scenario “A” in Table 2 depicts a work day in which the value of the survival payment is produced in four hours but the worker continues working an additional four hours. The first four hours of the work day are spent performing necessary labor dedicated to the reproduction of the worker while the second four hours are considered surplus labor. The value that he/she produces during the portion of the day performing surplus labor is called surplus value.
**SURPLUS LABOR**

**Table 1**  Company X: 1 employee $200 daily revenue.

<table>
<thead>
<tr>
<th>Labor inputs</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>100</th>
<th>120</th>
<th>140</th>
<th>160</th>
<th>180</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>$40</td>
<td>$80</td>
<td>$80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2**  The working day divided in two: necessary and surplus labor. SV, surplus value.

<table>
<thead>
<tr>
<th>Working day</th>
<th>$160 production value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Necessary labor</td>
<td>Surplus labor</td>
</tr>
<tr>
<td>$20/hour</td>
<td>$20/hour</td>
</tr>
<tr>
<td>A</td>
<td>$10/hour = 4 hours necessary labor</td>
</tr>
<tr>
<td>B</td>
<td>$12.50/hour = 5 hours necessary labor</td>
</tr>
<tr>
<td>C</td>
<td>$15/hour = 6 hours necessary labor</td>
</tr>
<tr>
<td>D</td>
<td>$17.50/hour = 7 hours necessary labor</td>
</tr>
</tbody>
</table>

In scenario “A” the worker produced a surplus value of $80. As the wage rises (incrementally in scenarios B, C, and D), we see a decrease in surplus value (from $60 to $20 respectively).

The boundary between what is considered “necessary” and what is considered “surplus” labor varies between enterprises and geographic location and it is often intensely negotiated among economic agents. As shown in Table 2, an increase in survival payment results in a decrease in surplus labor. If the minimum wage rises or the employer decides to increase the wage to $12.50 per hour or $100 per 8 hour day, as depicted in scenario “B,” surplus labor would decrease by one hour. Thus, when other variables remain constant, an increase in the survival payment is equal to the decrease in surplus value. In scenario “B,” we see a reduction of surplus value in the amount of $20 per worker per day. Instead of increasing the wage (or survival payment), the employer could decide to donate surplus to charity, invest it in the business, keep it for him/herself, or return some of it to workers in the form of profit share or bonuses. From a Marxist perspective, regardless of what the owner decides, the enterprise in this example is considered capitalist because decisions regarding surplus labor reside with the owner of capital.

Surplus labor analysis reveals a fundamental antagonism between workers and owners of capital in a capitalist context because profit originates in the unremunerated labor hours of surplus labor. In productive industries, surplus value is maximized by lengthening the working day and minimizing the wage (again see Table 2). The profit motive thus drives the capitalist business owner to minimize what is deemed “necessary labor” and maximize “surplus labor” by minimizing the wage and expanding the work day. Conversely, workers
fight for shorter days and higher wages, which, as we see in Table 2, decrease surplus labor from which profit is derived. Throughout the history of industrial capitalist development, labor movements have fought for greater wages and limits to the length of the work day.

Due partly to the success of organized labor in the developed regions of the world, some labor-intensive manufacturing production has shifted from places with highly paid union workers and an 8 hour day to places in which the working day is longer, survival payments are smaller, and labor is not unionized. Workers in China, for example, tend not to be unionized and may require $1.50 per 10 hour work day while workers in Canada or the United States might require $80 for an 8 hour day. Given nonlabor costs of production and production output equal to those of the business depicted in Table 2, a worker earning $1.50 per day would produce the value of his/her necessary labor in one and a half minutes. The remaining 9 hours and 58.5 minutes of the work day would be spent performing surplus labor generating surplus value of $198.50.

Surplus labor is performed, and surplus value produced, by workers in a variety of economic arrangements or processes, not just capitalist. For example, a worker cooperative may look very similar to a capitalist enterprise. Worker members perform necessary and surplus labor and produce surplus value just like waged workers in a capitalist enterprise. However, from a Marxist perspective a worker cooperative is considered a noncapitalist communal enterprise because the workers themselves appropriate and distribute the surplus value produced by their own surplus labor. Unlike capitalist businesses in which decisions regarding surplus labor are made by the nonproductive owners of capital, in a worker cooperative, members negotiate decisions with their coworkers about their own surplus labor and the surplus value they produce.

The concept of surplus labor is essential to understanding the origin of profit from a Marxist perspective. It is a key concept in critical geography concerned with labor and capitalism because profit, the sole motivation of conventional business, is derived from surplus labor. Surplus labor accounting can be used to measure the rate of exploitation in the form of unremunerated labor hours in a given enterprise. It also provides an understanding of geographic trends in a profit-driven enterprise as well as economic diversity and possibility. For if a capitalist process is simply one in which surplus value is appropriated by nonproductive owners as profit, the economy appears far from definitively capitalist.

Surplus labor accounting demonstrates that even in industrialized economies people participate in multiple rather than exclusively capitalist modes of production and exchange as workers, owners, producers, appropriators, and distributors of surpluses. Feminist economists and geographers distinguish over half of economic processes worldwide in both rich and poor countries as noncapitalist. Thus understanding surplus labor illuminates capitalism and its exploits as well as the ways through which an economy is constructed through feudal, slave, capitalist, gift, volunteer, and communal among other economic processes.

SEE ALSO: Exploitation; Marxist geography

Further reading

SURPLUS LABOR


Surveillance

Hille Koskela
*University of Turku, Finland*

Liisa A. Mäkinen
*University of Helsinki, Finland*

**Surveillance and geography**

In these times of increasing tension and insecurity, surveillance has become an important topic in many academic fields, including geography. While surveillance studies is a growing and genuinely multidisciplinary field of research, scholars from each discipline have their unique perspectives whereas geographers focus on space. In terms of sustaining order, it matters who is where and when – hence, the combination of space and time is essential. Technologically advanced systems use positioning or define geolocation – practices which also have their foundation in spatial existence. As a matter of fact, most forms of surveillance are spatial in one way or another. Nonetheless, there are two contexts that have caught the attention of academic geographers: border control and the monitoring of public urban space.

This entry proceeds from some historical notes and essential definitions toward metaphors and theories. Metaphors such as “1984” or “the panopticon” have commonly been used for describing the manners of surveillance. Nevertheless, critical examinations of these metaphors have helped in developing academically significant social and spatial theories. Thereafter, the entry goes on to highlight two research lines within geography. The political geography of surveillance focuses on the changing nature and role of borders and border control. Urban geography, on the other hand, examines surveillance practices in public urban space and claims that they produce a risk of spatial segregation. Last, a brief look at counter-surveillance shows what developments are likely to take place in the future.

**Defining surveillance**

The definition of surveillance as a practice is far from easy to elucidate. Generally, surveillance is understood to mean observing people’s behavior, activities, and communications for the purpose of controlling, managing, directing, or protecting them. Surveillance is often perceived as something or someone looking from above. The French origins of the word support this perception: surveillance combines two French words, “sur” – meaning “from above” – and “veiller” – meaning “to watch.” Hence, literally the word would translate “watching over.” However, while surveillance can include visual control, it also includes various other forms of information gathering and management. In terms of location, surveillance can take place “on the site” where something is happening at the moment, but it can as well include monitoring from a distance by detaining electronically transmitted data. Surveillance can be used for intelligence gathering, for preventing and investigating crime, and for protecting citizens and borders.

While the bodies which currently conduct surveillance are on the leading edge of developing technology, the history of surveillance has been
rather “low-tech.” Humans have monitored each other by various methods long before current technologies were established. Postal interception, eavesdropping, and telephone tapping have been among the first surveillance techniques. Furthermore, espionage has long been associated with an institutional effort by governments spying on potential or actual enemies. Mostly this has been for military purposes – identifying enemies coming from “outside” – but there are notorious examples of countries monitoring their own citizens – identifying the enemies “within.” Former communist East-European countries had an intense net of informers who surveilled their fellow citizens. Currently, similar practices take place in many nondemocratic countries around the world and, all the more, significant amount of surveillance technology is developed by military institutions. Surveillance tends to increase “in layers”: the old forms do not disappear when new ones are invented, but stay “underneath” the more sophisticated forms. This has led to a widespread intensification of surveillance.

People persistently perceive surveillance as centralized, as if there would be one database and one monitoring station managed by the authorities such as security guards or law enforcement executives. However, this is quite often not the case. When it comes to video surveillance, cameras operated by the private sector have outnumbered those used by the authorities for a long time. The majority of cameras are operated by corporations or other private parties. Furthermore, the authorities currently have very little control over how and where surveillance is used. Surveillance is increasingly conducted by people other than the police and security guards, for example, by janitors, shopkeepers, bus drivers, nurses, teachers, and so on. Corporate surveillance – both in physical and online environments – is used for consumer profiling and marketing. While corporate surveillance has its own specific purpose, it does not take place in a vacuum but information is often conveyed to government agencies.

With cheaper production techniques, surveillance cameras are simple and inexpensive enough to be used in home security systems and for everyday surveillance. Private individuals have been active in applying new technologies. Home security systems with alarms, cameras, and private policing services have become increasingly common. Also, mobile phones enable instant photographing and tracking. Both the responsibility for evaluating risks and the responsibility of implementing control have increasingly been moved away from different institutions toward individual citizens. David Lyon (2014, 72), accurately states that “[i]t is time to stop thinking of surveillance merely as a top-down phenomenon where ‘they’ monitor ‘us.’ Surveillance has become an everyday social experience, from a serious security issue to an incessant demand for data from numerous organizations to a playful part of mediated relationships.”

There are so many different constantly developing types and forms of surveillance that formulating any solid definition is challenging.
While these practices are overlapping in many ways, a rough categorization of different types increases understanding. Visual surveillance – video surveillance, often also called closed circuit television (CCTV) – is likely to be the form of surveillance people most often face in their everyday lives. Surveillance cameras are so widespread, especially in Western cities, that visual control has become a permanent condition. Video surveillance is also used for traffic control. Cameras can identify and track vehicles as they move around. Future development is likely to increase the amount of facial recognition software and other advanced video analytics which will then identify and track individuals everywhere they go. Visual control does not always need human workforce: rapidly developing algorithmic surveillance is able to identify many kinds of “risky” behavior without human intervention. Pictures and video can be gathered from aerial vehicles such as helicopters or unmanned vehicles called “drones.” Military surveillance is a typical setting for aerial vehicles but there are applications for civil use. In fact, helicopters – and increasingly also drones – are used by the police both for searching criminals and for “crowd control” in demonstrations and other mass gatherings.

Biometric surveillance is a term defining various technologies which measure human biological characteristics for authentication, identification, or screening purposes. These technologies include, for example, documentation of fingerprints, DNA profiling, facial recognition (which identifies people based on the unique configuration of a person’s facial features), and iris recognition (which is based accordingly on the uniqueness of each person’s eye). Identification of people is most crucial when crossing international borders, but some countries have internal ID cards and identification systems. Regular use of identification forms a cumulative electronic data storage, which can then be used for profiling a particular individual.

Dataveillance is the term most often used for surveillance by means of personal data systems. Dataveillance means the systematic and automatic monitoring (of actions or communications) of individuals or groups through information systems that collect personal data. These personal data systems store a “copy” of the individual. They reduce the physical body into binary categories through which a coded counterpart, “data double,” is formed. However, the information that is gathered and coded can always only be fragmented and incomplete and so can the data double. Mistakes are easily made in the data input and these mistakes can be challenging to correct. The data double is not a mere theoretical concept, but in certain events it can “take over” one’s material body. This is the case, for instance, when a person ends up on a blacklist of international travel restrictions without knowing the reason why. Data enables, but also restricts, and surveillance data is very difficult, if not impossible, to correct or erase.

One currently prevalent form of surveillance takes place in the Internet and focuses on the activities of individuals who are online. Cookies track users across Internet sites and collect information about browsing practices. This information can then be used, for instance, in targeting advertisements for that person. Monitoring and restricting Internet use is also a powerful political tool. Internet censorship is widespread, especially in nondemocratic countries, where data traffic on social media sites and other networking tools, such as Facebook, Twitter, and MySpace, are heavily monitored or even banned. In actual fact, this censorship could be conceptualized as a global indicator of democracy and freedom of speech.

Surveillance can also specifically focus in physical space and be based on identifying location and movement. Radio frequency identification
(RFID) commonly called “tagging” is the use of small electronic devices which can be incorporated into artifacts, animals, and so forth, for the purpose of tracking by radio waves. Using RFID tagging on people is a controversial but emerging trend. Global positioning system (GPS) is another way of identifying location and movement. GPS tracking system is a space-based satellite navigation system that provides detailed information on locations. Mobile phones can also be used in order to collect geolocation data. Besides being used for surveillance, people readily use GPS systems themselves. For instance, dating-related mobile phone applications based on GPS data are widely used by people who give out their own location and in return gain information on potential spouses in the vicinity. The types of surveillance mentioned here are only a few, but already one can begin to understand the scope of varying political, societal, and personal implications succeeding these practices.

Metaphors and theories

The “Big Brother” metaphor is often associated with surveillance. This metaphor refers to George Orwell’s novel, *Nineteen Eighty-Four*, that was published in 1948. The novel describes a fictional totalitarian society with a mass surveillance system involving human operatives, lay informants, and “telescreens” in people’s homes. This novel has had a significant influence on the public imagination on surveillance: it is both based on and further nourishing the idea of centralized control.

In academic contexts, surveillance has commonly been compared to the panopticon: a prison house designed by an architect called Jeremy Bentham in the late eighteenth century. The design consists of a round structure with cells on an outer ring surrounding a single guard tower situated in the center: from the cells it is impossible to see whether or not someone is looking from the tower. However, it was not the design itself that made this prison a metaphor for modern disciplinary societies, but the panopticon’s social meanings were pointed out by a French philosopher and social theorist, Michel Foucault, in his book *Discipline and Punish* (1977).

Since Foucault’s interpretations, the panopticon has been described as an ideal architectural figure of modern disciplinary power. Its unique structure creates a consciousness of permanent visibility as a form of power, where no bars, chains, and heavy locks are necessary for domination any more. Constant surveillance makes the
exercise of power almost automatic: people are controlled, categorized, disciplined, and normalized without any clear purpose. Foucault saw this as an accurate technology of power which was designed to solve the problems of surveillance.

It is understandable why panopticon has charmed scholars in surveillance research: the metaphor has often been found to be a useful tool when analyzing the functioning of surveillant power. For instance, video surveillance has a similar feature as panopticon does: it is a technological solution designed to solve the problems of surveillance in urban space. Through surveillance cameras the panoptic technology of power has been electronically extended, almost like cities would become “enormous panopticons” (Koskela 2000).

However, as surveillance practices expand and the theoretical understanding of surveillance increases, the panoptic framing of research starts to become more restrictive than enabling. It seems that even though panopticon, like “Big Brother,” has influenced surveillance research in its early years, present-day surveillance is too complex to be explained simply by these metaphors. The Big Brother metaphor, for instance, is nowadays more often connected to “surveillance entertainment” and performing for the camera than it is understood as the grand ideal of centralized control. Similarly, panopticon fails to take into consideration the complex roles of watchers and watched in decentralized surveillance. Nevertheless, Foucault’s work does show that space has a crucial role in the exercise of power: there is a reciprocal relationship between power and space.

The political geography of surveillance

The terrorist attack on New York’s World Trade Center on 11th September 2001 has continuously been articulated as a derivation of contemporary surveillance culture. In terms of global terrorism, 9/11 was nothing new. Yet, in terms of global surveillance, it made a difference. The United States – with all its political power – has been able to intensify surveillance on a new level. Any local threats there might have been were combined with global threats. The use of the discourses of terrorism has become both politicized and anchored in global and national policies. This has caused tension in international relations and change in Western foreign policies, which now increasingly emphasize national strategies for protecting territory. Further, the global and local definitions of security have become increasingly overlapping, leading to a “rescaling and reterritorialization of security as both a concept and a practice” (Coaffee and Murakami Wood 2006, 504). Surveillance is used to protect the territoriality of the state. The state can use surveillance both to control access to its territory and to resist unwanted political activity inside its boundaries. These practices together form a particular political geography of surveillance with new attention being paid to borders.

Figure 3 Surveillance cameras are typically a part of contemporary urban renovation projects, St Petersburg, Russia. Photo by Hille Koskela.
SURVEILLANCE

Borders are an important topic in surveillance research as they are the focal points of intense surveillance practices. Security issues which previously were defined as being in between the states have now geographically changed: the conceptualizations of local, regional, and global have become fuzzy. The more eagerly people seek to protect their own and gain national security, the tougher becomes the national border control. The contemporary geopolitical forces are forcing the rise of a new authoritarianism, heightened border anxieties, hostility toward refugees, and the legitimization of violence. Hence, many ethnic and minority groups face tight control procedures as well as routine harassment and racial and religious profiling. At worst, minority groups are deemed to pose a threat to national security.

Borders are constantly challenged and redefined. Boundary-producing practices, discourses, and performances both constitute and are constituted by diverging elements of power, such as authority, coercion, domination, seduction, and manipulation (Paasi 2009). Borders no longer exist merely in the far edges of a territory but they are located deep within the territory (in international airports, inland harbors, and equivalent points of entry) or far outside it (in points of origin from where it is possible to travel into that territory).

Overelaboration of the proactive and preemptive border control strategies – such as transnational databases of foreigners, profiling of potentially threatening nationalities, or biometric control on the borders – mean that surveillance has increasing influence on human mobilities. “Intelligent surveillance” includes smart passports with biometric information such as fingerprints, iris, and facial recognition features. This will change our concept of both the dynamics of present global flows and that of belonging.

Being able to move freely and being entitled to a positive sense of place have become privileges.

The urban panopticon

The number of surveillance cameras, and the amount of urban space under video surveillance, have grown vastly in recent decades. Video surveillance is changing the nature of urban space by making it more defended (Koskela 2000). Many contemporary scholars have pointed out that the principle of video surveillance is much the same as the principle of the panopticon: to be seen but never to know when or by whom.

Urban space surveillance has become most common in publicly accessible urban spaces, such as streets and squares, transport terminals, and shopping malls. In these spaces surveillance has emerged as a means of reducing crime and fear of crime. It not only aims to protect property but also tries to reduce violence and to achieve better safety and inviolability for people. One reason for the popularity of video surveillance is its easiness and assumed effectiveness. Compared to patrolling by foot, video surveillance enables the same amount of workers to oversee larger spaces.

Despite all the policing with surveillance cameras, it remains uncertain whether surveillance cameras really reduce crime, and if so, how significantly. Research reveals contradictory results as, in addition to not having any effect at all, CCTV systems have been found to both decrease and increase crimes in separate areas under research. The positive effect of these systems also tends to fade as time goes. Furthermore, there is evidence that surveillance causes “crime displacement”: while in some cases the areas under surveillance might become safer, the adjacent areas not covered by cameras can become more dangerous. In relation to property
crime, which can be compensated, the gains of surveillance are quite obvious. In relation to violent crime, they are less so. Video surveillance operates mainly backward: it helps more in solving crime than preventing it. In the event of, for instance, street violence, surveillance might have no effect in preventing the event. This is a common reason for doubt and mistrust.

Crime prevention, however, is not the only reason for urban space surveillance. Control of urban space – public and private policing, access control, and surveillance – aims to “normalize” it – making the space appealing to “normal people” instead of “suspicious groups.” The more there are “normal” people in urban places, the less space is there believed to be for panhandlers, wanderers, and other unwanted groups. Thus surveillance contributes not only to what but to who is allowed in public space. The contradiction is that it is always a matter of debate about who is considered normal and therefore allowed. While the idealistic aim of urban surveillance is to make public space safer and more available to different groups, surveillance in fact targets various groups differently, and specifically focuses on those groups deemed “suspicious.” This heterogeneous assortment of suspicious groups include, but is not limited to, youths, political activists, people of color, people of different origin or religion, homeless, and sexual minorities. Surveillance creates knowledge based on certain assumptions, categories, and technical abilities. Since urban surveillance is widely used to monitor those perceived as deviant, it entails an element of discrimination by default. Surveillance data is used for social sorting, in which less privileged populations are disproportionately stigmatized, discriminated against, or excluded. Thus, surveillance data is no less political or value-laden than any other information.

Surveillance in urban space is also relational: in addition to considering “what” is normal or “who” is normal, it is a matter of debate “when” is normal. This means that if something is considered to be appropriate behavior in a particular time and place, it might be regarded as “deviant” in another context. The norms change depending on time and place but also according to gender, age, race, color of skin, sexuality, and so forth. People targeted by surveillance have different understandings of their relations to society; some are strong, others are marginalized and threatened. The condition of being under surveillance, accordingly, can make people react in different ways.

Emotionally, there is a big difference between being looked at by someone directly and being looked at through the lens of a surveillance camera. The variety of feelings surveillance evokes is enormous: those being watched may feel guilty for no reason, embarrassed or uneasy, irritated or angry, or fearful; they may also feel secure and safe (Koskela 2000). Clearly the experience of surveillance is also dependent on whether or not the persons know they are watched. If someone discovers that they are exposed to surveillance involuntarily, it “provokes emotional, psychoanalytic and corporeal responses which are sometimes stultifyingly profound” (Ball 2009, 644).

Quite often, however, people’s feelings toward surveillance are ambivalent or even indifferent. What is characteristic of surveillance design is its paradoxicality – while everything (and everyone) under surveillance is becoming more visible, the forces (and potential helpers) behind this surveillance are becoming less so. Ambivalence toward urban space camera surveillance might simply be because there is “no identifiable ‘watcher’ or perceivable ‘control’ being asserted” through the cameras (Ball 2009, 641).

At the societal level, the expansion of urban surveillance and protection has been claimed
to lead to a vicious circle of defense. Although increasing security might make some people feel safer, it can also create increasing fear, racist paranoia, and distrust among people. A city that is designed to look dangerous starts to feel dangerous. Furthermore, this development might lead to a “fortress” city where some parts of the city are heavily guarded and controlled while others are left to their own devices. As a result, the original effort to make public space safer and more used might actually lead to the opposite.

Counter-surveillance and criticism

In much of surveillance research the role of surveillance technology has been perceived either overly optimistically (embracing the prospect that all social improvement is technology driven) or overly pessimistically (believing that all surveillance leads to totalitarianism). Surveillance is often a violation of privacy, and as such is opposed by various civil liberties groups and activists. However, surveillance can also be understood as a “social experience” (Lyon 2014) and as such is a part of the daily lives of many. As there are numerous different and developing types of surveillance, there are at least as many approaches for dealing with that surveillance.

Obviously, the purpose of surveillance is to exercise power: to curb “deviant” behavior, to prevent crime, and to keep a space safe. The Foucauldian way of understanding surveillance treats power as something that is exercised: those who can see and who are in the position to see are considered more powerful than those being monitored. However, this apparent control engenders other forms of power, both intended and unintended. The politics of seeing and being seen are complex: who has the right to look and whom will be looked at? What kinds of power relation and structures shape the space under surveillance, and how?

Furthermore, the politics of seeing is not merely about the right to look, but the possibility to do so. In the 2000s, surveillance equipment has become readily available to an increasing amount of people. Camera phones, portable video cameras, webcams, and other visual recording technologies have become more common, are more often connected online, and can instantly link images to global information flows. Although not all of these technologies are explicitly surveillance devices, people who carry them can easily turn into surveying subjects. People are wary of being passive targets of an ever-increasing surveillance and instead seek to play a more active role in producing, circulating, and consuming visual material. And as more people have access to surveillant devices, the differences between watcher and watched, controlled and controllers fade. People start to take surveillance into their own hands, either aiming to resist the official forms of surveillance or wanting to create their own practices of monitoring.

Resisting surveillance can take many forms. At the societal level, one can implement resistance by challenging the legitimacy of various surveillance measures, making surveillance practices more visible and recognizable and attempting to raise public debate about them. Critical people can also take surveillance equipment into their own use, turning surveillance against the authorities. For instance, video activism as a form of counter-surveillance targets police and other authorities in demonstrations and such in an attempt to prevent possible police brutality and provide evidence in such cases. On a personal level, resistance can be lived out by aiming to avoid surveillance systems in private life – to be anonymous despite surveillance. These societal and personal forms of resistance have been conceptualized as counter-surveillance.
While counter-surveillance always involves some form of criticism toward surveillance or a certain kind of political agenda, people also adopt surveillance equipment into their own use without a specifically critical stance. This can include, for instance, producing or watching visual material for one’s own purposes or simply playing with surveillance equipment. In surveillance research this type of “resistance” has been labeled “hijacking surveillance” (Koskela 2011). More than an organized form of political resistance, hijacked surveillance is about random witnessing (and recording) of incidents.

The transparency fostered by surveillance no longer operates in a single direction as people are involved in various counter-surveillance practices that can reveal official misconduct. Similarly the surveillant power is no longer merely in the hands of the authorities, but anyone, anywhere, can become a surveillant.

**SEE ALSO:** Biosecurity; Crime; Emotional geographies; Political geography; Privacy, personal privacy; Public space; Security; Urban geography

### References


Surveying

Andrew J. S. Hunter
University of Calgary, Canada

Surveying is the science of collecting data about the location of geographic features, and drawing those features on a map, plan, or digital model. Geographic features include natural landforms, lakes and rivers, bridges, roads, and other things that have been built by people, and even abstract things such as property boundaries. In this context surveying is all about measurement, the technologies, and methods used to locate a feature in two- or three-dimensional (3-D) space, and to determine its size. Common forms of surveying include land surveying (determining where (property) boundaries are), topographic surveying (collecting 3-D data about the shape of the Earth and the location of features such as rivers, lakes, roads, trees, etc.), and “geodetic” surveying (the measurement of the size and shape of the Earth, and the setting up of triangulation networks that define coordinate systems to which topographic details can be mapped).

Today, wherever there is land development, there is also surveying, whether above or below ground, on or in water. Surveying ensures that new developments fit with the old.

Surveying requires knowledge of mathematics, astronomy, geography, physics, mechanics, metrology, statistics, geophysics, and other scientific disciplines. Practitioners of surveying, surveyors, also require a theoretical and practical understanding of scientific measurement devices. Surveying locates features on the ground by measuring angles and distances with respect to a datum, a geodetic coordinate system, or a map projection. With recent progress in technology, global positioning systems (GPS), for example, time is also critical. To measure accurately, surveyors are most concerned with the effect of the atmosphere, as changes in temperature, pressure, and humidity will affect measurements. Surveyors are also concerned with the geometric configuration of measurements to ensure that the necessary positional accuracy can be achieved through a network of measurements.

The earliest historical records suggest that surveying began in Egypt around 1400 BC. Herodotus, the fifth-century BC Greek historian, recorded that the ancient Egyptian pharaoh Sesostris divided the Nile Valley into equal-sized plots and allocated them to Egyptians. He then raised the revenue he needed by levying a tax on each landholder. Annual flooding of the Nile River would wash away portions of the plots and surveyors would be sent to replace the boundaries and calculate how much land landholders had lost. These early surveyors were known as the dragger of the rope, since they measured distance using a schoinion (a measuring cord) 100 cubits long (150 ft or ca. 45.7 m), with knots every 12½ cubits (18.7 ft or ca. 5.7 m).

The science of geometry and trigonometry formalized by early Greek mathematicians such as Euclid and Archimedes is believed to have arisen, in part, out of the practical demands of surveying. Greek and Roman surveyors worked exclusively in Euclidean geometry using the concept of similar triangles to compute areas and volumes. Hero of Alexandria developed several treatises, including the Dioptra, which discussed field survey methods, survey calculations, and the drawing of plans. Many of the
surveying methods are still recognizable in today’s survey practices. The *Dioptra* also discussed the first survey instrument, a Hellenistic period dioptra, a precursor to today’s theodolites. Julius Africanus, Anonymous Byzantinus, and al-Karaji also wrote similar works. The Romans developed a number of survey instruments, the *libra aquaria* (water leveler or balance) being the most well known, which was used for leveling during the construction of engineering works.

The profession of the ancient surveyor, whether Greek or Roman, consisted of four distinct categories.

1. The land surveyor (in Greek, *geometres* or *geodaiestes*; in Latin *finitor, mensor, agrimensor*, or *gromaticus*) carried out local surveys on the ground. They recorded the shape of existing fields worked by landholders to calculate their area so that taxes could be levied. They might divide land into rectangular plots for allocation to settlers, or they might set out a grid of streets for construction in a town. In a military context they laid out fortifications or *castra*. A land surveyor was concerned solely with horizontal measurement. There was little need for measurement of heights, as all “cadastral” survey work was assumed to be on a plane.

2. The cartographic surveyor (*chorographos, geographos*) made maps, generally of large areas covering potentially many regions. From this work arose inquiry into the shape of the Earth, so that large areas fit together on the map; and the heights of mountains, for orientation and as points of interest on the maps.

3. The military surveyor (*mensor*) provided information to commanders and engineers. They developed indirect measurement techniques because the features they were measuring were typically occupied by hostile groups: the height of a city wall, so that ladders or a siege tower of the right height could be built; or the width of a river, so that a pontoon bridge could be built to the correct length.

4. The engineering surveyor (*mensor* or *liberator*) surveyed the terrain so that roads, aqueducts, irrigation channels, navigable canals, tunnels, mines, harbor works, and so on, could be constructed.

As is the case today, these categories were not exclusive. Surveyors used the same instruments and similar techniques for each type of survey. Astronomers who, particularly after the development of spherical trigonometry, wanted to measure angular distances between stars and planets also shared in the use of many of the instruments and techniques. History records that Eratosthenes estimated the size of the Earth ca. 200 BC. He reasoned that, because of similar geometries, measuring the distance between the two Egyptian cities of Alexandria and Syene (he believed they lay on a similar longitude) and measuring the length of the shadow cast by a pole of known length at Alexandria (he had observed that the sun was directly over Syene at noon on the summer solstice), he could compute the angle subtended at the center of the Earth between the two cities, and thenceforth compute the earth’s circumference.

As land has become scarce its value has increased, as has the demand to establish precise boundaries. The need for public works has also ensured that the value of surveying to society has remained high. The ongoing call for economic growth, and the use and preservation of earth’s resources will continue to drive the demand for surveying.

Surveying and mapping tools have now evolved to the point where the traditional instruments – theodolites, the level, the chain
or steel band – have almost been completely replaced by high-tech instruments, such as robotic total stations that can measure horizontal and vertical angles and distances automatically; and global navigation satellite systems (GNSS) such as the Global Positioning System (GPS), which can provide accurate data for virtually any type of survey. Terrestrial and airborne laser scanning instruments combine automatic distance and angle measurement to compute dense 3-D clouds of points that can quickly be converted to 3-D models. New aerial cameras and remote sensing systems can collect digital images that are then processed to obtain spatial information. Integrated systems that combine GPS, inertial measurement units, digital cameras, and other sensors can map items within tens of meters of a road at highway speeds, and the data collected can be provided to users as georeferenced coordinates overlaid on high-quality imagery. With these changes in technology, the term geomatics has now become synonymous with surveying, and covers all methods of measuring and collecting data about the physical Earth and our environment, and processing the data so that they can be used for planning and management of the land, the sea, and any structures thereon. Surveying and geomatics also includes research into the practices and instrumentation to develop them further.

As noted, surveying has a long history. From the earliest times, it has been necessary to mark boundaries and divide land. Today, surveying remains indispensable because of the need to map the Earth above and below sea level; to prepare navigation charts for use in the air, on land, and at sea; to establish property boundaries of private and public lands; to develop databases of land-use and natural resource information that aid in management of the environment; and to determine facts about the size, shape, gravity, and magnetic fields of the Earth.

Surveying and geomatics continues to play an important role in many branches of engineering. Surveys are required to plan, construct, and maintain highways, railway lines, rapid transit systems, buildings, bridges, tunnels, canals, residential land development, storm water and sewerage systems, pipelines, mines, and so on. Surveying methods are also commonly employed to lay out industrial assembly lines to minimize the wear and tear of moving parts, and to guide the fabrication of large equipment such as planes and ships where separate pieces have been assembled at different locations, and must fit together. Surveying also plays a role in many other industries today where accurate measurement is required – agronomy, archaeology, geophysics, meteorology, and seismology, to name a few.

SEE ALSO: Global navigation satellite systems; Ground-based LiDAR; LiDAR; Photogrammetry: 3-D from imagery; Sonar; Synthetic aperture radar

Further reading

Sustainability science

Robert W. Kates
Independent scholar

Sustainability science is an emergent field of science whose origins are found both in elite institutional efforts to bring environment and development together and in an outpouring of published studies related to sustainability or sustainable development. Sustainability science is characterized by use-inspired research and as such similar to the agricultural or health sciences. It is defined by the practical problems of sustainability which it addresses and by theories and models of interactions between natural and social systems. It is clearly transdisciplinary in structure integrating research from the natural, biological, social, and engineering sciences. Above all, sustainability science seeks to link knowledge with action on the identified problems of sustainability.

Origins of sustainability science

A series of major reports and international meetings document the first efforts to bring together environment and development beginning with the 1980 International Union for Conservation of Nature (IUCN) World Conservation Strategy, followed by the emergence of sustainable development in the 1987 Bruntland report, Our Common Future, and acted upon at the 1992 first Rio Earth Summit. Sustainable development, while politically successful, had little science content until the preparation of the US National Academy of Sciences 1999 report, Our Common Journey, that called for the development of a sustainability science and followed in 2003 by its journal, Proceedings of the National Academy of Sciences USA (PNAS), creating a sustainability science section now containing hundreds of research articles.

Figure 1 also graphs the individual authors of articles and conference proceedings written in English between 1974 and 2010 with “sustainability” or “sustainable development” in the title, abstract, or keywords. There were over 20,000 papers, published mostly in biology, engineering, and social science journals, with 37,000 distinct authors based in such diverse institutions as corporations, governments, NGOs, and universities (large and small) in 174 countries and 2200 cities. From the 1990s on, the number of authors of articles grew rapidly, doubling about every 8 years, with its worldwide authorship, including many developing countries. By analyzing the author’s interconnections, the field seems to unify around 2000, when the size of interconnected clusters of authors began to grow and the number of disconnected clusters of authors and reflecting disciplines rapidly declined.

Sustainability science as use-inspired research is often illustrated by Donald Stokes’s quadrant model of scientific research contrasting the quest for fundamental understanding with considerations of use by society. The quest for fundamental understanding is called, by some, basic science in contrast with applied science characterized by user needs. But a third option
is use-inspired basic research as exemplified by such researchers as Louis Pasteur and such fields as the agricultural or health sciences. Within this category, sustainability scientists address use-inspired problems of sustainability, perhaps emphasizing interactions between natural and social systems, perhaps emphasizing applied solutions to practical problems, or both.

Sustainability science as interactions between natural and social systems

Basic overviews of interacting natural and social systems have been used for many years labeled variously as nature–society, socio-environment, socio-ecology, or human environment and often presented as box–and–arrow diagrams rather than as operable models. Geographers have played major roles in developing this view of nature–society relationships and related research. This research, which is a study of how interactions between natural and social systems emerged in global change, notes the key roles of eight geographers whose discipline “by training, bridged social and natural sciences and had a strong spatial focus; furthermore, they had a cadre of scientists with strong leadership attributes” (Mooney, Duraiappah, and Larigauderie 2013, 3).

These box–and–arrow diagrams of nature–society interactions usually contain a societal or human subsystem, a natural or environmental subsystem, and other external systems that impact or modify these subsystems. Analysts differ on the important elements of each. Judging from the literature, much of nature–society systems study focuses on the environmental subsystem, especially its production of ecosystem or environmental services (provisioning, cultural, supporting, or regulating). Important elements of social subsystems are population, technology, governance, and economy, but these are seldom interacting and studied primarily as separate elements. If present, they are introduced into models usually as external inputs or scenarios.

As a research field with a theoretical perspective on coupled nature–society systems, sustainability science has evolved a set of seven core questions and research themes. These are as follows.

- What shapes the long term trends and transitions that provide the major directions for this century?
- What determines the adaptability, vulnerability, and resilience of nature–society systems?
- How can theory and models be formulated that better account for the variation in nature–society interactions?
- What are the principal tradeoffs between human well-being and the natural environment?
- Can scientifically meaningful “limits” be defined that would provide effective warning for nature–society systems?
- How can society most effectively guide or manage nature–society systems toward a sustainability transition?
- How can the sustainability of alternative pathways of environment and development be evaluated?
Sustainability science as applied problems of sustainability

The core questions and research themes can also be asked of specific places or sectors of human activity where applied problems of sustainability occur. Six initial sectors were first identified in 1987 by the World Commission on Environment and Development as global challenges for sustainable development: population and human resources, food security, species and ecosystems, energy, industry, and urban. With the addition of water, these have appeared in various guises, the latest being water, energy, health, agriculture, biodiversity, and urban.

Scientists and technologists who wish to help provide knowledge useful in addressing such sustainability problems do so best when these problems are determined collaboratively with potential users and policymakers, who often have their own traditional or practitioner knowledge to contribute. Such coproduction of knowledge enhances the search for solutions and the likelihood that use-directed knowledge will actually be used.

Transdisciplinary sustainability science

There are three kinds of major interactions between traditional disciplines when trying to understand or address a scientific or practical problem. In its simplest form, multidisciplinarity, which is when various disciplines present their separate perspective of the problem and the research on it found within their discipline. A more interactive form is interdisciplinarity, where relevant disciplines are brought together, usually in teams, to conduct joint research on the problem. Much of sustainability science functions in this way. But, increasingly, many sustainability scientists aspire to transdisciplinarity, moving beyond the approaches, training, and skill sets of their disciplines to become problem oriented rather than discipline oriented. In so doing, they seek to master and use the skills and approaches needed to address the problem, from whatever sources these may be derived.

Sustainability science in 2015

Since its unification as a scientific field in early 2000, sustainability science has become rooted in the scientific enterprise with new research, publications, institutions, and educational opportunities. As noted, its research increasingly seeks solutions to problems along with deepening theoretical perspectives in coupled nature–society systems. It has a very large published library of work, much appearing in new dedicated journals. Centers and institutes and a growing number of degree programs dedicated to sustainable development, sustainability, or sustainability science are found in many parts of the world and in their variety reflect the regional differences in how the science is pursued and that no common sets of needed competencies for students have yet emerged.

But sustainability science is a young science and there is much more work to be done. Most important is moving knowledge into action. The practical problems of sustainability are increasingly studied in joint undertakings with practitioners from local communities, industry, government, and civil society, but there is a lag between research and implementable solutions offered or problems solved. (A set of videos of such problem-solving in the state of Maine in the United States is available at www.mpbn.net/Telelevision/LocalTelevisionPrograms/SustainableMaine.aspx.) The next major metric
SUSTAINABILITY SCIENCE

for sustainability scientists should be in practical problems solved.

SEE ALSO: Environment and development; Environmental science and society; Socio–nature; Sustainable development

References


Further reading


Sustainable cities

Pierre Filion
University of Waterloo, Canada

Sustainability is about securing the conditions for the long-term existence of an organism or system. In its urban application, it concerns primarily the relation between cities and the environment, although the concept of sustainability can also pertain to social and economic requirements for the durability of urban systems. Different motivations prompt cities to adopt environmental sustainability trajectories. Cities can strive to maintain or improve the quality of their own environment to provide healthy living conditions and thus assure the wellbeing of their residents. Another reason for cities to engage in such initiatives is to preserve resources essential to their long-term survival. Cities can also commit to a lessening of their contribution to global environmental damage – most notably emissions of greenhouse gases. In this instance they assume the role of instruments of global sustainability. The emphasis in this entry is on the environmental, rather than social and economic, dimensions of sustainability, and mostly on sustainability issues confronting cities in developed countries.

Sustainable development and its application to cities

The expression “sustainable development” was coined in the 1987 report of the United Nations World Commission on Environment and Development. The Commission was chaired by Gro Harlem Brundtland, a former prime minister of Norway. The report, entitled Our Common Future, defined sustainable development as “development that meets the needs of the present without compromising the ability of future generations to meet their own needs.” The concept has been a catalyst for environmental consciousness in many parts of the world, and was soon adopted as a standard to strive for in many areas of human activity, including the planning of cities, the construction of urban infrastructure, and the delivery of urban services. It is important to note, however, that sustainable development has remained more an object of discussion than a source of radical policymaking and human behavior transformation.

With over 50% of the world population residing in urban areas, and given the concentration of economic activity in these locales, it is to be expected that cities will be a foremost target of sustainable development strategies, whether deployed on a local, regional, national, or international scale. The impact of cities on the environment is further accentuated by the fact that they are primary sites of consumption, which causes them to draw resources from extensive hinterlands. Relationships between cities and their environment have always been critical of their existence. Jared Diamond (2005) has chronicled how throughout history human-caused deterioration of the environment and an inability to adapt to environmental change have provoked the collapse of civilizations and their cities. The origins of the planning profession and of present regulations guiding urban growth can be traced back to efforts to safeguard public health in
the severely degraded environmental conditions of nineteenth-century industrial cities. People living in these cities were victims of repeated epidemics and overall poor health, in large part due to contaminated water, foul air, and crowded living conditions. Progressively, over more than a century, public policy addressed these perils. These policy responses have culminated in the elaborate urban development control mechanisms found in the cities of rich countries, which can be partly credited for their sanitary conditions (Porter 1999). Still, one must not lose sight of the fact that the cities of poor countries, and especially their informal settlements, experience living conditions comparable, if not worse than, those of early industrial cities.

The lens through which environmental impacts are perceived has widened over time. From a consideration of immediate health effects, the perception of these impacts has broadened to include long-term health consequences, and the regional and then global environmental consequences for cities, along with the repercussions of these consequences on humans. In the 1960s and 1970s, much attention was paid to the health effects, many of them long-term, of the rapidly deteriorating air quality, and of regional impacts of water pollution. In rich countries these problems were partially addressed by relying on technical fixes: fitting catalytic converters to car emission systems, using cleaner heating fuels, and building sewage treatment plants. Sustainable development introduced a longer-term and more comprehensive perspective on the relationship between cities and the environment. It raised concern about environmental consequences of all aspects of urban life and development, and led to environmental proposals ranging from less energy-demanding patterns of development to recycling programs.

Efforts at measuring the environmental load of urban areas in terms of the resources they consume and the space needed to absorb the waste they generate have cast further light on the global environmental repercussions of cities. One method for representing these impacts of cities is the urban ecological footprint, which transposes into spatial units the carrying capacity of the Earth needed to support cities. Studies adopting this approach conclude that humanity consumes resources at a rate that is 1.5 times in excess of the carrying capacity of the Earth, thereby mortgaging coming generations. Worse still, if the entire world achieved the standard of living of rich countries, resource and absorption capacities of several Earths would be needed (Rees 2006).

Since the late 1970s, mounting evidence of global warming and its association with human activity has cast a sobering light on the limited absorption capacity of planet Earth. Here again, cities come out as foremost contributors to this environmental threat.

The sustainable development of cities presents itself differently depending on whether they are in rich or poor parts of the world. McGranahan and Satterthwaite (2002) have formulated an urban environmental transition theory, which links different environmental agendas to distinct economic development stages, and echoes the evolution of environmentalism in countries of the Global North since their industrialization. Environmental attention in poor countries is monopolized by water supply and sanitation. As countries industrialize, they become confronted by the pollution generated by industrial production and consumption, especially automobile use. Finally, in their postindustrial phase, cities are allured by green agendas driven by concern over greenhouse emissions and other side effects of consumerist lifestyles. Perceived in this fashion, the scope of environmental preoccupations becomes ever larger as countries make their way through the different stages of the transition, from a household to a planetary
Sustainable city models

There are different versions of sustainability. Weak versions accommodate adjustments to production and consumption, which do not challenge their fundamental mechanisms. Cities maintain their present form and dynamics, but see their environmental footprint somewhat reduced by recycling programs, improved home insulation, and more energy-efficient automobiles, for example. Strong versions of sustainability are more ambitious. From their perspective, sustainable development should involve the redefinition of all human systems. Many researchers believe that nothing short of such radical transformations can avert the crises on the horizon, such as global warming. This thinking reverberates on urban planning. One manifestation of this is the designing of new cities that profoundly curtail carbon emissions, by lessening dependency on coal, gas, and oil. The environmental footprint of such cities would be considerably smaller than the prevailing urban norm.

Masdar City, a new town in Abu Dhabi in the United Arab Emirates, has been designed by Foster and Partners to host 50,000 residents in a near carbon-free environment. Another example is the proposal for a new Chinese city with a population of 80,000 planned by Adrian Smith and Gordon Gill, which would be fully pedestrian-oriented, allowing all destinations to be reached within 15 minutes on foot. Great City, located close to Chengdu, will produce 60% less carbon dioxide per capita than other Chinese cities do. An early prototype of an environmental community is Arcosanti, the brainchild of Italian architect Paolo Soleri, which has been under construction in the Arizona desert since 1970. The town, to house 5000 people is intended to demonstrate how cities can function in close harmony with nature.

These efforts at rethinking urban development according to its impact on the environment are consistent with the tendency to design new urban forms that address one specific issue. The purpose of the Garden City at the turn of the twentieth century was to provide an alternative to crowded industrial cities. Likewise, the Radiant City, consisting of towers surrounded by green space, purported to address the unhealthy living conditions of the densely built cities of the early to mid-twentieth century (Fishman 1982). The present tendency is to formulate new urban models according to sustainable development principles.

Specific sustainable urban interventions

Given their extreme rarity and small scale, we cannot expect new cities entirely designed around sustainable principles to contribute much, collectively, to an improved global environment. Moreover, their high construction costs are likely to cause them to become environmental havens for the rich only (Hodson and Marvin 2010). More likely to have a positive influence on global environmental conditions are more narrowly focused, but also widely adopted, measures promoting urban sustainability within existing or new urban areas conforming to prevailing land-use standards. These measures concern multiple aspects of the urban phenomenon, including those exemplified in the next paragraphs. The resources the erection of the built environment requires, and the energy to heat or cool them, make buildings a natural target for environmental measures. In addition, the built environment of a city
SUSTAINABLE CITIES

affects consumption choices and, more generally, lifestyles. High density can encourage public transit patronage and walking. Meanwhile, spacious living environments are more conducive to the accumulation of consumer goods than are smaller ones. One way of decreasing the environmental impact of buildings is to rely on infill development, which consists of surgically redeveloping parcels of the urban texture rather than demolishing and rebuilding wide swaths of the city. The replacement of large urban sectors results in the loss of the resources and energy locked in the existing built environment and the channeling of additional resources and energy towards the construction of the subsequent buildings. Least environmentally harmful is the retrofitting of existing, rather than the erection of new, structures.

The LEED (Leadership in Energy and Environmental Design) certification rates buildings according to their respect of the environment. Certification involves an evaluation of over 100 aspects of buildings, including water and energy efficiency, material selection, and innovation in design. Launched in 1998 in the United States, LEED certification has spread to many parts of the world.

An innovative approach to the creation of an urban built environment that consumes less natural resources and energy is to turn to biology to inspire design – a process called “biomimicry.” The outcome of this approach, which is adopted mostly in rich countries, is new types of building that borrow from biological forms and processes: roofs that espouse the shape and structure of leaves, for example.

Transportation represents the other principal source of energy consumption in cities. As expected, urban sustainability efforts aim at abating reliance on automobiles, the form of transportation that is the foremost consumer of energy and contributes most to carbon emissions. Public transit improvement and expansion along with programs encouraging walking and cycling thus fall under the sustainable transportation rubric by virtue of their reduced energy demands relative to the car. Transit Oriented Development and New Urbanism are examples of efforts to combine urban form and transportation systems in an attempt to create public transit- and walking-conducive urban settings (Schiller, Bruun, and Kenworthy 2010).

Urban sustainability efforts also target the production and consumption of energy. The main issues arising from energy generation and use are the destruction and contamination of extraction and generation sites (coal mines, oil fields, hydraulic fracturing, large reservoirs for hydro-electricity, nuclear power plants), air pollution, and above all, contribution to global warming. From an energy production perspective, advocates of sustainability call for a rapid transitioning from methods that carry a heavy environmental burden – those relying on coal, oil, and natural gas, or nuclear power, whose system failure comes with tragic consequences – to processes that are in harmony with nature – for the most part, solar and wind power generation. Solar power lends itself to a microproduction of electricity at the individual building level, marking a shift away from a concentration on generation in large power plants. On the consumption side, attempts are made to reduce demand through higher-efficiency devices. Until now, however, the aggregate effect of these measures on energy consumption has been limited.

Other initiatives striving to enhance urban environmental sustainability include increased reliance on green infrastructures, which depend as much as possible on natural processes. A vivid illustration of this renewed respect of nature is the change in the treatment of streams in urban areas. Until recent decades, the dominant approach
involved burying them in storm sewers, which caused serious deterioration of water quality and, in combination with the impervious surfaces that cover much of cities, heightened flood risks. In North America, the tendency is now to “daylight” streams (by bringing them back to the surface), and naturalize areas abutting them.

There is also rising interest in the quality, availability, and environmental implications of food within urban areas. One major food-related issue is the concentration of food outlets in increasingly large premises, causing the spread of “food deserts,” especially in low-income sectors. At the same time, efforts are made to bring food production close to urban consumers, be it through the 100-mile diet encouraging the purchase of food produced close by to reduce transportation needs and assure freshness, or through a proliferation of urban gardens.

Finally, there are also measures being taken to reduce the waste produced by cities. Many municipal administrations have adopted recycling programs, which contribute to divert waste from landfill sites. There is a great deal of difference in the goals and achievements of these programs. Few cities approach the 80% diversion rate attained in San Francisco. Most ambitious are proposals for zero-waste and zero-emission closed-loop systems whereby at the end of their consumption phase all the components of goods are transformed into new products, which then share the same fate at the end of their own consumption period. Closed-loop production and consumption are an example of biomimicry, as they attempt to reproduce natural cycles which take place without generating waste.

The question that emerges from the consideration of all above-mentioned sustainable urban interventions is whether, in the event of their full implementation, their cumulative effects are sufficient to make cities as environmentally benign as if they were designed from the start around sustainability principles. There is no clear-cut answer to this question. Past and present experience, however, points to the much more likely adoption of incremental narrow-focused approaches to urban sustainability than a radical redesigning of urban areas.

The present state of urban sustainable development

The reality of urban development rarely coincides with the commitment to sustainability expressed in plans. In North America, the resulting discrepancy is currently illustrated in the gap between a near-universal adherence of metropolitan-scale plans to public transit, recentralization (downtown expansion and the creation of networks of dense and multifunctional nodes), and overall compact development on the one hand, and the persistence of automobile-oriented sprawl on the other. Prominent among the obstacles preventing metropolitan plans from achieving their objectives are closely intertwined relationships between land use and transportation. Such interrelations constitute a powerful path dependence, making it difficult to substantially modify either land use or transportation systems because transformative policies must have a simultaneous large-scale effect on both land use and transportation to induce desired changes.

Incidentally, just as North American planners are becoming concerned about the adverse consequences of automobile-dependent urban land uses, this urban pattern is gaining popularity in new middle-class suburbs in different parts of the world. Issues related to car-reliance in large cities are increasingly universal phenomena.

There is also resistance to change from interests deeply vested in current patterns of urban land use. Additionally, global consciousness of the
need for urban sustainability as global warming progresses does not necessarily translate into coordinated responses. In fact, often the opposite happens. Some nations, notably the United States, Canada, and Australia, point to the inaction of other countries to justify scaling back efforts at containing their greenhouse gas emissions.

Over the past decades, sustainable development initiatives have mirrored the sharp discrepancy between the financial capacity of developed and developing countries. For example, public transit strategies of cities in the Global North emphasize expensive rail systems; light rail transit in many instances. Meanwhile, Latin American cities, notably Curitiba, Brazil, and Bogotá, Bolivia, have based their public transport strategy on more affordable bus rapid transit systems (Hidalgo and Graftieaux 2008).

Debates on sustainable cities

Beside the uncertainty about how to transform sustainable urban development principles into actual urban change, there are questions about which are the most effective urban sustainability formulas. One such question concerns the opposition between polar perspectives on optimal urban form from a sustainable perspective. The first proposes compact, high-density settlements to minimize the footprint of cities and lessen energy requirements and emissions, especially by reducing dependence on the automobile. In this vein, David Owen (2009) celebrates Manhattan as a model of sustainability. In his view, despite the wealth it concentrates, the exceptionally dense and public transit-reliant urban form of Manhattan accounts for much lower per capita energy consumption than is registered in other US urban areas. The opposite approach calls for a blending of urban development with natural systems. The outcome is a low-density urban form structured around protected natural features. This view was first articulated in the 1960s by Ian McHarg and is now promoted by the landscape urbanism movement (Waldheim 2006). The problem with this option is an enduring reliance on the automobile, with ensuing high energy needs and environmental repercussions. Perhaps, in a not so distant future, the generalization of electric traction will reduce some environmental downsides of the car.

Recently, sustainability has been criticized for the change of emphasis it brings to the study of urban issues. In this regard, Michael Gunder (2006) has noted an association between the rise of interest in urban sustainability and reduced attention to social equity.

SEE ALSO: Environment and urbanization; Environmentalism; Sustainable development; Urban geography; Urban planning: human dynamics

References


Further reading

Sustainable development

Rob Krueger
Worcester Polytechnic Institute, USA

Every generation comes around to the question at some point, at least in so-called developed countries: How long will the material conditions exist to support our economic systems and lifestyles? Sustainable development has been one conceptual means of organizing our responses to this question. For a generation, those advancing the idea of sustainable development have sought to render visible previously “unseen” environmental and social costs associated with the development and allocation of resources. Sustainable development represents the first effort by world leaders to explicitly consider the broader environmental consequences of economic development and the geographical distribution of the benefits derived from it. However, students of geography know that this approach is not as simple as the champions of sustainable development had hoped or thought it would be. Sustainable development is a social construct – this means the concept was founded in a certain political-economic context that represents the product of interactions between people and institutions with varying interests and access to power. When its various forms are examined and judged, what should be understood, above all, are the relations of power and the impacts they have on humans and nonhumans, so that they can be viewed with a sharp, critical eye and their dangers, as well as their promise, identified. Indeed, sustainable development is a framework for organizing economies, but it is also a lens through which to view these relationships (see Political economy and regional development). This entry will explore the concept of sustainable development at the macro and local scales, as it emerged as a response to earlier beliefs (also social constructs: see Social constructionism) about societal and economic development, contemporary social relations, and economic imperatives.

Sustainable development and its antecedents

Sustainable development is a modern concept only in name. Humans have attempted to live sustainably for millennia. Hunter-gatherer societies were nomadic because they followed food. Once humans learned agricultural techniques, sustainability took on another meaning: the ability to produce adequate food for the social group year on year. During the Industrial Revolution, the English were concerned about establishing a sustainable energy supply. As European human populations grew, philosophers and economists raised concerns about resource scarcity. How long could nature support agricultural and then industrial endeavors and aspirations? Scholars such as Malthus, Ricardo, Jevons, and the French physiocrats all wondered – and sometimes strongly disagreed – about the capacity of land and other resources to support human communities. In these different examples, and in our own recent history, diverse understandings of sustainable development have been shaped by our different social contexts and institutional systems, technologies, and authority figures. This point can be illustrated with an important historical example.
SUSTAINABLE DEVELOPMENT

Sustaining resource access, and sustaining “Western” values in a Cold War context

With the full onset of the Cold War in the early 1950s, uncertainty about the scarcity of natural resources resulted in tangible effects. Commodity prices rose dramatically during the first year of the Korean War (1950–1953). Western leaders asked themselves whether they had access to enough raw materials to defeat the Soviet Union – not for military use, but to provide industry with the materials needed to meet the rapidly expanding demand for goods and services since the end of World War II. The Paley Commission, named after its chairman William Paley, president of the Columbia Broadcasting System (CBS), was appointed by the US president to examine commodity needs based on population projections to the year 1975. Using commodity price as an indicator, the Commission found that US demand would grow significantly and that supplies of some resources were adequate to support the economy. However, the Commission found other raw materials, as a matter of price, to be increasingly scarce. Thus the report suggested that an adequate supply of these materials could be sustained if it was acquired from other sources, but at an increasing cost. Therefore, according to the report, the government would have to identify and procure new supplies while at the same time searching for substitutes. Two elements of this report informed future discussions about sustainable development. First, the Paley Commission recognized the difference between absolute scarcity and relative scarcity. Relative scarcity arises from the lack of control over existing resources. The second was the doctrine of technological substitution, which means that technologies could change to require fewer or different resources to support the same function. Both of these points would later figure in the version of sustainable development famously proposed in the United Nations report *Our Common Future*, or the Brundtland Report (World Commission on Environment and Development (WCED) 1987).

In response to its insight about relative scarcity, the Commission looked to identify government actions that might secure the resources necessary to sustain Western economies and their militaries. For decades, Western economic powers had been funding explorers to identify and map the locations of key natural resources: oil in the Middle East, copper in Chile, and gold in South Africa, for instance. At the same time (the 1950s to the 1970s) these same nations were undertaking a process of relationship building and policy exports. Key aspects of these policy exports were based on W.W. Rostow’s ideas about stages of economic development, which culminated in “modern” or “industrialized” economic forms. These stages provided Western governments with a framework for creating model regimes through policy interventions, aid packages, and consultations for countries seeking industrialization. As a result, the West nurtured new political regimes in Iran and in Chile, and maintained a convivial relationship with the apartheid regime in South Africa. In these ways, for the sake of securing natural resources in scattered regions across the globe, the US-led Western nations made various alliances that blatantly defied their own explicit principles about democracy and human equality.

The second important insight of the Paley Commission, the doctrine of technological substitution, proved to have even more powerful and long-standing implications. Indeed, the notion that technology will provide solutions to environmental and economic problems remains a core belief for many. Technological substitution was popularized by Paley’s Commission and codified a decade later by the academic economists Harold Barnett and Chandler Morse in their book, *Scarcity and Growth* (see Environment
and resources, political economy of). Barnett and Morse (1963) echoed the Commission’s findings that, under conditions of increasing scarcity, rising raw material prices would trigger innovation and a process of technological substitution. These concerns revived discussions from the previous century by neoclassical economists about the possible decoupling of material scarcity and economic development. Of course, the notion of sustainable development as it is now conceived results from myriad conditions, but the arguments of the Paley Commission and of Barnett and Morse would have had a lasting effect on the political, economic, and discursive underpinnings of the sustainable development discourse. In contrast to the arguments put forth in the Paley report, which characterized economic development as a deliberate effort requiring risky and unethical international policies, the sustainable development concept argued that development is a sociopolitical process that works on its own in some deterministic way.

With this background in place, the entry now turns to this concept and various attempts to define it.

Simply sustainable development

The concept of sustainable development proposed a set of organizing principles for future economic growth. The phrase itself was coined in the 1980s, and shaped by the environmental movement and a recognition that our material relationship to nature was harmful to humans and the earth systems on which they depend. The environmental movement resulted from industrial disasters and risks in the United States and elsewhere, which led to debates about the connections between economic growth and environmental degradation. Debates also took place in academia—most prominently in the Romanian economist-mathematician Nicholas Georgescu-Roegen’s radical ideas linking economic systems to the Second Law of Thermodynamics. The concept of scarcity took a central position in the internationally recognized Club of Rome group (see Environment and development). In 1972 the club published its now famous system dynamics model which, for them, demonstrated the “limits to growth” because of the world’s finite ecological capacities. A few years later Herman Daly, father of ecological economics, published a now well-known work that called for a steady-state economy, one characterized by mildly fluctuating population growth and energy input. These new perspectives were informed not only by environmentalism but also by a systems thinking approach. In contrast to their predecessors, these analysts did not see single resources as being scarce inasmuch as they understood the systemic capacity of the planet’s ecosystems.

The term “sustainable development” entered public discourse in the early 1980s when the International Union for the Conservation of Nature and Natural Resources (IUCN) published its World Conservation Strategy (1980) (see Conservation and capitalism). Like its more widely known counterpart, the Brundtland Report, the IUCN report addressed the escalating, environmentally destructive, and systemically damaging approach to the use of natural resources. The report cited a long list of hazards and disasters, including soil erosion, loss of agricultural lands, industrial pollution, deforestation, ecosystem degradation, species extinction, and more. It also called for a new global ethic grounded in the interrelatedness of global development. Questioning the reasoning of previous approaches, this ethic charged all nations with identifying coordinated global strategies that outline strategic environmental
SUSTAINABLE DEVELOPMENT

conservation measures. Most importantly, the interrelatedness of global development required that development and environmental conservation be considered equally with mutually reinforcing goals.

Environmentalists celebrated the Brundtland Report (WCED 1987) as a large step forward in thinking on economic development, the environment, and the human condition. The report was grounded in concepts of “limits”: economic and technological activities can be improved but we are still limited by the “biosphere to absorb the effects of human activities” (WCED 1987, 8). It distinguished itself from its immediate predecessors by asserting environmental quality and social equity as fundamental goals of development for both impoverished and wealthy nations. Where the IUCN report limited its concern to environmental degradation and identified conservation as the remedy, the Brundtland Report argued that growth must be revived in developing countries because that is where the links between economic growth, the alleviation of poverty, and environmental conditions operate most directly. Yet, developing countries are part of an interdependent world economy; their prospects also depend on the levels and patterns of growth in industrialized nations. (WCED 1987, 51)

Given that its progenitors recognized the interdependence (between developing and industrialized nations, between economic growth, social justice, and environmental stewardship), it is not surprising that the report was explicit about the role of social and economic conditions in shaping sustainable and unsustainable behavior. To quote directly:

Sustainable development is a process of change in which the exploitation of resources, the direction of investment, the orientation of technological development, and institutional change are all in harmony and enhance both current and future potential to meet human needs and aspirations. (WCED 1987, 46)

It may be easy to overlook the dramatic change in environmental discourse exemplified in Brundtland, but it should be noted that in this context, “human needs and aspirations” refer to more than the natural resources needed to meet the demands of consumers. For the first time, the discourse of sustainable development was paying explicit attention to the quality of all human life. The concept of social justice had entered the development conversation in a potentially powerful way.

Brundtland famously defined sustainable development as “development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (WCED 1987, 46). The report set a new standard by which the success of any development program should be measured. To support this broad goal, the Brundtland Commission suggested a system of governance that extended from the global to the local scale. A key feature of the proposal was an ecology-first approach to governance. In other words, policymakers would develop regulations and policy to set the stage for economic activity that was first and foremost ecologically appropriate. Brundtland also called for a system of regulations to be developed to address existing problems of environmental protection and resource management. These were to be defined by a rigorous risk assessment using sophisticated indicators and other methods of assessment. Finally, these actions would be based on the participation of stakeholder groups and marshaled evidence so that deliberations would be inclusive and actions appropriate.

In some ways, the Brundtland Report reshaped the key features of the sustainable development debate. While scarcity remained a topic of
concern, the report largely replaced the concept with *natural capital*. In this formulation, the cost of a commodity such as gold or copper or “rare earths” includes the cost of depleting or destroying ecosystem services. Here the concept of natural capital supplanted scarcity as the ultimate value because the concept of scarcity itself has no relationship to the environment where ore or other resources are hosted. Consider as well how we think about climate. We do not talk about the scarcity of clean or carbon-free air. But we do think about the atmosphere’s carrying capacity of carbon and the effect of this on our lives and livelihoods. With the Brundtland Report, the task of defining the environment and identifying the beings affected by its transformation entered the political sphere as a site of debate.

The shift from resource scarcity to *ecosystem integrity* was underwritten by the assumption that economic systems can innovate to: (i) reduce resource intensity (using less raw material in a product); (ii) substitute one resource for another; and (iii) find new sources (e.g., by recycling), among others. Hence, the Brundtland concept of sustainable development is fundamentally based on growth and technological substitution. Where we once believed that economic growth is limited by the Earth’s physical limits, Brundtland taught us to see that our practice of technological innovation constantly expands those limits.

These points are best illustrated through the example of ecological modernization. Like the pre-1970 scarcity and growth debate, the sustainable development debate emerged from a certain political-economic context. The relationship between that particular historical context and the terms of the debate helps us to understand sustainable development as a set of practices as well as goals embedded in a particular context.

**Ecological modernization and sustainable development**

Cold War détente between the NATO countries and the communist bloc, coupled with efforts to contain the geopolitical expansion of the USSR, helped to define the scarcity and growth paradigm of sustainable development. In the 1980s this struggle, though very real as a suite of foreign policies, opened up a new front in domestic debates around economic development and growth. In particular, after the fall of the Iron Curtain in the late 1980s the role of the state came increasingly into question. What was perceived as centralized control of economies, and therefore development, via nation-states began to give way to state restructuring and an increasing role for private enterprise to regulate itself and the market to be the adjudicator of economic, social, and environmental relations, not the nation-state (see Political economy and regional development). These political and economic changes influenced how sustainable development would be implemented. Specifically, the theory of *ecological modernization* motivated and underpinned sustainable development for the first time as a realized policy discourse and not simply a plan.

In the 1980s, ecological modernization theory (see Ecological modernization) was being discussed by a relatively small group of scholars in the social sciences, primarily in Germany and northern Europe. Ecological modernization was, and remains, the theory that environmental problems should be addressed through the transformation of production via the development and implementation of new technologies and processes that were less resource-intensive than their predecessors. This differs from Brundtland in its focus on environmental concerns. Further, economic actors such as entrepreneurs and the private sector, rather than the state, were to be the drivers of this transformation.
A macro-economic structural change, ecological modernization identified sectors capable of combining higher levels of growth with lower levels of material input. For the adherents of this approach, industrialization and technological substitution and innovation offered a way out of ecological crisis. And, unlike earlier responses to this crisis, ecological modernization was more than a theory. In the 1990s it became a key part of the European Union’s Environment Action Programme (see Environment and development). Key aspects of the approach as codified in policy focus on the precautionary principle, integrated regulation (carrot and stick), and voluntary mechanisms. Indeed, countries like Germany, the Netherlands, and Japan are said to be paradigmatic examples of this approach in practice.

Ecological modernization was also rhetorically similar to sustainable development. Advocates for both concepts shared an interest in maintaining the environmental integrity of the Earth through continued but less resource-intensive growth. While the Brundtland approach to sustainable development called explicitly for intergenerational equity, ecological modernization implicitly sought to address these concerns by advocating for the use of fewer resources now so that more will be available in the future (see Conservation and capitalism). Both visions relied heavily on technological innovation to reduce environmental impacts. And, while sustainable development goals appeared in countless official documents worldwide, it was the vision of ecological modernization, of “greening the economy,” that drove practice. Ecological modernization became the operant of macro-level sustainability.

Ecological modernization as a theory of social change and a set of policy prescriptions emerged from established and newer thinking about how the economy works. The Paley Commission could not have anticipated the shrinking role of the government, which by the 1990s had diminished from the architect of economic and social activity to merely one actor among many in a governance structure. A restructured government sits well in ecological modernization. As a result, more recent efforts in sustainable development look very different from earlier approaches. First, ecological integrity has largely been reduced to a series of technological fixes. Technology, ecological modernizers believe, can mediate, replicate, or augment natural processes. Furthermore, prompts for addressing these issues come not from a state regulation but from either social movements or consumer demand, with the market determining which environmental systems are integrated into the economy. When, for example, was the last time we heard of the importance of biodiversity outside the potential benefits that “Big Pharma” can derive from it (“Big Pharma” being a critical term used to describe global pharmaceuticals firms)? Thus abstract forces define what the environment is and whether and how it is integrated into the larger social system of economic activity.

Today, while many politicians, policymakers, and academics would argue that market forces, which are assumed to be based on the decisions of individual consumers, may seem more democratic and thus “just,” evidence suggests that one cannot achieve justice through markets. Some people have fewer economic or social resources now than before; therefore their interests are not well represented among the forces that drive sustainable development. As a result, certain groups – for example, women in Ghana where water is scarce, elderly people in Chicago who are vulnerable to heatwaves, and the world’s non-whites and economically disadvantaged – bore and continue to bear a greater burden of environmental risks in the world’s cities. Brundtland’s sustainable development seeks the inclusion of social justice concerns within these discussions. It
is not, however, capable of changing the broader set of social relations that help to determine who suffers from inequity and who does not.

As a normative construct, Brundtland’s sustainable development established a framework for measuring policy goals. However, as the brief discussion suggests, how those ideals are implemented has not been a straightforward process and has been heavily dependent on a broad social and political-economic context. Sustainable development remains a shaky three-legged stool. Every once in a while policymakers and thought leaders meet in an attempt to repair the design flaws, but these are only addressed conceptually, and not in the political sphere. To further explore this notion and the role of justice in sustainable development, the discussion now turns to local level sustainable development.

From macro- to local-scale sustainable development

A discussion of sustainable development at a local scale is important for a variety of reasons. First, it gives us an opportunity to further ground the ideas from Brundtland. The Earth Summit in Rio in 1992 extended and grounded Brundtland into an actionable agenda both globally and locally. Here, at the largest gathering of high-level officials ever, policymakers deliberated over the goals of Agenda 21 in general, and Local Agenda 21 in particular. This discussion focuses on Local Agenda 21.

Rio, Local Agenda 21, and sustainable development

The Brundtland Report was produced in 1987. Five years later, in 1992, at the Earth Summit in Rio de Janeiro, delegates from around the world discussed Agenda 21, particularly the concept of subsidiarity. They decided that subsidiarity was the principle that decisions should be made at the most local level possible, where decision-makers are closest to the people affected by their decisions. To adopt this principle of subsidiarity, Agenda 21 architects proposed Local Agenda 21 (LA 21). Chapter 28 codified LA 21, and recognized that local governments play a key role in bringing about sustainable development across scales. LA 21 principles emphasized that local authorities needed to make considerable changes to their policy-making approaches so they could incorporate the perspectives and views of a range of interests. Chapter 28 of Agenda 21 states:

Each local authority should enter into a dialogue with its citizens, local organizations and private enterprises and adopt a “LA 21.” Through consultation and consensus building, local authorities would learn from citizens and from local, civic, community, business and industrial organizations and acquire the information needed for formulating the best strategies. The process of consultation would increase household awareness of sustainable development issues. Local authority programs, policies, laws and regulations to achieve LA 21 objectives would be assessed and modified, based on local programs adopted. (UNCED 1992)

Key to LA 21 was a broad-based, multistakeholder planning process that balanced the economy, social equity, and the environment. Proponents believed the success of LA 21 would also turn on the ability of a local authority to redirect its policies, laws, and regulations to align with the principles that emerged from the planning efforts. By 1996, according to LA 21 architects, cities should have engaged the public in a consultation process and achieved consensus around a local sustainability program.

The International Council of Local Environmental Initiatives (ICLEI), the international environmental nonprofit organization, spearheaded the effort to assist communities in
developing their LA 21 activities. Indeed, ICLEI had hatched the LA 21 idea a year before the 1992 Rio Summit. The goal of ICLEI’s LA 21 Campaign is to “to build a worldwide movement of local governments and associations dedicated to achieving sustainable development through participatory, multi-stakeholder sustainable development planning and the implementation of resulting LA21 action plans” (ICLEI 2002).

The process was quite simple. By 1996, 1119 communities in Europe had initiated LA 21 programs. To accomplish this goal, in 1994 ICLEI organized the European Sustainable Cities and Towns Campaign, which led to the Aalborg Charter. Two years later, ICLEI implemented steps for the Aalborg Charter, which was ratified under the Lisbon Action Plan. Since Lisbon, ICLEI has worked with hundreds of communities on their LA 21 plans and organized thousands of communities.

A survey published in 2002 by ICLEI showed that some 6400 communities from 113 countries worldwide had become involved in LA 21 activities over the previous decade (ICLEI 2002). The popularity of LA 21 as a local and regional development strategy was again affirmed when 400 European communities, representing 100 million Europeans from 36 countries, signed the Aalborg Charter in 1994. The Aalborg Charter, which was facilitated by ICLEI as part of their European Sustainable Cities and Towns Campaign, committed its signatories to develop local sustainability action plans in parallel with LA 21.

Another reason for the success of LA 21 in Europe is that its agenda was supported by the functions of the state. Despite calls for fundamental change, the adoption of LA 21 was not a total paradigm shift in Europe. It emerged in the context of existing state functions and regulatory mechanisms. Typically, these occur along two axes, market-oriented “regulation” and state-based command and control. In Europe, state-based regulation, while waning, remained a dominant force that influenced development patterns and land use. This is to say that local, regional, and national authorities have the capacity to clearly direct outcomes of urban development through performance standards and other forms of command and control regulation. The Dutch, for example, set minimum density standards at 33 units per hectare in an effort to promote compact cities with smaller ecological footprints. In the Netherlands, as in other places like Germany and Denmark, the rural fringe of cities is not regarded as transitional and up for grabs for development. Thus, market forces in these countries could be closely regulated so that they align more appropriately with the dictates of the LA 21 process, in particular with considerations of social equity.
One important consequence of neglecting a just and inclusive process is the predictable absence of equitable outcomes when many kinds of stakeholders are excluded from the process. In contrast to European cities and towns, North American cities did not adopt the LA 21 process in large numbers. Roughly 20 US cities are LA 21 cities. This is not to say that US cities did not adopt sustainable development principles. As with other sustainable development policies, US governments at the local, regional, and national scales eschewed the global standard and came up with their own approach. The US approach relied on mechanisms that were incentive-based, voluntary, and largely developer-driven (by the market, not the state), rather than process-driven. This American variant of local sustainability is often referred to as “smart growth” (Krueger 2007). With its roots in the state of Maryland, it spread rapidly across the United States. Over 500 US cities and towns have adopted principles of smart growth. Indeed, smart growth now extends beyond US borders as policy entrepreneurs have exported it to Europe and Asia. To understand not only the rapid spread of smart growth policies and practices but also the absence of concerns about justice, the political-economic context must be considered once again.

Economic engines: urban sustainability, LA 21, and “greening the city”

Many urban theorists have agreed that cities are the main site for economic development. Since the mid-1990s, in the context of an unprecedented development boom in many cities around the globe, the process of urbanization has been a focal point of economic development. The sustainable development discourse has shaped the way many cities have been remade. It is hard to find a city that does not mention sustainability as part of its identity. In the 2000s the mode of urban development – remaking long derelict urban cores into vibrant urban utopias and sites of consumption – was synonymous with urban sustainable development. Development plans go by the monikers of “compact urban development,” “transit-oriented development,” “smart growth,” and the like. They rely on existing infrastructures, they remediate and regenerate brownfields, they often include public transit lines, and they reinvest in communities in need. Coincidentally, these landscapes appealed to the imagery of the urban utopia for the creative class – artists, inventors, entrepreneurs, and so on. It has been argued that the creative class was an important ingredient in urban development because these people are the new raw materials for the entrepreneurial, new economy city (Krueger 2007). Advocates for the sustainable city claimed it as not only a worthy goal but also a driver in shaping a city for the new economy; they proposed that it would stimulate the kind of work needed to support the supposedly less resource-intensive information and service economies, which were the engine of the postindustrial city, or the eco-modern city. Those adopting this development approach also believed that such practices made them more competitive in the global labor market. The Local Agenda 21 process, grounded in environmental and social concerns, had been supplanted by a developer-driven “sustainable city.” Like ecological modernization on the macroscale, urban areas “greened” but did so for purely economic reasons. Urban sustainable development had arrived, but not as it had been imagined by its ethically motivated architects in the early 1990s.

No matter which moniker one chooses – “urban sustainable development,” “LA 21,” “smart growth,” “compact urban development” – sustainable development has been
adopted as a set of policy goals at the urban scale. We now have more mass transit nodes, more “green” buildings, more viable ecosystems in our cities than ever before. These are all good things. The value of being “green” is indisputable in many cities. Freiburg, Germany, proudly calls itself the greenest city in Europe, as does Växjö, Sweden. Despite the accolades and media attention, these cities have not realized what it means to be a “sustainable city.” The actions of these cities have brought economy and environment closer together in ways that are in line with the norms established by Brundtland. Many commentators have noted that for many cities the single most important element in a sustainability effort revolves around environmental concerns (Warner 2002). These environmental aspects coupled nicely with economic development visions, making them palatable to city policymakers and boosters. While environment-led development remains an elusive goal, in the last round of urban regeneration these two goals articulated by Brundtland and operationalized at the local level have more closely aligned visions that, a generation earlier, were seen to be competing.

Reasserting justice and equity in sustainable development?

While economic and environmental concerns have begun to converge in recent years, concerns for social justice have all but disappeared from conversations about sustainable development. Sustainable development, particularly in urban settings, has privileged environmental goals over social ones. An example of a macro-level sustainable development policy is reducing emissions from deforestation and forest degradation (REDD), which is designed to encourage local stewardship. At the urban scale environmental injustice is flourishing in the wake of gentrification (see Gentrification), even when adjacent to sustainable development projects. Environmental justice and sustainable development advocate and scholar Julian Agyeman (2013, 5) notes that “a truly sustainable society is one where wider questions of social needs and welfare, and economic opportunity, are integrally connected to environmental concerns.” With colleagues he coined the phrase “just sustainability” as a label for those activities that consider all three components of sustainability: environment, economics, and justice. Because social equity and justice are not integral parts of the sustainable development agenda, Agyeman and others explicitly remind scholars and practitioners that what often looks like “proper” urban sustainable development is only a creative use of the environmental and economic priorities without any attention to social justice. Environmental justice commentators provide a trenchant critique of sustainable development, especially in urban contexts in Western Europe and North America. By bringing together the sustainability literature with the environmental justice literature, these authors seek to create a vocabulary for political opportunity and mobilization both at the grass roots and in local government. To date, this work has focused largely on the organizational principles of movement leaders rather than the extensive politics of economic development.

There remains a critical need to examine how sustainable development could be mapped onto the current geography of neoliberal capitalism and, more importantly, where the opportunities for broader political engagement with sustainability may present themselves. By recalling that some groups of people are disempowered, marginalized, or silenced in the policy process, and by rendering the concept of justice explicit, we can place justice and equity in an equal position alongside environment and economy. (More on this later.)
Urban greening has been central to the urban sustainable development discourse. Green space has increased and/or improved in many cities in the name of sustainable development. Green space has been preserved from development because of its aesthetic, recreational, and ecological value. In addition to these green infrastructures, bike lanes, bike share programs, shared car schemes, additional miles of trams and public transport, and smarter public transport have been developed. Planners have developed “green infrastructure toolkits” to help mobilize and disseminate how cities might turn unmanaged or derelict space into sources of economic value and cost savings through relying on ecological services. Finally, inner-city redevelopment and brownfield rehabilitation have all contributed to this convivial relationship between green spaces and the city.

Adopting green practices in cities makes sense in terms of transaction costs and other development costs as they contribute to the city’s economic competitiveness. Local and regional urban spaces market their greening strategies to attract new workers and desirable high-tech jobs. Many policymakers believe that, in return, these jobs have broader advantages throughout the local economy. Freiburg, Germany, for example, has developed policies to support its claim to be Europe’s greenest city. The city is known for its extensive network of local and regional bicycle lanes, tram service, high quality of life, and history of progressive politics. Using these as points of evidence, city boosters thus claim that Freiburg represents an ideal sustainable city (for contrary views see Freytag, Gössling, and Mössner 2014).

However, deep injustices accompany urban sustainable development. In the process of gentrification, regenerated and newly “sustainable” property tends to rise in value. In most cases these redeveloped properties were formerly rental units or properties in low-income neighborhoods. Consequently, many residents get priced out of the housing market in their own neighborhoods. This process of gentrification transpires both when the market is left to regulate, as often happens in the United States, and through state-led gentrification, such as in London’s East End. Forced out of marginal housing, these people must move to other marginal housing further from their jobs, causing additional hardships like longer commutes, higher fuel costs, time away from family, and lower-quality housing.

In addition to economic forces, political practices also contribute to the exclusion and victimization of disempowered groups. Picking and choosing which aspects of sustainable development they will adopt to serve their needs, urban elites contribute to the process of victimization begun by impersonal market forces. When business associations, civic leaders, and others come together in the name of urban sustainable development, they introduce green buildings or the like in ways that benefit themselves, not the larger community. Finally, and perhaps most perniciously, the very idea of poor people defending their rights within these debates has been discredited by the notion that urban sustainable development discourses are postpolitical – supposedly no longer about power relations – because justice is now an actual rather than merely a rhetorical component of sustainable development.

Sustainable development … really?

While sustainable development is discussed in city halls around the world, from Bangladesh to Boston and from Casablanca to Carlisle, the concept mutates to fit disparate political, economic, temporal, and geographic contexts. In its most recent and prominent form, in a world that is increasingly wedded to notions
SUSTAINABLE DEVELOPMENT

of market-led development, it is a project of ecological modernization. Whether in local form or in more macro-policy approaches (e.g., Millennium Development Goals), sustainable development has integrated with capitalist social relations while offering the promise of a better world, not business as usual. And, in some ways, it has delivered on this promise: today we are much less concerned about scarcity per se than about finding the next technological alternative (e.g., natural gas, oil shale, nuclear power). At the same time, this cultural shift has reinforced past ideologies, such as technological determinism and continuous growth through innovation (Schulz and Bailey 2014).

Concepts and frameworks, including sustainable development, are useful in that they provide a new set of norms for society to appeal to. However, new frameworks alone don’t always advance us toward the better world we hope to create. To make the best use of new norms, we must recognize how they are introduced and adopted, not only in their own terms, but in the context of the prevailing ideology and conceptual framings of how the world works – understanding, for instance, what happens to considerations of social justice when economic development becomes the driving force of sustainable development. This entry has tried to reveal these relationships and consequences while offering a history of the development and evolution of sustainable development discourse. The final section looks toward a new approach that might lead to a “post”-sustainable society.

Toward a post-sustainable development society?

As the entry has shown, natural and social scientists, and particularly geographers, have, for 40 years, examined sustainable development from the perspective of ecological modernization. For example, economic geographers continue to focus on technological innovation and substitution by examining businesses like green manufacturing and environmental services, as well as supply chains and other economic networks. Scholars thus rely on the traditional development paradigms, which is to say they rely on traditional growth paradigms (or business as usual). For Schulz and Bailey (2014), this focus also dominates most of the recent scholarly work on the economic impacts of climate policy and energy transitions. While focusing on the (spatial) dynamics of emerging industries (e.g., related to renewable energies), underlying innovation processes, regional economic trajectories, or international trade, scholars overlook the implications of these development trends for alternative growth models and sufficiency imperatives. The final section will briefly introduce one of these alternative development scenarios.

Post-growth?

Can our seemingly open-ended attraction to economic growth be sustained, even as more and more people are recognizing that this model creates more problems than it addresses? The push for continuous growth undermines economic security, communities, the environment, a sense of place and continuity, and even mental health. Increasingly, policymakers at a variety of spatial scales are mobilizing new indexes, such as the Happy Planet Index, the Human Development Index, and the Index of Sustainable Economic Welfare.

Alongside these new ways of measuring sustainability and prosperity is an alternative approach to development that eschews growth in its current form. This concept, “post-growth,” should be understood as a departure from dominant growth paradigms in the sense of Serge
Latouche’s *décroissance*: that is, as a rejection of the maxim that private and societal prosperity can be ensured only via a continuous growth of materially and monetarily measurable economic performance. Tim Jackson’s (2009) book *Prosperity without Growth* captures the orientation of a transition toward sustainable lifestyles and economic systems envisioned by the post-growth concept. Jackson and Latouche place strong emphasis on distributive justice in growth and wealth, both at the level of international and development policies and within individual national economies.

Drawing on the work of Georgescu-Roegen, especially his steady-state economics theory, Latouche has refocused the discussion of environmental sustainability, suggesting that the objective of economic growth has endangered the goal of environmental sustainability and banished the principle of justice. His principle of *décroissance*, which rejects growth-oriented forms of production and consumption patterns, has been influential in the Italian, Spanish, and French sustainability movements. In this way, Latouche warns us against the trap of ecological modernization, rejecting its belief that a transition to a sustainable economy can be possible within present market principles. Latouche and the post-growth approach point to a possible way forward, exposing new analytical terrain and the limits of previous work. Bearing in mind this history of sustainable development, its promises and pitfalls, we can begin again to look for new alternatives to what sometimes seems an intractable global challenge.

**Summary**

This entry has examined sustainable development as a historically embedded process that is part of larger social, economic, and institutional contexts, each of which has implications for how sustainable development is defined. Sustainable development has been explored as a historical artifact that takes on different meanings in different places and times (e.g., from Barnett and Morse to Brundtland). The powerful hold some concepts, such as technological determinism, have exhibited as different sustainable development perspectives have evolved has also been explored. Sustainable development has been examined in different contexts, such as ecological modernization. The entry also described how different legs of sustainable development’s three-legged stool are privileged based on their relationship to economic thinking and policy, and concluded with a brief discussion of what are referred to as post-sustainability perspectives. Overall, the entry showed that sustainable development is an elastic concept whose meanings and practical implications reflect varied views on the global environmental problems (and opportunities) facing humanity.

**SEE ALSO:** Conservation and capitalism; Ecological modernization; Environment and resources, political economy of; Political economy and regional development

**References**


SUSTAINABLE DEVELOPMENT


Further reading

Sustainable transport

Bradley W. Lane
Joseph Beeler
University of Kansas, USA

The concept of sustainable transport

Sustainability has become one of the most important interdisciplinary issues in society. The concept of sustainability has developed with our awareness of the finite nature of many resources and of the negative externalities of our activities. The concept has permutated almost every aspect of society, and the notion of sustainability has become a key component of business and government actions in a wide variety of sectors. The interconnected nature of sustainability likewise makes it a difficult and controversial set of concerns to address.

It is unsurprising, then, that one of the most prominent recent developments in transport has been the concept of sustainable transport. From long-term efforts to regulate emissions to more recent attempts to develop and integrate alternatives to petroleum for fuel, sustainability has become an implicit component throughout transport. Despite this, defining sustainable transport has proved to be a challenge. One of the first definitions comes from Black (1996, 151) who, drawing from the Brundtland Report of 1987, describes it as “transport that satisfies the current transport and mobility needs without compromising the ability of future generations to meet those needs.” Many questions persist as to what constitutes those needs. Most transport modes and activities are sustainable; unsustainability develops when demand for those modes becomes excessive (Black 2010).

Sustainability is often referred to as having three pillars: energy, environment, and equity. These pillars are likewise prominent in discussing sustainability in transport. Most discussions of sustainability in transport begin with describing what makes transport unsustainable. There are three primary realms of issues that contribute to the unsustainability of transport; these are energy issues, airborne issues, and land-based issues, which are briefly summarized in Table 1.

Sustainability issues have a long history in transport. The use of timber for navies and seafaring exploration was linked with deforestation in the United Kingdom and North America, and the use of horses for urban travel was linked with congestion and the disposal of manure. Industrialization placed additional pressures on timber supplies for fuel and as a construction material, and the growing usage of coal exacerbated air pollution. The use of chemicals as materials, such as lubricants, led to new forms of chemical pollution. Rapid urbanization in industrial cities led to congestion and urban air pollution, which were partially mitigated by the development of the first urban transit systems. However, the rapid diffusion of the automobile in the following decades contributed to urban air pollution, climate change, congestion, safety, and energy issues. In the mid-twentieth century governments began to recognize the negative externalities of motor vehicle use and to introduce policies to mitigate them. Marine transport is relatively sustainable, but there are issues with congestion at or near
Unsustainability in transport: airborne issues

Global climate change

To understand the interaction of transport and climate change requires a brief overview of some relevant atmospheric chemistry. There exists a family of naturally occurring chemical components that, when released, become suspended in the atmosphere and create a naturally occurring phenomenon called the greenhouse effect. The greenhouse effect is necessary to keep the planet warm enough to sustain life. However, a problem arises when greenhouse gas emissions exceed the ability of the Earth to absorb these emissions. The resulting increase in the atmospheric content of the gases traps more heat near the Earth’s surface, causing the overall average temperature of the planet to increase. The impacts on the climate and ecology of the Earth from this warming make up the phenomenon called global climate change.

Human activities in transport (or elsewhere) were not enough to influence the greenhouse effect until after the Industrial Revolution, when the burning of fossil fuels for energy began to increase at an unprecedented scale. As transport mechanized, coal and petroleum were developed into its primary energy sources. The incomplete combustion of these fossil fuels results in the emission of greenhouse gases, and over time the worldwide growth of transport has resulted in emissions levels that exceed the Earth’s ability to absorb them, and resulted in an increase in the atmospheric content of these gases. The primary
such pollutant from transport is carbon dioxide, though nitrogen oxides and methane are also contributors. In the United States, transport accounts for 27% of greenhouse gas emissions (EPA 2016).

Complicating the effect is the residence time of gases. Residence time is difficult to calculate because of the varying rate at which gases decay, and the effect of emissions sources and absorption rates on atmospheric content. Carbon dioxide can last in the atmosphere for between 5 and 200 years, nitrogen oxides for over 100 years, and methane for over a decade (IPCC 2014). Because of this, the full effects of efforts to deal with global climate change will not be felt for generations after their implementation.

One impact of increasing global average temperature is the increasing frequency and severity of major weather events. Damage to transport infrastructure will likewise become more frequent and more severe. Facilities that are at low elevations or near bodies of water will also see damage and require repair or relocation. These include much of the world’s port facilities and railway infrastructure. The flow of goods, as well as the travel of individuals, could experience significant disruption as sea levels rise and severe weather events increase in frequency. The movement of agricultural regions in response to changing climate should impact infrastructure usage, requiring new roadway, railway, and terminal infrastructure, as well as the repair of existing infrastructure from additional use.

**Urban air quality**

Emissions from petroleum usage contribute to the degradation of urban air quality. These include carbon monoxide, sulfur dioxide, nitrogen oxides and dioxides, ozone, and solid particulate matter, all of which are released from burning gasoline in internal combustion engines. These “criteria pollutants” are the focus of urban air policy because of their negative impacts on human health when released in amounts that are toxic or lethal, such as impairing the function of the cardiovascular system and damaging the respiratory system. Vehicle emissions are greatest when operating in stop-and-go traffic, when the combustion of gasoline in engines is least efficient. Policies place the burden of compliance on automakers through tailpipe emissions standards, and on urban areas through ambient air quality standards, both of which must be met to receive the necessary federal government funding for roadways.

A major criteria pollutant in the past was lead. Lead was introduced as an additive to gasoline in order to reduce engine knock in the 1930s. Lead exposure has been found to interfere with cognitive development and to induce blood disorders in children. Despite the negative health effects of lead emissions, and awareness of these effects decades earlier, lead was not removed from gasoline completely until the 1970s and 1980s in the developed world. Even then, the primary motivation for the removal of lead from gasoline was its effect on catalytic converters, which it tended to clog.

Urban air pollution has seen significant improvement in the latter half of the twentieth century in the United States and the developed world. In the United States, urban motor vehicle congestion reached critical stages; some of the most notable effects were seen in Los Angeles, California. This led to the state of California developing more stringent air quality standards than had previously existed, standards that have continued to evolve. US states may choose to adopt either the federal Environmental Protection Agency (EPA) standards or the California standards, but they cannot produce their own policy. Currently, nine states have adopted the California policy.
SUSTAINABLE TRANSPORT

Urban air quality is a critical unresolved issue in the developing world, which continues to urbanize rapidly. Automobile ownership rates and usage are increasing precipitously, and emissions standards are far less stringent than those in the developed world. The cities of China are particularly glaring visual examples of the extent of this problem.

Unsustainability in transport: land-based issues

Motor vehicle safety

One of the most implicitly ignored components of sustainability in transport is safety, particularly in the form of fatalities from motor vehicle accidents, which account for 94% of all the fatalities in transport. Automotive accidents claimed 32,719 lives in the United States in 2013 (IIHS 2015), and are among the leading causes of death for children and for adults under the age of 35. Data indicate that there were approximately 1.24 million deaths in motor vehicle crashes worldwide in 2010 (WHO 2015). Despite this, there is relatively muted public and policy attention toward traffic fatalities, except when there are high-profile incidences of malfeasance by government or industry.

Traffic fatalities face two disparate trends. In the developed world, traffic fatalities have been dropping steadily for nearly 50 years, even as vehicle miles traveled have increased. However, worldwide, traffic fatalities are increasing due to increasing vehicle miles traveled in the developing world coupled with more lax safety standards, lower driving education, and poorer road quality.

Traffic fatalities are caused by human error, environmental factors, or design flaws in the vehicle or the roadway. Many countries in the developed world have made significant strides in reducing motor vehicle fatality rates through enhancing safety standards, and automakers have made tremendous advancements in safety through vehicle design and engineering. To address fatalities, a distinction exists in policies that try to reduce fatalities from a crash, and policies that try to reduce the frequency of a crash. Most policies in the United States, for example, have been focused on safety measures that increase the survivability of an accident, while Sweden’s Vision Zero policy plan targets the eventual elimination of all fatalities through both vehicle safety standards and improvements to the environmental and design characteristics of the road network.

Congestion

Congestion is one of the more controversial and challenging aspects of unsustainable transport. This consists largely of motor vehicle congestion, though congestion from other sources such as air travel and surface freight movement raise similar issues. A generally accepted definition of motor vehicle congestion comes from the European Conference of Ministers of Transport (ECMT), which defines congestion as “the impedance vehicles impose on each other, due to speed–flow relationships, in conditions where the use of the transport system approaches its capacity” (Black 2010, 64).

Congestion causes include traffic accidents or roadway incidents, weather, special events, and economic activity. Many tools have been deployed to mitigate or reduce congestion. These have included actions directed toward the road network, such as expansion of the roadway system, improving operational treatments such as traffic signaling and motor vehicle incident clearance, and introducing high-occupancy vehicle lanes, congestion pricing, and toll roads. There is
also an array of strategies to reduce motor vehicle use through behavioral change. These include land-use planning designed to increase the use of nonmotorized modes and transit modes, expansion of the transit system, teleworking, carpooling initiatives, and increasing flexibility in work schedules to alleviate peak period congestion. The evidence indicates that, despite massive investments in roadway infrastructure, we cannot build our way out of congestion, and these can even induce additional travel and congestion. Thus far, most of the efforts to mitigate congestion through behavioral change have had limited effect.

Of particular recent interest are intelligent transportation systems (ITS). Various forms of vehicle automation, including self-driving cars, have the potential to reduce or eliminate human error as a cause of traffic accidents and to reduce the distance between vehicles while still traveling safely, thus increasing road capacity. It may be that we are just scratching the surface of the potential of applications of this type of technology in transport, and its effects on sustainability, while potentially large, are unknown. There is also a line of thinking in transport that congestion is an inevitable by-product of economic growth and development, and that a do-nothing approach will place sufficient pressure on congestion (through the loss of economic activity from congestion) that the problem will eventually solve itself. Some also argue that congestion is inevitable and that losing congestion is undesirable because it represents the loss of economic activity.

**Land use in sustainable transportation**

Land use and transport are interdependent. Transport policy has largely focused on enhancing mobility through facilitating motor vehicle use and the expansion of roadway infrastructure. This is particularly strong in the United States and the developing world, though it is also found in the rest of the developed world. This policy has contributed greatly to the other causes of unsustainability in transport discussed here, particularly in relation to petroleum use, congestion, and fatalities. This has also contributed to inequity issues by, effectively, requiring automobile ownership for participation in society. The land uses that serve automobile-oriented development are described as “urban sprawl,” whereby low-density, single-use development is built with ample land for parking, motor vehicles are the only suitable mode for travel, and improving vehicle flow is done through expansion and management of the roadway system.

Efforts to improve sustainability through land use have gained significant traction since the late twentieth century. In the United States, efforts called “New Urbanism” and “Smart Growth” represent attempts to encourage land uses and transport modes that discourage motor vehicle usage, and replace it with other, more sustainable forms of travel. These development forms are more conducive to walking and biking because they mix land uses and build safe infrastructure for nonmotorized modes of travel. Many of these developments are built in connection with and around new transit stations, and this “transit-oriented development” (TOD) makes public transit trips (often by light rail systems that stop at these stations) more conducive as well.

The evidence supporting the effect of increasing sustainability through modifying land use is, at best, mixed. These developments appear to capture people who are choosing to live in a denser urban form, as well as bringing about modal shifts away from motor vehicle usage. However, these developments are too few in number and too small in size to influence the whole of the city in which they are located, and most of the evidence indicates that they
have done little to mitigate citywide congestion. Many of these types of developments (and the light rail systems built to support them) have been undertaken as much for economic development purposes as for anything related to enhancing transport sustainability. Land-use strategies remain among the most complex and controversial approaches to dealing with sustainable transport.

Unsustainability in transport: energy

Petroleum dependency

Perhaps the primary issue with sustainability in transport is reliance on petroleum as its primary energy source. Petroleum comes from refined crude oil, which is unsustainable because of its being finite in supply, nonrenewable, and polluting. Petroleum deposits exist in a “petroleum window” of underground caverns of numerous shapes and depths. Most of the petroleum that has been extracted is the “sweet crude” deposits that are relatively easy to reach. The remaining supplies that are relatively easy to extract are located in places that are geopolitically unstable. What remains exists in difficult-to-access subterranean pockets that are too expensive to drill unless oil prices are sufficiently high to make extraction profitable. However, North America and parts of Central Asia experienced a boom in supplies in the early twenty-first century due to advances in fracking technology, where water is flushed into these deposits to help draw out the petroleum.

Estimates vary considerably of when worldwide petroleum supplies will be exhausted. A key concept in oil production is “peaking,” where yearly production totals begin to decline after reaching a maximum (or peak) because of dwindling global supplies. Oil production in the United States is believed to have peaked sometime in the 1970s, and global production is expected to peak sometime in the second or third decade of the twenty-first century, though the discovery of new fields and current advances in fracking technology make identifying an exact date difficult.

Most of the world’s transport supply runs on various forms of gasoline, a petroleum derivative. Diesel gasoline, natural gas, and ethanol–gasoline blends are also used. Petroleum is also a heavily used lubricant in the internal combustion engine (ICE), the most common type of engine in transport. Petroleum became the key energy source for transport because of several advantages it has over other alternatives. First, it has the highest energy density of any fuel in transport, meaning it takes the least amount to go the farthest compared to other energy sources. Storage is also relatively easy, and other energy sources, such as electricity and natural gas, have much greater competing uses (such as home and commercial heating and cooling) than petroleum (Greene 2004).

Further driving unsustainability is increasing the worldwide demand for petroleum. Globalization contributes to this by increasing the volume of shipments of goods and services over longer distances to a broader array of locations, and a rise in the overall standard of living in developing countries contributes to increasing the amount of transportation individuals do as well. Globalization also appears to contribute to transport unsustainability through inequity, as much of this economic growth is spatially and socioeconomically segregated.

A final issue influencing transport sustainability with petroleum is its price. The costs of crude oil, gasoline, and natural gas are some of the most influential components of the worldwide economy. Their costs, supply, and demand can be subject to significant fluctuations that in turn influence activity in sustainable transport. Rising costs and tightening demand increase interest in
developing alternatives and efficiency, changing behavior, and extracting remaining crude supplies that were previously too expensive to be profitable. When supply increases and costs drop, interest in these changes subsequently wanes.

Solutions and alternatives in energy issues

Some of this unsustainability in transport is being mitigated by increasing efficiency. This involves increasing the efficiency of the use of the remaining supply of petroleum, and developing alternatives for replacing petroleum – efforts which began decades ago and are still in progress. The regulation of petroleum in transport for sustainability-related purposes dates back to the responses to acute urban air pollution problems, and the oil shortages in the 1960s and 1970s instigated by the Organization of the Petroleum Exporting Countries (OPEC). In this era, regulations were introduced in the United States requiring the average fuel economy of all the cars an automaker sold to meet a certain threshold. Similar regulations were introduced elsewhere in the developed world, though the focus of reducing fuel use through these policies ranges across increasing vehicle efficiency, regulating emissions, and diversifying the fuels an automaker can offer. Lastly, governments are gradually acknowledging the connection between human carbon dioxide emissions and global climate change, and are beginning to use this as motivation for energy policy.

Nonelectric alternative fuels in transport

Natural gas has a relatively long history as an alternative fuel in transport. It is used extensively in commercial freight transport, and in some household vehicles in rural areas. Natural gas offers several advantages that partially explain its longer history of use. It has a higher energy density than other alternatives, one more comparable to gasoline. Refueling and storage procedures are similar to those of gasoline, where an onboard tank can be refilled in a relatively short time. However, natural gas is not a sustainable replacement for gasoline. It is another nonrenewable, carbon-based fuel, and features the same negative externalities as burning gasoline (described earlier). Additionally, it suffers from several existing competing uses, in particular for indoor heating.

Hydrogen and other forms of fuel cells are also on the horizon, and may hold the greatest promise of meeting the energy density requirements of transport that other alternatives currently cannot meet. However, major questions remain about the cost competitiveness and technical viability of hydrogen fuel cells to power individual automobiles.

Key current advancement in sustainable transport: electric alternatives to gasoline

Significant developments have occurred worldwide in electrified forms of surface transportation in the early twenty-first century, making it one of the most notable and widespread activities in sustainable transport to date. Electrification is a rather old concept in transportation. The first and most popular cars of the late nineteenth and early twentieth centuries were electric, and they outperformed other fuels. However, they fell out of favor in the early twentieth century, and experienced numerous false restarts in the 1970s, 1980s, and 1990s. The current rebirth of electric vehicles (EVs) probably began with the introduction of the Prius by Toyota in the late 1990s, which was the first commercially available car to feature a hybrid gasoline–electric powertrain. This powertrain consists of two motors, a gasoline-powered engine and an electric motor powered from a battery pack. The two motors
work in tandem: the battery drives the vehicle when it is sufficiently charged, and, when it is depleted or insufficient to power the car, the vehicle then adds power from or switches over completely to the gasoline engine. The battery is recharged through use of the gasoline engine and through a regenerative system that captures energy from friction applied in braking the wheels of the vehicle. Hybrid engines have increased in application, and major automakers now offer hybrid engines in a variety of models, though large gains in fuel economy have thus far been limited to smaller cars, similar in size to the Toyota Prius.

Vehicle powertrains that rely more exclusively on electricity have recently seen notable advancement. In 2011 two of the first widely commercially available cars were released, highlighting the different approaches to electrification in surface transport. The Chevy Volt is a plug-in hybrid electric (PHEV) compact sedan that runs off of an electric battery which, once depleted, transitions to power from a backup gasoline motor. The motors work independently of each other and not in tandem. The Volt uses the electric motor exclusively until the battery is depleted, at which point it switches over to the gasoline motor for power. At roughly the same time as the Volt was released, Nissan introduced a plug-in electric (PEV) subcompact called the Leaf. It runs off an electric motor exclusively, powered by a battery pack that recharges mostly through being plugged into an external outlet (though it does also recharge to an extent during use through a regenerative braking system).

The sustainability benefits of electric vehicles come from reducing petroleum demand and consumption, and thus reducing the negative externalities from petroleum reliance already discussed. Additionally, maintenance costs on electric motors are considerably less than on ICE motors, as they do not have nearly as many moving parts and don’t require services like oil changes or transmission repair. The user cost of recharging via electricity is currently significantly less per mile driven than refueling with gasoline, and public charging stations also exist on a limited but growing basis.

There are several critical issues limiting the proliferation of electric vehicles. First and foremost is the purchase price of electric vehicles which, even with government incentives to automakers and rebates to purchasers, still remains considerably higher than ICE counterparts. Additionally, electric vehicles face a perception problem that they are small and underperforming, and fit only a niche market of politically and environmentally inclined consumers.

Primary concerns with vehicle use include the distance the vehicle can travel without being recharged. Many current PEV models can only travel less than 100 miles on a single charge, much less than the distance an ICE vehicle can travel on a single tank of gas. Some models, such as those produced by innovative automaker Tesla (which produces only electric vehicles), have ranges up to 300 miles on a single charge, but currently these models are very expensive (retailing from $75,000 to $100,000). Many automakers and battery manufacturers are working to advance technology to extend the battery range and lower the cost of electric vehicles.

Another use issue with EV proliferation is the recharging process itself. EVs currently recharge through one of three different types of power connectors. They can charge through a conventional, 110V outlet (such as is common in US households), but fully recharging the battery can take up to 12 hours. They can also charge through less common but more powerful 240V outlets (such as is common in Europe and many other countries), and this can reduce the full recharge time to 3–6 hours. Lastly, there is “fast charging,” which can replenish a depleted battery in 30 minutes to an hour. However, fast
charging is currently fairly limited, and questions remain as to the long-term effect on battery life.

Electrification of surface transport may be the biggest current activity addressing the unsustainability of transport. Governments throughout the developed world as well as in China have introduced a vast array of policies designed to increase the adoption of electric vehicles. These include setting targets for the proliferation of EVs in their domestic fleet, requiring automakers to produce a certain number of EVs, incorporating EVs to meet fleet fuel economy or emissions requirements, incentivizing consumers to purchase EVs through tax rebates or purchase price reductions of vehicle cost, incentivizing EV purchase through subsidizing user costs such as recharging, parking, high-occupancy vehicle lane access, and registration fees, and, lastly, investing in research and technological development in the automobile, battery, and supply chain sectors to encourage industrial development and economic growth in these sectors.

Despite this activity, serious questions remain about the sustainability impacts of the source of electricity and the effect of increasing electric demand. Most of the electricity in the world comes from the burning of fossil fuels, usually coal. Though in abundant supply, coal is a finite resource that features the same negative externalities as burning other fossil fuels. Additionally, burning coal releases sulfur dioxide, which contributes to acid rain. However, increasing demand for electricity may also continue to drive current existing investments in development of renewable sources of electricity, such as solar and wind power.

Many argue that current energy grids and plant capacities cannot handle increasing demand, particularly at peak daytime periods of usage at the beginning and end of the traditional workday. Initial evidence indicates (and, interestingly, many electric utility providers believe) that much of this increase in demand will be realized not during peak periods but during off-peak periods, such as through charging the car while it is parked overnight on a weekday or at the weekend. There is a significant supply of electricity, particularly through coal-fired plants, that is unused during off-peak periods, and increasing demand for electricity during these times would represent a profit increase without additional infrastructure investment. A way to further incentivize this would be to implement time-of-day charging, where the cost of electricity goes up or down based on the demand for electricity during that time of the day.

There are broader questions about the overall ecological impact, or “carbon footprint,” of the electrification of surface transport. Some evidence indicates that EV proliferation contributes to increasing pollution from generating electricity. EV motors and batteries use certain rare earths and metals such as nickel and lithium, and it is argued that deployment of the resources and equipment necessary to extract these materials results in a larger net impact on the environment than maintaining the status quo. However, calculating a carbon footprint is a complicated and somewhat controversial process, and, to a certain extent, a carbon footprint critique ignores the impacts on resource extraction that were required to become dependent on petroleum in the first place, as well as what would be a necessary transition away from petroleum dependency in transport.

Most of the preceding discussion involves surface transport in motor vehicles. Energy supply as a sustainability question in transport also applies to railroad, maritime, and air transport. Railroad transport is powered primarily by diesel, coal, or externally supplied electricity delivered via overhanging wires. Civilian maritime transport is fueled from on-board diesel, and air transport is fueled from high-octane forms of gasoline.
There is relatively little discussion here of alternative fuels in these modes, partly because there is relatively little in the way of advancement in alternatives for them. Railroad and maritime transport are considered relatively sustainable despite their reliance on nonrenewable energy sources because their distance traveled per unit of fuel used is many times greater than that of motor vehicles. Rail transport is also partially electrified already. Air transport does not have any significant advancement toward alternative fuels at this time, though airplanes have become more efficient and less polluting with time.

Despite the advances in discoveries of existing supplies of petroleum and improvements in efficiency of its use, petroleum faces an undeniable truth of having a finite supply. Sometime in the twenty-first and twenty-second centuries, alternatives to it will be found and petroleum replaced before the supply is exhausted. Whether this happens via a transition made up of technological advancements, policies, and behavioral change, through international conflict to secure remaining supplies, or a combination of the two, remains to be seen.

Acknowledgments

Much of this writing draws from William R. Black’s *Sustainable Transportation: Problems and Solutions* (2010).

SEE ALSO: Built environments; Climate adaptation/mitigation; Climate change policy; Environment and urbanization; New Urbanism; Road transport; Sustainable cities; Sustainable development; Transport and development; Transport geography; Transportation planning; Transport policy; Transport technology; Transportation and land use; Urban transit

References


Further reading


Sweden: Geografiska Förbundet (GF) (Geographical Association)

Founded: 1918
Location of headquarters: Stockholm
Website: http://geof.se
Membership: 150 (as of 2013)
President: Camilla Eriksson
Contact: geoforbundet@gmail.com

Description and purpose

The Geografiska Förbundet was founded in 1918 and is an organization for anyone interested in geography. It has approximately 150 members and is located at the Stockholm University.

The main purpose of the association is to share geographic knowledge and to encourage young geography students. A scholarship is awarded each year to support students’ traveling, for example, field trips related to studies. The student who receives the scholarship presents his/her work in the form of a lecture for the members of the association. The association also arranges other lectures and excursions.

Current activities or projects

The Geografiska Förbundet arranges several lectures during the year. The lectures cover different areas within the subject of geography and are usually open to nonmembers.

Each year a few guided excursions are arranged to places of physical or human geographical interests, such as national parks or cultural sites. Longer excursions to other countries have been arranged in cooperation with similar organizations. All activities are carried out with a strong geographical focus.

Brief history

The GF was founded by geography students at Stockholm University as the Swedish Society for Geography and Anthropology (SSAG) excluded students from its board. The founders of GF initiated a platform for student-driven activities aimed to promote the development of geography as a discipline.

Submitted by Camilla Eriksson
Sweden: Svenska Sällskapet för Antropologi och Geografi (SSAG) (Swedish Society for Anthropology and Geography)

Founded: 1877
Location of headquarters: Stockholm
Website: www.ssag.se
Membership: 500 (as of December 31, 2013)
President: Sten Hagberg
Contact: ssagmail@yahoo.se

Description and purpose

The Swedish Society for Anthropology and Geography (SSAG) aims to support the development of anthropology, geography, and other closely associated disciplines. SSAG supports interactions between academics in these fields and wider society and has links with similar societies outside Sweden. The protector of SSAG is the king of Sweden and the association collaborates with the Royal Swedish Academy of Sciences. SSAG is governed by an elected council with scholars from different universities in Sweden.

Journals or major publication series

Geografiska Annaler Series A, Physical Geography
Geografiska Annaler Series B, Human Geography
YMER (SSAG’s year book)

Current activities or projects

Among other activities, the SSAG organizes meetings with lectures on topics of member interest. The society stages an awards ceremony on Vega day, April 24, and a symposium is held on a topic chosen by the medalist.

SSAG supports research work through financial grants to primarily young scholars. SSAG also cooperates with local and professional associations of geographers and anthropologists in Sweden and is a member the European Geographical Society (EU-GEO).

Brief history

SSAG was founded under its present name in 1877 as an extension of the activities of the Association for Anthropology founded in 1873. This was a time of great exploratory expeditions and the SSAG took a particularly active part in the scientific description of the Arctic region highlighted by the “Vega” voyage through the passage north of Siberia led by Adolf Erik Nordenskiöld in 1878–1879.

In recent years SSAG is more dedicated to the popularization of the three disciplines and to current topics, for example, environmental degradation, physical planning, political geography, and questions of underdevelopment. SSAG has also become more involved in international relations through outreach activities and the development of Geografiska Annaler to become international research journals.

Submitted by Brita Hermelin
Switzerland: Verband Geographie Schweiz/Association Suisse de Géographie/Associazione Svizzera di Geografia (ASG) (Swiss Association of Geography)

Founded: 1881
Location of headquarters: Berne
Website: www.swissgeography.ch
Membership: 600 (as of December 31, 2013)
President: Hans-Rudolf Egli
Contact: hans-rudolf.egli@bluewin.ch

Description and purpose

The Swiss Association of Geography (ASG) was founded to promote geography in science, teaching, practice, and in the public. The ASG is a member of the Swiss Academy of Sciences (SCNAT; see www.scnat.ch) and represents the concerns of geography under the “Geosciences.” The ASG is also a member of the International Geographical Union (IGU; see www.igu-net.org). The association accomplishes its goal through promotion and sponsorships of young researchers, coordination between the different institutions and organizations of geography, the planning and performance of congresses and excursions together with the departments of geography, the advancement of school geography and corresponding technical didactics, and so forth.

Journals or major publication series


Current activities or projects

ASG activities focus on the coordination among the different geographical societies in Switzerland. Some of these societies are also politically active with respect to the geography teaching at the different school levels. The ASG financially supports young researchers for visiting congresses abroad and also supports national and regional conferences. The ASG collaborates with cantonal authorities with respect to environmental and spatial planning, nature protection, and similar issues.

Specific activities are listed among the individual societies affiliated to the ASG.

Brief history

The ASG is the successor organization of the Verband Schweizerischer Geographischer Gesellschaften that was founded in 1881. It is the coordinating link of several geographical organizations in Switzerland. Among these organizations are the SGAG (Schweizerische Gesellschaft für Angewandte Geographie, founded in 1967), the VSGG (Verein Schweizer
SWITZERLAND

Geografielehrpersonen, founded in 1911), the Swiss Geomorphological Society (founded in 1946), the JUGS (Junge Geographie Schweiz), and all geography departments of Switzerland and the regional (cantonal) societies (among them the Geographic and Ethnological Society Zurich that was founded in 1889).

*Geographica Helvetica* is the official journal and is cosponsored by the Swiss Academy of Sciences (SCNAT). The organizing institutions are the Geographic and Ethnological Society of Zurich (GEGZ) and the Swiss Association of Geography (ASG). *Geographica Helvetica* is part of the cultural heritage of Switzerland and is a multilingual journal (German, French, Italian, English).

Submitted by Markus Egli
The increasing complexity of human activity and its impact on the planet Earth demand a comprehensive understanding of how the various dynamic processes, natural and man-made, interact and, in particular, how they affect the natural and socioeconomic environment we are living in. A fundamental prerequisite in order to gain information about the spatial and temporal variation of phenomena of interest is an adequate means of measuring the current state of our environment at different scales, ranging from the local to regional to global scale. Over the last decades, remote sensing technology and the resulting products have become a major source of spatial information for monitoring and planning tasks assumed by local authorities, governmental agencies, supranational and intergovernmental organizations, as well as private industry.

Envisioning the experience of standing on a high building, a tower, a mountain top, or sitting next to the window on an airplane and looking down at the Earth, it is obvious to everyone that observing the Earth from a distance reveals a wealth of information at a single glance – information that is not readily available when standing on the ground in flat terrain. Thus, capturing this information in a systematic way by means of adequate sensors on board aircraft and spacecraft is a consequential next step to just “looking at the Earth” with our own remote sensing device, the human eye.

Remote sensing encompasses all methods that consist of retrieving information about the Earth’s surface by means of measuring and interpreting (energetic) fields that emanate from the Earth. Thereby, electromagnetic radiation that is back-scattered or emitted by the Earth serves as a carrier of information. These definitions are analogously valid for remote sensing of other celestial bodies. Remote sensing provides a means to not only overcome, by far, the very limited resolving capability of a human eye in terms of both the spatial and the spectral resolution, but it also allows for the capture of information that is carried by the nonvisible parts of the electromagnetic spectrum.

With respect to the source of electromagnetic radiation, two classes of remote sensing systems can be identified: passive remote sensing, where electromagnetic radiation from a natural source (e.g., the sun, thermal radiation, etc.) that is back-scattered or emitted by an object of interest is detected, and active remote sensing, where an artificial source of radiation is part of the remote sensing system. Active remote sensing systems are further distinguished into light detection and ranging (LiDAR) systems and radio detection and ranging (radar) systems. LiDAR systems use a laser device as their source of radiation; thus, they work in the optical domain. With radar systems, microwave antennas are employed for transmitting electromagnetic waves and receiving their back-scattered echoes. Electromagnetic waves within the frequency interval from 300 MHz to 300 GHz are commonly termed microwaves. As the term “ranging” suggests, these two types of active systems measure the time delay between
transmitting a signal and receiving the portion of the signal that is back-scattered towards the sensor. The propagation time is related to the distance (range). By combining this information with the sensor’s position and attitude, the geolocation of the object where the scattering process occurred can be determined. In this basic mode, the cross-range resolution of the system is governed by the beam footprint, which is a function of the wavelength, the aperture of the sensor, and the range distance. In the context of radar remote sensing, the resolution obtained in this way is typically low as a consequence of the large wavelengths of microwaves (m to mm), as compared to the wavelengths in the optical domain (μm to nm), for instance. In the next section a concept termed aperture synthesis that alleviates this adverse property is discussed in some detail. The high-resolution imaging radar technique based on this concept is called synthetic aperture radar (SAR).

When thinking of remote sensing as an operational monitoring tool, imaging radar technology has a number of attractive assets. The source of radiation, the radar antenna, is man-made and, therefore, well controlled. It can be deployed in the required manner, limited only by the degree of freedom and mission constraints of the aircraft or spacecraft that carries it. As a further consequence of the artificial source of radiation, the data acquisition schedule is not restricted by external factors such as daylight or weather conditions. Microwaves experience very little attenuation when they propagate through the Earth’s atmosphere. But while only being mildly attenuated, microwaves are subject to polarization rotation (Faraday rotation) and propagation delays in the ionosphere depending on the total electron content; they are subject to propagation delays in the troposphere as a function of the hydrodynamic condition. Microwaves penetrate clouds and, within a substantial range of the microwave frequency band (at wavelengths of the order of decimeters and larger), radar imaging is possible even in case of heavy precipitation. Further, synthetic aperture radar, unlike optical remote sensing, is the only imaging remote sensing technique where the spatial resolution of the resulting imagery is not a function of the distance between the sensor and the object of interest.

**Aperture synthesis in a nutshell**

In this section, a short overview of the general concept of aperture synthesis is given. In particular, aperture synthesis within the context of radar remote sensing is highlighted. These explanations are accompanied by a brief sketch of further relevant basic concepts of radar imaging. For an in-depth treatment of the subject, there are a number of excellent and comprehensive textbooks on synthetic aperture radar signal processing (Curlander and McDonough 1991; Carrara, Goodman, and Majewski 1995; Soumekh 1999; Cumming and Wong 2005) and on Fourier array imaging (Soumekh 1994). Unless indicated otherwise, the subsequent paragraphs are based on material compiled from Soumekh (1994), Born and Wolf (2009), and Demtröder (1999).

**Diffraction**

For many applications it is critical that remote sensing data are available at a high level of detail. Striving for a high spatial resolution has, therefore, always been one of the key challenges in remote sensing. As for any imaging system, there is a fundamental physical limit that determines the maximum resolving performance that can theoretically be achieved by a remote sensing
system. This limit is imposed by the physical phenomenon of diffraction. Waves of any kind (electromagnetic waves, sound waves, ocean waves, etc.) are subject to diffraction when they interact with an object. In the context of the fundamental resolution limit of imaging systems, the diffraction pattern that occurs when a wave front impinges on a slit, a rectangular, or a circular aperture, is of interest. The real-world equivalents would be the antenna array (real aperture) in the case of radar or the slit dimension in optics.

For remote sensing systems, where the physical aperture $D$ is typically very small compared to the distance $R$ from the object under examination, a type of diffraction called Fraunhofer, far-field, or plane-wave diffraction defines the fundamental theoretical resolution.

According to the Huygens–Fresnel principle (Born and Wolf 2009), every point on a wave front is a source of a secondary wavelet. Assuming a 1-D or 2-D aperture, which lies in the tangent plane of the spherical wave, and assuming the far-field case, a planar wave front “impinges” on the aperture at one instance of time. Thus, an infinite number of secondary radiating point sources, which oscillate in phase, occupy the planar wave front within the planar aperture $S$. The radiation stemming from a secondary radiating point source within the aperture $S$ experienced at a point in space with position vector $r$ is then:

$$E_{sec} = \frac{1}{R} a(l) e^{i\omega t - ik |\vec{r} - \vec{r}_s(l)|} dl$$  \hspace{1cm} (1)$$

where $R = |\vec{r} - \vec{r}_s|$. $\vec{r}_s$ is the center of the aperture, $a(l)$ is an amplitude function, $i$ is the imaginary unit, $\omega$ is the angular frequency, $t$ is the time, $k = 2\pi/\lambda$ is the wave number, $\lambda$ again is the wavelength, and $dl$ represents an infinitesimally sized patch of the planar aperture $S$. Hence, the total radiation experienced at the spatial point defined by $\vec{r}$ is obtained by integrating over all secondary radiating point sources within the aperture $S$:

$$E(x, y) = \frac{1}{R} \int_{S} a(l) e^{i\omega t - ik |\vec{r} - \vec{r}_s(l)|} dl$$  \hspace{1cm} (2)$$

which is called the Fresnel–Kirchhoff diffraction integral. Plane waves that are subject to diffraction when passing an aperture (optical case) and planar wave fronts that impinge on, or are transmitted, by a planar antenna are equivalent and, therefore, yield similar diffraction patterns.

In terms of the fundamental resolution of remote sensing systems, we are interested in the diffraction patterns that occur in the far-field case. (A more general (using neither the Fresnel nor the Fraunhofer approximation) treatment of the radiation pattern for dish-type radar antennas can be found in Soumekh (1999).) Essentially, the far-field case is obtained by approximating the nonlinear dependence of equation 2 on

$$R = |\vec{r} - \vec{r}_s(l)|$$

$$= \sqrt{(x - x_s(l))^2 + (y - y_c(l))^2 + (z - z_c(l))^2}$$

using a Taylor series expansion. Choosing the coordinate system in such a way that $z_c(l) = 0$, that is, the $z$-axis runs orthogonal to the aperture plane, then the Taylor series expansion of $R$ yields:

$$R = \sqrt{z^2 + (x - x_s(l))^2 + (y - y_c(l))^2}$$

$$= z \left(1 + \frac{(x - x_s(l))^2}{2z^2} + \frac{(y - y_c(l))^2}{2z^2} + \ldots\right)$$  \hspace{1cm} (3)$$

Stopping the expansion after the quadratic terms, the Fresnel, or near-field, diffraction is obtained ($E$ is the electric field strength) (equation 4):

$$E(x, y, z) = \frac{e^{i\omega z}}{az} \int_{S} E(x_c(l), y_c(l)) e^{ik((x - x_s(l))^2 + (y - y_c(l))^2)/2z^2} dl$$  \hspace{1cm} (4)$$
SYNTHETIC APERTURE RADAR

The Fraunhofer, or far-field, diffraction is obtained by expanding the quadratic terms of the second-order Taylor series expansion of \( R \) (equation 5):

\[
R \approx z \left( 1 + \frac{x^2 - 2xy}{2z^2} + \frac{y^2}{2z^2} - \frac{xy}{z} - \frac{y^2}{z^2} + \frac{x^2 + y^2}{2z^2} \right)
\]

\[
E(x, y, z = z_0) = E_0 \sin \left( \frac{\pi D_X}{\lambda z_0} x \right) \sin \left( \frac{\pi D_Y}{\lambda z_0} y \right)
\]

Similarly, for a circular aperture of diameter \( D \) the Fraunhofer diffraction pattern is obtained

\[
E(r, z = z_0) = E_0 \frac{2J_1 \left( \frac{\pi Dr}{\lambda z_0} \right)}{\pi Dr}
\]

where \( r = \sqrt{x^2 + y^2} \) and \( J_1 \) is the Bessel function of the first kind, integer order one.

An interesting and, in practice, a very useful property of Fraunhofer diffraction is the fact that the amplitude distribution of the Fraunhofer diffraction pattern is directly proportional to the Fourier transform of the aperture shape.

This becomes obvious when equation 8 is rewritten for the rectangular aperture case, for instance (equation 12)

\[
E(x, y, z = z_0) = E_0 \int_{-D_y/2}^{D_y/2} \int_{-D_x/2}^{D_x/2} \frac{e^{-i(kx_0 + ly_0)}}{\sqrt{x_0^2 + y_0^2}} \sin \left( \frac{\pi D_X}{\lambda z_0} x \right) \sin \left( \frac{\pi D_Y}{\lambda z_0} y \right) dx_0 dy_0
\]

which is easily identified as the 2-D Fourier transform of the rectangular function \( \text{rect}(x, y) \):

\[
\text{rect}(x, y) = \begin{cases} 
1, & |x| \leq D_x/2, |y| \leq D_y/2 \\
0, & |x| > D_x/2, |y| > D_y/2
\end{cases}
\]

In Figure 1, the intensity distributions of the Fraunhofer diffraction patterns are shown for three different shapes of apertures: a single vertical slit, a rectangular aperture, and a circular aperture. The following is notable: a constant field amplitude within a rectangular aperture of the dimensions \( D_x, D_y \) yields the same diffraction pattern as two orthogonally running slits of infinite length and widths \( D_x, D_y \), respectively. The horizontal apertures of both the single
Figure 1  Fraunhofer (far-field) diffraction patterns obtained from the following planar apertures: (a) a single vertical slit of horizontal width $D_x = D$ and infinite vertical extension $D_y = \infty$; (b) a rectangular aperture of dimensions $D_x = D$, $D_y = 2D$; (c) a circular aperture of diameter $D$. The grayscale map has been adjusted to increase the visibility of the diffraction pattern.

slit and the rectangular aperture, which lead to the far-field diffraction patterns shown in Figures 1(a) and 1b, are of the same size, $D_x = D$. Thus, the central, or main, lobe of the diffraction pattern has the same width in the horizontal dimension in both cases. The vertical extension $D_y = 2D$ of the rectangular aperture, however, is twice the size of the horizontal aperture, which leads to a main lobe that is only half as wide as compared to the horizontal dimension. The far-field diffraction pattern depicted in Figure 1c results from a circular aperture of diameter $D_r = D$.

In conclusion, for rectangular apertures the intensity distribution of the far-field diffraction pattern is sinc-squared-shaped ($\sin^2(\xi) = \sin^2(\xi)/\xi^2$) in either dimension, and a circular aperture leads to a pattern in the form of the so-called airy function $(2J_1(\rho))^2/\rho^2)$. In the following, the Fraunhofer diffraction pattern is also called the point spread function (PSF) or the impulse response function (IRF) of an imaging system.

The fundamental resolution limit of imaging systems

In the context of imaging systems, the resolution is a measure for the separability of two point targets (dirac pulses). The width of the main lobe of the Fraunhofer diffraction pattern defines the fundamental resolving performance of an imaging system. Looking at the diffraction patterns given in Figure 1, it can be observed that the first dark line or circle defines the width of the main lobe. In terms of mathematical expressions, the dark lines correspond to the roots of the sinc and airy function, respectively. The roots of the sinc function are found at:

$$\xi_0 = \pm n\pi, \quad \text{for } n = 1, \ldots, \infty \quad (13)$$

The outer boundary of the main lobe can be expressed in terms of the aperture $D_x$ and the wavelength $\lambda$ A by combining equation 9 and equation 13:

$$\frac{\pi D_x x}{\lambda z_0} = \pi \quad (14)$$
SYNTHETIC APERTURE RADAR

Further, using \( x/z_0 = \sin(\delta_\theta) \), where \( \delta_\theta \) is called the divergence angle, equation 14 can be rewritten as:

\[
\sin(\delta_\theta) = \frac{\lambda}{D_x} \quad (15)
\]

In optics, a common empirical definition of resolution is the so-called Rayleigh criterion. According to the Rayleigh criterion, two point sources are considered to be just resolved if the center of the main lobe of the IRF of the first point source coincides with the first root of the IRF of the second point source – a circular aperture is assumed. The first root of the airy function is found at \( \rho_0^0 = 1.220\pi \) (Born and Wolf 2009). Analogous to equation 15, this yields

\[
\sin(\delta_\theta) = 1.22 \frac{\lambda}{D_x} \quad (16)
\]

where \( \delta_\theta \) is the angular resolution that could theoretically be obtained if no other resolution-degrading effects were present.

In radar remote sensing (and also for other modern imaging systems) a slightly different definition of resolution is used, as the sensitivity of these devices is usually better than imposed by the Rayleigh criterion. Commonly, and also throughout this work, the \(-3\,dB\) width of the main lobe of the intensity distribution of the Fraunhofer diffraction pattern is used as a definition of the angular resolution, which is obtained by solving

\[
10\log_{10} \left( \frac{I(\pi D/\lambda \sin(\delta_{\theta dB}))}{I_0} \right) = -3 \quad (17)
\]

for \( \sin(\delta_{\theta dB}) \). \( I(\cdot) \) is the intensity distribution of the Fraunhofer diffraction pattern. For the rectangular aperture case, and considering the two-way propagation the relationship is then

\[
\sin(\delta_{\theta dB}) = 0.64 \frac{\lambda}{D} \quad (18)
\]

where \( \delta_{\theta dB} \) is the \(-3\,dB\) angular resolution and \( D \) is the size of the aperture in the particular dimension of the rectangular aperture. The threshold of \(-3\,dB\) corresponds approximately to half the maximal intensity \( I_0 \).

In remote sensing, one is eventually interested in the spatial resolution, that is, the resolution that is achieved when observing an object from a particular distance. The spatial resolution \( \delta_{3dB} \) that is obtained by illuminating an object using one single radar antenna of aperture \( D \) at a distance \( R \) is

\[
\delta_{3dB}(R) = a \sin \left( 0.64 \frac{\lambda}{D} \right) R \quad (19)
\]

With respect to radar remote sensing, two basic essentials can be drawn from these reflections:

1. A smaller aperture (antenna) has a lower directivity, a lower angular resolution, and a larger beam width.
2. The final spatial resolution is directly proportional to the distance of observation \( R \), which means that the spatial resolution deteriorates with increasing distance of observation.

Synthesizing a larger aperture

The basic concept of aperture synthesis involves appropriately combining the electromagnetic waves, which are back-scattered from an object of interest, as they are measured at different sensor locations. Thereby, an aperture that is a hundred to several thousand times larger in size than the actual aperture of a real antenna can be simulated. As a consequence, the main lobe of the synthetic antenna is extremely narrow, leading to a high resolution in the dimension where the aperture synthesis is performed. In SAR remote sensing, this is usually the direction of the motion vector of the sensor, called the azimuth direction. In multibaseline SAR configurations, such as SAR...
In synthetic aperture tomography, a synthetic aperture is also built in the normal direction, which is the direction orthogonal to both the azimuth direction and the line of sight.

Bearing in mind the Huygens–Fresnel principle once more, where the real aperture is thought of as being composed of an infinite number of secondary wavelets, which radiate in phase, this thought can be extended to subsequently placing antennas to form an array, where for all individual antennas an infinite number of secondary wavelets radiate in phase. However, as the synthetic aperture increases, the Fraunhofer approximation is no longer valid. As a consequence, the phase of the electromagnetic waves that are recorded at the different sensor locations has to be altered such that the wave fronts that are echoed by the object of interest interfere constructively for the location at which the object is situated.

Given that premise, the beam divergence angle $\theta_{SA}$ obtained for the synthetic aperture of length $L$ can be written, analogously to the real aperture case (equation 15), as:

$$\sin(\theta_{SA}) = \frac{\lambda}{L} \quad (20)$$

Using the fact that $x/z_0 = \sin(\delta_\theta)$ and setting $x = L/2$, the length of the synthetic aperture $L$, which varies as a function of the range distance, can be expressed in terms of the wavelength $\lambda$ and the aperture $D$ of the real antenna:

$$\frac{L/2}{z_0} = \sin(\delta_\theta) = \frac{\lambda}{D} \quad (21)$$

Combining equation 20 and equation 21, gives the following expression (neglecting tapering factors) for the divergence angle $\theta_{SA}$ of the synthetic aperture:

$$\sin(\theta_{SA}) = \frac{D}{2z_0} \quad (22)$$

The maximal spatial resolution $\delta_{SA}$ obtained with the full synthetic aperture is then:

$$\delta_{SA} = \sin(\theta_{SA})z_0 = \frac{D}{2} \quad (23)$$

Hence, the beneficial characteristics of aperture synthesis can be summarized as follows.

- A very high spatial resolution of half the aperture size of the real antenna is obtained in the dimension in which the synthetic aperture is built. A smaller real aperture (antenna), which has a lower directivity, that is, a larger beam width, is favorable in order to obtain a higher resolution because individual radar echoes measured at different azimuth positions can be combined only as long as the object lies within the beam width of the real antenna.
- Most importantly, the resolution is independent of the distance of observation.

This analysis is valid for an imaging mode called stripmap mode and the additional constraint is that the direction in which the antenna is pointing is orthogonal to the synthetic antenna array during the whole data acquisition period. In contrast, in spotlight mode imaging the antenna beam is steered, mechanically or electronically, such that the scene of interest is illuminated over a longer period and, thereby, the resolution in azimuth direction can be further enhanced (Soumekh 1999; Carrara, Goodman, and Majewski 1995).

Similar to aperture synthesis in the azimuth direction, a synthetic aperture can also be formed in the normal direction, thus yielding a 2-D synthetic aperture. This concept is called multibaseline SAR interferometry or SAR tomography. In practice, the main difference lies in the fact that multiple sensor trajectories, ideally running parallel, are required to build the synthetic aperture. This topic is not discussed here.
So far, discussion has been about imaging resolution in two dimensions of the planar antenna, assuming that an infinitesimally short pulse consisting of a monochromatic wave accurately indicates the propagation delay to the object of interest. For pulsed radar systems, a high resolution in range direction (line of sight) is obtained by typically using a frequency-modulated pulse (chirp) of a length of several microseconds; each radar echo is then compressed later by matched filtering with a replica of the original frequency-modulated pulse. Without going into further detail here, the resolution $\delta_r$ in range direction for the case of a linear frequency modulation leading to a resolution that is inversely proportional to the bandwidth $B$ is:

$$\delta_r = \frac{c}{2B}$$

where $c$ is the propagation speed of the electromagnetic wave.

This brief introduction to aperture synthesis and imaging resolution concludes here but a detailed introduction to the signal model of synthetic aperture radar systems is available elsewhere (Curlander and McDonough 1991; Carrara, Goodman, and Majewski 1995; Soumekh 1999; Cumming and Wong 2005).

Radar remote sensing is only one of many fields where imaging by aperture synthesis is performed. Examples of such imaging systems include very large baseline interferometry (VLBI) in radio astronomy, synthetic aperture sonar (SAS) in the maritime environment, ultrasonic synthetic aperture systems in noninvasive diagnostic medicine, as well as seismic sounding in geophysical exploration (Stolt 1978). In astronomy, even optical aperture synthesis (OAS) using an array of telescopes is possible today, a notable difference being that the large aperture cannot be synthesized by digital post-processing but has to be correlated by a dedicated optical system (e.g., the Very Large Telescope, VLT).

Case studies have also been made for future OAS Earth observation systems from a geostationary orbit. Similar concepts are found in speech and acoustic signal processing and wireless communication systems, mostly under the name of direction-of-arrival estimation (DOA). The fundamental framework of all these techniques is called array signal processing.

As diverse as are the fields where aperture synthesis is applied, so too is the terminology. Thus, a number of terms that essentially all describe resolution enhancement by means of aperture synthesis are used; they are listed here nonexhaustively: (azimuth) focusing, (digital) beamforming, matched filtering, correlation with a reference function, wave equation migration, back-projection, and direction-of-arrival estimation.

**Imaging paradigms in synthetic aperture radar data processing**

Imaging essentially means inverting the measurement process in an adequate way in order to obtain a representation of the objects of interest.

Over the years, a number of different algorithms have been developed or adapted for SAR image processing. This section gives a very brief overview of three imaging paradigms that are commonly used in SAR signal processing. An in-depth review of available SAR focusing algorithms is outside the scope of this entry but a number of excellent textbooks are available that cover this topic at length and in great detail: Curlander and McDonough (1991) and Cumming and Wong (2005) highlight and compare a number of focusing algorithms from a spaceborne SAR perspective and, therefore, predominantly for narrow band width and narrow beam width systems. Carrara, Goodman,
and Majewski (1995), Jakowatz et al. (1996), and Soumekh (1999) all treat SAR signal processing from an airborne SAR perspective. Thus, compensation of irregular motion receives considerable attention in these works.

Basic focusing steps

With the exception of the chirp-scaling-based algorithms, all SAR imaging techniques can be split into a sequence of focusing the data in the range (fast-time) direction (range compression, range-matched filtering) and synthesizing a large aperture in the azimuth (slow-time) direction. In Figure 2 an example of an airborne SAR dataset is given; it visualizes the SAR data as amplitude representations at different stages of the image formation process: (a) before any manipulation, that is, the data are displayed as taken by the sensor before focusing (raw data); (b) at an intermediate stage, after range matched filtering (range-compressed data); and (c) after aperture synthesis has been performed (azimuth-focused data).

In the range-compressed data representation shown in Figure 2b, the 2-D hyperbolic nature of the target signal history becomes obvious: the target appears to be smeared along a hyperbolic trajectory. This nontrivial range/azimuth coupling is exactly what makes synthetic aperture imaging a challenging task. The signal contributions of a point target appear at range distances that vary from pulse to pulse along the azimuth direction, an effect called “range cell migration” (RCM), or “motion through resolution cells.” In addition, perturbation of the signal history in the airborne case, caused by irregular motion of the sensor and varying propagation delays due to atmospheric conditions, further complicate the image inversion.

In the following sections, the most common SAR imaging algorithms are discussed briefly in the context of three imaging paradigms: the range-Doppler paradigm, the wave number domain reconstruction paradigm, and the tomographic paradigm.

Range-Doppler paradigm

Following this paradigm, efficient processing is achieved by making use of the fact that, in SAR remote sensing, the sensor is in motion. Thus, the signal of a point target experiences a Doppler frequency shift depending on the projection of the relative velocity vector onto the connecting line between the sensor and the target. Algorithms that can be assigned to the range-Doppler paradigm are the range-Doppler algorithm (Cumming and Bennett 1979; Wu, Liu, and Jin 1982; Barber 1985), the chirp scaling algorithm (Raney et al. 1994), as well as further extensions of the chirp scaling algorithm such as for squinted geometry (Davidson, Cumming, and Ito 1996), the airborne case including motion compensation (Moreira and Huang 1994), and a more generalized formulation for processing of air- and spaceborne stripmap SAR data and even ScanSAR data (Moreira, Mittermayer, and Scheiber 1996). Another chirp-scaling-based framework for higher order approximation of the hyperbolic range equation, called the generalized frequency domain algorithm, is described by Zaugg and Long (2009).

The basic range-Doppler algorithm involves the assumption of a narrow band width and a narrow beam width, as well as a parabolic approximation (corresponding to the Fresnel approximation of the diffraction pattern) of the hyperbolic shape of the signal history. A further, distinct feature of the range-Doppler algorithm is that the range cell migration correction (RCMC) is performed in the range (time)-Doppler (frequency) domain where targets lying at a certain range distance are superposed as they have the same Doppler
Figure 2  Exemplary SAR data representations (SAR data acquired by the airborne E-SAR sensor of the German Aerospace Center (DLR)): (a) raw data, as measured by the sensor; (b) range-compressed data (i.e., after range-matched filtering); (c) azimuth-focused data (i.e., after aperture synthesis). Note the hyperbolic shape of bright objects as they appear at different range distances in subsequent pulses within the range/azimuth geometry of the range-compressed image. Evidently, the signal processing steps that are needed to build the synthetic aperture are of a 2-D nature (in contrast to the 1-D range compression).
characteristics. RCMC is performed using an interpolation scheme, the choice of which considerably affects the quality of the final image.

The chirp scaling algorithm avoids that critical and time-consuming interpolation step. It involves basically multiplying the SAR data in the range-Doppler domain with a quadratic phase function (chirp scaling) in order to equalize the range cell migration to a reference range, followed by a range compression and a secondary range compression (SRC) in the wave number domain. The SRC is strictly correct only for one reference range and it is updated as a function of the Doppler frequency. The SAR data are imaged without any data interpolation within this focusing procedure. The chirp scaling algorithm in its standard formulation employs a parabolic approximation of the hyperbolic range history. Varying the Doppler rate with range is possible. The RCMC is carried out differentially by means of the chirp scaling operation in the range Doppler domain.

The accuracy of both, the range-Doppler algorithm and the (extended) chirp scaling algorithm, is limited for higher band widths and wide beam widths. While the generalized frequency domain algorithm is theoretically suitable to process SAR data from systems with wide beam widths, large band widths, or low carrier frequencies, no concept that can handle severe deviations from a linear sensor trajectory has been presented so far.

In Walker (1980) the so-called polar format algorithm was introduced; this algorithm was later reformulated from a tomographic point of view. The polar format algorithm is particularly useful for spotlight mode SAR data processing.

**Wave number domain paradigm**

The wave number domain algorithm, also called range migration algorithm, or \( \omega - k \) algorithm (Rocca 1987; Cafforio, Prati, and Rocca 1991; Soumekh 1992; Milman 1993) allows for an approximation-free image reconstruction based on the exact inversion (as far as the phase history is concerned) of the wave equation. The only approximation involved in this reconstruction scheme is the principle of stationary phase, which is used for an explicit evaluation of the Fourier transform of the signal model. The principle of stationary phase is valid in the standard monostatic scenario except in the extreme case of frequencies close to zero (Zaugg and Long 2009). Another exception where the principle of stationary phase is not valid is an extreme bistatic case, such as a combined spaceborne-airborne bistatic data acquisition Wang et al. (2008).

The characteristic feature of this algorithm is a nonlinear coordinate transform also termed Stolt mapping (Stolt 1978), which stems from a seismic migration technique.

Theoretically, the \( \omega - k \) algorithm is a very elegant and approximation-free solution of the SAR image reconstruction problem. In practice, however, several drawbacks become evident; for instance, the exact \( \omega - k \) algorithm lacks the possibility to account for a range varying adjustment of the Doppler rate, which renders it to be an option only for airborne SAR remote sensing. Therefore, approximations of the exact \( \omega - k \) algorithm have been developed for small band width and narrow beam width spaceborne systems. One of these approximative solutions is called monochromatic \( \omega - k \) developed by Rocca, Prati, and Monti–Guarnieri (1989). It is an approximation of the strict \( \omega - k \) algorithm in two ways: The nonlinear Stolt interpolation is replaced by a linear phase term in the range-Doppler domain. Consequently, only the carrier frequency is mapped correctly; hence, the attribute is monochromatic. In addition, a quadratic approximation is used instead of the exact transfer function. The main advantage of
the monochromatic $\omega - k$ algorithm is that it needs no interpolation and, therefore, is very efficient.

Another interpolation-free imaging algorithm deduced from the wave number domain algorithm is the range stacking algorithm (Soumekh 1998), where a (range-varying) matched filter is explicitly calculated for each range bin, followed by a slow-time inverse Fourier transform. In the original wave number domain algorithm, by contrast, the matched filter is applied in the 2-D wave number domain for one reference range. Thus, the Stolt mapping of variables $k_r = \sqrt{4k^2 - k^2_u}$ requires a delicate interpolation of nonlinearly spaced data points to regular grid points in the $(k_r, k_u)$ domain. The range stacking algorithm is computationally more expensive than the wave number domain algorithm, but due to the fact that the azimuth focusing can be performed independently for each range bin, the image formation can be parallelized with ease. In addition, it features the advantage that no interpolation artifacts are present in the final image.

In the airborne case, however, another problem may occur in the sense that strong deviations from a linear trajectory, especially in combination with rugged terrain, cannot be corrected for adequately.

**Tomographic imaging paradigm**

In a pioneering paper, Munson, O’Brien, and Jenkins (1983) highlighted the similarity of spotlight mode SAR data and tomographic data as obtained from, for example, X-ray computerized tomography devices. They gave a 2-D formulation of spotlight SAR imaging in terms of the projection slice theorem known from computer-aided tomography (Kak and Slaney 1988). The signal received at each antenna position is interpreted as a portion of the Fourier transform of a central projection of the imaged area. This interpretation allows for a geometry-based or Radon-transform-based understanding instead of a Doppler-shift-based understanding of the synthetic aperture imaging process. In that paper (Munson, O’Brien, and Jenkins 1983), the polar format algorithm (Walker 1980) originally formulated in a range-Doppler sense is reinterpreted and opposed to a back-projection algorithm. Jakowitz and Thompson (1995) have extended the 2-D view towards a 3-D tomographic formulation of spotlight mode SAR. However, the polar format algorithm is again an approximation of the exact image formation process in the sense that a plane wave approximation is used. This limits the scope of application to narrow beam width systems and spotlight processing of limited-sized patches in terms of the target area (Soumekh 1999).

Another instance of the class of tomographic imaging algorithms is called time-domain back-projection (TDBP). In contrast to all frequency-domain based reconstruction methods, TDBP theoretically allows for a nonlinear aperture shape and incorporation of terrain models into signal processing in a stringent way (Frey *et al.* 2009). The idea of back-projecting the measured data pulse by pulse onto a 2-D or 3-D image space is a very natural solution of the inversion problem: For each sensor position, the range-compressed signal is spread (back-projected) along the spherical shape of constant range in the monostatic case, or along the ellipsoidal shape of constant bistatic range in the bistatic SAR case. Basically, the coherent superposition of the back-projected signals at different range distances and for all antenna positions within the synthetic aperture (as obtained at the pixel locations of the reconstruction grid) yields the final focused SAR image.

It has been stated by Ulander, Hellsten, and Stenstrom (2003) that for airborne wide beam width systems, in particular, an extreme case is
the CARABAS II sensor (Hellsten et al. 1996): the TDBP supersedes the frequency domain algorithms. The TDBP can be regarded as a generalization of the standard beam forming or delay–and–sum approach, on the one hand, or matched filtering of a known signal embedded in Gaussian noise by means of a spatially varying reference kernel for each pixel location on the reconstruction grid, on the other hand.

In summary, depending on the parameters of the SAR sensor (carrier frequency, band width, beam width), platform type (spaceborne, airborne, ground based), and type of application, the appropriate SAR imaging algorithm must be selected. The fast frequency-domain-based algorithms, which rely on simplifications that trade an accurate model of the azimuth reference function for speed, are typically valid for spaceborne narrow band width, narrow beam width SAR sensors, and standard data acquisition modes, such as monostatic stripmap SAR. At the other end of the wide spectrum of SAR imaging algorithms, the time-domain back-projection-based algorithms are found. These algorithms are typically used for airborne sensors (subject to nonlinear platform motion), very wide band width, wide beam width SAR sensors at low carrier frequencies, as well as for intricate data acquisition geometries such as curvilinear or circular SAR, bistatic SAR imaging, or, potentially, for future geosynchronous SAR systems.

Acknowledgments


SEE ALSO: Cryosphere: remote sensing; Digital Earth; Earth system science; Geocomputation; Geodesy; Microwave remote sensing; Natural hazards and disasters; Radar remote sensing

References

Frey, Othmar, Christophe Magnard, Maurice Ruegg, and Erich Meier. 2009. “Focusing of Airborne
SYNTHETIC APERTURE RADAR


Further reading


The discipline of system dynamics was founded by Jay W. Forrester when he utilized principles of control theory to inform hiring, inventory, and production decisions in industrial systems (Forrester 1961). Following from this foundation, the problem-solving approach of system dynamics has been widely applied to the study of business and managerial problems in industrial systems. Forrester’s work became well known when his WORLD model (Forrester 1971) was described in *The Limits to Growth* (Meadows et al. 1972). The WORLD model characterized the global socioeconomic system to study the impact of population growth and consumption on society and the environment. A central insight was that exponential population growth would overshoot and collapse due to resource constraints if no action were taken to avert this scenario. This unpopular conclusion garnered media attention and induced a vocal backlash from critics. However, the book’s fundamental tenet that unlimited growth is not possible on a finite Earth has withstood the test of time: recent editions of the book have updated the original model with new information about population growth and resource use.

System dynamics offers a powerful way to aid human decision-making. Humans make assumptions about the causes and effects of phenomena in a system as well as how the system functions. These assumptions, also known as mental models, help individuals make sense of the system and, thus, influence the decisions to regulate or intervene in the system. Limitations of the human mind for attention, memory, recall, and information processing about dynamically complex problems often produce deficiencies in mental models, such as incomplete causal reasoning and misperceptions about feedbacks. System dynamics provides a way to address these deficiencies with specific methods for representing, testing, and ultimately modifying mental models. Information about a system may be processed and interpreted using different strategies, resulting in very different decisions. Over time, these decisions will change system structures and the subsequent system behavior. A central principle of system dynamics is to study a problem from different perspectives so as to broaden the boundaries of mental models to design better strategies and structures.

The methodology of system dynamics is used to address problems embedded in systems that exhibit dynamic complexity. Complex systems often exhibit dynamic and nonlinear behavior. Such system behavior produces uncertainties, complicating efforts to understand and regulate complex systems. Policy changes and corrective actions can lead to unintended consequences because of feedback processes and delays in the system. A system dynamics model may be used to explore policies for management of these systems.

Complex systems exist in many forms: physical, biological, social (human), natural, or coupled human–natural (variously termed...
Socioecological, human–biophysical, human–environment). Complexity research often involves explorations of the resilient and adaptive capacity of systems in response to stresses and shocks. Familiar to geographers who use geographic information systems (GIS), the term “system” refers to a set of interacting components connected in a distinctive structure. The system forms a boundary-maintaining entity that can be embedded within other systems, establishing a hierarchical structure. A functional system is one that balances the overall purpose with the subpurposes of individual system components and embedded systems.

A system’s behavior is generated by interactions among its elements and the context in which they operate. The model’s structure, in terms of stocks and flows interconnected via feedback loops, determines its characteristic behavior over time. Common patterns of system behavior are exponential growth, decay, goal seeking, oscillation, S-shaped growth, and combinations such as overshoot and collapse.

Central to system dynamics are the concepts of stocks and flows. Stocks are accumulations that integrate their net inflows (the sum of all inflows minus outflows) and, therefore, change only in response to their flows. Stocks are also known as state variables, levels, buffers, inventories, compartments, reservoirs, reactants, or reaction products. Stocks may represent tangible (e.g., water level in a reservoir) or intangible (e.g., customer satisfaction) quantities. Flows refer to rates of change over time and are expressed in the units of the stocks to which they are connected per unit time (i.e., days, months, years). Flows are encoded as actions that increase or decrease the quantity contained in the stock. Because flows occur over time, they are not directly observable. In contrast, stocks reflect the state of the system so that if a snapshot were taken of the system, the stocks would be observable (if also visible). Since they characterize the state of the system, stocks provide information as a basis for action that ultimately feeds back to affect future flows. Those flow rates, in turn, affect future values of the stocks, completing a feedback loop.

Given linkages between stocks and flows, system dynamics models facilitate exploration of time delays that occur when there are time lags between inputs and outputs. Delays tend to introduce instability into the system, affecting material as well as information flows in a system. Material delays are produced by capacity constraints on the flows into and out of stocks, such as those of a manufacturing process. Information delays reflect the time needed to adjust public perceptions or beliefs in response to new information about the state of the system as contained in the stocks. All delays involve stocks, so that stocks endow the system with inertia and memory.

A common metaphor invoked to convey the meaning of stocks and flows is the level of water in a bathtub (stock) as a function of input/outputs of water (flows). As an open container with flows into and out of it, a bathtub creates a stock with water levels changing as water flows into and out of the bathtub. The faucet (water flowing in) and the drain (water flowing out) together regulate the amount of water in the bathtub, which characterizes the state of the system and is observed as level (depth) of water in the bathtub. The flows into and out of the bathtub are controlled by the humans who turn the faucet on or off and close or open the drain. Information about the depth of the water in the bathtub relative to its size is used to decide when to turn off the faucet.

Mathematically, stocks are expressed as integrals that integrate their net inflow for each time step. The equation for the quantity of the stock at time $t$ is as follows:

$$\text{Stock}(t) = \int_{t_0}^{t} [I(s) - O(s)]ds + \text{Stock}(t_0)$$
In this equation, $I(s)$ and $O(s)$ indicate the inflow and outflow rates at any point in time, $s$, between the initial time ($t_0$) and the current time ($t$). In an equilibrium condition, the corresponding differential equation would be expressed as: \[ \frac{d(\text{Stock})}{dt} = I(t) - O(t). \]

While a system dynamics model is mathematically equivalent to a system of coupled ordinary differential equations, it is also characterized by the use of descriptive variable names, visual icons that distinguish stocks from flows and auxiliary variables, and causal information arrows that identify the presence of complete feedback loops. Figure 1 illustrates this iconography, with a box to represent the stock of population, pipeline icons with valves for flows, curved arrows for causal information mechanisms, and cloud icons to indicate sources or sinks. Sources and sinks function as stocks of infinite capacity that are beyond the scope of the problem being represented.

Linear systems can be split into subsystems whose behaviors can be expressed with exact analytical solutions (general solutions to systems of equations). For linear systems, the rate of flow into a stock is constant over time and the sum of parts (the behaviors of subsystems) is equal to the whole (the behavior of the overall system). In contrast, for nonlinear systems the parts connect in such a way that their sum does not equal the whole.

Simulation modeling is well suited to the study of nonlinear behavior. For example, the simulation experiments of Edward Lorenz led to the complexity principle of “sensitive dependence upon initial conditions” or path dependence (Lorenz 1963). These simulation experiments demonstrated that a third-order deterministic system of ordinary differential equations interconnecting three weather state variables could generate chaos if their parameters were perturbed slightly from the rounding of digits when printing simulation results after each day’s worth of experiments.

System dynamics models usually operate under deterministic assumptions, although probabilistic functional forms are feasible. Embedding stochastic behavior into a system dynamics model renders it harder to interpret. Rigorous methods of parameter variation, sensitivity testing, and calibration are instead performed to build confidence in the structure and conditions of a system dynamics model.

Various computer software platforms (i.e., Vensim, Powersim, Stella/iThink, AnyLogic) are available to construct computer simulation models via the graphical language of system dynamics. These software platforms facilitate the practice of system dynamics using a graphical interface with distinct iconography for stocks, flows, and information feedback. Vensim was used to construct the models represented in the figures of this entry.

As a simplified representation, a system dynamics model has utility when it can explain why characteristic behavioral patterns are being produced by the system. Identifying which structures and conditions generate the observed behavior reveals leverage points for policy change. Leverage points signal opportunities to intervene in the system where relatively small changes can be introduced so as to encourage larger-scale changes in system behavior to
achieve desired goals. System dynamics models can be used to assess the impact of different courses of action in the short term and long term. A fundamental insight of system dynamics, and systems thinking more generally, is that the problems and the solutions are often endogenous to the system in that the most effective change comes from within.

The modeling process

Although there are different ways to model a problem using system dynamics, the general modeling process usually involves the following steps (Sterman 2000):

1. defining the problem;
2. developing a dynamic hypothesis (causal map) of the problem;
3. formulating the model with stock and flow structures;
4. calibrating and testing model boundaries;
5. considering the impact of different policies and scenarios.

The stages of the modeling process may be interwoven, as it may prove valuable to revisit hypotheses (step 2) after beginning the model formulation (step 3), or to perform model testing (step 4) on scenarios (step 5) that have been articulated. Some system dynamics practitioners begin with stock–flow structures (step 3), and different steps are emphasized depending on the purpose of the modeling effort. The process of modeling is iterative, with each cycle of model design and experimentation resulting in new insights for modelers, model users, and problem stakeholders. In addition to the model itself, the modeling process can provide a mechanism for gaining insights about how a problem came to exist and what actions should be taken to address this problem. After a model is developed, discrepancies in simulation results and expectations may motivate new hypotheses and additional experiments to better understand alternative hypotheses.

In step 1, the problem is defined and characterized as dynamic in nature. The system dynamics approach emphasizes modeling of a particular problem, not the system itself (which would be more difficult to bound without a problem orientation). Useful system dynamics models are those designed to investigate policy issues to solve a specific problem that exhibits dynamic behavior.

To define the problem, system variables relevant to the problem of interest that exhibit dynamic behavior are first identified. Identifying system variables (elements that vary over time) helps determine the model scope, that is, what to put in the model. Drawing out reference modes and time horizons for system variables is part of articulating the problem dynamically. Observed behaviors over time are characterized graphically as reference modes from which to compare simulation results. Reference modes, also known as behavior over time graphs, characterize the dynamic pattern as a point of reference during the modeling process. The x-axis of the reference modes spans the time horizon to be simulated. The length of the time horizon will influence the perspectives on the problem. Longer time horizons allow for consideration of system delays and long-term impacts of feedback processes that contribute to unexpected system behaviors, which are often overlooked. The y-axis of the behavior over time graph should represent a magnitude either on a continuous scale (e.g., 0–100) or a categorical scale (e.g., low, medium, high). It is common practice for different trends over time to be charted for hoped, feared, and expected behaviors of key state variables in the reference modes.
An illustration of reference modes is provided for the example of technology diffusion in Figure 2. Diffusion, also referred to as dispersal or spillover, is a dynamic process. The reference mode for the diffusion process is the S-shaped curve, as for the population of adopters (green line) in Figure 2. The S-shaped (sigmoidal) curve or logistic growth curve is characterized by a slow initial incline, followed by a rapid, uniform acceleration until about the midpoint of the curve, at which point there is continuous rise that progressively decelerates until leveling off at an upper limit. S-shaped growth is generated by a system where the interaction between reinforcing and balancing loops is nonlinear; there are no significant time delays in balancing loops; and the upper limit is fixed. The “tipping point” of diffusion processes occurs when there is a shift in loop dominance from reinforcing feedback (growth) to balancing feedback (market saturation). This point is characterized by the maximum adoption rate (dotted line) in Figure 2.

Dynamic patterns of system behavior are produced by time delays and interactions between feedback loops, particularly nonlinear interactions. As outlined in Ford (2010), common reference modes characterizing patterns of system behavior include reinforcing feedback for exponential growth; balancing feedback for exponential approach (goal-seeking behavior); negative feedback with significant time delays for oscillation; nonlinear interaction between reinforcing and balancing feedbacks with no significant time delays and a fixed upper limit for S-shaped growth; S-shaped growth, but with significant time delays for S-shaped growth with overshoot; and S-shaped growth, but with a diminishing upper limit over time for overshoot and collapse.

In step 2, a dynamic hypothesis is developed as a plausible explanation for system behavior. Causal maps (also known as causal loop diagrams) provide a systematic way to develop dynamic hypotheses and identify important feedback loops. Because the modeling process is iterative, the dynamic hypothesis serves as a working theory that may be modified in future iterations. As a conceptual framework, the dynamic hypothesis provides an important link between theory and model design.

The system dynamics approach aims to find endogenous explanations for observed phenomena within the structure of the system. Therefore, causal maps of dynamic hypotheses are composed of cycles created by linking variables through causes and effects to create feedback loops.

The process of building a causal map generally begins with the identification of which variable best characterizes the state of the system. This variable is then rendered endogenous by linking relevant system variables in cause and effect relationships. Upstream factors or causes are identified, followed by the downstream factors or effects that are associated with the problem. Lastly, links are made between upstream and downstream factors. In terms of downstream factors, modelers using system dynamics should...
think beyond expected and immediate outcomes. During this step of the modeling process, modelers should consider the difference between short-term and long-term outcomes, as well as possible unintended consequences of policy changes. The linking of downstream and upstream factors should yield complete feedback loops. Richardson (1991) illustrates numerous examples of feedback mechanisms implicit in social theory.

In causal maps, links between variables are represented with arrows that indicate the direction of causation. Each arrow between pairwise relationships is also labeled with a polarity, a positive (+) or negative (−) sign. The polarity denotes whether variable X increases or decreases relative to variable Y, ceteris paribus (i.e., holding all else equal). For each pairwise relationship, the modeler asks: if X were to increase, would Y increase or decrease compared to what it would otherwise have been? A positive sign (+) can be interpreted as follows: all else remaining equal, an increase in variable X increases variable Y above what it would otherwise have been. A negative sign (−) can be interpreted as follows: all else remaining equal, an increase in variable X decreases variable Y below what it otherwise would have been. These relationships hold when the system is responding to a decrease in variable X, so that a positive causal link means the effect variable Y decreases it below what it would otherwise have been. Similarly, a negative causal link means that the effect variable Y increases in response to a decrease in X. An even number of negative (opposite) relationships or a positive product of polarities indicates a reinforcing feedback loop, whereas an odd number of negative (opposite) relationships or a negative product of polarities indicates a balancing feedback loop. An important check when constructing causal maps is to think through the logic of a feedback loop to determine whether it is reinforcing or balancing.

When appropriately presented, causal maps can reveal and communicate potential mechanisms driving the dynamic behavior via feedback loops. An example causal map is illustrated in Figure 3 for a hypothesis of the diffusion behavior depicted in Figure 2. The left-hand loop is denoted “B” for balancing feedback in the saturation of the potential market as increased adoption rate (a flow, in units of people per year who adopt the technology) reduces the stock of potential adopters and, thereby, diminishes the pool of potential adopters not previously
exposed to the technology. The right-hand loop is labeled “R” for reinforcing feedback exhibited by growth of the market due to word of mouth as an increase in adopters fuels the probability that a potential adopter will encounter adopters, further increasing the adoption rate.

System variables may be exogenous or endogenous. Endogenous variables are embedded within the system boundary, whereas exogenous variables impinge upon the system from outside the system boundary. The variables in the causal map expressed in Figure 3 are endogenous. In step 3, each endogenous variable mapped in step 2 is formally related to its causes with an explicit equation. Exogenous variables are defined as parameters that can be controlled to create distinct simulation experiments. Units are assigned and checked for dimensional consistency.

In step 3, a computer simulation model is formulated using stock and flow structures interconnected with feedback loops. Formulating a computer simulation model involves specifying assumptions as parameter settings and equations encoding decision rules.

Figure 4 illustrates the full stock–flow model corresponding to the causal map depicted in Figure 3. Particularly, the diminishing stock of potential adopters is seen to be part of the balancing feedback loop, and the growing stock of adopters fuels the reinforcing feedback loop through word of mouth about the innovation that is modulated with the contact rate parameter. This example is adapted from Sterman (2000).

The equations for the innovation diffusion model depicted in Figure 4 are:

\[ A(t) = \int_{t_0}^{t} [ar(s)]ds + A_0(t_0) \]
\[ P(t) + A(t) = n \]

\[ P(t) = \int_{t_0}^{t} [-ar(s)]ds + n - A_0(t_0) \]
\[ ar(t) = cr \times i \times P(t) \times \left( \frac{A(t)}{n} \right) \]
where: \( A \) = number of adopters (people); \( A_0 \) = initial number of adopters (people); \( n \) = total population (people); \( P \) = number of potential adopters (people); \( ar \) = adoption rate (people/day); \( cr \) = contact rate (people/person-day); \( i \) = probability of adoption after exposure to innovation (dimensionless); \( P_0 \) = initial number of potential adopters, \( n - A_0 \) (people); and \( j \) = probability of encountering adopters, as fraction \( A/n \) (dimensionless).

Variables are classified as stocks, flows, auxiliaries (intermediaries), and parameters (Table 1). By definition, the functional form of a stock is an integral of the net inflow. In addition, each stock is assigned an initial condition.

Auxiliaries, such as “encounters made by potential adopters” and “probability of encountering adopters” in Figure 4 and Table 1, are intermediate variables used for representing a part of the problem that corresponds to particular mathematical operations. By breaking down larger equations into smaller ones, auxiliaries enable the model logic to be more transparent and intuitive to the model user.

Because it involves transforming a conceptual model into clearly defined structures and decision rules in a computer simulation model, step 3 may expose gaps in problem articulation (step 1) and resolve potential discrepancies in the logic of the causal map (step 2). Modifications may need to be made, as there may be limitations in estimating parameters and initial conditions, but the model purpose and model boundary should guide the construction of the simulation model.

Attention should be given to checking the consistency of units, referred to as dimensional analysis, when formulating the model. Unit checking is a standard option in system dynamics software. After the model equations are completely formulated and the model has been checked for unit consistency, the model is simulated.

Although modelers often justly derive much satisfaction in getting their models to run after investing effort into model design, it is critically important to budget sufficient time for the last two steps of the modeling process, namely testing and analysis. Step 4 involves model testing. In

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Term</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential adopters</td>
<td>Stock</td>
<td>( P(t) )</td>
<td>People</td>
</tr>
<tr>
<td>Adopters</td>
<td>Stock</td>
<td>( A(t) )</td>
<td>People</td>
</tr>
<tr>
<td>Adoption rate</td>
<td>Flow</td>
<td>( ar(t) )</td>
<td>People per day</td>
</tr>
<tr>
<td>Contact rate</td>
<td>Parameter</td>
<td>( \alpha )</td>
<td>People encountered per person per day</td>
</tr>
<tr>
<td>Initial adopters</td>
<td>Parameter</td>
<td>( A_0(t_0) )</td>
<td>People, where ( P_0(t_0) = n - A_0(t_0) )</td>
</tr>
<tr>
<td>Total population</td>
<td>Parameter</td>
<td>( n )</td>
<td>People</td>
</tr>
<tr>
<td>Probability of adoption</td>
<td>Parameter</td>
<td>( i )</td>
<td>Fraction of exposed potential adopters who adopt innovation</td>
</tr>
<tr>
<td>Encounters made by potential adopters</td>
<td>Auxiliary</td>
<td>( e )</td>
<td>People per day = ( P(t) ) * ( \alpha )</td>
</tr>
<tr>
<td>Probability of encounters with adopters</td>
<td>Auxiliary</td>
<td>( j )</td>
<td>Dimensionless = ( A(t)/n )</td>
</tr>
</tbody>
</table>
this step, model boundaries are first tested by adjusting parameter values to extreme settings to determine where the model logic holds and where it breaks down. Then the simulation model may be calibrated by adjusting or “tuning” parameter values to match an objective function comparing the simulated results to empirical data, as in the case of time series. There are several tests that can be performed but the fundamental tests are comparison of simulated behavior to historical behavior (reference modes), robustness under extreme conditions, and sensitivity of uncertain parameters. These tests are conducted to assess model performance and utility, that is, to determine where the model performs well and where the model breaks down. Shock tests for model robustness involve setting extreme model parameter values, while sensitivity tests involve setting a range of model parameter values. In both robustness and sensitivity testing, of interest is observing how the simulated behavior over time responds to different parameter changes and its comparison to the reference mode. Model testing can indicate structural flaws in the model, such as errors that were introduced when formulating the model (step 3), or critical model components that are missing. Furthermore, in addition to evaluating model credibility, documentation of results from model testing improves the transparency of the model.

The graph in Figure 5 illustrates a sensitivity test indicating outcomes for the stock of adopters that occur when changing the parameter setting for contact rate. Confidence intervals for the range of outcomes are shown with the color legend. Increasing contact rate to the maximum value hastens market saturation to about 20 days, but slowing it to the minimum value slows the pace beyond the 120-day horizon of the graph.

Model calibration is appropriate if a model is expected to provide a forecast (as opposed to used for exploratory or pedagogical purposes). The base case under default model settings is illustrated in Figure 5 with a single blue line saturating soon after 30 days. This behavior is faster than the reference mode depicted in Figure 1. Therefore, a step toward calibration in this case would be to lower the average contact rate until the base case slows enough to match the reference mode.

Once confidence in the model has been established through the process of exploring model behavior, the model can become useful as tool for decision support and policymaking. In step 5, policies and scenarios are designed and analyzed. In this stage of the modeling process, the model is used to explore how system behavior responds to alternative policies and scenarios. Here, policies refer to system-wide interventions, whereas scenarios refer to particular plausible cases of interest (e.g., hoped and feared market behavior in response to a new technology). Because the modeling process is iterative, scenario development can be undertaken at the beginning of the modeling process, along with problem definition. Scenario analysis, however, involves the use of a reasonable model to explore implications of policy shifts.
Scenario analysis is performed to recommend policy shifts that would induce desired outcomes. Sensitivity testing on a wide range of scenarios should be performed because interactions of different policies may produce unanticipated interferences or synergies. High-leverage policies are often system interventions that can change dominant feedback loops and delays. An important objective at this final stage of a modeling effort is to design proactive rather than reactive policies, which can alter system structure to address the root of the problem.

Advanced system dynamics analysis involves identification of which feedback mechanisms dominate behavior at different points in time. In the example of technology diffusion, reinforcing feedback of market growth precedes the dominance of balancing feedback from market saturation. Feedback loop dominance can have a profound impact on the equilibrium or end state of the system, especially in two interesting ways, through tipping points and path dependence. Shifts in dominance of reinforcing feedback loops to balancing feedback loops (or vice versa) create tipping points, also known as bifurcation points. Path dependence is present when small changes or random system shocks early in the process significantly impact the equilibrium or end system states. Path dependence tends to arise when there are dominant reinforcing feedback loops in the system.

Outlook and extensibility

System dynamics is one of several methods of computational modeling employed in systems science (others include GIScience, ABM and social network analysis) and, as such, may be fruitfully applied to the study of complex systems. Simulation models provide a way to conduct experiments in a virtual world that may be unethical, harmful, or too resource intensive to conduct in the real world.

Since models are necessarily simplifications of real-world systems, all aspects of the system may not be captured in models. The utility of models is in the identification of the core elements and processes of the system associated with the problem of interest. The difficulty of addressing problems in complex systems is, in part, due to the distance in space and time between the causes and outcome. An advantage of computer simulation models is the compression of space and time.

Whether tangible (e.g., water level in a reservoir) or intangible (e.g., trust in a relationship), stocks reflect aggregate or average attributes at a specific scale. Although the contents of each stock are assumed to be homogeneous and randomly mixed, stock and flow structures are often replicated to represent distinct population subgroups. Such subgroups can be used to reflect spatial distribution of population by neighborhood or region.

Stocks can be considered as compartments or containers with an implicit spatial dimension and scale. Therefore, stock–flow system dynamics may be spatially extended to represent quantities like population that vary with place. The stock–flow structures of system dynamics may be extended using interlinked cellular automata to enable raster-compatible simulation (BenDor and Metcalf 2006). These structures may also be embedded in individual agents, or may characterize the environment in which agents interact. Standard agent-based modeling software (RePast, NetLogo, AnyLogic) now includes the capacity for representing stocks, flows, and information feedback. BenDor and Kaza (2012) explore the potential for characterizing spatially explicit reference modes for use in models involving spatial system dynamics.
Because the path dependence of complex systems is attributable to their sensitive dependence on initial conditions, an argument can be made for the importance of characterizing environmental context when simulating such systems. In a system dynamics model, geospatial context would affect the initial level of the stocks, even if the hypothesized feedback mechanisms were considered universal.

Acknowledgment

The authors acknowledge NIH (OBSSR/NIDCR) R01 award DE023072 for partial support in the writing of this article.

SEE ALSO: Agent-based modeling; Cellular automata; Cultural diffusion; Disease diffusion; Finite-element and finite-difference modeling; Representation: dynamic complex systems; Spatiotemporal analysis; Validity and verification

References


Taiwan: Xhong guo di li xue hui (The Geographical Society of China located in Taipei (GSC-Taipei))

Founded: 1934
Location of headquarters: Taipei
Website: www.geogsoc.org.tw
Membership: 350 (as of August 2014)
Chair: Jinn-Guey Lay
Contact: jglay@ntu.edu.tw

Description and purpose

The GSC-Taipei serves to publish the Zhong guo di li xue hui hui kan (Bulletin of the Geographical Society of China); organize local conferences, coordinates research and teaching, and special lectures; organize local and international excursions; and organize international conferences. In addition, the GSC-Taipei provides logistical support for our members who undertake research and educational endeavors.

Journals or major publication series

Zhong guo di li xue hui hui kan (Bulletin of the Geographical Society of China)

Current activities or projects

IGU conferences that have been organized by GSC-Taipei include Transnational Lives: Feminist Perspectives on Citizenship, Home and Belonging in 2007, Island Development in 2013, and Geomorphology & Society in 2014.

In 2014, the society supported projects such as a training program of environmental education for Yehliu Geopark (from the Ministry of Transportation and Communications) and the Geography Olympiad for Senior High School Students in Taiwan (from the Ministry of Education.)

Brief history

GSC-Taipei was founded in 1934 in Nanjing and reestablished in 1951 in Taipei 2 years after the civil war in China. It was formally registered with the Ministry of Interior, Republic of China in Taiwan, in 1977, with the goal of developing geography as a scientific discipline.

Submitted by Jin-Guey Lay
Technology and development

James T. Murphy
Clark University, USA

In the public’s perception, technology and development are commonly viewed as being in lock-step with one another, with development driven by technology and technology serving as a key indicator of development. In reality this somewhat tautological relationship belies the fact that technological change is often highly problematic with respect to its socioeconomic and environmental implications as it may exacerbate inequality, uneven development, ecological degradation, and/or social exclusion. A critical understanding of the drivers, dynamics, implications, and geographically uneven distributions of technology and technological change is thus an important component of development studies and practice, and an area in which geographers contribute critical insights.

While there are many definitions of what technology is, it is perhaps best understood as an expansive and multifaceted phenomenon as Metcalfe (2010, 154) notes:

By technology, we can mean a body of understanding of cause and effect in human minds, as with the codified realizations of productive knowledge in operating manuals, blueprints, recipes, scientific papers and so on; the capacities and skills that permit action, whether individually or in cooperation with others, not all of which is written down; and the purposefully organized and designed built structures within which action takes place – the realized, human built world.

Viewed in this manner, technology is much more than simply artifacts, computer algorithms, and/or built systems or facilities, it is cognitively and often tacitly constructed and embedded deeply in social, cultural, economic, and political systems. Moreover, technology is an inherently geographical phenomenon in that it emerges out of particular contexts, is often diffused spatially, can be scaled up or down (e.g., from the local to the global), and has uneven geographies of use, significance, and impact.

With respect to development, technology is viewed by most as an essential driver and determinant of socioeconomic, cultural, environmental, and political change. Economically, technology can increase productivity through improvements to the efficiency of production, exchange, logistics, and service-provisioning activities while speeding up innovation-development and knowledge creation processes. Alternatively, technology can exacerbate socioeconomic (e.g., class) differences, create uneven development within and between countries, and/or further the exploitation of labor through more intensive, real-time monitoring of workplaces. Culturally, technology has a profound effect on the norms, conventions, identities, and traits that help to constitute particular social groups, often destabilizing these and creating generational divisions. Environmentally, technology can contribute in significant ways to greener and more sustainable societies or exacerbate ecological degradation through intensified or expanded impacts locally and globally. Politically, technology can have democratizing effects (e.g., see the so-called Facebook revolutions in the Middle East) or it can facilitate enhanced forms of repression or surveillance by
TECHNOLOGY AND DEVELOPMENT

state authorities. All told, the impacts of technological change on development are complex, contingent, sometimes unpredictable, and often highly uneven geographically. Technology can thus serve as a valuable lens through which one can examine the context-specific and multi-scalar drivers of, and obstacles to, more even, sustainable, and just forms of development.

Neoclassical views on technology and its role in/for development

The most prevalent perspective on technology is that it is principally an economic phenomenon; one that is a key driver of economic growth created in response to scarcity or in pursuit of profit. Neoclassical (orthodox) economists generally view technological change as manifest in productivity improvements – in the form of labor savings, increased output, decreased costs, and/or scale economies that enhance the efficiency of production. Early frameworks (Solow 1957) viewed technology as an external factor (or residual to) in econometric models, but one that plays a crucial role in shifting national production functions such that higher levels of output (i.e., GDP) are possible. With Romer’s (1990) work this perspective shifted somewhat as technological change (understood in part as human capital) was conceptualized as being endogenous to the economy, determined in large part by investments in education, research, and innovation. As knowledge – a nonrival good manifest in increasing levels of human capital – becomes more widespread, it creates positive externalities able to sustain long-run growth (development) through measureable productivity improvements driven by market forces and profit-maximizing entrepreneurs and firms.

Although countries that are early innovators or leading technological centers have an initial advantage with respect to their development, the neoclassical perspective holds that there will be, over time, a “natural” convergence between lower-tech and higher-tech places provided the former increase their rate of capital (human, physical) investment such that productivity rises at a higher rate than that of so-called mature economies. Increasing productivity will enhance growth, increase national output, and reduce poverty as all boats are lifted through the surpluses generated. Particularly important for this developmental process are productivity improvements in agriculture which can free up labor for work in manufacturing sectors able to add more value, accelerate the pace of technological change, and create the kinds of localization and urbanization economies that enhance a country’s comparative advantage in global trade. In sum, the neoclassical view of the technology-development nexus conforms to Rostow’s modernist perspective of a singular/linear trajectory; one that reflects the evolution of advanced industrial economies during the nineteenth and twentieth centuries.

Contextualizing technological change: institutional, systemic, and constructivist perspectives

In the neoclassical view, technology is essentialized in four significant and problematic ways. First, technologies, manifest as productivity improvements and are seen as being inevitably positive/progressive given that they drive economic growth (i.e., development). Second, technologies are developed by utility or profit-maximizing agents who make rational decisions about how to resolve socioeconomic challenges with only the most productive solutions surviving in the context of market-based competition. Third, technologies are principally
artifacts, skills, algorithms, and so forth that quantifiably increase productivity and/or the scale/scope of economic production; thus meaning that they are best understood as tangible/codified “things” that can be diffused or spread from one context to another with the same effect or outcome. Fourth, the economic impacts (i.e., productivity contributions) that a technology makes are generally assumed to be the same everywhere, thus making the neoclassical view by and large ageographical.

Such perspectives on technology have been challenged effectively by social scientists who view them as acontextual, oversimplified, abstract, and/or narrowly focused on codified (patented, westernized) scientific discoveries and discontinuous shifts in production functions, rather than more incremental or tacit changes that are often not manifest in tangible or material innovations (Rosenberg 1976). The push-back against neoclassical understandings of technology and its role in development emerged in response to the many failures – environmental, social, public health, and economic – that accompanied modernization and rapid technological change/diffusion in the post-World War II period (e.g., thalidomide and DDT poisoning). Rather than deterministically assuming that modern technology must create development, new questions emerged regarding who technologies were for, how/why they are made/developed, and what their constitution and impacts can tell us about societies and economies more generally. Such inquiries began as explorations by heterodox economists into the evolution of technologies and innovation-development systems, eventually leading to the emergence of the field of science and technology studies.

For heterodox economists, the direction, pace, scale, and impacts of technological change are shaped fundamentally by the institutional and geographical environments where innovations are developed; the paradigm, cognitive frame, or “gaze” that technologists deploy in developing their ideas; and forces and factors in the wider societal context (Dosi 1982). Context-sensitive perspectives such as these provide a helpful foil to technologically deterministic frameworks as they shift the focus from artifacts to institutions, culture, built and natural environments, political economies, and sociotechnical systems wherein technologies are developed, innovation diffusion occurs, and/or path dependencies or development trajectories evolve. Particular emphasis is placed on the governance of national and regional systems of innovation, networks, and institutions within and through which new technologies and industries are developed (Freeman 1995). The development of such systems is shaped by multiscalar and context-specific relationships and struggles between state agencies, policymakers, firms, civil society actors, universities, and innovators who seek to reproduce or transform national or regional economies through technological change. In-depth analyses of these networks, institutions, and governance processes can provide key insights regarding the challenges of overcoming path dependencies and so-called lock-ins that make alternative development trajectories difficult to achieve.

Heterodox perspectives on the evolution of regional economies helped to spawn the emergence of the field of sociotechnical transitions research with its particular emphasis on the prospects for more environmentally and socially sustainable forms of development (Truffer and Coenen 2012). This work assesses the prospects for, and potential drivers of, technology-driven and systemic shifts (transitions) away from business-as-usual development pathways; trajectories that generate excessive waste, pollution, and other environmental and social externalities (e.g., climate change, inequality, poverty).
Transition researchers examine the regimes and innovation systems associated with key service-provisioning sectors – such as energy, water, housing, sanitation, food, and transportation – in order to determine if and how greener technologies might destabilize and redirect them toward more sustainable outcomes. Importantly, this work contextualizes regimes and technological innovation systems in the wider context of countries, regional authorities, societies, and the global economy in order to more fully understand the obstacles to, and potential drivers of, sustainable development (Geels and Schot 2007).

A key question for transition scholars is how to coordinate and manage the diverse actors, infrastructures, institutions, and markets governing a sociotechnical regime such that business-as-usual technologies are replaced by greener alternatives that might enable a transition toward long-run sustainability. Work in this regard has examined increasingly transitions processes in the Global South in order to determine, in particular, if developing regions can establish alternative sociotechnical systems such as those that are not dependent on carbon-based fuels.

Beyond innovation systems and sociotechnical transitions research, others have contextualized technologies by critically examining the ways in which they are developed and embedded in particular social, cultural, political, and historical contexts that shape significantly the direction and substance of their contributions to development. Considered in this manner, technologies are seen as social constructions developed, promoted, supported, and/or diffused by states, firms, and societal groups to achieve or reproduce desired power relations, profits, and/or socioeconomic conditions. Through analyses of these dynamics, constructivist perspectives strive to overcome technologically deterministic views in order to understand the coevolution of technological change and socioeconomic development at a variety of scales – from individual artifacts to sociotechnical systems and cultures (Bijker 2010). Research in the constructivist tradition typically focuses on one or more of the following analyses: (i) deconstruction of the technology itself in order to understand different interpretations of its use; (ii) an assessment of how the technology was constructed in the first place; and/or (iii) an examination into the role that particular social groups played in framing and constructing the technology.

Constructivist studies reveal critical insights regarding the contingent origins of technologies, their persistent use despite better alternatives, the unexpected ways in which they are absorbed into contexts that differ from where they originated, and their contribution to path dependencies with respect to industrial, regional, and community development. For example, although ultrasound imaging technologies have proven to be highly significant tools for prenatal medical care, their diffusion into some societies (e.g., India) has led to higher incidences of sex selection given that male children are more valued in patriarchal settings. Other examples include the persistence of the QWERTY keyboard in a post-typewriter world, despite its inefficiencies, and the ways in which energy, production, and transportation systems worldwide are generally locked in to carbon-based fuels despite our knowledge of the climate-change crisis and the other negative social and environmental externalities associated with fossil fuel extraction, processing, and consumption. Through in-depth, qualitative analyses of such lock-ins and path dependencies, constructivist perspectives demonstrate the challenges of overcoming dominant technological paradigms through the diffusion and deployment of improved technologies, particularly in contexts that differ in significant ways from those where they were developed. Such insights are crucial in order to move beyond deterministic
views of technology’s role in development and to design more appropriate (i.e., context-specific, distributive) technologies to meet the needs of the poor and disempowered.

**Radical and alternative perspectives on technology’s role in/for development**

Beyond contextualized and constructivist analyses of the technology-development nexus there is a significant body of work that takes a much more critical view, particularly of technology in its mainstream or dominant forms. Radical (Marxist), feminist, and postcolonial/postdevelopment scholars have made important contributions in this regard, reminding us of technology’s tendency to (re)produce repressive, unequal, and/or exploitative socioeconomic systems. Moreover, such scholarship highlights the ways in which technologies and knowledges situated/produced outside the westernized, capitalist context are marginalized by mainstream development institutions and states. In response to such failures some have called for alternative technology development and diffusion systems, ones that value indigenous knowledge, welfare distribution, the environment, and community wellbeing above growth, profit, and individual measures of productivity.

For Marxists, the technologies produced and diffused in the context of capitalism are fundamentally the outcome of the class struggles associated with control over the means of production (Smith 2010). As capital increased its power historically, its need to extract surplus value from labor has resulted in technological changes that tend either to substitute machinery for workers or to deskill them such that wages can be compressed. The net result is that technology helps to disempower the working classes and enhance the ability of capitalists to reap profits through the exploitation of living and so-called dead labor (i.e., that which is embodied in the machinery used in industries). Absent a class revolution whereby workers take control over the means of production, and the scientific/technological apparatus, capitalist-driven technological change will reproduce and exacerbate inequality, deprivation, and exploitation despite the improvements that seem to accompany modern innovations.

Rather than focusing on class, feminist analyses reveal the sociocultural embeddedness of technologies and their subsequent bias or tendency to reproduce patriarchal systems that essentialize or marginalize women’s roles in society. For Merchant (1980), such biases and inequalities accelerated during the Enlightenment as a new value system emerged in the West, one that sought to order society and nature as if they were machines which could be technologically engineered in rational, predictable, and productive ways. Women’s role in the societal machinery was reduced to that of social reproduction (e.g., child rearing, household labor) with the powers-that-be in society valuing technologies that sustained this status quo by restricting women’s personal freedom, access to technology, and mobility in all its forms. Feminist scholars further demonstrate that technologies – artifacts, the built environment, and knowledge in all its forms – should be understood as phenomena inscribed with valences or biases that reflect societal values regarding gender roles and relations, among others (e.g., class relations) (Wajcman 2010). Consideration of such biases or tendencies, and their embedded nature in particular contexts, is essential in order to understand the implications of technology development and diffusion for women’s empowerment and wellbeing. Rather than viewing technology as being inevitably problematic with respect to such concerns, some (e.g., Haraway 1991)
view today’s hyper-technological/information age as one laden with political, economic, and social possibilities for women, manifest in new identities and opportunities that can destabilize essentialized notions of gender roles and create political space for difference and diversity.

A third major critique of mainstream views on technology-for-development is concerned with the ways in which the Western scientific and technological project marginalizes subjects and knowledges existing/produced in the Global South. These critiques began with concerns over the loss of indigenous forms of knowledge (IK) that evolved in situ, but which were devalued or delegitimized by Western science and its prioritization of universal, globalized, and rationally engineered technologies (Agrawal 1995). More recent work has been inspired by postcolonial and actor-network theory (ANT) which seeks to disrupt the “purity” of Western science and technology through studies that demonstrate its messy, contingent, and often highly subjective nature, while highlighting the value of IK and alternative forms of knowledge as means to create more emancipatory, distributive, and collective technologies in/for development. A key concern of these scholars has been attempts by mainstream development institutions to co-opt, codify, and decontextualize these technologies in ways that make them more compatible with Western science and development practices (Briggs and Sharp 2004). For example, some have raised concerns over initiatives such as the World Bank’s “Knowledge Bank” (https://openknowledge.worldbank.org), a database that strives to collect, codify, and centralize information about successful livelihood, production, and development practices such that they are available globally. Critics charge that initiatives such as these can either distort the often tacit and localized knowledge associated with such practices such that it adheres to the ideology and priorities of mainstream development institutions (e.g., growth, individual productivity, market-led development), or exclude altogether practices associated with radical or alternative perspectives on development.

Beyond offering critiques of mainstream and capitalist-oriented views of the technology-development nexus, radical, feminist, and postcolonial perspectives have inspired initiatives to protect IK and support appropriate technologies to empower the poor and disadvantaged. In general terms, appropriate technologies are those designed to meet the direct/subsistence needs of people and their communities better, rather than foster economic growth and individuated forms of productivity. Inspired initially by E.F. Schumacher’s (1973) book Small is Beautiful, appropriate technologies privilege public/social utility, human values, justice, accessibility, and context specificity over scale, universality, and market driven forms of diffusion. As such, they are often associated with social movements and grassroots organizations which seek to delink the wellbeing of communities and their environments from the technologies and forces of global (corporatized) capitalism such that livelihoods can become more autonomous and resilient. Developing appropriate technologies demands a finer grained geographical understanding of the wider social, cultural, and political contexts wherein livelihoods are practiced, one that puts the priorities, capabilities, and wellbeing of communities first (e.g., see Via Campesina’s notion of a peasant agroecology at http://viacampesina.org/en/). Some have called for the development of a postcolonial technoscience as a means to more systematically develop technological alternatives to facilitate more just, distributive, and progressive forms of socioeconomic development in the Global South and beyond (Anderson 2002).
Development through technology diffusion: from modernization to translation

For geographers and others, a traditional concern with respect to the technology-development nexus has been the processes through which technologies diffuse across space and over time. This is a fundamentally geographical question, one that spawned the “diffusion of innovations” literature starting the 1950s. Early diffusion theorists – E.M. Rogers (1962) and Torsten Hägerstrand (1967) – emphasized the cultural and behavioral drivers of technology diffusion, arguing that the geographical spread of modern innovations depends upon whether individuals in peripheral places are able to embrace change and shift away from traditional practices and understandings of their place in the world. Viewed from this perspective, the (technological) diffusion of modernization is dependent upon changes in individual behavior that can, over time, aggregate into community and regional-scale transformations. In spatial terms, à la Hägerstrand (1967), information and social communication are crucial and it is essential for practitioners (e.g., agriculture extension officers) to establish centers or hearths from which an innovation can spread. For Rogers (1962), technology driven development objectives can be achieved through a process of getting local opinion leaders and others to adopt and effectively use new technologies such that other community members follow suit and eventually replace the “old” ways (e.g., peasant farming) with modern, productive, and efficient techniques.

As Figure 1 demonstrates, the early diffusion literature (especially Rogers 1962) presumed that there is a “normal” diffusion pattern or distribution of behavioral types in any given setting, one that determines the pace and scale of a technology’s spread. The goal for development practitioners is to find ways to increase adoption levels (i.e., to provide sufficient information or communication in order to change behaviors) such that a majority of the population uses the technology consistently and effectively. Once an (early) majority is reached, the diffusion process can proceed in a more “natural” fashion as late adopters and so-called laggards eventually take up the innovation. Considered in this manner, technology diffusion – as a means for development – depends to a large extent on spatial and

---

social proximity, spurred on when neighbors or individuals of similar standing copy one another's behavior. Such forms of proximity thus serve as central explanatory factors with respect to the diffusion of innovations across space and over time.

Diffusion studies were further enhanced by Brown (1981) who extended early frameworks in order to better account for the context wherein adoption occurs. Rather than focusing solely on the demand-side of technology adoption (i.e., individual behaviors), Brown's framework considers supply side factors such as markets, institutions, and infrastructure in determining one's access to new innovations. Moreover, his work sought to explicitly address the implications of innovation diffusion for development and to assess the impacts that levels of development have on the diffusion process. This work highlighted and conceptualized both the contingent nature of the technology-development nexus and the context-specific factors that create sometimes highly uneven patterns of technology uptake, particularly in the Global South.

In many respects, the behavioral perspective in diffusion studies remains pervasive in mainstream development studies despite numerous critiques of the technological determinism and binary categorizations (e.g., traditional versus modern) built into Rogers's framework. More critical, poststructurally inspired, studies of technology diffusion emerged along with the radical and alternative perspectives discussed above. In addition to explaining why certain social and cultural groups (e.g., poor people, women, etc.) “fail” to adopt modern innovations despite their efficiencies and utilities, this work provided in-depth, contextualized, and ethnographic research into technology adoption and absorption processes. Particularly important is work inspired by ANT which, among other things, examines critically the ways in which technologies move across space and how they are integrated into distanced social, political, and economic systems such that they contribute to new or existing forms of order (Law 1992). Considered in this manner, diffusion occurs through a process of “translation” whereby the technology is “assembled” with other aspects of the social and material worlds actors inhabit such that it helps to generate positive or negative development outcomes. Importantly, the ANT perspective, while bestowing a degree of autonomy and agency on technologies themselves, is keen to understand and describe the power relations that enable translation to occur, and the effects these have on material and social outcomes.

For example, de Laet and Mol's (2000) careful, ANT inspired, analysis of the diffusion of the Zimbabwe bush pump in Africa reveals the technology’s origins and fluid meanings as it is “translated” across space and over time. Through a detailed ethnography of this seemingly “modest” technology, they demonstrate its adaptability and flexibility with respect to the contexts where it is distributed. Such fluidity, as they term it, means that the pump can be translated or “enrolled” into a diversity of cultures, communities, and sociotechnical systems throughout Africa and thus help to generate effects such as cleaner drinking water, fewer water-borne illnesses and disease, and the social capital that comes with the collective endeavor of installing and maintaining the pump (see Figure 2). As such, they demonstrate how the effective and widespread diffusion of so-called appropriate technologies depends on their “firmness” or mutability with respect to geographical context, the latter of which is preferable as it makes technological absorption more likely.

To conclude, the technology–development nexus is by no means a unidirectional or predictable relationship given the myriad factors and contingencies that shape innovation and
diffusion processes. For human geographers this is thus an important area of inquiry given its unevenness, the multiscalar nature of technology development processes, and the crucial role that spatial relationships play in shaping diffusion dynamics. The ability to move beyond neoclassical and modernist abstractions regarding what technologies are, and how they contribute to development, has been made possible through geographically sensitive and critical inquiries into the nature of technologies themselves, the drivers of their development and diffusion, and the context-specific factors that determine their accessibility, utility, and appropriateness.

**SEE ALSO:** Actor-network theory; Gender and development; Indigenous knowledge; Modernity; Power and development; Technology; Technology spillover

**References**


Technology spillover

Yifei Sun
*California State University, Northridge, USA*

Peilei Fan
*Michigan State University, USA*

Technology spillover refers to the unintentional technological benefits gained by firms due to the actions of other firms. Though spillover can happen among firms that belong to the same industry, geographers are most concerned with spillover that occurs among firms co-located in the same geographical space. As such, technology spillover is a phenomena associated with agglomeration. Most research has focused on spillover of foreign direct investment (FDI) on domestic firms in emerging economies. It is hoped that FDI can strengthen the technological capabilities and innovation of domestic firms in developing countries so that such sustainable development can be achieved in the long term. Domestic firms in developing countries can benefit from FDI through competition, demonstration, supplier–client networking, and labor mobility. First, domestic firms can imitate foreign products in order to reduce market uncertainties. Domestic firms can also learn the management practices and marketing strategies of the foreign competitors. Second, the spillover effect can also be fulfilled through labor turnovers from foreign firms to local firms. When workers are relocated to local firms, they can apply the management experiences and skills that they learned from foreign firms. However, there could also be negative impacts; domestic firms may lose highly skilled labor because foreign firms in general offer higher salaries and better working conditions than domestic firms and so attract local talent. Third, domestic firms can benefit from FDI spillover through backward linkages (contracts between foreign affiliates and domestic suppliers) or forward linkages (interactions between foreign suppliers of intermediate input and domestic suppliers) with foreign firms. Domestic firms can learn from foreign customers on product development, quality control, and technological improvement. Domestic firms can also improve their product quality by using intermediate products from foreign companies. Finally, foreign firms also stimulate competition in the domestic market. Fierce competition can urge domestic firms to use more advanced technologies or new technologies and
resources in order to survive the competition, which may also lower the demand for local products or even force domestic firms out of the local market.

Absorptive capabilities, the technological gap, and technology spillovers

Empirical results on spillovers have shown conflicting results, from positive impact to no impact and negative impact. The most consistent results concern the moderating effects of a firm’s absorptive capabilities (Cohen and Levinthal 1990, 128), which refers to “a firm’s ability to recognize the value of new, external information, assimilate it, and apply it to commercial ends.” Absorptive capability is more than just effort and investment in research and development (R&D) but more importantly includes a firm’s combinative capability, which is a function of the firm’s level of prior related knowledge. Absorptive capability is more than just the sum of individual skill or knowledge levels, and can be acquired, retained, and updated through an accumulative process.

Most recent research on spillovers along this line has further examined the technological gap between foreign firms and domestic firms. It is argued that there may exist a nonlinear relationship between the technological gap and spillover effects. Spillover reaches a peak when the technological gap is of medium size: if the technological gap is either too big or too small, the potential spillover will be lower. When the technological gap is too big, the technology becomes too advanced for the local firms to learn. When the technological gap is too small, local firms become less interested/motivated to make efforts to learn from foreign firms.

Technology spillover: a social process

Technology spillover is a social process that involves multiple actors situated in a particular social context. Research has demonstrated that organizational and cultural proximity between the sending and receiving firms are important in understanding the process of technology spillovers. Technology spillover relies on a shared “interpretation system,” “shared systems of meaning,” and shared “organizational focus,” since people “interpret, understand and evaluate the world differently” (Nooteboom et al. 2007), and such dimensions are summarized as “cognitive proximity,” which denotes “a broad range of mental activity, including proprioception, perception, sense making, categorization, inference, value judgments, emotions, and feelings.” However, the relationship between cognitive proximity and technology spillovers is not necessarily linear, and it is argued that the relationship between absorptive capability and cognitive distance may follow an inverted U shape; however, such research is only preliminary and more solid work is needed to clarify the issues here.

Geographical context and technology spillover

Traditional research on spillover has been spaceless and assumes that spillover occurs among firms that are located in the same industry within a country. Knowledge spillover tends to be geographically bounded within regions where the new economic knowledge was created (Feldman and Audretsch, 1999). Geographical proximity enhances the interactions between local and foreign firms and the spillovers should be affected as well. Demonstration or imitation effects are more likely to occur when foreign
and domestic firms are co-located. It is easier for local firms to learn advanced technologies or collect market information from their neighbors. Similarly, frequent labor turnover between foreign and domestic firms is more likely to occur when they are located in proximity to one another. Geographical concentration can also enhance interfirm linkages. Domestic firms are more likely to become the suppliers or customers of foreign firms. Similarly, firms are likely to experience competition from local competitors, foreign or domestic. Recent research has examined FDI spillover at different geographical scales: from the national to regional and local scales (Sun and Du 2010; Zhang et al. 2010). It is interesting to note that studies at different geographical scales reveal great differences in their findings, even studies that are situated in the same country.

Conclusions

Technology spillover is a social process of learning. It does not occur automatically and does require intentional and sustained efforts made by the receiving firms. As such, it requires firms to invest in internal capabilities that enable them to identify the opportunities and internalize the technologies that are in their geographic and relational space. Research on spillovers should expand from the firm-centered perspective to the social perspective that incorporates the geographical context in which the learning takes place. Such research has the potential to further our understanding of spillovers from foreign firms to domestic firms. Finally, the current research that focuses on spillovers of foreign firms from advanced economies to firms in developing countries should be expanded. With more and more firms going global, it is necessary to examine how foreign and domestic firms can learn from each other.

SEE ALSO: Global commodity/value chains; Industrial linkage

References


From *tekhnē* (meaning craft, or art) and *logos* (the application of reason), technology might suggest the “art of reason.” But it has just as often tended to mean quite the reverse, referring instead to the rational principles that govern the construction of artefacts, “the reason of art rather than the art of reason” (Ingold 2000, 294). In some senses, this reversal in meaning captures the ambivalence of technology. For many, technology signifies hope, freedom from drudgery, and the chance for a better life. But the application of technological reason to human dealings with themselves and with the planet can also lead to dehumanized and environmentally damaging results. Hydroelectric dams, IT networks, genetic modification – indeed anything we choose to call technology – can suggest tragic outcomes as well as hope for a better future.

Part of the problem is that technology, as the downstream “application” of reason or science to real-world issues, can imply a particular and rather reductive understanding of the world. It can tend to treat the Earth as a vast machine, and “through a rational scientific understanding of its principles and functioning, this machine [can] be harnessed to serve human interests and purposes” (Ingold 2000, 294–295). While this machine metaphor seems to suggest a human mastery over nature and continued human progress, the lessons from the last century have been that the world rarely behaves in this fashion, and the social and ecological costs of technologies are as common as its benefits.

An inspiration for this kind of critique of technology comes from the twentieth-century German philosopher Martin Heidegger. While Heidegger saw technology as ambiguous, he nevertheless wrote of the tendency for technical rationality to produce an impoverished understanding of the world and an instrumental treatment of both humans and nonhumans. For Heidegger, technical rationality can degrade the Earth and humanity, producing a “tear in being” and “enframing” living beings and other components of an environment as mere resources or what he called standing reserve (Heidegger 1978). He was writing in part in opposition to Bolshevist and other ideological investments in technological modernity. His warnings were meant to disrupt any blind faith in technological solutions to problems of scarcity, hunger, and labor. For Heidegger, technology was descent as much as ascent, and we need to remember this ambiguity lest we fall into the trap of technological optimism.

More starkly, perhaps, for recent interpreters of Heidegger, technology seems inevitably to divide Being from beings (or entities), and sets up a distinction between the proper and the improper, the hand-written and the typed, “man” and human. This distinction involves an ordering of both human and nonhuman lives that makes it easier to take or disqualify those lives (so technology makes everyone and everything look the same, and in so doing we enter a world of expendable, dehumanized beings). In biopolitical renderings (see Biopolitics), technology is something that makes some human and nonhuman lives of lesser value and able to
be sacrificed for some supposedly greater good. It is a dark side to technology that can help us to think through the atrocities and tragedies of the disastrous twentieth century. The holocaust, ethnic cleansings, mass animal slaughter, and environmental disasters all characterize this most technological of centuries.

These dehumanizing effects of technologies remain key debates in geography. As the question of technology is posed less in terms of a debate between Bolshevism and National Socialism, and more so in terms of neoliberalism (see also Neoliberalism and the environment), globalization, and disaster capitalism, there has been a return to an interest in the essences of technology, and a thanatopolitical (or towards death) reading of technology and biopolitics. At the same time, though, more affirmative treatments exist, in some ways reinvoking the notion of technological progress as, for example, the widespread adoption of social media and Internet technology provide the conditions of possibility for new kinds of politics (Hardt and Negri 2005). The former offers warnings concerning technical rationality (for instance, in relation to geoengineering) while the latter looks for the political openings that a change in technological consciousness may present.

The problem with some of these accounts is that the future already seems mapped out, for better or for worse. Some would regard this as technological determinism, or the ascription of essential qualities to a technology that then dictates how it will be used. Such a view denies geographical variation and seems to foreclose political action. In contrast, more empirically grounded work, which attends to the role of technologies in generating or transforming political situations in ways that are not preset or determined in advance, seems to offer an alternative. Indeed, it may be that the more recent attempts in geography and cognate social sciences to bring together the lively materialisms associated with feminist inspired science and technology studies, such as those of Donna Haraway (Haraway 1985), start to open up a technopolitics that is both situated and attenant to the dynamics of socially embedded technologies (see also Feminist geography). This more empirically focused work tends to return “technology” to an art of reasoning and involves attending to the broad sweep of activities involved in generating knowledge (which is already in this sense technological), and is part and parcel of acting in the world. Technology, here, becomes hopeful once again, though certainly not in an idealistic or utopic sense. Aiding from ethnographic studies concerned with the practical “doing” or “enacting” of science and technology, we start to see the art of reasoning as an embodied, distributed, and often fraught process. Technologies can, in this sense, open up debate and politics, and can form the issues over which technological publics come into being.

Studies of science and technology in practice foreground the interrelations and intra-actions between people, tools, machines, and other nonhumans, showing how their assembly, or configuration, has significant effects on outcomes (Barad 2007) (see also Actor-network theory). Underestimating the complexity of entanglements between people, electricity grids, financial investment, and climate can, for example, lead to catastrophic failures. Realizing the agency of all these and more can help to re-enchant and possibly improve such relations (Bennett 2001). Technology in this sense is no more than the quality of its relations. As a result, an intermaterial analysis or a more spatial, empirically focused geography of technology has taken shape with a concentration on the uncertainties and indeterminacies of technologies in complex and power-laden fields.
Technology as a process, one that of necessity requires listening to and acting with a whole range of both human and nonhuman others, is indeed an art. More significantly, its futures are not already determined as there is scope here for doing technology in ways that are more, as well as less, progressive. Rather than technological determinism (the sense that a particular technology always leads to particular outcomes), here the politics of technology, while always dangerous, monstrous, and liable to co-option, is nevertheless more open and to be worked out.

SEE ALSO: Biotechnology; Geoengineering/climate intervention; Indigenous technical knowledge; Interdisciplinarity and geography; Posthumanism; Uncertainty

References


Further reading

Tectonic geomorphology

Lewis Andrew Owen
University of Cincinnati, USA

Tectonic geomorphology is one of the youngest branches of geomorphology, clearly merging the study of geomorphology and the relative new geologic discipline of tectonics. Not only does tectonic geomorphology combine these two immense fields, but it also incorporates the study of geodesy, geophysics, geochemistry, geochronology, archaeology, and paleoclimatology. In essence, tectonic geomorphology examines the relationship between Earth’s internal (tectonic/endogenetic) and external (surface/exogenic) processes, their role in landscape development, and their implications for tectonic and geomorphic hazard mitigation. Owen (2013b) provides a perspective on the development of tectonic geomorphology as a discipline and emphasizes how this evolving area of study has provided a far deeper awareness of the importance of tectonics in helping to drive the study of landscape development and surface processes. In addition, tectonic geomorphology has helped force awareness of the strong relationships between endogenetic and exogenic processes, which have in turn led to many new geologic and geomorphic paradigms.

As Owen (2013b) points out, there is no formal definition for tectonic geomorphology, but it is broadly defined as the study of landforms produced by tectonic processes such as faulting, warping/tilting, folding, uplift or subsidence, and the application of geomorphology to the study of landscape evolution and tectonic problems. Some authorities include volcanic processes and landforms under the heading of tectonic geomorphology (Owen 2013a), while others do not (Burbank and Anderson 2012/2001). Tectonic geomorphology as a term did not become part of the geosciences lexicon until the 1980s and it was not until 1998 that it was used as a keyword for articles cited in ScienceDirect (Owen 2013b). Moreover, tectonic geomorphology was not mentioned, or only given fleeting consideration, in encyclopedias and textbooks until the middle of the first decade of the twenty-first century, when it became a more widely recognized area of study.

Tectonic geomorphology has been enhanced and boosted since the late 1990s with the development of new technologies including remote sensing, GIS science, geophysical exploration methods, enhanced microcomputing, and geochronology. A search for tectonic geomorphology in ScienceDirect today yields more than 13,000 articles and more than 80,000 results listed in a search of Google Scholar. The publication of Burbank and Anderson’s textbook, Tectonic Geomorphology in 2001 helped establish tectonic geomorphology’s place in the mainstream of geomorphology and tectonics. The Treatise on Geomorphology (Shroder 2013) published in 2013 dedicates one of its 14 volumes solely to tectonic geomorphology (Owen 2013a).

Although considered a young branch of geomorphology, the study of tectonic geomorphology (but not called that, of course) has been an integral part of landscape evolution studies for some time. Such concepts as the “Davisian cycle of erosion” and various models of “landscape evolution” have included the role of uplift and subsidence driven by endogenetic processes.
Aspects of tectonic geomorphology have also been included under the broader themes of neotectonics and active tectonics during the latter part of the twentieth century. Of particular note was Vita-Finzi’s (1986) book, *Recent Earth Movements*, which provided a modern overview of neotectonics and included many aspects of tectonic geomorphology, though not specifically referred to by name. Keller and Pinter’s (1996) book, *Active Tectonics*, also provided a wealth of information on tectonic geomorphology, including a chapter called “Landforms, Tectonic Geomorphology, and Quaternary Chronology,” with other chapters focusing on geomorphic indices, rivers, coastlines, and mountain building, all pertinent to tectonic geomorphology.

Quaternary science, particularly the study of rates and magnitude of climate change, underwent a mini-revolution during the latter part of the twentieth century which continues into this millennium. The realization that climate change throughout the Quaternary was very rapid and often of high magnitude, influenced thoughts and studies on the likely responses of surface processes to these changes and helped drive renewed interest in climatic geomorphology. As a consequence, researchers began to explore the connections among climate, tectonics, and landscape evolution. This helped strengthen interest in tectonic geomorphology and studies of landscape development, particularly the polygenetic origin of landscapes.

Today, tectonic geomorphologists have an impressive arsenal of tools at their disposal. Most tectonic geomorphic studies begin with a detailed analysis of landforms and the landscape, which usually requires the acquisition of topographic and/or other remote sensing data (Figure 1). Topographic data such as LiDAR can be digitally manipulated to enhance and reveal subtle topographic features. In recent years, the development of Google Earth has proved particularly useful, and in many cases has replaced more the more traditional analysis of landforms using stereoscopic aerial photography (Figure 2). Improved global positioning systems are also enabling very quick and accurate maps to be constructed, and comparisons to be made over time. In recent years, stereoscopic analysis of photographs can now be easily and quickly done on microcomputers, in particular enabling maps to be made quickly and accurately without the need for expensive surveying equipment.

Improvements and advances in Quaternary geochronology have been a particular boon for understanding the ages of landforms and the rates of processes underlying tectonic geomorphology; these include the use of radiocarbon-dating, terrestrial cosmogenic nuclides (TCN), optically stimulated luminescence dating, U-series geochronology, and various systems of thermochronology. The ability to accurately and precisely date tectonic landforms is critical to assessing rates and styles of deformation, and landscape evolution. In essence these methods are used to determine the timing of formation or deformation of landforms. Some of the methods can be used to quantify rates of denudation and the timing and duration of sediment burial, such as those using TCNs. Thermochronology allows exhumation histories to be determined for large-scale topography. In all cases, the power to unravel the tectonic geomorphic history of a study area comes from combining high-fidelity age control with high-resolution remotely sensed imagery and/or surveying/map-making. Only by combining these two datasets can the tectonic geomorphic history be fully understood.
Tectonic landforms are diverse and pervasive throughout the world but are most evident along active plate margins (Figure 1). In fact, it may be argued that all landscapes are essentially a consequence of tectonics because tectonic processes are partially responsible for the distribution of the bedrock that constitutes the ground on which landforms develop. However, tectonic landforms are generally considered to be those landforms that have formed in direct response to geologically young crustal movements. The definition of “geologically young” can vary given different perspectives and geologic/geographic locations. A good definition may be a landform/landscape that has formed over a timescale of about $10^{-3}–10^7$ years, the duration over which geomorphic processes are considered to operate.

Continental-scale (around $10^{3–4}$ km$^2$) tectonic landforms include mountain ranges, plateaus, volcanic island arcs, ocean trenches, basins, great escarpments, drainage systems, and linear valleys, among others. Such features can be readily observed from space and in the field (Figure 1). On a regional-scale ($10^{1–3}$ km$^2$), tectonic landforms include marine terraces, fault-bounded mountain fronts, transverse and longitudinal drainages, grabens and volcanic centers (Figure 2). Local-scale (around $10^{-3}–10^{-1}$ km$^2$) tectonic landforms include fault scarps, shutter ridges, sag ponds, landslides, sand
Figure 2  Regional-scale tectonic landforms. (a) LiDAR data (gray) overlain on an oblique Google Earth image of the part of the right lateral strike-slip Banning Fault near San Gorgonio Pass in Southern California. The Banning Fault is part of the San Andreas Fault system, and this image illustrates offset alluvial fans, impressive fault scarps, and fault valleys. This example illustrates some of the rapidly evolving geographic and geodetic technologies/tools that are enabling tectonic geomorphology to be more readily studied. (b) View south along the San Andreas Fault near Wrightwood in Southern California. The San Andreas Fault runs along an impressive valley at this location, with the Mojave Desert to the east and the San Gabriel Mountains to the west. The city of Palmdale can be seen in the Mojave Desert in the middle part of the view. The relative motion between the North American (to the east) and the Pacific Plate (to the west) is about 30 mm/year along this stretch of the San Andreas Fault. (c) View northwards into Death Valley. As the lowest point in North America, Death Valley is a pull-apart basin formed by normal and strike-slip faulting that bound its eastern margin.
blows, volcanoes, lava flows, and river terraces (Figures 2 and 3). The smallest-scale (around $10^{-1}$ m$^2$ and less) features include such phenomena as fissures, cracks, individual rock falls, displaced streams, small terraces (marine or river) and liquefaction structures (Figure 3). These tectonic landforms have been described in much detail with appropriate examples in Burbank and Anderson (2012/2001) and Owen (2013a).

Owen (2013a, 2013b) identified several main areas of emphasis within tectonic geomorphology. These are modified and summarized below.

**Landscape and tectonic evolution of active plate margins**

Tectonic geomorphic studies have been undertaken on nearly every, if not all, active plate margins. In particular, there has been much interest in quantifying geomorphic rates ($10^0$–$10^5$ years) of vertical and horizontal crustal displacement to help define tectonic models and for seismic hazard assessment. Much emphasis has been placed on comparing geomorphic with geologic ($10^2$–$10^6$ years) and geodetic ($10^{-1}$–$10^2$ years) rates of displacement, particularly along fault systems such as those along the Gulf of California–San Andreas transform plate boundary, for seismic hazard mitigation (Frankel and Owen 2013). Other areas of focus include the development of landscape evolution models, most notably for the Himalayan–Tibetan orogen and southern and southeastern Alaska (Koons, Zeitler, and Hallet 2013).

**Mountain building**

Tectonic geomorphic methods have been applied in numerous mountain belts to help understand their evolution and to help identify which areas are more tectonically active. In recent years, it has become increasingly apparent that young mountain belts are the product of localized interactions among tectonics, climate and erosion (Koons, Zeitler, and Hallet 2013). A major focus of interest is whether steady state can be achieved in mountain belts and whether tectonics and/or climate control the width and asymmetry of mountain belts. Tectonic geomorphic studies are adding much to this debate, as discussed in detail in Koons, Zeitler, and Hallet (2013). Moreover, it is becoming apparent that tectonic geomorphic evidence, for example, in the form of topography, terraces, and drainage systems may provide important insights into possible influences of crustal flow on the geomorphic and tectonic evolution of continental plateaus such as Tibet and the Colorado Plateau (Koons, Zeitler, and Hallet 2013).

**Development of fault and fold systems**

The development of fault and fold systems has attracted much attention on a variety of scales ($10^2$–$10^8$ m$^2$). The tectonic geomorphology of many continental-scale strike-slip faults has been studied in much detail, including those along the Pacific–North American plate boundary and those within the Himalayan–Tibetan orogen (Frankel and Owen 2013). Measuring and dating offset landforms has enabled rates of displacement to be quantified on geomorphic timescales. Particular emphasis has been placed on what geomorphology can reveal about fold development, especially for seismic hazard mitigation because of the growing recognition that large earthquakes occur within fold belts, often on blind thrusts. In addition, examining drainage development as well as the deformation of river terraces along traverse drainages has been particularly insightful.
for understanding the relationship between deformation and landscape development.

Evolution of passive margins, continental interiors, and plateau uplift

Tectonic geomorphic studies of passive margins, continental interiors and plateau uplift have provided important new insights into how continents evolve. Of particular note has been the study of continental-scale landforms, such as great escarpments and plateaus, along passive margins. Blenkinsop and Moore (2013) summarize the diversity of passive margins and discuss how they reflect variations amongst thermal, isostatic, flexural, buckling, and dynamic (mantle) sources of stress that affect margins.

Study of the evolution of continental interiors has focused much on the very long-term geomorphic evolution of shield areas, particularly examining how drainage systems develop. Examining plateau development has also benefited greatly from tectonic geomorphic studies. This has included studies that aim to define rates of plateau uplift, quantify paleoelevation, and document drainage development. Many mechanisms have been proposed for plateau uplift including denudational unloading, and uplift associated with hotspots and delamination; these are discussed in detail in Blenkinsop and Moore (2013). A further area of research related to plateau development concerns the feedback between plateau uplift and global climate change.

Volcanic landforms and processes

Volcanoes are among the most dynamic landforms on our planet, yet tectonic geomorphic studies on volcanoes are still very much in their infancy. Nevertheless, the tectonic geomorphology of volcanoes has provided important insights into our understanding of the evolution of volcanic islands and continental volcanic arcs along subduction zones. Moreover the development of volcanic centers over hotspots has provided valuable information for tectonic modeling. Studies often focus on volcanic hazards, particularly as there is a growing awareness of the potential for volcanoes to have local, regional and global catastrophic effects. Volcanic eruptions may trigger other catastrophic

---

**Figure 3** Small- and local-scale tectonic landforms. (a) Tectonically formed cracks in alluvial fan surfaces near Antofagasta that were produced during the 1995 Antofagasta earthquake in Chile. (b) Small offset drainage channel on an alluvial fan in Death Valley displaced by the right-lateral strike-slip Northern Death Valley Fault. (c) and (d) Fault scarps produced during the 2010 El Mayor–Cucapah earthquake (magnitude 7.2) across an active alluvial fan (c) and along a mountain front (d). (e) Bedrock fault scarp of the Sierra Nevada Frontal Fault in Owens Valley, California, that defines the eastern margin of the Sierra Nevada and westernmost part for the Basin and Range Province. (f) Low-angled normal fault defining the western margin of the Black Mountains in Death Valley. These mountain slopes are produced by low-angled normal faults in Death Valley and have become known as turtlebacks because they resemble the shape of a turtle’s shell. (g) View of the Hattian Bala landslide in northern Pakistan from its debris pile. This landslide was triggered during the October 8, 2005 Kashmir earthquake resulting in the burial of three villages and more than 700 fatalities. (h) View looking across the city of Orotava in Tenerife. This city is built within a giant landslide scar that formed on the slopes of the Rico de Teide, a volcano that rises to 3728 m above sea level. The submarine run-out from this landslide stretches for about 100 km beneath the sea and the volume of debris is about 1000 km³. The landslide age is younger than about 0.78 Ma and may even be younger than 0.2 Ma.
Figure 3
TECTONIC GEOMORPHOLOGY

tectonic-geomorphic hazards including landsliding and tsunami. As such, real-time monitoring by such research group/organizations such as US Cascade Volcano Observatory is becoming increasingly important, especially in areas near large population centers such as in the Pacific Northwest of the United States, Japan, Indonesia, and the Philippines.

Paleoseismology and seismic hazard assessment

Over the last few decades paleoseismology has emerged as its own subdiscipline within geology, aimed at determining the timing and magnitude of past earthquakes. Ultimately, paleoseismology can be used to help determine earthquake recurrence intervals for seismic hazard mitigation. Much paleoseismological work depends on tectonic geomorphology to identify active faults that can be trenched to provide a record of past earthquakes by measuring and dating displaced sedimentary strata. However, paleoseismology has been also extended into the use of landslides, rock falls, and precariously perched boulders to determine the timing of past earthquake events.

A holistic knowledge of geomorphology is essential for seismic hazard mitigation. The effects of earthquakes on people, their animals, and property extend far beyond the shaking and collapse of structures, and can include landsliding, liquefaction, flooding, and tsunami. In particular, earthquake-triggered landslides can be particularly devastating, as was illustrated by the 2005 Kashmir and the 2008 Sichuan earthquakes that resulted in many thousands of fatalities. Earthquake-triggered landslides may also block drainages to create metastable dams behind which large lakes may form. These have the potential to fail catastrophically, as was the case with the Hattian Bala slide produced by the 2005 Kashmir earthquake in Pakistan (Figure 3g). The construction of a sluice through the Hattian Bala helped drain the lake and reduced the effects when the dam eventually burst. Careful geomorphic mapping can help identify slopes that might potentially fail. In addition, mapping may aid in identifying flood zone and run-up zones for tsunami, and areas of potential liquefaction. Such maps may become integral parts of geographical information systems that may be used by planners, policymakers, and relief agencies to help reduce the impact of tectonic hazards.

Tectonics, climate change and erosion, and polygenetic landscapes

Arguably the most exciting aspect of tectonic geomorphology is the growing realization of the polygenetic nature of landscapes, and the complex interactions between endogenetic and exogenetic processes. This is particularly well illustrated with the tectonic aneurysm model that was initially developed to help explain the exceptionally high topography and rapid exhumation rates of the Nanga Parbat–Haramosh syntaxis at the western end the Himalaya. This model argues for localized uplift caused by enhanced glacial and fluvial erosion and the exhumation of metamorphosed crustal rocks, weakening the lithosphere, and further enhancing bedrock uplift (Koons, Zeitler, and Hallet 2013). Key issues in the evolution of polygenetic landscapes include examining the role of tectonic and isostatic uplift, the nature of exhumation rates, climate control and the distribution and timing of erosion, the role of glaciers and relief production, focused erosion and spatial coincidence, geomorphometry and erosion modeling, and equilibrium concepts such as topographic steady state. In addition, human influences are now contributing to the development of polygenetic landscapes and
this factor poses greater challenges for predicting the nature of future geomorphic change.

In summary, tectonic geomorphology is an exciting and expanding new branch of geomorphology, which has added much to our understanding of the dynamics and interplay of Earth’s interior and surface processes. In particular, it is allowing the complex variability between regions at different scales, and over different timescales (particularly ranging from $10^0$–$10^6$ years) to be examined in great detail and more holistically. The broader impact of tectonic geomorphology is great, as it has the potential to aid in tectonic hazard assessment and mitigation. This is important as human population increases dramatically in tectonically active areas. In particular, the largest population on our planet live in areas that are likely to experience exceptionally large earthquakes and volcanic eruptions, notably, China, India, Indonesia, Japan, Philippines, and western United States.

SEE ALSO: Digital elevation model and digital surface model; Environmental hazards; Fluvial erosional processes and landforms; Geodesy; Geomorphic hazards; Geomorphic systems; Geomorphological mapping and geospatial technology; Glacial erosional processes and landforms; Hillslopes; Landscape; Mass movement processes and landforms; Models in geomorphology; Mountain geomorphology; Quantitative geomorphology; Radiometric dating/techniques

References


Temperate forest ecosystems

Shannon McCarragher
The University of Tennessee at Chattanooga, USA
Lesley S. Rigg
University of Calgary, Canada

Temperate forest ecosystems are found throughout the middle latitudes in both the Northern and Southern Hemispheres, although they occur much more extensively throughout the northern mid-latitudes than the southern. Temperate forests are perhaps the most exploited and degraded in the world, primarily as a result of unsustainable harvesting and development, clearing for agriculture and livestock, pollution, and introduced pathogens.

There are seven major mid-latitude regions in which the temperate forests of the world can be found: Europe, eastern North America, Asia (specifically North China, Korea, Japan, and far-eastern Russia), southern South America, southeastern Australia, Tasmania, and New Zealand. For the purpose of this entry, these regions will be broadly discussed by hemisphere: Northern Hemisphere and Southern Hemisphere.

Generally, the mid-latitude temperate forest communities can be classified into three broad categories:

- temperate deciduous forests, dominated by broad-leaved trees that lose their leaves seasonally and support a wide and complex variety of flora and fauna;
- temperate coniferous forests, composed of evergreen trees and a simple two-layer vegetation structure; and
- temperate mixed forests composed of both coniferous and deciduous trees.

Biomass is the total mass (usually dry weight) of the vegetation within a forest and is often used to estimate the amount of energy stored in the vegetation. The rate at which biomass is produced by the forest plants each year is known as net primary productivity. Generally, biomass and net primary productivity levels vary from location to location. North American temperate forests, for example, have high biomass levels that steadily increase until the community is about 200 years old and a net primary productivity that ranges from 500 to 1200 g/m²/year. By comparison, the net primary productivity of a rainforest is 1000–1500 g/m²/year, while deserts can be as low as 150 g/m²/year.

Temperate forests’ underlying bedrock, geology, and soils differ from region to region. The two most common soils found beneath the temperate forests of the world are Alfisols and Ultisols. Alfisols, which tend to be found in cooler regions, are grayish in color at the surface due to the loss of some minerals as water moves through the soil. They have a relatively thin topsoil layer, clay accumulations at depth, and generally have a high fertility. Ultisols, on the other hand, tend to develop in warmer regions on stable landscapes and have red or yellow hues from the iron oxides present within them. They are usually less fertile, since many of their primary minerals have been broken down and removed as a result of a process called weathering. One of the most important
aspects of soil is the nutrients which the plants use for growth. In temperate forests most of the plant nutrients (nitrogen, calcium, sodium, etc.) are contained in the leaves.

Northern Hemisphere

The Northern Hemisphere contains most of the world’s temperate forests. The following are common families of trees, listed alphabetically by genus, that make up these forests: maple (Acer spp.), birch (Betula spp.), beech (Fagus spp.), oak (Quercus spp.), willow (Salix spp.), and elm (Ulmus spp.). These broadleaf deciduous trees lose their leaves every fall and don’t resume growth until the following spring when increasing temperatures and day lengths trigger their leaf development (see Figure 1). Prior to this trigger, light levels are high on the forest floor. Spring ephemerals or short-lived plants that lack woody tissue in their stems (also known as herbaceous) take advantage of the high light levels by growing quickly and flowering in early spring before the leaves on the deciduous trees fully develop and shade the understory. Some conifers or evergreen trees, such as fir (Abies spp.), red cedar and juniper (Juniperus spp.), spruce (Picea spp.), pine (Pinus spp.), cypress (Taxodium spp.), and hemlock (Tsuga spp.) can also be present within the northern temperate forests.

Physical environment basics

The temperate forests in the Northern Hemisphere have four changing seasons (i.e., winter, spring, summer, and fall) and are the second rainiest biome after rainforests. The average yearly precipitation in temperate forests is 75–150 cm (30–60 in). Since precipitation is usually abundant due to high summer rainfall and large winter snowfall events, the vegetation is generally affected more by temperature than by precipitation.

The average temperature in temperate deciduous forests is 10°C (50°F). Typically, winter (November–April) in the north is characterized by cold weather, but conditions are strongly influenced by latitude and continental characteristics. Summers are mild, averaging about 21°C (70°F), while winter temperatures are often well below freezing. In North America, cold dry arctic air dominates the winter months.
TEMPERATE FOREST ECOSYSTEMS

and the average winter temperature can be as low as −10 °C (14 °F) or up to 10 °C (50 °F), depending on the location. Summer averages are usually warm and humid with July averages ranging from 20–23 °C (68–73 °F). In Europe, winter temperatures can range between 0.6 and 7 °C (33–45 °F) while summer temperatures tend to range from 11–20 °C (51–68 °F) and rarely exceed 30 °C (86 °F). Europe is wettest in the western regions and becomes progressively drier in the regions to the east that are located on the continent’s interior. The coastal zones of Asia from equatorial Southeast Asia to the subarctic eastern Siberia have sufficient rainfall and, as a result, have given rise to a sequence of five forest formations, two of which are temperate: cool temperate deciduous broad-leaved forest and warm temperate evergreen broad-leaved forest.

**Major communities**

In general, the deciduous forests in the Northern Hemisphere have five distinct layers. The first layer is the canopy trees, which range from 18 to 30 m (60–100 ft) tall. Maple (Acer spp.), elm (Ulmus spp.), and oak (Quercus spp.) trees are just some examples of those that can be found in the canopy. Beneath the canopy is a distinct subcanopy, composed of younger, small trees and saplings. The shrub layer sits below the small trees and saplings and includes mountain laurel (Kalmia latifolia), huckleberries (Gaylussacia spp. and Vaccinium spp.), dogwoods (Cornus spp.), viburnum (Viburnum spp.), and many others. The herbaceous layer occurs below all the others and is made up of small non-woody plants such as wildflowers, ferns, and grasses. Many short-lived herbaceous species, called spring ephemerals, have early leaf development that permits them to complete important stages of their lives prior to full tree canopy development; these include trilliums (Trillium spp.), trout lilies (Erythronium spp.), and mayapples (Podophyllum peltatum) (see Figure 2). Mosses and lichens often form a blanket ground layer on the forest floor, depending upon moisture and canopy cover.

In Europe, many of the forests have been highly disturbed or entirely cleared by human activity and agriculture production (see Figure 3). Most of the remaining forests in Europe can either be classified as mixed temperate forests, primarily found in Western Europe, or temperate evergreen woodlands, primarily located in the mountainous and Mediterranean regions of central and southern Europe. The European temperate forests are generally composed of fewer species when compared to other Northern Hemisphere forests, even though their densities (number of trees per unit area) and basal areas (area occupied by tree trunks and stems) are reported to be similar to those of old-growth forests in eastern North America.

Oaks, particularly pedunculate oak (Q. robur), dominate regions with lower elevation throughout Britain and northwest Europe. As elevation increases, beech (Fagus sylvatica), sessile oak (Q. petraea), and birch species (i.e., Betula pendula and B. pubescens) become dominant. A variety of oaks (i.e., Q. lusitanica, Q. pubescens, Q. trojana, and Q. cerris), chestnut, and beech (Fagus spp.) dominate the southern Mediterranean temperate forests of Europe. The dominance of oak throughout Europe is in part a result of centuries of planting monospecific oak stands. Other common tree species found throughout Europe include hornbeam (Carpinus betulus), field maple (Acer campestre), and elm (Ulmus procera), although many were damaged by Dutch elm disease (Ophiostoma ulmi). Often, an open shrubby understory can be found below the European temperate forest canopy.

The North American and Asian temperate deciduous forest communities are much more diverse than their European counterparts. Between 100 and 150 different deciduous tree
species compose the North American temperate forests, each with their own unique distribution. In the north, American beech (*Fagus grandifolia*) and maples (*Acer saccharum, A. saccharinum, A. rubrum, A. platanoides*) mixed with some conifers, specifically eastern hemlock (*Tsuga canadensis*) and others, are abundant in the forests, with conifers becoming more dominant as the community transitions into the immense boreal forests of Canada. In the south, oak (*Quercus alba, Q. rubra, Q. marilandica*), hickory (*Carya cordiformis, C. ovata*), maple (*Acer saccharum, A. saccharinum, A. nigrum, A. rubrum*) and American basswood (*Tilia americana*) dominate the forests. This general forest pattern is due to a north–south temperature gradient and further species differentiation occurring along an east–west moisture gradient, with more drought tolerant species dominating the western edges of the ecosystem. Deciduous trees become restricted to wetter drainages and open woodlands, eventually transitioning in the west into what was formerly a tall-grass prairie. On the northern and southern fringes of the community evergreen trees dominate with needle-leaf conifers in the north and broad-leaved evergreens in the south.

This distinct pattern of forest species assemblages has shifted through time. Forest responses to past climate change are well documented and significant. At the height of the most recent glaciations much of the area in North America, Europe, and Asia that is currently occupied by the boreal and mid-latitude forests was ice covered. As climates warmed, tree species migrated northward out of their southerly refuges (an area where populations of species can survive during unfavorable environmental conditions) and shifted into geographic locations and ranges currently associated with the mid-latitude forests. Studies of pollen distribution demonstrate that plant species in North America shifted range in response to natural climate warming over the past 18,000 years. Pollen records found in North America show that maple (*Acer* spp.) and birch (*Betula* spp.) species reached their current northerly limit approximately 6000 and 7000–10,000 years ago, respectively (Davis 1983). By 7000 years ago, within the Great Lakes region of North America, beech (*Fagus* spp.) had migrated into the oak/hickory (*Quercus* spp./*Carya* spp.) forests of lower Michigan, and hemlock (*Tsuga* spp.), beech (*Fagus* spp.), and white pine (*Pinus* spp.) formed extensive stands throughout southern Ontario. In Europe, the extent of the shift was more pronounced than in North America with only pockets of boreal and
Figure 3  An example of a highly disturbed and cleared area used for agricultural and grazing purposes in Europe (Ireland). Source: Photo by Shannon McCarragher.

deciduous species surviving in small populations within protected locations. Pollen studies similar to those conducted in North America suggest that the past 10,000 years of forest change in northeastern Sweden, showed a northward expansion of species (Quercus spp., Ulmus spp., and Tilia spp.) associated with warmer climates. Ultimately, this shift was mediated by a cooler period around 3200 years ago during which spruce (Picea spp.), pine (Pinus spp.), and birch (Betula spp.) established.

However, projections of human-induced climate change over the next several hundred years suggest that warming may be much faster than tree species have experienced in the past 18,000 years – in fact, probably over the past two million years. This rapid rate of climate change may severely impact many species’ ability to reproduce and persist in regions where they are currently found. Most significantly, numerous species may ultimately be unable to migrate to locations more suited to their species-specific ecological tolerances. Collectively, this problem may result in altered forest communities comprised largely of species with wide ecological tolerances. For example, sugar maple (Acer saccharum), an ecologically and economically important species that ranges across much of eastern North
TEMPERATE FOREST ECOSYSTEMS

America, will most likely be affected by altered climates. Increases in temperature could eventually lead to increased growth rates for adult sugar maple (Goldblum and Rigg 2005) and potential migration over longer time frames. In Scandinavia, East Asia, and North America, a northward movement of mid-latitude deciduous forest is modeled to occur at the expense of the southern boreal forest.

Adaptations and disturbances

Due to the marked cold winter season, nearly all plants in the Northern Hemisphere have a winter dormant season. For many shrubs and trees, the dormant season includes the loss of their leaves. During the shorter days and cooler weather of autumn, green chlorophyll in the leaves is reduced and begins to decompose, leading to discoloration that reveals brilliant oranges, yellows, pinks, and reds (variable by plant species) (see Figure 4). Actually, these colors were present in the leaves all year long, but had been hidden by the green pigment of the chlorophyll. Toward the end of the growing season, deciduous trees try to recover as many nutrients (such as nitrogen, phosphorus, and potassium) from leaves as they can before they drop them. The cycle of deciduousness relies upon the decomposition of the leaf litter deposited on the forest floor and the return of any remaining nutrients back to the soil for use by soil fauna (e.g., insects, earthworms, amphibians) and flora (e.g., trees, wildflowers, shrubs). The combination of retracted and stored nutrients allows for new leaf production each spring when the longer days and warmer weather signal to the trees that it is time to begin photosynthesizing again. Spring also triggers the growth and germination of various plants in the herbaceous layer.

If trees in the northern temperate forests did not undergo this deciduous cycle and instead kept their leaves throughout winter, the water in the leaves would freeze into ice, damaging the leaves and leaving the plant vulnerable to pathogens. In colder biomes, where species retain their leaves (needles) throughout the cold season, the plants are equipped with adaptations that prevent freezing. The short growing season within these colder biomes, however, inhibit the ability of nutrients to be returned to the soil via decomposition. Long-lived leaves are common in temperature extremes, both cold and hot, where tree species produce very high quality leaves. Trees in the deciduous forests produce comparatively low quality leaves that are easily decomposed on the forest floor when dropped. The herbaceous plants growing in the temperate forests do not simply lose their leaves; instead, they completely die back to the ground surface during the winter and regrow each spring.

Fire is not very common in North Hemisphere mid-latitude forests, except on the fringes (for
example, along the North American western and northern boundaries). Rather, the dominant forms of disturbance in the Northern Hemisphere are storms (Fischer, Marshall, and Camp 2013). Typhoons and hurricanes are common in both East Asia and northeastern North America, while windstorms occurring in the winter are dominant in central Europe. Tornados are also a destructive force within mid-latitude forests, especially in North America, although their damage tends to be confined to narrow strips of land that lie within their paths of movement. Ultimately, storm events tend to have a fairly isolated impact on the vegetation, leveling anywhere between a single individual to several hundred trees, but in doing so they create openings in the canopy, known as canopy gaps. The disturbance patches in central Europe tend to be the smallest when compared to East Asia and North America, respectively. Even though these gaps occur infrequently (1% per year), the increased light levels within them tend to play an important role in promoting germination and seedling development. The subcanopy, composed primarily of trees in varying stages of succession to the canopy, usually benefits the most from the increased sunlight within these gaps. However, many herbaceous species are also well adapted to take advantage of the conditions within canopy gaps. Ultimately, these gaps occupy about 10% of the forest area and support turnover within the forest community, whereby older trees are replaced by younger, newly established trees. As a result of this turnover, Northern Hemisphere temperate forests are generally composed of unevenly aged stands of trees, rather than the evenly aged stand of trees that may develop in regions where larger disturbances, like fire, cause an entire community to reestablish all at once.

Temperate forests are one of the most altered biomes on our planet, having been seriously affected by human settlement and clearing. While early human populations most likely played a key role in altering forest structure, it is the movements, more recently, of colonial and postcolonial Europeans that have had the greatest impact on these forest communities. In both Europe and North America, the northern boreal forest remains largely intact, while the mid-latitude forests have been extensively utilized for many centuries for farming and habitation. This is not to say that the boreal forest, especially at its southern margins, is not highly managed for timber and other activities, but in comparison, the history of human land use is more pronounced in the temperate forests.

These activities have led to the decline of this biome throughout most of the world. In the early eighteenth and nineteenth centuries, enormous tracts of forest land were cleared for agriculture, but the ground was rocky and soils were minimal. Following the opening of the American Midwest, nearly all the northeastern farmlands were abandoned and many of them have since returned to second growth forests. However, there are very few remnants of original forests remaining from which to piece together the original character of the forest. The American chestnut (Castanea dentata) was previously a canopy dominant tree in these forests until its virtual extinction in the 1930s due to a fungal disease (Cryphonectria parasitica). Evidence suggests that the chestnut blight fungus may have been accidentally carried in on Chinese chestnut trees being imported through New York. After entering the United States, the spores of the fungus were quickly carried by the wind and spread throughout the rest of the temperate deciduous forests. Today, American chestnuts exist only in small numbers as shrubs and rarely survive to maturity.

Pollution is a key factor within the temperate forest as exposure to subtle, but long-term,
TEMPERATE FOREST ECOSYSTEMS

pollutants such as acid deposition weakens trees leaving them susceptible to disease, insect outbreaks, and extremes in weather conditions. Researchers monitoring sugar maple forests in southeastern Canada and northeastern United States note changes in nutrient levels in the woody tissue of the trees (Jones 2006). They have linked these changes in wood chemistry to changes in environmental chemistry, specifically the progressive acidification of forest soils due to increased amounts of acid precipitation at higher latitudes. The long-term impacts on forested communities include forest decline due either to nutrient deficiencies or soil toxicity and dieback of species that are particularly sensitive to changes in soil environmental chemistry or species growing in marginal soils. Many commercially important tree species (e.g., maple, pine, and spruce) are experiencing decline, and in many areas the affected trees are the dominant tree species in their respective forest communities. Their decline not only results in changes to the relative distribution and dominance of those tree species within the forest community, it also impacts the ecological network of species that rely on them. For example, in some declining sugar maple stands, researchers have noted a decrease in the number of early season butterflies and moths, while others have reported increased stress in birds nesting in the canopies of declining sugar maple stands due to increased exposure.

The introduction of exotic or non-native plant and animal species to temperate forests can create a disturbance, as the new arrivals compete for food and habitat space, and can potentially displace and threaten native species. An introduced species becomes known as an invasive species when it is non-native to the ecosystem in which it has established, has the ability to become the dominant species within that community, and is causing, or is likely to cause, harm to human health, the environment, and/or the economy (Devine and Fei 2011). Ecosystem services, or the processes and natural cycles by which ecosystems fulfill and sustain humanity, are often also disrupted. Kudzu (Pueraria lobata), mile-a-minute or knotweed (Persicaria perfoliata), and Amur or bush honeysuckle (Lonicera maackii) are a few examples of exotic invasive plants that present an ecological threat to regions throughout North American temperate forests (see Figure 5). Invasive insect outbreaks are quite common in northern temperate forests and often result in an increased vulnerability or exposure to pathogens and diseases. The gypsy moth (Lymantria dispar) was accidentally introduced to the United States in the late 1860s and has since caused the defoliation (removal of leaves) of over 80 million acres of forested land. Dutch elm disease has negatively affected most elm trees growing within all Northern Hemisphere temperate forests and was spread by a variety of bark beetles that transmitted the spores of the fungus (Ophiostoma ulmi, formerly Ceratocystus ulmi) to healthy elm trees. Asian longhorned beetles (Anoplophora glabripennis), walnut twig beetles (Pityophthorus juglandis), and the emerald ash borer (Agrilus planipennis) are a few more examples of pests that pose serious threats to the hardwood trees found within northern temperate forests. It is important to note that not all exotic species are invasive; some actually prove beneficial to the environment and economy, such as cattle, potatoes, and honeybees. Multiple strategic management programs have been developed and initiated in order to detect, prevent, and control the harmful invasive biological disturbances.

Southern Hemisphere

The temperate forests of the Southern Hemisphere differ significantly from those in
the Northern Hemisphere and are not as geographically extensive. They are characterized by milder winters and are composed of evergreen broad-leaved forests, rather than deciduous broad-leaved forests found in the cooler Northern Hemisphere. Tree species in the Northern Hemisphere also exhibit a higher degree of frost hardiness than those in the Southern Hemisphere, but lack the fire tolerance that is necessary for life in the south.

**Physical environment basics**

Temperate forests in the Southern Hemisphere span climates from coastal maritime (South America and New Zealand) to drought prone (interior South America and Australia). The coastal maritime climate of the Southern Hemisphere is characterized by a mild temperature range and high annual precipitation (greatest in the winter). The summer drought becomes less severe as precipitation increases from north to

**Figure 5** An example of the invasive species Amur honeysuckle (*Lonicera maackii*) found in a northern temperate forest understory. The inset shows its distinguishing characteristic: the flower/fruit stems are shorter than or equal to the leaf stem lengths. Source: Photos by Shannon McCarragher.
TEMPERATE FOREST ECOSYSTEMS

In the south, average monthly precipitation can exceed 100 mm (4 in.) even during the driest months. The wettest conditions prevail on west-facing slopes, which catch onshore clouds.

In South America, the slopes of the Andean Cordillera receive more than 4000 mm (157 in.) of precipitation per year. Above about 800 m (half mile) on the Coastal and Andean Cordillera, winter precipitation falls as snow. When moist, warm air flows over a mountain range, it condenses to form clouds and precipitation on the windward slopes. After the precipitation falls, the air continues to flow over the mountain range and descends down the leeward side, becoming warm and dry. The warm, dry air produces what is called a rainshadow effect – significantly lower rainfall totals and higher air temperatures on the leeward side of the mountain range and beyond, as compared to the windward side. In South America, the Argentinean side of the Andes mountain range and the central depression beyond are examples of regions that have significantly lower rainfall totals as a result of the rainshadow effect. Average annual temperatures in this region range from 6–10°C (43–50°F).

In New Zealand, temperatures range between around 5°C (41°F) in the winter to around 14°C (57°F) in the summer. In some areas of New Zealand, annual precipitation levels can reach 9670 mm (381 in.) while the temperate forest regions of western Tasmania, Australia, are considerably drier and rarely receive annual precipitation levels over 4000 mm (157 in.).

Major communities

Plant communities on the Australian continent, over the geologic time scale, have largely evolved with tropical climatic regimes. Australia's vegetation has a very high proportion of endemic species, many of which can resist drought, fire, or both. Two genera dominate the tree communities: there are more than 700 species of *Eucalyptus* and over 900 species of *Acacia*. The broad-leaved temperate forests of Australia are some of the most distinctive in the world. They display a remarkable uniformity in that the principal tree genus, *Eucalyptus*, dominates. In no other region of the world does one genus so dominate the forests. They range from some of the tallest hardwoods in the world, mountain ash (*E. regnans* – 100 m; see Figure 6), to small woodland species found across Australia and form the treeline at elevation in the mountains of the southeast. In eastern Australia, the deciduous snow gum tree (*E. pauciflora*) defines the boundary between the subalpine forests and the treeless alpine zone.

Due to the aridity of Australia, large areas are without tree cover. The total forested area of Australia is 147 million ha, which represents approximately 19% of the continent. The greatest area of Eucalyptus forest occurs in eastern Australia in a nearly continuous band from the Tropic of Cancer in Queensland to South Australia and extending to Tasmania – an amazing 3000 km (1864 mi.) extent, north to south. River red gum (*E. camaldulensis*) occurs along river courses in the more central portions of Australia. The temperate eucalyptus forests in Victoria and Tasmania are composed primarily of *Eucalyptus* varieties, such as mountain ash (*E. regnans*), white gum (*E. viminalis*), and Australian oak (*E. oblique*), while tree ferns and small trees dominate the understory. In southwestern Australia, karri (*E. diversicolor*; see Figure 7), marri (*E. calophylla*), yellow tingle (*E. guilfoylei*), and red tingle (*E. jacksonii*) dominate. These forests have similar understories to those found in Victoria, mostly tree ferns, small trees, and shrubs. Conifers in these forests include Patagonian cypress (*Fitzroya cupressoides*), which grow in poorly drained lowland sites or in wetter mountain zones and can live to an age of 3000 years.
TEMPERATE FOREST ECOSYSTEMS

Figure 6  Mountain ash forest near Wallaby Creek in Victoria, Australia, with trees reaching over 90 m tall and a person for scale. An understory of tree ferns, small trees, and shrubs can also be seen. Source: Photo by Lesley S. Rigg.

The coastal temperate forests (in Australia, South America, and New Zealand) are dominated by the southern beeches (*Nothofagus* spp.). Myrtle beech (*Nothofagus cunninghamii*) is found in Tasmania and parts of southern Victoria. These Australian forests have a close affinity with other Southern Hemisphere forests of the same genus in Chile, New Zealand, and Argentina. Tasmania has the highest percentage of forest cover of any Australian state. In the Australian wet temperate forests other dominant species include the following conifers: king billy pine (*Athrotaxis selaginoides*), huon pine (*Lagarostrobus franklinii*), and celery-topped pine (*Phyllocladus asplenifolius*).

The temperate broad-leaved forests of New Zealand are characterized as podocarp-hardwood forests, meaning they are primarily composed of trees from the *Podocarpaceae* taxonomic family. Common species include matai (*Prumnopitys taxifolia*), rimu (*Dacrydium cupressinum*), kauri (*Agathis australis*), and kahikatea (*Dacrycarpus dacrydioides*). The forests have multistoried canopies with a podocarp dominant emergent (or tallest) layer that can reach 40–60 m (131–197 ft) high and a main canopy below that can reach 25–30 m (82–98 ft). Common subcanopy trees include kamahi (*Weinmannia racemosa*), rata (*Metrosideros robusta*), and mountain beech (*Nothofagus solandri*) in the wetter regions. In general, these forests are found at low altitudes toward the north of the country where rainfall is greater and soils are more fertile. The south island of New Zealand is primarily composed of beech (*Nothofagus* spp.) forests that tend to form in the subalpine areas in pure stands. Beech and podocarps are found coexisting in the warm and moist areas of New Zealand.

The temperate evergreen forests of South America occur primarily in Chile with minor extensions into neighboring Argentina. The temperate forests range from 37°S to about 46°S. The northern limit of this forest is determined by aridity while the cold temperatures determine the southern limit. This region of Chile varies topographically from west to east. The Cordillera mountain range (1000 m or 3280 ft) occur near the western coast, while the Andean Cordillera mountain range lies on the eastern portion of Chile and is twice as high as its western counterpart. Between these two mountain ranges lies a valley, known as the central depression, in which the elevation ranges from 100–300 m.
Three broad categories of plant communities exist within these ranges. Forests dominated by evergreens, such as Chilean laurel (*Laurelia sempervirens*), and tineo (*Weinmannia trichosperma*), occur in the central depression and low elevation forest of the coastal Cordillera. The deciduous roble beech (*Nothofagus obliqua*) is also found scattered throughout the central depression and low elevation forest, but the evergreen species are more common. As the altitude increases in the Andean Cordillera, coígües (*Nothofagus dombyri*) and Chilean tepas (*Laurelia philippiana*) become the dominant forest species. Epiphytes (plants that live on other plants) and lianas (woody vines) cover the tree trunks.

**Adaptations and disturbances**

Fires are the dominant disturbance in the southern temperate forests and they have been occurring since before the arrival of humans. The most common fires are surface fires that kill the ferns/grasses/shrubs, but merely burn the bark on the trees. Intense fires, called bushfires in Australia, kill the canopy tree species and are responsible for clearing massive areas of forest during the dry season. Many species in these forests are dependent on fire for regeneration and have therefore developed adaptations to survive them. Some trees have highly flammable bark and leaves that, in combination with the accumulation of leaf and woody litter, assist in their ability to burn. Other adaptations to fire include thick bark and the ability to resprout.
from lignotubers (a protected outgrowth at the base of the trunk containing vegetative buds). Stand-destroying fires of the Southern Hemisphere result in large, evenly aged stands of trees, in contrast to the unevenly aged stands in the temperate forests of the Northern Hemisphere.

Following fires in the temperate evergreen forests of South America, bamboo (Chusquea auila) will occupy the site for some years until trees reestablish. Beyond fires, South American forests are prone to earthquakes, which in turn trigger large mass movements (landslides) on the steep slopes that are found in the Andes. Landslides in South America are similar to fires in Australia as large even-aged stands regenerate after the disturbance. Evidence from disturbed forests where Nothofagus species are growing suggests that these trees require large-scale disturbance to regenerate. Agricultural practices in South America required open land, which in turn required the clearing and destruction of vast areas through Southern Hemisphere temperate deciduous forests. Nothofagus forests were hit particularly hard by clearing for agriculture, while other trees, particularly the raúlí (Nothofagus alpina) and roble beech (Nothofagus obliqua) were selected and cleared for use in charcoal production and ship building.

The Southern Hemisphere forests are prone to other disturbances, such as avalanches, cyclones, and introduced grazers (e.g., possums and deer). In New Zealand, the introduction of exotic animals to the island habitat by both Maori and Europeans (Australia) has wreaked havoc on the temperate forests. The tree species were adapted to grazing by native birds, including the flightless Moa whose extinction was linked to human arrival, but they were not adapted to grazing by arboreal possum or deer. Browsing by the introduced Australian brush-tailed possum (Trichosurus vulpecula) is generally accepted by most ecologists as the main sources of forest dieback, especially for forest dominated by rata (Metrosideros spp.) and kamahi (Weinmannia racemosa).

**SEE ALSO:** Biodiversity; Biogeography; Biomes; Climate change and biogeography; Dendroclimatology; Ecosystem; Ecosystem services; Global climate change; Soils of humid mid-latitude landscapes

**References**


**Further Reading**

TEMPERATE FOREST ECOSYSTEMS


Temperature measures the intensity of heat within the physical world and thereby determines, influences, or limits most geographic features and processes. As primary examples, temperature (1) determines the solid, liquid, or vapor state of water, (2) influences heating and cooling needs within the built environment, and (3) limits the geographical range and distribution of many plant and animal species. Temperature also is a critical variable in analyzing climatic variability and change.

While heat refers to the total reservoir of kinetic energy of a substance, temperature refers to the average kinetic energy of a substance. As a comparative example, the water in a small lake in northern Wisconsin during summer will have a higher temperature than that of the water in nearby Lake Superior, but Lake Superior has a far greater overall heat content.

In many cases, what is referred to as “temperature” is the air temperature, usually measured at a standard height of about 1.5 m above the ground. Many other measures of temperature are routinely made and these include (1) profiles of air temperature from the near-surface to the upper atmosphere (made with instruments on weather balloons known as radiosondes), (2) sea-surface temperature (measured either by satellite sensors or by “in situ” sampling), and (3) soil or deep-earth borehole temperature profiles. The focus here primarily will be on near-surface air temperature; however, many of the concepts can be extended to other measures of environmental temperature.

Controls on temperature

Within the natural environment, air temperature is most closely related to the availability of solar radiation. Therefore, temperature has strong relationships with variables that influence solar radiation, such as latitude, cloud cover, aerosols, and surface albedo. The annual and daily cycles of solar radiation availability create corresponding cycles of temperature, although the peaks of the annual and daily temperature cycles lag the solar radiation cycles, as temperature is more closely related to net radiation than solar radiation alone.

The initial heating or cooling of air temperature near the Earth’s surface occurs via conduction with the surface below. While incoming solar radiation (insolation) is the primary heat source for the atmosphere, reflection and long-wave radiation emission by the surface offsets solar input. These factors added together are known as net radiation. Net radiation indicates the overall amount of radiative heat gained or lost by the surface. A net radiation gain generally means warming and net radiation loss generally means cooling of the surface and, subsequently, of the atmosphere that is in contact with the surface.

The thermal climate of the Earth over recent millennia has varied only slightly, primarily because the amount of solar radiation reaching the Earth is relatively stable. In addition, the incoming and outgoing radiation from the Earth’s surface as a whole are nearly in balance.
Figure 1 Time–latitude plot of top–of–the–atmosphere (extraterrestrial) solar radiation (left) and annual mean of top–of–the–atmosphere solar radiation (right), both expressed as a percentage of the solar constant. The tropics receive relatively uniform solar input, while many higher latitude regions have high solar inputs during summer months only when sun angles are higher and day length is particularly long. At high latitudes, substantially longer path lengths of radiation through the atmosphere (not shown) further reduce incoming solar radiation.

However, large variations in solar radiation occur daily and seasonally because of the geometry of the Earth and its orbital characteristics within the solar system. These characteristics – such as the Earth’s orbit around the sun, axial tilt, and rotation, along with latitude – govern the angle at which solar radiation enters the Earth–atmosphere system. When totaled over the course of a day and expressed as a percentage of the solar constant, insolation at the top of the atmosphere at the poles varies from near zero during the winter solstices. Contrastingly, tropical areas receive nearly 30% of the solar constant at nearly all times of the year, resulting in year–round homogeneous temperatures (Figure 1).

By governing the amount of absorbed and reflected solar radiation and the amount of absorbed and emitted longwave radiation, characteristics of the Earth’s surface and atmosphere exert a strong influence on temperature. The greatest influence of the surface on net radiation is albedo. Snow is the most effective natural reflector of sunlight, whereas most land surfaces
and open water are efficient absorbers. Even small differences in land-surface albedo, however, can be great enough to cause large spatial variations in temperature during sunny days.

Emissivity is another important surface property because of its influence on longwave radiation. The amount of longwave radiation emitted by a surface \( E \) is dependent on both its temperature \( T \), in Kelvin, and its emissivity \( \varepsilon \) and is given according to the Stefan–Boltzmann law as

\[
E = \varepsilon \sigma T^4,
\]

where \( \sigma \) is a constant \( 5.67 \times 10^{-8} \) W m\(^{-2}\) K\(^{-4}\). Emissivity varies from 0 to 1, with natural surfaces generally having higher values than man-made surfaces (see Oke 1987, table 1.1). The effects of emissivity are at work 24 hours a day, whereas albedo is only a factor when sunlight is present. As a result, emission of longwave radiation plays a particularly important role in cooling processes at night but is overshadowed by the effects of solar radiation during the day. Higher nighttime temperatures often occur in areas with lower emissivity than adjacent areas where it is higher. Snow has a very high emissivity, which, combined with its high albedo and ability to insulate the atmosphere from ground heat below, makes it a very efficient cooling device in the climate system. Inversion of the Stefan–Boltzmann law to solve for temperature,

\[
T = \left[ \frac{E}{\varepsilon \sigma} \right]^{1/4},
\]

is used to estimate land- and sea-surface temperature from satellite- or aircraft-based sensors that measure longwave (thermal infrared) emission.

Like the Earth’s surface, the atmosphere has a variable albedo and emissivity. Clouds and other atmospheric constituents, such as small atmospheric particles or liquid droplets known as “aerosols,” typically increase albedo and act as cooling processes. Conversely, greenhouse gases – such as water vapor, carbon dioxide, and methane – affect the absorptivity and emissivity of the atmosphere and act as warming processes. Even though greenhouse gas concentrations are relatively low, they have important direct radiative effects that are compounded by positive feedback mechanisms – processes that have acted to make Earth substantially warmer and more habitable than it would be without greenhouse gases. As a result, changing concentrations of greenhouse gases and atmospheric aerosols are very influential in regional and global climate change.

Water vapor is the most influential of the greenhouse gases in the atmosphere because of its high concentration compared to the others. Water vapor that is converted to liquid or solid form in the form of clouds, however, typically increases atmospheric albedo. The net effect of clouds on near-surface temperature is dependent on their optical thickness and their height. For instance, high cirrus clouds tend to lead to higher temperatures over clear-sky conditions because of their transparency to solar radiation and ability to absorb longwave radiation and re-emit it to the surface. Low clouds tend to be less transparent and have a higher albedo, leading to lower temperatures compared to clear-sky conditions. In addition, natural and anthropogenic aerosols and particles can linger for hours to days and lead to cooling. Volcanic eruptions can result in large-scale atmospheric albedo increases and temporary global cooling. The most influential of these volcanoes inject gases and particles into the stratosphere where they persist for months to several years.

The physical geography of the Earth’s surface affects temperature spatially through topography and the distribution and configuration of land and water bodies. An important property of water is its role in dampening temperature variation. This is because of its high specific heat, which is about five times that of most land surfaces. Water absorbs energy when there is a surplus of net radiation and re-releases it when there is a deficit of net radiation, acting to
DECREASE THE RATE OF TEMPERATURE CHANGE RELATIVE TO SURFACES WITH LOWER SPECIFIC HEATS.

WATER ALSO PLAYS A KEY ROLE IN MODULATING TEMPERATURE VIA LATENT HEAT EXCHANGE. WHEN WATER CHANGES PHASE, AN EXCHANGE OF ENERGY IS INVOLVED THAT CAN OFFSET BACKGROUND TEMPERATURE TENDENCIES. PHASE CHANGES FROM A LOWER ENERGETIC STATE TO A HIGHER ONE – SUCH AS MELTING OR EVAPORATING – CONSUME LATENT HEAT ENERGY SO THE WARMING THAT LEADS TO THE PHASE CHANGE IS OFFSET. EVAPORATIVE COOLING IS A GOOD EXAMPLE: FORESTS OR OTHER VEGETATED AREAS OFTEN HAVE A LOWER TEMPERATURE THAN ADJACENT NONVEGETATED AREAS BECAUSE THE TRANSPERSION PROCESS LOWERS THE AIR TEMPERATURE AS WATER IS EVAPORATED. CONVERSELY, PHASE CHANGES FROM A HIGHER ENERGETIC STATE TO A LOWER ONE – SUCH AS CONDENSING OR FREEZING – RELEASE LATENT HEAT ENERGY SO THAT THE COOLING IS OFFSET. CITRUS FRUIT GROWERS MAKE USE OF THIS PROCESS BY SPRAYING LIQUID WATER WHEN TEMPERATURES APPROACH THE FREEZING POINT. THE FREEZING PROCESS ADDS LATENT HEAT TO THE ATMOSPHERE AND RESULTS IN A LOCAL WARMING OF THE AIR THAT PROTECTS THE FRUIT CROPS. THROUGH THESE LATENT HEAT PROCESSES, THERE IS NO RESULTANT CHANGE IN TOTAL ENERGY IN THE SYSTEM, DESPITE A CHANGE IN TEMPERATURE. AS A RESULT OF THESE PROPERTIES OF WATER, LOCATIONS THAT ARE NEAR OR DOWNWIND OF WATER BODIES, OR THOSE WHERE ATMOSPHERIC HUMIDITY IS HIGH FROM EVAPORATION OR TRANSPERSION FROM VEGETATION, HAVE LOWER ANNUAL AND DIURNAL VARIATIONS IN TEMPERATURE. SEMIARID AND ARID REGIONS HAVE A VERY HIGH DIURNAL TEMPERATURE RANGE BECAUSE OF THE LACK OF WATER VAPOR, VEGETATION, AND TRANSPERSION.

DIFFERENTIAL HEATING OF THE EARTH’S SURFACE AND THE RESULTING TEMPERATURE CHANGES GIVE RISE TO WIND AND AIR CIRCULATION FROM LOCAL TO GLOBAL SCALES. COLD AIR IS DENSER THAN WARMER AIR, WHICH MEANS COLD AIR TENDS TO DISPLACE WARM AIR VERTICALLY WHEN THEY ARE ADJACENT. AT THE GLOBAL SCALE, TRANSPORT OF COLDER AIR FROM THE POLES EQUATORWARD AND WARMER AIR IN THE TROPICS POLEWARD SERVES TO ALTER TEMPERATURE FROM THAT WHICH NET RADIATION ALONE WOULD PRODUCE AT THE SURFACE. ATMOSPHERIC CIRCULATION ALSO INFLUENCES AND INTERACTS WITH OCEAN CURRENTS, WHICH EXCHANGE LARGE AMOUNTS OF HEAT LATITUardinALLY. IN ADDITION TO THE LATITUDINAL GRADIENTS OF OCEAN TEMPERATURE, THE ORIENTATION OF THE OCEAN BASINS AND CONTINENTS HELPS TO DIRECT OCEAN CURRENTS AND PRODUCE THE PATTERNS THAT WE SEE TODAY. FOR INSTANCE, THE GULF STREAM TRANSPORTS WARM WATER FROM THE TROPICS TO WESTERN EUROPE, HELPING TO PRODUCE A CLIMATE THAT IS MUCH WARMER THAN AREAS AT SIMILAR LATITUDES IN EASTERN CANADA OR EASTERN RUSSIA WHERE COLDER OCEAN CURRENTS EXIST (FIGURE 2).

TOPOGRAPHY HAS A STRONG INFLUENCE ON TEMPERATURE BECAUSE OF THE EFFECTS OF ALTITUDE ON PRESSURE (AND THEREFORE TEMPERATURE), BUT SURFACE FEATURE SHAPES AND SIZES ALSO PLAY A ROLE. THE SLOPE AND ASPECT OF A SURFACE CAN ALTER THE INCIDENCE ANGLE OF SOLAR RADIATION, SUCH THAT SLOPES THAT FACE AWAY FROM THE SUN ARE COOLER THAN SLOPES FACING TOWARD THE SUN. THIS EFFECT CAN BE SEEN IN MID-LATITUDE REGIONS WHERE STEEP SLOPES FACING EQUATORWARD ARE WARMER AND OFTEN HAVE VEGETATION INDICATIVE OF A DRIER AND WARMER CLIMATE THAN SLOPES FACING POLEWARD, PRIMARILY BECAUSE EVAPORATION IS DIRECTLY RELATED TO TEMPERATURE.

AT HIGHER ELEVATIONS, THE MEAN TEMPERATURE IS COOLER BECAUSE ATMOSPHERIC PRESSURE IS LOWER, AND THEREFORE THERE IS LESS MOLECULAR KINETIC ENERGY AT THESE ALTITUDES. HOWEVER, BECAUSE THE ATMOSPHERE IS THINNER, GREATER INSOLATION OCCURS AT HIGH ELEVATIONS DURING THE DAY AS THERE IS LESS ATMOSPHERE ABOVE TO REFLECT OR SCATTER INSOLATION. ON THE OTHER HAND, THERE IS A GREATER LOSS OF LONGWAVE RADIATION AT NIGHT BECAUSE THE THINNER ATMOSPHERE DOES NOT ABSORB AND RETURN AS MUCH OUTGOING LONGWAVE RADIATION. THESE PROCESSES OFTEN RESULT IN A GREATER DIURNAL TEMPERATURE RANGE AT HIGH ELEVATIONS. FACTORS LIKE PERSISTENT CLOUDS THAT LIMIT NET RADIATION OR WINDS IN EXPOSED MOUNTAIN
slopes that enhance atmospheric mixing, however, can reduce the diurnal temperature range.

Topography and elevation can work together to increase the diurnal temperature range. Valleys and enclosed basins tend to collect cool air at night, since it is denser than warmer air and thus drains to topographic lows. During the day, surface heating usually mixes this chilled air and then is further warmed from positive net radiation. These conditions require clear skies
and dry air for a greater temperature variation with the presence of wind, clouds, or high water vapor concentrations dampening the effect. Valley environments can experience large seasonal temperature differences from surrounding areas, especially where winter snow or enhanced cloud cover occurs. The effects of snow and cloud on surface and atmospheric albedo give rise to the extreme annual temperature variation seen in nonglaciated cold environments, such as the interior regions of Asia and North America, where atmospheric humidity is normally low.

Global patterns of air temperature

Global-scale maps of monthly mean values of near-surface air temperature clearly show the influence of latitude, topography, and proximity to water bodies discussed above (Figure 2). In response to solar radiation variations, isotherms (lines of equal temperature) tend to be predominantly east–west. Major mountain ranges such as the Andes in South America, the Himalayas in Asia, and the Rockies in North America are very influential in global-scale patterns of temperature. Beyond the expectation of lower temperatures at high latitudes, the high elevations of the ice sheets on Greenland and Antarctica act to further reduce air temperature substantially. Some of the largest spatial gradients of air temperature occur in North America and eastern Asia during January – primarily because solar radiation gradients are largest in the winter hemisphere and both of these continents have vast interiors that are distant from oceanic sources of heat and water vapor. North America and eastern Asia also have substantial snow cover at higher latitudes, which further cools northern regions relative to southern ones. A quantitative example from North America shows that the rate of change of air temperature with latitude during January is about 1.0 to 1.5°C per degree of latitude, whereas July air temperatures decrease only 0.4 to 0.6°C per degree of latitude (Figure 3). These latitudinal rates of change of air temperature are the cause of variability in weather systems across the continent, but also are influenced by the same weather patterns, illustrating the nonlinear nature of the climate system.

Temperature and climate change

Changes and variability in near-surface air atmospheric temperature are among the most important and tangible indicators of climate change. Reliable liquid-in-glass thermometry developed in the mid-eighteenth century produced relatively consistent instrumentation during the historical record. High-quality data from microwave-based satellite sensing of temperature are available from about 1979 onward (Hurrell and Trenberth 1996). As temperatures have a wide range of variability, the analysis of temperature changes with time is simplified by averaging it over space and time. Global and hemispheric-scale averages of monthly or annual averages of temperature are typically used to monitor the energetic state of the near-surface environment. Often, these large-scale spatial averages are integrations of air temperature over land and sea–surface temperature over the oceans. Latitudinal and topographic influences on temperature are usually considered to be independent of climate change. As a result, to reduce the spatial variability associated with latitude and topography, air-temperature anomalies are created by removing a local mean from an air-temperature time series. Converting air temperatures to anomalies produces a smoother spatial field while preserving the temporal variability of each location. The temporal variability of nearby locations that are at very different elevations,
Figure 3  Latitudinal transects of monthly mean air temperatures during January and July at 120°W and 90°W in North America (same data as Figure 2). The gradient of air temperature with latitude (slope) in January is 2–3 times that of July. In both months, the more westerly transect (120°W) shows a weaker latitudinal gradient due to its closer proximity to the Pacific Ocean. Data from Willmott and Matsuura (2009).

The data, for instance, becomes much more comparable after conversion to anomalies (Figure 4). Once temperatures are converted to anomalies, they can be interpolated to a regular grid and averaged accordingly. Analysis of global-scale air-temperature anomalies show that global
Figure 4  Time series of January mean air temperature (°C) at two locations in Colorado, USA, from 1903 to 2012 (data from Historical Climate Network, National Climatic Data Center). Boulder is located at an elevation of 1672 m while Dillon is at 2763 m. The two locations have January mean air temperatures that differ by approximately 9°C, but removing the long-term January mean produces anomaly time series that have similar interannual variability. As a result, air-temperature anomalies are used for comparative analysis and to estimate climatic change across regions that have wide ranges in elevation or latitude. Data from Historical Climate Network, National Climatic Data Center.

and hemispheric rates of recent warming range from 0.05 to 0.25°C/decade (Figure 5; Morice et al. 2012). Concentrations of greenhouse gases are increasing and they are mixed well in the atmosphere, but changes in temperature in the historical record have been larger during nights and winters, especially at high latitude where net radiation often is negative (Meehl et al. 2009; Robeson 2004; Seneviratne et al. 2012).

Human alterations of land cover also influence temperature, primarily by changing surface albedo, emissivity, and the availability of surface water for evaporation and transpiration (Pielke et al. 2011). For instance, many urban heat islands arise from the introduction of materials and urban geometry that result in increased absorption of solar radiation, decreased longwave radiation loss, and decreased evapotranspiration relative to natural landscapes. These processes, along with anthropogenic waste heat, frequently combine to increase air temperature over cities, with the effect usually being larger at night (Oke
Figure 5  Time series of globally and hemispherically averaged temperature anomalies from 1881 to 2012 (HadCRUT4 data from Hadley Centre at UK Met Office and Climatic Research Unit at University of East Anglia). Both hemispheres have similar interannual variability, but the Northern Hemisphere has warmed slightly more rapidly in recent decades.

1987). Other larger-scale anthropogenic changes to the land surface include deforestation and agricultural activities. The direction in which temperature is forced is dependent on what landscape was altered because albedo, emissivity, and evapotranspiration can be either increased or decreased. For instance, irrigated farmland in a desert environment can lower the temperature through increased evapotranspiration. Deforestation can increase daytime temperature and decrease nighttime temperature through decrease in evapotranspiration and subsequent water vapor concentrations. Overgrazing removes vegetation and exposes soil, which reduces transpiration and water vapor while also lowering emissivity, with the net effect being higher daytime and nighttime temperatures.

SEE ALSO: Agroclimatology; Arid climates and desertification; Atmospheric aerosols; Atmospheric/general circulation; Climate change, concept of; Climate and societal impacts; Climatology; Earth’s energy balance; Global climate change; Global dimming/brightening; Hydrologic cycle; Lake climates; Land-use/cover change and climate; Microclimatology; Mountain climatology; Oceans and climate; Paleoclimatology; Polar climates; Snow cover changes; Urban climatology; Water and climate change
TEMPERATURE

References


Further reading


Meanings are not fixed but are historically, geographically, and socially specific. This simple, but insightful lesson of the postmodern and poststructural turn seems quite appropriate when looking at the notion of “territory.” In its original Latin use, *territorium* referred to the agricultural and grazing land surrounding a human settlement. With the rise of the modern state in Europe, around the fifteenth century, the term has increasingly been associated with the state. Among Anglo-American geographers, territory is today often conceptualized as a bounded space over which a form of political authority is exercised, although it can also be used as a generic synonym of place and region. A similar multiple usage can also be found in French and Italian, but within these geographical traditions territory tends to be treated as a lived space more than as a space monopolized by a politico-institutional power. And yet, if we move out of the Western world, we will see how territory cannot be easily translated into other languages. “Territory” is therefore a polysemic term and the fact that it is a key notion in several disciplines (geography, law, ethnology, ecology, anthropology, etc.) makes the task of pinning down its meaning even more dependent on the context within which it is used. For the purpose of this entry, territory will therefore be analyzed within a limited disciplinary and linguistic scope. Preference will be given to the geographical literature in English and French. This choice might sound too confining, but within both the anglophone and francophone geographical literature territory and territoriality have been extensively researched, producing the prevailing interpretations that are more widely used today within the international scientific community.

**Early steps on territory**

In his historical account of the political and legal institutions of the Western world, Carl Schmitt (2003, 74), echoing the linguist Jost Trier, affirmed: “In the beginning was the fence. Fence, enclosure, and border are deeply interwoven in the world formed by men, determining its concepts.” *Nomos*, the legal principle around which, according to Schmitt, the first political institutions were created, is a spatial term which recalls indeed the notion of fence. *Nomos* stands for the primeval form of land appropriation and delimitation – the first partition and classification of space, which gives orientation (*Ortung*) and order (*Ordnung*) to space.

This interpretation is particularly relevant in considering territory, which today is still mainly defined as a portion of the Earth delimited by borders, over and through which some form of power is exercised. In one of the first geographical reflections on the notion of territory, Jean Gottmann (1973) implicitly echoed Schmitt’s definition of territory as the fruit of partitioning and organization through political processes. In the interpretation of this cosmopolitan geographer, born into a Jewish family in Ukraine, raised in Paris, and active in the United States, France, and England, territory
TERRITORY AND TERRITORIALITY

is the basic unit in the political organization of space that defines the relationships between a group of people and their habitat. Gottmann treats this relationship in psychological terms, as a “psychological reflex” that characterizes all people, thus suggesting, somewhat uncritically, that territorialization is a sort of inescapable condition of humanity. Further elaborating on this point, Gottmann talks of territory as a psychosomatic device – a rather obscure notion that, in the intention of its author, should capture people’s psychological oscillation between a search for security and a search for opportunities. These are indeed the two functions with which Gottmann believes territory is endowed: territory serves as both a shelter and a springboard of opportunity. While the former clearly resonates with forms of control and exclusion, the latter opens up terrain for collaborations and relationships across boundaries. Territory in this latter sense is a space associated with a socioeconomic project, namely the pursuit of the “good life” of people, their happiness, and their wellbeing. In Gottmann, territory is therefore not only a bounded space imposed by a sovereign power on a portion of the Earth that constrains and encages people. True, for Gottmann there can be no territory without sovereignty, as the two are intimately linked; however, territory is something more than a state space; it is a peopled space. Territory comes into existence first and foremost thanks to and for people. It is less a top-down than a bottom-up construct, as it aims to satisfy the needs of those who inhabit it. This is an important point which has gone astry in later academic production, often concerned with the oppressive character of territory. Equally important is the fact that, although territory signals a unique relation between a group of people and the land they inhabit, this relationship is not necessarily cast in identity terms. Rather, what is generated is a shared modus vivendi, which, along with shared interests, works as a common denominator among fairly diversified peoples – an insight equally lost in the successive treatment of territory as mere national homeland.

Another important early step on the notion of territory was Edward Soja’s (1971) essay on the political organization of space. For the American geographer, this organization revolves around three main functions: (i) control over distribution, allocation, and ownership of resources; (ii) maintenance of order and enforcement of authority; and (iii) legitimization of authority through societal integration. Human or social territoriality is a primary basis for the political organization of space. Territoriality provides an essential link between society and space, in terms of identity, concentration of activities, and exclusion of undesirable outsiders. Although Soja believes that men (sic) are territorial animals and that territoriality affects human behavior at all scales of social activity, he also cautions against making direct analogies between animal and human territoriality. He argues that human territoriality is a cultural and behavioral phenomenon that varies historically and geographically in its structure and functions. Similarly, the boundaries that define territory as a demarcated space should also be evaluated in historical perspective, as they might not necessarily coincide with precise lines on the Earth’s surface.

Both Gottmann and Soja stand as isolated contributions within an academic production that most often treated territory and territoriality as self-explanatory terms. Their emphasis on territory as organizational space made by and for people did not find a large following, as scholars largely looked at territory as the product of the state, often at the expense of people, and as a defining principle of nationhood.
Human territoriality

In the Anglo-American geographical literature, territory has long been treated as a juridico-political concept associated with sovereignty—that is, a key component of the modern state. Not surprisingly, it has been monopolized by political geographers. According to Kevin Cox (2002), territory and territoriality actually should be regarded as the defining concepts of political geography. Ironically, though, the most compelling reflection on these two terms came in the 1980s from someone who was not a political geographer, Robert D. Sack (1986).

Before his seminal work on human territoriality opened up the notion of territory to a plurality of scales and actors, both terms referred mainly to the state.

For Sack (1986, 1), human territoriality can be best understood “as a spatial strategy to affect, influence or control resources and people, by controlling an area.” This means that human territoriality cannot be considered as a natural instinct. Since the 1960s, the majority of social scientists had approached territoriality with the lenses of ethology. Following the insights of Konrad Lorenz, territoriality was regarded as an innate feature of all living species, humans included. Also, Soja – whose definition of territoriality as a behavioral phenomenon associated with the organization of space into demarcated territories clearly anticipated Sack’s definition – felt obliged in his essay (Soja 1971) to offer a lengthy review of the literature on animal territoriality. Within this scholarship, the most cited work was *The Territorial Imperative* by the American anthropologist Robert Ardrey (1967). For Ardrey, human beings are driven by an inherent need to occupy, control, and defend the space which they inhabit—a territorial imperative that would inevitably lead to aggressive and conflictual relations between groups. Similar considerations were also present in another influential book of the time, *The Hidden Dimension*, by the American anthropologist, Edward T. Hall (1966), who introduced the notion of proxemics, which was to prove very popular among geographers.

Moving from a behaviorist approach, Sack argues that territoriality should be regarded neither as an instinct nor as an essentially aggressive move. Territoriality is instead a spatial strategy, and therefore it can be turned on and off according to the circumstances and the interests of the given actor. In other words, the same space can be a territory, that is, a bounded controllable place, at one time and not at another time. Let’s imagine, for instance, an open green space in the outskirt of a city. It might be assumed that usually this space can host a variety of leisure activities: lying, strolling, playing games, and so on. The day a ticketed event is organized, this space might become a territory: a fence is built to delimit the event area and the flow into this area is regulated by a given authority. In Sack’s (1986, 19) words, territoriality is “the attempt by an individual or group to affect, influence, or control people, phenomena, and relationships, by delimiting and asserting control over a geographic area. This area will be called the territory.” The importance attributed to both scale and boundaries is clear in this definition. Contrary to that strand of political geography, which has long associated territory with the scale of the state, Sack introduces a plurality of scales at which territory can be conceived. Potentially, any space can become a territory: from the microspaces of the home (e.g., a given room declared off-limits to children by their parents) to large supranational spaces (e.g., the Schengen Area). Similarly, the actors involved are not only state actors, but any individual or group that can produce and enforce a territorializing strategy. This involves the demarcation of
clear boundaries, which, once trespassed, might lead to various consequences to the trespasser, including their physical removal or suppression. Territory is the result of bounding strategies, and its existence is dependent on the establishment, maintenance, and enforcement of boundaries. These boundaries are a form of social communication, as they are indeed signs that give meaning to place, mediating the interaction between individuals, groups, and organizations. According to Sack, territoriality is a form of spatial behavior associated with some universal traits, as they occurred throughout history and across most of the world, and with some specific traits as well, which are distinctive to historical periods and geographical contexts. In this sense, territoriality and territories are not given, but they are the product of people’s conscious agency – what Sack calls the social construction of territoriality. This idea was developed further in the seminal studies of Anssi Paasi (1996), who detailed the social construction of territories around four interrelated processes: the physical shaping of a territory, its symbolic constitution, its institutionalization, and its final establishment.

**De- or reterritorialization**

With the emergence of poststructuralism as a major paradigm in geography in the 1980s and 1990s, territory underwent a severe criticism. The fact of being defined by a physical boundary was clearly at odds with the new understanding of place as an open, unbounded, relational space. Territory was too structural a notion in an epoch that started privileging ideas of hybridity, porosity, and blurriness. While place was regarded as progressive and empowering, territory was dismissed as a site of oppression, exclusion, and confinement. The inside/outside demarcated by a territorial boundary came to be questioned by a growing numbers of scholars. International relations (IR) theorists in particular challenged the analytical purchase of a modern, exclusive form of territoriality. Going beyond this Westphalian model of state territoriality, they suggested the notion of “unbundled territoriality” as a more insightful way to explore the condition of postmodernity in international politics (Ruggie 1993). Associated with this notion was the idea of the emergence of a plurality of overlapping and competing authorities and jurisdictions, giving way to a sort of new medievalism. This post-Westphalian or postmodern territoriality was seen to best capture governing practices adopted within the European Union (EU), which was heralded as the first postmodern or postnational polity, articulated around a form of nonhierarchical, multilevel governance, which defied any separation between domestic and international politics.

In geography, the most sustained critique of modern territoriality came from John Agnew (1994), whose “territorial trap” exposed the ahistorical and decontextualized view of a world reified into fixed units of sovereign space – a view also labeled “methodological nationalism.” Within this conception, the territorial state was thought to exist prior to and as a container of society. To be accurate, Agnew’s criticism was not targeting territory per se but the conceptual isomorphism between state, sovereignty, and territory. One of Agnew’s later major arguments would indeed be the idea that effective sovereignty is not necessarily predicated on and defined by the strict and fixed territorial boundaries of the state.

In the 1990s the excitement over the end of the Cold War and the transformations broadly associated with globalization produced new narratives calling for “the end of history,” “the end of geography,” and “the end of territories.” Transnationalism and cosmopolitanism gained
momentum, also spurred by the perception that the economic, political, and ecological risks societies were facing were too large to be effectively addressed by individual states. Responding to this deterritorializing frenzy, geographers observed, echoing Deleuze and Guattari, that the claim that territories were disappearing was wrongly placed, and that deterritorialization was necessarily associated with reterritorialization, as these were two sides of an ongoing territorialization process. Thus, on the one hand, geographers, among other scholars, acknowledged the “hollowing out” of the state caused by processes of denationalization, destatization, and internationalization. Yet, on the other hand, they mapped the emergence of new territories at the supra- and subnational scales, paying particular attention to (world) cities and regions. In this sense, globalization entailed not the implosion of territories but their explosion, with the territory of the state now being only one of many competing territories.

**Territoire**: a francophone perspective

While anglophone geography has clearly privileged a statist understanding of territory, francophone geography has followed a different path. Loosely informed by the Vidalian tradition of regional geography, with its focus on the unique relationship between a group of people and the space they inhabit, *territoire* (the French term for territory) has emerged mainly as a social and symbolic concept. True, the legal-political connotation – territory as the space divided, possessed, and controlled by a political authority – has also been present in francophone geography. But some of its most renowned representatives (e.g., Claude Raffestin, Joël Bonnemaison, Bernard Debarbieux, Jacques Lévy, Guy Di Méo, and Jean-Luc Piveteau) have generally preferred exploring notions of *territoire* and *territorialité* beyond their connection to the state. Overall, *territoire* has been theorized as a cultural space of belonging, imbued with geo symbols and individual and collective memories: a lived space (*espace vécu*), wherein people are linked to their habitat in an ecological relationship. Within this perspective, *territoire* is defined not so much by a principle of material appropriation, as it is in anglophone geography, but by a principle of cultural identification. It has to do with being more than having, since it is a central concept in processes of identity formation. As a *milieu de vie* (living environment), *territoire* allows people to give sense to their lives and to make sense of what surrounds them. *Territoire* is permeated by ethical, spiritual, symbolic, and affective values, and at times it is conceptualized, as it is in Gottmann (1973), as the product of a presocial, psychological human need for rootedness. This ecological and cultural territory is generally kept separated from the political territory, and in some instances it is also said to precede it. The focus, rather than being on the production of territory by political institutions, is on the everyday production of territories (*territoires du quotidien*), in a move which takes territory away from political geography and places it at the center of social and cultural geography.

It is clear that the francophone conception of territory comes close to how place has been conceptualized in anglophone humanistic geography. Not surprisingly, both share the same Heideggerian perspective on the human subject’s mode of being, which is always being-in-the-world or being-in-place (*Dasein*); however, while anglophone geography has in time moved away from an essentializing link between culture and place, also introducing issues of power, economy, and politics in the making of place, a great deal of francophone geography
TERRITORY AND TERRITORIALITY

still seems to approach both the notion of identity and the relationship between identity and territory in a rather unproblematic way.

One of the most interesting departures of francophone geography in relation to mainstream anglophone geography is its understanding of the notion of territoriality. Instead of a strategy to control people and goods by controlling an area, *teritorialité* is a relational and semantic system: it is aimed not at keeping people and things off a space, but at making space meaningful. This view first originated in the work of Claude Raffestin (1980), a Franco-Swiss geographer who, earlier and more than others, offered a comprehensive study of *territoire* and *teritorialité*. Combining Lefebvre’s insights into the production of space with Foucault’s notion of power relations, Raffestin conceives of *territoire* as the result of the projection of labor – understood as a combination of energy and information – by an individual or collective actor onto a given space. For Raffestin, *territoire* is constructed out of a pre-existing space on the basis of an actor’s program (a set of intentions and objectives) which signifies that space. Largely influenced by semiotics, Raffestin pays particular attention to the system of signs (“semiosphere,” to use Yuri Lotman’s term) which is available for any given culture and from which the actor draws the informational resources necessary for action. Thus, *territoire* is first and foremost a semiotized space, which can manifest itself at a plurality of scales. This emphasis on the system of signs also explains why Raffestin overlooks the notion of physical boundaries. While these are essential in a political-institutional understanding of territory, they are treated by Raffestin instead as a mere subset of the more general notion of “limit.” Quoting Wittgenstein, Raffestin (1980, 170) affirms that “the limits of my world are the limits of my language.” The extension of a *territoire*, therefore, is not necessarily geographical but semantic. Similarly, *teritorialité* is defined by Raffestin as a system of relations, exchanges, and flux that the territorializing actor entertains with both the physical environment (exteriority) and the social environment (alterity). There is here an explicit attempt to rewrite geography in light of social theory, anticipating the path which also anglophone geography would have then followed. Yet, until recently, Raffestin’s ideas have remained largely confined within francophone geography, with the only relevant exception of Italy, whose geographers have been the first to build on and to add empirical evidence to Raffestin’s theoretical framework.

**Territory: a new renaissance?**

After living in the shadow of place from around the mid-1990s, territory has since 2009 attracted new interest among geographers. This revival has taken various directions, from historical accounts of its origin and use to reconceptualizations aimed at overcoming its strict governmental character. One of the most ambitious projects on the historical excavation of the notion of territory is *The Birth of Territory* by Stuart Elden (2013). In this book and in numerous earlier journal articles, Elden proposes putting territory back in its historical context. Criticizing the largely ahistorical treatment of territory and territoriality in the social and political sciences, Elden offers a meticulous history of the notion, meaning, and use of territory by surveying an extensive amount of historical, political, and literary texts, from ancient Greece to eighteenth-century Europe. On this basis, Elden argues that in Western thought the modern concept of territory is closely tied to the notion of jurisdiction first theorized in the late Middle Ages by the Italian glossators of Justinian’s *Corpus Iuris Civilis* and then refined, in terms of exclusive sovereignty, in seventeenth-
and eighteenth-century Europe. In his reading, territory is a historically and geographically specific form of political organization associated with the gradual emergence of the modern state. Moving onto a more conceptual digression, Elden maintains that territory is both “land” and “terrain,” but also more than them. As a land, territory stands for a relation of property, providing sustenance to its people. As a terrain, territory is both the stage and the stake of conflict over land. Yet, neither a political-economic nor a political-strategic understanding can fully capture the specifics of territory as a political technology for measuring land and controlling terrain. Territory as a technology is the historical product of a series of advances in geometry, cartography, and land-surveying that have allowed for the effective establishment of linear boundaries. This quantification and measurability of space is what made territory possible (although in his book Elden specifies how territory as a political technology should be understood, in a Foucauldian sense, as an art or technique of the political rather than mere technology). For this reason, Elden maintains that boundaries are a second-order problem when it comes to territory, as territory is primarily a product of calculable space – although it might be argued that between the introduction of these calculable techniques in the seventeenth century and the emergence of a state system based on modern, exclusive territoriality in the nineteenth century, as the most recent historiography suggests, there is an important gap, which demands further historical investigation. Moreover, as maintained by Antonsich in an exchange with Elden in Progress in Human Geography, to exclusively associate territory with modern techniques of calculability risks obliterating other forms of territorial organizations that existed before, during, and after the rise of modern territory (Antonsich 2011; Elden 2011). True, The Birth of Territory also surveys a great deal of these political spatial arrangements. Yet, in the name of a rigorous etymology, Elden avoids labeling them territory.

The value of Elden’s work, though, consists not only in offering a comprehensive historical account, but also in reading territory in light of the Foucauldian analysis of power. Territory is about the reification and naturalization of space – as well as of the sociospatial relations that insist on it – by the state. This Foucauldian “territory effect” has been read by Elden in conjunction with the etymology of “territory,” whose Latin term, territorium, derives from the verb terrere, meaning to frighten, to terrorize. Although Elden (2009) engages with the debate surrounding this etymology, he privileges an analysis of the concrete, practical link between territory and terror as a way to interrogate more broadly contemporary geopolitics in an epoch of “war on terror.”

An alternative excursus in the history of political institutions is offered by Saskia Sassen (2006). For Sassen, territory equates with a bordered space imbricated with authority and rights into what she calls organizational assemblages. Contrary to Elden, she takes territory as a transhistorical category which cannot be reduced to the modern state. Her concern is therefore to map variable instantiations of territory across time. Once liberated from its exclusive association with the modern state, territory acquires for Sassen an analytical power that would allow exploring the formation of new types of bordering capabilities in the age of globalization.

A similar approach also characterizes the work of Marco Antonsich (2009). His aim is to rescue territory in an epoch where place has emerged as the dominant scientific paradigm. However, rather than reproducing notions of territory as a symbolic resource in nationalist discourses or as a control device in the hands of the state,
TERRITORY AND TERRITORIALITY

Antonsich proposes looking at territory as a space of political belonging beyond the sociospatial congruence of nation and state. This approach would also pay due attention to the agency of people, who are not passive pawns waiting to be controlled, contained, or disciplinized by the state but active players in the reproduction of territories.

Another attempt at reconceptualizing territory is that of Joe Painter (2010), who, along with Jessop, Brenner, and Jones (2008), suggests exploring its construction in close relation to the notion of network. Rather than treating territory and network as two rival principles of spatial organization, Painter shows how the former is actually the product of networked sociotechnical practices, building on his consolidated investigation of prosaic geographies of the state.

Besides a continuing interest in more traditional aspects of territory (Murphy 2012), the revival of this notion can also be associated with the “discovery” of Raffestin by anglophone geography (see Klauser 2012). The relational approach to territory pioneered by the Franco-Swiss geographer has also found new interpreters (Brighenti 2010), who further push the idea that territory does not revolve around space but revolves around a program of relatedness between individuals and that, in order to be implemented, territory has to be grounded.

Finally, a newly published journal, Territory, Politics, Governance seems to confirm the new interest surrounding the notion of territory today. Even in this case, territory is treated in a pluralist vein as a signifier of a political spatial arrangement not exclusively associated with the state. According to the journal’s editors, territory should be regarded not as a mere cage within which people are contained by the state, but as a medium for fulfilling positive projects from changing society to administering public goods. Forty years after Gottmann’s (1973) original

reflection on territory, this move has brought us back to where we started.

SEE ALSO: Borders, boundaries, and borderlands; Frontiers; Geopolitics; Nationalism and geography; Place; Political geography; Sovereignty; State, the

References


Further reading

Text and intertextuality

Jonathan M. Smith  
Texas A&M University, USA

Geographers began to use “text,” “textuality,” and “intertextuality” as theoretical terms in the late 1980s, as part of what was then called the new cultural geography. These terms denoted concepts that had originated more than 20 years before in literary theory, and were adopted by geographers who felt a need to go beyond the humanistic hermeneutics of the 1970s. Geographers employed these terms and concepts in the interpretation of a wide range of objects, including scientific and literary documents, geopolitical discourses, maps, and landscapes, one result being a spreading conviction that the meaning of such objects is arbitrary, constructed, and endlessly contestable. Because the theory of which these terms were part represented meaning as unstable, provisional, partial, and vulnerable to deconstruction, these geographic studies were said to be poststructuralist. In its early days, Marxist geographers were among the harshest critics of poststructural geography; however, over the course of the 1990s, the two traditions merged to form critical geography.

The concept of text was not invented by literary theorists in the late 1960s. Nor was textual interpretation unknown among geographers prior to the late 1980s. Indeed, the concept of a text as an object of interpretation has a long history in the Christian and humanistic scholarship of the West.

In traditional hermeneutics, a text is a document subject to critical analysis and elucidation.
TEXT AND INTERTEXTUALITY

most often sacred scripture, but also, at times, a classical manuscript or legal document – and the commentary or interpretation was either an exposition of recondite terms and obscure passages, or an application of the “principal truth” of the text to a contemporary question of thought, taste, or conduct. This pattern began to break down for three reasons.

First, critics began to regard many more documents as texts, which is to say, as documents worthy of critical interpretation. At first they turned their attention to works of modern literature, many of which were replete with difficult points and obscure passages, and then later, in the late 1960s, they extended their commentary to anything – verbal or nonverbal – that could be made an object of interpretation. Thus critics began to speak of buildings, events, images, and landscapes as “texts.”

Second, the didactic authority of texts was diminished, and sometimes denied. The word “lesson” originally denoted a reading from scripture, and the relation between readers and texts such as the Bible or Plato’s Republic was understood as a tutorial relation. The authors of these texts were authorities, in one instance God himself, so men read these texts in the expectation of learning something from them. Under modernism, with its condescending view of the past and admiration for science, the relation between reader and text changed from a tutorial relation to what might be called an interrogatory relation. The text was not an oracle; it was a specimen.

Third, criticism became a creative activity. In the classical pattern, the critic was a humble conduit who conveyed the meaning of a great authority – perhaps the greatest authority – to a modern audience. His job was to make sure nothing had been added to, or subtracted from, the original, oracular message. In the modern period, however, it was common for a critic to discover in a text meanings the author of the text had never dreamed of. The exaltation of the critic over the author was complete when theorists such as Julia Kristeva and Roland Barthes began to challenge “authorial intention” and to speak of the author as a fiction produced by his or her text.

With this background in hand, we can define the theoretical terms “text,” “textuality,” and “intertextuality” as they have been used in geography since the late 1980s. A text is simply an object of interpretation, whether or not it has any words in, on, or beside it. An artifact becomes a text when a critic subjects it to interpretation. It makes no difference what the persons who fabricated the artifact were trying to express when they fabricated it, or whether they were trying to express anything at all, since authorial intentions are, at best, one reading among many. A text is not a product of authorial intention, but rather of critical intention.

Textuality is the property of being susceptible to interpretation. Applied to an artifact of verbal communication such as a book, it denotes the semantic properties of the book, as distinct from its physical properties, such as its weight, or economic properties, such as its price. The term is, however, most frequently employed to denote the property of being susceptible to interpretation in artifacts that were not intended, and are not generally recognized, as communicative gestures or utterances. To say that a landscape or an event has textuality is, therefore, simply to assert that it can be read or interpreted.

Critics coined the term “intertextuality” in the late 1960s to denote the state, allegedly common to all texts, of being joined in a dynamic and interrelated galaxy comprising all possible texts. The galaxy itself is called the intertext or social text. More specifically, the theory of intertextuality proposes that the meaning of any particular text is produced, and changed, by its shifting relations to other texts in the intertext. Text is,
of course, broadly defined, so the intertext takes in, not only verbal documents, but also images, objects, and events. In fact, it has often been claimed, most famously by Jacques Derrida, that there is “nothing outside the (inter)text.”

In theory, the texts of the intertext are represented as seething and stirring like particles in the vortex of Epicurus, their ever changing resonances, reactions, and relations resulting in an “endless play of signification.” The metaphor of a “semiotic machine” is sometimes used to suggest the robotic production of meaning by the revolutions of the intertext, which acts autonomously. In practice, of course, it is the critic who constructs (and deconstructs) the meaning of a text because it is the critic who proposes (or dismisses) the context in which the text is interpreted, and who thus produces (or deactivates) the signifying resonances, reactions, and relations. As we have seen, criticism has become a creative activity.

The main point to notice about the doctrine of intertextuality is that, on this doctrine, a meaning is never absolute. Another way to say this is that no meaning is a true meaning. Every meaning is partial, provisional, and dynamic. An “endless play of signification” means that it is impossible to fix or establish the significance of any text. No reading should be “privileged,” which is to say regarded as authoritative or correct, because every reading is the result of placing the text in an arbitrary context. As one geographer described it, “the appeal to context … involves the installation of borders that provide a secure frame within which calculations of an otherwise unbounded textuality can be contained,” but such borders are always arbitrary because “textuality has no beginning or end, it is … limitless” (Barnett 1999, 280–281, 284).

This arbitrariness of context gives rise to what is often called the “indeterminacy of meaning.” An early discussion of the idea by geographers described this as “the liberation of meaning in the post-modern world, the freedom of intertextuality with which we characteristically invert signs and symbols to recycle them in different contexts and thus transform their reference” (Cosgrove and Jackson 1987, 98). Another geographer similarly wrote that “intertextuality accounts for the apparent instability of meanings because their contexts change so rapidly” (Burgess 1990, 144). It was the idea of the arbitrariness of context and the indeterminacy of meaning that caused scholars in the 1980s to speak of a “crisis of meaning,” and the general theory that supposedly captured this crisis of topsy-turvy significance was called, expressively enough, antifoundationalism, poststructuralism, or deconstruction.

It must be understood that poststructuralist geographers did not look on this “crisis of meaning” with despondency and alarm. On the contrary, far from toiling to resolve the crisis, poststructuralist geographers did everything they could to make it worse.

This is why.

When we look out on the world, it appears to us fraught with meanings, and it appears that these meanings are stable, structured, and grounded in the very nature of things. We produce this meaningful world in an act of interpretation that places the objects of our interpretation (i.e., texts) in what we are accustomed to believe are their proper contexts. Taken together, these customary contexts or frames of interpretation are the structure or foundation of our interpreted world. However, according to the theory of poststructuralism, this world is an illusion because there are no proper contexts, and hence no true or stable meanings. What is more, this illusory world is a sinister illusion because it is constructed by powerful social interests, who enforce these customary contexts, and who thus stabilize these apparent meanings, in a form that they find satisfying or advantageous. As one
TEXT AND INTERTEXTUALITY

ageographer put it, “effects of social power are understood to take the form of wholly arbitrary stabilizations of the necessary indeterminacy of meaning” (Barnett 1999).

The purpose of poststructural geography, therefore, is to destabilize or “deconstruct” such meanings by announcing the arbitrariness of “proper” contexts, destroying the appearance of true foundations, and transporting humankind beyond the structure and into the “freedom of intertextuality.” Poststructuralism may therefore be thought of as a republication of Nietzsche’s “transvaluation of values,” not only because it aims to overthrow the established order (which it posits as arbitrary), but also because it ultimately grounds meaning in the will. When the will aims to produce aesthetic satisfactions, we may speak of the poetics of poststructural interpretation; when the will aims to produce rhetorical advantages, we may speak of its politics.

Much of what geographers have written about text, textuality, and intertextuality is at a very high level of abstraction, but these concepts have also been put into practice in several lines of empirical research. There were, for instance, several inquiries into the textual practices, hermeneutic protocols, and reading strategies that were dominant in geography’s research literature, travel writing, and geopolitical discourse. In the last instance there was mention of “deconstructing the textuality of foreign policy” by exposing the arbitrary and tendentious “discursive expressions” and “representational practices that are foreign policy” (O Tuathail 1993).

Poststructuralism also stimulated interest in critical readings of the “textuality of maps.” A map is, of course, a symbolic representation that conveys meaning much like a verbal text, but poststructural geographers were not concerned with map reading as conventionally understood. To read a map as a representation of a portion of the Earth’s surface was, after all, to read it in the context of the “discourse” of scientific inquiry. But a map could be read, and poststructural geographers believed that it should be read, in the context of ideological suspicion and exposure. It should be read as a representation that serves some powerful interest by stabilizing meaning in a form that satisfies or serves that interest.

The textual metaphor was most often employed by geographers who were interpreting landscapes. When they spoke of landscape as a text, they were, first of all, drawing attention to its textuality or susceptibility to interpretation. This was not an altogether original idea, since geographers had long spoken of “reading” or “interpreting” the cultural landscape, but using the term “text” allowed them to adopt (some might say smuggle ashore) the entire apparatus of poststructural interpretation, with its doctrines of intertextuality, indeterminacy, and the construction and deconstruction of meaning by power. The method was used most effectively by the geographers James and Nancy Duncan, who applied the poststructural concepts of text and intertextuality to landscape and thereby represented landscapes as transformations of ideology. The textual approach to landscape was thought especially promising because, viewed in this manner, the landscape appeared not only as an ideological representation, but also as a uniquely cunning representation that masked its ideological content. This was because “the cultural landscape serves to naturalize or concretize – to make normal – social relations as embodied in the various discourses and their combinations” (Schein 1997, 676).

Not all geographers welcomed the textual metaphor and its attendant theoretical doctrines. Many traditional cultural geographers scoffed at the jargon, which they dismissed as a pretentious way to talk about the established practice of reading the landscape. They were, no doubt, sometimes correct about the pretension, but
they were slow to grasp the radical implications of indeterminacy and deconstruction. The most strenuous objections to the textual metaphor and its attendant doctrines were raised by Marxist geographers who complained that it is a form of idealism that diverts attention from material relation, that encourages a life of ironic detachment rather than revolutionary action, and that cannot, in any event, explain anything.

On Marxism, meanings are mere reflections and shadows of material social relations. They are, as it were, symptoms of an underlying disorder; so to explain material social relations as a consequence of textual meanings is, on Marxism, to get causality backwards. It is an error akin to explaining an infection as an effect of inflammation. This is why, rather than speaking of the textuality of physical objects, Marxists urged geographers to speak of the “physicality of texts.” When they did this, they were encouraging greater attention not to the size and weight of books, but rather to the material relations that, according to Marxism, produce the consciousness of the authors and the content of books.

Ultimately the dispute between Marxists and poststructuralists was overcome in a dialectical synthesis that combined textuality and physicality in a hermeneutic cycle. This synthesis began with a denial of the “structural binary that separates discourse and materiality into separate ontological domains,” and ended with an affirmation that “in all practices, textuality and materiality are inextricably intertwined” (Braun 1996, 717). As with all dialectical syntheses, the validity of this argument is far from self-evident, but the utility of combining Marxism and poststructuralism is obvious. Both theories are subversive and revolutionary, both operate on an assumption that the world of appearances is an evil sham sustained by power, and both proceed in the faith that this evil sham can be exposed and defeated through theoretical redescription and critical analysis.

“Text,” “textuality,” and “intertextuality” are still used as terms of art in the geographic literature, but they are no longer seen as daring or new. The term “intertextuality” retains currency because it is so well adapted to describe the production of meaning in cyberspace and our ubiquitous media environment, neither of which existed when it was first coined (e.g., Cupples and Glynn 2013). The general theory of intertextuality also continues to underwrite claims that interpretation is characterized by “essential incompleteness,” that “ultimately, there is no final, “true” reading,” and that “texts are apt to be received, reworked and resisted” according to the “presuppositions” – which is to say, the contexts – in which they are read (Tyner 2005; Eriksson 2008).

SEE ALSO: Critical geography; Cultural turn; Discourse; Landscape iconography and perception; Literary geography; Poststructuralism/poststructural geographies

References

TEXT AND INTERTEXTUALITY


Therapeutic landscapes

Allison M. Williams
McMaster University, Canada

The concept of therapeutic landscape is recognized by health geographers as one of the key contributions made to the study of health. Wilbert Gesler is credited as the pioneer of the idea, defining therapeutic landscapes as those changing places reputed to have an “enduring reputation for achieving physical, mental and spiritual healing” (1993: 171). He highlighted a number of therapeutic landscape themes, which he categorized as those having either “inner/meaning” (i.e., natural setting, built environment, sense of place, symbolic landscapes) or “outer/societal context” (i.e., beliefs and philosophies, social relations and/or inequalities, and territoriality). Gesler examined these themes in a number of traditional healing sites, including Epidauros (Greece), Lourdes (France), and Bath (England). Gesler’s thinking was subsequently embraced by a growing number of geographers and other health researchers (representing such diverse disciplines as nursing, kinesiology, medical anthropology, and sociology), many of whom were interested in present-day places for health. Although initially intended to encompass landscapes known for their therapeutic qualities, and therefore of interest to those experiencing ill-health, the concept has evolved to include places recognized as having health-promoting effects (Williams 2007). The idea has provided health geographers and others with a tool to better understand the characteristics of place that contribute to the making of healing places and symbolic landscapes, and to the marketing of such places.

Building on Gesler’s original work, scholarship has encompassed a large number and range of present-day applications, such as natural wilderness environments, and environments characterized as having a mix of natural and built attributes, such as monastery retreats, holy wells, pilgrimage sites, children’s camps, and yoga ashrams. Further, there has been a call for exploring sites that provide relief from the fast-paced, stress-filled lifestyles that characterize the Western world. These many contemporary applications continue to highlight the therapeutic place characteristics that contribute to healing; such characteristics are many and wide-ranging, contributing to better understanding the lesser-known mental, social, and spiritual elements of these sites.

Much of what has been learned from the physical places known for health has informed the practical application of the concept to a wide range of health-care sites and services. Some of the more common sites for exploration have included hospitals, residential care facilities, continuing care retirement communities, purpose-built residential homes for young adults with complex disabilities, respite centers, and the home as a setting of care. Research on hospital design for the mentally ill in the United Kingdom (Wood et al. 2013) provides an excellent application. A recent application of these ideas involves medical tourism sites, where forms of health care are combined with travel and leisure. Addressing the health-care spaces and places for particular groups, often in unique sites, frequently focuses on marginal and special populations, including addicts. Further, a small number of applications
THERAPEUTIC LANDSCAPES

have been made to specific cultural groups, such as indigenous peoples.

The symbolic component of the therapeutic landscapes concept is central to research on “fictive geographies,” as this is what reflects people’s perceptions and meanings of both illness and health. Employing literary analysis of fiction, a tool borrowed from cultural geography, has equipped health geographers to interpret literature using the therapeutic landscape concept as an analytic framework to better understand how place impacts health.

The variable nature of landscapes, with regard to their capacity for restoration and/or risk, suggests that places can be simultaneously healthful and hurtful. Such uneven landscape perceptions clearly illustrate the everyday nature and variable therapeutic qualities of place, while also recognizing the larger social, political, and economic context in which they are situated. For example, Masuda and Crabtree (2010), in their work on Vancouver’s (Canada) highly stigmatized Downtown Eastside, reveal a contradictory therapeutic response to environmental injustice, illuminating how the inner-city landscape operates as both restorative and risky for society’s most marginalized.

The performative nature of therapeutic landscapes is more recently highlighted in the rituals involved in the use of places characterized as therapeutic (i.e., healing rituals related to water, song, prayer). Similarly, there is a growing literature examining the mobile and embodied construction of therapeutic landscapes via walking modalities and other forms of travel.

The emergent topic of bluespace highlights the specific importance of water and sky in healing places, accentuating the known healing properties of water as one of the most important landscape elements.

SEE ALSO: Health geography; Health and wellbeing; Medical tourism; Place

References


Thiessen polygons

Ikuho Yamada  
Chuo University, Japan

Thiessen polygons, also known as Voronoi polygons, are generated around a set of points in a given space by assigning all locations in that space to the closest member of the point set. Any location in a Thiessen polygon is closer to the corresponding point inside it than to any other member of the point set. A spatial tessellation generated in this way is called a Voronoi diagram. The Voronoi diagram, as well as the Delaunay tessellation, which is its dual (as will be explained below), is an interdisciplinary concept that has been used extensively in many different fields including, but not limited to, archaeology, astrology, cartography, computational geometry, ecology, geography, geology, marketing, meteorology, physics, and urban and regional planning. Because the concept of the Voronoi diagram is fairly simple and has potential use in many applied fields, there is no surprise that it has been discovered independently in different contexts and that a variety of terms have been coined for it. While some of the terminologies will be introduced later in relation to a brief history of the Voronoi diagram, Voronoi-based terminology will be used hereafter since it seems to be the most extensively used in the literature.

Each Voronoi polygon may be seen as an area of influence of a point in a given set. For example, Voronoi polygons generated for a set of points representing fast-food restaurants can be used as approximated market areas of the restaurants given that foods and services provided by each restaurant are practically the same so that it is reasonable for customers to visit the nearest one. The concept of the Voronoi diagram is also utilized for the purposes of spatial interpolation, point pattern analysis, and locational optimization, as well as for models of spatial processes.

Brief history

As mentioned above, the concept of the Voronoi diagram has such a broad range of potential applications that it has been developed and (re)discovered independently in many research fields. This section discusses a brief history of the concept of the Voronoi diagram, mainly relying upon a comprehensive textbook on the Voronoi diagram by Okabe et al. (2000).

Diagrams that resemble the Voronoi diagram are found in the work of French philosopher and mathematician René Descartes (1596–1650). The diagrams are used in his discussion on cosmic fragmentation, but with no specific explanation about their construction. The first comprehensive presentations of the concept can be traced back to Peter Gustav Lejeune Dirichlet (1805–1859), a German mathematician, and Georgy Feodosevich Voronoy (1868–1908), a Ukrainian mathematician. They utilized a special form of the Voronoi diagram in their studies on positive definite quadratic forms. “Dirichlet domain” and “Voronoi region,” each reflecting their names, respectively, are arguably the most frequently encountered terms for the concept of the Voronoi polygon.

A more practical application is found in the work by Alfred H. Thiessen, an American meteorologist, after whom the term
“Thiessen polygon” was created. He used Thiessen/Voronoi polygons to obtain an improved estimation of regional rainfall averages (Thiessen 1911), which is one of the first examples of using the Voronoi diagram for spatial interpolation. Meteorologists as well as geographers and social scientists in related fields tend to prefer the term Thiessen polygons in their two-dimensional applications. Application of the Voronoi diagram for spatial interpolation has also been carried out in the context of estimation of ore reserves from observations at bore holes. In this type of application Voronoi polygons are simply called “area of influence polygons.”

The concept of the Voronoi diagram was rediscovered by Eugene Wigner and Frederick Seitz, Hungarian and American physicists, respectively, in their study of metallic sodium (Wigner and Seitz 1933). The term “Wigner-Seitz regions” is commonly used among physicists. Independent development of the concept of the Voronoi diagram continued in the fields of crystallography, cryptography, ecology, biology, and astronomy until the late twentieth century. While computational burdens of constructing Voronoi diagrams for large-scale data had long limited empirical applications, advancement in computational geometry that was stimulated by that in computer science since the 1960s has facilitated development of efficient algorithms starting around the 1980s.

Whereas the development and early applications of the concept of the Voronoi diagram mostly occurred in natural sciences as outlined above, applications in social sciences can also be found as early as the late nineteenth century in epidemiology and linguistics: a study of a cholera outbreak in London and a study of dialect variations in southwest Germany, respectively. Analyses of service/market areas, human territorial systems, and point patterns remain the major areas of applications in social sciences. In epidemiology, for example, the Voronoi diagram has been used to estimate service provision areas of medical facilities. Zwarenstein, Krige, and Wolff (1991) used such service provision areas to evaluate inequality in medical services between racial groups, and Tanser et al. (2001) used them to assess the hypothesis that patients choose the nearest medical facility, both in South Africa.

### Definition and use of the Voronoi diagram

For a given set of distinct, isolated $n$ points, $P = \{p_1, \ldots, p_n\}$, in a Euclidean plane, $R$, a Voronoi diagram is constructed by assigning all locations in $R$ to the closest member of $P$. An example of the Voronoi diagram is shown in Figure 1. This construction is referred to as “Voronoi tessellation.” The Voronoi diagram consists of $n$ polygons, each associated with a point in $P$. The polygons are referred to as “Voronoi” or “Thiessen polygons,” and each of the Voronoi polygons is convex with possible exceptions along the boundary of $R$. They

![Figure 1 Voronoi diagram.](image)
THIESSEN POLYGONS are collectively exhaustive in $R$ and mutually exclusive except for their boundaries; in other words, the Voronoi polygons exhaustively fill the plane $R$ with no overlap. In construction of the Voronoi diagram, a location that is equally close to two or more members in $P$ is assigned to the multiple members, which collectively form the boundaries of the Voronoi polygons. A boundary of the Voronoi polygon consists of line segments, which are referred to as “Voronoi edges.” Two Voronoi polygons that share the same Voronoi edge are considered to be “contiguous.” A vertex of a Voronoi polygon, that is, an end point of a Voronoi edge, is referred to as a “Voronoi vertex.” The points in $P$ are called “generators,” “generator points,” or “sites” in some literature, and the set $P$ is called the “generator set.” The related terminology is indicated in Figure 1.

Mathematically, the Voronoi polygon associated with $p_i$ (or simply, the Voronoi polygon of $p_i$), $V(p_i)$, is defined as shown in equation 1.

$$V(p_i) = \{x \mid ||x - x_i|| \leq ||x - x_j|| \text{ for } j \neq i, \ j \in I_n \}$$ (1)

where $x$ is a location vector for a point, $p$, $||x - x_i||$ is a distance between points $p$ and $p_i$, and $I_n$ is a set of integers from 1 to $n$. Then, the Voronoi diagram generated by $P$ (or simply, the Voronoi diagram of $P$) is given by the set

$$\mathcal{V} = \{V(p_1), \ldots, V(p_n)\}$$ (2)

So far, the Voronoi diagram has been discussed in relation to the two-dimensional Euclidean plane. However, the concept can readily be extended into the $m$-dimensional Euclidean space. The Voronoi tessellation in this case splits the space into nonoverlapping, space-filling polyhedrons instead of polygons. Moreover, Voronoi diagrams may be constructed in a space different from the Euclidean space; for example, Voronoi diagrams on a sphere, a cylinder, or a cone are possible. Because the majority of applications in geography and related fields remain in the planar Euclidean space, the following discussion focuses on the planar situations.

A dual diagram of the Voronoi diagram is called the “Delaunay triangulation.” For a given planar Voronoi diagram with three or more generator points that are not on the same line, a Delaunay triangulation is constructed by joining all pairs of generator points of which Voronoi polygons share the same edge. Figure 2 shows the Delaunay triangulation that corresponds to the Voronoi diagram in Figure 1. The resulting Delaunay diagram consists of nonoverlapping triangles, which are referred to as the “Delaunay triangles.” The Delaunay triangles are known to be as close to equilateral as possible. Edges and vertices of the Delaunay triangles are referred to as “Delaunay edges” and “Delaunay vertices,” respectively. Every Delaunay vertex is a generator point in $P$ by definition. The Delaunay triangulation is widely applied in the context of spatial interpolation of a spatially continuous variable. For instance, suppose that the generator set $P$ represents a set of weather stations taking rainfall measurements. Then imagine that each station is assigned a height that is proportional to
the measurement at that station. The Delaunay triangles whose vertices are projected to these heights form a surface estimating a continuous distribution of rainfall.

A Delaunay triangulation may also be constructed directly from the set of points \( P \). For each possible pair of members of \( P \), examine its circumscribed circle. Only when there are no points inside the circle, connect the two members. A resulting diagram after all possible pairs have been examined is the Delaunay triangulation. As a dual diagram of the Delaunay triangulation, a Voronoi diagram can be constructed from a given Delaunay triangulation, too. More specifically, it can be defined by perpendicular bisectors of the edges of the Delaunay triangles as Voronoi edges and corresponding circumcenters as Voronoi vertices.

Since the Voronoi tessellation assigns every location in a given space to the closest generator, one of the most straightforward uses of the Voronoi diagram is to identify the closest facility among a set of facilities given as a generator set. A facility location problem is thus a natural application field of the Voronoi diagram. As mentioned earlier, a market area analysis is another application field assuming that customers are likely to visit the closest store. The Voronoi polygons may be used as approximated market areas of generators so as to estimate the number of potential customers and/or to plan a location of a new restaurant.

In cartography, the Voronoi diagram is used to create a choropleth map. When each generator is given a certain property value, a choropleth map may be based on the value itself or on a density measure computed with the area of the Voronoi polygons if the property value is a count of some spatial objects, like a population. In the simplest case that each generator represents a single object like a person, 1 divided by the area of the Voronoi polygon indicates a density, for instance, a population density. The Voronoi diagram can also be used to solve the largest empty circle problem. For a given set of points \( P \), the problem is to find the largest empty circle of which center is in the convex hull of \( P \). Such a circle is given as the largest of the largest empty circles centered at the Voronoi vertices generated by \( P \). When \( P \) represents retail facilities, for example, the largest empty circle suggests an area that is the most inconvenient for shopping. Furthermore, Voronoi-like diagrams are often found in natural structures, so that Voronoi diagrams can be used to model spatial processes to form them. Crystallization is an example of such processes. Further applications will be introduced later in conjunction with extended versions of the Voronoi diagrams.

Extensions of the Voronoi diagram

The Voronoi diagram discussed so far is the most fundamental form of its kind and it has been extended in many directions since the early 1970s. This fundamental form of the Voronoi diagram is referred to as the “ordinary Voronoi diagram.” The basic idea to form a Voronoi diagram is that every location in a space is assigned to a member of a generator set, and resulting sets of locations assigned to all members of the generator set are collectively exhaustive and mutually exclusive except for their boundaries. Various forms of Voronoi diagrams are therefore possible depending on the rule of assigning a location to a generator and the generator set. An element in the generator set may be a point, a set of points, a line, a polygon, and so on. They may also be weighted differently reflecting a certain property that they possess; for example, the population of a city and the number of beds in a hospital. In the case of the ordinary Voronoi diagram, the assignment rule is to assign
THIESSEN POLYGONS

a location to the closest generator and the type of generator is a point with an equal weight.

The line Voronoi diagram is a Voronoi diagram generated for a generator set, \( L = \{L_1, \cdots, L_n\} \), of which elements, \( L_i \ (i = 1, \cdots, n) \), are line segments or geometric elements consisting of connected line segments. Line segments \( L_i \) may be curved, but they are assumed not to intersect with one another. Defining that the distance from an arbitrary point, \( p \), to a line segment \( L_i \) is the shortest-path Euclidean distance and denoted as \( d_i(p, L_i) \), the line Voronoi region associated with \( L_i \) is given by equation 3.

\[
V(L_i) = \{ p | d_i(p, L_i) \leq d_j(p, L_j) \text{ for } j \neq i, j \in L_i \} \tag{3}
\]

Note that the term “Voronoi region” is used instead of “Voronoi polygon” because each resulting region in this case is not necessarily a polygon. The area Voronoi diagram can be defined analogously. Okabe and his colleagues illustrate how the line and the area Voronoi diagrams can be utilized in the context of urban analysis to define areas of influence of road segments and parks, respectively. Okabe and Yoshikawa (1989), for instance, applied the line Voronoi diagram to investigate the potential impact of vehicle emissions on roadside trees along highways.

While the ordinary Voronoi diagram deals with a set of points \( P \) as a generator set, a family of Voronoi diagrams named the “higher-order Voronoi diagrams” considers a set of all possible subsets of \( k \) \((2 \leq k \leq n)\) points out of \( P \). Letting \( P_i^{(k)} \) be a subset of \( k \) points in \( P \), that is, \( P_i^{(k)} = \{p_{i1}, \cdots, p_{ik}\} \) and \( p_{ij} \in P \), a Voronoi polygon, \( V(P_i^{(k)}) \), of the “order-\( k \) Voronoi diagram” is composed of all locations that are closer to points in \( P_i^{(k)} \) than to any other points in \( P \). The order of the \( k \) points is not indifferent here; that is, point \( p_{i1} \) may be the closest from a location \( p \) in \( V(P_i^{(k)}) \), or it may be the \( k \)-th closest. When the order is taken into account, the resulting diagram is referred to as the “ordered order-\( k \) Voronoi diagram.” Applications of the higher-order Voronoi diagrams may be found in a facility location problem where identifying the \( k \) nearest facilities is critical; for example, a system to dispatch emergency vehicles would require identifying the second-nearest station when more than one vehicle is needed and/or vehicles in the closest station are unavailable.

Another direction of extensions is to consider an assignment rule based on the \( k \)-th nearest member of the generator set \( P \), rather than the nearest one. A Voronoi diagram constructed by assigning all locations in a space to the \( k \)-th nearest member of \( P \) is called the “\( k \)-th nearest-point Voronoi diagram.” A special case of \( k = 1 \) corresponds to the ordinary Voronoi diagram, and that of \( k = n \) is referred to as the “farthest-point Voronoi diagram.” Such Voronoi diagrams can also be used in the facility location problem mentioned above; the farthest-point Voronoi diagram in particular may be found in applications with a min-max criterion.

Other families of extensions can also be achieved by modifying the way in which a distance between two locations is determined. The Voronoi diagrams discussed above are based on the Euclidean distance, which almost implicitly assumes that one can take a straight path between any two locations in the space. The “shortest-path Voronoi diagram” is developed to deal with a situation where there exist some obstacles that prevent straight-line movement between locations. The obstacles may be natural geographic features such as rivers and lakes or artificial ones such as gated communities and large parks that one cannot go through. The underlying concept of the shortest-path Voronoi diagram can further be extended to handle a Voronoi diagram constructed in a bounded region for which a shortest path traversing

...
outside of the region is unrealistic. An obvious example in the context of geography is an island.

The distance between two locations may be weighted according to a certain property assigned to the locations. A family of Voronoi diagrams constructed based on a weighted distance is referred to as “weighted Voronoi diagrams”; the multiplicatively weighted distance and the additively weighted distance are often used. Other types of distance measures such as Manhattan distance and network distance facilitate further extensions. A Manhattan distance is defined as the sum of the absolute differences in the x- and the y-coordinates; a Voronoi diagram with the Manhattan distance would reflect a situation where movement between locations is restricted to a grid-like street network like the one in Manhattan. A network distance is measured along a street network, which may be seen as a more generic, realistic representation of movement in a city than the Manhattan distance. A Voronoi diagram in a network space is referred to as the “network Voronoi diagram.”

While construction of the network Voronoi diagram is much more complex than the ordinary Voronoi diagram, recent advancements in geographic information systems (GIS) have greatly contributed to practical implementation of its construction algorithms. SANET (Spatial Analysis along Networks), a toolbox for network-based spatial analysis in GIS, provides a tool for the network Voronoi diagram (Okabe, Okunuki, and Shiode 2006).

SEE ALSO: Interpolation: areal; Point pattern analysis; Spatial analysis; Spatial modeling; Voronoi diagrams; Spatial tessellations

References


Time geography and space–time prism

Harvey J. Miller
The Ohio State University, USA

Time geography is a constraints-oriented approach to understanding human activities in space and time. Time geography recognizes that humans have fundamental spatial and temporal limitations: people can physically only be in one place at a time and activities occur at a sparse set of places for limited durations. Participating in an activity requires allocating scarce available time to access and conduct the activity. Constraints on activity participation include the location and timing of anchors that compel presence (such as home and work), the time budget for access and activity, and the ability to trade time for space in using mobility or information and communication technologies (ICTs). Time geography is a physical not a behavioral theory; it highlights the necessary spatiotemporal conditions for human activities, but does not explain the sufficient events that lead to specific activities. But since these necessary conditions vary by individual and situation, time geography supports an approach to understanding human and environmental systems that recognizes individual histories and the importance of geographic context.

Time geography is consistent with some core ideas in fields such as geography, transportation, urban science, social sciences, and environmental sciences. These include an integrated perspective on human and physical phenomena, the need to build macro-level explanations from micro-level processing, and situating human activities within context. Basic time geographic concepts, such as events being sparsely distributed in time and space, limited time availability, and trading time for space to access activities, seem mundane, since they are common and correspond with everyday experience. But this is why time geography is needed: these seemingly banal but utterly crucial factors in our scientific explanations of human behavior should not be neglected. Time geography provides a framework that demands recognition of the fundamental constraints underlying human experience and also provides an effective conceptual system for keeping track of these conditions.

Time geography originates from Professor Torsten Hägerstrand (1916–2004), a Swedish geographer who spent his career at the University of Lund. He nurtured the ideas for a long time, but time geography emerged dramatically to the international scientific community with a now-famous 1969 presidential address to the Regional Science Association (Hägerstrand 1970). Hägerstrand was concerned that human geography and regional science were neglecting much of what comprises a livable world. He also wanted to provide a counterbalance to increasing specialization and fragmentation in science, technology, and administration by offering a more holistic view of human activities. Hägerstrand believed in the integrative power of the regional approach in geography, but felt that it needed to be more inclusive and rich. Time geography upgrades the regional approach by describing the “bare skeleton” of spatiotemporal conditions and constraints that emerge from the interplay of historical and geographical factors.

Time geography is an active and flourishing research domain one half-century after its initial
conceptualization in the 1960s. This ranks time geography among elite and enduring scientific ideas. Time geography has endured for good reasons. It is beautiful: the geometry of the space–time path and prism (discussed below) and their relationships are intuitive and appealing. It is elegant: it can help to explain much with few basic principles. It is robust: it can help us understand a wide range of phenomena in human and linked human-environmental systems. It is sensitive: it treats people as individuals and recognizes social differences across a wide range of factors (gender, age, socioeconomic status, culture) as well as geographic context. It is ecological: it connects the individual to the aggregate, balancing nomothetic law seeking with context and situation, as well as balancing agency and structure. Finally, it is practical: Hägerstrand was prescient when thinking about how to keep track of the basic spatial and temporal existential facts of human activities. It is now possible to collect, store, manage, and analyze individual-level data on mobile objects and human activities with ease and power that would seem magical from the perspective of the 1960s.

Classical time geography

Time geography recognizes three major types of constraints on human activities. Capability constraints limit the activities of individuals through their own physical capabilities and/or available resources. People need to conduct maintenance activities such as eating and sleeping; these require time and place. Also, individuals with private automobiles can generally travel faster than individuals who walk or rely on public transportation. Coupling constraints define where, when, and for how long an individual has to join with other individuals for shared activities such as work, meetings, and classes. Authority constraints are fiat restrictions over particular space–time domains. For example, a shopping mall or gated community can make it difficult and illegal to enter at designated times, while a public street cannot.

Space–time path

Activities such as personal and domestic maintenance, work, shopping, health care, education, and recreation are sparsely distributed in time and space; they are available for limited duration at relatively few locations. Participating in activities requires trading time for space to access these locations at their available times. The space–time path highlights these requirements. Figure 1 illustrates a space–time path between activity stations. Stations are places where activities can occur; classical time geography treats these as tubes designating their locations in space and availability in time (e.g., work hours, operating hours for a store, appointments, scheduled lectures). The classical space–time path focuses on physical mobility and interaction. Virtual interaction via ICTs is possible; for example, a telephone call can be represented as a connection between two space–time paths. However, virtual interaction is muted relative to physical interaction in the classic theory.

Time geography also classifies activities based on their flexibility for an individual. Fixed activities are those that cannot be easily rescheduled or relocated (e.g., work, meetings), while flexible activities can be more easily rescheduled and/or occur at more than one location (e.g., shopping, recreation). These categories can be arbitrary; for example, a software developer can code in an office or a cafe. Nevertheless, the dichotomy provides an effective means for understanding how the location and timing of some activities condition accessibility to other activities. Fixed activities act as space–time anchors because other
activities must occur at the temporal gaps between fixed activities.

**Space–time prism**

The *space–time prism* (STP) highlights the influence of space–time anchors on the ability to participate in flexible activities. The STP is the envelope of all possible space–time paths between known locations and times. Figure 2 illustrates a planar STP. In this case, two *space–time anchors* frame a prism. Anchors correspond to known locations and times for the mobile object. These are often (but not always) the locations and times of fixed activities that compel presence. A maximum travel speed represents the object’s mobility capabilities; in classic time geography this is a uniform across space and time. (Time geography uses the term “velocity,” but this is incorrect since velocity implies magnitude and direction. “Speed” implies magnitude only and is, therefore, the more appropriate term.) Given these anchors and the speed limit, the prism defines the envelope all possible space–time paths between the anchors. The spatial footprint of the STP is the *potential path area* (PPA); this is the region in space that is accessible to the moving object.

The prism in Figure 2 is general since it accounts for stationary activity time and has two anchors that are spatially separate. A prism without stationary activity time consists only of its *time-forward cone* rooted at the first anchor and its *time-backward cone* rooted at the second anchor. The STP will have a larger volume and PPA, as no stationary activity means more time to be mobile. The two anchors may also be coincident spatially; this prism consists of two right cones instead of oblique cones as in Figure 2. A prism may also have only one anchor. An STP with its first anchor only is a time-forward cone delimiting all destinations that can be reached from that origin within a specific time limit. Conversely, a prism with its second anchor only is a time-backward cone delimiting all origins that can reach that destination within a specific time limit.

The STP measures *accessibility*: the ability for an individual to travel and participate in activities and the amount of time available for activity participation at locations. An activity at a station is not feasible unless that station intersects with the prism spatially and temporally, the latter for at least as long as the minimum activity time required. This delimits the subset of opportunities in an environment that is available to a person based on their STP constraints.

The STP provides a dramatically different image of accessibility than place-based accessibility measures, such as those based on spatial interaction (“gravity”) models or simply counting the number of opportunities near a person’s...
home and work location. The locations and timings of fixed activities vary by age/life cycle stage, socioeconomic status and culture. For example, STP-based accessibility measures capture gender differences in accessibility due to household organization that are missed by place-based accessibility measures such as home–work commute length. Place-based accessibility measures assume everyone at a place, such as home and work, has the same space–time scheduling constraints, while people-based measures derived from the space–time prims recognize individual capability constraints that are masked through homogenization by place. Time geography also facilitates understanding how the temporal organization of service and trading hours can have differential impacts beyond the locations of services and businesses.

Bundling and intersections

Time geography recognizes two types of relationships between paths and prisms. Bundling refers to the convergence in space and time of
two or more paths. Path bundling is necessary (although not sufficient) evidence of shared activities and individuals meshing their space–time activities to participate in projects. Bundling can occur when objects are in motion or stationary; examples of the former include public transportation and ride-sharing. Intersection is the condition of two or more time geographic features sharing some locations in space with respect to time. Two or more people cannot physically meet unless their STPs intersect, or a path is within an STP.

Bundling and intersections are necessary conditions for the emergence of broader space–time activity systems, such as a university or a city. Time geography is an ecological theory; it considers interactions between individuals and between the individual and the aggregate. Bundling and intersections require individuals to synchronize (coordinate over time) as well as synchronize (coordinate over space). Coordination also occurs at multiple scales. Individuals must conduct projects consisting of sequenced activities linked by mobility events to meet larger goals such as hosting a dinner party. Cities and regions are time systems balancing the supply of available time and demands on that time. This is a rich and intricate conceptualization of cities and societies; Alan Pred memorably referred to a “ballet of adjustments” as disruptions propagate through a time system due to activities and projects changing to meet the new space–time requirements for participation and interaction.

Analytical time geography

Classical time geography is conceptually rich but limited analytically. The rise of geographic information systems (GIS) motivated renewed interest in time geography, particularly in relaxing strict assumptions such as the maximum speed constraining an STP being uniform in space and time. The development and deployment of location-aware technologies (LATs), such as the global positioning system (GPS), mobile phones, and radiofrequency identification (RFID) chips, have greatly expanded capabilities for collecting data on mobile objects and have led to the development of mobile objects databases and mobility mining or exploratory analysis of mobile objects data. These scientific and technological developments led to the development of analytical time geography and supporting GIS software.

Path analytics

Location-aware technologies do not generate space–time paths directly; they generate a temporal sequence of spatial locations that are used to construct the path. There are several ways to generate this temporal sequence. Event-based recording captures the time and location when a specified event occurs; for example, a person texting or calling using a mobile phone. Time-based recording captures mobile object positions at regular time intervals; this typifies GPS receivers. In change-based recording a capture occurs when the position of the object is sufficiently different from a previous location; this includes dead-reckoning methods as well as some mobile objects data technologies that avoid recording locations to manage data volume. Location-based recording occur when a mobile object comes close to locations where sensors are located; examples include stationary radiofrequency identification and Bluetooth sensors.

The simplest and most common way to generate a path is linear interpolation: assume the object followed the straight-line segment between recorded locations. This works well for time-based and change-based recording with high capture frequencies. Event-based recording creates more issues since events are often not
very frequent, and the spatial resolution can be coarse and variable; this is often the case with mobile phone data resolved only to membership in a service cell. Even coarser are paths derived from location–based recording; these are often simply sequences of visits at the sensor locations. Space–time paths can also be matched to other spatially referenced data such as transportation networks.

The ease of collecting space–time path data from location-aware technologies often comes at the expense of path semantics: details about the moving objects such as the reasons for mobility behavior. Semantics can be recovered by overlaying paths with other georeferenced data. This method can produce errors related to data inaccuracies and intrinsic ambiguities. For example, GPS positional error means that it can be hard to tell if a person is inside or outside of an activity location, for example, a coffee house. If the person is inside the coffee house, what was she doing – dining, working, socializing, or some combination of the above? Often, there is no unambiguous link between locations and activities, especially in an era with near-ubiquitous access to information and communication. Methods for recovering path semantics include decomposing the trajectory into a sequence of moves and stops, and annotating these sequences based on map matching with background geographic information. Also available are advanced data mining techniques such as machine learning algorithms.

Table 1 summarizes the fundamental measurements available from space–time paths. The primitive position allows the derivation of distance, direction, and spatial extent in the object’s movement; an example is spatial range methods. From these primary derivatives the spatial distribution of the object, changes in direction, and the path’s sinuosity can be derived. From the temporal primitives instant and interval the duration of the object’s movement and its speed, as well as the temporal distribution of the object’s movement and changes in duration, can be derived. Primary derivatives from combining spatial and temporal intervals include the object’s speed (a scalar) and velocity (a vector), and corresponding secondary derivatives acceleration and an approaching rate with respect to a location, boundary or region.

**Prism analytics**

It is difficult to analytically describe the entire STP. However, it is easy to describe its spatial extent at a moment in time; this can serve as

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Primitive</th>
<th>Derivatives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Primary</td>
<td>Secondary</td>
</tr>
<tr>
<td>Spatial</td>
<td>Position</td>
<td>Distance</td>
<td>Spatial distribution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Direction</td>
<td>Change of direction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spatial extent</td>
<td>Sinuosity</td>
</tr>
<tr>
<td>Temporal</td>
<td>Instance</td>
<td>Duration</td>
<td>Temporal distribution</td>
</tr>
<tr>
<td></td>
<td>Interval</td>
<td>Travel time</td>
<td>Change of duration</td>
</tr>
<tr>
<td>Spatiotemporal</td>
<td>Speed</td>
<td>Acceleration</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Velocity</td>
<td>Approaching rate</td>
<td></td>
</tr>
</tbody>
</table>

the basis for a wide range of prism analytics. STP parameters are \( \{x_i, x_j, t_i, t_j, s_{ij}, a_{ij}\} \) where \( x_i, x_j \) are the first and second anchor locations with associated departure and arrival times \( t_i, t_j \) respectively, \( s_{ij} \) is the maximum travel speed, and \( a_{ij} \) is the stationary activity time. At a moment in time \( t \in (t_i, t_j) \), the spatial extent of an STP (denoted by \( Z_{ij}(t) \)) is the intersection of three convex spatial sets: (i) the future disc \( f(t) \) comprising all locations that can be reached from the first anchor by time \( t_i + t \); (ii) the past disc \( p(t) \) encompassing all locations at time \( t \) that can reach the second anchor by time \( t_j - t \); and (iii) the geo-ellipse \( g_{ij} \) that constrains the prism locations to account for any stationary activity time:

\[
Z_{ij}(t) = \{f(t) \cap p(t) \cap g_{ij}\}
\]

(1)

\[
f(t) = \{x \mid \|x - x_i\| \leq (t - t_i) s_{ij}\}
\]

(2)

\[
p(t) = \{x \mid \|x - x_j\| \leq (t_j - t) s_{ij}\}
\]

(3)

\[
g_{ij} = \{x \mid \|x - x_i\| + \|x - x_j\| \\
\leq (t_j - t_i - a_{ij}) s_{ij}\}
\]

(4)

Figure 3 provides an illustration. This definition of the STP is not limited to 2-D space. In 1-D space, the sets described by equations (2)–(4) are line segments. In 2-D space, the discs are circles and the geo-ellipse is an ellipse. In 3-D space, the discs are spheres and the geo-ellipse is a spheroid. There are scalable methods for calculating these objects and their intersections. Also, only one or two of the sets are relevant at any time, since the future and past disc change in size and may be enclosed by the other two objects. For example, with a general prism as in Figure 2, it is only necessary to solve (in the following order) a disc, a disc–ellipse intersection, an ellipse, a disc–ellipse intersection, and finally a disc. Similarly, finding path–prism intersections only requires testing if a point lies within a disc, ellipse, or a disc–ellipse intersection. Finding a prism–prism intersection requires solving for the intersection of two, three, or four of the sets based on the prisms’ morphologies at that moment in time. The worst case is a four-set intersection involving two discs and two ellipses (Miller 2005).

Over time, the future disc traces the time-forward cone with an apex at the first anchor, the past disc traces the time-backward with an apex at the second anchor and the geo-ellipse is a cylinder. Over time, the prism in space is a lens-like set that traces the PPA footprint. These dynamics can be computed at discrete moments in time using the static construction described above and reconstruct the corresponding spatiotemporal region for the prism.

Error and uncertainty in paths and prisms

Error in space–time paths results from two sources. Measurement error occurs when the recording of a mobile object’s location has noise or uncertainty. The left-hand side of Figure 4 illustrates this for one segment of a space–time path; the captured locations have some degree of spatial error. This is equivalent to the problem of error in line segments and polylines, a well-established topic in the GIS literature. Sampling error occurs when the captured locations undersample an object’s movement pattern. This creates a spatial uncertainty region equivalent to the STP over time. Therefore, a space–time path with sampling error can be viewed as a sequence of linked STPs. The right-hand side of Figure 4 illustrates this; the black curve is the actual path of the mobile object, the dashed red polyline indicates the interpolated movement, and the blue ellipses indicate the spatial uncertainty region surrounding each paired sample locations.
Combined measurement and sampling error in a space–time path is equivalent to an STP with measurement error. Therefore, if it is possible to solve the measurement error problem for STPs it is also possible to solve the problem of combined measurement and sampling errors in space–time paths. One strategy is Monte Carlo simulation, which involves generating multiple prism realizations based on sampling from the stochastic error distributions. Monte Carlo simulation is effective for theoretical investigations but it is cumbersome for applications: it is awkward to run prism error simulations when executing STP-based mobile objects database queries or measuring accessibility in a model.

Spatial error propagation theory provides an analytical strategy for analyzing prism error. Given a direct geographic measurement \( \mathbf{x} = \mu_x + \varepsilon_x \), where \( \mu_x \) is the true location vector and \( \varepsilon_x \) is an error vector, it must be determined how that error propagates through a geographic operation \( \mathbf{y} = f(\mathbf{x}) \). If \( f(\mathbf{x}) \) is nonlinear, the error propagation is approximated using the first-order partial derivatives of the function. In the case of the analytical prism, the relevant functions are the intersections of the boundaries of spatial objects described by equations (2)–(4). These boundaries are:

\[
\begin{align*}
f_i : & \quad \|\mathbf{x} - \mathbf{x}_i\| - (t - t_i)s_{ij} = 0 \quad (5) \\
p_j : & \quad \|\mathbf{x} - \mathbf{x}_j\| - (t_j - t)s_{ij} = 0 \quad (6) \\
g_{ij} : & \quad \|\mathbf{x} - \mathbf{x}_i\| + \|\mathbf{x} - \mathbf{x}_j\| - (t_j - t_i - a_{ij})s_{ij} = 0 \quad (7)
\end{align*}
\]

A problem is that equations (5)–(7) define these boundaries implicitly rather than explicitly (that is, in the form \( f(x, y) = 0 \) rather than \( f(x) = y \)), meaning that solving for the intersection points requires finding the roots of high order polynomials. This difficulty can be resolved using implicit function methods that allow calculation of the required first order partial derivatives without having their explicit functions. This allows the error propagation from STP parameters to the constructed STP and to STP intersections to be estimated analytically. However, some of the required circle and ellipse intersection cases are
still unsolved and the techniques are not tractable beyond two STP intersections. Still required are scalable approximations and heuristics for STP error based on these analytical techniques. The analytical techniques can also serve as bounds for improving the efficiency of simulation-based techniques.

The methods discussed above estimate error propagation from measured STP parameters to the constructed object. Another strategy is to construct objects that encompass the possible STP consistent with the prism parameters and their errors. A rough STP consists of the upper and lower bounds on an STP. A reliable STP is the set of space–time locations where an individual can conduct an activity and meet the second anchor constraints with a specified probability.

Properties of the prism interior

The STP is traditionally a binary concept; all locations within the prism interior are considered to be equally accessible. However, this is a simple characterization that masks intricate properties of the prism interior. Intuitively, it would be expected that the distribution of visit probabilities would be unequal: locations that are near the space–time axis connecting the prism anchors are more likely to be visited than locations near the prism boundary since there are more possible paths through the former.

One way to model visit probabilities within a planar STP is to use the theory of random walks (RWs). An RW is a stochastic process in discrete space and time; given a spatial lattice of discrete locations, an RW involves, at each time step, a random choice of direction. Results from RW theory confirm intuition that locations near a prism anchor in a forward-time or backward-time cone are more likely to be visited; visit probabilities follow a bivariate multinomial distribution centered on the prism anchor.

However, movement within an STP with two anchors is not completely random; the object needs to move from the first prism anchor to second anchor within the time budget. A directed random walk (DRW) is an RW process with directional biases. A DRW process constrained by an STP requires updating the RW direction biases at each step to account for the remaining time to travel to the second anchor given the speed limit. This suggests a visit probability distribution similar to a bivariate normal distribution that is centered on the axis connecting the anchors and moves along that axis with respect to time.

A technique for modeling directed movement in continuous time and space is Brownian bridges (BBs). A BB is a continuous stochastic process between two known values; it has been applied widely to model animal movement between recorded locations. While BB can capture directionally biased random movement between prism anchors, it is not constrained by a maximum speed and, therefore, does not capture the STP boundary; it is still possible for the object to travel outside the prism boundary. A truncated Brownian bridge (TBB) imposes a maximum speed on a BB process, fully representing STP constraints. Figure 5 illustrates visit probabilities at a moment in time within the PPA of a planar STP based on a TBB process. Figure 5 shows the two prism anchors with the spatial projection of the space–time axis connecting the anchors. The gray area is the PPA for the STP; these locations have zero probability of being occupied at this moment in time. The blue, yellow, orange and red colors show increasing probability that the locations may be occupied by the object at that moment in time. These results are also consistent with the assumption of a bivariate random distribution whose center moves with time along the axis connecting the
anchor points; however, rather than assumed it is derived from fundamental movement principles.

Another strategy for estimating visit probabilities within planar STPs is to use spatial interpolation to infer unknown object locations from the known locations at the prism anchors. This assumes that the most likely locations are along the axis between the anchors, but as seen above this assumption is consistent with theory. *Time-geographic density estimation* (TDE) integrates kernel density estimation with the constraints imposed by an STP. This method can generate the fine-grained movement patterns of objects between anchors derived from sparse tracking data.

**Other types of space–time prisms**

A planar STP has interesting properties and is useful for applications where it can be assumed that objects move in constrained space with a uniform maximum speed. However, in some applications the assumptions of unconstrained space and uniform maximum speed are unrealistic; examples include vehicles within a transportation network or movement across terrain. In addition, the STP assumes unrealistic motions such as infinite rates of acceleration and deceleration.

A *network time prism* (NTP) is a prism subject to network routes and allowable speeds that vary by network arc and, possibly, time. Figure 6 illustrates an NTP for a network embedded in 2-D space with time along the third orthogonal dimension (Kuijpers and Othman 2009). The green subportion of the network arcs is the *potential network area* (PNA), the network analog of the planar PPA; these are the accessible locations within the network. The red polygon comprises the full NTP; this is the envelope of possible space–time paths between the anchors constrained by the network.

Calculating an NTP involves three steps. A precomputation step uses the planar PPA as an upper bound on the NPA; this speeds computation by eliminating network nodes that cannot possibly be in the NPA. It then solves for the shortest path tree in this subnetwork twice: once for travel from the first anchor and a second time for travel to the second anchor. The procedure assigns the earliest arrival time and latest departure time at each node. The second step builds the spatial footprint of the NTP by testing if only one of an arc’s end nodes is included in the NPA and (if so) solving for the NPA boundaries within the edge. The third step computes the full spatiotemporal region of the NTP based on the earliest arrival times and latest departure times at network vertices. These arrival and departure times dictate, for each node incident to an arc, whether the mobile object can traverse the full arc to an adjacent node or can only traverse partway and return to the original node. These cases correspond to the
Figure 6  A network time prism.

rectangular and triangular regions in Figure 6, respectively. The computational bottleneck in the algorithm is the shortest path trees in the precomputation step; although this is only applied to a subportion of the network, this subnetwork can be large, as can the number of prisms to be processed. However, the Boolean query about whether two NTPs intersect can be solved analytically.

Similar to an STP, the parameters of an NTP can also be measured with uncertainty. An anchor region is the set of all possible locations and times for a prism anchor within a network arc. These regions are not necessarily continuous in space or time and can have nonuniform probabilities. There are two ways to calculate an NTP with anchor regions. The first way is to calculate the envelope of all NTPs having an anchor point within a given anchor region. The second way is to calculate, for any space–time point, the probability that an NTP with given anchor regions contains that point. As with the standard NTP, the computational bottleneck is the precomputation phase involving shortest path calculations.

TIME GEOGRAPHY AND SPACE–TIME PRISM

A field time prism (FTP) is a planar STP where travel speeds (magnitude) or velocities (magnitude and direction) vary continuously across space. Fields are useful for describing movement across terrain or through water and air that are subject to currents. Fields can extend the NTP by treating edges as 2-D regions with continuously varying speeds or velocities due to factors such as traffic. In these cases, space–time paths have unobserved components corresponding to minimum cost curves through an inverse speed or velocity field rather than straight line segments through a uniform plane. The FTP generalizes the prism concept: the STP and the long standing concept of isochrones (curves of equal travel time) are special cases of the FTP. It also links time geography to the continuous transportation or urban fields tradition in quantitative geography and regional science.

Space–time prisms as described above are physically impossible; they assume that the mobile object can instantly accelerate and decelerate at locations such as the prism anchors and the prism’s sharp corners. Physical limitations on acceleration and deceleration mean that an STP is an overestimate of the true space–time region accessible to the object. While this may not matter for some STP applications, kinetics can make a difference for applications such as microscale movement (e.g., pedestrians, athletes), active transport modes such as bicycling, movement through media such as airplanes and ships, and animal behavior. It can also make a difference for applications in sustainable transportation, as vehicle fuel consumption and emissions are often dominated by acceleration events.

A kinetic time prism (KTP) is the set of all possible kinetic paths between two locations and times, where a kinetic path is a space–time path that obeys physical limits on acceleration and deceleration. Figure 7 illustrates a KTP in
1-D space and time, overlaid with a classic STP. Solving a KTP is scalable in 1-D space and time but complex in 2-D space. In 2-D space, only one-quarter of the prism must be solved; the remainder can be obtained through point and reflection symmetries. But calculating the first quarter of the KTP requires solving parametric functions describing the 1-D prism rotating around the line connecting the prism anchors. This can only be accomplished if it is assumed that the object’s initial heading is unknown; still open is the case where the object’s initial heading is known.

Path and prisms collections

With the proliferation of location-aware technologies generating mobile objects data, it is often the case that there are large collections of space–time paths and prisms to analyze. With these collections, it is often useful to summarize the paths or prisms through clustering (finding groups of similar paths or prism) or aggregation (forming a composite representative paths or prisms). It may also be desired to search through a collection of paths and prisms for other cases that resemble a reference path or prism. This requires methods for calculating path and prism similarity.

Path similarity is the degree of correspondence between two space–time paths. Geometric similarity measures focus only on the geometry, ignoring sequence and time; these include Euclidean and Hausdorff distances. Euclidean distances, such as the average, minimum and maximum distance are intuitive and scalable; however these measures are sensitive to noise and outliers. The Hausdorff distance is the maximum distance from a location in one path to the closest location in the other path. However, the Hausdorff distance can be misleading, as it does not take into account the temporal sequence of the path locations, only their geometry.

Dynamic similarity measures include Fréchet distance, dynamic time warping, longest common subsequences, and edit–distance functions. Unlike Hausdorff, the Fréchet distance takes into account the sequence of locations in the path and is, therefore, better suited to mobile objects. One way to think about the Fréchet distance is to imagine two space–time paths representing a person walking a dog: the Fréchet distance is the shortest leash that connects the two. Dynamic time warping measures the similarity between two sequences or trajectories based on the effort involved to stretch or compress time to get the sequences to match. Least common subsequence (LCSS) measures similarity based on the length of least common subsequence between two sequences. Edit-distance functions are a generalization of LCSS; these measure similarities between sequential patterns based on the cost of the insertion, deletion, and substitution operations required to transform one sequence into the other.

Other methods for analyzing collections of space–time paths include path clustering methods and spatial field methods. Path clustering methods find groupings of similar space–time...
paths in collections of mobile objects; these are often based on path similarity measures such as the ones already discussed. Spatial field methods translate movement patterns of objects into fields or surfaces that summarize mobility and activity frequency by geographic location, allowing the identification of “hot spots” or locations with high mobility activity.

Similar to space–time paths, it may desired to analyze a collection of space–time prisms for patterns. This may involve clustering or aggregating prisms based on morphology, finding prisms similar to reference prism, and summarizing and aggregating prisms to discover synoptic patterns. However, to date there has been little research on these questions; it is a much less developed research topic than path comparison and summarization. A fundamental problem is to develop efficient numeric measures that summarize meaningful prism properties. It is easy to calculate basic properties such as prism volume and PPA size for classic prisms, as these have an elegant geometry consisting of cones and cylinders. More generally, measures are required that can capture a wide range of morphological properties. These properties may be straightforward for classic prisms, but more complex prisms such as NTPs, FTPs, and KTPs, or prisms with fuzzy or probabilistic boundaries have more complex shapes and, consequently, are more difficult to describe analytically.

One possible strategy for measuring planar space–time prism morphology is shape analysis or quantitative measures summarizing geometric form. Shape analytical techniques typically reduce geometric forms to a single numeric indicator or a small set of indicators; for example, roughness, perforation, and elongation. However, most shape measures are specific to 2-D space: required are scalable shape measures that can handle 3- or 4-D space (treating time as static) and the evolution of shape (treating time as dynamic).

A possible way to summarize NTP structure is through network indices based on graph theory. Network indices summarize network properties such as connectivity, the distribution of arc lengths/costs, and properties of shortest paths and tours within the network. Graph theory and network analysis is a very well established field that has been newly reinvigorated due to the rise of network science to analyze phenomena such as social network and cities. A research frontier is determining which existing network indices describe relevant properties of NTPs, and the development of new network indices tailored for NTPs.

Joint accessibility and time ecology

As mentioned, time geography recognizes bundling and intersections among paths and prisms as indicators of shared activities. The next step beyond calculating bundles and intersections are the higher-level properties associated with space–time coordination and dynamics.

Collective motion methods focus on individual object movement patterns within the context of a larger group of mobile objects. Distance-based measures search for collective patterns, such as flocks, by searching for moving objects that are densely connected in space given user-defined distance and time thresholds. Relative motion methods consider the individuals’ directions to detect collective patterns such as flocking, leadership and convergence in a group of moving objects. Joint accessibility measures use prism–prism intersections as a basis for analyzing the potential benefits to coordinated activity participation.

Static measures of collective motion and joint accessibility do not directly address the ultimate
goal of understanding the dynamic processes though which shared and interlinked activities form and intertwine in space and time. This highlights a weakness of time geography: its inability to specify an explanation of human activity that goes beyond the implications of space–time constraints limiting possible activities. Time geography can explain why infeasible activities did not occur, but it cannot explain why some feasible activities occurred while others did not. However, it would be difficult to build a dynamic, processes-based time ecology theory based only on the negative space implied by prisms and stations. A more promising way forward is to link time geography with modeling approaches such as activity-based travel demand models and agent-based models that naturally capture the emergent properties associated with individuals’ interactions over space and time. The natural fit between activity-based analysis and time geography is well recognized and well utilized in many models. However, the linkages between time geography and agent-based modeling are less developed, somewhat surprisingly given their complementarity.

**Time geography and virtual interaction**

Classical time geography recognizes the possibility of virtual interaction; for example, Hägerstrand (1970) shows a telephone call between two space–time paths. However, virtual interaction is neglected relative to travel and physical interaction in classical time geography. This is understandable given the era when time geography was initially formulated. But in the contemporary world virtual interaction is more pervasive and cannot be ignored when discussing physical mobility and activities.

Table 2 is a typology of possible communication modes based on space–time constraints.

<table>
<thead>
<tr>
<th>Temporal Presence</th>
<th>Spatial Telepresence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Synchronous</strong></td>
<td></td>
</tr>
<tr>
<td><em>SP</em></td>
<td><em>ST</em></td>
</tr>
<tr>
<td>Face-to-face</td>
<td>Telephone</td>
</tr>
<tr>
<td>conversation</td>
<td>Television</td>
</tr>
<tr>
<td></td>
<td>Radio</td>
</tr>
<tr>
<td></td>
<td>Teleconferencing</td>
</tr>
<tr>
<td><strong>Asynchronous</strong></td>
<td></td>
</tr>
<tr>
<td><em>AP</em></td>
<td><em>AT</em></td>
</tr>
<tr>
<td>Trail signs</td>
<td>Mail</td>
</tr>
<tr>
<td>Note left on</td>
<td>E-mail</td>
</tr>
<tr>
<td>office door</td>
<td>Newspapers</td>
</tr>
<tr>
<td>Geocaches</td>
<td>Webpages</td>
</tr>
</tbody>
</table>


*Synchronous presence* (SP) includes face-to-face conversations; this requires people involved to be physically present or proximal at the same time. *Asynchronous presence* (AP) requires co-location in space but not coincidence in time; examples include notes left on an office door or in a geocache. *Synchronous telepresence* (ST) requires coincidence in time but not in space; it includes media such as telephones, television, radio, and teleconferencing. *Asynchronous telepresence* (AT) does not require co-location in space or coincidence in time; it includes mail, e-mail, messages, and webpages. Classical time geography focuses on SP, recognizes but does not develop ST, and mostly ignores AP and AT.

There are at least two strategies for incorporating the wider range of communication modes into time geography. One strategy is to treat these as spatial and temporal relationships between space–time paths and prisms. SP requires the prism to intersect both spatially and temporally, while AP interactions only require spatial coincidence between the prisms (in other words, they share the same locations but at different times). ST requires temporal but not spatial coincidence.
(the prisms share some interval in time but not locations in space), while AT only requires one prism to precede the other in time.

Another strategy for extending time geography is to make communication technology explicit using time geographic entities and derive space–time constraints based on those entities. A portal is a type of space–time station (Figure 1) where a person can access communication services. It includes a point location, and an access range and time intervals when the service is available. Portals correspond to real world entities such as wired Internet connections (a point location with zero range), wireless access points, and cellular telephone base stations (both characterized as point locations with different ranges). An individual can access a communication service only if his or her paths or prism intersect with the service footprint of an appropriate portal. When this occurs, it generates a message window or an interval of time when a message could be sent or received. Message windows can be compared using temporal predicates such as before, during, and after to determine which communication events are feasible. This approach is consistent with the perspective that temporal constraints dominate spatial constraints on telepresence.

Discovering time geographic knowledge

As the size of mobility datasets become richer and voluminous, the challenges move beyond developing scalable techniques for computing low-level time geographic properties and relationships to discovering higher-level time geographic knowledge. Mobility mining is a set of activities and techniques for discovering novel knowledge hidden in moving objects data. This involves three major activities: (i) trajectory reconstruction and management; (ii) space–time knowledge discovery; and (iii) space–time knowledge delivery.

Trajectory reconstruction and management requires processing the raw mobility data to obtain the space–time paths of the mobile objects as well as data structures and access methods for processing these paths efficiently. These include specialized data warehouse designs, mobile object indexing methods, and semantic trajectory compression methods for reducing storage requirements. Preserving locational privacy through security, anonymization, and other protocols is also a concern.

Space–time knowledge discovery involves processing the trajectories to discover patterns such as clusters and behavioral rules. As noted, path and prism similarity methods can be used in conjunction with clustering methods to provide synoptic summaries of large mobile objects databases. Collection motion methods can also scale to large mobile databases. Another technique is space–time association rule mining that discovers rules describing how objects move among a set of regions over time. Sequence mining techniques search for temporal patterns in mobile objects data. Periodic pattern mining search for recurrent patterns in sequential data includes finding locations that are repeatedly visited by mobile objects and the recurrent movement patterns between these locations. Under specific conditions, causal relationships can also be inferred from sequential pattern in movement data.

Space–time knowledge delivery from movement data requires methods for transparent management of the knowledge discovery process and linking discovered knowledge to real-world semantics. A strategy for managing mobility mining is through visual analytics. Visual analytics is an extension of scientific visualization, but rather than seeking insights into data, visual analytics seek insights into how data are processed during exploration and analysis. Visual
analytics for moving objects include techniques for analyzing and comparing entire trajectories, the distribution of properties within trajectories, synoptic visualization, and investigating movement within geographic context.

Connecting movement patterns to real-world movement semantics requires a conceptual system for describing low-level patterns and their relationships with higher-level behavior. System dimensions can include movement parameters (direct measures and derivatives; Table 1), the number of objects (one, group, or a cohort of similar objects such as people aged 20–30 years), path type (semantically continuous or discontinuous), influencing factors (the object’s intrinsic properties, spatial constraints on movement, environmental factors, and the influence of other agents), and scale/granularity of the data (spatial and temporal scales, temporal granularity). Generic patterns are low level and shared with many different types of mobile objects. These can be primitive (based on a single mobility parameter) or compound (based on multiple primitives and/or inter-object relations). Behavioral patterns are high level and specific to the object type. Examples include pursuit/evasion, fighting, flocking, and leader/follower. Unlike generic patterns, behavioral patterns are open-ended and can be added to the framework as more empirical movement patterns are discovered in different domains (Dodge, Weibel and Lautenschütz 2008).

SEE ALSO: Accessibility, in transportation planning; Agent–based modeling; Behavioral geography; Big data; Data model, moving objects; Geographic data mining; Geographic information science; Geographic information system; Geolocation services; GIS for transportation; Graph theory; Information and communications technology; Information technology and mobility; Network analysis; Regional geography; Representation: time; Representation: trajectories; Transport geography; Transport networks; Transport technology; Uncertainty

References


Further reading


Time–space convergence

Donald G. Janelle
University of California, Santa Barbara, USA

Time–space convergence (TSC) refers to the rate at which the travel time between places declines in response to transport innovation (Table 1). Analogous to the velocity concept, it is a surrogate for the impact of technological investments in overcoming distance. The original formulation of TSC was linked to the concept of spatial reorganization of settlement systems and socioeconomic institutions (Janelle 1969) and to the enhancement of place utility (the ability of a place to capitalize on its location in meeting its needs and to compete successfully within its local region and more broadly within national and global contexts).

As a component of the process of spatial reorganization, TSC is seen as a significant motivating factor in reshaping the relationships between places and in altering socioeconomic landscapes at local through global scales. The reciprocal concept of human extensibility is, in essence, the dual of time–space convergence, in that it considers how space-adjusting technologies (transportation and communication) are used by individuals and institutions to extend their presence and influence beyond their immediate geographical locations.

TSC has been used to evaluate equity in accessibility gains through public transport services at urban and regional scales and to investigate information diffusion and the historical transformations in the hierarchy of places in shaping emergent urban systems. Cost–space convergence was documented by Abler (1975) and convergence measures have been used to document advances in place connectivity through communication and virtual technologies. TSC is often characterized as one of the forces contributing to time–space compression – a separate concept introduced by David Harvey to invoke the experiential impress of capital accumulation on people and places.

Average rates of change (as shown, e.g., in Table 1) do not fully capture the complexities of the convergence process. Convergence is uneven across space and time because of variation in the abilities of places to attract transport and communication investment, and because of temporal discontinuities in the process of technological innovation, thereby benefiting some places over others. Whereas convergence takes place primarily at regional and global levels, time–space divergence is more common at urban scales owing to congestion. Also, regional variation in patterns of convergence and divergence frequently align with the socioeconomic status of individuals, groups, and places. These TSC attributes and their implications for research in the social sciences are reviewed in Janelle (2014).

**SEE ALSO:** Accessibility, in transportation planning; Distance; Spatial organization and structure; Transport technology
### Table 1  

<table>
<thead>
<tr>
<th>Travel mode</th>
<th>Year</th>
<th>Travel time (minutes)</th>
<th>Average convergence (min p.a.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stagecoach</td>
<td>1776</td>
<td>5760</td>
<td></td>
</tr>
<tr>
<td>Train</td>
<td>2010</td>
<td>260</td>
<td>5500 min/234 years = ≈23.5</td>
</tr>
<tr>
<td>Motorway</td>
<td>2010</td>
<td>410</td>
<td>5350 min/234 years = ≈22.9</td>
</tr>
</tbody>
</table>

#### References


Tobler’s first law of geography

Nigel Waters
University of Calgary, Canada

This discussion of Tobler’s first law (TFL) of geography begins with a formal definition and an account of its origins in an article published by Tobler in 1970. This is followed by a section that parses and explains the implications of all the components of TFL: whether it was/is the “first” law, whether it is a “law,” and how distance and relatedness might be measured. The extent to which TFL can be accommodated within the classic models of human and urban geography is considered next. TFL can be made operational through spatial interpolation techniques and these are explained in the ensuing section. TFL has been considered to be highly original but the subsequent segment examines “first laws” in other disciplines. Finally, consideration is given to geographical contexts and circumstances where TFL fails and where other intellectual frameworks and methodologies are more informative. Throughout the discussion reference is made to observations and commentaries of the Forum published in the Annals of the AAG (Association of American Geographers) in 2004 which remains the most detailed examination of TFL to date (Sui 2004; Barnes 2004; Goodchild 2004; Miller 2004; Phillips 2004; Smith 2004). It seems fitting to allow Tobler to have the last word and so concluding section is devoted largely to Tobler’s “Reply” to the participants in the AAG Forum (Tobler 2004).

Definition of Tobler’s first law of geography

Tobler’s first law of geography was originally “promulgated” (this seems more appropriate than the more mystical “invoked” preferred by Sui (2004)) in a paper published in the journal Economic Geography (Tobler 1970). Tobler’s article described a model of population growth for Detroit from 1910 to 2000. The model produced a simulation that was presented as a wire-frame, three-dimensional block diagram. Tobler suggested that if the simulated surfaces were viewed at 16 frames per second then a new surface should be generated for every month of the simulated time period in order to maximize viewing effectiveness. Many of the spatial characteristics of the model were provided in a paper published a year earlier and cited in the 1970 article. A total of 1150 grid cells were used, each being 1.5 miles square. To operationalize his simulation, Tobler stated that he would “invoke the first law of geography: everything is related to everything else, but near things are more related than distant things.” Tobler had earlier argued that “everything is related to everything else” but he added the additional proviso to make the model more localized and thus more tractable. Temporal influences on the population at a particular point in time came from the population during the previous decades, declining by 50% for each passing decade. Spatial influence came from the adjacent cells and this influence was modeled.

Tobler’s paper and his first law have been immensely influential. His 1970 paper has now been cited over 3700 times in the academic and research literature. In June 2003, a Google search
TOBLER’S FIRST LAW OF GEOGRAPHY

of “Tobler’s first law of geography” produced “at least 150 returns” (Sui 2004, 274), but in April 2016, the same search produced almost 8000 returns. Interest in TFL appears to have been growing exponentially.

The law unmasked, deconstructed, and parsed

To understand the importance of Tobler’s law we must question what exactly a first law in the discipline of geography might imply, and ask if other academic disciplines have laws and, more specifically, if they are the most important laws, that is, “first laws.” We should ask whether this was the “first” such law in geography. We need to know what is meant by “everything.” We need to ask how we determine whether two “things” are “related” and what we mean by a “thing.” Finally, we should make explicit how we measure distance and what is “near” as opposed to being more “distant.”

Determining the importance and law-like status of TFL

When Tobler said that this was “the first law of geography,” he was not suggesting that this was the first law to be declared. He was arguing that this was the most important law in geography and that distance was the most important variable governing the influence of one entity on another. In his case study of Detroit, the variable of interest (the “thing”) was persons aggregated into populations, but the implication of TFL is that most geographic variables would show a distance-decay effect.

It is important to determine whether others had suggested there should be laws in geography. Therefore the question must be asked whether this was the first time any scholar had suggested that there be a “first [most important] law” in geography. The answer is that possibly it was, but in the years before 1970 there had been an intense interest among scholars in defining laws and theorems in geography. This concern for formulating laws was a consequence of the increasingly bitter debate between Richard Hartshorne and William Bunge about the nature of geography (Lewis 1965). Hartshorne maintained that geography’s role was to describe the uniqueness of geographical regions and that, according to Lewis (1965, 24), “geographers study individual cases rather than construct scientific laws.” Bunge argued that entities studied in geography were not unique, in the sense that they had similarities that could be studied and characterized and that these similarities could be measured and exploited allowing for the construction of functional models that produced informative explanations and predictions, and therefore a scientific approach to geography. Bunge maintained that the traditional science disciplines did not always seek to argue that cause and effect or deterministic laws were essential. Thus, it might be reasoned that a stochastic law could yield useful predictions in a more scientific, mathematically formulated approach to geography. This is fundamentally the view of Miller (2004, 284), namely “that science accepts the concept of empirical laws, or compact descriptions of patterns and regularities … [that] are not required to be immutable truths.” Phillips (2004, 290) also argues for a less rigid definition of a “law,” citing Cole and King’s classic text, Quantitative Geography, which provides six definitions of a law that “are used in geography, several of which would apply to TFL.” Neither Miller nor Phillips is particularly concerned with the fact that Tobler claims his proposition to be a “law,” but for Smith (2004, 294) this is a primary concern. He argues that had Tobler claimed his proposition to be a rule rather than a law then it would have attracted little attention.
TOBLER’S FIRST LAW OF GEOGRAPHY

Measuring distance

To understand TFL it is important to recognize that there are various ways to measure distance. For many years, geographers have recognized that models and analysis can be formulated in both Euclidean and network space. From the earliest days, location-allocation models have been formulated in both types of space. For example, Leon Cooper formulated his solutions in Euclidean or continuous space while Seifollah Louis Hakimi determined optimal facility locations on a graph or network that is in discrete space (Waters 1999). Most of the classic, theoretical models of human geography (see discussion below) were formulated in continuous space. More recently, due to the availability of software such as SANET (Spatial Analysis on a Network) and the network analyst routines in ArcGIS, human geographers have evaluated distances along networks such as transportation routes. SANET, for example, permits network autocorrelations to be evaluated allowing the researcher to determine if “near things,” such as traffic accident densities, are “related” or spatially autocorrelated along the links of a transportation system.

Regardless of whether we choose Euclidean distance or network distance, it is crucial to be able to discriminate between what is a near thing and what is a more distant thing. Measuring distance is much more complicated than simply making a choice between continuous space and discrete space. It is so much more complicated that in 2006 Michel Marie and Elena Deza produced a comprehensive and exhaustive Dictionary of Distance. This was quickly replaced, in 2009, by an even more extensive survey, the Encyclopedia of Distances (EOD), and a second edition followed three years later (Deza and Deza 2012). Deza and Deza demonstrate that measuring or estimating the distance between two points is of fundamental interest to researchers in a wide variety of disciplines and that these disciplines include...
TOBLER’S FIRST LAW OF GEOGRAPHY

those that are interested in using estimates of
distance in the physical world where distance
may be measured in spatial units (e.g., miles) or
in time or cost of travel, and those disciplines that
use distance as a measure of similarity or related-
ness in a virtual or variable space (the latter being
commonly used for finding clusters of observa-
tions). Of special interest to geographers in the
EOD are the sections on: Distances in the Inter-
net and Web (characterized by links in virtual
networks); Distances in Ecology, Biogeography,
and Ethology; Distances in Geography; Distances
in Economics and Human Geography (p. 561
references Tobler’s first law); and Distances and
Similarities in Data Analysis (especially, for
example, cluster analysis – see below).

Measuring similarity, dissimilarity,
and association in space

If a researcher is to determine that “near things
are more related than distant things” then
the researcher must be able to establish this
relationship using an objective, reproducible
statistic that measures “relatedness.” This statistic
might be some form of correlation coefficient,
such as Pearson’s product moment correlation
coefficient, $r$, or Spearman’s rank correlation
coefficient, $r_s$; a measure of association, such as
Goodman and Kruskal’s $\lambda$ or Cramer’s $V$; or a
measure of dissimilarity between two locations,$x$ and $y$, determined as the sum of the values
obtained from their differences on each variable
$i$ standardized to have a mean of zero and a
standard deviation of 1.0, to insure that each
variable has potentially the same influence in
calculating the dissimilarity (relatedness) between
the locations. With correlations and measures of
association (i.e., measures of similarity), a larger
number indicates that the two locations are more
positively related. With dissimilarity measures, a
larger number means that the two locations are
less “related,” that is, further apart in a single or
multivariable space. Correlations and measures of
association may be used as measures of strength
of the relatedness of a variable or variables at
two locations, and are usually constrained to a
range of $+1.0$ through $0.0$ to $−1.0$ or from $+1.0$
through to $0.0$, allowing for easy comparisons
of, in this case, the relatedness between different
pairs of locations. Correlations and measures of
association are commonly accompanied by an
associated statistical test (for $r$ and $r_s$ this is a $t$
test and for $\lambda$ and $V$ a chi-squared test or $G$
test) allowing the researcher to determine whether an
observed statistic is statistically significant or not.

Gould (1970) was the first to cite TFL in a
paper, in the same issue of Economic Geography.
He noted a problem with traditional measures
of relatedness such that, when used with spatial
data, they frequently violated the assumption of
independence. Indeed TFL states that spatial data
are defined by their spatial dependence or spatial
autocorrelation. Researchers in geography have
had two responses to this conundrum. We can
use TFL and the spatial dependence in our data as
an opportunity to model spatial autocorrelation
as in kriging (see below), or by calculating and
interpreting local indicators of spatial association
(LISA) statistics to gain insights as to what this
implies for our understanding of the variable(s)
under investigation. Alternatively, we can remove
the spatial autocorrelation using more recent
methodologies such as spatial regressions, as
exemplified in software packages such as GeoDa.
An alternative approach to the modeling of
near things is to use geographically weighted
regression (GWR) models (and, more recently,
geographically weighted discriminant analysis
(GWDA) and geographically weighted principal
component analysis (GWPCA)), which allow the
researcher to model local influences using dis-
tance or a predefined number of points that spec-
ify exactly what is “near” and what is “related.”
The Center for Spatial Studies at the University of California at Santa Barbara maintains a webpage on spatial statistics and spatial econometrics software resources including GeoDa. For R-coded software packages that perform spatial statistical analysis, researchers can access the Comprehensive R Archive Network (CRAN) where they can download the R software for various operating systems (Linux, OS X, and Windows) as well as the CRAN Task View for the analysis of spatial data, a website maintained by Roger Bivand (Bivand 2015). This is the most comprehensive set of routines for the analysis of spatial relatedness available anywhere and includes R code for most of the procedures mentioned here. Spatial autocorrelation is thus both a problem and an opportunity for the researcher using geographical data, but with so many new spatial analytical software tools now available and often embedded in geographic information system (GIS) packages such as ArcGIS, the former has become increasingly less of a concern.

Tobler’s first law and the classic models of human and urban geography

The classic models of human geography, including Von Thünen’s model of agricultural activity around a centralized market town in an isolated state, Christaller’s central place theory (CPT) defining the settlement geography of hierarchical urban systems, and Lösch’s *The Economics of Location*, all produce economic landscapes where TFL is consistently shown to be insufficient. Distance is important in Von Thünen’s model, but the model produces rings of agricultural specialization that are homogeneous internally but where the land use between adjacent rings (land-use patches which are indeed very “near things”) may be quite unrelated. CPT produces a hierarchy of settlements superimposed on a background of uniform agricultural activity. Thus a large urban center will be located within a rural setting and the largest centers will be spaced the furthest apart, and it is these centers where similar “high-order” goods and services will be found. A hierarchy is inimical to TFL, at least in terms of *geographical* distance, and so are borders and boundaries (see below for further discussion of these topics). Lösch’s analysis produced sectors of economic activity which alternated between high and low densities of settlements and thus, according to Lösch’s theoretical analysis, adjacent sectors would be less similar in this respect and, presumably, in their population and other socioeconomic characteristics.

Geographers and other social scientists have identified a number of models to characterize the internal structure of cities. The best known of these are the concentric ring model of Burgess, the sector model of Hoyt, and the multiple nuclei model of Harris and Ullman, all three of which were reconciled by Bourne and Murdie’s factor analysis of the internal structure of the city of Toronto, Canada (Perry, Crew, and Waters 2013). It might be argued that TFL applies within the Burgess rings, or within Hoyt’s sectors, or inside Harris and Ullman’s ethnically homogeneous nuclei, but again, on the borders of these zones, there might be radically different residential communities or zones of economic activity living adjacent to each other. These inconsistencies in the validity of TFL are in part a function of scale. When a researcher concentrates on the “big picture” and “zooms out” from the map, variables that exhibited a gradual decline with increasing distance from a given location may, at a more detailed scale, simply appear as uniform concentrations with sharp boundaries where the spatial autocorrelation declines to zero. Thus first-order trends may appear as second-order concentrations.
TOBLER’S FIRST LAW OF GEOGRAPHY

Less theoretical, more applied, empirical models have fared better, especially those used in transportation geography. Step 2 of the four-step transportation planning model (Waters 1999) is based on the premise that commuters moving out of trip production zones where they live and toward trip attraction zones where they work will attempt to minimize the length of their journey to work. Trip distributions are a practical example of TFL that is based on the concept of a doubly constrained, spatial interaction model where trip numbers are directly proportional to the product of the number of commuters in a production zone and the number of jobs in an attraction zone, and, most significantly, inversely proportional to some power of the distance between the two zones. Put simply, the shorter the distance between the production and attraction zones, eteris paribus, the more commuter trips will result; all traffic analysis zones (TAZs) are related but near zones are more related than distant zones.

Spatial interpolation and TFL

Spatial interpolation is the science of estimating the values of a surface at unsampled locations within a region based on a set of sampled values at various locations and is commonly found in GIS software such as ArcGIS and Idrisi. The value of a surface at an unsampled location is dependent on the values at nearby sampled locations. Spatial interpolation is a way to operationalize TFL. It assumes that the values of a surface exhibit some form of spatial autocorrelation.

Distance-weighted averaging

A common method of spatial interpolation is to use a distance-weighted average. Decision variables in a distance-weighted average involve the number of sampled values to include, usually four, six, or eight. In Tobler’s original words, this determines just how “parochial” or localized the estimating function should be (in kernel-based spatial interpolation the kernel can be limited by distance or number of points used) – the more localized, the more irregular the surface of estimated points. A second decision variable is whether there should be a spatial bias in choosing the nearest four, six, or eight points, that is whether they should be the nearest point in each one of four quadrants, six sextants, or eight octants, respectively, the object being to reduce directional bias in the estimation. The distance-weighting function may be exponential or simply 1/d where d is the distance from the unsampled point or 1/d^n where n is some power of distance. This allows “near things to be more related than distant things.” The degree to which a distance-weighted interpolation is parochial depends on both the number of points included in the averaging process and the distance function used. In theory, all sampled points might be included (where everything is related to everything else) and this would produce a relatively smooth surface, the degree of smoothness being a function of how quickly the influence of distance declined. Trend surface analysis uses a global polynomial with two independent locational variables (X and Y; easting and northing; longitude and latitude) to estimate values of the dependent variable (elevation, population, as it was in Tobler’s 1970 paper, or some other spatially varying variable) using a single equation. The polynomial can be first, second, or third order, or some higher number producing surfaces with additional inflections each time the order increases. Tobler referenced Richard Chorley and Peter Haggett’s early use of such trend surface models.
Kriging

Kriging is another commonly used method of spatial interpolation. Kriging allows the researcher more objective control over the variables implicit in TFL, more objective control, that is, than the subjective choices used in distance-weighted averaging. Specifically, in kriging everything is not “related to everything else” because the “relatedness” ceases when the range is reached. The range is a distance over which points no longer covary. The extent to which points covary is measured by the covariance function that is estimated by a variogram. The variogram can take a large variety of functional forms, similar to the changing influence of distance in a distance-weighted averaging. The variogram is an estimation of just how strongly “near things” are related. It also shows how quickly “relatedness” declines to zero. Universal kriging (as opposed to simple kriging) allows the geographic researcher to model the influence of spatial trends in the data since “near things” in one direction may be “more related” than near things in another direction. For example, data in a north-south direction might be more highly correlated than in an east-west direction.

In many instances, near things may be related not only spatially but through the influence of one or more variables. Thus population density in a city such as Detroit might be influenced by many variables. Yeates (1965) found that land values at any given location in Chicago were influenced by distance from the central business district, distance from the nearest regional shopping center, distance from Lake Michigan, distance from the nearest elevated subway system, population density, and percentage of nonwhite residents at the given location. Thus the land values at any location might be “related” to the land values at nearby locations, as TFL states, but this relationship would be influenced by these other variables and would thus be better modeled by co-kriging than by simple kriging or regression as suggested by Yeates.

First laws in other disciplines

The idea of a first law may be found in other disciplines, often with partially humorous intent. Examples abound on the Internet: “The first law of Politics is that it is always about politics”; the first law of sociology: “Some do, some don’t.” Science provides some rigorous, universally accepted examples, such as the first law of thermodynamics. Geologists have given the name “uniformitarianism” to the idea that the natural laws governing geological processes that operate today also operated throughout geological history. More wittily, one researcher has suggested that the first law of geology should be that “the rocks remember while liquids and gases forget.” Since kriging and, to a large extent, geostatistics were developed in geology, it is not surprising that TFL has recently been claimed by geologists, without acknowledgement: “According to the first law of geology ‘everything is related to everything, but near things are more related than distant things’” (Bayraktarli and Faber 2011). It has been suggested that biology’s first law should be that there is a tendency for diversity and complexity to increase in evolutionary systems. It is interesting to note that evolution itself is hindered by TFL, for the primary method of speciation, allopatric speciation, requires the isolation of two breeding populations by geographical barriers (mountains, rivers, oceans), migration, or habitat fragmentation. Perhaps TFL should be classified as a dangerous idea, since, like globalization, it inhibits diversity and complexity. Tobler (2004) himself cites a number of “first laws” from other disciplines (some closely allied to geography and some not) including Ernst Georg Ravenstein’s well-known laws of migration and George Zipf’s
**TOBLER’S FIRST LAW OF GEOGRAPHY**

law, which stated that the frequency of a word in a large corpus of text would be inversely proportional to its frequency rank. This is also known as a power law or the rank-size rule, the latter suggesting perhaps more reasonably a diminished status. Power laws have been applied extensively to the population ranks of cities in numerous countries and to the study of social networks in geography to detect scale invariance, but unfortunately they can be derived in many different ways and therefore tell us little about generative processes (Waters 2013).

**Spaces where Tobler’s first law fails**

**Borders, boundaries, and barriers**

TFL will fail (or provide an incomplete explanation) wherever there are natural barriers similar to those that have aided the process of evolution discussed above. It will also fail wherever there are political or administrative barriers. A striking example may be found in the land use on the border between some parts of Haiti and the Dominican Republic, on the island of Hispaniola. Even the land use north and south of the forty-ninth parallel of latitude, forming the border between the United States and Canada, may show dramatic differences. In many countries, census divisions and political constituencies have boundaries that are deliberately constructed to maximize internal homogeneity and external heterogeneity in terms of socioeconomic characteristics and voting behavior. Ordinarily, having a political unit composed partly of an urban area and partly of a rural area would be considered to be gerrymandering. Finally, some countries may erect barriers between themselves and neighboring countries. Examples include the former Berlin Wall separating East and West Berlin, the Mexico-US barrier or border fence, and the Israeli West Bank and other separation barriers. The existence and impact of such separation barriers have been prominent in fiction, including the celebrated work of Gloria Naylor who, in her first and second novels respectively, located the African American communities of Brewster Place and Linden Hills bordering each other but separated by a wall. “Linden Hills is a posh upper-middle-class settlement, Brewster Place the last stop on the road to the bottom in American society… with a great distance between them even as they are both black” (Christian 1990). TFL stops at the wall in both the real and the fictional world. TFL does not readily explain ethnic segregation (Perry, Crew, and Waters 2013). Having noted these examples at micro-geographical scales, it is important to state that some researchers have found evidence suggesting “that national cultures are organized geographically; similar cultures are found in neighbouring countries, and dissimilar cultures in countries that are far apart” (Gelade and Dobson 2009).

**Hierarchical spaces and networks**

Diffusion and migration processes often progress hierarchically. Thus a diffusion of an innovation may be passed and adopted from one individual of influence to another. Once these individuals have adopted the innovation it may then be adopted by a process of filtering down through the hierarchy of their networks of influence. Such hierarchical diffusion may also occur across space. Thus an innovation may be adopted by higher-order centers in a central place hierarchy and may then, in a second wave of adoptions, be embraced by individuals in lower-order centers. Indeed, the whole adoption process might filter down through the entire urban system in a hierarchical fashion. Geographical research (Bowen and Laroe 2006) has shown that the temporal spread of diseases such as severe acute respiratory syndrome (SARS), which, in 2003, spread from...
its origins in China to 25 other countries and to Taiwan, is primarily due to airline network accessibility. Bowen and Laroe estimated the parameters of a multiple regression model where the dependent variable was the number of days from the first case of the disease in a newly SARS-infected country (SIC) until the disease was contained on July 5, 2003. Thus the higher the number of days, the earlier the outbreak, that is to say, the more quickly the disease moved from China, its origin, to the country concerned. Six independent variables were included in the regression model: the number of Chinese ancestry residents of a country as a percentage of the total population; a binary variable that was assigned a 1 for the 11 countries with the highest foreign direct investment in China and a 0 for the remaining countries; per capita gross national income; a country’s 2003 population; a measure of accessibility from Beijing using airline network schedules; and distance between the SIC and China. Only the airline network accessibility and the number of Chinese ancestry residents as a percent of the total population were statistically significant. Distance itself was not, and so it can be argued that TFL did not apply to the diffusion of the 2003 SARS outbreak unless we measure “distance” as a function of network accessibility. In addition, the medical geographer would need to take past migration patterns into consideration as well. Bowen and Laroe cite other examples of disease propagation via transportation routes, including the transmission of the fourteenth-century outbreak of bubonic plague along sea trade routes and three outbreaks of cholera in the United States during the middle of the nineteenth century along existing trade routes. A county not on a trade route would not be infected, regardless of “nearthiness.”

Social media and social networks

Social media (SM) and social networks (SNs) have weakened the influence of distance, contributing to the so-called death of distance and the weakening of TFL. Waters (2013) has reviewed both the literature on the death of distance, including Cairncross’s book on this topic, and the research on the impact of SM and SNs on distance-decay functions within social networks. The primary attraction of SM and SNs, and even the Internet itself, has been to weaken or eliminate the influence of geography and distance and, if this were true, Waters (2013) has argued that a second law of geography might replace the first: “Everything is connected to everything else, but things more closely connected [in network space] are more related — and geography may well be irrelevant.” In SN space there is also a problem with defining what is local. Christakis and Fowler (2011) have discussed this issue and describe research that demonstrates that it is our friends’, and friends of friends’, and friends of friends of friends’ behavior that affects us, and it is these linkages that determine whether we are obese, happy, or smoke, or have heart disease, and presumably other characteristics as well. We may be connected to everyone else in the network, but influence ceases to be significant with three degrees of separation (Waters 2013). Sociologists have identified homophily as “the principle that a contact between similar people occurs at a higher rate than among dissimilar people” (McPherson et al. 2001). Although their observation relates to social networks they argue that Geography is the most important of a number of influential independent variables.

However, the question remains as to whether distance is now irrelevant. Or perhaps more alliteratively: “Is spatial special when it’s social?” (Waters 2012). The existing research (which is expanding rapidly as SN companies such
as Facebook and LinkedIn utilize their own research teams) implies that the more “social” as opposed to “professional” the SN is the more distance still matters. Tobler (2004, 308) himself suggested a second law that “the phenomenon external to [a geographic] area of interest affects what goes on in the inside” – essentially a problem of system closure presumably because “everything is related to everything else.”

Various researchers (for a complete discussion see Waters (2013)) have fitted distance-decay functions to “friends” and connections within SNs. For example, the probability that two cell phone users are connected is proportional to the Euclidean distance separating them raised to the power $-2$. This is the classic distance-decay function used in a gravity model and is reminiscent of the findings of Mackay (1958). Mackay’s seminal study of long-distance phone calls between cities in the Province of Quebec, Canada, showed a distance-decay exponent of $-0.9$, but between cities in the Province of Ontario it was $-1.7$, a steeper distance-decay function. Rapid drop-offs in call interactions occurred between Quebec-Ontario city pairs in which case the telephone traffic was about one-fifth what it was between pairs of cities in Quebec. An even more dramatic drop-off was noted between Quebec-US city pairs, where the traffic was about one-fiftieth of the Quebec pairs. Such differences presumably occurred because there were fewer contacts across provincial and national borders and because of higher long-distance charges. Cell phone rates may also influence such long-distance interactions and cross-jurisdictional calls. These provincial and national borders do not prevent interaction; rather, they modify the process making it impossible to use a single function to model it. Waters (2013) notes that one study of the SNs, LiveJournal, determined that the average user had eight friends and that, again on average, 5.5 were “geographically influenced” and lived nearby with a distance-decay function where the exponent of distance was approximately $-1$, while the remaining 2.5 friends were a result of nongeographic processes, the former geographically influenced connections being a partial vindication of TFL. Other research cited by Waters also suggests that SNs are influenced by a mix of geographic and social influences. Such research has suggested that “node locality” or network-based metrics be combined with “geographic clustering metrics” (and presumably socioeconomic factors similar to the past migration patterns elucidated by Bowen and Laroe) to provide a modeling framework with exceptionally high explanatory power but lower generality than Tobler’s Detroit model.

Tobler: the last word on Tobler’s first law

Daniel Sui, the convener of the 2003 AAG panel session on TFL, asked Waldo Tobler to review and respond to the participants’ written reviews of his law. Tobler (2004) introduced his comments by saying that the goal of his model and of TFL was to emphasize simplicity and therefore generality. This allowed Tobler to predict population growth in Detroit, a modest goal but a goal that is most useful in transportation planning and for urban budgeting wherever these budgets are based on population numbers. Tobler states that Barnes was correct in observing that he was “restricting … [himself] … to local effects.” Unfortunately, this is not what the law states when it declares “that near things are more related than distant things,” implying a gradual decline in relatedness with no sharp break. However, spatial interpolation methods such as kriging, with its concept of the range, spatial smoothing’s use of kernels and geographically
weighted regressions, discriminant analysis, and principal component analysis all allow analysis to be restricted to the local. Tobler confines his subsequent comments to three issues: (i) What defines a law and does TFL satisfy the definition? (ii) Is it true “everything is related”? (iii) How should “near” be defined?

Laws

Unlike Smith (2004), Tobler is satisfied with, and quotes, Richard Feynman’s relatively undemanding definition of a law from his 1967 book, *The Character of Physical Law*, as an educated guess, where the consequences have been computed and then tested by experiment, experience, or observation. If the test conforms to reality then it is a law, and if not, the scientist must guess again. This is a very low bar and would not satisfy Smith’s criteria described above. Tobler’s definition of a law is simply based on predictive power, not explanation, but this, according to Tobler himself, and also Miller (2004), was enough for Newton’s law of gravity to be universally accepted. However, Tobler does not appear to have validated his own predictions for his Detroit population model.

Tobler admits that other well-known social scientists and geographers have found the idea of laws in social science to be simply unacceptable (e.g., Anthony Giddens and Bent Flyvbjerg, references in Tobler (2004)). Nevertheless, it is difficult not to conclude that often what passes for a law might be better referred to as a principle (e.g., Zipf’s principle of least effort), or perhaps a rule (e.g., the rank-size rule now known as one form of the power law), or simply an empirical generalization such as Roger Shepard’s observation concerning multidimensional scaling: “if a solution exists, probably it exists in two dimensions” (cited in Tobler 2004).

Everything is related

Tobler concedes that everything is not usually related to everything else and, if it is, then the influence of the most distant relationships is usually not significant or meaningful. So TFL is not strictly correct and geographers have long accepted this in their modeling activities by applying locally constrained models, measures of association, and spatial interpolations including LISA, GWR, GWDA, GWPCA, and variations on kriging and distance-weighted averaging, all of which have been mentioned above.

Near things

Tobler admits that his use of the term “near things” was “equally ambiguous” (Tobler 2004, 306). He noted that distance can be measured in a variety of ways but was, of necessity, unaware of the Dezas’s contribution since both their *Dictionary* and *Encyclopedia of Distance* were published subsequent to Tobler’s “Reply.” Tobler does, however, note that the variogram (see above) is a way to assess both relatedness (covariance) and nearness (range). Tobler commends Goodchild (2004) for speculating on other “laws” and “principles” that are fundamental to geographic information science and for contemplating situations where TFL does not apply, which might include the chaotic drainage in the landscapes of southeastern Alberta, South Dakota, and New Mexico.

Tobler concludes, modestly, by noting how the eminent statistician Sir Ronald Fisher made a similar statement to TFL in 1935. He also expresses his surprise that TFL should have aroused so much debate when he was “just having fun doing an animation” (Tobler 2004, 308).

**SEE ALSO:** Borders, boundaries, and borderlands; Distance; Distance decay;
TOBLER’S FIRST LAW OF GEOGRAPHY

Geography of evolution; Geostatistics; Interpolation: inverse-distance weighting; Interpolation: kriging; Location-allocation models; Neighborhood; Network analysis; Spatial analysis; Spatial epidemiology; Urban geography

References


TOBLER’S FIRST LAW OF GEOGRAPHY


Further reading

Topological relations

M. Andrea Rodríguez
Universidad de Concepción, Chile

Topology is typically seen as the most abstract spatial structure. It is considered essentially qualitative because it is independent of quantitative measures such as the distance or length of common boundaries. Topological relationships, such as “C is inside D” or “C touches E,” are those relations that are invariant with respect to continuous transformations of the underlying space, namely, translation, rotation, and scaling. They represent basic notions of connectivity between spatial objects and have spurred much research in the spatial information community.

Formalization of topological relations

In the formalization of topological relations there are two well-known approaches: axiomatic theories (generally logic of first order) and point-set theories.

The region connection calculus (RCC) by Randell, Cui, and Cohn (1992) is a well-known axiomatic theory based on a single binary relation and a set of axioms. The primitive “connected” relation $C(x, y)$ is true if the topological closures of the nonempty spatial regions $x$ and $y$ are connected. Then, other relations can be formally defined as expressions in first-order logic using the relation $C$. This formalization derives a set of eight topological base relations (Table 1), which are called the RCC-8 set of relationships. While this theory is rich in its expressive power and has captured the attention of the research community, it falls short under considerations of computational implementation.

The same set of topological relations between connected regions (i.e., regions with connected boundaries and connected interiors) was independently proposed by Egenhofer and Franzosa (1991) as being very important for geographic information systems (GIS) and spatial databases. In point-set theory, an object is represented by its point-set topological elements. The basic strategy is to determine the point-set intersection between an object’s topological elements and then to derive constraints to filter out impossible cases.

Egenhofer and Franzosa defined binary relations according to the emptiness and nonemptiness of the four intersections between interiors (○) and boundaries (δ) of regions; hence, it is called the 4-Intersection Model. The empty/nonempty topological invariant is the simplest and most general invariant, so any other invariant may be considered a more restrictive relationship classifier.

It is straightforward that the sixteen possible intersections between interiors and boundaries form a jointly exclusive and pairwise disjoint (JEPD) set because, for any pair of regular regions, one and only one of the relations holds. However, if spatial regions are defined as regular nonempty proper subsets of $\mathbb{R}^2$ with connected interiors, then only eight of the sixteen relations are realizable in the plane. This set of eight relations is called the high resolution set or the base set of topological relations. The eight relations are: “$x$ is disjoint or disconnected from $y$” (Disjoint), “$x$ touches $y$” (Meet), “$x$ overlaps $y$” (Overlap),
**Table 1** Topological relations defined by RCC (Randell, Cui, and Cohn 1992).

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC(x, y)</td>
<td>$\equiv \neg C(x, y)$</td>
<td>$x$ is disconnected from $y$</td>
</tr>
<tr>
<td>P(x, y)</td>
<td>$\equiv \forall z[C(z, x) \rightarrow C(z, y)]$</td>
<td>$x$ is part of $y$</td>
</tr>
<tr>
<td>PP(x, y)</td>
<td>$\equiv P(x, y) \land \neg P(y, x)$</td>
<td>$x$ is a proper part of $y$</td>
</tr>
<tr>
<td>EQ(x, y)</td>
<td>$\equiv P(x, y) \land P(y, x)$</td>
<td>$x$ equals $y$</td>
</tr>
<tr>
<td>O(x, y)</td>
<td>$\equiv \exists z[P(z, x) \land P(z, y)]$</td>
<td>$x$ overlaps $y$</td>
</tr>
<tr>
<td>PO(x, y)</td>
<td>$\equiv O(x, y) \land \neg P(x, y) \land \neg P(y, x)$</td>
<td>$x$ partially overlaps $y$</td>
</tr>
<tr>
<td>DR(x, y)</td>
<td>$\equiv \neg O(x, y)$</td>
<td>$x$ is discrete from $y$</td>
</tr>
<tr>
<td>EC(x, y)</td>
<td>$\equiv C(x, y) \land \neg O(x, y)$</td>
<td>$x$ is externally connected to $y$</td>
</tr>
<tr>
<td>TPP(x, y)</td>
<td>$\equiv PP(x, y) \land \exists z[EC(z, x) \land EC(z, y)]$</td>
<td>$x$ is a tangential proper part of $y$</td>
</tr>
<tr>
<td>NTPP(x, y)</td>
<td>$\equiv PP(x, y) \land \neg \exists z[EC(z, x) \land EC(z, y)]$</td>
<td>$x$ is a nontangential proper part of $y$</td>
</tr>
</tbody>
</table>

“$x$ is equal to $y$” (Equal), “$x$ is covered by (or a tangential proper part of) $y$” (Covered_by), “$x$ is inside (or a nontangential proper part of) $y$” (Inside), and the converse of the two latter relations (Covers and Contains). The definitions can be found in Table 2 and are illustrated in Figure 1. For example, for nonempty regions $x, y$, Meet($x, y$) is true if and only if all of $\delta(x) \cap \delta(y) \neq \emptyset, x^c \cap y^c = \emptyset, \delta(x) \cap y = \emptyset$ and $x^c \cap \delta(y) = \emptyset$ simultaneously hold.

As the eight relationships correspond to the eight base relationships of the RCC-8 theory, there seems to be a natural agreement about what is considered to be a high level of resolution of topological relations. However, there are other sets of topological relations that are obtained by the disjunction of high resolution relations. For example, a relation Within captures the Equal, Inside, or Covered_by relation of the high resolution set. A smaller set in this condition is said to be a coarse resolution set of topological relations.

The 4-Intersection Model was later revised to incorporate the analysis of regions, lines, points, and combinations of them using an algebraic topology with primitive geometric objects called “cells” and their aggregation into cell complexes. In this model, a point is a 0-cell in $\mathbb{R}^2$, a line is a sequence of connected 1-complexes in $\mathbb{R}^2$, and a region is a 2-complex in $\mathbb{R}^2$ with a nonempty connected interior. The binary topological relation between two cells is derived from the
set intersection of the interior (\(\,\)), boundary (\(\partial\)) and exterior (\(\,\)) of cells, which was called the 9-Intersection Model (equation A).

\[
M(x, y) = \begin{pmatrix}
\delta(x) \cap \delta(y) & \delta(x) \cap y^+ & \delta(x) \cap y^- \\
x^- \cap \delta(y) & x^+ \cap y^- & x^+ \cap y^+
\end{pmatrix}
\]

(1)

Based on the set intersections, it is possible to identify \(2^9\) mutually exclusive combinations from the nine intersections, where each of these intersections has two possible values: empty and nonempty. This leads to 8 realizable topological relations between arbitrary regions, 10 between regions with disconnected boundaries, 33 between lines, and 19 between a line and a region.

Notice that by taking into account the efficiency to derive topological relations, the 4-Intersection Model (i.e., intersections that consider only the interior and boundary) can be sufficient for modeling topological relations between \(n\)-cells if their boundaries are connected and codimensions are zero, that is, cells are of the same dimension. However, when objects are embedded into higher dimensional space, it is necessary to examine whether or not boundaries and interiors are subsets of the other objects' closure.

Addressing the analysis of topological relations on a surface, which is neither the planar 2-D nor the 3-D space, the work in Egenhofer (2005) systematically derives the topological relations between two nonempty regions on the surface of a sphere. It analyzes first if all eight topological relations in \(\mathbb{R}^2\) could be realized on the sphere and then it analyzes the completeness of this set by examining constraints on the 9-Intersection combinations that could lead to new topological relations. A first conclusion of this work is that the 9-Intersection Model serves to distinguish the same eight relations on the sphere. However, new topological relations between regions on a sphere are found. Indeed, there are eleven, instead of eight, topological relations between regions on the sphere. Figure 2 shows examples of the three additional topological relations realizable on the sphere, which are characterized by the fact that the union of regions forms the entire sphere.

Taking a more general approach, the 9\(^+\)-Intersection Model in Kurata (2010) is an extension to the 9-Intersection Model that distinguishes topological relations between arbitrary objects by patterns in nested matrices. It defines a set of constraints on this matrix, which are applicable to arbitrary pairs of objects in various spaces. In the 9\(^+\)-Intersection Model, matrices are of the form \(M^+\) in equation 2, where the bracketed elements are matrices by themselves and \(\delta_i(x), x_i^+,\) and \(x_i^-\) are the \(i\)-th subparts of an object \(x\).

\[
M^+(x, y) = \begin{pmatrix}
\delta_i(x) \cap \delta_j(y) & [\delta_i(x) \cap y^+] & [\delta_i(x) \cap y^-] \\
[x_i^+ \cap \delta_j(y)] & [x_i^- \cap y^+] & [x_i^- \cap y^-] \\
[x_i^- \cap \delta_j(y)] & [x_i^- \cap y^+] & [x_i^- \cap y^-]
\end{pmatrix}
\]

(2)
An example of this model is shown in Figure 3 (Kurata 2010). It shows $M^+(x, y)$, where $x$ is a directed line (DLine) with distinguished start and end points of the boundary and $y$ is a region.

The 9+-Intersection Model derives nine general constraints upon which it defines a set of admissible intersection matrices. Table 3 shows a summary of the number of topological patterns derived by this model in the 2-D ($\mathbb{R}^2$), 3-D ($\mathbb{R}^3$), and on a 2-D sphere ($\mathbb{S}^2$).

There are also efforts to translate the topological relationships in a continuous space (i.e., vector space) into relations in a discrete space and applicable to a raster representation of the space. The raster space $\mathbb{Z}^2$ is seen as a rectangular array of points or pixels. The work in Egenhofer and Sharma (1993) defines a raster object as an extended object that is bounded. An extended object is a finite proper subset of $\mathbb{Z}^2$, where two points in the extended object are connected if there is a connected path between them (i.e., a path formed through the 4- or 8-neighbors). In this context, a boundary of a raster region is a Jordan curve that separates the interior from the exterior. For this work, raster regions have nonempty boundaries and interior, and their
TOPOLOGICAL RELATIONS

\[
\begin{pmatrix}
\delta_1 (x) \cap \delta(y) & \delta_1 (x) \cap y^c & \delta_1 (x) \cap y^- \\
\delta_2 (x) \cap \delta(y) & \delta_2 (x) \cap y^c & \delta_2 (x) \cap y^- \\
x^c \cap \delta(y) & x^c \cap y^c & x^c \cap y^- \\
x^- \cap \delta(y) & x^- \cap y^c & x^- \cap y^- \\
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\emptyset & \emptyset & \emptyset \\
\emptyset & \emptyset & \emptyset \\
\emptyset & \emptyset & \emptyset \\
\end{pmatrix}
\]

**Figure 3** An example of the topological relation between a directed line \((x)\) and a region \((y)\) in the 9+-Intersection Model (Kurata 2010).

Table 3 A summary of derived topological patterns of the 9+-Intersection Model in the \(\mathbb{R}^2, \mathbb{R}^3\), and \(\mathbb{S}^2\) (Kurata 2010).

<table>
<thead>
<tr>
<th></th>
<th>(\mathbb{R}^2)</th>
<th>(\mathbb{R}^3)</th>
<th>(\mathbb{S}^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point–Point</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Point–DLine</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Point–Region</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Point–Body</td>
<td>—</td>
<td>3</td>
<td>—</td>
</tr>
<tr>
<td>DLine–DLine</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>DLine–Region</td>
<td>26</td>
<td>45</td>
<td>26</td>
</tr>
<tr>
<td>DLine–Body</td>
<td>—</td>
<td>26</td>
<td>—</td>
</tr>
<tr>
<td>Region–Region</td>
<td>8</td>
<td>43</td>
<td>11</td>
</tr>
<tr>
<td>Region–Body</td>
<td>—</td>
<td>19</td>
<td>—</td>
</tr>
<tr>
<td>Body–Body</td>
<td>8</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

**Figure 4** The boundary, interior and exterior of a raster region: (a) admissible raster region with four-connected interior and boundary and (b) inadmissible raster region with disconnected interior. Adapted from Egenhofer, M.J., and J. Sharma. 1993. “Topological Relations Between Regions in \(r^2\) and \(z^2\).” In Advances in Spatial Databases, Third International Symposium, SSD’93, 316–336, © With permission of Springer.

interior and exterior are four-connected, that is, they are composed of cells that are adjacent by row or by column (Figure 4).

From the analysis of the set intersection in the discrete space, all eight JEPD topological relations defined by the 9–Intersection Model that can be realized between two regions in a continuous space can be also realized between two raster regions. In addition, there are eight more topological relations between bounded raster regions. Even more, another three topological relations can be identified in \(\mathbb{Z}^2\) if considering whether or not empty boundary–boundary intersections are neighbors. Figure 5 shows topological relations for raster regions that are not realizable in the continuous space, excluding those cases that can be distinguished when boundary–boundary intersections are neighbors. This figure omits converse relations, which can be derived from cases Figure 5a, Figure 5b, and Figure 5c, respectively.

The work in Egenhofer and Sharma (1993) identifies the following mappings of relations from the raster to the vector space: Relation (a)
in Figure 5 and its converse relation map to the Inside and Contains, respectively. In a similar way, relation (b) and its converse relation map to Covered by and Covers, respectively. Finally, relations (b), (c) and their converse relations in the raster space map to the Overlap in the vector space.

Despite the wide use of the 9-Intersection Model and its extensions, the model is not exempt from suffering some criticism. In the 9-Intersection Model, not all intersections are possible and there may be several object intersections that are topologically equivalent. The 9-Intersection Model was originally designed for 2-D objects and cannot distinguish a large number of topological relations between complex 3-D objects because it ignores that the interiors, boundaries, and exteriors may also be complex objects. Even more, it has been argued that the 9-Intersection Model suffers a basic contradiction in the definition of boundaries and interiors when objects of lower dimensions are embedded into higher dimensions. In particular, when a line, which is a 1-D entity whose boundary is composed of the two end points, is embedded into a 2-D space, the interior is connected to the exterior. This contradicts that the boundary separates the interior from the exterior of an entity. To overcome this previous drawback of the 9-Intersection Model, other point-set formalisms for topological relations consider other topological invariants that can distinguish more topological relations (Zlatanova, Rahman, and Shi 2004).

A different formalization of topological relations replaces the exterior of an entity by its Voronoi diagram. This results in thirteen relations between areas, eight between lines, thirteen between a line and an area, three between points, four between a point and a line, and five between a point and an area. Later, a Voronoi-based algebra was also proposed. This formalization utilizes set operators (i.e., union, intersection, difference, difference by, symmetric difference) to distinguish spatial relations between neighboring spatial objects, and several values are the results of these operators: content, dimension, and number of connected components.

From a mathematical point of view, the extended model for topological relations (EMTR) defines topological relations between any two objects in GIS. From this perspective, a topological relation is basically an equivalence relation that splits the relation between two objects into different partitions. Instead of using the value of empty/nonempty of the 9-Intersection Model, the EMTR uses for convex sets \( x \) and \( y \) the number of connected components of elements in the matrix:

\[
\begin{pmatrix}
  x \cap y & x \setminus y \\
  y \setminus x & \delta(x) \cap \delta(y)
\end{pmatrix}
\]
Table 4  Definition of topological relations by the Open Geospatial Consortium.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disjoint ((x, y))</td>
<td>True if (x \cap y = \emptyset)</td>
</tr>
<tr>
<td>Touches ((x, y))</td>
<td>True if (x \cap y \subseteq (\partial(x) \cup \partial(y)))</td>
</tr>
<tr>
<td>Equal ((x, y))</td>
<td>True if (x = y)</td>
</tr>
<tr>
<td>Within ((x, y))</td>
<td>True if (x \subseteq y)</td>
</tr>
<tr>
<td>Contains ((x, y))</td>
<td>True if (y \subseteq x)</td>
</tr>
<tr>
<td>Overlaps ((x, y))</td>
<td>True if (x \cap y \neq \emptyset, x \cap y \neq x \neq y), and (\dim(x \cap y) = \dim(x) = \dim(y))</td>
</tr>
<tr>
<td>Crosses ((x, y))</td>
<td>True if (x \cap y \neq \emptyset, x \cap y \neq x \neq y), and (\dim(x \cap y) &lt; \max(\dim(x), \dim(y)))</td>
</tr>
<tr>
<td>Intersects ((x, y))</td>
<td>True if (x \cap y \neq \emptyset)</td>
</tr>
</tbody>
</table>

For nonconvex sets \(x\) and \(y\), this model also considers the fundamental group of \(x \cup y\). The fundamental group is a concept of the algebraic topology that captures information about the basic shape or holes of a topological space. By considering the property of connectivity and fundamental group, the topological relations between objects can be described by an infinite sequence of matrices.

The dimensional model is another model for representing spatial relations. The model is based on the dimensional elements of spatial objects and their dimensional relations. The \(\alpha\)-dimensional element of an object \(A\) corresponds to the set of all points of \(C\) that have order 0. For example, a triangle in \(\mathbb{R}^2\) contains points of order 0 that are vertices of the triangle, points on the edges are of order 1, and points inside are of order 2. Consequently, elements in the 3-D space are of order 0, 1, 2, and 3. Then, to define the topological relations between spatial objects \(A\) and \(B\), the dimensional relations between elements in \(A\) and \(B\) are checked from the highest to the lowest dimensional order. The dimensional relations can be partial, total, or not existing, depending on the intersection between dimensional elements. They can also be classified by distinguishing basic relations that consider whether or not the dimensional elements intersect, extended relations that consider the dimension of the intersection, and simplified relations where some dimensional relationships are aggregated.

The major drawback of the alternative approaches to the 9-Intersection Model is that there is no clear semantic interpretation of all relations derived from these models and, in consequence, how they relate to common sense in practical use.

Topological relations in query languages and for information retrieval

Topological relations are very important components of spatial query languages in databases. Topological relations are inserted as filter conditions of join or range queries. For example, they are useful to specify a join query that retrieves rivers that cross a city. They also serve to specify the valid states of a database through the definition of integrity constraints. For example, it could be specified that the boundary of two counties must be disjoint or touch each other. Thus, every time there is an update on the database, the system should check if there is a pair of counties that are internally connected. These types of constraints are called semantic integrity constraints, which are typically implemented by using query languages.
TOPOLOGICAL RELATIONS

Current spatial database management systems have adopted the standard by the Open Geospatial Consortium (www.opengeospatial.org/). Table 4 provides the definitions of the topological relations extracted from the Open Geospatial Consortium simple feature specification. In this table, given a geometry \( x, \partial(x) \) indicates its boundary, and \( \dim(x) \) its dimension, where \( \dim(x) = 0 \) if \( x \) is a point, 1 if it is a curve, and 2 if it is a surface.

It can be seen from this table that current query languages use a coarser set of topological relations than the set defined by the 9-Intersection Model. These relations, such as the coarser relation Within, are not part of the basic and pairwise disjoint basic relationships of models. However, they are useful in practical cases.

In common spatial databases, retrieval by spatial relations is done on the basis of information consisting of objects stored in relational tables and organized by thematic layers with spatial indexing methods (e.g., R-Tree). This technique answers queries as cascaded spatial joins and is restrictive to the type of objects and relations. Unlike querying databases where data are retrieved by an exact matching between the query and the stored data, spatial-information retrieval focuses on similarity-based searching of spatial configurations. In this context, queries can be specified with a visual language that represents a scene by a set of objects and a set of constraints defined by the topological relations between objects. Similarity-based search has a common strategy: (i) definition of the set of spatial relations that can be used in a query, (ii) definition of a similarity measure of spatial relations, and (iii) implementation of a search algorithm for similarity retrieval.

In the typical similarity-based retrieval of spatial configurations, similarity measures compare the topological relation between objects with respect to a topological constraint. Using relations defined by the 9-Intersection Model or RCC, a similarity measure compares topological relations using the semantic distance between relations as defined by a conceptual neighborhood graph (Egenhofer and Al-Taha 1992). This type of measure suffers the disadvantage that it does not distinguish pairs of objects that, holding the same topological relation, correspond to different spatial configurations. For example, it does not distinguish between a pair of disjoint objects very close to one another and a pair of disjoint objects very far apart. Other types of similarity measures use a quantitative approach to compare relations between objects based on distances, angles, or overlapping areas (Godoy and Rodriguez 2004). This later type of measure can only be applied when comparing pairs of geometries.

In general, spatial information retrieval for a variable number of objects and constraints is a hard computational problem due to the size of the search domain of possible solutions. This means that there is no algorithm that can run in a reasonable time and that finds the set of objects that best match the query. As a consequence, studies take approximation approaches. Most of these studies use heuristics and approximation algorithms based on evolutionary algorithms. They start with a set of possible solutions and then apply heuristics and transformations on the initial solutions to improve the solutions in each iteration of the algorithm.

Reasoning about topological relations

Two important reasoning problems with topological relations are the topological consistency problem and the realizability problem (Grigni, Papadias, and Papadimitriou 1995). These are important problems in applications where you
have to derive possible scenes from qualitative descriptions based on topological relations or when you want to evaluate the consistency of specifications.

A set of topological relations defined over a set of spatial objects is said to be topologically consistent if there are no contradictions among the topological relations. For example, three objects $A$, $B$, and $C$ and the relations Covers($A$, $B$), Covers($B$, $C$), and Disjoint($A$, $C$) are topologically inconsistent, as the first two relations imply that $A$ and $C$ cannot be disjoint.

A fundamental concept to topological consistency is the notion of composition of topological relations. The composition of two topological relations $T_1(A, B)$ and $T_2(B, C)$ over a common object $B$, denoted $T_1(A, B) \otimes T_2(B, C)$, enables the derivation of the topological relation between objects $A$ and $C$. The composition of the eight base topological relations between regions is shown in Figure 6, which is the same derived from the 9-Intersection Model and RCC. In Figure 6, each topological relation is represented as a dark node in a graph. If, for example, Overlap($A$, $B$) and Covered_by($B$, $C$) are true, then from the composition table it is obtained that either Overlap($A$, $C$), Covered_by($A$, $C$) or Inside($A$, $C$) is also true. The composition between two relations, say Overlap and Covered_by, is denoted by:

\[
\text{Overlap} \otimes \text{Covered_by}\\
= (\text{Overlap} \lor \text{Covered_by} \lor \text{Inside})\\
= \{\text{Overlap, Covered_by, Inside}\}
\]

A more formal definition of topological consistent is then defined by a consistency network as follows. Given a set of spatial objects $Ob$ and a conjunctive expression of the form $\psi = \bigwedge_{i=1}^{n} T_i(A_i, B_i)$, where each $T_i$ represents a disjunction of base topological relations between objects $A_i$ and $B_i$ belonging to $Ob$ (i.e., alternative base topological relations between $A_i$ and $B_i$ — recall that base relations are the high relations (JEPD) topological relations), $\psi$ is said to be topologically consistent if there exists a network of base topological relations (denoted by $\pi$), that is, if there exists a function $TR : Ob \times Ob \pm \pi$ such that:

1. A topological relation $T$ is defined for every $A, B \in Ob$.
2. For every $A \in Ob$, $TR(A, A) = \text{Equal}$.
3. For every $A, B \in Ob$ such that $T(A, B) \in \psi$, $TR(A, B) \in T$.
4. For every $A, B, C \in Ob$, $TR(A, C) \subseteq TR(A, B) \otimes TR(B, C)$.

The following example illustrates a case where topological consistency is satisfied. Given a set of objects $Ob = \{A, B, C\}$, the topological expression $\psi = T'(A, B) \lor T(A, C)$, where $T' = \{\text{Meet, Equal}\}$ and $T = \{\text{Equal, Inside, Covered_by}\}$, is topologically consistent because there exists a function $TR$ that can be defined such that $TR(A_i, B_i) = \text{Equal}$ for every $A_i$ and $B_i$ in $Ob$. Indeed, $TR(A, B) = \text{Equal} \in T'$, $TR(A, C) = \text{Equal} \in T$ and for every $A_i$, $B_i$, and $C_i$ in $Ob$, $TR(A_i, C_i) \subseteq TR(A_i, B_i) \otimes TR(B_i, C_i)$ because $\text{Equal} \subseteq \text{Equal} \otimes \text{Equal}$.

The following example, in contrast, illustrates a case where a set of topological relations does not satisfy topological consistency. Consider the set of spatial objects $Ob = \{A, B, C, D\}$ and the topological expression $\psi = \text{Covered_by}(A, B) \land \text{Covered_by}(B, C) \land \text{Inside}(C, D) \land \text{Covers}(A, D)$. The possible topological relations between $A$ and $C$ must be defined by the common relations in the compositions $\text{Covered_by}(A, B) \otimes \text{Covered_by}(B, C)$ and $\text{Inside}(C, D) \otimes \text{Covered_by}(D, A)$. However, by the composition table (Figure 6), $\text{Covered_by} \otimes \text{Covered_by} = \{\text{Covered_by},$
TOPOLOGICAL RELATIONS

Inside} and Covers \( \otimes \) Includes = \{Includes\}. Since Covered\_by \( \otimes \) Covered\_by \( \cap \) Covers \( \otimes \) Includes = \( \emptyset \), there exists no function TR that satisfies the conditions for consistent network.

There are some cases in which, even if the topological relations are consistent, the spatial objects may not be realizable due to reasons of planarity, that is, there are no objects that can be

---

Figure 6 Composition of basic topological relations. Adapted from Egenhofer (1994). © 1994, with permission from Elsevier.
drawn on a plane and satisfy topological relations. The problem of determining if these objects exist is the realizability problem (Grigni, Papadias, and Papadimitriou 1995). The conjunctive realizability problem consists in determining if there exists a planar model that satisfies a conjunctive topological expression like $\psi = (\text{Equal}(A, B) \land \text{Disjoint}(B, C) \land \text{Disjoint}(A, C))$.

Formally, given a set of objects $Ob$ and a conjunctive topological expression $\psi = \wedge_{i=1}^{n} T_i(A_i, B_i)$ where each $T_i$ represents a disjunction of base topological relations between objects $A_i$ and $B_i$ belonging to $Ob$, the realizability problem consists of determining whether there exists a planar model for each object that satisfies $\psi$.

If a topological expression is not topologically consistent, then it directly follows that it is not realizable. There are some cases, however, where the expression is topologically consistent but it is not realizable on a plane. From a computational point of view, there is no known algorithm that can efficiently determine if a topological expression is realizable. Consequently, the topological consistency problem is generally used as a good approximation algorithm for the realizability problem.

Another interesting concept when reasoning about topological relations is the notion of a conceptual neighbor. The conceptual neighbor refers to relations that are semantically closer to each other. This is useful for grouping relations and detecting similar relations, which have practical applications to compare spatial configurations and select appropriate terminology when people communicate about specific spatial configurations. Figure 8 shows the eight base topological relations organized in a graph that connects conceptual neighbors derived from the concept of gradual changes (Egenhofer and Al-Taha 1992). For example, Disjoint and Meet are two neighboring relations in this graph and, therefore, they are conceptually closer than Disjoint and Overlap. Using this graph a similarity measure compares topological relations using the semantic distance between relations.

In a similar way, there exist conceptual neighboring of relations between lines or between regions defined by interval relations in $\mathbb{R}^1$, region relations in $\mathbb{R}^2$ and regions on the sphere $\mathbb{S}^2$.

**Natural-language spatial relations**

With the goal of making GIS closer to the way users communicate about topological relations, the work in Shariff, Egenhofer, and Mark (1998) studies the geometry associated with natural-language spatial terms. Understanding
TOPOLOGICAL RELATIONS

the use of natural-language spatial terms is important when trying to capture the common-
sense knowledge about the physical world, which is a central area of naive geography.

Focusing on the 19 region–line relations defined by the 9-Intersection Model, three types
of measures model the metric refinements of the topological relations between regions and lines:
(i) splitting, which determines how the interior, boundary, and exterior of a region is divided
by the interior and boundary of a line, and vice versa; (ii) closeness/nearness, which deter-
mines how much in common two regions have; and (iii) approximate alongness, which assesses
the length of the line and perimeter when the interior of a line does not coincide with
the boundary of a region. In total, they define six different measures of splitting, four of closeness,
and two of alongness. To clarify the idea of these metrics, Figure 8 shows one measure for each
type that is defined as a dimension-neutral ratio.

This model of metric refinements over the topological relations was calibrated for 59
English–language spatial terms (Shariff, Egenhofer, and Mark 1998). Major conclusions of
this work are that topological relations are more important than metric refinements when char-
acterizing spatial terms and that many spatial terms fall in the same topological relation but
not with equal metric parameters.

The work in Xu (2007) also studies natural-
language spatial terms but focuses on linear
objects. It uses the 9-Intersection Model and
metric indices to formalize natural-language spatial terms between two line objects. Metric
indices in this case are direction (local and global angles), splitting (i.e., intersect splitting, along-
ness splitting, and interior traversal splitting), distance (i.e., short and long distances), and
overlap ratio. Unlike the work in Shariff, Egen-
hofer, and Mark (1998), however, in this case topological and metric indices are used together
to classify spatial terms with a decision tree. Only
the indices that are found to better classify terms
are then used in spatial queries. As a conclu-
sion, this work found that topological relations
together with metric indices are better to classify
spatial terms than just topological relations alone.

SEE ALSO: Geographic information science;
Qualitative spatial and temporal representation
and reasoning; Spatial concepts

References

of Binary Topological Relations.” Journal of Visual
Languages and Computing, 5(2): 133–149.
Egenhofer, M.J. 2005. “Spherical Topological Rela-
tions.” Journal on Data Semantics, 2: 25–49.
Topological Relations.” International Journal of Geog-
cal Relations between Regions in r2 and z2.” In
Advances in Spatial Databases, Third International
Symposium, SSD’93, Lecture Notes in Computer
“Topological Inference.” In Proceedings of the 14th
International Joint Conference on Artificial Intelligence (IJCAI ’95), 901–907. San Francisco, CA: Morgan Kaufmann.
Kurata, Y. 2010. “9+ Intersection Calculi for Spatial
Reasoning on the Topological Relations between Heterogeneous Objects.” In GIS ’10 Proceedings of
the 18th SIGSPATIAL International Conference on
Advances in Geographic Information Systems, 390–393.
New York: ACM.

Further reading

Place names weld together language and territory. Place names that label, denote, and identify places belong to the language of space and to the poetics and politics of the landscape. Importantly, as elements of language spoken and inscribed in gazetteers and ordinances and on maps and public signs, place names are enmeshed in a web of significations.

Intended to call attention to the existence of particular places, the primary function of place names is to signify geographical features, such as countries, seas, mountains, creeks, towns, or streets, and thereby differentiate between them. Giving a name to a place entails that the place is recognized as an entity distinguishable from other places. Essentially, place names are signs of civilization: geographical features belonging to elemental wilderness are largely nameless. However, conceiving geographical features as places entitled to be named possibly entails a degree of abstraction. Can a mountain be definitively delineated? What constitutes a street? Where exactly is the source of a river? Moreover, while “real” place names signify places in the physical world, real and imaginary place names also flourish in the fictional worlds of literature, cinema, and computer games.

Many place names such as Long Island, Dead Sea, or Yellow River are descriptive, which can be potentially misleading. The Pacific Ocean can be stormy. The Red Sea is not red. Myriad place names commemorate events or other places, such as Alexandria, Egypt; Washington, DC, USA; Piazza Venezia, Rome; and Pariser Platz, Berlin. Place names do not necessarily designate the properties and character of the place they refer to. They may be words that are intrinsically void of lexical sense. However, since place names are associated with human experiences and agendas, they not only denote locations but possibly also connote reputations and call to mind certain images and associations. This is clearly evident in the case of (in)famous places, whose names have become household names. As a place name, Las Vegas denotes a city in Nevada and connotes its reputation as Sin City. A reference to New York City also connotes its reputation as a vibrant metropolis that never sleeps. Sometimes place names connect with events and are invested with historical memories. Examples include Verdun (a battlefield in France, 1916) or Yalta (conference, 1945). Place names may also become metaphors for social milieux, cultural phenomena, and lifestyles that, at least in their early evolution, were connected with the location. The meanings associated with place names are historically dynamic; place names may acquire new connotations, while older ones may lose their power of evocation.

Naming is a prerequisite for converting space into place. Place names are not mere signifiers but are actively involved in place-making. The place-making capacity of place names are manifest in their power to stir the imagination, to evoke the idea of place, and to conjure a sense of place. It is evident in their ability to elicit personal and communal memories that associate places with real and fictional biographical and cultural experiences.
Human curiosity and fascination with the origin and meaning of place names are manifest in stories societies tell about the origin and meaning of place names. In Maori lore, stories about the journeys of ancient heroes also recount the names they gave to features of the land. The Old Testament contains numerous etiological stories about the origin and meaning of place names.

As it developed in the nineteenth century, linguistic orientation and philological discipline underlie the traditional European approach to the study of place names (Stewart 1975). The aim of such research has mainly been to establish etymology: tracking the development of place names since their earliest documented occurrence and possibly unearthing the provenance of place names in relation to patterns of migration and settlement. With its overwhelmingly linguistic orientation, the traditional European approach to the study of place names has been fixated on older names and almost exclusively focused on taxonomy and etymology. For lack of historical records and documentary sources, etiology, namely the causes underlying name-giving, was almost inevitably a highly speculative matter.

That place names belong to the academic discipline of geography seems self-evident, yet within academic geography studying place names was traditionally consigned to the margins of scholarship (Rose-Redwood, Alderman, and Azaryahu 2010). Whereas the kind of “applied” toponymy advanced by national boards and United Nations agencies has focused, respectively, on the standardization of toponymic nomenclatures nationally and internationally, academic geographers interested in social theory largely considered the traditional approach to the study of place names as not worthy of serious scholarship.

The resurgence of academic interest in toponymy since the 1980s has largely been anchored in and connected to applying a critical approach to the study of place names and especially procedures and patterns of place naming, where issues of power reign supreme (Berg and Vuolteenaho 2009; Rose-Redwood, Alderman, and Azaryahu 2010). The “critical turn” in toponymy as a field of academic study has largely been concerned with recent, well-documented place-naming procedures. Notably, it has also been manifest in the growing number of studies dedicated to issues related to the potentially contested urban and possibly national politics of street naming. Being a feature of urban modernity, street names had almost completely been ignored by traditional toponymic studies.

Crossing disciplines, the critical approach to the study of place names has been applied by geographers inspired by earlier works on landscape symbolism as well as by historians, political scientists, sociologists, and anthropologists. Underlying the critical approach to the study of place names is the insight that place names belong to structures of power and discourses of identity. The critical orientation of geographic inquiry into toponymy has treated place naming as an arena of politics. Accordingly, recent critical scholarship on toponymy has directed attention to how the politics of place names and place naming in particular reproduce power relations and ideological agendas in public space.

This line of research has shed light on the relationship between place (re)namings and identity politics, such as when contested place names and place-naming patterns are embedded into larger struggles for recognition and legitimacy and are emblematic of them. Studies have yielded significant insights into the shaping of toponymic regimes in colonial societies and in nationalist contexts. Along this line of research several studies have focused on toponymic changes following power shifts, most conspicuously in postcolonial and post-communist societies. When belonging
to the symbolic politics of regime change, the renaming of streets demonstrates how processes of political change and ideological reorientation at national level are instrumental in reshaping urban geographies of everyday life.

Within the growing literature on critical toponymy, several studies have focused on toponymic commemoration and especially on street naming as a strategy of public commemoration where possibly competing visions of history shape (re)naming priorities and practices. These studies have shown that when invested with commemorative capacity, toponymic inscriptions become urban sites of memory that weld together power, remembrance, and language in public space. Some recent studies have directed attention to the politics of displaying toponymic inscriptions in public signage in bilingual societies. Another line of investigation has explored the commodification of toponyms in conjunction with tourism and as a place-branding strategy and the association of place-naming practices and priorities with creating social distinction and prestige for communities and commercial ventures.

Based on the understanding that (re)naming is an expression of power, critical toponymy emphasizes the politics underlying place naming as a potentially contested spatial practice susceptible to involvement in struggles for symbolic control of public space. Importantly, the emphasis on (re)naming procedures that direct the formation of modern toponymic regimes has introduced unequivocal etiological concerns and considerations into current theoretically informed and critically oriented study of place names.

**SEE ALSO:** Cultural politics; Cultural turn; Historical geography; New cultural geography; Place; Point-of-interest databases and gazetteers; Political geography

**References**


**Further reading**

Tourism

Daniel H. Olsen
Brigham Young University, USA
Dallen J. Timothy
Arizona State University, USA

Tourism entails temporary travel outside people’s usual place of residence from one area of the Earth to another and all the elements of supply and demand at home, in transit, and in the destinations that facilitate this mobility and create an experience. People have traveled for leisure and other purposes for centuries. The foundations of modern tourism have been attributed to the Grand Tour, wherein wealthy Europeans traveled to various European countries for education, health, and pleasure. While travel for leisure purposes initially was an elite activity, increases in discretionary income, the development of cheaper and more efficient transportation, the growth of various tourism market niches, and the institutionalization of the act of travel into Western cultural lifestyles and schedules (e.g., spring breaks, reading weeks, gap years, and the growing number of holidays and paid vacation days) has led to the democratization of tourism for virtually all socioeconomic classes in the Western world.

Since World War II there has been a dramatic increase in domestic and international travel. In 2015, 1.185 billion international tourist arrivals were reported, generating approximately US$1.241 trillion in international tourism receipts (see Table 1). Tourism also represented 6% of the world’s exports, contributing to approximately 1 out of every 11 jobs worldwide, or 8.7% of total employment (including jobs indirectly supported by tourism) in 2015. With international travel expected to increase 3.3% a year until 2030, and with emerging destinations expected to see a 4.4% increase in international arrivals during the same timeframe, tourism has become a crucial component in the economic development strategies of national, regional, and local governments in their quest to diversify economies and enhance employment opportunities (World Tourism Organization 2015).

While geographers were among the earliest tourism scholars, the academic study of tourism has become increasingly multidisciplinary, in part because of the increasing complexity of various types of human mobility, the emerging industries that have been organized to cater to their needs, and the increasing economic, sociocultural, and ecological impacts of tourism. Considering that tourism is inherently a spatial phenomenon, and while geographic concepts such as complementarity, intervening opportunity, transferability, distance decay, and gravity models are key in understanding tourism flows between and within destinations, the study of tourism as a subject of analysis seems to occupy a liminal or relatively minor specialist position within academic geography (Gibson 2008). This is no doubt so because many tourism geographers are employed outside of geography departments (e.g., in business, tourism and recreation, community development, sport management, environmental planning, and cultural studies), even though many of them continue to undertake geographical research.

Geographic research on tourism began in the 1920s, and contributions by geographers
Table 1  International tourist arrivals and tourism receipts by world region, 2015.

<table>
<thead>
<tr>
<th>Region</th>
<th>International tourist arrivals</th>
<th>International tourism receipts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>53 million (5%)</td>
<td>US$36 billion (3%)</td>
</tr>
<tr>
<td>Americas</td>
<td>191 million (16%)</td>
<td>US$274 billion (22%)</td>
</tr>
<tr>
<td>Asia</td>
<td>278 million (23%)</td>
<td>US$372 billion (30%)</td>
</tr>
<tr>
<td>Europe</td>
<td>609 million (51%)</td>
<td>US$512 billion (41%)</td>
</tr>
<tr>
<td>Middle East</td>
<td>54 million (5%)</td>
<td>US$47 billion (4%)</td>
</tr>
</tbody>
</table>


to the study of tourism continue to grow and influence contemporary thinking, with some of the best known, prolific, cited, and respected tourism scholars being geographers. In the early twentieth century, geographic studies of tourism were highly descriptive, like many human geography studies of the time. The 1950s saw the beginnings of tourism-focused research in area studies, human–environment interactions and spatial analyses, which tended to dominate tourism research by geographers until the 1980s. The 1980s and 1990s saw an additional shift in focus to the sociocultural, economic, and environmental effects of tourism’s growth (Butler 2004), tourism planning and policy, and community–based development. The 1990s and the new millennium saw the expansion of research by tourism geographers into many emerging themes, including safety and security, cultural heritage management, indigenous rights, poverty alleviation, rural development, urban renewal, niche tourism in relation to gender, sexuality, spirituality, and religion, and amenity migration, addressing a range of issues central to geography such as power and empowerment, human mobility, globalization processes, neoliberalism, and technology.

While geographers continue to contribute to the increasingly fragmented study of tourism through various theoretical and applied perspectives and approaches (see Wilson 2012), many of these traditional and modern contributions can be categorized into at least three key areas: explaining the spatial complexities of tourism development, human–environmental interactions in tourism, and critical and sensuous geographies of tourism.

The spatial complexities of tourism

Because of the importance of tourism in Western economic and capitalist economies, early work by geographers revolved around the economic elements of tourism. Initially, geographers described, mapped, and modeled the spatial patterns of domestic and international travel patterns and flows, segmented tourist attractions, and identified different tourism niche markets based on people’s motives for travel and geographic origin. They explained the marketing and promotion of new tourism spaces and discussed how increasing numbers of travelers affected the creation of new commodity chains and tourism labor markets. Geographers have studied the role of international law, the supranationalization of the world, and the neoliberal growth factors that either facilitate or inhibit international tourism. Geographers have also tried to explain the effects of tourism on changing land-use patterns within urban and rural contexts, tourism development in wilderness areas and along coastlines, and the
spatial dimensions of tourist attraction systems. Different forms of land-use development unique to tourism, such as tourism zones, resorts, and touristic enclaves have also been examined as well as a number of issues related to tourism planning, growth and destination management, destination life cycles, community-based tourism, entrepreneurial behavior, transportation development, and the rise of second homes in the context of the particularities of different places.

Tourism is supposed to improve the livelihoods of people in destinations through the improvement of a country’s balance of payments, the attraction of both local and foreign investment, the creation of new economic linkages, the “multiplier effect” beyond tourism zones and related businesses, and increased employment opportunities; but such is not always the case. Transnational tourism businesses and cruise lines, for instance, tend to encourage high levels of economic leakage in many lesser developed destinations, and considering that most jobs in the tourism sector are highly seasonal, and therefore usually part time, and consist of low-wage, low-skill, and frequently dangerous work, tourism is therefore seen as contributing to, rather than helping with, current global spatial and economic disparities. Therefore geographers investigate pro-poor tourism and bottom-up approaches to development as a way to understand tourism as a potential tool to help alleviate poverty in developing countries.

Place is an essential ingredient of tourism, particularly from the perspective of destination development and marketing. Geographic approaches have led to increased knowledge about people’s attachment to place and sense of place in tourism settings, both from tourists’ and residents’ perspectives. From a marketing perspective, place branding has come to the fore in recent years as a critical factor in a location’s success as a tourist destination. Place-making and location branding, as well as meanings and identities associated with place attachment, while initiated by geographers and embracing geographical concepts and spatial complexities, have gained significant currency among non-geographers as well and form a common thread in tourism planning and marketing studies.

Finally, technological developments have provided methodological and virtual tools and cyber spaces for contemporary geographic research in tourism and elucidate additional spatial complexities in global tourism. Geographic information systems (GIS) are currently routine tools for assessing impacts and locational valuations. Social media, including Facebook, blogs, Youtube, photo sharing sites, Google Earth, and other neogeographic spaces are overflowing with volunteered geographic information about destinations and travel experiences and are now among the most popular cybernetic laboratories for tourism research. Global positioning systems-based technology, for example, has been used in recent years by geographers to understand tourists’ spatial behavior patterns in parks, historic cities, and in other heritage contexts.

Human–environment interactions in tourism

While the competitive economic elements of tourism continue to dominate much tourism scholarship, because of their interests in human–environment interactions, geographers have a strong track record of examining the direct and indirect impacts of the sector on the physical environment. The relationships between tourism and the physical environment are complex and symbiotic in many respects. On one hand, the natural environment is a critical tourism resource, whether climate- or landscape-based, that creates widespread demand
for destinations. In many cases, the development of tourism naturally leads to the protection of a destination’s natural resources, so that they remain viable in the long term (e.g., national parks and protected areas). On the other hand, the natural and cultural resources that need to be protected are the same attractions tourism utilizes most. Frequently these resources are fragile or located in ecologically sensitive areas and as increasing numbers of visitors interact with them the environment can be literally “loved to death.” Loss or deterioration of such sites is a real danger in both built and natural environments. Geographers have also examined other negative results of tourism on the environment, including changes in biodiversity, water use and quality problems, waste generation, site erosion, reduced landscape aesthetics, and reduced air quality, and how these issues affect the destination residents’ quality of life. Research by geographers shows that these impacts go beyond pure ecology to include economic implications when environmental damage decreases the competitiveness of tourist destinations, particularly those that are most reliant on environmental resources for their tourism development. While many impact studies are undertaken at site- or destination-specific scales, geographers are now highlighting global trends related to the planetary impacts of tourism development. In particular, they are now investigating the impacts of travel on global climate change, which in the future is predicted to alter not only the physical locations of tourism but also the types of markets and demand for destinations and their products, such as those located on islands and coastlines that are especially vulnerable to climate change and rising oceans.

Sustainability is one of the most important considerations in the long-term viability of tourism. Geographers have been key players in discussions about sustainable tourism development and management, particularly in the contexts of protected natural areas and various other ecosystems (e.g., marine environments, coasts, mountains). Although most discussions surrounding sustainability and tourism have focused on natural settings, geographers have been instrumental in broadening the debate into social and economic realms as well. Social and economic carrying capacities are now viewed alongside ecological capacities as vital elements of sustainable development and limits of acceptable change. Geographers also investigate various philosophical, indigenous, and religious views of the natural environment, and have been active in developing environmental ethics and codes of conduct in tourism’s use of natural and cultural resources. These efforts have assisted in improving corporate social responsibility and the establishment of eco-certifications for tourists, guides, tour operators, transportation providers, and lodging establishments.

Critical geographies of tourism

In recent years there have been calls to move beyond traditional views of tourism as a structurally and spatially circular economic practice and the statistical inclination of tourism studies to apply more critical and deconstructed analyses of the moral, cultural, and social dimensions of tourism within wider global trends, including increases in technological, political, and trans-local cultural exchanges. The critical turn in tourism studies coincides with the “cultural turn” in the social sciences, in that scholars have begun to reexamine tourism as a cultural phenomenon in terms of social representation and identity formation. Additionally, critical geographic views of tourism emphasize the ways
in which tourism reinforces dominant power structures and relationships at different scales.

As part of this critical approach, geographers have used postcolonial views of development to argue that tourism is neo-colonialist, because it creates dependency relationships between countries in the Global South where travelers holiday and the wealthier countries of the Global North where they live. From this perspective, neocolonialism suggests that low-income countries act as “pleasure peripheries” for tourists from high-income countries. In so doing, tourism is seen to increase economic, social, and racial divisions between tourists and destination residents – with tourists/tourism being the oppressor and the local community the oppressed – by marginalizing tourism workers from tourists in terms of access to capital and local resources necessary to improve livelihoods, leading in many cases to observed acculturation processes in the Global South (Gibson 2008).

The work of geographers has also noted that tourism marketers and promoters act as cultural brokers and tend to construct idealized images of both tourist spaces and cultural groups that cater to the whims of tourists, resulting in inaccurate representations of culture through the process of commoditization. Commodified, packaged, and staged resident–tourist encounters focus on affirming outsiders’ misconceptions of a particular culture rather than focusing on “authentic” cultural objects and ways of life. Geographers have also noted the performative nature of both tourist and tourist worker behavior in tourism spaces, and how governments have appropriated religious and indigenous cultural heritage to represent their regimes to tourists in a way that legitimizes their authority. Such practices have led some geographers to examine tourism ethics in the context of appropriate cultural and social encounters between tourists and local residents.

Tourists, rather than inherently searching for “staged authenticity” through the act of travel that objectifies cultures and places and utilizes tourism as a means of escape from their everyday lives, are increasingly viewed as modern reflexive subjects that negotiate their identities through the embodied consumption of places, experiences, and encounters with the Other. Geographers therefore have begun to analyze ways in which tourists interact with the signs, icons, and souvenirs created by tourism, as well as the multisensuous, multisensory, and embodied practices in which tourists engage as they try to make sense of what they experience in touristified spaces. At the same time, geographers acknowledge that tourism spaces are social constructs and have examined how tourists, not just tourism marketers and promoters, act as place-making subjects in the (re)construction of tourism spaces.

Future directions

There is virtually no area of the Earth’s surface that remains untouched in some way by the influence of tourism. With the number of people traveling expected to increase in the long term, tourism is poised to have an increasing vital impact on the economies, environments, and societies of the world. Geographers will continue to play a key role in the aforementioned fields as well as in understanding the importance and impacts of virtual travel on embodied tourism experiences; the continued use and importance of GIS and remote sensing in tourism development decision-making; the increased tension between competitiveness and sustainability; the spatialities of changing and mobile tourist and cultural identities; the development of extra-terrestrial tourism (e.g., space tourism, deep ocean exploration); safety and security in the contemporary world of terrorism.
and other threats to travel; migration, diasporas and displaced identities; and the mapping and examination of new patterns of travel from formerly low-income countries (e.g., China, Russia) where larger population cohorts now engage in international tourism. Tourism geographers will also continue to be active in the new mobilities focus in human geography, in which tourism is seen as part of people’s everyday mobility patterns and where tourism is only one manifestation of a multiplicity of diverse travel mobilities.

SEE ALSO: Climate and societal impacts; Competitiveness; Cores and peripheries; Critical spatial thinking; Environment and development; Environmental degradation; Environmental planning; Global environmental change: human dimensions; Neoliberalism; Regional inequalities; Virtual geographic environments

References


Further reading


Trade and regional development

John S.L. McCombie
University of Cambridge, UK

Regions show a great deal of heterogeneity in terms of their production structure. This derives from differences in natural resources or in relative factor endowments, namely the abundance of capital (both private and public) relative to the workforce. A corollary of this geographical specialization of production is the importance of trade, as without trade there could be little or no specialization. Much of the early work on regional development and trade treated regions as equivalent to nations, but within a common currency area, and drew on the pure theory of international trade.

The theory of comparative advantage, first put forward by David Ricardo in 1817, still remains an influential explanation of trade. Suppose that there are two regions, the East and the West, each with two industries. Labor is the only input, and the East has greater productivity in both industries. Labor is the only input, and the East has greater productivity in both industries. The remarkable result of the theory is that trade will benefit both regions, although there is nothing that ensures equality of the gains from trade. The East will specialize in the production of that good where its relative (comparative) advantage in terms of productivity is greater and the West will produce the other good. However, the theory is best seen as a static theory of resource allocation and assumes full employment. It does not explain why the productivity levels differ between regions and assumes constant returns to scale. As such it says little about regional growth except that if there is labor mobility, then there will be migration of workers to the East which will end up producing both commodities, with the West neither. With a common wage rate in both regions, the East, with the higher level of productivity, will have lower unit costs and, hence, lower prices than the West. Even if the West has an offsetting lower wage rate so that prices do not differ, labor will still migrate to the East because of the higher wages there.

The next development in trade theory was the factor endowments (Heckscher–Ohlin) theory. This introduced capital as a factor of production and assumed that each region has the same technology and the production functions once again exhibit constant returns to scale. The theory predicts that regions that have a high capital to labor ratio will export goods that require more capital relative to labor, and vice versa. The rationale is that regions with a relative abundance of capital will have lower profit rates and hence should export those goods that use a lot of capital and where the regions have a cost advantage. Early tests in the 1950s found a paradox in that the United States, a relatively capital-rich country, tended to export relatively labor-intensive goods. A number of explanations were put forward to explain this, including the fact that the United States had a highly skilled labor force and an allowance should be made for its human capital. Tests of the Heckscher–Ohlin theorem are fraught with data problems (especially concerning estimates of regional capital stocks and lack of statistics for interregional trade flows). Generally speaking, the results indicated that factor endowments explain some, but not
much, of the pattern of interregional trade. Even without factor migration, under certain assumptions, the two-country model predicts factor price equalization, but for which there is little empirical evidence. The three-country case has more ambiguous results.

There are a number of serious shortcomings of the theory. The first is that much of trade is now intraintustry, or intrafirm, which the Heckscher–Ohlin theory cannot account for. For example, intraintustry trade in the mid-2000s accounted for between 27 and 44% of world trade, depending upon the level of statistical disaggregation. The degree of intraintustry trade has steadily increased since the 1960s, suggesting a greater internationalization of supply chains with globalization and outsourcing, and this is likely to be reflected at the regional level. Second, the theory does not explain why, in the first place, some regions are relatively capital rich and others poor, especially when factor mobility is taken into account. And finally, and most importantly, there is much evidence that regional production is subject to increasing returns, especially agglomeration economies. Indeed the very existence of cities bears testament to this, because if there were constant returns to scale, there should be a much more uniform spatial distribution of production to minimize transport costs.

A major recent development in the analysis of trade and geography is the use of mainstream economic tools to analyze the causes of the location of spatial activity. The approach of the “new economic geography” (NEG), or “geographical economics,” is based on what is seen as rigorous economic microfoundations. It adopts the “representative agent” approach where the optimizing decisions of single agents (both firms and consumers) are modeled. These models are then taken to apply directly to the macroeconomy (and aggregation problems are ignored).

The pioneer in this field was Krugman (1991, 2009), whose approach was based on the observation that economic activity is highly spatially concentrated. Moreover, areas of manufacturing, such as the northeastern manufacturing belt of the United States, have maintained their dominance long after the importance of any geographical advantages, such as a location on rivers or seaboards or the presence of raw materials, have disappeared. Furthermore, industries also tended to be highly clustered, and this is as true of many older industries as it is of the new hi-tech industries. The reason for this is the presence of increasing returns which includes agglomeration economies, namely the external benefits arising from firms locating next to each other. (For a survey of the evidence see Rosenthal and Strange 2004). This leads to the spatial concentration of industry, while transport costs lead to a dispersion.

The NEG approach was based on the development of techniques that enabled increasing returns to scale and product differentiation to be formally modeled using the monopolistic or imperfect competition model developed in the 1930s by Joan Robinson and Edward Chamberlin. The approach is based on the existence of product differentiation, which explains intraintustry trade. While traditional trade theory could explain specialization of production, intraintustry trade until recently presented problems. Intraintustry trade could be largely explained by this “love of variety.” Because of increasing returns, it is more efficient for each region, or country, to specialize in one, or a few, of the various different products of an industry and to trade them, rather than to produce the whole spectrum. Intraintustry trade can also be explained by globalization and the greater spatial division of labor both between regions and between countries. In the latter
case, the routine assembly is often outsourced to low-cost countries.

The NEG models are often mathematically complicated. However, the fundamentals of the approach have been illustrated in a simple numerical example by Krugman (1991) using aggregate output (rather than a variety of different products) and transport costs. He assumes that there are two regions, namely, East and West. On the one hand, there is an immobile agricultural labor force that is divided equally between the two regions. The manufacturing labor force, on the other hand, is mobile and its distribution depends on the location choices of the firms. Production is subject to increasing returns that result from fixed costs and there are transport costs of shipping goods between the regions. The position is analyzed from the viewpoint of a single firm deciding where to locate. If transport costs are high relative to fixed costs then the other firms will have established plants in both regions and the labor force will be equally split between the regions. Hence, the firm will likewise find it most profitable to set up two plants, one in each region. The extra fixed costs arising from having two plants will be less than the total transport costs of shipping goods from a single plant located in one of the regions.

However, with low transport costs, then it will pay all the firms to locate in one region (to minimize their fixed costs) and pay the costs of shipping to those agricultural workers in the other region. The firm under consideration will similarly find it most profitable to follow suit. Thus, over time, as transport costs fall relative to the fixed costs, so firms will agglomerate in one region, through a process similar to Myrdal’s cumulative causation, but which region they choose may be due to small initial advantage or chance factors. In this sense, history matters. The approach is generalized to analyze other centripetal forces, beyond those caused by fixed costs, such as knowledge spillovers, thick labor markets, and the spatial concentration of suppliers. These cause firms to concentrate whereas such factors as increasing land rents and congestion costs lead to centrifugal forces and dispersion (Fujita, Krugman, and Venables 1999). Krugman and Venables (1995) have extended this model considering the path of wage rates. When transport costs are high, the regional structure of production and the wage rates are similar. But as transport costs fall, one region will gain an advantage because of proximity effects and deindustrialization will occur in the other region. The result is a wage disparity. But with a further fall in transport costs and globalization, wage rates will converge.

Trade, in this approach, is incorporated into the analysis through transport costs, and provides an explanation of why production becomes concentrated in terms of equilibrium (albeit multiple possible) outcomes. As such, it explains the specialization of production, but does not say much about regional growth, or development, per se. Neoclassical endogenous growth models based on the aggregate production function have been extended to allow the interaction through trade between, usually, two regions. The standard Solow growth model treats technological change as occurring exogenously and as a pure public good. Endogenous growth theory, as its name suggests, attempts explicitly to model technical change, either the result of learning-by-doing and a function of capital accumulation (the $AK$ model) or through an R&D (research and development) sector. Trade is, however, rather tangential in the spatial adaption of these models which emphasize, for example, the distance of a region from the technology frontier, and technological spillovers. Donaghy (2009) provides a survey of this literature.

It is somewhat ironical that Krugman (2009) has himself questioned the relevance of the NEG
model to regional economies in the developed countries, noting that the degree of specialization in industry the United States reached its zenith in the 1930s and since then has been steadily declining (Kim 1998). The same pattern is reflected in the increasing globalization of international trade with falling transport costs. Trade is growing between the low- and high-income countries, reflecting comparative advantage and differences in labor costs. There is also the outsourcing of some stages of production from the developed to developing countries. It seems that the NEG now has greater relevance for regions in the new industrializing countries, such as the regions on China’s seaboard which are enjoying rapid growth, reflecting their greater access to overseas markets.

A number of theories have emphasized the role of exports as an important factor in determining the growth of a region or city. The approach was developed, most notably, by Kaldor (1970) and his discursive argument was formalized in the canonical model of Dixon and Thirlwall (1975). The Kaldorian approach concentrates on the factors determining the growth of the demand for a region’s exports, which in turn drives output growth. A faster growth of demand will lead to a faster growth of supply, and the additional resources will be met, inter alia, by the net in-migration of labor and increased capital accumulation. There will also be a faster rate of induced technical change through, for example, learning-by-doing. Hence, it is a Keynesian theory of long-run growth, but where the focus is on the structure of exports and whether they are products for which world (including other regions’) demand is rapidly growing. It thus stands in contrast to the supply-oriented NEG models, where the role of demand does not explicitly feature.

Kaldor also stressed that production is subject to increasing returns (defined broadly to include static and dynamic increasing returns to scale, agglomeration economies, and induced technical change) as evidenced by the Verdoorn law. This is a macroeconomic production relationship and, in its simplest form, is given as \( p = \alpha + \nu q \), where \( p \) is productivity growth and \( q \) is output growth. In the equation, \( \alpha \) is the exogenous growth of productivity and \( \nu \) is the Verdoorn coefficient, which most empirical studies find takes a value of about one half. (These regression studies often use regional data for the states of the United States, China’s provinces, or the European Union (EU) NUTS (nomenclature of territorial units for statistics) regions.) This means that a one percentage point increase in output growth causes an extra growth of productivity of half a percentage point and an increase in employment growth of a similar magnitude. If there were constant returns to scale, then the rate of increase in employment would be the same as output and the Verdoorn coefficient would take a value of zero. The results provide confirmation at the aggregate level of the microeconomic studies that find agglomeration economies important.

Recent estimates of the Verdoorn law extend this approach, most notably explicitly including the growth of the capital stock. For example, Angeriz, McCombie, and Roberts (2009) tested the law for the EU regions over the period 1986–2001, within a more general spatial econometric framework and including other regressors to capture spatial spillovers. They find that the Verdoorn coefficient of 0.67 gives a substantial estimate of “encompassing returns to scale” (i.e., including induced technical progress and the effects of learning-by-doing) of just over 3. The spatial spillovers are also statistically significant.

The cumulative nature of regional economic growth takes the following form. A faster growth of regional output leads to a faster growth of productivity, through the Verdoorn law. A more rapid growth of productivity, in turn, leads to
improving regional price competitiveness. This, in turn, results in a faster growth of exports. Prices are determined by a markup pricing model, and the growth of nominal wages is assumed to be constant across the regions, determined by national bargaining and comparability norms. This implies that the faster the growth of productivity, the more slowly a region's prices will grow, compared with other regions. Export growth is determined by an export demand function and is a function of the rate of change of relative prices and the growth of the region's export markets, often proxied by the growth of the latter's income. The faster growth of exports leads to a further increase in the growth of regional output through the dynamic Harrod foreign trade multiplier.

Whether the regional growth in the model converges to a steady state, or follows an explosive path, depends on the values of the various parameters of the model, including the size of the Verdoorn coefficient and the values of the price elasticity of exports and the elasticity of output growth with respect to export growth. Dixon and Thirlwall (1975) considered that these values are such that there will be a convergence, whereas Kaldor disagreed. As regional growth rates clearly cannot accelerate indefinitely, in the latter case, supply-side factors come into play to restrict the rate of growth. These include the lack of a sufficient degree of labor mobility with resulting labor shortages bidding up regional wages and reducing the region's price competitiveness. Other growth-restricting factors include greater congestion costs and increasing urban land rents. Nevertheless, the outcome of trade flows is a core-periphery situation with relatively high- and low-productivity regions and greater open, and disguised, unemployment in the latter.

A shortcoming of this approach is that it is possible for, say, the growth of regional imports to permanently diverge from that of exports. But no country (or region) can run an increasing deficit on the current account indefinitely. In the case of a country, the net overseas debt to gross domestic product (GDP) ratio will eventually become so large that international borrowing to finance the deficit will dry up. (This is roughly when the ratio of net overseas debt to GDP ratio reaches 60%.) Consequently, a country is constrained to grow in the long run at a rate where the current account is roughly in balance. Likewise, regions have a balance of payments with other regions and countries. While these payments are not recorded in the official accounts, the economic consequences arising from a region's performance in its external markets are just as real for a region as for a country. “Regional problems are ‘balance-of-payments problems,’” as Thirlwall (1980) puts it. The regional balance of payments is similar to those in the national accounts with the exception that there is no category of a change in foreign reserves. Moreover, interregional fiscal transfers can be important.

The importance of this financial constraint has been often overlooked and this is due to the fact that a national banking system may obscure the problems faced by unbalanced trade. Moreover, there cannot be a (regional) currency crisis. The fact that with a national banking system, regional branch bank loans do not have to be matched by deposits may give the erroneous impression that financial flows induced by a current account deficit are automatically accommodated indefinitely. But as the case of the post-2008 eurozone crisis shows, countries cannot borrow indefinitely from other countries, even when they are all in a common currency area.

Suppose that there is a collapse in regional export earnings. It is likely that there will be initially an attempt by consumers and firms to maintain expenditure by borrowing from other regions. However, a reevaluation of the prospects of the region by the banks and the decrease in
the region’s collateral as its assets progressively become mortgaged or sold will lead to a rising regional risk premium. Eventually, there may be credit rationing. There is, consequently, only a limited scope for a region to offset a decline in the performance of its export sector by borrowing from other regions. The only exception to this is through the automatic stabilizing transfers such as unemployment benefits and other fiscal transfers, but these are likely to be relatively small. (See Dow 1986.)

The cumulative causation model can be extended to take into account the regional balance-of-payments constraint. It is now assumed that prices of exports are set in national markets and hence price competition is limited. (In an international context there is evidence that the price elasticities are low and the Marshall–Lerner conditions are only just satisfied, and/or that the terms of trade do not greatly alter in the long run because of oligopolistic pricing policies.)

Under these circumstances, the balance-of-payments-constrained growth of a region \(y_B\) is given by the simple rule known as Thirlwall’s law (see Thirlwall 2012 for a survey). This is

\[
y_B = \frac{x}{\pi} = \varepsilon \frac{z}{\pi},
\]

where \(x\) is the growth of the region’s exports, and \(z\) is the growth of the region’s exports markets. The parameters \(\varepsilon\) and \(\pi\) are the world income elasticity of demand for the region’s exports and the region’s income elasticity of demand for imports. Differences in the values of these admittedly highly aggregated parameters between regions reflect all aspects of the nonprice competitiveness of the region. These include the structure of production, for example whether the region specializes in, say, the production of goods and services, such as hi-tech products or financial services, for which “world” demand is rapidly growing, or alternatively heavy industry for which world demand is falling. If the ratio \(\varepsilon / \pi\) is, say, 3, then this means that if the region’s export markets are growing at 3% per annum, the growth of the region will be 9% when the balance of payments is in equilibrium. Alternatively, if the region is specializing in a declining industry and the ratio is one half, then the growth of the region will be about half that of its export markets. Thus, this approach, although highly aggregated, emphasizes the disparities in the structure of output and the drivers of productivity as key elements of regional development. The cumulative nature of growth is now generated by a faster growth leading to increased nonprice competitiveness, through increasing the income elasticity of demand for the region’s exports and reducing its demand for imports, although this may be a slow process. This is a change in emphasis from the cumulative causation model discussed above, where changes in relative prices are important. Hausmann, Hwang, and Rodrik (2007), in their study of “what you export matters,” have shown that an index of the sophistication of exports (a concept closely related to the income elasticity of demand for exports) in an international context is closely related to the growth of exports and output.

If the growth of exports increases, *ceteris paribus*, the region will initially experience a growing trade surplus. However, with multiplier effects and favorable entrepreneurial expectations, this will induce a faster growth of investment, construction, and services. This will continue until the faster growth of imports that this induces brings the balance of payments back into equilibrium, although at a faster rate of growth of regional output. These are the workings of the Hicks super-multiplier, of which the Harrod foreign trade multiplier is a component.

There is an essential difference in this approach compared with that of the NEG. While the latter also emphasizes the diversity of production and increasing returns to scale, regions can sell as much as they can produce. While the model
explains the concentration of production and its change, it is not designed to explain fully, say, the steady decline of Detroit since the late 1970s, due to the collapse of its major export earnings, the automobile, and the failure of the city to diversify.

Because of data problems, including the absence of interregional trade flow data, most of the tests of this approach have used national and international data. There is now substantial evidence that confirms the importance of export growth and the balance-of-payments constraint. See, for example, the articles summarized in Thirlwall (2012, 343–344, table 2). Data for the US states have been used to estimate export demand functions and hence confirm the importance of demand, although not explicitly within a balance-of-payments-constrained growth framework (e.g., Cronovich and Gazel 1998). D’Acunto, Destefanis, and Musella (2004) test a version of the model for 20 Italian regions using a two-stage approach. First, they find a statistically significant relationship from the growth of exports to that of manufacturing for the North and Central regions, but generally not for the South (the Mezzogiorno). The second stage shows that the differences in these export growth rates could be explained by the constraining effect of supply-side factors.

Rowthorn (2010) uses the export-base framework and finds that a substantial cumulative decline in the North’s export base accounts for much of the UK North–South divide. Moreover, he shows that increased migration from the North to the South is likely to be of limited help in closing the gap.

Porter’s work on “competitive advantage” and clusters presents a detailed empirical examination of the role of spatial specialization in growth. Porter’s notion of a cluster is an attempt to operationalize the spatial nature of production and externalities broadly defined. Porter (2003, 562) defines a cluster as “a geographically proximate group of interconnected companies, suppliers, service providers and associated institutions in a particular field, linked by externalities of various types.” His concept has been subject to a number of criticisms, such as the wide variation in the spatial extent of different clusters and the fact that they are similar to much earlier geographical empirical work on “industrial districts” and related concepts. While Porter does not explicitly set out a theory of regional growth, it is clear he implicitly adopts a form of export-led growth theory (see Porter 2003, 572, n. 13, where he cites a number of studies of this theory). Using data for the 172 US Economic Areas, Porter (2003) finds that wages in the traded sectors are about two-thirds higher than in the local sector and he argues that the causation runs from the wages in the traded sector to the local sector. The implication is that any regional economic development policies need to be particularly focused on traded clusters, because these not only support higher wages but also drive local employment and local wages. This approach provides the foundations for the Kaldorian approach.

SEE ALSO: Cumulative causation, endogenous growth, and regional development; Growth theory; New economic geography; Political economy and regional development; Regional development models; Regional and interregional trade in producer services; Uneven regional development

References


Further reading


Trade unions

David Jordhus-Lier
University of Oslo, Norway

Trade unions have drawn interest from geographers for decades, and the discipline has helped explain their spread through history and their recent decline in Western countries. While many geographers have studied trade unions, very few have offered a definition of the concept. This might seem like a paradox, given that labor geography (see Labor geography), which has emerged as a distinct subdiscipline since the early 1990s, has attempted to redefine trade unionism through its empirical and conceptual contributions. A trade union is an organization of, and for, workers. Trade unions mobilize for worker rights and conditions, often through collective bargaining. Importantly, all trade unions are worker organizations but not all worker organizations are trade unions. While strikes (work stoppages) pre-date the establishment of formal trade unions, they have been associated with trade union organization since the start of the industrial era. The strike threat remains one of the most effective weapons of organized labor, particularly when coupled with negotiations where workers can present their demands. This entry will first briefly introduce trade unionism as a phenomenon of our contemporary world, and then discuss it as a subject of study within the geography discipline. Geographers have studied trade unionism in a variety of ways. Here, the focus is on four dimensions of trade union geographies: the spatial distribution of organized labor since industrialization; the ways in which trade unionism has been embedded in identities, places, and societies; the concept of scale as a lens through which to understand trade unions’ successes and challenges; and, finally, the organic relationship between geographers and trade unionists, and how this tie has shaped political analysis.

Trade unions across the globe

Reports on the global state of trade unionism vary according to different interpretative frames and range from pessimistic readings of universal decline to proclamations of global revival. Even the most basic facts and figures are difficult to obtain and remain subject to interpretation. The International Labour Organization (ILO) found that, 20 years ago, 164 million workers of a globally estimated 1.3 billion-strong workforce were members of a trade union. Since then, the proportion of paid workers with a trade union membership (union density) in Organisation for Economic Co-operation and Development (OECD) countries has dropped from 20.8% in 1999 to 17.0% in 2012. However, this picture is geographically very uneven. The world’s largest national union federation, the All-China Federation of Trade Unions, saw a sharp increase in membership, which was reported at 239 million in 2010. Outside China, in comparison, the International Trade Union Confederation (ITUC) and the World Federation of Trade Unions (WFTU), the two main worldwide trade union confederations, reported a membership on their websites totaling nearly 260 million members. The highest union density in the OECD region can be found in Iceland and...
TRADE UNIONS

Finland, where more than 70% of the workforce belong to a trade union. Nordic countries were also near the top of the chart measuring workdays lost to industrial action in the 2000s, joined by countries such as France, Belgium, Canada, and Spain. Such statistics do, however, undergo dramatic annual variation. While Canada lost an annual average of 186 workdays per 1000 employees in the period 1997–2006, the South African Department of Labour reported that the country had lost as many as 1593 workdays per 1000 employees during the turbulent strike year of 2010.

However, neither trade union membership, nor union density, nor strike action equals trade union influence. While the Chinese trade union movement is impressive in numbers, and is arguably undergoing liberalization, it is still dominated by a one-party state and has “not been able to assert itself as a legitimate outlet for the new kinds of grievances that are surfacing today, particularly among young workers” (Zhu, Warner, and Feng 2011, 141). The human rights nongovernmental organization Freedom House still characterizes the Chinese labor regime as repressive. In fact, in a 2010 review, this organization listed 40 out of 165 countries reviewed as “repressive” or “very repressive” in relation to worker rights. In these countries, few independent unions exist, they are not allowed to mobilize politically, and labor law is either repressive or neglected. Middle Eastern and Northern African countries are, together with countries in the former Soviet Union region, over-represented in the lowest categories. At the other end of the spectrum, 41 countries are listed as “free.” Western Europe is the region with the highest degree of freedom for workers, where all countries get the highest rating (all Western European countries receive status 4 on a scale of 0–4 in the Freedom House report, with the exception of Turkey, which does not belong to the Western European region). Progressive labor laws, independent unions, and authentic collective bargaining agreements characterize these countries. Still, influential trade unionists in these countries have voiced concerns over falling membership and union density rates. The anti-union agenda of conservative administrations in the United States and the United Kingdom in the 1980s, and the subsequent wave of reform-minded governments across Europe in the 1990s, contributed to a weakened trade union movement in the part of the world that hitherto had been the bastion of organized labor. Conversely, several emerging economies in the Global South underwent political transitions which enabled their suppressed labor movements to consolidate their position in the workplace and legitimately seek political power. The growth of the South African, Brazilian, and South Korean labor movements thus provided a counter-narrative to the experiences of European and American unions, and has encouraged unionists and scholars alike.

Union landscapes: the spatial diffusion and retreat of trade unionism

While labor history is by now a well-established discipline whose hall of fame includes prominent names such as E.P. Thompson, there is also an interesting literature within and beyond human geography that has attempted to spatialize the history of labor movements. In Southall’s (1988) study of the early beginnings of unionism, he found that the spread of the labor movement pre-1850 was facilitated by traveling artisans across the United Kingdom, in turn stimulating the establishment of national trade unions in certain sectors. In other sectors, such as cotton and coal, however, the pattern was instead a development of strong localized union cultures.
In other words, the history of labor has simultaneously contained the strategy of crossing space and developing place attachment since its inception. Southall argues that the evolution of early trade unionism was closely linked to the so-called friendly societies in the nineteenth century. Links between trade unions and other civil society organizations are also found in sociologist Hedström’s (1994) study of the Swedish trade union movement of the early twentieth century. In particular, certain free churches and temperance organizations facilitated the spread of Swedish trade unions. These close ties between trade unions and civil society organizations would later attract the attention of geographers studying the contemporary era.

As Marxist geography became dominant in Anglo-American geography departments, scholars attempted to use their analytical tools to understand and explain how class struggle, epitomized by trade unions, shaped the postwar economic map. The shift of production from Northern US states to greenfield sites in Southern states was also given a geographical treatment by Peet (1983). He argued that the concentration of working-class populations in Northern cities had given birth to a strong trade union movement in key manufacturing industries, which then spread to other sectors in the prewar period. Before this process was allowed to spread south, the 1947 Taft–Hartley Act placed strict limitations on organizing, making the American union landscape “more or less set by the time of capital to worker militancy in the late 1940s and early 1950s” (Peet 1983, 124). This paved the way for wage differentials which later, when the United States faced economic crisis and competition from East Asian export economies, would trigger an increase of manufacturing jobs in Southern states, at the expense of stagnation and decline in the North. More nuances were added to this mapping exercise by Clark and Johnson (1987), who argued that in addition to national legislative reforms and foreign competition, the geography of unionism has been shaped by local factors such as union representation elections and the strategy of union locals.

In the United Kingdom, a similar exchange took place in the 1980s, with Marxist geographers wanting to understand how the shift from Fordism to post-Fordism, as postulated by the regulation school, had influenced the geography of trade unionism. In a 1984 article in *Marxism Today*, Massey and Miles argued that deindustrialization (see Deindustrialization), organizational changes in the economy, and the relocation of production had, together with a sharp public sector growth, created a new trade union geography. The old industrial districts of trade unionism had experienced decline, due to the “massive changes in the sectoral composition, technology and geographical organisation” of key industries (Massey and Miles 1984, 20). In short, economic restructuring had dispersed trade unions and shifted their stronghold away from their traditional heartlands. Martin, Sunley, and Wills (1993) added a new chapter to this spatio–historical narrative by showing how this trend was reversed during the Conservative Thatcher government’s attack on the trade unions in the 1980s. While the “spatial dispersal” thesis might describe economic restructuring and decentralization in the 1970s, subsequent union decline created a different pattern. Throughout the 1980s, the union heartlands had – despite an absolute decline in both jobs and union members – kept their union density, in contrast to the union experiences of other areas, where new economic sectors had emerged (Martin, Sunley, and Wills 1993). These accounts of union diffusion and retreat shared a fine-grained understanding of trade unionism as simultaneously local, through activism with strong attachments to particular places, and translocal, as actors and
TRADE UNIONS

ideas traveled between workplaces and towns. The notion of union dispersal across social and material landscapes brought regional science together with Marxist analysis of class struggle. Still, certain dynamics seemed to problematize the assumption that proximity and spatial dispersal were the key ingredients of union geographies. First, international labor migration has challenged the trade union movement. Union landscapes being transformed by migrants were not a new phenomenon. In fact, Mitchell, in his seminal book *The Lie of the Land* (1996) chose as his subject of study migrant agricultural workers in California in the 1930s. These workers were instrumental in building an organized resistance against exploitation on the industrial farms across California in this stage of US agricultural development, and the response from employers and lawmakers was consequently to construct a localized discourse vilifying “outsiders,” in an attempt to secure labor control in this critical stage of the agricultural value chain. Labor migration complicated assumptions of union dispersal through close ties and geographical proximity, because migrants often had close ties to people far away while experiencing social and cultural distance with fellow workers nearby. Trade unions in Western cities have throughout history often failed to organize migrants, who in turn have chosen to rely on other networks for their security and representation.

Second, a globalizing world economy was being pulled together in what Harvey labeled the time–space compression, which affected some places and certain sectors more than others. Trade unions organizing workers in ports and logistics, for instance, have created networks mirroring the stretched nature of their industry, tying workers in one country to others working in a port on the other side of the planet (Castree 2000).

Third, around the turn of the millennium, the impact of communication technologies on trade unions started to draw the attention of activists and social scientists (e.g., Diamond and Freeman 2002). Instant connections across the globe opened up new possibilities for campaigning, illustrated by the fact that the online news and campaigning service LabourStart now claim to coordinate more than 700 trade union websites in 27 languages across the world. But as social media become part of our lives, more deep-seated changes are also likely to take place. Geographers have been late in responding to these virtual union landscapes. Still, it seems clear that in an age when we find friends on the Internet, and when solidarity can be built through Facebook groups, trade unions will again undergo spatial transformation.

Embedded unionism: geographies of difference and union renewal

In the late 1980s and early 1990s new orientations made their mark in the geography discipline – sometimes referred to by the related, but not synonymous, labels of poststructuralism, critical geography, and “the cultural turn.” With them, studies of trade union geographies also changed their focus. Rather than exploring their spatial extent, or their effects on capitalist space, scholars started asking critical questions about who they represented and how they related to other suppressed subjects. Encouraged by feminist and radical theory, categories of class were seen in relation to those of gender, ethnicity, and sexuality. While some critical geographers chose to let other working-class subjects receive the focus of their attention, others did study trade unions, but through a critical lens. With these discussions already underway in parts of the trade union movement, there were inspirational case studies to learn from. What is here labeled “embedded unionism” refers both to actual
changes in the trade union movement, that is, union renewal, and to the broadened scope adopted by scholars who study it.

The trade union has historically been seen as a society of working men, but this is a gross oversimplification. Female unionism has a history that stretches back at least to the Lowell Female Labor Reform Association (LFLRA), established in Massachusetts in 1840s to protect single women working in the textile mills (Dublin 1975). In a seminal study of US waitresses in the early twentieth century, Cobble (1991) portrays progressive single-gender unions in the hospitality industry. By pursuing what she terms “occupational unionism,” waitress unions protected female waitresses from unemployment. However, they did little to protect them from losing their jobs. Instead, they could offer them new employment opportunities in the sector in their capacity as union hiring halls. Thus they managed to secure the consent of the hospitality workforce while serving the flexibility requirements of employers.

These early and encouraging examples notwithstanding, there is little doubt that the trade union movement has remained patriarchal until the present. Interestingly, the historical organization of working women challenges the mainstream narrative of the labor movement, as pointed out by Reimer (1999). She argues that if we focus on the experiences of male labor in late Fordism, one might argue that the flexibilization and deregulation of employment since the 1970s represent a gradual shift towards risk exposure and deregulation. Tracing the trends for female labor, on the other hand, reveals a much less linear pattern. Multiple job-holding, part-time work, and casual contracts were nothing new to women, and Fordism had done little to change that (Reimer 1999). On the other hand, this period of union decline in male-dominated manufacturing coincided with a period of growth in public sector employment, meaning new opportunities for waged work for many women and a concomitant rise in unionization (Standing 1999).

Munro (1999) argues that certain sociospatial conditions have made the unionization of women particularly difficult in the private sector. For instance, while female workplaces in the public sector were larger, reminiscent of the conditions enabling the successful organizing of male industrial workers during Fordism, women in the private sector were often working part-time, in isolation, and in smaller units.

Turning her focus away from the workplace and towards the union itself, Jepson (2005) has studied how female farm workers have been represented by the United Farm Workers in South Texas. Her account found that women tended to become active in the more socially oriented parts of the trade union organization, the educational center of the union known as the campesino. However, once women became actively involved, these spaces were increasingly domesticized and detached from the politics of the trade union. By repoliticizing the campesino in the late 1970s, women were able to create a “‘differential space’ that challenged both the class-based structure of larger South Texas society and the masculinist practices within the movement” (Jepson 2005, 698).

The challenge of labor migration mentioned earlier is not something which simply stretches the social networks of workers. It also creates more heterogeneous working-class populations, particularly in big cities, which presents social actors with new challenges in overcoming difference. Trade unions might accommodate strongly localized identities, and even racist attitudes, which inhibit them from becoming representative worker organizations in a multicultural society. Ethnic divisions in labor markets historically have often been entrenched and
reproduced by trade unionism. One of the most infamous cases took place in Johannesburg in the 1920s, when white mine owners wanted to reform their racially discriminatory labor regime from a “job color bar” to a “wage color bar”: in other words, black workers were allowed to fill labor shortages in higher positions but would still not enjoy the wages of white workers, in order to maintain profit margins. The white trade unions chose to respond in an equally reactionary fashion, by fighting to keep the job color bar to fend off competition from black workers (Johnstone 1971). Fifty years later, black trade unions would form an integral part of the anti-apartheid movement leading the country towards political transition.

Recently, geographers have studied innovative attempts by socially oriented trade unionism (see Unionism, community) to overcome ethnic difference through their mobilization. In a 2001 article, Pastor explores community mobilization in one of the world’s most culturally diverse cities, Los Angeles. The Labor/Community Strategy Center attempted to organize around an issue which affected the mobility of working-class people in need of access to their regional labor market. By focusing on improved bus services, they united the interests of low-income, mainly minority-based, neighborhoods across Los Angeles. What became known as the Los Angeles Bus Riders’ Union also challenged institutional racism and fought for spatial justice. Pastor is not alone in suggesting that there is an embryonic, but significant, reorientation within the Anglo-American trade union movement signaling a more inclusive approach to diversity. Perhaps as acts of necessity, trade unions challenged by conservative policies and neoliberal reforms have chosen to abandon business-oriented models of unionism in favor of social mobilization.

At one level, community-oriented unionism can simply be understood as a way for unions to reach out to the local communities of workers, and to civil society organizations beyond the trade union movement. But it can also be thought of as a more fundamental shift from a politics of production to embracing a politics of consumption. This also has an explicit spatial dimension, as global production networks bring producers in one country into relationships with consumers in faraway places. Hartwick (2000) explores these politics, and how consumers are alienated from the producers of the goods they consume. Johns and Vural (2000, 1210) explore how consumers and trade unions can rebuild solidarity with exploited producers despite “geographic and social disjunctures.” In other areas of life, such as public services and health, consumption and production often take place in a shared physical location, giving trade unions organizing workers the ability to build strong community alliances. Through empirical examples and theoretical work, geographers have attempted to show how trade unions can (re)gain legitimacy among new worker subjects, and in new cultural and socioeconomic contexts. Thus, geography has not simply given us a better understanding of the state of trade unionism, but also contributed to redefining its political agenda.

Scalar union strategies: from local to national to glocal

The historical development of trade union movements in countries across the world has taught us important lessons about how economic and political spaces are ordered. In the human geography discipline, this has nurtured an ongoing debate about the concept of Scale (see also Scale and anti-scale). In its simplest definition, the scale of trade unionism refers to how trade
unions are articulated from the local to the global level. Scale also refers to how agreements, contracts, and jurisdictions are socially constructed. Herod made the point that different scales tend to be produced in relation to each other. For example, when the localized union cultures of early industrial Britain evolved into national union federations, there were particular strategies of local organizations and actors that shaped the mandate of the national union federations. Over time, this influence can work in the opposite direction.

The trade union councils (Landsorganisationene, LO) of the Scandinavian countries are emblematic of the way the trade union movement in many parts of the world evolved into bureaucratic and hierarchical institutions with a strong reliance on the national scale. Typically, LO-Denmark, LO-Sweden, and LO-Norway will have workplace union representatives reporting to district and regional offices, and finally to a national leadership. The national offices of Scandinavian trade unions are very influential in shaping the political direction of the trade union movement. Like other civil society associations and nongovernmental organizations, they tend to mirror the scalar organization of the state apparatus, which forms the key territorial organization of space. In other countries and industries, the development of trade unions has created other scalar configurations, such as the United Auto Workers union, which from the 1930s until an organizational split in 1985 organized US, Canadian, and Puerto Rican workers in the same union (Holmes and Rusonik 1991).

For organized labor in the Nordic countries, strong national federations have facilitated workers’ leverage in collective bargaining, not least due to a high unionization rate and a well-organized employer side. This institutional power has also given the trade union movement significant influence over political decision-making, made possible by their historically close links to national union federations and labor parties. While this model has proven effective in these economies over a sustained period of time, attempts to export a tripartite system of industrial relations have not led to similar successes in contexts where national organizations lack the backing of strong social movements.

As globalization has increased competition between labor markets, driven by the dominance of multinational corporations, national union federations have been put on the defensive. Consequently, activists and scholars alike have been eager to call for a rescaling of union strategies towards the global. Labor internationalism is not a new phenomenon. In addition to the two worldwide union confederations mentioned in the introduction, there is an established organizational infrastructure which includes several industry-specific global union federations, as well as the Council of Global Unions established in 2007. Even though these umbrella organizations have seen some recent success in developing global campaigns and in attempting to secure international framework agreements with multinational companies, their potential impact on the global economy remains unfulfilled (Fairbrother and Hammer 2005).

Nonetheless, union activism at a global scale goes far beyond these formal organizations. Some scholars argue that a “new labor internationalism” can be discerned which distances itself from the hierarchical, bureaucratic, and Eurocentric “old labor internationalism” represented by the institutions above (Webster, Lambert, and Bezuidenhout 2008). One such example is the World Social Forum, which since 2001 has gathered civil society organizations and progressive trade unions annually to discuss global challenges. A more limited and strategic initiative is the Southern Initiative on Globalization and Trade Union Rights (SIGTUR), which is
anchored in particular national trade unions in South Korea, Australia, South Africa, Brazil, India, and other emerging economies in the Global South. According to Webster, Lambert, and Bezuidenhout (2008), this initiative, while still in its infancy, represents an alternative to the European model of internationalism anchored in social compromise and corporate dialogue. International solidarity was also the subject of an insightful study by Johns (1998), who critically examined the intentions and motivations of US unions initiating transnational campaigns. She showed labor internationalism can be motivated by protectionist policies: is, fighting to raise wages in labor markets competing with US workers. This strategy is what Johns labels “accommodationist solidarity,” which she characterizes as regressive in class terms. Alternatively, workers can be motivated by so-called transformative solidarity that transcends place-based interests “in favor of their wider class interests in the international economy” (Johns 1998, 270).

Parallel to debating labor internationalisms, geographers have been exploring new configurations of union power and worker mobilization at subnational and local scales. The importance of communities and neighborhoods has been highlighted by many case studies in labor geography. However, Brecher and Costello (1990) make an important point when they state that, rather than a historical innovation, community unionism should be seen as a return to the early days of trade unionism in the nineteenth century. Beyond the immediate surroundings of the workplace and the home, urban regions are arguably becoming more important as mobilizing arenas for organized labor. For instance, living wage campaigns are based on a distinct rationale that the wages should be able to carry “the costs of social reproduction and care in a particular place” (Wills and Linneker 2014, emphasis added). As costs vary between places, and between rural and urban areas, both strategic and moral reasons can be used to legitimate urban living wage campaigns. By organizing all workplaces across an urban labor market, living wage campaigns can “take wages out of competition” (Walsh 2000). In a ground-breaking study of the living wage campaign in Baltimore, Walsh argues that low-wage workers, whose main threat might not be interregional spatial competition and the threat of relocation, might find this a particularly useful strategy.

As geographers have developed the concept of scale since the 1990s, and debated its relevance, labor geographers have moved away from attempting to locate the ideal political scale in favor of more relational and multiscale approaches. In other words, there is no “right scale,” and political actors need to be attentive towards how different scales are related. Hence, the Liverpool dock workers’ strike in 1995–1998 showcased fundraising with nonunion organizations across the United Kingdom as well as solidarity actions by dock workers around the world. Castree (2000) stressed that this would not have been possible without pre-existing organizational infrastructure at the local and national scale. In a similar fashion, Tufts (2007) suggested that a “spatial circuit of union renewal” was emerging in the hotel sector in Toronto, where support for organizing efforts at the workplace level was bolstered by city-level community action, national coordination, and, finally, international support campaigns. In short, unions do pursue explicitly scalar strategies to secure decent income and working conditions, but these strategies should not be reduced to a quest of scale-jumping.

Union–geographer exchanges

Of course, the salient role of trade unions in geographical research is not simply a result of
analytical curiosity. On the contrary: since the
discipline was radicalized through Marxist, fem-
inist, and postcolonial geographies in the 1980s,
geographers have established close links to their
research subjects, in part motivated by polit-
ical sympathies. The long-term collaboration
between UK geographer Jane Wills and the Lon-
don Citizen campaign, and the research involve-
ment of Canadian geographer Steven Tufts in a
Toronto local of the trade union UNITE HERE
represent only two of many examples.

A question worth asking is whether such
close-knit relationships allow geographers to
maintain academic rigor. The potentials and
challenges of combining academic and activist
subjectivities have been continuously discussed
in seminars and in written form. Such exchanges
of experiences and understanding are part of
a broader critical geography agenda, where
political sympathies are openly announced
while methodological and analytical procedures
are documented and criticized. This balance
between relevance and rigor has arguably given
us accounts of working-class mobilization of
a depth and richness that would be difficult
to attain through a more distanced research
approach. Academic involvement in union
activism combines peer-reviewed publications
with forms of engagement directly relevant
to the labor movement – often directly chal-
lenging their agenda. The journal Antipode has
become an important outlet for geographical
research on working-class mobilization, not
simply because geographers have been able
to publish committed, activist research, but
because unionists themselves have been actively
involved as authors of labor geography. In this
way, union–geographer exchanges share the
ambition of the People’s Geography Project of
US geographer Don Mitchell: “[to] make crit-
ical, radical geography useful to people in their
everyday lives and a resource for those engaged
in the struggle for social and economic justice”
(People’s Geography Project 2016).

SEE ALSO: Civil society; Class; Critical
geography; Economic geography; Labor
geography; Migrant labor; Precarious work;
Rights, labor

References

Brecher, J., and T. Costello. 1990. Building Bridges: 
The Emerging Grassroots Coalition of Labor and

Castree, N. 2000. “Geographical Scale and Grassroots
Internationalism: The Liverpool Dock Dispute,

US Union Elections 1: The Crisis of US Unions
and a Critical Review of the Literature.” Environ-

Cobble, D.S. 1991. Dishing It Out: Waitresses and Their
Unions in the Twentieth Century. Urbana: University
of Illinois Press.

Unionism Prosper in Cyberspace? The Promise of
the Internet for Employee Organization.” British
Journal of Industrial Relations, 40(3): 569–596.

Dublin, T. 1975. “Women, Work and the Fam-
ily: Female Operatives in the Lowell Mills,

Unions: Past Efforts and Future Prospects.”
Relations Industrielles/Industrial Relations, 60(3):
405–431.

Hartwick, E. 2000. “Towards a Geographical Poli-
tics of Consumption.” Environment and Planning A,
32(7): 1177–1192.

the Spatial Diffusion of Swedish Trade Unions,
1890–1940.” American Journal of Sociology, 99(5):
1157–1179.
TRADE UNIONS


Trade, FDI, and industrial development

Godfrey Yeung
National University of Singapore

International trade is the exchange of goods and services, with currency (normally in terms of US dollars, but also in euros and Japanese yen) as the medium of exchange, across national borders (although it could be conducted in forms of barter exchange: the exchange of goods and services with the same perceived value by two partners). Foreign direct investment (FDI) is the direct investment into production or services by a company across national sovereignty for the purposes of entering into the host country’s market and/or expanding or relocating their existing business from the host country.

International trade and FDI influence the direction and value of commodity flows, services, and financial and human capital, both in their countries of origin and at their destinations. The balance between outward (FDI outflows) and inward (FDI inflows) FDI has an impact on the balance of the capital account and the volatility of exchange rates (especially between US dollars, the currency for international trade, and other countries’ currencies). FDI inflows could bring in capital and expertise in manufacturing technologies and management know-how but could also crowd out domestic investment and cause an appreciation of the exchange rate in host countries’ currency. FDI outflows could lead to a repatriation of income generated from foreign investment and the depreciation of the currency exchange rate, which affects the balance of trade (depending on the price elasticity of demand for domestic versus imported commodities) in home countries. The stock and flow of international trade and FDI are thus regarded as the catalysts or sledgehammers for industrial and regional development, and their relationship has long been a bone of contention for academics and policymakers.

The general trends in FDI and international trade are presented in the next section before a review on mainstream theories on international trade and FDI, followed by a brief overview of the contributions of social scientists and economic geographers on the relationship between trade, FDI, and industrial development. An overview of state policies on FDI and international trade and the responses of transnational corporations (TNCs) on such policies concludes this entry.

The general trends in FDI and international trade

The recent global financial crisis has had a significant impact on the trajectory of FDI and trade. FDI inflow into developed countries had declined by 32%, from US$820 billion in 2011 to US$561 billion in 2012, a level that was reached a decade ago (Table 1). The European Union (EU) accounted for two-thirds of the global decline in FDI. However, developing countries played an important role and accounted for more than half (52%) of the US$1.35 trillion global FDI inflows for the first time in 2012. Notably, China and Hong Kong became the second and third most popular destinations of FDI in 2012, directly behind the United States. Brazil and the Russian
Table 1  FDI inflows in different regions, 1990–2012.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td>207,362</td>
<td>1,413,169</td>
<td>1,490,966 1,408,537 1,350,926</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Developed economies</td>
<td>172,514</td>
<td>1,141,586</td>
<td>975,754 696,418 560,718</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>104,403</td>
<td>728,479</td>
<td>685,182 429,230 275,580</td>
<td></td>
<td></td>
</tr>
<tr>
<td>European Union</td>
<td>97,297</td>
<td>701,824</td>
<td>648,495 379,444 258,514</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other developed Europe</td>
<td>7,106</td>
<td>26,655</td>
<td>36,687 49,785 17,066</td>
<td></td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>56,004</td>
<td>380,802</td>
<td>253,555 226,991 212,995</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other developed countries</td>
<td>12,107</td>
<td>32,306</td>
<td>37,016 40,197 72,143</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Developing economies</td>
<td>34,777</td>
<td>264,545</td>
<td>452,023 637,063 702,826</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>2,846</td>
<td>9,621</td>
<td>39,589 43,582 50,041</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asia</td>
<td>22,658</td>
<td>156,581</td>
<td>295,276 400,687 406,770</td>
<td></td>
<td></td>
</tr>
<tr>
<td>East Asia</td>
<td>8,820</td>
<td>125,490</td>
<td>141,243 214,604 214,804</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>12,821</td>
<td>22,641</td>
<td>64,275 97,898 111,336</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latin America and Caribbean</td>
<td>8,940</td>
<td>98,050</td>
<td>116,092 189,855 243,861</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transition economies</td>
<td>71</td>
<td>7,038</td>
<td>63,189 75,056 87,382</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Compiled from the UNCTAD, FDI/TNC database (http://www.unctad.org/fdistatistics).

Federation were ranked fourth and ninth, respectively, which was vastly different from 1990 when European or North American countries dominated the top 10 countries for FDI inflows (UNCTAD 2013a).

Developing countries played an increasingly important role and accounted for about one-third of global FDI outflows in 2012 (Table 2). With the outward FDI increased from only 1% of the global flow (US$7 billion) in 2000 to 10% (US$145 billion) in 2012, the BRICS (Brazil, Russian Federation, India, China, and South Africa) countries have become important recipients and origins of FDI. The EU had, however, only recorded US$323 billion of FDI outflows in 2012, which was about 40% of the value reached in 2000. Although only about two-thirds of the income from global FDI was repatriated back to the home countries, developing countries retained the highest proportion of their FDI income (40%) (UNCTAD 2013a).

The significant rise in FDI also contributes to the increasingly fragmented and dispersed production processes where intermediate goods and services are traded across sovereign boundaries. This leads to the double counting of value in international trade, for example, 28% of US$19 trillion global gross exports in 2010 was counted twice. It is estimated that the value of international trade (after adjustment for double counting) was about 30% of developing countries’ gross domestic product (GDP), on average. The corresponding rate for developed countries’ GDP is 18% (UNCTAD 2013a).

International trade had been increasing rapidly until the outbreak of financial crisis in 2008–2009 (Tables 3 and 4). Although the
Table 2  FDI outflows in different regions, 1990–2012.

<table>
<thead>
<tr>
<th>Region/economy</th>
<th>Value in current prices (millions of US dollars)</th>
<th>2005–2007</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>pre-crisis average</td>
</tr>
<tr>
<td>World</td>
<td>241 421</td>
<td>1 240 316</td>
</tr>
<tr>
<td>Developed economies</td>
<td>229 583</td>
<td>1 090 846</td>
</tr>
<tr>
<td>Europe</td>
<td>139 341</td>
<td>863 805</td>
</tr>
<tr>
<td>European Union</td>
<td>130 571</td>
<td>809 238</td>
</tr>
<tr>
<td>Other developed Europe</td>
<td>8 770</td>
<td>54 568</td>
</tr>
<tr>
<td>North America</td>
<td>36 219</td>
<td>187 304</td>
</tr>
<tr>
<td>Other developed countries</td>
<td>54 023</td>
<td>39 737</td>
</tr>
<tr>
<td>Developing economies</td>
<td>11 838</td>
<td>146 273</td>
</tr>
<tr>
<td>Africa</td>
<td>659</td>
<td>1 534</td>
</tr>
<tr>
<td>Asia</td>
<td>10 943</td>
<td>94 565</td>
</tr>
<tr>
<td>East Asia</td>
<td>9 574</td>
<td>82 108</td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>2 328</td>
<td>8 972</td>
</tr>
<tr>
<td>Latin America &amp; Caribbean</td>
<td>222</td>
<td>50 179</td>
</tr>
<tr>
<td>Transition economies</td>
<td>0</td>
<td>3 197</td>
</tr>
</tbody>
</table>

Source: Compiled from the UNCTAD, FDI/TNC database (http://www.unctad.org/fdistatistics).

The significant increase of exports from developing countries also altered the pattern of economic outputs in the global economy. The share of developing countries in global GDP had more than doubled, from 17% in 1990 to 36% in 2012, while the corresponding share of developed economies had decreased simultaneously, from 79 to 60% (UNCTAD 2013b).
Table 3  Exports in different regions, 1990–2013.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td>3,495,693</td>
<td>6,452,624</td>
<td>12,216,177</td>
<td>15,300,328</td>
<td>18,785,055</td>
</tr>
<tr>
<td>Developed economies</td>
<td>2,534,230</td>
<td>4,239,362</td>
<td>7,206,405</td>
<td>8,242,724</td>
<td>9,555,877</td>
</tr>
<tr>
<td>Europe</td>
<td>1,663,665</td>
<td>2,592,048</td>
<td>4,943,882</td>
<td>5,504,130</td>
<td>6,444,566</td>
</tr>
<tr>
<td>European Union 27</td>
<td>1,563,742</td>
<td>2,448,943</td>
<td>4,667,991</td>
<td>5,172,108</td>
<td>6,055,741</td>
</tr>
<tr>
<td>North America</td>
<td>521,759</td>
<td>1,058,865</td>
<td>1,415,317</td>
<td>1,666,377</td>
<td>2,037,614</td>
</tr>
<tr>
<td>Other developed countries</td>
<td>348,806</td>
<td>588,449</td>
<td>847,207</td>
<td>1,072,217</td>
<td>1,073,698</td>
</tr>
<tr>
<td>Developing economies</td>
<td>843,085</td>
<td>2,059,257</td>
<td>4,553,497</td>
<td>6,436,620</td>
<td>8,412,306</td>
</tr>
<tr>
<td>Africa</td>
<td>104,923</td>
<td>147,903</td>
<td>372,907</td>
<td>521,464</td>
<td>599,525</td>
</tr>
<tr>
<td>Asia</td>
<td>589,761</td>
<td>1,538,421</td>
<td>3,482,606</td>
<td>5,016,322</td>
<td>6,686,685</td>
</tr>
<tr>
<td>East Asia</td>
<td>280,961</td>
<td>779,294</td>
<td>1,860,873</td>
<td>2,725,754</td>
<td>3,619,456</td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>144,148</td>
<td>430,167</td>
<td>763,891</td>
<td>1,050,066</td>
<td>1,270,352</td>
</tr>
<tr>
<td>Latin America &amp; Caribbean</td>
<td>148,400</td>
<td>372,932</td>
<td>6,309,469</td>
<td>898,833</td>
<td>1,126,096</td>
</tr>
<tr>
<td>Transition economies</td>
<td>118,378</td>
<td>154,005</td>
<td>456,275</td>
<td>620,984</td>
<td>816,871</td>
</tr>
</tbody>
</table>

Source: Compiled from UNCTAD, UNCTADstat (http://unctadstat.unctad.org/).

Table 4  Imports in different regions, 1990–2013.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td>3,607,430</td>
<td>6,634,891</td>
<td>12,453,648</td>
<td>15,419,272</td>
<td>18,779,083</td>
</tr>
<tr>
<td>Developed economies</td>
<td>2,669,485</td>
<td>4,637,188</td>
<td>8,061,807</td>
<td>8,926,884</td>
<td>10,172,436</td>
</tr>
<tr>
<td>Europe</td>
<td>1,724,480</td>
<td>2,628,997</td>
<td>5,016,685</td>
<td>5,563,591</td>
<td>6,175,117</td>
</tr>
<tr>
<td>European Union 27</td>
<td>1,625,188</td>
<td>2,507,462</td>
<td>4,797,630</td>
<td>5,305,131</td>
<td>5,876,356</td>
</tr>
<tr>
<td>North America</td>
<td>641,357</td>
<td>1,505,232</td>
<td>2,249,428</td>
<td>2,373,769</td>
<td>2,807,607</td>
</tr>
<tr>
<td>Other developed countries</td>
<td>303,648</td>
<td>502,959</td>
<td>795,694</td>
<td>987,524</td>
<td>1,189,712</td>
</tr>
<tr>
<td>Developing economies</td>
<td>797,814</td>
<td>1,918,053</td>
<td>4,046,868</td>
<td>6,018,888</td>
<td>7,965,661</td>
</tr>
<tr>
<td>Africa</td>
<td>94,658</td>
<td>129,967</td>
<td>311,577</td>
<td>478,621</td>
<td>627,717</td>
</tr>
<tr>
<td>Asia</td>
<td>573,173</td>
<td>1,392,999</td>
<td>3,079,209</td>
<td>4,631,139</td>
<td>6,157,879</td>
</tr>
<tr>
<td>East Asia</td>
<td>268,089</td>
<td>745,115</td>
<td>1,658,281</td>
<td>2,526,502</td>
<td>3,379,445</td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>162,346</td>
<td>380,640</td>
<td>688,622</td>
<td>953,359</td>
<td>1,246,150</td>
</tr>
<tr>
<td>Latin America &amp; Caribbean</td>
<td>129,983</td>
<td>395,086</td>
<td>656,082</td>
<td>909,128</td>
<td>1,180,066</td>
</tr>
<tr>
<td>Transition economies</td>
<td>140,131</td>
<td>99,650</td>
<td>344,972</td>
<td>473,500</td>
<td>640,985</td>
</tr>
</tbody>
</table>

Source: Compiled from UNCTAD, UNCTADstat (http://unctadstat.unctad.org/).
The concept of comparative advantage was first introduced by Adam Smith in 1776 and elaborated by David Ricardo in 1817. Instead of focusing on the cost ratios or absolute cost advantages between two trading partners, Ricardo argued that the cost ratios of producing commodities within each country needed to be considered. Comparative advantage is defined as the capacity of a region/country to produce a specific commodity at a lower opportunity cost (the value that one has to forgo) than that of another region/country. Rather than each country trying to produce all the commodities they need, Ricardo argued that countries can gain from trade if they produce commodities for which they have the comparative advantage and then trade them for commodities from other countries.

The canonical neoclassical Heckscher–Ohlin–Samuelson (H–O–S) model of international trade was developed by Eli Heckscher and Bertil Ohlin from the Ricardian conceptualization of comparative advantage and later enhanced by Paul Samuelson. To address concerns about the restrictive assumptions of Ricardian comparative advantage, Ohlin (1967/1933) relaxed the single-factor assumption in the two-country, two-commodity, and two-factor (labor and capital) model and argued that comparative advantage, geographical specialization, and trade are driven by regional factor endowments in Interregional and International Trade. According to the Heckscher–Ohlin theorem (H–O theorem), countries export products that make intensive use of their production factors (labor and capital) and natural resources that are relatively abundant and relatively inexpensive. In other words, a capital-abundant country exports capital-intensive goods to labor-abundant countries while importing labor-intensive commodities in return.

Neoclassical economists argue that the international division of labor (the specialization of production on a national basis) is a natural consequence of the spatial differentiation of labor productivity (Ricardian comparative advantage) and (Heckscher–Ohlin) factor endowment. As the production factors (labor and capital) are mobile within countries but not between countries (as stipulated in the H–O theorem), and markets are perfectly competitive, factor prices will equalized across countries as a result of international trade, according to Samuelson’s factor price equalization theorem. Therefore, the neoclassical theory of international trade hypothesizes that incomes across countries will converge and development will be even. In other words, free trade promotes economic growth and reduces regional disparities across space.

However, the H–O–S model is unable to reconcile the uneven economic growth observed between developed and developing countries since World War II. The H–O–S model still suffers from a set of unrealistic assumptions: that the production function has constant returns to scale (not accounting for technological advancement), production factors are immobile across space (countries), and markets are perfectly competitive.

To address the lack of solid empirical evidence for the convergence of income across countries hypothesized by the H–O–S theorem, international trade economists, notably Paul Krugman, relax the constant returns to scale assumption and argue that the divergence of income between developed and developing countries is due to the increasing returns to scale of investment among the former in the new trade theory. Paul Romer and Robert Lucas also incorporated increasing
returns to capital arising from the accumulation of knowledge in new (endogenous) growth models.

The H–O–S theorem, augmented by the new trade theory and new growth models, laid the foundation for the neoliberal type of trade and capital account liberalization proposed in the Washington Consensus, which international institutes, especially the International Monetary Fund (IMF), until very recently, insisted that developing countries adopt before receiving financial assistance.

**FDI and industrial development**

The eclectic international production paradigm of FDI (the ownership-location-internalization or OLI paradigm) developed by John Dunning is a synthesis of five relevant economic theories: macroeconomic theory, industrial organization theory, location theory, product life-cycle theory, and internalization theory (see Yeung 2001 for detailed explanations). The OLI paradigm postulates that ownership-specific, locational, and internalization advantages are the dominant factors affecting the pattern of FDI.

The rationales of TNCs for investing in host countries directly through FDI rather than through contractual resource transfers or exportation are presented in Table 5. The ownership-specific advantages (O) are prerequisites for contractual resource transfers, exportation, or FDI by TNCs in their home countries, that is, the exclusive production and technical know-how and managerial and marketing skills possessed by TNCs in their home country are superior to those of rival firms in the host country. Through exporting to or investing directly in the host country, TNCs in their home country can reduce their transaction costs and internalize the externalities in factor or product markets. These internalization advantages (I) are larger than the benefits from licensing production know-how to foreign firms. Moreover, TNCs can relocate from their country of origin due to the pull factors of locational endowments (L), such as when the low labor costs and preferential investment incentives available from the host country are stronger than those in the home country. In other words, TNCs select FDI because it can reap the locational advantages enjoyed by local industry in the host country, where direct exportation and licensing cannot be obtained (see Yeung 2001).

Moreover, Dunning argues that the change in net capital inflow or outflow of a country is determined by its changing economic structure in terms of industry and firm-specific OLI compared with other countries. In fact, he suggests that countries tend to pass through five different stages of development according to their propensity to net inward or outward investment. This propensity is determined by

---

**Table 5** The OLI characteristics of FDI, exports, and licensing.

<table>
<thead>
<tr>
<th>Means of TNCs</th>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ownership</td>
</tr>
<tr>
<td>FDI</td>
<td>✓</td>
</tr>
<tr>
<td>Exports</td>
<td>✓</td>
</tr>
<tr>
<td>Contractual resource transfers</td>
<td>✓</td>
</tr>
</tbody>
</table>

the extent and pattern of the ownership-specific advantages of each country’s indigenous firms, its location-specific advantages, and the internalization advantages relative to the OLI endowments of other countries.

There have been some important additions to the OLI paradigm in response to the significant changes in the global economy during the last two decades. In addition to nonequity alliances or contractual relationships (including outsourcing), Dunning and his associates have highlighted the importance of social relational capital – such as lack of crime and corruption – as institutionally related location advantages for TNCs. Dunning and Lundan (2008) eventually incorporated Douglass North’s (1990) seminal work on the institutional environment – the formal rules (such as laws and regulations) and informal constraints (such as behavioral norms) that govern the OLI advantages – into the eclectic paradigm to allow it to explain the dynamics of FDI flows and home and host countries’ institutional environment effectively.

**Controversies of conventional theories**

The flows of capital and commodities across space could have negative impacts for developed and developing countries. The outflow of FDI, especially the relocation of manufacturing to lower-cost developing countries in East Asia, led to the deindustrialization (“hollowing-out”) and the subsequent massive redundancies in the textiles and clothing, iron and steel, and ship-building industries in the Midlands and the North of England, and the Rust Belt in the Northeast and the East North Central states in the United States in the 1970s and 1980s. The need for developed countries to import their previously locally manufactured goods to sustain the consumption also created persistent trade deficits with East Asian countries (see the previous section).

For developing countries, the significant inflows of FDI and the subsequent boom in international trade also have their costs. The inflow of FDI could crowd out local entrepreneurs who do not have access to the same level of incentive policies implemented by their own countries. These policies could also develop the dependency on foreign capital for economic development in developing countries.

The dominance of TNCs based in developed countries in the economies of developing countries fuels further debates about the effects of inward investments and uneven trade relationships between developed and developing countries. In *Kicking Away the Ladder*, Ha-Joon Chang (2002) highlighted the historical paradox that advanced economies (like the United States and the United Kingdom) used interventionist and protectionist economic policies to accumulate capital during their own early stages of economic development yet bar developing countries from following similar pathways today. He further pointed out that the imposition of a common institutional standard on developing countries with different social norms, cultural values, and conditions is the implementation of a “one-size-fits-all” institutional transplantation with only a 5–10-year transition period. Joseph Stiglitz also joined the debate, arguing that international institutions, such as the IMF, should not force trade and capital liberalization on developing countries without taking local institutions into consideration.

Neo-Marxist social scientists go further, highlighting the exploitative nature of free trade and the subsequent dependency relationship between developed and (newly independent, especially former colonies) developing countries. Rather than focusing on a lack of integration into international trade, as argued by neoliberal
trade economists, dependency theorists, notably Andre Gunder Frank, argue that developing countries are perpetually locked into a world system of core–periphery dependency where the core (developed) countries dominate the advanced technologies and global political and economic institutions while the developing countries are locked into unfavorable terms of trade and remain exploited.

Rather, as a result of FDI and free trade, Neo-Marxist geographers have typically highlighted the relocation of manufacturing sectors from developed to developing countries and the subsequent unequal exchange and the accumulation of capital in core countries. In other words, there is no development for developing countries as TNCs are extracting the surplus value of labor from developing countries through FDI and international trade. Other economic geographers and social scientists examine the increasing importance of international trade and FDI for industrial development from two main perspectives: the agglomeration of economic activities and their development, including the role of knowledge transfer; and the linkages between different economic actors in host and home countries for industrial upgrading.

**Agglomeration, knowledge transfer, and industrial upgrading through trade and FDI**

Industrial agglomeration, the concentration of similar and complementary industrial sectors in a specific area, has long been the focal point of research into trade and FDI flows. Following Marshall’s seminal work on agglomeration, it is generally argued that firms in similar and related sectors benefit from positive externalities of co-location, the external economies of scale, such as cost saving (from shared infrastructures, etc.), knowledge spillover, inter-firm linkages, and the availability of skilled labor due to spatial proximity. International trade and FDI have a great impact on how and where economic activities have tended to cluster in the increasingly globalizing world since the 1980s. Researchers have examined this phenomenon from different perspectives and scales, from the inequality of industrial and economic development in the postcolonial world to industrial districts, knowledge transfer, and the governance of value chains, and so on. Economic geographers, such as Allen Scott, along with business scholar and economist Michael Porter, and Paul Krugman have pointed out the important role the spatial agglomeration of industries has on the globalizing economy.

The ease of overseas procurement through international trade has shaped the characteristics of regional economic activities and their performance. Manufacturers in developed countries are no longer able to compete on cost alone as generic products are readily available from developing economies at much lower costs. The move away from the mass production of Fordism to a flexible form of organization of production was a direct response to the change in consumer demand for differentiated and trendy products in the increasingly globalized world. Regional suppliers are crucial for the “just-in-time” delivery system under the flexible specialization whereby specialized and quality parts and components have to be delivered in frequent, small-volume batches to the assembly plant (Dicken 2011), and this subsequently promotes the development of logistic sectors and the intraregional trade, in addition to the importation of parts and components and the exportation of finished products from and to developed countries.

Constantly upgrading technological capacity is crucial to maintaining the competitiveness of an industry in the knowledge-based economy. The study of how trade and FDI impact on
knowledge transfer and industrial upgrading normally involves examining the institutional factors that play an important role in facilitating various channels of learning and innovation, from the diffusion of tacit knowledge through face-to-face communication and knowledge spillover derived from the economies of agglomeration, to the localization of learning activities. Economic agents have to absorb newly acquired knowledge either through internalizing codified knowledge with the help of tacit knowledge acquired earlier, or through socialization involving the exchange of tacit knowledge (Storper 1995; Dicken 2011).

By referring to various case studies, researchers argue that both Marshallian and Jacobsian externalities are crucial for the development of localized or urbanized economies and the corresponding state policies on FDI and international trade (see Beaudry and Schifferauerova 2009). The specialization of specific industries within industrial districts could be due to Marshallian externalities where potential benefits are generated through intra-industrial knowledge spillovers. The potential benefits of Jacobsian externalities are generated through the co-location of different industries in cities and subsequent inter-industrial knowledge spillovers which stimulate innovation through the interaction of ideas from diverse industrial sectors. As cities are sources of innovation and diverse industrial fabric in close proximity to each other, this fosters opportunities to imitate, share, and recombine ideas and practices across industries. Industrial diversification rather than specialization is crucial for Jacobsian externalities and this could have significant implications for the preferential policies on FDI and international trade (see below).

The importance of highly tacit and idiosyncratic know-how actually creates specific issues of property rights for firms. This is especially the case when the know-how is embedded in the firm’s human capital and thus property rights are difficult to establish, that is, whether the firm or a certain individual working in the firm has ownership of the knowledge. It is therefore in the interests of the firm to convert the knowledge into a codified and transferable form that can be patented, trademarked, or copyrighted. This explains why the regime of intellectual property rights is important for the transfer of codified knowledge, and the importance of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) under the World Trade Organization (WTO) on international trade, especially the unbalanced terms of trade between developed and developing countries where the former normally have the technological/knowledge advantages.

To reconcile the asymmetrical relationships between TNCs in developed countries and their suppliers (local firms) in developing countries, economic sociologists Gereffi, Humphrey, and Sturgeon (2005) introduced five specific types of governance for global value chains (GVCs) to unpack the relationship between actors and industries across space in the 1990s. They argued that the type of product and how and where they are produced is determined by the capabilities of suppliers, the extent of efficient codification, and the transmission of complex information and knowledge. To highlight the nexus of the interconnected functions and operations of firms and nonfirm institutions where goods and services are produced and distributed globally, Coe et al. (2004) proposed the concept of global production networks (GPNs) to highlight the asymmetry of the value capture processes associated with various industrial activities across space as manufacturing activities, core knowledge, and technologies are still ultimately controlled by various TNCs and their first-tied subcontractors. This explains the continuous divergence
of growth between developed and developing countries despite the massive relocation of labor-intensive manufacturing industries to developing countries since the 1970s.

State policies and TNC strategies for FDI and international trade

State policies can impact the flows of FDI and trade. The state policies on FDI and international trade include unilateral incentives, from taxation holidays (i.e., a waiver and/or a discount of profits tax) to direct and indirect subsidiaries (e.g., financial supports for investment in certain high-technology sectors and the provision of physical infrastructures). Another commonly adopted policy is the establishment of special economic zones (SEZs), especially for (high-technology) industrial clusters, normally anchored by foreign-invested companies and supported by locally financed enterprises. The aim for these policies is to encourage the inflow of FDI, and the affiliated expertise on manufacturing and management, and promote export-led growth. Furthermore, central banks could buy US treasury to depress the value of its own currencies against the international trade currency to enjoy more competitive production costs in terms of US dollars (but this strategy could only be effective when other countries are not implementing the same strategy, otherwise there could be a race to the bottom with competitive devaluation of currencies).

To facilitate the export-oriented industrialization, especially the development of infant (new) industries, both developed and developing countries have adopted a number of bilateral and/or multilateral agreements, notably the trade blocs formed by the EU, the North American Free Trade Agreement (NAFTA), and the Association of Southeast Asian Nations (ASEAN). Some countries also adopted unilateral policies to regulate trade through the adoption of tariffs (import tax) and export subsidiaries (normally in the form of a refund of value-added/sales taxes).

Despite the adoption of WTO free trade treaties by most countries, there is a frequent deployment of nontariff barriers to trade (NTBs), notably the antidumping and countervailing measures by various trade partners under the WTO’s provision of protecting local markets for importation of commodities with predatory (below cost) pricing. Moreover, developed countries are increasingly adopting other technical barriers of trade, especially in the form of regulations of health and safety and intellectual property rights, against imports from developing countries. One such regulation is the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) adopted by the EU in 2006, which includes specific regulations on the production and use of chemical substances with regard to their potential impacts on both human health and the environment.

In response to state policies on the regulation of flow of capital and commodities, TNCs relocate or expand their manufacturing sites from their home countries into host countries to enjoy the lower production costs and/or conduct intra-TNC trade, including the use of transfer pricing to exploit the spatial differentiation of taxation regimes. Another common strategy adopted by TNCs is to channel the capital through offshore financial centers (i.e., tax heavens), and this partially explains why the tiny British Virgin Islands, with an area of 153 km² and a population of around 28,000, recorded US$64 billion of FDI inflows (ranked fifth in the world, above the United Kingdom, Australia, the Russian Federation, and Canada) and US$42 billion of FDI outflows in 2012 (UNCTAD 2013a).
The flows of capital and commodities across spaces lead to conflicts between trading partners. This is illustrated by the processing and assembling trade where the high value-added economic activities are conducted and retained in home (developed) countries while the lower value-added economic activities are conducted by subcontractors in host (developing) countries. For instance, less than 10% of the typical high-end mobile phone retail price is created in China but its export value is included in the Chinese export statistics. Chinese suppliers are generally facing an uphill battle for upgrade as they have to rely on the importation of higher value-added parts and components for the assembly of finished products for exportation. Moreover, there are significant discrepancies between China’s trade data and world declared trade data, partly due to the methods of how trade data are recorded (export data are recorded on a freight on board (f.o.b.) basis whereas import data are recorded on a cost, insurance, and freight (c.i.f.) basis) and partly due to the inflated value of re-exports of Chinese merchandise by trading companies in the Hong Kong Special Administrative Region (SAR) of China; for example, more than 95% of export value from Hong Kong in 2012 was subsequently re-exported to other destinations.

SEE ALSO: Comparative advantage; Dependency theory; Flexible specialization; Global commodity/value chains; Global production networks; Industrial districts

References


Further reading

Dunning, John H. 1981. “Explaining the International Direct Investment Position of Countries: Towards a Dynamic or Developmental Approach.”
TRADE, FDI, AND INDUSTRIAL DEVELOPMENT

Weltwirtschaftliches Archiv (Review of World Economics), 117(1): 30–64.

Traffic sensors

Yang Yue
Shenzhen University, China

Traffic sensors are those devices that indicate the presence or passage of vehicles and provide vehicle volume, vehicle speed, and other data that support traffic management, such as signal control, freeway mainline control, ramp metering, and electronic toll collection (Klein, Mills, and Gibson 2006). Traffic sensors can be generally classified as either fixed sensors or mobile sensors. Most traffic sensors are fixed, either embedded in the pavement, such as inductive loops, or hung overhead, such as cameras. Examples of mobile traffic sensors are probe vehicles that record vehicle location and speed by GPS. Since fixed sensors often have limited coverage, mobile sensors are used to complement fixed traffic sensors.

Fixed sensors

In-roadway sensors

In–roadway sensors are those that require installation on, are embedded in, or require installation below the road surface. The two most widely used in–roadway sensors are inductive loop detectors and magnetic sensors.

Inductive-loop detectors are saw-cut into the pavement and can detect vehicles passing or arriving at a certain point, for instance, approaching a traffic light. Since its introduction in the early 1960s, the inductive-loop detector has become the most utilized sensor in a traffic management system because of its relatively low cost. An inductive-loop detector uses a moving magnet to induce an electrical current in a nearby wire to detect the passage or presence of a vehicle (Figure 1). When a vehicle passes over the loop or stops within the loop, the vehicle induces eddy currents in the wire loops, which decrease their inductance. The decreased inductance actuates the electronics unit output relay or solid-state optically isolated output, which sends a pulse to the traffic signal controller signifying the passage or presence of a vehicle.

By sensing vehicle passage and presence, inductive loops can be used for queue detection, vehicle counting, speed measurement, and, thus, traffic signal control. Other basic traffic flow parameters, such as volume, occupancy, headway, and gap, can also be inferred from algorithms. In general, this technique provides one of the best accuracies for count data compared with other commonly used techniques, and is a common standard for obtaining accurate occupancy measurements (Gordon and Tighe 2005). However, since inductive-loop detectors are saw-cut into the pavement, the analysis of a malfunctioning inductive-loop detector can be difficult. They are easily affected by pavement failure and damage caused by construction activities, rain, and snow.

Magnetic sensors are an alternative to inductive-loop detectors. They are usually placed underneath a paved roadway or bridge structure. A magnetic sensor is a passive device that detects the presence and passage of a vehicle by measuring the perturbation in the Earth’s quiescent magnetic field caused by ferrous metal objects (Figure 2). Two types of magnetic field sensor are used for traffic flow parameter measurement. The two–axis fluxgate magnetometer
detects stopped and moving vehicles, and the induction or search coil magnetometer typically detects moving vehicles.

Two-axis fluxgate magnetometers detect changes in magnetic field anomaly produced by a ferrous metal vehicle. Their outputs are connected to an electronic unit that measures the output voltage, that is, the magnetic signature of a vehicle. When the voltage exceeds a predetermined threshold within the detection zone, the presence or stopping of a vehicle is detected (Klein 2001).

Induction or search coil magnetometers generate a voltage when a moving ferromagnetic object perturbs the Earth’s magnetic field. Therefore, induction magnetometers only detect moving vehicles and cannot detect stopped vehicles. However, multiple units of some magnetic detectors can be installed and utilized with specialized signal processing software to generate vehicle presence data.

**Overhead sensors**

Overhead sensors are mounted above the surface of the roadway, or alongside the roadway, a certain distance from the target traffic lanes; these may be, for example, video image processors (VIPs), microwave radar, ultrasonic, passive infrared, and laser radar sensors. In some wide-area surveillance systems, overhead sensors are mounted on tall buildings, radio towers, and even aerial platforms.
Video detectors are based on real-time video image processor systems to report vehicle presence, vehicle type, volume, lane occupancy, and speed for each class and lane (Romero et al. 2011). Video detector systems normally perform two main tasks: (i) the inference of road geometries and (ii) vehicle and obstacle recognition. Even without its association with VIPs, closed-circuit television (CCTV) has become an important component for traffic management. High-definition and megapixel cameras have been used in CCTV.

The core uses of VIPs for traffic surveillance are vehicle detection, classification, and tracking algorithms. Traffic flow parameters are calculated by analyzing successive video frames. Vehicle speed can be estimated by identifying vehicle travel times within the detection zones. Vehicle tracking can potentially provide a link between travel time and origin–destination pair information. VIPs that track vehicles may also have the capability to register turning movements and lane changes (Klein 2001). The effectiveness of VIP algorithms can be affected by weather conditions, shadows, and daytime or nighttime artifacts (JPL 1997). An effective VIP should meet the following requirements (Kastrinaki, Zervakis, and Kalaitzakis 2003):

- automatic segmentation of each vehicle background and other vehicles, so that all vehicles are detected;
- able to detect correctly all types of road vehicles, motorcycles, cars, buses, construction equipment, trucks, and so on;
- function in a wide range of traffic conditions: lights, congestion, varying speeds, different lines;
- function under a wide variety of lighting conditions: sunny, cloudy, sunset, night, rain, and so on.

Microwave radar sensors: radar was originally an acronym for “RAdio Detection And Ranging.” Most roadside microwave radar sensors utilize the X-band frequency of 10.525 GHz. The effective area, that is, beam width, of a radar sensor is controlled by the size and the distribution of energy across the aperture of its antenna. Higher frequencies illuminate smaller ground areas with a greater spatial resolution given the size of antenna. When a vehicle passes through the antenna beam, a portion of the transmitted energy is reflected back towards the antenna. The energy then enters a receiver that calculates traffic volume, speed, and vehicle length. Thus, microwave radar sensors are usually used to detect vehicle speed, or in electronic toll collection and automated truck weighing applications that need vehicle length information. Two types of microwave radar sensor are used in roadside applications: those transmitting continuous wave (CW) Doppler waveforms and those transmitting frequency modulated continuous waves (FMCW).

CW Doppler radars measure vehicle speed using the Doppler principle by transmitting a stable frequency signal with respect to time. The frequency of received signal increases when a vehicle moves toward the radar and decreases when a vehicle moves away from the radar. Thus, CW Doppler radars can detect vehicle passage and measure vehicle count and speed, but cannot detect stopped vehicles, that is, vehicle presence.

FMCW radars detect vehicle passage as well as vehicle presence. The transmitted frequency of the sensor is constantly changing with respect to time. They can detect stopped vehicles and provide measurements of lane occupancy, vehicle count, speed, and vehicle length. FMCW radars can also use Doppler to calculate the speed of moving vehicles (Edde 1993).

Infrared sensors emit and/or detect infrared radiation to measure the heat of vehicles or
TRAFFIC SENSORS

passengers. Most infrared sensors only measure infrared radiation, rather than emitting it, and thus are known as passive infrared (PIR) sensors; IR sensors that emit and detect are known as active infrared (AIR) sensors.

AIR sensors illuminate detection zones with low-power infrared energy in the near infrared region of the electromagnetic spectrum at 0.85 mm. PIR sensors do not generate or radiate any energy for detection purposes. Instead, they work entirely by detecting the energy given off by other objects, such as vehicles and road surfaces. PIR sensors with a single detection zone measure volume and lane occupancy by sensing vehicle passage and presence. Those with multiple detection zones can also measure vehicle speed and length. Real-time signal processing is needed to analyze the signals for the presence of a vehicle. In traffic management applications, passive sensors may encounter performance degradation due to (i) glint from sunlight, (ii) inclement weather, and (iii) atmospheric particulates.

Acoustic/ultrasonic sensors can be used for vehicle detection and in driver-assisting systems. Similar to radar, the sensors produce a beam of sound and evaluate the echo that is received back by the sensor. Ultrasonic sensors generate high-frequency sound waves (between 25 and 50 kHz), which are above the human audible range. Sensors calculate the time interval between sending the signal and receiving the echo to determine the distance to an object (Ni 2013).

Most ultrasonic sensors operate with pulse waveforms and provide vehicle count, presence, and occupancy information. Pulse-shape waveforms measure distances to the road surface and vehicle surface by detecting the portion of the transmitted energy. When a distance other than that to the background road surface is measured, the sensor interprets that measurement as the presence of a vehicle. Acoustic/ultrasonic sensors cover only a short range and have slow response times. Accuracy is limited by the surface of the objects. Temperature change and extreme air turbulence may also affect the performance of ultrasonic sensors.

LiDAR (Light Detection and Ranging) uses ultraviolet, visible, or near-infrared light in the form of a pulsed laser to measure ranges (variable distances) and image objects. A basic LiDAR system consists of an emitter, mirror, and receiver (Figure 3). The emitter sends out a laser beam that bounces off a mirror that rotates along with the cylindrical housing at a certain speed. After bouncing off objects, the laser beam returns to the mirror and is bounced back towards the receiver, where it can be interpreted into data.

LiDAR has been used extensively to make high-resolution maps in geomatics, archaeology, geography, geology, forestry, remote sensing, atmospheric physics, airborne laser swath mapping, laser altimetry, and contour mapping. Thus, it is often mounted on top of a driverless car in order to generate an accurate map of the car’s surroundings. It is the heart of object detection and obstacle avoidance. In transportation applications, LiDAR sensors produce 2-D and 3-D imagery of vehicles to identify vehicle width, speed, and height, and thus, for example, can be used on toll roads for vehicle classification. They can also provide vehicle presence, volume, speed, length assessment, queue measurement, and classification at traffic signals.

In summary, sensors using different detector technologies are useful in various transportation applications to provide a wide variety of detection parameters and classification information from a single location. The inductive-loop detector is, by far, the most widely used sensor in traffic control systems. Meanwhile, overhead sensors have provided an alternative to in-roadway sensors. The traffic flow parameters measured with overhead sensors
satisfy the accuracy requirements of many freeway and surface street applications. However, a requirement for in-roadway sensors continues, though the installation and maintenance of in-roadway sensors can disrupt traffic and pose a safety risk to the installers. Besides axle counting and weigh-in-motion applications requiring sensors under or on the road surface, cost and safety issues associated with mounting overhead sensors are reasons for choosing or installing in-roadway sensors. Aesthetic consideration is another consideration. Thus, sensor type, mounting height and location, vehicle mix, road configuration, and sensor viewing angles must be analyzed with respect to the intended application (Klein, Mills, and Gibson 2006).

**Mobile sensors**

In contrast to fixed sensors, which collect traffic parameters from a single location, mobile sensors allow for continuous data collection, such as probe vehicle and cellular phone positioning data. They make large-scale data collection possible, including those where data were previously unavailable.

**Probe vehicle**

The probe vehicle is also known as floating car. It is primarily designed for vehicle dispatch, travel time collection, and route guidance applications. The probe vehicles can
be personal or public transit, or commercial vehicles (such as taxis). Probe vehicles may be equipped with several different types of electronic transponders or receivers. For example, a signpost-based system that relies on transponders attached to roadside signposts can be used for tracking bus locations. Electronic tags, such as RFID tags, are used for automatic vehicle identification for those vehicles equipped with tags. These tags communicate with roadside transceivers to identify unique vehicles and collect travel times between transceivers (Turner et al. 1998). Nowadays, GPS has become the most widely used device. The vehicle’s location, speed, and heading are obtained by GPS and transmitted to a control center in real-time.

Cellular phone

An emerging potential source of traffic flow data is from cellular telephone companies who monitor the transmitting status of telephones that are engaged in conversations in support of wireless-enhanced all-automatic location identification. The location of these telephones can potentially be made available to traffic management agencies to anonymously track vehicles on a noncooperative basis. The travel speed or travel time of cellular telephone users can be determined by time-difference-of-arrival techniques. This information can also assist in estimating congestion and travel time over wide areas, while protecting the identity of the telephone subscriber (Turner et al. 1998, Klein, Mills, and Gibson 2006). However, privacy issues become a concern, and coordination is often necessary between the agency responsible for the system’s operation and the agency that would like to utilize the system for travel time data collection.

Other sensor platforms

One emerging source of traffic monitoring data is from satellite, aircraft, and unmanned aerial vehicles. Information gathered from these platforms can be used to estimate arterial and freeway traffic characteristics either in real-time or over long time scales and large geographic areas, especially in those situations that cannot be easily accessed by humans and vehicles, such as estimating road damage caused by an earthquake.

Another source of traffic monitoring data is vehicle-based sensors, for example, event data recorder and on-board diagnostics. These sensors are not regarded as typical traffic sensors, but they may become important sources for collecting traffic data in the future with the continued evolution in the self-driving car, connected vehicle, and cooperative vehicle-infrastructure system.

SEE ALSO: LiDAR; Sensor networks, wireless

References


Trajectories: analysis

Noam Shoval
The Hebrew University of Jerusalem, Israel

Data about movements of objects are often collected as trajectories in space and time, that is, the movement of each object is recorded as a series of geographic locations with respective time stamps, when the object moves in the basic three-dimensional (3-D) framework of our physical world, defined by geographic space and time.

Trajectories in the geographic context usually refer to the space–time paths of people or cars and are mostly related to the conceptual framework of time geography proposed by Torsten Hägerstrand (1970) that integrates time, as a limited resource, into the thought on spatial behavior.

Qualitative analysis of trajectories

One of Hägerstrand’s most profound professional and disciplinary achievements was the ability to represent space and time in a single diagram, unlike an ordinary map, but rather like a snapshot, that reproduces a moment frozen in time. The result was his, now famous, time-geographical diagrams: notational (representational) systems, which formed the basis of much of the subsequent work in the field of time geography, particularly in the realm of analysis and interpretation. The diagrams (Figure 1) consist, as a rule, of two axes: a time-axis and a space-axis, thus making it possible to trace in graphic terms individual time budgets. The effective range of each person is described by a prism, or a series of prisms, whose shape is dependent upon the aforementioned capability constraints. Hence, every pause, regardless of the activity involved, will cause the prism’s (or subprism’s) range to shrink in direct proportion to the time spent at said stop. But there are also other wider structural features, specific to the social systems within which individuals operate, which, as has long been recognized, help shape people’s time budgets and activity patterns (Neutens, Schwanen, and Witlox 2011).

The new computation abilities that have become available within the past two decades have opened new possibilities in the ability to plot space–time data. Kwan (2000) was among the first to produce a “space–time aquarium” within a 3-D geographic information system (GIS) environment using individual activity travel diary data. These images have improved the capability to reveal the characteristics of space–time patterns belonging to different population subgroups, and they have also assisted in improving the ability to identify common patterns.

However, one basic problem in trajectories analysis is the limited ability to aggregate space–time paths in order to create generalized types of trajectories composed of varied activities in order to create patterns fashioned on a quantitative basis while taking into account the sequential element (Andrienko and Andrienko 2010). Previous attempts with quantitative pattern aggregation methods, mainly by transport researchers, did not manage to tackle the issue of the sequential element. Understanding the sequence of activities in space and time allows one to understand an additional integral dimension of activity and to recognize patterns that exist within this dimension.

Geography, like many other research and application fields, has moved in recent decades from being a data-poor field to being a data-rich field. This transition happened with the development of fields such as GIS and remote sensing, and today public and private sector agencies are creating larger and larger datasets. This change made it imperative to develop new data mining tools to extract and construct knowledge from the huge databases that otherwise will remain undeciphered. The introduction of location devices, such as cellular phones, has created databases of human activities in space–time in great detail, but due to their large size there is need for new techniques in order to be able to extract patterns of behavior from within those large databases.

**Quantitative analysis of trajectories**

Attempts to aggregate behaviors into patterns and to analyze behaviors have mostly been descriptive; only a handful of studies introduced analytic methods to tackle the new type of high-resolution data available. The realm of trajectory pattern analysis offers many tools that can contribute to the analysis of spatial behavior. Research in this field grows out of the increasing availability of spatiotemporal data and the lack of techniques to analyze it (D’Urso and Massari 2013). The basic entities in this line of research are moving objects’ trajectories that have a defined beginning and end time, and are divided into movement segments by stops–pauses in movement, identified in accordance with the scale of analysis. Yet, when considering semantic trajectories, one must consider background geographic information in order to understand and model trajectory patterns.

Trajectory pattern analysis attempts to resolve three major issues: how to measure similarities between trajectories, how to recognize clusters of similar trajectories, and how to treat the sequential element of a trajectory (Grinberger et al. 2014). Many measures have been developed to identify the similarity of trajectories—distance-based measures, dynamic time wrapping, and longest common subsequence being some of the most frequently used (Chen, Özs, and Oria 2005). A variety of clustering methods are also used in pattern analysis, following the methods used for clustering of static data: direct methods that create clusters according to a central value or entity ($k$-means, $k$-medoids, fuzzy $c$ means); agglomerative methods, which use a hierarchical tree to cluster trajectories together until a certain threshold is reached.
(hierarchical clustering); divisive methods, which work in a top-down manner, dividing an entire database into a number of clusters; density-based methods, which extend the cluster’s reach until a certain threshold number of objects is crossed; and so on. In order to deal with the elements of time and sequence, three approaches are available (Liao 2005): a raw-data based approach requiring substitution of standard similarity measures with ones that take time into account; the feature-based approach, in which raw data are converted into a feature vector, upon which standard clustering methods can operate; the model-based approach, which likewise converts raw data into prespecified model parameters.

The analysis of trajectories and pattern mining has proven to be useful in many areas aside geography. The fields of travel research and transport planning may be the most relevant fields, with some examples being aggregation of trajectories into trajectory patterns, based on regions of interest (predefined or discovered from data), and the sequence and duration of movement between them (Giannotti et al. 2007); identification of mobility profiles – the routine routes of an individual based on data-defined stops and the sequence between them (Trasarti et al. 2011); and the work done by Renso et al. (2013), which combines a predefined mobility behavior ontology and movement patterns aggregated from trajectories (measured, synthetic, or semantic) in order to classify patterns into behavior classes. Lately, this form of analysis has proven useful when implemented within the framework of time geography, as is done by the Activity Pattern Analyst, developed by Chen et al. (2011).

Sequence alignment methods

Based on the principle of comparing sequences (strings), sequence analysis was developed during the 1980s for use in the natural sciences and was utilized primarily by biochemists to, among other things, analyze DNA sequences. It was adapted for use in the social sciences sometime toward the end of the 1990s (Wilson 1998). Unlike conventional quantitative methods, sequence analysis, as its name suggests, tackles the problem of sequences directly. Sequence analysis could, if applied properly, complement the methods used to collect, present, and analyze temporal and spatial data, particularly the data amassed in the course of empirically based time geography research.

Traditional quantitative methods band similar objects together on the basis of specific shared characteristics, but they cannot expose the hidden patterns buried within sequences (Wilson 1998). Unlike the more generally accepted methods of sequence comparison in which the distance between two sequences of activities is calculated by means of Euclidian-based geometry, such as Euclidean distance, city-block distance, or Hamming distance (the number of positions in which corresponding elements are different), sequence analysis computes the distance between the two on a “biological” basis (Bargeman, Joh, and Timmermans 2002).

The algorithm used in sequence alignment to measure the degree of similarity between two sequences utilizes three elementary operations: insertion, deletion (two operations, which are, on occasion, referred to singly as an “indel”), and substitution (switching the places of two characters). By applying these three processes to one of the sequences, one string is made identical to the other. The more operations needed to make the sequences identical, the longer the distance between the sequences. Hence, the longer the distance between groups, the smaller the similarity. Thus, sequence alignment methods measure the degree of difference between two sequences in terms of their element composition and sequence.
Humanspatial behavior is the sum of three parallel dimensions: the “what,” the “when,” and the “where.” The “what” describes the activity taking place, the “when” the temporal dimension of said activity, and the “where” its spatial aspects. Geographical analysis combines these three dimensions to reproduce and understand part of the complex reality of human behavior. Those social scientists who have used sequence analysis to date have tended to focus on the relationship between the “what” (activity) and the “when” (time). In their work, the temporal dimension forms the basis of the sequences analyzed, with each unit (character or word) representing an activity carried out within a specific time frame. Incorporating information such as the frequency, duration, and timing of an activity, the temporal dimension is as a rule analyzed by traditional quantitative methods (Bargeman, Joh, and Timmermans 2002). The location component “where” can be approached in two different ways. The first is “categorized” locations, descriptions of the functionality of the location, such as home, work, and store. Although these words describe locations, they describe the functionality of the location and not the geographical location. These locations are also not unique. Many different homes exist in different locations, and their geographical location cannot be derived from this kind of location description. (For examples of research applying this approach, see Bargeman, Joh, and Timmermans 2002; Stovel and Bolan 2004; Shoval and Isaacs 2007; Shoval et al. 2015.) The second way to approach the location component of activity is by using geographic locations. This approach creates locations that are unique in the sense that each set of geographic coordinates symbolizes a specific location in space.

The method offers an effective means to extract sequence patterns from trajectories of human activities and by doing so may also present new ways of analyzing such data. It is equally worth noting that the relatively new and more accurate digital methods of collecting spatial data (i.e., GPS) produce databases that are characterized by extremely high temporal and spatial resolutions. This reinforces the observations to the effect that recent developments in the field of location aware technologies (LAT) and location-based services (LBS) could trigger an even wider resurgence in time-geographic studies (Miller 2005; Wilson 2008).

SEE ALSO: Representation: trajectories; Time geography and space–time prism

References


Transaction costs are costs associated with effecting an economic transaction, either through market exchange between two or more legally distinct economic actors, or internally within a single organization (firm, or, more generally, “hierarchy”).

Costs for market-mediated transactions might include:

- the cost of gathering information concerning the availability, price, quality, and so on, of particular commodities (goods or services);
- the costs associated with identifying potential customers;
- the costs associated with estimating the reliability and credit of suppliers and buyers;
- the costs associated with negotiating the terms of an exchange, including price, delivery date, and terms of payment (and of setting these out in the form of an agreed contract);
- the cost of completing a transaction (marking or collecting payment); and
- the costs accompanying nonmarket transactions within an economic actor (like a single firm). These are mainly the costs of organizing and coordinating complex, multistep production processes in-house.

The idea of transaction costs was proposed by Ronald Coase (Nobel Laureate in economics) in 1937. In his paper *The Nature of the Firm*, he defines transaction costs as the most obvious costs of organizing production through a price mechanism. It’s the cost of finding all relative prices. In other words, transaction costs are the expense of the negotiating and signing of every transaction that has happened in the market, and other costs associated with using the price mechanism (Coase 1937). Market and firm are two kinds of organizing forms of economic activity, and the cost of market-mediated transactions provides a need for the existence of the firm. Firms can reduce market transaction costs through strengthening vertical integration (Matthews 1994), which will increase management costs at the same time. So the boundary of the firm is determined by the market transaction costs and firm organization costs.

Based on Coase, Oliver Williamson, the 2009 Nobel Laureate in economics, analyzed and investigated the determinants of transaction costs, enriching and perfecting transaction cost theory. Because of his valued contribution, transaction cost economics was established and transaction costs have been widely used in the economic analysis paradigm. *The Economic Institutions of Capitalism* (Williamson 1985), his writing on transaction costs, has become an economic classic, and he has become the master of transaction cost theory.

Williamson divided transaction costs into costs before and after contract business (Williamson 1985). The former includes the costs associated with drafting a contract, negotiating the contract, and guaranteeing the contract can be performed. The latter includes maladjustment costs, bargaining costs, constructing costs, performing costs, and guaranteeing costs.
As to why transaction costs exist, there are subjective reasons and objective reasons. Subjective reasons include bounded rationality and opportunism. Bounded rationality means that people participating in trading activities will be subject to physical, intellectual, emotional, and other types of restriction, which will limit the benefit maximization. Opportunism refers to people involved in the transactions consciously acting with the purpose of seeking self-interest, which increases mistrust and suspicion. Objective reasons include asset specificity, uncertainty, and frequency of transaction. Asset specificity means that the investment asset of the exchange does not have market liquidity. In other words, the cost of investing in the asset is difficult to recycle or convert to another use once the contract is terminated. The uncertainty means that the transaction process is very complex, which will generate a lot of uncertainty and risk. As to frequency, the higher the frequency of transactions, the higher the corporate management and bargaining costs.

However, the definition of transaction costs among academics is not completely consistent, and neither is the scope that transaction costs cover. The term “transaction costs” was first and clearly proposed by Kenneth Arrow in 1969. He argued that uncertainty and negative externalities exist in people’s communicative behavior. As a result, the social system emerges. The costs of using the system are transaction costs (Arrow 1969). Yoram Barzel argues that transaction costs are the costs of defining and maintaining property rights (Barzel 1997), while the broadest definition of transaction costs is made by Steven Cheung, who argues that transaction costs are all kinds of costs that can’t exist in a Robinson Crusoe economy, namely “the system cost” (Cheung 1987). Not surprisingly, there are also many other scholars who have created their own definitions of transaction costs. Although there are differences between these definitions, they are the same in essence.

Though derived from economics, transaction costs have also been widely used in economic geography. In the mid-1980s, Allen Scott, the American economic geographer, introduced transaction cost theory in urban and regional study to analyze industry clusters, forming the new industrial location theory which had a far-reaching influence. He argued that projects with low unit transaction costs would organize production in places with a cheap labor force in the form of enterprise branches (Scott 1983, 1986). By contrast, those projects with high unit transactions costs would locate in big cities or economic centers in the form of market transactions. That is to say, different kinds of firms have different location patterns. Some will agglomerate in big cities while others choose to stay in smaller cities, which will directly influence the urban and regional form.

Transaction cost theory can be applied to understand the microfoundation, structure, essence, and evolution of industry clusters (Iammarino and McCann 2006). In order to cope with sharp fluctuations in the external environment, firms should manage input–output relations with other firms. This will generate transaction costs and create pressure for agglomeration (Storper and Venables 2004). Besides, when an industry cluster forms, it can significantly reduce transaction costs in many ways.

First, geographical factors will influence transaction costs (Scott 1988). After firms agglomerate within a district, the space transaction costs will decrease accordingly (McCann 1995). Second, firms in an industry cluster can share common service providers and suppliers. Being geographically close to each other can help suppliers reduce transaction costs, which will also help firms in the cluster reduce the cost of finding potential suppliers and customers (Bathelt, Malmberg,
and Maskell 2004). In addition, it’s easy for firms to cooperate frequently when they agglomerate within a district. This will increase trust between them and promote information-sharing to enable nonroutine transactions (Harrison 1992; Storper 1997; Nooteboom 2006). Furthermore, the local culture that results from firms’ interactions will bring intangible restrictions to the trading activities between enterprises, which will also reduce transaction costs.

SEE ALSO: Accessibility, employment; Firms; Industrial agglomeration; Industrial complex; Industrial districts; Industrial linkage

References


Transcripts: coding and analysis

Meghan Cope
University of Vermont, USA

Transcribing

The term “transcripts” refers to the product of transcribing, which is the process of researchers (or specialists they have hired) creating a textual version from an audio or video recording of some kind of interaction, media report, or research event. Transcribing is commonly used in qualitative research projects when researchers want a written version of their interactions with participants, or from other audio sources such as radio reports, advertising, speeches, or television shows, for the purposes of analysis. Common qualitative research interactions that lend themselves to recording and transcription include focus groups, interviews, oral histories, participatory mapping sessions, photovoice, and any other technique which involves audio content that might be recorded and turned into text. Another form of transcripts that should also be acknowledged here is the transcription that is typically employed in archival work, when, for instance, a researcher creates a typed, digital copy of historical documents written by hand, such as letters, diaries, or other manuscripts. While the process of transcribing may be somewhat different – straining to read faded or archaic handwriting rather than straining to hear a mumbled word – the resulting document is in essentially the same format and can be analyzed in similar ways to transcripts made from audio material. Because there are fewer technical and decision-making points to concern the researcher working with historical manuscripts, many of the comments in this entry focus on contemporary work involving audio recordings, but the analysis principles remain largely the same between various data sources.

The technical aspects of audio recording have become simplified with the development of digital audio technology. Most cellphones can now be used as audio recorders, though for higher quality recordings special equipment may be desired, such as multidirectional microphones placed in the middle of a table during a group exercise. Although researchers will still be faced with the challenges of managing audio files (saving them, organizing them, keeping metadata on them, and insuring their security), digital files are still an improvement over the flimsy, heat-sensitive, and cumbersome audio tapes that were in use until quite recently (and, perhaps, still are in some places). Digital files should be backed up on alternate drives, the “cloud,” or other secure locations, so as not to mistakenly delete hundreds of hours of interviews with the swipe of one cup of tea spilled on the researcher’s laptop. As with all data that may be personal, politically sensitive, or whose violation could potentially endanger the researcher or her/his informants, security measures should be taken seriously. A final note on technology: while there are various software programs and applications that can at least partially automate transcription, they are (as of 2016) still limited to being essentially dictation transcribers – that is, they need to be “trained” by the speaker and, thus, are really only useful for things like transcribing.
audio-recorded field notes, or other instances when only one person (usually the researcher her/himself) is talking. Of course, the field of digital technologies is changing very quickly and it may well be possible to completely automate transcription with sufficiently high levels of accuracy in the coming years.

The practice of transcribing, therefore, involves creating, as accurately as possible, a fair written version of the verbalizations generated in a research encounter with participants, or gleaned from other sources of talk, such as media reports. Transcription typically also involves researchers paying attention to other, nonword audible expressions, such as pauses, laughter, scoffing, sighs, emphasis, and talking over one another. These can be very important for interpreting the meanings of the text and, thus, should not be ignored. At another level, researchers often need to account for movements, facial expressions, directions of the gaze, and other “body language” performed by participants; this can be done either from video recordings or based on written notes taken by the researcher or an assistant during the audio-recorded interview/focus group. There is a system of commonly used markings to indicate these nonverbal expressions, which is reprinted in the appendices of many methods texts (see Silverman 2011, 465–466 for some notation conventions of transcription).

A quick review of the many methodology discussion boards and online forums (Listserves) reveals that there are numerous approaches, many potential pitfalls, and countless spirited debates about various strategies of transcribing, the full extent of which is beyond the scope of this entry, but included here are a few cautionary findings. First, there are professional transcription services that can be hired at varying costs and – undoubtedly – with varying levels of skill. Second, there is a lot of discussion and debate on the topic of (abovementioned) digital transcribers, from phone apps to full software packages, which again have varying costs and levels of accuracy. Third, the biggest downfall for both professional (human) transcription and automated digital transcription is the misinterpretation of sounds and verbalizations because the transcriber is not familiar with the local place names, acronyms, technical terms, or other specialty references that are typically sprinkled throughout interviews and research sessions, and may not adequately represent nonverbal elements that can greatly impact the meanings of the words. This leads to a final observation: through all the online debates about the merits of various outsourced and automated transcribing approaches, the gold standard seems to be that the researcher does her/his own transcribing, or at least performs a careful check on each transcription. There are two reasons most frequently cited for this: (i) the researcher is going to have the best knowledge of those place names, acronyms, and special terms that are inevitably uttered in interacting with research participants, and will be able to discern more of the implied meanings of pauses, laughter, and so on; and (ii) by doing their own transcribing (or careful review of the text while listening to the recording), researchers get closer to their results than is possible just by reading someone else’s transcription and – indeed – they will likely even get a start on the process of sorting, organizing, and analyzing data by noticing patterns and themes just through the transcribing process.

Of course, transcribing can be tedious and is certainly time-consuming. Transcribing one’s research recordings is even, in some circles, a bit of a rite of passage for new scholars, generating the requisite “survival” stories to be shared and sympathized with afterward. These are all points to consider when approaching such a task. Most books on qualitative research methods, in geography as well as other disciplines, have
sections on the joys and sorrows of transcribing, as well as helpful hints on techniques (Kitchin and Tate 1999).

**Putting transcripts to use: making sense of texts**

Diaries, notes and letters, policy statements, newsletters, open-ended survey questions, even photos, performances, and art installations are rich sources of insight that have been creatively employed by human geographers and other social/cultural researchers. However, as the diversity of qualitative research methods has increased over the past two decades, particularly with the inclusion of more participatory methods and the expansion of digital forms of representation, researchers are increasingly in situations where even this broad array of sources is expanded. The digital age is generating new sources of qualitative data, such as tweets, social media posts, sketch maps, images, blogs, video, and so on; these are increasingly spatially referenced, allowing new dimensions of contextual understanding. Thus, the remainder of this entry focuses on methods of coding and analysis that can be used for transcripts, but acknowledges that the organization and analysis functions can be applied to a much wider set of diverse data forms.

While transcribing itself presents some technical challenges, the value of investing time and energy into producing good quality transcripts comes through once the researcher begins trying to make sense of the data. There are many ways to analyze and understand transcriptions, and more or less systematic ways of organizing and understanding the raw material. One common approach to making sense of transcripts, which will be the main focus here, is to construct a system of *codes* as a way to comprehend data. Codes are comprised of *categories* of data and *themes* that the researcher is interested in. For categorization, data are examined and grouped based on similarities (e.g., all mentions of “part-time work” or all experiences related to migration), and each category is assigned a code. For thematic coding, the data are assigned codes based on instances related to themes that are important to the research project (e.g., themes of employment or transnational identities). Beyond just thinking of codes as “categories” or “themes,” there are many other ways to structure a coding system, such as distinguishing between descriptive and analytic codes, “latent” and “manifest” codes, and considerations about how codes relate to each other (are they hierarchical, networked, or parallel?). Many excellent sourcebooks cover the issue of creating coding structures more thoroughly than is possible in this space (Corbin and Strauss 2007; Dey 1993; Miles, Huberman, and Saldana 2014; Saldana 2013; and, in geography, Cope 2009, 2016; Kitchin and Tate 1999). Many of these authors reiterate what Coffey and Atkinson identify as the strengths of coding, which go well beyond mere organization and search functions:

> Coding need not be viewed simply as reducing data to some general, common denominators. Rather, it can be used to expand, transform, and reconceptualize data, opening up more diverse analytical possibilities… The general analytic approach here is not to simplify the data but to open them up in order to interrogate them further, to try to identify and speculate about further features … it is intended to expand the conceptual frameworks and dimensions for analysis. *Coding here is actually about going beyond the data, thinking creatively with the data, asking the data questions, and generating theories and frameworks.* (Coffey and Atkinson 1996, 29–30; emphasis added)

However, a few guidelines and suggestions on coding can provide some starting points.
TRANSCRIPTS: CODING AND ANALYSIS

First, consider the purposes of coding. At its most basic, coding is about breaking data into smaller pieces and then recombining those pieces in order to identify and explore relationships and discover new connections. Imagine how tedious and long a paper would be if you included the full transcripts of, say, twenty interviews or eight focus groups and expected your readers to hold all those accounts in their heads as you made your conceptual argument! Instead, coding allows us to break down those transcripts, distill raw data down to the most pertinent findings by splitting them into constituent parts, and make clear connections between our empirical and theoretical discoveries. Watson and Till phrase it as: “An iterative analytic practice, coding allows the ethnographer to break apart, relate, and recombine the materials generated in the research process” (Watson and Till 2010, 128, emphasis added).

More specifically, coding helps researchers to: organize their data in more digestible forms, perform searches and find key phrases, pull all the instances of one code together for comparison, analyze the copresence of real-world events with certain settings or characteristics, clarify patterns and the relationships between patterns and the underlying processes that create them, and – ultimately – generate new knowledge about the social world. At a different level, coding represents a systematic and rigorous attempt to do justice to the research participants and the time and effort invested by all the various players in the research experience. For all of these reasons, coding is increasingly recognized to be an important component of analysis, with inherent and significant elements of critical reflection, well beyond data organization and processing.

The second guideline is a cautionary note: do not assume that you know all the best codes to use right from the start. While some codes will have their roots in the original research design and the intent of the researcher(s), others will emerge and be developed in varied ways. Indeed, codes sometimes spring directly from the data (referred to as in vivo codes), often from phrases used by research participants. As an example, in a project with low-income women during the 1990s retrenchment of welfare benefits, the phrase “struggling constantly alone” was used as a code, which had come from one of the women participating in one of our focus groups. Although it was somewhat lengthy for a code, “struggling constantly alone” succinctly represented many of the women’s sentiments and experiences, and fostered a way of embedding the participants’ words into the analysis. This is an example of a thematic, emerge from the data. For example, in a project on teenagers’ everyday spaces and mobility, my research partner and I knew from related literature that teens’ communication with family and friends (to coordinate get-togethers, find a ride, or just check in) would be important, but we also discovered from our open-ended surveys and focus groups with teens that there is a hierarchy of the modes of communication: phone and e-mail were used for communicating with adults (parents, teachers, coaches), while texting, “chat,” and instant messaging were preferred for communicating with peers (Cope and Lee 2016). Thus, in examining and coding our data, we started to pay attention not merely to the method of communication (which served as one code) but also where it fit in the hierarchy of technological intimacy (a separate code). Being open to the emergence of new or unexpected patterns need not disprove one’s initial sense of what is important but, rather (as in our case), can enhance one’s understanding of social phenomena.

The third guideline follows on the flexibility mentioned above: do not assume that you know all the best codes to use right from the start.

While some codes will have their roots in the original research design and the intent of the researcher(s), others will emerge and be developed in varied ways. Indeed, codes sometimes spring directly from the data (referred to as in vivo codes), often from phrases used by research participants. As an example, in a project with low-income women during the 1990s retrenchment of welfare benefits, the phrase “struggling constantly alone” was used as a code, which had come from one of the women participating in one of our focus groups. Although it was somewhat lengthy for a code, “struggling constantly alone” succinctly represented many of the women’s sentiments and experiences, and fostered a way of embedding the participants’ words into the analysis. This is an example of a thematic,
analytic code; it is fairly abstract and required a certain amount of interpretation and judgment on the part of the researcher. Thus, it was quite different from more categorical codes such as, for example, whether or not someone has children under 18 in their household or their employment status. However, when going back to review all the segments and statements for the women participants that had been coded “struggling constantly alone,” it provided a rich entry point into the data and allowed patterns and relationships to be seen between that analytic code and the more categorical information collected about the women (Cope 2001). Indeed, the women with multiple children under 18 who were single parents and marginally employed were revealed to be “struggling constantly alone” in diverse contextual settings, but with the commonality of having few social or economic resources and many demands on their time and energy.

There are many strategies and methods available to begin coding but, beyond the basic conventions, researchers tend to develop their own practices. Some people code “by hand,” meaning using note cards, colored pencils, or highlighting markers, lists of codes and relationships, whiteboard diagrams, “sort and pile” techniques, and other tangible means of analysis. Other researchers use computer-aided qualitative data analysis software programs (CAQDAS, such as Atlas.ti, HyperRESEARCH, NVivo, etc.), which operate on the same basic principles but allow the digital storage, organization, coding, memoing, and analysis of data. In a comparison of coding manually and electronically, Basit (2003, 143) found that “the choice will be dependent on the size of the project, the funds and time available, and the inclination and expertise of the researcher.” So, a small project involving a limited amount of data and one researcher might lend itself to notecards and colored pencils for the purpose of expediency, while a multiyear project with many types of data and multiple research partners and students might lend itself to the use of a CAQDAS. However, what Basit calls the “inclination” of the researcher matters regardless of the size of the project. For example, Watson and Till’s (2010) chapter on ethnography and participant observation includes an interesting review of why each of them adopts their method of coding (Watson uses Atlas.ti while Till employs a hard-copy strategy), as well as some interesting step-by-step accounts of how they actually perform their coding.

Finally, employing transcription and coding can generate a more rigorous research practice overall by introducing a systematic approach to the data. This does not mean the analysis should be mechanical or unresponsive to the unexpected; rather, a faithful reproduction of audio files to the written record and the subsequent construction and employment of a rigorous coding system generates insights that are both innovative and defensible. Having a systematic method of coding and analyzing data has the following advantages: (i) it allows researchers to incorporate a broader variety of data sources (text, photos, sketches, tweets, etc.) into the analysis; (ii) it fosters the success of projects involving multiple researchers (Weston et al. 2001); and (iii) it works with a wide range of types of projects from small, exploratory pilot studies by one person to large-scale, multiyear, multi-investigator projects.

Of course, coding transcripts is not the only way to make sense of data. Other methods of analysis include critical discourse analysis, narrative analysis, content analysis (a quantitative approach), and hermeneutics. Transcripts are useful in all of these. As an example, consider Dixon’s employment of critical discourse analysis (Dixon 2010), where her aim is identifying the ways society is “enmeshed” in structures, mechanisms, and events, all of which are conveyed...
and learned through discourse (communication, manner of dress, economic exchange, daily routines, laws/regulations, etc.). While the mechanics of interpretation and meaning may differ, in the end, what social researchers are seeking is the rigorous interpretation of data and the production of knowledge.

Connecting to the conceptual

Transcribing and coding oral and textual data was, even as recently as the turn of the millennium, a pioneering method for human geographers. With the expansion and deepening of qualitative investigations, and a turn toward more critical theoretical approaches, these techniques have, perhaps, become more common. Latham argued there was still much to be done, though, in cultural geography:

Over the last couple of decades we have seen something of a revolution in ways we frame what it is that geography is concerned with. We have seen that it is as much about discourses as about “actual” events; that things that seem small and everyday can be as interesting and complex as phenomena that appear much larger and more general; that our own ways of writing the world are bound up with that world’s constitution. But we do not seem to have made much progress in rethinking what this should mean to us as researchers. (Latham 2003, 1993–1994)

However, it might be suggested that in the decade or more since Latham wrote these words that we have indeed made a great deal of progress on this very question, both in terms of who we are as researchers and in terms of how our research is conducted and its place in the world. Feminist and other critical perspectives have inspired more reflexive epistemologies, the legitimation of broader ranges of topics (including the “mundane,” everyday ones), as well as the methods of investigation and researcher positionality as respectful participation with people who are experts in their own lives. With the cultural turn in human geography, the broader focus on everyday life and mundane social/cultural production of space(s), and the increased attention paid to historically marginalized populations, transcription and coding have become important tools for exploring the lives and spaces of those who do not appear in historical or contemporary records in the same forms as the elite. Overall, then, transcribing and coding are useful methods of data management and analysis for projects at varying levels of abstraction and diverse scales: is the research question about how a national leader frames her/his state’s geopolitical position in official speeches or is it about how children navigate school lunch-time politics? In either instance, careful attention to, recording of, and analysis of actions, verbalizations, expressions, and material props can lead to better insights and the practices of transcription and coding play an important role in both projects.

SEE ALSO: Ethnography; Qualitative data, acquisition; Qualitative information: representation

References


Further reading


Transnational environmental governance

Frank Biermann
*Utrecht University, Netherlands*

Environmental governance is the collective steering of societal processes by public and private actors in order to prevent, mitigate, and adapt to environmental change. Transnational environmental governance describes such steering processes that involve two or more countries.

There are a variety of additional or alternative terms used in the literature that emphasize different aspects: *global environmental governance* denotes environmental policymaking that addresses concerns of all countries or that involves all countries. Examples are current negotiations on climate change, biodiversity depletion, overfishing, and the pollution of the oceans. *Intergovernmental environmental governance* describes such processes if they involve only governments. *Earth system governance* is a new paradigm that has emerged in recent years, emphasizing as policy goal the protection of the entire Earth system from transformations brought about by human activities. Climate change is a prime example of Earth system governance, given the global interconnectivity and interdependence of the climate system.

Transnational environmental governance is part of the broader domain of governance studies and governance theory. Importantly, the notion of governance is different from traditional concepts of government. Governance, as opposed to government, is not confined to states and governments as sole actors, but involves myriad public and private nonstate actors at all levels of decision-making, ranging from networks of experts, environmentalists, and multinational corporations to intergovernmental bureaucracies. Governance also often implies notions of self-regulation by societal actors, of private–public cooperation in the solving of societal problems, and of new forms of multilevel policy (Jordan 2008; Adger and Jordan 2009).

As a research field, transnational environmental governance is inherently interdisciplinary. It transcends traditional concepts of environmental policy, given that the current human-made perturbation of the entire Earth system encompasses new problems that range from global changes in biogeochemical systems to the loss of biological diversity. Transnational environmental governance must also address the entire system of multilevel governance, from local to global levels. As a consequence, a complete understanding of transnational environmental governance requires the involvement of the full range of social science disciplines, from the study of local decision-making to international relations.

**Actors of transnational environmental governance**

Transnational environmental governance is marked by a high degree of complexity that involves a large array of institutions and actors. The most important actors are still national governments. National governments are key agents to sign up to international environmental treaties, and to subsequently employ the required means at their disposal to implement these commitments. However, the power and influence of
national governments differ fundamentally. For example, China’s population alone is greater than the combined populations of the 159 least populous countries. While China and India together account for roughly 37% of humankind, the 100 least populous countries represent only 3.9% of the world population and 6.4% of the global gross domestic product. The current system of multilateral decision-making, however, does not sufficiently take into account such vast differences in the size and relevance of countries, leading to recent calls to introduce new types of decision-making in transnational environmental governance, for example, by majority decisions that would weigh the votes of governments by population size or the financial contributions that they represent.

In addition to national governments, transnational environmental governance is characterized by the increased participation of nonstate actors (Dellas, Pattberg, and Betsill 2011). Environmentalist groups have managed to alter the foreign policy of powerful states, to initiate new global rules, or to influence global negotiations. Networks of scientists have assumed a new role in providing complex technical information that is indispensable for policymaking on issues marked by analytic and normative uncertainty. One example is global science assessment programs, notably the Intergovernmental Panel on Climate Change, which involves thousands of scientists from all over the world (Gupta et al. 2012). Business actors have taken a more prominent direct role in transnational environmental governance as well. Many corporations have assumed a direct influence as immediate partners of governments, for example in the framework of the United Nations and of the Global Compact on corporate social responsibility that major corporations have concluded with the UN.

Also important are the many intergovernmental organizations that are active in transnational environmental governance. Many international organizations have developed some activities and programs on environmental governance, including, for example, the Food and Agriculture Organization of the United Nations (FAO), the United Nations Industrial Development Organization (UNIDO), the International Maritime Organization (IMO), and the United Nations Educational, Scientific and Cultural Organization (UNESCO). In addition, most international environmental treaties have set up a secretariat to assist in implementing the agreement, with some of these bodies having evolved into sizable international organizations, such as the secretariats to the United Nations Framework Convention on Climate Change (UNFCCC) and the Convention on Biological Diversity (CBD). To coordinate all these activities, in 1972 governments established the United Nations Environment Programme (UNEP) as part of the general United Nations organization. The effectiveness of this relatively small program, however, has been repeatedly put into doubt in recent years, with many observers arguing that the program should be replaced by a much larger and more powerful independent international organization, such as a world environment organization. The creation of such a new organization is currently supported by numerous countries, including all member states of the European Union and of the African Union. However, other countries, including powerful actors such as the United States, still argue against a new UN agency.

### Transnational institutions

Key instruments of transnational environmental governance are transnational institutions that set standards for participating states and other actors. At present, about 1100 environmental treaties concluded by governments are in force.
Many studies since the 1980s have investigated the emergence, maintenance, and effectiveness of such multilateral agreements, for example the United Nations Framework Convention on Climate Change or the Montreal Protocol on Substances that Deplete the Ozone Layer. Key questions are the conditions that allow for the quick and effective negotiations of transnational institutions, and the primary factors that can explain high levels of implementation, or the failure, of such agreements.

Given the large number of transnational environmental institutions, the interlinkages between different institutions have become a major focus of attention and given rise to concerns about an increasing “fragmentation” of transnational environmental governance. While some studies have argued in favor of a diversity of approaches and a positive evaluation of this growing fragmentation, many others are critical and call for political measures to strengthen the consistency and coherence of the overall system of transnational environmental governance. Important new research has also addressed possible political responses to problems of governance fragmentation and ways forward in “interplay management.”

In addition, transnational environmental governance remains inconceivable without continuous policymaking at national and subnational levels. Transnational standards need to be implemented at the local level, and the global norm-setting requires local decision-making to put into place the frames for global decisions. This results in the coexistence of policymaking at the subnational, national, regional, and global levels in most areas of transnational environmental governance, with the potential of both conflicts and synergies between different levels of regulatory activity. As for the general fragmentation of transnational environmental governance, the emergence of multilevel environmental governance from local to global levels has also given rise to much recent research on optimizing governance effectiveness and on resolving potential problems and conflicts.

In addition to institutions set up by governments, nonstate actors are also becoming part of norm-setting and norm-implementing institutions and mechanisms in transnational environmental governance. Examples are the Forest Stewardship Council (FSC) and the Marine Stewardship Council (MSC), two standard-setting bodies that have been created by major corporations and environmental advocacy groups without the direct involvement of governments. It sometimes seems that traditional intergovernmental policymaking through diplomatic conferences is being replaced by such networks of nonstate actors, which are regarded by some as more efficient and transparent. On the other hand, the distribution of such networks is often linked to the particular interests of private actors who have to respond to their particular constituencies, and serious questions about the legitimacy of private standard-setting remain. For example, policymaking by private organizations such as corporations and environmentalists cannot relate back to democratic elections or other forms of direct representation, and is often perceived as being biased in favor of the interests and concerns of the rich countries in the North.

Public–private partnerships have received much attention since the 2002 Johannesburg World Summit on Sustainable Development (Pattberg et al. 2012). Some observers see transnational public policy networks as an innovative form of governance that addresses the deficits of interstate politics by bringing together key actors of civil society, governments, and business. In this perspective, transnational public policy networks are important new mechanisms to help resolve a variety of current governance
deficits. Others, however, view the new emphasis on such partnerships as problematic, since voluntary arrangements could privilege more powerful actors (in particular industrialized countries and major corporations), would consolidate the privatization of governance and dominant neoliberal modes of globalization, and also lack accountability and legitimacy.

**Equity in transnational environmental governance**

Transnational environmental governance has to operate in a nation-state system that is marked by substantial differences in wealth and power. The richest 20% of humanity account for 76.6% of total private consumption, while the poorest 20% consume merely 1.5%; 842 million people lack sufficient food. The poorest 25% of humanity have no access to electricity. One billion people lack sufficient access to water, and 2.6 billion have no basic sanitation. This global inequality has further increased over the last decades. In 1960, the richest 20% of humanity were 30 times wealthier than the poorest 20%, but this ratio had increased to 74 times by 1997. Transnational environmental governance has to address this context of vast inequalities between people and countries.

One consequence of this global inequality has been that in many institutions of transnational environmental governance, developing countries have insisted on various exemptions. A general rule of transnational environmental governance widely accepted today is the principle of “common but differentiated responsibilities and capabilities,” which was agreed on by governments at the 1992 United Nations Conference on Environment and Development (UNCED) in Rio de Janeiro. This principle is operationalized, for example, in a provision of the Montreal Protocol on Substances that Deplete the Ozone Layer which allows developing countries to delay implementation of the phasing out of such substances by 10 years. Industrialized countries also agreed to compensate the full agreed incremental costs that developing countries would incur in complying with this treaty. In the UN Framework Convention on Climate Change, industrialized countries agreed to take the lead in addressing global warming, and quantified obligations to reduce emissions of greenhouse gases apply at present only to industrialized countries. Since the mid-1990s, a variety of international funding mechanisms have been set up to support poorer countries in their efforts to support transnational environmental governance. One major example is the Global Environment Facility, which is jointly operated by the World Bank, the UN Development Programme, and the UN Environment Programme and which has, since the mid-1990s, allocated US$9.2 billion to developing countries, along with over US$40 billion in cofinancing. In the area of climate policy, developing countries can rely on financial support by an Adaptation Fund, a Special Climate Change Fund, a Least Developed Countries Fund, and, since 2010, a Green Climate Fund.

However, the funding levels to which industrialized countries are committed are not sufficient to cover all the additional costs that developing countries will incur. In addition, the normative principles applicable to global environmental issues are still contested between North and South. Developing countries argue, for example, that industrialized countries should accept a historical responsibility for past greenhouse gas emissions, and should strive to reduce their emissions to levels that equal the much lower emissions of developing countries on a per capita basis, before developing countries have to accept any quantified targets. Industrialized countries, on the other hand, argue that all
countries should implement some environmental policies, and that all efforts by industrialized countries would be pointless if the larger developing countries such as China or India were to continue increasing their emissions through the rapid development of their fossil fuel-based industries and consumption patterns. Overall, these different perceptions of the responsibilities that states have in transnational environmental governance, combined with the vast differences in wealth and welfare, continue to result in substantial deadlocks in negotiations on issues as diverse as climate change, deforestation, and chemicals, with no easy compromises in sight.

The future of transnational environmental governance

In sum, transnational environmental governance is today one of the most important domains of international relations. Its overall effectiveness, however, remains insufficient. A number of key indicators show that planetary systems are moving outside the range that has been known for the last 800,000 years, with potentially disastrous consequences for humankind (Rockström et al. 2009). This calls for renewed efforts by decision-makers in both public and private agencies to strengthen the overall system of transnational environmental governance in order to maintain stable conditions on our planet that allow for a sustainable coevolution of human and natural systems. A number of concrete proposals have been published in recent years, drawing on broader assessments of the social science research in this field (e.g., Biermann et al. 2012). Yet the implementation of reforms is lagging behind. According to recent calculations by the World Bank, all current commitments by governments are likely to lead to a global warming of more than 4 °C by 2100, with a 20% likelihood of more than 4 °C. If current political commitments are not followed through, a warming of 4 °C could occur as early as the 2060s. This is clearly an unacceptable scenario. More political will and efforts in this policy domain are thus urgently needed.

In this sense, effective transnational environmental governance is likely to be transformative. Business as usual will hardly prevent critical transitions in the Earth system. Technological revolutions and efficiency gains might, in theory, provide a solution within the current systems of distribution and consumption, but this is not likely. Instead, to be effective, transnational environmental governance will need to directly address the key concerns of social change. Notably, in a highly divided world, transnational environmental governance poses fundamental questions of equity and allocation within and between nations. Transnational environmental governance also raises important questions about the legitimacy and accountability of public action. The entire system of global governance and world politics is likely to be affected by the changes that are required in transnational environmental governance or, as it is increasingly referred to, earth-system governance. In times of progressing climate change, the accelerating depletion of biological diversity, and growing levels of persistent organic pollutants all over the world, the current system of independent decision-making by 190 nation-states seems increasingly outdated. The success of transnational environmental governance is thus inherently interlinked with the evolution of the overall architecture of world politics. More political engagement and progress toward a “constitutional moment” in international politics (Biermann et al. 2012) seems to be the need of the hour.

SEE ALSO: Anthropocene and planetary boundaries; Climate policy; Environment and
the state; Environmental governance; Governance and development; Multilevel governance

References


Further reading


Transnationalism

Katharyne Mitchell  
University of Washington, USA

The Latin root trans means across, and transnationalism literally refers to that which takes place across national boundaries. The term is most often used to describe the material ties, networks, and sustained relations that operate in an ongoing fashion between actors across national borders; however, it can also function to describe the cross-border symbols and ideas that flow between people and institutions to create multidimensional transnational spaces.

As a general term, “transnationalism” is often linked with globalization. The similarities are many, as both terms can be used to describe processes that are centuries old which also display attributes particular to the post-World War II period. These include the global production and increasingly rapid movement of commodities and information across national borders, and the transmission of cultural practices and ideas in an ever widening gyre. Contemporary globalization and transnationalism are enabled by innovations in transportation such as airplanes and container shipping, as well as advances in computing, telecommunications, and the Internet.

Globalization refers to the widespread interconnections of people, places, and things and, as such, it is a process that can be traced back to the twelfth century, if not earlier. But contemporary globalization manifests these interdependencies in a denser, tighter, and more accelerated framework—what some have termed time–space compression. It is also characterized by a rhetoric of necessity and inevitability, in which the term itself serves to propagate its own functional workings vis-à-vis the accelerating circulation of finance and commodities and the declining regulatory role of the neoliberal state (Sparke 2013).

In this regard transnationalism can be seen as an idea that overlaps significantly with contemporary globalization, but which also contains its own modalities and trajectories. One of the key areas in which transnationalism displays these differences is when it is applied to migration. Scholarly theory linking transnationalism with migration developed in the early 1990s as a new way of thinking about migrant movement, integration, and attachment. Among the early adopters were Basch, Glick Schiller, and Szanton Blanc (1994), who used fresh terminology to describe what they were observing in their research. Rather than taking a dichotomous view of assimilation versus non-acculturation, these scholars noted the ways in which contemporary migrants seemed to be variably integrated in both sending and receiving societies. The migrants forged ties and allegiances across borders with more than one national community and sustained “multi-stranded social relations [linking] together their societies of origin and settlement” (Basch, Glick Schiller, and Szanton Blanc 1994, 6).

As a result of this particular theoretical framing, beginning in the late 1980s, the term began to connote the practices and experiences of nonstate actors and institutions that extend across national borders. At this time it referred primarily to a kind of highly networked, spatially stretched experience where lives were lived simultaneously in more than one place. In contradistinction to the processes associated with economic globalization, however,
these transnational activities often occurred in a highly regulated and sometimes antagonistic relationship to the state. While the neoliberal state’s regulatory role has greatly diminished vis-à-vis the processes of economic globalization, it still retains the authority and legitimacy to police the transnational flows of human actors and social institutions in the name of national security. And today, while most states have facilitated the free flow of commodities and money, they have simultaneously restrained or otherwise controlled the movement of laboring bodies.

These insights help to create new ways of thinking about transnationalism as a plural, multidimensional process, yet one that is still intricately involved with neoliberal governance, especially involving the nation-state. States remain active in policing borders and membership, and retain great social and political power vis-à-vis cross-border migrants and institutions. Nevertheless, boundaries are inevitably leaky, and their diffuse and variable permeability lends the possibility of subversion to the rules and norms of containment in a number of different registers.

The subversive potential of transnationalism as a concept and set of practices has resonated with recent historical investigations of pan-regional events and cross-border relationships such as the Haitian Revolution and the Black Atlantic. Critical scholarship on these transnational moments in space and time has upended universal history and normative assumptions about nation, race, and revolution (Gilroy 1995). The term has also been implemented in more recent critical scholarship examining the expanding alliances forged through feminist, indigenous, and pan-African movements worldwide. It now broadly signifies an opposition to methodological nationalism and a reconfiguration of the national imaginary – opening up the potential to evaluate new political, economic, and cultural forms, alliances, and systems of governance that exist across and beyond the nation-state.

**Transnational migration, institutions, and communities**

In migration studies, where the term “transnationalism” has been employed most frequently and consistently, the concept has propelled greater interest in the multidirectional processes that simultaneously embed people, places, and things in two or more societies. It has enabled research into the geographies of transnational social fields, which produce spaces and topographies in novel configurations. And it has spurred novel ways of thinking about community and the uneven impacts of global remittances.

A remittance is the transmission of money across national borders by a worker to his or her home community. It generally involves money transferred by transnational migrants from the wealthier nations, where they are working, to the less developed countries from where they originated and where they still have strong family ties. Remittances have become an important source of revenue for individuals and also for countries. In 2012 remittances exceeded $500 billion, with over $400 billion of that going to developing countries. This figure has increased rapidly since the mid-2000s.

The importance of remittances cannot be overestimated, as they account for the largest source of foreign currency in numerous countries and often exceed international aid, as well as official sources of capital inflow. While there has historically always been a flow of money from migrants to their extended families and communities back home, the scale of contemporary remittances indicates changes that reflect both the extent of current transnationalism and also the economic shifts occasioned by neoliberal globalization.
Some of the negative effects of neoliberal free trade agreements such as NAFTA (North American Free Trade Agreement) include widening economic gaps between many rich and poor nations and also the socioeconomic dispossessions and dislocations of certain populations, such as Mexican farmers. As a result, numerous individuals have been forced to move from less developed countries to find work in wealthier ones. At the same time, new kinds of technologies have enabled the simultaneous embedding of lives in more than one national locale. This multilocality has led to family strategies involving splitting productive and reproductive labor; work outside the home is conducted abroad by one family member, while those who remain behind raise children and care for the elderly. This strategy produces a strong incentive to invest in both communities, with money sent home in the form of remittances literally providing for the reproductive care of the household, and often for the larger community as well.

Scholars interested in transnational migration have investigated the impact of remittances on individuals, communities, neighborhoods (including the physical forms of buildings and general aesthetics), gender relations, culture, and the rise of new associations and nongovernmental organizations. They have also looked at the implications of remittances on politics, including the reactions of both sending and receiving states to these major financial flows. Often remittances are used to fund businesses as well as to support family, and governments have responded with various efforts to promote the continued inflow of remittance currency. These include establishing social and economic institutions that provide needed services to the migrants themselves, as well as offering matching funds to encourage continued investment.

In addition to state institutions offering aid and the various forms of regulatory control, numerous intermediary organizations have also emerged alongside the development of transnational actors and practices. Levitt (2001) has examined the rise of religious, political, and civic organizations that span national borders, facilitating transnational life and belonging. In her study of the cross-border lives of people inhabiting the social space between Miraflores, a town in the Dominican Republic, and Jamaica Plain, a neighborhood in Boston, Levitt identified both the legal and extra-legal businesses and informal organizations that arose to aid the transnational villagers. These entities helped the migrants work with or around the law, adapt to unfamiliar cultural mores, negotiate the educational system, and deal with racism, among other things. As with many similar studies, the research indicated a transformation of both family and work life in both venues, and the emergence of shifting ideas around gender and race.

While much of the early work on transnationalism emphasized laboring migrants working in the global North and sending remittances home to the global South, significant attention has also been paid to the movements of the wealthy. For the transnational elite, cross-border movements occur in many directions and take many forms, but the relative lack of friction at the border distinguishes their mobility from the movements of the less skilled. The transnational business class, in particular, is highly mobile, often possessing more than one passport and with cultural and educational capital that is transportable from one nation-state to another.

It is critical to study the movements of these parachuting transnationals for a number of reasons. This elite group often has a disproportionate impact on the communities in which they locate their work or residences because of their wealth and savoir faire. They might simultaneously oversee factory production in southern China, research and design in Singapore, and be involved
(by mobile phone or Skype) with their children’s education in Canada. Moreover, their views on politics, culture, or education may carry undue weight in all three settings. In a study of Hong Kong migrants in Vancouver, British Columbia, for example, Mitchell (2004) showed how their presumed tastes in homes influenced the types of house styles and gardens built by developers hoping to make money from this wealthy and highly kinetic population.

In this sense, both the economic practices and the cultural tastes and attributes of transnational entrepreneurs and hi-tech professionals are implicated in the transnational social spaces that they help to create through their global movements. In addition to political and cultural influences, their activities and experiences can shape the ways in which commodity chains are organized, ethnic entrepreneurship functions over time and space, and even how ethnic communities are integrated into both local and global economic networks.

Expanding research on transnational social spaces now goes beyond a primary focus on migrants and their communities to encompass social formations and other nonstate organizations forming sustained ties across national borders. Rather than just looking at transnational relations involving market-based interdependencies, however, this work also emphasizes the emergence of a global civil society (Faist and Ozveren 2004). These dense and lasting cross-border formations are often institutionalized through systems of law, regulatory agreements, or through forms of status or policy. But sometimes they remain informal, institutionalized via performance and repetition rather than through formal structures and conventions. In each case, however, the relationship between these institutions and practices is generally multidimensional, involving differing levels of governance – including local, regional, national, and supranational scales.

Transnational governance, civil society, and culture

Transnational governance is a critical feature of contemporary society, as it involves policy coordination across national borders, primarily by nonstate actors and nongovernmental organizations (NGOs). This form of transnationalism is distinct from supranational governance (such as the International Criminal Court) because it does not subsume national institutions but, rather, operates more informally to address specific problems impacting more than one area of jurisdiction. It has become particularly important to study transnational governance in an era of global climate change and other profound cross-border problems, as new actors and networks have emerged alongside more traditional national, international, and intergovernmental organizations.

Environmental studies often provide a good lens through which to study this type of political transformation as they give a glimpse into the new geographical spaces and conflicts being negotiated with transnational governance networks. One recent example involves the struggles over the Arctic in the context of melting ice and unclear territorial control. As different nations attempt to assert their sovereign rights to this territory and its resources, other transnational actors have also arisen to mediate the disputes and/or call attention to the region and its issues. These competing, cross-border, and multiscaled networks of actors came into direct confrontation on September 19, 2013, when Russian security forces took over the Greenpeace ship _Arctic Sunrise_ by force, and charged the international 30-person crew with piracy. (Greenpeace had been attempting to highlight the critical nature of the Arctic region and the harm being caused by overfishing and the production of dirty fossil fuels.) In this case, the nonstate actor
was perceived as a direct threat to the sovereign might and right of Russia to govern the territory and was harshly punished as a result.

The Greenpeace debacle indicates both the high stakes involved in these types of conflicts, and also the proliferation and complexity of issues involving transnational jurisdiction in the contemporary era. In addition to environmental problems, these include the movements and rights of migrants who may now have differently scaled forms of citizenship depending on their legal rights and sense of belonging to local, national, and/or supranational levels of governance. Since the mid-1990s numerous scholars have explored the meanings and transformations of citizenship in the context of transnational migration and transnational governance, where rights and responsibilities have become increasingly complex. For example, a person may enjoy the right to vote in a local election or may participate in a discussion about the provision of local services but be denied formal national citizenship; in other cases, an individual may be awarded *jus soli* citizenship (citizenship awarded on the basis of birth in a country) in some member nations of the European Union (EU) but not others.

Transnationalism is intimately connected to the rise of these kinds of differential regimes of citizenship, and also to the changing relationship of the state to its most mobile citizens and denizens. How can the state control the finances and maintain the allegiance of its citizens who may be living *across* one or more national borders? Some states have offered the prospect of dual citizenship, while others rely on more punitive sanctions and regulations, particularly around property ownership and inheritance laws. However, in almost all cases, none of these laws or policies is set in stone but they are, rather, characterized by expedient transitions from one form of citizenship formation to another, depending on the larger political and economic context.

Similarly, transnationalism calls into question some of the long-standing assumptions about the relationship of sovereign states to their territorial borders. Given the expansion of activities across national boundaries, are there presumed limits to a state’s right to survey, monitor, and otherwise spy on its own citizens living elsewhere? This issue has become a critical one in the context of recent media leaks indicating massive US-based spying overseas by the National Security Agency (NSA). The lack of definitive territorial borders to the contemporary nation-state’s sovereign power is related to questions surrounding state interventions in the “internal” affairs of other sovereign states. What is the spatial extent of sovereignty? This issue was especially germane subsequent to the United States’ invasion of Iraq in 2003 (see Elden 2009).

One of the responses to the latter question is drawn from the discourse on terror and security that burgeoned after the al-Qaeda attacks in the United States in 2001 and the London bombings in 2005. The discourse centers on the risk and danger to national security of so-called failed states. Failed states are those with chaotic, poorly governed populations and corrupt or inefficient state institutions—a situation, it is presumed, that enables the internal chaos to flow out and across national borders in ways that adversely affect other nation-states. In the discourse of many powerful states, failed states relinquish their right to territorial sovereignty because they pose a danger to other states, serving as a potential channel of risky and unacceptable cross-border flows of drugs, weapons, and/or terrorists.

Nongovernmental organizations have arisen to both aid and resist state actions on cross-border flows and on the more general issues and problems associated with transnationalism. Together they form what has been termed a kind of
transnational civil society. The organizations range from the types of institutionalized global resisters such as Greenpeace, to globally active but less “activist” entities such as Médecins sans Frontières (Doctors without Borders), to transnational feminist networks such as the Sangtin Writers Collective (Swarr and Nagar 2010).

Transnational civil society also embraces less formal relationships and practices. These can be formed on the basis of ties such as diaspora connections, or social justice movements, or both. One example involves the African diaspora and social movements against racial injustice. Marable (2008) argues that multiple, diffuse pan-African circuits and information highways enabled the development of “transnational blackness” as a social site for resistance and change. Historically these circuits included the movements and relationships of black activist-intellectuals in the countries of Africa, the Caribbean, and the Americas in particular, but they are also linked to broader contemporary struggles against racial injustice worldwide.

In Marable’s (2008) analytical framework the manifold paths of cultural, political, and intellectual exchange in the African diaspora are linked together in a transnational network through their convergence on the highway to racial justice. This type of theorizing exemplifies much of the current scholarship on transnationalism. “Transnationalism” is a term that is often employed to challenge norms of containment and fixity, and to highlight movement, resistance, and boundary transgression.

Human geographers have engaged with these and similar ideas, often emphasizing the spatial component involved in transnational flows. Some have focused on the cultural spaces formed through cross-border commodity chains; others have investigated how transnational migrants and transnational spaces are mutually constitutive. A consistent thread of the scholarship is the importance of engaging the complex dynamics of transnationalism as both a process-based and spatially oriented way of thinking, and an ongoing phenomenon related to the circulation of people, goods, ideas, and exchanges across national borders.

**SEE ALSO:** Globalization; Migration: international; Nationalism and geography; Neoliberalism; Political geography

**References**


Transport and development

Jean-Paul Rodrigue
Hofstra University, USA

The importance of transportation for development

Because of its intensive use of infrastructures, the transport sector is an important component of the economy, impacting on the development and welfare of populations. A relationship between the quantity and quality of transport infrastructure and the level of economic development is apparent. When transport systems are efficient, they provide economic and social opportunities and benefits that result in positive multiplier effects such as better accessibility to markets, employment, and additional investments. When transport systems are deficient in terms of capacity or reliability, they can have an economic cost such as reduced or missed opportunities and lower quality of life.

At the aggregate level, efficient transportation reduces costs in many economic sectors, while inefficient transportation increases these costs. In addition, the impacts of transportation are not always intended and can have unforeseen or unintended consequences. For instance, congestion is often an unintended consequence of the provision of free or low-cost transport infrastructure to users. Transport also carries an important social and environmental load, which cannot be neglected. Assessing the economic importance of transportation requires a categorization of the types of impacts it conveys. These involve core (the physical characteristics of transportation), operational, and geographical dimensions.

- **Core.** The most fundamental impacts of transportation relate to the physical capacity to convey passengers and goods and the associated costs of supporting this mobility. This involves the setting of routes enabling new or existing interactions between economic entities.

- **Operational.** This dimension involves improvement in the time performance, notably in terms of reliability, as well as reduced loss or damage, which implies a better utilization level of existing transportation assets benefiting its users as passengers and freight are conveyed more rapidly and with fewer delays.

- **Geographical.** This involves access to a wider market base, where economies of scale in production, distribution, and consumption can be improved, and increases in productivity from access to a larger and more diverse base of inputs (raw materials, parts, energy, or labor) and broader markets for diverse outputs (intermediate and finished goods). Another important geographical impact concerns the influence of transport on the location of activities.

The added value and employment effects of transport services usually extend beyond employment and added value generated by that activity; indirect effects are salient. For instance, transportation companies purchase a part of their inputs (fuel, supplies, maintenance) from local suppliers. The production of these inputs generates additional added value and employment
in the local economy. The suppliers in turn purchase goods and services from other local firms. There are further rounds of local re-spending which generate additional added value and employment. Similarly, households that receive income from employment in transport activities spend some of their income on local goods and services. These purchases result in additional local jobs and added value. Some of the household income from these additional jobs is in turn spent on local goods and services, thereby creating further jobs and income for local households. As a result of these successive rounds of re-spending in the framework of local purchases, the overall impact on the economy exceeds the initial round of output, income, and employment generated by passenger and freight transport activities. Thus, from a general standpoint the socioeconomic impacts of transportation can be direct, indirect, and induced.

- **Direct impacts.** The direct benefits are mostly related to capacity and efficiency improvements that impact users and operators, particularly in terms of time and cost savings. Corporations involved in the provision of transport services earn an income and are paying wages to their employees.

- **Indirect impacts.** The indirect benefits are mostly related to accessibility gains and better economies of scale. While employers and the retail sector (as well as other activities such as institutions) gain better access to labor or customers, the customers of freight transport services (distribution centers, manufacturing, retailers) derive some productivity gains that are the outcome of better transport services. The owners of land and activities also usually derive higher rents from the increasing intensity of passenger and freight traffic taking place in the vicinity. Both passenger and freight traffic also convey additional demands for goods and services (e.g., fuel, maintenance, repairs, insurance). Freight-related activities also benefit from a wider range of suppliers for its inputs and markets for its outputs.

- **Induced impacts.** The induced benefits are mostly related to economic multipliers and increased opportunities. A society benefits from increased mobility since individuals have a wider range of options for their activities and the associated social opportunities (education, leisure). An economy usually becomes more competitive, attracts new and expanded economic activities, and has more complex distribution networks.

Transportation links together the factors of production in a complex web of relationships between producers and consumers. The outcome is commonly a more efficient division of production by an exploitation of geographical comparative advantages, as well as the means to develop economies of scale and scope. The productivity of space, capital, and labor is thus enhanced with the efficiency of distribution and personal mobility. Economic growth is increasingly linked with transport developments, namely infrastructures, but also with managerial expertise, which is crucial for logistics. Thus, although transportation is an infrastructure-intensive activity, hard assets must be supported by an array of soft assets, namely management and information systems. Decisions have to be made about how to use and operate transportation systems in a manner that optimizes benefits and minimizes costs and inconvenience.

**Transportation and economic opportunities**

Transportation developments that have taken place since the beginning of the Industrial
Revolution have been linked to growing economic opportunities. At each stage of human societal development, a particular transport technology has been developed or adapted with an array of impacts (Figure 1). Five major waves of economic development can be suggested where a specific transport technology created new economic, market, and social opportunities (Figure 2).

- **Seaports.** These are linked with the early stages of European expansion from the sixteenth to the eighteenth century, commonly known as the age of exploration. They supported the early development of international trade through colonial empires, but were constrained by limited inland access.

- **Rivers and canals.** The first stage of the Industrial Revolution in the late eighteenth and early nineteenth centuries was linked to the development of canal systems in Western Europe and North America, mainly to transport heavy goods. This permitted the development of rudimentary and constrained inland distribution systems.

- **Railways.** The second stage of Industrial Revolution in the nineteenth century was linked to the development and implementation of rail systems enabling more flexible and high-capacity inland transportation systems. This opened up substantial economic and social opportunities through the extraction of resources, the settlement of regions, and the growing mobility of freight and passengers.

- **Roads.** The twentieth century saw the rapid development of road transportation systems, such as national highway systems, and of automobile manufacturing as a major economic sector. Individual transportation became widely available to mid-income social classes, particularly after World War II.

**Figure 1** Socioeconomic benefits of transportation.

<table>
<thead>
<tr>
<th>Passengers</th>
<th>Freight</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Users</strong></td>
<td><strong>Operators</strong></td>
</tr>
<tr>
<td>Time and cost savings (e.g. commuting).</td>
<td>Time and cost savings (e.g. deliveries).</td>
</tr>
<tr>
<td>Income from transport operations.</td>
<td>Income from transport operations.</td>
</tr>
<tr>
<td><strong>Employers/retail</strong></td>
<td><strong>Customers</strong></td>
</tr>
<tr>
<td>Wider access to labor or customers.</td>
<td>Productivity and added gains.</td>
</tr>
<tr>
<td>Rent income.</td>
<td>Rent income.</td>
</tr>
<tr>
<td>Goods and services to users.</td>
<td>Goods and services to operators.</td>
</tr>
<tr>
<td><strong>Society</strong></td>
<td><strong>Economy</strong></td>
</tr>
<tr>
<td>Improved mobility.</td>
<td>Formation of distribution networks.</td>
</tr>
<tr>
<td>Increased social opportunities.</td>
<td>Attraction of economic activities.</td>
</tr>
</tbody>
</table>

Increased competitiveness.
This was associated with significant economic opportunities to service industrial and commercial markets with reliable door-to-door deliveries. The automobile also permitted new forms of social opportunities, particularly with suburbanization.

- Airways and information technologies. The second half of the twentieth century saw the development of global air and telecommunication networks in conjunction with economic globalization. New organizational and managerial forms became possible, especially in the rapidly developing realm of logistics and supply chain management. Although maritime transportation is the physical lynchpin of globalization, air transportation and IT support the accelerated mobility of passengers, specialized cargoes, and their associated information flows.

Prior to the Industrial Revolution, economic opportunities were limited by the low capacity to move commodities over long distances, and most activities were much localized in scale and scope. The Industrial Revolution unleashed greater economic opportunities, initially with the development of inland canal systems, steamship services, and then railway systems. Passenger and freight transportation expanded as well as production and consumption, while new markets and resources became available. In many instances, the development of a transportation mode built on the opportunities developed by another, such as maritime and canal shipping. In other situations, the growth of a new mode of transportation precipitated the decline of others, such as the collapse of many inland canal networks in the late nineteenth century because of competition from railways.

The development of the mass production system at the beginning of the twentieth century increasingly relied on the commercial opportunities introduced by road transportation, particularly the automobile. Later in the twentieth century, globalization benefited from the synergy of maritime shipping, roadways, railways, air, and telecommunications, all of which supported
integrated transport systems and supply chain management. Economic opportunities became global in scale and scope, particularly because of the capacity to maintain an intricate network of trade and transactions through transport systems. More recently, new opportunities have arisen with the convergence of telecommunications and information technologies, supporting a higher level of management of production, consumption, and distribution, as well as a more efficient mobility of passengers. This process, building upon the advantages conferred by other transportation modes, will account for a significant share of economic opportunities of the first half of the twenty-first century.

No single transport mode has been solely responsible for economic growth. Instead, modes have been linked with the economic functions they support and the geography in which growth was taking place. The first trade routes established a rudimentary system of distribution and transactions that would eventually be expanded by long distance maritime shipping networks and the setting of the first multinational corporations managing these flows. Major flows of international migration that occurred after the eighteenth century were linked with the expansion of international and continental transport systems that radically shaped emerging economies such as in North America and Australia. Transport played a catalytic role in these migrations, transforming the economic and social geography of many nations.

Transportation has been a tool of territorial control and exploitation, particularly during the colonial era, when resource-based transport systems supported the extraction of commodities in the developing world and forwarded them to the industrializing nations of the time. The goal to capture resource and market opportunities was a strong impetus in the setting and structure of transport networks. More recently, port development, particularly of container ports, has been of strategic interest as a tool of integration to the global economy, as the case of China illustrates. There is commonly a direct relationship between foreign trade and container port volumes, so container port development is commonly seen as a tool to capture the opportunities brought by globalization.

Due to demographic pressures and increasing urbanization, developing economies are characterized by a mismatch between limited supply and growing demand for transport infrastructure. While some regions benefit from the development of transport systems, others are often marginalized by a set of conditions in which inadequate transportation plays a role. Transport by itself is not a sufficient condition for development. However, the lack of transport infrastructures can be seen as a constraining factor on development. In developing economies, the lack of transportation infrastructures and regulatory impediments are jointly impacting economic development by conferring higher transport costs, but also by causing delays rendering supply chain management unreliable. A poor transport service level can negatively affect the competitiveness of regions and corporations and thus have a negative impact on regional added value and employment. In 2007, the World Bank published its first ever report to rank nations according to their logistics performance based on the Logistics Performance Index. Investment in transport infrastructures is thus seen as a tool of regional development, particularly in developing countries.

The standard assumption is that transportation investments tend to be more wealth-producing investments, as opposed to wealth-consuming investments such as services. Still, several transportation investments can be wealth consuming if they merely provide convenience, such as parking and sidewalks, or service a market
size well below any possible economic return, with for instance projects labeled “bridges to nowhere.” In such a context, transport investment projects can be counterproductive by draining the resources of an economy instead of creating wealth and additional opportunities. Efficient and sustainable transport markets and systems play a key role in regional development, although the causality between transport and wealth generation is not always clear.

Transport investments also tend to have declining marginal returns. While initial infrastructure investments tend to have a high return since they provide an entirely new range of mobility options, the more the system is developed, the more likely additional investment would result in lower returns. At some point, the marginal returns can be close to zero or even negative, implying a shift of transport investments from wealth producing to wealth consuming. A common fallacy assumes that additional transport investments will have a similar multiplying effect to that enjoyed by the initial investments, which can lead to capital misallocation.

**Types of transportation impacts**

The relationship between transportation and economic development is difficult to formally establish and has been debated for many years. In some circumstances transport investments appear to be a catalyst for economic growth, while in others economic growth puts pressures on existing transport infrastructures and incites additional investments. In a number of regions around the world, transport markets and related transport infrastructure networks are seen as key drivers in the promotion of a more balanced and sustainable development, particularly by improving accessibility and the opportunities of less developed regions or disadvantaged social groups. To begin with, there are different impacts on the transport providers (transport companies) and the transport users. There are several layers of activity that transportation can valorize, from a suitable location that experiences the development of its accessibility through infrastructure investment to a better usage of existing transport assets through more efficient management. This is further nuanced by the nature, scale, and scope of possible impacts.

- Timing of the development varies as the impacts of transportation can precede (lead), occur during (concomitantly), or take place after (lag) economic development. The lag, concomitant, and lead impacts make it difficult to separate the specific contributions of transport to development. Each case appears to be specific to a set of timing circumstances that are difficult to replicate elsewhere.
- Types of impacts vary considerably as the spectrum ranges from the positive to the negative. In some rarer cases transportation impacts can promote economic development while in others they may hinder a region by draining its resources in unproductive transportation projects.

There is no straightforward relationship between transport and economic development as the level of impact and its time sequence can vary based on a location and its socio-economic characteristics. This leads to five potential relationships (Figure 3).

1 Weak relationship. Although transportation supports economic and social activities, no specific causality can be expressed. This is particularly the case for infrastructures that have been implemented a while ago and have become embedded in the regional economy. Their lead role can no longer
be asserted but it does not mean that transportation is not important as it is still a fundamental component supporting interactions within the economy.

2 Positive/lead impacts. This represents the best case scenario where investments and the presence of infrastructures convey economic growth for a region, namely the expansion of production and consumption. This process commonly takes place when new infrastructures are built to access resources or new markets, which then triggers a wave of investments.

3 Lag and positive impacts. The development of the transport system follows economic development. A good example would be fast paced development, as seen in Pacific Asia (notably China), where investments in transport infrastructure do not keep up with the substantial growth of the traffic generated by rising mobility and new globally linked manufacturing functions. Such a situation could eventually impair future growth prospects as the existing infrastructure is no longer able to satisfy the demand.

4 Lead and negative impacts. This commonly involves infrastructure investments made with the expectation of triggering development, but failing to meet those expectations. Many negative outcomes of such strategies have been observed. For instance, infrastructure can be built at great cost while failing to generate significant additional traffic, leaving the community with a substantial debt that cannot be recovered and that will drain regional wealth. On the other hand, transportation could generate traffic, but the new accessibility benefits external economies with improved access to the regional market. Local resources, either physical (commodities) or human (emigration), can also be “drained” away by transport improvements.

5 Lag and negative effect. This represents the worst case scenario where, in addition to the negative consequences of transportation investments on the economy (drain on resources), these investments are occurring after the downward spiral began and may even accelerate the process.

Cycles of economic development provide a revealing conceptual perspective about how transport systems evolve in time and space as they include the timing and the nature of the transport impact on economic development. This perspective underlines that after a phase of introduction and growth, a transport system will eventually reach a phase of maturity through geographical and market saturation. There is also the risk of overinvestment when economic growth is credit driven, which can lead to significant misallocations of capital, including in the transportation sector. The outcome is a surplus capacity in infrastructures and modes, creating deflationary pressures that undermine profitability. In periods of recession that commonly
TRANSPORT AND DEVELOPMENT

follow periods of expansion, transportation activities may experience a setback, namely in terms of lower demand and a scarcity of capital investment. Since transport infrastructures are capital-intensive fixed assets, they are particularly vulnerable to misallocations and malinvestments.

Transport, as a technology, typically follows a path of experimentation, introduction, adoption, diffusion, and, finally, obsolescence, each of which stages has an impact on the rate of economic development. The most significant benefits and productivity gains are realized in the early to mid-diffusion phases, while later phases face diminishing returns. Containerization is a relevant example of such a diffusion behavior as its productivity benefits were mostly derived in the 1990s and 2000s when economic globalization was accelerating. Many technologies go through what can be called a “hype phase,” with unrealistic expectations about their potential and benefits, and many are eventually abandoned as the technology proves ineffective at addressing market or operational requirements, or is simply too expensive for the benefits it conveys. Since transportation is capital intensive, operators tend to be cautious before committing to new technologies and the significant sunk cost they require. In addition, transport modes and infrastructures are depreciating assets that constantly require maintenance and upgrades. At some point, their useful life span is exceeded and the vehicle must be retired or the infrastructure rebuilt. Thus, the amortization of transport investments must consider the life span of the concerned mode or infrastructure.

Transportation as an economic factor

Contemporary trends have underlined that economic development has become less dependent on relations with the environment (resources) and more dependent on relations across space. While resources remain the foundation of economic activities, the commodification of the economy has been linked with higher levels of material flows of all kinds. Concomitantly, resources, capital, and even labor have shown increasing levels of mobility. This is particularly the case for multinational firms that can benefit from transport improvements in two significant markets.

- **Commodity market.** Improvement in the efficiency with which firms have access to raw materials and parts as well as to their respective customers. Thus, transportation expands opportunities to acquire and sell a variety of commodities necessary for industrial and manufacturing systems.
- **Labor market.** Improvement in access to labor and a reduction in access costs, mainly by improved commuting (local scale) or the use of lower-cost labor (global scale).

Transportation provides market accessibility by linking producers and consumers so that transactions can take place. A common error in assessing the importance and impact of transportation on the economy is to focus only on transportation costs, which tend to be relatively low, in the range of 5–10% of the value of a good. Transportation is an economic factor of production of goods and services, implying that it is fundamental in the generation of goods and services, even if it accounts for a small share of the input cost. This means that, irrespective of the cost, an activity cannot take place without the transportation factor. Thus, relatively small changes in transport cost, capacity, and performance can have substantial impacts on costs, locations, and performance of economic activities. An efficient transport system with modern infrastructures favors many economic
changes, most of them positive. The major impacts of transport on economic processes can be categorized as follows:

- **Geographic specialization.** Improvements in transportation and communication favor a process of geographical specialization that increases productivity and spatial interactions. An economic entity tends to produce goods and services with the most appropriate combination of capital, labor, and raw materials. A region will thus tend to specialize in the production of goods and services for which it has the greatest advantages (or the fewest disadvantages) compared to other regions as long as appropriate transport is available for trade. Through geographic specialization supported by efficient transportation, economic productivity is promoted. This process is known in economic theory as comparative advantage.

- **Large-scale production.** An efficient transport system offering cost, time, and reliability advantages enables goods to be transported over longer distances. This facilitates mass production through economies of scale because larger markets can be accessed. The concept of “just-in-time” has further expanded the efficacy of production and distribution with benefits such as lower inventory levels and better responses to shifting market conditions. Thus, the more efficient transportation becomes, the larger the markets that can be serviced and the larger the scale of production. This results in lower unit costs.

- **Increased competition.** When transport is efficient, the potential market for a given product (or service) increases, and so does competition. A wider array of goods and services becomes available to consumers through competition, which tends to reduce costs and promote quality and innovation. Globalization has clearly been associated with a competitive environment that spans the world and enables consumers to have access to a wider range of goods and services.

- **Increased land value.** Land that is adjacent to or serviced by good transport services generally has greater value due to the utility it confers to many activities. Consumers can have access to a wider range of services and retail goods while residents can have better accessibility to employment, services, and social networks, all of which translates to higher land value. In some cases, transportation activities can lower land value, particularly for residential activities. Land located near noise and pollution sources such as airports and highways will thus be impacted by corresponding diminishing land value.

Transport also contributes to economic development through job creation and its derived economic activities. Accordingly, a large number of direct (freighters, managers, shippers) and indirect (insurance, finance, packaging, handling, travel agencies, transit operators) areas of employment are associated with transport. Producers and consumers take economic decisions on products, markets, costs, location, and prices which are themselves based on transport services, their availability, costs, and capacity.

**SEE ALSO:** Development; Economic geography; Factors of production; Transport geography

**Further reading**

TRANSPORT AND DEVELOPMENT


The principles of transport geography

The unique purpose of transportation is to overcome space, which is shaped by a variety of human and physical constraints such as distance, time, administrative divisions, and topography. Jointly, they confer a friction on any movement, commonly known as the friction of distance, or friction of space. These constraints and the friction they create can only be partially circumscribed. The extent to which this is done has a cost that varies greatly according to factors such as the distance involved, the capacity of the modes, and the infrastructures supporting the flows and the nature of what is being transported.

A few core principles of transport geography can be brought forward.

Transportation is the spatial linking of a derived demand

Transport takes place as a result of other economic activities, and links the spatial components of this demand. For instance, commuting is the spatial linking of labor flows resulting from the demand for labor at some locations (e.g., commercial districts) and its supply from others (e.g., residential districts). What is different about transport is that it cannot exist alone, and a movement cannot be stored. An unsold product can remain on the shelf of a store until bought, but an unsold seat on a flight or unused cargo capacity on the same flight remain unsold and cannot be brought back as additional capacity later. In this case, an opportunity has been missed because the amount of transport being offered has exceeded the demand for it. The derived demand of transportation is often very difficult to reconcile with an equivalent supply. This principle underlines that a market economy could not function without the capacity of transportation to link supply and demand.

Distance is a relative concept involving space, time, and effort

How distance is perceived is a function of the amount of effort made to overcome it. These efforts can be assessed in terms of spatial distance, time distance, or other effort measures such as cost or energy spent to overcome distance. Distance can take the form of a simple Euclidean distance considered as a straight line between two locations, but it can also include more complex measures such as transport and logistical distances (Figure 1). Transport distance accounts for the existing structure of the transport network. In a simple form involving only one mode, it is a routing exercise considering the shortest path between two points. In a more complex form, it concerns the set of physical activities related to transportation, such as loading, unloading, and trans-shipment, are considered. The logistical distance encompasses all the tasks required so that a movement between two locations can take place. It thus includes physical flows, but also a set of activities necessary for the management of these flows. To use an air travel example again, a ticket would first need to be purchased,
commonly several weeks in advance. Other common time and cost tasks relate to checking in, security checks, boarding and disembarking, and picking up luggage. Thus, a three-hour flight may in reality be a movement that is planned several weeks in advance, and its full realization can take twice as much time if all the related logistical activities are considered.

Space is a support and a constraint for mobility

Space is the formal context in which mobility takes place, so it can be a support for mobility as it will shape the nature and structure of the transport network. This includes topography, hydrography, and climate. How space acts as a constraint for transport is often relative and paradoxical. For instance, a river acts as a constraint for land transport systems such as roads and railways, but represents the physical support for fluvial transportation (if the river is navigable). At the same time as the atmosphere is the physical support for air transport operations, the weather can be a constraining factor under specific circumstances (e.g., snowstorms, thunderstorms, hurricanes). Since the Earth is a sphere, this fundamental characteristic also impacts transportation, particularly long-distance maritime and air transport.

The relation between space and time can converge or diverge

Every form of transport involves the consumption of a unit of time in exchange for a given amount of space. Over time this process has mostly converged, implying that a greater
amount of space can be reached in the same amount of time (or the same amount of space can be reached in a lesser amount of time). This is jointly the outcome of technological improvements as well as a better capacity and extent of transport infrastructures. Convergence is, however, a spatially and socially uneven process because it impacts the accessibility of locations differently. For instance, transport infrastructure is not laid up uniformly because of economic and political imperatives, and segments of the population will experience a greater improvement in mobility because of their socioeconomic status. The relation between space and time can also diverge when congestion starts to be significant and each additional unit of movement results in additional delays. For instance, traffic in congested urban areas currently moves at the same speed that it did 100 years ago by horse carriage. Air transportation, despite having dramatically contributed to the space–time convergence is also experiencing growing delays. Flight times are getting longer between many destinations, mainly because of takeoff, landing, and gate access delays. Airlines are simply posting longer scheduled flight times to factor in congestion.

Circumnavigation is a good proxy for assessing space–time convergence (Figure 2). Prior to the introduction of the steamship in the mid-nineteenth century, circumnavigating the globe took about one sailing year, a journey greatly delayed by rounding the Cape of Good Hope and the Strait of Magellan. The late nineteenth and early twentieth centuries saw a series of innovations that would greatly improve circumnavigation, notably the construction of the Suez (1869) and Panama (1914) canals, as

![Figure 2](image-url)
TRANSPORT GEOGRAPHY

well as steam propulsion. Circumnavigation was reduced to about 100 days at the beginning of the twentieth century and to 60 days by 1925 with fast liner services. The introduction of the jet plane in the second half of the twentieth century reduced circumnavigation to about 30 hours if two (or three) direct and connecting long-range flights can be booked.

A location can be a central or an intermediate element of mobility

Locations are said to be central when they act as generators (origins) or attractors (destinations) of movements. Locations are labeled as intermediate when movements are passing (transiting) through them on their way to other locations. Ports and airports are often intermediate locations because they act as gateways or hubs within a complex transport network. While a hub is a central location in a transport system, with many inbound and outbound connections of the same mode, a gateway commonly implies a shift from one mode to the other, such as a shift from maritime to land transport. A gateway performs an intermodal function (between modes) while a hub is mostly transmodal (within a mode).

To overcome geography, transportation must consume space

Transportation infrastructures are important consumers of space, which includes the pathways (e.g., roads and rail lines), their rights-of-way, as well as the terminals. The more extensive a transport system, and the higher the level of mobility, the more extensive its consumption of space. For instance, in highly motorized cities, roads, and parking spaces can consume up to 50% of the land. Globalization has been linked to the setting up of massive terminal facilities such as container ports and airports. While the space consumed by road infrastructure is mostly linked to local and regional activities, the space consumed by rail, port, and airport terminals is linked to activities taking place at a higher scale.

Transportation seeks massification but is constrained by atomization

Transport systems are most effective when they develop economies of scale, particularly in terms of the loads (passengers or freight) they carry. Massification involves conveyances with higher capacity and supported by larger terminals. However, more often than not, the first and last segments of a transport sequence require atomization, which implies that loads must be consolidated and deconsolidated. For transport modes, atomization represents the smallest load unit that can be effectively transported. The person is obviously the smallest load unit for passenger transportation, while a parcel or a component are the smallest load units for freight transportation. So, the higher the level of massification the more complex atomization becomes. Individuals and customers tend to prefer the convenience of atomization while carriers favor massification and the economies of scale it confers.

Velocity is a modal, intermodal, and managerial effort

Velocity does not necessarily mean speed, but the time it takes for a passenger or a unit of freight to move across a complete transport sequence. For instance, the speed advantage of air transportation is undermined if a passenger spends several hours between connecting flights. Therefore, the velocity of passengers or freight is a joint consideration of the effectiveness of the respective modes involved, as well as that of the intermodal operations connecting the modes. In addition,
the complexity of transport systems requires an effective management of operations, such as scheduling. All of these jointly contribute to improving the velocity of flows carried by transport systems. Many transportation modes, particularly maritime and rail, have not shown any significant speed improvements in recent decades, with the exception of high-speed rail services.

The principles of transport geography underline that there would be no transportation without geography and there would be no geography without transportation. The fundamental purpose of transportation is geographic in nature because it facilitates movements between different locations. The goal of transportation is thus to transform the geographical attributes of freight, people, or information from an origin to a destination, in the process conferring on them an added value.

The spatial relevance of transportation

Transport plays a role in the structure and organization of space and territories. In the nineteenth century, the purpose of the emerging modern forms of transportation, mainly railways and maritime shipping, was to expand spatial coverage with the creation, expansion, and consolidation of national markets. In the twentieth century, the objective shifted to selecting itineraries, prioritizing transport modes, increasing the capacity of existing networks, and responding to the mobility needs — and at a scale that was increasingly global. In the twenty-first century, transportation must cope with a globally oriented economic system in a timely and cost-effective way, but also with several local problems such as congestion and capacity constraints.

Transport is an indispensable component of the economy and plays a major role in supporting spatial relations between locations. Transport creates valuable links between regions and economic activities, between people and the rest of the world. The relevance of transportation derives from the interactions of its core components, which are the modes, infrastructures, networks, and flows.

- **Modes** represent the conveyances, mostly in the form of vehicles that are used to support the mobility of passengers or freight. Some modes are designed to carry only passengers or only freight, while others carry both.
- **Infrastructures** are the physical support of transport modes, where routes (e.g., rail tracks, canals, or highways) and terminals (e.g., ports or airports) are the most significant components.
- **Networks** are systems of linked locations that are used to represent the functional and spatial organization of transportation. These systems indicate which locations are connected and how they are serviced. Within a network some locations are more accessible (have more connections) than others (which have fewer connections).
- **Flows** are the movements of people, freight, and information over their respective networks. Flows have origins, intermediary locations, and destinations.

These components are fundamental for transportation to take place, but they also underline that geography, in spite of significant technological, social, and economic changes, remain a salient force shaping transportation. Transportation as a multidisciplinary endeavor can be approached through several fields of inquiry, some of which are at the core of transport geography, such as transport demand, nodes, and networks, while others are more peripheral, such as natural resources, political geography,
TRANSPORT GEOGRAPHY

and regional geography. Yet, they all contribute to the understanding of transport activities and their impacts on the economy, society, and the environment. These impacts are historical, social, political, economic, and environmental.

Historically, transport modes have played a role in the rise of civilizations, in the establishment of trade and commercial relations, in the development of societies, and also in national defense. Roman roads and Chinese canals are salient examples. Transportation offers a valuable perspective to understanding the history of a region or nation. Socially, transport modes facilitate access to health care, welfare, and cultural events, thus performing a social service. They shape social interactions by favoring or inhibiting the mobility of people. Transportation thus supports and may even shape social structures. Politically, governments play a critical role in transport as sources of transport investment and as regulators of transport operations. The political role of transportation is undeniable as governments often subsidize the mobility of their populations (highways, public transit, etc.). While most transport demand relates to economic imperatives, many transportation infrastructures have been constructed for political reasons, such as national accessibility or job creation. Transport thus has an impact on nation-building and national unity, but it is also a tool-shaping policy.

Economically, the evolution of transport has been linked to new opportunities and economic development. It is an industry in its own right. The transport sector is also an economic factor in the production of goods and services. It contributes to the value added of economic activities, facilitates economies of scale, and influences land value and the specialization of regions. Transport is both a factor that shapes economic activities, and that is also shaped by them. Environmentally, despite the manifest advantages of transportation, its environmental consequences are also significant. These include negative impacts on air and water quality, noise level, and public health. All decisions relating to transport need to be evaluated by taking into account the corresponding environmental costs and how they can be mitigated. Transport is therefore a dominant factor in contemporary environmental issues, including climate change.

Substantial empirical evidence indicates that the importance of transportation is growing. The years following World War II have seen a considerable growth of transport demand from individual (passengers) as well as for freight mobility. This growth is the result of the larger number of passengers and amount of freight being moved, but also of the longer distances over which they are carried. Recent trends underline an ongoing process of mobility growth, which has led to the multiplication of the number of journeys involving a wide variety of modes that service transport demands. Even if several transportation modes are very expensive to own and operate (ships and planes, for instance), costs per unit transported have dropped significantly over the past few decades. This has made it possible to overcome larger distances to further exploit the comparative advantages of space. Despite the lower costs, the share of transport activities in the economy has remained relatively constant over time. Therefore, the demand for transport infrastructures has expanded for both their quantity and quality. Roads, harbors, airports, telecommunication facilities, and pipelines have expanded considerably to service new areas and add capacity to existing networks.

Consequently, transportation has an important role to play in the conditions that affect global, national, and regional economic entities. Transport is a strategic infrastructure that is so embedded in the socioeconomic life of individuals, institutions, and corporations that it is often invisible to the consumer, but it is
always part of all economic and social functions. This is paradoxical, since the perceived invisibility of transportation derives from its efficiency. If transport is disrupted or ceases to operate, the consequences can be dramatic, such as workers unable to reach their workplace or parts not being delivered to factories, or, worse, inadequate food distribution.

Transportation in geography

Geography seeks to understand the spatial order of things as well as their interactions, particularly when the spatial order is less evident. Transportation is one element of this spatial order which is influenced by geography at the same time as it has an influence on it. For instance, the path followed by a road is influenced by regional economic and physical attributes, but once constructed the same road will shape future regional developments, including urbanization. Transportation is of relevance to geography for two main reasons. First, transport infrastructures, terminals, modes, and networks occupy an important place in space and constitute the basis of a complex spatial system. Second, since geography seeks to explain spatial relationships, transport networks are of specific interest because they are the main physical support of these interactions.

Transport geography, as a discipline, emerged as a branch of economic geography in the second half of the twentieth century. In earlier considerations, particularly in commercial geography (late nineteenth and early twentieth centuries), transportation was an important factor behind the economic representations of the geographic space, namely in terms of the location of economic activities and the monetary costs of distance. These cost considerations, which could be logically articulated, became the foundation of several geographical theories such as central place theory and location theory. The growing mobility of passengers and freight justified the emergence of transport geography as a specialized field of investigation.

In the 1960s transport costs were formalized as key factors in location theories, and transport geography began to rely increasingly on quantitative methods, particularly in network and spatial interactions analysis (Taaffe, Gauthier, and O’Kelly 1996). However, from the 1970s technical, political, and economic changes challenged the centrality of transportation in many geographical and regional development investigations. The strong spatial anchoring effect of high transportation costs receded and decentralization was a dominant paradigm that was observed within cities (suburbanization), but also within regions (Tolley and Turton 1995). The spatial theory foundations of transport geography, particularly the friction of distance, became less relevant, or less evident, in explaining socioeconomic processes. As a result, transportation became underrepresented in economic geography in the 1970s and 1980s, even if the mobility of people and freight and low transport costs were considered as important factors behind the globalization of trade and production (Keeling 2007).

Since the 1990s transport geography has received renewed attention, with new realms of investigation. The issues of mobility, production, and distribution became interrelated in a complex geographical setting where the local, regional, and global became increasingly blurred through the development of new passenger and freight transport systems (Hoyle and Knowles 1998). For instance, suburbanization resulted in an array of challenges related to congestion and automobile dependency. Rapid urbanization in developing economies underlined the challenges of transport infrastructure investment for private as well as collective uses. Globalization
supported the development of complex air and maritime transportation networks, many of which underpin global supply chains and trade relations across long distances. The role of information and communication technologies was also being felt, often as a support or as an alternative to mobility. All of these developments were linked to new and expanded mobilities of passengers, freight, and information (Knowles, Shaw, and Docherty 2008).

This “new transport geography” (Keeling 2007) is based on the premise that transportation is a system supporting complex relationships between its core components: nodes, networks, and demand (Figure 3). Transportation primarily links locations, often characterized as transportation nodes. They serve as access points to a distribution system or as trans-shipment or intermediary locations within a transport network. This function is mainly serviced by transport terminals where flows originate, end, or are trans-shipped from one mode to the other. Transport geography must consider its places of convergence and trans-shipment. Transportation networks consider the spatial structure and organization of transport infrastructures and terminals. Transportation geography must include in its investigation the structures (routes and infrastructures) that support and shape movements. Transportation demand considers the demand for transport services as well as the modes used to support movements. Once this demand is realized, it becomes an interaction which flows through a transport network. Transport geography must evaluate the factors affecting its derived demand function.

Analysis of these concepts within transport geography relies on methodologies often developed by other disciplines such as economics,
mathematics, planning, and demography. For instance, the spatial structure of transportation networks can be analyzed with graph theory, which was initially developed for mathematics. Further, many models developed for the analysis of movements, such as the gravity model, were borrowed from the physical sciences. Multi-disciplinarity is consequently an important attribute of transport geography, as in geography in general.

New, expanded, or receding spatial relations are produced by transport systems. A better understanding of these spatial relations is essential in enabling private and public actors involved to mitigate key transport constraints, such as capacity, transferability, reliability, and the general integration of transport systems through intermodalism.

As all activities are located somewhere, each location has its own characteristics which potentially generate a supply and/or a demand for resources, products, services, or labor. A location will determine the nature, the origin, the destination, the distance, and even the possibility of a movement to be realized. For instance, in addition to consuming resources, a city provides employment in various sectors of activity. Locations require the exchange goods, people, or information, which creates a complementarity. This implies that some locations have a surplus while others have a deficit. The only way to reach an equilibrium is by movements between locations that have surpluses and locations that have demands. For instance, there is a complementarity between a store (which has a surplus of goods) and its customers (which demand goods). Movements generated by complementarity occur at different scales, depending the nature of the activity. Scale illustrates how transportation systems are established over local, regional, and global geographies. For instance, home-to-work journeys generally take place at a local or regional scale, while the distribution network of a multinational corporation is most likely to cover several regions of the world.

Consequently, transport systems consume land and support the relationships between locations at a scale that is increasingly global. Transport geography provides a multidisciplinary perspective to understanding the complexity of transportation and how spaces support and hinder mobility.

SEE ALSO: Air transport; GIS for transportation; Graph theory; Maritime transport; Transport and development; Transport networks; Transportation history; Transportation and land use; Urban transit

References

Transport networks

César Ducruet
*Centre National de la Recherche Scientifique (CNRS), France*

Transport networks may be defined as sets of connected nodes and links allowing for the circulation of individuals and/or commodities. Traditionally, transport networks correspond to physical infrastructure such as roads, railways, canals and rivers, streets, and subway lines, while the concept also includes airline and maritime networks in which nodes are (sea)ports and airports, as well as telecommunications networks transmitting information and knowledge flows, such as telephone lines, the Internet, and similar technologies (e.g., telegraph, mobile phones). In fact, the definition of transport networks can be extended to anything that supports circulation, as natural scientists analyze, for instance, termite mounds and fungal networks just as geographers and engineers study road networks. Tero et al. (2010) have particularly demonstrated that biological networks such as those built by slime molds have “comparable efficiency, fault tolerance, and cost to those of real-world infrastructure networks – in this case, the Tokyo rail system.” While the nature of nodes may vary from one network to the other, those are often crossings, stations, terminals, cities, or countries, and so on, while links can be of a totally different nature, from physically grounded tracks to more ephemeral flows. In all cases, transport networks belong to the category of “spatial networks” that comprises many other networks such as so-called technical networks (e.g., power grids) and any network being physically (and often geographically) grounded in space (see Barthélémy 2010). Like any other network, and according to the language used by graph theory, a branch of mathematics describing the topological properties of networks, a transport network may be planar or nonplanar (i.e., with or without crossings between links or “edges”), directed or undirected, and weighted (i.e., where links carry a certain quantity, e.g., of traffic, distance, or time) or not. A multimodal transport network may be understood as a “multigraph” where two given nodes are connected via at least two links of different nature. However, the concept of transport networks does not entirely belong to graph theory; it includes other approaches, such as transport history, spatial analysis, economics, or engineering, where nodes, links, and flows are considered by means of other methods, such as discourses, policies, actors, governance, planning, and dedicated geographical information systems for transport (GIS-T).

Traditionally in geography, transport networks have been the focus of scholars willing to apply quantitative methods to geographical research. Transport networks and flows received foremost attention in the 1960s by US geographers. They particularly developed the idea of ideal-typical sequences of transport network development and applied several graph-theoretical methods to road and railway networks (Taaffe and Gauthier 1973). The most renowned model describes six successive phases inspired by the evolution of transport networks in West Africa (see Rodrigue 2014 for a useful online illustration). In a first phase (scattered ports), a set of ports along a coastline is characterized by homogeneity in...
traffic volumes and connection levels based on local resources. The second phase (penetration lines and port concentration) witnesses an expansion of the hinterland benefiting only a minority of such ports through new transport lines such as rail or canal. The subsequent phases reinforce a path-dependent process by which initially favored ports continue to grow in proportion to the importance of their inland accessibility as the corridor emerges gradually through economies of scale. Although this model has been updated to take into account overseas linkages among ports and more recent trends of deconcentration, there have been many critiques of it, given the linear dimension of the successive phases which ultimately always lead to concentration in the core and decline in the periphery. Similar criticism emerged from (trans)port specialists about early urban models, such as the central place theory, due to its inability to incorporate deviations caused by specific transport configurations (sea-land, intermodal), as in the works of British geographer James Bird. In the early days, however, limited computational power largely explained the reliance upon small, planar networks easily extracted from city and country maps, while detailed – and often more voluminous – data about transport flows remained hardly accessible.

Two central and recurrent themes are the topology of transport networks and the accessibility and centrality of their nodes. The topology may refer to the differences between the actual layout of a given transport network and its simplified representation as a graph, that is, straight lines (or edges) between nodes (or vertices). The graph itself is often understood as a matrix, sometimes called an origin-destination (O-D) matrix, where links either exist or do not between nodes ordered in lines and columns (Figure 1). Several measures express a variety of topological features based on such a graph or matrix (Kansky 1963), such as the size of the network (number of nodes and links), its density (number of links per node), completeness (proportion of existing links in the maximum possible number of links), cliquishness (number of cycles), transitivity (as measured by the average proportion of connected adjacent neighbors in the maximum possible number of connected neighbors), and so on. Using the graph or matrix to calculate network indices makes it possible to apply any graph-theoretical measure, such as those found in social network analysis (SNA) and complex systems, despite fundamental differences in the nature of the studied networks. Some measures may not be interpreted in the same way, depending on whether people, firms, terminals, or neurons are concerned. Nevertheless, transport networks are common objects of an emerging “science of networks,” notably since physics and computer science started to analyze such networks with renewed tools and concepts in the late 1990s (Newman 2010). Other aspects of transport networks have thus been explored, such as their scale-free and/or small-world dimensions, the probability for nodes to connect with other nodes of comparable size (assortativity), the average shortest path length, the “rich-club” coefficient, and so on. While the degree distribution follows a power-law function in scale-free networks, small-world networks are characterized by the existence of dense communities (or subgroups of nodes). The rich-club dimension refers to networks in which nodes with many links connect primarily with each other. Great care must be taken, however, in comparing different networks via such methods, because network indices are often biased by network size. Such measures better apply to the evolution of a single network over time, or to the comparison between actual and optimal configurations. For instance, a given (planar) transport network, if weighted by cost or another metric, is best understood when its two
possible extreme situations are confronted, that is, when all possible links are built (cf. greedy triangulation) or when only the optimal links are retained (cf. minimum cost spanning tree). Novel evidence could have been proposed based on such methods when comparing cities of the world through the configuration of their streets, in terms of planning models, urban architecture, and overall circulation efficiency (see Barthélémy and Flammini 2009).

At the node level, measures are similar and may look at the immediate environment of nodes by the number of adjacent neighbors (degree centrality), which correspond to the number of connected airports, for instance. In a road or railway network, cities or stations can be compared according to their distance from (or eccentricity to) other nodes, their occurrence on shortest paths, and so on. In a weighted network, it is also interesting to look at the proportion of traffic depending on the biggest link; in the case of maritime transport, it revealed the extent to which certain ports heavily depended on a close competitor despite their high traffic volume (e.g., Shenzhen vs. Hong Kong). Questions might concern whether larger nodes are more central in the transport network and what the respective roles are of local and global centralities. Certain findings could have pointed at interesting functions of some nodes: in an air

Figure 1  Graph representations of networks and matrices.
TRANSPORT NETWORKS

Transport network some nodes, despite having few links, can still be highly accessible due to their bridging role between large regions, while larger and more connected nodes often connect across farther topological and kilometric distances. In turn, the existence of such “regions” in a transport network is revealed by means of various techniques of clustering nodes based on certain attributes of connectivity. Other areas of inquiry might include whether best connected airports or seaports are geographically close, or how commuting flows can help in delineating functional urban areas and/or urban regions. Clustering algorithms applied to transport networks, such as modularity, have, among other results, verified the underlying regional patterns of global airline and maritime flows. Single linkage analysis, for instance, was applied to telecommunication flows as early as the 1960s by US geographers in order to delineate so-called nodal regions in a given urban system.

Complementary methods of network coupling based on the concept of multigraphs have explored how different transport networks share similarities but also vulnerabilities. This is where the network analysis of transport meets more practical problems of disruptions, rerouting, and congestion, but also optimization. Questions include where a new node or link should be placed in order to make the transport network(s) more robust to all kinds of failures and targeted attacks, and how two or more transport networks should be connected so as to limit such shocks. Such questions are particularly relevant in the case of networks being largely dependent upon a few large, critical nodes (e.g., Suez and Panama Canals for sea transport). Other aspects to be explored via coupled networks include the study of potential intermodal shifts, the co-evolution of transport networks, and the multimodal accessibility of cities, among others. Geographers have been particularly keen to analyze transport network dynamics in various contexts, such as the impact of deregulation on airline networks, of railway development on urban accessibility, and so on. Notably, Bogart (2009) is a good example of a study combining a dynamical and a multimodal approach when looking at the evolution of canals, roads, and ports during the English Industrial Revolution. This work notably demonstrated the positive effect of road development on canal development, but a less significant role of ports, probably due to their earlier existence. It also suggested that investing in a canal would be less profitable when nearby road improvements were initiated due to higher uncertainties about potential profits from canal tolls.

Overall, graph-theoretical approaches to transport networks attracted a lot of criticism from more applied research due to their inability to include more complex aspects of transport, such as cost, time, strategies, and scenarios, but also graphical aspects. Indeed, a recurrent problem of transport networks is their visualization and cartography. Recent advances in graph visualization algorithms have provided certain solutions to represent, for instance, airline networks as planar networks using so-called edge-bundling tools, in order to limit the fuzziness of mappings caused by the intermingling of numerous, space-constrained links and nodes in a large transport network (Lambert, Bourqui, and Auber 2013). In turn, exploring the multifaceted dimensions of transport networks, such as their territorial embedding, had mobilized methodologies other than graph theory alone. One good example is the raster-based analysis of global urban accessibility taking into account no less than 10 different layers in a GIS understood as “friction surface components” (Nelson 2008). Despite its fundamental graph-theoretical essence, a work on the road accessibility of European seaports integrated several qualitative approaches.
aspects such as national regulations in terms of speed limits and resting times for truckers, the presence and cost of highway tolls, the types and capacities of main road arteries, and the potential delays of loading/unloading at intermodal junctions (Chapelon 2006). In the case of urban streets, more qualitative aspects were introduced by “space syntax,” which proposed to define links based on the continuity of avenues and boulevards from a cognitive point of view, rather than alignments of arbitrarily separated segments (Hillier and Hanson 1984). Such an approach could, by transforming lines into nodes and nodes into lines, reveal a hidden dimension of the network, the hierarchy of streets, which was not visible via the more traditional, planar perspective. Many measures did not provide extraordinary new results, since in general the traffic volume of transport nodes, which is known from classic sources, is highly equivalent to any centrality measure: large seaports or airports have many links and vice versa. Rare studies, however, have truly investigated the spatial determinants of the centrality of transport nodes. These include statistical evidence about the role of urban and regional socioeconomic characteristics, mostly in the realm of airline networks, where cities are nodes and aircraft trips are links. Main results point to noticeable interdependencies between centrality and the wellbeing of localities, for instance in Europe based on airline data and urban indicators, often concluding a circular causation between urban and transport development. The spatial dimension was also included, perhaps more by physicists than by geographers, so as to demonstrate rather trivial, albeit fundamental, properties of transport networks whereby larger nodes connect more distant nodes on average. Such works shared some affinities with the regional science perspective on transport hubs where cost and time are central parameters.

Given such limitations, and despite the improvements brought by GIS-T and natural sciences, geography and other social sciences have increasingly defined transport networks as elements of wider structures. The shift from quantitative/structural to more behavioral approaches is one main explanation (Waters 2006), but another is the fact that network analysis mostly considers transport networks as simplified and self-sufficient entities developed in abstract spaces. New concepts that emerged in the late 1990s, such as global production networks and value and commodity chains, have in common a disregard of pure physical flows, and are more concerned by the “flesh” than the “bones” when dealing with actors, mobility, governance, and territorial issues at various scales. A transport network can thus be approached via more complex and related configurations, planned or not planned, such as corridors and gateways, although the two latter concepts are often subcomponents of the transport network itself.

Another way to place transport networks into a wider analytical perspective has been proposed by urban geographers, in implicit or explicit ways. Early urban models, such as central place theory and up to the point of the new economic geography (NEG), approached transport networks in a rather abstract fashion while conferring on them vital importance in the emergence of cities and hubs within and between regions. Nevertheless, most NEG studies keep considering cities and regions as equally spaced, thus somewhat excluding the friction exerted by transport networks on mobility. Studies of urban systems and systems of cities were more keen on analyzing urban development through the more concrete configuration of transport networks, such as railway networks. From such a perspective, the transport network is used as a proxy for – and a material illustration of – wider
interurban interactions and dynamics in terms of spatial expansion, functional diversification, hierarchical transformations, and innovation cycles. Simulation methods are used to extract meaningful properties about both transport and urban dynamics and to compare estimates with real population data across many decades or centuries. Yet, there are very rare attempts to integrate urban or regional attributes into transport network analysis and/or modeling, such as that by Schintler et al. (2007) analyzing the vulnerability of road and railway networks in Florida through a mixture of various methods and approaches. More efforts are thus needed to push this integration further in order to better meet current challenges in transport and society.

SEE ALSO: GIS for transportation; Graph theory; Maritime transport; Network analysis; New economic geography; Rail transport: freight; Rail transport: passenger; Road transport; Routing and navigation; Transport geography; Visualization

References


Further reading


Transport and mobility are essential to human existence. As modes of transport have become more numerous and technologically sophisticated, the complexity of transport provision has increased dramatically in recent years. While contemporary transport systems convey numerous benefits, they also come with considerable costs, including economic, environmental, and social externalities. Making informed decisions about the provision of transport is guided by the development of policy from both the public and private sectors.

The policy context

In general, policy can be defined as a definite course or method of action or inaction selected (by a government, institution, group, or individual) from among alternatives to guide and usually determine present and future decisions. Policy, especially in the governmental sphere, is often expressed in legislation or regulation, or in the form of policy declarations (white papers, executive orders, etc.). Policy is often, but not always, closely linked to planning, which involves efforts to define a problem and to develop potential solutions in a systematic fashion to improve the quality of decision-making. Policy sets the goals and objectives to address an issue or problem and then specifies the rules and principles by which decisions are to be made, while planning is more concerned with the assessment of policy options and the measures needed to reach particular goals and to resolve specific problems.

Accordingly, in the transport realm, transport policy can be defined as a set of constructs and propositions established to achieve objectives related to the functioning and performance of a transport system (Rodrigue 2013). The provision of transport can be the domain of either public agencies or private companies (or both), depending on the institutional frameworks of the countries involved and the historical development of particular transport modes. Countries such as the United States have a history of private sector transport provision, especially in the railroad, maritime, automobile, and airline industries, while state control of transport has been more typical of countries such as France, Japan, China, and Russia. In almost all countries, though, the reality is a mix of public and private ownership and operation. While private companies may provide many transport services, they often rely on public provision of physical infrastructure, traffic control, regulation, and safety. So, while transport policy receives inputs and generates outputs from both the public and private sectors, it is typically governments (at all levels) that establish and coordinate transport policies for their jurisdictions.

As transport systems have increased in scale, speed, and technological complexity, the challenge of creating and maintaining effective transport policies has grown over time. Governments and private companies have limited resources to invest in transport provision, and,
given the considerable investment in infrastructure and operations needed to deliver modern mobility, the stakes for establishing strategic transport policies are high. The time and effort required to create effective transport policies can be considerable, but the cost of inadequate or obsolete policies would be even greater. Since the cost of policy shortcomings usually manifest over time while the costs of creating and improving policy are immediate, there is a perverse incentive for organizations to underinvest in policy development when resources are scarce.

The transport policy agenda

Transport is vital in supporting societal and economic development by providing access to needed goods and services. Because of the significant burden on society that can arise from monopolistic tendencies across several transport modes, which encourage mobility providers to raise prices and lower productivity, transport policy needs to ensure that the general public interest is protected by promoting access to and the availability and affordability of transport services. Transport also has a significant national security dimension since it supports the movement of people and goods into and out of nationally controlled territories and spaces.

The increasing speed and volume of contemporary transport operations has driven an increase in the number of transport fatalities and accidents over time; thus, the promotion of safety in transport operations is a critical concern that requires policy attention. The adoption of new technologies in areas such as autonomous vehicles and systems invites new policies. Negative environmental externalities – burdens that are not borne directly by the user – from transport operations include local air pollution, global greenhouse gas emissions, water pollution, noise, land take, and wildlife habitat encroachment. Governments need to coordinate transport with environmental and energy policies to mitigate these and other negative effects. Societal equity concerns also require thoughtful transport policies to ensure that minimal levels of access are available to all members of society, regardless of socioeconomic status or class, and that the burdens of negative impacts are not concentrated on marginalized and vulnerable populations within a society (e.g., low-income and minority communities adjacent to highways, airports, seaports, or railroad yards).

Deregulation, liberalization, and privatization

Beginning in the late 1970s, public regulatory and administrative structures behind transportation policy, and in some cases the direct management and operation of transportation services that had been created over the past 100 years, began to be transformed to meet a new hierarchy of policy goals. Economic assessments of the transport sector showed mounting evidence of inefficiency and stagnating productivity that were attributed to the effects of regulation. Claims were made that, over time, special interests, usually connected to producers but also including labor organizations, had succeeded in “regulatory capture” to displace the public interest with their own concentrated benefits.

The prescription offered by economists to make transportation perform better was less government intervention, and conservative political leaders from President Ronald Reagan in the United States to Prime Minister Margaret Thatcher in the United Kingdom became enthusiastic supporters of deregulation and privatization in transportation and other policy
domains. Policy redesign to prioritize market forces over government planning was accelerated by a broader liberalization of international trade and finance in the late twentieth century that accelerated the demand for moving people and goods across the globe. The end of the Cold War also served to decrease concerns about national security, which had often been used to justify greater government oversight of transportation ownership and operation.

From the deregulation of air transport, rail, and trucking in the United States and the privatization of national airlines across Europe and Asia, to the sale of publicly owned railway systems that crisscrossed Australia, Canada, Mexico, the United Kingdom, and other countries, the role of government in the business of transportation has declined since the 1980s. And, with reduced economic regulation, the productivity of railroads and airlines has increased considerably during these years, along with an increase in the number of mergers, bankruptcies, and strategic alliances. Profitability has been more erratic as many firms have struggled to stay afloat while trying to maximize revenues and reduce costs. With a shrinking level of engagement and expertise in government, transportation policymaking has become more reactive, responding to crises in transportation after extreme events such as the terrorist attacks of September 11, 2001 (“9/11”), which led to an increase in government intervention in the provision of transportation security.

The implementation of deregulation and privatization policies has resulted thus far in several general outcomes, though there is some variation across industries and countries. Overall, these policies have encouraged more companies to enter the transport industry, but there has also been greater turnover as most of the new entrants have been unable to stay in business. After one or more periods of new firm entry in deregulated markets, there has been a tendency toward increased consolidation through mergers and alliances. Increased competition from new service opportunities for existing operators, as well as new entrants, has resulted in generally lower prices for shippers and consumers, at least until consolidation creates more concentrated markets that provide more pricing power for carriers.

The spatial outcomes of liberalization policies have seen increased geographic concentration in the large cities and key hubs of transport networks, and a reduction in the level and quality of service in many peripheral communities. The most notable spatial manifestation of airline deregulation has been the rise of hub-and-spoke networks, which provide airlines with the ability to create many one-stop connections through their principal hubs. Most large full-service carriers, including many former national flag carriers, have adopted hub-and-spoke networks which have allowed them to concentrate passenger flows at key hubs for both domestic and international feed traffic. Many of these full-service carriers, such as American Airlines, Delta Air Lines, United Airlines, Lufthansa, and Air France–KLM, are now members of global alliances (such as Star Alliance, Oneworld, and SkyTeam) with other airlines from countries throughout the world. In contrast, low-cost air carriers have focused on shorter-haul point-to-point networks, favoring direct services between larger cities and tourist destinations. In the maritime and railroad industries, liberalization policies and technological improvements have allowed carriers to tap inherent economies of scale, resulting in larger ships, double-stack trains, and more efficient intermodal operations. Increased geographic concentration at large port facilities, which can accommodate the much larger cargo ships, typify the contemporary movement of freight. Deregulation has facilitated the entry of
numerous trucking companies into the industry, which handle the movement of freight from the major consolidation terminals to the customer’s final destination.

Privatization initiatives have been extended to infrastructure provision, including airports, seaports, highways, and transit systems. With declining levels of government funding for transport infrastructure, private firms have become more heavily involved in infrastructure design, construction, financing, operation, and maintenance in the form of public–private partnerships. User payments, facilitated by electronic systems of fee collection, have become important sources of funding to supplement gasoline and other more traditional tax revenues.

The changes in the roles and responsibilities of public sector participation in the transportation sector have created a growing asymmetry of information and capacity for policy development between governments and firms. Relative to the private sector, public agencies and decision-makers know less about transportation system performance now than they did in the twentieth century. For example, financial data on carrier costs, revenues, and profits which used to be disclosed to regulatory bodies or reported by state-owned firms has become less available, since private operators and infrastructure owners shield such data under the doctrine of commercial confidentiality. Third-party logistics providers and supply chain managers also sign confidential contracts with carriers and terminal operators, further limiting public knowledge about the movement of goods.

This compartmentalization of knowledge about overall transportation system performance affects private firms as well as government actors. Private firms have almost as few incentives to share commercial information with one another as they do with government. This knowledge gap can inhibit long-range planning efforts in both the public and the private sectors.

Sustainability

Developing sustainable transportation policy poses a contemporary challenge that reveals how the constraints of piecemeal information can inhibit shared understanding of potential solutions to serious risks. The risks of extreme weather events, and the role of transportation in contributing to anthropogenic climate change, are becoming increasingly evident, though no less politically contested. What cannot be denied is that transportation lags behind other major sources of greenhouse gas emissions in reducing these atmospheric outputs. Manufacturing, housing, and power generation have found more ways to reduce the carbon intensity of their activities than the transportation sector has. If these trends continue, transportation will eventually be impacted by climate risks, triggering some form of crisis and response.

The failure to address climate change and the concurrent vulnerability of carbon-based energy supplies, particularly oil which powers over 90% of global mobility, could lead to significant disruption in the growing demand for mobility that was associated with previous shifts toward market-led policy. Success in reconciling mobility with sustainability could inaugurate new policy instruments and organizational structures that evolve to shift the policy paradigm beyond its free-market priority to a greater role for government to contribute to future policy development.

In the realm of urban transportation, policies concerned with sustainability have favored more investment in public transit, bicycling, walking, telecommuting, and reduced travel overall, while being less supportive of automobile-oriented highway systems. Combined with land-use policies that seek to limit urban sprawl and
encourage higher-density mixed-use activity centers, sustainable transport policies are intended to reduce vehicle-kilometers traveled, petroleum consumption, and atmospheric emissions that contribute to both global warming and local air quality concerns.

Safety and security

Increasing transport activity, especially road transport, has had a dramatic effect on transport safety worldwide. While transport-related fatalities and accidents have been decreasing over the last several decades across more developed countries, they have been increasing for many developing countries, mainly as a result of increased levels of motorization and inadequate attention to safety concerns. Transport policies that have been enacted to improve safety include the establishment and enforcement of speed limits, improved transport equipment standards such as seat belts and other accident controls, improved transport infrastructure design, enforcement of laws that prohibit the operation of transport vehicles while under the influence of alcohol and other drugs, and limiting the hours of work for commercial transport operators.

The events of 9/11 and subsequent incidents have refocused attention on the need for greater transportation security. In the United States, the Transportation Security Administration was created after 9/11 and policies such as the screening of passengers and increased inspection of freight were enacted to reduce risks from potential terrorist attacks using transport systems. Coordination between domestic transport security agencies as well as international organizations such as the International Civil Aviation Organization (ICAO) and the International Maritime Organization (IMO) have been a prominent feature of transport policies throughout the world.

Conclusion

While much of the transport policy agenda in the 1980s and 1990s focused on economic opportunities such as deregulation and privatization, increased attention on environmental sustainability and security risks have dominated the twenty-first century agenda thus far. It is likely that these concerns, as well as those of new technologies, will continue to be major issues in transport policy. Petroleum consumption, greenhouse gas emissions, and climate change will continue to be major challenges especially for the transport sector. Government policy must play an important role in coordinating and incentivizing private firms and public agencies to develop transport vehicles and systems that use nonpolluting renewable fuels in a safe and efficient manner. Security concerns are also unlikely to diminish over time, especially the continuing threat of international terrorism targeting transport operations, as well as disputes regarding territorial control and access to air space, land, and water.

SEE ALSO: Transport and development; Transport geography; Transport networks; Transport technology; Transportation history; Transportation and land use; Transportation planning

References

Further reading


Developing the technology of mobility: three dimensions of accomplishment

Moving beyond one’s immediate surroundings to access new territories, resources, and markets has been a primeval source of inspiration for humankind’s efforts to create and improve transport technology. While the means used to enhance mobility have evolved considerably over millennia, the goal of moving people and goods further, faster, and more easily continues to motivate advances in transport technology. These development efforts can be more clearly understood through examining three dimensions of technology that support the movement of people and goods: (i) propulsion, (ii) vehicles, and (iii) infrastructure.

Achieving propulsion advances

Without propulsion, mobility options are quite limited. Thus, humans have regularly sought to harness bigger and better means of propulsion for moving themselves and their goods. Propulsion technology encompasses the range of mechanisms used to convert energy into motion. The original and oldest means of human locomotion comes from moving about under our own power.

Since human migrations out of the African savannah that began sometime between 100,000 and 300,000 years ago, our ancestors have reached all the continents except Antarctica on foot.

Self-powered propulsion was supplemented by animal power once dogs were domesticated 12,000 to 14,000 years ago, with horses and oxen adding considerable power to animal propulsion options around 4000 years ago. Evidence of sails being used to harness the wind and propel boats dates back 4000 to 6000 years (Block 2003). Thus, long before the first motor was primed to power mobility, humans have been eagerly seeking alternative sources of mechanical energy to substitute for their own efforts at locomotion.

Propulsion technology reached a new stage of obtaining inanimate power in the early nineteenth century when the steam engine successfully converted the heat from combustion, another technology discovered by ancient humans, into mechanical energy. These external combustion engines, in which boilers piped steam into large cylinders, propelled a considerable increase in mobility by rail and water during the nineteenth century, particularly in Europe and North America. But their size and weight proved challenging, necessitating the use of tracks to guide land-based vehicles powered by steam. By the late nineteenth century, the internal combustion engine (ICE), which was much lighter and more efficient in its conversion of fuel into mechanical energy, enabled vehicles to operate on the open road, ushering in the era of automobility. Electric motors were also introduced around the same time as ICEs. To date, they have played a more limited role in powering transport, filling specialized niches propelling ships, submarines, and railroad trains.
Recently, concerns about energy efficiency and environmental impacts have led to the adoption of electric motors in automobiles, as both hybrids and battery electric vehicles.

The Wright brothers were first to apply the ICE to rotating a propeller that could lift an aircraft off the ground in 1903. And while piston engine propeller aircraft continue to operate today, jet engines introduced during World War II rapidly became the preferred propulsion source for both commercial and military aviation. Also appearing during World War II were rocket motors that could lift missiles into the upper atmosphere (Anderson 1985). During the “space race” of the 1950s and 1960s, engineers in the United States and Soviet Union developed rockets capable of powering orbital and interplanetary space travel. More recent design efforts have sought to reduce the cost of rocket propulsion to enable commercially viable travel into space.

Improvements in carriage technology proceeded incrementally over centuries. In fifteenth-century Europe, suspension systems of leather straps and iron chains offered a major advance in vehicle stability and capacity. In the eighteenth century, steering mechanisms were added to vehicles’ design. And in the nineteenth century, automotive and bicycle pioneers advanced versions of the horseless carriage as extensions of the same design that had evolved incrementally from the Egyptian chariot.

No evidence of wheeled vehicles has been uncovered from the ancient civilizations of North or South America. Although Mexican archeological excavations have yielded ancient wheels, they were not used on vehicles, but were instead attached to small objects that could have been toys or objects of worship. Before European contact, the vehicles used for transportation in the Americas were quite rudimentary. When French explorers reached the Great Plains of North America in the early eighteenth century, they found first nations people moving freight over land by means of an A-frame litter which they named a travois. These freight-bearing
litters were sometimes pulled by dogs, and later by horses.

Moving over (and under) water has been greatly aided by vehicle designs. Boat building has been dated back to 8000 BCE, with dugout canoes carved out of tree trunks found in Africa and Europe. Lighter boats made with tree bark hulls may have been used even earlier, but they were not as sturdy and thus would not leave an archaeological record. A boat constructed 7000 years ago from reeds and tar has been unearthed in modern-day Kuwait. By 3000 BCE, Egyptians had built larger boats that carried people and goods along the Nile, and by 1500 BCE Egyptian ships were sailing the Mediterranean and Red Seas.

The Phoenicians pioneered trade routes around the Mediterranean between 1200 and 800 BCE by deploying square rigged sailing ships for freight transport and a longer and narrower vessel with a sharpened bow that could be used to ram other vessels during maritime battles (Basch 1969). These ships were slowly scaled upward in size and capacity across Europe over two millennia, while the Chinese built even larger ocean going junks with up to four masts during the twelfth through fifteenth centuries BCE (Levathes 1996).

Traveling underwater, particularly during military combat, had inspired elementary diving equipment as early 413 BCE, when The Peloponnesian Wars described divers taking part in the siege of Syracuse. In 1620, Cornelius van Drebbel invented the first submarine while working for King James I of England. It successfully traveled 6 m (20 ft) below the surface of the Thames River, but was not adopted by the Royal Navy. Submarine design evolved incrementally during the eighteenth and nineteenth centuries, becoming a mainstay of naval fleets only after the self-propelled torpedo was invented in 1866.

The first marine vessel made entirely of metal was an iron barge built by England’s John Wilkinson in 1787. During the 1830s, the Laird shipyard in Birkenhead constructed the first all-metal paddlewheel steamboats, combining innovations in propulsion and vehicle technologies. Today’s steel-hulled ships have been incrementally adapted from the Laird’s design and production.

Vehicles used in air travel predate motorized aviation by 120 years. In 1783, a balloon with gondola designed by the Montgolfier brothers carried two men over Paris in the first documented episode of untethered human flight. Airships were the first vehicles to make use of motorized propulsion in the skies when a propeller-engine-equipped balloon designed by Jules Henri Giffard flew 27 km from Paris to Trappes in 1852. By 1891, Otto Lilienthal had flown a glider in Germany that marked the beginning of the use of craft that were heavier than air. Less than a decade after the Wright brothers’ first flight, in 1912 the Avro Type F monoplane provided the first enclosed cockpit to shelter its occupant from the elements. Pressurized cabins extended the heights that aircraft could reach, beginning with the Boeing 307 Stratoliner, which entered service in 1938. By the 1950s, the technology to protect occupants traveling in the skies had been extended to spacecraft that could maintain a livable atmosphere in the harshness of outer space.

By expanding the capability to shelter people when in motion along, above, or below the Earth’s surface, transportation vehicle technology has proven capable of extending the range of human mobility to reach increasingly inhospitable frontiers. While the dream of interplanetary and even interstellar mobility would certainly benefit from advances in propulsion and communication technology, the vehicles that could reach such extreme destinations are
TRANSPORT TECHNOLOGY

no less plausible than the aircraft and spacecraft in use today would have appeared to people in the preindustrial past. Designing and building vehicles reveals considerable skills that humans have developed in extending the comfort and security of their immediate environment.

Building effective infrastructure

A third dimension of transportation technology development is revealed by the infrastructure that has been developed to facilitate mobility. Infrastructure construction reveals not only the technical capability of various stages of human development, but also the dimensions and dynamics of human community since these creations are shared by many users. The paths that people make when moving from one place to another reveal some of the earliest evidence of ancestral human behavior that created transportation infrastructure. In Tanzania, the Laetoli footprints excavated by Mary Leakey show bipedal hominins from 3.8 million years ago following in one another’s footsteps. While the jumbled fossilized footprints proved challenging for paleontologists to interpret when seeking to understand the evidence from humans’ first bipedal ancestor, they clearly reveal those ancestors were creating pathways that eased the journey of those who took the same route. The oldest footprints that can be definitely attributed to humans dated from around 117,000 years ago and were preserved in Langebaan, in South Africa.

Ancient walking tracks have been discovered across Africa and Europe, but when goods began to be dragged over the ground with sleds and travois, improving the geometry of these routes required additional planning and skill. Ridge-ways were trails that followed high ground to avoid swamps and water crossings. Reaching these heights required creating gradients to access the ridge, and evidence of such graded trails dates to 3000 BCE in England and Germany (Lay 1992). England’s Ridgeway National Trail preserves one such route to this day. Another driver of improved infrastructure was urban settlement, where the density of travel by people and vehicles called for more durable surfaces. Evidence of paved streets can be found dating from 3000 to 4000 BCE in excavations of the cities of Harappa and Mohenjo-Daro in the Indus Valley of Pakistan as well as the Mesopotamian city of Ur in modern-day Iraq.

Extending these roads beyond cities enabled consolidated control of large territories. The First Persian Empire created the Western world’s earliest intercity expressway around 500 BCE, featuring a 2699 km Royal Road commissioned by King Darius I that linked Susa in modern-day Turkey to Sardis, near the head of the Persian Gulf. The Greek historian Herodotus wrote that messengers could traverse the length of this road in 90 days, delivering goods and communications across the Persian Empire at a speed and scale unprecedented for the time. Persia’s Royal Road was subsequently incorporated into the Roman Empire’s more extensive road network, which displayed considerable advances in both engineering and pavement technology along with the legal framework to plan and build highways across Europe, the Middle East, and North Africa (French 1998).

Roman civil engineering introduced standardized dimensions and geometry to road infrastructure, along with the use of consistent and durable pavement materials, including concrete. Romans also advanced the use of bridges and tunnels in road infrastructure and formalized the design of sidewalks as a regular component of street design (Van Tilburg 2007). The Roman legal code offers the oldest documentation to date of principles and regulations on road building and maintenance. The Romans built many long and straight roads and established towns...
TRANSPORT TECHNOLOGY

at regular intervals along the roads. Evidence of these Roman roads and town origins can be found today throughout Italy and other parts of Europe. Roman achievements produced an advanced infrastructure that was not surpassed until the eighteenth century in Europe, with the exception of tar pavement, a breakthrough that was made in Iraq around 800 CE and applied to the roads of Baghdad.

France led Europe into the modern era of road building by rediscovering and improving the engineering and legal techniques that the Romans had created. In 1716, France launched its “Grand Corps des Ponts et Chausées” which was the Western world’s first professional road-and-bridge-building bureaucracy. The Corps was responsible for designing and constructing France’s national road network and continues to play that role along with other engineering responsibilities today. During the eighteenth and nineteenth centuries, England created decentralized administrative units to plan, build, and maintain intercity roads. These turnpike trusts borrowed funds to build infrastructure and then collect tolls to repay the debts. The British turnpike trustee who left the greatest technical legacy was John Loudon MacAdam who improved upon the Roman technique of creating a gravel subgrade between the ground and the exposed pavement. He also invented the mix of crushed stones and tar to produce a durable road surface that has made his name synonymous with pavement the world over.

Railroad infrastructure grew out of the early road technology where carts were guided along ruts that became progressively worn into the ground by the passage of wheels. An early example of such a tracked roadway was the Diolkos, a 6–8.5 km paved trackway across the Isthmus of Corinth (Werner 1997). Built sometime between 700 and 600 BCE and used regularly through the first century CE, this trackway enabled both ships and cargos to be rolled across the isthmus between the Gulf of Corinth and the Saronic Gulf, avoiding the long and rough sail around the Peloponnese. Tracks were used to roll heavy loads out of mines from the sixteenth–eighteenth centuries in Europe, and by the end of the eighteenth century English horse-drawn railways were moving coal from mine heads to canal docks several miles away. The first three decades of the nineteenth century saw a rapid evolution of railroad infrastructure, with England leading the world in developing the first recognizably modern railroad operation, the Liverpool and Manchester Railway, which opened in 1830. Today, rail infrastructure serves a considerable diversity of mobility, supporting the fastest movement of commercial transport vehicles on the Earth’s surface through high-speed rail systems in Asia and Europe as well as moving the heaviest loads of freight on land along the railroads of North America.

Port infrastructure dates back to at least 2550 BCE. Archeologists have excavated a 6 km² wharf complex on the Gulf of Suez at Wadi el-Jarf in Egypt from this era (Tallet and Marouard 2014). Building piers, docks, and warehouses to store arriving and departing cargo has increased in scale up to today’s busiest port infrastructure in Shanghai. Canals appear to have first been dug for irrigation in Mesopotamia and the Indus Valley, but they were being used as a transport corridor to enable boats to bypass rapids and waterfalls along the Nile River by around 2300 BCE. The advantages of water transport in the era of non-motorized transport inspired ongoing efforts to build such infrastructure in the ancient world. For example, the Chinese built their Grand Canal between Beijing and Hangzhou, a distance of 1794 km, by connecting shorter canal stages between around 300 BCE and 609 CE (Davidson and Brooke 2006). China’s Grand Canal is still used today and remains the
TRANSPORT TECHNOLOGY

world's longest. But the highest volume of ship traffic to use such infrastructure passes through the 193.3 km Suez Canal, which was completed in 1869 and allows Asian traffic to reach Europe without sailing around the Cape of Good Hope in South Africa, and the 77.1 km Panama Canal, completed in 1914, which allows ships to transit directly from the Atlantic to the Pacific oceans bypassing the long journey around Cape Horn in Argentina and Chile. A third Panama Canal channel is being built to expand the capacity of this waterway.

Unlike ports, airport infrastructure does not predate motorized mobility. The oldest airport still in use dates from 1909 and is located in College Park, Maryland. At this airfield, Wilbur Wright offered the first training flights to US military personnel. As with ports, the width and length of the runway have expanded, along with terminal and support buildings and ground transport access, but the basic principles of aerodrome design have not changed. The first spaceports were built during the Cold War to launch missiles and later manned spacecraft. To date, there are only four spaceports that regularly launch vehicles into outer space: the Baikonur Cosmodrome in Kazakhstan, the Cape Canaveral Air Force Station (now known as the Kennedy Space Center) in Florida, USA, the Guiana Space Centre in Kourou, French Guyana, and the Jiuquan Satellite Launch Center, in the Gobi Desert, Inner Mongolia, China.

Our infrastructure legacy reveals a great deal about our evolving society and its approach to meeting mobility needs. From paths that were informally carved out of the wilderness and shared by our nomadic ancestors to paved roads that have been designed and administered by ancient empires and modern states, infrastructure design and development reflect the ideas that our predecessors held about pooling their skills and resources to make travel easier, not just for themselves but also for others. As the scale of that challenge mounted, with growing technical complexity and economic cost needed to produce increasingly sophisticated infrastructure, the mechanisms for sharing those burdens were created. From the compulsory labor of slaves in the ancient world to the public finance of expressways, ports, airports, and spaceports today, infrastructure building demonstrates how society has pooled its talents and resources to create shared spaces that enable mobility to be more productive and effective.

Contemporary technologies

Transport has benefited from numerous technology innovations in recent years. In freight transport, the practice of intermodalism, that is, the ability to move goods seamlessly from one mode to another in order to optimize the strengths of individual modes, has been facilitated by several technological innovations. The introduction of the standard container, based on the twenty-foot equivalent unit (TEU), has revolutionized freight transport by facilitating the loading, unloading, and transferring of goods from one mode to another. Together with the mechanized gantry crane, the time it takes to unload a large cargo ship is now measured in hours rather than days. These technologies have contributed to the increasing size of container ships well beyond the limits imposed by the length of the locks in the Panama Canal, resulting in a new generation of “post-Panamax” ships.

The current expansion of the Panama Canal will allow larger ships to be able to use the canal, which is expected to lead to container ships from Asia bypassing US west coast ports in favor of US east coast and gulf coast ports. At the ports, containers can be moved directly from a ship to a rail car or a truck where they can be transported
to inland terminals and eventual destinations. The innovation of double-stacking of containers on rail flat cars has greatly increased rail freight capacity, contributing to the recent resurgence of the freight railroad industry. Intermodal connectivity has been facilitated by improved information technology, such as electronic data interchange (EDI) which has permitted more seamless coordination of freight movements.

The growth and sophistication of information technology infrastructure have had a huge impact on contemporary transportation. Major investments in intelligent transportation system (ITS) development have led to smarter vehicles and infrastructure systems that have automated many elements of the transport process. The operation of automobiles and other motor vehicles has been facilitated by technologies such as cruise control, automated safety systems, and global positioning systems. A considerable amount of progress has been made recently in developing driverless vehicles, led by Google’s pioneering efforts, and aircraft technology has improved greatly so that many of the pilot’s tasks have been automated. A long-term goal of ITS is the development of integrated vehicle/infrastructure systems that will allow smart vehicles to operate in automatic fashion on smart infrastructure, thus improving efficiency and safety greatly. Information on traffic conditions is now routinely collected and disseminated to improve the flow of traffic, so that vehicle operators or traffic controllers can make better decisions about specific routes to take to avoid congestion. Automated ticketing on air, rail, bus, and other passenger transport systems has facilitated the use of these modes, and automated tolling on highways has made the collection of user fees much easier.

The use of technology has been extended to improving the sustainability of transport systems. Development of alternative fuel vehicles, such as electric or hybrid automobiles, has contributed to greater fuel efficiency and allowed transport to utilize more renewable sources of energy rather than continued reliance on fossil fuels. Technological improvements in internal combustion engines have also helped to reduce harmful emissions of local air pollutants and global greenhouse gases. Continued research and development in using alternative fuels, improving fuel efficiency, and reducing emissions from aircraft, motor vehicles, ships, and rail locomotives has the potential to significantly reduce transport’s carbon footprint and other negative externalities.

SEE ALSO: Transport geography; Transport policy; Transportation history; Transportation and land use

References


**Further reading**

Transportation and land use

Selima Sultana
University of North Carolina at Greensboro, USA

Transportation serves as a fundamental linking factor between people and for moving goods and information, and has spatiotemporal dimensions. Hence, throughout history transportation investments have had a tremendous impact on land use, yet defining land use is a complex task. In general, land use is the conversion and management of the natural environment for human use. Examples might include the conversion of forests into housing developments with accompanying road networks, or of grasslands into farm land.

The relationship between transportation and land use has been a dynamic process in which the two are interdependent and inseparable. Historically, changes in transportation systems, modes, and speeds have created opportunities to overcome distance, which have been linked to changes in land-use patterns. Similarly, changes in land-use patterns influence travel demand and induce changes in transportation systems. Thus, the interactions between transportation and land use are often referred to as having a “chicken-and-egg” connotation with an essentially inseparable outcome regardless of which change occurred first.

This entry focuses on the nature of these complex processes of interactions between transportation and land use for the twenty-first century. It will first provide an overview of the transport and land-use connection in a historical context. It will then focus on the challenges that have arisen from the dominance of automobile-oriented land-use patterns.

Transportation networks as strategies for spatial interactions of land use

Transportation has always been closely related to land-use developments throughout the world. Preindustrial patterns of compact and dense land use reflected the difficulty of travel and the reliance on slow modes of transportation such as walking. Innovations in transport technologies and their respective transport modes that prevailed at different time periods have had a profound influence on land-use patterns. These relationships can be summarized by reference to specific technological and historic periods.

Before the 1800s, waterways and road networks mostly connected coastal areas and had major influences on early land-use patterns. Transport by waterway was substantially more efficient in terms of time and cost than transport over land, especially because of the limited road capacity that prevailed at that time. As a result, settlement patterns were prevalent along waterways because they were able to maintain economic, political, and cultural activities by creating trade relationships, even over longer distances. As such, the most dominant cities of that time emerged along the waterways in places such as Mumbai, Suzhou, Lisbon, Athens, and Istanbul. Similarly, many early American cities developed along the waterways as they provided access to the interior. Jamestown, Charleston, Savannah, Boston, New York City, and Philadelphia are salient examples.
TRANSPORTATION AND LAND USE

With the development of the steamboat in the 1810s, waterways even became a prevailing force in American settlement patterns. Since steamboats could travel upstream, inland river towns, such as St Louis, New Orleans, Chicago, Milwaukee, and Buffalo, were formed along the major rivers and lakes. They were connected by waterways to the larger port cities of the Atlantic and Gulf coasts.

The development of rail networks made the hinterland more accessible, especially in North America. The setting of new rail connections between 1830 and 1920 resulted in significant land-use changes in the American hinterland with the establishment of rail-based cities such as Kansas City. Unlike waterways, which could only connect locations along navigable rivers or canals, railroads could create their own pathways and were also more efficient than waterways because of their speed and reliability. As a result, rail transportation emerged as a new factor for developing lands between waterways and the hinterland by facilitating better access and connectivity to the existing waterway networks. Since it was easier and cheaper to ship crops and livestock to markets from farms close to a rail station, public lands were given to railroads to capitalize construction and create markets along the lines of railroad constructions. Once railroad spurs were built, land agents were hired by the railroads to recruit settlers as quickly as possible by underlining the benefits of the agricultural land adjacent to the newly constructed railways. To recruit new settlers, railroads carried buyers on a special land-seeking train, with the train ticket price included in the price of the land. Marketing of land using railroads was so popular that by the 1870s railways had become the principal mode of transport for passengers and freight in the United States. Not only did railroads lead to changes in land-use patterns nationwide, they also shaped the development of land use at the local level. For instance, the railroad station became the nexus of many urban cores, resulting in the decline of waterways and many small river towns.

The mass production of the automobile from the early 1920s changed the land-use dynamic around the world and to a greater extent in the United States. Automobiles allowed people to travel faster and wherever roads could be built, hence allowing the development of land in areas that were previously isolated, underserved or underserved by the railroads. The “Autostrade” network, the world’s first intercity motorway, was built in Italy in the early 1920s and eventually linked all regions of Italy with over 6000 km of highways. The “Autobahn” is another example, the first highway system in the world, built in Germany during 1913–1921. This highway system was later used to expand and build the country’s economic growth and financial corridor allowing for transportation of goods from all areas of Germany and beyond. After World War II there was a great deal of interest in launching major highway construction projects in the United States. By the 1950s, the federal government underlined that infrastructure based on roads and interstate highways would be in the strategic interests of the country, particularly to improve the connectivity of its major cities. This led to passage of the Federal Highway Act of 1956 that allowed substantial federal funding to build the Interstate Highway System (IHS). The effects of the IHS were dramatic. The interstate highway network offered tremendous flexibility to individual businesses to deliver goods by truck and resulted in a significant decline in rail transportation. By 1970, it was apparent that interstates carried more vehicles than anyone had anticipated. The IHS and increased automobile use had significant impacts on land use in urban areas across the United States.
Urban modes of transportation and land use

Transportation serves as a fundamental linking factor between a person’s residence and his or her everyday activities within urban areas. Where an individual chooses to live can be decided based on the availability and cost of transportation and the distance to work or to other activities. The innovation of new transportation technologies over the centuries promoted changes in the spatial organization of land uses within and in the vicinity of urban areas as new modes of transport came into use. It is common to categorize the influence of transportation modes on land-use characteristics by different eras (Adams 1970).

Walking and horse car era (before 1890)

The urban land use of this early era was very compact and dense (75–100 people per acre), with most destinations within a 30-minute walk, or within 2 miles, as people were only able to walk or ride horse cars to move within a city. The separation between home and workplace was unnoticeable as factory owners often built their homes next to their factories; artisans and storekeepers lived above or behind their workshop or store. Even though in the mid-1800s railroads were available, they were principally developed for transportation between cities, not within cities, and that led to intense competition for land around railroad stations. Factories, warehouses, and shops were able to pay higher rents for the most accessible areas within cities and they were located in close proximity to railroad stations.

Electric streetcar era (1890–1920)

By the late nineteenth century, cities were very crowded and that increased demand for more transportation. Alternative means of propulsion began to be experimented with, using cables (cable cars) and then electricity (electric streetcars or trolleys). Horse cars (and most cable cars) were eventually converted to electric streetcars, and many new streetcar lines were built in almost every city and in large towns. These systems had very dense grid networks with sidewalks, much like bus routes today. They ran on most major streets, and sometimes on every street in downtown areas. In this era, the land uses in most cities were still compact, but the residential fabric of cities changed radically by expanding along new outlying trolley corridors for the increase in accessibility these areas offered. There were some signs of change in housing styles as well from multifamily to single-family and two-family housing types of land use, but they were built on narrow lots on a grid pattern with sidewalks connected to the main street, and land values decreased away from streetcar lines.

Early auto era to freeway era (since 1920s)

Over the last 100 years automobiles have dominated transportation systems in developed countries, especially in the United States, pushing the urban and suburban boundaries into the surrounding rural areas from the city cores of the nineteenth century. Low-density housing development started outside the city with wider housing lots for driveways and garages as automobiles gained popularity in the 1930s. One notable consequence of this spatial structuring of land use was the increasing separation between workplace and residence, but railroads were expanded to these communities for commuting to city jobs.

The construction of post-World War II freeways accelerated access to plentiful and inexpensive land in rural areas and led to the
TRANSPORTATION AND LAND USE

construction of more roads and low-density residential areas in suburban and exurban areas. This came to be called “suburban sprawl,” with wide lots on various types of curvilinear road networks with differing land uses located far from each other leading to the growing use of automobiles for every trip. Due to the decline of automobile costs coupled with affordable housing options, many workers were amenable to living further from their place of employment, sometimes more than 30 miles from work. This distance necessitated long commute times, and automobiles became by far the most commonly used mode of transportation for commuting to work, which led to the decline of public transportation. Indeed, many systems, especially regional transit, have in many cases seen large decreases in funding because of a perceived lack of need leading to limited investments being made in maintenance and expansion of public transit.

Beginning in the 1950s, geographical expansion of suburban growth was uncontrollable, which led to significant changes in land-use patterns and the distribution of population accompanied by a dramatic increase in automobile numbers. By the end of the 1970s, the uncontrolled outward suburban growth reached a point where more Americans lived in suburbs than in core areas of cities (Table 1). This trend still continues today even though many core areas of cities saw some population gains in the past two decades. Today the average American commutes for a little less than 26 minutes and 16 miles each way. Some 10.8 million (8.1%) US workers commute for more than 60 minutes, and another 600,000 (0.8%), labeled “mega-commuters,” travel for 90 minutes or more and at least 50 miles each way for work (AASHTO 2013).

<table>
<thead>
<tr>
<th>Year</th>
<th>Cities</th>
<th>Suburbs</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950</td>
<td>32.9</td>
<td>23.2</td>
<td>43.9</td>
</tr>
<tr>
<td>1960</td>
<td>32.3</td>
<td>30.6</td>
<td>37.0</td>
</tr>
<tr>
<td>1970</td>
<td>31.4</td>
<td>37.2</td>
<td>31.4</td>
</tr>
<tr>
<td>1980</td>
<td>30.0</td>
<td>44.8</td>
<td>25.2</td>
</tr>
<tr>
<td>1990</td>
<td>31.3</td>
<td>46.2</td>
<td>22.5</td>
</tr>
<tr>
<td>2000</td>
<td>30.0</td>
<td>50.0</td>
<td>20.0</td>
</tr>
<tr>
<td>2010</td>
<td>33.0</td>
<td>51.0</td>
<td>16.0</td>
</tr>
</tbody>
</table>


developed uses of land have expanded by more than 1 million acres annually since 1982 (Alig and Ahearn 2006). It is estimated that there were 73 million acres of urban and built-up land in 1982, and this grew to 111 million acres in 2007, a 56% increase. One may argue that this expansion of urban land uses results from an increase in population, mainly from immigrants, and an increase in wealth combined with certain lifestyle choices. However, population and income pressure are not consistent with the ratio of urban land consumption, which expanded at a faster rate than population growth until the 1990s (Figure 1). Urban boundary expansions reached the point where rural communities started seeing these urban expansions as a negative change in their communities, raising concerns over unchecked growth. Eventually, many rural communities adopted land-use controls designed to make suburbanization difficult and costly. Many rural land-use regulations allow by right (without regulatory approval) uses that support a main use, such as agricultural districts that typically allow a number of uses that would need special permission in urban areas. During this time, regulation on lot sizes also tended to be greatest in urban areas. Land-use controls
that mandated large minimum lot sizes had immediate effects on urban land-use expansion, reflected in Figure 1.

Concurrently, since the 1970s, commercial land uses, such as mass concentrations of office parks adjacent to retail-based industries and skyscraper buildings along the intersections of major highways, have emerged as a distinctive urban-like phenomenon in America’s suburban landscapes. Commercial suburbanization was so rapid that, by the 1980s, job growth (about 150%) in suburbs outpaced population growth in the United States. Today, typically there are more jobs located in suburban locations than within the core urban areas (CTPP 2011). In general, these are areas of lower building densities than downtowns, with large buildings widely separated from others by grass and parking lots with an excessive use of land area. For example, the Cumberland-Galleria and Perimeter commercial areas in the Atlanta metropolitan area have a larger land area than downtown while also having lower population and worker densities. Roadways in these areas often lack sidewalks, which, combined with the low densities, often make it obligatory for workers to drive between
adjacent buildings. These areas typically have limited public transportation services (if any) due to the lack of density and because their locations along expressway and freeway corridors allow for easy access by automobile. This trend continues today, but even farther out in exurban locations where low-density, scattered, and small-scale commercial development such as retail and wholesale is also visible.

Transportation and classical urban land-use models

Classical urban land-use models are the starting point for conceptualizing the continuum of land use in relation to transportation. Burgess’s concentric zone land-use theory (CZLUT; 1925), Hoyt’s radial sector or wedge land-use theory (SWLUT; 1937), and Harris and Ullman’s multiple nuclei land-use theory (MNLUT; 1945) are among the best known. Burgess’s CZLUT viewed the city as a dynamic place evolving from a center point of greatest accessibility of transportation to a series of concentric zones, or rings, like a two-dimensional bull’s-eye with the center zone containing the central business district (CBD) as well as the transportation, social, and cultural hub of the city (Figure 2). According to the model, urban land use is based on distance from this hub in relation to transportation costs and has six rings: CBD, a factory zone, the zone of transition, the working-class residential zone, a
middle-class residential zone, and the commuter zone (usually wealthy single family homes).

In CZLUT, land uses are in competition for space based on their ability to pay for higher valued locations near the center of the city, with the CBD having the highest value, due to its having the highest transportation accessibility, and therefore occupying the most valuable position. Outside the CBD, the secondary ring was a transitional land-use area with mixed uses, including some manufacturing and low-income or slum housing for greater access to transportation. Residential density continues to fall as the rings move outward, with an inverse relationship between residential categories and income. The poorest residents of the city live nearest the CBD to reduce the transportation cost and the wealthiest reside at the furthest edge of the city while incurring higher costs in transportation.

Burgess envisioned an expanding CBD exerting land-use conversion pressure on the surrounding district; this in turn pushed outward on the other districts. In this and later monocentric models, changes in transport costs determine changes in the relative value, and therefore use, of land at specific distances. The concentric zone model therefore follows bid rent curves. Bid rent curves illustrate how the most expensive land is at the central point of a city (e.g., the CBD) and that land value goes down as distance from the CBD increases. While this logic fit the constraints of the first half of the twentieth century, CZLUT focuses almost entirely on the CBD, residential, and factory land uses with no real consideration of the nuances of institutional uses, wholesale and retail sales uses, or office uses, or even ownership characteristics of land, as Burgess essentially envisioned all of these uses to be contained within the CBD.

Two other classical models are SWLUT and MNLUT (Figure 3). These theories are more representative of actual spatial development in urban land use, and both of these theories take the same classifications and land uses of CZLUT and apply them to wedges or sectors radiating out from the central business district, but MNLUT recognizes that an urban area may have multiple hubs of activity outside the central business district. Hoyt’s SWLUT aligns land-use zones along major transportation routes such as highways, rail lines, or ports, but also takes into account geographic realities such as water bodies and the landscape (Figure 3a). SWLUT proposed several major organizing principles of land use. Relative location shapes how the urban form develops in a city. Corridors of industry and warehousing tend to be surrounded on both sides by sectors of working-class housing. Middle-class housing acts as a buffer between industrial/working-class areas and the main sector of elite neighborhoods. Affluent households and transportation have a large impact on city structure. Once the CBD is established and the corridors of industrial development are laid out, affluent households develop adjacent to the most desirable sites. High-status areas then expand along transportation lines.

In this theory, sectorial development occurs with generational shifts. Each new generation of affluent households buys or builds at the current urban “edge” of the city, far from the poorer sectors. These new affluent generations sell off their old family homes, most likely to families of slightly lower status. In this way, sectors are banded into residential zones of different ages, styles, and conditions, moving outward from the oldest areas in the city center to newest areas at the city edge. Over time, high-rent sectors tend to grow outward along major transportation routes. Affluent areas also grow along ridges of high ground, away from the risk of floods and where the best views are located. Hoyt also recognized that neighborhoods go through changes
over time through filtering and vacancy chains. The chain of residential moves initiated by the construction of new homes for the affluent leads to the filtering of households up the housing scale and the consequent filtering of houses down the social scale. This process creates a vacancy chain in which higher-income families move out and lower-income groups fill in houses and areas left behind. The vacancy chain keeps going until it stops at an old house no one wants to move into. This has been the case in the inner city for a long time and the same pattern is taking place in the inner-ring suburbs, the areas immediately outside of central cities. Unlike CZLUT, SWLUT does not address commuter zones.

MNLUT is a schematic representation of the same major categories of land use with several economic centers, based on the decentralization of commercial and industrial nodes beyond the CBD (Figure 3b). MNLUT argues that automobile-based suburban nodes of commercial and industrial activity are not arranged in any predictable fashion, except in relation to surrounding land uses. These nodes may develop around transit stops or highway intersections. If they were office and retailing centers, they would attract middle-income residential development, whereas if they were industrial centers, these nodes would attract working-class residential development. If the node was an airport, it would most likely attract shippers such as

---

**Figure 3** (a) Hoyt’s sector model; (b) Harris and Ullman’s nuclei model. Rodrigue 1998–2016. (Reproduced by permission of Dr Jean-Paul Rodrigue.)
FedEx, hotels, and working-class residents. The mutual attractiveness and mobility allowed by the potential income of each node drives what develops there. The CBD may be separated from the transportation, education, or cultural hubs, thereby creating a scattered land-use development effect. Heavy industrial uses in this model are located on the periphery of the city. The result is an irregularly shaped patchwork of land uses across the urban landscape, which is remarkably present in today’s cities with polycentric and dispersed patterns. American classical models are all essentially center-oriented and do not attempt full descriptions of all land usage. Moreover, these earlier models did not consider the impact of newer technologies such as information and communication technology (ICT) on land uses.

**ICT and land use**

As a rule of geography, if distance between people increases, social interactions decrease as the cost of transportation increases, and this explains why compact land uses were necessary in early cities. Automobiles and freeways have reduced the cost of travel and allowed for greatly expanded cities. Communication costs have fallen and the ability to transmit information (bandwidth) has greatly increased. Some scholars suggest that the Internet and new telecommunication technologies can be a viable option for communication and, therefore, substitute for physical transport. Many tasks that once required personal travel, such as going to the bank, arranging insurance, filing tax forms, searching for library books, and even taking classes or working, can now be done electronically at home. Technologically deterministic futurists have made predictions based on these ICT trends that transportation will not matter for locational decisions of households and firms as people can access anything from anywhere at any time. Commuting distance will no longer be a consideration for residential and firm location, so people will be able to live anywhere or even nowhere near their workplace, therefore travel will be substituted for. This may essentially lead to a weakening link between transportation and land uses.

There appears to be no evidence so far that ICT has much of an influence on land use nor on travel behavior. There is some evidence, however, that back office activities such as routine, repetitive data processing work that follows standardized procedures does not necessarily need to take place in cities or nearby suburbs and can be performed anywhere around the world. Front office activities – highly skilled and specialized occupations – still require face-to-face contact as innovation and creativity thrive in areas with daily intellectual contacts between workers, and still require a center location in the cities. Whether communications could substitute travel or transportation is also questionable as evidence shows that the emergence of e-commerce has actually triggered a surge in goods shipments from businesses to consumers, which is far from substituting transportation or reducing travel. Certainly, ICT complicates the straightforward linkage between land use and transportation and this effect should be integrated to better understand transportation and land-use modeling in future research.

**Challenges for transportation and land use in the twenty-first century**

During the past several decades there have been growing concerns over climate change issues to which automobile-dependent transportation systems with associated low-density land uses are
TRANSPORTATION AND LAND USE

a contributing factor. This has resulted in a general consensus on the need to reduce greenhouse gas (GHG) emissions from transportation – the second largest contributor (about 30%) of total GHG emissions worldwide. These concerns have made the relationship between transportation and land-use even more complex. A recent emphasis has been on encouraging land-use planning policies in ways that will alter the possible relationship between urban form or built environment and travel behavior. Hence, the smart growth and new urbanism concepts emerged in the early 1990s as an explicit alternative to sprawl-oriented development, aimed to integrate urban land use and transportation development in a way that limits outward expansion of cities by preserving open space and promoting density, diversity, variety of transportation options, and job–housing balance – much like historic European cities (Congress for the New Urbanism 2001). Therefore the consensus currently favors promoting compact land-use patterns with high-density, pedestrian, and transit-oriented cities where every trip does not have to be initiated by automobile.

Since then considerable attention has been devoted to understanding how the neighborhood built environment, including density and land use, impacts mode choices, especially on nonmotorized modes of transportation. There is clearly an agreement in the literature that residential environments with a mix of land-use types (such as residential, parks, and small shopping centers), higher population densities with accessible transit, a well-connected street network with short distances between intersections, and a pleasant walking environment (with good sidewalks, shade, and nice views) as well as a lack of parking spaces in commercial areas are associated with higher levels of nonmotorized mode choices (walking and cycling). Residents of these neighborhoods own and use automobiles less and have a reduced carbon footprint.

Today, thousands of new urbanism communities are being developed throughout the country and many downtowns are now being rejuvenated by utilizing this concept. To a certain extent new urbanism has succeeded. New urbanist community residents are far more likely to walk to where they shop relative to their counterparts in traditional suburban neighborhoods, and more likely to use transit if available, but they still use cars for commuting. This has raised ongoing debates among academics and policymakers as to whether it is possible to change car use by creating compact land-use patterns. In reality, American cities are not well served by transit networks. In addition, despite the unequivocal goal of income diversity for new urbanism, many critics have pointed out that most new urbanist communities are artificial, exclusive, and expensive. The nature of the style is affordable only by those in the upper middle income to upper economic classes. As a result, two of the most significant challenges that new urbanism communities face today are to create socially diverse (in terms of mixed income) neighborhoods and to provide transit connection outside these developments to the rest of the areas in the cities. So far there has been no easy fix.

Moreover, it is indisputable that the majority of American urban growth is driven by low-density suburban and exurban expansions, and their existence cannot be just ignored. It is also clear that the opportunity to completely wipe out the suburbs and build new cities and towns with a compact philosophy is impractical. Historically, density tended to increase by locating housing and commercial businesses near transit stops and stations. For that reason transit–oriented development (TOD) has become a possibility to utilize most effectively these existing suburban land-use patterns by creating higher density
regional employment and residential centers connected by transit. The diameter of these types of developments is usually one-half mile, with stations spaced one-half to one mile apart. Residents of this community can rely on rail, and are able to walk or cycle between the stations and their homes. Despite decades of strong resistance to increasing density in suburban areas, there seems to be already some evidence of cultural shift in suburbanized areas, as reflected by a growing urgency in the implementation of smart growth and new urbanism development concepts in land-use planning. If this trend continues, it is expected that TOD could be more favorable to many suburbanites in coming years.

SEE ALSO: Edge city; New Urbanism; Rail transport: freight; Suburbanization; Sustainable transport; Urban transit

References


Further reading


Transportation history

Ka-chai Tam
Hong Kong Baptist University, China

Transportation refers to the means of transporting people, freight, and information from one place to another. In addition, transportation history also focuses on the factors that influenced the setting of transport routes, the emergence and diffusion of transport technologies, and even the biographies of the key persons responsible for the discovery of trade routes or the invention of key technologies, for example.

Walking obviously marks the first step in transportation history. A reliance only on human legs implies movements taking place at a slow pace of about 5 km per hour, and the distance that can be travelled in a single day would be no more than 40 km on good roads. A more constraining environment could easily reduce this range. That people needed to return to their stable shelters before dusk would also have restricted their travel distance by half or more. Therefore, in the past the range of human activities was highly limited by the distances that could be crossed and the amount that could be carried by foot. The emergence and extension of civilizations was, in the earlier stages, closely related to transportation. Transportation is considered one of the four essentials of human life alongside clothes, residence, and food according to the traditional popular Chinese wisdom of civilization, for example.

Three inventions are indeed key to the initial historical impacts of transportation on human activities before the Industrial Revolution: the understanding and application of the principle of buoyancy and wind power for water transportation; the domestication of animals as a source of power in transportation; and the invention of the wheel, which increased the capacity and efficiency of land transport.

Water transport systems

Exploitation of the buoyancy of wood and other materials, crafted into vessels, represented a crucial development in transportation. Early propulsion systems were rudimentary, for example using pole or a paddle to push a vessel over shallow water. The presence of a native population in Australia as far back as 40 000 to 45 000 years ago is indicative of early forms of simple seafaring vessels that enabled people from Eurasia to travel to that isolated continent. The reed boat is one of the older vessel designs as used for river navigation in Egypt, Mesopotamia, and India around 7000 years ago. Later, the galley was an early example of a seafaring vessel that appeared around the Mediterranean by 800 BCE. Using paddles as the main mode of propulsion, the galley was effective for calm sea as well as river navigation. The large crew required to power it, however, limited its capacity and range, as it was necessary to carry food and water to sustain the rowers. It was good for short trips within the Middle Sea but not a practical means of transportation in the blue water. Rowing was also harsh work, and was often done by convicts or slaves; early naval warfare was waged to capture slaves for galleys. The galley was one of the critical elements that helped the Greeks...
TRANSPORTATION HISTORY

to establish their colonial empires in the eastern Mediterranean, and later the pillar of the Roman navy. In the early modern period, galley navigation also contributed to the reform of the penal system in southern Europe as a poena extraordinaria could replace a death sentence with a galley sentence.

The sail is a convenient means of using wind power for propulsion for prolonged periods of time and is thus suitable for deep sea navigation. In most parts of the world, sails have been made from canvas of linen, hemp, or cotton. However, canvases were subject to wear and tear and to degradation by sunlight and water absorption, and hence needed to be repaired and changed frequently. Vessels in the Far East employed bamboo mat instead of canvas which was more resistant to tearing, waterproof and buoyant, and easier to repair en route.

The square sail, which is more effective in downwind sailing, is one of the two basic types of sailing rigs. It was invented in Mesopotamia around 4500 BCE to 6000 BCE and later independently developed in China and Ecuador. The other type is the triangular fore-and-aft sail invented in the Mediterranean region around 1000 CE. It is set completely aft of a mast and placed parallel to the ship’s keel, enabling it to take wind on either side. The triangular sail performs much better than the square one for upwind and crosswind sailing.

Inadequate wind strength and unpredictable wind direction are major constraints for blue water navigation. From the sixteenth century onward, the use of combined square and triangular sails by the European ships significantly enhanced navigation by allowing ships to be steered in several directions in a prevailing wind. The wide application of the compass and watertight compartments, both invented by the Chinese but soon adopted by the Arabs, Indians, and Europeans in seafaring, also contributed to more accurate navigation and hence survival in the ocean.

Facilitated by these new elements, the Europeans were then able to establish long-distance trade networks that went beyond the limits of the Mediterranean. While Arab and Indian merchant fleets frequently harbored around the Indian Ocean, by the tenth century Chinese merchants were also exploring the South China Sea and reaching as far as the East African coast to establish trade and diplomatic relations. At the peak of China’s maritime dominance in the early fifteenth century, Admiral Zheng He (1371–1433) led a large fleet of 317 vessels manned by 28,000 crewmen on seven major expeditions in these maritime regions, just prior to the rise of the Iberian sea powers. Arab, Indian, and Chinese merchants, each of whom had already established important elements of a world trade system, traded goods and exchanged information and ideas. Later, in the seventeenth century, European powers took over these trade routes, established centuries beforehand, in addition to their own elements and discoveries, and consequently created their global maritime empires.

The domestication of transport animals and the wheel

Unlike waterborne transport developments, which took place in different parts of the world independently, breakthroughs in land transport developed mainly in the Eurasian steppes, which offered large tracks of flat land and native species adapted to travel long distances. The nomadic inhabitants of this region employed ox wagons, horseback riding, and chariots to extend their reach. By exploiting the strength of large mammals they were able to travel much faster and with heavier loads than on foot, and to increase the speed and capacity of transport. Nomads of
the Eurasian steppes became traders, conquerors, and wanderers who imposed great changes on the peoples around them through the course of human history.

The ox, one of the first mammals used for transport, was domesticated from 9000 to 7000 BCE. In addition to providing meat and milk, the animals were employed in agricultural activities because of their muscular strength. When equipment was invented and put to use, the ox also became a means of transportation. The domestication of horses probably took place in the Caspian Sea region around 4800 BCE, well after sheep, goats, pigs, and oxen had been domesticated. Horseback riding was developed well after the horse’s domestication, as it took a long period of time to control this less docile animal effectively. In contrast to the slower pace of oxen, the superior speed of the horse was ideal for faster travel, either for pulling vehicles or for horseback riding. From then on, the horse became the dominant form of land transportation until the Industrial Revolution.

The invention of the wheeled vehicle is more complex, involving the combination of the wheel, the axle, and the frame to create a vehicle with load-bearing moving parts. This form of transportation appeared around 4000 BCE, when Eurasian kingdoms reached a level of technical expertise and craftsmanship in the manufacturing of wooden and metallic objects. This device required precisely crafted parts to ensure that the wheels fitted through the nave, as well as calculating the correct size of the wheels and the vehicle to ensure that it could carry a particular load and travel at speed. Four-wheeled wagons were effective at carrying heavier loads on flat surfaces, while two-wheel carts where easier to pull and able to travel over more difficult terrains. It took centuries of experimentation in different areas to determine the right forms of vehicles to fit their respective transportation needs. Oxen were used to carry wagons with heavier loads over a flat surface, while horses were employed to pull lighter vehicles at a faster speed and in more difficult conditions. Incidentally, the introduction and diffusion of the wagon from 3500 BCE had significant impacts on the settlement pattern of farming communities in Europe, as it allowed individuals to carry manure out to the fields and to bring firewood and other supplies back home, and therefore reduced the need for cooperative communal labor beforehand.

Not unlike many other inventions in human history, some major transportation means were transformed into military equipment. Horse carriages, for example, were equipped with armor to become chariots and the horse was saddled by the cavalry. The military application of transportation significantly contributed to empire building around the Eurasian steppes. For instance, at the peak of their power around 1300 CE, the Mongol empire occupied most of the Eurasian continent, linking Europe to the Far East through what came to be known as the Silk Road, which had been established centuries earlier. Under the umbrella of Mongol dominance, exchanges of cargoes, scientific ideas, religions, and, more subtly, diseases between Europe, Central Asia, and East Asia reached unprecedented levels.

Besides the domestication of horses in temperate climates, camels and donkeys were also tamed from 4000 BCE in the savanna and desert regions of North Africa and the Arabian Peninsula as means of transport that were suited to these harsher environments. Their capacity to carry goods over difficult terrains unsuitable for wheeled vehicles made them the backbone of interregional trade that bridged parts of Eurasia separated by deserts.
TRANSPORTATION HISTORY

Roads and canals

Without roads travelers usually walked, rode, or drove to their destinations along the easiest paths using the shortest distances. The most frequented paths naturally became the first roads. To support empire building, road networks were artificially designed as an effective means to transport troops, collect taxes, trade, and deliver messages. In the major empires these systems usually focused on the imperial capital. For instance, by 500 BCE the Persians had built highways to link the many provinces of their empire. This legacy was surpassed by the Romans who created an extensive road system that is one of the hallmarks of their rule: the 113 provinces of the late empire were connected by 372 major roads, and one-fifth of the 400,000 km of Roman roads were stone-paved. In contrast, around 200 BCE, the first emperor of China, Ying Zheng (r. 260–210 BCE), constructed a network of very wide roads using the rammed earth technique. After prolonged use, the surface of the Chinese earthen roads would be eroded away by the wheels to form parallel ruts, which facilitated carriage transport with higher speeds and stability, so it became necessary for the imperial government to enforce a standardized cart gauge all over China. Many of the subsequent empires established post stations and hostels beside roads to enhance the convenience and security of traveling within the realm.

The earliest canals were built by the Egyptians around 2000 BCE to link the Nile River to the Red Sea. In Europe, the Dutch and the Venetians began to build canals during the Middle Ages to cope with the marshy terrain. These human-constructed water courses were implemented wherever an extensive river system was present and where their construction was justified by economic and political imperatives. Land transport was far more expensive than river transport. For example, it was estimated that the cost of land transport in China by ox wagon was 50 times that of river transportation. It is not surprising that China, from the seventh century onward, built the largest canal system in the world from south to north to connect most of its river systems flowing eastward from the Himalayas. The Grand Canal linked Beijing with Hangzhou over a total length of 1794 km, connected by a system of double-gate locks where barges could be raised or lowered to different water levels. Similar designs appeared in the Netherlands in 1373. Canal systems formed the backbone of inland transportation in many countries on different continents before the introduction of the railways in the nineteenth century.

The age of discovery

Prior to the eighteenth century, all transportation means were very slow by modern standards, and their main forms had remained unchanged for centuries. The major concern of premodern transportation was cost. Following the advancement of navigation technologies and the accumulation of travel experiences, safety and risks could be more accurately evaluated for long-distance transportation. Maps and log books were widely circulated. Celestial navigation tools were also employed alongside the compass to accurately estimate the location of ships.

“The age of discovery” is a Europe-centric interpretation of exploration activities that took place from the fifteenth century onward, when the rising colonial powers of Europe, either in competition or in cooperation with one another, began to build a global trade network. The Portuguese were the pioneers in exploring the Atlantic coast of Africa from 1418; the
The Industrial Revolution and modern transportation developments

New means and systems of transportation have emerged at a far faster pace since the Industrial Revolution started in the Great Britain in the eighteenth century. This involved greater speeds and capacities through the development of mechanical forms of propulsion. The steam engine was devised by the Scottish mechanical engineer James Watt to activate water pumps in mines, but it soon became evident that, by installing the engine directly inside a wheeled vehicle or ship and letting the steam power rotate the wheels or paddles, the steam engine could provide motion power for transportation.

There were many experimental forms of the steamship before the American Robert Fulton invented the first one used for regular passenger services in 1807. His design involved setting up a paddle wheel on each side of the hull. Since steam power was much more reliable than wind, sailboats began to be replaced by sailless steamships. This process accelerated in the second half of the nineteenth century when iron and steel were widely employed in ship construction. By 1838 the paddle wheels began to give way to the fan-shaped screw propeller which enabled faster and more stable propulsion. In the early twentieth century, the marine diesel engine superseded the steam engine, and almost all ships are now powered by petroleum.

To save on the cost of transportation, steam and diesel ships were built far bigger than the sailboats in the past, using stronger building materials and more powerful engines. The container ship was introduced in 1955 and has become the ubiquitous form of transportation, particularly because these standardized units can be carried by both inland and maritime transportation systems with great efficiency. The growing size of all types of sea vessels, including containerships, often necessitated the use of entirely new port sites outside the central business areas, where most ports had originally emerged. Many ports established on shallow sites during the age of sail were no longer adequate for the new giants. This changed the dynamics of ports toward sites that could handle larger cargo levels, and also transformed urban planning. Some traditional port cities in the past lost their importance; for
TRANSPORTATION HISTORY

example, Liverpool, once a major English port, declined with containerization.

The growing reliance on petroleum, particularly after World War II, created a huge demand for fossil fuels. The application of economies of scale in maritime shipping led to the introduction of large tanker ships carrying oil from producing countries to large consumption markets. The guarantee of sea routes for oil transportation became a national security priority of all industrial powers, and also a major source of conflict during and after the Cold War.

The first steam-powered locomotive was introduced in 1801 for road use, but turned out to be impractical as the roads at the time were too rugged for such machines, even though road conditions were improving in Great Britain through the construction of toll roads and the use of the macadamized process (1816) whereby roads were paved with single-sized aggregate layers of small stones, with a coating of binder as a cementing agent. The first practical steam-powered locomotive was invented in 1814 by George Stephenson who rested the engine on railways. The rail had been employed for moving coal from mines since the seventeenth century and was originally pulled by horses. Early forms of railways were made out of wood or stone, but from the 1760s they were replaced by cast iron plates. The low friction between steel wheels and rails enabled the mechanically powered locomotive to become efficient means for transporting people and freight at high speed and low cost.

Railway systems were soon established in many countries, and subsequently became the primary form of mechanized land transportation. Bridges and tunnels were constructed to facilitate the smooth operation of the system, and segments linking port cities, industrial regions, and resource extraction areas completely changed the landscape alongside them. Rail became a critical factor shaping urban systems and the urban spatial structure. For example, satellite towns linked by the railway could be built around major metropolitan areas, and the urban area thus expanded in line with the tracks.

The coal-fueled steam engine was first replaced by the diesel engine, which in turn was taken over by the electric engine on several rail segments, particularly on passenger trains. In environmental terms, the electric engine is much cleaner and can be used underground, which is better suited to subway systems supporting urban mobility in industrial countries. High-speed rail systems traveling at over 250 km per hour were first developed in Japan and Europe. Today there are huge high-speed rail construction projects in China and other emerging economies. In spite of its efficiency, the dominance of rail transportation was challenged by the automobile and the airplane in the second half of the twentieth century.

The first modern automobile was designed by the German inventor Karl Benz in 1885. Cars soon replaced horse carriages as a dominant form of individual transportation, especially after Henry Ford was able to mass-produce cars using the assembly line system in 1908, which made cars more affordable. After World War II the automobile became more affordable by the middle class and even the working class in some developed countries. In 2010 the number of automobiles had risen to over 1 billion; this figure is still increasing thanks to the rapid development of newly industrialized countries such as India. Highways and other roads were built to increase the connectivity and accessibility of nations, so that in theory almost any location can be reached by motor vehicle. The heavy consumption of petroleum by the internal combustion engine, however, constitutes a serious threat to the environment because of the air pollution and carbon emissions caused by burning fossil fuels, which are nonrenewable.
and will probably be exhausted within a few decades.

The airplane is perhaps the most innovative means of transportation. While only a dream before the Industrial Revolution, unlike the train, car, or steamboat, the airplane was not developed from any nonmachine-powered predecessors. Although hot air balloons existed in the eighteenth century, the diesel-powered plane was invented by the Wright Brothers only in 1903. Commercial airborne passenger travel became popular after World War II, with technical improvements making it faster, safer, and more cost-effective. In the 1950s air travel overtook the ocean liner as the dominant mode of long-distance passenger transportation.

**Conclusion**

Transportation history underlines the impacts of transport innovation on economic and social activities, notably their scale and frequency. It is undoubtedly a critical factor influencing all aspects of civilization past and present. Transportation development since the Industrial Revolution have changed the relative size of the world, and also imposed striking impacts on human social and economic activities. Our modern ways of residing, eating and even clothing are results of the new modes of transportation. Indeed, in the past two centuries transportation has turned out to be faster, safer, much cheaper and more comfortable, but nowadays pollution and sustainability have emerged as relevant serious concerns. Future improvement in transport technology will likely focus more on environmental issues: the greener the better.

**SEE ALSO:** Containerization; Empire; Maritime transport; Population mobility and regional development; Ports; Rail transport: freight; Rail transport: passenger; Sustainable transport; Transport and development; Transport geography; Transport networks; Transport technology

**Further reading**


Chen Zhongdan. 2005. *Tushuo jiaotong fazhan shi* [A Pictorial History of Transportation Development]. Hong Kong: JPC.


Transportation planning

Eric A. Morris
Clemson University, USA

What is transportation planning?

Transportation planning manages the finance, politics, economics, design, construction, routing, and operation of transportation systems. Although there is considerable private sector involvement in transportation, typically “transportation planning” refers to the public provision of goods such as streets and highways, mass transit like buses and subways, maritime facilities such as ports, and air travel facilities such as airports and air traffic control systems. There is considerable overlap between the fields of transportation planning and transportation engineering; for example, both engineers and planners study travel behavior in the present and use it to predict travel patterns in the future. However, in general, transportation engineers are more focused on the design and construction of physical facilities, and on studying the mechanics of traffic flow. Planners focus more on issues like finance and economics, policy and politics, administration, public processes, transportation’s effects on the environment, and the two-way interplay between transportation and land use.

The early historical development of transportation planning

Transportation planning is not new. The city of Mohenjo-Daro in modern Pakistan, constructed in about 2600 BCE, had a gridded street pattern that is perhaps the first recorded instance of the conscious planning of a transportation system. However, a formal science of transportation planning would not emerge until the twentieth century (Brown 2006). At the beginning of this period, planning was dominated by “City Beautiful” principles. This movement was inspired by work such as the plan for Washington, DC, in the late eighteenth century, and particularly the plan for the reconstruction of Paris beginning in 1853, in which much of the dense, chaotic warren of streets that organically developed from the Middle Ages onward would be replaced by the broad, grand boulevards studded with monumental architecture that characterize Paris today. In addition to Washington, City Beautiful plans were implemented in Chicago, Cleveland, Detroit, Canberra, Melbourne, and many other cities.
TRANSPORTATION PLANNING

Transportation planning’s response to the automobile

However, growing motor vehicle traffic congestion ultimately led planners to shift their focus from aesthetics to improving the flow of traffic. Transportation planning transitioned to become a more “scientific” field, which used empirical data on travel demand and the movement of traffic. Early developments included innovations such as the stop sign, yield sign, pedestrian island, one-way street, pedestrian crosswalk, traffic circle, municipal traffic code, and, ultimately, the stop light. Next came the “Major Traffic Street Plans” that flourished in the 1920s. These plans focused on segregating different kinds of traffic (i.e., separating pedestrians, autos, and streetcars), creating street hierarchies that channeled auto traffic off of minor residential streets and onto major thoroughfares, and street widening, straightening, rationalization, and signalization.

The ultimate expression of this focus was the development of the freeway. It featured limited entrance and exit points, the elimination of at-grade intersections using over- and underpasses, and multiple wide lanes with gentle slopes and curves. Key milestones were the Italian Autostrada, the backbone of which was completed in 1926; Germany’s Autobahn, which saw its basic network built under the Nazis during the 1930s; and the American Interstate Highway System, on which large-scale construction began during the mid-1950s.

Transportation planning and the development of mass transit

Mass transit was introduced beginning with omnibuses, large horse-drawn carriages, in the 1820s. The next major milestone, which increased vehicle speed, safety, and capacity, occurred in the mid-1850s when the carriages, now called “horsecars,” were moved onto rail track embedded in the streets. In the 1890s, horses were replaced with electric propulsion (the electric streetcar), and, in the largest cities, “metros” were constructed with trains running in tunnels or on elevated track.

Transit systems in the United States were generally overseen, licensed, and sometimes financed by municipal governments, but they were typically owned and operated by private firms. Since their fares in the late nineteenth century were often beyond the means of the working classes and poor, transit was in the beginning a private enterprise patronized by the well-off as opposed to a public enterprise disproportionately patronized by the poor, as it is today. Although through World War II streetcar and metro systems thrived, competition from the auto, customer dissatisfaction, and the poor economic foundation of the industry resulted in hard times for mass transit in the postwar period; this was particularly the case in the United States, where mass motorization occurred several decades before it did in other nations. As a result, by the 1960s mass transit was suffering a severe loss of patronage and financial woes that threatened the survival of the industry. Its survival was guaranteed only by government ownership, which became widespread in the United States in the second half of the twentieth century.

The revolution in transportation planning

The government takeover of transit was an effect of a transformation in planning principles. This began in the late 1960s and early 1970s in part
due to widespread public opposition to the
disruption caused by the construction of urban
freeways, a growing environmental movement,
oil price shocks due to supply disruptions in
1973 and 1979, and the desire to arrest inner
city decline and check mass suburbanization
and sprawl. Moreover, at this time it became
increasingly clear that transportation planning
was prone to mistakes and improper strategies
that post ante revealed its fundamental flaws;
trying to control and predict traffic flows and
the nature of transport development was, and
remains, challenging, as was evidenced by the
ever-rising surge of traffic congestion which
seemed to defy all efforts to control it.

As a result of widespread dissatisfaction with
the state of transportation planning, a paradigm
shift took place in the field. Meyer (2000)
outlines the key points of this transformation in
transportation planning principles.

Transportation planning shifts from a
quantitative to a more qualitative focus

First, there has been waxing emphasis on qual-
itative planning principles as opposed to purely
quantitative ones. As in the past, transportation
planners today perform quantitative analysis of
present and future conditions. They conduct
large-scale surveys where respondents record
their travel behavior to inform complex fore-
casting efforts. The traditional method of such
forecasting is the “four-step” model. The first
step involves breaking a region into zones, and
calculating the number of trips each zone will
produce based on characteristics such as popula-
tion and the level of auto ownership. The second
step involves determining where the trips from
each zone will go, using the assumption that
each zone will tend to send more trips to zones
that are larger, denser, and closer. The third
step involves using “discrete choice” models to
predict the mode (e.g., auto, transit, walking)
that will be used for each trip. The fourth is
loading those trips onto a representation of the
transportation network to determine which
specific routes the trips will take.

The four-step model has been widely used for
decades, but it has long been known to have
deficiencies due to its abstraction, imprecision,
and inability to mimic many potential changes
in the transportation system. “Disaggregate
models,” which create idealized households
based on travel demand data, and attempt to
simulate these households’ actual travel patterns,
have long been recognized to be conceptually
superior. Disaggregate modeling is gradually
replacing four-step modeling, but this process
has been slow due to the appeal of the relative
simplicity of the four-step model, its ease of use,
and a degree of inertia among planning agencies.

However, given the complexity of the prob-
lem, it remains clear that even the best-informed
modeling is prone to error, and even willful
misrepresentation. Megaprojects such as urban
freeways and metro systems routinely and even
systematically come in over budget (Flyvbjerg,
Holm, and Buhl, 2003), and, in the case of
metro systems, nearly always have traffic levels
far below initial projections (Flyvbjerg, Holm,
and Buhl, 2005).

In response to this, and thanks to the recog-
nition of transportation’s broader impacts on
society, other “qualitative” concerns have begun
to share the stage with quantitative ones in recent
decades. For example, while the sole purpose
of modeling was once thought to be reducing
travelers’ time and monetary costs, planners
now consider a wider set of concerns such as
neighborhood cohesion and transportation’s
effects on social life, economic development,
public health, and so on. This change is generally
hailed by transportation thinkers, public officials,
TRANSPORTATION PLANNING

and the general public, though some question how concepts as vague as “livability,” which are difficult to define, let alone quantify, can reliably guide policy.

Planning focuses on alternative modes

Second, there has been a shift from an overwhelming focus on auto travel to planning for alternative modes such as mass transit, bicycles, and walking. For both financial and political reasons, the era of mass freeway building, at least in the developed world, has come to an end. At the same time, there has been a dramatic expansion of mass transportation, with the number of metro systems worldwide rising from 48 in 1970 to over 190 today, with 38 more currently under construction. The major reason for this has been the extension of government ownership of, and public subsidy for, mass transit. Virtually all metro systems in the world today are publicly owned and operated. This has dramatically increased transportation planners’ responsibility, resources, and influence, and enhanced their potential ability to produce more positive outcomes in terms of congestion and travel time, environmental impacts, quality of life, and much else. It has also, however, brought challenges to the field, as planners struggle to balance political, administrative, economic, financial, technological, engineering, and operational demands in running complex transportation systems. Moreover, efforts to persuade the public to substitute transit travel for auto travel have had mixed success.

In addition, transportation planners today work closely with land-use planners to create walkable and bikeable environments with higher densities, a mix of land uses (where a half century ago strict separation of land uses was considered the ideal), good pedestrian infrastructure, and appealing aesthetics. It is hoped that these areas will reduce auto trips, although how much they actually do so, and why, is a subject that is still being debated.

Transportation planning becomes integrated with land-use planning

Third, there has been increasing focus on using transportation to shape land use. In the past, city form was considered largely fixed, and transportation planners sought simply to accommodate anticipated future development. Today, there is a richer understanding of the symbiotic development of transportation systems and land uses. For example, cities such as New York and London, whose formative years were in the eras when walking and mass transit were dominant, have developed high densities and lively centers, while cities that developed when the automobile was the dominant mode, such as Atlanta, Phoenix, or Houston, grew with low densities and high levels of dispersion.

In light of this, transportation planning and land-use planning are now much more tightly integrated. Transportation and land-use models are now used in an iterative fashion, with transportation patterns predicting land-use developments, and land-use developments conversely predicting future use of the transportation system. Moreover, transportation and land-use planning are now more tightly coordinated, for example in the case of “transit-oriented developments” that aim to create dense, walkable communities around transit stops. On a larger scale, decision-makers increasingly build – or choose not to build – transportation facilities to channel growth into desired areas. For example, planners have increasingly eschewed building highways to outlying suburban and exurban areas in an effort to arrest urban sprawl.
Despite near universal acceptance that land use influences travel (Ewing and Cervero 2010), and that transportation systems shape land use (Giuliano 2004), the empirical evidence suggests that these links depend on context and have limits. For example, there is good evidence that new transit investments can powerfully stimulate growth in healthy central business districts with limited parking. However, to this point mass transit has not consistently shown the ability to reenergize blighted areas or cause new, denser development in low-density, outlying areas.

Transportation planners emphasize shaping travel demand

Fourth, planners now focus less on accommodating travel and more on managing it. Through the first two-thirds of the twentieth century, planning operated according to the “predict and provide” principle, presuming growing auto travel as a given and seeing the role of transportation planning as accommodating this growth. Today, transportation planning also emphasizes managing, and in many cases discouraging, auto travel. Europe has been in the lead in this campaign, with aggressive steps to heavily tax fuel and car purchases; levy tolls on autos in congested city centers; restrict parking, particularly in city centers; limit the construction of new highways; and so on. Transportation demand management programs, which attempt to change behavior by, for example, offering incentives to employees to share rides, take transit, walk, or telecommute, have also been tried, though with mixed success. At the same time, governments worldwide heavily subsidize public transit travel to promote it as an alternative to driving (as well as a means of promoting mobility for those unable to own and operate autos).

Transportation planning focuses on environmental impacts

Fifth, there is now intense interest in reducing transportation’s harmful effects on the environment. Transportation planners have always been cognizant of these, going back to the late nineteenth century when a pollution crisis developed due to reliance on horses in urban areas. However, a relaxed attitude toward the environment, plus ignorance about many of the detrimental impacts from emissions from motor vehicles, led to severe auto-generated pollution problems as early as the mid-1950s, when smog blanketed cities such as Los Angeles. These problems began to be addressed in the 1960s and especially in the 1970s. In the United States, the Clean Air Act of 1970 set in motion a chain of regulations that set limits on major auto-generated pollutants and penalized metropolitan areas for failing to meet standards. America’s Corporate Average Fuel Economy (CAFE) standards dramatically improved (and continue to improve) fuel economy and reduce emissions. Moreover, efforts to promote alternative travel modes such as bicycling, transit, and walking are often touted on the basis of their environmental benefits. By and large, tremendous progress has been made in reducing air pollution from autos and trucks, and smog levels have plummeted, at least in developed nations that have adopted high standards and are not experiencing rapid growth in vehicle use. However, it is unlikely that the question of transportation and the environment will ever be “solved,” if for no other reason than that past history shows that expanding knowledge of the environment and human health is perpetually resulting in the discovery of new hazards, such as ultrafine particulate matter from diesel engines and greenhouse gas emissions.
TRANSPORTATION PLANNING

Transportation planning exploits technology

Sixth, transportation planning increasingly focuses on developing and deploying technology to solve problems such as fuel consumption and emissions. Thanks to the CAFE standards, a new auto in the year 2025 will average roughly four times the mileage per gallon of its counterpart in the early 1970s. This is primarily being done with improvements in car body design, transmissions, tires, and the internal combustion engine; many believe that continued improvement in the near future will be dramatic thanks to the replacement of gasoline by electric batteries, natural gas, cellulosic ethanol, or hydrogen fuel cells. Since greenhouse gas emissions are proportional to fossil fuel use, all of these developments may ameliorate, though not solve, the problem of transportation’s effects on the climate. Also, new regulations are considerably reducing the emission of particulate matter from diesel trucks. Catalytic converters and lead-free gasoline have improved air quality to a great degree. Car safety has vastly improved, in part thanks to new technology like antilock brakes, airbags, stability control, rear cameras, and so on. Finally, many now believe that autos that drive themselves will ultimately have dramatic positive impacts on congestion and travel time, fuel consumption, crashes, air pollution, urban sprawl, and much else. It should be noted that these changes have not come about through market forces alone; for example, the evidence shows that automakers and their customers would not have sought out the current high levels of fuel economy without government regulation and the transportation planners who help craft it.

Despite considerable successes, however, the question of whether technology and the regulations that promote it will be sufficient for solving the problem of transportation’s negative impacts is a subject for debate. While one stream of thought is that the most severe problems caused by the auto can be addressed by improving it, another school of thought is that the auto is inherently deeply flawed, and that transportation planning should focus on replacing it with alternate forms of mobility and developing land-use patterns that reduce the need for auto travel. Undoubtedly, transportation planners in the future will reach some middle ground, depending on the context in which they operate.

The transportation planning process embraces public participation

Seventh, public participation has become a key element in the transportation planning process. At the time of the construction of the American Interstate Highway System, government had almost limitless power to displace homes and businesses, with compensation but with little possibility of appeal. Without these powers this massive system could never have been built, but in time citizens bridled as neighborhoods were riven by intrusive new freeways. In response to this, a “freeway revolt” movement began, setting in train a trend toward greater public input into the planning process. Today, planners in the developed world conduct extensive public outreach, seeking input on potential policies. This has great advantages in that it can prevent arbitrary, unjust, and counterproductive action on the part of planners. However, some maintain that not-in-my-backyard sentiment, ingrained aversion to change, and the ability of small numbers of passionate people to sway the process, coupled with the fact that most potential beneficiaries of transportation projects are too uninformed or apathetic to lobby, has sapped planners’ ability to develop needed transportation interventions. In short, some feel that the price
of checking overweening government may have been the creation of government paralysis. Still, few would prefer to return to an era where transportation planners ran roughshod over residents. In addition to working with the public, transportation planning also involves cooperation between numerous levels of government. The variety of arrangements for political control over, and funding of, transportation services and infrastructure varies considerably across the world, but typically it is a joint effort of national, state/provincial, regional, and local governments. As such, the transportation planning and finance system is often of great complexity, necessitating considerable political coordination and the navigation of daunting bureaucratic obstacles.

The future of transportation planning

Problems of political will can be simpler in developing nations, particularly those with less public input into the decision-making process, and this is where the demand for new transportation infrastructure is greatest. It remains to be seen what choices nations such as China and India will make. On the one hand, these countries have the opportunity to learn from the mistakes that developed nations made during their historical evolution, and they can adopt the latest technology and planning principles. On the other hand, the historical record shows that with rising wealth comes dramatic public demand for automobiles, highways, and suburban development, and in addition it is far from clear that transportation planning even in the developed world has “solved” many of the key problems. Given the developing world’s huge population and rapidly rising prosperity, developments in transportation planning there promise to have dramatic repercussions not just for those nations but for the world as a whole.

SEE ALSO: Air transport; Ports; Sustainable transport; Transport and development; Transport networks; Transport policy; Transport technology; Transportation history; Transportation and land use; Tropical geography; Urban transit

References


Further reading

Travel geographies

Perla Zusman
Hortensia Castro
University of Buenos Aires, Argentina

Traveling may be defined as the movement of people from one place to another, while the experience of travel can involve a large range of voluntary and involuntary social practices, such as exile, pilgrimage, scientific exploration, and tourism. This movement across space involves a transformation of one’s subjectivity, one’s vision of the traveled to and visited places and societies, and one’s own place of origin, which, on return, may be perceived or experienced in a different way. The interaction between individuals and places new to them opens the door to the development of new ways of looking at and learning from places. The idea of traveling is also usually reserved for engaging in journeys that could shape an alien gaze; in this connection, the displacement that can occur in traveling involves an experience of discovery and invention of oneself and others (individuals, cultures, places). Within this framework, travel geographies can be understood as the study of the different practices connected with traveling and both the imaginary and the material geographies that are created through the displacement that travel can engender.

Traveling also creates travelers. Looking at how a relationship with a visited place is established, Todorov (2004/1989) identified a collection of types of travelers who represent variants of universal types. Among them he distinguishes: the assimilator, who wishes to change others so that they become similar to her/him; the tourist, a visitor usually in a hurry who prefers to engage with sites as opposed to human beings; the impressionist, who has more time than the tourist, and who extends her/his horizon to human beings and takes home not only photographic or verbal clichés, but also her/his own writings, sketches, or paintings; and the outsider, who makes a permanent comparison between her/his place of destination and her/his own country, and who may establish either a detached or an attached relationship with others.

Travel narratives in geography

Travel narratives, to the extent that they record and recreate a displacement experience, have historically been considered within academic geography as a source of objective information offering “true” empirical knowledge of visited places, especially in the case of texts generated by explorers, naturalists, and scientists. In the 1990s, with the influence of poststructuralism and the “cultural turn,” traveling and travel narratives started to be studied as a cultural product, contingent on historical and geographical factors, as both an expression and a means of representation of desires and fantasies. Critical approaches to travel entailed, in turn, an increase in the type of narratives under study. For example, in addition to canonical texts, such as those by the explorer Alexander von Humboldt (1769–1859) and the naturalist Charles Darwin (1809–1882), which express a relationship between traveling and scientific knowledge, there has also been a focus on
impressionistic or biographical narratives, such as those by the writers François de Chateaubriand (1768–1848) and Bruce Chatwin (1940–1989), and on narratives by women such as Mary Kingsley (1862–1900) (Blunt 1994). One focus in critical travel studies is the text itself, in terms of the forms of representation in narratives – a frequent object of interest being the narratives of Enlightenment naturalists: the knowledge strategies at stake, and the form and style in which these narratives were laid out. Conversely, other studies have turned their attention to the relationship between text and context, and delved into travel narratives as an active element of the production of knowledge about places.

Studies of travel narratives initially situated journeys in discussions that considered traveling as an individual practice, undertaken by “romantic travelers” whose narratives focused on impressions, desires, and inner needs, as opposed to traveling as a group (whether political, economic, cultural, and/or scientific) or on institutionalized ventures. For example, the early nineteenth-century exploration undertaken by Humboldt throughout South America, is frequently quoted as an example of a spirit of individual initiative, while his links to various institutional networks are ignored. And yet, in addition to being authorized by Charles V of Spain to travel across South American territories, Humboldt capitalized on the outcome of his trip by becoming a firm member and an important participant in French scientific and political circles.

Since the 1970s, travel literature has entered the fields of study of postcolonial geographies (in the anglophone world) and territorial formation (in Latin America), as well as of historical geography in general, enlarging the horizons of geographical knowledge. Accordingly, the political meanings attached to travel narratives became a focus of academic attention, particularly their role in the construction and reproduction of situations of domination. In examining political meanings, some scholars have focused on the deconstruction of travel narratives by naturalists and scientists in order to show that, under a “strategy of innocence” (Pratt 1992), such narratives could harbor political implications, which could in turn lend support to the creation of a consensus on imperial actions by Latin American nation-states against indigenous peoples. Literary critique, and in particular the texts by Said, Orientalism (1978) and Culture and Imperialism (1993), have contributed to the development of this perspective by proposing that travel narratives, in addition to texts by novelists and pictures or paintings from outside the European milieu, produce geographical representations that have supported the domination of colonized peoples. Said (1993) has shown how travel narratives have rewritten the history and geography of native populations, (re)constructing them from a perspective based on exoticism and curiosity in order to hide the most perversive effects of imperialism, namely oppression and violence.

Travel narratives and geographical imaginaries

A key topic dealt with by such studies has been that of the geographical imaginaries of colonized territories, in which territories were the other of imperial centers, presented as places of insecurity and disorder as opposed to those exemplifying the ideals of security and social and aesthetic order (Serjé 2005). For example, the idea of tropicality contributed to shaping the equatorial zones in America, Africa, and Oceania as environments that were supplementary to Europe and available to meet its needs and desires. They were represented as benign and prodigal, exuberant, and even paradisiacal places
in the narratives by Christopher Columbus (ca. 1451–1506), Louis Antoine de Bougainville (1729–1811), James Cook (1728–1779), and Humboldt, or as sick and hazardous places by Joseph Conrad (1854–1924) and Daniel Defoe (ca. 1660–1731). In Latin America, geographical imaginaries of the desierto (desert; a term used for areas with physically heterogeneous conditions such as Patagonia, or even fertile regions such as the Pampa) and the sertão (the Brazilian backlands) described these spaces as only inhabited by indigenous peoples with a social and economic system different from those of the West. Both of these imaginaries depicted these areas as “demographic vacuums” or “unpopulated lands” in an attempt to endorse their appropriation. In particular, throughout the process of the constitution of the nation-states of South America toward the end of the nineteenth and the beginning of the twentieth centuries, travel narratives shaped the Andes mountain range into a wall or barrier where it had traditionally been a space heavily crossed by exiles, guides, and shepherds (Hevilla 2007).

Imperial and nation-state networks could also give rise to dissimilar geographical imaginations for the same time and place. For example, travelers such as Ludwig Brackebusch (1849–1906), Eduardo Holmberg Jr (1852–1937), and Juan B. Ambrosetti (1865–1917), contracted by the Argentine government between the end of the nineteenth and the beginning of the twentieth century to survey the highlands of northwest Argentina, described in their texts the mineral wealth and potential oases in these areas. Other narratives, such as those by the American explorer Isaiah Bowman (1878–1950), who traveled across the same places, highlighted the harsh limitations imposed by extreme nature (particularly water scarcity) on the use of the environment and the population of these areas. Although it was such Western geographical imaginaries that became dominant, narratives and cartographies contained vestiges of representations from native populations. Such is the case in relation to indigenous toponomies, describing the perceptions and valorizations of native populations concerning their surrounding environment, the survival of which can be considered a strategy of resistance against dominant geographical imaginations.

Some authors have studied the ways in which local elites, seeking independence from the colonial power, gave new meanings to the imaginations of imperial travel narratives; however, this resignification did not imply a complete break with the European imagination. Several Spanish American authors writing from the end of the nineteenth century, such as Domingo Faustino Sarmiento (1881–1888) and Andrés Bello (1781–1865), used and adapted narratives from European travelers to advance claims for local autonomy that largely focused on the diversity and exuberance of South American “nature.” Nevertheless, their writings did not imply a rejection of the Enlightenment values and ideas of civilization that these elites shared with their European counterparts.

Other studies have shown how the ideas developed and disseminated through travel narratives produce and reproduce materialities. Such studies have focused on an analysis of the physical characteristics of the landscapes visited by travelers, as well as on their material forms of representation. Additionally, they have recognized the active role played by the ideas communicated through narratives on the transformation of places, as in the case of the export of quinine plantations to tropical India whose properties had been described by La Condamine (1701–1774) in his narratives on South America, or in the production of a “visibility space” around the Nile after the construction of platforms to secure panoramic views for
TRAVEL GEOGRAPHIES

Tourist consumption, which, based on romantic narratives, turned “an ancient Egypt” into “an exotic Orient” (Gregory 1999, 146).

Travel narratives and intersectionality

Geographical analyses of journeys and travel literature since the 1990s have also assumed that class, gender, and ethnic differences influence the ways in which a journey is envisaged, the type of routes followed, and the form of how a relationship with “others” is established, thus contributing to identity construction processes.

Concerning displacement as a mark of class distinction, research has been conducted on the Grand Tour, the trip undertaken by young European elites during the eighteenth and nineteenth centuries across ancient and Renaissance sites, including Greece, the south of the Italian peninsula, and also to South America (as in the case of Evelyn Waugh). Such journeys were conceived and implemented as an educational tool that reaffirmed the learned status of elite individuals, who went on to produce canonical pictorial representations of Mediterranean landscapes, particularly from Italy, such as the Coliseum, the Tivoli waterfalls, and Mount Vesuvius.

Literature produced by women travelers and their associated geographies awakened the interest of feminist geographers. In general, the emancipatory character of trips is recognized, even though, in some cases, an ambivalence is observed in connection with colonial discourses, which reveals the limits of emancipation. For example, the travel narratives of her trips through Argentina by the journalist Ada Elflein (1880–1919) in La Prensa – one of the most prominent newspapers in Argentina in the first decades of the twentieth century – invited women to abandon their homes and embark on adventure. The household space, however, was replaced in her narrative by the homeland space, a milieu containing the sense of safety similar to that inspired by home, to be praised only through direct contact with its breathtaking landscapes (Servelli 2014).

Globalization and less expensive travel costs have awakened a new interest in traveling and travel narratives. Today, many male and female travelers publish accounts of their experiences in journals, newspapers, or on the Internet, using canonical narratives to define their routes and to recreate their geographical imaginaries (or call them into question). Additionally, some intellectuals and scientists have recreated the experiences of past explorations to revisit sensations and imaginations. For example, the Paraná Ra’anga philosophic trip was organized in South America from March to April, 2010, by Graciela Silvestri, a landscape specialist, and sponsored by the Spanish International Cooperation Agency network of cultural centers. The term “philosophic trip” mirrors the name used in the age of Enlightenment to refer to travels designed to explore and “discover” the economic potential of colonial territories. Based on a river journey, the expedition attempted to follow the same route taken by the Spaniard Juan de Ayolas (1493/1510–1538) over the de la Plata, Paraná, and Paraguay rivers from 1536 to 1537, as described by the writer Ulrico Schmidl (1510–ca. 1580). A group of scientists, writers, and artists made this journey in order to reconsider the common “figures” (“figure” referring to one of the meanings of the Guarani word ra’anga (Silvestri 2011), such as external configurations or the appearances of things, the construction of which is shaped by anything from technical skills to feelings) representing the Paraná River as well as to stimulate discussions on current topics and problems concerning this river. From on the interactions that developed during this journey, knowledge exchanges were established between the different
actors living on the shore of the Paraná River and the members of the expedition.

There remains a need to deepen the search for and the study of narratives by subaltern populations that were part of the colonial encounter, whether these be represented by native populations or the people who accompanied travelers as guides, servants, or interpreters. The absence of these voices is likely to be explained by the small number of written narratives they produced, although these could be substituted for by a compilation of the oral testimonies of their descendants.

Images and discourses about places, originating in trips made during the eighteenth and nineteenth centuries, remain current today and are recreated by narratives of the present. Is it possible to overcome the exoticism embodied in these imaginations, which are so distant from current realities? Maybe a different kind of trip and traveler profile will unfold that can explore a horizontal and nonhierarchical relationship with visited people, which can produce a narrative reflecting a negotiation between the images that insiders and outsiders have of places.

SEE ALSO: Cultural geography; Historical geography; Orientalism/Occidentalism; Postcolonial geographies; Text and intertextuality; Tourism

References


Further reading

An ecotone refers to a boundary or transition zone that is bracketed by more homogeneous vegetation types. Ecotones are created by environmental gradients that ultimately govern the physiological and reproductive range limits of species from adjacent ecosystems. Given the relatively high levels of stress imposed on species growing at or near the edge of what may be referred to as their bioclimatic envelope, the dynamics of ecotone boundaries are widely considered to be sensitive proxies for climate variability. The width of these biogeographical transition zones normally corresponds to how steep the local environmental gradients are, with more narrow ecotones approaching an abrupt line produced in environments with particularly sharp gradients and broader ones indicating a more gradual influence over space. In either case, however, the clear visual contrast provided by forest trees abutting communities of low-stature vegetation makes treeline ecotones one of the most easily recognized ecotones in existence.

Yet, considering the varying importance of environmental gradients in creating and governing tree regeneration and growth within these conspicuous vegetation boundaries or transition zones, this entry will be separated into two sections that deal exclusively with (1) arctic and alpine treeline and (2) savanna ecotones. These two categories of treeline environments occur in distinct biogeographic settings worldwide, and thus an individual rather than a group perspective can help articulate the processes responsible for maintaining their dynamic existence.

### Arctic and alpine treeline ecotones

#### Terminology and distribution

The arctic treeline (also referred to as the subarctic or polar treeline) is a very broad ecotone that forms a high-latitude circumpolar belt around the Northern Hemisphere between approximately 60°N and 70°N. It forms the low-elevation transition zone between the boreal forest biome to the south and arctic tundra biome to the north. The width of the arctic treeline ecotone demonstrates remarkable variability across relatively level terrain, as it ranges from a few kilometers to over 100 km. Although moderate relief can play a role in how wide these ecotones are, the fundamental reason for such expansive transition zones relates to the gradual decrease in temperature along a latitudinal gradient, particularly compared to the relatively sharp decreases in temperature found along steep elevational gradients. In addition, the southernmost boundary of continuous permafrost exerts a strong influence on the spatial extent of arctic treeline ecotones.

Alpine treeline (also referred to as upper treeline) ecotones form a high-elevation transition zone from subalpine forest to alpine tundra. Thus, arctic treeline ecotones form the northernmost extent of forests globally while alpine...
TREELINE ECOTONES

Treeline represents the highest elevational extent. Given that the main prerequisite for alpine treeline is a mountain that attains an altitude beyond which trees can successfully establish and reproduce, the geographic distribution is widespread, ranging in latitude from the subarctic to the tropics. Alpine treeline ecotones can vary in width but are relatively abrupt compared to the arctic treeline, especially on the steep mountain slopes characteristic of the European Alps and parts of the Rocky Mountains in the United States. In areas dominated by more gentle topography, the alpine treeline ecotone can span hundreds of meters and, generally speaking, limitations from permafrost are minimal and are confined to subarctic regions.

Terminology plays an important role in treeline ecotone research, although a cross-comparison of studies is often complicated by the fact that researchers do not adhere to a universally-accepted terminology in defining the spatial dimensions of a particular treeline environment. For instance, the term “timberline” normally refers to the northernmost or uppermost elevational extent of continuous forest where the treeline ecotone begins. Depending on how abrupt the transition from forest to tundra is and what species are present, this can be a challenging baseline to delineate in the field. This is especially true when the ecotone is composed of more open-grown species like pine (Pinus sp.) or larch (Larix sp.), compared to relatively dense stands of more shade-tolerant spruce (Picea sp.) and fir (Abies sp.). The “treeline” or “tree limit” then refers to the furthest extent of trees with an upright growth form, with the treeline ecotone existing between these two delineations. Another term used in treeline studies includes “tree species line,” which is intended to convey the existence of both krummholz and seedlings that may exist beyond the limit of upright trees. Krummholz is a German word and is used with reference to trees with a twisted or gnarled growth form that represents a phenotypic response to particularly harsh climate conditions often related to local wind and snow regimes. Distinguishing characteristics of krummholz also include a predominantly horizontal growth form resembling a large, multistemmed shrub rather than a single-stemmed upright tree.

Considered together, these terms are used to describe one of two main definitions used in treeline ecotone research. The first has roots in landscape ecological investigations of the treeline, and specifically defines the arctic and alpine treelines as the transition zone extending from closed subalpine or northern boreal forest (also called timberline or forest line) to the uppermost or northernmost extent of scattered and stunted individuals, regardless of their height. This definition of a treeline ecotone differs from the other commonly employed method where the uppermost or northernmost boundary of treeline is based on trees reaching a minimum height, such as 2 or 3 m. The rationale for confining ecotone boundaries to the elevational or latitudinal extent of trees with a given height instead of the point beyond which successful tree establishment occurs is typically based on the depth of winter snowpack in a specific environment and whether trees (seedlings and/or saplings) protrude above this protective layer or are sheltered beneath it. The basis for requiring a minimum height of 3 m, for example, also ensures the differentiation between tree crowns closely aligned with prevailing atmospheric conditions and low-stature vegetation, which are more capable of strongly modifying their microclimate. Potential arguments against using a specific height criteria when defining treeline ecotone boundaries relate to the tremendous variability in localized wind–snow interactions throughout alpine and arctic treeline landscapes, species-specific differences in vertical growth rates, and the pivotal
importance of understanding the spatiotemporal patterns of seedling establishment (e.g., growth to 3 m) within treeline ecotones. Moreover, this dichotomy in how a treeline ecotone is defined fits within a long-standing historical context, where observations and causal mechanisms were traditionally approached from one of two spatial scales: fine-scale local variability versus a more uniform broad-scale approach aimed at reaching a global perspective.

History of treeline research

Arctic and alpine treeline ecotones have captivated the research interests of geographers, ecologists, and naturalists alike for centuries. During the European Renaissance, for example, Leonardo da Vinci (1452–1519) observed that certain organisms were specific to a particular elevational belt on Monte Rosa in the Italian Alps. This rich history has in turn produced a remarkably extensive literature. Written descriptions of the alpine treeline ecotone date back to at least 1555, when Conrad Gessner (1516–1565) from Zurich published “description montis fracti” in which he attributed the elevational zonation of vegetation communities in the front ranges of the Swiss Alps to a decrease in temperature and growing season with elevation. This early period in treeline research extends through the time of the famous geographer (naturalist and explorer) Alexander von Humboldt (1769–1859). Humboldt was an unmistakably keen observer and, according to Körner (2012), he clearly viewed the alpine treeline as a life form boundary and common bioclimatological reference point, whereby the location of all other elevational vegetation belts were positioned relative to its occurrence. Moreover, Humboldt made some foundational linkages in the early nineteenth century (1805) regarding the global occurrence of the treeline and not only how the elevational position of the ecotone varied by latitude but how different vegetation belts were caused by shifting isotherms as elevation increases along a mountain slope. These insights strongly influenced his ideas of global plant formations and were instrumental in the eventual formation of biogeography as a viable subdiscipline of geography. Given that Humboldt’s observations remain valid today and because his understanding of alpine treeline linked observation with causality, he is perhaps the most distinguished of early treeline researchers. In summary, although this early period of treeline ecotone research lasted for over three centuries, its legacy is largely defined by observational insights derived from regional and local geographic studies.

Systematic treeline research began seeking to build on earlier observational work approximately 150 years ago. This transition period can be thought of as a time when an increasing number of observational studies hinted toward heat deficiency as the primary reason for the formation of arctic and alpine treelines. Collective efforts to prove this, however, often necessarily relied on the interpolation of meteorological data from weather stations far removed from treeline environments, especially in high-latitude areas of the Arctic, which in many instances, remains problematic today. During this time, climatological-based investigations identified the coincidental overlap between the arctic treeline boundary and the 10°C isotherm of the warmest month, July. Other increasingly systematic investigations centered on the mountain-mass effect in raising the elevational extent of the alpine tree-line and, correspondingly, the upper limit of the snow line, agriculture, and human settlements, along with early attempts to link precipitation with high-elevation treeline ecotones. Central to this transition into the more systematic modern period, however, was the work of Däniker in 1923, who was the first to pay special attention
TREELINE ECOTONES

to local ecological conditions when explaining the patterns he saw at the alpine treeline in the European Alps (Holtmeier 2009). What is perhaps most impressive about Däniker's work is that he had virtually no data, yet, through ecological reasoning based on comparison and plausibility, he reached the conclusion that a lack of sufficient heat limits tree growth at the upper treeline ecotone (Körner 2012).

Experimental approaches supplemented by the use of field instrumentation in the early 1930s marked a shift toward a more fine-scale understanding of the alpine treeline. This collective initiation of case studies identified a suite of governing mechanisms that may explain local-scale variation such as winter desiccation, but do not contribute further to a global understanding of the life form boundary that the alpine treeline represents (Körner 2012). Shortly thereafter, catastrophic avalanches in the Alps during the winters of 1951–1952 and 1953–1954 prompted the development of extensive research programs in Austria and Switzerland to assess the protective function of trees positioned within the alpine treeline ecotone which had been largely deforested for various land uses, including the creation of alpine pastures, mining, burning, salt works, and timber harvesting (Holtmeier 2009).

Another positive outcome of the avalanches was the visionary establishment of research stations in close proximity to the alpine treeline in 1953 (Obergurgl, Austria) and 1959 (near Davos, Switzerland) where experimental field studies on topoclimate, plant communities, soils, snow fungi, mycorrhiza, and species-specific reforestation efforts could be carried out and carefully observed (Holtmeier 2009). Similar field stations exist at far northern latitudes near the arctic treeline (e.g., Abisko, Sweden) that further underscore the ecological benefits of a permanent location to conduct fine-scale field experiments.

Deviations from the more fine-scale experimental approach during this period can be found in the broad-scale comparative approaches of Daubenmire and Troll who, in the opinion of Körner (2012), championed a biogeographical approach rooted in the Humboldt tradition that emphasizes the global nature of treeline ecotones. In both cases, this sort of treeline research focused on climate correlates that could explain the distribution of the alpine treeline. For example, Daubenmire (1954) meticulously plotted the elevation of the alpine treeline along the entire length of the American Cordillera and discussed the broad-scale variation within the context of temperature, wind, snow, and air mass climatology influences. Akin to previous work, his writing strongly hints at the likely overriding influence of heat in spite of the regional and local variations attributed to oceanic and wind/snow influences, respectively. Oceans influence the treeline through the trajectory of maritime air masses that lead to lower alpine treeline elevations because of cooler summer temperatures and normally shorter growing seasons because of heavier snow totals, in terms of both total accumulation and water content, especially compared to more continental settings, which have relatively dry, powdery snow. Troll (1973) was similarly placed within the context of bioclimatological studies, yet his global comparisons focused on high-elevation mountain environments in general rather than on the alpine treeline exclusively. As a result, it can be argued that his thorough consideration of periglacial processes in shaping the landscapes at or near the alpine treeline represents a precursor to more contemporary biogeomorphological studies. Either way, these broad-scale studies revived the global emphasis on treeline ecotones.

A later emphasis on more fine-scale studies accompanied the publication of Tranquillini's *Physiological Ecology of the Alpine Timberline*
(1979), which is widely considered a classic. This work represented a synthesis of Tranquillini’s experimental research on the ecophysiology of tree species growing at the alpine treeline in the Austrian Alps, in which he argued that the formation of the treeline is ultimately dependent on an increasingly unfavorable heat balance with gains in elevation or latitude. He supported this assertion with evidence of higher alpine treelines on south-facing slopes, which in the Northern Hemisphere receive more insolation than contrasting north-facing slopes, and the overall rise in treeline elevation when moving along a latitudinal gradient from the poles to the tropics. Tranquillini’s work also underscored how complex the effect of temperature is on plant physiological processes and that annual tree growth is likely limited by a minimum temperature threshold. These topics remain relevant in the literature today, including his understanding of how the dominating influence of temperature can in turn be modified by local variations in solar radiation, moisture, and wind. For those with ecophysiological interests, this book clearly reinforced the importance of studying tree growth (Körner 2012). However, the experimental work described by Tranquillini was discounted by others for the nontransferability of its results to distant treeline environments characterized by pronounced heterogeneous site conditions (Holtmeier 2009). Disagreements over the importance of Tranquillini’s pioneering ecophysiology work point to a natural divide within the treeline research community which, to a certain extent, remains evident today.

Context for modern treeline research

Global climate change and the burgeoning field of global change ecology have perhaps provided the strongest motivation for continued research efforts at both arctic and alpine treeline ecotones worldwide. This is because, in the presence of suitable substrate conditions for tree establishment, the location of arctic and alpine treeline ecotones is primarily controlled by heat deficiency. This means that the physiological processes responsible for tree growth and establishment at the uppermost or northernmost extent of the ecotone are temperature-limited and unable to occur successfully beyond a certain latitude or elevation. The onset and foreseeable continuation of warming related to global climate change is therefore expected to facilitate successful tree establishment in areas further north or upslope from where the arctic and alpine treelines currently extend. As a result, identifying the existence of and determining the causal factors for treeline advance have dominated the contemporary literature on both arctic and alpine treeline ecotones.

The premise for widespread treeline advance with warming is supported by paleoecological studies based on tree rings, radiocarbon dating of remnant wood in treeline ecotones, and plant macrofossil and pollen assemblages from lake sediment cores, which all demonstrate a relatively tight coupling between treeline extent and climate variability throughout the Holocene (ca. 10 000 years BP). During this time frame, treelines expanded beyond the present ecotone boundaries during a warming period initiated during the early Holocene (e.g., 10 000–6000 years BP in the Rocky Mountains) and returned to approximate positions comparable to the early twentieth century beginning around 3000 years BP. Although reconstructions focused on more recent changes in the elevational extent of the alpine treeline are possible using a variety of field and geospatial computer techniques, that treeline advance is confined to specific treeline types is underemphasized in the literature.
Climatic treelines essentially refer to environments where tree establishment and eventual treeline advance beyond the current ecotone boundary are unimpeded by local topography or geomorphological conditions, such as steep and rocky slopes, absence of soil development, or frequently disturbed avalanche tracks. Based on extensive field research in the Rocky Mountains in the United States, identifying mountain peaks with “climatic” treelines can be challenging because of the quantity of geologic treelines, where the uppermost extent of subalpine forest borders rocky landforms (e.g., talus slopes or sheer rock) that are incapable of supporting treeline advance, regardless of how amenable climate conditions become. Anthropogenic treelines also exist and refer to treeline ecotones impacted by human activities; these normally occur on more gentle terrain. Without knowledge of the previous land-use history of an ecotone, anthropogenic ecotones can visually resemble climatic treelines, especially on mountain peaks with a sizable expanse of palatable alpine species for grazing, as is the case in portions of the Sangre de Cristos in northern New Mexico. And, in some cases, the spatiotemporal patterns of tree establishment during the reforestation of a previously disturbed treeline resemble what one might expect from a climate-induced shift. Thus, a paradox can exist when climatic treelines are supposed to contain the most climatically sensitive trees positioned in relative disturbance-free environments, yet, because they are typically located along slopes or near peaks that are more traversable than those close to the geologic treeline, the influence of climate may be obscured by land-use legacies. Similar challenges arise at the arctic treeline when the influence of grazing reindeer and of climate on birch (Betula sp.) forests are juxtaposed. Local site histories are, therefore, important to consider when examining treeline advance, even when at climatic treelines.

Treeline advance is a term that has been linked to phenotypic and genotypic responses to climate. A phenotypic response refers to growth form changes, most notably when vertical leaders form on horizontally oriented krummholz in the presence of less severe climate conditions. This sort of treeline change or advance is most commonly reported in studies that define ecotone boundaries by minimum height, and has been a notable component of treeline change in the Scarps Mountains of Sweden and at the arctic treeline in eastern Canada. A genotypic response refers to successful tree establishment from seed beyond the current ecotone boundary and therefore typifies what is commonly thought of as treeline advance. Considering that both responses rely on favorable climate conditions, they are not mutually exclusive and may provide concurrent examples of treeline advance across a landscape or mountain slope.

The advancement of trees into tundra is often cited as evidence for the climatic sensitivity of trees within treeline ecotones. When treeline advance is considered from a regeneration standpoint, this process is contingent on several consecutive stages, beginning with the production of viable seeds. Although this ignores the likelihood of vegetative layering (e.g., root sprouting), which is common to certain treeline species such as black spruce (Picea mariana) at the arctic treeline and subalpine fir (Abies lasiocarpa) or quaking aspen (Populus tremuloides) at the alpine treeline, sizable treeline advance is contingent on the dispersal of viable seeds beyond the treeline boundary. Seed production alone is no guarantee of treeline advance without an ability to disperse effectively. Most seeds at the treeline are wind-dispersed (anemochorous), but some five-needle pines, including whitebark pine (Pinus albicaulis) in North America and stone pine (Pinus cembra) in Europe, have their highly palatable seeds cached by a specific subspecies.
of nutcrackers (*Nucifraga* sp.). Since birds cache multiple seeds in a manner that seems to facilitate establishment, these pines are often found in isolated spatial clusters well above the timberline boundary. Wind dispersal can spread seed rain up to a few hundred meters, depending on local wind regimes and topographic roughness; and, depending on the relative proximity of alpine treeline ecotones to bountiful seed sources at a lower elevation, dispersal can be more limiting to treeline advance at the arctic treeline (Holtmeier 2009).

Following successful dispersal, seeds require a suitable microenvironment for establishment, which can vary by species and in accordance with local climate and geomorphological conditions. A newly established seedling is highly susceptible to mortality and even a single year with extreme winter cold is capable of causing the dieback of seedling and sapling cohorts from several consecutive years. At the opposite end of the radiation spectrum, current trends and future projections of warming foretell the development of heat-induced moisture stress if rising temperatures are not accompanied by parallel increases in precipitation. The potential switch in dominant limiting factors, from the classic temperature-limited paradigm to one of moisture stress, has already been documented at various arctic and alpine treelines worldwide (e.g., the arctic treeline in Alaska and the alpine treeline in the Andes). Elsewhere, sharp rises in temperature triggered an abrupt regime-shift increase in tree establishment. A regime shift refers to the tipping point or measurable change stemming from an ecological threshold response, and results from external climate forcing and/or internal feedbacks more closely related to fine-scale biotic interactions. Considered together, these examples illustrate the meaningful impacts of climate variability exceeding critical threshold values and, as a result, either accelerating or retarding the pace and pattern of treeline advance.

**Considering spatial scale**

The relative role of climate in governing tree establishment is contingent on fine-scale, site-specific attributes capable of facilitating, modulating, or possibly overriding the influence of abiotic inputs. Examples of these more local-scale factors include slope aspect and corresponding soil moisture regimes, the structural form or spatial pattern of tree establishment, properties of soil and geologic substrate, geomorphic processes and landforms, herbivory, invasive forest pathogens, and land-use history. In spite of these wide-ranging factors to consider alongside climate, the expanding literature focused on studying tree establishment, and related treeline advance can be broadly classified into one of two categories: (i) climate–vegetation studies and (ii) process–pattern studies. Of course, neither approach is necessarily ideal; rather, each category often focuses on a particular scale of analysis. Combining these two categories, on the other hand, may help to form a holistic understanding of treeline ecotones that would otherwise be unattainable.

Climate–vegetation studies at the treeline have traditionally focused on broad regional scales where the influence of climate in general and temperature in particular are more easily discernible because governing factors increase in complexity when moving from broad to fine spatial scales. Conversely, process–pattern interactions have largely been restricted to local- and landscape-scale studies where the dominating influences on tree establishment are discussed from a bottom-up perspective. A central theme from this line of research that is relevant across
TREELINE ECOTONES

multiple spatial scales deals with positive feedback switches. Positive feedback in this case refers to the facilitative effects provided by the establishment and the sustained presence of a tree within the treeline ecotone, which in turn ameliorates locally harsh conditions and favors additional establishment. Krummholz patches, inanimate objects like boulders or terrace risers, and topographic depressions can also initiate positive feedback loops at both arctic and alpine treelines. Once formed, the neighboring biotic interactions of positive feedback can override the importance of climate inputs in particularly harsh treeline ecotones characterized by persistently strong winds or drought stress. This then creates a longer lag time in treeline response to climate compared to ecotones where positive feedback is not needed for treeline advance and trees are, in turn, more closely aligned with changes in the climate system.

Spatial scale is a historically relevant theme in treeline research that remains at the forefront of treeline advance studies rooted in landscape ecology or macroecology frameworks where cross-scale, process–pattern interactions are emphasized. Although considerable commonalities exist between these two frameworks, the latter can be distinguished by its stronger emphasis on broad spatial scales (e.g., regional or continental) and top–down climatic controls on process–pattern interactions operating at finer spatial scales (e.g., local). These linkages can be examined in a reciprocal manner as well, which provides the opportunity for a comprehensive understanding and synthesis through both a top–down and a bottom–up approach at spatial scales extending beyond the level of a landscape (e.g., mountain range). A macroecological framework is, therefore, useful for linking climate–pattern–feedback interactions across broad latitudinal gradients (e.g., Elliott 2011), which satisfies the reported need for treeline studies to address both regional and local-scale responses to climate variability while also permitting direct comparisons across large geographic areas.

Savanna ecotones

Tropical savanna ecotones exist in tropical and subtropical regions centered around the 30°N and 30°S latitudes (MacDonald 2003). They occupy the transition zone between wet tropical forests toward the equator and semiarid tropical steppes and deserts toward the poles. This broad ecotone is primarily located in Africa and contains the famous Serengeti Plains of Tanzania and Kenya, but can also be found in parts of South America (referred to as campos or llanos), South Asia, and Australia (referred to as brisalow). Globally, savannas occupy roughly 15 million km² or 40% of the terrestrial landmass. However, permanent land-use practices, such as grazing, are permanently removing savanna grasses and perpetuating the desertification of this biome while simultaneously facilitating the encroachment of the Sahara desert.

Mean annual temperature ranges from 20° to 30°C with an approximate 10°–15°C seasonal range between summer and winter (MacDonald 2003). Total annual precipitation varies from 50 to 150 cm given that the latitudinal position of tropical savannas generally corresponds to the position of subtropical high-pressure cells. These are warm and dry, moisture-limited environments with a winter dry season up to seven months long. The wet season is therefore short in comparison and is initiated when the migrating equatorial cyclone belt referred to as the Intertropical Convergence Zone (ITCZ) arrives and forces air aloft through surface convergence. The pronounced wet and dry seasons of the Serengeti dictate the migration routes for an
unparalleled number of mammal fauna in general and ungulates (hoofed mammals) in particular, with over 90 different species present. Rainfall patterns dictate not only animal migration routes but also areas where trees can exist.

Savanna ecotones are composed of xerophytic vegetation that has evolved to minimize water loss. Within these dry ecotones, the proportion of tree cover varies dramatically depending on moisture availability, competition with grasses, herbivory, and fire. This leads to some areas with only sporadic tree growth and woody shrubs within a broader grassland matrix. In other, more moisture-rich areas, such as riparian zones and well-drained uplands, tree density can resemble a closed canopy forest. Still elsewhere, vast expanses of treeless grassland can exist. Grasses grow quickly once the rainy season starts and can grow to 2 m in height, which shades out tree seedlings of a much smaller stature. Once trees establish, however, they develop a deep tap root to access groundwater because surface rainwater is typically consumed by grasses with more shallow root systems. Trees are characteristically flat-topped in savannas, and this expanded horizontal canopy helps reduce grass cover by shading out the grass understory, which reduces competition for water and the accumulation of highly flammable fine fuels. In Africa, trees are deciduous and shed their leaves in response to drought stress. Outside of Africa, however, trees and shrubs are primarily evergreen with sclerophyllous leaves, indicating a leathery or waxy texture to help prevent water loss through transpiration (MacDonald 2003).

Disturbance is common in savannas and tree cover is often related to grazing, browsing, and fire interactions. Fires occur on an annual basis, and their destructiveness to trees depends on whether they occur at the beginning or end of the dry season, with early-season fires shown to improve tree cover compared to late-season ones, which burn more intensely and in turn kill trees. Trees are equipped with thick bark to protect against both water loss and fire, but the intense heat produced by late-season burns can override this adaptation. African savannas are unique with respect to supporting both grazers, which feed primarily on grasses and forbs, and browsers, which prefer woody vegetation (Holdo, Holt, and Fryxell 2009). Wildebeests are a common grazer and, according to Holdo, Holt, and Fryxell (2009), elephants are a particularly destructive browser because of how quickly they can reduce tree abundance by toppling canopy trees and preventing shorter ones from reaching full size. In addition, the loss of trees creates an opportunity for grasses to expand, which, owing to the competitive advantages of grasses, creates a bottleneck for further tree establishment while concurrently creating more highly flammable, fire-prone grasses. This leads to more widespread fires and tree loss across the landscape unless wildebeests are able to remove enough fine fuels through grazing and thus modify the impacts of fire (Holdo, Holt, and Fryxell 2009).

SEE ALSO: Biodiversity; Boreal forest ecosystems; Climate change and biogeography; Community/continuum in biogeography; Grassland ecosystems; Mountain biogeography; Polar region ecosystems; Tropical savanna ecosystems

References


Elliott, G.P. 2011. “Influences of 20th-Century Warming at the Upper Tree Line Contingent on
TREELINE ECOTONES


Further reading

A hurricane is a tropical cyclone over the Atlantic Ocean with maximum sustained winds of at least 74 mph (33 m s$^{-1}$ or 64 kt). In this entry the discussion of hurricanes will serve as the basis for a broader exploration of tropical cyclones. Large-surface low-pressure systems are classified as tropical, extratropical, or subtropical. A tropical cyclone is a low-pressure system that forms over tropical or subtropical oceans. It has a well-developed cyclonic (i.e., counterclockwise in the Northern Hemisphere or clockwise in the Southern Hemisphere) circulation at the surface and a warm core in the middle and upper troposphere. Tropical cyclones are driven by the release of latent energy, and the strongest winds occur relatively close to the center. Frontal boundaries are not attached to a purely tropical cyclone. An extratropical cyclone is a low-pressure system that develops over oceans or land in the middle and high latitudes. It generally forms along existing frontal boundaries and has a cold core in the middle and upper troposphere. Extratropical cyclones are driven by the baroclinic conversion of gravitational potential energy to kinetic energy as cold, dense air moves under warmer, less dense air. The strongest winds may occur at a large distance from the center of an extratropical cyclone. A subtropical cyclone exhibits some characteristics of tropical cyclones and some of extratropical cyclones. It forms over subtropical or tropical oceans and has some deep convection. However, it lacks a central core of convection and it gets energy from baroclinic conversion and the release of latent energy. A subtropical cyclone may have a cold core in the upper troposphere and the strongest winds may occur at an intermediate distance (i.e., several hundred kilometers) from the center.

Tropical cyclones are given various classifications according to their geographic locations. Tropical cyclones are designated as tropical depressions, tropical storms, or hurricanes over the Atlantic, Eastern North Pacific, and Central North Pacific Oceans. Similar low-pressure systems are called tropical depressions, tropical storms, or typhoons over the Western Pacific Ocean. Designations of depressions and cyclonic storms are used over the Northern Indian Ocean, while similar systems are categorized as tropical depressions, tropical storms, or tropical cyclones over the Southwest Indian Ocean.

Formation of hurricanes

Formation of a hurricane requires a pre-existing disturbance, sufficiently warm sea surface temperatures (SSTs), and light winds in the upper troposphere. The pre-existing disturbance may be of tropical, subtropical, or extratropical origin. The original disturbance consists of an area of low pressure at the surface and it may contain some regions of convection. The low serves as a focus for convergence and as a source of low-level vorticity (the tendency for air to rotate cyclonically). As air at the surface flows over a warm ocean, evaporation transfers large amounts of latent energy into the air. The SST in a tropical
ocean has to be at least 26°C to provide enough energy to support the development of a hurricane. When the air converges in the low-pressure system, it begins to rise. Condensation in deep convective clouds converts latent energy into internal energy. Enough latent energy is released to warm the temperature in the middle and upper troposphere, and light winds in the upper troposphere allow for the formation of a warm core over the center of the system. The warm core enhances upper-level divergence, which pumps mass out of the developing cyclone and allows the pressure at the surface to decrease. As the pressure decreases, the pressure gradient force increases and accelerates the air into the center of the developing storm. Stronger winds cause more evaporation and extract more latent energy from the ocean. A positive feedback intensifies the developing low-pressure system. It is classified as a tropical depression and given an alphanumeric designation when a well-organized closed circulation at the surface is identified. A tropical storm is given a name when the maximum sustained wind speed reaches 39 mph (17 m s$^{-1}$ or 33 kt). It becomes a hurricane when the maximum sustained wind speed is greater than or equal to 74 mph (33 m s$^{-1}$ or 64 kt).

Several different weather systems may serve as the initial disturbance for the formation of a hurricane. A tropical disturbance is any discrete area of organized convection that forms and moves over the tropics or subtropics and persists for more than 24 hours. Tropical (easterly) waves are elongated troughs of low pressure in the lower troposphere that are generated as tropical easterly winds flow around mountains in eastern Africa. Tropical waves are characterized by an inverted-V shape that appears in the surface isobars and may be apparent on satellite images. The waves form every one to four days during the summer and early autumn. They move westward across sub-Saharan Africa and emerge over the Atlantic Ocean. Approximately 10–20% of tropical waves develop into tropical cyclones over the Atlantic Ocean. The majority of tropical waves dissipate without becoming a tropical cyclone, but some of them cross into the Eastern North Pacific Ocean and intensify into hurricanes. The southern end of some tropical waves interacts with the Intertropical Convergence Zone (ITCZ) and a cluster of thunderstorms breaks off from the ITCZ. A few of these ITCZ disturbances emerge into the easterly flow at low latitudes and intensify into hurricanes over the Atlantic Ocean. In portions of the world with well-developed monsoonal circulations, such as the Western North Pacific Ocean, a monsoon trough or gyre may be the initial source of low-level convergence, convection, and vorticity.

Extratropical and subtropical weather systems can make a transition into tropical cyclones during cooler seasons. Nearly stationary frontal boundaries, old extratropical cyclones, midtropospheric cutoff lows, mesoscale convective systems, and subtropical cyclones have transformed into hurricanes. These weather systems initially have a cold core at some levels in the atmosphere, a large low-level circulation, and scattered convection at some distance from their center. It can take days for them to be transformed into a tropical cyclone. When they move over the ocean and sit over relatively warm SSTs, evaporation increases the latent energy in the air. An additional transfer of internal energy from the warmer ocean into cooler air reduces horizontal temperature gradients and existing frontal boundaries weaken. Latent energy is released in convection and starts to warm the atmosphere in the middle and upper troposphere. If convection is generated near the center of circulation, a warm core develops and the system exhibits a more tropical cyclone-like structure with the strongest winds located closer
to the center. This tropical transition results in the low being reclassified as a tropical cyclone. In some years, typically when El Niño conditions exist, nearly half of tropical cyclones over the Atlantic may develop as a result of this type of tropical transition.

Structure of hurricanes

Hurricanes exhibit a structure that is distinct from most other types of cyclones. Showers and thunderstorms are organized in bands of rain that spiral cyclonically around the center of circulation. Local regions of convergence in the low-level wind flow generate enhanced rising motion and create the bands of precipitation. Air sinks in a broad area outside of these rainbands and in a narrow region just inside the band. The pattern of vertical motion creates relatively precipitation-free zones between rainbands. A larger primary or principal rainband may spiral into the center of a tropical cyclone. The strongest horizontal winds may occur within the primary rainband of a tropical depression or tropical storm. When rising air reaches the upper troposphere, it diverges anticyclonically and spreads a canopy of cirrus clouds over the tropical cyclone. As a tropical storm intensifies into a hurricane, the primary rainband may wrap around the center of circulation. Subsidence, possibly associated with convection, creates a clear eye at the center of the hurricane. The subsidence does not extend to the surface and the lower one or two kilometers of the eye are comprised of moister air and may include low clouds. The eye is surrounded by a circular eyewall of strong convection. The tallest clouds and the strongest winds in a hurricane are typically found near the inner edge of the eyewall. The average diameter of an eye in an Atlantic hurricane is 46 km, but hurricane Wilma (2005) had an eye with a diameter of 3 km and some hurricanes have exhibited eyes with diameters of over 100 km.

In more intense hurricanes a rainband may wrap completely around an existing eye and eyewall to create concentric eyewalls. Willoughby, Clos, and Shoreibah (1982) presented a conceptual framework to explain eyewall replacement cycles. Initially the strongest winds and convection occur in the original, inner eyewall, and a secondary wind maximum is present in the new, outer eyewall. A clear area caused by subsidence between the two eyewalls is called the moat. As more air converges and rises in the outer eyewall, convection weakens in the inner eyewall. This causes the wind speeds to decrease in the original eyewall and the hurricane weakens temporarily and the radius of maximum winds (RMW) expands to the outer eyewall. As the inner eyewall dissipates, the outer eyewall begins to contract and the hurricane begins to reintensify. If the eyewall replacement cycle continues uninterrupted, eventually the outer eyewall will move to the previous radius of the original inner eyewall and the hurricane will regain its prior intensity.

Knaff, Kossin, and DeMaria (2003) identified a group of annular hurricanes. These hurricanes have large eyes surrounded by a symmetrical ring of convection. The eyes in annular hurricanes have an average diameter of 86 km, which is almost twice the size of eyes in average Atlantic hurricanes. Convection in outer rainbands is much less active than in other hurricanes and annular hurricanes have a “truck tire” or “donut” appearance on satellite imagery. It is still not clear what causes hurricanes to develop an annular structure, but it has been suggested that mesovortices inside the eye may mix energy and momentum until a balanced axisymmetric structure is generated. Annular hurricanes tend to maintain a steady intensity longer than other hurricanes and can pose a challenge for forecasters.
Motion of hurricanes

Many tropical cyclones develop in easterly flow located on the equatorward side of the subtropical high-pressure systems. The easterly flow extends throughout most of the depth of the troposphere and initially steers the storms toward the west. A hurricane will begin to move more poleward as it nears the western end of the subtropical high. As the hurricane moves toward higher latitudes, it begins to enter the zone of westerly flow in the middle and upper troposphere. The westerly flow causes the hurricane to recurve back toward the east. The forward motion of a hurricane will often accelerate as it moves poleward. The precise steering level for a specific tropical cyclone will depend on the intensity of the storm and its thermal environment. Strong storms over the warm SSTs in the tropics will have deep convection and will be steered by the winds throughout the troposphere. An average of the winds between 850 hPa and 200 hPa may give a good estimate of the short-term motion of these hurricanes. The convection in a storm over cooler SSTs will not be as tall and those hurricanes are more likely to be steered by the winds in the middle troposphere (e.g., 500 hPa). Tropical cyclones that are weaker and exhibit mainly shallow convection are steered by the winds in the lower troposphere (e.g., 850 hPa). When the winds at the steering level are weak, a hurricane may become stationary or complete several loops before resuming a steadier forward motion.

Dissipation of hurricanes

Since the primary source of energy for a hurricane comes from evaporation of water from the ocean, movement over land greatly reduces the energy supplied to the storm. Increased friction over land contributes to the weakening of the circulation and movement over mountains can quickly dissipate a hurricane. However, cases have been documented where tropical cyclones maintained their intensity or reintensified modestly when they were well inland. Andersen and Shepherd (2013) performed a global analysis of inland tropical cyclones. They identified a small group of cases they called tropical cyclone maintenance/intensifying (TCMI) systems. The TCMI systems were most prevalent over Australia. They also reported another group of cases they labeled neutral/hybrid that occurred mainly over the United States, China, and India. It is hypothesized that large upward fluxes of latent and internal energy from the surface maintained the circulation and warm core of the inland tropical cyclones.

When hurricanes move poleward over the oceans, cooler SSTs result in less evaporation and hurricanes begin to weaken. Stronger upper-level westerly winds and a more baroclinic environment may cause a hurricane to make an extratropical transition into an extratropical cyclone as it moves poleward. Some hurricanes are sheared apart when they move into a region where there is a large difference in the winds in the lower and upper troposphere. The increased vertical shear can result from an increase in the low-level easterly flow caused by a surge in the trade winds, or from a rapid increase in the flow in the upper troposphere. In either scenario the low-level circulation may become detached from the upper-level warm core and it dissipates in two or three days.

Geographical distribution and frequency of tropical cyclones

Globally, an average of 86 named tropical cyclones form each year. Around 47 tropical cyclones reach hurricane intensity each year. Although there is some year-to-year variability,
the average global frequency of tropical cyclones is relatively stable. A large area of warm SSTs, known as the Western Pacific Warm Pool, contributes to a favorable environment that results in the formation of more tropical cyclones over the Western North Pacific (WNP) Ocean than over any other basin. Approximately 26 named tropical cyclones originate each year over the WNP. The second most active basin is the Eastern North Pacific where an average of 17 named tropical cyclones and nine hurricanes form in a year. The highest frequency of tropical cyclone formation per unit area occurs over the Eastern North Pacific. An average of 12 named tropical cyclones and six hurricanes form annually over the Atlantic Ocean, Caribbean Sea, and Gulf of Mexico. There are an average of 11 named tropical cyclones near Australia, nine over the Southwest Indian Ocean, seven over the South Pacific Ocean, five over the Northern Indian Ocean, and one over the Central Pacific Ocean in a year. For a long time it was thought that hurricanes did not occur over the South Atlantic Ocean. However, in March 2004 a low-pressure system east of Brazil underwent a tropical transition and struck the Brazilian state of Santa Catarina with hurricane force winds. This storm was subsequently given the name Hurricane Catarina and it represents the first documented instance of a tropical cyclone over the South Atlantic Ocean. In rare instances small intense low-pressure systems form over the Mediterranean Sea that have a structure and some characteristics similar to those found in tropical cyclones.

**Hazards accompanying hurricanes**

High winds associated with hurricanes present a danger to life and property. In the strongest hurricanes winds may approach 200 mph (90 m s\(^{-1}\)) and can do damage to even well-constructed buildings. Strong winds blowing toward the shore can create a rapid rise of water level called a *storm surge*. The storm surge can increase water levels by several meters in weaker tropical cyclones and by 10–15 m in the strongest hurricanes. Indentations in the coastline associated with bays or rivers may funnel water into particular locations and cause locally higher water levels. Storm surges are the greatest hazard at the coastline and have been responsible for some of the greatest loss of life in tropical cyclones. A storm surge associated with a hurricane killed thousands in Galveston, Texas, in 1900, which is the largest loss of life in the United States from a hurricane. A storm surge with a tropical cyclone over the Bay of Bengal killed over 100,000 people in Bangladesh in 1970. Heavy rainfall and the flooding it causes is often the greatest hazard in inland locations. Rainfall rates can approach a meter per day in large slow-moving tropical cyclones. Steep topography and deforestation increase the speed of surface runoff and increase the potential for flooding of rivers and streams. Floods caused by Hurricane Mitch killed over 10,000 people in 1998. Some hurricanes spawn tornadoes when they make landfall. Increased friction at the surface slows the wind and it changes direction, while the winds initially are less affected several thousand meters above the surface. The vertical wind shear created by the differences in wind speed and direction causes rotation in the updrafts of some thunderstorms in the rainbands around a hurricane. In some cases this rotation is strong enough to spin up a tornado. Most tornadoes generated in hurricanes are relatively weak and short-lived. However, they are capable of doing damage and causing casualties. Many hurricanes create only a few tornadoes, but several have produced large numbers of tornadoes. The circulation around Hurricane Ivan generated 117 tornadoes.
when it made landfall in the United States in 2004.

Hurricane and tropical cyclone warnings and classifications

The World Meteorological Organization (WMO) has assigned the responsibility for monitoring the oceans and issuing warnings for tropical cyclones to Regional Specialized Meteorological Centers (RSMCs). Each RSMC has the responsibility for a designated area of the Earth. The National Hurricane Center (NHC) monitors the North Atlantic and Eastern North Pacific Oceans, The Central Pacific Hurricane Center (CPHC) monitors the Central Pacific Ocean, the Japan Meteorological Agency (JMA) covers the Western North Pacific Ocean, the Indian Meteorological Department (IMD) handles the North Indian Ocean, and Météo France issues warnings for the South West Indian Ocean. In addition, Tropical Cyclone Warning Centers provide warning information for smaller regions, and national meteorological agencies may provide specific advisories for their countries.

Each RSMC may use a different system for classifying tropical cyclones based on their maximum sustained wind speed. The NHC and CPHC designate tropical cyclones as tropical depressions, tropical storms, or hurricanes. Those RSMCs also use the five categories of the Saffir–Simpson Scale to specify the intensity of hurricanes based on their maximum sustained wind speed. Category 1 and 2 storms are sometimes considered to be minor hurricanes while Category 3, 4, and 5 hurricanes are considered to be major hurricanes. The JMA classifies tropical cyclones as tropical depressions, tropical storms, severe tropical storms, or typhoons. The IMD uses a classification system that includes depression, deep depression, cyclonic storm, severe cyclonic storm, very severe cyclonic storm, and super cyclonic storm. Météo France’s classification scheme ranges from tropical depression through moderate tropical storm, severe tropical storm, tropical cyclone, and intense tropical cyclone to very intense tropical cyclone.

Climatological cycles of hurricane activity

Hurricane activity appears to vary on timescales running from intraseasonal to decadal. Attempts have been made to link the variability to other known atmospheric cycles. Maloney and Hartman (2000) postulated a link between the frequency of hurricanes over the Eastern Pacific Ocean and the phase of the Madden–Julian Oscillation (MJO). The MJO is a 30–60–day oscillation in convection and outgoing longwave radiation that moves west to east across the tropics. In one phase of the MJO, enhanced low-latitude westerly winds in the lower troposphere increase the vorticity around tropical disturbances and upper-level divergence may contribute to the development of tropical cyclones. The North Atlantic Oscillation (NAO), which is a function of the pressure difference between the subtropical high near the Azores and low pressure near Iceland, has been linked to changes in the location of formation of tropical cyclones and their subsequent tracks. Elsner, Liu, and Kocher (2000) concluded that when the subtropical high is located farther south and west, the NAO will be negative and hurricanes will take more westerly tracks across the lower latitudes of the Atlantic Ocean. If the subtropical high is located farther north and west, then the NAO will be positive and
hurricanes are more likely to recurve before they reach the East Coast of the United States.

Gray (1984) presented observational evidence that the El Niño Southern Oscillation (ENSO) phenomenon in the tropical Pacific Ocean had an impact on the formation of hurricanes over the Atlantic Ocean. An El Niño is associated with above normal SSTs over the tropical Central and Eastern Pacific Ocean. Enhanced convection associated with the warmer SSTs pumps mass into the upper troposphere and strengthens the westerly winds occurring in that portion of the atmosphere. Those stronger westerly winds flow across the Atlantic Ocean and produce increased vertical wind shear, which inhibits the formation of tropical cyclones. The frequency of hurricanes also seems to exhibit a longer multidecadal cycle associated with the Atlantic Multidecadal Oscillation (AMO). It has been hypothesized that the AMO represents a multidecadal oscillation of SSTs in the North Atlantic and thus modulates hurricane activity on the same timescale. There has been a significant increase in the frequency of hurricanes over the Atlantic Ocean since 1995, which may be associated with the active portion of the AMO.

Climate change and hurricane activity

The possible impact of climate change on hurricane activity has been a subject of some debate. Early research which focused on an increase in SST suggested that warmer SSTs would produce more and stronger tropical cyclones. However, the situation is more complicated than that and a number of environmental factors affect the development and intensification of hurricanes. Some coupled global atmosphere/ocean models have produced a reduction in the worldwide number of simulated tropical cyclones when the concentration of carbon dioxide and the SST were increased. The greatest decrease occurred in the Southern Hemisphere and the changes in hurricane activity over the Atlantic Ocean were less clear. In some models a decrease in the humidity in the middle troposphere and a reduction in upward motion were linked to a decrease in the number of tropical cyclones over some basins.

There has also been a suggestion that a warmer climate could lead to a reduction in the speed of the upper-level westerly winds in the middle latitudes. This would be consistent with the concept that the upper-level westerly winds are partly a function of the pole-to-equator temperature difference. If those westerly winds were weaker, there would be less wind shear, which could potentially result in stronger hurricanes at higher latitudes. Less cold air in the middle latitudes could also result in tropical cyclones moving further poleward before extratropical transition occurs. Any changes to the global wind pattern could also alter the origin and types of disturbance that produce tropical cyclones and the paths of those storms. Much more research needs to be done before a definitive statement can be made about the impact of climate change on future hurricane activity.

SEE ALSO: Atmospheric/general circulation; Environmental risk analysis; Natural hazards and disasters; Oceanic circulation; Weather, extreme

References


Tropical geography

Daniel Clayton
University of St Andrews, UK

In his 1947 primer *Les pays tropicaux*, the French geographer Pierre Gourou envisioned tropical geography as a disciplinary specialism dealing with the physical and human geography of the tropical zone – usually understood as the region of the Earth around the equator between the Tropic of Cancer (23° 26′ 16″ N) and the Tropic of Capricorn (23° 26′ 16″ S), and distinguished from surrounding (temperate and polar) regions in cosmographic, environmental (climatic and biogeographic), and sometimes ethnographic terms. But Gourou asked a profound question which pushes what he had in mind into a much wider critical frame: Does “the tropics” augur a certain way of seeing or style of inquiry; or should the idea of “tropical geography” be conceived more restrictively, as a branch of geographical inquiry that deploys (and if need be adapts) conventional disciplinary tools to tropical settings?

Gourou went down the latter route, and during the 1950s and 1960s tropical geography emerged as a distinct research area within geography, most evidently in those countries that had (or once had) tropical colonies – America, Belgium, Britain, France, Germany, and Portugal. Significant – although, with the exception of Bordeaux, never sizeable – ‘schools’ and ‘laboratories’ of tropical geography were established in these Western countries, and in West and East Africa, Brazil, and South and Southeast Asia. Gourou, Paul Pélissier, Gilles Sauter, Orlando Ribeiro, and Robert Steel (to name only some of the most prominent figures) undertook fieldwork. By the 1970s there were also a number of textbooks on tropical geography besides Gourou’s (which became a staple of university geography courses on the subject).

Gourou adapted the regional paradigm in French geography pioneered by Vidal de la Blache (with its chiefly rural focus on “man and milieu”) to the rice-growing peasant civilization of the Tonkin Delta in French Indochina, and his regional monograph (the product of six years of fieldwork) is still considered one of the finest in this tradition. Gourou was also a vehement critic of environmental determinism, which imbued early twentieth-century American and German geography, and viewed the field-based analysis of what he termed “techniques d’encadrement” (local landscape molding techniques), and the cartographic analysis of population densities, as keys to unlocking the nature and mysteries of tropical geographies. However, tropical geography was conceived in other ways. Gourou’s *Les pays tropicaux* advocated a zonal understanding of the tropics, and geographers (such as Charles Fisher, Dudley Stamp, George Kimble, and Gourou himself) also produced works on “monsoon Asia,” “tropical Africa” and “Amazonia.” Most works in tropical geography proceeded (as was customary in regional geography) from an analysis of physical geography to that of human geography. But this research niche was also sometimes deemed to lie at the interface between a “regional” and a more “systematic” (or integrated) geography (although the conceptual difficulties surrounding these competing claims to know in unique and general ways were hardly explored), and it had strong links with the
TROPICAL GEOGRAPHY

subfields of population geography and economic geography.

Furthermore, American, British, and Portuguese geographers generally had a more utilitarian concern with the economics of tropical development than their French counterparts, who tended to focus more on land-use practices. Much postwar tropical geography was thus palpably “applied,” and, in tandem with other fields (especially agronomy), cultivated the tropics as a laboratory within which new forms of scientific management and academic expertise could be brought within the armature of the late colonial state and newly independent governments. Indeed, some of Gourou’s work on the Far East was commissioned by the French ministry of foreign affairs, and his postwar fieldwork on tropical Africa was funded by the Belgian government. His expertise was also sought by the United Nations (Bowd and Clayton 2016). These imbrications of knowledge and power in this era of decolonization extended a much longer imperial history, stretching back to the Spanish and Portuguese in the Amazon, and British and Dutch in the East Indies, in which scientific deliberations and aesthetic reactions to the tropics were bound up with commercial and political objectives (Driver and Martins 2005). However, by the 1970s radical geographers inspired by Marxism and feminism, and concerned with issues of dependency and underdevelopment, deemed tropical geography conceptually staid and politically reactionary, and since the late 1970s the appellation has for the most part been supplanted by others that resituate “the tropics” in alternative critical frames (development geography, political ecology, environmental management, postcolonial studies, and globalization).

But let us backtrack to the 1950s. For it is in the overlap between geography and decolonization, and Western intellectual inquiry and Western dominance that the other side of Gourou’s question comes into view: the equation of the tropics with the “exotic,” and the priming of an exotic tropical world for material and intellectual appropriation. Since the start of modern European empire-building in the fifteenth century, tropical climate and vegetation — a warm, moist climate year-round, albeit with rainfall distributed unevenly in wet and dry seasons in monsoon regions, and falling more evenly through the year in rainforest and subtropical (savannah) ecosystems — have been lodged in Western perception and experience as potent emblems of environmental otherness, and out of this has come a close and enduring set of connections between tropical geography and power. It has been argued that “the tropics” solicited a particular set of Western sensitivities: ones revolving primarily, and in the most visceral of ways, around the senses (particularly sight, but also sound) and the vulnerability of the human body and mind to heat and humidity, and inimitably tropical diseases (Driver and Martins 2005).

David Arnold (2000) and now geographers have used the term “tropicality” to describe these sensitivities and the imaginative geography (akin to Edward Said’s idea of Orientalism) they install. Tropicality is not simply a Western style of imagining the tropics — as the West’s unruly environmental other; as the opposite of a temperate (regular, moderate, and composed) Western nature. It has also served as a way of “handling” (as Said put it in relation to the Orient) the tropics — licensing exploitation and “intemperate” behavior. Furthermore, since many of the figures who have been extolled as modern geography’s intellectual forebears (explorers and naturalists such as James Cook, Alexander von Humboldt, and Charles Darwin) focused their energies on the tropical belt, contemporary geographers with postcolonial critical sympathies have found it at once important and difficult to disentangle
their discipline’s heritage and public image from a long history of tropical othering and its strong imperial overtones (Driver and Martins 2005).

However, tropicality is an ambivalent construct and its connection to geography’s imbroglio with empire can be overplayed. In Gourou’s work and a much greater swathe of Western experience and learning, the tropics have been exoticized in both positive and negative terms: on the one hand, as edenic, picturesque, fecund, alluring, and promiscuous (Paul Gauguin’s Tahiti paintings in many ways the epitome of this vision); and on the other hand, as stifling, disease-ridden, deceiving and degenerative, a green hell, savage wilderness, land of illusory riches, “the heart of darkness” (to draw on Joseph Conrad’s modern fable about Western civilization lapsing into primitivism in the jungle peripheries of the capitalist and colonial world). Gourou’s work provides a good example of this ambivalence. He romanticized the rice-growing delta areas of the Far East as beautiful and harmonious, but represented equatorial Africa in pejorative terms, as “suffering from a certain number of inferiorities” (Gourou 1947, 176). Critics accused him of casting a stifling geographical spell on the tropics. At the same time, he was skeptical about the benefits of colonialism and Western style development in the tropics. He identified poor soils and deleterious agricultural practices as inherent sources of tropical backwardness; yet he also saw them as “problems” that had been exacerbated by colonial and development practices (see Bowd and Clayton 2016). In other words, while it is in some ways easy to wield a postcolonial critical agenda which situates the tropics and bodies of work like Gourou’s in a Western imperial imaginary, and see tropical geography as part of geography’s maleficent and lingering Eurocentrism, the story of tropicality and its relation to tropical geography is complex.

Work on tropicality has added an important environmental dimension to critical debates about the development of Western science, medicine, and colonial power, and encouraged critical and historical reflection on the categories and practices of tropical geography. However, negative tropicality is being given a new lease of life in the work of prominent public intellectuals, such as Jeffrey Sachs, who are talking again about “tropical backwardness” and the humanitarain “curse” of the tropics. The annals of The Malayan Journal of Tropical Geography (launched in 1953 and renamed the Singapore Journal of Tropical Geography in 1980) provide a rich record of (especially anglophone) research in this area, and in recent decades has published essays on tropicality and the place of the tropical world in a broader postcolonial geography.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Development; Environmental discourse; Imaginative geographies

References

Tropical geomorphology

Avijit Gupta  
University of Wollongong, Australia  
National University of Singapore

Tropical geomorphology is a branch of geomorphology that studies landforms and processes in the tropics and subtropics. The tropics may be perceived as an area of radiative surplus between the two anticyclonic circulations near the latitudes of 30°N and 30°S. This is an area over which the Hadley Cell operates, giving rise to easterly trade winds which converge at the equator. The tropics are considered a climatic region but the only common meteorological component across this belt of low latitudes is temperature. Considerable climatic variations are found across the tropics, primarily in rainfall. The annual total, the seasonal distribution, and occasional synoptic disturbances all vary across the tropical zone which, depending on the amount of moisture available, can be divided into two basic units: the humid and the arid tropics. About half of the surface of our planet lies in the tropics. The tropical oceanic expanse is huge and plays an important role in influencing the world climate. A number of geomorphic processes operate in the tropics over a range of rock types and geological structures, resulting in a wide variety of landforms. The processes are similar in kind to those found in other regions, but in the tropics they operate at different rates and with varying intensities.

In the geologic past, the present tropical landmasses (South and Central America, Africa, the Indian subcontinent, Southeast Asia, and Australia), a number of islands, and the cold continent of Antarctica constituted a single landmass known as Gondwana or Gondwanaland. Gondwana, located in the cold southern latitudes, was the southern part of a single world continent, Pangea. Pangea broke up about 200 million years ago, and the different parts of Gondwana drifted slowly towards the present warmer locations via plate tectonics to create the present distribution of land and sea in the low latitudes. This entire history is reflected in the tropical landforms: the origin from Gondwana, the break-up, the drifting of individual landmasses towards the equator, the impact of geomorphic processes after reaching the tropical zone, and currently, the changes caused by anthropogenic activities. The tropics exhibit a wide variety of landforms, and no single representative template exists.

Traditional tropical geomorphology

We know very little about the geomorphological knowledge that existed before the arrival of Western maritime powers in the tropics. From the last decade of the eighteenth century, Western scientific explorers including Alexander von Humboldt, Alfred Russel Wallace, and Charles Darwin visited the tropics with great astonishment, enthusiasm, and insight. They were revolutionizing biological sciences, but they also left perceptive accounts of regional geology and geomorphology. Our knowledge was further extended by geologists such as Sapper, Passarge, and Bornhardt from Germany and Dana from the United States. However, even up to the 1970s, the landforms and operating
TROPICAL GEOMORPHOLOGY

Geomorphic processes in the tropics were considered to be primarily controlled by climate. The work of Peltier, De Martonne, and Büdel was associated with this period. At the same time, the teaching and research carried out in the tropical countries used textbooks based on extra-tropical examples. Two concepts were heavily used: the Davisian cycle of erosion and the concept of climo-morphogenetic regions (Gupta 1993).

Modernizing tropical geomorphology

The climate-based approach of tropical geomorphology changed from the second half of the twentieth century. The writings of L.C. King in South Africa emphasized the role of geological history and rock types in understanding the ambient landscape. A number of excellent case studies, such as Ruxton and Berry on granite weathering in Hong Kong, Simonett on earthquake-generated slope failures in New Guinea mountains, and Coleman’s account of the Brahmaputra River in Bangladesh, highlighted the new process-based approach. Several regional collections of case studies were published, and accounts of tropical landforms and geomorphological processes became easily available, reducing the exoticness of tropical geomorphology. This was also the time of diffusion of modern mainstream concepts and techniques to geomorphologists of tropical residence (Gupta 1993 and references therein).

The diffusion of the new concepts of process geomorphology resulted in a rapid increase in the quantity and quality of research. Several textbooks in geomorphology with a tropical slant were published. Examples from the tropics were no longer considered unusual, and even used as background information for theory building. The Andes and Himalaya were studied repeatedly as sediment sources, and the Amazon and Ganga-Brahmaputra river systems monitored in order to determine sediment storage and transfer in large river systems (Dunne et al. 1998; Kuehl et al. 2005; Meade 2007). Mainstream geomorphic processes were seen to operate in the tropics but at different rates. For example, denudation rates were found to be unexpectedly high in many areas.

Although climate acts as an important control in both humid and arid tropics, landforms and operating processes are dependent also on other factors. Gardner et al. (1987), writing about certain locations in Central America and the Caribbean, highlighted three items from a long list of regional geomorphic processes and landforms. These are karst, alluvial fans, and tectonism along convergent plate margins, being the most characteristic of the region and representing the dominance of lithology, a rapid fluvial depositional process, and neotectonics. In other parts of the Caribbean, the local landforms demonstrate the dominant role of hurricanes or volcanoes. The fascination and challenge of tropical geomorphology lies in the recognition of such diversity within the region.

A number of world maps on suspended sediment yield, published since the 1980s, demonstrate the rapidity and efficiency of tropical geomorphic processes (Milliman and Syvitski 1992 and references therein). These show very high rates for most of the humid tropics, especially in high-relief areas. Milliman and Syvitski interpreted relief as a surrogate for tectonism, and suggested that large sediment yields from active orogenic belts are due to “the entire tectonic milieu of fractured and brecciated rocks, oversteepened slopes, seismic and volcanic activity” (Milliman and Syvitski 1992, 539–540). The high sediment yield is also due to intense tropical rain falling on the tectonic mountains.
Table 1  Tropical geomorphology: a summary description.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major controls</td>
<td>Location of tectonic belts, volcanoes, cratons, alluvial valleys, deltas</td>
</tr>
<tr>
<td></td>
<td>Climate, mostly rainfall</td>
</tr>
<tr>
<td></td>
<td>Vegetation</td>
</tr>
<tr>
<td></td>
<td>Land cover changes</td>
</tr>
<tr>
<td>Geomorphic processes</td>
<td>Weathering</td>
</tr>
<tr>
<td></td>
<td>Mass movements</td>
</tr>
<tr>
<td></td>
<td>Rivers, frequently seasonal and prone to flooding</td>
</tr>
<tr>
<td></td>
<td>Tropical karst</td>
</tr>
<tr>
<td></td>
<td>Glacial and fluvial processes on high mountains</td>
</tr>
<tr>
<td></td>
<td>Coastal (including mangroves and coral reef) processes</td>
</tr>
<tr>
<td></td>
<td>Wind and river processes in the arid tropics</td>
</tr>
<tr>
<td>Quaternary history</td>
<td>Pleistocene glaciation in mountains</td>
</tr>
<tr>
<td></td>
<td>Climate change</td>
</tr>
<tr>
<td></td>
<td>Sea level changes</td>
</tr>
<tr>
<td>Present and future changes</td>
<td>Anthropogenic changes</td>
</tr>
<tr>
<td></td>
<td>Global warming and climate change</td>
</tr>
</tbody>
</table>

Source: Based on Gupta (2011), p. 10, Table 1.1.

The natural geomorphic processes, however, are not necessarily the only ones in operation. Parts of the tropics have a long history of human occupation, leading to modification of the natural landscape. Such alterations accelerated from the 1950s, and natural landforms and processes became progressively less extensive in areal coverage. The destruction of tropical rain forest has increased the annual sediment yield from less than $10^2$ t km$^{-2}$ to more than $10^3$ t km$^{-2}$ in Southeast Asia. Transfer and storage of such increased volume of sediment sequentially affect hillslopes, gullies, rivers, and coastal forms. Thus, erosion and sedimentation in the tropics is a function of relief, climate, and land use: all three.

Tropical geomorphology combines scientific research with the application of its findings for the betterment of people (Gupta 2011). Recent research in tropical geomorphology highlights three areas (Table 1):

1. the major controls in tropical geomorphology: geology, landforms, climate, and land use
2. the operating geomorphic processes (mainly dealing with the passage of water and sediment from the mountains to the coast), and inherited Pleistocene changes
3. anthropogenic alteration of the natural rates and processes, including the effect of climate change

Major controls in tropical geomorphology

A variety of landforms occur across the tropics, depending on their history. Part of the tropics are inherited from Gondwana, forming uplands or basins at various levels, underlain by old metamorphic rocks that are difficult to erode such as found in the Brazilian Shield. The old uplands are also formed by flood
basalts and granitic bodies which give rise to flat-topped plateaus (Deccan Plateau of India) and steep-sided rounded hills (Main Range of Malaysia), respectively. Convergent plate boundaries, and the active sides of the moving continental plates, created mountains such as the Andes and Himalaya. Sediment from such mountains filled large structural downwarps to form wide plains such as the valleys of the Amazon and Ganga. Sediment then traveled to the passive side of the continental plates to build coastal plains and deltas over the shallow continental shelf. Geological structures and rock types induced small-scale landforms against this background (Gupta 2011).

Active mountains, positioned by plate tectonics, are eroded heavily to become primary sediment sources for large rivers such as the Amazon, Ganga, Brahmaputra, and Mekong. More than 90% of the Amazon’s sediment is derived from the Andes and its foothills, and carried along its 6000 km length to the South Atlantic (Meade 2007). Short streams that drain the mainly volcanic steep slopes of the islands of Southeast Asia and the Indian and Pacific Oceans may carry a very high sediment load. The annual contributions from their basins have been estimated to be 3000 tkm$^{-2}$.

In contrast, little sediment is carried by short steep streams flowing down the drier western slopes of the Andes. Ancient and metamorphosed stable rocks of most of tropical Africa produce very little sediment. In tropical Africa, rivers that drain the high relief associated with rifting in East Africa transfer higher sediment discharges, but not high by world standards. Old rocks of South Asia or Brazil also produce a low volume of sediment.

Climatic characteristics do affect geomorphology in the tropics but not as deterministically as was presented in the past. For example, the warm and wet climate of the humid tropics accelerates the process of weathering, and the high-intensity rainfall gives rise to slope failures and flooded rivers in the humid tropics.

Over the oceans, the easterly trade winds operate throughout the year with minor latitudinal shifts and small variations in intensity, bringing rain primarily to the eastern slopes of islands and land masses. In the higher tropics, an alternating wet and dry season operates over the continents and coastal waters due to a seasonal reversal of wind systems known as the monsoon. Rainfall from moisture-bearing winds is triggered by orographic lifting, convergence, and atmospheric disturbances. Depressions bearing heavy rain tend to arrive embedded in the monsoon system of winds. Arriving in the middle of the wet monsoon, when the soil is already saturated and rivers high, these events frequently cause slopes to fail and rivers to flood. The general pattern of rainfall in the tropics is periodically disrupted by the El Niño Southern Oscillation (ENSO), causing certain years to be wetter or drier, and causing floods or droughts.

Tropical cyclones (also known as hurricanes or typhoons) are large atmospheric disturbances which develop from depressions over warm oceans in summer. These huge revolving storms are carried in the trade winds, traveling west from their area of origin over oceans. They are associated with very strong winds, spiralling clouds, heavy rainfall, and storm surges. Daily rainfall in a tropical cyclone may reach several hundreds of millimeters. If the tropical cyclones slow down or confront orographic barriers, the cumulative rainfall over several days may reach very high totals. Such storms, associated with extensive slope failures, massive floods, and catastrophic coastal damage, are recognized as high-magnitude, low-frequency events capable of causing large-scale landform alteration in days.

Tropical rainfall can be strikingly erosive. Hudson (1971), examining the erosive nature
of rainfall, mentioned that rainfall intensity seldom exceeds 75 mm per hour in temperate countries but intensities double that amount are common in the tropics. The surface of the land used to be protected by the extensive cover of tropical forests. Significant destruction of natural vegetation began with the start of agriculture at the beginning of the Holocene, and within a few thousand years forests were being depleted at a fast rate. From the mid-twentieth century, deforestation was accelerated by demand for timber and for land for pasture or plantations of rubber or oil palm. Land surface exposed to the direct impact of the rain became subjected to accelerated slope erosion and enhanced deposition of sediment in rivers and coastal waters.

Geomorphic processes in the tropics are influenced by three types of climatic variations. Near the equator it is nearly always warm and humid, weathering is advanced, and unless the natural vegetation has been removed, a deep soil cover is in place. A seasonal component exists away from the equator resulting in variations in streamflow and arrival of floods. Weathering and soil formation are limited in the arid tropics where rivers carry water only for a short period after a rare rainstorm.

Operating geomorphic processes in the tropics

A range of geomorphic processes, the majority of which also occur elsewhere, operate at varying rates in the humid and arid tropics. The following description highlights processes that operate efficiently in the tropics, giving it a characteristic landscape. Weathering operates differently between the arid and the humid tropics. Rocks break up physically into a shallow layer of small clasts in the arid sector whereas the hot and humid climate of the humid tropics accelerates chemical changes in minerals. The thick weathered layer near the equator, the reddening of soils on igneous rocks and limestone, and the characteristic layers in weathered granite with sand and clay on top and large corestones in depth, all are very characteristic of the humid tropics due to intense chemical weathering (Figure 1).

Deep soft weathered layers fail as a rotating mass following rain, the process being called rotational slides or slumps. The thickness of the weathered zone, however, is frequently interrupted by impervious layers of precipitated material. The generic term for such layers is “duricrust,” but they are also specified by the constituent material and called silcrete, calcrite, ferricrete, and so on. This impermeable layer allows the weathered mass above it to store water, become saturated, and fail. Such failures are translational or planar slides which also happen at the contact between the weathered mass and the hard rock underneath.

Many types of mass movements may happen on tropical slopes but the common ones are creep, planar and rotational slides, and debris flows and debris slides. Weathered material tends to creep in the rain forest. Slides are common on slopes, and debris flows occur on hillslopes following high-intensity rain. In the humid tropics, the failures are moisture driven, and slopes fail in the wet season when a rainfall event acts as a trigger to the antecedent moisture already stored in the weathered material. In contrast, slope failures are episodic events in the arid tropics, happening only after earthquakes or rare heavy cloudbursts.

Slope failures in tropical mountains contribute most of the river sediment, irrespective of the scale of the basin. Larsen and Torres Sánchez (1992), working in a small river basin in eastern Puerto Rico, calculated that 81% of the total
sediment transported out of the basin originated in mass movements. The pattern is similar in very large basins including the Amazon (Meade 2007).

The sediment is transported to the coast by rivers. Rivers of the humid tropics follow the general rules of fluvial geomorphology that determine their size, shape, slope, and channel pattern. The rivers in the seasonal tropics, however, may need to adjust also to (i) seasonality in discharge and (ii) episodic large floods. These floods, often driven by tropical cyclones or disturbances, periodically affect river systems in selected regions, such as the Indian subcontinent, the Caribbean, or tropical Australia (Wohl 1992; Gupta 2011). Such rivers display a nested set of channels of different sizes in order to adjust to floods and seasonality. Working in Australia, Pickup and Rieger (1979) showed that regional channel forms result from a series of discharges rather than a single one. This nested pattern has been reported for other rivers of the seasonal tropics. The largest one carries the episodic high-magnitude flood, the middle one the wet season high flow, and the small one the dry season flow. The sediment and the channel forms are stored inside the banks of the largest channel, and are eroded and removed in high-magnitude floods and rebuilt in the interflood period (Figure 2). This may not happen if seasonality is not pronounced, the river carries only fine sediment, or it is geologically controlled.
A large amount of sediment is carried by rivers of the humid tropics, giving rise to characteristic depositional forms. These include big alluvial fans called megafans at the highland-lowland contact of tectonic mountains in the seasonal tropics, wide river valleys filled with sediment (Figure 3), and large deltas. Climate and biological activities provide tropical coasts and wetlands with a distinct appearance with the presence of mangroves, dry salt flats (sabkhas), and coral reefs.

A number of features, such as star-shaped depressions (cockpit karst), cone-shaped rises, and towers, are characteristics of the limestone topography in the humid tropics, and occur along with the karst features found elsewhere. The tropical features are attributed to high temperature and precipitation, dense vegetation, and high production of organic acids. Good examples occur in South China, Sarawak, Yucatan, Jamaica, and other places (Gupta, 2011 and references therein).

Nearly half of the tropics is semiarid or arid. Wind action and sand dunes cover about half of the arid tropics; the other half is in rock, often under a thin veneer of gravel, sand, and riverine silt. Extremely variable rainfall translates into episodic sheetflows and stream discharges, and the dominant geomorphic process on rock is flowing water, in spite of the general aridity. The channels and depositional forms reflect a nested arrangement related to the past flood history.

Tropical geomorphology also carries a Quaternary past. Tropical landforms and geomorphic processes were affected mainly by enhanced glaciation in the mountains, repeated climate change, and repeated changes in sea level. Such events in the Pleistocene explain the presence of relict alluvium in tropical river valleys, paleo-channel of major rivers, and the origin of deltas and coastal plains.

**Anthropogenic alteration of the natural rates and processes**

Anthropogenic transformation of the physical environment modifies landforms and processes. Case studies such as Rapp *et al.* in Tanzania, Dunne in Kenya, and Douglas *et al.* in Danum Valley, Sabah, indicated a remarkable increase in runoff and sediment yield following the removal of or change in vegetation cover. With removal of vegetation, soil and regolith are restructured, natural slopes altered, a dense network of gullies established, and erosion rates and sediment yields significantly increased. A combination of anthropogenic alteration and fragile landforms
locally gives rise to very high sediment yield. The resulting erosion and sediment transfer not only impact proximal slopes and rivers but also distal deltas, coasts, and coral reefs. This may happen both episodically and cumulatively.

Dams are barriers to the downstream passage of water and sediment, forcing rivers to adjust their morphology and behavior below dams. A large number of dams, some huge, have been built across tropical rivers since the mid-twentieth century. Aswan High Dam on the Nile, Kariba Dam on the Zambezi, and Xiaowan Dam on the Mekong are examples. The practice is still continuing, although dams are now hardly being built in the developed world. Without these impoundments, the blocked sediment would have built bars and floodplains downstream, and ultimately reached the coast. Rivers below dams need to adjust to a decrease in sediment and a controlled water release. This alters their form and function, and results in a negative impact on riverine ecology.
Anthropogenic alterations have occurred in the tropics for a very long time. However, since the middle of the twentieth century, high population and economic development have strikingly modified landforms and geomorphic processes. This suggests that the tropical geomorphologist may possess a problem-solving capability in environmental management. This is well illustrated in two areas: urbanization and climate change. Urbanization often degrades the physical environment by enhancing flood frequency, slope failure, and subsidence in cities (Gupta 2011). Furthermore, a number of tropical cities are located in or have spread to hazardous areas. A knowledge of geomorphology is crucial in managing such settlements (Gupta and Ahmad 1999). The present ongoing climate change will affect landforms and geomorphic processes in the tropics. Anticipating such changes and finding the ameliorating techniques could be the most fascinating part of tropical geomorphology in the future.

**SEE ALSO:** Environmental degradation; Fluvial depositional processes and landforms; Fluvial erosional processes and landforms; Geomorphology: history; Karst processes and landforms; Mass movement processes and landforms; Tropical cyclones and tropical climate; Weathering processes and landforms

**References**


Further reading


Tropical rainforest ecosystems

A.H. Armstrong
Universities Space Research Association – GESTAR, NASA Goddard Space Flight Center, USA

Defining a tropical rainforest

The term “tropical rainforest” (tropische Regenwald), was first used by A.F.W. Schimper in his classical Plant Geography (Schimper 1898) to describe “the climax, or potential natural vegetation” found in humid equatorial climates (Richards 1996). In the broadest of definitions, a forest must include the following attributes to be considered a tropical rainforest.

2. Receive rainfall regularly throughout the year (200–1000 m or more per year).
3. Remain warm and frost free all year long (mean temperatures are between 70° and 85°F, 21°C and 31°C) with very little daily fluctuation.

However, for anyone who has ever visited, studied, or even heard the term tropical rainforest, what comes to mind is a dense evergreen forest of large, extremely tall trees, interspersed with vines and epiphytes, teeming with all sorts of faunal life, from insects and birds to primates and large carnivores. Essentially, we think of a place synonymous with the word jungle. Noted evolutionary biologist Charles Darwin gave an account of what it is like to behold a tropical rainforest for the first time during his first shore visit from the Beagle to the Atlantic coast forest in Salvador, Brazil, on February 29, 1832. In his journal he wrote: “Delight … is a weak term to express the feelings of a naturalist who, for the first time, has wandered by himself in a Brazilian forest. The elegance of the grasses, the novelty of the parasitical plants, the beauty of the flowers, the glossy green of the foliage, but above all the general luxuriance of the vegetation, fill me with admiration … The noise from the insects is so loud, that it may be heard even in a vessel anchored several hundred yards from the shore … To a person fond of natural history, such a day as this brings with it a deeper pleasure than he can ever hope to experience again.” (in Whitmore 1998)

The occurrence of tropical rainforest ecosystems

Tropical rainforests occur on all three of the continental land areas that fall within the boundaries of the tropics (Figure 2). About half of the total global land area (4 × 10^6 km^2) is found in the American and neotropical rainforests, which comprise about one-sixth of the total broadleaf forest area in the world. The rainforests occurring in the Americas are divided into three discrete areas, the largest being that of the Amazon and Orinoco basins. The second extends from the Pacific coasts of Ecuador and Colombia, through Central America to the northernmost boundary in the southern state of Veracruz, Mexico. The smallest (and most disturbed) area includes the Atlantic coast of Brazil, which has by some accounts been over 99% destroyed (Whitmore 1998).
Figure 1. Photos of rainforests. Clockwise from top left: lowland Madagascar rainforest canopy; vines and canopy emergent in Corcovado National Park, Costa Rica; slash and burn agriculture in eastern Madagascar; deforestation in eastern Madagascar; giant epiphytes in eastern Madagascar; stream in primary rainforest, eastern Madagascar. Photos taken by the author.
The second-largest area of rainforest in the world occurs across tropical Asia. Centered on the Malay Archipelago, the land area included covers about $2.5 \times 10^6\text{ km}^2$. This includes Indonesia, whose total forest area is second only to Brazil. The Malay Archipelago, also called Malesia, extends northward to continental Southeast Asia, Burma, Thailand, and Indochina. The extent to which tropical rainforests extend into the southern Himalayas in upper Burma, Assam, and southern China is under debate as there is not a clear line. At the southern end of the boundary are the coastal rainforests of Queensland, Australia.

Africa has the smallest land area of tropical rainforest: about $1.8 \times 10^6\text{ km}^2$, centered on the Congo River Basin (Whitmore 1998). Cameroon and the Central African Republic are located at the northern limits of tropical rainforest in this region. Gabon, Equatorial Guinea, Congo, and the Atlantic Ocean comprise the westernmost boundary. The Democratic Republic of the Congo has the largest land area of African tropical rainforests. To the east, Madagascar and the associated islands of Reunion, Mauritius, the Comoros, and the Mascarenes are located at the eastern edge of the region’s rainforests. These rainforests have been extensively destroyed over the past 40 years, with
TROPICAL RAINFOREST ECOSYSTEMS

only 7–8% of intact forest remaining (Green and Sussman 1990).

Though tropical rainforests now occur in these three regions, this was not always the case. Paleo scientists believe that tropical rainforests began evolving prior to continental drift. Since all present-day tropical rainforests lie on lands that once made up Gondwanaland, biogeographers highlight the significant role continental drift played in determining the ranges and patterns of species distribution among tropical rainforests. With the major evolution of flowering plants occurring before Gondwanaland began to drift apart, today upwards of 334 genera (59 families) of flowering plants are essentially pantropical (Whitmore 1998). The timing of the shift in continental locations also played a role in speciation of tropical rainforests. Madagascar, for example, was separated from Africa about 160 mya and remained part of present-day India until about 80–90 mya when it began to drift into its current location. Scientists believe that it is for this reason that Madagascar’s rainforests are more like Malasian rainforests than African rainforests in stature and species richness (Grubb 2003; Armstrong, Shugart, and Fatoyinbo 2011). Madagascar has unparalleled levels of endemism as a result of the long period of isolation that occurred from when it broke off from India and drifted back toward the African continent. Upwards of 85% of floral and faunal species are found nowhere else on planet Earth (Goodman and Benstead 2003).

Similarities in flora that occur between the three regions of tropical rainforest do so because they were all once part of Gondwanaland. High levels of species diversity are thought to have evolved across tropical rainforests in part because the tropics did not experience the mass extinctions that occurred in ecosystems at higher latitudes during periodic ice ages. The location of the landmasses that support present-day rainforests, in addition to the timeline of continental drift, is the source of many of the complexities of floral distribution, evolution, and interaction that have occurred across time and space. Some genera, such as Campnosperma, have a typical Gondwanan distribution, meaning that they are found in present-day Panama, Colombia, Brazil, Madagascar, and the Seychelles, and from India to Micronesia. Their absence from continental Africa is thought to be due to mass extinctions that occurred during periods of drying out that happened in more recent geologic history. Climate influences have given rise to both distinctions (like that of Campnosperma) that exist between rainforest regions and similarities of species distributions along climatic gradients within the regions. Geography, continental drift, climate, and local environmental gradients are all important factors when relating to tropical rainforests worldwide.

Characteristics of the tropical rainforest ecosystem

An important universal feature of tropical rainforests is that they have a multistory canopy structure. Unlike forests at higher latitudes, rainforest tree species vary significantly with respect to maximum potential height. Most tropical rainforests are comprised of three to four canopy layers including: the canopy emergents, the main canopy, the understory, and the forest floor (the latter two are considered together in some forests). However, multiple variations in layering can exist and some forests may have as many as five or six distinct canopy layers, while others have as few as three. The significance of distinct layers is that the vertical complexity gives rise to a larger variation in microenvironments or habitats than is found in temperate or boreal forests. Spatial complexity in both the vertical
and the horizontal directions leads to an increase in diversity of niche space, resulting in high levels of species biodiversity.

Tropical rainforests are known to have some of the most nutrient-poor soils on planet Earth. While this may seem counterintuitive with regard to the abundance of vegetation types that are supported by these soils, it follows that the majority of nutrients found within these ecosystems are stored in the vegetation. As the tropics were not subjected to the glaciation that occurred during the ice ages, tropical rainforest soils have been subjected to millions of years of weathering. In addition, heavy rains have leached most of the nutrients out of the soil. Furthermore, nutrients that are present in these soils are often in forms that are not accessible by plants. High humidity and warm year-round temperatures result in rapid decomposition rates. Thus, soils of tropical rainforests contain significantly less organic matter than those of temperate forests. The nutrients that are released into the soil by rapid decomposition following a treefall are quickly taken up by nearby trees.

Ecologists have long recognized soils as an important feature in rainforest environments. As noted above, nutrient cycling in forested ecosystems is characterized by a set of direct and indirect feedback mechanisms between soil and vegetation (Brand and Pfund 1998; Castellanos et al. 2001). Biogeochemical cycles in forests involving successive transfers of nutrients among compartments or pools (i.e., canopy, litter, humus, soils, etc.) have been shown to vary greatly depending on soil type (Schlesinger 1997; Vagen, Andrianorofana, and Andrianorofana 2006). Major essential nutrients include: carbon (C), nitrogen (N), phosphorus (P), calcium (Ca), and potassium (K). Iron (Fe), aluminum (Al), magnesium (Mg), and sodium (Na) are considered secondary nutrients that are often important components of cycling of many of the primary nutrients. Recent research on overall patterns of mineral cycling in soils of tropical rainforests clearly indicates that patterns of nutrient cycling in tropical forests are as diverse as they are complex. Forests growing on nutrient-poor, acidic soils — such as those often found in tropical rainforest ecosystems — tend to exhibit tightly closed nutrient cycling in which the forest subsists almost entirely on the stock of nutrients circulating through the soil, biomass, and litter (Richards 1996). Soil texture has been shown to strongly influence nutrient availability and retention, particularly in highly weathered soils of tropical rainforest ecosystems (Silver et al. 2000). As defined by percentage of sand, silt, and clay, texture can affect the ability of a soil solution to retain carbon, water, and nutrients, thus greatly impacting many hydrologic and biogeochemical processes in forest ecosystems (Jenny 1980; Silver et al. 2000).

Vegetation formation in tropical rainforest

Physical environmental variables including elevation, climate, and soil dictate which types of trees and plants are able to grow, compete, and adapt to change within different landscapes. Tropical rainforests are a broad category of forests encompassing a range of physiognomic properties with respect to size, stature, density, tree species, and presence of vines and epiphytes. A combination of physical and physiognomic properties dictates how rainforests are subdivided into four rainforest types, including: semi-evergreen rainforest, lowland evergreen rainforest, lower montane rainforest, and montane rain forest. There are an additional nine other types of forests found in the tropics, but as they do not fit the criteria mentioned above, they are not considered to be rainforests.
TROPICAL RAINFOREST ECOSYSTEMS

Tropical lowland rainforests

Tropical lowland rainforests occur in humid climates where water stress is intermittent or absent, and at a range in elevation from sea level to 1200 m. Some of the highest tree species diversity worldwide is found in lowland rainforests (Table 1). Lowland rainforests are characterized by the presence of dense closed canopy stands of evergreen trees in at least three layers, with a canopy often exceeding 30 m and emergents exceeding 45 m in height. The majority of canopy emergent trees are uncommon species that individually do not contribute more than 1% of the total number in a given forest (Whitmore 1998). In addition to a multilayered canopy, a common feature of lowland forests is a lack of plant species diversity; ground vegetation is often sparse and mainly made up of tree seedlings. Some of the largest trees may be deciduous or semi-deciduous, but with staggered leafless periods that do not affect the evergreen nature of the canopy as a whole. Trees with buttresses are common; other structural features such as prop roots may be present but less common. Other common tree features include: cauliflory and ramiflory, tree species with pinnate, mesophyll leaves, woody vines that are both free hanging and bole hugging, and both sun and shade epiphytes (Richards 1996).

In addition, there is a relatively low diversity of large-leaved monocots, a feature that has been commonly attributed to acclimation or adaptation to poor soils and low soil-water supply (Grubb 2003). Palms and pandan species are found in varying levels of presence depending on location. For instance, there are 130 species of palms (Palmae) found in Madagascar lowland rainforests, with 11 of the 19 genera endemic (the seven remaining genera contain one species each). Similarly, the pandan (Pandanaceae) family is well represented with 70 species found mostly in the eastern forests of Madagascar (Goodman and Benstead 2003). This is in contrast to the large areas of tropical East Africa, where there are fewer than 20 species of palms and pandans combined, and they are confined to topographically constrained wet sites (Grubb 2003).

Tropical montane rainforest

Montane rainforests are subdivided into lower and upper forest formations. Occurring between elevations from 1200 m to 1500 m, tropical lower montane rainforests exhibit different structure and physiognomy than lowland rainforests. These forests are characterized by a lower stunted, often paler-colored, but more even canopy. Trees tend to be more slender, with gnarled limbs and dense subcrowns. Buttresses are uncommon or small, cauliflory and ramiflory is rare, leaves are mesophyllic, vines are highly uncommon, but nonvascular and vascular epiphytes are abundant. The canopy height is usually between 15 and 33 m, depending on the topography and soil strata, with very few emergent trees that are up to 37 m tall.

From 1500 m to the treeline (ca. 4000 m), upper montane rainforests ecosystems are distinctly smallest in stature. Canopy height decreases with increasing elevation, and ranges from 1.5–18 m. Emergent trees are not usually present; when they do occur they usually only grow to a maximum of about 26 m tall. In general, the forest is characterized by a smaller dense canopy of microphyll leafed tree species that may be heavily swathed in bryophytes and filmy ferns. Cauliflory is absent, very few or no vines exist, and the occurrences of epiphytes are frequent and abundant.

On small mountains, lowland rainforest neighbors upper montane rainforest. However, on large mountains, lower montane rainforest occurs in between and defines a broad ecotone
between lowland rainforest and upper montane rainforest. The Eastern tropical forests are very mountainous and many studies exist of the extensive zonation of these forests. The Andes also exhibit strong zonation, though less area has been fully characterized and described. The most comprehensive studies of elevational zonation in tropical rainforest types exist for the forests of Jamaica. By contrast, montane rainforest systems are least prevalent in Africa, where they only occur in Cameroon and on the eastern edge of the Congo River Basin (Whitmore 1998).

**Tropical semi-evergreen rainforest**

The African continent has the most extensive seasonally dry or semi-evergreen rainforest on the planet. Semi-evergreen rainforests are tall closed canopy forests with emergents upwards of 45 m in height. The canopy is comprised of a mixture of evergreen and deciduous trees. Though deciduous tree species may make up about 30% of the main canopy tree species, they are not all necessarily leafless at the same time. Species richness is high, but second to that of lowland rainforests. Trees with buttresses and other structural features are common in both evergreen and deciduous tree species. The height of both the canopy and emergents tends to be slightly less than in lowland rainforests. Tree bark tends to be thicker, and cauliflory or ramiflory is rarer. Woody vines and epiphytes with ferns and orchids tend to be very abundant, though varying regionally. In addition, bamboos are often present.

Whereas zones of lowland and montane tropical rainforests are defined by altitude, regular annual periods of moisture stress differentiate semi-evergreen rainforests. Because moisture stress can be influenced by rainfall seasonality,
particular local soil conditions, and microclimate effects, semi-evergreen forest stands can appear in a less ordered degree of zonation than what appears on a mountainside, for instance. Semi-evergreen rainforest occupies the lower Amazon and most of the African forest block, including the entire Congo River Basin. It forms as outlier buffer forests around the edges of many lowland rainforests in the Eastern rainforest block, as well as in India and most of the Australian tropical rainforest. It probably was also the most prevalent rainforest type in continental Southeast Asia, but this area in general has been reduced to almost nonexistence by human activity. Heterogeneously mixed evergreen and semi-evergreen forest stands are thought to comprise the rainforests of the middle Americas. Given the predicted drying out of the tropics as a consequence of global climate change, semi-evergreen forests with their deciduous tree species may prove to be a successful adaptation strategy that will lead to their expansion into previously evergreen rainforests in the future.

Tree species patterns and diversity

The structural complexity of tropical rainforest ecosystems worldwide and the discerning of underlying environmental factors that shape community interactions within them have long been considered important topics of study among ecologists. In many cases, the relationships investigated between environmental conditions and plant communities have provided a basis for explaining vegetation patterns, but mostly at local and occasionally regional scales (Holdridge 1967; Box 1981; Smith and Huston 1989; Clark and Clark 2000; Leigh et al. 2004). In 1982, noted rainforest ecologist T.C. Whitmore asserted “that forests of the world are fundamentally similar, despite great differences in structural complexity and floristic richness, because processes of forest succession and many of the autecological properties of tree species, worked out long ago in the north temperate region, are cosmopolitan. There is a basic similarity of patterns in space and time because the same processes are at work” (Whitmore 1998).

Whitmore’s conclusion, that the fundamental patterns of succession in forests have evolved in a similar manner globally despite their varied development through geographic space and time along broad environmental gradients, has become a widely accepted tenet in ecosystem ecology. Because all trees and plants rely on the process of photosynthesis to grow and survive, they are all constrained by the same basic needs of light, water, and nutrients. Thus, there is a certain amount of homogeneity of growth and competition processes that exist in forests around the world.

In rainforest ecosystems, species richness has given rise to differing competitive strategies among cohorts of trees and fine-scale patterns are often constrained by local environmental processes such as nutrient limitations. High levels of biodiversity and year-round growing seasons cause patterns in gap recovery phases and larger-scale environmental processes to be increasingly difficult to decipher in tropical rainforests. High biodiversity in tropical rainforests is thought to be caused by a combination of the following attributes: (i) the absence of mass extinction events during the ice ages that affected higher latitudes and resulted in forests that are millions of years old; (ii) year-round growing seasons; (iii) the absence of water and temperature limitations; (iv) nutrient limitations that vary spatially and are extreme in some forests. These attributes have resulted in large variations in competitive strategies and niche differentiation.
Biodiversity is measured in numerous ways. Two common ways to quantify species diversity in a given forest are: (i) by calculating the importance value, and (ii) by relating the Shannon–Wiener index with evenness. In the first, species richness and abundance for each species is used to calculate an importance value index (Curtis and McIntosh 1951; Kent and Coker 1994). An importance value index is often used because it is a better measure of biodiversity than frequency and dominance calculations alone. It combines and quantifies the three most commonly used vegetation descriptors.

A second way to enumerate overall species diversity in a given forest site, the Shannon–Weiner index makes the two assumptions that: (i) individuals are randomly sampled from an infinitely large population, and (ii) that all the species from the community are included in the sample (Kent and Coker 1994). In quantifying vegetation descriptors such as species diversity and forest structure variables, scientists attempt to discern patterns of species distribution across forests and relate them to other rainforests. The commonalities that exist among rainforest trees shed light on our understanding of how rainforests have grown and evolved over millions of years, and perhaps give us insight into how they can be conserved and protected in the face of our changing planet.

The imperiled future of tropical rainforests

The importance of tropical rainforest ecosystems with respect to the health of the entire global ecosystem cannot be understated. In addition to the unparalleled species diversity as discussed above, tropical rainforests are in many ways an untapped resource of information about everything from the evolution of genetic diversity and fundamentals of the planet’s past ecosystems to understanding the role this diversity plays in an ecosystem’s resilience in the face of natural disturbance. From the human perspective, the genetic diversity in tropical rainforests remains a reservoir of information about uses for forest products, from medicinal knowledge about plants to high-yield energy foods. Currently, hundreds of species of trees and plants are currently used in thousands of pharmaceuticals worldwide. Information reported by the US National Cancer Institute states that 70% of plants found to be useful in the treatment of cancer are found only in tropical rainforests (www.cancer.gov). Yet, only a fraction of a percent of rainforest species have been analyzed for their potential use in drugs or products.

Human product potential only encapsulates a small amount of the importance of tropical rainforests. By some estimates, as much as 50% of the terrestrial species on planet Earth are found in rainforests, and this figure does not include migratory birds that winter in the tropics. Fundamentally, tropical rainforests help maintain global rain and weather patterns by providing important intermediate water storage; much of the water that evaporates from the trees returns in the form of rainfall. This intermediate storage provides a buffering capacity against floods, droughts, and erosion in these regions where annual rainfall amounts are so high.

At the forefront of the global climate change discussion is the increase in carbon dioxide and greenhouse gases in the Earth’s atmosphere. Rainforests store a significant portion of the Earth’s carbon, while producing a substantial percentage of the planet’s oxygen (Figure 3). Saatchi et al.’s global carbon stock assessment from 2011 estimates that 247 gigatons of carbon – more than 30 years’ worth of global emissions of fossil fuel burning at the current annual rate – are stored in the tropical rainforests.
TROPICAL RAINFOREST ECOSYSTEMS

located across Latin America, Africa, and South-east Asia (Saatchi et al. 2011). Carbon dioxide given off into the atmosphere from deforestation constitutes 20% of total annual anthropogenic greenhouse gas (GHG) emissions.

Yet, as our knowledge of the importance of tropical rainforests has increased over recent decades, so too has the rate of deforestation of these ecosystems on a global scale. Rainforests are in trouble; deforestation worldwide was measured to be 5.44 million ha per year between 2000 and 2005 (Hansen et al. 2008). Deforestation rates have varied regionally across space and time, however. Brazil’s deforestation rate was measured to be 3.1 million ha per year (2000–2005), which amounts to a 1% loss per year. This annual rate was approximately 18% higher than in the previous five years. The worst year of deforestation on record was 2004, when 27 000 km² were lost. Following 2004, rates declined steadily until 2012, when a 28% rise in deforestation area was recorded (BBC 2013). Upwards of 70% of the deforestation occurs as a result of the land being converted to cattle pastures. Previously that percentage was as high as 91% (between 1970 and 2000); however, in recent decades Brazil’s farmers have increasingly converted rainforest to cropland (Barreto et al. 2006; Melillo et al. 1993). Brazil is currently the world’s second-largest producer of soybeans (behind the United States).

The second-largest rainforest in the world, Congo’s rainforest has the lowest human density surrounding the forest as compared to other rainforests worldwide. In addition, the area has been checkered by conflict in recent decades. This instability has insulated the ecosystem from large-scale commercial logging and agriculture, but smaller-scale slash-and-burn agriculture and illegal logging operations persist. Between 1990 and 2000, Congo’s deforestation rate remained relatively constant at about 311000 ha (roughly 0.20%) annually. Since 2000, satellite imagery has shown this rate to decline by one third (FAO 2011). In addition, West African and Madagascan rainforests have shown similar decline in deforestation rate, with Africa’s overall rate dropping from 592000 ha per year in the 1990s to 288000 ha per year during the past decade. Mayaux et al. (2013) note that

![Figure 3](image)

**Figure 3** World biomass map in (tonnes of carbon per hectare). Saatchi et al. 2011. Reproduced by permission of the National Academy of Sciences.
Africa’s rate of deforestation is a product of its circumstances, including “the almost complete absence of agro-industrial scale clearing,” which accounts for more than half of tropical forest loss worldwide.

Indonesia has the third-largest area of tropical rainforest on the planet and the second-highest deforestation rate, of approximately 1.9 million ha annually between 2000 and 2005, and 1.17 million ha in 2008 (Indonesia UN-REDD National Joint Programme Document 2009). With 68% of its landmass classified as forest (131.3 million ha), this deforestation rate is equivalent to a loss of 0.9% each year (Ministry of Forestry 2008). A 2010 report by the National Council on Climate Change suggests that 85% of Indonesia’s GHG emissions stem from land-use activities, with 37% due directly to deforestation (National Research Council 2001). Whereas the majority of deforestation in the Amazon has been for conversion to cattle pasture, and in the Congo for subsistence agriculture, in Indonesia over the past 20 years deforestation has been driven predominantly by agricultural expansion of oil palm plantations (Singh and Bhagwat 2013).

Conservation tools: ecological modeling, remote sensing, and REDD+

The scientific tools available to measure and track rainforest species spatial extent has advanced considerably in the past few decades. With studies on the onset and impact of climate change for tropical rainforest ecosystems ever increasing, so too are the strategies to conserve them. The ability to accurately assess aboveground biomass and associated carbon stocks across tropical rainforest landscapes is a topic at the forefront of current scientific investigation and methodological refinement. Two important tools in the advancement of carbon assessment and monitoring globally are those of high-resolution forest modeling and high-resolution satellite imagery. Ecological models have the ability to incorporate fine-scale in situ vegetation measurements. When these models are run with datasets from high-resolution satellite image data, scientists are developing improved larger-scale estimates of rainforest biomass and reducing uncertainties in our overall understanding of tropical rainforest dynamics.

Ecological forecasting has become an essential tool used by ecologists toward understanding the dynamics of growth and disturbance response in threatened ecosystems such as rainforests. The application of growth and process models for forest ecosystems has been widely studied in temperate forest ecosystems since the 1980s. However, there are some inherent difficulties that exist in the development of rainforest-specific modeling frameworks, including: (i) the immense tree species diversity in tropical rainforests (upwards of 500 species per ha in some forests); (ii) the fact that the ages of trees within a rainforest are often not known and are difficult to measure due to the inability to apply methodologies such as dendrochronology or monospecific plantation measurements and to the absence of long-term datasets; (iii) the fact that the lifespan of tropical trees is very long, as compared to temperate trees, and can vary widely among species with similar light requirements; and (iv) the fact that there is a lack of long-term monitoring data relative to the age of the forests and for existing datasets (the observation period is usually under 20 years) (Whitmore 1998; Richards 1996; Huth and Ditzer 2000).

Individual-based gap models, which simulate the growth and process of every tree, provide the most accurate virtual depiction of a rainforest ecosystem; however, they are often restricted to local spatial scales by processor limitations. Gap models are hierarchical in that the higher-level
patterns observed (i.e., population, community, and ecosystem) are the result of the integration of plant responses to the environmental constraints defined at the level of the individual (Shugart and Woodward 2011). Recent increases in the capability of remote sensors to detect some of the patterns generated by individual tree processes has increased the potential to apply these gap models over larger areas of rainforests through tree demographic averaging. The use of remotely sensed datasets to inform ecological models brings a new level of application to both; models can now be steered by datasets to track actual landscapes (rather than virtual) through time, and modeled data can be used to predict and validate landscape change signatures in areas where satellite images previously had high uncertainty.

Repeat satellite observations of vegetation patterns in two dimensions have revealed much about the state and dynamics of the global biosphere. In tropical rainforests, satellite imagery has provided unprecedented access to previously underexplored areas too remote to monitor extensively. Coarser grain products such as MODIS (Moderate Resolution Imaging Spectroradiometer) and Landsat are used in conjunction with ground-truthing datasets to make accurate maps that classify land cover by forest health and land use. The high periodicity of measurement makes the large-scale tracking of land-use change, deforestation, and biomass flux quantifiable at multiple scales from the 1980s to the present.

The more recent application of higher-resolution datasets such as LiDAR (Light Detection And Ranging) is considered the most accurate and consistent measurement of vegetation height and structure on the scale of the individual tree. The shape of the LiDAR waveform is a function of canopy height and vertical distribution of foliage, trunks, twigs, and branches at varying heights within the LiDAR footprint. A number of studies have shown that LiDAR metrics describing the waveform are proportional to forest biomass over a range of ecosystem types (including high-biomass tropical forests), although biomass accuracy varies with forest types and terrain slopes (Drake et al. 2002; Lefsky et al. 2005; Anderson et al. 2009). In addition, tree height can also be measured with radar data using a technique known as Interferometric Synthetic Aperture Radar (InSAR) (Graham 1954). Scientists can also integrate Landsat, L-band Synthetic Aperture Radar (ALOS/PALSAR), and LiDAR (GLAS/ICESat) to develop a spatially explicit approach to modeling carbon stocks across a gradient of forest disturbance (from mature forest to recently disturbed) (Simard et al. 2006; 2008). Recent advances in remote sensing technology allow viewing the biosphere in three dimensions, and provide refined measurements of horizontal as well as vertical structure of forests.

The availability of satellite imagery, and to a lesser extent data models, has provided an accessible monitoring capability to private entities, corporations, universities, and governments. A combination of this data access, an increasing awareness of the link between greenhouse gas emissions and tropical rainforest deforestation, and the catapulting of the imminence of climate change into the global consciousness by the Intergovernmental Panel on Climate Change’s (IPCC) 2007 report has led to a number of programs that combine environmental science with an economic-based approach toward mitigating rainforest deforestation (IPCC 2007). While still relatively new in both scope and implementation, these programs provide economic incentives to governments and private entities toward conserving rainforest ecosystems worldwide. Born out of the concept that many of the nations that contain the world’s rainforests
TROPICAL RAINFOREST ECOSYSTEMS

are developing and trying to enter the global marketplace, the Reducing Emissions from Deforestation and forest Degradation (REDD) program aims to create an incentive for these countries to protect, better manage, and wisely use their forest resources, ultimately contributing to the global fight against climate change. The program was launched in 2008 and is a multi-organizational collaboration of the Food and Agriculture Organization of the United Nations (FAO), the United Nations Development Programme (UNDP), and the United Nations Environment Programme (UNEP). The UN-REDD program’s goal is to make forests more valuable remaining intact than they would be if cut down. By creating a financial value for the carbon stored in trees, UN-REDD creates a market out of carbon storage. Once the carbon is assessed and quantified through forest inventory and scientific measurement, the final phase of REDD involves developed countries paying developing countries carbon offsets for their standing forests. Further, REDD+ is a climate change mitigation solution that goes beyond UN-REDD strategies to incorporate the role of conservation, sustainable forest management practices, and in some cases enhancement of forest carbon stocks toward reducing greenhouse emissions (www.un-redd.org).

Conclusions

Anthropogenic factors driving deforestation in the tropics are causally linked to the economic livelihoods of the cultures living within them. As noted by the World Bank’s 2011 Economic Prospects report, the majority of the world’s tropical rainforests are located in countries that are classified as “less developed” or “least developed” in terms of gross domestic product (GDP), infrastructure, industrial potential, and other indices (Figure 4). As shown in Figure 4, the majority of the world’s remaining rainforests occur in developing nations, many of which have a majority of their populations living in rural areas where pressure on forest resources is extremely high. With the increasing awareness of the linkages between deforestation, local livelihoods, and in many cases globalization of commodity markets, however, the green activist “Save the rainforest!” campaigns of the 1980s have evolved into a more holistic conservation approach.

Whereas the campaigns of the latter part of the twentieth century were aimed at increasing international awareness, today organizations such as Conservation International (CI) and World Wildlife Fund (WWF) combine a scientific approach in discerning target areas called hotspots with a more collaborative relationship at the local and government levels. In addition, market-based strategies such as REDD+ provide economic incentive-based protection mechanisms to threatened ecosystems. Even so, deforestation rates in the Amazon River Basin, for instance, match those of the 1980s, and government-based programs aimed at protection remain difficult to enforce at both the local and the international level.

Equally important in our understanding of how best to manage and conserve the world’s remaining tropical rainforests is an appreciation of the dynamic relationship that exists between the endemic flora and fauna at the ecosystem level. Comprehensive ecosystem studies that include species inventories, structure, growth, diversity information, and comparisons to other rainforests worldwide provide a foundation for future research toward understanding both floral and faunal patterns of existence and interaction. In addition, scientific research serves to provide a basis of comparison to future studies grounded in wider-scope, cutting-edge techniques such as remote sensing of forest extent, 3-D structure,
TROPICAL RAINFOREST ECOSYSTEMS

Figure 4  This figure depicts the income disparity between developing and developed nations according to annual GDP. The size of the red (developed nations) and blue (developing nations) spheres relates to the remaining forest in 2014 in square kilometers of forest. The annual GDP (in thousands of US dollars) in 2014 for 55 countries is shown with the percentage of the total population living within rural areas for each nation. The blue spheres represent the majority of the world’s tropical forests and include (in order of forest size): Brazil, Democratic Republic of the Congo, Indonesia, Peru, Mexico, Colombia, Angola, Bolivia, Zambia, Venezuela, Tanzania, Mozambique, Papua New Guinea, Myanmar, Gabon, Republic of the Congo, Malaysia, Laos, Guyana, Thailand, Suriname, Zimbabwe, Vietnam, Ecuador, and Madagascar. The red developed nations, only one of which contains tropical rainforest (all others contain other forest types), include: Australia, Austria, Canada, Denmark, Finland, France, Germany, Ireland, Israel, Japan, Netherlands, New Zealand, Norway, Sweden, Switzerland, United Arab Emirates, United Kingdom, United States, Iceland, and Belgium. Data from http://www.worldbank.org/en/publication/global-economic-prospects.

and ground-truthing of forest-wide biomass estimates for inclusion into the emerging field of economic incentive-driven conservation approaches such as that of the UN Framework Convention on Climate Change’s (UNFCCC) Agreement on Reduced Emissions from Deforestation and Degradation (REDD+) and the UN-REDD programs.

SEE ALSO: Biodiversity; Climate change and biogeography; Ecosystem; Land-use/cover change and climate

References


Clark, D.B., and D.A. Clark. 2000. “Landscape Scale Variation and Forest Structure and Biomass in


TROPICAL RAINFOREST ECOSYSTEMS


Schimper, A.F.W. 1898. *Planzengeographie auf Physiologischer Grundlage*. Bonn, Germany: G. Fischer. (Translated to English in 1903 as *Plant Geography upon an Ecological Basis*.)


Tropical savanna ecosystems

Kasturi Devi Kanniah
Universiti Teknologi Malaysia

Jason Beringer
University of Western Australia

Coverage

Savannas are a major component of the world’s vegetation, covering 20% of the land surface in the lowlands of tropical and subtropical latitudinal zones between 30°N and 30°S. They are generally found in a transitional zone between evergreen tropical rainforest and mid-latitude deserts. A large proportion of savannas are located in Africa, with an estimated 15.1 million km², or 50% of the African continent. This is followed by South American savannas, which cover 2.1 million km² and are mostly located in Brazil, Colombia, Venezuela, and Bolivia. Australian tropical savannas are the third-largest tropical savannas, covering nearly 2 million km², or 25% of the country. Savannas are also found in other parts of the world in small pockets, such as in India, Southeast Asia, Central America, and Pacific islands. Some distinct formations of savannas also occur in temperate (Southeast Australian temperate savannas), Mediterranean (oak tree savannas in California), and montane eco-regions.

Definition

Savannas are ecosystems in seasonal tropics comprising a distinctive dual canopy system. The understory vegetation consists of mainly herbaceous cover, whereas overstory plants are comprised of trees and/or shrubs with varying proportions of canopy cover (Hill, Román, and Schaal 2011). Based on overstory tree crown cover, savannas can be classified as (i) savanna woodland (50–80% of crown cover), (ii) open savanna woodland (20–50%), (iii) open savanna (5–20%), (iv) isolated trees (0–5%), and (v) pure grassland (crown cover 0%). Savannas have a distinct composition of trees and grasses, with contrasts in functional types (e.g., C₃ trees and a C₄ herbaceous layer). As such, the trees and grasses compete for water resources and interact with each other in a way that eases water stress (competition/facilitation mechanisms) and allows coexistence and maintenance of the spatial structuring of this ecosystem.

Seasonality of rainfall (pronounced dry and wet seasons in a year), hot temperatures, and high solar radiation determine the structure and vegetation composition of savannas. In the dry season, water deficit in the soil causes the understory vegetation to dry out and senesce, leaving only the overstory trees to contribute to productivity. Increased rainfall in the wet season encourages the growth of herbaceous vegetation that forms a more or less continuous understory canopy cover.

Savannas generally grow in nutrient-poor soils that are: (i) highly leached, often sandy and

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0802
lateritic soils (Oxisols or ferralitic soils), and (ii) base-rich montmorillonitic black clay soils (Vertisols). Generally, these soils are frequently disturbed through fire and herbivory and thus have low cation exchange and very low phosphorus and nitrogen concentrations but are high in iron and aluminum.

**Carbon cycle in savannas**

The carbon budget of savannas consists of the following processes (Figure 1).

1. Gross primary productivity (GPP) is defined as the sum of photosynthesis by all leaves at the ecosystem level (Chapin et al. 2006). CO₂ fixed by trees, shrubs, and grass species in savannas through photosynthesis directly supports savanna growth and produces organic matter. GPP is the primary conduit of carbon transfer from the atmosphere to the plants on land. The energy used in fixing carbon in plants is consumed by animals (grazing) and soil microbes in the ecosystem.

2. Net primary productivity (NPP) refers to the net primary production after accounting for autotrophic respiration (Rₜ). Organic matter produced by plants in savannas through photosynthesis is released back to the atmosphere through Rₜ, which converts carbohydrates into CO₂ and releases energy that is used for growth and maintenance of plants. NPP also refers to the annual plant growth. Globally, the average NPP of savannas is estimated to be 7.2 ± 2.0 tC ha⁻¹ year⁻¹ (average 95% confidence interval) (Grace et al. 2006), which is approximately 64% of the NPP of rainforests (Grace et al. 2006). Savannas account for approximately 66 Pg of the global carbon stored in biomass (13% of all terrestrial carbon).

3. Net ecosystem exchange (NEE)/net ecosystem production (NEP). The instantaneous net exchange of carbon between the atmosphere and the ecosystem is known as NEE and the net accumulation over days or longer is termed NEP. NEE/NEP is calculated as the difference between GPP, Rₜ, and Rₜ (heterotrophic respiration). Rₜ refers to the respiration by nonphotosynthetic organisms such as decomposers (bacteria and fungi) and herbivores. NEP and NEE are widely used as indicators of the amount of carbon accumulated or lost (medium-term storage) by an ecosystem. Generally, root carbon stocks of savannas tend to be slightly higher compared to the above-ground carbon. Soil carbon stocks are high compared to those of the vegetation.

4. Net ecosystem carbon balance (NECB). The amount of carbon remaining after accounting for longer-term disturbance regimes is NECB. Carbon stored in savanna ecosystems is also released to the atmosphere through disturbances such as fire (see section “Fire,” below), herbivory, storms, cyclones, and fluvial transport of dissolved organic carbon (DOC), dissolved inorganic carbon (DIOC), and particulate organic carbon (POC) (see section “Other disturbances,” below). Savanna fires are estimated to contribute CO₂ emissions of between 2 and 4 Pg C year⁻¹ (Bowman et al. 2009). The carbon lost due to disturbance is mostly regained during the subsequent period of plant regrowth, unless the area is converted to pasture or grazing land for cattle (Grace et al. 2006), and savannas are considered to act overall as carbon sinks, taking up an estimated 0.5 Gt C year⁻¹ (Scurlock and Hall 1998).

5. Net biome productivity (NBP). NECB integrated spatially across a biome is termed
TROPICAL SAVANNA ECOSYSTEMS

Figure 1  Schematic diagram showing the typical carbon cycle processes in savannas.

NBP and it is representative of the long-term biome carbon balance. Estimates of NBP show that tropical savannas and grasslands are a sink for carbon.

Factors controlling carbon cycle in tropical savannas

A comprehensive review on the environmental determinants of savanna productivity was given in Kanniah, Beringer, and Hutley (2010) and they identified the major environmental drivers as radiation, rainfall, temperature, and nutrient availability (discussed below) which is then modified by disturbance and longer-term climate change (discussed in the following sections).

Radiation

Radiation is the primary driver of productivity in ecosystems, and savanna regions are typically found in lower latitudes where radiation is plentiful. Nevertheless, in Australian savannas it has been shown, through a land surface model sensitivity analysis of leaf area index (LAI) and soil moisture content inputs, that modeled GPP was limited by light interception rather than water availability. Radiation may be modified by cloud or aerosols (Kanniah, Beringer, and Hutley 2010) from frequent burning, and both may impact on vegetation productivity. Although a reduction in radiation can decrease plant productivity, it has been shown worldwide that the concomitant changes in diffuse radiation can enhance photosynthesis/productivity because diffuse radiation can illuminate light-limited understory vegetation canopy more evenly by penetrating multi-layer canopies (Kanniah et al. 2012; Figure 2).

For savannas, the quantity of solar radiation is more critical for photosynthesis in the wet season, especially when cloud cover is high, although generally this area receives a higher level of radiation compared to many other ecosystems (Kanniah, Beringer, and Hutley 2013a). However, enhanced light-use efficiency (LUE) under cloudy skies can be counterbalanced by the decline in total radiation. In Africa, Le Roux (2006) indicated that solar radiation was potentially a limiting resource for savanna growth, especially at the end of the fast growing season in July and August (wet season) when cloud cover was high.

Rainfall

The intensity, duration, and changes in patterns of rainfall can affect the structure, composition, and primary production of savannas both spatially and seasonally. During the wet season, topsoil moisture is available for any plant species.
Grasses and trees compete for topsoil moisture, but the deeper water is available only to deep rooted species such as the evergreen trees. Herbaceous plants with shallow roots dominate the superficial horizon of soil substrate, and thus have high productivity in the wet season when soil water is high. After the wet season ends, the topsoil water content drops and can remain low during the dry season. The deeper soil water content also drops, but the rate is slower, therefore it maintains higher soil water potential for longer periods in the dry season. In the dry season, soil water deficit causes the herbaceous layer to diminish, but trees with deep roots are able to extract groundwater, and hence become photosynthetically active.

The highly seasonal pattern of available moisture in savannas is reflected in a strong seasonal cycle in productivity. During the wet season months of February and March in northern Australian savannas, average NEP has been recorded as 3 gC m$^{-2}$ day$^{-1}$ (Figure 3), when LAI was at its peak of around 2.3 due to the rapid growth of C$_4$ grasses (grass LAI approximately 1). Wet season fluxes (December to April) were responsible for 72\% of the mean annual NEP of 3.6 tC ha$^{-1}$ year$^{-1}$. In the dry season, a net uptake of 1 gC m$^{-2}$ day$^{-1}$ was maintained by the ecosystem, largely due to evergreen trees accessing deep soil water, and hence driving photosynthesis.

The length and intensity of the wet season also determine the structural composition of savanna vegetation spatially. Long-term moisture availability is a critical driver of ecosystem structure and composition. For example, along the North Australian Tropical Transect (NATT) Eucalyptus and deciduous species dominate in the wetter regions (12.50°S, where rainfall was 1622 mm year$^{-1}$) and there is a transition to Acacia spp.
Figure 3  A strong seasonal response to rainfall on the seasonal patterns of savanna gross primary productivity (GPP), net ecosystem exchange (NEE), and respiration. Solid triangles show fire events. Beringer et al. 2007, 7. © John Wiley & Sons, Ltd. Reproduced by permission of John Wiley & Sons, Ltd.

in lower rainfall regions (17.73°S, where mean rainfall was 643 mm year⁻¹). At the same time, the grass stratum changes from tussock grasses with ephemeral leaves to hummock grasses with evergreen and sclerophyllous leaves. In these lower rainfall regions the amounts are smaller, with high interannual variability due to isolated convective storms. Such changes are an evolutionary response that enables vegetation in drier sites to withstand drought via water-conserving strategies. Changes in structure, composition, and water use have a profound effect on the spatial and interannual variation in productivity.

Analysis of carbon accumulation by different vegetation types within the savanna region in the Northern Territory, Australia (Kanniah, Beringer, and Hutley 2013b), showed that closed forest occurring in wet and often fire-proof environments assimilated (GPP) 4–6 times more carbon than grasslands and Acacia woodlands that grew in arid environments (<600 mm annual rainfall). Annual rainfall and interannual variability in rainfall was found to
TROPICAL SAVANNA ECOSYSTEMS

exert a significant influence on GPP for different vegetation types. Arid ecosystems had a higher interannual variation in GPP (>20%) compared to woodlands and forest (5%).

Rainfall also strongly influences woody plant cover of savannas, especially in arid and semi-arid regions that receive mean annual rainfall of <650 mm. In these regions, woody cover and tree-grass coexistence are determined predominantly by rainfall. Although rainfall was found to influence tree basal area (surrogate for woody biomass) of savannas in the continents of Africa and Australia, in South America there was no relationship between rainfall and tree basal area, which could be due to the narrower range of rainfall that savannas occupy.

Rainfall and evapotranspiration together drive the available soil moisture, which is also an important driver of soil respiration and soil carbon storage in arid and semiarid savannas. The arrival of the first rains at the beginning of the wet season can cause savannas to become a strong source of carbon for a short period of time due to high ecosystem respiration (Beringer et al. 2007).

Temperature

Generally, savannas receive plenty of solar radiation and thereby maintain higher annual mean and minimum temperature. In Australian savannas, the daytime temperature is more than 25°C in the wet season; meanwhile, the optimum temperature for savanna photosynthesis (tree species) in southern Africa is 30°C. Nevertheless, high temperatures due to climate change may decrease the productivity of savannas. For instance, at Mongu, Africa, the daytime canopy carbon flux of Kataba forest was seen to be reduced by 50% when air temperature was above 27°C (Scanlon and Albertson 2004).

Temperature can also influence fire in savannas. At warmer sites, such as in Australia and Africa, temperature can cure fuels faster, thus facilitating frequent fire. High temperature can also promote the photosynthesis and growth of C₄ pathway grasses and thereby large accumulation of biomass for fire.

Nutrient availability

Savanna ecosystems are largely driven by the interactions between water and nutrients (nitrogen, phosphorus, and calcium). Due to the low fertility of Australian soils, many native trees such as Eucalyptus have an array of nutrient-conserving mechanisms, such as the ability to extract almost all nutrients from dying leaves. Consequently, there is a low decomposition rate of litter. In South American savannas the acidic and infertile soils constrain the distribution and structure of savanna vegetation.

Nutrients are cycled by termites in savannas, especially in the dry season when microbial decomposition almost stops. In African savannas, nitrogen and clay content in the soil affect woody plant cover negatively, probably due to increased growth of the grass layer under increased levels of nitrogen availability. This has significant implications for savanna structure under future environmental change (i.e., nitrogen deposition), where woody savannas may change into grassy savannas.

Disturbances

Fire

Fire and land clearing are major natural (lightning in late dry seasons) and/or anthropogenic (prescribed burning) activities in savanna landscapes that can modify the structure and physiological function of savannas. In Australian savannas, up to 75% of the landscape burns annually.
The direct impact of fire on the savanna carbon cycle includes the emission of CO$_2$ (including trace gases) from biomass burning. Over the period 1997–2013, Australian tropical savanna fires have been estimated to contribute 0.088 Pg C year$^{-1}$ to this total (through all gases), which represents an average of 0.46 tC ha$^{-1}$ year$^{-1}$ across the entire Australian savanna biome (Beringer et al. 2014; Figure 4).

Indirectly, fire reduces productivity following scorching due to the loss of functional leaf area and the high respiration associated with canopy rebuilding. Fire mainly consumes the understory herbaceous or grassy vegetation and scorches the overstory canopy in the dry season, when the soil water content is reduced dramatically. Consequently, canopy leaf area, canopy photosynthesis, and evapotranspiration are reduced and savannas switch from a small carbon sink into a carbon source following fire.

The effect of fire is most prominent in mesic savannas because increasing moisture availability promotes grass production and tree basal area. Consequently, wetter areas are capable of supporting more frequent, high-intensity fires by providing fuel load. Although moisture can increase tree basal area to a maximum value, fire prevents the accumulation of tree basal area and maintains it at a lower value. Thus, mesic savannas are more likely to be fire dependent (tree-grass coexistence) than savannas in arid regions. The role of fire in reducing savanna woody cover at the global scale is well documented. In African and Australian savannas, continental-scale analysis showed that fire emerged as the second-most important predictor (after rainfall) of woody cover. The effect of fire on savanna vegetation is dependent on many factors, such as the seedling and sapling growth rate, fire resilience traits, and fire intensity. A recent review (Beringer et al. 2014) elucidated the impact of fires on savannas from leaf to ecosystem scale on long-term regional carbon budgets (Figure 4).

Although fire can cause a loss of carbon from savanna, over the longer term it could contribute to soil carbon storage through the production of slowly decomposing black carbon. Grace et al. (2006) estimated an accumulation of black carbon of approximately 0.07 tC ha$^{-1}$ year$^{-1}$ in the soil of savanna ecosystems. If savannas were protected from fire, they could store more carbon in biomass due to the encroachment of woody species, whereas complete fire control can replace savanna species with forest trees. Fire suppression has resulted in an increase in tree density.

Other disturbances

In addition to fire (largest disturbance pathway), savannas are also subjected to carbon loss due to disturbance processes by herbivory (termites, insects, and grazers), extreme storm and cyclone events, and fluvial transport of DOC, DIOC, and POC. In Australian savannas, carbon may be lost due to grazing by native and feral animals, although grazing pressure is generally low.

Climate change impacts on savannas

Atmospheric CO$_2$ concentration

An increase in atmospheric CO$_2$ concentration is expected to enhance woody plant cover in savannas compared to grasses due to their low CO$_2$ saturation levels ($C_3$ photosynthetic pathways). Preliminary findings of the investigation of the impact of CO$_2$ enrichment on the savanna carbon cycle in Australia show a rapid (103%) increase in above-ground grass NPP for 550 ppm CO$_2$ relative to ambient controls in the first full growth season (Stokes et al. 2005). This increase was believed to be a short-term enhancement of growth and water-use efficiency (WUE) in
TROPICAL SAVANNA ECOSYSTEMS

**Figure 4** The net biome productivity (NBP) for Australian savanna (tC ha\(^{-1}\) year\(^{-1}\)). Evapotranspiration (ET) and precipitation (P) are in mm year\(^{-1}\). Uncertainties of the estimates are shown as * low, ** medium, and *** high (based on expert’s opinion). Beringer et al. 2014. Reproduced by permission of John Wiley & Sons, Ltd.

Response to the initial application of the CO\(_2\) treatments. However, long-term studies have revealed limited impacts on soil carbon sequestration in Southern African C\(_4\) grassland (Stokes et al. 2005). Nevertheless, increasing temperatures under high levels of CO\(_2\) may reduce the sink strength of savannas by enhancing heterotrophic respiration (i.e., loss of soil carbon to the atmosphere). In a worst case scenario, tropical forests are expected to be encroached by savannas and grasslands are to be replaced by arid systems as a consequence of warming and reduced WUE.

**Woody thickening**

Woody thickening, the increase in standing biomass of woody species, is a global phenomenon that is most commonly observed in arid and semiarid regions and therefore has been observed in savannas and shrublands. Impacts of woody thickening on biogeochemical cycling and C stocks are regionally substantial and globally significant. Thus the conversion of grasslands to closed woodlands could represent an increase in terrestrial carbon sink. The causes of woody thickening include changes in land-use practice, changes in fire regime and grazing and cessation of clearing and abandonment of formerly managed grass and cropland. Changes in climate and in atmospheric CO\(_2\) concentration may also drive woody thickening.

**Climate change and multiple factors**

The current and future links between climate, biomass, and fire interactions of trees and grasses can be explored using dynamic vegetation models. A study by Scheiter et al. (2014) modeled the effect of precipitation seasonality, fire return interval, and fire timing on carbon stocks under a scenario where, by 2100, CO\(_2\) concentration increased to approximately 700 ppm, air temperature increased by 4°C, and mean annual precipitation increased by 13%. Under that scenario, both tree and grass biomass increased, by 18.6 and 0.7 tC ha\(^{-1}\) year\(^{-1}\), respectively. Factorial simulations that varied climate drivers
showed that most of the increase in tree biomass was due to the CO₂ concentration increase, whereas grass biomass increased due mainly to the temperature increase. Increasing the seasonality of precipitation shortened the growing season and decreased biomass in both trees and grasses, with a feedback to lower fire intensities due to lower fuel loads. Changing fire timing to occur in the early dry season or reducing fire frequency both increased the tree biomass, by 20.3 and 4.9 tC ha⁻¹ year⁻¹, respectively. Increases in tree biomass were due to woody thickening, and competition between trees and grass resulted in slightly decreased grass biomass. The fire intensity increased with longer return times (due to increased fuel loads) but decreased in early dry season fires (due to higher fuel moisture content).

From the existing studies, it is clear that future climate change will have a large impact on the structure and function of savannas. Appropriate management is therefore essential if we are to prevent degradation of these resources. Managing the savannas for the best outcomes for sequestering greenhouse gases is also becoming increasingly important.

**SEE ALSO:** Climate change and biogeography; Disturbance in biogeography; Ecosystem; Grassland ecosystems; Treeline ecotones

**References**


Tunisia: Association des Géographes Tunisiens (AGT) (Association of Tunisian Geographers)

Founded: 1977  
Location of headquarters: Tunis  
Website: www.agt.org.tn  
Membership: 240 (as of 2013)  
President: Ben Boubaker Habib  
Contact: hboubaker@yahoo.fr

Description and purpose

The Association of Tunisian Geographers (ATG) is a scientific association founded in 1977 by young geographers working in university and public departments in the field of land planning and regional development.

The main objectives of the association are: the promotion of geographical research and knowledge in Tunisia; the dissemination of geographical knowledge within academic and general public; ensuring the participation of Tunisian geographers in international events; and the promotion of Tunisian geography at the international level.

The ATG’s main initiatives are: editing the periodical “Géographie et Développement”; several events planned per year, including an international congress, Arab congresses, and national conferences and meetings; scientific fieldtrips inside and outside of Tunisia; and workshops on methodologies and techniques in geography.

Journals or major publication series

Geographie et developpement

Current activities or projects

Activities of the association of the last few years were: the 31st International Geographical Congress “IGC-Tunis2008”, held in August 2008 with the main theme as “Building Together our Territories”; the 7th International Olympiads of Geography “IGEO Carthage 2008” held in Tunis in August 2008; the 9th Colloquium of Maghrebine Geography held in Sfax in April 2007; regular meetings of geographers from the Maghreb Region (Algeria, Libya, Mauritania, Morocco and Tunisia), with a main theme of “Dynamics of Coastal Spaces”; 12e Journées Scientifiques du Réseau Télédétection de l’A.U.F.-Tunisie 2010, held in Monastir in November 2010 with the main theme of “Télédétection, Changements Climatiques et Environnements”; and regional geographical days held in Medenine in October 2011 with a main theme of “L’espace local dans les régions arides: Connaissance, Aménagement et Développement.”

Submitted by Adnane Hayder
Turkey: Türk Coğrafya Kurumu (TCK) (Turkish Geographical Society)

Founded: 1941
Location of headquarters: Istanbul
Website: www.tck.org.tr
President: Ahmet Ertek

Description and purpose

The society aims to promote and improve the study and research of geography in Turkey and other places, to facilitate geographical surveys of Turkey using appropriate scientific methods, and to enable Turkish geographers to remain in touch with global geographical development.

Journals or major publication series

Turkish Geographical Review: it is published twice a year and has been granted national reference publication status.

Current activities or projects

The society regularly hosts professional workshops for academicians and training courses for geography teachers as well as annual national geographical congresses at various locations around the country. It organizes geography field excursions for geography teachers and international trips to different continents. In September 2011, the society celebrated its 70th anniversary with a geographical congress in Istanbul where national and international organizations participated. The society participates at international geography Olympiads, such as Koln 2012 and Krakow 2014. TCK is the host organization for the 34th International Geography Congress (IGC) of IGU 2020 in Istanbul, Turkey.

Brief history

In June 1941, the Turkish Ministry of Education pioneered the first formal gathering of Turkish geographers that led to the founding of the Turkish Geographical Society, established as a non-government organization through a decision ratified by the Council of Ministers. The headquarters of TCK was in Ankara from 1943 to 1988, but then relocated to Istanbul. TCK has been a member of the International Geographical Union (IGU) since 1949 where it represents Turkish geographers.

Submitted by Barbaros Gönençgil
Turkey: Türkiye Coğrafyacilar Derneği (Turkish Association of Geographers (TAG))

Founded: 2012
Location of headquarters: Balikesir
Website: www.cd.org.tr
Membership: 170 (as of January 31, 2014)
President: Yılmaz Ari
Contact: yari@balikesir.edu.tr

Description and purpose

The Turkish Association of Geographers (TAG) was established in 2012 by a group of Western-influenced young academics with a view to bringing Turkish academic geography in line with the international geographical world. The main purposes of the association are to improve the status of geography, to disseminate geographic research and knowledge, and professionalize the discipline. Although it is a young nongovernmental organization (NGO), it is the largest geographical association in the country. The association holds annual meetings, educational programs, and runs national and international projects. It publishes the Turkish Journal of Human Geography, Turkish Journal of Physical Geography, and the Turkish Journal of Geographical Education in addition to a biannual bulletin.

Journals or major publication series

Beseri Coğrafya Dergisi (Turkish Journal of Human Geography). www.cd.org.tr/?bcd&changelang=EN

Current activities or projects

The Turkish Association of Geographers coordinates or participates in national and international projects that aim at advancing geographic knowledge (www.cd.org.tr/?projeler), publishing books, newsletters, journals and reports (www.cd.org.tr/?dergiler), and providing training activities for its members (www.cd.org.tr/?kariyer). TAG also organizes annual meetings in different places throughout Turkey where scientific exchange takes place. One of the priorities of TAG is to internationalize Turkish geography through participation in international meetings, developing international projects, and cooperation with international geographical bodies such as the International Geographical Union (IGU), the European Association of Geographers (EUROGEO), and the American Association of Geographers (AAG). The association has just finalized the Piri Reis project (http://pirireis.dicle.edu.tr) which will connect Turkish geographers with European colleagues.

Brief history

As scientific geography initially developed in Turkish universities in the early 1900s with the efforts of first generation of geographers educated in the West or by Western geographers who migrated to Turkey for several reasons. They established academic geography, started geographic research activities in universities, and institutionalized the discipline with a number of
TURKEY

activities. However, the practice of geography did not develop further for decades after that initial brilliant start and the Turkish academic geography was until recently characterized by traditional, descriptive regional geography.

When the Turkish government decided to establish a number of universities in the 1990s, funding was provided for geography students to seek graduate education abroad resulting in a number of students moving to the United States, United Kingdom, and Germany. These students received their graduate degrees and returned to Turkey in the early 2000s to take up posts in these new universities. They, together with some local geographers who were aware of the shortcomings of traditional geographic practice, decided to reform Turkish geography and the establishment of the TAG was one of the results of that reformation movement. The TAG works to bring contemporary approaches and issues into the agenda of Turkish geographers by meetings, conferences, workshops, educational programs, new books, and journals.

Submitted by Yılmaz Ari
Ukraine: Instytut Heohrafiyi NAN Ukrayiny (BGN/PCGN) (Institute of Geography, National Academy of Sciences of Ukraine)

Founded: 1964/1991
Location of headquarters: Kiev
Website: http://igu.org.ua
Membership: 91 (as of January 1, 2014)
President: Leonid Rudenko
Contact: L.G.Rudenko@rambler.ru

Description and purpose

The institute is the leading Ukrainian institution that provides advanced geographical research. There are six departments in its structure: Landscape Science, Paleogeography, Geomorphology, Human Geography, Nature Management and Sustainable Development, and Cartography.

Its activities include fundamental and applied research; PhD students’ training; development and revision of standards, laws, plans for governmental authorities, and public administration; and conferences. The institute sponsors research in GIS and atlas mapping, sustainable development, neotectonics and morphostructures, paleogeography of the late Cenozoic, landscape research and planning, and urban and regional development. The institute is the head office of the National Committee of Ukrainian Geographers and the Association of Ukrainian Geomorphologists.

Journals or major publication series


Current activities or projects

As an academic organization, the BGN/PCGN focuses on fundamental and applied research, supporting 10 to 20 projects annually. Its current projects focus on issues of modern and paleo landscapes’ assessment, multipurpose GIS mapping and modeling, urban spaces, regional development, spatial planning, and measures of sustainable development. Among the most significant projects in recent years have been the “Atlas of Natural, Technological, Social Hazards and Risks of Emergency in Ukraine,” “Landscape Planning in Ukraine,” and “A Sustainable Development Strategy.” The institute also supports international and national activities through its publications (monographs, Ukrains’kyi Heohrafichnyi Zhurnal), conferences, and scientific exchange programs.

Brief history

The institute was established in 1964 as the Scientific Department of the National Academy of Sciences (NAS) of Ukraine and received its new name – the Instytut Heohrafiyi NAN Ukrayiny – in 1991, although work in the sphere of geographic research had been carried out within the NAS of Ukraine beginning in 1918. During its complicated history the institute and its members have been recognized frequently for their achievements and outstanding accomplishments, including awards, thousands of
UKRAINE

publications, and hundreds of successful projects. In 1993, the institute received its first State Prize in the field of science and technology for the series of monographs “Geographical Foundation of Regional Nature Use in Ukraine.” In 2009 it was awarded its second for the National Atlas of Ukraine, published in 2007. The Atlas is an encyclopedic work prepared in response to a decree of the president of Ukraine to summarize the state-of-the-art spatial information needed by executive and legislative agencies of government to formulate the policy development of the state.

In accordance with its mission – to make geography an important part of administration and public life – BGN/PCGN have been collaborating with similar national and foreign institutions, ministries and governmental agencies of Ukraine, regional authorities, and NGOs.

Submitted by Eugenia A. Maruniak
Ukraine: Ukraine’ke geografichne tovarystvo (Ukrainian Geographical Society)

Founded: 1873
Location of headquarters: Kiev
Website: www.geokyiv.org
Membership: 3300 (as of April 2013)
President: Yaroslav Oliynik
Contact: geokyiv@gmail.com

Description and purpose

The Ukrainian Geographical Society is a nonprofit scientific organization founded to encourage the organization, development, and coordination of geographical research; the use of this research to solve problems of socioeconomic development; to optimize human–environment interactions, encourage geo-ecological education, promote geographical knowledge; and encourage interest in geography among young people. Members include scholars, educators, researchers, students, and employees of government, business, and nonprofit organizations. The society has regional divisions and subcommittees. The society sponsors meetings, conferences, and seminars as well as the publication of research papers and periodicals and the organization and support of scientific expeditions.

Journals or major publication series

Ukrayins’kyy geografichnyy zhurnal (Ukrainian geographical journal). http://ukrgeojournal.org.ua/en

Kyyivs’kyy geografichnyy shchorichnyk (Kiev geographical almanac). http://geokyiv.org/vidannya/kijivskij-geografichnij-shchorichnik

Current activities or projects

The society’s main event is held every four years. The 11th annual conference was held in Kiev in 2013 with the theme on “Ukraine: Geography of goals and opportunities,” and with 200 participants.

In recent years the society’s activities have focused mainly on the development and publication of the national atlas of Ukraine; the national collection of nautical maps and the oceanographic atlas of the Black and Azov Seas; the historical atlas of the Ukraine; the comprehensive atlas of Lviv; the completion of a set of local studies (in particular, the publication of a series on the “native places of Ukrainians,” “anthropogenic landscapes of Podillia,” etc.); the promotion and improvement of the teaching of geography courses; personnel training and retraining; the celebration of the anniversaries of prominent geographers (N. Miklouho-Maclay, P. Chubynskyy, etc.); the submission of proposals to state and local governments for the development of geographical science and education, regional policy and administrative reform, the development of tourism and resorts, and so forth; and the systematization of geographical terminology.

Brief history

The Ukrainian Geographical Society traces its origin to the South-Western Department of Imperial Russian Geographical Society founded in 1873 in Kiev. The Ukrainian branch of
UKRAINE

The Geographical Society of the USSR was organized in 1947. The first meeting of the Society was held in Kiev in 1964. In the 1950s and 1960s it focused on the problems of physical and economic-geographical regionalization, thematic mapping of Ukraine, integrated water management, the impact on atmospheric processes, environmental protection, training of geographical personnel, and improvement of teaching of geography. In the 1970s and 1980s, the issues of the planning of economic development, rational environmental management, and spatial planning were important.

The Geographical Society of Ukrainian SSR was renamed the Ukrainian Geographic Society in 1992. From the 1990s to the present conference themes have been: “Think about the future. Take care of nature,” “Ukraine and global processes: the geographical dimension,” “Geographical problems of sustainable development,” and “Geography in the information society.”

Submitted by Kostyantyn Mezentsev
Uncertain geographic context problem

Stephen A. Matthews
Pennsylvania State University, USA

Across the social, behavioral, and health fields there has been interest in the effects of broader social and geographic contexts on human behavior. Perhaps the best known conceptual framework in the social sciences is Bronfenbrenner’s (1979) ecological systems theory of child development. In his theoretical model, Bronfenbrenner introduced a perspective and a typology consisting of multiple, overlapping, individual, and environmental contexts relevant to a child’s development. That is, variables and processes relevant to the development of the child were seen to operate simultaneously at multiple levels. In refinements to the ecological systems theory, Bronfenbrenner also made more explicit the temporal processes and the variability in the influence of multiple contexts over a person’s lifespan. Similar models and theoretical frameworks that pertain to understanding the contextual effects on individual behavior exist in other fields too. For example, a multilayered perspective on the influence of place – or geographic context – is reflected in the socioecologic model, a model that has emerged as a dominant theoretical framing in the epidemiology and public health fields (Glass and McAtee 2006).

The renewed interest in geographic context has been fueled by the recognition that scientific explanations for the human condition based solely on individual-level data may be insufficient. Health researchers, in particular, have long suspected that where you were born, where you have lived, and where you currently live matters, and the general scientific interest in health effects of place dates back at least over two millennia to the writings of Hippocrates (Macintyre, Ellaway, and Cummins 2002). “Putting people into place” (Entwisle 2007) and utilizing analytical approaches that attempt to shed light on the processes by which place “gets under the skin” (Taylor, Repetti, and Seeman 1997) have led to a proliferation of multilevel studies across the social, demographic, and health sciences. Indeed, open almost any social science, demographic, and health journal within the last decade and the reader will likely find several articles on multilevel modeling applications exploring the relations between individual-level outcomes and attributes of a geographically defined context. Interestingly, the majority of the multilevel studies to date link an individual to a single geographic context, most typically the individual’s current residential neighborhood (variously defined; see below). Moreover, the relative lack of data on residential histories means that many of these studies are by design cross-sectional and as such can only examine the relationship between individual outcomes and the attributes of their current residence or geographic context (Kemp 2011).

It is the use of just one representation of place and the lack of attention to the temporal duration individuals spend within places that may result in either an over- or an underrepresentation of actual exposure to the geographic context or contexts where individuals live, work, and play (Basta, Richmond, and Wiebe 2010; Chaix et al. 2009; Cummins et al. 2007; Matthews 2008; Rainham et al. 2010; Zenk et al. 2011). The potential inability to detect what Diez Roux and

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0599
Mair (2010) refer to as “true causally relevant” geographic contexts is the crux of the uncertain geographic context problem (UGCoP), introduced and labeled by Kwan (2012a) (see also Kwan 2012b).

In the seminal paper on UGCoP, Kwan identified and focused on two sources of contextual uncertainty: (i) the uncertainty associated with the spatial configuration of geographically defined contexts, and (ii) the uncertainty about the timing and duration of exposure to these contexts (expanded on below). Briefly then, the effects of geographic contexts on individual behaviors may be affected by how these geographic contexts are geographically delineated and the extent to which these areal units deviate from the true geographic context relevant to the processes and outcomes being studied. Kwan’s paper is a timely addition to the literature and while only recently published we can anticipate that the UGCoP will help generate more careful decision-making about how researchers across the social and health sciences might measure contextual exposures in space and time.

It is important to note that the conceptual and methodological issues and the challenges associated with the UGCoP can be traced back to critical discussions within several disciplines, including geography (e.g., time-geography) and epidemiology (e.g., the socioecologic model), as well as public health, sociology, demography, and other disciplines that have all embraced place-based multilevel modeling applications (see Glass and McAtee 2006). However, even though the discipline of geography has a long-established literature on fundamental spatial concepts, spatial behavior, and time-geography (Golledge and Stimson 1997; Hägerstrand 1967), and multilevel models and spatial data usage continue to proliferate across the health sciences, there is a clear need to provide improved guidance on spatial thinking, spatial concepts, and spatial theories. As noted earlier, there are several theoretical models of health and society that explicitly incorporate multiple, hierarchical, and nonhierarchical levels, and many of these (e.g., the socioecologic model) are adopted as a theoretical backdrop to the study of say a health or wellbeing outcome. It is ironic, then, that closer scrutiny of such multilevel theoretical frameworks uncovers a disjuncture between the processes being modeled and how these same processes may act in the real world. That is, while socioeconomic models make frequent reference to distal factors such as social structure, social environments, social/political/economic conditions, policy, and environmental resources and constraints, it is rare that the level or geographic context is defined as anything other than a neighborhood or community. Moreover, rarely is the definition of a neighborhood or a community specified. Kwan’s research, arguably more than that of any other researcher, has helped promote greater awareness of the UGCoP and has further reinforced the need to take geographical contexts seriously.

Implications of the UGCoP

The choices researchers make regarding both geographic contexts and exposure to geographic contexts are not benign. Kwan argues that the lack of association between the contextual variables and the individual outcome in a multilevel model might be due to a failure to account for the mismatch between the observed and the true geographic context relevant to the individuals under study. As Kwan states “[T]he spatial uncertainty and dynamics of geographic context associated with the UGCoP greatly complicate any examination of the effect of contextual influences” on individual outcomes (Kwan 2012a, 962). As shown in Figure 1, adapted from
### UNCERTAIN GEOGRAPHIC CONTEXT PROBLEM

<table>
<thead>
<tr>
<th>True contextual effect</th>
<th>Observed contextual effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has effect</td>
<td>Contextual units correct</td>
</tr>
<tr>
<td></td>
<td>Inference correct</td>
</tr>
<tr>
<td>No effect</td>
<td>Contextual units incorrect</td>
</tr>
<tr>
<td></td>
<td>Inference obscured contextual effects</td>
</tr>
<tr>
<td></td>
<td>Type II error (false negatives)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>True contextual effect</th>
<th>Observed contextual effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has effect</td>
<td>Contextual units correct</td>
</tr>
<tr>
<td></td>
<td>Inference spurious associations exist</td>
</tr>
<tr>
<td>No effect</td>
<td>Contextual units correct</td>
</tr>
<tr>
<td></td>
<td>Inference correct</td>
</tr>
</tbody>
</table>

**Figure 1** Inferential errors due to the UGCoP. Source: Adapted from Kwan (2012a) with the permission of the Association of American Geographers, www.aag.org.

Kwan (2012a, 962), being unable to identify and measure the true geographic context might lead to inconsistent results and inferential errors. While the true and observed contextual effects can be in broad alignment, Type I errors (false positives) can arise when the contextual variable of interest has no effect on the outcome variable but a significant, and ultimately spurious, association between them is observed. Similarly, Type II errors (false negatives) will arise when the association between the contextual variable and the individual outcome exists but is not observed. For many research applications the true geographic context is not known (i.e., unobserved) but it has been most often assumed to operate at the local level, often the smallest administrative or aggregate level, at which data are most easily available.

Important questions arise as a result of thinking about the potential for Type I and II errors associated with the UGCoP. Frumkin (2006) among others has suggested several questions that deserve attention: What theoretical model should underlie definition and measurement of place? By what mechanism(s) does place affect health? How far does place extend? How should the relevant qualities of place be defined and measured? Kwan also would likely add questions associated with the measurement of the movement of people across geographic
context and the measurement of the dynamics of geographic context. Ultimately, perhaps the most fundamental question, especially relevant to the goal of generating a better understanding of how place “gets under the skin,” is quite simply, whether all variables of interest, and the processes and mechanisms they represent, operate at the same level, and do so consistently across time. That is, are the processes related to the individual outcomes we seek to understand adequately captured by the definitions of geographic context – or place – that are typically used in our analytical models?

Conventional geographic contexts and the UGCoP

In much empirical research that explicitly analyzes individual outcomes within a geographical context – demographic and health outcomes related to fertility, risky behaviors, morbidity, mortality – the measurement of place is conceptually naïve. The adoption of a naïve view of place is in part because the approach fits the available data and analytical models. Data availability and data format are arguably the most important reasons why the majority of the published empirical research examining the role of geographic context uses the residential neighborhood (typically defined by a postal or census unit; in the United States as a ZIP code or a census tract). This is the conventional approach within sociology, demography, and other social sciences; indeed, there has been somewhat of a fetish for the residential census tract to be used as the single correct scale in the study of neighborhood effects. Interestingly, even though researchers may be studying the same individual outcome of interest (e.g., obesity), other disciplines, especially in the health sciences, are increasingly using alternative definitions of geographical context. In recent years we have seen the emergence of both distance- and network-based buffers being used to define a person-centered or egocentric local residential neighborhood (Chaix et al. 2009). Given the variability in definitions of place, and by extension the measurement of attributes within those places, it is perhaps not surprising to find inconsistent results across studies (Feng et al. 2010). What both the administrative and the egocentric views of geographic context have in common is the use of discrete bounded areas, whether defined as a census tract or a buffer, and as such they reduce individuals to actors who cannot escape the “local trap” (Cummins 2007). The discrete view of the world is reinforced by geographically based contexts and multilevel data structures; these ignore the exposure to places traced out by people as they navigate their complex lives in continuous space.

Functionally useful geographic contexts

Before moving on to discuss the complexity of measuring spatial behavior and contextual exposure, it is important to note that multilevel models are appropriate tools in specific nested and discrete contexts; in Figure 1 this is represented by the two cells in which the contextual units are appropriately defined and as such the inferences are likely to be correct. For example, multilevel modeling in the social sciences grew out of their application to education outcomes, typically studying school children within classrooms within schools. The context of the classroom and the school is perfectly appropriate and also the levels follow a strictly nested relationship. Similarly, others have used multilevel methods to look at economic, social, and political organizations such as when
examining employee performance within firms or individual beliefs within congregations and denominations. Again the higher-level context is relevant to the outcome being investigated.

Moving beyond organization or group contexts it is important to note that appropriate discrete geographic contexts also can be identified. Policymakers and social scientists are often interested in policy environments and as such some geographical contexts make analytical sense. At one extreme, the modern nation-state provides one of the clearest examples. The nation-state has well defined legal boundaries, within which all citizens and visitors are bound to the laws and regulations of the geographic context. Below the nation-state there are jurisdictions also that retain an element of control over local policies and practices (e.g., US states). In the United States, state-level policy can determine the broad parameters associated with access to and provisioning of welfare and health-care services, levels of taxation, or whether same-sex couples can marry, to name but a few. Similarly, it can also be argued that some geographic contexts also have an intuitive appeal vis-à-vis individual behavior. In the latter, one might consider fairly coarser-scale units of analysis such as labor and housing market areas as being appropriate for the study of employment and housing outcomes. Moreover, researchers argue that several finer-scale functionally meaningful units of analysis exist (Siordia and Matthews 2016). For example, counties, minor civil divisions, census-defined urban places, and school districts are in many instances highly salient – or functionally relevant – geographic contexts. It is within these local government-related geographic contexts that both service provision (revenue spending) and financial obligations (taxation) relationships exist between individual residents and the place of residence.

### Contextual exposures in continuous space: short-term mobility

As implied by the UGCoP, a major challenge for researchers is the need for flexibility in allowing for individual exposures across different spatial and temporal scales. The lack of attention to the boundaries of geographic contexts and how they invisibly dissect everyday life is particularly curious given what we already know about movement and mobility in everyday life (e.g., over short timescales, commuting to work and journeys for food shopping typically take the individual out of their residential tract). Individuals’ exposure spaces are typically bimodal in time and space, with clusters of time and activities around specific hubs or nodes (such as home and place of work, or for children, home and school), and some activities cluster around journeys and routes between the main nodes. As research in behavioral geography (Golledge and Stimson 1997) has repeatedly shown us, we live in a continuous world, not one bounded by arbitrary boundaries (Chaix et al. 2009; Matthews 2011). Kwan (2012a), in her discussion of recent literature, also identifies the pervasiveness of nonresidential movements, whether the studies utilize census data, travel surveys, or ethnographic materials. Indeed, in the last decade Kwan has been the leading proponent of activity space research, initially based on the use of transportation survey data (Kwan 2000) then later in mixed-method research and qualitative geographic information systems (GIS) (Kwan and Ding 2008).

In Kwan (2012a) the way to address the UGCoP focused on the use of GPS and wireless technologies to help identify the daily activity space of the individuals in question, that is, to actually measure an individual’s exposure to the multiple places they might visit during a typical day (i.e., both the local residential area and
UNCERTAIN GEOGRAPHIC CONTEXT PROBLEM

extra-local places). This is important as attributes calibrated for experienced exposure areas may be less correlated with each other than at the residential or local neighborhood level, potentially allowing researchers to differentiate among potential mechanisms (Zenk et al. 2011). Kwan (2012a) suggests examining the potential effects of UGCoP by using sensitivity analysis methods to assess how different delineations of contextual units might affect the measurement of contextual variables and study results. Researchers in physical activity research, more so than in geography (see Shoval et al. 2014), are beginning to use GPS data to create person-specific activity space summary measures (i.e., convex hull, standard deviational ellipses, buffers, and kernel density estimation) to define environmental or place-specific exposure. The analysis of activity space environments will allow researchers to think in a more spatially sophisticated manner about the linkages between inequality and the conditions of contextual environments: namely, how the conditions within one’s activity space shape both exposure to risks and access to resources and opportunities.

It is also worth noting that paralleling the focus on objectively measured activity spaces, using GPS, researchers might, where possible, collect information on perceptions of place(s); that is, measuring environmental features within a perceptual space may offer a more accurate and comprehensive characterization of environmental exposures than one based solely on an administrative unit (see Vallée et al. 2010). Within sociology, researchers have observed that individuals living in close proximity to each other and sharing similar socioeconomic and demographic characteristics do not necessarily share a common definition of their residential neighborhood (for a review see Matthews 2008). Interestingly, an individual’s own definition of their residential neighborhood also may change over time as length of residence increases but also across an individual’s lifespan as the utilization of resources and services changes with age; what was once an important feature of the local neighborhood as a child may lose significance during adolescence or as an adult as spatial knowledge and boundaries expand (or even contract).

In the social sciences and other fields, often lacking GPS data, researchers may examine adjacent or proximal census tracts, compare multiple definitions of place, or explore clustering techniques to create homogeneous geographic contexts (see Matthews and Yang 2013). In one of the most well-cited and innovative studies of recent years, Inagami, Cohen, and Finch (2007) examined extra-local effects – in neighborhoods where people shopped – as well as simultaneously examining local residential neighborhood effects. Inagami and colleagues’ work was based on the Los Angeles Families and Neighborhood study which collected geocoded activity location data spanning a narrow range of basic activities and locations.

Contextual exposures in discrete space: long-term mobility

The importance of time and dynamics is relevant for both tracking the individual and for measuring change or dynamics in the attributes of geographic contexts. Kwan (2012a) focuses on the use of GPS to track movement over comparatively short periods of time. The focus on UGCoP (Kwan 2012a) lies mainly with short-term temporal coverage and the collection of data on activity spaces over comparatively short periods of time: periods of up to 7–10 days (occasionally an intervention study might collect GPS data before and after the intervention or perhaps at multiple intervals to examine temporal effects such as seasonal influence on spatial
behavior). One problem with GPS data, perhaps more so in the first generation GPS studies, was the inability to necessarily show that the observation window of a few days captured all of the relevant geographic contexts an individual might visit. As GPS battery and storage capacity have improved, the collection of longer sequences of days, say 14 or more, likely provides a much more complete picture of typical movements and activity spaces. As tracking expands across temporal scales, researchers may need to adapt and think differently about the implications of UGCoP. It is worth noting that the possibility of compiling data on exposomes—a set of exposures experienced over an individual’s lifetime—may not be so farfetched (Wild 2005). Wild seeks better precision in the measurement of individual exposures coupled with spatial and temporal precision. This will not be an easy task given known constraints on resources and available data, but scientific practice is changing rapidly and we can begin to envision ever more detailed data being gathered that will facilitate the tracking of human movement and exposure from cradle to grave. In a world of “big data,” “volunteered information” (e.g., Facebook, Twitter), biomarker data, real-time tracking, surveillance and other sensor data, and smartphones, this idea of compiling intensive longitudinal data on individuals over extended time periods is increasingly possible; this is especially the case in studies of movement, networks, and interactions (see González, Hidalgo, and Barabási 2008; Raento, Oulasvirta, and Eagle 2009; Palmer et al. 2013).

While technological developments in wireless tracking devices and new social media place increasing emphasis on fine-scale temporal and spatial units of analysis, it is worth noting that in the social sciences there is a tradition of collecting longitudinal data, including long-running birth cohort studies or longitudinal panels of adolescents, working-age, and older adults. As researchers begin to think about UGCoP they might be well served by other literatures and disciplines in the social sciences with a research tradition that focuses on the individual and on the sequencing of events they experience (e.g., migration histories), the durations of events or exposures to specific circumstances (e.g., job retention, housing tenure, cancer survivorship), and on transitions (e.g., entry into marriage, the labor market, or retirement). Within the United States, the ability to utilize linked administrative records at secure research data centers will also enhance the ability to study exposure to sequences of places over a lifetime or part of a lifetime. Similarly, Scandinavian countries have an established record of collecting detailed migration data associated with all residential moves. In these latter examples, the temporal and spatial specificity may be crude (annual data and at the scale of the housing market) but these types of data structures can inform the calculation of lifetime exposure via the reconstruction of detailed residential histories.

Measuring attributes in discrete and continuous contexts

As stated earlier, it is important to note the tendency to use relatively small geographic units to define context (e.g., census tracts), and in the analytical model, to use variables based on these areas (e.g., the unemployment rate) to examine associations between contextual variables with an individual outcome (e.g., teen pregnancy). However, let us briefly consider the unemployment rate of a census tract. Most fundamentally, the unemployment rate in a specific census tract does not occur in a vacuum and thus embedding of an individual within a single tract disregards the absolute and relative positioning of the
tract and the patterning of unemployment at other higher-level contextual scales. That is, census tracts are embedded in nested and even non-nested larger geographical regions, and the processes or drivers of unemployment can emerge from macro-level factors that may be operating at the level of the labor market areas, which are in turn embedded within regional and national economies. Residential neighborhoods are not isolated islands, rather they are connected to other neighborhoods and they are embedded within larger contexts that typically remain undefined, unmeasured, and unexamined.

Some of the contextual variables we use are more complex than unemployment. Consider race/ethnic segregation, a critical component of any study of racial inequality and racial health disparities. Race/ethnic segregation can be measured in multiple measures, as many as 20 different indexes, grouped into five key dimensions of segregation: evenness, exposure, concentration, centralization, and clustering. Moreover, as in the case of unemployment, we know that segregation exists at multiple scales simultaneously, embedded in local, urban, regional, and national contexts. Race/ethnic segregation is typically calibrated from a single common source, in the United States, from data available from the US Census Bureau. However, other variables can be calculated, and calculated in different ways, from a variety of data sources and for different geographical contexts. A good example of this concerns the myriad of ways in which the food retail environment or food availability can be measured (see NCCOR 2014). Briefly, some data sources provide counts, densities, and other aggregate attributes of food store types (variously defined) for different geographical contexts, typically units such as a ZIP code or census administrative unit. Increasingly, researchers on food environments are also working with point data; the data on food stores comes as a geocoded point. With this format the analyst can aggregate and generate density scores or measures of the diversity of different store types for almost any geographic context they wish, or generate density and/or accessibility surfaces, the latter being yet another data format. In theory, point data enhances flexibility, but the aggregation of such data and manipulation to match areal units also can potentially lead to the calculation of different scores for input into a model.

These are just a few examples of the complexity of variable measurement, how they vary by measure and geographic context, to say nothing of utilizing such data for calculating change scores for examining dynamic contexts.

Moving from monadic contexts to dyadic and polyadic contexts

The patterns of association among different individuals and groups of people with place suggest that the use of restricted definitions of place (i.e., residential neighborhoods) is missing part, and often significant parts, of the narrative on the connections between people and place. If researchers are to better understand social and health outcomes they need also to think more creatively about the complexity and heterogeneity of human mobility, and about individual exposure to risks and access to resources across time and space. This kind of perspective is increasingly necessary as the human scale of activity no longer appears to coincide with the local scale of the residential neighborhood (Cummins 2007; Basta, Richmond, and Wiebe 2010; Zenk et al. 2011). Indeed, one of the potential reasons why neighborhood effects are found to be small may be due to the heterogeneity in spatial behavior of individuals living in the same context (Chaix et al. 2009; Matthews 2011). As where and when people spend their time will vary across individuals, new
notions of context, and by extension contextual exposure, are increasingly likely to be operationalized through individualized measures that allow exposure levels to vary even for individuals within the residential context. Kwan’s UGCoP could not be a clearer statement on this issue.

Place and health are intertwined, but conventional approaches typically adopt somewhat limited views of the complex relationship between people and place, and more specifically places. The UGCoP challenges researchers to move research away from single exposure or monadic perspectives. A dyadic perspective or framework would see researchers focus on an examination of exposure at two primary places, such as work and home, or borrowing from both spatial and multilevel modeling might examine exposure to adjacent geographical contexts (via estimation of, for example, spatially lagged variables) or examine exposure at processes and measures calculated for two higher levels, either nested or non-nested (e.g., individuals within tracts within counties, or individuals within nonhierarchical geographies such as school districts and health service areas; complex multilevel models can handle both nested and non-nested multiple membership models). A more complete perspective on contextual exposure would adopt a polyadic framework, seeking to examine exposures to multiple places. GPS data collection represents one form of data required for a polyadic perspective where the focus is on short-term measurement periods. Similarly, data from a longitudinal study such as a birth cohort also could fit under the polyadic umbrella, as migration histories could be reconstructed to examine the timing, sequence, and duration of residence and thus exposure to multiple contexts across a longer time period: the human lifespan. While there are examples of three-level data structures (e.g., nested predefined units such as people within census tracts within counties), it is still rare that researchers explore linking individuals to multiple, hierarchically nested or hybrid places.

Examining the effects of multiple places, specifically extra-local or nonresidential places, is an emerging area of research that lies at the intersection of multilevel analyses, geospatial data, and spatial analysis. We will inevitably see more use of wireless technologies and tracking devices to capture human spatial behavior, but so too in the next decade we will read about the adoption and use of other analytical techniques that can help model an individual’s exposure to multiple contexts. Nonspatial disciplines will use techniques such as multiple membership models that permit assigning individuals to multiple non-nested contexts, and this will help unpack UGCoP-related issues.

SEE ALSO: Behavioral geography; Migration: internal; Modifiable areal unit problem; Neighborhood; Neighborhood, conceptual; Neighborhoods and health; Spatial context; Time geography and space–time prism

References


UNCERTAIN GEOGRAPHIC CONTEXT PROBLEM


UNCERTAIN GEOGRAPHIC CONTEXT PROBLEM


Further reading

Uncertainty

Stephen C. Guptill

US Geological Survey (Retired)

A digital spatial dataset represents a model of geographic reality. These models incorporate both the locational and the nonlocational aspects of the phenomena of interest. An observation describes a property of these phenomena. However, every observation is subject to some uncertainty. The uncertainty of the observation is the error that is associated with the result or value of that observation. Taken in total, the uncertainty of the observations conveys information about the quality of the dataset.

Although the real world of geographical variation is infinitely complex and often uncertain, it must be represented digitally in a discrete, deterministic manner. In one method of organizing or describing geographic reality we think of the world as a space populated by objects or features of various kinds. Features have attributes that serve to distinguish them from each other. Buildings, bridges, roads, streams, grassland, and counties are examples of features. Any point in the space may be empty or occupied by one or more features. Feature data models use graphic elements (points, lines, and polygons) to describe the geographic location and extent of the phenomenon under study. These features may exist in one combined database, or may be separated according to a theme or variable into a number of layers or classes. Most digital map data are in this form.

Instead of describing discrete geographic objects, geographical phenomena can be represented and organized by recording the locational pattern of one variable over the study area – that is, as fields of data. When the spatial interval of this spatial sampling is regular (e.g., square), the resulting matrix of observations is called a raster data structure. Locational information is embedded in the data structure; each cell of raster data is located at a unique position. The value of an observation at the cell location is recorded for each and every cell in the area. Digital terrain models, digital images, and land-cover maps (derived from remotely sensed data) are typically stored in this structure.

Whether organized as objects or fields, it is the set of observations that hold the information. An observation tells us about a property of a feature of interest at a specific point in time. Observations can be made with measuring instruments resulting in scalar results, such as locations, elevations, or reflectance values. Or observations can be the result of human or algorithmic interpretation of a feature of interest, resulting in nominally valued objects, such as roads, streams, or buildings. The number of observations about an object can be quite large, and may include information about the various characteristics of the feature of interest (e.g., the number of lanes of a highway) and their relationships to each other (e.g., directional flows on streams, or turn restrictions in road networks). However, for every observation there is always a margin of error or doubt.

Axes of uncertainty

Uncertainty in geographic data is measured along at least four different conceptual axes, including space (location), time, theme (content...
UNCERTAINTY

characteristics), and logical consistency. The space axis characterizes positional accuracy; the time axis indicates temporal accuracy and validity; the theme axis reflects semantic ambiguity, delineation ambiguity, and attribute accuracy; and the logical consistency axis concerns conformity to a data model, internal consistency, and completeness. These qualities are not independent, however. Indeed some data quality aspects, such as logical consistency, directly refer to the interrelationships of elements measured on the other three axes. Not all of these quality elements have well-defined quality metrics associated with them. Metrics of semantic accuracy, for example, are not well understood and some of the measures used for other quality elements may be inadequate.

Location

Uncertainty along the space axis usually concerns the accuracy of geographic coordinates in two- or three-dimensional space. These scalar measurements can be collected directly with a wide variety of instruments, such as global positioning system (GPS) receivers (of varying accuracy and precision), surveying instruments, and geodetic recorders. Many of these instruments have been tested and calibrated and the uncertainty values associated with the measurements are known. However, some locational data is not collected directly, but results from a secondary process. For example, the elevation at a particular location may be calculated from a stereopair of aerial photographs. The final elevation value is the result of complex series of photogrammetric calculations that take into account camera positions, lens distortions, and relief displacement, along with other factors. Each of these elements has an uncertainty factor associated with it, and estimates of the total error introduced in the process can be made. In another example, the geographic coordinates of linear geographic features such as roads might be collected by tracing that feature from a paper map or orthophoto image. Uncertainty in the coordinate values can be introduced by the accuracy of the tracing device, the skill of the operator doing the tracing, the locational accuracy of the feature on the source material, and a myriad of other factors. In such cases, it is difficult to estimate the amount of error introduced by each step of the process, and the uncertainty is usually determined by testing the values in the end product.

Time

The time axis appears straightforward; simply record the time when the observation was made. The first source of uncertainty involves the granularity of the time measurement to associate with the observation. GPS timings are measured in nanoseconds, but a date may be sufficient for describing a land-cover condition. The second confounding factor is that a geographic dataset is usually comprised of a collection of objects, each of which may have a separate “observation” time associated with it. The variance of these times could be measured in months or years. Furthermore, the dataset takes time to produce and assemble, and the final product might not be released until some later time. For example, consider what appears to be a seamless, uniform, cloud-free image of a large geographic area, such as a county or state. The question arises as to what time stamp would be provided with this product. This product is a mosaic made up of a number of separate images, collected at different times, processed and merged together before being published. Typically, all that is provided is the date of publication or release. Clearly, this single date does not accurately represent the temporal characteristics of the product. To
compound the problem, a person might trace a road segment into a database using this image. The question then arises as to what temporal stamp is associated with this road segment. A third aspect involves the concept of temporal validity and whether an observation made at time $x$ is still valid or useful at a later time $y$. Often, observations are assumed to be valid until replaced by newer ones: for example, when users get a new edition of a road database for their car navigation system. However, there are many instances using historic or time series data where all of the observations are required, not just the ones that are currently valid. It is incumbent on data producers to associate sufficient temporal attribution and uncertainty measures with the spatial data that they provide. Determining the temporal uncertainty of a dataset after the fact by an independent data quality evaluation is difficult.

Theme

The theme axis reflects the uncertainty with respect to the content of the dataset: what the phenomena being observed are, and what observations are being made on those phenomena. For users to understand the nature of the data there must be a complete specification of the phenomena being described. Comprehensive data specifications include the definition of the observed phenomena as well as definitions of their attributes and attribute values. Consider the simple example of a geographic dataset consisting of towns and temperatures, such as shown on a weather report. First, we need a geographic location for the named town, for example, “Washington, DC,” second, the measurement scale for the temperature value (Fahrenheit or Celsius) is needed, and third, the question must be asked as to whether the temperature value represents the current temperature (in which case we need the time of the reading), or the maximum or minimum value from yesterday, or a prediction for the future temperature. Lacking any specific guidance, a user will make assumptions about each of these items, but clearly there is a degree of uncertainty associated with each of them.

The term “uncertainty,” in a scientific sense, implies that we can quantify the amount of doubt about each of the observations. Along the theme axis, the items of concern are difficult to quantify. It is less about the uncertainty value of a given observation, but rather more about the ambiguity of what phenomena are being observed (semantic ambiguity). This dilemma is particularly acute when describing phenomena modeled as geographic features or objects, and so deciding what is really meant by terms such as “forest,” “marshlands,” “residential land,” or “bridge.” Consider features labeled with a common place term, such as a stream/river (which we define as a body of flowing water), and whether a database of features called “streams” contains or distinguishes the following: a perennial stream (it contains water throughout the year except for infrequent periods of severe drought); an intermittent stream (it contains water for only part of the year, but more than just after rainstorms and at snowmelt); or an ephemeral stream (it contains water only during or after a local rainstorm or heavy snowmelt). Additional source of ambiguity stems from defining the boundaries of the objects being described: for example, where the edge of a wetland or forest is, or the shoreline of a lake. Even “hard-edged” objects, such as a road, have a level of ambiguity (e.g., whether or not the gravel shoulder is part of the road; where the gravel ends and the dirt begins). Clearly, additional information is required for a complete understanding of the data. Hopefully, the definitions and specifications of the data producers
Uncertainty

(which are not always publicly available) match to some degree the expectations of the user.

To help solve this data-definition problem, a feature data specification can be used to provide a detailed definition of each feature and how it is modeled. This technique is used by national mapping agencies and commercial spatial data producers. The feature data specifications define the domain of features that may be represented in a spatial database. For each feature in the feature data specification, a domain of attributes is defined. Each attribute is assigned a value, taken from the domain of values specified by the attribute definition authority. In addition to the definitions, guidance is given on how a given feature is delineated, when it is collected, and how it is represented in the spatial data model. However, user-contributed or volunteered geographic information is usually collected and provided without regard to adherence to a set of specifications and procedures. As a result, this increases the level of ambiguity and uncertainty in datasets comprised of such data.

Logical consistency

Logical consistency issues concern the degree of conformance to a specified data model. For example, geographic features are often represented by point (or node), line (or edge), and polygon (or face) geometric elements. The relationships between the nodes, edges, and faces are called topological relationships. In simple terms, the topological relationships are that edges start and end only at nodes, and that faces are enclosed by a series of edges. Attention is given to neighboring geographical entities when coding edges. For instance, the topological description of each arc includes not only the starting and ending bounding nodes but also the left and right designations of the co-bounding regions. In this way an edge is directed: that is, the first coordinate is that of the start node and the last coordinate that of the end node. Each face record might also include the identification of any interior islands or, if the polygon is itself an interior island, the identification of the single exterior polygon. This data structure not only accommodates the natural pattern of mapped features efficiently, but also provides the information needed to model networks and to facilitate automated editing and manipulation of the database. Computer algorithms can check to see if the data conform to the specified set of topological relationships.

Another aspect of consistency deals with the consistent, complete collection of all instances of the phenomena (and their associated attributes) that are specified in the data model definition—for example, if the data specification includes driveways that serve two or more houses as a part of the database, or whether all instances of such driveways have been collected. Incomplete collections or collections that do not strictly adhere to the collection specifications add a great deal of uncertainty to the quality of the data.

Other specific relationships, termed “semantic relationships,” are used to express interactions that occur between features that are otherwise difficult to represent with locational data, topological relationships, or feature descriptions. Such relationships can include the following: to flow, the relationship that occurs among the two-dimensional features composing a river system; to connect, the relationship that occurs (at junctions) among the two-dimensional transportation features composing a road network; to bound or to be bounded, the relationship that occurs between the perimeter of a city and its city limits; to be composed of, the relationship that occurs between a designated route of travel and the transportation features that compose it, or between the stream and pond features that comprise a watercourse; and to be above, the
relationship that occurs between the features in an overpass or underpass of a freeway interchange. The intent of all of these relationships is to model the interactions between features sufficiently to support geographic information system (GIS) applications and automated cartographic-product-generation requirements. Failure to provide information about these interactions among the data elements could increase the uncertainty level of the information that is provided.

Evaluating uncertainty

To evaluate uncertainty, a variety of evaluation procedures can be used. Data quality evaluation methods are divided into two main classes: direct and indirect. A direct evaluation method evaluates the quality of a dataset based on a comparison of the items within the dataset with reference information. The direct evaluation methods are classified as internal or external. Internal direct data quality evaluation uses only data that reside in the dataset that is being evaluated. For example, topological consistency checking of the road network is an example of an internal data quality evaluation. External direct quality evaluation requires reference data that are external to the dataset being tested. The reference data are accepted as representing the “ground truth.” In this case, an example would be that the positional accuracy of United States Census Bureau TIGER (topologically integrated geographic encoding and referencing) roads might be evaluated by comparing the roads visible in high-resolution orthoimagery, or field personnel with high-precision GPS devices could retrace the road segment. For both external and internal evaluation methods, one of two inspection methods may be used: full inspection or sampling. A “full inspection” tests every item in the population and is most appropriate for small populations or for tests that can be accomplished by automated means. Sampling tests are performed on subsets of the geographic data.

An indirect evaluation method infers or estimates data quality using information about the data, such as lineage information. This type of external knowledge may include, but is not limited to, data quality overview statements or other quality reports on the dataset or data used to produce the dataset. It may be estimated from knowledge about the source, tools, and methods used for capturing the data and evaluated against procedures and specifications that represent the “best practice” for this product. If indirectly evaluated data quality has been reported, it should be accompanied by a description on how it was determined. In some cases it might be misleading or not even possible to report indirectly evaluated data quality as quantitative results. In those cases the data quality may be described in textual form using a descriptive result. As such, it is often subjective. For these reasons, direct evaluation methods are preferred to indirect evaluations.

Data quality evaluation procedures can be applied at various points in a product’s life cycle. At the production stage, the producer may apply quality evaluation procedures as part of quality control. The description of the applied quality evaluation procedures, when used for production quality control, may be reported as lineage metadata including, but not necessarily limited to, the quality evaluation procedures applied, the conformance quality levels established, and the results.

A quality evaluation process might be used upon completion of production to report data quality results. These results may be used to determine whether a dataset conforms to its product specification or not. If the dataset passes inspection (composed of a set of quality
UNCERTAINTY

evaluation procedures), then it is considered to be ready for use. Most data producers apply quality evaluation methods as part of data processing procedures. However, formal documentation of these quality processes and results are usually not available.

Uncertainty in geographic analyses

Geographic information systems utilize spatial data to provide a wide variety of analytical results. However, it is not standard practice to utilize uncertainty information (should it exist) in the algorithms used to generate results: for example, in determining the level of uncertainty that can be placed on a set of routing instructions from a navigation product, or how often the identified road will not exist, an illegal turn will be specified, or an erroneous destination will be provided. Each of these probably results from errors in the underlying data. Data producers should have knowledge about the uncertain levels in their data and could tell the user that there is, say, one chance in a thousand that the driving instructions contain an error.

Many GIS-processing algorithms utilize multiple sets of often disparate data, each with their own uncertainty characteristics. However, the results of GIS analyses have often been presented as definitive to decision-makers (e.g., this is the best site for the municipal water well), with little or no qualification. One would expect that decision-makers would insist on qualifying statements on any results presented to them. The methods for doing this are not well understood. For example, due to the cumulative effects of error, it has been conjectured that, as the number of datasets used in a GIS analysis increases, the accuracy of the result decreases. This runs counter to common practice, where multiple datasets are used with little regard for whether they are contributing more to the information content or to errors in the result.

Formal characterization of spatial data quality

The ISO (International Organization for Standardization) Technical Committee 211 Geographic information/Geomatics (ISO/TC211) has developed a series of standards, the ISO 19100 series, which deals with standardization in the field of digital geographic information. The standards contain a detailed set of terms, definitions, procedures, and organizational hierarchies to describe and report the various aspects of spatial data quality, including measurements and observations. These terms are used in the section that follows to provide a formal characterization of the elements of spatial data quality.

The information on data quality, which is encapsulated in metadata records, can be applied at varying data granularities, from individual feature instances to entire databases. A quality description can be applied to a dataset series, a dataset, or a smaller grouping of data located physically within the dataset sharing common characteristics so that its quality can be evaluated.

The quality of a dataset is described using the following components: data quality elements, data quality sub-elements, and data quality overview elements.

Data quality elements and sub-elements

Data quality elements and sub-elements, where applicable, are used to describe aspects of the quantitative quality of a dataset. The completeness element describes the presence and absence of features, their attributes and relationships. It has sub-elements of: commission – excess data
present in a dataset; and omission – data absent from a dataset.

The logical consistency element characterizes the degree of adherence to logical rules of data structure, attribution, and relationships (data structure can be conceptual, logical, or physical). It has sub-elements of: conceptual consistency – adherence to rules of the conceptual schema; domain consistency – adherence of values to the value domains; format consistency – the degree to which data is stored in accordance with the physical structure of the dataset; and topological consistency – correctness of the explicitly encoded topological characteristics of a dataset.

The positional accuracy element details the accuracy of the position of features. The sub-elements include: absolute or external accuracy – closeness of reported coordinate values to values accepted as or being true; relative or internal accuracy – closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true; and gridded data position accuracy – closeness of gridded data position values to values accepted as or being true.

The temporal accuracy element reports the accuracy of the temporal attributes and temporal relationships of features. It has sub-elements of: accuracy of a time measurement – correctness of the temporal references of an item (reporting of error in time measurement); temporal consistency – correctness of ordered events or sequences, if reported; and temporal validity – validity of data with respect to time.

The thematic accuracy element describes the accuracy of quantitative attributes and the correctness of nonquantitative attributes and of the classifications of features and their relationships. The sub-elements include: classification correctness – comparison of the classes assigned to features or their attributes to a universe of discourse (e.g., ground truth or reference dataset); nonquantitative attribute correctness – correctness of nonquantitative attributes (e.g., correctness of attribute values such as “road name” or “pavement type”); and quantitative attribute accuracy – accuracy of quantitative attributes.

Additional data quality elements may be created to describe a component of the quantitative quality of a dataset not addressed in the ISO 19100 standards.

Data quality overview elements

The data quality overview elements are used to describe the nonquantitative quality of a dataset. The overview elements are: purpose – describes the rationale for creating a dataset and contains information about its intended use; usage – describes the application(s) for which a dataset has been used; and lineage – describes the history of a dataset and, in as much as it is known, recounts the life cycle of a dataset from collection and acquisition through compilation and derivation to its current form.

Lineage may contain two unique components. These are: source information describing the parentage of a dataset; and process step or history information describing a record of events or transformations in the life of a dataset, including the process used to maintain the dataset (whether continuous or periodic), and the lead time.

The ISO 19100 standards can be used when identifying and reporting quality information, evaluating the quality of a dataset, developing product specifications and user requirements, and specifying application schemas. The quality of a dataset is described using a combination of data quality elements and data quality overview elements.

Data quality elements, together with data quality sub-elements and the descriptors of
UNCERTAINTY

data quality sub-elements, describe how well a dataset meets the criteria set forth in its product specification, and provide quantitative quality information about the dataset. Data quality overview elements are critical for assessing the quality of a dataset for a particular application that differs from the intended implementation. Theoretically, a dataset (from the data quality perspective) may be as small as a single feature or feature attribute contained within a larger dataset. When data quality measures are applied to collections or aggregates of individual features, those measures tend to serve as screens or filters. In other words, for example, all of the features in this collection have a positional accuracy greater than 3.5 m, or all of the features have passed a logical consistency check.

The ISO standards recognize that quantitative and nonquantitative information may have associated quality and the analysis of quality information may include a measure of confidence or reliability. This type of information is then recorded in a quality evaluation report.

SEE ALSO: Cartographic modeling; Data model, F-objects and O-fields; Data quality standards; Data structure, raster; Data structure, vector; Geocoding; Geographic information system; Representation: fields; Spatial analysis

Further reading


Underclass theories

Bradley S. Gardener
Middlebury College, USA

The underclass, a concept popularized in the United States in the mid- to late twentieth century by social scientists and journalists, describes persons living in the most severe poverty. Historically, the underclass, referencing the Marxist concept of the *Lumpenproletariat*, is black and urban, and lives in the inner city. Underclass theories are mainstays in US and UK policy circles, and have helped produced a retrenchment of state policies related to public assistance and affirmative action. Originally contrived as a way to identify Americans, regardless of race or geographic location, who lacked class mobility, underclass theories have largely been positioned as a part of a larger discourse that blames the victim for their state of impoverishment.

Underclass theories and debates are enlivened by the following questions: Who are the poorest of the poor? Where do they live? Why do they continue to stay in poverty? And perhaps most importantly, what are the policy solutions for eradicating this kind of poverty? In brief, according to most variations of underclass theory, these answers are found in identifying the poorest of the black poor, and more recently Latinos in the United States (in an international context this can refer to any number of racialized minority groups). Here, if poverty is pathology, the welfare system encourages people not to work. Thus, the best solution is to cut the safety net and force the poor to find legitimate employment. Out of a need to survive, it is reasoned, this poor population will be forced to take the attitudes and norms associated with mainstream society. Only then will they be able to escape from what supposedly traps them, a vicious cycle of poverty.

Early uses of the term underclass are attributed to Swedish social scientist Gunnar Myrdal, who is best known for his studies of race relations and assimilation in the United States. Myrdal first used the term underclass in a series of lectures he made at prominent US college campuses between 1962 and 1963. These lectures were eventually collected into a book called *Challenges to Affluence*. In a society characterized by wealth and social mobility, particularly among white European immigrants and their offspring, Myrdal was concerned about Americans who were relegated to multigenerational poverty (Myrdal 1963).

Myrdal’s focus, like that of many of his successors, would be on employment or waged labor in the formal economy. He concentrated on several categories related to work in his definition of the underclass, which included the unemployed, the underemployed, and the unemployable. Although he identified structural conditions as the primary cause of poverty, he still focused on problematic attitudes around legitimate work, a keystone of future behavioral-centered theories of the underclass.

Around the same time Myrdal lectured about the underclass, Oscar Lewis, an anthropologist and ethnographer, wrote a book about his research with poor Mexican families. While Myrdal focused on the nation as a whole, Lewis studied the behaviors and attitudes of individuals and families living in poverty. His...
book *The Children of Sanchez* demonstrated that culture greatly influenced the cyclical nature of intergenerational poverty (Lewis 1961). His primary point was that the kinds of behaviors people adopt in order to cope with poverty often prevent them from escaping from it. The findings of his research would be debated heavily in the years following its publication. The culture of poverty is very influential to behavioral variations of underclass theory.

In the early 1980s journalists such as Ken Auletta, who first started using underclass in the highbrow magazine *The New Yorker*, further advanced behavioral strands of underclass theory. Drawing comparisons to Jacob Riis, Auletta sought to reveal everyday realities of the underclass (Auletta 1982). His New York research identified the underclass as being black, poor, and urban. Auletta stressed the values of the underclass, distinguishing them from other impoverished people based on what he saw as aberrant or antisocial behavior. Auletta also wrote a book on the subject, *The Underclass*, which would appear side by side with a plethora of similarly themed magazine articles and news programs, most of which made a racialized spectacle out of extreme poverty.

More than any other scholar, William Julius Wilson is credited for framing the debate around the underclass in academic circles (Wilson 1987). His classic book, *The Truly Disadvantaged*, published in 1987, produced hundreds of follow-up studies. Through a sociologist's lens he documented and explained the most concentrated African American poverty in Chicago. He argued that large-scale changes in the economy having to do with the location of jobs and the disappearance of legitimate work were strongly related to the antisocial behaviors of ghetto residents, which included having children out of wedlock, welfare dependency, and criminal activity.

Key to his argument was the increased social and spatial mobility of African Americans in the post-Civil Rights period. He argued that black ghettos suffered a brain drain. Civil Rights legislation allowed the most educated members of the community to migrate into more integrated neighborhoods, leaving those left behind with fewer connections to legitimate work.

Perhaps the most useful concept to come out of Wilson's work for geographers is “spatial mismatch,” a term that continues to be a focus of geographers who do work on poverty. In this case, poverty is partly a function of the geographic distance between jobs and skill sets. As Rust Belt cities deindustrialized, manufacturing jobs that provided opportunities for social mobility disappeared. This theme gave a geographical dimension to the underclass, as Wilson focused on the lack of legitimate job opportunities in areas of concentrated poverty. Feminist geographers have used Wilson's concept to analyze gender, space, and employment.

Political scientist Charles Murray focused on the behavioral side of Wilson's argument. Unlike Wilson, he deemphasized the economic causes of the underclass related to globalization, and instead concentrated on the excesses of the welfare state. In his influential book *Losing Ground: American Social Policy 1950–1980* he created a series of laws about human behavior that prefigured his later work with Richard Herrnstein, *The Bell Curve*, where he argued that intelligence was related to race and geography (Herrnstein and Murray 1994). In *Losing Ground*, Murray argued that the state and what he described as generous welfare programs were responsible for creating and reproducing the underclass (Murray 1984). To eliminate the underclass, he suggested that public assistance programs such as welfare should be severely cut or eliminated. His research and recommendations foreshadowed significant policy changes in the United States and the
United Kingdom. Murray is particularly relevant in the UK context, as he was invited there to do research about poverty and the underclass.

Underclass theories are relevant because of their widespread adoption by policy circles. In academic and political discourse, the underclass became a signifier for a type of place and person, one that was poor, not because of structural reasons, but rather, because of flawed culture or individual behavior. This ideological shift towards individual behavior and responsibility led Prime Minister Margaret Thatcher famously to say, “There is no such thing as society.” Murray’s theory of the underclass prefigures and influences what scholars call the Neoliberal turn.

Around the same time that the underclass concept became popular, actor Ronald Reagan ran for president of the United States. In his campaign speeches he promised to decrease the size of government and punish the undeserving poor. He popularized the welfare queen, a gendered and racialized character who cheated tax payers by spending food stamps on alcohol, buying luxury items with welfare checks, and having kids with multiple men in order to exploit government programs. Reagan’s assault on the undeserving poor, coupled with claims that affirmative action gave jobs and seats in universities to people of color based on skin color, and not merit, provided a wave of support to hamper gains made through the civil rights movement, and the Great Society, a set of programs launched by President Lyndon B. Johnson in the mid-1960s to fight social inequality.

In the United States these debates culminate in a significant restructuring of the welfare system in what is popularly known as the “End of welfare as we know it.” In the early 1990s, Republican politician Newt Gingrich and others conspired to change the way the American public thought about welfare and the role of the state, creating a series of words, expressions, and concepts that were antithetical to leftist solutions to poverty (Peck 1998). Their efforts successfully reconfigured debates around welfare. Being pro-welfare in any capacity became a political faux pas. Along with the support of President Bill Clinton, the welfare system was hampered.

Today, the social science literature on underclass theories reveals two major schools of thought. The structural theory argues that the underclass cannot compete in the formal economy because good jobs for unskilled laborers have relocated from cities to suburbs and overseas due to revolutions in transportation and communication technology, and the automation of work has deskillled workers to create a bifurcation between the rich and poor. Coupled with multiscalar forms of racism, the underclass is trapped in poverty because of forces beyond their control.

Alternatively, the behavioral theory of underclass tells variations of the following narrative. Poor blacks reside in ghettos where they live in opposition to mainstream norms. They engage in irresponsible sexual activities producing single female heads of households, depend on welfare and government assistance for basic needs making it disadvantageous for them to find legitimate work or develop a good work ethic, and lack the ability to think towards the future, practicing instant gratification. These behaviors are part and parcel of a culture that glorifies crime and antiestablishment behavior. This culture reigns supreme and keeps underclass groups in an endless cycle of poverty. Individual behavior and flawed culture are the primary reason why the underclass is not socially mobile.

In both cases there is a geographic or spatial aspect to renditions of the underclass. The underclass lives in highly concentrated urban areas, primarily in Rust Belt cities where formal economic opportunities have diminished. Scholars also use the concept of geographic isolation to explain the persistence of the underclass. Spatial
separation from mainstream norms, education, and jobs creates a potent culture of poverty, a set of behaviors that make it very difficult to escape poverty. In *American Apartheid*, Massey and Denton argue that underclass conditions are produced by high rates of segregation between blacks and whites (Massey and Denton 1993).

For geographers, there is a special emphasis on the spatial strategies that are used to deal with dangerous or at-risk populations, such as those identified as underclass. Kay Anderson’s seminal work on Vancouver’s Chinatown demonstrates how space is used as a tool to manage groups seen as a threat to the health of a city or nation (Anderson 1989).

The underclass and its geographical correlate, the inner city, fall into this paradigm. Critical scholars identify public housing, urban renewal, and the ghettoization or warehousing of African Americans in the United States as spatial strategies for dealing with the underclass. Parks, highways, and bodies of water often naturalize boundaries between urban neighborhoods and highlight the otherness of the people who live in these areas. One spatial strategy used to deal with the underclass is to destroy the neighborhoods where they live. In Chicago, the infamous Robert Taylor Homes were demolished less than 50 years after they were built.

The underclass concept travels well in academic, journalist, and policymaking circles. Its wide and varied use means that the definition of underclass largely depends on the context in which it is deployed. It has been used by social scientists on the left to decry structural racism and inequalities wrought by global capitalism, and on the other hand, by the right, to lobby for the end of welfare. The malleable nature of the underclass concept has allowed it to jump to the international scale, where it has frequently been used to question the behavior of what are perceived to be unruly minority groups. That being said, the scholarly evolution of underclass theories is well documented.

Leftist academics have taken issue with the term underclass, and have attacked the presuppositions on which behavioral versions of underclass theory are based, primarily that deficient behaviors produce poverty. The most direct critique of this underclass theory was from Loïc Wacquant, a French student of William Julius Wilson. He took an ethnographic approach to studying the underclass. Wacquant’s critique of underclass theories mirrored other scholars who saw it as a clumsy term that unfairly grouped all urban African Americans into a negatively valued category. He worried about the implication of placing a diverse group into a narrow category, arguing that poor African Americans in Chicago did not see themselves as part of an underclass. Rather, he saw that academics and state actors were unfairly placing this category upon them from above.

The underclass concept has also been critiqued for not actually referring to a class, lacking a common economic relationship or the capacity for enacting collective social action. Marx’s writing about the *Lumpenproletariat* was a precursor for this argument. In the *Eighteenth Brumaire*, Marx attempted to calculate the political potential of this group, and found little hope (Marx 2005/1852). By most accounts, Marx saw them as a heterogeneous group of thieves, pimps, and ragpickers, who lived off the detritus of the current political economic order.

Another critique of the underclass concept has to do with heavy policing and incarceration. Scholars of the prison industrial complex have long pointed to the devastating effects of removing men from urban communities.

Although debates over underclass theories dimmed in the early 2000s, there has been a shift in the terrain on which it is discussed. The largely urban and black character of the
underclass has been expanded to include Latino groups, particularly undocumented Mexican workers, spawning the social science term “the rainbow underclass.” The expansion of the underclass category reflects anxieties around undocumented workers (Cameron, Cabaniss, and Teixeira-Polt 2012). The location of racialized poverty has changed from the inner city to the border. Although the geography and groups identified as underclass have changed, the link to individual behavior has not. The underclass now also refers to “illegals,” who it is claimed receive undeserved medical and financial benefits of living in the United States, not because they refuse to work, but rather because they do not pay taxes.

Popular talk-show hosts and right-leaning intellectuals have recently taken up underclass theories in reference to the proposed amnesty of undocumented Mexican migrants. Rush Limbaugh, for example, recently said that the Democratic Party wants to give voting privileges to 11 million illegal immigrants because they need a permanent underclass to stay in power. Defending the research of Jason Richwine, who argued in his doctoral dissertation that the IQ of immigrants is below that of white Americans, Pat Buchanan is concerned that Hispanics threaten to become a permanent underclass in the United States because of their lack of intelligence.

To conclude, it is appropriate to talk about indigenous resistance to underclass theories. What have the people that the underclass targets, primarily people of color living in urban areas, done to resist claims that they have a flawed culture?

Scholars see hip hop as a form of indigenous resistance, as it communicates injustices to the world outside underclass areas, and helps solidify locally based resistance against the retrenchment of the welfare state associated with underclass research and policy recommendations.

Currently, a popular hip hop concept is “work.” Although it has different meanings depending on context, one popular iteration has to do with work as drugs. “Pushing work,” is another way of discussing the sale of illicit drugs. In underclass theories, selling drugs is seen as the antithesis of work. In this case, selling drugs is work. The illegitimate becomes legitimate. By reappropriating the concept of work, youth of color are talking back to social scientists and politicians who think they are lazy.

In the black community, the link between behavior and poverty is a highly contested issue. These debates center around the role that parenting, behavior, culture, and role models play in class mobility. Right-leaning African Americans such as Bill Cosby feel that poor African Americans have dishonored the civil rights tradition by not holding up their end of the bargain, engaging in behavior such as buying their kids expensive sneakers, being poor role models, not emphasizing academic achievement enough, and even giving their kids African names. Hip hop artists such as Ice T and scholars such as Michael Eric Dyson have trenchantly critiqued Cosby’s link between behavior and poverty, trying to shift the debate about the underclass towards issues of job availability, economic realities, and institutional racism.

SEE ALSO: Globalization; Poverty; Race and racism; Subculture of poverty; Unemployment and “underclass”

References

UNDERCLASS THEORIES


Further reading

Unemployment and “underclass”

Michael Samers
University of Kentucky, USA

The persistence of severe unemployment in the world’s poorest countries, the last 40 years of stubbornly high unemployment in much of the European Union (EU), and the global repercussions of periodic economic and financial crises have all ensured that unemployment remains high on the agenda of political institutions and publics around the world. Higher unemployment in especially the wealthier countries over the last six years has also seemed to increase concerns about a continual “underclass,” which in the twenty-first century is sometimes referred to as “the precariat.” Unemployment may be an “invention” of economists and others, relative to changing understandings of work since the nineteenth century (Salais, Baverez, and Reynaud 1986), it may also be implicitly coded in negative and moralistic terms, but it affects millions with significant consequences for individuals’ physical and psychological wellbeing. Similarly, the underclass or the precariat may be the creation of social thinkers of various political stripes but both describe conditions and processes that affect countless numbers of people across the globe.

Though unemployment is calculated differently in different countries, and even within countries depending on political winds, it is usually defined by the number of people actively seeking work who cannot find work. If an individual has given up seeking formal work, or is self-employed, then s/he cannot be considered as unemployed. It is typically measured by direct surveys (such as the Current Population Survey in the United States) or by other estimations. Unemployment statistics are often broken down into further categories: by age (e.g., “youth unemployment”) or other social categories (such as disability, ethnicity, gender, or “race”); in spatial terms (e.g., national unemployment or local and regional unemployment); temporally (e.g., temporary unemployment, long-term unemployment, cyclical, seasonal); or by other qualities (e.g., frictional, hidden, and structural). Official unemployment statistics notoriously do not measure the unemployment of undocumented migrants (which is one element of hidden unemployment) although estimates or surveys may fill in the void. In poorer countries, unemployment is not only difficult to measure because of inadequate or heavily manipulated survey and reporting practices, but it may explain very little about people’s actual economic activity, since the extent of informal employment is so considerable.

The reasons cited by economists for increases in unemployment change over time and space. In the 1970s and 1980s, for instance, it became a common argument that slow growth and structural unemployment (“Eurosclerosis”) plagued the European Union, not just because of industrial restructuring, but also because EU labor markets suffered from “labor market inflexibilities” relative to especially the United Kingdom and the United States. In contrast, the United States (and to a less extent the United Kingdom) appeared to produce innumerable low-wage service sector jobs owing significantly to highly flexibilized labor markets. Dunford (1996) argued against what he saw as this
limited “Eurosclerosis” argument, and instead underlined how labor market regulation and the conditions of reproduction shaped different levels of regional productivity and thus employment levels in certain European regions. More generally, explanations for unemployment in the wealthier countries have tended to stress some mixture of “human capital” variables (e.g., levels of education), flawed macro-economic policies, the role of family support systems, welfare policies, and other gender-inflected labor market institutions, as well as racial discrimination and disadvantage. Since the mid-2000s and certainly after 2008, other explanations for high unemployment in at least the wealthier countries have surfaced, including the crisis itself, insufficient nationally oriented mass consumption owing to income inequality and the globalization of consumption; technological innovation in destroying “moderately paid” clerical and manufacturing jobs, international competition (e.g., the debated impact of “offshore outsourcing”), and a heavily disputed “skills gap” or “mismatch” between the skills required by new jobs and the skills of jobseekers.

In poorer countries, there is a significant difference between explanations for unemployment by economists and those of other social scientists, including geographers. For the former, quantitative analyses focus mainly on the effects of trade liberalization, (real) exchange rates, and government policies. For other social scientists, it is more common to explain unemployment by referring to crises (e.g., the Asian financial crisis) or to “neoliberal” structural adjustment policies beginning in the late 1980s – in many ways this terminology reflects a more critical stance that emphasizes the unequal effects of trade liberalization on poorer countries in terms of poverty and unemployment, but also structural adjustment’s impact on declining state resources for antipoverty, pro-employment programs. Furthermore, considerably more attention by noneconomists is accorded to whether declining employment in rural areas and rural to urban migration has an effect on (un-)employment in cities. This is often combined with both critical reflections on the significance of “human capital” (e.g., the inability of formal educational credentials to ensure employment or acceptable employment), and on the differentiated social impacts of unemployment, especially the relationship between gendered work expectations, the division of labor, and unemployment differences between men and women. Finally, over the last 15 years especially, there is far more research on youth unemployment, especially in African and Latin American countries.

Beyond the macro-economic arguments that seek to explain global, macro-regional, or national differences in unemployment, many geographers and sociologists (along with economists) have contributed to understanding subnational geographical differences in unemployment, especially in the wealthier countries. Such explanations are diverse. Some have interrogated the so-called spatial mismatch hypothesis (e.g., Holloway 1996), or neighborhood effects to analyze (un)employment. The spatial mismatch hypothesis refers to the gap between the location of available jobs in the suburbs and the location of the poor and the jobless, especially in the United States. The debate evolved into whether there is in fact a spatial mismatch, a skill mismatch, or even an “automobile mismatch.” Outside the United States, for example, in the United Kingdom, Houston (2005) notes a reinforcing relationship between a skills and a spatial mismatch. The question remains to what extent such debates are still relevant in the United States in particular, with the revival of some central cities. Other explanations for local (un)employment include insufficient or deficient social networks, local labor market institutions, and the need for specific forms of often locally
oriented “cultural capital” at the point of hiring, which some immigrants may lack (Bauder 2006).

In the 1980s, socially and spatially concentrated unemployment and poverty in US “inner cities” led to adoption of the term “underclass.” Some viewed this as a racialized and pejorative term which stereotyped the African American urban poor as “work-shy.” Others believed the term accurately reflected a “class” of people below the “working classes,” chronically or permanently divorced from the world of work. Wilson (1996) argued that the concept could be legitimately mobilized, provided that one viewed the underclass as a structural problem associated with deindustrialization and unemployment in US cities, and not strictly a behavioral problem associated with a particular group of people. Since the 1980s, the seemingly growing mass of long-term unemployed, chronically (under-)employed, or informally employed with diminished social benefits soon gave rise to another concept: the precariat. This frequently amorphous term is not without critics, and recalls the nineteenth-century formulations of Marx, namely the “reserve army of labor” and the lumpenproletariat, or more recently “the socially excluded” (in the EU). The choice of terms is always political, but probably most social scientists agree that the explanations for, and the often dire consequences of, severe unemployment vary over space and time.

SEE ALSO: Gender, work, and employment; Inequality; Labor geography; Underclass theories

References


Further reading


Uneven regional development

Jamie Peck

University of British Columbia, Canada

Uneven development is a foundational concept in political economy, especially its Marxian variants, where it denotes asymmetrical and exploitative relations between classes, companies, and countries, all of which are understood to be endemic, not transitory, features of capitalist development. When referring to the subnational scale of cities and regions, uneven development signifies an out-of-equilibrium economy, characterized by the dynamic coexistence of regionalized growth and localized decline, by “slash and burn” logics of political-economic colonization and abandonment, by unequal interactions and asymmetrical power relations between cores and peripheries, and by qualitatively variegated forms of dis/connection to the matrix of transnational economic development. Specific patterns and forms of uneven regional development are always historically and conjuncturally contingent, but the production of uneven regional development itself, as a generalized condition, is a recurring and indeed systematic phenomenon. Political economists recognize this to be a defining (necessary) feature of capitalist development. Orthodox economists, while they use neither the term nor the concept, have found it increasingly necessary to accommodate the fact of uneven spatial development (albeit in the more depoliticized terms of agglomeration economies and path dependence).

It was Leon Trotsky who first invoked the notion of uneven development, which was central to his understanding of “permanent revolution.” Trotsky rejected simplified stage models of capitalism (and socialism) for their “pedantic schematism,” insisting instead that “[u]nevenness [is] the most general law of the historic process” (Trotsky 2008/1932, 5). This called attention to the complex (and mutually constitutive) relations between leading and lagging economies, connecting the advanced (or accelerated) political-economic development of some regions to the exploitative underdevelopment of others. Ernest Mandel (1979, 34) suggested that this “innovating idea” was later “widely assimilated,” including into more orthodox scholarship, but as Neil Smith and others have since reflected, the notion that (geographical) difference is everywhere is easily trivialized, sometimes purporting to account for practically everything while actually explaining almost nothing. The concept of uneven development, Smith (2006, 180) pointed out, had been “subject to a remarkable lack of serious analysis” in the half-century after its premature inauguration as a “law” of Marxist theory in the 1920s: for all its enigmatic allure, the idea seemed to have “passed into arcana.” Jon Elster (1986, 56), for his part, took the view that the concept belongs to that “class of Marxist notions whose suggestiveness is equaled only by their elusiveness.”

Elster had failed to take account of developments in radical geography, however, where much more was being made of this suggestive concept, especially by those who had since the late 1970s been working to “spatialize” various currents of Marxian theory. Among other things, this line of work was opening up questions of uneven development at scales other than the national. Specifically, the notion of
UNEVEN REGIONAL DEVELOPMENT

uneven urban and regional development, which was treated no more than obliquely in the founding texts, would be the focus of sustained (and productive) attention (see Harvey 2006/1982; Dunford and Perrons 1983; Massey 1995 Smith 2008/1984). This included fundamental work on spatial divisions of labor and the fast-changing geographies of industrial restructuring; on the extra-regional causes of “regional problems” like structural unemployment and long-term economic decline; on gentrification, suburbanization, and metropolitan transformation; and, in more abstract terms, on the contradictory logic of capital switching, devaluation crises, and the dialectics of fixity and motion in the accumulation process.

In economic geography, in particular, the idea that political-economic transformations, capitalist and otherwise, are necessarily and inescapably characterized by uneven spatial development would later acquire something approaching foundational status, even if this was (sometimes) observed no more than implicitly, as a condition of existence for studies of localized economic formations, processes, and practices. Here, the received view that (real-world) economies vary significantly (and sometimes systematically) across space – rather than tending toward a convergent, universal, or transhistorical form – is broadly shared, not to say axiomatic. The widespread utilization, across this field, of urban and regional (i.e., subnational) case studies and units of analysis, of network-based understandings of distantiated-yet-connected nodes and modes of development, and of now-basic concepts such as the spatial division of labor, with its implications for moving landscapes of profitability, employment patterns, social regulation, wealth, and socioeconomic calculation – all can be seen to reflect a generalized ontological presumption of uneven development, and a geographically differentiated world of distinctive “local” economies with constitutive outsides, variously positioned in relation to one another.

This loose consensus around the issue of uneven spatial development (either as an active analytical proposition or as a “background” condition for localized studies) is one that holds across most of the heterodox and eclectically pluralist field economic geography today, although it tends to occupy a more awkward (and relatively marginal) position in orthodox economic theory (as it does, for that matter, in neoliberal policymaking). In the latter, more mainstream discussions, despite considerable evidence to the contrary, there is continuing faith in the universal reach of singular economic “laws” and market forces, in equilibrating tendencies and long-term dynamic efficiencies, and in the role of global integration as a source of (upward) convergence, typically married to a commitment to remove political “obstacles” and institutional “barriers” to what should otherwise be a freely functioning market economy. Empirically, uneven regional development must be acknowledged, of course, but it is understood to reflect a transitory state, linked to inherited (or natural) endowments, which in the long run will be eroded (away) by competitive tendencies for the equalization of factor returns and convergence toward more “balanced” and efficient development. Geographically inclined economists are now exploring the economics of spatial agglomeration and “clustered” forms of (industrial) development, although they tend to do so by relaxing only some of these orthodox assumptions.

In comparison, where they are embedded in political-economic theorizing, notions of uneven regional development are more disruptive, more unsettling; they index conflict and contradiction, inconstancy and inequality, discontinuity and divergence. They also raise a host of political questions, about imperialism and colonialism, about the interconnections
between wealth creation and poverty generation, about the problems of “external control,” about exploitative or “dependent” modes of development, about the causes and consequences of localized “underdevelopment” and the (re)production of “regional problems,” about the role and responsibilities of global cities, about free-riding suburbs and bankrupt inner cities, about unsustainable environmental exploitation, about the distributional outcomes of different growth models or policy paradigms, and about sociospatial equity and global social justice. In contrast to the empty promises of competitive convergence through market exchange and lift-every-boat models of neoliberalized growth, this is anything but a smooth ride.

Unevenness in theory

Critical of Stalin’s notion of “socialism in one country,” Trotsky’s ideas about what he called “uneven and combined development” positioned trajectories of socialist revolution within the context of capitalism’s emergent world system. Contrary to the view that transitions to socialism were more likely to emerge from (political conditions in) “advanced-stage” capitalist countries, propelled by deepening economic contradictions and prompted by the limits of reformism, Trotsky argued that the volatile circumstances of so-called backward countries like Russia were potentially more conducive to revolutionary change. (He saw no common evolutionary pattern, and no universal stage model that could be applied to each and every society.) The progressive but uneven integration of the world capitalist system meant that different countries (or regions) would exhibit qualitatively distinctive trajectories of development, their political and economic prospects being shaped through mutual interdependencies and interactions of various kinds (for example, involving trade, geopolitical relations, technological diffusion, and cultural ties), both with one another and with the emergent totality of the world system, duly reflected in particularized patterns of class power and politics. Nonlinear “leaps” and orthogonal turns are consequently (always) possible, if rarely predictable.

Capitalism … prepares and, in a certain sense, realizes the universality and permanence of man’s [sic] development. By this the repetition of the forms of development by different nations is ruled out. Although compelled to follow after the advanced countries, a backward country does not take things in the same order … Savages throw away their bows and arrows for rifles all at once, without traveling the road that lay between those two weapons in the past. The European colonists in America did not begin history all over again from the beginning … The development of historically backward nations leads necessarily to a peculiar combination of different stages in the historic process. Their development as a whole acquires a planless, complex, combined character. (Trotsky 2008/1932, 4)

These conditions were held to bestow an ironic “privilege of historic backwardness” on some countries with a “slow tempo” of development, in the sense that lagging economies, those “coming along at the tail end of the European nations,” might find themselves in the vanguard of a stage-hopping leap from capitalism to communism (Trotsky 2008/1932, 5, 74). Even if some of Trotsky’s formulations sound crude today, and even if their political context may strike some as idiosyncratic or even anachronistic, they appeal to a sophisticated, dialectical understanding of uneven development as a relational phenomenon. Countries (and regions) have their own, distinctive trajectories and histories, often combining the old with the new, but these are shaped through mutual interaction and
UNEVEN REGIONAL DEVELOPMENT

long-distance connectivities with other places. And they occupy conjuncturally specific positions within (while at the same time recursively shaping) the evolving world system, understood as an emergent totality rather than a pre-formed container, or source of top-down determination.

Trotsky’s first-hand observations on the historical geographies of Russia at the turn of the twentieth century were instrumental in shaping this understanding of uneven and combined development. At the time, the country contained a few islands of relatively advanced industrialization, mostly confined to the big cities (and often linked to various forms of military production, financed by European powers). Yet these were located in a sea of feudal underdevelopment, as Russia remained a largely peasant society, its relatively weak bourgeoisie existing as an adjunct to the Czarist state. This conjuncture was characterized as an unstable “amalgam of archaic with more contemporary forms,” the combination of “backward cultures” and uneven capitalist integration prompting alternating surges and setbacks, festering political-economic tensions, and all manner of contradictory frictions, further stressed by the “whip of external necessity” (Trotsky 2008/1932, 5). Under these conditions, Trotsky reasoned that social transformation would be led not by Russia’s nascent capitalist class, by virtue of its lack of political independence from the autocratic state, but by a revolt originating in the urbanized workforce. (It was the sprawling Putilov works in St Petersburg – a complex of railway, artillery, and steel factories which employed more than 20000 workers – that duly became the epicenter of a wave of strikes, organized by the Petrograd Soviet, that were the seedbed for the Russian Revolution of 1917.)

The Petrograd Soviet retains its place in history, in the pantheon of “urban-based class struggles,” from the Paris Commune of 1871 to the wave of civil-rights, anti-capitalist and antiwar protests in 1968, and from the Prague Spring to Occupy Wall Street (Harvey 2012, 115). These highly uneven geographies of localized insurrection are hardly predictable, needless to say, but neither is their spatiality entirely random. Historically, cities have served as vital basing points for the intertwined territorial (re)organization of capital accumulation, social reproduction, and state capacities; they are also sites where the pressures of intensified growth and restructuring periodically reach boiling point. These concerns have been central to David Harvey’s long-term project of constructing an historical-geographical materialism, grounded in Marxism but at the same time amplifying and elaborating what he sees as the distinctive spatialities of capitalism – uneven development being understood to be, simultaneously, a motor of the system, an expression of some of its most dogged contradictions, and a trigger for both reformist policies and disruptive politics. If the period 1898 to 1917, and the purview of Trotsky’s initial analysis, might be regarded as “the birth of uneven development proper in the global political economy” (Smith 2006, 185), the experiences of the following century would lay to rest the notion that geographical unevenness was merely the disorderly residue of a pre-capitalist past, subject to erosion into historical irrelevance by the modernizing diffusion of market relations. As the work of David Harvey, Neil Smith, and others has demonstrated, (historically) new forms of uneven spatial development were being produced by capitalism itself, layered on top of inherited spatial inequalities, and not by accident or happenstance, but in accordance with the “moving contradictions” of profit-driven accumulation (Harvey 2014).

It was Harvey who first worked out a “capital logic” understanding of uneven development in his Limits to Capital (2006/1982). Here, he
demonstrated that, through the ceaseless ebb and flow of the accumulation process, the mass of fixed capital tends to build up over time, relative to circulating capital, piling up in places in the form of relatively fixed infrastructures, such that the conditions for profitable expansion in one phase may hamper development in the next. In a précis of his argument, Harvey (2014, 155) has explained:

Capital has periodically to break out of the constraints imposed by the world it has constructed. It is in mortal danger of becoming sclerotic. The building of a geographical landscape favourable to capital accumulation in one era becomes, in short, a fetter upon accumulation in the next. Capital therefore has to devalue much of the fixed capital in the existing geographical landscape in order to build a wholly new landscape in a different image ... Capital creates a geographical landscape that meets its needs at one point in time only to have to destroy it at a later point in time to facilitate capital's further expansion and qualitative transformation. Capital unleashes the powers of "creative destruction" upon the land.

Capitalism was birthed into a world of uneven development, in the prolonged transition from feudalism, but it would inhabit and remake this world, increasingly, in its own image. Having originally encountered a "variegated geographical environment ... encompass[ing] great diversity in the munificence of nature and in labour productivity," the forces unleashed by capitalism would "attack, erode, dissolve and transform much of the pre-capitalist economy and culture" (Harvey 2006/1982, 416). These sources of inherited (or "imported") uneven development had been earlier explored by Rosa Luxemburg, who maintained that capitalism was destined to remain contradictorily codependent on its "outside," those extra-capitalist (socioecological) worlds that precede and exceed the accumulation process, while at the same time providing a less than completely manageable source of raw materials, labor, and markets for a crisis-prone accumulation system. Increasingly, imperialism would become the primary means to annex the noncapitalist world, Luxemburg believed, the ravages of which marked out not only the "new frontiers" of production but ultimately the final frontier of capitalism (like Lenin, she saw imperialism as the highest, but also terminal, stage of capitalism). Any "leveling" effects of capitalism would, for Luxemburg, be attenuated by the fact that the accumulation system must live off and through its "noncapitalist surroundings" (Luxemburg, quoted in Smith 2006, 184).

It has been Harvey's distinctive contribution to reveal how uneven geographical development is "actively produced" by forces intrinsic to capitalism, in the sense of being integral to both the logic and the contradictions of the value form: just as the "accumulation of capital and misery go hand in hand, concentrated in space," so also the fastest-growing cities and regions are prone to rising costs, negative externalities, and restive class politics, generating (counter)tendencies for spatial dispersal (Harvey 2006/1982, 416, 418), wherein the nonrepeating cycle continues. On the basis of a searching examination of these interrelated dynamics of capital deepening (in place) and dispersal (through space), Harvey proposed his "third cut" theory of crisis formation, which recognized the systemic normality of relentless restructuring, place-specific devaluation, and localized capital flight, revealing that bust in one location and boom in another are both manifestations of the same macro dynamic.

These are drivers of creative destruction as a geographical process. Capital's insatiable need for "spatial fixes" results in a susceptibility not only to inertia but to stagnation, against which ever more violent "switching crises" seek to effect an exit from despoiled locations, where profit
uneven regional development

and productivity growth had hit their limits, opening up new fronts of accumulation. “Crises build, therefore, through uneven geographical development,” being not so much alleviated as displaced, “mov[ing] capital’s systemic failings around from place to place” (Harvey 2006/1982, 431; emphasis added; 2014, 161). Outward displacement consequently reflects, but does not suspend, inherent crises in the value form; the problems may be moved around, from region to region, resulting in temporary fixes, but the contradictions go all the way up.

If Harvey always regarded his spatialization of Marxian theory in *The Limits to Capital* as “incomplete,” he would credit his student, Neil Smith, with taking the next steps. Smith’s *Uneven Development* sought comprehensively to resuscitate the titular concept, cutting a new theoretical path between a Marxian tradition that had come to rely on underdeveloped (if not bourgeois) conceptions of space, and received approaches in “academic geography,” which acknowledged spatial differences but had a habit, at the same time, of evacuating these of meaningful historical and political content. His would be a sustained theorization of uneven geographical development, not as the fount of deterministic laws, but as a contradictory driver-cum-dynamic of the accumulation process. “Uneven development,” for Smith (2008/1984, 206), was “social inequality blazoned onto the geographical landscape,” a condition that was not simply encountered by an always-expansive capitalism, but which was systematically exploited by and reproduced through capitalist social relations. His analysis entailed, in particular, an understanding of the *scaled* nature of uneven spatial development (from the regional and the intra-local to the global and the international) as a vital dimension of capitalist development. Echoing Harvey, Smith maintained that uneven spatial development could not be reduced to a transitory stage, en route to competitive convergence, as in “flat Earth” versions of the orthodox globalization narrative, and it was also much more than a mere byproduct, or unfortunate side-effect, of capitalist growth. More than just a “container” for the otherwise-standardized operations of the capitalist accumulation process, space and scale were *themselves* seen to be continuously restructured, as arenas of political struggle, and as the means and media for accumulation dynamics themselves. Taking cues from Henri Lefebvre, Smith understood space and scale to be actively (re)produced through the dynamics and contradictions of capitalism.

Smith was writing in the early 1980s, when the perverse logics of capital accumulation, devaluation, and spatial switching were being exposed in especially vivid ways. The “Fordist” model of accumulation, based on the integration of mass production and mass consumption, which had been dominant for three decades after World War II within the “advanced capitalist” countries, had begun to unwind. This threw the gears of industrial growth into reverse, as “capital flight” to low-wage, de-unionized locations in Latin America and Asia, by way of the extension of production chains and the reorganization of multinational capital, left behind deindustrialized regions in new “rustbelts” across the US Midwest and northwestern Europe. Parallel (and inter-related) processes were also at work across the (shifting) scales of metropolitan capitalism, taking the form of suburban economic growth, racialized patterns of inner-city decline, and “frontier” struggles over gentrification as simultaneous moments of displacement and revalorization. These dynamics were taken to represent, at the same time, a manifestation of capital’s never-ending “restructuring of geographical space,” and the historical specificities of a particular moment of crisis, as one pattern was violently displaced by another. “Capitalism is always
transforming space in its own image,” Smith explained, expressing its restlessly dialectical logic in constantly evolving and contradictory forms:

The logic of uneven development derives specifically from the opposed tendencies, inherent in capital, toward the differentiation but simultaneous equalization of the levels and conditions of production. Capital is continually invested in the built environment in order to produce surplus value and expand the basis of capital itself. But equally, capital is continually withdrawn from the built environment so that it can move elsewhere and take advantage of higher profit rates. The spatial immobilization of productive capital in its material form is no more or less a necessity than the perpetual circulation of capital as value. Thus it is possible to see the uneven development of capitalism as the geographical expression of the more fundamental contradiction between use-value and exchange-value. [The resulting pattern] in the landscape is well known: development at one pole and underdevelopment at the other … at a number of spatial scales. (Smith 2008/1984, 208, 6).

In what Smith would disarmingly present as his “seesaw” theory of capitalist dynamics, the movements of profit-seeking capital were seen, simultaneously, to be producing and exploiting the shifting “profit surface,” constantly recalculating, reorganizing, and then flowing in the direction of the highest returns — a phenomenon especially evident at the urban and regional scales. Like a plague of locusts, capital would descend on a place, devour everything of value, and then move on. Meanwhile, “in the process of restoring itself after one plague the region makes itself ripe for another” (Smith 2008/1984, 202).

A parallel concern with the vagaries of profit-led accumulation and abandonment at the regional scale can be found in Doreen Massey’s influential formulation of the spatial division of labor. Here, however, the focus was less on the singular dynamics of capitalism as an integrated system, and more on the variety of restructuring strategies exhibited by different fractions of capital (or sectors, like electronics, finance, and garment manufacturing), along with the diverse array of regional conditions, “problems,” and conjunctures that are both an empirical outcome of, and a causal factor in, these ceaseless geographies. Massey’s conception of the spatial division of labor owed a debt less to Harvey than to a different strain of Marxian scholarship (after Hymer 1972), in which the evolving corporate structure of the multinational (and therefore “multilocalional”) firm found a spatial analogy in the unequal and asymmetrical relations between headquarters regions, back-office sites, and branch-plant locations. More fundamentally, Massey’s approach also spoke to a more intricately “relational” understanding of capitalist spatiality, which took the specificities of contingent variation, in identified localities and regions, as more than empirical particularities, or “noise,” around common structural tendencies. For Massey, spatial divisions of labor were “more than just new patterns, a kind of geographical re-shuffling of the same old pack of cards,” because they entailed much deeper forms of restructuring, along with “whole new sets of relations between activities in different places, new spatial patterns of social organisation, new dimensions of inequality and new relations of dominance and dependence” (Massey 1984, 8).

For Massey, regions occupy unique positions within wider divisions of labor, their political cultures and class and gender profiles reflecting (in “sedimented” form) the cumulative effects of the succession of historical roles that those places have played in different production networks, corporate hierarchies, industries, investment portfolios, and “rounds of accumulation.” Each round of accumulation (or pattern of investment in, say, heavy engineering or business services) is associated with particular, emergent geographies
UNEVEN REGIONAL DEVELOPMENT

of (profitable) production and employment. This framework allows, simultaneously, for the situated analysis of regional economies (in relation to local histories as well as interregional relations) and for a geographically sensitive understanding of sectoral change (in relation to the specificities of technological innovation, employment relations, and managerial strategies in different industries). In one inventive (re)formulation, the spatial division of labor framework is represented not as a pack but as a game of cards (Gregory 1989), where the four suits refer to different rounds of investment, while the hierarchy of (local) functions, from low-skill assembly work through professionalized research and development to managerial control, is indicated by the face value of each card. Each (local) player must work the odds as best they can, with the hands that they are dealt, the succession of hands adding up to what amount to winning or losing runs. On its face, each card is unique, though at the same time its value and role in the game reflect the (complex combination of) relations with the other cards.

Massey positioned her argument against the kind of “two-dimensional” analyses of industrial and employment change found in neoclassical economic geography (where the geography of jobs was correlated with the geography of wages or skills, for example), maintaining that “space is not a passive surface on to which the relations or production are mapped,” and calling instead for the conceptualization and interrogation of what she called the “spatial structures of production,” recognizing the inextricability of the social and the spatial, of production and politics (Massey 1984, 68). She also sought to differentiate the approach in Spatial Divisions of Labour from “mechanistic Marxist’ insensitivity” to particularity, and from styles of analysis that look at the world as if it were merely the pre-determined product of a set of laws and tendencies. Such approaches leave little scope for real conflict and struggle, still less for surprise and setback … [T]he geography of industry is an object of struggle. The world is not simply a product of capital’s requirements. (Massey 1984, 70, 6–7).

Relative to the more abstract treatments of Harvey and Smith, Massey’s involved a closer-focus analysis of shifting class, gender, and employment relations in the British economy, enabling her to draw out the particularities of uneven geographical development associated with (Keynesian) modernization in the 1960s and (neoliberal) monetarism in the 1980s. She was writing, from a position of active political engagement, during a wrenching period of accelerating deindustrialization, resulting in waves of factory closures and massive job losses, especially in the north and west of the country, and culminating in signature political events like the coalminers’ strike of 1984–1985 and the abolition of the municipal-socialist Greater London Council in 1986.

This was also a time, however, when alternate geographies of class relations and new models of economic growth were being forged, which was reflected in the over-layering of new spatial divisions of labor, favoring sectors like business and financial services, and regions like London and the southeast in particular. The “savage harvest” of what David Harvey (2006/1982, 429) had earlier portrayed as capital’s inherent contradictions was there for all to see, but neither the form nor the contradictions of what became known as “restructuring” were being worked out in singular form. This meant that there was not one but a range of localized consequences – for labor unions, for community groups, for economic-development agencies – and not one but a range of potential political openings and opportunities, as well as threats. There were
consequently both political and analytical reasons to take “localities” seriously. Anticipating a theme that Massey (1995, 326) would explore in subsequent work, this attention to the specificities of local and sectoral change was part of an effort to “think the politics of place and uneven development in a different way.”

Problematizing regions

“Region” may be one of the most unstable signifiers in human geography. Around the middle of the twentieth century, the field was practically defined by its preoccupation with regions, and the kind of ideographic descriptions typical of regional geography in which uniquely local characteristics stood for themselves, if they were not being implicitly explained by other (uniquely) local characteristics. The positivist search for “spatial laws” in the 1950s and 1960s represented a reaction against this long-established tradition, in the sway of which peopled places were reduced to spaces of rational economic calculation and cost minimization, as the field was brought closer to neoclassical economics. In human geography, the subsequent turn toward Marxism and then to a wider array of critical social theories, from the 1970s, placed “the region” in question once again, (re)politicizing the regions at the same time (see Dunford and Perrons 1983).

Even in Harvey’s early formulations, regions mostly appeared as epiphenomena of the deeper currents of uneven geographical development under capitalism, being reflected in the relative fixity of the built environment, the territorially of social infrastructures, and so forth. He was not one to deny the play of “human sentiments” in the animation and aggravation of regional (class) conflicts, but his initial concern had been with the “material basis” of these recurrent conflicts, as reflected in the “circulation process of capital itself” (Harvey 2006/1982, 419). As Ray Hudson has argued, however, it is one thing to point out that the production of “problem regions” – like his own home region of the northeast of England, with its history of coalmining, shipbuilding, heavy engineering, and steel production, all in long-term decline – is “inscribed into the inner logic of the capitalist mode of production,” but it is quite another to account for concurrent factors, such as state policy, let alone the array of embedded and acculturated sociopolitical conditions, as expressed and experienced in actual regions.

Although Harvey’s analysis shows that there will inevitably be “problem regions” … because of causal mechanisms that are an integral part of the process of uneven development, it does not and cannot reveal which regions will become “problematic.” This is a contingent matter, as the switch in the position of the North East from “core” to “periphery” in the world economy and the emergence of this aspect of uneven development as a political question exemplifies. (Hudson 1989, ii).

Hudson’s point was not simply that politics can never be reduced to, or “read off from,” the dynamics of uneven capitalist development, no matter how acutely analyzed, but also that the contradictory spatialities of the state exhibited relative autonomy too. The state is not just a manager but also a maker of uneven spatial development, or what has been called “state space” (see Brenner 2004). Even if, as a matter of methodological principle, “chaotic conceptions” like the region should not be objectified, even if the meaning and significance of regions should be emergent from (rather than presupposed by) a theorization of uneven geographical development, they are nevertheless subject to objectification by state policymakers and by other political actors. Clearly, the state was acting in, and on, “problem regions” like Hudson’s
UNEVEN REGIONAL DEVELOPMENT

northeast, the vacillating fortunes of which have been profoundly shaped by a barrage of regional policies and spatially targeted interventions.

These questions around the politics of uneven regional development, state policy, and political responsibility had earlier been posed most emphatically by Doreen Massey in her seminal article, “In What Sense a Regional Problem?” It was here that Massey first introduced the concept of the spatial division of labor, “in order to make a point,” the point being that problems of capital should not be (mis)represented as problems of particular regions, due to workforce deficiencies of some lack of entrepreneurial spirit: historically, since capital accumulation has always involved the “opening up of some areas, and the desertion of others [then] the ‘regional problem’ is not a problem produced by regions, but by the organisation of production itself” (Massey 1979, 234, 242–243). “Solving” the regional problem, then, could not be reduced to a question of spatial distribution or some technical matter of building a better policy mousetrap, in the manner of Keynesian reformism, since regionalized unemployment had deeper roots in aggregate demand deficiency, technical change and the deskilling of labor, the intensification of international competition, and the restructuring of multinational production systems.

In what sense, in other words, was this a regional problem? The commonplace tactic of blaming (problem) regions for their own plight confuses regionalized effects with endogenous causes, evading (rarely innocently) the real answer to Massey’s rhetorical question, “who pays?” Then there are the related questions of who benefits, and where? As subsequent analyses of regressive social and spatial redistribution under Thatcherite neoliberalism and its stealth program of “privatized Keynesianism” would reveal, there were arguably more compelling reasons to characterize the affluent southeast of England as “England’s problem region” (John, Musson and Tickell 2002), in the light of its ill-regulated financialization, its overheating housing, commercial, and labor markets, and its overburdened infrastructures and governance failures (Peck and Tickell 1995; Allen, Massey, and Cochrane 1998). The privileging of regions like the southeast of England was also sapping the developmental potential of the English provinces, and its so-called problem regions.

In few places are these questions of the costs and consequences of regionalized economic growth – not to say the contradictions of uneven regional development – being played out more dramatically than in contemporary China. It was Mao who said that “[n]othing in this world develops absolutely evenly” (quoted in Smith 2006, 182), a truism that few would have questioned during his time as the head of the Chinese Communist Party (1945–1976), but which has been strongly reaffirmed in the dramatic forms of uneven development that have accompanied China’s subsequent embrace of globalizing capitalism. The country’s ongoing transformation from Maoism to marketization has been associated with spiraling inequalities, expressed in both social and spatial terms, and in historically new forms of uneven geographical development – both planned and unplanned. China’s post-1980s growth model has been predicated on uneven regional development, the management, containment, and reform of which has in recent years become a political priority.

Under reformist Premier Deng Xiaoping (1982–1989), China initiated the historical process of substituting one (stage) model of development with another. The declaration that China remained in the “primary stage of socialism” foretold a gradual departure from policies deemed appropriate for “mature” socialist societies, such as the Soviet bloc. The alternative, “socialism with Chinese characteristics,” sought
to accelerate economic development through concentrated capital investment, selective liberalization, and export-led growth in Beijing’s very own favored regions, the Pearl River Delta and the southeastern seaboard, on the coat tails of which the lagging regions of the rural interior and west were supposed to catch up in accordance with the “ladder-step” (tiaodilun) doctrine of market-led modernization. The ladder-step concept is a Chinese variant of orthodox growth-pole theory, which holds that economic development is enabled through the spatial concentration, localized agglomeration, or regional clustering of key industries and activities, the benefits of which are expected to trickle down (or out and back) to less-developed regions, sometimes via secondary growth poles, over time. In the Chinese case, accelerated development along the southeastern coastal areas has been facilitated by geographically targeted, experimental reforms (including in special economic zones like Shenzhen, Shantou, and Zhuhai), which has been the source, in addition to cheaply manufactured goods for global markets, of demonstration and backwash effects for less-developed regions.

China’s (market-friendly) plan has therefore been to achieve the historic goal of “higher” and more egalitarian development by means of an intensification of uneven regional development. These means have certainly been exploited to the full, but the historic goal remains elusive:

By endorsing “socialism with Chinese characteristics” and emphasizing efficiency over equity, Deng became a major force behind the reversal in regional policy. He support[ed] the concept of “stage of development,” which legitimizes uneven regional development as a natural and inevitable outcome of the development process. Once the stigma about uneven regional development is removed, [Chinese policymakers were] then free to advocate comparative advantage, regional specialization, regional division of labor and export-led economic growth. (Fan 1997, 632).

The fact that coastal development in China has been reliant on a vast “floating population” of migrant laborers, impelled by virtue of rural underdevelopment and economic displacement to seek work in the factories of the coastal growth zone – a population mostly stripped of not only workplace rights but residency rights (which attach to home regions under the hukou or household-registration system) – underscores the point that the country’s explosive rates of economic growth since the 1990s have been achieved through uneven regional development. Looming threats not only to economic sustainability but to social and political stability are surely not unrelated to the official embrace, since the late 1990s, of compensatory efforts to bring about a reduction in regional inequities, including massive programs of urbanization and industrialization across the western interior. The often overriding imperatives of capital accumulation, however, continue to drive spatially concentrated patterns of growth in China, often compounding the advantages of the first-mover regions on the top rung of the ladder. In this geographically differentiated context, regional growth models have been proliferating under a mode of regulation that combines centralized political control with devolved institutional experimentation, models that each seek to maximize the profit-making potential of distinctive positions within evolving international divisions of labor (Zhang and Peck 2014).

In contrast to many other “transition” economies, China has never been compelled to accept the policy advice of the multilateral development agencies like the World Bank; it has not been subject to externally imposed “structural adjustment.” China has fashioned its own path toward marketization, in the context of a significant degree of state integrity. Yet ironically,
some will now invoke the country’s experience as proof of the benefits of “free market” agglomeration. Policymakers are now being urged to make the most of the growth-accelerating effects of deregulated urbanization and concentrated investment, uneven regional development being represented as a natural stage on the path to more advanced (and ultimately convergent) forms of economic development. This echoes the long-standing contention, derived from classical economics, that trade-driven specialization and spatial concentration hold the key to long-term dynamic efficiencies. These arguments have acquired newfound legitimacy in the field known as new economic geography – as practiced by the likes of Masahisa Fujita, J. Vernon Henderson, Paul Krugman, and Jeffrey Sachs. By relaxing neoclassical assumptions with respect to increasing returns and perfect information, the new economic geographers account for the persistence of uneven spatial development and entrenched core–periphery patterns by reference to scale economies, comparative advantage, natural endowments, and the cumulative effects of path-dependent growth. Uneven regional development has duly been accommodated anew within mainstream policy orthodoxies.

While somewhat less idealized than the most parsimonious strains of neoclassical theorizing, the new economic geography once again invokes an almost featureless, rational-actor world, largely devoid of social institutions, politics, and power relations. No surprise, then, that this offshoot of orthodox economics has generally been at odds with the temperamentally heterodox field of economic geography “proper” (Peck and Sheppard 2010), where a range of political-economic, cultural, institutional, and socioeconomic theories continue to hold sway, and where uneven regional development is rarely equated with a rationalist path to prosperity, or with stage models, but instead stands as a reminder of the contingent, contradictory, and ultimately always incomplete character of profit-driven accumulation. Consistent with their closer acquaintance with the granulated realities of actually existing economies, uneven regional development remains, for economic geographers, not only a recurring empirical fact but a principled theoretical axiom.

SEE ALSO: Imperialism; Industrial restructuring; International division of labor; Location and multiplant firms; Marxist geography; Neoliberalism; Power and development; Regional inequalities

References


Unfree labor

Siobhán McGrath
University of Durham, UK

As of 2012, the International Labor Organization (ILO) estimated that there were 20.9 million people in some form of forced labor worldwide. Since in the early 2000s, advocacy around so-called “new slavery” has burgeoned alongside private and public anti-trafficking initiatives, contributing to renewed interest in understanding how and why unfree labor plays a role in contemporary economic dynamics. To date, the topic has not been widely studied within labor geography but some geographers are beginning to insist that it is, in fact, central to the questions that labor geography is concerned with (Strauss 2012).

With respect to contemporary labor relations, three of the terms used above—“new slavery,” forced labor, and unfree labor—are used as broad, overarching categories. These terms overlap in many respects, and empirical examples may fit the criteria for more than one category. Yet it is important to distinguish among them and their respective definitions.

Firstly, for Bales (1999), the term new slavery denotes a relationship constituted by three elements: exploitation, violence (or its threat), and loss of free will. New slavery is thus not limited to institutions such as chattel slavery. Rather, it encompasses a range of relations and is, therefore, closely aligned with concepts such as unfree and forced labor. Those in the new slavery school of thought tend to align with the work of Bales and the above definition of new slavery. While the literature on new slavery has arguably helped draw attention to the issue, it has also been faulted for describing a wide range of contemporary practices through the use of a term (slavery) normally associated with a much narrower set of historical practices.

Secondly, forced labor is mainly used as a politico-juridical term, a key reference for which is ILO Convention No. 29 (1930). The convention sets out two elements of forced labor: involuntariness (or a lack of consent) and the menace of a penalty. The ILO has interpreted initial consent as nullified, however, in instances of fraud, deception, or retention of identity documents. In 2012, the ILO published a set of indicators as guidance for interpreting the two elements in practice. (ILO publications on the topic are listed here: http://www.ilo.org/global/topics/forced-labour/lang--en/index.htm.)

Finally, the term unfree labor is generally used by those, including but not limited to Marxist scholars, seeking to understand its role within the wider economy. Unfree labor is understood to involve labor which: (i) is used by someone other than the person providing it, that is, an employer, but (ii) is obtained without allowing workers the type of “free” choice they are believed to exercise in “normal” labor markets. From a Marxist perspective, unfree labor within the context of capitalism is, therefore, doubly dispossessed: workers in unfree labor have neither access to the means of production nor the ability to personally commodify their own labor power.
UNFREE LABOR

Marxist theorization of unfree labor

Marxist scholarship theorizing unfree labor pre-dates the interest in the topic that has emerged over the past decade. Marxists interpret the “freedom” to participate in a labor market as masking the fact that workers are forced to sell their labor in the first place, having been dispossessed or “freed” of the means of (re-)production. Understood in this way, “free” labor has been central to Marxist understandings of capitalism as a class system. The fact that unfree labor, lacking even the opportunity to participate in a labor market, continues to exist under capitalism has thus been contentious. The primacy of “free” labor as the core feature of capitalism has led to numerous debates among Marxists from the 1970s onwards, hinging on the extent to which the existence of unfree labor indicates a noncapitalist mode of production.

One Marxist explanation of unfree labor in the context of capitalism lies in theories of primitive accumulation. These posit unfree labor as part of the process of dispossession that paves the way for capitalism. Another explanation is semi-feudalism, which treats unfree labor as evidence that a transition to capitalism is still underway. Both explanations have been challenged by fellow Marxists, who argue that unfree labor appears to be compatible with or even produced out of the dynamics of capitalism. While it may be possible to understand primitive accumulation as an ongoing process resulting from capitalism’s continued expansion, semi-feudalism appears to have few proponents today.

Miles’ work takes a slightly different approach. He examines how forms of unfree labor – indenture, slavery, convict labor, forced indigenous labor, domestic servitude, and labor in the apartheid system – may be articulated to a capitalist mode of production, or connected to it through economic relations. According to Miles, unfree labor may be reproduced as a means of maintaining the conditions for capitalism to function. Miles points to historical contingency, racism as a relation of production, and the state’s role in mobilizing labor within these articulations (Miles 1987). Finally, Brass’s theory of deproletarianization sees unfree labor as emerging from class struggle – in which those seeking to exploit labor impose unfreedom as a means of curtailing the development of class consciousness and resistance among workers (Brass 1999).

Critical studies of unfree labor

Some recent work on the topic cannot be placed neatly into either the Marxist framework or the “new slavery” school and has, therefore, been referred to as constituting a new school of thought: critical studies of unfree labor (McGrath and Strauss 2015). Many scholars associated with this school have questioned binary approaches to policy and analysis that place concepts such as “new slavery,” forced labor, and unfree labor in opposition to “free” labor. Such binaries are seen as, firstly, depending upon a liberal notion of contract and, secondly, treating unfreedom as an exception. According to this critique, widespread forms of exploitation, precarity, and degrading conditions of work are, therefore, portrayed as less pressing, rationalized through market ideologies in which workers are seen to have “chosen” this work. While the figure of 20.9 million workers in forced labor is shocking, for example, the vast majority of workers are, by virtue of their exclusion from this category, presumed to be “free.”

Marxian analyses of unfree labor are not equally subject to the above critiques, but the free/unfree binary may still be limiting. Social reproductive work involving unpaid labor provided by members of families and communities,
for example, is integral to the capitalist economy but not commodified. It therefore does not fit neatly into either the “free” or “unfree” categories according to the definitions used above. Rather, the extent and nature of unfreedom within these forms of labor would need to be assessed through measures other than the degree of commodification.

While some scholars therefore reject terms such as forced, unfree or “slave” labor altogether, many hold that unfree labor represents one end of a spectrum of exploitation (Lerche 2007). Building on this, a multidimensional view may allow for an assessment of different types of (un)freedom as well as different conditions of work (McGrath 2013a). Moving beyond binary approaches means that all forms of labor relations can potentially be subject to scrutiny in terms of the extent and types of (un)freedom that characterize them and how this, in turn, structures the negotiation over conditions of work. Such an approach would necessitate a challenge to two key ideas which inform much popular understanding as well as a considerable amount of academic analysis: first, the idea that chattel slavery is the ideal-type of unfree labor against which all other forms of labor are measured; and, second, that workers in forced labor, unfree labor, and new slavery are entirely lacking in agency.

“New World” slavery and unfreedom over time

In the Western imagination, chattel slavery is generally seen as not only the most egregious form of unfree labor but also as the “classic” form against which others should be measured. Chattel slavery was of course a critical institution in the development of colonialism and post-colonialism in the Americas. It can be understood as the institutionalized ownership of one person by another, in which this fictitious ownership is violently enforced and thus made real through what Patterson (1982) terms “social death.” (Other forms of slavery have existed which are not structured around property rights.) Historically, of course, chattel slavery in the Americas depended on forced migration of African peoples, and racialization was an explicit and core aspect of this system.

Seeing chattel slavery in the Americas as the ideal-type of unfree labor, though, engenders a somewhat ahistoric view which fails to recognize the variety of forms of unfreedom in labor relations. It is not uncommon for media reports on contemporary manifestations of unfreedom in labor relations to include phrases such as “slavery is back” or has “re-emerged.” Yet varieties of unfreedom can be identified before, during, and after the period of institutionalized chattel slavery in the Americas, in the “New World,” and elsewhere. In the United Kingdom and many former British colonies, for example, master and servant laws still structured labor relations in the nineteenth century, while vagrancy laws were used in many countries as a way to force people to stay in one place and work. “Free” labor as normally understood was, therefore, more the exception than the rule. And, in addition to chattel slavery, colonial regimes also imposed penal transportation, facilitated indentured labor migration, depended upon bonded labor, and imposed forced labor for economic development projects. Nor were all of these practices immediately eradicated with the inauguration of post-colonial regimes. Across the aforementioned labor relations, there were different forms and degrees of choice over who workers would work for and where, how long they would work for them, where they could go, and what relationships they could maintain “outside” of work.

This emphasis on chattel slavery also tends to be accompanied by abolitionist narratives.
UNFREE LABOR

(perpetuated through fictionalized historical portrayals, commemorations, and museum exhibits) that frequently downplay resistance and rebellions by enslaved people, instead emphasizing the actions of enlightened non-enslaved individuals who were involved in either abolitionist movements or in national governments. The abolition of slavery is, therefore, commonly portrayed as part of a march of progress. Yet in many cases when chattel slavery was legally abolished, the expansion of free labor did not immediately follow. In much of the British Empire, legal provisions for “apprenticeships” effectively prolonged enslavement after the legal date of abolition. In the United States, the convict leasing system that emerged in the South meant that former slaves were frequently imprisoned (often on groundless charges) and offered as cheap labor to local employers while dispossession also led to sharecropping arrangements for many former slaves. New or expanded systems of indentured labor emerged in many parts of the world after slavery was legally abolished, and practices of bonded labor continued. Albeit recognizing moment of sudden and dramatic change, some scholars seek to uncover historical continuities and evolutions as well, thereby countering the oversimplified notion that slavery ended and has since re-emerged (Quirk 2011; Kothari 2013).

The question of agency and debates around trafficking

Underlying many of the debates around the terms and concepts described above and about the question of binaries, spectrums, or dimensions is another issue: the question of agency (Rogaly 2008). Much of the writing on forced labor and “new slavery” treats workers as lacking agency altogether – sometimes explicitly so. This lends credence to policy responses that are apolitical and isolated from wider issues of power and regulation. These solutions are often framed in terms of “punishment, protection, and prevention” but fail to account for the rights, needs, and aspirations of workers. As a result, there is a poor record of progress in most of the world relative to the amount of resources dedicated to such efforts. Labor geography may have more to contribute here, as it is squarely focused on workers’ agency. Challenges to the idea that unfreedom equates to an absence of agency have been growing. Workers’ trajectories include various moments in which they exercise agency. Workers may initially seek work and/or migrate in an attempt to improve their social and economic circumstances, experience unfreedom, but then engage in struggles which result in improvements. These improvements may appear minor but nonetheless be significant. For example, workers may be earning incredibly low wages but may exercise agency in demanding some degree of health and safety protections.

Questions of agency and representations of “victims” as in need of rescue are particularly vexed in discussions of “sex trafficking.” Human trafficking has become a key policy concern since the early 2000s, particularly in the Global North, where it is frequently equated with contemporary “slavery” – overtly so by the US government. The ILO estimated in 2005 that (only) 20% of forced labor results from trafficking. As of 2012, they estimate that 68% of those in forced labor are working in the private economy in a range of activities not related to sexual exploitation, including construction, manufacturing, agriculture, and domestic work, with 22% in “forced sexual exploitation,” and 10% in “state-imposed” forced labor. Nonetheless, trafficking garners more attention than other forms of forced or unfree labor, and there is disproportionate attention given
to the sex industry in particular. This is both reflected in and reinforced by the emphasis the media gives to trafficking, particularly sex trafficking.

The United Nations (UN) Palermo Protocol to Prevent, Suppress and Punish Trafficking in Persons, Especially Women and Children defines trafficking as:

the recruitment, transportation, transfer, harbouring or receipt of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation.

According to the Protocol, trafficking may result in forced labor, slavery and “practices similar to slavery.” However, it also includes an ambiguous reference to “exploitation of the prostitution of others” as well as going beyond forced labor to include the practice of organ trafficking. Trafficking is generally understood to involve movement and/or a labor market intermediary and may involve workers crossing (or being transported across) national borders. It is frequently conflated with smuggling and unauthorized labor migration. But even where the definition is used precisely, discussions around trafficking are, in most countries, inextricably linked with debates around immigration. While some see restrictions on labor migration (within the context of contemporary neoliberal capitalism) as the principal reason that migrant workers are made vulnerable to unfreedom, political responses frequently start from the premise that reducing migration will reduce trafficking, and therefore promote even greater legal restrictions on labor migration.

The over-emphasis on trafficking – especially “sex trafficking” – over other forms of forced or unfree labor also creates a perception that trafficking, and by extension forced labor, affects mainly “women and children.” The coherence of placing women and children into a single category has been challenged (Anderson and O’Connell Davidson 2002). The ILO instead disaggregates by sex and estimates that 45% (11.4 million) of those in forced labor are men and boys, and has highlighted the issue in its 2014 report on the economics of forced labor. It has been pointed out that many stereotypes of women as helpless and in need of rescue are reproduced through the trafficking discourse (and to some extent through discourses on forced and unfree labor more widely). The degree to which the image of a stereotypical “victim” can be used to gain attention and resources is reflected in recent revelations that the stories of high-profile “victims” such as Somaly Mam were largely fabricated. Further, and echoing historical concerns over “white slavery,” there have been heated debates over whether all prostitution is “forced” (Doezema 2010). The US government, for example, positions itself as a global leader in the fight against trafficking through its annual trafficking in persons report and through distributing hundreds of millions of dollars in anti-trafficking funds worldwide since 2001. It requires recipients of these funds (and HIV/AIDS prevention funds) to take a stance against prostitution. Thanks to a Supreme Court decision in 2013, the requirement no longer applies to US-based organizations but can be applied abroad.

In addition to laws, policies, and government programs, an anti-trafficking industry has been spawned in which new or existing nongovernmental organizations aim to raise awareness, prevent trafficking, and/or provide services to those who have experienced trafficking. For many scholars, activists, and advocates, the resources and attention that have been mobilized
under the guise of fighting trafficking in recent years have largely failed to prevent labor abuses or to advance migrants’ rights. In fact, many argue that more harm than good has been done. This is based on the contention that anti-immigrant policy is more easily justified where governments are able to claim they have addressed concerns about trafficking. As anti-trafficking policy in particular often prioritizes prosecution over the needs of “victims,” those identified as trafficked often receive little assistance and, in some cases, may even experience harm as a result of anti-trafficking efforts – for example, as a result of police raids on brothels.

Rather than imagining situations in which workers’ agency is entirely absent, then, it has been argued that a more useful starting point is to see unfreedom (and freedom) in labor relations as an exercise of power (McGrath and Strauss 2015). Restricting workers’ options and mobility is, in essence, exercising control over labor (whether this is aimed at increasing the rate of exploitation, ensuring labor is available when needed at a low cost, or other reasons). The mechanisms through which this power is exercised vary across time and space. Further, no one moment is decisive: whether workers are able to participate in a labor market is not the only moment at which the balance between freedom and unfreedom will be determined. The exercise of power in labor relations is not just a matter of how individuals treat each other, for these dynamics are situated within wider political economic contexts.

Mechanisms of unfreedom in contemporary labor relations

The particular forms of unfreedom that are produced and reproduced in labor relations today are constructed out of contemporary socioeconomic dynamics, not least processes of neoliberalization (LeBaron and Ayers 2013) and the fragmented and dispersed processes that characterize contemporary globalized and globalizing capital (McGrath 2013b). Exclusions from or weakening of labor regulations can create an environment in which unfreedom is more easily imposed. Labor migration resulting, in part, from economic change is generally matched with restrictions on workers’ rights to employment if they migrate across national borders. And the power dynamics created through outsourcing and subcontracting in many global production networks can result in pressure on suppliers to reduce costs or improve productivity and flexibility while also partially shielding the most profitable firms from responsibility for the results in cases where workers endure unfreedoms and degrading conditions of work. Thus, the dominant policy approach, which suggests reducing workers’ vulnerability through education and anti-poverty interventions along with government implementation of legal initiatives that target “slavery” and trafficking in isolation from the wider context of labor rights and labor regulations, is clearly inadequate.

While the aspects of neoliberalization and globalizing capital are one key part of producing and reproducing unfreedom in contemporary labor relations, the mechanisms through which this is achieved vary. Trafficking, bonded labor, and other categories seek to define particular systems of unfree labor associated, respectively, with a particular set of mechanisms by which unfreedom is imposed. Debt is a key mechanism, perhaps even the most common mechanism used today, and such debt is often clearly induced. Yet, in South Asia bonded labor is further intertwined with the caste system and caste discrimination. Processes of racialization, which mark certain bodies as suited for more difficult or dangerous conditions of work and able to sustain themselves with lower levels of compensation, play a role
in justifying unfreedom in many contexts. In the case of trafficking, debt (often induced) also tends to play a key role, but so too does migration status (in itself racialized). And it is not just undocumented or unauthorized workers who face unfreedom in labor relations. In many cases, workers’ “legal” status ties them to an individual employer in that they are not allowed to change jobs under the terms of their visas. The kafala system in the Gulf States, Jordan, and Lebanon is the prime example of this but some guest-worker schemes and domestic worker visas in the United Kingdom, the United States, and elsewhere tie workers to their employers. This highlights the role of the state and the threat of state violence in structuring relations of unfreedom. Interpersonal violence and other forms of abuse by employers or labor market intermediaries are mechanisms by which unfreedom is enforced in some cases, but contrary to some analyses they are not the hallmark of unfreedom. Restrictions on physical mobility, sometimes independent of migration status (or residency status in the case of China’s hukou system), also facilitate unfreedom. Isolation, as a feature of some remote workplaces or as an intentional strategy, can contribute to restricted mobility. Threats may be used as well, and/or identity documents may be retained. But restrictions on mobility also include arrangements in which workers are housed in employer-provided accommodation.

A number of other forms of unfree labor exist today. The penal system may be the context for unfree labor. Child soldiers are generally considered to be in unfree labor. Religious beliefs may be manipulated as a mechanism of imposing unfree labor. The restavek system in Haiti and similar systems elsewhere rely on (and manipulate) traditions in which extended family members would take in children of families who were struggling to meet their needs. The sociocultural, economic, legal, and institutional contexts therefore matter in terms of which mechanisms of unfreedom (and freedom) are available and to what extent. For this reason, different categories of unfree labor are constructed in particular moments and places, and systems of unfreedom in labor relations have changed over time. A number of scholars have described forms of “neobondage” in India today, for example, in which relations are less personalized than had previously been the case. There is, then, considerable space to further the study of unfreedom in labor relations from a perspective that is sensitive to temporality, space and place, and the relationships between movement and mobility.

**SEE ALSO:** Gender, work, and employment; Labor geography; Labor migration; Migrant labor; Precarious work; Regulation/deregulation; Rights, labor; Social reproduction

**References**


UNFREE LABOR


Unionism, community

Steven Tufts
York University, Canada

The origin of the term “community unionism” dates back to the 1960s when it was used to refer to students organizing in poor city neighborhoods in the United States. It was adopted by geographers in the 1990s as a way of discussing community–union coalitions in the emergent field of labor geography (see Tattersall 2010; Tufts 1998). Amanda Tattersall (2010, 20–21) defines community unionism as: coalitions between unions and nonlabor groups; union strategies that organize workers on the basis of identity rather than place of work; and place-based organizing strategies that mobilize communities around extra-workplace issues. These strategies can range from campaigns to organize racialized domestic workers to struggles against local factory closures to larger conflicts with the state over the regulation of labor (e.g., living/minimum wage campaigns) and the withdrawal of public services (e.g., education and health-care funding).

There are debates over community unionism in terms of the concept itself and its political efficacy. First, there are conceptual issues of how to define community, whether as place, identity, or set of institutions. Such distinctions are important in determining coalition partners and common areas of concern. Second, there is the question of scale and the extent to which common concerns may be too place-specific to build broader geographical coalitions or may fail to focus on the local concerns that often inspire initial action.

Third, there is debate over the limits and power of coalitions as tensions often arise between the needs of communities and workers (e.g., local environmental groups and unions representing workers in resource extraction industries). Here, structures are required to ensure consistent reciprocity among labor and nonlabor groups (Holgate 2009). Building organizational capacities and structures to accommodate such tensions are necessary to allow the community unions to accumulate political power and maintain coalitions after issue-based campaigns have ended.

There are barriers to building effective community unions. Foremost are the unequal resources that different groups can bring to coalitions. Unions most often have greater organizational and financial resources at their immediate disposal given that they have access to revenues from membership dues. At the same time, community groups have fewer resources but may have the necessary expertise and relationships with marginalized communities that unions often lack. The challenge is to establish democratic decision-making bodies that give all members a strong voice. Tattersall (2010) has developed a framework for the categorization of community–union coalitions with “deep coalitions” being the most developed, integrated, and rarest formation.

SEE ALSO: Community; Labor geography; Power; Scale

References

UNIONISM, COMMUNITY


United Kingdom: Council of British Geography (COBRIG)

Founded: 1988
Location of headquarters: London
Website: www.cobrig.org.uk
Membership: 10 constituent organizations/agencies
Honorary Secretary: Peter Wood
Contact: peter.wood@ucl.ac.uk

Description and purpose

Established in 1988 to coordinate the promotion of geographical education across the United Kingdom, COBRIG links organizations in England, Scotland, Wales, and Northern Ireland where there are separate systems of education, research, and public policy. The Royal Irish Academy Committee for Geographical Sciences has had observer representation since 2012. Members promote geographical knowledge and education in the British Isles through public policy, in school teaching, and in higher education. COBRIG acts in partnership with member organizations on matters concerning the status of the discipline across the British Isles.

Current activities or projects

The council holds twice yearly meetings to review development across the British Isles in schools and university geography teaching and research, and other developments affecting the public contribution of the discipline. It facilitates communication between member organizations and supports meetings of school and higher education representatives to review developments affecting geographical education and research in the British Isles.

Brief history

The Council of British Geography coordinates the promotion of geography in the British Isles and, if necessary, can also act on matters concerning the status of the discipline in education, research, and public policy where geographers’ views need to be made known.

In recent years, COBRIG’s member organizations have developed closer cooperation in various ways. In 1995, the Royal Geographical Society and Institute of British Geographers were merged. The Royal Scottish Geographical Society and Scottish Association of Geography Teachers also work increasingly closely together. Member organizations currently represented are:

- The Royal Geographical Society (with the Institute of British Geographers)
- Royal Scottish Geographical Society
- University of Wales Council on Geography
- The Royal Irish Academy Committee for Geographical Sciences
- Geographical Association (in England and Wales)
- Scottish Association of Geography Teachers
- Welsh Joint Education Committee
- Her Majesty’s School Inspectorate in each country
- Committee of Heads of Geography in Higher Education
- British Science Association Section E (Geography)

COBRIG supports cooperation between these organizations through twice-yearly meetings to
review educational research and other developments affecting the public contribution of the discipline. Issues of regular concern include the status of geography in the school curriculum across the British Isles, the quality of school teacher training, and trends in numbers taking schools examinations and applying for university degree courses in geography.

Submitted by Peter Wood
United Kingdom: Royal Scottish Geographical Society (RSGS)

Founded: 1884  
Location of headquarters: Perth, Scotland  
Website: www.rsgs.org  
Membership: 2500 (as of 2014)  
Chief Executive: Mike Robinson  
Contact: enquiries@rsgs.org

Description and purpose

The Royal Scottish Geographical Society is an educational charity which promotes an understanding of the natural environment and human societies and how they interact making the connections between people, places, and the planet.

Established in 1884, we have a long and distinguished history of supporting geographical education, research, and exploration. We are the leading organization in Scotland providing geographical understanding on contemporary issues which shape our future.

Journals or major publication series

Scottish Geographical Journal. www.tandfonline.com/loi/rsgj20#.VmXvCfl96VM

Current activities or projects

The Society’s activities focus on the following.

Inspiring people – an annual program of illustrated public talks all around Scotland awarding prestigious medals that celebrate achievement and running a public visitor and education center.

Making Connections – contributing to scientific and policy debate.

Promoting science – encouraging geographical study and research, publishing an academic journal, awarding small grants, and running conferences.

Exciting learning – using our collection, expertise, volunteers, and networks to help support formal and informal education and working at policy and practical levels to promote geographical teaching and learning in schools.

Conserving heritage – maintaining an important historical and contemporary collection of geographical artifacts, images, maps, books, journals, papers, and other publications.

Brief history

The Royal Scottish Geographical Society (RSGS) was founded in 1884 and is based in Perth. The society has a membership of 2500 and aims to advance the understanding of geography worldwide by supporting education, research, expeditions, through its journal (Scottish
UNITED KINGDOM: ROYAL SCOTTISH GEOGRAPHICAL SOCIETY (RSGS)

Geographical Journal) and its newsletter (The Geographer).

Originally based in Edinburgh, the society moved to offices in Glasgow before relocating to Perth in 2009 where the society rejuvenated the city’s oldest building The Fair Maid’s House. This is now a public visitor and education center.

Submitted by Fraser Shand
United Kingdom: Royal Geographical Society (with the Institute of British Geographers) (RGS-IBG)

Founded: 1830
Location of headquarters: London
Website: www.rgs.org
Membership: 16,500 (as of 2014)
Director: Rita Gardner
Contact: director@rgs.org

Description and purpose

The Royal Geographical Society (with IBG) (RGS-IBG) has one objective, the advancement of geographical science. In the twenty-first century the society does that through securing and promoting the position of the discipline in education, fieldwork, higher education and research, and raising its profile with the public and government; advancing geographical knowledge and understanding and sharing new scholarship with a wide range of audiences, including global scholars, teachers, students, policymakers, businesses, and the public; supporting the discipline’s practitioners (professional geographers and students) with training, mentoring, networks, and professional accreditation; and serving a vibrant membership.

Journals or major publication series


Current activities or projects

Annual International Conference: three-day event attracting about 2000 scholars worldwide, planned by the society’s 28 specialist research groups.
Grant giving for research and scientific expeditions: about £170,000 annually in small grants mainly to those at earlier career stages.
Professional development and online resources for school teachers: about 1.5 million downloads in 2014.
Influencing policy: at the heart of government reforms of geography school curricula and the revised framework for geography in higher education.
Inspiring next generation: 600 ambassadors talk to 30,000 pupils each year.

Brief history

The RGS-IBG, the UK’s learned society and professional body for geography, is a dynamic institution with a distinguished history and modern reputation. Founded in 1830, the RGS-IBG is the largest geographical society in Europe and one of the most active in the world.
UNITED KINGDOM

The society exists for the “advancement of geographical science.” Today it is involved in a range of activities supporting geographical research, education and outdoor learning, public engagement, and knowledge transfer to policy. The society actively advises on and represents the discipline and is governed by a council of trustees elected from its fellowship.

The society’s collections are an unrivalled repository tracing 500 years of geographical discovery and research. Included are more than one million maps and atlases, 500,000 images (photographs, sketches, art works, and films), and books and manuscripts. Recent initiatives have focused on “unlocking these archives” – enabling public access.

The society was responsible for both the incorporation of the study of geography in schools (at the turn of the twentieth century) and the first university positions in the discipline. Today, the society continues actively to influence educational debates and supports teachers and students in schools and the academic community through grants, journals, conferences, specialist research groups, and the society’s medals and awards.

Broader initiatives to promote the wider understanding of geography include a rich program of public lectures, support for fieldwork and expeditions, debates, walks, exhibitions, and the popular Geographical magazine. Society briefings, proactive media work, and events serve to place geography firmly in policy debates about the future of places, environments, and communities.

Submitted by Rita Gardner
United Kingdom:
Geographical Association (GA)

Founded: 1893
Location of headquarters: Sheffield
Website: www.geography.org.uk
Membership: about 5500 (as of 2013)
Chief Executive: Alan Kinder
Contact: akinder@geography.org.uk

Description and purpose

The Geographical Association is a subject association with the charitable objective of furthering geographical knowledge and understanding through education. It is a lively community of practice with over a century of innovation behind it and an unrivalled understanding of geography teaching. The GA was founded in 1893 in Oxford, England to share ideas and learn from each other. Today, the GA’s purpose is the same. The GA’s activities are financed principally by membership subscriptions, but also by the sale of educational resources, conference and CPD event delegate fees, and externally funded curriculum development project work.

Journals or major publication series

Journals

GA Magazine
Geography
Teaching Geography

All journals available online at www.geography.org.uk/Journals

Handbooks and other publications

The following are key reference books for teachers of geography, PGCE students, and geography teacher educators.

Geography Through Enquiry (2013) Margaret Roberts.

Current activities or projects

Meeting professional and educational needs: The GA provides support and guidance for geography educators, principally in the UK but with members across 60 countries. This may be grouped under the subheadings: professional development events and networks, publication of resources, provision of geography “quality marks,” and bespoke consultancy services.

Demonstrating the value of geographical education: The GA maintains close relationships with UK central government and directly advises on curriculum matters. It engages with key stakeholders to influence public debate and policy toward geographical education. The association works in partnership with other organizations to ensure that geography’s contribution is maximized and well understood, including through educational research.

Brief history

The GA was founded in 1893 in Christ Church, Oxford, by Sir Halford Mackinder and ten
UNITED KINGDOM: GEOGRAPHICAL ASSOCIATION (GA)

others. Douglas Freshfield was the first president of the association in 1897. In 1901 *The Geographical Teacher* (renamed *Geography* in 1927) became the first GA journal. Membership was opened to non-UK geographers in 1920. In 1930 Professor Sir L. Dudley Stamp led the Land Utilisation Survey of Britain. In 1932 the first edition of *Handbook for Geography Teachers* was published. In 1975 the *Teaching Geography* journal was launched as a practical guide for teachers and in 1987 *A Case for Geography*, edited by Bailey and Binns, helped secure geography as a national curriculum subject. The GA website was launched in 1998 and in 2005 the *GA Magazine* (incorporating *GA News*) was launched. From 2006 to 2011 an Action Plan for Geography, led jointly by the GA and the RGS-IBG, was put in place to boost geography education in England and Wales. From 2013 to date, the GA forms part of the Global Learning Programmes for England and Wales to improve global understanding and development education in schools.

Submitted by Alan Kinder
United States of America: American Geographical Society (AGS)

Founded: 1851
Location of headquarters: New York
Website: www.amergeog.org
Membership: 800 (as of 2014)
Executive Director: John Konarski III
Contact: gs@amergeog.org

Description and purpose

The American Geographical Society (AGS) advances and promotes geography in business, government, science, and education. Our goal is to enhance the nation’s geographic literacy so as to engender sound public policy, national security, and human wellbeing worldwide. Since 1851, AGS has been a leading advocate for geography in the United States and around the world. AGS stands for better analysis and decision-making in business and government based on better understanding of how real-world geography affects society, economics, infrastructure, and politics. AGS stands for better science and education based on explicit recognition of the spatial and temporal contexts that shape the real world and influence how it works.

Journals or major publication series

FOCUS on Geography. http://americangeo.org/focus-on-geography

Current activities or projects

A large portion of the work at AGS takes the form of research and special projects. Since its founding in 1851, original research has been a distinguishing hallmark of AGS. Every year AGS participates in research studies either as the primary organization conducting the research or as a supporting organization of research under the direction of an AGS member or an AGS affiliated organization. In addition to original research, AGS has a commitment to supporting special projects such as conferences, lectures, publications, and other projects that help AGS fulfill its mission. Currently, some examples of AGS activities include: the highly regarded worldwide Bowman Expeditions, continued development of the World Standard for Cartographic Representation, updating of the AGS Geographic Knowledge and Values Survey, continued recognition for geographical excellence through the highly prestigious AGS Medal/Awards program and Fliers and Explorers Globe signings, updating the Independent Accuracy Assessment of LandScan Global Populations Data, development of the AGS and MapStory Foundation Ambassador Network, and hosting the AGS Annual Meeting and Geographic Symposium.

Submitted by John Konarski
United States of America: Cartography and Geographic Information Society (CaGIS)

Founded: 1981
Location of headquarters: Mt. Pleasant, SC
Website: www.cartogis.org
Membership: 250 (as of 2014)
Executive Director: Eric Anderson
Contact: cagisxd@gmail.com

Description and purpose

The Cartography and Geographic Information Society (CaGIS) is composed of educators, researchers, and practitioners involved in the design, creation, use, and dissemination of geographic information. CaGIS provides an effective network that connects professionals who work in the broad field of cartography and geographic information science both nationally and internationally.

Journals or major publication series

Cartography and Geographic Information Science. www.tandfonline.com/toc/tcag20/current#

Current activities or projects

The CaGIS is the US member of the International Cartographic Association (ICA). CaGIS supports participation in ICA conferences and activities. CaGIS also hosts the AutoCarto series of biennial conferences that focus on cartographic and geospatial research. The society will host the 28th International Cartographic Conference of the International Cartographic Association July 2–7, 2017, in Washington, DC (www.icc2017.org).

The CaGIS Distinguished Career Award honors the accomplishments of senior professionals in areas of cartography and GIS or GIScience. CaGIS sponsors annual scholarships to students in research programs. An annual map competition for professionals and students is held to recognize significant design advances in cartography.

Submitted by Terry Slocum
United States of America: National Council for Geographic Education (NCGE)

Founded: 1915
Location of headquarters: Washington, DC
Website: www.ncge.org
Membership: 2750 (as of December 31, 2013)
President: Susan Hume
Contact: ncge@ncge.org

Description and purpose

The National Council for Geographic Education (NCGE) was founded in 1915 to enhance the status and quality of geography teaching and learning. NCGE supports geography teaching at all education levels. Its members include K-12 teachers, curriculum directors, professors of geography and education, authors, researchers, students, business representatives, and others who support geography education.

NCGE fulfills its mission through its journals, online newsletter, publications, awards program, annual conference, conducting and gathering research, developing curricular resources, providing professional development opportunities, and partnering with national and international organizations.

Journals or major publication series

*Journal of Geography*. www.ncge.org/journal-of-geography
*The Geography Teacher*. www.ncge.org/the-geography-teacher

Geography for Life: National Geography Standards. www.ncge.org/geography-for-life

Current activities or projects

NCGE facilitates communication and professional development among geography educators. The annual National Conference on Geography Education (www.ncge.org/conference) and live webinar program (www.ncge.org/webinars) strive to enhance the preparation of geography educators with respect to their knowledge of content, techniques, and learning processes. NCGE works to develop, publish, and promote learning materials and curriculum through its journals and publications (www.ncge.org/publications), while also recognizing exceptional supporters and teachers of geography with its annual awards program (www.ncge.org/awards). NCGE supports, gathers, and conducts research (www.ncge.org/research-in-progress) and works in collaboration with organizations that have similar goals.

Brief history

The National Council for Geographic Education (NCGE) was started by George J. Miller of the State Normal School in Mankato, Minnesota, to fill a gap in the field of geographic education. The idea gained support nationally and the first meeting of the organization was held on December 31, 1915.

Originally called the National Council for Geography Teachers (NCGT), the purpose was to increase the effectiveness of geography teaching in America. Miller was concerned with both the quantity and the quality of geography education. His vision was that educators at all levels
be both subject-area experts and trained in the most advanced teaching methods. In 1956, the NCGT changed its name to the National Council for Geographic Education to reflect the idea that not all education occurs in the classroom.

Submitted by Zachary R. Dulli
United States of America: National Geographic Society

Founded: 1888  
Location of headquarters: Washington, DC  
Website: www.nationalgeographic.com  
Membership: 8.45 million (as of January 2014)  
President: Gary Knell  
Contact: http://press.nationalgeographic.com/media-contacts/

Description and purpose

National Geographic Society is one of the world’s largest nonprofit scientific and educational organizations. With a mission to inspire people to care about the planet, the member-supported society offers a community for members to get closer to explorers, connect with other members, and help make a difference. The society reaches more than 500 million people each month through National Geographic and other magazines, National Geographic Channel, documentaries, films, books, DVDs, radio, maps, exhibitions, live events, interactive media, and merchandize. National Geographic has funded over 10,000 scientific research, conservation and exploration projects, and supports an education program promoting geographic literacy.

Journals or major publication series


Current activities or projects

National Geographic’s grant programs have funded over 10,000 scientific research, exploration, and conservation projects around the world. The society’s Big Cats Initiative focuses on stemming the decline of big cats in the wild and protecting their habitats.

National Geographic’s Genographic Project uses DNA analysis to help answer questions about human origins and migratory history to better understand the connections and differences that make up humankind. The society’s education activities include a classroom magazine, national and international geography competitions, state geographic alliances, and professional development resources for teachers. The society’s public programs include exhibitions, speaker series, films, and special events.

Brief history

The National Geographic Society was founded in Washington, DC, on January 13, 1888, by 33 of the city’s scientific and intellectual leaders who had gathered to consider “the advisability of organizing a society for the increase and diffusion of geographic knowledge.” Lawyer and financier Gardiner Greene Hubbard was elected the society’s first president. He was succeeded in 1897 by his son-in-law, inventor Alexander Graham Bell, who was followed by four generations of the Grosvenor family. The current president is Gary E. Knell, who took office in January 2014.
In October 1888, the first issue of *National Geographic* magazine was published. The first local-language edition of the magazine was published in Japan in 1995. There are now nearly 40 local-language editions, and the magazine is read by 60 million people globally in print and digital editions.

Research, conservation, and exploration have been at the heart of the society’s work since it sponsored its first expedition in 1890 to explore and survey Alaska’s Mount St Elias. The society has funded more than 10,000 scientific projects around the world, including Robert Peary’s expedition to the North Pole in 1909, Hiram Bingham’s expeditions to Machu Picchu from 1912 to 1915, Richard Byrd’s pioneering flight over the South Pole in 1929, primate research by Dian Fossey and Jane Goodall, investigations into human origins by the Leakey family, ocean research by Jacques Cousteau, and dinosaur discoveries by Paul Sereno.

National Geographic television programming debuted with the National Geographic Special “Americans on Everest” on CBS in September 1965. In September 1997, National Geographic entered the cable television market, and National Geographic channels are now available in 440 million households in 171 countries in 48 languages.

As part of an initiative to promote geographic literacy and improve geography education in schools, the society launched a Geography Education Program in 1985. Offshoots of this are the annual National Geographic Bee and the biennial international geography competition, the National Geographic World Championship.

Today, 126 years after its founding, the society is one of the world’s largest nonprofit scientific and educational institutions, reaching more than a half billion people each month. With a mission to inspire people to care about the planet, the multimedia organization reflects the world and explores new frontiers through its magazines, television programs, films, radio, books, DVDs, maps, exhibitions, live events, interactive media, travel and education programs, and merchandise.

Submitted by Bob Dulli
United States of America: North American Cartographic Information Society (NACIS)

Founded: 1980
Location of headquarters: Milwaukee, WI
Website: www.nacis.org
Membership: 528 (as of January 1, 2014)
President: Tanya Buckingham
Contact: tanya@nacis.org

Description and purpose

The North American Cartographic Information Society (NACIS) is a collegial collection of map enthusiasts. The aim of the society is to improve the quality of mapmaking and map literacy by creating opportunities for collaboration between everyone who works with cartographic materials and data. The society’s various initiatives bring together specialists from academia, private industry, government agencies, and the media so that we can all be better at what we do and have more fun doing it.

Journals or major publication series

*Cartographic Perspectives*. www.cartographicperspectives.org

Current activities or projects

The society’s annual meeting brings together cartographers, map librarians, researchers, and other map enthusiasts where members trade knowledge and form new collaborations.

To keep the conversation going beyond the conference, NACIS hosts CartoTalk (www.cartotalk.com), a cartography forum with a worldwide membership.

The society’s triannual journal, *Cartographic Perspectives*, is the only free and open-licensed cartography publication, giving contributors the opportunity to reach a broad audience. Our biennial anthology, the *Atlas of Design*, recognizes some of the world’s best recent cartographic works selected from an international competition.

MapGiving organizes cartographers for pro bono work: creating maps for nonprofits, educational materials, and free data sets.

Submitted by Tanya Buckingham
United States of America: Open Geospatial Consortium (OGC)

Founded: 1994
Location of headquarters: Wayland, MA
Website: www.opengeospatial.org
Membership: 472 (as of March 5, 2014)
President: Mark Reichardt
Contact: mreichardt@opengeospatial.org

Description and purpose

The OGC is an international consortium of more than 470 companies, government agencies, research organizations, and universities participating in a consensus process to develop publicly available geospatial standards. OGC standards support interoperable solutions that “geo-enable” the web, wireless and location-based services, and mainstream IT. OGC standards empower technology developers to make geospatial information and services accessible and useful with any application that needs to be geospatially enabled.

Journals or major publication series

The documents published by the OGC are software interface and encoding standards, best practices, public engineering reports, discussion papers, reference models, and white papers. These are all free and available to the public at www.opengeospatial.org/standards.

Current activities or projects

The OGC convenes quarterly Technical Committee and Planning Committee meetings (www.opengeospatial.org/ogc/programs/spec) at locations that alternate between North and South America; Europe, the Middle East and Africa; and Asia Pacific. At these meetings OGC domain working groups and standards working groups discuss interoperability requirements and advances and vote on standards and related best practices.

The OGC Interoperability Program (www.opengeospatial.org/ogc/programs/ip) runs test beds, pilot projects, interoperability experiments, and concept development studies. These fast-paced initiatives enable geospatial technology users and providers to work collaboratively in an agile development environment to develop, evolve, test, demonstrate, and validate candidate geospatial standards under marketplace conditions.

Brief history

The OGC was founded in 1994 to engage geospatial companies and major user agencies in a consensus process to define interface standards that would enable easy and direct sharing of geospatial data. From the first approved OGC implementation standard in 1997 and the first OGC Interoperability Program testbed (Web Mapping Testbed) in 1999, the consortium has expanded to provide a broad array of standards and initiatives. OGC standards have been implemented in hundreds of commercial and open source geoprocessing products deployed in communities and organizations around the world.
UNITED STATES OF AMERICA: OPEN GEOSPATIAL CONSORTIUM (OGC)

the world. OGC standards are key elements in sensor webs, location services, Earth imaging networks, climate models, disaster management programs, and national spatial data infrastructures around the world. To ensure consistency across the Internet and web ecosystem, the OGC maintains alliance partnerships with many other standards development organizations and industry associations who work closely with the OGC on a wide range of topics such as indoor/outdoor location integration, sensor fusion, 3-D urban modeling, emergency communications, smart grid, location based marketing, aviation, meteorology, the Internet of things, points of interest, and the Semantic Web.

Submitted by Mark Reichardt
Founded: 1925  
Location of headquarters: Washington, DC  
Website: www.iswg.org  
Membership: 487 (as of December 31, 2013)  
President: Ann Imlah Schneider  
Contact: headquarters@iswg.org

Description and purpose

The Society of Woman Geographers is a non-profit association designed to provide a medium of intellectual exchange among traveled women engaged in the study of geography and its allied arts and sciences, to further geographical work in all its branches, to spread knowledge of the world and its people, to encourage geographical research, and to provide fellowship programs that aid outstanding women graduate students in geography and its related fields.

Current activities or projects

SWG supports and recognizes women engaged in the study of geography and its associated arts and sciences through its headquarters’ resources (oral histories, library, and museum), its awards and honors program, and its fellowships for graduate students. Intellectual exchange and interaction occur through SWG’s group meetings and its Triennial conference when members make presentations about their research, work, or explorations. Selected members also carry the society’s flag on expeditions of such unusual character that successful accomplishment adds distinction to the society and makes a permanent contribution to the world’s geographical knowledge.

Brief history

Four exceptional women explorers founded SWG in 1925. Marguerite Harrison, Blair Niles, Gertrude Shelby, and Gertrude Emerson Sen created a group to bring together women who shared ambitions and interests in exploration and achievement. No women’s organization existed at the time to share experiences, exchange knowledge derived from field work, and encourage women pursuing geographical exploration and research. In naming their organization the Society of Woman Geographers, the founders intended the word “geographer” in its broadest sense.

Today some 500 members are connected in ways the founders could not have imagined. SWG has grown from a single group in New York to a total of six nationally, adding Washington DC (the headquarters), South Florida, Chicago, Southern California, and the Bay Area. Each group meets several times (between September and May) to hear from a member or guest reporting on research or unusual travel. In addition, SWG has at-large members living in 36 states and corresponding members, who reside outside of the United States, currently in 32 countries on five continents. Every three years, SWG holds an international conference, the Triennial. At this meeting SWG awards a Gold Medal and an Outstanding Achievement Award to members of particular distinction.

Submitted by Ann Imlah Schneider
United States of America: The American Association of Geographers (AAG)

Founded: 1904
Location of headquarters: Washington, DC
Website: www.aag.org
Membership: 11 000 (as of December 2015)
Executive Director: Douglas Richardson
Contact: gaia@aag.org

Description and purpose

The American Association of Geographers (AAG) is a professional and scholarly association representing educators, researchers, and practitioners in geography. Founded in 1904 as a nonprofit scientific and educational society, its 11 000 members share interests in the theory, methods, and practice of geography. The AAG is governed by an elected council, which sets overall policy and appoints an executive director who manages the affairs of the association.

AAG hosts approximately 9000 attendees each year at its annual meetings, sponsors international workshops and specialty conferences, and conducts research programs on a wide range of geographical topics. The AAG also publishes some of the world’s most distinguished and influential scholarly journals, newsletters, books, and research reports. Publications include its flagship journals and the monthly AAG Newsletter.

Representing a discipline with a strong international perspective, the AAG also supports and recognizes geographers around the globe through its many professional development programs and prestigious grants and awards programs. Scientific, research and educational programs are conducted internationally and scientific exchange and interaction occurs through the AAG’s annual meetings, its leading scholarly journals, its newsletters, an extensive network of online interactive media and discussion groups, and other collaborative activities. The AAG outreach and research network consists of more than 30 000 colleagues worldwide.

Journals or major publication series

GeoHumanities. www.aag.org/cs/publications/journals/gh
The Professional Geographer. www.aag.org/cs/publications/journals/pg

Current activities or projects

The AAG’s current activities include multiple international educational programs and workshops, the creation and dissemination of professional development and careers resources, and substantial grants and awards programs.

The AAG also sponsors 70 specialty groups which promote geographic research and discussion on specific regional or topical geographic research areas. The AAG’s nine regional divisions organize annual meetings and other activities within their regions. The AAG also developed the capacity to carry out large-scale research programs and now conducts numerous funded cutting-edge research projects which help advance the discipline of geography.
UNITED STATES OF AMERICA

Brief history

In 1904, an emerging need for a society dedicated to the study of geography as a discipline encouraged William Morris Davis and other academic geographers to found the American Association of Geographers. In 1948, the AAG was amalgamated with the American Society of Professional Geographers (ASPG). From a charter membership of 48, the association has grown to more than 11,000 members, now representing geographers and related disciplinary researchers in universities, governments, international agencies, NGOs, and the private sector.

In 2016, the AAG was renamed the American Association of Geographers to reflect the increasingly international composition of its membership.

As the scientific, societal, and university context within which the discipline of geography exists has evolved, the American Association of Geographers has continued to lead, adapt, and evolve in order to sustain and strengthen geography’s intellectual and practical contributions to the needs of today’s rapidly changing world.

Submitted by Douglas Richardson
United States of America: The GIS Certification Institute (GISCI)

Founded: 2004
Location of headquarters: Des Plaines, IL
Website: www.gisci.org
Membership: 6000 (as of December 2015)
Executive Director: Bill Hodge
Contact: info@gisci.org

Description and purpose

The GIS Certification Institute (GISCI) is a tax-exempt, not-for-profit organization that provides the geographic information systems (GIS) community with a complete geospatial certification program. GISCI offers participants, from the first early years on the job until retirement, a positive method of developing value for professionals and employers in the GIS profession. GISCI offers an industry-wide, internationally recognized, software agnostic GIS certification available to geospatial professionals around the world.

Brief history

From 2002 to 2004, several leading organizations, including the Urban and Regional Information Association (URISA) and the American Association of Geographers (AAG), began exploring the prospect of developing a GIS certification program and both URISA and the AAG had undertaken concrete plans to do so. In 2004, at the invitation of URISA, the AAG decided to join URISA and two other geospatial organizations, the National States Geographic Information Systems Certification Institute (GISCI) and the University Consortium of Geographic Information Science (UCGIS), to form a wholly separate and independently incorporated organization, GISCI, to develop and administer a nationwide GIS certification program. The four founding member organizations of GISCI each appointed two board members to serve as the governing Board of Directors of GISCI and to build the new organization. In 2010, the Geospatial Information and Technology Association (GITA) became the fifth member organization of GISCI, and the latest organization to join is the Geographic and Land Information Society (GLIS) in 2013.

From 2004 through the present, a great deal of work have been undertaken by the GISCI founding member organizations and their board members to develop and build the GISCI as a viable organization, to develop myriad procedures and practices for certification, to collectively develop the GISCI as a credible certification entity for the GIS community, and to recruit new geographic information system professionals. During the past several years, the board members of GISCI have expended a great amount of effort, thought, and funds to develop an examination as a core part of the GISCI GISP application process. This examination-based GISP application process was implemented in 2015.

GISCI’s ongoing purpose is to maintain high standards of professionalism, competency, and integrity within the GIS profession and community and to promote ethical conduct within it.

Submitted by Bill Hodge
United States of America: University Consortium for Geographic Information Science (UCGIS)

Founded: 1995  
Location of headquarters: Ithaca, NY  
Website: www.ucgis.org  
Membership: 55 (as of 2014)  
Executive Director: Diana S. Sinton  
Contact: dianasinton@ucgis.org, execdir@ucgis.org

Description and purpose

The University Consortium for Geographic Information Science (UCGIS) is a nonprofit organization that creates and supports communities of practice for geographic information science (GIScience) research, education, and policy endeavors in higher education and with allied institutions. UCGIS advances research in the field of GIScience, expands and strengthens GIScience education, advocates policies for the promotion of the ethical use of and access to geographic information and technologies, and builds scholarly communities and networks to foster multidisciplinary geographic information systems research and education. UCGIS serves as a professional hub for the academic GIScience community.

Current activities or projects

The Geographic Information Science & Technologies, Body of Knowledge project (2006) was the first comprehensive effort to inventory and document the knowledge that comprises this field. In 2014, UCGIS coordinated a community effort to revise the GIS&T BoK.

At its annual symposium, academic thought-leaders share and discuss research activities, educational initiatives, and the policy landscape. Professional networking is a key component of this gathering.

UCGIS hosts workshops on various technical, theoretical, and educational aspects of GIScience.

UCGIS webinars highlight innovative projects and share formative ideas to support the GIScience higher education community.

Submitted by Diana Sinton
United States of America: Gamma Theta Upsilon (GTU)

Founded: 1928
Location of headquarters: New Orleans
Website: www.gammathetaupsilon.org
Membership: 60,000 initiated since 1928 (as of 2013)
President: Burrell Montz
Contact: www.gammathetaupsilon.org/contact.html

Description and purpose

Gamma Theta Upsilon (GTU) is the international honor society for geography. Its goals are to (i) strengthen professional interest in geography through student and professional training and academic experiences beyond the classroom and laboratory, (ii) advance the status of geography as a discipline for study and investigation, (iii) encourage high quality student research, (iv) serve as an outlet for research publication, and (v) administer financial support for furthering study and/or research in the field.

Journals or major publication series


Current activities or projects

GTU supports student research and training that advances the discipline of geography. To that end, the organization awards scholarships to undergraduates and graduate students pursuing degrees in geography. In 2013 GTU began a partnership with the American Association of Geographers (AAG) to offer conference travel awards for student members of GTU. GTU also sponsors the Visiting Geographical Scientist Program, which helps institutions bring well-known geographers to speak at colleges and universities. GTU publishes Geographical Bulletin twice annually. The Bulletin is an outlet for students as well as others in the field.

Brief history

Gamma Theta Upsilon traces its origin to the efforts of Dr Robert G. Buzzard. While a graduate student at Clark University in the mid-1920s, Dr Buzzard was influenced by the Clark University Geographical Society (CUGS). CUGS was comprised of students and graduate-level alumni of Clark University. Seeing the value of organizations like CUGS, Dr Buzzard established a geography club at Illinois State Normal University where he served as a professor of geography. On May 15, 1928, Gamma Theta Upsilon was chartered within the state of Illinois as a professional geography fraternity with 33 inaugural members.

For the next three years, GTU thrived as a local organization. At the same time proposals were developed for expanding GTU into a national organization and in the spring of 1931 charters were granted to four additional chapters with Illinois State’s chapter being known as the Alpha Chapter. Since that time, more than 200 GTU chapters have been established, and new chapters are added almost every year. In 1972, as a method to strengthen ties among chapters and among all members of Gamma Theta Upsilon, regional divisions were established that correspond generally to the regional divisions of the American Association of Geographers.

Submitted by Tom Wikle
An unmanned aerial vehicle (UAV) is a flying object without a human pilot onboard. It can be piloted either by a human operator on the ground using, for example, a radio controller (RC) manually, or by an onboard computer-controlled pilot system, the so-called autopilot, autonomously or in a semi-automatic mode. Other terminologies referring to more or less the same thing as UAV are, for example, remotely piloted aircraft (RPA), remotely piloted air system (RPAS), and drones, which were used in the early days and mostly in the military, and nowadays the widely accepted unmanned aerial system (UAS) or unmanned aircraft system (UAS), which emphasize the system components.

Typically, a UAS consists of two main components: a UAV or aerial platform and a ground control station (GCS), which communicate with each other via data links. Onboard a UAV there are, generally, a RC receiver, a data transmitter and receiver pair, different sensors — for, for example, engine temperature, rotation rate, fuel, and power supply — and an autopilot for autonomous missions. The autopilot, the “brain” of a UAV, is mainly composed of a computer, a GNSS (global navigation satellite system) receiver, and an IMU (inertial measurement unit) with some supporting sensors such as barometer and magnetometer. Based on GNSS signals, the IMU, and other sensor data the computer generates pilot commands for the aircraft. Corresponding to the onboard configuration, a GCS may consist of, for example, an RC for a human pilot, a data transmitter and receiver pair for communication, and a computer with ground control software for mission planning, real-time flight state viewing, and/or generating pilot commands.

UAVs can be classified in different ways. Based on the working principle, UAVs can be divided into two categories: fixed-wing and rotary-wing aircraft. Fixed-wing aircraft, commonly known as airplanes, are those which use wings that generate lift caused by the vehicle’s forward airspeed and the shape of the wings. An example of fixed-wing UAV is the Wing Loong.

Rotary-wing aircraft or rotorcrafts, commonly known as helicopters, are those that use lift generated by wings, also called rotor blades, that revolve around a mast. Depending on the type of rotor system, helicopters can be of, for example, main and tail rotor, coaxial rotor, tandem rotor and multirotor design. They fall into the category of vertical take-off and landing (VTOL) aircraft. Some examples of rotary UAVs are shown in Figure 1. A classification based on lighter or heavier than air and powered or unpowered can be found elsewhere (Eisenbeiss 2009).

UAVs can also be classified according to their capabilities and functionalities. For example, based on flight altitude, endurance, data link range, maximum take-off weight (MTOW), payload and size, Bento (2008) presented a comprehensive classification of UAVs. According to Bento, there are four main categories: MAV (microaerial vehicle)/mini
UNMANNED AERIAL VEHICLE (UAV)

UAVs, tactical UAVs, strategic and special task UAVs, and seven subcategories, amongst others, MALE (medium altitude and long endurance) and HALE (high altitude and long endurance). From a geodata acquisition point of view, the payload capacity of a UAV is of special interest, as the sensing system configuration onboard defines the scope of its application. Based on the payload capacity, UAVs can be classified into ultra-light (<2 kg), light (<10 kg), medium (<50 kg), and heavy (>50 kg) categories.

UAVs have a long history of use in the military. The first recorded use of UAVs dates back to as early as the mid-1800s, when Austrians sent off unmanned air balloons filled with bombs to attack Venice. The first pilotless aircraft were built during and shortly after World War I. During the course of World War II, various early types of UAVs were built and used as targets for anti-aircraft gunners training as well as attack missions, for example, drones from the “Radioplane Company” of Reginald Denny, which built nearly fifteen thousand drones for the US army during World War II. While during the Cold War period, the successes of drones as training targets and for attack missions was
extended to data collection and reconnaissance missions. The US military used drones to collect radioactivity data from nuclear tests. During the Vietnam War, the US air force launched 3435 reconnaissance drones over North Vietnam and its surrounding areas, at a cost of about 554 drones lost.

Battlefield UAVs have been developed since the 1980s. As communication technology advanced, real-time reconnaissance, including video and data, became possible. Long endurance UAVs, including solar powered, were also developed. The longest endurance was the QinetiQ Zephyr Solar Electric, which stayed in the air for 336 hours and 22 minutes from July 9–23, 2010.

In surveying and mapping, the early attempts to use UAV as platform employed balloons, kites, pigeons, and rockets around 1900. Whittlesey (1970) used a tethered balloon for archaeological documentation. One of the earliest experiments in photogrammetry using a fixed-wing UAV was carried out by Przybilla and Wester-Ebbinghaus in 1979 (Przybilla and Wester-Ebbinghaus 1979). Wester-Ebbinghaus was also the first to use a rotary-wing UAV for photogrammetric purposes in 1980 (Wester-Ebbinghaus 1980). Vozikis (1983) presented an example of using a balloon-mounted medium-format camera (a Hasselblad 500 EL/M) for photogrammetry. Marks (1989) gave another example of using a balloon for photogrammetry of historical sites. Wanzke (1984) used a hot-air ship for stereo photogrammetric documentation.

As consumer electronics advanced, radio controlled model planes and helicopters were built in such ways that they could be used as platforms to carry cameras to photograph the ground from the air. For example, Tokmakidis et al. proposed a manually controlled model helicopter carrying a Hasselblad camera to photograph the Tholos in Delphi (Tokmakidis et al. 1998) to map an archaeological excavation (Skarlatos, Theodoridou, and Glabenas 2004; Tokmakidis and Scarlatos 2002). In 2000, Zischinsky, Dorfner, and Rottensteiner (2000) used a model helicopter to take images of a historical mill in order to generate a 3-D model. The RMAX rotary-wing UAV system from Yamaha was used for various surveying tasks (Eisenbeiss 2003). The first UAV-borne LiDAR (light detection and ranging) system for 3-D mapping was carried out at Carnegie Mellon University (Thrun, Diel, and Haehnel 2003). Nagai was also among the first to use a UAV integrated with LiDAR scanner, camera, and GPS/IMU for mapping (Nagai et al. 2004).

In 2004, at the ISPRS (International Society for Photogrammetry and Remote Sensing) congress in Istanbul, Turkey, a new intercommission working group – IC WG I/V for autonomous vehicle navigation, including all aspects of navigation of spacecraft, aircraft, UAVs, land vehicles, robots, ships, and underwater vehicles – was initiated. Since then, more work has been conducted using UAVs for surveying and mapping; most of this work was presented at the 2008 Beijing congress of the ISPRS. Everaerts (2008) gave an overview of UAVs used in photogrammetry during the period 2004–2008.

In 2011, the first unmanned aerial vehicle in geomatics conference, UAV-g-2011 (http://www.geometh.ethz.ch/uav_g/index), was held in Zurich; the second unmanned aerial vehicle in geomatics conference, UAV-g-2013 (http://www.uav-g.org/), was held in Rostock, Germany in 2013. The open access journal of Remote Sensing also published a special issue “Unmanned Aerial Vehicles (UAVs) Based Remote Sensing” in 2012 (https://www.mdpi.com/journal/remotesensing/special_issues/uav).

Compared to traditional platforms, such as manned airplanes or satellites, UAVs have the following pro and cons:
UNMANNED AERIAL VEHICLE (UAV)

- safe for humans, especially in high risk situations and inaccessible locations, such as mountainous and volcanic areas, flood plains, earthquake and highly radioactive disaster sites;
- high image resolution (both temporal and spatial) due to low flying altitude;
- hover and low speed;
- easy operations, no need for special launch and recovery facilities (runways);
- low cost, high efficiency;
- small coverage;
- small payloads.

As UAVs were developed for military use from the start, military uses are still the major part of their application; these can fall into three categories: (i) target and decoy – providing ground and aerial gunnery a target that simulates an enemy aircraft or missile; (ii) reconnaissance and surveillance – providing battlefield intelligence; (iii) combat – providing attack capability for high-risk missions.

In civil or commercial applications, UAVs are used as delivery platform or platforms to carry various sensors for data collection. Sensors include, but are not limited to:

- digital cameras, from small consumer products to high-end professional digital cameras;
- video cameras;
- infrared cameras, thermal cameras;
- multispectrum sensors;
- hyperspectral sensors;
- mini synthetic aperture radar (SAR);
- LiDAR;
- vertical cavity surface emitting lasers (VCSELs).

There are many applications of UAVs. It is not the intention here to provide a complete reference list but rather examples. Geoscience is one of the major civilian applications for UAVs, as traditional data acquisition in this area is either satellite or airborne imagery. With the help of UAVs, the range of data acquisition is extended to very low altitude, thus expanding the range of airborne remote sensing from space to near ground. The low flying altitude also provides the possibility to obtain both spatial and temporal high-resolution images, which was not possible before.

UAVs are well suited as image acquisition platforms for 3-D reconstruction, DEM (digital elevation model) or DSM (digital surface model) generation and orthophoto generation. Digital camera technology and onboard GPS and IMU data have made georeferenced full motion video (FMV) in real-time or near real-time, and being viewed within a GIS, possible. This adds another layer of information to the system and is very valuable to many time-critical decision-making situations.

Environmental surveying and monitoring is another good example of UAV application. The ease of operation of UAVs can provide quick and frequent flights allowing land and water monitoring. Dense point clouds can also be used for excavation volume computation and natural resources documentation. Archaeology and cultural heritage site mapping and documentation are typical UAV tasks. 3-D documentation and mapping of sites and structures are easily achieved with a low-altitude image-based UAV survey platform. UAV-borne LiDAR systems can be used for forest inventory, forest fire detection, crop growth monitoring, and precision farming, as well as for the inspection of oil and gas pipelines and power lines.

UAVs specifically designed for cargo and logistics operation, in particular, in time-critical situations or otherwise inaccessible regions, provide extra value to the logistical business. For example, Amazon is testing drones to deliver packages to customers (http://www.foxbusiness.
In scientific research, UAVs also play an important role (Marris 2013). Scientists in various fields nowadays use UAVs as tools to collect data that were difficult or not possible to collect before. Vision-based navigation of UAVs is one of the research areas that interests many researchers. A research group at the University of Oxford, UK, uses fish-cameras mounted on drones to “find” drones themselves using visual information. There is a widespread “TED” talk video by Vijay Kumar from the University of Pennsylvania in Philadelphia that demonstrates multiple quadrotors, that is, MAVs, performing synchronized tasks (Kumar 2012).

Other civil and commercial applications include: search and rescue, traffic surveillance and domestic policing, videography and the filmmaking industry, communication relays – one of the more interesting applications of UAVs. In 2014, Google and Facebook each brought companies that make high altitude drones, Titan Aerospace and Ascenta, respectively (http://techcrunch.com/2014/03/27/facebook-drones/, http://techcrunch.com/2014/04/14/google-acquires-titan-aerospace-the-drone-company-pursued-by-facebook/). It was reported that one of the intended uses of both drones is communication-relay, for example to bring the Internet to the third world via drones.

Regulations concerning UAVs

The practices of UAVs in the civilian sector, especially in low-altitude range (below 500m), are very country specific. Due to safety and security considerations, many countries prohibit UAV flying over populated areas, for example, cities. In other places, for example, in Europe, UAV flights have to be backed up by a qualified pilot with line of sight (LOS), who can take over control of the UAV at any time if it is necessary. In Europe, UAVs with an MTOW over 150 kg have to be certificated by the EASA (European Aviation Safety Agency), while UAVs with an MTOW less than 150 kg are regulated by national authorities. For example, in the United Kingdom, the Civil Aviation Authority (CAA) issued guidance “Unmanned Aerial Vehicle Operations in UK Airspace” in 2004 (CAA 2004). Regulations are one of the biggest barriers that limit the expansion of uses for UAVs in the civilian sector. Safety is the main concern. On the other hand, the rapid development and advances of UAV technologies have made it difficult for the regulatory authorities to “catch up.”

Communities

There are a number of organizations and communities related to UAVs, including governmental, industrial, academic, and open source concerns:

- The European Association of Unmanned Vehicles Systems (EUROUVS)
- The Unmanned Aerial Vehicle Systems Association (UAVS), https://www.uavs.org/
UNMANNED AERIAL VEHICLE (UAV)

- Paparazzi: open-source autopilot hardware and software, http://wiki.paparazziuav.org
- DIY drones community, www.diydrones.com
- APM: open-source autopilot hardware and software, http://plane.ardupilot.com
- Pixhawk: open-source autopilot hardware and software, https://pixhawk.org/
- Dronecode: an open source, collaborative project that brings together existing and future open source drone projects under a nonprofit structure governed by The Linux Foundation: https://www.dronecode.org/about

SEE ALSO: Geographic information system; Global navigation satellite systems; Photogrammetry: 3-D from imagery

References


The term unsupervised classification originated from the machine learning field, a branch of artificial intelligence, as a paired concept contrasted to supervised classification. Both techniques have the common objective of categorizing data into a number of classes or clusters. This can be formulated as

\[ F(X) = Y \]

where \( F \) is a classification algorithm (or a classifier, or a classification rule), \( X \) is data, and \( Y \) is categorical class labels that have been or will be assigned to data. A primary difference between unsupervised and supervised classification lies in use of training data in the process of obtaining \( F \). Training data

\[ S_{\text{train}} = \{(x_i, y_i)\} \]

refer to data samples whose class labels \( y_i \) are known for each data point \( x_i \) prior to the analysis. Supervised classification “trains” the classifier \( F \) by analyzing the training data and applies the trained classifier to other data samples that do not have class labels yet.

On the other hand, unsupervised classification classifies data samples without such a priori information on the data–class association. What unsupervised classification relies on instead is data information that is present in the data themselves, namely inherent properties such as similarity between data samples or density. While the forms of unsupervised classification vary in many algorithmic aspects, the underlying ideas of the algorithms are the same – determining the regions where the samples are densely populated (high similarity) by making them disconnected in the sparsely populated regions (low similarity). This explains why the terms clustering analysis and unsupervised classification can be used interchangeably.

Major unsupervised classification algorithms

Hundreds of unsupervised classification algorithms have been proposed in the literature of various scientific and engineering disciplines. Accordingly, it is difficult to categorize and review all the published algorithms. Three major frameworks that embrace unsupervised classification algorithms that are frequently used in the studies are introduced here.

Iterative algorithms

Iterative unsupervised classification algorithms seek for optimal classification of data by repeating certain procedures that are expected to improve classification results. One of the most widely used iterative algorithms is K-means clustering algorithm. The key process of the algorithm is first to assign each sample a class label of the class closest to the sample.
UNSUPERVISED CLASSIFICATION

\[ F(X) = \arg \min_{C_i} |X - X_{C_i}| \]

\[ X_{C_i} : \text{mean value of class } C_i \]

Once this assigning process is complete for every sample, the class means (centers) are then updated using the new assignment status. The class assignment and the mean computation procedures are repeated until a stopping criterion is satisfied. The overall scheme is illustrated in Figure 1.

K-means is relatively scalable (i.e., capable of processing large amounts data) and is easy to understand. However, it is sensitive to noisy data and outliers. Since the class means are required in a numeric form, categorical data cannot be processed by K-means. Classes having unbalanced sizes also cause problems since K-means tends to result in equal-size clusters.

The K-means algorithm might appear to be heuristic, but it has a mathematical foundation connected to the Gaussian mixture model (GMM) and the expectation-maximization (EM) algorithm. The EM algorithm is an iterative method for finding the maximum likelihood estimates of parameters in statistical models. Like the EM algorithm, K-means converges to local optima and has no guarantee of global convergence. To avoid convergence to local minima, a heuristic approach can, for example, be employed by repeatedly applying K-means with difference initializations (i.e., different arrangement of class means in the first iteration).

Different distance measures can be used for computing the distance between samples. While Euclidean distance is by far the most popular distance measure, diverse measures, such as spectral angular mapper (SAM) for high-dimensional data, can be employed depending on the type of applications and data.

Another popular iterative algorithm is ISO-DATA. ISODATA has a similar framework as K-means, but with an advanced feature that allows clusters to merge and split during the iteration. Clusters are split if the standard deviation

Figure 1  Illustration of K-means algorithm for \( k = 3 \).
of samples in a cluster is larger than a user-defined threshold and merge if the distance between cluster centers is less than a user-defined threshold. The advantage of ISODATA is that users do not need to assume the exact number of clusters beforehand. Disadvantages are that ISODATA can be extremely time consuming for unstructured data and may converge to local optima as in the case of K-means.

**Spectral clustering**

Unsupervised classification that is based on graph theory is referred to as spectral clustering. In spectral clustering, a data graph is first constructed by connecting data samples, that is,

\[ G = (V, E) \]

where V is vertices representing data points and E represents edges between data samples. Edges are then weighted by similarity between vertices. Classification takes place by finding an optimal partition of vertices such that the weights of edges between the clusters are minimum (maximum dissimilarity), while the weights inside the clusters are maximum (minimum dissimilarity). The solution is obtained by solving the eigenproblem of a graph Laplacian L. There are many variations in graph Laplacians, but one simple definition is

\[ L = D - W \]

where D is a matrix whose i-th diagonal entries are filled with \( d_i = \sum_j w_{ij} \) and W is the weight matrix given as \( W = \{ w_{ij} \} \).

Spectral clustering has shown good results on relatively complex data sets and has an advantage in that the implementation is easy, requiring only a standard linear algebra without the need to assume any statistical models. The disadvantages are that the computation in solving an eigenproblem may be demanding for a large data set and the classification results are sensitive to the choice of parameters.

**Hierarchical clustering**

Hierarchical clustering is an unsupervised classification algorithm that exploits a hierarchical structure. Two approaches, agglomerative (bottom up) and divisive (top down), are possible. In the agglomerative approach, each data sample forms a single cluster in the beginning and pairs of clusters merge as the level in the hierarchy goes up. The divisive approach starts from one cluster that contains all data samples and proceeds by splitting clusters successively until each data sample forms a cluster. Cluster pairs are selected for merging or splitting based on a distance measure and a linkage criterion between clusters. As in the iterative algorithms, various distance measures can be used for computing the distance between samples; Euclidean, Mahalanobis, Manhattan distances, and cosine similarity. Linkage criteria determine which pair of samples (two samples each of which is selected from each cluster) is to be used for representing the distance between the clusters. Complete linkage criterion defines the distance between clusters as the distance between samples, each of which is from each cluster, that are maximally distant compared to the other sample pairs. Single-linkage criterion uses a minimum distance and average linkage uses the mean distance.

One of the advantages of hierarchical clustering is that the number of classes does not need to be predefined before the analysis. Instead, users can decide the optimal number of classes by analyzing the hierarchy, which is usually visualized in a dendrogram. An example of a dendrogram for human genetic diversity is presented in Figure 2. This advantage comes at the cost of computational load, which stems from its greedy–manner implementation (the distances of newly merged or split clusters to other clusters need to be compared every time the merge/split occurs). Hierarchical clustering is a decisive algorithm and has the disadvantage
Applications of unsupervised classification

Applications of unsupervised classification are countless in various disciplines of social and natural science. Regardless of types of data to be analyzed, the development of typologies is probably the most popular use of unsupervised classification. A clinical study that was performed to analyze the relationship between dietary patterns and body is such an example (Newby et al. 2003). In this study, 459 healthy men and women were classified based on their dietary intake into five clusters – healthy pattern,
white-bread pattern, alcohol pattern, sweets pattern, and meat-and-potatoes pattern – and their characteristic variations in body mass and waist circumference were analyzed for individual clusters.

In the field of geography, unsupervised classification is often combined with geographical location of data to investigate the spatial patterns of the data. Since most unsupervised classification is usually designed without spatial consideration, a separate process is needed for integrating the geographic information into the framework. The easiest solution is to treat the spatial information as one of the data attributes. In that case, a careful normalization of the spatial information is required to balance with other attributes. The second approach is to use unsupervised classification schemes that have a spatial consideration from the beginning (e.g., image segmentation).

There are numerous applications of unsupervised classification to remote sensing. Unsupervised classification has been effectively used for identifying spatially congruent regions in airborne and space-borne images. Examples include land cover (e.g., vegetation, agricultural fields, forest area, mineral fields, snow cover, etc.), oceanic events (e.g., algal blooms, ocean currents, sea ice, cold pool, etc.), and atmospheric features (e.g., cloud cover, aerosol density, sand dust, typhoon trajectory, smoke plumes, etc.). The results of unsupervised classification are used not only as independent final outcomes when it has sufficient interpretability, but also can be used as a prior step of supervised classification when users have no clues on which class categorization will be meaningful for the data to be analyzed.

**SEE ALSO:** Hyperspectral remote sensing; Machine learning; Microwave remote sensing; Optical remote sensing; Supervised classification

**References**


**Further reading**


Urban biogeography

C.Y. Jim
University of Hong Kong, China

Urban biogeography explores the spatiotemporal intricacies of the nature-in-city realm. Half of the world’s human population now lives in cities. The built environment is apparently hostile to nature, yet it offers many overt and latent opportunities for nature to thrive. Urban greenery provides cost-effective means to improve environmental quality and improve the quality of life of 3.6 billion people. It offers a fertile research theme for scholarly and application-oriented inquiries (Jim 2011a).

Humans shifted from the primitive nomadic tenure to settlements to seek relief from nature’s vicissitudes. After acquiring security and solace in the artificial ambience, humans began to miss their natural companions. A selected cohort of plants with practical functions such as edible harvests and ornamental characters was cultivated in domestic spaces. Artificial selection of plants with naturally selected and human-favored traits provided the basis to screen species for a secluded horticultural palette. Plants growing spontaneously in incidental ruderal sites would have fueled the inspiration to drive the innate desire to emulate nature in urban spaces.

The rise of aristocracy initiated large-scale inclusion of nature in human abodes. They engendered the means and taste to create extensive private gardens of stately homes to satisfy horticultural, recreational, hunting, and status-symbol quests. Such human-modified green spaces generated rather novel conditions for selected plants to exist under human care and protection from competitors. Landscape design was developed to regiment nature and present nature on human terms. Nature is allowed to get close to humans in a tamed and filtered form.

The Industrial Revolution pulled people en masse to cities, which trapped multitudes of factory workers in abysmal living conditions. The negativity of extreme severance from nature led to the institution of public parks. They were pioneered by the first purpose-built urban park, the Princess Park in the Liverpool suburb of Toxteth, England, in 1843. Henceforth, the urban park institution has been adopted worldwide as a de rigueur infrastructural component of town planning to embellish cities and encourage salubrious outdoor recreation. Some urban parks were converted from former royal hunting grounds, whereas many were created from scratch.

Before the advent of urban parks, people and nature found ample chances in miscellaneous green sites embedded in the city. Urban parks denote a formal category in the urban green spaces (UGS) typology. They tend to be larger and receive a higher level of design and management inputs. Their creation and maintenance demand massive human inroads into key issues such as floristic membership, means of sustenance, and exclusion of unwanted organisms (weed and pest organisms). They are accompanied by a plethora of informal UGS with less human intervention, permitting nature to flourish of its own accord as habitat islands.

With reference to abiotic and biotic ecosystem components and constituent ecological linkages, processes, and functions, UGS are subsumed under a hemeroby sequence in terms of naturalness vis-à-vis human mediation: (i) remnant or
inherited natural enclaves left by default or design; (ii) incidental ruderal pockets left unallocated or unused in the course of urbanization; (iii) emulated nature which are created green areas that duplicate nature in terms of species composition and biomass structure; and (iv) artificial green sites that follow a human-dominated landscape style characterized by manicured and simplistic design. In urban ecological planning, a wide range of UGS can be preserved or created. Few natural enclaves in cities can sustain their pristine conditions and most sites carry the telltale symptoms of human modifications.

Diverse UGS furnish a broad spectrum of habitats to accommodate a correspondingly varied flora–fauna assortment. A large UGS can embrace a heterogeneous habitat mosaic. Each habitat type supports a species assemblage with some unique members in the core and edge respectively. The interface between habitat types generates transitional or ecotone habitats to attract biota somewhat different from the core. Permutations of core-cum-ecotone species assemblages present significant aggregate urban biodiversity, which can sometimes exceed the pre-urbanization state or surrounding countryside. UGS management should respect inherent ecological makeup and functions and allow nature to run its own course. Overzealous management can degrade habitat conditions for survival of species and ecological processes.

By biogeographical provenance, the urban flora is divided at the high level into native (indigenous) and exotic (alien or introduced). Natives can dwell in natural or anthropogenic sites (apophytes). Exotics can similarly exist in natural or anthropogenic sites. Some may reproduce of their own accord in the new home to become naturalized (adventives). Exotics reaching a recipient site before 1500 are archeophytes, and after, neophytes. Natives or exotics growing in human-modified seminatural sites are ruderals.

Some ruderals originated as cultivated garden plants, and subsequently spread spontaneously as garden escapees. UGS are prone to advertent and inadvertent human disturbances to induce gradual or drastic changes in species composition and ecosystem integrity. At the macroscale, frequent exchanges of people and goods via different transport modes across city boundaries provide opportunities for species introduction. Native and exotic species dwelling in the city’s countryside can move into urban areas. Exotic species can leapfrog directly into cities in jump dispersal, bypassing intermediate migration steps. Such expansion of species’ geographical range involves multiple modes of dispersal through multiple pathways over different time periods.

At the microscale, the built environment and visitor impacts can impose unnatural changes in UGS (Jim 2011b). Seed dispersal within the city could be hampered by unnatural and daunting biogeographical barriers of tall and tightly packed buildings and hostile environmental conditions lying between source and recipient habitats. Without adequate recruitment to replenish the young age bracket, some urban woodlands are suffering from a degenerating age structure dominated by old trees. Such poor regeneration dynamics are unfavorable to long-term sustainability. The heavy recreational use of natural sites may cause trampling, compaction, erosion, and vegetation damage to modify species composition. Visitor can bring seeds unintentionally on their clothing or bags. Sensitive species could be outcompeted whereas aggressive species can take over their niches. Invasion by aggressive exotics into a natural site can trigger fundamental floristic changes. Ecosystem succession could be arrested by deleterious human and natural drivers.

Nature-in-city can maximize holistic ecosystem services to more people. The benefits
are readily enjoyed by residents living in close proximity to UGS. The spatial spillover and penetrative effects of some ecosystem services could favor a larger hinterland. The main ecosystem services include (i) cooling due to latent heat absorption of evapotranspiration; (ii) cooling due to shading of solar radiation; (iii) absorption of some gaseous air pollutants in the process of gas exchanges associated with photosynthesis and respiration; (iv) dry deposition by gravitation settlement or wind impaction of particulate pollutants on leaves followed by rainfall washing into the soil; (v) absorption of carbon dioxide in photosynthesis leading to fixation in tissues and carbon sequestration; (vi) generation of fresh nascent oxygen in photosynthesis to improve air quality; (vii) infiltration of rainwater into the soil to reduce loading on urban stormwater drainage system; (viii) recharge of groundwater aquifer in the urban watershed; (ix) filtration and physicochemical cleansing of water as it passes from soil to groundwater; and (x) habitat islands and corridors as forage and breeding sites for urban wildlife to enhance urban biodiversity.

It is perhaps paradoxical that highly natural enclaves often provide excellent ecosystem services, yet they usually do not require human inputs and management. The manicured UGS that require costly inputs to create and maintain, however, furnish less ecosystem services. It is important for municipal authorities to preserve high-quality natural enclaves in new urban developments. In developing countries, the practice of obliterating nature in new development areas, and then laboriously create new UGS at high cost but with low ecosystem service value, should be replaced by keeping nature’s gifts.

The ecological or naturalistic design can be adopted to emulate nature’s multiplicity, complexity, variability, changeability, and flexibility. The species choice should follow a high degree of fidelity to the relevant local or regional ecosystem. UGS can cater to both people and nature, and contribute to nature conservation and urban and regional biodiversity. It can be exemplified by nurturing the climax of urban woodlands to achieve the most complex and self-sustaining urban ecosystems with reference to native-species content, vegetation cover, biomass structure, leaf area index, and species richness. They demand notably less capital cost and recurrent management inputs than formal UGS, but provide biodiversity hotspots with significant ecosystem services.

In location terms, UGS are found in the urban core, urban fringe, suburb, and countryside envelope. Besides the ground level, greenery is increasingly enlisting the neglected potential of the vertical dimension. The huge total surfaces of buildings are amenable to vegetation cover. With innovative design and materials, green roofs can be installed on rooftops, terraces, and podiums and green walls can decorate low to high floors. The skyrise gardens can adopt different plant growth forms, from turfgrass to herb, shrub, and tree, representing various biomass structure and ecosystem services. Sky woodland can establish on roofs with proper design of soil and drainage and sufficient roof-slab loadbearing capacity. Three-dimensional greening is important for compact cities which lack ground-level UGS space. UGS patronage is notably influenced by accessibility. Most people prefer to walk less than 10 min or less than 400 m to reach a venue. UGS in core areas with high population density can serve more people living in the hinterland. However, core areas usually do not have sufficient land for liberal UGS provision. Green roofs are highly accessible to residents and workers of the building and surroundings.

The geometry (size and shape) of UGS can be optimized based on landscape ecology principles. UGS denote patches occluded in the built-up urban matrix. Individual sites should preferably
be linked by habitat *corridors* or *greenways* to enhance their connectivity. Moreover, they could be linked to natural ecosystems in the city’s surrounding countryside. The equilibrium theory of island biogeography (MacArthur and Wilson 1967) stipulates that well-connected UGS can support more species and are more resistant to species’ decline or extinction. The habitat continuity permits propagules, animals, and their genes to literally flow and exchange among daughter urban and parent extra-urban ecosystems. The spatially connected natural patches and corridors can create a *green network* to permeate the urban matrix, taking ecosystem services to people’s doorsteps. Such a configuration can reduce the distance between residents and UGS, rendering them more accessible and more used. Small gaps between patches could be crossed by wildlife, using a series of sites as *stepping stones* to penetrate the city. Urban development pressure, however, can shrink and dismember UGS to induce undesirable habitat isolation and fragmentation, which can trigger biotic pauperization.

Studies of UGS benefits have extended from environmental-ecological to socioeconomic. UGS should adopt a socially inclusive design to attract users from different strata of society. They can contribute to social interaction, social cohesion, social integration, and social harmony. By encouraging outdoor activities they foster a healthy life style to nurture happy, healthy, successful, and productive citizens. The economic benefits include increasing property value to raise government tax revenues which could be plowed back to develop the community. People are willing to pay more to acquire residential properties to fulfill a hedonic wish of proximity to UGS and pleasant green views. UGS are increasingly factored into green infrastructure development of cities, expected to ameliorate the urban heat island effect, and contribute to climate adaptation and climate resilience of cities.

**SEE ALSO:** Alien and native species; Biodiversity; Built environments; Climate adaptation/mitigation; Ecological footprint; Ecosystem services; Environment and urbanization; Ethnobotany; Green infrastructure; Island biogeography; Nature conservation; Urban ecology

**References**


**Further reading**


What is urban climate?

Human settlement, through the development of towns and cities, affects both the properties of the surface and that of the atmosphere above them. Urban regions therefore develop distinct urban climates in which the mean characteristics of the atmosphere within, above, and downwind of cities are modified by the city. These modifications act upon the regional and large-scale climate of a location that is determined by the latitude, elevation, and continental interior or coastal location and further modulated by the specific topographic setting of the city (e.g., valley bottom, lake shore). Urban climates represent an unintentional modification of climate. Understanding these changes is important because they provide a basis for minimizing negative effects of cities on their local climate and maximizing positive aspects that can improve the environmental quality and sustainability of the urban environment.

The urban surface

Much of the urban climate influence arises because of changes to surface characteristics in cities. The surface is important because it is the site at which energy, mass, and momentum are exchanged with the atmosphere. Compared to natural areas, cities alter the structure, cover, and fabric of the surface. Urban surface structure is determined by the arrangement and spacing of structural elements such as buildings and trees (Figure 1; Table 1) and is distinctly three-dimensional. The cover represents the main fractional components of vegetation, built-up, paved, bare soil, and water areas. The climatic properties of the fabric include radiative, thermal, roughness, and moisture properties. The cover and properties vary across the city and can be categorized into local climate zones (Figure 1; Table 2). Cities also incorporate a metabolism associated with city functions. Important metabolism processes include emissions of heat, water vapor, and pollutants that come from energy and water use and industrial processes.

The urban atmosphere

The bottom portion of the troposphere in contact with, and directly influenced by, the Earth’s surface is known as the atmospheric or planetary boundary layer (PBL). The PBL height varies with time and location; it is highest by day (1–3 km) under fair weather when convective turbulence from the surface is strongest. It is lower as winds and cloud cover increase (Figure 1a). At night, as convective mixing from the surface subsides, a shallow stable layer of air develops upwards from the surface that may be several hundreds of meters in height above which is an overlying residual layer that is effectively decoupled from the surface.

Over large cities, the PBL is influenced throughout its depth by the underlying urban surface and is referred to as the urban boundary
layer (UBL). The UBL is typically slightly higher than rural PBL by both day and night and exhibits modified characteristics (temperature, humidity, winds, pollution). The UBL is advected downwind by the regional wind as a plume of modified air (Figure 1a). This extends the influence of urban climate to larger scales.

The bottom 10% of the UBL is known as the surface layer. Here, the urban surface plays an
Table 1  Components of the urban surface

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facet</td>
<td>Roof, road, wall, lawn</td>
</tr>
<tr>
<td>Element</td>
<td>Distinct structural elements such as buildings and trees</td>
</tr>
<tr>
<td>Street or urban canyon</td>
<td>Air volume above a street that is bordered by buildings on either side</td>
</tr>
<tr>
<td>Block</td>
<td>Grouping of building elements defined by adjacent urban canyons. Maybe regular or irregular in shape</td>
</tr>
<tr>
<td>Neighborhood or LCZ</td>
<td>Homogeneous land use and land cover characteristics that define local climate zones (LCZ)</td>
</tr>
<tr>
<td>City</td>
<td>Entire urban area</td>
</tr>
</tbody>
</table>

important role in the climate. From the ground up to a height of about twice the average height of the main surface elements (buildings, trees) is the roughness sublayer (RSL, Figure 1a). In this layer, the wind patterns, receipt of sunlight, temperature, pollutant concentrations, and so on, show microscale variations due to the nearby buildings and trees. The urban canopy layer (UCL), where most of the population lives and where most urban activities takes place, lies within the RSL and is defined by the mean height of the surface elements (Figures 1b, 1c). Above the RSL lies the inertial sublayer (ISL). Here, mixing of the atmosphere is sufficient to blend the microscale variability of the RSL below so that its characteristics are representative of the underlying urban surface. Within the ISL we can make representative neighborhood-scale measurements of the urban climate.

Physical basis of urban climates: energy balance

Energy balance

The physical basis for urban climates is the energy and mass exchanges between the surface and the atmosphere. The energy balance of a surface is a conservation of energy equation. It represents the radiative, convective, and conductive exchanges between the surface, atmosphere, and substrate. For urban surfaces, because of their three-dimensional structure, we consider the energy balance of a volume (Figure 2) written as (neglecting advection):

\[ Q^* + Q_E = Q_H + Q_E + \Delta Q_S \]

\( Q^* \) (units, W m\(^{-2}\)) represents the energy supplied to the urban system from the net absorption of radiation from the sun and atmosphere. It is positive by day (net energy gain) and negative at night (net energy loss). \( Q_E \) is the anthropogenic heat flux from human energy use. Energy transfers between the surface and atmosphere occur as sensible (\( Q_H \)) and latent (\( Q_E \)) heat through turbulent convective mixing of air, and as the change in heat stored via conduction in the urban materials (\( \Delta Q_S \)). Terms are defined as positive when the flux is away from the surface (into the atmosphere for \( Q_H \) and \( Q_E \) and into the substrate for \( \Delta Q_S \)) and negative when toward the surface.

The energy balance is driven largely by the net radiation (\( Q^* \)). The radiation budget of a surface

\[ Q^* = K \downarrow - K \uparrow + L \downarrow - L \uparrow \]
Table 2 Important properties of urbanization that affect urban climate with sample values for two urban local climate zones (LCZ) and one rural. Change* indicates the general difference of urban from rural values. Based on Stewart and Oke (2012).

<table>
<thead>
<tr>
<th>Property</th>
<th>Description and units</th>
<th>Change*</th>
<th>LCZ 1</th>
<th>LCZ 6</th>
<th>LCZ D</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Geometric and surface cover properties</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sky view factor</td>
<td>Amount of sky hemisphere visible relative to that of an unobstructed surface</td>
<td>−</td>
<td>0.2–0.4</td>
<td>0.6–0.9</td>
<td>&gt;0.9</td>
</tr>
<tr>
<td>Aspect ratio</td>
<td>Ratio of building height to street width</td>
<td>+</td>
<td>&gt;2</td>
<td>0.3–0.75</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Building surface fraction</td>
<td>Ratio of building plan area to total plan area</td>
<td>+</td>
<td>0.4–0.6</td>
<td>0.2–0.4</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Impervious surface fraction</td>
<td>Ratio of impervious plan area to total plan area</td>
<td>+</td>
<td>0.4–0.6</td>
<td>0.2–0.5</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Pervious (vegetation, bare soil, water)</td>
<td>Ratio of pervious total plan area to total plan area</td>
<td>−</td>
<td>&lt;0.1</td>
<td>0.3–0.6</td>
<td>&gt;0.9</td>
</tr>
<tr>
<td>Height of roughness elements</td>
<td>Geometric average of building and tree or plant heights (m)</td>
<td>+</td>
<td>&gt;25</td>
<td>3–10</td>
<td>&lt;1</td>
</tr>
<tr>
<td>Aerodynamic roughness length</td>
<td>Extrapolation of logarithmic wind speed profile to zero(m)</td>
<td>+</td>
<td>≥2</td>
<td>0.25–0.5</td>
<td>0.03–0.1</td>
</tr>
<tr>
<td><strong>Thermal, radiative, and metabolic properties</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surface thermal admittance</td>
<td>Ability of a surface to accept or release heat Units: J m$^{-2}$ s$^{-1/2}$ K$^{-1}$</td>
<td>+</td>
<td>1500–1800</td>
<td>1200–1800</td>
<td>1200–1600</td>
</tr>
<tr>
<td>Albedo</td>
<td>$K \uparrow / K \downarrow$</td>
<td>−</td>
<td>0.1–0.2</td>
<td>0.12–0.25</td>
<td>0.15–0.25</td>
</tr>
<tr>
<td>Anthropogenic heat</td>
<td>Heat flux density (W m$^{-2}$) from urban functions</td>
<td>+</td>
<td>50–300</td>
<td>&lt;25</td>
<td>0</td>
</tr>
</tbody>
</table>

is composed of inputs of shortwave radiative energy received from the sun ($K \downarrow$) with wavelengths of energy < 3 μm and longwave radiation from the atmosphere ($L \downarrow$). The upwelling radiation energy consists of reflected shortwave radiation ($K \uparrow$) and emitted and reflected longwave radiation ($L \uparrow$). The temperature of the surface largely determines $L \uparrow$; surfaces have temperatures that only permit them to emit longwave, not shortwave radiative energy.

All the radiation budget components, $Q_H$, and $Q_E$ terms may be directly measured from instruments mounted on tall towers in the ISL, while $Q_F$ is estimated from inventories, models,
or derived as a residual. The storage heat flux term is most often determined as a residual while \( \Delta Q_A \) is assumed to be zero over a homogeneous urban land-use area. Urban climates can also be studied using numerical models that use equations to represent the surface–atmospheric exchange processes. Such models are important to urban-scale weather forecasting and as a planning tool for helping design more sustainable cities.

**How does urbanization affect the energy balance?**

Figure 3 illustrates the energy balance for three sites along an urban–rural gradient in Basel Switzerland. Relative to most natural surfaces, urbanization leads to an increase in \( Q_H \) and \( \Delta Q_S \), and a decrease in \( Q_E \). The increase of \( Q_H \) is associated with the increase of impervious surfaces in urban areas and the fact they tend to be warmer and drier than the vegetated surfaces they replace. Greater turbulence over cities due to the interaction of winds with the rough urban surface also helps to enhance the sensible heat flux. The ratio of observed \( Q_H/Q_E \) (Figure 4) increases substantially as vegetated cover is reduced (and impervious surface cover increases) for sites of varying intensities of urbanization. Exceptions are cities in very dry climates where significant irrigation and planting may decrease the ratio. The three-dimensional surface structure of cities, which enhances absorption of radiation, and thermal properties of built materials enhance the uptake of heat by conduction and increases \( \Delta Q_S \) by day. This provides a store of heat, which is then released at night acting as source of heat for the urban area.

Changes to \( Q^* \) are often small because changes to the individual incoming and outgoing terms tend to offset (Table 3). Polluted atmospheres reduce \( K \downarrow \) at the surface. Absorption of sunlight along with warming from below increases the temperature in the UBL and increases \( L \uparrow \). Urban surfaces typically absorb more \( K \downarrow \), in part due to the 3-D surface structure that acts to trap the sunlight more effectively and in part due to changes in the reflectivity of urban materials. This reduces \( K \uparrow \). Warmer urban surfaces increase \( L \uparrow \).

The anthropogenic heat flux adds a contribution that is not present for natural surfaces. It comes from the use of energy for space heating and cooling, traffic, and electricity use (Figure 2). Its magnitude varies with the intensity of urban development and is sensitive to the large scale climate and seasonal variations. In the mid-latitudes, \( Q_F \) is larger in winter and smaller in summer (Table 2). \( Q_F \) is also relatively more important in winter because \( Q^* \) is then smaller. On a daily basis, \( Q_F \) is regulated by the cycles of human activity.

**Urban modifications to climate**

Changes to the surface radiation budget and energy balance in urban areas lead to modification of all the atmospheric characteristics that define weather and climate (Table 3).
Temperature

Urban areas are generally warmer than their non-urbanized surroundings. This so-called urban heat island (UHI) effect has been studied for well over a century. The relative warming extends upwards from the surface to the top of the UBL and downwards into the subsurface. The most studied urban heat island is that in the canopy layer air (UHI\textsubscript{UCL}) which is most easily accessed for measurement and represents the air volume in which the bulk of urban activities occur. Satellite-based sensors can provide an urban-scale depiction of the temperatures of the surfaces they view that is referred to as the surface heat island (UHI\textsubscript{SFC}).

The UHI\textsubscript{SFC} forms because the temperatures of most built materials are warmer than those of the natural surfaces they replace. The energy balance of the built infrastructure directs more energy into sensible rather than latent heat and leads to higher temperatures by both day and night.

In the canopy layer, the UHI\textsubscript{UCL} is best expressed under calm and clear weather conditions when differences between urban and rural energy balances that drive the heat island are largest (Figure 5). Under such conditions, rural areas cool quickly in the late afternoon and following sunset as heat conducted from the ground is lost by radiative cooling to the sky. In
urban areas, the rate of cooling in the UCL is slowed because the view of the sky for surfaces is obstructed, buildings provide a sheltering effect and there is a large store of sensible heat released slowly from the built materials (Table 4). This generates a nocturnal UHI\textsubscript{UCL} that grows

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Modification of radiation budget, energy balance, and climate elements in urban areas. Adapted from Oke (1997).</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Change</td>
</tr>
<tr>
<td>Radiation budget terms</td>
<td></td>
</tr>
<tr>
<td>Incident solar (shortwave) radiation</td>
<td>$-$</td>
</tr>
<tr>
<td>Reflected shortwave radiation</td>
<td>$-$</td>
</tr>
<tr>
<td>Atmospheric (longwave) radiation</td>
<td>$+$</td>
</tr>
<tr>
<td>Outgoing longwave radiation</td>
<td>$+$</td>
</tr>
<tr>
<td>Energy balance terms</td>
<td></td>
</tr>
<tr>
<td>Convective heat flux</td>
<td>$+$</td>
</tr>
<tr>
<td>Evaporation (latent heat flux)</td>
<td>$-$</td>
</tr>
<tr>
<td>Storage heat flux</td>
<td>$+$</td>
</tr>
<tr>
<td>Anthropogenic heat flux</td>
<td>$+$</td>
</tr>
<tr>
<td>Climate parameters</td>
<td></td>
</tr>
<tr>
<td>Air Temperature</td>
<td>$+$</td>
</tr>
<tr>
<td>Surface temperature</td>
<td>$+$</td>
</tr>
<tr>
<td>Humidity</td>
<td>$+$</td>
</tr>
<tr>
<td>Wind speed</td>
<td>$-$</td>
</tr>
<tr>
<td>Wind direction</td>
<td>$+$</td>
</tr>
<tr>
<td>Turbulence intensity</td>
<td>$+$</td>
</tr>
<tr>
<td>Cloud</td>
<td>$+$</td>
</tr>
<tr>
<td>Cloud</td>
<td>$+$</td>
</tr>
</tbody>
</table>

(Continued opposite)
### Table 3 (Continued)

<table>
<thead>
<tr>
<th>Change</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>−</td>
<td>Decrease in fog occurrence, especially dense fogs in warmer urban atmosphere</td>
</tr>
<tr>
<td>+</td>
<td>Increase in fog, especially light fogs, when aerosol pollution is high</td>
</tr>
<tr>
<td>−</td>
<td>Reduced due to pollutants and haze droplets</td>
</tr>
<tr>
<td>+</td>
<td>Increase over and downwind with summer convective weather</td>
</tr>
<tr>
<td>−</td>
<td>10–30% due to warmer urban atmosphere</td>
</tr>
<tr>
<td>+</td>
<td>Localized snowfall due to water vapor and ice nuclei additions from major point sources, possible increases in polar cities</td>
</tr>
<tr>
<td>+</td>
<td>Particularly downwind; hail fall frequency $\approx +30%$</td>
</tr>
<tr>
<td>+</td>
<td>More cloud-ground flash densities downwind; decreased % of positive cloud-ground flashes</td>
</tr>
<tr>
<td>+</td>
<td>Increase related to use of energy and sensitive to the particular energy type used. Fine particulate matter and ground-level ozone are of particular importance</td>
</tr>
</tbody>
</table>

© 1997 Routledge, reproduced by permission of Taylor & Francis Books, UK.

---

**Figure 5** Cross-section of the urban heat island under calm, clear conditions for surface and canopy layer temperatures by (a) day and (b) night. Temporal development of the canopy layer heat island showing urban and rural temperatures (c) and the resultant $UHI_{UCL}$ (d). Source: adapted from Voogt (2002) and Oke (1982).
Table 4  Suggested causes of the canopy layer urban heat island. Adapted from Oke (1982).

<table>
<thead>
<tr>
<th>Energy balance term</th>
<th>Urban features</th>
<th>Urban effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased solar radiation absorbed</td>
<td>Canyon geometry</td>
<td>Increased surface area and multiple reflection</td>
</tr>
<tr>
<td>Increased longwave radiation from atmosphere</td>
<td>Air pollution</td>
<td>Greater absorption and re-emission</td>
</tr>
<tr>
<td>Reduction in longwave radiation loss from surface</td>
<td>Canyon geometry</td>
<td>Reduced sky view factor</td>
</tr>
<tr>
<td>Heat added by human activity</td>
<td>Buildings and traffic</td>
<td>Direct addition of heat</td>
</tr>
<tr>
<td>Heat stored in urban fabric</td>
<td>Construction materials</td>
<td>Increased thermal admittance</td>
</tr>
<tr>
<td>Reduced evaporation from surface</td>
<td>Construction materials</td>
<td>Increased waterproofing</td>
</tr>
<tr>
<td>Reduced turbulent heat loss</td>
<td>Canyon geometry</td>
<td>Reduced wind speed</td>
</tr>
</tbody>
</table>

Humidity

Spatial patterns of humidity in urban areas are complex owing to the patchiness of urban vegetation, variability in irrigation practices, and large point sources of anthropogenic vapor emissions from industrial or power generation facilities. At night in the UCL, cities are often characterized by an absolute increase in moisture content. Urban relative humidity on the other hand may decrease because the increase in the saturation vapor pressure with temperature exceeds the increase of the actual moisture content. By day, the humidity influence is seasonally dependent in temperate climate cities – urban areas tend to be somewhat less humid than their surroundings in the summer, but more humid in winter. The summer daytime decrease is related to the loss of vegetated surfaces in cities and the reduced evaporation combined with greater mixing; the patchiness of vegetation influences the spatial patterns of humidity along with the extent to which irrigation is used. At night, rural cooling under conditions that favor UHI formation leads to dewfall that depletes the moisture content of the near surface atmosphere. In cities, the warmer urban atmosphere reduces dewfall and
can support some nighttime evaporation. Combustion processes and heat exchangers also add humidity via the latent heat component of the anthropogenic heat flux. In winter, cities are commonly more moist by day as well as at night. In high latitude cold climate cities, the addition of anthropogenic moisture can lead to ice fogs that seriously impact visibility.

The UBL is similar to the UCL in its humidity characteristics. The geographical setting of a city – whether it is near a coast or embedded within a desert – can affect the urban–rural humidity differences. Cities in climate regions outside the temperate zone tend to show similar patterns with the exception of cities in desert climates that may be more moist both day and night as urban sources of moisture contribute to higher humidity.

Wind

Urban areas affect winds at scales that range from the patterns of flow around an individual building up to regional-scale impacts on wind speed, direction, and turbulence within the UBL above and downwind of a city. Urban modified winds are of particular importance to understanding dispersion and transport of pollutants in urban areas. In general, the urban surface is “rougher” compared to its surroundings. This influences the speed, direction, and turbulent energy of the wind at all heights within the urban atmosphere. Air flow in and over cities is also impacted by spatial patterns of temperatures (i.e., the UHI), which induce pressure differences and can create thermally forced winds.

Within the UCL, wind flow is affected by the spacing and dimensions of buildings. Buildings force air flow to divert around and over them and create a downwind “wake” that includes a reverse circulation vortex in the immediate lee of the building (Figure 6a(i)). As the building spacing decreases, the wakes of individual buildings begin to interact and the surface is said to be “rougher” so that it slows the average wind more. As spacing decreases further, wakes merge and eventually the flow becomes almost decoupled from that above the buildings. Very tall buildings can bring higher

Figure 6  Urban flow regimes (a) associated with increasing density of development and the profile of mean wind speed $\bar{u}$ above an urban area (b) showing the extrapolation (dashed line) to the roughness length ($z_0$). The profile is displaced upwards to a height $z_d$ because of the tall roughness elements. Source: modified after Oke (1987).
velocity air downwards to street level with implications for pedestrian safety.

When the buildings are closely spaced and form a relatively narrow canyon a distinct vortex flow forms within the canyon when the wind is at right angles to the street orientation (e.g., Figure 6a(iii)). When the relative orientation of the wind shifts to intermediate angles, a corkscrew-type flow pattern results within the canyon and for winds approximately parallel to the canyon, the flow will be channeled along the street. In particularly narrow street canyons multiple-stacked vortices may be present. The structure and configuration of the vortices can be affected by temperatures of the canyon surfaces as warmer buoyant air tries to rise along heated walls. At night, cold dense air can form on roof tops and drain into canyons leading to mixing that helps to flush air from the canyon. Vehicle traffic within the UCL can also generate turbulence. Flows of denser cooler air out of urban parks and forested areas into surrounding neighborhoods can also occur.

Above the UCL, within the ISL the wind speed shows a characteristic logarithmic profile that is displaced upwards at a distance \( z_d \) by the urban canopy. The roughness length \( z_0 \) is a measure of the overall aerodynamic roughness of the urban surface within a neighborhood. It varies (see Table 2) with the height and spacing of the buildings and the presence of other surface elements such as trees.

Within the UBL, the air flow is more turbulent and slower relative to the rural PBL when regional winds are strong. The slowing air above the rougher city induces convergence of air which leads to uplift and contributes to the doming of the UBL top (Figure 1). The additional friction over a large city produces a turning effect so that winds turn cyclonically within the UBL. This can create additional zones of uplift (sinking) associated with air that convergences (diverges) horizontally. Temperature differences associated with the UHI can drive a thermally induced circulation when regional winds are calm or very light. Air will flow from cooler rural areas into the city, rise, and then return to the rural area aloft and sinks. The strength of this circulation is dependent on the temperature differences and stability; more unstable conditions facilitate the vertical motions associated with the circulation.

**Clouds and precipitation**

Cities have long been suspected of modifying the cloud and precipitation over and downwind of them because they are known to affect upward motion, the presence and properties of aerosols that act as precipitation nuclei, and because individual large sources of water vapor, heat, and pollutant emissions (e.g., power plants and industrial sources) have been associated with cloud formation. However, much of the cloud and precipitation experienced by a city is determined by larger-scale synoptic processes and observing networks are often too sparse for assessing patterns at the urban scale. Thus urban effects on cloud and precipitation are often subtle and difficult to detect.

In the mid-latitudes, there is good evidence that large cities show some increase in cloud cover and precipitation during summer season convective weather events when clouds are more strongly coupled to surface and boundary layer processes. The increase in convective clouds is coupled with an increased incidence in thunderstorms and hail over and downwind of cities, especially when regional winds are light. These weather events, in combination with the modified urban water budget that increases runoff due to impervious surface cover, can be important to the incidence of flooding events in cities. Lightning detection networks, radar, and satellite-
based observations have all contributed to the identification of spatio-temporal patterns of urban precipitation effects (e.g., Shepherd 2013).

The physical basis for the increased frequency of summertime convective precipitation involves multiple factors but a key element is enhanced uplift over and downwind of cities associated with convergence of UBL winds driven by the urban heat island. Surface roughness influences may also play a role; tracking of individual storms cells suggests some cells may divert or split to move around cities and remerge on the downwind side affecting the location and intensity of precipitation patterns. The addition of urban aerosols may also affect the precipitation process. In some circumstances, cloud growth may be invigorated, updrafts are stronger, cloud electrification is enhanced, and precipitation is increased downwind.

In the winter season, cold climate cities can experience a reduction in the frequency of freezing rain and snow events because of urban warming in the boundary layer converting the phase and potentially the amount of precipitation received (through sublimation and evaporation of precipitation in the warmer UBL). Under very cold and humid conditions, individual point sources of water vapor and aerosols can lead to clouds and possibly even snowfall (Koenig 1981) or to enhance naturally existing clouds and their precipitation.

In all climates, the geographical setting of the city influences the spatial patterns of precipitation and must be considered because topography may influence patterns of precipitation. The setting also determines the background aerosol characteristics, the level of which affects the impacts of urban additions of aerosols.

Air pollution

The urban atmosphere is generally more polluted. Urban air pollutants include those directly emitted into the atmosphere (primary pollutants), usually as gases or particulates, and those that form from reactions in the atmosphere (secondary pollutants), such as ground-level ozone that forms in photochemical smog. Air pollution spans a wide range of scales. Indoor air pollutants come from combustion processes, cooking, smoking, and emissions from the ground and building materials. Their concentrations are controlled by the strength of the emission, removal rate, and the ventilation of the interior space. In the UCL, traffic is an important source of emissions, along with emissions from buildings, and strong spatial variations in pollutants are noted, especially with respect to distance from roadways. Leakage from pipes and storage sites also contribute as may accidental or intentional releases of hazardous materials. In the UBL pollutants are mixed upward from below, injected from tall stacks, and can form in situ from secondary reactions. Biogenic emissions of volatile organic carbons come from urban vegetation. Long-lived pollutants and those that form from secondary reactions extend urban climate changes to larger scales because some pollutants are transported or form well beyond the city. Of particular note is the enrichment of urban atmospheres with CO₂. The use of energy in cities drives large emissions of CO₂ from combustion that typically exceeds the uptake by photosynthesis in urban vegetation. The resulting increased concentration of CO₂ in the urban atmosphere is not sufficient to provide significant local warming but the long residence time of this pollutant means that cities are providing important contributions to the global increase in atmospheric CO₂ concentrations.

Air pollution is important to the health of urban residents and also modifies other aspects of urban climate. Air pollutants, especially particulate matter, reduce the incident solar radiation, increase the fraction of scattered
diffuse) radiation, and reduce the ultraviolet wavelengths. The absorption of shortwave and absorption and emission of longwave radiation by pollutants alter the temperature profile of the boundary layer and affects stability of the atmosphere. The addition of particulate matter acts as nuclei for droplet formation and affects visibility, cloud, and precipitation formation and can alter cloud electrification processes.

**Urban climate and global climate change**

Cities both affect and are affected by global climate change. Cities occupy only 2% of the global land surface yet have impacts that extend to the global scale. This is because urban activities are responsible for a large fraction of the greenhouse gas (GHG) emissions that drive global anthropogenic climate change. The warmth of the UHI, however, does not significantly affect global climate. Increased temperatures from global climate change, superimposed with the UHI, will increase the risk of hazardous thermal conditions in summer heat wave situations and may also impact air pollution concentrations. The increased likelihood of intense rainfall events with global climate change will differentially impact cities, since they significantly increase the amount of impervious surfaces that lead to large volumes of runoff, and we must also consider urban enhancements of precipitation from large cities.

Cities will be required to adapt to climate change. Changes to city location, layout, neighborhood design, and individual building technology may all contribute. For example, green and cool roof technologies can both help buildings adapt and to mitigate GHG emissions associated with building energy demand. The design of neighborhoods and cities, including their density and configuration, can use urban climate knowledge to provide solutions that minimize the impacts of climate change on the city and reduce the demand for energy that drives GHG emissions. Understanding urban climates is therefore a growing need as the majority of the world’s population now resides in urban areas.

**SEE ALSO:** Built environments; Climate adaptation/mitigation; Microclimatology; Sustainable cities; Temperature; Urban ecology; Water budget

**References**


URBAN CLIMATOLOGY


Further reading


Urban ecology

Ian Douglas
University of Manchester, UK

The intellectual and social context

Urban ecology is the study of urban areas as a series of habitats for human beings and other organisms; the relationships between all these organisms under changing urban environmental conditions; and the interactions of living things with natural and human-made flows of energy, water, and materials, both deliberate and accidental, in the varied and diverse conditions of the world’s towns and cities. It integrates both theoretical and applied aspects of the natural and social sciences in analyzing, understanding, and managing the diverse ecological conditions encountered in urban areas.

Urban ecology is firmly rooted in ecology, the science studying the processes influencing the distribution and abundance of organisms, the interactions between organisms, and the interactions between organisms and the transformation and flux of energy, matter, and information. It embraces all the sociopolitical, cultural, and economic factors that affect ecological processes in urban areas, paying particular attention to the way those factors continually modify the abundance and diversity of organisms found in particular urban places. Urban ecology also examines specific ecological situations within towns and cities, as well as how urban demands for food, water, energy, and materials and disposal of wastes affect other ecosystems at all scales from entire oceans to individual organisms.

People from many academic and technical backgrounds have written about nature in cities, urban ecosystems, the flora and fauna of towns, the species colonizing newly vacant urban land, and the human benefits from contact with wildlife in built-up areas. Broadly, four general themes emerge from these diverse points of view:

- The ecology of health in urban areas: towns and cities as habitats for humans and disease vectors, including the role of urban nature in human wellbeing.
- Ecology in towns and cities: urban areas as habitats for wildlife, plants, and other organisms, from coyotes in Californian cities to crocodiles in Darwin, Australia, and monkeys in Malaysian towns, and including molds on food and kestrels on skyscrapers.
- The ecology of cities as a whole: the flows of energy, water, materials, and information into, within, and out of urban areas and their internal and external impacts; differentiation within urban areas; rural–urban gradients; and the analysis of complex urban socioecological systems.
- Ecology for cities: enhancing the ecosystem services provided by urban green infrastructure to help build resilience to change, including environmental change, and to increase urban sustainability and improve community cohesion.

The public health theme looks at urban areas as habitats that may improve or threaten human health and wellbeing. Good urban design can reduce the urban heat island effect that leads to...
excess deaths during extreme hot spells. Urban street trees and gardens may alleviate stress and anxiety for some people, and create fear among others. Conversely, poor housing and sanitation can support disease vectors that carry malaria and dengue fever as well as lead to the spread of cholera and viral gastroenteritis. Climate change is allowing disease vectors to extend their range, especially with large urban areas often being warmer than their surrounding countryside. Many zoonoses (diseases transmitted from animals to humans under natural conditions) come into urban areas with pets, farm animals, and imported meat. Thus analysis of ecological conditions reveals urban habitats for disease-bearing organisms, helps establish where risks to health are increasing, and heightens awareness of how social factors, such as poor housing and sanitation, lead to conditions that favor human ill-health.

Urban ecosystems continue to be a major focus for an immense range of biological work on particular urban influences on plants, animals, fungi, and other organisms. Local natural history societies developed in the mid-nineteenth century, many of them concentrating on plants and animals in and immediately around urban areas, for example the London Natural History Society. From such beginnings, ornithologists, for example, have built up great databases on changes in urban bird populations that enable their behavior and ecology in towns and cities to be analyzed in detail. Similarly many invasive plant and insect species are well documented, particularly in cities where large areas of vacant land were created by wartime bombing or industrial dereliction. Such studies often focus on particular species or types of species, but many also examine interactions with the built environment, pollutants, and human behavior. This is the classic urban ecology of Europe: the study of the interactions of various organisms with each other and with their diverse urban surroundings. Similar work is found in many other parts of the world, especially in China (Li et al. 2008) and India (Nagendra and Gopal 2011; Khera, Mehta, and Sabata 2009), and particular attention has also been given to the social aspects of urban biodiversity in developing countries (Cilliers 2010).

The view of urban areas as single systems integrated with their immediate surroundings (the city-region) sees the city-region ecosystem as a life-support system for all the organisms within the urban settlement. It provides ecosystem services such as water supplies, sources of aggregates, areas for landfill, recreation zones, watershed protection, greenhouse gas uptake, biodiversity, and some food supplies. This point of view is often discussed in terms of the ecology of cities or as urban metabolism (a somewhat criticized organic analogy).

The same urban settlement can be seen as creating demands for energy, food, water, and materials that may have widespread, even global, implications. Accessing supplies of coal, natural gas, or oil, of grain, fruits, vegetables, fish, meat, and other foods, and of manufactured goods from distant regions has a global outreach with impacts on other ecosystems, from fine particles of plastic frozen in Arctic ice to residues of agricultural chemicals and mining wastes contributing to the eutrophication of rivers, lakes, and reservoirs in distant countries. This global impact, or ecological footprint, is one expression of the widespread effects that these complex socioecological systems have around the world.

**History of urban ecology**

Urban ecology has diverse roots. One stems from nineteenth-century public health investigations of the relationships between human living conditions, the spread of disease, and the wide variety of disease vectors in urban areas leading to a
vision of cities as socioecological systems that still resonates strongly with twenty-first-century views of urban ecologists about people and nature in urban areas.

A second strand of urban ecology stems from landscape design and urban planning. Changed ideas about housing and sanitation in the nineteenth century included the belief that exercise in the open air improved people's health. Thus, the goals of the first public parks in England, Paxton's Princes Park in Liverpool and his Birkenhead Park across the Mersey Estuary, built in the 1840s, were to provide open space for the benefit of townspeople and local residents in areas being rapidly built up, and to develop attractive designed landscapes as a setting for the suburban residences. Frederick Olmsted took on board some ideas from a visit to Birkenhead in his 1857 plans for Central Park in New York. The many important ecological attributes in such urban landscape formation were incorporated into Olmsted's sophisticated design combining wastewater management and recreational amenity in the Riverway and Back Bay Fens in Boston, Massachusetts, a forerunner of today's sustainable drainage.

Eventually this stimulated ideas of planning new towns as garden cities, beginning with Ebenezer Howard's garden cities at Welwyn and Letchworth in England at the end of the nineteenth century. The subsequent garden city movement influenced the design of suburbs, new towns, and federal capital cities. It had no particular emphasis on planning for wildlife or children's access to urban nature, but it did have a city-region approach emphasizing a green belt of intensive food production around each garden city. The emphasis was on healthier living spaces for people, workplaces close to homes, and easy access to high-quality community services. Even so, garden city designs, with their diversity of urban planting schemes, gardens, and horticulture, created mosaics of habitats for other organisms, producing novel combinations of species in emerging ecosystems. Workers' housing with individual gardens, for example the 1930s council estates in the United Kingdom, arose from the garden city movement. Crises such as World War II saw parks, sports grounds, and golf courses used for growing a wide range of crops in patches of land ranging from domestic vegetable plots to large wheat fields. This intensive agriculture helped to increase urban biodiversity and plant productivity. After 1945, most of the open spaces were returned to their recreational uses, but the urban food-growing tradition remained strong in Europe, with many residents choosing to cultivate family gardens or allotments. The value of urban nature for human physical and mental health remains a central concern of government and civil society organizations involved with urban green spaces and urban wildlife today.

Urban agriculture is even more significant in poorer countries, where urban growers supply as much as one-third of all the food consumed in some cities. Here food growing is necessary for survival, often using public lands and illegally occupied vacant areas, and forming a key component of the urban ecology of most of the world's fastest-growing cities. The value of urban nature for human physical and mental health remains a central concern of government and civil society organizations involved with urban green spaces and urban wildlife today.

By the mid-twentieth century, surveys of introduced and invasive species began to appear in Europe, and specialized work on synanthropic plants and animals in urban areas began to form a third foundation of urban ecology. The rapid ecological changes on bombed sites after 1940 added a further stimulus to modern urban ecology. Derelict bombed sites were invaded by unusual plants and insects, creating new habitats
for birds and leading to novel combinations of flora and fauna. Major studies in many European cities examined the results of this tremendous natural experiment and stimulated notions of novel ecosystems, some of which were considered worth preserving so that city dwellers could see natural vegetation close to their homes. However, the work was essentially on the rubble left after bombing and on derelict industrial land. It concentrated on the newly naturalized plants, the dispersal strategies of particular species, succession under various site conditions, and the formation of new plant communities. It produced the first peak of activity in urban ecology and had a major impact on attitudes to nature in urban areas. Such was the biodiversity on some of these sites that many local groups campaigned to have some remnant areas protected as ecology parks or urban nature reserves. The demands for regeneration and rebuilding after 1945, and the emphasis on creating national parks and protecting the countryside, drew attention away from urban wildlife. However, industrial change, following the replacement of coal by natural gas and the decline of heavy manufacturing industry, left much derelict land that was colonized by unusual combinations of plants. Old factory sites and coal mine waste tips were converted in the public open spaces, and the reclamation of derelict sites created new opportunities for the deliberate planning of new ecosystems. Restoration ecology developed. People began to see the opportunities for allowing natural recolonization to proceed at some sites. By the 1970s urban ecological projects were underway in several European cities and urban ecology parks began to be established. A book on Nature in Cities by Ian Laurie was published in 1979 and the first European symposium on urban ecology was held in Berlin in 1982. The journal Urban Ecology began publication in 1975, and was incorporated into Landscape and Urban Planning in 1986. Now there are at least eight scientific journals concentrating on urban ecology.

Somewhat detached from European developments, studies of urban forests and wetlands in North America developed after the 1969 US Environmental Protection Act stimulated work in urban forestry that has played a major role in establishing methods of evaluating urban ecosystem services and encouraging investment urban greenways and green infrastructure. The 2009 report on Planning the Urban Forest: Ecology, Economy, and Community Development (Schwab 2009) set out the following benefits of the urban forest:

- **environmental**: providing green infrastructure; treating stormwater runoff; shading and cooling the urban heat island; reducing air pollution; providing wildlife habitat;
- **social**: health benefits; environmental justice;
- **economic**: value of trees in business district; influence on property prices; added value from parks.

These benefits interest urban ecologists and indicate how people have come to examine interrelated urban environmental issues from many different disciplines. They also demonstrate how the concept of green infrastructure has become prominent since 2001. So the examination of particular species and special habitats in cities has now grown into a much broader consideration of the place of nature in the city and its benefits to urban inhabitants.

The fourth root of urban ecology extends into the view of the city as a whole ecosystem with inputs and outputs and with its own metabolism. Ideas on cities as dependent ecosystems expanded from William Cobbett’s 1829 description of the impact of London’s food and timber requirements on life in villages 100 km away, to Wolman’s landmark 1965 paper...
on the metabolism of cities. Their ecological significance was first emphasized by the 1972 publication on Sydney as an ecosystem. A new standard was set by Stephen Boydén’s team’s work on Hong Kong in the 1970s. Their book provided an intellectual framework for the research; analyzed empirical data from a wide range of sources to understand the dynamics of the city; explained both ecological and social problems affecting the dynamics of the urban area; and set out the first model for integrating social and ecological process into an urban socioecological system. It became a pilot project of UNESCO’s Man and the Biosphere Programme Project 11 (MAB 11), in which ideas of analyzing urban energy, water and materials budgets, and their ecological impacts eventually expanded into both the industrial ecology and the urban ecology of the early twenty-first century. MAB 11 helped to develop views on the ecological aspects of urban systems, particularly in Rome where the psychological dimensions of urban relationships with nature received much attention. The dependency of urban areas on external ecosystems became well understood, particularly through the readily grasped, but sometimes misinterpreted, concept of the urban ecological footprint. Probably the most significant element of the urban footprint approach is the comparative per capita area of land needed to support an inhabitant of a particular city at its present rate of consumption. Around the year 2000, it ranged from 0.8 ha per person in an informal settlement in Delhi, India, to 11.0 ha per person in Calgary, Canada.

The urban ecological footprint is a modern example of the organic analogy or allegorical use of ecology. An earlier instance arose at the University of Chicago in the first decades of the twentieth century, where pioneering ecologists developed ideas of evolving plant communities and plant succession in the dune systems at the southern end of Lake Michigan. These biological ideas were noted by Chicago sociologists who saw possibly similar changes in human communities and the successions of occupation of various parts of the city by different social groups. This human ecology of urban areas, known as the Chicago School, was termed “urban ecology” by the 1950s. Urban ecologists then studied the spatial organization of the urban community and the changing patterns of segregation and occupation of different districts by various social and cultural groups. With the advent of computers to analyze large quantities of census tract data, urban ecology within sociology developed into social area analyses and factorial ecologies, which engaged many geographers and sociologists in the 1950s and 1960s. While this work continues in public health investigations, others subsequently concentrated on how the socioecological character of their nonwork relationships and interactions influence urban people’s lives.

Since 1990, urban ecology has moved into an era of high-level cross-disciplinary comparative and analytical work, building on MAB 11 and given a major boost by the establishment in 1997 of two US research sites, the Central Arizona–Phoenix Long-Term Ecological Research (CAP LTER) project and the Baltimore Ecosystem Study (BES). The BES aims to understand metropolitan Baltimore as an ecological system, bringing together researchers from the biological, physical, and social sciences to collect new data and to synthesize existing information on how both the ecological and the engineered systems of Baltimore work. CAP LTER is a multidisciplinary, urban ecological investigation of the socioecological systems in central Arizona. The basic questions guiding the research are: How do the services provided by evolving urban ecosystems affect human outcomes and behavior, and how does human
URBAN ECOLOGY

action (responses) alter the patterns of ecosystem structure and function, and ultimately urban sustainability, in a dynamic environment?

Modern urban ecology

Today, a broad cross-disciplinary endeavor is helping to change the appearance of urban areas by improving understanding of urban ecosystem functioning and assisting the design and management of urban areas for human wellbeing, biodiversity, and other ecosystem services. For urban ecology such work is a challenge, requiring the coupling of science with practice to address many of the wider issues faced by humanity.

Although modern urban ecology emphasizes the integrated view of complex socioecological systems, scientific publications tend to deal with parts of systems and it is convenient to examine the contemporary scene in terms of the four themes set out in the introduction: the ecology of health in urban areas, ecology in cities, ecology of cities, and ecology for cities.

Ecology of health in urban areas

The relationships between people and nature in cities are many and varied. Humans have profound effects on species survival, population structure, reproduction, behavior, and evolution; and, at the same time, nature also has profound effects on the health and wellbeing (both positive and negative) of people. As urban areas increase in number and size, so the frequency and intensity of interactions between humans and urban wildlife also increases. These interactions can be with individual animals or plants, or with the habitats in which these species live (such as parks, gardens, and ponds).

The diseases carried by animals (disease vectors) and the diseases occurring in animals that can be passed on to humans are some of the main hazards associated with animals. Attention to zoonoses has greatly increased with the rapid expansion of intercontinental air travel and the movement of animals around the world. Disease can be transmitted rapidly. Climate change is also permitting disease vectors to move into areas that were previously climatically inhospitable. Increasing urban populations, pet ownership, travel, and activities that bring people closer to wildlife are raising the possibility of the rapid transmission of wildlife diseases to, and within, urban environments. Particularly severe in their consequences are exchanges between people and nonhuman primates such as herpes simian B virus and monkey pox (through animal bites) and Ebola (through hunting and butchering). More significant in terms of numbers of people, especially in urban areas, are the diseases carried by nonprimate animals that are close to humans, including dogs and rats. Alligators and large snakes, birds, mosquitoes, and ticks all carry diseases or pose other health risks. Rats (Rattus sp.), for example, cause several possibly fatal diseases including Weil’s disease, salmonella, tuberculosis, cryptosporidiosis, E. coli, and foot and mouth disease (Aphthae epizooticae). Urban birds exploit the built environment extensively. However, they also spread disease by spores from their droppings that are carried on the wind and inhaled as dust. Symptoms of the diseases involved, such as chiamdiosis, a virus similar to influenza, and psittacosis, which is similar to pneumonia, may be minor in healthy adults but are a more serious problem for those with low immunity.

Many insects are disease vectors. The role of mosquitoes in malaria transmission is well known, but they can also carry other diseases, such as dengue fever. Ticks carry many diseases of which Lyme disease and tick-borne encephalitis are among the best known. Some diseases are carried by more than one species of tick, and some ticks transmit more than one disease. Venomous
Spiders are a hazard to many people and insect vectors carry diseases, including malaria, that affect people and other animals, whether those animals are pets, domesticated, or wild species.

Some plants cause more direct harm to people. Plant sap sometimes contains toxic chemicals known as furanocoumarins, which in the presence of sunlight cause phytophotodermatitis: a reddening of the skin, often followed by severe burns and blistering. The burns can last for several months, but even the skin can remain sensitive to light for many years. Toxic plants and fungi may enter the human food chain by being mistaken for edible varieties.

Much emphasis has been placed on the positive aspects and health benefits of urban greenspace. The health effects of nature are usually discussed in terms of stress reduction and attention restoration theories. Even passive viewing of natural environments after psychophysiological stress helps to reduce stress and may later confer health benefits. Contact with nature, or natural views, can confer psychological wellbeing benefits through attention restoration. Thus stress recovery and attention restoration may be central to understanding other psychological wellbeing benefits gained from contact with vegetated urban spaces.

As people have become more aware of the dangers of obesity, heart disease, and high cholesterol levels in urban societies, the lists of the positive health and wellbeing aspect of activity and exercise, of the gentlest kind, in blue- and greenspaces in towns and cities have grown markedly. These benefits are now being expressed in the terms of hypotheses that urban ecology research should be able to test, confirm, or refute.

The way vegetation is arranged and managed in parks and neighborhoods considerably affects how people engage in health-promoting physical activities. Theories relating to visual aesthetics can be linked to visual concepts that are important in landscape design and landscape architecture as they shape the landscapes that are produced. Often the most important issue is that of the perceived safety of the place used for the physical activity. High levels of vegetation do not necessarily equate with more physical activity. The pattern of this vegetation is critical; if vegetation is planted and allowed to grow in ways that limit the ability to visually and physically access one’s immediate surroundings, then fears of being in a potentially unsafe environment may override any motivation to use the space for physical activities.

Urban agriculture provides multiple ecosystems services, helping to improve health by providing better nutrition and by the exercise involved in cultivating the crops. In more developed countries urban household gardens supplement dietary intake and reduce overall household expenditures on food. In poorer cities, urban agriculture provides food that people cannot afford to buy.

Recreational gardening helps people to relax and releases stress. It is widely used in prisons and hospitals to help inmates and patients recover. General public health is improved by involvement in converting run-down, unkempt, garbage-laden urban neighborhoods into vegetable gardens. Such beneficial interactions of ecological and social change are now at the heart of urban ecology. When low-income neighborhoods and market gardeners become involved in transforming their urban landscapes and claiming for themselves a sense of pride and place, urban agriculture becomes a forceful empowerment strategy for community participation and social change.

Poor housing conditions create urban habitats that affect public health. They are associated with a wide range of health conditions, including respiratory infections, asthma, lead poisoning,
injuries, and mental health. Specific relationship- 
ships are well understood; for example, hospital 
admissions for chronic obstructive pulmonary 
disease (COPD) are often higher from crowded 
older buildings in the heart of a city where urban 
heat island intensity, and vehicular traffic and air 
pollution levels, are also higher than elsewhere.

Ecology in cities

Modern urban ecology drew its inspiration 
from detailed studies of urban flora and fauna, 
particularly in Berlin where Herbert Sukopp 
established the Institute of Ecology, which 
concentrated largely on botany and vegetation, 
with a strong emphasis on the phytosociology 
of urban areas (Sukopp 2008); and in Warsaw 
where Maciej Luniak (2004) and colleagues 
concentrated mainly on animal ecology. Com-
prehensive biotope mapping in Berlin provided 
the ecological basis for a land-use plan that 
incorporated nature conservation. No other 
city has been subject to such close ecological 
investigation, which has laid firm foundations 
for studying plant ecology in urban areas. In 
Warsaw, since about 1970, at least 12 new bird 
species and two mammal species have colonized 
densely built-up areas. In Europe as a whole, 
the most well-known species that have been 
more successful in urban areas than in their 
native natural habitats are the blackbird, magpie, 
hooded crow, kestrel, grey squirrel, striped field 
mouse, rabbit, and red fox. This process of synurbanization shows that there are chances for 
some kind of coexistence between nature and 
the expansion of urban civilization. Nevertheless 
most cities show some loss in species diversity 
as many birds and mammals disappear from the 
new suburban areas where they once lived.

In terms of flora, central European urban areas 
are rich in alien plants, which account for a mean 
of 40% of all species, their representation being 
13.7% higher than their representation in the 
total species pool available in the region (Pyšek 
1998). Compared to natural habitats, the differ-
ence is greater, a group of Czech nature reserves 
having an average of 9.8% aliens. The cities are 
thus the habitat that is richest in alien species, 
and serve as an important source of aliens (often 
invasive) for smaller settlements and further 
spread into rural areas.

The changes in urban flora depend on the 
pattern of succession on particular substrates, 
emphasizing the importance of natural and 
artificial soils (often termed “made ground”) 
in the character of naturally colonized urban 
vegetation. For example, in the Czech Republic, 
after 12 years of succession, native woody species 
dominated ruderal (landfill) sites, with Populus 
tremula on nutrient poor sites and Sambucus nigra 
and Salix caprea on moderately rich nutrient 
sites. Similar trends over five years were noted 
in a Berlin experiment, with Betula pendula and 
Populus nigra succeeding on sandy soils, and 
annuals giving way to biennials and perennials 
on topsoils. Work of this type generally leads 
to two hypotheses: the urban environment and 
land transformation create unique environmental 
conditions to which species have to adapt to 
survive and mature, and the loss of native species 
and increase in non-native species, with increas-
ing urbanization, leads to biotic homogenization 
with the same urban-adaptable species being 
found in widely separated cities.

The hypothesis of unique environmental 
conditions needs testing by paying attention to 
urban geomorphology and urban soils, looking 
at the ground on which the city is built and in 
particular paying attention to the deposits and 
surface materials created by past urban land uses. 
In this way urban ecology has links to geological 
concerns for the Anthropocene and geomor-
phological concerns about artificial landforms 
and sediments. Above all, the species occupying 
particular sites reflect the local biogeochemistry.
Gardens are a particular type of Anthropocene environmental change. They reflect cultural aspirations, human needs for food and relaxation, ambitions to acquire the novel and exotic, and desires to express personality and wealth. Sometimes they are made up entirely of materials, including soils and rock, brought onto the site and are planted with exotic non-native species. At others they are but the slightest modification of what would have been there naturally. Detailed work from 1972–1986 in a garden in Leicester, England, noted 2204 species, including 422 species of plants, 1602 insects, 121 other invertebrates, and 59 vertebrates. Further work on the mosaic of habitats and landscape in suburban gardens has found 1166 vascular plant species, of which 30% were natives, together with some 80 lichen species and 68 bryophyte species across 61 gardens in the United Kingdom. Generally along rural–urban gradients, the species richness of some major taxa (particularly plants and birds) commonly exhibits a peak in suburban areas where there is greater habitat heterogeneity, much of it contributed by suburban gardens.

The manipulation of habitats in urban areas provides opportunities to experiment in designing with and for nature. These planned interventions include restoration ecology and creative conservation, and lead to recombinant ecology. Restoration ecology seeks to provide a scientific basis to inform the practitioners working to restore areas affected by overexploitation, farming, industry, or natural catastrophe. The conservation of existing habitat and the restoration of degraded, damaged, or destroyed habitat, through restoration ecology, provide ways to address the trend for habitat loss, species extinction, and the decline of ecosystem services. Practitioners of ecological restoration employ a wide range of techniques including removal of non-native species and weeds, revegetation of disturbed areas, reintroduction of native species using stock with a known local provenance (to ensure, as far as possible, that the genetic diversity of the introduced plants and animals is as close to the naturally occurring population as possible), reforestation, erosion control, daylighting streams, as well as habitat and range improvement for targeted species. Restoration ecology uses ecological theories that guide landscape managers and others to put degraded and disused land into a better ecological condition which will provide habitats for plants and wildlife. Restoration ecology draws on a wide range of ecological concepts, including those of disturbance, succession, fragmentation, ecosystem function, adaptive capacity, and resilience (Pickett, Cadenasso, and McGrath 2013).

Restoration can aim simply to restore and enhance biodiversity, or it can be seen as integral to urban regeneration and environmental improvement, for example in assisting towns and cities to adapt to climate change. Specific examples of the benefits of ecological restoration include the development of wetlands to control water distribution and quality, and of wildflower meadows to enhance the attractiveness of former mown urban grassland, and the establishment of urban forests to provide both wildlife habitat and the potential for wood production by coppicing. Ecological restoration is of great value in areas of changing land use, such as abandoned landfills and mine waste tips, contaminated land, and former industrial sites.

Creative conservation is the introduction of species to new situations and the reintroduction of species to habitats where they have become extinct. In many cases it is simply conserving nature in an adaptive fashion so that ecosystems can respond to changing environmental conditions. In others, it is the reworking of land to introduce either animals that have been bred in captivity for reintroduction to the wild, or plants, particularly wildflowers, that have been
URBAN ECOLOGY

prepared in nurseries or experiment stations to yield seeds or seedlings to be sown in specially prepared areas.

In the United Kingdom creative conservation establishes novel ecosystems that may be totally different from what existed before urbanization or may reintroduce native species onto previously used urban sites (brownfield land). It differs from restoration ecology in not directly attempting to put areas back to some ideal previous condition, but to make an area more attractive to both the local human population and wildlife, such as planting wildflowers meadows around the apartment blocks and row houses of social housing estates. It redesigns the urban landscape for both ecological and social goals.

Ecology in cities thus involves classic botany, zoology, and habitat analysis, but looks in particular at the diversity of ecological situations in urban areas and at ways of improving and managing those habitats for the benefit of wildlife and people. In creating novel ecosystems and deciding what to do about those that have developed naturally, urban managers have to balance people’s landscape tastes and preferences against municipal and private budgets for greenspace care and management. In so doing they are designing future urban geographies under constraints that are often external to their own environment and beyond their control.

Ecology of cities

Studies of the ecology of cities are interdisciplinary and multiscale, incorporating both the ecological and the human dimensions of urban ecosystems. They comprise three broad categories of thinking:

1. ecology of the physical urban system, including the wide range of people-impacted or modified environmental patterns and physical flows, affecting organisms and their habitats, in and through the city, and through urban systems on a regional, national, and global scale;
2. ecology of the social-economic urban system, involving the range of human activities that are in various ways associated with ecosystem and environmental processes, including industrial ecology, ecological design, and ecosystem service markets;
3. ecology throughout the human–environment system, embracing social, economic, cultural, and political patterns and relationships in the human ecology sense of a community of interdependent processes in both human systems and ecosystems.

The CAP LTER studies in central Arizona examine how, in a dynamic environment, the services provided by evolving urban ecosystems affect human outcomes and behavior, and how human actions (responses) alter patterns of ecosystem structure and function, and ultimately urban sustainability.

Ecology for cities

Urban ecosystem services involve scientific, philosophical, political, and practical concerns for urban ecology because, although it is clear that nature provides benefits to humanity and that without those benefits human development cannot continue, organizations promoting nature in cities are being asked to quantify the value of urban greenspaces in terms of the provision of those services. Many of the services are appreciated only by sections of the total population, so valuation is linked to preferences that are difficult to establish equitably. Many of the commonly cited environmental benefits of urban greenspaces remain poorly supported by empirical evidence. Benefits to one group of
people may be ecosystem disservices to another. Some social groups seldom use particular parks, and for many individuals urban woodlands can be “landscapes of fear.” Thus the complexity of urban socioecological systems impinges directly on applying urban ecology.

Local issues of urban design are but one concern. Urban ecology has a central role in the debate about urban futures. Preparing urban settlements for the future involves both creating new eco-towns or eco-cities embodying sustainable lifestyles and smart technologies, and adapting existing settlements to environmental, social, and economic changes by increasing their resilience and sustainability. Green technologies and green spaces are integral to both processes, involving everything from green roofs and green walls to sustainable drainage and urban wildlife reserves.

Globally, urban ecologists are asked to contribute to establishing sustainable and resilient urban areas. Planners, architects, designers, decision-makers, and society expect urban ecologists to be able to offer guidance based on locally relevant ecological information; and, from that, they expect to derive general principles that can be applied more widely, even globally. This is no small challenge. It requires an understanding of the ecological patterns found within urban areas and the processes operating in them, which are very different from the traditional study of urban ecology.

Urban ecology is evolving rapidly with developments in both theory and practice. In working with complex socioecological systems, it is developing new opportunities for sound practice and good urban management. Some of the key elements of an emerging philosophy of adaptive practice for ecological design to help develop existing urban areas into resilient future cities (Pickett, Cadenasso, and McGrath 2013) include:

- using ecological knowledge to improve ecological urban design and to avoid dependence on green, or “eco,” ideology;
- recognizing that urban agglomerations or metacities (complexes of cities, suburbs, peri-urban and ex-urban areas) are spatially extensive, dynamic mosaics, in which flows, distant connections, historic legacies, and the impacts of innovation, redevelopment, and crisis act in localized explicit ways;
- seeing how design affects flows, creating both obvious and inconspicuous or unintentional feedbacks, some of which have the potential to enhance the adaptive capacity of affected areas;
- avoiding the unintended or inconvenient negative effects of designs that produce environmental injustice for sectors of the population;
- using art as a powerful tool for linking design with its ecological implications, to engage the public and decision-makers in the environmental flows and feedbacks that exist in urban systems;
- using socially and politically constructed narratives to help incorporate ecological awareness and knowledge into design processes;
- anticipating changes resulting from feedbacks and from wider-scale factors such as climate, human and biotic migrations, and shifting economic investments in order to include adaptive processes and structures in metacities (urban agglomerations of over 10 million inhabitants) by design as a key tool for resilience;
- using landscape as a shared medium, coupled with the consideration of designs as experiments, to make the learning aspect of design more explicit;
- assessing the ecological effects of proposed projects as part of a process of evaluation–feedback–revision in design and construction.
• constantly appraising impacts of urban growth, change, and adaptation on the hydrological systems to cope with floods and droughts, contamination, treatment, storage, sea level rise and coastal change.

Urban ecology for future cities

Urban ecosystems depend on both natural and anthropogenic substrates, and are highly affected by air, soil, and water pollution. They help to control runoff and reduce the urban heat island effect. They contribute to human health. Although some air and water pollutants may be trapped by trees and aquatic vegetation, in terms of air quality the benefits of urban tree planting programs may be overstated. The great scientific challenges in urban ecology are coupled with the social, cultural, political, and economic challenges of fitting proposed actions into socioecological systems. Studies of urban socioecological systems, for example in Baltimore, United States, and in Cape Town, South Africa, reveal that many urban ecological hypotheses, such as that biodiversity is lower in urban areas than in rural areas, are not valid. They emphasize that the diversity of urban areas has arisen through cultural and social diversity and the different values people place on particular urban greenspaces. Within large urban settlements, biodiversity and biogeochemical characteristics may differ along gradients between the truly rural and the intensely built-up urban core. New frameworks for urban ecology are being proposed to cope with this diversity and complexity, but they are not necessarily transferable from one continent to another in an age where the largest and densest urban populations tend to be found in the high-density but low-rise cities of Asia and Africa. The human impacts of environmental and ecological conditions and changes are most severe in the poorest cities, but the bulk of research on urban ecology is carried out in the more affluent cities. As in many areas of geography, the most salient urban ecological work in Asia, Africa, and South America is on immediate applied issues, aiming to alleviate the worst conditions and to improve health and wellbeing through environmental and changes. Many lessons can be learned from past experiences elsewhere, but all too often political action only happens after a local disaster. Urban ecology can help to avoid some of the disasters, but without relevant local case studies it is difficult to persuade decision-makers that things can be improved. There are real opportunities and challenges in setting new priorities for urban ecology, for action not criticism, for solving problems not political theorizing.

Globally, urban ecologists are asked to contribute to establishing sustainable and resilient urban areas. Practitioners and society generally expect urban ecologists to offer guidance based on locally relevant ecological information, and from that to derive general principles that can be applied more widely, even globally, despite the major differences in natural and social conditions affecting the world’s urban places.

SEE ALSO: Built environments; Chicago School; Climate change adaptation and social transformation; Ecological footprint; Ecosystem services; Health and wellbeing; Sustainable cities

References


Further reading


Urban elites

Ferenc Gyuris
Eötvös Loránd University, Hungary

Dominance, division of labor, and differential evaluation exist in all societies. There are always differences between individuals as well as social groups in their ability to command various kinds of resources, even if the appearance and level of these differences vary on a broad scale. Furthermore, most people have a different level of command over various resources. A simple and important concept is Bourdieu’s (1998) notion of the forms of capital and social space, which interprets social position as a function of economic and cultural capitals. In this interpretation, while industrialists may own much economic capital but little cultural, artists may have the opposite combination.

In reality, a much broader set of resources exists than is suggested by this simplistic model. Hence, the number of possible combinations is very high, and most social actors belong to categories with limited command over certain resources. There is always a limited number of actors, however, who are able to mobilize a wide array of resources. This narrow stratum is the (power) elite, which can be identified at various geographical scales. In the urban context, this powerful group of actors is called the urban elite. This is the group whose interests and decisions have a decisive influence on which policies are issued, which economic decisions are made, and which growth image comes to dominate in the city.

Urban elites, of course, are never homogeneous. They are fragmented, and some of their interests can be conflicting. The force that unifies them is the shared interest in sustaining their economic and political position, which is manifest in their support of initiatives and decisions they all expect to benefit from. In contemporary cities of the neoliberal age, this point involves treating the city as a “growth machine,” which means forming “growth coalitions” (Logan and Molotch 1987) with other actors within an urban elite and promoting economic growth as if it were the most important city goal. This argument today is frequently utilized to justify “urban privatism” (Peck 1995), the privatization of the urban commons and communal services such as public spaces, water supply, or education, and deregulation and liberalization in the urban economy, including the management of the urban labor force and the housing sector.

In this context, a multitude of institutional means is put in place to reinforce an urban elite’s special position. These include the legal system or law enforcement organizations, which enable in the postdemocratic city the use of techniques aimed at keeping social tensions under control. Hence, neoliberal urban elites strongly rely on “governing through crime” (Simon 2007), that is, the legal and discursive criminalization and severe punishment of activities and behaviors that elites consider undesirable.

Moreover, urban elites hold key positions in economic institutions as well as in the religious, educational, and civic realms, which play a key role in indoctrinating people to accept prevailing conditions. Crucial elements in this are advertising economic growth as the central imperative for the survival of cities, as well as creating a common identity for urban inhabitants through fabricating collective memories about
Urban elites are usually demographically distinctive in the city. Besides being disproportionately affluent, owning property, and having easy access to the levers of local power, they are also highly educated and predominantly male, and embody the dominant racial category of that city and society. Hence, urban elites are highly genderized and racialized, and their residences are segregated in urban space.

Although urban elites occupy a leading position in urban society, they are also embedded in a complex power hierarchy of various actors at different geographical scales, who also have the ability to shape urban realities. Even political leaders and influential economic actors in the city have to consider and adapt to events at the national and global scales. Metropolitan governments, for example, cannot ignore the actions and expectations of the national government or multinational institutions such as the United Nations, the World Bank, or, in Europe, the European Union, even if these conflict with their own desires. Furthermore, though urban elites are much more powerful than other urban actors, they are also strongly reliant on, and to some extent even constrained by, the latter, for without other actors the city could not function, either in economic or in political terms. Therefore, urban elites never have exclusive mastery over urban realities.

Power relations among urban elites are also dynamic and ever evolving, which often muddies the sense of who constitutes the most powerful political institution in the city. Powerful institutions can, under certain circumstances, quickly lose power. New institutions can step in to fill the void, creating an ever-changing political reality of who rules the city. Such institutions, it follows, must read political trends and processes in the city meticulously to stay relevant and in power. Without identifying emerging political and social trends, currently powerful political institutions place themselves at risk of losing touch.

In the current neoliberal world, with a historically unprecedented intensification of transnational flows of information, capital, and labor, urban elites are much less tied to a single city in geographical space. They are increasingly becoming part of a transnational elite, which is reliant on a number of places, mainly urban ones, where, for example, their children can be educated and their economic partners and political allies reside, and where they organize their leisure activities. The globalization of urban elites also makes them increasingly open to adopting globalized urban policies with the potential of political success. This results in extreme geographical mobility of neoliberal policies, such as those of the creative city.

The exclusive position of urban elites and the resulting urban inequality raise important normative questions. On the one hand, divisions of labor attributing more power to certain individuals are necessary prerequisites for the functioning of society and are thus present in urban social systems. It would be misleading to suppose that the interests of urban elites necessarily conflict with those of other urban actors, and that urban elites themselves unavoidably change the situation of other social strata in the city for the worse. For these reasons, the emergence of urban elites seems unavoidable and, in some ways and under certain circumstances, desirable. On the other hand, however, the mechanisms that determine who can join urban elites, who can remain in the group, and who is dropped should be subject to continual critical investigation.

Similar important questions are what activities urban elites should engage in (in moral terms),
are allowed to do (in a legal sense), can do (in technical terms), and actually do in a given situation, and to what extent these are in accordance with the expectations of disparate social groups. These questions are clearly open and debatable, both in neoliberal and non-neoliberal cities. To define related legal and institutional structures and practices and to continually fit them to the constantly changing views and interests of society are necessary and difficult tasks that should be addressed by urban elites.

SEE ALSO: Growth machines; Power; Social capital; Urban regimes

References


Further reading


Urban geography

Matthew B. Anderson  
*Eastern Washington University, USA*

The principal focus of urban geography has been and continues to be the city, and how political, economic, and social forces and human–environmental relations flow and reverberate through urban landscapes. Urban geography is among the largest and most diverse subfields of human geography, as urban geographers have been influenced by numerous (and often competing) philosophical and epistemological approaches to the study of the city. There are just as many ways to define and understand the city as there are to discuss the ways in which it has been examined by urban geographers.

The city is widely identified as a location marked by a concentration of people, economic producers, services, and physical infrastructural elements. At the same time, it is widely seen as a political and administrative entity, a space of flows (i.e., money, goods, and people), a setting through which social relationships are established, sustained, and evolve, and where everyday lives unfold. The discipline is best summarized as concerned with untangling this complexity, with understanding the myriad urban environments that the majority of people in the world call home, and with explaining the varying “socio–spatial similarities and differences that exist between and within urban places” (Pacione 2005, 17–18).

Urban geography also has a dynamic, interdisciplinary character. The discipline has historically reached beyond the confines of geography, frequently engaging with and incorporating insights and perspectives from urban planning, sociology, or economics. Urban geography can be conceptualized as an organic entity, coevolving in dialectic relation with other disciplines, both within and beyond geography. It might seem that urban geography therefore lacks a center, or core set of identifiable or coherent characteristics. This is not so, as what distinguishes urban geographers from urban scholars in other disciplines is the centrality of what is often referred to as the spatial perspective, perhaps the key defining feature of geography as a whole. This concept ties urban geographers together as the common thread that harnesses the discipline’s integrative power.

But what is this spatial perspective? There is no simple answer, as the significance and even meaning of space varies, from the Cartesian conception of space in, for example, Walter Christaller’s central place theory to Henri Lefebvre’s conception of space as a social product. In general, however, and in the context of urban geography specifically, it is a perspective that identifies and treats space as an important analytic concept for understanding and explaining the development of urban landscapes, from sprawling suburban subdivisions to disinvested urban cores and glossy downtowns, and how these landscapes are experienced by users and inhabitants. The process of urbanization is also a material force, the manifestation, or spatial expression, of the broader social system within
URBAN GEOGRAPHY

which it is embedded, that both shapes it and is shaped by it.

Positivism, spatial analysis, and modernist urban planning

During the early twentieth century, urban geography was a much smaller discipline. This began to change around 1950, when a rapidly increasing proportion of the world's population shifted to urban locations. As a result, greater scholarly interest in the city and the process of urbanization emerged. In many ways, this early history of urban geography is reflected through the history of human geography in general, as most subfields evolved through a succession of dominant epistemological paradigms that marked the discipline as a whole, such as environmental determinism, possibilism, and the region (Cresswell 2013).

Positivism and spatial analysis

During this time, urban geography was still a minor research area, which began to change during the so-called quantitative revolution in human geography. To be considered a legitimate spatial “science,” an elusive goal for decades, urban geographers adopted the philosophy of positivism and the development of quantitative methods, models, and testable theories of human spatial behavior. Here a central tenet is that only things that can be experienced as observable and countable facts can and should count as the empirical foundation of true and objective knowledge, the substance of a science where hypotheses are to be formulated mathematically (Cresswell 2013).

Much of this research, led by Brian Berry, William Bunge, Edward Ullman, Peter Haggett, and David Harvey, happened to be situated within the urban context, invigorating urban geography as a result. Supported by new computer-based technologies, the possibility of identifying general patterns through statistical data and replicable (and falsifiable) hypotheses led geography down the path of a “hard” and empirically grounded science guided by the “principles and practices of scientific method” (Hubbard 2006, 28). In the process, urban geography became a more prominent subfield geared toward the search for general laws in the spatial structure and organization of the city.

Spatial analysis was also influenced by the principles of neoclassical economics, introduced to geographers by European scholars such as Walter Christaller, Alfred Weber, and August Losch. The primary origins of this influence resided in the location theory of the nineteenth-century economist Johann Heinrich von Thunen, who sought to understand the location and spatial distribution of economic activity. Building on this background was Christaller’s *Central Places in Southern Germany* (1966/1933), which posits that clusters of economic activity function as “central places” that service surrounding areas. A key assertion was the notion that people typically travel the least distance to acquire the goods and services they need, and that certain types of economic clusters typically locate in close proximity to important resource base or critical mass of consumers. For example, Pittsburgh is located where it is because of its proximity to the coal deposits and mining operations in Pennsylvania and West Virginia. This location, according to the theory, is what underpinned this city’s early manufacturing of steel.

In short, the locational patterns of human behavior, from local-scale interaction and movement to the spatial ordering of entire urban landscapes, was understood as a function of the friction of distance, people minimizing the costs of transportation and communication, otherwise known as the principle of least net effort.
This principle was treated as a central explanatory variable, and underpinned the assumptions, models, and theoretical formulations (i.e., spatial interaction theory) of many scholars working in this tradition, from William Alonso’s oft-cited bid rent theory to the methodological expositions outlined in David Harvey’s landmark text Explanation in Geography (1969).

The Chicago School of urbanism

This approach to urban geography was also influenced by early twentieth-century urban sociology, particularly the Chicago School of urbanism. The Chicago School was composed of a number of scholars working in or affiliated with the sociology department at the University of Chicago during the 1920s and 1930s, and included Robert Park, Ernest Burgess, and Louis Wirth. Inspired by nineteenth-century European scholars, notably August Comte and Émile Durkheim, this school of thought pioneered the positivist approach to understanding social and morphological change in the industrialized city. Much of their work was based on quantitative analyses of dense descriptive statistics derived from participant and visual observation. It also highlighted Chicago as the paradigmatic modern city (see Park, Burgess, and McKenzie 1925).

The Chicago School presented a model of urban growth and change, perhaps best reflected in Ernest Burgess’s concentric-ring diagram of 1920s Chicago. Depicting the sociospatial differentiation of Chicago, the diagram visualizes a theory of how cities were understood to grow and evolve over time, marking different zones dominated by particular land uses and functions, such as working-class housing, industrial factories, commerce in the central business district, affluent housing at the periphery, and the oft-cited zone in transition of newly arrived immigrants. This model, often referred to as a scientific urban ecology, conceptualized the city as a living organism, with the different zones of the diagram reflecting the different organs that functioned within (and in the service of) the broader, living entity of the city.

This approach also adopted Darwinian explanations for the social behavior and patterns that were being observed. Replete with biological metaphors, people were seen to be enmeshed in competitive struggles to survive. The city’s spatial ordering (represented by concentric rings) can be understood as the outcome of this supposedly natural competition. This process was made explicit by Park: “Reduce all social relations to relations of space and it would be possible to apply to human relations the fundamental logic of the physical sciences” (Park 1936, 7).

For Chicago School scholars, this assumption explained why particular social-class groups tended to live in the same neighborhoods, as the manifestation of intergroup competition where particular groups found it more advantageous to congregate together as a competitive strategy. This schematic also formed the basis for the theory of invasion and secession, the phenomenon where one group is observed to move into, or “invade,” another group’s neighborhood, resulting in their relocation, or succession, to another area, thereby pushing the city further outward into the periphery and reproducing this checkered, hyperdifferentiated urban landscape. This process was interpreted as a natural process of neighborhood change.

This model also competed with others, such as Homer Hoyt’s sector model. Both models, however, are now considered overly simplistic and limited in explanatory strength. The Chicago School was nonetheless an important reference point in the subsequent development of urban geography as a positivist spatial science during the 1950s and 1960s, as it represented among the first attempts to draw conclusions
about the ever-changing spatial structure and social patterning of the modern urban landscape (Hubbard 2006).

The behavioral critique

The quantitative revolution proved important for catapulting human geography in general, and urban geography in particular, to the status of a legitimate science. The positivist approach became an object of critique, however, particularly in response to the rather poor predictability of the models this research generated (Pacione 2005). While these were designed to account for all aspects of human spatial behavior, much seemed to fall through the cracks. Despite the general patterns explained by the principles of distance decay, minimization of transportation costs, and so on, there still existed a notable degree of variability that could not be accounted for by these models.

This concern led to a shift toward what became known as behavioralism, an attempt to highlight the role of human agency, particularly people’s cognitive or decision-making processes, as a contingent variable that mediates general patterns. As an outgrowth of positivist spatial analysis, this shift was perhaps most influenced by Kevin Lynch’s *The Image of the City* (1960), where the mental map concept was first introduced. Deployed here was a method of cognitive mapping designed to visually construct people’s spatialized imaginings of the city using questionnaires and interviews. This method inspired a range of scholars, such as Julian Wolpert and Larry Brown, who sought to understand people’s less than utility-maximizing behavior by illuminating the rather messy, subjective, and often inaccurate imaginings people held of urban environments.

Through the use of more qualitative methods, it was asserted that the previously too simplistic models of the city could be strengthened to better account for variability, from people’s migratory and spatial mobility patterns to their residential preferences and perceptions of city spaces. This critique, although still situated within the positivist frame, ultimately provided the foundation for the subsequent development of humanistic urban geography (to be discussed in the next section), an epistemological approach that departed from positivism altogether.

Modernist urban planning

With the emergence of the modern industrial city, the necessity of managing and accommodating the spatial demands of a rapidly growing and urbanizing industrial capitalism became the purview of an emergent professional practice: urban planning. This would become another source of influence and inspiration in the development of urban geography, as both disciplines have historically placed a shared emphasis on the spatial dimension and constitution of social life in the modern city; however, during the early to mid-twentieth century, urban planning remained somewhat disconnected from urban geography. While some research in urban geography was meant for planning audiences (i.e., Christaller’s central place theory and Kevin Lynch’s cognitive mapping), only recently have these two communities of scholars and practitioners entered into more substantive dialogues.

Haussmann’s restructuring of Paris during Second Empire France (1852–1870) is typically noted as the first significant example of modern planning – sanitized water, plumbing, and sewage systems, street lighting, monumental civic squares, and the city’s famous network of wide boulevards – and deeply influenced the ways in which the supposedly rational internal ordering of cities was to be imagined and materialized. In North America, this influence
was first manifest through the City Beautiful Movement of the early twentieth century (Hall 2002). Perhaps the most prominent figure of this movement was the planner Daniel Burnham, who designed the reconfiguring of many cities (i.e., Chicago, Cleveland, Washington, DC, and San Francisco, among others) using Haussmann’s Second Empire Paris as the template.

This mode of planning, while it was rational to civic elites and industrial capitalists, ultimately resulted in numerous problems, most notably felt by the urban poor. In short, the living conditions of the working class were neglected, and they suffered immensely as investment was concentrated in the construction of grand avenues, central city monuments and civic spaces, and the servicing of more affluent residential areas. Famously chronicled by Frederick Engels in *The Condition of the Working Class in England* (1844), this worsening state of affairs – starvation, poverty, disease – ignited the garden cities counterplanning movement led by Ebenezer Howard. Based on a polycentric network of garden cities linked via mass transit, Howard envisioned more equitable social processes engineered through spatial planning, with each garden city linked with mass transit through radial concentric rings of avenues, small-scale enterprises (including agriculture), and central civic spaces.

Although this vision never really came to fruition, its influence has been felt through the twentieth century across Europe, North America, and beyond. For instance, it is here that the neighborhood unit in urban planning first emerged, and later influenced the regional planning of Clarence Perry, Lewis Mumford, and others. This vision, however, was subordinated to the emergence of what is often referred to as modernist urban planning, a paradigm and vision of the city that dominated much of the twentieth century. Modernist urban planning, sharing much in common with spatial scientists, was also based on the understanding of humans as rational economic beings, and held that similarly rational planning designs could help improve lives and contribute to “progress.” Reflected most famously in the Swiss planner Le Corbusier’s grandiose designs of high-rise living marked by right angles and standardized monolithic facades amid plentiful parkland and high-volume expressways, and implemented by neo-Haussmann technocrats, such as New York’s Robert Moses, the influence of this planning paradigm on the urban landscape has been profound.

From the sprawling cookie-cutter suburban landscapes that dot the periphery of most North American metropolitan regions to the public housing high-rises that marked many now deeply stigmatized inner-city spaces, the characteristics and consequences of the modernist urban landscape – municipal fragmentation, inner-city decay, automobile dependency, and deepening sociospatial polarization – have also become the context and substance for what much critical scholarship in urban geography and planning has since been about. The critique of this trend, initiated perhaps most notably by Jane Jacobs in her classic text *The Death and Life of Great American Cities* (1961), sought to recover the importance of the neighborhood unit as the basis for the flourishing of urban life. Planning should be a bottom-up, rather than a top-down, affair, with emphasis placed on, for example, close proximity between work and home, higher-density living, and a vibrant and walkable street life. The influence of Jacobs and Mumford, in particular, has since gained considerable momentum and has deeply informed much of urban design reform over the past few decades, most notably the principles of New Urbanism.

The year 1970 was a watershed in human geography. Spatial analysis was now under assault from numerous critical commentators, whose critiques took one of two different forms, each shaped by different philosophical influences. Both were stimulated, in part, by a sweeping rejection of positivism and the notably less than scientific assumptions and problematic interpretations that marked the spatial analysis tradition. Yet they were also markedly at odds with one another.

Humanistic urban geography

Extending the behavioral critique of spatial analysis, humanistic urban geography emerged in the 1970s and flourished into the 1980s. This tradition’s influence in urban geography and beyond is considerable and remains strong today. Spurned was the supposedly delusional rationalism that had previously pervaded the work of urban geographers, sociologists, and planners. To humanists, positivism sharply limited the kinds of questions that could be asked and, ultimately, the types of knowledge that could be produced. And, rather than merely studying the urban world, it was also having very real and destructive impacts, echoing the critics of modernist planning. For instance, urban landscapes were observed as being steamrolled by a globalizing homogeneity that reflected an increasingly technocratic understanding of the city shared by planners and spatial scientists and envisioned from the same towering elitist throne.

For humanists, what were inherently subjective interpretations and assumptions about the world masqueraded in the form of statistics and models, the positivist veneer of objectivity. For spatial analysts, knowledge was not being produced by subjective humans, but via a science (including supposedly impartial scientists) that could overcome such subjectivity. But, for humanists, such thinking led to an impoverished conception of human beings, whereby people were reduced to mere predictable things that circulate through the city according to some rational logic – what many now refer to pejoratively as homo economicus. To these critics it was no surprise that the models predicted human behavior so poorly.

Building on this critique, the humanistic approach adopted a different and alternative epistemology for the study of the city. Concerned less with hypothesis testing of only observable and quantifiable facts, humanism was (and remains) especially influenced by phenomenology, a philosophy about the “essence” of things which drew its inspiration from a variety of philosophers, particularly Edmund Husserl. According to phenomenology, everything has an essence, a core of what makes things what they are, which can be ascertained only through deep immersion in and experience of these things (Cresswell 2013).

For urban geographers, this new direction focused on the concept of place, and on how places, defined as spaces inscribed with human meaning, are experienced (i.e., home, workplace, civic square, or public park). As notably presented by Yi-Fu Tuan in his landmark text *Space and Place: The Perspective of Experience* (1977), the myriad ways in which people experience the spaces they encounter (and the meanings they attach to them) are important in understanding the nuanced contours of urban life.

Urban political economy

The humanistic tradition was critiqued by another parallel movement that emerged during the 1970s, urban political economy, which
was informed especially by Marxist analysis. Humanists purportedly allocated too much agency to individuals in their experiences and shaping of urban landscapes. Humanist research was also considered too apolitical, silencing uneven power relations between, for instance, elite capitalist actors and working-class minority populations. Where was the power of structural forces and influences? Where was the recognition of profound constraints imposed on actors and institutions in the everyday realities of capitalism? The humanistic perspective, according to these scholars, had very little in the way of answers to these questions and, as such, was deemed too simplistic an approach.

During the 1970s a revival of critical social theory and Marxism swept across universities in the English-speaking world. In geography, this “radical turn” was particularly influenced by an emphasis on the structural forces at work in shaping the world around us (including our perception of it). In continental Europe, the Marxist tradition had remained comparatively strong through the twentieth century, propelled by a number of philosophers, notably Manuel Castells and Henri Lefebvre. In this setting, urban political economy emerged as a major intellectual force in urban geography. The figure who spearheaded this movement in geography was, ironically, one of the key promoters of spatial analysis a few short years earlier, David Harvey.

Amid the world’s volatile social climate in the 1960s, including the Civil Rights Movement in the United States, the numerous urban uprisings and violent protests against the Vietnam War across North America and Europe, and the economic evisceration of many inner cities (particularly in North America), Harvey (1973) argued that geography risked falling into irrelevance. In short, both spatial scientists and humanists were deemed incapable of asking the kinds of questions that could shed light on the state of affairs.

Urban political economy conceptualizes cities as economic engines that are internally shaped by political and economic processes, though always mediated by local actors and institutions. They are thus sites where both local and multinational businesses and civic elites make and remake landscapes to produce and sell goods and services. In the process, governing actors (i.e., mayors, developers, and financial institutions) are understood as keenly responsive to changing economic realities and sociopolitical conditions, and continuously impact and revolutionize the spatial structure of the city in order to maximize profitability.

Spatial patterns of cities, in this perspective, are not merely the outcome of a natural competitive process. Rather, they are a manifestation of a central feature of capitalism, what Neil Smith (1984) termed “uneven development.” Poor neighborhoods, for instance, are poor precisely because wealthy neighborhoods are wealthy; these two phenomena are not unrelated. Why? Because poverty is not only an outcome of, but also a necessary input to, the inner workings of an economy driven by the capitalist mode of production. It is an outcome that represents the materialization of the exploitative relationship between capital and labor. But it is also a vital precondition for the source of profit, the surplus value maximized and appropriated from workers by capitalists through the labor process. A perpetually underemployed population, thus, maintains favorable conditions for profitability.

In short, uneven urban development is understood as the spatial expression of a deep structural force. Deindustrialization, for instance, is the flight of capital investment from high-rent urban districts in the Global North to much cheaper labor supplies overseas, for example in China. This flight is a means to restore and enhance profit margins, a process that comes at the expense of the very places where
URBAN GEOGRAPHY

investment had previously been fixed. In this way, capitalism requires some places to be less developed than others, the consequences of which – homelessness, despair, deepening inequality – were critically examined through the 1980s by a plethora of urban geographers and sociologists, such as Michael Dear, Jennifer Wolch, Loïc Wacquant, and William Julius Wilson, among others.

Richard Walker similarly chronicled uneven movement at the metropolitan scale in the context of mass suburbanization in North America. Here, the flight of middle-class whites, jobs, and resources during the 1950s and 1960s to the suburban periphery contributed to the decimation of inner-city districts (also see Beauregard 2006). This spatial fix later became regional, as Southern and Western cities (i.e., Houston, Atlanta, Phoenix) grew rapidly at the continued expense of the declining cities of the Northeast and Midwest, the so-called Rust Belt. In short, capital, in its search to restore profit margins, became attracted to the cheaper land, fewer regulations, and nonunionized labor that marked much of the Sun Belt. This process was not confined to North America; Doreen Massey (1984) chronicled a similar multiscalar uneven development in the United Kingdom, as the manifestation of shifting spatial divisions of labor.

In these ways, poverty and inequality are not some accident or natural phenomenon, but actively made and remade conditions endemic to the economic system within which we live. For the spatial scientist, these conditions may have appeared to be natural, but for the Marxist this is only the surface appearance. The Marxist tradition was also not just content to produce knowledge. Like the humanist critique, the Marxist approach considered objectivity an illusion and that the spatial scientist was ultimately (wittingly or not) advancing elite agendas under the cover of objectivity. But, for Marxists, academics should also actively embrace the politics of their work by intervening in the world, changing it for the better. For these scholars, to do otherwise would be to support bourgeois practices.

In a series of articles by Harvey (1989a), urban political economy was significantly developed through a critical re-evaluation of urbanization in North America. By adopting Marx’s methodological approach of historical materialism, Harvey cast his analytic gaze on the dialectical relationship between urbanization and capitalist development. The role of technological innovations in transportation and communication was seen as transforming the urban landscape periodically – repeatedly demolishing and recreating it on a new and expanded scale. Urbanization unfolded by navigating a tension between fixity and mobility: as soon as one phase of capitalist investment materializes in the urban landscape, in the form of transportation infrastructure, housing stock, hospitals, office towers, shopping malls, and so on, this landscape becomes a prison that inhibits further growth precisely because it is (relatively) immobile. To accommodate the always enhancing flow and mobility of capital, capitalist elites, to paraphrase one of Harvey’s oft-cited passages, must walk a knife’s edge between, on the one hand, preserving the flow of value through the existing urban landscape and, on the other, having to tear down this very landscape to meet the dictates of an inherently expansionary capitalist economy.

In contrast to the humanist and the spatial science approaches, social conflict, has been a central concern in urban political economy, and is widely interpreted as a function of resistance to the effects of capitalist-induced inequality. People do not necessarily desire to live in a city that is functionalized as an incubator for capitalist profits. Rather, and usually in contrast, people desire decent housing, jobs, communities, and quality of life. And, as urban economies ebb
and flow according to the broader economy’s crisis-prone rhythms, the ability of workers, capitalists, and city governments alike to achieve their myriad and competing goals is subject to these changing conditions, with the effect of further intensifying conflict.

The humanistic critique and urban geography diversified

The influence of urban political economy has been nothing short of profound, and it remains a dominant perspective in urban geography today. It has also become quite diverse, incorporating other theoretical frames and insights in addition to Marxism. The radical turn was felt across human geography and influenced the rise of environmental studies, political ecology, feminist geography, and beyond. Urban geography was also, in turn, influenced by each of these other subfields, with scholarly inquiry in urban political economy gradually informed by insights from feminist scholars, political ecologists, and political geographers, as well as those working in emergent interdisciplinary fields, such as cultural studies.

The Marxist approach did face criticism, particularly from humanists, in the 1980s. Articulated perhaps most forcefully by David Ley and James Duncan, everything seemed to be explained by or reduced to the logic of capital. The supposed inner logic and large-scale structures are portrayed in a mystified way, with research findings ultimately resting more on belief than on empirical data. In this context, Marxism was also proclaimed as depriving people of agency: mysterious structural forces, rather than human nature, were now dictating human motivations and behavior.

This critique was levied perhaps most notably in the context of gentrification, the process of urban redevelopment whereby a low-income population is displaced by a more affluent one moving into the area. This phenomenon, although widespread and generalized today, had just begun to be noticed by urban scholars in the 1980s. Neil Smith, a student of Harvey, famously explained gentrification as the closing of what he called the “rent gap” (Smith 1996), the gap between existing land values (i.e., from low-income apartment units) in a given neighborhood and the potential values of that same land under different uses (i.e., high-income condominiums). The more depressed the land values, particularly in favorably located neighborhoods, the more likely gentrification is to unfold.

According to this thesis, gentrification should be expected in the most disinvested neighborhoods, but this has generally not been the case. Ley (1996) responded by asserting that Smith’s explanation is focused entirely on the supply side and, as a consequence, is severely limited. He argued for the importance of demand-side considerations, that gentrification is more a function of a mass of consumers seeking specific “cultural” amenities in particular neighborhoods. As such, developers must necessarily respond more to the specificities of this consumer demand than the size of the rent gap. This debate raged through the 1980s and 1990s, with gentrification now widely understood as a product of both supply and demand forces. In much Marxist thought, the rent gap is a reality, but it is mediated and complicated by particular consumption tastes, local business climates, the (re)packaging of ethnic histories and historic housing stocks, and stigmatized racialized spaces (i.e., public housing).

This debate was situated within the broader “cultural turn” that unfolded across the social sciences in the 1980s as a response to a prevailing Marxism considered too dominated by the influence of structuralism. Urban political
URBAN GEOGRAPHY

economy responded by incorporating these insights and adopting a wider methodological and analytical toolkit, which has resulted in a considerably more nuanced and diversified perspective. The outcome is an urban geography that is now quite eclectic, characterized not by the dominance of one or two paradigmatic epistemologies, but by different groups of scholars working within particular topical areas and informed by myriad cross-cutting theoretical approaches.

Contemporary urban geography (1990 to the present)

Contemporary urban geography is a discipline marked by its pluralism and multidimensionality. Spatial analysis, propelled by advances in technology, such as geographic information systems (GIS), has rebounded from the critical onslaught experienced through the 1970s and 1980s, and evolved into a more diverse and nuanced mode of inquiry. The more recent emergence of critical GIS in urban studies is a notable example of this (see the work of Sarah Elwood, Rina Ghose, Sara McLafferty, and Matthew Wilson), from research that uses GIS to examine shifting patterns of socioeconomic segregation to the differential access to resources (e.g., healthy food and health care) experienced among minority populations. Both humanistic and Marxist approaches have also been increasingly nuanced by the insights of the other as well as by a variety of new influences, such as postmodernism, post-structuralism, actor-network theory, critical race theory, and feminism. This section introduces some of the most significant thematic areas of contemporary urban geography and concludes with some brief remarks on the present moment.

The Los Angeles School of urbanism

During the 1980s it had become increasingly clear that metropolitan regions, particularly across North America, were developing in a notably polycentric fashion. No longer were cities structured around high-density cores and low-density, primarily residential, suburban peripheries. The urban/suburban dichotomy no longer made sense as the suburbs were urbanizing in the form of edge cities, and central city areas were increasingly targeted for gentrification and an entertainment-oriented “revitalization” in the form of elaborate consumption zones not unlike those of suburban shopping complexes. Urban landscapes were also increasingly marked by gated communities, hyperpoliced spaces, and a more pronounced socioeconomic polarization.

According to scholars of the emergent Los Angeles School of urbanism, cities had ceased to reflect the now outmoded concentric-ring model of the modern industrial city offered by the Chicago School. Urbanization was seen to be unfolding through a new historical layer of development, which, for these scholars, required a new set of theoretical and analytic tools to understand.

Led by Michael Dear, Edward Soja, Mike Davis, and a host of other predominately Marxist urban scholars working in Los Angeles, a new paradigm for understanding the so-called postmodern city was formed to understand the ways cities were developing (Soja 1989; Dear 2000). Los Angeles, in contrast to the classic modern industrial city of Chicago, was the empirical setting for this new mode of empirical and theoretical analysis, and was portrayed as the new prototype of the emergent postmodern city. Even in Chicago, this hyperdifferentiated layer of notably decentralized and disorderly development was superimposed on the relics of the older industrial landscape.
The Los Angeles school was based on a wide range of influences. Perhaps the most significant of these was postmodernism, which entailed a further rejection of modernist science by asserting that no objective truth could ever be obtained because all knowledge claims reflect as much the subjectivities of the researcher as the supposed truths being revealed, whether about the essence of places, powerful economic forces, or laws of spatial mobility. Michael Dear was a central proponent of this new urban geography, which incorporated insights from (and critiques of) social theorists as diverse as Henri Lefebvre, Michel Foucault, and Fredric Jameson into a new urban theory built on a taxonomy of postmodern urbanisms based on Los Angeles (Dear 2000).

Not all who were linked to the Los Angeles School were as eager to embrace the postmodern turn, however, which was impacting the social sciences in general through the 1980s. Mike Davis, for instance, rejected it, while Allen Scott and Michael Storper both tended to situate their work on emergent patterns of regionalization within the broad transition from Fordist to post-Fordist modes of production, which were marking the developed world during the 1970s and 1980s. Others criticized the claim of Los Angeles as paradigmatic and the notion that this emergent urban form represented some radical break from the past. To Harvey (1989b), the seeming disorder and chaos of Los Angeles was mere surface appearance, confusion arising from the ensuing spatial reworking of urban landscapes in accordance with the evolving dictates of an inherently expansionary capitalist economy.

This debate unfolded through the 1990s and early 2000s and is now extensive, with many of the insights stimulated by the Los Angeles School, such as the increasing polycentric and kaleidoscopic topology of metropolitan regions, now generally accepted. It also forced scholars to reflect on the part played by their own biases and subjectivities in the very knowledge they were producing. The incorporation of postmodernism into urban geography was also accompanied (indeed even influenced) by a host of other emergent epistemological approaches to the study of the city, from posthumanist studies of animal geographies (see Wolch 2002), to feminist approaches to the study of social reproduction in the household, and the experiences of women and children (see the work of Melissa Gilbert, Sallie Marston, Susan Hanson, Meghan Cope, Dolores Hayden, and Cindi Katz, among others).

**The production of space**

A number of texts by the French urban theorist Henri Lefebvre were translated into English for the first time during the early 1990s, most notably *The Production of Space* (1991/1974), which propelled another new kind of urban geography. For Lefebvre, space is a social product, an active ingredient in the constitution of the dominant mode of production driving a given society, whether feudalism, capitalism, or state socialism. Space, from this perspective, is both a constitutive variable and a mirror of the broader social system that produces it, including entire urban landscapes. In this capacity, produced social spaces also function as signifiers of symbolic meaning, or the spatial manifestation of ideological content, from, for example, themed residential housing districts (e.g., Thames Town in Shanghai) to public housing towers, urban parks, or monuments like the Empire State Building or the 9/11 memorial site in New York City.

The impact of Lefebvre, across human geography in general, and urban geography in particular, has been nothing short of profound, stimulating entire new areas of inquiry into the
URBAN GEOGRAPHY

relationships between space, the city, and capitalism. This “spatial turn” in urban geography has ranged from those engaged in the Right to the City movement and the politics of public space to those examining the myriad processes whereby social meaning is inscribed within particular spaces (see Cresswell 1996; Mitchell 2003; Merrifield 2013). In this way, Lefebvre explicitly rejected the structuralism that characterized much previous Marxist-inspired research.

Urban governance

Urban political economy was now a much more diverse perspective in the 1990s, as established Marxist insights melded with new conceptual approaches, particularly the poststructuralist emphasis on discourse. Incorporating insights from a variety of philosophers, such as Michel Foucault, Pierre Bourdieu, and Antonio Gramsci, a hybrid urban political economy perspective was developed to examine what was observed as a broad transition unfolding in the city, from urban managerialism to entrepreneurialism (Harvey 1989c). City governments were now partnering more with private actors – developers, business elites, and financial institutions – to develop policies aimed at resuscitating the city from industrial decline by stimulating private capitalist investment in the urban built environment. Now on their own to balance their budgets, city governments were spurning established social welfare provisions in favor of renewing central city areas as sites for capital accumulation (Merrifield 2014).

Scholars examining this “new urban politics” (Cox 1993), such as Jamie Peck, Roger Keil, Gordon MacLeod, and David Wilson, emphasized the discursive means of producing consent to legitimate this agenda. With mayors and developers cast as salvationists and the racialized poor demonized, the global trope was identified as a fear-based discursive formation that justified gentrification and central city refurbishment, a necessary endeavor in the face of the new harsh realities of globalization. A variety of conceptual frames – regime theory, regulation theory, and growth machine theory – were developed to study this emergent form of governance, the broader regimes of accumulation and regulation within which it is situated, and the sociospatial effects of these strategic discursive tactics (Jonas and Wilson 1999).

This research thrust also consisted of a broader critique of globalization. Identified more as a discursively constructed reality than anything else, globalization was understood less as the mysterious, actor-less force commonly portrayed by governing elites than as a process mobilized by and through the profit-seeking acts of these very governing elites. This reality was characterized by Erick Swyngedouw as “glocalization,” a term that overcomes the long-standing local/global binary that ultimately works to the advantage of the mainstream conception of globalization and the seemingly innocent localities subjected to its blistering impacts and dictates.

During the 2000s, this mode of inquiry inspired further critical theorizations of what was increasingly understood as the discursive division between culture and economy. The result was another conceptual frame – cultural economy – notably developed by Ash Amin and Nigel Thrift, Bob Jessop, and Allen Scott. These scholars have examined both the cultural dimensions of urban economies and the economic constitution of urban cultures. Parallel to this work is the influence of actor-network theory, derived from the French theorist Bruno Latour, which conceptualizes the city as a space of flows constituted by relations, networks, ideas, and material objects as signs and signifiers of a multiplicity of functions and meanings. Here, assemblages of sociomaterial relations are both
constituted within and extend well beyond the scale of the city (Amin and Thrift 2002).

Neoliberalism and the city

In the early 2000s a growing emphasis on the concept of neoliberalism emerged as an outgrowth of the burgeoning literature on urban governance. Neil Brenner and Nik Theodore’s edited volume, *Spaces of Neoliberalism: Urban Restructuring in North America and Western Europe* (2002), inaugurated what has now become a voluminous literature on the neoliberalization of the city, a process of political-economic restructuring that has been unfolding since the 1980s. Neoliberalism is a political ideology and ensemble of human sensibilities that privileges small government, deregulation, defunding of social programs, individual culpability, and the unleashing of the “free market.” Strongly promoted as the solution to the recession of the 1970s, it was first institutionalized by the Reagan and Thatcher administrations in the United States and the United Kingdom, respectively, and has since spread across the world.

The previously observed shift from government to governance was situated within the broader process of restructuring. The effect has been the drastic sociospatial transformation of urban landscapes, beginning with the rolling back of existing welfare state sociospatial structures (i.e., public housing, labor unions), followed by the rolling out of new institutional forms, policies, and practices guided by neoliberal principles, such as central city revitalization, mixed-income redevelopment, historical preservation, participatory governance, and tax-increment financing. These practices are not inherently negative, but evidence suggests that when they are mobilized within neoliberal agendas the outcome has generally been the gentrification of, and the eviction of low-income populations from, central city districts.

Research on urban neoliberalism has emphasized both its top-down dissemination from power centers (conservative think tanks, politicians, ideologues) and its contingent, uneven manifestation across urban contexts (see the work of Neil Brenner, Jamie Peck, Nik Theodore, Jason Hackworth, Roger Keil, and David Wilson, among many others). More than a simple monolithic ideology, neoliberalism is a complex, variegated institutional formation that is always mediated by local actors in a multiplicity of ways, perpetually mutating in relation to broader social, political, and economic conditions. It has also become deeply embedded in the social and physical form of the contemporary city, disciplining local populations and providing civic elites with the discursive resources and guiding rationality – from Richard Florida’s creative class to Andrés Duany’s New Urbanism – for justifying policies and practices geared toward facilitating upscale redevelopment.

Much of this research has been situated within urban political economy. Yet, within this frame, neoliberalism has been conceptualized in a number of different ways, from a set of policy prescriptions to a hegemonic ideology and mode of governmentality, with each understanding informed by a particular melding of neo-Marxian and poststructuralist insights. The literature on neoliberalism and its multifaceted impacts on the city now constitutes a significant proportion of critical research in urban geography and cognate disciplines.

World city theory

Another area of research that emerged in the 1990s is world city analysis. The sociologist Saskia Sassen first coined the concept of the global city in her now classic text *Global City: New York, London, Tokyo* (1991). Sassen provided a framework to analyze the function and
importance of cities within the global capitalist economy, and presented a conceptualization of globalization as constituted by a complex hierarchy and global network of metropolitan regions, linked together as interconnecting nodal points within this broader globalizing economy. For Sassen, New York, London, and Tokyo were the three alpha global cities in the world, the primary command and control centers propelling the global capitalist economy.

Loosely influenced by the world-systems theory of Immanuel Wallerstein, a number of urban geographers, notably Peter Taylor and his colleagues at Loughborough University, have further developed this mode of analysis since the turn of the twenty-first century. Of primary importance have been the methodological questions surrounding which variables are most important in measuring global city status and how cities are best categorized and ranked. One of the key insights is that population size does not necessarily equate with global city status. Some of the most populated metropolitan regions in the world, such as Mexico City, Lagos, and São Paulo, are not particularly high-ranking cities. Although the variables included and their degree of importance are debated, global cities are calculated by their political experience (innovations in political, economic, and business practices); their cultural experience (the ethnic diversity of the population); their information exchange (the existence of mass media and communications infrastructure for global networks); their human capital, particularly the existence of high-quality educational institutions and skilled workers in key sectors (finance, legal, and information); and the degree of business activity that takes place in them (the volume of transactions, commodities, and money that flows through sea- and airports and container facilities, and the number of multinational corporations and stock exchanges).

As the global economy expands and evolves, so does the global urban system, as some cities, such as Shanghai, Hong Kong, Beijing, and Moscow, have risen in the ranks while others have declined in importance. The shift in global hegemony from North America to East Asia is reflected in the rise of East Asian cities in the rankings. However, by some accounts, Tokyo has fallen, while other analysts have not only retained Tokyo but added Paris as a fourth alpha global city. This area of inquiry continues to attract urban scholars interested in charting the macroscale geography and evolution of the global urban system.

Urban political ecology

Urban political ecology also emerged in the early 2000s as a prominent research area within urban political economy. Its focus is ecological and environmental social issues informed predominately by Marxist insights. As such, urban political ecology can be understood as a point of intersection between physical, environmental, and political-economic domains. The city, and urbanization in particular, is a point of emphasis and is conceptualized as an always unfolding process propelling what many have termed the “metabolic” relationship between nature and the capitalist mode of production.

Developed by Erick Swyngedouw, Maria Kaika, and Nik Heynen, this approach draws inspiration from Neil Smith’s (1984) conception of nature not as a pregiven terrain on which humans exist but as a human construct itself. Based in Enlightenment notions of progress, the very concept of nature is anthropocentric, and stems from the ideological removal of humans from the so-called natural world of raw materials and nonhuman forms of life. In short, since the Industrial Revolution, nature has been conceived as an input of production, the source of
all raw materials, resources, commodities, and thus wealth. The capitalist system, essentially, is underpinned by this appropriation and transformation of nature by human labor and propelled by innovations in what is an always evolving, metabolic relationship.

Urbanization, it follows, can be understood as the process of perpetual shaping and reshaping of elaborate urban environments, or humanly constructed natures. Cities are not just social landscapes, the realm of human civilization divorced from some pristine nonhuman world, but rather are socio-natural landscapes, assemblages of people, raw materials, and resources. They are underpinned by myriad human–environment relations, from the oil that fuels human mobility to the coal that supplies electricity, the materials used to construct buildings, transportation systems, and the communication infrastructure, and the food and water (often transported over significant distance) that provide the very basis for human life and further development. Moreover, carbon dioxide is pumped into the atmosphere, fueling global warming, within or in support of the capitalist urbanization process. These various urban–nature relations are now well documented within this frame, with notable texts by William Cronon (1991) and Matthew Gandy (2002) frequently cited.

In short, the dominant Western conception of nature is based on a problematic premise, the division between the human and the nonhuman worlds. In this way, urban political ecology is also a critique of environmental conservation movements, whose object of conservation is typically some romanticized or pristine version of the same nonhuman nature. Thus, rather than promoting the conservation of nature, urban political ecologists tend to ask what kind of relationship with nature humans should construct, and in what kind of urban environments we want to live. At the core of urban political ecology are social justice, and the envisioning of a more progressive human–nature relationship and a socio-natural urban environment that privileges the majority of its human (and nonhuman) inhabitants rather than the contemporary city, which is fashioned by and for elite capitalist interests.

**Racial economy**

Only recently has the issue of race become important in urban political economy. While it was an important topic of study in urban sociology and critical race studies for decades, following David Wilson, Ruth Gilmore, Laura Pulido, and others, the specific connections between race and political economy have been only minimally examined. Racial economy, to Wilson (2009), is a perspective that explicitly analyzes race as a constitutive force within the unfolding of urban economies. Here, race is understood as a social construct rather than as an objectively existing thing, and functions as a semiotic resource drawn on by governing actors in their rhetoric justifying and shaping urban redevelopment agendas.

Research in racial economy is now diverse, from studies examining the role of race in narratives of industrial decline to the disproportionate targeting of the racialized poor by the rapidly growing prison industrial complex and the deepening marginalization of low-income racialized populations in the neoliberal period. Much of this work tends to be marked by qualitative methods such as ethnography, deconstruction of texts, and historical analysis. In general, race is conceived as a human construction and an expedient opportunity structure through which governing actors pursue their profit-seeking interests. In the process, such actors are seen to navigate people’s passions and predilections concerning race, as a racialized cast of characters (i.e., the demonized drug dealer or, more
recently, the civically responsible entrepreneur and corresponding imaginative spaces (public housing towners, ghettos, and gentrified townhomes) are discursively choreographed to the tune of central city revitalization.

Another area of inquiry within this perspective stems from the work of Elvin Wyly, Daniel Hammel, Phil Ashton, and Daniel Immergluck, among others. Here, for example, spatial analyses of quantitative datasets are mobilized to reveal the perpetuation of discriminatory lending patterns by financial institutions through the dissemination of subprime mortgages to previously untapped populations, such as Latinos and African Americans. Reflecting a kind of critical positivism (Barnes 2009), the analytic tools of spatial science are here melded with critical social theory to illuminate the uneven power relations operating behind a financial industry mobilized less by extending opportunities for homeownership to historically excluded populations, as was frequently pronounced, than by channeling accumulated financial capital (in the form of subprime mortgages) to new outlets for capitalist investment.

Current trends: the post-2008 crisis and planetary urbanization

The current moment tends to be marked by the influence of the global financial crisis of 2007–2009 and the myriad uneven impacts these events have had on cities around the world. Particularly noteworthy has been a plethora of commentary concerning the fate of neoliberalism amid what many have interpreted as a crisis of its own making. Despite initial pronouncements of the demise of neoliberalism, a consensus has emerged that not only have we not entered a post-neoliberal world, but that this mode of rationality has only been enlivened as a means of justifying business as usual (i.e., fiscal austerity, further cuts to social programs).

Others, such as Eric Sheppard, Andy Merrifield, and Neil Brenner, have proposed that significant reform is needed to the existing arsenal of theoretical tools urban scholars have at their disposal for understanding the city. For some, such models are disproportionately based on the cities of the Global North, particularly in the United States and the United Kingdom, which should be clear from this entry. In short, urban theory is in need of revision through analysis of the Global South, particularly the most rapidly urbanizing cities in East and South Asia, cities that do not necessarily reflect the various paths of development taken in the Global North. Indeed, much urban research now emphasizes the experience of cities in, for example, China, India, and postsocialist Central and Eastern Europe.

Moreover, the process of urbanization is now planetary, with more than half the world’s population living in areas classified as metropolitan. Words like “city” and “urban” have lost their former meanings as some of the largest metropolitan regions – Tokyo, Shanghai, Moscow, and Los Angeles – now extend hundreds of miles in any direction. Even rural communities are now shaped by and plugged into the global urban network. The world is becoming increasingly urbanized and linked through social and mass media. As such, resistance movements have also become global in that, everywhere from the global Occupy movement to the Indignados in Spain, from the revolts in Tahrir Square in Cairo to Syntagma Square in Athens, Iran, Tunisia, and beyond, social media and a greater global awareness among people in distant places can swiftly mobilize a critical mass of people in any public square. These emergent dynamics of urban resistance are now increasingly the focus of scholarly inquiry and debate.
Others have argued for more scholarly attention devoted to what Jennifer Robinson (2005) has referred to as “ordinary” cities. Here, understanding the city and urban processes is too reliant on overly studied large global cities like New York, Chicago, or London. The experiences of these cities do not necessarily reflect those of smaller, medium-sized cities, and may, in fact, be more exceptional than paradigmatic. The critique of this assertion, however, has been that all cities exhibit some broader structural patterns combined with local contingencies, and that a focus on ordinary cities, while not unimportant, may not reveal anything particularly novel or thought-provoking in how we understand the contemporary city.

This entry has provided an expansive account of both the historical development of urban geography and the major areas of inquiry that currently constitute the contemporary discipline. It does not, of course, capture everything, as urban geography is now far too diverse to comprehensively review in a single entry of this length without omissions; however, it is hoped that this entry can be used as a point of departure to further investigate the conceptual approaches and historical periods that have been discussed, as well as the key texts cited and recommended.

SEE ALSO: Actor-network theory; Central place theory; Chicago School; Cultural economy; Cultural studies; Edge city; Feminist geography; Gentrification; Geographic information science; Ghetto; Global cities; Glocalization; Growth machines; Homelessness; Humanistic geography; Los Angeles School; Marxist geography; Neighborhood; New Urbanism; Place; Postmodernity; Public-participation GIS; Public space; Race and racism; Rent gap; Rust Belt cities; Spatial analysis; Suburbanization; Urban elites; Urban managerialism; Urban political ecology; Urban politics; Urban redevelopment; Urban regimes; Urban uneven development

References


Further reading

Haggett, Peter. 1965. Location Analysis in Human Geography. London: Edward Arnold.


Urban managerialism developed as a framework for applied urban-scale policy research in the 1970s. As originally conceptualized, proponents of what became known as urban managerialism advocated for a focus on a class of professionals who, by virtue of their positions in state or parastate organizations, had “control” over the allocation of scarce resources in cities. These professionals were given various labels: gatekeepers, controllers, or simply managers. Conceived specifically in analyses of the operations of the British welfare state by sociologists, urban managerialism became embedded in methodological and theoretical debates and practices at the intersection of urban sociology, human geography, and urban studies for about 20 years.

It is important to note, for the sake of clarity, that the phrase managerialism was also used to denote a phase of urban governance. Here, the function of the local state was seen to be as manager of the provision of services (known as collective consumption) as well as a variety of place-based economic development functions at the urban scale. In this account, as the epoch of global neoliberalism and late capitalism unfolded, urban managerialism gave way to new forms of governance with state actors engaged in an array of public–private partnerships. This new phase of governance was captured with the concept of urban entrepreneurialism (Harvey 1989).

Here we will focus on scholars who worked with the original and subsequent conceptualizations of urban managerialism. They used this methodological lens and framework for analysis to examine the interplay between individual autonomy and organizational constraint in urban-scale governance which also, coincidentally, came to include the types of governance changes seen in the transition to urban entrepreneurialism.

The phrase urban managerialism originally coalesced in the writings of British urban sociologist Ray Pahl – particularly in the classic book Whose City? (Pahl 1970). This widely read book catalyzed debates over the interplay between a focus on collective consumption and work-place or production-based class struggles. Through analyses of access to, and distribution of, scarce urban resources and facilities, Pahl isolated the role of “controllers” in both the public and the private sector. In this earliest formulation, the scarce resources and facilities were often in the housing sector, with urban managers – or gatekeepers, as they were sometimes called – directly involved with the provision of housing units themselves. But it also included actors linked to sales and marketing, finance, and management. The housing sector studied included both owned and rental units. Many facets of the character of “controllers” were also examined, including their individual value systems and political and class ideologies. In Pahl’s formulation, which has endured in other scholars’ work, inequality through the uneven allocation of resources at the urban scale influenced the life chances
of individuals and groups who were already subjected to inequality in the world of work (the production sector) and therefore deserved an explicit and sustained research effort.

Early debates about this approach to the analysis of urban space and resources revolved around the tension between proponents of a progressive welfare state, which included Pahl, and others, including David Harvey, who saw these urban-scale allocations of resources as a sideshow to the “main event” of the inequalities systematically produced by capitalism. Critics questioned both the authority of local-scale actors and their autonomy to act. Pahl himself had struggled with this contradiction as well as with more focused critiques which questioned which specific managers were worthy of study and for what reasons. In fact, in a revised version of the original book and in other writings, he reformulated his original thesis to move away from a direct causal role for managers. It’s fair to say that for his critics, Pahl’s adaptations further undermined the relevance of the urban managerialist framework itself. This was especially the case in the United Kingdom as the welfare state was dismantled or hollowed out through privatization or through the imposition of market-based rules and structures which privileged narrow quantitative measures of “success.” This retrenchment of the state occurred under the leadership of both the Conservative Party and the Labour Party.

Several scholars, however, continued to focus on the interplay between the broader power relations of capitalism and the actions of gatekeepers or managers. Peter Williams was a key contributor to the collation and evaluation of empirical work in the late 1970s and early 1980s, and to a brief resuscitation of urban managerialism in the United Kingdom. In the United States, meanwhile, urban geographer David Wilson was also actively applying aspects of the approach to studying urban change (Wilson 1987). Wilson’s empirical investigations included both urban revitalization and the implementation of the United States Federal Community Development Block Grant program, which focused on housing, infrastructure, and other urban-scale redevelopment. He argued for the continued potential of an examination of the complexities of the blending of the individual life paths of gatekeepers with the institutional and organizational development of urban resources. The resulting accounts and analyses of the decision-making processes of managers, and their impact, recognized that while specific historic and material circumstances constrain the behavior of “managers,” they do not determine them. While considering broader political and economic changes and the structural design and rules of individual programs and local conditions, there was still room for explorations of the intentional actions of managers who had direct access and control of “scarce resources.” This brought Wilson and others full circle to the analytical power of some of sociologist Pahl’s original formulations and to rich, place-based accounts of urban change.

It is important to note that the focus of this later work was not restricted to distributional aspects of housing and other so-called scarce resources but was also applied to aspects of commercial redevelopment, urban revitalization, and neighborhood change. Here we see the analysis of the constrained behaviors of actors by Wilson and others finessing the either/or dichotomy of structure and agency which was the focus of intense theoretical debates at the time. This middle-level analysis was a contribution to debates and discussions of the hollowing out of the state, the development of new forms of local growth machines and place-based urban redevelopment, and the critical examination of state and parastate organizations. The lens
of urban managerialism and other facets of organizational analysis were deployed to examine the structure–agency contradiction through a synthesis with tenets of urban political economy.

The earlier tensions between studies focused on collective consumption and “sites of production” ameliorated somewhat as an ultra-commodified urban real estate sector, in all of its complexity, became the site for windfall profits, speculative investment, and place-based displays of wealth and power. It also provided cheek-by-jowl evidence of economic and social inequality as a context for interrogations of place-based economic and political processes.

The attentions of scholars in urban geography moved much more toward the analyses of urban entrepreneurialism and public–private partnerships for economic growth and away from a distinct focus on managers or gatekeepers per se. As a result, the explicit lens and conceptual label of urban managerialism has diminished in visibility in more recent times. However, insights from urban managerialism were deployed in analyses of the housing and service economies of the socialist and transitional economies of Eastern Europe and China.

Most recently, though less explicitly, examinations of urban governance driven by critical network theory, new institutionalism, and critical policy studies, reveal a deep and continuing concern with the complex interplay between organizational structure and regulatory and programmatic design, and with the autonomy and authority of “controllers” at various spatial scales. These studies include examinations of the provision of housing and other services but also broader topics related to urban redevelopment and local-scale sustainability. They are typically embedded in the discourses of stakeholder processes, and the interplay or even interchangeability between state actors as place-based entrepreneurs and capitalists as place-based managers in local applications to urban-scale social and economic change.

SEE ALSO: Urban politics; Urban redevelopment; Urban regimes

References


Further reading


The establishment of multiple racial and ethnic groups in cities often results in a patchwork of neighborhoods that can aptly be described as an urban mosaic. The urban processes that lead to an urban mosaic were described as far back as the Chicago School, by the sociologists Park, Burgess, and McKenzie (1925). The Chicago School sociologists were a group of human ecologists who used Chicago as their laboratory to learn how immigrants locate in the city. At the time of their writing, Chicago was a major port of entry for European immigrants (i.e., Italians, Germans, Polish, and Irish) who settled in distinct and separate neighborhoods. The human ecology principle of competition served as the primary organizing agent where, under the pressure of competition, ethnic groups carved out residential and functional niches within the city. The effect was to segregate people of similar origin into relatively homogeneous residential and functional subareas of the city.

Following this tradition in urban studies, the phrase “urban mosaic” was adopted as a title of a book in the mid-1970s (Timms 1975) to illuminate residential patterns in US cities. Yet, 20 years earlier, the quantification of city patterns using the techniques of social area analysis had begun to reveal neighborhood patchworks. Social area analysis attempted to describe and explain ethnic clusters within cities. Shevky and Bell (1955) first initiated social area analysis in the study of Los Angeles and San Francisco, using three groupings of variables they labeled social rank, family status, and ethnic status. The social area analysis method became widely used in the 1960s and was enhanced by computers that could handle large census datasets. The technique became known as factorial ecology (Berry and Kasarda 1977). Factorial ecology lasted as a leading methodology to investigate the city through the 1970s and 1980s, with the following common findings of the variable groupings:

- social rank or economic status exhibited sectoral characteristics;
- family status exhibited a concentric ring pattern; and
- ethnic status variables formed clusters.

Timms (1975) overlaid these findings of social space onto the physical space of the city to demonstrate a pattern he referred to as the urban mosaic. Today, the urban mosaic pattern persists in cities, but much of the contemporary discussion is embedded within the segregation literature, with debates ranging from those who see a rise of diverse multi-ethnic neighborhoods to those who suggest that the segregation of racial and ethnic groups persists.

SEE ALSO: Urban ecology

References


URBAN MOSAIC


Urban planning: human dynamics

Bruce Appleyard
San Diego State University, USA

Historically, one of the main forces driving the creation of physically proximal collective human settlements (villages, towns, cities, and metropolises) is the utility gained from interpersonal and inter-collective exchanges of goods, services, knowledge, sociocultural support for reasons of prosperity, security, love, learning, enlightenment, satisfaction, hope, and so forth.

As interpersonal communication is a central requirement of these exchanges (gestures, spoken and written language), early advances in personal telecommunications (telephone, fax, and computer) and mobility (automobiles) loosened spatial requirements for people to live and work in close proximity to one another to conduct such exchanges face-to-face. However, recent rapid increases in the availability and power of mobile hand-held devices enabling dynamic, on-demand exchanges in urban environments, improving people’s ability to navigate and execute the exchange and management of goods and services to meet their wants and needs (also known as “urban access”), is central to the sharing economy phenomenon emerging in the early twenty-first century and may significantly transform human settlement and behavior patterns (Appleyard 2013; Figure 1).

Origins of urban planning

Modern urban planning originated in response to the poor and overcrowded living conditions of the mid-nineteenth century cities, rapidly transformed by the Industrial Revolution. Public response was galvanized by writers such as Friedrich Engels in the United Kingdom and newspaper reporters (also known as “muck-rakers”) such as Jacob Riis in the United States who exposed abuses in tenement housing in his 1890 book How the Other Half Lives.

Modern urban planning grew as a profession to improve the health, safety, and welfare of overcrowded industrializing cities by focusing on technical and political processes guiding such things as the orderly use of land and the provision of infrastructure (streets, water, and sewer). In the mid- to late twentieth century, sustainable development, livability, and social equity came to represent ideal outcomes of all urban planning goals.

Early on, several key movements formed around the redesign of cities. One of the most notable city redesign efforts was Georges Eugene Haussmann’s redevelopment of Paris, between 1853 and 1870, undertaken during the rule of Napoleon III. Drawing on Baroque, axial planning ideas with terminating vistas, Haussmann carved new wide boulevards through the city linking important public places (parks, civic buildings, and rail stations). The boulevards were lined with trees and mandated six stories of housing over groundfloor retail
The ubiquitous use of GPS-enabled mobile hand-held devices that can facilitate dynamic, on-demand exchanges, and knowledge acquisition, such as how much time it will take until a bus arrives, are improving people’s ability to navigate urban environments and improve the quality of a transit rider’s experience. Photo by Bruce Appleyard.

An innovatively new municipal sewer system was also built beneath the boulevards. At the time, Haussmann’s work was admired worldwide as the model of city modernization. However, following the excesses and failures of modern urban renewal projects in the 1950s, Haussmann’s work has been criticized as an expression of positivism and political subjugation for how it cleared out old working-class neighborhoods that were areas of unrest, displacing many, mostly poor, people. Nevertheless, the boulevards of Paris are viewed as an important invention in urban planning as they not only opened up vast areas of the inner city to the public, providing sunlight, space, and “breathing room,” but they also forced the middle and upper classes to confront the reality of urban, working-class poverty as people from the dense surrounding urban areas spilled out onto the boulevards. The boulevards also introduced modern traffic to the inner city by providing wide, unencumbered roadways allowing private carriages to move at much faster speeds than previously possible (Larice and Macdonald 2012).

In addition to Haussmann’s modern redevelopment of Paris, other urban planning movements grew in response to overcrowding and poor living conditions of the Industrial Revolution, as follows:

- the parks movement,
- the city beautiful movement,
- the city efficient movement,
- the garden city movement,
- the modernist movement,
- New Urbanism and a place agenda.

The parks movement

During the time of Haussmann’s reconstruction of Paris, the works and writings of Frederick Law Olmsted Sr (1822–1903) bore the beginnings of the US parks movement. In 1858, Olmsted, with his partner Calvert Vaux, won the competition to design New York City’s Central Park, an 843 acre open space carved out of the city’s street grid pattern (designed in 1813). Viewed as the founder of modern landscape architecture, Olmsted believed that naturalistically designed parks with gentle landscapes offering views of pastures, meadows, and calm waters could play a central role in counteracting what he saw as the debilitating physical and mental stresses of living in crowded and congested urban settings.

The city beautiful movement

The 1893 Columbia Exposition in Chicago, led by Daniel Burnham, melded both the parks and city beautiful movements. Sponsored by local civic groups, and led by activists (mostly women), the implementation of these
movements advocated for improving the public realm through such elements as street lighting, fountains, ornamental plantings, shade trees, public art, and classically designed civic buildings and centers. While short-lived (mostly between 1900 and 1910) and derided for being concerned only with shallow design tastes of urban elites, the city beautiful movement is credited with historic preservation movements of the 1970s and maintaining focus throughout the city efficient and modernity movements on valuing urban aesthetics.

The city efficient movement

City beautiful began to give way to the city efficient or city practical movements in 1909, where many speakers at a 1909 Washington, DC, national conference on city planning prioritized efficiency and economy over beauty. Critics of “city beautiful” had varying complaints, such as largely ignoring solutions for the problem of overcrowded urban housing. But the prevailing beliefs were that city planning should foremost make cities more efficient, particularly in transportation and land use. Although city planners often ridiculed the city beautiful movement after 1909, elements of the city beautiful existed in dozens of city plans drawn up during the 1910s, such as park systems and civic centers. But the emphasis of city plans increasingly shifted to traffic and transportation.

Eventually, city planners would seek two types of authority. The first involved increasing cities’ powers to finance improvements, such as by raising debt limits and issuing special assessments for projects. The second involved greater public control over private property. The most notable result was the development of zoning laws. Zoning, initially known as districting, placed limits on the type of development that could take place in a particular area. Starting in California in 1908, at least 20 cities had adopted zoning proponents of the city efficient movement and had attempted to distance themselves from the city beautiful movement, but the movements were in many ways two steps on the road toward modern urban planning. For the first time in the history of the United States, a profession developed around the goal of shaping the development of the urban environment. Although efforts had been made in the nineteenth century to guide city building, the city beautiful and city efficient movements were the first attempts to deal with the issue on a citywide scale rather than on a piecemeal basis by 1917.

While different, both the city efficient and the city beautiful movements are viewed as major steps toward modern urban planning. Both represented initial attempts to deal with the issue on a citywide scale rather than on a piecemeal basis.

The garden city movement

The garden city movement was one of the most influential planning and design ideas of the late nineteenth century for providing a remedy to the perceived ills of the crowded and congested conditions of the nineteenth century city. The garden city concept was most influentially articulated by Ebenezer Howard in his 1898 book, To-morrow: A Peaceful Path to Real Reform, which was then reissued in 1902 as Garden Cities of To-morrow. In sum, Howard’s garden city concept provided a new vision of urban life that responded to the poor conditions of an industrial city in the following ways.

1. It envisioned new settlements that brought production and workers out of the tenements of the cities.
2. It also balanced agrarian ideals without ignoring the rise and utility of the capitalist, manufacturing economy, as shown
in Howard’s the “three magnets” diagram, shown in Figure 2.

Howard’s “three magnets” represent elements of both town and country, but showing a clear preference toward creating human settlements balancing town/country characteristics. More than just a physical plan for design, Howard’s vision included details on a garden city’s government/organizational structure and financing. For example, Howard provided guidelines that a limited dividend company should borrow money and buy land in the relatively cheap countryside and that the increased value brought on by development should help fund the enterprise. He also believed that each garden city should be limited in size – no larger than 1000 acres, or 32,000 in population – and surrounded by a large greenbelt with farms as well as some institutions. Once a garden city would reach its limits, new garden cities should then be created as part of a regional network connected to other garden cities and to the central city by rail transit lines.

Unlike other utopian visionaries, Howard saw his ideas put into practice, although in somewhat compromised form. In England, Letchworth (1900) and Welwyn (after World War I) garden cities were built. In the United States, new deal greenbelt towns, Sunnyside (Queens, New York) and Radburn (New Jersey), are the closest examples.

In the United States, concerns with the rise of the automobile and how “raging streams of traffic” were impacting neighborhoods led Clarence Perry to build on the garden city concept proposing the formation of “neighborhood units” within which schools, local streets, and parks would be located and protected from through traffic in such a way that “children should never be required to cross a main traffic street on the way to school” (see Figure 3).

As only parts of the physical concepts of both the garden city and neighborhood unit were mostly adopted in practice, they served as a foundation for suburban development the world over. In an overwhelming number of cases, these new neighborhoods lacked key benefits of both Howard’s and Perry’s visions for such elements as a balance between jobs, housing, and retail; organized around transit; walkability; and equitable access to all.

City planning’s modernist movement

The Swiss/French architect Le Corbusier (1887–1965) envisioned bold, rational, efficient, and technology based approaches to urban development. Le Corbusier provided powerful images of high-rise residential “towers in the park” and elevated highways serving segregated...
flows of traffic swooping through and out of cities. In the United States and elsewhere, these ideas spurred countless central city urban renewal projects, displacing thousands of people, mostly of low-income and minority groups. Le Corbusier was also a founding member of the Congrès Internationaux d’Architecture Moderne (CIAM), or International Congresses of Modern Architecture (created in 1928 and disbanded in 1959). CIAM espoused its modernist philosophy in its 1933 manifesto, the Charter of Athens.

After World War II, the decentralizing forces of suburban development exploded driven by aspects of both the modernist approach (building grade-separated freeways) and garden city ideals and fueled by technological advances (personal automobiles, telephones). In the United States, urban planning practice specifically facilitated suburbanization through zoning with exclusionary provisions (minimum lot sizes, lack of multifamily housing); the building of the interstate highway system and designing local streets solely around automobile travel (in many cases lacking sidewalks entirely); and US government policy favoring home ownership by insuring home mortgages (Figure 4).

New Urbanism and the rise of a “place” agenda

A number of urban theorists, dissatisfied with the increasing homogeneity and soullessness of mid-twentieth-century urban spaces, in both garden city-inspired suburbs and modernist inner-city urban renewal projects, began questioning current approaches. Early on, Jane Jacobs soundly refuted Le Corbusier’s ideas and those of the modernist movement in her 1961 seminal book, *The Death and Life of Great American Cities* (Jacobs 1961).

In the 1960s, with the rise of a hyper-networked society in view, some thought that the importance of city centers and quality urban places may fall by the wayside. Webber’s 1964 paper “Urban Place and the Non-Place Urban Realm” introduced the idea of “community without propinquity” (which is nearness in space, time, or relationship). He suggested that technological advances in communication and mobility would directly undermine some of the main historic forces pulling people to agglomerate in urban settlement, mainly the utility people gain from exchanges over knowledge, goods, services,
the arts, and so forth. Webber argued strong city-center metropolises would disperse toward suburban clusters, determined primarily by social links and economic networks in a “non-place urban realm.”

While there are findings of dispersion and centralization on both sides, people’s preferences for quality urban places and city centers remain, at least to a certain extent (Florida 2002, 2014). Given this, it is therefore important to study the influence technology, such as GPS-enabled mobile devices (smartphones) and their ability to facilitate people’s “civic operation” and exchanges in and around urban environments, so we can better understand and plan for future human settlement and behavior patterns.

In 1982, Donald Appleyard and Allan B. Jacobs authored *Toward an Urban Design Manifesto* (Appleyard and Jacobs 1982), which took on the Charter of Athens and the garden city concept for their overly strong and unrealizable utopian programs in response to the physical decay and social inequities of the industrial city. Instead they called for an urbanism based upon social objectives and upon how people live and experience cities and space, espousing more human-scaled development. Inspired by this and other writings, urban designers sought ways to
recreate what were felt to be the best physical qualities of traditional neighborhoods and small towns – mixed uses, local shopping, connected street grids, community parks, rear alleys, and front porches.

Initially referred to as “traditional neighborhood development” (TND), New Urbanism, and, later, as “smart growth,” one of the catalytic events in this movement occurred in 1991 when some of the nation’s most influential urban designers and architects gathered in Judy and Michael Corbett’s home in Davis, California, to conceive the Ahwahnee principles which, in 1993, would provide the foundations for the professionally based Congress of New Urbanism (CNU). In 1996, CNU published its charter of New Urbanism which espoused 27 guiding principles organized according to three interrelated spatial scales: metropolis, city, and town; neighborhood, district, and corridor; and block, street, and building. The New Urbanist theorist Peter Calthorpe re-conceptualized the garden city idea in the form of transit oriented developments (TODs), placing prominence on linking urban settlements by public transit networks.

Broader goals for urban planning: social equity, sustainable development, and livability

While social reformers, such as Jane Addams, pioneered social programs to help the poor immigrant populations around the turn of the twentieth century, the urban planning profession focused more on the technical, aesthetic, and political processes to guide things like the orderly use of land and the provision of infrastructure such as streets, water, and sewer. During the latter half of the twentieth century, however, social equity, sustainable development, and livability emerged to represent ideal outcomes of the sum of all planning goals.

Social equity emerged anew in response to discriminatory and unfair housing practices and the displacement of poor, minority communities for urban renewal projects, interstate highways, and other major projects. One of the first to conceptualize “advocacy planning,” Paul Davidoff (1930–1984), argued that a planner could speak for the interests of a group or individuals’ affected plans, but such a planner may have limited influence and not be included in the planning process (Davidoff 1965).

In 1961, the federal government took action in response to problems of the urban poor by passing the 1961 Housing Act which had provisions for housing, public transit, and open space acquisition. In 1964, President Johnson led the passage of four major pieces of legislation for urban populations, including the Civil Rights Act, the Economic Opportunity Act, the Housing Act (which provided $750 million for urban renewal, but also included funds for renovation of existing homes), and the Urban Mass Transportation Act (UMTA) which provided $375 million mass transit.

In response to the urban riots of the mid-1960s, President Johnson convened an advisory commission on civil disorders (Kerner Commission 1968) to study the causes of the riots in order to guide federal policies. The report released on February 29, 1968, made clear and concise findings on how the riots were the result of systematic discriminatory practices against African Americans, denying them critical opportunities and fostering their racial isolation in inner city ghettos.

In light of the riots and Kerner Commission findings, Congress in the Spring of 1968 considered new fair housing legislation, but was unable to make significant progress until April 4, 1968, the day Dr Martin Luther King Jr was assassinated. Since the summer of 1966, Dr King had been strongly in the fight for fair housing.
On April 10, 1968, Title VIII of the Civil Rights Act of 1968 (Fair Housing Act) passed, prohibiting discrimination concerning the sale, rental, and financing of housing based on race, religion, national origin, and sex.

By the mid-twentieth century, air, water, and chemical pollution emerged as major national issues in sustainability and the rise of the environmental movement. One of the most influential books to jumpstart the environmental movement was Rachel Carson’s 1962 *Silent Spring*, which documented the damaging effects of DDT and other pesticides on birds (particularly songbirds) and other species. Up until then, such chemicals were regarded as epitomizing progress through technology. By 1965 DDT was outlawed in the United States. Meanwhile air and water pollution had emerged as serious problems. In 1969, Cleveland’s Cuyahoga River caught fire and Los Angeles appeared to be continually overtaken by air pollution. Following Rachel Carson’s lead, there were increasing studies of the damaging health effects of air and water pollution.

The rapid rise of the environmental movement during this time was an unprecedented political and social paradigm shift in US history marked by the millions who celebrated the first Earth Day on April 22, 1970. By the mid-1970s, most Americans identified themselves as environmentalists. The following are some of the major policy milestones. On January 1, 1970, the National Environmental Policy Act passed, requiring full disclosure and scientific analysis of the environmental impacts of general policies and projects. In his 1970 State of the Union address, Richard Nixon declared clean air and water legislation to be at the center of his domestic agenda, leading to the bipartisan passage of the Clean Air Act (CAA) in 1970 and the Clean Water Act (CWA) in 1972. Nixon also argued for creation of a one-stop environmental policy agency, leading to the establishment of the Environmental Protection Agency on December 2, 1970.

In 1972, the environmental movement entered worldwide prominence at the UN Conference on Human Development in Stockholm where the international community first met to consider global environment and development needs, leading to the creation of UNEP, the UN Environmental Programme. In 1987, the UN created the World Commission on Environment and Development (also known as the Brundtland Commission), which produced the 1989 report, *Our Common Future*. This report presented the UN’s definition of sustainable development with the central theme being that growth “meets the needs of the present without compromising the ability of future generations to meet their own needs.”

In the first decade of the twenty-first century, in order to combat climate change, California passed several climate action initiatives requiring the public agency to lower greenhouse gas (GHG) emissions.

In the 1960s and 1970s, *livability* emerged as an objective to go beyond minimum standards of environmental quality and social equality to include a broader range of characteristics allowing people to achieve higher levels of quality of life. One of the first works to clarify and popularize the concept of livability was the 1981 book *Livable Streets* by Donald Appleyard, which revealed the detrimental impacts automobile traffic had on the quality of life on streets for other users. This work laid the foundation for later work by his son, Bruce Appleyard, who proposed a set of livability ethics which argued that one’s pursuit of livability (such as a driver speeding down the street in a car) should not be allowed to unduly impact the livability of those trying to more humanly experience the street environment, such as walking, talking.
with neighbors, or resting (Appleyard, Ferrell, Carroll, and Taecker 2014).

On June 16, 2009, livability was formalized as federal policy when the secretaries of HUD (US Department of Housing and Urban Development), USDOT (US Department of Transportation), and EPA (US Environmental Protection Agency) introduced the six livability principles of the sustainable communities partnership, which sought to promote transportation choices (walking, bicycling, transit); housing affordability; access to employment, education, and service opportunities; target federal policies and funding within existing urban areas served by transit – also referred to as “smart growth” or transit oriented development (TOD) – and, finally, to coordinate federal policies and funding with all levels of government to plan more effectively for future growth. Many efforts have been made to capture and cross-reference associated metrics and indicators of livability, sustainability, and social equity. One of the most promising initiatives has been the livability calculator which provides a comprehensive understanding of a community’s performance in relation to the six livability principles of the EPA/HUD/USDOT sustainable communities partnership (Figure 5).

**Origins of human dynamics**

The term “human dynamics” finds some of its origins in the study of personality dynamics and statistical physics and refers to a body of work begun in 1979 by Dr Sandra Seagal and her associates on the distinct inner processes in the way people inherently learn, assimilate information, relate, communicate, approach tasks, problem solve, contribute to others, respond to stress and trauma, and maintain health and wellness. The early research into these fundamental distinctions in people emerged as a result of a discovery related to the human voice; namely three frequencies that corresponded to a high, middle, and low frequency. These three frequencies, or capacities of a person, were organized into the following principles – mental (objective), emotional (relational), and physical (practical).

The current concept of “human dynamics” emerges from the convergence of new developments in spatial science, mobile technology, big data, and social behavior research. According to Dr Ming-Hsiang Tsou, professor and founder of the HDMA center at San Diego State University, “Human Dynamics is a transdisciplinary research field focusing on understanding the dynamic patterns, relationships, narratives, changes, and transitions of human activities, behaviors, and communications” (Tsou 2015).

**Disruptive innovation and the sharing economy improving our urban access**

The advances in hand-held devices and their ability to improve people’s ability to navigate...
Figure 6  Geo-locating Twitter feeds related to transit corridors in the Washington DC metro area, based on transit corridor livability typology work by Appleyard, Ferrell, and Taecker (2016). Integrated corridors are green; transitioning corridors are yellow; emerging corridors are red.

Urban environments as well as to engage interpersonal exchanges central to the emerging “sharing economy” – in short, improving our “urban access” – are becoming major forces influencing human settlement and behavior patterns. Such innovations can be viewed as part of the “spatial turn” or the humanities-based quantitative study of “place” as facilitated by the emerging technological advances such as GIS and access to data on mobility patterns and preferences from dynamic mobile devices (see also Figure 6). These new capabilities can also be viewed as a “disruptive innovation,” a term first coined by Bower and Christensen in 1995 and later, in 1997, by Christensen. A disruptive innovation is an innovation that helps create a new market and value network that eventually disrupts an existing market and value network (over a few years or decades). Such innovations improve a product or service in ways that the market does not expect, typically, first, by designing for a different set of consumers in a new market and later by lowering prices in the existing market. And since these new technologies create and/or serve emerging market segments that are not being served by existing stewardship systems, the field of urban planning and human dynamics (UPHD) seeks to also help
decision-makers and institutions understand how to appropriately integrate these new capabilities into current practices.

SEE ALSO: Urban climatology; Urban ecology; Urban elites; Urban geography; Urban managerialism; Urban mosaic; Urban redevelopment; Urban renewal; Urban transit

References


Further reading


Urban political ecology

Nik Heynen
University of Georgia, USA

Theoretical foundations of urban political ecology

Urban political ecology (UPE) has a distinctly Marxist origin. The analysis of urban form and social process within UPE is an effort to better incorporate the ways capitalism not only produces cities but also produces nature in cities at the same time. The Marxist geographic logic that led to the development of UPE can be clearly connected to the work of David Harvey, Neil Smith, Henri Lefebvre, and Manuel Castells, as well as other key scholars. While the explicit attention to urban form is what is most recognized in discussions of UPE, wrestling through the centuries-old struggles to reunite the ideas of nature and society is just as central to UPE and especially important for recognizing the early Marxist influence. This rings clear when Smith (2006, xii) said:

Enlightenment thinkers from Newton to Kant, Adam Smith to Montesquieu – and many others – answered the demand to understand the externality of nature. This is neither to diminish their contributions necessarily, nor to cast aspersions on their accomplishments, but it is to contextualize the ideological power that their contributions came to have. In many ways, the stunning question today – still almost unaskable – is not how to reconcile nature and society, how to understand the “interaction” of nature and society – the agenda of most in the environmental movement – but how western ideologies could have got to the point of flattering themselves so successfully that they were somehow separate from nature.

As the rapidly growing literature demonstrates, a central contribution of UPE asks this very question: How do we pull back the Western mindset regarding the separation from nature in its urban form? UPE scholars have sought to do this from its earliest efforts through urbanizing discussions about “metabolism.” These scholars have also pushed toward a deeper understanding of how the metaphorical rendering of metabolism offers a way to cut through the dualisms that have historically been problematic for the discussions of nature and society. However, UPE scholars have also taken seriously the egalitarian potential that is embedded within a robust conceptualization of urban metabolism. This is another distinguishing characteristic of this literature. In seeking to roll back the historically imposed separation of urban nature and urban society, many working within UPE have sought to replace it with an ideology of urban nature’s revolutionary potential.

The idea of metabolism goes back to Marx within socio-natural theory. This idea of creativity continues to be expressed through interrelated and interconnected socio-natural urban processes. Their resulting uneven configurations are a main contribution of UPE. UPE scholars have been working to articulate urban metabolism as a dynamic process. Central to this is showing how new sociospatial formations, entwinings of materials, and collaborative enmeshing of socio-nature emerge through human labor and nonhuman processes simultaneously.
Too narrow a focus on “ecological” processes within urban ecology opened up a critique that helped facilitate the development of UPE. Much of the urban ecological approach excludes the role of uneven social power relations. These critiques opened pathways for thinking about the urban interactions between political economy and nature using a more holistic and dynamic approach. To this end, David Harvey notably suggested:

It is inconsistent to hold that everything in the world relates to everything else, as ecologists tend to, and then decide that the built environment and the urban structures that go into it are somehow outside of both theoretical and practical consideration. The effect has been to evade integrating understandings of the urbanizing process into environmental-ecological analysis. (1996, 427)

However, a study originating in urban ecology necessitates that primacy be given to urban space, form, and process. Alternatively, ecological processes, while implicit in much of urban cultural research, often tend to simply play the role of backdrop for other spatial and social processes.

**Becoming urban political ecology**

While in the past there have been ideas, logics, and empirical research that sounded like what we call UPE today, it was not until 1996 that Erik Swyngedouw named “urban political ecology” as such. Swyngedouw explicitly articulated the significance of urban studies traditions up to that point, while simultaneously highlighting their intersection in the theoretical strands of political economy, political ecology, and science and technology studies (STS). This is an important moment in thinking about the city, as Swyngedouw discussed uneven development in a synthetic language bringing together “representational,” “discursive,” “ideological,” “material,” and “biochemical” constellations of power relations through the notion of urban metabolism. If there is a founding passage for what we call UPE it was when Swyngedouw suggested:

In the city, society and nature, representation and being are inseparable, integral to each other, infinitely bound-up, yet simultaneously this hybrid socio-natural “thing” called the city is full of contradictions, tensions and conflicts … Only over the past few years, a rapprochement has begun to assert itself between ecological thinking, political-economy, urban studies and critical social and cultural theory. This may provide the ferment from which a new and richer urban ecology or urban political-ecology may germinate. (1996, 65–66)

As engaging urban nature seemed a logical continuation of Marxist approaches to the city, it made sense to many urban theorists interested in the environment when Keil (2003) suggested that urban political ecology is the appropriate response to Lefebvre’s call for creating an “urban science for an urban world” (2003, 23). Here, there started to be a more deliberate focus on the historical lack of attention to how social power relations contributed to uneven development by focusing on how political and cultural economy shape and are shaped by urban environments (see Swyngedouw and Heynen 2003). Access to water, food, nonpolluted air, and green space within uneven matrices of class, race, gender, age, and physical ability would all pose the sort of examples urban political ecological investigations started to engage.

Not seeing political ecology in the city was a gap not just for those working in urban ecology but also for those who have worked in the much more historically rooted tradition of political ecology. Indeed, many engaged in political ecology have not acknowledged the impact of cities
within their framing of political ecology despite the now common tropes of the twenty-first century connecting “the urban century” with “the Anthropocene.” Like the blind spots within urban ecology, the blind spots within political ecology, according to Braun (2005), are simultaneously related to both the long tradition of social scientists not being comfortable with thinking about cities as part of nature and to the heavy attention within political ecology, at least historically, on rural “third world” forms of nature.

During the 2000s, urbanists took up this challenge and produced a series of important books and articles that collectively constitute some of the most influential foundations of UPE. In these texts can be seen the articulation of important thoughts about urban nature that go beyond much of the urban ecological discussion that came before it. These texts started to create more theoretical space and coherence toward an expanded way of thinking about nature in cities. These ideas helped form what today are the taken-for-granted (if not always agreed-to) principles of UPE.

Cronon’s (1991) *Nature’s Metropolis: Chicago and the Great West* paved the way for the development of UPE in important ways. Cronon did this by illustrating early on how the production of socioenvironmental changes between the city and its hinterland always results in the continuous production of new forms, dynamics, and interconnections within urban nature. He illustrates, in rich and nuanced historical context, how the interlocking spatial relations of unfolding urban social and physical environmental conditions develop and are always part of ongoing urban–rural systems. Cronon also helped to situate the role of political power within these socio–natural dynamics within Chicago. By bringing class, ethnicity, race, and gender into the articulation of Chicago’s environmental history, Cronon helped to integrate different ways of seeing the power struggles through which varied individuals and groups of individuals shape their local environments.

Another important book that prefigures the development of UPE is Matthew Gandy’s (2002) *Concrete and Clay: Reworking Nature in New York City*. Just as Cronon focused on Chicago’s socio–natural history, Gandy thinks through the ways in which New York City offers examples for seeing social struggles over urban nature transform both those involved and the city itself. More precisely, he argues that “Nature has a social and cultural history that has enriched countless dimensions of the urban experience. The design, use, and meaning of urban space involve the transformation of nature into a new synthesis” (Gandy 2002, 2). As Gandy helps to illustrate, attention to the processes that give way to uneven development, and thus uneven nature, is necessary to the work of compressively thinking through urban nature. These struggles, through this sort of textured analysis, led UPE away from apolitical understandings of urban ecology.

Kaika’s (2005) *City of Flows* offers some important departures from both Cronon and Gandy as another exemplar book often thought of as foundational to the development of UPE. Cronon and Gandy approached their respective North American cities from diverse processes of socio–natural interaction, ranging between processes as diverse as agriculture, lumber, and transportation in the case of Cronon, and water, landscape, poverty and pollution for Gandy. Kaika, however, focuses her analysis on the relationship between modernity and water as a way of understanding the urbanization of nature and UPE. Struggles over water access and infrastructure quickly became one of the most concentrated areas of inquiry as UPE continued to evolve. Much of Kaika’s empirical effort focused on “decipher[ing] the historical geographical process through which modernity
discursively constructed the modern city and the modern home as autonomous ‘space envelopes’ independent from natural and social processes” (2005, 4). For Kaika, as well as many others working in this tradition, much attention has gone beyond simply trying to articulate that the nature/society dualism is intellectually problematic, but rather they demonstrate the political implications of the Enlightenment era ways of seeing the world and the city.

Robbins’s (2007) Lawn People: How Grasses, Weeds, and Chemicals Make Us Who We Are offers up yet another innovation for the early formation of UPE through his theoretical and empirical attention to the role of ideas in shaping urban nature. For Robbins, “apolitical ecology” has too long dominated understandings of the decisions and behaviors central to shaping urban nature. Instead, through a grounded study of the prominence of the lawn in US cities, Robbins works against the traditional apolitical approach within urban ecology. His “apolitical approach” includes:

(1) a focus on free individual choices, (2) a propensity to assign culture a driving role in understanding group behavior, (3) a predisposition to think of economic activities and the behavior of firms as meeting consumer demand, and (4) an inclination to think about human actions, whether those of individuals or companies, as sovereign relative to the influence of nonhuman actors, objects, and animals. (Robbins 2007, 4).

As such, his book helps demonstrate “the mutual tyrannies of urban political ecology” (2007, 13) by showing how in urban nature, “each of the separate pieces is not independent, but instead made to be the way it is by virtue of its relationships to all the other parts” (2007, 14).

If these books, and others, served as the empirical and theoretical foundations of what has become UPE, there were also key articles that have helped name UPE. Early on, Roger Keil (2003) started to do the work of sorting and situating the internal differences of what was clearly becoming a subdisciplinary domain of geography’s much larger tradition of nature–society research. In an important review essay in the journal Urban Geography, Keil was one of the first to say:

There is now a rich body of literature with much diversity of origin and some diversification across various academic and philosophical traditions. In contrast to the more policy-oriented and problem-solving work on urban sustainability … the UPE literature is characterized by its intensely critical predisposition; critical is defined here as the linking of specific analysis of urban environmental problems to larger socioecological solutions. This necessitates, at a minimum, some modicum of indebtedness to radical and critical social theory. It is no coincidence then, that the emerging field of UPE has many of its multiple roots in the intellectual traditions of fundamental social critique: eco-Marxism, eco-feminism, eco-anarchism, etc. It is also, however, indebted to a neo-pluralist and radical democratic politics that includes the liberation of the societal relationships with nature in the general project of the liberation of humanity. (2003, 724).

Similarly to Keil’s authoritative first progress report, Swyngedouw and Heynen’s (2003) introductory essay in a special issue of the journal Antipode on UPE also worked to knit together the research and insight that was by that time increasingly being referred to as UPE, even as some authors were not actively using that language. In this essay, they suggest:

Urban political ecology provides an integrated and relational approach that helps untangle the interconnected economic, political, social and ecological processes that together go to form highly uneven urban landscapes. Because the power-laden socioecological relations that
go into the formation of urban environments constantly shift between groups of actors and scales, historical-geographical insights into these ever-changing urban configurations are necessary for the sake of considering the future evolution of urban environments. (Swyngedouw and Heynen 2003, 914).

In this language we see another defining feature of much UPE, which like Robbins, insists on a logic that seeks to “untangle the interconnected.” Untangling the interconnected also related to the discussion in another early and important essay related to UPE published in Progress in Human Geography by Braun. In it Braun suggests:

This concern with contextualization, as well as its focus on the processes and practices that produce uneven and spatially differentiated environments also differentiates UPE from the large environmental justice (EJ) literature, which has been dominated by a liberal discourse of “distributional justice” rather than a radical analysis of capitalist urbanization. (2005, 644).

Situating the difference between UPE and EJ, in the earliest phases of UPE development was important in terms of showing the trajectories of different kinds of intellectual and political projects. It is worth saying, however, that while there are some differences between UPE and EJ there are also productive similarities. In part this is because, as UPE continues to be defined through empirically grounded studies, its coherence continues to open up toward productive partnerships with other like-minded scholarship. Braun foreshadowed this to some degree in his article when he said:

The rapid growth of UPE is likely to continue, with its definition and scope expanding accordingly. This is true in part because political ecology does not name a single, coherent field, or seek its own distinctive theory. This renders it open-ended, perhaps for the good. To the extent that urban political ecology interrogates the complex and interrelated socioecological processes that shape cities and their uneven urban environments, there is no reason to exclude Mike Davis’s … splendid writings on Los Angeles, or David Harvey’s … reflections on nature, justice and difference, even if neither uses the term. (Braun 2005, 644).

One of the most referenced collections within UPE is an edited book by Heynen, Kaika and Swyngedouw (2006) entitled In the Nature of Cities: Urban Political Ecology and the Politics of Urban Metabolism. One reason for the attention to this collection has both to do with the timing of its publication at a moment when UPE was maturing. Another reason the collection has attracted attention relates to the rich melding of theoretical and empirical research conducted by many of the scholars who supported the need to think about the politics of urban nature in more comprehensive and interconnected ways. The timing meant that particular statements in the book, backed up by robust collaboration across empirical terrains, both helped foster the ongoing development of UPE and also gave a name to something that could be critiqued and expanded on. A particular passage from the introduction of the book, like other defining passages in other early works, concisely sums up what UPE is about:

An urban political ecological perspective permits new insights in the urban problematic and opens new avenues for re-centring the urban as the pivotal terrain for eco-political action. To the extent that emancipatory urban politics reside in acquiring the power to produce urban environments in line with the aspirations, needs, and desires of those inhabiting these spaces, the capacity to produce the physical and social environment in which one dwells, the question of whose nature is or becomes urbanized must
be at the forefront of any radical political action. (Heynen, Kaika, and Swyngedouw 2006, 15).

While there has now been considerable effort to develop UPE, Loftus’s (2012) Everyday Environmentalism: Creating an Urban Political Ecology is one of the first monographs that explicitly engages UPE. Loftus develops both the history and the scope of UPE by grounding his empirical project in South African hydro-struggles. The book pushes the evolution of UPE in useful and provocative ways. The range of urban environmental processes used to organize the project is important for showing the possibilities for producing a more democratic urbanized planet. Loftus utilizes qualitative research on South African water politics in an effort to show how grassroots struggles within informal settlements of Durban help to embody UPE. His empirical insights invigorate the politics of possibility within UPE as deliberately framed. The “everyday” desires, possibilities, emancipatory potentials, and dangers that swirl within this narrative push an explicit understanding about the articulation of “urban” and “political” and “ecology.” Loftus’s research expands the reach of UPE by showing, through a blend of theory and grounded politics, who benefits and who suffers through the particularities of the production of urban nature.

As the confluence of nature, political economy, and the city continue to evolve we can see more distinction and diffusion of the research done within UPE. The earliest and most significant research within UPE related to urban hydroscapes and how uneven socio-natural power relations configured these landscapes differently. The foundations provided in this work help open up other ways of articulating nature in the city through a wide spectrum of more diverse dynamics (e.g., air pollution, urban forests, wildlife corridors, gentrification, insects and pesticides, gardens and food, hunger, health). Much of this work responds to something that Smith said:

A dialectic is at work here. As part of a broader political movement, an urban political ecology can help integrate a politics of nature into a more established “social” politics; at the same time an ecologically enhanced politics focusing on the productive metabolism of nature can further exoticize the absurd separation of nature and society while denying any anti-social universalism of nature. (2006, xv).

Provoked by both critique and support, efforts continue to better connect UPE with political ecology more generally, as well as literature on environmental and food justice and other important threads of nature-society scholarship. In an article expanding on the UPE of food justice Agyeman and McEntee (2014, 211) suggest:

Our central proposition is that urban political ecology (UPE) offers up a broader theoretical frame for advancing and retaining the potency of the field of Food Justice while also allowing the movement to progress in alleviating the problems of hunger and inadequate food access.

This move toward great expansion and inclusion of more and more urban dynamics helps to illustrate that UPE is likely to continue to develop into the future. Agyeman and McEntee go on to say:

UPE’s emphasis on hybridity, scale, and commodity relations has the potential to ensure that FJ [food justice] remains true to focusing on outcomes and processes as well as symptoms (e.g., immediate needs such as inadequate access to food) and causes (e.g., structural inequalities) of food injustice within the current neoliberal system. (2014, 211).
Again, seeing the application and expansion of the dialectical logics that led to the formation of UPE in the first place is another productive signal regarding the growth of this way of thinking about urban nature.

**Ten key ideas in urban political ecology**

Despite the theoretical differences that are increasingly defining new work in UPE, and given the encyclopedic purposes of this current entry, it would be useful to recount 10 core principles that increasingly characterize much of UPE scholarship which Swyngedouw developed after his 1996 *Capitalism Nature Socialism* essay.

1. **Environmental and social changes codetermine each other.** Processes of socioenvironmental metabolic circulation transform both social and physical environments and produce social and physical milieus (such as cities) with new and distinct qualities. In other words, environments are combined sociophysical constructions that are actively and historically produced, in terms of both social content and physical-environmental qualities. Whether we consider the making of urban parks, urban natural reserves, or skyscrapers, each of these contains and expresses fused sociophysical processes that contain and embody particular metabolic and social relations.

2. **There is, therefore, nothing a priori unnatural about produced environments like cities, genetically modified organisms, dammed rivers, or irrigated fields.** Produced environments are specific historical results of socioenvironmental processes. The urban world is a cyborg world, part natural, part social, part technical, part cultural, but with no clear boundaries, centers, or margins.

3. **The type and character of physical and environmental change, and the resulting environmental conditions are not independent of the specific historical social, cultural, political, or economic conditions and institutions that accompany them.** It is concrete historical-geographical analysis of the production of urban natures that provides insights into the uneven power relations through which urban “natures” become produced, and that provides pointers for the transformation of these power relations.

4. **All sociospatial processes are invariably also predicated on the circulation and metabolism of physical, chemical, or biological components.** Nonhuman “actants” play an active role in the mobilizing of socio-natural circulatory and metabolic processes. It is these circulatory conduits that link often distant places and ecosystems together and permit relating local processes with wider sociometabolic flows, networks, configurations, and dynamics.

5. **These metabolisms produce a series of both enabling and disabling social and environmental conditions.** Indeed, these produced milieus often embody contradictory tendencies. While environmental (both social and physical) qualities may be enhanced in some places and for some people, they often lead to a deterioration of social and physical conditions and qualities elsewhere.

6. **Processes of metabolic change are, therefore, never socially or ecologically neutral.** This results in conditions under which particular trajectories of socioenvironmental change undermine the stability or coherence of some social groups or places, while the sustainability of social groups and places elsewhere might be enhanced. In sum, the political-ecological examination of the urbanization process reveals
the inherently contradictory nature of the process of metabolic circulatory change and teases out the inevitable conflicts (or the displacements thereof) that infuse socioenvironmental change.

7 Particular attention, therefore, is paid to social power relations (whether material or discursive, economic, political, and/or cultural) through which metabolic circulatory processes take place. It is these power geometries and the social actors carrying them (out?) that ultimately decide who will have access to or control over resources or other components of the environment, and who or what will be positively or negatively enrolled in such metabolic imbroglios. These power geometries, in turn, shape the particular social and political configurations and the environments in which we live. Lefebvre’s “right to the city” also invariably implies a “right to metabolism.”

8 Questions of socioenvironmental sustainability hereby become fundamentally political questions. Political ecology attempts to tease out who gains from and who pays for, who benefits from and who suffers (and in what ways), from particular processes of metabolic circulatory change. It also seeks answers to questions about what or who needs to be sustained and how this can be maintained or achieved.

9 Political-ecological perspectives seek to unravel the nature of the social relationships that unfold between individuals and social groups and how these, in turn, are mediated by and structured through processes of ecological change. In other words, environmental transformation is not independent of class, gender, ethnic, or other power struggles.

10 It also seeks to question the actual processes of metabolic reconstruction and recasting, and advocates a position on sustainability that is achieved by means of a democratically controlled and organized process of socioenvironmental (re)construction. The political program, then, of political ecology is to enhance the democratic content of socioenvironmental construction by means of identifying the strategies through which a more equitable distribution of social power and a more inclusive mode of the production of nature can be achieved.

The ideas contained within these principles continue to inform efforts to better connect and better embody the ways in which urban nature is struggled over and reproduced in the image of those struggles. While there are many important ways that the false dualism between nature and society has been addressed, UPE offers a way of addressing this issue with specific attention to urban socio-natural form and spatial processes.

SEE ALSO: Environmental (in)justice; Marxist geography; Political ecology; Urban geography

References


Urban politics are complicated. Urban politics is often treated as a distinctive form of politics and “the urban” itself is often taken as a self-evident concept. Neither is quite so straightforward.

The contemporary notion of urban politics has its roots in Manuel Castells’s work on “the urban question” in the early 1970s. In Castells’s structuralist formulation, capitalist society is expressed through the spatial forms of its economic, political, and ideological systems. These systems take material, spatial form in association with the production activities of factories, mines, offices, and so on; the consumption activities of housing, stores, recreation, and so on; and the circulation and exchange activities of banks, roads, telecommunications, and so on. The spatial structuring of production, reproduction, and circulation and exchange has “return effects” on society, giving rise to localized urban problems, urban planning conflicts, and ultimately urban politics, including urban social movements, which address collective consumption issues. Urban politics, in other words, has its roots in the spatial expression of a capitalist society’s structural contradictions. Urban politics are not merely epiphenomenal, however, as they influence the actions of the state, which intervenes in cities and attempts to ameliorate conflicts around collective consumption.

With the publication of his classic book, The City and the Grassroots (1983), Castells signaled a major theoretical shift in his approach to urban politics. While he continued to highlight conflicts around collective consumption as the defining concern of urban politics, Castells dropped the structural focus, instead emphasizing the concerns and conflicts of everyday life. For Castells, the city was the basis of the territorial organization of urban politics (and especially urban social movements), focusing on three dynamics: (i) conflicts over the city as a space of use-values versus a space of exchange-values; (ii) conflicts over neighborhood cultural identity, neighborhood cultural autonomy, and communication; and (iii) conflicts over territorial or neighborhood self-management. Based in everyday experience at the city or neighborhood scale, these collective-consumption-based conflicts were largely defensive in nature. Urban politics were about the defense of localized lifeworlds from commodification and state control, and were to be viewed as a very specific type of territorially based politics, both reactive and inward-looking.

In the same era that Castells and other urbanists were analyzing the urban politics of collective consumption, global changes in state regulation and economic and class structure were underway. The Fordist regime of accumulation gave way to post-Fordism or, even more chaotically, after-Fordism. Post-World War II processes of economic restructuring had accelerated in the 1960s, and became widely felt by the 1970s when declining profits, rising unemployment, and declining state revenue led to what James O’Connor (1973) called “the fiscal crisis of the state” and Jürgen Habermas (1975) labeled a “legitimation crisis.” Cities were forced to cut funding for services, maintenance, and infrastructure investment, bringing many of
them to the brink of bankruptcy and under-mining support for the welfare state. The urban politics of collective consumption gave way to a very different outward-looking type of politics oriented to the attraction of new capital investment and tax base enhancement. An early influential analysis of this shift was Harvey Molotch’s classic 1976 article, “The City as a Growth Machine.” To Molotch, research based on the traditional definitions of what an urban place is has had very little relevance to the actual, day-to-day activities of those at the top of local power structure[s] whose priorities set the limits within which decisions affecting land use, the public budget, and urban social life come to be made. (1976, 309)

Molotch argued that those at the top of local power structures were increasingly promoting growth rather than satisfying collective consumption demands, and that growth was being pursued at the expense of competing localities. Beggar-thy-neighbor strategies were driven, first and foremost, by landed elites who stood to gain substantially from increasing property values and rents. Landed elites frequently led growth coalitions that included other business interests, which also benefited from local economic expansion, such as local newspapers. While Molotch acknowledged the contested nature of growth politics and pointed to countercoalitions challenging the ideology of growth, growth coalitions became the norm while retrenchment of the welfare state and its collective consumption functions deepened. These transformations were part and parcel of the slow creep of neoliberal urban policy which entailed rolling back state functions, responsibilizing the individual, and demanding lower taxation to facilitate continued private consumption in the face of declining incomes. A critical milestone in this trend was the passage of the populist “People’s Initiative to Limit Property Taxation” referendum, commonly known as Proposition 13, in California in 1978. Proposition 13 signaled the turn away from an urban politics of collective consumption to an urban politics of low taxation and economic growth.

By the early 1980s, right-wing national governments pursuing neoliberal policy agendas had been elected in many countries in both the Global North and the Global South. Deregulation, the “freeing” of markets, downloading central state functions to local states, and the pursuit of economic growth became hegemonic. Economist Paul Peterson’s book, City Limits (1981), argued that cities were now shaped by external structural forces and faced the necessity of adopting policies to attract capital investment and highly skilled labor. Urban governments had little scope to address issues of collective consumption and redistribution. In this economic-deterministic view, politics had essentially become irrelevant since all elected city officials, regardless of political allegiance, faced the same powerful structures. Harvey’s (1989) analysis addressed the sweeping social consequences of this transformation: accelerated gentrification, increased investment in facilities that enhance interurban competition, the rise of place marketing, the shift to consumerist spectacles for the wealthy, increasing income inequality, and the rise of public–private partnerships.

Yet, while much analysis focused on economic development strategies and politics of cities, it was clear that urban politics and policy now needed to be understood in a broader context. Cities did not adopt policies in a vacuum and the road to urban entrepreneurialism was riddled with contradictions and conflict. Drawing from French regulation theory, many urbanists came to a broader understanding of the post-Fordist crisis as both ideological and inherently geographical. Drawing from Gramsci, regulation theory posits hegemonic ideologies underpinning regimes of
accumulation and modes of regulation, allowing accumulation to proceed on the basis of provisional “grand compromises.” The Keynesian welfare state, relatively secure employment, and rising wages underpinned the Fordist grand compromise. Post-Fordism has yet to arrive at a grand compromise and it remains unclear whether the ideology of “free markets” and individual responsibilization will be sufficient to hold the post-Fordist regime of accumulation together. What is clear, however, is that the shift from Fordism to post-Fordism has entailed a massive geographical restructuring of regulatory arrangements. Under Fordism national forms of regulation were dominant, with the national state promoting the general welfare. Under the regulatory logic of post-Fordism the nation-state has been hollowed out – programs and funding cut – and many of its functions rescaled. In some cases regulatory functions have been shifted to supranational institutions, but in many other cases functions and regulatory responsibilities have been downloaded to local states, which were often lacking the fiscal capacity to carry out these new responsibilities (Peck and Tickell 1994). It is in this context that the shift to urban entrepreneurialism and growth machine politics must be understood.

It should be emphasized that the new scalar set of regulations was a necessary precondition to creating a relatively autonomous urban growth politics. Urban entrepreneurialism did not arise in the same way everywhere, and in many parts of the world it was relatively muted, raising the question of whether the turn to urban entrepreneurialism had its roots in anything that could be called “urban.” Whether there is such a thing as a specifically urban process and, in turn, specifically urban politics, was the topic of a lively debate in the 1980s between David Harvey and Peter Saunders. Recounting this debate, Kevin Cox (2001, 760) argues that, while politics is necessarily spatial, “there does not have to be ‘a relatively autonomous urban politics.’” Instead, Cox stresses that the scalar structure of the state sets the preconditions for the politics of urban development. In highly decentralized federal states like the United States, city is pitted against city in the quest to attract capital investment, making possible an urban politics of interurban competition; however, strong central states that directly fund municipal functions may mute or even preclude interurban competition and its associated politics.

While the work of Cox, Harvey, and others demonstrates the importance of the state’s scalar structure in shaping urban politics, the question remains of how and why certain actors become politically active. Harvey’s work on the mobility of capital and labor is an important starting point for understanding how and why particular actors are likely to become involved in urban development politics. Building from Harvey’s work, Cox (1993) develops a particularly powerful and nuanced analysis of what he calls the “new urban politics” (NUP). Cox proceeds by questioning sweeping claims for the hypermobility of capital and the immobility of labor, and points out that different fractions of capital and labor differ in their mobility. Some fractions of capital (and labor) may be highly mobile, allowing them to address conditions not to their liking by relocating to more favorable locations. Other fractions of capital and labor, however, may be locally dependent, and therefore their ability to relocate is limited, which leads them to engage in struggles in the local political arena. This formulation calls into question Molotch’s original thesis that growth politics is led by landed elites. In Cox’s analysis, growth politics tends to involve actors whose interests are highly locally dependent – not only landed elites – and to give rise to a variety of cross-class coalitions rooted in local dependence. Political actors
are constantly “negotiating the contradiction between fixity and mobility” (Cox 2001, 761), and this negotiation can take place at a variety of scales, not only the urban.

The epistemological question of what “the urban” is and how it is constituted has been raised by numerous scholars. As early as 1970, Henri Lefebvre asserted that the issues that cities face can no longer be analyzed through the lens of the bounded, independent city. Instead, the commodification and bureaucratization of urban life – what Lefebvre termed the conflict between abstract space and social space – has become a global problem and, indeed, urbanization processes have themselves become global. These processes, Lefebvre (2003/1970) argued, need to be understood relationally. More recently, scholars such as Neil Brenner and Christian Schmid have advanced similar arguments. Brenner and Schmid assert that three macro-trends undermine the capacity of areally delimited models to adequately capture the processes that shape “the urban”: (i) “new geographies of uneven spatial development mutually produce one another at all spatial scales, from the neighborhood to the planetary” (2015, 151–152); (ii) the “basic nature of urban realities … has become more differentiated, polymorphic, variegated and multi-scalar than in previous cycles of capitalist urbanization” (152); and (iii) “the regulatory geographies of capitalist urbanization have likewise been undergoing profound, rapid mutations” (153).

Combined, these three trends call into question any analysis based on a single areal unit such as a city. Brenner and Schmid’s point, much like Lefebvre’s, is that we have to look beyond the city to understand the processes, including political processes, that shape cities. Indeed, many urban political activists understand this point. As Cox (1998) has argued, there is a scale division of politics and the problems of cities are not always best addressed at the scale of the city. In some cases, urban issues may be addressed at the scale of the neighborhood, and in other cases at the regional or even the national scale. In other words, the scales at which urban political issues might be engaged are not necessarily the scales of the social relationships of the city.

Nonetheless, while the spaces of urban political engagement may extend well beyond the city, the conditions that give rise to collective political mobilization are most commonly found in cities. As spaces of diversity, density, and frequent encounter, cities are “spaces of politicization” (Miller and Nicholls 2013). Experiences of oppression, the formations of social bonds and social capital, and the working out of shared identities, common understandings, and political strategies most commonly occur in cities (Martin 2003). Patterns of interaction and encounter that foster politicization are structured by the characteristics of urban places. As John Agnew (1987, 28) explains, three dimensions of place shape the nature of politics: through locale, the settings in which social relations are constituted; through location, the area encompassing the settings of social interaction as they are shaped by social and economic processes operating at wider scales; and through sense of place, the local “structure of feeling.” Urban places structure and facilitate (and, in some cases, hinder) social contact, social bonding, alliance building, and resource mobilization. It is from these places that political initiatives to address the processes shaping urban life are launched, whether the specific issue at hand is highly localized, such as a neighborhood conflict over a development proposal; city-wide, such as protests against city-wide policing practices; or global in scope, such as protests against growing income inequality and austerity policies, as exemplified by the Occupy movement.

An understanding of urban politics must start from an understanding of urban social life and its complex geographical constitution (Nicholls,
Miller, and Beaumont 2013). The processes of politicization are not necessarily geographically congruent with the processes that structure cities and the problems urban residents face. Urban politics grows out of the social relationships, social bonds, and experiences of urban residents; they decide which grievances to politicize and where to direct those grievances. While the imperative of growth in a capitalist economy and the geographical structure of the state always figure prominently in the politics of capitalist societies, they do not define the totality of urban social life. Urban politics is ultimately grounded in the diversity of the city but cannot be reduced to a particular issue or areal unit such as the city.

SEE ALSO: Civil society; Class; Community; Cultural politics; Difference; Environmental movements and protest; Globalization; Inequality; Place; Political geography; Regional political movements; Scale; Social capital; Social movements; Spatial social networks; State, the

References

Urban redevelopment

Martine August  
*Rutgers University, USA*

Alan Walks  
*University of Toronto, Canada*

Urban redevelopment is the process of re-establishing the profitability of devalorized and declining urban districts, using a mix of residential, commercial, office, or entertainment-related land uses. Often called “urban regeneration” or “urban renewal,” and encompassing infill, brownfield, and intensification projects, urban redevelopment is carried out by public or private actors (or through public–private partnerships). The process typically involves the remaking of urban space for more affluent users, both physically (through demolition and reconstruction) and symbolically (with the discursive renaming and reimaging of place). Urban redevelopment is often promoted as a way to reverse economic decline, attract new investment, and stimulate middle-class vibrancy. It is also likely to promote unjust social outcomes – through displacement – that are disproportionately experienced by low-income and racially marginalized communities.

Theoretically, urban redevelopment can be understood as one aspect of physical and social change in the broader play of capitalist processes of urbanization. Typically targeting central urban areas, redevelopment is driven by the same set of forces that produce deindustrialization, inner suburban decline, and peripheral suburban expansion. According to Gotham (2001a, 2), these processes are “part and parcel of the continuous reshaping of the built environment to create a more efficient arena for profit making.” Harvey (1982) describes such reshaping as “uneven geographical development” resulting from the contradictory dynamics guiding the production of space under capitalism. These dynamics can be understood as a “see-saw of uneven development” (Smith 1984), in which capitalists invest where profit margins are high, only to experience diminishing returns as competitive pressures erode the local advantages of place. At this point, capital seesaws away from disinvested sites to new areas with hot property, only to seesaw back again at some future time when the first location has once again become profitable. In this theory, disinvestment and decline are directly related to redevelopment, playing a crucial role in creating areas that are ripe for reinvestment. As areas decline, according to Smith (1984), a “rent gap” emerges between the profits that can be reaped from a currently devalorized land use and the potential profits that could be gained from a “higher and better” redeveloped land use. Once the rent gap widens to a certain point, redevelopment and gentrification become possible.

Urban redevelopment can also be understood as a part of the capitalist process of creative destruction. Harvey (1978) explains how capitalism creates a built environment “in its own image” to facilitate accumulation. As technology improves and the needs of capital change, however, this landscape of roads, buildings, and infrastructures begins to present a barrier to continued accumulation. In order to create a
built environment suited to present conditions, it must destroy the physical assets that exist. As Harvey (1978, 83) put it:

under capitalism there is, then, a perpetual struggle in which capital builds a physical landscape appropriate to its own condition at a particular moment in time, only to have to destroy it, usually in the course of a crisis, at a subsequent point in time.

Urban redevelopment targeting “outmoded” housing, factories, port facilities, shopping outlets, transportation networks, and the like destroys these relics from capitalism’s past to create new built environments for future accumulation. While the “creative” aspect of urban redevelopment is often celebrated, critical scholars focus on its destructive impacts and the social dislocations experienced by low-income and marginalized groups (including racialized communities, women, seniors, people with disabilities, and LGBTQ communities) who are displaced by processes of capitalist reinvestment. Indeed, in a number of developed nations, urban redevelopment is fundamentally ordered by racial inequalities, which it reinscribes in urban landscapes (Gotham 2001b). Rather than meaningfully addressing the impacts of segregation, inequality, poor housing quality, and disinvested infrastructure in poor communities, urban redevelopment often works instead to shift residents of these communities elsewhere. However, it is also possible for socially just forms of urban redevelopment to occur.

Urban redevelopment discourses

Urban redevelopment often leads to gentrification, although the latter term is rarely openly used by public or private actors promoting urban redevelopment. Instead, a range of other linguistic terms work to justify and promote projects that might otherwise provoke political resistance. Terms such as “revitalization,” “rejuvenation,” “upgrading,” “renovation,” or “renewal” carry more positive connotations among the lay public, implying that urban neighborhoods with aging infrastructure and housing, which are often occupied by poor and racially marginalized people, are not “vital” and are in need of improvement. This rhetoric often masks the violent and unjust processes of residential displacement and dispossession that redevelopment entails. When discussing urban redevelopment, critical scholars argue that language is important, and insist that we hold on to the word “gentrification” when appropriate in order to politicize the process and foreground the realities of associated class- and race-based inequalities.

Early urban redevelopment: Engels and Haussmann

Patterns mirroring contemporary urban redevelopment were already being described and critiqued in the nineteenth-century city, notably in Friedrich Engels’s pamphlet The Housing Question (1988/1872). Industrial cities, Engels noted, tended to house their working classes in substandard, cramped, and overpriced tenements near centrally located employment. With time and neglect, these buildings would deteriorate, depress land values, and be demolished to make way for new structures. According to Engels, “They are pulled down, and in their stead shops, warehouses, and public buildings are erected,” in the process enriching landlords and displacing poor tenants. In London, Manchester, Liverpool, Berlin, and Vienna—everywhere—the effect was the same: “the result is that workers are forced out of the centre of towns towards the outskirts; that workers’ dwellings … become rare and expensive and often altogether unattainable.”
(1988/1872, 319). “Slum clearance” in this era offered landlords the opportunity to capitalize on the rent gap, and created attractive and sanitized districts for middle-class homeowners and consumers. Dispersal of the urban poor also soothed the worries of the bourgeoisie, who feared the spread of communicable diseases and the threat of social unrest. This early approach to urban redevelopment was not a real or just solution, according to Engels. He argued that the bourgeoisie settled the “housing question” by merely shifting the problem elsewhere, “in such a way that the solution continually poses the question anew” (1988/1872, 365).

A famous early model for large-scale, planned urban redevelopment was the renovation of Paris carried out by Baron Georges von Haussmann, Louis-Napoléon Bonaparte’s appointed Prefect of the Seine. Between 1853 and 1870, “Haussmannization” cut straight wide boulevards through Paris’s medieval alleyways, and introduced modernized sewage, water, and gas systems; new parks; train stations; public markets; and squares – all in line with an imposing and uniform architectural style. Urban redevelopment in this case was motivated by public health concerns – overcrowding and poor sanitation had led to multiple cholera outbreaks – but was also a way to enhance social control over the “dangerous classes,” which were concentrated in these areas and often rose up in revolt. Wide streets were intended to prevent the construction of barricades and to provide easy access for the marching military to swiftly put down future attempts at revolution. Urban redevelopment under Louis-Napoléon had another aim, with parallels in the contemporary era. The broad streets, regularized facades, and imposing terminal vistas were meant to create a spectacle, and to convey an image of the French capital as grandiose, elegant, and modern. Despite its visual appeal, redevelopment was socially destructive, displacing poor Parisians to outlying districts with overcrowded housing and poor sanitation, hence shifting the problem elsewhere and creating it anew.

Postwar urban redevelopment: urban renewal, slum clearance, and public housing

In the postwar period, urban redevelopment took the form of “urban renewal,” an approach that eviscerated inner urban areas with vast swaths of demolition, highway building, commercial and office development, and public housing construction. After World War II, the suburbanization of housing and decentralized industry dominated patterns of urban growth, often aided by government policies subsidizing both highway construction and middle-class homeownership. The inner city suffered tax loss and state-led disinvestment over this period, and in many US cities was abandoned to low-income and racially marginalized residents who were systematically excluded from good jobs and suburban homeownership. Inner-city disinvestment was paired at this time with sporadic urban redevelopment under the auspices of federal urban renewal programs (in the United States, launched as Title I of the 1949 Housing Act), which enabled government expropriation of “blighted” private land. This land was then sold at a reduced price to private actors willing to develop it for a “higher and better” use. Business elites and real estate interests profited massively from this program, building high-rise office towers, convention centers, shopping malls, and luxury apartments in the space of cleared “slums.” Renewal often introduced a modernist architectural and design landscape to inner cities, with wide highways, segregated land uses, and unornamented concrete towers.
Urban renewal and highway development resulted in the demolition of far more homes than were replaced, and displaced tens of thousands of residents and businesses (Gotham 2002). Postwar urban redevelopment also intensified and reproduced patterns of residential racial segregation, and disproportionately targeted racialized communities for dislocation and upheaval. In the United States, nearly 70% of the homes demolished for renewal were black-occupied, leading opponents to dub the program “Negro Removal” (Gotham 2002, 88; Greer 1965). Displaced families did not typically receive any assistance, and were shuffled into increasingly scarce and overcrowded housing markets. For black home-seekers in the United States, the reality of redlining, racially restrictive covenants, real estate steering, and white homeowner activism limited options to within the “color line” of segregated cities (see Figure 1). Echoing Engels, Greer (1965, 55) argued that urban renewal amounted to “slum shifting” and “pushing people around,” without properly addressing their housing needs. While it was perhaps most egregious in the United States, similar patterns are evident in cities across the developed world.

Postwar urban redevelopment programs also included the creation of public housing, to address housing shortages resulting from years of depression and war, and to improve the safety and sanitation in inner-city “slum” housing. In North America, fierce resistance from real estate interests ensured that public housing construction was limited, and built to minimum standards, so as not to compete with the private housing market. In the United States, public housing siting policies have been criticized for concentrating projects in low-income and African American inner-city districts, intensifying racial segregation. In Canada, federal urban renewal targeted racially marginalized inner-city populations for displacement, produced far fewer affordable units than were demolished, created stigmatized public housing in low-rent areas, and was massively profitable to real estate and business interests. The modernist design of postwar public housing has also come under fire, for its lack of relation to surrounding communities and building styles, and its distinctively austere functionalist style – now commonly associated with poverty, disadvantage, and stigmatization.

Postwar urban redevelopment programs, ostensibly intended to address slum housing and revitalize inner cities, were in reality programs that provided “federal subsidies to the private sector to encourage urban reinvestment” (Gotham 2002, 87). Urban renewal programs ended in 1969 in Canada and in 1974 in the United States, amid mounting criticism. Conservative writers railed against its bureaucratic nature, while neighborhood activists fought against expropriation, demolition, and displacement, and decried the destruction of historic cityscapes, modernist architecture, and authoritarian top-down planning. The rising tide of “new left” activism...
associated with civil rights, antiwar movements, and environmentalism ushered in a new political moment supportive of more citizen-oriented, participatory efforts toward urban redevelopment. While urban renewal was halted, the era of “postmodern” urbanism would present new opportunities for capitalist urban redevelopment and new rounds of dispossession and displacement.

**Contemporary urban redevelopment: postmodern revitalization and gentrification**

The contemporary context for urban redevelopment emerged after the global economic crisis of the mid-1970s and the period of global economic restructuring that followed. In Western advanced capitalist nations, this crisis ushered in a shift from the Fordist–Keynesian “regime of accumulation” that sustained decades of postwar economic growth to a new period of “flexible accumulation” and neoliberal ideological hegemony. This new era is marked by a range of trends, including a weakening of the collective power of labor, the introduction of flexible production systems, a rise in both precarious low-paying jobs and high-paying service employment (in finance, real estate, and insurance), deindustrialization in the Global North associated with increasing corporate mobility, and increasing financialization of the economy (Harvey 1989). Politically, these shifts have been guided and accompanied by the economic theory of neoliberalism and its emphasis on individualism, limited government, unencumbered markets, private property rights, and free trade. Neoliberal policies have underpinned the rollback of Keynesian social policies and institutions through budget cuts, privatization, deregulation, and enforced programs of austerity; and the rollout of neoliberal statecraft and regulations, including market-friendly policies and the aggressive disciplining and regulation of marginalized groups (Peck and Tickell 2002). A final shift associated with this era has been the “glocalization” of regulation, in which national state powers have shifted both upward to global institutions (such as the IMF and the World Bank), and downward to municipalities. Such a process often results in a rescaling of responsibility, in which the downloading of social services is not accompanied by sufficient resources, limiting municipal capacities and encouraging them to pursue entrepreneurial strategies coupled with antilabor policies and budgetary austerity.

For cities, this era has marked a shift in urban governance practices, from the managerial approaches of the 1950s and 1960s, which focused on providing public services, to an entrepreneurial model focused on facilitating private sector development and attracting economic development in a context of global intercity competition (Harvey 1989). Neoliberal urban governance involves creating favorable policies for investment (using tax breaks, subsidies, “empowerment zones,” etc.) in order to attract production functions (especially key command and control functions of the new global economy) and consumption dollars from tourists and middle-class urbanites. Cities are now active in public-private partnerships (absorbing considerable risk) and aggressive self-promotion, through “place marketing” and city branding exercises.

**Postmodern urban redevelopment: consumption landscapes and place-making**

In this context, a new style of urban redevelopment emerged in the 1980s and 1990s, targeting
the disinvested inner-urban landscapes of the “postindustrial” city for revitalization. With the loss of traditional manufacturing and production functions, cities began to support the construction of consumption-oriented attractions related to tourism, culture, professional sports, and entertainment. Aging downtowns, warehouse districts, and old port facilities have made way for themed shopping malls, convention centers, aquariums, luxury hotels, ballparks and stadiums, theaters, and festival marketplaces (Hannigan 1998). More recently, urban redevelopment has been boosted by a discourse promoting “creative cities” under the assumption that economic development follows middle-class consumers who work in “creative” jobs in white-collar management and information fields. Luring office- and knowledge-related production facilities, and creating consumption-oriented landscapes for leisure and entertainment to attract tourists and “creatives,” are a new preoccupation for revitalization-minded urban boosters and officials.

This style of postmodern urban redevelopment was promoted to stimulate investment in aging downtowns, and to catalyze spin-off economic impacts and “trickle-down” benefits. Since the 1980s, projects such as these have proliferated, becoming ubiquitous in North American central cities and beyond. Places like Winnipeg, St. Louis, and Birmingham have tried to capture the success of early adopters – duplicating Boston’s “festival marketplace” at Faneuil Hall (Figure 2), the “waterfront redevelopment” at Baltimore’s Inner Harbor, London’s Canary Wharf, or Bilbao’s successful “cultural revitalization” with its branch of the Guggenheim Museum. According to Harvey (1989, 10), this “repetition and serial adoption” of redevelopment strategies is a feature of the neoliberal era that results from interurban competitiveness to attract increasingly mobile capital investment. Competition between cities acts as an external coercive force, compelling municipal decision-makers to latch on to redevelopment trends or risk missing out. Ironically, the result is repetitive sameness rather than distinctive urban attractions. Rather tragically, Harvey argues, any gains that come from redevelopment are likely to be ephemeral, as imitation and innovation in other cities swiftly erode the competitive advantages of place. Competition for private investment also leads to a very narrow range of policy options for city governments, who are focused on tax cuts for business while engaging in a neoliberal race to the bottom in the areas of social policy and welfare provision.

Contemporary urban redevelopment has been criticized for favoring urban elites and private partners who profit directly from the development projects that are subsidized and facilitated by the public sector. In cases where redevelopment enhances the urban tax base, city administrators and landowners also win out. When revitalization doesn’t have the economic impacts that are expected, however, it is city taxpayers who end up subsidizing the loss. This outcome is typical, for example, of stadium-based revitalization: while countless...
cities have subsidized sports facilities to rejuvenate their downtowns and to keep or lure in sports teams, studies show again and again that economic impacts are illusory. Some argue that the democratic process is another casualty of entrepreneurial urban development because it elevates the profit-oriented goals of private developers over the general interests of diverse urban constituencies. Finally, despite hopes that economic benefits might trickle down to marginal and low-income urban residents, studies find that redevelopment tends to enrich development partners while displacing the urban poor. In the case of Baltimore’s famous (and serially replicated) Inner Harbor, for example, redevelopment did nothing to address the high rates of poverty, underemployment, and housing distress facing African American neighborhoods near the site (Levine 1987). Unfortunately, unequal outcomes appear to be as serially replicated as the museums, aquariums, and revamped waterfronts that are the hallmarks of postmodern urban redevelopment.

Place marketing, spectacular cities, and securitization

An important dimension of contemporary urban redevelopment in the neoliberal era is its tendency to focus on qualities of place, by creating urban districts that are themed, spectacular, and aesthetically appealing (Hannigan 1998). This focus on place is understood as a result of intense interurban competition, in which elite tourists, footloose high-tech industry, and its creative employees can pick and choose which place offers the best package of urban amenities. A focus on creating “livable,” safe, and beautiful urban environments, with parks, coffee shops, cool bars, upscale boutiques, and art galleries is seen as paramount for cities seeking to edge out the competition. Urban redevelopment is accompanied by place marketing, image-making, and efforts to sell a city (or the city’s crafted brand) as a commodity. As central areas are symbolically remade and rebranded into “fantasy cities” (Hannigan 1998) and theme parks for the affluent, the reality of resistance, diversity, and poverty in urban spaces must be both symbolically and physically removed. Urban redevelopment since the mid-1970s has been accompanied by “revanchist” (Smith 1996) policies and policing to control public spaces, criminalize homelessness, and keep “undesirable” or “deviant” users from compromising the image, which disproportionately targets the poor and people of color. “Quality of life” bylaws in many urban areas limit panhandling, loitering, sitting or sleeping on the ground, skateboarding, squeegeeing, and so on in an effort to sanitize urban space for the maximum comfort of affluent consumers and residents.

Gentrification and public housing redevelopment

The move toward central-city redevelopment is increasingly linked to the broader process of gentrification. Gentrification is defined as “the transformation of a working-class or vacant area of the central city into a middle class residential and/or commercial use” (Lees, Slater, and Wyly 2008, xv). While initially viewed as a marginal and strictly residential process, gentrification scholars now view it as the “leading edge” of global neoliberal restructuring, an all-encompassing process consisting of “the class remake of the central urban landscape” (Smith 1996, 39), or the “production of space for progressively more affluent users” (Hackworth 2002, 815). Beginning in the global cities of advanced capitalist nations in the 1970s and 1980s, and
stretching worldwide after the 1990s, gentrification has altered patterns of urban development, shifting the “land value valley” of devalorization, as well as traditional immigrant-reception functions, outward from inner cities into postwar suburbs, which are becoming the new urban districts of racialized urban poverty, territorial stigmatization, and disinvestment.

Early cases of gentrification began to occur when urban renewal was still the dominant type of urban redevelopment, and when discourses of inner-city decline were hegemonic. In London, England, and in US cities such as Boston, New York, and Philadelphia, “pioneer” gentrifiers (artists, bohemians, and risk-taking middle-class homeowners) began to upgrade housing and to displace working-class residents in disinvested central neighborhoods in the 1950s and 1960s. By the 1970s and 1980s the phenomenon had become more widespread, but was still seen as an anomaly in the context of trends toward suburbanization and inner-city decline. Even so, the process was becoming increasingly corporate and state-driven, as developers followed (and even preceded) “pioneers,” and as banks “green-lined” areas for reinvestment and municipalities worked to facilitate the process. Resistance and opposition movements emerged to fight back against gentrification and the social and cultural displacement it entailed. This resistance was at times militant (Smith 1996), and in some cases effective in slowing the process and forcing municipalities to preserve affordable housing and stop gentrification-related evictions.

Gentrification re-emerged after the recession of the mid-1990s with a new set of characteristics, and has diffused outward from global city cores into distant neighborhoods and downward along the urban hierarchy into cities previously untouched (Hackworth 2002). Early areas to gentrify have since been re-gentrified via processes of “financification” and “super-gentrification” (Lees, Slater, and Wyly 2008). Since the 1990s, big corporate developers have replaced individual owner-occupiers as initiators of the process, often forcing urban redevelopment into areas requiring larger capital investment to overcome barriers to property value appreciation. The role of the state as an active player in facilitating and encouraging redevelopment via gentrification has expanded since the 1990s. In the United Kingdom and the United States, national state programs (such as the United Kingdom’s “urban renaissance” programs and the United States’ HOPE VI program) have encouraged gentrification, and entrepreneurial city governments enthusiastically promote it. Indeed, with gentrification now generalized in the urban landscape, it has been globally adopted as urban policy by decision-makers and urban elites. Critics of gentrification point to its negative social impacts on the less affluent, working poor, and racially marginalized groups, who tend to experience displacement and exclusion as property values rise.

Since the early 1990s a new trend in urban redevelopment has involved the demolition of postwar public housing, the dispersal of public housing residents, and the construction of mixed-income communities in their stead (Crump 2002). This incarnation of urban redevelopment involves the extensive reach into areas previously thought “un-gentrifiable” (Hackworth 2002), including “ghetto” neighborhoods, single-room occupancy (SROs), and public housing. In the United States, public housing redevelopment was popularized by the federal HOPE VI program (1992–2010), which provided funding to local authorities (much like urban renewal) to demolish public housing and “deconcentrate” poor (predominantly African American and Hispanic) residents. HOPE VI redevelopment targeted many projects in gentrifying areas and resulted in a net loss
of social housing units and considerable displacement. Public housing redevelopment is not only an American phenomenon; it has been adopted around the globe, in the United Kingdom, Australia, Europe, Hong Kong, Singapore, and Canada as a “best practice” for eliminating postwar projects and facilitating gentrification by “social mix” (see Figure 3). In these places too, redevelopment can be violent and distressing, marked by unscrupulous and threatening behavior, and “revanchist” politics targeting low-income tenants (August 2014; Slater 2013). Community resistance has emerged to counter the destructive impacts of this new approach to urban renewal. In the United States, a grassroots coalition called the Right to the City Alliance published *We Call These Projects Home* (RTTC Alliance 2010), a community-driven research report arguing for the protection and expansion of public housing and the extension of housing rights. These demands envision a socially just approach to urban redevelopment, which prioritizes the needs of low-income communities (including women, LBGTQ communities, and people of color) over the needs of capital.

**Socially just urban redevelopment**

While geographers have shown that urban redevelopment has typically been associated with gentrification, displacement, and intensified...
spatial inequality, this does not have to be so. It is possible to envision a socially just urban redevelopment, in which the urban poor have the right to remain in their homes and neighborhoods, to send their children to high-quality schools, and to secure livelihoods within their communities. This vision would require interventions into capitalist market processes. Social housing can be built in wealthy and gentrifying areas to tip the balance away from concentrated wealth, targeting “redevelopment” toward rich areas, with the benefits also flowing to the poor. Inclusionary zoning can force developers to include a certain proportion of affordable units in their market-rate projects. A socially just form of urban redevelopment would set minimum requirements for such units at 50%, ensuring that wealthy households cannot dominate community governance. Redevelopment should incorporate a mix of housing styles and sizes, so that households of differing needs, life stages, incomes, and sizes can age in place. Similarly, urban redevelopment needs to include employment lands, for industry as well as community services, daycare, and health services—providing jobs and amenities within walking distance. Low-income, racialized residents and underrepresented groups should be guaranteed a voice on local councils making decisions on such matters, ensuring that urban redevelopment will be of benefit to the full range of people in a community. Such a democratization of urban redevelopment would necessarily mean limiting the reach of private sector investors and elites in determining city building processes. Otherwise, urban redevelopment cannot address contemporary urban problems, but will instead necessarily revert to the same basic pattern that Engels noted back in 1872, which is a hallmark of capitalist urban redevelopment—displacing the poor to make room for the rich.

SEE ALSO: Gentrification; Housing; Rent gap; Residential segregation; Slum; Urban renewal; Urban uneven development

References


Urban regimes

David Wilson

University of Illinois at Urbana–Champaign, USA

Urban regimes and its conceptual partner, urban regime theory, has been a dominant perspective in urban geography, political geography, and the social sciences for more than three decades. From its origins as a framework to explain city development and its relations of power in US cities, the concept has been applied across a diversity of settings: at the neighborhood level, city level, and metropolitan level. The perspective has been used to examine a multiplicity of questions, including how machine-style political formations stay in power, whether or how various interests (e.g., poor African Americans, poor Latinos, alternative sexualities, working-class communities) are incorporated into governing coalitions, how programs like urban renewal and public housing are established and implemented, and how uneven development in cities is deepened through government programs.

This work identifies urban regimes as amalgams of city-based institutions that construct, manage, and regulate the social, political, and spatial domains of cities. Such regimes — city governments, builders, developers, realtors, businesses, corporations, chambers of commerce — are seen to operate as semi-monolithic political entities in order to advance goals and plans. Their actions — plans, programs, initiatives — reflect the simultaneous drive to make cities ideal sites for capital accumulation and to successfully reproduce regime formation and maintenance. Successful regimes are formed when their dominant narratives successfully suggest that such amalgams and their actions are fruitfully and progressively advancing the best interests of the city and its populations.

The notion of urban regimes is inseparable from its underpinning theoretical base, urban regime theory. Regime theory views power as fragmented and regimes as the collaborative arrangements through which local governments and private actors assemble the capacity to govern. The original focus of urban regime theory, rooted in 1970s work, was the connections between locally based urban amalgams of institutions and the structure of the world economic system. Here such regimes were theorized as mediators and interpreters of transcontinental economic and political structural features. This focus soon shifted to the connections between US city regimes and the American capitalist economy. Fainstein and Fainstein’s (1983) work was extremely influential. To them, the character of urban regimes is derived from two structural features of the political economy: local government dependency on property taxes for fiscal solvency, and production of commodities being primarily privatized. These structural features, to Fainstein and Fainstein, explain the dominance of business, privatization, and property value interests in local politics.

More recently, regime theory has deepened to incorporate the realities of how city regimes must be responsive to political struggles, changing political and cultural conditions, and contradictions in local, regional, and national economic formations. As poststructuralist and postmodern sensibilities have suffused urban studies since the turn of the twenty-first century,
regime theory has seamlessly incorporated these concerns. Yet, still, the regime perspective continues to embrace a political economy orientation that rejects both pluralist assumptions that governmental authority is adequate to make and carry out policies, as well as structuralist assumptions that economic forces determine policy. Regime analysis, observed Clarence Stone (1989), should always be anchored in a foundation that examines the interconnections between the inheritance of structural forces, how these are interpreted, responded to, and modified, and the human decision to act. Urban regimes, to Stone and to current practitioners of this perspective, are structured but actively building and structuring constellations of power.

SEE ALSO: Neoliberalism; Race and racism; Uneven regional development; Urban geography

References


Further reading


Urban renewal

David L. Prytherch
Miami University (of Ohio), USA

Urban renewal is a set of policies and planning practices addressing urban problems through government-led redevelopment. Urban renewal specifically connotes slum clearance policies in mid-twentieth-century North America, but is inseparable from wider urban redevelopment and gentrification processes.

Urban renewal, like biological metaphors often justifying it, dates back centuries. But urban changes accompanying industrialization prompted more systematic public interventions, particularly to relieve congestion. The Haussmann Plan for renovating Paris is one nineteenth-century example, but advocates of progressive tenement reform and modern(ist) social housing justified “slum clearance for its own sake” (Zipp 2012) in the twentieth century.

New public health and social sciences offered emergent welfare states techniques for analyzing urban problems, and planning provided the tools to address them.

Urban renewal took specific policy shape in the United States with the Housing Act of 1949, which promoted community development and “elimination of substandard and inadequate housing through the clearance of slums and blighted areas.” Grants encouraged local authorities to draft plans identifying slums and “deteriorating areas,” purchase or condemn them for redevelopment, and construct public housing. A revised Housing Act of 1954 shifted emphasis from public housing to “urban renewal” and commercial redevelopment. Urban Renewal Authorities across American cities targeted high-density, mixed-use, and typically minority neighborhoods for replacement by modernist commercial development and high-rise housing projects. But housing construction lagged, and millions – mostly African Americans – were displaced before what some called the “federal bulldozer.”

As impacts of these policies grew, so did critiques. Urbanist Jane Jacobs framed her book The Death and Life of Great American Cities (1961) as “an attack on current city planning and rebuilding” (p. 3), likening renewal to the “pseudoscience of blood-letting” (p. 13). Sociologist Herbert Gans highlighted the inequitable emphasis on developer over resident needs, and historian Robert Caro argued renewal created new slums as quickly as it cleared the old. By the 1970s, these critiques and public protest helped shift federal policy away from modernist planning and toward redeveloping existing neighborhoods through block grants.

But urban renewal continues today, if under different names, and remains subject to debate by policy actors and urban scholars. Cities today partner with private capital in strategic investment, planning, and use of eminent domain to encourage gentrification and mixed-use development. This “new urban renewal” (Hyra 2012) may emphasize historic preservation over clearance, but “serial forced displacement” continues apace (Fullilove and Wallace 2011). And the high modernist ideals and practices of urban renewal still enjoy a prominent planning role in industrializing and urbanizing countries like China, where historic neighborhoods – like Beijing’s hutong – are razed for high-rise development.
URBAN RENEWAL

Urban renewal may have specific connotations, but can be found wherever public planners encourage redevelopment and displacement in the name of blight removal and modernization.

SEE ALSO: Modernization theory; Urban planning: human dynamics; Urban redevelopment

References


Urban resilience

Richard Friend

University of York, UK

Urban Climate Resilience in Southeast Asia Partnership (UCRSEA)

Urban resilience has become a ubiquitous motif but one that has emerged from disparate conceptual and historical roots, and is being applied with a diverse range of definitions, interpretations, and objectives. The rise of urban resilience brings together two agendas; the growing interest in and concern for an urbanizing world, along with a sense of a future that is increasingly uncertain and fraught with risk, in relation to global ecological change and to economic and political uncertainty. In order to understand urban resilience, it is necessary to unpack these two separate strands and to scrutinize the ways in which they have become entwined.

That the world is increasingly urban is widely accepted, even if the extent and significance of such urbanization is very much a matter of critical debate. With statistics in 2008 pointing to the majority of the world’s population being urban, a critical turning point in the ecological and social history of the world has been reached. This is interpreted in different ways – from an evolutionary perspective, in which the transition from rural to urban is an inevitable, necessary, and positive progression, to a debate framed around the growing importance of the globalized economy and the ways that cities are linked across territories and geographies, to representing an era beyond the nation-state. Cities and urbanization are powerful metaphors.

At the same time, resilience has emerged as a way of framing responses to the vicissitudes, uncertainties, and risks of life, whether on a personal, psychological level or at the ecological level, in the face of climate change and disasters and of emerging economic and security concerns. As a consequence of the prevalence of large-scale natural disasters, the economic crisis of 2008, and the spread of global terror and security fears, resilience has become a metaphor (Olsson et al. 2015) that appears to have crept into all facets of contemporary life at a time when life itself is portrayed as inherently uncertain.

The 2000s have also witnessed a series of major “natural” disasters that have struck urban centers in both the Global North and the Global South – Hurricanes Sandy and Katrina, Typhoon Haiyan, and the floods of 2011 in Thailand. Each of these events portrayed the implications of shocks and crises on large urban centers vividly. Intergovernmental organizations, international financial institutions, national and local governments, and large charitable foundations have all embarked on a range of initiatives targeting urban resilience to climate change and disasters, creating funding mechanisms to leverage investment finance to pay for building urban resilience. Urban considerations, and the language of urban resilience are also being taken up in critical global policy commitments such as the Sendai Framework for Disaster Risk Reduction, Sustainable Development Goals (SDGs), the United Nations Conference on Housing and Sustainable Urban Development (Habitat III), and the Paris Agreement of the United Nations Framework Convention on Climate Change. In the parlance of international development commitments and policy, achieving resilience is both an end in itself and a means to an end.
With so much of the world now framed as being urban, it is not surprising that the interest in resilience should be coupled with the urban world. Yet, in considering resilience and its application to the urban context, it is important to distinguish between resilience as a theory and concept with its own intellectual (if diverse) roots, and resilience as discourse that generates a whole set of connotations and meanings in its own right. Drawing this distinction between the theory and discourse of resilience becomes all the more critical when the two terms are used in combination. Much of the confusion and more heated debate around urban resilience has been caught between these two very distinct threads.

The theoretical concept of resilience has a long tradition of complex systems thinking, largely based around the ecological sciences but also found in the disciplines of psychology and engineering. Each of these intellectual fields influences its application in the urban context. Across these disciplines the significance of systems and functions is center-stage. Resilience thinking from the ecological perspective argues that complex systems are shaped by universal fundamental, natural laws that allow them to adapt to changing circumstances and, if necessary, transform themselves. Ecological systems possess qualities of dynamic equilibrium, thresholds, and self-organization (Olsson et al. 2015; Gunderson and Holling 2001). This interest in complex systems that have their own boundaries, structures, and functions and that can change and adapt, yet still maintain their core function, fits neatly with approaches to cities as organic systems, with their own structure and metabolism, yet with a whole set of fragilities, vulnerabilities, and risks. The psychological sciences tend to adopt resilience as an innate human characteristic, partly shaped by social and environmental factors, that in some circumstances requires intervention and nurturing to allow it to come to the fore. Building resilience allows individuals to function as full human beings, overcoming adversity, tragedy, and trauma; disturbances that are themselves seen as part of the human condition. From both disciplinary perspectives, resilience is thus an innate, natural feature of social-ecological systems and a universal characteristic of life. In many ways the engineering perspective on resilience is less problematic; the boundaries of the system, and the social purpose, function, and structure of an engineering initiative, are clearly defined as part of its design, which itself is a product of social and economic processes. Thus, identifying the boundaries of the system, and maintaining structure and function in the face of shocks and crises, are less problematic.

Images of both ecological and engineered systems are often alluded to in how resilience theory is defined and applied in the urban sphere. A common thread running through the theory of urban resilience lies in the application of at least some aspect of systems-based approaches to the contemporary urban sphere (Elmqvist 2014). The city is presented in some way as a system itself. This has been framed in different ways. The city as a system has been an enduring metaphor in urban studies. Ecological approaches have also been applied to the urban system. The link with the natural world also led to the city being presented as an urban metabolism, an ecological system with quantifiable flows of energy materials and information. This general approach has been developed into a view of cities bringing the natural and social world together, in line with concepts of the social ecological, in which the city is a combination of biophysical and social factors.

The urban systems approach has other traditions. It is argued that contemporary urbanization is distinct from other periods in human history because of its dependence on complex systems of infrastructure and technology that
allow for the scale and intensity of urbanization, and create linkages and networks of cities across specific territories and geographies. The engineering perspective has a clear link to approaches to urbanization that prioritize the built environment and the dependence on infrastructure and technology that characterize contemporary urbanization (Graham and Marvin 2001).

There are clear parallels with earlier approaches on networks of cities and global perspectives that highlight the ways in which cities are globally networked and interlinked, whether through infrastructure and technology (e.g., fiber-optic cables) or through flows of natural, financial, or labor resources.

Work focused specifically on urban climate resilience, an influential field in urban resilience, presents urban systems as composed of the interactions between (i) ecosystems, infrastructure, and technology; (ii) institutions (the rules, norms, regulations, and daily practice); and (iii) actors, that is, organizations, communities, households, and individuals (Tyler and Moench 2012; da Silva, Kernaghan and Luque 2012). Grounded in the theory of social ecological systems, this approach presents urban climate resilience as a contested process that brings together different interests and stakeholders, and that requires dialogue and different types of knowledge (Tyler and Moench 2012). For some authors it is an essentially political process. From this perspective, building resilience cuts across these three dimensions of urban systems, each with its own set of characteristics.

Such systems-based approaches to the urban phenomenon also points to emerging points of vulnerability not only as lying in shocks and crises that are place-based but also as being related to fragilities and potential failures that reside within multiscale systems themselves. Again, this kind of perspective appears in different forms and with different interests. It is argued that many cities are facing critical natural resource constraints: drawing core resources from increasingly distant hinterlands, they are yet unable to meet the needs of many citizens, and increasingly under strain as they expand and intensify. The infrastructure systems on which cities depend, and the ecological systems that make these viable, are increasingly vulnerable. While a shock or crisis may originate in one specific location, the way in which such systems are interlinked and interlocked creates impacts that cascade from one location to another, in some cases cascading across the globe through the systems that link different geographies.

These concerns also appear in a theory of urban conflict and resilience more overtly focused on military security. From this perspective, emerging urban systems create a whole new series of security threats that are argued to be irregular, hybrid, and nested (Kilcullen 2012). Such security threats are directly related to the connectedness of urban systems and to the trend for urbanization to occur in hazardous coastal locations. This again reminds us that, while climate change and disasters have been fairly persistent concerns running through urban resilience literature and debate, thinking around resilience has long been prominent in security theory.

Urban resilience theory, and the systems perspectives that underpin it, also draw on particular concepts of governance. Resilience has been closely associated with approaches to ecological governance that highlight issues of scale, arguing the need for multiscale, polycentric, participatory, and learning-oriented approaches. Similar basic approaches have been taken on in the urban resilience literature. Some emphasize issues of process and public learning (Reed et al. 2013) and others the need for institutional planning processes that are better able to deal with uncertainty and risk. Such systems-based approaches to urban resilience operate alongside...
a more individualistic dimension, where the adaptations that are argued to be most plausible, rather than most needed, are said to be realized at the individual, household, and community levels rather than through formal mechanisms of the state.

Much effort in the urban resilience literature has also been directed at identifying and measuring resilience. With so much debate as to what urban resilience means, there is also an effort to define the conditions of urban resilience, and especially to find globally applicable indicators and metrics. In this way, characteristics of resilience are outlined, for example redundancy, safe failure, diversity, flexibility, which can be measured quantitatively according to universally applicable indicators, and which allow for both self-assessment and comparison across different urban areas.

While urban resilience has certainly been in the ascendant in recent years, a number of critiques have also emerged, of both the theory and the concept of urban resilience, but perhaps more often of the discourse and policy narratives with which it is associated. These critiques have emphasized the importance of political economy perspectives of urbanization, pointing to the constraints of systems approaches and their application to the urban context, and to the limitations of resilience more broadly.

Resilience theory has been widely critiqued for its inadequate consideration of politics and power (Béné et al. 2012; Tanner et al. 2015). When applied to an urban context, this criticism relates to the ways in which cities create social and political relations, to cities as contested arenas, and to cities as performing symbolic functions in creating physical embodiments of meaning and value. For some authors, the urban (climate) resilience agenda has been poorly informed by theories of urbanism or of urbanization and the city, with inadequate attention to historical and political economy perspectives. From this perspective, urban resilience needs to take on board how urbanization is both a product and a necessary condition of capitalism that allow for the agglomeration of assets of production in specific locations and the transformations of social and economic relations of production and exchange.

Both urbanization, and by extension resilience-building efforts, inevitably support certain interests and groups at the expense of others. Depending on how the city is defined, and by whom, the resilience of the city does not necessarily mean a just and equitable outcome. Critics argue that, without a theoretical approach that is grounded in history and political economy, it is not possible to contextualize or to analyze the urban.

Despite this rich intellectual history, much of what appears as urban resilience is not always so clearly grounded in theory and concepts. Across the development and humanitarian sectors, resilience has emerged as a loose discourse, in which there are competing efforts to define what it means and what it implies rather than attempts at conceptual clarity. With the plethora of competing definitions, and public campaigns to say what resilience means, it is a term that means all things to all people, and eventually risks meaning nothing at all. It is in the discourse of urban resilience that the murkiness is greatest. Yet the discourse also takes on some of the core principles from resilience theory, but as metaphor with connotations rather than as precise theoretical application.

The influence of the discourse of urban resilience is thus not so much in terms of what it defines but what it symbolizes and the connotations that surround its use – and how such discourse can help shape policies, legitimizing certain kinds of solutions, knowledges, and actors while delegitimizing others. As Raymond Williams (1976) wrote of keywords in policy
discourse, their pervasiveness and endurance relate to the combination of the everyday with the technical, and their essentially positive associations. Being resilient is clearly a good thing, and so it is difficult to be against resilience.

The discourse of resilience draws as much on its everyday meaning as its more technical meaning. Connotations of coping with adversity and of stoicism in the face of repeated shocks and crises lend themselves to a whole set of urban policy framings, with a range of political implications. When the discourse of systems is introduced, this becomes additionally problematic, with urban resilience easily lending itself to policy narratives that privilege the protection of existing systems without considering how these systems are constructed and reinvented or whose interests they serve. Given resilience’s connotations of withstanding and bouncing back from shocks and crisis, and the need for tradeoffs in order to maintain system structure and function, there is less consideration of the systemic causes of such shocks and crises. As such, when resilience enters policy discourse, it tends to be more readily associated with a conservative or a more reactionary political agenda (Peyroux 2015; Tierney 2015).

The communitarian and individualistic arguments that emerge from urban resilience can similarly be problematic. These arguments have two quite distinct sets of interests. For some commentators, community is held up as the driver of radical, grassroots change: the mechanism to counter the extremes and injustices of both the market and the state. Yet individualistic framings are also in line with arguments around “autonomous adaptation” with the responsibility for action at individual, household, and community level. As such, urban resilience can be very much in line with neoliberal arguments, about the failings of the state, and the inadequacy of formal government systems to deal with complex, multiscale challenges. Such arguments tend to privilege markets and individual responses.

Emphasized as a human quality to be nurtured in individuals, households, and communities, the concept of resilience takes responsibility for shocks and crises out of the public sphere, thus lending itself to a narrative of coping and bouncing back – even when the system might be broken (Harrison 2012). For others, that the discourse of urban resilience has emerged during a period of global austerity is no coincidence (Slater 2014). Resilience narratives can easily legitimize the state pulling back from its responsibilities (Harrison 2012; Peyroux 2015; Tierney 2015).

Along with the discourse of resilience, a more persistent urban policy discourse has emerged that is shaped by interpretations of what cities mean as they become metaphors for a whole range of political, economic, and social values as centers of business and economic activity and thus of innovation; the need for infrastructure; and the business case for building smart cities. These shifts lend themselves to large programs of investment-led resilience building, of bringing in the finance to build a resilient future.

This discourse has also been challenged around issues of power and knowledge, with a critical counternarrative of urban resilience emerging (Slater 2014). Resilient cities are not necessarily cities that are socially just or equitable – and indeed critics would argue that arguments around resilience operate by diverting attention away from these more overtly social and political considerations. In many ways, the grounding of resilience theory in systems theory, and the connotations of innate universal natural conditions reinforce a discourse and range of policy narratives of a resilient city that ignore such issues. Yet other interpretations again highlight the political dimension of urban resilience, and
framing the problem as fundamentally one of governance (Friend and Moench 2015).

In other disciplines, resilience has been presented as part of a continuum (Elmqvist 2014). For example, in the literature and practice of disaster risk reduction, resilience is a necessary but not a final condition. Ultimately, the reduction of the risk of disasters requires transformative solutions that address the root causes of vulnerability; where systems are broken they need reconfiguring. Increasingly both the urban resilience, or perhaps more accurately the urban futures debate, and the climate change debate have begun to scrutinize the trajectory of economic development, in which urbanization plays a central role. A more critical body of theory has emerged arguing that both urbanization and climate change are features of capitalism and that, in order to avert global ecological catastrophe, we need a transformative urban agenda that is climate compatible and also socially just, equitable, and inclusive.

A common notion across these debates is that we are entering a critical and uncharted period in human history, and are thus caught up in a struggle about reconfiguring social life that is very much about an urban future, in relation not just to the city space itself but also to what the increasingly globalized world should look like and for whose benefit. Perhaps these struggles of the urban future are as significant as similar, more familiar, struggles over citizenship, nationhood, and nationality; social justice, equity, wellbeing, and rights; and human–nature interactions.

SEE ALSO: Cities and development; Climate adaptation/mitigation; Desakota; Discourse; Environment and urbanization; Global cities; Multilevel governance; New Urbanism; Political ecology; Public policy; Resilience and human geography; Rural/urban divide; Social resilience and environmental hazards; Sustainable cities; Urban ecology; Urban geography; Urban political ecology; Urban politics; Urban uneven development; World cities

References


Urban transit covers the realm of nonpersonalized transportation in cities. Known by many terms including “mass transit” and “public transportation,” it refers to the task of moving people throughout urban environments in shared vehicles where origins and destinations are accessed by vehicle stops that can be either at predetermined stops or stations, or at ad hoc locations, both types of which can feature varying levels of infrastructure. The challenge of the mass transit of people is a formidable one in transportation, and immense amounts of money and time have been spent on utilizing urban transit to alleviate congestion, move individuals in a timely and efficient manner, and advance the interaction of transit with land use and economic development. This entry reviews the evolution of this process and the different modes of urban transit that have been utilized. It ends with a discussion of how urban transit varies by region, and of the future outlook of urban transit technologies, use, and deployment.

Evolution of urban transit

In the United States, urban transit has been described by Peter Muller (2004) as developing in four overlapping eras of the spatial evolution of cities: the walking/horsecar era (1800–1890), the streetcar era (1890–1920), the recreational automobile era (1920–1945), and the freeway era (1945 to present). Though this timetable is obviously US-centric, urban transit across the world can be seen developing through these stages at different times and places.

Urban development has occurred as a function of the speed and carrying capacity of its transportation system. Therefore, urban transit has evolved as a function of the relationship between the technologies available for transportation with the type of urban development supported by the characteristics of that travel technology. The level of economic development of a city and its host country; its cultures of mobility, development, and personal space; and the structure, legitimacy, and efficacy of its political institutions also influence the evolution of a city’s urban transit system.

The walking/horsecar era

Before the middle of the nineteenth century, most American and European cities were small, highly clustered settlements where citizens were able to walk from the center of the city to the edge in 30 minutes or less. After the early 1830s, the Industrial Revolution placed pressure on the boundaries of the city by increasing population and population densities, innovating building construction to allow for taller structures, and creating intense pollution in areas of the city near factories. It also provided the railroad, whose steam- and coal-powered locomotives provided a means to power larger-capacity vehicles at speeds faster than walking along dedicated rail tracks. Along with the horsecar (where horses pulled passenger trolleys along existing roads), the railroad provided the speed and carrying
URBAN TRANSIT

capacity to expand daily travel in and out of the inner city to new “suburban” residential development built around stops that featured more single-family homes and settlement at lower population densities. The railroad also helped introduce and then make commonplace the act of “commuting” from home to place of work, whereas before people usually lived near or in the same buildings as where they worked. These introductions provided an opportunity for middle-income citizens to move away from the congested city center, and helped start the gradual dispersion of the urban center (Muller 2004). Some cities in the developing world experienced this stage of urban development and transit deployment, especially those that were parts of colonial powers, at similar times or in the ensuing decades after the Industrial Revolution spread through the developed world.

The electric streetcar era

By the early 1890s, the electric trolley or streetcar was the most common mode of intraurban transit in the United States. Systems progressed with similar speed in European cities, and their impact on the development of urban areas was profound. This new ease of movement further encouraged rapid outward expansion of the city, with new land utilized for home development and city subcenters along trolley lines and trolley stops, and saw the emergence of specialty districts for commerce, industry, and transportation. Running on electricity, streetcars lacked the point-source pollution of the locomotive railroad’s smokestack and the horsecar’s equine waste, and served to help clean up urban areas dramatically. Trolley lines stretched out to the edge of the city, allowing for the rapid expansion of urban development while improving access to all the essential activity nodes of the city. These lines would later go on to pave the path for major boulevards (and later freeways), which aided the proliferation of the automobile (Muller 2004). Colonial cities in the developing world saw parallel developments, while such established, fixed guideway transit systems were largely bypassed in urban areas, which were growing rapidly under modern globalization, in favor of bus and informal, small-scale means of urban transit.

The recreational automobile era

It was the early 1920s before automobile ownership began to spread beyond early adopters to the general population. The roaring twenties and the economic expansion that preceded the Great Depression, as well as assembly-line innovation and Henry Ford’s Model T helped to democratize automobile ownership and establish it as an economic and cultural aspiration. During this period up to the end of World War II, the trolley lines and subways of urban areas were at their maximum use. Every urban settlement of any size had an established streetcar network, and the combination of relatively limited spatial expansion of the suburbs (compared to what was to come), extensive transit infrastructure, accessible mixed-use development near widespread transit stops, and limited automobile ownership and highway infrastructure (but still enough to alleviate some transit congestion) has led to this being called the point where “intrametropolitan transportation achieved its greatest level of efficiency” (Muller 2004, 69). Modern European cities more closely resemble this now, though the proliferation of the automobile within them is still much greater than what cities in this era saw. Many rapidly growing cities of the developing world (in particular, those in China) are currently going through a version of this stage, though their population densities are so great and their automobile growth so rapid that they
are experiencing choking congestion and urban air pollution, and much more of their urban transit is by bus than rail.

The freeway era

Production efforts for World War II and the preceding 15 years of economic depression hindered the expansion of the automobile industry. However, in conjunction with the petroleum industry, the automobile industry boomed once the postwar economy transitioned from military production to the production of consumer goods. Additionally, in 1956, the Interstate Highway Act was implemented, which would provide high-speed infrastructure to maximize the speeds with which the automobile could travel. A rising standard of living, along with numerous other cultural factors and institutionalized policies (such as incentives for homeownership and mortgage loans) helped prioritize individual automobile travel as a social status symbol and transportation necessity. As urban areas grew, they rapidly decentralized, and single-family housing proliferated. Cities encouraged radial expressways around their central business districts in the 1950s and 1960s, believing that it would increase accessibility to downtown by speeding up travel. The proliferation of these freeway systems sped up residential and commercial decentralization (Muller 2004).

This changing physical and social environment of transportation had a profound and lasting impact on urban transit. As transit use declined with increasing automobile use, petroleum and automobile companies purchased private transit operators, removed their extensive track systems, and replaced them with buses (or motor carriages), which were more flexible in routing and had smaller carrying capacities. The service quality and extent failed to match those of their rail-based predecessors, and, with rapidly dropping population and housing densities, transit use declined to near zero in many areas and private transit operations almost ceased to exist. Urban transit was reinstated as a publicly managed social service to provide transportation to those who couldn’t afford or were unable to own an automobile in American cities in the 1960s and 1970s. At this point, urban automobile congestion began to reach critical levels, and in the 1970s and 1980s urban transit first began to be reintroduced to urban areas as a way to combat this congestion and to help redevelopment in declining central city areas. This continued into the 1990s as heavy rail, and then later light rail and dedicated busways, began to proliferate in US cities, and bus services grew to help feed some of these lines. Rising gasoline prices also placed pressure on automobile use, and by the early 2010s urban transit use was at its highest level in the United States since the early postwar period.

Many European cities had a different experience, with the evolution of urban transit after World War II in large part due to the extensive damage as a result of the conflict. In the reconstruction of these cities, historical city centers were preserved and urban transit prioritized to a greater extent than in postwar US urban growth. Proliferation of the automobile has long been an issue in European urban areas, but greater constraints on automobile use, funding for transit and taxation of automobiles, and tighter coordination of land uses with transit investment meant that urban transit has always enjoyed a much larger role and presence compared to its American counterpart.

Cities in the developing world often find themselves in a hybrid version of the recreational automobile era and the freeway era. Rising standards of living, rapid economic and population growth, and automobile ownership as a sign of social status encourage its use, but extraordinary wealth
disparity and poorly functioning state institutions hinder management of this growth and of urban transit as a means of alleviating the problem. As such, many cities in the developing world are characterized by extreme urban congestion.

Urban transit vehicular modes

Urban transit is deployed through an array of technologies. These technologies vary by cost and powertrain, as well as by carrying capacity, speed at which the vehicle can travel, and the means of piloting the vehicle. They also vary by infrastructure needed for the transit vehicle to operate and the nature and frequency of stops used to access the transit vehicle. Lastly, transit operations vary by the degree of structure and formality of the entity operating the transit service. Some modes, particularly in developing countries, play an instrumental role in the transportation of people around busy urban areas despite not being formalized by the state or by public transportation agencies.

Bus transit

Perhaps the most recognizable and ubiquitous form of urban transit is the bus. Buses come in many shapes and sizes, but are characterized by having a single driver and a mix of seats and standing room for passengers. Buses have often run on diesel fuel, though in the developed world there has been a transition to cleaner burning fuels such as liquefied natural gas (LNG) or hybrid electric motors. Buses usually run fixed routes and pick up passengers at predetermined stops, though ad hoc and on demand services are also deployed. They usually run in traffic with other cars and share roadway infrastructure with private transportation. Many buses have signal priority and dedicated spaces to pull into stops to improve their ability to maintain a schedule and ease their impact on urban congestion.

Light rail transit (LRT)

Light rail transit most closely resembles the vehicles of the electric streetcar era. These systems usually operate single cars or short trains along exclusive rights of way at ground level, though sometimes they also operate in mixed traffic. Additionally, they can operate on dedicated aerial or subterranean structures. The flexibility of LRT is what distinguishes it from other forms of fixed guideway modes of transportation (Transportation Research Board 2000). Light rail can take on many forms and names (modern streetcars, trolleys, and trams often feature LRT characteristics). Some systems are very short, serve almost exclusively tourist or recreational trips at low speeds, and operate with other traffic; others resemble heavy rail systems more closely, operating in long multicar trains at high frequency along dedicated tracks, with extensive infrastructure around stops. Light rail has been a popular choice to help alleviate urban congestion and instigate urban development around stops and central cities in the developed world, particularly in North America, and to a lesser degree in European cities.

Heavy rail transit (HRT)

Heavy rail transit systems have separate rights of way from all other vehicular and foot traffic, with sophisticated signaling and high platform loading. They operate with heavier loads of traffic at higher speeds in larger cars and with less frequent stops than light rail. Heavy rail transit can also be defined by its power source, which is usually supplied electrically by a third rail. Heavy rail commonly runs underground in the form of subways so as not to interfere with already
dense street surface congestion, though it sometimes appears on the surface and may also run on elevated structures within a system. Sometimes referred to as a metro, heavy rail is usually limited to densely congested corridors of the largest cities in a country, and is an essential part of a city’s transportation system in these places. New York, London, Paris, Mexico City, Tokyo, and Moscow are all examples of places with extensive, heavily used metro systems.

**Commuter rail transit (CRT)**

Commuter rail systems are electric or diesel-powered railways for distance travel between a central urban city and its surrounding suburbs. Services on commuter rails are operated on a regular basis under contract for the sole purpose of transporting passengers within or between urbanized areas and their outlying areas. They are not usually meant to influence urban development; instead, they are meant to serve it and to alleviate congestion in highly traveled corridors. A specific characteristic of commuter rail systems is their offer of multitrip tickets from station to station, usually with only one or two main stations within the urban core of a city.

**Bus rapid transit (BRT)**

Bus rapid transit can be described as having all the characteristics of a rail-based system without running on actual rails. BRT is a high-capacity rapid transit system that operates on a dedicated, fixed guideway lane. As with a train, passengers are picked up and dropped off at stations, and the grade-separated lane gives the BRT the ability to avoid automobile traffic. Many BRT systems also include an intelligent transportation system (ITS), which tracks BRT vehicle locations, controls traffic signals, and predicts arrival times in order to streamline the process. BRT is increasingly popular among urban policymakers because its infrastructure costs are cheaper than those of light rail and other fixed guideway modes.

**Cable car**

A cable car is a railway-based individually controlled transit vehicle similar to a trolley or a streetcar, which uses a moving cable system located below the street surface and is powered by a central offboard motor. An iconic example of cable cars is the much loved system in San Francisco, California. Cable cars are also capable of traveling on steeper gradients than light or heavy rail, because they are pulled by cable rather than trying to push their way up gradients, which also explains their use in places, like San Francisco, which have many sharp elevation changes. Most cable car systems are used primarily for tourism, because the small cars cannot handle a heavy load of traffic such as those of light or heavy rail systems.

**Aerial tramway**

An aerial tramway consists of an electrically powered system of cables that carry suspended, powerless vehicles. While the most common use of aerial tramways can be seen at ski resorts as gondolas or in a mountainous area that draws a lot of tourism, some cities use aerial tramways as a quick way to move people up and down highly mountainous urban areas. One successful example of aerial trams as urban transport can be seen in Portland, Oregon, which introduced the Portland Aerial Tram in 2007 in order to ease traffic congestion for patients and employees of the Oregon Health and Science University, the largest employer in Oregon (Greenhalgh 2013). Medellín in Colombia has also used aerial
URBAN TRANSIT

tramways as a form of urban transit across its mountainous cityscape.

Ferry boat
Ferry boats are most commonly used in coastal areas to transport people (and sometimes vehicles) across large bodies of water that divide urban areas. Most ferry boats are powered by diesel or steam, and have the ability to carry a large amount of weight, which is useful when commuters driving across bays or harbors need the use of their vehicle once they reach the ferry’s destination. New York City and Seattle are examples of areas that feature ferry boats as essential parts of their transportation systems.

Taxi
Taxis are individual cars that are usually part of a larger fleet working for a private company, as opposed to a state-funded and organized agency. The distinguishing characteristics of taxis include their ad hoc, point-to-point pickup, metering charge by distance, and usually small carrying capacity. They are extremely common in urban areas, and are often used when people prefer the convenience and privacy of individual transport that picks them up and drops them off exactly at their desired points of origin and destination. Many traditional taxis operate by the driver traveling through high traffic areas and picking up passengers who signal to them. Increasingly, taxis also operate in response to demand, where a trip is arranged ahead of time. Recent developments have included the rise of car-sharing services such as Uber, which have challenged the traditional structure of the taxi cab company. Other types of taxis which are more common in cities in developing countries include the jitney, which is a shared taxi between the size of a passenger vehicle and a bus, and a rickshaw, which is a cart pulled by one or more people on foot or tricycle.

Regional differentiation

Urban transit deployment varies considerably across different regions of the world. Many factors drive these differences. Economic development and standard of living are important influences. Cultural factors, policies, and institutional structure also influence distinctiveness, as does physical geography. These are discussed in greater detail by region in the rest of this section.

North America
North America has already been largely covered in this entry. Urban transit in the United States and Canada has been closely tied to the development of the automobile and (in particular in the United States) national-level policies designed to facilitate automobile mobility and single-unit homeownership. With industrialization in the late nineteenth and early twentieth centuries, urban population densities grew to extraordinary levels. Since the automobile boom of the mid-twentieth century, private transportation has become an ever-growing necessity in North American culture. The Interstate Highway Act is one of the most important public works projects in American history because of its influence on the modern structure of cities, through facilitating the suburbanization of housing and single-use development by the rapid expansion of roadway infrastructure after World War II. Urban transit has had a revival in use and prioritization in quite a few North American cities in the late twentieth and early twenty-first centuries, but it still constitutes a small proportion of overall travel in these places.
Europe

Urban transit in Europe, as in North America and East Asia, has a long history embedded within the evolution of transportation technology and the spatial characteristics of the city. Major differences from North America include the influence of the age of cities and the rebuilding of many European cities after World War II. As in the United States, the history of urban transit began as a response to industrialization in the eighteenth and nineteenth centuries, and its effect in creating very large urban densities of population. Partly as a result of the natural density of the old European cities, the negative effects of heavy automobile transportation like air and noise pollution became more acute than in their North American counterparts after World War II. Many policies have been enacted that regulate vehicles in congestion-free zones during certain periods of the day, and that make gasoline consumption and automobile purchase considerably more expensive, which explains to some extent the smaller number of vehicles on the road compared to in the United States (European Commission 2015). Nearly 60% of Europeans use public transportation in some form, compared to a mere 5% of Americans (Benfield 2011).

Australia

Australia has urban transit very similar to its developed world counterparts, though it more closely resembles American than European cities. Sydney and Melbourne have large, advanced fixed guideway systems complementing bus systems. Since the end of World War II, most of the urban cities in Australia changed drastically from small, close-knit cities, to relatively sprawling and auto-dependent. In the 50 years between 1955 and 2005, the number of registered cars increased from 2.5 million to over 11 million, while the number of citizens who depend on trains or buses has steadily decreased (Australian Bureau of Statistics 2005). As in the United States, there has been uptake in the use of urban transit and an acknowledgment of the role it can play in transportation and urban sustainability since the early 1980s.

China

Urban transit in China, through much of its history, consisted largely of informal, small-capacity modes like the rickshaw. Increasing urbanization and industrialization saw the arrival of some bus and streetcar systems in a few cities. In the latter half of the twentieth century, with China’s economic boom, came a rush of urban development that put rapidly growing pressure on its nascent urban transportation systems. Rapid, sustained urban growth and modernization have contributed to urban China experiencing American-style traffic conditions, featuring expansion in automobile travel and, with it, traffic and urban air pollution at scales and speeds far greater than what was seen in the developed world. For example, by 2004 the average speed of traffic on major highways around Shanghai had slowed to 10 km per hour, and this kind of congestion is now commonly seen in Chinese cities that were small towns only a decade ago. Many Chinese cities have begun to address these issues by developing and implementing innovative plans for expanding road infrastructure, subway and light rail systems, rapid bus transit vehicles that bypass automotive congestion, and magnetic levitation systems (Peng 2011). Today many citizens living and working within the urban core of a large city also depend on walking and a metro system to get around, though car use grows with an increasing standard of living.
**East Asia**

Japan and South Korea offer distinct urban transit environments characterized by their post–World War II redevelopment, limited urban land availability, and high population densities. Though they have extensive urban roadway networks and automobile congestion, they are also characterized by dense urban development and extensive railway and busway networks designed to serve these extraordinary densities. Their municipalities are also characterized by high economic standards of living, and there is no stigma attached to the use of mass transit as there is in some other developed countries. Rather it is seen as an essential part of the fabric of urban life. Other municipalities in Southeast Asia, such as Singapore, Hong Kong, and Kuala Lumpur also fit this description, featuring heavily used metros, high density, and relatively high standards of living.

**Central, South, and Southeast Asia**

Much of the rest of urban Asia doesn’t have the same level of economic development as China, Japan, or South Korea, and as such have a wide array of urban transit modes, which are influenced by the culture and policy within each country. Urban transit within bustling Southeast Asian cities is highly informal and more dependent on the private sector to provide it. Aside from rail systems, which are one of the consistently formalized transportation systems across Southeast Asia, modes of transportation range from rickshaws and three-wheeled covered bicycles, to buses and overnight trains. In addition to the cities of Southeast Asia mentioned earlier, the densely populated major cities of India also feature extensively used heavy rail systems, which have helped alleviate massive roadway congestion and feature a wide mix of public and private transportation modes. In places such as Vietnam, one of the most common short-distance modes of urban transit is by scooter, or “moto.” Shared taxis are also commonplace across southeast Asian countries (Dean 2011).

**South and Central America**

Though they lack the level of economic development and standard of living of their North American counterparts, many cities in Central and South America have extensive and well-organized urban transit systems, some of which offer better accessibility than similar North American cities. Governments in Central and South America are fraught with issues of institutional legitimacy, which hinder development in many of these places. Urban transit was once exclusively based on buses, private taxi systems, and private vehicles, but since the 1970s many cities have implemented BRT systems, trams, and aerial tramways, and have pioneered innovative solutions such as outdoor escalators to enable pedestrians to climb the steep hills of those cities embedded in undulating terrain more easily. Mexico City and Sao Paulo are examples of cities that feature extensive heavy rail systems, though much of Mexico City’s system actually runs on rubber-tired vehicles. Curitiba, Brazil, pioneered the BRT style of transportation in the 1970s when the city opted to build infrastructure for “long buses” instead of a subway or light rail system. This saved the city hundreds of millions of dollars and, alongside concomitant land-use policies, had a positive impact on urban congestion and development that many other cities have tried to emulate (Khana and Khana 2011). Outdoor escalators were installed in the slums of Medellín, Colombia, to enable citizens to walk to the city center easily (Henley 2013). Many cities across South America, like Bogotá in Colombia and Rio de Janeiro in Brazil, have begun to implement new and innovative urban...

---

*URBAN TRANSIT*

Japan and South Korea offer distinct urban transit environments characterized by their post–World War II redevelopment, limited urban land availability, and high population densities. Though they have extensive urban roadway networks and automobile congestion, they are also characterized by dense urban development and extensive railway and busway networks designed to serve these extraordinary densities. Their municipalities are also characterized by high economic standards of living, and there is no stigma attached to the use of mass transit as there is in some other developed countries. Rather it is seen as an essential part of the fabric of urban life. Other municipalities in Southeast Asia, such as Singapore, Hong Kong, and Kuala Lumpur also fit this description, featuring heavily used metros, high density, and relatively high standards of living.

Central, South, and Southeast Asia

Much of the rest of urban Asia doesn’t have the same level of economic development as China, Japan, or South Korea, and as such have a wide array of urban transit modes, which are influenced by the culture and policy within each country. Urban transit within bustling Southeast Asian cities is highly informal and more dependent on the private sector to provide it. Aside from rail systems, which are one of the consistently formalized transportation systems across Southeast Asia, modes of transportation range from rickshaws and three-wheeled covered bicycles, to buses and overnight trains. In addition to the cities of Southeast Asia mentioned earlier, the densely populated major cities of India also feature extensively used heavy rail systems, which have helped alleviate massive roadway congestion and feature a wide mix of public and private transportation modes. In places such as Vietnam, one of the most common short-distance modes of urban transit is by scooter, or “moto.” Shared taxis are also commonplace across southeast Asian countries (Dean 2011).

South and Central America

Though they lack the level of economic development and standard of living of their North American counterparts, many cities in Central and South America have extensive and well-organized urban transit systems, some of which offer better accessibility than similar North American cities. Governments in Central and South America are fraught with issues of institutional legitimacy, which hinder development in many of these places. Urban transit was once exclusively based on buses, private taxi systems, and private vehicles, but since the 1970s many cities have implemented BRT systems, trams, and aerial tramways, and have pioneered innovative solutions such as outdoor escalators to enable pedestrians to climb the steep hills of those cities embedded in undulating terrain more easily. Mexico City and Sao Paulo are examples of cities that feature extensive heavy rail systems, though much of Mexico City’s system actually runs on rubber-tired vehicles. Curitiba, Brazil, pioneered the BRT style of transportation in the 1970s when the city opted to build infrastructure for “long buses” instead of a subway or light rail system. This saved the city hundreds of millions of dollars and, alongside concomitant land-use policies, had a positive impact on urban congestion and development that many other cities have tried to emulate (Khana and Khana 2011). Outdoor escalators were installed in the slums of Medellín, Colombia, to enable citizens to walk to the city center easily (Henley 2013). Many cities across South America, like Bogotá in Colombia and Rio de Janeiro in Brazil, have begun to implement new and innovative urban...
transit infrastructure to alleviate congestion and provide accessibility for their rapidly growing populations and economies.

Africa

Africa’s urban transport systems, while well developed in countries like South Africa, are the least-developed transportation systems in the world. Many African cities have encountered a surge in growth with urbanization, which has led to the need for stronger infrastructure to handle rapidly growing need for mobility in their population. Most cities have not been able to keep up with this rapid growth and motorization, which leaves many roads unpaved, overwhelmed with car congestion, and lacking in traffic control. Rampant corruption and issues of institutional legitimacy also plague African urban transport management. Because of the infrastructure shortfalls in many countries, the use of large buses or rail transit is not feasible, so the use of minibuses, shared taxis, and motorcycles is promoted by cities (Kumar and Barrett 2008). Bus systems account for up to 44% of all transport in many cities, while personal vehicles account for only 7% of daily transportation. While bus transport is by far the most popular mode of motorized transportation, walking in urban areas accounts for 47% of daily travel within a city. There is also a great amount of walking in addition to many bus trips (Howe and Bryceson 2000).

The future of urban transit

Urban transit serves many different purposes across the world. In many American cities, it is little more than a social service for those who cannot afford to own an automobile. In the large dense cities of the developed world, it is a vital and frequently used component of a thriving metropolis. In much of Europe, it is a standard way of travel within a modern transport system that is coordinated with land-use management. In the burgeoning urban environments of the developing world, it operates through a diverse array of formal and informal forms that struggle to keep pace with a rapidly growing population.

Technological development has always influenced transit. Future developments of alternative fuels, and remote power and control of transit vehicles, should continue to permeate urban transit operations. The replacement of fossil fuels in transit operations with electrification, natural gas, and even hydrogen fuel cells continues, and in many aspects urban transit is far ahead of private automobile transport in advancements in the sustainability of its energy sources.

Perhaps the primary set of questions for transit in the near future involves operation and management. In the United States, political and economic trends have put increasing pressure on the ability of governments to fund essential operations such as transit. Urban transit also has to contend with a negative image and associations for much of its car-driving populace; continued investment in capital projects, technological advancement, and operation and management are therefore always an uphill battle. Many local governments have had success in attracting investment for fixed guideway transit projects to help revitalize urban development around its stops, but these serve a fairly limited subset of the population. Elsewhere in the developed world, urban transit is far more established and extensive, and land use is more suited to access by transit. Despite this, car ownership and use in Europe is rising, as preference for the convenience, luxury, and status of the private automobile grows despite considerably higher costs through taxation and the regulation of automobile use.
Lastly, the developing world provides what is likely the best opportunity for transit to serve a population and influence land use. Most cities in the developing world are growing at exponential rates. This growth into previously rural land and the redevelopment of underutilized land, coupled with an existing low standard of living and long-standing use of urban transit and other nonmotorized modes of transport, could provide a blank slate for a transit-oriented development that discourages automobile use. However, organized public transit faces an uphill battle in many urban areas in the developing world. Institutions in these places are often poorly organized and function extremely inefficiently. Car ownership is also widely seen as an essential status symbol and component of economic development and personal achievement. The population density in many of these places includes undesirable living conditions reminiscent of extreme versions of urban development during the Industrial Revolution, whereas suburban-style development is viewed as a positive and desired goal. The momentum of Western-style suburban, auto-oriented development in the developing world may already be so great that the outcome is inevitable and the latter no longer has the option of developing a transit-oriented urban model.

SEE ALSO: Mobility gaps; Rail transport: passenger; Road transport; Sustainable transport; Transport geography; Transport networks; Transport technology; Transportation and land use; Transportation planning

References


Further reading

Urban uneven development

Christopher Mele
University at Buffalo, USA

The concept of urban uneven development refers to the particular spatial expression at the city or metropolitan scale of a larger process of uneven development under capitalism. Uneven development is defined as asymmetrical capitalist social relations of production across geographic space at varying scales (hemisphere, nation, region, urban area, etc.). Central to the theory of uneven development is the principal connection between the investment of capital in one or more places over time: the concentration of capitalist growth and investment in one area (e.g., the rise of Sun Belt cities in the United States) requires a correlate underdevelopment in another (e.g., deindustrialization and the decline of Rust Belt cities in the United States).

Uneven development provides two essential interrelated functions to modern capitalism: as a sociospatial process of accumulation, tied to a broader theory of capitalism related to the production of space for purposes of surplus extraction, and for accumulation, through the production of space as a commodity itself within the second circuit of capital (investment in production). The uneven production of space for accumulation in the second circuit of capital is an important subarea of research in critical urban studies in the disciplines of geography, sociology, politics, and cultural studies. Research on urban uneven development focuses primarily on the production of urban space as a commodity for capitalist accumulation. This entry summarizes the two aspects of uneven development tied to accumulation that are the basis of the concept of urban uneven development, and discusses the driving roles of the state and real estate sectors.

Uneven development is a necessary precondition for the continued accumulation of capital and, therefore, the historical development of capitalism itself. The spatial form of development acts directly on the reproduction of relations of production: the spatial expression of land values, labor costs, technology, and transportation necessary for surplus value extraction and capital accumulation. Periodic crises in accumulation therefore entail both a social reorganization of production and the (re)development of new spaces for restructured surplus extraction. The unevenness in the social relations of production between spaces triggers investment capital to move in order to seek higher rates of return, resulting in urban uneven development. Recent technological innovations, the expansion in multinational corporations, and the restructuring of economic boundaries between nation-states have facilitated the acceleration in the global circulation of capital, information, commodities, and labor. The acceleration of flows enhances the prospects of accumulation crises and the restructuring of space. Global capitalism not only links different places across the globe to surplus extraction more efficiently but also subjects them to recurrent bouts of uneven development due to the ease of movements of capital.

The second, linked, component of uneven development emphasizes the production of space itself as a commodity for accumulation, subject to crisis. It developed out of a social justice perspective that took hold within critical geography in the wake of the global proliferation of
urban social movements in the late 1960s and
the spatial polarization and social inequalities in
Western cities in the 1970s. Since the 1970s,
uneven development has formed the theoretical
basis for critical research on suburbanization,
urban decline, and gentrification, thus connect-
ing urban development to the larger process
of uneven development. At the center of this
research agenda is the conception of urban space
itself as a core component of capitalist accumula-
tion in addition to acting as a space for the social
relations of production (as summarized earlier).
The critical geographer David Harvey’s core
contribution to urban uneven development is
his analytical linking of the social production of
space to the circulation of capital. To Harvey, the
key to the functioning of the capitalist accumu-
lation process is the ability of capitalists to shift
investments between three interrelated circuits
of capital. The primary circuit consists of finance
capital invested in commodity production, such
as manufacturing and services. The second
circuit is fixed capital, consisting of land and
the built environment. In the third or tertiary
circuit, capital is embedded in scientific research
and knowledge and in social expenditures for
the reproduction of labor power. Because of its
enhanced liquidity, investment in the first circuit
is optimal for the return of profit. But continued
investment in finance capital lowers productivity
rates and creates a crisis for profitability, prompt-
ing capitalists to seek a higher return of profit
elsewhere. Investment in the fixed capital circuit
provides an alternative – what Harvey refers to as
a “spatial fix”: the flow of investment capital into
the production of space. The reverse process is as
tenable; during periods of spatial disinvestment
capital “fixed” in the land and the built envi-
ronment can flow into the first or third circuits.
Alternatively, capital can remain in the second
circuit but shift to another space (geographic
switching). The free movement of capital in
search of higher investment returns produces
recurrent patterns of spatial unevenness; as one
place is subject to disinvestment and declines, the
other benefits from (re)investment and thrives,
and so forth. The operation of the second circuit
of capital provides the material basis for uneven
development, as capital is deployed in the spec-
culation, development, or disinvestment of the
built environment.

Building on the work of Harvey, the geogra-
pher Neil Smith’s research has added significantly
to the understanding of uneven development at
the metropolitan–urban scale. Smith’s theoretical
contributions are best conveyed through a
synopsis of his empirical analyses of the aban-
donment and gentrification of the urban built
environment. Urban uneven development in the
late twentieth century involved the systematic
incremental flow of investment capital out of
and back into the built environment. Given that
the built environment is a fixed commodity and
not prone to easy liquidity, capital flows tend
to occur “below the surface” (such as the spec-
culation in empty lots or deteriorated buildings
for later development) and sequentially (such
as reinvestment block by block over time). The
value of urban space to capitalists, according to
Smith, consists of its actual (at a given moment)
land rent and potential (at some point in time
in the future) land rent; in terms of investment
flows, the potential rent is the more important
of the two values. Smith refers to this difference
between actual and potential land rent as the rent
gap. Gentrification is a mostly invisible process
(from abandonment to speculation to actual
redevelopment) of closing the rent gap. Smith
showed how disinvestment and reinvestment in
urban space are intrinsically part of the same
process: the uneven development of urban space.
In addition to a political economic analysis,
Smith pointed to the importance of social and
cultural factors that mediate or accelerate urban
uneven development, including state policies
relating to the urban poor and homeless and grassroots social struggles and resistance to displacement pressures, showing that urban uneven development is neither a mechanistic nor an inevitable force of capitalism.

Urban uneven development is a process contingent on the sometimes volatile interplay between structural forces (e.g., government policy, land values) and various agents and stakeholders with different interests (e.g., existing residents, gentrifiers, and neighborhood groups). The process of “switching” in the second circuit of capital involves a number of institutions and actors, including developers, speculators, investors, realtors, government officials, and financial institutions. These actors can form growth coalitions to advance their collective interests in development. Cohorts of residents (e.g., working-class renters, early gentrifiers, artists, students, middle-class homebuyers, etc.) participate in uneven development as stakeholders in favor of change or in resistance to it. State intervention in urban uneven development facilitates investment capital flows through various direct and indirect actions, including financial subsidies (e.g., tax abatements, eminent domain, issuing bonds, etc.), governance (e.g., police patrols, clean streets, etc.), administrative support (e.g., zoning variances, building permits, etc.) and infrastructure projects (e.g., construction of roads, sidewalks, etc.).

Recent work in critical geography on neoliberal urbanism places urban uneven development at the center of its analysis of urban political economy. Neoliberalism is an ideology and a mode of governance based on the primacy of free-market principles, including privatization, entrepreneurialism, and enhanced competition between places, to further capitalist accumulation. Neoliberal urbanism refers to governance practices that privilege market mechanisms, including limiting state oversight and regulation of the private sector, reducing expenditures on social welfare, and forming public–private growth partnerships, to create conditions favorable to capital reinvestment. The increased dependence on market forces in urban governance and the expanded control of the private sector in the planning, financing, and implementation of local economic growth has enhanced urban uneven development. Greater competition between localities to attract investment capital (e.g., development incentives and subsidies) enhances urban uneven development and the potential negative social and economic outcomes for communities.

**SEE ALSO:** Cities and development; Critical geography; Rural/urban divide; Suburbanization; Urban redevelopment

**Further reading**


User-centered design is a design methodology and philosophy in which the needs, goals, and success of the end user are considered. User-centered design can be applied to any type of system, object or product intended for human use, but the term is used most frequently in connection with computer information system design.

In user-centered design, the ease with which a system can be understood by its intended users, how easily they can complete tasks with it, and how satisfied they are with their use of it, become the key measures of the success of the design. User-centered designs typically require less documentation and training to be successfully used because they have been designed specifically to match the expectations and abilities of the users. Although user-centered design is closely associated with the discipline of user interface design, it can cover almost every aspect of a system, not only its user interface. Elements of a system such as its documentation, how data are entered and structured within the system and made available for use, how it is deployed and presented within an organization, and how the system use fits into the business goals of an organization, can all have user-centered design applied to them.

The term “user-centered” design was made popular by Donald Norman, a cognitive scientist working in the field of system and product usability. He first referred to “user-centered system design” in 1986. His bestselling book, The Design of Everyday Things (Norman 1988) (originally published as the Psychology of Everyday Things), introduced the concept to a wide audience.

Successful user interfaces meet users’ expectations. The application of human factors and ergonomic design principles result in displays and controls that are clearly visible and easy to interact with. User-centered design ensures that displays and controls make sense to the user. A key tenet of user-centered design is that to be intuitive the user interface should match the user’s mental model of the task they are performing.

Norman gives the compelling and memorable example of the cooktop found in most kitchens. Cooktops typically have four cooking rings: two at the front nearest the user, and two at the back closer to the kitchen wall. However, the knobs with which the user controls the rings are often arranged in a straight line. This is a bad user interface for this system because there is no immediately obvious relationship (also called “mapping”) between the location of the knob and the location of the ring it controls. The spatial arrangement of the knobs does not match that of the rings they control. This basic user interface design flaw makes it hard to use the system and causes the user to frequently turn the wrong ring on.

User-centered design recognizes that the user’s ability to understand how to use controls for the cooktop’s rings is an important design problem that must be solved before the cooktop is put into production for users to struggle with. When the
Figure 1  A cooktop with a user interface that is not user-centered. There is not an obvious spatial relationship between the rings and the knobs that control them. Icons are usually added next to each knob to try and make this clearer to the user. (Diagram by Rupert Essinger after Norman.)

four controls are presented in a linear arrangement, there is no obvious relationship of which one controls which ring (Figure 1).

In a kitchen cooktop in which user-centered design has been applied, the controls are arranged as a group in the same pattern as the rings: two in the front and two at the back. This natural mapping of the position of the controls to the position of the rings is immediately understandable by the user because it matches the task as they see it (Figure 2).

When a user interface is not user-centered, documentation is required. In the cooktop with the linear controls, the documentation is usually provided in the form of little icons under each knob designed to show the user which ring each knob controls. These icons are not needed when the knobs are arranged in the same way as the rings: two in the front and two in the back. This illustrates another key tenet of user-centered design: when user interfaces match the user’s model successfully, less documentation is required. The goal of user-centered design is to create systems that are successful because they are intuitive and self-explanatory, and that therefore reduces the need for end-user documentation and training.

It may seem obvious that any system aimed at users should be designed with the user in mind. And especially now that the term “user-centered design” is well known, it is unlikely that a system engineer or designer would admit that their system has not had some amount of user-centered design or at least user-centric thinking applied to it. But modern system development processes vary in the extent to which they involve user-centered design techniques. These techniques include persona-based or role-based design where the needs and workflows of typical users are considered, usability testing of prototypes and development versions of systems with sample users, the involvement of designers with specific usability and human interaction expertise and qualifications, and successive refinement based on feedback from early adopters and beta testers.

User-centered design techniques

Applications and systems may not have one idealized user. User-centered designers realize that they are often not designing for one user, but for a number of different types of users. An application’s users will often be segmented by designers into smaller role-based groups, with specific needs based on the information they require and the tasks they want to perform.
Sometimes those needs may pull the user interface design in different directions (Garrett 2003). For example, novices learning new work tasks may find it easier to learn a new application if it presents its workflow as a sequence of simple steps. Experts who have mastered getting the work done, however, may view the sequence of steps enforced by a linear user interface as tedious. These expert users may need a user interface that provides rapid access to a breadth of functions the application provides.

User-centered design recognizes that it may not always be possible to meet both sets of user needs with a single interface solution. One option is to provide different ways for the different user groups to accomplish the same task. Another option is to focus on one group of users and exclude support for the other. This kind of decision has consequences on many of the other choices the designer will need to make as they craft the user interface.

Once an application’s users have been identified, the central focus of user-centered design is to provide a solution that solves the users’ real problems. The first step involves researching and understanding the activities and tasks that the application’s users want to accomplish. This is followed by identifying software features that support those activities. If a feature does not address a real user activity, it does not make it into the user interface design specification. Features that support key user activities get the highest priority in the design specification, and are expected to be both easy to find and easy to accomplish in the user interface.

User-centered design recognizes that users are much better at providing critical feedback in response to a prototype user interface, that is, something they can see, than they are in providing up-front detailed descriptions of what they need or seek. To address this designers often jump-start the process by beginning with nonfunctional mock-ups or storyboards illustrating what the screens could look like, and then change them in response to the users’ feedback. The benefits of starting with mock-ups or storyboards include:

- rapid, iterative prototyping to support key user tasks and “use cases”;
- prioritization of features and workflows before actual software coding starts;
- users and subject matter experts may feel more comfortable giving feedback on mocked-up prototypes rather than systems that are already in development.

After the mock-ups are implemented, the application will often go through additional usability testing and acceptance testing before it is released, with successive iterations used to refine the user interface and functionality.

Once the application is completed and deployed, user-centered design can still continue. The ability of users to perform their tasks and activities with the system can be reviewed and assessed. Plus, as users gain more experience with the application, they can provide practical feedback and requests for improvements reflecting their actual usage. This is extremely beneficial as it:

- helps focus the priorities and attention of the development team, and is a key input in the planning and design of future iterations of the application;
- supports the refinement and development of any complementary documentation and training both for users and technical support specialists working with them;
- enables managers and user representatives to discuss enhancements to the application in terms of user tasks and roles rather than technical software features;
USER-CENTERED DESIGN

- allows, once the application is completed and deployed, user-centered design to still continue.

There is a large body of literature about user-centered design, its methodologies, and user interface design. There is also an International Organization for Standardization (ISO) standard dealing with the ergonomics of human-system interaction (ISO 2008).

User-centered design in geographic information systems (GIS)

User-centered design can be applied to geographic information systems (GIS), GIS applications, online mapping systems, and websites designed to provide access to maps, geographic information, and data. A successful GIS implementation involves not only the use of GIS software, but database design, data acquisition, the development of workflows, best practices and management, data publication, project-based analysis, cartographic design, and application development. User-centered design can be applied to each of these key areas with the goal of ensuring that a system fits with the needs of the people who will use it, rather than the other way round. For example, when an organization plans how data will be acquired, collected, and maintained in its GIS, it will want to ensure that these workflows and policies are easily understandable by the people who will be performing them, and the tools and training provided to different personnel based on their roles, from data collection in the field to database administrator, are successful.

SEE ALSO: Cognition and spatial behavior; Geographic information system

References

Validity and verification

Ashton M. Shortridge
Michigan State University, USA

The late statistician George Box famously noted that, “all models are wrong, but some are useful.” This perspective is a critical one for understanding validity and verification (V&V). V&V engage the general problem of determining how closely data and models reflect reality. This problem is a central scientific concern: it encompasses philosophical issues, representation challenges, sampling and measurement concerns, and forms the basis of inferential statistics. This entry considers V&V within the spatial domain, a context that introduces particular opportunities and challenges. While it is convenient to separately discuss the validation of data and models, the distinction is somewhat artificial: the consensus in geographic information science is that data are the consequence of a considerable modeling process.

Data

The terms “validity” and “verification” are used in the context of spatial data production, and are closely tied to data quality standards. Validity is the degree to which the produced dataset meets agreed-upon data quality standards. Validation is the process by which this is done. Verification is typically reserved for efforts to assess whether data were produced following an agreed-upon protocol, and is concerned with the sources and methodologies used in the data production process. Taken together, V&V are used to evaluate the positional and attribute accuracy of spatial datasets as well their completeness and consistency, all key elements of spatial data quality.

Data validation

In the spatial data quality standards domain, accuracy assessment is a validation issue (Veregin 1999). The US National Standard for Spatial Data Accuracy (NSSDA) specifies a generally applicable validation approach to assess horizontal and vertical positional error from a large feature dataset.

1 Select a representative subset of test locations from the data.
2 Obtain an independent dataset of higher accuracy (“reference”) for those locations.
3 Calculate an accuracy metric based on the differences between the data and the reference positions.

As an illustration of the NSSDA standard validation procedure for horizontal positional accuracy, consider a dataset of point locations representing the centers of 346 vernal pools, as interpreted from digital aerial imagery. Vernal pools are ephemeral wetlands that provide important springtime habitat for many kinds of plants and animals. In the first step, 20 of these pools are randomly selected, for a sample size just larger than 5%. In the second step, the selected locations are visited, and the geographic coordinates of each pool’s center are carefully surveyed. This set of observations is the reference. The coordinates from the dataset and from
VALIDITY AND VERIFICATION

the ground survey for these pools (showing just five to save space) are presented in Table 1.

Accuracy metrics summarizing the horizontal positional error distribution include the sample mean error, the mean squared error, and the root mean square error (RMSE) of the sample. RMSE (Equation (1)) is a widely reported metric. This is the square root of the mean squared error, in the horizontal units of the coordinates, for the N sample points. The larger the RMSE, the greater the discrepancy between the data-reported and reference positions. If the error is unbiased, RMSE corresponds to the standard deviation of error for the sample (note that by custom the denominator is $N$ rather than $N-1$).

$$
\text{RMSE} = \left( \frac{\sum (x_{data} - x_{reference})^2 + (y_{data} - y_{reference})^2}{N} \right)^{0.5}
$$ (1)

If one assumes positional errors are normally distributed in $x$ and $y$, then the 95% horizontal error can be calculated (Equation (2)). This value represents the radius of a circle that, when centered on each coordinate in the dataset, can be expected to contain the actual location for 95% of the points, providing the underlying assumptions are met. This accuracy measure is called the NSSDA in the United States spatial data quality standards community and is commonly reported in metadata; it has been criticized for relying on assumptions which may frequently go unmet (e.g., normality and independence of errors, random sampling) (Zandbergen 2008).

$$
\text{NSSDA 95\%} = 1.708 \times \text{RMSE}
$$ (2)

For nominal data such as land cover/land use, a similar approach is used: reference information from another source, such as ground survey or higher-accuracy classified imagery, is obtained for a representative sample, and summary metrics based on class error or confusion are calculated (see Modeling uncertainty in categorical fields).

Table 1  Vernal pool test point coordinates, reference coordinates, and positional errors. All measures are in meters. Squared differences in $x$ and $y$ are reported for each test point, as well as the sum of squared differences. Summary statistics in the bottom right report the mean square horizontal error, the root mean square error (RMSE), and the NSSDA 95% statistic.

<table>
<thead>
<tr>
<th>Rec</th>
<th>$x$</th>
<th>$y$</th>
<th>$x_{Ref}$</th>
<th>$y_{Ref}$</th>
<th>$Diff_x$</th>
<th>$Diff_y$</th>
<th>$Diff_{x^2}$</th>
<th>$Diff_{y^2}$</th>
<th>$Diff_{x^2+y^2}$</th>
<th>$D_{x^2} + D_{y^2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>248551</td>
<td>934559</td>
<td>248580</td>
<td>934528</td>
<td>-29</td>
<td>31</td>
<td>853</td>
<td>967</td>
<td>1820</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>205000</td>
<td>898044</td>
<td>205013</td>
<td>898054</td>
<td>-13</td>
<td>-10</td>
<td>169</td>
<td>102</td>
<td>271</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>195811</td>
<td>931726</td>
<td>195783</td>
<td>931719</td>
<td>29</td>
<td>7</td>
<td>817</td>
<td>49</td>
<td>865</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>204800</td>
<td>898837</td>
<td>204799</td>
<td>898811</td>
<td>0</td>
<td>27</td>
<td>0</td>
<td>702</td>
<td>703</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>205010</td>
<td>895545</td>
<td>205010</td>
<td>895521</td>
<td>0</td>
<td>24</td>
<td>0</td>
<td>591</td>
<td>591</td>
<td></td>
</tr>
</tbody>
</table>

Mean 586.19

RMSE 24.21

NSSDA 41.90

All measures are in meters. Squared differences in $x$ and $y$ are reported for each test point, as well as the sum of squared differences. Summary statistics in the bottom right report the mean square horizontal error, the root mean square error (RMSE), and the NSSDA 95% statistic.
Measures such as RMSE, 95% error, percent correctly classified, and related metrics are global indicators of validity. Local properties, including spatial or other patterns in the degree of positional accuracy, are not captured by these metrics, nor are other metrics that could assess local validity used or reported by mandate through metadata standards.

Data verification

While accuracy assessment is a validation issue, other aspects of data quality fall under the purview of verification. Verification is an integrated component of the data production process for large geospatial data producers, and is concerned with ensuring that the data are developed to align with established protocols. Core data characteristics that can be verified include the following:

**Resolution:** data may have spatial, temporal, and spectral (in the case of imagery) resolutions. Each of these dimensions plays a key role in determining the scale at which features or processes may be characterized, and it is clearly important for data to offer resolution commensurate with the desired process scale. Spatial resolution is straightforward to report for raster data structures given its relationship to pixel or cell size, though the effective resolution may be coarser than the cell size. For vector data, scale or resolution of the source of the digital data (e.g., paper map, imagery) may determine the resolution, or digital resolution may be dictated by the effects of subsequent processing.

**Consistency:** this term is concerned with the degree to which key data characteristics match their intended properties. If a vector data product is expected to obey certain topological rules (e.g., planar topology, in which every line crossing must have a node, polygons cannot overlap, and so on), then the data must follow those rules. Logical consistency of attributes implies that stored thematic properties of features should not conflict with one another.

**Completeness:** data are intended to represent some real-world features. Completeness considers the degree to which all real-world features, as specified in the data model ontology, are represented by the dataset. This feature specification is typically quite constrained. Consider a “building footprint” polygon dataset in a municipal GIS database. Buildings eligible for inclusion would be constrained: only those within the city limits, only those present in imagery from a particular year, and only those above some minimum area might be included.

Like validation output, verification results generated by large public data producers are intended both for internal quality assessment/quality control (to determine, for example, if data products generated by an external contractor meet specific standards) and for external communication to data users via metadata, to satisfy the “fitness for use” criterion. Metadata standards for reporting V&V findings are generally mandated by government and therefore can vary from country to country and also across regional and local jurisdictions; however, there is substantial consistency among many of them. These mandates can indirectly influence V&V-based data quality reports for nongovernmental producers as well.

Models

Few words have as many distinctive meanings as *model*, but here the term is defined as a computer-based simulation of one or more social
or environmental processes. A model attempts to capture either the key mechanisms of the process or the characteristic patterns manifested by the process. Models ingest data and parameter inputs and produce an output intended to reflect the actual process. Models may be dynamic: such models reflect behavior over multiple periods of time, and may therefore seek to characterize changing relationships and patterns over that time. Regardless of the type of model, its output may be regarded as data – often spatial data – and can serve as inputs to other models. Figure 1 presents a generalized computational model. It might be a hydrologic model characterizing both surface and groundwater flow within a stream basin over time. In that case, input data include precipitation time-series data, a digital elevation model (DEM), and soil data. Model 1 characterizes the flow and infiltration properties of the basin, while Model 2 characterizes surface and groundwater response to specific precipitation inputs, producing output maps of properties of interest (e.g., saturation, flow rate) at particular intervals.

In many computational modeling contexts, validation and verification may refer to distinct aspects of model performance assessment. However, these terms are not used consistently across all scientific domains: in weather forecasting, verification can refer to the comparison of model output to what actually occurred, while in statistical modeling, validation is reserved for this context. In algorithm assessment, verification is an evaluation process that determines if the algorithm has been implemented correctly, while validation is reserved for determining if the algorithm is appropriate for the application task. Similarly, the hydrologic modeling community uses validation to refer to the comparison of model output with reference data, and verification to refer to the assessment of the correctness of the model’s implementation. The latter definitions will be used in this entry, as they parallel use of the terms in the data section. Figure 2 characterizes the modeling process and the role of validation and verification in it. The simuland is the environmental or human process that is to be modeled. A conceptual model is derived for this process. The conceptual model is implemented as a computational model. This model is used, with input data and parameters, to generate results. Validation is the comparison of the results with the simuland, typically through reference data sampled from the simuland. Verification is the
Figure 2  Verification and validation in the modeling process. Adapted from Figure 10.2 in Petty 2010, with permission of John Wiley & Sons, Ltd.

comparison of the computational model with the conceptual model.

Model validation

Validation of spatial models can be done by pointwise comparison of modeled output to reference data, in a similar fashion to data accuracy assessment. This procedure is straightforward for any modeling context where reference data is available, such as high-resolution weather forecasts. In addition to RMSE, the coefficient of determination, or R², is often used as a model validation metric (Arnold et al. 2012). Let \( \{y_1, y_2, \ldots, y_n\} \) be reference measures at \( n \) locations, and \( \{m_1, m_2, \ldots, m_n\} \) be the model predictions at the same \( n \) locations. Then \( R^2 \) measures the ratio of the sum of squares of the residuals to the total sum of squares of the data (Equation (3)).

\[
R^2 = 1 - \frac{SS_{res}}{SS_{tot}} 
\]  

(3)

where:

\[
SS_{tot} = \sum (y_i - \bar{y})^2
\]

\[
SS_{res} = \sum (y_i - m_i)^2
\]

\( R^2 \) varies from 0 to 1 and indicates the goodness of fit of the model predictions to the reference observations.

However, pointwise approaches ignore spatial components of accuracy that have important consequences for verification. Considerations such as the spatial autocorrelation of model errors may undermine classic assumptions of independence of model residuals. Further, the match of scale-dependent spatial structure of predictions to reality, as well as the geographic displacement of a modeled spatial feature from its actual occurrence, is not well-handled by pointwise measures (Gilleland, Ahijevych, and Brown 2009). As an illustration, consider a forecast for a storm event. Suppose that the storm does in fact occur at the indicated time with the predicted magnitude, but its position is just to the north of the predicted location. A pointwise accuracy measure such as RMSE or \( R^2 \) will indicate that the model has performed quite poorly, while other important but unassessed aspects of the forecast capture reality with good fidelity. To account for this, spatial filtering has been used with comparison metrics to validate model performance for field-based output. Alternative, object-based approaches determine the degree of spatial deformation needed to most closely match model-generated feature positions to reference, or use fuzzy, rather than crisp-set comparison methods. Figure 3 illustrates a representative spatial validation typology of field- and object-based validation methods.

Many models use empirically derived parameters that capture key environmental properties (e.g., a hydrologic model may use soil particle size, vegetation cover type, and surface slope to drive flow speed, infiltration rate, and other characteristics). Realistic values for these parameters may vary over space and time, and may in fact be unique to a particular context. For example, output of the Schelling Model,
an agent-based model that characterizes urban segregation, is affected by many environmental and household parameters (Crooks, 2010). *Model calibration* is the process of identifying reasonable parameter values, while validation of those parameters generally requires reserving a subset of observations from the calibration process for independent assessment.

**Model verification**

Verification refers to the implementation of the conceptual model, typically in a computational
setting. Representative verification questions include the following.

- **Does the algorithm reflect the process described in the conceptual model?** In hydrologic models, for example, equations reflecting the physics of water flow across surfaces (“overland flow”) and within streams (“channelized flow”) are commonly implemented. The computational model must correctly implement and solve the equation. This may be difficult to assess when a heuristic is used: an approximation of the conceptual model in situations where solving the model might be computationally complex, time-consuming, or impossible.

- **Was the algorithm coded correctly?** This is a central concern in software engineering, and a range of procedures and approaches have been developed to address it. Logical errors, especially in contexts outside of the typical range of inputs and conditions, are particularly difficult to identify and correct.

Verification is an essential element of model development. Typically it is evaluated by implementing the model on synthetic data with known solutions. The boundary between verification and validation is particularly thin for activities such as parameter assessment.

Models are typically complex, with multiple input datasets and starting parameters. The model may be well validated and verified but still perform poorly if the input data are not also carefully assessed via V&V approaches. Input data fitness for use must be carefully assessed: if critical factors such as accuracy or spatial resolution do not match model specifications, it is unreasonable to expect quality model results. As the old saying goes: garbage in, garbage out.

**SEE ALSO:** Data quality standards; Geographic information science; Metadata; Modeling uncertainty in categorical fields; Modeling uncertainty in digital elevation models; Spatial sampling; Uncertainty; Visualizing uncertainty; Volunteered geographic information: quality assurance

**References**


**Further reading**

VALIDITY AND VERIFICATION


Value added

C. Michael Wernerheim
Memorial University of Newfoundland, Canada

Value added is a measure of the production that occurs within an industry. It bears a close relationship to the gross domestic product (GDP), as GDP can be estimated as the sum of values added for all producing entities in the economy. GDP is the most widely used indicator in national accounts, as it summarizes in a single estimate the value of the output (goods and services) produced by all firms, nonprofit institutions, government bodies and agencies, and households in a given country or region during a given time period. The concept of value added makes a profound contribution to the computation of GDP in that by design it eliminates the double-counting that would otherwise occur in the aggregation of production from the micro (firm) level to the macro (national) level. GDP is above all else an instrument for measuring the magnitude, growth, or decline of the economy as a whole, and for determining the economic process that accounts for the composition of the national product. In value added, national accountants thus have a measure of GDP that is at once theoretically meaningful and empirically measurable.

Measuring the output of a single firm is comparatively simple. Measuring the output of an entire country is a complex matter that requires innovative concepts and the use of a vast array of data. Consider first the value added concept at the firm level. If all firms in a given industry produced all the inputs they used in production, the ratio of value added to sales would be one. This is generally not the case. Instead, they rely in varying degrees on firms in other industries for inputs (raw materials, intermediate and/or semi-fabricated goods and services), implying a ratio of value added to sales that is less than one. The value added (or production) of each firm in each industry may thus be aggregated across the entire economy. However, if purchased inputs were not subtracted from the firm’s sales prior to aggregation, double-counting of inputs would occur. To see the importance of this, consider next the estimation of output at the macro-level.

To begin, there are three equivalent approaches to GDP: the output (production) approach; the final demand (expenditure) approach; and the income (supply) approach. The important point here is that all three measures give the same total for GDP. (For a more detailed discussion of GDP, the reader is referred to the reference sources shown.) The output approach is best suited to explicating value added, as it defines gross domestic product as:

\[
\text{GDP} = \sum \text{value added}
\]

As value added in turn equals production minus intermediate inputs, then:

\[
\text{GDP} = \sum \text{production} - \sum \text{intermediate inputs}
\]

That is, rather than crediting each firm with the value of its sales (or output), national accountants count only the value that the firm adds to that of businesses supplying its intermediate inputs. By deducting the portion of intermediate input usage from that of sales ensures that the input portion is not counted more than once when the output produced by one industry is re-sold as intermediate input.
to another industry. The result is a measure of the value that is “added” to each product or service by firms at each stage of domestic production.

To illustrate, take the example of a smartphone manufacturer. Value added is the revenue from sales of smartphones less the amount the manufacturer pays for plastics, metals, electronics, circuit boards and other hardware, producer services such as software and engineering services, and other inputs. The way in which sales of the intermediate inputs “wash out” in the computation of GDP according to the production or value-added method is described by Hall and Taylor (1986, 35) as follows: “If a firm sells a final product, the sale appears in that firm’s value added but nowhere else. By contrast, if the firm sells its output as an input for another firm, that sale appears negatively in the other firm’s value added. Thus, when the two firms are added together in computing GDP, sales of intermediate products effectively cancel out.”

Furthermore, it follows that by eliminating double-counting, the resulting measure of aggregate value-added-based production is also independent of how firms are organized, that is, independent of the number, size, and type of economic units that account for national output. This matters because corporate structure and industrial organization are subject to change over time.

It must be emphasized that precise measurement of economic activity on a large scale is virtually impossible; hence, partial data, estimates, and statistical methods are necessary to approximate economic activity. As suggested earlier, the method for estimating GDP thus involves considerable reliance on large numbers of surveys, and administrative (secondary) data initially collected and compiled by central governments for other purposes. Sources include private industry associations, public statistics agencies, income tax returns, value added tax (VAT) declarations, regulatory reports, imputed rents, business and government transfer payments, corporate profits, depreciation, records of employee compensation, dividend income, income statements by firms, and the value added by the services of general government. The uneven quality of this data gives rise to inevitable measurement errors and, hence, a statistical discrepancy across the three independent measures of GDP. Technically, the statistical discrepancy is the “product side” minus the “income side” of national accounts.

As for the theoretical equivalence of the product side and the income side of the accounts, it should be noted that the income side includes only those transactions (i.e., events where money changes hands) involving the generation of value added. National income excludes, for example, procurement of raw materials, loans, gifts, and government transfers. The reason in this case is that these activities do not generate net output in the domestic economy, even though these and similar activities do, at some point, involve the exchange of money.

Similarly, domestic product does not include sales and purchases of existing assets (real or financial), as it is defined on a value-added basis. Domestic product is thus an index of the total volume of newly produced goods and services in the domestic economy. The present definition of GDP is internationally agreed upon, which explains its extensive use by policymakers, researchers, and the media at large. Finally, GDP is estimated annually or quarterly, but it can also be obtained monthly.

**Value added tax**

An important use of value added is in the calculation of a widely implemented tax often referred to as a value added tax. A VAT is a so-called
“multistage” tax calculated as a percentage of the value of a firm’s sales minus the tax the firm has paid on its purchases of inputs produced by other firms. In effect, the firm thus pays tax on the value it has itself added to these inputs, and not on the value created at earlier stages in production (on which the tax has already been paid). VAT is an important source of tax revenue for governments in many countries worldwide. In European countries, for example, they are typically levied at rates exceeding 20%. Although such rates can appear high, most tax experts argue that consumption taxes (VATs) are preferable to income taxes in that the former discourage saving and investment less. This is because the latter tax income twice; first on the original income, and then again on any earnings from savings and investment of that income. This matters greatly as private savings and investment underpin economic growth.

Value-added services and location

It has been argued that lower value-added corporate functions, such as processing activities, logistics, and supply functions, are more likely to relocate abroad, typically to emerging economies, whereas high value-added service functions, such as R&D, are more likely to locate in advanced economies (Daniels 2007). This centralization of high value-added services in global cities, and the decentralization of low value-added services into the global low-wage periphery, is evidenced by analysis of global commodity chains, various types of studies of agglomeration economies, and also by sectoral studies of back offices, call centers, and offshore banking. The common denominator is the processing of standardized information and the use of relatively unskilled labor. These empirical observations and findings have also been predicted by various theoretical approaches, such as actor-network theory, which purports to explain the global division of labor and the simultaneous centralization and decentralization of different kinds of value-added services. A central conclusion of actor-network theory is that, notwithstanding the geographical segregation of value-added services on the basis of their value-added content and the types of skills and knowledge utilized, both high and low value-added services are embodied in people and embedded in a complex mix of local and international contexts. The result is a restructuring of the service economy globally, resulting from, in part, novel means of production and delivery via telecommunications networks (Warf 2007).

Parenthetically, it should be noted that R&D has historically been treated in national accounts as current expenditure on intermediate inputs and has, therefore, not entered GDP. However, starting with the current version of the system of national accounts (2008 SNA) (EC et al. 2009), R&D is treated as capital formation, except where the activity does not benefit the owner, in which case it continues to be treated as intermediate consumption. This development promotes the role of services in the modern economy. Services are now increasingly seen as a key component of global value chains, as they facilitate value-added activities in different markets. However, this enhanced role played by producer services gives rise to new conceptual problems in measuring the value added of such services. This is because they increasingly form part of entire complex platforms of highly integrated manufacturing and service functions. The value added by the services components can thus be difficult to disentangle and measure in such applications. The value added by producer services is difficult to measure also in situations where these services are not used as one-off inputs in the production process but may be
VALUE ADDED

re-used, perhaps in an entirely different context. Examples include various consultancy services and market research.

Value added and industrial structure

The greater the reliance of a firm in a given industry on firms in other industries, the lower will be the ratio of value added to sales, and vice versa. In the industrial organization literature, higher ratios of value added to sales are, therefore, sometimes seen to reflect greater vertical integration – a measure of the vertical linkages in the supply chain, achieved either by forward or backward merger or investment in a given industry.

Value added and income distribution

In light of the preceding discussion, it should be understood that value added per unit of labor and wages, though correlated, are not the same. All else being equal, the higher the value added per worker, the higher the wage rate, and vice versa. More to the point, the value added typically exceeds the wage. This follows from the way these measures are computed. Since labor and other inputs (such as capital) cooperate to produce value added at each stage of the production process, the computation of value added per worker must include the payment to labor (the wage bill), as well as the return to owners of capital. Introducing a small amount of notation serves to clarify this point. Under perfect competition, profits in excess of a normal return to entrepreneurship (an opportunity cost) are zero by definition, and total revenue equals total cost.

Write this as:

\[ p \cdot Q - (w \cdot L + r \cdot K + v \cdot M) = 0 \]

where price \( p \) times output \( Q \) represents total revenue, the wage rate \( w \) times the labor input \( L \) denotes the wage bill, the rate of return to capital \( r \) times the capital usage \( K \) denotes the payment to capital, and the per unit cost of intermediate inputs \( v \) times input usage \( M \) represents the corresponding factor payment. By definition, value added is total revenue less the cost of intermediate inputs, as shown in equation 1.

\[
VA = \sum production - \sum intermediate inputs = pQ - vM = wL + rK
\]

Value added per worker is thus:

\[
\frac{(w \cdot L + r \cdot K)}{L} = w + r \cdot K/L > w
\]

That is, value added per worker exceeds the wage. Moreover, the more capital intensive the production, the greater is this effect all else being equal.

Value added: further applications

It should be clear from the discussion that value added plays a major role in the system of national accounts and, consequently, in a host of related applications, such as the national and regional input–output tables (Isard 1960; Leontief 1986). The concept of value added has important uses in other areas as well. For example, in regional analysis it is sometimes used as a base to compute the so-called location quotient – a measure indicating the location (or a region’s share) of an industry relative to the geographic distribution of its labor productivity in the aggregate (Isard 1960).

SEE ALSO: Factors of production; Firm foundation and growth; Industrial linkage;
References


Further reading


Venture capital

Jun Zhang
University of Toronto, Canada

Venture capital (VC) is a special form of financial institution designed to provide financial support and business assistance to emerging firms with high growth potential, particularly in technology sectors (Gompers and Lerner 1999; Cumming 2010). Venture capital firms are professionally managed equity investors: they provide money and value-added services to emerging firms in exchange for an equity stake in the business, expecting the market value of the shares to grow exceptionally quickly in the time frame of a few (typically three to seven) years. Returns to VC are expected to be obtained through a “liquidity event,” ideally in the form of an initial public offering (IPO) of the business in a stock market or a trade sale: a merger or acquisition.

Venture capitalists are highly selective, usually on the lookout for some extremely rare combination of innovative technology, a well-developed business model, and an impressive management team, providing an emerging firm with the potential for rapid growth. Such opportunities are most likely in technology sectors, and VC thus has become almost synonymous with technology finance in sectors such as computers, the Internet, semiconductors, telecommunications, electronics, biotechnology, clean technology, and health care. Given the high level of risk, and that investments are illiquid and require an extended time frame for realization, venture capitalists usually carry out detailed “due diligence” analysis prior to investing. They also nurture investee firms in many ways to help them move faster in developing products and bringing them to market.

Economic geographers have become increasingly interested in VC, because it is believed to be a key engine driving technological innovation and entrepreneurship. It is also an area of increasingly active public policy aimed at creating or stimulating VC funds to boost technological innovation and economic development, especially in laggard regions. Venture capital is characteristically a highly localized practice, whose investments have been highly concentrated in a few hotspots such as Silicon Valley. At the same time, VC is highly mobile and has become increasingly globalized. This intricate interaction between spatial centralization and decentralization has posed new challenges for geographers.

Fundamentals of venture capital

Venture capital can be defined as dedicated pools of capital managed by independent, professional institutions that focus on equity or equity-linked investments in young firms with expected high growth potential. This definition distinguishes VC from three other forms of investment: (i) nonventure private equity investment, in the form of buyouts, restructure, and mezzanine funds targeting relatively mature firms with stable cash flows and limited growth potential; (ii) public equity investment in mature firms with high liquidity and low risk; and (iii) individual angel capital, commonly understood as wealthy private individuals who use their own money to provide the first round of external
VENTURE CAPITAL

finance and hands-on business support to get a company established. Institutionalized VC firms can also play the role of angels by offering seed money and entrepreneurial support to budding firms. In Europe the terms “private equity” and “venture capital” are often used interchangeably, but it is important to differentiate them. Venture capital funds are managed by investors with entrepreneurial experience and industrial knowledge (frequently with a related technological background). They invest in early-stage ventures, actively participating in their investee companies to provide value-added services. By contrast, private equity funds are managed by individuals with a background in investment banking or other financial organizations, and with short-term investment horizons. Whereas VC funds are heavily involved with the start-up companies in their portfolio, in both a monitoring and an advisory capacity, private equity firms, working with more mature firms, do not have such imperatives. Private equity investors are more susceptible to speculation, and face much fewer spatial constraints than venture capitalists. Thus they imply entirely different geographies.

Most VC firms are structured as limited partnerships, with general partners (GPs) who manage the firm and serve as investment advisers for the VC funds raised. These GPs and related investment professionals are often referred to as venture capitalists. Investors contributing money to VC funds are known as limited partners (LPs), and consist of wealthy individuals and financial institutions with large amounts of available capital, such as state and private pension funds, university financial endowments, foundations, insurance companies, and pooled investment vehicles (termed “funds of funds”). Venture capital funds typically have a prespecified investment focus: for example, location, sector, or stage of business development. A fund is closed once the expected amount of money is raised from LPs, initiating its life cycle. For most funds the investment cycle is generally three to five years, after which the focus shifts to managing and making follow-on investments in an existing portfolio. Most VC funds have a fixed life of 10 years, with the possibility of limited extensions to accommodate private companies still seeking liquidity. Venture capitalists are compensated through a combination of management fees and carried interest. Typically, the GPs receive an annual management fee equal to around 2% of the committed capital made by the LPs. Carried interest is a share of the profits of the fund (20–30%), paid to the VC fund’s GPs and professionals as a performance incentive. The remaining profits are paid to the LPs.

There are two other types of VC fund: captive and government. Captive VC funds are partly or wholly owned by a parent company, which may be a bank, a financial institution, or a large transnational corporation. A few large nonfinancial companies, particularly technology companies such as Intel and Microsoft, have their own corporate VC subsidiaries which invest strategically, to complement their internal research and development activities, to gain access to new technologies and/or products, or to limit competitive threats. Government VC funds are policy tools aimed at filling local gaps in private VC, accelerating economic growth and employment, or commercializing technology. One US example is Small Business Investment Companies. Publicly owned or funded VC firms often face restrictions in where their investments can be located.

Venture capital investment is constituted through an inherently interrelated cycle of fundraising, investing, divesting, and new fundraising: a staged process of screening, investment, monitoring, management assistance, and liquidation or exit. Venture capitalists generate
their deal flow mostly by tapping their networks. Entrepreneurs will get warm referrals to venture capitalists through a respected and trusted member of the latter’s network: Unsolicited proposals rarely capture the venture capitalist’s attention. When competition becomes fierce, however, venture capitalists also actively seek deals in a number of ways, including forums, university events, and even cold-calling interesting prospects. Through screening, venture capitalists quickly decide whether a potential venture merits further evaluation. Most businesses seeking VC do not meet investors’ investment criteria and are declined. Those passing the initial screening stage are scrutinized in detail. In this due diligence phase venture capitalists examine the various dimensions of the venture, including quality of the management team, the business model, the product and its market segment, and the firm’s financial standing. Particularly critical is evaluation of the firm’s founders. Due diligence takes a considerable amount of the venture capitalist’s time, and may last a few months in each case.

If due diligence is completed to the venture capitalist’s satisfaction, he or she and the entrepreneur enter into negotiations on the specific terms of the investment. Several issues are pertinent in this stage, including valuation, contract provisions providing protection against agency risks, staging of future investment rounds, and board representation and oversight. Given the inherent uncertainties associated with any venture investment, venture capitalists often contract to protect themselves from agreeing to an inflated valuation. Rather than providing the entire sum of money at the outset, it will be staged over time as various milestones are met. This provides the investor with the option to revalue or abandon the investment in light of unfavorable new information, or to increase the capital committed in case of positive progress by the investee company. Venture capitalists typically take at least one seat on the board of directors of the firm to gain control over key decisions.

Venture capitalists spend around half of their time monitoring and supporting the companies in their portfolio, focusing particularly on early-stage companies. They use their expertise and networks to provide value-added services to investee companies, including monitoring financial performance and general business operations, serving as a sounding board, assisting in the acquisition of additional funding, helping with recruitment for, or replacement of, the management team, advising on strategic planning and key decision-making, coaching inexperienced entrepreneurs, and assisting with IPOs. There is a positive correlation between the level of the venture capitalist investor’s sectoral expertise and value added. It is fairly common for VC investments to be syndicated between several VC funds, especially in second and subsequent funding rounds. Syndication allows venture capitalists to pool expertise, diversify their portfolios, and share information and risk. However, any syndicated investment requires a lead investor – usually the initial investor – to perform the key monitoring and support tasks.

The VC industry is plagued by various conflicts as a result of informational asymmetry. Differences in opinion and conflicts of interest arise between venture capitalists and entrepreneurs as well as other parties (angels, coinvesting venture capitalists, creditors, exit counterparties). Coordination failure can even lead to litigation, although this is uncommon. While conflicts can be addressed in complex contracts, such contracts remain incomplete and are often insufficient to curb opportunistic behavior. Indeed, many of the terms in venture capitalist–entrepreneur agreements designed to protect the venture capitalist may invite opportunistic behavior by them at the expense of the entrepreneur and
other equity investors. As a result, reputational concerns play a major role in preventing venture capitalist opportunism.

The final stage in the investment process is to secure profitable exits for their investments. As a rule of thumb, if venture capitalists can realize spectacular successes in just one or two of every 10 investments, this suffices to guarantee expected returns and compensate for losses from failed investments. An IPO, in which the shares of the venture capitalist and potentially other shareholders are sold on the stock market, usually produces the highest return, but this may be feasible only for the best companies in a portfolio. Alternatively, a trade sale is pursued, especially for poorly performing investments, setting a lower valuation that can be immediately paid in cash or shares of the acquiring company.

Geographies of venture capital

The first independent VC organization, American Research and Development Corporation (ARD), was incorporated in Boston in 1946. Venture capital was regarded more or less as an American phenomenon until the late 1980s, when a significant VC industry emerged in Western Europe and Israel. The Internet boom at the end of the 1990s triggered rapid worldwide growth in the VC industry; the bursting of this dot.com bubble at the end of the decade resulted in a dramatic decline in the number of VC firms and the amount of VC. The market has gradually recovered to the pre-dot.com level in the core countries such as the United States and the United Kingdom, but has yet to return to its peak of 2000.

Research into the geography of VC commonly finds that the geographical concentration of VC funds closely overlaps with the location of high-technology industry. Numerous studies have documented this co-location. Martin et al. (2005) thus define spatial proximity effects in the VC industry as the tendency of, or necessity for, VC firms to focus a significant proportion of their investment in enterprises in their own immediate region. This reliance on proximity is considered to result from the very nature of VC operations. Apart from searching for new deals and assessing business proposals through a densely knit community with frequent face-to-face interactions, the bulk of a venture capitalist's time is normally spent serving on the board of directors and offering frequent managerial counsel to portfolio companies. Venture capitalists must also devote a large amount of time to fundraising under very high time pressure, and to monitoring, advising, and managing work more effectively when portfolio firms are located nearby. This implies that there are strong incentives for VC firms to locate in places that offer them the highest concentration of profitable investments.

But decentralized VC distribution also exists. Historically, many VC organizations were formed through spin-offs from, or at least linked to, existing financial institutions, and consequently tend to be based in or near major financial centers. Indeed, VC funds in high-tech clusters used to be outposts, established by investment institutions located in major financial centers. While the clustering of VC funds in high-tech centers evolved through feedback effects, VC funds continue to exist in non-high-tech financial centers such as New York, exporting their capital to established high-technology regions elsewhere. In short, venture capitalists are both mobile and local, developing many ways to overcome the liabilities of distance, particularly through investment syndication (Sorenson and Stuart 2001).

Meanwhile, lucrative investment opportunities in core countries have been decreasing; the
global decentralization of VC is underway. Such key target sectors as the Internet and telecoms have matured in the core, passing their innovative peak, with new breakthrough technologies yet to emerge (Mason 2009). The 2007 economic crisis in the core economies further cut into VC business, driving mainstream venture capitalists to search for investment opportunities in emerging markets. While the mature VC markets in the United States and the United Kingdom continue to evolve, traditional regional or national geographical boundaries are diminishing as more venture capitalists look to new regions, especially China and India, for new ideas, markets, and entrepreneurs.

In addition to investment syndication, VC companies are globalizing by establishing special funds dedicated to foreign markets and opening satellite offices in foreign hotspots. New VC clusters in emerging economies, such as Beijing and Shanghai, have been rising out of their pre-existing industrial bases, generating sizable investment targets for venture capitalists and successful track records of VC investments (Zhang 2011). The lack of strong regulatory and normative institutions in many emerging markets means that venture capitalists have to place more emphasis on employing personal networks for their operations. Combined with underdeveloped formal information infrastructures, and shortage of entrepreneurial experience and technological competency, VC investments there are generally constrained by even stronger proximity effects than in the advanced economies.

Spatially uneven VC development is also a product of variegation in institutional and policy contexts. The VC cycle from fundraising to investing and divesting is a highly regulated process. Thus VC operation is strongly shaped by the characteristics of the national and regional financial markets and institutional arrangements. The rise of VC in the United States was critically enabled by regulatory changes that removed restrictions on pension funds from investing in VC, reductions in capital gains tax rates, and the creation of NASDAQ in 1971 as a stock market for young firms with minimal listing criteria (Lazonick and Mazzucato 2013). Stock market-oriented markets in the United States and the United Kingdom are clearly more favorable to the development of VC than bank-centered markets such as in Germany and Japan (Sunley et al. 2005). The recent ascendance of VC in China is also associated with regulatory liberalization and institutional emulation of the United States (Bruton and Ahlstrom 2003; Zhang 2011).

It is not clear, however, to what extent VC is necessary and conducive to spurring innovation. Lazonick and Mazzucato (2013) argue that the socially devised, politically charged institutional arrangements that support VC have questionable implications for innovation and income distribution; the profits of innovation have been privatized but its risks socialized. In the United States, venture capitalists’ most favored investment targets in electronics and biotechnology were largely based on technologies generated through massive risk-taking state investments via National Institutes of Health, the National Science Foundation, the Department of Defense, and so on. The rewards gained by venture capitalists may thus be disproportionately high relative to the actual risks taken at the cost of taxpayers, especially for those late entrants in the development of a sector, after most of the real uncertainty and risk has already been absorbed by the public sector. Moreover, VC investments have propelled the spread of executive compensation through stock options, and have resulted in speculative or fraudulent management behavior, and massive corporate spending on stock buybacks aimed at boosting stock prices at the expense of research and development.
Simply supplying VC does not automatically create its own, high-tech-based demand. Very few businesses are capable of meeting the demanding investment criteria of VC, and the structure of venture investments is inappropriate for many young firms. Pisano (2006), for example, argues that, given the 10–20-year time frame for developing biotechnology products and the lack of profitability of the industry as a whole, biotechnology is not a suitable sector for VC. Also, the minimum amount of VC investment is simply too large for most small firms. For other parts of the world, it thus might be naive to copy the poorly understood Silicon Valley VC model without also funding and nurturing the underlying state-funded knowledge base on which VC in the United States has always depended. It is implausible to expect VC to lift innovation in the vast majority of less developed and economically lagging regions; it can never supplant other wellsprings of innovation – vibrant universities, corporate research laboratories, and policies supporting ordinary businesses, basic education, and labor training.

SEE ALSO: Economic geography; Finance and development; Financial geography; Globalization; Industrial agglomeration; Industrial geography; Innovation and regional development; Institutions and development; Networks, social capital, and development; Political economy and regional development; States and development; Technology and development; Uneven regional development

References

“Vertical integration” refers to the merging together of two or more businesses that are at different stages of production – for example, a meatpacking company and a chain of supermarkets. One most famous case of vertical integration is the expansion of Andrew Carnegie’s steelmaking empire in the nineteenth century. His company, Carnegie Steel, controlled not only the mills where the steel was made but also the iron ore mines and the coal mines, the ships that transported the iron ore, the railroads that transported the coal to the mills, and the coke ovens where the coal was cooked. Another good example is the restructuring of the oil industry in the second half of the twentieth century, where a handful of multinational oil companies, including Exxon, Shell, and BP, began to actively establish control along the entire supply chain from locating deposits, drilling and extracting crude oil, transporting the oil around the globe, refining it into petroleum products, to distributing the fuel to company-owned gas stations. It is believed that the benefits of vertical integration come from the greater capacity it gives organizations to control access to inputs. In other words, by establishing ownership along the supply chain, firms are able to control the cost of production, the quality of their product, and the delivery times of the inputs.

Two associated concepts that often come with the term “vertical integration” are forward integration and backward integration. Forward integration refers to the merging in the direction of further on into the production process (and thus closer to the point of sale). Merging with something further back in the process (if a food manufacturer were to merge with a farm, for example) is known as backward integration.

Vertical integration can be contrasted to “horizontal integration,” which refers to the merging together of businesses that are at the same stage of production, such as two supermarkets or two meatpacking companies. The integration of two organizations that are in completely different lines of business is referred to as “conglomerate integration.” One good example is Korean conglomerate Samsung, whose businesses range from electronics, heavy machineries, food processing, insurance, and shipbuilding to advertising.

The issue of vertical integration was at the center of the flexible specialization debate within Marxist geography during the 1980s. Regulation school scholars, such as Piore and Sabel (1984), Aglietta (1979), Scott (1988), and Lipietz (1987), argued that the transformation of capitalism during the 1970s can be understood as a transition from “Fordism” to “post-Fordism.” According to their view, from the interwar period to the mid-1970s the advanced capitalist countries were dominated by Fordism – by mass production of standardized goods in huge factories owned by large, vertically integrated global firms. Towards the end of that period, however, the system had begun to run into crisis: mass markets had begun to break up as consumers tired of standardized products; labor resistance
VERTICAL INTEGRATION

had built up; industries met inherent technical barriers, such as line balancing problems; and the system was too rigid to cope with the uncertainty of economic recession. Meanwhile, a new model of flexible specialization, or post-Fordism, emerged: small batch production in interlinked, specialized small firms, flexible in organization, work process, and output, and tending to concentrate spatially into industrial districts. As production processes become increasingly vertically disintegrated, firms are now becoming more responsive to changing tastes, and to the tendency for consumer products to support the construction of personal identities.

One of the primary examples that regulation school scholars use to make the case for post-Fordism is the Third Italy. The term refers to clusters of small firms and workshops developed in the 1970s and 1980s in the central and northeast regions of Italy, including Tuscany, Umbria, Marche, Emilia-Romagna, Veneto, Friuli, and Trentino–Alto Adige/Südtirol. Each region specialized in a range of loosely related products and each workshop usually had 5–50 workers (often less than 10). The wide range of products in each cluster suggested a shift from economies of scale to economies of scope. Additionally, these small firms and workshops were known for producing high-quality products and employing highly skilled and well-paid workers. The workshops were also very design oriented and multidisciplinary, involving collaboration between entrepreneurs, designers, engineers, and workers. This unique industrial structure was in sharp contrast to the large-scale mass production systems seen in Turin, Milan, and Genoa (the First Italy) and the underdeveloped South (the Second Italy).

Sayer and Walker, in The New Social Economy (1992), criticized the thesis of flexible specialization for being “long on speculation and short on coherence.” They pointed out that the binary histories implied in the flexible specialization thesis led scholars to only select cases and data that would fit into the story. The result of this dualistic rhetoric is the reduction of the richness of the history. For example, it is problematic to characterize the pre-1970 capitalism as “Fordism,” as vertically integrated firms constituted only a small part of the economy even in advanced countries. Moreover, the examples that the regulation school scholars used were too selective to build up the case for post-Fordism. While fascinated with the batch production of Third Italy or Los Angeles, regulation school scholars have ignored the mass production system in Japan, which was a real threat to the manufacturing powers of the United States and Europe. Last but not least, the need to respond to the diversified consumer demands alone cannot explain vertical disintegration. The Japanese postwar economy, for example, is highly vertically disintegrated: its manufacturing is carried through multiple layers of subcontracting, which are governed by industrial groups. The products that this system is good at producing, such as automobiles, however, are for the mass market. There is, therefore, no clear causal relationship between the market and industrial organization.

Sayer and Walker argued that the shift towards vertical disintegration is more likely the result of innovation in organization. Here, “organization” refers to the coordination of different production stages. A firm that looks vertically-integrated may actually be very weakly organized. By the same token, production systems that are formally vertically disintegrated may have strong interfirm organization. Thus, while there is evidence of an increase in vertical disintegration, an increase in “vertical organization” in both vertically integrated and vertically disintegrated cases is of more significance. The term “innovation,” on the other hand, emphasizes the need of
an evolutionary perspective on the change of industrial structure. They argued that industrial structure is not the product of some timeless logic of rational choice, but the consequence of organizational evolution and learning that characterize capitalism. Such an evolutionary perspective recognizes the role of place and local happenstance in the development of new forms of industrial organization. The extreme physical restrictions on space and the population density of Japan, for example, have been more congenial to agglomerated, highly organized, and disintegrated production than has the American context, in which giant, decentralized, multidivisional companies prospered.

**SEE ALSO:** Economic geography; Flexible specialization; Fragmentation of production; Industrial geography; Industrial restructuring; Marxist geography; New industrial space

**References**


Vietnam: Viện Địa lý, Viện Hàn lâm Khoa học và Công nghệ Việt Nam (Institute of Geography, Vietnam Academy of Science and Technology)

Founded: 1993
Location of headquarters: Hanoi
Website: http://ig-vast.ac.vn
Membership: 122 (as of 2015)
Director: Lai Vinh Cam
Contact: lvcamminh@yahoo.com, qlthvdl@gmail.com

Description and purpose

The institute is one of the research organizations of the Vietnam Academy of Science and Technology. The institute focuses on the various research tasks in the field of applied geography, such as the integrated study of the natural environment and resources including socioeconomic problems for the establishment of the scientific foundation for socioeconomic planning in Vietnam; participation in appraisal of projects on natural resource exploitation, spatial distribution of population, and territorial organization; and research, assessment, and forecast of environmental changes under human activities including natural processes to support sustainable development and disaster mitigation. Postgraduate education is also one of the functions of the institute.

Journals or major publication series

Journal of Sciences of the Earth

Current activities or projects

Activities include:
- the development of sustainable socioecological models for Tay Nguyen Highland;
- development of the scientific foundation to resolve conflicts in water exploitation in Tay Nguyen Highland;
- integrated study of soil degradation and desertification in the southern part of central Vietnam;
- integrated study and assessment of natural resources and the demography situation in the border area between Vietnam and Laos in Kontum and Attapeu provinces for sustainable development and spatial planning of the population; and
- acting as the principal organizer of the 31st Asian Conference on Remote Sensing, held in Hanoi in 2010.

Brief history

The institute has a long history which dates back to 1957 when the idea for the current format of the institute was formulated under the Ministry of Education. This first body was named the Research Committee on Literature, History, and Geography. In 1961, a team was formed to research biology and geography under the State Committee of Science. Two important research projects were conducted: physical geography zoning in the north of Vietnam and spatial distribution mapping of population and ethnic groups in the north of Vietnam. In 1967, the Department of Geography under the Institute of Natural Science, State Committee of Science and Technology, was founded. In 1975, the Vietnam National Center of Science was established and geography research was moved...
to the Institute of Earth Sciences. In 1987, the Center of Geography and Natural resources was founded. This center brought together geographers from several research units to enhance capacity in research implementation. Finally, in 1993, the center became the Institute of Geography and a national authority in geography and environment research. The institute plays an important role in the development of a scientific foundation for various socioeconomic planning tasks in Vietnam. Environmental assessment of important infrastructure and hydropower station development have also been conducted by the institute.

Submitted by Lai Vinh Cam
Violence

Michael J. Watts
University of California, Berkeley, USA

Problems of violence still remain very obscure. (Sorel 1961/1906, 60)
What Sorel remarked sixty years ago … is as true today as it was then. (Arendt 1969, 3)

During World War II, a German officer apparently visited Pablo Picasso in his Paris studio, where he was confronted with what he saw as the modernist chaos of Guernica, perhaps the defining artistic representation of contemporary war and violence. “Did you do this?” he asked, whereon Picasso was said to have calmly replied, “No. You did this.” Picasso’s accusation – your politics is responsible for this violence – speaks directly and powerfully, of course, to several defining aspects of the twentieth and twenty-first centuries: interstate and civil wars, state violence and terror, and endemic political conflict. It has been calculated that war-related killings in the twentieth century exceeded 105 million, including 62 million victims. For Isaiah Berlin, the twentieth century was “the worst century there had ever been” (quoted in Mazower 2002, 1158). For every person killed, many more were subject to a multiplicity of physical forms of violence: torture, injury, brutality, persecution, imprisonment, starvation, and so on. This was the century of secret police, concentration camps, gas chambers, labor camps, and the calculated liquidation of specified populations. In the five years after the end of the Cold War — when it was plausibly expected that the fall of the Berlin Wall would bring to an end the raft of Cold War proxy and imperialist wars in the four decades that had followed the Allied victory in 1945 – there were 177 violent conflicts in 79 countries (Cramer 2006, 1). Even though data seem to suggest that deaths from interstate and civil wars have declined since the mid-1980s, there is a proportionate increase in so-called new forms of conflict (militias, terrorist groups, transnational gangs, piracy, organized crime) and a recognition of the conflict–security–development nexus associated with fragile and conflicted (and often highly populated) states such as Nigeria, Pakistan, Indonesia (World Bank 2011). The UN High Commissioner for Refugees’ (UNHCR) annual report on Global Trends for 2014 documents that the number of worldwide refugees and displaced peoples – precisely the “collateral damage” of these conflicted states – had been at the highest level ever recorded: the number of people forcibly displaced at the end of 2014 had risen to a staggering 59.5 million compared to 51.2 million a year earlier and 37.5 million a decade ago (UNHCR 2015). The increase represents the biggest leap ever seen in a single year, and the situation is likely to worsen. Globally, one in every 122 humans is now either a refugee, internally displaced, or seeking asylum (if all refugees constituted a country, it would be the world’s twenty-fourth biggest).

It is no surprise, given the proliferation of political violence – indeed of all manner of forms of violence, some of which do not sit easily within the circumference of “the political” – that violence has come to be seen as central to an understanding of modern politics (think of Thomas Hobbes, John Locke, Max Weber, Hannah Arendt) and to stand at the heart
of much social science theorizing, including geography. It is now commonplace to acknowledge that the great arc of modernization – or, for that matter, the Enlightenment – bears all the marks of coercion, social conflict, and struggle (Tilly 1992); that organic nationalism gave birth to genocidal forms of nationalism and ethnic cleansing; and that state violence and terror figure centrally in what might be called modern mass or totalitarian violence, whether perpetrated in the name of fascism, communism, or state building (Arendt 1973).

It needs to be said immediately that such views do not go uncontested. Psychologist Steven Pinker’s recent book, *The Better Angels of Our Nature* (2012), argues that violence has declined, that the premodern societies were “shockingly violent,” that brutality was woven into the fabric of premodern social life, and that the frontier of violence has receded because of our deepening empathy, self-control, and morality; the spread of government, literacy, trade, and cosmopolitan worldviews; and our deepening powers of reason. Whatever the empirical scope and trajectory of such brutality over the past couple of centuries, the term “violence” itself is capacious. It can refer to specific sorts of events (e.g., interstate, civil, ethnic, religious, electoral, political, communist, revolutionary, guerilla, insurgent, imperialist, terroristic), as well as to a variety of forms and analytical categories (e.g., racial, queer, psychological, structural, silent, symbolic, repressive, normalized). All of this makes any entry purporting to capture geographical and social scientific scholarship on the subject fraught and exceedingly difficult.

Two related but rather different ways of thinking about violence can be deduced from the word’s etymology (Bufacchi 2005). Violence is derived from the Latin *violentia* (vehemence), meaning a passionate and uncontrolled force. Force is typically seen to be constitutive of violent acts: the *Oxford English Dictionary* defines the word as the exercise of physical force so as to inflict injury or harm and to cause damage to persons or property. Violence, said John Dewey, the American pragmatist, is force gone wrong. Not all acts of violence deploy force, of course – persons might be killed by the using of chlorine gas, say, as in contemporary Syria. But violence is necessarily destructive and is often defined by intentionality. A simple relation between force and violence has, of course, been questioned. Hannah Arendt (1969, 1973) rejected the purported synonymy of force and violence, and suggested that we should instead focus on the relations between violence and power. Force is a potentiality or a capability – for Arendt it is a sort of energy – while violence is an act, something that is perpetrated and enacted. In the same way, force carries none of the moral and pejorative opprobrium of the concept of violence.

There is a second and more capacious meaning of violence, however, which turns on the word’s relation to violation from the Latin *volare* (infringement). There is here an association between violence and words such as “dishonor,” “violation,” and “emotional or psychological injury.” Violence as violation is a more capacious subsuming force but also one acknowledging and emphasizing the transgression of norms and rights. Bufacchi (2005) identifies three sorts of violations: first, violations of personal rights and of the dignity of the person; second, violations of rights to ourselves; and, third, violations of human rights in the fashion articulated by peace theorist Johan Galtung in his notion of structural violence and in the way that much of the humanitarian international and human rights communities see their mission.

The differing circumferences of both notions of violence – one aimed at excessive and destructive force, the other pointing toward...
what Randall Collins (2009) would call violent “situations” capable of embracing conditions of threat, fear, and intimidation which imply psychological harms – call attention to quite differing theoretical traditions, including those deployed by geographers. Violence in these differing senses also stands in quite different relation to conceptions of power (repressive vs biopolitical, for example). Much social scientific scholarship turns on the character and consequences of the putative state monopoly of the means of violence – Weber's classic definition of the state and the state’s ability to declare states of exception in moments of crisis. Virtually any state increasingly obsessed with security – not least in a post-9/11 paranoid world in which the American empire is seen to be under direct threat – is inclined to terror by the provocation of terror. The relation between state power and the desire to identify new or old enemies (internal and external), and the state power’s complicated connection to authority, law, and the military, stand at the heart of much modern political violence and naturally of much social scientific (and geographical) analysis.

These points of theoretical contrast – and the degree to which they invoke contrasting theories of the role and capabilities of the modern state in the perpetration of violent acts, whether war, genocide, counterinsurgency, or repression – can be glossed over briefly in reviewing a quartet of foundational theorists of violence: Hannah Arendt, Frantz Fanon, Walter Benjamin, and Michel Foucault. Arendt argues that, while they are related, power and violence are distinct concepts: the former arises from the consent of groups, “the human ability not just to act but to act in concert” (Arendt 1973, 143), whereas violence requires neither numbers nor consent but relies on implements that are capable of multiplying their strength. Two less important but related concepts in Arendt are strength and force. Strength Arendt defines as belonging to the individual and as distinct from power, which is a property of groups. Force is used to designate the energy released by physical or social movements. Violence, according to Arendt, has the role of ensuring the stability of international relations, which means national independence and the claim to unchecked and unlimited power in foreign affairs as identified by nations (Arendt 1973, 107). Arendt concludes that, as the power of violence in international relations diminishes, it is enhanced and intensified in domestic relations. A government that relies entirely on violence has no power. Tyranny is both the least powerful and the most violent form of government (Arendt 1973). Power stands as a counterpoint to violence, and where violence interacts with power it inevitably destroys it. In the same way, violence can never create power (Arendt 1973, 152).

Conversely, Fanon (1969) draws on both conceptions of violence in his account of the postcolonial condition and of the relations between violence and human emancipation. His notion includes force, physical injury, coercion, militancy, and, importantly, psychological injury. Settler colonialism was itself a violent imposition, and decolonization, faced with the alienated condition of the “native,” required both the physical freeing of a territory from a colonizer but also the psychological freeing of an alienated and subjugated consciousness. Decolonization was always a violent phenomenon for moral and practical reasons (settler colonialism was ruled through violence; only through violence could humanity be restored), and only through violent decolonization could the native (and colonial society) be purified. Fanon takes on the question of race and racism seen as central to totalitarianism but inverts Arendt’s logic to see the emancipatory potential of violence.
Benjamin (1978) emphasizes the complicity between the order of violence and the law. He counterposes natural law (law as a pure means to an end, that is to say, an act is violent to the extent that its ends or objectives are unjust) against positive law (in which the means of a certain action comes under scrutiny regardless of justness). Benjamin focuses on the question “Under what conditions is it possible to distinguish between legitimate and illegitimate violence?” He distinguishes between law-making and law-preserving violence, both of which are irreducibly connected by power. At the heart of his analysis is the state (rather than nonstate actors) and the notion that violence reaffirms the law and the law reaffirms violence. There is always violence at the heart of every form of political and legal authority.

Foucault’s (2010) genealogy of power also allows us to see how society settles each of its violations within a system of rules. Power (as for Benjamin) is not an instrument or a capacity (repression); it must be freed from the law and sovereignty. Power is seen through the lens of violence. War was woven into the fabric of society and, says Foucault, it operates as an ontological schema for interpreting the world, a grid of intelligibility. In the modern period, violence and war find form in a specific set of operations of power, namely biopolitics: violence is now inscribed in modern societies through the administration of life, through forms of regulation and calculation, in which wars are fought by states on behalf of the populations they administer. Biopower is organized around the principle of security and the preservation of life.

Running across these four figures is an interrogation of the relations between the state, power, law, authority, and violence. Each tradition has been drawn on, in different ways, by geographers to grasp violence in particular places and times, to locate violence in a specific space–time nexus (Gregory 2011). Geographic work on what one might call “economies of violence” (Watts 2013) has been especially shaped by three formative events and processes. The first is the close relation between the discipline of geography and the twin forces of state formation and empire (Smith 2003). Exploration, cartography, the deployment of geographical knowledges in forms of militarism, and imperial conquest have been central themes in geographical scholarship, whether through geopolitical theory, the biographies of key geographers and geographic institutions (Smith 2003), or analyses of colonial and postcolonial state building (Bryan and Wood 2015). A second is of course the “war on terror,” post–Cold War militarism, and the deepening invasiveness of the national security state particularly after 9/11 (Gregory and Pred 2007; Paglen 2009; Kosek 2010. The third has been a body of work inspired by Edward Said’s observation that:

Just as none of us is outside or beyond geography, none of us is completely free from the struggle over geography. That struggle is complex and interesting because it is not only about soldiers and cannons but also about ideas, about forms, about images and imaginings. (1993, 7)

These struggles have focused on multiple spatial and territorial arenas, from the city to indigenous territories, to liberated insurgent zones, to prisons, to the domestic sphere, to the violence of neoliberalism and the wageless class, to more recently the violence associated with police brutality and the Black Lives Matter movement. This considerable body of work draws on both conceptions of violence and varied notions of power (repressive, disciplinary, biopolitical, and patriarchal) in addressing a raft of geographical problems.

Some of the richest geographical work on violence has emerged from two theaters of conflict: the first surrounds resources and the
environment, where political ecology figures centrally as a frame for exploring violent and conflicted struggles over access to resources (Peluso and Watts 2001), and the second addresses the so-called new forms of post-Cold War conflict (militias, transnational gangs, ethnic gangs, organized crime), particularly in the “failed states” of the Global South. The World Development Report 2011: Conflict, Security, and Development (WDR 2011) argues that violence may not just be a cause of poverty but perhaps the central cause for the “bottom billion of the population” (World Bank 2011). One and a half billion people currently reside in areas affected by what WDR 2011 calls fragility, conflict, or large-scale organized criminal violence. Central to the condition of the poor is, in short, a violence trap. Peace confers conditions in which economies can grow; conversely, poverty is concentrated in states riven by civil war, ethnic conflicts, and organized crime. WDR 2011 asks what spurs the risks of violence, why conflict prevention and recovery have proven so chimerical, and what can be done by national leaderships and international development organizations alike to restore stable development in conflict-torn and fragile states, or what the Clinton administration in the 1990s famously called “failed states.” The central message of the report is that “strengthening legitimate institutions and governance to provide citizen security, justice and jobs is crucial to prevent cycles of violence,” a key point being that violence is recursive and high rates of recidivism plague war-torn countries.

There is a dialectics of violence: conflict can cause poverty (death, injury, and displacement reduce life chances) as much as poverty can cause violence: if you are poor, it pays to rebel. As a result, not a single low-income conflict-affected or fragile state has achieved a single Millennium Development Goal. And, finally, the cyclical and repetitive nature of conflicts, identified by Paul Collier (2007) and his colleagues in their earlier work, produces a geographical and politico-economic condensation of violence and its destructive consequences. In other words, the territorialization, or spatialization, of organized violence falls disproportionately in some parts of the Global South, with devastating effect. For every three years that a country is affected by violence, poverty reduction lags by 2.7 percentage points. The figures are mind-numbing: globally, 42 million people are displaced by violence, and two-thirds of those who are undernourished, without potable water, and without primary education reside in conflict-affected states. The indirect costs of conflict – in which children and women bear a disproportionately high burden – are mind-boggling: the indirect medical costs of violence in Brazil are 40 times higher than the direct costs (World Bank 2011). While these putatively new forms of violence are not exclusively the preserve of resource-dependent (or resource-“cursed”) political economies, the intersection of postcolonial state building and resource-rich states (see the so-called resource curse literature: Collier 2007), some of which are confronting the challenges of global climate change, has provided an exceptionally rich setting for geographic study (Dalby 2009).

As Philippe Le Billon’s important book on resource conflicts (2012, 1) points out, the 1990s were crucial because, even if the trend line of the number of armed conflicts began to fall, those that remained seemed increasingly “tied to resource sectors” while the armed groups themselves came to resemble “militarized commerce” rather than “ideological sponsorship.” All of this was framed, says Le Billon, by the rapid restructuring of the global economy, the “third wave” of democratization, the liberalization of the resource sector, the volatility of commodity prices, and the arrival of new
resources companies from “emerging countries, chiefly China.” These observations were (and remain) part of a more expansive body of scholarship and policy work addressing the changing face of conflicts in general, variously framed in terms of new and old wars, of the rise of asymmetrical warfare, of the disappearance of 1960s-style guerilla movements and their attendant political commitments and ideological orientations, or the various species of “new conflicts.” Le Billon focuses exclusively on resource sectors (although his analysis has implications for how we think about armed nonstate actors and patterns of violent or contentious politics more broadly) and how resources influence the likelihood and course of armed conflicts. He provocatively claims that all resource sectors influence conflicts in the same way – through location, mode of production, revenue accessibility – and that resources themselves do not cause conflicts. Rather, he argues, resources are hybrids (products of joint social and economic processes) that contribute to shaping social relations and are in turn expressive of social relations. Le Billon offers “three dimensions” of his analysis – vulnerability (economic dependence on a resource), risk (social relations of production that overdetermine conflict over access, control, and extraction of resources), and opportunity (how resource properties provide differing opportunities for insurgents and armed conflict) – which “engender various types of violence.” He addresses three of these: unrealized potential (failure to harness and spread wealth), environmental violence, and multiple forms of violence associated with armed conflict (Le Billon 2012).

Geographers of resource conflicts share the field with armies of economists and political scientists, many of whom have been drawn to study civil wars of various sorts through large-N studies. Le Billon and other geographers (e.g., Richards 1998) are critical of the econometric approaches and three broad sets of conventional political economic mechanisms linking resources and armed conflicts: the institutional weakening effect (the orthodox resource curse analysis), the motivational effect (grievance and greed theory), and the opportunity effect (rebel finance). The strength of a geographical analysis is that it links political economy, the material qualities, and differences of resources (location, abundance, material form, forms of extraction, and so on) while simultaneously being attentive to their properties as commodities (their social life, the ways in which resources cannot be represented or discussed just in terms of the meanings attached to them but also in terms of how they might be fetishized). Resources are social processes demanding what Le Billon calls “thick historical and geographical contextualization” (2012, 13). History matters through memory and identity, through the timing of resource discovery and insertion into institutional and political economic arrangements, and geography itself is a historical product (Watts 2013). By emphasizing vulnerability, risk, and opportunity, geographers have generated a typology of resource conflicts that turn on whether the resource in question is proximate or distant from sources of economic and state power, and whether the resource is at a point or diffuse. Le Billon’s (2012) matrix yields four conflict outcomes: coup d’état (proximate point), secession (distant point), mass rebellion (diffuse–proximate), and warlordism (diffuse–distant). The originality of this sort of analysis is that it takes the materiality of the resource – for example, how the “oilyness” of oil makes a difference – very seriously, and Le Billon attempts to draw out these implications for the onset, duration, and severity of the conflicts. He concludes that resources can motivate and finance both before and during the onset of violence, and can also shape the internal
coherence and cohesion of the movements themselves.

Geographic work has delved deeply into the material worlds of specific resources understood as commodities within global capitalism: oil, timber, and diamonds have figured centrally. Vulnerability, risk, and opportunity all play a role in the case of oil, as did the location of oilfields and the “institutional setting.” In the case of diamonds, diamonds are not a rebel’s best friend because they can also set out constraints and barriers (tactical, logistical, financial) for belligerents. In relation to timber, there is no statistical evidence that forests make any difference to the onset of risk or to the duration of conflict. Insurgent groups may opportunistically use forests as advantageous terrain, and there may also be links between forests and rebellions through particular histories, institutions, and capabilities of forest-dwelling peoples.

Political ecologists, and geographers generally, have resisted the urge to offer “a theory of violence,” or Olympian claims about the powers of particular resources, or indeed easy solutions and prescriptions. Political ecology properly encourages a sensitivity to history, space (at various scales), and the institutional and political economic settings into which resource exploitation and resource revenues are inserted, while being attentive to the properties and qualities of the resources. It is this complex overlapping and nesting – in effect, a recognition of the simultaneous coexistence of various forms of conflict in and around the very same resources that collectively comprise a vastly complex social field of violence – that geographers are exploring. There is interesting new work of this sort on resources, state-making, and frontier violence, on transnational gangs, and on the geographies of counterinsurgency. These sorts of complexities – that is, the historically and spatially granular contextualizations they invoke – sit uneasily with much of the social science research running large-N regressions. As Gadda (2007, 5–6) put it in describing not violence but crime:

the apparent motive, the principal motive was, of course, single. But the crime was the effect of a whole list of motives, which had blown on it in a whirlwind … and had ended by pressing into the vortex of the crime the enfeebled “reason of the world.”

SEE ALSO: Biopolitics; Carceral geographies; Crime; Frontiers; Geopolitics; Political ecology; Power; Power and development; Race and racism; Resource curse; Resource extraction; Sovereignty; State, the; Territory and territoriality; War

References


VIOLENCE

Virtual geographic environments

Min Chen  
*Nanjing Normal University, China*  
*Chinese University of Hong Kong, China*

Hui Lin  
*Chinese University of Hong Kong, China*

Guonian Lu  
*Nanjing Normal University, China*

After Michael Batty first used the term “virtual geography” in 1997 (Batty 1997), the term “virtual geographic environments” (VGEs) was formally proposed by Lin and Gong as a concrete object of study in the discipline of virtual geography (Lin and Gong 2001). The term “virtual geographic environment” stems from two other well-known terms, “geographic environment” and “virtual environment.” A geographic environment refers to the portion of Earth’s surface on which creatures (especially human beings) live. It consists of natural (e.g., water, air, and soil) and social (e.g., human behavior) factors related to this geographic space, and it is continuously changing as a dynamic system. The virtual environment, which is also called a virtual world, digital world, or electronic environment, is widely used to describe virtual spaces that are built based on visualization and virtual reality (VR) technologies; recently, many of these spaces have been developed through the Internet. Users can participate or become immersed in these virtual environments by “feeling” and “exploring” spaces that do not exist in real life or that cannot be reached at the local time and location. Although many types of virtual environments are rooted in game playing, they have potential usefulness for scientific research (Bainbridge 2007). Many applications have been appeared in various fields, such as house design, urban planning, and social network analysis. With the two terms, a VGE can be regarded as a typical virtual-based geographic environment that allows users to “feel the geographic scenarios in person” and to “know the geographic laws beyond reality” (Lin et al. 2013b).

In this context, since the mid-2000s, more researchers have begun to study this new branch of geographic information science (GIScience) from a theoretical or a technical perspective (Lin et al. 2013a, 2013b). Although the main concept remains the same, the development process indicates that VGEs have experienced a gradual evolution in terms of their content and functions. Overall, three primary stages are apparent.

The period from 1998 to 2002 can be regarded as the embryonic period of VGEs. The term “virtual geographic environments” was coined at a conference and later published in the conference proceedings (e.g., Gong and Lin 1998). However, no clear definition of it was given at the time. After three years, the term was formally proposed in Lin and Gong’s article “Exploring Virtual Geographic Environments” (2001), which described VGEs as “environments pertaining to the relationship between post-humans and 3-D virtual worlds.” Five types of space (i.e., Internet space, data space, 3-D geographic space, personal perceptual and cognitive space, and social space) were defined to characterize VGEs. Lin and Gong also mentioned the “georeferenced virtual environment” to “allow
VIRTUAL GEOGRAPHIC ENVIRONMENTS

distributed users to congregate virtually on the Web and interact with 3-D graphical worlds to explore the Earth’s geographic phenomena and processes in an immersive or semi-immersive way.” In their opinion at that time, a “georeferenced virtual environment” was a component of VGEs. Although their idea of VGEs was slightly clearer than the former concept, it appears that the meaning was so broad that it was difficult to understand the main purpose of providing such a VGE. A gap still existed between VGEs and geography. In contrast, the definition of “georeferenced virtual environment” suggests that it has a use for the exploration of geographic phenomena and processes on Earth. It is easy to understand and appears to focus on solving geographic problems.

The following six years (up to 2008) are regarded as a continuous exploration stage. During those years, in order to study the relationship between VGEs and geography, scholars interpreted VGEs from the perspective of geographic language as a basic tool to represent geospatial information that is related to both the physical and social aspects of an environment. By considering multidimensional expression and multichannel interaction as typical characteristics, they regarded VGEs as the result of the evolution from verbal language through text, maps, and geographic information systems (GIS). This evolution is a reflection of the continual improvement of spatial information communication (Lin and Zhu 2006). Accordingly, studies on VGEs then focused on how to provide a virtual environment that corresponds to the real world to allow users to better understand it through a visual or immersive experience. At this stage of its evolution, a VGE is similar to the previous “georeferenced virtual environment.” Although the objective of VGEs was narrowed down, its contribution to geographic problem-solving emerged. This concept is more practical for today’s development of VGEs.

The years 2009–2015 are regarded as the explosive stage of VGEs, as an increasing number of scholars have conducted research on VGEs. For example, Goodchild (2009) proposed that volunteered geographic information (VGI) could be an important data collection mode for the development of a VGE; Mekni (2010) discussed the use of agents to simulate human behavior in VGEs; Priestnall et al. (2012) highlighted that VGEs can be employed in education and geographic learning; and Konecny (2011) summarized the challenges in the study areas of VGEs. Although the different perceptions were linked to various research approaches and achievements, this period saw the clarification and clearer conception of VGEs. To provide a summary of the previous research over nearly one decade, Lin, Chen, and their colleagues (Lin et al. 2013a, 2013b, 2015; Chen et al. 2015) redescribed their proposed VGEs as a new generation of geographic analysis and computer-aided geographic experiment tools. VGEs are “a type of typical web- and computer-based geographic environment” built “by merging geographic knowledge, computer technology, virtual reality technology, network technology, and geographic information technology,” “with the objective of providing open, digital windows into geographic environments in the physical world, to allow users to ‘feel it in person’ by a means for augmenting the senses and to ‘know it beyond reality’ through geographic phenomena simulation and collaborative geographic experiments.” Thus, a VGE can “contribute to human understanding of the geographic world and assist in solving geographic problems at a deeper level” (Lin et al. 2013b). Compared with the ideas in the former two stages, the latest definition of a VGE has a closer relationship to geography. The construction of a
VGE is currently targeted at geographic understanding and problem-solving through not only geovisualization and human immersion but also geosimulation and geocollaboration. To some extent, this definition has provided an answer to one of the “big questions” proposed by Cutter, Golledge, and Graf (2002), namely, “what role will virtual systems play in learning about the world?” The definition also coincides with three of Goodchild’s five future points about GIScience: multidimensional visualization, dynamic phenomenon simulation, and public participation (Goodchild 2010; Lin et al. 2013b).

Structure and components of VGEs

To achieve a better understanding of real geographic environments through VGEs, the features of real geographic environments should be addressed. Three keys characteristics are noteworthy: (i) a geographic environment is a comprehensive system consisting of natural factors (e.g., soil and water), social factors (e.g., human), and their interactions; (ii) it is a dynamic system because geographic processes change over time; and (iii) in this environment there are generally two types of geographic objects, discrete objects with boundaries (e.g., a tree or building) and continuous fields without boundaries (e.g., the air). These characteristics require substantial detailed functions of VGEs. Thus, to date, a VGE must support (i) geovisualization and human immersion to assist in acquiring a visual image and to foster participation; (ii) geosimulation for reproducing and predicting dynamic geographic processes; and (iii) geocollaboration of experts from various fields to conduct comprehensive geographic analyses and experiments through the Internet.

The current structure of a complete VGE is shown in Figure 1. The foundational technologies are computer technology, VR technology, network technology, and geographic information technology. However, the most important basis is the thorough consideration of geographic knowledge during the building and implementation processes. For example, in some game-like virtual environments, a man can fly everywhere at any instant, which is not possible according to common geographic knowledge in the real world. A VGE is built to solve geographic problems so that these common laws should be integrated into the VGE as restraints. Based on these foundational technologies and factors, four components (i.e., the data, modeling and simulation, interactive, and collaborative components) and two cores (i.e., a geodatabase and a geographic process model base) should be equipped within a complete VGE. Finally, the virtual geographic scenarios are built for the public to immerse themselves in and to provide their spatial knowledge, and for the researchers conducting collaborative geographic experiments.

The functions of the four components have been discussed in detail by Lu (2011) and Lin et al. (2013a). A short summary is provided here. Geographic data (geodata) are the basic resource for geographical scene expression and geographic problem analysis. The collection, management, and usage strategies of geodata are always key components of geographic analysis tools. In a VGE, the data component is designed to take responsibility for geodata management and organization. Today, with the rapid development of photogrammetry and remote sensing technologies, mass geodata from the sky to underground can be acquired by different professional methods with various tools. At present, more timely social data (e.g., behavior choice) can be provided by the public as a result of the wide distribution of sensors and the wider availability of natural data (e.g., the temperature and air quality of a location). Although the
accuracy and precision of this type of data are sometimes still worthy of discussion, there is no doubt that the data will contribute to the construction of VGEs (Goodchild 2009). In the context of mass multisource and heterogeneous (e.g., differences in data structure, semantic, scale, or coordinate system) geodata originating from the real geographic environment directly or indirectly, putting these data back into the computer via data integration to form a virtual environment for representation, simulation, and prediction is a significant challenge. Many types of geodata models in traditional GISs were designed for visualization (2-D/3-D) and spatial analysis; thus, a gap still exists between traditional geodata models and efficient geosimulation (Lu 2011). If these issues are solved, then the functional geodatabase, which is a core base of the VGEs, can be developed and fully utilized as a foundational support to other functions.

The current research priorities of GIScience are changing from static snapshots to dynamic complex processes (Goodchild 2010). A static scenario is easily described, but it is powerless for trend predictions. Geographic modeling and simulation are important for gaining a better understanding of these geographic processes. Studies in this field have produced various geographic process models, many of which have been used for hazard analysis, weather forecasting, and flood warnings. However, except for the developer and owner, other researchers typically

Figure 1 Current structure of a complete VGE. Modified from Lin et al. 2013b.
have difficulties in understanding and reusing these models, which causes a problem of “model islands.” Thus, a type of model reuse mode should be promoted. In practice, one needs to buy the model resources or related software and deploy them on a server, even for small experiments. This process is costly and inefficient. If these models can be packaged and shared in a network and provided by a unified architecture, then the fee can simply depend on the times that the model users request. Under the precondition that the safety of both the data and model are strictly protected, this approach will not only benefit model users but also allow model owners to promote and test their models. The modeling and simulation component in a VGE aims to help researchers share and reuse the existing model resources in a VGE conveniently without focusing on the configuration of the network and hardware. Thus, it can be regarded as a bridge between the geographic process model and distributed model users. Technological problems related to this component are under consideration (e.g., Wen et al. 2013), and the geographic process model base, which forms the second core of VGEs, will be construed based on this component. Thus, in a VGE one may rent and implement a single model through the Internet after it is packaged and integrated into the geographic process model base, regardless of the user’s budget or technological knowledge of the computer’s configuration environments. Combined with the functions that are provided by the data component and the interactive component, a VGE can provide the simulation result and corresponding visual or sensible “real” geographic processes that are calculated by the process models. Moreover, because of the complexity and comprehensiveness of the geographic environment, scientific geosimulation urgently needs the collaborative efforts of multidisciplinary models (e.g., rainfall, soil erosion, and evaporation and diffusion models). These models must be coupled in a systematic workspace. According to Lin et al. (2013a), the modeling and simulation component will also provide multidisciplinary researchers with the opportunity to create a conception model based on the complex geographic processes in a VGE. Then different unit models can be linked to conduct a comprehensive simulation to explore a cross-cutting geographic problem. The results can be examined in terms of intuitive feeling and common understanding.

The modeling and simulation component of a VGE reflects the design of the tools that support multichannel geointeraction. This component is the bridge between the users and virtual scenarios. For the users, interaction can be divided into perception (through vision, hearing, touch, smell, and taste) and manipulation (through a mouse, gesture, or voice). Related research on human–machine interaction (HMI) is increasing in various areas, such as VR and artificial intelligence. Those achievements have been fully referenced in the development process of a VGE. However, specific challenges remain to realizing geointeractions in a VGE in order to achieve a “natural” feel and comfortable manipulation. For example, a lake may release different smells when it is polluted by different sources or under different levels. When such a lake is rebuilt in a virtual environment, for an immersed avatar, the various smells may need to be provided to warn a user of the pollution conditions. In another example, a mouse can be used to rotate or zoom in or out in traditional GISs. In the VGE, beyond this mode, if the user wants to control his or her immersed avatar by voice, it is necessary to translate the verbal commands (e.g., turn left, toward the south) into manipulative commands. Moreover, the costs of the related interaction tools must be reduced if more participants are expected to be involved in the geographic exploration tasks using VGEs.
VIRTUAL GEOGRAPHIC ENVIRONMENTS

Collaborative work is common in today’s scientific research. Geocollaboration aims to bring different users together to contribute their knowledge and to work cooperatively. This popular approach may be explained by several factors. First, the geographic environment is a complex system that covers various natural and social factors. Thus, the research of the geographic environment requires a closer collaboration of multidiscipline experts. Second, during a task, one person may have basic data only; the suitable process model may be held by another person, whereas another person may know the empirical parameters of a certain experimental area. Collaborative work is also important in such situations. In a VGE, the collaborative component is designed to provide related tools to support such collaboration. When performing collaborative work, the actors of different participants should be defined first, and the total workflow should be designed. Then, every task and subtask must be clearly divided and assigned to different actors. The necessary tools are requested for each task (e.g., provide the metadata of the original data resource, set the model parameters). Conflict detection and communication strategies are also critical during this process. These issues are the main research topics in the development of a collaborative component of a VGE.

Contributions of VGEs

Geographic experiments and analysis

As mentioned, the main goal of developing a VGE is to provide a tool, or a workspace, for geographic experiments and analysis. This geo-oriented conception may distinguish VGEs from other digital worlds, such as digital Earth and virtual games. Geographic experiments aim to simulate the natural and social worlds under one explanatory umbrella (Matthews and Herbert 2008). Conducting geographic experiments in VGEs will provide the following advantages.

First, although traditional geographic experiments have played an important role in geographic research, they are limited with regard to specific times, environments, and scales. A traditional geographic experiment often employs a fieldwork method or conducts process analysis based on a physical model. In the fieldwork method, it is difficult to control the surrounding environment (e.g., unexpected rain), and experiments that aim to make predictions are difficult to conduct. In the model-based method, the shortage of field data may affect the accuracy of a simulation. Moreover, a large region study area in a traditional geographic experiment also requires a significant amount of money and human resources. A VGE can fill in these gaps to some extent. With a VGE, geographic scenarios may be developed at multiple scales, from global to local. Thus, researchers can zoom in or out to visit the study areas and perform experiments at various scales. They can add different factors to the study area to view the outcomes under certain assumptions (e.g., set a virtual rain or flood over a virtual farmland or village). A visual image can be obtained through this low-cost experimental mode through the use of geographic process models at the appropriate scales. Using the tools designed in VGEs, researchers can then adjust the parameters to obtain different experimental results and choose a suitable one to explain the real-world scenario.

In addition, human behavior has proven to be an important influence on the changing geographic environment. Although data on human behavior may be collected and sorted through questionnaires, surveys, or statistical materials, there is an alternative method to fill this gap. Today, psychologists often test human feelings and behaviors by placing volunteers in small virtual spaces
This mode can be referenced by geographic experiment as a strategy to involve human factors. Generally, two types of users are involved with VGEs: professional users (including scientific researchers and decision-makers) and public users. Even without professional knowledge, public users are encouraged to enter into virtual scenarios through multichannel interactive tools to gain their own perceptions and experience. Through this means, they can also indicate their geographic knowledge and living choices. For example, to conduct a human–environment interactive experiment, a virtual polluted small town may be created in a VGE, and 1000 virtual people placed in it. Among these people, 900 may be simulated by ruled agents and the remaining 100 by invited volunteers as avatars, through multichannel interactive tools. These volunteers can perceive the surrounding environment and attempt to make personal suggestions through their judgments. They may control their avatars to change the conditions in the virtual environment by, for instance, planting virtual trees or reducing the number of virtual cars. Avatars and agents can interact with each other as well as with these virtual geographic factors. After recalculation by the corresponding process models, a comprehensive result is identified that considers the natural and social impacts, which will facilitate further comprehensive experiments by researchers (Chen et al. 2013).

Finally, decision-makers generally need to have a plan before starting large projects in the real world (e.g., demolishing a dam or building a large petrochemical plant). Traditional decisions may be made based on documents, reports, and figures, but a recent trend has been for decision-makers to use virtual worlds to obtain a more realistic projection (Bainbridge 2007). Researchers with professional knowledge often conduct some experiments and build various virtual scenarios as proposed solutions (e.g., in city planning) beforehand. Then the decision-makers may choose one scenario as the final solution from the several proposed solutions. However, limitations still exist because it is difficult for decision-makers to comprehend the overall concept proposed by a solution. Thus, the potential problems during the entire process have not been fully identified. With a VGE, decision-makers can more easily identify a solution in collaboration with researchers because, apart from the professional formulas and parameters, they are confronted with the same virtual environment. Decision-makers can join in the experiments directly and ask the researchers to perform potential actions in the virtual environment (e.g., remove 60 chimneys from the southeast sector of a city or reduce the traffic flow in the northwest sector of a town) to determine the outcome. The results are represented after each step is conducted and recomputed, and the solution can be adjusted in a timely manner. Using this mode, less time is required, and the efficiency of the decision-making process can be greatly enhanced.

Geographic knowledge sharing

Geographical thinking and reasoning are important ways of exploring the world and its inhabitants. They produce geographic knowledge of the world’s natural and social phenomena. Golledge (2002) noted that the nature of geographic knowledge over the past 50 years has evolved from phenomenal to intellectual, such as understanding why and how in addition to what and where. Hence a change is required from geographic form representation to geographic process analysis.

A main feature of VGEs is that they own a geographic process model base. A geographic process model is an important foundation for geographic simulation. In some ways, it can
also be regarded as a type of representation of geographic knowledge because the inner mechanism and related formulas in geographic process models are generally produced from a summary and induction of former geographic knowledge (Wen et al. 2013). As discussed earlier, the modeling and simulation component of a VGE promotes a share and reuse mode of geographic process models while performing simulation tasks. When these geographic process models are shared, the inner geographic knowledge can also be shared accordingly.

Comprehensive and complex geographic research requires collaborative work and process models and geographic knowledge from multidisciplinary experts. One challenge is that different experts often have specific understandings of the problems and are accustomed to conducting their research under their own direction and using tools with which they are familiar. In collaborative working, a workspace that can help to achieve a common and familiar understanding is critical. A VGE can provide users with a virtual environment that corresponds to the real geographic world with visual objects and lively phenomena which are familiar and easily recognized. Compared to tools that use abstract or professional symbols to represent geographic information, the VGE can create a better and more universal understanding to enable multidisciplinary experts to communicate with each other and to encourage knowledge sharing (Lin et al. 2013b). In a collaborative study (e.g., a simulation conducted through setting different parameters or the model is modified collaboratively), existing knowledge can be shared and new knowledge may be produced. A network-based VGE is useful for the knowledge sharing even when researchers are scattered across the globe.

**Big data in GIScience**

In 2007 Jim Gary proposed data-intensive science, also known as the fourth paradigm of science, initiating a new chapter in the development of scientific research. Data-intensive science calls for a special data-centered method that integrates the previous stages of experiment, induction, and simulation into one process. Researchers are encouraged to explore various phenomena and to summarize rules by collecting, storing, analyzing, and presenting these big data, which originate from everywhere and at any time. Nature and Science have published special issues on big data. In 2012 the US government, under the guidance of President Obama, proposed its “Big Data Big Deal,” and the US National Science Foundation (NSF) tried to lead the federal efforts in big data.

The same situation occurs in GIScience, where mass data can be acquired by various means (e.g., remote sensing, surveying and mapping, global positioning system (GPS), and sensor networks). Although various data are acquired from different tools with different formats (some data may even be unstructured or its semantic may be missing), a common feature is that these geodata are from the real world and can be coded with a real location at a point in time. Accordingly, location can be regarded as a special factor for the gathering and sorting of these big data, and is the most important reason for maps being used over the years. Today, data are characterized by three or multidimensions (e.g., time dimension), which require a more powerful tool as the data interface and support further presentation and analysis. A VGE would provide such a workspace because it aims to provide digital mirrors of the real world (Lin et al. 2013b). Based on this background information, all data that exist in the real world can be computerized to form a virtual environment with visible and analyzable
information. Specifically, environmental data that are collected by sensors (e.g., air quality, humidity, and temperature) can be inserted into a virtual environment directly through location matching. Using integrated process models and analysis methods, geoeperiment, geoinduction, and geosimulation can be applied to the same virtual environment that is familiar to human beings.

Moreover, social data are also a type of geographic data because the discipline of geography includes both natural and social factors. Social data can now be easily produced by volunteers, (e.g., travel behavior in a city, networking among friends in different areas, or evaluation of a shop or a school), and these data can be provided through personal digital assistants (PDAs) or smartphones. By these means, citizens can easily be involved in contributing to and using big data. Examples have appeared in OpenStreetMap and Google Earth. A VGE can also provide a visual and accessible interface that allows users to enter mass social data (often in real time) into one virtual environment. Combined with the data collected by professional methods, these data can be displayed in the virtual environment to reflect dynamic situations that occurred in the real world. Thus, the data are more useful and easier to understand, and comprehensive geographic phenomena and rules can also be explored and analyzed in this visual way.

Potential contributions to other disciplines

Potential contributions of VGEs to specific disciplines can be established based on the advantages discussed. Here, remote sensing and global change are presented as examples.

The study of fusing multisource, multiresolution remote sensing data into time-series super-resolution images has become increasingly popular in recent years. These studies explain the natural geographic process using mechanistic models and empirical models. However, remote sensing images do not yet fully reflect dynamic geographic phenomena because of the fixed resolution and the precision problem of remote sensing data. One possible solution is to integrate remote sensing data with geographic process models. Using a VGE, remote sensing data and other geodata can be linked together to form a perceptible environment and human visualized thinking can easily be introduced to contribute to the interpretation of remote sensing images. More importantly, a VGE’s geographic process model base provides a bridge between remote sensing data and geosimulation, which can enhance the efficiency of remote sensing information. On the one hand, multisource, multiresolution remote sensing data can be used as parameters for geographic process models and as comparative data to assist the optimization and validation tasks during the simulation. On the other hand, after simulation, the results can be used as supplementary information to facilitate the interpretation of geographic phenomena using remote sensing data.

Global change is a comprehensive discipline involving many factors such as natural, social, economic, diplomatic, and legal ones. Collaboration between different disciplines and departments and various interest groups is urgently needed to research global change. However, the current situation is not conducive to this. For example, an Earth system model is the most important tool to explain global change and the related impacts of factors. At present, unique terms are associated with different models and specific simulations, so it is difficult for different research teams to conduct collaborative research. The successful application of VGEs in geographic research can provide an opportunity for collaboration in global change studies. The critical task is to integrate Earth system models
VIRTUAL GEOGRAPHIC ENVIRONMENTS

into VGEs. Once the geographic process model base is extended to an Earth system model base, VGEs can provide a suitable workspace for collaborative global change studies.

Conclusion

With the development of geographic science and the progress of geographic understanding, corresponding geographic analysis tools are needed to satisfy the requirements of modern users. From acquiring geographic information to studying geographic objects and their relationships, and then exploring dynamic geographic phenomena and processes, the different phases require different tools, for example, the evolution from the use of maps to GIS and finally to VGEs. The use of a virtual environment for the effective management and analysis of the real world has been shown to be a workable method (Bainbridge 2007), as have VGEs for geographic analysis and research (Chen et al. 2015).

Many tools have been approved after several years (or decades) of evolution, including maps and GIS, and VGEs are still evolving. Detailed information and problems have been discussed elsewhere (e.g., Lin et al. 2013a). This entry has focused on the expectations for future development. A VGE is an integrated workspace based on many theories and various technologies, and critical developments in future may see a greater integration of those theories and technologies. It is very important that multidisciplinary researchers engage in this meaningful work and make their contributions to collaborative efforts. When various types of knowledge originating in the real world can be seamlessly integrated into the virtual world, VGEs can provide us with a better opportunity to explore and understand our intriguing world.

SEE ALSO: Big data; Cognition and spatial behavior; CyberGIS; Digital Earth; Geographic information science; Geography education: digital and online trends; Geography and the study of human–environment relations; Public-participation GIS; Representation: 3-D; Representation: dynamic complex systems; Virtual reality; Visualization; Volunteered geographic information

References


Goodchild, Michael F. 2009. “Virtual Geographic Environments as Collective Constructions.” In Virtual Geographic Environments, edited by Lin


Further reading

Virtual reality

Weitao Che
Chinese University of Hong Kong, China

The term “virtual reality” (VR) was used for the first time by Jaron Lanier in 1989. Generally speaking, VR is often referred to as a virtual environment and is a computer technology that enables users to view or immerse themselves in an alternate world. VR is the ability of the user of a constructed view of a limited digitally encoded information domain to change their view in three dimensions causing an update of the view presented to any viewer, especially the user (Fisher and Unwin 2003).

By using real-time computer graphics, users can experience a computer-generated environment as if it is real and they are a part of it. Users of virtual reality systems can move about the computer-generated 3-D scene and explore their surroundings. For example, users can walk through the corridors and into the rooms of a virtual museum to view the exhibits. In addition to navigating through the environment, users are able to interact with it. Objects can be touched, lifted, manipulated, and moved (e.g., users could rearrange displays in the museum). Finally, virtual reality systems are often, but not always, immersive. Immersion refers to the perception that the user is fully surrounded by the virtual environment (Pierce and Aguinis 1997; Stuart 2001). The surroundings of the physical environment are blocked out typically through the use of a head-mounted display and many senses are captivated through images, sounds, and touch to create a realistic virtual environment. According to Heim (Heim 1998), there are three I’s of virtual reality: immersion, interactivity, and information intensity. Immersion comes from devices that isolate the senses sufficiently to make a person feel transported to another place. Interaction comes from the lightning ability of the computer to change the point of view of the scene as fast as the human organism can alter its physical position and perspective. Information intensity is the notion that a virtual world can offer special qualities such as telepresence and artificial entities that show a certain degree of intelligent behavior. The immersion of VR depends not only on the display but also on the interaction between computer and human. The interaction devices have evolved from the mouse and keyboard to the joystick, digital glove, digital helmet, and now even the brain (Lécuyer et al. 2008).

VR allows a user to interact with computer-simulated environments and is often used in a wide variety of applications involving immersive, highly visual, and 3-D environments. This computer-generated world may be either a model of a real-world object such as a house or an abstract world that does not exist in a real sense but is understood by humans such as a chemical molecule or a representation of a set of data. This computer-generated world might be in a completely imaginary science fiction world. Further, because input devices sense the reactions and motions of the operator, the computer modifies the synthetic environment accordingly, creating the illusion of interacting with, and thus being immersed within, the virtual environment. A recent development of VR has been in the area of networking and the Internet. Networked virtual environments and 3-D interfaces to the Internet are among the latest applications of
VIRTUAL REALITY

VR in a growing telecommunications market. Most current virtual reality environments are primarily displayed either on a computer screen (Van Oppenraaij et al. 2009) or through wearable displays (Santos et al. 2009; DeFanti et al. 2009). Some advanced haptic systems also include tactile information (Reif and Walch 2008). Users can interact with a virtual environment either through the use of standard input devices such as a keyboard and mouse or through multimodal devices such as a wired glove (Reif and Walch 2008). Some examples of VR definitions are as follows:

The terms virtual worlds, virtual cockpits, and virtual workstations were used to describe specific projects. In 1989, Jaron Lanier, CEO of VPL, coined the term virtual reality to bring all of the virtual projects under a single rubric. The term therefore typically refers to three-dimensional realities implemented with stereo viewing goggles and reality gloves. (Krueger 1991)

I define a virtual reality experience as any in which the user is effectively immersed in a responsive virtual world. This implies user dynamic control of viewpoint. (Brooks 1999)

It is a simulation in which computer graphics is used to create a realistic-looking world. Moreover, the synthetic world is not static, but responds to the user’s input (gesture, verbal command, etc.). This defines a key feature of virtual reality, which is real-time interactivity. (Burdea and Coiffet 2003)

Based on computer science and relevant scientific technologies, virtual reality (VR) produces a digital environment in which visual perception, sense of hearing, and sense of touch are highly similar to those of actual environment within a certain range. With the help of necessary equipment to interact and interfere with objects in a digital environment, users may have feelings and experiences corresponding to those in the actual environment. Virtual reality (VR) is a scientific method and technology created during the exploration of the nature by human beings to understand, simulate, and better adapt and use the nature. (Zhao 2009)

History: chronological milestones

- In the 1860s, VR developed roots when 360-degree art through panoramic murals began to appear. An example would be Baldassare Peruzzi’s piece titled “Sala delle Prospettive.”
- In 1929, Edwin A. Link invented a type of flight simulator to give passengers the feeling of flight, and this was the first attempt by humans to simulate or emulate physical reality. With the constant development of control technology thereafter, a succession of various simulators came into being.
- In the 1930s, “Pygmalion’s Spectacles” (1935) by Stanley G. Weinbaum describes a goggle-based virtual reality system with a holographic recording of fictional experiences, including smell and touch.
- In 1938, the origin of the term “virtual reality” appeared. The French playwright, poet, actor, and director Antonin Artaud, in his seminal book The Theatre and Its Double (1938), described theater as “la réalité virtuelle,” a virtual reality in which, in the words, “characters, objects, and images take on the phantasmagoric force of alchemy’s visionary internal dramas” (Davis 1998).
- In 1956, Morton Heileg developed a simulator called Sensorama, which is a motorcycle emulator that provided the user with a unique combination of 3-D visuals, stereo sound, vibration, wind sensations and city smells. He advanced some basic thought of VR technology in the Sensorama simulator patent in 1962. The development of electronic technology and the miniaturization of computers facilitated the development
of simulation technology and finally computer simulation technology appeared (Zhao 2009). Unfortunately, Sensorama was not a commercial success (Kalawsky 1993).

• In 1965, Ivan Sutherland suggested that a display could be built that would ultimately provide computer-generated images so realistic that they would be indistinguishable from the real thing. Then, he proposed the ultimate solution of virtual reality: an artificial world construction concept that included interactive graphics, force feedback, sound, smell, and taste (Kalawsky 1993).

• In 1966, Thomas A. Furness III introduced a visual flight stimulator for the air force.

• In 1968, with the help of his student Bob Sproull, Ivan Sutherland created the first virtual reality and augmented reality (AR) head-mounted display (HMD) system – The Sword of Damocles. The Sword of Damocles was primitive both in terms of user interface and realism, and the HMD to be worn by the user was so heavy it had to be suspended from the ceiling. The graphics comprising the virtual environment were simple wire-frame model rooms.

• In 1968, Ivan Sutherland produced a device that is considered the first head-mounted display (HMD) device at Harvard University. He based his design on two small CRTs (cathode ray tubes) mounted on a head band. This system was unique in that it also incorporated a head position sensing system (Kalawsky 1993).

• In 1971, the first prototype of a force-feedback system was realized at the University of North Carolina (UNC).

• In 1975, Krueger nominated the phrase of “artificial reality” as “a conceptual environment, with no existence.” In this system the silhouettes of the users grabbed by the cameras were projected on a large screen.

The participants were able to interact with one another thanks to the image processing techniques that determined their positions in the space of a 2-D screen. Due to the constraints of computer technologies during the 1960s and 1970s, the development of these types of technologies was not very rapid. The related thoughts, concepts, and technologies still stayed in the brewing and formation stage (Mazuryk and Gervautz 1996).

• In 1975, a company called LEEP Systems Inc. (formerly known as Pop-Optix Labs) was founded by Eric Howlett to develop wide-angle lenses for 3-D still photography applications. In 1979, he designed the large expands, extra perspective (LEEP) optical system. LEEP optical systems are now used in almost every virtual environment headset manufactured in the world today (Kalawsky 1993).

• In 1977, the Aspen movie map was created at MIT. The program was a crude virtual simulation of Aspen, Colorado in which users could wander the streets in one of three modes: summer, winter, and polygons. The first two were based on photographs (the researchers actually photographed every possible movement through a street grid of the city in both seasons), and the third was a basic 3-D model of the city.

• In the 1980s, with the development of computer technologies, especially the update of the PC and the computer network, VR technology made significant headway. Several typical VR devices and systems came up during this stage, including DataGlove in 1981–1982, visually coupled airborne systems simulator (VCASS, an advanced flight simulator) in 1982, virtual visual environment display (VIVED) in 1984, virtual interactive environment workstation
VIRTUAL REALITY


- In 1989, Jaron Lanier, the founder of VPL Co., put forward the phrase “virtual reality,” which was generally accepted by researchers and became the specific title of this scientific technology field (Zhao 2009).

- In 1989, BOOM, commercialized by the Fake Space Labs, was a small box containing two CRT monitors that can be viewed through the eye holes. The user can hold the box to the eyes and move through the virtual world while the mechanical arm measures the position and orientation of the box.

- In the 1990s, with the breakthrough and rapid development of computer technology and high performance computation, human–machine interaction technology and equipment, computer networks and communication, as well as huge demands in the significant application fields such as military drill, aeronautics, and astronautics and complicated equipment research, VR technology came into a rapid development stage.

- In 1990, VR technology was discussed at the SIGGRAPH conference held in Dallas, USA, where the main contents of VR technology research were settled – formation technology of real-time 3-D graphics, interaction technology of multisensor, high-resolution display technology, and so forth.

- In 1991, the virtual environment configurable training aid (VECTA) was used with one of the fastest available graphics systems and the highest resolution LCD helmet-mounted displays to give a display that compared more than favorably with other virtual environment systems being demonstrated throughout the world (Kalawsky 1993).

- In 1991, cave automatic virtual environment (CAVE) was invented at the Electronic Visualization Laboratory at the University of Illinois (www.evl.uic.edu/pape/CAVE). More recent VR systems based on the CAVE are the ImmersaDesk (www.evl.uic.edu/pape/CAVE/idesk/) and the IWall. Another display device created for virtual reality is GeoWall, which is a practical 3-D display system that combines advanced graphics PCs and specially calibrated digital projectors to display 3-D presentations with ease and versatility. Using 3-D and high-resolution technology, GeoWall allows audiences to view and interact with spatially immersive content.


- In 1994, Burdea and Coiffet published a book named *Virtual Reality Technology*, in which they used 3I (Immersion, Interaction, and Imagination) to generalize the basic characteristics of VR (Zhao 2009).

- In recent years, there was another huge revolution in the field, namely augmented reality (AR), a technology that “presents a virtual world that enriches, rather than replaces the real world” by means of see-through HMD that superimposes virtual three-dimensional objects on real objects (Mazuryk and Gervautz 1996).

Input devices: tracking devices

One of the three I’s defining virtual reality stands for interactivity. To allow human–computer interaction, it is necessary to use special interfaces designed to input commands of a user into
the computer and to provide feedback from the simulation to the user (Kalawsky 1993). That is, input devices allow the users to interact with and control objects in the virtual environment (Denby and Schofield 1999). Commonly used input devices include tracking devices, the 3-D mouse, and data gloves.

A tracking device reports its location or orientation to the computer. Typically, there is a fixed piece at a known position and additional unit(s) attached to the object being tracked. The position and orientation of the head of the viewer is the absolute minimum of information that immersive VR requires for the proper rendering of images. Additionally, other parts of the body may be tracked, for example, hands to allow interaction, chest or legs to allow the graphical user representation, and so forth. Three-dimensional objects have six degrees of freedom (6DOF): position coordinates (x, y, and z offsets) and orientation (e.g., yaw, pitch, and roll angles). Each tracker must support these data or a subset. In general, there are two types of trackers: those that deliver absolute data (i.e., total position/orientation values) and those that deliver relative data (i.e., a change of data from the last state).

These devices monitor the position and orientation of the head, hand, and/or body of the user. When the user turns his or her head, a tracking device located in a head-mounted display or shutter glasses (discussed later) detects the motion, determines the new position of the head, and updates the viewpoint of the 3-D environment to create the illusion that the environment is staying still while the user moves. Trackers attached to the back of gloves or located in a 3-D mouse detect hand position. In addition, body position can be determined with trackers sewn into a bodysuit or exoskeleton. In the future, eye trackers may be available that will update the view when the eyes of the user move. Many types of trackers are available, including mechanical, optical, ultrasonic, and electromagnetic (Aguinas, Henle, and Beaty 2001).

The most important parameters of 6DOF trackers to evaluate the performance of the different types of position tracker for the given application are as follows (Meyer, Applewhite, and Biocca 1992; Bhatnagar 1993).

- **Update rate** – the rate at which position and orientation measurements are reported by the tracker to the host. Higher update rate values support smoother tracking of movements, but require more processing.
- **Lag (latency)** – the amount of time between the real (physical) action of the user and the beginning of transmission of the report that represents this action. Lower values contribute to better performance.
- **Accuracy** – the difference between the actual 3-D position of the object and the position reported by tracker measurements.
- **Resolution** – the smallest change in position and orientation that can be detected by the tracker. Smaller values contribute to better performance.
- **Range** – the working volume within which the tracker can measure position and orientation with its specified accuracy and resolution, and the angular coverage of the tracker.
- **Jitter** – the change in tracker output when the tracked object is stationary.
- **Drift** – the steady increase in tracker error with time.

**Magnetic trackers**

A magnetic tracker is a noncontact position measurement device that uses a magnetic field produced by a stationary transmitter to determine the real-time position of a moving receiver...
VIRTUAL REALITY

element. As the most commonly used tracking devices in VR, magnetic trackers typically consist of three parts: a static part as the emitter (source), a number of movable parts as the receivers (sensors), and a control station unit. The assembly of emitter and receiver is very similar: they both consist of three mutually perpendicular antennae. As the antennae of the emitter are provided with current, they generate magnetic fields that are picked up by the antennae of the receiver. The receiver sends its measurements (nine values) to the control unit that calculates position and orientation of the given sensor.

There are either alternating fields of 7–14 kHz (for AC magnetic tracker) or pulsed fields (for DC magnetic trackers). The fields penetrate the receiver producing a signal that consists of nine voltages (three for each of the orthogonal transmitter fields). DC magnetic trackers add another three voltages obtained when the transmitter is turned off. These voltages correspond to the local value of the Earth’s DC magnetic field. The receiver consists of three small orthogonal coils when AC magnetic fields are used and three magnetometers (or alternatively Hall effect sensors) when DC magnetic fields are used. The receiver voltages are sampled by an electronic unit which uses a calibration algorithm to determine the position/orientation of the receiver in relation to the transmitter. These data packets (three positions and three rotation angles) are subsequently transmitted to a host computer via communication lines. If the receiver is attached to a remote moving object, then the computer can indirectly track the motion of that object relative to the fixed transmitter.

The major advantage of the magnetic tracker is that it has no line of sight restriction and wireless systems have become available thereby reducing encumbrances on the participant. However, there are also some limitations. One of the limitations is that metal in the environment can cause magnetic interference. Another limitation is the short range of the magnetic field that is generated.

Acoustic (ultrasonic) trackers

An acoustic tracker is a noncontact position measurement device that uses an ultrasonic signal produced by a stationary transmitter to determine the real-time position of a moving receiver element. Acoustic trackers usually have three components: a transmitter, a receiver, and an electronic unit similar to their magnetic counterparts. The difference is that the transmitter is a set of three ultrasonic speakers mounted approximately 30 cm from each other on a rigid and fixed triangular frame. Similarly, the receiver is a set of three microphones mounted on a smaller rigid triangular frame (Burdea and Coiffet 2003). There are two types of acoustic trackers. The acoustic trackers either use time-of-flight (TOF) or phase-coherent measurements to determine the distance between a pair of points. TOF trackers (e.g., Logitech 6DOF Ultrasonic Head Tracker, Mattel Power Glove) measure the flight time of short ultrasonic pulses from the source to the sensor. PC trackers compare the phase of a reference signal with the phase of the signal received by the sensors. The phase difference of \(360^\circ\) is equivalent to the distance of one wavelength. The difference between two successive measurements of phases allows computation of the distance change since the last measurement. As this method delivers relative data (so the error tends to accumulate with time), development of PC trackers was abandoned (Mazuryk and Gervautz 1996).

Optical trackers

An optical tracker is a noncontact position measurement device that uses optical sensing to
determine the real-time position/orientation of an object. Of the number of ways this can be achieved, the most common is to make use of a video camera that acts as an electronic eye to “watch” the tracked object or person. The video camera is normally in a fixed location. Computer vision techniques are then used to determine the object’s position based on what the camera “sees.” In some cases, light-sensing devices other than video cameras can be used (Sherman and Craig 2002). Generally, the light sensing can be divided into three categories (Mazuryk and Gervautz 1996).

- Beacon trackers – this approach uses a group of beacons (e.g., LEDs) and a set of cameras capturing images of the pattern beacons.
- Pattern recognition – these systems do not use any beacons, they determine position and orientation by comparing known patterns to the sensed patterns.
- Laser ranging – these systems transmit the laser light that is passed through a diffraction grating onto the object.

The limitation of optical trackers is that the line of sight between the tracked person or object and the camera must always be clear.

Mechanical trackers

A mechanical tracker consists of a serial or parallel kinematic structure composed of links interconnected using sensorized joints. The first tracker used in a VR simulation was the mechanical arm that supported the CRT-based HMD developed by Sutherland. The motion of the head of the user was tracked with regard to the ceiling arm attachment. The dimensions of each link segment are known and used by the direct kinematics computational model stored in the computer. This model allows the determination of the position and orientation of one end of the mechanical tracker relative to the other, based on the real-time reading of the tracker joint sensors. By attaching one end of the tracker to the desk or floor and the other to an object, the computer can track the 3-D position of the object relative to the fixed end of the arm.

Body tracking

Body tracking is the ability of the VR system to sense the position and actions of the participants. The particular components of movement that are tracked depend on the body part and how the system is implemented. The body parts and techniques of body tracking commonly used in VR applications include: (i) head tracking, (ii) hand and finger tracking, (iii) eye tracking, (iv) torso tracking, (v) foot tracking, and (vi) tracking other body parts (Sherman and Craig 2002).

Head tracking allows proper rendering of images from the point of view of the user. The advantage of the head tracking is that a motion parallax cue can be provided, which improves the depth perception. One more important aspect can be considered: the visual acuity of the eye changes with the arc distance from the line of sight.

Tracking the hand and fingers is generally done to give the user a method of interacting with the world. In multiparticipant space, hand gestures can also provide communication between participants. A hand can be tracked by attaching a tracker unit near the wrist or through the use of a tracked handheld device. A glove input device is used to track the positions of the fingers of the user and other flexions of the hand.

In general, most important eye-tracking technologies can be grouped as: (i) limbus tracking; (ii) image tracking, and (iii) electro-oculography (EOG) (Mazuryk and Gervautz 1996).
VIRTUAL REALITY

Very few VR applications actually track the torso of the participant, although when an avatar of the user is displayed, this avatar often includes a torso with certain assumptions made about its position based on head and hand positions. However, the torso is actually a better indicator of the direction the body is facing than either the head or hands. The bearing of torso might be a better element to base navigational direction on than head or hand positions (Sherman and Craig 2002).

Tracking the feet provides an obvious means of determining the speed and direction a user wishes to travel. The obvious method of determining foot movement is to track the position of each foot. The most common way to track the position of each foot is to use magnetic trackers. Other body parts can be tracked and used to control various aspects of a virtual world. These items include body functions, such as temperature, perspiration, heart rate, respiration rate, emotional state, and brain waves, which could be measured simply to monitor the condition of the participants as they experience a world, or they might be used to control the world— to determine what experiences are most relaxing, and so forth.

3-D mice and bats

This basic and simple user interaction tool is in general a joystick-like 6DOF device that can be moved in space by hand, equipped with a tracker sensor to determine its position/orientation, and a few buttons that may trigger some actions. Some 3-D mice may be equipped with a thumb ball for additional movement control.

The 3-D mouse is a handheld device used to navigate the user through the virtual environment and interact with objects. To navigate the environment, the gaze of the user determines the direction and clicking a button on the mouse moves the user forward. The mouse can also lift and move objects in the virtual environment. When the user collides with an object, a mouse button is pushed to lift the object, and the object then moves with the user until a button is pushed to drop the object. As previously mentioned, position trackers located in the mouse determine the hand position of the object and, subsequently, the user.

Data gloves

Gloves are devices that measure the real-time position of the fingers (and sometimes wrist) of the user in order to allow natural, gesture-recognition-based interaction with the virtual environment. The measurement of finger flexion is achieved with the help of various sensors. VPL DataGlove, Virtex CyberGlove, Mattel PowerGlove, Pinch Glove, the 5DT Data Glove, and the Didjiglove are representative gloves that have appeared in VR systems. The use of gloves allows the user richer interaction than the 3-D mouse because hand gestures may be recognized and translated into proper actions. Additionally gloves are equipped with a tracker attached to the wrist of the user to measure the position and orientation of the wrist.

These devices indicate the position of the fingers and hand, interact with the virtual

Other physical input devices

In addition to trackers that capture movements of the users, many other physical input devices were developed to make human–computer interaction easier and more intuitive. For full freedom of movement, 3-D input devices seem the most natural. Attached to our body or handheld, 3-D input devices are generally used to select, move, and modify virtual objects. Devices range from simple handheld objects to large cockpit-style platforms in which the user can sit.
VIRTUAL REALITY

environment, and signal commands to the computer through gestures. Gloves typically have thin fiber optic sensors sewn into the cloth that bend and stretch when the fingers move (Biocca and Delaney 1995) and communicate finger movement to the computer. A position tracker, usually attached to the back of the wrist, monitors the position and orientation of the hand and can be used to create a 3-D representation of the hand of the user in the virtual environment. When the glove collides with an object, the computer can be signaled to lift the object and move it. Gloves can also be designed to provide force and tactile feedback. Finally, bodysuits or exoskeletons are available, which function similarly to the glove, but they track movement of the entire body and display a virtual body to which the users can relate (Mazuryk and Gervautz 1996).

Speech recognition (audio input)

Speech recognition systems are becoming increasingly practical, and they could provide an excellent opportunity for natural communication with computer systems. Some speech recognition systems use speaker-independent speech recognition while others use training where an individual speaker reads sections of text into the speech recognition system (speaker-dependent). These systems analyze the specific voice of a person and use this voice to fine-tune the recognition of the speech of that person, resulting in more accurate transcription.

Output devices and representation

Output devices convey information from the computer to the user about the virtual environment. There are three groups of output devices: visual displays, haptic devices that convey force and tactile information to the body of the user, and audio devices. Virtual environments can be displayed and viewed by users in many different ways. Regardless of the display used, the display must create realistic scenes that correspond to what the user would see as he or she navigates the virtual environment. Common methods are shutter glasses, head-mounted displays (HMDs), binocular omni-orientation monitor (BOOM) systems, and cave automatic virtual environment (CAVE) systems. Shutter glasses create the 3-D effect for some desktop virtual reality systems. A 3-D display monitor shows alternate right and left images at a fast rate and the shutter glasses alternately allow light to reach the eyes. When a left image appears on the display monitor, the shutter glasses receive a synchronizing signal from an infrared device placed on top of the monitor to shut the right lens and when a right image appears, the glasses are signaled to shut the left lens. The right and left images are fused together by the brain to produce the 3-D image. The HMD presents the virtual environment through a device mounted on the head of the user that can resemble a helmet with a visor in front of the eyes or a scuba mask. The HMD contains a separate display for each eye so that each eye sees a different view of the same image. As with shutter glasses, these images are fused together by the brain to produce the 3-D effect. HMDs also have position trackers to signal to the computer the position of the head so that the view of the virtual environment can be updated to match the head movements of the user. HMDs may also have sound devices such as headphones or earphones.

BOOM systems have become a popular alternative to the HMD because users do not support the image display on their heads. Rather, the BOOM involves a viewer mounted on a stand, which the user holds to his or her face to view the virtual environment. The user can
VIRTUAL REALITY

sit or stand and can move the viewer using handles to observe different aspects of the virtual environment. In addition, the user can easily shift between the virtual and real environment because no equipment must be put on and taken off, as with the HMD. Although motion may be slightly restricted due to the design of the BOOM, the quality of the graphics display is higher than the HMD.

A CAVE system allows one or more users to be completely immersed in the virtual environment. Users enter a small room with large video projection walls, which are used to surround users on all sides with computer-generated images. Users wear shutter glasses to view the environment and a position tracker is mounted on the glasses to determine head position. Haptic devices provide output from the computer to the user by simulating force and tactile feedback. Users can “feel” force, tension, friction, pressure, temperature, and speed of objects. Force-feedback devices convey information about the resistance of surfaces (e.g., prevents users from walking through walls) and the gravity, weight, and solidity of objects. This feedback is provided through mechanisms such as joysticks, steering wheels, and handgrips. As mentioned earlier, gloves and exoskeletons or bodysuits can be designed to convey force feedback. Tactile feedback provides information about the surface and texture of objects in the virtual world. Many types of approaches can be used to communicate touch in the virtual environment. Pneumatic devices simulate touch by inflating and deflating air pockets, air jets, or air rings that are attached to gloves and touch the fingers of the user. Vibrotactile methods signal touch through vibration on the skin of the user. Electrotactile devices send pulses of electricity through electrodes touching the skin. Finally, functional neuromuscular devices directly stimulate the nervous system, but are rarely used because of their invasiveness. Audio devices are not a necessary component of most virtual reality applications, but they enhance the reality of the experience and help create immersion by replacing sounds from the physical environment with sounds produced in the virtual environment. Headphones are usually built into head-mounted displays. Additionally, earplugs, which are lighter and less intrusive, can be used. Simple audio systems play back digitally recorded sounds into the headphones of the user or through desktop speakers. Higher end systems use spatialization or 3-D sound that creates the illusion of sound originating outside the head of the user, coming from a particular location away from the user, or moving through the virtual environment.

Visual display devices

Human visual perception is generally considered to be the primary means of gaining information about physical spaces and appearances of objects. Realistic rendering of 3-D graphics and image in real time is the most important means to achieve VR visual perception and a core technology to construct a virtual environment. Realism and real time are two important safeguards for the sensation of system immersion and interactivity in VR, and a pair of prominent conflicts at the same time. Realistic rendering of 3-D graphics and image in real time is mostly about improving realism and real-time capabilities.

The visual interface of VR often provides the most salient and detailed information regarding the virtual world. In general, a complete visual interface consists of two primary components: (i) a visual display surface and an attendant optical system that paints a detailed, time-varying pattern of light onto the retina of the user; and (ii) a system for positioning the visual display surface relative to the eye of the operator. However, the current technical limitations make both the optimal use of normal visual sensory
capabilities and the real-time display of detailed, continuous-motion imagery a difficult challenge for VR developers. In particular, the inadequate optical arrangements between the display and the user’s eyes or the inadequate channel capacity in the display unit may produce spatial distortion and image degradation. For these issues, none of the many display options that currently exist are completely adequate across all applications.

Two major classes of visual display systems are available: head-mounted and off-head displays. Head-mounted displays (HMDs) are physically coupled to the head of the operator by mounting display hardware on a helmet or headband worn by the user.

The most frequently used display type for HMDs is the backlit liquid crystal display (LCD). In many HMDs, all the visual output is generated by computer. The small size of displays used in HMDs brings with it small field of view (FOV). To enhance the viewing range special optics may be used such as LEEP or Fresnel lenses. Both of these approaches require a pre-distortion of the image that will be viewed through the special optics. Wide field optics is used, for example, by VPL Research for the construction of their HMDs. Beyond LCDs, a virtual retinal display (VRD) has been proposed. A prototype of a VRD developed at the HIT-Lab uses a modulated laser light that projects the image directly onto the retina of the user. Although application of this system in practice is currently questionable, VRD offers great potential quality improvement – the goal of the project is to achieve a resolution of 4000 × 3000 and a FOV bigger than 100°.

VR visual output devices

Different types of VR systems – from desktop to full immersion – use different output visual displays that can vary from a standard computer monitor to a sophisticated HMD. The following section will present an overview of the most often used displays in VR.

- 3-D glasses: the simplest VR systems use only a monitor to present the scene to the user. However, the “window onto a world” paradigm can be enhanced by adding a stereo view by use of LCD shutter glasses. LCD shutter glasses support a 3-D view using sequential stereo: with high frequency they close and open eye views in turn, when the proper images are presented on the monitor. An alternative solution uses a projection screen instead of a CRT monitor. In this case polarization of light is possible and cheap polarization glasses can be used to extract proper images for each of the eyes. A head movement tracking can be added to support the user with motion parallax depth cues and increase the realism of the images presented.

- Surround displays: an alternative to standard desktop monitors is large projection screens that offer not only better image quality but also a wider field of view, which makes them very attractive for VR applications. The total immersion demand may be fulfilled by a CAVE-like display, where the user is surrounded by multiple flat screens or one domed screen. Ideally total immersion would support a full 360° field of view. The disadvantage of such projection systems is that they are large, expensive, fragile, and require precise hardware setup.

- Binocular omni-oriented monitors (BOOM): developed and commercialized by Fake Space Labs, BOOMs are complex devices supporting both mechanical tracking and stereoscopic display technology. Two visual displays (for stereo view) are placed in a box mounted to a mechanical arm. The box can be held by the user and the monitors can be watched through
VIRTUAL REALITY

two holes. As the mechanical construction supports usually counter-balance, the displays used in the BOOMs need to be neither small nor lightweight. Therefore, CRT technology can be used for better resolution and image quality.

- **Head-mounted (coupled) displays (HMDs):** HMDs are headsets incorporating two small monitors placed in front of the eyes of the user. The images are presented to the user based on his/her current position and orientation as measured by a tracker. Because the HMD is mounted to the head of the user, the HMD must fulfill strict ergonomic requirements: it should be relatively light, comfortable, and easy to put on and take off. Like any visual display, the HMD should also have the best quality. These demands force engineers to make hard trade-offs. Consequently, the prices and quality of HMDs vary dramatically: from approximately US$800 for a low-cost, low-quality device to about $1 million for a high-tech military HMD.

- **HMDs can be divided into two principal groups:** opaque and see-through. Opaque HMDs totally replace the view of the user with images of the virtual world and can be used in applications that create their own world such as architectural walkthroughs, scientific visualization, games, and so forth. See-through HMDs superimpose computer-generated images on real objects, augmenting the real world with additional information. Most of the HMDs currently available on the market support stereo viewing and can be driven either with PAL or NTSC monitor signals.

**Audio displays**

Hearing is the second greatest source of perception of human beings next to vision and approximately 15% of human perception information of the objective world depends on hearing. Hearing cannot only accompany the sound for visual pictures but also complement information out of vision to enhance the sense of space and realism of the virtual world. Sound greatly enhances the ability of the participant to become mentally immersed in the world. From ambient sounds that give cues to the size, nature, and mood of the setting to sounds associated with particular objects or characters nearby, aural representation is key to user understanding and enjoyment (Sherman and Craig 2002).

Sound can powerfully enhance the human perception ability. As an addition to the visual information, auditory information can offer several benefits like:

- an additional channel for the communication data;
- the ability to perceive information that is outside the visual display;
- the ability to alert the user and focus signals that attract the user or warn him/her;
- spatial orientation cues;
- the possibility for parallel perception of numerous information streams.

In the simplest case, sound can indicate completion of some tasks or signal that some conditions have been met (like collision with objects, placing of bodies, etc.) without cluttering the screen. This technique has been used for a long time in desktop systems and includes click-keyboards or desktop sounds (including speech audio) confirming certain actions or system events. For these purposes a monaural sound is sufficient. In VR, however, more convincing three-dimensional auditory displays can be used to simulate distance, direction, material, and spatial information about the environment. A successful area of application is architectural visualization. Sound cues lead to better
VIRTUAL REALITY

spatial impression and allow visual and acoustic evaluation of building models, which is a significant improvement for walkthrough applications.

To generate these spatial cues convincingly, basic knowledge of the human sound localization system is required. The duplex theory accentuates the two primary hints that play a leading role in this process: interaural time difference (ITD) and interaural intensity difference (IID).

In addition to these two basic cues, other aspects should also be considered. Spatial information reaching our ears depends strongly on the distance from the source, environment geometry, and material properties of objects in the vicinity, which influences the creation of echoes. At least the pinnae (outer ear) of the listener have influence on spectral soundwave shaping by reflecting and refracting it slightly. All these subtle effects altering the sound received by the ear create the effect of spatial sound perception.

To synthesize the sound that matches human perceptual abilities artificially is a complex and computationally expensive task. Moreover, the proper synchronization of sound and visual events is extremely important not to disorient the user by contradictory cues. Three basic steps are required for successful simulation of virtual acoustic environments (Astheimer 1995).

- **Sound generation** – every action in the real world generates some sound (steps, object collisions, etc.). For acoustic simulation sound can be either generated (synthesized) or sampled and played back.

- **Spatial propagation** – this is the most expensive part of the whole simulation. Spatial propagation involves calculation of spatial propagation of sound waves through the environment. Ideally all the reflections (echoes) from different objects and walls should be considered, according to the material properties and surface shapes of these objects.

- **Mapping of parameters** – the calculated parameters should be mapped onto proper sounds that will be delivered to the listener through headphones or speakers. To do this, all rendered chunks (impulses) should be convoluted together by the use of special filters, to produce one sound. The directional effect can be achieved by use of ITD and IID cues and calculation of head-related transfer functions (HRTFs).

Haptic displays

Both vision and audition presentation provide noncontact perception information, while haptic (force/tactility) representation provides contact perception information of virtual objects and the virtual world. The participants could obtain vivid force/tactility experience in contact interaction with virtual objects and their senses of reality, immersion can be directly enhanced and the application field of VR could be expanded.

Haptic representation has the features of: (i) the cues it provides about the world are the ones that are most trusted by the cognitive system when confusing or conflicting information is presented to the senses (seeing is believing, but touching is knowing); and (ii) the perception happens only local to the user, on or near their skin or inside their body.

Haptic sensations perceived by humans can be divided into two main groups.

1. **Kinesthetic (force) feedback** – forces sensed by the muscles, joints, and tendons.
2. **Tactile feedback** – includes feedback through the skin, such as sense of touch, temperature, texture, or pressure on the skin surface.

These perception issues are extremely important when performing some precise manipulation.
VIRTUAL REALITY

tasks. Manipulating every object in the real world always causes a collision between the hand and the object, which is perceived as haptic feedback. Therefore, the many dexterous manipulators and some data gloves are equipped with devices simulating these sensations. A remote interaction with fragile objects (e.g., human eye surgery or laboratory tasks) could not be completed accurately without proper haptic cues.

Kinesthetic or force feedback is used when object placement and manipulation requires a proper force, which is a quite natural phenomenon for humans. To increase the naturalness of interaction in VR, some devices are equipped with force feedback, including a variety of manipulators from simple gloves to sophisticated and mechanically complex exoskeletal hand masters.

In addition to these hand mounted manipulators that are supposed to simulate the real interaction with objects, several studies have proven that presence of force feedback increases efficiency of various placing and manipulation tasks. Therefore, many devices were developed: from the simplest devices such as joysticks or desktop mice with kinesthetic feedback, through force simulation with the help of tense strings in the UNC GROPE project. All of them confirmed qualitative improvements of interaction – haptic display together with visual display can enhance the perception and understanding of natural phenomena such as scalar or vector fields. A recent prototype of GROPE consists of a ceiling-mounted arm coupled with a computer and was used by the chemists for a drug-enzyme docking procedure. The application in other areas such as investigating molecules or protein property seems to bring significant advantages.

Tactile feedback is much more subtle than force feedback and therefore more difficult to generate artificially. The simulation may be achieved with the help of vibrating nodules, inflatable bubbles or electro rheological fluids (fluids whose viscosity increases with an applied electric field) placed under the surface of a glove. All these currently available technologies are unable to provide the whole bandwidth of sensory data we obtain through our skin. They generate the indication of touch with some surface but do not allow the user to recognize its structure.

Modeling methods

The topics of geometric modeling include virtual world space definition, virtual observer location, perspective projection, 3-D modeling and clipping, 3-D space curves, 3-D boundary representation, geometrical transformation, and modeling transformations for translation, scale, reflection, and rotation. The topics of physical modeling include illumination models, reflection models, shading algorithms, radiosity and realism, and collision detection. The topics of kinematics modeling include picking, flying, scaling virtual environments (VEs), and the dynamics of numbers. The topics of behavior modeling include free-form deformation, shape and object inbetweening, the animation of objects, animating the virtual environment (VE), particle systems, interactive using navigation, selection, and/or manipulation for a convincing simulation.

Geometric modeling

Geometric modeling describes the shape of virtual objects (polygons, triangles, vertices, splines) as well as their appearance (surface texture, surface illumination, and color).

- Modeling based on depth range image. Each pixel of range image saves the range information of the first intersection point of
related light and the scene shot, the modeling process of which involves the acquisition, registration, surface reconstruction, and repair of range image.

- Image-based modeling method. The image-based modeling method uses photos to construct vivid 3-D models, which can be divided into active and passive. The active type obtains 3-D information of objects actively through the control of the illumination of the scene, which has high reconstruction accuracy and an easy algorithm. However, artificial reconstruction of clues, whose operating process is complicated, is required. The passive type receives the intensity of the information of scenes passively, and then conducts 3-D reconstruction through analyzing passive clues of images such as tone, shadow, focal length, texture, and parallax. This method barely restricts the scale and position of modeling scenarios, but has a poor accuracy and complicated algorithm.

- Material illumination modeling method. Computer graphics normally uses texture mapping technology to map the 2-D image of the surface reflection property of a reflected object onto the 3-D model surface. This method only represents the color information of the object under certain specified illumination and visual points, but many surface reflection properties of objects possess a large amount of high gloss information. Simple texture mapping technology is unable to realistically present plentiful light reflection and refraction effects of objects. Therefore, the research on acquisition and modeling technology of surface material properties has significant meaning in the realism modeling field.

- Volume illumination model. The volume illumination model is the mathematical description of physical phenomena, such as light intensity change, when the light goes through voxel, serving as the basis for direct volume rendering. The current normal volume illumination models are the active-attenuation model, the varying density emission model, and material classification and hybrid model.

- Field modeling method. Physical fields such as electromagnetic field and gravitational field are all invisible. However, some VR applications often require the distribution, range of action and effects of certain fields in the virtual environment (such as the scanning of a radar beam) to be delivered to users through visualization. Therefore, the research on field modeling could be divided into visualization and appearance modeling. Fields such as vector field and scalar field usually use grids for modeling. The grids mainly refer to a regular grid, rectangular grid, curvilinear grid, nonstructural grid, scattered data, and so forth.

**Kinematics modeling**

Kinematics modeling determines the location of 3-D objects with respect to a world system of coordinates as well as their motion in the virtual world. Object kinematics is governed by parent–child hierarchical relations, with the motion of a parent object affecting that of its child. A further aspect of kinematics modeling is that the camera image needs to be transformed and projected in the 2-D display window in order to provide visual feedback to the user.

**Physical modeling**

Physical modeling integrates the physical characteristics of the object, which include weight, inertia, surface roughness, compliance (hard or soft), deformation mode (elastic or plastic), and
VIRTUAL REALITY

so forth. These features, together with object behavior, bring more realism to the virtual world model. The computation load required by physical modeling is assigned to the haptic rendering pipeline.

- Rigid body motion modeling method. If each point of the object cannot be translated or rotated correspondingly to any other point of this object, then this object is called the rigid body. This modeling method only has to consider the position and direction change of rigid body motion modeling in environment. The object’s deformation can be neglected. The content referred to includes mainly the motion simulation, collision detection, connection, and constraint of the rigid body.

- Flexible object modeling method. In the real world, many objects have no rigid body characteristics. They deform during movement or interaction, namely they are flexible objects. The flexible object modeling refers mainly to geometrically based methods, physically based method, and collision detection of flexible objects. According to the shape and appearance, the modeling of geometrically based method usually adopts a catenary method and the B spline method, resulting in a limited effect on realistic simulation. This section discusses mainly the physically based flexible object modeling method and collision detection.

- Physically based modeling method. This method divides the flexible objects into particles in cells and treats flexible object motion as a particle motion under various physical quantity effects such as force and energy. Grid structures of flexible objects are of discrete type and continuous type. The point particle model, particle model, and spring-point particle model deal with the former type, while the spring deformation model, air aerodynamic model, wave propagation model, finite element model, and so forth, deal with the latter type.

- Virtual human motion modeling method. The virtual human is an important component of the virtual environment. The research on virtual human motion modeling addresses mainly the acquisition and treatment of motion data and control of motion.

Behavior modeling

Behavior modeling is the “modeling of human behaviors or representations required to be expressed in military emulation” (Pew and Mavor 1998). It is the major content of autonomous object research in the virtual environment, which is dated from the research on computer-generated forces (CGFs) in the virtual battlefield (Zhao 2009).

Most of the current VR systems also model virtual humans, called avatars. Avatar is defined as “a user embodiment in a collaborative virtual environment” (Gerhard, Moore, and Hobbs 2004), which has agency (an ability to perform actions) and is controlled by a human agent in real time (Bell 2008). In the sense of shared virtual environment, the more surprising developments are coming from the impressive successes of Second Life (http://secondlife.com; Warburton and García 2009), putting avatars into the 3-D environment and extending their applications via their scripting languages, which have the potential to move how we share, visualize, and communicate geographic data to a new collaborative level (Hu et al. 2010).

New types of virtual reality

Mixed reality

Mixed reality (MR) refers to the merging of real and virtual worlds to produce new environments
and visualizations where physical and digital objects coexist and interact in real time. Not taking place only in the physical world or the virtual world, but a mix of reality and virtual reality, encompassing augmented reality and augmented virtuality (Milgram and Kishino 1994).

In an MR system, the user can interact with and visualize both real and virtual objects in the same context. For example, users could interact with a real-world anesthesia machine as a tangible user interface and visualize an abstract simulation of the gas flows in the context of the real machine. Through this merging of virtual (the abstract simulation) and real (the anesthesia machine) spaces, such a system might also enable the user to mentally merge the different knowledge types – abstract knowledge learned from the simulation and concrete knowledge learned from the real world (Quarles et al. 2008).

Cross reality (dual reality)

Proposed by the MIT Response Environmental Group (ResEnv), cross reality is defined as the ubiquitous mixed reality environment that comes from the fusion of sensor/actuator infrastructure and 3-D virtual environments for an environment resulting from the interplay between the real world and the virtual world, as mediated by networks of sensors and actuators. The difference between cross reality and AR/MR is that the latter is always restricted to head-worn or other wearable/mobile devices, while cross reality is a ubiquitous environment (Lifton 2007).

3-D online virtual world (Second Life, etc.)

As three-dimensional (3-D) virtual worlds such as Second Life and OpenSimulator (OpenSim; http://opensimulator.org) become more mature and technically advanced, they can be used for serious applications that are driven by research (Von Kapri et al. 2009). OpenSim is an open source multiplatform, multiuser 3-D application server that can be used to create a virtual environment (or world) that can be accessed through a variety of clients on multiple protocols. Both Second Life and OpenSim depend on user-created content and have programming interfaces to create new functionality. The networked virtual environment allows for intuitive interaction.

Online 3-D has become increasingly popular in recent years. The 3-D Internet can be defined as one that is navigable in three dimensions rather than two. Instead of websites, you have explorable regions. Instead of 2-D text chatting, you have 3-D avatar chats (Barnes 2010).

SEE ALSO: Augmented reality; Virtual geographic environments

References


Visibility analysis

Juan M. Domingo-Santos
Rubén Fernández-de-Villarán
University of Huelva, Spain

University of Huelva, Spain

The very definition of the word “landscape” characterizes it as “a view or prospect of natural inland scenery, such as can be taken at a glance from one point of view.” Thus, it is clear that for land-use planning and landscape management, visibility is a key factor for decision-making regarding visual impact. In urban areas, dominated by vertical structures, visibility is also a major concern in relation to touristic interest, traffic control, security, and citizen wellbeing, among other issues.

Initially, it is important to recognize that visibility is a complex concept that is used in different fields and deals with multiple factors. As an example, in meteorology, visibility concerns the transparency of the atmosphere, and is defined in terms of viewable distance, whereas in geography it refers mainly to the arrangement of objects in the landscape but it is influenced by other factors, such as the atmospheric conditions, as discussed in this entry.

Thinking of visibility in geographic terms, visibility analyses determine those portions of the landscape that can be seen and the content and composition of available views. With respect to what “can be seen,” it is assumed that the observers have normal vision and there are average-to-good observing conditions. In any case, the “can be seen” idea invites us to think of visibility in probabilistic terms, that is, as a parameter that could be analyzed with a fuzzy approach (Anile et al. 2003), rather than a well-defined deterministic function.

Visibility is based on the existence of a clear line of sight between two points, that is, there are no obstacles that would block light traveling from one point to the other. Beyond this evident requirement, there are many other factors constraining visibility that have a strong influence on what the observer actually “sees;” that is, human perception is a combination of optical physics and atmospheric effects, together with psychological and cultural factors. Gross (1991) identifies the following as the main factors influencing the visual impact of objects on the landscape: visual acuity of the human eye, color contrast with the background, atmospheric extinction (the manner in which light is absorbed and scattered), and surface area covered by the image of the object on the retina. These visibility factors are covered in more detail here.

Visual acuity

Visual acuity is the ability of the eye to see fine detail, technically defined as the spatial resolving capacity of the visual system. This ability is mainly limited by optical aberrations, caused by diffraction in the eye lens and by photoreceptor density in the eye. It is also influenced by other factors including refractive error, illumination, contrast and the location of the retina being stimulated (Kalloniatis and Luu 2007).
VISIBILITY ANALYSIS

Diffraction aberrations

If a point source is taken, the image will be distributed on the retina as a point spread function due to distortions created by diffraction of light passing through the optics of the eye. The point spread function follows an Airy disk pattern that presents a central maximum of light and a series of rings of diffracted orders with very low and decreasing intensity of light (Figure 1).

Two point sources produce two point spread functions at the back of the eye. Rayleigh’s criterion states that two points or lines are just resolved if the peak of one point spread function lies on the first trough of the other point spread function; otherwise, the points are seen as a single element (Figure 2).

Dimension of the retinal mosaic

Retinal cone density is another limiting factor. In this case, two points are resolved if there is a row of unstimulated cones in between rows of stimulated cones (where cones are the basic photoreceptor cells on the retina that enable humans to see). Based on cone spacing, a maximum of approximately 0.47 arc minutes is the best possible resolution, which is well above conventional clinical measures, ranging from 1 to 0.63 arc minutes (Colenbrander n.d.), as this measured spacing does not take into account the optics of the eye or post-receptor neural processing.

Other conditions affecting visual acuity

In addition to the two main limiting factors cited above, there are particular observing conditions and eye disorders that may affect visual acuity. These factors should not be considered in general visibility analysis but they can reduce visibility further. The most significant are briefly described (Kalloniatis and Luu 2007).

- Refractive error: visual diseases such as myopia and hyperopia cause images to be focused away from the retina, affecting resolution.
- Pupil size: a mid-size pupil (3–5 mm) is optimal; smaller or larger sizes decrease resolution.
- Illumination: visual acuity is greatly affected by the level of background luminance.
• Exposure time of the target: although there is no simple function relating acuity and exposure time, this may be significant for detection of linear shapes.
• Area of the retina stimulated: visual acuity is greatest at the center of fixation. At a distance of five arc minutes from the center of fixation, there is a measurable loss in visual acuity, this rising to around a 25% loss at 10 arc minutes (1/6 of a degree).

**Color contrast and atmospheric extinction**

Colors are coded under different standardized systems, or color spaces, which allow the precise reproduction of images and the quantification of differences that will make an object visible against a background. For a given color space, the color difference between two objects is expressed by the Euclidean distance between the color components. More generally, a color space links the physical properties of light, particularly their electromagnetic signatures, to how they are perceived by humans.

Taking the CIE-Yxy color space and considering that color components will be influenced by the atmospheric attenuation, related to distance \( r \) to the observer, the aforementioned Euclidean distance for the color space CIE76, Delta-E or \( \Delta E \) can be expressed as:

\[
\Delta E = \sqrt{C(r)^2 + (\Delta x(r))^2 + (\Delta y(r))^2}
\]  

where the luminance contrast \( C(r) \) depends on the luminances of the object \( Y_O \) and the background \( Y_H \):

\[
C(r) = \frac{|Y_O(r) - Y_H(r)|}{Y_H(r)}
\]

and the color distance between chromaticity coordinates \( x \) and \( y \) in CIE-Yxy color space is given by:

\[
\Delta x = x_O - x_H; \quad \Delta y = y_O - y_H
\]

where \((Y_O(r), x_O(r), y_O(r))\) is the color location of the object in the color space, attenuated by distance \( r \), and \((Y_H(r), x_H(r), y_H(r))\) is the color location of the background, attenuated by distance \( r \).

Atmospheric attenuation of light (Overington 1976) is caused by scattering and by absorption. Specifically, light is scattered by small particles and water droplets suspended in the atmosphere (aerosol scattering), as well as by molecules of air (Rayleigh scattering). Absorption generally depends on atmospheric composition and the wavelength of the light but is generally negligible for the visible region on a clear day when water content is low. Attenuation follows an exponential law in homogeneous air:

\[
F(r) = F_0 e^{-\sigma_e r}
\]

where \( F_0 \) is the original light flux at the source, \( r \) is the distance from the source to the observer, and \( \sigma_e \) is the extinction coefficient.

Scattered light from other light sources entering the viewing path (air light) will increase the apparent luminance of the object. Considering a homogeneous horizontal path, taking \( Y_{Hi} \) as the sky luminance, this apparent luminance of the object may be specified as (Overington 1976):

\[
Y_O(r) = Y_{O0} e^{-\sigma_e r} + Y_{Hi}(1 - e^{-\sigma_e r})
\]

To now consider the contrast of an object against its immediate background, the luminances of the object and background may be subtracted from equation 5 yielding:

\[
(Y_O(r) - Y_H(r)) = (Y_{O0} - Y_{H0}) e^{-\sigma_e r}
\]

Considering

\[
C(r) = \frac{|Y_O(r) - Y_H(r)|}{Y_H(r)}
\]

and
VISIBILITY ANALYSIS

$$C_0 = \frac{|Y_{G0} - Y_{H0}|}{Y_{H0}}$$ \hspace{1cm} (7b)

The apparent luminance contrast may be written as:

$$C(r) = C_0(r) \frac{Y_{H0}}{Y_{H}(r)} e^{-\sigma_e \cdot r}$$ \hspace{1cm} (8)

The extinction coefficient for specific atmospheric conditions can be obtained from meteorological data providing the meteorological optical range (MOR), defined as the length of path of atmosphere required to reduce the luminous flux in a collimated beam from an incandescent lamp, at a color temperature of 2700K, to 5% of its original value (See also Visual variables). From that definition, when \( r = \text{MOR} \), equation 4 will provide the extinction coefficient:

$$\frac{F(r)}{F_0} = 0.05 = e^{-\sigma_e \cdot \text{MOR}}$$ \hspace{1cm} (9)

Then

$$\sigma_e = \frac{2.995}{\text{MOR}}$$ \hspace{1cm} (10)

These equations for attenuation assume horizontal views in a homogeneous atmosphere. In case of nonhorizontal line of sight, attenuation decreases with height and an “equivalent extinction coefficient” must be obtained, by integration along the viewing path considering the standard atmosphere layers involved (Overington 1976).

Line of sight: viewsheds and isovists

Given a topographic surface, two points \( p_1 \) and \( p_2 \) are said to be mutually visible if and only if the interior of segment \( p_1 \) \( p_2 \) lies strictly above the surface (De Floriani and Magillo 2003). The segment is defined as the line of sight (LoS) and both points \( p_1 \) and \( p_2 \) must be on or above the surface.

In natural areas, sloped landforms dominate over vertical structures and normally there is a single height (\( z \) coordinate) for each pair of horizontal coordinates \( x, y \). In that case, the definition given for visibility along the LoS would be correct and, for a given location, the viewshed can be defined by finding the LoS for every other point within the area of interest (the target points). If the land surface rises above the LoS, then the target is out of sight; otherwise, it is in sight and is within the viewshed (De Floriani and Magillo 2003).

Given a viewpoint \( V \) on a terrain, the viewshed of \( V \) is the set of points on the surface that are visible from \( V \). The local horizons for the viewpoint are the loci of points on the terrain that are visible from \( V \) and that block the view of points lying immediately beyond them. The global horizon of \( V \) is the most distant point of the terrain that is visible from \( V \) (Figure 3) (De Floriani and Magillo 2003).

Viewsheds are calculated (Figure 4) from a model of the terrain built up from a finite set of data in Euclidean space \((x, y, z)\), which is called a digital elevation model (DEM); the points can be either defined on a regular \((x, y)\) grid or scattered. A DEM based on the elevations \( z \) of those points represents a surface that is an approximation to the real topographic surface.

GIS consider two main types of DEM: triangulated irregular networks (TINs) and regular square grids (RSGs); a review by De Floriani and Magillo (2003) describes both types, as well as various algorithms that can be used to obtain viewsheds from such models. TIN models provide a better representation of the terrain for the same amount of computer memory but RSG models are easier to operate with different layers of information (Figure 5).

In terms of horizontal visibility, the Earth’s curvature has the effect of decreasing terrain elevation with distance; in flat areas, this effect must be taken into account for long views (Figure 6). GIS
Figure 3  Viewshed in a given direction. Terrain points $q_i$ can be observed as long as more distant points are observed at angles higher than any point nearer the observer. That is, a point, for example, $q_2$, at an angle $(\alpha_2)$ lower than the angle of observations $\alpha_p$ of a point $p_1$ nearer to the observer is not visible. Points $p_1$ and $p_2$ are local horizons; point $p_2$ is also a global horizon because it has the highest observing angle for a visible point. The difference in height $[q_i - k_i]$ is the local offset. When the local offset is negative ($q_2$), the point is not visible and vice versa ($q_3$).

Figure 4  Left: viewshed from viewpoint $V$ (indicated by a yellow arrow), calculated using a GIS. Gray areas are not visible. Purple areas are visible. Right: 3-D recreation of the terrain model from an elevated point of view at a distance from $V$. The viewshed from $V$ is overlaid.
VISIBILITY ANALYSIS

Figure 5  Upper plots: GIS representations of a digital elevation model (DEM) using a triangulated irregular network (TIN) (left) and a regular square grid (RSG) (right). Lower plots: 3-D detail of a TIN model (left) and a RSG model (right).

calculations of viewsheds can be adjusted using the terrain elevation correction:

$$\Delta z = \frac{d^2}{D_e}$$ (11)

where $d$ is the planimetric distance between observed and observation points, and $D_e$ is the Earth’s diameter ($2R_e$) (Figure 6, left).

Light refraction partly compensates for the curvature effect. The combination of the two effects (Figure 6, right) is approximately evaluated in GIS calculations of viewsheds by the equation:

$$z_v = z - \frac{d^2}{D_e} + 0.13 \frac{d^2}{D_e} = z - 0.87 \frac{d^2}{D_e}$$ (12)

where $z_v$ is the terrain elevation value taken for the viewshed calculation, corrected for the $z$ value given by the DEM (ESRI 2008).

In urban areas, where landforms tend to be flat, vertical walls and buildings provide the main restrictions to visibility. In such contexts, the concept of the isovist is used. As illustrated in Figure 7, the isovist is defined as the set of all points visible from a given viewpoint in space with respect to an environment. The analysis of isovist and derived magnitudes (isovist fields) are used in architecture for identifying surveillance paths, generating spatial diversity, by changing sensations such as spaciousness, and identifying inconspicuous viewpoints with good visibility (Davis and Benedikt 1979).

Inspired by isovist, the Viewsphere approach created by Yang, Putra, and Li (2007) uses GIS-based 3-D visibility analysis that integrates buildings and terrain to create a 2.5-D surface model; the algorithm obtains a set of indices based on the volumes of space perceived in relation to the total volume of a hemisphere centered at the viewpoint, with an observation radius decided by the observer. Specifically, the authors use two indices: the Viewsphere index ($I$), which evaluates visible volumes from the 2.5-D model and can be taken as an indicator of perceived density; and the Sky view factor ($F$), a ratio between the volumes under and above the horizon line, giving an indicator of the openness of the view.

A similar approach is called the multiscale implicit TIN; it allows multiscale terrain models to be integrated with 3-D features. The improvements in the analysis of intervisibility achieved through this approach, as well as the limitations, are discussed in Kidner et al. (2001).

The image area on the retina: visual exposure

With reference to the surface area covered on the retina, Llobera (2003) defines visual exposure as “a measure of the visible portion of whatever is the focus of the investigation.” Specifically, visual exposure of an object can be described in terms of the surface covered by its image projected onto the observer’s retina (Figure 8).
Figure 6  Left: effect of distance and the Earth’s curvature on terrain elevation in terms of horizontal visibility. The apparent decrease $\Delta z$ in elevation means that certain points in the line of sight are not, in fact, visible from the viewpoint and this translates to a reduction in the viewshed area. Right: the combined effect (red line) of terrestrial curvature and bending of the line of sight caused by light refraction ($\Delta z_r$).

Figure 7  Zenith view of the isovist of point $x$ in a room. White lines are opaque walls.

Considering that the observer’s field of view can be projected onto a sphere of unit radius, the best way to measure the projection of objects on the view of the observer will be to consider segments of a sphere, which are called solid angles. Equivalent to arcs of circumference, measured in degrees or in radians in 2-D space, portions of a sphere in 3-D space are measured in solid angle units, namely steradian (sr). The sphere subtends $4\pi$ sr and an infinite plane $2\pi$ sr.
The visual exposure of objects is influenced by two main variables, distance and position relative to the observer. These variables are the main constituents of perspective. Regarding distance, Figure 9 illustrates the inverse linear relationship between distance and apparent height of objects. For two identical objects of height $h$, the 1-D vertical viewing angles or apparent heights of the object ($\alpha_1$ and $\alpha_2$, in radians) are approximately given by:

$$\alpha_1 \approx \frac{h}{d_1}; \quad \alpha_2 \approx \frac{h}{d_2}; \quad \Rightarrow \frac{\alpha_1}{\alpha_2} \approx \frac{d_2}{d_1} \quad (13)$$

where $d_1$ and $d_2$ are the distances to the observer. The horizontal component or apparent widths $\beta_1$, $\beta_2$ will change by the same factor. Let $w$ be the average width of the object, then:

$$\beta_1 \approx \frac{w}{d_1}; \quad \Omega_1 = \alpha_1 \beta_1 \approx \frac{hw}{d_1}; \quad \Rightarrow \frac{\Omega_1}{\Omega_2} \approx \frac{d_2}{d_1}$$

$$\beta_2 \approx \frac{w}{d_2}; \quad \Omega_2 \approx \frac{hw}{d_2}; \quad (14)$$

where $\Omega_1$ and $\Omega_2$ are the 2-D surface or solid angles (in steradians) projected on the sphere of unitary radius.

Regarding relative position of the observer and the object, in Figure 10 it can be seen how visual exposure decreases when the object is inclined at an angle $\phi$ from perpendicular to the LoS. The approximate change in visual exposure will be:

$$\alpha_1 \approx \frac{h}{d} \, sen \left( \frac{\pi}{2} \right) = \frac{h}{d};$$

$$\alpha_2 \approx \frac{h}{d} \, sen \left( \frac{\pi}{2} - \phi \right) = \frac{h}{d} \, cos \phi \quad (15)$$
The relative position of the object and the observer has a strong influence on visual exposure. As the angle between the object and the LoS decreases (from \( a_1 \) to \( a_5 \)), so does the vertical component of the solid angle, and with it the visual exposure from high (\( b_1 \)) to very low (\( b_5 \)).

where \( h \) is the height or length of the object and \( d \) is the distance, this being a good approximation when the size of the object \( h \) is small compared to the distance \( d \).

A precise procedure for assessing visual exposure using GIS was developed by Domingo-Santos et al. (2011) using RSG models. This procedure can be applied for single or multiple observation points. For the case of a single observation point, the steps followed are: (i) the division of each cell into two triangles, (ii) the filtering of cells out of the viewshed, and (iii) the calculation by vector algebra of the solid angle of each triangle and the total solid angle of each cell. In terms of raster cells, visual exposure would be the measure of how much each cell occupies the field of view of an individual at any location, measured in solid angle units. In the case of multiple observation points, the total visual exposure of a given cell can be obtained by summing up the set of raster layers obtained for each viewpoint; the result will provide a good indicator of general visibility of each cell. The procedure is briefly described in the following paragraphs.

**Cell division**

As various authors have suggested (De Floriani and Magillo 2003), triangles can be used to
VISIBILITY ANALYSIS

provide a good approximation to a real terrain surface. In Figure 11, each square grid cell is divided along a diagonal to obtain two triangles. The height of each triangle vertex is calculated as the mean elevation of the four cells around the vertex. In this way, triangles give a better approximation to the real terrain than the stepped raster cells.

Obtaining viewsheds

For each viewpoint, $V$, the viewshed of the DEM is obtained using the standard algorithm provided by the GIS software.

Calculation of the solid angle for each visible cell

The calculation of solid angles of each triangle seen from viewpoint $V$ can be performed for the triangulated model using the algorithm of Van Oosterom and Strackee (1983):

\[
\tan(\Omega_1/2) = \frac{|a \ b \ c|}{(abc + (a \cdot b) \cdot c + (a \cdot c) \cdot b + (b \cdot c) \cdot a)} \quad (16)
\]

\[
\tan(\Omega_2/2) = \frac{|a \ d \ c|}{(adc + (a \cdot d) \cdot c + (a \cdot c) \cdot d + (d \cdot c) \cdot a)} \quad (17)
\]

where $\Omega_1$ and $\Omega_2$ denote the solid angles of triangles $ABC$ and $ADC$ (Figure 12); $a$, $b$, $c$, and $d$ are the vector positions of $A$, $B$, $C$, and $D$ taking $V$ as the coordinate origin (that is, vectors $VA$, $VB$, $VC$, and $VD$); $|a \ b \ c|$ denotes the determinant of the matrix obtained by writing the vectors side-by-side; $a$, $b$, $c$, and $d$ are the magnitudes of each vector (origin-point distance); and $(a \cdot b)$ denotes the scalar product of vectors $a$ and $b$.

Depending on the position of the triangles $ABC$ and $ADC$, the values of $\Omega_1$ and $\Omega_2$ can be positive or negative and, hence, absolute values must be used to calculate the total solid angle:

\[
\Omega = |\Omega_1| + |\Omega_2| \quad (18)
\]
where \( \Omega \) is the solid angle of the cell in steradians (sr). Given that the absolute value of \( \frac{1}{2} \Omega_1 \) or \( \frac{1}{2} \Omega_2 \) could be greater than \( \pi / 2 \), the inverse function for the tangent function must be the four quadrant arctangent function (atan2).

This procedure implemented on a GIS platform will produce a new raster grid (the same size as the DEM) with the \( \Omega \) values for each raster cell; this is the visual exposure map from point \( V \).

Whenever multiple viewpoints need to be considered (e.g., along roads or tracks, or in urban areas), one visual exposure map will be produced for each viewpoint. The set of raster maps can be summed using a map algebra tool to obtain a cumulative map (Llobera 2003) of visual exposure for the selected viewpoints.

Due to the uneven distribution and very low values of most solid angles, it may be very useful to change the scale for the visual exposure values, to optimize the visualization of the results (Figure 13) (Domingo-Santos et al. 2011).

**Calculation of visibility values**

The final procedure for the evaluation of visibility may depend on the object that it is intended to identify in the landscape. For instance, a clear-cut forest area or a quarry may produce such a sharp color contrast to the background that only MOR and visual exposure may need to be taken into account for evaluating visibility.

On the other hand, when analyzing visibility of elements like overhead power lines and pylons, the background is a matrix onto which parts of the object are projected, and hence more complex approaches may be required like those proposed by Gross (1991) or Zewe and

---

**Figure 13** Visual exposure calculated from the indicated viewpoint. Gray areas are not visible. All colored areas belong to the viewshed but have different levels of visual exposure: purple areas are highly visible; while exposure decreases in yellow and blue areas; and green pixels are the least visible. Other visibility factors presented in this work, like atmospheric extinction or visual acuity, may have great influence on the visibility of areas with low visual exposure.
VISIBILITY ANALYSIS

Koglin (1995). These authors obtain an estimate of the visual impact of an object, at a particular observer location, referred to as the specific visibility $S''(r)$. This value is calculated as an integral over the solid angle describing the surface area of the retina $\Omega$ taken up by the image of parts of the object $d\Omega$, each of them weighted with the color contrast to the background (influenced by distance $r$ that causes atmospheric extinction), $\Delta E(r)$, and this is normalized by the observer's surface area $\Delta A$ or $dA$ (usually $\Delta A = 1 \text{ m}^2$):

$$S''(r) = \frac{1}{dA} \int_{\Omega} \Delta E(r) \cdot d\Omega \quad \text{(unit } S'' : \text{ sr/m}^2)$$

(19)

Using this specific visibility, the same researchers define the total visibility $S$ as an estimate of the entire visual impact caused by an object in the

Figure 14  Changes in visibility due to vegetation screening and changes in the viewshed defined by the GIS. The line of sight (LoS) corresponds to the maximum vertical angle as read by the GIS software from a corrected digital elevation model (also called digital surface model, DSM). Top: The visible area increases owing to the height added to the terrain level by the vegetation canopy. Bottom: Once the DEM is corrected, the GIS viewshed algorithm will identify cells behind $c_2$ as nonvisible; in reality some of these cells ($c_3 - c_6$) might be visible “through the woods” (red LoS). Domingo-Santos et al. 2011. Reproduced from Elsevier.
landscape. This is obtained by integrating over all possible $S''$ within the viewshed area $A$:

$$S = \int_A S'' \cdot dA \quad \text{(unit $S$: sr)} \quad (20)$$

Zewe and Koglin (1995) also define other useful parameters, namely the region-specific visibility and the route-specific visibility.

### Limitations in viewshed and visual exposure evaluation based on GIS

Using GIS data, there may be several sources of error or constraints in the calculation of viewshed and visual exposure: (i) simplification and aggregation in raster elevation data; (ii) factors that may modify the viewshed, such as vegetation screens near the observer (Figure 14), height of the observer, and the presence of unmapped human artifacts, which cannot be modeled in advance by a traditional DEM model (Anile et al. 2003); and (iii) a data structure that produces a 2.5-D model (there being only one possible $z$ value for any particular $(x, y)$ location) and, hence, an inability to deal with vertical structures such as trees, bridges or buildings (Figure 15) with any accuracy (Bishop, Wherrett, and Miller 2000).

Simplifications and aggregations from the raster model may be overcome by using more accurate DEMs or multiresolution models, which help to reduce computational complexity by adapting the level of resolution in terrain representation to the requirements of specific visibility calculations (De Floriani and Magillo 2003; Kidner et al. 2001).

Unmapped obstacles and vegetation constraints may have significant effects on visibility. Given this, to obtain visibility estimates that are as realistic as possible, the vegetation canopy and other obstacles should be considered as modifiers of the terrain, yielding corrected terrain models, known as digital surface models. The high precision elevation data obtained from airborne laser scanner (LiDAR) data provide extremely accurate tree heights, when available.

**Figure 15** Left: vertical view (aerial photography) of a rock bridge. Right: side view of the same image re-created by means of a 3-D viewer based on a DEM. As can be seen on the right, the DEM is not able to provide an adequate view of two objects with the same $x$, $y$ coordinates (the bridge and the terrain under it). The same viewer provides highly sophisticated 3-D images for many urban areas that overcome these limitations. Reproduced from Google Earth © 2015, Google.
The last of the aforementioned constraints of GIS-based analysis can be overcome through the use of 3-D visualization technologies, which have become useful and popular tools for design purposes, evaluating public preferences, and for making decisions regarding forest management or the visual impact of artificially constructed structures. Indeed, 3-D models can provide an observer at a specific location with the visual magnitude of the visible part of any object. Photorealistic applications like Google Street View® also provide an easy and affordable tool for visibility analysis.

However, 3-D and photo viewers can be very time consuming when dealing with many observation points. In such cases, the fall back is GIS analysis, which can provide maps of all locations from which an object can be seen, as well as cumulative viewsheds of “times seen” or total visual exposure of each cell from a large set of locations.

SEE ALSO: Digital elevation model and digital surface model; Representation: 3-D; Spatial resolution; Visual variables; Visualization; Visualizing uncertainty

References


Visual variables

Robert E. Roth
University of Wisconsin–Madison, USA

Overview of the visual variables

The visual variables describe the graphic dimensions across which a map or other visualization can be varied to encode information. The visual variables originally were described by French cartographer and professor Jacques Bertin (1918–2010) in the 1967 book *Semiologie Graphique*. The English translation, *Semiology of Graphics*, was released in 1983 and today is recognized as a seminal theoretical work in both cartography and the broader field of information visualization. The visual variables were inspired by Bertin’s reading of semiotics, or the study of sign systems. Semiotics seeks to understand how one object comes to stand for another object, and therefore helps cartographers think about the way in which a map symbol (the sign vehicle or signifier) representing a geographic phenomenon or process (the referent or signified) comes to mean something (the interpretant) to the map user. French semiotics was developed in linguistics as a method for deconstructing a language into its basic units in order to allow for description and comparison across different language sign systems. Following this approach, Bertin’s visual variables describe the basic building blocks of a map or other visualization. By reducing the map into its constituent graphic elements, unclear or polysemic map symbols can be identified and redesigned, improving cartographic communication. Further, such deconstruction reveals the value judgments and power relationships implicit in the map design process, allowing for critical analysis of the social construction and cultural negotiation of meaning within the cartographic sign system.

Although the visual variables are an influential theoretical framework for cartographic design and research, the contents of visual variable taxonomies vary considerably by scholar (Tyner 2010). Figure 1 synthesizes several notable contributions within cartography into a set of twelve visual variables, adapted from MacEachren et al. (2012). While Figure 1 provides point symbol examples for each of the visual variables, the visual variables can be used to encode information about line and area features as well. Importantly, the twelve visual dimensions described here are only considered visual “variables” when functionally manipulated to encode information. Each of these visual dimensions also can be manipulated as design embellishments to improve the aesthetic quality of the map or visualization.

Bertin (1983) originally identified seven visual variables that can be manipulated to encode information.

1 Location describes the position of the map symbol relative to a coordinate frame. Location is considered an “indispensable” visual variable and takes visual primacy over the others. In cartography, location typically signifies the position of the map symbol relative to a projected spatial coordinate system, meaning that location primarily is used to represent the spatial component of information in cartographic design. However, location can be used to
**Visual Variables**

<table>
<thead>
<tr>
<th>Visual Variable</th>
<th>Associative</th>
<th>Selective</th>
<th>Nominal (non-ordered)</th>
<th>Ordinal (ordered)</th>
<th>Numerical (quantitative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Y</td>
<td>Y</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>Size</td>
<td>N</td>
<td>Y</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>Shape</td>
<td>Y</td>
<td>N</td>
<td>G</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>Orientation</td>
<td>Y</td>
<td>Y</td>
<td>G</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Color hue</td>
<td>Y</td>
<td>Y</td>
<td>G</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Color value</td>
<td>N</td>
<td>Y</td>
<td>P</td>
<td>G</td>
<td>M</td>
</tr>
<tr>
<td>Texture</td>
<td>Y</td>
<td>Y</td>
<td>G</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>Color saturation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arrangement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crispness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transparency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Y = yes; N = no; G = good; M = marginal; P = poor; hatched = n/a

**Figure 1** Visual variables and their syntactics. Figure derived from Bertin (1983), MacEachren (1995), and MacEachren et al. (2012).
represent attribute information, such as the use of perspective height in prism maps. Further, isoline maps manipulate the visual variable location, interpolating attribute values to produce a spatial surface and then weaving isolines through locations of equal attribute values within this surface. In planimetric, 2-D representations, map symbols near the optical center tend to rise to figure, resulting in more immediate visual interpretation by the map user and, thus, landing them in a higher place of importance in the “visual hierarchy.” In contrast, map symbols near the periphery of the page tend to recede to ground, falling to the bottom of the visual hierarchy. In oblique or 3-D representations, map symbols that are perceived as nearer or taller tend to rise to figure.

2 Size describes the amount of space occupied by the map symbol. Size is the primary visual variable manipulated in proportional symbol maps and value-by-area cartograms. Size also is manipulated in flow maps that scale the thickness of the flow lines to an attribute value. Larger map symbols tend to rise to figure.

3 Shape describes the external form (i.e., the outline) of the sign vehicle. The visual variable shape is essential to the design of qualitative point symbols commonly used in reference mapping. The shape of these map symbols can vary from highly abstract, such as circles, squares, or triangles, to highly iconic, directly mimicking the referent represented by the map symbol. Map symbols that are more complex or less compact tend to rise to figure.

4 Orientation describes the direction or rotation of the map symbol from “normal.” The normal orientation typically is relative to the map’s neatline (either explicitly included or inferred by negative space), but in some cases it can be relative to the projected spatial coordinate system (e.g., relative to the graticule) or another baseline. Multivariate glyph symbols typically make use of orientation to differentiate among the represented attributes. Orientation also is manipulated in flow maps to represent the directionality of flow. Proximate clusters of map symbols that have the same orientation (either aligned or misaligned to normal) tend to rise to figure as a single group. Otherwise, individual map features that are misaligned to normal tend to rise to figure.

5 Color hue describes the dominant wavelength of the map symbol on the visible portion of the electromagnetic spectrum (e.g., blue, green, red) and is one of three visual variables associated with the perception of “color.” Color theory occupies a substantial space in the cartographic canon (Brewer 2005) and is particularly relevant for choropleth mapping as well as other forms of reference and thematic maps that use colors to designate categories or classes. A qualitative or spectral color scheme manipulates color hue while controlling the other components of color; as explained below, such color schemes are appropriate for nominal choropleth maps only. Red map symbols tend to rise to figure while blue map symbols tend to recede to ground.

6 Color value describes the relative amount of energy emitted or reflected by the map symbol. Variation in color value results in the perception of shading, or areas of relative light (high emission or reflectance of energy) and dark (low emission or reflectance of energy). Accordingly, color value is sometimes referred to as “lightness”
in color theory. Manipulation of color value is important for choropleth maps depicting ordinal or numerical information. A sequential color scheme adjusts color value in one direction, sometimes crossing over two or three different color hues to improve discriminability. A diverging color scheme adjusts color value in two directions away from a critical midpoint, with each direction denoted by color value changes within a different color hue. A variation on a spectral scheme with the green hues removed also is effective as a diverging color scheme. The figure–ground relationship for color value is relative to amount of light and dark areas on the map. Map symbols that are dark tend to rise to figure on maps that are mostly light (e.g., with a white background), while map symbols that are light tend to rise to figure on maps that are mostly dark (e.g., with a black background).

Texture describes the coarseness of the fill pattern within the map symbol. It was once common to manipulate texture in choropleth maps using halftone or dithering techniques in order to mimic the appearance of shading; as described above, contemporary choropleth maps instead manipulate the visual variables associated with color due to advances in modern printing and digital display devices. Caivano (1990) describes texture as a higher-order visual dimension with three constituent components: the directionality of the texture units (related to the visual variable orientation), the size of the texture units (related to the visual variable size), and the density of the texture units (approaching the perceptual effect of shading associated with the visual variable color value). Regarding the latter component, map symbols with a denser texture tend to rise to figure.

Bertin’s set of visual variables was extended by Morrison (1974) to include two additional variables used in cartographic design. 8 Color saturation describes the spectral peakedness of the map symbol across the visible spectrum, and is the third of three visual variables associated with the perception of color. Bold or saturated colors emit or reflect energy in a concentrated band of the visible spectrum, whereas pastel or desaturated colors emit or reflect energy evenly across the visible spectrum. From a design standpoint, therefore, color value can be conceptualized as the amount of black in a map symbol while color saturation can be conceptualized as the amount of grey in a map symbol. Color saturation also is referred to as “chroma,” “intensity,” and “purity” in color theory. Bold, saturated map symbols tend to rise to figure while pastel, desaturated map symbols tend to recede to ground.

Arrangement describes the layout of graphic marks constituting a map symbol. The visual variable arrangement varies from regular (i.e., graphic marks are perfectly aligned in a grid-like structure) to irregular (i.e., graphic marks are randomly placed or coalesce into clusters). Arrangement differs from the visual variable texture in that all textures are assumed to be arranged regularly, regardless of the initial direction, size, and density of the texture. Dot density maps vary in both arrangement and texture density, a higher-order visual dimension described as “numerousness” by Nelsen (2000). Map symbols with irregular, and particularly clustered, arrangements tend to rise to figure.
Finally, MacEachren (1995) identified three additional visual variables whose manipulation is made easier through digital production methods. MacEachren (1992) originally grouped these three visual variables under a single technique called “focus” in context of uncertainty visualization, but ultimately acknowledged each component as a visual variable given the potential application to other forms of cartographic representation.

10 Crispness describes the sharpness of the boundary of the map symbol. Crispness also is referred to as “depth-of-field” and “fuzziness” in information visualization. MacEachren et al. (2012) found that crispness was the most effective visual variable for representing uncertainty in the context of point symbolization. Map symbols with a crisp boundary tend to rise to figure while map symbols with a fuzzy boundary tend to recede to ground.

11 Resolution describes the spatial precision at which the map symbol is displayed. The visual variable resolution relates to the topic of generalization in cartographic design, which describes the meaningful removal of detail in the map design as the complexity of the real world is abstracted to fit the reduced scale of the map. Resolution as a visual variable leverages different levels of abstraction to encode information, rather than the typical purpose of generalization. In a raster depiction, resolution refers to the coarseness of the grid size. In a vector depiction, resolution refers to the amount of detail (in terms of nodes and edges) in the linework. Map symbols that are depicted in a relatively high level of detail tend to rise to figure.

12 Transparency describes the amount of graphic blending between a map symbol and the background or underlying map symbols. MacEachren (1992) originally referred to transparency as “fog” to suggest a partially opaque barrier impacting the clarity of the underlying map symbols. Transparency is the primary visual variable manipulated in value-by-alpha maps (the “alpha” channel indicating transparency in computer graphics), an alternative to the value-by-area cartogram that visually equalizes enumeration units by adjusting their opacity rather than their size (Roth, Woodruff, and Johnson 2010). Opaque map symbols tend to rise to figure.

Perceptual basis of the visual variables and their syntactics for mapping

An important characteristic of the visual variables is that they are processed pre-attentively, or in an immediate and preconceptual manner at the sensory level of the human eye. Accordingly, the direct translation of Bertin (1983) is “retinal” variable rather than “visual” variable. Thus, the visual variables are “seen” perceptually rather than “understood” cognitively. Principles of perceptual psychology predict how each visual variable is processed by the eye–brain system, and therefore inform use of one visual variable over others for cartographic design. Bertin postulated four such ways that visual perception may inform application of the visual variables in maps and other visualizations, describing these four properties as “levels of organization.”

Bertin’s (1983) first level of organization is associative perception. With an associative visual variable, variations in the visual dimension are perceived with equal weight, allowing for the eye to perceive all map symbols with the same variation as a group (i.e., as associated). For instance, the eye is not drawn to one color hue
**Figure 2** Visual variable levels of organization. Figures (a)–(d) and (e)–(h) depict the same attribute information. Figure based on Bertin (1983).
over the other in Figure 2a or one shape over
another in Figure 2b, allowing for the perception
of all map symbols having the same color hue
or shape as an associated group; as a result, the
eye is likely to see a series of horizontal rows
in Figures 2a and 2b. Because no variation
dominares, an associative visual variables allows
the eye to attend visually to other visual variables
that also might vary in the visual scene (see the
later discussion on visual variable conjunctions
and bivariate mapping). Bertin believed location,
shape, orientation, color hue, and texture to be
associative visual variables. With a dissociative
visual variable, one variation dominates visual
perception, with the eye drawn to this variation
over others. In Figure 2c, the eye is drawn to
the darker color values due to the contrasting
white background, while in Figure 2d, the eye
is drawn to the larger sizes. As a result, the
eye perceives a vertical gradient in Figures 2c
and 2d, rather than a set of horizontal rows as
in Figures 2a and 2b, despite all four figures
encoding the same information. Variation in a
dissociative visual variable inhibits attention to
other visual variables that may vary in the visual
scene. Bertin believed size and color value to be
dissociative visual variables.

Bertin’s (1983) second level of organization
is selective perception. With a selective visual
variable, the eye is able to focus individually
(i.e., attend selectively) upon each variation of
the visual variable across the visual scene and
ignore the other variations. In other words, it is
relatively easy to isolate visually the distribution
of a particular category of map symbol across the
map when symbolized using a selective visual
variable. For example, it is easier to see the
distribution of red symbols in Figure 2e than the
distribution of hexagonal symbols in Figure 2f,
despite the pair of figures encoding the same
information. Bertin believed shape to be the
only visual variable that was not selective. This
property makes shape useful when each map
symbol should be interpreted individually, such
as in the context of large qualitative point symbol
sets. However, shape is not useful when scanning
for broad patterns across all map symbols, an
important goal of most thematic mapping.

Bertin’s (1983) third level of organization
is ordered perception, while his fourth level
of organization is quantitative perception.
Variations in ordered visual variables are per-
ceived as ranked, with the eye pre-attentively
interpreting one variation as “more” or “less”
than another variation. For example, the green
symbols in Figure 2e are not perceived as
“more” than the purple symbols, just different.
In contrast, the darker symbols in Figure 2g are
perceived as “more” than the lighter symbols,
given the white background. Bertin believed
location, size, color value, and texture to be
ordered visual variables; MacEachren (1995)
later argued that color saturation, crispness,
resolution, and transparency also are strongly
ordered visual variables and that texture is only
marginally ordered. Quantitative perception
extends ordered perception, allowing for the
estimation of numerical values from variations
in quantitative visual variables. For example, the
darker symbols in Figure 2g are perceived as
“more” than lighter symbols, but it is difficult
to estimate how much more without use of a
legend. In contrast, it is possible to estimate how
much “more” the larger symbols in Figure 2h are
than the smaller symbols. Bertin believed quan-
titative perception to be restricted to location
and size only.

Importantly, Bertin’s (1983) third and fourth
levels of organization inform MacEachren’s
(1995) visual variable syntactics. In semiotics,
syntactics describes the relationship of sign vehi-
cles to one another. In cartography, syntactics
prescribes the use of a visual variable given the
level of measurement of the attribute information. Unordered visual variables – such as color hue, orientation, and shape – are appropriate for encoding nominal information. Visual variables that are ordered but not quantitative – such as color value, color saturation, crispness, resolution, and transparency – are appropriate for encoding ordinal information. Finally, visual variables that are quantitative – such as location and size – are appropriate for encoding numerical information but also can be applied for ordinal and nominal information given their visual dominance. Figure 1 reproduces MacEachren’s visual variable syntactics, designating each visual variable as good, marginal, or poor for nominal, ordinal, and numerical levels of measurement.

Visual variable conjunctions and bivariate mapping

Maps or other visualizations can make use of a conjunction of two visual variables. Conjunctions can be applied for redundant symbolization, strengthening the graphic encoding of one attribute, or representing multiple attributes in a bivariate display. For bivariate maps, a conjunction can be homogeneous, using the same visual variable in two different ways, or heterogeneous, using two different visual variables to represent the pair of attributes. In cartography, this distinction also relates to the map symbol’s dimensionality (point, line, polygon, and volume), with a homogeneous conjunction manipulating the visual variables at the same symbol dimensionality and a heterogeneous conjunction manipulating the visual variables at different symbol dimensionalities. The images in Figure 3 encode the same pair of attributes, but Figure 3c represents them using a conjunction that is homogeneous, Figures 3a, 3b, and 3d using a conjunction that is heterogeneous by visual variable, and Figure 3a using a conjunction that is heterogeneous by symbol dimensionality.

As with individual visual variables, perceptual psychology informs the syntactics of visual variable conjunctions. The most promising work to date is based on the sensory characteristic of selective attention, or the ability to attend to only one visual variable while ignoring others. Visual variable conjunctions exhibit one of four conditions of selectivity, each of which is appropriate for a different mapping context (Nelson 2000).

First, a separable conjunction describes a bivariate map or visualization in which selective attention of both attribute encodings is uninhibited. With a separable conjunction, the distribution of each attribute (“X” and “Y”) can be “seen” without one restricting the other. Separable conjunctions are produced when using the visual variable shape, given its status as the only nonselective visual variable. Conjunctions that are heterogeneous by symbol dimensionality also tend to be separable, meaning that a thematic combination of choropleth or isoline (i.e., areas) with dot density or proportional symbol (i.e., points) will result in a separable conjunction. In Figure 3a, it is easy to attend to the distribution of either attribute (“X” and “Y”) but relatively difficult to determine their spatial correlation (“+”). A separable conjunction also is created when combining size and color value on the same symbol dimension, as both visual variables are dissociative; thus, shaded cartograms and shaded proportional symbol maps are considered separable. Elmer (2013) recommends use of a separable conjunction when mapping two independent variables that have incongruous scales (e.g., different units of measure, different methods of normalization, different classification breaks), as there is no assumed correlation and the user is forced to attend to each attribute scale individually.

The conceptual opposite of a separable conjunction is an integral conjunction. With integral conjunctions, selective attention is possible on an
Figure 3  Visual variable conjunctions and conditions of selectivity. All maps depict the same attribute information. Figure derived from Elmer (2013).
emergent – or gestalt – visual dimension (“+”) but selective attention of each original attribute encodings (“X” and “Y”) is inhibited. In this situation, it is easy to “see” where the two attributes are the same or different on their respective scales, but it is difficult to attend to each attribute individually. Recommended color schemes for bivariate choropleth maps make use of an integral conjunction, with the arrangement of colors producing an emergent dimension of increased color value as both attributes increase. In Figure 3b, it is easy to determine where both attributes are high or low in tandem by attending to changes in color value (“+”). However, it is difficult to attend to only one attribute at a time; for instance, both New Mexico and Oregon have the same value in the “X” variable, but this is not easily interpreted due to the integral conjunction. Elmer (2013) recommends use of an integral conjunction when the correlation between dependent attributes is more important than the attributes themselves, as the gestalt dimension will focus the user upon this correlation.

Separability and integrality represent ends of a continuum of conjunctive selectivity, with different conjunctions falling somewhere upon this continuum in terms of visual strength. A configural conjunction describes a situation falling in the middle of this continuum, where a gestalt dimension exists (“+”) but the original attribute encodings are not fully inhibited (“X” and “Y”). Here, it is possible to attend to each attribute but there is also a visual cue aiding interpretation of the correlation between the attributes. Homogenous conjunctions that make use of split symbols are likely to be configural, as they exhibit a gestalt dimension of being “in-phase” (i.e., having the same value for each half of the symbol) or “out-of-phase”. In Figure 3c, it is possible to attend to either side of the split graduated symbol (“X” and “Y”) but also easy to interpret where both attributes are high (e.g., Utah) or low (e.g., Arizona) together, as the two halves reform the original circular shape. Elmer (2013) recommends use of a configural conjunction when mapping two independent attributes that have congruous scales, as the condition of being “in-phase” implies the same attribute value on the same attribute scale.

Finally, an asymmetrical conjunction describes a situation in which a nonlogical gestalt dimension emerges in addition to the positive correlation, as is the case with integral and configural conjunctions. Here, the nonlogical emergent dimension (−) inhibits the ability to “see” one of the two attributes (“X” or “Y”), producing an imbalanced visual effect in which the reader tends to interpret one attribute over the other. Asymmetrical conjunctions tend to be produced when using size or color value with other visual variables, given their status as dissociative visual variables. An asymmetrical conjunction also is produced with a value-by-alpha map (Figure 3d), with the attribute encoded using color (“X”) visually equalized by the attribute encoded using transparency (“Y”). Elmer (2013) recommends using an asymmetrical conjunction when one attribute (i.e., the variable of interest) is more important than the other (i.e., the equalizing variable).

The visual variables remain a central framework for empirical research on cartographic design, informing both the experimental trials and controls. A similar approach rooted in semiotics has been taken to understand dynamic representation in animated maps (DiBiase et al. 1992), sonic representation in multimodal maps (Krygier 1994), haptic representation for handheld devices (Griffin 2002), and interactivity (Roth 2012).

SEE ALSO: Cartographic design; Choropleth map; Color theory; Representation and presentation; Visualization; Visualizing uncertainty
References


Visuality

Antje Schlottmann
Goethe University Frankfurt, Germany

The term “visuality” is broadly used to describe an individual sensation, or a particular competence of an organism to perceive its environment. In the context of social theory, visuality encompasses visual practices of appropriating (spatial) reality and established ways of looking at the world – by society, cultural and social groups, or disciplines.

The field of visual geographies can be defined with regard to the role of the visual in the study of spatial issues. In this sense, it is not a disciplinary subfield but rather a research perspective focusing on the visual dimension of geographical issues. The theoretical basis of visual geographies provides opportunities for reflection on the relationship between all kinds of mental and material images, visual practices, and space (Schlottmann and Miggelbrink 2009).

Geography has long been understood as a visual discipline in the sense that it (technically) produces and uses material images and relies on modes of observational practice. In recent years, poststructuralist and constructivist approaches have established visual culture as a major field of study. The ensuing visual turn in geography is a manifestation of the attempt to deal critically and reflectively with geographical visualizations (and the associated idea of “pure vision”) and addresses the significance of visuality in the constitution of spatiotemporal realities (Thornes 2004). Subsequently, there has been a growing awareness that producers and interpreters of geographical visualizations are, at the same time, involved in a production of space, which they usually perceive as unavoidable, “natural,” and “obvious.” Cartographic products, maps in particular, have been condemned for claiming objectivity in “pinning down” the material manifestations of social practices. Critical cartography argues furthermore that visual media facilitate the powerful production and fetishizing of space and that they disguise the causes of spatially manifest(ed) phenomena by merely displaying them. On the other hand, increased attention is paid not only to the visible but also to the invisibilities of the sociospatial. Visuality, seen both as the act of seeing and the materialization of views, is referred to as a contingent, selective, and in its core discursively embedded, practice.

Concepts of the relation between image and reality diverge, thus reflecting the underlying social theories and epistemological premises. Poststructuralist approaches emphasize the independence and arbitrariness of signs. Media theory approaches often cling to a similarity theorem, especially when stressing the manipulative character of images, that is, the way in which they “distort” reality. Constructivist perspectives highlight generative and structuring effects of images. Nonrepresentational approaches tend to refrain from determining the relationship between image and reality in support of a performative idea of visuality as an embodied, material, and often politically charged realm.

The interrelation of visuality and spatiality makes room for an infinite set of research foci. Work on the constitution of space through material images focuses on the role of material culture, on images for everyday use or advertising in
structuring of external spaces. Through this lens, associated spatial concepts are metrical with regard to the regulation of bodies and vessels, or social with regard to the structuring of publicity and privacy, of exclusion and inclusion. In contrast, work on the constitution of space “within” images is concerned with the symbolic representation and hence the established sociocultural interpretations of space (e.g., spatial entities such as town, country, home, landscape, inside, outside, etc.). Applied datasets may consist of maps, landscape paintings, press photos, or montages for advertising purposes for instance. In this context, pragmatic approaches relate to spatial images as reflections of contemporary social needs. They put emphasis on (strategic) visual evocations of specific interpretations of space (“natural open space”; “urban transport space”) under social conditions of globalization. Research is centered on hidden locational logics and propositional truth claims of images (“there it looks like this!”) and on representations of discursively formed and conventionalized spatial meanings (“exotic landscape,” “sublime mountains,” “secure residential area,” etc.). Research in this area even includes deconstructing stereotyped constructions of spatial identities in visual media.

Work on the visual constitution of reality also deals with the alleged documentary and informative character of images and their evidential nature, based on a constitutive analogy of perception and image. Related studies analyze the *documentary politics of truth*, in which the image and its appropriation of space are primarily means to the end of creating evidence and legitimation. Images, such as photographs, satellite images, films, or even surveillance videos, often become guarantors of veracity. An image of a place where something is happening or has happened appears to have a certain evidential quality, based on a general sense of objectivity attributed to the mediating technology. Awareness of the fact that images are in principle susceptible to manipulation is only the obverse of the belief in their character as an accurate reproduction of reality and in the authority of vision as an ontological precondition.

Another research perspective focuses on how the production of, access to, and appropriation of space through images affect and alter social relations. Scholarly work on image icons such as “the blue marble,” the ultimate symbol of Earth as a “subjected object” (Cosgrove 2006), investigates to what extent such *visiotypes* penetrate general communication and discursively shape contemporary world views or the idea of the human. A related strand of research (un)covers the interrelation of behavioral patterns and forms of visual surveillance in a wide range of places, for example in public squares and transport, in nursery schools and house entrances. It deals with subject constructions created by video surveillance and with the extent to which new technologies of visual power serve new forms of governance. Other inquiries focus on the visual availability of spaces and places and the people acting in or with them.

Opposing the visual to the embodied and visuality to materiality has been criticized from a non- or more-than-representational perspective in recent years (Rose and Tolia-Kelly 2012). Likewise, phenomenological theorists such as Pierre Bourdieu, Edmund Husserl, Alfred Schütz, or Maurice Merleau-Ponty always claimed that semiotic theories create the problem of representationality by presuming the sign character of images, that is, their referentiality. In this phenomenological point of view, often adopted by human geographers, images exist once material objects confront the viewer (are “presented”) as something that is (solely) visible. On the other hand, Helmut Plessner’s concept of an uncorrupted first-order experience related to the concept of the lived body (*Leib*), as the
ultimate taken-for-granted experience of oneself as an embodied, material being, is questioned. Corresponding work conceptualizes visuality as a powerful agent in the discursive structuration of sensations of, for example, space. In this sense, visuality bears a hidden “somatic power” which is subject to critical research into visual advertising effect.

Geographical research into the visual constitution of sociospatial realities relies on methods originating from the fields of iconology, art history, or, more recently, visual culture. On the other hand, visual methods have traditionally been used for qualitative (ethnographic) research in human geography and its subfields such as cultural, political, or urban geography (Rose 2011). Videotaping the physical performance of politicians, for instance, may highlight issues of gendered spaces. Self-directed photography provides genuine insights into the individual life-worlds (Lebenswelten) of, for example, the inhabitants of a particular urban quarter and the spatial identities they constitute. More inclusive, participatory involvements of academia and the public, for example at the interface of cultural geography and arts, should provide new pathways toward a reflective handling of the representational and natural character of visuality – including the problematic dualism of observers and observed (Urry and Larsen 2011).

In the field of geographical education, participatory visual geomedia such as open street map promise to overcome the dualism of producers and consumers of space. Such visual tools offer creative modes of geographical education and call for more-than-analytical competences. However, there is growing awareness of the fact that all forms of visuality in educational processes – be it the material images offered in school textbooks, the displays of navigation systems, digital maps, or instructed gazes on excursions – demand critical reflective visual literacy as a basic educational goal.

**SEE ALSO:** Critical geography; Critical GIScience; Cultural studies; Participatory geographies; Phenomenology; Visualization

**References**


**Further reading**

Visualization

Anthony C. Robinson
The Pennsylvania State University, USA

Visualization

Computerized graphics and representation techniques to support information communication and knowledge discovery became known as visualization in the early 1980s, concordant with the development of computing systems that could render complex graphics and support basic forms of interactivity with digital datasets. The fundamental idea behind visualization is to transform data from tabular collections of observations into visual artifacts that can prompt thinking. Where humans may have a hard time interpreting long lists of numbers by visual scanning alone, those same data projected into a graphical representation can help reveal elaborate stories. In that sense, cartography has been a form of visualization for a very long time. Today, we understand visualization to refer to the use of digital methods for representing and interacting with data.

Many of the formative examples in visualization grew from data gathered in scientific experiments and modeling. In 1987, a special issue of the Computer Graphics journal on Visualization in Scientific Computing (McCormick, DeFanti, and Brown 1987) helped crystallize this stream of work into what is now known as scientific visualization. Scientific visualization applications include volume rendering, medical visualization, flow visualization, and chemical visualization. A common attribute of many scientific visualizations is an emphasis on revealing three-dimensional (3-D) structures. As a result, most of the datasets in question represent physical phenomena, and often also include temporal elements to indicate change over time.

Visualization of datasets describing more abstract types of information, including non-numerical data, developed in parallel with scientific visualization, in a topic stream that became known as information visualization. Where scientific visualization focused on representing 3-D phenomena relevant to experiments and modeling, information visualization sought the design of techniques to develop insights from more abstract data sources including financial data, demographic data, and text collections. The lines between scientific and information visualization are not always clear, however, and many relevant problem contexts require the integration of datasets emerging from scientific experiments as well as more abstract information collections. Information visualization has become a significant influence on geographic visualization since the 1990s, and information visualization design patterns for interaction, system components, and representation techniques are commonly found in geographic visualization tools. Ben Shneiderman's work to define a taxonomy for the ways in which different types of tabular data can be analyzed in information visualization (Shneiderman 1996) continues to serve as a strong influence for the general design of geographic visualization tools.

Today, there are thriving academic and professional communities around scientific visualization, information visualization, and a
burgeoning host of subfields that include knowledge visualization and visual analytics. Since a great deal of visualization concerns datasets that include a geographic component, geographers have developed a substantial literature on geographic visualization, frequently referred to as geovisualization.

**From visualization to geovisualization**

The primary connection between visualization and geography has been developed by academic cartographers. References to visualization in the cartographic sciences exist as far back as the early 1950s, but it was the rapid development of scientific visualization in the 1980s that prompted geographers to adopt concepts from that realm and apply them in the mapping sciences. In 1990, David DiBiase was among the first to elaborate on ways in which mapping could support scientific visualization (DiBiase 1990). DiBiase’s so-called swoopy diagram (Figure 1) describing the process of visual thinking and visual communication as a sequence moving from private to public realms is frequently cited as a seminal work in geovisualization. A core aspect of this model is the connection between maps and spatial representations and the process of visually enabled thinking. Maps have the ability to spark knowledge construction and to facilitate

---

**Figure 1** Model of the process of visual thinking and visual communication in cartographic visualization. DiBiase (1990).
problem-solving, not just to communicate what has already been determined through analysis.

Related theoretical developments by MacEachren and Ganter (1990) and Taylor (1994) focused on cognitive models to explain what happens when people use maps to conduct visually led analysis. They emphasized the need for research to understand what prompts people to find patterns using maps, to detect outliers, to understand data quality and uncertainty, and to formulate or confirm hypotheses. In 1995, Alan MacEachren published How Maps Work, which elaborates on these and other aspects of the coupling between cognition, perception, and the use of geographic representations to construct knowledge. A key contribution of this work, first published in an earlier 1994 book chapter, is the Cartography³ theoretical model for geovisualization (Figure 2). This model proposes the realm of possibilities for geovisualization tasks, interaction types, and users. A cross-cutting aspect of this model is its integration of DiBiase’s earlier theory describing the process of visual thinking and communication as facilitated by maps.

Subsequent theoretical developments in geographic information science (GIScience) have drawn further connections between the process of science and the actions one might take with geographic visualization to move between stages in that process. Gahegan (2005) proposed the model shown in Figure 3 which provides for a fully flexible framework in which users could move in sequence through each stage, or move across and between stages in a less prescribed manner. This model offers key leverage points where geographic visualization can be used to support science in its multiple phases of development and refinement.

Geovisualization for exploration and knowledge discovery

The implementation of these theories into practice resulted in an explosion of new tools and techniques in the 1990s and 2000s. Animated maps were particularly popular early on, as emerging computer graphics technologies made it possible to generate animations of spatial data. An important next step was to allow users to control the dynamic behavior of these maps, which saw the development of interactive mapping techniques. Linked displays are another hallmark of many early (as well as contemporary) geovisualization efforts. For example, a scatter plot showing observations may be interactively coupled to a map of where those observations were made, and the interface provides visual linking through user-controlled brushing and highlighting to reveal relationships. Whereas traditional cartographic design ended with the creation of a final product for information dissemination, geovisualization sought the development of interactive maps that could be constantly re-created and manipulated by users as they explore and analyze datasets.

A major emphasis in theoretical models and applications for geovisualization has been on supporting user-driven exploration and knowledge discovery. This corresponds closely to key portions of theoretical models proposed by DiBiase (1990) and MacEachren (1995). The emergence of large spatiotemporal datasets in the 2000s and beyond has spurred the need for new methods by which analysts can make sense out of what they have collected. In many cases, analysts may already have hypotheses in mind that they wish to confirm or refute, but in other instances they may not have any preconceived notion regarding what of interest can be retrieved. Geovisualization provides a

visually led framework for using interactive maps and linked data views to begin unpacking complex datasets and formulate hypotheses about patterns. The underlying theory posits that by approaching problems this way we are able to leverage the enormous power of the human visual system to prompt higher-order thinking and problem-solving. In contrast, manual browsing of raw data tables is an untenable method for prompting this kind of thinking, and only the
most basic outlier detection is possible for most users.

Geovisualization recasts the production of maps from a one-way, intensive design process into an iterative framework that allows users to dynamically create (and re-create) maps and corresponding statistical graphics. As patterns emerge, users can flexibly refine their visualization in order to pursue multiple paths of inquiry. Cartography then becomes a process that happens in real time, and it may be driven heavily by end users who may have no formal cartographic expertise. Many early geovisualization examples grew from a focus on enabling support for animated mapping, a technique that had been proposed in the 1950s and demonstrated as early as 1970. Through animated maps, change over time may be shown, and trends may become evident that are difficult to otherwise discern. As animated mapping matured in the 1980s and 1990s, increased support for user controls to direct animation became more prominent, as did studies to evaluate the impact of these interaction techniques. Animated maps helped push academic cartography toward better solutions for spatiotemporal problems, and this work coincided with the spread of new
datasets that provided temporal as well as spatial dimensions.

As new forms of animated maps were developed to include a wide range of user controls for their parameters, eventually it became possible for a fully dynamic paradigm to emerge. This meant that, instead of deciding when to launch an animated sequence or controlling its speed and other parameters, users could cause the system to display any type of change at any time, in an order of their choosing. Geovisualization at this point adopted what is known today as a coordinated-view interface design, where an input device controls how representations change in a visualization system. The most common manner for coordinated-view systems to operate is for the user to move a mouse cursor around the screen and direct changes to the views by hovering over an observation, clicking on an observation, or selecting multiple observations with a click and drag. To implement this type of interface, visualization tools use brushing, filtering, and other dynamic controls to alter representation methods and analyze data.

A priority for these systems was to support the rapid refinement of spatial information displays, making maps now much less of an ultimate result, and instead more of a transient product that was intended to be changed constantly (Dykes 1997). These coordinated-view systems were the product of a unique synthesis of design patterns from computer science, exploratory data analysis techniques developed by statisticians, and cartographic design and representation methods coming from geographers.

The development of coordinated-view systems also sparked a wave of new research efforts to evaluate their utility and refine their usability. Understanding users became a major priority in this work, and researchers in geography began to incorporate practices from human–computer interaction science in order to evaluate new geovisualization systems and compare them to previous approaches. This work continues today, and it remains a challenge to measure the impact of a fully dynamic coordinated-view environment on problem-solving with real datasets. Traditional measures of efficiency and effectiveness frequently fall short when the goal of an application is to prompt a hypothesis or to help someone interrogate an existing hypothesis.

**Geovisualization toolkits**

The late 1990s and early 2000s saw the growth of several notable geovisualization software systems that implemented both animated mapping methods and coordinated-view interfaces. These systems include CommonGIS, GeoVISTA Studio, the GeoViz Toolkit, and the GeoAnalytics Visualization Toolkit (GAV Toolkit). What these systems had in common is their ability to link together multiple views on spatial data in a fully interactive and dynamic manner. Each of these systems is also data agnostic; they are designed to support visual inquiry into tabular spatial data for any number of possible use-case contexts.

The CommonGIS project began in 1998 at the Fraunhofer Institute in Germany, building on several earlier toolkits with the goal of making a user-friendly geovisualization system that could transform geographically referenced statistical data into actionable information (Andrienko, Andrienko, and Voss 2005). CommonGIS had the ambitious goal of making geospatial data accessible and usable by anyone working from any place. To achieve this goal, CommonGIS was crafted around interactions to support data exploration and the automated creation of maps and statistical graphics. CommonGIS allowed users to explore attributes, dynamically manipulate query parameters, and see the impacts of
these changes in real time through linked maps and statistical representations (Figure 4).

GeoVISTA Studio (Figure 5) began development in the late 1990s in a parallel effort to CommonGIS by researchers at the GeoVISTA Center at The Pennsylvania State University. GeoVISTA Studio, like CommonGIS, was intended to serve as a reconfigurable geovisualization toolkit that could support spatial data exploration and analysis in a coordinated, multiple-view architecture (Takatsuka and Gahegan 2002). A unique attribute of GeoVISTA Studio was its ability to be reconfigured on a component-by-component basis through a graphical user interface that leveraged Java programming technology to automatically link elements and generate new composite applications. This mechanism of geovisualization application construction was made possible in part by the design of common component-level coordination methods that allow visualization elements (a dynamic choropleth map and a scatterplot, for example) to broadcast and receive messages about user-driven highlighting, selection, or other interactions.

Progress on geovisualization technological frameworks, graphics rendering quality, and spatial data handling throughout the late 1990s and early 2000s set the scene for an evolutionary step forward in the late 2000s. The GeoViz Toolkit (Hardisty and Robinson 2011) and the GAV Toolkit (Jern et al. 2007) are two examples of second-generation geovisualization environments that emerged in that era, offering superior performance, tighter integration with the World Wide Web, and improved usability. An important step forward for these toolkits was the complete transition from desktop applications to web-embedded applications. One example of a web-based geovisualization environment is the GAV Toolkit (Jern et al. 2007). The GAV Toolkit features many of the same types of visualization components as its predecessors; however, a key difference is that they are usable inside a web browser. The GAV Toolkit is also designed to be configured in advance of its use, so end users are not deciding which views to display and their general layout. The GAV Toolkit is notable for its deployment in OECD eXplorer, a custom geovisualization application designed to support end-user visual analysis of Organisation for Economic Co-operation and Development (OECD) statistical data (Figure 6).

Figure 4  A screen capture from CommonGIS, showing the result of a mouse selection on a scatterplot of demographic data that coordinates with selection of specific countries on a map. Adapted from Andrienko et al. (2002). Reprinted by permission of Taylor & Francis Ltd, www.tandfonline.com.
Figure 5  A screen capture of the GeoVISTA Studio coordinated-view geovisualization environment. Screen capture provided by Alan MacEachren, GeoVISTA Center, Department of Geography, The Pennsylvania State University. Reproduced by permission of The Pennsylvania State University.
Figure 6  OECD eXplorer screen capture showing correlation among seven socioeconomic indicators in Western Europe. Screen capture provided by Mikael Jern, National Center for Visual Analytics, Linköping University. Reproduced by permission of Linköping University.
Coupling computational and visual methods in geovisualization

As available spatiotemporal datasets grew in size and content diversity during the 1990s and 2000s, so too did the need for geovisualization systems to incorporate computational methods to help users detect patterns, isolate outliers, and model possible results. While visually enabled analysis remained (and remains) a key feature of visualizations of all types, today’s tools frequently feature automated or semiautomated mechanisms for determining what users should look at in the first place. Therefore, the intersection between geovisualization and spatial data mining is an active area of contemporary research.

Early examples of coupled computational–visual systems in the geovisualization community sought to suggest potential patterns of interest to analysts through the use of hierarchical cluster detection and correlation matrices (Guo, Peuquet, and Gahegan 2002). In addition to the advantages associated with automated pattern detection methods for very large and complex datasets, a core goal for these early integrative efforts was to help users understand how spatial data mining methods work, and the impacts of changes on their parameters. Through geovisualization toolkits, users can manipulate the settings associated with data mining algorithms and see changes in their results in real time to evaluate their impact.

Computational techniques in geovisualization frequently focus on dimensional reduction tasks. A common problem today is that thematic datasets often contain hundreds of potentially interesting variables, and combinations of those variables may turn out to be significantly correlated (or not correlated) with one another. Visual inspection alone, even with sophisticated geovisualization tools, is not enough for users to discover and explain everything of interest in such datasets. Therefore, researchers have the goal of reducing dimensions to make these complex interrelationships more visible. A notable example of this scenario emerged with work to synthesize the textual content associated with thousands of Association of American Geographers presentation abstracts. Embedded within these data were large numbers of topical keywords, place references, and author affiliations. Skupin and Fabrikant (2003) used a computational method called a self-organizing map (SOM) to reduce this complex data into clusters that could be projected into what looks very much like a typical relief map, albeit with regions, shading, and labels that correspond to thematic similarity inside the text from presentation abstracts (Figure 7). This method is commonly known as spatialization, which is defined as the translation of nonspatial information into a spatial representation such as a map.

In addition to dimensional reduction techniques, computational methods in geovisualization have also included space–time cluster analysis methods, geographically weighted regression analysis, sequence alignment analysis (borrowed from genetic science) for event data, and named-entity recognition methods for identifying place names in text media. In many instances, geographers are adapting computational methods from other domains to account for the complexities of spatiotemporal analysis. In other cases, such as with geographically weighted regression, methods are emerging directly from geographic science. It remains an essential (albeit difficult) goal for geovisualization tools to make it as easy as possible for end users to flexibly manipulate computational methods and see results in real time to evaluate their utility.
3-D geovisualization and virtual environments

The influence of scientific visualization on geographic visualization extended to a focus early on in the development of 3-D representations and virtual environments (Dykes, MacEachren, and Kraak 2005). Many scientific visualizations feature 3-D representations of physical phenomena, and since the subject of geography is quite often the Earth, the use of 3-D virtual globes, landforms, cities, and immersive environments has naturally been a significant focus of geovisualization. For example, imagery from remote sensing platforms and digital elevation and/or building models can be combined to construct graphical representations that simulate 3-D environments. Thematic information can also be integrated into these environments to support analytical work that requires an understanding of all three physical dimensions.

While 3-D virtual environments in geovisualization first emerged in the 1990s, perhaps the most influential example of such a tool came in the form of Google Earth in 2005. Google Earth was the first public release of a virtual globe originally created by Keyhole, Inc., in 2001. Google acquired Keyhole in 2004, and the launch of Google Earth to the general public in 2005 led to the rapid popularization of virtual environment visualization. Its hallmark is its elegant usability, offering aesthetically appealing virtual views of the planet that appear to seamlessly change as you zoom in or navigate. Today, there are many competing virtual globe platforms, nearly all of which allow users to easily pan, zoom, and query areas on the planet. The basic pattern of mouse-driven navigation used in Google Earth is
now a de facto standard for most contemporary 3-D virtual environments.

The temporal dimension is also a major focus for 3-D geovisualization research and technology development. The space–time cube, first proposed by Hägerstrand (1970), added time as a possible third dimension on top of 2-D maps, resulting in a 3-D cube representation. This method has been adapted and extended by many others, and is implemented today in research and commercial geovisualization software for use cases where understanding the temporal dimension of spatial data is particularly important. Most often the interface for a space–time cube allows users to tilt, rotate, zoom, and filter their data.

Virtual environment research in geovisualization also includes a focus on immersive experiences where users wear special equipment or enter into engineered rooms where the perception of going inside a virtual environment can be created. Early work in these areas was promising but heavily limited by the computational power and equipment engineering required to deliver a fully interactive and immersive user experience. Today, such technical hurdles are much less of an issue, and immersive geovisualization is poised for a resurgence in the form of fully virtual as well as augmented reality systems that overlay digital information on top of what we see in the real world.

A key cross-cutting challenge associated with 3-D geovisualization lies within understanding the extent to which 3-D representations and interfaces can help users in real-world analytical tasks. Many studies have attempted to unpack the differences associated with usability and utility in the context of 3-D geovisualizations, and the strongest signal from this work is that three dimensions works well for showing phenomena that are actually 3-D in reality, such as landforms and structures in urban environments. Users have a harder time doing analysis with 3-D representations of thematic data that are overlaid on top of 3-D landforms or cities, so while such methods have a strong visual impact on users, they do not actually perform as well as they might with simpler 2-D representations or multiple 2-D linked views.

Today, 3-D geovisualization is complemented by a variety of new web technologies that support 3-D graphics in web browsers without the need for special software extensions or high-performance computing. Products such as ESRI CityEngine allow designers to construct urban development scenarios to evaluate the impacts of new construction proposals, and the resulting 3-D models can achieve very high levels of realism while retaining fully dynamic interaction in a simple web browser. This stream of work in geovisualization is also intersecting today with a burgeoning range of technical and scientific efforts to improve the ways in which city infrastructure and individual buildings themselves are managed and maintained. As more and more structures become embedded with sensor technology, new roles emerge for the use of virtual environments to help us understand interrelated systems in our world.

The development of geovisualization in academic and professional practice

Formal evidence of the development of geovisualization has come through a substantial body of published research beginning in the 1990s in journals such as Cartography and Geographic Information Science, Cartographica, International Journal of Geographical Information Science, and The Cartographic Journal. In 1995, the International Cartographic Association (ICA) approved the formation of the Commission on Visualization to focus on research on nascent developments in dynamic mapping tools and techniques. This commission, led by Alan MacEachren at The Pennsylvania State University and Menno-Jan
Kraak at ITC in the Netherlands, helped spur the early development of key research themes in geovisualization. In 1999, the ICA approved the continuation of this commission, which was retitled as the Commission on Visualization and Virtual Environments, to signal the inclusion of virtual reality systems as a core theme. In 2007, the Commission was renamed the Commission on Geovisualization to reflect the widespread adoption of that term to describe this stream of research.

Techniques from geovisualization have also made their way into professional practice through implementation in desktop and web-based geographic information system (GIS) software. For example, ESRI’s ArcGIS for Desktop tools include dynamic filtering and linked maps with statistical graphics, as well as 3-D visualization and virtual globe methods. Sophisticated web mapping platforms such as Google Maps, CartoDB, and GeoCommons have featured animated and interactive map methods that were pioneered by academic cartographers in the early 1990s. Whereas those early examples proved conceptual possibilities, contemporary professional systems have implemented these methods with large datasets and made them usable by novices with no technical overhead aside from the use of a standard web browser.

The rise of geovisual analytics

Starting in the mid-2000s, researchers in the information, scientific, and geographic visualization communities began developing a new stream of visualization science around a field named visual analytics. This new field was defined as the “science of analytical reasoning facilitated by interactive visual interfaces” (Thomas and Cook 2005, 4). The development of this new stream of visualization research was heavily influenced by failures to synthesize and understand incoming evidence in the wake of the 9/11 terrorist attacks in the United States. Several of the chief proponents for the first visual analytics research agenda included federal science and homeland security entities from the United States.

A distinguishing characteristic of visual analytics is the focus on explicit support for analytical reasoning. This signaled a deliberate shift from a focus on supporting visual exploration, as had become quite common by the mid-2000s in information and geographic visualization research and applications. Therefore, the aims of visual analytics research could include empirical investigations on how users make sense of their data through the use of visual interfaces, as well as the development of new visual and computational methods that could directly leverage or augment these sense-making processes. The visual analytics movement brought together cognitive and perceptual study methodologies, the development of new visual interfaces for complex data streams, and real-world problem contexts such as crisis management, crime analysis, financial fraud detection, network security, and epidemiology.

A subset of visual analytics work began concurrently in the field of geography, with many geovisualization researchers shifting to what was called geovisual analytics, an extension of visual analytics focused on problems that require geographical analysis and visualization approaches (Andrienko et al. 2007). An early effort was made to connect geovisual analytics to decision support tasks, where relevant problems in crisis management and other domains were most likely to require careful attention to geographic data and analysis techniques. The broader visual analytics community has eagerly embraced the importance of geography in many relevant task domains and datasets. For example, the annual IEEE (Institute of Electrical and Electronics Engineers) Conference on Visual
Analytics Science and Technology has held a synthetic data analysis challenge contest since 2006, with each year featuring at least one dataset that includes a geographical component. Examples of geovisual analytics applications include tools to evaluate cluster results in cancer epidemiology, to extract and geolocate topics from collections of news articles and social media streams, and to detect patterns in movement in very large datasets of mobile phone usage logs. Recent work has also begun to integrate methods for detecting unusual deviations in streaming data sources, and to develop predictive methods for anticipating future events. These aspects represent portions of analytical reasoning that are among the most important to support for end users, but are also the most difficult to develop in interactive visual interfaces. While many early examples of visual analytics work focused on supporting national and homeland security contexts, today there is an increased focus on supporting citizen science and decision-making, business analytics, and digital arts and humanities. Problems in these domains frequently involve the synthesis of complex and imperfect datasets and require systems that can be leveraged by nontechnical subject matter experts.

Research frontiers in geographic visualization

A substantial collection of published research agendas exists today to guide geovisualization and geovisual analytics investigations (e.g., MacEachren and Kraak 2001; Andrienko et al. 2007). What each of these agendas has in common are key aims to develop new geographical representation methods, new forms of interaction with geographic data, novel approaches for evaluating usability and utility, deeper knowledge on cognitive and perceptual processes that humans leverage when doing visually led analysis in geography, and innovative methods for computational analysis that can help identify patterns or predict future outcomes. Emerging contexts for these research aims include the need to understand patterns in complex collections of spatiotemporal data, such as that gathered en masse by satellites or generated in millions of social media messages each day.

The development of new representation methods to visualize spatial data is a primary focus of contemporary visualization research in geography. Innovation in representation has grown from initial discoveries on animated choropleth and point symbol mapping techniques to include new methods that attempt to show diverse data types (such as text, photographs, and video) in their spatiotemporal context. Much work today is aimed at the development of new representational methods that can scale elegantly to help users make sense of massive numbers of observations. Ongoing challenges include the need for new geovisualization tools that can represent and render useful very large collections of text data from social media sources that have location references, and visual representations that help users understand streaming sources of movement data collected by mobile devices.

A key aspect of leveraging geovisualization to solve real-world problems involves the coupling of data representation methods with computational algorithms that have the power to extract potentially interesting patterns, cluster similar groups of observations, or detect and filter significant outliers. Early visualization and geovisualization work focused on making it possible for users to simply use their eyes to detect and interrogate patterns in their data. This is still a key part of the value proposition for visualization today; however, the science has progressed to include a strong focus on the use of computational methods for automatically
suggesting what deserves visual attention. A wide range of pattern analysis, modeling, and cluster detection methods are in development today, with a particular focus on methods that can detect significant changes in streaming data sources. Whereas datasets of interest in the 1980s, 1990s, and early 2000s were primarily static sources, today we are faced with enormous, dynamic data streams. Another major area of growth is in the creation of pattern-matching techniques whereby end users can specify an important general pattern that could be found in a large dataset, and have the visualization system return the patterns that exist that are most similar to that idealized model.

Interaction with geovisualization tools remains a major focus of ongoing research. A long tradition of academic cartography has centered on the interface between people and maps, and recent technological paradigm shifts toward powerful personal computers and mobile devices has raised new and vital questions that require new science investigations to answer. This research stream had its origins in early geovisualization work to explore ways for controlling animated and dynamic maps. Today, there is a great deal of inquiry around touch-based mobile map interfaces and coordinated-view visualization tools that bring together sometimes dozens of representation methods and blend a variety of data types. We have much to learn yet about the limits of human–computer interaction when it comes to manipulating and making sense of mapped information. Today, users can manipulate map interfaces through keyboards, pointing devices, touch, speech, and their eye movements. While some methods are more common than others, there remain many open questions regarding when and how to use each interaction mode to best fit domain-specific tasks.

A related research stream revolves around the empirical evaluation of geovisualization tools and techniques. Drawing heavily from human–computer interaction and usability engineering domains, research on evaluation in geovisualization has attempted to understand the gaps between research prototype methods and what may work in the real world for end users. This work can be broken down into formative and summative studies, with the former focusing on knowledge elicitation to develop designs for new tools with user inputs, and the latter focusing on comparing the resulting tools to existing approaches to measure their utility and usability. Formative approaches in geovisualization evaluation include case studies, focus groups, interviews, card sorting, think/talk aloud protocols, and paper prototyping. Summative approaches in geovisualization evaluation include task analysis, surveys, heuristic assessment, eye movement analysis, and interaction logging. Many geovisualization evaluation efforts involve a series of iterative steps to conduct both formative and summative assessments, blending together a variety of these approaches to triangulate results.

In addition to interaction science and conducting usability or utility evaluations, cognitive and perceptual research helps to characterize and establish the fundamental processes that direct user engagement with geovisualizations and the information that can be derived from them. Visualization is of particular interest to GIScience investigators with cognitive and perceptual research backgrounds as the analysis contexts in geographic visualization frequently place a heavy burden on end users as they attempt to make sense out of large collections of diverse spatiotemporal data. Visualization to support pattern detection and sense-making is often of interest to cognitive scientists, and the visual affordances of these tools provide opportunities for a wide range of perceptual studies as well. Technologies such as skin response measurement and eye movement analysis allow researchers to
explore what stimulates users when they work with geovisualization tools. Frequently, these types of datasets are then used by cognitive scientists to synthesize patterns of interaction, introspection, and exposition by users in order to understand the spatial analytical process and conceptions of spatial knowledge.

**SEE ALSO:** Cartographic design; Exploratory spatial data analysis; Geographic data mining; Geovisualization of social media; Spatialization; Web-mapping services

**References**


Further reading


Visualizing uncertainty

Kate Beard  
*University of Maine, USA*

The visualization of uncertainty in spatial data was identified as an important topic and an original National Center for Geographic Information and Analysis research initiative in the 1990s. One of the main challenges of this research initiative was to first identify conceptualizations and understanding of uncertainty in its different forms. Uncertainty has many definitions, not all of which fully agree. One understanding of uncertainty is as an expression of confidence in our knowledge about something. In contrast to explicit definitions, uncertainty may be better characterized by identifying its sources. These include lack of data, incomplete knowledge, measurement errors, sampling errors, natural variability in the phenomena being measured, and problems of definition. Fisher (1999) presents a useful categorization of uncertainty centered on the difficulties of defining an object or phenomenon. His categories build on Klir and Yuan (1995) and include three cases. In the first, objects or phenomena are assumed well defined, uncertainty about them stems from errors, and is addressable through probability theory. In the second case, an object or phenomenon cannot be well defined, the uncertainty is attributed to vagueness, and may be addressed by fuzzy set theory. In the third case, uncertainty arises from ambiguity or confusion about how an object or phenomenon should be classified. In such cases, an object or phenomenon may be well defined but disagreement exists over classification or different interpretations of evidence (Klir and Yuan 1995). Such interpretation or classification differences today may be understood as a problem of semantic heterogeneity, a topic being addressed in part by semantic web technologies.

In the mid-to-late 1980s, as interest and technical capabilities in GIS expanded, early work on spatial data quality identified several dimensions of quality which can be equally well be considered dimensions of uncertainty. These dimensions included positional and attribute accuracy, completeness, and consistency. Uncertainty in geospatial data is thus characterized as a multidimensional problem. Uncertainties in the different dimensions (spatial, temporal, thematic) can arise independently, but frequently the dimensions exhibit interdependencies that create extra challenges for characterizing and ultimately visualizing uncertainty.

The balance of concern over dimensions of uncertainty and the interactions among dimensions can vary substantially for particular phenomena of interest, within different domains, and in different context settings. For some spatial phenomena, it is possible to be quite certain about the identity or attributes (identity of a forest fire) yet be uncertain about positional locations or spatial geometry (a forest fire’s location and shape). Change over time introduces temporal uncertainty and limited data sampling in time leaves open substantial temporal uncertainty (e.g., about how a fire progressed). With the introduction and now wide availability of GPS, it is quite easy to obtain precise spatial positions yet be uncertain about attributes at those locations. For example, we could be quite certain about the location and delineation of a town’s boundary, and be quite certain
about the population and other demographic characteristics of the town (a data sampling and collection problem). In other cases the uncertainty in location and identity or attribute are interdependent. For example wetlands are notoriously difficult to delineate thematically and spatially and the task is made more difficult by variations over time. With wetlands the problem is largely one of vagueness although there are also problems in competing definitions and classifications. Spatial autocorrelation, the spatial dependence in a variable over space, is an important dependency that complicates uncertainty assessment. The multidimensionality of uncertainty and the common occurrence of dependencies among uncertainty dimensions in geospatial data has been an on-going challenge in characterizing and visualizing uncertainty.

The different dimensions and sources of geospatial data uncertainty can all be important but, over the last several decades, the relative emphasis on different dimensions of uncertainty has been changing. In the early days of GIS, much attention focused on positional uncertainty. With new technologies, including GPS, more robust and comprehensive spatial reference networks, and higher spatial resolution satellite sensors, positional uncertainty contributed by measurement error has been substantially reduced, making it relatively less critical relative to other sources of uncertainty. Time series data and higher resolution temporal data have become more widely available, heightening concerns about temporal and combined spatial and temporal variability and attendant uncertainties. New data collection technologies have also opened up the potential for average citizens to become data collectors, thereby introducing new sources of uncertainty due to lack of any systematic standards for such data collection. The integration of data from such mixed sources with differing provenance opens new challenges for uncertainty assessment and modeling.

The next sections describe methods for characterizing and modeling uncertainty as a prerequisite for visualizing uncertainty. The entry then moves on to approaches for visualizing uncertainty. The last section reviews some recent evaluations of the effectiveness of these visualizations in conveying uncertainty to users.

Characterizing uncertainty: measures of error

Substantial research has addressed uncertainty representations associated with measurement error and error propagation (Heuvelink and Goodchild 1998; Mowrer and Congalton 2000). When uncertainty is a function of measurement error, several numerical representations in the form of scalars, confidence intervals, and probability distributions are possible. Measures such as root mean square error and circular map error for positional error have been traditional measures used to express uncertainty. These can capture certain aspects of uncertainty but not all. For example, root mean square positional error does not capture shape uncertainty.

Characterization of error-based uncertainty relies primarily on probability models. For a geospatial data object, its dimensional components \((x,y,z)\) coordinates, attributes) are considered random variables whose uncertainty is completely specified by a probability distribution function (pdf). When statistically dependent uncertainties are present, these dependencies must be accounted for; an advantage of probability models is that they can support interdependencies or correlation between dimensions. In such cases, joint probability distribution functions replace separate marginal probabilities distribution functions for each dimension.
Heuvelink, Brown, and van Loon (2007) provide a comprehensive summary of probability models for spatial and temporal dependencies in attribute uncertainties, for correlated attribute uncertainties, and for positional uncertainties that may be statistically dependent in space and time, as well as in their \( x \), \( y \), and \( z \) coordinate dimensions.

Once a probability-based uncertainty model has been specified, it can be used to create multiple simulations of alternative, equally probable realizations of uncertain objects, attributes, or fields.

Spatial uncertainty for example can be modeled through the generation of a set of simulated maps, each of which is consistent with supporting information.

Two recent and useful contributions to the modeling and representation of probabilistic uncertainty include the data uncertainty engine (DUE) (Heuvelink, Brown, and van Loon 2007) and a conceptual model and proposed XML encoding called UncertML. The DUE prototype software tool is applicable to spatial representations that include points, collections of points that behave as rigid objects or deformable objects, and to continuous or qualitative attributes. It supports assessing uncertainties, storing them in a database, and generating realizations for inclusion in Monte Carlo uncertainty propagation studies. The UncertML model is designed to support interoperable uncertainty descriptions for the web environment. Currently the OGC standard web coverage service and web feature service have no way to include uncertainty information with the data. UncertML provides encoding options for summary statistics and probability distributions including joint and marginal distributions as well as mixture models.

Despite several advantages to probability-based models, a number of challenges remain in developing a full suite of models to account for the various cases of interdependencies between spatial, temporal, and thematic dimensions. Also, open questions remain on how to efficiently capture and store the uncertainty information from such models.

**Characterizing uncertainty: representation of vagueness**

Vagueness presents a different form of uncertainty and expectations for different treatment.

Uncertainty related to vagueness has been traditionally modeled using fuzzy set theory. Fuzziness differs from the randomness of probability theory. For example fuzziness measures the degree of occurrence of an event or degree of membership in a class not whether an event occurred or a class exists. Fuzzy set theory supports vagueness or indeterminacy by defining a continuous membership function in the range \([0,1]\) where a value of one signifies definite membership, a value of zero signifies definite nonmembership and intermediate values express gradations of membership. A membership value of 0.2 indicates relatively weak membership while a membership value of 0.8 indicates a stronger degree of belonging to a class. Fuzzy methods have been quite widely used to represent geospatial uncertainty. Early work by Leung (1992) employed fuzzy set theory to model uncertainty in multinomial fields, a common geospatial representation for classified data which includes land cover, soil type, wetlands, and bedrock units. For such data, classes are frequently not well defined or not clearly separated and, thus, well modeled by fuzzy methods.

Fuzzy membership applies most directly to thematic classes but in raster representations, pixel class membership can capture boundary or positional uncertainty as well as classification.
uncertainty. Instead of a single class map (such as a land cover map), fuzzy representations result in the creation of multiple class membership maps, one for each class. Each fuzzy class map conveys one possible version of a categorization. The outcome is thus somewhat analogous with the multiple realizations generated by stochastic simulation from probability based error models.

**Visualization methods for uncertainty**

For geographic information systems which rely heavily on maps and graphics in analyses, visualization provides a logical method for incorporating and conveying uncertainty. Numerous visualization strategies for uncertainty have been proposed and developed (MacEachren et al. 2005). However, it is also the case that no universally accepted standards, best practices, or generally accepted metaphors have yet emerged.

When some measure of uncertainty is available as a variable for display, several graphic symbolizations are possible. Statistical graphics such as boxplots, scatterplots, and histograms can be used to visualize and explore uncertainty but these graphic representations alone do not effectively capture spatially varying patterns in uncertainty. For geospatial data, representation of uncertainty in map form is generally desired in order to convey the spatial location and spatial variation in uncertainty. The depiction of uncertainty borrows substantially from traditional cartographic visualization considerations for selecting an appropriate match between uncertainty measures and visual variables. Of the classic set of visual variables (Bertin 1983) – color (hue, saturation, and value), texture, shape, size, and orientation – the dominant choice for depicting uncertainty has been color and manipulations of its dimensions. Uncertainty representation methods have employed lightness and considered both lighter and darker values for representing uncertainty. Using saturation pure hues are used to represent areas or values with the least uncertainty and desaturated hues to express greater uncertainty. Hengle (2003) manipulated value and saturation jointly to effectively “whiten” areas with greatest uncertainty. These approaches implicitly or explicitly use the analogy of fog in which the data has the appearance of being obscured where it is most uncertain. Other visual variables considered for representing uncertainty include crispness (e.g., contour crispness), focus, and transparency.

Glyphs, compound symbols combining multiple visual variables (e.g., Chernoff faces), have been used frequently to represent multivariate data and have been suggested for representing the multiple dimensions of uncertainty (Pang, Wittenbrink, and Lodha 1997).

All of the above methods face the challenge of how to effectively combine the data representation with the uncertainty representation, which becomes a significant problem for geospatial data. Representation of uncertainty in lower dimensional data is simpler and more common, such as error bars on charts. For geospatial data, the data visualization typically fills the two dimensional space and clever solutions are required to accommodate the uncertainty depiction.

In a GIS context, an uncertainty representation is often viewed as another layer that can be displayed in relation to the data representation. Three alternatives have been identified for relating data and uncertainty in 2-D settings. One option is to display the data and its uncertainty as two separate maps in a side-by-side or adjacent display. In a static context such a display requires a user to make an association between the same locations in the adjacent images. To help users in exploring and assessing the uncertainty, such side-by-side displays need a form of linkage that allows the same object or location in the data
view to be connected to the same object or location in the uncertainty view. For example, selecting a set of pixel locations in the data view can highlight the corresponding pixels in the uncertainty view.

A second option is a sequential presentation in which data and uncertainty views exist as separate maps layered on top of each other and a user can toggle between the two views. The user again has the problem of associating the same positions in the image sequence. The third option is an integrated view in which the data and uncertainty representations are combined in a single representation. In this case data values are mapped to one visual variable and uncertainty values to another. Texture overlays or bivariate color schemes provide methods to accomplish this integration. A hue–saturation–intensity (HSI) color model is a specific example of a bivariate color model in which data values (predictions) map to hues and the uncertainty in the data value (prediction error) maps to saturation of the hue.

Interactive visualizations provide more options to address the display complexity issues. Some interactive methods allow users to select the variable (data or uncertainty) and choose the visual variable to map the variable to depending on what they wish to emphasize. Other interactive approaches allow users to control what data is shown based on setting an uncertainty threshold. As the user interactively adjusts the threshold, data value locations that exceed the uncertainty threshold disappear from the view.

Visual representations of uncertainty become more complex with increasing complexities in the uncertainty representations. The stochastic simulation approach, which offers a conceptually effective means to characterize uncertainty, presents significant challenges for visualization. The Monte Carlo simulations generate large numbers of realizations to capture a full range of variation, leaving the problem of how to synthesize and make visual sense of a large collection of map versions. In the case of Monte Carlo simulation, the uncertainty is not explicitly quantified but implicit in the range of variation expressed across the realizations. The set of simulated maps can be summarized in a static display of the probability of exceeding a particular threshold or some measure of the spread of the posterior distribution. Summary statistics or images, however, lose characterization of the spatial variation. A suggested alternative is to animate the set of realizations but to be effective an animation requires some ordered sequencing of the realizations. These animations should ideally convey an impression of areas that remain stable (have low uncertainty) relative to regions that show large fluctuations.

The visualization of fuzzy set memberships has similar visualization challenges and no widely adopted standard for visualization. The general goal is to visualize multiple class memberships and the confusion between classes. Hengle (2003) describes a color mixture model in which pure classes get a pure hue, locations with mixed but similar classes get transitional hues, and locations with mixed and very different classes transition to white to express the greatest uncertainty about class identity. Pixel mixture methods have also been investigated for fuzzy data representation. In such approaches, pixels are divided into subpixels and each subpixel is assigned a class and the number of subpixels assigned to a class is proportional to the membership value of each class.

Fuzzy classifiers can produce multiple realizations indicating degree of membership. In an example of fuzzy classification of land cover classes, pixels are chosen at random and assigned a cover type based on the membership of the pixel for that cover type and a large number of realizations are produced. Sets of these fuzzy
clairved maps can be viewed one at a time, combined using separate color schemes or animated. Both random and serial realizations have been proposed for such maps. Pixels with dominant membership in one class maintain a consistent color in the animation while pixels with variable membership change color assignments and appear to flicker. The animations are useful for conveying an overall impression of uncertainty. To get more specific uncertainty information for a location, interactive enhancements use a linking strategy to link statistical graphics with image locations.

Assessment of uncertainty visualizations

The goals of visualization methods range from making users aware of the presence of uncertainty to creating quantitative expressions from which an amount of uncertainty can be assessed, to ultimately a deeper spatial understanding. The overall expectation is that visualizations can aid users in better understanding uncertainty, comprehending its possible impacts on their analysis and interpretation and aid in better decision-making.

A number of studies have evaluated the usability of visualizations, user preferences for techniques, and correctness of user interpretation of the intended message (Davis and Keller 1997; Evans 1997; MacEachren et al. 2005). As noted above, visualizations of geospatial uncertainty that attempt to combine data and uncertainty can become very complex and, in many cases, may simply confuse the user, an outcome not always explicitly tested for. Evaluation studies have differed in the number and type of uncertainty visualization methods tested and in the criteria measured (visual appeal, speed of comprehension, correctness of interpretation, ease of use, and overall effectiveness of various visualization techniques). Results have been mixed and because of the wide variation in the test settings no clear patterns on visualization approaches have emerged. Potential dependencies include the user’s level of experience, the task or analysis question, the type(s) of uncertainty, and the user domain. Some studies found that user experience level made a difference in preferences and interpretation while others did not. Some commonalities among the findings suggest that uncertainty as a texture overlay seems to be effective and correctly interpreted by users, and color value (lightness) appears to be more easily interpreted and better associated with uncertainty than saturation. Animations were evaluated in a number of studies and, dependent on the type of animation, some were found to be effective in some contexts, while others were not.

A number of studies sought to test the underlying assumption that inclusion of uncertainty information can contribute to better decision-making by framing the evaluation of uncertainty depictions in decision-making contexts. Leitner and Buttenfield (2000) conducted a study that framed the uncertainty visualization assessment in a site planning decision-making context and their findings indicated that maps depicting uncertainty led to more correct location decisions. However, a comprehensive understanding of the impacts of visualizing uncertainty on decision-making is still an open question.

Unfortunately, several study findings point to limited usefulness of many current uncertainty visualization techniques. An evaluation of multivariate glyphs, for example, found that users were confused and dissatisfied by too much visual clutter imposed by the glyphs.

An analysis by Zuk and Carpendale (2006) of eight uncertainty visualizations suggested a need for greater attention to human factors and cognitive psychology literature on uncertainty and decision-making to better understand the
implications of increased cognitive load in processing complex visual representations.

The empirical studies offer some insights but, unfortunately, many of them are not comparable on task setting, visualization techniques, user experience, and complexity of the data, among other variables, and thus a clear picture of the effectiveness of uncertainty visualization has not been determined.

In studies to date, samples have typically been small and a number of results have been contradictory, leaving a number of open questions. There is, in fact, substantial uncertainty on how well various visualization methods actually communicate uncertainty to the users and the degree to which users are able to assess the uncertainty information correctly.

Conclusions and future directions

Substantial progress has been made in modeling uncertainty and approaches to visualizing it. However, routine use of uncertainty visualization in geospatial analysis contexts remains rare. A general consensus among researchers in the field is that no single visualization method is sufficient to convey all the dimensions and subtleties of geospatial uncertainty.

Assessment studies collectively have not yet provided clear guidance on best practices for communicating uncertainty visually. Because of domain specific aspects, task dependencies, high dimensionality, and difficult to model interdependencies among uncertainty dimensions, uncertainty visualization remains very challenging. Context dependencies are a significant issue and more systematically designed experiments would help to assess the performance of uncertainty visualizations in different contexts.

Experts in visualization and experts in uncertainty modeling have perhaps not interacted as closely as they might in addressing the problem as well as working with cognitive scientist and psychologists in evaluating human factors. A fairly substantial gap still exists between visualization techniques developed largely independently of specific characteristics of uncertainty models and similarly uncertainly models developed largely without visual representations in mind.

The context dependencies may be such that greater attention to the dimensions of uncertainty and their relevance for different data, domains, and tasks warrant very flexible methods for users to indicate and prioritize methods for uncertainty depiction. Many of the approaches are able to indicate a presence of uncertainty but do not yet help with discriminating different types and forms of uncertainty. In other words, uncertainty due to vagueness cannot be discriminated from uncertainty due to errors. Because of the many different dimensions and types of uncertainty this is another challenging problem.

Couclelis (2003) makes the important point that information or knowledge products such as maps do not carry evidence of their own inadequacy. She notes that, “Bad knowledge and erroneous beliefs do not necessarily look or behave differently from good unless and until some special situation arises.” This point argues for continued proactive development of methods to better convey data, information, and knowledge uncertainty through visual means or otherwise.

SEE ALSO: Fuzzy classification and reasoning; Modeling uncertainty in categorical fields; Modeling uncertainty in digital elevation models; Qualitative information: representation; Visualization

References


Volcanic processes and landforms

Saeideh Gharehchahi
Texas State University, USA

Volcanoes are among the most spectacular landscapes on the Earth. They have enormous impacts on human life, ranging from providing fertile soil to producing fatal catastrophes. Volcanic mountains are mainly considered a potential threat for the population living around them. The presence of unconsolidated materials accompanied by the possibility of seismic activities and explosive eruptions make them a potential threat for populations living in their vicinity.

As they are a part of geomorphic systems, geomorphologists have played an important role in studying and classifying the processes and landforms related to volcanic activity. All geomorphic systems can be considered the result of endogenous (tectonic and volcanic) or exogenous (geomorphic) processes. Endogenous variables or tectonic processes are the result of internal energy originating from the Earth’s core, which ultimately influences the processes and structures on the Earth’s surface.

Intrusive and extrusive volcanic processes are associated with the modern theory of plate tectonics. In this model, the hard outer layer of the Earth is a set of broken plates that interact along the margins. These plates include the African, North American, South American, Antarctic, Australian–Indian, Eurasian, and Pacific plates (Huggett 2006) (Figure 1). Some of these plates are covered by oceans and the others carry continents.

There are three types of boundaries through which tectonic plates converge or diverge. Divergent boundaries are those in which plates are being extended by divergent forces as new materials create new oceanic crusts causing the plates to develop. In transform boundaries plates move along main faults or fractures, whereas at the convergent boundaries or active margins two plates collide and subduction occurs. In fact, as a result of this subduction, extreme pressure and heat melt the rock and push the resultant magma toward the surface (Figure 1). The active margins are the location of volcanic mountain belts and high rates of earthquakes; however, some volcanic activities can be found in the middle of the plates. To explain this type of activity, scientists postulate a hypothesis which relates its origin to hot centers known as mantle plumes (Goudie 2004). A volcano above a hotspot is not active forever, because plate motion will carry a volcano past the hotspot and the volcano becomes inactive; however, a new volcano may develop over the hotspot. The most active volcanic regions to which this theory applies are Hawai‘i, Réunion, Yellowstone, and Iceland.

Volcanic and plutonic processes

Volcanic processes refer to the extrusion of molten rocks, known as magma, onto the Earth’s surface or intrusion of magma into crustal material beneath the surface. Volcanoes and the resulting landforms vary depending on the
eruptive behavior. Explosive or pyroclastic volcanoes blow pieces of solid rock from the vents, whereas effusive eruptions contain molten rocks, which extrude less violently onto the surface in the form of lava flows.

The nature of a volcanic eruption is determined by the chemical composition of the magma (molten subsurface rocks), its temperature, and the amount of dissolved gases within it. All of these factors have a large influence on the ability of magma to flow, which is associated with the viscosity of magma.

Silica-rich felsic magmas have a tendency to be cool in temperature, which directly results in high-viscosity magma (resistant to flow), whereas mafic magmas tend to be extremely hot and less viscous. Thus, high temperature causes lower viscosity and high silica content leads to a high level of viscosity. As molten rock rises toward the surface, the pressure and temperature decrease and gases release. This process results in violent eruptive explosions (Petersen, Sack, and Gabler 2010). Granitic (or rhyolitic) magmas consist of 70% silicate. Andesitic magma contains 60% silicate and has a substantial content of incorporated gas with a moderate tendency to flow, whereas basaltic magmas contain 50% silicates which tend to vent gas more readily. These lavas are more effusive. As the basaltic lava solidifies, a network of cracks develops, resulting in columnar features that are roughly hexagonal (Figure 2b).

Silicic, viscous, and cooler magmas expel solid material into the air in various sizes. These rocks, called tephra, are explosive fragments of volcanic
rocks or lava ejected by eruptive processes and classified into three types of pyroclastic material. Ashes are sands or gravels less than 4 mm in diameter, lapilli are between 4 and 64 mm in diameter, and blocks are larger than 64 mm (Huggett 2006).

**Volcanic landforms**

Lava flows are outpourings of magma that result from effusive eruptions and flow with low viscosity and high temperature over a broad area. The mineral composition of lava is a determining
factor in the type of resulting landforms. Basaltic lava is the most dominant form of lava that contributes to the creation of surface features (Figure 2c).

Extremely fluid lava can easily spread across the surface for long distances before cooling. In this case, the outer surface layer comes in contact with the atmosphere and solidifies while the moving body underneath results in a rosy and wrinkling crust called pahoehoe (Figures 3 and 4). With increasing distance from the source, pahoehoe breaks into sharp-edged tubes. This accretion along margins due to less heat and great viscosity is known as aa, which is a common effusive form in the Hawaiian Islands. Lava flows do not always pour out from volcanoes, but can erupt without any explosive activity through linear vents called fissures. Very fluid basaltic lava erupted from fissures is able to scatter over a large area before solidifying, which creates geomorphological landscapes called flood basalts. In some continental regions, these flat surfaces are elevated and result in basalt plateaus. The Deccan Traps of central India, the Siberian Traps, and the Columbia River Plateau of western North America are examples of basaltic plateaus.

Figure 3  Ropy pahoehoe lava. Photo by Garry and Susan Hayes, 2004. Reproduced by permission of Garry and Susan Hayes.
The accumulation of lava flows results in a type of volcano with a low slope and convex profile like a shield, which is accordingly named a shield volcano (Figure 5a). Unlike the more explosive volcanoes that are more dangerous, shield volcanoes are built by relatively quiet effusive eruptions. The most prominent shield volcanoes in the world are located on the Hawaiian Islands, with the largest being Mauna Loa. It reaches 4170 m above sea level and is 17 km from the Pacific floor to its summit.

The smallest type of volcano is the result of an explosive eruption of pyroclastics, rather than an accumulation of lava from repetitive eruptions. These are known as cinder cones (Figure 5b). They have a bowl-shaped crater made of pyroclastic materials. At the time of eruption, because of high levels of gas, cinder fragments blast into the air and accumulate around the volcanic vents. Distinctive cinder cones can be found in Mexico, Idaho, and New Mexico. The most famous one, named Paricutin, in central Mexico, grew out from a new vent in 1943 to a height of 100–400 m in a year.

The most common type of volcano, named stratovolcanoes or composite cones, are conical volcanoes made of accumulated layers of pyroclastic materials (tephra, ash) and lava (Figure 5c). They are distinctive, with a simple symmetrical cone and a concave profile as seen on Mount Damavand (Figure 2a), the highest volcano in Iran and Asia (5610 m), and Mount Fuji (3776 m) in Japan. Stratovolcanoes are common...
features where an oceanic plate subsides below a continental plate or another oceanic plate.

Silica-rich magma forms acid lava which is extremely viscous. Where this type of lava moves toward the volcanic cone due to high silica content and outgassing of fluid magma, the viscosity will increase and magma will not be able to travel far from the vent (Figure 5d). As a result, a steep-sided and jagged dome-shaped volcano emerges with a collar of unstable debris around it. Due to the growing pressure of gas over time, this type of volcano, called a plug dome, often erupts explosively.

In 2008, Mount Chaitén, a plug dome in southern Chile, erupted for the first time in more than 9000 years. The eruption began with the emission of ash, seismic activities, and pyroclastic flows while two vents developed in the old dome. The government evacuated the city when defensive works failed. From July 2008, Chaitén continued to erupt, with a growing lava dome and ash columns up to 3000 m. In 2009, the dome collapsed, and pyroclastic materials flowed 5 km from the central vent. The risk of mudflows and flooding is still high in this area, so the government has abandoned the city and relocated the residents.

The most common feature of all volcanoes are the chambers through which magma rises to the surface. The main chamber is the central vent. During the eruption, the volcano chamber may evacuate enough to trigger the collapse of the volcanic structure, leaving behind a distinct depression crater called a caldera. Yellowstone Caldera, the best known example, erupted most recently approximately 650,000 years ago during which released materials covered a substantial part of North America.
Intrusive processes

Intrusions or plutons include those molten rock materials that solidify or cool before reaching the surface of the ground. These igneous processes are collectively referred to as plutonism.

Large intrusions are sited deep and appear on the surface by the uplift of the land they intrude or by the process of erosion. They become a distinctive landscape because of their resistance to erosion compared to the other, weaker surrounding rocks. These resistant rocks often form the core of mountain chains.

Intrusions are classified by their shape and their location in relation to the rock layers they intruded (Figure 6). Large intrusions are circular in shape. Small intrusions without any particular shape are called stocks and cover an area less than 100 km². Large ones that cover an area greater than 100 km² and mostly have granite in their composition are known as batholiths. They are associated with metamorphosing or melting adjacent rocks while they develop under the land surface. The best examples of batholiths can be found in California; for example, the Sierra Nevada Batholith is a continuous granitic form. Another larger batholith is located in the Coast Mountains of western Canada, and stretches for about 1800 km to southeastern Alaska.

The other type of intrusion is related to the injection of magma into fractures and between two layers of sedimentary rock without changing the surrounding rocks. A laccolith is associated with the horizontal flow of magma between two layers of rock in a bump shape. These sedimentary layers are often eroded and leave a resistant feature of mountains or hills. They form like a pipe in which the cap is connected to the source of magma by a stem. The Henry Mountains of Utah are one of the typical laccolith sites in North America.
There are also other smaller intrusive features that can be exposed by erosion. One of those small forms is termed a sill, created by intrusion of magma in a flat shape between rock layers without pushing overlaying layers upward. Sills are known as harder strata in geology. They may develop to be hundreds of meters thick, such as some unique sills in Tasmania, but usually they are between 10 and 30 m thick.

The pressure on magma can cause a vertical injection into the fractures within a rock body. These wall-shaped structures, named dikes, are always younger than their surrounding rocks. However, tectonic processes can change the form of dikes from vertical to horizontal sheets. Volcanic necks are a kind of dike which results from magma solidifying in the central vent with the weak layers of cone eroded over time and the neck exposed as a resistant remnant standing above the Earth’s surface. In Europe, volcanic necks are used as a unique place for constructing castles or chapels, such as Saint Michel d’Aiguilhe chapel in France or Trosky Castle in the Czech Republic.

Volcanic hazards

Volcanoes are a collection of processes and hazards in mountain regions, but they also include further hazards involved with eruptive activity. Direct hazards are the result of explosive activity whereas indirect hazards are the by-product of volcanic activity. Volcanoes may remain dormant for many years, but at the time of activity the consequences of hazards are severe and enormous. They contribute to the rapid change of the landscape and cause disruption by interrupting drainage systems and slope processes. Because of the combination of unconsolidated materials, such as pyroclastic fragments and ashes, and steep slopes that are highly fractured and faulted, slope hazards are common (Thouret 2010).

Pyroclastic flows are one of the direct hazards that are caused by explosive eruptions. They are among the most dangerous volcanic hazards because they are a moving body with high velocity and contain a large amount of pyroclastic materials and gas which can reach a temperature of 1000°C. In 1902, the eruption of Mount Pelée completely flattened the town of St Pierre in Martinique and killed some 30,000–40,000 people. Despite enormous eruptive activity prior to the event, the government did not take defensive measures. At that time the wrong perception of risk caused the death of the entire population.

Rockfalls, landslides, and debris avalanches accompany or follow eruptive activity. They are the collapse of a section of slope in which a huge mass of deposits moves rapidly downslope and destroys everything in its path. One of the largest debris avalanches recorded in history is associated with the eruption of Mount St Helens, in the United States, in 1980. An initial earthquake caused the north side of the mountain to bulge. A second earthquake, of magnitude 5.1, led to a massive collapse.

Lava flows move slowly, but some of them can be sufficiently fast to cause destruction and fatalities. In 2002, Goma, in the eastern part of Congo, was destroyed by lava which spread through the center of the city. Lava destroyed 4500 houses and buildings and killed some people.

Tsunamis are the other deadly events resulting from seismic activity prior to or simultaneous with the incidence of volcanic eruptions, which transfer a high level of energy to sea waves. Because many shorelines are densely populated, they can severely affect coastal populations. The 2011 earthquake off the coast of Tohoku in Japan triggered powerful seismic waves which reached...
heights of up to 40 m. In this catastrophic event, 15,893 people lost their lives.

Lahars are a mixture of hot pyroclastic materials and rock debris, which start flowing down a steep slope and are able to melt snow and glaciers on their way. This will result in a muddy and fast-flowing stream. The rock boulders carried by lahars make them devastating while the muddy flows spread over the populated plains and far away from the source. The eruption of the Nevado del Ruiz volcano in 1985, in Colombia, South America, resulted in a tragic event. Pyroclastic flows melted the glaciers on the way and formed a lahar moving down the slope at a speed of 60 km per hour. It buried the town of Armero, killing more than 20,000 people.

SEE ALSO: Geomorphic hazards; Mass movement processes and landforms; Mountain geomorphology; Tectonic geomorphology

References


Further reading

Voluntarism and the voluntary sector

Mark Skinner
Trent University, Canada

Andrew Power
University of Southampton, UK

The voluntary sector is a complex and increasingly complicated sphere of twenty-first-century society. It comprises a myriad of diverse and heterogeneous self-governing associations of people who have a mutual interest in providing services, support, education, advocacy, self-help, and, in some cases, activism and resistance. The sector is also referred to, often synonymously, as the third sector (in between the state and market), nonprofit sector (versus the for-profit sector), community sector (alongside government and business), nongovernmental sector (alongside government and business internationally), or the social economy (in between the private and public spheres of activity). Each term is derived from particular historical–geographic contexts, such as the prevalence of the nonprofit sector in North America, social economy in continental Europe, and nongovernmental (in relation to international) development in the Global South.

Some, but not all, within the voluntary sector are charities. Many have statutory sources of funding. Some rely primarily on voluntary fundraising and/or grant-in-aid funding, while others have detailed contracts/service agreements with state agencies. Given the heterogeneous nature of voluntary organizations, they can be differentiated by their size and organizational structure, and range from small community groups to international development organizations.

Volunteers, those individuals who contribute their time freely, are essential to the voluntary sector. Most voluntary organizations involve volunteers formally in some fashion, in front-line and/or governance roles. Although not established for financial gain, many voluntary organizations also have paid employees and can be highly professional. There is, however, growing recognition of the scope and importance of individuals who volunteer their time informally, without association to voluntary organizations. Taken together, the actions of the voluntary sector and volunteers, both formal and informal, are rooted in the broader philosophy of voluntarism – that is, noncoerced action, usually for public benefit, outside the constraints of the state. In principle, this means that voluntary bodies do not depend on statute or the market for their existence or purpose.

Yet, as has become evident over the past quarter of a century, voluntarism and the attendant activities of the voluntary sector and volunteers are increasingly intertwined with the state and private interests as welfare societies have evolved. Given the importance and complexity of the sector, the study of voluntarism involves understanding the spatially uneven and place-specific nature of the voluntary sector and volunteering. Interest in geographical dimensions is long-standing and was renewed recently as part of a broader voluntary turn across the social sciences over the past two decades. At the heart of what is
now referred to as “geographies of voluntarism” has been an effort to tackle the multidisciplinary challenge of understanding voluntarism in an era of globalization, welfare state restructuring, and demographic ageing. Such work is set against the backdrop of a widespread rediscovery of voluntarism within political debates on the role of the voluntary sector. Largely confined to the Global North, debates on where the voluntary sector is positioned vis-à-vis the state have long been a focus in many countries, particularly within the context of welfare and care systems worldwide. In the early twenty-first century, national and subnational governments are increasingly turning to voluntary organizations for the delivery of many services, including local transport, libraries, information and advocacy, and health and social care. Debates over how the voluntary sector operates have become increasingly relevant. At the same time, the voluntary sector (and volunteers) in many countries has always played a role parallel with the state. Nonetheless, for much of recent history, the sector has operated outside the state’s remit, as family and friends’ associations, and community (parish)-based and mutual charities. The increasing ways that the sector has become involved – and invoked, even co-opted – by the state has led to a growing recognition of the crucial importance of understanding how voluntary sector activities are manifest geographically. The new scales, places, and spaces that the sector increasingly occupies give rise to wholly new spatial dimensions, as well as prompting new theoretical and methodological debates about the sector’s future role in society. Such debates, however, remain predominantly focused on the Anglo-American-Australian/New Zealand context of voluntarism in the Global North.

Evolution of voluntarism and the voluntary sector

The voluntary sector has evolved from the primary to a supplementary and even subservient provider of welfare services and support alongside the state. Historically, the voluntary sector has had a long-standing role within the provision of support for those in need in society, including many who were poor, destitute, sick, or old. Partly in resistance to state involvement in the lives of individuals and families, voluntary services traditionally have included almshouses, private philanthropy-based institutions, and religious-based charities, as well as extending to the administration of hospitals and early mental health asylums. This primary faith-based role in alleviating destitution continues today, especially in countries of the Global South.

The widespread repositioning of the state as the primary provider of welfare, with the voluntary sector playing a more supplementary role, began in the early nineteenth century with the establishment of publicly funded asylums across the United Kingdom, Europe, Scandinavia, and North America, as well as the Soviet Union. It was further entrenched in the welfare states that were established across Anglo-America and Western Europe following World War II. Even in strongly Catholic countries such as Spain, Portugal, and Ireland, state provision of services gradually extended into many areas of public life. In essence, the policy of viewing voluntarism as the “extension ladder” was promoted, with the idea that the voluntary sector would provide specialist services or add-ons to those provided publicly; however, this supplementary role within the welfare state would change with the growing mistrust in what many saw as paternalistic public services in the postwar era.
In the last quarter of the twentieth century, a further repositioning of the voluntary sector was sparked by fiscal crises and a broader neoliberal-inspired rolling back of the welfare state via restructuring of health-care and welfare systems. Attendant devolution, divestment, and downloading within welfare states coincided with deinstitutionalization and other community-oriented policies to create more demand and responsibility for voluntary sector support. For instance, people with mental health and other issues, who had previously been cared for by the state in public institutions, now had to rely on the “patchwork quilt” of volunteer-based community services that became the prominent form of service delivery throughout the 1980s and 1990s. The patchwork model of voluntary support was emulated throughout health and social care, leading to policy concerns about service coordination and accessibility, as well as uncertainty about the sustainability of volunteer-based support in general.

Interest in the geographical implications of the evolving position of voluntarism vis-à-vis the state began to emerge as geographers, initially in North America, then in the United Kingdom and elsewhere, traced how volunteer services in the neoliberal era were located in ghettoized, inner-city metropolitan areas, which were often stigmatized places, given the punitive view of the welfare class that used their services. Out of this situational work arose a major contribution to the political economic critique of public service restructuring – Wolch’s (1990) pivotal conceptualization of the voluntary sector as a shadow state, with attendant questions about its role in, and capacity for, delivering health-care and welfare services. The term “shadow state” refers to the repositioning of the voluntary sector from providing a complementary or supplementary role to being seen as an alternative and, in some cases, a subservient (or co-opted) provider of health and community care services. Underpinning this work was a wider belief that public welfare retrenchment was eroding the centralist and universalist foundations of support services. In turn, many scholars argued it would lead to a more fragmented, decentralized, and ad hoc provision of support, reliant on a piecemeal voluntary sector. Wolch’s conceptualization provides a critical, and still relevant, appraisal of how far voluntary sector providers may come to look more like public sector providers. This trend is particularly apparent in the increasing professionalization of much of the voluntary sector, and its corporate positioning within or alongside the state that is evident across the range of voluntary organizations today.

The shadow state phenomenon has arguably become more pervasive since the turn of the twenty-first century, as public service restructuring and the emphasis on a greater role for the voluntary sector has emerged as a defining feature of welfare state reform across Europe, North America, and elsewhere. It is possible to identify the macro-level shifts in how the sector has been recognized and co-opted through different political governance models, including neoliberal contexts. For instance, the United Kingdom’s “third way” agenda in the 1990s and early 2000s sought to reconcile right-wing and left-wing politics by advocating a synthesis of right-wing economic and left-wing social policies. Under this political project, which has been emulated elsewhere, the voluntary sector was to encompass all nonprofit organizations, including social enterprises, cooperatives, and charities of all shapes and sizes. This development coincided with a large rise in the profile of the sector and an upward shift in the scope and scale of voluntary action. Alongside this change in policy and discourse came structural changes to support state and voluntary sector partnerships, which marked a new “settlement” (or compact) in the
voluntary–state relationship, one that promises a further repositioning of the voluntary sector as partner in the provision of services.

In response to the renewed focus on the voluntary sector, a growing number of geographical studies in Canada, New Zealand, the United Kingdom, and the United States have focused on the implications for voluntarism of the rationalization, privatization, and devolution of health-care and welfare services. These effects have been borne out in terms of how and where organizations operate and with whom they work. This particularly productive body of work, exemplified in the groundbreaking book by Milligan (2001), has explored the implications of restructuring processes for local communities and individuals, and traced the changing roles and responsibilities of voluntary organizations, volunteers, and families. The impacts of this evolving context have been examined through a range of voluntary sector services, including home care, community support, counseling, literacy, sport and health promotion, HIV/AIDS care, hospices, hospitals, long-term care, mental-healthcare, urban homeless shelters, and rural elderly care.

As noted, most of the current literature is focused on the Global North context of voluntarism, even though the positioning of the voluntary sector is known to have evolved quite differently in many countries across the Global South as well as in post-Soviet countries. In Eastern Europe, for instance, the voluntary sector remains a primary provider, as the state has continued its residual “last resort” role in the form of institutions and comparatively punitive welfare services. In contemporary communist countries, such as China, the role of voluntarism and the voluntary sector traditionally has been restricted to state-sanctioned activities, mostly in relation to culture and the environment. In the broader Global South context, voluntarism has evolved to comprise primarily nongovernmental organizations (NGOs) that, in some countries, fund and/or provide the majority of health-care services. The development of a voluntary sector comprised of NGOs in the Global South, however, has seen phenomenal growth and proliferation. Their contribution has been linked primarily with development projects but they also have opened up a political space at the local level as an alternative to state-led structures and mechanisms. At the same time, NGOs are often integrated into national and global systems in ways that voluntary organizations in the Global North often are not. In addition, there is the phenomenal growth of “volunteer tourism,” as explored through the work of Sin (2010), driven by social initiatives like sustainable development, ethical consumerism, and “first world responsibilities to the third world.”

Scales, places, and spaces of voluntarism

The previous discussion offers an account of the ways in which voluntarism and the voluntary sector have evolved around the world and the different positions they have occupied in relation to the state, and the ways in which such positioning has been conceived and critiqued. Turning to geographical dimensions, the voluntary sector has been “mapped” empirically in comparative contexts, for instance, in the interdisciplinary Johns Hopkins University “Global Civil Society in Comparative Perspective” project. As noted by Milligan and Conradson (2006) in their seminal anthology of the field, this interdisciplinary work confirms the spatially uneven landscapes of voluntarism and volunteering that have emerged between and within the Global North and Global South at the turn of the twenty-first century. This comparative work also reveals important gaps in knowledge around volunteering internationally, most notably
in regards to gender and development studies, population ageing and community development, and environmental governance.

Alongside these international comparisons is a body of work that has sought to trace the scales, places, and spaces of the voluntary sector and volunteering within national borders, and the resulting “geographies of voluntarism” that have materialized at the national, local, and individual levels. Geographers have been attuned to the process of voluntary organizations “jumping scale” from local to national and, indeed, from national to international scales (e.g., European Union, United Nations). At each scale, there are varying benefits and risks involved in terms of funding, level of political involvement, and ability to make change. This work usefully illustrates the varying spatial practices and strategies of voluntary organizations, and how such organizations try to best adapt these to suit the needs of their clients. Some organizations are very sensitive to scale: for instance, some disability groups feel their activities become diluted and/or they become too distant to their client groups by jumping scale from local and more service-driven activities to national and more political activities.

At the national scale, various attempts at mapping the voluntary sector have been undertaken within new, nationally focused research initiatives, such as the Birmingham (UK) Third Sector Research Centre. This work reveals discernible patterns of voluntary sector activity on the landscape, shaped by different political contexts. For example, Mohan and Bulloch (2012) point to the dispersed and fragmented nature of the “civic core” of voluntary activity across the United Kingdom, which they found to consist of one-third of the population, which provides the vast majority of volunteering hours and participation in civic associations, as well as the amount given to charity. Members of the civic core, they discovered, are drawn predominantly from the most prosperous, middle-aged, and highly educated sections of the population, and are most likely to live in the least deprived parts of the country. This is a crucial point, as governments are increasingly turning to the sector to fill the gaps in welfare support for the most needy. The concept of civic core is used to reveal the gap between the government rhetoric associated with policy and the reality on the ground, where many “charity deserts” occur. In the UK context, for instance, the current “big society” government agenda has sought to redefine the relationship between voluntarism and society by calling on individuals to play a larger role in volunteering. This political project deploys a narrative associated with the middle-class volunteer of the Victorian “golden age” of voluntarism, and of all parts of society contributing regardless of class, religion or gender. It also emphasizes and seeks to champion the ubiquity of “everyday” volunteering, as seen in support of libraries, school activities, and sports, for instance, which account for a significant proportion of volunteer hours within national surveys.

In other national contexts, such as Canada and the United States, this rhetoric has become a typical narrative alongside (and perhaps associated with) the wider welfare retrenchment and restructuring of the voluntary sector discussed in the previous section, and is having profound material outcomes for voluntarism and the shape of the voluntary sector on the ground. At the subnational scale, there have been growing expectations and pressures to fill gaps in urban welfare services and to occupy an increasingly public stage in terms of having to provide wider community services. There has also been a corresponding push towards individualism, which has been propelled by the rise of self-directed support. Self-directed support refers to growing expectations that welfare service recipients will
become more autonomous and responsible for their own care and wellbeing. Such expectations are arguably characteristic of neoliberal constructions of economic life and have implications for expectations of volunteers. As DeVerteuil (2012) argues, however, the supportive approaches by volunteers and voluntary organizations to groups deemed vulnerable (e.g., people with mental health problems, the homeless, the elderly) have been downplayed by many mainstream accounts that have become largely fixated on the punitive accounts of injustice in the city – particularly within contexts where the residual neoliberal welfare state reigns. In this sense, injustice must coexist with and depend upon more supportive, charitable currents within urban space. Voluntary sector support at a very fundamental level can be seen to facilitate the broader processes of neoliberal capitalism to continue.

At the local scale, the role of voluntarism has emerged as a means of explaining disparities in local manifestations of health and care, as well as spaces of resistance to structural changes in health-care systems. Geographic work has also examined how voluntary organizations and volunteering have become embedded in particular places. There is an underlying concern for understanding the importance of “place” in explaining how and why the voluntary sector is manifest quite differently among and within various types of communities, locales, and environments. Geographers have increasingly advocated for a nuanced understanding of the role of voluntarism in the formation of places, for instance, in the World Health Organization (WHO)’s global age-friendly cities initiative, which identifies volunteering as a key determinant of successful community development in an era of population ageing. Attention to the compositional and transformative nature of place (as a physical and social phenomenon) has led to a greater understanding of the local dynamics of voluntarism. This development is seen not only in metropolitan and urban settings but also in various rural and more remote areas, where volunteer-based support is seen to be a crucial community and individual asset. Indeed, the place-embeddedness of the voluntary sector and volunteering is fundamental for understanding the uneven landscapes of voluntarism that have emerged in recent decades. Key studies have also examined the physical, representational, and emotional aspects of place in relation to voluntary care activities from the household, to the community, to the school or clinic, and beyond. As seen in work by Skinner and Joseph (2011), the underlying emphasis is on the social construction of place and the transformative potential of voluntarism in place formation, such as the role of older volunteers in leading initiatives to create age-friendly rural communities, thus advancing understanding of the place-embeddedness of voluntarism. It also reveals the ways in which people respond to the power dynamics associated with health and care in their particular locales, such as in resistance to health-care restructuring (Milligan 2001).

At the micro (individual) scale, the embodied dimensions of volunteering at the individual level have also garnered research, from accounts of front-line volunteers in urban homeless shelters, to the experiences of volunteer health promoters in rural parts of South America, and to biographical accounts of the life stories of volunteer activists. As is evident in the contributions to a special issue of *Health & Place* edited by Skinner and Power (2011), the underlying emphasis is on understanding how voluntarism shapes, and is shaped by, the experiences of people and the places in which they volunteer. For instance, Milligan, Kearns, and Kyle (2011) reveal how those involved in and committed to voluntarism invoke memories and develop strong personal narratives about place-based
struggles, and reflect on the importance of place in shaping their activist roots.

To further understand the (uneven) landscape of the voluntary sector across scales and within places, an appreciation of the different spaces the voluntary sector occupies is required, as well as of the place-specific ways in which voluntarism occurs. There are myriad spaces and places which the voluntary sector occupies, uses, manages, and resists. Understanding voluntarism as a multifaceted process that operates across urban, economic, political, and social spaces has become increasingly useful for the development of informed policy at various levels, from the national state to the regional government and even more locally. In terms of the spaces of welfare within the city (e.g., voluntary homeless shelters, drop-in centers), for instance, geographers have argued that these sites have a lot more to offer marginalized groups than basic survival or a “last resort” service. They are also a “space of refuge” away from stigma and where normal rules do not apply and it is claimed that everyone is welcome regardless of their background. These spaces also serve other purposes: to utilize facilities (relevant for those with more complex and multiple challenges); to socialize; to kill time and to gain information. It must be acknowledged, however, that these settings can also be problematic, where clients can often experience infantilization, stigmatization, patronization, and exclusion due to the complex codes of behavior and subtle forms of policing that are often interwoven within these sites. Other work has focused on understanding how the voluntary sector can occupy a “space of resistance” to the growing pressure to take on new and greater roles in services provisioning on the ground as well as on wider social justice issues. Building on the seminal work of Wolch (1990), Milligan (2001), and others, there have been increasingly sophisticated theorizations of, and empirical inquiries into, voluntarism as a response to structural and socioeconomic changes, such as restructuring health-care systems, urban homeless shelters, and ageing rural communities, as already noted.

Expanding on these lines of inquiry, new forms of voluntarism are being observed within a range of social movements that are responding in different ways to state practices around the world. For instance, scholars examining volunteer activism in the wake of the 2007–2008 global financial crisis – and the attendant calls for austerity measures among Western governments, the IMF, and the World Bank – have highlighted the varying and contested spaces of participation and resistance that have evolved in response to the further rolling back of the state in terms of social welfare provisioning in different societies, and in different socioeconomic climates. Movements such as Occupy and Anonymous often reject the notion that twenty-first-century state restructuring policies are simply governmental responses to the hardship associated with fiscal crisis (as in the rhetoric of “keep calm and carry on”). They argue rather that policies such as austerity are neoliberal political projects in their own right that exacerbate this hardship. Much of this work has been informed by (post-) Marxist geography, led most prominently by the seminal work of David Harvey. He has examined the resistance movements and revolutionary urban struggles that have mobilized in reaction to the effects of global capital accumulation, including volunteer-based activist groups. Work in this vein, including studies of the Arab Spring revolutions in the Middle East and North Africa, has examined how resistance movements have collectively taken over public space and, through their actions, made it more inclusive and accessible. In this way, what makes a space public is often not its preordained “publicness” but is the result of various strategies and acts of resistance that often involves volunteers.
and volunteering. While broader in scope than voluntarism, this work nonetheless illustrates how civil society (and the voluntary sector) can mobilize in response to utopian visions of society, as well as to issues concerning the organization of everyday economic and social life.

Geographers have also become more attuned to the democratic spaces opened up by volunteers, activists, and organizations seeking to represent the identity of, and make claims for rights and recognition for, constituent groups (e.g., lesbian, gay, bisexual, and transgender (LGBT) people, ethnocultural minorities, older persons). This relational geography reveals a complex set of voluntary spaces where small charitable bodies and highly professionalized nonprofit organizations coalesce. A core concern of this work is how such spaces of voluntarism seek to reshape the citizenship status of individuals, and to examine the motivations and strategies behind the coalescence of different interest groups around, for example, defense, mutual support, culture or language preservation, and resistance. These varying motives often relate to the broader civil society context that voluntary groups occupy. For example, during the “troubles” in Northern Ireland, local people sought to withdraw from the hostility of wider society and create mutual societies based on strong religious grounds. Meanwhile, gay people have often mobilized for mutual support and resistance (for example, groups such as Stonewall), in an attempt to cultivate more safe spaces and to maintain the notion of queer lifestyles. While work has focused on the importance of voluntarism in these fields, others have criticized these movements for their essentializing tactics of creating group identities and their political positioning in relation to the state. As an indirect outcome of their identity politics, to fit in within these groups often involves performance of certain roles. Often they create duties and obligations for members, who may feel under constant surveillance to ensure that they behave in an appropriate way (e.g., Muslim women).

Conclusion

The growing interest in and research into volunteering across the social sciences in the early twenty-first century is evident in the multidisciplinary and widespread policy concern about the implications of the uneven landscapes of voluntarism. This concern has been propelled, in no small part, by the enduring gaps in knowledge about the geographical dimensions of voluntary sector activities and volunteering. In response to this gap, much work has sought to understand voluntarism in an era of globalization, welfare state restructuring, and demographic ageing. As part of this work, the evolution of geographical conceptions of and empirical investigations into voluntarism (from the shadow state to spaces of resistance) has highlighted the crucial differences that “scale,” “place,” and “space” make to understanding the processes, outcomes, and implications of voluntary sector activities. This recognition of the place-specific ways in which voluntarism occurs is evident within particular contexts (from metropolitan to remote) and among various scales (from household to global). Underlying the contributions of geography and geographers, however, are the implicit connections to concepts of scale, place, and space that link geographies of voluntarism and provide the foundation for the emerging and, arguably, increasingly explicit spatial turn in voluntary sector studies more generally.

Despite, or perhaps by virtue of, this work, the lack of consensus surrounding the defining scope and character of the voluntary sector remains a challenge. Given its broad range, there are myriad ways in which its constituent organizations, groups, and volunteers (both
formal and informal), can shape and be shaped by the wider changes occurring in the early twenty-first century. This recent period offers both challenges and opportunities for the sector, in terms of how it responds to socioeconomic, demographic, and environmental change. Looking forward, the idea of the postwar welfare state’s supplementary positioning of the voluntary sector is rapidly becoming a distant memory. Yet, given the now ubiquitous nature of welfare reform and the ongoing political project to roll back the state, debate over the role of the voluntary sector and volunteering will continue to be at the forefront of public policy worldwide. Geographical knowledge, especially in regards to other aspects of voluntarism, such as gender and development studies, population ageing and community development, ethnocultural and religious identities, and the environment, will be an essential part of understanding the proliferation of new forms of voluntarism in response to these challenges.

SEE ALSO: Civil society; Governance and development; Social capital

References


Further reading


Volunteered geographic information: quality assurance

Mordechai (Muki) Haklay
University College London, UK

Volunteered geographic information (VGI) originates outside the realm of professional data collection by scientists, surveyors, and geographers. Quality assurance of such information is important for people who want to use it, as they need to identify if it is fit for purpose. Goodchild and Li (2012) identified three approaches for VGI quality assurance: the “crowdsourcing” approach, which relies on the number of people that edited the information, the “social” approach, based on gatekeepers and moderators, and the “geographic” approach, which uses broader geographic knowledge to verify that the information fits into the existing understanding of the natural world. In addition to the approaches that Goodchild and Li identified, there are also the “domain” approach, which relates to the understanding of the knowledge domain of the information, the “instrumental observation” approach, which relies on technology, and the “process-oriented” approach, which brings VGI closer to industrialized procedures. First we need to understand the nature of VGI and the source of concern regarding quality assurance.

While the term “volunteered geographic information” is relatively new (Goodchild 2007), the activities that this term describes are not. Another relatively recent term, “citizen science” (Bonney 1996), which describes the participation of volunteers in collecting, analyzing, and sharing scientific information, provides the historical context. The collection of accurate information by nonprofessional participants has been an integral part of scientific activity since the seventeenth century and likely before (Bonney, Shirk, and Phillips 2013). Therefore, when approaching the question of quality assurance of VGI, it is critical to see it within the wider context of scientific data collection and not to fall into the trap of novelty. Thus, approaches to ensure that such information is of suitable quality, and the handling of information about quality aspects, can be based on well-established methods that were developed in the past.

Yet, this integration needs to take into account the insights that have emerged within geographic information science (GIScience) research over the past decades. Within GIScience, it is the body of research on spatial data quality that provides the framing for VGI quality assurance. Van Oort’s (2006) comprehensive synthesis of various quality standards identifies the following elements, which are commonly considered when spatial data quality is discussed.

- **Lineage** – description of the history of the dataset and how it was collected and evolved.
- **Positional accuracy** – how well the coordinate value of an object in the database relates to the reality on the ground.
- **Attribute accuracy** – as objects in a geographical database are represented not only by their geometrical shape but also by additional attributes, this measure evaluates how correct these values are.
- **Logical consistency** – the internal consistency of the dataset, in terms of topological
correctness and the relationships that are encoded in the database.

- **Completeness** – a measure of the lack of data; that is, an assessment of how many objects are expected to be found in the database but are missing, as well as an assessment of excess data that should not be included. In other words, how comprehensive the coverage of real-world objects is.

- **Semantic accuracy** – linking the way an object is captured and represented in the database to its meaning and the way in which it should be interpreted.

- **Usage, purpose, and constraints** – this is a fitness-for-purpose declaration that should help potential users in deciding how the data should be used.

- **Temporal quality** – this is a measure of the validity of changes in the database in relation to real-world changes and also the rate of updates.

Naturally, the definitions above are shorthand and aim to explain the principles of geographical information quality. The burgeoning literature on spatial data quality provides more detailed definitions and discussion of these aspects.

While some of these quality elements might seem independent of a specific application, in reality they can only be evaluated within the context of use. For example, when carrying out analysis of street lighting in a part of town, the question of completeness becomes specific regarding the recording of all street-light objects within the bounds of the area of interest; whether the dataset includes these features for another part of the settlement is irrelevant for the task at hand. This is true for all other quality elements. The scrutiny of information quality within a specific application to ensure that it is good enough for the need is termed “fitness for purpose.” Fitness for purpose is a central issue with respect to VGI.

To understand the reason that geographers are concerned with quality assurance of VGI, we need to recall the historical development of geographic information and, especially, the historical context of geographic information systems (GIS) and GIScience development since the 1960s. For most of the twentieth century, geographic information production was professionalized and institutionalized. The creation, organization, and distribution of geographic information was completed by official bodies such as national mapping agencies (e.g., France’s Institut géographique national) or national geological bodies (such as the British Geological Society) which were funded by the state to carry out mapping activities. As a result, the production of geographic information became an industrial scientific process in which the aim was to produce a standardized product – commonly a map. Due to financial, skill, and process limitations, products were engineered carefully so they could be used for multiple purposes. Thus, a topographic map could be used for navigation and also for urban planning and many other purposes. Because the products were standardized in terms of scale or the type of geographical features that were recorded, detailed specifications could be drawn against which the quality elements could be tested, and quality assurance procedures could be developed to ensure that the surveyor collected the necessary information on the ground and the cartographer tested her map before sending it to the printing press. This was the backdrop to the development of GIS and to the conceptualization of spatial data quality.

The practices of centralized, scientific, and industrialized geographic information production lend themselves to quality assurance procedures that are deployed through organizational or professional structures, and explain the perceived challenges with VGI. For example, procedures for appropriate calibration
VOLUNTEERED GEOGRAPHIC INFORMATION: QUALITY ASSURANCE

of photogrammetric equipment can be drawn up and then applied every time an operator starts a work session. Centralized practices also support employing people with the focus on quality assurance, such as going to the field with a map and testing that it complies with the specifications that were used to create it. In contrast, most of the collection of VGI takes place outside organizational frameworks. The people who contribute the data are not employees and seemingly cannot be put into training programs, asked to follow quality assurance procedures, or expected to use standardized equipment that can be calibrated from time to time. The lack of coordination and top-down forms of production raises questions about ensuring the quality of the information that emerges from VGI.

To consider quality assurance within VGI requires understanding of some underlying principles that are common to VGI practices and so differs from organized and industrialized geographic information creation. For example, some VGI is collected under conditions of scarcity or abundance in terms of data sources, number of observations, or the amount of data that is being used. As noted, the conceptualization of geographic data collection before the emergence of VGI was one of scarcity, where data is expensive and complex to collect. In contrast, in many applications of VGI the situation is one of abundance. For example, in applications that are based on microvolunteering, where the participant invests very little time in a fairly simple task, it is possible to give the same mapping task to several participants and statistically compare their independent outcomes as a way of ensuring the quality of the data. Another area where the abundance concept can be useful is in the development of software for data collection. While in previous eras there would be inherently one application that was used for data capture and editing, in VGI there is a need to consider multiple applications, as different designs and workflows can appeal and be suitable for different groups of participants.

Another underlying principle of VGI is that, since the people who collect the information are not remunerated or in contractual relationships with the organization that coordinates data collection, a more complex relationship between the two sides is required, with consideration of incentives, motivations to contribute, and the tools that will be used for data collection. Overall, VGI systems need to be understood as sociotechnical systems in which the social aspect is as important as the technical part.

In addition, VGI is inherently heterogeneous. This is due to the principle above – because the process is not controlled in a top-down manner it cannot follow attempts to capture all the information in an orderly way. In large-scale data collection activities such as the census of population there is a clear attempt to capture all the information about the population over a relatively short time and in every part of the country. In contrast, because of its distributed nature, VGI will vary across space and time, with some areas and times receiving more attention than others. An interesting example has been shown in temporal scales, where some citizen science activities exhibit “weekend bias” as these are the days when volunteers are free to collect more information. For more details, see Volunteer geographic information.

Because of the difference in the organizational settings of VGI, a different approach to quality assurance is required, although, as noted, in general such approaches have been used in many citizen science projects. Over the years, several approaches have emerged and these include “crowdsourcing,” “social,” “geographic,” “domain,” “instrumental observation,” and “process-oriented” approaches.
The “crowdsourcing” approach builds on the principle of abundance. Since there are a large number of contributors, quality assurance can emerge from repeated verification by multiple participants. Thus, if several participants deliver Global Positioning System (GPS) tracks for the same road while driving their car and all the tracks contain the evidence of fast driving, we can be confident that the road is there and suitable for private vehicles. Even in projects where the participants actively collect data in an uncoordinated way, such as the OpenStreetMap project, it has been shown that, with enough participants actively collecting data in a given area, the quality of the data can be as good as authoritative sources. The limitation of this approach is when local knowledge or verification on the ground (“ground truth”) is required. In such situations, the crowdsourcing approach works well in central, highly populated or popular sites where there are many visitors and therefore the probability that several of them will be involved in data collection rises. Even so, it is possible to encourage participants to record less popular places through a range of suitable incentives.

The “social” approach also builds on the principle of abundance in terms of the number of participants, but with a more detailed understanding of their knowledge, skills, and experience. In this approach, some participants are asked to monitor and verify the information collected by less experienced participants. The social method is well established in citizen science programs such as bird watching, where some participants who are more experienced in identifying bird species help to verify observations made by other participants. To deploy the social approach, there is a need for a structured organization in which some members are recognized as more experienced and are given the appropriate tools to check and approve information.

The “geographic” approach uses known geographical knowledge to evaluate the validity of the information that is received by volunteers. For example, by using existing knowledge about the distribution of streams from a river, it is possible to assess if mapping contributed by volunteers of a new river is comprehensive or not. A variation of this approach is the use of recorded information, even if it is out of date, to verify the information — for example, by comparing how much of the information that is already known about a location also appears in a VGI source. Geographic knowledge can be potentially encoded in software algorithms that evaluate new information in light of existing knowledge.

The “domain” approach is an extension of the geographic one and, in addition to geographical knowledge, uses a specific knowledge that is relevant to the domain in which information is collected. For example, in many citizen science projects that involve collecting biological observations there will be some body of information about species distribution both spatially and temporally. Therefore, a new observation can be tested against this knowledge, again algorithmically, to help ensure that it is accurate.

The “instrumental observation” approach removes some of the subjective aspects of data collection by a human, who might make an error, and relies instead on the availability of the equipment that the person is using. Because of the increase in availability of sufficiently accurate equipment, such as the various sensors that are integrated in smartphones, many people keep, in their pockets, mobile computers with the ability to collect location, direction, imagery, and sound information. For example, image files that are captured in smartphones include,
in the file, the GPS coordinates and time-stamp which, for a vast majority of people, are beyond their ability to manipulate. Thus, the automatic instrumental recording of information provides evidence of the quality and accuracy of the information.

Finally, there is a “process-oriented” approach, which brings VGI closer to traditional industrial processes. Under this approach, the participants go through some training before collecting information, and the process of data collection or analysis is highly structured to ensure that the resulting information is of suitable quality. This can include provision of standardized equipment, online training, or instruction sheets, and a structured data recording process. For example, volunteers who participate in the US Community Collaborative Rain, Hail & Snow Network (CoCoRaHS) receive a standardized rain gauge, instructions on how to install it, and online resources to learn about data collection and reporting.

Importantly, these approaches are not used in isolation and any given project is likely to see a combination of them in operation. Thus, an element of training and guidance for users can appear in a downloadable application that is distributed widely; therefore the method that will be used in such a project will be a combination of the process-oriented and the crowdsourcing approaches. Another example is the OpenStreetMap project, which, in general, gives limited guidance to volunteers in terms of information that they collect or the location in which they collect it. Yet, a subset of the information collected in the OpenStreetMap database about wheelchair access is completed through the highly structured process of the WheelMap application in which the participant is required to select one of four possible settings that indicate accessibility. Another subset of the information, recorded for humanitarian efforts, follows the social model in which the tasks are divided between volunteers using the Humanitarian OpenStreetMap Team (HOT) task manager, and the data that is collected is verified by more experienced participants.

The final, and critical, point for quality assurance of VGI is fitness for purpose. In some VGI activities the information has a direct and clear application, in which case it is possible to define specifications for the quality assurance elements that were listed above. However, one of the core aspects noted above is the heterogeneity of the information that is collected by volunteers. Therefore, before using VGI for a specific application, there is a need to check for its fitness for this specific use. While this is true for all geographic information – and even so-called authoritative data sources can suffer from hidden biases (e.g., lack of update of information in rural areas) – the situation with VGI is that variability can change dramatically over short distances – so while the center of a city will be mapped by many people, a deprived suburb near the center will not be mapped and updated. There are also limitations that are caused by the instruments in use – for example, the GPS positional accuracy of the smartphones. Such aspects should also be taken into account, ensuring that the quality assurance is also fit for purpose.

SEE ALSO: Data quality standards; Metadata; Uncertainty; Volunteered geographic information

References

Further reading


Volunteered geographic information

Daniel Sui  
*Ohio State University, USA*

Jonathan Cinnamon  
*University of Exeter, UK*

Origin and meaning

Volunteered geographic information (hereafter VGI) is defined as “geographic information acquired and made available to others through the voluntary activity of individuals or groups, with the intent of providing information about the geographic world” (Elwood, Goodchild, and Sui 2012, 575). The growth of VGI in the past 10 years is a result of several related technological advances and scientific practices, such as Web 2.0, Geoweb, spatial media, neogeography, citizen science, crowdsourcing, and open science (Rice *et al.* 2012). Goodchild (2007) coined the term VGI to describe what he viewed as a different form of spatial data production that contrasted with more conventionally produced geographic information, and to emphasize the potential role of VGI in augmenting our knowledge of the geographic world. VGI often differs from conventionally produced forms of geographic information, including the types of information produced and the approaches used to acquire it, the methods and techniques for working with it, and the social processes that mediate its creation.

Despite the increased attention to VGI in recent years, voluntary spatial data collection efforts are not new and can be traced back before the age of Web 2.0. A host of related movements have helped enable VGI as a phenomenon. One of the earlier examples was the Christmas Bird Count starting back in 1900. In geography, Bill Bunge’s “Geographical Expeditions” involved local residents in counter-mapping and spatial data production efforts in Detroit in the 1970s (Bunge 1971). In the United Kingdom, the BBC’s Domesday Project in the 1980s used volunteers and community groups to assemble a digital spatial data archive (Openshaw, Rhind, and Goddard 1986). The explosive growth of VGI in the past decade is a result of several converging technological advancements. The traditional high cost of spatial data collection—which necessitated its production either by government agencies or corporations—has now been vastly reduced due to broad access to the portable global positioning system (GPS) devices, location-aware smartphones and other handheld/wearable devices, broadband Internet, and the availability of user-friendly mapping software.

In addition to declining costs and the emergence of accessible mapping technologies, another catalyst of VGI in the past decade has been the desire and need to produce user-generated content and maps about phenomena that are not covered by conventional data production activities. These include transient and ephemeral phenomena in time-sensitive situations, information that is less-readily quantifiable, or phenomena of interest to individuals and groups, but which have perhaps less significance to national mapping agencies and other traditional producers of spatial data.
Furthermore, during the past five years, interests in VGI have been merging with research agendas of the broader phenomenon of big data. As part of the big data deluge, VGI is also of interest to the private sector, which has taken an increasingly significant role in the production, archiving, dissemination, and use of geographic information in general. The private sector is also beginning to cash in on citizen data production efforts. This new trend raises a series of social, economic, and political issues regarding the production and use of VGI in society.

Types of VGI

There have been several efforts to develop VGI typologies according to the purpose of VGI (allocentric vs egocentric) or the nature/content of VGI (implicit vs explicit) (Craglia, Ostermann, and Spiniatti 2012; Engler, Scassa, and Taylor 2014). However, in the age of big data, the boundary between VGI and other types of geographic information/spatial data is rapidly blurring. In fact, as a special type of geographic information, VGI shares multiple commonalities with the traditional geographic information we have been dealing with so far. In general, VGI can be loosely grouped into three types: geospatial framework data, gazetteer/place name data, and miscellaneous geotagged content data (text, audio, photo, video, etc.) about various places/points of interest. Furthermore, VGI can be used to augment existing authoritative geospatial databases of various kinds.

A growing number of VGI initiatives are developing geospatial framework data. In general, framework data are the most common data themes geographic data users need and typically include seven themes: geodetic control, orthoimagery, elevation, transportation, hydrography, governmental units, and the cadaster. The framework data cover relatively static phenomena and are commonly used for municipal services and administration, geopositioning, and wayfinding. Until recently, the production of geographic framework data has been the exclusive responsibility of government agencies. Not all seven types of framework data are particularly well-suited to the VGI or crowdsourcing approach. For example, maintenance of the primary geodetic control system and the compilation of accurate elevation data require a high level of expertise in surveying, mapping, and advanced photogrammetry. Major investments in equipment are also a prerequisite for performing these tasks successfully. Furthermore, delineations of land tenure and ownership of governmental units such as voting districts or school districts are structured by laws and government policies. As such, citizen volunteers can only play limited roles (if any) in producing most types of geographic framework data although road network data is a notable exception. The primary responsibility for data compilation for most framework data will likely remain with professionals such as licensed surveyors with advanced training in mapping and photogrammetry.

The production of transportation and road network data using the VGI approach has been successful. OpenStreetMap (OSM), perhaps the best example of geospatial framework data collected by volunteers, is a collaborative global effort to produce a detailed, full coverage digital map of the world’s road networks (and some other features) through voluntary effort. Volunteers capture the locations of transportation infrastructure and topographic features by using GPS and tracing satellite imagery. Once compiled, these data are freely available as rendered online maps or for download as data files under the framework of creative commons. Users are free to use these data and to develop their own
In addition to the abovementioned VGI projects that have been designed to parallel conventional data collection efforts, there exists great potential to harness VGI as a data source for improving existing products. The impact of VGI on this process could be profound with potentially billions of sensors around the globe contributing to improving existing maps and databases. Instead of intermittent coverage by professionals when time and funds allow, VGI could offer more timely observations by distributed volunteers acting as an early warning system for local changes. There are a few examples that provide a template for how this approach could be more widely applied, by both the private and public sectors. The United States Geological Survey (USGS) has a program called the National Map Corps (http://nationalmap.gov/TheNationalMapCorps), which encourages members of the public to update the National Map. Corporations are widely using this model as well, including Google who are enlisting the help of the public through their MapMaker program and “report a problem” feature. These hybrid spatial data production activities could be particularly valuable since they could benefit from the strengths of both approaches, the low-cost and distributed nature of VGI and crowdsourcing initiatives and the oversight and quality control of conventional data production methods (Cinnamon 2014).

In addition to framework and gazetteer related data, VGI has also been used to produce attribute data – information about local phenomena. These nonframework VGI initiatives assemble data about some phenomenon for which the spatial distribution or patterns are significant and compile the observations of many contributors. This may be one of the most exciting aspects of VGI, the ability to produce data on any theme or phenomena of interest, including information
that has not previously been recorded (Cinnamon and Schuurman 2013). An emerging source of VGI may exist due to the emerging popularity of Apple iHealth, Google Fit, Jawbone UP, and other products that are producing personal health data (such as heart rate, calories, weight) that are geotagged. Although these types of data are personal and private, they could become a goldmine for health-related research if producers are willing to share and forgo privacy concerns.

Many VGI efforts produce thematic data using interactive mapping interfaces on the web, such that users of the site may contribute information to the map and dataset or gather information from it. Well-known platforms such as Google Maps and the Google API (application programming interface) can be modified to create map-based data collection interfaces for use in these projects. Although these platforms are a very commonly used data production tool, this type of VGI can be produced with more analog methods including text messaging. The Ushahidi platform can collect VGI contributed via text messages, which is then compiled and disseminated via an online map interface. In addition to interface and text message based strategies, VGI is also produced by georeferencing user-generated content online, typically through the practice of geotagging existing media such as audio/sound data (murmurtoronto.ca), photos (http://queue.acm.org/detail.cfm?id=2212756), and videos (Mills et al. 2010). Flickr, for example, encourages users to contribute geotagged photographs that can then be retrieved through a map interface along with text descriptions and links to other sources. This method of VGI production has grown significantly due to the automated georeferencing procedures in apps such as Twitter and due to the locating capabilities of handheld multimedia devices that automatically assign coordinates, location information, and attribute data to photos and other media.

Perhaps even more significantly, Google’s new mantra “Google Maps = Google in Maps” has made all the geotagged online VGI content searchable through Google Maps, which is revolutionary not only for geographic data production, but also for the design and development of search engines. More than just an easy way to georeference photos, this is potentially a valuable source of VGI. Twitter enables users to tag their tweets with locational information, a function that has already fostered interesting new possibilities for understanding the spatial patterns of very immediate or quickly changing situations, such as threat from wildfires or student-led occupying central protests in Hong Kong. A recent popular social media-based video collecting site is Vine (https://vine.co), a site that allows users to geotag all videos collected. While this type of VGI production typically occurs through users applying geotags to their own content, a growing number of geosocial networking applications are passively recording location information from their users, which requires very little if any input from the user. Many of these services are “opt in” meaning the user has to agree to share their location and activities (e.g., FourSquare, Instagram, Path, Vine, etc.), however, it is not always clear what the location information produced by the app’s users will be used for by the company. Given the range of ways that VGI is produced and the potential for producing new, previously unrecorded datasets, VGI has become an important area of geographic information science (GIScience) research. This represents a “wikification” of GIS (Sui 2008), a broader societal transformation in how geographic information is created and used. Exciting research both about and with VGI has been reported in recent years, as illustrated below.
Research about VGI

VGI quality assessment

Crowd-sourced data and VGI have been labeled by some as the scientific equivalent of fast-food, due to its ubiquity and relatively low cost to produce (Spielman 2014). This conceptualization poses an interesting question for the scientific community: is it healthy and safe to consume crowd-sourced data for research purposes given the limited ability to know its accuracy and credibility? The general consensus within the research community is that VGI offers an alternative mechanism for acquiring and compiling geographic information, but suffers from a general lack of quality assurance. Before applying VGI in any meaningful way, one must understand steps to assess the quality of VGI in hand.

In general, spatial data quality includes the following seven elements: position accuracy, attribute accuracy, logical consistency, completeness, lineage, temporal quality, and usage/purpose constraints. Goodchild and Li (2012) described three approaches for VGI quality assurance, which are described as the crowdsourcing, social, and geographic approaches respectively. As an extension of these three approaches, Haklay (see his entry in this volume on Volunteered geographic information: quality assurance) further articulated three additional approaches for VGI quality assurance – the domain approach, the instrumental observation approach, and the process-oriented approach. Bordogna et al. (2014) reported a linguistic approach for VGI quality assessment. It should be pointed out that the boundaries between these different approaches for VGI quality assurance are not so clear cut. One of the major problems among the existing approaches for VGI quality assurance is that they are blind to the sources of errors and do not differentiate at which stage of the data life cycle these errors occur.

To further address the VGI quality issue, VGI quality assurance research should be tied to the big data quality assessment techniques for maximum data usability (Loshin 2014). Like other types of big data, errors of VGI can be introduced in any stage of VGI life cycle, and they may be attributed to protocol and participants errors. In the broader citizen science literature, Wiggins et al. (2011) identified a total of 18 mechanisms for data quality control and validation in citizen science. These 18 mechanisms consider stages of data life cycle and sources of errors. The existing VGI quality assurance approaches as reviewed by Haklay (Volunteered geographic information: quality assurance) focus on the data collection stage and participants’ errors only. Additional work is needed to develop approaches to assess errors introduced in other stages of the data life cycle beyond the data collection stage. Moving beyond the participants’ errors, efforts are also needed to tackle the various protocol errors (e.g., poor sampling) in VGI production.

As spatial data quality assurance is shifting towards a more user-centric perspective, the precise meaning of data quality, accuracy, and uncertainty may be quite different from their traditional meanings in the context of surveying and mapping. Spatial data accuracy 2.0 (Goodchild and Hill 2008) and spatial metadata 2.0 (Kalantari et al. 2014) emphasizes more on the fitness of purpose in assessing the quality of VGI. In addition to the technical quality and accuracy (Esmaili, Naseri, and Esmaili 2013), social and political credibility is also becoming increasingly important in assessing VGI quality (Flanagin and Metzger 2008). All these new developments may further complicate the VGI quality assessment process.
VGI use methodology: from mapping parties to GeoHackathon

VGI use methodology has exhibited a few defining characteristics that are quite different from the traditional use of spatial data. The new emerging VGI practices are not only now increasingly conducted in the mode of open science and open GIS (Sui 2014), but they are also closely tied to Daniel Pink’s (2006) six senses of the whole new mind (design, story, symphony, empathy, play, and meaning). Sui (2015) documented that new GIS themes (of which VGI is an integral part) in geodesign, map stories, synthesis, critical GIS, geogames, and platial GIS coincide with Pink’s whole new mind remarkably well. Conventional GIS development has predominantly concentrated on cartographic, quantitative, computing, and spatial database-oriented with the goal of efficiency, which tends to be closely associated with the left-side/slow thinking capabilities of the human brain. In contrast, the uses of VGI focus more on narrative, qualitative, story-telling, and synthesis-oriented with the goal of equity, which tend to be more closely associated with the right-side/fast thinking capabilities of the human brain. Evidently, the increasing VGI applications have the potential to enable GIS users to transcend the enframing nature of technology and lead them to explore new territories with greater sensitivities.

The diverse information that is being generated as VGI presents a number of challenges for developing methodologies to make use of it. Although conventional methodologies for spatial analysis and modeling are still useful and used for VGI applications, a major shift in VGI use methodology is the transition from traditional analysis to mashup and synthesis (Elwood, Goodchild, and Sui 2012). Admittedly, synthesis is much more challenging than analysis from a methodological perspective, due to its involvement with data in multiple formats and media (number, text, oral story, photo, video, simulation, etc.). Geographers’ current efforts to develop a more eclectic approach by linking diverse quantitative and qualitative methods may be tapped for VGI applications. Unlike the spatial analysis of a previous era that was often conducted by a lone analyst, synthesis of VGI tends to be much more participatory, through a mixing and remixing of multiple data and methods, as demonstrated in VGI applications in several recent disaster-relief efforts. Instead of seeking truth, the new mashup efforts focus more on developing narratives about various locales. VGI production and use have evolved from the mode of mapping parties (http://wiki.openstreetmap.org/wiki/Mapping_parties) during the early days to GeoHackathon in most recent years (www.upsingapore.com/geohackathon; www.cura.umn.edu/visualizingneighborhoods). As more and more people, both GIS professionals and citizens, are getting involved in VGI production and use, we have witnessed the growth of a new level of volunteerism and civic activities within the broader geospatial community, as demonstrated by voluntary organizations such as giscorps.org or mapaction.org.

VGI use and legal issues

Similar to many other advances in geospatial technologies, the development of VGI has also brought in a new set of legal issues, ranging from intellectual property to liability, and from defamation to concerns about privacy. According to Scassa (2013), these legal VGI issues may manifest in different ways depending on the stakeholders concerned. The operator or producer of a product, website, or application that relies upon VGI will need to be aware of various potential legal issues, and should take
care to address these issues in licenses and other documentation related to the site. The extent to which they need to do so and how they do so may depend on whether they are public or private actors, or whether they operate commercially or noncommercially. VGI contributors should be aware of the license terms under which they provide content. Although they are unlikely to hold copyright in small quantities of information, creative material provided in other forms, most notably photographs, video, or text, is likely protected by copyright. In some cases, contributors of VGI may be institutions or organizations that provide data sets, and there may be a copyright that can be asserted in an original selection or arrangement of data.

Sites that incorporate VGI typically have terms of use or licenses that function as a contract with the user. The terms of these documents are important as they may govern issues such as the right to use or incorporate the information on the site into other works/products/websites and whether users may use the material for commercial purposes. For example, Creative Commons (2012) and Open Data Commons (2012) each have licenses that limit users to noncommercial uses of licensed content. Such licenses may require that any downstream uses of the contents of the site be licensed under the same terms and conditions. Other license agreements may set out the rules regarding attribution: a user of licensed data may be required to attribute the source of the data or to do so in a specific format.

Privacy and surveillance are significant concerns associated with the phenomenon of VGI. Geographic information contributed by users to a website or database enables the ability for location tracking and monitoring, which may lead to an invasion of privacy and the potential for harm (Andrejevic 2007). Many social media sites and location-based services record the locations of their users in space, which is sometimes actively provided by the user, however, in other cases it is produced passively without the user fully understanding this or its consequences. Some location-based services such as FourSquare utilize this location information in developing their service, while others compile it for use by other companies for targeted geo-demographic marketing purposes (boyd and Crawford 2012; Wilson 2012). Willingly or unwillingly sharing one’s location – whether for the purposes of contributing VGI or not – is a serious safety concern, and presents one of the most serious issues implicated with the recent integration of mobile and location technologies (Elwood and Leszczynski 2011). In crowdsourcing and VGI projects, users should be made aware of what the data will be used for, and any potential privacy infringement related to their contributions.

Research with VGI: using VGI as a data source for various applications

A wide variety and rapidly increasing number of projects have been developed in recent years to take advantage of the phenomenon of VGI and crowdsourcing. In some cases, the projects have been developed in order to produce geographic information where none existed before, and in other cases they are motivated by a desire to augment or complement existing formalized sources of data. A range of examples of each motivation is provided below. For more examples, refer to the report by Haklay et al. (2014), a comprehensive collection of crowdsourcing examples including VGI projects from around the world.

Four wildfires in 2010 in California provide an interesting example of this type of activity, and serve to emphasize the value of VGI as a mechanism for creating and sharing geographic
information in time-critical situations, a point made earlier in the context of framework data (Goodchild and Glennon 2010). The Jesusita fire (May 2009) generated large amounts of Internet traffic, as citizens posted text descriptions, photographs, and other kinds of information within minutes of the fire’s outbreak. Individuals and groups of citizens used the framework to compile this information on a minute-by-minute basis into situation maps, showing the location of the fire front, areas under evacuation orders, and other relevant information. These maps served for many citizens as the timeliest source of information, supplanting official sources that could not keep up as well with the rapidly evolving situation. By linking various wildfire VGI with Twitter, social media has been proven to be an effective way to disseminate early warnings in emergency situations (http://newscenter.sdsu.edu/sdu_newscenter/news.aspx?newsID=75219).

“Road Watch in the Pass” is an example of a citizen science project that produced VGI about wildlife sightings along a stretch of highway in British Columbia, Canada (Lee, Quinn, and Duke 2006). The project was set up to allow citizens to share sightings of wild animals including deer and moose for the purposes of conservation and public safety, given the problem of collisions between wildlife and vehicles in this area. Locations along the highway where wildlife was sighted were identified on a web-based map. The data produced was used to augment reports from conventional sources, therefore increasing knowledge about solutions for mitigating animal–vehicle collisions at high-risk locations.

The Geo-Wiki platform (www.geo-wiki.org) is a VGI project that enables volunteers from around the world to contribute to a global land classification project (Fritz et al. 2012; Fritz et al. 2009). Participants evaluate the accuracy of land-cover classification in existing land-cover maps, by comparing them with Google Earth imagery and through their own knowledge of local areas. This project has been successful in harnessing the wisdom of the crowd to settle disputed land-cover classifications in existing maps. The value in this approach is the speed by which land-use classification can be verified and the low cost compared to conventional approaches.

The United States Geological Survey has a crowdsourcing program that collects local reports of seismic activities from the public. The Did You Feel It program was set up to provide an additional source of information that could be collected quite rapidly in the aftermath of an earthquake event. Any web user can provide reports of local seismic activity on the website (http://earthquake.usgs.gov/earthquakes/dyfi/), which can then be used for logistical planning and recovery. This program was initially restricted to the United States only, but is now set up to collect reports worldwide.

FixMyStreet is a website and mobile app that allows citizens to report various urban infrastructure problems such as potholes, graffiti, and abandoned vehicles in the United Kingdom (www.fixmystreet.com). This project – operated by the civil society organization mySociety – uses a web-based platform to allow members of the public to produce VGI pertaining to observations in their local area. The user enters the postal code of the area in which they wish to report an observation, selects the specific area on the map, and then adds a textual description of the problem. Users can also view reports made by other users and all reports are then forwarded on to local councils.

A project organized by the City of Boston is using a passive crowdsourcing approach to produce VGI related to street conditions called StreetBump, in which citizens are contributing information about potholes and other road quality concerns (www.streetbump.org). StreetBump
is an app that the user runs while driving, which is able to automatically collect information on bumps in the road through a smartphone’s built-in GPS and accelerometer. The app is designed around the idea of “gamification,” in which contributors score points if the pothole is fixed. If three bumps are recorded at the same location, this site is identified and added to the list of locations to be addressed by the city.

The US National Park Service is developing the Places of Interest project (www.nps.gov/npmap/tools/places), which will utilize a VGI approach to produce geographic information that will be added to their park maps to keep them up to date and accurate. The project will also contribute VGI to the development of OpenStreetMap, which the National Park Service already uses in their digital mapping projects. The plan is to develop an interface for the public to contribute geographic information related to the national parks – such as trails, tourist sites, and park infrastructure – and have a data validation procedure in place to ensure the quality of data. If relevant and accurate, the data produced will be added to the databases of both the National Park Service and OpenStreetMap.

Map Kibera is a VGI project started in 2009 that empowered local residents of the Kibera informal settlement in Nairobi, Kenya to produce a detailed local area map (http://mapkibera.org). Despite the extensive size and number of people living in this area, no previous, comprehensive map existed, which made planning and community organization a challenge. Prior to this project, Kibera was a “blank spot on the map,” which visually underscored the socioeconomic disparity in access to and provision of geographic information in Nairobi. The project provides local residents with GPS receivers to record streets, buildings, and points of interest, which are then added to the OpenStreetMap database.

The Humanitarian OpenStreetMap Team (HOT) has carried out a number of projects that demonstrate the value and credibility of VGI produced by trained volunteers in a structured way. A number of projects have been undertaken around the world, including recently in Ulaanbaatar, Mongolia (http://hot.openstreetmap.org/projects/mongolia_mapping_ulaanbaatar). The local government in Ulaanbaatar has teamed up with HOT to engage local residents to produce a comprehensive map of roads and urban infrastructure such as street lighting, bus stops, buildings and landmarks, parks, schools, and hospitals. Training and oversight is provided by HOT to increase the accuracy and potential value of the data and maps produced.

Summary and conclusion

Rapid advances in technology, along with changing demands for geographic information, have dramatically altered the environment within which this information is produced and accessed, effectively removing the economies of scale and assured patterns of use that allowed national mapping agencies such as the US Geological Survey (USGS) or Ordnance Survey (OS) in the United Kingdom to evolve and flourish. Perhaps the most important contributor to the advancement of VGI is the growth of the web, particularly the increasing participation of its users in generating its content in the age of Web 2.0 along with the development of mobile technologies and handheld devices such as portable GPS devices and smartphones. Through the successful stories of VGI reported so far, it is abundantly clear that VGI is rapidly becoming a valuable data source in low-resource areas or time-sensitive situations.
VOLUNTEERED GEOGRAPHIC INFORMATION

In addition to its contributions to environmental monitoring via citizen science, VGI has also promoted transparency in government operations and fundamental changes in business practices and models in various industries with the emergence of the sharing economy. Citizen sensors all over the world have rejuvenated a higher level of volunteerism.

With the growing heterogeneity of VGI and its contributors, the quality, credibility, and uncertainty of VGI also increase. This can be challenging to communicate with the public or for people who are used to working with authoritative data sources in a traditional top-down model (Haklay et al. 2014).

As this entry is written (late 2014), it is believed that the development of VGI and its applications so far have followed the storyline of a long tail along multiple dimensions according to the anecdotal evidence we observed. Despite the lack of quantitative measures, the long-tail nature of VGI has manifested at least in the following two aspects. First, as far as the spatial data infrastructure is concerned, traditional authoritative data collected by government agencies still serve as the “head” (large data for several key application areas) whereas VGI is serving the “long tail” (a variety of relatively small data for a diverse range of applications). Second, as far as the VGI production as a whole is concerned, contributors or players in VGI production also follow a power-law distribution: a relatively small number of contributors account for a disproportional percentage of the VGI production. One interesting development during the past five years is that VGI is, in fact, only a branch of the big data stream and is also merging rapidly with the big data deluge. Not surprisingly, as more and more big data with spatial and temporal stamps are available, boundaries between VGI and non-VGI are increasingly blurred day by day. Perhaps we are rapidly approaching the day when VGI will be synonymous with big data or will be largely indistinguishable from more conventionally produced spatial data. VGI has proved to be a valuable source of data for multiple exciting applications, but at the same time it also raises profound legal, ethical, social, and political issues regarding its collection, retention, dissemination, and use.

SEE ALSO: Big data; Information synthesis

References


Elwood, Sarah, Michael F. Goodchild, and Daniel Z. Sui. 2012. “Researching Volunteered Geographic Information: Spatial Data, Geographic Research,


Further reading


Stephens, Monica. 2013. “Gender and the Geoweb: Divisions in the Production of User-Generated


Vulnerability

Robert McLeman
Wilfrid Laurier University, Canada

Vulnerability refers to the potential to experience loss or harm. The term appears widely across the natural and social sciences. In geography and related disciplines, vulnerability is commonly used when describing the relationship between humans and the natural environment from a socioecological systems perspective, where changes in one part of a human–environment system have the potential to stimulate changes elsewhere in the same system or in other, linked systems. Vulnerability is employed in a relational sense, that is, the vulnerability of something to something else, such as the vulnerability of a farming population to drought. Vulnerability (of a system, or a unit or actor within a system) is determined by the nature of the adverse stimulus (often referred to as exposure); the sensitivity of that system, unit, or actor; and its capacity to adapt (Smit and Wandel 2006). In the UN Framework Convention on Climate Change (UNFCCC) the term “vulnerability” has important legal implications, including an obligation that signatories will assist vulnerable countries in adapting to the impacts of climate change. The UNFCCC has consequently had a significant influence on the volume and direction of vulnerability research by geographers, and will continue to do so. The International Geographical Union’s (IGU) Commission on Population Geography has a strong, ongoing interest in demographic dimensions of vulnerability, which reflects new directions in vulnerability research being pursued by geographers.

Ecologists began using the term “vulnerability” in the early twentieth century to describe potential stimulus–response relationships in natural systems (e.g., effects of predation on bird populations (Errington 1934)). By the 1970s the term had been adopted by researchers studying the human impacts of natural hazards, with human ecology (or political ecology) scholars advancing its development in the 1980s through research on the causes of soil erosion, land degradation, and famines (e.g., Blaikie 1981). In the 1990s vulnerability became a key concept for describing the human impacts of anthropogenic climate change. Current vulnerability research reflects these combined influences of natural science–derived systems theory and social science–based human ecology (Adger 2006). In socioecological systems approaches to vulnerability, the human world and the natural environment are seen as being interconnected sets of complex and dynamic systems, where changes experienced in one have consequences for the other. The ability of a system to successfully absorb perturbations or to adapt to changing conditions without experiencing significant disruption, or changing from one state to another, is described as resilience (Holling 1973). Resilience, which has its own associated scholarship focusing on system transformation and learning (Miller et al. 2010), helps in the identification of vulnerability. For actors, groups, or populations within a given socioecological system, a condition of vulnerability develops when potential perturbations or changes exist that exceed the resilience of the system. Vulnerability increases positively with the strength or force
of the perturbation; similarly, the greater the sensitivity to that perturbation, the greater the vulnerability. Conversely, the better adapted systems (or actors or units within them) are to perturbations, the lower their vulnerability (Figure 1). Because human and natural systems are continually changing in response to internal and external stimuli, the nature of and degree of vulnerability are also continually changing.

A simple example of the impacts of climate change on farming illustrates the functional relationship between exposure, sensitivity, and
adaptation in the formation of vulnerability. Most farms are potentially vulnerable to droughts, given the inherent nature of agriculture, but not all farms are equally so. Some types of farm operation are more sensitive than others to temporary shortages of water, such as dairy farms, because cattle must consume large volumes of water daily. Similarly, fresh vegetables require larger, more regular amounts of water than most grains. The degree of exposure to drought varies according to local and regional climatic conditions and patterns, and to site-specific factors such as soil quality and drainage. Farmers may employ a range of adaptations to reduce their exposure and sensitivity to drought, such as planting drought-tolerant crop varieties, using groundwater for irrigation or watering livestock, buying crop insurance, pooling resources with other farmers, or supplementing household income through off-farm employment. The availability and accessibility of adaptation options vary from one place to another depending on governance structures, institutional arrangements, market prices, technologies, and other factors operating across multiple spatial, temporal, and jurisdictional scales. With climate change threatening to increase the frequency and severity of droughts in many regions, the vulnerability of farmers will increase in the absence of building greater adaptive capacity (Smit and Wandel 2006).

This functional understanding of vulnerability, and the related concepts of resilience and adaptation, have been used by geographers to explore contemporary socioenvironmental challenges including food (in)security, the spread of HIV/AIDS, disaster risk reduction, environmentally related migration, and the potential impacts of climate change. Such applications have brought attention to the uneven social distribution of vulnerability and its root causes in the unequal access of individuals and households to resources, wealth, social capital, and other assets (Bohle, Downing, and Watts 1994). The least powerful, most marginalized members of society are often the most highly exposed to environmental risks and have the least adaptive capacity. This susceptibility is regularly seen following natural disasters, when rates of death, injury, property damage, and displacement show distinct patterns along lines of race, gender, age, and poverty. For example, when Hurricane Katrina struck New Orleans in 2005, the poorest residents (disproportionately composed of African Americans and minorities) were living in neighborhoods most damaged by flooding, were most dependent on underprepared government authorities for assistance, and had the least individual capacity to adapt to post-storm increases in the cost of housing. They were consequently more likely to be permanently displaced (Fussell, Sastry, and VanLandingham 2010).

Theoretical and empirical research continues to grow in the study of the physical vulnerability of locations and infrastructure to natural hazards, the social dimensions of vulnerability, and their interactions. An emergent area of research investigates how vulnerability to environmental change can be both a driver and an outcome of human population processes. It has long been known that human population growth (and accompanying demand for resources) is an important driver of environmental change. A disproportionate amount of that growth is occurring in regions with high rates of land degradation and where exposure to natural hazards and climate change risks are high. Within exposed populations, rapid changes in age distributions, fertility rates, and migration behavior affect social networks, social capital, gender roles, and related attributes that influence the adaptive capacity of households and communities (McLeman 2010).
VULNERABILITY

Vulnerability researchers employ many different qualitative and quantitative methods, often mixing them to achieve particular results. Because vulnerability research is often aimed at supporting policy and planning, there is growing emphasis on creating tools that enable the visualization of vulnerability. GIS-based technologies are particularly valuable for these purposes. The UNFCCC negotiation process generates ongoing demand for research on regional and sectorial vulnerability to sea level rise, ecological changes, increased frequency/severity of extreme events, and other impacts of climate change, with increasing attempts being made to measure and quantify vulnerability. Researchers also seek to develop future vulnerability and adaptation scenarios that incorporate more sophisticated models of future socioeconomic and demographic trajectories (Stern et al. 2013).

Coming decades will see continued growth in vulnerability-based research generally, a result of its conceptual simplicity, methodological utility, and widespread acceptance by policymakers. High rates of urbanization in developing regions will demand greater research attention on urban vulnerability to environmental risk and change. The linkages between environmental vulnerability and mobility, food security, and political instability will receive growing research attention, as will questions of governance structures best suited to deal with such challenges. Vulnerability research is unlikely to consolidate around a particular set of methodological approaches anytime soon, but is instead likely to see ongoing expansion and innovation in methods and techniques being used. Rapidly evolving technologies like mobile computing, online social media, and crowdsourcing apps are opening new avenues for data collection and analysis, and for public participation in vulnerability research.

SEE ALSO: Climate adaptation/mitigation; Climate change adaptation and social transformation; Climate change policy; Climate and societal impacts; Human ecology; Natural hazards and disasters; Social resilience and environmental hazards; Social vulnerability and environmental hazards

References


War

Benedikt Korf
University of Zurich, Switzerland

It has been proposed that geographers – or geography as a discipline – have been better at studying war than peace. The French geographer Yves Lacoste famously titled one of his books on geopolitics: *La géographie, ça sert, d’abord, à faire la guerre* (Lacoste 1976) – that geography serves primarily to wage war. This statement might partially apply to certain “geographic” techniques, such as cartography, geographical information systems, and remote sensing, which are all important tools in waging war, even though this was not necessarily the original purpose of their development. But the complicity of geographers with agendas of war-making has often been more deeply implicated in political projects of colonialism, imperialism, Nazism, and territorial expansion in the nineteenth and first half of the twentieth century. These geographers have been explicitly apologetic to the pursuit of war. It is fair to say, however, that this might have been true for some geographers, but not for all, as there have also been important figures in geography promoting peace rather than war (Megoran 2011).

While there continues to be a group of military geographers who explicitly provide analysis for military establishments, critical geographers today rather seek to provide an analysis of the phenomenon of war as a contribution to a critical theory of society, state-making, capitalism, imperialism, and globalization. Such critical theory requires a careful analysis of war and its relation to peace, to politics, to society, and to capitalism.

Clausewitz’s puzzle and beyond

“War is the continuation of politics by other means,” wrote the Prussian military general and theoretician of war Carl Clausewitz in 1832 – as an afterthought to the Napoleonic wars (Clausewitz 2005/1832). These wars had rampaged through most of Europe’s landscape, causing death and destruction. But they had also triggered political reform in some countries, caused national awakening, fostered democratic and republican movements across Europe, and, finally, had brought the restoration of monarchical order as a result of Napoleon’s defeat on the battle ground. Not surprisingly, then, for Clausewitz, war was a chameleon: depending on context and perspective, its effects can be colored very differently. But Clausewitz suggests that war is even more than a chameleon; it is a total phenomenon, composed of a trinity of (i) violence, hatred, and enmity, (ii) the play of chance and probability, and (iii) subordination.

Much controversy has emerged around the question of whether Clausewitz is still relevant in the twenty-first century, and although it is not our task here to write a critique of Clausewitz, this controversy indicates the complexity of the phenomenon of war. Clausewitz developed his theory of war in a historical era when wars were often fought between states, or rather dynasties, as many of these states were not yet nation-states in a modern sense but rather dynastic monarchies. This type of warfare reached its height in the two world wars of the twentieth century, with large armies confronting each other on the battlefield. This is also how “conventional” war is usually defined:
WAR

as a military confrontation between two or more countries, states, or political collectivities, and their combat forces. In line with this state-centered view of war, there is a set of legal prescriptions in international law on how to fight such wars: how to declare them, how to end them, and how to conduct them. An important distinction in these prescriptions is that between combatants and civilians, and there are legal definitions of what a “proper” war is. Proper war is also subject to moral philosophical debates to distinguish “just” from “unjust” war, by defining conditions and causes that do or do not justify the pursuit of war.

There are at least three problems with this view of proper war: first, there are many wars (they have in fact become the vast majority in recent decades) that do not fit into these categories. These “other” wars are then often attributed with additional labels to distinguish them from “conventional” war: for example, civil wars (intrastate wars, fought out between a government and insurgents, or between different opposing militant groups), partisan wars (fought by resistant groups against an occupying force), wars of liberation (fought against colonial powers, for instance), and terrorism. With the end of World War II, and more so after the end of the Cold War, many observers suggest, this latter type of war has become dominant, while the number of interstate (conventional) wars has decreased significantly.

The second problem with the definition of “proper” war is that many, if not most, interstate, conventional wars were much messier than the categories attributed to describe them. Many of them displayed elements of civil war, partisan war, counterinsurgency war, and so on, besides and beyond the combat field of the conventional armies with their tanks, soldiers, artillery, aircraft, and other military technologies.

The third problem is that the definition of proper war also defines the antithesis of war: peace. This definition marks a clear boundary between war and peace: as soon as war is declared, peace is over, and as soon as a ceasefire is declared, peace has returned. But in reality this boundary is often fuzzy – we might find ourselves in a situation of not war anymore but not (yet) peace either, or of simmering, low-intensity warfare, or of gang violence, of militant rule. David Keen (2000) therefore asked: War and peace – what’s the difference? In many situations today, this difference has become blurred.

For the philosopher Michel Foucault, “peace itself is a coded war” (2003, 51). Foucault suspected Clausewitz to have inverted the original proposition that politics is the continuation of war by other means, rather than, as Clausewitz had suggested, that war was the continuation of politics by other means. Foucault writes: “Right, peace, and laws were born in the blood and mud of battles” (2003, 50). According to Foucault, war is not something that comes after politics has failed, but politics itself is war. War is not a remnant of a premodern past, but continues to permeate societies – to purify their inner social body and polity. To understand the full extent of Foucault’s provocation, we need to contrast his argument with that of Thomas Hobbes in Leviathan (1996/1651). Hobbes saw the question of war as an institutional problem. War would inevitably emerge in a “state of nature” – a presocietal condition where nobody holds a monopoly of violence. Intrastate war, to Hobbes, was a problem to be solved by state building that would bring about a monopoly of violence and thus security and social peace. Here, we can see how Foucault and Hobbes differ: for Foucault, war continues to permeate society even after a state has established a monopoly of violence.
Warscape

With the end of the Cold War, many observers expected an end to armed conflict. But surprisingly, the number of armed conflicts increased after 1990, at least in the realm of public media attention. Many of these wars of the 1990s and thereafter irritated modern observers. They were subsequently called “new wars”: new not in the sense of hypermodern or advanced, but rather new as different from conventional (“old”) wars. Paradoxically, these new wars brought the premodern past back onto the battlefield in the sense of a return of “archaic” forces of “primitive” or “primordial” warfare. These new wars were often labeled as “asymmetric”: they were not fought between states, but between states and a variety of violent nonstate actors; they were archaic, brutal, and primitive, involving large-scale massacres (Rwanda, Bosnia), mass-scale sexual harassments, and erratic violence against civilians (Congo, Sierra Leone, etc.). These new wars were said to be predatory, involving rampant warlords (Congo, Sierra Leone, etc.), and a blurring of the distinction between combatants and civilians, between war and business, between war and peace. Yet, at the same time, it became clear that these wars were not an exception to, or disconnected from, globalized modernity; in fact, the flows and forces of globalized capital were intricately related to the political economies of “new wars” (Kaldor 2012). Then came 9/11 and with it the two US-led wars in Afghanistan and Iraq, which have consumed much attention of critical geographers interested in war. Many observers see this as an asymmetric war between a global power (the United States and its allies) and nonstate actors (Taliban, insurgents, tribal groups), where the former has a super-hegemony in terms of military technology and nevertheless fails to systematically control the battlefield and bring the war to a sustained end. Violence has continued to unsettle large parts of Afghanistan, northern Pakistan, and Iraq, despite the presence of military personnel, and the utilization of hypermodern military and intelligence technologies and counterinsurgency tactics.

What we find in those battlefields, and in different figurations in other wars, is a complex “warscape”: a landscape of violence, destruction, coercion, suppression, fear, suffering, and of excitement, enrichment, opportunity, technological advancement, and social transformation. Amid and besides the struggles on the battlefield, there is a more silent space of struggle; this is the space of all sorts of noncombatants who happen to live in warscapes. Violence has ambiguous, often contradictory, effects on agency at different times, in different places, and for different individuals: war is not a matter of “all terror all the time all over the place” (Korf, Engeler, and Hagmann 2010, 386). Inhabiting a warscape is thus not simply a question of coping with violence or adapting to new power holders; it is entangled in multiple and differentiated agendas of social struggle, life projects, and negotiations over livelihoods. It is therefore difficult to draw a clear line between the social conditions of war and those of nonwar as social actors continue to struggle throughout both conditions in a peace-to-war continuum, which has become fuzzy in many of today’s wars: there is no clear beginning or end to a war, but rather a diffuse grey zone of hidden, erratic forms of violence and more open combat. Similarly, it is often not possible to distinguish neatly between perpetrators and victims, as some noncombatants may take profit out of war, while others might experience tremendous suffering, and the same individual might at some point in time belong to the former and at another to the latter category.
Synchronicity of the nonsynchronous

Warscapes are not confined to a locale, but are embedded in transnational networks of accumulation, financial capitalism, technological grids, and blurred boundaries between “archaic,” “modern,” and “hypermodern” (futurist) subjectivities, objects, technologies, and practices of warfare. These blurred boundaries encapsulate what the German philosopher Ernst Bloch has described, in a different context, as “synchronicity of the nonsynchronous”: all these objects, elements, and subjects ought to be in different historical times; they seem to belong to different historical eras of warfare – the past, the present, the future. But they are all co-present in the same historical moment – the present condition; and they may even be co-present in the same warscape (e.g., Afghanistan, Iraq), while in others, one particular historical time may dominate our perception of the battlefield: for example when violence in the Rwandan genocide was labeled as “archaic” or when drones are seen as futurist elements in present-day warfare.

This synchronicity of the nonsynchronous can be mapped along a continuum from the body to the drone. The most “archaic” subjectivity in these warscapes seems to be encapsulated in the figure of the suicide bomber, and her object of war: the body. At the other end of this continuum is the drone that wages war through an unmanned robot vehicle that is steered from afar – where the warrior operates from a distance. The “heroic” dimension of manly combat that has been an important element of “archaic” warfare has thereby escaped into the grids of the technologies of the cyborg warfare of a “post-heroic” society. Warfare is steered from afar, but modern technologies make the battlefield present on the computer screen – an intimacy from a distance that is mediated through hypermodern technology (Gregory 2010). It links the messy warscape of blood, dust, and struggle with the clean office of the drone operator at a safe distance from the place of combat.

The geography of war

Geographers have been involved in studying warscapes from a variety of angles – and there has been a proliferation of geographical work on war and violence in the last decade. Not surprisingly, there have been differentiated approaches, methods, and theories guiding different geographical scholars in studying warscapes – and they have also focused on different types of warscape. One group of geographers has focused on warscapes with a heavy involvement of the US military, in particular on the battlefields of Afghanistan and Iraq. The Palestine–Israel conflict, the escalation of violence in Gaza, and violence in the Middle East more broadly have also attracted the attention of geographers. But there are many other contemporary warscapes – beyond these hotspots – that geographers have studied: Sri Lanka, Nepal, East Timor, the Horn of Africa, Sierra Leone, Rwanda, Colombia, Mozambique, Angola, and many more. Some geographers have studied past atrocities and warscapes: for example, the legacy of Cold War violence, the geography of civil war, mass violence, and other atrocities in countries such as Cambodia during the Khmer Rouge and Rwanda and its genocide, but also East Timor, Vietnam, and Laos. In these latter sites, mass-scale violence has been a historical, often traumatizing experience.

The US intervention in Afghanistan and Iraq as part of the global war on terror has drawn a lot of attention from geographers during the past decade. These wars have been high up in global media discourses and in the critical, self-reflective engagement of geographers who
felt obliged to challenge their home countries’ implication in massive military operations in various countries and sites across the globe. In these warscapes, where the United States has sustained a massive military presence over a long time period, fighting an asymmetric war against Taliban or other insurgents, the synchronicity of the nonsynchronous objects, technologies, and subjectivities has been pronounced in its most peculiar way. In the Afghan warscape, for example, hypermodern technologies (drones, cyberwarfare) were deployed on the battle-ground amid the troops and machinery of “conventional” modern armies (soldiers, tanks, aircraft) with “tribalist” elements of war, the Kalashnikov fighting of warlords and elements of partisan war (road blocks, insurgents’ attacks, counterinsurgency). The “archaic” mode may also include blood killing, revenge, and “irrational” violence: in the case of Rwanda, killing with machetes rather than with guns. At the core, these objects and technologies of warfare appear to belong to different historical times. But also the subjects fighting war are attributed to different historical epochs, with tribal leaders and warlords being consigned to a “premodern” era, soldiers representing the modern (the present), and the “cyborg” of hypermodern technologies blurring the boundaries between human and nonhuman and bringing the future into the present.

A number of geographical studies resonate with these observations. A growing field of study in geography has looked into the emerging battlespace of the war in Afghanistan as a network-centric warfare, where hypermodern or “futurist” military technologies are mobilized to produce a “surgical” type of warfare that combines military intelligence from remotely sensed surveillance with targeted killings, often through drones or robots. Derek Gregory calls this the deadly embrace that reconfigures the relation between war, distance, and intimacy: the soldier is not immersed physically in the battlefield. The battlefield is now digitized on the computer screen. It is distant and yet intimate, as modern surveillance technologies appear to bring the battlefield’s configuration back to the operator’s office. These new technologies of warfare foster an unbounding of military activities across the war – what Gregory calls “everywhere war” (Gregory 2011). But this type of war occurs alongside other types of military technologies and violence, such as paramilitary and terrorist violence, as well as more conventional military operations, which require large-scale logistics to transport troops and weaponry for large distances. For Gregory, this “everywhere war” can break out and emerge in very different sites and in different types of war, beyond the battlefields of US interventions in the Middle East.

The proliferation and increasing fuzziness of war is captured in the concept of “battlespace,” which “prefigures a boundless and unending process of militarization where everything becomes a site of permanent war … [It] has no front and no back and no start and no end” (Graham 2009, 389). For Steve Graham, permanency signifies the emergence of a new military urbanism: the world’s main battlegrounds are now becoming urban, rather than the dusty, remote, rural frontiers that are normally associated with warfare (Afghanistan, Iraq, Somalia, Sierra Leone, etc.). The everyday spaces, sites, and infrastructures of the urban have become the main targets and threats for attacks. Insurgents, terrorists, and other nonstate militant actors lurk invisibly in the everyday spaces of urban life; they exploit and target the circulation and flows within and between cities – air travel, the Internet, financial flows. The new urban militarism represents a process of insidious militarization of urban life, a securitization of the city and an urbanizing of security. Urban everyday life becomes
WAR

the Achilles heel of irregular and asymmetric warfare – the urban becomes a target of violence, but it is also transformed to garner possibilities of violence through it, for example in the attempts of the US military or Israeli forces to demodernize urban societies through the destruction of infrastructure. Graham identifies a militarization and securitization of modern urban life in the proliferation of surveillance technologies, in the increasingly militarized dealing with immigrant sites as potential dangerous spaces in the city, and in the militarized veins of popular culture. For Graham, then, war is everywhere and everything. If we accept such a position, there is, however, a certain danger that the concept of war loses its value as an analytical category: if everything is war, the thing that is not war becomes not only blurred and indistinguishable from war but vanishes from political life – as a social reality and as an aspiration.

Another influential study, by James Tyner (2009), *War, Violence, and Population: Making the Body Count*, has invited population geographers to embrace more systematically structural forms of violence and war in the study of populations. Tyner studies the politics of violence in the genocidal atrocities of Rwanda and Cambodia as well as in the Vietnam War. He frames this violence as a form of biopower in a Foucauldian sense: death and destruction are often seen as being productive as a politics of life that protects or enables certain lives and populations. Tyner’s work resonates with feminist geographers’ insistence on and careful study of the body politics and the gendered geographies that warscape produces (Giles and Hyndman 2004): bodies count (i.e., the politics of the body) rather than simply a body count (of how many people are dead or injured, i.e., an aggregate sum of war’s collateral damage). Geographers have analyzed the geographical imaginations framing this politics of the body, which can take various dynamics and forms: gendered subjectivities might become un-settled and re-enforced through warrior images of female fighters or child soldiers. Postcolonial inequalities might be reasserted through differential practices of implementing security for “different” bodies – “local” versus “international” personnel, military versus civilian (Fluri 2011) – or in the media attention and public grief attributed to “different” bodies – the loss of individual US soldiers, missing or dead, is acknowledged, while the dead or missing Afghan or Iraqis are reported in the hundreds or thousands.

Another prominent field of study among geographers has been the problematic “resource wars” (Le Billon 2008). These are wars where the accumulation of resource rents from high-value mineral resources (diamonds, oil, etc.) have provided incentives for warlords to keep war ongoing, but in a way that does not disturb, but rather promotes, their business interests. War enables those warlord entrepreneurs to monopolize resource rents, through securing access both to the sites of exploitation and to the sites of transnational trading. The convoluted transnational entanglements of Western consumers, international companies, and “local” warlords has been captured in the concept of “blood diamonds”: the consumer good of diamonds have a past that is rather bloody, excavated in the minefields of some of the most protracted wars in West and Central Africa. Again, there seems to be a kind of nonsynchronicity at play here: in the liberal minds of Western observers, trading is thought to promote peace rather than war. These bloody resource wars are “archaic,” a savage type of “premodern” business practice, that is nevertheless part of global capitalism, its production networks, and its consumer culture.

The problem with the literature on “resource wars” has been that some of this work has been found to purport a soft or “quasi” environmental
determinism by linking resource abundance with political violence – or by overemphasizing the economic incentives of warfare, while downplaying other dimensions of political struggle. Again, the analytically interesting point in this form of political violence is the entanglement of political struggle, insurgent logistics, and battleground dynamics with the highly complex transnational spaces of the marketization of mineral resources, such as oil, diamonds, and gold, which links the global borderlands of capitalism with its metropolitan centers. And yet, the local battlegrounds within which those mineral resources are exploited retain their own characteristics and dynamics, which can be described as the “telluric geographies” of partisan, insurgent warfare, requiring a terrain of operation, a complicated entanglement of insurgents and civilian population with often fuzzy boundaries, shifting sites of control, violence and atrocities, and the affective geographies of fear, hope, and collectivity (Korf 2011). One important element in these geographies is the often persistent ungovernability of these sites, which are characterized by overlapping and contradictory governable spaces with different rules, governing logics, and claims to loyalty (Watts 2003).

This brief survey of geographers’ work on war highlights some of the most innovative contributions of geographers to the interdisciplinary field of war and conflict studies. Three characteristic features can be identified: first, most geographers engage with warscapes. The word “scape” signifies the fluid and irregular shapes of warscapes across different sites, including their transnational entanglements of spaces of militarization with the global capitalist frontier. Second, in one way or another, they reflect the synchronicity of the nonsynchronous that shapes such “scapes” – the moral imaginations of past, present, and future that are ascribed to different technologies, objects, and subjectivities in these sites. And third, they engage and weave together the three elements of war that Clausewitz branded as the “trinity” of violence, chance, and suppression.

Clausewitz wrote that war is a total phenomenon. For geographers studying war, there is an epistemological challenge in capturing this totality of the phenomenon, while not neglecting the intricate entanglements of localized social practices that are spread across different sites and that are deeply implicated in the everyday experiences of warfare. Geographers have contributed significantly to better understand, empirically and conceptually, different sites, practices, and technologies pertinent in these complex warscapes. While much work has emphasized that such entanglements exist and has started to conceptualize them, studying these intricate entanglements empirically remains a challenging task to be accomplished more fully.

After Clausewitz: the irritation about an irritation

Clausewitz wrote that war is more than a chameleon. In late-modern wars, it seems that what makes war a chameleon is the synchronicity of the nonsynchronous. This nonsynchronicity is irritating to the eyes of the Western observers. It reminds the late-modern (Western) subject about some remnants of a premodern past that is co-present in contemporary wars. And these remnants of the past are not “neutral” in normative terms: the (cognitive) irritation about their presence comes from the (negative) normative ascription of these pasts, their subjectivities, objects, technologies, and practices. This “past” warfare is said to be “irrational,” “inhuman,” “archaic,” “brutal,” “uncivilized”: the seemingly “futile” heroism of the suicide bomber
who sacrifices his (or her) body and life in war appears to be an expression of madness. There is a latent racism at work here, because those brutal, uncivilized fighters are primarily located in the Global South, in Africa (the “dark continent”) and some parts of Asia. Modern armies, and especially the futurist technologies, tend to remove the bodies of their own soldiers from the battlefield (the drone approach) or surround the body of the soldier with the newest technologies to protect it from physical harm. And yet, the presence of the premodern warrior subjectivity, objects, and technologies forces even the most modern and futurist army to keep the soldier subject on the battleground.

It is not only the phenomenon of this synchronicity of the nonsynchronous in warscapes that is troubling; the fact that it troubles us should invite critical geographers to uncover the moral imaginations tied to these representations of wars. The question is why certain technologies of killing are attributed with normative superiority. These moral imaginations that cause irritation about the nonsynchronous elements in the present place the “premodern” elements of warscapes onto an inferior normative level (primitive, cruel, brutal, inhumane) compared to the “regulated” space of conventional warfare that is represented as “rational,” “legal,” and “humane.” This representation all too often justifies Western military interventions in the name of security, democracy, development, and humanitarianism – to protect the inhabitants of these warscapes from their own “lack of civilization.” These normative imaginations should be irritating – but are not necessarily surprising. As Foucault reminds us, politics itself is war – and global politics often amounts to a war over the geographical imaginations of our postcolonial present.

SEE ALSO: Colonialism, decolonization, and neocolonialism; Geopolitics; Military geography; Natural resources and human conflict; Peace; Peacekeeping; Political ecology; Political geography; Postcolonial geographies; Postconflict geographies; Security; Violence

References


Further reading

Waste and waste management

Josh Lepawsky
Memorial University of Newfoundland, Canada

What is waste?

In everyday language waste usually refers to something to be discarded as unwanted, unusable, or as a byproduct of some process. It may also be used to describe careless, frivolous, or extravagant behavior. But waste defies easy definition. We may know it when we see it, but presupposing a universal definition would be a mistake. Waste is always situated and infused with a vast range of value judgments that are at once ethical, spiritual, political, and economic as well. Thus, as much as waste may seem mundane, it is also of cosmological significance.

Types of waste

There are several broad categories of waste generally recognized: industrial waste, municipal (or household) waste, and hazardous or special wastes such as radioactive and biohazardous waste notable for their particularly problematic risks. These broad categories of waste can each take solid, liquid, and gaseous forms which have their own attendant management approaches and infrastructure.

Broadly speaking, industrial waste refers to any discards resulting from heavy-duty manufacturing activities that cannot be reused and must be handled for disposal in some manner. Sometimes the term also includes waste arising from resource extraction and processing (e.g., mining and smelting). Municipal waste refers to the materials discarded as part of daily life by individuals and households. It is often synonymous with “garbage” and “trash.” Hazardous waste is any material deemed to pose a substantial risk of harm to people and environments. As might be expected, what counts as “substantial” is a source of considerable uncertainty and debate. “Special” wastes often refer to byproducts of nuclear power generation and of medical applications because of the severe risks for harm posed by the materials involved (e.g., radioactive waste or pathogenic material), the protocols for handling them as waste, and the management techniques used to mitigate their risks once placed in some form of disposal facility.

Specific definitions of each of these waste types vary between countries. For example, in Canada, the major categories of solid waste include oil sands tailings (645 million metric tons), mine waste rock (256 million metric tons), mine tailings (217 million metric tons), livestock manure (181 million metric tons), and municipal solid waste (34 million metric tons) (Statistics Canada 2012, 11). Thus asking how much waste is produced in Canada reveals absolute numbers, but it also reveals that what counts as solid waste in these figures includes waste from only some activities of primary production (e.g., mining and livestock agriculture) and excludes manufacturing altogether.

It also reveals that the portion of the waste stream that is typically understood as garbage (municipal solid waste, or MSW), makes up a mere 2.6% of the total. In the United States, similarly MSW makes up only about 2.1% of
disposed solid waste (MacBride 2012, 242), while in the European Union (EU), household MSW is 8.8% of waste totals. There are difficulties in making such side-by-side international comparisons because countries use different categorization schemes. However, these figures highlight that the majority (between 91% and 97%) of waste generated or disposed of in these regions is attributable to sources such as primary production, agriculture, and manufacturing, rather than what individual consumers and households throw in the trash bin.

Management strategies typically treat each type of waste as separate waste problems. Yet relying on such divisions has consequences for how waste is framed as a management problem. For example, when MSW data are tallied separately from industrial solid waste (ISW), vast amounts of waste arising from energy production, manufacturing, and distribution are left out of the picture. In this sense, distinguishing between ISW and MSW reinforces a false separation between industrial systems that produce commodities for individual consumption.

Per capita measurements used to make different populations comparable do their own work in terms of making certain versions of waste as a management problem visible in particular ways and not others. They do this by dividing waste arising over an entire population, and combining vastly different processes (e.g., personal consumption, byproducts of manufacturing, institutional, commercial, and construction activities) into an undifferentiated mass that is then divided equally among the population of a given territory (usually a city, region, or a country). One of the results is that individuals become the bearers of responsibility for masses of waste that they have little or no control over. At the same time, the idea that waste generation is an individualized postconsumer problem is reinforced while obfuscating the role of waste arising from raw material extraction, manufacturing, and distribution, processes that occur well before consumers even purchase commodities.

Waste generation grows as economies grow, and the prospects for separating waste generation and economic growth remain ambiguous (Fell, Cox, and Wilson 2010). In broad terms, international comparisons reveal that, as income increases, so does the amount of waste. The evidence suggests this relationship is stepped. In other words, instead of waste generation increasing linearly or exponentially with economic growth it increases in a series of rather abrupt jumps. Furthermore, the largest increases in waste generation occur in the shift from “low” to “medium” income or from “medium” to “high” income situations; “high” income countries, generally speaking, produce more overall waste than low or medium income countries do, but the rate at which waste generation increases is more rapid in low and medium income countries whose economies are growing. Also, generally speaking, as income level increases the composition of the waste stream changes: organic matter declines but metals, plastics, glass, and especially paper increase. As national income increases, there is a corresponding shift in waste management practices toward recycling and incineration. Future projections of waste generation indicate that it will increase most in low income and upper-middle income countries. As their wealth and populations increase, low income and middle income countries are respectively projected to experience an annual increase in per capita waste generation of more than 56% and 48% over 2010 levels by the year 2025. In contrast, annual per capita waste generation growth rates in high income countries will likely be below 10% of 2010 levels by the year 2025.

The designation of wastes as hazardous raises its own complexities. For example, many household wastes meet or exceed toxicity thresholds
for industrial hazardousness (e.g., paint, batteries, cleaning fluids), yet they are often excluded from hazardous waste lists for practical reasons – if all municipal household waste were treated as hazardous waste, it would vastly outstrip the financial capacity of municipalities to deal with it (Wynne 1987). Indeed, the thresholds used to define hazardous waste are themselves subject to discretionary decision-making. In the 1980s, for example, the United States Environmental Protection Agency (US EPA) had to decrease by tenfold its standard threshold for toxicity of waste lest almost all domestic sewage count as hazardous waste and massive costs to manage it would be incurred. This is an example of the effects of pragmatic management where definable thresholds of harm can be determined. However, in many cases such thresholds cannot be determined: for example, for endocrine disruptors such as bisphenol-A (BPA) in common plastics. With these chemicals any level of exposure has observable effects that can be considered a form of harm.

Waste management

In its contemporary forms, waste management refers to the governance and handling of solid, liquid, and gaseous materials externalized by some actor (e.g., a person, group, organization, or economic sector). The need to handle the discards of human activity is an ancient one, but waste management in the contemporary sense of the term emerged with the urbanization and industrialization of the nineteenth century. Even in Canada and the United States, contemporary norms of municipal waste management such as curbside pickup are of very recent origin.

The postwar period of industrialization in the United States and Europe witnessed profound shifts in economic production in industry and the household that in turn led to broad-scale shifts toward the acceptance by consumers of increasingly disposable products. One of the key indicators of this shift was the emergence of prepared preserved and frozen foods and their attendant packaging (see Figure 1). Paper, aluminum, and, later, plastic posed new challenges for waste management beyond their increase in overall and relative volume. Plastics in particular added material and chemical complexity to the waste stream, making such strategies as recycling and composting more complicated and costly.

The post-World War II period saw rising incomes with attendant changes in waste generation. Substantial changes in the waste management industry also ensued. Up until the 1960s, garbage handling in US and Canadian municipalities was done by a mishmash of public and private enterprises; cities were their own garbage markets and little integration existed across business segments (e.g., hauling versus disposing) or geographical territory. But postwar urbanization presented a serious waste management problem for cities as the growth of business and apartment buildings resulted in large quantities of waste in densely populated areas. The era also saw an increasing trend toward disposable packaging.

A key technological innovation that partially facilitated the shift from garbage business to waste management industry was containerized collection. In 1939, George R. Dempster patented what would become known as the “Dumpster” – a trade name often used interchangeably with the now familiar metal container with which municipal and other garbage is stored prior to collection (see Figure 2). Containerization of garbage collection meant a single driver could collect, empty, and return each container, thereby cutting collection costs by as much as 50%.
More critical examinations of waste management point out that most waste management strategies do not seriously challenge the underlying economics of growth within which they are situated. Treatment and disposal (T&D) of waste can in principle be part of an integrated waste management scheme that works to reduce the tonnage, toxicity, and heterogeneity of produced wastes, but in practice this is difficult (Liboiron 2013). Reliance on T&D infrastructure, especially when private in whole or in part, reduces or eliminates incentives to reduce waste put through the system. One only has to imagine the paradox of a private waste hauling firm actively finding ways to reduce the material throughput of the waste its customers contract it to haul to a landfill, or a privately owned landfill operation actively discouraging the use of its facilities by helping to reduce the amount of waste sent to it by its customers.

At issue here is that customers of T&D-based waste management systems are buying a service. The T&D operator is paid to provide that service and, in so doing, receives both money and the wastes to be dealt with. This arrangement can create a situation in which the customer has no knowledge of whether the waste has been handled properly after it has been collected. At the same time, T&D operators are confronted with the problem of a lack of control over the quality of the materials they handle: their customers are trying to rid themselves of materials that are of zero or negative value to them, so they have little incentive to incur additional

Figure 1  Composition of municipal solid waste, 1960–2012. Source: United States Environmental Protection Agency.
WASTE AND WASTE MANAGEMENT

costs by sorting those materials. Moreover, if and when technological or price changes emerge that increase the economic value of the materials T&D systems manage, the materials can shift from being waste to being valuable commodities and can thus become exempt from waste management regulations, despite no change in tonnage or toxicity (see Wynne 1987).

It might seem ideal to organize waste management schemes such that waste production is avoided altogether or the tonnage, toxicity, and heterogeneity of the wastes so produced are minimized and their costs internalized to those that manufacture the products that will eventually become waste. Legislation in the EU and elsewhere (e.g., Canada and Japan) has gone some distance to instituting extended producer responsibility and the “polluter pays” principle. But remainders of some sort will inevitably be produced, not all or even most of which may be reusable as inputs into other production processes. Consequently, ideals such as extended producer responsibility and the “polluter pays” principle – where manufacturers internalize the costs of waste management and pollution related to their products – often meet with limited success.

Critics also argue that current schemes such as curbside recycling are small compared to the scale of waste production by industrial manufacturers or by industrial consumers. For example,

Figure 2  Dumpster in Los Angeles. Source: commons.wikimedia.org.
since municipal solid waste comprises about 3% of total waste arisings in Canada and the United States, and less than 10% in Europe (see above), household recycling can achieve little in terms of overall waste reduction. Some argue an emphasis on recycling results in the “recycling trap,” a situation in which gains are not made from waste prevention and reduction because emphasis is placed mostly or entirely on recycling. When this happens, waste prevention or reduction strategies such as making products longer lasting, more repairable, and/or more modular (which would permit ease of replacing components, rather than whole products) fall by the wayside (Ackerman 1997; MacBride 2012).

More broadly, recycling rates can increase and the recycling of materials can become more efficient, but such changes do not necessarily lead to reduced waste generation, as a consequence of the so-called Jevons’s Paradox. The phenomenon is named after William Stanley Jevons, a nineteenth-century English economist. He observed how technological changes that increased the efficiency of coal burning led to increased coal consumption resulting from those very gains in efficiency – since less coal was needed to do the same work, it became cheaper to use coal, thus total demand for coal increased and outstripped the savings gained from efficiency. Similarly, even as contemporary industrial processes become less materially and energy intensive, aggregate growth in those processes typically outstrips gains from efficiency. Thus while each unit of economic activity might generate less waste, the overall amount of economic activity grows and so does the overall amount of waste.

Types of waste management

Contemporary waste management is typically framed using a hierarchy of options. In order of preference, these options include waste reduction, reuse, and recycling, with disposal (usually by landfill) being the least preferred. Each option includes a wide range of possible management techniques.

Landfilling

One of the oldest and most common forms of waste management is open dumping on land. Today, landfill is still the most common form of waste management in many countries, though considerable variation occurs (see Figure 3). A modern landfill, however, involves significant technological engineering and intervention and is a relatively recent occurrence. In 1965 the United States passed the Solid Waste Disposal Act which made open dumping illegal and mandated regulations and standards for landfills. In Canada, where solid waste is primarily a provincial jurisdiction, Ontario passed the Waste Management Act in 1970 with similar implications as the US legislation. Together these laws precipitated a “garbage crisis” in the sense of a manufactured scarcity by making landfills far more expensive, when they had previously been cheap and plentiful. Municipal governments, largely responsible for managing household waste, were suddenly confronted with legislation that drastically affected their budgets. Larger firms had spent years acquiring landfill space in the run-up to the passage of the US law. These larger firms were able to present themselves as a solution to the “garbage crisis” faced by municipal authorities which were unable or unwilling to finance the costs of new landfill facilities required by the new regulations. The business of handling waste, once fragmented between thousands of small operators typically serving single urban markets, increasingly became a vertically integrated and international industry of multibillion-dollar firms (see Table 1).

Though landfills might be thought of as final resting places for solid waste, they too produce
their own gaseous and liquid waste products in the form of landfill gas and leachate. Landfill gas is principally comprised of carbon dioxide and methane, but the material complexity of landfill constituents and their interactions with microbes and other conditions in the landfill can result in a wide range of other gaseous constituents as well. With appropriate technology at least some methane, a potent greenhouse gas, can be captured for use as an energy source in electricity generation.

Leachate is any liquid with dissolved and/or suspended substances that may cause environmental harm. Leachate is mostly formed from precipitation or groundwater sources that subsequently percolate through a landfill. Like landfill gas, leachate can be particularly problematic because of the material complexity of landfill constituents. For example, in the

![Figure 3](image-url) Management of municipal waste in the OECD, 2011. Source for data: OECD.
late 1980s approximately 7,000,000 chemicals were known to exist, of which 80,000 were used commercially, and 1,000 new chemicals entered commercial use each year. Yet, even if all available laboratory testing capacity on Earth were devoted full-time to testing the toxicity of those chemicals newly entering commercial use, only about half could have been tested (Wynne 1987). Today there are over 85,000,000 known organic and inorganic chemical substances potentially available for industrial uses. Thus, the chemistry of manufactured things that are eventually discarded and which may end up in landfills involves great uncertainty about the impacts of those chemicals in action with biological and geological processes long into the future.

It is commonly understood by waste engineers that even a well-designed landfill will need to be properly maintained in perpetuity to avoid all the possible risks associated with the biological and geological changes its contents will undergo.

**Air pollution abatement**

Air pollution abatement techniques involve neutralizing or mitigating the release and/or toxicity of gaseous waste byproducts of a variety of activities such as smelting, manufacturing, power generation, and transportation. Catalytic converters on cars, for example, reduce the toxicity of car exhaust through a chemical reaction. So-called scrubbers are used to control the particulate and/or chemical composition of exhaust from industrial processing and energy generating stations. Scrubbers work via wet or dry means. Wet scrubbers use water or other chemical solutions to remove particulate matter and thus reduce or eliminate the toxicity of exhaust effluent. Dry scrubbers use some form of material that can absorb acid gases and a filtering system to remove particulates.

Air pollution abatement techniques have their own waste generation problems. Catalytic converters, for example, rely on precious metals like gold, palladium, and platinum which must be mined. Those mining activities entail their own implications for waste generation (e.g., the production of waste rock or “overburden”) and waste management (e.g., for mitigating acid mine drainage). Scrubbers eliminate or reduce air pollution at the source, but their byproducts are solids and liquids that themselves must be

**Table 1** Top 10 waste management firms, 2014.

<table>
<thead>
<tr>
<th>Company name</th>
<th>Headquarters</th>
<th>Subsidiaries</th>
<th>Net sales (billions US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suez Environment</td>
<td>Paris, France</td>
<td>53</td>
<td>15.1</td>
</tr>
<tr>
<td>Waste Management</td>
<td>Houston, USA</td>
<td>497</td>
<td>13.6</td>
</tr>
<tr>
<td>Republic Services</td>
<td>Phoenix, USA</td>
<td>581</td>
<td>8.1</td>
</tr>
<tr>
<td>Rumpke Consolidated</td>
<td>Cincinnati, USA</td>
<td>8</td>
<td>2.2</td>
</tr>
<tr>
<td>Clean Harbors</td>
<td>Norwell, USA</td>
<td>111</td>
<td>2.1</td>
</tr>
<tr>
<td>Progressive Waste</td>
<td>Vaughn, Canada</td>
<td>37</td>
<td>1.8</td>
</tr>
<tr>
<td>Waste Connections</td>
<td>Woodlands, USA</td>
<td>138</td>
<td>1.6</td>
</tr>
<tr>
<td>Industrial Services of America</td>
<td>Louisville, USA</td>
<td>2</td>
<td>0.19</td>
</tr>
<tr>
<td>US Ecology</td>
<td>Boise, USA</td>
<td>13</td>
<td>0.17</td>
</tr>
<tr>
<td>Avalon Holdings</td>
<td>Warren, USA</td>
<td>10</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Source: Data from LexisNexis Academic.
treated and/or disposed of if they cannot be reused in some way. In this sense, air pollution abatement turns one kind of waste management problem – airborne pollution – into other kinds of waste management problems – solid or liquid pollution – and moves them around (e.g., from the exhaust source to a landfill). This kind of transformation and transposition challenge is a feature common to many forms of waste management.

**Waste water treatment**

Waste water treatment includes a vast range of applications. Like air pollution techniques, they all involve neutralizing or mitigating the effects of releasing water deemed to have harmful consequences for people and/or ecosystems. Probably the most important and familiar form of waste water treatment is that for sewage. Strategies to keep waste water separate from potable water have had major benefits for human health. Though often invisible, engineering for sewage has significantly impacted the built form of some of the most iconic urban settings. For example, in the mid-1850s the flow of the Chicago River was reversed and much of the city’s built landscape was raised by ten feet to accommodate a new sewage system.

Waste water management solutions developed in the nineteenth century were influenced by the miasmatic theory of disease (Melosi 2008). The theory held that noxious air from rotting organic matter was responsible for the outbreak of various epidemics sweeping burgeoning industrial cities. In England, reformers such as Edwin Chadwick were advocates of Jeremy Bentham’s utilitarian philosophy. Bentham believed that right action is that which delivers the most happiness for the most people. Utilitarianism in this form provided a philosophical basis for a variety of state-based interventions into waste management for public health, such as the separation of potable water sources from sewage. In turn, these major infrastructural projects entailed the organization of new administrative structures to govern hygiene and disease prevention through waste management. For example, the Metropolitan Sanitary Commission, created for Greater London in 1847, replaced the seven existing and separate sewer commissions with a single administrative body. Later, in 1875, the Public Health Act was a key event representing the first time that the British national government took responsibility for protection of its citizens’ health. Waste water management and state formation, then, are tightly linked (more broadly, see Goldstein 2013).

**Incineration**

Incineration involves the combustion of waste materials to reduce their mass and volume. Incineration schemes for managing industrial and urban waste streams emerged in the United Kingdom in the late 1860s. Known as “destructors,” they quickly became synonymous with poor air quality, but as early as the 1880s technical improvements led to higher incineration temperatures, which improved air quality and enabled the production of steam for energy generation. Today’s industrial incinerators involve high-temperature combustion and can significantly reduce both the mass and volume of waste material, thereby substantially reducing, but not eliminating, the need for landfills. They may also be especially important for mitigating the hazardousness of particular waste streams with pathogenic and/or toxicological risks since high temperatures can destroy pathogens and chemical bonds.

The use of incineration as a waste management approach is on the rise in Europe, where bans against landfills untreated solid waste have been
enacted. It has also been granted tax exempt status as a form of renewable energy generation by the US Environmental Protection Agency. Nevertheless, incineration is also highly controversial. Though the high temperatures of modern incinerators can destroy the chemical bonds of particularly toxic substances such as dioxins, those temperatures are reached using furnaces burning oil. Thus, the status of incineration as a renewable and sustainable form of energy generation is questioned by some. Also, though incinerators can break down chemicals such as dioxins, those same chemicals can reform from exhaust gases as they cool and are released into the air. For these and other reasons, new incineration plans often face resistance in the locale where they are to be sited. They are also increasingly resisted in transnational networks of activism (Davies 2005).

Recycling

Recycling describes a variety of practices that divert discards away from waste streams and toward various reuse applications. The diversion of materials is done using one or a combination of strategies including separation at source (e.g., by private citizens) or in a materials recovery facility (MRF) where employees sort recyclables from nonrecyclables. The term “recycle” first entered into use in the 1920s to describe the reuse of oil byproducts in the petroleum industry, and now recycling has become one of the most common forms of environmental behavior among households and individuals. Rates of individual and household participation in recycling climbed quickly in the United States, more than doubling between 1960 and 1990, but since then the rate of increase has stagnated. Currently, around 34% of all US municipal solid waste arisings are recycled but other countries claim higher (e.g., Germany, Austria, and Korea) and lower (e.g., New Zealand, Turkey, and Chile) rates (see Figure 3).

Understood in broader terms, recycling has a rich geohistory. For example, scavenging for potential value in waste has been an important practice throughout human history (Medina 2007). It would be a mistake to assume that scavenging, even in its contemporary forms, is solely a response to extreme poverty and inherently defined by exploitation. In many instances around the world scavengers who collect and return various materials to the production stream can regularly earn several times the legally mandated minimum wage, although the work is often precarious and risky. Many industrial processes, both historically and today, substantially depend on scavenging to recover what others have cast off as waste and return it to production (Minter 2013; Figure 4). For example, in 2011 the US scrap recycling industry processed 135 million metric tons of material, which is more than the amount of material recycled by US households. Scrap steel accounted for at least 83% of inputs to iron and steel foundries globally between 2008 and 2011. Globally, scrap metals comprised the inputs for production of more than 40% of copper, 35% of lead, 33% of aluminum, and 30% of zinc (Bureau of International Recycling, n.d.).

Recycling can also play more than an economic function during times of material scarcity. For example, the salvaging drives by civilians in the United Kingdom and the United States during World War II that were designed to supply material for those nations’ war efforts were as much about material recovery as they were about encouraging the morale of civilians on “the home front” (Cooper 2008; see Figure 5). Similar uses of recycling and its affective or emotional geographies underwrote communist nation-building efforts in Hungary and other Soviet satellites between 1948 and 1984 (Gille 2007).
Biodegradation

Biodegradation is the chemical breakdown of waste materials by micro-organisms under aerobic (i.e., in the presence of oxygen) and/or anaerobic (i.e., in the absence of oxygen) conditions. Familiar forms of biodegradation include composting and anaerobic digestion. Composting has been a key component of agricultural systems for centuries. Today it is an increasingly popular form of municipal waste management. Industrial-scale composting of organic municipal solid waste began to emerge in the 1980s, mostly in Western Europe. The European Commission claims, for example, that over 14% of per capita municipal waste generated in the EU was treated via composting and digestion in 2012. Anaerobic digestion has been a common part of sewage treatment processes since the nineteenth century. More recently it has been applied as a management approach for organic industrial and municipal solid waste. Both forms of biodegradation are useful on a number of fronts: they divert waste from landfills, saving space and mitigating the production of landfill gases and leachate. They can be added to soil to improve agriculture and home gardening; and they can have synergies with solid waste and waste water treatment systems when biodegradation is used to treat the organic fractions of these
two waste systems. One of the challenges facing biodegradation as a management practice is the difficulty of keeping out contaminants (e.g., plastics and chemicals) that reduce the efficiency of the biodegradation process itself and/or result in outputs with toxic characteristics, making them unsuitable for reuse in agriculture.

Waste trading

At its most general, waste trading is a commercial transaction involving the movement of waste arising in one place for treatment and/or disposal in another. The term is usually used to indicate movements that cross jurisdictional boundaries either within countries (i.e., as a form of domestic trade) or between countries (i.e., as a form of international trade). The waste trade can involve many steps of treatment that recover materials for reuse. It is not simply a matter of moving material to where it is cheapest to dispose of it, though that is an important aspect of waste trading. New York City, for example, treats around 5 billion liters of sewage per day. Once treated and dewatered, the remaining solids are sold to private firms, which trade them domestically and internationally for use as fertilizer.

Like other forms of waste management, the waste trade can be controversial. Indeed, a key international agreement known as the Basel Convention was a response to international concern raised over ocean dumping and the dumping of waste from “developed” countries in “developing” countries (Kummer Peiry 1995). The convention intends to halt transboundary shipments of hazardous waste from what it calls Annex VII countries (OECD, European Community, and Lichtenstein) to all other signatory countries, but various issues complicate the picture. First, that portion of the convention that seeks to halt Annex VII to non-Annex VII shipments (called the Ban Amendment) is not yet in force. Second, while Annex VII signatories would be prohibited from shipping hazardous waste to non-Annex VII signatories, nothing in the convention prohibits waste trading between countries within each category. Given that non-Annex VII territories include a very wide diversity of countries in terms of wealth and technological capacity for waste management (e.g., much of Asia and all of Africa), the convention is limited in its ability to stop waste trading from “rich” to “poor” countries.

International regulations like the Basel Convention are promulgated in an attempt to halt or reduce undesirable international trade in waste by making it illegal. Yet the very presence of national or international regulation on waste
treatment and disposal opens up the potential for profit to be made from skirting those laws and thus the potential for waste crime. Estimates suggest that organized crime syndicates earn US$20–30 billion annually from illicit trade in waste. Passing and enforcing stronger regulations may help to reduce the problem; however, any such efforts also have to overcome the substantial challenge of detection. The world’s major ports handle millions of containers per year and properly inspecting them for illegal waste shipments, even in a single port, is impossible.

Geographic research on waste and waste management

Geographic research on waste and waste management engages with a wide diversity of issues, reflecting the discipline’s breadth (see Moore 2012 for an excellent review). Among the most important themes is environmental justice (EJ), a broad concept that captures the idea that every person has an equal right to a safe and healthy environment. Research in this vein examines whether there are consistent patterns in the spatial distribution of a variety of activities related to waste and waste management, such as trading, treatment, storage, and disposal, that vary systematically with sociodemographic variables such as race and class. Results have been very mixed. Some studies contend that cause–effect relationships between the location of undesirable waste management activities and the location of particular sociodemographic groups are not consistent and cannot be clearly linked to systematic prejudice (Cutter 1995). One key complicating factor is that widely different results can be obtained merely by changing the geographical scale used in the data analysis. This involves the modifiable area unit problem (MAUP) which arises from the aggregation of data from one scale (e.g., households) to another (e.g., census tracts) and can make it more difficult to identify the effects on different economic or racial groups. Another complicating factor is the broader issue of problem definition: for example, whether inequality can be understood to have a generally applicable geographic unit of analysis, or even why hazardous waste management problems should inherently be specified as problems about locating facilities rather than problems derived from capitalist accumulation and externalization strategies (Lake and Disch 1992).

A second key area of geographic research on waste and waste management focuses on various aspects of both individual and collective environmental behavior and its governance. Briefly, environmental behavior is action construed by individuals and groups to be environmentally beneficial (e.g., recycling) while governance in this context refers to the regulation of that behavior through the daily practice of norms and/or formal governmental policy. A key finding from this kind of research is that a diverse set of environmental values, along with situational and psychological variables, influences or motivates actual environmental behaviors (e.g., recycling, waste minimization, and reuse). These findings are notable for their importance for governance since they suggest the possibility of, and need for, a variety of subtle policy approaches that might be employed to nudge individual and collective behavior toward more environmentally beneficial ends (Barr, Gilg, and Ford 2001).

An important focus of the environmental behavior research is recycling and waste minimization. Like other areas of geographic research about waste and waste management, the work on recycling and waste minimization is highly diverse. One stream of such research attempts to understand the underlying factors that might account for the observable gap between people’s stated support for environmental action and their
actual behavior. Recycling has become widely accepted as a positive environmental action, but the degree to which individuals and households actually undertake it strongly depends on the ease and convenience with which it can be done.

Related to this stream is research documenting links between recycling, citizenship, and identity. Research along these lines tends to examine recycling as part of a gamut of practices that perform personal and civic identities deemed to be ethically good. Here, practices may or may not be specifically about attitudes toward nature or the environment, but may instead be understood as part of broader conceptions of being a good person or citizen. An associated area of research takes recycling and community-based waste management strategies, particularly but not exclusively in the Global South, as examples of paths toward greater socioeconomic inclusion and justice (and, as such, it resonates with environmental justice research). This line of research challenges the received wisdom that modern consumer societies are thoughtless throwaway societies. Indeed, people engage in a wide variety of practices that route discarded items away from the waste stream, even if achieving environmental ends is not necessarily the point. These practices include hand-me-down and hand-around relations of exchange for things such as clothes (particularly for children) and other household items. These practices typically occur outside formal recycling systems and may or may not be explicitly framed as environmental behaviors by those who engage in them, but may instead be about generating and maintaining relations of care, community, and the like. Indeed, some studies have found procedures that route materials away from the waste stream to be a widely practiced norm (rather than an exception) for most classes of discards except food and packaging.

Research on recycling and community-based waste management practices in the Global South document the key role such practices play as livelihood strategies and paths out of poverty. It also challenges assumptions that the lives of those who engage in scavenging for recycling are defined by marginalization and victimization. This research area also documents the tangible environmental benefits achieved by the work performed by people in these circumstances, which are often informal and undervalued (Gutberlet 2013).

A third area where geographic investigations of waste and waste management have offered novel insights is industrial ecology. Industrial ecology describes an interdisciplinary approach to understanding the industrial economy, in whole or in part, as networks of material and energy flows in which the byproducts of one process are incorporated as resource inputs into another, thus generating zero waste. Practitioners in this field are typically devoted to finding technological solutions to maximize the efficient use of those material and energy flows underpinning industrial activity, while simultaneously achieving goals of sustainable development. The main contributions by geographers to the industrial ecology literature have been critical assessments of its history and of the very large gaps between industrial ecology ideals and their actual implementation (Gibbs and Deutz 2005). For example, industrial agglomerations of Victorian England exhibited features of what would today be called industrial ecology. Such agglomerations were fueled by coal, residuals from which included coke, ammonia-rich water, and coal tar. Victorian era blast furnaces producing iron left behind slag at a ratio of 2:1. Initially managed as forms of unwanted discard, commercial uses were eventually found for each of these materials (Desrochers 2009). Carbon-rich coke powered railway engines;
ammonia-rich residual water became a basic input to the pharmaceutical, dyeing, printing, and fertilizing industries; coal tar was found to be useful as a wood and metal preservative and as an intermediate input to rubber production and dye making; and blast furnace slag could be used in copper smelting, road building, building materials, and steel production. Indeed, the commercial uses of slag turned the massive accumulations of it on the British countryside (the “slag heaps”) into a lucrative export.

Geographers have also challenged as flawed abstractions the tropes and metaphors of nature, ecology, and natural systems that typically underpin the intellectual and policy justifications for industrial ecology. One problem underlying the actual implementation of industrial ecology is that the costs of standard forms of waste management are minimal relative to overall business operating costs and thus do not provide enough of an incentive to change; for example, by co-locating and/or sharing inputs and outputs with other businesses. Also, generating the interfirm trust necessary for taking on the financial risks of industrial ecology projects is difficult to instigate through policy. Where relatively successful industrial ecology projects have occurred, it seems that initial gains were made via firms willing to participate in initiatives that were economically low risk, but with environmental benefits (e.g., shared waste water treatment). Cooperation developed during these low risk initiatives then forms a basis for riskier and more substantive interfirm cooperation to achieve broader industrial ecology goals. In contrast, in the Global South, industrial ecology agglomerations manage flows of waste sourced globally rather than locally and exist in circumstances where waste management costs are minimal (e.g., ship breaking in Bangladesh, see Gregson et al. 2012). These examples strongly contrast with the metaphors and tropes about nature and ecology commonly associated with industrial ecology projects in the Global North, such as being “clean and green.” By comparison, recovering secondary materials from waste in the Global South, such as ship breaking and electronics recycling, can be quite dirty and messy.

A fourth and more recent theme of geographic research on waste and waste management focuses on tracing the “afterlives” of particular classes of commodities, such as clothes, cars, electronics, food, and container ships cast off or discarded as waste (Alexander and Reno 2012). Following the trajectories of such things, geographers have analyzed patterns of exchange premised on the rekindling of value out of waste. The empirical results have also been used to push and challenge major theoretical concepts and approaches that cut across a variety of substantive areas of research in geography and beyond. For example, global production networks (GPNs) and global value chains (GVCs), two major theoretical approaches in economic geography and economic sociology respectively, typically trace value creation from the point of manufacture to the point of sale. But studies of the trade in discarded clothing, cars, electronics, and ships have demonstrated that the usual conceptualization of value in these approaches needs to be substantially modified to account for additional rounds of value capture and creation that occur after the final sale of new commodities. These studies also suggest differences in governance between the coordination found in global recycling networks and the firm-driven commodity chains for products consumed in the Global North.

In a different vein, studies of the afterlives of food shift the analytical focus toward infrastructure and away from individual attitudes, behaviors, and awareness (Evans 2012). Thus, instead of food waste being a problem stemming from lack of awareness that could be ameliorated by informational campaigns, researchers document food
Waste and waste management raise far-reaching and fundamental questions that have ethical dimensions, such as what is a good or right action with respect to waste and waste management (Hawkins 2006). Such questions involve both spatial and temporal dimensions. Spatially, questions about good or right action might involve investigating concerns for justice around siting decisions for waste treatment and disposal facilities; or they might examine the ethical implications of the mobility of waste for treatment (e.g., recycling) or between source(s) and sink(s) (e.g., persistent organic pollutants into food webs). Temporal dimensions focus on the ethical questions raised by the implications for future generations of waste management strategies being practiced today. Such temporal questions are at first glance most acute with respect to nuclear waste. For example, the Waste Isolation Pilot Plant at Yucca Mountain, New Mexico, was approved under regulations that mandated a 10,000-year time horizon. More common forms of waste and waste management, such as the landfilling of household waste, raise similar issues. This is because landfills agglomerate such a wide range of organic and inorganic matter and lifeforms (especially bacteria) in combination with biophysical and geophysical processes that the potential consequences for future human and/or other forms of life is essentially unknowable and necessitates long-term management (Hird 2013).

Two broad approaches to the spatial and temporal aspects of waste ethics exist. The first proposes to avoid waste altogether and is exemplified by the zero waste movement. Zero waste is an ideal that aims to prevent waste from being generated in the first place. It derives its intellectual and practical inspiration from conceptualizations of cycles of energy and material in natural systems. It focuses on upstream practices such as design for durability, repairability, and reusability; it also advocates for lifestyle changes that reduce consumption. As such, achieving zero waste would entail fundamental shifts in the production and distribution of commodities. Substantive policy proposals that would achieve such ideals already exist. Two
important examples are steady-state economics and degrowth. Steady-state economics entails matching population, consumption, and carrying capacity. Degrowth entails equitably downscaling production and consumption while supplanting efficiency with sufficiency to increase human wellbeing and ecological conditions.

A second approach to the spatial and temporal aspects of waste ethics claims that avoiding waste altogether is unrealistic, thus the solution is to “waste well” (Lynch 1990). This approach to waste ethics also draws conceptual and practical inspiration from understandings of nature and is compatible with the ideals of zero waste. Wasting well would entail designing commodities for durability, modularity, ease of reuse, repair and recycling, ease of disassembly, and benign disposability (e.g., by mandating “green” chemistry, which designs toxicity out of products before they are put on the market, as an industrial norm).

SEE ALSO: Consumption; Deindustrialization; Environment and waste; Externalization; Geography and the study of human–environment relations; Industrialization; Nature; Wicked problems

References


WASTE AND WASTE MANAGEMENT


Further reading

The water system is highly dependent on the climate system. As precipitation and air temperature are the main inputs of the water system, changes in the spatial and temporal distribution of precipitation and air temperature will change how much water goes back to the atmosphere and how much remains in the soil, and thus be available for runoff. Rising air temperature increases evapotranspiration over water bodies and land surfaces, which will further reduce the water level of lakes and streams, soil moisture, groundwater flow, and thus streamflow. Even without changes in precipitation, warming can alter the whole water cycle, thus lowering runoff in dry, warm months. Changes in the volume of water in different reservoirs within the cycle have significant impacts on water-dependent industries and ecosystems. Indeed, climate change affects many different aspects of water resources, and the impacts will vary over space and time, reflecting the affected region’s land use and management conditions. Hence, climate adaptation strategies in the water sector should consider each location’s unique social and environmental conditions (see Figure 1).

Climate change impacts on water quantity

Water supply is projected to change as the spatial and temporal distribution of precipitation shifts. However, the impacts of climate change on regional water supply depend on the response of each basin to the direction of climate change. The sensitivity of each basin is largely a function of the location of the basin that influences the climate regime, basin soils and geology, land cover, and basin morphometric characteristics such as shape and slope.

Global hydrologic model simulation results show that climate change is likely to amplify the existing uneven distribution of water supply, both spatially and temporally. In humid temperate climates in mid-latitudes, where the basin runoff process is largely controlled by snow accumulation and subsequent melt, rising air temperature is projected to increase spring streamflow but reduce subsequent months’ streamflow as snowpack reduces in the late spring and summer months (Chang et al. 2012). In the equatorial wet tropics, annual streamflow is projected to increase in general, while in the dry tropics and in the Mediterranean climate, it is projected to decline. These changes are associated with change in the precipitation amount in these areas as the wet tropics are expected to receive more precipitation overall, which
can cancel the effort of evaporative loss due to rising air temperature. In glacier-fed rivers such as those in the Himalayas, the Andes, and the Sierra Nevada, warming can temporarily increase summer flow in the next few decades, but it will eventually decline as glacier ice contribution diminishes (Cisneros et al. 2014).

Changes in precipitation seasonality will also increase the temporal variations of streamflow in mid-latitudes. Seasonality of flow is likely to be amplified with drier summers and wetter winters. In humid marine west coast climates, the wet season flow is projected to increase, while the dry season flow is projected to decrease further. With increased evapotranspiration and reduced soil moisture, the recharge rate will decline, reducing groundwater supply. While there is high uncertainty in estimating the potential changes in groundwater availability from climate change due to groundwater’s long residence time, groundwater-dependent regions, typically already water-stressed in drier climates, are vulnerable to changes in the timing and magnitude of precipitation (Georgakakos et al. 2014).

**Climate change impacts on water quality**

Changes in air temperature and streamflow characteristics will affect the physical, chemical,
and biological aspects of water quality. Stream temperature is particularly sensitive to such changes in flow amount and timing. Together with warmer air temperature, decreased summer flow will raise stream temperature, reducing dissolved oxygen content in the water, thus negatively affecting the survivability of cold-water fish species such as salmon and trout. In lakes and reservoirs, as hot days extend, thermal stratification increases, lowering the mixing of top warm waters with cold bottom waters, which results in the excessive growth of algae in the top layers, which in turn depletes oxygen and kills fish (Georgakakos et al. 2014). When higher water temperatures are combined with calm and warm dry spells, it can increase the formation of blue-green algae (Arnell 1996). However, groundwater-fed streams or lakes may be less affected by climate change because summer baseflow is likely to be maintained, even though some reductions of the absolute amount of summer flow are inevitable.

Increased seasonality resulting from increasing streamflow variability will also increase sediment and nutrient loads during high-flow seasons (Praskievicz and Chang 2011). As channel erosion increases, sediment will be deposited in areas further downstream. During the low-flow season, reduced flow can, together with rising air temperature, increase pollutant concentrations because a higher water temperature can precipitate the chemical reactions of these pollutants. After a prolonged dry period, a small rainfall or snowmelt can transport a large amount of nutrients that have accumulated in the soil (Whitehead et al. 2009). Wastewater treatments may need to be upgraded to cope with such climate change-induced alterations in water quality. As some climate models predict higher tree mortality and more frequent wildfires, additional nutrients or sediments may be delivered to streams after forest fires. Changes in flood frequencies and magnitudes resulting from climate change will increase the capacity of streams to carry large sediments, exponentially increasing sediment loads during high-flow events.

Climate change impacts on droughts and floods

Climate change not only means changes in flow but also implies more frequent droughts and floods. Climate models predict increases in the number of dry days while the intensity of precipitation increases in mid-latitude humid climates. Such changes in precipitation patterns will bring more frequent urban flash floods in already developed areas that have high amount of impervious surface areas. As urban development continues in the coming decades, the magnitude of urban flash flooding is likely to increase. In large watersheds, riverine flooding, which occurs when channel capacity can not hold the excess water and thus spills over to the river bank and adjacent low-lying areas, is also likely to increase when prolonged wet periods saturate soils (Georgakakos et al. 2014). The 1993 Mississippi River flood exemplifies such a case. As the intensity of tropical storms is likely to increase under warming scenarios, coastal areas are prone to flooding that results from storm surges. Cities located at the mouth of large river basins will be exposed to more frequent coastal flooding as sea level rises and storms bring inland as well as coastal flooding. Many infrastructures such as power plants, roads, drinking water, and wastewater treatment plants, which are located near rivers or coastal areas, are particularly vulnerable to flooding.

While extended extreme droughts (multiyear droughts) may not necessarily increase in humid temperate climates, short-term droughts (3 or 6 months) are projected to increase under some
climate change scenarios (Cisneros et al. 2014). Even without changes in annual flow, changes in the timing and magnitude of seasonal flow will exacerbate summer droughts in some areas, particularly where the gap between water supply and demand is already high. In arid climates, extreme droughts are more likely to occur under a warming climate, as evidenced by the American southwest and southeast in the past decade. The social impacts of such prolonged droughts can be severe in areas where alternative drought mitigation strategies are not available.

Climate change impacts on water demand

Rising air temperatures will also increase municipal and irrigational water demands. Historical analysis of some large municipal water providers shows that daily water production is highly positively associated with daily maximum temperature and negatively related to antecedent precipitation amounts (Miller and Yates 2006). The effects of this climate variability and change on water demand will vary by region. The sensitivity of water consumption to warming is higher in arid climates than in humid temperate climates (Breyer, Chang, and Parandvash 2012). Outdoor water use in these hot arid climates will increase further as urban residents attempt to reduce the temperature by applying more irrigation water. As the temperature rises, soils get drier earlier in the growing season, so crop water needs to be increased as well. Thus, irrigation water demand will rise, although plants may use water more efficiently to minimize transpiration under harsh climates. Water withdrawals for cooling power plants can be affected, as warmer waters can reduce the efficiency of cooling. With an expected decline in summer water supply, an increase in competition between different water users is likely in many river basins that are already stressed by other nonclimatic factors such as ongoing development and population growth.

Climate change impacts on water-related ecosystem services

Changes in water quantity and quality and in the frequency of floods and droughts will affect water-dependent ecosystems and thus water-related ecosystem services, the benefits that people obtain from nature. Changes in the timing and magnitude of flow will shift water-provisioning ecosystem services, such as drinking water supply and hydropower generation. Additionally, regulating ecosystem services will be affected. Changes in sediment and nutrient loads or concentrations will increase the cost of water treatment for drinking or irrigation. Increased sediment supply from upstream may fill in reservoirs or dams quickly, adding to the cost of dredging and reducing the lifespan of dams. During flood events, water-borne disease may spread over wider geographical areas quickly, adding to the cost of public health care. During the low-flow season, the increased concentration of sediment can block navigation or irrigation channels, adding to the cost of removing silts. Diminished groundwater supply will increase the cost of pumping for irrigation and desiccate wetlands, limiting wetland functions and services. Particularly in areas where hydrologic and thermal regimes shift abruptly, increased stream temperature will negatively affect aquatic species and thus diversity because of the limited time allowed for species to adapt to such rapid change. In those areas, invasive species may take advantage of the changing climate and replace native species. Other water-dependent cultural ecosystem services such as skiing, sport fishing, kayaking, and waterscape aesthetics may
degrade as snow, water availability, and water quality decline. While there is high uncertainty in estimating the value of these water-related ecosystem services, as society’s demand for these services increases in the future, their value is also likely to increase.

Possible climate change adaptation strategies in water resources

While climate change is occurring at a global scale, the impacts of climate change will vary by region and will be realized differently across different scales. Therefore, effective climate adaptation strategies should consider the regional and scalar differences of each region’s environmental and social systems. A wide combination of structural and nonstructural policy may help relieve the potentially negative consequences of climate change in the water resource sector. New dam operation rules can minimize the disruption of hydropower generation while maintaining environmental flow and preventing downstream floods. Aquifer recovery storage can draw excessive winter water into the groundwater system and pump out during the low-flow season for meeting the needs of irrigation water. Best management practices in farmlands can reduce nonpoint source pollution, while urban green infrastructure can retain more storm runoff and nutrient or sediments during storm events. Smart land-use planning and landscaping practices in urban areas can reduce residential outdoor water consumption in the summer. Early flood or drought warning systems can prepare communities to cope with extreme hydrologic hazards. Some portions of floodplains can be restored by planting trees, reducing peak flow in downstream areas while providing habitat in stream bank areas. Floating homes or amphibian houses can be more proactive adaptation strategies in coastal areas where both sea level and storm surges are projected to increase in a warming world. For these adaptation strategies to be successful, concerted coordination across different sectors at different levels is essential.

SEE ALSO: Climate adaptation/mitigation; Climate change adaptation and social transformation; Climate change, concept of; Climate and societal impacts; Ecosystem services; Global climate change; Hydrologic cycle; Water conservation; Water quality; Water resources and hydrological management

References

Miller, K., and D. Yates. 2006. *Climate Change and Water Resources: A Primer for Municipal Water*
**WATER AND CLIMATE CHANGE**


Water and human rights

Olen Paul Matthews  
*University of New Mexico, USA*

Drinking water and water for personal sanitation are essential for human life, health, and dignity. Access to water for these purposes is the focus of an evolving international human right. In 2010 the United Nations (UN) General Assembly passed a resolution recognizing water as a human right (United Nations 2010). The vote on this resolution was 122 in favor with 41 abstentions. The United States, Canada, and some members of the European Union were among those abstaining. The abstentions occurred because the full parameters of this right are still being developed, and the impacts are uncertain.

The need for a specific human right to water has evolved for several reasons. In the various declarations of human rights made since 1945, water has not always been explicitly singled out, though it is mentioned in some conventions. An explicit human right would encourage investment in water and sanitation services. Access to water is a global problem with 768 million people lacking safe drinking water and 2.5 billion without adequate sanitation facilities (WHO 2013). Framing water as a human right recognizes water’s importance to human dignity and the need for equal access. If water is a human right, then investments will be made equitably. Another reason for the evolution is the backlash against private service providers that propelled a series of confrontations starting in the 1990s. The protestors asserted that *water is a human right, not a commodity*. This anti-privatization movement has been very vocal in bringing a human right to water to global attention.

A human right to water is often a polarizing issue. This is partly because the *right* is associated with specific local conflicts. Even though the conflicts that focus attention on the issue are local, a human right to water is an international right. Human rights must be accepted by the international community, and this often requires compromise.

### Water as a human right in context

The issue of access to drinking water has deep roots, especially in arid areas (Salzman 2006). In traditional Jewish water law, surface waters were common property (open access) available to anyone for drinking water purposes (see also Water rights). Wells on the other hand were the common property of the community (limited access common property). Although everyone had a right to drink before water was used for agricultural purposes, community members had a preference over outsiders if there was an extreme shortage. Islamic traditions also give drinking water a preference over other uses, with drinking water available to everyone. Australian Aboriginal traditions required outsiders to *ask first* for drinking water, with permission usually granted. Other parts of the world have similar traditions. Although the provision of drinking water as a human *need* has a strong tradition, treating it as a *commodity* also has a long history.
WATER AND HUMAN RIGHTS

The growth of large urban areas created water supply problems which local wells and surface sources could not overcome. Solutions were often expensive and a variety of public and private (sometimes both) supply systems developed depending on circumstances. In ancient Rome citizens were provided free water in a variety of public basins, but almost half the total supply went to private destinations where a fee was paid (Salzman 2006). By the late 1500s London licensed a water delivery system that delivered water for a fee. The system was built mostly with private money. By the late 1700s cities in the United States were sufficiently large to require more sophisticated delivery systems. Philadelphia, New York City, Boston, Chicago, and Cincinnati developed a variety of fee-based water supply systems. Some were public and others were private (Armstrong, Robinson, and Hoy 1976). Today, urban water delivery supported by user fees is the norm, with both public and private systems supplying water.

Many parts of the world have difficulty funding the kinds of water supply and sanitation systems needed to meet health standards. International sources can provide the needed capital, but this often comes with strings attached. By the 1980s the International Monetary Fund and the World Bank had adopted neoliberal policies promoting the privatization of public services such as water utilities. Some of the private firms, after investing in water supply and sanitation systems, raised the prices charged to a level beyond the consumer’s ability to pay. In Bolivia and Argentina protestors refused to pay their bills and took to the streets. Eventually, the private firms were ousted. The protestors claimed a human right to water under international law. But their anti-privatization and anti-commodification rallying cries do not represent current trends in international law. A human right to water is neutral over public or private delivery systems (Murthy 2013). A description of international law will help in understanding the current situation.

International law

International law, and in particular that associated with human rights, differs from the law that governs nations. In general, international law governs the relationships between States (countries). International human rights law goes one step further, creating an obligation governments have to their own citizens. Human rights bring individuals into international law. Because States are sovereign, international law is largely (not completely) consensual. This means States must agree to be bound by the international law in question. Without this consent enforcement is problematic.

International law comes from four sources. Not all are of equal importance and some are not recognized by all States (countries). Conventions and customary international law receive the greatest amount of acceptance. The other two, general principals and secondary sources such as the writings of legal scholars, have persuasive value in some instances. Only the first two will be discussed here. Conventions are written agreements (treaties and other international instruments) where the parties have expressly consented to be bound by the terms. A treaty can be between two or more parties, and only those parties are bound. In contrast, customary international law is based on the common practice of States (what they actually do), and the belief they are bound to behave that way by international law (opinio juris). Consent is implied in this instance. If a State behaves contrary to what some consider customary law, it shows lack of consent. Acquiescence, doing nothing to object, can be used to infer consent. Thus when a right is claimed under international law, States frequently object in order to show they do not consent.
Statutory law binding all States is not common internationally, although the European Parliament creates binding statutory law for member states. The UN Security Council does have power to create binding law, but each permanent member has veto power. A UN resolution may or may not be binding and certainly would not be on States voting against such resolutions. The language of a particular resolution controls, and many are considered soft law since they do not have language forcing any action. Declarations or resolutions containing nothing more than an aspirational goal or a statement of principle are not hard law (enforceable law). Even if resolutions are soft law, they can show a trend toward an emerging customary international law.

A common way for customary international law to emerge is through the use of UN conventions. A UN convention, unlike a treaty between two States, is meant to be agreed to by a large number of countries, at which point it shows evidence of customary international law. For example, through a lengthy process of negotiation a draft convention is created. This draft is reviewed by UN member States, and those that agree will sign and ratify the convention. After a specified number of States have ratified a convention it is considered to be in effect, although it may not be binding on all. This is typical for the human rights conventions that implicitly recognize a human right to water.

Human rights

Human rights, as a legally binding obligation States have toward their citizens, date from the end of World War II. Although the concept of natural law has been around much longer, states were not bound to any particular standard of behavior. Natural law recognizes that all people are entitled to specific rights and freedoms regardless of man-made laws. Natural law can include a right to self-determination, life, liberty, property, or the pursuit of happiness, depending on the philosophical interpretation. The US Constitution’s Bill of Rights recognizes many of these natural rights and others. In 1948 the UN adopted the Universal Declaration of Human Rights (UDHR), which includes 30 articles defining fundamental freedoms and liberties. Although this is not a binding treaty, the signatory States (all members of the UN) pledged themselves to achieve this aspirational goal. As such, the document presumptively has legal effect whether it is enforceable or not. A right to a standard of living adequate for health purposes is included. This right can only be achieved with access to water, which by implication is included in the right.

Subsequent to the UDHR, several other UN human rights conventions addressed specific aspects of human rights with water either implicitly included or with some explicit reference to water. The International Covenant on Economic, Social, and Cultural Rights (1966) declares that people cannot be deprived of means of subsistence. Water is implicitly included in a right to subsistence. The Convention on the Elimination of All Forms of Discrimination against Women (1979) recognized a right to adequate living conditions, including access to water and sanitation. The Convention on the Rights of a Child (1989) recognizes a right to clean water to protect children’s health. The Convention on the Rights of Persons with Disabilities (2006) assures equal access to clean water. All but the last convention have entered into effect because a sufficient number of countries have ratified them. This does not mean they are universally binding and enforceable, although they may represent aspirations for what customary international law should include. For example, the United States
WATER AND HUMAN RIGHTS

has not ratified any of these conventions. Two problems exist. One is with the uncertainty created by the convention language, and the other is with different importance States assign to the meaning of a human right.

Not all States mean the same thing when they use the term human right. One position looks on human rights as an acknowledgement of an important human need that ought to be addressed. Another views human rights as a legal entitlement obligating a government to act, even though it means a reallocation of scarce resources (Donoho 2012). There is a significant difference between a need and a legal entitlement. A legal entitlement gives an individual a mechanism for enforcing the right against the government. Needs are aspirational, as reflected in most human rights conventions where the government obligation is limited to available resources or all appropriate means. This language allows governments to interpret the convention as a goal rather than an entitlement.

Current status as customary international law

The 2010 UN resolution expressly declares water as a human right. The resolution is nonbinding but is an expression of high aspiration. No treaty or convention expressly creates this human right. The real issue is whether a human right to water has reached the point where it is customary international law (Cavallo 2012). Commentators argue either that customary law status has already been achieved or that it is in process. The main issue is whether the right will be defined in terms of a need or an entitlement. Enforcement changes depending on the approach (Beail-Farkas 2013) as do obligations between States. The parameters of customary international law are determined by examining state practice (how States actually behave). State practice is used to establish opinion juris (a State’s belief that it has an obligation to follow the law) and requires consistent and uniform behavior.

State practice can be determined by both State behavior and the actions of international bodies. For example, ratification or failure to ratify the human rights conventions discussed above shows an individual State’s position with regard to state practice. The reasons for non-ratification or the exceptions made as part of the ratification process are important in determining the contours of an individual state’s acceptance or denial. Also, state practice can be determined by examining the international soft law created by international bodies. International soft law includes resolutions by the UN Human Rights Council and other UN documents, statements from international conferences, and regional or national human rights documents. All these together provide evidence of state practice. At present, the lack of uniformity and consistency makes it difficult to determine the parameters of the human right to water. Several examples will be used to illustrate the complexity in establishing customary international law.

The right to safe drinking water and sanitation has been recognized expressly or implicitly in regional and national legal instruments. Regional treaties in Asia, Africa, America, Europe, and the Middle East are one example (Cavallo 2012). Typically, these are in regional human rights agreements and obligate a state to provide basic services, eliminate discrimination, and provide equitable access. A right to water is also being recognized in national constitutions. South Africa, Uganda, Argentina, and others have explicitly recognized the right while in India it is implicit in their constitution. Court decisions in South Africa have extended this to include a duty to provide free water under some circumstances (Bluemel 2004). California joined
this growing movement in 2012 by legislatively enacting a bill establishing a right to “safe, clean, affordable, and accessible water” sufficient for “human consumption, cooking, and sanitary purposes” (California Water Code 2013).

International conferences have generated a large number of declarations and resolutions addressing this issue. When these conferences have official State representation they reflect state practice and build the case for an evolving customary international law. Additional weight is added when the conference is sponsored by the UN. For example, the UN Water Conference in Mar del Plata (1977) recognized that all people have a right of access to drinking water sufficient for their basic needs. This was reaffirmed in Agenda 21 (1992), the Programme of Action of Cairo (1994), the Istanbul Declaration of Human Settlements (1996), the Johannesburg Declaration (2002), and most recently the Rio + 20 Conference (2012).

The UN has made many declarations on drinking water and sanitation (Cavallo 2012), but the most recent is an explicit recognition of a human right. In 2010 the UN General Assembly recognized “the right to safe and clean drinking water and sanitation as a human right that is essential for the full enjoyment of life and all human rights.” The resolution recognizes the body of soft law discussed above, and then calls upon the international community “to provide financial resources, capacity-building and technology transfer, through international assistance and cooperation” (United Nations 2010). Although 122 States voted for this resolution, 41 abstained. Some of the States voting for the resolution qualified their vote by saying it was subject to domestic legislation or to resource limitations (Cavallo 2012). Many of the states voting against the resolution felt it would interfere with the work by the Human Rights Council in Geneva. The Council is in the process of fleshing out the parameters of this human right, as will be discussed below.

Although the discussion above is not exhaustive of the number of statements made declaring a human right to water, it shows sufficient state practice to confirm the existence of the right. The elements of the right are also beginning to take shape mostly within the UN Human Rights Council. Any convention solidifying this right will probably be proposed by that source.

**Attributes of a human right to water**

The General Assembly resolution does not have much detail describing the elements of the human right. The details are most fully developed in General Comment 15 (Committee on Economic, Social, and Cultural Rights 2002) and Human Rights Council Resolution 15/9 (2010). The Committee on Economic, Social, and Cultural Rights and the Human Rights Council are part of the administrative structure of the UN. Although neither the resolution nor General Comment 15 is binding, together they foreshadow what might appear in a proposed international convention.

General Comment 15 defines freedoms and entitlements associated with the right. It outlines governmental obligations, defines what acts violate the right, and requires a process for enforcement. Freedoms include a right of access to existing supplies. Arbitrary disconnections, interference with existing delivery systems, and contamination violate the freedom of access. Entitlements include equal opportunity to enjoy water and water adequate for human dignity. Adequate water depends on the circumstances but includes a consideration of availability, quality, and accessibility. Availability includes sufficient water for drinking, personal sanitation, and domestic uses. Quality includes both
health-based standards and palatability requirements such as taste and odor. Physical, economic, nondiscriminatory, and information accessibility are all required. Physical accessibility is based on distance to the supply (immediate vicinity), quality of the facilities and services, and personal safety of the users. Economic accessibility requires that services and facilities be affordable to everyone. Nondiscriminatory accessibility promotes equality and prohibits discrimination for any reason. The list is extensive and includes sexual orientation, social status, HIV/AIDS (human immunodeficiency virus/acquired immune deficiency syndrome), and other discriminatory grounds. Special attention must be shown to groups who traditionally have had difficulties accessing water. Information accessibility includes a right to both receive and give information during water management decisions.

General Comment 15 creates government obligations designed to achieve the freedoms and entitlements. Two obligations are effective immediately: governments must guarantee the right can be exercised without discrimination, and steps must be taken to fulfill the right. The language describing implementation gives governments some flexibility. Implementation must be “deliberate, concrete, and targeted” toward achieving the goal. Governments have a duty that is “constant and continuing” and must “move as expeditiously and effectively as possible.” This language requires constant progress, but does not require governments to exceed their financial capability. Obligations to respect, protect, and fulfill are included. The respect obligation grants equal access to adequate water, recognizes customary and traditional water rights, and prohibits pollution from government facilities or actions. The obligation to protect requires governments to pass legislative measures that will prevent third parties from interfering with the right. This is aimed specifically at nongovernmental service providers, and requires legislation that guarantees equal access to safe and affordable water. This provision does not prevent the privatization of water services, but does provide consumers protection against inequities. The obligation to fulfill suggests a broad range of actions, including water planning, education on hygiene, and water conservation. Most importantly it addresses the issue of affordable water. Payment for water services must be equitable regardless of whether services are publicly or privately provided. Income supplements, low-cost and subsidized water rates, or even free water may be required depending on circumstances. It should be noted that subsidized or free water is a government obligation and not the obligation of a private service provider.

States are required to adopt a national water strategy in order to achieve the level of services required by the right. A process must be put in place that can monitor progress. States must take “necessary and feasible steps” in order to be in compliance. A distinction is made between inability and unwillingness. A State that is unwilling to use resources to achieve the right violates the requirements of General Comment 15. Resource constraints can make it impossible for a State to comply. This inability to comply requires justification, but States have discretion in determining the suitability of specific measures. Progress reports must be given to a UN committee who will evaluate whether goals are being achieved. Any person or group who is denied this right should have access to a legal remedy at the national and international levels. Legal remedies include compensation, restitution of prior conditions, and guarantees that the violation will not reoccur.

Obligations can extend beyond a country’s boundaries. Technical and financial assistance should be provided to other countries, especially by States that are economically developed.
The operative parts of the Human Rights Council Resolution 15/9 are less than two pages long. Several things are made clear by their repeated emphasis. While drinking water and sanitation services can be delegated to third parties, States are not exempted from their obligation to protect a human right to water. Non-State or third parties can provide services, but they must do so in a way that is transparent, nondiscriminatory, and accountable. Planning and implementation processes must be transparent and have meaningful stakeholder participation. A regulatory framework is needed to insure all providers fulfill the obligations, with regulatory institutions being given the ability to monitor and enforce the regulations. Non-State service providers are specifically singled out and required to engage stakeholders proactively and to actively incorporate human rights into their operations. Non-State providers must also provide a grievance process.

Clearly, States have the primary responsibility for protecting human rights, but third-party, non-State service providers are allowable. Although the terms “privatization” and “commodity” are not used, private service providers are acceptable. But States must regulate and enforce regulations to ensure non-State providers protect human rights and ensure stakeholder participation and protection.

**Consensus, ratification, and enforcement**

The above description provides insight on the current parameters of the international human right to water, but General Comment 15 and Human Rights Council Resolution 15/19 are not binding international law. Both were produced within the administrative structure of the UN. Neither of them is an agreement that can be ratified by individual States. At best they illustrate evolving customary international law. When the human right to water was specifically adopted by the General Assembly, the resolution recalled (acknowledged) General Comment 15 but not in the active part of the resolution. Acknowledgement (recalling) is not the same thing as a full discussion and the creation of a convention. Consensus on the shape of this human right has not yet been achieved. A draft convention is needed to stimulate discussion that in time can form a consensus. Once a proposed convention is in place, States must consent by ratifying it before a binding legal obligation is created. Even if a convention receives sufficient support to be in effect, human rights conventions have a weak record of enforcement.

Reaching a consensus may be difficult, as was shown by the 41 abstentions in the General Assembly vote. Many States, including the United States, felt the Resolution was premature because there were many uncertainties and a lack of consensus. Problem areas include: private service providers, obligations a State has to other States, philosophical differences States have in defining human rights, and international supervision. One reason for the 41 abstentions on the General Assembly Resolution (United Nations 2010) was the lack of clarity on the private sector role in service delivery. Anti-privatization and anti-commodification were rallying cries to bring world attention to a human right to water. From that perspective the private sector has no role in delivering water and sanitation services. Human Rights Council Resolution 15/19 (2010), which came out after the General Assembly Resolution (United Nations 2010), makes it clear that private sector service delivery is compatible with a human right to water. States have an obligation to ensure compliance regardless of the delivery mechanism. In fact many corporations are in favor of a human right to water because they view the investments this will require as a
business opportunity. In spite of the current UN position the debate is not a closed issue.

Inter-State obligations are unclear. General Comment 15 requires States that are more developed to provide financial and technical assistance to others. In addition, States should provide water to other countries to help them realize the right to water. Many countries are already involved in providing international assistance and are uncertain if more will be required. In voting in favor of the General Assembly Resolution (United Nations 2010) a few countries made an exception, stating the provision would not apply to them (Cavallo 2012). If this obligation survives into a final proposed convention, it will most likely be stated as an aspiration rather than an obligation. States are not likely to agree to provide water to another State. Such a reallocation could interfere with existing rights and have substantial costs (Donoho 2012).

States have substantially different philosophical approaches to human rights. Three examples are used with a realization that there are many nuances that fall in between. In one view a human right is an individual entitlement enforceable against the government with sanctions possible. The right is judiciable in an international or national forum, with real consequences including a reallocation of resources. From a different perspective, the right is an expression of human need best fulfilled by creating safety nets negotiated by a country’s political institutions. This approach uses available resources to progressively achieve specified goals. It allows the consideration of economic needs other than drinking water and sanitation. These two approaches are not completely compatible, but both are found in General Comment 15. The third approach views an endorsement of human rights as window dressing that can be ignored. Governments do not always mean what they say.

International supervision is also problematic. General Comment 15 envisions a process with States reporting progress to the UN where it can be evaluated. Enforcement by international judicial bodies is also proposed. Both these have elements that can be viewed as infringing on a State’s sovereignty.

Any proposed convention must reach some form of consensus on these items, making it unlikely that everything currently in General Comment 15 will be included. General Comment 15 was drafted by an unelected body without full debate. Ratification will be difficult in some countries, depending on the final convention language. Consensus and support for ratification will need to give States flexibility to protect their sovereignty. For example, the United States seldom consents to human rights conventions. A convention is a treaty which requires ratification in the US Senate. Even if the President signs the treaty/convention, consent cannot occur without Senate approval.

If a convention is negotiated and ratified by a sufficient number of countries to be in effect, enforcement is still a difficulty. Human rights conventions are notoriously weak on enforcement provisions, giving each State substantial freedom to make their own interpretation of a convention’s meaning. Convention terms tend to be malleable with substantial wiggle room. In addition, economic and social human rights are generally nonjudiciable, meaning individuals cannot get redress in a national or international forum. The weakness of international enforcement of human rights is well known even for States that have ratified conventions. But declaration of a human right has moral authority, and can be persuasive in influencing political processes and investments. Ultimately, if a State has the political will, enforcement will occur.

At the national level more progress has been made. The South Africa Constitution of 1996
recognizes a human right to water. Subsequent court decisions have upheld this right, leading to investments in infrastructure. In India a right to water and sanitation has been implied by the courts. Other States have made similar choices. The trend toward bottom-up national approaches will likely continue as negotiations on an international human right to water progress. Most commentators classify the right as soft international law or arguably customary international law. The real challenge is establishing the parameters of the right and reaching a consensus.

SEE ALSO: Geopolitics of the environment; Human rights; Waste and waste management; Water: drinking; Water rights

References


Further reading


Water banking

Sheryl Luzzadder-Beach
Jonathan M. Flood
University of Texas at Austin, USA

Water banking enhances the storage component of the water budget, to overcome times of water shortage, when human water use demand exceeds the available water supply. This is most critical in arid regions where there is high demand both for drinking water and for agricultural uses. Water banking is also necessary for reducing water withdrawal impacts on natural terrestrial and aquatic habitats. California is a prime example of a US state with numerous water banking programs. Other states with major water banking programs include Texas, Colorado, Idaho, Nevada, and Arizona.

Physical water banking programs

Physical water banking involves the actual transfer of water from one region to another, or capture of excess surface runoff into storage in a groundwater aquifer or surface reservoir. It may also involve retention of water from one season to use in a later season, to meet the restrictions of water law. For example, in Texas, California, Arizona, and Nevada, groundwater banking is practiced for a number of reasons (Maliva 2014; Megdal and Dillon 2015). It is a form of artificial aquifer recharge that stores water that may not be put to beneficial use until a later date. The water is stored in an aquifer in order to maintain a future right to withdraw the water (Maliva 2014; see Water rights). Water is introduced into the aquifer either by injection wells or by spreading the water over the surface and allowing infiltration and percolation to the water table (Maliva 2014). Banked aquifer water may also be treated to meet use quality criteria (Megdal and Dillon 2015). The physical benefit of this system is that water is stored in overdrawn freshwater aquifers, and, by adding water, the aquifer’s pore spaces are prevented from collapsing, which helps to avoid ground subsidence (see Aquifers). Another benefit is to even out the supply when water is seasonally scarce. There may be impacts on stream flow that is connected to the water table, so coordinating with surface water management is important, especially in times of drought. Because groundwater law can be different from surface water law even in the same jurisdictions, care must also be taken in regulating, measuring, accounting for, and monitoring groundwater banking activities (withdrawals and deposits), especially if surface water is part of the equation. Water banking is also used for ecological habitat restoration, as applied on the Snake River in Idaho, where instream flows are managed for endangered species (Green and O’Connor 2001).

Virtual water banking programs

Unlike physical water banking, virtual water banking programs involve trading water use credits or water withdrawal rights on a market. Water banking matches buyers and sellers, and follows the market value of water across and between regions (Maliva 2014). For example, urban and agricultural waters have different prices, values, markets, and desired qualities. Lepper (2006,
368) further differentiates water markets into exchanges, transfers, and groundwater “augmentation” programs. Transfers involve a change of ownership or the lease of water rights, whereas exchanges involve temporary trades or changing the timing of the use within the frame of water rights law, as practiced in the Arkansas Valley (Colorado) water banking program (Lepper 2006). Likewise, Southern California has a complicated system of conjunctive water use to follow demand and exchanges between surface water systems, reservoirs and groundwater systems, and multiple water districts (Pulido-Velasquez, Jenkins, and Lund 2004). This is conceptualized using economic engineering network flow models, since multiple uses and outcomes in geography and time are optimized. California sees urban water demands alone growing from 10.8 km$^3$ in 1995 to a forecasted 14.8 km$^3$ in 2020. Conjunctive uses in Southern California provide about 3.1 km$^3$ a year, in comparison to the urban demand (Pulido-Velasquez, Jenkins, and Lund 2004). Conjunctive use and water banking occurs in Southern Idaho, with the one of nation’s earliest water banking programs formed in the droughts of the 1930s and formalized by law in 1979 (Lepper 2006; Ghosh, Cobourn, and Elbakidze 2014). This system marketed water rights or entitlements by sale or lease (Lepper 2006, 367). Many orders of magnitude less than California’s urban water demands, the number of trades or exchanges in Idaho grew from less than 0.001 km$^3$ of water in 1995 to nearly 0.071 km$^3$ of water in 2013 (Ghosh, Cobourn, and Elbakidze 2014).

Water markets are also practiced internationally, as exemplified in Spain (Palomo-Hierro, Gómez-Limón, and Riesgo 2015). Spain’s water markets became legally possible over the last two decades, in response to higher water demands and a growing economy. Informal markets had existed for a long time prior to the 1999 legislation. The Water Law of Spain formed water centers, and allows for transfers between users, and the transfer of rights for private water use, although water itself is in the public domain. Water rights cover both groundwater and surface water use, and are usually granted for 75 years. The transfers can be time-limited, and water authorities can intervene for the greater good. Transfers can benefit urban, agricultural, and industrial (e.g., hydropower) uses. Droughts in Spain in 2005 and 2008 drove a large number of transfers, and interbasin transfers were also accomplished in these trades (Palomo-Hierro, Gómez-Limón, and Riesgo 2015).

Conclusions

Although water itself may be in the public domain, its use and distribution can be bought, sold, leased, and lent. Physical and virtual water banking can alleviate conditions of water shortage, protect aquifers from collapse, and prevent ground subsidence. Water banks can form a market of exchange and balance of water costs between users and regions. Water banking can also alleviate unevenness of water in time, and allow users to put water to beneficial use in later times than the normal water cycle and water law might allow. Interactions of surface and groundwater must be considered by careful management of water resources that are physically banked in aquifers, and virtually traded and applied.

SEE ALSO: Aquifers; Water budget; Water economics; Water rights

References


“What happens to the rain” is a fundamental topic for geography. Rainwater, as freshwater, is essential for human and almost all of terrestrial plant and animal life. With the exception of seafood, nearly all human food is directly, or indirectly, produced through natural vegetation or forage crops, dependent on rainwater. Public water supply through surface water, and groundwater, is dependent on rainfall, in many dry locations from tens of thousands of years ago during the cooler, often wetter episodes of the Pleistocene epoch. With the exception of eolian processes, water is a fundamental agent in sculpting the Earth’s terrestrial surface, whether as flowing water or ice, or as an agent in the physical or chemical weathering of surface materials.

Placing time and quantitative values on the components of the hydrological cycle is critical for managing water resources. This is exactly what is entailed in generating a “water budget” or “water balance” for a site or an area. Thus, the water budget involves placing quantitative values on the destination of the precipitation ($P$) that is evapotranspired ($ET$), becomes surface runoff ($RO$), becomes groundwater ($GW$) plus a change in surface or groundwater storage ($\Delta S$). Over a longer term the water balance is thus (Dunne and Leopold 1978):

$$P = ET + RO + GW \pm \Delta S$$

However, over the shorter term additional components need to be included:

$$P = I + ET + OF + SM + GW \pm \Delta S$$

where $I$ is interception, $OF$ is overland flow, and $SM$ is soil moisture in the vadose zone (Gurnell 2000). Over the longer term, those three items are evapotranspired, become runoff, or become groundwater.

Water budgets can be calculated for a point location, in which values are in water depth (mm or inches) or can be calculated for an area such as a stream drainage basin or an aquifer in volumes of water ($m^3$ or acre-feet).

In the context of time, water budgets can be calculated for any specific period depending upon the purpose of the analysis. When calculating specific irrigation needs for a crop, water budgets are often calculated for a day, a week, a season, or a year. When analyzing water supply or climate, water budgets are typically calculated for a specific year or for “mean annual” conditions which in the United States is the preceding three decades, for example 1981–2010.

The basic outlines of the hydrologic cycle have been commented upon since antiquity. Jewish King Solomon wrote in Ecclesiastes 1:7 that: “All the rivers run into the sea, yet the sea is not full. To the place from which the rivers come, they return again.” In Greek civilization, Plato relied upon a giant, underground reservoir called Tartarus, that desalinized seawater and then, drawn by the sun, lifted the water to sustain springs and stream baseflow. Aristotle rejected his mentor’s model and described the outlines of the hydrologic cycle, but without quantitative data fit their concepts into pre-existing models about the symmetry of nature. In Roman times, Vitruvius postulated the “hot bath model,” in which water vapor rose from the hot water and...
condensed on the cool ceiling above (Biswa
1970).

In the early scientific era, Bernard Pallisy in *Discours admirables* (Admirable Discourses) qualitatively describes the precipitation-based hydrologic cycle. Taking that concept, Pierre Perauiq measured precipitation, stream flow, and spring flow in the Seine River basin and stated that precipitation alone was sufficient for those water outputs. A generation later, Edmond Halley measured evaporation rates providing the final component for generating a quantitative whole Earth hydrologic cycle (Biswa 1970).

Detailed hydroclimatic analyses began in the 1800s. Rain gauges and methods for measuring stream flow provided quantitative values for analysis, but were limited by the inability to calculate land surface water losses through soil evaporation and plant transpiration (Mather 1991). Initially estimates of combined evaporation and transpiration, “evapotranspiration,” were obtained by simple accounting of liquid water inputs and outputs, with evaporation (*ET*) being: $ET = P - RO + GW$ (Thornthwaite and Mather 1955).

Developing methods for calculating evaporation and combined evapotranspiration has been more difficult. Lysimeters, in effect a flowerpot on a scale, were first used in the late 1880s, where evapotranspiration was equal to the loss of the weight of the “flowerpot” over time. Since then lysimeters up to the size of small buildings have been used to measure evapotranspiration for specific types of vegetation and crops. In the twentieth century infrared spectrometers and cameras have been used to measure evapotranspiration.

Because of the difficulty in actually measuring *ET*, the development of models to estimate *ET* has been critical to the application of water budgets. The agricultural motivation of these efforts resulted in the concurrent desire to calculate optimal water need for vegetation, that is, potential evapotranspiration (*PET*). A critical step in the development of water basin studies was the incorporation of potential evapotranspiration into water budget studies. By developing models to estimate potential evapotranspiration, hydroclimatologists could calculate the supplemental irrigation needs for a given crop.

In a series of papers beginning in 1931, C.W. Thornthwaite developed methods for calculating potential evapotranspiration and actual evapotranspiration based upon temperature, latitude, precipitation, and day of the year. This work facilitated the development of classification systems, first based upon temperature and the ratio of precipitation to evaporation, and then precipitation to potential evapotranspiration. Based upon precipitation and potential evapotranspiration he developed water budgets that included actual evapotranspiration that was supported both by direct precipitation and by soil moisture utilization. In the 1950s, Thornthwaite developed water budget models that provided values for soil moisture recharge, soil moisture utilization, deficit, and surplus over varying time periods. The developed water budget models were directly applicable to irrigation applications, as they calculated the water deficit as the difference between actual and potential evapotranspiration. These later methods were then developed into the WATBUG Fortran IV computer program in the late 1970s (Mather 1991). In the “Internet age” there are online water budget calculators for Thornthwaite methods and others such as Penman *PET* (Ponce 2014).

Developing water budgets has provided an essential tool for analyzing the size of closed lakes in dry regions. Edmond Halley (1715, 298) stated: “Now I conceive that as all of these lakes receive rivers, and have no exit or discharge, so it will be necessary that their waters rise and cover the land, until such time as their surfaces
are sufficiently extend, so as to exhale in vapour that water which is poured in by the rivers, and consequently that lakes must be larger or smaller according to the quantity of the fresh water that they receive …”

Lake size is a function of basin runoff plus lake precipitation and evaporation from the lake surface as a function of its size and the lake evaporation rate. This reasoning has been used to calculate lake sizes during the Pleistocene “pluvials” and to analyze the effects of inflow diversions to modern closed lakes such as the Salton Sea, Aral Sea, and Dead Sea.

Human impacts on natural water budgets, initially inconsequential in pre-domestication times, have dramatically increased. Initially irrigation distributed and hastened the evapo-transpiration of local stream flow from within a basin. More recently, major water imports and exports have created highly altered budgets and hydrological regimes. For example, in the western United States, water of the Colorado River basin is diverted from the tributary Animas River to the Rio Chama of the Rio Grande basin. Farther north in Colorado, Colorado River water is transferred to the Plate River to supply water for the burgeoning Denver metro region. One of the largest water transfers involves the importation of water into coastal Southern California that receives water from the Colorado River, the Owens River of the eastern Sierra Nevada., and the Feather River of the northern Sierra Nevada. Linton (2008) observed that such dramatic water transfers limit the utility of simple regional applications of water budgets because such water transfers create entirely non-natural environments in both the source and the destination regions.

The recent increase in municipal use of groundwater has provided impetus to generate water budgets for aquifers. Aquifers can be considered the subterranean equivalent of drainage basins with inputs, outputs, and changes in storage. However, the difficulty in directly measuring these amounts justified resistance to managing groundwater resources until recently. In a famous case decided in 1904 involving a suit to receive compensation for a lowered water table, the Texas Supreme Court ruled that “Because the existence, origin, movement, and course of such water, and causes which govern and direct their movement are so secret, occult, and concealed … an attempt to administer any set of legal rules in respect to them would be involved in hopeless, uncertainty, and would therefore be practically impossible” (Kaiser 1986, 31). Gradually agencies responsible for aquifers and scientists studying them have developed at least reasonably accurate water budgets for aquifers. Initially these budget models were based upon solving for surface water inputs and outputs and the concurrent changes in aquifer volume. Fluorescent dye studies facilitated the calculation of relative inputs and outputs from aquifers.

The power of geographic information systems (GIS) to analyze and model three-dimensional space facilitated the development of GIS-based groundwater models to quantify components of groundwater budgets. The legal mandates to manage the water resources of specific aquifers have sponsored the development of “water availability models.” Such models are by nature complex as they need to accommodate different hydrologic heads, waters of different densities, and aquifer units of varying specific yield and conductivity. In an echo of the 1904 Texas Supreme Court decision, and in spite of significant technological progress, Kresic (2007) warns that such models remain simplifications of reality and can never be 100% accurate on account of the myriad of variables and uncertainties affecting groundwater.
WATER BUDGET

SEE ALSO: Agroclimatology; Climatology: history; Groundwater; Hydroclimatology and hydrometeorology; Hydrologic cycle; Hydrology: history; Soil mass balance; Soil water; Water resources and hydrological management

References


Water conflicts

Jacob D. Petersen-Perlman
Aaron T. Wolf
Oregon State University, USA

Water has been the source of disputes since the Neolithic revolution, when humans settled down to cultivate food between 8000 and 6000 BCE. This is reflected within the English language: the word “rivalry” is derived from the Latin rivalis, meaning “one using the same river as another.” The ancient roots of the word are still evident today in many situations, as riparians – countries or provinces that have the same river in their territories – are often rivals over shared waters. This is evident today with countries downstream complaining of upstream users, such as Pakistan about India, Iraq and Syria about Turkey, and Egypt about Ethiopia.

Rivalries such as these lead to water being managed for mitigating conflict, as all water management serves multiple objectives and navigates between competing interests. Within a nation, these interests – domestic users, farmers, hydropower generators, recreational users, ecosystems – are often at odds, and the probability of a mutually acceptable solution falls dramatically in proportion to the number of stakeholders. Add international boundaries and the odds fall farther still.

Without a mutually agreed solution, parties with interests in shared water resources can find themselves in dispute or even violent conflict, either with each other or with state authorities. Still, these water-related disputes must be considered in broader political, ethnic, and religious contexts. Water is rarely the single or even the dominant cause of conflict, but competition for scarce or shared water resources can exacerbate existing tensions between countries and therefore must be considered within the larger context of conflict and peace.

The nature of water conflicts is often similar across disparate spatial and temporal scales. So too are many of the solutions to these conflicts, which humans have developed to address water shortages and to cooperate in its management. In fact, cooperative events (water-related events between riparian countries ranging from minor verbal support to voluntary unification into one nation) between riparian states outnumbered conflicts by more than 2:1 between 1945 and 2008. Water has also been shown to strengthen ties, develop cooperation, and even prevent conflict. In some cases, water has been one of the few paths for dialogue in otherwise heated conflicts. In regions that are politically unsettled, water may serve as an essential component of regional development plans, which serve as de facto conflict-prevention strategies.

While the underlying reasons for controversy related to water may be numerous, including power struggles and competing development interests, all water disputes can be attributed to one or more of three issues: quantity, quality, and timing (Figure 1). No matter at what scale the conflict occurs, the key to managing, understanding, and preventing future water-related conflicts can be found in the institutions established to manage water resources.

WATER CONFLICTS

Understanding conflicts over water is particularly relevant, as many basin boundaries do not conform to political boundaries. The world’s 276 international surface water basins that include political boundaries of two or more countries constitute 47% of the Earth’s land area. These basins are also home to around 40% of the world’s population. Many of these basins have more than two countries that contribute to its area, demonstrating the high level of interdependence; for example, the Nile is shared by 12 countries, while the Danube has 19 riparian countries. Thousands more basins are divided at subnational levels.

The nature of water conflict

Because of the prevalence of these water bodies spanning boundaries across the globe, scholars have debated whether disputes between nations due to the transboundary nature of water may lead to violent conflict; the debate appears to be centered on how “transboundary water conflict” is defined. Wolf (2000) has argued that, while there is a growing literature describing water as a historic and, by extrapolation, a future cause of warfare, a close examination of case studies cited in this literature reveals looseness in the classification categories; in other words, it is often unclear how authors define water “wars.” De Stefano et al. (2010) found only 38 acute disputes (i.e., those involving water-related violence) between 1948 and 2008; of those, 31 were between Israel and one or more of its neighbors, with no violent events occurring between Israel and its neighbors after 1970. This means that most of the disputes were (i) political tensions or instability rather than formal acts of warfare, and (ii) instances where water was a

Figure 1 Reasons for transboundary water-related conflictive events and cooperative events (TFDD 2011).
tool, target, or victim of armed conflict. While these disputes are certainly important, they do not constitute water “wars.” Wolf (2000) also cited many incidents of violence related to water at the subnational level; these incidents were generally between indigenous peoples, water-use sectors, and/or provinces. Others suggest that, while the likelihood of war over water is small, a long history of violence associated with transboundary water resources exists, and that future pressures, including (but not limited to) population and economic growth and climate change, could increase tensions. A few characteristics that make water likely to be a source of strategic rivalry between nations include the degree of water scarcity; the extent to which the water supply is shared by more than one region or state; the relative power between basin states; and the ease of access to alternative freshwater sources (Gleick 1993). Nations have cut off access to shared water supplies for various political and military reasons; they have also aimed for new water supplies through aggressive military expansion. Gleick (1993) cited inequalities in water use as another source of many regional and international frictions and tensions.

Though water supplies and infrastructure have often served as military tools or targets, no states have gone to war specifically over water resources since the Mesopotamian city-states of Lagash and Umma fought each other in the Tigris–Euphrates Basin in 2500 BCE. Instead, official cooperation over water has been much more common, with more than 3600 treaties signed between 805 and 1984 CE. Though most of those treaties concerned navigation, treaties signed more recently have addressed water management, including flood control, hydropower projects, or allocations in international basins. Since 1820, more than 680 water treaties and other water-related agreements have been signed, with more than half of these concluded since the mid-1960s.

The likelihood of international water conflict

With the proliferation of water-related agreements, especially in recent years, the likelihood of escalated, even violent, conflict over water resources is one that scholars have debated over the past decades. This debate falls within the framework of hydropolitics (Waterbury 1979), which relates to the ability of geopolitical institutions to manage shared waters in a manner that is politically sustainable, that is, without tensions or conflict between political entities (McNally, Magee, and Wolf 2009).

Examining transboundary water system resilience within a hydropolitical context leads to the concept of hydropolitical resilience, defined as the complex human–environmental systems’ ability to adapt to permutations and change within these systems. In contrast, hydropolitical vulnerability is the risk of political dispute over shared water systems (Wolf 2005). Characteristics of a basin that would tend to enhance hydropolitical resilience to change include international agreements and institutions, a history of collaborative projects, generally positive political relations, and higher levels of economic development. At the same time, facets that tend toward vulnerability include rapid environmental change, increased hydrologic variability, rapid population growth or asymmetric economic growth, major unilateral development projects, absence of institutional capacity, the potential for the “internationalization” of a basin, and generally hostile relations (Wolf 2005). When examining characteristics that would tend to enhance or detract from hydropolitical resilience in combination, it becomes clear that the settings
WATER CONFLICTS

of hydropolitical conflict are most likely with major water projects, such as dams, diversions, or diversion schemes, built in the absence of agreements or collaborative organizations, that can mitigate the transboundary impacts of these projects (Petersen-Perlman et al. 2012).

Variables that increase the likelihood of conflict

The aforementioned variables (or combination of variables) may cause actors in the river basin to move from nonconflict into conflict. Wolf, Stahl, and Macomber (2003, 2) suggested the following relationship between change, institutions, and likelihood of conflict: “The likelihood of conflict rises as the rate of change within the basin exceeds the institutional capacity to absorb that change.” This suggests that two important dimensions exist in transboundary water disputes: (i) a high rate of change in the system; and (ii) a low capacity of institutions in the system to absorb change, for example, low hydropolitical resilience.

Change that may lead to low hydropolitical resilience is manifested in several ways, one of which is the unilateral construction of large water infrastructure in transboundary basins. One prominent example of this unilateral construction occurred in the Tigris–Euphrates Basin, where in 1990 Turkey finished construction of the Ataturk Dam, the largest of 21 dams constructed for the Southeastern Anatolia Project (GAP), and interrupted the flow of the Euphrates for a month to partly fill the reservoir. Despite the Turkish government’s warning of a temporary cutoff of flow, Syria and Iraq’s governments protested that Turkey now had a water weapon that could be used against them. Those fears were prescient, as later that year the president of Turkey, Turgut Ozal, threatened to restrict water flow to Syria to force it to withdraw support for Kurdish rebels operating in southern Turkey (Gleick 1993). Another example of water infrastructure-related conflict is the 1986 case of North Korea announcing plans to build the Kungansan Dam on a tributary of the Han River upstream of the South Korean capital of Seoul. Gleick (1993) suggested that this led to fears in South Korea that the dam could be used as a weapon to flood Seoul through sudden releases of the reservoir. South Korea built a series of levees and check dams upstream of Seoul to try to mitigate potential impacts. More recently, rhetoric from leaders of downstream nations has escalated over Ethiopia’s construction of the Grand Ethiopian Renaissance Dam, arising from concerns over how the dam would impact flows of the Nile for downstream countries. These are strong examples that feature how water management decisions from an upstream riparian can be used (or viewed) as a political weapon to impose its will on downstream co-riparian nations.

Dam construction alone, of course, may not be the only condition for interstate water conflict to occur. Wolf (1998) hypothesized that the aggressor would have to be both downstream and the basin’s hegemon. He also stated that the upstream riparian would have to launch a project which decreases either water quantity or quality, knowing that their actions will antagonize a stronger downstream neighbor. The downstream power would then have to decide whether to launch an attack, and would have to weigh not only invading, but also occupying and depopulating, the entire watershed to prevent any retribution (Wolf 1998). Also, both countries in this scenario would most likely not be democracies, as two democracies have never been at war with one another (Wolf 1998). Climate change may also heighten hydropolitical
tensions within countries, between countries, and/or within river basins. In addition to climate change possibly altering the quantity and timing of flow within river basins and increasing water scarcity, there may also be detrimental indirect, negative effects of reducing food availability and increased exposure to new disease vectors. In combination, these can undermine the legitimacy of governments, damage local and national economies, and have adverse effects on human health. These direct and indirect changes may alter the hydropolitical resilience of the basin; as a consequence, river basins without robust water-related treaties and institutions may be more vulnerable to tension and conflict.

Variables that decrease the likelihood of international conflict

Now that the causes of hydropolitical conflict have been explored, it is natural to discuss what may lessen its likelihood. It is important to note that water cooperation may still happen in times of international tensions; there are several examples of countries actively at war with one another that may still concurrently find a way to cooperate over water. For example, the Mekong Committee, established by the governments of Cambodia, Laos, Thailand, and Vietnam as an intergovernmental agency in 1957, exchanged data and information on water resources development throughout the Vietnam War. Israel and Jordan held secret “picnic table” talks regarding the Jordan River following the unsuccessful Johnston negotiations of 1953–1955, even though they were at war from the time of Israel’s independence in 1948 until they signed the 1994 peace treaty. The Indus River Commission, created by the signing of the Indus Waters Treaty between India and Pakistan in 1960, survived two major wars between the two countries in 1965 and 1971. All 12 Nile Basin riparian countries are also currently involved in senior government-level negotiations to develop the basin cooperatively, despite disagreements.

Building institutional capacity by signing treaties and creating river basin organizations is described as a mechanism to decrease the likelihood of hydropolitical conflict. Giordano and Wolf (2003) cited characteristics that would make a treaty over water effective:

1 an adaptable management structure (including flexibility, allowing for public input, changing basin priorities, and new information and monitoring technologies);
2 clear and flexible allocating criteria;
3 equitable distribution of benefits;
4 detailed conflict resolution mechanisms.

Though there is certainly evidence to support these observations, there may also be inherent weaknesses of certain consent-building relations in water. Riparians can exploit treaties in a number of ways; treaties are not easily enforceable, can be structured to reflect (or exacerbate) existing inequalities between riparians, and can lead to a lack of participation from nonsignatory riparians (Zeitoun and Warner 2006). Perhaps the mere presence of a treaty or treaties does not indicate hydropolitical resilience alone; indeed, perhaps Zeitoun and Warner’s declaration that “the absence of war does not mean the absence of conflict” (2006, 437) may also extend to how the presence of agreements does not signify the absence of conflict. It may also be the parties engaged in treaties/institutions themselves, rather than institutional content or presence, that are at the heart of their success. These assertions are healthy critiques but, within the context of violent conflict, the relationship of institutional
WATER CONFLICTS

capacity and decreased violent conflict still stands (e.g., Wolf, Stahl, and Macomber 2003).

The nature of intranational disputes

The principles for avoiding international water conflicts have been discussed; but it is also important to see how the nature of water conflict changes at the national scale. While the literature on transboundary waters often treats political entities as homogeneous monoliths – “Kyrgyzstan feels …” or “Syria wants …” – analysts have recently identified the pitfalls of this approach, showing how subsets of national actors have different values and priorities for water management. In fact, the history of water-related violence includes incidents between water-use sectors, tribes, urban and rural populations, and states or provinces. Some research even suggests that the likelihood and intensity of violence increases as the scale becomes more local. Worldwide, local water issues revolve around core values that often date back generations. Groups such as irrigators, indigenous populations, and environmentalists, for example, may all view water as tied to their way of life, which is increasingly threatened by new demands and uses for cities and hydropower.

Interstate water conflicts have led to fighting between downstream and upstream users, such as between Native Americans and European settlers and along the Cauvery River in India. These water-related disputes can engender civil disobedience, acts of sabotage, and violent protest. In 2001 the United States Bureau of Reclamation closed the headgates of the Klamath Project in order to protect an endangered fish. This cut off water deliveries to irrigators, leading to large protests. In the Indian state of Orissa, 30,000 farmers clashed with police in December 2007 as a result of the government’s decision to allow a large number of industries to source water from the Hirakud Dam, while the farmers depended on this water for irrigation. Fifty protesters were injured in the confrontation with police. From 1907 to 1913, in California’s Owens Valley in the United States, farmers repeatedly bombed an aqueduct diverting water to the fast-growing city of Los Angeles.

National instability can also be provoked by poor or inequitable water services management. Disputes may arise over system connections for suburban or rural areas, service liability, and especially prices. In most countries, the state is responsible for providing drinking water. Even in cases where concessions are transferred to private companies, the state usually remains responsible for the service. Water-supply management disputes therefore usually arise between communities and state authorities. Protests are particularly likely when the public suspects corruption in water services management or that public resources are being diverted for private gain.

There are both similarities and distinct differences inherent in national and international water conflicts. The differences are more often stressed, but just how different the two settings are is open to debate. The following subsections deal with some common assumptions.

Institutions and authority

National cases are often played out in relatively sophisticated institutional settings, particularly in the West, while international conflicts can be hampered by the lack even of an institutional capacity for conflict resolution. It has been argued, though, that even sophisticated institutions have often not been amenable to relinquishing the traditional, usually legal, approaches to resolving water conflicts, effectively presenting the same challenges as the international setting.
Law and enforcement

The United States and other countries have, over the years, established intricate and elaborate legal structures to provide both guidance in cases of dispute, and a setting for clarifying conflicting interpretations of that guidance. International disputes, in contrast, rely on poorly defined water law, a court system in which the disputants themselves have to decide on jurisdiction and frames of reference before a case can be heard, and little in the way of enforcement mechanisms. One result is that international water conflicts are rarely heard in the International Court of Justice. Likewise, of the international cases presented here, only the Mekong Committee has used the legal definition of “reasonable and equitable” use in its agreement.

In the legal realm too, it has been argued that the differences between national and international disputes are more apparent than real. Given the myriad of legal venues open to disputants, and ambiguities of court jurisdiction, creative lawyers can effectively hamstring legal challenges for years, essentially creating a de facto lack of legal authority.

Presumption of equal power

“All are equal in the eyes of the law” is a common phrase describing national legal frameworks. No such presumption exists in international conflicts, where power inequities define regional relations. Each of the watersheds presented here includes a hegemonic power which brings its power to bear in regional negotiations, and which often sees agreements tilt in its favor as a consequence. Here, too, it has been argued that unequal resources, usually financial or political, result in real world inequities finding their way into the national settings of conflict resolution.

The best alternative to a negotiated agreement (BATNA)

A difference commonly pointed out between national and international disputes is that, in national water conflicts, war is not usually a realistic BATNA. While it may be true that intranational water wars are not likely, the same is increasingly accepted as being true of the international setting. While shots have been fired, both nationally and internationally, and troops have been mobilized between countries, no all-out war has ever been caused by water resources alone. As one analyst familiar with both strategic issues and water resources has noted, “Why go to war over water? For the price of one week’s fighting, you could build five desalination plants. No loss of life, no international pressure, and a reliable supply you don’t have to defend in hostile territory” (personal communication).

While real differences do exist between the national and international settings for water conflict resolution, these distinctions may not be as great as is often thought. The fortunate corollary to this is that many of the successes of alternative dispute resolutions in the national realm may be more applicable to the international setting than is commonly argued.

The difference makers: strong institutions

Strong institutional capacity that can adapt to change can prevent conflict at both the intranational and international scales. Institutions responsible for managing water resources have to be strong to balance conflicting interests over allocation and to manage water scarcity, which is often the result of previous mismanagement. These institutions can even become a matter of dispute themselves. In international river basins, water management typically fails to manage conflicts when there is no treaty spelling out each
nation’s rights and responsibilities with regard to the shared river, nor any implicit agreements or cooperative arrangements.

Similarly, at the national and local levels, it is not the lack of water that leads to conflict but the way it is governed and managed. Many countries need stronger policies to regulate water use and enable equitable and sustainable management. Especially in developing countries, water management institutions often lack the human, technical, and financial resources to develop comprehensive management plans and ensure their implementation.

Institutions may help to solve disputes

Most disputes, regardless of scale, are resolved peacefully and cooperatively, even if the negotiation process is lengthy. Cooperative water management mechanisms can anticipate conflict and solve smoldering disputes, provided all stakeholders are included in the decision-making process and given the necessary information, trained staff, and financial support to act as equal partners. Cooperative management mechanisms can reduce conflict potential by:

- providing a forum for joint negotiations to occur, thus ensuring that all existing and potentially conflicting interests will be taken into account during decision-making;
- considering a variety of perspectives and interests to reveal new management options and to offer win–win solutions;
- building trust and confidence between parties through collaborating and joint fact-finding; and
- making decisions that have a much higher likelihood of being accepted by all stakeholders, even if consensus cannot be reached.

Cooperation-inducing design

Building an environment for cooperation may prove difficult, especially in an environment where a history of cooperation is absent or lacking. Wolf (1995) listed general guidelines for cooperation-inducing implementation, using the pre-peace treaty Jordan Basin as his case study.

- Control of one’s major water sources: It is necessary to address both past and present grievances as a prerequisite for market-driven solutions. As such, an initial “disintegration” of the basin is recommended.
- Cooperation opportunities may be hidden in the details of each party’s bargaining mix.
- Water basin development can proceed from “small and doable” projects to ever-increasing levels of cooperation and integration, remaining always on the cutting edge of political relations.

One step that is often proposed in the beginning stages of cooperation is data and information exchange. Yet, it is important to remember that data collection should be purposeful, as data collection for its own sake may not be particularly useful. Strategic joint fact-finding between parties engaging in a project may serve as an important catalytic tool for developing political buy-in and fostering participation. Joint fact-finding also lowers the perceived risks of cooperation, as it has low sovereignty infringement and transaction costs. Factors that promote this exchange of data and information include the presence of compatible needs, absence of legacies of mistrust, increasing water resources stress, perceptions of mutual benefit, external pressure and funding, comparable levels of institutional capacity, popular and political concern about water resources management,
and functional formal or informal cooperative arrangements (Chenoweth and Feitelson 2001).

**Track I and track II diplomacy**

While any project can be designed to induce cooperation, it is also necessary for all parties involved to go through diplomatic processes to agree on a framework. This is where an outside party may help to play a role in the process. Though third parties alone cannot create a conducive political environment, they can provide incentives to cooperate both directly and indirectly through playing a brokerage role, by:

- providing technical competence and examples of best practices;
- assisting in negotiation and mediation skills, including the provision of legal and other water experts; and

Four different strategies of third-party support can be identified (Mostert 2005): track I diplomacy (cooperation); track II diplomacy (collaboration); track III diplomacy (transformation); and continuing support. Track I diplomacy involves supporting the conclusion of a formal agreement between parties, typically through mediation and facilitation. Track II diplomacy aims at arriving at feasible development strategies on the ground through promoting informal dialogues, research and studies, and capacity building. Track III diplomacy addresses policies at both national and local levels, which are typically at the root of transboundary water problems. Finally, third parties may provide the financial support necessary to sustain cooperation, through funding for a river basin organization or loans for development projects. None of these strategies is mutually exclusive; for instance, track II diplomacy efforts may build the pathway to more formal track I discussions.

**SEE ALSO:** Environmental governance; Geopolitics; Natural resources and human conflict; Rivers and river basin management; Water security

**References**


WATER CONFLICTS


Water conservation

Heejun Chang  
*Portland State University, USA*

Water conservation means efficiency in use or all measures to reduce water consumption. The goal of water conservation is to use water more efficiently in order to meet the needs of humans and ecosystems. Water conservation is a pressing global environmental issue because the world freshwater supply is likely to diminish as the planet warms up and precipitation decreases in some parts of the world, while urban, agricultural, and environmental water demands continuously increase as population and income grow. Additionally, in some parts of the developing world, access to safe drinking water is very limited. Because water conservation implies any institutional, technological, and behavioral changes in water use as they relate to broad social changes, and drivers of these vary by each water use sector, changes by individual sector must be identified. Additionally, the spatial dimension of water conservation needs to be examined because the determinants of water use and conservation vary over space, and reduced water use in one area may induce increased water use in another region.

Urban residential water conservation

Urban residential water use is tightly associated with both macroscale factors (e.g., land use or institution) and microscale individual factors. These factors also interact with each other across scales, which shape complex water use and conservation patterns in urban areas. Densely developed urban areas use less water per household than sparsely developed suburban areas. In many old European compact cities where small lots and green spaces are typical, water consumption rates are two or three times lower than those of other relatively new North American or Australian cities (Sauri 2013). Therefore, changing the density of new development can have significant potential for water conservation.

Even within the same metropolitan area, a wide range of spatial variations in water use exists. Relatively affluent suburban residents in newer developed areas use more water than old dense inner city residents (House-Peters and Chang 2011). This is tightly associated with larger outdoor spaces in suburban houses that require lawn maintenance and irrigation during the summer. Lawn irrigation accounts for approximately a third of the total residential outdoor consumption in North American cities (EPA 2012). In some instances, neighboring residents or neighborhood organizations (e.g., homeowners’ associations) are under pressure to maintain green lawns, which contributes to higher household outdoor water consumption.

However, progress in water conservation has been made in some cities through the introduction of mandatory (regulatory) or voluntary (incentive) water conservation programs. Mandatory conservation programs include water curtailment during severe droughts. A case study of Seattle, Washington, shows that water curtailment is most effective with higher income, larger lot size, larger living space, and smaller household size (Polebitski and Palmer 2013). Institutional change such as changing building code may help...
achieve some water conservation programs; for example, installing new water-saving devices are now mandatory in new homes. Additionally, some cities encourage denser development in newer homes to reduce outdoor water consumption.

Voluntary water conservation programs include customer incentives to reduce outdoor water consumption by promoting the use of weather-sensitive water meter or lawn aeration programs (through rebate in the next water bill from a water provider). Other voluntary programs include the promotion of xeriscape landscaping that uses local native plant species that are adapted to local climate and thus require little water for maintenance. Another approach to achieving water conservation is capacity building that provides necessary information on water-saving devices to the target population that is interested in improving water use efficiency. A study in Santa Clara Valley in California, United States, shows that program participants were able to reduce water consumption by 18% (Reed 2012).

Technological changes have reduced water consumption. New dishwashers, washing machines, low-flush toilets, low-flow showerheads, and faucet aerators can substantially reduce indoor water consumption, while smart water meters or drip irrigation devices can help decrease outdoor water use. Such changes need to be linked to institutional or behavioral changes to be effective, as the adoption of such new technology depends on a wide array of institutional and behavioral factors. Water metering for individual household can be implemented so that urban residents are fully aware of their water consumption patterns and can identify potential water leakage problems.

Other behavioral changes help achieve water conservation targets. Changing water use habits rather than replacing existing water use devices enables urban residents to use water more efficiently. In arid climates, residents can avoid automatic lawn irrigation by applying water more sensitively in response to fluctuating local weather conditions. This will result in substantial reduction in outdoor water consumption. Other behavioral changes in indoor water use include using dishwashers and washing machines when they are full instead of partially filled. Since a wide difference exists between household perceptions of water use and actual water use efficiency, there is much room for new conservation practices in residential water use.

As urban water consumption is closely associated with water infrastructure and urban development patterns, a distributed approach has been suggested for achieving water conservation goals. Urban areas typically rely on transferred water outside of the city, so implementing an on-site water reuse systems can promote alternative sources of water for indoor and outdoor water consumption. These relatively small-scale structures include rainwater harvesting, storm water harvesting, graywater reuse, and on-site wastewater purification systems. Rainwater harvesting is one of the most common practices that is gaining population. Graywater, which comes from household wash water from kitchen, bathtubs, or washers, can be reused for other applications. On-site wastewater treatment systems can minimize evaporative loss and replenish groundwater.

Industrial water conservation

Globally, industrial water use makes up less than 10% of total water use, and the consumption has leveled off or declined in most developed countries. Most of industrial water is returned back to the water system, but the quality of water is typically degraded. Since the enactment of the Clean Water Act of 1972 in the United States,
many industries have started using water more efficiently and discharging less polluted water by recycling and reusing wastewater, which typically accompanies a pretreatment of wastewater (Vickers 2002). Water recycling is the reuse of water for the same purpose, which may require treatment before reuse (EPA 2012). Water reuse is the use of wastewater or reclaimed water from one use (treated wastewater) to another use (landscape watering).

The amount of industrial water used in one region is subject to the specific type of industry in the region because each industry has different water needs. Food-processing, paper mill, chemical, refined petroleum, or metal industries use a large amount of water, while service or knowledge-intensive creative industries use less water. Additionally, while water use generally increases with industrial production, the relationship between water use and production may be either linear or nonlinear depending on the type of industry (Vickers 2002). As such, any changes in regional industrial restructuring have implications on the amount of industrial water that can be saved. Some industries (e.g., food processing) incorporate water into a product (consumptive use), while others (e.g., power plants) return water back to the environment once it is used.

Together with new regulations on water efficiency requirements, technological changes have allowed many industries to conserve water. New improvements in water efficiency in heating and cooling systems, plumbing fixtures, and appliances, as well as federal mandates in water conservation at federal facilities have contributed to the reduction in industrial water consumption. Because a large amount of industrial water is used for cooling, recycling such water has a big potential to save industrial water. A modern recirculating cooling system uses water multiple times for cooling (EPA 2008) rather than using a single-pass cooling system.

Other behavioral and economic incentives can help reduce industrial water use. Like residential water conservation, industrial water conservation can be achieved by monitoring its uses regularly and applying water at the right amount when needed. Metering individual equipment that uses water can also help to identify water use efficiency over time. Increasing the cost of water service has somewhat mixed results. While increasing the costs of water and sewer service offers incentives to promote water conservation, some firms may divert their main water supply from public to private wells, which may not necessarily lead to the reduction in industrial water use (Vickers 2002).

**Agricultural water conservation**

Agriculture is the largest sector of human use of water, with world irrigation water withdrawals accounting for approximately 70% of the total human use of renewable water resources annually. The share of agricultural water in total water consumption is declining, with the absolute amount of water used for agriculture falling in some parts of the world. Approximately half of irrigation water is lost either by evaporation during water transfer or by leaking from infrastructure. Because demand for food production continuously increases and climate change can further increase crop water demand (see Water and climate change), conserving water in the agricultural sector has been a major issue.

The efficiency of irrigation depends on irrigation technology. Surface irrigation (also called gravity or flood irrigation) typically pumps water to the upper end of a field, allowing water to move downward by gravity, but it wastes a lot of water in runoff (Pimentel et al. 2004). This is the most common irrigation method, which
WATER CONSERVATION

has been practiced since ancient times. Sprinkler irrigation applies water through the air from sprinkler heads, distributing water more evenly than gravity systems. While it helps reduce runoff, sprinkler systems are more expensive than gravity systems (Vickers 2002). Drip irrigation uses much (up to 50%) less water than surface irrigation because water drips slowly to the plant root zone through narrow tubes. While drip irrigation minimizes evaporative loss and eliminates runoff and thus helps in water conservation, it is expensive, requires a constant water supply, and needs filtration to avoid clogging up the narrow water delivery tubes (Pimentel et al. 2004).

Improved irrigation technologies and their adoption have contributed to the efficiency of agricultural water use in most developed countries around the globe. A shift from gravity flow irrigation to center-pivot sprinkler or micro-irrigation has decreased the amount of irrigation water, particularly in the dry western United States since the late 1990s. More efficient pressure irrigation systems such as drip, low-pressure sprinkler, or low-energy precision application systems contributed to the reduction in agricultural water use. Even though the total irrigated area increased, total agricultural water use declined in the western United States. Additionally, in gravity-irrigated areas, which account for a relatively small area, less water-intensive management practices have been implemented. These practices include special furrowing techniques, shortened furrow lengths, and laser leveling of fields.

Agricultural practices as they relate to crop production and irrigation behavior can help achieve water conservation. First, farmers adjust water needs to specific crops because each crop has a different irrigation need during the growing season. Second, farmers apply organic mulches or leave biomass to minimize water loss from the soil surface and improve water percolation to the soil (Pimentel et al. 2004). Finally, farmers can reduce irrigation water by carefully selecting the timing and amount of water application using weather information, soil moisture monitoring, and rainfall sensors.

Other practices can also reduce the amount of water needed for irrigation. For example, crop residue management can capture precipitation more effectively and reduce runoff, thus reducing the water requirement for crop irrigation. Additionally, conservation tillage can increase percolation and infiltration, and decrease the amount of water evaporated from the soil surface. Applying these land-based best management practices can work better with pressurized sprinkler irrigation systems than with gravity irrigation systems since farmers apply water when they plant crops in traditional fields operated by gravity irrigation systems.

Educational, financial, and regulatory incentives can encourage farmers to adopt such water-efficient irrigation and farming practices (Vickers 2002). When farmers are well informed about the economic and environmental benefits of water conservation and are given direct technical assistance, they are more willing to adopt new irrigation technology and to change their water use behavior to achieve water conservation. Economic incentives such as the establishment of new pricing structures that reflect the true cost of water delivery and water uses (e.g., multiple block rates) and water rebate programs, help farmers reduce irrigation water use. Regulatory incentives include the flexible supply of water, facilitation of a water market that allows voluntary transfer of water between users, and the promotion of the conjunctive use of groundwater and surface water supply and recycled water.
Scale of effective conservation

When water conservation programs are launched, it must be considered that household- or farm-level conservation may or may not conserve water at another scale such as the sub-basin or the whole watershed scale. Given that some portions of lost water from lawn or farm irrigation may end up with either returning surface runoff or recharging groundwater, water conservation at the individual household or farm level may actually reduce the amount of water available for other purposes such as environmental flow or navigation. This is particularly the case in arid climates where water-saving irrigation technologies can actually increase water depletions at the sub-basin scale, limiting the utility of water conservation programs (Ward and Pulido-Velazquiz 2008). However, when agricultural runoff is polluted, increasing irrigation water-use efficiency at a farm scale can help achieve multiple benefits for farmers, ecosystems, and society. Farmers can draw less water for irrigation, lose less water, and return less polluted water while maintaining both crop yields and water quality in main streams.

SEE ALSO: Water and climate change; Water economics; Water resources and hydrological management; Water: urban

References


Water: drinking

Raj Rajagopal
Michael Wichman
*University of Iowa, USA*

Edwin Brands
*University of Minnesota, Morris, USA*

Definitions, availability, distribution of, and access to drinking water

Water that is made fit for human consumption with minimal short- or long-term harm is called potable or drinking water. Waters in lakes and rivers with surface runoff from rain and snowmelt are referred to as surface water, whereas, water below the ground with rain and snowmelt infiltration and percolation is referred to as groundwater or aquifers. Groundwater and surface water serve as the two major sources of drinking water.

Geologic evidence indicates that water of volcanic origin has been flowing on Earth for most of its 3.8 billion years of existence. Water is essential for all life forms and thus makes the Earth unique in many ways compared to other planets in the solar system. As shown in Table 1, over 96.5% of the Earth’s water is contained in the 1.34 million km³ of the oceans (Shiklomanov 1993). All the ocean water is saline. For an aqueous water sample, concentration of other substances in it, such as salts, is often measured as mg/L (milligrams per liter) or ppm (parts per million), and for low concentrations also as μg/L (micrograms per liter) or ppb (parts per billion). The salinity of ocean water is around 3.5% or 35 000 ppm. For practical purposes, this water without significant modifications is unfit for human consumption and for most agricultural and industrial uses.

In general, dissolved salt in fresh water is less than 0.1%, brackish or estuarine water is in the range of 0.1–3%, and ocean water is in the range of 3–5%. As shown in Table 1, ocean and saline groundwater comprise 97.5% of Earth’s water, which is unfit for human consumption without major treatment. The remaining 2.5% of the Earth’s water, which has low salinity, is termed “fresh water.” Over two-thirds (68.7%) of this fresh water is inaccessible and locked up in polar regions, and the remaining 31.3% is made of fresh groundwater (30.06), ice/permafrost (0.86), lakes (0.26), soil moisture (0.05), atmospheric water (0.04), swamps (0.03), rivers (0.01), and biological water (0.003). Due to the presence of rivers across the global landscape, a column of other water components in relation to the total global river volume (unit of 1) is constructed and presented in the last column of Table 1. For every single drop of fresh river water, there are 43 drops in lakes, 4967 such drops in groundwater, and 637 243 drops of saline water (Table 1).

After rivers, the second largest visible component of fresh water is the lakes. The 91 000 km³ held by freshwater lakes constitute a mere 0.26% of total fresh water. The spatial distribution of the 10 largest freshwater lakes on this planet, Lake Erie, and the remaining lakes as a group is shown in Table 2.

Collectively, Lake Baikal (25.93), the Great Lakes (Superior, Michigan, Huron, Ontario, and Erie combined, 25.06), and Lake Tanganyika...
### Table 1  Estimated global water distribution.

<table>
<thead>
<tr>
<th>Volume in km³</th>
<th>% of fresh water</th>
<th>% of all water</th>
<th>Relative weight to rivers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oceans, seas, and bays</td>
<td>1 338 000 000</td>
<td>96.538</td>
<td>631 132</td>
</tr>
<tr>
<td>Ice caps, glaciers, and permanent snow</td>
<td>24 064 000</td>
<td>68.697</td>
<td>1.736</td>
</tr>
<tr>
<td>Saline groundwater</td>
<td>12 870 000</td>
<td>0.929</td>
<td>6 071</td>
</tr>
<tr>
<td>Fresh groundwater</td>
<td>10 530 000</td>
<td>30.061</td>
<td>0.760</td>
</tr>
<tr>
<td>Ground ice and permafrost</td>
<td>300 000</td>
<td>0.856</td>
<td>0.022</td>
</tr>
<tr>
<td>Freshwater lakes</td>
<td>91 000</td>
<td>0.260</td>
<td>0.007</td>
</tr>
<tr>
<td>Saline lakes</td>
<td>85 400</td>
<td>0.006</td>
<td>40</td>
</tr>
<tr>
<td>Soil moisture</td>
<td>16 500</td>
<td>0.047</td>
<td>0.001</td>
</tr>
<tr>
<td>Atmosphere</td>
<td>12 900</td>
<td>0.037</td>
<td>0.001</td>
</tr>
<tr>
<td>Swamps</td>
<td>11 470</td>
<td>0.033</td>
<td>0.001</td>
</tr>
<tr>
<td>Rivers</td>
<td>2 120</td>
<td>0.006</td>
<td>0.000</td>
</tr>
<tr>
<td>Biological water</td>
<td>1 120</td>
<td>0.003</td>
<td>0.000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1 385 984 510</td>
<td>100.000</td>
<td>100.000</td>
</tr>
</tbody>
</table>


### Table 2  The volume of water held by the top 10 freshwater lakes of the world and Lake Erie.

<table>
<thead>
<tr>
<th>Name</th>
<th>Continent</th>
<th>Volume (km³)</th>
<th>% of total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baikal</td>
<td>Asia</td>
<td>23 600</td>
<td>25.93</td>
</tr>
<tr>
<td>Tanganyika</td>
<td>Africa</td>
<td>19 000</td>
<td>20.88</td>
</tr>
<tr>
<td>Superior</td>
<td>North America</td>
<td>12 232</td>
<td>13.44</td>
</tr>
<tr>
<td>Malawi (Nyasa)</td>
<td>Africa</td>
<td>7 775</td>
<td>8.54</td>
</tr>
<tr>
<td>Michigan</td>
<td>North America</td>
<td>4 918</td>
<td>5.40</td>
</tr>
<tr>
<td>Huron</td>
<td>North America</td>
<td>3 538</td>
<td>3.89</td>
</tr>
<tr>
<td>Victoria</td>
<td>Africa</td>
<td>2 760</td>
<td>3.03</td>
</tr>
<tr>
<td>Great Bear</td>
<td>North America</td>
<td>2 292</td>
<td>2.52</td>
</tr>
<tr>
<td>Issyk-Kul (Ysik-Kol)</td>
<td>Asia</td>
<td>1 738</td>
<td>1.91</td>
</tr>
<tr>
<td>Ontario</td>
<td>North America</td>
<td>1 639</td>
<td>1.80</td>
</tr>
<tr>
<td>Erie</td>
<td>North America</td>
<td>483</td>
<td>0.53</td>
</tr>
<tr>
<td>All others (estimated)</td>
<td></td>
<td>11 025</td>
<td>12.12</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></td>
<td>91 000</td>
<td>100.00</td>
</tr>
</tbody>
</table>

NB: Lake Erie was included so as to have a complete representation of the five lakes that constitute the Great Lakes, USA. Data from LakeNet 2014; NOAA 2014; and Shiklomanov 1993.
WATER: DRINKING

Water in its three major forms of solid (ice), liquid (water), and gas (vapor), has been cycling the Earth’s surface and subsurface for billions of years influenced by physical, chemical, and biological processes. This water movement process is referred to as the water or hydrological cycle, as shown in Figure 1 (Ohio DNR 2011). This cycle is one of the most important natural phenomena and serves as the driving force behind many other life cycle processes on this planet. The heat energy from the sun is the main external fuel that gets this process going on a continuous basis.

One of the most contested debates in recent decades is the influence of human activities (especially carbon emissions), as opposed to nature, on changes to Earth’s surface temperature. A concomitant debate also surrounds the impact of such changes on spatial and temporal variability and uncertainty of extreme weather events such as floods, hurricanes, and droughts.

Drinking water, to varying degrees, may be obtained from several components of the hydrological cycle (Figure 1). Key components of the cycle include evaporation, transpiration, condensation, precipitation, snowfall, surface runoff, infiltration, percolation, and groundwater seepage (USGS 2014). Solar heat enables evaporation of water from oceans, lakes, and rivers and also enables the release of water as vapor from vegetation to the atmosphere through transpiration. As moist air rises in the atmosphere, it cools and condenses to form clouds which then fall to oceans, freshwater bodies, or land surfaces as precipitation/rain and snow. Of the precipitation falling on land, a portion becomes surface runoff that finds its way to creeks, streams, rivers, and lakes, eventually evaporating or emptying into oceans. The remainder infiltrates the Earth’s surface to become groundwater, the majority of
which percolates and seeps into surface water bodies such as rivers, lakes, and oceans. This entire process has been repeating for billions of years with some uncertainty and significant variability over space and time. For example, at Cherrapunji, India, it rained 2300 cms in 1861, whereas at Arica, Chile, there was no rain for 14 years (USGS 2014).

Creeks, streams, and rivers are characterized by their watersheds. The watershed of a river is the area of land that contributes to the river’s flow. The final outfall from the river to a higher order river or the ocean will be represented by waters from the entire watershed. The term “watershed” is used to describe smaller units of land that drain into creeks, streams, and rivers, and the term “river basin” is considered to include many watersheds that are a part of a larger river such as the Mississippi, Nile, Amazon, or Ganges. Hydrologically and spatially defined catchment areas such as watersheds and river basins have played a major role in the administration, planning, protection, and management of freshwater resources based on landform and landscape characteristics and use, in addition to politically derived administrative boundaries.

**Drinking water and regulations**

**Drinking water and public health**

Sanitation, drinking water, and public health are closely connected. On this planet, 2.5 billion people do not have access to adequate sanitation,
783 million do not have access to clean water, and 85% of the population lives in the driest half of the planet. An estimated 3.5 million die each year due to inadequate sanitation, water supply, or hygiene. Up to 90% of wastewater in developing countries flows untreated into rivers, lakes, and highly productive coastal zones. Industry dumps an estimated 300–400 million tons of polluted wastes into water bodies, and nitrate from agriculture is one of the most common chemical contaminants in surface and ground waters (UN Water 2013). Mining, forest operations, power generation, rapid urbanization, land-use change, climate change, and population growth have also become serious stressors on the quality of available fresh waters.

In addition to human wastes, naturally occurring contaminants such as fluoride and inorganic arsenic are found in high concentrations in a number of countries such as Argentina, Chile, China, India, Mexico, the United States, and Bangladesh. In Bangladesh, 20 million and 45 million people are at risk of being exposed to arsenic concentrations greater than their national standard of 50 \( \mu \)g/L and the WHO guideline value of 10 \( \mu \)g/L, respectively (WHO 2011; Flanagan, Johnston, and Zheng 2012).

Combined with the sewer systems of the mid-to late nineteenth and early twentieth centuries, filtration and eventually disinfection of water contributed to near-elimination of diseases such as cholera and typhoid that annually killed significant numbers of urban dwellers. Although it was nearly eradicated by the mid-1900s, cholera has recently reemerged as a pre-eminent public health threat in certain densely populated developing nations: cholera and other diarrheal diseases cause up to 5% of deaths worldwide. Acute waterborne illnesses are relatively uncommon in developed nations, where the focus is on reducing exposure to low concentrations of organic and inorganic chemicals.

### Drinking water treatment

Through the early to mid-1800s drinking water systems in developed nations consisted mainly of pipes for delivering water (e.g., New York and London). Modern technologies for contaminant removal and remediation include filtration, flocculation and sedimentation, and disinfection. Flocculation is utilized to coagulate particles that settle out of the water as sediment in the treatment process. Alum, metal salts, and synthetic organic polymers are often used as flocculants. The clear water above the sediment is filtered to remove smaller particles. Typical filters are gravel, sand, and charcoal. Public water systems are often further treated with a disinfectant such as chlorine, chloramines, chlorine dioxide, ozone, and ultraviolet (UV) radiation to control biological contaminants in the distribution system. Additional treatment such as ion exchange may also be used to remove inorganic ions and activated carbon to remove organic contaminants (US EPA 2016a).

Various small-scale or household treatment systems may also be utilized to treat private well water or provide further treatment of tap water. These systems include simple particulate filters, water softening devices to remove hardness or carbonates, anion exchange to remove contaminants such as nitrate, and activated carbon filtration, distillation, reverse osmosis, and others to treat specific contaminants (State Hygienic Laboratory at The University of Iowa 2013).

Strategies for providing safe drinking water include source water protection and treatment of drinking water to meet health-related water quality standards. A form of pollution prevention, source water protection has been increasingly employed in international and
WATER: DRINKING

national water protection legislation including the European Union’s Water Framework Directive and the United States’ Safe Drinking Water Act. As was the case in historical hydraulic empires, large cities worldwide have continually expanded their influence over available water resources nearby to develop and sustain supplies for their growing populations. Some, such as New York City, have opted for a watershed protection approach which has proven more cost effective than filtering the 1.2 billion gallons provided to its 9 million residents each day.

Despite significant advances in water treatment technologies, drinking water quality is still strongly linked to the quality of the source water. Many countries actively promote source water protection to prevent pollution of groundwater, lakes, rivers, and streams that serve as drinking water feeders for local communities.

In developed countries, only a very small part of the treated piped water is consumed or used in food preparation. Most of the rest is used for toilet flushing, washing, bathing, and landscape irrigation. As a conservation and/or economic measure, dual water supply systems are employed by some for separating such uses.

Drinking water standards

Most if not all nations have adopted some form of health-based drinking water standards that specify the maximum allowable concentration of various classes of contaminants, including inorganic and organic compounds, pathogens, sediment, and radioactive substances. In the United States, drinking water standards are set by the Safe Drinking Water Act (US EPA 2002), in the European Union by the Drinking Water Directive (EU 1998), in Canada by Guidelines for Drinking Water Quality, in India by the Bureau of Indian Standards, and in Australia by Drinking Water Guidelines. Many nations in South America base their drinking water standards on WHO guidelines, and drinking water standards in China are also based on WHO guidelines (WHO 2011). Regulated contaminants include metals such as lead, mercury, and chromium, volatile organic compounds (VOCs) such as benzene and 1,2-dichloroethane, semi-volatile organic compounds including pesticides and polycyclic aromatic hydrocarbons (PAH), and bacteria such as coliforms and E. coli, as well as viruses and radionuclides.

Table 3 compares selected drinking water standards published by the US EPA (2016b), the WHO (2011), and the European Union (EU 1998). Review of the comparison chart shows that there are many similarities in regulated contaminants, but there are also differences in compounds regulated as well as regulatory concentrations allowed. For example, the compound atrazine, a common agricultural triazine herbicide, is included in the US national primary drinking water regulations at a concentration of 0.003 mg/L and in WHO guidelines for drinking water quality at a concentration of 0.002 mg/L, yet it is not a regulated parameter in the European Union. Arsenic is a regulated contaminant by all regulatory bodies at a concentration of 0.01 mg/L.

Drinking water: current and future challenges

There are numerous current and future challenges in the realm of drinking water, including issues of access to resources, available quantity of water, and threats to water quality. Gleick and colleagues (2012) peg the basic human daily need for water at 50 L per person, which includes 5 L of water for drinking and the balance for hygiene, sanitation, and food preparation. In 2010, the United Nations recognized the human
right to clean drinking water and sanitation via Resolution 64/292. The human right to water contrasts with the legal right to use water, which, with some notable exceptions, is often determined by ownership of overlying or adjacent land, connection to an official water supply system, or purchase of a vessel containing water. Societies worldwide developed local approaches to protecting source waters and allocating drinking water, but most modern legal principles for water access and use are based on prior appropriation, riparian rights, or permit systems.

More than 80% of people without access to improved drinking water reside in slums and rural areas of developing nations (WHO/UNICEF 2013). For some, drinking water is often obtained by walking several kilometers carrying containers weighing 20 kg or more. This burden often falls on women and children whose health, education, and general wellbeing suffer as a result (Gleick et al. 2012; Holden 2014).

Lack of access is partly due to mismatches in distribution of fresh water and human populations: for example, India and China together account for nearly a third of global population but possess less than an eighth of global freshwater resources. Access to safe drinking water also varies widely with level of economic development, ranging from less than 50% in many developing nations (sub-Saharan Africa and Oceania) to nearly 100% in many developed nations. Large populations in South Asia, Southeast Asia, sub-Saharan Africa, and Oceania lack access to improved drinking water (Figure 2). In addition to differences in access, average per capita consumption also varies widely between developed (500–800 L/day) and developing (60–150 L/day) nations, and in some extreme cases may be as low as 20 L/day (WHO/UNICEF 2013).

Access to drinking water may also be affected by conflict over freshwater resources. There are several current and longstanding disputes over international river basins including the Nile, Jordan, Tigris-Euphrates, and Colorado. In some cases, there are productive international agreements (e.g., Nile River Basin Initiative) and related processes in place.

Climate change is likely to impact drinking water quality and quantity in at least two ways: (i) via changes in precipitation patterns, and (ii) via sea level rise. General forecasts for impacts of global climate change invariably include an intensification of climate extremes and more rapid circulation of water within short-term components of the water cycle (e.g., precipitation and evaporation). Increased average temperatures are forecast to reduce snow pack in high elevations of the Rockies (United States), and to continue rapidly melting glaciers in the Alps (Europe), Andes (South America), and Himalaya (Central Asia), which will reduce the storage capacity of these “water towers” that provide a significant part of major rivers’ (e.g., Amazon, Colorado, Ganges, Brahmaputra, Rhine, and Yellow) flow. For low-lying small island nations and coastal regions, sea level rise increases the risk of inundation of drinking water infrastructure and fouling of scarce drinking water resources.

Regulation, governance, and economics of drinking water are also significant factors in determining access to and quality of drinking water. There is no standard regulatory practice to protect drinking water supplies; prevailing community norms suggest that local participation in protection and provision of drinking water may be ideal. In many cases, pollution of drinking water sources is regional or global in scale, and is thus outside the control of state, national, river-basin or watershed-level governance. Expensive centralized treatment and distribution systems are but one model for protecting public health and providing reliable drinking water.
### Table 3  Comparison of selected drinking water standards (values in mg/L unless otherwise noted).

<table>
<thead>
<tr>
<th>Analyte/parameter</th>
<th>US EPA Maximum contaminant level (MCL)</th>
<th>WHO Guideline value</th>
<th>European Union Parametric value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acrylamide</td>
<td>*</td>
<td>0.0005</td>
<td>0.0001</td>
</tr>
<tr>
<td>Alachor</td>
<td>0.002</td>
<td>0.02</td>
<td>pesticides†</td>
</tr>
<tr>
<td>Aldrin</td>
<td>0.00003</td>
<td>0.00003‡</td>
<td></td>
</tr>
<tr>
<td>Alpha/photon emitters</td>
<td>15 pCi/L</td>
<td>0.5 Bq/L</td>
<td></td>
</tr>
<tr>
<td>Antimony</td>
<td>0.006</td>
<td>0.02</td>
<td>0.005</td>
</tr>
<tr>
<td>Arsenic</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Atrazine</td>
<td>0.003</td>
<td>0.1</td>
<td>pesticides†</td>
</tr>
<tr>
<td>Benzene</td>
<td>0.005</td>
<td>0.01</td>
<td>0.001</td>
</tr>
<tr>
<td>Benzo[a]pyrene</td>
<td>0.0002</td>
<td>0.0007</td>
<td>0.00001§</td>
</tr>
<tr>
<td>Beta photon emitters</td>
<td>4 mrem/y</td>
<td>1 Bq/L</td>
<td></td>
</tr>
<tr>
<td>Bromate</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Cadmium</td>
<td>0.005</td>
<td>0.003</td>
<td>0.005</td>
</tr>
<tr>
<td>Carbofuran</td>
<td>0.04</td>
<td>0.007</td>
<td>pesticides†</td>
</tr>
<tr>
<td>Carbon tetrachloride</td>
<td>0.005</td>
<td></td>
<td>0.004</td>
</tr>
<tr>
<td>Chlordane</td>
<td>0.002</td>
<td>0.0002</td>
<td>pesticides†</td>
</tr>
<tr>
<td>Chromium</td>
<td>0.1</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Copper</td>
<td>1.0‖, 1.3‖</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Cyanide</td>
<td>0.2</td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>1,2-Dibromo-3-chloropropane (DBCP)</td>
<td>0.0002</td>
<td></td>
<td>0.001</td>
</tr>
<tr>
<td>1,4-Dichlorobenzene</td>
<td>0.075</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>1,2-Dichloroethane</td>
<td>0.005</td>
<td>0.03</td>
<td>0.003</td>
</tr>
<tr>
<td>Di(2-ethylhexyl)phthalate</td>
<td>0.006</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>Dieldrin</td>
<td>0.0003</td>
<td></td>
<td>0.00003‡</td>
</tr>
<tr>
<td>Endrin</td>
<td>0.002</td>
<td>0.0006</td>
<td>pesticides†</td>
</tr>
<tr>
<td>Epichlorohydrin</td>
<td>*</td>
<td>0.0004</td>
<td>0.0001</td>
</tr>
<tr>
<td>Ethylbenzene</td>
<td>0.7</td>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td>Fecal coliform and E. coli</td>
<td>MCL‡</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluoride</td>
<td>4.0, 2.0‖</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Heptachlor</td>
<td>0.0004</td>
<td></td>
<td>0.00003‡</td>
</tr>
<tr>
<td>Heptachlor epoxide</td>
<td>0.0002</td>
<td></td>
<td>0.00003‡</td>
</tr>
<tr>
<td>Lead</td>
<td>0.015‖</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Lindane</td>
<td>0.0002</td>
<td></td>
<td>0.002</td>
</tr>
<tr>
<td>Mercury</td>
<td>0.002</td>
<td>0.006</td>
<td>0.001</td>
</tr>
</tbody>
</table>
### WATER: DRINKING

<table>
<thead>
<tr>
<th>Chemical</th>
<th>US EPA</th>
<th>WHO</th>
<th>EU 1998</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methoxychlor</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitrate (as N)</td>
<td>10</td>
<td>11.3</td>
<td>11.3</td>
</tr>
<tr>
<td>Nitrite (as N)</td>
<td>1</td>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td>Pentachlorophenol</td>
<td>0.001</td>
<td></td>
<td>0.009</td>
</tr>
<tr>
<td>Radium 228 and 228 (combined)</td>
<td>5 pCi/L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selenium</td>
<td>0.05</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Simazine</td>
<td>0.004</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>Styrene</td>
<td>0.1</td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td>Tetrachloroethene</td>
<td>0.005</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Toluene</td>
<td>1</td>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td>Total coliforms</td>
<td>5%**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trichloroethene</td>
<td>0.005</td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td>Tritium</td>
<td></td>
<td>10000 Bq/L</td>
<td>100 Bq/L</td>
</tr>
<tr>
<td>Uranium</td>
<td>30 μg/L</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Vinyl chloride</td>
<td>0.002</td>
<td>0.0003</td>
<td>0.0005</td>
</tr>
<tr>
<td>Viruses (enteric)</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xylenes—total</td>
<td>10</td>
<td></td>
<td>0.5</td>
</tr>
</tbody>
</table>

* US EPA – Each water system must certify annually, in writing, to the state (using third-party or manufacturers certification) that when it uses acrylamide and/or epichlorohydrin to treat water, the combination (or product) of dose and monomer level does not exceed the levels specified, as follows: Acrylamide = 0.05% dosed at 1 mg/L (or equivalent); Epichlorohydrin = 0.01% dosed at 20 mg/L (or equivalent).
† Pesticides means: organic insecticides, organic herbicides, organic fungicides, organic nematocides, organic acaricides, organic algicides, organic rodenticides, organic slimicides, related products (inter alia, growth regulators), and their relevant metabolites, degradation, and reaction products. Only those pesticides which are likely to be present in a given supply need be monitored.
‡ The parametric value applies to each individual pesticide. In the case of aldrin, dieldrin, heptachlor, and heptachlor epoxide the parametric value is 0.00003 mg/L.
§ Polycyclic aromatic hydrocarbons – specified compounds are: benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(ghi)perylene, and indeno(1,2,3-cd)pyrene.
|| US EPA National Secondary Drinking Water Regulations are nonenforceable guidelines regarding contaminants that may cause cosmetic effects (such as skin or tooth discoloration) or aesthetic effects (such as taste, odor, or color) in drinking water.
# US EPA Action Level, if more than 10% of tap water samples exceed the action level, water systems must take additional steps.
** US EPA surface water treatment rules require systems using surface water or groundwater under the direct influence of surface water to (1) disinfect their water, and (2) filter their water or meet criteria for avoiding filtration so that the following contaminants are controlled at the following levels: Cryptosporidium – 99% removal for systems that filter. Unfiltered systems are required to include Cryptosporidium in their existing watershed control provisions, and Giardia lamblia – 99.9% removal/inactivation, viruses – 99.99% removal/inactivation, Legionella – no limit, but EPA believes that if Giardia and viruses are removed/inactivated according to the treatment techniques in the surface water treatment rule, Legionella will also be controlled. Turbidity – for systems that use conventional or direct filtration, at no time can turbidity (cloudiness of water) go higher than 1 nephelometric turbidity unit (NTU), and samples for turbidity must be less than or equal to 0.3 NTU in at least 95% of the samples in any month. Systems that use filtration other than conventional or direct filtration must follow state limits, which must include turbidity at no time exceeding 5 NTU. HPC – no more than 500 bacterial colonies per milliliter.
Source: Data from US EPA 2016a; WHO 2011; and European Union 1998.
Other decentralized approaches that draw upon traditional knowledge and experiences in collective water resource management and strategies, such as rainwater harvesting for household use, may be necessary in urban areas in both developing and developed nations. Unfortunately, such approaches are often not supported or encouraged by existing legal frameworks. A wider array of reliable means and flexible policy frameworks for providing safe drinking water may be necessary to meet future demands.

Treated wastewater is increasingly viewed as a “new” source of water for industrial, agricultural, and domestic use. Treated wastewater makes up more than 10% of total water demand in Kuwait, Israel, Cyprus, Qatar, and Singapore, and some 20 million m$^3$/day are reused in 43 nations. In the United States, treated municipal wastewater has the potential to account for up to a quarter of domestic supply. Beginning in 1969, Windhoek, Namibia, has had a history of using treated wastewater for potable use. In 2002, the reclamation plant capacity was increased to 5.5 million gallons per day (MG/D) or 21 000 m$^3$/day. Singapore also has systems in place for the direct reuse of treated wastewater for potable use, but this practice is relatively uncommon. Indirect reuse of treated water (e.g., injection into an aquifer prior to use as drinking water) is much more common; as of 2005, more than 2000 European cities had indirect reuse systems.
Desalination is a major technology that has been employed to remove salts from seawater, brackish water, or saline groundwater. It is energy intensive, costly, and creates significant environmental impacts. Often, it is employed by regions and nations with access to plenty of energy, such as the Middle East, wealthy nations, such as the United States, Japan, Spain, Australia, and Israel, and nations with large coastal populations and water scarcity, such as India and China. In mid-2015, over 300 million people in 150 countries were estimated to obtain 87 million m$^3$/day of treated water (23 billion gallons per day) from over 18,400 desalination plants (IDA 2015). In recent years, the desalination capacity of various regions has grown considerably with much of the increase devoted to industrial, agricultural, or municipal users, based on the needs of the particular region.

Privatization of water supplies may also affect access to improved drinking water. Water has been viewed through much of modern history as both commodity and public good. During Roman times, for example, those paying for household piped water subsidized the construction and supply of water to public wells. There is considerable contemporary debate concerning private versus public ownership of drinking water systems, particularly in developing nations. Some argue that private corporations would be able to operate systems more efficiently and effectively, while others assert that privatization leads to violations of the human right to drinking water. Notable examples of both failures (price increases, decreased access to drinking water, and ultimately protests and violence leading to cancellation of contracts, as in Cochabamba, Bolivia) and success stories (significant reductions in child mortality in poor areas of Argentine cities in the 1990s) make privatization fertile ground for debate.

Drinking water quality issues include biological (e.g., reemergence of cholera) and chemical threats that are exacerbated by ever-expanding industrial activity, growing human population, increased consumption, and land-use changes. Many if not most commercially significant compounds can be found in surface waters, and therefore in drinking water drawn from these sources. More than 100,000 substances are in regular commercial use, and health effects of many compounds are not known, as toxicological studies have only been completed on a small fraction of the total universe of substances. Combined or synergistic effects of consuming combinations of low concentrations of myriad substances are also unknown.

There continues to be significant concern associated with contaminants that may be present in drinking water that may not be removed by standard water treatment processes, or that are not regulated and thus not included in regulated monitoring programs. These contaminants include, but are not limited to, pharmaceutical compounds, personal care products, pesticides, flame retardants such as polybrominated diphenyl ethers (PBDEs), fabric protectors such perfluorinated compounds (PFCs), and the environmental degradates or these, as well as pathogens. Of particular concern are contaminants that have been characterized as endocrine disruptors, which are associated with adverse developmental and reproductive effects on fish and wildlife, and possibly humans. Some persistent compounds may travel via the food chain or atmospheric circulation to contaminate drinking water thousands of miles from their origin. Treating drinking water to remove substances such as pharmaceuticals is partially effective with conventional treatment technologies such as coagulation and filtration, and higher removal rates can be achieved with advanced techniques.
such as ozonation, nanofiltration, or reverse osmosis (WHO 2010).

Existing as well as new forms of fossil fuel exploration must be balanced with the protection of drinking water sources. Hydraulic fracturing in particular poses significant threats to drinking water sources via contamination of aquifers and surface water bodies with methane and injected fluids. Existing regulatory frameworks may be inadequate for coping with the long-term and potentially irreversible damage to some drinking water sources.

Due in part to public concern about the safety of drinking water, bottled water has become a rapidly growing industry. Global bottled water consumption nearly tripled between 1997 and 2011, from 21 billion to over 60 billion gallons, with the top 10 nations consuming nearly 75% of bottled water and the fastest growth in China, Mexico, the United States, and Brazil. In 2011, the highest per capita consumption, of nearly 65 gallons, was in Mexico. In general, there is less regulatory oversight of bottled water than of piped drinking water, but there may be some cases in which bottled water is the only viable alternative.

SEE ALSO: Groundwater; Rivers and streams; Surface water; Water and climate change; Water conflicts; Water and human rights; Water quality; Water rights; Water security

References


Further reading


Websites

http://www.cdc.gov/healthywater/drinking/
http://blueplanetnetwork.org/about/
http://thewaterproject.org
http://water.epa.gov/drink/
http://water.worldbank.org/related-topics/drinking-water
www.thepacificinstitute.com
Water is a special resource vital for human survival and has long been considered a human right. Recognition of this fact has been established in several international conventions (United Nations 1992). However, increases in water scarcity have led to calls for using economic tools to quell demand. This has resulted in attempts to reconcile these two seemingly disparate viewpoints. The Dublin Statement on Water and Sustainable Development recognized that “Water has an economic value in all its competing uses and should be recognized as an economic good,” which seemingly put to rest the debate over whether water is an economic good. At the same time, the Statement recognized (Principle Four): “Within this principle, it is vital to recognize first the basic right of all human beings to have access to clean water and sanitation at an affordable price. Past failure to recognize the economic value of water has led to wasteful and environmentally damaging uses of the resource. Managing water as an economic good is an important way of achieving efficient and equitable use, and of encouraging conservation and protection of water resources.”

Since its ratification, considerable attention has been placed on determining what constitutes an “affordable price” for water in developing nations. Considerable efforts have been made by nongovernmental organizations to build capacity in developing nations to allow individuals to attain adequate amounts of water per day for drinking and sanitation purposes. Often this takes the form of public wells and low-cost mechanisms to purify surface water to make it potable. The projected impacts of climate change are leading to increased calls to strengthen these efforts. These goals were also codified in the United Nation’s Millennium Project. The goals of providing adequate drinking water have been integrated in the larger Millennium Project Goal 1: “Eradicate extreme hunger and poverty.”

The benefits of using economic tools to manage water demand have strong theoretical support. However, transfers of water between voluntary parties is often undermined or complicated by a series of difficulties, including the lack of infrastructure to move bulk water, excessive legal fees associated with gaining approval to transfer, and temporal delays (Matthews 2003; Gould 1988).

Water economics can be classified into three general categories: domestic use, municipal and industrial use, and agricultural uses. Each category has its own managerial issues associated with pricing structures. For example, cities struggle with determining appropriate pricing for water for domestic purposes. The goal is to maintain availability of a minimal amount of low-cost water for those with limited incomes. The emerging trend is to provide an economic incentive for conservation. Traditional water pricing was often fixed per unit or decreased per unit with additional water used. Newer research suggests water has been priced suboptimal in the past. To promote conservation, newer systems with rapidly escalating pricing based on quantity of use are being adopted; these maintain a low cost for initial water use, followed by a rapidly increasing cost per unit used.
In semiarid areas agriculture represents the largest use of water diverted from a watercourse. Because of the quantity of water extracted for agriculture, water economics plays a substantial role in shaping not only food production but also water availability for all other purposes, including ecosystem protection. At the same time that water pricing can encourage water conservation when the price of improving efficiency is lower than the price of water, agricultural subsidies can distort market signals.

Water pricing can impact what crops are grown in a given area. Agricultural subsidies provide artificially low-cost water to farmers in an effort to keep food prices low. However, they simultaneously discourage conservation by distorting the true value of the water. The result is artificially high levels of production, and a reduced incentive to conserve water. The increase in production has an economic benefit for some areas. However, this practice distorts water allocation schemes, and can reduce the amount of water left in river systems for ecological protection. The overutilization of water for agriculture also has impacts on water quality, as runoff from agricultural fields often contains high levels of fertilizer and pesticide residues.

Numerous concerns and pitfalls exist in the transfer of bulk water. These include carriage losses, or loss of water in transit from the basin of origin to its new location, ecological impacts in the basin of origin, and the socioeconomic impacts of the basin “losing” a portion of its water supply. This creates a difficult paradox: as Tregarthen (1983) once quipped, “An economist might be defined as someone who doesn’t see anything special about water”; however, the needs for reallocation are pronounced. Every reallocation is unique, meaning the exact impacts are not always clear when a transfer is proposed. The ideals of the market economy call for goods to be homogenized as much as practicable for economic efficiency. For market signals to work properly, market participants must fully understand what they are buying or selling—without this understanding an “informational asymmetry” develops (Cooter and Ulen 1997).

Water is unlike many other resources because of its fungible nature. Its ability to move in time and space, and change forms from liquid water to water vapor, makes defining ownership problematic. In the case of water, the myriad property rights criteria (i.e., when a user can use water, from where the water can be taken, how much must return to the stream), along with the uncertain ecologic and socioeconomic impacts of transfers, preclude a homogeneous water market. Research efforts have codified the difficulties in standardizing water transactions (Krutilla 2010; Gould 1988). Brennan and Scoccimarro (1999) cited the lack of clear water rights as a primary deterrent to water transfers in Australia. With similar issues apparent in other semiarid areas such as Spain and the United States, it is clear this is not an issue unique to one particular nation’s property rights laws.

Commoditization of water, along with the subsequent reallocation, poses threats to third-party users by reducing the quantity of water that returns to a stream. To prevent such harm, water managers can impose requirements necessary to ensure the transfer will not impair other water users. Treating water like any other commodity does elicit some rational concerns. Considering the “public interest” before approving water transfers can raise issues of equity and fairness. Critics claim such an imposition constricts an individual’s property rights, and impacts their return on investment.

Restrictions on transboundary reallocations often have a protectionist slant requiring legal intervention to determine the legitimacy of the prohibition. For example, most western
states in the United States imposed restrictions on transfers of water across state lines. These statutes have been the subject of considerable legal review and in most cases have been ruled an impermissible burden on commerce. Such protracted legal battles are not limited to intranational disputes. Many nations have similar restrictions on water transfers. Debate and legal wrangling over international water marketing efforts are likely to be an increasingly frequent occurrence in the twenty-first century.

Impacts to third parties are rarely considered when other commodities are bought and sold—it is logical to ask whether water should be held to a different standard (Sax 1965). Management of water is necessary at least to the level required to ensure downstream water users with vested property rights are not unfairly injured by another party’s water transactions. Beyond this standard the role of government in the spatial and economic use of water is unclear. Most nations have some level of zoning ordinances for land use. At issue is whether restrictions on water use, which have an impact on the broader landscape, are just a natural extension of these kinds of land-use regulations. Still, many critics question the appropriateness of government intervention over water use if the uses in question are “reasonable” (i.e., not wasteful). Also in question is whether it is the responsibility of government to decide what the rural landscape should “look like.” These are simultaneous philosophical and resource management questions that will substantially impact water use and pricing in the twenty-first century. The open-ended nature of these questions will lead to different approaches taken in different political jurisdictions, be it across national or provincial boundaries. While the lack of standardization may be problematic, these natural case studies will provide water resource managers with opportunities to study the efficacy of various techniques, and use adaptive management to enhance water management.

Geographers have long played a critical role in managing water scarcity and studying water usage patterns across the globe. Enhanced technologies within the geographic information sciences allow more robust tracking of water use, and monitoring of areas experiencing water shortages. Coupled with the increased use of economic tools, allocation of water in both rural and municipal areas impacts how and where water is used. Water demand is predicted to intensify to meet the growing needs of rising populations and increases in per capita food consumption. It can be argued that all commodities markets are “messy” with constantly oscillating valuation of the resource. Pricing water at a level that maintains its availability as a basic human right for the poor, while still encouraging water conservation and protecting minimum streamflows, will be a primary natural resource management issue in the twenty-first century.

SEE ALSO: Droughts and water shortages; Water rights

References


Water engineering

Paul F. Hudson
LUC The Hague, Leiden University, Netherlands

Water engineering has been a fundamental dimension of anthropogenic impacts on the environment since humans first started to live upon river banks. The subject spans a range of topics and is especially divided into coastal and freshwater environments, with the latter primarily focused towards hydraulic engineering. This entry considers the subfields of water engineering from the perspective of surface water, and especially alluvial rivers. Hydraulic engineering involves a multitude of discreet actions and structures designed to alter the functioning of flowing rivers for the benefit of humans. Alluvial rivers are physical systems which adjust to changes in sediment and energy, both of which are impacted by water engineering structures.

The earliest comprehensive and large-scale water engineering occurred in the Middle East and North Africa for crop irrigation within the Tigris–Euphrates and Nile Deltas. New sophisticated approaches to water engineering continue to rely upon some of the same principles laid out centuries or even millennia ago. Surface water engineering requires knowledge of basic fluid mechanics to computationally analyze the force of water and its interaction with sediment and the channel boundary layer. Additionally, hydraulic engineering requires knowledge of statistics to forecast time series of stream discharge and suspended sediment data. Common analytical tools apply mathematics and statistics in computer-based hydraulic models (e.g., Hydrologic Engineering Center - River Analysis System (HEC-RAS)), which combine hydrologic, topographic, and channel geometric data within the framework of a geographic information system (GIS). Additionally, much water engineering requires fieldwork, including topographic surveying and leveling, in addition to more job-specific tasks such as stream gauging and sediment sampling. These tools may be applied to, for example, estimating stream power, reservoir storage capacity, sediment transport, or irrigation potential, or to forecasting the impact of dam removal on downstream channel geometry and aquatic habitat.

Inventory of major water engineering structures and impacts

Surface water engineering includes a comprehensive range of activities and structures (Table 1), such as dams and reservoirs, groynes (wing dikes), meander cutoffs and channel straightening, dikes (levees), dredging (cut and fill), bridges and culverts, bifurcations, flow diversions and sluice gates, and revetments. As such, actual engineering structures may range in size from the smallest of culverts (<1 m) placed within small first-order channels in headwaters, to dams and reservoirs, among Earth’s largest human-constructed features.

Groynes

Groynes are elongated rigid structures extending outwards from the river channel bank, emergent at low stage but inundated at moderate to high water levels. The goal of engineering groyne
WATER ENGINEERING

fields is to create a deeper self-cleaning channel, such that it maintains a navigable channel with minimal need for dredging. Morphologically, groyne are oriented nearly perpendicular to stream flow and most often located in shallow channel reaches on the inside of elongated bends and riffles, whereby they trap coarse sediments. Groyne are often placed in a series of relatively closely spaced sets. Groyne are usually semipermeable, constructed from concrete, granite, or basalt. Along large rivers intensively utilized for economic activities, such as the Ohio River, groyne are an ever-present component of the riverine environment.

In addition to reducing fish habitat, because of their resistance to flow, groyne also increase flood stage levels. Indeed, an important measure of the new integrated flood management program along the Dutch Rhine entitled “room for the river” requires that groyne fields are lowered from 0.5 m to 1.0 m to reduce flow resistance at high flow events, which is expected to lower flood stages by between 6 cm and 10 cm.

Dikes (levees)

Floodplain embankment by dike construction for flood control is among the oldest and most common forms of water engineering. Earthen dikes have been constructed for millennia along large rivers, providing flood control and access to agricultural lands, water resources, transportation, and aquatic environments. Development and increased population density follows floodplain embankment, such that dikes are also associated with increased flood risk. Dike construction should be informed by a thorough understanding of flood frequency, floodplain hydraulics and sedimentology, and subsidence, in addition to the ecological values of the local floodplain environment.

Underseepage is a major geotechnical concern along dikes underlain by highly permeable sands and gravels, especially with long duration floods, and can result in sand boils and local flooding along the backside of dikes. Environmentally, dikes are associated with a multitude of negative impacts, such as restricted hydrologic connectivity, degradation of aquatic habitat, and subsidence of local floodplain reaches on the backsides of dikes. A major component of many modern integrated floodplain management plans is to create more “room for the river” by moving dikes further away from the channel, thereby increasing storage capacity for floodwaters while also restoring exchanges of nutrients and fish between rivers and floodplain aquatic habitat.

Revetments

Channel revetments are constructed to prevent bank erosion and are placed along the eroding channel bank, especially cutbanks. Revetments should extend from above the average high water line to between the toe of the river bank or channel thalweg to prevent scour and undermining. Revetments are often constructed of natural materials, such as willow saplings, for small rivers, or of articulated concrete pads for large rivers. Concrete revetments increase local channel velocities and shear stress, and are mainly detrimental to fish diversity. Their effectiveness in maintaining channel stability has resulted in revetments being a major component of the environment along intensively utilized alluvial rivers.

Dams and reservoirs

The most substantial water engineering projects are dams and reservoirs. The purpose of riverine dam construction varies, and includes hydroelectric generation, flood control, agriculture, recreation, and consumption by humans and industry (Graf 2006). Dams are constructed from concrete or earthen materials, with the latter being particularly important along lowland
Table 1  Major features of water engineering and potential impact.

<table>
<thead>
<tr>
<th>Engineering action</th>
<th>Primary justification</th>
<th>Intended effect</th>
<th>Potential adverse impact</th>
<th>Hydraulic/morphologic location</th>
<th>Hydrologic events (low, moderate, high)</th>
<th>Timescale for impact</th>
<th>Spatial impact (upstream, local, downstream)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow diversion</td>
<td>Flood risk</td>
<td>Reduce downstream flood risk</td>
<td>Channel bed aggradation, overbank burial of wetlands</td>
<td>High in water column</td>
<td>High</td>
<td>1–10 years</td>
<td>Local</td>
</tr>
<tr>
<td>Groynes (wing dikes)</td>
<td>Navigable and stable channel</td>
<td>Narrow channel, increase depth</td>
<td>Disruption of sediment budget, trapping of coarse bed material, channel bed incision</td>
<td>Channel bed, banks</td>
<td>Low, moderate, high</td>
<td>10 years to permanent</td>
<td>Local, downstream</td>
</tr>
<tr>
<td>Cutoffs and straightening</td>
<td>Flood risk</td>
<td>Reduce frequency and duration of flooding</td>
<td>Knickpoints and channel incision, channel widening, disconnection of floodplain habitat</td>
<td>Meander bend</td>
<td>High</td>
<td>5 to approx. 50 years</td>
<td>Upstream, local, downstream</td>
</tr>
</tbody>
</table>

(Continued opposite)
| Dikes (levees) | Flood risk | Reduce flood extent | Reduction of connectivity, floodplain accretion, increased flood risk due to increased settlement and economic activities | Floodplain | High | 10 years plus | Local |
| Revetments | Channel stability | Prevent bank erosion | Thalweg scour, disruption of sediment budget | Channel banks (cutbanks) | Moderate, high | 10 years | Local, downstream |
| Dredging | Navigation | Increase channel depth | Disruption of sediment budget | Bars, riffles | Low | Instantaneous | Local, downstream |
| Dams (main-stem) | Flood risk, irrigation, human/commercial consumption | Prevent flooding, support agriculture and economic activities | Disruption of sediment budget, upstream storage of sediment, disruption of flow regime, downstream channel incision, habitat degradation | Channel bed, banks | Low, moderate, high | 10–50 years | Upstream, local, downstream |
rivers. Dams may also be created by natural processes, such as earthquake-triggered mass wasting events which result in large volumes of material blocking river valleys. Several general categories of dams include check-dams, low flow dams, and large main-stem dams. While large main-stem dams constructed upon large rivers receive much publicity, small dams located on low-order rivers are nearly ubiquitous across many parts of Earth’s riverine landscapes.

Dams alter the downstream hydrologic regime by reducing high flows and increasing low flows. Dams also trap sediment behind reservoirs, with sediment trap efficiencies commonly exceeding 95%. The altered hydrologic regime and reduction in sediment load result in downstream channel incision, degrading aquatic habitat. Simulated flow events and sediment flushing can be an effective management regime, but are dependent upon the physical setting and dam operation.

Globally there are currently two main contrasting trends concerning dams: construction of large dams in Asia, Africa, and South America, and removal of small dams in North America and Europe. The science of understanding the downstream impact of dam removal on rivers is in its infancy, with important questions pointed towards understanding the effects of released sediment on channel and floodplain habitat, including the magnitude of change and the timescale for recovery.

Meander bend cutoffs

Meander cutoff for channel straightening is among the most commonly utilized engineering procedures for flood control, but is also used to increase the navigability of rivers heavily utilized for commerce (Gregory 2006). Many meandering rivers in the Northern Hemisphere have been straightened by the process of artificial cutoff. Cutoffs are created by dredging and cutting a ditch across a point bar, especially at an elongated meander neck, during the low flow season. Because the sandy point bar material is noncohesive, the ditch rapidly enlarges as stage levels increase, further accelerating inflow and subsequent incision and widening of the new channel cut. Within several years the cutoff is usually complete, depending upon the sedimentology and subsequent streamflow events. Well-known examples of rivers straightened by artificial meander cutoffs include the Yangtze, the Rhine, the lower Mississippi, and the Kissimmee River in Florida. Meander bend cutoffs were an especially common engineering practice in the United States during the early half of the 1900s, and were seen as part of a broader management program for river “improvement” to reduce flood risk and increase the depth of navigable waterways. Along many meandering rivers, cutoffs have achieved their desired goal of reducing flooding. Unfortunately, the practice has had profound unintended geomorphic and environmental consequences, requiring further management and engineering.

Meander bend cutoffs reduce channel length (and sinuosity) and locally increase channel gradient, thereby causing a channel knickpoint. Hydraulically this is important because the increased channel gradient increases shear stress. The combination of the greater shear stress and higher velocity results in channel bed incision, dependent upon the channel substrate and bed material. Along many rivers which have undergone multiple cutoffs within a relatively short period of time, knickpoints migrate upstream and result in extensive channel degradation. Following channel bed incision, channel banks then become over-steepened and subsequently erode and input large volumes of sediment into the channel. Thus, upstream migration of knickpoints frequently results in downstream channel bed widening and aggradation. This
presents a new problem to river management authorities, which may now need to increase both the frequency of dredging and the construction of revetments to prevent bank erosion and groynes to “train” the channel.

Because meander cutoffs for flood control result in reduction in overbank hydrologic events, the natural hydrologic connectivity is substantially diminished, resulting in significant adverse consequences to floodplains and associated aquatic environments. Management strategies to reconnect rivers to their floodplains are oriented towards increasing freshwater pulse events into aquatic environments. This may involve a gamut of engineering options, including sluice gates, diversion structures, side channel creation, and floodplain lowering, as well as breaching of floodplain dikes.

**Sequence of water engineering**

The timescale for engineering to occur ranges greatly according to the sophistication and scale of the feature, as do the benefits to humans and the adverse impacts to the river environment. Planning should take into account several distinct phases, including (i) initial planning, (ii) construction, (iii) implementation, and (iv) abandonment and/or removal of structures.

The **planning** phase should include a multidisciplinary team of scientists and stakeholders, and establish hydrologic and environmental baseline conditions of the setting. This often includes a review of hydrologic and sediment databases from government agencies, as well as of the geomorphic and environmental setting. The **construction** phase is completed by engineers with heavy machinery, and should be organized with a view to minimizing the construction footprint. Many water engineering activities are closely synchronized to the hydrologic regime, implying a seasonality of construction activities. The **implementation** phase awaits actual hydrologic events to test the effectiveness of the structure, which may then require further calibration. Constructing sluice gates, for example, to manage or restore freshwater inputs into floodplain water bodies often requires additional calibration to optimize sediment inputs. **Abandonment** of structures is common and unfortunately many rivers are littered with relict engineering structures (local dikes and bank protection works), which may continue to influence active hydrologic processes and effectively become part of the modern environment. In some instances prominent features are being removed, such as dikes and some dams, but this depends on the budget and management priorities of government agencies which oversee such activities.

**SEE ALSO:** Fluvial depositional processes and landforms; Fluvial erosional processes and landforms; Geomorphic hazards; Geomorphic systems; Geomorphic thresholds; Rivers and streams; Water resources and hydrological management

**References**


**Further reading**


Water quality

Susanna T.Y. Tong
University of Cincinnati, USA

Water is a precious and limited resource. It covers 75% of the Earth’s surface, of which 2.5% is fresh water, and only 0.77% is accessible. But it is essential to all life forms. Not only does it support aquatic ecosystems, but also human survival is intricately related to the availability of clean water supply for both consumptive and nonconsumptive uses. By supporting domestic, industrial, agricultural, recreational, and other economic activities, such as power generation, transportation, and waste processing, the quantity and quality of water is of paramount importance to a society.

Under natural ecosystems, the quality of water is related to an array of physical, chemical, and biological attributes of the watershed and the waterbody. But with anthropogenic disturbances to the natural environment, the quality of many surface water bodies is deteriorating. Water pollution is a state in which the water has an excessive amount of a certain type of constituents that render the water harmful to human health and the environment and is unusable for humans.

Since the Industrial Revolution, more affordable products have been manufactured, and our standard of living has been improved, but we have generated many forms of wastes. Without much treatment, chemical wastes from industry and domestic sewage were directly disposed of in the receiving water bodies. The American waterways became cesspools and open sewers. The drinking water was contaminated.

In 1972, to address the water quality problems and to ensure that the nation’s waters are swimmable and fishable, the United States enacted the Clean Water Act (CWA), which was subsequently amended in 1977 and 1987. Under the auspices of the Act, the US Environmental Protection Agency (USEPA) is authorized to regulate the quality of the nation’s water. The legislation has provided administrative controls, stipulated regulations for controlling water quality, and provided funds to defray the costs of establishing treatment facilities. The paradigm of water quality management has gradually shifted from the human use strategy to the ecological integrity strategy. The earlier strategy is based on the assumption that water bodies have certain assimilative capacities. The common notion is that dilution is the solution to pollution. The new strategy is to restore and maintain the biological, physical, and chemical integrity of the waterways. In 1976, the USEPA set up the National Recommended Water Quality Criteria, providing standards to assess pollutants to protect aquatic life and human health. Since then, the criteria have been revised. The current criteria provide standards to assess 150 pollutants. Under the Safe Drinking Water Act, the USEPA has set standards for drinking water quality. It requires states to conduct source water assessments for each drinking water intake. The National Pollution Discharge Elimination System (NPDES) was also created, and all industry and wastewater treatment plants are required to have permits to discharge any form of pollutants to water bodies. Moreover, the CWA requires the establishment of the Total Maximum Daily Load (TMDL) program. If a water body is impaired, the states are required
to identify the sources of the pollutant, estimate assimilative capacity of the water body to the pollutant, and determine the maximum allowable discharge pollution load. However, even with these regulations, water pollution in the United States is ubiquitous. According to the 1998 National Water Quality Inventory report to the Congress (USEPA 2000), 40% of the surface water bodies in America do not meet the water quality standard. Contaminated by pathogens, nutrients, and chemicals, they are unfit for fishing and swimming. In the developing world, more than 780 million people do not have access to improved water (UNICEF and WHO 2012).

Types and sources of pollution

There are two major types of water pollution: point source pollution and nonpoint source pollution. Point source pollution comes from identifiable sources and is easier to monitor and regulate. An example is the chemical wastes from industry. It often contains various types of organic and inorganic chemicals, such as polychlorinated biphenyls (PCBs), cleaning solvents, detergents, and metals. Many of these chemicals are toxins and are not soluble in water. Instead, they are soluble in fats. They tend to adhere to soil colloids and reside in the substrate of the aquatic ecosystem. They are persistent and prevalent in the environment and can bioconcentrate in the ecosystem and bioaccumulate in the food chain. When contaminated fish are consumed, they can cause human health problems. Lead and mercury, for example, can affect the nervous system in humans.

In addition to industrial discharge, commercial and domestic wastewater is another important point source pollutant. Wastewater contains debris (grit, sand, gravel, plastic), particulate organic matter (fecal matter, toilet paper, food wastes), dissolved organic and inorganic matter (soaps, detergents, urine, nutrients, and wastes from the degradation of bacteria, algae, and zooplankton), pharmaceuticals, and personal care products. In developing countries, the wastes may be minimally treated or may not be treated at all. Hence, the most common pollutants in these areas are waterborne disease-carrying vectors. Examples are bacteria, viruses, parasites, and pathogens, which are commonly found in human and animal excrements. According to the World Health Organization, 3.4 million people die every year from waterborne diseases. Typhoid and cholera are rampant in developing countries. Coliforms, *Escherichia coli*, and fecal streptococci have been used as fecal indicators for sewage contamination and the presence of other harmful pathogens. Besides diseases, organic matter from human and animal feces and other biodegradable wastes, such as leaves and grass, can cause water pollution problems. During biodecomposition, while the bacteria and detritus feeders consume the organic wastes, they also use up the dissolved oxygen (DO) in the water. When dissolved oxygen is consumed faster than it is replenished from the atmosphere, the amount of oxygen will be depleted. Most fish and macroinvertebrates require an oxygen concentration of more than 4 mg per liter. The lack of DO will result in hypoxia and a dead zone, and the aquatic ecosystem will not be able to support sensitive and endemic species. Many higher aquatic life forms will die and the water will become turbid. Biological oxygen demand (BOD) is a measure of the amount of organic matter in water. A high BOD in a water body signifies not only a high amount of organic matter but also a high probability of low DO. Similar to BOD, chemical oxygen demand (COD) is an indirect measurement of pollution and an indicator of the amount of DO in water.
It measures the amount of oxygen consumed during the chemical decomposition process of the organic and inorganic contaminants.

Although the wastewater in the industrialized world is collected and treated, sewage treatment is not a thorough process, and discharge from municipal treatment plants and onsite septic tanks is often a major source of contamination. In America, municipal sewage typically undergoes two processes. The primary treatment employs mechanical methods and sedimentation to remove debris and other suspended matter. The secondary treatment is a biological process. It allows aerobic or anaerobic micro-organisms, including bacteria, decomposers, and detritus feeders, to biologically disintegrate the organic wastes. When the effluent is discharged, it may still contain contaminants. For instance, the discharge from electric utilities often contains chromium, which in its hexavalent species is carcinogenic. But because of cost, chromium is not treated by water treatment facilities. Instead, the treatment of industrial contaminants is left to the industry, and its discharge is regulated by NPDES. Likewise, without tertiary treatment, nitrogen and phosphorus are not treated. In small watersheds, point source discharge from the municipal sewage treatment plants can be a dominant contributor of nutrients (NRC 2008). Pharmaceuticals and personal care products are emerging contaminants, which are yet to be regulated. When disposed from the sewage treatment plants, some of these products are found to affect aquatic fauna. For example, high levels of estrogen have been attributed to infertility in alligators in Florida. In Europe, the Urban Waste Water Treatment Directive (UWWTD) requires member states to implement tertiary sewage treatment for cities with a population larger than 150,000. Today, many Nordic countries and Western European countries employ tertiary sewage treatment. Consequently, there is a marked decrease in nitrogen and phosphorus discharge from the sewage plants in these countries.

In many older cities in the United States, storm water is routed through the combined sewers, which collect both the sanitary sewage and the urban runoff. This technology originated in England. While it will work well if the amount and intensity of rainfall are relatively consistent throughout the year, the system may fail if the precipitation patterns are irregular. In times of heavy rain, excessive precipitation may flood the combined sewer system, and the discharge may overwhelm the sewage treatment systems. Combined sewer overflow (CSO) can cause serious point source pollution as it may contain not only surface runoff from rain but also untreated raw sewage. This has been a major problem for many cities in the Eastern United States and in the Midwest, such as Cincinnati in Ohio.

Another point source pollutant is hot water discharge from utility plants, which can cause thermal pollution. A plume of hot water can act as a barrier for migration of aquatic organisms. While it may promote the growth of micro-organisms, such as algae, it can affect the metabolic rate or cause stress to other aquatic species. Besides, it can change the thermal stratification, lowering the thermocline below the compensation point where light cannot pass through. As a result, the depth of the nonproductive zone with warm water but no light and oxygen will increase. Moreover, a higher temperature may affect the chemical reactions and change the solubility of gas. The DO in streams will be decreased. Hot water release is a common water pollution problem in many industrialized countries. In the United Kingdom, half of the waterway is used for cooling in power generation and other industrial processes. Although thermal pollution is often related to hot water release, cold water discharge can cause thermal pollution.
as well. When cold water from the bottom of the reservoir is discharged into a nearby stream that is warmer, it can upset the ecosystem. In New South Wales, cold water pollution affects almost 3000 river kilometers.

In the United States, the regulation of interstate transport of oil is vested in the Federal government; it is beyond the jurisdiction of states. But seepage from oil wells and oil pipes and accidents of oil tankers may occur within local communities and can lead to oil pollution. Many petrochemicals, such as benzene and toluene, are toxic to aquatic organisms. Through ingestion, inhalation, or absorption, aquatic life can be harmed. In contact, oil can clot the gills of fish and coat the feathers of water fowls. A layer of oil covering the surface water will also reduce gaseous exchange between the water and the air as well as the amount of sunlight infiltrating into the water.

Since the enactment of the CWA, pollution from point sources has been greatly reduced, and the nation’s water has been improved. The levels of toxins in many water bodies, such as the Great Lakes, have been reduced (ASIWPCA 2004). Some rivers and lakes have also been restored. To cite an example, Chesapeake Bay used to have problems of hypoxia, and eutrophication was a commonplace during the summer months. In 1996 and 1997, *Pfiesteria* was found in the Bay. In partnership with the local states and the USEPA, concerted basin-wide efforts have been dedicated to the reduction of nutrient and sediment loadings. It was an expensive program, but the Bay is gradually recovering. Moreover, throughout the process, new knowledge has been gained in terms not only of scientific discovery but also of the efficacy of various legislative regulation programs and restoration efforts.

Nevertheless, the CWA is not adequate to protect water quality as its goal is mainly concerned with regulating point source pollution, and it does not provide any enforcement mechanisms for controlling nonpoint source pollution. Nonpoint source pollution occurs when rainfall scavenges the contaminants from the land surfaces and washes them to the receiving water bodies. Since nonpoint source pollution is more scattered and may be derived from a large watershed encompassing various states, it remains a challenge to monitor and control.

In the United States, the leading source of nonpoint source pollution and surface water impairment is from agriculture. Depending on the tillage, cropping systems, fertilizer and chemical use, animal stocking intensity, and methods of disposing of animal wastes, runoff from farmlands and feedlots is often contaminated by nitrates, ammonium, phosphates, pesticides, herbicides, bacteria, animal wastes, and sediments. As more wetlands are being drained and the riparian buffer strips removed for urban development, runoff from farmlands is not intercepted and retained and flows directly downstream. In many watersheds, such as the Mississippi, where fertilizers are commonly used, runoff from farmlands is the most dominant source of nitrogen and phosphorus, about six or even 10 times greater than the background natural levels. Nutrients are essential to both aquatic flora and fauna. But excessive amounts of nutrients can stimulate undesirable growth of photosynthetic algae, dinoflagellates, protists, and cyanobacteria. These phytoplankton and zooplankton consume an abundant amount of oxygen, and when they die, oxygen will also be used by bacteria to biologically degrade their dead remains. Known as cultural eutrophication, this process of nutrient enrichment can lead to algae bloom and the depletion of DO. Moreover, some phytoplankton may secrete toxins. The dinoflagellate *Pfiesteria* has been found to cause ulcerative lesions on fish. By damaging the skin, it can predispose the fish to other pathogens.
The cyanobacterium *Microcystis* also produces the liver toxin microcystin, which when ingested or exposed can cause adverse human health effects.

Sediment is another common pollutant in American streams, rivers, and lakes, much of which has a nonpoint source origin. Soil erosion from farmlands and barren soils contribute a large amount of sediments and debris. When washed to the waterways, suspended matters and total dissolved solids will increase the turbidity of the water, deterring the penetration of light into a water body and reducing the amount of primary productivity. Water will become murky, and aquatic animals will not be able to locate food. Sediments can also clot the gills of fish, suffocate eggs, larvae, and young fish, and destroy the habitat and breeding grounds for many aquatic fauna.

Together with nutrients and sediments, herbicides and pesticides are often leached from farmlands to receiving water bodies. In the United States, about 90% of the wells sampled by the US Geologic Survey (USGS) are contaminated by pesticides. The agricultural pesticide dieldrin is highly toxic to aquatic organisms. It affects the immune system of fish, and in high concentrations it can kill them. If contaminated fish is consumed by humans, it can cause headache, dizziness, vomiting, convulsion, and even kidney damage. Atrazine, a commonly used herbicide in cornfields, is found to cause reproductive problems and tissue abnormalities in fish. In humans, methaemoglobinemia (blue baby syndrome), gastric cancer, and non-Hodgkin’s lymphoma have been linked to ingesting water contaminated by fertilizers and pesticides.

In urban areas, storm water runoff from construction sites and impervious surfaces, including rooftops, roadways, and lawns, also contains debris, rubber fragments, heavy metals, road de-icer, gasoline, motor oil, and nutrients. First flush phenomenon is the condition of the first few inches of urban runoff after the commencement of a rain event. The runoff is usually highly contaminated. But as the rain continues, the runoff will generally become less polluted.

Drainage from coal and mineral mines can also contaminate surface and groundwater supplies. Sediments and toxic chemicals in mine ores, such as heavy metals, when exposed during and after the mining activities, can be scavenged and washed to surface water. Recently, hydrofracking has become more popular as a means to obtain an alternative energy source. But the use of chemicals to fracture deep rock strata to collect natural gas, and the disposal of such chemicals, can instigate water quality problems. Besides toxic chemicals, drainage from mines and landfills (commonly known as leachate) can be highly acidic.

Atmospheric deposition can contribute to nonpoint source pollution as well. It may contain contaminants from natural (such as volcanic eruptions and forest fires) or anthropogenic sources. Transportation, fossil fuel combustion, electric power generation, fertilizer production, and other industrial activities emit various species of sulfur and nitrogen in the air. When washed out by rain and deposited on waterways, they form acid precipitation. Acid runoff and acid rain can dissolve other heavy metals and can disrupt growth and reproduction of aquatic organisms. In aquatic systems, a pH value below 5 or above 10 often signifies some form of pollution.

Salinization can be another form of nonpoint source pollution. In many coastal areas, a high water table keeps pressure in the aquifer. Because of the difference in salinity and density between the fresh water and the seawater, fresh water from the groundwater table extends to the ocean, thereby acting as a buffer between the salty water and the fresh water and protecting the underground water from contamination. However, with the abstraction of groundwater or subterranean fluid, the water level in the
groundwater table drops, reducing the pressure in the water table. Consequently, seawater will intrude into the aquifer, causing salinization and contaminating the groundwater. Over-irrigation under a hot and semiarid or arid climate can cause salinization as well. As water is evaporated from the highly soluble sodium and calcium compounds, their residues will be left behind in the soil. Together with the less soluble sulfates and carbonates, soils can become saline. Urban runoff from roadways may be enriched with salts from road salts and de-icers. The excessive amount of salts in soil and surface runoff will ultimately be transported to receiving waters, which can affect stream biota, ecosystem health, and drinking water supply.

Measurement and assessment of water quality

To measure water quality, the traditional method is to conduct in situ water quality samplings and chemical analyses. Although these methods can provide good records of the water constituents, the measurements will only reflect the water quality conditions in one location at one time. Automatic ambient water quality sampling can overcome some of this problem by providing continuous temporal measurements. However, many water quality data loggers will be required to provide a large spatial coverage.

To have better spatial and temporal coverage of the water quality conditions, some scientists use remote sensing imagery. By mounting different sensors, such as Landsat Thematic Mapper (Landsat TM) and Enhanced Thematic Mapper Plus (ETM+), Moderate Resolution Imaging Spectroradiometer (MODIS), Satellite Pour l’Observation de la Terre (SPOT), National Oceanic and Atmospheric Administration – Advanced Very High Resolution Radiometer (NOAA-AVHRR), IKONOS, Quickbird, Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), and Hyperion, on board airplanes or satellites, a constantly updating and almost continual coverage of the landscape and water body in different wavelengths and spatial and temporal resolutions can be captured. The images acquired from these sensors can be used in the detection of snow cover, evapotranspiration, land use, crop type, and vegetation change, as well as in water availability and consumption analyses. Additionally, when the spectral signatures of the imagery are related to the ground-truth water quality data, a near real-time registry of the water quality can be attained. This method works particularly well as the first approximation of the amounts of sediments and chlorophyll in the water body and can be used to monitor soil erosion and algal bloom and facilitate water quality management.

Since analytical chemical information collected from field surveys cannot reflect internal exposure and body burden, some scientists prefer to use biotic data. They are simpler to use and acquire. Biosurveys of flagship or umbrella species can provide valuable information on pollution exposure–response relationships. When multispecies are examined, the information on species composition and abundance in a biotic community can be compiled into metrics and indices. One such index is the Index of Biological Integrity (IBI). The Ohio EPA has derived an IBI consisting of a matrix of 12 criteria of fish communities, including the presence and abundance of sensitive and tolerant species, to depict the instream water quality in Ohio. Other states in America also have similar fish indices to indicate the quality of the aquatic environment. Another index, the Invertebrate Community Index (ICI), uses data on benthic invertebrates to portray the water quality conditions. It consists of 10 metrics depicting the health of the
macroinvertebrate communities. In addition, some scholars have used the community structure and the richness of bacteria taxa to reveal the health of water bodies and catchments. These indices are constructed based on the established knowledge of the ecology and synecology of the biotic communities. The assumption is that the presence and performance of biotic organisms are indicative of the environmental conditions. At a subspecies level, the body burden of the contaminant found in the biota is used to depict the water pollution level. It is also used as the basis for decision-making by the food advisory committees. Visual symptoms in the biotic organisms are good indicators of the water pollution status. Indices, such as DELTA (which denotes deformities, erosion of fins, lesions, tumors, and abnormalities of fish), can be used to indicate the presence of toxins, fungus, and nutrients in water.

Biosurveys and bioassessments are good tools for rapid biotic evaluations. The data collected can be used in empirical, statistical, and stochastic analyses to determine the degree of contamination, the causes of pollution, and the biotic response to the impacted environment. The results are useful in understanding the causal pathways.

With the advent of computer technology and the availability of metadata sets, a variety of empirical and processed-based hydrologic and water quality computer models have been used to simulate the fate and transport of pollutants, water quality, and the potential consequences of various water planning schemes. Commonly used watershed-scaled models are the Hydrologic Simulation Program – Fortran (HSPF), Soil and Water Assessment Tool (SWAT), Spatially Referenced Regression On Watershed attributes (SPARROW), Agricultural Non-Point Source model (AGNPS), Areal Non-Point Source Watershed Environment Response Simulation model (ANSWERS), MIKE-SHE (System Hydrologique European), Water Evaluation And Planning System (WEAP), and Long-Term Hydrologic Impact Assessment (L-THIA). Underpinning the conceptual framework of these models are the hydrologic principles and empirical interrelationships of different variables, including climate, soil, geology, land use, vegetation growth, latitude, topography, basin hydrology, channel morphology, water chemistry, and pollutant transport. Some of these models are integrated with geographic information systems (GIS), expediting data acquisition and processing. The models are often used to assess the current water quality conditions, ascertaining whether or not the water body is meeting the water quality standards.

Future challenges in water quality management

According to the Intergovernmental Panel on Climate Change, the increase in fossil fuel burning and deforestation is causing an increase in atmospheric carbon dioxide, a greenhouse gas, and inducing global warming. At local scales, a higher summer temperature and consequently greater evaporation may increase convective precipitation. A higher winter temperature may lead to a lower amount of snow in higher latitudes and altitudes and earlier snow melt. The changes in global circulation patterns may affect the precipitation regimes in different parts of the world. Some regions may have a higher amount of precipitation, while others may experience a lower precipitation. The variation in precipitation will change the amount of surface runoff and the concentration and loading of pollutants. Hence, climate change is anticipated to affect the hydrologic and nutrient cycles and impact not only water supply, demand, and availability, but also water quality. Furthermore, climate change is likely to change the frequency and
WATER QUALITY

severity of extreme hydrologic events. Intense storms, hurricanes, droughts, and floods may exacerbate water quality problems. Storm events may have a shorter duration, but they may be more intense with a higher erosive power. Dilution, flushing, and delivery of sediments, nitrates, and phosphorus will be increased. The higher temperature and the changes in flow regime can also modify the fate and transport of pollutants.

Other anthropogenic activities, such as channelization, installation of storm drains, and land-use/land-cover changes from forested lands to farmlands and urban areas, can affect the porosity and permeability of the land surface and the retention, interception, and infiltration of water. As such, the surface and subsurface water pathway (surface runoff, groundwater recharge, evapotranspiration, the flashiness of storm hydrograph, and the magnitude of peak flow) will be altered. As the water balance in the catchment is changed, water quality will also be degraded. With the increase in population, many watersheds are urbanizing at a rapid rate, and it is inevitable that there will be more bacteria and higher levels of COD, BOD, and phosphate in these areas, which unequivocally will affect water quality and impair aquatic ecosystems.

Moreover, future urbanization may work in concert with climate change or other changes, amplifying or moderating the frequency and magnitude of floods and droughts, leading to concomitant changes in water pollution. The rise in the use of biofuels to meet the increasing energy demand may also increase the use of fertilizer and the risk of nitrate contamination. In the face of global environmental changes, there is a need to predict the plausible hydrologic and water quality conditions under different future scenarios. However, due to uncertainties of the future as well as in the quality of data and accuracy of the analytical methods, formulation of appropriate future water management plans remains a difficult endeavor.

Mitigation strategies and sustainable water resources management

To mitigate and adapt to the future hydrologic and water quality impacts of global changes and ensure quality water for future use, both point source pollution and nonpoint source pollution have to be controlled. More stringent regulations for industrial and sewage discharge have to be implemented. Improved sewage treatment to remove nutrients or other contaminants can be helpful to reduce pollution from municipal sewage plants. Best management practices (BMPs) can be employed to control nonpoint source pollution. For farmlands, conservation farming practices, such as no till and crop rotation, may help to reduce the amount of sediments and nutrients leached to the streams. To entice the farmers to practice conservative measures, the BMP strategies can be integrated with incentive programs or payments for agricultural services. In urban areas, porous pavements, rain barrels, retention ponds, and buffer strips can be installed to reduce the amount of urban surface runoff and improve its quality. Though expensive, distillation, microfiltration, and reverse osmosis can be used to desalinate saline water.

Today, many watersheds in America do not have a coordinated program for water quality management that involves different states and institutions. But a watershed transcends political boundaries. Hence, better coordination is needed. To meet the challenges of future demand of high quality water in a cost-effective manner and to ensure an adequate supply of clean water, it is essential to link water management with water needs arising from changes in population, energy generation, economy, land use, and climate. Long-term
regulatory programs should be established. Such programs should couple monitoring and water sustainable assessments with management goals. Water policies should involve plans for water appropriation, water-use management, and water conservation, for example, grey water reuse and recycling and the use of water-saving faucets and trickle irrigation systems. Strategies may include water pricing, water trading, and return flow credits. Moreover, it is essential to protect the quality of water, which may entail sustainable landscape management, riparian zone restoration, river management, and pollutant reduction. Green techniques that emulate natural processes, such as the restoration of wetlands, should be considered in tandem with the traditional hard engineered solutions to derive water management decisions so as to optimize the results and achieve water quality objectives. More importantly, there is a need to have a better understanding of the interplay of the biophysical, water quality, and socioeconomic variables in a watershed. Data, knowledge, and skills from different disciplines should be integrated. As such, a holistic and systematic water quality monitoring program and a unified national approach for sustainable water resources management can be implemented.

SEE ALSO: Environmental degradation; Environmental policy; Global environmental change: human dimensions; Hydrologic flow models; Sustainable development; Water: drinking; Water resources and hydrological management; Water: urban

Further reading


Howarth, Robert W., and Roxanne Marino. 2006. “Nitrogen as the Limiting Nutrient for Eutrophication in Coastal Marine Ecosystems: Evolving Views over Three Decades.” Limnology and


Water resources and hydrological management

James L. Wescoat Jr
Massachusetts Institute of Technology, USA

Water resources refer to surface waters that flow across the land and groundwater that percolates beneath the surface, which are subject to various types of human activities and hydrological management (i.e., human measurement, use, modification, and control). Key hydrologic processes in the water cycle include precipitation as snow and rain, interception of rainfall by plants, infiltration into soil surfaces, evaporation and transpiration by plants, snowmelt and runoff of these fresh waters into streams, lakes, and wetlands, along with erosion and sedimentation of those surface water bodies, recharge of groundwater aquifers, and ultimately discharge into saline coastal waters (Figure 1). Hydrology has close relationships with climatology, glaciology, geomorphology, and oceanography. Each hydrologic process is shaped by regional interactions between the Earth’s energy budget, climate, landforms, and land uses which themselves range from forests to grasslands, croplands, industrial areas, and urbanizing landscapes. Each of these types of human settlement modifies hydrologic processes in ways that affect water quantity, quality, and use. Although the distribution and flows of water resources change in dynamic ways, water is not lost from Earth’s hydrologic cycle. It remains as part of the world water balance.

Human activities convert hydrologic processes into “resources” that serve various social purposes. Water resources are therefore defined as systems of supply, demand, treatment, use, recycling, and reuse, until water returns to the atmosphere or ocean. Water supplies are obtained by collecting rainfall, impounding water in upland reservoirs for distribution to lowland areas, pumping groundwater, or desalinating saline waters for various human uses. While such supplies are sometimes regarded as the resource, it is human water demands that drive the search for supplies and transform hydrologic processes into water resources.

In economic terms, water demand can be defined as the quantities of scarce water resources that are required by competing water uses. However, this formulation is complicated in most places by the absence of water markets, objections to treating water as a commodity, and distorted water pricing. Some water uses are highly subsidized while others, such as in-stream flows for ecosystem protection, have no markets or policies that assign a value to them.

Water policies and politics may thus subsidize some uses and undervalue others, which can increase water withdrawals, inefficiency, and waste. Conservation policies seek to establish standards of water use, pricing, and pollutant discharge. Such policies fit within complex frameworks of water governance that comprise water regulations, institutions, and politics that reflect and respond to changing social values. Some societies have developed elaborate systems of private water rights while others rely on state regulation, community governance, or a combination of public, private, and community control. New water demands and values are continually created. For example, demand for bottled beverage products is expanding, as are the aesthetic values associated with water conservation.
Modern domestic water use usually requires treatment through filtration, disinfection, pH control, and contaminant removal prior to use. The growth of municipal and industrial uses has led to increasing withdrawals of water for drinking, power plant cooling, and waste disposal. Although these uses consume little water through evaporation, they alter water quality in ways that require wastewater treatment. Irrigation agriculture by comparison consumes a large proportion of the water withdrawn through evapotranspiration by crops, and it returns a relatively small proportion to streams and aquifers, along with increased concentrations of salts and agricultural chemicals. Recreational and ecological uses of modified rivers require that some level of minimum in-stream flows be maintained.

Water management strives to mitigate hazards, especially floods and droughts. Hydrologic processes also contribute to mudslides, accelerated erosion, land subsidence, storm surge, dam failure, and water-borne disease. Taking the example of floods, it has often been said, in news accounts and insurance policies, that flood events are “Acts of God.” However, geographer Gilbert F. White (1945) transformed the field of hazards management by documenting how flood losses are “acts of man.” Human beings have a long history of occupying floodplains inundated by river channels for purposes of water transport, fisheries, land development, and other benefits. Investment in flood protection through dams and levees can induce further floodplain settlement, and thus greater exposure to less frequent but
more damaging events. White also showed how nonstructural measures of flood risk reduction, such as warning, evacuation, insurance, and preparedness, can reduce losses. Subsequent research has shown how social groups marginalized in terms of race, class, or ethnicity have been pressured to live in vulnerable floodplain locations (Mustafa 2013). Hazards mitigation policies strive to rectify such injustices, reduce exposure to extreme events and losses of life and property, and increase resilience (i.e., the capacity to recover from disaster).

Water resource problems and hazards generate conflicts, and they can be aggravated by conflicts over other issues. Ancient Mesopotamian, Roman, and Asian water laws record conflicts over scarce water supplies for irrigation, fishing, and domestic water use. They stipulate the obligations of one water user to another and the penalties and remedies for various types of damage. The development of modern nation-states from the seventeenth century onwards has led to formal transboundary conflicts between countries and among the subnational states within countries. Water conflicts range from formal declarations of war to lesser uses of force and legal disputes. Although water wars as such have been extremely rare in history, lesser conflicts are frequent and intense (Gleick et al. 2014). Conflict resolution at all scales is thus a central function of water governance.

No discussion of water conflict would be complete without some reference to the role of social power relations and politics, and to historian Karl Wittfogel’s (1981/1957) controversial hypotheses about the role of water in the origin of cities and states. Wittfogel argued that despotic empires arose through forced construction of large irrigation projects in arid regions, while more democratic societies were associated with decentralized water systems and private property. These arguments were proven to be false, but they drew attention to the importance of social and political dimensions of water management. Unequal power relations contribute to the impoverishment of irrigators at the tail end of canals, place heavy burdens on women to fetch water and cope with water-borne disease in the home, and increase exposure of low-status water users to water pollution and flooding (Mustafa 2013). Wittfogel and his critics helped usher in a comparative global perspective on water resources (Wescoat 2000).

### Estimating the world’s water supplies

Aside from very small losses and gains in the outer atmosphere, the Earth has a finite amount of water. It circulates in dynamic ways while sustaining a balance at the global scale. Early global estimates of how much water there is on Earth were made in the late-nineteenth century and were revised in the mid-twentieth century by Soviet geographers, who also estimated runoff for each of the continents, in a book titled *World Water Resources and Their Future* (L’vovich, 1979).

These estimates were further refined by Russian hydrologist Igor Shiklomanov, who increased the proportion of groundwater to ice and glaciers, and added the category of biological water (Table 1). The largest volume of water in the world is in the oceans (93.5%). In terms of fresh water, the main volume is held in ice caps, glaciers, and permafrost (69%), followed by groundwater (30%). Surface waters in soils, rivers, and lakes constitute less than 1% of global fresh water. And biological water, including that which constitutes over 60% of the human body, is a mere 0.003% of the fresh water total.

However, these static figures are misleading because some water bodies cycle much faster...
than others. Their renewal periods (i.e., length of time required to discharge and replace all of the water in that condition) vary from hours to thousands of years. The annual flux shown in the far righthand column is the amount of each type of water renewed over the course of one year. Those quantities underscore the significance of rapidly circulating atmospheric and ocean water, followed by soil moisture and comparable volumes of water circulating through ice, lake, and river forms. Particularly interesting is the massive circulation of biological water (1647.1 × 10³ km³/year), which is almost three times the annual circulation of the atmosphere. This large volume of biological water movement gives support to the emerging subdiscipline of ecohydrology, which strives to bridge the ecological and hydrologic sciences.

Remote sensing and world data centers for water supplies

These gross estimates of global water stocks and flows are changing as new hydrologic data are produced by remote sensing satellite technologies. The entire suite of water resources measurement technologies and computational methods is evolving dramatically (National Research Council 2012). For example, new satellite sensors are planned to transmit near real-time data on soil moisture and evapotranspiration. Global synthesis of these measurement and modeling efforts has come in part through international scientific projects such as the Global Energy and Water Experiment (GEWEX). Initiated in 1990, GEWEX has gone through three cycles of program development focusing on global data, hydroclimatology, and land atmosphere interaction modeling. In contrast with earlier world water balance estimates, based on

Table 1 Distribution of the world’s water.

<table>
<thead>
<tr>
<th>Component</th>
<th>Volume of total fresh and saline water (10³ km³)</th>
<th>% of total fresh and saline volume</th>
<th>% of freshwater volume</th>
<th>Renewal period (years)</th>
<th>Annual flux 10³ km³/year</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ocean</strong> (saline)</td>
<td>1 338 000</td>
<td>93.5</td>
<td>–</td>
<td>2 500</td>
<td>535.2</td>
</tr>
<tr>
<td><strong>Fresh water</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groundwater</td>
<td>33 930</td>
<td>2.46</td>
<td>30.1</td>
<td>1 400</td>
<td>16.7</td>
</tr>
<tr>
<td>Soil moisture</td>
<td>16.5</td>
<td>0.001</td>
<td>0.05</td>
<td>1</td>
<td>16.5</td>
</tr>
<tr>
<td>Ice and glaciers</td>
<td>24 064</td>
<td>1.65</td>
<td>68.7</td>
<td>1 600–9 700</td>
<td>15–2.5</td>
</tr>
<tr>
<td>Permafrost</td>
<td>300</td>
<td>0.022</td>
<td>0.86</td>
<td>1 0000</td>
<td>0.03</td>
</tr>
<tr>
<td>Lakes</td>
<td>280</td>
<td>0.019</td>
<td>0.04</td>
<td>8 days</td>
<td>589</td>
</tr>
<tr>
<td>Swamp water</td>
<td>11.5</td>
<td>0.0008</td>
<td>0.03</td>
<td>5</td>
<td>2.3</td>
</tr>
<tr>
<td>Atmosphere</td>
<td>12.9</td>
<td>0.001</td>
<td>0.04</td>
<td>16 days</td>
<td>48.4</td>
</tr>
<tr>
<td>River water</td>
<td>2.12</td>
<td>0.0002</td>
<td>0.006</td>
<td>Several hours</td>
<td>1 647.1</td>
</tr>
<tr>
<td>Biological water</td>
<td>1.12</td>
<td>0.0001</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1 454 193</td>
<td>100</td>
<td>100</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Source: Modified from Shiklomanov and Rodda 2003, 13, 17.
periodic ground measurements, remote sensing provides more information on spatial and year-to-year variation, interactions, and trends (e.g., Trenberth et al. 2007). In addition to GEWEX, international scientific organizations have created world data centers that may be consulted for records on various components of the hydrological cycle, such as the following.

- Global Precipitation Climatology Project (GPCP),
- National Snow and Ice Data Center (NSIDC),
- Global High-Resolution Soil-Water Balance (Consortium of International Agricultural Research Centers – Consortium for Spatial Information, CGIAR-CSI),
- Global Runoff Database (GRDB),
- National Climate Data Center’s (NCDC) global climate databases.

Rapid advances in data storage, data mining, and modeling will also transform twentieth-century approaches to global hydrology and water resources.

**Local water balance analysis**

Similar water balance methods, also known as water budgets, have practical applications at local scales. For example, climatic water balance analysis compares two key variables – precipitation (snow and rainfall in millimeters) and potential evapotranspiration (the maximum depth of water in millimeters that would be evaporated or transpired by vegetation when there is unlimited water available). These variables are plotted over time (Figure 2). When precipitation (P) is greater than potential evapotranspiration (PET), soil moisture is recharged and surplus water begins to run off (dark blue). When PET exceeds P, soil moisture is drawn down (yellow), there is a deficit, and plants undergo stress (orange). Annual water balance studies have been used for local irrigation, forestry, and water supply planning. Multi-year water balance analysis with shorter time steps are used to assess water supply variability and drought hazards.

Although water balance analysis appears simple, estimating the evapotranspiration term is challenging. The earliest models tried to use only air temperature data, but that neglected key processes of energy and water transfer at the Earth’s surface. The Penman–Monteith equation is the most widely used equation today: in addition to temperature it includes measures of solar radiation, humidity, and heat transfer in the ground and atmosphere. Its widespread use for agricultural purposes has led to the creation of public databases such as the United Nations Food and Agriculture Organization’s CLIMWAT and CROPWAT webdata and software.

**Figure 2** Climatic water budget diagram for 0.5 degree grid that includes Cambridge, MA, USA. DEF, deficit; −DST, decrease in soil water storage; +DST, Increase in soil water storage; SURP, surplus. University of Delaware. http://climate.geog.udel.edu/~wimp/. (Reproduced by permission of Professor Cort Wilmott, University of Delaware).
Hydrological measurement and management

Each component of the water balance poses resource opportunities and hazards that need to be measured and managed, and it is useful to examine each one of them.

Precipitation

Often the first process considered in the hydrologic cycle is precipitation in the form of rainfall. Although there are examples of ancient rain gauges, the modern instrumental record for most areas of the globe only dates back to the start of the twentieth century. Precipitation records are complicated by weather station histories that change location, land-use context, measurement technology, or recording methods (e.g., human observation, automatic recording, or telemetric data transmission). Various techniques are used to convert these measurements at a point in space into areal and volumetric data, for example, interpolating observations between stations, constructing representative polygons (Thiessen polygons) on flat terrain, or mapping equal rainfall lines (isohyets) on more varied terrain. Enormous effort has been devoted to designing rain gauges that dampen wind effects and other sources of measurement bias. As noted above, measurement is being revolutionized by remote sensing analysis of storms for both weather forecasting and climatology. Remote sensing is especially valuable for estimating the spatial and temporal distribution of intense storm cells and low-frequency storms in arid regions.

Precipitation management remains difficult. Various experiments with cloud seeding using silver iodide and other precipitation nuclei have been attempted since the mid-twentieth century with little in the way of controlled results. More widespread are attempts to estimate probable maximum precipitation to plan for floods based on estimates of maximum storm intensity and the time of concentration for rainfall to reach the stream. Record precipitation measurements range from 38 mm per minute and 1870 mm per day in Cilaos, Reunion, in the Indian Ocean to 26461 mm per year at Cherrapunji, India.

Precipitation management in the early twenty-first century is focusing on rainwater harvesting and climate change. In the field of climate change, precipitation variability remains among the great uncertainties, in terms both of shifts in precipitation means, and of changes in rainfall variability and extreme events (IPCC 2012). Rainwater harvesting, by comparison, is an ancient method of rooftop and surface runoff collection and storage, which is being rediscovered and adapted for new contexts and uses. India is a leader in the field, but there are centers of innovation in all regions of the world (Agarwal and Narain 1997). In addition to rainfall quantity, water harvesting is affected by changes in precipitation quality and the dry deposition of particulates which require diversion of the first flush of runoff, sedimentation, filtration, and storage. The interception of precipitation varies by different tree species, vegetation structure, and building facades in urban environments – all of which can cause a 10–40% change in the amount of rainfall that reaches the soil and the pathways by which it does so.

Snow and ice

Snow and ice hydrology are crucial for river and reservoir forecasting in high latitudes and high-altitude headwaters, but they are even more challenging to measure. Issues range from the complexity of snow physics and snowpack dynamics to snow water chemistry and alpine water quality. The geographic distribution of snow is concentrated in landscapes above 40’N
and on the windward face of high elevation ridges (e.g., the Hindu Kush-Karakorum-Himalayan mountain ranges). The percentage of precipitation that falls as snow tends to increase linearly with altitude. The liquid water contained in snowpack is a function of its density (e.g., fresh powdery snow may have a density of 10% in which 10 cm of snowfall would contain 1 cm of liquid water). Melting depends upon energy and heat transfer through the snowpack. Snow management includes meltwater forecasting using temperature and energy budget models, as well as physical structures such as snow fences. Redistribution of snow through windblown drifts and avalanching complicate snowmelt runoff modeling.

Avalanche management has developed methods for prediction, warning, controlled triggering of dangerous accumulations, and construction of snow retention structures and bridges. At the other end of the spectrum, low snowfall years have given rise to snowmaking technologies to support skiing and winter sports industries. Cooperative management of snowmelt processes has been documented among villages in the Karakorum region of northern Pakistan (Kreutzmann 2000). Concern is escalating worldwide over glacial retreat associated with climate change in most regions of the world, from the small disappearing glaciers of the Andes to large gliated regions of the eastern and central Himalayas, the Greenland ice sheet, and arctic and Antarctic sea ice, which are closely tracked to assess the implications of sea level rise.

**Infiltration and soil moisture**

In nonurbanized gentle grasslands and forest landscapes, most precipitation infiltrates into the soil. The maximum rate of infiltration depends upon soil texture, grain size, crusting, crumb structure, grading, and organic matter. On relatively flat vegetated loamy soils, rainfall infiltrates up to the moisture-holding capacity of the soil, when all of its pores are filled, after which water ponds on the surface and begins to run off or evaporate. Soils with high sodium content in arid regions (sodic soils), or iron content in the humid tropics (laterites), may have surface soil structures that repel water, especially when compacted by grazing or human activities. Once infiltration occurs, percolation of moisture downward through the soil by gravity continues until it reaches an unconfined water table, recharging groundwater aquifers and causing the water table to rise until the soil is saturated. Percolation is accelerated by plant root holes, fauna tunnels, rock fracturing, and lenses of porous soil; it is obstructed by rock layers and dense hardpan layers in a soil that may be broken up or compacted by different cropland cultivation methods. Infiltrated water may also drain laterally through the upper layers of a soil in a process known as interflow, emerging in springs or stream discharge.

When drainage ceases, the soil is said to be at field capacity, which retains moisture available to plant roots. Managing the soil moisture used by plants is a primary determinant of agricultural food and fiber production. It is also important for managing slope stability, erosion, sedimentation, and landslides affected by the additional weight and lubricating effect of soil moisture.

Saturated soils and ponded water in flat areas create wetlands whose boundaries are subtle and dynamic. Throughout history, enormous human effort has been devoted to soil drainage and filling for land development, cultivation, and health reasons. Drainage of wet (hydric) soils may involve cutting open channels or installing porous drain pipes at regular intervals; protecting low-lying lands by earthworks known as polders, levees, or bunds; and pumping drainage waters
over those barriers. These drainage practices accelerate the subsidence of soil surfaces, especially in soils with high organic content that decomposes when drained, leading some landscapes to subside below sea level and become vulnerable to coastal storm surge as well as terrestrial flooding. In agricultural regions, drainage water may be laden with chemical leachates. In addition to their effects on pollution, algal blooms, and reduction of oxygen (eutrophication) in local water bodies, agricultural chemicals can accumulate in regional food chains and contribute to large-scale oxygen reduction (hypoxia) in coastal waters.

An estimated half of the world’s presettlement wetlands have been lost through drainage and filling for agricultural and urban land development. Wetlands are defined by the presence of water, hydric soils, and aquatic vegetation—all of which fluctuate seasonally and from year to year in what is termed the “hydroperiod” of the wetland. Wetlands have been deemed unhealthy from antiquity, due in part to the habitat they provide for insects but also to erroneous theories of poisonous vapors (miasma). However, the late twentieth century witnessed a dramatic reversal of attitudes toward wetlands, brought about in part by scientific understanding of their rich ecological productivity and economic value (e.g., for fisheries, waterfowl, flood protection, and stormwater treatment). Wetland protection efforts have increased at every level, from local advocacy to the international Ramsar Convention on Wetlands of International Importance in 1971. Despite increasing recognition of wetland values, delineating them for protection and regulating land development remain highly contested processes. Conservation of degraded wetlands has had some success, while attempts to create new wetlands are more difficult and less productive in terms of ecological values and services.

**Evapotranspiration**

Soil moisture is subject to evaporation and transpiration through plants. Evaporation from open water surfaces such as lakes, ponds, and reservoirs has important effects on microclimate and storage. Evaporation from oceans is the principal source of atmospheric water vapor (approximately 85%). Evaporation from terrestrial water bodies is smaller but significant for water supply planning. Efforts to reduce evaporative water losses through shading are successful in small-scale cisterns and covered water reservoirs. Floating objects and films designed to reduce evaporation on medium-sized water bodies are often blown to the edge of the water body. Evaporation from upper soil horizons occurs through capillary action and heating. It is particularly significant on bare cultivated soils at the start of the growing season. It can be reduced by mulching, reduced tilling, and composting, which have been practiced from prehistoric times to the present.

Transpiration from natural vegetation and irrigated plants represents the highest consumptive use of water worldwide. Transpiration rates vary greatly by species, climate, cropping practices, and atmospheric heat stress. Plant water uptake is also influenced by rooting patterns, microbial transport, and soil porosity. When plant water stress is initiated, a host of chemical and physical signals operate to conserve water and reduce plant stress. Wilting and defoliation can be followed by plant mortality due to hydraulic failure in the root, stem, or leaf zones, carbon starvation, or stress-induced disease.

In light of food security implications, a great deal of agronomic research has focused on managing plant water requirements. The Green Revolution in the late twentieth century developed hybrid plants that have greater yields but also greater water and nutrient requirements. Subsequently, plant breeding has focused on
drought tolerance as well as yields and disease resistance. A cultivation method known as the System of Rice Intensification (SRI) is producing higher yields with fewer, more carefully tended plants that have lower total water requirements. Efforts are being made to test intensification techniques for other crops such as wheat as well.

Evapotranspiration from phreatophytes (deeply rooted floodplain plants) such as *Tamarix* can be a major source of water losses along river corridors, especially on impounded rivers favored by invasive plants. This problem has led to eradication methods that further disrupt riparian ecology, and to watershed experiments that aim to increase water yield by changing tree to grass cover, a method that at best increases short-term yields while also aggravating erosion and wildlife habitat disturbance.

By far the greatest potential for managing evapotranspiration involves irrigation management. Every irrigation variable from scheduling to water delivery, application methods, and drainage water reuse can help increase water use efficiency. For example, precision land leveling ensures more uniform application rates, water-course lining reduces seepage losses, and drip irrigation delivers water directly to the root zone. Evapotranspiration-based irrigation scheduling uses automated controllers with on-site weather station data, soil moisture sensors, and dynamic crop coefficients to deliver the right amount of water for each stage of plant growth. Excess irrigation runoff reuse reduces agricultural chemical runoff. However, care must be taken that conservation does not increase salinity concentrations or impact downstream users.

**Groundwater**

Water that percolates below the root zone continues to drain through the soil until it reaches either an unconfined water table lying on top of an impervious soil layer, a limestone (karst) flow channel, or an aquifer confined between two impervious layers. Depending upon aquifer slope and saturation conditions, percolation may be lateral as well as downward. Groundwater ultimately discharges into springs, streams, or coastal waters. Most soil and rock strata are complex in materials as well as stratigraphy, which complicates the measurement and modeling of groundwater quantity and quality. Aquifer drainage may or may not follow surface drainage patterns. In addition, the timescales of aquifer flows are much longer than those of surface waters, which makes groundwater supplies difficult to calculate through simple water balance methods.

Most ancient societies relied upon shallow hand-dug groundwater wells. Expansion of large-scale canal irrigation in the nineteenth and twentieth centuries led to dramatic changes in shallow groundwater conditions, including waterlogging and salinity in some irrigated regions of the world. Diffusion of pumping technologies in the mid-twentieth century reduced waterlogging in some areas, but greatly depleted streams, traditional irrigation systems, and groundwater levels in others. The rate of groundwater depletion accelerates when energy and pumping costs are subsidized (Birkenholtz 2009), and it is so dramatic in some regions that it has been inferred from satellite measurement of variations in the Earth’s gravity in some regions.

Depletion is by no means the only challenge. Saline groundwater caused by natural geological deposits and leaching of salts by irrigation has diminished soil fertility from the Colorado to the Amu Darya basins. In the latter basin, the Aral Sea has receded so far that it has left salt and agricultural chemical residues along the former sea margin. These deposits are transported by wind erosion, causing various respiratory and internal diseases. Careless lining of oil and water
pump shafts has also led to mixing of saline and freshwater aquifers. Groundwater contamination by industrial wastes raises enormous concerns due to the technical difficulty of cleanup, and the long renewal times of most aquifer waters, which range from hundreds to tens of thousands of years. Leaching of contaminants from surface wastes is one concern. Injection of drilling fluids for oil, gas, and hydraulic fracturing is another. The latter is extremely controversial as careless drilling can easily contaminate shallow aquifers in this underregulated industry. Modeling the flow, transport, and fate of groundwater contaminants is technically difficult, and the consequences of groundwater’s misuse are likely to be as serious as those of its overuse.

Runoff into springs, streams, and rivers

Excess stormwater runs off into streams and rivers, generating high flows and sometimes floods. Soil moisture and groundwater maintain the base flow of springs and streams, long after precipitation storm events have occurred. Springs were deemed sacred in ancient Greece and Rome, as the sites of nymphs and magical grottoes. At the same time they were pragmatically tapped by venators (Latin for “hunters”) who opened their flows into aqueducts that served major cities and landscapes.

The smallest tributaries emerge from springs and join other streams to form perennial rivers in all but the most arid regions. Arid and dry mountainous regions, by contrast, have intermittent channels (called arroyos, torrents, or wadis in different cultures), which carry occasional thunderstorms with high velocity and erosive power. River channel planning on large perennial rivers such as the Danube and Rhine Rivers in Europe has centuries-long records of engineering, development, and pollution. Many rivers were channelized and impounded by dams. While the removal of small dams is increasing in some regions, such as the United States, construction of large dams is accelerating in China, the Himalayas, and the Andes. Efforts to restore streams to fishable, swimmable, and ecologically productive systems is now a specialized professional field with impressive examples but limited in regional-scale implementation. While the transportation of hazardous substances was addressed early in European water treaties, that did not prevent continuing processes of urban and industrial contamination from occurring.

Streams are the most accessible water bodies and are thus the best known components of the water budget. Even so, they present complex challenges for scientific management of multiple uses, flood hazards, and water quality. Channel networks vary by terrain, from the most common dendritic (branching like a tree) patterns on alluvial soils to rectangular patterns in geologically block-faulted terrain and streams that radiate from volcanic mountains.

In each case, stream flows respond to precipitation storm events through peak flows that rise during the initial time of concentration in the watershed, and wane as surface water runoff declines. Stream discharge is influenced by watershed size, shape, steepness, land cover, soil depth, and channel configuration. Channel cross-sections and alignments are affected in turn by slope, suspended sediment load, bed load, and riverbank materials, which produce characteristically meandering, braided, or rock channel sections. Some channel morphologies exhibit the self-similarity of fractal systems. Each channel type and configuration has a floodplain shaped by former river channels, deposits, terraces, and natural levees that reflect its history of flood events.

Human activities also shape complex river systems in the following ways:
WATER RESOURCES AND HYDROLOGICAL MANAGEMENT

- changes in land use that accelerate runoff, erosion, sedimentation, and river pollution;
- floodplain use for farming, settlement, and waterfront development;
- river crossings by bridges, fords, and ferries;
- stream channel use for domestic drinking, fishing, and boating purposes;
- channel engineering (straightening, widening, armoring) for drainage, flood control, navigation, and commercial shipping;
- dam construction for multiple purposes of flood control, water supply, and hydropower;
- withdrawals of river water for irrigation, municipal, and industrial use;
- wastewater disposal from all human activities, from agricultural runoff to urban sewer discharge, industrial releases, and accidents; and
- planned and accidental introduction of exotic and invasive aquatic species.

These river modifications must be examined within a comprehensive approach to floodplain management. While the regulation of land uses adjacent to stream channels has proven politically difficult to implement, nonstructural measures such as warning, evacuation, and flood proofing are now diffusing widely.

Deltaic, estuarine, and coastal waters

As rivers approach the sea, they shift from tributary to distributary systems. The distributary channels have reduced slopes and velocities that discharge coarse sediments on the levees that guide them and fine sediments in the low-lying swamps and marshes between them. They meet tidal processes that redistribute those sediments and mix the salt and fresh waters in brackish zones that extend miles inland. These brackish mixing zones (estuaries) tend to be some of the most ecologically productive waters in the world for plankton, shellfish, and fish spawning. Management of estuaries is made challenging by their complex flows, biogeochemistry, and mixing processes. Human actions compound these difficulties by the construction of roads and infrastructure that block freshwater flows and sediment transport. River engineering confines channels within artificial levees that reduce the overbank flooding and sediment deposition that sustain delta growth. This leads to accelerated coastal subsidence and erosion of all but the main delta channels. These channels often have major ports that discharge partially treated wastes from municipal and industrial facilities. The Rhine delta has gone through a long succession of flood control and management regimes, the most recent of which strives to “make room for the river.” Other deltaic regions in North America, Asia, and Africa are studying the Rhine’s example closely for lessons that may be adapted to other regions (National Research Council 2013).

Water resource management at different scales

In everyday life, water management begins at the personal and household levels of early morning excretion, bathing, and drinking. It continues through the day in manifold uses of water to support different forms of livelihoods, waste disposal, and reuse. Each of the hydrologic processes discussed above is involved and affected. For thousands of years communities have developed diverse ways of managing water resources, hazards, and water conflict. They are shaped by water management practices at larger regional, national, and increasingly international scales, which interact with one another in complex ways. In recent decades, specialized water planning has developed at each of these scales, from the global to the local.
Global and international water management

International river problems date back at least to the mid-seventeenth century with the advent of modern nation-states. Early water conflicts among nations dealt primarily with issues of navigation on major rivers. By the eighteenth century, international river organizations were organized to address physical, economic, and political conflicts over shipping. By the nineteenth century, water scientists and planners in Europe and elsewhere began to realize shared interests in knowledge about water resource management and water-related hazards. They formed international nongovernmental organizations (INGOs), such as the International Meteorological Organization in 1879 and the International Institute of Agriculture in 1905, both of which became United Nations (UN) agencies in the mid-twentieth century. As international concerns grew over water quality, environmental impacts of large water projects, and safe drinking water and sanitation for all, new UN agencies and NGOs were created to address them.

The early twenty-first century is witnessing growing concerns about global water problems, including global climate change (Gleick et al. 2014). As populations grow, gross per capita water supplies are declining relative to the water scarcity threshold of 1000 m³ per capita per annum, especially in arid regions of the world, though the significance of these gross estimates is challenged by those who underscore the greater problem of unequal access to water by different social groups. The United Nations Millennium Development Goals sought to reduce the estimated 1 billion persons without adequate access to safe drinking water by half by 2015, and this goal was reportedly met ahead of schedule (WHO/UNICEF 2014). Far less progress has been made in assisting the estimated 2 billion persons worldwide who lack access to safe sanitation. Deaths from flooding are declining, albeit with spikes during extreme events, but the economic losses of flooding worldwide continue to rise due to increasing floodplain development. Child mortality due to water-related diseases is declining as a result of oral rehydration therapy and antibiotics, though the frequency of those diseases and their related malnutrition effects remain high. These positive trends in water management reflect the purposeful efforts of millions of people working in the water sector.

Water is arguably the major natural resource concern associated with global climate change, along with its connections to energy and food security. Although there is a scientific consensus about climate warming trends, there is much greater uncertainty regarding trends in precipitation and extreme water hazards. Fears of international water wars are often repeated, notwithstanding research on the historical rarity of large-scale military conflicts over water. Most agree, however, that there is a high frequency of social conflict over water uses that sometimes becomes violent. Conventionally, modern water planning worldwide has distinguished four major categories, or subsectors, of water withdrawal and use:

- agricultural water use – for irrigation, livestock, and aquaculture;
- industrial water use – for thermoelectric power production, mining, and commercial purposes;
- municipal water use – for domestic and related urban purposes; and
- in-stream water uses – for recreation, fisheries, and navigation.

However, the practice of water planning requires that they be addressed jointly, in part due to competition and conflict between subsectors and in part due to recycling and reuse...
of wastewater. It is encouraging that total water demand forecasts in the late twentieth century exceeded actual withdrawals in the year 2000. However, forecasts in the first decade of the twenty-first century project a 50% increase in withdrawals by 2025 in some studies and by 2050 in others (UN Water 2014).

The UN Food and Agriculture Organization compiles the most comprehensive country data on water use in its AQUASTAT database. It relies on self-reported country data, which are variable in quality and often estimated rather than measured. AQUASTAT includes five main categories of data:

1. geography and population – baseline areas and populations for normalizing water use data per hectare and per capita;
2. water resources – estimates of country’s supplies from precipitation, renewable domestic and transboundary surface and groundwater quantities;
3. water withdrawals – categorized by type of use (agriculture, municipal, and industrial); and source (fresh water, groundwater);
4. irrigation and drainage – areas under irrigation, by source of water, technology, and type of drainage; and
5. environment and health – limited in this database to waterlogging, salinity, nutrition, and water-borne disease.

Worldwide, agriculture is by far the largest water use in terms of withdrawal from natural water sources (70%), though regions such as Europe withdraw far more water for municipal and industrial purposes (79%). As urbanization increases worldwide, there is pressure to increase agricultural water productivity (“more crop per drop”) and to transfer water from agriculture to urban use. Urban populations are also demanding increased regulation of agricultural water pollution.

The UN also publishes a World Water Development Report periodically on a wide range of indicators (Table 2), which are broader in social and environmental scope than assessments focused on water use alone.

The Global Water Partnership (GWP) is an important international NGO that seeks to integrate international water programs. It was created following the Dublin Conference on Water and Environment in 1991 and the UN Conference on Environment and Development at Rio de Janeiro in 1992. Those conferences propounded four guiding principles for world water management: (i) water is a finite and vulnerable resource; (ii) water is an economic good; (iii) water management should occur with public participation at the lowest possible administrative level; and (iv) women play a central role in water management. GWP has subsequently advocated an approach known as Integrated Water Resources Management (IWRM), which strives to more closely link water development with improved environmental, economic, and governance approaches. It holds triennial conferences on major issues and themes. In 2014, the main thematic foci were as follows.

Table 2 World Water Development Report indicators.

<table>
<thead>
<tr>
<th>Level of stress on the resource</th>
<th>Governance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Settlemenrts</td>
<td></td>
</tr>
<tr>
<td>State of the resource</td>
<td></td>
</tr>
<tr>
<td>Ecosystems</td>
<td></td>
</tr>
<tr>
<td>Health</td>
<td></td>
</tr>
<tr>
<td>Food, agriculture and rural livelihood</td>
<td></td>
</tr>
<tr>
<td>Industry and energy</td>
<td></td>
</tr>
<tr>
<td>Risk assessment</td>
<td></td>
</tr>
<tr>
<td>Valuing and charging for the resource</td>
<td></td>
</tr>
<tr>
<td>Knowledge base and capacity</td>
<td></td>
</tr>
</tbody>
</table>

WATER RESOURCES AND HYDROLOGICAL MANAGEMENT

- Climate resilience and water security.
- Transboundary water security.
- Food and water security.
- Urbanization and water security.
- Energy and water security.
- Ecosystems and water security.

GWP meetings and the IWRM approach have been increasingly challenged by social and environmental activists who regard it as a neo-liberal approach to economic growth, trade, top-down planning, large infrastructure finance, and commodification of scarce water supplies and services.

Every six years, the Intergovernmental Panel on Climate Change (IPCC) provides a perspective on global freshwater impacts, adaptation, and vulnerability. Its Working Group II report on the impacts of climate change in 2014 underscored concerns in which there was “high confidence” scientifically that glaciers, permafrost, and ice-caps are melting (IPCC 2014). A special report on water-related disasters and extreme events reported more variable levels of scientific confidence about the effects of climate change on floods and droughts (IPCC 2012).

Global databases on floods, droughts, and other hydrologic disasters have been compiled by organizations such as the Centre for Research on the Epidemiology of Disasters (CRED). The trends in Figures 3 and 4 reflect increased reporting, but they also indicate increased floodplain encroachment and investment, and may indicate more extreme storm events as well. However, the most severe impacts associated with water hazards occur where governance has broken down in areas of war, civil strife, and involuntary population displacement.

Such crises have contributed to international movements to recognize a universal human right to water and sanitation (Sultana and Loftus 2012). In 2010, the UN passed Resolution 64/292 on the Human Right to Water and Sanitation, according to which each person must have access to water that is sufficient (50–100 liters per capita per day), safe (World Health Organization water quality standards), acceptable (for domestic use), accessible (within 1 km of the home), and affordable (no more than 3% of household income). The resolution was supported by 122 countries.


![Figure 3](image)


![Figure 4](image)
Some larger international regions have formulated integrative water policies. One of the most advanced of these is the European Union’s Water Framework Directive adopted in 2003 (European Commission 2007). The Water Framework Directive calls for common monitoring programs by 2008, river basin planning by 2009, and good water status in all rivers, lakes, coastal waters, and groundwater by 2015. However, an interim progress report indicated that most countries of the European Union still report that a majority of their surface water bodies are “at risk” or have “inadequate data.”

Transboundary water conflict resolution

Some transboundary water disputes go to international tribunals and courts. These include conflicts over upstream water development (e.g., dams, industrial discharge, increased runoff) that adversely affect downstream water availability and environmental quality, or downstream development that forecloses upstream navigation and future water uses.

Three major breakthroughs occurred in the late twentieth century that addressed transboundary water disputes. The UN Convention on the Law of Non-Navigational Uses of International Watercourses was completed in 1997, after more than 20 years of negotiation (McCaffrey 2007). By formulating the concept of “international watercourse systems,” this agreement struck a balance between river basin planning, which was rejected as too broad in scope by some riparian nations, and river channel planning, which was seen as too narrow. It developed the concept of an “international watercourse system” to encompass a network of surface water bodies and associated aquifers that cross or run along an international boundary. The Convention calls for a balance between the principles of “equitable and reasonable utilization,” and “no significant harm.” It requires prior notification of development plans to neighboring states. Although only 36 states were parties to the Convention as of 2016, that was sufficient for it to enter into force as the current legal perspective of the UN.

The Transboundary Freshwater Dispute Database (TFDD) at Oregon State University was a major academic development. This web database is searchable by country, river basin, water issue, and other variables. It provides the full text of historical water treaties online. It also includes an international water events database for 1950–2008, which compiled accounts of international water news, and ranked them from major conflict (−7, war) to major cooperation (+7, voluntary territorial unification). Interestingly, most events fell on the positive side of the scale, indicating that cooperation has been more common than conflict (Wolf 2007). Some scholars disagree, arguing that a lack of extreme events in the database does not necessarily signify the absence of severe conflict, especially in regions of power imbalance. The TFDD database includes transboundary agreements between states within a country, which can be just as intense and protracted as international conflicts.

UNESCO (the United Nations Educational, Scientific and Cultural Organization) has established an international Centre for Water Law, Policy, and Science in Dundee, Scotland, which emphasizes international water law research with students from around the world. As indicated by its title, it seeks to link legal research with policy analysis. Similarly, UNESCO also supports an International Hydrology Education (IHE) program based in Delft, The Netherlands, which addresses issues of water management as well as scientific hydrology. International education at such centers plays an important role in advancing the theory and practice of international water research and conflict management.
National water resources management

Most countries have national agencies that deal with water resources, although these are often fragmented across separate agencies responsible for navigation, flood control, irrigation, fisheries, drinking water, pollution control, and so on. Smaller countries with strong central governments are sometimes able to concentrate water knowledge and authority at the national level (e.g., The Netherlands, Singapore). The Netherlands has had a combination of local water boards with national coordination for almost a millennium. Although a relatively new nation, Singapore faces reductions in water supplies from Malaysia and is therefore establishing strong programs for water development and management, ranging from potable water reuse to ecological stream restoration, wetland treatment, and coastal water impoundment for water supply.

Some countries have developed integrated water programs because of major historical events. For example, after the dismantling of South Africa’s apartheid regime, representatives of the country traveled the world studying national water law and policy approaches. Beginning in the 1990s, it adopted policies for both cost recovery and access to basic water services by the poor. Another exceptional case occurred in Australia where a lengthy drought followed by floods led to national intervention for water rights reform in 2010–2012. Brazil stands out for its comparative international studies of national drought policy alternatives that have been undertaken to inform its own policy. While these innovations face continuing challenges, they constitute advances in national water policy.

The situation is more complex in federal systems of government where constitutional authority for water matters rests primarily at the subnational state level, even when financial and technical resources may be concentrated at the national level. The United States, for example, has failed to adopt a coherent national water policy since the 1970s (Christian-Smith and Gleick 2012). Notwithstanding the strong water research and management capabilities of the US Geological Survey, National Oceanic and Atmospheric Administration, Environmental Protection Agency, US Army Corps of Engineers, and other agencies, these capabilities remain fragmented at the national level and only partially coordinated with counterparts at the state level.

India also has a federal system of government with water responsibilities divided between the state and national levels. The government of India adopted a new National Water Policy in 2012 which has stimulated vibrant debates in state, national, and civil society forums. After surveying the array of water problems that India faces, the policy calls for a new water framework law and outlines policies for adaptation to climate change, enhancing water supply, demand management and water-use efficiency, water pricing, environmental protection, floods and droughts, project planning, transboundary rivers, database development, and so on. Implementation committees are now devising ways to achieve this broader scope of national water policy.

In the face of increasing water demand, some nations are pushing ahead with large dam construction, notwithstanding domestic and international opposition based on environmental costs, seismic risks, and large-scale human displacement. Examples include China’s Three Gorges Project, India’s Sardar Sarovar dam on the Narmada River, and Turkey’s Southeastern Anatolia Project (GAP, Guneydogu Anadolu Projesi). Some of these emerging economies have the internal financial resources to build their own dams, while countries in weaker financial positions continue to rely upon international
development banks and bilateral aid to finance large dam projects.

Watershed and river basin planning

The field of scientific hydrology began with local experiments in the seventeenth century which showed that rainfall over a watershed was sufficient to produce springs and streams, as opposed to ancient arguments by Aristotle and others that springs originated from deep groundwater sources. The French geographer Philippe Buache (d. 1773) mapped physiographic basins of the continents and oceans. Mapping river basins was important in processes of exploration and settlement, for example, by Joseph Nicollet (d. 1843) and John Wesley Powell (d. 1902) in the American West, and English geographer James Rennell (d. 1830) in British India.

The concepts of watershed and river basin are closely related. Both refer to a drainage area delimited by a ridge (e.g., a mountain ridge) that directs and discharges runoff to a common outlet. Although sometimes used interchangeably, the term “watershed” tends to mean a relatively smaller drainage used for agriculture and forestry, while the term “river basin” refers to a larger-scale drainage area managed for multiple purposes of hydropower, navigation, and water storage for urban and industrial uses. Watershed management has been associated with movements to decentralize water management while river basin planning has accompanied large-scale public infrastructure development. Both approaches strive to link water management with related land resources. A common challenge that they face is defining the “navigable waters” of the basin, as that has been a criterion for determining the scope of public rights of water access and environmental regulation since Roman times. Some jurisdictions follow a strict definition of waters that are navigable in fact, while others include creeks and wetlands. During the late twentieth century, watershed and river basin planning have both given increasing emphasis to water quality and ecosystem health.

Smaller-scale watershed management is practiced in uncounted areas around the world. Early European watershed planning was linked with scientific forestry and hill torrent control. The United States has witnessed the formation of thousands of watershed organizations dedicated to local environmental planning and water quality improvement since the late twentieth century. India’s participatory watershed management program has focused on upland resource protection for erosion control, which in some regions has pitted forest hill tribes and pastoralists against lowland farmers. Watershed management is especially effective at reducing erosion, runoff, and flooding from regularly recurring storm events.

At the larger scale, the iconic example of integrated river basin development is the Tennessee Valley Authority (TVA) in the United States. The TVA began with the conversion of a dam and munitions factory constructed in World War I to peacetime use, and it ultimately encompassed a broad range of innovative programs for rural electrification, agricultural development, and hydropower generation to alleviate rural poverty. The hydropower options were initially assessed in basin-wide surveys carried out by the US Army Corps of Engineers. However, to address the broader range of river basin development issues the federal government created a new type of river basin “authority,” which is a federal corporation dedicated to comprehensive natural resources planning and economic development.

The TVA model was partially successful, but it was never replicated in the United States.
for political reasons. Critics called it “grass-roots bureaucracy” and “creeping socialism,” dismissing its contributions toward more integrated water resources planning. Over time, its water and agriculture programs declined, and it devolved into a regional power utility with coal and nuclear as well as hydropower generation. However, the TVA has been emulated by river basin authorities in other countries, for example, the Damodar Valley Corporation in India.

One of the largest examples of river basin development outside the United States involves the Indus River basin in South Asia. TVA director David Lilienthal wrote an early article proposing a collaborative approach to river basin development between India and Pakistan. Eight years of negotiation led to an Indus Waters Treaty in 1960 based on the opposite principle, namely the partition of three tributary rivers to India and three to Pakistan. However, these negotiations also contributed to the establishment of a Water and Power Development Authority (WAPDA) in Pakistan, which is like a TVA on a national scale. WAPDA led the development of large-scale water storage and canal infrastructure to mitigate the effects of partition and develop the agro-economic base of the country. To support those planning efforts, the World Bank created a series of sophisticated Indus basin optimization models to evaluate water investment alternatives. The Indus basin models have more recently been used for climate impact assessment and adaptation.

The Rhine, Danube, Yellow, Nile, and Colorado Rivers have been laboratories for water development on their respective continents. Some experiments have failed. The Colorado River Compact of 1928 was based on flawed estimates of average basin discharge that used a short instrumental record that coincided with the wettest period in centuries. Other Colorado River experiments have had mixed effects. The shift from single-purpose single-means dams to multipurpose multiple-means reservoirs, such as the Hoover and Glen Canyon Dams, constituted an engineering advance, on the one hand, that proved damaging to environmental, social, and cultural resources on the other. Integrated river basin development came to mean a chain of jointly operated reservoirs and infrastructure rather than a process for progressive social and environmental management (Wescoat and White 2003). Over the course of decades, the Colorado acquired a complex “law of the river” that is fascinating in its institutional complexity, but insufficiently responsive to changing social values and climatic extremes.

Several river basin experiments are trying to address these constraints. For example, adaptive environmental management is an innovative method for transforming acrimonious battles over dam operations, fish and wildlife protection, indigenous water rights, and other issues. In large regulated water systems, adaptive management focuses on clarifying stakeholder objectives and supporting stakeholder design of restoration experiments to enhance multiple purpose management. It argues for learning by doing, citizen science, and a jointly scientific and political approach to water and environmental management. It has been attempted in the Columbia River, the California Bay Delta, Glen Canyon Dam, the Florida Everglades, the Danube River basin, and scores of other projects. After two decades of investment and experimentation, the institutional results are encouraging as they have helped shift water management from conflict toward collaboration, but the long-term ecological benefits remain to be established.

As noted above, Europe has emphasized transboundary basin planning in its Water Framework Directive. While most European countries have complied in principle and on paper, progress on the ground varies enormously. On the Mekong
River, the lower basin riparian countries of Thailand, Laos, Cambodia, and Vietnam established a Mekong Commission to pursue joint interests in river basin management. Although international conflict in the third quarter of the twentieth century constrained this vision of joint development, the Commission was able to sustain a basic level of cooperation even during war, particularly in the field of flood forecasting (Sneddon and Fox 2007). What was less anticipated was the acceleration of upper Mekong dam development by China.

Urban and community water management

The largest body of water management experience occurs at the local level, from households to communities and cities. Roman water law and waterworks provide some of the best documented ancient water systems, though there are ancient water texts and archaeological waterworks in many regions of the world (Mithen 2012).

Modern cities adopted different approaches to water management. For example, eighteenth-century Philadelphia created a public water department, while New York City granted a charter to a private company to provide water to its residents (Gandy 2002). The latter approach failed to provide a safe, reliable service when the company diverted its revenues to more lucrative nonwater investments, which led to a public utility approach in much of the United States. Interestingly, the late twentieth century witnessed a new wave of private water utility arrangements in European and UK cities that with strong public regulatory oversight of pricing, quality, and disconnection of low-income residents provide high levels of service (Bakker 2010). In England and Wales, the Water Services Regulation Authority (Ofwat) has undertaken this role since 1989. However, movements to privatize urban water systems in developing countries, often with the involvement of Western water companies, have generated greater controversies due to weak regulation and inequitable service, pricing, and provisions for low-income access to water for basic needs (Swyngedouw 2004).

Cities affect regional water quality and stormwater runoff in ways that were not initially taken up by water planners, but became major water management issues in the late twentieth century. Linkages between urban planning, conservation, and ecological approaches to stormwater and floodplain management are being widely adopted and becoming specialized fields of water research and professional practice.

While a majority of the world’s population now live in cities, greater water challenges continue to face those living in rural areas. Access to safe water and sanitation in rural areas consistently lags behind access in urban areas (WHO/UNICEF 2014). Even when they have adequate water supplies, rural communities rarely have the resources to regularly test water quality or treat water effectively at the source and at home.

While considerable progress has been made in community-based water management, communities are not inherently cooperative or equitable. For example, rural domestic water and sanitation challenges are highly gendered, with women responsible for most of the fetching, carrying, and managing of household water supplies at great physical, time, and economic cost. Inadequate local supplies can reduce hand washing and bathing, and thus aggravate the incidence of sanitation-related diseases, which women also manage (Harris 2009). These burdens are accentuated by male migration to cities for wage labor, which can shift the irrigation work formerly undertaken by men to women as well.
Groups who are marginal in economic or social status may be excluded from community water sources and decision-making processes that affect them. They may also be concentrated in areas subject to greater water hazards, for example, floodplains, hillslopes, open drains, and the tail end of canals that receive less reliable water (Mustafa 2013).

These unequal power relations at the household and community levels have parallels at each of the larger levels of governance discussed above. Cities are able to outbid and annex rural areas to obtain a larger regional share of scarce regional water supplies. Hydropower development generates more financial revenues than other river basin uses and garners a larger share of investment. Nationally and internationally funded river basin development historically prevailed over community-based watershed management, though that has been effectively challenged in places with strong water users associations. Upstream users have an inherent advantage over weaker downstream neighbors, though a powerful downstream riparian country such as Egypt can threaten to block development in a weaker upstream country such as Ethiopia. Negotiations to mitigate and adapt to the effects of climate change require an understanding of water management alternatives that connect these different levels of analysis, from the local to the global.

Future water resources

As in previous eras, new water resources are being discovered, developed, and managed in the twenty-first century. Water balance analysis has been extended from individual sites and watersheds to whole industries (e.g., bottled drinks) and their supply chains of growers, processors, and transporters. Some of these studies involve “water footprint” analysis which strives to optimize the amounts of blue water (surface supplies), green water (soil and plant water), and gray water (recycled wastewater) that are used (Hoekstra et al. 2011).

Tracking the flows of goods and services that have consumed water, known as “virtual water,” is a growing field of analysis. Virtual water trade was pioneered by geographer Tony Allan (2011). From early estimates of water used in imports and exports, virtual water research is incorporating increasingly detailed modeling of water use in value chains affected by climatic processes, product inputs and substitutions, prices, and regulatory policies (Konar et al. 2011). To better understand the linkages between hydrologic and social processes, a new subfield of socio-hydrology is developing. Like the field of ecohydrology, which bridges hydrologic and ecological sciences, socio-hydrology strives to link the social sciences, hydrologic sciences, and management (Sivapalan et al. 2014).

The turn of the current century may also be remembered for its discovery of new terrestrial and extraterrestrial waters. There have been recent freshwater discoveries on Earth, including those of the upper mantle and submerged coastal continental shelf. The quantity of water in the hydrous mantle (410–660 km below the surface) is disputed, though a recent study identifies the water-holding capacity of the hydrous mantle transition zone as approximately 1% by weight (Pearson et al. 2014). Water in the upper mantle has been associated with the lubrication of continental plates. Another recent discovery of fresh water has been found buried beneath the coastal continental shelves (Post et al. 2013). These deposits of meteoric groundwater could be substantial but are likely to be nonrenewable except where they continue to receive continental recharge. Should they be tapped, they could face problems of salinity ingress. All of these sources contribute to the Earth’s water budget, from
atmospheric humidity to precipitation, freezing, melting, infiltration, evapotranspiration, recharge, and runoff, each of which is challenging to measure and monitor at a global scale.

The origins of water on Earth is a related question. Although it is often stated, not quite correctly, that no new quantities of water are created on or lost from Earth, such processes take place on such long geological timescales as to have little bearing on water management. The origins of Earth’s water include early cooling of Earth’s mass 4 billion years ago, which created a gaseous atmosphere that held water in a vapor phase. Some of Earth’s water came from numerous water-bearing comet and meteoric collisions, some from Earth’s hydrous mantle, and some from early forms of photosynthesis. Some water input from space continues at an estimated rate of 0.01 km³ per year (Shiklomanov and Rodda 2003). Conversely, the Earth loses small amounts of water vapor through evaporation beyond its atmosphere (ranging from 0.03 to 0.27 km³/year, or approximately 0.1 km³/year). More water is lost than gained, and as the sun becomes hotter, the Earth’s liquid water supply is expected eventually to evaporate.

The search for extraterrestrial water, especially liquid water, is driven by intense curiosity about life beyond Earth. The results to date include identification of frozen water on some planets such as Mars, in the atmospheres of suns such as CW Leonis, in the brilliant tails of comets, and on diverse moons. These studies tap a range of evidence, from landform analysis to absorption spectroscopy for detecting water vapor and theories of habitable zones, which bear comparison with early geographic thought on the ecumene of Earth. Water vapor is believed to be present in most planetary and solar atmospheres. However, inferences about liquid water and oceans on other planets and moons remain tenuous to date. The utility of such knowledge for hydrologic management on Earth is also tenuous, but that is not to downplay the potential power of the imagination in adapting these astronomical sensing and visualization tools and ideas for water resource management on Earth.

SEE ALSO: Aquifers; Droughts and water shortages; Freshwater resources: past, present, future; Groundwater; Hydrologic cycle; Plant-water interactions; Rivers and river basin management; Water and climate change; Water conflicts; Water rights; Watersheds

References


Harris, L. 2009. “Gender and Emergent Water Governance: Comparative Overview of Neoliberalized Natures and Gender Dimensions of Privatization, Devolution and Marketization Gender.” *Place and Culture*, 16(4): 387–408.


Further reading


Water rights

Olen Paul Matthews
University of New Mexico, USA

Property rights are defined by relationships, and with water rights the relationships are particularly geographic. A water right defines the ways a water right holder can use water (human–environmental relationships) and the relationships within the network of right holders who have an interest in the same water. As a water molecule moves through the hydrologic cycle, it is subject to a sequence of rights that are geographically interconnected. The use of water at one place and one point in time can impact the quality and quantity of water available to other down-cycle right holders. Thus, water rights define not only the uses allowed, but the obligations to other users. Water rights allow users to transform landscapes by developing irrigated agriculture, supplying municipal and industrial water users, and generating hydro-electric power. Water rights allow uses that can dramatically change the hydrologic cycle’s integrity (quality and quantity). Some uses can cause environmental degradation while other water rights preserve in-stream flows and prevent harm to the cycle’s integrity. Understanding water rights is fundamental to understanding a region’s geographic patterns, especially where water is in short supply.

The nature of water rights

Water rights are complex and easily misunderstood primarily because water, unlike most other property, is naturally mobile. This mobility leads to two interrelated but fundamental misconceptions – all water rights are privately owned (it’s my water) and all rights are exclusive (I can do what I want with it). When most people think of a property right, they envision private ownership of land (their suburban home) or personal property (their car or computer). But water rights are not always private, at least not in the way other property can be, and water is not owned in the traditional way that term is used. Private property ownership has specific attributes (a bundle of rights) that generally include a right of exclusive possession, a right to use, a right to sell or transfer, and other items depending on the particular property conceptualization. Even when landownership is as complete as possible (the greatest number of sticks in the bundle of rights) the owner still has obligations to other property owners and is subject to government controls such as zoning, subdivision, and building regulations. Rights associated with water use are even more constrained. Water is mobile and therefore inherently shared, precluding a similar bundle of rights. Water is not subject to exclusive possession and may or may not be subject to sale or transfer. The closest water comes to being private property occurs when a consumer purchases a bottle of drinking water. All the attributes of private property apply to that bottle of water until a short time after it is consumed. Because exclusive possession is missing from the water rights bundle, a water right is usually defined as a usufructory right (a use right) rather than an ownership right. In addition, private water rights can be either attached to the ownership of land as in the riparian rights doctrine found in the eastern United States or
WATER RIGHTS

completely detached from landownership as in the appropriation doctrine found in the western United States.

Water rights can also be public and common property (a private right shared in common with others). Common property rights are a form of private right held by a limited number of interests. Landowners surrounding a lake or along a river may have a property right in the water body. This limited pool of users must in some way share in the use of water. Riparian property owners in the eastern United States and acequias (traditional irrigation districts in New Mexico) are examples.

Public rights can take three forms – the commons, regulatory rights, and government rights. With the commons everyone has free and open access. The commons is not the same as common property, which has a limited pool of users with access. Overuse of the commons can lead to private property rights, regulatory rights, or a combination of both. Regulatory rights are based on command and control laws and permits. Government rights in water are distinctive in that they generally serve some public purpose but are held in the name of the government. They can have substantial impact on other rights, as with the federal navigation servitude (protecting the navigable capacity of the nation’s waters) or federal reserved rights (providing water on federal reserved lands). To make things more complicated, a single molecule of water passing sequentially through the hydrologic cycle could conceivably have all these different property rights connected to it at different points in time.

When the commons is overused one solution is a system of private property rights allocating the resource. Theoretically the system of private rights will reduce conflict and prevent overuse. A private rights approach to water is complicated by its mobility and the strong interest the public has in water use. Any system of private water rights must define the relationships between competing users to prevent conflicts. But defining relationships requires a full understanding of the interconnections within the hydrologic cycle. At the time the common law system developed in England and the United States, the linkages between the elements of the hydrologic cycle were unclear. Private rights were based on landownership and the water adjacent to or under the land. As hydrologic science evolved, the connectivity between the cycle’s elements forced a redefinition of the relationships between users and a modification in the water rights structure.

Some parts of the hydrologic cycle fit easily within a private property regime but not all. For example, it would be difficult to create a private water right in soil moisture that is separated from landownership. Creating a private property right to atmospheric moisture would be difficult unless the water can be captured. Many private water rights allow uses that modify the integrity of the hydrologic cycle. The integrity of the hydrologic cycle includes physical (temperature, volume, movement/speed, etc.), chemical, biologic, and radiologic elements. It also includes social, cultural, religious, and political elements. Property rights associated with protecting the integrity of the hydrologic cycle are generally, but not always, public rights. Historically, many of these public rights were unregulated, leading to pollution and other overuse. Balancing these public and private interests is complex, and the complexities increase as the value of ecosystem services becomes better understood.

Different property rights regimes produce different outcomes in balancing the complexities found within water management systems. This gives policymakers choices. Unfortunately, most water right systems were established when hydrologic science was in its infancy; in addition, a different set of norms and values were associated with water use at that time. For that reason,
existing property systems are often unresponsive to current policy needs. The result has been a regulatory overlay modifying existing water rights.

Water use is controlled by a combination of property rights and regulations. Regardless of the system, certain fundamental elements must be present. Water rights are generally defined within a specific portion of the hydrologic cycle, but they need to recognize the interconnections within the cycle. The right should specify who can use the water and who can enforce the right. The public’s interest in water use must be balanced with any private rights. Rules for establishing a right and rules for using water (uses allowed, location of use, quantity, time period, etc.) are required. Terms of the entitlement need to be established. Is it permanent, temporary (for a period of years), or subject to termination under specific circumstances? Rules for selling, modifying, or transferring water are needed and must examine the potential impacts on third parties and the public. Rules are needed to allocate water between competing uses when shortages occur. Many of these water right elements are interrelated and complex. An understanding of water rights begins with the ways different water uses and users change the integrity of the hydrologic cycle. This provides a framework for understanding property relationships.

Water use and the integrity of the hydrologic cycle

Water uses differ in their impact on the hydrologic cycle’s integrity. Volumetric consumption, controlling flowing water (speed, direction, etc.), and using water for waste disposal can substantially alter the cycle. Consumptive uses include irrigation, municipal and domestic uses, and industry. Rights to consume water have a long history and are extremely important in arid areas. When most people think of water rights they generally envision a consumptive use. Irrigation is by far the largest, with 80% of the consumptive use in the western United States being attributed to it. In the West these consumptive water rights are private rights which can be sold under specific rules. Not all consumptive uses are allocated under a system of private rights.

Controlling water flow also changes the integrity of the cycle, but the issues that result are different. Flow can be controlled to capture the energy created by flowing water (water mills and hydroelectric power), for flood control (dams and levees), to store water for later consumption (municipal, industrial, and irrigation uses), for navigation improvement (dredging, channelization, and canals), and to drain wetlands. Some of these uses are subject to private rights but others require communal or governmental approaches. Two examples illustrate different property rights approaches. The energy from flowing water has long been used in milling operations. By 1800 Britain’s available mill sites were almost completely occupied. Mill dams backed up flowing water, interfering with the upstream mills, and diverted water deprived downstream mills of flow. Litigation was common and expensive (Getzler 2004). The result was a property rights system designed to protect flowing water to the mills. All riparian landowners shared a common property right to use the flow. Since the right was associated with landownership, it was considered a private right attached to the land. Flood control is the second example. Property owners along the lower Mississippi River built individual levees to prevent flooding. Inconsistent levee heights between adjacent land, or no levee at all, required a different scale of response to flooding. Local flood control districts were formed, but even these were inadequate for major floods.
WATER RIGHTS

State and federal projects were eventually implemented, but remnants of older systems are still in place. Flood control does not lend itself to a private rights approach.

Waste disposal, historically, was looked on as a public right open to all. Rivers were frequently used to dispose of industrial and mining wastes as well as untreated urban effluent. Downstream property owners frequently objected, especially as scientific knowledge increased, proving that public health dangers resulted from traditional disposal practices. In order to protect the public’s interest (right) in clean water, regulations placed limitations on disposal practices. Some regulations were absolute prohibitions and others allowed disposal through permit systems. These permits grant regulatory rights which in some circumstances could be considered private regulatory rights. Markets for these private rights can be created if policymakers so chose.

Other human uses have less dramatic impacts on the hydrologic cycle’s integrity. Navigation and recreation will be used to illustrate this point. Historically, free navigation has been an important public right, especially on waters where commercial navigation is possible. Interference with the right is often protected by government actions. For example, the Magna Charta (1225) prohibited the construction of fish weirs that interfered with navigation. In the United States the federal navigation servitude protects the public’s right of navigation. This federal constitutional power, exercised for the benefit of the public, requires permits before any actions that interfere with the navigable capacity of a water body. On the other hand, the eastern US riparian rights doctrine gives riparian landowners a private right to “warf out” (build a dock). This private right may not always be exercised since federal permission is required on waters subject to the navigation servitude. Recreational uses (fishing, swimming, boating, etc.) are considered public for many waters, but not always. For example, in Britain public fishing is only allowed in waters affected by the ebb and flow of the tide. Fishing in other waters requires the consent of the landowner. In the United States public access for recreational purpose varies by state. Some states have expansive public rights, but not all. Overuse can lead to the imposition of permit systems such as a fishing license or rafting permit. In some circumstances these may become marketable property rights.

Other uses (nonuses) have virtually no direct impact on the integrity of the hydrologic cycle. Humans enjoy/use water for scenic, aesthetic, and spiritual purposes. Historically, water rights systems did not address these uses, leaving them as part of the unregulated commons. Increasingly, some kind of public interest is being asserted. Indirectly, these uses are protected by establishing in-stream water rights or by establishing minimum stream flows. These rights may be public or private depending on circumstances, but most derive from some form of legislation. The federal Wild and Scenic Rivers Act and the Endangered Species Act are examples. Water left in-stream can provide valuable ecosystem services. Establishing rights to ecosystem services is an evolving theoretical topic.

**Legal relationship between water and the user**

The relationship between water and the user depends on the subset of the hydrologic cycle subject to the right, the nature/security of the legal entitlement, the rules for establishing a right, and the operational rules for using water. Different kinds of property rights may be associated with water in different parts of the hydrologic cycle. These rights may ignore the interrelationships that exist. Water rights do not
follow a single drop of water moving through
the hydrologic cycle. Water in a stream may
have one set of rules and groundwater another.
Atmospheric water may have no rules at all.
To illustrate, a single drop of water is followed
through the hydrologic cycle. As atmospheric
moisture, the water drop is an unregulated
commons, at least as far as water quantity is
concerned. Cloud seeding may require a permit
or a license for cloud seeders, but rights are
not granted to the water produced. Proving the
causal connection between the cloud seeder’s
actions and a specific molecule of water is dif-
ficult, as is proving harm from cloud-seeding
operations. From a water quality perspective,
the water drop is subject to regulatory rights.
Statutory permit systems are designed to prevent
acid rain and other atmospheric pollution that
may impact the integrity of the hydrologic cycle.
Once the raindrop hits the surface some of it
will infiltrate, becoming soil moisture, and some
may run across the surface, eventually reaching
a watercourse. Soil moisture is a land right and
from a practical viewpoint is really usable only
by the landowner. If the water drop infiltrates
reaching an aquifer, water rights can be estab-
lished because the water can be extracted and
separated from the land. Once the water reaches
a stream it may also be captured, and private
water rights can be established. Diffused surface
water or runoff may be treated as a land right,
but it is possible to separate the water right from
the land right. Generally, private water rights
are limited to groundwater aquifers and surface
water in streams, lakes, and rivers. Potential
conflicts exist between these two kinds of rights
when established on an interconnected source.

Entitlement refers to the right’s legal status
/license, permit, title, permanent, temporary,
etc.). Security of entitlement refers to the potential
for losing a right. Security of entitlement is
not the same as actually receiving water for
use (security of delivery). Ownership of land may
be the basis of a permanent entitlement, for
example, when the federal government has a
water right on reserved land. If a water shortage
occurs, then any right established prior to the
federal act of reservation will have a preference
for delivery. Thus, the federal right is still in
place, but water may not be delivered until the
shortage is alleviated. Ownership of riparian land
in the eastern United States contains a similar
permanent entitlement, but the rules for sharing
do not guarantee a specific volume of water. The
riparian landowner has an entitlement (right)
equal to all other riparian landowners. In the
western United States the entitlement is based
not on landownership but on the acts of appro-
priation. In times of shortage prior rights have
a preference over those established subsequently,
meaning that delivery is not guaranteed. Title
still exists and delivery will occur after all prior
rights have been satisfied.

In the West a water permit or license may grant
a right to use water, but this permission does not
necessarily mean the title is secure. Water rights
must be adjudicated through a judicial or admin-
istrative process. The adjudication process exam-
ines all the use claims in a watershed to determine
their validity and establish their relative prior-
ity. Many western states are still completing the
adjudication process. This process can invalidate
rights with existing permits.

Although some entitlements are permanent,
others are limited by time or are conditioned in
some way. For example, many regulatory rights
have permits that must be renewed after a period
of years. Renewal may be based on meeting spec-
ified conditions that take into account changing
circumstances. Other rights can be lost because of
actions by the right holder. In the western United
States a water right may be lost for nonuse. “Use
it or lose it” is the common saying. Also, the use
must be beneficial, meaning that wasteful uses
WATER RIGHTS

will not qualify. In riparian rights states, title can be lost if a lot that was once riparian (adjacent to water) is subdivided and sold. Any portion that no longer touches the water body loses the riparian right. Public rights exercised in an unregulated commons can also be lost. Recreational access, for example, can also be lost through overuse. Too many rafters could lead to a permit system with a limited number of permits available.

The rules for establishing a water right depend in part on the kind of right involved. Some rights are an attribute of owning riparian land and require only ownership of the land (eastern United States riparian right). Federal reserved rights require the federal government to reserve federal land for a specific purpose. This act of reservation creates a water right sufficient to accomplish the purposes of the federal reservation. Unregulated public rights (the commons) exist without a process for establishing rights. For this reason overuse is possible. When the commons is regulated, rights are established by rules derived from the regulations. Some of these regulated rights can take on attributes of private property. A good example of this is the appropriation doctrine in the western United States.

Under the appropriation doctrine a water right is established by intentionally diverting water from a stream and applying it to a beneficial use. The earliest versions of this doctrine resolved conflicts between users by creating common law property rules. These rules separated the water right from the ownership of land. Miners during the California gold rush did not own the land and, prior to the passage of federal mining laws, were trespassers. The miners also wanted to move water from one watershed to another. The common law riparian rights doctrine was inadequate for the miner’s needs. The result was a new set of common law rules – the appropriation doctrine. The basic principles of these rules were incorporated into the early water codes. By the time the Reclamation Act (1902) passed, many states also required permits. Water still had to be diverted from a stream and applied to a beneficial use. Permits could be denied if another water right would be harmed or if it was in the public interest. Public interest in 1900 does not mean the same today, but at that time included uses that were economically beneficial.

The operational rules are the primary element of any water right system. They define the uses allowed, quantity, location, time, cost, and conditions. Some water rights have very specific and narrow uses such as a right to navigation (federal navigation servitude). Others uses are based on vague standards. Federal reserved rights are based on the purpose of the reservation, which for national forests includes timber harvest and watershed protection. This vague description makes it difficult to quantify the volume associated with the right. In riparian rights states, the use must be reasonable, and in appropriation doctrine states the use must be beneficial. Determining reasonableness requires a balancing process between competing riparian rights. Beneficial uses were traditionally limited to economic (profit-oriented) uses, but many states have expanded definitions to include in-stream uses that are environmentally beneficial or that provide recreation and aesthetic benefits. If the right is based on a permit, it may specify in detail the uses allowed. The other operational rules may be very general or they may be clearly specified in a permit. A federal Indian reserved right has a volume of water associated with it that is sufficient to irrigate all the practicably irrigable acreage on the specified reservation. This standard was established by the Supreme Court in 1964, requiring an agricultural survey of each reservation to reach a specific volume.

Problems can also occur in defining the operational rules. Quantity could be based on a diversionary entitlement (amount taken from
the stream) or a consumptive use (the volume consumed). Location includes place of use and point of diversion. If the quantity is measured at the place of use rather than the point of diversion, incentives for efficient delivery systems may be missing. Changes in place of use, point of diversion, point of return flow, time of use, or any other element of the operational rules can impact other water rights. These third-party effects need to be considered before making any change in the operational rules for a particular right (Gould 1988). Operational rules also include any costs assessed to the water use. Generally, users only pay for the cost of delivery and not for the value of the water. Often the cost of delivery is subsidized, especially for irrigation. Operational rules can be conditioned to meet efficiency standards or to conserve water.

When people think about a water right it is often the relationship between the user and the water that first comes to mind. This brief discussion gives the flavor of that relationship. More complex are the interrelationships within the network of users.

Defining relationships between users

The interconnected system of water rights requires a system of rules to resolve conflicts. Conflicts occur for many reasons. Water shortages, controlling water flow, and waste disposal will illustrate the range of rules used in defining relationships.

Water shortages

A drought can cause a water shortage that interferes with a water right. The security of the user’s delivery is a critical element in determining the value of the right. If a right holder has a low security of delivery in an area prone to drought, they should not plant an orchard requiring regular water deliveries. Property rights systems use three basic approaches for allocation water in times of shortage – preferences, flexible criteria, or proportionality. A fourth approach, absolute ownership or rule of capture, allows the right holder to do what they want with water; the consequences to others are not considered. Although this type of right has been claimed in international disputes and in disputes between US states, it is not the current law. The doctrine still has some remnants when applied to groundwater. Nominally, several US states claim that if a landowner pumps groundwater they can use it as they please.

Preferences can be based on use, time, or region (spatial). With a use preference the specified use receives water first in time of shortage. This is the type of preference most people think of when the term “preference” is used. If water is in short supply, then some uses should get water before other uses. Use preferences frequently recognize the importance of domestic uses such as drinking water. Other preferences can be based on important economic needs. For example, Idaho grants mining a preference over other uses in organized mining districts. In New Mexico’s traditional acequias, orchards have a customary preference over annual crops. Preference systems allow policymakers to rank the uses that will receive water in time of shortage. Almost all systems have some kind of preferences. An international human right to water will create a preference for drinking water and sanitation.

In appropriation doctrine states, a temporal preference, based on the date the water right was established, sets the order in which users will receive water in times of shortage. “First in time, first in right” is the mantra. An early priority (right established before others) thus has a high security of delivery, making the right more valuable than one with a recent priority.
Strict interpretation of the doctrine would treat all beneficial uses the same, whether the use was for a hospital or an irrigated pasture. The results can be inequitable and economically inefficient. To overcome this problem, water can be rotated between users so that all get a share at different points in time. This has always been a common practice in irrigation districts. Additionally, some western states have limited use preferences. For example, municipal or domestic uses may be granted a temporary statutory preference if a shortage occurs, but compensation may be required. Accommodation and augmentation plans, as well as contracts, have been used to avoid a strict application of priorities.

Spatial preferences limit the place water may be used. These preferences are often stated as area-of-origin or local public interest preferences. Some of these preferences prevent water exports from the preferred area. The protected spatial unit can be a watershed, a county or other political unit, or an undefined local place. For example, if a local economy develops around irrigated agriculture and the water is transferred to another location, substantial harm can occur. Farms will no longer be viable, businesses will be closed, the tax base will erode, and depopulation will occur. The Owens Valley in California illustrates this problem. Water once used for irrigation now goes to Los Angeles. In response, California statutes place area-of-origin limitations on water movement.

In the United States some states have tried to limit water exports to protect their own economies. These antiexport statutes have constitutional problems. The commerce clause creates a level playing field for the movement of goods and services between states. Because water is considered an article of commerce, spatial restrictions must pass the legal tests developed under the commerce clause. Local public interest and area-of-origin statutes are also designed to protect economic interests which bring them into direct conflict with the commerce clause.

Flexible systems for resolving disputes require balancing multiple factors. Balancing can create a degree of uncertainty when looking at security of delivery, because balancing is done on a case-by-case basis. Balancing requires looking at other users and uses within the system. A good example is the riparian rights doctrine based on reasonable use, as found in some eastern states and the reasonable use doctrine applied to groundwater. Flexible doctrines have also been used in disputes at different scales as seen with equitable apportionment (disputes between US states) and equitable utilization (international disputes). The riparian rights doctrine requires a more detailed examination because of the misconceptions describing it in the literature.

The riparian rights doctrine grants an equal right to all riparian landowners contiguous to the water body. This equal right refers to the entitlement each landowner has but not to the quantity of water delivered to them. An equal right does not mean an equal or proportionate share of water. A riparian right (entitlement) grants adjacent riparian landowners a right of access to water, a right to build a wharf, a right to use water without transforming it (the hydrologic cycle’s integrity is maintained), a right to consume water, the right to accretions, and ownership of the bed of non-navigable waters. Many of these rights are nonconsumptive, and some early court decisions implied the doctrine was designed to protect nonconsumptive uses under the natural flow doctrine. The natural flow required water to be left in the stream in its natural condition. This doctrine prevented consumptive uses and was modified to allow consumptive uses that are reasonable (the reasonable use doctrine). Reasonableness is an open-ended list of social, economic, environmental, religious, and cultural uses that benefit
WATER RIGHTS

the adjacent land. If insufficient water is available for all coequal riparian landowners, a balancing test is used where the harms and benefits of each use are evaluated. If some scheme for sharing the resource is possible, the water will be shared between users. If sharing is impossible, one user could obtain the entire available supply. If the use is socially desirable, is economically beneficial, and benefits a large number of people, that use may be preferred over another that does not fit this degree of desirability. A new use on a stream may be preferred over an older use. In addition, values change over time. What was considered the most valuable use in 1900 may not be today.

Where reasonable use is the legal test, security of delivery changes with the circumstances. This applies to both surface waters subject to riparian rights and groundwater where the reasonable use doctrine applies.

When the delivery of water is based on flexible standards, the results can be equitable but unpredictable. What one person thinks is fair may be inequitable from another perspective. In addition, equity, fairness, or reasonableness can change with time.

Proportionality is the last method discussed under security of delivery. Proportionality means that water is divided based on specific criteria. The criteria could be based on population, amount of land owned, number of shares in an irrigation district, or a percentage of flowing water. This applies to both surface waters subject to riparian rights and groundwater where the reasonable use doctrine applies.

US water rights controlling flow have roots in both common law and constitutional/statutory law. Many of the common law rights developed as a result of conflicts between water users. The constitutional and more recent statutes generally protect some public right. Rights associated with flowing water occur in two parts of the hydrologic cycle – streams/rivers (watercourses) and diffused surface water.

Diffused surface water (runoff) is legally different from the same water molecule once it reaches a stream. In western appropriation doctrine, water can be captured by the landowner without obtaining a water right. Once the water reaches a watercourse, defined by having a bed, banks, and reasonably regular flow, the appropriation doctrine applies. Check dams, contouring the landscape, and planting vegetation to inhibit runoff are considered to be rights associated with landownership. Western
WATER RIGHTS

states are beginning to connect surface water rights with groundwater rights, but diffused surface water is excluded. Colorado is an exception and has recently extended the connectivity within the hydrologic to include harvesting rainwater from rooftops. Diffused surface water also creates drainage issues. The problem here is too much water.

Three different common law doctrines developed in drainage cases. The common enemy rule allows landowners to take whatever measures necessary (speeding up flow, erecting barriers) to prevent harm to their land. This absolutist form of property right ignored potential harm to neighboring property. The natural servitude rule is the opposite approach. Changes in a drainage pattern that harm another’s property, upslope or downslope, violate the doctrine. Both doctrines create inequities so the trend today is a reasonable use rule. This flexible doctrine allows a balancing of harms and benefits between competing property interests.

Flow conflicts within a water course can create tension between public and private rights. The appropriation doctrine allows a river to be completely consumed, leaving the river bed dry. But, if the water is subject to the federal navigation servitude, the public’s right to navigation can be protected. Public rights to navigation and the protection of ecosystem services are designed to keep water in-stream and to protect the integrity of the hydrologic cycle. The navigation servitude is based on a federal constitutional power and has been extended to the tributaries of navigable waters. Consuming or diverting a tributary’s water can be prohibited if the action affects the river’s navigable capacity. This federal power has been used to stop irrigation projects and can impact a water right established under state law. In addition, dredging and filling navigable waters requires a permit and can be prohibited if navigation will be harmed. More recently, statutes have been designed to put more water in-stream for a variety of purposes. Some western states have modified the diversion requirement of the appropriation doctrine to allow in-stream appropriations. In other instances minimum stream flows are created that keep a specific volume in-stream. This is done either through reserving a minimum flow or by rules for dam management. Many of these minimum flow requirements stem from the Endangered Species Act. Competing uses may be curtailed in order to provide water needed for species survival.

Waste disposal

Water quality disputes have a long common law history. Doctrines of nuisance, negligence, and trespass have all been used by individual property owners to protect their property from harm. Public interests have also been protected through a public nuisance doctrine. The riparian rights doctrine of reasonable use includes water quality. Reasonable use and nuisance, public or private, require a balancing of the harms and benefits of the various interest holders. Thus, a property owner who wanted an injunction to stop a pulp mill had to show that the harm to their property was greater than the benefits of the mill. Historically, courts did not favor the individual very often. The best bet for the property owner was to prove a trespass where damages might be possible.

Common law solutions did not work very well because individuals had to bring a law suit. In addition, as the harm to public health and the environment was better understood, pressure increased to change the approach. The result is a regulatory structure. Point source pollution is controlled through a permit system while nonpoint sources are controlled through land-use planning.
Responding to challenges in water management

Although property rights define relationships, conflicts are inevitable. Conflicts occur for many reasons and some are difficult to resolve. This is especially true because of the fragmentation within the system and enforcement problems. In addition, property rights systems do not always have good mechanisms for responding to change.

Enforcement

The first step in enforcing a water right is to determine the source of the right. Not all water rights are equal. When right holders compete for the same water, they may win or lose depending on the source of the right. The federal navigation servitude trumps state water rights. Water is considered an article of commerce under the US Constitution, giving the federal government regulatory powers over water. That power was exercised with the Clean Water Act and many other federal environmental statutes. Under the Supremacy Clause, the exercise of a federal constitutional power is supreme over conflicting state laws. At the same time, states have created property rights in water. These rights are constitutionally protected, requiring compensation to be paid if the property is taken. The question of federal regulations taking water rights is unresolved.

Enforcement also requires the identification of the user or beneficiary of a right, the right's title holder, and the party that can enforce the right. These are not always the same. For example, the title to the federal navigation servitude is in the federal government, and the public is the beneficiary. The federal government can enforce the servitude and in some circumstances users may have legal standing to bring a lawsuit. If individual right holders must take action to enforce a right, enforcement may never occur. Inertia must be overcome before action is taken. If court action is required for enforcement, it could be expensive and beyond the financial ability of the right holder. Public rights are problematic because of the difficulty of identifying parties that can act in the public's name. In some instances individuals or public interest groups have standing, which is a legal concept that will allow them to bring a lawsuit. Even if a group or individual has standing, inertia must still be overcome and costs can be high. Government agencies are generally given power to enforce public regulatory rights. At times statutes may give standing to individuals or public interest groups to compel government action if they fail in their enforcement obligation. The Clean Water Act is an example. In some instances enforcement may be impossible because of insufficient budget or lack of legal clarity.

Fragmentation

Fragmentation occurs because the system of water rights does not integrate the hydrologic cycle, water rights are subservient to land rights, and water quantity is not integrated with water quality. The appropriation doctrine is a good example of an unintegrated approach to the hydrologic cycle. In the early 1900s most appropriation doctrine states had a system for establishing water rights for water in a watercourse. Groundwater rights came much later, with separate codes for establishing rights even though the basis was still the appropriation doctrine. Different priority systems were in place, which was fine as long as surface water and groundwater were disconnected. As the interconnections became better known, courts were asked to enforce priorities for the interconnected rights. Colorado and New Mexico were the first
WATER RIGHTS

states to do so. Except in Colorado, where rainwater catchment is included, the rest of the hydrologic cycle is ignored.

In addition, water in many parts of the hydrologic cycle is controlled by land rights. These land rights are dominant over water rights and are exercised without regard to them. Diffused surface water can be captured in appropriation states without obtaining a water right. Land surfaces can be modified to prevent soil erosion, while at the same time increasing infiltration. No water right is required even though water flow may be decreased in an adjacent stream. Orchards can be planted which stick their roots into an aquifer without obtaining a water right. The trees can consume water that would otherwise move downgradient to a well with an existing right. The water right holder would have no recourse because the orchard does not require a water right. The dominance of land rights over water rights makes water planning difficult. Land-use planning and water planning are generally done at different scales and by different planning bodies. Integrating the two planning processes is difficult.

Managing water quality and water quantity is also fragmented. The Clean Water Act is federal, while state law establishes water rights controlling water quantity. Even when states administer the Clean Water Act, it is usually a different agency than the one administering water rights. Fragmentation makes conflicts between water right holders inevitable.

Responding to change

New demands for water and changing preferences for water use require a water rights system capable of reallocating water from one use to another. Two reallocation methods can be used – markets and regulation. As seen from the discussion, regulatory rights have been used to prevent harm to the public’s interest in water. Once in place, these regulatory rights may not have sufficient flexibility to respond rapidly to new challenges. Regulatory rights are generally a response to a particular water problem such as pollution. As seen earlier, the regulation may ignore the integrated nature of interlocking rights while solving the specified problem. Piecemeal problem-solving, although typical, need not be the only regulatory approach. Integrated response to change requires a long-range planning process.

Markets are another way of reallocating water in response to changing circumstances. Markets exist for appropriation doctrine rights in the West and could be developed for regulatory rights with permits. Because of the high degree of connectivity between water uses, a free market in water would be inappropriate. Western water rights cannot be sold if other water rights will be harmed or if there is injury to the public’s interest in water. Both these factors have been in place since the codification of water rights 100 years ago. Even so, for most of this time the evaluation of the public’s interest has been perfunctory. As can be seen by the broad nature of the public’s interest, a thorough evaluation should occur. Protecting the public interest requires more market rules than are in place in most water rights systems. Markets without rules will not work because of the substantial number of interrelationships that exist within water rights systems. The rules should be designed to achieve specific policy goals. Identifying such goals requires a planning process.

SEE ALSO: Ecosystem services; Hydrologic cycle; Legal geography; Property and environment; Water and human rights; Water resources and hydrological management
References


Further reading


Water security

Mark Giordano
Georgetown University, USA

Water security as a concept emerged in the early twenty-first century as a new rallying cry of the epistemic water community and quickly found its way into water policy and practice. In contrast to almost no mention of it before 1990, there are now (at the time of writing) over 400 peer-reviewed journal articles from across the social, natural, and medical sciences using the term (Cook and Bakker 2012), with the majority published since 2010. The concept is now so ubiquitous among water scientists and practitioners that it regularly forms the basis of major conferences (e.g., Water Security, Risk and Society Conference, Oxford, 2012; Securitization of Water Discourse Conference, Jerusalem, 2013; Stockholm Water Week, 2013; World Water Week, 2015) as well as of books and journal special issues, all of which examine what water security means and whether it can be achieved.

The concept of water security has made an equally rapid rise in policy and practice, with multiple organizations now using it as a basis for awareness raising, project design, and bureaucratic structuring. The United Nations increasingly links its programs, such as those related to climate change, with water security, and has framed the Global Environmental Facility’s water strategies around the idea. National governments have taken similar approaches. In the United States, for example, the Environmental Protection Agency now has a Water Security Division; the US intelligence community organized its 2012 assessment around water security; and the Agency for International Development’s 2013–2018 water strategy has set the achievement of water security as its primary goal. Nongovernmental organizations (NGOs) are following similar paths, with the Nature Conservancy, for example, highlighting water security as a key challenge and Harvard University establishing a Water Security Initiative.

While there is widely shared interest in water security as a concept within organizations ranging from grassroots NGOs to the world’s largest military and the United Nations, there is not always a shared understanding of its meaning. One end of the spectrum thinks of it in terms of existential threats to nation-states, the other in terms of individual welfare. What led to the emergence of water security as a concept? What are the purposes to which it is put? How is it defined? This entry examines these questions and summarizes the current thinking on water security.

Origins of water security

Until the early 1990s, reference to water security was limited to narrow questions related to legal allocation entitlements (e.g., “Is my water right secure?”). Since then the concept has extended in two distinct directions that reflect two separate sets of changes in conditions and understanding. The first is related to climate change, with water security, and has framed the Global Environmental Facility’s water strategies around the idea. National governments have taken similar approaches. In the United States, for example, the Environmental Protection Agency now has a Water Security Division; the US intelligence community organized its 2012 assessment around water security; and the Agency for International Development’s 2013–2018 water strategy has set the achievement of water security as its primary goal. Nongovernmental organizations (NGOs) are following similar paths, with the Nature Conservancy, for example, highlighting water security as a key challenge and Harvard University establishing a Water Security Initiative.

While there is widely shared interest in water security as a concept within organizations ranging from grassroots NGOs to the world’s largest military and the United Nations, there is not always a shared understanding of its meaning. One end of the spectrum thinks of it in terms of existential threats to nation-states, the other in terms of individual welfare. What led to the emergence of water security as a concept? What are the purposes to which it is put? How is it defined? This entry examines these questions and summarizes the current thinking on water security.

Origins of water security

Until the early 1990s, reference to water security was limited to narrow questions related to legal allocation entitlements (e.g., “Is my water right secure?”). Since then the concept has extended in two distinct directions that reflect two separate sets of changes in conditions and understanding. The first is related to climate change, with water security, and has framed the Global Environmental Facility’s water strategies around the idea. National governments have taken similar approaches. In the United States, for example, the Environmental Protection Agency now has a Water Security Division; the US intelligence community organized its 2012 assessment around water security; and the Agency for International Development’s 2013–2018 water strategy has set the achievement of water security as its primary goal. Nongovernmental organizations (NGOs) are following similar paths, with the Nature Conservancy, for example, highlighting water security as a key challenge and Harvard University establishing a Water Security Initiative.

While there is widely shared interest in water security as a concept within organizations ranging from grassroots NGOs to the world’s largest military and the United Nations, there is not always a shared understanding of its meaning. One end of the spectrum thinks of it in terms of existential threats to nation-states, the other in terms of individual welfare. What led to the emergence of water security as a concept? What are the purposes to which it is put? How is it defined? This entry examines these questions and summarizes the current thinking on water security.
The securitization of water

Since at least the early 1990s, there have been growing fears in the popular and academic press that disputes over waters shared by two or more countries could lead to conflict and even “water wars.” The increasing emphasis on water and international conflict fit within a broader change already occurring in international relations on the conception of national security. National security had traditionally been viewed as safety from external military threats, but by the 1980s the idea began to take hold that national security could be endangered not only by foreign military power but also by internal disruptions, including natural disasters, a degrading environment, and resource scarcity. This new line of thinking “securitized” the environment, and its proponents highlighted the potential interconnection between resources, social turmoil, and intra- and international conflict. In fact, water is perhaps second only to oil as the resource most associated with conflict and most often “securitized.”

The water securitization discourse now takes at least three distinct forms. Fischhendler (2015) describes two, which he terms strategic and tactical. Strategic securitization relates to internationally shared waters and the environmental, economic, and political interdependencies they create, which are associated with direct conflict between states and the classic idea of water wars. Tactical securitization relates to the connection of low-politics aspects of internationally shared waters (e.g., data sharing) between states and other, high-politics issues. Tactical interconnection is not portrayed as leading to interstate conflict but as a possible exacerbating factor in existing interstate conflicts or a new avenue for international cooperation.

A third form of securitization discourse relates to the secondary or spillover impacts of intranational water outcomes on the relations between states. If natural disasters, including droughts and floods, increase the risk of violent internal conflicts, those conflicts may spill over to neighboring states. This logic has been applied to the crisis in the Darfur region of Sudan, where drought has been linked to civil war, and then to fighting with neighboring Chad. Others believe the relationship between natural disasters and civil unrest is less clear and may in fact run in the opposite direction. In Pakistan, for example, while the large-scale floods of 2010–2011 and slow government response have been linked to a weakening of the government and a rise of fundamentalism, other evidence actually suggests that the floods brought increased credibility to the state. Given Pakistan’s geopolitical context, there are direct implications for regional stability (e.g., “security”) whichever view is correct.

Pushing this view to its logical extreme, if water impacts a state’s internal security, and the security of that state is held to be of strategic importance to a second state, wherever its location, then water issues anywhere in the world can be considered as security threats in any other part of the world. Aid or other resource transfers and interventions are then justified. As the US Agency for International Development’s 2013–2018 water strategy puts it, “Ensuring the availability of safe water to sustain natural systems and human life is integral to the success of the development objectives, foreign policy goals, and national security interests of the United States” (USAID 2013, 1).

Water security as human security

While one epistemic and policy community has focused on water security in terms of international relations, another has, separately but in parallel, considered it as a part of a more broadly defined human security. In 1994 the United Nations Development Program published a landmark report that de-emphasized
both traditional military security as well as gross national product as measures of welfare. The resulting “freedom from fear” and “freedom from want” approach incorporated food, health, environmental, and other “securities” within an overarching framework of human security.

In the initial conception, water was subsumed under other securities, most notably through the placement of potable water access under the environmental security rubric. The role of water within the human security discourse grew over time, though, in particular with increased understanding of the global challenge of ensuring sufficient freshwater supplies to meet growing agricultural needs, by far the largest consumer of water. The 1997 Congress of the International Water Resources Association, for example, focused on water scarcity as a critical issue in food security. While for some water scarcity continued to be viewed as a factor in ensuring human security through food supplies, for others it morphed into water security and became an end in itself. For example, the “Ministerial Declaration of The Hague on Water Security in the 21st Century,” presented at the 2000 World Economic Forum, recognized the role of water in basic needs, the food supply, and ecosystem maintenance, but made water security its goal. As stated in the document “together we have one common goal: to provide water security in the 21st century” (World Water Council 2000, para. 1). The Food and Agriculture Organization of the United Nations similarly declared that water security should be the main goal inspiring the international agenda for the twenty-first century (FAO 2000).

As water security became an end rather than a means, it took on different meanings for different authors and practitioners with differing backgrounds and goals. Cook and Bakker’s (2012) review identifies water availability (focused on drinking water and sanitation), human vulnerability to hazards (e.g., flooding and possibly drought), human needs (focused on food production and availability), and sustainability as four key water security themes that have emerged in recent years. A review by Clement (2013) also found substantial variation in how the term “water security” has been defined and utilized.

Water security has also found its way into another new discourse, that of the water, food, and energy nexus. Allouche (2011) described water and food as two separate “securities” but also described how they have been interconnected, and others have highlighted an even broader web of securities and how water security is a function not only of local and international water policies but also of policies outside the sector. Recent high-level conferences in Bonn (2011) and a follow-on in Chapel Hill (2014) have further promoted and put forward ideas on the interconnections between water, food, energy, and other “securities.”

Usage: synonyms, instruments, and agendas

As the concepts of water security have grown and evolved, so too have the uses to which the term is put. This section describes three usage themes that can be discerned from the literature and as applied in the organizations that use the term.

Water (in)security as a synonym for water scarcity

Despite the substantial discussion of water security, in practice the term is often used primarily as a synonym for water scarcity or other water problems. In other words, “water security” is simply employed to signal that a paper or an organization is going to address issues, but it is not pursued further as an idea in itself. An illustrative textual example is the 100-page FAO
WATER SECURITY

publication, *New Dimensions in Water Security* (2000). The document uses the term “water security” only three times. It is used in the title and once to refer to crops having enough water in the growing season. The third use occurs where one chapter mistakenly refers to another as being titled “The Turn of the Screw – Social Adaptation to Water Security” when the actual title is “The Turn of the Screw – Social Adaptation to Water Scarcity.” This small mistake suggests a perceived interchangeability between the terms “scarcity” and “security” in common usage. An institutional example of how the term “water security” is used to signal a focus on water problems is provided by the United States Agency for International Development’s Water and Development Strategy 2013–2018 (USAID 2013). The opening paragraphs of both its introduction and conclusion highlight water security as the strategy’s key goal. However, the body of the document does not use the term even once. Instead it focuses on the range of water problems USAID hopes to address.

Water security as a new agenda?

A number of authors have already questioned whether the water security concept reflects an evolution in thinking about water problems or a paradigm shift, or is simply a new buzzword to replace old concepts. For example, Clement (2013) showed how the water for food discourse has shifted away from the earlier ideas of water productivity and toward water security. However, while the shift expanded the water productivity debate, she argues that its water security manifestation still focuses on technical issues and still fails to take into account the power and politics that are the key to water outcomes. A perhaps more common discussion is whether water security is simply a rebranding of integrated water resources management (IWRM). The chief institutional proponent of IWRM, the Global Water Partnership, was actually an early advocate of the water security label (GWP 2000) and continued to promote it as a way to achieve cross-sectoral cooperation in water management, a goal normally associated with IWRM itself.

How these changes are reflected in practice can be seen in the reframing over time of the International Water Management Institute’s (IWMI) recent strategic planning documents. As a member of the Consultative Group on International Agricultural Research, IWMI’s main work relates to water as a means of agricultural development and poverty reduction. This is consistent with the organization’s official vision of “Water for a Food-Secure World,” as articulated in its 2004–2008 strategic plan (IWMI 2004), which highlights IWMI’s contribution to the food security agenda through the promotion of two paradigm shifts in water management: increasing water productivity at the basin level and focusing on integrated natural resources management. To support the change, a new research theme was established which emphasized water productivity improvement, basin-scale approaches, and cross-sectoral (i.e., integrated) thinking on water management. While the plan mentions food and environmental security multiple times, water (in)security is mentioned only once, and then as a synonym for water scarcity (“water poverty or water insecurity”).

In the follow-up 2009–2013 plan, the vision of water for a food-secure world remains (IWMI 2009). However, the path for achieving the goal has changed, with IWRM elevated to a “guiding principle” for all work and water productivity improvement no longer having an explicit home but still highlighted throughout the document as a key goal. Food security, and water’s role in it, continue to be frequently mentioned but water security itself is mentioned only once.
in terms of security in “water access rights” for the poor. In the current 2014–2019 plan (IWMI 2014), IWMI’s official vision has been changed to “A Water-Secure World,” and thus from water as a means to achieving food security to water security as an end in itself. Water productivity is no longer mentioned, though agricultural productivity and water’s contribution to it are, and IWRM has disappeared entirely.

Water security as an instrument

Fischhendler (2015) has described how (traditional) securitization of any issue, including water, can be used as an instrument to increase public awareness and perceptions of the important issues and to create a sense of urgency to mobilize funds. Allouche (2011) has demonstrated in the case of water how scholars and policymakers have used “water security” to raise the profile of water through its linkage with violent conflict at local, national, and international scales. Similarly, Clement (2013) describes how water has been viewed in (human) security terms as an instrumental means to move it higher up the international policy and aid agendas, for example to justify development bank investment in large-scale water storage structures.

How the traditional and human security aspects of water securitization have been combined and used instrumentally can be seen in the United States’ Environmental Protection Agency (EPA). The EPA’s main purpose with respect to water is to protect its quality as required by the 1972 Clean Water Act. Its responsibilities included eliminating high releases of industrial pollutants into water, preventing additional water pollution, and ensuring that surface waters meet the standards necessary for sports and recreation. However, in a post-9/11 world, perceived sources of water quality threats expanded to include terrorism. The EPA then took on the role of the agency in charge of water and water infrastructure protection from traditional and now terrorist threats. This change raised the agency from one working primarily with domestic polluters and users of water to one working at the core of US diplomatic and military agendas. The EPA’s mandate was even expanded by Homeland Security Presidential Directive 9 to include international intelligence-gathering, since the agency must “develop robust, comprehensive, and fully coordinated surveillance and monitoring systems, including international information, for … water quality that provides early detection and awareness of disease, pest, or poisonous agents.”

Definitions, scales, and methods for assessing the state of water security

Since the concept of water security now has multiple meanings and uses, it is not surprising that it has been formally defined at different times in many different ways, as illustrated by the following quotations, which have been selected to illustrate key variations and to highlight those most commonly cited or promoted.

A condition in which there is a sufficient quantity of water, at a fair price, and at a quality necessary to meet short and long term human needs to protect their health, safety, welfare, and productive capacity at the local, regional, state, and national levels. (Kaplowitz and Witter 2002, 379)

[An overarching goal in which] every person has access to enough safe water at affordable cost to lead a clean, healthy and productive life, while ensuring the environment is protected and enhanced. (GWP 2000, 12)

[Water security] involves the availability of water in adequate quantity and quality in perpetuity
to meet domestic, agricultural, industrial and ecosystem needs. (Swaminathan 2001, 35)

[Household water security is] the reliable availability of safe water in the home for all domestic purposes. (WHO 2003)

The state of having secure access to water; the assured freedom from want of, or poverty for, water for life. (Wouters 2005)

The availability of an acceptable quantity and quality of water for health, livelihoods, ecosystems and production, coupled with an acceptable level of water-related risks to people, environments and economies. (Grey and Sadoff 2007, 545)

Availability of, and access to, water sufficient in quantity and quality to meet the livelihood needs of all households throughout the year, without prejudicing the needs of other users. (Calow et al. 2010, 248)

Sustainable access, on a watershed basis, to adequate quantities of water, of acceptable quality, to ensure human and ecosystem health. (Norman, Bakker, and Dunn 2011, 54)

Water security is a tolerable level of water-related risk to society. (Grey et al. 2013, 4)

The capacity of a population to safeguard sustainable access to adequate quantities of and acceptable quality water for sustaining livelihoods, human well-being, and socio-economic development, for ensuring protection against water-borne pollution and water-related disasters, and for preserving ecosystems in a climate of peace and political stability. (UNU 2013, vi)

Two overarching observations come from a reading of the list. First, each definition is derived from human security rather than securitization perspectives of water security, though some connection between the two is provided by the final UN definition. This is perhaps in part because there is a broader understanding in the international relations literature of what water resource securitization means, and water is just one example of many environmental and other concepts that can be securitized. It may also be in part related to the second observation, which is that none of the definitions are based on how the term “water security” is used but rather on how author(s) will use it or like it to be used. This perhaps demonstrates the perceived power of reframing water discourse and the instrumental uses to which water security discourse is put.

Looking more specifically at definition content reveals substantial variation. For example, the spatial scales used to frame or define water security range from the household (WHO 2003) to a population (UNU 2013) to society (Grey et al. 2013). In one of the earliest definitions, water security is held to exist when there is security simultaneously across scales, from the individual to the national level (Kaplowitz and Witter 2002). In at least one case, water security is linked to the favored scale of hydrology, the watershed (Norman, Bakker, and Dunn 2011), but other definitions are silent on scale (Grey and Sadoff 2007).

While a few definitions (e.g., Kaplowitz and Witter 2002) consider at least vague notions of temporal scale (e.g., “long term”), most either ignore time dimensions or invoke sustainability language (Norman, Bakker, and Dunn 2011; UNU 2013). These latter two approaches ignore possible tradeoffs between short- and long-term water security, for example the tradeoff between the use of fossil groundwater today and the option for its use in the future.

Definitions also show considerable variation in what it means for the unit of interest to be water-secure. Some define water security as the existence of physical water supplies (Kaplowitz and Witter 2002; WHO 2003; Grey and Sadoff 2007). Others define security in terms of access
to those supplies (Norman, Bakker, and Dunn 2011), while yet others define water security not in direct relation to water but as an ability to manage water-related risk (Grey et al. 2013) or, strangely, as having the capacity to be water-secure whether or not security is the chosen outcome (UNU 2013). A final point is that many definitions conceptualize water security in terms of both human and ecosystem uses. Whether the focus on water-secure ecosystems is related to the inherent (or existence) value of the ecosystems or to the ecosystems’ contributions to other aspects of human security is unclear.

There have been limited attempts to use the definitions to measure the state of water security. Most attempts have used largely biophysical approaches based on the availability of water resources and population. While this may be the only way to realistically assess and map at regional or global scales, it cannot account for the social and political factors that determine water access and therefore whether an individual or a nation is water-secure. As a result, they miss explanations of why, for example, water outcomes (water security) in Israel or Jordan may be different from those in India, why differences exist between a commercial farm in South Africa and one on a former homeland, or why there may be differences between individuals within a single city, community, or state. Calow et al. (2010) use the example of groundwater to describe how this shortcoming to examine processes at smaller scales and including human agency can skew assessments of water security across spatial and temporal scales. Physical assessments of water security based on where water is physically “produced” also ignore the impacts of globalization such as the virtual water (the water used in the production of or “embedded” within products) trade in the grain trade and the role of multinationals.

**Conclusion**

A water security discourse has grown and evolved rapidly since the mid-2000s and now has two strands of meaning, one related to security in the traditional international relations sense and the other related to the broader concept of human security. The multiple meanings have led to multiple uses of the term. In some cases water security is used as a new synonym for water scarcity or water problems in general. In others it is used to help redefine and perhaps invigorate existing water agendas. And in yet others it is used instrumentally to bring urgency and resources to water issues. As use of the term has grown, so too have definitions, which now span spatial and temporal scales and measure the nature of water security variously in terms of physical water resources, access to those resources, and risks associated with those resources.

In the securitization sense, the creation of water security, along with other environmental securities, has given legitimacy to the expansion of high politics and military affairs into what was previously a more mundane water field. In the human security sense, water security discourses have been employed to draw more attention to particular works on water, to raise awareness of water problems, and to redefine water agendas and resource flows. While in some cases this may be done to serve narrow interests, in most it is more likely a tactical adjustment in recognition of the power of alternative language. Tony Allan (2003), for example, has highlighted how his earlier concept of “embedded water” did not gain traction until it was renamed “virtual water.” Likewise, water security has more resonance and urgency than earlier terms such as “water productivity” and “integrated water resources management” (IWRM). Everyone understands the need to be water-secure, even if they are unsure of its exact meaning.
When it is combined with food, energy, and other “securities,” the power of water security may be even greater. The 2011 “Water, Food and Energy Security Nexus” conference in Bonn was successful in gathering policymakers and professionals together from far beyond traditional water circles and in ways that earlier framings around water productivity or Integrated Water Resources Management conferences could not have done. As Cook and Bakker (2012) observe, water security “thus holds promise as a new approach to water management.” Interestingly, though, the 2014 follow-up to the Bonn conference, held in Chapel Hill, North Carolina, was titled “Water, Food and Energy Nexus.” The word “security” was dropped, perhaps suggesting that another reframing has already begun.

While a water security discourse may have brought new actors and awareness to the water field, it may not have changed the content yet. Assessments and proposed solutions to water security challenges still tend to focus on the technical aspects and to be couched in terms of efficiency, continuing the assumptions that water outcomes are primarily related to physical supplies, population levels, and technology options rather than power, politics, and access. The questions of who exactly is water-insecure and what are the underlying causes of insecurity remain largely ignored.

**SEE ALSO:** Environmental (in)security; Food security; Water conflicts; Water and human rights; Water rights

**References**


Water has influenced the settlement and development of cities and it will determine their future vitality. The quality and availability of water have always been fundamental to settlement and population agglomeration and critical for travel, trade, industrial development, and energy production. Natural water supply is a limiting factor for urbanization, and technological developments over millennia have sought to augment local supplies. Advances in the generation and regulation of freshwater supplies include topographical modification to direct surface flows to consumers, aqueduct development to distribute water long distances, groundwater extraction, impoundments for storage, as well as other forms of mechanized water-delivery infrastructure.

Climatological patterns dictate the distribution of water seasonally and generate not only water resources, but also water hazards. Relatively regular seasonal changes in rainfall amounts and stream flow are often punctuated by unpredictable extreme conditions and are enhanced by construction of impervious surfaces (buildings and pavements) that increase runoff quantities and rates. Technological advances have changed the implications of flood and drought in cities as residents and regional authorities have sought engineering remedies for the impacts without surrendering to nature's caprice. Major projects have been undertaken to channelize and drain wetlands, to construct levees to protect settled areas, to straighten river channels to speed excess water flow through urbanized areas, and even to drill deep tunnels beneath cities to prevent flooding.

The physical availability and presence of water does not determine how it is used or treated. Historically, the management of water has ranged from unconscious and frivolous waste and maltreatment to careful use by those cognizant of its limits. Water can be used, employed in an activity, but remain unchanged in state or quality, or it can be consumed, made unavailable to users downstream by changing either its quality or its state. The former can be regarded as borrowing the work of water as it flows through a community. The latter is a depletion of the resource. How water is used or consumed by a culture can determine the long-term prospects for its cities, and is the focus of the efforts of urban water managers.

Prior to the nineteenth century, awareness of the relationship between water quality and the health of urban residents was superficial, and water was a relatively loosely managed resource. Water meets the basic needs of drinking, cooking, hygiene, cleaning, and sanitation. The rules of use and the standards for water quality have changed through the ages and are often dictated by cultural norms. Ancient Greeks and Romans built extensive water conveyance systems throughout their cities, and rudimentary devices were often developed to provide protective piping for distribution to aristocratic villas and public baths, but the water in these systems was typically gravitationally fed and often carried human waste and refuse off to rural and coastal environs. While the city of London in the United Kingdom developed an extensive water-supply infrastructure, beginning in medieval times,
water was not commonly distributed to private residents in modern cities until the middle of the nineteenth century. Prior to this time, it was typically available from neighborhood wells and drawn with buckets or by hand-operated pumps, from plaza fountains to which water was piped or physically transported, where personal containers were filled, or from barrels filled at springs or reservoirs and transported by draft animal to urban consumers.

The growth of early nineteenth-century cities, like Boston, Massachusetts, began to outstrip local water supplies. Consumption extended to the rural fringes and beyond, and eventually sites in the Connecticut River Valley in central Massachusetts were appropriated for reservoirs to meet the needs of the city. The emergence of metropolises in the United States by the 1850s established the exploitation of distant regions for water. New York City constructed reservoirs upstate and piped untreated surface water to the city. In the following decades the city promoted the establishment of Catskill Park (1885) in east central New York state and even Adirondack Park (1885) in northeastern New York state to protect critical hydrological source regions and thus its water supply. The case of San Francisco after the devastation of the earthquake and conflagration of 1906 demonstrates the political nature of urban thirst. City leaders called for a diversified water infrastructure for the city and committed to construction of reservoirs in the coastal range to the north and south of the city as well as in the Sierra Nevada far to the east. Hetch Hetchy Valley in Yosemite National Park was their prime site for a reservoir, and its eventual grant to the city water system illuminated the conflicting values of conservationists and those who held preservationist views, as well as the power held by urban boosters over rural communities and other interests.

The urban growth of the American West during the first several decades of the twentieth century was enabled by the frenzied development of hydroelectric dams and reservoirs after the Newlands Reclamation Act was passed in 1902. The dams were initially built for rural irrigation to “reclaim” land for agricultural use, but ultimately the program morphed into energy- and water-supply projects for cities like Las Vegas, Nevada, and the greater metropolitan region of Los Angeles, California. Cities like Phoenix, Arizona, relied on federally endorsed projects to extract water resources from far-off regions. Los Angeles, for instance, exploited the Owens Valley, east of the Sierra Nevada, and the Colorado River, on California’s eastern border, to promote its growth, to the detriment of the source regions. And, despite finite urban water supplies in desert regions, water was often consumed without care. Land-use practices and aesthetic preferences of western cities mimicked those of more humid eastern cities. Extensive planting of non-native vegetation demanded intensive irrigation to overcome normal precipitation deficits. As water demand increased with urbanization, the costs of water rose as well, but failed, until rather recently, to promote frugality and more indigenous landscaping. Tension grew between thirsty urban areas and water-intensive industrial-scale agricultural operations. Changes to either urban behaviors or agricultural approaches have occurred only when one’s power exceeded the other’s in state or federal governments. Pricing schemes, subsidies, and the establishment of watershed-defined ownership are tactics aimed at stopping urban–rural water conflicts during the twentieth century.

Cities, even in more humid environments, are finding it more difficult to meet the demands of their populations. States that encompass the Great Lakes region of the United States and
Canada have denied requests to allow the transfer of water beyond the boundaries of the lakes’ contributing watershed despite the volume of water in the watershed. Interbasin water-transfer restrictions have emerged to reflect the stress of the growing need for water in cities that have surpassed their local and regional supplies. Even the city of Waukesha, Wisconsin, a mere 20 miles from Lake Michigan, has not been allowed to transfer lake water across the divide 15 miles west of the lakeside city of Milwaukee to surmount its limitations. Milwaukee and the Great Lakes Compact states are concerned about setting a legal precedent.

Urban densities intensified during the twentieth century with industrialization and the construction of transportation infrastructure to fuel economic growth. Segregation by wealth and ethnicity was fostered by investment in streetcar, rail, subway, and eventually highways and improved surface roads. This uneven development was also reinforced by the extension of public services. Water and sewerage piping were quickly retrofitted in established upper-income parts and more slowly in working-class sections of cities, and increasingly presaged the sale and development of residential and commercial sections of growing cities. Modernization included bringing privies indoors, and expecting stronger piped-water pressure and higher-quality and better-tasting water. Until the 1890s, water consumed in cities was untreated and was often contaminated by sewage. Sanitary sewers to carry off human waste and storm runoff were designed and built in Chicago, Illinois, and Brooklyn, New York, in the 1850s, but few other cities built them until germ theory revolutionized understanding of the sewage–health nexus. Unfortunately, the sanitary sewers of most cities often incorporated industrial wastes into the mix and simply discharged the aggregate into adjacent rivers, lakes, or the ocean in the belief that, once diluted, the sewage would be harmless.

In 1948 the US government enacted the Federal Water Pollution Control Act to provide financial support to facilitate the construction of wastewater treatment facilities for communities of all sizes. Urban areas required greater financial backing than states typically could provide, so federal loans and grants established plants to provide primary and secondary treatment of residential wastewater. Integrated plumbing and safer piping became standard in American homes. However, the push to completely remove lead piping from waterworks and the treatment of industrial effluent did not occur until the 1980s in the United States.

According to the World Health Organization (WHO) and United Nations Children’s Fund (UNICEF), as of 2011, 96% of the world’s urban population has access to improved water sources (WHO 2015). Access for the US urban population nears 100%, as for urban areas of most of the rest of the industrialized countries. The trends in access have been rather steady for the Global North, and significant improvements have been made in less industrialized countries of Africa and Asia, some increasing urban access by more than 70% (e.g., in Afghanistan) since the 1990s. In former Eastern Bloc countries, however, access diminished with the collapse of the Soviet Union in the early 1990s. Most have not rehabilitated their water systems fully.

Aging infrastructure, increasing costs of water provision, and generally difficult economic conditions around the world have hastened the push for nongovernmental organizations and international aid agencies to step in to build infrastructure, as well as the privatization of water services in some regions and countries. The need for fresh water continues to grow with population and economic development. The
production of sewage has similarly continued apace. The WHO reports that only 82% of the world’s urban population had access to improved sanitation facilities in 2015 and the greatest need today remains throughout Africa and in some small island states where only 15% to 50% have access (WHO 2015).

Both water and sanitation have required more investment to increase supply, build infrastructure and desalination plants, construct more treatment facilities, and increase regulatory enforcement. Emerging megacities and metropolitan urban agglomerations have also revealed the artificiality of political boundaries for natural resource management. Governments may be challenged to negotiate their own needs within the contexts of other urban neighbors’ needs and rural water demands. As water and sewer systems grow in size, they grow in complexity and require increasingly complicated approaches. But funding has fallen behind. Alternative mechanisms have emerged to meet water supply and water quality demands in economical and more efficient ways.

Besides repair to extant systems, late twentieth- and early twenty-first-century approaches to water use include conservation, recycling, gray-water use, rainwater harvesting, and environmentally appropriate landscaping and development. Water conservation and hazard-mitigation projects have often sought to undo what had been done in the past. For instance, the majority of flood management in the United States throughout the twentieth century dictated the engineering of rivers to hasten the movement of water past urban areas, but today’s urban floodplain managers are working to re-establish more natural river environments and to enhance the absorption and water-holding capacity of urban stream networks. Rooftop gardens are being constructed on tall buildings to reduce runoff, to exploit available rainfall, to reduce solar energy absorbed by modern architecture, and even to recreate habitat for urban wildlife. In many Sun Belt cities of the United States, municipal governments have mandated that xeriscaping replace water-intensive lawns and gardens and that outdoor water use for landscaping be reduced significantly.

Recently, many urban areas around the world have sought sustainability of their water use in a quest for consumption practices that fit within a region’s natural supply to ensure a long-term renewable and permanent supply of water. But a major complication that is yet to fully manifest itself is climate change. Cities will experience a diverse set of trends influenced by many variables: extremes of temperature or precipitation may increase or decrease in frequency or intensity, and regular patterns may shift throughout the hydrological year; durations of hot, cold, wet, or dry periods may expand or contract; snowfall may diminish or snow may fall instead as rain and reduce the amount of mountain-stored moisture available to urban streams; the timing of seasonal weather patterns may change or persist. All of these factors will influence the amounts and qualities of water supplies throughout the year. The generally dependable long-term record of water availability is no longer assured, and planning and management of urban water now involves new uncertainties and unknowns.

SEE ALSO: Industrialization; River basin management and development; Sustainable cities; Water and climate change; Water conservation; Water economics; Water engineering; Water quality; Watersheds

References

Further reading


A watershed is a land surface area encompassing all points from which streamflow at a given location is derived. Watersheds are fundamental natural, hydrological, and ecological subdivisions of the Earth’s surface with system attributes and functions having great relevance to human existence. The term “watershed” is often used synonymously with “drainage basin,” particularly in the United States, and also with the term “catchment,” although the latter is frequently used to emphasize land surface conditions rather than the combined surface and subsurface hydrological system. The word “watershed” has been traced to Germany (wasserschiede; water divide), where it was in use by about 1400 CE. It was in widespread English usage by 1800 CE. However, the concept of a collecting area for waters contributing to river flow is probably much older. Watersheds have been described as fundamental geomorphic units (Chorley 1969) for several reasons. They are fully contiguous and space filling; all points on the land surface fall within some watershed. They are convenient units of system organization, with divides representing the natural and easily established positions of open system boundaries. Within-system fluxes of materials are easily conceptualized, if not always easily quantified within watersheds. Furthermore, water is the single most important of all resources for human existence, and a watershed-based system of land area organization is considered to provide the least ambiguous and most sustainable land parcel arrangement for resource assessment, development, and conservation.

Watershed extent and delineation

The spatial extent of a watershed is defined by its point of outflow on a stream. In practical application, the point of outflow, and thus watershed size, is usually determined by specific data needs. For example, the point may be a location from which water will be removed for various uses (municipal supply, irrigation, power plant cooling, etc.), or a location of flood hazard. More generically, stream junctions may be used as delineation points; thus, when speaking of a river’s watershed, all points upstream of its confluence with another water body (river, lake, or ocean) will be included.

The line delimiting the spatial boundaries of a watershed defines the drainage divide, and usually corresponds with the crests of topographic ridges (Figure 1). Locating the boundaries of watersheds, the procedure known as watershed delineation, can be accomplished by hand-tracing on a topographic map, or by using flow directions derived from computer-based geospatial analyses of digital elevation models. On a topographic map, tracing begins at the designated point of outflow, from which lines are drawn immediately upslope and perpendicular to contours, and then extended around the catchment limits until they join, encircling all upstream tributaries. These topographic divides may not always precisely correspond with the groundwater (phreatic) divides beneath the land surface, in which case their placement will be an approximation. This is particularly relevant...
where geological conditions produce unpredictable directions for groundwater movement, such as in karst environments.

Although the upper limit of watershed size is ultimately constrained by the size of a continental land mass (Table 1), there is no objectively defined lower limit. Every watershed can be subdivided into smaller subwatersheds, which can be further subdivided ad infinitum. Thus, the minimum size of a watershed is technically unlimited. More practically, however, the lower limit of watershed size is set by the sorts of problems requiring a watershed-scale analysis. For example, small-plot soil loss studies may require delineations of hillslope areas measurable in units of several square meters. But other lower limits might be proposed, such as the smallest contributing area required for the existence of a perennial stream channel.

**Table 1** Drainage areas of the largest nontributary river basins on each continent (excluding Antarctica).

<table>
<thead>
<tr>
<th>Continent</th>
<th>Watershed</th>
<th>Drainage area (km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>Congo</td>
<td>3 822 000</td>
</tr>
<tr>
<td>Asia</td>
<td>Ob’-Irtysh</td>
<td>2 975 000</td>
</tr>
<tr>
<td>Australia</td>
<td>Murray-Darling</td>
<td>1 072 000</td>
</tr>
<tr>
<td>Europe</td>
<td>Volga</td>
<td>1 380 000</td>
</tr>
<tr>
<td>North America</td>
<td>Mississippi-Missouri</td>
<td>3 221 000</td>
</tr>
<tr>
<td>South America</td>
<td>Amazon</td>
<td>7 180 000</td>
</tr>
</tbody>
</table>

Source: Czaya 1983.
Physical characteristics

Watersheds are composed of land surface areas and their underlying regolith and bedrock, in addition to stream channels connected to form the channel network. The land areas may be further subdivided into sloping uplands (hillslopes of variable size and shape), and flat bottomlands consisting of river floodplains and terraces. More detailed geomorphological parsings are also possible. A number of useful descriptors for watersheds have been proposed, many since the appearance of Robert Horton’s classic (1945) “Erosional Development of Streams and their Drainage Basins: Hydrophysical Approach to Quantitative Morphology,” a broadly recognized and influential publication focusing on the quantitative process geomorphology of watersheds. Among Horton’s groundbreaking contributions are the infiltration-excess mechanism of overland flow generation, a theory of drainage network development, and empirical laws of drainage network composition, relating the numbers and mean lengths of streams in a watershed to stream order.

Stream order, a widely used index of stream size, is found by assigning the smallest headwater streams an order of one (“first-order”). Progressing downstream, order is incremented by one only after two streams of equal order join (Figure 2). Later, Ronald Shreve proposed stream magnitude as an alternative to stream order to account for the substantial water inputs of excess streams (those whose additions do not increase stream order). Headwater streams are assigned a magnitude of one, then the magnitudes add progressively downstream at each junction such that the magnitude of any stream segment equals the total number of magnitude-one streams that are upstream (Figure 2). Horton also introduced the concept of drainage density, the total length of streams in a watershed divided by watershed area, which is a measure of stream frequency that is of fundamental significance in explanations of watershed material fluxes. Many other descriptors, including shape indices (circularity, elongation), relief characteristics, and topological properties of stream networks, are also frequently used in watershed analyses.

Watershed hydrology

The continuity equation \( I - O = \Delta S \) (input minus output equals change in storage) is a basic expression useful in water and sediment budgeting at both watershed and channel-reach scales. In watershed-scale water budgeting, input is via precipitation, outputs are via stream discharge (the rate of water volume outflow, or runoff) and evapotranspiration, and change in storage is via deep groundwater recharge and soil moisture storage. In dryland environments, both precipitation amounts and the infiltration of precipitation into bare, crusted soils are low, and little water is available for deep percolation on hillslopes. In such cases most streams will be influent, with groundwater primarily being recharged by streamflow infiltrating the channel bed. In contrast, much of the large
amounts of precipitation in humid environments will infiltrate hillslope regolith, subsequently following both deep and shallow subsurface pathways towards stream channels (Figure 3). Thus, the variable potential for water infiltration into the ground, known as the *infiltration capacity*, is a critical measure for understanding the rainfall–runoff response of watersheds during storms. Infiltration capacity can be expressed as a volume, or more commonly as a depth of applied water that moves into the soil per unit time.

The infiltration process is influenced by three interdependent subprocesses: (i) the movement of water across the plane separating the ground surface from air, (ii) soil moisture storage, and (iii) the continued transmission of water below ground (percolation). The boundary between air and regolith is diffuse because air extends downward into porous materials when pore spaces are not occupied by pre-existing (antecedent) moisture. However, the ground surface plane is important because kinetic energy transferred from raindrop impact can produce surface sealing only in a shallow (between 1 mm and 1 cm thick) zone below the regolith surface. Surface sealing occurs when smaller soil particles are splashed into large pore spaces on the soil surface, obstructing downward movement of water, and reducing infiltration capacity. For cases in which the precipitation rate (expressed in the same units as infiltration) is greater than the infiltration capacity of soil, the difference in the two rates will accumulate on the ground surface, first filling small depressions (depression storage), which, if the process is continued, may
spill over to produce a thin sheet or series of shallow rivulets constituting overland flow. This model for the generation of surface runoff was first proposed by Horton (1945), and is referred to as the infiltration-excess overland flow mechanism, or in honor of Horton, Hortonian overland flow.

Precipitation that infiltrates into soil may percolate downward to recharge the water table, or it may begin to flow laterally downslope through the soil as throughflow. The reemergence of underground water-flows onto the surface, known as return flow, often takes place near slope bases where ground may already be near saturation due to continual receipts from upslope drainage. Overland flow produced in these saturated zones is referred to as saturation overland flow, which combines both return flow and direct precipitation onto saturated ground. The spatial extents of these zones expand and contract through time, changing the effective storm runoff contributing area of the watershed. This is known as the variable source area model of overland flow generation, which was introduced by Hewlett and Hibbert (1967) to explain why streams can rapidly flood in humid regions even when most or all of storm precipitation infiltrates in upland areas.

Changes in stream discharge through time are depicted in hydrographs (Figure 4). The characteristics of hydrographs, including peak discharge, the delay between rainfall and the discharge peak (known as the lag time or lag-to-peak), and the duration of flows elevated by storm precipitation (known as the time base of stormflow), are controlled by the intensity and duration of the rainfall event, factors such as slope and soil cover, which affect infiltration capacity and general hillslope hydrology, watershed size and shape, and properties of the stream network such as streamflow travel distances (and associated times) and the geometric arrangement of streams. Hydrograph properties change as a flood wave moves downstream. Water storage effects are often critical, with two extremes possible (Figure 4c). Translation of the flood wave under conditions of minimal storage results in a hydrograph shape that changes little as the flood wave moves downstream, an outcome most likely to be important in fully confined stream channels with no floodplains, and steep valley walls. Where channels are unconfined, sinuous, and bordered by large floodplains, floodwaters, storage potential is higher, leading to lower peak discharge and longer time base, a process paralleling storage of floodwaters in reservoirs behind dams. Along rivers too, this in-channel and overbank storage and slow release process is known as the reservoir effect. Although either effect may be dominant at a given site, most rivers display behaviors intermediate between the two.

**Watershed erosion**

Water-flows throughout watersheds are capable of doing work by eroding and transporting sediment. Sediment can be in either particulate or dissolved (ionic) form and is derived from both surficial (or near-surface) erosion of hillslopes and erosion of stream channel bed and banks, and from both ground-level and subsurface areas via chemical weathering. Sediment transport occurs in one of three fundamental forms: bedload (particulates transported in constant or intermittent contact with the channel bed), suspended or wash load (particulates transported in suspension in the moving, turbulent water column), or dissolved load (fully dissolved ionic materials).

The amount of sediment transported during a particular flood event depends on many factors, but the most fundamental are the size of flood...
WATERSHEDS

and the nature of sediment sources. The removal of weathered and eroded sediment mass from upland areas resulting in ground surface lowering is known as denudation. Hillslope erosion may be accomplished by both mass wasting (gravitational movement of rock or regolith) and overland flow (sheet and rill erosion). Mass wasting is controlled by dynamic changes determining the ratio of driving forces (the downslope component of gravitation acting on Earth materials) to forces resisting motion (friction and cohesion). Sheet and rill erosion by overland flow is controlled by the relative rates of precipitation and infiltration capacity and hillslope position, or in the case of saturation overland flow, by throughflow mechanisms resolving near the bases of hillslopes.

Surficial erosion by water on hillslopes is a two-stage process: (i) detachment of particles from the ground surface, and (ii) transportation of those particles in overland flow. Detachment can be produced via the application of shear stress by overland flow, or by the transfer of kinetic energy possessed by falling raindrops upon impact with the ground. The magnitude of stresses imposed by overland flow is tied to the depth of flow, which is often limited on upland surfaces, particularly near the crests of ridges where cumulative flow additions from successive upslope positions is not great. Horton referred to this topslope zone as a belt of no erosion, only downslope of which sufficient depths of water could occur to initiate sheetwash erosion and incision of small channels (rills). However, in some regions, overland flow rarely reaches depths capable of producing effective detachment. In all types of areas, raindrop impact, sometimes referred to as rainsplash, can be an effective detachment mechanism, although it is severely limited in areas with dense ground cover. The combination of rainsplash detachment penetrating through shallow overland flow and the subsequent removal of detached particles in that flow is particularly effective and has been referred to as rainflow or rainwash erosion (Morgan 1986).

The far-reaching onsite and offsite consequences of surficial erosion, including soil fertility losses and excessive sedimentation in streams (siltation), have prompted many decades of soil erosion research leading to several methods for its prediction. The most commonly used models for soil loss prediction are the Universal Soil Loss Equation (USLE; Wischmeier and Smith 1965) and its global variants. The USLE-type model integrates the effects of rainfall and soil characteristics, topography, vegetation

Figure 4 (a) Components of a hydrograph; (b) hypothetical stormflow hydrograph from an urban watershed, showing decreased lag time and time base, and high peak discharge; (c) hypothetical hydrographs for translation and reservoir effects. Original artwork by author.
cover, and erosion control practices in a simple multiplicative structure created from large amounts of empirical erosion data. Although these empirical models do not directly simulate the physical processes of erosion, their spatially distributed implementation at the watershed scale using geographic information systems is straightforward and potentially a powerful tool for watershed management in both agricultural and even urban settings, if properly refined.

Upland erosion is often of overriding importance to watershed-scale erosion, given the areal dominance of uplands in the watershed, and the concentration of land use in these areas. However, fluvial processes leading to stream channel bank and bed erosion are also usually substantial in magnitude, and potentially of primary importance in landscapes with stable (or stabilized) uplands. Like erosion by overland flow, erosion of the stream channel boundary is contingent on the shear stresses applied to the bed by flowing water, which are proportional to the depth of flow and gradient (longitudinal slope) of the channel. These constantly changing stresses may be compared at any given time to the critical shear stress required to move sediment particles of sizes represented in the channel boundary to determine the likelihood of erosion. The initiation of particle movement into streamflow is known as entrainment, which is a stream channel precursor to transportation. Stream bank erosion differs from bed erosion in that banks are subjected to alternating inundation and aeration during and between floods and, when inundated, covered by lesser water depths. Furthermore, banks are often composed of finer and more texturally varied or stratified sediment than channel beds, and may be stabilized by vegetation. Banks may be exposed to rainsplash when above water, and given the relative steepness of banks and the resulting increase in downslope splash distance, rainflow erosion is capable of contributing some sediment to sediment loads when banks are less than 90 degrees. However, much greater stream bank erosion is typically produced by mass wasting of banks undercut by streamflow.

Watershed sedimentation

The continuity equation is also useful for thinking about watershed sediment budgets. A sediment budget is a structure which seeks to quantitatively define what happens to sediments generated by watershed erosion, and is an important tool for understanding short- and long-term geomorphological dynamics of watersheds. This knowledge is critical for understanding past, present, and future environmental impacts of land use, and making decisions regarding the rehabilitation of degraded streams and general watershed management. The US Environmental Protection Agency lists excess sediment as the single most common contributor to stream degradation nationwide.

Inputs in fluvial sediment budgets include all forms of erosionally produced sediments. Output of sediment is known as sediment yield, which is defined as the mass of sediment exported from a watershed per unit time (metric tons per year). To enable comparisons between watersheds of different size, the sediment yield per unit drainage area, known as the specific sediment yield, is used. Determining sediment yield values from different watersheds has been an important research focus over many decades because of the need to link excessive sediment loads and the problems they can cause to land-use practices. Erosive land uses such as row-crop agriculture almost always result in increased sediment input into streams, which manifests in the form of higher sediment yields at watershed outlets. However, strong direct relationships between
WATERSHEDS

the quantities of erosion in the uplands and contemporaneous sediment yield may not occur. Particulate erosion and sediment transportation in watersheds are highly intermittent, with travel distances during storm events ranging in magnitude from meters (or less) for coarse fractions on hillslopes, to many kilometers for finer sediments suspended in streamflow. Temporary shifts from sediment in transport to storage ($\Delta S$) may occur at many different landscape positions, and stored sediment may be remobilized over widely varying timescales. Potential storage locations include on and at the base of hillslopes, on terrace and floodplain surfaces, and in sediment benches and bars within stream channels, each having its own remobilization period.

In one of the best known illustrations of the importance of sediment storage, Trimble (1999) found that sediment yield rates from the Coon Creek watershed in Wisconsin (United States) remained virtually unchanged over almost 150 years, despite major changes in sediment inputs from upland erosion and tributaries. The modification of the relation between erosion and sediment yield, mediated by storage dynamics, is often expressed as the dimensionless sediment delivery ratio (SDR), as shown in equation 1.

$$SDR = \frac{\text{Watershed sediment yield}}{\text{gross watershed erosion}}$$

Thus the term sediment delivery refers to the proportion of gross catchment erosion that manifests as sediment yield in some period of interest. This should not be interpreted necessarily to mean that the actual sediment particulates being measured at the watershed stream outlet are the same particulates that were eroded during the period of monitoring. It is possible that all eroded materials in a given storm or other period of observation are stored for periods longer than that of observation. As long as the denominator is defined to include all locations of erosion (stream channel erosion as well as upland erosion), the sediment delivery ratio must vary between 0 and 1 (0–100%). Although sediment delivery ratios may be calculated over any time interval, individual storm periods are probably too short to give a representative value, and temporal averaging in proportion to catchment size has been advocated (Richards 1993).

Sediment yields and SDR values commonly decline downstream as watershed area grows. Explanations for these trends focus on declines in the numerator of SDR (sediment yield), and two competing explanations have been offered (Richards 1993). One explanation is that the sizes of storage compartments and opportunities for sediment storage increase with distance from more steeply sloping headwater areas. Downstream increases in floodplain width allow for greater overbank sediment storage as well as hydrograph-attenuating storage of floodwaters which, due to shallower depths, have reduced velocity, promoting deposition. The toeslopes of valley walls adjoin the flatter floodplain rather than the streambank, and the abrupt gradient decrease promotes deposition of colluvium (mass wasting sediment). Furthermore, the valley walls become more distant on average from channels due to intervening floodplain areas, and their sediments are less likely to reach rivers traveling over these gentle intervening floodplain surfaces.

The second explanation is that steeper headwater areas are often the dominant sediment sources, and thus sediment production (via erosion) per unit area of watershed decreases downstream due to decline in average land surface slope. However, the evidence in support of this hypothesis is relatively weak and explanations invoking increasing sediment storage opportunities are more widely accepted.

SDR is a black-box concept, meaning that erosion and sediment yield are the only data required for its determination. A detailed knowledge of
intervening sediment storage dynamics is not necessary. Yet, detailed spatially explicit information on all aspects of sediment delivery is critical for understanding the impacts of watershed processes on landscape change, water quality, land resources, and riparian ecosystems. Inasmuch as sediment storage indicates a disconnection, however temporary, in sediment flows through a watershed, SDRs index connective properties of watersheds which are a vital concern for current analyses of sediment cascades. Watershed connectivity can be viewed in terms of both hydrologic and sedimentological connections between adjacent land surface forms. Connectivity in watersheds depends on climatic characteristics, hillslope runoff potential, topographic positional attributes affecting energy and slope continuity, and flow-path distance. Thus, connectivity variables are similar to those important in explaining sediment yields.

The origins of watersheds

Prior to the twentieth century, few comprehensive explanations of watershed origin had been offered. The geographical cycle of W.M. Davis and similar theories of landscape evolution remained influential in the early 1900s. With increased emphasis on quantitative process geomorphology beginning around 1940 came new ideas on how drainage networks, and by necessity their contributing (watershed) areas originate. Robert Horton, in his seminal (1945) paper, was one of the first to compose a comprehensive theory of drainage network development to go along with his ideas on overland flow and the belt of no erosion. According to Horton, once overland flow becomes deep and fast enough (i.e., downslope of the belt of no erosion), rills are formed. Because of minor differences in hillslope or divide shape, one of the initial rills is likely to become larger than adjacent rills and begin capturing their flow. Horton referred to this process as cross-grading, and depicted it as a hierarchical process in which captured rills become tributaries and undergo the same differential enlargement process, progressively dividing up the former slope until drainage areas are no longer sufficient to support rilling. Captured tributary flow progressively enlarges the trunk rill to gully size, and eventually to a size necessary to be regarded as a stream.

Horton’s model may accurately describe the origins of some smaller drainage networks in dryland environments, and poorly vegetated and thin-soiled areas such as shale badlands, where infiltration excess (Hortonian) overland flow is the dominant runoff mechanism. Stream networks in humid, soil-mantled, and well-vegetated landscapes where infiltration excess overland flow is uncommon do not fit the Horton model. In these environments networks are more likely to originate in association with zones of saturation overland flow near slope base. Headward growth of stream channels often occurs along subsurface drainage lines by processes like gully wall sapping that are strongly influenced by groundwater. Larger networks may develop in more complicated ways, and multiple mechanisms of drainage network initiation and growth may be operative in a single watershed.

The long-term evolution of watershed morphology includes not only initiation and growth of stream networks, but also their contraction. Glock (1931) used topographic map data and ergodic reasoning (space-for-time substitution) to infer that drainage network evolution has distinct morphological stages, including: (i) an initiation phase in which the first poorly integrated channels are created on sloping land, (ii) an extension phase characterized by rapid elongation (headward growth) and elaboration (addition of
small tributaries), (iii) the attainment of maximum extension, and (iv) a final period of integration characterized by the loss of stream length (and reduction of drainage density) driven by relief reduction, slope decrease, and the increasing influence of hillslope processes that close out channels. Despite decades of research involving field studies, scale (hardware) models, and other approaches, a comprehensive understanding of drainage development remains elusive, and increasingly, sophisticated computer-based modeling is required to fruitfully explore the issue.

**Land-use impacts on watersheds**

Land-use practices such as timber harvesting, row-crop agriculture, and urban development impact watershed functions in distinctive ways. Forest clearance for timber sale or conversion of land to agricultural or urban uses decreases evapotranspiration losses in the short term, resulting in increased streamflows. Depending on the method of timber extraction, the land surface may undergo variable changes in ground stability leading to surfcial erosion, or more commonly of greater impact along logging roads, large hillslope mass failures. Row-cropping requires prior removal of all vegetation, and the loss of ground cover and destruction of natural soil structure by tillage exacerbate surface sealing by rainsplash, ultimately increasing overland flow and surfcial erosion rates. Urbanization involves large increases in watershed surface covered by pavement or other materials that are impervious to rainfall, greatly reducing average infiltration capacities. As a result, a large proportion of rainfall in urban areas is converted to overland flow, then carried rapidly through a partially buried stream network made up of stormwater pipes, before emptying into nearby streams. Hydrographs in urban streams are characterized by short lag times and time bases, and high peak discharges (Figure 4b), which in most cases lead to erosional enlargement of channel cross-sections. High peak discharges also represent a significant flood hazard for neighborhoods proximal to valley bottoms within urbanized areas.

Wolman (1967) tracked changes in watershed sediment yields encompassing the sequential transition from forested, to agricultural, and finally to urbanized conditions as these occurred in the Maryland Piedmont (United States) over the last two centuries. He found that sediment yields during peak agriculture were at least 20 times greater than under the prior forested conditions. Peak sediment yields began declining with reductions in cultivated area and the advent of soil erosion control practices between 1925 and 1960. Urbanization brought sediment yields at least 80 times larger than under forested conditions due to erosion on urban lands severely disturbed by construction activities. Elevated sediment yields were quickly reduced following construction due to land surface stabilization under pavement. Recent studies of urban sediment yields point to renewed channel erosion as the dominant source of sediment following the achievement of a stable impervious surface density. Other studies have noted that urban land disturbance is not steady state, because areas may undergo periodic redevelopment indefinitely through time. In many places, large quantities of sediment from historical anthropogenic erosion phases are stored in channels, on floodplains and behind small dams, and later remobilized. Withdrawal of this “legacy sediment” from storage after erosive land uses have diminished can maintain elevated watershed sediment yields for decades to centuries.

In watersheds, water quantity is inseparable from water quality, and land-use-driven hydrologic and sedimentation changes are usually paralleled by negative impacts on water
chemistry. Pollutants derived from sources distributed across watershed surfaces are known as non-point-source pollutants (NPS). In addition to excessive sediment, these include fertilizers, pesticides, and bacteria in agricultural watersheds, and fertilizers, petroleum hydrocarbons, bacteria, and heavy metals in urban watersheds. NPS pollution is now the largest cause of water quality impairment in US streams. A number of structural and nonstructural solutions to NPS problems exist, although legislation focusing on NPS pollution, such as Section 319 of the Clean Water Act in the United States (1987), is of relatively recent origin.

Watershed-based planning

Land and water resource planning applied at the scale of entire watersheds is a relatively recent occurrence in natural resource history. The famous nineteenth-century explorer, naturalist, and geographer John Wesley Powell was a strong advocate of watershed-based planning and settlement in the dry-climate western United States. Despite Powell’s efforts, watershed-based planning was not widely implemented in the United States during his lifetime, although an increased focus on integrated watershed planning was to emerge shortly after his death in 1902. One of the first major examples of comprehensive large watershed planning is the creation of the federally funded Tennessee Valley Authority (TVA) in 1933. The TVA’s multiple purpose mandate was to build dams to control flooding on the Tennessee River, enable navigation, generate hydroelectric power, and promote economic development. Modern watershed management at a smaller scale may also have a number of aims, including flood hazard mitigation, provision of water supply, and reduction of environmental hazards posed by agricultural and urban water pollution. Water resource and ecological problems caused by poor land-use practices have primarily been approached through activities focused on particular points on stream reaches where problems are acute or interfere with other needs. Although it has long been recognized that watershed-scale conditions are responsible for many environmental, natural hazard, and water resource problems, it has proven more difficult to proactively solve the complex issues of land-use regulation and change throughout a watershed than it is to attempt remediation of point manifestations of these problems as they arise.

Despite the now common recognition that watersheds are ideal spatial units for considering resource decisions, the physical integrity of their boundaries is often compromised by interbasin water transfers. Large-scale water diversions over vast distances supply water for dryland irrigation and cities such as Los Angeles, California; yet similar interbasin transfers of shorter distance have become more common in humid environments as well, sometimes resulting in water wars like those in dryland areas.

SEE ALSO: Fluvial depositional processes and landforms; Fluvial erosional processes and landforms; Geomorphic hazards; Hillslopes; Hydrologic cycle; Infiltration; Rivers and river basin management; Soil erosion and conservation; Water resources and hydrological management

References


**Further reading**


Weather affects all facets of life on Earth. Atmospheric elements such as precipitation, temperature, and wind are ordinarily considered resources supporting habitats and rudiments necessary for animal and plant survival and well-being. However, when a particular weather element deviates from a normal band of tolerance for life, it transitions from a resource into a hazard. These extremes combine with human and environmental sensitivities to create negative consequences and, in some cases, disasters.

Weather extremes are typically referred to as “storms,” which are atmospheric disturbances driven, in large part, by the foundations of our Earth-atmospheric system including: differential heating of the surface of the Earth by the sun, the heating of the lower atmosphere by energy transfer from the Earth’s surface, and Earth’s rotation and varied topography. Storms are important in redistributing excesses in energy found in the Earth-atmospheric system; however, in trying to seek an insatiable energy balance, these storms generate weather elements such as pressure extremes, high winds, intense precipitation rates, and other hazards that can be detriments to society.

Large-scale storms typically arise as regions of low pressure that are characterized by winds rotating cyclonically (hence, their meteorological name, “cyclones”). These cyclones derive their classification – extratropical or tropical – depending on their meteorological and geographic origins. Smaller, or mesoscale, disturbances can create localized hazards that are often driven by deep, moist convection in the troposphere, or thunderstorms. However, in the absence of these atmospheric disturbances, sustained quiescent weather conditions can, somewhat paradoxically, lead to climatological extremes such as drought or heat waves.

Extratropical cyclones

Extratropical cyclones are the primary genitor of weather extremes in the mid-latitudes (30°–70°). Studies as far back as the late nineteenth century recognized the important effects these cyclones have on the day-to-day variability and hazard-sapes in extratropical regions. Extratropical cyclones are large swirling areas of low pressure that are characterized by air mass fronts and are generated dynamically by the Earth’s rotation in combination with the instability caused by the uneven heating of the Earth’s surface. Storm initiation, or cyclogenesis, is provoked by upper-level disturbances, or troughs, embedded in regions of enhanced winds within the jetstream, or jetstreaks, that create diverging air currents near the tropopause. Cyclogenesis is climatologically focused where extreme gradients in orography (leeside of a mountain range) or low-level atmosphere temperatures (warm ocean current with nearby cold landmass) exist. Intense extratropical cyclones occur most frequently in the Southern Hemisphere near New Zealand, east of South America, and in the southern
WEATHER, EXTREME

Indian Ocean and in the Northern Hemisphere near east-central North America, the northern Atlantic, Europe, Far East and northwest Pacific, and south-central Asia.

In the high latitudes, cyclones with similar spiraling structures to their mid-latitude counterparts, but smaller in spatial (≈100 km in diameter) scale, are known as polar lows or, colloquially, arctic hurricanes. These events are particularly dangerous to mariners due to their forecast difficulty and rapid onset of extreme winds and snow. In the Northern Hemisphere, these severe events most commonly form and affect the areas near the Davis Strait and Labrador Sea, Baltic Sea, and Gulf of Alaska and Bering Sea.

Extratropical cyclones have the propensity to rapidly develop and intensify after their inception, lasting from a few days to over a couple weeks; the lengthier lifetimes occur when a cyclone becomes “cut-off” from the jet stream and gradually dissipates. Climatologically, cyclones tend to dominate the cool and transition seasons in each hemisphere, promoting a wide spectrum of hazards, from blizzards and ice storms on the cool side of the cyclone to tornado outbreaks, lightning, hail, severe nontornadic winds, and flooding in the cyclone’s warm sector. Cyclones cover extensive geographies, with each storm encompassing from one to three thousand kilometers at maturity and traversing wide expanses of their hemisphere. Not all extratropical cyclones produce the spectrum of weather extremes, but due to their great size, regularity, and strong dynamics, they are more responsible for weather hazards than any other storm phenomenon. In particular, events that undergo explosive cyclogenesis, or exceptionally rapid deepening of a storm’s center of low pressure, are disproportionately more likely to produce hazards due to their intense dynamics. These events are classified by meteorologists as “bombs” and, in the Northern Hemisphere, are most frequent during the cool and transition seasons along the Canada and US east coast, northern Atlantic and northwest Europe, as well as the north-central and far northwestern Pacific.

Extratropical cyclones and their affiliated fronts can promote extensive changes in sensible weather over short periods of time. Severe winter weather most frequently manifests on the poleward side of cold and warm fronts in extratropical cyclones. If an extratropical cyclone draws in a polar air mass that is sufficiently deep and cold, it can be branded a “winter storm” containing life-threatening cold temperatures, high winds, heavy snow, ice pellets, and/or freezing rain. Winter weather hazards are capable of negatively impacting public safety and human systems, with high-end events damaging infrastructure and utilities, halting transportation systems, restricting socioeconomic activities, and harming or killing livestock and cultivated plants. In particular, blizzards and ice storms can paralyze regions, impacting millions of people and causing billions of dollars in agricultural, property, utility, and productivity losses, as well as casualties (e.g., frostbite, hypothermia, or injuries sustained in weather-related vehicle crashes). Extreme cold air outbreaks into the middle and lower latitudes occasionally follow strong extratropical cyclone passages in the cool season. These cold waves can result in billions of dollars in economic loss (primarily in the agricultural sector) and can cause thousands of fatalities in exceptionally vulnerable regions.

The term blizzard is loosely applied to any heavy snowstorm; though, few snowstorms exceed official blizzard criteria employed by government forecast agencies. For example, the US National Weather Service (NWS) defines a blizzard as an event that has snow and/or blowing snow reducing visibility less than four tenths of a kilometer and sustained winds of 15.6 m s$^{-1}$ for at least 3 hours. A global blizzard
climatology does not exist, but these events are most commonly experienced poleward of the climatological track of intense, cool season extratropical cyclones, with formation occurring on the leeside of mountain ranges and within or adjacent to arctic air mass source regions. For example, in North America, blizzards most commonly affect the interior portion of the continent including large expanses of the north-central United States and central Canada prairies. On average, one to two blizzards per year can be expected in these regions.

Though less frequent than their interior continental counterparts, blizzards can be generated by powerful low-pressure systems fostered by dramatic low-level temperature gradients that predominate where warm coastal waters border cool landmasses. In North America, these coastal lows develop near the northern Gulf Coast or Cape Hatteras and intensify as they move to the northeast United States and Atlantic Canada. These coastal extratropical cyclones, called nor’easters, are particularly noteworthy due to their litany of hazardous effects – from blizzard conditions, hurricane-force winds, precipitation type variability and extreme rates, to coastal flooding and erosion. On average, approximately 12 of these coastal cyclones occur per cool season, with two or three producing significant impacts on the amplified population and infrastructure found along the east coast. Similar extratropical cyclones occur regularly during the cool season across northern Japan and northwestern Europe.

Drizzle and rain falling into air near the surface at or below the freezing point can promote ice accumulations that may have significant influences on energy and transportations systems. While minor ice accumulations, or glazes, can lead to vehicle crashes and pedestrian accidents, significant accumulations (≥6.35 mm) produced by ice storms can cause incredible damage to power distribution systems, as well as commercial and personal property due to felled trees. In North America, climatologies reveal that freezing rain and drizzle occur in about a quarter of all winter events, and are most frequent during the cool season (December–March) across central and eastern portions of the United States and Canada, Alaska, and the northern shores of Canada. Freezing precipitation events and ice storms have a propensity to develop poleward of fronts or within cold air damming or trapping patterns. In the instances of damming or trapping patterns, the cold air becomes entrenched in the low levels of the atmosphere due to airflow blockage caused by terrain barriers such as mountains. This unique effect occurs on the east side of orographic barriers in central (Rockies) and eastern (Appalachian) North America, Europe (Alps), South America (Andes), and the Far East.

Severe local storms

Smaller scale, but equally impressive, weather phenomena occur as a result of the formation of deep, moist convection. When atmospheric conditions are favorable, specifically in regards to the vertical variation in temperature and wind, thunderstorms can form that are capable of producing frequent lightning, large hailstones, copious rainfall, damaging winds, and even tornadoes. In the United States, the NWS defines these severe thunderstorm events as hail with a diameter of ≥2.5 cm, wind gusts of ≥26 ms⁻¹, and/or a tornado. Commonly, these events are found within the warm sector of extratropical cyclones. However, during the warm season, severe thunderstorms are often independent of extratropical cyclones, developing wherever the basic ingredients required for their formation juxtapose.
Three ingredients are necessary for the formation of a thunderstorm, including a moist layer of sufficient depth in the low or mid-troposphere, an unstable atmosphere, and a lifting mechanism. Essentially, thunderstorms form as theoretical volumes of air accelerate upwards due to differences in density when compared to surrounding air. This process is efficient at relieving the accumulation of potential energy caused by differential heating of Earth’s surface. While thunderstorms are most common in tropical regions, these storms are often short-lived and rarely meet severe weather criteria. In order to obtain more organized thunderstorms capable of producing severe hazards, an additional ingredient must be added – wind shear. This shear, which is a change in wind speed and/or direction with atmospheric height, promotes a separation in the thunderstorm’s updraft and downdraft region, as well as the development of distinctive internal dynamical processes. The addition of wind shear increases the strength and longevity of thunderstorms, often engendering so-called supercell thunderstorms, which are the strongest thunderstorms on Earth. These relatively rare thunderstorms are characterized by their powerful rotating updrafts, or mesocyclones, and ability to produce Earth’s most violent weather. In fact, most strong tornadoes and severe hailstones originate from supercell thunderstorms.

Large clusters of thunderstorms that undergo a mode of self-organization are known as mesoscale convective systems. Found in many tropical and subtropical areas around the world, these complexes are very beneficial to regions dominated by rain-fed agriculture. However, these thunderstorms systems can cause floods, extreme lightning rates, and extensive severe nontornadic wind events known as derechos.

The central United States is home to the world’s highest frequency of severe thunderstorms. The relatively flat topography of the Great Plains allows for little obstruction between warm, humid air positioned over the Gulf of Mexico and its interaction with hot, dry air from the high terrain of the Chihuahuan Desert and cool, dry air originating from the Canadian Prairies. If sufficient wind shear overlaps these intersecting air masses, severe thunderstorms can result. Statistically, this is most likely to happen in the United States during the period April through June when over half of all annual severe weather reports occur. Tornadoes, which are arguably the least understood of all hazards, are also most common in this region and exhibit a relatively predictable annual cycle of occurrence. Throughout the cool season, tornadoes, while climatologically rare during this time of the year, are most frequent in the extreme southern portions of the United States due to that region’s proximity to warm, moist air from the Gulf of Mexico. During early spring, tornadoes surge in frequency and tend to be most common in the southern Great Plains before migrating to the northern Great Plains and Canadian Prairies during summer. This occurs as solar declination shifts toward the Tropic of Cancer, forcing wind shear and lifting mechanisms affiliated with the polar jet stream and its disturbances toward higher latitudes.

Globally, other locations (e.g., Argentina, Australia) also exhibit similar seasonal cycles of tornado occurrence. However, since severe weather reports are biased to areas of population, climatologists favor analyzing data related to the aforementioned ingredients necessary for their formation. Global climatologies of severe weather ingredients reveal local maxima concentrated in equatorial and southern Africa, southern Europe, the central United States, southern Brazil, northern Argentina, eastern Australia, and adjacent to the Himalayas. Such environments are most common in transition seasons, when
low-level temperature gradients are maximized and extratropical cyclones are abundant.

Severe weather can have large socioeconomic consequences, especially during so-called outbreak events of significant magnitude, frequency, and spatial coverage. For example, severe thunderstorms in the southeast United States during April–May 2011 spawned tornadoes responsible for US$17.3 billion in damages and at least 350 fatalities across 20 states. Extremely devastating tornadoes such as these may rank as EF4 or EF5 on the enhanced Fujita tornado intensity scale. While only 2% of all recorded US tornadoes are rated EF4 or EF5, they are responsible for nearly 68% of all tornado fatalities. In extremely vulnerable countries, the death toll from tornadoes can be enormous; for instance, in April 1989 nearly 1300 people were killed in Bangladesh in a single event. The recent increasing trend of economic losses from severe thunderstorms and tornadoes can be attributed to societal changes rather than any detectable increase in event frequency. However, recent research has indicated that the potential exists for an increase in severe thunderstorm environments under future anthropogenically driven climate change scenarios.

**Hydrologic extremes**

Floods are one of the costliest and deadliest hazards due to their relatively high rate of occurrence and the extensive human habitation of river floodplains and coastlines. Nearly 10% of the global population lives in flood-prone areas. According to the International Disaster Database, non-geophysical floods were responsible for nearly 4000 disasters from 1979 to 2013, affecting approximately 3.3 billion people and killing 230,000. An estimated US$600 billion in damages were blamed on floods during this period. Historically, floods have been immense killers, particularly in parts of Asia. The deadliest natural disaster (excluding pandemics and famines) in the history of the world occurred in central China along the Yangtze and Huai Rivers in 1931, when between 1 million and 4 million people were killed.

Floods are classified based on their geography, sources, and rapidity. The deadliest events are associated with quick-onset coastal storm surges accompanying tropical and strong extratropical cyclones (geophysical events such as earthquakes and sub-marine landslides can cause surges, or tsunamis, as well) and flash floods, which are localized, short-duration events often produced by flood control structure failure and/or high rainfall rates in thunderstorms. Due to their swiftness and localized nature, surge and flash flood location and magnitudes are difficult to forecast. Widespread floods are more predictable and can transpire over days, weeks, or even months when excessive rain and/or snow melt occurs in river basins. The intensity, extent, and frequency of precipitation events are important flood determinants, but other contributing factors such as watershed orientation in relation to the precipitation footprint, orography and slope, infiltration rates and saturation of soils, vegetation, land use (e.g., urbanization, deforestation, fallow fields), snow melt and ice jams, and flood control structures (e.g., levees, dams) can influence the occurrence and degree of flood.

Tropical cyclones produce exceptional rates of rainfall, leading to flash and widespread floods not only at landfall locations, but inland areas affected by decaying and remnant systems. Climatologically, rainfall contribution from tropical cyclones is most pronounced over southeast North America, Central America, the Caribbean, countries abutting the Bay of Bengal, far southeast and eastern Asia, northern Australia, and Madagascar. Floods can also be
triggered by monsoons, which are seasonal reversals in regional wind circulations and precipitation produced by stark contrasts in surface heating over continents and nearby oceans. The monsoon is most prominent in south Asia, but weaker (and sometimes, incomplete) seasonal reversals in wind and precipitation occur in west and sub-Saharan Africa, the Mojave and Sonoran deserts of North America, northern Australia, and parts of Indo-China into Japan. In the last case, a front between poleward cool, dry air and equatorward warm, moist air can become quasi-stationary during the warm season. This front, known as the Mei-Yu, can trigger the continual development and training of large thunderstorm complexes, exacerbating rainfall totals. The front has a tendency to align itself along the Yangtze River basin, which can result in enormous and deadly floods. In this region, seven flood events were responsible for over 100 000 deaths each during the twentieth century.

Though tropical cyclones and monsoons are important genitors of floods, most damaging and casualty-producing flood events are caused by localized heavy rains produced by thunderstorms. Climatologies of floods caused by these intense rain events are difficult to assess due to data inconsistencies, lack of reporting, and overrepresentation of cases in some countries. However, reported floods caused by torrential rainfalls are most frequent in East and Southeast Asia, the Indian subcontinent, eastern Africa, Central America, eastern United States, and Europe.

Conversely, extended periods of precipitation deficiency can lead to drought. Drought is a particularly difficult concept to recognize and comprehend due to its slow onset, multifaceted effects on and interactions with human-environmental systems, complexity in identification due to the lack of a universal causative source, and its relativity to each locale’s hydroclimatology. In vulnerable regions, drought can prompt disasters with immense death due to famine and prolonged military conflict that, ultimately, has resulted in some of the world’s greatest human tragedies. Like floods, droughts can occur in almost every region of the globe, but the most devastating human calamities instigated by drought in the past century have occurred in the sub-Saharan Sahel and China. Similar to droughts, heat waves are weather extremes that occur absent storms. These events are particularly deceptive due to their high human mortality without visceral images of damage.

Tropical cyclones

Tropical cyclones, which include hurricanes (Atlantic and eastern Pacific Ocean basins), typhoons (western Pacific), and severe cyclones (Indian), are the most destructive storms on Earth. When these spiraling storms interact with developed landscapes, significant losses to economies and life can occur. For instance, the deadliest weather disaster in the past half century was Cyclone Bhola, which devastated the northern Bay of Bengal coastline in 1970. The costliest weather disaster in history was Hurricane Katrina, which ravaged the central Gulf Coast of the United States in 2005.

Tropical cyclones develop from clusters of thunderstorms that persist over tropical ocean waters in an environment with little or no vertical wind shear. The Coriolis force (due to the rotating Earth) causes the incipient thunderstorm complex to rotate; this force becomes negligible at latitudes less than 5° and, for this reason, tropical cyclones do not form near the equator. Tropical cyclones obtain energy for formation and sustenance from warm seas through sensible and, most importantly, latent heat transfer. Warm oceans readily evaporate water and, through this process,
heat is extracted from the sea surface. This heat is reserved, or held latent, until it is re-released into developing tropical thunderstorms when water vapor condenses in storm updrafts. The tremendous rates of condensation that occur in a tropical cyclone promote a warm-core system, indicating that temperatures in the center of the cyclone are warmer than the areas at the same altitude surrounding the core; this is counter to extratropical cyclones, which feature cold cores.

A tropical cyclone structure includes a tight center of descending air that forms a small area absent of convection known as the eye. Surrounding the calm eye is a ring of deep convective torrents called the eyewall. It is within this relatively narrow region where the storm's most significant hazards, including wind and surge, transpire. Outside of the eyewall, spiraling rainbands encircle and converge toward the center of the storm. These bands include squalls of wind and intense precipitation and, in some land-falling cases, tornadoes. Mature tropical cyclones are characterized by deep low pressures at their core, promoting concentrated pressure gradients and, in turn, low-level winds that can cause extreme damage and encourage the formation of deadly storm surges.

On average, there are approximately 86 named tropical storms (greater than $17 \text{ m s}^{-1}$ sustained winds) per year globally. Of those 86, 47 reach the minimal threshold (greater than $33 \text{ m s}^{-1}$ sustained wind) to be considered a hurricane, typhoon, or tropical cyclone. Two thirds of all tropical cyclone activity occurs in the Northern Hemisphere, with the northwest Pacific basin the most active (mean of 26 named storms per year), followed by the northeast Pacific (17), Atlantic (12), southwest Pacific (10), and southwest Indian (9) ocean basins. Interannual variability in the number of events per basin is significant and is dependent on the seasonal character of sea surface temperatures, the strength of the wind flow in the troposphere, and ambient humidity in development regions.

The northwest Pacific is the only basin that experiences year-round tropical cyclones. Tropical cyclone activity in other basins is most common after a seasonal maximum in the solar declination. For instance, in the Atlantic and eastern Pacific basins, hurricane activity extends from late May or early June to the end of November, with peak activity typically in August and September. The north Indian (south Indian and southwest Pacific) basin typically experiences cyclones from April to December (October to May). Most tropical cyclones are characterized by westward motion during their formative and mature stages due to their usual position equatorward of semipermanent, oceanic subtropical ridges and embedment in the easterly trade winds. Thereafter, some events may curve poleward around the western flank of subtropical ridges, interact with the mid-latitude westerlies, and even transition into powerful, cold-core extratropical cyclones.

Tropical cyclone strength is most often classified based on the Saffir-Simpson scale, which ranks tropical cyclones from category 1 (at least $33 \text{ m s}^{-1}$) to 5 (exceeding $70 \text{ m s}^{-1}$) based on sustained wind speed. While all tropical cyclones have the potential to cause disasters, more intense landfalling events are commonly characterized by devastating winds, storm surges, and flooding rains. Indeed, a disproportionate amount of damage and loss of life is caused by “major” hurricanes (category 3 or greater on the Saffir-Simpson scale) and “super” typhoons (equivalent of category 4 or 5). Increases in coastal population, built environment, and vulnerabilities have led to amplifying impacts, especially as measured in people affected and economic loss.
WEATHER, EXTREME

SEE ALSO: Climate and societal impacts; Climatology; Environmental hazards; Hydroclimatology and hydrometeorology; Natural hazards and disasters; Social vulnerability and environmental hazards

Further reading


Weathering processes and landforms

Tyler J. Thompson  
Ronald I. Dorn  
Arizona State University, USA

This entry relates the appearance of bare-rock landforms to the processes of rock decay. (The term “rock decay” is used rather than “weathering” throughout this entry, because “weathering” no longer functions as a useful scientific term (Hall, Thorn, and Sumner 2012), even though it is still widely used.) This entry introduces the topic by presenting a few classic examples connecting form to process. This is followed by an explanation of why bare-rock landforms occur. Then, case studies illustrate how the appearance of bare-rock landforms relates to rock decay. The last section details the importance of scale in rock decay.

Figure 1 compiles a few classic landforms often taught in introductory physical geography. In these classes, tafoni (1a) are typically linked with salt precipitation. Pressure release shells – sometimes incorrectly called exfoliation – form when enough overlying rock material erodes away to result in the formation of meter-thick shells (1b). Wildfire suddenly heats the water in the outer shell of a rock enough to spall scales off boulders (1c). Erosion of grussified granite exposes core stones that pile into tors (1d). Accumulation of silica, manganese, and iron inside of weathering rinds results in case hardening (1e). Root growth exerts sufficient pressure to open pre-existing fractures (1f). All of these examples occur in sites where soil does not cover bare rock.

The fundamental difference between landforms of mostly bare rock (Figure 1) – the focus of this entry – and landforms covered with soil and regolith derives from an imbalance between the rate at which rock decay produces rock fragments \((P_r)\) and the rate of transport of these rock fragments \((T_r)\). Bare-rock slopes occur in deserts where \(T_r > P_r\) – stripping the rock of its cover of particles. In contrast, soil- and regolith-covered slopes occur in wetter areas where \(T_r < P_r\) – and plants play a key role of holding produced particles in place. A key determinant in whether landscapes are soil-covered or consist of a lot of bare rock is the rate at which rocks decay and soil is produced to cover the bare rock. This entry focuses on locations where soil production has not outpaced mass wasting, resulting in the exposure of rock.

Case study 1: petroglyphs and rock decay

Petroglyph rock art panels display a wide range of forms linked to rock decay (Dorn et al. 2008). Figure 2 illustrates seven of these forms generated by different decay processes: (a) the splintering form of subparallel millimeter-scale cracks (Dorn et al. 2013); (b) spalling of centimeter-thick scales; (c) tafoni growth; (d) block separation along opening joints; (e) lithobiont-enhanced decay (Viles 1995); (f) weaknesses along bedding planes; and (g) salt precipitation resulting in the flaking of a petroglyph panel face.
Figure 1  Classic forms associated with rock decay: (a) tafoni in sandstone at Timna, Negev Desert; (b) pressure release shells in granodiorite, British Columbia; (c) scaling of diorite from a wildfire, Arizona; (d) tor of granite, Sonoran Desert in Arizona; (e) case hardening of sandstone at Petrified Forest National Park in Arizona; (f) root pressure widening fractures in gneiss, Sonoran Desert, Arizona. Image widths approximately 3 m, 12 m, 2 m, 8 m, 0.5 m, 0.2 m for a–f, respectively.
Figure 2  A variety of different processes degrade rock surfaces, illustrated here for sites with petroglyphs: (a) the splintering process on silicified dolomite, South Australia; (b) spalling of joints in sandstone, Utah; (c) tafoni formation through dissolution of grain-cementing agents in sandstone, Arizona; (d) spalling of joints opened by dirt cracking and calcrete wedging, Utah; (e) lithobiont-related crumbly disintegration of sandstone, Utah; (f) enhanced granular disintegration associated with joints that align with bedding planes in sandstone, Arizona; and (g) salt precipitation enhancing flaking of sandstone, Wyoming.

Microclimatic control and mineral assemblage permit communities of various epilithic lichens (Figure 2e) and micro-organisms to inhabit the area between the surface and weathering front of the rock (Viles 1995). The acids generated by bacteria include citric, oxalic, nitric, sulfuric, and others. These acids accelerate dissolution rates of minerals. Where these micro-organisms live within the substrate (whether on the surface or in rock fractures, tunnels, or pores) may impact the resultant decay. For example, ivy overhanging historic buildings can cause deterioration through exploiting fractures (root wedging), but may also moderate the microclimate of the wall reducing thermal stresses. Lichens also moderate the thermal stresses on blocks by storing moisture in their thallus and protecting the surface from temperature changes. Contrasting with these instances of bioprotection, chemical decay rates on olivine were shown to be greater underneath lichens due the combined effects of
higher moisture retention and the secretion of organic acids (Brady et al. 1999).

Figure 3 illustrates the complexity of decay processes that can result in the flaking of a sandstone surface. First, a joint fracture opens slightly, allowing water movement and accumulation of silica glaze along the joint walls. Second, the overlying block erodes the joint face and exposes it to the accumulation of rock varnish and iron film rock coatings. Third, the silica, iron, and manganese in these coatings dissolve and remobilize into the underlying sandstone. This case hardens the outer shell of the sandstone, but it also inhibits water from flowing out of the rock; the case hardening thus acts as a “dam” that allows capillary accumulate underneath the case hardening. Eventually, enough decay takes place underneath the case hardening to reach the point where the surface flakes away.

Case study 2: limestone in different environments

Limestone dissolution typically proceeds along the lines of the chemical reaction: \( \text{CaCO}_3 + \text{CO}_2 + \text{H}_2\text{O} \rightarrow \text{Ca}^{2+} + 2\text{HCO}_3^- \), leading to the basic karst landforms of sinkholes, caves, and karren (White 1990) that occur in a wide variety of terrestrial environments. Consider the small-sized features such as flutes (Figure 4a), runnels (Figures 4c and 4d), slabs called clints separated by deep fissures called grike (Figure 4b). Karren occurs on bare surfaces in virtually all climatic settings. Substantial differences, however, do occur in the larger karst landforms found in drylands and the wet tropics. Stone forests (Figure 4d) and tower karst (Figure 4e) are forms that develop in the wet tropics, even if climatic changes can sometimes shift them into other environments. In contrast,
Figure 4  Limestone karst dissolution landforms in different environmental settings: (a) flutes in semiarid Trans-Pecos Texas; (b) clints and grikes in semiarid northern Israel; (c) large flute associated with rare fluvial discharge in arid drainage basin, eastern California; (d) stone forest in China; (e) tower karst in China; (f) cuesta face, northern Arizona; (g) anticline, Iran.
Arid and semiarid regions tend to have only minimal development of relief associated with dissolution (e.g., deep sinkholes, deep valleys between towers, and forest pinnacles). Arid and semiarid limestone tends to stand out as higher relief relatively resistant to erosion, such as the cliff-forming Kaibab limestone of Arizona (Figure 4f) and anticlines in Iran (Figure 4g). The reason for the tendency of dryland limestone to be more resistant to dissolution clearly relates to the abundance of moisture and biotic acids that enhance dissolution rates.

Scale cannot be ignored in the study of rock decay. Understanding forms at different scales – from individual grains to the scale of outcrops – is of critical importance in diagnosing the current susceptibility of rock-face features to erosion and therefore near-term stability. The system of the rock art stability index (RASI) (Dorn et al. 2008) illustrates the concurrent decay and erosion across the range of visible scales. For example, in Figures 2g and 3, varnished rock surfaces display scaling (erosion of centimeter-scale clasts) along the fringes of the varnish, while the weathering rind erodes granularly beneath the case-hardened surface. Additionally, detached blocks are subject to undercutting through block orientation in rockfalls (Figure 2d). The idea of RASI rests in analyzing current processes and evidence of past processes that influence the threat of erosion to rocks, allowing field scientists to contribute to rock art conservation.

However, connecting processes at the finest scales to the forms seen in the field is both difficult and rarely considered. Consider form of splintering (Figure 2a). The splintering form resembles the pages of a book that went through a cycle of wetting and drying. At the hand-sample scale (Figure 5a), splintering appears to correspond with the development of aligned micron-scale fractures that carry capillary

Figure 5 The weathering form of splintering of a silicified dolomite, South Australia: (a) at the hand-sample scale; (b) a micron-scale backscattered electron microscope image; (c) a nanoscale transmission electron microscope image.
WEATHERING PROCESSES AND LANDFORMS

Micron-scale quartz silt formation

Nanoscale silica spheroid deposition

Varnish at the surface

Case hardening by Fe film

Weathering rind (weakness)

Figure 6  Visualization of nanoscale decay placed within broader spatial scales of rock decay phenomena. Examples presented from nano to landscape scales are: (a) nanoscale silica spheroids a few tens of nanometers across from silica glaze in Tibet (high-resolution transmission electron microscopy (HRTEM) image); (b) micron-scale silt formation from quartz weathering in Arizona (back-scattered electron (BSE) image); millimeter- to centimeter-scale rock coatings and weathering rinds illustrated from (c) Wyoming (BSE image) and (d) Death Valley (case hardened rock shelter); meter-scale forms of (e) a mushroom rock, Arizona, and (f) limestone karst stone forest, Kunming, China; and kilometer-scale landscapes of (g) a salt-encrusted marine terrace, Peru, and (h) varnish-coated alluvial fan, western China. Reproduced courtesy of NASA.
water. Then, at the micron scale (Figure 5b) the fractures transporting capillary water align with nanoscale (Figure 5c) fractures. When these fractures do not align all the way down to the nanoscale, the splintering form does not occur (Dorn et al. 2013).

Landforms generated through rock decay processes are not restricted to any particular scale (Figure 6), but the scale at which a process operates is very important. The new frontier in rock decay studies rests at the nanoscale, where research has been slow to link across scales (Dorn et al. 2013). Nanoscale formation of rock coatings, such as silica glaze (Figure 6a), influences weathering rind and case hardening formation at the micron scale (Figure 6c). At micron scales, for example, weathering processes may dissolve (Figure 6f) and weaken the boundaries between mineral grains (Figure 6b). The cumulative effect of these minute processes may lead to patterns of granular disintegration on the stone (Figures 6d and 2e), the coating of slopes with salt (Figure 6g), and the production of materials that generate alluvial fans (Figure 6h). The environment, be it the microclimate under a boulder, the soil accumulation within a rock fracture, landscape geochemistry, or broad-area precipitation patterns, sets the bounds of rock decay and erosive interaction for the landform.

The ever-changing balance between rock decay processes and erosion generates the great variety of landscapes seen on Earth’s surface. The forms seen in bare-rock landscapes reflect most closely rock decay, since erosion rates far exceed decay rates in these settings. Forms seen in the field result when decay proceeds far enough to result in detachment. Thus, in some settings, geological processes can control mineral compositions and textures that in turn help define chemical decay rates. In other settings, tectonic histories and material properties govern joint density and orientations, thus providing avenues for solution processes between joint walls and planes of future mass wasting failure. Organisms can play both protective and accelerative roles. Rock decay across scales ultimately sculpts landforms by generating points, planes, or zones of relative physical strength or weakness that are then exploited by agents of detachment and then transport.

SEE ALSO: Fluvial erosional processes and landforms; Karst processes and landforms; Mass movement processes and landforms; Periglacial processes and landforms; Tropical geomorphology

References


Further reading


Web-mapping services

Michael P. Peterson
University of Nebraska at Omaha, USA

Cartography has evolved through the years to incorporate new tools and methods of data acquisition, map analysis, and presentation. The discipline can trace its origins to Greek scholars over 2000 years ago, with a history marked by a number of intellectual accomplishments that have furthered understanding of the world and its representation. Beginning in the early 1960s, the computer started to be used to make maps. Initially, paper maps were “digitized,” literally converted to numbers. Eventually, these digitized maps would be distributed between computers through the Internet. Along the way, new and related areas of study developed including remote sensing – the use and analysis of imagery taken by aircraft and satellites – and geographic information systems (GIS) – the input, manipulation, and analysis of geographic information by computer. Today, cartography and these new areas of study are put under the umbrella of “geospatial technology.” A central activity in all areas of geospatial technology is the making of maps (Peterson 2014).

The Internet is a global computer network linking computers in different continents that are thousands of miles apart. Its development was revolutionary for maps, similar to the invention of printing. Maps were not duplicated in mass until the mid-1400s. Before this, they were reproduced by hand and very few existed. As a result of printing, more people had access to maps and, thus, had a better understanding of the world. Like printing, the Internet increased the availability of maps, but the Internet went further and combined the printing and distribution in a single step. One of the major benefits of this new age of mapping is that maps could be made available to the user in a fraction of the time required to distribute maps on paper.

Interaction through the Internet is made possible through an interface that allows the user’s client computer to interact with a program running on a server. Access to the resources of a remote computer is the basis of the Internet. Every time that we use the Internet, we are interacting with a distant computer – increasingly located in massive data centers.

Introduced in 1991, the World Wide Web (WWW) is a particular protocol of the Internet conceived at the European Particle Physics Laboratory (CERN) located near Geneva, Switzerland. It was intended to assist researchers in high-energy physics by linking related documents. The developers wanted to create a seamless network in which textual information on high-energy physics from any source could be accessed in a simple and consistent way. Sir Tim Berners-Lee, director of the World Wide Web Consortium (W3C), played a major role in designing the system.

As originally conceived and implemented by CERN, the World Wide Web consisted only of text. Two university students in the United States, Mark Andreesen and Eric Bina, working at a supercomputer laboratory at the University of Illinois at Urbana-Champaign, recognized that the web would have limited acceptance as a text-only system. They added the display of graphics, sound, and video with the Mosaic web browser introduced in March 1993.
WEB-MAPPING SERVICES

As conceived by Andreesen and Bina, the WWW fostered a series of developments for the delivery of maps. Web mapping services is sometimes narrowly defined as a set of protocols that assist in the delivery of maps and the underlying information. A somewhat broader interpretation is taken here to include any web-based technology that assists with the generalization, manipulation, analysis, and representation of a map. It is perhaps best to begin with a chronological development of these services.

Development of web mapping

The online map is a product of the web. Although maps were distributed through the Internet before the introduction of the graphical World Wide Web, it was the graphical web that made it possible for large numbers of people to access both static and interactive maps. By the end of the 1990s, it was estimated that 200 million user-defined maps were distributed within webpages on a daily basis.

Client/server

The client/server architecture is the major distributed computing model (Figure 1). In this system, clients request services that are provided by servers. The server may be viewed as a dominant computer that is connected to many client computers with fewer resources. In this system, the user’s client computer communicates with a server through a specific protocol. Most often, the server resides in a data center that houses a cluster of computers that respond to user requests. A client computer requests these services using the Internet as the medium of communication.

The distributed model provides an open and flexible environment in which a wide variety of client applications can be distributed and used by large numbers of computing devices, including mobile phones. Client/server computing has already reshaped the way computers are used and is affecting nearly every facet of our lives. Some predict that all computer applications in the future will be “in the cloud” on a system of distributed computers. An example is Google’s G Docs, a series of online applications for the creation and distribution of word processing, spreadsheet, and presentation documents. Amazon Web Services (AWS) provides a platform for bringing applications to the cloud.

Google’s Chrome operating system is an example of a cloud-based appliance. Installed on a basic, sub-$200 computer with a small amount of memory and disk space, the computer is merely an interface to the resources of the cloud. The user accesses programs in the cloud and the files that are created are stored there as well. The Chrome operating system, distinct from the Chrome browser, is an example of a client with very few resources. Most people access the cloud with a Windows or Macintosh computer that has much more computing resources than needed.

Figure 1 The client/server model. The client/server architecture is an example of a centralized architecture where the whole network depends on a central server to provide services.
WEB-MAPPING SERVICES

Servers reside in data centers that contain a large number of computers and are housed in specialized buildings, usually without windows. They require uninterrupted power and diesel generators provide emergency backup in case of a power failure. A lead–acid battery backup system is also in place to temporarily power the computers. It has been reported that, in the case of a natural disaster such as an earthquake, Google data centers in California have contingencies to acquire diesel fuel by helicopter to continue operations. It would be difficult for any individual to implement these types of backup contingencies to maintain the operation of a server.

The main program for serving web content is Apache. Begun in 1995 as a series of “patches” (thus, the name “Apache”) to a hypertext transfer protocol (HTTP) server running at the National Center for Supercomputer Applications (NCSA) at the University of Illinois. Apache is now open source software maintained by the Apache Software Foundation (http://www.apache.org/). The program is available on multiple platforms, including Windows, Unix/Linux, and Mac OS X. In 2013, approximately 60% of all web servers were using the Apache web server, with most of the remaining 40% using a combination of commercial software from Microsoft and Google, and Nginx – the latter also a free and open source project (Netcraft 2013). The Microsoft server has made major gains since 2013 with the growth of Microsoft Azure, a cloud-based server environment. In cloud computing, multiple servers can be implemented on a single computer.

Interactive online maps

On-demand web maps began appearing soon after the introduction of the Mosaic browser in 1993. An early online mapping system was developed by Steve Putz (1994) at the Xerox Palo Alto Research Center (PARC). His Map Viewer program allowed the user’s client computer to create maps from a geographic database. Each interaction with Map Viewer would request a map from a server that was zoomed in on a specific point (Figure 2). The server would respond with a new map that was embedded into a new webpage.

A new era in online mapping was introduced with MapQuest’s user-defined street maps in 1996. Using a standard client/server technology, the user’s client computer would make a request for a specific map, then MapQuest servers would respond to the request by drawing the map from a database of points and lines, converting this to a grid-based raster format and delivering the resultant map within a webpage. Each request for an additional map, at a different zoom level or centered at another point, would

**Figure 2** Xerox Parc Map Viewer was an early example of an interactive web map. The user interacted with a program on a server through the common gateway interface (CGI). The user was able to generate a map of the world at different scales. The resultant map was converted into a graphic file and inserted into a webpage.
WEB-MAPPING SERVICES

Figure 3 A 2001 version of the MapQuest webpage. Dominated by advertisements, the map constitutes only a small part of the webpage. The maps of southern Florida are at three different zoom levels. A total of ten zoom levels were available. Reproduced from MapQuest © 2001 MapQuest – Portions.

result in another server request that would produce another map that would be embedded in another webpage that would update the page on the user’s computer (Figure 3).

Although the entire MapQuest mapping process was fairly fast, there was always a wait for the server to respond. Any zoom or pan required waiting for the server to produce another map that was inserted into another webpage. Response times would also be subject to Internet traffic so a request for a map might take considerably longer when traffic was heavy. Maps would be produced very quickly during the overnight hours but the service would be slow at peak times during the day. This variability in response times was found to be more annoying by users than the generally slow response.

MapQuest quickly became the largest publisher of maps on the Internet. By 1999, MapQuest was responding to 20 million daily map requests (Peterson 2001). Developed by a map publishing company, GeoSystems, MapQuest was a major business success and was soon purchased by the Internet giant AOL in 1999 for $1.1 billion dollars (Rohde 2000, 64). MapQuest held the distinction of having the greatest market share among mapping sites until 2009, when it was replaced by Google Maps.
MapQuest competed effectively against corresponding mapping services from Yahoo!, and Microsoft (Bing). The phrase “I’ll MapQuest it” was used to describe how to find a location, even if using a site other than MapQuest.

**Tile-based mapping**

Google Maps, introduced in 2005, revolutionized the online mapping landscape. Known for its search engine, Google effectively added a map-based search through Google Maps. By not including advertisements around the map, like MapQuest, they left more room for the map on the computer screen. More importantly, from a map user’s perspective, Google Maps streamlined how maps were interacted with.

The delivery of a Google map is based on two major ideas: (i) asynchronous JavaScript and XML (Ajax); and (ii) image tiling. Ajax was the culmination of many years of effort to re-shape how clients communicate with the server. Essentially, AJAX maintains a continuous connection with the server – exchanging small messages in the background even when the user has not made a specific request. This leads to faster server responses. AJAX might be thought of as an application that works in the background of a browser page to anticipate what the user might want, ready to communicate with the server to respond to a request. Operations in Google Maps that are particularly assisted by AJAX include zooming and panning, the most common form of interaction with maps.

Image tiling had been used since the early days of the World Wide Web. In comparison to text, images require more storage and, therefore, take longer to download. A solution is to divide the image into smaller segments, or tiles, and send each tile individually through the Internet. These smaller files often travel faster because each can take a different route to the destination computer (Sample and Ioup 2010). On the receiving end, the tiles are reassembled in their proper location on the webpage. With a moderately fast Internet connection, all of this occurs so quickly that the user rarely notices that the image is actually composed of square pieces. With slower connections, the individual tiles are clearly evident. In a dramatic transformation, all of the other major online map providers – MapQuest, Yahoo, Microsoft (Bing) – converted from the standard client/server interface – popularized by MapQuest – to the AJAX, tile-based method of map delivery within a short time after Google Maps was introduced in 2005. The projection used by all of these online map providers is a slight variation of the Mercator projection, termed the Web Mercator.

Many tiles are needed to implement tile-based mapping. Each is 256 × 256 pixels and requires about 15 KB to store in the PNG format. Table 1 shows the number of tiles that are used for 20 levels of detail (LOD), or zoom levels, and the associated storage requirements and storage costs. With 20 LODs, approximately one trillion tiles are needed for the world. At an average of 15 KB per tile, the total amount of memory required is 20 petabytes, or 20,480 terabytes. No single computer currently has this much storage capacity.

Storing tiles is expensive. If the cost of data storage is estimated at about US $100 per terabyte, storing the entire one trillion tiles on disk drives would be about US $2 million (Table 1). In order to achieve faster response times, there is strong indication that data centers use faster random-access memory (RAM) to cache the map tiles. At $30 per gigabyte, storing all of these tiles in RAM would cost more than US $629 million. Data centers likely use a combination of disk drives and RAM. This would result in a cost of map storage at each data center somewhere between approximately US $2 million and US $629 million. A still faster storage option...
WEB-MAPPING SERVICES

Table 1  The number of tiles, storage requirements, and estimated storage costs used by a tile-based online mapping system to represent the world at 20 different levels of detail (LOD) or zoom levels.

<table>
<thead>
<tr>
<th>Levels of detail (LOD)</th>
<th>Number of tiles</th>
<th>Ground distance per pixel (in meters)</th>
<th>Storage requirements at 15 kilobytes per tile</th>
<th>Disk storage costs at $100 per terabyte ($)</th>
<th>RAM memory storage costs at $30 per gigabyte ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>78,272</td>
<td>60 kilobytes (KB)</td>
<td>0.000006</td>
<td>0.002</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>39,136</td>
<td>240 KB</td>
<td>0.00002</td>
<td>0.007</td>
</tr>
<tr>
<td>3</td>
<td>64</td>
<td>19,568</td>
<td>968 KB</td>
<td>0.0001</td>
<td>0.03</td>
</tr>
<tr>
<td>4</td>
<td>256</td>
<td>9,784</td>
<td>3.75 megabytes (MB)</td>
<td>0.0004</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>1024</td>
<td>4,892</td>
<td>15 MB</td>
<td>0.001</td>
<td>0.44</td>
</tr>
<tr>
<td>6</td>
<td>4,096</td>
<td>2,446</td>
<td>60 MB</td>
<td>0.006</td>
<td>1.76</td>
</tr>
<tr>
<td>7</td>
<td>16,384</td>
<td>1,223</td>
<td>240 MB</td>
<td>0.02</td>
<td>7.03</td>
</tr>
<tr>
<td>8</td>
<td>65,536</td>
<td>611.50</td>
<td>960 MB</td>
<td>0.09</td>
<td>28.13</td>
</tr>
<tr>
<td>9</td>
<td>262,144</td>
<td>305.75</td>
<td>3.75 gigabytes (GB)</td>
<td>0.37</td>
<td>112.50</td>
</tr>
<tr>
<td>10</td>
<td>1,048,576</td>
<td>152.88</td>
<td>15 GB</td>
<td>1.46</td>
<td>450.00</td>
</tr>
<tr>
<td>11</td>
<td>4,194,304</td>
<td>76.44</td>
<td>60 GB</td>
<td>5.86</td>
<td>1,800.00</td>
</tr>
<tr>
<td>12</td>
<td>16,777,216</td>
<td>38.22</td>
<td>240 GB</td>
<td>23.44</td>
<td>7,200.00</td>
</tr>
<tr>
<td>13</td>
<td>67,108,864</td>
<td>19.11</td>
<td>968 GB</td>
<td>93.75</td>
<td>28,800.00</td>
</tr>
<tr>
<td>14</td>
<td>268,435,456</td>
<td>9.55</td>
<td>3.75 terabytes (TB)</td>
<td>375</td>
<td>115,200.00</td>
</tr>
<tr>
<td>15</td>
<td>1,073,741,824</td>
<td>4.78</td>
<td>15 TB</td>
<td>1,500</td>
<td>460,800.00</td>
</tr>
<tr>
<td>16</td>
<td>4,294,967,296</td>
<td>2.39</td>
<td>60 TB</td>
<td>6,000</td>
<td>1,843,200.00</td>
</tr>
<tr>
<td>17</td>
<td>17,179,869,184</td>
<td>1.19</td>
<td>240 TB</td>
<td>24,000</td>
<td>7,372,800.00</td>
</tr>
<tr>
<td>18</td>
<td>68,719,476,736</td>
<td>0.60</td>
<td>960 TB</td>
<td>96,000</td>
<td>29,491,200.00</td>
</tr>
<tr>
<td>19</td>
<td>274,877,906,944</td>
<td>0.30</td>
<td>3.75 petabytes (PB)</td>
<td>384,000</td>
<td>117,964,800.00</td>
</tr>
<tr>
<td>20</td>
<td>1,099,511,627,776</td>
<td>0.15</td>
<td>15 PB</td>
<td>1,536,000</td>
<td>471,859,200.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1,466,015,503,700</strong></td>
<td><strong>0.15</strong></td>
<td><strong>20,480 terabytes or 20 petabytes</strong></td>
<td><strong>2,048,000</strong></td>
<td><strong>629,145,600</strong></td>
</tr>
</tbody>
</table>
would be to use a graphical processing unit (GPU). These devices are specifically designed to store and manipulate images and transfer image data much faster than computer memory. Map storage on GPUs would be at least twice as expensive as on RAM, or nearly US $1.3 billion for the whole world.

These data storage requirements and costs are only for the map. The satellite view, with tiles in the JPEG format, requires approximately the same amount of storage space. Google also stores multiple versions of the satellite image from different dates. Each Google data center would likely have a copy of the map and satellite images, and any other map that is provided – such as the terrain view (offered at only 15 levels of detail). Combining all of these data storage costs provides some indication of the importance placed on maps by Google and other companies.

In 2010, Google switched from raster to vector tiles for some types of map delivery. One impetus for the change was the increased use of maps on mobile devices. Limited by a monthly data plan, mobile users noticed that any application involving maps quickly consumed this monthly allocation. A less data-intensive method was needed to distribute maps to these devices. While the satellite and terrain views are still sent as raster tiles, Google adopted a vector approach for its map layer. Rather than sending tiles as pre-rendered PNG files, the map is sent as a series of lines and shapes. While the primary benefit of this vector approach is a reduction in the amount of data to create a map, it also facilitated changing the style of the map by specifying different colors for features like roads.

Map mash-ups

Probably the most important development in mapping during the first decade of the twenty-first century was the introduction of online mapping tools in the form of the application programmer interface (API). APIs are specialized libraries of computer code that are accessible through the Internet. Soon after the introduction of Google Maps, the company made a library of routines available that would allow for the creation of custom online maps. Users could map their own points, lines, and areas on a Google Map and make these maps freely available to others. The data used for mapping often came from other websites, thus the term “mash-up” to indicate the melding of data and mapping tools to create new presentations of information.

An early application was the mapping of apartment listings from Craigslist, a free service for selling goods and services. New businesses were born simply by combining free data from one site and free mapping software in the form of a mapping API and mashing them together. Websites like MapsKrieg and HousingMaps are examples of this type of combination (search: MapKrieg, HousingMaps). Mash-ups have made it possible to map data that had never been mapped before.

Map mash-ups have had a major influence on how spatial information is presented. One particular advantage of using an API from a major online mapping site is that the maps represent a standard and immediately recognizable representation of the world. Overlaying features on top of these maps provides a familiar and comfortable frame of reference for the map user. This has created a different way of making thematic maps. In the past, thematic maps limited the display of reference information such as cities and transportation networks – partly for simplicity and to emphasize the distribution being mapped. In doing so, the cartographer may have left out critical information to help the user understand the locational aspects of the spatial pattern. It should be noted, however, that adding locational information may also detract from the spatial pattern that is trying to be communicated.
WEB-MAPPING SERVICES

Google Maps API

Introduced soon after Google Maps in 2005, the Google Maps API is by far the most commonly used of any currently available API, whether for mapping or otherwise. The API consists of a series of functions that control the appearance of the map, including its scale and location, and any added information in the form of points, lines or areas and associated descriptions.

The use of the Google Maps API is essentially free, provided the site does not charge for access. Google places a limitation on the number of maps that can be served. A site cannot generate more than 25 000 map loads a day for 90 consecutive days. A map load is one map displayed with the Google Maps API. Once loaded, the degree to which a user interacts with a map has no impact on the map load number.

It would be extremely difficult for the average user of the Google Maps API to reach 25 000 map loads. Even if a site were to go “viral” with a topic that generates considerable interest, it would need to sustain 25 000 map loads per day for 90 consecutive days before the limit would be reached. Usage limits can be placed on a site so that it does not exceed that number. If the site consistently exceeds 25 000 maps a day, Google charges US $0.50 for each 1000 map views beyond this limit. Serving 100 000 Google maps a day would cost $37.50 (75 000/1000*0.5) a month.

Specialized Google Maps API web services have additional usage limits, including:

- directions – provides directions in text form – limited to 2500 a day;
- distance matrix – returns travel distance and time – limited to 100 elements per query and 2500 a day;
- elevation – elevation at points – limited to 2500 requests per day where each request returns up to 512 elevations for a total of 1 280 000;
- geocoding – converts a street address to latitude and longitude – limited to 2500 a day;
- places – returns business establishments and other points of interest around a point – requires an API key and limited to 1000 requests a day.

A Google Maps API key is a numeric code that registers your site with Google. It is not needed for normal applications and would only be required if the usage limits are exceeded, or the Places web service is used. A $10 000 a year pay service, Google Maps API for Business, provides an unlimited number of map downloads and no limits on these specialized services. Over 99% of Google Map API users pay nothing for the use of the service.

The example in Figure 4 shows the JavaScript code and API calls for displaying a simple map that is centered at a specific location. The zoom level, which can range from 0 to 21, is set to 15 under “myOptions.” The center is defined with a specific latitude and longitude value, and the ROADMAP option is selected to define the map style. All of the API calls are made in the “initialize” function. This function is called with “onload” within the body of the HTML file.

There are four basic types of maps that can be displayed:

- MapTypeId.ROADMAP displays the default road map view;
- MapTypeId.SATELLITE displays Google Earth satellite images;
- MapTypeId.HYBRID displays a mixture of normal and satellite views;
- MapTypeId.TERRAIN displays a physical map based on terrain information.

The example in Figure 5 places the basic Google upside-down teardrop marker at the center of the map. The initial zoom level is 15.
The event.addlistener option sets the zoom level to 17 when the marker is clicked. The title text, “Hello World,” is displayed when the mouse is hovered over the marker.

**Overlays**

Overlays are map layers that are placed over a map, satellite, terrain, or digital globe view to provide some type of additional information.
Overlays can be done in two different ways. The first is as a series of points or flags that are placed on top of the map. Google, for example, uses the upside-down teardrop as a standard point symbol to indicate a location.

The second type of overlay is a series of tiles that have the same size and dimension as the base tiles, but these overlay tiles are made transparent so that you can see the tile underneath. Whatever part of the tile is opaque becomes superimposed on the underlying map (Figure 6). Overlaying transparent tiles in this way is faster than overlaying individual points or lines. Most views are transparent tile overlays with a particular theme such as traffic, webcams, or photos.

Another major advantage of vector tiles is that they allow text to be separated from the underlying map as its own overlay. Previously,
the cities in China and Japan were written with local characters. It was not possible to change the text to a western script because the text was “burned-into” each raster tile. By separating the text, it is now possible to switch between western and local scripts (Figure 7).

Online GIS

The growth of geographic information systems (GIS) has been phenomenal since the 1980s. These systems essentially combine a database with query and mapping capabilities. By the late 1990s, web interfaces began to be developed for these systems. Online GIS systems initially only implemented a subset of the overall GIS system functionality.

In online GIS, the user is presented with a webpage that allows a query to be made of a database. The resultant map is then displayed within a webpage and further controls are provided to perform another query. Response times are typically slower than with online mapping systems because the map is usually drawn in the background and then converted to a raster image using the older client/server approach.

The user interface in online GIS is generally not very intuitive and it can sometimes take considerable effort to navigate. Panning and zooming are not as cleanly implemented and this, among other things, makes the experience less than satisfactory. Although often frustrated, a motivated user is usually willing to accept a poor user interface and slow response times in order to get a specific piece of information.

Free and open source software (FOSS) has also become available for online GIS. The most commonly used is MapServer, originally developed at the University of Minnesota. It has been implemented around the world for a variety of online mapping applications. Continual development of MapServer is accomplished through a loose association between developers from various parts of the world.

As with other online GIS implementations, including commercial versions, the person who installs MapServer creates the specific user interface. Many poor decisions that detract from the usability of the program are made during this process. In addition, like most older client/server forms of online map delivery, MapServer sites
tend to be slow. While it is possible to implement a tile-based, AJAX approach, a large number of MapServer sites still use the older client/server method. With minimal server resources, these sites will likely never achieve the map display speed or general user-satisfaction of commercial online mapping systems that have larger data centers, more resources, and are based on the tile-based method of map distribution.

Open standards and open source

A considerable amount of geographic information has been placed into GIS databases since these systems came into widespread use in the 1980s. In order for this information to be useful to more people, standard methods were needed for both “pulling” and disseminating the information from GIS databases. In 1999, the Open Geospatial Consortium (OGC) defined a set of standards for distributing geographic data.

The OGC is an international consortium of more than 420 companies, government agencies, research organizations, and universities established to create a publicly available geospatial interface and format standards. OGC standards provide developers with a way to make geospatial information and services accessible and useful. The Open Source Geospatial Foundation, or OSGeo, is a not-for-profit organization whose mission is to support, promote, and endorse the collaborative development of open source geospatial technologies and data. The foundation provides organizational, legal, and financial support to the broader open source geospatial community. It also serves as an independent legal entity to which community members can contribute code, funding, and other resources. The code and other contributions are maintained for public benefit. OSGeo also serves as an outreach and advocacy organization for the open source geospatial community, and provides a common forum and shared infrastructure for
improving cross-project collaboration. The foundation’s projects are freely available and usable under an open source license (OGC 2014).

The development of standards is a central activity of the OGC. Standards are intended for a common system of communication. The OGC defines open standards in the following way:

- freely and publicly available – they are available free of charge and unencumbered by patents and other intellectual property;
- nondiscriminatory – they are available to anyone, any organization, anytime, anywhere with no restrictions;
- no license fees – there are no charges at any time for their use;
- vendor neutral – they are vendor neutral in terms of their content and implementation concept and do not favor any vendor over another;
- data neutral – the standards are independent of any data storage model or format;
- defined, documented, and approved by a formal, member driven consensus process – the consensus group remains in charge of changes and no single entity controls the standard.

The OGC’s open standards specify interfaces and encodings that enable interoperability between geoprocessing systems from different developers, whether proprietary, application developers or active in open source projects (OGC 2014).

Open standards contrast with open source in that the latter involves some type of code. As defined by opensource.org, the distribution terms of open source software must comply with certain criteria.

- Free redistribution. The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.
- Source code. The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well publicized means of obtaining the source code for no more than a reasonable reproduction cost preferably, downloading via the Internet without charge. The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms such as the output of a preprocessor or translator are not allowed.
- Derived works. The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.
- Integrity of the author’s source code. The license may restrict source code from being distributed in modified form only if the license allows the distribution of “patch files” with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of software built from modified source code. The license may require derived works to carry a different name or version number from the original software.
- No discrimination against persons or groups. The license must not discriminate against any person or group of persons.
- No discrimination against fields of endeavor. The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research.
- Distribution of license. The rights attached to the program must apply to all to whom the
program is redistributed without the need for execution of an additional license by those parties.

- License must not be specific to a product. The rights attached to the program must not depend on the program’s being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program’s license, all parties to whom the program is redistributed should have the same rights as those that are granted in conjunction with the original software distribution.

- License must not restrict other software. The license must not place restrictions on other software that is distributed along with the licensed software. For example, the license must not insist that all other programs distributed on the same medium must be open source software.

- License must be technology-neutral. No provision of the license may be predicated on any individual technology or style of interface.

Web services

The basic purpose of the OGC standards was to both facilitate the distribution of data and make layers of information more accessible. A series of standardized services were defined for supplying geodata universally across any platform. With standardization, map data services are able to interact with and display maps through an Internet-based interface.

Prior to standards being established, extracting information from a GIS database required interacting with a variety of large and complicated databases. The OGC streamlined this process by placing the burden for the extraction of data on the server. As defined by OGC, the web mapping service consists of two functions: (i) GetCapabilities that defines the capabilities of a server such as defining the supported file formats, the available map layers, and the method of display; and (ii) GetMap that tells the database what is needed. The database reads a request and creates the map-based data from the requirements that have been defined by GetCapabilities. The requested data package is then sent to the web mapping service.

In addition to the two primary functions just discussed, most web mapping services support a handful of other functions as well. For example, GetFeatureInfo sends specific information about locations on the map, such as the name of a road or the altitude at a location. Another function, GetLegendGraphic, provides information about symbols used on the map.

The OGC standard has led to the development of a variety of services, including:

- Web map service (WMS) – georeferenced map images typically in the form of raster tiles (PNG, GIF, or JPG). The tiles can also be delivered in a vector format. Requests are made using a standard web URL address.

- Web coverage service (WCS) – a geographical area that can be overlaid upon a map but cannot be edited or analyzed. WCS is used to transfer coverages that consist of objects such as data points, pixels, or paths defined with vectors.

- Web feature service (WFS) – allows requests for geographical features, essentially providing the information behind the map. WFS allows features to be queried, updated, created, or deleted by the client. This data is usually provided in an XML format like GML.

GeoServer

As the reference implementation of the web feature service standard, GeoServer also implements
the web map service, web coverage service, and web processing service specifications using open standards set forth by the OGC. Written in Java, GeoServer allows users to share, process, and edit geospatial data and allows for great flexibility in map creation and data sharing. Built on Geotools, an open source Java GIS toolkit, GeoServer has become standard method of connecting existing information to web-based maps.

The main advantage of GeoServer is that it permits the actual sharing and editing of the data that is used to generate the maps. Others can incorporate this data into their applications, freeing data and permitting greater transparency. GeoServer can display data on any of the popular mapping applications such as Google Maps, Google Earth, Yahoo Maps, and Microsoft Virtual Earth. In addition, GeoServer can connect with traditional GIS architectures such as Esri ArcGIS.

GeoServer reads geographic data in a variety of database and nondatabase formats, including:

- **PostGIS** – an open source software program that adds support for geographic objects to the PostgreSQL object-relational database.
- **Oracle Spatial** – a separately licensed option component of the commercial Oracle database. The spatial features in Oracle Spatial manages geographic and location-data in a native type within an Oracle database.
- **ArcSDE** (spatial database engine) – is a server–software subsystem that is marketed by Esri and enables the usage of relational database management systems for spatial data. The spatial data may then be used as part of a geodatabase, a database that is optimized to store and query data that represents objects defined in a geometric space.
- **DB2** – a family of database server products developed by IBM. These products all support the relational model but in recent years some products have been extended to support object-relational features and nonrelational structures, in particular XML.
- **MySQL** – is the world’s most widely used open source relational database management system (RDBMS) for the administration of websites.
- **Shapefile** – a popular geospatial vector data format for geographic information system software developed and administered by Esri as a mostly open specification for data interoperability among GIS software products.
- **GeoTIFF** – a public domain metadata standard that allows georeferencing information to be embedded within a TIFF file, a computer file format for storing raster graphics images.
- **GTOPO30** – is a digital elevation model for the world, developed by the United States Geological Survey (USGS). It has a 30-arc second resolution of approximately 1 km/pixel, and is split into 33 tiles stored in the USGS DEM file format.
- **ECW** – (enhanced compression wavelet) is a proprietary wavelet compression image format optimized for aerial and satellite imagery.
- **MrSID** – a multiresolution seamless image database developed and patented by Lizard Tech for encoding of georeferenced raster graphics, such as orthophotos.
- **JPEG2000** – an image compression standard and coding system created by the Joint Photographic Experts Group committee in 2000 with the intention of superseding the original discrete cosine transform-based JPEG standard (created in 1992) with a newly designed and more efficient, wavelet-based method.

### Databases

Everything on the Internet relies on a database. Defined as an organized body of information that is arranged for ease and speed of access, databases
WEB-MAPPING SERVICES

are simply the way information is stored and retrieved. In most cases, before anything is put onto a webpage, some type of information for that page has been extracted from a relational database.

Although many different ways of storing data have been proposed, the two most common in general usage are the flat file and the relational database. Spreadsheet programs like Microsoft Excel™ are based on the flat file concept. All records in a flat file are stored in the same number of fields in a table format. A flat file is very good for summing columns of data, or sorting data with a certain theme, but complex queries are difficult to perform. In contrast, a relational database may be viewed as multiple tables of data that are tied together by a key field or fields. The relational database results in less duplication of data and, more importantly, the ability to easily query the data.

Relational databases are the most common and the basis of almost all information presented through the web. All relational databases use SQL, often referred to as structured query language, a computer language designed to input, manipulate, query, and extract data. Development of relational databases began in 1970 by IBM, eventually resulting in a product called SEQUEL. The associated language to manipulate the data became known as SQL and was standardized by the American National Standards Institute (ANSI) in 1986. The latest standard defines a language called XQuery to query data in XML documents. SQL is essentially a set of statements that result in certain actions on data stored in relational tables. SQL is what makes a relational database work.

New features have been added to SQL to incorporate object-oriented functionality and extensions to handle spatial data. In 1997, the OGC published the OpenGIS® Simple Features Specifications For SQL, a document that proposed several conceptual ways for extending SQL to support spatial data. With version 5.0.16, MySQL supports a subset of spatial extensions to enable the generation, storage, and analysis of geographic features.

Spatial extensions refer to a SQL language that has been augmented with a set of commands that code geometry types such as points, lines, and areas. The specification describes a set of SQL geometry types, as well as functions on those types, to create and analyze geometry values. The initial spatial extensions implemented by MySQL were only a subset of those proposed by OGC. More recent versions of MySQL incorporate more of these spatial extensions.

A geographic feature in SQL is anything in the world that can be specified by location. The following are all examples of geographic features:

- an entity, such as a mountain, a pond, a city;
- a space, for example, a town district or the tropics;
- a definable location, such as a crossroads where two streets intersect;
- a location associated with a spatial reference system, which describes the coordinate space in which the object is defined.

The GEOMETRY spatial extension supports any type of point, line or area feature. Other single-value types include POINT, LINESTRING, and POLYGON that restrict their values to a particular geometry type. Still other data types hold collections of values, including MULTIPOLYGON, MULTILINESTRING, MULTIPOLYGON, and GEOMETRYCOLLECTION. GEOMETRYCOLLECTION can store a collection of objects of any type. The remaining collection types (MULTIPOINT, MULTILINESTRING, MULTIPOLYGON) restrict collection members to those having a particular geometry type.
With the spatial extension, latitude and longitude are no longer input as separate fields but as a single entity.

Database programs

The most commonly used relational database software for webpages is MySQL. Pronounced “my sequel” or “my ess cue el,” MySQL is an open source relational database management system (RDBMS). It is used for all types of database applications including e-commerce, airline reservation systems, and online social networks. MySQL is known for its performance, reliability, and ease of use. It can handle large databases with billions of entries. MySQL continues to be adopted by organizations as they discover that it can handle their database needs at a fraction of the cost of commercial software. Any http address that ends with PHP has likely consulted a MySQL database before constructing the webpage.

MySQL is free and open source software (FOSS). FOSS presents a major challenge for commercial software companies because it makes similar software available for free that companies sell for a great amount of money. While no one company is responsible for open source software, there may be many thousands of installations and millions of users. In considering the long-term viability of any software, it is always best to stay with those that have the greatest number of users, whether commercial or open source.

While still free, MySQL is technically no longer open source because it is now owned and supported by Oracle, the largest database company in the world. While the company has stated that it will keep MySQL open and free, many open source developers are switching to another database package called PostgreSQL (PostgreSQL Global Development Group 2010). In many ways, the two databases are very similar – both being based on SQL. At one time, MySQL was considered to be faster but with fewer features while PostgreSQL was seen as slower but with more features. With improvements in both databases, these characterizations are no longer accurate. MySQL has added functions, such as spatial queries, and PostgreSQL has dramatically improved its speed. PostgreSQL with the PostGIS extension is more SQL compliant while MySQL has greater support by Internet service providers. The most important consideration for many is that while PostgreSQL has many sponsors and developers, it is not controlled by any one company – especially one as influential as Oracle.

PostGIS adds support for geographic objects to PostgreSQL (PostGIS 2011). In effect, PostGIS “spatially enables” the PostgreSQL server, allowing it to be used as a backend spatial database for geographic information systems (GIS). PostGIS follows the OpenGIS Simple Features Specification for SQL and has been certified as compliant with the “Types and Functions” profile (PostGIS 2011). As a more specialized service, PostGIS is not commonly offered by web hosting sites, especially those offering minimal services at no cost.

Preprocessors

PHP, or PHP Hypertext Preprocessor, is a server-side scripting language that works with a database. It constructs dynamic webpages on the fly by combining html with information derived from a database. It is the most popular language to create websites competing against other commercial and open source languages like ASP.NET, Java, ColdFusion, Perl, Ruby, and Python. As of 2014, 82% of all web servers use PHP as a server-side programming language (W3Techs 2013).

As an interpreted language, PHP does need to be compiled into executable code. PHP code cannot execute locally on a client computer. This means
WEB-MAPPING SERVICES

that PHP code is not viewable by choosing the source code option for a webpage. The webpage has itself been written by PHP and what is seen by viewing the source code is only the output. Because of this, it is less transparent than JavaScript code. The advantage for programmers is that they can hide their code on the server so nobody else can see it. In that sense, it becomes like any compiled computer language, like C++.

While PHP is the most popular preprocessor, its use varies considerably in different parts of the world. Over 90% of all server-side applications in Europe use PHP but that rate is only 40% in China. Microsoft servers are much more common in China, prompting greater use of Microsoft's ASP.NET (W3Techs 2013).

Web services and the cloud

The cloud will have major influence on the future development of web mapping services. Wikipedia defines cloud computing broadly as: “Cloud computing is web-based processing, whereby shared resources, software, and information are provided to computers and other devices (such as smartphones) on demand over the Internet.” As there was confusion over the terms, the National Institute of Standards and Technology (NIST) was asked to develop a common definition for cloud computing:

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics.

1 On-demand self-service. A consumer can unilaterally provision computing capabilities without requiring human interaction with each service’s provider.
2 Broad network access. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).
3 Resource pooling. Computing resources are pooled to serve multiple consumers with resources dynamically assigned and reassigned according to consumer demand. The customer generally has no control or knowledge over the exact location of the provided resources.
4 Rapid elasticity. Capabilities can be rapidly and elastically provisioned.
5 Measured Service. Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service.

Key aspects of cloud computing are pooling and elasticity. These two characteristics are especially relevant to web mapping services.

Summary

Web mapping services are continually evolving. The major work since 2007 has been to decrease the amount of map related data transmission to adapt to limitations of mobile devices. Open source offerings have also improved while commercial entities have struggled to innovate and differentiate their offerings to provide added value for the paying customer.

The new technology has fundamentally changed the processing and distribution of maps. It has also changed the way maps are used. Multiscale pannable (MSP) maps are now the norm and will forever dictate how maps are presented and used. Mobile devices are adding
ubiquitous blue dot to show the user’s current position, creating a MOMM (Me on My Map) representation.

Maps are a major part of the Internet. Determining location, and navigating, are firmly tied to the Internet. Most people would now be lost with a paper map. Web mapping services is the basis of our new relationship with spatial information.

SEE ALSO: Cloud computing; Geographic information system; Open Geospatial Consortium standards; Overlay, graphical; Representation; Routing and navigation; Scale

References


Further reading

Welfare geography

Andrew Kirby
Arizona State University, USA

Welfare economics emerged after World War II with the goal of determining individual and collective utility. Welfare geography began to develop as a related but distinct subfield in the 1970s, with an emphasis upon collective outcomes manifested in, and measured across, spatial units.

The fundamental logic of Anglo-American geography after 1945 was the examination of spatial difference. Embracing physical and social phenomena its concerns were not social outcomes, but rather differentiation and its measurement – the essence of regional geography. With an increased focus on social science research after 1945, however, geographers began to develop normative principles and empirical techniques. Theories of settlement and industrial location, allied with increased computing power, generated a burgeoning field committed to identifying “what and where.”

A fundamental tension existed within what was increasingly labeled as human (rather than regional) geography. Political sophistication, driven by the civil rights movement and the Vietnam War, caused an increasing number of academics to modify the disciplinary focus to “what, where, and so what?” (or “who gets what, where”). Here, then, we find the basic but essential foundation of welfare geography.

The 1970s was a period in which spatial difference was analyzed in varied ways. Social segregation was documented at different spatial scales, ranging from broad analyses, between and within nations, through to the smallest units for which data were available, namely census tracts in cities. In large measure these were data-driven exercises: the empirical analysis was shaped by the computing packages available, the data that had been collected from public sources, and the scales at which the information was tabulated. Scale was neither a concept nor a construction but a taxonomic device defined by those collecting the secondary data. It is worth noting that what was typically identified as the “quantitative revolution” produced a predisposition toward technique at the expense of a preference for fieldwork (which had tended to migrate to physical geography). Bunge’s Detroit Expedition and Institute, which sought to collect primary data on social difference in the incendiary years 1968–1971, was perhaps a unique example.

Around the time that the Vietnam War ended in 1975, there were, then, several disparate strands of research investigating sociospatial difference. Some was focused upon heavy secondary data manipulation, using social and housing statistics for census tracts in cities, known as the factorial ecology approach. The intention of this research was to identify commonalities in social space, which was connected to abstract principles of social area analysis. In contrast were approaches that marked the emergence of the social indicators movement, which looked explicitly at the manner in which spatial units displayed variations in income and ethnicity, and outcomes such as mortality. In addition, there was a growing questioning of the mechanisms that drove such differentiation: an early instance of this was Harvey’s research on redlining and the manner in which financial institutions shaped the housing market and reproduced racial constructs in urban space (Harvey and Chatterjee 1974).
This research began to diverge in dramatically different ways. The work that was increasingly identified as “positivist” by its detractors pursued technical sophistication, and ultimately contributed to an instrumentalist focus on social and spatial outcomes. The more radical approaches—grounded in the extremes of urban politics in the American city—began to move toward a sophisticated understanding of economic inequalities, expressed via Marxian analysis. And a third way also emerged, that ultimately claimed for itself the explicit term “welfare geography.” This research was most closely associated with David M. Smith who, after stints in South Africa and Florida, was long based at Queen Mary College, London.

Smith published an influential paper in Geography, asking “who gets what, where and why?” (Smith 1974). This work connected the empirical interests of those mapping difference to specific mechanisms, including the operation of the market but also, importantly, the activities of the state apparatus. In this early paper, and then in detail in his later work, Smith laid out a series of propositions. These had been summarized in his first book: “As geographers we have a special role—a truly creative and revolutionary one—that of helping to reveal the spatial malfunctionings and injustices, and contributing to the design of a spatial order of society in which people can be really free to fulfill themselves in a secure social setting where the rights of all are respected” (Smith 1973, 121).

Fears of the imminent collapse of the Western economies, marked by inflation and high unemployment, and political dislocation extending to urban terror, emphasized the pressing need for theories of the state, its local components, and the mechanics of state expenditure. Some of this research had explicit spatial components, such as the UK Inner City initiatives that were very influential in the late 1970s and early 1980s. Other research sought correlations between expenditures and social outcomes, but there was a general sense that much of this work was asking the wrong questions. Increasingly, research sought to position spending and social outcomes in much broader modes of understanding, and this was also true of political science and sociology, where a European neo-Weberianism was especially influential. Some of this research was presented in a summary of a joint NSF–SSRC initiative that occurred between 1981 and 1983 (Kirby, Knox, and Pinch 1984). While this work was innovative, its multidisciplinary nature inevitably placed it at the margins of all the subjects involved, while its political nature gave it a politicized taint. Smith’s work was, for example, seen as both “radical” and rather marginal. Those who actually engaged with welfare geography found its ambitions to be outstripping its abilities.

Welfare geography was ineluctably of its time, caught up with assessments of inputs and outcomes, out of step with the tenor of practical politics, expressed most visibly in the Reagan–Thatcher alliance but most damagingly in the emergence of neoliberal austerity budgets and a roll back in public expenditures. Clearly, empirical work on budgets and outcomes, and theoretical work on state formations might have constituted a valuable counter-weight to this trend. This was also the moment, however, when the postmodern turn began to manifest.

The devastation wrought by postmodernism can only now be assessed across the disciplines. Its impact upon geography did not occur in isolation from other influences, but collectively the result was an emphasis upon critical theory and a concomitant retreat from social and political theory on the one hand, and from empirical work on the other. Postmodernist urbanism (expressed, for instance, in the Los Angeles school) generated an infatuation with text, identity, and design at the expense of more
WELFARE GEOGRAPHY

grounded empirics. Bibliometric research shows few geographic contributions to the study of welfare published after the close of the twentieth century, and equally few citations.

Welfare geography has not left a lasting contribution, either to the broader study of welfare or to the narrower context of geography. As shown in Krugman’s work, economics has demonstrated little interest in geography’s extant body of research. Within the latter, the shift toward identity and emotion has marked a move away from collective outcomes and the manner in which they can be achieved. Subjective assessments of life quality now appear to be much more important than objective measures.

Perhaps the one lasting residual is at the international scale, where comparison between the have and the have-nots is a constant reminder of the issues that link global development and human progress. Smith himself has continued to emphasize the importance of justice in this context.

SEE ALSO: Los Angeles School; Marxist geography; Radical geography

References


Further reading

Wetland biogeography

Joe R. Melton
*Climate Research Division, Environment and Climate Change Canada*

Wetlands are regions that are frequently inundated or saturated for extended periods of time such that they are a distinct ecosystem—separate from the upland and deep-water water systems that they connect. As wetlands form an ecotone between terrestrial and aquatic systems, they provide unique habitats for diverse flora and fauna and many important ecosystem services including flood mitigation of rivers and streams, protection of clean water, buffering of coastlines from high energy waves (such as from typhoons or tsunamis), and habitat for a great diversity of plant and animal species. These ecosystem services have led wetlands to be termed the “kidneys of the landscape” as they receive both water and waste from upstream natural and human sources (Mitsch and Gosselink 2007). When compared to adjacent terrestrial and aquatic systems, wetlands are generally among the most productive ecosystems on the planet (though not all are highly productive, for example, peatlands and cypress swamps have low productivity). Wetlands commonly have anoxic soils which slows the decomposition of organic matter. This slow decomposition over time results in wetlands storing an immense amount of carbon in their soils. Decomposition of this carbon produces proportionately more methane, a potent greenhouse gas, during respiration than upland systems, making wetlands the largest natural source of methane.

Many different definitions of wetlands exist for scientific, legal, and management purposes but they generally share three common criteria. One, wetlands have the presence of water at the soil surface, or within the root zone, either continuously or for extended periods of time. Two, wetlands have unique soil conditions that differ from adjacent uplands, commonly containing hydric soils (soils that have anoxic conditions). Lastly, wetland vegetation and biota are adapted to wet conditions while flooding-intolerant biotas are absent.

Globally, wetlands cover about 7–10 million km² (about 5–8% of the land surface) (see Figure 1). Wetlands are found predominantly in boreal and tropical regions with lesser amounts in temperate latitudes (Figure 2). The general controls on wetland occurrence are climate and geomorphology. While many cultures evolved in concert with wetlands (e.g., the Babylonians, Egyptians, and Aztecs all delivered water via wetlands), and indeed paddies used in rice cultivation are a form of domestic wetlands, regions with high human impact have lost about 50% of their wetlands. Wetlands are generally less impacted by humans in sparsely populated areas, but have suffered up to complete removal in densely populated areas. Wetland destruction has occurred for a variety of purposes—some examples include filling-in for residential or industrial developments; peat mining; drainage for agriculture, mosquito control, or forestry; and indirect effects through the construction of dams.

Wetlands are often divided into three main categories: (i) coastal wetlands, which include salt marshes, tidal freshwater marshes, and mangrove swamps; (ii) peatlands; and (iii) freshwater
swamps and marshes. As the physical and biotic characteristics of these classifications grade between each other, characterizing a wetland in practice can be somewhat arbitrary with the same term used for different systems in different regions.

Coastal wetlands constitute a relatively small fraction of total global wetlands (ca. 5–10%). Salt marshes occur primarily in high latitudes while mangrove swamps become dominant in tropical latitudes. Some characteristic salt marshes occur in the Mediterranean Sea, a famous example being the Camarague in the Rhone River delta. San Francisco Bay also has salt marshes (now over 95% destroyed). Mangrove swamps occur in the Florida Everglades and the extensive Sine Saloum Delta in Senegal. Coastal wetlands are heavily influenced by two primary factors: salinity of the water and the energy provided by tidal forces. Tidal energy influences processes such as gas exchange, sediment deposition, fluxes of mineral and organic matter, and removal of toxins while salinity creates unique habitats for organisms that are able to withstand the stress of salinity ranges from freshwater (yet still influenced by tidal forces) to approaching that of seawater. Wetland plant adaptations to saline conditions include exclusion and excretion of salt, pneumatophores, production of viviparous seedlings, and prop roots.

Peatlands, also termed moors and muskegs, are the most common wetland type globally, covering some 2.4 million to 4.1 million km², primarily in the humid northern boreal regions. Peatlands include fens and bogs as part of a succession in which fens can transition to bogs over time. Bogs commonly have low nutrient availability with no significant water source besides rainwater.
Figure 2 Predominant occurrence of wetlands. Source: Kaplan (2007) in Bergamaschi et al. (2007).

Bogs typically have peat deposits dominated by acid-loving vegetation such as mosses and can be forested or open. The Pocosins of the Carolinas is an example of an evergreen shrub bog. Fen-dominated peatlands are characterized by a flow-through water source and are covered by sedges, reeds, and grasses. There is an extensive body of research conducted on peatlands, especially in the extensive peatland complexes of the Hudson’s Bay Lowlands in Canada and the West Siberian Lowlands in Russia.

Freshwater marshes and swamps are typically differentiated based upon the dominant vegetation type such as: marshes (herbaceous plants), swamps (woody vegetation, trees, or shrubs), sedge or wet meadows (shallow marshes), and wet prairies (intermediate between a marsh and a meadow), among other subcategories. Freshwater swamps cover approximately 1.1 million km² globally and vary from almost continually inundated to only sporadically. Few tree species can exist in continually saturated soils, however, some exceptions exist such as the tupelo/gum (Nyssa spp.) and cypress (Taxodium spp.) prevalent in the swamps of the southeastern United States. The Great Dismal Swamp (southeast Virginia and northeast North Carolina, United States) is an example of a predominantly bald cypress-gum swamp. Trees are also the dominant vegetation in many riparian zones, such as floodplains, that are not continually saturated but are flooded often enough to be classified as wetlands. Freshwater marshes do not have extensive peat deposits like peatlands, but have a shallow-water regime and soft-stalked emergent aquatic plants. The Mesopotamian marshlands of southern Iraq and Iran were originally a large freshwater marsh complex but have been heavily damaged by human land use.

SEE ALSO: Ecosystem services; Riparian ecosystems; Soils of wet and hydric landscapes; Wetlands hydrology

References


Further reading

WETLAND BIOGEOGRAPHY


Wetland hydrology

Tim Beach
Jonathan M. Flood
University of Texas at Austin, USA

Wetlands are defined based on soil types and ecosystems and on the presence of water, meaning duration, level, and quality of water. A wetland can be perennially wet or seasonally wet, depending on multiple classifications. This entry considers the hydrologic conditions inherent in the types, functions, evolution, landscape connections, and human alterations and trends of wetlands. Hydrologic conditions play a paramount role in creating and maintaining the structure, function, and unique physiochemical conditions that define wetlands. Water level, hydropattern (or hydroperiod), water chemistry, and residence time are the primary hydrologic variables influencing most wetlands. Wetlands, like all terrestrial ecosystems, are the complex product of the Earth’s physical spheres within the wider concept of the critical zone, which is the slice of Earth from the top of the vegetation to the bedrock (Figure 1).

Water level within wetlands is principally a function of hydrologic input (i.e., precipitation, runoff, stream and groundwater inflow), hydrologic output (i.e., evapotranspiration, stream and groundwater outflow, and percolation), and geomorphology in terms of landscape position and sediment flow (Figure 2). Water content directly influences the dissolved oxygen within submerged soil matrixes, which, in turn, influences the structure and distribution of plant and animal communities within the wetland. Deeper water levels and more biological oxygen demand (BOD) generally lead to more anaerobic soil/sediment conditions. Many emergent plants cannot survive these oxygen deficiencies in their root zones. Shallow waters have a greater capacity for oxygen delivery and exchange with the submerged matrixes and generally host a greater variety of emergent plant species and other primary producers. Water level and anoxic conditions drive and maintain the creation of hydric soils; both influence the biotic community (Figure 3).

The majority of the world’s wetlands do not maintain a perfect balance between hydrologic inputs and outputs over the course of a year. Water level tends to fluctuate over time in most wetlands and, typically, mirrors larger-scale seasonal patterns in precipitation and temperature or even longer term aquifer-related fluxes. Wetland hydropattern (or hydroperiod) is a measure of temporal variability in water level. A hydrograph is the common method for plotting flux in water level as a function of time, which could be over a short- or long-term scale, depending on whether the interest is in variation over a flood cycle or an annual cycle. The constancy of the hydrologic pattern from year to year characterizes each type of wetland and represents a degree of regularity that determines the types of plant and animal communities. For example, monsoon or strongly seasonal tropical watersheds go through long flooded periods on floodplain wetlands in the Amazon Basin called Igapô and Várzea forests (Junk et al. 2011).

The hydrologic residence time is a measure of the amount of time it takes for a hydrologic input to pass through a wetland and exit the system. Hydrologic residence times may be different for different sections of a wetland.
WETLAND HYDROLOGY

Figure 1  Wetland spheres.

Figure 2  Wetland, terrestrial, and aquatic systems.
Water may run faster through the deeper, better connected sections of a wetland, while shallow systems remain still and stagnant for greater periods of time. Water level within an entire wetland may fluctuate collectively over a year but different sections with differing residence times will create dissimilar conditions. The network of biotic and abiotic processes at work in wetland systems complicates specific measurements but general hydrologic residence times are a useful metric. Residence time influences the amount of dissolved and particulate material (nutrients and toxins) entering and exiting a system. In coastal wetland systems the residence time directly impacts salinity concentrations.
Scales and classes

The scale of wetland hydrology ranges from less than a hectare vernal pool in California or Prairie Pothole in Minnesota to the nearly 20 million hectare Pantanal wetlands of South America and the seven million hectare Llanos de Moxos wetlands of Bolivia (McClain 2002), the largest protected wetlands on Earth. All wetlands have diachronic and synchronic characteristics, indicating their evolution over time and in context with surrounding environments. Wetlands are usually both evolutionary over time and ecotonal over space because the hydrology that maintains them is always changing. The US Clean Water Act (33 CFR 328.3 (b); 1984) defines wetlands as “those areas that are inundated or saturated by surface or groundwater at a frequency and duration sufficient to support, and that under normal circumstances do support, a prevalence of vegetation typically adapted for life in saturated soil conditions.” Cowardin also added soil to this definition of wetlands by including the notion that water was also a dominant element for soil development and soil ecosystems in wetlands (Cowardin et al. 1979).

In soil terms, wetlands with consistent saturation usually have Histosol (dominantly organic matter) soils with O horizons and sometimes L (limnic) horizons of open water deposition. Where water tables are seasonal, they have many other soil types often with the aquic descriptor term of reducing conditions (see Soil water). Wetland soils often have the horizon designation of “Bg” or “Cg” with “g” signifying “gleyed” or reducing conditions. In groundwater terms, wetlands reside where the phreatic zone is near the surface, though they may seasonally be in the vadose zone (Figure 3). Since saturation is a defining feature, an area with a high water table within the capillary fringe would not necessarily be a wetland, though many such areas become saturated or inundated for some part of the year.

Geography

The distribution of wetlands based on latitude shows a peak in area in the mid to high latitudes (45–70° N) in the Northern Hemisphere and a secondary peak in the equatorial zone (Figure 4). The main reason for this distribution is the vast area of formerly glaciated lowlands with deranged drainage in North America and Eurasia. No equivalent peak occurs in the Southern Hemisphere’s high to mid latitudes, ironically in the latitude of the great Pantanal wetland, due to the small area of land and deranged drainage in this hemisphere. The area of wetlands decreases in the lower latitudes of the Northern Hemisphere through the global belt of arid climate to about 10° N. Wetland area rises again in the equatorial zone with increased precipitation and runoff of this rainiest zone of the planet but also with well-developed and efficient drainage networks. Poleward of 45° S and 70° N wetlands are tiny areas because most of this zone is open water or glacial ice (Figure 4).

Types

The international Ramsar Convention classifies wetlands based on hydrology, which include 12 marine/coastal, 20 inland, and 10 human-made wetland types (www.ramsar.org). The Ramsar Convention, officially titled the Convention on Wetlands of International Importance, is an international agreement signed in Ramsar, Iran, in 1971, which now has 169 signatories. The treaty’s three pillars are to work together internationally for wise use, list and safeguard wetlands of international importance (more than 2200), and work together on transboundary wetland
Figure 4  Global distribution of wetlands.
WETLAND HYDROLOGY

systems. Ramsar has a comprehensive definition of wetlands by including marine classes and human classes of wetlands.

The geomorphic controls on wetlands include any agents that create depressions or high water tables, and most wetlands have a geomorphological cause of their high water table. The Ramsar classification describes the coastal geomorphic classes of saline (e.g., sea grass beds), brackish (e.g., intertidal marshes), or freshwater (e.g., lagoons). Ramsar also describes karst sinks and fluvial systems such as deltas, floodplains, springs, and geothermal areas. To this can be added other sinks created by landslide dams, a host of glacial depressions such as the Prairie Potholes, aeolian blowout hollows, fault lines, especially tension faults with sinking blocks, and bolide impacts.

The US Environmental Protection Agency (EPA; www.epa.gov/wetlands/) categorizes wetlands succinctly as marshes, swamps, bogs, and fens. Marshes, for example, can be tidal and nontidal and are usually covered by pH neutral water. They feature soft stemmed, emergent plants anchored in saturated soils. Examples include both deeply and perennially inundated lands like the Everglades to more seasonally inundated (after spring snowmelt) Prairie Potholes. Tidal marshes include the lower, strongly tidal zone and the upper marsh.

Swamps have woody plants and have seasonal water tables ranging from occasional flooding to a saturated soil surface. These may be dominated by maple trees (Acer spp.) in the higher latitudes of the United States and Canada or Bald Cypress, White Cedar, and Tupelo trees in subtropical areas of the United States. An iconic type is the nearly year-round Bottomland hardwood swamps with buttressed Bald Cypress trees and oak and sweet gum trees, as with the Mississippi River valley. These swamps usually have alluvial and organic, nutrient-rich soils. Another globally important type is Mangrove swamps, which act as barriers to coasts but are endangered by hydrology from sea level rise and freshwater abstraction and rampant development.

Bogs usually are precipitation dependent, have mosses, especially sphagnum moss, and develop peat layers. They may form either as lakes fill in from the sides or as wet upland develops a sphagnum moss surface that traps moisture, termed paludification. The formation of such blanket bogs occurs naturally or as deforestation removes the heavy transpiring trees, leaving wetter conditions pioneered by water retaining sphagnum moss. The vegetation is often acidic, which imparts a low pH to these wetlands, and has low levels of decomposition, which makes them hotspots of carbon sequestration. These occur both in poleward latitudes and in more subtropical conditions as with the pocosins of the southeastern United States, which are on uplands with shallow water tables and evergreen shrubs and trees.

Fens are groundwater- and runoff-fed, peat-forming wetlands that are less acidic and more nutrient rich. They are more biodiverse with woody or herbaceous vegetation and, like bogs, often occur in higher elevations and latitudes. These may become bogs if the peat growth separates them from runoff, and they grow to form in just the acidic mossy vegetation.

Wetland hydrology functions

Wetland environments support, generate, and interact with a large web of biotic and abiotic processes that influence distant components of the biosphere, atmosphere, and hydrosphere (Figure 1). Wetlands provide many ecosystem services as a result of their dynamic hydrology: storing water and delivering the water downstream and/or allowing aquifer recharge through infiltration and percolation; serving to decrease
the impacts of flooding and storms by storing water and moderating storm energy; acting as sediment sink to remove sediment loads from streams; providing habitats for many species; and serving as nurseries for numerous fish, insect, reptile, bird, arthropod, and amphibian species. Wetland bodies and hydrologic corridors serve as transportation networks for species and economic activities. Wetlands are important for plant domestications and paddy rice agriculture and have an ancient and worldwide history of human uses.

**Human interactions with wetland hydrology**

Humans have a long history of both positive and negative interactions with the hydrology of wetlands of many types (Beach and Luzzadder-Beach 2013). Cultures around the world have actively and complacently used perennial wetlands for numerous resources, including hunting, fishing, gathering, and transportation. Humans have also developed intensive agricultural systems in wetlands around the world. Intensive systems include active manipulation of wetlands to maintain their water levels, and to drain water, associated with “miasma” or disease, to make dry, healthier farmlands (Larson and Kusler 1979).

The Ramsar list of internationally important wetlands has both significant natural wetlands like the Ria Celestun, Yucatan, Mexico, where groundwater interfingeres with saltwater in an estuarine channel on the northwest coast of Yucatan, and human-made ones like Xochimilco in Mexico City, where Aztec farmers built chinampas fields between canals into a lake. These two examples show the importance of hydrology on wetlands in many ways. First, lowered water-flow in the Ria Celestun results in salt-water intrusion inland and a rearrangement of habitats and reduced foraging for charismatic species like flamingos. Second, ancient Aztec farmers dammed out water of lower quality and maintained water heights to grow a wide diversity of crops, which farmers have carried on to the present. Beyond Ramsar, UNESCO lists several cultural wetlands as world heritage sites, including the Subak system of rice terraces of Bali and the rice terraces of the Philippines.

To build wetland field systems means the maintenance of hydrology, and they represent widespread global success because they range from 45°N to 40°S and sea level to nearly 4000 m. Mainly the hydrology is to water crops but it is also for moderating temperatures in places like the Highlands of Mexico and Bolivia. In the tropical lowlands of Mesoamerica and South America were vast wetland field systems that may have persisted until European conquest or declined with the varied fates of civilizations, such as those as the Maya civilization a millennium ago (Beach *et al.* 2015).

The iconic chinampas of Xochimilco and other wetland field systems around the world relied on maintaining field plots between canals and using the canals for maintaining soil water through irrigation and capillarity from high water tables (Beach and Luzzadder-Beach 2013).

In contrast, paddy rice wetlands in Asia generally relied on growing rice in wet paddies with the monsoon water up to a phase in their phenology. Rice can transport oxygen from the air to roots but even rice has limits to the flooding and needs to dry out for the final maturation of the seed to harvest. Other cereal crops need better drainage, and the indigenous wetlands of the Americas required raising beds above the water table or lowering the water table for drainage for their maize, beans, squash, and other crops. The latter implies alteration of hydrology through canals, subsurface drains or tiling either horizontally or vertically, and diversions.
Another important human impact on wetlands is from deforestation, because grasslands and croplands have about one-third less evapotranspiration than do forests (Woodward et al. 2014; Zhang, Dawes, and Walker 2001). In many environments, 9–12% of global wetlands, deforestation leads to the water equivalent of 10–15% more precipitation annually. Moreover, if these areas stay deforested, they should stay wetter (Woodward et al. 2014). Countering this increased water in environments, especially in tropical forests, is decreased precipitation over time linked to the decreased transpiration into the atmosphere, seen in large-scale empirical and model studies (5–10% for Congo basin simulations in Akkermens, Thiery, and Van Lipzig 2014; Lawrence and Vandecar 2015).

All types of wetlands have declined around the world (https://www.fws.gov/wetlands/), often due to hydrological changes such as groundwater table drawdown, tiling and other forms of drainage, and vegetation changes. Ramsar states 87% of world wetlands has been lost since 1700 CE and wetland losses have accelerated in the twentieth and twenty-first centuries, with a meta-study concluding a 64–71% wetland loss since 1900 CE. Since the 1980s, these rates have slowed in the United States, Canada, and Europe but continue unabated in Asia (Davidson 2014). The ecosystem services from wetlands have also declined by almost 10 trillion dollars (in 2011 dollars) from 1997 to 2011, caused by declines in area (Costanza et al. 2014).

Many wetlands will be vulnerable to climate change from greater water abstraction, higher evapotranspiration, and sea level rise (Zedler and Kercher 2005). In permafrost regions, climate warming is especially severe (IPCC 2013), which is leading to loss of permafrost and new types of wetland or to losses in wetlands in some places, depending on their changing water table levels and water balance inputs. Permafrost is likely to decline further, 30–70% by 2100 (Schuur et al. 2013).

The interaction between sea level and wetlands takes many forms. Wetland drainage leads to higher sea levels, likely rising too fast for wetlands to maintain their equilibrium by growing upward with peat accumulation. Moreover, many wetlands will not be able to migrate inland in response to sea level rise because of competing land uses like urban infrastructure and even future seawall construction.

Many regions are also changing in the face of accelerated sea level rise, such as the Chesapeake Bay region. Other land-use changes, such as deforestation and reforestation and the removal and reintroduction of beavers, have changed wetland hydrology (Pollock, Heim, and Werner 2003). Future sea level rise may occur at too rapid a rate for wetlands to adapt to new states, and many wetlands may suffer further erosion of their functions, but wetlands have a proven ability to counter coastal erosion (Spalding et al. 2014; Barbier et al. 2013). There are many wetland enhancement programs to help meet these changes and mitigate the impacts of storms on coastlines in developing countries, but not in undeveloped ones (Denny 1997).

Human impacts are also in the form of wetland restoration and mitigation, which always requires restoration of hydrology to restore some part of the function of a wetland. Wetland restoration efforts include passive measures designed to halt damaging actions and active measures that primarily focus on the restoration of the hydrology in order for biological recovery to occur (Kauffman et al. 1997). Several iconic wetland restoration sites, such as the Kissimmee River in the United States and the “Garden of Eden” (Iraqi marshes), successfully re-engineered hydrologic inputs to increase wetland areas, but the loss of wetlands to hydrologic changes far
WETLAND HYDROLOGY

outweighs these successes worldwide. Unfortunately, the databases are lacking for wetland losses for much of the developing world and there is, thus, a strong need for understanding trends in this most important part of global environment.

The future of wetland hydrology

Multiple factors are likely to influence the hydrology of wetlands over the next decades. One factor is likely climate change, especially in the high latitudes of the Northern Hemisphere, where so many wetlands occur in the zone with the highest rates of warming. Here, more permafrost will melt but increased evapotranspiration without equal precipitation increases will reduce wetland area. Further wetland drainage and land-use conversion will likely continue, with the net effect of lowering water tables, although the initial response to deforestation is often raised water tables. In many tropical areas, deforestation will also likely decrease rainfall, which may transform some wetlands to drylands. Drainage of wetlands continues in many forms, from channel stabilization to tiling and canalization. Sea level rise will inundate many wetlands; in some places, new wetlands will form inland but with reduced ecosystem functions (IPCC 2014). In other places, new sea walls will counter wetland formation to maintain private and public property. Some coastal wetlands can keep up with lower rates of sea level rise but sea level rise in the next decades may be too fast.

SEE ALSO: Geoarchaeology; Geomorphic systems; Paleoecology; Permafrost: definition and extent; Soils in geomorphic research

References


Further reading

Whiteness

Alastair Bonnett
Newcastle University, UK

White and whiteness are terms used to describe the skin color of people of European heritage and, hence, to depict them physically as a race. From the late eighteenth century this identification became tied to the development of ideas of racial difference and hierarchy in which the possession of whiteness was used as: (i) evidence of natural, innate, superiority; and (ii) the norm and standard of racial identity against which others (the “nonwhites”) are measured.

Once it had been established as the central symbol of European racial power, whiteness began to take on an important function in the organization of colonial societies. Three basic roles may be discerned, the importance and practice of each varying considerably between different countries and regions and none of which should be taken to imply that “whites” can be approached as a homogeneous group. The first is ideological: whiteness was used as a justification for the power and prestige of European heritage peoples, as well as constituting a site of solidarity and identity for this group. In this way a racial bond was established, a bond which ensured that white people had a cultural and economic stake in aligning themselves with other whites. Second, whiteness became a mechanism for the organization of society. This function is particularly apparent in the logic of colonial economies, where access to different occupations and labor organizations; familial, cultural, and social affiliations; and economic leverage (i.e., what one’s labor was worth) was often determined by one’s position vis-à-vis whiteness (either by way of a white/nonwhite binary split or through a gradual loss of socioeconomic power the further away from whiteness one was imagined to be). Third, whiteness developed a psychological dimension: the formation of personality through the construction of self and other became bound up with a racialized set of fears and desires concerning the nature of whiteness (especially in respect to notions of purity and defilement) and its loss or acquisition; however, it is worth reiterating that these processes were all complicated by established and, sometimes, entirely distinct dynamics of ethnic separation that meant that whiteness was cross-cut with religious distinctions (such as between Catholic, Protestant, and Jew) and ethno-national differences. These distinctions and differences suggest that polemical generalizations about the behavior or status of “white people” rarely sustain detailed historical and geographical analysis.

Nevertheless, not only has discussion of “whiteness” been dominated by political and campaigning voices, but many of these interventions have been highly perceptive. One of the first critical investigators of whiteness as a social and historical phenomenon, the African American writer and activist W.E.B. Du Bois, writing in the early 1930s about the 1860s and 1870s, noted of “the white group of laborers” in the United States, that,

while they received a low wage, were compensated in part by a sort of public and psychological wage. They were given public deference and titles of courtesy because they were white. They were admitted freely with all classes of white people to public functions, public parks, and the best schools … The newspapers specialized on
WHITENESS

news that flattered the poor whites and almost utterly ignored the Negro except in crime and ridicule. (Du Bois 1992, 700–701)

Du Bois also exemplifies another of the ironies of whiteness: that whiteness is often more visible to people who feel whiteness acting against them than to white people themselves.

Whiteness is not, however, a static phenomenon. Its changing cultural and economic role across the world shows it to be a persistent but flexible current. Much antiracist discussion of whiteness over the past 30 years has been dominated by a concern to show that whiteness is still relevant and that the hierarchies and inequality associated with it remain important. The defensive quality of these assertions is the result of widespread yet diverse arguments and processes that suggests that white racism is not the force it once was; that distinctions of culture and religion have become more pressing than “old-fashioned” ideas of color, and/or that, in a multipolar world, ideas about the origin of race and racism need to be diversified beyond a sole focus on the West and its global impact.

The shifting nature of whiteness is also apparent from a study of its history and its changing deployment in the discipline of geography. The imperial geographies of the late nineteenth and early twentieth centuries constantly strove to map out the limits and possibilities of white settlement and colonial control. However, it would be a mistake to read these applications of racial environmental determinism as indicative of a confident and unchallenged white supremacism. Rather, they formed part of a larger debate on a perceived crisis of white authority and colonial legitimacy. The focus within this body of geographical research was, after all, on the limits of white rule. From 1880 to 1930 a considerable literature arose on the “perils” and problems of whiteness, its vulnerability to a myriad of challenges. Common sites of crisis identified were internal racial strife (the “fratricidal” nature of World War I emerged as the prime illustration), the rising power of the nonwhite world (the outcome of the 1904–1905 Russo-Japanese war quickly became the principal example), the racial treason of Bolshevism, and the rise of the “underman” (i.e., the white proletariat).

From the late 1930s onward, the idea of openly celebrating whiteness – of talking about it as a positive identity, as something to be proud of – was gradually dropped within British public discourse and, increasingly if unevenly, across Europe and North America. At the same time, the idea of “the West” and “Western” (terms that had been developing contemporaneously with the crisis of whiteness literature) began to come to the fore. The same pattern can be witnessed within academic geography. White attitudes became things to be studied, and eventually “whiteness” became a term to be problematized (indeed, it can sometimes appear to be the sole object of antiracist scrutiny). But “the West” and “Western” emerged as categories to be deployed in a taken-for-granted fashion, understood not as constructs but as coherent and meaningful expressions.

Despite its declining status within public discourse, white identity continues to be significant, within both everyday and popular culture (including among Western antiracists) and in coded or euphemistic forms within government discourse. The function of whiteness in symbolizing or producing “normal” space in the West rests on the social fact that nonwhite people are far more likely to be visible, to be considered to be “out of place” than whites. Ingrid Pollard’s photographic meditation on her (black British) presence in the English countryside provides a straightforward illustration. Her images work by producing an immediate but uncomfortable thought: that black Britons’ place is in the city. John Urry (1995, 27)
asserts that there is a “racialisation” of the phe-
nomenology of the urban; and this works partly in
England through the contrasting high valuation
which is placed upon the English countryside,
which is taken to be predominately white.”

Yet, even within the “cosmopolitan” and
“multicultural” city, whiteness has its territories
and spatial limits. Recent studies affirm that it
is only when “white places” and “non-white
places” come into contact that the former
become visible, the reality of white space
becoming legible against a darker background.
Thus, for example, Wendy Shaw’s ethnographic
research on Sydney suggests that,

Away from the stark black/white racialised
boundary near The Block [i.e., the Aboriginal
identified area of Redfern], where the space of
whiteness absorbs other ethnicities, whiteness
appears to fade into ethnic neutrality. Away from
the Aboriginal “other,” whiteness is not so visible
... whiteness strengthens and consolidates against
the presence of The Block. (2001, 8)

The geography of Britain, the United States,
and Australia is also racialized at other spatial
levels, such as the street (McGuinness 2000), the
shopping center (Jackson 1998), the suburb (Watt
1998), and the town. The role of the town as a
kind of haven of the white urban, defined against
the menacing presence of the “cosmopolitan”
city, appears to be particularly important in
parts of the United States (Kobayashi and Peake
2000). Anoop Nayak (2003) has researched the
intersections of globalization, youth culture,
and whiteness in both city and suburban loca-
tions. A methodological literature that challenges
researchers to reflexively identify the influence of
whiteness on research assumptions and practice
within the discipline of geography has also begun
to emerge.

It is one of the ironies of both contem-
porary ethnic and racial studies and cultural
geography in the West that these two scholarly
traditions – which might be assumed to offer a
resolutely antiparochial frame of reference – have
tended to restrict the focus of their inquiries
to Western countries. The shift in the latter
half of the twentieth century, away from the
racist assumptions of colonial era geography
and anthropology and toward a scholarship of
self-doubt and self-questioning, may help to
explain this reticence to “explain” other soci-
eties. However, although it remains limited in
scope, study of the role of whiteness outside the
West has attracted a growing interdisciplinary
literature. One of the foci of this work is on the
interplay of neoliberalism and whiteness. This
interest has developed, in part, because it has
been noted that, despite the almost universal
abandonment of explicit doctrines of white
supremacy, and the adoption of antiracist rhetoric
as the lexicon of legitimacy by institutions the
world over, “cultural whiteness” continues to
be globally influential. The patterns and paths
of resistance to this process are diverse, yet if any
one attribute of the white racial norm stands
out from the last century it is its capacity for
adaptation. Today it appears that the durability
of whiteness is a function of its changing rela-
tionship with modernity. It is pertinent to note
that a biologically identified view of “the white
race” as the physical carrier of modernity was
central to its early development. This association
was most starkly apparent within the successive
campaigns of governments throughout South
America in the late nineteenth and early twenti-
eth centuries to modernize both their economies
and their social structure by “pouring in white
blood” through assisted immigration schemes
for Europeans. In the latter half of the twentieth
century this phase declined in significance and
a new point of emphasis within the relationship
between whiteness and modernity was forged. In
contemporary South America the white ideal is
“enforced,” not necessarily through the physical
WHITENESS

presence or even the agency of white people, but rather through a close association between the aspirational lifestyle of consumer capitalism and the actual lifestyle and culture of the white West. The white ideal is not limited to South America. However, to date, Latin Americanists have provided the most thoroughly researched explications of this process. In 1997 the Peruvian activist Patricia Oliart identified an association of the internationalization of economic and media interests with the reinvention of the white European as the symbol of modernity, as the corporeal marker of social progress and physical attractiveness.

What is happening now is that racism is coming back stronger ... Money counts again now and the way you look, we have lots of gyms, that we never had before, all classes doing aerobics and dying their hair, like a blond hair ... it doesn't matter if you're not white, you can look white, you can become white, you can wear nice shoes, you can dye your hair, you can get a great body, and if you don't do that, you have a ponytail, wear ethnic skirts or whatever, then that's your problem. (Oliart 1997, quoted in Bonnett 2002, 355)

Oliart’s observations indicate that whiteness is being connoted as a lifestyle, symbolically tied to the pleasures of a consumption-led identity (pleasures such as “freedom” and “choice”). A further implication of this process is that, far from being an archaic ideology from a discredited past, the white ideal is being reimagined and appropriated. In its new guise, this racial and cultural archetype is less identifiable as the foreign belief system of white supremacists. It is offered instead as part and parcel of a widely disseminated aspirational agenda – a process that entails its availability to local reinterpretation, mutation, and transgression as well as emulation.

SEE ALSO: Antiracist geography; Identity; Inequality; Race and racism

References

Wicked problems

Steve Rayner
University of Oxford, UK

Origins

Horst Rittel and Melvin Webber (1973) first identified wicked problems in the late 1960s. Professors in the planning department at the University of California, Berkeley, they contrasted the relatively straightforward challenges of the late nineteenth and early twentieth centuries, such as public health and basic sanitation, with the more difficult challenges facing planners in the latter half of the twentieth century, such as the redevelopment of entire neighborhoods.

Rittel and Webber (1973) contrasted complex problems of social policy with more straightforward puzzle-solving kinds of activities typical of mathematics and the natural sciences. They also recognized the value conflicts arising from the increasing heterogeneity of contemporary communities and emerging popular protest against expert formulations of policies being imposed on populations without their consent. Wicked problems have also been described as messy problems (Ackoff 1974).

Characteristics

Different authors list different characteristics of wicked problems, but the main ones can be summarized as follows. First, wicked problems are often symptomatic of deeper problems and often display circularity, as in explaining educational problems by poverty, poverty by social class, and social class by educational achievement. Second, wicked problems offer little room for trial-and-error learning: once a neighborhood has been demolished, it can never be restored. Third, they don’t present clearly defined alternative solutions; indeed available solutions are often used to define the problem, as in the idea of monetizing ecosystem services to address “market failure” in environmental protection. Fourth, they are characterized by contradictory certitudes, in that there are many different diagnoses of the problem and many different prescriptions for solving it, often ones that are incompatible; for example, the recent banking crisis has been blamed on both too much and too little regulation. Fifth, wicked problems tend to have redistributive implications for entrenched interests, such as those of the fossil fuel industry in relation to greenhouse gas mitigation.

Responses

Nancy Roberts (2000) proposes three strategies for “taming” wicked problems: (i) simplify the issues and apply routines; (ii) use expertise to control resources; (iii) open the problem up to stakeholders. Each strategy offers an ideal solution from the respective standpoints of hierarchical, competitive, and egalitarian forms of social organization. However, they represent contradictory strategies, so any chosen strategy can only succeed by overriding the interests and values of those advocating either of the others.

In contrast, cultural theorists in the tradition of Mary Douglas advocate “clumsy solutions”
or hybrid strategies in which all of the voices – hierarchical, competitive, and egalitarian – are heard and responded to (e.g., Verweij and Thompson 2011). As such, clumsy solutions are likely to emerge from negotiation or conflict rather than be planned from the outset.

In the final analysis, wicked problems have no definitive solutions. They can only be managed more or less well through “settlements” (Peck and 6 2006) that endure for a while before the problem reasserts itself in a new form that requires renegotiation.

**SEE ALSO:** Urban redevelopment; Urban renewal

**References**


**Women explorers**

Sarah L. Evans  
*University of the West of England, UK*

A lady explorer? A traveller in skirts?  
The notion’s just a trifle seraphic  
Let them stay and mind the babies, or hem our ragged shirts  
But they mustn’t, can’t and shan’t be geographic.  
*(Punch 1893, quoted in Maddrell 2009, 32)*

These lines, which appeared in the popular journal *Punch* in June 1893, satirized a debate then central to contemporary UK geography. In 1892, after years of debate, the Council of the Royal Geographical Society (RGS) decided for the first time to admit women as fellows, electing a total of 22 women to their fellowship between 1892 and 1893, including the traveler and explorer Isabella Bird. This brought the RGS into line with the policies of many of the United Kingdom’s regional geographical societies, including the Royal Scottish Geographical Society, and with other national geographical societies such as the American Geographical Society and the Paris and Vienna geographical societies (see Maddrell 2009).

The decision to admit women to the RGS prompted a backlash among its more conservative fellows, however, whose subsequent campaign resulted in fellowship being once again closed to women in 1893, a situation which lasted until 1913 (Bell and McEwan 1996). Despite British women’s involvement in the United Kingdom’s regional geographical societies, and in geographical teaching associations, their exclusion from the leading UK geographical society restricted the opportunities available to them for exploration and geographical fieldwork. Similarly, although the Association of American Geographers (AAG) admitted women from its foundation in 1904, they were few in number, with many female applicants rejected and excluded in accordance with the contemporary gendered standards for geographical practice. As the satirical ditty cited earlier states, at the heart of this issue was the notion of a “lady explorer,” and the question of whether a woman could – or should – be “geographic.”

Exploration, as a particular form of geographical knowledge production, has played a key role in the history of geographical thought and practice, shaping the development of geography as an academic discipline and leaving its marks on the modern world. In this context it refers to European exploration of other parts of the world from the fifteenth century onward, a key and central part of the geographical tradition. Exploration therefore encompasses a diverse set of concepts and practices, which nonetheless share some common themes, as have been identified by numerous historical, feminist, and postcolonial geographers in recent years. These could include a sense of novelty, of being somehow a pioneer and the first to reach the area being explored or to make a particular discovery; a sense of risk, danger, and endurance; and a sense of making a contribution to knowledge or of conducting scientific research, broadly defined. These elements of exploration were often closely linked with European imperial dominance and colonization of the places being explored. As the work of feminist researchers has shown, these tropes and practices have also often been strongly

---

*The International Encyclopedia of Geography.*  
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.  
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.  
DOI: 10.1002/9781118786352.wbieg1019
gendered as male: the element of discovery was often expressed in terms of the male domination of landscapes cast and depicted as female; the element of risk and danger was tied to the performance of heroic masculinity, while the element of scientific achievement was linked to norms of manly objectivity (Rose 1993). As a result, women have had a complex and at times uneasy relationship with the exploration tradition.

Although exploration is only one particular, and contested, aspect of geographical thought and practice, it continues to capture the contemporary popular imagination. The figure of the explorer continues to dominate popular understandings of geography past and present, as well as to trouble the contemporary academic discipline with imperial and colonial legacies. Who counts – who is allowed to count – as an explorer is a question that is still central to many of the debates on this topic, and particularly those around gender and exploration. In the popular imagination, the figure of the explorer is still often gendered male; one has only to think of the many books, films, and anniversaries of iconic expeditions led by heroic figures such as David Livingstone, Ernest Shackleton, and Edmund Hillary, who together make up a veritable – and decidedly male – pantheon of exploration.

Although the popular figure of the explorer is gendered male, there has also been a long-running tradition of interest in apparently “plucky” or intrepid female travelers and explorers, beginning with popular press interest in their exploits and continuing to the present day. Approaches to these women have varied considerably over time. Earlier work tended to position them as oddities, or somehow eccentric, while early feminist research in the 1970s and 1980s sought to recover these perceived lost heroines and pioneers, an example of a somewhat naive feminist empiricist approach that took a largely uncritical stance on these women and their complexities, as well as leaving existing understandings of exploration and geography unchallenged. More recent work in feminist historical geography and related fields has sought instead both to situate these women in their own contexts, and to interrogate the underlying assumptions that govern and shape exploration as a form of geographical knowledge production. It also seeks to avoid the construction of a female canon of exploration, and to go beyond simply adding women explorers to the history of geographical thought and practice. This approach draws on a number of literatures, including the feminist epistemology of situated knowledges and positionality, and the development of a contextual approach within the history of geographical thought and practice, which is now very well established.

This research has revealed women’s varied and diverse participation in the exploratory tradition, and the ways in which they engaged with diverse forms of exploratory thought and practice, from exploratory travel of the kind undertaken by the British women Gertrude Bell, Isabella Bird, and Freya Stark in the late nineteenth and early twentieth centuries, to the more scientific, systematic work undertaken by British women such as Gertrude Caton-Thompson, Evelyn Cheesman, and Stella Worthington in the early to mid-twentieth century. As shown by these examples, much of the existing research focuses on white Western women explorers, and particularly British and American women, and there is considerable scope for expanding this research further to include women from other countries and traditions.

While there are distinctions between the explorer and other, closely related categories such as traveler, geographer, field scientist, climber, and aviator, the boundaries between these terms are blurred and expansive – not clear borders but frontier zones. There are also several examples of women engaging in all these forms
of exploration-related geographical practice, beginning in the early eighteenth century with women such as Lady Mary Wortley Montagu. Posted to Turkey with her ambassador husband, Montagu wrote letters home that provide a fascinating account of her travels in the region. In anthologies of Western female travel writers, Montagu is often positioned as launching the genre of women’s travel writing, and as a pioneer of women’s participation in exploration. As well as holding this pioneering status, some of her experiences would find echoes in the experiences of other later female explorers, allowing us to explore some key themes within the broader topics of female explorers.

The first of these parallels is the fact that, like Montagu, many of the Western women involved in exploration were from upper and upper middle-class backgrounds until well into the twentieth century. They had sufficient available resources and the leisure time to undertake extensive travel, as well as to participate in the social networks that enabled and supported their exploratory travel. By the late nineteenth century, many of these women explorers had also participated in higher education, enabled by their elite status. Examples include Gertrude Bell and Katherine Routledge, both educated at the University of Oxford in the late 1880s (although, in line with gendered restrictions then in operation, both were not permitted to graduate), and Freya Stark, educated at Bedford College in the 1920s.

Gender shaped the experiences of many of these women. Montagu’s gender played a key role in enabling the particularly novel elements of her account, by allowing her access to the women’s quarters or harem in places that she visited, spaces from which male European travelers and explorers were barred. Gertrude Bell and Freya Stark reported similar experiences during their exploratory travels in the Middle East in the early twentieth century. In many other cases, their gender served to limit and constrain women’s ability to participate in exploration or to be recognized as explorers producing geographical knowledge. Women were barred from many professional opportunities, such as fellowship of the RGS, or full access to higher education, until the end of the nineteenth century.

Even when such professional barriers fell, gendered roles and expectations continued to limit women’s access to opportunities. This limitation can be seen, interestingly, in women’s engagement with tropes around risk, danger, and exploration, particularly in their subsequent publications about their travels, which have given rise to the popular image of the “plucky” lady explorer. For example, in her books about her 1890s expeditions in West Africa, Mary Kingsley both describes the thrilling dangers she encounters, and downplays her own courage in facing them. A similar strategy is deployed by Evelyn Cheesman in her books about her 1930s expeditions in the southwest Pacific, positioning herself as simultaneously brave yet appropriately ladylike. These women had to tread carefully between adhering to gendered expectations of appropriate conduct for a woman while also meeting reader expectations of excitement, danger, and adventure, to demonstrate their status as explorers and not merely travelers. Other examples include the British desert explorer Rosita Forbes, and the French Alexandra David-Neel, who traveled through Tibet in the 1920s when it was still closed to foreigners.

Another important aspect of the positionality of many of these women was that they were not only women explorers, but white women explorers, who need to be considered within their appropriate colonial or postcolonial contexts. Montagu’s text expresses many Orientalist tropes (although, in depicting the
active agency of the Eastern women that she meets, she also subverts them), as does the work of a nineteenth-century travel writer, Harriet Martineau. Many women explorers were active participants in colonizing projects, acting to achieve colonial, imperial, and nationalist aims. For example, a number of women, including Stella Worthington, Kate Ricardo, and Janet Owen participated in a series of British expeditions in the 1920s and 1930s to East Africa, which sought to survey the Great Lakes geographically and ecologically, with a view to using their fish stocks to feed a growing colonial workforce. These women gained status from their raced position as white women, where their gender might otherwise have constrained them.

Another parallel between Montagu’s experiences and those of other women explorers is the fact that she was traveling alongside her husband. Although interest has often focused on the women who traveled without (European) male companions – such as Mary Kingsley, Freya Stark, and Gertrude Bell – many others traveled with male partners, relatives, and colleagues – such as Florence Baker, Katherine Routledge; and Louise Arner Boyd, an American polar explorer who organized and funded a number of expeditions to the Arctic in the 1930s. Whether female explorers were accompanied by male Europeans or not, exploration and expeditionary work were fundamentally collaborative in nature, drawing on the expertise and labor of local people in addition to the European expeditionary team. Boyd’s work suggests another important dimension of women’s involvement in exploration, that of helping to fund and support such endeavors, without directly participating in the work “in the field.” A good example is that of Lady Jane Franklin, the first woman to receive the Royal Geographical Society’s Gold Medal, in 1860, in recognition of her extensive efforts in organizing and funding a series of expeditions in search of her husband, Sir John Franklin, who disappeared alongside the rest of his crew and expeditionary team while in search of the Northwest Passage in 1845.

**SEE ALSO:** Colonialism, decolonization, and neocolonialism; Exploration; Feminist geography; Gender; Intersectionality; Positionality

### References


### Further reading


Women in geography

Linda J. Peake
York University, Canada

Although women’s presence in the discipline of geography and engagement in geographical research may appear to be firmly entrenched, they are unevenly and often precariously positioned within the discipline. In the Anglo-American world there are substantive differences between women along the lines of geographical location, language, subfield within the discipline, age, racialization, and level of job security. Even for those most privileged – white, middle-class, and able-bodied women in the Anglo-American world in tenure and permanent positions – there are often marked disparities vis-à-vis their relation to male colleagues. Studies have shown that both subtle and structural barriers (many of which emulate those in other disciplines) to equality exist for women, especially women marked by race and disability; sexism, racism, able-bodiedism, and a general “discomfort with difference” (Falconer Al-Hindi 2000, 699) still permeate the discipline.

Notwithstanding this gloomy scenario, women have a long and somewhat hidden historical presence in geography. Although geography as a field of study has a history going back to the time of the ancient Greeks, women did not appear in it as individuals until the nineteenth century. And prominent women geographers who existed outside of the formal institutions and societies of geography that started to emerge in the nineteenth century have a fragile existence in accounts of the emergence of geography as a discipline. In the preprofessional phase of the discipline, prior to the emergence of geographical associations, women’s engagement in geography was primarily undertaken by women travelers and explorers, most of whom were European and North American (Maddrell 2009). Their involvement in geographical practices of travel and exploration, however, did not lead to an acceptance of their ability to produce legitimate geographical knowledge, and they were commonly denied entrance into professional geographical associations. The Royal Geographical Society (RGS) finally admitted women as fellows in 1892; in the following year it admitted a further 22 women before closing its membership to women until 1913 (see Women explorers). Other national geographical societies in Europe and the American Geographical Society in the United States also admitted women in the late nineteenth century. Perhaps the best-known female academic geographer of this period was Ellen Churchill Semple, a founding member of the Association of American Geographers (AAG) and its president in 1921. She studied anthropogeography with Ratzel and was involved in the environmental determinism debates in the late nineteenth and early twentieth centuries. Well into the twentieth century, gendered roles and expectations limited women’s ability to contribute to “professional” geographical knowledge production. The AAG had women members from its inauguration in 1904, albeit very few until its merger in 1948 with the short-lived American Society for Professional Geographers (ASPG) (Miller 1950) and the disbanding of the AAG’s “elitist and exclusionary membership policies [which] until the late 1940s required a record of original research,
nomination by two sponsors, and affirmation by
90 percent of voting members” (Monk 2004, 5).
The work of the feminist geographer Jan Monk
(2004) has revealed, however, that AAG policies
did not prevent women’s entry into the discipline
in the United States, and that there have been
hundreds of women geographers since the late
nineteenth century, though most have heard
of few of them beyond the aforementioned
Ellen Churchill Semple. Many of these women
worked as professionals – as teachers, librarians,
administrators in government departments, and
editors – although employment in universities
was far less easy to break into. As Monk (2004, 1)
has pointed out in relation to geography in the
United States: “Histories of American geogra-
phy have tended to concentrate on geographic
thought and on the men who have been seen
as major figures in research.” By choosing to
focus instead on the professional practices of
American geography in the nineteenth and
twentieth centuries she has unearthed a wealth
of long-forgotten women geographers, some
of whom have been trailblazers of radical social
change, including, among others, Thelma Glass
(1916–2012), the African American civil rights
activist who was active in the Women’s Politici-
Cal Council in Montgomery, Alabama, which
sparked the 1955 Montgomery bus boycott and
the significant ruling a year later that segregated
buses were unconstitutional (Monk and George
2004). Yet Thelma Glass, who helped initiate the
geography program at Alabama State University,
where she taught for over 30 years until her death
in 2012, was unknown within the discipline of
geography except to a relatively small number of
geographers.

Women started to trickle into the (overwhelm-
ingly white and male-dominated) discipline of
geography at the university level after the 1960s
and 1970s. In Europe the 1968 student revolts
marked a demand for left-wing reform. In
North America, engaging in social and political
movements advocating for black and gay civil
rights and an end to imperialism in the Global
South, including the Vietnam War, but also in
the Peace Corps and second wave feminism,
marked a changing context for women (Monk
2004). There are now over four decades of data,
from both quantitative and qualitative studies,
including personal accounts, documenting the
position of women and gender imbalances in the
discipline. The most comprehensive overviews
come from surveys from the United States, the
United Kingdom, Canada, and more recently
Japan, South Korea, Germany, Norway, Israel,
and Australia (Johnson 2009).

From the early 1970s there has been an interest
in recording and questioning the underrepre-
sentation of women in geography (particularly
at the level of the professoriat) (Zelinsky 1973;
McDowell 1979; Monk and George 2004). In
1973, for example, women accounted for
only 3.4% of the faculty in US and Canadian
graduate geography departments, with only one
female full professor, which Zelinsky (1973)
attributed to wider patterns of sexism as well
as particular institutional rules, traditions, and
biases. In the United Kingdom in 1978 a survey
of its departments of geography (with an 83%
response rate) found that women comprised
only 7% of all teaching staff, only three women
were full professors, and 43% of departments had
no women staff at all (McDowell 1979). This sit-
uation was to improve in the late 1970s and early
1980s, which were a period of consolidation for
white women in Anglo-American geography;
however, the later 1980s and 1990s cannot be
characterized as a glowing affirmation of the
relentless rise of women in geography. Common
barriers of gender-based discrimination were
(and still are) encountered during the academic
job search and in the hiring and the promo-
tion and tenure processes, as evidenced by the
increasingly smaller numbers of women the
further one proceeds up institutional hierarchies
(Falconer Al-Hindi 2000).

In the late 1980s the institutional picture
remained virtually as dismal as in the preceding
two decades in many respects. It was proving
very difficult to diversify the academic profession
in North America. Although there had been an
increase in the number of women gaining faculty
positions, they were still significantly underrep-
comprised only 10.6% of faculty, and there was
still only one female full professor (Mackenzie
1989). The numbers of women in permanent
academic ranks in US geography departments
still constituted only 8% of geography faculty
members. Women comprised a quarter of the
new doctorates and low-ranking professors but
only 10% of associate and 3% of full professors,
in both large and small departments (Lee 1990).
A 1987 survey of geography departments in
North America also found that just over 5%
of academic geographers were people of color
(African Americans, Hispanics, Native Ameri-
cans, and Asians) (Shrestha and Davis 1988), and
a more recent survey of black geographers in
the United States put the number of female and
male practitioners at just over 60 (Darden and
Terra 2003).

Although white women in the Anglo-
American heartland of geography have made
increasing forays into the discipline in significant
numbers, women of color still find geography
institutionally racist and not only a space that has
proven difficult to enter but also one that some
of those who found a way in have subsequently
decided to leave. A survey of AAG members in
2004 found that

The shares of black and Hispanic members in the
AAG have risen since 1980, but are still minis-
cule, at 1.2 and 1.3 percent respectively in 2001.
The number of members identifying as Asian,
however, has grown rapidly over this period: in
2002, it accounted for approximately 5.5 percent
of the total membership. (Pandit 2004, 15)

Over a decade later, there is nothing to indicate
that these figures are significantly differ-
ent. Not only are women of color missing
from Anglo-American geography, but women
geographers of color outside of Anglo-American
geography are also being marginalized. While
multilingual, including English-speaking, women
of color in previously colonized countries – India,
South Africa, Singapore – are accorded some
visibility in geography, those in “Europe, Asia,
South and Central America, and Africa, are truly
on the edge of feminist as well as geographical
discursive spaces – a power geometry of knowl-
edge circulation that necessarily renders invisible
much work” (Johnson 2009, 54). These issues
concerning the composition of women in the
discipline are indicative of the consideration of
women’s representation being replaced by those
of the absence of women of color (Kobayashi
2006; Pulido 2002) and the hegemony of the
English language (Garcia Ramon, Simonsen, and
Vaiou 2006).

This is not to dismiss the invaluable contribu-
tions to the discipline that a number of individual
key white women geographers have made, pro-
viding intellectual leadership through teaching
and publishing, ensuring the appointment and
promotion of women, mentoring, networking,
supervision, and promoting the production of
(feminist) geographical knowledge, not least
through feminist journals: since 1994, Gender,
Place & Culture: A Journal of Feminist Geography,
and, since 2010, the Revista Latino-Americana
de Geografia e Gênero (Latin American Review
of Geography and Gender). Johnson (2009)
points out that the key women who entered
geography affected by the tumultuous times of
the 1960s, the 1970s, and early 1980s (before the
cultural turn), and who have remained figures
of influence, are now retiring or entering senior management positions. Their successors, she observes, “are coming of academic age in a very different political and theoretical era, one where gender is no longer fundamental … What is now emerging is therefore a very different feminist geography, and … such changes have seen the evacuation of the category by many” (Johnson 2009, 54). These changes are echoed in the organizations that women in geography occupy.

Women’s organizations in geography

The institutional framing that would allow women to address professional issues and for feminist approaches to prosper was first established at the national level in the early 1970s, notwithstanding that women constituted only a very small percentage of faculty members in that period. These bodies can be loosely differentiated as serving two functions, with many serving both. The first group consists of those that represent the professional interests of all women in the discipline and that typically engage in issues of affirmative action/equal opportunities to increase the representation of women and promote data collection to aid such efforts (some of it reported earlier in the entry). Second, there are organizations that promote research undertaken by women/feminist geographers. They either embrace all women geographers – including physical geographers – regardless of whether they engage in feminist research and are also open to male members, or tend to be run by feminist scholars and attract only women and men engaged in feminist research. It should be stressed, however, that, while a few sympathetic male geographers may join these groups, their numbers are extremely small, just as it is often still rare to see male geographers at conference sessions organized by feminist geographers. Active groups representing these two functions can be found in the United States, Canada, the United Kingdom, Japan, and a range of Western European countries, including the Netherlands, Spain, Germany, Austria, and Switzerland. Women/feminist geographers in Australia, New Zealand, India, Israel, Southeast Asia, Eastern Europe, Scandinavia, Africa, and South and Central America, among other places, tend to work together outside of any official women’s organizations (Morin 2009).

Nationally, the first body to be inaugurated was the Committee on the Status of Women in Geography (CSWG) of the AAG in 1971, although this happened more by accident than by design (Monk, personal communication; Monk 2004). In 1979 the AAG Specialty Group on research on women and gender, Geographic Perspectives on Women (GPOW), was launched, and the AAG also adopted a bylaw on affirmative action. A similar thing happened in Canada a few years later, not least because a number of Canadian feminist scholars who were to play important roles were outside Canada (mostly in the United Kingdom and United States) pursuing PhDs and did not return until the early 1980s, when the Canadian Women and Geography Study Group (CWAG) of the Canadian Association of Geographers (CAG) was formed in 1982. In the United Kingdom, the (then named) Women and Geography Study Group (WGSG) of the Institute of British Geographers (IBG) was also formed in 1982, preceded by the formation in 1980 of the Women and Geography Working Party (Bowlby and Tivers 2009). Strong transnational linkages existed between a number of feminist geographers in Canada, the United States, and the United Kingdom. Indeed, the establishment of the WGSG was predominantly due to the presence in the United Kingdom of a Canadian feminist
geographer, Suzanne Mackenzie. Suzanne was a charismatic presence whose warmth, wit, and expansive nature influenced a whole generation of women geographers to embrace feminism. All three of these organizations – GPOW, CWAG, and the WGSG, now renamed the British Gender and Feminist Geographies Research Group (GFGRG) – remain very active, publishing geographical research from feminist perspectives, bestowing awards, funding named lectures, maintaining a feminist geography bibliography, and facilitating the exchange of information and ideas through websites, newsletters, conference sessions, workshops, reading groups, and listservs (geogfem@lsv.uky.edu has a largely North American and international audience and gfgrg@jiscmail.ac.uk has a UK audience). Similar groups exist in a few other countries, such as the German-speaking Specialty Group of Gender and Geography, which mostly represents women in Germany, Switzerland, and Austria (www.ak-geographie-geschlecht.org/), and the small but vibrant Study Group on Gender and Space/Place in the Association of Japanese Geographers. The Australian equivalent of these groups – the Gender and Geography Study Group within the Institute of Australian Geographers – closed down in 2004, a potential sign of the inability of such groups to be self-sustaining in countries where the number of younger feminist geographers in the discipline is insufficient.

A number of loosely affiliated groups further support women in geography in the United States (Morin 2009). The Washington-based Society of Women Geographers (SWG) is a professional and social organization established in 1925 by women who were accomplished explorers. Its membership now also includes academics from the AAG and the American Geographical Society (AGS), and it bestows a number of awards on women who have made distinguished contributions to knowledge about the world (Monk 2004). The National Council for Geographic Education (NCGE) has a women’s special interest group that provides awards for research in geography education by women graduate students. Other professional networks to which women geographers belong include the Association for Women Geoscientists (www.awg.org), the Earth Science Women’s Network (http://eswnonline.org), and the GIS Lounge–Women in GIS (www.gislounge.com/women-in-gis).

Subnational organizations are not a common format adopted by women geographers, but one such well-established body is the Association of Pacific Coast Geographers Women’s Network (http://apcgweb.org/womens-network), which offers scholarships and mentoring opportunities to its members.

The Commission on Gender and Geography of the International Geographical Union (IGU) (http://igugender.socsci.uva.nl) is the only women’s geography group whose mission and purpose is international in scope. The IGU was formally initiated in 1922 (although it has met since 1871; Morin 2009), and the commission was established with study group status in 1988, then granted standing as a commission in 1992. It is now one of the most active groups in the IGU; in 2014 it was recognized by the IGU as its “best” commission (Monk, personal communication). The commission was the initiative of Jan Monk, an Australian feminist geographer who conducted her career in the United States and has been a stalwart supporter of feminist geography and mentor to many feminist geographers globally. The geographical unevenness of organizations for women within institutions of geography has partially been offset by the work of the IGU, which has been of particular importance for women geographers in countries where there is no strong heritage of feminist work, such as Taiwan, for example, where feminist geographers are present but few
WOMEN IN GEOGRAPHY

in number, and Africa, where no organizations for women in geography exist and where the commission has held meetings in the anglophone countries of Ghana and South Africa. The IGU Commission has also been important in Southeast Asia, where the regional geography association (SEAGA) does not support a distinct geography group related to women, although in East Asia the Association of Japanese Geographers Study Group on Gender and Space/Place has strong links to the IGU Commission. The IGU is also increasing its presence in Latin America, where women’s groups in geography also do not exist, although there are vibrant networks of women geographers (see Peake (1989) and Garcia Ramon and Monk (2007) for global overviews of women in geography).

In its role of bringing together feminist geographers from many countries, the IGU Commission now has 670 members from over 50 countries. While it has done much to counteract the Anglo-American hegemony of the field, the majority of its members are from countries already well appointed with women’s organizations – the United States, United Kingdom, Canada, and Spain – and women from India are also well represented (although there is no women’s group in the National Association of Geographers of India). The context where institutional groups supporting women in geography are lacking, such as in India and other countries in the Global South, is one where geography as a discipline has not always been prominent in university curricula, is still largely descriptive in focus, and is valued primarily as a subject for training teachers or policymakers (Morin 2009). In such instances there is little political space for women’s organizations or for feminist work in geography to flourish. The countries of Australia and New Zealand; Brazil, Argentina, and Mexico in Latin America; France, Germany, Italy, and Switzerland in Western Europe; the Scandinavian countries of Sweden, Norway, and Finland; Israel in the Middle East; Japan in East Asia; Singapore in Southeast Asia; and South Africa are also moderately well represented in the IGU Commission. Membership is lowest in Eastern Europe (including Russia), East Asia (except Japan), Southeast Asia (other than Singapore), and Africa (except South Africa). The commission organizes several meetings each year and maintains its intellectual network through its website, working papers, and an excellent newsletter (igugender@list.arizona.edu), compiled by Jan Monk and published in English, French, and Spanish.

In addition to national and international bodies, there are also formal and informal cross-national networks of feminist geographers. For example, the Erasmus Network for graduate teaching existed within the European Union for nine years between 1990 and 1998, consisting of individual feminist geographers from six universities in five countries (United Kingdom, Spain, Denmark, the Netherlands, and Greece). Coordinated by Joos Droogleever Fortuijn, an active IGU Commission member, the network organized courses on gender and geography each year. More recently, the Great Lakes Feminist Geography Collective (Fem-geog-network@listserv.uoguelph.ca) has been formed, a group of nearly 30 feminist geographers from universities surrounding the Great Lakes in North America, which in a 2013 meeting brought together for the first time feminist geographers working in Canada and the United States. It has already been extremely productive, serving to energize a number of its members.

There are also, of course, informal academic groups that exist outside of the formal institutions that represent geography. Perhaps the best known are the Supporting Women in Geography (SWIG) groups in the United States
(and present but less prominent in Canada), which have been set up in various departments of geography, and at which professors, postdocs, and graduate students meet informally to engage in various activities that support the intellectual, professional, and personal participation of women in the discipline and provide mentoring opportunities for women who are pursuing academic and professional careers in geography (there is a list of SWIG groups at http://www.aag.org/diversity/gender). SWIGs have now spawned parallel professional groups built on the same ideas and philosophies. In 2007, Supporting Women in Geography Geographical Information Systems (SWIGGIS) (http://swiggis-austin.org) was formed in Austin, Texas. This is the first group of its kind in the professional realm. Its members are trained women geographers who work in the field of geographical information systems (GIS) in private businesses, public institutions, and local and state agencies.

Finally, informal professional and friend-based feminist networks operate globally. These networks do much to explain the development of particular research foci and, with technological developments, do not always require face-to-face meetings to flourish. These informal networks may also represent a more common future mode of organization that is no longer underpinned by the notion, which arose out of second wave feminism, of women as a unified group intent on a (reformist) politics of inclusion via separate organizations for women (Johnson 2009).

Whither forward?

Although there has been a more or less constant expansion of formal and informal organizations run by women geographers since the 1970s, they create geographically uneven networks – they are particularly thick on the ground in Western Europe and North America – often serving to better link in those who are already well linked. They have been successful in highlighting the low numbers of women in the discipline, in aiding the development of feminist work, in contributing to ways in which geographical knowledge is produced and disseminated, and in some cases in highlighting the Anglo-American hegemony of geography. What the range of women’s organizations and networks reflects, moreover, is the diversity of politics, practices, types of knowledge production, and modes of organizing in various countries (which may or may not speak to the status of women in the discipline in those countries). It is hardly surprising then that, despite the encompassing work of the IGU Commission, political divides between feminist geographers, across nationality and race in particular, have prevented the development of any robust transnational feminist praxis via these organizations. Indeed, whether existing groups can maintain their presence in some countries is in doubt. Women of color in white-dominant countries are less inclined to join what are often, in effect, white women’s organizations. And younger women, both white and of color, are often more attracted to more encompassing groups that work with different others across issues or to different political and theoretical projects of poststructuralism, post-colonialism, antiracism, and queer organizing in which gender is no longer the fundamental organizing category. The latter point is currently evidenced in the name changes being considered by some of these organizations from the ubiquitous but singular and knowable category of “woman” to encompass gender, with a nod to being more inclusive of trans women and of men.

While some may see the demise of women’s organizations as detrimental to women’s progression in the discipline, others see these
organizations as only serving to benefit white women and as speaking to the past, and therefore do not regret their dissolution. Peake has argued that the influx of women and second wave feminism into geography have left its racialized composition virtually undisturbed:

Notwithstanding feminist geography’s call to difference and inclusion, our failure to leave our own history uninvestigated subjugates the role that women of colour have played in it. It also speaks to a complacency in the failure to tackle the current racialized composition of our practitioners’ body as well as to investigate our interests in avoiding the truth about such an injustice, for it verges on institutional apartheid. Affirmative action is needed to address the whiteness of our field[;] without it we will only be able to reproduce ourselves in our own image and our role as an epistemic community will be far less rich and credible. (Peake 2015, 265)

Pulido (2002), among others, has also argued that a lack of diversity has deleterious effects on the breadth of geography’s intellectual production. Structural change along the lines of affirmative action is needed, including, for example, interventions into job descriptions, strategies of recruitment, and mentoring of women of color. Pulido (2002) states, however, that in addition there needs to be a fundamental change in the disciplinary culture to make geography more open and welcoming to all groups. Achieving this culture requires women in geography to remain engaged in viewing the composition of their discipline as an inherently political project that speaks to broader challenges to inequality.

SEE ALSO: Antiracist geography; Feminist geography; Gender; Gender, work, and employment; Race and racism; Race, work, and employment; Women explorers

References


Miller, E. Willard. 1950. “A Short History of the American Society for Professional Geographers.”


Work–life balance

Diane Perrons
London School of Economics and Political Science, UK

Work–life balance, as a problem, entered the academic and policy arenas as the scale of women’s participation in the paid economy increased. While gratuitous attention was given to gender justice, reflecting long-standing feminist campaigns, policy discussion centered on how resolving the tension between paid work and caring would increase firm competitiveness and economic growth. The Organisation for Economic Co-operation and Development (2002, 5), whose mission is to “achieve the highest level of sustainable growth,” argued that “family friendly policies are a goal in themselves” but “getting the right policy in place will promote other societal goals … they will allow aggregate labour supply and employment to be increased.” Likewise the World Bank promoted gender equality as “smart economics” and recognized the importance of childcare to support women’s increased involvement in the economy.

Neither of these organizations contemplated a lateral redivision of domestic labor between women and men in line with Nancy Fraser’s (1997) ideal Universal Carer Model. More explicitly, Gøsta Esping-Andersen (2009, 169) considered “sufficient change in male behavior to substitute effectively for the decline in female domestic work unlikely.” More progressively, he argued that state support for carework would increase social efficiency by enabling low- as well as high-income women to enter the workforce and prevent gender equality in employment from being a middle-class privilege.

Apart from some of the Nordic countries, however, state childcare provision remains low in the Global North and Global South, leaving households to find their own resolution to work–life balance, which they do in class-differentiated ways. These resolutions sustain social inequalities within countries as high-income households can afford private care to facilitate the dual-earner model while lower-income households opt for a one-and-a-half-earner model, to the detriment of women’s careers and long-term incomes.

As caring and domestic work are highly labor-intensive and considered a natural talent of women, they are characterized by low pay and as such increasingly depend on a socially disadvantaged workforce, ethnic minorities, and migrant labor, a practice which sustains inequalities between people and places. In the Global North and Global South, maids and nannies from lower-income households, regions, and countries provide an important source of labor, as discussed in the global care chain literature. Such migration represents an individual solution to structural inequalities and exacts an emotional cost as migrants provide care while jeopardizing their ability to care for their own children.

While seemingly a small matter of decision-making within households, work–life balance raises a whole range of issues relating to economic and social relations between people and places and with respect to deep-seated gendered and racialized norms regarding the value of and suitability for different forms of work. Securing work–life balance requires resolving tensions between...
competing work-life scenarios of the free competitive disembodied worker presumed within economic policy and required by firms and the embodied worker for whom permissive work–life policies, parental leaves, and working time regulations are advocated in social and equality policies. Research in the information technology sector (James 2014) shows that resolving these tensions can be a win-win scenario for firms and workers alike, though realizing these potential gains more broadly remains elusive, leaving households to revolve their own problems with the associated gender, social, and geographical inequalities that this implies.

SEE ALSO: Care work; Domestic workers; Gender, work, and employment

References


Further reading


World cities

Kathy Pain
University of Reading, UK

World cities are generally understood to be leading global centers of economic and political power. Widely seen as internationally competitive urban concentrations of business, with specialized expertise, knowledge, and finance, they are undoubtedly a major manifestation of global resources and capital.

The world cities phenomenon has captured the imagination of scholars and policymakers, especially after Peter Hall’s book *The World Cities*, published in 1966, brought this powerful concept to widespread international attention. Fifty years earlier, Patrick Geddes’s *Cities in Evolution* (1915) had depicted world cities as a manifestation of the competitive international economy and, prophetically, as the generators of competitive regions. Roderick McKenzie went on to describe them in 1927 as world “centers of gravity” within a new form of global urban organization characterized by relations of dominance and subordination. It is this foundational conceptual framing of world cities as both an outcome and a generator of competitive spatial relations that has sustained discourse about this evolving expression of modern global capitalism for almost a century. Unsurprisingly, contemporary global economic integration has acted as a spur to intense world city competition, yet, at the same time, the ethics of the spatial centralities and power geometries which manifest the global system of world cities remains subject to contestation and vigorous debate.

In the globalizing world economy that has been emerging since the 1970s, the processes by which economically competitive world cities and their regions are generated and sustained have become increasingly complex. Facilitated by advances in information and communications technologies (ICT), the pace of change has accelerated to a once unimaginable speed. Fast forwarding to the twenty-first century, in comparison with the world metropolises described in the early world cities literature, contemporary cities are interconnected by a veritable blizzard of digitized information and financial flows. Developments in both communications and transportation have given specialized financial and linked business and professional services – first referred to as prominent in world cities by Hall and then, in 1986, by John Friedmann – an even more critical role in determining the way the present-day global system of cities functions. However, the specific link between the global dispersion and concentration dynamics of “new economy” finance, insurance, accountancy, law, management consultancy, and advertising “advanced producer services,” and the world cities phenomenon, was only made in the latter years of the twentieth century. The theory that, as offices servicing global capital extend worldwide, leading world cities gain a new role as strategic sites in economic globalization was not advanced until 1991 when Saskia Sassen’s *The Global City: New York, London, Tokyo* was first published.

Drawing on Manuel Castells’s (1996) theorization of the role played by advanced producer services in the contemporary network society, Peter Taylor and coresearchers in the Globalization and World Cities (GaWC) Research Network have empirically specified the global...
system of cities as a “world city network.” They have revealed a major reshaping of world city power geometries during the first decade of the twenty-first century in the form of a shift from localized and even national political control of cities to new de facto controls exercised by world money and information flows. This chimes with the prophetic vision of McKenzie who in 1927 pointed to coming changes in global urban organization “at variance with the existing political structure” as a consequence of increasing connectivity.

Alongside these global changes, a linked contemporary dynamic has been emerging at a more localized scale. Evidence for the global integration of extensive city regions in the new economy since the turn of the twenty-first century has sparked parallel empirical research into global centralities. This is not only focusing on the highly clustered business quarters of major world cities but also encompasses the proximate urban landscapes of nearby towns and cities. Interlinked by multimodal transportation infrastructures as well as virtual communications, these polycentric global regions such as the London “South East England mega-city region” integrate multiple surrounding urban centers of diverse size (Hall and Pain 2006). Theorists had already noted a similar process in the nineteenth-century “megalopolis” along the northeast coast of the United States and in the so-called “blue banana” in northwestern Europe. However, recent investigations of both of these extended urban regions have identified a further very high level of intraregional connectivity in global advanced producer services networks.

These multinodal global formations are sites for major flows of international skilled labor, knowledge, and finance that increase, as opposed to diminishing, economic returns and have a positive impact on GDP. This new scale of global activity has been described as creating the new engines of the world economy that threaten to overturn the primacy of nation-states. However, as is the case for the global system of world cities, new divisions of labor and polarizations are emerging at a local scale. The new foci for global centralities are generating spatially differentiated effects. These are illustrated by Doreen Massey’s 2007 analysis of the globalized spatial relations and political dilemmas facing contemporary London in World City.

This brief commentary on a topic that has been extensively explored in a large literature has necessarily been highly selective. Precedence has been given to a small selection of the key theoretical and empirical developments that provide an overview of the transformative processes and main lines of argument that have underpinned recent debates.

Origins of the world city concept

For Hall, who first brought the concept of the world cities to international attention, these complexes were “great cities, in which a quite disproportionate part of the world’s most important business is conducted” (1966, 7). A concentration of professional expertise, specialized knowledge, and finance was enabled by superb transportation and communications infrastructures. As centers for advanced business and professional activities, knowledge, and learning, as well as governmental organizations, the power of world cities extended far beyond national borders. Hall identified not only major metropolises such as London, New York, Tokyo, Paris, and Moscow as world cities, but also multinodal urban regions such as the Randstad (Netherlands) and Rhine-Ruhr (Germany). He recognized early on the potential
world competitiveness of extensive polycen-
tric urban regions, an idea elaborated 40
years later by Hall and Pain in The Polycentric
Metropolis (2006).

Drawing on the pioneering regional perspec-
tive developed by Geddes, Hall’s world cities
analysis identified the significance of the coming
shift from early capitalism to “finance capitalism”
for a dramatically changing regional landscape.
In particular, he drew attention to the physi-

cal spreading out of globally competitive city
growth and its consumption patterns, includ-
ing individual travel by car. These city-region
developments, and the challenges they present
for governance and planning, have become even
more apposite worldwide today.

By way of contrast, and drawing on theories of
a new international division of labor (NIDL) in a
world economy where people and finance capital
flow into and out of the city, in his World City
Hypothesis, Friedmann (1986) focused on the
global economic system as a whole and the city
distinctions it gives rise to. In prioritizing world
economic development as the key explanatory
factor in urban analysis, he emphasized the
importance of rapidly expanding financial and
“high-level” business services in creating “global
financial articulations” such as London, New
York, and Tokyo that were the world “basing
points” for global capitalism. Different localities
were identified as having varying global and
subglobal roles in the “spatial organisation and
articulation of production and markets,” result-
ing in a complex spatial “hierarchy of world
cities.” Certain “primary” centers for major
capital accumulation in “core” countries, such
as New York, were distinguished as carrying out
“all of these functions simultaneously,” whereas
other cities had multinational or national func-
tions and roles. Importantly, the cities in this
global hierarchy were, for Friedmann, all part of
a network.

Compared with the world-systems framework
in which relations between core and periph-
ery were conceived as process— as opposed to
place-specific, Friedmann’s world city hierarchy
was unequivocally spatial, with North American,
Asian, Southeast Asian, and Western European
subsystems, the latter including Hall’s exten-
sive Randstad and Rhine-Ruhr world cities.
Referring to research such as the early study of
New York by Sassen (2001/1991), he also drew
attention to consequent crises inflicted on the
state by the social and fiscal impacts of interna-
tional capital in primary core and semiperipheral
world city locations. The insights provided by
Friedmann have influenced the thinking of many
subsequent writers.

The world city as process versus place

Building on Friedmann’s analysis, the 1991
edition of Sassen’s The Global City was a seminal
contribution to the world cities literature. The
timely attention by Sassen to the process by
which prominent financial centers such as New
York, London, and Tokyo acquire distinctive
roles as agglomerations of command functions in
specialized business services also had a distinctly
spatial focus (Sassen 2001/1991). Her analysis of
the simultaneous dispersion and concentration
dynamics of specialized businesses servicing
multinational corporations worldwide informed
a wave of subsequent empirical studies into
the spatialities of global business organization
and practices, and their societal impacts. The
insights provided into the drivers behind the
centralization of the global command and con-
trol functions of advanced producer services and
the implications for the management of the new
space economy have had an ongoing influence
on urban research to the present day.
But a new way of understanding the production of global city space was introduced by Castells in 1996. In his *Information Age* trilogy, Castells advanced the seminal theory of the emergence of a new “network society.” The city network concept had previously been alluded to by other writers but Castells added a distinctive perspective in the shape of the rising influence of new economy producer services as key generators of increasing network relations between cities. In the context of late twentieth-century developments in ICT, he saw such services as active agents in a global space of flows that was superseding the mosaic space of places appropriated by states. The importance of this theoretical leap is illustrated by John Allen in a less well-known contribution to the world cities debate in 1999, “Cities of Power and Influence.” This analysis went against the grain of previous (and prevalent ongoing) interpretations of city power relations, hypothesizing that, in the network society, power can no longer be understood as “fixed” in space or as belonging to any one city, but must instead be seen as distributed through intercity networks. Allen’s challenge to competing world city narratives that present power as deeply embedded in place has, unfortunately, been somewhat overlooked in subsequent debate.

For example, in *World City*, Massey (2007) observes London predominantly through the geographer’s lens of a “global sense of place,” focusing on the concentration of wealth and resources there and the associated local and global polarizations. Her deliberately accessible writing style leads to an inference that London, the city, is the agent actively shaping globalization and generating its uneven economic and power geometries. Allen’s interpretation of world city power as distributed through networks of economic actors, and flows within and between cities, is less in evidence. Massey’s plea to London’s government leaders to adopt a deterritorialized politics that recognizes the city’s moral responsibilities perhaps overestimates the capacity of the bordered state to reshape power relations that are mobile in intercity networks. This critical question for world city policy will be returned to shortly.

### A world network of cities

Taylor deduced that cities can only be understood through their worldwide network relations. According to network theory, cities are “not a place but a process”; therefore, according to Taylor, writing in 2004, they must be studied by quantifying their interlinkages and connectivity as nodes in a world city network that services global capital. This theoretically inspired empirical departure marked a late twentieth-century turning point in world cities analysis in two ways. First, the analysis recognized cities as shaped by or as the outcome of a social process, as postulated by earlier pioneer urban sociologists, in which city relations and connectivity are constructed by business networks that straddle and use them to conduct global business. Second, empirical network analysis allowed the dynamic distribution of connectivity across the global system of cities to be measured robustly for the first time. Furthermore, the numerous quantitative network analyses initiated by GaWC and others have been supplemented by qualitative studies that shed light on the relations and practices of the advanced producer services that generate city network connectivity and the consequent functional interlinkages and flows between cities.

What can be learned from the results of this research is that the power of contemporary world cities cannot be assessed simply on the basis of place-based city attributes but must take into account intercity relations in dynamic global
networks. This new way of thinking about city relations is acknowledged by Sassen in her 2001 edition of *The Global City* and in subsequent publications which credit the partial deterriorialization of spaces of centrality in digital networks. But Sassen’s view that spatiality that “pivots on” global networks remains interdependent with a major territorial concentration of resources illustrates an important spatial contradiction that has underpinned ongoing controversy about the ethics of world cities.

To put the city network thesis into context, Castells too is acutely aware of the differential spatial impacts that globalization encompasses. Albeit modern society is increasingly networked, as in previous eras, it remains an expression of deeply territorialized political relations with unequal impacts. Castells’s dystopic world vision presented in the final 1998 volume of his *Information Age* trilogy engages with the deeply divisive inequalities prevailing within, as well as between cities worldwide. Spaces of fragmentation and political contestation have not disappeared, even in the world’s richest cities. Therefore a major challenge, it would seem, is the need for the politics of cities to engage with the networked space of contemporary intercity relations and flows, which are now more volatile than ever before.

**Cities of towers**

In the complex networked world economy, the concentration of value-adding functions has been ongoing and has become intertwined with new modes of “grounding” finance capital in the physical structure of cities.

In his 2009 book *Towers of Capital*, Colin Lizieri demonstrates how, in major world international financial centers where real estate interacts with other globalized economic activities, advanced producer services and office property markets are critically interlinked in the production of the evolving world city network. World cities are the sites for increasing major international financial investment, encouraged by the financial innovations introduced by their specialist transnational labor that have modified the relationship between office property development and its commercial funding. As the creation of new investment vehicles, such as private real estate funds or real estate investment trusts (REITs), has spread the risk on high-value property developments, the financialization of city real estate has accelerated. In this way, the capital that flows through financial and business services networks is stored in the physical city infrastructure such as the office buildings that they occupy.

Meanwhile, high-revenue streams associated with city “signature” or “starchitect”-designed landmark office space, literally build the evolving vernacular of competitive global capitalism. The integration of global finance and city real estate in world cities therefore facilitates the circulation of mobile international capital and also fixes it in specific city locations. Put simply, the workspace of the contemporary service economy is molding the politics of the world’s cityscapes.

Lizieri’s account is particularly pertinent to recent debate about the exposure of world cities to global financial risk. Counterintuitively for some commentators, London’s office market has proved resilient to the post-2008 global financial crisis. But a stream of qualitative world cities research suggests that this actually resonates with Lewis Mumford’s view that “the nature of the city is not to be found simply in its economic base: the city is primarily a social emergent” (1938, 6). Mumford complained that “people treat pragmatic abstractions such as money, credit, political sovereignty, as if they were concrete realities that had an existence independent of human conventions” (1938, 7).
WORLD CITIES

As already noted by Hall in the 1960s, world cities are used by a “significant proportion of the richest members of the community” (1966, 8). They are the places where highly paid labor with specialized skills congregate, tacit knowledge is exchanged, relationships are developed, global deals are brokered, and foreign direct investment is concentrated. London’s transnational constitution generates global work. A major high-value financial transaction taking place in the City of London can involve participants in, say, New York, Tokyo, or Shanghai. And it is the globalizing strategies of London’s transnational capitalist city users that have, for more than a decade, helped to sustain its supreme world city network connectivity, also creating investment confidence in its real estate markets. Together with its time zone and other place-related advantages, a politics of openness to foreign labor and firms since the 1980s, has undoubtedly been instrumental in sustaining London’s high-value office real estate market.

In the context of increasingly agile transnational capital investments, neoliberal strategies to attract inward investment through entrepreneurial development projects have also more directly promoted the commodification of city real estate. A progressive literature has drawn attention to the active role now played by the state in the mobilization of city space for the operation of global capitalism and its elite consumption practices. These developments highlight the huge resources that are necessary for globalized capital accumulation in world cities, endorsing Sassen’s (2001/1991) view that, in the parts of the city where major structures of the world economy are located, a “strategic new spatiality” is constructed. They also demonstrate that the wording of cities is at least partially cogenerated by political interventions. So, although Massey underemphasizes the role of London’s network relations, her call for a more progressive politics of space which looks “both within and beyond the city and [holding] the two things in tension” is pertinent (2007, 191).

The mobile world city

Nevertheless, world urbanization, increasing market liberalization, and economic integration ensure major transformative world city realignments and, as foreseen by McKenzie in 1927, “new centers of dominance are arising.” Twenty-first-century world city geographies are being reshaped in two important ways.

First, new strategies adopted by global advanced producer services are increasing the connectivity of many places in the world city network, leading to a dramatic global repositioning of some cities. Rapidly globalizing cities in emerging economies now accommodate a large number of strategic functions and are catching up with the “big three” global cities of the late twentieth century: London, New York, and Tokyo.

Second, as these rising world cities are gaining important global roles and economic status, a new landscape of functionally interlinked world city-regions is emerging. The “great urban regions” first described by Hall as world cities are subject to diverse development processes with morphologically “monocentric” regions, such as London, developing extensive globally connected mega-city regions that feature functional complementarities and polycentricity, as described by Hall and Pain in 2006. Meanwhile, in regions long regarded as polycentric, such as the Randstad and the Rhine-Ruhr, cities such as Amsterdam and Dusseldorf have become global nodes surrounded by urban centers exhibiting sectoral specialization and lower levels of global connectivity. Thus, new geometries of global agglomeration are arising.
Box 1 World cities in the global financial services network, 2010.

1. London
2. New York
3. Hong Kong
4. Singapore
5. Shanghai
6. Tokyo
7. Paris
8. Beijing
9. Sydney
10. Toronto
11. Los Angeles
12. Dubai

Source: Analysis by Dr Sandra Vinciguerra, University of Reading, based on ESPON (n.d.) and GaWC (www.lboro.ac.uk/gawc/).

In the meantime, on the other side of the world and half a century ago, Shanghai was only a minor player on the world cities economic stage when the Yangtze River Delta (YRD) was first described as one of six giant world city regions. The city has since experienced a staggering rise in world city network connectivity and prominence under China’s Open Door modernization project. During just one decade, Shanghai has risen to fifth position and Beijing to eighth position in the world city network for financial services (see Box 1). In consequence, London and New York are now highly connected to China through their global financial and linked advanced producer services, forming a new prominent, cross-border geography of superconnectivity within the world city network. And the major, rapid development of Shanghai is being accompanied by sweeping functional and economic changes across the once urban industrial/agrarian YRD megacity region. Mobilized by transnational corporate and state entrepreneurial strategies in an ambivalent political economy landscape, the iconic skyscraper landscapes typical of Shanghai are also being rolled out across China as part of a vast metropolitanization project that will provide new infrastructures for global capitalism.

The system of world cities is undoubtedly in a state of flux and hybridization as the asymmetric spatial relations of urbanized global capitalism are redrawn in a network landscape (see Figure 1). Differential work and wealth creation, environmental, ecological, or food and water security geometries have yet to play out across evolving megacity region and rural/urban divides in the longue durée.

What can we take from this brief trawl through a range of key contributions to world cities scholarship? The world cities phenomenon is defined by a plurality of complex cross-cutting relational geographies and power geometries with equity implications. The contradictions it presents remain challenging to address. On the one hand, the conception of cities as constituted by a new logic of network flows generated by economic agents informs understanding of the cross-border integration of world cities but, on the other hand, states are complicit in the mobilization of world cities as focal points for the operation of global capitalism. The metaphorical power of the world city concept has undoubtedly been a driver for burgeoning contemporary territorial competition. A reflexive approach to interpretation that takes into account the temporalities of world city transformation in a dynamic geopolitical context is therefore clearly necessary.

SEE ALSO: Cores and peripheries; Corporate spatial organization and producer services; Global cities
Figure 1  World city network of advanced producer services, 2010. Source: Analysis by Dr Sandra Vinciguerra, University of Reading, based on ESPON (n.d.) and GaWC (www.lboro.ac.uk/gawc/).
References


World-systems theory

Salvatore Engel-Di Mauro
State University of New York at New Paltz, USA

Foundations of world-systems theory

Geographers, since at least the 1980s, have borrowed from, contributed to developing, and critiqued (Flint and Shelley 1997) what should be more precisely called the world-systems perspective. It is an analytical framework or paradigm encompassing diverse, sometimes conflicting, macro-social approaches premised on certain broadly unifying theoretical positions. It emerged in the early 1970s mainly in response to prevailing capitalist explanations of worldwide inequalities, among other related processes, in a context of openings and heated debates within institutional or party-line Marxism (especially over the concept of “ Asiatic mode of production” and the causes of the feudalism-capitalism transition). In fact, the critique challenged both capitalist and predominant Marxist approaches because it pointed to the inadequacy of stage-based formulations of progress (especially for non-European histories) and the national-scale emphasis (the exclusive use of internal processes to explain social change) shared by mainstream capitalist and leftist theories and politics. The alternative offered, the world-systems approach, crystallized through a paradoxical confluence of neo-Marxist perspectives, especially from within the Latin American Dependencia (Dependency) School, and the partly anti-Marxist history approaches of the French École des Annales (the Annales School).

Elaborating on the unequal exchange thesis developed within the UN Economic Commission for Latin America, neo-Marxist Dependencia proponents, during the 1960s, studied the effects of external imperial forces from the metropole (core) on prospects for economic development (in the modernizationist sense) in regions like Latin America. These regions were argued to be structurally dependent on the metropole and, thereby, forced into the periphery of the world economy (Amin 1974). Some dependentistas were also directly involved in policy formulation and implementation, privileging national state control of investments and favoring such policies as land ownership reform. Various US-backed military dictatorships cut short such reform processes, especially following the murderous 1973 military coup in Chile. Many dependentistas experienced such dictatorships directly, some even losing their lives. At least some aspects of dependency theory were thereby radicalized (e.g., the proposition of delinking from the world economy).

In contrast, French Annales historians represented an academic response to what they presumed a mainly academic endeavor, that of how to study history. The 1929 establishment of the Annales d'Histoire Économique et Sociale (Annals of Economic and Social History, eventually renamed Annales. Histoire, Sciences Sociales, or Annals. History, Social Sciences) generally marks the beginning of this approach. Expressing dissatisfaction with treating history as a series of events, where formal political aspects drew most scholarly attention, these historians drew from multiple fields of knowledge, including
the environmental possibilist regional geography of de la Blache (hence, there are some geographical roots to world-systems perspectives). The Annales School alternative was a focus on material culture and long-term processes behind quotidian activities and attitudes, within what would eventually be called a world-economy composed of different societies. However, their particular brand of materialism is expressly antithetical to Marxist perspectives and largely unconcerned with power issues, especially class relations (cf. Braudel 1977). While most of its founding proponents lived through the Nazi dictatorship and one of its main figures (Marc Bloch) was executed by Gestapo officers, the Annales School perspective has not contributed to addressing coeval power relations in any direct way, such as the relationship between France and the Algerian independence struggle.

By the early 1970s, the world-systems framework combined aspects of the above-discussed schools of thought to provide the foundations for an innovative approach that situates national within broader geographical and historical processes (e.g., the world-economy of the Annales School) and attends to the relations of power behind large-scale social inequalities (e.g., the unequal exchange, core-periphery concepts from the Dependencia School). It joined a world-scale core-periphery model of uneven development with a long-term historical view to overcome narrow understandings of social change or struggles, which are still often treated as if they unfold ahistorically within disconnected countries. Implicit from the start in this conceptualization were a geographical orientation (e.g., emphasizing interlinkages between places and scale of analysis) and the quest for the unification of knowledge production modes (social and physical sciences, humanities).

Four premises guiding world-systems analysis

What emerged out of such alchemy of contrasting approaches was a set of theoretical premises guiding a general framework called world-systems analysis. First, world-systems are the most important units of analysis and general explanatory variable for social change, especially in the long term. The hyphenation in the term “world-system” actually serves to signal this explanatory focus on spatially and temporally delimited worlds, composed of multiple elements (e.g., polities, cultures) integrated within a single system. A world-system is thus defined as an entity encompassing different state-based and stateless societies interlinked and integrated through a division of labor, which implies that any form of labor counts, including unwaged labor in a capitalist world-system. For systems predicated on stateless societies, the term mini-system is employed. A world- or mini-system cannot be treated as operating independently of its scale-producing elements, which create the world- or mini-system in the process of interaction. The analytical framework is, therefore, both systemic and relational. Accordingly, the unevenly experienced historical emergence of capitalism, still the primary focus of attention, is explained as an intrinsically expansionistic system integrating other world- and mini-systems, often by force, into its periphery. Hence, what occurs in one area cannot be assumed to be independent of processes happening elsewhere (Amin 1974; Hall 2000).

Second, and a main premise of all world-systems approaches, is that social change cannot be explained through internal dynamics alone but by understanding how societies are interconnected with others as parts of a changing whole. This means that scale (at least the social one) is not static, but made through networks of social relations.
In fact, the scale of the specifically capitalist world-economy has increased historically, unlike any other system, to cover the entire physical world. The large-scale that is often referred to as “global” has been produced through capitalist expansion. Consequently, world-systems research focuses on both the effects of such world-system dynamics on the internal processes of the societies (or subsystems) that compose it (ultimately, all the human actors within it), and the influence of change within societies (subsystems) on the overall characteristics of that world-system itself (Hall 2000).

Third, and a major theory within the world-systems paradigm, is that over long historical periods there developed, out of multiple succeeding mini-systems, world-economies, and world-empires, a capitalist world-system (with a single world-economy) that marks the first planet-encompassing world-system in human history. The rise and fall of different world- and mini-systems and related multiple-scaled social changes are the product of the mutual constitution of societies (the parts) and world-/mini-system (the whole). In this manner, the insights of the Dependencia School are refined by addressing the existence and maintenance of political economic inequalities between and within countries. Wallerstein is especially interested in the long-term emergence, development, and demise of self-contained world-systems (if state-based) and, to a much lesser degree, mini-systems (if stateless) in different parts of the world. This long-term perspective, elaborated from the works of Braudel and others from the Annales School, allowed for an understanding of the specificity of the capitalist world-system and for the elucidation of the processes whereby wealth accumulation could become concentrated in some regions at the expense of others. The main mechanism, already considered by Dependencia theorists, is unequal exchange, benefiting the core countries unevenly and largely the ruling classes of the semi-periphery and periphery (Wallerstein 2011/1974, 347–357).

Fourth, a world-system (presumably not so much mini-systems) develops structurally into zones having uneven influences over the system as a whole. These zones are called core, semi-periphery, and periphery, which are relational concepts. In a capitalist context, core, semi-periphery, and periphery are spatial categories used to describe tendencies present in the world economy and the relative world-system position of places integrated therein. Core places are those characterized by capital accumulation and attendant combinations of high economic, military, and political influence in the world-system. Peripheries are impoverished places of largely imposed economic dependence on core regions and that may have marginal status in the capitalist world-system. Semi-periphery describes an intergrade world-system position of dependence on core countries and influence over some countries in the periphery.

The enormous breadth of studies undertaken under world-systems analysis spans most fields of social science, including human geographies, and increasingly the sciences concerned with paleoenvironmental change (Babones and Chase-Dunn 2012). World-systems analysis covers a vast array of topics, from explanations of long-term historical change at different scales, including indigenous peoples’ histories and world-systems, to issues of racism, gender relations, colonialism, urbanization, and warfare. Within geography, it is especially the political subfield of human geography that has benefited most from the insights of world-systems approaches, particularly through a de-emphasis of the national scale and through such concepts as hegemony cycles and regional power differentiation (core, semi-periphery, periphery) and their application in conjunction
with economic processes. In the case of the latter, analytical categories and studies of commodity chains, cyclical phases (e.g., Kondratieff waves), and phase-dependent technological innovation, among others, also markedly influenced economic geography.

Critiques and countercritiques of world-systems analysis

World-systems approaches have been critiqued since their very introduction and from diverse theoretical perspectives. One critique is the privileging of scale to the world-system level at the expense of processes at other scales. This is especially of concern to those taking individual actors and everyday activity as the primary research entry point. Through this and other lenses, world-systems analysis seems to prioritize structural explanations or to espouse another form of structuralism, rather than enable forms of inquiry that are more context-sensitive or consider agency and/or contingency. Another is the assumption that the geographical expansion of capitalist relations is coextensive with a capitalist world-system. This is related to conflating relations of exchange with interactions among modes of production. Another is the disproportionate attention to a presumed Europe-centered capitalist world-system and the virtual absence of such oppressive processes as racism, patriarchy, heteronormativity, and arguably even class struggles in world-systems analyses. Aside from these problems, there are some underlying issues in the world-systems framework that derive from problems in the Dependencia and Annales schools out of which the framework was wrought. Namely, they are the setting of system boundaries sufficiently clear for analytical purposes (both in terms of space and time) and a tendency to reduce processes to the activity of social forces influential at the world-system level (it is not so much agency in general as to whose agency), which can lead to missing world-changing processes coming through other scales of analysis. Some of the more potent charges include functionalism (tautology) and determinism.

Critiques of world-systems analysis have been in themselves problematic (Wallerstein 2011/1974, xix–xxviii). There are issues of problem misidentification and theoretical mischaracterization. One example is Laclau’s denial of the possibility of articulation of modes of production and insistence that capitalist systems must have a largely proletarian presence, which implies that capitalism is receding as wage employment becomes increasingly a privilege even in the most industrialized countries (see a rebuttal in Wallerstein 2011/1974, 127). The charge of functionalism confuses analytical categories used in world-systems theories for explanatory statements. Core, semiperiphery, and periphery and accumulation processes are analytical categories used not to explain but to describe a world-system. The content of these analytical categories has recently been challenged for their social reductionism but they have not been abandoned. They remain useful if understood through context-specific spatio-material mechanisms (Bunker 2003, 238–239). World-systems theories attempt to explain the evolving characteristics of and rise and fall of different world-systems, requiring careful and detailed historical analysis at many scales, including microsocial (see Straussfogel in Hall 2000). This theory is also contested in that some maintain that such a global world-system emerged much earlier and/or is not as Europe-centered as once thought (Frank and Gills 1993). However, in all these perspectives, the starting unit of analysis remains at the world-system scale (a historically variable unit in terms of Earth surface area and time period covered).
The imputation of determinism is similarly misplaced. Wallerstein (2011/1974), for instance, has made it very clear from the beginning that world- and mini-systems develop and change as a result of shifts within and between their components. The characteristics described and explained in world-systems research involve the understanding of co-determination of all forms of social relations, including cultural, political, and economic aspects of ethnic, class, and gender relations (Hall 2000). World- and mini-systems are also open systems, tending to be thereby dynamic, complex, and unpredictable. Far from reducing social phenomena to macroscale processes, world-systems approaches consider and attempt to integrate potentially all possible scales and aspects of social relations, cultural, political, economic, following in the footsteps of the Annales School. The historically informed accounts of world- and mini-system processes accord with the recent understanding of scale in human geography as socially constructed, as the product of power relations. The global, for example, is a recent phenomenon brought about by capitalist development and expansionism over the past 500 years or so and the construction of the global scale has simultaneously involved multiple social processes, including non-capitalist relations and Indigenous Peoples’ agency (Hall and Fenelon 2009).

There are, nevertheless, some fundamental shortcomings in world-systems perspectives, though not necessarily in terms of the theoretical framework itself. For example, studies are still rare that treat as fundamental to world-system dynamics the processes of, for example, gender and heteronormativity. However, since the 1980s there have been ways in which, for example, household and patriarchal processes have been fruitfully analyzed in relation to resistance to neoliberal policies or to colonialism, including indigenous peoples’ struggles, to furnish more effective explanations of world-system processes (Dunaway 2001). Similarly, and despite early works on environmental degradation, most world-systems research does not address biogeophysical processes, although much work has emerged to redress this situation. In part, the problem lies not just in the framework itself, but in social science approaches generally. As Bunker (2003) pointed out, the socially-based, analytical categories of core-, semiperiphery, and periphery must be defined according to context-specific spatio-material mechanisms (e.g., the highly uneven spatial distribution of metalliferous ores and the variable material conditions for their extraction, processing, use, and transport). He therefore called for a “new historical materialism.” Work is being done to redirect world-systems research agendas towards ecological processes with human world-systems as component parts of major interest. Recent advances hold much promise in explaining the interrelations between larger-scale social and environmental dynamics and smaller-scale phenomena. For example, the combination of world-system analysis with insights from environmental sociology has been increasingly adopted to explain shifts in global to local anthropogenic environmental impacts, but the work has really just begun (for newer developments in the wide variety of world-systems analyses, see Babones and Chase-Dunn 2012).

SEE ALSO: Annales School; Colonialism, decolonization, and neocolonialism; Cores and peripheries; Dependency theory; Development; Geopolitics; Global commodity/value chains; Globalization; Scale

References


Zoogeography

Michael N. DeMers
New Mexico State University, USA

Zoogeography is the subdiscipline of biogeography that aims to explain the distributions of animals and, through doing so, to further explain the relationships of these distributions to other geographic phenomena. The science differs from the broader biogeography in that it focuses on fauna rather than flora. This distinction is important because fauna have the capability to adapt not only by adaptation but also by migration to habitats that are more conducive to species survival. There is also a significant difference between traditional zoogeography, which focuses on the patterns of zoological phenomena and emphasizes historical as well as environmental distributional factors, and ecological zoogeography, which tends to restrict its examination to ecological factors. While the distinction is there, it is becoming more blurred with time.

The historical roots of zoogeography are inextricably linked with those of ecology and the research and musings of Ernst Haeckel, who not only introduced ecology in his general zoology text (1866) but also gave a new name and new content focus for what would eventually become zoogeography. His view was that the spatial distribution and organization of organisms should be a new discipline. He called that discipline chorology.

According to Udvardy (1969), chorology included the description of not only geographic and topographic habits and the distributional limits of species but also the elevational and bathymetric zonation of species. Interestingly, Haeckel (1866) believed that, while biogeographers of the past had studied these factors, they lacked the necessary underpinning of Darwin's evolutionary theory to explain their findings. In effect, Haeckel was emphasizing the zoogeographical method as a means of understanding biological phenomena.

Alfred Russel Wallace, a contemporary of Charles Darwin, developed his own theory of evolution independently and was the first to establish a clear distinction between the biologist and the zoologist. His Geographical Distribution of Animals (1876) identified the zoologist's work as "geographical zoology" and the geographer's as "zoological geography." For geographical zoology he advocated a classification of the distribution of the world's animals by taxonomic groups, but still made an allowance for biological interactions and historical factors (Udvardy 1969). Wallace's book was the definitive textbook on zoogeography until the mid-1950s.

Like today's zoogeographers, Wallace did not limit his work to extant species. He included evidence derived from the fossil record to discuss mass extinction, evolution, and migration which resulted in the geographic distribution of modern animal species. Wallace was well aware of the mass extinction of megafauna in the latter portion of the Pleistocene. This led him to believe that we are currently in a "zoologically impoverished world, from which all the hugest, fiercest, and strangest forms have recently disappeared" (Wallace 1876, 150). He eventually believed that the most likely cause of the rapid extinctions was not so much glaciation as the impact of humans. Such working hypotheses, and opposing hypotheses regarding species extinction, are also part of modern zoogeography.
ZOOGEOGRAPHY

Zoogeographic regions

Using the zoological geographic approach, Wallace was among the first to identify the distinct regionalization of the Earth’s fauna. This regionalization was based on very few species and partly on his famous study of the Malay Archipelago, in which he identified unique faunal assemblages in the Indonesian islands of Bali and Lombok in the south and of Borneo and Sulawesi in the north. This line separated what he defined as the Asian (Oriental) and the Australian zoogeographic regions. Later named Wallace’s Line (Figure 1), it clearly showed marked differences in fauna in islands that were a mere 20 miles apart. It is also part of the general zoogeographic regions (more accurately, realms) which, at the time, in addition to the Oriental and Australian, include the Nearctic, Palaearctic, Ethiopian, and Neotropical realms (Figure 2). The appeal of this system for geographers is that it coincides well with the limits of continents or major topographic regions. Unlike biomes, which are defined exclusively by ecological factors, zoogeographic realms are delineated by the presence or absence of particular assemblages of animals that are then explained by ecological or other factors. In a way, they are two sides of the same coin.

The general structure of the zoological regions and the methods by which they are derived have changed over the years, and many attempts have been made to make it less subjective and more quantitative. These continue to the present, and consider ancestral species relationships but seldom focus on explicit phylogenetic relationships. Phylogenetic trees are increasingly available for more and more species, which contain considerable additional information that is useful for grouping species assemblages into zoogeographical regions at a global scale. This approach clearly demonstrates that the original concept of zoogeographic regions devised by Wallace is still a focus of research. With increasingly robust datasets, including phylogenetic and phylogeographic data, zoogeographic regionalization research is likely to continue for some time to come.

Dispersal and migration

While analyzing zoogeographic regions as they exist today is one focus of the geographer’s research, the changes in distributions of individual species is also an area of research common to geographers and zoologists. Ecological zoogeographers are more likely to focus on dispersal, the spreading of individuals from their parents, siblings, or social group. Geographical zoogeographers are more likely to follow the approach of Wallace in that they look at movements of large numbers of animals from one place to another, often abandoning original environments in favor of more ecologically suitable ones. This, of course, results in different zoogeographic patterns that intrigue the geographer as they search for causal factors or, alternatively, for the impacts of these migrations on the new habitats. These pattern changes can be examined over any temporal framework, involving recent environmental changes or paleoenvironmental changes. Such studies often include a range of such changes from short-term ecological forcing functions to those involving long-term climate change and continental drift.

In examining individual dispersal or large-scale migration, some general factors play a role. Forcing functions – external factors that cause species to migrate seasonally, relocate, or become locally extinct or extirpated – and large-scale environmental changes that either displace whole faunal assemblages or cause their extinction are topics of concern to the zoogeographer. Each
species also exhibits its own level of environmental plasticity. For example, the cockroach has adapted to a symbiotic relationship with humans; ants seem to be able to survive a wide range of environmental conditions; and locusts adapt to changes in food availability and landscape patchiness by both changing their physical form and adopting a social structure for survival. Once a species is forced to relocate, it must face obstacles to relocations such as physical barriers (e.g., streams, lakes, or mountains); distance barriers, whereby the nearest compatible habitat might be thousands of kilometers away; or ecological barriers, where there simply is no compatible habitat at all. Each of these is a concern to the pure zoogeographer hoping to understand shifting patterns, and to the applied zoogeographer who uses this knowledge to plan for environmental change.

The methods by which species disperse also vary, and are often affected both by their mode of locomotion and by the physical features they encounter. Generally there is a steady movement or diffusion into adjacent areas because of proximity, or a jump dispersal may take place, often suddenly, thousands of kilometers from the original habitat. Flying species, for example the cattle egret, are best adapted to jump dispersal because of their mode of locomotion and their ability to cover large distances and overcome barriers. Dispersal can also be aided by corridors (e.g., roads, riparian areas, habitat edges), which act very much like highways for the movement of species. Closely spaced islands or habitat patches
Wallace’s map showing the zoogeographic regions of the world.
may act as stepping stones for this type of dispersal as well. Whatever the mechanism, dispersions may also be classified as invasions if the habitat is already occupied. This results in the possibility of niche overlap and conflict, which may result in the extirpation or elimination of an existing species, especially if the invading species is exotic. The devastating simplification of insect fauna as a result of the mass migration of fire ants into the east coast of the United States is one example. From just these few examples, it is clear that an understanding of zoological displacement is critical to conservation and land management efforts.

**Areography**

The importance of the environment to dispersal is part of a broader topic called areography, which examines how species interact with the geographical areas in which they reside. Interestingly, Rapoport (1982) uses the term “areography” synonymously with “chorology,” which was an early term for zoogeography itself. The primary topics of areography examine the impacts of the shapes, orientations, and dimensions of geographical features on the organisms that inhabit them; longitudinal impacts on species diversity, range compaction, and delimitation and compaction of species range areas. Beyond the identification of these geographic patterns, areography attempts to create new methodologies for their quantification and analysis.

Among the topics most relevant to zoogeography is the peninsular effect, which describes how species richness decreases as the shape of a peninsula narrows. This is the result of a reduction in the amount of habitat. Areography studies the methods of delineating and compacting species ranges to most accurately represent both location and species integrity. It evaluates geographical range and distance in relation to the number of species that are likely to occur. As suggested, the barriers to dispersion are also a topic of interest to the areographer and include examination of pest species.

Among the most well known of concepts to arise from areography is Rapoport’s rule, a hypothesis that the latitudinal range of species is smaller at lower than at higher latitudes. Among the explanations for this phenomenon is the suggestion that at very high latitudes species are restricted as a result of glaciations having wiped out species that would normally have narrow ranges and climatic variability, while areas with seasonal variability accommodate species with greater climatic tolerance and therefore a wider latitudinal range.

**Species adaptation**

Because species vary widely over the Earth’s surface, it stands to reason that they will encounter wildly different persistent environmental conditions. By extension, it stands to reason that such environmental differences result not only in different zoogeographical regions, but also in the species themselves, over time, adapting to such conditions through the evolutionary process. One ecogeographic principle associated with this, called Bergmann’s rule after the German biologist who described the pattern, holds that the body mass of birds and mammals (endotherms), and apparently other taxonomic groups as well (e.g., the ant *Leptothorax acervorum*), tend to be larger and bulkier in colder regions and smaller in warmer regions. Apparently, this is an adaptation of species as a way of conserving or of radiating body heat, depending on the climate. The degree to which Bergmann’s rule holds true is still a topic of zoogeographic research.
Animal species have also adapted in other ways to different environmental conditions. Allen’s rule, for example, indicates that body shapes and proportions in endotherms adapt to climatic temperature regimes either by minimizing exposed surfaces or by maximizing surfaces, depending on whether the adaptation is meant to minimize or maximize heat loss. Examples of this include long and thin appendages (ears, tails, limbs, snouts, etc.) in hot climates and shorter and thicker appendages in colder climates.

Other adaptations to geographic factors include the ability to store food and water in deserts, as in a camel’s hump; a massive layer of fat and thick insulating fur which allow polar bears to endure extreme cold; and the long arms of monkeys in tropical rainforests to allow them to use the dense vegetation as a means of locomotion and to avoid ground predation. Adaptations may also involve behavior rather than physiological modifications, such as the development of social structures (e.g., wolves), aestivation and hibernation during seasonal extremes, or seasonal migration to follow food sources. Species adapt to changing conditions through physical appearance (e.g., the British peppered moth) or even how they live within their environment (e.g., nest orientation to avoid harsh winds). An awareness of these adaptations not only informs the geographer about the ecology of an area, but may also explain how variations in zoogeographic regions affected by climate change can take place.

Zoogeography and evolution

Geography has an impact on the processes and patterns of evolution and speciation. The adaptations discussed in the previous section are likely a result of such evolutionary processes. It is generally understood, for example, that small populations that are geographically separated from their original parent population have a smaller composite genetic makeup. Because of the smaller genetic makeup, such founder populations frequently experience more genetic mutations, thus encouraging speciation when the mutations are advantageous in the new environment. This concept is known as the founder principle and the result is called the founder effect. Alternatively, the part of the population that remains also has a smaller genetic composition, called a bottleneck, which can also result in the increased likelihood of speciation for the same reasons.

In some cases, particularly where the splitting of continents is involved, large populations can become geographically isolated. These changes are called vicariance events, and the speciation that can result is the subject of vicariance zoogeographers. As with the founder effect, the isolation of the portions of the populations impacts the genetic composition, thus encouraging speciation. Given the much larger size of the populations, the process of speciation is likely to take a considerably longer time than those developing via the founder principle.

While the founder principle and vicariance zoogeography explain the divergence of one species into two, cases exist where speciation produces more than one new species. This process, called adaptive radiation, results when species evolve from a common ancestor to fill all environmental niches in the newly colonized region. Among the classic examples of adaptive radiation are the marsupials of Australia, which evolved to occupy niches occupied by placental mammals on other continents.

These results all suggest that the species survive through some form of genetic mutation. The alternative is that there is a loss of all individuals from taxa, a process called extinction. Extinction can be local, occurring when a species or taxon
disappears from one or more geographic locations but persists in others. Global extinctions refer to the elimination of an entire taxon throughout its range. The American bison is an example of a local extinction, while the passenger pigeon is a classic example of a global extinction in that these species no longer exist. Interestingly, both of these examples are fundamentally a result of human intervention, a factor that will play a larger role in zoogeographic research as time goes on, together with the impacts of rapidly changing climates.

Island biogeography

Perhaps one of the most well-known and among the first ever attempts at employing the experimental method in zoogeographic research was that of MacArthur and Wilson (1967), who focused on the determination of species numbers on a newly created island. Their theory was that the number of species on an undisturbed island would be a function of immigration and extinction. These processes are impacted by the distance of the source of immigrants from the target island. This distance effect suggests that the more remote the target islands are from the source areas – usually nearby continental mainlands and nearby islands – the fewer the immigrants.

The other component of island biogeography theory holds that the rate of extinction of a species that has colonized the island is related to the area of the island on which it established itself. This relationship, called the species–area curve, takes the general form illustrated in Figure 3. This is well established as a general theory; however, there are modifications to the original. For example, the isolation affects not only immigration rates but also extinction rates. Populations that are less isolated are less likely to go extinct because more individuals from the source population continue to add to the population. This is called the rescue effect. Likewise, while the size of the island affects the extinction rate, it also has an impact on immigration rate. The target effect suggests that species may actually select larger islands over smaller ones because they have more resources and available niches; or the larger islands may simply acquire more species and more individuals of each species simply because they are larger.

In addition to these modifications, there are other influencing factors, such as the degree of isolation (distance to mainland or nearest neighbor), time of isolation, and size of island where larger islands frequently result in greater species diversity. The habitat suitability of the target island also modifies the general theory, especially with regard to climate (e.g., tropical versus arctic); initial plant and animal composition if, for example, they were previously attached to a larger landmass; and the current species composition. Additional influencing factors include location relative to ocean currents that influence the availability of nutrients, fish, bird, and seed flow patterns. Serendipity – pure chance – obviously modifies the likelihood of species arriving or the interactions of other factors. Finally, humans produce often profound modifications of the original theoretical model.
ZOOGEOGRAPHY

Macroecology

Brown and Maurer (1989) coined a new term, “macroecology,” to describe what they considered a new discipline which focused on the study of the relationships between organisms and their respective environments at large spatial scales. For those who examine animals primarily and do so at large scales, this acknowledges the relationship between zoogeography and macroecology. The differences are useful for both disciplines. The focus of macroecology is on an understanding of the mechanisms underlying patterns of biodiversity at large spatial scales. The questions likely to be examined by the macroecologist include how changes in global climate impact wildlife populations and their distributions, the impacts of environmental conditions on species richness and diversity, and latitudinal gradients in species diversity. The typical research agenda of the macroecologist is surprisingly similar to that of the traditional zoogeographer but with a distinct ecological emphasis.

Conclusion

Zoogeography is the science of the distributions of the Earth’s fauna. Regardless of the name under which it is pursued, the field ultimately focuses on the distribution, movement, evolutionary impacts, and ecology of animals as they are distributed around the Earth. While the term itself is less often used today, its study continues and increases in importance, particularly with rapid changes in climate and anthropogenic influences on animal habitats throughout the Earth.

SEE ALSO: Biogeography: history; Ecogeography/macroecology (range and body size); Geography of evolution; Island biogeography; Phylogeography and landscape genetics

References


Further reading


Zoogeomorphology

David R. Butler
Texas State University, USA

Biogeomorphology, a subdiscipline of the field of geomorphology, examines how geomorphic processes (e.g., glacial ice, running water) affect the distribution of plants and animals, but also studies the effects of plants and animals on the landscape. Zoogeomorphology is a subfield of biogeomorphology that specifically focuses on the study of animals as geomorphic agents. The term “zoogeomorphology” was coined and defined by Butler (1992) as the study of animals as geomorphic agents, using the impact of grizzly bears on mountain slopes in Glacier National Park, Montana, as an example (Figure 1). That paper examined two different digging processes associated with grizzly bears, food excavation and seasonal hibernacula creation, from which a minimal annual sediment budget associated with bear excavations was calculated. Three years later, the book Zoogeomorphology – Animals as Geomorphic Agents (Butler 1995) was published. That book synthesized research on the geomorphic impacts of hundreds of species of animals, both invertebrates and vertebrates, into one volume. A related concept, ecosystem engineering (Jones, Lawton, and Shachak 1994), developed contemporaneously in the ecological literature. Zoogeomorphology studies the roles of animals in landscape formation and decay, whereas in ecosystem engineering ecologists study environmental change caused by biota and the consequences for the engineer, other organisms, and ecological processes. Not all ecosystem engineering has geomorphic impacts, however; a bird’s nest constructed entirely of twigs is an example of ecosystem engineering, but without direct geomorphic impacts.

Animal impacts on the landscape

Animals geomorphologically affect the landscape in a variety of ways, including trampling, wallowing, digging for food, excavation of burrows and dens, construction of mounds, ingestion of soils (geophagy) and rocks (lithophagy), mixing of soils, and even stabilizing sediments so that they are not eroded. Naylor (2005) categorized the suite of biogeomorphological processes, including those carried out by animals, as bioerosion, biostabilization, and bioconstruction. Statzner (2012) examined the role of a variety of animals that live in streams and provided examples of both sediment stabilization and sediment destabilization and subsequent movement downstream. This work, based on many years of field and laboratory experiments, is a thorough and meticulous presentation. The geomorphic impacts of animals may be direct or indirect. Direct geomorphic impacts include the aforementioned processes of burrowing, digging, and other processes whereby animals are directly moving soil, sediments, and/or rocks on the landscape. Indirect geomorphic impacts include the effects of trampling, whereby soils are compacted and surface infiltration of water is reduced such that runoff rather than downward percolation results.
The effects of trampling by domestic animals were examined in detail by Trimble and Mendel (1995). They illustrated how trampling produces compacted soils that in turn preclude infiltration and instead produces rapid runoff and erosion. They also described direct geomorphic impacts in describing the chiseling effects that cattle hooves have along stream margins where large amounts of sediment can be carved off and introduced into, and subsequently removed by, streams. Butler (1995) described a variety of large mammal species, including bison, elk, bighorn sheep, mountain goats, elephants, rhinoceri, and hippopotami, that create distinctive trampled trails on the landscape as well as areas of sediment removal associated with mud wallowing and resting in day beds. Some areas of localized trampling and wallowing are associated with natural salt outcrops where animals come seasonally to ingest salt-rich sediments (geophagy).

Soil mixing, or biopedoturbation, typically is accompanied by some form of mound production (Richards 2009; Wilkinson, Richards, and Humphreys 2009). Butler (1995) illustrated numerous examples of biopedoturbation with associated mounding, including the work of ants and termites, crayfish, mound-building birds, and a variety of mammals, including (but not limited to) gophers, marmots, and prairie dogs. Richards (2009) discussed the astonishingly high rates of sediment bioturbation produced by a specific genus of ant. Wilkinson,
Richards, and Humphreys (2009) offered an extremely thorough review of biopedoturbation studies, covering studies from the late 1800s to the mid-2000s. Burrowing for shelter, food acquisition, and production of birthing chambers also produces bioturbation as well as extensive underground tunnel systems and surface spoil mounds at the mouths of burrow entrances (Figure 2). Such tunnel systems in turn may profoundly affect surface drainage, capturing, and localizing water runoff into underground passageways. The burrowing behavior of birds has been studied in association with colonial seabirds that employ burrows for birthing chambers as well as protection from the elements. Entire islands (Butler 1995, and references therein) have been honeycombed and undermined by such burrowing to the point where some islands virtually sink into the sea.

Geomorphic impacts of beavers

One of the most significant zoogeomorphological agents is the beaver – in North America, the Canadian beaver, *Castor canadensis*, and in Europe, the European beaver, *Castor fiber*. Many more Canadian beavers currently exist than do European beavers, but historically their influences may have been equally important. Butler (1995) presented an entire chapter devoted to the geomorphic effects of beavers that is a useful and accessible entry to the topic and the literature pre-dating the mid-1990s. Known for their unique ability to construct dams, beavers and the variety of dam types they build are the focus of a short but very useful study (Woo and Waddington 1990) that also illustrated the effects of beaver dams on water retention and evaporation rates.
Gurnell (1998) offered a valuable examination of the specific hydrogeomorphological results of beaver-dam construction, as well as a valuable literature review. Beaver dams and associated pond creation (Figure 3) entrap water and therefore drastically reduce stream velocity, which in turn induces sedimentation over years to decades within individual ponds. The creation of a pond also elevates the local water table, with attendant ecological benefits for riparian species. The reduction in stream velocity results in reduced stream erosion and entrenchment downstream of beaver dams. In a variety of locations in the American West, beavers have been reintroduced into environments where they had been historically removed by trapping in an attempt to assist in reducing stream erosion and entrenchment and to increase and enhance riparian habitat.

Beaver meadows have been described in the literature since the 1930s and were briefly summarized by Butler (1995), but the specifics of the sedimentation necessary for their construction have not been clear. Polvi and Wohl (2012) showed that beaver meadows in Colorado contain sediments laid down over thousands of years and that those meadows are comprised

Figure 3  A beaver dam, approximately 1.5 m high, in Glacier National Park, Montana, USA. The pond impounded behind the dam was approximately 1.3 m deep when examined in the field. A second, downstream, pond is visible at lower left; that pond was impounded by a second dam out of sight, approximately 50 m downstream. Both ponds drained catastrophically in 1995 after a heavy thunderstorm over the adjacent mountains.
of sediments equally derived from stream sedimentation and from sedimentation in beaver ponds.

The geomorphic effects of European beavers are increasingly examined in the literature as that species continues to expand its range. Nyssen, Pontzeele, and Billi (2011) examined the role of reintroduced European beavers in the Belgian Ardennes and illustrated how their reestablishment is altering stream hydrology throughout the area. The geomorphic effects of the European species of *Castor* are basically the same as those of the North American species, that is, dam building and pond retention with attendant effects on riparian habitat and local water tables, reduction of stream velocity and discharge downstream of dams/dam sequences, and canal and bank burrow excavation. Sedimentation rates and patterns in European beaver ponds also seem similar to those described from North America, based on a limited number of studies to date.

Beaver dams occasionally fail catastrophically during periods of intense rain and/or rapid snowmelt. The failure of one dam can lead to a cascading effect, washing out additional dams downstream, that is, causing a domino effect. In North America, several examples exist where outburst floods from beaver-dam failure have washed out highways, clogged culverts leading to additional flooding, and washed out railroad bed sediment leading to train derailments. Several fatalities have resulted from washouts and floods associated with beaver-dam failure.

**Future zoogeomorphology**

Climatic change and human impacts have profoundly altered the zoogeomorphic signatures across a variety of landscapes and will continue to do so into the foreseeable future. As climates change, animal ranges also necessarily shift. Zoogeomorphically active animals may see their range reduced or even disappear; conversely, other active zoogeomorphic agents may see a range expansion. Very little research has been carried out in assessing how climate change will affect zoogeomorphic processes in terms of distribution or severity.

Human impacts, through human-caused drastic reductions in animal populations, through reduction of animal habitat via changing land uses, or by introducing zoogeomorphically active species into habitats where they escape and expand unchecked, need further examination as well. Certainly the range and strength of the zoogeomorphic impact of North American bison have been drastically reduced by human action; similarly, prairie-dog “towns” cover only a minute fraction of the geographic range they covered prior to European contact. In Australia, the introduction of non-native species has led to dramatic shifts in zoogeomorphic impacts. The introduction of the rabbit to Australia not only resulted in drastic overharvesting of vegetation in the continent’s semiarid and arid interior, but the burrowing actions of the rabbits have had impacts on surface runoff by channeling overland flow into underground tunnel systems.

**SEE ALSO:** Animal geographies; Biogeography; Biogeomorphology; Disturbance in biogeography; Environmental hazards; Fluvial depositional processes and landforms; Geomorphic hazards; Geomorphology: history; Mountain geomorphology; Treeline ecotones; Zoogeography

**References**

**Zoogeomorphology**


**Further reading**

